O'REILLY®

Data and Democracy

How Political Data Science Is Shaping the 2016 Elections

Andrew Therriault

Strata+ Hadoop WORLD

Make Data Work

strataconf.com

Presented by O'Reilly and Cloudera, Strata + Hadoop World helps you put big data, cutting-edge data science, and new business fundamentals to work.

- Learn new business applications of data technologies
- Develop new skills through trainings and in-depth tutorials
- Connect with an international community of thousands who work with data

Job # D2044

Data and Democracy

How Political Data Science Is Shaping the 2016 Elections

Andrew Therriault

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Data and Democracy
by Andrew Therriault

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editor: Marie Beaugureau
Production Editor: Dan Fauxsmith
Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

August 2016: First Edition

Revision History for the First Edition
2016-08-31: First Release

978-1-491-95905-3

[LSI]

Table of Contents

Prologue: The Role of Data in Campaigns

Andrew Therriault

Andrew Therriault was the Democratic National Committee's Director of Data Science from 2014 to 2016, leading a team that developed voter targeting models and other analytic tools used by thousands of Democratic campaigns.

Watching news coverage of the 2016 presidential race, it's easy to imagine that this year's election is unlike any other that's come before. Behind the scenes, though, this year's campaigns are continuing a trend that has been underway for more than a decade. Modern campaigns in the US have become more reliant on data with each election cycle, using analytics to drive everything from overall strategy and messaging to individual voter contacts and advertising. This evolution was well documented in the 2008 and 2012 presidential campaigns, and in 2016, the technology has spread to "downballot" campaigns for state and local offices that never before had access to such sophisticated tools.

At the same time, a data-driven mindset is also shaping how campaigns are covered in the media. To bring sanity to the deluge of often-conflicting poll numbers, survey aggregators use forecast models to combine results and produce a clearer estimate of where each contest stands. And by digging deeper into filings that record how money flows through each campaign, investigative journalists can better understand what's going on both within the campaign and across the electorate.

This compilation brings together leading experts on the role of data in electoral politics, to share insights learned across nearly a century of combined experience. These contributors span a range of technical areas and bring together perspectives from Democrats, Republicans, and members of the media. What the contributors have in common, though, is that they are all highly skilled practitioners who have an intimate understanding of the technical challenges in their fields. Bringing these voices together to explain political data for a tech-savvy audience, our goal is to leave readers with a more detailed and comprehensive understanding of the topic than has ever before been gathered in one publication.

A Primer on Modern Election Campaigns

Before we begin, it'll help to lay out a common framework in which elections take place, to anchor the discussions that follow. For all the drama and complexity surrounding election contests, every part of a campaign's work is ultimately designed to impact the same thing: the final vote totals. Working backwards from there, this gives a campaign two avenues for swaying the outcome: first, influencing who shows up to vote, and second, affecting how they vote when they get there. All of the other numbers in a campaign—fundraising totals, ad airings, poll results, etc.—ultimately matter mainly because of their connection to who votes and how.

In terms of overall strategy, then, campaigns' options are pretty straightforward: they want to convince eligible voters to support their candidates, then they want to make sure those supporters show up to vote. Almost everything else on a campaign is built around these goals. From firm supporters who are certain to vote, campaigns solicit donations and recruit volunteers to help with persuasion and get-out-the-vote efforts aimed at other voters. Campaigns then use both volunteers and paid staff to reach out to other voters directly, and develop messaging strategies and advertising campaigns to reinforce these efforts. And as Election Day approaches, they rely on polling and forecast models to allocate resources and adapt their tactics to fit the latest political environment.

The starting point for all data on a modern campaign is a voter file. This is a list of all registered voters in a given state or district, compiled by the local and state election authorities. A typical voter file contains voters' contact information, basic demographic characteris-

tics, historical voter turnout records, and (in most states) partisan affiliation. From there, campaigns supplement voter files with their own data on things like contact history and donations, as well as external data from other campaigns, public records, and commercial data vendors.

To reach these voters, campaigns develop messaging strategies based on polling and other research. These messages are sometimes tested and refined through experimental testing, and then the final messages are turned into specific ads and talking points. Those campaign-produced ads (often referred to as "paid media") are customized for TV, radio, print, mail, and online usage, and reinforced through talking points used in interviews and other news coverage ("earned media") that the campaign gets for free.

As the election season heats up, media sources also do their own research into campaigns' activities and performance. The most savvy of these journalists use many of the same techniques campaigns use, compiling large datasets and building advanced models to better understand the state of each election. Over time, these media not only provide new information to their audiences, but also help shape perceptions of the campaigns and candidates by translating complex patterns into findings that are easier for the public to digest.

What to Expect

Over the coming pages, you'll hear firsthand accounts of what this work looks like in 2016, examples of the many challenges practitioners still face, and predictions about how political data science will continue to evolve in the coming years. Taken together, these pieces illustrate just how essential data has become to today's elections. To be sure, this trend echoes that seen in many other fields, but the rapid pace and high stakes of politics have inspired a degree of innovation that few other industries can match.

No matter how much you're interested in politics for its own sake, these contributors' stories will likely offer insights that are relevant to your own areas of expertise. As editor, I hope that you all get as much satisfaction out of reading these great pieces as I did from assembling them. Enjoy!

CHAPTER 1

Essentials of Modeling and Microtargeting

Dan Castleman

Dan Castleman is cofounder and Director of Analytics at Clarity Campaign Labs, a modeling and targeting consulting firm for progressive candidates, political coalitions, corporations, international campaigns, and charitable groups. His expertise covers strategic program targeting, data analysis, modeling, and polling. One of the leading modelers in Democratic politics, Castleman has built voter targeting models for the Democratic National Committee, Democratic Congressional Campaign Committee, and Democratic Governors' Association, as well as dozens of individual campaigns and advocacy groups.

In the Democratic Party today, modeling and microtargeting have become ubiquitous. They are no longer peculiar curiosities confined to only a few forward-thinking campaigns, but de facto parts of any modern campaign with sufficient resources. While this leap forward can at least partly be attributed to coverage of their use by the Obama campaign in 2012, individual-level models have been used by Democratic campaigns since as early as 2004. Indeed, in many ways modeling and microtargeting are simply an extension of tactics employed to target voters at the aggregate level for decades. Over the past ten years, I have built countless models and advised dozens of political campaigns in how to use them. In that time, I have seen firsthand their rise in use, what they can (and can't) do, and what methods for building them work the best.

Although the two terms are often used interchangeably, "modeling" and "microtargeting" have some important distinctions. Modeling is the practice of using algorithms and observed data to build statistical or machine learning models, in order to predict unobserved actions or preferences. Modeling for political campaigns is most often done at the individual level using voter files, which combine official voter registration lists with other data sources, such as consumer records, campaign contact histories, and other proprietary information collected by the campaign and party.

Microtargeting, meanwhile, refers to the process of making campaign targeting decisions at the individual level—for instance, selecting which voters to mail campaign literature to—and is generally informed by these voter models. While these practices are clearly intertwined, they are not inseparable. Modeling can still be done on other types of data (e.g., collections of exit polls or precinct-level vote totals), and microtargeting can be done in the absence of models (e.g., making selections based on characteristics included in voter files or consumer data).

What Models Can and Cannot Do

The practice of using models and microtargeting has made a profound and important impact on how campaigns are run. But despite what is often said and written about them, they do not offer unpopular politicians any magic ability to win over electorates that dislike them. And while they can provide an accurate snapshot of the electorate's overall preferences, they are not an efficient way to generate the horse-race numbers that the media and public are most interested in. What they can do—and what we ultimately seek most from them—is to improve the efficiency of how campaigns run their programs.

While *efficiency* sounds like a rather mundane goal, for political campaigns it is of the utmost importance. Campaigns have to gain the support of a majority of often-large electorates, but are constrained in terms of money, volunteers, and the ever-looming election date. To help use these limited resources as wisely as possible, campaigns have to be smart about how they target voters. For example, looking at the breakdown of survey responses across subgroups can give a campaign an idea of the types of voters who are potential supporters or areas where a candidate's popularity is weak. The same

can be done historically by examining past election results at the precinct level. Modeling allows us to apply a more robust methodology to identify targets and provides even greater efficiency gains when applied at the individual level.

Selecting Voters to Target

Since campaigns' targeting needs drive their modeling needs, decisions about what models to build are based on larger strategic assumptions. When developing their overall strategies, campaigns typically think about voters in two main dimensions: support for their candidate and likelihood of voting. For campaigns it is not just important to identify whom voters support and whether they are persuadable, but also whether or not they will take their support to the polls. Figure 1-1 demonstrates how these two measures are connected and further divides voters into the three groups on which campaigns will want to focus their resources.

