Editorial

US elections: How could predictions be so wrong?

Journal of Marketing Analytics (2016). doi:10.1057/s41270-016-0010-2

INTRODUCTION: THE AWAKENING

The American public has gotten used to statistics – lots of them. Every sports event that gets broadcasted is accompanied by a whole slew of statistics. As commentators comment on individual players' performance, frequent reference to those stats provides a rich context to every move they make. But besides sports, statistics and predictive modeling have penetrated many other areas of life, too. When we surf the web, we are used to getting recommendations (people who bought X, also looked at Y). Crowdsourced data, like collaborative filtering where buyers help other consumers choose wisely, are used for all kinds of ratings. They help us choose where to eat, where to shop, what movies to watch, or where to go on holiday, and many, many other everyday decisions.

More examples of where Statistics have permeated everyday life are weather forecasts, for instance. We are now used to the chance of precipitation for a given day, or even the time of the day. When medical treatments or drugs are evaluated for effectiveness, we habitually make statements like 85–95 per cent of lung cancer cases are smoking related. In genetics, we associate a particular genetic make-up with the odds of developing certain symptoms. Insurance premiums are calculated on the basis of statistics about your age, where you live, gender, and your lifestyle characteristics. We consider these applications of statistics normal, and as such they have become a part of everyday life.

On the morning of the 9th of November 2016, the day after the presidential elections, it was a rude awakening. Despite all forecasts pointing to Hillary Clinton as the new president of the United States, America woke up to Donald Trump as president elect. The question on many peoples' minds was: "How could everyone have gotten their predictions so wrong?!?" This surprising victory has forced many questions around statistics, Big Data, election polling, and quite frankly many other areas of life that have been driven and directed by data for so long. People have grown accustomed to relying on statistics. What was different here? How could this happen?

As the famous Yankee baseball player, then coach, and later contemporary 'philosopher' Yogi Berra has pointed out: "Predicting is hard, especially when it's about the future." The general population, and also broadcasters and news reporters, took the polling aggregators as gospel. And since they all pointed in the same direction, favoring Clinton as the likely winner (some by a whopping 99 per cent probability), the election outcome seemed a lock in. So what went wrong to cause this spectacular breakdown of statistics?

With 20/20 hindsight of the erroneous predictions, lots of explanations and analyses for this aberration have been floated. Some pointed to Trump's social media presence, others to lower than usual voter turnout.

Upon further analysis, there probably isn't much truth to low voter turnout as an explanation, but alas, it has repeatedly been suggested. Since the outcome was so unexpected in light of all the predictions, quite a lot has been written about this remarkable presidential campaign.

TRADITIONAL ELECTION POLLING

For decades, election polling has been an honorable profession that has gotten lots of air time, of course especially around major elections. In the weeks and months leading up to an election, major news channels invite experts and often engage with professional agencies to run proprietary polls. Historically, Gallup was the household name for elections, and they have been doing this kind of work for over 80 years. As time goes by, and with scientific advances, this work has become more sophisticated. The number of agencies that offer such services has grown quite a bit (e.g., Pew Research Center, Gallup, Princeton Election Consortium, Harris, CBS News, and New York Times). There is a long history of tracking relative popularity of candidates, and these polls often begin many months before the actual elections.

For the last decade or so, probably even more so after the publication of his bestseller "The Signal and the Noise: Why So Many Predictions Fail, But Some Don't" (2012), Nate Silver has become the "go to" expert, a household name. Nate Silver has been on a roll. For the presidential elections of 2008, he predicted 49 of 50 states correct. And then in 2012 he got all 50 correct! Along comes the Clinton-Trump race in 2016. Right up to the end, Nate Silver gave Clinton a 70 per cent + chance of winning, and a forecasted 301.6 electoral votes. After the dust had settled we see that in 2016 he got 5 states wrong, and missed the mark on electoral votes by some 70 votes.

