# Campaign Strategy in an Age of Information Abundance: Data-Driven Campaigning for Congress

by
Levi Bankston

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy
(Political Science)

at the
UNIVERSITY OF WISCONSIN–MADISON
2023

Date of final oral examination: 07/12/2023

The dissertation is approved by the following members of the Final Oral Committee:
Barry C. Burden (Chair), Professor, Political Science
Eleanor Neff Powell, Associate Professor, Political Science
Katherine J. Cramer, Professor, Political Science
Kenneth R. Mayer, Professor, Political Science
Joshua P. Darr, Associate Professor, Communications (Syracuse University)

# Abstract

Many think big data transformed politics. Both political observers and scientists have fixated on cutting-edge innovations in campaign data and technology coming from the top of the ticket. Numerous books and articles detail the influx of individual-level voter records compiled into large-scale databases that enabled high-resourced presidential campaigns to microtarget their outreach messages at smaller and smaller segments of the electorate. Existing scholarship assumes that these presidential practices have diffused down the ballot to reshape how lower-level campaigns communicate with voters. Yet no study to date tests these claims.

This dissertation expands our understanding of data-driven campaigning by providing the first comprehensive overview of the electoral information environment. This project reveals how new sources of information have not fundamentally altered electoral politics. Even equipped with highly detailed information on voters and advanced statistical models, most campaigns lack the resources to engage in highly personalized outreach efforts and still must address strategic considerations that have long-defined politicking. I argue that the arrival of large voter databases has increased the efficiency of voter outreach activities but exacerbated longstanding tendencies to reduce voters to nothing more than electoral math.

Central to my contention that individual records intensify strategic considerations are consistent party-level differences in how campaigns interact with sources of electoral information. This dissertation uncovers how Democrats and Republicans operate in vastly different data environments. Democrats not only share a party-wide voter database but also a mutual data culture that reinforces an approach to campaigning concentrated on leveraging data to maximize the return on investment for their outreach activities. Republicans meanwhile have not coalesced around a single platform or approach to data-driven campaigning. These observed differences lead to a divergence in campaign-level data preferences. Republicans continue to prefer the inferences coming from traditional polls and surveys, while Democrats default to accessing individual-level data on voters.

To make these claims, this dissertation combines conversations with campaign professionals with careful analysis of millions of spending records made by thousands of campaigns for the U.S. House of Representatives between 2006 and 2018. These conversations help make sense of my findings and ensure my results reflect the realities of contemporary campaigning. I undertake an extensive review and refinement of this substantial number of records to recover campaign-level spending on data and outreach that, until now, have been unavailable to scholars. With these verified spending records, I provide the first thorough examination of electoral information marketplaces and campaign-level spending patterns over a period of marked technological change.

# Table of Contents

# List of Tables

# List of Figures

# Acknowledgments

No one achieves anything alone. This dissertation is the product of all the people who supported me. I am indebted to you.

My Ph.D. would not have been possible without Barry Burden. I don't think I could've had a better advisor. He never delayed giving me detailed feedback or empathetic advice. We worked together on research, in the classroom, and at the Elections Research Center. He taught me how to be a researcher and a communicator.

This dissertation succeeded because of the rest of my committee: Eleanor Powell, Katherine Cramer, Kenneth Mayer, and Joshua Darr. Ellie took the time to provide invaluable chapter-by-chapter feedback. Kathy and Ken helped me think about the bigger picture and the broader implications of my research. And Josh made sure undergraduate Levi lived up to Ph.D. standards.

This project also directly benefited from 14 anonymous individuals who took the time to speak to me about campaigns and voter data. I would also like to acknowledge OpenSecrets and their work maintaining the data that served as the foundation for my analyses. Thanks also to the Department of Political Science and Elections Research Center for financial support.

I am also grateful for colleagues past and present who contributed to my success. My work with Kathleen Searles and James Garand along with their guidance while an undergraduate at LSU set me up to achieve. This dissertation also benefitted from feedback at the American Politics Workshop. Thanks in particular to Blake Reynolds and Sabrina Roof who served as discussants.

Grad school would have been insufferable without good friends: Andrew McWard, Philip Bunn, Marcy Shieh, Alison Garfield, Peter Tirella, Sunny Cho, Doug Lewis, Monica and Evan Busch, and everyone else I met in Madison.

There's also no chance I would have received this title without my fiancé Chelsea Thibodeaux. She reassured me in doubt. Questioned me when I was nitpicky. But never stopped loving me.

My journey started with the love of my family. My grandfather, grandmother, and sister raised me, and they always supported me.

# 1 Framing Voter Data and Campaign Strategy

On June 18, 2017, ironworker and Democratic union activist Randy Bryce launched his bid to unseat Speaker of the House Paul Ryan for Wisconsin's first Congressional district in the southeastern corner of the state with a YouTube video. The video juxtaposed President Donald Trump congratulating Ryan on the successful House vote to repeal the Affordable Care Act with Bryce comforting his ailing mother as she describes her battle with multiple sclerosis. "It's like hot knives going through you," his mother says while noting the thousand-dollar price tag for one of her medications. "I think it's time – let's trade places, Paul Ryan," Bryce says at the end of the video with an image of him standing at a worksite wearing his hard hat and denim jacket adorned with his union's logo. "You can come work the iron, and I'll go to D.C."[1] The video quickly went viral garnering hundreds of thousands of views and netting hundreds of thousands of fundraising dollars for his campaign within its first two weeks. Bryce, embracing the moniker "Iron Stache," then made various national media appearances and toured the country holding fundraisers and raising millions of additional dollars to challenge the two-decade incumbent.

Following his announcement, the Bryce campaign decided it wanted to know what voters thought about their candidate and his opponent. They commissioned a poll of likely voters in the district in August and a second follow-up poll later in December, totaling around $18,000. Releasing the results of the December survey to media outlets, the poll had the incumbent Ryan up by only six points at 46 percent compared to Bryce capturing 40 percent of the hypothetical election day vote. The poll also revealed that nearly 80 percent of voters said they did not know

[1] https://www.youtube.com/watch?v=F6zAyPRbels&ab_channel=RandyBryceforCongress

enough about Bryce to have an opinion about him, yet 50 percent of voters said they had an unfavorable view of the Speaker of the House, ten points lower than the August poll. Bryce had a positive interpretation of what the poll meant. "The shine has completely gone from Paul Ryan in the district," he told a reporter about the polling results.[2] Despite being in a district where Ryan had won reelection the previous year by a 35-percentage-point margin, the Bryce campaign believed Ryan's sinking approval paired with the opportunity to introduce their Iron Stache candidate as a viable challenger might just make the speaker more vulnerable than many suspected.[3]

At the same time the Bryce campaign invested in polls to track how support shifted in the electorate, his campaign also paid thousands of dollars each month to maintain access to a big voter database provided by NGP VAN, the leading provider of data and campaign technology among Democratic candidates. Access to individual-level data and, importantly, donor-level data would help Bryce continue his record-breaking fundraising levels from small donations from over 100,000 individuals. He would pair his campaign's donor maintenance with voter outreach programs that the campaign planned and managed with the Voter Activation Network (VAN) software. As the software does for nearly every Democratic campaign across the country, the VAN would help Bryce and his staffers and volunteers figure out whose door to knock on, phone to call, or address to send a letter to by identifying likely Democratic primary voters.[4]

Bryce said he planned to win his primary election against Democrat Cathy Myers, a school board representative, and general election against the House speaker door by door and vote by

[2] https://news.yahoo.com/paul-ryan-faces-hard-hat-wearing-latino-democratic-challenger-011950958.html
[3] https://www.politico.com/story/2017/12/10/paul-ryan-reelection-democrat-internal-poll-randy-bryce-288639
[4] https://www.vice.com/en/article/8xva7p/can-a-union-man-with-dollar14-million-take-down-paul-ryan

vote. Yet his grassroots strategy would also receive outside support. The Democratic Congressional Campaign Committee selected Bryce for its "Red to Blue" program,[5] adding more fuel to the candidate's fundraising fire. Senator Bernie Sanders would also endorse Bryce, making him the only candidate to receive both the DCCC's official backing and the former presidential contender's progressive endorsement.[6] The union activist who had previously lost bids for the Wisconsin State Assembly and Senate was well-positioned to defeat his Democratic primary opponent. While his opponent Myers had also raised over one million dollars in a year seen as a referendum on Trump's Republican Party, she still trailed Bryce who would spend millions just on television ads inside and outside of the district.[7]

The dynamics of the race would shift in April when Paul Ryan announced his retirement four months before the partisan primary. Without Ryan's incumbency advantage and the ten million dollars in his war chest, the chances of a Democrat winning Wisconsin's first district became slightly better but still a longshot in the Republican-leaning district that Trump won by ten points the previous cycle. The Ryan camp made sure to clarify that his departure had nothing to do with Bryce, citing their internal polling showing a 20-point Ryan victory.[8] Replacing Ryan would be Bryan Steil, a corporate attorney and an appointed member of the University of Wisconsin's Board of Regents. Steil, who had worked for Ryan in the past as his legislative aide, immediately received the former speaker's endorsement. Although a field of Republican primary opponents emerged, Ryan's previous employee quickly solidified support among party elites and primary voters, raising hundreds of thousands of dollars from donations large and small and outpacing his

[5] https://www.wispolitics.com/2018/dccc-chair-lujan-recognizes-randy-bryce-as-part-of-red-to-blue-program
[6] https://www.newyorker.com/news/news-desk/the-fall-of-wisconsin-and-the-rise-of-randy-bryce
[7] https://www.cbs58.com/news/randy-bryce-running-ads-outside-of-wisconsin
[8] https://www.jsonline.com/story/news/2018/04/11/paul-ryan-retirement-news-wisconsin-house-congress-speaker-janesville/506737002/

Republican opponents.[9]

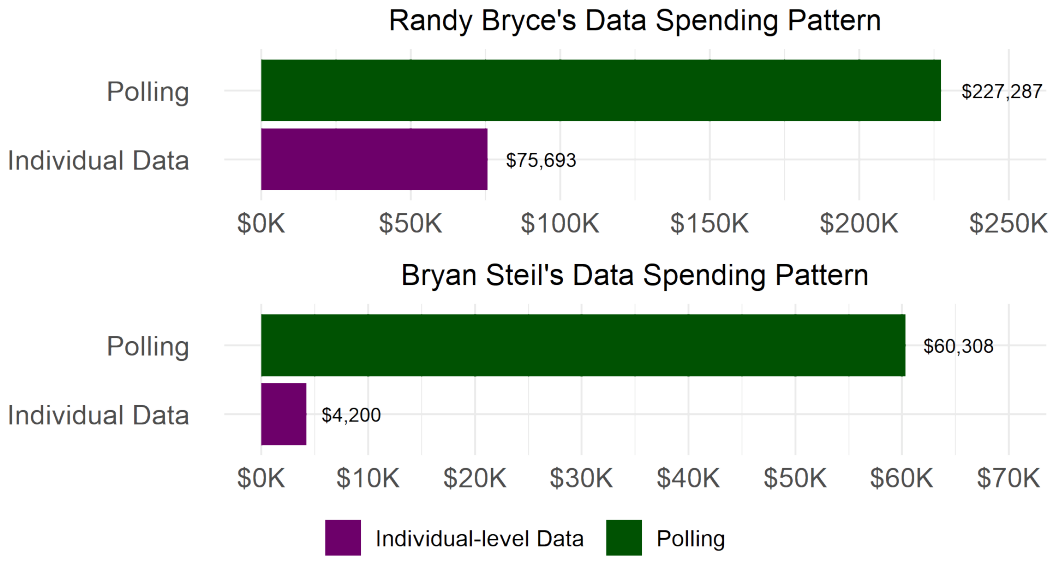
Steil would spend very little on learning about primary voters compared to Bryce's purchases in the Democratic contest. Political commentators at the time described the 37-year-old Republican as the clear frontrunner in the primary and the only serious contender.[10] Steil was just as confident as the onlookers. The Republican running to be Ryan's replacement did not field a single survey prior to the primary date, according to spending records he reported to the Federal Election Commission (FEC). The only purchase he made related to obtaining data on the district was $1,800 for access to a database from the nonpartisan firm Aristotle International, which specializes in providing campaigns with individual records on voters. By the primary election day on August 14, Bryce in contrast had spent over $100,000 on polls of the district and thousands a month on continued access to the VAN individual data platform.

Although the contest was closer on the Democratic side, Bryce and Steil would win their primary elections with comfortable margins at over 50 percent of the vote. Into the general election, the two candidates' data spending patterns would continue to differ. Figure 1.1 reveals what the candidates spent on learning about the electorate over the course of their primary and general contests. Bryce not only spent more overall, but he also spent proportionally more on individual-level data. In contrast, even though Steil purchased less electoral information, he spent most of his data-related budget on surveys of the electorate and next to nothing on individual-level records on voters.

[9] https://www.cnbc.com/2018/06/18/paul-ryan-endorses-bryan-steil-for-his-wisconsin-house-seat.html

[10] https://www.gazettextra.com/news/politics/bryan-steil-running-for-congress/article_90fba97a-c881-55d1-934f-be382f49d423.html

**Figure 1.1 Candidate Data Spending**

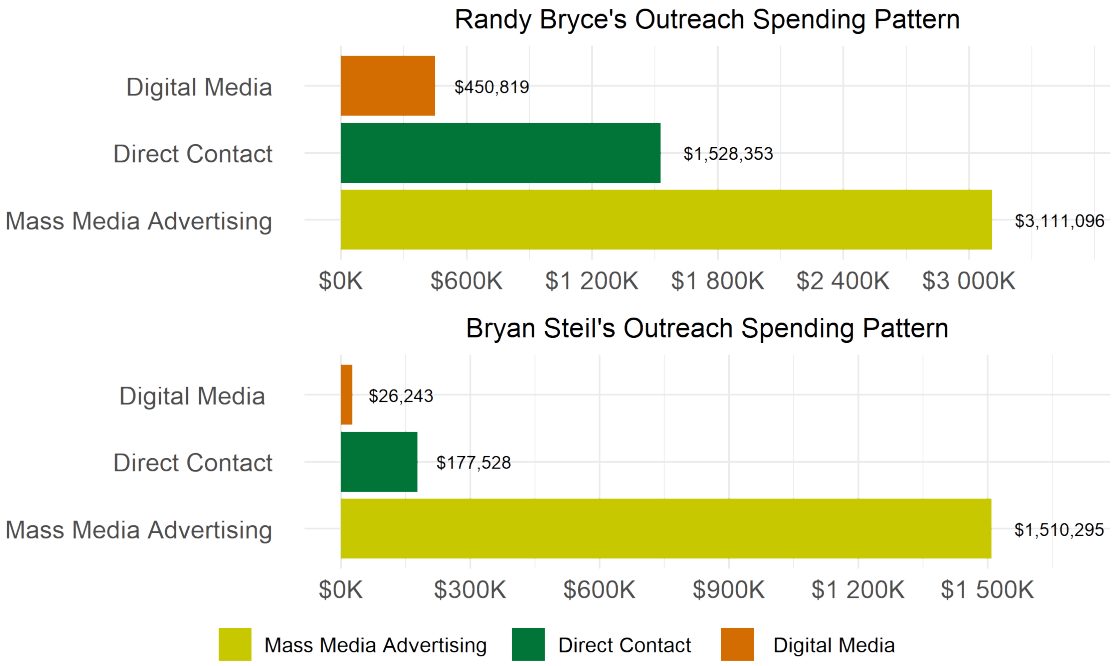



The candidates' differences in data spending patterns also extended to their outreach strategies to communicate with voters. Steil would almost entirely communicate with voters through television advertising, spending approximately $1.5 million as seen in Figure 1.2. His mass media strategy would be supplemented with an additional $2.6 million television ad buy made by the Congressional Leadership Fund (CLF), a Ryan-backed Super PAC. The outside spending against Bryce would paint the ironworker as a criminal, having himself been arrested nine times for, among other things, drunken driving, driving with a suspended license, and failure to make child support payments. His police officer brother was even featured in one of the CLF's ads calling him a "deadbeat" and endorsing his Republican opponent.[11] While Bryce spent a significant $3 million on mass media advertising buys, his outreach strategy was more concentrated on talking to voters directly. Figure 1.2 shows that Bryce spent significant amounts of his overall outreach budget on digital media advertising and direct contact, such as canvassing, door-knocking, and direct mail.

[11] https://www.jsonline.com/story/news/politics/elections/2018/11/06/wisconsin-1st-congressional-district-election-results-bryce-vs-steil/1857418002/

Bryce's outreach strategy appeared in line with his stated commitment to win "by knocking on every door and talking to every voter."[12]

**Figure 1.2 Candidate Outreach Spending**




Despite his commitment to persuasion, Bryce would lose to Steil by over 10 percentage points. All the money the outsider Democrat would spend on learning about and communicating with the district would not be enough to overcome the fundamental Republican advantage in a district previously held by the Speaker of the House.

## 1.1 Overview of Project

The amount of voter data available today is greater than at any point in history. While election campaigns have always used information about the electorate to develop and implement strategy,

[12]https://www.facebook.com/436981596658189/photos/a.443796612643354/636793553343658/?type=3&paipv=0&eav=AfanQcLV4gCNWwvP9aaavxRyXfXz4kpXJBNVCn-wZeXBThxa2rxNxZztKyO_WfmlrYs&_rdr

the amount and diversity of data have continued to increase exponentially. Today's campaigns confront an array of data options and must make decisions over what they use to craft a winning strategy. These data on the electorate are a combination of both old and new sources. Campaigns need to blend the signals they get from past election results, Census data, and polls with increasingly available individual-level information on voters gathered from state voter files and consumer records. All these data must then be transformed into politically relevant information and folded into a campaign strategy.

How campaigns deal with all these potential sources subsequentially determines how they perceive and engage with the electorate. They must make decisions about what information on voters is relevant and what is not. These understandings always fall short of the complexities of reality, but they are critical to the decisions that campaigns make when crafting an electoral strategy. Who are our supporters? How do we mobilize them? Who are the voters in the middle? How do we persuade them? Who will not support us? Do we ignore them? The current voter information environment can provide answers. Yet those answers are fundamentally shaped by the data available to campaigns, which data they choose to prioritize, and how they interpret that data.

The example from Wisconsin's first congressional district highlights how campaigns must consider many different strategic factors when determining what information to purchase and how to communicate with voters. Did Randy Bryce purchase relatively more individual-level data to support his commitment to door-to-door, ground-game efforts and spend thousands on polls simply because he had the money to spare? Why did his opponent Bryan Steil choose to spend just a few thousand on access to voter-level data yet still purchase several surveys for tens of thousands of dollars? To be sure, the candidates faced different strategic decisions. Even a Democrat with millions of dollars would face a tough, uphill battle regardless of who the Republican opponent

was. Yet both candidates were motivated to gather information to help them plan and implement their electoral outreach strategies, albeit in different amounts and proportions.

This dissertation expands our understanding of the voter information environment by providing the first systematic account of the data that campaigns further down the ballot have come to rely upon. Scholarship on this topic devotes most of its attention to the newest, cutting-edge data and technology available to well-resourced presidential campaigns. Although important to understand, presidential contests are highly unrepresentative of the thousands of lower-level campaigns waged more frequently across the country. I expand the scope of study by investigating down-ballot variation in data-driven campaigning for the U.S. House of Representatives. My project documents the extent to which House campaigns prioritize individual-level voter data primarily based on state voter rolls versus information provided by representative surveys of their electorate. A wider breadth of inquiry provides a more comprehensive picture of the current voter information environment and the extent to which presidential practices permeate down the ballot and across different strategic contexts.

Explaining the acquisition and impact of electoral information sources is crucial to understanding contemporary campaigning. In his treatise on voter databases, Eitan Hersh (2015) argues that "To understand how campaigns endeavor to perceive their supporters, one must look at the data they use to inform their perceptions" (30). I take a similar view. Campaigns can *only* understand voters by the types and kinds of data they can collect on them. Sources of electoral information, how they are interpreted, and how they are used are the antecedents to many outcomes of central concern in American politics. Going back to Richard Fenno's (1978) seminal work observing House members in their district, he motivates his research by asking, "*What does an elected representative see when he or she sees a constituency?* And, as a natural follow-up, "*What*

*consequences do these perceptions have for his or her behavior?*" (xiii). My broader theoretical motivation comes just before his: *What influences a campaign's view of the electorate?* The downstream consequences of electoral information impact everything from whom in the electorate candidates communicate with to whose views are represented.

Empirically, my dissertation addresses one central question: *How do campaigns use different sources of electoral information to inform voter outreach strategy?* To answer it, I combine interviews with analysis of millions of expenditure records. My inquiry begins with qualitative interviews to unpack how campaign professionals perceive and interact with different data sources to craft and implement their voter outreach strategy. In chapter two, I document how large individual-level voter databases and statistical models that reduce voters to a few predictive scores and demographic characteristics have reinforced longstanding perceptions that the electorate is comprised of different strategic segments. Additionally, my conversations reveal that, while Republicans and Democrats agree on how to segment the electorate into strategic components, they take different approaches to combining disparate data sources in practice. Building on these initial exploratory interviews, I provide the first comprehensive examination of voter data and outreach spending patterns by leveraging millions of verified Federal Elections Commission (FEC) expenditure records reported by congressional campaigns between 2006 and 2018 to answer three interconnected questions:

1. *How have congressional campaigns' data spending patterns evolved?* Leveraging approximately 3.5 million spending records from 2,500 major party candidates between 2006 and 2018, chapter three details my refinement of these records to generate novel insights into the data preferences of congressional candidates. I examine the marketplaces for electoral information, the cost of different sources, and campaign-level data spending

patterns. My descriptive findings detail how congressional campaigns operate in disjoint partisan marketplaces and how their data preferences have diverged over time. Democratic candidates have come to prioritize large individual-level voter databases, while Republicans maintain a consistent preference for using polls to craft their electoral strategies.

2. *What explains congressional campaigns' data spending patterns?* Building on my descriptive findings, chapter four explains these observed data spending patterns by accounting for various strategic factors that campaigns confront when crafting their communication strategies. I reveal how a candidate's party affiliation remains a consistent differentiator of campaign-level data preferences. Democratic congressional candidates are more likely to spend their limited budgets on acquiring individual-level data whereas Republicans instead prioritize polling. Yet campaign data spending patterns are also responsive to other traditional considerations like financial resources, competitiveness, and incumbency advantage.
3. *What impact do congressional campaigns' data spending patterns have on their voter outreach strategy?* In chapter five, I find that campaigns' mix of outreach spending toward mass media, digital advertising, and direct contact is responsive to their data spending patterns, suggesting that different data preferences have campaign-level consequences. As campaigns prioritize large voter databases, they are more likely to communicate with voters directly and in a more personalized manner than is possible with traditional mass appeals. Yet longstanding strategic considerations persist and even mitigate the effect of data preferences.

Weaving together my mixed methods, I argue in my final chapter that developments down the

ballot are not simply the result of new innovations in campaign technology. Rather, the diffusion and impact of data technologies are gradually realized and the product of strategic campaign-level considerations and party-level differences in data priorities. Lacking the resources at the top of the ticket, House campaigns waited for the arrival of third-party data vendors to process and package voter-level records before making them a core part of their electoral strategies. Despite the availability and affordability of voter-level information, the two major parties appear to be embedded in different voter information environments. While Republican congressional candidates have made greater strides toward incorporating individual-level voter databases, they have not made them central to implementing their outreach strategy. By contrast, Democrats throughout the country are united not only by a shared voter database but also by an approach to data-driven politics reinforced by a party culture in part inspired by the success of the two Obama presidential campaigns. Notwithstanding the impact of voter data, classic campaign considerations such as financial resources, incumbency advantage, and electoral threat still dominate campaign communication decision-making. As a result, I contend that new data sources have not fundamentally altered strategic electoral considerations that campaigns have long faced. Rather, the influx of large voter databases has primarily increased the efficiency of the direct contact activities designed to maximize the return on campaign-level communications and reinforced perceptions of voters as mere numbers necessary to win an election.

The remainder of this chapter is divided into two sections. The first section reviews the historical and scholarly records of developments in electoral information technology and data-driven campaigning up to and throughout my period of investigation starting in 2006 and ending in 2018. I then turn to a discussion of my methodological approach.

## 1.2 From Best Guess to Big Data

For most of American history, campaigns had a difficult time gathering information about the electorate. Politicians kept track of public opinion and their electoral chances through informal means, relying on attendance at political rallies, media reports, conversations with constitutes, opinions of other political elites, and claims of interest groups (see Brown and Halaby 1987; Bryce 1921; Eisinger 2000, 2003; Fenno 1978; Geer and Goorha 2003; Ginsberg 1986; Herbst 1993; Kernell 2000). Dating back to the 1800s, politicians and journalists also conducted straw polls – informal surveys taken orally or on paper – to gauge the popularity of a politician or policy (Herbst 1993; Smith 1990). Unfortunately for the candidates at the time, these early forms of information were frequently inaccurate because they reflected the views and priorities of only the politically engaged instead of a diverse range of voters (Geer 1996). Given the lack of alternatives, however, campaigns had little choice but to weave together these skewed sources to gain insight on voters.

The arrival of scientific polling represented a significant improvement in the accuracy of campaign perceptions about voters. With roots in the 1930s, survey and sampling methods would improve drastically over the following decades, especially with the addition of the random probability sample, and provide increasingly representative estimates of public opinion along with more accurate election forecasts.[13] Presidents were among the earliest adopters of robust private polling operations. The first was Franklin Roosevelt who used surveys to inform his policy decisions while in office. John Kennedy and subsequent presidents made survey research central to both campaigning and governing. In the 1970s, for instance, Richard Nixon spent millions of

[13] One of the most important developments was the implementation of the probability sample or random sample where each individual in the population of interest has an equal (or at least known) chance of being interviewed. See Converse (1987) and Rossi and colleagues (2013) for historical developments in survey and sampling techniques.

dollars on hundreds of polls to keep tabs on the policy preferences of the electorate and to help shape his strategy to communicate with voters, a trend that continued in subsequent administrations and among presidential candidates (Eisinger 2003; Jacobs and Shapiro 1995).

Although the earliest account of a House candidate commissioning a poll was in 1946 (Roll 1982; Sabato 1981), many members of Congress remained skeptical of polls and the sampling techniques used to find respondents for many decades. Most representatives instead preferred to rely on metrics such as letters from constituents for decades after polling had become routine for presidential candidates (Herbst 1993). When Kennedy was in office in the early 1960s, only around 10% of congressional candidates hired a pollster (Harris 1963). Medvic (2001) estimates that as late as 1984, only 15% of House candidates contracted a polling consultant. But by 1990, that number had risen to around 26% and reached 46% in 1992. Toward the end of the century and into the early 2000s, other estimates indicate that the number of House candidates who hired a pollster fluctuated between 50% and 60% in each election cycle (Herrnson 2004; Monson 2004).[14] In part driving the uptick was the decreased cost of polls resulting from increased sampling efficiency and computing power. Moving from mainframe to personal computing in the 1990s ballooned the number of firms offering political polls, increasing competition and further driving down their costs (Selnow 1993). The number of polling firms would continue to grow. While the exact number is difficult to estimate, Goidel (2011) reports that between 1997 and 2006 the number of marketing research and public opinion firms according to the Census Bureau economic data increased from approximately 4,000 to 5,500.

Political consultants were integral to polling's proliferation down the ballot. Sabato (1981)

[14] The number of publicly released trial heat polls conducted would follow a similar trend with a large increase in the late 1980s. Between 1980 and 2000, the number of public polls would increase 900% from fewer than 30 to nearly 250 (Traugott 2005).

provides one of the earliest accounts of their rise. He describes how political professionals introduced and, quite literally, sold many candidates on the benefits of survey research. For their fee, partisan pollsters provided campaigns with more than mere survey crosstabs. As Sabato writes, "For the modern national pollster is far more than an objective data collector or a mere engineer or statistician. He is an analytical interpreter, a grand strategist, and to some, a Delphic oracle" (73). The campaign consulting industry arising in the 1970s pioneered what Agranoff (1977) described as a new style of politics in which political professionals run campaigns with the assistance of polls and computers. Subsequent scholarship examining the professionalization of politics documented the continued growth of consultant-sold polling technologies and their diffusion into campaigns at all levels (see Burton, Miller, and Shea 2015; Burton and Shea 2010; Johnson 2007, 2010, 2016; Shea 2001).

The political science literature identifies four varieties of polls defined by when campaigns purchase them and their length: exploratory, benchmark, brushfire, and tracking. Candidates or party committees often commission exploratory polls before candidates assemble their campaigns to help a candidate decide whether to enter a race. Their primary purpose is to establish a sense of the overall race context and, if there is an incumbent, their vulnerabilities. After testing the waters, campaigns, with the help of consultants and pollsters, will field a lengthy benchmark survey with a sample size of around a thousand voters. They provide a starting snapshot of the electorate and political environment and serve to guide overall campaign plans. These baseline surveys ask an array of questions about the candidate and their opponent, including name recognition, support, and perceived candidate qualities. They also ask voters extensive questions about voters' issue positions and test the appeal of different campaign messages. Pollsters then analyze the data to create a summary of candidate, issue, and message support across different income and

demographic groups. Following the benchmark, campaigns sometimes field short follow-up surveys called tracking polls that collect only essential information, such as support and demographics, to measure if campaign activities had an impact. Brushfire polls by contrast are a bit longer than tracking polls but shorter the benchmarks and help campaigns test new messaging or issue strategies before implementing them, especially for mass media advertising (Burton, Miller, and Shea 2015; Monson 2004; Stonecash 2008).

Polls also help candidates segment the electorate into demographic and geographic groups. A practice that emerged in the 1980s, political pollsters paired the results of their surveys with Census data to target geographic clusters (Weiss 1988). The practice of geo-demographic targeting saw more use in the 1990s when the Census Bureau began to publish statistics at the block level. Many political consultants of the era (and afterwards) defined voters in largely sociodemographic terms (e.g., age, income, education, race, and ethnicity) based on the results of surveys. Political marketers used their survey results to create different demographic voter profiles and then layered those profiles onto neighborhoods with similar characteristics. The geographic clusters often received names such as "Pools & Patios," "God's Country," or "Downton Dixie-Style" (Weiss 1988, 229). Campaigns used these geo-demographic archetypes as an early form of microtargeting. Consultants would classify neighborhoods as potential targets for their outreach activities and create customized GOTV or persuasion messaging based on its characteristics (Baldwin-Philippi 2018; Issenberg 2013; Kreiss 2012).

The spread of polls represented a fundamental shift in the levels of uncertainty that politicians faced. Reflecting on the changes to campaigning brought about by polls, Roll (1982) writes: "Old-style political advice – subjective, arguable, derived from intuition, gut feeling, and past experience – has given way to the objectivity and finality of cold, hard figures, vividly presented and perceived

as realities in their own right" (73). In a similar vein, Geer and Goorha (2003) argue that all of American political history can be divided into two information eras: poorly informed before polls and well informed after polls (see also Geer 1996; Geer 1991). As the authors argue, however, this is not to say that polls provide perfect information, only that they represent a significant improvement compared to the often-biased indicators of the past. For instance, instead of straw polls and rally attendance, candidates can conduct representative surveys of their constituents to gain an up-to-date snapshot of their levels of support and electoral chances. They also permit candidates to identify systematically which issues voters care about and which types of voters care about those issues rather than having to rely on informal conversations or media reports (Stonecash 2008).

Polls past and present have potential biases that campaigns must pay attention to. Pollsters and survey methodologists divide threats to accurate representations of public opinion into two categories: sampling and non-sampling. Sampling error is the mathematically calculated deviation of the sample statistic from the true population and is often included in the reporting of results as the margin of error. Non-sampling errors are all other threats, including measurement, coverage, and nonresponse errors. Measurement errors reflect biases caused by the questions themselves, such as in question wording, question order, or response bias when the respondent consciously or unconsciously provides inaccurate responses. Both coverage and non-response errors are the results of the sample not reflecting the population of interest. Coverage errors occur especially when a sample misses important groups or subpopulations. Similarly, non-response errors are when particular groups are less likely to be represented in the sample because they refuse to take the survey or cannot be reached (see Groves 1989; Groves et al. 2009; Weisberg 2005 for discussion of survey methodology and the total survey error approach).

Out of all potential errors, contemporary public opinion scholars and professional pollsters have been most concerned with coverage and non-response. The decline of landline phone usage in the United States along with the decreased likelihood of individuals answering their phones raised concerns in the early 2000s about the representativeness of samples (Goidel 2011).[15] A series of studies from Pew found that, even with efforts to include cell phones, response rates declined from approximately 30% in the year 2000 to 6% in 2018 (Kennedy and Hartig 2019). *The New York Times*, in a polling partnership with Sienna College, reports an even worse overall completed survey rate in 2018 with only 46,000 completed surveys out of 2.8 million calls, approximately a 1.5% completion rate.[16] Such difficult-to-reach respondents and response rates approaching zero are in stark contrast to decades prior when landline usage in the United States topped 90% and response rates were upwards of 70% in the early 1980s. Both coverage and non-response errors are not only a threat to the accuracy and validity of the estimates but also contribute to their price, as a poll's cost increases as respondents become more difficult to find and interview. Polls can get quite expensive. Burton and colleagues (2015), for instance, quote the cost of a full-length benchmark poll conducted at the beginning of a campaign to measure baseline support, issue preferences, and voter attitudes at between $25,000 and $30,0000.

To address the difficulty of finding respondents, some polling firms began to offer Internet surveys with either recruited or opt-in samples in the 2010s. In the academic and market research survey space, the highest quality Internet polling firms, such as Gfk and YouGov, draw samples probabilistically and provide respondents with financial incentives to participate. Internet surveys

[15] Concerns over coverage and non-response biases are a perennial issue among academics and practitioners. For instance, Steeh (1981) documents increases in survey refusal rates increasing from near zero in the 1950s to upwards of approximately 25% in 1979. Likewise, the polling industry had a similar debate during the transition from face-to-face and mail surveys to phone-based approaches (Ansolabehere and Schaffner 2018).

[16] https://www.nytimes.com/2018/11/03/insider/midterms-polling-upshot-voters-elections.html

are significantly cheaper than phone surveys with prices as low as $10 to $20 per person, although a geographically restricted sample, such as a congressional district, is significantly more expensive compared to a national sample. Nevertheless, Internet surveys are typically less than half the cost of phone polls (Ansolabehere and Schaffner 2018). The growing consensus is that Internet samples can be representative and accurate by approximating demographic balance with weighting or quotas (Baker et al. 2013; Prosser and Mellon 2018), although there remains skepticism about the accuracy of sub-national Internet surveys (Keeter 2018). Most political consultants are also hesitant and prefer either random digit dialing (RDD) of phone numbers in the district or rely on registration-based sampling (RDS) techniques using the voter file as the population to sample (Burton, Miller, and Shea 2015).

Concerns over the accuracy of pre-election "trial heat" polls have more broadly permeated into the public consciousness even if evidence for their overall decline is limited. Many scholars argue that modern polling is no less accurate than in the past (see Prosser and Mellon 2018). One robust cross-national study even finds that the accuracy of national polls compared to election outcomes has remained consistent between 1942 and 2017 (Jennings and Wlezien 2018). Despite the empirical evidence, perceptions of polling's accuracy have declined in recent years, especially following Donald Trump's unexpected victory in 2016. While polling postmortems were not unanimous, they generally agreed that 2016's miss was the result of state-level polling falling victim to a mix of unintentional coverage and non-response errors (see Kennedy et al. 2018). A 2017 survey found that only 37 percent of registered voters have a great deal or good amount of trust in public opinion polling.[17]

[17] https://www.huffpost.com/entry/most-americans-dont-trust-public-opinion-polls n_58de94ece4b0ba359594a708

Just as polls became commonplace down the ballot, a new voter registration regime would spark further innovation in political information technology. The 1990s marked the policy beginnings of large databases containing every registered voter in the United States with the passage of the National Voter Registration Act (NVRA). The 1993 "Motor Voter" law is best known for stipulating that states must allow citizens to register to vote when applying for or renewing a driver's license. Less known, the law also required states to "conduct a general program that makes a reasonable effort" to remove ineligible voters from their voter registration lists and suggested a framework for doing so (Eckman 2021). The NVRA brought attention to what political practitioners had known for a long time: many states' voter files were in poor condition. They also were difficult to obtain. Approximately half of the states at the time relied on decentralized paper-based systems kept by various local officials. Despite the NVRA inspiring some states to establish statewide voter files and improve list maintenance, many states failed to do so, resulting in the continued poor quality of many files, if they existed at all (Hersh 2015; Hillygus and Shields 2009).

Only later when computing technology had advanced and a national controversy over election administration sparked reform did the states clean up their voter registration rolls and make them more accessible. Enacted in response to the problems of the 2000 presidential election, the Help America Vote Act (HAVA) of 2002 made sweeping changes to voting systems and election administration. Among its numerous provisions, including providing federal funds to upgrade voting equipment and establishing the Election Assistance Commission, the law also sought to establish standards for voter registration and strengthen the list maintenance requirements proposed by the NVRA. HAVA required each state to create "a single, uniform, official, centralized, interactive computerized statewide voter registration list" by 2006. Although twelve states missed the deadline to create their databases, that midterm election year marked the

beginning of widespread access to voter file records across the country. By 2008, every state apart from North Dakota[18] was providing a government-subsidized and maintained list of registered voters (Hersh 2015; Hillygus and Shields 2009). Most states provide voter registration lists to campaigns and political consultants at no cost or a nominal fee. For instance, the Election Assistance Commission reports that 11 states provide access to voter files for free, while the cost for the voter file in the remainder of states ranges between $20 and $37,000 with an average cost of approximately $3,000.[19]

Decades before HAVA's mandated statewide voter files, campaigns, parties, and consultants had made attempts to collect large lists of voters. Campaigns have accessed lists of registered voters to help them with their outreach activities since their genesis more than a century ago (Gosnell 1926). As early as the 1960s, members of Congress were taking advantage of new technological advances such as magnetic tape and punch-card computing to store and analyze information on individual voters. State and national parties also played a central role in this data collection and processing. In 1966, for instance, the Republican National Committee (RNC) began to provide state parties with assistance in setting up early centralized computing efforts to store and process individual records. The Democratic National Committee (DNC) at the time meanwhile provided state and local party organizations as well as congress members with access to computerized mail addressing and labeling technology along with training sessions on how to use voter files (Chartrand 1977).[20] Such piecemeal and fragmented data initiatives continued over the next few decades. Democratic consultant Robert Blaemire (2012), for example, recounts how

[18] North Dakota was exempt from the requirement because the state does not require voter registration.

[19] https://www.eac.gov/sites/default/files/voters/Available_Voter_File_Information.pdf

[20] The historian Jill Lepore (2020a, 2020b) provides a detailed historical account of these early innovations in the 1960s that applied a social scientific approach to new data sources.

political data vendors who contracted with state parties began painstakingly collecting voter rolls from municipalities and cleaning them to create as complete and accurate statewide voter lists as possible. But it was not until the late 1980s that the DNC made an initial effort to encourage state parties to build their own databases, compared to Republicans who began decades earlier (Johnson 2007).

The 1990s saw the first attempts to create a nationwide database of registered voters. As detailed in historical accounts by Daniel Kreiss (2016) and Rasmus Nielsen (2012), the RNC established Voter Vault in 1995 largely to support the national party's already robust direct mail fundraising efforts. The party database was a collection of the voter files that most state parties had been collecting since the late 1970s and into the 1980s (Huckshorn 1976). By the 1990s, Selnow (1993) documents through interviews with political operatives that the Republican party was the first mover in creating a national voter database. Similarly, Goodhart (1999) surveys state party organizations and finds that Republicans devoted considerably more resources to maintaining and updating their voter files compared to Democrats. In exchange for sending their lists to the national committee, the Republican state parties would receive back a cleaned and enhanced version of the voter file with additional information added in such as verified phone numbers and other commercial data. These early data collection efforts, largely meant to support presidential campaigns, evolved into a fully interactive database platform with 175 million records in the early 2000s. Coming online in 2002, Voter Vault's interface was piloted during the midterm contests to be refined for George W. Bush's 2004 reelection bid. His campaign in coordination with the national party would make further investments and heavily rely on the individual-level data platform to engage in extensive voter identification and microtargeting. After its development, lower-level campaigns paid the RNC a fee to access and export the database, and the price would

vary depending on the specific data requested and the level of access granted.

Lagging Republicans, the Democratic Party did not attempt to coordinate the creation of a large national voter database until the 2000s. Although state parties maintained voter registration records to various extents (Goodhart 1999), the DNC's first attempt to compile their files was after Al Gore's Electoral College defeat in 2000. To rival Voter Vault, the DNC rolled out its first voter database, DataMart, and fundraising database, Demzilla, in 2004.[21] The Democrats' plan was like the RNC's: collect, clean, and augment voter files provided by the state parties. The plan failed. As Nielsen (2012) argues, the project was abandoned after 2004 because many state parties did not buy into the services, and the data in the system was often inaccurate and filled with errors. Motivated in part by John Kerry's defeat in 2004 and the collapse of his presidential campaign's data infrastructure, the DNC pivoted to a new database called VoteBuilder in 2006. Partnering with the partisan firm Voter Activation Network (VAN), the DNC created one standardized national data tool and interface (Kreiss 2012). Arriving after the passage of HAVA, "the VAN," as the interface is referred to, also benefited from increasingly accurate and easy-to-access registration lists. The VAN saw widespread adoption and sign-on from nearly every state party by 2008 (Hersh 2015).

What had been early Republican innovation in data and technology would become inertia. In two books, Daniel Kreiss (2012, 2016) chronicles the evolution of campaign technology in both the Republican and Democratic parties between 2006 and 2014 through interviews with presidential staffers. Tracing back to Howard Dean's failed bid to become the Democratic presidential nominee in 2004, Kreiss documents how the Democratic Party and its network of

[21] Demzilla was technically the name of donor database, and the voter file was called DataMart. Most operatives, however, referred to both systems as "Demzilla" (Nielsen 2012).

partisan actors, consultants, and companies began to invest much more into the party's shared data and technology infrastructure compared to their Republican counterparts. Benefiting from the VAN interface (the company that makes the interface was later renamed NGP VAN after a merger) and other data initiatives in the extended party network, the Obama campaign in 2008 would build atop these past investments and continue to improve them in his 2012 campaign. One such celebrated innovation coming out of Obama's reelection campaigning was Project Narwhal, which merged data from various databases and was updated in real time up until election day.[22] Veterans of the Obama campaigns would institutionalize many of the innovations by creating partisan firms to support lower-level campaigns. What they created was a shared voter data ecosystem with near universal sign-on from Democratic campaigns in addition to many progressive causes.

In stark contrast, the Republican party saw little improvement to its shared data infrastructure, which remained for years in technological stasis. Even by 2014, the Republican party lacked a system on par with the VoterBuilder database and its VAN platform that was continually updated with data from across the country. Instead, RNC's Voter Vault, (later renamed GOP Data Center) was a static database with lists that could only be manually exported and could not be updated within the system. This lack of feedback was a critical flaw. As early as 2008, lower-level Democratic campaigns were adding back into VoteBuilder response data such as updates on partisan leanings, candidate preferences, and likelihoods of turnout from their interactions with individual voters (Nielsen 2012). The lack of response data and integration is the primary reason that even well-resourced Republican campaigns came to rely on third-party data vendors rather than the national party beyond 2014 (Kreiss 2016). Hatch (2016) reiterates these findings in a 2011

[22] Project Narwhal is often contrasted with Obama's opponent Mitt Romney's ORCA. The ORCA system was meant to rival Narwhal and was designed to be a comprehensive voter outreach tool that integrated data from multiple sources. The system saw limited use and was plagued with glitches, completely failing on election day.

survey of state parties that reveals little investment in data technology among Republicans compared to robust data-related spending by Democratic state officials, leading most Republican candidates to rely on outside companies rather than the party for their data needs.[23] Thus, unlike their Democratic counterparts, Republicans had (and continue to have) a much more fragmented voter data ecosystem with limited integration between campaigns and even official party organizations.

Whether coming from party-contracted platforms or other data vendors, voter databases and their interfaces have several commonalities. As mentioned, the foundation of the modern voter-level database is the statewide voter registration rolls collected from state election agencies. Although HAVA required each state to create a centralized list, states vary in the types of information they require voters to provide when registering. Every state with a voter file at least provides a voter's name, official address, and election turnout history information on whether an individual turned out in past elections. Varying between states, some record demographic information on voters, such as age, gender, and race, while others disclose voters' partisan affiliation either through official party registration or through participation in party primaries (Spencer and Ross 2019). Because the files include a voter's address, data vendors can layer on geographic information such as precinct voting returns and data culled from the U.S. Census like neighborhood characteristics. Data vendors also append additional individual-level information to

[23] As of 2023, at the time of the writing of this dissertation, the current party-contracted platforms evolving from these earlier efforts to build a nationwide list of voters are NGP-VAN (formerly the Voter Activation Network) and GOP Data Center (formerly Voter Vault). Not every data vendor is contracted with a party, however. For instance, Republican and conservative groups can access data provided by the firms Data Trust, i360, and CMDI (Crimson), which all maintain separate databases. Meanwhile, Democrats and liberal groups almost entirely rely on the NGP VAN, which other party companies, such as TargetSmart, use to build models and assist with microtargeting. There are also firms that serve both sides, such as L2 (formerly Labels & Lists), Political Data Intelligence (formerly Political Data, Inc.), Polis, Aristotle International, and NationBuilder, among others.

these voter registration records, including verified mailing addresses and phone numbers, missing demographic data, and even consumer spending patterns. Data vendors clean and merge all these individual data sources to compile up to and beyond a thousand pieces of information on each voter (Issenberg 2012, 2013).

Political data vendors leverage the millions of rows and thousands of columns of voter data to generate predictive scores on voters with statistical models. Nickerson and Rogers (2014) divide these propensity scores into three types: behavior, support, and responsiveness. Scores are scaled commonly to range from 0 to 100. They most importantly predict the likelihood that an individual will vote, donate money, or volunteer to work on a campaign. Support scores predict the likelihood that an individual will support a particular candidate, party, or issue. Finally, responsiveness scores anticipate how an individual will respond to direct campaign contact, which can be informed from past campaign interactions or even randomized field experiments. Political data analysts can create voter scores with various data points. Yet the most valuable variables for predicting political tendencies tend to be those publicly available coming from the voter file, basic demographic and socioeconomic information, and response data rather than consumer data such as home ownership or current subscriptions (Hersh 2015; Nielsen 2012). Campaigns use a combination of these scores to help them determine whom to contact and whom to ignore and, along with other available data points, what message to send them.

Access to individual-level records and derived predictive scores have certainly made it easier to find voters, but campaigns still have a few potential pitfalls to attend to. In contrast to the geodemographic archetypes and neighborhood-level outreach strategies of the 1980s and 1990s, contemporary campaigns microtarget specific households and individuals, contacting some and ignoring others who reside on the same street. Ample individual-level data also help campaigns

avoid ecological fallacies – making inferences about individuals based on group characteristics – and instead, as Kreiss (2016) refers to it, perceive voters as “whole citizens” through various data sources.

Threats to gaining a complete picture of voters come primarily from missing or incomplete data. First and perhaps most apparent, voter databases are based on the state voter registration records, meaning that unregistered voters are in many cases excluded. A Pew study published in 2018 found that voters who self-reported being certain they were not registered to vote were able to be matched to five major voter databases at rates of between 4% and 50% (Igielnik et al. 2018). Second, as mentioned, states vary in what fields of information they require voters to provide when registering. Hersh (2015) documents how campaigns in states that provide party affiliation and racial identity information differ in their outreach strategy compared to campaigns in states without these two pieces of individual-level information. Third, the quality of data varies by individual. Historically marginalized groups, including racial minorities and individuals with lower socioeconomic status, are less likely to be listed in commercial voter files. And if they are listed, disadvantaged populations are more likely to have incorrect information, especially their physical address, not only because they are less likely to be registered voters but also because they generate less consumer data (Igielnik et al. 2018; Jackman and Spahn 2021).

At a more theoretical level, individual-level data are also biased in the sense that they can only provide a partial view of the individual voter. Hersh (2015), for instance, develops the "Perceived Voter Model” that posits campaigns understand voters through the data they acquire. Although modern voter files may have thousands of columns of individual-level information, the predictive models that generate voter propensity scores are largely a function of public records such as state voter rolls and census data rather than consumer records, which have at best a tenuous relationship

to most political outcomes. As a result, campaigns have largely a simplistic view of voters and lack precise attitudinal and issue-specific data for most voters. This means persuasion – both the classification of voters as persuadable and the act of convincing on-the-fence voters – remains elusive even for high-resourced campaigns.

Beyond access to large voter databases, modern campaigns have access to a variety of digital technologies to help keep track of voters and target their outreach communications. At the presidential level, digital strategy includes the creation of customized digital tools. For instance, many heralded the 2008 Obama campaign for creating the MyBO online organizing platform, and his 2012 reelection campaign was the first to develop a mobile phone app (Bimber 2014). Permeating both of Obama's campaigns was also a culture of testing, such as A/B email testing where the campaign randomly assigned two competing messages to determine which is more effective (Kreiss 2016; McKenna and Han 2014).

While there have been attempts to link these digital trace data together with voter file records, such as when the Obama campaigns would drive supporters from social media to their websites to record their details (Kreiss 2012, 2016), much of what even high-resourced presidential campaigns can collect from social media is often limited.[24] Take for example the infamous Cambridge Analytica scandal where the namesake data firm harvested tens of millions of Facebook profiles under false pretenses and shared their results with the Trump campaign. Cambridge Analytica used the data on approximately 87 million Facebook users to create "psychographic" profiles to microtarget and influence voters. Claims of the near-magical success of this microtargeting based on psychological characteristics abounded, but the academic inquiry into the capabilities of such

[24] Though not focusing on political campaigns, a doctoral thesis by Malik Momin (2018) demonstrates the biases in digital data and the difficultly in making generalizable claims based on digital records alone.

Internet records instead suggested limited integration and success. The fallout of the Cambridge Analytica scandal also resulted in changes to the data privacy policies of Facebook and other major platforms (Simon 2019).

Below the level of the presidency, the use of digital tools is much less sophisticated. Through interviews with congressional staffers between 2010 and 2014, Baldwin-Philippi (2016) details how the culture of analytics and testing that dominates top-ticket campaigning is less widespread further down the ballot. Aside from a few highly resourced senate campaigns, she reveals how lower-level campaigns interact with streams of digital data in rather informal ways. For instance, congressional campaign staffers do pay attention to Internet metrics such as social media engagement, web traffic, and email interactions. Yet outside of customer relationship management (CRM) systems – software that campaigns largely use to manage and target emails – digital data and insights are disjointed from other analytical efforts. More importantly, major platforms such as Facebook and Google do not allow advertisers (and by extension campaigns) to access their user data outside of their within-platform ad purchasing and targeting tools (Jungherr 2018). Even still, campaigns make extensive use of microtargeting menus to target their digital ads based on various characteristics, such as user demographics, locations, search histories, and their platform engagement within the websites (Kim et al. 2018).

Scholars have come up with various terms and definitions to describe the changes to politics brought about by big data in the past two decades. Nielsen (2012) details how campaigns, with the assistance of voter data platforms, pursue "personalized political communication," especially in their on-the-ground direct contact efforts. Kreiss describes the current era as "technologically intensive" in which "parties and campaigns have invested considerable resources in technology, digital media, data, and analytics to not only keep pace with these changes but also actively shape

technological contexts and define what twenty-first-century citizenship looks like" (3). Baldwin-Philippi (2018) explains "data-driven campaigning" as "using large data sets to either target messages to particular populations or test the efficacy of variations of messages and a variety of goals" (2). For my purposes, although I use these terms interchangeably, this project does not ascribe to a particular definition of what counts as "technologically intensive" or "data-driven campaigning" but rather documents the different sources of information that congressional campaigns rely on to craft and implement their voter outreach strategy.