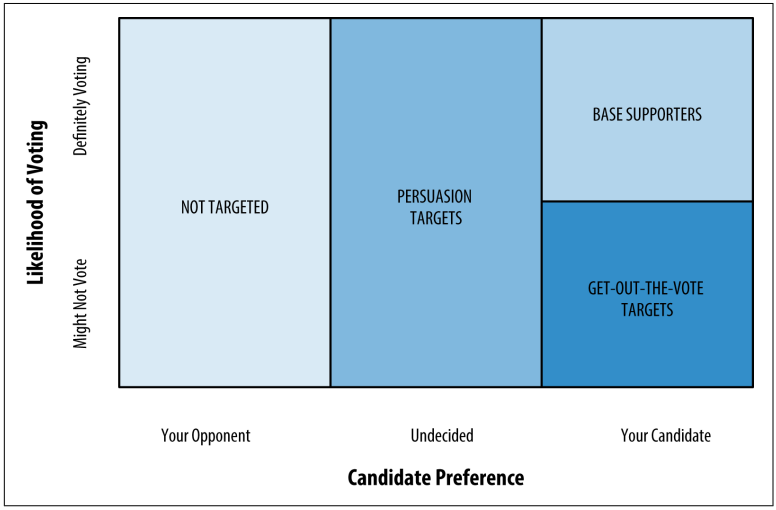

Figure 1-1. Basics of voter targeting in campaigns

When targeting voters who are highly likely to turn out and support their candidate ("base supporters"), campaigns will seek financial contributions and recruit volunteers. Those with high support but uncertain likelihood of voting become targets for "get-out-the-vote" efforts to increase their turnout. And those with uncertain support will be targeted for persuasion. For the remaining voters—those highly likely to support the opponent—a campaign's limited resour-

ces are best spent elsewhere, as these voters are not very likely to respond to the campaign's efforts.

This strategic thinking explains why turnout and support scores are the most common (and in many cases, the most useful) models for campaigns, but there are many other types of models that can also be valuable for campaigns. Table 1 lists some of the most common types of voter models, many of which are available pre-built and scored nationally by firms like Clarity.

Table 1-1. Common types of voter targeting scores

Type of Score	Description
Party	Likelihood of identifying with one political party or another
Turnout	Likelihood of voting in a given election
Candidate Support	Likelihood of supporting a specific candidate
Issue Position	Likelihood of holding a given position on a specific issue (marriage equality, gun control, etc.)
Volunteer / Donation Propensity	Likelihood of being a campaign activist or donor
Demographic / Behavioral	Likelihood of being a college graduate, gun owner, Fox News watcher, etc.
Persuasion	Likelihood of being responsive to campaign persuasion

How Models Are Made

Most models are built using survey data—and lots of it.[1] For example, a candidate support model will be informed by a short survey with a large sample size. Unlike traditional "horse-race" election polls (like those put out by newspapers and TV networks), which usually include between 400 and 1,000 responses, modeling surveys typically use anywhere from 2,500 to 20,000 or more responses, depending on the size of the geography and the type of model. A support model can be built on something as simple as a head-to-head candidate support question, and the responses used directly as the model's outcome. For more complex characteristics, such as a

1 Turnout models are the most notable exception: the standard approach is simply to model turnout in comparable past elections (as recorded in official voter files) to make inferences about who is most likely to turn out in the next election. Other models, such as those for donation or volunteer propensity, can also be built on similar directly observed behaviors.

voter's persuadability, models may use multiple questions in aggregate or estimate treatment effects from randomized experimental tests.

The end result of most models is a "score" that we can append to the voter file used by the campaign for targeting. Most scores represent estimated probabilities (of supporting a candidate or turning out to vote, for example), but others may give percentile or decile rankings, especially for outcomes that are unlikely to be well calibrated (often because they are built on low-frequency events, such as volunteering or donation).

As mentioned before, models are not magic. They can help *find* persuadable voters, but they cannot *make* voters persuadable or make a poor message more effective. And while the models we build get increasingly accurate every year, they will always be probabilistic, which means they will always have some degree of error. For example, appended consumer records can improve models, but they are far less informative than you might think: the data points provided are often sparse, vaguely sourced, or just plain wrong.

This uncertainty affects the methods and algorithms we use. With hundreds or thousands of potential features to include, the most important step in modeling is often the feature selection process. Furthermore, the machine learning algorithms that work the best for our needs are ones that scale well with a large number of disparate features and can discover nonlinear patterns. Tree-based models are most common, particularly as ensembles of trees (such as random forests) or when ensembled with other types of models. But every campaign's needs are different, and even more rudimentary algorithms, such as linear and logistic regressions, are still used for certain models.

Like much of the data science community, the political modeling field is increasingly turning to open-source tools such as R and Python, although some practitioners use commercial packages like Stata. In addition to developing a community of data scientists with expertise with these tools, we have also invested resources in building systems that allow us to work with large datasets efficiently and quickly, in order to create and score models in a streamlined fashion. These systems combine custom ETL and analytics tools with terabyte-scale data warehouses of voter data, allowing us to update models frequently so that the already-short timelines of campaigns

are not held up by our processes. After all, our models are only useful to a campaign if they have time to employ them, so efficient turnaround is key to providing value.

Looking ahead, there are still new techniques to be explored, methodologies to be improved, and challenges to overcome. For example, we have been working on enhancing our methods for providing accurate and useful models for not just presidential and statewide campaigns but for smaller campaigns as well (such as those for state legislature or local offices). To help these downballot campaigns in 2016, the Democratic National Committee has put together a suite of "off the shelf" models available to all their campaigns nationwide, and we at Clarity have been working on finding ways for multiple smaller campaigns to pool their resources and develop customized models that wouldn't otherwise be affordable or feasible. In this unusual election cycle, we are constantly striving to improve, adapt, and understand the implications of the ever-changing political winds on modeling for all races up and down the ticket.

CHAPTER 2

Data Management for Political Campaigns

Audra Grassia

Audra Grassia leads political engagements at Civis Analytics. She has spent the majority of her career working on political campaigns across the country, including statewide, local, and federal races. Additionally, Audra was the Deputy Political Director for the Democratic Governors Association during the 2014 election cycle.

"Big Data in politics is just a fad!…said no one. Ever."

This is the joke a non-data, political colleague of mine recently made while we were discussing the future of electoral politics. Her joke was correct in its sarcastic tone: we all know big data in electoral politics is here to stay. But it wasn't too long ago when some believed it would fade into the background as the next new, shiny thing came on the market.

The reality is that data is now in the background—and the foreground. It now informs every campaign decision and is talked about by pollsters, pundits, and even candidates. The level of granularity and the breadth of insight that one can produce using all of the data at our disposal is unmatched, and we are rapidly growing our understanding of who voters are, how to reach them, and what we say to them. Our ability to draw actionable insights from our data has been enhanced as data and analytics technology becomes more available, sophisticated, user-friendly, and decision-maker centric.

Just because we now have access to more information about voters than ever before, though, does not mean it is easy to use all of that data to inform decisions. In this article, I will discuss three broad challenges when using data to inform decisions in a campaign environment, and suggest ways to think about data management infrastructure—from the outset—in order to minimize those challenges.

The Challenge: Building a Smart, Scalable, and Compatible Data Infrastructure

First, there is the people gap. Managing, using, and understanding data takes technical training, and no matter how many people we put through data-training programs, the dearth of well-trained, skilled data scientists and analysts working in politics is not likely to change rapidly. Additionally, political decision-makers are becoming increasingly sophisticated in their consumption of data and analytics products. They are asking more interesting and involved research questions and expecting answers faster. So while demand for data and analytics services increases, the supply of skilled talent is slow to follow.

Second, there are disparate sources of data, all structured in different ways, living in different systems that don't talk to one another. This challenge isn't unique to campaigns. However, the speed and efficiency with which we need to unify and make use of that data is tied to a hard deadline (Election Day), which (almost) never changes.

Finally, given the first and second challenges, there is not a well-defined process for directing the right data to the right stakeholders. It will take a few more cycles before incorporating data and analytics into all aspects of a campaign becomes seamless and replicable. By then, there will undoubtedly be several more data sources to integrate, which is all the more reason to build a data infrastructure that is scalable and smart from the outset.

Seizing Opportunity: Five Ways to Maximize Use of Data on a Campaign

Thinking smartly about data management infrastructure can greatly enhance a campaign's ability to overcome challenges and ultimately contributes to a successful campaign.

To provide a bit of context, a data-driven culture encourages collaboration between previously siloed departments. What happens when a finance director sees that field data can help build a model that will increase their success at fundraising? They'll want more field data, and the campaign will raise more money. What happens when communications sees that their messages aren't resonating with digital donors? They'll want to figure out how to adjust their message—and then they will want to test it.

Good data management infrastructure is key to building a unified vision of campaign efforts in order to fully understand how resources are being allocated, where those resources are being most effectively used, and when/where resources might need adjustment.

