To get better at forecasting big political events, we need both better data and sharper reporting, a clearer read on the numbers, and a more penetrating portrait of on-the-ground realities. —Aaron Timms

THE RESULTS OF RECENT VOTES, particularly the U.S. presidential election and the United Kingdom’s referendum on leaving the European Union (better known as Brexit) left many surprised. In both cases, the postdecision lament went something like this: “In this age of big data, how could the pollsters and pundits have been so wrong in their predictions?” • I’m just a language guy, so I don’t pretend to have an answer, although surely it’s an open question whether polls, with their sample sizes in the few thousands, count as “big” data. (Polling statistics probably fall more under the rubric of medium data.) Perhaps that was the problem: If preelection and pre-referendum analyses could have accessed data points in the millions, then might the results have been less surprising? Or perhaps what’s needed isn’t big data on its own but an approach that takes advantage of the many new types of data that are available. • For example, fast data refers to data that requires near-instantaneous access or analysis or that is relevant for only a very short time. It’s an example of hot data, which is used constantly, so it must be easily and quickly accessible. On the opposite side of the information coin we have slow data, which accumulates over a relatively long time, meaning that at some point it might become long data, which extends back in time hundreds of years. It’s an example of cold data, which is used relatively infrequently, so it can be less readily available. Whether fast or slow, hot or cold, the information isn’t much use to anyone if it’s dirty data, which is incomplete, inconsistent, or just plain wrong. • In the same way that dark matter is an unseen but very large part of the cosmos (some estimates peg dark matter at 27 percent of the mass of the universe), dark data represents the unseen but very large part of the data that most corporations collect and store. It’s “dark” because corporations don’t use it for analysis, insight, or decision making. Some of it is transient data, such as unused sensor data or temporary network routering information, and live data, such as a user’s (changing) GPS coordinates. Occasionally, this data can produce perishable insights: valuable data that has a very short shelf life (such as when you detect that your roaming user is wandering past one of your brick-and-mortar storefronts). The opposite is target-rich data. When tagged, processed, and analyzed, this data offers its owner valuable, long-term insights. Perhaps there’s a way to synthesize both the big picture and the small—that is, to somehow combine both big data and our contributions to that data: small data, which arises out of our everyday actions. We’d have to navigate some dangers. For example, we’d need to ensure that our data does not become cubed data, where a third party shares our data with another third party and it becomes impossible to predict where it ultimately ends up or how it will be used or interpreted. We’d need some assurances that third parties practice responsible data, information that is used and shared sensitively and humanely.

A more promising approach might be one that utilizes thick data, which combines both quantitative and qualitative analysis. The pundits could take a cue from narrative medicine, which uses the story of a patient’s illness combined with traditional medical practices as a way of understanding, diagnosing, and treating the illness. Rather than making guesses about what people will do—for example, that rural voters would stay home on Election Day or that people who told pollsters they’d vote “Leave” would do the opposite in the voting booth—pundits could actually talk to people and listen to their stories rather than just looking at the numbers. Call it narrative data.