The scholarly and historical records thus far have taken stock of new developments in the storage, processing, and implementation of campaign information technology at different points in time. To summarize, innovations in campaign information technology can be characterized by four observations:

1. *Campaigns have access to more data on voters than ever.* Compared to the pre-poll and pre-database eras, campaigns in the new millennium have better perceptions of their potential voters both in the aggregate and at an individual level. Although concerns over the accuracy and the cost of polls have risen since the early 2000s, campaigns continue to rely on representative surveys of the electorate to help them learn timely insights into where their candidate stands with voters, what issues voters care about, and which voters care about those issues. Modern campaigns have access to large voter databases compiled and maintained by political data vendors that collect and make sense of various individual-level data sources, including state voter files, census data, consumer records, and past voter interactions. The development of comprehensive national voter databases was not only the result of increases in computing and storage technologies but also regulatory changes, especially since HAVA passed in 2002 and required states to create a single centralized

voter registration file. Lastly, while campaigns have access to a variety of digital tools, such as CRMs and ad targeting on social media platforms, they are often unable to integrate the digital data from voters' online activities into individual-level data systems or campaign-level decision-making.

2. *Campaigns relied on a network of partisan political actors to diffuse innovations.* Whether official party organs or partisan consultants, party-linked actors contributed to the current electoral information environment. Candidates were skeptical of polls until partisan consultants sold them on the value. Republicans had an advantage in the size and scope of their shared national voter registration database as late as 2004 because of early investment by the RNC in collaboration with state parties. Yet Republican sophistication stagnated. Beginning in the 2006 and 2008 election cycles, Democratic campaigns began to adopt the VoteBuilder database and the VAN interface, which was built and maintained by an officially contracted partisan firm. This shared voter database and software allowed Democratic campaigns and progressive causes to not only come up with lists of voters but also manage other aspects of their campaigns and input back into the system response data from their interactions with voters. In contrast, the Republican rival database, Voter Vault and the later renamed GOP Data Center, lacked such capabilities, leading to a fragmented Republican userbase and the rise of rival Republican and nonpartisan voter data platforms.
3. *Campaigns down the ballot looked to presidential campaigns as sources of innovations.* While not always first-movers, well-resourced presidential campaigns spurred developments in electoral information technologies by having the resources to invest and innovate. They were the first to adopt routine polling operations for planning their campaign outreach strategies. Early efforts to develop national voter databases were largely

in support of presidential campaigns. Most importantly, however, races atop the ticket served as the epitome of data-driven politics. They were training grounds for partisan consultants and staffers who went on to institutionalize innovations in networks of partisan vendors and data ecosystems.

4. *Campaigns must still pay attention to the limitations of data sources.* The results of a poll or a voter's modeled propensity score are simplifications of reality. Every source of information has potential sources of objective bias. Polls can falter because of sampling, coverage, non-response, or other errors. Contemporary voter databases, as large as they are, can systematically misrepresent the electorate because of unlisted and mislisted voters. Crucially, every source of electoral information only provides a partial picture of voters and strategic campaign contexts. The political information age is not one of perfect information but characterized by different electoral information sources with their own biases and limitations.

What existing scholarship misses is a comprehensive examination of how and why campaigns further down the ballot incorporate various forms of data into their decision-making. This is because much of the existing research has concentrated on what are truly atypical campaigns in American politics – those for president of the United States occurring every four years and costing billions of dollars. Surprisingly few studies have investigated the incorporation of electoral information technology in lower-level campaigns. Existing research includes, for instance, surveys of state party information technology (Goodhart 1999; Hatch 2016) and qualitative fieldwork and interviews with congressional campaigns and staffers (Baldwin-Philippi 2018; Nielsen 2012). As they reveal, the incorporation of new technologies into lower-level congressional campaigns is often haphazard and uneven (Nielsen 2012). Even well-financed campaigns at the congressional

level often lack the resources necessary to match sophisticated top-of-the-ticket practices and may even forgo data altogether (Baldwin-Philippi 2016, 2018, 2019).[25] This limited scholarship suggests that the uptake and diffusion of data-driven campaigning is not an inevitable conclusion but contextual on a host of other factors.

This dissertation seeks to build on and expand existing understanding with its focus on congressional campaigns. Several unanswered questions remain: Have congressional campaigns systematically shifted toward incorporating more individual-level data into their strategic decision-making and moved away from using polling in recent years? Have they done so evenly? In other words, do congressional campaigns vary in the sources of electoral information they rely on? If so, what explains this variation? And as a result, do different sources of electoral information have an impact on congressional campaigns' communication strategies with voters? Or are they unresponsive to shifts in the voter information environment? To answer these questions, I examine a period of marked technological change in campaign information technology starting in 2006 with HAVA's initial deadline for states to create a statewide voter registration list up until 2018 – years following the innovations of the two Obama campaigns and party-led efforts to create shared voter data ecosystems.

## 1.3 Methodological Approaches and Analyses

This dissertation tackles these remaining questions related to congressional campaigns' use of electoral information by combining interviews with observational analyses of millions of campaign expenditure records. I believe these outstanding questions and the lack of scholarly

[25] Investigations outside of the United States have found similar variations in uptake and sophistication of data-driven campaign practices, especially among lower-level campaigns (Anstead 2017; Belfry Munroe and Munroe 2018; Kefford et al. 2022).

knowledge on lower-level campaigns require both a systematic examination of observable patterns in data-informed campaigning and contextualization of these analyses through conversations with political practitioners involved in data-driven decision-making.

The starting point for my research is qualitative interviews. I conducted 15 semi-structured interviews with political practitioners between March 2020 and March 2021 during the COVID-19 global pandemic. The unprecedented public health measures in response to the outbreak of an unknown virus resulted in interrupted work, shuttered public spaces, a rattled economy, and a "new normal" connected online but isolated indoors. All these challenges occurred simultaneously with a presidential reelection year in which traditional campaign staples from rallies to door-to-door canvassing were completely upended. Rather than focusing on the novelties and changes caused by the virus, I consciously chose to structure my interviews (and this dissertation) around pre-pandemic politicking. While certainly on the mind of respondents, they were encouraged to be explicit about what differentiated the "before times" from the current situation when discussing how they used different sources of data to inform their voter outreach strategies.

I recruited my sample of campaign professionals through professional networks using a snowball sample. A snowball sampling method is a respondent-driven approach that begins with a convenience sample of easy-to-access contacts and then expands as respondents suggest other participants. Because professional political networks are organized by party, I chose the initial "seeds" of the snowball sample to be one Republican and one Democratic campaign professional. I was introduced to them through prior relationships and previous research on projects unrelated to this dissertation. This respondent-driven approach is a practical sampling method that attempts to maximize variation in relevant respondent attributes (Handcock and Gile 2011; Heckathorn and Cameron 2017). Respondents were encouraged to recruit others different from themselves, and I

sought out new seeds to increase variation based on partisanship, experience, and demographic characteristics, detailed in chapter two. The only other requirements to be included in the sample were at least five years of political experience and having worked with at least one statewide or congressional campaign, either as a consultant or staffer.[26]

I conducted the interviews via telephone and online video conferencing platforms. These semi-structured interviews involved asking respondents a predetermined set of open-ended questions to provide them the latitude to articulate their responses (Harvey 2011). [27] Questions fell into three categories: background, voter data, and campaign strategy. Background questions were brief. Most of their experience and demographic information was collected from online sources. Campaign strategy questions were about voter segmentation and modes of contact. Voter data questions prompted respondents to discuss how they conceptualize voter data and their views on the relative value of different sources to inform their outreach strategy. Appendix A contains the questionnaire. Interviews lasted approximately 30 minutes on average, and respondents were asked as many of the same open-ended questions as possible. I prompted respondents with follow-up questions when necessary to increase the comparability of responses. Respondents consented to participate orally and were made aware that their participation was voluntary and that their responses would remain confidential. Respondents were also told that the audio from the interviews would be recorded and transcribed.

To analyze the transcripts, I adopted Luker's (2010) inductive approach to qualitative analysis

[26] I chose to interview consultants in addition to former staffers in recognition of the fact that political consultants play an integral role in shaping campaign strategy and to increase variation in my sample.

[27] My reasoning for choosing open-ended questions is two-fold. First, the research question concerns both the perceptions and experiences that elite political actors have when dealing with voter data and coming up with a voter outreach strategy. Second, campaign professional are generally less receptive to close-ended questions and prefer to explain their reasoning in detail (Aberbach and Rockman 2002)

and based my analysis on its application by other scholars (e.g., Kreiss, Lawrence, and McGregor 2018; McGregor 2020; Van Duyn 2018). This methodological approach seeks to expand under-studied phenomena by examining evidence in the context of existing literature while also exploring extensions to existing scholarship. Specifically, I read the transcribed interviews to identify similarities and differences in responses to my interview questions. Based on my close reading and categorization of responses, I then sought out relevant literature and teased out new concepts to make sense of how practitioners think about sources of voter data in the context of planning and implementing voter outreach strategies. In a sense, my application of this approach can be thought of as an extended literature review guided by the real-world perceptions and experiences of individuals who engage with sources of electoral information and help to make decisions about how to reach and communicate with voters.

Building on these interviews, the remainder of this dissertation's empirical evidence comes primarily from cleaned and verified records of spending by congressional campaigns reported to the Federal Election Commission (FEC). The FEC has required federal campaigns to report itemized expenditures electronically since 2004. The independent regulatory agency defines an expenditure as "A purchase, payment, distribution, loan, advance, deposit or gift of money or anything of value made for the purpose of influencing a federal election."[28] Most itemized expenditures (also known as itemized disbursements) disclose the purpose, date, and dollar amount of the disbursement as well as the name and address of the payee. For example, records indicate that on August 31, 2018, Republican candidate Bryan Steil running against Democratic candidate

[28] House campaigns are required to report their operating expenditures when it is greater than $200 or when payments made to the same payee during the two-year election cycle add up to $200. Additionally, contributions of goods and services for free or less than market value, known as in-kind contributions, must be reported as an operating expenditure even though the campaign did not expend the money to receive them. See https://www.fec.gov/help-candidates-and-committees/making-disbursements/operating-expenditures-candidate/

Randy Bryce in Wisconsin's first congressional district spent $1,800 on a "database" from Aristotle International, a data firm based out of Washington, D. C. that offers campaigns access to a national voter file. Because most records are not classified by the FEC, the nonprofit OpenSecrets supplements FEC records with extensive coding of the records. Their coding, however, is insufficient for my tracking of data-source spending over time. Detailed in chapter three, I undertake a laborious but necessary refinement of these records to document the sources of electoral information campaigns purchased over a seven-cycle period between 2006 and 2018.

My investigation focuses on general election campaigns for the U.S. House of Representatives. These are the most common kinds of federal campaigns. The 435 contests represent a wide variety of campaigns in a variety of contexts. District competitiveness, for instance, can range from safe districts represented by a long-serving incumbent being challenged by a sacrificial lamb of an opponent to open-seat contests in a district that has flipped every cycle since the last time redistricting occurred. Districts also come in many shapes, sizes, and compositions. Contrast the ten-square-mile congressional district in New York City covering densely populated and majority black and brown neighborhoods in Upper Manhattan and the Bronx to Alaska's at-large district covering a diverse constituency spread out over 650,000 square miles. Existing scholarly work dominated by attention to presidential contests has left such potential variation understudied. Most importantly, compared to Oval Office contests, congressional campaigns face greater resource constraints that force them to make difficult choices about which sources of electoral information they acquire and which voter outreach strategies they choose to pursue. Across chapters three, four, and five, I explore these and other potential sources of variation at the candidate, campaign, and district levels.

I restrict the expenditure file supplemented by OpenSecrets to only House campaigns between

2006 and 2018.[29] As mentioned, this period covers the rollout of HAVA's statewide voter file mandate and the cycles following the expansion and uptake of large voter databases. Limiting the records to unique House campaigns in my seven-cycle period resulted in 4,176,486 itemized expenditures across 6,347 campaigns. Because party-contracted databases are an important aspect of the evolution of individual-level voter databases, I also confine the sample to Democratic and Republican candidates to make direct comparisons of differences between candidates of the two major parties easier. Similarly, I only examine major-party candidates who won their primary and made it to the general election to make candidates more comparable within each cycle.[30] With these exclusions, the total number of expenditures reduces to 3,429,292 reported by 2,515 unique campaigns (4,910 unique candidate-cycle campaigns) across all seven cycles and amounting to approximately $6 billion.

To analyze these records, I rely on a revealed preference approach to understand the ways in which campaigns prioritize different sources of voter data and plan their voter outreach strategies. Dating back to economist Paul Anthony Samuelson (1938), this approach to understanding human behavior is based on the idea that preferences can be "revealed" through purchasing patterns after accounting for an individual's income and the cost of a product. It assumes that consumers first have preferences for different products and second attempt to maximize their utility by choosing the best combination of goods and services. Among other assumptions, the revealed preference approach also assumes that consumers have complete information, thoroughly consider all their

[29] The original OpenSecrets file contained spending records on every federal campaign, including those for president and senate, and independent expenditure made by outside groups like political action committees post-2012. I chose House campaigns because they occur every two years, have the largest sample, and most variation in contrast to Senate campaigns where only approximately 33 occur in each cycle. I leave the analysis of other federal campaigns for future research.

[30] Campaigns for non-voting representatives of U.S. territories are also removed. In states that have runoff elections, such as Louisiana, I take the second-round vote totals when available.

alternatives, and consider decisions in similar ways. For my purposes, I do not consider all the formal economic assumptions in my analyses but do take the perspective that campaign purchasing patterns – in this case the relative amount of money spent on different data sources and types of outreach communications – are indicative of campaign-level preferences. For example, returning to Wisconsin's first congressional district, the Democrat Bryce spent around 25% of his entire data budget on individual-level data compared to 75% on polling. By comparison, his Republican opponent Steil spent only 6.5% of data-related spending on access to voter databases but 93.5% on surveys of the electorate. In this race, we would conclude that Bryce had a greater overall preference toward individual-level voter data because he spent *relatively* more compared to his opponent Steil.

By leveraging a mixed-methods approach to my analyses, this dissertation provides a robust overview of how campaigns engage with different sources of voter data for various aspects of planning and implementing their voter outreach strategies. The rest of the dissertation develops over five more chapters. While chapter two differs from the others in its qualitative methods and approach to inference, I draw on insights from my interviews throughout subsequent chapters. The concluding chapter seeks to marry these two approaches' understandings and provide the first comprehensive picture of how congressional have come to rely on data sources over time, why campaigns acquire sources of information, and what impact the sources have for campaign communication strategies with voters.

# 2 Putting Data in Perspective

Before collecting the spending records that serve as the primary source of analysis in my subsequent chapters, I conducted interviews with campaign professionals to help orient my understanding of how campaigns harness different sources of electoral information. These conversations help guide the rest of the dissertation. They reveal how practitioners take different approaches to interpreting and combining the results of polls and inferences coming from large voter databases. While practitioners from both parties have commonalities in their strategic segmentation of the electorate, their parties' collective orientations toward campaign technology and differences between the parties in their shared data infrastructure lead to divergent perceptions. Democrats are united not only by a single party-contracted voter database platform but also by a data culture that reinforces its hegemony across campaigns and party-linked organizations. Republicans meanwhile not only lack a unified voter database but have competing approaches to interpreting and making sense of different sources of information.

This analysis can be thought of as an extension of the first chapter's review of the historical record. By talking with campaign decision-makers, I account for the real-world constraints faced by lower-level campaigns throughout the country. Throughout this chapter, I compare their responses to existing literature to add nuance to our understanding of technology-oriented campaigning. Before turning to my comparisons, I first briefly document my interviewee's characteristics. The following section unpacks practitioners' perceptions of the electorate and the sources of information that shape those perspectives. Next, I detail the different approaches that practitioners rely on to integrate data sources into their decision-making.

## 2.1 Sample and Analysis

My interview sample comprises political consultants and former campaign staffers with at least five years of experience who worked on statewide or congressional campaigns. In total, I have transcripts from semi-structured interviews with 14 political professionals. I give each respondent a pseudonym to maintain their anonymity.[31] Their experience levels range from presidential campaigns to local elections, and most continue to advise political campaigns. Figure 2.1 arrays respondents' pseudonyms by years of political experience. I interviewed four Republicans and ten Democrats of which 10 were men and 4 were women. Most respondents had between 5 and 25 years of experience with the median being 9.5.[32]

Another way to describe sample diversity is by their experience and backgrounds. Figure 2.2 lists the respondents' current job roles at the time of the interview. Many had backgrounds in voter data management, analysis, and sales. The second most common backgrounds were general consultants and direct contact specialists. The general consultants I spoke with worked for all-in-one consulting firms that provide a variety of strategic, research, and media buying services and often contract with campaigns from start to finish. Direct contact specialists represent field organizers and direct mail consultants, all of whom now work as professional consultants but started as low-level volunteers and staffers. Two respondents also specialized in social media and have experience on campaigns and as digital consultants. The final two interviewees were a campaign communications director currently employed as a campaign spokesperson and a paid

[31] The pseudonyms are some of the most popular names in the United States according to the Social Security Administration. They do reflect any racial or ethnic background. I intentionally exclude this identifying information.

[32] As confirmed by many interviewees, the lower average age is the result of many people "aging out" of the campaign-side of politics largely because the demanding schedules, excess travel, and other typical work-life-balance issues. As result, most of my respondents did not have many connections to recommend individuals who were older and had more experience.

media consultant who makes expensive television ad buys.

**Figure 2.1: Interview Sample by Party, Gender, and Experience**

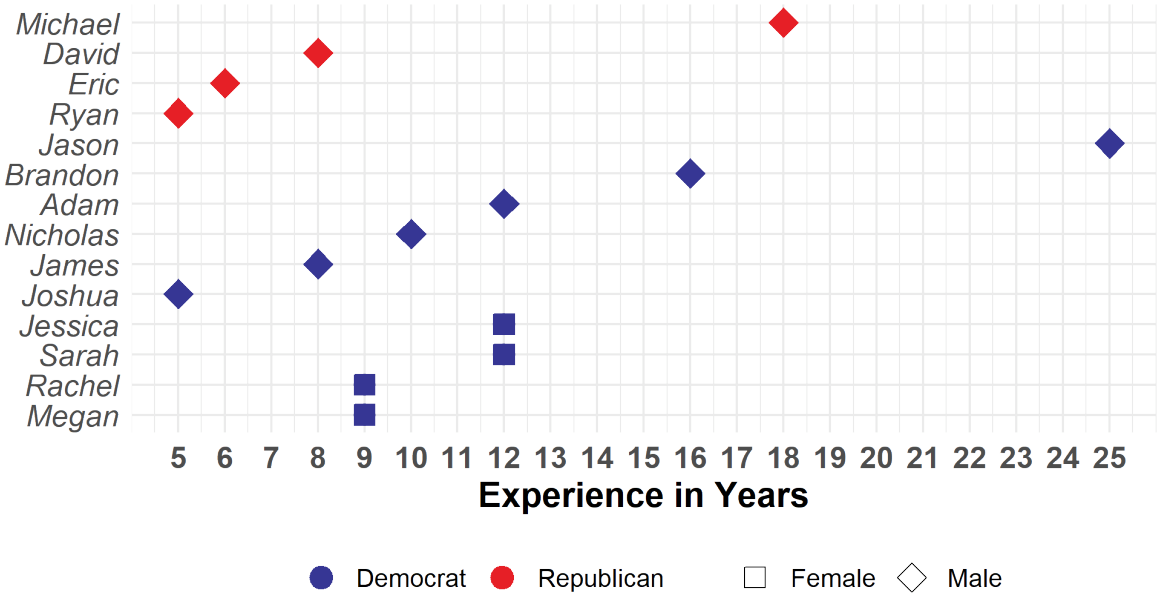
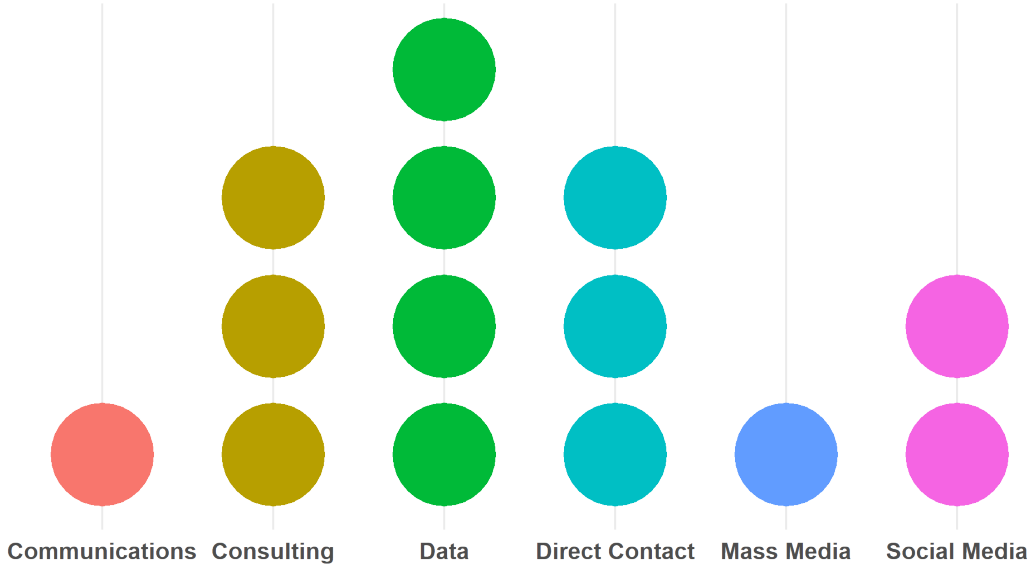



**Figure 2.2 Respondent Specialization**



As explained in chapter one, my analysis of interview transcriptions seeks to contextualize previous literature by examining the similarities and differences in responses to questions. Because respondents were asked the same set of questions in semi-structured interviews, I can directly compare their responses to questions to gain comparable insights. Similarly, throughout my analysis, I will refer to respondents' party, gender, years of experience, and background. While I do not use characteristics as "independent variables" in a statistical sense to explain the "dependent variable" of their responses to my questions, I do use some of their characteristics to draw contrasts and tease out differences between political practitioners' perceptions of voter data and their impact on campaign voter outreach activity. Ultimately, I take the view that my sample's diversity is an advantage to understanding their perceptions as much as its lack of diversity limits the generalizability of my findings.

## 2.2 Old Goals, New Data, and Electoral "Buckets"

Although new data sources have transformed how political experts learn about their electorates, contemporary campaigns still face many of the same strategic choices that existed in the era of pre-data politicking. This first section describes how large, individual-level voter databases have made targeting of direct contact activities more efficient, helping campaigns move beyond precinct and neighborhood-level targeting. Yet my conversations also highlight how campaigns continue to rely on an electoral segmentation strategy that has been a staple for years prior. While the practitioners I spoke to reiterate previous research suggesting differences in partisan data ecosystems, my findings document how both parties share a common strategic language.

The political science literature has studied turnout and candidate choice extensively, but most studies underemphasize the extent to which these two characteristics dominate voter outreach

strategy. Many studies find that voters who participate in elections are more likely to turn out in subsequent contests, and non-voters are likely to establish a habit of not participating. This creates two groups of habitual voters – those who always vote and those who always abstain (Gerber, Green, and Shachar 2003; Green and Shachar 2000; Miller and Shanks 1996; Verba and Nie 1972). Others take a longer-term view of electoral participation, noting that individuals tend to ratchet up levels of engagement as they age and have more opportunities to habituate voting (Plutzer 2002). Still, many scholars contend that voting should not be described as a habit because most Americans are "casual" or "intermittent" voters who vote in some elections but abstain from others (Fowler 2006; Niven 2004). Notwithstanding these debates, numerous field experiments demonstrate that campaigns are more successful at increasing turnout when they concentrate their "get out the vote" efforts on voters with middle to high-voting propensities. Conversely, campaigns struggle to mobilize individuals with low vote propensities who only turnout seldomly or not at all (Arceneaux and Nickerson 2009; Enos, Fowler, and Vavreck 2014; Green and Gerber 2019; Parry et al. 2008).

In addition to turnout, a long tradition of research characterizes voters' choice between candidates as largely predetermined before general election campaigning begins. The majority of the electorate relies on their partisan identification and retrospective evaluations of the incumbent party during general election contests (Fiorina 1981; Holbrook 1996; Vavreck 2009; Zaller 1992). Campaigns have few opportunities to convince voters to switch sides as most voting decisions are based on partisan and social identities rather than ideology or policy (Achen and Bartels 2017; Campbell et al. 1960; Kinder and Kalmoe 2017) Upwards of 80% to 90% of voters are not open to changing their minds in presidential contests (Berelson, Lazarsfeld, and McPhee 1954; Keith 1992; Mayer 2008; Shaw 2008), but the number of persuadable voters increases in less salient, down-ballot contests (Ambadjes 2014). Even if persuadable voters make up a relatively small part

of the electorate and debates continue on how to define these "independent," "undecided," or "swing voters" (Mayer 2007, 2008; Shaw 2008; Smidt 2017), campaigns often devote outsized resources to identifying and contacting these on-the-fence voters, especially in closer contests (Burton, Miller, and Shea 2015).

My interviews with political practitioners reveal that the observations coming from academic inquiry are fundamental to campaign perceptions of voters. Practitioners described how they routinely simplify voters based on turnout and support dimensions to place them into "buckets." Both Republican and Democratic practitioners said they orient nearly every general election strategy around classifying voters into three broad groups: who supports us, who is on the fence, and who is against us. Megan who works in sales at a Democratic data firm explained that these classifications are: "The only thing to understand. The industry is divided by these buckets." Republican social media consultant Ryan also described the importance of thinking about voters in terms of these broad electoral segments:

> I use the same analogy frequently. I say, you know, it doesn't matter what your religion or background is. If Jesus Christ descended from heaven in front of the whole world and everyone saw it on TV. And he told everyone to vote for your guy. Thirty percent of people still wouldn't do it. That's kind of what I have to impart to them. It doesn't matter what you do, or what anyone says. There are people who are just not going to vote for you, so it's important for candidates and consultants and anyone studying this to understand that that's absolutely a factor. It doesn't matter what level you're at. You have to understand that there are people who are definitely going to vote for you. There are people who you're going to have to work on. And then there's people who are just not going to vote for you.

Across my interviews, practitioners stated that the most important distinction between voters

was "turnout" versus "persuasion." Although using alternate terminology,[33] some of the earliest accounts of voting behavior dating back to the 1940s described how campaigns are most successful when encouraging supporters to turn out or when persuading uncertain voters to support their candidate. The idea of so-called "conversion" where a campaign convinces a party stalwart to switch to the other side is, in contrast, extremely rare (Lazarsfeld, Berelson, and Gaudet 1948). These classic views of voters are ingrained among political professionals. Most, if not all, of a campaign's attention is focused on voters perceived to be mobilizable or persuadable. And nearly every campaign with even minimal resources thinks about both groups. As Jason, the veteran Democratic consultant with 25 years of experience, explained:

> They're happening every campaign, but it's a question of how much you're doing and how much of your resources you're putting toward it. If those voters exist who are being persuaded – whether or not they will vote for a specific candidate or vote for a specific party – and if there are voters who exist that are deciding whether or not they're going to vote at all, then that's just reality. You're communicating with them in some way. It's just a question of whether you're doing it well and in an organized manner. But most campaigns with any resources are devoting some level of resources toward the persuasion piece and some of the resources toward the mobilization piece. And you have to make some judgments to whether it is the best use of your resources.

With a strikingly similar response, Republican communications director Eric explained that the goal of mobilization and persuasion segmentations is to make sure the campaign has enough votes to win the election:

[33] Lazarfeld, Berelon, and Gaudet (1948) refer to "persuasion" as "activation" and mobilization as "reinforcement."

> You'll hear people say, "Oh, it's a base election" or "This is a swing election." The reality is that every campaign is both. It's just a question of to what degree. What percentage of your vote is going to come from properly mobilizing your base versus get-out-the-vote efforts? What is turnout going to be? And what percentage of it is going to be swing, persuading the swing electorate? I mean obviously that varies from election to election. You got to do both. If you can't mobilize and mobilize your base to turn out and you can't persuade that electorate that could go either way, then you're not going to win. You have to motivate your base and win over that tiny silver of five, six, seven percent swing.

Daniel Shea (1996) details a similar "segmentation analysis" in his campaign management textbook that merges political science and practice. The basic logic of segmentation is to array the electorate on a support spectrum and divide it into the partisan vote at each end as distinct from the potential toss-up vote near the center of the continuum. The partisan vote can be further divided into "base" versus "soft" with the latter having the potential to become the swing vote. Either formally based on predictive calculations or informally informed by mental tallies and intuitions, the job of the campaign manager is to estimate support among these groups and their propensity to turnout. For instance, the Republican data analyst David unknowingly paraphrased Shea's textbook:

> What it comes down to is that your mobilization is what we call the "hard Reps" or people who we know are going to vote Republican no matter what. But it's not all Republicans. And then we have the persuasion, which is going to "soft Reps" which would be Republicans living with Democrats or independents. Or independents would also be in the persuasion. Or the soft Dems would be Democrats that vote for Republicans. That is a big universe of people. That's what we call the "swing universe" here. That's one way to look at it. And then you have the hardcore Dems that we're never going to reach.

Democratic social media consultant James put segmentation in more specific terms referencing

voter registration records:

> I think the first thing you want to do is to figure out what the audience sizes are, and probably the first thing you want to think about is party registration if you are in a state where that data is available. Just to get a sense of okay, "In my electorate how many Democrats are there, how many Republicans? How many independents?" And beyond that some basic demographic data. And beyond that, figure out that path to – generally it's fifty-one percent – but a plurality. Whatever will get you there in the most efficient way possible is where you want to go.

Reinforcing these perceptions of the electorate as segments are contemporary sources of individual data on voters that help campaigns target their outreach activities. Former staffers, especially on the Democratic side, found it difficult to separate the idea of segmenting and targeting from the voter scores produced by statistical models. For instance, Democratic data analyst Adam said that for finding voters, "If I had to take two pieces of data with you to my grave, it's probably a standard partisanship model and a standard turnout model." As discussed in the first chapter, especially after 2006, campaigns began to rely on voter-level propensity scores predicted from large individual-level databases. On the Democratic side, the almost-universally shared party database is called VoteBuilder and the interface to access it is called the VAN. Joshua who works for a large Democratic PAC explained how campaigns often choose persuasion targets in the software:

> Tactically, we rely heavily on the models in the VAN. What is their probability of supporting your candidate? Someone with a score of 100 is the most likely to support our candidate, and someone with a score of 0 is the least likely. The cut points are different, each model is weighted differently, and the distributions change. But for the most part, something like people with scores between 30 and 70, these are the persuadable voters. And in party registration states, you'll see that most of them are not Democrats and Republicans. They are the independents who fall into those universes. That's kind of your base layer of people that we're speaking to.

Campaigns use these scores that Joshua referenced to determine who is in the "targeting universe." In this case, he described how campaigns typically come up with their persuasion universe, but the same logic applies equally to figuring out which voters need to be mobilized. An example turnout universe, as my interviewees explained, would be support scores in the range of 80 to 100 and turnout scores starting at approximately 80 down to around 50, depending on the contest. For instance, Megan who works in data sales explained that the turnout universe is typically "high support and high to mid turnout." Commonly, staffers and consultants use the VAN to "cut lists" of voters by specifying voter score ranges and other individual-level characteristics such as demographics to target their canvassing, phone-banking, or direct-mail efforts. They also can add back in response data based on their interaction with voters, such as indications of candidate support from questions asked during contact.

Practitioners also told me how their targeting universe depended on several strategic race considerations. The most important considerations concern the candidate at the top of the ballot and campaign-level resources. The Democratic PAC director Joshua described how many down-ballot campaigns do not have the time or money to make a meaningful impact on turnout, especially in presidential election years, and may instead focus their attention on persuasion:

> Campaigns tend to make a lot of strategic decisions even though the headwinds are away from them sometimes. The presidential election is going to drive turnout and voting behavior. And it's going to lessen the importance of the down-ballot races to the voters. Not because they're not important but because they're just going to get less attention. All of the data in the world and good strategic decision-making will not solve the fact that the presidential election is driving everything. There're just big national things where you're pissing into the wind, for lack of a better term.
>
> There are very few undecided voters in the Trump-Biden race. The composition might change, but there's not some enormous amount of people who are clueless about what they're going to do. But the further down the ballot you go, the more persuadable voters are. There are lots of people who will vote for Trump and a congressional Democrat. There are lots of people who vote for Biden and a Republican.

Democratic consultant Jason compared presidential and midterm differences directly and comes to a similar conclusion that congressional campaigns can often rely on the top of the ballot to foot the bill for mobilization:

> You have to make some judgments to whether it is the best use of your resources. If you're a congressional campaign running in a presidential year, it's really not on your back to be doing the heavy mobilization lift. That can be done by others. And you would see a disproportionate assignment of resources into the different buckets related to persuasion. But if you're a gubernatorial candidate in a non-presidential year, then you're going to have to carry that burden. A lot more of your money is going to have to go on the mobilization track. The top of the ticket can foot the bill for mobilization in that regard.

Voter files, models, and predictive scores are integral to contemporary politicking, yet many Democratic practitioners also expressed their discomfort with the data-oriented mindset of their party. For instance, Brandon who works in paid media said, "I think the left has become obsessed with modeling itself to death." This sentiment was heightened among field organizers and those

who oversaw direct contact efforts. For instance, Sarah, who consults in relational organizing meant to facilitate people motivating their friends and family to vote (and vote for the Democrat) said:

> I don't put a ton of stock into models anymore because, frankly, I think we should have a movement that's not just about an election. And then also look at what we've been doing: narrowing, narrowing, and narrowing and communicating to smaller and smaller parts of the electorate. It's not working.
>
> We lose some of the ground because on the left we hit fifty-plus-one, and then we stop organizing until the next election comes around. That's why the Democrats are screwing ourselves.

Rachel, who works as a field organizer, expressed a similar sentiment about the negative side effects of Democratic data culture:

> I think that some of these models are not super well suited. I also think when we tell ourselves that these formulas work that it is a kind of safety blanket. We don't have to critically examine every step along the way because we agreed on a set of assumptions, built a beautiful slide deck, and wrote an accompanying memo. We all read it. We all understand it. We all agree, so we don't have to really revisit that.
>
> It doesn't matter what your organizers or volunteers are screaming about on the ground, we did the thinking already and we know. We have information that they don't have. So far that hasn't worked out so good for us.

In Daniel Kreiss' book *Prototype Politics,* he argues that perceived party-wide failures can serve as catalysts for technological innovation, such as when Democrats perceived they suffered an unexpected presidential loss in 2004 and subsequently invested heavily in the shared voter database that would become the VAN. The frustrations felt by some respondents reflect internal party tensions about the utility of the data-driven tactics in the wake of Hillary Clinton's surprising defeat. Clinton's loss to Trump left Democrats baffled without agreement on a cause or solution

(Masket 2020). Some organizers placed the blame on an over-reliance on data and modeling. The relational organizer Sarah summarized her critique as: "We're making it about math instead of people."

As the field organizer Rachel's comments make clear, however, the data-oriented culture of the Democratic party runs deep. Party professionals and activists agree on a wide set of assumptions even if some question the utility of big voter data. When observing Democratic campaigns in 2008, Nielsen (2012) details how many staffers trusted and relied on the voter models and propensity scores, regardless of if they did not understand how they worked. Even so, some campaigns and organizations, especially local ones at the time, chose to opt out of the VAN's targeting scheme and manage their own lists. My conversations reveal that opting out is simply no longer an option. Some in the Democratic Party question the hegemony of the VAN, but these predictive scores continue to guide their targeting and outreach practices.

Republicans talked about the importance of taking into consideration voter scores too. But unlike their Democratic counterparts, the Republicans I spoke to did not mention any sort of shared data culture in their party network. They mentioned the opposite – different firms with competing data sources and models. For example, Michael, who runs an all-in-one Republican consulting firm where campaigns can get polls, individual-level data, and consulting, spoke about the different data sources that go into statistical models:

> With how volatile politics is today, having the data and the modeling to go in and make sharper assumptions about what kind of voters are in each household is incredibly important. Everyone has different modeling. That's where it gets into it. The important stuff is vote history. That's critical. Voting propensity is critical. And then age, race, income. And then depending on how big the race is or depending on the resources you have, getting into consumer data and other stuff we have like that is effective.

Even the Republican data analyst David, who got his start in politics back in 2012, spoke about how even by 2018 (the most recent federal election cycle at the time of our interview), his fellow partisans were relying on a mix of individual data sources with varying degrees of accuracy. As he says, Republican operatives often must cobble together different sources of individual data and lists themselves, and there are few established data-sharing practices across Republican campaigns.

> I found when I first started at the party, we had roughly 20 million identification flags on voters in California. And identification flags are single identification flags, meaning like I asked you, "Do you prefer Coke or Pepsi?" And you told me, "Whatever." That's one kind of flag. We have 20 million of these identification flags in California. However, the data quality was so bad that we couldn't trust it. I had to end up scrapping almost all of it. I kept like a million flags. It depends on the sources. That data had come from campaigns over the years that had not kept good notes.
>
> Then you have voter data coming from the voter file, the secretary of state. Their file is not very good. The reason is that the secretary of state only updates their file twice a year. Meaning, twice a year they reach out to counties, get an update, and that's their file. That isn't very accurate because places like LA County in 6 months the entire county has changed. You can't use the secretary of state's file.
>
> Then you have companies like PDI, political data in California, who basically every single Republican consultant and candidate in California uses because they literally go to every single county monthly and get the voter file and update it. So, they have all the data in real-time, and they are not relying on the secretary of state.
>
> But the RNC who wants to come in and help with congressional candidates, they're updating from the secretary of state. Their data is out of date. When they try to give us models on things, they're giving us models based on information that is out of date and probably includes people who are dead or no longer voting.

The conflicts over voter data among Republicans in California that David described are in stark

contrast to their opponents. Many Democrats had access to the VAN application as early as 2008. Kreiss (2016) found that the Republican network of data vendors and consultants remained fragmented as late as 2014. Different Republican campaigns and organizations used (and continue to use) different sources of individual-level data and software.[34] Republicans continue to have many choices besides the platform provided by the RNC. For instance, they can use other Republican-only platforms like i360 or CMDI (Crimson), in addition to many other prominent non-partisan vendors like Aristotle International.

Though Republicans are not united by a common database, my conversations with them suggest that individual-level data still reinforce their perceptions of the electorate as comprising different voter segments. In several interviews, Republican practitioners referred to "affinity scores" to determine voters' level of support. Ryan, who works as a social media consultant, told me that the most important individual-level information on voters for him was vote history, affinity scores, and contact information, such as an address, email, or phone number. Similarly, David later explained that to figure out their mobilization and persuasion targets, a campaign typically turns to similarly predicted voter scores:

> We look at the data and the models. And we say: "What percentage of the electorate are Republican voters? What percentage are weak Republican voters? Which percentages are very likely to vote for Republicans if they come out, but they don't turn out because they're low propensity, high partisanship, lean but low propensity – all that stuff.

Even though access to individual-level data reinforces the segmentation of voters into different universes, campaigns have for years attempted to divide the electorate into these conceptual

[34] For instance, Kreis (2016) documents how in 2014 even Senate Minority Leader Mitch McConnell chose to export records from the party database GOP Data Center and instead use a third-party campaign management application named NationBuilder to run his campaign (177).

buckets with the electoral information available at the time. The 18-year veteran consultant Michael told me about the old precinct-based targeting strategies common in the early 2000s and before:

> Back when I was taught in my first campaign manager school, we used this formula called ORVS – Optimal Republican Voting Strength. Basically, it was a way to score precincts, and we would approach everything by precincts. We would say, "If we assume fifty percent of these people are Republicans because they voted in Republican primaries in the past and fifty percent are Democrats because they voted in Democratic primaries in the past," then we might not want to touch that. Let's only touch the precincts that are sixty and above. In doing that and trying to apply it, you're leaving hundreds of voters out of your target.
>
> Where now, instead of looking at a big precinct analysis to drive your strategy – if you have the resources and you can do it – I'm not concerned about precincts. I'm concerned about individual voters. Having that knowledge and that data and being able to target people down to the individual, even individuals within a household, is incredibly effective. I think if people are not doing that, then you're really running at a disadvantage. You have to remember, especially when you get in these general elections, for a lot of these independents we're using precinct-level data to assume that they are more likely to be Republicans because they live in that precinct.

Many scholars have noted the importance of the shift from precinct-based targeting to individual voters (e.g., Baldwin-Philippi 2018; Hillygus and Shields 2009; Issenberg 2013; Kreiss 2012). As mentioned in chapter one, campaign consultants beginning in the late 1980s and into the 1990s combined sociodemographic data from the census at the block level with precinct voting returns to engage in geo-demographic targeting of specific streets and neighborhoods. Especially after the development of national party databases, many academic and popular accounts described a new era of campaigning that permitted campaigns to "microtarget" individuals. As described by

journalist Sasha Issenberg (2013) and political scientists Sunshine Hillygus and Todd Shields (2009), early 2000s microtargeting involved consultants combining consumer and product purchasing information with compiled national party voter files. In a similar way that geo-demographic targeting generated archetypes such as "God's Country" and Downton Dixie-Style," these new individual-level prototypes informed by consumer data labeled voters with descriptors such as "Bible Believers" or "Republican Intelligentsia" (Issenberg 2013, 132). Michael mentioned this transition and the arrival of competing Republican firms engaging in microtargeting:

> When I was just starting out in consulting, microtargeting was this huge, hot deal. And Karl Rove had used it. All these microtargeting firms had started popping up. Of course, microtargeting is good. There's still some level of microtargeting that is still around and how it was built then.

Even with the shift toward voter databases generating predictive support and turnout scores, many practitioners were quick to push back on the notion that the arrival of individual-level data had fundamentally changed their strategic considerations when communicating with the electorate. Instead, they explained how individual-level data are only one factor among many to consider. When I asked specifically about the transition away from precinct-based targeting while mentioning Issenberg's book, the Democratic data analyst Adam quickly contested the notion that politicking had been completely overhauled by individual-level data:

> It's a great fucking fantasy world. But it's not the world that any modern campaign I know is living in. Because we know that to reach all the people that we care about in a really persuasive way, we can't limit ourselves to being that picky. There is a reach-precision tradeoff. You can deliver exactly the message you want to exactly the people that you care about, but the problem is that you're only going to be able to do it for 25% of the people who will decide an election. And there are 40 different reasons that are worth understanding but are not worth unpacking here. You can also reach 100% of the people you care about but in a really imprecise way by only doing earned media and only doing broadcast.
>
> The idea that every decision a campaign makes will be individually targeted is divorced from an understanding of how campaigns work. There are plenty of mediums that can. You can send mail as individual, but there are so many different ways that campaigns reach people.
>
> But more importantly, there are only so many different ways that campaigns can delegate their dollars. It's not shocking to you, but more than half of campaign dollars are spent on paid media – television. And it is more true today than it was 20 years ago that you can target paid media at the individual level, but there is a whole bunch of it that you can't. Campaigns will usually be very sad and frustrated if they want to make the perfect the enemy of the good and only deliver their message in a tailored, individually targeted way.
>
> Love and respect Sasha, but he was spending too much time talking to [the 2012 Obama campaign's chief analytical officer] Dan Wagner and missed the forest for all the other ways campaigns get their message out.

In other words, as much progress has been made with individual-level targeting, voter databases can only inform a small portion of campaign strategy. As Adam explained later in our interview:

> I think that there are certain things for which voter data is much more rarely going to inform. For example, we were pitching a U.S. Senate candidate not long ago. There were certain parts of the deck in the presentation where we said, "We think voter data needs to have a huge say in these handful of decisions." But like you, who you are, where you come from, and what your biography is – if you want to be the next U.S. Senator and you need voter data to answer that question – you probably shouldn't be running in the first place.

The veteran Republican strategist Michael similarly explained how data, regardless of its sources, can only be incorporated into decision-making after establishing who the candidate is and how the campaign can message him:

> The most important thing is the candidate. More than ever, the candidate is at the top of the list. I think messaging is second. If you don't have a good message, it doesn't matter how good your data is. It doesn't matter. Candidate, messaging, and data I think is third. That's the hierarchy that I put on campaigns, and that's how we go through campaigns. Who is the candidate? What are their strengths? How do we position them? How do we define them? What is our messaging? We have to make damn sure that that's on point. And then after, we have our candidate, and we have our message. Now, who do we talk to? And how do we know with confidence that we're talking to the right people to get the number of votes we need to win and reach our vote goals?

In short, campaign strategists have long sliced the electorate into different groups to maximize their return on investment from contacting voters. While Republicans and Democrats may be embedded in different party data cultures and voter information environments, they both share a common strategic perception of electoral segments defined by a combination of turnout and support propensities that has been intensified by the development of large voter databases. Both parties agree that they must leverage sources of electoral information, whether polls or individual-level data, to separate a complex electorate into understandable components meant for voter

outreach activities.

## 2.3 Putting the Pieces Together in Practice

This section builds on the last by documenting how practitioners perceive "voter data" and their methods of combining different data sources during a general election campaign. My conversations reveal how practitioners often have vastly different conceptualizations of voter data largely because of their everyday experiences dealing with sources of electoral information. Because Republicans are not united by a common database or an approach to data-driven campaigning, their integrations of sources such as polls and individual-level data are more often makeshift compared to their Democratic counterparts who have technical systems to combine inferences from disparate sources.

Much of the scholarship about down-ballot campaigns' interaction with different data sources comes from ethnographic and interview-based qualitative investigations from political communication scholars Jessica Baldwin-Philippi and Rasmus Nielsen. They focus on how staffers engage with both digital technologies, such as social media and email, along with their access and use of large individual-level databases maintained by third-party data vendors. Combined, they make two crucial contributions. First, the descriptions of data-driven campaigning reported in the press and the rhetoric coming from political data vendors describing their sophistication are often at odds with their haphazard applications on the ground (see also Simon 2019). Second, lower-level campaigns' lack of aptitude with data and technology is not merely a function of practitioners lacking experience and data know-how but also the result of limited resources. Even the most well-funded campaigns do not have the staff, time, or cash to integrate and make use of all the data and information available (Baldwin-Philippi 2016, 2016, 2017, 2018,

2019; Nielsen 2012).

Voter data continue to mean different things to different people, and political practitioners are no exception. I asked each of them to define "voter data" in their own terms. Their conceptualizations fell into two categories. The first was limited and solely referenced databases based on state voter registration files. David, the Republican data analyst, defined the term as information that voters "provided themselves to a registrar of voters, the secretary of state, or to a campaign." The view was more comprehensive and considered many different sources of electoral information. For instance, the Democratic data analyst Adam conceptualized voter data as "an amalgamation of data that I acquire from various governmental, civic, consumer, and political sources" that also include aggregate information that is otherwise "very challenging to get at the individual level." Table 2.1 provides two contrasting definitions of voter data. Field organizer Rachel's definition is the most limited. She described how her conception is essentially only the public voter file and the information that states record during registration and voting. At the other end of the spectrum is Jason. He took a broad view of voter data as any potential source that can be merged to make predictions about voters. These sources included individual-level data either collected from secretaries of state, by the campaign, or elsewhere in addition to the results of surveys and aggregate information related to the district and precinct.

**Table 2.1 Example Definitions of Voter Data**

| Limited Example | Comprehensive Example |
|---|---|
| Really, it's just like the public voter file. It's a list of people that are registered to vote, and then we also have the information on how frequently they vote, whether they vote in primaries, generals, whether they were registered, and when they were registered.<br><br>All of that kind of paints a little bit of a picture that we can use to make decisions about which tactics we'll deploy on them, which universes we'll put them into, how we reach those voters, and whether they're voters we want to reach out to in the first place.<br><br>*– Rachel, Democratic Field Organizer* | I look at it two ways. There are individual-level voter data, and there are collective voter data. We tend to target down to the individual, and we do.<br><br>But we also do model up from that and do still rely on traditional public opinion data. Voter data exist at multiple different levels, and we use it at multiple different levels.<br><br>There're access points for that too. There's individual voter data that is based on actual contact, and then there's individual voter data based off modeling processes and then applied to individuals. Technically, I wouldn't call that individual-level voter data because it is based off aggregate voter data and analytics, which again is important.<br><br>*– Jason, Democratic General Consultant* |

In a technical sense, Jason's definition comes closest to how political data vendors build predictive voter scores. James, the Democratic social media consultant, described the process of integrating multiple information sources as "layering" where the "foundation" is the state voter file. After acquiring registration rolls, data vendors clean the contact information on file, such as verifying addresses or phone numbers. Jessica who works at a direct mail consulting firm told me how their firm relies on the U.S. Postal Service's National Change of Address database to either correct or at least remove bad addresses. Data vendors next purchase and merge information from consumer records from companies such as Acxiom or Nielsen. Consumer records may include purchasing patterns identifying such categories as gun or pet owners or real estate records indicating home ownership, but political data vendors are most interested in sociodemographic

information like estimated income and race. Another common source of individual-level information is preexisting donation databases that track which voters have contributed to parties, campaigns, and other party-aligned causes and interest groups. The last piece of individual information is response data from voters' previous interactions with campaigns. Layered on top of the combined individual-level data, as Jason explains in Table 2.1, is aggregate data, including precinct-level election results and other geographic-based data coming from the Census, including income and race estimates to help data modelers proxy these characteristics at the individual level. James described how his information is then leveraged to create predictive scores on voters:

> I would base the definition on the information that can be pulled from a state's voter file. That's the foundational piece of information about a voter. That's usually not a whole lot. Then on that foundation campaigns and consultants have layered on additional data that they have pulled from other sources – whether that be consumer data, data from past campaigns. Polling data, that's layered onto that base layer of voter file information, which is usually just like party affiliation, gender, age, and location.

Political data vendors build models and predictive scores with common multiple regression techniques or machine learning models such as random forests that are trained using the results of surveys (Nickerson and Rogers 2014). The basic process is to take a random sample of voters from the database, run surveys that ask issue and support-related questions for specific candidates or generic partisan identification, and then build a model based on the survey results to predict scores for the rest of the database population. Democratic data analyst Joshua described how he builds a partisan support model: "It's as simple as you survey a group of people and literally just ask them a straight-up [partisan] ID questions, and then we'll do that with a pool of like 3,000 or 4,000 people and then they're stratified by age, race, and gender." As discussed in the first chapter, although firms can potentially add thousands of variables to the models, the most valuable pieces

of information tend to be what can be collected from the voter file and basic demographic characteristics. Data director Nicholas described how the VAN partners with other Democratic data firms like Catalist or Targetsmart to build models using as few as 10 variables:

> I honestly don't think there's a lot that we need. The voter file Catalist and TargetSmart have is like thousands of columns wide. It's huge. And like even when people are building models, they're using like only 20 or 30 of the listed variables. Well, it's probably closer to 10. I think that past vote history is always important, demographics, and socioeconomic data. That might be where some of the more commercial stuff is useful. If there's a way to know if someone has graduated college or what their income levels are, that sort of thing is helpful too. Obviously, those aren't always great, but those are the biggest things of value.

In a practical sense, the main reason for different conceptualizations of "voter data" is the result of how political professionals interact with data sources. On the Democratic side, most practitioners are not involved in the process of creating voter scores. Instead, they rely on the predictive scores produced by data analysts they have never met. Brandon who makes a living facilitating Democrats' mass media purchases told me how lower-level campaigns have little choice but to lean on voter scores to help them target their direct contact activities given that their lack of resources precludes more sophisticated data processing and analysis:

> In terms of modeling and all that stuff, that is more of a tactical implementation for a campaign. Honestly, most campaigns can't afford to have a dedicated data analytics person on them. If you're in a good, targeted race, you'll have the support of a party committee or one of the party committee's consultants who helps with that stuff. But otherwise for campaigns, it's more about executing the tactics rather than crafting the strategy.

As a result, most Democrats not involved directly with data analytics conceptualize voter data as individual-level records because that is what they see when they access the VAN even if voter

scores are modeled from a variety of data sources at different levels of aggregation.

Some of the Democrats I spoke with talked about the importance of paying attention to other sources of information such as traditional polls and even qualitative feedback, but most of them explained how the VAN should be (and is) the primary way in which their campaigns seek to understand and segment the electorate. For instance, general consultant Jason with 25 years of experience said:

> Certain qualitative things like focus groups, online surveys, things like that. I think that has more utility in defining your campaign strategy than the analytics do. I think that's something that should probably be front-end for most campaigns. And then analytics come in and tell you how you can do something or what about it you can do, and they help you to refine a strategy that may have a lot of blunt edges. But when the analytics are the front-access point for strategy, you tend to run into problems.