So what should you consider when planning your campaign data infrastructure? Ideally, decision-makers will consider five factors as they are building a campaign, but the principles behind them can be revisited as a campaign (and the data) grows:

1. **Let each tool (and person) do what it does best.**

   On any campaign, there are a number of tools meant to both manage and use data, with specific software designed for fundraising, digital outreach, regulatory compliance, accounting, and field management. There is no single system that will accomplish what you need for all users and use cases. Therefore, the question becomes, what tools are best for their users and can interact through syncing? Having multiple tools for collecting or managing data isn't a problem as long as there is a way to then aggregate, analyze, and report on all of the data in order to answer key questions.

   Thinking about your data management tools as Legos is a helpful way to consider this problem. Imagine you have a tub of plastic Legos, and in that tub are a few oddly shaped wooden blocks. While the wooden blocks may add some interesting dimensions to whatever you are trying to build, they ultimately won't click into place with the other Legos, and you'll be left with a structure that needs to be cobbled together in a less than ideal way. This is the same danger you face if you don't consider interoperability between your systems before you decide to use them.

2. **Distribute access to your data and data tools.**

   Let's take the toy analogy one step further. Now imagine that Legos are all individual pieces of data, and you have a big clear tub full of them. They are all different colors, shapes, and sizes. Individually, and even all jumbled up in the box, those Legos are not particularly useful or interesting. The clear tub allows you to view those Legos, but if you cannot get the Legos out to build something, then what's the point? Data is no different. There are many systems that allow you to input information. Many of those systems will even help you enhance the data from other sources (e.g., consumer data sources). What happens next is where things get tricky.

   Some systems allow you to take your data—in almost any form you decide is best—and do whatever you need to with it, pending any legal/contractual limitations. Other systems severely restrict access to the raw data and instead provide the information in the form of hard-to-manipulate reporting tools. Depending on your organizational needs, the more restrictive version can limit the usefulness of the data. It prevents easy unification with other data systems and hampers the ability for people within your organization to find creative ways to conduct analysis. It is like locking the lid of the clear tub holding the Legos and only giving the key to the system vendor.

   There was a time when this model made sense, when what we could do with our data was as limited by bandwidth and server size as it was by the lack of trained analytics professionals. Leaders ten years ago didn't grow up in a culture of data and were therefore only asking questions that could be answered with the limited amount of available information. Simple reports helped them make informed decisions. But as the role of data in society grows and as we have infinitely more access to infinitely more data, using a system that is severely restricted puts organizations at a disadvantage. Instead, seek out vendors that provide APIs, webhooks, or backend SQL database access. Direct access to underlying data will give you flexibility as your organization scales.

3. **Get your data together and automate.**

   Imagine that you just hired your first data director. She has already proven her value ten-fold by producing interesting insights visualized in a way that your entire team can really

understand. You wish you could clone her, but we are at least a few decades away from such technological advancement. Given that reality, finding ways to automate all of the "data munging" tasks she needs to do *before* she gets to drawing insights is the most efficient way to exponentially increase a smart analyst's time and value to the organization.

Automation is often one of the most difficult things to achieve. As discussed in #2, you might have a number of different tools you are working with to achieve specific tasks. Understanding the compatibility of the job-specific tools with one another and your overarching analytics platform is important. To the greatest extent possible, you want to use tools that allow you to automate rote tasks (e.g., syncing data to and from given sources, cleaning data, reporting).

4. **Create clear standard operating procedures (SOPs) for each department's use and handling of data, and ensure those SOPs work together.**

   While you might have different tools you use to capture data from different sources or departments, put together a system that is consistent across the organization. A fundraiser should be using the field tool if they go out canvassing, while field staff should be using the fundraising tool if they are making donor calls. Managers should be aware of the data management tools from other departments and, for functions that go across departments (e.g., digital fundraising), deciding at the outset the business rules around which system should be used and how the data should flow.

   You might, for example, have a digital platform that will allow email blasts to be sent out for recruiting volunteers and for raising money. The platform should make sure action taken by potential volunteers is shared with the field team and donation information is shared with the finance team appropriately.

   Any SOPs should start with the department heads, as they understand the best and most practical ways their tools should be used. However, someone with organizational oversight (e.g., an operations director or data director) who understands the need for departments to work together should ultimately ensure that the tools and methods used will allow for the greatest integration of data down the road.

5. **Listen to all levels of the campaign, not just the leadership.**

   Innovation is often driven from the bottom rather than the top. People in leadership might be too far removed from the day-to-day of data management and analytics to really understand the added value of doing things differently. As leaders, it's important to empower people who are responsible for generating insights from data to find faster, cheaper, and better ways to answer the core questions of a campaign. While some methods might require investments, good ideas can come from anywhere. Facilitating a creative culture that listens to and entertains innovative solutions will help build a collaborative environment where everyone feels invested in the data infrastructure.

Applying These Principles on Real Campaigns

Ultimately, a few critical choices and habits can make the difference between just having a lot of data and being a data-driven campaign or organization. Adopting as many of these five principles as possible when building your data management infrastructure will increase your likelihood of achieving the latter, and a data-friendly culture needs to be adopted from the ground up and championed from the top down.

Fortunately, you don't have to recreate the wheel. These same challenges have been dealt with by campaign managers and staff for more than a decade, and there are tools available that can make their solutions' implementation, adoption, and collaboration easier. If you're not sure where to start, seek out the advice of those who have done this in the past. Many of those who faced these same challenges in previous election cycles are now working to help others in the same position, and their knowledge and experience can help future campaigns make smart decisions from the outset and avoid having to learn these lessons the hard way.

CHAPTER 3

How Technology Is Changing the Polling Industry

Patrick Ruffini

Patrick Ruffini is a cofounder of Echelon Insights, a survey research and analytics technology company. He has been applying data and technology to political campaigns for more than a decade. Ruffini helped establish one of the first full-fledged digital operations in Presidential politics for Bush-Cheney '04, led digital strategy for the Republican National Committee, and later founded and grew Engage, one of the leading digital agencies on the right.

In recent elections, high-profile examples of polling missing the mark have people asking whether traditional polling is as reliable as it used to be. These questions reached a fever pitch after the UK's recent vote to leave the European Union, with immediate pre-referendum polls pointing to a win for remaining in the EU; the 2015 UK general election, where nearly every pollster missed the Conservative Party's comfortable win; and the stunning 2014 primary loss of House Majority Leader Eric Cantor, who had led by as many as 30 points in pre-election polling.

These questions come at a time when it is getting much harder and more expensive to reach a representative sample of adults over the telephone, the dominant way of reaching survey respondents for decades. Response rates have declined from 36% in 1997 to just 9% in 2012, according to the Pew Research Center. Additionally, respondents increasingly only use cell phones, which are twice as

expensive to reach, since the government mandates that these numbers be dialed manually. The increased cost of calling mobile devices disproportionately affects access to young people between 18 and 29 years old and Hispanics, two groups where more than 60% of the population doesn't even have a landline telephone, according to statistics compiled by the Centers for Disease Control.

These difficulties are leading consumers of polling to look for alternatives, from Internet surveys to digital sentiment analysis to big data analytics. The polling industry as we know it today will look very different ten years from now, and we already see the transformation underway in this year's elections.

Despite Challenges, Polling Isn't Dead Yet…

First, the good news: Despite highly publicized "misses" and the challenges facing the industry, there is little evidence that polling has gotten less accurate over time. This primary season, polls have largely been on the mark. The Republican polls showed a strong lead for Donald Trump throughout. When analysts strayed from the raw polling numbers, conjecturing that Trump would wither under a barrage of establishment attacks, they were usually wrong. On the Democratic side, polls showed a stable and consistent national Hillary Clinton lead with few surprises in state-level contests.

The polling industry has also adapted, conducting more interviews via cell phone or online. Internet polling in particular has proven very popular as a way to counteract the rising costs of dialing cell phones. In the Republican primary, Internet polling rose 257% over 2012, counteracting a 15% drop in live telephone polling and contributing to an overall 37% increase in national polling, according to an analysis by The Huffington Post. Nearly half of all national polls in the 2016 primary race were Internet polls.

…But Understanding *How* to Poll Is More Essential than Ever

Recently, we've been seeing a rise in state-level media polls conducted using phone numbers collected from state voter rolls, a best practice used by campaign pollsters. Calling phone numbers at random, or random-digit dialing, has been the traditional way of ensuring that polls are representative of the general population. This

method is time-tested and is still often used in national surveys, but can fall short in low-turnout primaries when knowledge of the respondent's past participation matters more. Polls were most off the mark this year on the Republican side in Iowa, where just 6% of Iowa adults voted, when many pollsters dialed randomly or didn't know who had attended caucuses in the past.