How is it possible that someone like Nate Silver, with such a fabulous track record in predicting election outcomes (and some other fields, too, like Baseball), all of a sudden lost his hand? And it wasn't just that Nate Silver missed the mark, he missed it by a landslide! In all fairness, it wasn't only Nate Silver who got the outcome wrong. He just happened to be more visible as the poster child of near perfect election predictions. A glance at the election forecasting landscape, however, tells us that almost all polls and predictions showed Clinton as the new president, and often by a considerable margin, from 70 to 99 per cent.

WHY IS PREDICTING ELECTIONS SO HARD?

Presidential elections happen once every four years. What that implies, is that even if you have been gathering data for 20 years, there are still only five elections you may have covered. And even if you have been gathering data for 20 years, it is highly unlikely that you will have complete time series for all of the variables that you gathered. What is much more realistic is that you have gathered some data for parts of that time span.

To summarize this issue: although gathering data for as long as 20 years seems quite impressive (and I would agree it is), you will still only have five elections that you were able to "predict," which makes it –from a statistical perspective – a rather sparse dataset. What makes the problem of data sparsity even worse is that for many of the variables that you would like to use to predict the upcoming election outcome, you will not have the full 20 years' worth of history. If you consider something like "active Facebook usage" an important variable to predict, then obviously that information was not available 20 years ago, because Facebook didn't exist, yet.

A universal problem with any sampling method is that we calculate statistical inferences on the basis of the assumption that the survey responses are a random, non-

biased selection of the larger population. The risk for bias is high, though, when less and less people are willing to participate in research. The election turnout has been hovering around 60 per cent for the last decade, and even with incentives (which have their own unintended side effects, bias being one of them) a researcher would be lucky to have a response rate of, say, 10 per cent. A low response rate doesn't necessarily mean the results are bad, it merely implies that the risk of bias is greatly elevated.

In many cases, for statistical and technical reasons, you will need to impute ("estimate") missing values from your time series. No matter how clever your data fusion algorithms may be, they will still remain only approximations of reality. As Pyle (1999) has outlined, this is a tricky, and very risky activity. It is unavoidable that time series imputation casts a doubt on *any* prediction that is based on such data. Yet after the missing data have been replaced by best guesses, the risky consequences of doing so can be easily forgotten….

A second factor that makes predicting elections hard is that generalizing – which is the hallmark of *any* kind of prediction – is tricky because the longitudinal time span your data cover is easily long enough for relations between variables to have changed. This type of change in reality happens everywhere, and always makes predicting hard, as Yogi Berra already pointed out. Once you compound it with the problem of data sparsity, it becomes clear why so-called "model engineering" (the analyst inputting domain knowledge to override statistical evidence in a particular data set) is both necessitated as well as extra risky. Let's look at an example from another field to illustrate this problem.

MODEL ENGINEERING

Application credit scoring is used, for instance, when someone applies for a credit card, a cell phone contract, car loan, etc. Over time, the meaning or significance of certain variables that are used to predict will be subject to change, so-called "population drift." For instance, 20 years ago, if someone applied for an unsecured loan (like a credit card) one might have asked for the applicant's phone number, and whether he or she had a land-line and/or a cellphone. For data pertaining to applicants some 20 years ago, we "know" that not having a land-line is associated with a higher risk of credit default. We also know that reality has changed, in the meantime, and nowadays the profile of customers who do own a cellphone, but do not own a land-line, has clearly changed from what that was 20 years ago. That is an intuitive example of what population drift looks like, in practice. In many other cases, we may observe statistical evidence for population drift, but no common sense explanation may be available, yet.

Datasets that are used for credit scoring are (almost) always known to be very sparse. In this case, the sparsity is caused by the fact that there are many more applicants who continue to pay as agreed, than there are customers who acquired a credit product and wound up in arrears. This, of course, is good from a business perspective, but makes the predictive modeling exercise particularly challenging. The defaulted records are the minority class, and the relative sparsity of them makes predicting hard: random fluctuations between credit defaulters make it difficult to arrive at accurate productions about that risk. To improve on that prediction, experienced model builders resort to "model-engineering" (Van den Berg and Breur, 2007).