James who currently works in the digital space spoke to me about the value he sees in analyzing traditional crosstab survey results to inform how to segment the electorate with voter scores and sociodemographic information for fieldwork:

> I think polling and specifically diving into the crosstabs of a poll are going to be incredibly effective. The voter file is like the physical people and the numbers in terms of how many people in the demographic physically exist, and then the polling, for me, is an indication of how those demographics care about the candidate and using that polling to inform how I can slice an audience in the voter file.

Nevertheless, the overwhelming sentiment among Democrats was to prioritize the VAN even if they recognized its limitations and sometimes questioned its usefulness. Many expressed a perceived disconnect between modelers and organizers. Brandon who makes a living facilitating Democrats' mass media purchases told me how his biggest issue with data-driven practices is not the raw individual records themselves but rather the opaque process of generating predictive voter

scores that are appended to those records. “I trust voter data. I don’t necessarily trust the people who create scores or apply scores.” Democratic social media consultant James who got his start as a field organizer told me about how his perception of voter data changed during his eight years in politics as he graduated from field organizer to professional consultant:

> As a field organizer, you think the data is bullshit because you’re calling all the numbers and getting all the disconnects and getting all the wrong numbers. Doing that honing of the voter file manually, you have a different perception of voter data when you’re making 300 calls a day and getting all the wrong numbers. And in that position, you really don’t have much ability to decide the strategy. I think once I got beyond that point, the voter file is definitely important in that in most elections it is not going to change radically from when you start working on the campaign to the end.

Even field organizer Rachel who bemoaned the data-driven culture of the Democratic party told me how every campaign should “open up VAN once in a while and actually bother to cut a list.” She added:

> We really put data on a pedestal partly because it’s technical and not everyone has the expertise, so I think we’ve seen qualitative feedback take a back seat and also gut instinct take a back seat. And those things are also important and worth paying attention to. It is an important and absolutely necessary tool, but it should always just be one tool in the arsenal. It should never come at the expense of any of those other [qualitative] things.

Compared with Democrats, all four Republicans described a much more informal, campaign-level system of interpreting different sources and gathering lists that did not solely rely on predicted voter scores. Even the data analyst David, for instance, described polling and individual-level voter data as being disjoint from one another and when to trust the signals of one versus the other:

> A really great example is polling versus individual contact information. We would have polling in an election that would say one thing. Then, when we talked to the voters – a sample of a thousand people in an area – they would say a different thing. We would have different projections of what an election was looking like at the time. In those situations, I trust the voters. I trust the people I talked to. Polling tends to rely on samples and to rely on weighting to get an accurate picture, and it talks to an N of 600 from a district. Versus if I talk to 40,000 people in a district, I have a much better idea of what people look like, so I tend to trust my voter data more than polling.

All the Republican practitioners also described rather ad hoc systems of dealing with individual-level data in contrast to the more technical integrations reported by Democratic data analysts. Data analyst David described how dealing with different lists and sources of individual-level data requires "a little bit of political acumen." He further clarified: "Every company has consultants and data people who have worked in the field for a while that know what to trust and what not to trust and know the questions to ask." Republican communications director Eric described how he has dealt with multiple different databases himself:

> You name it. I've probably used it in terms of voter databases. It's how I've integrated it into messaging. In a primary campaign, it's great to have the voter data because then I can throw out people who don't vote in primaries. If it's a presidential primary or an off-year primary, you'll have to look at folks who voted in 2 or 3 of the last 4 primaries and focus on them rather than focusing on those less likely to vote. Same thing if I have consumer data or some third-party data that they've signed a pro-life petition or are an NRA member. And there are lists out there that exist on all those things. Then I can target those people with specific appeals.

Eric continued and described a scenario in which his campaign had to resolve conflicting information signals in a competitive race with "gut" decisions:

> Well, it's scary. A lot of it is, at the end of the day, there are times that you have to go with your gut, but you try to have as many data points as you can. In fall of 2018, a poll showed the race tightening, and we were down several points. And then we were down one. And then we were up one. And later we were down one, so it kind of locked up there. As it moved from us being down several points to us being up one, we knew to trust that not only because the [public polling agency] poll was showing it, but because our data was showing it. Our data modeling was showing it. It was happening at the same time that we were running messages that tested well, so you have a couple data points there. You know that your modeling is showing it. You know that the messages you are running are the ones that you thought might do that sort of thing.
>
> Usually, a campaign will do an internal poll three, maybe four times throughout the campaign cycle. But sometimes at the end, you'll have what they call brush-fire polling, you know, rapid-fire polling. Or you have a data tracker poll, which is kind of a variation on that. And the daily tracker polls will drive you nuts because you're looking at it every morning to see if we're down, and it's nerve-racking.
>
> But as you get into those final weeks, you probably got your modeling. You probably got a recent internal poll that you did sometime in the fall. You probably got a recent [public] poll. And you've probably got some form of daily tracking or brush fire. So, all you can do is see how much of the data is moving in the same direction. If there's something that's going in the opposite direction, can you explain how that is? If you can, great. If you can't, that's a problem. Between all those datasets, you try to determine if you're headed in the right direction.

As discussed in chapter one, campaigns traditionally fielded polls of different lengths throughout the election season to learn more about their levels of support across different demographic groups in the electorate. Eric's recounting of a particularly close race reveals his mindset. Political practitioners must combine poll results, individual-level data, various statistical models, and gut feelings together to inform campaign strategy. He views them as disjoint pieces of information that must be interpreted holistically in the mind of the strategist. What he does not

reference is any technical processes or advanced statistical modeling to derive insights from the streams of electoral information.

Ultimately, my conversations with professionals on both sides of the aisle have different approaches to dealing with electoral information. The biggest practical difference is the extent to which practitioners have control over the interpretation and implementation of different sources of data. As Republican practitioners explained, they often must determine the strategic value of different data on a case-by-case basis. Whether choosing an individual-level data vendor or weaving together the results of traditional polling, most of their data-based decisions are highly situational and made in response to their personal views of an evolving race context. In contrast, Democratic consultants and practitioners exercise much less direct control and instead must rely on statistical modeling conducted by outsiders who do not work on the campaign. As explained by Democratic data analysts, these models mathematically combine the results of polls and surveys, but most operatives only see the predictive scores they generate displayed as individual-level data in the VAN.

## 2.4 Summarizing Data Perceptions

Campaigns have access to more data on voters than ever, but how on-the-ground practitioners make sense of all of it varies from person to person and race to race. While practitioners shared a common perception of the electorate as made up of different strategic segments based on underlying propensities to support a candidate and turn out in an election, they had different beliefs about the value of data sources to inform their voter outreach strategies. Echoing the scholarly and historical records of data-informed campaigning discussed in chapter one, this chapter reveals how, even by 2020 when I began conducting interviews, the two parties maintained different

approaches to the practical application of electoral data sources. Democratic perceptions of the electorate are difficult to separate from the party's shared VoteBuilder database both in terms of conceptualizing the electorate as "buckets" of persuasion and mobilization targets and in their use of the VAN application to segment the electorate for the direct contact activities. While Republicans also think of the electorate in terms of these two segments, their combination of different sources is much more ad hoc relying on the decision of campaign staffers to choose which sources of individual-level data to use and how to interpret the results of polls and surveys.

Two metaphors I was told summarize the persistent differences between Democratic and Republican practitioners in their beliefs regarding the extent to which campaigns should incorporate strategic considerations and qualitative feedback into campaign-level decision-making. In the first, Democratic direct mail consultant Jessica discussed the importance of individual-level data in reference to an episode of the NBC television series *Parks and Recreation.* In the episode, the character Leslie Knope is running for city council and becomes obsessed with winning over the vote of one man who described her as "not the kind of person you can go bowling with" during a focus group.

> Data is really important. And I know you know that, and I know that. The biggest struggle I have with candidates as a strategist is communicating those buckets. I think that that's the hardest thing. And do you watch Parks and Recs? I think about the episode when Leslie is obsessed with that guy who doesn't like her in the focus group and then tricks him into going bowling with her. That's very typical of candidates. They don't want to lose people, or they think they will be the one to change people. That's the struggle with candidates. But when you involve the candidates, that's when they want to throw the data out the window.

In other words, Democrats have a data culture that is reinforced not only at a structural level with a party database shared by candidates and progressive causes across the country but also at a

practitioner and campaign level. Even the Democratic practitioners who were critical of the data-oriented culture of their party saw their critiques as ultimately falling on deaf ears. In many ways, what political communication scholar Rasmus Nielsen (2012) described as the "dominant targeting scheme" defined by large individual-level databases, predictive models, and staffers reinforcing their application has become the *only* scheme among Democratic operatives.

Contrast Democrats evangelizing the VAN to their candidates with Republican consultant Michael's view that he plays the role of an offensive coordinator in American football. Polls are akin to what is happening on the field, and then his role is to come up with a play and implement it using the voter file.

> I equate it to football. You know the offensive coordinator with the laminated sheet that has all the colors on it, calling the plays? That's really what it looks like more times than not. A poll says these three messages are really good here, but then you have work to do with these people and you have work to do with these people. These people want this message. And these people want this message. You can get subsets of the voter file for different demographics, and you got ways to deliver the message they need to hear.

In many ways, the Republicans I spoke with described their role as a complicated challenge of combining and weaving together polls and other information to come up with a campaign-specific playbook to respond to an evolving race context. Republicans not only lack a shared party database but also a consensus on how to approach data-driven politicking.

Having listened to political practitioners describe the various factors that potentially impact their use of different data sources in detail, the remainder of this dissertation provides the first systematic investigation into the types and kinds of electoral information that campaigns have come to rely on. In the next chapter, I turn to detail my refinement of millions of expenditures and track the diffusion of data technologies into congressional campaigns over a period of marked

innovation.

# 3 A Complete Picture of Data Sources

The qualitative evidence from the last chapter suggests that campaigns further down the ballot have come to rely on large individual-level voter databases that help them slice the electorate into different strategic segments, especially since HAVA mandated statewide voter registration lists in 2006. As detailed in the first chapter, many accounts have documented innovations in data technologies over time, but the bulk of their attention has concentrated on unrepresentative presidential campaigns with massive budgets to invest in novel data practices and thousands of staff and volunteers to implement their electoral outreach strategies. This chapter creates a complete picture of different data sources and campaign technologies by examining millions of spending records made by congressional campaigns. To my knowledge, it represents the first effort to decompose a substantial number of candidates' spending patterns to determine the specific electoral information sources they purchased.

This chapter unfolds in four sections to provide a comprehensive overview of an electoral information environment that is defined by partisan differences. As previewed in chapter one, the primary purpose of the following analyses is to reveal campaign data preferences by examining their marketplaces, costs, and campaign-level spending patterns. I first detail my meticulous refinement of FEC records to recover approximately 50,000 data-related expenditures from 3.5 million total records produced by 2,500 unique congressional campaigns between 2006 and 2018. With these verified records of data spending, I document the partisan marketplaces of voter data vendors and polling firms. I describe how both parties have networks of partisan companies that maintain individual-level data and polling markets. The central difference between Republican and Democratic markets is the level of competition and stability over time. Republican markets more

closely mirror free markets with more companies and less consolidation. Next, I leverage the data-related expenditures to provide the first cost estimates of electoral data. While Democrats pay slightly more for both sources of information, the price differences are marginal and most likely the result of unique partisan market forces. The last section details campaign-level variation in the data source spending patterns, which I further unpack in the next two chapters. Democrats running for Congress have come to embrace individual-level data much more than their Republican counterparts who continue to spend more of their limited budgets on polling. Altogether, these analyses help to elucidate why the partisan practitioners I spoke with described different data-driven practices.

## 3.1 Refining Expenditure Records

To create a complete picture of data-driven campaigning for Congress, I undertook an extensive, year-long review of approximately 3.5 million campaign expenditure records reported to the Federal Elections Commission (FEC) by major party candidates running for the House of Representatives between 2006 and 2018. As detailed in the first chapter, my seven-cycle sample included 2,515 unique campaigns (4,910 unique candidate-cycle campaigns) with 3,429,292 reported expenditures approximating $6 billion. Table 3.1 reports the number of districts and candidates by election cycle. The table also provides the total number of expenditures reported by House candidates and their cumulative overall spending across both -the primary and general elections. Expenditure counts fluctuate over time but remain relatively consistent with an uptick in 2018. A similar trend appears in overall spending. The 2018 House election cycle was the most expensive midterm election up to that point in history. The bulk of this spending came from Democratic candidates who outspent their Republican opposition by approximately 38%,

according to estimates from the OpenSecrets.[35]

**Table 3.1 Candidates, Expenditures, Spending by Election Cycle**

| Cycle | Candidates (N) | Total Expenditures (N) | Total Spending ($) |
|---|---|---|---|
| 2006 | 691 | 396 K | $761 M |
| 2008 | 688 | 496 K | $863 M |
| 2010 | 707 | 500 K | $886 M |
| 2012 | 688 | 493 K | $889 M |
| 2014 | 709 | 499 K | $844 M |
| 2016 | 689 | 462 K | $793 M |
| 2018 | 738 | 584 K | $1,273 M |

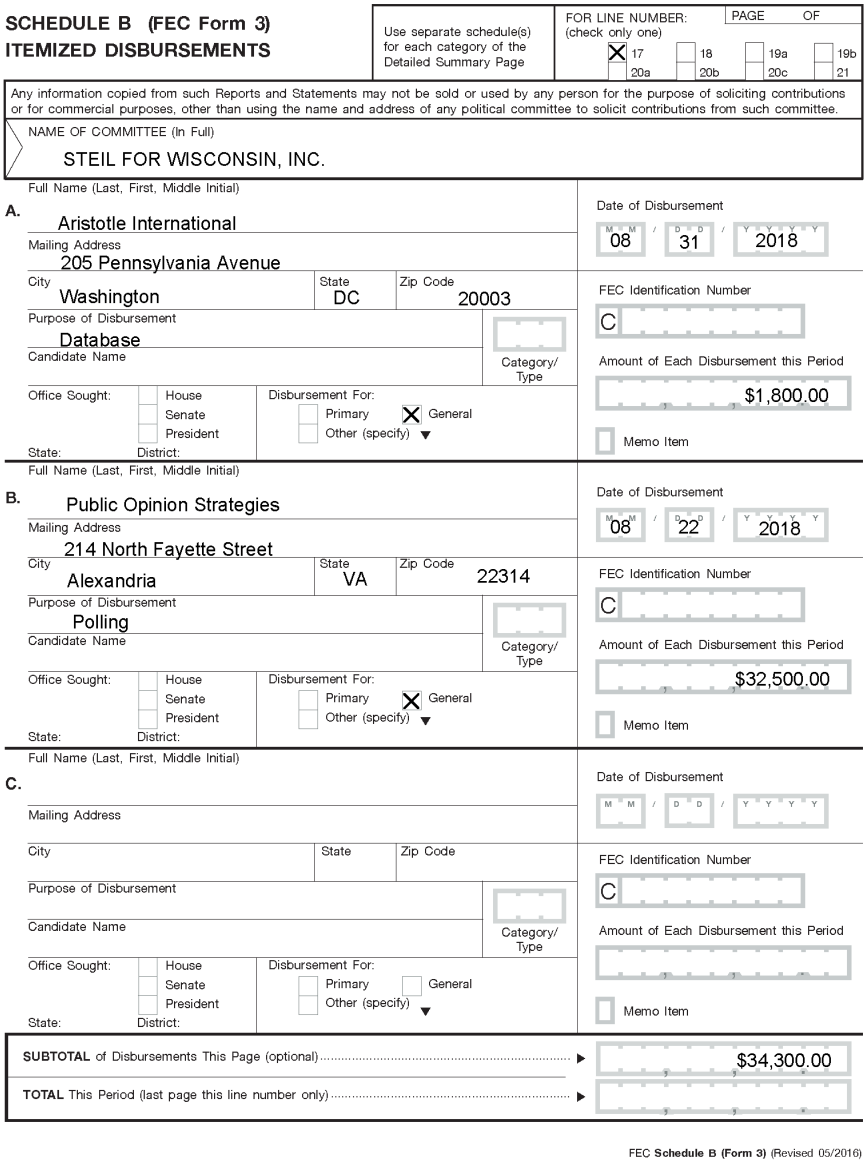
The FEC requires House campaigns to report making disbursements for operating expenditures for any purchase over $200 or if purchases total more than $200 from a single payee in an election cycle.[36] Nearly every campaign files its expenditures reports with the FEC electronically. Figure 3.1 is an example of the required itemized disbursement Form 3 submitted to the FEC by Bryan Steil for data-related purchases during the general election campaign period for Wisconsin's first congressional district in 2018. Most expenditures disclose the purpose, date, and dollar amount of the disbursement as well as the name and address of the payee. FEC guidance requires campaigns to provide the purpose of the disbursement to the extent that a person could reasonably interpret what was purchased when considered alongside the identity of the recipient. The FEC also provides campaigns with a list of 12 expenditures category codes, such as "administrative/salary/overhead" or "media" expenses, but most campaigns do not bother to fill out category codes when making their disbursement reports, as exemplified in the blank "Category/Type" fields in both of Steil's itemized expenditure examples from Figure 3.1. On

[35] https://www.opensecrets.org/elections-overview/cost-of-election Note my sample reports lower total spending because I restricted my estimates to Democratic and Republican candidates who ran in the general election.

[36] https://www.fec.gov/help-candidates-and-committees/filing-reports/operating-expenditures/

average, congressional campaigns leave the category field blank in approximately half of their itemized disbursements every cycle.

**Figure 3.1: Example of FEC Form 3 Itemized Disbursements**



SCHEDULE B (FEC Form 3)
ITEMIZED DISBURSEMENTS

Use separate schedule(s) for each category of the Detailed Summary Page

FOR LINE NUMBER: (check only one)
☒ 17 ☐ 18 ☐ 19a ☐ 19b
☐ 20a ☐ 20b ☐ 20c ☐ 21

PAGE OF

Any information copied from such Reports and Statements may not be sold or used by any person for the purpose of soliciting contributions or for commercial purposes, other than using the name and address of any political committee to solicit contributions from such committee.

NAME OF COMMITTEE (In Full)
STEIL FOR WISCONSIN, INC.

A.
Full Name (Last, First, Middle Initial)
Aristotle International
Mailing Address
205 Pennsylvania Avenue
City: Washington | State: DC | Zip Code: 20003
Purpose of Disbursement: Database
Candidate Name:
Category/ Type:
Office Sought: ☐ House ☐ Senate ☐ President
Disbursement For: ☐ Primary ☒ General ☐ Other (specify)
State: District:
Date of Disbursement: 08 / 31 / 2018
FEC Identification Number: C
Amount of Each Disbursement this Period: $1,800.00
☐ Memo Item

B.
Full Name (Last, First, Middle Initial)
Public Opinion Strategies
Mailing Address
214 North Fayette Street
City: Alexandria | State: VA | Zip Code: 22314
Purpose of Disbursement: Polling
Candidate Name:
Category/ Type:
Office Sought: ☐ House ☐ Senate ☐ President
Disbursement For: ☐ Primary ☒ General ☐ Other (specify)
State: District:
Date of Disbursement: 08 / 22 / 2018
FEC Identification Number: C
Amount of Each Disbursement this Period: $32,500.00
☐ Memo Item

C.
Full Name (Last, First, Middle Initial)
Mailing Address
City | State | Zip Code
Purpose of Disbursement
Candidate Name
Category/ Type
Office Sought: ☐ House ☐ Senate ☐ President
Disbursement For: ☐ Primary ☐ General ☐ Other (specify)
State: District:
Date of Disbursement: M M / D D / Y Y Y Y
FEC Identification Number: C
Amount of Each Disbursement this Period:
☐ Memo Item

SUBTOTAL of Disbursements This Page (optional) ▶ $34,300.00
TOTAL This Period (last page this line number only) ▶

FEC Schedule B (Form 3) (Revised 05/2016)

The nonprofit OpenSecrets supplements FEC records with extensive coding and cleaning to provide greater insights into campaign spending patterns. OpenSecrets' stated purpose is to track

money in politics and understand its impact on elections.[37] One of its transparency efforts is to better categorize FEC expenditure records. OpenSecrets researchers review the purpose and payee to provide an appropriate classification for each expenditure.[38] These categories include many specific spending categories, for instance, such as "travel," "postage/shipping," "salary," "web ads," television ads," "campaign materials," and "polling." In total, OpenSecrets provides 46 expenditure categories.

OpenSecrets' efforts to clean and categorize FEC records are highly valuable but require yet more processing. For my purposes, I am interested in their categories related to electoral information sources.[39] Specifically related to different data sources are the OpenSecrets' categories of "Polling & Surveys" and "Campaign Data & Technology." The former category captures polling expenditures well, but campaign-side errors inputted into the purpose description along with ambiguous purpose descriptions make even OpenSecrets' best-faith effort to classify the latter data and technology expenditures quite a difficult task. Additionally, because I am interested in individual-level voter data, it was necessary to verify that campaign technology was specifically related to large voter databases.

The most comprehensive study of expenditure records to date that relied on a keyword search was unable to distinguish between different forms of data beyond documenting spending on polls (Limbocker and You 2020). Other automated content analysis methods and machine learning

[37] https://www.opensecrets.org/about

[38] OpenSecrets also seek to reduce the number of duplicate records present in the unprocessed FEC records. https://www.opensecrets.org/campaign-expenditures/methodology

[39] As detailed in Appendix B, I employ a similar effort to classify and verify voter outreach records, which serve as my primary dependent variable in chapter five. As a quality control check on OpenSecrets' own categorization, I include categories that are seemingly unrelated, such as events, administrative costs, and other overhead, excluding only contributions and transfers to other political entities like parties and committees and staff salary. A separate review of these remaining OpenSecrets categories indicates that they do not include expenditures related to either sources of electoral information or voter outreach.

applications have only limited success given the short text of the purpose description and the campaign-side errors when inputting expenditures (Williams, Gulati, and Zeglen 2020). Most studies of FEC expenditure records as a result focus their investigations on a limited number of records or candidates often within specific cycles or on a subset of consulting firms and vendors that campaigns hire (Cain 2013; Kolodny and Dulio 2003; Martin and Peskowitz 2015, 2018; Nyhan and Montgomery 2015).

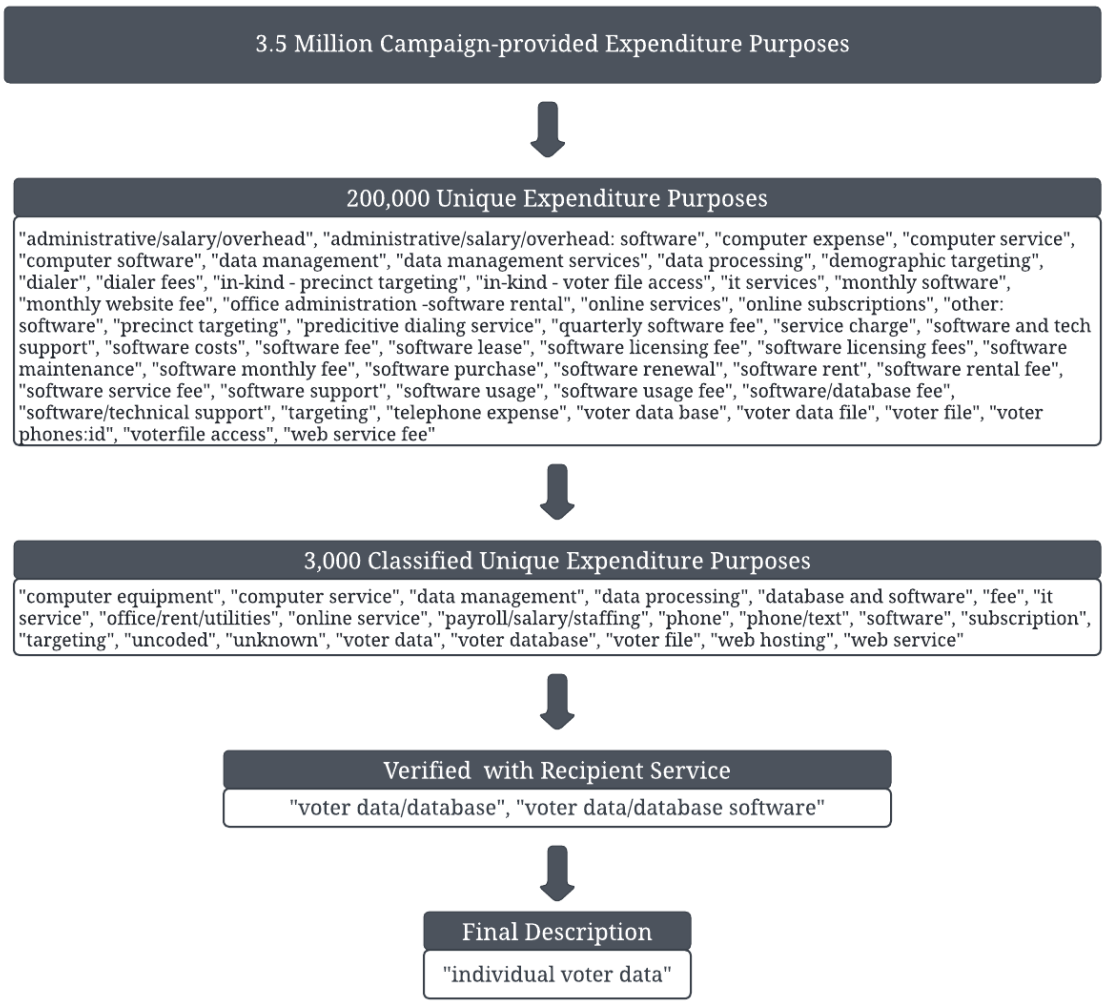
Given the challenges with other classification methods, I undertake a thorough review of a couple hundred thousand distinct purpose descriptions, which represent millions of expenditure records, and thousands of data vendors and polling firms across. I first reduced the 3.5 million campaign-provided expenditure purpose descriptions to unique entries provided by congressional campaigns in the "Purpose of Disbursement" field of FEC Form 3, seen in Figure 3.1. Figure 3.2 provides an example of classifying "individual voter data" to illustrate the entire refinement procedure detailed below and in Appendix B.

Specifically, within each of the seven cycles, I first filtered the congressional expenditure records to include all OpenSecrets-provided categories that could be potentially related to sources of electoral information. These categories included administration, campaign activities, fundraising, media, technology, strategy, and unclassified expenditures.[40] I then limited my review to only the unique words and phrases listed as the "Purpose of Disbursement," a total of approximately 200,000 unique descriptions across the seven cycles. I reviewed each of the unique purpose descriptions individually within each cycle and classified them into an appropriate category. For example, I classified descriptions such as "in-kind – voter file access," "voter data

[40] The remaining categories are mostly contributions to other political entities (e.g., other campaigns and parties). In total, my filtering reduced records from approximately 3.5 million to 1.4 million records across all cycles.

file," "voter file," and "voterfile access" as "voter file." In total, I generated 3,000 manually classified unique purpose descriptions.

**Figure 3.2 Illustration of Expenditure Review for "Individual Voter Data"**




As an additional quality control check, I verified all data-related expenditures by authenticating if the listed payee offered political polling or access to individual-level voter databases at the time of the reported purchase. I investigated each payee with more than three data-related expenditures in each cycle. I then confirmed if the organization provided the service through an examination of

their archived website on a date near the reported purchase date.[41] For example, I determined that Aristotle International, listed in Figure 3.1, provides campaigns with access to individual-level voter data. Their 2018 archived website[42] advertises that the data vendor provides campaigns with access to hundreds of millions of "highly customized voter, consumer, and state contributor lists anytime, anywhere."

In total, I recovered 50,000 verified expenditures related to either individual-level databases or polling and surveys that amounted to approximately $209 million, a figure that represents 3.5% of the total $6 billion in spending across seven cycles. Of the $209 million, House campaigns spent approximately $60 million on individual-level data and $150 million on polling between 2006 and 2018. The spending gap between polling and individual data vendors is partially attributable to the fact that polling firms field surveys in addition to providing analysis and consulting in their service costs (Martin and Peskowitz 2018). Individual voter data firms by contrast often provide their data directly to campaigns in forms such as voter lists or access to an interactive voter database restricted to their district, leaving the strategic decisions on how to utilize such data up to the campaigns (Hersh 2015). Polling is also inherently more costly even if it was possible to distinguish between the data collecting and consulting costs.[43]

## 3.2 Voter Data Vendors and Polling Firms

This section documents the companies and organizations that powered data-driven campaigning

[41] https://web.archive.org/

[42] https://web.archive.org/web/20180901091937/http://aristotle.com/data/datasolutions/

[43] I do not attempt to account for or distinguish between these data versus consulting costs for two reasons. First, it is impossible to make this distinction accurately given the campaign-provided expenditure description do not distinguish it. Second, the poll, its interpretation, and resulting advice from the consultant are integral to the aggregate picture campaigns gain of potential voters.

between 2006 and 2018. As chapter one reviewed, partisan consultants and party-aligned vendors are integral to contemporary campaigning. Whether selling candidates on the value of polls in the 1980s or creating the large voter databases that became commonplace in the 2000s, political professionals have convinced and literally sold congressional candidates on the value of new campaign technologies. Leveraging the millions of refined expenditures from the previous section, I document how both parties maintain separate individual-level data and polling markets that differ in both levels of competition and stability over time. These firm-level findings quantify the different partisan electoral information ecosystems and contextualize the different approaches to data analysis uncovered in my previous qualitative interview chapter.

Connecting the academic literature to my conversations with political professionals in chapter two, I documented how the two parties have different data cultures that differ in their data interpretation practices and the extent they share a common individual voter database. Prior investigations suggest that a network of Democratic actors trained in presidential campaigns spurred on the adoption of the VAN interface, while Republicans operate in a much more fragmented voter data environment characterized by limited data-sharing practices (Kreiss 2016). One study of consulting firms suggests the opposite dynamic within the polling industry in which Republican politics is dominated by a few firms while Democrats hire a wider variety of polling outlets (Grossmann 2009). Specifics aside, Democrats and Republicans operate within distinct partisan consulting networks that disseminate advancements in campaign tactics and technology internally (Kreiss and Jasinski 2016; Nyhan and Montgomery 2015). An outstanding empirical question is the extent to which congressional campaigns have coalesced around different partisan voter data vendors and polling firms and the degree to which these organizations dominate electoral information markets.

Figures 3.3 and 3.4 figures reveal how the individual-level data and polling marketplaces are comprised of distinct firms and characterized by different levels of within-party competition. The figures respectively display the top ten voter data vendors and polling firms across my period of investigation. The bars indicate total purchases in dollars made by congressional candidates arrayed from top to bottom by each vendor or firm's total revenue. While my extensive review of FEC records indicated that general consulting firms sometimes provide campaigns with both polling and individual-level data services, Figures 3.3 and 3.4 indicate that such full-service political data companies are rare in the marketplace. Each bar is also shaded based on the percentage of purchases made by Democratic candidates with red shades indicating more Republican purchases and blue shades indicating more Democratic purchases. Only a few firms service both Republicans and Democrats, and most work with congressional candidates from one party or the other.[44]

[44] My review of each of these vendors archived websites during my extensive review of FEC records reveals that many political pollsters and voter data vendors do not explicitly state their party affiliation. Some such as Aristotle International are even explicitly nonpartisan. I choose, however, to refer to firms in this section as "Democrat" or "Republican" because most if not all of their revenue comes from Democratic or Republican congressional candidates.

**Figure 3.3 Top Voter Data Vendors by Party**

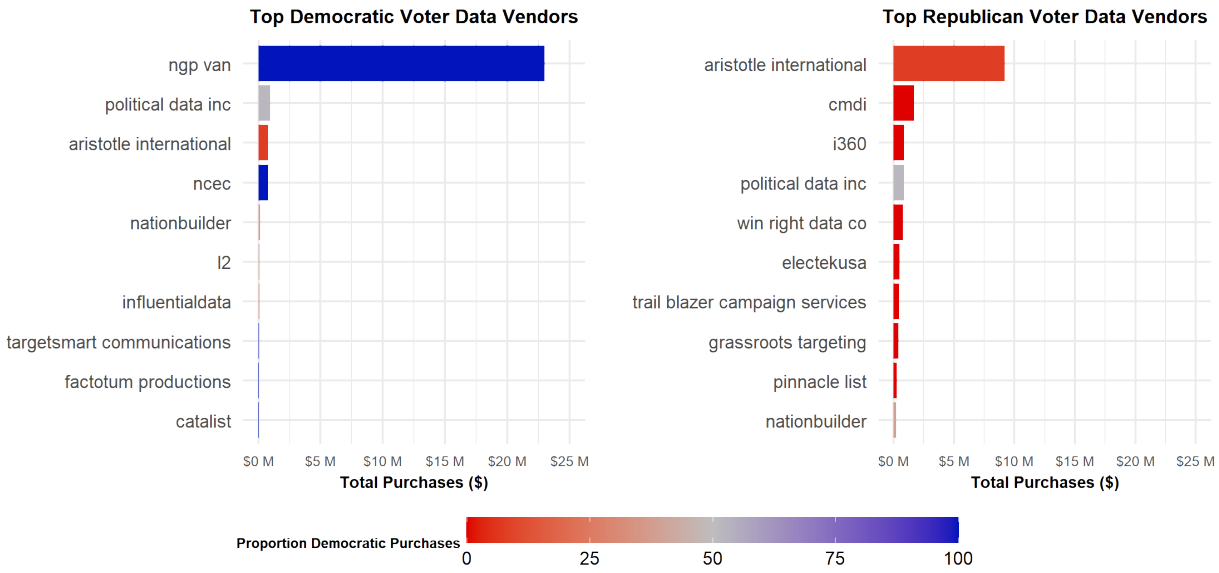
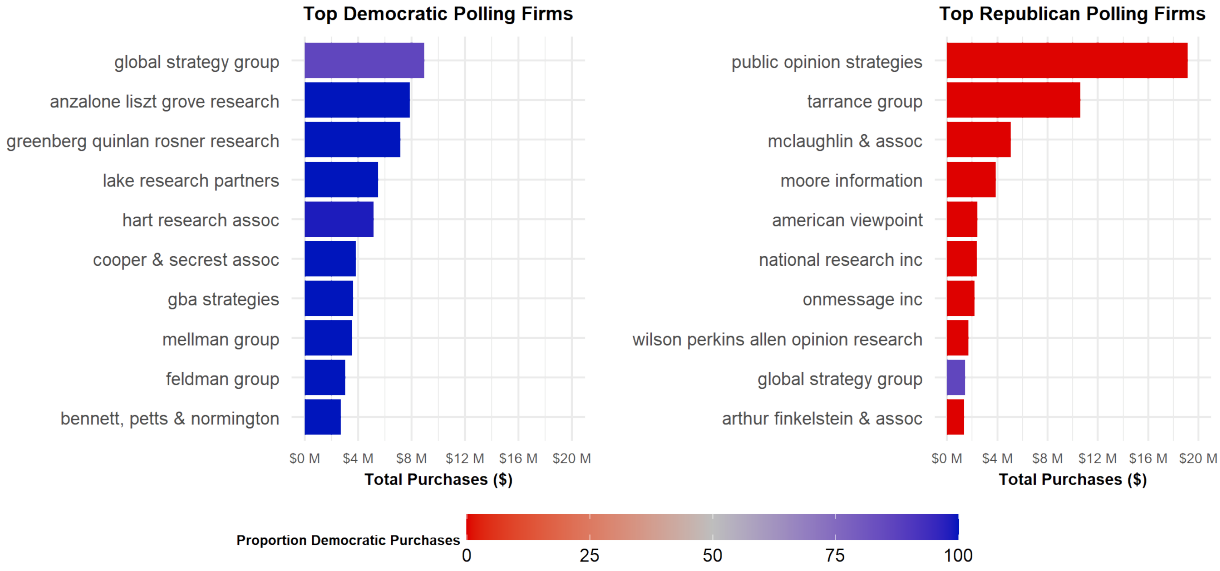



**Figure 3.4 Top Polling Firms by Party**




As seen in Figure 3.3, the individual-level data markets for Republicans and Democrats are dissimilar in terms of vendor prominence and consolidation. The vendor NGP VAN that manages the namesake VAN platform and VoteBuilder database dominates Democratic purchases of individual-level data approximating $23 million period-wide. By comparison, Republican

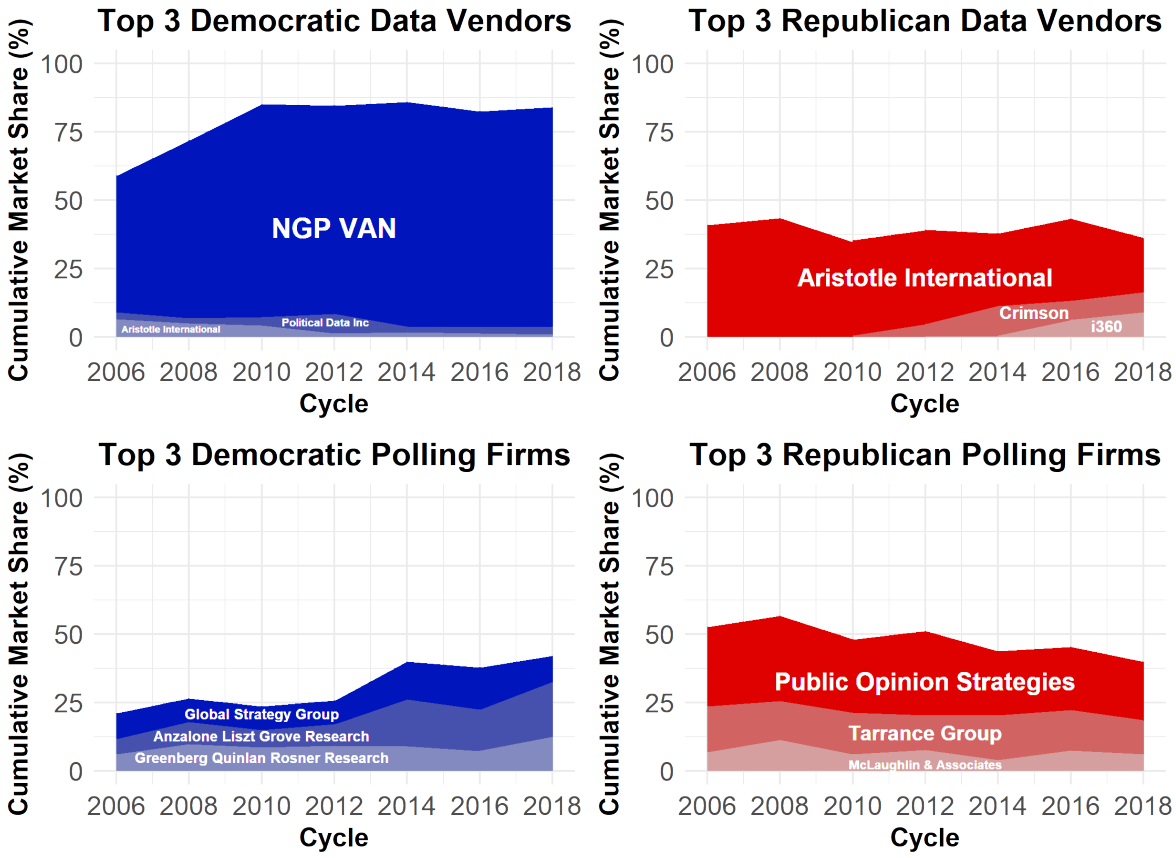
congressional candidates rely on a variety of different firms, including most prominently the nonpartisan Aristotle International, Republican data firm CMDI (also known as Crimson), and i360, a data firm founded by the prominent Republican donors Charles and David Koch. Both Democrats and Republicans make data purchases from Aristotle International and Political Data Inc (PDI), a nonpartisan firm that specializes in maintaining an up-to-date voter database exclusively in California, but only PDI has comparable sales revenue from both parties' congressional candidates.

Figure 3.4 depicts a polling marketplace with higher levels of competition. In relative dollar terms, more firms compete both in the Democratic and Republican polling sectors. Contrary to the partisan differences in the individual-level data market, the Republican rather than Democratic polling industry is more consolidated, although not to the level of market dominance held by NGP VAN. The top Republican polling firms Public Opinion Strategies, Tarrance Group, and Mclaughlin & Associates took in around $19, $11, and $5 million, respectively, over seven cycles. Contrast the top three Republican pollsters with Democratic outlets that each made between $7 million and $9 in revenue. As with individual-level data, most large polling firms work with one party or the other. The exception is Global Strategy Group which is nevertheless more likely to work with Democrats but also maintains Republican clients, as indicated by its bar's purple hue.

Partisan-level differences in market competition also persist over time. Figure 3.5 plots the market share of the top three firms measured as a firm's individual-level data or polling revenue divided by total market sales revenue in each cycle between 2006 and 2018. Since 2010, NGP VAN has maintained an 80 percent individual-level data market dominance among Democrats

with next-to-zero competition from Political Data Inc and Aristotle International. On the other hand, the Republican individual-level data market has increased in competition over time. The nonpartisan firm Aristotle International lost approximately 20% of its market share between 2006 and 2018 with the arrival of Crimson and i360 which acquired just under 10% of the market each. In contrast to the markets for voter data vendors, the Democratic and Republican polling markets are relatively stable across the entire period. The top three firms fielding surveys for Democratic congressional candidates maintained roughly equal competition across the period with all three slightly increasing their collective overall market share. Republican polling vendors meanwhile held a persistent market share with the top three firms' percentages of revenue in proportion to their rank.

**Figure 3.5 Market Share Over Time, 2006-18**



Across my period of analysis, I find that a small percentage of vendors and firms account for a large portion of purchases in both the individual-level data and polling partisan markets. Such empirical phenomena are commonly described by a "power law" relationship where one quantity varies as a power of another.[45] Power laws are extremely common in both the natural and social sciences. For instance, power laws map the distributions of earthquake magnitude, city population, word frequency, and income. Power law behavior is also typical in economic markets and generally follows a pattern where the market dominance of a firm (measured as revenue) is inversely proportional to its rank. This specific instance of power law known as "Zipf's law" predicts that free markets, absent external factors, result in the first-ranked firm having about twice as much revenue as the second-ranked firm, three times as much as the third-ranked firm, and so on (Gabaix 2009, 2016). When considering the Democratic voter data market, for instance, Zipf's law would expect that Political Data Inc should generate half the revenue and Aristotle International should generate one-third of NGP VAN.

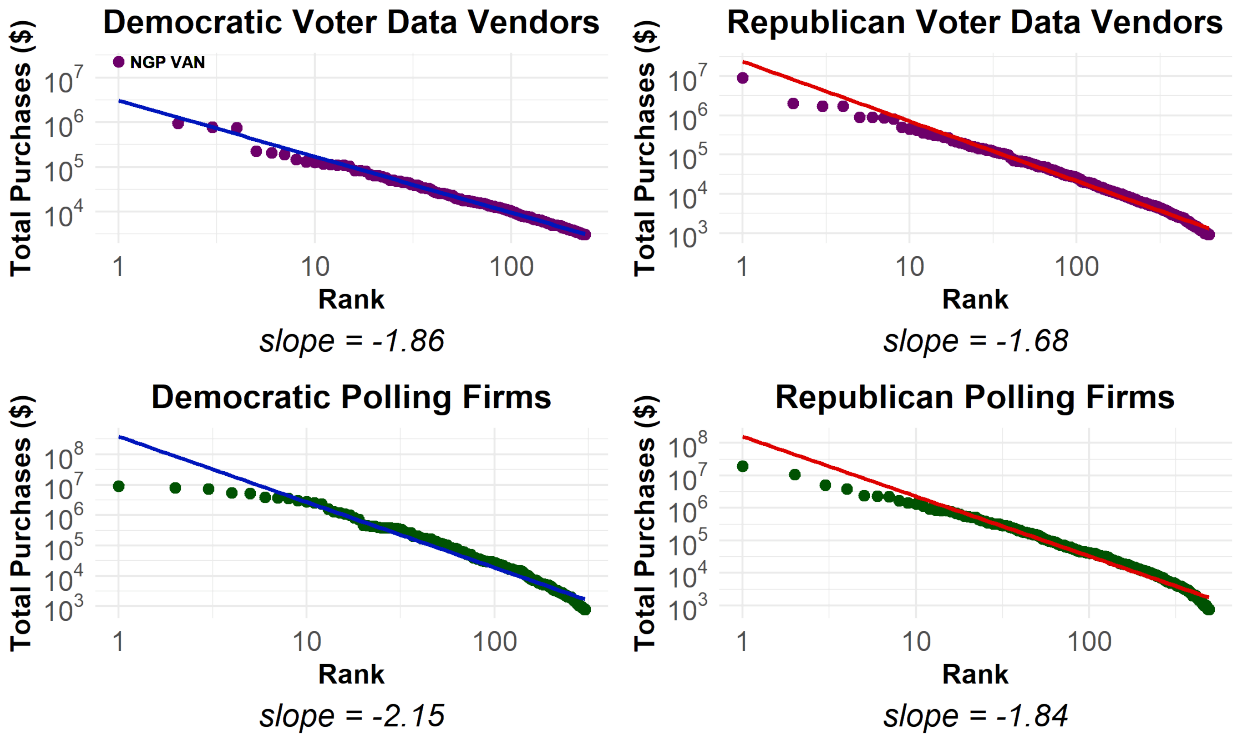
Figure 3.6 provides log-log scatter plots of total purchases in dollars and the rank of voter data vendors and polling firms split up into Democratic and Republican markets. Scatter plots with axes scaled to log base 10 are useful because dots following a straight line indicate power laws. Additionally, Figure 3.6 reports the slope of the power law fit to assess the closeness of the line to the Zipf's law slope of -1. All four markets approximate power-law behavior[46] except among the largest firms. NGP VAN (labeled in the figure) is an outlier among Democratic data vendors and has sales revenue outside of the predicted power law trend line, which fits the patterns for the

[45] Formally, a power law relationship takes the form $Y = kX^{\alpha}$ where $Y$ and $X$ are the variables of interest and $\alpha$ is the power law exponent (scaling parameter).

[46] A goodness-of-fit test via a bootstrapping procedure as recommended by Clauset, Shalizi, and Newman (2009) implemented in the R package poweRlaw returns p-values indicates no statistically significant differences between the four market distributions and synthetic power law distributions (Gillespie 2020).

remaining Democratic vendors. The largest Democratic polling vendors also buck the power law trend given their more equivalent purchasing totals. By contrast, both the Republican individual-level data and polling markets follow the power-law line closer among top firms.

**Figure 3.6 Power-Law Behavior of Voter Data and Polling Markets**




Combining insights from the above analyses, the Republican and Democratic electoral information markets primarily differ in the extent to which they mirror other free markets commonly found in the economics literature (Gabaix 2009, 2016). The Republican markets more closely reflect Zipf's power-law distribution defined by proportional random growth over time. Firms with greater capacity for growth are more likely to capture greater market share over time with the caveat that random events or "shocks" can lead to fluctuations. The Republican individual-level data market closely follows the proportional random growth model when the firms

Crimson and i360 entered the market and invested heavily in voter databases and interfaces.[47] Likewise, the Republican polling market approximates a stable power-law behavior over the entire period with firms having revenue proportional to their rank. Democratic markets by contrast do not follow the same trend. NGP VAN far outpaces its expected revenue proportional to its rank, and the largest Democratic polling firms control roughly equal proportions of the market.

In short, Republican and Democrats not only have different data cultures but different data markets. The NGP VAN is a near monopoly within the Democratic individual-level data market. While Democratic congressional campaigns sometimes purchase individual-level data from other vendors, nearly every Democratic candidate has access to the VAN. Republicans by contrast are likely to purchase access to individual-level data from a broader set of vendors. These marketplace differences help account for why the Republicans I interviewed in chapter two described their hands-on approach to combining different individual-level data sources for the needs of specific campaigns. Additionally, this firm-level analysis makes evident that the party-level effort by Democrats to invest in shared party-wide data infrastructure was extremely successful, and the position of its current VAN platform prevents market competition from other voter database platforms. This contributes to Democrats sharing a party-wide perspective and approach to data-driven campaigning, as detailed in chapter two.

## 3.3 The Cost of Voter Data and Polling

This section builds on the previous and details the cost of campaign-provided voter data and

[47] Specifically, Crimson first launched its campaign management software in 2008 and continued to invest and expand the system in subsequent cycles (https://www.cmdi.com/our-history). Similarly, i360 started in 2009 and sought to build a national voter database outside of the party infrastructure, which it launched and refined in the years following (https://www.i-360.com/company/our-story/).

polling expenditures. Opaque pricing and varied billing practices have made accurate cost estimates of individual-level data and polling difficult to find.[48] Closely examining approximately 50,000 verified data-related expenditures recovered from my extensive review of FEC records, I find that Democrats pay slightly more on average for polling but essentially the same to access individual-level data with some variation between 2006 and 2018. The causes of the partisan cost differences are difficult to determine but may derive from opaque pricing practices and disjointed marketplaces.

As emphasized throughout this dissertation, scholarship examining the data-related spending patterns of congressional candidates is limited, and this dearth equally applies to pricing estimates. A few studies to date have examined cost differences between Republican and Democratic consultants, but none have estimated the cost of electoral information. One study examining television advertising finds that Republican media firms charge more than their Democratic counterparts while at the same time providing worse media-buying services (Martin and Peskowitz 2018). Other research has tracked the cost of polling among congressional candidates without attention to potential party-level cost differences (Limbocker and You 2020). Exact cost estimates aside, my review of scholarship in chapter one revealed partisan asymmetry in campaign technology innovation and adoption. Paired with the fact that firms operate in separate markets, the costs of electoral information likely differ at a party level even if the direction of the difference is unclear based on prior studies.

Creating accurate estimates of polling and individual-level data costs requires accounting for common campaign-level billing and reporting practices. First, the cost to access an individual-

[48] The opaque market applies to both researchers and to political campaigns themselves. There is not publicly available pricing information for polling contracts or for when campaigns purchases access to large individual-level databases.

level database should be much lower than the cost of a poll. Campaigns commonly access individual-level databases through a software-as-a-service (SaaS) model and pay monthly subscription fees that vary depending on the level of access, including the number of individual records (rows) and fields of individual-level information (columns). This subscription-based payment plan is how the two top firms NGP VAN and Aristotle International bill for access to their voter data interfaces.[49] Additionally, campaigns have options to purchase a la carte data services either from the same vendor or other specialized voter data and modeling companies. Although less common, campaigns may also purchase lists of voters based on the value of the entire list or through a pay-for-voter model. By contrast, campaigns purchase polls by paying a flat fee for a polling firm to field and analyze a survey. Alternatively, campaigns can maintain a contract with a polling firm for an entire campaign paid either in one lump sum or in large increments. Taken together, the variation in pricing and billing makes the recovery of reliable estimates from campaign-provided records challenging.

To account for different payment practices, I add weights to the cost estimates based on the campaign-level frequency of reported expenditures in each cycle. Substantively, individual-level data expenditures from campaigns that report fewer individual data expenditures are weighted *less*, while polling expenditures from campaigns with fewer polling-related expenditures are weighted *more*. This weighting accounts for the fact that accurate polling estimates are more likely to come in bulk payments, while individual-level data estimates are more likely to be reported piecemeal. Specifically, individual-level data expenditures are weighted by the frequency of reported expenditures related to individual data. For example, an individual data expenditure from a

[49] Most data vendors do not advertise their costs. This is especially true for NGP VAN's services and software with its various pricing plans and contracts with the national and state parties. The leading voter data provider for Republicans Aristotle International similarly does not publicly advertise its plans or pricing.

campaign that reports 10 payments related to individual-level data is 10 times the weight of a campaign that reports a single purchase. Conversely, polling expenditures are weighted by the inverse of the expenditure frequency (one divided by the total number of polling-related expenditures in a cycle), so that, for example, an expenditure from a campaign with one polling-related expenditure is weighted twice as much as one from a campaign that reported two poll-related payments in the same cycle.

