FROM THE EDITOR

Dear readers,

The "lead water incident" is the most serious health issue in Hong Kong. In this issue of science@HKU, we report the toxicology of lead. We also discuss the possible ways of heavy metal contamination of our water sources.

Yours sincerely,
Professor Hoi Fung CHAU
Chief Editor

FROM THE EDITOR

Statistics in the Era of Big Data

By Dr Philip L H Yu, Department of Statistics and Actuarial Science

Possibly you may not know that we are producing data every day, from birth to death. Starting from 2014, Google began collecting genetic and molecular information from thousands of healthy volunteers, in an attempt to allow medical experts to predict the onset of diseases such as heart disease and cancer. The company Evolv extracted millions of applicant data from job openings in hundreds of the world’s largest companies such as Xerox and Starwood Hotels & Resorts, with an aim to find suitable talents for such organizations. Although you may not want to volunteer in the Google project or to find another job, you possibly have installed Facebook, a popular social media app in your smart mobile, and the daily data on your personal connections and sharing have actually been uploaded to the Facebook server. This offers Facebook a wealth of data to identify the “influencers” in each social group, so that advertisers can maximize their market spending by targeting these “influencers”.

There is no doubt that the world is now experiencing a proliferation of big data whose volume, variety and velocity continue to reach unprecedented levels, capturing and analyzing data have tremendously influenced how decisions are made and how resources are allocated in various fields, as the accuracy of data processing can affect the results variably. However, the market is facing a shortage of talents who can employ techniques and theories drawn from mathematics, statistics and computer science to analyze big data, to formulate data-driven strategies and to communicate with stakeholders effectively. Someone call them "data scientists", the sexiest job of the 21st century called by Davenport and Patil (2012).

Statistics is the science of studying uncertainty and learning from data. Statistics is said to be the most mature of the data sciences as discussed in a whitepaper written by the Big Data R&D Initiative working group of the American Statistical Association (ASA) in July 2014. In the Royal Statistical Society’s President Address delivered by Professor Peter J Diggle in 2015 (Diggle 2015), he argued that data science is not just a new name for statistics. In fact the title of the ASA whitepaper "Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society” reveals that the data scientists should be able to integrate computer technologies and statistical techniques to help tackle the many big data challenges.

In this article, I will briefly introduce several recent research projects which demonstrate the essential contribution of statistics in big data research and applications.
Nowadays, various social media platforms such as Facebook, Twitter and Instagram are getting huge worldwide penetration. According to a study in 2013, people spent more than 27% of internet time on social networking (Experian 2013). The popular use of social media platforms attracts a lot of attention in many areas including product advertisement, sentiment analysis and election prediction.

The communications among users in the social media platforms can then form a network graph with interconnected nodes where each node represents a user and a line connecting two nodes may indicate an interaction, a friendship or a contact. Sometimes, an arrow will be drawn on the line to indicate the direction of contact or information flow.

Given the social communication activities made among the users, some users may be more influential in a social network in the sense that they tend to post more messages than the others. As these messages are mainly text documents, various text mining and computational linguistics tools such as statistical natural language processing, content categorization, etc., are essential methods of quantifying and analyzing textual data. See Yu and Tang (2015) for example.