MIT Sloan Management Review

The Big Impact of Small Data Errors

Big Idea: Data & Analytics • **Blog** • August 29, 2017 • Reading Time: 3 min

Sam Ransbotham

Even when the data itself is solid, mistaken connections are sometimes made during analysis. But with vigilance, managers can avoid a data mishap.

A 2008 news report led with reports of Russian tanks and troops surging into Georgia — accompanied by a map mistakenly depicting the invaded territory as Savannah, Georgia, rather than the homonymous Eastern European country. While the story was correct, Google News' map selection was not **(https://news.slashdot.org/story/08/08/09/213236/google-news-has-russian-army-invading-savannah-ga)**.

Born and bred in Georgia as I am, any initial fears for my kinfolk were quickly dispelled when a cursory investigation revealed that the map of the *country* of Georgia would have been a better image, not the map of the U.S. state of the same name. This was a simple case of misidentifying the value "Georgia" when associating the news data with the map data.

But other cases of data misidentification are not so simple and carry greater consequences. To the chagrin of many people sharing names with others with more nefarious tendencies, a false match to the no-fly list **(https://en.wikipedia.org/wiki/No_Fly_List#False_positives)** can be quite an inconvenience. Images, for example, can be linked to people. Advances in image enhancement and processing are yielding growing prowess in facial recognition — and growing concerns about misidentification **(https://www.fastcompany.com/3069264/congress-fbi-face-recognition-real-time-street-lineup)**.

In a recent, painfully public episode, online vigilantes used the copious amounts of image data from the recent Charlottesville protests to identify participants — and in at least one case, it seems that an unrelated person was swept up in the fervor **(http://www.npr.org/sections/alltechconsidered/2017/08/17/543980653/kyle-quinn-hid-at-a-friend-s-house-after-being-misidentified-on-twitter-as-a-rac)**. The consequences for this person were significant, causing him to hide until the emotional upheaval subsided. Yet we've been here before: In the aftermath of the Boston Marathon bombing, several people were similarly misidentified from image data **(http://nymag.com/daily/intelligencer/2013/04/wrongly-accused-boston-bombing-suspects-sunil-tripathi.html)**. The consequences for incorrect linking (such as linking image data to the wrong person) can be far more serious than including the wrong map in a news report — something easily corrected.

In these individual cases, there was someone to notice and complain about the misidentified data. The error could be highlighted for further scrutiny and correction, however difficult, could begin. But with large volumes of data to link, the chances are higher that such misidentifications go undetected.

For example, data processing steps in handling genetic data can cause genes like "Septin 2" (abbreviated as "SEPT2") to be interpreted as "Sept. 2" — and get dropped in subsequent analysis as the records then silently fail to match reference data about the gene **(https://www.washingtonpost.com/news/wonk/wp/2016/08/26/an-alarming-number-of-scientific-papers-contain-excel-errors)**. More commonly, numbers such as "01234" are frequently used to uniquely identify items in databases, such as customers or policies. But "01234" stored in string format won't necessarily match record "1234" stored in numeric format when analysts join database tables — and SQL by default drops records without corresponding matches.

Misidentifications such as these are far more insidious. Data silently disappears from analysis, requiring vigilance to notice. Experienced developers and analysts know to be alert for these instances, but as analytics becomes more pervasive, the experience necessary won't likely travel as quickly — just as vigilantes may lack vigilance.

Given the potential for misidentified data, what can managers do?

Invite scrutiny. Thinking critically about analysis and the data behind it can help. For example, "data antagonists" can be valuable in countering confirmation bias **(http://sloanreview.mit.edu/article/for-better-decision-making-look-at-facts-not-data/)**. But the probing from an adversarial stance **(http://sloanreview.mit.edu/article/detecting-bias-in-data-analysis/)** can also help uncover misidentification. What processes does your organization have to detect misidentification? Does your organizational culture encourage the sort of questioning that helps root out these types of errors?

Accept some inevitability. Given the vast quantities of data available about many aspects of business, it is impractical — if not impossible — to detect every possible misidentification. Some mistaken associations are bound to occur. In some cases, it may not make a difference. If the errors are sparse and close to random, then the analysis may not suffer — even incorrect data provides value **(http://sloanreview.mit.edu/article/analytical-value-from-data-that-cries-wolf/)**. But systematic errors are far more troublesome — like missing everyone born in New England because of the tendency for Social Security numbers there to begin with a leading "0," which gets dropped when the SSN is converted to a true number. Is there potential for systematic error in the values your organization relies on to make linkages?

Consider the context. The Georgian map misidentification seems to have had few consequences — it likely mattered little that readers briefly had the wrong Georgia on their mind. Contexts like these may not merit significant oversight. But other misidentifications may lead to loss of resources, reputation, liberty, or life. These clearly deserve more extensive processes to detect mismatches. How important are specific analytics results to your organization or your organization's customers?

ABOUT THE AUTHOR

Sam Ransbotham is an associate professor of information systems at the Carroll School of Management at Boston College and the *MIT Sloan Management Review* guest editor for the Data and Analytics Big Idea Initiative. He can be reached at sam.ransbotham@bc.edu and on Twitter @ransbotham.