As expected, congressional campaigns pay much less for individual-level data than they do polling. Table 3.2 reports weighted and unweighted descriptive statistics for data-related expenditures. Across all cycles, congressional campaigns spent approximately $61 million on individual-level data and $150 million on polls and surveys. Individual data payments are spread over three times more reported expenditures than polling, documenting their different pricing structures. The cost difference is also true at the expenditure level as seen in the estimates of the median and mean. The median cost of a poll is approximately 10 times that of individual-level data. Mean estimates are inflated on average because of the high variation, as indicated by the large standard deviations. In the weighted estimates, the standard deviation of individual-level data is $4,548 and polling is $10,584. While both numbers suggest large outliers, individual-level data prices are more widely distributed. Given the high level of dispersion, the median is a more appropriate measure of central tendency for both distributions. The weighted median approximates $750 for individual-level data and $11,000 for polling.

**Table 3.2 Individual Data and Polling Expenditures**

| Source | Weights | N | Sum | Median | Mean | SD |
|---|---|---|---|---|---|---|
| Individual Data | No | 37,328 | $61 M | $850 | $1,631 | $3,660 |
| Polling | No | 12,318 | $150 M | $10,000 | $12,215 | $10,755 |
| Individual Data | Yes | 37,328 | $61 M | $750 | $1,734 | $4,548 |
| Polling | Yes | 12,318 | $150 M | $11,245 | $12,956 | $10,584 |

Figure 3.7 illustrates the patterns from Table 3.2. The figure overlays the distributions of individual-level data and polling expenditures across all seven cycles split up by congressional candidates' partisanship. The horizontal axes are scaled to log base 10 to better capture the high-cost outliers in the untransformed scale. The polling distributions are shifted further right along the horizontal axis compared to the individual-level distributions because of higher costs associated with fielding a survey, as also indicated by the dashed line representing the overall weighted median. Individual data distributions are also more dispersed than the polling distributions indicating a wider range of prices. Both parties' individual-level data expenditure distributions are also bimodal. Although difficult to determine, this may be the result of different tiered voter database subscriptions with lower-level peaks representing more limited access with fewer columns of data on voters. The polling distributions by contrast are unimodal and less dispersed.

**Figure 3.7 Estimated Costs of Individual Data and Polling**

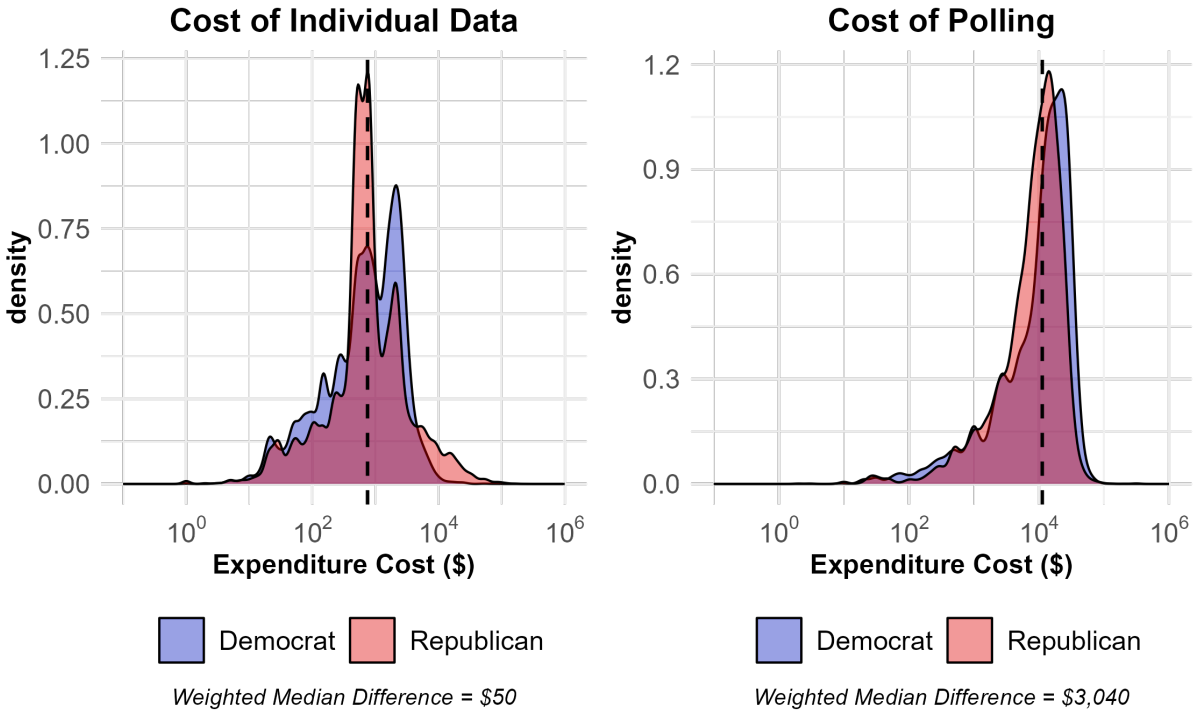
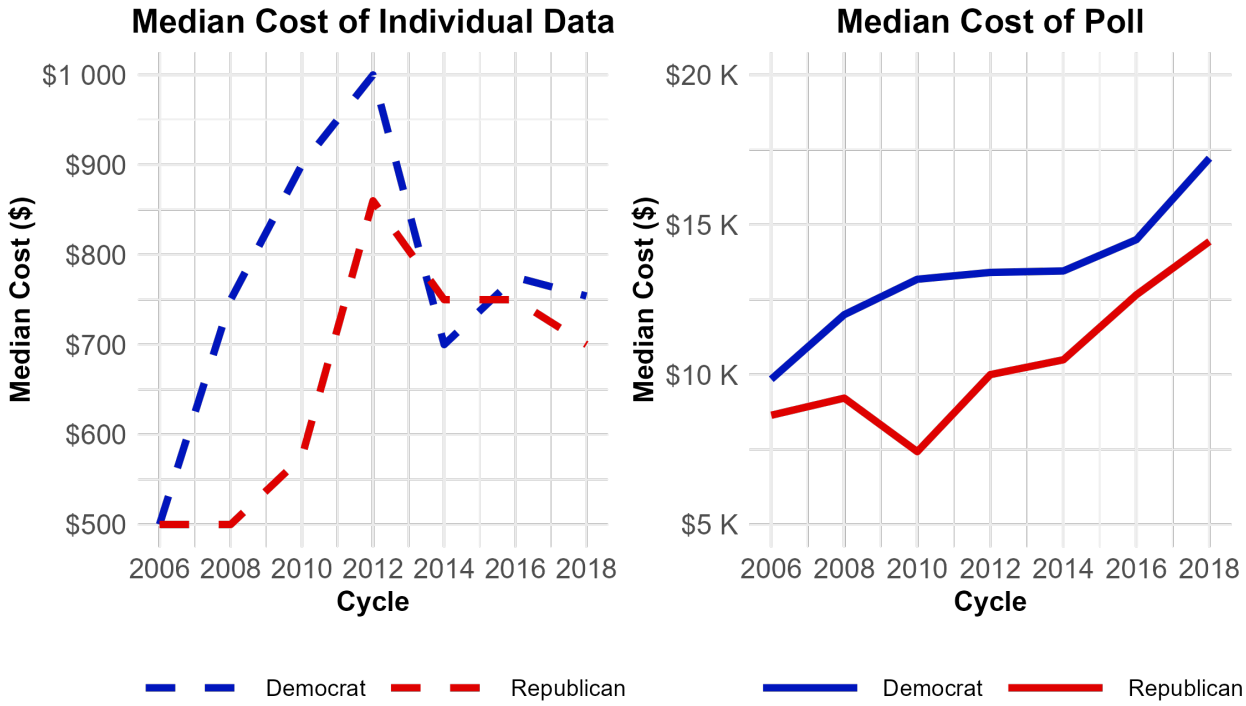



Figure 3.7 also reveals some relatively small party-level differences in individual data and polling prices. Visually, Democrats pay slightly higher prices on average for both individual data and polling, as indicated by the blue peaks shifted right along the horizontal axes. Two tests of weighted median difference using the nonparametric Wilcoxon rank-sum test reveal that these differences are statistically significant with p-values that include fourteen leading zeroes ($10^{-14}$). Although statistically different, a difference of approximately $50 and $3,000 may not make much substantive difference either at an aggregate level or a campaign level. For many campaigns, subscription payments for individual-level data are treated as operating expenditures much like wages or rents. By contrast, polling is akin to a capital investment in research and development where the higher price may not make much of a difference after a campaign has already decided to pay for a survey. All told, however, price differences between partisan marketplaces may contribute to differences in campaign-level allocation of resources to one source versus the other,

as explored in the next section. Appendix C provides a corresponding unweighted version of Figure 3.7.

To investigate overtime variation, Figure 3.8 plots the weighted median cost of individual-level data and polling expenditures in each cycle between 2006 and 2018. Looking at the trends among both parties, the figure reveals an increase in the cost of both data sources. While increasing over the entire period, individual-level data prices fluctuated reaching a peak in 2012 and then coming back down to between $700 and $800 on average in 2018. While difficult to ascertain, the lower cost may have resulted from the individual voter data industry switching to a SaaS pricing model, although the shift may well represent actual price decreases. Certainly, both Democrats and Republicans are paying less on average for individual-level voter data over time, which sees consistent price deflation during the period rather than consistent inflation. The price of polling by contrast inflates consistently over the period from around $10,000 to closer to $15,000. As mentioned in chapter one, polling response rates have fallen significantly during my period of investigation drastically increasing the price of phone-based surveying methods – the gold standard among political pollsters. The figure makes clear that these price increases were passed on to congressional candidates in both parties. Appendix Figure C2 provides individual-level data and polling cost estimates in real 2018 dollars, adjusted for inflation based on mean Consumer Price Index estimates from the Bureau of Labor Statistics. Inflation-adjusted calculations reveal similar trends. The real cost of individual-level data decreased over time, while the real cost of a poll increased, outpacing standard measures of consumer product inflation.

**Figure 3.8 Median Cost of Individual Data and Polling Over Time, 2006-18**




Democrats and Republicans have become closer to one another in how much they pay for individual-level data, while the price Democratic candidates pay for polls remains consistently higher. Democrats paid slightly more to access voter databases in 2008, 2010, and 2012, but the median price levels off and became essentially equal for both parties in subsequent cycles. The fact that the individual-level data trend lines for Democrats and Republicans are similar suggests that both markets respond to similar market forces. A parallel overtime analysis of the median cost of data-related expenditures purchased from the top two individual data vendors for each party is provided in appendix Figure C3. It reveals that the nonpartisan firm Aristotle International, which is the largest vendor among Republican candidates, consistently prices its voter data products at a lower price point than the NGP VAN. The partisan polling industries similarly have tracked trend lines, but Democrats pay approximately $1,000 to $5,000 more per poll on average than their

Republican counters in each cycle.

These uncovered slight differences in party-level prices between 2006 and 2018 are difficult to explicate even with my refinement of data-related expenditure records. Higher Democratic reported prices for individual-level data in the earlier cycles may be the result of the near monopoly status of NGP VAN compared to higher levels of competition among Republican data vendors. Similarly, why Democrats pay more for polls on average is not immediately clear, but it may be the result of market structure with three firms each having roughly equal market share compared to the Republican polling industry that more closely reflects a free market. Ultimately, explanations for the reported difference in pricing are difficult to ascertain and may result from either pricing structures that different partisan firms have or the result of unmeasured market forces outside the scope of this investigation.

## 3.4 Campaign-level Spending on Voter Data and Polling

While understanding the marketplaces and cost of voter data and polling are important to gain a better picture of the data-driven campaign environment, this final section homes in on understanding campaign-level spending patterns. After finding an appropriate measure to understand campaign-level spending patterns, I detail how congressional campaigns have increasingly taken up individual-level data to help them implement their voter outreach strategies. Yet I also find that party-level differences persist. Democratic congressional campaigns made individual-level data a priority long before their Republican counterparts who continue to prefer polls.

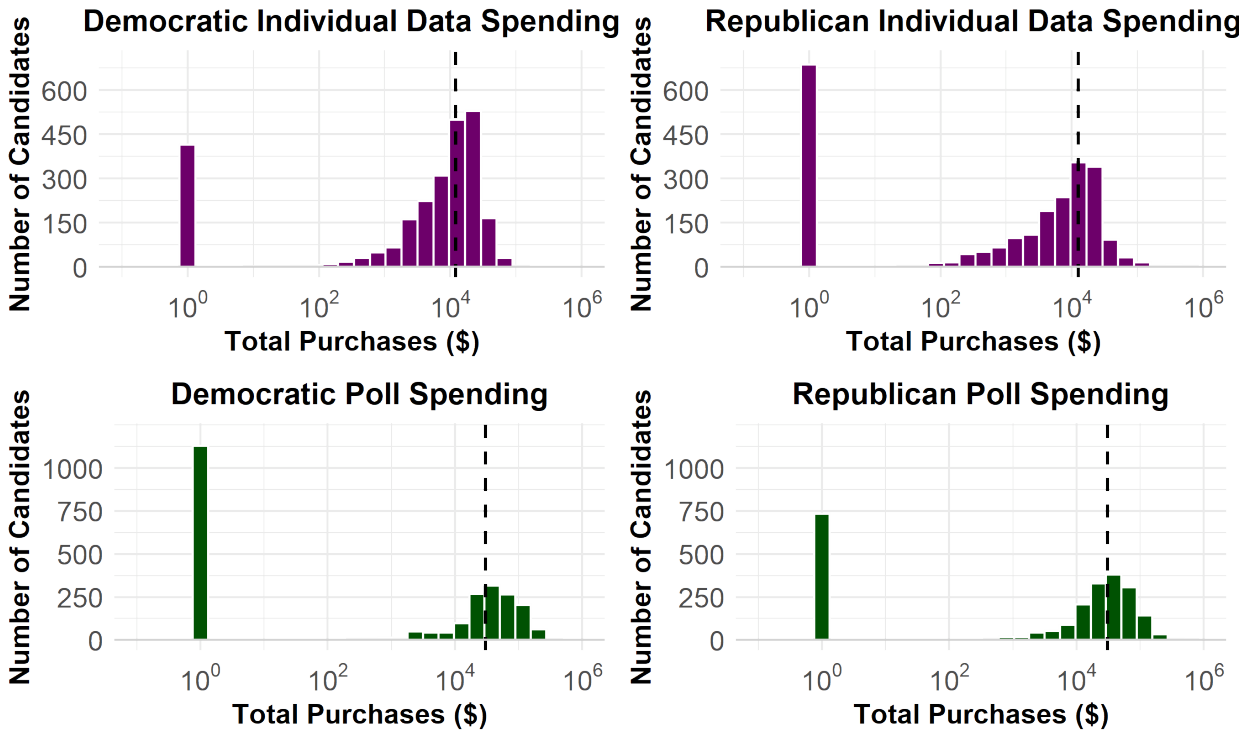
Direct comparison of campaign spending patterns requires a comparable unit of analysis not only across campaigns but also across cycles. As previewed in the first chapter, I rely on an

analytical strategy akin to the revealed preference approach more common in the economics literature. Given the difficulty of computing specific dollar amounts, as mentioned in the previous section's weighting discussion, I choose to focus my inquiry on understanding the relative amount of money the congressional campaigns spend on one data source or the other rather than explaining overall levels of spending. Specifically, I measure campaign-level data preference by calculating the *proportion* of spending on either individual-level data or polling divided by the total amount a campaign spent on either source across the entire election cycle.

Figures 3.9 and 3.10 illustrate the differences between analyzing spending in dollars versus relative spending measured as percentages. Both figures present the distributions of campaign-level spending on both sources divided by partisanship. The resulting four plots in Figure 3.9 document similar findings to my cost analysis but also crucially reveal that not every congressional campaign chooses to purchase data. After scaling to log base 10, spending in dollars approximates a normal distribution with a lower bound skew across all four figures with the evident exception that many campaigns often choose to spend nothing on individual-level data or polling. This choice to invest in data sources is an important revelation not only because of the impact it has on actions that campaigns take but also because it suggests many potential factors are at play when campaigns choose to make investments in electoral information sources. A closer examination of the vertical axes depicts how many campaigns will forgo polling altogether, while a larger number will purchase individual-level data. The different distribution heights are almost certainly a direct result of the differences in prices reported in the previous section. Polls are expensive. Many campaigns lack either the cash on hand to purchase a survey or their understanding of strategic considerations like levels of support and vulnerability within their district leads to decision-making sans surveys. Individual-level data, on the other hand, have a lower cost of entry and, as inferred by spending

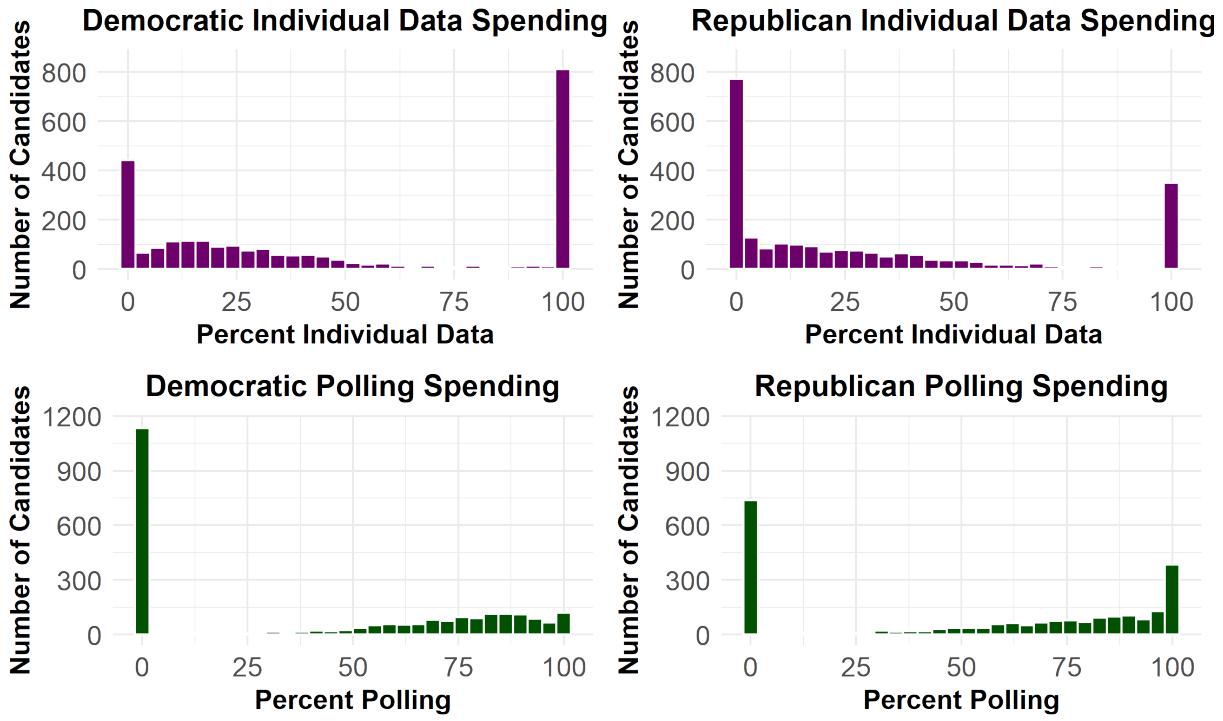
patterns, are more broadly accessed and utilized by many more campaigns.

**Figure 3.9 Campaign-level Data Spending in Dollars**




To better examine campaigns with and without data spending, Figure 3.10 plots the same figures but with spending measured as the percentage of spending on either individual-level data or polling out of the total amount spent on both sources. In addition to documenting data-less campaigns, the figure also highlights the extent to which some campaigns choose to only purchase one source or the other. For many campaigns, the choice is between investing in individual data or not. Other campaigns invest in a mix of polls and individual data, as indicated by all campaigns with percentages between 0 and 100 in either distribution. Campaigns are less likely to invest all their money into polling, though some do. The top and bottom of the mean spending distributions again emphasize a party-level data preference. Democrats are more likely to rely on only individual-level information, while Republicans are more likely to devote a larger portion to polls.

**Figure 3.10 Campaign-level Data Spending in Percentages**




Probing campaign-level spending over time further reinforces the importance of considering proportions of spending rather than pure dollar amounts. Figures 3.11 and 3.12 provide the mean spending in dollars and percentages, respectively, between 2006 and 2018. The figures plot individual-level data spending with dashed lines and poll spending with solid lines. Color indicates party. Only examining spending in dollars depicts an evolution that is equally matched with congressional candidates in both parties purchasing the same amount of each source on average. According to Figure 3.10, the typical Republican candidate and the typical Democratic candidate differ by no more than $5,000 in individual-level data spending in a cycle, and Republicans spend more on individual-level data starting in 2012 and afterward. Meanwhile, the polling trend lines indicate that the two parties vacillate in spending with Republicans again spending more starting in 2012.

**Figure 3.11 Campaign-level Data Spending in Mean Dollars Over Time, 2006-18**

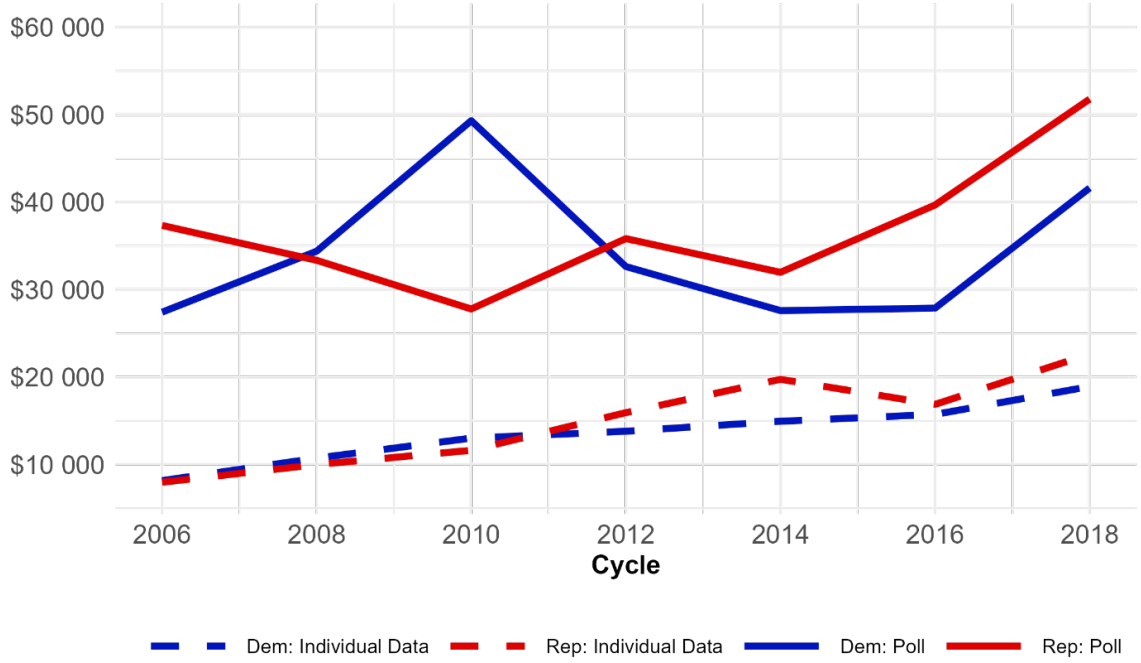



**Figure 3.12 Campaign-level Data Spending in Mean Percentages Over Time, 2006-18**

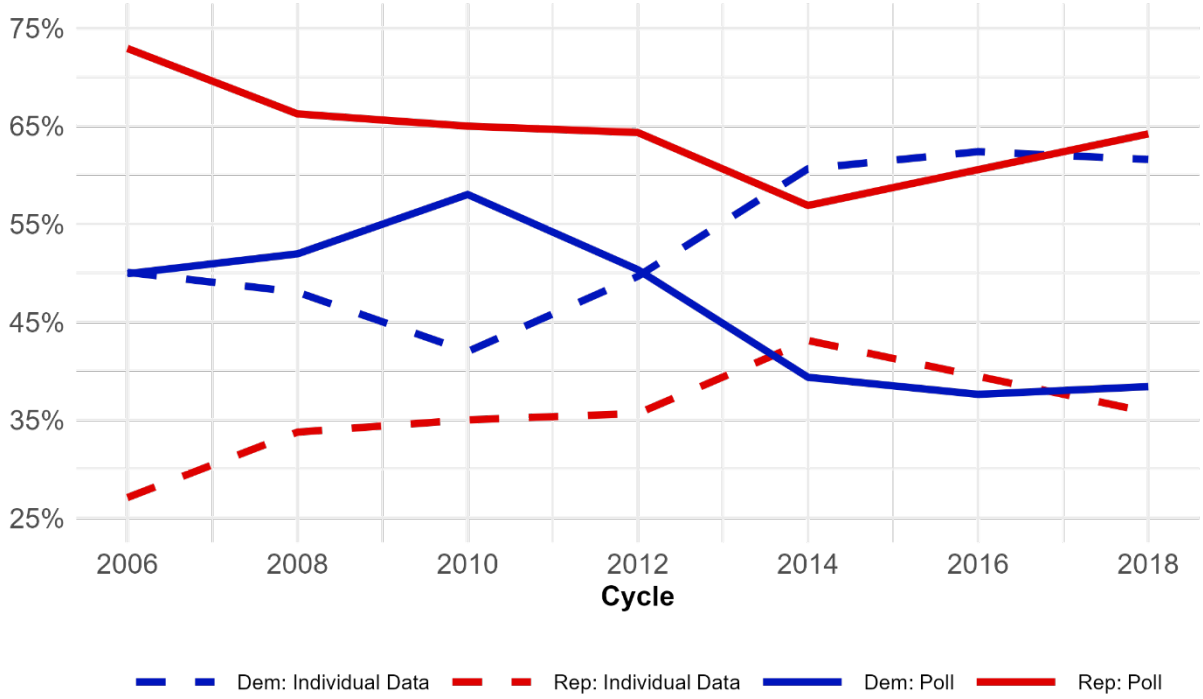


Figure 3.12 depicting mean percentages over time reveals a different narrative than uncovered by examining the mean dollar amount. At a campaign data allocation level, Democrats were spending larger portions of their data budgets on individual-level data since the beginning of the period in 2006. Specifically, Democrats vacillated between a data spending equilibrium and prioritizing polling as comprising the majority of data expenditures on average. A shift occurred in 2014. Democrats began to apportion more of their data budgets to acquire individual records. By comparison, the average percentage spent on either source never converges for Republicans with only a moderate increase in the percent spent on individual-level voter data. This disparity is important because, unlike seen with total spending in dollars, the typical Democrat and Republican campaigns diverge in how they allocate their data budgets. Thus, these different party preferences are not simply the result of increased spending but, instead, Democrats prioritizing information coming from individual-level data over that of polling-provided information beginning in 2014.

The 2014 Democratic transition toward more individual-level data among congressional campaigns is an important addendum to the story about the diffusion of voter data most often told from the perspective of presidential campaigns. Not being able to afford the heavy investment into the acquisition and processing of these individual-level sources themselves, most House campaigns, particularly Democrats, waited until the arrival of third-party data vendors to provide cheaper access to voter lists to help inform their outreach strategy. Republicans, in contrast, devote less of their spending to records based on the voter file and continue to prefer polling with only a slight decrease in the proportion of their data budgets they allocate toward polling across the entire period.

## 3.5 Summarizing Over Time Trends

This chapter finds persistent partisan differences in the marketplaces and costs of electoral information that also extend to the spending patterns of congressional campaigns over a period of marked technological change and innovation between 2006 and 2018. Existing research was limited in methodological scope and unable to provide a complete picture of electoral information sources across different campaigns and time. Distorting the image of data-driven campaigning was the literature's focus on presidential campaigns. Unlike Oval Office contests with the financial resources and expertise to invest in the creation of large-scale voter databases and experiment with novel analytical techniques, congressional campaigns (and by extension those below them on the ballot) must instead rely on a network of partisan polling firms for their surveys and third-party data vendors to process and package the voter file for them.

Examining the companies that comprise electoral information markets and their data-related products quantifies past assumptions that up-ballot differences transfer down the ballot. For the Democratic party, the Obama campaigns of 2008 and 2012 served as turning points that transformed their data infrastructure and reshaped strategic party culture. Obama's widely perceived success led to the creation of new firms and approaches to voter outreach incubated in his campaign and institutionalized by his former staffers forming new data firms. The scholarly and historical records reviewed in chapter one suggested that the Democratic party became unified around a dominant targeting strategy made possible by the shared VAN interface. No longer having to re-cobble different individual-level data sources together at the start of each campaign, I learned from my interviews that Democratic Party operatives came instead to trust in the predictive scores modeled by outside, third-party analysts that reduce voters to a few numbers.

This chapter's analyses make clear the empirical reality of the faith Democrats put into the VAN and its voter records. Access to the VAN platform is not only the default choice for Democratic congressional campaigns but, in most regards, the only choice. Democrats may pay slightly more on average to access voter databases but likely receive a better return on their investment given data-related economies of scale and its universal use among operatives. Republicans, by contrast, have not coalesced around one provider of individual-level data and predictive scores at the levels of both the national party and congressional campaigns. This fragmented individual data ecosystem may lead to inefficiency because of the often haphazard and ad hoc analysis of both individual-level data and polls undertaken by Republican campaign practitioners. Yet Republican campaign consultants and practitioners exercise much greater control of the sources of electoral information they acquire and how they are combined to inform their outreach strategy.

Integrating the results of my interviews with the present descriptive analysis of tens of thousands of data-related spending records verified from millions of congressional expenditure records makes clear that Republicans and Democrats operate in vastly different data ecosystems. Different partisan marketplaces and party-level orientations toward electoral information sources have contributed to the campaign-level divergence in data preferences since HAVA's statewide registration roll requirement accelerated the scale of large voter databases. Democratic congressional campaigns have come to rely on the VAN and prefer spending their limited budgets on individual-level data over polling. Candidates in the Grand Old Party, on the other hand, continue to prefer the inferences derived from traditional surveys to help craft their outreach strategies. This party-level difference in data preferences is an important aspect to emphasize because it not only affects what information campaigns obtain, but as I learned in my interviews,

how campaigns are run.

In the next chapter, I continue to refine this image of data-driven campaigning by accounting for the strategic considerations campaigns face when developing a strategy to communicate with voters. Accounting for the on-the-ground realities impacting strategic decision-making serves as yet another test to determine if the congressional candidates exist in different partisan data ecosystems and cultures.

# 4 Zooming in on Campaign-level Data Spending

In the previous chapter, I discovered how data-driven campaigning for Congress evolved separately for Democratic and Republican candidates in terms of their marketplaces for electoral information, the prices they pay for data, and ultimately how they choose to allocate their limited budgets. Combining those findings with the interview insights from chapter two, my inferences so far suggest that a consistent differentiating factor in how campaigns interpret and utilize different sources of electoral information is dependent on the party they belong to. This chapter continues to leverage verified data-related expenditure records to further reveal the data preferences of congressional campaigns across different strategic contexts. Understanding these revealed preferences is central to understanding how campaigns interact with the current voter information environment and clarify what was previously an incomplete picture of voter data's down-ballot diffusion.

Most existing studies examining data-related technological innovation and diffusion have focused almost exclusively on presidential campaigns. Following every presidential election since 2008, scholars have written books and articles to describe how sophisticated campaigns at the top of the ticket gathered more information on voters to tailor and target their appeals at crucial segments of the American electorate (e.g., Bimber 2014; Endres 2020; Epstein 2018; Hersh 2015; Hillygus and Shields 2009; Kreiss 2012, 2016; McKenna and Han 2014; Trish 2018). The overwhelming attention paid to these atypical campaigns occurring every four years, costing billions of dollars, and involving millions of voters contacted by thousands of staff and volunteers obscures the reality of data-driven campaigning in the thousands of other contests happening throughout the country. Campaigns for Congress occurring every two years are not merely less

sophisticated than presidential contenders but face a different set of circumstances, strategic considerations, and, importantly, resource constraints that force them to make tough decisions about whether to invest in electoral information and how to allocate their limited budgets if they make the data investment. My extensive review of existing literature in chapter one discovered only a few qualitative studies exploring the many factors that go into the data acquisition calculus of lower-level candidates (Baldwin-Philippi 2018; Nielsen 2012), despite recognition of down-and-top ballot differences in data-driven campaigning and calls to investigate them (Baldwin-Philippi 2019; Kefford et al. 2022).

This chapter fills in the empirical gap in current scholarship by systematically examining the data purchasing patterns of thousands of congressional campaigns. I first detail which strategic factors candidates running for the House may consider when investing in electoral information. The following section details my modeling choices. Modeling budgeted spending devoted to either individual-level data or polling requires special attention to statistical assumptions and model parameters. Next, I detail my results. I find that a candidate's party affiliation remains a consistent differentiator in campaign-level data spending patterns across a wide variety of models and specifications. The robust impact of party is attenuated at higher levels of spending where candidates of both parties are more likely to exhibit similar preferences for individual-level data and polling. Notwithstanding the effect of party, this chapter reveals how congressional data spending patterns also respond to strategic considerations of overall campaign spending, competitiveness, and outside spending. In short, party matters but so do many other perennial aspects involved in strategic campaign decision-making.

## 4.1 Potential Factors Explaining Voter Data Investment

While the previous chapter's findings suggest party-level differences in data purchasing patterns, many more considerations go into a campaign's decision to invest in different sources of electoral information. Table 4.1 describes the six categories of variables that I consider in my multivariate analysis as well as their source. They include campaign-level financial resources, outside spending by other groups, strategic considerations of competitiveness, candidate characteristics of incumbency and party, district-level information, and finally two variables that measure the quality of each state's voter file. Because of limited scholarship, I do not develop testable hypotheses about the impact of each variable apart from party. Instead, the purpose of the following analysis is to uncover the correlates of data spending patterns to reveal differences in preferences. I provide brief justifications for including each variable informed in part by existing literature but also by reasonable assumptions gained from my exploratory analyses of the same records and my conversations with campaign professionals.

The amount of money a campaign can spend is an obvious consideration central to understanding data purchasing patterns. In my interviews with campaign professionals, they emphasized how financial resources are the number one consideration not only when coming up with a voter outreach strategy but also in determining the kinds of data campaigns will purchase. Campaigns do not only decide how much they want to spend on polls versus individual-level data but also whether to purchase them in the first place. Even well-resourced congressional campaigns lack the staff, time, and money to invest in sophisticated data-driven practices seen in presidential campaigns. Most do not, for instance, build new voter databases from scratch, generate their own statistical models, or develop customized voter tracking and targeting tools (Baldwin-Philippi

2018). Instead, as detailed in chapter one, campaigns pay outside data vendors to collect data on voters and build targeting applications and polling firms to field and analyze surveys.

**Table 4.1 Explanatory Variables**

| Variable | Description | Source |
|---|---|---|
| | *Financial Resources* | |
| Log Spending | Log total spending calculated as the sum of all expenditures | OpenSecrets |
| | *Outside Spending* | |
| Log Outside Spending For | Total log independent expenditure spending expressively advocating for the election of the candidate | FEC |
| Log Outside Spending Against | Total log independent expenditure spending expressively advocating for the defeat of the candidate | FEC |
| | *Competition* | |
| General Vote Difference | The absolute vote percent difference between the top two candidates in the general election | FEC |
| Competitive Primary | Competitive primary election defined as candidate receiving less than 90 percent of the vote in the primary election | FEC |
| | *Candidate Characteristics* | |
| Incumbent | Indicator of candidate incumbency | FEC |
| Democrat | Indicator of candidate party | FEC |
| | *District-level Factors* | |
| % CVAP Turnout | Percent of district citizen voting age population that turned out in general election | FEC, Census ACS 1 year |
| % White | Percent of district white | Census ACS 1 year |
| % College + | Percent of district with bachelor's or higher | Census ACS 1 year |
| % Urban | Percent of district classified as urban | Census ACS 1 year |
| Median Income | Median income of district | Census ACS 1 year |
| Median Age | Median age of district | Census ACS 1 year |
| | *Voter File Indicators* | |
| Party Registration | State voter file provides partisan information in any form (primary voting or partisan registration) | State Laws |
| Race Information | State voter file contains information on race | State Laws |

Additionally, as detailed in chapter three, individual-level data and polling have different costs

associated with them. While campaigns can access individual-level voter records with an affordable subscription as low as a few hundred dollars a month, polling comes with a much higher price tag typically upwards of $15,000 that will prevent cash-strapped campaigns from purchasing them altogether. To account for spending, I calculate the log of total spending as indicated in the OpenSecrets records by summing all expenditures records per campaign per cycle. Calculating the log of total spending accounts for high-spending outlier campaigns.

Outside spending may also affect campaign data spending patterns. The FEC requires organizations that expressively advocate for the election or defeat of a clearly identified candidates to disclose their independent expenditures.[50] Individuals, groups, corporations, labor unions, political committees, and party committees make independent expenditures. Especially since the Supreme Court's ruling in *Citizens United v. FEC* in 2010 and the rise of independent expenditures-only committees (commonly referred to as "Super PACs"), the amount these outside groups spend to influence elections has ballooned. Super PACs can raise and spend unlimited sums of money in support or opposition of candidates, which have comprised a larger portion of overall election spending every subsequent midterm and presidential cycle. Although these outside organizations are not permitted to coordinate[51] formally with campaigns and parties, they often signal their spending strategies to one another by disclosing publicly available information about purchase intentions, especially related to mass media purchases that comprise the large majority of their spending (Herrnson et al. 2013). Super PACs can also potentially publish the results of research they have conducted, such as polls, as public information. To capture the potential

[50] https://www.fec.gov/help-candidates-and-committees/making-independent-expenditures/

[51] The FEC defines coordination as "Coordinated means made in cooperation, consultation or concert with, or at the request or suggestion of, a candidate, a candidate's authorized committee or their agents, or a political party committee or its agents." https://www.fec.gov/help-candidates-and-committees/making-independent-expenditures/

influence of outside spending, I calculate the total log spending of independent expenditures made for and against each candidate. The FEC provides these expenditures for every cycle and contest in my sample. Between 2006 and 2018, independent expenditures, including those made by Super PACs and party committees, made for and against the election of House candidates numbered 83,394 and totaled approximately $1.6 trillion.[52]

Beyond spending, campaigns likely react to levels of electoral competition in the district. For instance, candidates facing a strong challenger may purchase polls more frequently to get updates on their current standing compared to their opponents. As I learned in my interviews with campaign professionals, competitive races also lead to more fine-grained segmentation of persuadable voters. This may lead to an increase in the cost of individual-level data if campaigns are either purchasing specific lists of voters or paying for data companies to develop persuasion models specific to their race. I was told these race-specific persuasion models are even more common in competitive primary elections where models require much more detailed individual-level information to predict and differentiate support for candidates of the same party. As a result, I measure both the competitiveness of the primary and general elections because my data-spending dependent variables include expenditures from both. Specifically, general election competitiveness is defined as a vote margin of less than five percent between the top two vote-getters. Primary election competitiveness is an indicator variable measuring any time a candidate receives less than 90 percent of the vote. Only a modest amount of competition is required for

[52] The FEC also requires the disclosure of "electioneering communications" for organizations that name a "clearly identified" candidate in communications made by individuals or "other persons" (corporations and labor unions) 30 days before a primary or 60 days before a general election a reaching 50,000 or more people. According to OpenSecrets, electioneering communications that do not engage in express advocacy comprise a negligible portion of outside spending. For instance, in 2018, electioneering communications comprised less than 1 percent of outside spending. I exclude electioneering communications as a separate category because of their less common use. https://www.opensecrets.org/outsidespending/fes_summ.php

candidates to alter their strategy compared to elections in which the rivals are trivial (Burden 2004).

Candidate characteristics – namely party affiliation and incumbency status – are also likely correlated with data spending patterns even after accounting for financial resources and other factors. The main finding of this dissertation so far indicates that Republicans' and Democrats' acquisition and use of data sources differ in a myriad of ways. Compared with the other variables, a strong hypothesis is that a candidate's party is a robust determinant of their data spending patterns. Compared to Republicans, Democrats are expected to devote larger portions of their data budgets toward individual-level data purchases even after accounting for all the other variables in the model. To test this, I include a binary variable to indicate the candidate's party. Additionally, incumbency may also affect the kinds of information that campaigns purchase. Incumbents may be less likely to field polls, for instance, because they have more familiarity with their standing in the district. Incumbency is similarly measured with an indicator variable for legislators who are seeking reelection.

Campaigns may also consider the demographic characteristics and physical boundaries of the district when trying to make inferences about the electorate. Past research, for instance, indicates that the quality and completeness of the voter file in most states varies systematically with voter-level socioeconomic characteristics (Igielnik et al. 2018). Whiter, older, and wealthier voters are more likely to be listed and listed correctly, while racial minorities, young people, and the poor are likely to be unlisted or mislisted (Jackman and Spahn 2021). I proxy these variables at the district level through measures of the percent white, percent who have attained at least a college degree, median household income, and median age coming from one-year estimates provided by the U.S. Census American Community Survey (ACS). Beyond these measures, I also capture the percent urban estimate of district constituents because of a possible positive relationship between

individual-level data prioritization and urbanity. A higher population density certainly makes direct contact involving on-the-ground field operations more feasible, so campaigns might also invest in the individual data that helps them implement that strategy.

Another district characteristic to consider is turnout. As with the strategic consideration of competitiveness, the number of people expected to turn out may change the decision of any investment and the choice between either polls or voter file records. The models include the turnout rate of the citizen voting age population (CVAP) in the district calculated by dividing the FEC's official counts of total ballots cast by citizen voting age estimates from the ACS one-year survey. Relying on citizen voting age as the denominator provides a more accurate estimate of true turnout rates compared to the voting-age population (McDonald and Popkin 2001).

Finally, I also consider the availability of records in each state's voter file. Some states require citizens to provide more information than others when they register to vote. Of particular interest to campaigns are records related to partisan affiliation and the racial identification of the voter. Indicators of voters' party affiliation come in two forms from state voter files. The first is party registration. Voters register as a member of a political party. The second is primary data. The voters' choice to participate in primary elections is recorded. Seventeen states have both. Twenty-five states have one or the other. And eight states have neither. To proxy these differences, I create an indicator variable based on a review of state laws to account for these ten states in reference to the other forty. Likewise, I also account for a state requiring racial identification as part of the voter registration process. Racial identities are highly correlated with partisanship and make mobilizing based on racial identity much easier. An indicator variable captures the eight states, all

in the South, that collect racial information (Hersh 2015).[53] If what is available in the voter file affects if campaigns invest in and prioritize it as a form of data, then these variables are appropriate proxies to identify that relationship.

## 4.2 Modeling Data Spending Proportions

As explained in chapters one and three, the purpose of my expenditure record analyses is to reveal the preferences that campaigns have for different sources of electoral information. Rather than modeling the amount spent on electoral information sources, my dependent variable is the proportion each campaign spent on individual-level data or polling out of total data-related spending during each election cycle. For instance, if a campaign spent $8,000 on polling and $2,000 on individual-level data, then they would be represented in my dataset as having spent 0.8 of their data budgets on polls and 0.2 on individual-level data. By modeling the proportion of data-related spending, my analysis offers comparable insights into the preferences that campaigns have for different sources of electoral information across different races and cycles. Additionally, my unit of analysis is each unique candidate-cycle campaign restricted to major party candidates running in the general election (n = 4,910). I include in my proportion estimates polling and individual-level data purchased during both the primary and general elections because my focus is on how campaigns choose to allocate their limited data budgets throughout the entire election cycle. Furthermore, it is ambiguous whether spending on sources of electoral information before a primary election date is intended for strategic planning solely during the primary, for the general election, or both. For instance, a campaign could purchase a contract with a polling company for

[53] The states are Alabama, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and Tennessee. Mississippi and Tennessee do not require racial identifiers to be recorded but still list the field on voter registration forms, resulting in fewer residents providing it.

the entire election cycle at the start of a primary or similarly invest in access to a voter database early on that continues into the general election.

Because my variation of interest comes in the form of proportions, I rely on a specialized modeling technique that accounts for the bounded nature of these variables. Bounded continuous dependent variables are a challenge for common statistical techniques. Complicating my modeling specification more, the previous chapter's inspection of individual-level and polling percentages out of total data-related spending revealed U-shaped distributions with peaks at zero and one with fewer observations between zero and one. A common ordinary least squares (OLS) regression is inappropriate because it assumes that the dependent variable is continuous and normally distributed. Common approaches to modeling binary variables such as logit regressions fail to capture the distribution between zero and one. Other popular alternatives for modeling proportions, such as compositional modeling techniques, can account for the proportions between zero and one but are unable to model observations at zero and one.

Given common regression techniques are inappropriate, I use a statistical model rarely used in political science known as a zero-one-inflated beta regression (ZOIB).[54] I choose the ZOIB regression not only for statistical but also for substantive reasons. When it comes to campaign data budgets, campaigns can allocate all their resources toward one source or the other, or they can employ a mixed data investment strategy with some spending on polls and some spending on individual-level voter records. The decision to invest entirely in one source or the other is a considerably different decision from how much to spend on polling versus individual data. The

[54] The presence of an unequal number of zeros and ones in the dependent variable precludes using compositional or fractional modeling techniques able to estimate the individual-level data and polling regressions simultaneously. My review of extant compositional methods found no other viable alternative that can recover effect estimates for traditional hypothesis testing (see Alenazi 2022; Tsagris, Alenazi, and Stewart 2022; Tsagris and Stewart 2018 for reviews of current compositional methods and limitations).

ZOIB regression takes this into account by simultaneously modeling the data proportions with different equations within the same model. The model assumes that different variables explain the variation for proportions at one, zero, between zero and one, and the variance between zero and one. All four of these "components" of the proportion estimated together in the same model represent all the variations of interest when considering the limited financial resources campaigns choose to devote to their electoral information budgets. In other words, the ZOIB model recognizes that campaigns may respond to different strategic considerations, candidate characteristics, and district-level factors when deciding to go all-in on polls or individual-level data compared to when they are allocating their limited data budgets between the two sources.

In essence, the ZOIB model achieves simultaneous estimation by combining a beta regression with logistic regressions to account for the inflated bounds of the proportion. A standard beta regression requires that the dependent variable is the interval (0,1), meaning the proportion never reaches one or zero. A ZOIB regression, introduced by Ospina and Ferrari (2008), measures the dependent variable on the [0,1] interval and thus includes the entire range of possible values for the proportion of data-related expenditures spent on individual-level data and polling. The ZOIB regression accounts for both the excess of zeros and ones in the dataset and the continuous component between the bounds by simultaneously modeling the proportions as different components. Specifically, the ZOIB regression assumes that the dependent variable, *y*, is a closed interval ranging from zero to one. The probability density function of the proportion *y* is defined by:

$$p(y) = \begin{cases} p_0 & if\ y = 0 \\ p_1 & if\ y = 1 \\ \dfrac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} & if\ y \in (0,1) \end{cases}$$

where $y \leq 0 \leq 1, \alpha > 0 \; and \; \beta > 0$. $p_0$ is the probability of zero, and $p_1$ is the probability of one. The binary components $p_0$ and $p_1$ are modeled using a logistic regression. The continuous distribution of the proportions is modeled using a beta regression and function defined as:

$$B(\alpha, \beta) = \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}$$

where $\alpha > 0$, $\beta > 0$ are respectively the shape and scale parameters.

Ospina and Ferrari (2008) propose interpreting the ZOIB regression in terms of four components. In application, the model is re-parameterized to be defined by components $\mu, \sigma, \nu, and \; \tau$:

$$\mu = \frac{\alpha}{\alpha + \beta} \qquad \sigma = \frac{1}{\alpha + \beta + 1}$$

$$\nu = \frac{p_0}{1 - p_0 - p_1} \qquad \tau = \frac{p_1}{1 - p_0 - p_1}$$

$\mu$ is the mean of the beta distribution between zero and one. $\sigma$ is the variance of the beta distribution (proportional to the scale parameter). $\nu$ is the zero-inflated parameter capturing the probability of zero, and $\tau$ is the one-inflated parameter capturing the probability of one. In my application, $\mu$ represents the estimated proportions of individual-level data *or* polling that a campaign invests in after having purchased both data sources. $\sigma$ captures the variability of those bounded probabilities. Conversely, $\nu$ is interpreted as the probability of a campaign *not* investing in any polling or individual-level data, and $\tau$ signifies the likelihood that a campaign invests in *only* individual-level data or polling.

Each of the four model components is estimated as a function of all covariates discussed in the

previous section and with their effect estimates standardized for visual comparison. The models include variables related to financial resources, outside spending, competition, candidate characteristics, district-level factors, and voter file indicators. Additionally, the model includes state and cycle fixed effects to generate effect estimates applicable across the entire period of investigation between 2006 and 2018 and accounting for different state-level conditions that may affect data spending patterns.

Specifically, the components are specified as:

$$logit\left(\mu_{S_{ijt}}\right) = X_{ijt}b_1$$

$$logit\left(\sigma_{S_{ijt}}\right) = X_{ijt}b_2$$

$$log\left(\nu_{S_{ijt}}\right) = X_{ijt}b_3$$

$$log\left(\tau_{S_{ijt}}\right) = X_{ijt}b_4$$

The dependent variables are the components for data source *S* for candidate *i* in district *j* in election cycle *c*. The $\mu$ and $\sigma$ parameters rely on a logit link function to ensure bounded predictions between 0 and 1. The $\nu\ and\ \tau$ parameters are linked to the log function to maintain positive probabilities. While estimations for the $\mu$ component of the individual-level data and polling models are additive inverses of one another (opposite signs) and the $\sigma$ variance estimates are equivalent, the $\nu$ and $\tau$ estimates are not perfectly correlated because of the presence of an unequal number of observations reporting zero spending on individual-level data and polling because some campaigns choose to not purchase either source, as reported in the previous chapter. $X_{ijt}$ is the vector of independent variables listed in Table 4.1 as well as state and cycle fixed effects. Given the outsized role that spending plays in campaign-level decision-making and previously documented party-level

differences, the vector $X_{ijt}$ also includes an interaction term between campaign spending and party. $b$ represents the standardized coefficients to make their different scales directly comparable and ease interpretation within and across parameterized components. The effect estimates are standardized by dividing by two standard deviations, so each can be interpreted as if they are untransformed binary indicators (Gelman 2008).

