Why mining Big Data is far from crude

Yang Qiang

LONG BEFORE THE US presidential election campaign was won in November Barack Obama's team was highly confident about the outcome – thanks to a technique called data mining.

Using the technique, an accurate picture of each voter was built, how he or she was likely to vote on election day, and which way the result would swing. A team of data scientists working on a wide variety of inputs and historical records was continually updating their model.

This was based on people's preferences and behavior and was gathered from thousands of sources, including past voting records, responses to campaign issues, thousands of phone and online interviews, and how voters changed their views as a result of campaigning.

Volunteers not only had a continuously evolving picture of voters in each state by the week, they also had an assessment on how they may respond to speeches, campaign themes, and pivotal issues that could change their views.

The team also had statistical models on how effective each volunteer was in approaching an undecided voter. For example, a volunteer from California was more effective in convincing voters of a certain issue than others.

The story is becoming the norm in a new economy and at the center of it all is data.

Or rather Big Data, a term that refers to just about every digitally recorded fact about things around us: society, gadgets, videos we watch, the torrents of data traffic that move our money transactions, web searches, app usage, online courses that we take, just to name a few.

An analogy with the oil industry and Big Oil is apt. All this data acts like the crude that needs to be found, drilled and extracted, and then processed in a refinery. Instead of machines that extract and process oil from the earth, you have the analytic technique known as data mining – a collection of powerful techniques based on statistics, machine learning and data management.

The machines are operated not by engineers but by data scientists, comprising people from a variety of disciplines including computer science, artificial intelligence researchers, statisticians, data storage experts, social scientists and so on.

The knowledge learned from the data is put into the hands of politicians, scientists, educators and business managers.

Today, data mining is already part of our everyday life. When we use Google, behind the search button is a powerful data-mining engine that predicts who you are, what you intend to do with the information, and how adverts should be displayed to attract your attention.

When we purchase a product using our credit card, a powerful data-mining engine is at work to find out whether an imposter who has stolen your card number is using your card; the data model in this case is built on billions of transactions that people have executed in the past.

When we go through the border at Lo Wu and press our finger on the fingerprint machine, a model constructed by a data-mining algorithm confirms at lightening speed that it is you who is standing in front of the machine.

With the new age of Big Data, data-mining research is at its infancy. Even so, Hong Kong academics and industry are already regarded as world leaders in the field.

In our universities, there are researchers working on various aspects of data mining – from algorithms that can make accurate predictions based on data, including data from the web, videos and spoken language, to studies on how to protect user privacy while data mining.

We also have several industry labs including the new Huawei Noah's Ark Lab, which is working on projects aimed at changing the data-mining landscape.

• *Yang Qiang heads Huawei Noah's Ark Lab and is a professor at the Hong Kong University of Science and Technology*