Data from voter files is being used to update traditional methodologies for assessing who is and isn't likely to vote. Rather than only asking voters if they are likely to vote, we can see whether they actually have in the past. Then, rather than discard data from unlikely voters in a "likely voter" survey as pollsters traditionally do, we simply weight according to the share of sporadic voters traditionally in the electorate. In our pre-primary survey in South Carolina, we found that those with less than a 50% chance of participating supported Donald Trump at 36%, while those 95% likely to turn out supported him at 26%. In no turnout cohort did Trump lose. By appropriately weighting based on a voter's likelihood to participate, we were able to accurately project Trump's winning margin within 1%, while also releasing multiple sets of expected results based on high, medium, or low turnout. This approach gives consumers of polls a more accurate, scenario-based understanding of the interplay between turnout and election outcomes.

The Shift to Online Interviewing: Possibilities and Pitfalls

The new paradigm of online polling has led to exciting successes, and inevitably, novel challenges. Many of the reasons behind the shift to online polling are pragmatic: they cost less, typically have larger sample sizes, and have proven they can be accurate in forecasting election results. For example, the online survey technology company SurveyMonkey correctly forecast the 2014 US midterms and 2015 UK elections—which other pollsters missed—by randomly sampling users of its survey tools, people doing everything from filling out customer satisfaction surveys to choosing times for their next book club meeting.

Some online surveys are conducted of "non-probability" panels, meaning that respondents opt-in to take surveys and that every member of the public does not have an equal probability of being included (a basic tenet of traditional survey research). Yet, even

those Internet panels performed better than most traditional pollsters in 2012. In a review compiled by Nate Silver, four of the seven most accurate pollsters in 2012 conducted interviews online. By contrast, only one of the bottom ten performers used online polls exclusively, while another used online panels as a supplement to automated landline surveys.

While online polling is getting better every year, it isn't perfect. The number of people who take online surveys is rarely large enough to complete a reliable survey in Congressional or local elections. There are also vast differences in the composition of panels on lines of educational attainment and political engagement. Some panels are skewed heavily in the direction of more engaged citizens (who have actively volunteered to take surveys). This is true even if they're representative of the demographics of the country. A recent Pew Research Center study found that one company in particular (YouGov) was able to achieve significantly reduced bias and increased accuracy by weighting to education and civic participation in addition to traditional demographic variables.

For firms like Echelon Insights, the reasons to use online polling extend beyond capturing younger demographics or electoral accuracy. Much of the value in online surveys is in opening up new possibilities for the types of questions we can ask, and how.

Most polling is not about elections, and even the political polling done for election campaigns is not primarily about knowing who's winning. Campaigns and issue organizations poll to know *how* to win the argument. They poll to understand which messages will persuade voters. Properly breaking down and framing the argument—and wording questions accordingly—is critical to this task.

Additionally, the political industry has a "big data, small content" problem. Campaigns are ever-more precise in their targeting, but not always as rigorous as they should be at validating message effectiveness. Voter files that have been matched to consumer databases often come with hundreds or even thousands of variables that we can use to build target segments. And while much effort has been expended on discovering new niche audiences, campaigns often fail to close the loop by proving that a given message or ad actually works in the field. New techniques made possible through online surveys are closing this gap.

In fact, Internet polling may be uniquely suited to understanding what resonates with different populations. Telephone surveys can easily fatigue respondents forced to sit through lengthy 20-minute interviews designed to tease out every angle of an issue. The same questions can often be asked more efficiently online. Survey respondents can also rate campaign ads second-by-second online, in a mechanism similar to dial tests in focus groups, except that technology allows us to show this message to hundreds of people rather than the dozen or so in a physical focus group. At our firm, we've applied this same concept to any piece of text, letting respondents highlight text they like and strike through text they don't. Methods like these—which aren't possible over the phone—help organizations surgically hone their message before committing limited resources to TV ad production and paid media.

The Future as a Hybrid

We've heard the case for gloom and doom in the polling industry: response rates are plummeting, and new techniques aren't reliable enough yet. But upon closer inspection, this narrative doesn't hold water: as the industry's methods have shifted, accuracy hasn't suffered. If anything, these shifts have been a necessity to defend against the declining accuracy of what the industry used to do: call voters at random on landline phones.

Furthermore, the scope of what we can learn about public opinion and human behavior from data is growing daily, and what we think of as "polling" is only a tiny part of it. Voters exhibit behaviors we can learn from even when not in a voting booth or talking to an interviewer. Often, the outward signs of this are digital: typing their questions into the Google search box or sharing what they think on Twitter and Facebook.

Social media and search trends data is not truly representative of the general population, sure, but it can help us go places traditional methods can't. Day-of-the-election Google search share caught late swings to surging candidates and was accurate to within a few percentage points in early Republican primary contests. Language used in online conversation can help us craft better surveys by making sure our questions are reflective of how people talk about the issues. Media analytics, measuring the rate at which both traditional and new media sources mention candidates, were the canary in the coal

mine for Donald Trump, with a surge in media coverage and online interest preceding his initial rise in the polls and his improbable march to the nomination.

The polling industry in the future will be informed by the time-tested methods of the past, but will increasingly bring in newer disciplines, from modeling and statistics, to capturing and making sense of mountains of real-time digital data. Technology disruption in the polling industry isn't just coming. 2016 shows it's well underway.

CHAPTER 4

Data-Driven Media Optimization

Alex Lundry

Alex Lundry is the cofounder and Chief Data Scientist at Deep Root Analytics. He is one of the country's leading experts on media and voter analytics, electoral targeting, and political data mining, and has directed the data science efforts of two presidential campaigns.

It shouldn't surprise anyone that the Fox News Channel is a good place to find and advertise to Republican voters. But what may surprise you is the degree to which Republican campaigns rely upon Fox News for their television positioning. In the 2014 Texas Republican primary, nearly two-thirds of Republican advertisements on cable occurred on the network, and in the race for Lieutenant Governor a stunning 99.9% of all candidates' cable spots ran on Fox.

This rush for inventory on Fox News has two rather impactful negative consequences for Republican campaigns. First, their ads are being shown in an environment where they are bound to run up against their competitors. Second, and probably more consequential, the rates Fox can charge for a 30-second ad increase exponentially as demand rises for a limited supply of slots. Surely there are other, less-dense and less-expensive places to find Republicans on television.

In 2014, one Republican campaign that did things differently—the Abbott for Governor campaign—ran only 19% of their cable advertising on Fox. Why? They ran media optimization models against the Republican primary audience and identified programming that

gave them broad reach into their target but also balanced target density for each show against its cost efficiency. At the end of the campaign, a randomized controlled experiment revealed their targeted TV efforts to be the most impactful of all their campaign communications. Their buying habits were responsible for a 10.4 percentage point gain in net favorability of the candidate, more than twice as high as the next most impactful form of communication.

Why Media Optimization Matters

Data-driven media optimization has become increasingly common since the 2012 elections. Practically, this is a function of the continued dominance of television advertising in political campaigns. Kantar Media estimates that in 2016, $4.4 billion will be spent on television advertising in political races. Though TV ads account for 80% of a typical campaign's spending, until the uptake of media optimization tools, it was the least data driven. These techniques couldn't have come at a better time, as the way viewers interact with televisions has changed (and continues to change) dramatically.

At the dawn of the TV era, 68% of TV households watched *I Love Lucy*. But the top-rated show in 2013 was NBC's *Sunday Night Football*, only reaching 12% of households. The growth of cable is a primary driver of this change, and indeed, cable's share of overall TV viewership passed the 50% mark in 2001. Much of this is a function of the increasing number of channels available, growing from an average of 19 in 1984 to more than 189 today. And there's more original content: 1999 saw only 23 original scripted series on cable, while this year there are 180. Moreover, an increasing number of consumers are cutting their cable subscriptions and moving to streaming providers like Netflix or Hulu. But despite the growth of these alternatives, even single Millennials, who are least attached to traditional television programming, watch more than two hours of live television on average every day.

Even in this increasingly complex viewing environment, many advertisers continue to make suboptimal decisions about where to place their ads. The metaphors used by TV buyers speak volumes about their philosophy, using terms like "saturation," "carpet bombing," "spray and pray," or "burning it in." But in their defense, much of this philosophy is rooted in necessity, as one of the major encum-

brances to change in advertising has been the data powering it: the Nielsen household.

The Challenge of Measuring Viewership

The television ratings used to drive billions of dollars of advertising each year are built off of the Nielsen company's panel of individuals in each of the nation's media markets. Nielsen recruits a representative sample of Americans and either equips them with devices that will track their media consumption or asks them to keep detailed diaries of their usage. In each media market, Nielsen has a few thousand tracked respondents. This panel, while carefully recruited and maintained, suffers from a few key problems.

First, diary measurement of TV consumption suffers from inherent issues of precision. Respondents may put off their diary entries and fail to accurately recall what they watched, or they may give in to social desirability bias (the desire to answer questions in a way that respondents think reflects well on themselves) and say they watched *Downton Abbey* rather than *Real Housewives*. Their viewing habits may also be more complex than what they can enter in the diary—say, if they are prone to channel flipping. Unfortunately, many key political markets (for example, all of Iowa) are entirely diary-based. And perhaps most importantly, the relatively small samples in the Nielsen panels means viewing habits can only be derived for a few key demographics such as gender, age, and ethnicity, and going deeper into subgroups can be highly unreliable.