PREDICTING ELECTIONS AND MODEL ENGINEERING

With regard to the practice of model engineering, analysts who are trying to predict polls face a quandary: we know "the times they are a changin'," yet to preserve some of the

value of historical data collection you want to be rather conservative when it comes to attempting new ways to measure a particular candidate's success. If you change nothing, you may have a 'perfect' historical timeline, but a less relevant and less valid measurement today; yet if you dramatically change the method measurement, how you determine voters' preference too much, you will have very little reference to the historical data, despite getting a more valid measurement *today*. In this respect, pollsters find themselves between a rock and a hard place. In some ways, every new election is unique.

As an example of the effects of model engineering, Nate Silver's FiveThirtyEight had Clinton as steady favorite to win, throughout the campaign. By October 24th 2016, two weeks before the election, they predicted Clinton's chances of winning at 85 per cent. However, when they experimentally recalibrated (!) their model on the basis of more recent polls dating back to just 2000, their prediction for Clinton's chances rose to 95 per cent, according to Nate Silver. Obviously, this is one of those unfortunate scenarios where tweaking the model just a little bit further may have led to overfitting of data, ultimately resulting in a deteriorating performance of the prediction model.

As if all of these problems weren't hard enough, there are genuine concerns about sampling bias during election polls. Sampling bias in and of itself wouldn't be so problematic if it weren't for the fact that the bias itself can change from one election to the next. The sampling bias is compounded with population drift. As some have argued, the population distribution of landlines and cellphones has changed over the past four years, and a lot of polling is still done by phone, making this type of bias a pernicious source of error.

Besides sampling bias, which is intuitively easy to grasp, there is a principled challenge about probabilities that is less intuitive, well described by Chris Adamson in his blog "Probability and Analytics: Reactions to 2016 Election Forecasts." Before the election, Nate Silver gave Clinton a 71 per cent chance of victory. The day after the election, he had egg on his face, but does that mean his model was "wrong"? What his 71 per cent really means is that *if* the elections could have been repeated 10 times, Clinton should have emerged as the winner about 7 out of those 10 occurrences. But of course the elections are run only once, and after that there is only one, certain outcome. Given exactly one possible outcome, that *cannot* render the attribution of probability either right or wrong. That's just how statistics works, a fact that may be poorly understood by the population at large, but news reporters, too.

BIG DATA: THE NEW WAVE

Big Data has been a huge driver of investments and innovations in Silicon Valley for the last decade. Angel investors and established tech companies have invested billions of dollars in software and computer systems. These ventures were launched with the purpose of sifting through gargantuan volumes of data, in the hope of finding useful business insights and trends in consumer behavior. Criticasters have pointed to the risks that these hype cycles bring with them. When companies invest such large amounts of money and hold unreasonable expectations, it can be tempting to present unverified, improperly validated research results, to still the hunger for news, and nuggets of insight.

For those who have been working with Big Data, the dysfunctional patterns behind these unreasonable expectations are clear and obvious. Without critical thinking skills, and lacking adequate context to interpret research numbers, it is all too easy to influence stakeholders who have a vested interest in believing Big Data actually "works" with essentially false information. Those who have been involved in this kind of work know that the data never speak for itself. Data can never

be trusted at face value, or else you put yourself at grave risk of drawing completely inappropriate conclusions. Many (as well as your author) would agree that Big Data holds big promise, but it is a relatively new field, still in its infancy.

It can be no surprise that Big Data was going to be leveraged, too, to (help) forecast the 2016 presidential elections. One of the tenets of Big Data is that various data sources are combined and several of those sources are the result of peoples' digital footprint. Among others, this implies that voters do not perform *any* voluntary action in the process of generating these data. Website usage data, for instance, are created as a "byproduct" of someone surfing the web. The user doesn't have to "do" anything to enable creation of those data, other than his usual surfing behavior.

Big Data sources can be the byproduct of man–machine interactions, like a website visitor going about his usual behavior. But besides man–machine generated data, machine–machine interactions can also be leveraged in much the same way. When a browser makes a call ("page request") to a website, quite a few exchanges take place to serve up that page. Often information about the user's approximate location can be derived from this. That information might be the reason why you would see local advertising banners on a website, for instance.

