# How to handle Big Data

The Hollywood film *Moneyball* (2011) is about the Oakland Athletics baseball team and the attempt by its manager to put together a competitive team on a lean budget using data and computer analytics rather than depending on mere biases to recruit new players. The film stands out for focussing the spotlight on data science by showing that the art of data science is more about asking the right questions than just having the data.

It is difficult to imagine the great volume of data we supply to different agencies in our everyday actions, bit by bit through surfing the Internet, posting on social media, using credit and debit cards, making online purchases, and other things where we share information about our identity. It is believed that social media and networking service companies such as Facebook may already have more data than they are leveraging. There are infinite ways to slice and dice data, which itself is quite daunting as at every step, there is potential to make huge mistakes.

Careful data mining from Big Data might help understand our behaviour in order to facilitate planning. But there are examples of blunders being made with a load of information at one's fingertips. The problem with so much information is that there is a much larger haystack now in which one has to search for the needle.

Here is an example. In 2008, Google was excited about "Big Data hubris" and launched its much-hyped Google Flu Trends (GFT) based on online searches on Google for flu-related information with the aim of "providing almost instant signals" of overall flu prevalence weeks earlier than data out out by the Centers for Disease Control and Prevention (CDC), the leading national public health institute in the U.S. But much of it went wrong; GFT missed the 2009 swine flu pandemic, and was wrong for 100 out of 108 weeks since August 2011; it even missed the peak of the 2013 flu season by 140%. Google tried to identify "flu" with the search pattern. Usually, about 80-90% of those visiting a doctor for "flu" don't really have it. The CDC tracks these visits under "influenza-like illness". Understandably, the Net search history of these people might also be an unreliable source of information. While GFT was promoted as the poster project of Big Data, it eventually became the poster child of the foibles of Big Data. In the end, it focussed on the need for correctly using the limitless potential of Big Data through efficient data mining.

Data blunders often arise out of bias, low-quality data, unreliable sources, technical glitches, an improper understanding of the larger picture, and lack of proper statistical tools and resources to analyse large volumes of data. Moreover, Big Data invariably exhibits fake statistical relationships among different variables, which are technically called "spurious correlations" or "nonsense correlations". Relying too heavily on a particular model is also a common mistake in Big Data analyses. Therefore, the model should be wisely and carefully chosen according to the situation.

"Big data may mean more information, but it also means more false information," says Nassim Nicholas Taleb, author of *The Black Swan: The Impact of the Highly Improbable* However, today's world is obsessed with collecting more and more data while being inattentive to the necessity or capacity to use them. There is a possibility of getting lost in the waves of data. As a statistician, I firmly believe that unless there is serious need, we should restrain ourselves from collecting data as searching for the needle in haystacks should not be made unnecessarily difficult. The errors are bound to increase exponentially with more and more redundant information.

Mining and geological engineers design mines to remove minerals safely and efficiently. The same principle should be adopted by statisticians in order to mine data efficiently. Big Data is more

complex and involves additional challenge. They might involve the use of some skills involving analytics, decision-making skills, logical thinking skills, problem-solving, advanced computational expertise and also statistical expertise. So, using some routine algorithm is not enough. Too much reliance on available software is also a serious mistake.

So, where are we headed with so much data, most of which are useless? What is the future of so much reliance on data, where a lot of spurious correlations could dominate our lifestyle and livelihood? Let me bring in another Hollywood film, Spielberg's *Minority Report* (2002) which is set in Washington DC in 2054, where the 'PreCrime' police force is capable of predicting future murders using data mining and predictive analyses. However, when an officer is accused of one such future crime, he sets out to prove his innocence! Does data mining depend more on probabilistic guesswork much like the danger inherent in a game of dice? Does this Spielberg film depict the future of data mining? And is the future dystopian or Utopian?

*Atanu Biswas is Professor of Statistics at the Indian Statistical Institute, Kolkata*



Marriage is a civil contract — adultery or divorce should have only civil consequences

END