The New York Times

June 1, 2013, 8:00 am

Why Big Data Is Not Truth

By QUENTIN HARDY

The word “data” connotes fixed numbers inside hard grids of information, and as a result, it is easily mistaken for fact. But including bad product introductions and wars, we have many examples of bad data causing big mistakes.

Big Data raises bigger issues. The term suggests assembling many facts to create greater, previously unseen truths. It suggests the certainty of math.

That promise of certainty has been a hallmark of the technology industry for decades. With Big Data, however, there are even more hazards, some human and some inherent in the technology.

Kate Crawford, a researcher at Microsoft Research, calls the problem “Big Data fundamentalism — the idea with larger data sets, we get closer to objective truth.” Speaking at a conference in Berkeley, Calif., on Thursday, she identified what she calls “six myths of Big Data.”

Myth 1: Big Data is New

In 1997, there was a paper that discussed the difficulty of visualizing Big Data, and in 1999, a paper that discussed the problems of gaining insight from the numbers in Big Data. That indicates that two prominent issues today in Big Data, display and insight, had been around for awhile.

“But now it’s reaching us in new ways,” because of the scale and prevalence of Big Data, Ms. Crawford said. That also means it is a widespread social phenomenon, like mobile phones were in the 1990s, that “generates a lot of comment, and then disappears into the background, as something that’s just part of life."

Myth 2: Big Data Is Objective

Over 20 million Twitter messages about Hurricane Sandy were posted last year. That may seem sufficient for a picture of whom the storm affected. However, the 16 percent of Americans on Twitter tend to be younger, more urban and more affluent than the norm. "Very few tweets came out of Breezy Point, or the Rockaways," Ms. Crawford said. "These were very privileged urban stories." And some people, privileged or otherwise, put information like their home addresses on Twitter in an effort to seek aid. That sensitive information is still out there, even though the threat is gone.

That means that most data sets, particularly where people are concerned, need references to the context in which they were created.

Myth 3: Big Data Doesn't Discriminate

"Big Data is neither color blind nor gender blind," Ms. Crawford said. "We can see how it is used in marketing to segment people." Facebook timelines, stripped of data like names, can still be used to determine a person's ethnicity with 95 percent accuracy, she said. Information like sexual orientation among males is also relatively easy to identify. (Women are tougher to pinpoint.) That information can be used to determine what kind of advertisements, for example, that people receive.

It's important to remember that whenever people start creating data sets, these become fallible human tools. "Data is something we create, but it's also something we imagine," Ms. Crawford said.

Myth 4: Big Data Makes Cities Smart

"It's only as good as the people using it," Ms. Crawford said. Many of the sensors that track people as they manage their urban lives come from high-end smartphones, or cars with the latest GPS systems. "Devices are becoming the proxies for public needs," she said, "but there won't be a moment where everyone has access to the same technology." In addition, moving cities toward digital initiatives like predictive policing, or creating systems where people are seen, whether they like it or not, can promote lots of tension between individuals and their governments.

Sorry, IBM. Take that, Cisco. That goes for you, too, Microsoft, Ms. Crawford's employer. All these big technology companies have Smart Cities initiatives.

Myth 5: Big Data Is Anonymous

A study published in Nature last March looked at 1.5 million phone records that had personally identifying information removed. It found that just four data points of when and where a call was made could identify 95 percent of individuals. "With just two, you can identify 50 percent of them," Ms. Crawford said. "With a fingerprint, you need 12 data points to identify somebody." Likewise, smart grids can spot when your friends come over. Search engine queries can yield health data that would be protected if it came up in a doctor's office.

Myth 6: You Can Opt Out

Last December, Instagram, the photo-sharing site, changed its terms of service to allow it to share customer's photos more broadly, even use images in ads. What it didn't have was a paid option, in which a person could, for a fee, not be part of that. Even if that option existed, Ms. Crawford said, this would imply a two-tier system — people who could afford to control their data and those who could not. "Besides," she said, given the ways that information can be obtained in these big systems, "what are the chances that your personal information will never be used?"

Before Big Data disappears into the background as another fact of life, Ms. Crawford said, "We need to think about how we will navigate these systems. Not just individually, but as a society."