This is especially problematic when campaigns have largely moved away from basic demographic targeting in favor of individual voter-level predictive models. Campaigns no longer focus on winning "white suburban women," but instead focus their efforts on individual voters who are predicted as likely to be persuadable. Nielsen is incapable of providing viewing information for a group like this because they lack both the voter information that campaigns have and the sample sizes to do a meaningful voter match to their panel.

Fortunately, campaigns now have the ability to match their lists of voter targets with media consumption data taken directly from the devices that cable and satellite providers use to deliver programming. These set-top boxes were originally built only to push content to a household, but many are now able to pull back viewing information as well. In these instances, any time a viewer changes the

channel on a TV, a new row is created in a dataset that is date- and time-stamped and has the network they arrived at. The advantages of this data are significant, and it has all the hallmarks of a big data solution: the data has volume (it includes raw viewing data for millions of customers), it has velocity (new data is collected every day), and it has veracity (it contains directly observed logs rather than self reports).

Working with TV Data

Providers of set-top box data take many steps to protect their customers' privacy when sharing. Records can sometimes be matched directly to a campaign's list of targeted voters, but the data is matched via an independent third party and sent back anonymized so that no analyst can see what a specific household has watched. Additionally, many providers will go one step further and provide only aggregated information—campaigns will know how many of their targets were watching a particular program, but not which ones. The data also has varying degrees of latency that must be accounted for: some providers will deliver preliminary data within 24 hours that is then backfilled over one to two weeks, while others deliver data in batches two to three weeks afterward.

Once acquired, this data must be further manipulated to be useful. Many sources of viewing data are imbalanced in terms of the type of people that use the service or the geographic coverage of the provider. To account for these biases, analysts have to apply weighting similar to what you might do with a survey that didn't include enough young people. This process uses information about the households in our viewing data and adjusts the dataset to match our overall target group.

The final balanced sample is used to derive the three key metrics used by any political media buyer:

- the target rating, which estimates the percentage of the target audience that is watching a particular program
- the target index, which tells us whether the proportion of viewers that are in our target audience is above or below average
- the targeted cost per point, which serves as a normalized unit cost for ad placements so that we can measure the value we are getting out of each program

The Federal Communications Commission (FCC) mandates that all political ad buys be placed in a publicly accessible database online, and this opens up a new avenue for campaigns to be adaptive and opportunistic: competitive advertising data. Campaigns can see where and when both allies and opponents are purchasing ads. Frustratingly, the data is released as unstandardized PDFs, but combining OCR tools with human coding can get the data into an operable format. From there, matching this planning data to viewership allows campaigns to identify key metrics like share of Voice (the percentage of all ads being run that are coming from the campaign) between campaigns by media market, network, and daypart.

Campaigns also have access to retrospective data on what advertisements actually ended up airing. The data (compiled by Kantar Media) is limited to broadcast channels, but it gives a much more detailed look at when spots aired and their precise contents (issues, tone, and so forth), and it can similarly be matched to viewership data. This viewer-matching process can also be applied to "earned media" news coverage of a candidate or campaign: companies like Critical Mention and TVEyes track closed captions and can identify each time a candidate is mentioned.

Beyond these viewership and advertising data sources, a variety of other datasets—social media sentiment about particular shows, radio and online media consumption data, or guide data on future programming—can all help to identify promising ad opportunities.

Building Optimized Media Strategies

Having established key metrics for media-buying and assessed the overall media environment, campaigns have a number of options for creating an optimized flight of inventory. The nature of this optimization depends on the goals of the particular moment. Campaigns must choose between, or try to balance, reach and frequency. That is, for a given cost, an advertiser can choose between reaching many people a few times by targeting programs with high viewership or reaching a few people many times by purchasing more spots on programs with fewer viewers.

Moreover, campaigns are usually trying to do this with an advertising schedule that gives them both "horizontal balance" (good representation in each of the TV viewing dayparts) and "vertical balance" (a mix across all of the networks). This is usually an attempt to max-

imize a schedule's unduplicated reach—the calculation of the number of viewers that you are reaching at least once. And indeed, this becomes another important evaluative metric for the overall health of the ad buy.

Finally, campaigns can also use this viewership data alongside other data sources that help give them a holistic view of the political media landscape. Tracking earned media (free publicity gained through social media, word of mouth, news mentions, etc.), for example, would enable a campaign to identify not only which newscasts most often mention their candidate, but also how many swing voters were likely to be watching. With this information, the campaign could choose to run ads on the same programs to complement or respond to that coverage, or alternatively, to direct ad resources to other channels where the audience is less likely to hear about the candidate otherwise.

The complexity of media optimization underscores the need for analysts in the political space who possess a facility with varied datasets and can quickly process, clean, and merge these data with others. This means familiarity with cloud-based storage solutions like Amazon Web Services, distributed data processing platforms like Spark, query languages like PostgreSQL, and programming languages like R. At the same time, these analysts also need to understand the intricacies of the media buying process and how it fits into a broader campaign operation.

Media optimization, previously the province of simplistic demographic targeting, has had to grow in its complexity to match the nuances of modern media consumption. Fortunately, the data now exists to perform the necessarily sophisticated analysis to account for this, and it has come at a critical time. A new generation of marketers has grown used to the targeting and optimization capabilities of online advertising, and applying the same optimized approach to TV allows campaigns to make their largest budgetary outlay both more effective and more efficient.

CHAPTER 5

How (and Why) to Follow the Money in Politics

Derek Willis

Derek Willis is a news applications developer at ProPublica, where he maintains political data websites and APIs and does some reporting, too. He previously worked at The New York Times.

Campaign finance data is ultimately about behavior. Journalists follow the money in politics because it reveals the connections between donors and politicians and tells us something about how campaigns operate. Fundraising and spending data helps us see whether politicians actually do what they say, what each candidate's priorities are in running a campaign, and which people and organizations play major roles in campaigns. Even though the amounts of money can be distracting, campaign finance data is a reminder that politics is about people and their decisions.

As a journalist, what interests me is how campaigns raise and spend their money, and in particular how the Internet is changing that. I try to find stories in the data, keeping in mind that the best stories involve people doing things—sometimes fascinating or unique, sometimes banal or criminal.

Getting Campaign Finance Data

First, a brief history of federal campaign finance data in the United States. Whenever you hear, "This is the most expensive election in

history," know that "history" means since 1978, since that was when the Federal Election Commission (FEC) began collecting and publishing data that campaigns, journalists, and researchers can download and use. Here's how it works: political campaigns and organizations ("committees") register with the FEC, then file reports listing their contributions and expenditures on a regular schedule. The FEC takes that information, standardizes it, and releases weekly files covering a two-year election cycle on its FTP site.

For years, that was the only way to get FEC data: wait until Monday morning when updated files were published, then download them, and import them into your database of choice. And the practice continues today—rare among government agencies, the FEC believes in and practices backwards compatibility for its data files. You could, if you wanted to, maintain a local copy of every disclosed contribution since 1978. But there's a big catch: for years, that information had to be entered by hand by the FEC's contractors. (Because of a quirk in campaign finance law, reports by senate candidates still require this step.) That meant (and still means) a delay of weeks for some data, as some filings run thousands of pages long.

Thankfully, most committees now file electronic reports, which means we can see them almost immediately on filing deadlines. This is a very good thing for people interested in campaigns, and not just because we don't have to wait for data. Electronic filing means we actually get more data: expenditures, which the FEC did not keypunch for years, are now available, as are street addresses for donors and recipients, which the FEC avoided as a cost-saving measure when data entry was labor-intensive.

A small group of developers at *The New York Times*, the Associated Press, and ProPublica, among other news organizations, have written software to parse the electronic filings and make them more useful for journalism. But there's a catch here, too. The electronic filings are considered unofficial by the FEC until they are processed and released via FTP. And for good reason, too: they do contain mistakes, in terms of both substantive content and data formatting. So, as the adage goes: timeliness, accuracy, completeness—pick two, most of the time.

The Toolkit

The hard part about working with FEC data isn't the technical stuff. If you know SQL and a general-purpose language like Python or Ruby, it's not hard to work with the FTP data, which has solid documentation. More important are the lessons learned from working with the data and understanding the political system that it describes. That's why it can help to have software that is regularly interviewing the data, comparing new records to older ones and identifying patterns and outliers.

At ProPublica, where I maintain our campaign finance data, we use PostgreSQL for storage and a Ruby on Rails application to handle regular updates, common analysis tasks, and the API that we offer (the FEC has its own API, in beta). Almost everything is automated. We load new electronic filings every 15 minutes, and update some of the larger files produced by the FEC daily or weekly. We need the ability to make manual edits sometimes, but not often, and usually to fix an obvious error in the data.

