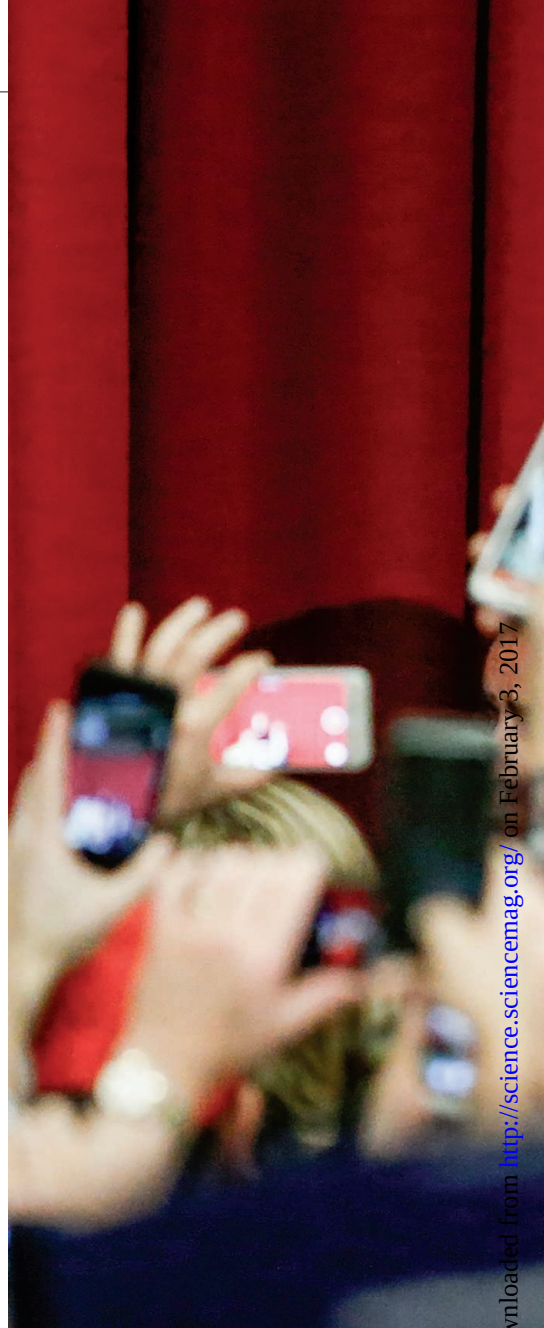
THE PULSE OF THE PEOPLE

Can internet data outdo costly and unreliable polls in predicting election outcomes?

By **John Bohannon**



Both polling and an analysis of pre–election night tweets failed to flush out Trump's hidden voters.

PHOTO: EDUARDO MUNOZ ALVAREZ/STRINGER/GETTY IMAGES

In an apartment on New York City's Upper West Side on 8 November 2016, Hernan Makse and several friends cooked branzino and sipped Chablis as they watched the U.S. presidential election unfold. They hopscotched between MSNBC and Fox News while keeping an eye on *The New York Times* website on a laptop. The *Times* was streaming live updates of its "presidential election forecast." It was still early, and results from key states had not yet come in. On a chart labeled "Chance of Winning Presidency" that reflected the polling data rolling in, Hillary Clinton bounced above 80%, leaving Donald Trump mired below 20%.

Makse, a statistical physicist at City University of New York, had placed a scientific bet on the outcome. The day before, his lab group had posted a research paper to arXiv, the online preprint repository. They had feverishly revised it to make the 4 p.m. deadline and publish on Election Day. Like the gauge chart on the *Times* website, they predicted who would become president. But whereas the *Times* used data from state-by-state polling, Makse's prediction was based entirely on data gathered from Twitter in the months leading up to the election.

If Makse's group nailed the election forecast, they would have reason to brag. Polling, whether done by phone or door-to-door, is extremely labor intensive and expensive: It fuels an $18 billion industry. And it has problems. Not only have response rates fallen to single digits, leaving pollsters to rely on a thin and biased sample of people, but also an analysis last year of more than 1000 polls found evidence of widespread data fabrication (*Science*, 4 March 2016, p. 1014). By contrast, Makse's group tracked the political opinions of millions of people directly, second by second, for months—and they got those data for free.

Twitter isn't the only online data stream that scientists are funneling into predictive models of everything from elections to street protests. The largest tech companies such as Facebook and Google generate data that are free for researchers to use, though with varying degrees of inconvenience. So Makse and many other social scientists are asking: Could online data enhance polling as a forecasting tool, or even replace it?

The election night verdict: not yet. As the evening wore on, Makse's forecast based on freely harvested tweets continued to match the pricey polling data, predicting a win for Clinton with 55.5% of the vote. But both forecasts got it wrong. Before their dinner was done, Makse watched as the projections on the *Times*'s data-driven blog, The Upshot, caught up with reality. "It was funny to see how at around 8 p.m.," he says, "they switched from 20% to 95% for Trump."

