Reports of the Death of Big Data Have Been Greatly Exaggerated

Daniel Castro | March 6, 2017

Can we still trust data? Following a series of surprising election results in 2016 that defied the predictions made by some of the world's most talented pollsters, some critics seem to think the answer is a resounding no. But while those who forecast elections certainly made mistakes, these failures do not mean that other uses of predictive analytics should be discarded. To the contrary, a close look at U.S. presidential election predictions shows that more investment in data, not less, is the way to avoid replicating these types of problems in the future.

Given that virtually every major pollster predicted before the election that Clinton would defeat Trump, some have suggested that the enormity of this mistake will lead to the demise of the entire voter projection industry. Others think the impact will be even more far-reaching. Patrick Tucker, the technology editor of Defense One, argues, "If 'big data' is not that useful for predicting an election, then how much should we be relying on it for predicting civil uprisings in countries where we have an interest or predicting future terror attacks?"

While these responses are understandable, they are still overreactions. The problems with these forecasts could have been avoided by remembering a few key principles.

First, predictions are not guarantees. Many of the election results, especially of the popular vote, are consistent with pre-election polling and differences fall within the margin of error. Any serious data scientist will admit that even the best models cannot be accurate 100 percent of the time. Uncertainty is a fact of life. Nobody could have predicted the impact that FBI Director James Comey's late-stage intervention would have on the final tallies. Organizations should use data to be better prepared for the future, but they should remember that they must still plan for contingencies. This will require data-literate executives who understand the assumptions in the underlying statistics.

Second, measure the right thing. This type of mistake is far too common. During the Vietnam War, the U.S. government used body counts to measure the military's progress, only to discover later that this was a poor indicator of the war's success. If existing data does not accurately measure a particular phenomenon, then additional data may be needed.

During the presidential campaigns, one of the few models to correctly forecast the election relied not on polling results, but on online behavior, such as social media engagement. These types of models also avoided the so-called " Bradley effect," where polls differ from election results because some voters are reluctant to tell pollsters their true opinion. Data-driven organizations need to routinely validate whether the data they have is an accurate measurement of reality.

Third, data quality matters. The pre-election polls misjudged how likely certain people were to vote and these underlying assumptions polluted the quality of the data used in most voter prediction models. Data quality should remain a top priority for organizations, as inaccurate or incomplete data can limit the utility of predictive models. Organizations should explore how new data sources, such as data produced by connected Internet of Things devices or social media, can provide better and more timely information than traditional sources.

Finally, models get better with age. One reason that pollsters got it wrong in 2016 is that they had few opportunities to refine their models — presidential elections only take place every four years. The best predictive models, such as those used to predict weather or recommend movies, have often gone through multiple iterations based on enormous data sets on predictions and outcomes. The only way organizations can get better at data analytics is by building models and then assessing them.

So while the past year offers many lessons on how not to use data, it is premature to suggest that these errors have shaken the pillars of today's data economy. Instead, it should be an important reminder that data-driven innovation requires organizations to invest time, talent and treasure to produce the desired results.

Government agencies that double down on data are likely to find themselves ahead of their peers in the future, while those that pull back may find themselves reliving 2016.

http://www.govtech.com/opinion/Reports-of-the-Death-of-Big-Data-Have-Been-Greatly-Exaggerated.html