As society and business are increasingly making the internet their primary means of communication, more and more of this machine–machine data will surface communication and travel patterns. Derived information, about location for instance, is a powerful driver of analytics. The geographic origins of tweets during this past election cycle have been a distinguishing feature to earmark non-human sources of tweets (so-called bots, more on that later).

The fusion of Big Data with more traditional methods of polling holds great potential for advances, especially for real-time tracking of activities and trends. Although the recent experience may have undercut trust in Big Data and statistics, there were notably data scientists leveraging Twitter data who were among the few, and also the early ones pointing to Donald Trump as the most likely president elect. Although this isn't proven technology, yet, many believe that the future of more advanced election forecasting lies here.

It has been pointed out that after Mitt Romney's loss in 2012, the chairman of the Republican National Committee, Reince Preibus, invested $100 million in the party's data. Trump's digital campaign managed to leverage these data by carefully targeting their ads. In order to do so, they built their own database called "Project Alamo," probably named after the battle of Alamo between Texians and Mexicans that took place in San Antonio, and was a pivotal moment in 1836 during the Texas revolution. The Project Alamo database would grow to a list of 220 million people in the US, holding 4000–5000 data points per individual, and is (still) owned by Trump.

These analytical efforts were used to build psychological profiles of voters. The Alamo database not only was obviously instrumental in targeting of ads, but also largely drove the location of Trump's campaign rallies. Steve Bannon, CEO of Trump's 2016 presidential campaign, is a board member of one of the leading Data Science firms, Cambridge Analytica. This can hardly be a coincidence. It has been said that in the final weeks of the campaign, their main focus was to selectively target likely democratic voters, in an attempt to persuade them *not* to vote. Although not scientific proof, in all swing states like Michigan, Ohio and Wisconsin, democratic voter turnout was indeed notably lower.

WHAT MADE THE 2016 ELECTIONS "DIFFERENT"?

In an attempt to explain a posteriori why all the traditional prediction methods had failed so spectacularly, many analysts have resorted

to describing structural differences between the 2016 elections and preceding ones. There is always a risk when doing so, because *given* that the expected election outcome didn't materialize (which we now know, after the fact), you may be tempted to use anything that is different between the 2016 elections and previous ones, and postulate that as an explanation for the erroneous predictions. Needless to say, we do so at great statistical risk. It is exactly these practices that gave data mining in its early stages such a bad reputation. The phrase "torture the data, until they confess" sometimes gets used to make reference to such questionable practices.

There is ample evidence that one of the defining differences between Clinton's and Trump's campaign was the role that social media played. Trump himself claims that Facebook was his defining advantage. Looking at the numbers, another thing is apparent with regard to campaign financing. The bulk of Trump's $250 million in fundraising, was generated online and most of that came from his followers on Facebook. He also leveraged the site to aggressively test some 40.000 to 50.000 variations of his promos per day, to figure out which strategy worked best according to Brad Parscale, who worked for Trump's digital campaign and heads Giles-Parscale from San Antonio, TX. According to Parscale, most of the $90 million digital spend went to Facebook, both for advertising as well as testing of messages.

Whereas Clinton spent more than $200 million on TV ads in the finals months of her campaign, Trump spent less than half of that. Although Clinton certainly has a digital presence, Trump's campaign was fully centered around digital, it was his primary communication channel. Going forward, it seems that acknowledging the power of social media will help candidates garner what is called "earned media": mainstream media coverage that candidates get for free, because it refers to their activity on social media. Trump broke all conventions with regard to traditional models of campaigning.

The sample sizes that a platform like Facebook can offer provided him with the means to constantly test and refine his messages, to tune into what people like and dislike. Facebook users have long been conditioned to "like" messages, and this habitual behavior greatly increases the conversion from reach to response. At its peak, the Trump team ran 175.000 message variations on a single day. Gary Coby, director of advertising and fundraising on Trumps campaign refers to this as A/B testing on steroids.