The internet, it seems, can't yet reliably take the pulse of the people. But Makse and many other social scientists are convinced that it eventually will—if only they can figure out how to translate terabytes of data into human intentions.

FORECASTING WHAT PEOPLE WILL DO, and why, is the essence of social science. Considering how hard it is to divine even a single person's behavior, scaling up prediction to a community or society seems like a nonstarter. "But in some ways that is an easier problem," says Taha Yasseri, a computational social scientist at the University of Oxford Internet Institute in the United Kingdom. He offers an analogy from physics: Although the movement of a single particle looks random, "the behavior of a gas made up of millions of particles is very predictable."

The idea that society can be treated like a physics problem has deep roots. In the 1950s, science fiction author Isaac Asimov conjured up a branch of science called psychohistory. With powerful computers and gargantuan data sets, he imagined, researchers would forecast not just elections,

but the rise and fall of empires.

A lifetime later, the computers and the data Asimov envisioned are becoming reality. But for now, polling—costly and inefficient as it is—remains the tool of choice for predicting group behavior such as elections. And a study of electoral races around the world on p. 515 suggests that polls are still reliable, despite last November's surprise.

Ryan Kennedy, a social scientist at the University of Houston in Texas, and colleagues focused on a data set of presidential elections. They avoided the complexity of comparing different government systems by limiting the study to elections in which voters chose a national leader directly, rather than, for example, through a party-based parliamentary system like the United Kingdom's. That filter left plenty of data: The final tally came to more than 500 elections from 86 different countries going back to World War II. To predict winners, Kennedy, David Lazer, a social scientist at Northeastern University in Boston, and his Ph.D. student Stefan Wojcik statistically modeled the elections using voter polling data as well as data on other factors that can tip elections: the country's economic prosperity, democratic freedoms—using a third-party measurement called a Polity score—and whether an incumbent was running.

They trained their models on data up to 2007 and then tested them on the most recent 8 years, totaling 128 elections. Overall, they correctly predicted the winner 80% to 90% of the time. And of all the indicators, polling proved the most powerful, by far. "We predict that reports of the death of quantitative electoral forecasts are greatly exaggerated," the authors quip. Others agree that for the time being, polling reigns. "If you're trying to predict a decision people will make, there's just no substitute for asking them directly," says Andrew Gelman, a statistician at Columbia University.

Yet Lazer, for one, believes our reliance on polling may not last much longer. "Canonical polling methods are in crisis," he says. One factor is people's growing impatience with pollsters; another is the death of the landline. You can't survey people if you don't know how to find them. Could a fire hose of data from the internet plug the gap? That holds "great promise," says Lazer, "but a lot of work has to be done before those approaches are validated."

One challenge is that it is hard to decipher people's motivations from their internet habits: that is, their web searches and

social media posts. If millions of people tweet sentiments supportive of a candidate or critical of an opponent, can it be deduced, reliably, how they will vote? Predicting people's behavior is tough, Yasseri says, "if you don't know what motivates them."

A promising test ground for probing motivation is Wikipedia, a website used by a remarkably broad swath of humanity as a one-stop shop for basic information on almost any topic. To see what Wikipedia's traffic might reveal about electoral outcomes, Yasseri and fellow Oxford researcher Jonathan Bright have been tracking the number of daily visitors to the Wikipedia pages devoted to political parties competing in the European Union's parliamentary elections every 5 years. Because the voters speak different languages, Yasseri and Bright gather data separately from each of 14 language versions of the site.

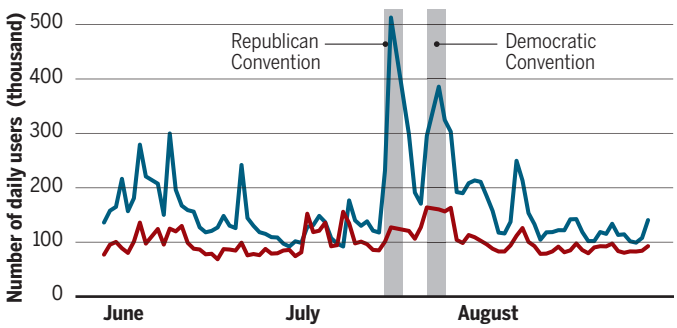
The number of visitors to each political party's Wikipedia page would not have reliably predicted who ultimately won seats in the 2009 and 2014 elections. "It's not that simple," Yasseri says. His theory is that voters are "information misers" seeking out the minimum information needed to make a decision. And indeed, they found, the most active Wikipedia pages tended to be those of newly formed parties, with visits peaking in the week before an election. By monitoring those pre-election spikes, Yasseri and Bright reported last year in *EPJ Data Science*, they were able to forecast the gains of the nationalist parties within France and the United Kingdom in their respective national elections. But what's needed, they say, is more information about those Wikipedia visitors to translate browsing choices into likely voting choices.