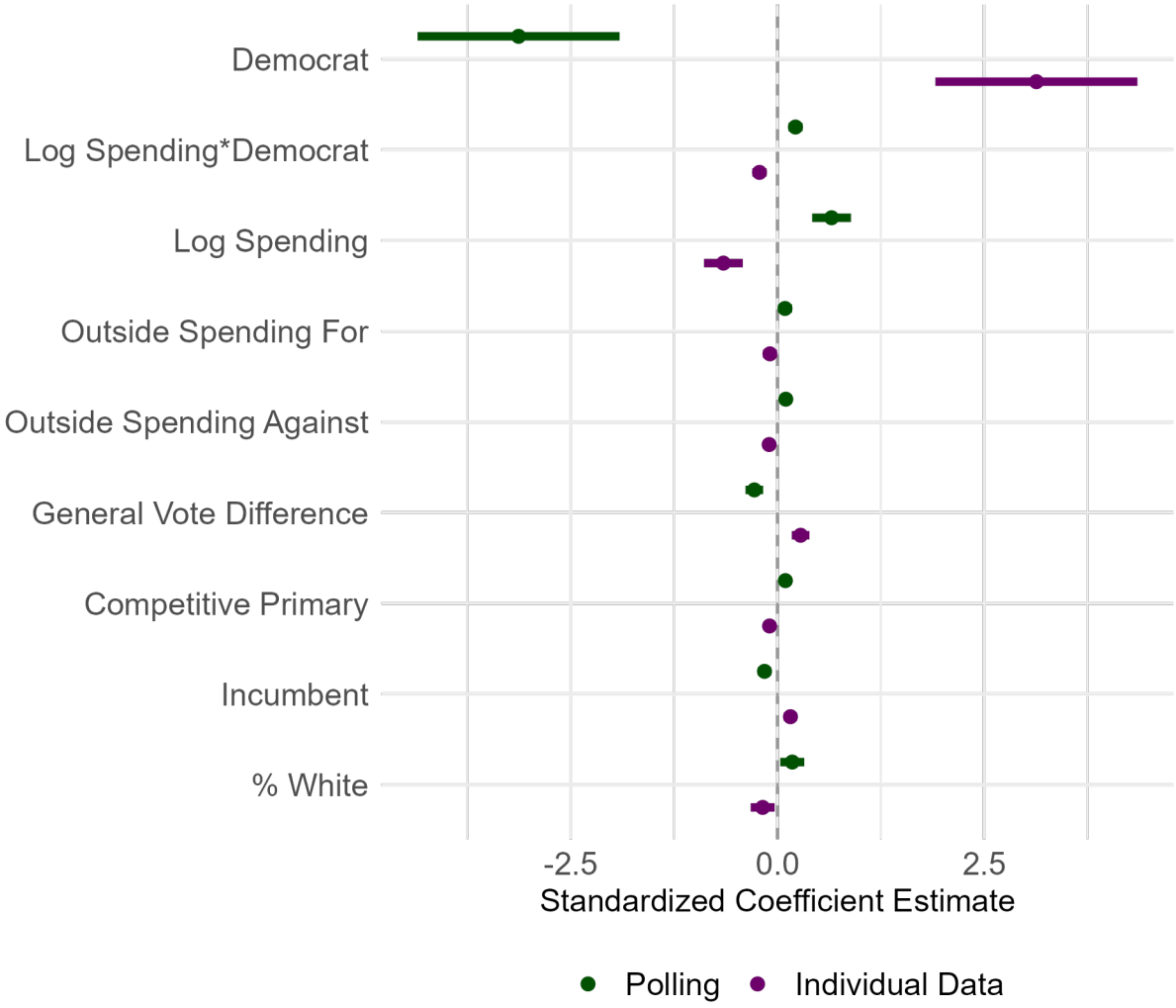
## 4.3 Explaining Campaign-level Data Spending Patterns

Having outlined the modeling approach, this section seeks to determine if party remains a consistent factor after accounting for various factors campaigns consider when crafting a campaign strategy. I proceed by examining the results of all four components of the ZOIB regressions separately before summarizing commonalities between them. While the components of the individual-level data and polling proportion models reported below are ostensibly separate, their estimation is simultaneous and thus their interpretation must be holistic. In line with the party divergence findings in the previous chapters, I find that a congressional candidate's partisan affiliation plays a persistent role in observed data spending patterns, but campaigns are also attentive to strategic consideration when choosing to allocate limited financial resources to learning about the electorate.

Figure 4.1 presents the standardized effect estimates for the mean proportion devoted to either individual data or polling, the $\mu$ parameter of the ZOIB regression. To ease visual inspection, the below figures only contain statistically significant (95% confidence interval) effect estimates, but the entire models can be found in Figure D1 in Appendix D. Because the figure includes standardized coefficients, visually inspecting their *relative* distance from zero indicates each variable's substantive impact on data spending patterns compared to the other variables. By nature

of the ZOIB estimation approach, the estimated effects on individual-level data and polling proportions are mirror images of one another. I include the mirrored estimates in Figure 4.1 because they differ in the zero and one components of the models (Figure 4.4 and Figure 4.5).

**Figure 4.1 Explaining Mean of Data Spending Patterns between 0 and 1 (μ)**



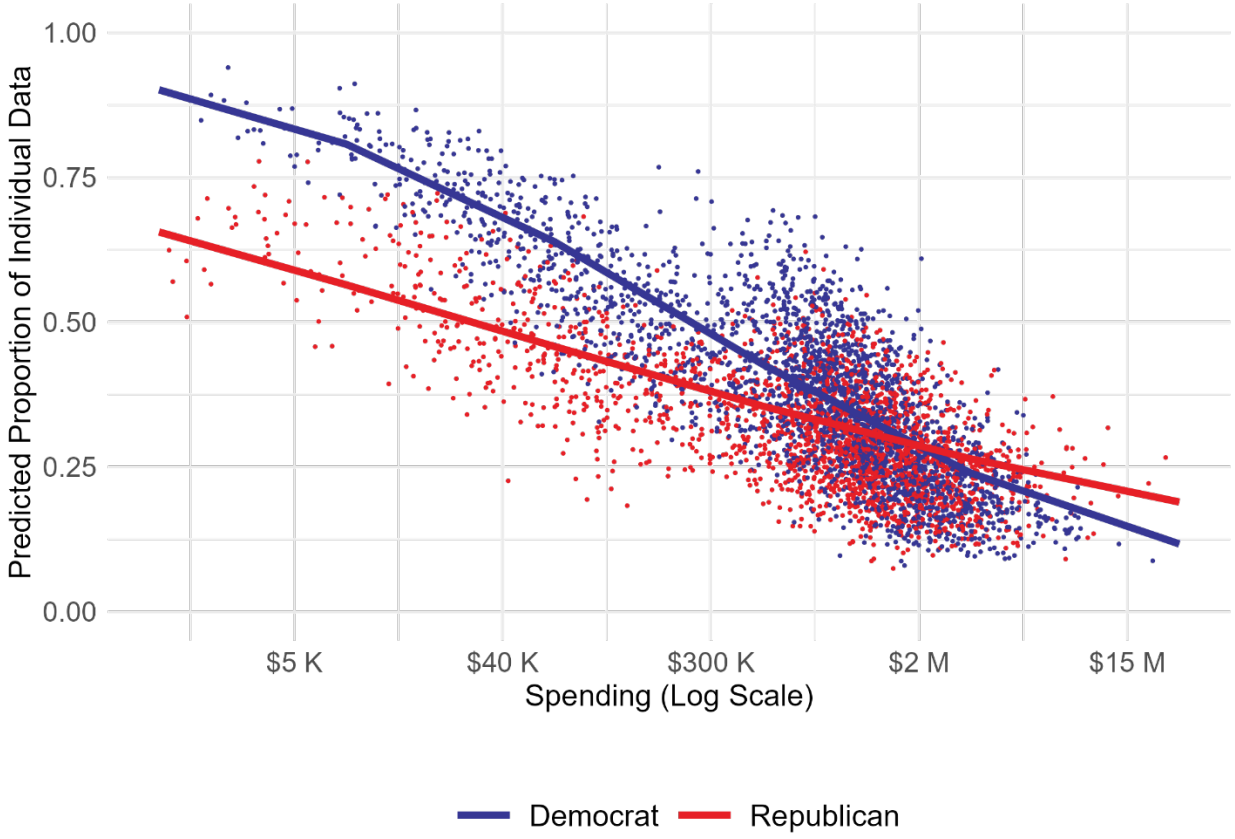

As seen in Figure 4.1, party has the largest relative effect of any statistically significant variable included in the model.[55] The direct, non-interactive effect of party, listed in the model as

[55] The party variable's larger confidence intervals compared to the other effect estimates are the result of inflated standard errors because of the inclusion of an interaction term.

"Democrat," is more than double the effect of total campaign-level spending. Higher levels of overall spending do correlate with data spending patterns, increasing the amount campaigns allocate toward polls while reducing spending on individual-level data, but party affiliation plays a central role in the decision to acquire sources of electoral information even after accounting for other strategic factors. Democrats running for the House are more likely to spend their data-related budgets on individual-level data, while Republicans prefer polling.

While Democrats choose to invest more in individual-level data compared to their Republican counterparts, closer inspection of the spending-party interaction term reveals a dampening and even potential reversal of the pattern at higher levels of spending. As campaign-level spending increases, Democrats begin to spend less of their data budgets on individual-level data and more on polling, while the opposite is true among Republican congressional candidates. To visualize this interaction, Figure 4.2 plots the predicted proportion of spending on individual-level data split up by party as logged total of campaign-level spending increases. Note that the relationship for polling proportions is the additive inverse and thus equivalent in the opposite direction. Additionally, the predictions do not include one or zero because the ZOIB estimation method models those two components separately.

**Figure 4.2 Predicted Data Purchasing Patterns by Campaign Spending by Party**




Whereas Democratic candidates running for the House are more likely to invest in individual-level data, these differences vanish at higher levels of campaign spending, as illustrated in Figure 4.2's fitted prediction lines. The highest-spending campaigns on both sides match one another in terms of their data spending allocation patterns. This is mainly because Democrats spend larger shares on polling as their total spending increases. As a Democratic campaign's total spending approaches $1 million or more, they will purchase more polls while their spending on individual-level data remains relatively constant, as a result, lowering the proportion of their data budgets spent on individual-level data. Republicans, already starting at lower base levels of spending on

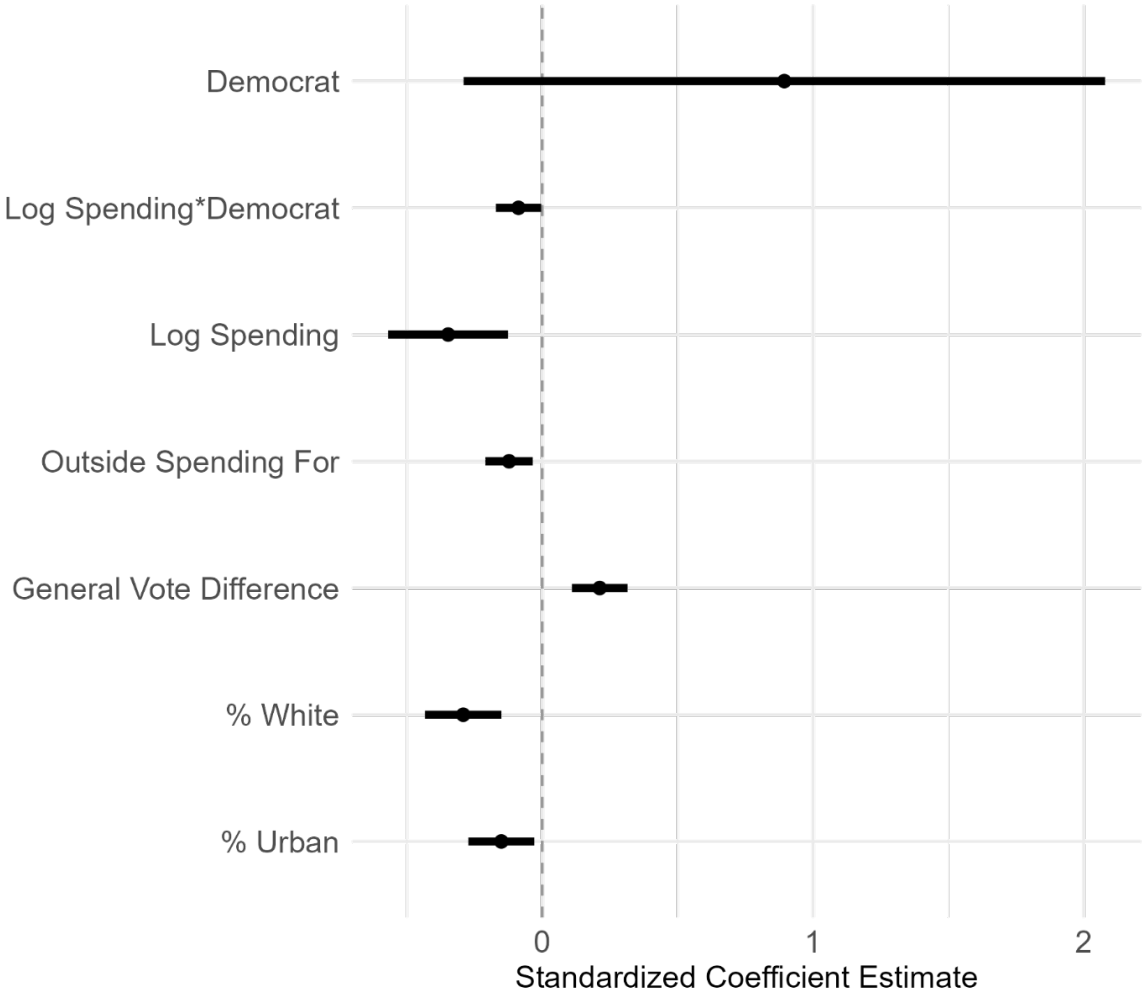
large voter databases, similarly increase their spending on polling when they have more money, but the shift in their proportion is not as dramatic as on the Democratic side. The conditional relationship between party and spending helps to illustrate that party-level differences in data spending do indeed appear to be robust but are also dependent on many other factors campaigns commonly confront.

Beyond spending attenuating the observed party-level differences, Figure 4.1 also reveals other factors that correlate with data spending. Overall, a congressional campaign's preference for data sources is also responsive to levels of outside spending by interest groups, strategic considerations of competitiveness, and incumbency status. Outside spending increases the relative amount campaigns spend on polling. Though difficult to deduce, one potential explanation, explored in the next chapter, is that outside spending makes polling more likely because increased outside attention and mass media spending makes it more difficult to gauge the relative impact of their campaign activities. Outside spending may well also serve as an alternative measure of district-level competition or race saliency. Likewise, although the two competitiveness measures have opposite signs, they both indicate that increased competition leads to greater spending on polls. Specifically, as general vote difference increases, campaigns spend less on polls, while a competitive primary increases a campaign's poll spend. Incumbent members of Congress also spend less of their data budgets on polls, likely a result of better familiarity with their standing in the district. After accounting for the effects of the other variables, district-level factors have less of an impact on a campaign's data source allocation patterns, with only racial composition, measured as the percentage of the district identifying as white, having an observable impact.

Many of the same strategic factors that influence the mean data spending patterns of congressional campaigns also affect the variability of their data spending patterns. Figure 4.3

reports ZOIB's σ variance component. The figure only reports one set of standardized coefficients because they are equivalent across individual-level data and polling models. While the direct, non-interactive effect of party affiliation does not reach traditional levels of significance, the interaction term reveals the conditional effect of party. At higher levels of spending Democratic congressional candidates have less variability in their data preferences compared to Republicans.

**Figure 4.3 Explaining Variance of Data Spending Patterns between 0 and 1 (σ)**




Total campaign spending reduces variability as well, while less competitive districts see more variability, as indicated by the positive effect of general vote difference. The effect of outside spending meanwhile is mixed. Increased outside spending in support of the candidate decreases

variability while spending against a candidate increases the variability of data spending patterns. The responsiveness of congressional candidates' data spending patterns to measures of spending and competitiveness in Figure 4.3 comport with the findings of the $\mu$ component. Campaigns respond to strategic considerations even with observable party-level differences. Meanwhile, candidates running in less racially diverse and more urban districts also have less variability.

The final two model components indicating a campaign's decision to *not* make any investment into polls and individual data or to *only* invest in either polls or individual data (to the exclusion of the other data source) reveal the continued impact of party affiliation but also highlight the complicating influence of spending, competition, and even some district characteristics. Figures 4.4 and 4.5 respectively present the effect estimates for component $\nu$, indicating zero spending, and component $\tau$, indicating a data spending proportion of one. Because some campaigns spend nothing on both sources, leading to excess zeros, the two figures' estimates do not necessarily mirror one another. In other words, the variables that correlate with campaigns spending *nothing* on either individual-level data or polls are not necessarily the same factors that relate to campaigns purchasing *only* individual data or *only* polls.

Figure 4.5 uncovers that the decision to do without electoral information sources is more dependent on spending, although party affiliation still has a discernable impact. It reports the statistically significant standardized effect estimates for the $\nu$ zero-inflated component of the ZOIB regression. Note that missing effect estimates in reported variables indicate results that do not reach a traditional level of statistical significance, which are reported in Appendix D. The baseline, non-interactive effect of partisanship indicates that a candidate being affiliated with the Democratic party is more likely *not* to make any purchase of individual-level data and less likely to *not* purchase a poll. In other words, after accounting for overall campaign-level spending and

the effect of other factors, Democrats are more likely to forgo individual-level but less likely to forgo polling. As with the other components, however, the relationship is conditional on total campaign-level spending. As spending increases, Democrats are less likely to go without individual-level data, while Republicans are less likely to go without polling. The responsiveness of the spending-party interaction term to data spending patterns comports with the observed patterns in the other model components, depicted in Figure 4.1 and Figure 4.3.

Relative to party, Figure 4.4 reports that the largest relative determinant of the decision to spend nothing on either source is spending. The decision to not make any data-related purchases is also correlated with a wider range of observable covariates than the previous two components. Outside spending again has an inconsistent effect but this time across different data sources with increased outside spending in support of the candidate making it less likely campaigns will spend nothing on polls, while increased spending by outside groups in opposition makes it more likely a candidate will not purchase individual-level data. The competitiveness variable exhibits a similar pattern reported above where, the more competitive a race is, the more likely a candidate will choose to make data-related purchases. Likewise, incumbents are more likely than challengers or candidates in open-seat races to go without either source. The initial decision to invest in either source of electoral information is also more responsive to district-level characteristics depending on the source. For example, more urban districts are more likely to not purchase a poll.

**Figure 4.4 Explaining Zero Data Spending (v)**

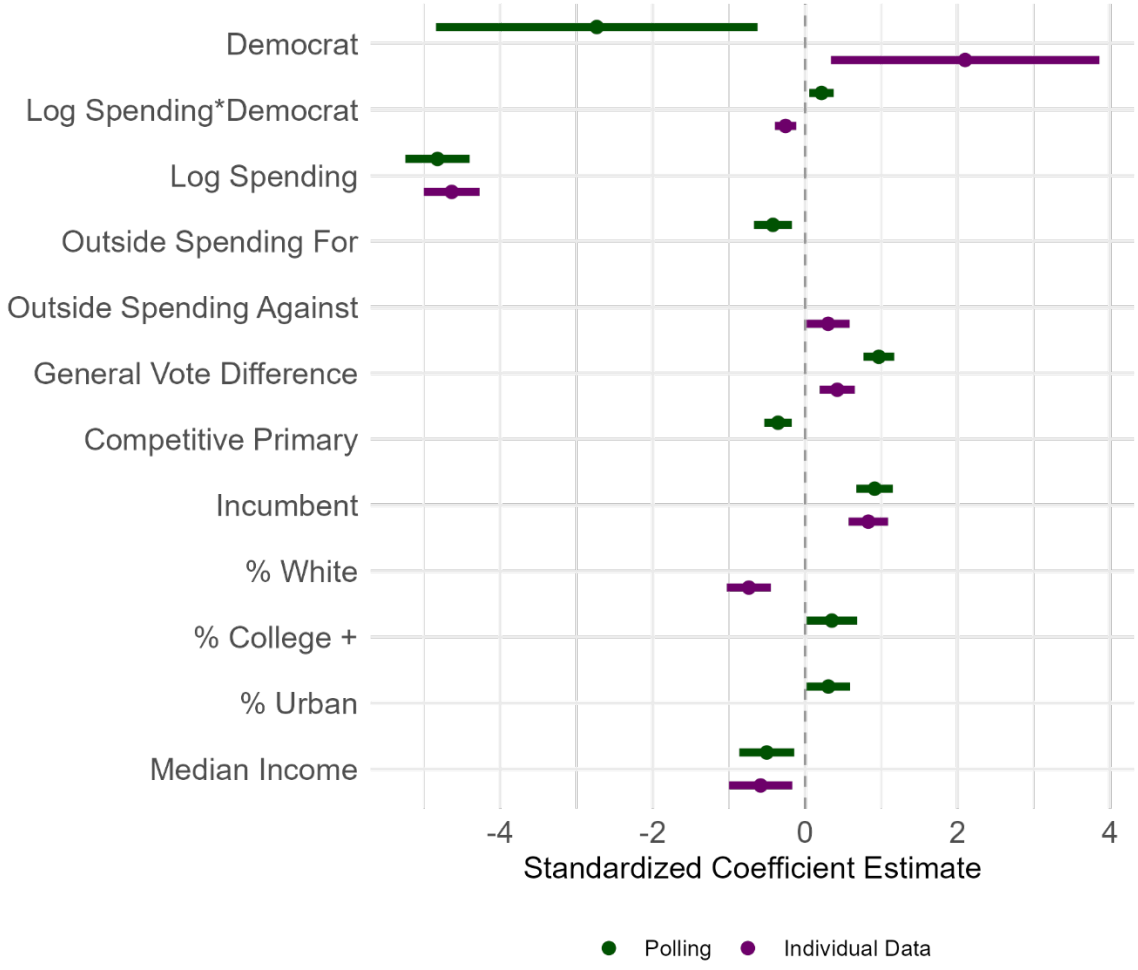
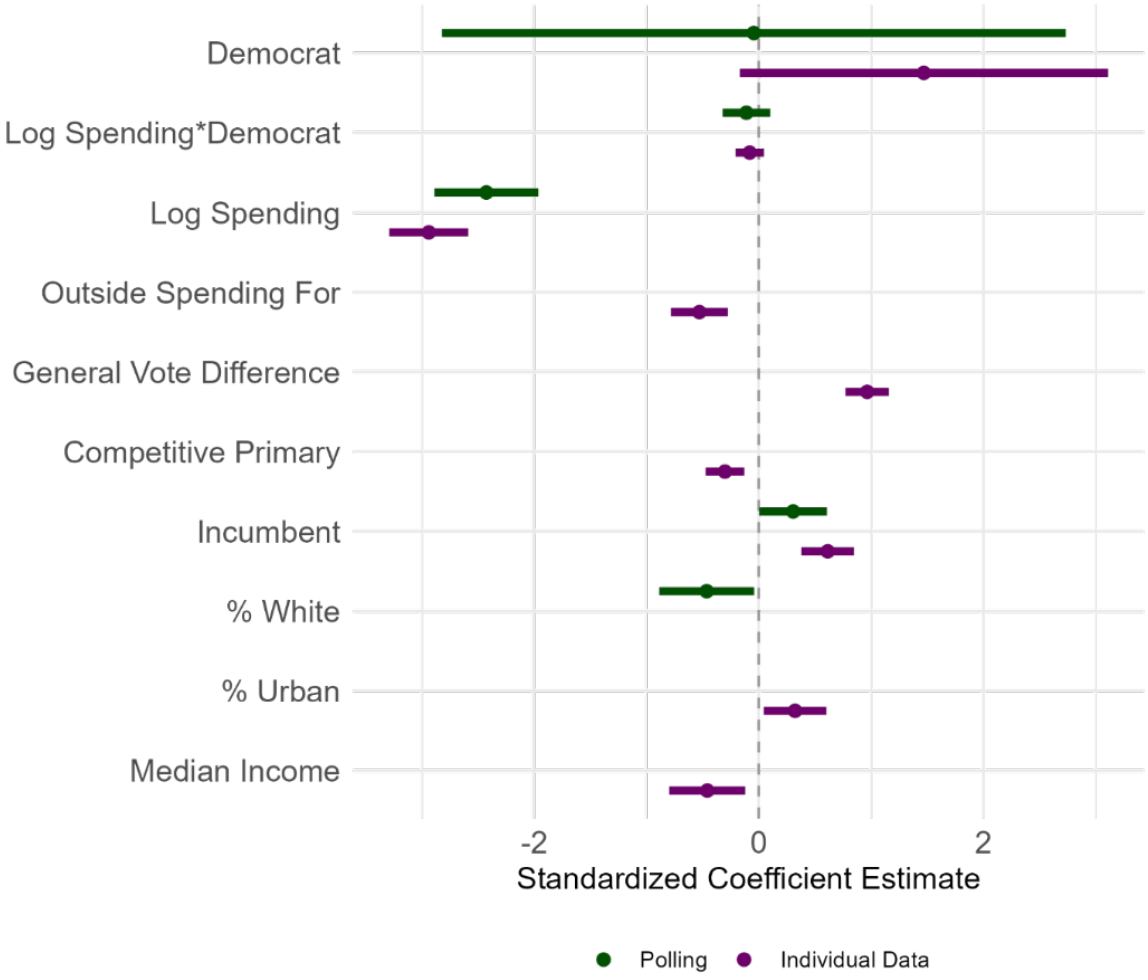



Figure 4.5 reporting the results of the τ one-inflated component reveals that partisanship has little to do with the decision to only purchase one source or the other. The figure only displays statistically significant results except, in this case, for the insignificant party effect estimates. Neither the baseline, non-interactive effect of party nor its interaction with spending reach statistical significance. Rather, total spending is the most important determinant. Campaigns that spend more are likely to employ a mixed data investment strategy, spending a portion of their budgets on polls and another portion on individual data. Likewise, competition stimulates a mixed data strategy rather than only purchasing one type.

**Figure 4.5 Explaining One-Source Data Spending (τ)**




Like findings related to the other components, incumbents are more likely to only purchase access to one data source, if they invest any money into electoral information at all. Unlike the mean and variance components, many more district-level characteristics appear to influence the choice of only investing in one source of information even after accounting for spending and strategic factors. For instance, campaigns operating in more urban districts tend to invest only in individual-level data compared to their more rural counterparts.

Overall, analyzing all four components of the carefully considered ZOIB regression modeling the proportion of spending candidates devote toward either individual-level data or polling

indicates that while party plays a major role in data spending patterns candidates also respond to other on-the-ground considerations. Campaigns do not occur in a vacuum. Whether choosing to gauge their electoral chances with polling or informing their outreach with access to large voter databases, congressional candidates must make their decisions considering financial constraints, spending by outside groups, and their strategic advantages or disadvantages. Most prominently, strategic concerns about the competitiveness of the district and outside dollars spent on the race induce candidates to acquire multiple sources of data. They not only pay for access to large individual databases, but the more competitive environment also appears to justify the high price tag of a poll. Likewise, incumbents tend to have different data spending patterns compared to challengers and open-seat candidates. Current congress members are less likely to purchase polls and employ a mixed data-informed strategy that includes both electoral information sources.

Unlike other considerations, my analysis indicates that candidates' data spending patterns are less responsive to district characteristics and the amount of information contained in the underlying state voter file. Candidates do sometimes adjust their data purchasing plans in response to their districts' characteristics, such as the correlation between urbanism and the decision to make an initial investment in individual-level data, yet a holistic interpretation considering all four model components suggests that most district characteristics have unrobust and inconsistent effects. Analysis of the four components provides no evidence that congressional campaigns change their data spending patterns because of state laws dictating the availability of individual-level information in the voter files. Candidates purchase access to voter-level data irrespective of whether the underlying voter file contains party identification information related to official party registration or partisan primary participation.

Altogether, the results suggest that a candidate's partisan affiliation is a key determinant of

campaign-level data spending patterns. To increase confidence in the persistence of party-level differences, Appendix D provides alternative specifications and robustness checks across all four components. Figure D2 reports a logistic regression where the dependent variable is an indicator of whether a campaign purchased any individual data or polling. Results indicate Democrats are more likely than Republicans to invest in individual-level data compared to polls, but, like the above findings, the initial investment is most influenced by overall campaign spending amounts. Figure D3 reports an OLS regression of dollar amounts spent on the two sources. It indicates that Democrats spend more on individual-level data but no differences in the amount spent on polling between the two parties. Figure D4 reports party effect estimates with an alternative specification of the ZOIB's μ mean modeled without specifying the other three components and including other combinations of the observed covariates. The effect of party remains positive across the models that iteratively add variables from a binary party model up to the full fixed effects μ model without the party-spending interaction. Figure D5 reports all four components simultaneously modeled without the party-spending interactions. While the effect of party is insignificant in the μ component, the other three parameters demonstrate an effect estimate consistent with other models where Democrats prefer individual-level data over polling. Figure D6 shows parallel results for party-level differences when restricting the sample to campaigns in districts with final general election vote margins of less than or equal to 5% (n = 400). The results are also robust to an alternative measure of district competitiveness. Figure D7 re-estimates the model components with a variable capturing the Democratic advantage measured as the Democratic presidential nominee's vote share in each congressional district in the immediately preceding contest.[56]

[56] Estimates of Democratic presidential vote share in congressional districts during the 2004, 2008, 2012, and 2016 contests come from Daily Kos. https://www.dailykos.com/stories/2022/11/14/1163009/-Daily-Kos-Elections-presidential-results-by-congressional-district-for-2020

## 4.4 Summarizing Campaign-level Data Spending Patterns

Scholars have taken stock of new developments in data-driven campaigning at different points in time. Going back to the innovations of the 1960s, Agranoff (1977) explains how advancements in punch-card computing led to a new style of campaigns defined by "combining basic secondary data sources (aggregate electoral statistics, census data), political data (canvass information, registration), and political judgment into a profile of the constituency" (123). Such descriptive accounts of innovations occurring at the top of the ticket have since dominated the study of campaign technology. Whether describing further advances because of increased computing power (Selnow 1993) or the shift from demographically and geographically oriented data to individual data and predictive modeling (Hersh 2015; Nickerson and Rogers 2014), scholarly knowledge of developments in down-ballot, data-driven campaigning has as until now only been accessible via elite surveys and qualitative case studies (Baldwin-Philippi 2018; Hatch 2016; Kefford et al. 2022; Nielsen 2012).

My extensive refinement of data-related spending records reported by congressional candidates and the close attention I paid to modeling their data preferences provide the most comprehensive overview of data purchasing determinants to date. Combining all the findings of my dissertation up to this point, I detail how party affiliation is a consistent point of separation in the acquisition of voter data. As uncovered in chapters two and three, Republicans and Democrats not only rely on different partisan firms to learn about the electorate but also have different approaches to combining and interpreting sources of information about the electorate. The shared VAN data platform across Democratic campaigns and party-linked organizations translates into operatives working under the same set of assumptions on when and where to leverage individual-

level records on voters. Republicans, on the other hand, lack not only a common platform but commonalities in their approach to merging voter-level data with other signals about the electorate.

Yet the story of campaign-level decision-making is more complicated than merely the different partisan data cultures or ecosystems. Campaign-level data preferences are also a function of ever-present concerns over the level of competition and cash on hand in addition to common advantages such as incumbency. It appears that campaigns pay little attention to the compositions of their districts or the information contained in underlying voter registration records when purchasing data. As nuanced in my interviews, this finding is partially attributable to the fact that most political practitioners have a predetermined approach to how they acquire and use data regardless of where the race occurs. Additionally, because campaign decision-makers often interact with third-party individual-level data cleaned and organized to reduce voters to a few predictive scores, they operate largely unaware of the impact that individual-level data availability may have on the accuracy of their inferences.

The story behind partisanship's influence over electoral information is far more complicated than its estimated effect in regression models. As discussed in chapter one, campaigns rely on a network of party-affiliated actors, including official party organizations, political consultants, voter data firms, and partisan polling vendors, to learn about and interact with their electorate. The observed party-linked preferences for sources of electoral information depends not only on the hundreds of small decision made within campaigns but are also the collective result of party-wide priorities and investment into campaign technology and innovation that are, like our contemporary political environment, divided by partisanship. The different approaches that the Republican and Democratic parties took to developing a shared system for storing and sharing data on voters in the early 2000s continue to shape whether their candidates make big data on voters central to their

campaign strategy nearly two decades later.

Still, the importance of uncovering persistent campaign-level differences between the two parties' candidates should not be understated. The Democratic party's official VAN platform was successful not because it was created but instead because Democratic campaigns and progressive groups around the country embraced it. One only has to go as far back in history as two years before the VAN emerged in 2006 to find Democratic campaigns and state parties around the country firmly rejecting attempts by the DNC to establish party-wide data sharing practices and a nationwide voter file. Their collective reorientation toward big data in politics came only when campaigns up and down the ballot, in part inspired by the success of Obama's two presidential campaigns, chose to make the VAN a core component of their decision-making and their on-the-ground efforts to communicate with voters. The Republican party never experienced such a collective reorganization of data priorities in part because their candidates continue to prefer polls over individual-level data, as the evidence in the chapter makes clear.

Left to explore is whether these collective differences in data priorities have observable consequences for how campaigns choose to communicate with voters. The next chapter unpacks the relationship between congressional campaigns' data spending patterns and their voter outreach strategy.

# 5 Zooming Out on the Impact of Voter Data

Whereas my investigations of data-related expenditures thus far have concentrated on understanding why campaigns vary in their data spending patterns, this chapter puts my picture of data-driven campaigning into a larger frame of reference by exploring the impact of individual-level data and polling on the outreach strategies employed by candidates running for the U.S. House of Representatives. As discussed in chapters one and two, campaigns leverage a combination of polling and individual-level data to help them make decisions about who in the electorate should be the attention of their outreach efforts. This process of segmenting the electorate into different "buckets" for mobilization and persuasion is foundational both at the planning and implementation stages of voter outreach. Just as campaigns choose whether to purchase sources of electoral information or not, campaigns also make decisions about whether to contact voters and how they wish to get their message across to them.

Unlike the previous chapters, political science scholarship has paid much more attention to campaign interactions with voters, yet few studies to date have been able to investigate the campaign-level relationship between data and outreach strategy. Numerous studies have examined campaigns' mass media advertising, digital media outreach, and direct contact efforts with various approaches to document not only spending patterns but also the causal effect of these different modes of outreach. Continuing to leverage my refinement of FEC spending records to make novel insights, I now turn to detail spending on different forms of outreach between 2006 and 2018 before examining whether the kinds of information campaigns purchase to learn about the electorate have a demonstrable correlation with outreach spending patterns after accounting for other observable strategic considerations. Before turning to my analyses, I first briefly review

existing literature investigating different modes of campaign outreach and develop hypotheses informed by prior scholarship and my findings in previous chapters.

In brief, this chapter uncovers that the growing availability of individual data between 2006 and 2018 has not overhauled how congressional campaigns choose to communicate with the electorate. While research focusing on presidential campaigns finds a shift toward field operations at the top of the ticket, these patterns are not mirrored down the ballot. Congressional campaigns are remarkably consistent over time in how they allocate their limited budgets across mass media, direct contact, and digital media. Levels of spending on direct contact remain stable across my period of investigation with the primary shift being an increase in digital communications at the expense of mass appeals. Yet mass media advertising continues to make up most of the money campaigns spend to communicate with voters. Notwithstanding these outreach spending patterns, I also find a persistent relationship between data preferences and outreach strategy that remains after accounting for the influence of common strategic considerations. Combining these results in the context of my findings in previous chapters, I argue that large voter databases have not upended politics but rather have reinforced longstanding practices that reduce voters to mere electoral math.

## 5.1 Reaching Out to the Electorate

Mass media advertising typically makes up the largest category of outreach expenditures in both presidential and congressional campaigns (Herrnson, Panagopoulos, and Bailey 2019; Wayne 2019). While political scientists disagree over the precise impact of these efforts, some studies suggest that political advertising venues such as television and radio can have small, marginal effects on voter behavior (e.g., Ansolabehere and Shanto 1995; Brader 2005; Fowler, Franz, and Ridout 2016; Franz and Ridout 2007; Goldstein and Freedman 2002; Johnston, Hagen, and

Jamieson 2004; Overby and Barth 2006; Panagopoulos and Green 2008, 2011), even if these effects are extremely short-lived (Bartels 2014; Gerber et al. 2011; Johnston, Hagen, and Jamieson 2004; Sides, Tesler, and Vavreck 2018; Sides and Vavreck 2013). Unlike political scientists, campaign practitioners appear to be much less concerned over the causal validity of mass appeals on television and radio or through print. Campaigns invest more than half of their limited outreach budgets on these traditional media outlets because they want to communicate with a large number of potential voters and believe it is an effective way to mobilize and persuade voters (Diamond and Bates 1992; Williams and Gulati 2018).

Campaigns also increasingly spend their limited resources on digital advertising across different online platforms such as Facebook, Google, and YouTube. As with mass media advertising, a large body of research from political science and related fields finds that these digital appeals have a small impact on voters' behavior and especially serve as effective tools for mobilizing supporters and gathering contributions (e.g., Aldrich et al. 2016; Ballard, Hillygus, and Konitzer 2016; Bimber 2014; Boulianne 2018; Chester and Montgomery 2017; Dimitrova et al. 2014; Haenschen and Jennings 2019; Jungherr, Rivero, and Gayo-Avello 2020). And while television expenditures continue to make up the largest proportion of voter outreach expenditures, digital ad buys have risen significantly since the early 2000s and make up a growing portion of campaign outreach budgets (Franz et al. 2020; Williams and Gulati 2018).

Campaigns also engage in direct and in-person outreach activities. While traditional and even digital advertising often reach wider segments of voters (Franz et al. 2020), direct contact activities are aimed at smaller segments and are designed – at least in theory – specifically for individuals (Burton, Miller, and Shea 2015). This type of outreach has seen greater attention from political scientists in recent decades. The literature is replete with field experiments measuring the causal

effect of different approaches to making phone calls, sending text messages, canvassing neighborhoods, or sending direct mail on various outcomes such as turnout, persuasion, and financial contributions. (Aldrich et al. 2016; Alvarez, Hopkins, and Sinclair 2010; Arceneaux and Nickerson 2010; Bankston and Burden 2023; Doherty and Adler 2014; Foos and John 2018; A. Gerber, Huber, and Fang 2018; Green and Gerber 2019; Malhotra et al. 2011; Panagopoulos 2009; Theocharis and Lowe 2016). The growing consensus not only among political scientists but also political practitioners is that these direct contact activities are among the most effective ways to affect electoral outcomes. These personalized outreach efforts are often seen as part of the "ground game" – local field operations that organize staff and volunteers to persuade and mobilize key voters. Studies of presidential campaigns have documented an increase in their fieldwork and offices beginning in the 2000s (Darr 2020; Darr and Levendusky 2014; Masket 2009; McKenna and Han 2014; Panagopoulos 2016).

As emphasized throughout this dissertation, the concentration on presidential contests to the exclusion of down-ballot races has left many assumptions about data-driven campaigning untested. Scholars have proposed a link between the increasing availability of individual-level data and the increase in direct contact activities. Moving from geodemographic targeting of precincts and neighborhoods to microtargeting voters based on individual-level records, especially following HAVA's mandate in 2006, many note that campaigns have come to rely increasingly on voter records and propensity scores to segment the electorate into targets of mobilization and persuasion for their mailers, door knocks, and phone calls (e.g., Hersh 2015; Issenberg 2013; Kreiss 2016; Nielsen 2012). Few studies, however, have investigated the effect of data sources on campaign behavior directly. For instance, Endres and Kelly (2018) find that the shift toward voter-level propensity scores made it less likely that Republicans interact with young voters. Similarly, Endres

(2020) details how the Romney 2012 campaign's persuasion targeting had some success when contacting cross-pressured Democratic voters. Several other studies have investigated the increase in presidential ground game activities under the assumption that the rise of large voter databases explains their uptick since the early 2000s (Darr 2020; Darr and Levendusky 2014; Masket 2009; McKenna and Han 2014; Panagopoulos 2016).

Drawing on this limited scholarship and the findings in my previous chapters, I center my expectations around the impact of data given the capabilities and affordances of each mode of voter outreach. In my interviews, practitioners spoke to me about a "reach versus precision" tradeoff where each medium's message and the number of targeted messages must be developed in consideration of the costs and the capacity to deliver personalized messages. In a similar sense, Burton and colleagues (2015), in their book bridging political science and consulting, describe this tradeoff in terms of "efficiency" versus "coverage." Efficient voter outreach is when campaigns successfully contact their intended mobilization and persuasion targets. They communicate with the voters most likely to maximize their return on investment. Coverage, by contrast, refers to how many potential voters campaigns contact out of the total number of targets available. These two theoretical metrics are often at odds with one another. For instance, campaigns can increase coverage and communicate with a large swath of the electorate through expensive mass media buys, but those communications will be less efficient than targeted direct appeals because they will reach intended and unintended targets (including many voters who do not even reside in the district).[57] As I learned in my conversations with campaign professionals, these conflicting considerations are at the forefront of outreach decision-making and guide their choices over which

[57] Advertisers and campaigns alike purchase ads in designated media markets with their own geographic boundaries. Television media markets in particular do not often align with congressional district boundaries and, in some instances, cross state lines.

sources of electoral information to prioritize.

Individual-level data should be most helpful and positively related to increased spending on direct outreach activities. These interactions with voters have the highest potential for message personalization and would be informed by voter-specific information. The most longstanding use of individual-level voter databases stretching back decades was to microtarget direct mail at different voter segments. Hillygus and Shields (2009), for instance, document efforts by George W. Bush's 2000 and 2004 presidential campaigns to target persuadable and swing voters with customized mailers. As discussed in chapter one, this practice dates back much further and even led Republicans to create a national voter registration database in the 1990s about a decade before their Democratic counterparts. These individual-level records have also fueled field organizing efforts. Nielsen (2012) describes how Democrats developed a unified approach to targeting centered around their shared VAN platform to assist their ground game efforts defined by using "people as media" to deliver tailored phone calls and door knocks.

Large voter databases should be less relevant to digital media purchases. As mentioned in chapter one, campaigns can use digital platforms such as Google, Facebook, or YouTube to help target their digital communications. While digital platforms allow for highly customized audiences and messaging (Kim et al. 2018), the digital data on these platforms are inaccessible to political campaigns to use for their own purposes. Instead, campaigns have restricted access to digital data only within the online platforms in the form of menu-style targeting applications. For example, campaigns can specify certain demographics they want to target and have access to other targeting specifying interests and behaviors. One practitioner I interviewed described these platforms as "walled gardens" because of their closed system of targeting and advertising. For instance, during the latter half of my investigation period between 2006 and 2018, Facebook permitted targeting

ads at specific voters based on lists of their names and emails but only provided campaigns with limited feedback on the number of individuals successfully targeted.[58] Beyond this limited integration, the digital data on prominent platforms are almost entirely disjointed from the individual-level data records based on state voter registration rolls and compiled into large databases. Although gaining much news attention, the 2016 Cambridge Analytica scandal in which the namesake firm accessed Facebook's digital data was an anomaly rather than a norm. Most campaigns only make limited use of the digital data in the form of user interactions generated by customer relationship management (CRM) systems for their email marketing activities.

Lists of voters and predictive individual-level scores are least useful to campaigns' mass media purchases. Advertising on television and radio or in print is almost entirely done with aggregate demographic information such as the gender, racial, or income makeup of viewers, listeners, or readers. The high price of production in most cases also precludes most campaigns from developing multiple different ads targeted at different demographic segments. Instead, the most valuable information for crafting these mass media appeals comes from the results of polls. Campaigns with enough money to produce ads and buy time slots should be more likely to invest in large-N surveys that can be analyzed in terms of the same sociodemographic factors that are available to them when making media buys. As discussed in chapter two, campaigns also commonly conduct longer-length brushfire polls prior to purchasing mass media to help them test potential messaging or better understand how electoral segments respond to different issues positions.

Taking all these factors into consideration, I expect that campaign-level investment into

[58] https://developers.facebook.com/docs/facebook-business-extension/fbe/get-started/custom-audience-onboarding

individual-level data should be correlated with their outreach purchasing strategies. A lack of correspondence between data purchasing patterns and outreach strategy would suggest that the influx of individual-level data over the past decades has done nothing to alter campaign-level determinations of how to reach out to voters. Alternatively, if the data sources campaigns purchase explains at least some of the variation in outreach strategy even after accounting for other strategic considerations, as many scholars have assumed, then the results would suggest that sources of information have empirical consequences for how campaigns interact with voters.

## 5.2 Outreach Spending Patterns

Continuing to leverage FEC spending records, this section turns to document patterns of outreach spending before modeling the effect of electoral information sources in the next. In contrast to studies of presidential campaign outreach, I find that congressional campaigns are remarkably consistent in their outreach preferences and employ similar strategies across my entire period between 2006 and 2018. While the previous chapter uncovered persistent party-level differences in data purchasing patterns, differences in congressional candidate spending on outreach do not appear to be divided along party lines. Republicans spend more than Democrats on their outreach efforts, but both parties' candidates do not have a strong preference for one mode over the other.

Appendix B details the parallel FEC records refinement procedure to produce verified outreach expenditures. In short, I reviewed tens of thousands of unique expenditures and placed those related to outreach into the appropriate categories of mass media, digital media, or direct contact. The bulk of mass media expenditures that campaigns reported are related to television ad buys and advertising consulting and production. Most digital media expenditures are either paid digital advertising or digital media consulting. Finally, the most common direct contact activity is direct

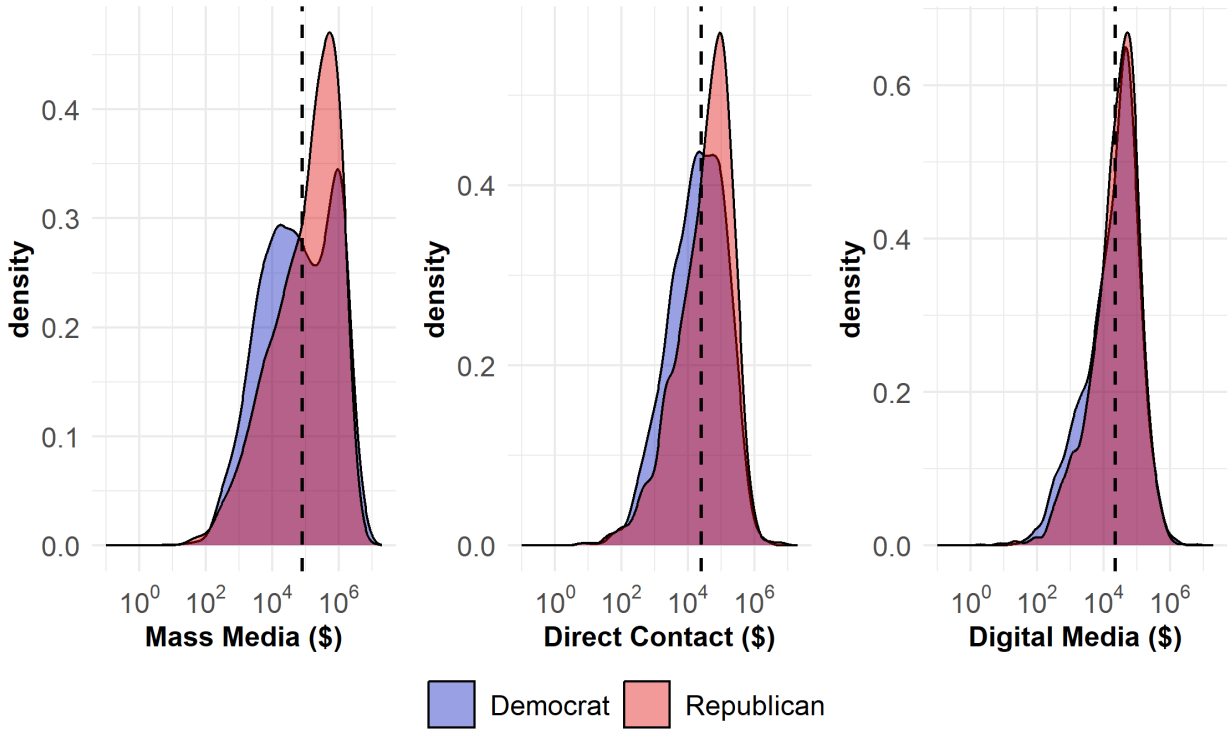
mail followed by phone calls or texts and the purchase of other supporting outreach and canvassing materials material, such as leaflets and literature.

Congressional candidates in both parties spent similar amounts of money on voter outreach between 2006 and 2018 with most of their budgets going toward mass media advertising. Figure 5.1 reports the distributions of campaign-level spending on different modes of outreach in dollars split up by party. I transform the horizontal spending axis to a log base 10 scale to better represent high-spending outliers. Campaigns spend more of their limited budgets on mass media advertising compared to other forms of outreach. As seen in the figure, the mass media spending distributions and its overall median (indicated by the dashed line) in the left panel of Figure 5.1 are shifted further right along the horizontal axis compared to the direct contact and digital media spending distributions in the two other panels.

Examining these campaign-level dollar spending amounts also reveals some marginal partisan differences. Republican congressional campaigns spend slightly more on mass media and direct contact compared to Democrats. Comporting with past research, the largest difference in the distributions is mass media spending where Republican campaigns spend more on average for their television and other mass media buys (Martin and Peskowitz 2018). The Republican distribution for direct contact is also shifted slightly right along the horizontal axis indicating that Republican congressional campaigns tend to spend more on direct contact. This difference is the result of Republican direct mail efforts, typically comprising a large portion of their direct contact expenditures, while Democrats' largest expenditure is on canvassing and field materials. Appendix E reports the distributions of spending split up by different types of direct contact. Lastly, the party distributions of campaign-level spending on digital media are nearly identical, reflecting the fact that campaigns commonly make purchases of digital ads themselves on Internet platforms rather

than relying on a partisan consultant.

**Figure 5.1: Campaign-level Spending on Outreach in Dollars**

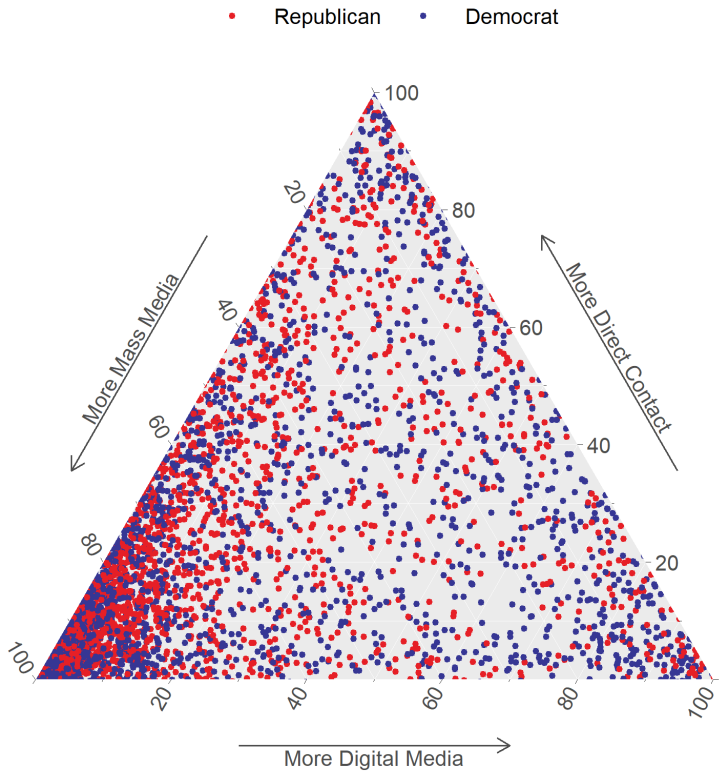



In line with the analyses in previous chapters, the goal of this chapter is to reveal the preferences of different campaigns across different electoral contexts. While I descriptively report the campaign level-spending estimates on different types of voter outreach, I contend that these are poor indicators of campaign preferences because of pricing variation across campaigns, districts, and cycles. As the case was with the data markets discussed in chapter three, the cost of outreach and related consulting also varies across parties with Democratic and Republican consultants charging different rates (Martin and Peskowitz 2018). As a result, I choose to examine the proportions of spending on each mode of outreach out of a campaign's total spending on outreach activities as an indicator of campaign-level preference.

Examining spending proportions reveals that campaigns again spend the bulk of their budgets

on mass media but, unlike dollar amounts, display little difference in the voter outreach strategies employed by Republican and Democratic congressional campaigns. Figure 5.2 depicts campaign-level outreach on mass media, digital media, and direct contact spending split up by party in a ternary plot across my entire period between 2006 and 2018.