This isn't exactly big data by industry standards: we've got more than 100 million rows of data covering contributions, expenses, and other records included in the FTP data dating back more than 10 years. But some individual filings, which can contain multiple record layouts in a single text file and include millions of individual contributions, can consume a lot of resources to load and process. And while bigger campaigns account for more of the data, fundraising occurs whether an election is considered competitive or not.

Part of the software process is translating the questions we have into code. Most of these are basic heuristics, but machine learning can play a role in standardization; *The New York Times* has developed software to help standardize donors based on the name and address information contained in filings. We've also adapted statistical tests to find connections among donors and recipients, using cosine similarity as a measurement of how alike two recipients are. The harder task is trying to learn enough from changes in the entire campaign finance system to look for things we've not seen before. A lot of those changes have come about thanks to the Internet.

What the Internet Has Changed

When Bernie Sanders began his presidential campaign in 2015, he had few of the traditional fundraising networks to rely upon. As a senator from Vermont, he had access to some individuals willing to give him donations of up to the $2,700 federal limit, and something of a national following, albeit small compared to some other politicians. But he didn't need one, thanks to online fundraising. Supporters could—and did—repeatedly donate small amounts to Sanders' campaign, mostly through ActBlue, a unique type of committee that serves as a conduit to Democratic campaigns and related organizations.

ActBlue provides campaigns with the standard tools of a digital campaign platform—hosting websites, managing email programs, and processing donations—but the group itself is organized as a nonprofit political action committee rather than a for-profit vendor. Because of that, FEC rules require ActBlue to report the details of all donations it processes for its clients, not just those from donors who have given a total of $200 or more (the standard for donations directly to campaigns or through commercial processors). This means that ActBlue's filings provide an unprecedented amount of data to analyze. The group's 2015 year-end filing, for example, was the largest single filing in the FEC's history, at over 3.1 *million* pages, and its monthly filings have averaged more than a million pages *each* in 2016.

Thanks to ActBlue's filings, we have a much better idea of who Sanders' financial supporters are, and likewise donors to other major committees (such as the Democratic Congressional Campaign Committee) that also use ActBlue's platform. By matching these contributions to voter file records, campaigns can use this information to identify voters receptive to these campaigns' messages. (By law, other committees cannot directly use Sanders' contribution data to solicit for their own fundraising, but they can use it to help improve what they know about politically active individuals.)

We can't see this in real-time, but committees file at least once every three months (with more frequent filings by larger organizations and closer to Election Day), and committees often amend previous filings in between to correct mistakes and account for refunds. We do see filings almost every day, and direct online access to the FEC's records makes it possible to examine not just the financial activity of

a single campaign or race but of the entire system of campaign finance. That's the next step for those of us who work with campaign finance data.

The Challenges Ahead

To study the system as a whole, we need complete data, and right now we don't have it. The most important single step that Congress could take is to require senate candidates to file their reports electronically. This would close a loophole that makes it nearly impossible to provide a complete picture before an election is held.

Another task that's getting easier is donor and recipient standardization. Both require a combination of machine learning techniques that enable comparisons of large numbers of records and manual research. Now that it's easier than ever to give smaller amounts, we need to pay more attention to donors who make dozens of small contributions, not just wealthy ones who can write checks for thousands of dollars at a time, which has long been our focus. Shifting towards examining these donors could make the system more approachable to more people, since most Americans don't make political contributions.

Spending data isn't just overlooked; it's also one of the clearest indications of political intent in campaigns, as opposed to what politicians might say. There's a phrase I've heard clergy members tell their congregations that applies here: "Don't tell me what you believe. Show me your checkbook, and I'll tell you what you believe."

Finally, we've long placed an emphasis on campaign finance in races for president and Congress, and less on state and local races. That's largely because there's a single federal system for campaign finance data, while each state runs its own for non-federal races. The single federal regulatory system is easier to deal with, but 50 disparate regulatory environments can provide opportunities for those looking to exploit the system for personal or political benefit.

Most people don't think of elections as "federal" or "state," and races at one level can impact the others. What we need to work on in the years ahead are systems that help make it easier to see the money connections between campaigns no matter where they are or who contributes to them. The open data example set by the FEC is a good

one for states and counties to follow (and some state and local authorities do that and more), even if it's not perfect in every case.

CHAPTER 6

Digital Advertising in the Post-Obama Era

Daniel Scarvalone

Daniel Scarvalone is the Associate Director of Research and Data at Bully Pulpit Interactive (BPI). BPI serves as the largest digital marketer for the Democratic party as well as working for major corporations and causes. Before joining BPI, he served as Director of Data and Modeling at the Democratic Congressional Campaign Committee in 2014 and as National Reporting Director for the Obama 2012 campaign.

President Obama's presidential campaigns revolutionized the way technology and data could be used together to identify and speak to voters' interests, reflecting the increasingly sophisticated application of digital marketing to politics. But in this election cycle, and for the last 20 years, political campaigns still spend more than 70 cents of every dollar to reach voters via broadcast television. While broadcast TV is an effective medium to talk to voters and change minds, there is great debate about how cost-efficient it is.

In politics, there are countless conversations about the importance of matching the "red bar" (the amount of money the opposition is spending) to the "blue bar" (the amount of money our allies are spending), irrespective of cost. This arms race means TV stations can charge an exorbitant price for ad inventory, especially as Election Day approaches. And this paradigm completely overlooks the effects of ad frequency—i.e., whether the 30th TV ad each week is as

effective as the 10th—taking away a key tool for running efficient campaigns.

As voters spend more time than ever online, digital advertising combines the targeting, persuasion, and measurement capabilities that move the most votes at the least cost. Modern campaigns can no longer win using TV ads alone. They must shift their dollars to mediums where they can quantify the bang they are getting for their buck. This shift is critical to what will make or break the future of political campaigns, as well as how corporations talk to consumers or advocacy organizations talk to supporters.

My firm, Bully Pulpit Interactive (BPI), is a digital advertising agency that specializes in public affairs, corporate reputation, social impact, and political campaigns. I work on the BPI Labs team, where I focus on connecting measurement and analytics to our digital advertising campaigns, and building out technical infrastructure throughout the entire firm.

How Digital Advertising Works in a Campaign

Digital has become an integrated operation throughout campaign structures and part of the core strategy of most campaigns, operating at the intersection of fundraising, communications, and voter contact efforts. Many parts of a digital advertising program are generally run in-house, from email writing to volunteer recruitment and mobilization.

For paid digital advertising, the process is often similar to that of television advertising: campaigns have the final say on the strategy, budget, and message of a buy recommended by consultants, but leave the execution and buying tactics to an outside firm. Like modeling, analytics, and television ad buying, the economies of scale that enable efficient digital buying and measurement aren't available to individual campaigns like they are for high-volume agencies like BPI.

At the outset of a campaign, the leadership will set an overall campaign budget and allocate that budget to different types of voter contact. After polling and research determines the messaging and audience strategy, the digital staff works with the rest of the consulting and leadership teams to define and execute media buys that attempt to register, persuade, mobilize, or turn out voters.

Efficient and effective digital advertising programs are informed by a few rules and principles:

1. **Let the audience be your guide.**

   Leverage the sophistication of political analytics programs to define the precise audience you want to reach, and then build strategies matching the consumption and behavioral patterns of that audience

2. **Tailor the creative.**

   Digital advertising allows campaigns to tailor creative elements of advertising to specific audiences at scale. Audiences can be shown the customized messages that will move them the most, instead of speaking to everyone with one megaphone. But only be as granular as your creative capacity allows—each individual voter doesn't need his or her own version of an ad.

3. **Don't just measure *how much*, measure *how well*.**

   Focus not on how much media was delivered, but instead on how that media actually changed the minds of their target audiences. Campaigns should be consistently integrating attitudinally based, experimentally-informed programs (EIPs) to measure the efficiency and effectiveness of every dollar they spend.

This last point is the key area of focus for digital marketers in 2016 and beyond, and the motivation behind many recent advances in technology and tools for advertising agencies on both sides of the aisle.

Using Experimentally-Informed Programs to Measure Effectiveness

Borrowing from concepts developed in the academic world, an EIP divides an audience into random subsets that are exposed to specific advertising conditions (treatments) or to no ads at all (control). Because these groupings are assigned at random, we can assume that their attitudes or activities would have been identical (statistical noise aside) absent any advertising. Therefore, any differences can be attributed to the advertising received by the treatment group.

EIPs follow the same basic principles as traditional A/B tests, but because the outcomes they measure are more complicated than basic click-through or purchase behavior, they also typically involve some kind of separate post-exposure measurement. Crucially, these measurement cycles take place within the advertising campaign itself—allowing for the applications of those insights to optimize the remainder of the program. EIPs also allow for the testing of multi-part treatments and other complex strategies.