FAKE NEWS

Ever since this recent presidential campaign, and even more so in the wake of its surprising outcome, there has been a lot of talk about the role that fake news might have played. "News" may traditionally have been defined as the carefully vetted, redacted, cross-checked, and verified broadcasting from news stations and (quality) newspapers. However, for the generation that grew up in the 21st century, the word "news" has taken on a new definition. There are many people who spend considerably more time on social media like Facebook and Twitter, than they do looking at traditional news broadcasts and reading quality newspapers. For this (mostly) young generation, "news" is by and large what shows up on their Facebook or Twitter timeline.

It is safe to assume that frequent Facebook users are generally aware that the information that is presented on their timeline gets selected and ordered by means of sophisticated algorithms. Since this is proprietary Artificial Intelligence (AI) technology, the way information gets organized and ordered will remain unknown to all but a very few engineers and data scientists at Facebook. For the rest of us, it remains a mystery how exactly Facebook

determines what we get to see, when, and why. What is obvious with items on your Facebook timeline, and what also holds for paid advertising, is that if more people are inclined to click on certain content, algorithms tend to favor such content, i.e., place it more prominently.

Peter Cohan, writing for Forbes Magazine, has suggested that a significant part of Facebooks income, potentially as much as *over half of its advertising revenue*, comes from traffic that is drawn to fake news stories. Facebook earns about 30 billion dollar per year with advertising, so if over half of that comes from fake news, we are talking about a whopping 15 billion dollar. If you understand better than others how the timeline gets organized, and have an advertising budget, there is opportunity to game this system. This isn't new, in much the same way Google is continuously altering its search engine algorithms to discourage smart, but not-so-ethical SEO advertisers.

Mark Zuckerberg has announced he intends to clamp down on fake news, but many critics have suggested his attempts are somewhat half-hearted. Given the magnitude of earnings Facebook would stand to lose, this ambivalence is much less surprising. Facebook is not (yet) in a position to simply bar sites that have published fake news, in the same way that Google simply stops displaying results from parties that violate its policies. For the sake of this discussion, we will define fake news as a story that is dressed up or appears as news, but that is based on false information. During this campaign, we have seen several inaccurate and vile attempts at smearing one or the other candidate.

Facebook is unlikely to ever give comprehensive insight into their site's traffic, so the real answer to how much of Facebook's advertising traffic is drawn to fake news will probably never be known. But an interesting recent case was thoroughly researched by the New York Times. On November 9, 2016, the day after the elections, a photograph of a bus in Austin, Texas was posted by Eric Tucker on Twitter. That message was retweeted some 16.000 times, and later went viral as a news story on Facebook.

The news story reported that supposedly a group of people had been paid to participate in a protest against president elect Donald Trump, showing a picture of a bus that was used to transport them. Two days later this "news" was corrected, it turned out that the bus on that picture had instead been used to drive delegates to a software conference by Tableau (Business Intelligence Visualization software). The initial (fake) story was shared 350.000 times on Facebook, and must have drawn millions of views. The correction that was posted two days later was shared a mere 3500 times, 100 times less! Apparently, the truth was much less interesting than a made up story about paid protesters…

Another interesting headline appeared which also proved to be completely false: "FBI Agent Suspected in Hillary Email Leaks Found Dead in Apparent Murder-Suicide." The fake news story suggested a family drama, where an FBI agent from Walkerville, MD, supposedly killed his wife, before committing suicide. This was a completely fabricated news story, published on a poorly assembled fake website that is clearly intended to mimic a legitimate news source. It got shared on Facebook over half a million times, though, and must have generated millions of views.

Since websites, like this fake one, have the potential to generate so much traffic, there is potential to make money. Simply run a Google AdSense marketing campaign and the traffic that is rerouted to third party commercial websites gets converted into cash for the referring website. Needless to say, passing on that internet traffic generates revenues, regardless of whether that referral was a genuine or fake news website. It is easy to see how these schemes can support themselves. The sheer fact that some news item has the potential to draw attention makes it a commercially viable topic to publish.

TWITTER BOTS

Although one might consider Twitter an "open" medium, the recent election has shown that news – if one would choose to refer to a Twitter timeline as "news" – can be influenced and distorted by tweets that have been generated by non-human contributors. In a study published a day before the election, it was found that some 400.000 bots were operating on Twitter, actively producing election-related content. They were tweeting and being retweeted, and together managed to generate approximately 20 per cent of all election-related messages. Researchers Bessi and Ferraro (2016) of the University of Southern California found that these bots were quite influential, capable of distorting online debate.