**Figure 5.2 Outreach Spending Patterns by Party**

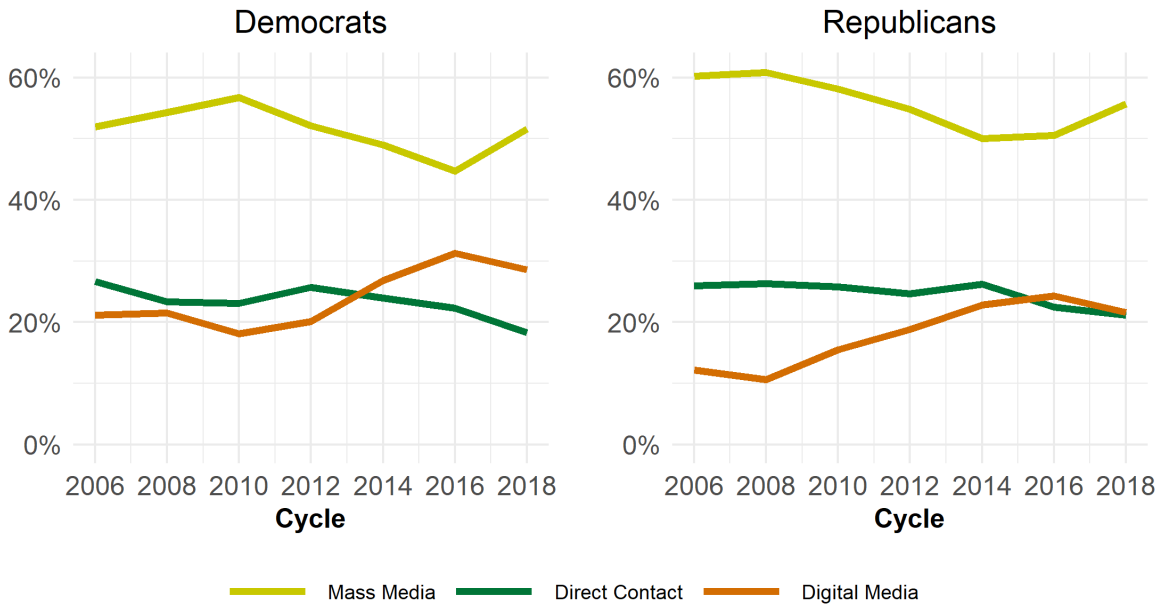



Observations closer to a corner indicate more spending on that mode. The three shares along the sides of the triangle add up to 100%. Points closer to the center indicate mixed-mode spending across all three sources. Observations along the side between the two outreach modes indicate only spending on those modes. The clustering of observations toward the bottom left mass media corner reveals that more campaigns choose to invest in mass appeals. Fewer campaigns will choose to only invest in direct contact or digital media. Based on the distribution of observation alone,

Republican and Democratic candidates do not appear to exhibit different outreach spending patterns with both parties having candidates who pay for a mix of outreach across mass media, digital media, and direct contact.

Over time, Republicans and Democrats exhibit similar trends, and congressional campaigns have made few changes to their electoral communication strategies between 2006 and 2018. Figure 5.3 plots the mean outreach spending in percentages over time again split up by party. Colors indicate outreach mode. Yellow, green, and orange represent mass media, direct contact, and digital media, respectively. Republican candidates spent marginally more of their outreach budgets on mass media advertising by just a few percentage points. Since 2006, congressional candidates from both parties have reduced their mass media advertising outlays slightly. Average direct contact spending patterns also appear relatively constant over time with approximately equal allocation across the parties. The largest difference appears in the percentage of spending on digital media for the typical Republican and Democrat. While increasing for both parties, Democrats allocate larger portions of their outreach budgets toward digital advertising over traditional mass media advertising. Regardless of these subtle differences, spending on digital media has made up a larger portion of the outreach budget for candidates of both parties. The figure suggests that this increased spending on digital media was because of a reduction in their preference for mass media more so than a shift away from direct contact efforts. Campaigns have shifted their dollars toward affordable and easy-to-access digital media platforms, especially Facebook and Google.

**Figure 5.3 Campaign-level Outreach Spending in Percentages over Time, 2006-18**




The persistence of spending across modes of outreach suggests that the dynamics affecting presidential campaigns and leading to an increase in their ground-game efforts may not be present further down the ballot. Presidential campaigns during this time frame increased their number of field offices and staff to engage with voters directly through retail politics staples like live phone calls and neighborhood canvassing. By contrast, a holistic examination of outreach spending patterns indicates that congressional campaigns kept constant in their efforts to produce customized appeals for voters even with the increasing availability of individual data to help with targeting. One probable explanation is that House campaigns have always devoted more to ground game activities compared to top-ticket campaigns and that the expansion of individual data has not necessarily shifted their reliance but instead made their ground game activities more precise, such as targeting individual households rather than entire neighborhoods. Paired with the fact that both parties' candidates have similar outreach allocations, these descriptive findings suggest that campaign-level preference for different modes of outreach is likely the result of other strategic

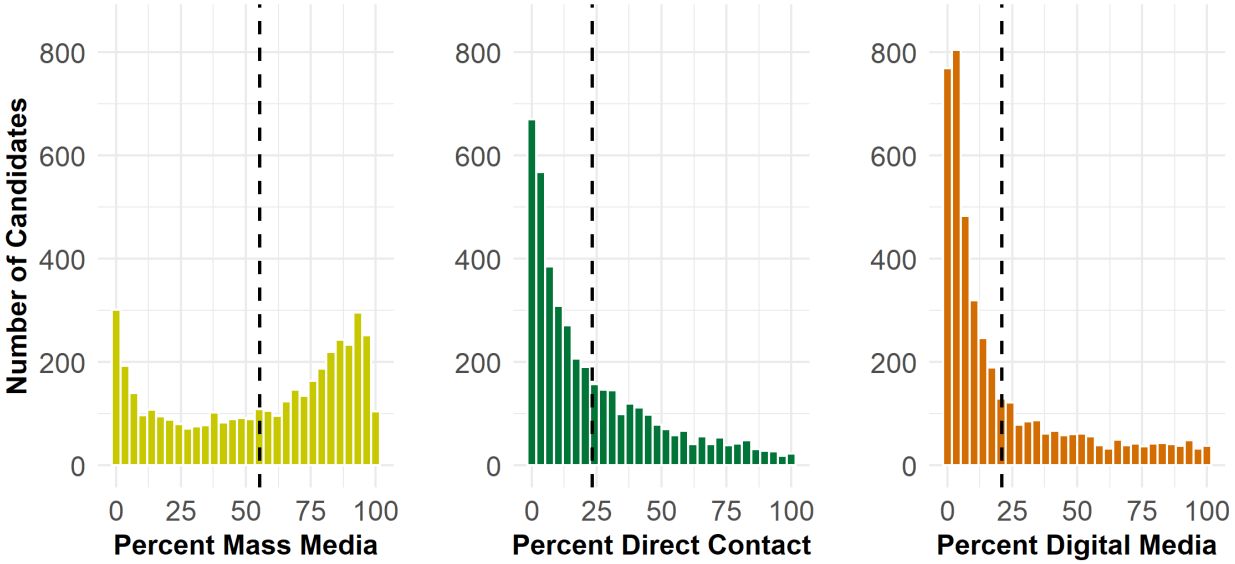
considerations in addition to the influx of voter-level records across the period.

## 5.3 Explaining Campaign Outreach

This section presents the results of carefully selected regression models to determine if sources of information about voters purchased by campaigns impact subsequent congressional campaign outreach strategy. I find that data purchasing patterns correlate with how campaigns choose to allocate their budgets across different modes of outreach. After accounting for a host of other considerations, campaigns that spend more of their limited data budgets on individual-level data will purchase fewer mass media ads and increase their direct contact efforts and, to a lesser extent, also increase their presence on digital media. While campaigns also change their outreach strategy depending on their overall spending, incumbency status, and competition, these findings emphasize that the sources of voter information that campaigns possess are consequential.

To model outreach spending patterns, I again draw on modeling techniques that consider the properties and distributions of measuring campaign preference for outreach as the proportion of spending on each mode out of total outreach spending. As explained in the previous chapter, most common regression techniques are inadequate for bounded proportions. The choice of a model requires not only attention to the dependent variables' bounds but also to their distributions. A visual inspection of the distributions reveals that the outreach modes are zero-inflated. Figure 5.4 plots my dependent variables transformed into percentages. Mass media spending is more equally distributed with fewer campaigns choosing to invest only in mass appeals, but a significant number do not purchase any. Many more campaigns do not report investing in any direct contact or digital media leading to the positive skew of the distributions.

**Figure 5.4 Campaign-level Outreach Spending Distributions**




Given their distribution, I use a zero-inflated beta regression to account for the large number of campaigns with no spending on outreach modes. The zero-inflated beta regression has the same properties as the ZOIB regression explained in chapter four, but the model has three components rather than four. Specifically, the model is reparametrized to be defined by the $\mu$ mean and $\sigma$ variance components of the beta distribution, while $\nu$ captures the probability of no investment into an outreach mode. The $\tau$ one-inflated component present in the previous chapter's models is excluded.[59] For interpretation purposes, the zero-inflated $\nu$ component represents a campaign's initial decision to invest in either mass media, direct contact, or digital media. The variables for $\mu$ mean and $\sigma$ variance components, on the other hand, explain variation after a campaign has decided to invest in a mixed-mode outreach strategy, spending part of their budgets on at least two out of the three modes.

The three model components are estimated as a function of individual-level data spending,

[59] Observations at the upper bound are transformed to approach 1 and are included in the estimation of μ and σ components.

financial resources, outside spending, competition, candidate characteristics, and district-level factors with cycle and state fixed effects. Individual-level data spending is measured as the percentage of individual-level data spending out of data-related spending.[60] Descriptions of the other variables are in the previous chapter's Table 4.1. I also add a variable that measures the extent to which television media markets overlap with each congressional district. Congressional district boundaries sometimes overlap with television media markets but can also cut across two or more. Candidates running in districts with higher market overlap purchase more television advertising because their constituencies will be more likely to view the ads (Campbell, Alford, and Henry 1984; Schaffner 2006). I created a variable that measures the proportion of the electorate contained within a given media market. Measured between zero and one, higher values indicate that the district population is spread across more media markets (Branton, Perkins, and Pettey 2019; Herrnson and Gimpel 1995).[61]

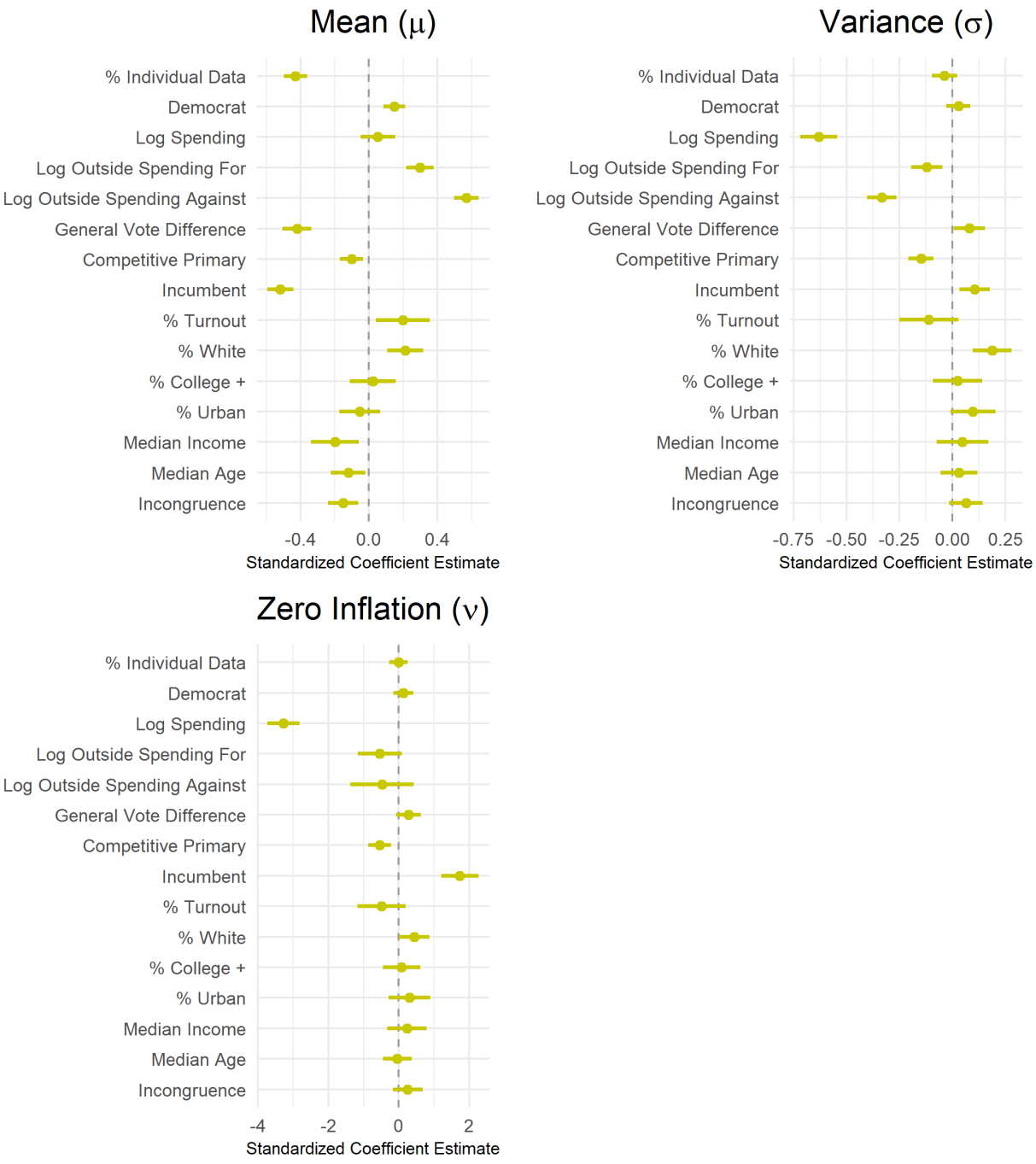
Figure 5.5 reports standardized effect estimates for the three components of the mass media zero-inflated beta regression model. I standardized the effect estimates by dividing them by two standard deviations, permitting direct visual comparison of their relative impact (Gelman 2008). Examining the effect of individual-level data across all three model components reveals that a campaign-level preference for large voter databases correlates with reduced spending on mass media for the mean ($\mu$) component but has no effect on the variance ($\sigma$) or zero-inflated ($\nu$)

[60] I do not include a measure of polling spending portion because the two variables are by nature of their measurement highly correlated and additive inverses of one another except among campaigns that spend nothing on both sources of electoral information. The zero term in the individual-level percentage measure captures campaigns that only invest in polls or spend nothing on either source. Additionally, because campaigns are more likely to invest in individual-level data over polling, the measure better captures variation in campaign-level data spending patterns.

[61] Specifically, incongruence is calculated with the following equation: $1 - \sum_{i=1}^{n}(p_i)^2$ where $n$ is the number of media markets a district contains and $p_i$ the proportion of the electorate in the media market.

components.

**Figure 5.5 Explaining Outreach Spending on Mass Media**



Whether campaigns choose to purchase mass media and the variation in their outreach allocation is largely a function of other strategic considerations. After having decided to spend their limited budgets to advertise on mass media, campaigns that invest more in individual-level data and less into polling spend less of their outreach budgets on mass media.

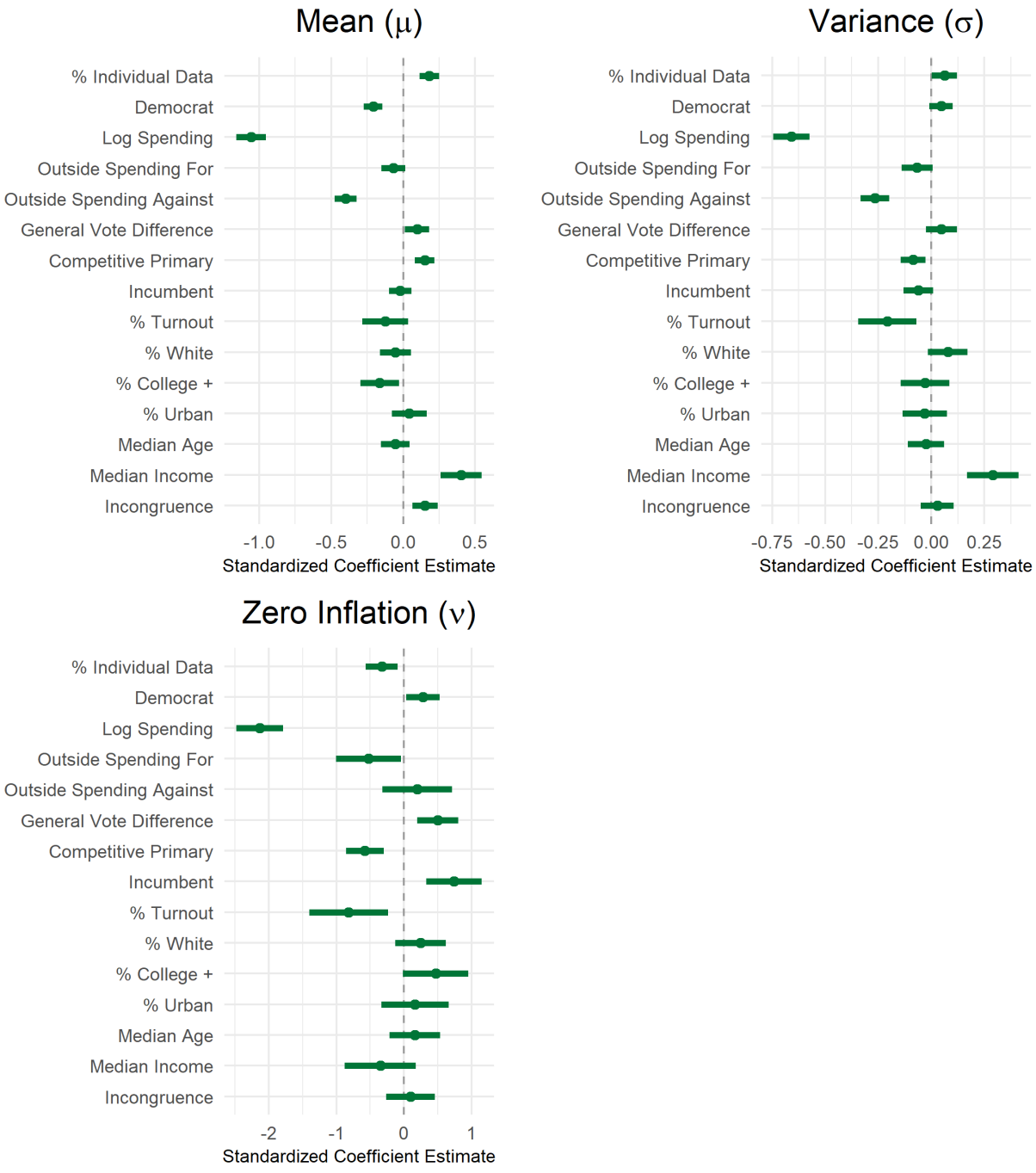
Levels of spending and competition along with incumbency are the primary factors affecting spending on mass media, with some responsiveness to district-level characteristics also apparent. The biggest consideration for whether to invest in mass media advertising is how much money a campaign has to spend, as indicated by the large negative standardized effect in the $\nu$ component. Conversely, once a campaign has decided to purchase advertising, financial resources have no discernable effect on its outreach allocation strategy. Instead, campaigns respond to levels of competition and outside spending when choosing to invest more in mass advertising like television. Incumbents are less likely to purchase any mass media ads, as indicated by the positive relationship between the incumbency variable and spending zero on mass media in the $\nu$ component. Even after deciding to purchase an ad, incumbents will still spend less on mass appeals. Lastly, district-level demographics and media market congruency also appear to impact their mass media outreach strategy. Candidates running in districts spread over a larger number of media markets are less likely to purchase advertising.

Unlike the descriptive results in the previous sections, a candidate's party affiliation appears to impact outreach spending patterns only marginally after accounting for other strategic considerations. Despite their consistent preference for individual-level data as shown in chapter four, Democrats invest more of their budgets on average into mass media advertising. Note, however, that party does not appear to be a determinant of whether campaigns make an initial investment into mass media advertising as seen in the $\nu$ component. Appendix Figure E2 reports

the model components re-estimated to include interactions between party, spending, and individual-level data variables. Including the interaction terms reverses the main effect of party. Democrats on average will spend less of their budgets on mass media appeals. Only among Democratic campaigns with higher levels of spending is there a discernable preference for mass media advertising. Additionally, the effect of individual-level data is not conditional on party or spending. Both two-way interactions between individual-level data and party and spending along with their three-way interaction term do not reach statistical significance. The observed relationships are also robust to measuring competitiveness as the Democratic advantage based on the most recent Democratic presidential candidate's vote share in the district, found in appendix Figure E3.

The relationship between data and direct contact spending patterns comports with the results of the mass media model. Figure 5.6 reports the portion of spending on direct contact regressed on the same set of explanatory variables for the three zero-inflated beta regression components. As seen in both the mean ($\mu$) and zero-inflation ($\nu$) components, higher levels of investment into individual-level data make it less likely a campaign will go without direct contact and that they will increase their investment into direct contact over the other two modes. Variation in direct contact spending patterns also increases slightly as campaigns purchase individual-level data. Note that this positive relationship between individual data and spending on direct contact activities is not conditional on party, as reported in the interactions found in appendix Figure E4. Figure E5 with the alternative measure of competitiveness based on Democratic advantage according to presidential vote share produces parallel results.

**Figure 5.6 Explaining Outreach Spending on Direct Contact**

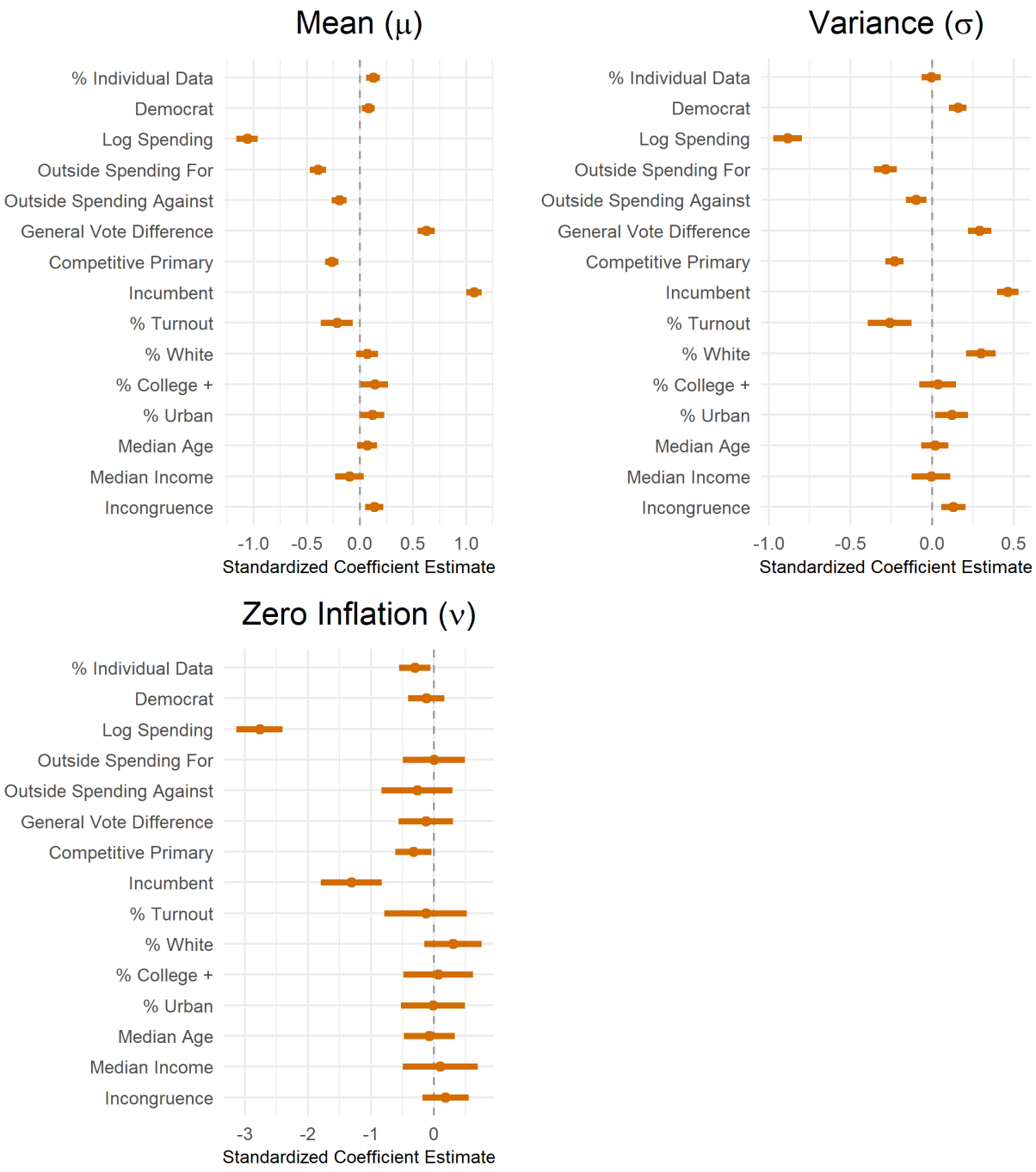


While individual-level data purchases have a consistent effect across all three model components, the largest relative determinant is the overall level of campaign spending. There is less responsiveness to incumbency and district-level characteristics than in the mass media model. Campaigns with more money are less likely to rely solely on direct contact activities to communicate with voters because they will choose to purchase mass media advertising. As levels of spending increase, campaigns also reduce the portion of their outreach budgets devoted to direct contact. Incumbency is a deciding factor on whether to invest in any direct contact. Like mass media purchases, incumbents are more likely not to purchase any direct contact as indicated by the negative effect estimate in the $\nu$ component. Fewer district-related variables appear to determine direct contact budget allocation, although districts with higher levels of media market incongruency correspond with increased spending on direct contact.

Partisan patterns in the direct contact model components also agree with the mass media model. According to Figure 5.6, Democrats are marginally less likely to invest in direct contact activities and spend less of their outreach budgets on these direct appeals to voters. As with the mass media model, including interaction terms between party, spending, and individual data (reported in appendix Table E3) reverses this observed relationship with only Democratic campaigns at higher levels of spending having less of a preference for direct contact activities. Yet interactions between individual-level data and party appear again to be near or statistically indiscernible from zero.

The final model explaining the proportion of outreach spending campaigns devoted to digital media is similar to the previous direct contact model, although the relative impact of individual-level data is reduced. Figure 5.7 reports the effect estimates for digital media budget allocation. Campaigns that invest more in individual-level data are more likely to make digital advertising purchases ($\nu$) and prefer to purchase more digital advertising in their mixed-mode strategy ($\mu$).

**Figure 5.7 Explaining Outreach Spending on Digital Media**



The effect of individual-level data, however, is much smaller than in the previous models, and campaign-level spending explains more variation in digital media proportions, indicated by the variable's large relative standardized effect across all three components. Appendix Figure E6 reporting estimated model components with party, spending, and individual data interactions again suggests no conditional effects of electoral information on outreach strategy. The findings are again consistent with the alternative Democratic advantage measure of competitiveness, reported in Figure E7.

As with the other two mode models, an initial investment into digital media is primarily the product of available funding and incumbency. The $\nu$ component suggests campaigns with lower levels of spending will not invest in digital media and that incumbents are less likely to make a digital investment compared to challengers or candidates running in open-seat contests. Once a campaign chooses to make digital media part of its outreach strategy, its relative allocation is a function of not only spending and incumbency but to a smaller degree levels of competition as with mass media and direct contact. Both measures of competition indicate that candidates in less competitive primary and general contests make digital media a larger portion of their voter communication strategy. Like when choosing to invest in direct contact, candidates running in districts spread across many more mass media districts are also more likely to turn to digital media, as seen by the positive effect estimate of incongruence in the $\mu$ component. Other district characteristics are less relevant to a campaign's digital prioritization having smaller effects size near or at zero, although some help to explain variation in the $\sigma$ variance component.

Across all three outreach mode models, campaign data preferences correspond with their choice of voter outreach strategy. Campaigns that spend more on individual-level data rather than polls are more likely to engage in direct and digital voter communications, while those who

prioritize polls tend to spend more on reaching out to the electorate with mass appeals. The impact of data purchasing patterns, unlike findings related to their acquisition in chapter four, does not appear to be dependent on a candidate's party affiliation. Instead, the prominent factors that determine a campaign's patterns of outreach are classic considerations of financial resources, incumbency advantages, and their opponent's electoral viability.

## 5.4 Summarizing the Impact of Data on Outreach

Scholars have long assumed that the rise of large voter databases caused a shift in campaign strategy from being oriented around broad appeals for mass audiences to tailored messages delivered through customizable communication channels. This chapter provides a more nuanced view. While patterns of data source acquisition correlate with how campaigns choose to allocate their outreach across different modes of outreach, the influx of voter records has not led to the wholesale reorientation of campaign communication strategies. Congressional campaigns still have a clear preference for mass media advertising, and few campaigns choose to communicate with voters solely at an individual level with personalized messages either through direct contact or digital media. They must consider their financial resources, spending by outside groups, electoral threat, incumbency advantage, and district-level characteristics while also deciding who in the electorate they will contact and how they will contact them.

Campaigns cannot reach all the voters they want solely relying on individual-level data and direct or digital contact. In an idealized voter information and outreach environment, every campaign message would be perfectly tailored for each voter given their political predispositions. Campaigns would have complete coverage and reach every single target they need to win an

election with the coalition that gives them more votes than their competitors. Their messages would also have perfect efficiency. They would only mobilize the voters who already support their candidate but need to be nudged to turn out. Their persuasion messages would be perfectly targeted at true swing voters with customized messages empirically tested beforehand to move the needle. The realities of data-driven campaigning could not be more divorced from this ideal. Campaigns operate with high levels of uncertainty even in an information environment that provides more data on voters than ever. As highlighted in my conversation with political practitioners, many cast doubt on the idea that voter data has fundamentally changed politics. Individual voter data can help implement contact efforts more efficiently, but they do not overwrite fundamental strategic considerations or communication limitations.

Considering my dissertation's findings holistically up to this point, I argue that campaign-level preferences for both sources of information and modes of electoral communication are the result of longstanding strategic considerations as much as they are influenced by having individual-level data available to target their outreach efforts more efficiently. What large voter databases, statistical models, and voter propensity scores have done is to strengthen preexisting conceptualizations among campaign professionals of the electorate being easily divisible into distinct segments for mobilization and persuasion. Especially among Democratic practitioners, the reduction of voters to two simple scores based on their predicted levels of turnout and support is not only routine but institutionalized through their shared VAN platform and a party culture that reinforces it. Republicans meanwhile might lack the data sharing practices that have enabled Democrats to establish a unified approach to data-driven campaigning but their methods of combining sources of individual-level data nevertheless still reduce voters to but a few simple scores that are then informally calculated to make decisions over which forms of outreach to

purchase.

In the final chapter, I combine the insights of my interviews with documented data and outreach preferences to argue that data-driven decision-making for contemporary campaigns is the result of durable party-level differences in data orientations and longstanding strategic considerations.

# 6 Developing a Picture of Voter Data and Campaign Strategy

This dissertation began with a motivating example from Wisconsin's first congressional district in 2018. Long-shot Democrat Randy Bryce faced off against establishment-embraced Republican Bryan Steil in an open-seat contest. Bryce's campaign was invigorated by a midterm election seen as a referendum on President Donald Trump's party, and Democratic donors big and small from around the country infused his campaign with outside money. Flush with cash, the union ironworker made a concerted effort to engage with and persuade the Republican-leaning district to send a Democrat to Washington for the first time in over 20 years. To support his plan, Bryce spent over $225,000 of his campaign funds on surveys of the district and approximately $75,000 to gain access to individual-level records on the VAN platform shared by Democratic campaigns around the country. His opponent Steil meanwhile spent much less on polls at around $60,000 and next to nothing at only $4,200 on access to individual voter records. Their differences in data spending appeared to translate to differences in their approaches to communicating with the electorate. While both spent the bulk of their budgets on mass media advertising, Bryce spent more than two times more across all modes of outreach and relatively more on digital media and direct contact efforts. Bryce's mixed-mode efforts would prove futile with his Republican opponent winning by over ten points. It is unlikely that any amount of cash or data on the electorate could have overwritten the fundamental Republican advantage in a district previously held by former Speaker of the House Paul Ryan.

To unpack the spending patterns in this preliminary example, my research began in chapter two with qualitative interviews of fourteen political professionals to understand how campaigns make sense of different sources of electoral information. From them, I learned that practitioners in

both parties conceptualize the electorate as comprising strategic mobilization and persuasion segments and that contemporary databases, with their millions of rows of voters and thousands of columns of characteristics, are integral to this strategic division of the electorate. While Republicans and Democrats share a similar hypothetical view of voters, they differ in their practical application of electoral information sources and the importance they collectively place on large voter databases.

Democratic practitioners are unified in their approach to data-driven campaigning because they not only share access to a single nationwide voter data platform but, more importantly, have a data culture that reinforces its dominance. Nearly every Democratic campaign and progressive cause throughout the country pays for access to the VAN. Staffers routinely segment the electorate based on propensity scores that reduce voters to just a few numbers. These voter scores – generated by specialized political data analysts who typically do not work on the campaign – combine the inferences of polls and big data to predict each registered voter in the country's individual chances of supporting a candidate and turning out in an election. Democratic campaign staffers and volunteers nationwide are trained to access the VAN and generate lists of voters based on these two scores to target their outreach activities. The VAN is fundamental to Democrats' perceptions of the electorate. How they define voter data, view voters, and target their outreach activities are inseparable from their communal database.

On the other side, the Republican party not only lacks a common data platform but also a cohesive strategy for combining and applying different sources of electoral information. Republican consultants compete with one another to sell campaigns on the quality of their voter data and different statistical models. They have diverse approaches to integrating poll results with voter-level data to assist campaigns in segmenting the electorate. Their methods range from

technical statistical models to ad hoc decision-making over how to interpret polls and combine them with different databases. Rather than being trained on one data platform, Republican candidates and staffers navigate a party data environment defined by many different data vendors and disparate lists of voters. They make list-by-list purchases and poll-by-poll decisions. Their inferences about the electorate come much more from their individual political intuitions and experiences rather than mutual assumptions about the value of big data.

With the findings of my interviews in mind, the rest of this dissertation analyzed millions of cleaned and verified expenditure records reported by thousands of congressional campaigns between 2006 and 2018 to provide a complete picture of data-driven campaigning. The expenditure records covered a period of marked change and innovation in large voter databases. Its start in 2006 corresponded with HAVA requiring states to provide computerized statewide lists of registered voters and the initial rollout of the Democratic party's VAN platform. The period also overlapped with the Obama presidential campaigns of 2008 and 2012. His campaigns spent millions investing in data and technology. Not only did many celebrate his campaigns as the epitome of data-driven campaigning, but his efforts helped to train a new generation of Democratic consultants and activists that would go on to institutionalize many big data practices for campaigns further down the ballot in subsequent cycles.

Starting in chapter three, my analysis of approximately 50,000 data-related spending records further elaborated on why the two parties have different approaches to data-driven politics. Republican and Democratic congressional campaigns first operate in dissimilar partisan markets for polls and individual-level data. Republican congressional candidates purchase surveys on the electorate and access to large voter databases in marketplaces that have levels of competition comparable to other free economic markets and industries. Unlike their Democratic counterparts,

the RNC's choice to only collect and sell a national voter list rather than contract with a single company led to a marketplace with many competing individual-level data vendors and rival voter targeting and campaign management platforms. Even the largest provider of individual-level data to Republican campaigns, Aristotle International, advertises itself as nonpartisan and maintains a separate national voter file and data platform. By contrast, Democrats running for the House have coalesced around one database platform and a few large polling firms. The NGP VAN and its namesake VAN platform together are a near monopoly within the market for individual data. It is the default option and, in most instances, the only choice for Democratic campaigns across the country, especially after 2014.

The end of chapter three and the entirety of chapter four revealed that the two parties' different data approaches and marketplaces are consequential for the data purchasing patterns of congressional campaigns. Even at the beginning of my period of investigation in 2006, Democratic congressional candidates were prioritizing individual-level data much more than Republicans who, by 2018, still spent greater portions of their data budgets on polls over access to voter databases. In contrast, Democratic candidates were spending more than half of their data budgets on voter-level records as early as 2014, the cycle immediately following Obama's 2012 presidential campaign and eight years after VAN's official party contract. Chapter four found that these partisan differences in campaign-level data continued to be a point of separation even when considering the impact of other strategic campaign considerations. While campaigns respond to ever-present concerns about their financial resources, outside spending from other groups, and the levels of competition they face from challengers, my carefully selected statistical models indicated that a candidate's party affiliation remains the largest determinant of the sources of information congressional campaigns purchase to learn about the electorate and execute their voter outreach

strategies.

The fifth chapter further emphasized how campaigns still attend to classic strategic concerns even in an electoral information environment that provides more data on voters than ever. I closely examined the influence of data spending and other strategic considerations on campaign-level preferences for communicating with the electorate. Both parties' congressional campaigns have remained surprisingly consistent in how they allocate their voter outreach budgets since 2006. Overall, Republican and Democratic campaigns have spent marginally more on digital media, but their patterns of outreach still indicate a preference for mass media, particularly television advertising. Campaigns also maintained similar levels of direct contact such as canvassing, phone-banking, and direct mail throughout my period of investigation. While campaign spending on individual-level data corresponds with more direct contact activities, a candidate's party is not a crucial factor influencing their mix of outreach activities. Instead, campaigns determine their voter outreach strategy based on strategic considerations that predate data – namely their financial resources, levels of outside spending and competition, and the size and shape of their district.

By providing the first systematic account of data and outreach-related spending from a large number of candidates, I hope that future scholars will have more descriptive evidence to inform their research. My approach to connecting real-world experiences and implications with robust analysis, once common in the political science literature (e.g., Fenno 1978; Mayhew 1974; Miller and Stokes 1963), has over the past few decades become increasingly rare. Many scholars, especially those studying American politics, have shifted their attention away from thick descriptions of broad political phenomena toward causally identifiable questions and theories tested with field, lab, survey, natural, and quasi-experiments (Teele 2014). Given the dearth of previous scholarship on how campaigns below the level of the presidency engage with sources of

electoral information, my project took a step back in the social scientific process to provide a descriptive account of the electoral information environment as it appears for most campaigns. Much is left to learn about data-driven campaigning further down the ballot, but now scholars have a much better picture to generate new questions and conclusions.

## 6.1 Reducing Voters to Electoral Math

This dissertation's main finding highlights how Republican and Democratic congressional campaigns operate in different data environments yet communicate with the electorate in similar ways. On their face, these claims appear to be at odds with one another. If Democrats can easily access (and likely have better) individual data, why would they choose to spend most of their money on untargeted mass media appeals where their big data on voters is largely useless? On the other hand, why would Republicans pursue similar levels of direct contact if they ostensibly must work harder to gather individual-level data and prefer the big-picture summaries coming from polling anyway? The answers to these questions lie in the realities of both campaigning and data. Ultimately, I argue that voter data have not upended politics but rather intensified strategic tendencies.

Campaigns have never communicated with every single voter about every single issue. Every election contest (past and present) starts with basic considerations about a candidate's potential number of supporters, how many possible toss-up votes exist, and the percentage of the electorate that will not vote for your candidate no matter what. Most campaigns aim to win the election, which often means winning one vote more than 50 percent. As emphasized in my interviews, these basic vote tallies of broad groups in the electorate are central to how political practitioners in both parties conceptualize, target, and reach out to voters. Campaigns want to estimate their number of

likely supporters and the number of potentially persuadable voters compared to the total number of votes they need to win the election. They then determine who in the electorate to contact, defined in terms of mobilization and persuasion, before calculating their needed vote totals.

If campaigns could, they would contact every supportive voter who potentially may not turn out and every persuadable voter who may vote for another candidate. They would create a customized message tailored to each individual voter's preferences that maximizes their return on investment in the form of an additional vote for their candidate. But campaigns can never reach every voter they want to in a tailored manner. Outside of small local contests, the electorate is too large, and campaign resources are too limited. Campaigns simply do not have enough money, staff, volunteers, or time to communicate with every voter even if they have access to thousands of individual data points. The reason my findings indicate that congressional campaigns continue to spend the bulk of their limited budgets on mass media like television is because they need a way to reach a large portion of the nearly three-quarters of a million voters living in most districts. Congressional campaigns also have increasingly turned to digital media, in part, because of the massive number of users on major platforms and the affordability of digital advertising. All the individual-level data in the world cannot give campaigns the additional resources and expertise they would need to provide each identified mobilizable or persuadable voter with a customized message intended for just that voter.

Voter communication strategies not only depend on a campaign's available resources but also on the context of the race. For instance, incumbents are less likely than their challengers and open-seat counterparts to make mass media appeals given, among other advantages, their higher name recognition in their districts. Contemporary campaigns also do not only compete with their opponent for broadcast time. Since the *Citizens United* decision in 2010, outside groups have spent

more than a trillion dollars influencing just House races. Their spending on mass media advertising often rivals and sometimes surpasses the candidates themselves. Sometimes campaigns have the money to spend on television but must adjust their outreach strategy to emphasize digital and direct contact because their district does not neatly overlap with designated media markets. Campaigns may even choose to forgo television altogether if their district is compact and urban enough, preferring instead to communicate with voters online and in person. My findings from chapter five indicate that all these strategic factors along with others are large determinants of congressional campaigns' voter outreach strategies.

Even if campaigns had unlimited resources and did not face other strategic considerations, they still would not be able to collect enough data or generate precise enough predictions on voters for only individually targeted outreach appeals. I learned from my interviews that, even though voter files from major data vendors are thousands of columns wide, most statistical models leverage as few as ten and rarely more than two dozen variables to generate propensity scores on individual voters. The most valuable data on voters comes from publicly available information in state voter rolls like party registration, partisan primary voting, turnout history, and other basic demographic information. While databases contain hundreds and potentially thousands of additional consumer data points recording attributes like home ownership or subscriptions, their power for predicting support and turnout scores for individual voters is negligible compared to other more readily available and politically relevant variables like race, gender, age, and income. These predictive challenges, paired with the fact that voter databases do not have individualized records of voters' preferences on specific political issues, make predicting the persuadability of individual voters difficult and generating a customized appeal for each individual voter a near impossibility.

Large individual-level voter databases also have limited utility for the bulk of campaign

outreach targeting decisions. Advertising on television, for instance, involves purchasing time slots in designated media markets on specific programs and channels. Campaigns typically have only limited information on the aggregate composition of television viewing audiences, rendering most of their individual-level data useless. Digital advertising faces its own set of targeting difficulties. Both Facebook and Google allow campaigns to upload lists of voters they wish to target but provide little information on how their matching process works and only aggregate summaries of the number of voters on the list that viewed the ad. I learned from my interviews how political practitioners are often frustrated by this matching feature on major digital platforms. They explained to me the concept of a "walled garden" in digital advertising. Large technology companies only provide limited access to their digital data and require advertisers to interact with users exclusively within their platforms and with the targeting features they choose to provide. If allowed, campaigns can import some of their own data, but they cannot export the proprietary digital data contained within these closed environments. All these data difficulties involved in targeting paid advertising result in most campaigns primarily leveraging their individual-level data on voters for their direct contact activities where they have full control over their data and a record of who receives a door knock, phone call, mailer, or other similar direct appeal.

Considering all these real-world complexities helps decipher why congressional candidates' preferences for electoral information depend on their party affiliation but their outreach strategies do not. Republicans and Democrats may operate in dissimilar data environments, but the outcomes of their segmentation schemes are similar. The data they draw on may differ. Democrats are more likely to depend solely on the voter propensity scores listed in the VAN and predicting voters' likelihood of casting a ballot and supporting a Democratic candidate. The typical Republican meanwhile develops a targeting strategy by fielding a poll and examining support among different

demographic groups before seeking out an updated list of voters matching those criteria. While the composition of their segments will surely diverge, the product of their segmentation is an electorate split into a group of supportive voters the campaign thinks need to be mobilized and another group of possible swing voters that need to be persuaded. Most campaigns will not have the time nor the resources to check the accuracy of their predicted targets. Additionally, because most campaigns depend on third-party data vendors and companies, they make their targeting decisions often unaware (or despite) the quality of the underlying data. Campaigns are forced to make the decision about which voters to communicate with and which voters to ignore whether they are dividing up the electorate based on basic demographic characteristics, support and turnout scores, or even customized persuasion scores built for a specific race.

Ultimately, I argue that big data on voters have not disrupted electoral politics but rather intensified the extent to which campaigns pay attention to electoral math and strengthened their perceptions of the electorate as divisible into units for mobilization and persuasion. Before polls and individual-level data, candidates had to make best-guess determinations of their electoral chances based on the results of past elections, how many voters typically turn out in similar contests, and intuitions about how levels of support may have shifted. Contemporary campaigns are better able to quantify their intuitions about shifting support and identify the voters that will contribute the most to their electoral math victory. At their core, polls serve to segment the electorate in demographic terms based on their levels of support for a candidate. The most important survey results for segmentation tell campaigns which groups of voters support their candidate and which groups might support their candidate. Even a basic voter list with only a few demographic characteristics often provides campaigns with enough information to find those voters whom they think to be supportive or potentially supportive. Advanced databases take more

of the guesswork out of targeting decisions by providing campaigns with easily accessible propensity scores estimating whether a voter needs to be mobilized or persuaded.

Recall from my interviews that both Republicans and Democrats shared the same hypothetical views of the electorate as divisible into "buckets," but it was only some of the Democratic practitioners who expressed discomfort with their party's approach to data-driven politics. Their discomfort and even definitions of "voter data" came from their personal experiences and everyday interactions with the VAN. Most of them spent their formative years working in politics while Democratic campaigns and progressive causes nationwide embraced big data and voter outreach strategies that prioritized metrics measuring the return on investment for every direct interaction with voters. They admitted the value of collecting, analyzing, and leveraging voter-level data but expressed discontent over a Democratic party they think is too obsessed with efficiency and campaigning only for votes that generate a 50-percent-plus-one electoral outcome. In their opinion, Democratic data culture reduced voters to mere numbers, exactly how they appear in the VAN.

While big data and voter propensity scores do not apply to every campaign decision, their prominence in voter targeting gives them an outsized impact on perceptions of the electorate. Contemporary campaigns can calculate the minimum number of votes they need to win an election based on expected turnout and then immediately decide the universe of individual voters they need to communicate with to achieve their win number. They can generate a list of voters and rank them based on their turnout and support propensity to determine which voters will provide them with the highest return on investment and concentrate their communication efforts on them. My results in chapter five correlating more spending on individual-level records with increased direct contact activities suggest that the sources of electoral information campaigns have access to matter for outreach decisions.

Access to individual-level records on voters does not replace longstanding strategic considerations, but it does exacerbate how campaigns view the electorate – counts that sum to victory.

# Appendices

## **Appendix A:** Interviewer Script and Questionnaire

**Introduction** (2 mins)

Hello… [small talk]

Thanks for taking the time to do this interview…

I just wanted to start telling you a bit about myself and the purpose of this interview…

I'm a Ph.D. student in the political science department at the University of Wisconsin-Madison. I work with the Elections Research Center here. And I primarily research campaigns and their contemporary voter outreach and communication strategies.

The purpose of this interview is to help inform my research into campaign outreach and voter contact. I want to understand campaigns from their perspective and the perspective of people who work with them.

I'm just interested to learn more about your background in politics and hearing your thoughts and opinions about different aspects of campaign strategy.

I also need to say…

Participation is completely voluntary, and you can refuse to answer questions or terminate the interview at any time.

All your responses will be confidential. Your responses may be included in a future academic publication but all identifiable information about you will be removed.

I'll also be recording the audio from this interview and will later transcribe it. After I transcribe the interview, the audio files will be deleted. Again, no identifiable data will be available in the transcript or in anything I may pull from it for my research.

Do you have any questions for me before we start?

Do you consent to participate in the interview?

Great, to respect your time, I will try to keep our call under 30 minutes, so let's get started with the questions.

**Background** (6 mins)

I'd like to start by asking you a few questions about your background in politics and then talk a little bit about your current position…These questions are just meant for me to understand your level of experience, any information you provide will not be used to identify you. 9

- First… could you tell me a little bit about yourself and your background in politics?
    - (if not answered, then ask…)
    - Experience
        - How many years have you worked in politics?
        -
        - What types of organizations have you worked with in the past?
    - Partisanship
        - Have you worked mostly with Republicans, Democrats, or both sides?
            - (If the response is ambiguous…) "Would you say that the organizations you've worked with are mostly aligned with the Republican (Democratic) party?"
- Thanks… could you tell me more about your current position/role and organization?
    - (if not answered, then ask…)
        - What are your responsibilities/ day-to-day look like?
        - And where (in which state) do you do most of your work?
        - How would you describe the relative size and resources of your organization in comparison to your competitors or your peers?

**Campaign Strategy** (10 mins)

Next, I want to talk to you briefly about campaign strategy. Some of the questions will be about your organization while other questions will be about your personal views.

What your organization does and what your thoughts are about best practices may be the same or they may be different. Please just let me know if they are the same or tell me how they differ. Everyone has a different opinion; I'd just like to hear yours.

*Outreach Goals*

So… the first set of questions is about your goals when reaching out to voters…

In political science, we often divide campaign outreach goals into three possible types: mobilization, persuasion, and conversion. Mobilization is turning out supporters to vote or do some other participation activity like a financial contribution. Persuasion is where you try to convince undecided or swing voters to support your candidate or position. And conversion is trying to convert the other side.

- Do you think this is a useful way to think about campaign outreach goals? And are they as distinct as we think?

- To what extent does your organization prioritize persuasion, mobilization, or conversion? Why?
    - Do you have different opinions or priorities than your organization?
- What are the core factors that determine whether you engage in persuasion, mobilization, or conversion (e.g., timing, money, election type, candidate/issue, opponent, geography, data/modeling)? How?
    - (if not answered, then ask…)
        - How does timing affect the use of these different types of appeals and whether they are related to persuasion, mobilization, or conversion?

*Types of Appeals*

Next, I'd like to ask about different types of appeals that you can send out to voters…

So again… In political science, we often divide types of voter appeals into two categories: mass appeals – such as television or radio ads – and direct contact appeals – phone calls, text messages, mailers, canvasing, and (sometimes) even digital ads.

- Do you think this mass-versus-direct appeal distinction is a useful way to think about campaign outreach? Why or why not?
- What kind of voter appeals does your *organization* engage in and prioritize? Why?
    - Do your views differ from your organization or are they the same?
- What are the factors that determine the kinds of voter outreach your *organization* engages in? And how do they affect your choices?
    - (e.g., timing, money, election type, candidate, opponent, geography, data/modeling)
    - (if not answered, then ask…)
        - How important is timing during the election cycle when choosing between different types of appeals?
            - (if not answered, then ask…)
                - What are the most important differences between primary outreach and general-election outreach?
        - Are your views on these factors the same as your organization's?
- If you were to characterize the types of voters your *organization* often reaches out to, how would you describe them?
    - (if not answered, then ask…)
        - How important is it that the voter is registered?
        - How important is their turnout history?
        - How important is it that the voter is likely to support your candidate/issue?
    - Do you and your organization have the same opinion about which kinds of voters to reach out to?

**Voter Data** (10 mins)

Now in this final set of questions… I'd like to ask you about voter data. Again, some of these

questions will be about your organization and others about your personal views.

- What do you think of when I say, "voter data"?
- What is the most important information that voter data provide? In other words, if you were to use voter data to target appeals at voters, what information would you rely on?
    - (e.g., support scores, turnout scores, demographic, geographic)
    - (if not answered, then ask...)
        - Does it depend on the goal of the outreach – either mobilization or persuasion – or the type of outreach?
        - What are the other factors that affect how much you rely on voter data?
- To what extent do you *trust* voter data? Why?
    - (if unclear, then ask…)
    - Do you think that voter data are accurate?
- Do you consider the information available on digital media platforms, such as Facebook, to be "voter data"?
    - Do you think that the information about voters on digital media platforms is valuable? And do you trust them?
- How does your *organization* weight the relative value of voter data to inform campaign strategy relative to other signals about the electorate or even "gut feelings"?
    - (if not answered, then ask…)
        - What are these other signals from the electorate?
    - Do your personal views about the value of voter data differ from the rest of your organization?
- Does timing during the election cycle affect the relative value or emphasis that your *organization* places on voter data?
    - (if not answered, then ask…)
        - Is voter data more valuable during the primary or general election?
    - Do your personal views differ?

Some people think that voter data today are different from the past… where in the past campaigns only had geographic-level information based on precincts or neighborhoods while today campaigns have individual-level data on voters…

- Do you think this distinction between geographic-level voter data of the past and individual-level data of today is a good way to conceptualize things? Why?
    - (if not answered, then ask…)
        - Would you say that your organization relies more on geographic-level data, individual-level data, or a mixture of both? Why?

