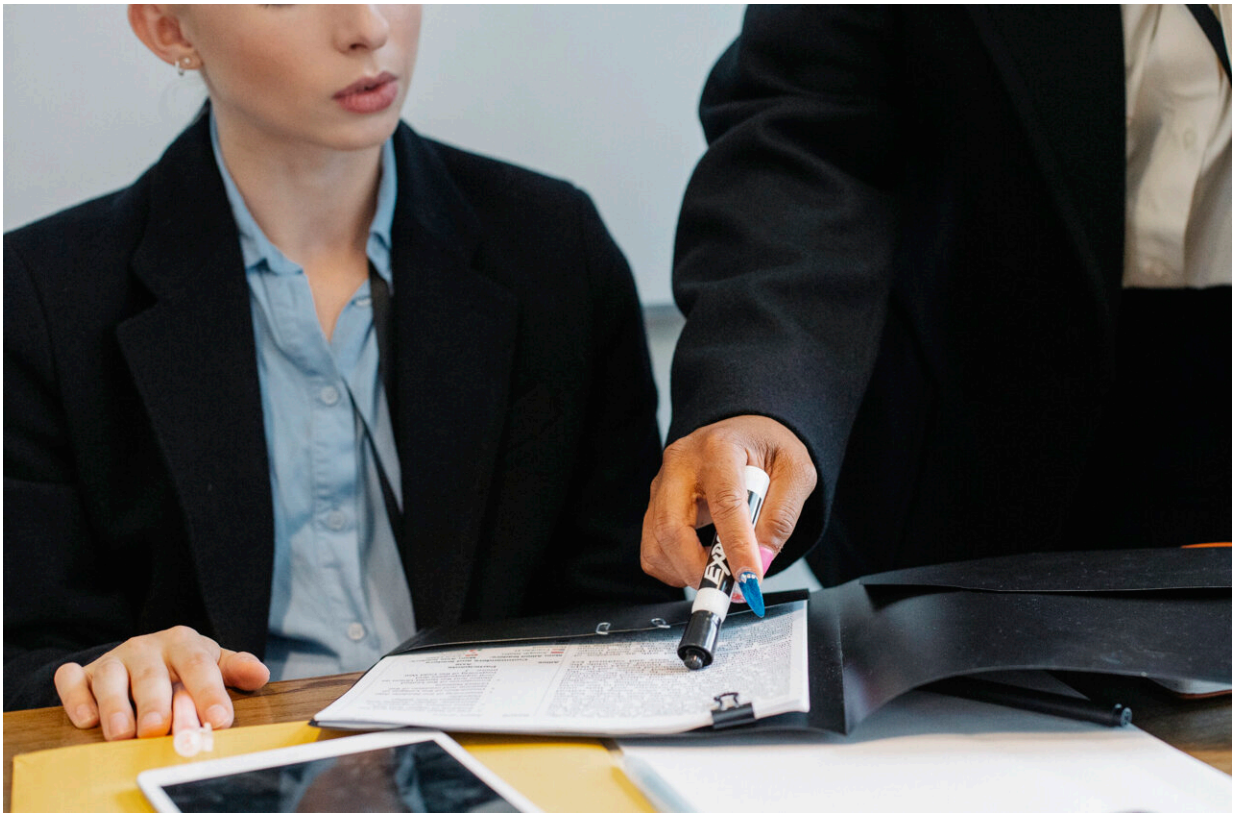
Big Data analyses depend on starting with clean data points

August 5 2015, by H V Jagadish



Credit: Sora Shimazaki from Pexels

Popularly referred to as "Big Data," mammoth sets of information about almost every aspect of our lives have triggered great excitement about what we can glean from analyzing these diverse data sets. Benefits range

from better investment of resources, whether for government services or for sales promotions, to more effective medical treatments. However, real insights can be obtained only from data that are accurate and complete, so it's critical to keep in mind how the data were collected.

Data scientists know the importance of accurate and complete data. After all, if the data itself is unreliable, you'll wind up making invalid conclusions based on your analysis.

To avoid that pitfall, one major cost for most data analysis projects comes from data preparation and cleaning – that is, finding and correcting errors in the data. These errors include incorrect values, missing entries, aliasing (where information about two distinct entities has been merged in error, for example, because two people have the same name) and multiple entry (where information about the same entity is split up, for example, because the name has been spelled differently for the same person). When data sets are small, the analyst can manually examine and validate each entry. With large data sets, we have to rely on computer-executed algorithms. The development of such algorithms is now a subfield itself.

