Source: *TODAY*, p24
Date: 26 June 2015

DISCOVERING THE ROOT CAUSES OF DATA TRENDS, AND CONSUMER AND FIRM BEHAVIOURS

Is data analytics about causes ... or correlations?

This commentary is part of a series in TODAY*'s Science section, created in collaboration with the National University of Singapore's School of Computing, that explores the computer science research projects conducted here.*

GOH KHIM YONG

Does eating more chocolate make a country's population smarter so it will win more Nobel Prizes? Or, is there a third causal factor that could explain the relationship between chocolates and Nobel Prizes? These are the kinds of questions that big-data analytics researchers like me attempt to answer.

Data science and big-data analytics are all the hype right now. The total global market value for big-data hardware, software and services is projected to be worth some US$114 billion (S$153 billion) by 2018.

However, many current data analytics efforts in both academia and industry focus on deriving correlational insights from large data sets which, while useful, have their pitfalls. One example is Google Flu Trends' failure since 2011 to accurately predict actual flu incidences from flu-related search phrases, despite the company's confidence in its ability to leverage on its data.

Another is the aforementioned bizarre relationship between chocolate and Nobel Prizes: A paper published a few years ago in the established New England Journal of Medicine claimed a link between chocolate consumption and the number of Nobel Laureates produced by a country.

A simple explanation is that correlation does not imply causation. Many big-data examples that have received popular attention, such as the Google case, may not have the most valid or comprehensive instrumentations or measurements for rigorous analysis, especially from a causal perspective.

In the case of the chocolate-Nobel Prize example, other, more important factors that could be causal in nature, such as the countries' Human Development Index and per capita income, were missing in the analysis.

UNDERSTANDING DATA TRENDS, BEHAVIOURS

My research in data analytics seeks to uncover the underlying causes of data trends, and consumer and firm behaviours. Such an approach transcends correlational analytics in big data by examining the theoretical explanations or uncovering the direction of causal effects.

My preferred research approach is to use a randomised field experimentation method with control-treatment groups of data samples. This can unequivocally attribute the outcomes of a phenomenon under study to the factor being manipulated in the "treatment" groups, such as consumers targeted specifically via mobile ads versus those in the "control groups", who are not given the promotion during the same time period.

While such approaches are considered the "gold standard" for scientific analysis in data analytics, they are hard to design and implement on a large scale in practice. Also, it has been cautioned previously that although such experiments work well for short-term business research questions, they may stifle creativity in solving long-term business problems.

In most practical situations, it may also not be possible to conduct such randomised experiments. As such, in many of my research projects involving observational data that have already been collected, I "simulate" the control-treatment group setting.

A causal approach to data analytics work is critically important as it creates confidence in the predictions generated. PHOTO: REUTERS

THE PSEUDO EXPERIMENTATION ANALYTICS METHOD

One way of doing so is via the propensity score matching method. Using this technique in social media marketing, I could quantify return-on-investment of Facebook marketing at an individual customer level, which is about S$25 per week of expenditure.

This was done by comparing the incremental expenditures of customers who have joined and interacted on the Facebook page of a retailer versus those who have not. I also ruled out other potential explanations, such as loyal customers who are inclined to spend more and social influences.

Another example is the use of instrumental variables to quantify the causal impact of a factor under study.

I used this method to examine the impact of newspaper reports on consumer sign-ups with the United States Do Not Call Registry. It is found that a 1 per cent increase in the number of news reports increased registrations by 0.018 per cent, and the impact on registration was higher for news reports that mentioned the number of other people registering.

A causal approach in data analytics work is critically important because it gives scientists, business practitioners and government administrators confidence in the predictions generated from changing different parameters. Needless to say, a deep understanding of the context under analysis is necessary. Without this understanding, even the most sophisticated data science technique can still fail badly.

● Dr Goh Khim Yong is an Associate Professor of Information Systems and Assistant Dean (Corporate Relations) in the School of Computing at The National University of Singapore.