Similarly, some people think of polling data – information about the opinions of different groups in the electorate – is different from individual-level information from voter files…

- What do you think about this distinction?
- To what extent does your organization rely on polling?

To my knowledge, some organizations rely on data from other political data vendors, while

others have in-house "analytics" or "modeling" …

- Which best describes your organization?
    - (if yes, then ask…)
        - How would you characterize your organization's analytics or modeling?
        - And how many resources does your organization devote to analytics or modeling?

Some states have more data on voters depending on their public information laws. Some states have voting history, partisan registration, primary voting, and even demographic information, other states have much less information.

- (If respondent worked in different states, then ask…)
    - Have you noticed this variation in the data?
- To your knowledge, do the states that you work in provide this information? Or are you unsure?
- And are you ever skeptical of information that you see in voter files?

**Final Questions** (2 mins)

- Is there anything that I haven't asked you, that you'd like to add about campaign strategy or voter data?
- Do you have any feedback or suggestions about the questions I asked you?
- If I have any more questions in the future, would you mind if I reached out to you – perhaps via email – to ask them?
- Do you know of anyone else that may be interested in being interviewed?

**Appendix B:** Details of Expenditure Classification Procedure

The goal of the procedure is to classify House-campaign-provided expenditures as related to data or polling-related spending and outreach-related expenditures. As mentioned in chapter three, the coding process begins with filtering records for review. I selected categories created by OpenSecrets that could potentially be related to either polling and individual data or various categories of voter outreach, such as mass media advertising, digital advertising, and direct contact efforts like canvassing, phone-banking, and mail. Table B1 lists the included and excluded categories used to filter the records. I include all categories related to voter data (i.e., Unknown data & tech, Admin data & tech, Unknown data & tech, and Campaign data & tech) and polling (i.e., Polling & surveys). Additionally, I include several other categories that could contain data-related purchases or are related to voter outreach, including administration, campaign materials, events, fundraising, media, digital media, advertising, and unclassifiable or unknown expenditures. The primary excluded categories are contributions from other political entities and salaries. In total, 1,444,292 records are selected for standardization across all cycles, representing approximately 42% of the overall universe of itemized expenditures.

**Table B1: Standardized OpenSecrets-Provided Categories**

| OpenSecrets Code | Short Description | Long Description | Standardized |
|---|---|---|---|
| A00 | Misc admin | Miscellaneous administrative | Yes |
| A20 | Admin data & tech | Administrative data & technology | Yes |
| A30 | Admin event expenses | Administrative event expenses & food | No |
| A40 | Admin travel | Administrative travel & lodging | No |
| A50 | Admin consulting | Administrative consulting | Yes |
| A60 | Accounting & legal | Accountants, compliance & legal services | No |
| A70 | Rent & utilities | Rent, utilities & office expenses | No |
| C10 | Campaign materials | Campaign mailings & materials | Yes |
| C30 | Campaign events | Campaign events & activities | Yes |
| F00 | Misc fundraising | Miscellaneous fundraising | Yes |
| F10 | Fundraising mailings | Fundraising mailings & calls | Yes |

| | | | |
|---|---|---|---|
| F20 | Fundraising data | Fundraising data & technology | Yes |
| F30 | Fundraising events | Fundraising events | No |
| F40 | Fundraising fees | Fundraising fees | No |
| F50 | Fundraising consult | Fundraising consulting | Yes |
| M00 | Misc media | Miscellaneous media | Yes |
| M01 | Media buys | Unspecified media buys | Yes |
| M10 | Broadcast ads | Broadcast ads | Yes |
| M20 | Print ads | Print ads | Yes |
| M30 | Web ads | Web ads | Yes |
| M40 | Media production | Media production | Yes |
| M50 | Media consulting | Media consulting | Yes |
| N10 | Contrib refunds | Contribution refunds | No |
| N99 | Non-expenditures | Other non-expenditures | No |
| R00 | Misc contributions | Miscellaneous contributions | No |
| R10 | Natl party contribs | Contributions to national parties | No |
| R20 | State party contribs | Contributions to state & local parties | No |
| R30 | Fedl cand contribs | Contributions to federal candidates | No |
| R35 | Contribs to JFCs | Contributions to joint fundraising committees | No |
| R40 | State cand contribs | Contributions to state & local candidates | No |
| R50 | Contribs to cmtes | Contributions to committees | No |
| S10 | Polling & surveys | Polling & surveys | Yes |
| S20 | Campaign data & tech | Campaign data & technology | Yes |
| S50 | Campaign consulting | Campaign strategy & communications consulting | Yes |
| T00 | Misc transfers | Miscellaneous transfers | No |
| T10 | Natl party transfers | Transfers to national parties | No |
| T20 | State party transfers | Transfers to state & local parties | No |
| T30 | Transfers to cands | Transfers to candidates | No |
| T50 | Transfers to cmtes | Transfers to committees | No |
| U00 | Unclassifiable | Unclassifiable | Yes |
| U10 | Unknown print & mail | Unclassifiable printing & shipping | Yes |
| U20 | Unknown data & tech | Unclassifiable data & technology | Yes |
| U30 | Unknown events | Unclassifiable event expenses | No |
| U50 | Unknown consulting | Unclassifiable consulting | Yes |
| U60 | Unknown supplies | Unclassifiable supplies & equipment | Yes |
| W10 | Salaries | Salaries, wages & benefits | No |
| Missing | | | Yes |

As explained in chapter three, the second step involved standardizing campaign-provided purposes in each cycle. I reviewed the unique campaign-provided purposes of the filtered expenditure records for each cycle and inductively created standardized descriptions. After every cycle was coded, I merged the standardized descriptions back with the entire sample. Using this

process, 205,840 unique campaign-provided descriptions ultimately reduced into 570 standardized descriptions. List B1 provides the standardized expenditures from the first step. Table B2 reports the results of the second step. The number of unique expenditure descriptions decreased between 2006 and 2008. The decrease in campaign-provided expenditure discrepancies happened in part because of the proliferation of compliance software that standardized some terminology. Yet the large gap between the campaign-provided description and standardized descriptions indicates the ongoing necessity of the refinement process.

**List B1 Standardized Expenditure Codes**

absentee ballot, access, administration, administrative consulting, advertising, advertising consulting, advertising materials, advertising placement, advertising production, agency fee, air bill, air travel, analysis, analysis consulting, apparel, appliance/furniture, art, assessment, auction item, audio production, audio service, automation system, av service, bags, balloons, ballot, ballot/filing fee, bank fee, banner advertising, banners, billboard advertising, biography, block walking, blogging, booklet, books, brm, broadcast advertising, broadcast consulting, broadcast email, broadcast production, brochures, building assessment/maintenance/materials, bulk mail, business cards, business consulting, business registration, buttons, cable expense, cable/internet/phone service, calendar, calendar advertising, camera, campaign software, candidate, canvassing, canvassing consulting, canvassing director, canvassing food/drink, canvassing software, canvassing/field materials, capital assessment, cards, cash, cell phone, check, child care, cleaning service, cleaning supplies, clipboards, coins, commercial, commercial production, communications, communications consulting, communications director/coordinator, communications equipment, communications software, community advertising, community consulting, community organizing/outreach, community public relations, compliance, compliance consulting, compliance software, computer consulting, computer equipment, computer service, conference, conference, conference call, conference fee, consultant lodging, consultant travel, consulting, contact, contract, contribution/donation, convention, convention advertising, coordination, coordination consulting, credit, credit card expense, credit card fee, crm, crypto, cups, data, data analysis, data card, data consulting, data entry, data licensing, data management, data management software, data plan, data processing, data service, data software, data storage, database, database consulting, database management, database service, database software, debt, decorations, demographics, deposit, design, design consulting, design software, digital advertising, digital advertising consulting, digital advertising email marketing, digital advertising production, digital communications, digital communications consulting, digital consulting, digital data, digital database, digital design, digital fundraising, digital fundraising consulting, digital list, digital marketing, digital marketing consulting, digital media, digital media consulting, digital media management, digital messaging, digital outreach, digital production, digital programming, digital promotion, digital service, digital support, digital video advertising, digital video production, direct mail, direct mail consulting, direct marketing, director, directory, directory advertising, disbursement, display advertising, district data, documents, door knocking, doorhangers, dues/membership, early voting, election day, email, email blast, email blasting, email consulting, email data/list, email data/list consulting, email data/list

maintenance, email event, email fundraising, email fundraising consulting, email hosting, email marketing, email service, email software, endorsement, entertainment, envelopes, equipment, event, event advertising, event flyers, event food/drink, event invitations, event materials, event space, event sponsorship, event supplies, event travel, facebook, facebook advertising, facebook endorsement, fans, fax, fax blast, fax service, fee, field consulting, field data, field director/manager/organizer, fieldwork, fieldwork consulting, fieldwork software, financial consulting, financial director, financial service, financial software, financial/compliance/legal, flag, flowers, flyers, focus group, food/drink, framing, fundraising, fundraising consulting, fundraising data, fundraising database, fundraising database software, fundraising email, fundraising email blast, fundraising event, fundraising event food/drink, fundraising event invitations, fundraising event mail, fundraising event space, fundraising event sponsorship, fundraising event supplies, fundraising fax, fundraising fee, fundraising food/drink, fundraising gifts, fundraising list, fundraising mail, fundraising event supplies, fundraising fax, fundraising fee, fundraising food/drink, fundraising gifts, fundraising list, fundraising mail, gotv canvassing, gotv consulting, gotv event, gotv mailers, gotv materials, gotv phone, gotv phone/text, gotv text, grassroots, grassroots consulting, grassroots organizing, ground travel, hair/makeup, handouts, hats, hosting, image licensing, in-kind contribution, individual, individual data, information, information cards, information consulting, instagram advertising, insurance, interest, internet, internet consulting, internet operations/work, internet/cable/phone service, interpreter, interview, invitations, issue consulting, it service, items, jewelry, journal advertising, keys, labels, leaflets, legal consulting, legal research, legal service, letterhead, letters, licensing, list, literature, lodging, logistics, logistics consulting, logo, magazine advertising, mail, mail consulting, mail data/list, mail service, mailbox, mailers, maintenance, management, management consulting, management software, mapping software, maps, marketing, marketing consulting, marketing materials, marketing research, marketing software, mass mail, mass text, materials, media, media buying, media consulting, media director, media materials, media monitoring, media placement, media prep, media production, media relations, media service, media training, meeting, merchandise consulting, messaging, messaging consulting, messaging software, mobile advertising, mobile canvassing, mobile communications, modeling, moving, name tags, nation builder, new media consulting, newsletter, newsletter advertising, newspaper, newspaper advertising, ngp, ngp data, ngp software, ngp van, ngp van data, non-expenditure, notary, office supplies, office/rent/utilities, online database, online endorsement, online service, operating expenses, operational/organizational consulting, operations, operations consulting, opposition research, organization, organizing, outreach, outreach consulting, page/program advertising, palm cards, pamphlets, parade, parade supplies, payroll/salary/staffing, pens, petition, phone, phone consulting, phone data/list, phone food/drink, phone service, phone software, phone/text, photography, pledge cards, podcast advertising, policy consulting, policy director, policy research, political advertising, poll watcher/worker, polling, polling consulting, postage/shipping, postcards, posters, posts, precinct targeting, presentation consulting, press relations, press release, print advertising, print media, printer, printing, production, program, program advertising, programming, promotion, promotional materials, prospectus, public relations, public relations consulting, publishing, push cards, rack cards, radio advertising, recording, records, recount, redistricting, registration, reimbursement, relational voting, remittance, rent, repair, report, research, research consulting, research database, research software, robocalls, robotext, sample ballot, satellite radio, scheduling, scholarship, security, security equipment, service, shirts, shredding, signs, slate cards, slate mailers, snapchat filter, social advertising, social media, social media advertising, social media consulting, social media director, social media endorsement, social media management, social media marketing, social media production, social media software, software, sound equipment, speech consulting, sponsorship, sponsorship advertising, statement, stationary, statistical modeling, stickers, storage, streaming service advertising, subscription, supplies, survey, survey consulting, tabloids, targeting, targeting consulting, tax, technology, technology consulting, telecommunications, telemarketing, telephone, television, television advertising, television advertising consulting, television advertising production, television consulting, television data, television production, tent, text, text data/list, text software, ticket, tip, tools, towels, tracking, training, transcription,

transfer, transition, translation, travel, treasurer, twitter advertising, uncoded, unknown, utilities, van, van data, video advertising, video advertising production, video production, video production consulting, video research, volunteer, volunteer coordinator, volunteer event, volunteer food/drink, volunteer gifts, volunteer lodging, volunteer supplies, volunteer travel, vote builder, voter data, voter data software, voter database, voter file, voter guide, voter information, voter list, voter mail, voter outreach, voter outreach consulting, voter outreach software, voter records, walk cards, web design/development/maintenance, web expense, web hosting, web service, web support, website, website consulting, website endorsement, writing, yard signs, youtube, youtube advertising

**Table B2 Count of Unique Descriptions and Standardized Descriptions per Cycle**

| Cycle | Unique Campaign-provided Descriptions | Unique Standardized Descriptions |
|---|---|---|
| 2006 | 57,450 | 389 |
| 2008 | 60,451 | 401 |
| 2010 | 63,238 | 422 |
| 2012 | 55,573 | 433 |
| 2014 | 44,959 | 439 |
| 2016 | 38,159 | 433 |
| 2018 | 38,601 | 520 |

The third step involves resolving expenditures with multiple purchases and collapsing the standardized descriptions into fewer categories. For some expenditures, campaigns report compound purposes, such as "voter outreach and website." Expenditures with multiple descriptions of this kind represent less than 1% of expenditure records. To resolve expenditures with multiple purposes, the standardized description is assigned based on which description has the highest median cost. With only a single description per expenditure record, I next collapse the standardized descriptions in List B1 to reduce again the number of unique descriptions. This collapsing step also eliminated many unique descriptions that are unrelated to voter data or voter outreach, being instead grouped into a description of "other." All descriptions that are potentially related to data and outreach are preserved, however. Grouping together standardized descriptions reduced the number of unique descriptions further to 127, which are provided in List B2.

**List B2 Collapsed Standardized Descriptions**

administration, advertising, advertising consulting, advertising materials, advertising production, analysis/research, billboard advertising, broadcast advertising, broadcast consulting, broadcast production, campaign software, canvassing, canvassing consulting, canvassing software, canvassing/field materials, communications, communications consulting, communications software, community organizing/outreach, consulting, contribution/donation, data entry/management/processing, data/database, data/database consulting, data/database software, design, design consulting, design software, digital advertising, digital advertising consulting, digital advertising production, digital data/database, digital fundraising, digital fundraising consulting, digital marketing, digital marketing consulting, digital media, digital media consulting, digital outreach, email, email consulting, email data/list, email data/list consulting, email service/software, event, event advertising, fax, field consulting, fieldwork, fundraising, fundraising consulting, fundraising data/database, fundraising data/database software, fundraising email, fundraising event, fundraising fax, fundraising mail, fundraising materials, fundraising phone/text, fundraising software, gotv, gotv consulting, gotv mailers, gotv materials, gotv phone/text, grassroots, grassroots consulting, issue consulting, list, mail consulting, mail data/list, mailers, marketing, marketing consulting, marketing materials, marketing software, media, media buying/placement, media consulting, media production, media relations, messaging, messaging consulting, messaging software, newsletter, newspaper advertising, other, other data, other digital, other media, outreach, outreach consulting, outreach materials, page/program advertising, phone/text, phone/text consulting, phone/text data/list, phone/text software, policy, polling/survey, polling/survey consulting, postage/shipping, print advertising, printing, promotional materials, radio advertising, signs, social media, social media advertising, social media consulting, social media software, software, television advertising, television advertising consulting, television advertising production, video advertising, video advertising production, volunteer, voter data/database, voter data/database software, voter information/records, voter list, voter outreach, voter outreach consulting, voter outreach software, web expense, website

In step four, I verify expenditures related to data and polling by investigating the primary purpose of the vendor providing them. Coding the primary purpose of data vendors first involves examining the recipients' names and standardizing the spelling of their names. As with the expenditure descriptions, slight discrepancies and misspellings are also present in the campaign-provided recipient names. For example, NGP VAN, a leading data vendor for the Democratic Party, had various listed names, such as "ngp," "ngp van, inc", "van," and "vote builder."

After standardizing vendors' names, I reviewed each vendor with more than three data-related expenditures in any given cycle. If possible, I assigned each vendor a type as primarily specializing in polling or individual-level data by examining the standardized descriptions and the vendor's archived website provided by archive.org. I also recorded if the vendor was unrelated to data or

polling or ambiguous. The recipient type also provided an additional check on the campaign-provided descriptions. Descriptions of "data" that come from unrelated recipients, such as AT&T providing cellular data, can be corrected. Or in the case of "software" being linked to a voter-data provider, such as Aristotle International, the expenditure can be correctly classified as "voter data software" rather than what would be indicated in the OpenSecrets-provided category of "unclassifiable data & technology." In total, the description-recipient review of data-related expenditures alters approximately 40,000 records across all cycles. Table 3.2 in chapter three provides counts of each vendor type, the number of expenditures they provided, and the total amount of money each type received.

In the final step, I decided on a case-by-case basis whether a standardized description could appropriately be placed into a finalized category. Some cases are less clear. For instance, many descriptions are described as related to "consulting" or more specifically, for instance, "data/database consulting." While these could potentially be separated into service and advice components, I choose to consider the description as related to the main category of interest. Thus, the examples would be grouped under "data/database" and "polling/survey." Likewise, I chose to code software-related descriptions into the main category of interest. List B3 provides the finalized descriptions.

**List B3 Finalized Descriptions**

advertising/mass media, campaign/outreach software, communications, consulting, contact, design, digital data, digital fundraising, digital media, email, event, fundraising, fundraising contact, fundraising data, individual/voter data, marketing, marketing materials, media relations, other, other data, other digital, other media, other research, polling/survey, postage/shipping, printing, promotional materials, software, website

Figure 1 summarizes these quality control procedures by revealing the top standardized descriptions within these two categories. Each square in the figure represents approximately 1%

of total spending across cycles. For instance, "software" makes up approximately 30% of spending for the finalized description "individual/voter data." "Data/database" amounts to nearly 25% and "list" comprises close to 9% of spending. "Remaining descriptions" refers to all standardized descriptions that comprise less than 1% by themselves. About 80% of finalized polling-related expenditures could be initially classified as "polling/survey." The ambiguous descriptions of "analysis/research" and "consulting" also are corrected and placed into the proper categories.

**Figure B1: Data-related Standardized Descriptions in Final Descriptions**

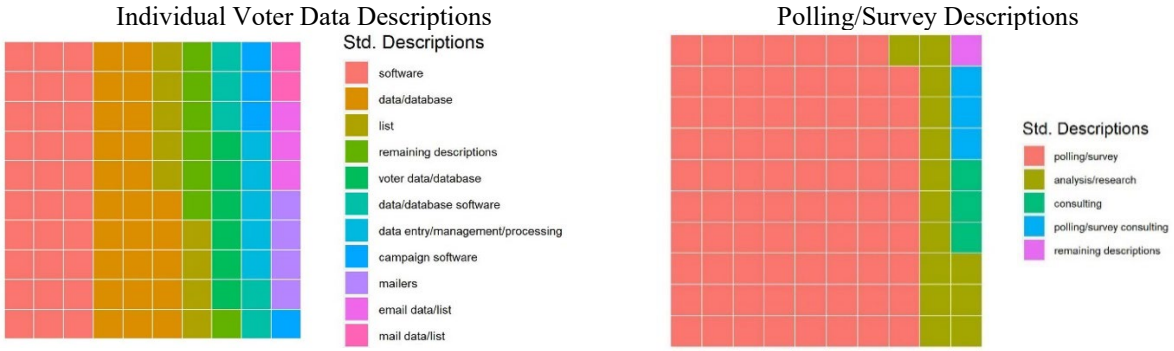
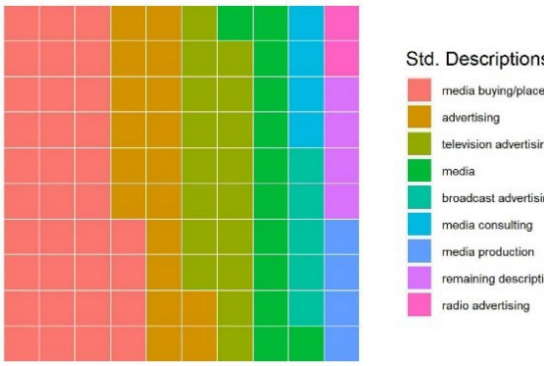


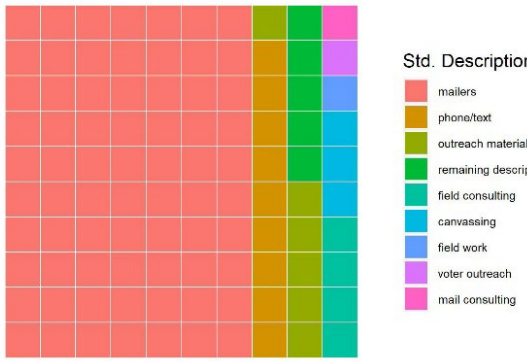

Figure B2 reveals the standardized descriptions within the final descriptions related to voter outreach. Around half of "advertising/mass media" finalized descriptions are originally standardized as "media buying/placement" and "advertising." Significant portions also include the descriptions of "television advertising" and generic "media" along with consulting and production. The finalized "direct contact" category is mostly comprised of "mailers" as well as "phone/text" forms of outreach. The largest portions of digital spending went to "digital advertising" and "digital media" or "digital media consulting."

**Figure B2: Outreach-related Standardized Descriptions in Final Descriptions**

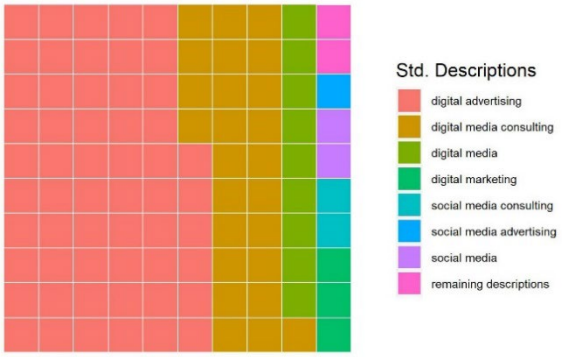
Advertising/Mass Media Descriptions




Direct Contact Descriptions




Digital Media Descriptions



**Appendix C:** Supplemental Analyses from Chapter Three

**Figure C1: Unweight Distributions for the Cost of Individual-level Data and Polling**

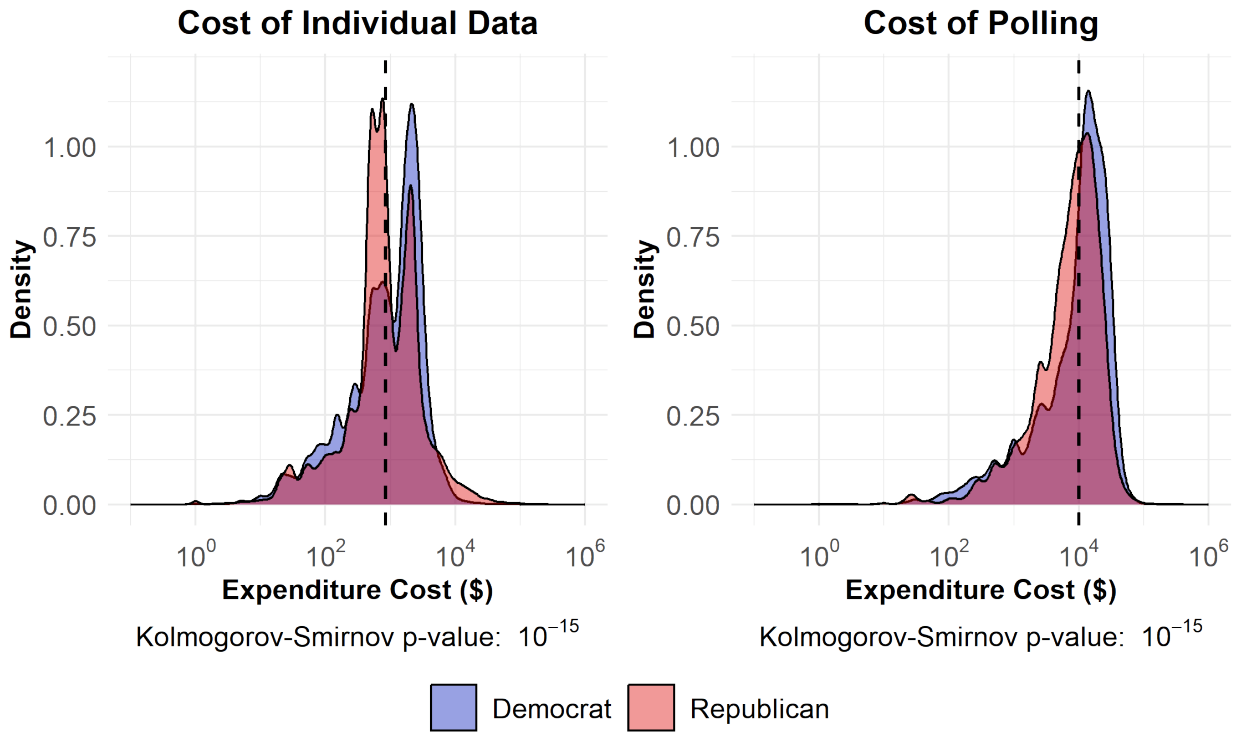


**Figure C2: Real Median Cost of Individual Data in 2018 Dollars, 2006-18**

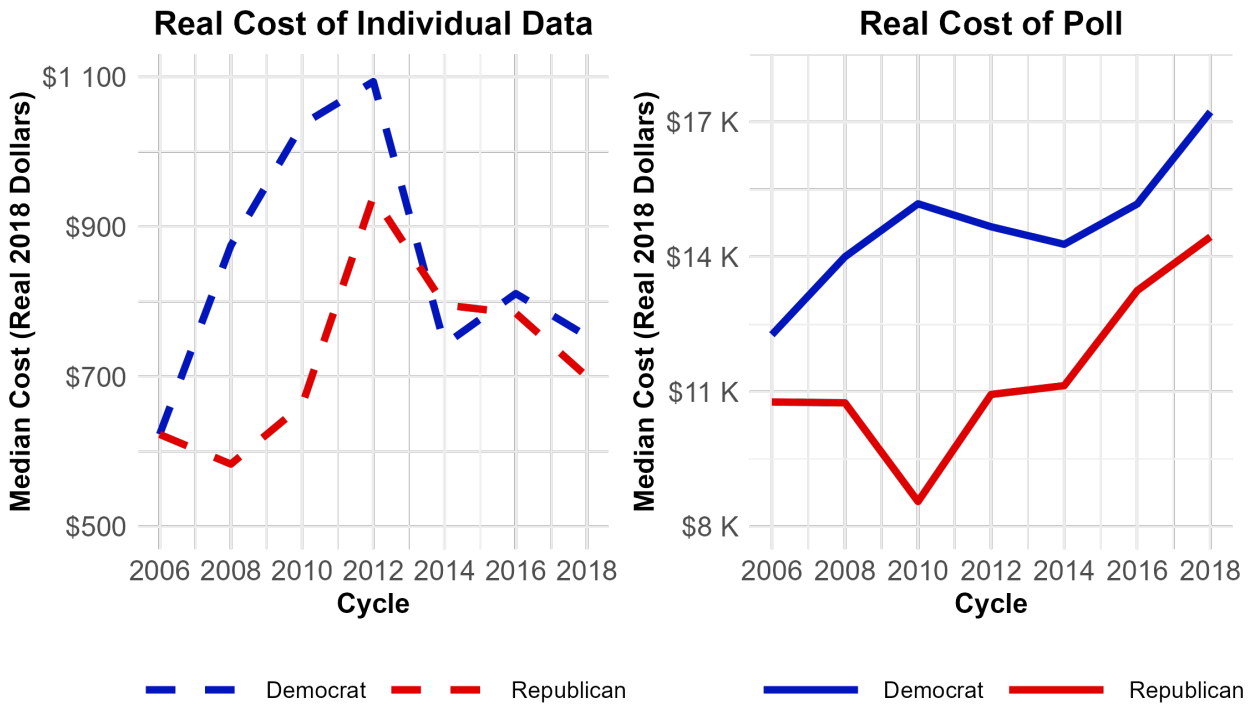
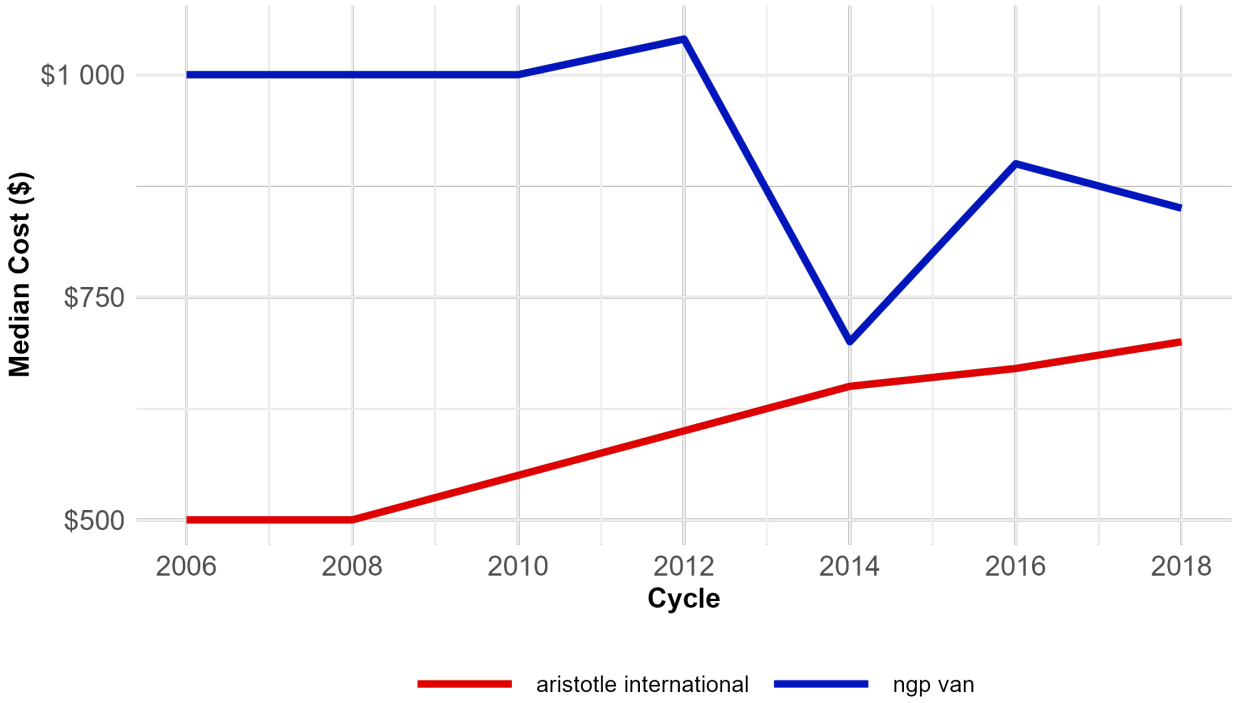


**Figure C3: Median Cost of Individual Data Among Top Two Firms, 2006-18**



**Appendix D:** Supplemental Analyses from Chapter Four

**Figure D1: Full ZOIB Models with Party-Spending Interaction**

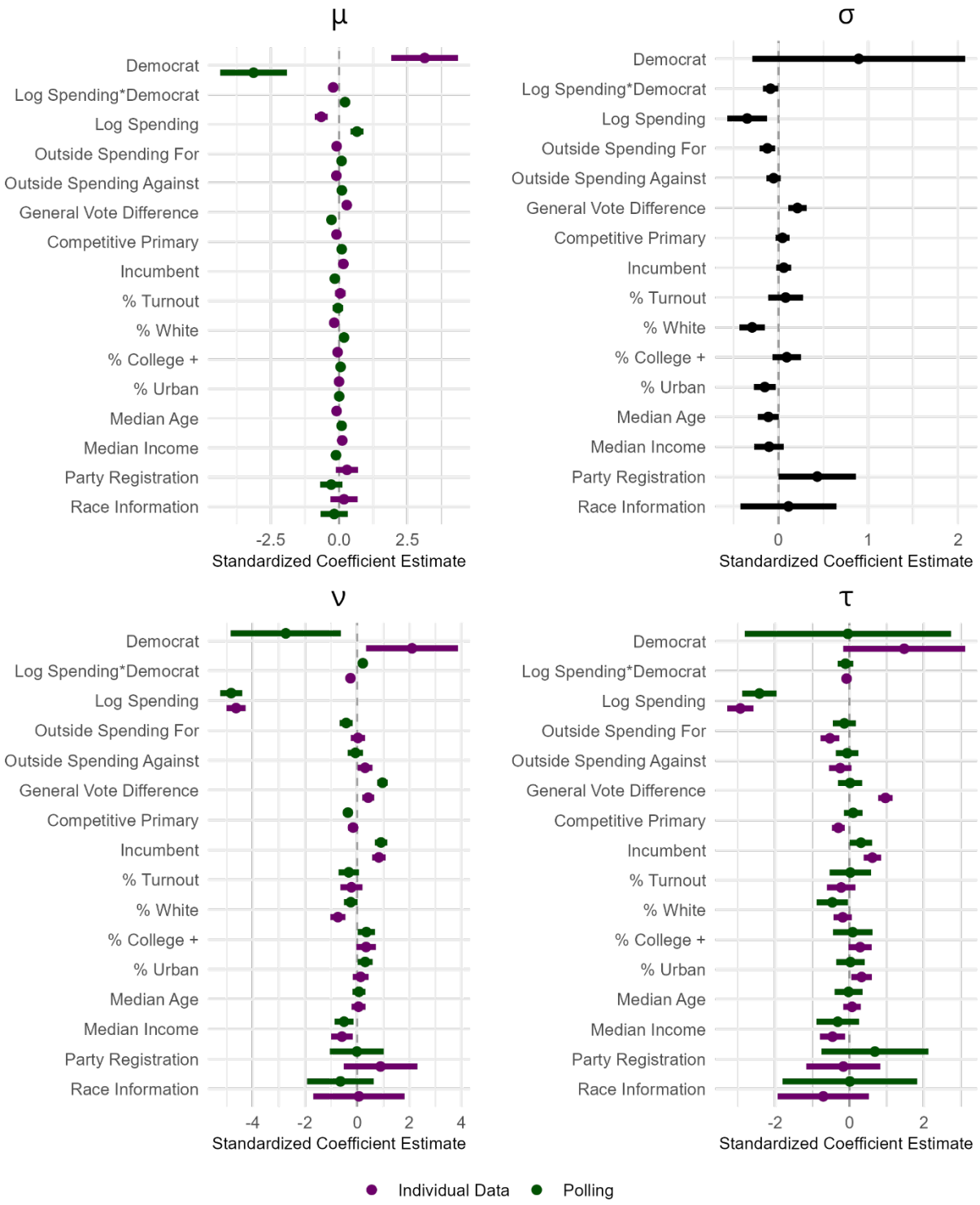


**Figure D2: Logistic Regression for Any Data Source Spending**

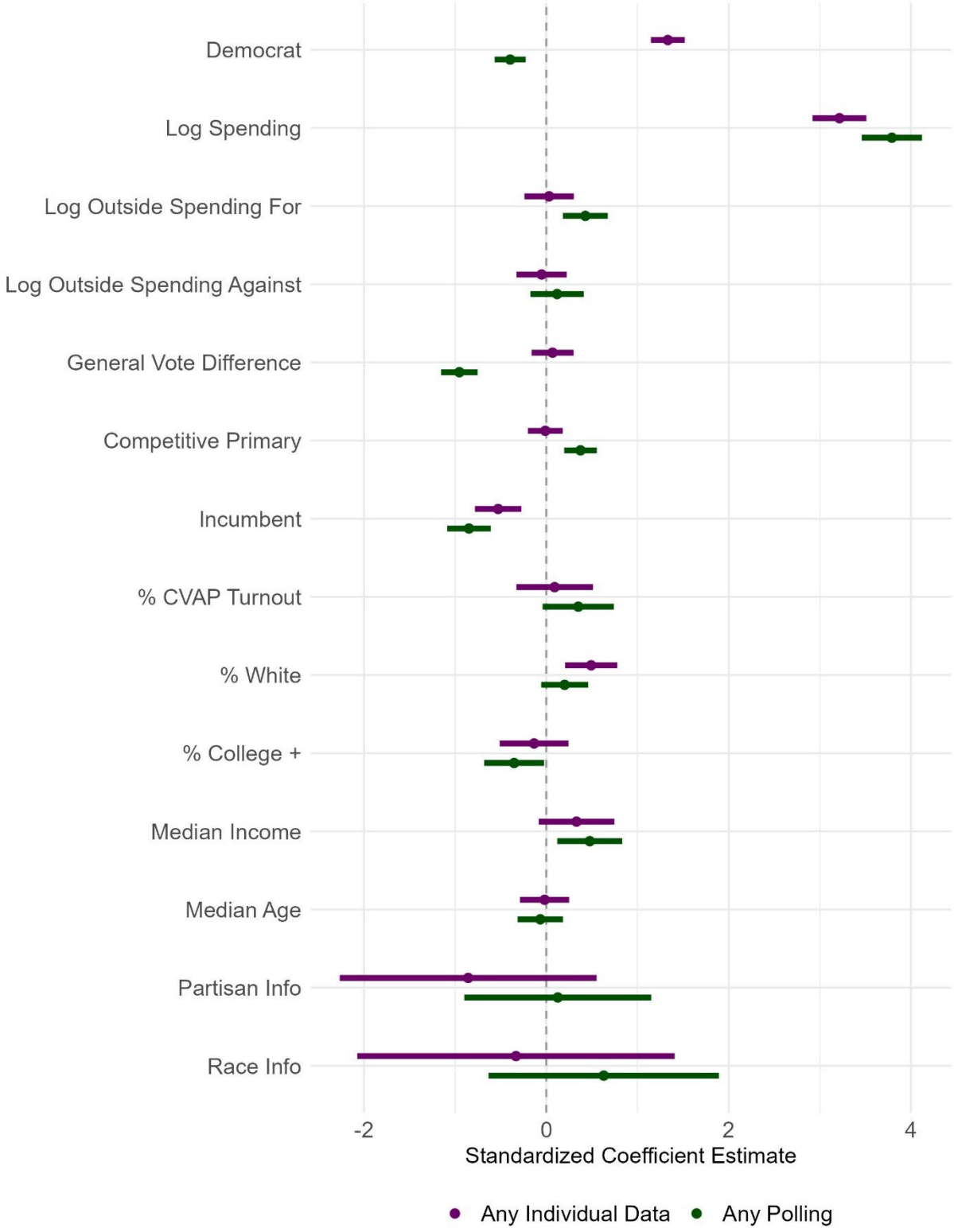


**Figure D3: OLS Data Source Spending Regression**

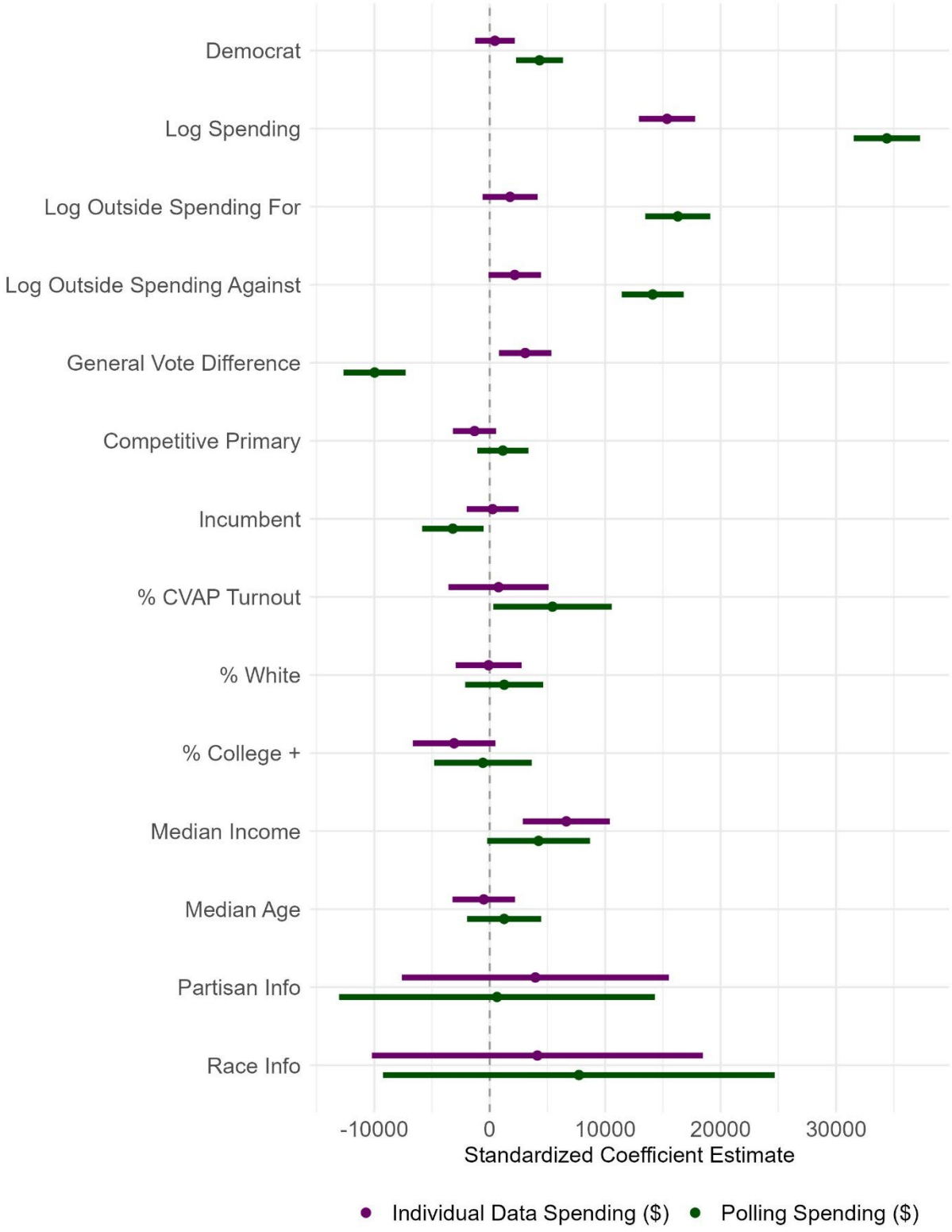


**Figure D4: ZOIB $\mu$-only Model Specification for Party Effect Estimate (Individual Data)**

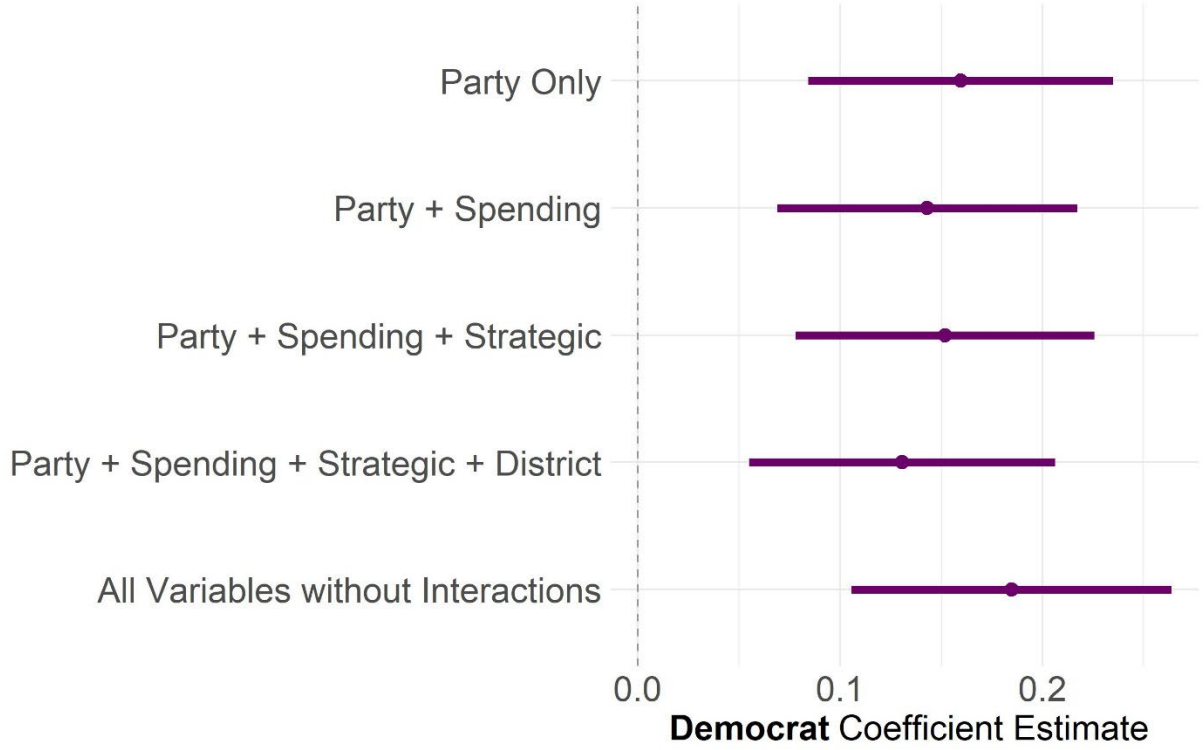


**Figure D5: Full ZOIB Model without Party-Spend Interaction**

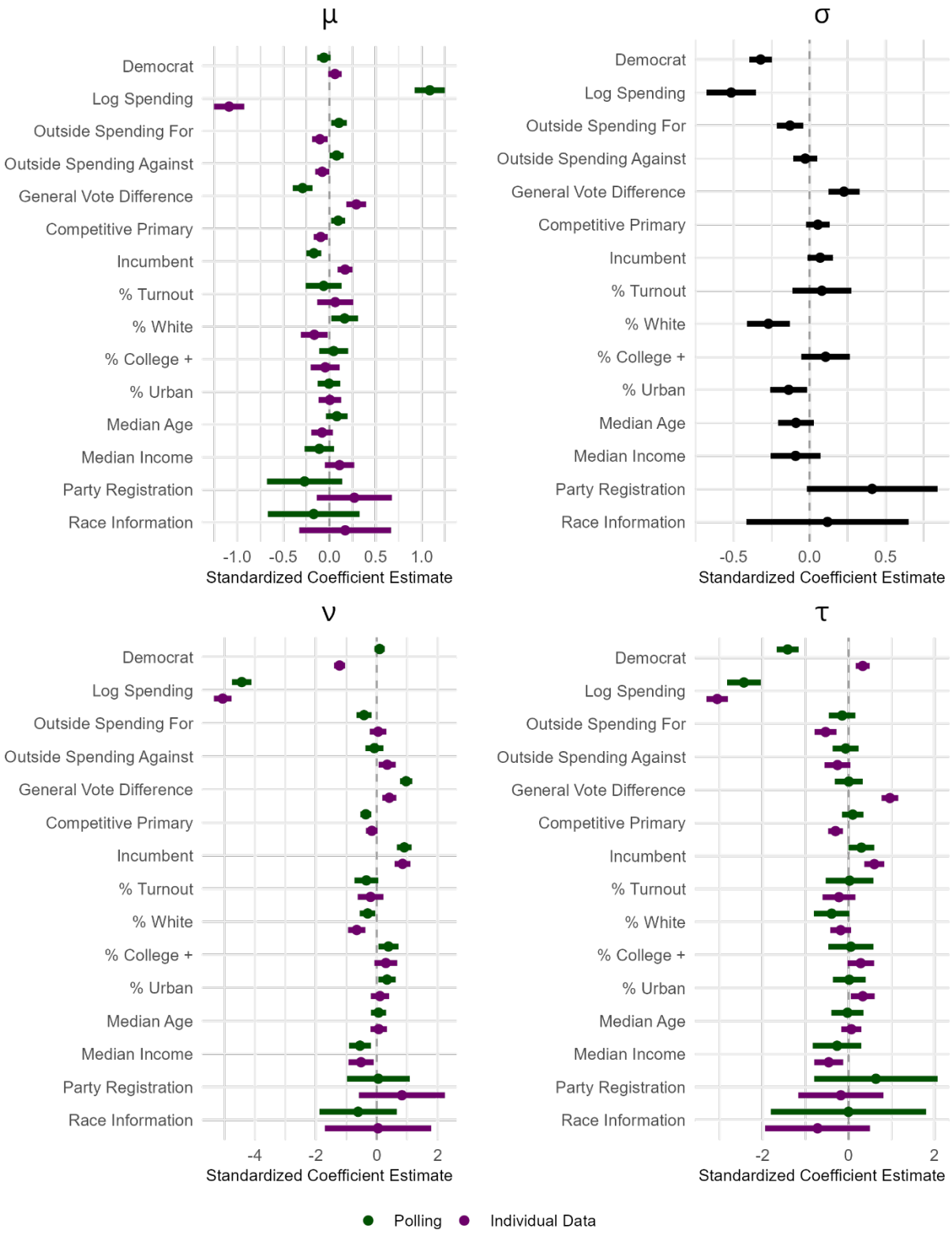


**Figure D6: Individual-Data ZOIB Model Restricted to General Vote Margin <= 5%**

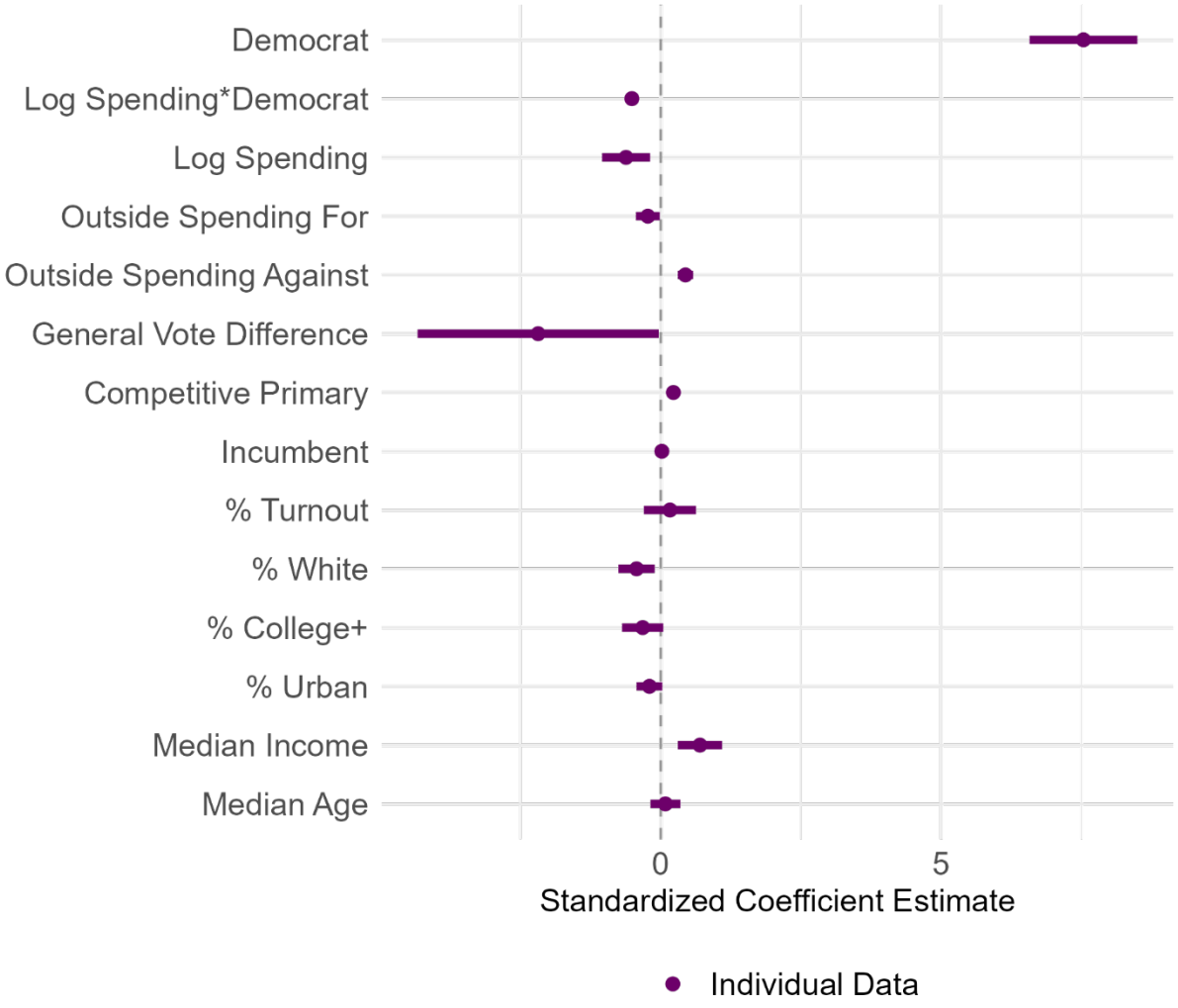
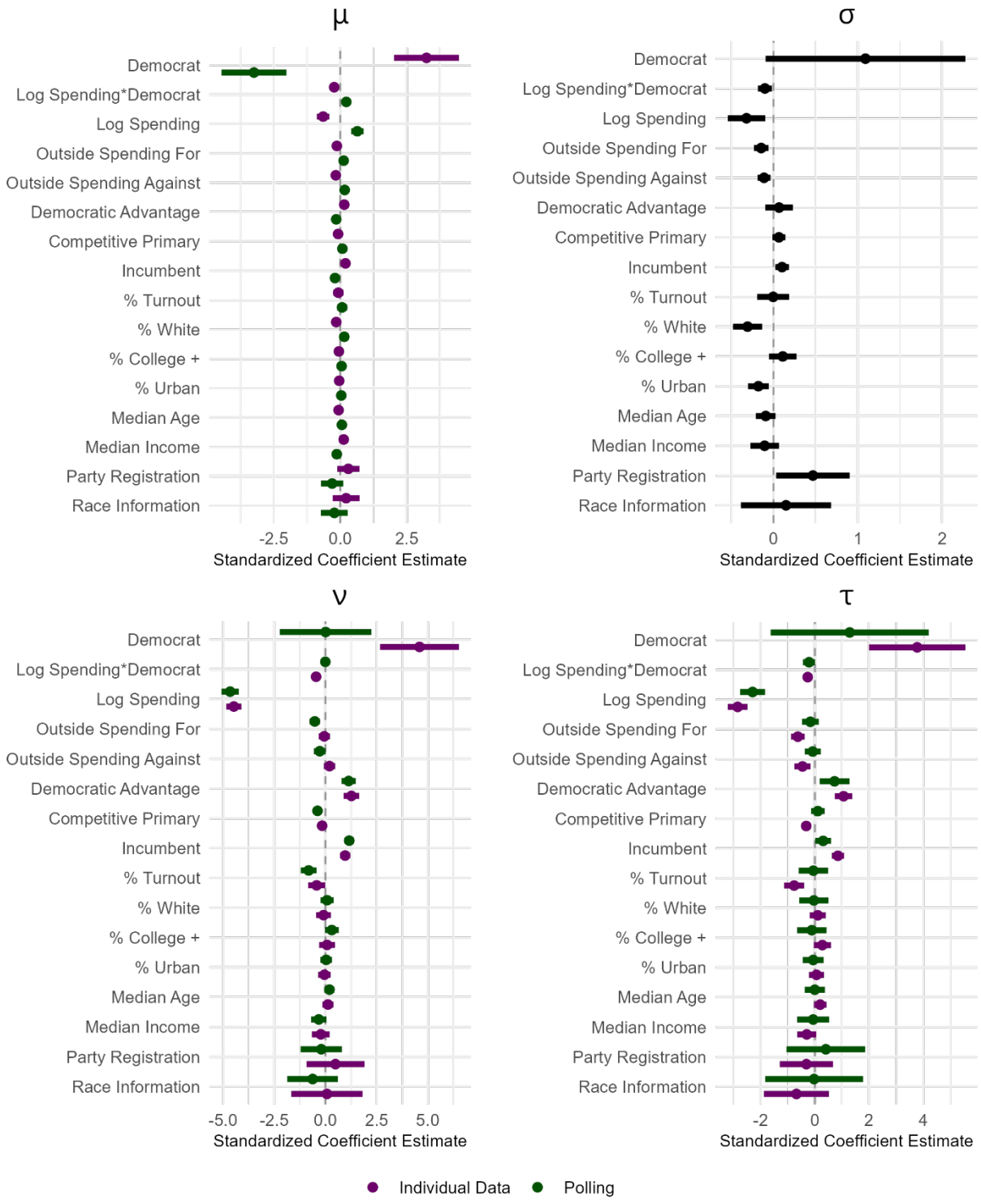