The old truism "garbage in, garbage out" is more apt than ever in this era of complex and gargantuan data sets – and the sometimes weighty consequences of trusting what they seem to imply.

How inaccuracies creep in

Errors in data can arise for a variety of reasons. For example, users often make mistakes when filling in web forms. Data cleaning software can verify that the zip code matches the street address, and possibly even correct it. So if the state has been entered along with the town in the city field (for example, "Plainfield, NJ" for city), data cleaning can move the state entry to the correct field. Or if a street has only house numbers

1–80, data cleaning software can flag as erroneous a house number entered as "125." Many inadvertent errors can be caught, and possibly fixed, by clever software.

Bad data entry isn't the only source of inaccuracies. One common place where errors arise is in linking data across data sets. Unless both data sets use a unique identifier – such as a social security number – with each entry, it is challenging to match entries across data sets: there are likely to be entries that wind up linked even though they should be distinct, and entries that are not linked even though they correspond.

Another frequent source of mistakes is when computer software creates table entries based on other, more complex, data. For example, if you write a review of a product, this may be condensed into one of a few buckets (eg, loved/liked/hated) along a few simple axes (eg, ambiance, food taste, service, value for money). The condensed form is amenable to quantitative analysis, which the original text form is not. But errors can be made in the process of condensing.

At least don't motivate people to lie

Dirty data are almost impossible to clean when errors are due to intentional user choice as opposed to inadvertent causes. Suppose you enter your neighbor's address as yours: clever software cannot catch this lie without knowing more about you – after all, the address entered is technically a valid entry, it's just not correct.

If we are to trust the results of analysis, we must ensure that the data collection procedures at least don't give users incentive to cheat.

Consider web forms that routinely ask us to fill out information about ourselves. Many users enter a bogus email address in these forms, perhaps for fear of possible spam mail. Some websites confirm the email

address entered, for instance, by sending a verification link that the user has to click. But such verification is expensive and unfriendly. The complementary approach is for the website to develop a reputation for trustworthiness so that users are willing to share their email addresses without worrying about the potential for misuse.

In fact, people (and businesses and other entities) will provide correct and complete data only if they feel they can trust the data collection. The US Census Bureau is able to collect high-quality data because it can assure citizens that what they report in the census will not be used for tax collection or any other such government purpose, other than statistical reporting. While it might be desirable to catch tax cheats and obvious that census data could greatly enhance the government's ability to identify them, laws in most countries prevent such use of census data, because the moment citizens know census data can be used for tax computation, they will be motivated to lie to the census-taker.

Big data can't outsmart high-stakes incentives to lie

Maybe you don't really care whether or not you get the right targeted weekly email highlighting sales of possible interest to you at a local chain store. But there are certainly other instances where the stakes for big data accuracy are much higher.

For instance, take the current spotlight on German privacy laws centered on the mental health of pilot Andreas Lubitz. He allegedly crashed a plane intentionally into the Alps and killed 150 people in March. Given his mental health, he probably should not have been flying an airplane. Some people advocate that his employer, Lufthansa, parent company of Germanwings, should have had complete access to Lubitz's mental health record and thus been able to keep him out of the cockpit before he had a chance to bring down a flight.

But weakening privacy laws would not reveal to authorities the true mental health of people like Lubitz. Rather, it would make it less likely that the official health record is a reliable record of fact. Someone like Lubitz, who is keen to fly and dreams of becoming a pilot, would likely do everything possible to hide any disqualifying condition from his official medical record if he knew it could be used against him. The incentive for omission and falsehood would undermine the ability to collect and use a reliable data set. In this case, privacy would be sacrificed without any safety payoff. Much better to keep the medical record data clean, and qualify pilots through tests run outside the formal medical system.

It's great for us as a society to make use of all the data resources we have. But it's important not to ruin the quality of this data resource in our enthusiasm to use it, even if with good intentions. Unless we are careful about how we deploy these big data sets, we'll collect data of poor quality – particularly so where there are individual points of concern, such as Lubitz's health record. The inferences we draw from big data are only as good as the individual data points we feed in.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Source: The Conversation