To put human prediction to the test, Yasseri is part of a European team building a "social data bank," which, like a genetic data bank, would provide deep information—demographics, health records, traces of online browsing, and even mobile phone data—on a slice of the population. To start, it will focus on the United Kingdom, Finland, Hungary, Spain, and Slovenia. "We have to figure out how to keep the data anonymous," Yasseri says. The hope is that tracking all the online behavior of relatively few people will allow researchers to deduce what motivates someone to visit a website, tweet, and, ultimately, vote one way or another. Once they solve the anonymity problem, he says, the team hopes to start predicting outcomes such as elections within a few years.

MAKSE HIMSELF IS TRYING to improve his Twitter-based model. The morning after Trump's election, he met his graduate students and postdocs in the lab. The mood was grim. "Most of them are foreigners," he says, and the anti-immigrant rhetoric of Trump's campaign had been bruising.

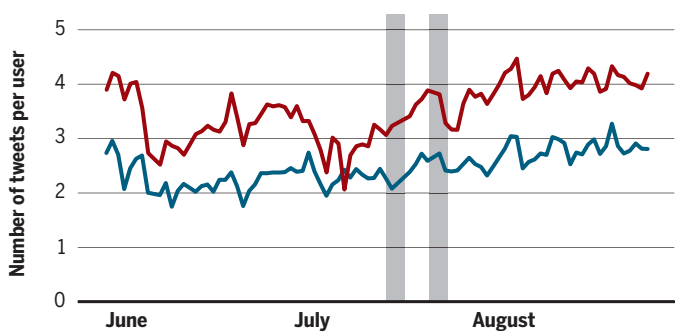
Conventional wisdom

During the party conventions last summer, the number of individuals who tweeted in favor of Hillary Clinton (blue) spiked, far outstripping those who backed Donald Trump (red)—in line with polls that had Clinton in the lead.



The few, the loud

After filtering out bots that retweeted pro-Trump sentiments, researchers found that Trump's backers tweeted far more often than did Clinton's. Election night predictions undervalued that indicator of stronger enthusiasm for Trump.



They did a postmortem on their Twitter study to look for signs they should have seen. Although the Twitter data were far easier to gather than the polls, they were harder to interpret, posing challenges that pollsters need never consider. For example, among the 73 million tweets about Trump or Clinton in the 4 months before the election, how many were written by humans? Twitter's platform allows "bots"—computer programs imitating humans—to take part in the online discussion. They aren't labeled as such, and to most observers they appear to be enthusiastic fellow voters, echoing political slogans and amplifying a point of view. Deploying such bots is like planting people in an audience to laugh at your jokes.

To use Twitter as a voter opinion poll, Makse's team had to detect all those bots and filter them out first. They did that before election night by analyzing not only the content and timing of the tweets, but also information about the accounts behind them. A telltale sign of a bot is an account that does not use one of the standard Twitter software clients and relentlessly retweets content from other accounts verbatim. As they debotted their data, a stark pattern emerged: Whereas the pro-Trump tweeters were riddled with bots, those supporting Clinton seemed almost exclusively human. The effect the bots had on the election remains an open question.

Another unknown is the number of Twitter users who are paid hacks. One of the most influential pro-Trump tweeters of all, based on Makse's analysis of the cascades of retweeting in Twitter's echo chamber, was @LindaSuhler. According to the account profile, it is a "Linda Suhler, Ph.D." The internet has no record of that person and direct Twitter messages from *Science* were never answered.

The close match last November between Twitter-based projections and voter polls could be a fluke. Only more elections will tell. But Makse suspects that both approaches have the same limitation: Certain voters were underrepresented. The Trump voters "just weren't there," he says. The rural U.S. population that propelled Trump to victory may have been the very people not using Twitter or answering pollsters' questions. Either that, he says, or the "shy Trump voter" theory is right: People who intended to vote for Trump tended to keep quiet about it. In retrospect, says Makse, the much higher intensity of tweeting from pro-Trump Twitter users versus pro-Clinton users (see chart, above) "was a red flag" for deep differences between them.

If those biases can be tracked, data from social media could increase the accuracy of election predictions, Makse says. But how much accuracy do we need? There is a downside to psychohistory, Gelman warns. If we could predict election outcomes with perfect fidelity, he says, the act of voting itself "would be meaningless." ■

CREDITS: (GRAPHIC) J. YOU/*SCIENCE*; (DATA) HERNAN MAKES AND ALEXANDRE BOVET, CITY COLLEGE OF NEW YORK