**Figure D7: Full ZOIB Model with Democratic Advantage Measure of Competitiveness**



**Appendix E:** Supplemental Analyses from Chapter Five

**Figure E1: Campaign-level Spending on Direct Contact in Dollars**

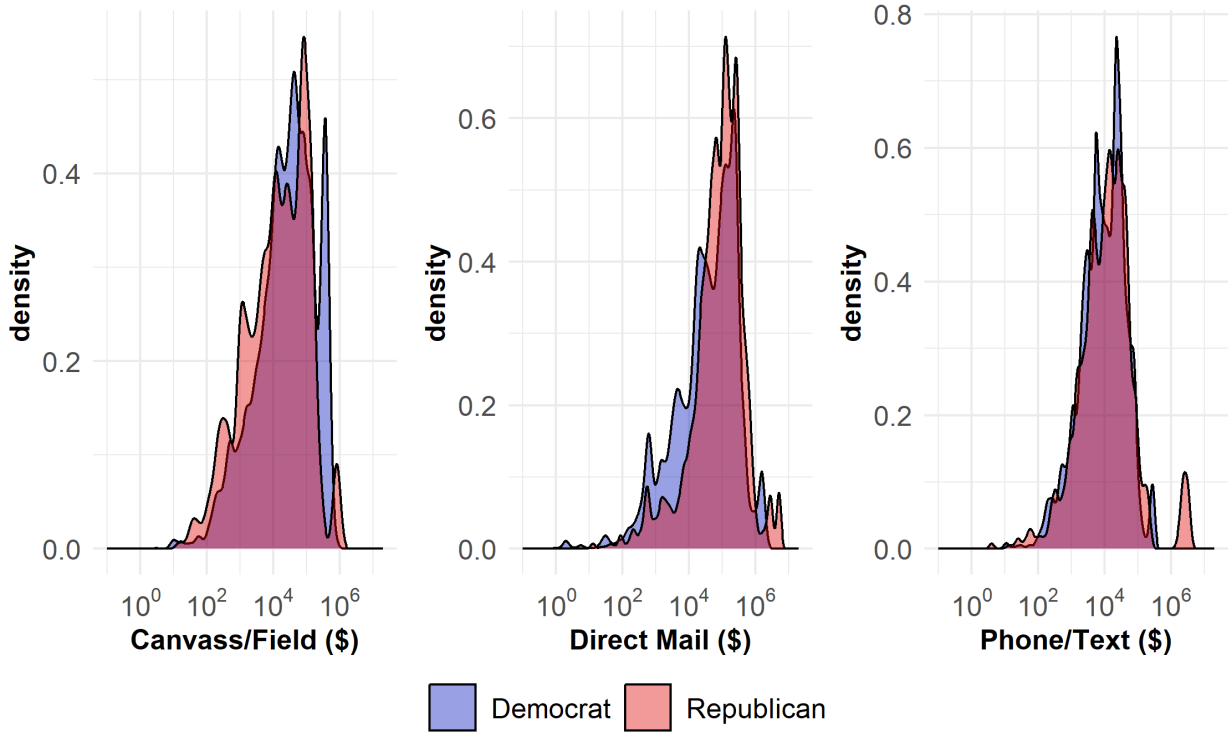


**Figure E2: Zero-inflated Mass Media Beta Regression with Interactions**

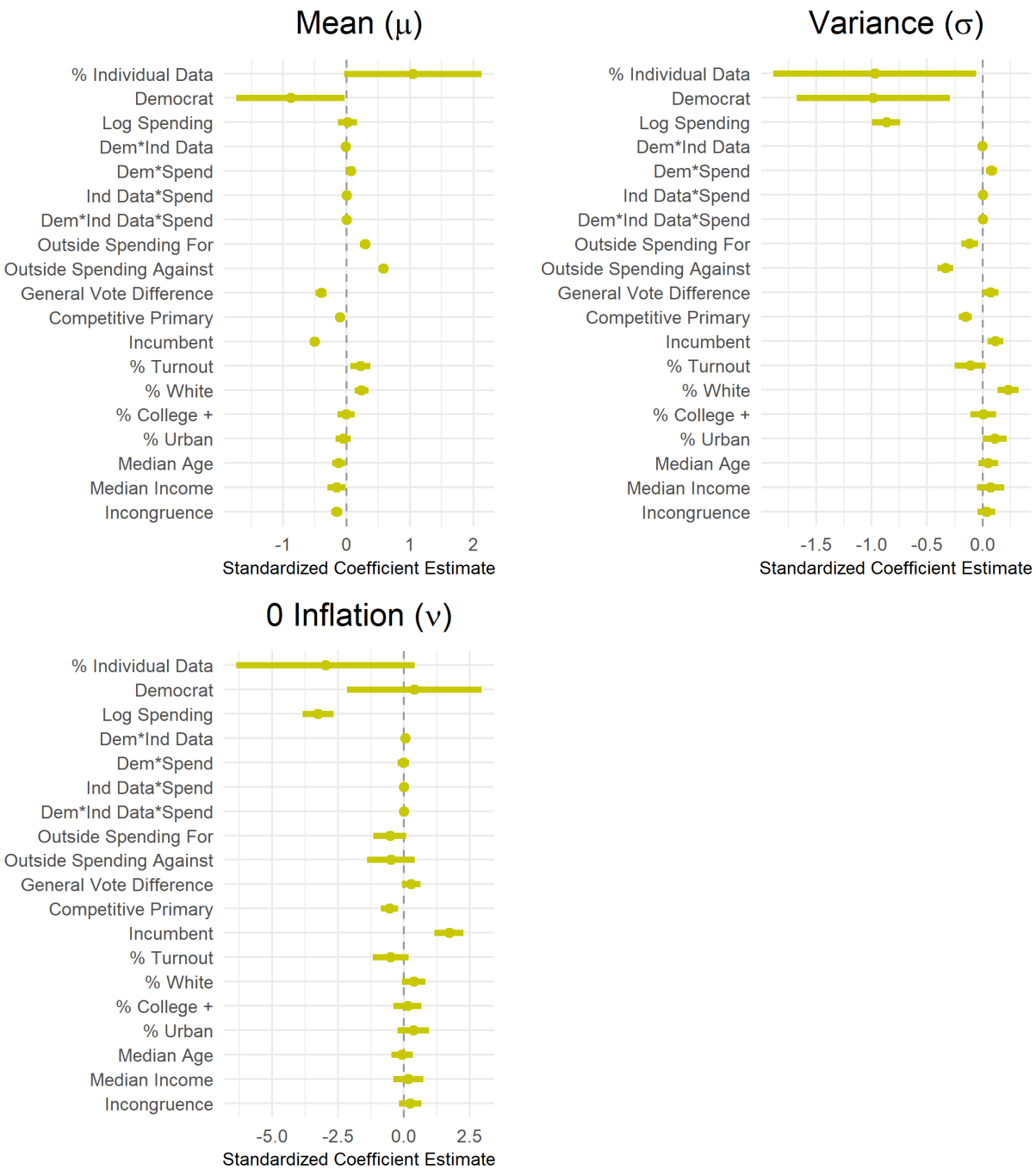


**Figure E3: Zero-inflated Mass Media Beta Regression with Democratic Advantage**

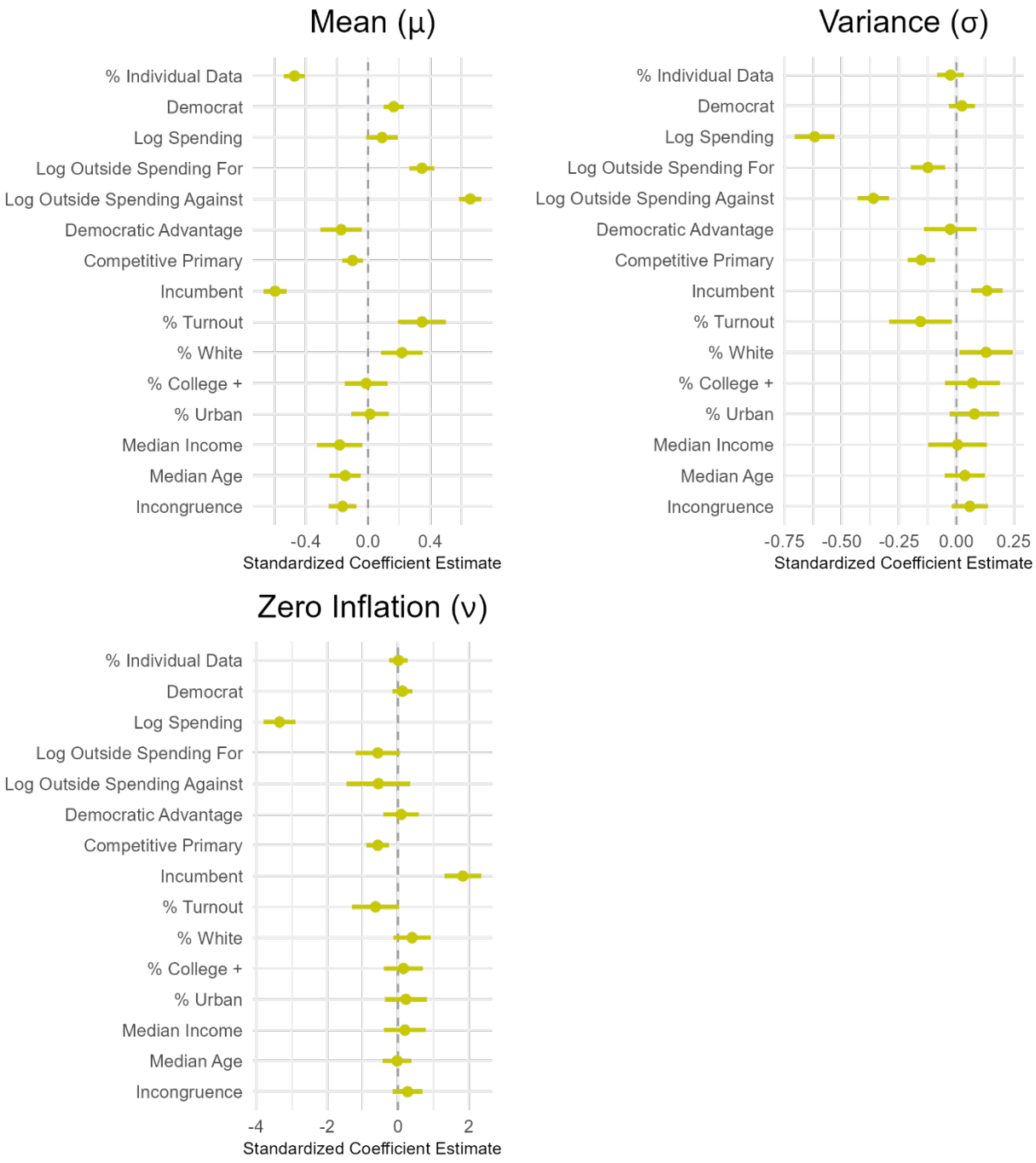


**Figure E4: Zero-inflated Direct Contact Beta Regression with Interactions**

**Figure E5: Zero-inflated Direct Contact Beta Regression with Democratic Advantage**

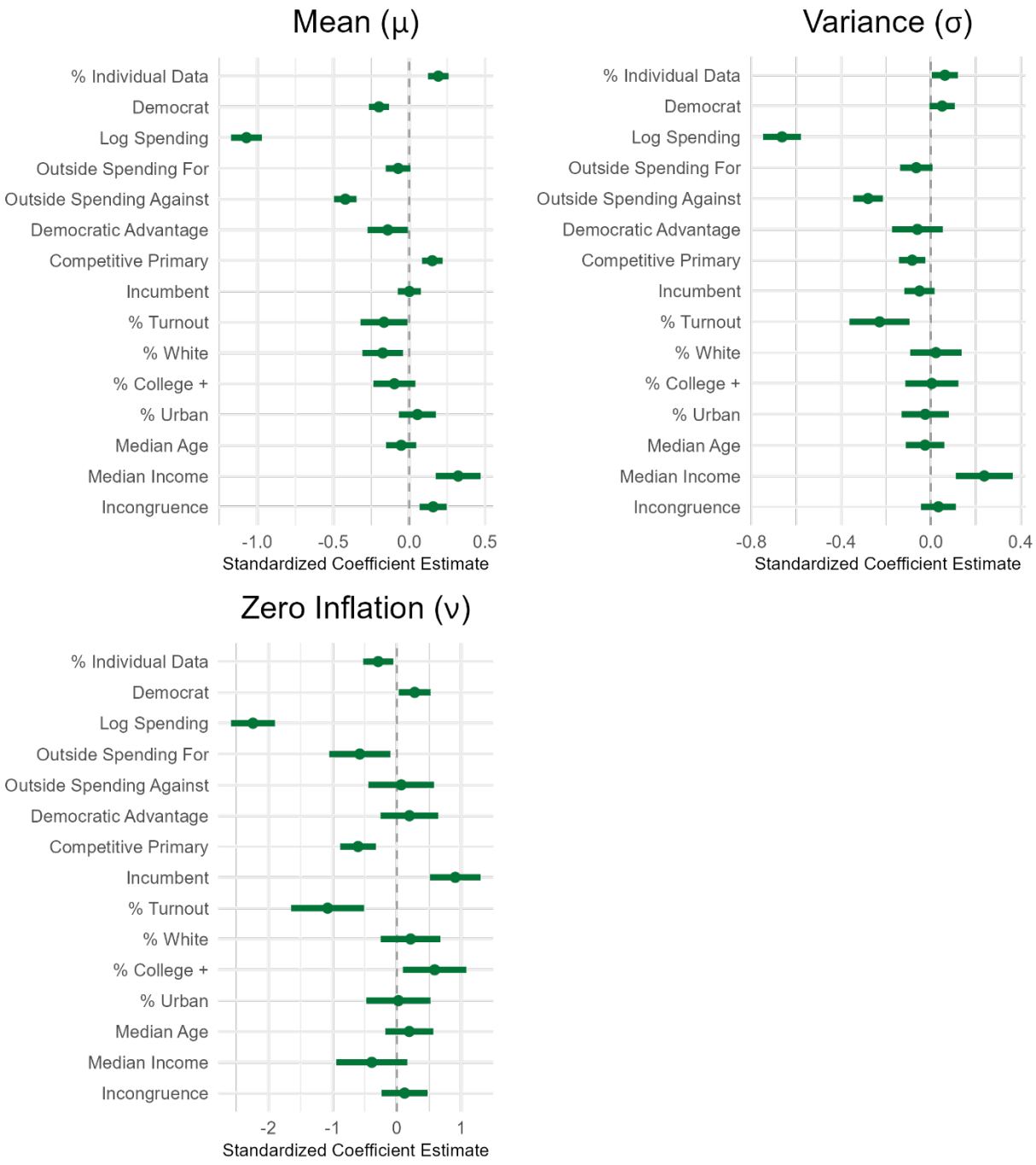
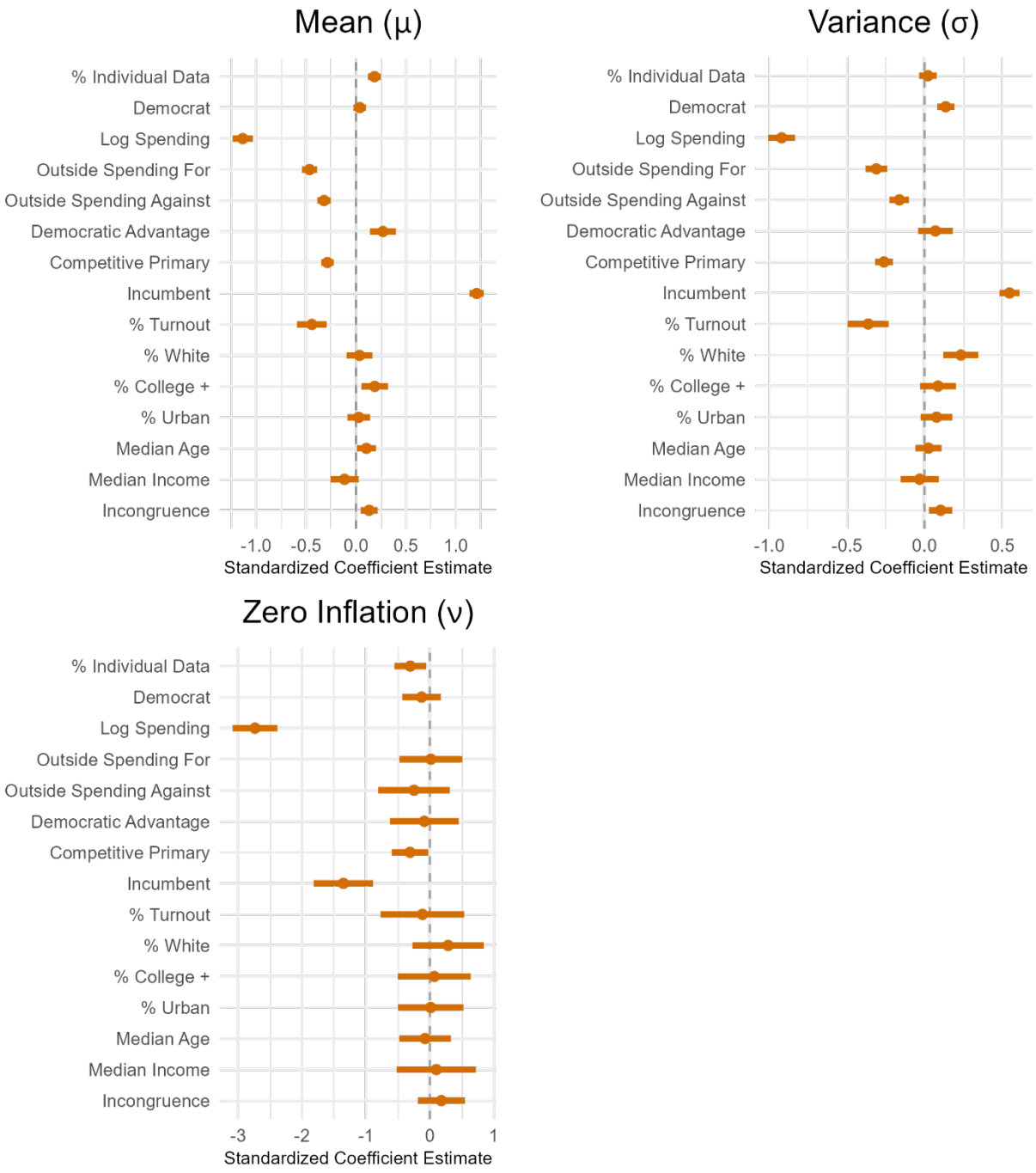


**Figure E6: Zero-inflated Digital Media Beta Regression with Interactions**

**Figure E7: Zero-inflated Digital Media Beta Regression with Democratic Advantage**



# References


Aberbach, Joel D., and Bert A. Rockman. 2002. "Conducting and Coding Elite Interviews." *PS: Political Science & Politics* 35(4): 673–76.

Achen, Christopher H., and Larry M. Bartels. 2017. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton: Princeton University Press.

Agranoff, Robert. 1977. *The New Style in Election Campaigns*. New York: Holbrook Press.

Aldrich, John H, Rachel K Gibson, Marta Cantijoch, and Tobias Konitzer. 2016. "Getting out the Vote in the Social Media Era: Are Digital Tools Changing the Extent, Nature and Impact of Party Contacting in Elections?" *Party Politics* 22(2): 165–78.

Alenazi, Abdulaziz. 2022. "Regression for Compositional Data with Compositional Data as Predictor Variables with or without Zero Values." *Journal of Data Science* 17(1): 219–38.

Alvarez, R. Michael, Asa Hopkins, and Betsy Sinclair. 2010. "Mobilizing Pasadena Democrats: Measuring The Effects of Partisan Campaign Contacts." *The Journal of Politics* 72(1): 31–44.

Ansolabehere, Stephen, and Brian Schaffner. 2018. "Taking the Study of Political Behavior Online." In *The Oxford Handbook of Polling and Survey Methods*, eds. Lonna Rae Atkeson and R. Michael Alvarez. Oxford University Press.

Ansolabehere, Stephen, and Iyengar Shanto. 1995. *Going Negative: How Political Advertisements Shrink and Polarize the Electorate*. The Free Press.

Anstead, Nick. 2017. "Data-Driven Campaigning in the 2015 United Kingdom General Election." *The International Journal of Press/Politics* 22(3): 294–313.

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

———. 2010. "Comparing Negative and Positive Campaign Messages: Evidence From Two Field Experiments." *American Politics Research* 38(1): 54–83.

Baker, Reg et al. 2013. "Summary Report of the AAPOR Task Force on Non-Probability Sampling." *Journal of Survey Statistics and Methodology* 1(2): 90–143.

Baldwin-Philippi, Jessica. 2016. "The Cult(Ure) of Analytics in 2014." In *Communication and Midterm Elections: Media, Message, and Mobilization*, eds. John Allen Hendricks and Dan Schill. New York: Palgrave Macmillan, 25–42.

———. 2017. "The Myths of Data-Driven Campaigning." *Political Communication* 34(4): 627–33.

———. 2018. "Data-Driven Campaigning." In *Digital Discussions*, eds. Natalie Jomini Stroud and Shannon C. McGregor. New York: Routledge, 179–202.

———. 2019. "Data Campaigning: Between Empirics and Assumptions." *Internet Policy Review* 8(4).

Ballard, Andrew O., D. Sunshine Hillygus, and Tobias Konitzer. 2016. "Campaigning Online: Web Display Ads in the 2012 Presidential Campaign." *PS: Political Science & Politics* 49(3): 414–19.

Bankston, Levi, and Barry C. Burden. 2023. "Voter Mobilization Efforts Can Depress Turnout." *Journal of Elections, Public Opinion and Parties* 33(1): 94–104.

Bartels, Larry M. 2014. "Remembering to Forget: A Note on the Duration of Campaign Advertising Effects." *Political Communication* 31(4): 532–44.

Belfry Munroe, Kaija, and H.D. Munroe. 2018. “Constituency Campaigning in the Age of Data.” *Canadian Journal of Political Science* 51(1): 135–54.

Berelson, Bernard R., Paul Lazarsfeld, and William McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University Of Chicago Press.

Bimber, Bruce. 2014. “Digital Media in the Obama Campaigns of 2008 and 2012: Adaptation to the Personalized Political Communication Environment.” *Journal of Information Technology & Politics* 11(2): 130–50.

Blaemire, Robert. 2012. “An Explosion of Innovation: The Voter-Data Revolution.” In *Margin of Victory: How Technologists Help Politicians Win Elections*, ed. Nathaniel G. Pearlman. Santa Barbara: Praeger, 107–19.

Boulianne, Shelley. 2018. “Twenty Years of Digital Media Effects on Civic and Political Participation.” *Communication Research* 47(7): 947–66.

Brader, Ted. 2005. “Striking a Responsive Chord: How Political Ads Motivate and Persuade Voters by Appealing to Emotions.” *American Journal of Political Science* 49(2): 388–405.

Branton, Regina, Jared Perkins, and Samantha Pettey. 2019. “To Run or Not to Run? U.S. House Campaign Advertising.” *Journal of Political Marketing* 18(3): 196–215.

Brown, M. Craig, and Charles N. Halaby. 1987. “Machine Politics in America, 1870-1945.” *The Journal of Interdisciplinary History* 17(3): 587–612.

Bryce, James. 1921. *Modern Democracies*. New York: Macmillan.

Burden, Barry C. 2004. “Candidate Positioning in US Congressional Elections.” *British Journal of Political Science* 34(2): 211–27.

Burton, Michael, William Miller, and Daniel Shea. 2015. *Campaign Craft: The Strategies, Tactics, and Art of Political Campaign Management*. 5th Edition. Santa Barbara: Praeger.

Burton, Michael, and Daniel M. Shea. 2010. *Campaign Craft: The Strategies, Tactics, and Art of Political Campaign Management*. 4th Edition. Santa Barbara: Praeger.

Cain, Sean A. 2013. “Political Consultants and Party-Centered Campaigning: Evidence from the 2010 U.S. House Primary Election Campaigns.” *Election Law Journal: Rules, Politics, and Policy* 12(1): 3–17.

Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter*. New York: Wiley.

Campbell, James E., John R. Alford, and Keith Henry. 1984. “Television Markets and Congressional Elections.” *Legislative Studies Quarterly* 9(4): 665–78.

Chartrand, Robert L. 1977. “Information Technology and the Political Campaigner.” In *The New Style in Election Campaigns*, ed. Robert Agranoff. New York: Holbrook Press.

Chester, Jeff, and Kathryn C. Montgomery. 2017. “The Role of Digital Marketing in Political Campaigns.” *Internet Policy Review* 6(4).

Clauset, Aaron, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. “Power-Law Distributions in Empirical Data.” *SIAM Review* 51(4): 661–703.

Converse, Jean M. 1987. *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.

Darr, Joshua P. 2020. “Polls and Elections: Abandoning the Ground Game? Field Organization in the 2016 Election.” *Presidential Studies Quarterly* 50(1): 163–75.

Darr, Joshua P., and Matthew S. Levendusky. 2014. “Relying on the Ground Game: The Placement and Effect of Campaign Field Offices.” *American Politics Research* 42(3): 529–48.

Diamond, Edwin, and Stephen Bates. 1992. *The Spot: The Rise of Political Advertising on*

*Television*. 3rd Edition. Cambridge: The M.I.T. Press.
Dimitrova, Daniela V., Adam Shehata, Jesper Strömbäck, and Lars W. Nord. 2014. "The Effects of Digital Media on Political Knowledge and Participation in Election Campaigns: Evidence From Panel Data." *Communication Research* 41(1): 95–118.
Doherty, David, and E. Scott Adler. 2014. "The Persuasive Effects of Partisan Campaign Mailers." *Political Research Quarterly* 67(3): 562–73.
Eckman, Sarah J. 2021. *Voter Registration Records and List Maintenance for Federal Elections*. Congressional Research Service.
Eisinger, Robert M. 2000. "Gauging Public Opinion in the Hoover White House: Understanding the Roots of Presidential Polling." *Presidential Studies Quarterly* 30(4): 643–61.
———. 2003. *The Evolution of Presidential Polling*. New York: Cambridge University Press.
Endres, Kyle. 2020. "Targeted Issue Messages and Voting Behavior." *American Politics Research* 48(2): 317–28.
Endres, Kyle, and Kristin J. Kelly. 2018. "Does Microtargeting Matter? Campaign Contact Strategies and Young Voters." *Journal of Elections, Public Opinion and Parties* 28(1): 1–18.
Enos, Ryan D., Anthony Fowler, and Lynn Vavreck. 2014. "Increasing Inequality: The Effect of GOTV Mobilization on the Composition of the Electorate." *The Journal of Politics* 76(1): 273–88.
Epstein, Ben. 2018. "How Innovative Was the Trump Campaign in 2016: A Historical Perspective." *SSRN Electronic Journal*. https://www.ssrn.com/abstract=3416832 (July 12, 2023).
Fenno, Richard F. 1978. *Home Style: House Members in Their Districts*. New York: Longman.
Foos, Florian, and Peter John. 2018. "Parties Are No Civic Charities: Voter Contact and the Changing Partisan Composition of the Electorate." *Political Science Research and Methods* 6(02): 283–98.
Fowler, Erika, Michael Franz, and Travis Ridout. 2016. *Political Advertising in the United States*. 1st Edition. Boulder, CO: Westview Press.
Fowler, James H. 2006. "Habitual Voting and Behavioral Turnout." *The Journal of Politics* 68(2): 335–44.
Franz, Michael M., Erika Franklin Fowler, Travis Ridout, and Meredith Yiran Wang. 2020. "The Issue Focus of Online and Television Advertising in the 2016 Presidential Campaign." *American Politics Research* 48(1): 175–96.
Franz, Michael M., and Travis N. Ridout. 2007. "Does Political Advertising Persuade?" *Political Behavior* 29(4): 465–91.
Gabaix, Xavier. 2009. "Power Laws in Economics and Finance." *Annual Review of Economics* 1(1): 255–94.
———. 2016. "Power Laws in Economics: An Introduction." *Journal of Economic Perspectives* 30(1): 185–206.
Geer, John G. 1991. "Critical Realignments and the Public Opinion Poll." *The Journal of Politics* 53(2): 434–53.
———. 1996. *From Tea Leaves to Opinion Polls: A Theory of Democratic Leadership*. New York: Columbia University Press.
Geer, John G., and Prateek Goorha. 2003. "Declining Uncertainty." In *Uncertainty in American Politics*, ed. Barry C. Burden. Cambridge, UK: Cambridge University Press.

Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." *Statistics in Medicine* 27(15): 2865–73.

Gerber, Alan, James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. "How Large and Long-Lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *The American Political Science Review* 105(1): 135–50.

Gerber, Alan, Gregory Huber, and Albert Fang. 2018. "Do Subtle Linguistic Interventions Priming a Social Identity as a Voter Have Outsized Effects on Voter Turnout? Evidence From a New Replication Experiment: Outsized Turnout Effects of Subtle Linguistic Cues." *Political Psychology* 39(4): 925–38.

Gerber, Alan S., Donald P. Green, and Ron Shachar. 2003. "Voting May Be Habit-Forming: Evidence from a Randomized Field Experiment." *American Journal of Political Science* 47(3): 540–50.

Gillespie, Colin. 2020. "PoweRlaw: Analysis of Heavy Tailed Distributions." https://CRAN.R-project.org/package=poweRlaw (July 12, 2023).

Ginsberg, Allen. 1986. *The Captive Public: How Mass Opinion Promotes State Power*. New York: Basic Books.

Goidel, Kirby, ed. 2011. *Political Polling in the Digital Age: The Challenge of Measuring and Understanding Public Opinion*. Baton Rouge: LSU Press.

Goldstein, Ken, and Paul Freedman. 2002. "Campaign Advertising and Voter Turnout: New Evidence for a Stimulation Effect." *The Journal of Politics* 64(3): 721–40.

Goodhart, Noah J. 1999. "The New Party Machine: Information Technology in State Political Parties." In *The State of the Parties: The Changing Role of Contemporary American Parties*, People, Passions, and Power, eds. John Clifford Green and Daniel M. Shea. Lanham, MD: Rowman & Littlefield.

Gosnell, Harold F. 1926. "An Experiment in the Stimulation of Voting." *The American Political Science Review* 20(4): 869.

Green, Donald P., and Alan S. Gerber. 2019. *Get Out the Vote: How to Increase Voter Turnout*. 4th Edition. Washington, DC: Brookings Institution Press.

Green, Donald P., and Ron Shachar. 2000. "Habit Formation and Political Behaviour: Evidence of Consuetude in Voter Turnout." *British Journal of Political Science* 30(4): 561–73.

Grossmann, Matt. 2009. "Campaigning as an Industry: Consulting Business Models and Intra-Party Competition." *Business and Politics* 11: 2–2.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: Wiley.

———. 2009. *Survey Methodology*. 2nd Edition. Hoboken, NJ: Wiley.

Haenschen, Katherine, and Jay Jennings. 2019. "Mobilizing Millennial Voters with Targeted Internet Advertisements: A Field Experiment." *Political Communication*: 1–19.

Handcock, Mark S., and Krista J. Gile. 2011. "Comment: On the Concept of Snowball Sampling." *Sociological Methodology* 41(1): 367–71.

Harris, Louis. 1963. "Polls and Politics in the United States." *The Public Opinion Quarterly* 27(1): 3–8.

Harvey, William S. 2011. "Strategies for Conducting Elite Interviews." *Qualitative Research* 11(4): 431–41.

Hatch, Rebecca S. 2016. "Party Organizational Strength and Technological Capacity: The Adaptation of the State-Level Party Organizations in the United States to Voter Outreach and Data Analytics in the Internet Age." *Party Politics* 22(2): 191–202.

Heckathorn, Douglas D., and Christopher J. Cameron. 2017. "Network Sampling: From Snowball and Multiplicity to Respondent-Driven Sampling." *Annual Review of Sociology* 43(1): 101–19.

Herbst, Susan. 1993. *Numbered Voices: How Opinion Polling Has Shaped American Politics*. Chicago: University of Chicago Press.

Herrnson, Paul S. 2004. *Congressional Elections: Campaigning at Home and in Washington*. 4th Edition. Los Angeles: CQ Press.

Herrnson, Paul S., and James G. Gimpel. 1995. "District Conditions and Primary Divisiveness in Congressional Elections." *Political Research Quarterly* 48(1): 117–34.

Herrnson, Paul S., Stacey L. Joyner, Christopher J. Deering, and Clyde Wilcox. 2013. *Interest Groups Unleashed*. CQ Press.

Herrnson, Paul S., Costas Panagopoulos, and Kendall L. Bailey. 2019. *Congressional Elections: Campaigning at Home and in Washington*. 8th Edition. Los Angeles: CQ Press.

Hersh, Eitan. 2015. *Hacking the Electorate: How Campaigns Perceive Voters*. New York: Cambridge University Press.

Hillygus, D. Sunshine, and Todd G. Shields. 2009. *The Persuadable Voter: Wedge Issues in Presidential Campaigns*. Princeton: Princeton University Press.

Huckshorn, Robert Jack. 1976. *Party Leadership in the States*. Amherst: University of Massachusetts Press.

Igielnik, Ruth, Scott Keeter, Courtney Kennedy, and Bradley Spahn. 2018. *Commercial Voter Files and the Study of U.S. Politics*. Pew. https://www.pewresearch.org/methods/2018/02/15/commercial-voter-files-and-the-study-of-u-s-politics/ (July 12, 2023).

Issenberg, Sasha. 2012. "How Obama's Team Used Big Data to Rally Voters." *MIT Technology Review*. https://www.technologyreview.com/2012/12/19/114510/how-obamas-team-used-big-data-to-rally-voters/ (July 12, 2023).

———. 2013. *The Victory Lab: The Secret Science of Winning Campaigns*. New York: Broadway Books.

Jackman, Simon, and Bradley Spahn. 2021. "Politically Invisible in America." *PS: Political Science & Politics* 54(4): 623–29.

Jacobs, Lawrence R., and Robert Y. Shapiro. 1995. "The Rise of Presidential Polling: The Nixon White House in Historical Perspective." *The Public Opinion Quarterly* 59(2): 163–95.

Jennings, Will, and Christopher Wlezien. 2018. "Election Polling Errors across Time and Space." *Nature Human Behaviour* 2(4): 276–83.

Johnson, Dennis W. 2007. *No Place for Amateurs: How Political Consultants Are Reshaping American Democracy*. 2nd Edition. New York: Routledge.

———. 2010. *Campaigning in the Twenty-First Century: A Whole New Ballgame?* New York: Routledge.

———. 2016. *Democracy for Hire: A History of American Political Consulting*. New York: Oxford University Press.

Johnston, Richard, Michael G. Hagen, and Kathleen Hall Jamieson. 2004. *The 2000 Presidential Election and the Foundations of Party Politics*. Hackensack, NJ: Cambridge University Press.

Jungherr, Andreas. 2018. "Normalizing Digital Trace Data." In *Digital Discussions*, eds. Natalie Jomini Stroud and Shannon C. McGregor. New York: Routledge, 24–54.

Jungherr, Andreas, Gonzalo Rivero, and Daniel Gayo-Avello. 2020. *Retooling Politics: How Digital Media Are Shaping Democracy*. New York: Cambridge University Press.
Keeter, Scott. 2018. "Are Public Opinion Polls Doomed?" *Nature Human Behaviour* 2(4): 246–47.
Kefford, Glenn et al. 2022. "Data-Driven Campaigning and Democratic Disruption: Evidence from Six Advanced Democracies." *Party Politics* 29(3): 448–62.
Keith, Bruce E. 1992. *The Myth of the Independent Voter*. Berkeley: University of California Press.
Kennedy, Courtney et al. 2018. "An Evaluation of the 2016 Election Polls in the United States." *Public Opinion Quarterly* 82(1): 1–33.
Kennedy, Courtney, and Hannah Hartig. 2019. "Response Rates in Telephone Surveys Have Resumed Their Decline." *Pew Research Center*. https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/ (July 12, 2023).
Kernell, Samuel. 2000. "Life Before Polls: Ohio Politicians Predict the 1828 Presidential Vote." *PS: Political Science and Politics* 33(3): 569–74.
Kim, Young Mie et al. 2018. "The Stealth Media? Groups and Targets behind Divisive Issue Campaigns on Facebook." *Political Communication* 35(4): 515–41.
Kinder, Donald R., and Nathan P. Kalmoe. 2017. *Neither Liberal nor Conservative: Ideological Innocence in the American Public*. Chicago: University of Chicago Press.
Kolodny, Robin, and David A. Dulio. 2003. "Political Party Adaptation in US Congressional Campaigns: Why Political Parties Use Coordinated Expenditures to Hire Political Consultants." *Party Politics* 9(6): 729–46.
Kreiss, Daniel. 2012. *Taking Our Country Back: The Crafting of Networked Politics from Howard Dean to Barack Obama*. New York: Oxford University Press.
———. 2016. *Prototype Politics: Technology-Intensive Campaigning and the Data of Democracy*. New York: Oxford University Press.
Kreiss, Daniel, and Christopher Jasinski. 2016. "The Tech Industry Meets Presidential Politics: Explaining the Democratic Party's Technological Advantage in Electoral Campaigning, 2004–2012." *Political Communication* 33(4): 544–62.
Kreiss, Daniel, Regina G. Lawrence, and Shannon C. McGregor. 2018. "In Their Own Words: Political Practitioner Accounts of Candidates, Audiences, Affordances, Genres, and Timing in Strategic Social Media Use." *Political Communication* 35(1): 8–31.
Lazarsfeld, Paul Felix, Bernard Berelson, and Hazel Gaudet. 1948. *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. New York: Columbia University Press.
Lepore, Jill. 2020a. *If Then: How the Simulmatics Corporation Invented the Future*. New York: Liveright.
———. 2020b. "Scientists Use Big Data to Sway Elections and Predict Riots — Welcome to the 1960s." *Nature* 585(7825): 348–50.
Limbocker, Scott, and Hye Young You. 2020. "Campaign Styles: Persistency in Campaign Resource Allocation." *Electoral Studies* 65: 102140.
Luker, Kristin. 2010. *Salsa Dancing into the Social Sciences: Research in an Age of Info-Glut*. Cambridge, MA: Harvard University Press.
Malhotra, Neil, Melissa R. Michelson, Todd T. Rogers, and Ali Adam Valenzuela. 2011. "Text Messages as Mobilization Tools: The Conditional Effect of Habitual Voting and Election

Salience." *American Politics Research* 30(4): 664–81.
Malik, Momin M. 2018. "Bias and Beyond in Digital Trace Data." Carnegie Mellon University.
Martin, Gregory J., and Zachary Peskowitz. 2015. "Parties and Electoral Performance in the Market for Political Consultants." *Legislative Studies Quarterly* 40(3): 441–70.
———. 2018. "Agency Problems in Political Campaigns: Media Buying and Consulting." *American Political Science Review* 112(2): 231–48.
Masket, Seth E. 2009. "Did Obama's Ground Game Matter?" *Public Opinion Quarterly* 73(5): 1023–39.
———. 2020. *Learning from Loss: The Democrats, 2016–2020*. New York: Cambridge University Press.
Mayer, William G. 2007. "The Swing Voter in American Presidential Elections." *American Politics Research* 35(3).
———, ed. 2008. *The Swing Voter in American Politics*. Washington, DC: Brookings Institution Press.
Mayhew, David R. 1974. *Congress: The Electoral Connection*. New Haven: Yale University Press.
McDonald, Michael P., and Samuel L. Popkin. 2001. "The Myth of the Vanishing Voter." *American Political Science Review* 95(4): 963–74.
McGregor, Shannon C. 2020. "'Taking the Temperature of the Room' How Political Campaigns Use Social Media to Understand and Represent Public Opinion." *Public Opinion Quarterly* 84(S1): 236–56.
McKenna, Elizabeth, and Hahrie Han. 2014. *Groundbreakers: How Obama's 2.2 Million Volunteers Transformed Campaigning in America*. New York: Oxford University Press.
Medvic, Stephen K. 2001. *Political Consultants in U.S. Congressional Elections*. Columbus: Ohio State University Press.
Miller, Warren E., and J. Merrill Shanks. 1996. *The New American Voter*. Cambridge, MA: Harvard University Press.
Miller, Warren E., and Donald E. Stokes. 1963. "Constituency Influence in Congress." *The American Political Science Review* 57(1): 45–56.
Monson, Joseph Quin. 2004. "Polling in Congressional Election Campaigns." The Ohio State University.
Nickerson, David W., and Todd Rogers. 2014. "Political Campaigns and Big Data." *Journal of Economic Perspectives* 28(2): 51–74.
Nielsen, Rasmus Kleis. 2012. *Ground Wars: Personalized Communication in Political Campaigns*. Princeton: Princeton University Press.
Niven, David. 2004. "The Mobilization Solution? Face-to-Face Contact and Voter Turnout in a Municipal Election." *The Journal of Politics* 66(3): 868–84.
Nyhan, Brendan, and Jacob M. Montgomery. 2015. "Connecting the Candidates: Consultant Networks and the Diffusion of Campaign Strategy in American Congressional Elections." *American Journal of Political Science* 59(2): 292–308.
Ospina, Raydonal, and Silvia L. P. Ferrari. 2008. "Inflated Beta Distributions." *Statistical Papers* 51(1): 111.
Overby, L. Marvin, and Jay Barth. 2006. "Radio Advertising in American Political Campaigns: The Persistence, Importance, and Effects of Narrowcasting." *American Politics Research* 34(4): 451–78.

Panagopoulos, Costas. 2009. “Partisan and Nonpartisan Message Content and Voter Mobilization: Field Experimental Evidence.” *Political Research Quarterly* 62(1): 70–76.

———. 2016. “All About That Base: Changing Campaign Strategies in U.S. Presidential Elections.” *Party Politics* 22(2): 179–90.

Panagopoulos, Costas, and Donald P. Green. 2008. “Field Experiments Testing the Impact of Radio Advertisements on Electoral Competition.” *American Journal of Political Science* 52(1): 156–68.

———. 2011. “Spanish-Language Radio Advertisements and Latino Voter Turnout in the 2006 Congressional Elections: Field Experimental Evidence.” *Political Research Quarterly* 64(3): 588–99.

Parry, Janine, Jay Barth, Martha Kropf, and E. Terrence Jones. 2008. “Mobilizing the Seldom Voter: Campaign Contact and Effects in High-Profile Elections.” *Political Behavior* 30(1): 97–113.

Plutzer, Eric. 2002. “Becoming a Habitual Voter: Inertia, Resources, and Growth in Young Adulthood.” *American Political Science Review* 96(1): 41–56.

Prosser, Christopher, and Jonathan Mellon. 2018. “The Twilight of the Polls? A Review of Trends in Polling Accuracy and the Causes of Polling Misses.” *Government and Opposition* 53(4): 757–90.

Roll, Charles. 1982. “Private Opinion Polls.” *Proceedings of the Academy of Political Science* 34(4): 61–74.

Rossi, Peter H., James D. Wright, and Andy B. Anderson. 2013. *Handbook of Survey Research*. Academic Press.

Sabato, Larry. 1981. *The Rise of Political Consultants: New Ways of Winning Elections*. New York: Basic Books.

Samuelson, P. A. 1938. “A Note on the Pure Theory of Consumer’s Behaviour.” *Economica* 5(17): 61–71.

Schaffner, Brian F. 2006. “The Political Geography of Campaign Advertising in U.S. House Elections.” *Political Geography* 25(7): 775–88.

Selnow, Gary W. 1993. *High-Tech Campaigns: Computer Technology in Political Communication*. Westport, CT: Praeger.

Shaw, Daron R. 2008. “Swing Voting and U.S. Presidential Elections.” In *The Swing Voter in American Politics*, ed. William G. Mayer. Washington, DC: Brookings Institution Press.

Shea, Daniel M. 1996. *Campaign Craft: The Strategies, Tactics, and Art of Political Campaign Management*. 1st Edition. Westport, CT: Praeger.

———. 2001. *Campaign Craft: The Strategies, Tactics, and Art of Political Campaign Management*. 2nd Edition. Westport, Conn: Praeger.

Sides, John, Michael Tesler, and Lynn Vavreck. 2018. “Hunting Where the Ducks Are: Activating Support for Donald Trump in the 2016 Republican Primary.” *Journal of Elections, Public Opinion and Parties* 28(2): 135–56.

Sides, John, and Lynn Vavreck. 2013. *The Gamble: Choice and Chance in the 2012 Presidential Election*. Princeton, NJ: Princeton University Press.

Simon, Felix M. 2019. “‘We Power Democracy’: Exploring the Promises of the Political Data Analytics Industry.” *The Information Society* 35(3): 158–69.

Smidt, Corwin D. 2017. “Polarization and the Decline of the American Floating Voter.” *American Journal of Political Science* 61(2): 365–81.

Smith, Tom W. 1990. “The First Straw?: A Study of the Origins of Election Polls.” *The Public Opinion Quarterly* 54(1): 21–36.

Spencer, Douglas M., and Bertrall L. Ross. 2019. “Passive Voter Suppression: Campaign Mobilization and the Effective Disfranchisement of the Poor.” *Northwestern University Law Review* 114(3): 633–704.

Steeh, Charlotte G. 1981. “Trends in Nonresponse Rates, 1952-1979.” *The Public Opinion Quarterly* 45(1): 40–57.

Stonecash, Jeffrey M. 2008. *Political Polling: Strategic Information in Campaigns*. 2nd Edition. Lanham, MD: Rowman & Littlefield Publishers.

Teele, Dawn Langan. 2014. *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*. Yale University Press.

Theocharis, Yannis, and Will Lowe. 2016. “Does Facebook Increase Political Participation? Evidence from a Field Experiment.” *Information, Communication & Society* 19(10): 1465–86.

Traugott, Michael W. 2005. “The Accuracy of the National Preelection Polls in the 2004 Presidential Election.” *Public Opinion Quarterly* 69(5): 642–54.

Trish, Barbara. 2018. “Big Data under Obama and Trump: The Data-Fueled U.S. Presidency.” *Politics and Governance* 6(4): 29–39.

Tsagris, Michail, Abdulaziz Alenazi, and Connie Connie Stewart. 2023. “Flexible Non-Parametric Regression Models for Compositional Response Data with Zero.” https://doi.org/10.21203/rs.3.rs-2006067/v1 (July 12, 2023).

Tsagris, Michail, and Connie Stewart. 2018. “A Dirichlet Regression Model for Compositional Data with Zeros.” *Lobachevskii Journal of Mathematics* 39(3): 398–412.

Van Duyn, Emily. 2018. “Hidden Democracy: Political Dissent in Rural America.” *Journal of Communication* 68(5): 965–87.

Verba, Sidney, and Norman H. Nie. 1987. *Participation in America: Political Democracy and Social Equality*. Chicago: University Of Chicago Press.

Wayne, Stephen J. 2019. *The Road to the White House 2020*. 11th Edition. Boston, MA: Cengage Learning.

Weisberg, Herbert F. 2005. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. Chicago: University of Chicago Press.

Weiss, Michael J. 1988. *The Clustering of America*. New York: HarperCollins.

Williams, Christine B., and Girish J. “Jeff” Gulati. 2018. “Digital Advertising Expenditures in the 2016 Presidential Election.” *Social Science Computer Review* 36(4): 406–21.

Williams, Christine B., Jeff Gulati, and Mateusz Zeglen. 2020. “Following the Money: Uses and Limitations of FEC Campaign Finance Data.” *Interest Groups & Advocacy* 9: 317–29.