Bessi and Ferraro (2016) found that humans and bots were both retweeted at about the same rate, which casts serious doubt if people are able to discern whether a source is human or not. On the basis of identical retweet probabilities, they concluded that people are not able to discern between real humans tweeting and bots generating these tweets. During the months preceding the last election, they discovered that nearly 75 per cent of bots were found supportive of the Republican candidate Donald Trump. Since bots have the potential for being much more active, once people interact with them they can become more influential. Their study showed that on average, bots were about twice as active as humans.

Social media campaigns that are launched with evident manipulative intent are sometimes referred to as astroturf or Twitter bombs. Because of the fickle nature of account assignment, usage of proxy servers, and outright deceptive behavior of bots, it is often near impossible to determine who is behind these activities (Kollanyi and *et al*, 2016). It can be challenging to identify with certainty whether an account is operated by a bot, or a human. To flag this, machine learning technology was applied. Bessi and Ferraro used a fairly conservative threshold to classify tweets as originating from humans, or not, using the well documented "Bot or Not" algorithm.

Based on their extensive research, covering over 20 million tweets, posted by nearly 2.8 million distinct contributors, Bessi and Ferraro convincingly demonstrated that bots can successfully engage in political debate, and have a material impact on human participants. As this technology evolves, and the smartest bots begin to behave more and more like humans, the potential for manipulation is a genuine threat to democracy that can materially affect election outcomes. Obviously, a scary perspective.

CONCLUSION

Predicting elections is hard. For various reasons, most importantly their relatively low frequency and high rate of change among predictive variables, predicting the outcome of elections with high certainty is a near impossible task. There is also uncertainty about sampling error, because when you rely mostly on telephone survey sampling (current practice), this introduces a bias when the possession of phones in the population changes over time. This seems to be the case since the widespread acceptance of cell phones. Sampling methods or else weighting schemes need to cater to that shift.

Because election data are so sparse, and population drift is a factor that threatens traditional methods, some degree of model engineering will be needed. How else can we blend in the Big Data models that are now being built on the back of, for instance, Twitter data alone? There simply isn't enough historical data to reliably incorporate important variables that pertain to social media like Facebook and Twitter. Few would deny that those factors have played a significant role in the recent elections, and there is already scientific evidence that these

influences date back at least 10 years (Howard, 2006) and also extend outside the geographical boundaries of the US (e.g., Gibson and McAllister, 2006, or Enli and Skogerbø, 2013). To incorporate these effects, at least some data imputation and model engineering will be called for.

Model engineering and its perils have not been documented much in the literature. Yet in areas with very sparse data sets, professionals in specialized functions, like the ones who build credit scorecards, are very familiar with it. While every credit scoring population is a little bit different, most analysts have broad experience. In this field, many data analysts have been building credit risk scorecards 10 + years, and have handled 100s of data sets. This experience gives data analysts abundant context to "insulate" themselves from idiosyncrasies in any particular data set. Presidential elections simply don't happen often enough to gather a similar amount of experience! Yet for a number of reasons, time series do require missing value imputations, and this alone invalidates even an objective assessment of the predictive model's confidence interval.

Big Data will surely impact the way we make predictions. As our world turns increasingly digital, and the means and technology to capture consumers' digital footprint continue to mature, a much richer picture of behavior emerges. Blending and combining all this information is a non-trivial task, but rewarding and very promising from the standpoint of dramatically improving our understanding and insight.

Part of what made the 2016 election so different from every previous one was the dominant role that social media played in reaching voters. Maybe even more importantly, some of these digital platforms allow extensive testing and honing of messages in an attempt to influence voter behavior. Some argue, after the fact, that Trump's team did a more convincing job of leveraging these new, Big Data opportunities.