Tools for Delivering Better Ads and Measuring Their Impacts

In order to convince campaigns, companies, and causes to buy into digital advertising, the industry has had to prove that digital advertising could raise money, mobilize activists, and provide engagement with its content. But the days of proving that only by impressions and clicks are over. Now, in addition to those metrics, there are others that matter: exposure time, creative heat mapping, and audible/viewable metrics to ensure clients are getting the most exposure for their investment.

A key decision point for any digital advertiser is choosing partners to provide the types of tools and metrics that are necessary to run a smart ad campaign and tell a story about effectiveness. Parts of the digital ecosystem are well monopolized, and custom-building a solution would be foolish. No one has built a person-based advertising platform that rivals Facebook's, a buying platform with breadth and reach that rivals Google's Doubleclick, or a matching platform that rivals Liveramp and Neustar. This parallels the situation for offline components of a campaign as well: an overwhelming majority of Democratic campaigns use VAN to record and measure their voter contacts, BSD to send their email, and Acxiom, Experian, or Infogroup to supplement their voter files with consumer records.

But in other areas, the ecosystem is sufficiently fragmented for agencies to have leeway in choosing with whom to partner to help detect fraud, to measure viewability, or to reach certain audiences. Across existing technology solutions, a lot of the metrics that have been effectively commoditized in the digital space don't effectively tell the story of the persuasive effectiveness of our campaigns, and whether advertising dollars actually moved the minds of our audiences. When we talk about moving minds, we want to know which features

of our campaign made voters more likely to retain a fact about a candidate, feel less favorably about an opponent, or shift their vote choice.

In this environment, BPI made the decision to build its own measurement tool, Vantage, which quantifies the persuasive impact of different types of digital marketing. By surveying individual users online and connecting those surveys to exposure to a specific digital campaign, we've begun to build a platform that can integrate the full range of advertising tactics outlined above and to learn how to apply those tactics for maximum effect.

None of what can be done now would have been possible in the technological ecosystem that existed 5 or 10 years ago. The ability to leverage scalable, modular pieces of infrastructure like Amazon Web Services and Google Cloud Storage are absolutely critical to building the technology that powers digital advertising. These solutions give us the ability to rapidly and efficiently scale what we do to meet the ever-evolving solutions in the ecosystem, without locking into an unsustainable cost curve.

Finally, there is some critical nuance to figure out which parts of our ecosystem can be automated and repeated, and which parts need to be iterated on and produced anew for each client. Measurement as a science belies easy cycles of productize-and-forget-it. Ad measurement technology has to be as tailored as the advertising that we run for a given client or vertical. And since the advertising ecosystem changes every quarter, the ad technology has to evolve as well.

Adapting to a Changing Campaign Environment

The most critical asset in broadening adoption of digital advertising isn't technology as much as it is the way we talk about it. If digital practitioners can't use measurement to demonstrate the value of their work, then it won't be taken seriously, no matter how novel the campaign or technology used to analyze it. What we do is just as important as how we talk about it and measure it.

The end result focuses on a clear ROI of the overall persuasive impact of a campaign, but also on the relative performance of different messages and tactics. For years, campaigns have responded to mail and field measurements that have put a precise cost-per-vote

on their programs. Now digital advertising can optimize for persuasion in the same way, and provide campaigns an apples-to-apples metric of cost-effectiveness across mediums and tactics.

Applying These Lessons for Non-Political Clients

Digital advertising should be tailored to the specifics of the client and the objective of the campaign, but many parts of our approach remain constant. Focusing on a defined universe of individuals—from a model, poll, or external list—is a critical component of planning, executing, communicating, and measuring the success of advertising. Whether a campaign is devoted to motivating someone to vote, or getting someone to buy a car, it is trying to generate actions that are difficult to move and measure at scale.

Whether voting for a candidate or purchasing a car, the actions an audience takes can be difficult to attribute to the specific digital advertising they have received. To meet this challenge, it is important to create specific messages within a campaign to move an audience through a "ladder of engagement"—for example, first getting someone to sign up for a mailing list, then priming them to pay attention to those emails, and ultimately inspiring them to make a purchase or donation. This allows us to measure success at each stage of the process and optimize the overall program toward the high-bar end goal. A deliberate focus on specific steps along the way helps us run better campaigns by making it possible to measure and optimize a program's effects before the final vote tallies or sales numbers are in.

What Comes Next

Future campaigns, especially at the downballot level, will continue to embrace and elevate digital considerations in campaign decision-making. More and more campaigns in 2018 and beyond will be led by staff with digital expertise or who have run digital programs.

In a technological ecosystem that's dominated at the top by industry giants like Google and Facebook, and fragmented at the bottom by hundreds if not thousands of ancillary solutions, the coming years will force advertisers to make difficult decisions about how to prioritize platforms and approaches to stay ahead of the curve. Digital

components of a campaign are dependent on technological partners to achieve scale and efficiency, and so decisions about advertising technology affect every facet of its program. Once those digital components are decided upon, you'll see advertisers begin to build budgets from the ground up (based on target audience size, ad vendors' capacity, and a proven understanding of where marginal dollars are best spent), instead of allocating budgets from the top down based on outdated rules (like spending every required dollar to achieve television parity, and then allocating the rest among a campaign's other functions).

And finally, digital survey technology—the actual tools that collect attitudes and quantify effectiveness—will continue to evolve. Campaigns will come to expect the same forms of feedback from the online world as they do the offline world, and that means being able to collect more demographic information, detailed feedback, and open-ended qualitative outputs from survey respondents. Digital advertisers that focus on developing their measurement solutions will be the ones best equipped to successfully participate in the discussions about campaign planning at the highest levels.

CHAPTER 7

Election Forecasting in the Media

Natalie Jackson

Natalie Jackson is Senior Polling Editor at Huffington Post, coordinating the Pollster section of the site. Her primary focus is on polling coverage and methodology, statistical methods, and using polls to forecast elections. Natalie has a PhD in political science from the University of Oklahoma, with heavy emphasis on statistics, survey methodology, and American politics.

Election forecasting has been around for many years in academic circles, but in the last few presidential election cycles it has seen tremendous growth in the media. Networks and other media outlets have gathered polls and made electoral projections for decades, but recently advanced statistical models have become more prominent in elections coverage. In 2008, Nate Silver debuted his FiveThirtyEight blog with his projection that then-Senator Barack Obama would defeat Senator John McCain, and the market for that kind of statistical model has expanded in each subsequent election cycle. In 2014, *The Huffington Post*, *The New York Times*, FiveThirtyEight (by then its own media outlet), *The Daily Kos*, and *The Washington Post* all had forecast models for the midterm senate races.

The process of election forecasting is complex, but the general public seems to love forecasts. A challenge, however, is getting that audience to understand that a forecast—for example, Candidate A has a 60% chance of winning—is not as simple as it sounds.

The Basic Mechanics of Election Forecasting

Most media forecast models use time-series methods to combine data into a rolling daily prediction of how likely a candidate is to win an election. Much of the data comes from publicly released pre-election polls, although there are many other data sources that can be included as well. The key difference between these forecast models and simpler polling aggregations is that forecasts project beyond the polls, whereas the aggregations simply average the most recent polls.

Every forecast is slightly different, but there are some basics that differentiate them: whether the modeling technique is Bayesian, and whether the model includes data other than polls.

Bayesian versus Frequentist Modeling

From a data perspective, the fundamental difference between a Bayesian forecast model and a non-Bayesian (frequentist) forecast model is the ability to incorporate "priors" into the model that quantify what's already known about the election. We usually know quite a bit about upcoming elections by learning from past elections, so this prior information can give us a place to start with our predictions.

For example, the 2014 senate forecast model I created for *The Huffington Post* used a Bayesian method: Kalman filtering. To come up with priors, I used the *Cook Report*'s estimates of whether the senate race was a "tossup" or was leaning, likely, or solidly Republican or Democrat. An analysis of past *Cook* ratings and election results provided information on the distribution of final vote shares for races assigned to each rating category in past election cycles. Those parameters served as the priors for the 2014 senate races, effectively incorporating all the information inherent in current *Cook* ratings as well as their past performance. As new polling data became available, it was used to update the priors and produce a distribution of possible outcomes for each race. The means of those "posterior" distributions were the vote share estimates, and the distributions themselves provided the estimates of uncertainty and the probabilities of each outcome.

Non-Bayesian modeling eliminates the priors and works with traditional regression techniques. The basic procedure is similar, how-

ever. Polling averages are typically calculated using some form of time-series model. If election fundamentals—other non-polling data that can help predict election outcomes, such as the state of the economy—are included, they are combined or modeled to get a fundamentals estimate. Then the polls and fundamentals are put together to generate a single outcome. These outcomes can still be expressed in probabilistic terms using the confidence intervals and standard errors that the models produce, so that both the Bayesian and frequentist models are reported similarly. It is only in reading the details of each model that the difference in techniques becomes clear.