We will never know for sure whether this was the decisive causal factor, but Trump's "digital first" strategy paid off for him in spades. His team of data scientists and media specialists leveraged all data available to them, which fitted will with Trump's personal preference for Twitter as his medium of choice. There is no indication that they themselves *generated* any fake news, but their campaigns probably benefitted more form it than Clinton's team did. In part because Clinton spent less on digital advertising, and in part because more Twitter bots supported Trump, and hence had more opportunity to influence discussion on Twitter and so-called social media echo chambers.

The net result of this change in campaign strategy was that Clinton, although certainly invested in digital, too, was focused more on traditional channels like TV ads. Trump, on the other hand, spent almost all of his campaign money (some $275 million) on Facebook, and to lesser extent Instagram. Given the magnitude of "buzz" and free publicity (called "earned channel") this generated, we can expect future campaigns to make similar inroads. Only time will tell what an optimal mix is, there. With 20/20 hindsight it is evident, though, that these changes played a role in misleading the outcomes of traditional forecasting techniques.

In part, the current disillusionment with election forecasting isn't so much a problem with statistics per se (although possibly the models did "underperform"), but more so a problem with peoples' incomplete understanding of how statistics *works*. This is at least as much a governance issue, as it is a problem with shortcoming in predictions. As the famous statistician George Box has stated: "All models are wrong, but some are useful."

In hindsight, we now know that the election forecasts in the closest states (typically called the "swing" states) were off, on average, by 3.4 per cent. Although that is a much bigger gap than we had seen in more than 20 years, similar margins of errors did

occur in the elections of 1988 and 1992. So, again, if you are familiar with the statistical implications of saying a candidate is a 70 per cent favorite to win, that (still) implies that if you could repeat such an experiment multiple times, you would expect to see some distribution of error margins. In that light, a 3 per cent + margin of error doesn't mean election polling is useless, it only means that you might have to wait a while before you are likely to see such big discrepancies again, as after this election.

With the rise of Big Data, it is to be expected that advances in that field will be leveraged even more in future campaigns. This will likely hold for both efforts to predict election outcomes, as well as efforts by candidates to influence the outcome in their favor. Mostly anecdotal evidence suggests that Trump's "digital first" strategy and his willingness to test and hone his messages in A/B testing may have given him the decisive advantage.

Most panel research methods have changed over the past decade or two, to reflect economic access to participants online. Much research work has also moved online to leverage the potential to scale projects quickly, and provide (much) quicker answers than traditional paper-and-pencil methods ever could. It is only natural to expect that advances in data blending and Big Data solutions will help improve the accuracy, as well as timeliness, of election forecasting results.

REFERENCES

Adamson, C. (2016) Probability and Analytics: Reactions to 2016 Election Forecasts. http://blog.chrisadamson.com/2016/11/probability-and-analytics-reactions-to.html

Bessi and Ferraro (2016) Social Bots Distort the 2016 U.S. Presidential Election Online Discussion. First Monday, Vol. 21, No 11. http://firstmonday.org/ojs/index.php/fm/article/view/7090/5653

Enli, G. and Skogerbø, E. (2013) Personalized campaigns in party-centred politics: Twitter and Facebook as arenas for political communication. *Information, Communication & Society* 16(5): 757–774.

Gibson, R. and McAllister, I. (2006) Does cyber-campaigning win votes? Does cyber–campaigning win votes? Online communication in the 2004 Australian election. *Journal of Elections, Public Opinion and Parties* 16(3): 243–263.

Howard, P. (2006) New Media Campaigns and the Managed Citizen. New York: Cambridge University Press.

Kollanyi, B. et al (2016) Bots and automation over Twitter during the first U.S. Presidential debate. *COMPROP Data Memo* 2016.1 (14 October), at http://politicalbots.org/wp-content/uploads/2016/10/Data-Memo-First-Presidential-Debate.pdf.

Pyle D. (1999) *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann.

Silver, N. (2012) *The Signal and the Noise*. New York: Puffin Books.

Van den Berg, B. and Breur, T (2007) Merits of interactive tree building – Part 2: How to do it. *Journal of Targeting, Measurement and Analysis for Marketing*. 15: 201–209.

Tom Breur
Marlborough, MA, USA
e-mail: tombreur2@gmail.com