Bayesian modeling requires different software tools than frequentist regression modeling. Most of the coding and analysis can be done in R, but you need special-purpose Bayesian modeling software such as JAGS (for Mac users) or BUGS (for Windows users) to execute the model. For frequentist modeling, depending on the exact type of model you choose, you can use a wide variety of statistical modeling software, including many open-source packages available in R, Python, and other common programming languages.

Fundamentals versus Polls-Only

Another key difference between types of forecast models in the media is whether the model includes so-called "fundamentals" about the election or is only using polls to construct the forecast. Fundamentals are generally anything besides horserace polls that contains information about how the election might turn out: indicators of the national mood, measures of the partisan makeup of the district or state, incumbents' previous win margins, measures of political experience for each candidate, fundraising and ideology scores for the candidates, or any other relevant metric.

There isn't any evidence that fundamentals dramatically improve a model's predictive power close to the election, but they do offer more long-term stability when the election is months away. Fundamentals provide more information about the electorate and the general election atmosphere, and they don't change as frequently as polls can a long time before the election.

Estimating the Electoral College

Most models use Monte Carlo simulations to estimate the final probabilities of a presidential candidate winning an election across the various states, or as in the 2014 models, to estimate the likelihood that a party will maintain or take over control of the Senate. The probability of each individual state race going Democratic or Republican is put into the simulations, and in a presidential election forecast, winning a state is converted to the number of electoral votes the winner would receive for that state; in a senate forecast the election in each state is counted as one seat. The Monte Carlo process simulates many different random elections and the proportion of times a presidential candidate gets more than 270 electoral votes or a party has 51 or more seats in the Senate is the final probability for the outcome of the contests.

Communicating About the Model

In media forecasting, it's not enough to have a good model. You have to be able to explain it to the audience. This task can be even more difficult than building the model itself. Explaining the uncertainty of probability-based forecasting to the general public is a task that has flummoxed scientists, and particularly weather scientists, for many years. It seems no matter how many times you remind the public that forecasts are based on uncertain probabilities, some people want to read the numbers as completely certain, and then castigate the analysts if the outcome is different from their expectations.

All of the major media forecasts for 2014 measured the outcome in terms of the probability that the Republicans would take over the majority in the Senate and splashed those numbers on a main forecast landing page. Most forecasters did explain the uncertainty of the forecasts, often in great detail. However, these discussions of uncertainty were typically buried in long discussions of the methods used to generate the estimates—which most people will not read all the way through—and the message was easily lost.

The (probably large) portion of the audience who went directly to the forecast pages, ignoring the methods explanations, saw numbers that declared how likely the Republicans were to take over the Senate without any explanation for what an 80% likelihood actually means. Presenting the numbers with the appropriate explanation of uncertainty, without requiring the audience to spend an hour read-

ing model details, is something public forecasters will need to work on in the future.

The Future of Election Forecasting

Media-produced forecasts might not proliferate much more, however. These kinds of forecasts that need to appeal to a broader audience could face a few issues if more outlets get into the game. First, presumably there is a finite audience that these media forecasts can appeal to because of their statistical complexity. Additionally, most forecasts generally come to the same conclusion as Election Day nears, so the utility of developing more forecasts is questionable. The skillset is expensive and complex.

At the same time, forecasting is likely to become more common as technological advances make statistical tools more user friendly, also similar to what happened with polls in the last few decades. It's possible that more forecasts would flood the market and the bubble of election forecasting popularity could burst, particularly if forecasts are not as accurate as they have been between 2008 and 2014. And, although the forecasts might already seem ubiquitous, 2016 is only the third presidential election cycle to feature forecasting. It remains to be seen whether forecasting is a temporary trend or will become a permanent fixture in election coverage.

Epilogue: The Future of Political Data Science

Andrew Therriault

The 2016 election cycle will be remembered for many things, but for those who work in politics, it may be best remembered as the year that political data reached maturity. For years, much of the "old guard" of political strategists resisted the growing influence of data and analytics, preferring to stick with a more traditional formula: instinct and experience built up over time. This conflict persisted even as the 2008 and 2012 presidential races showed the advantages data can provide, particularly when one campaign has a distinct technological edge. In 2016, though, it's clear that this fight is all but over, and the data side has won.

When it comes to technology, political data science is likely to follow a similar trajectory to the one playing out in the broader field of data science. If anything, the political field is particularly well-suited to rapid adoption of new technologies, since the election cycle allows for many organizations to completely overhaul their technology every 2 to 4 years. Though physical hardware is still used by many organizations, the adoption of cloud platforms like Amazon Web Services and Microsoft Azure has quickly become mainstream. Along the same lines, while commercial data analysis software such as Stata and SPSS was once standard, most organizations are now relying mainly on open source tools like Python and R. And as the scale of our datasets grows, traditional relational databases such as MySQL and Microsoft SQL Server are being replaced with distributed platforms such as HP Vertica, Amazon Redshift, and the Hadoop ecosystem.

For novice and veteran data scientists alike, working in politics is a great opportunity to quickly build skills working with some of the most advanced technology and techniques available. While it's probably too late for readers to become part of a campaign this year, you don't need to wait until 2020 to get involved. In 2017, there will be elections for governor and state legislature in both Virginia and New Jersey, mayoral races in many cities including New York, Los Angeles, Seattle, and Boston, and a variety of other local contests throughout the country. And in 2018, voters will elect 36 governors, 33 US senators, and all 435 members of the US House of Representatives. So even after the presidential race is over, the field of political data science will continue to grow and evolve for many years to come.

About the Authors

Andrew Therriault was the Democratic National Committee's Director of Data Science from 2014 to 2016, leading a team that developed voter targeting models and other analytic tools used by thousands of Democratic campaigns. Therriault has spent the past decade working on all types of political data as both a researcher and practitioner, and he received his PhD in political science from New York University in 2011. He was recently appointed as the City of Boston's first Chief Data Officer, leading a team that applies the tools and techniques of data science to improving municipal government.

Dan Castleman is cofounder and Director of Analytics at Clarity Campaign Labs, a modeling and targeting consulting firm for progressive candidates, political coalitions, corporations, international campaigns, and charitable groups. His expertise covers strategic program targeting, data analysis, modeling, and polling. Prior to starting Clarity, Dan helped found ISSI in 2007, where he grew and oversaw their analytic services and developed internal systems to streamline and optimize cutting-edge data-mining tools and technologies. Dan holds a dual BA in Political Science and Computer Science from Brandeis University.

Audra Grassia leads political engagements at Civis Analytics. She has spent the majority of her career working on political campaigns across the country, including statewide, local, and federal races. She has also spent time managing technical programs for two separate companies and Brigham and Women's Hospital in Boston. Additionally, Audra was the Deputy Political Director for the Democratic Governors Association during the 2014 election cycle. Originally from Sugar Land, TX, Audra graduated from the University of Texas at Austin with a BA in Government. She also received her Masters in Public Health from Washington University in St. Louis.

Patrick Ruffini is a cofounder of Echelon Insights, a survey research and analytics technology company. He has been applying data and technology to political campaigns for more than a decade. Ruffini helped establish one of the first full-fledged digital operations in Presidential politics for Bush-Cheney '04, led digital strategy for the Republican National Committee, and later founded and grew Engage, one of the leading digital agencies on the right.

Alex Lundry is the cofounder and Chief Data Scientist at Deep Root Analytics. Alex has worked as a political data scientist, pollster, microtargeter, data-miner, and data-visualizer for presidential candidates, national organizations and Fortune 50 companies. As a pollster, he has conducted surveys and focus groups on everything from the race for the White House to Hollywood movies. He is one of the country's leading experts on media and voter analytics, electoral targeting, and political data-mining, and has directed the data science efforts of two presidential campaigns.

Derek Willis is a news applications developer at ProPublica, focusing on politics and elections. He previously worked as a developer and reporter at *The New York Times*, a database editor at *The Washington Post*, and at the Center for Public Integrity and Congressional Quarterly. He began his journalism career at *The Palm Beach Post*. He is a cofounder of OpenElections, a project to collect and publish election results from all 50 states.

Daniel Scarvalone is the Associate Director of Research and Data at Bully Pulpit Interactive (BPI). BPI serves as the largest digital marketer for the Democratic party as well as working for major corporations and causes. Before joining BPI, he served as Director of Data and Modeling at the Democratic Congressional Campaign Committee in 2014 and as National Reporting Director for the Obama 2012 campaign. Daniel also previously worked with Catalist and NGPVAN.

Natalie Jackson is Senior Polling Editor at *Huffington Post*, coordinating the Pollster section of the site. Her primary focus is on polling coverage and methodology, statistical methods, and using polls to forecast elections. Natalie has a PhD in political science from the University of Oklahoma, with heavy emphasis on statistics, survey methodology, and American politics. Prior to joining *Huffington Post*, she worked as a survey consultant as a postdoctoral associate at Duke University and as senior analyst at the Marist College Institute for Public Opinion.