Guest Editorial
Big Data Analytics: Risk and Operations Management for Industrial Applications

BIG DATA research has been a popular research topic over the last few years. It not only generates enormous attention compared with other research trends in the past, but also covers diverse and wide disciplines in its applications. This probably leads to the fact that the growth of this community is unparalleled, and attention being drawn to this research "buzzword" is growing at an explosive pace. It is simply not a business jargon! Above assertion can be supported by Table I, which summaries the number of publications in recent years. The data were obtained by searching the term "big data" via three common scholarly databases. The search is simple and no screening was conducted, but the numbers are very representative and impressive.

TABLE I
NUMBER OF PUBLICATIONS WITH THE SEARCH TERM "BIG DATA" (ACCESSED ON 3/3/2016)

Year of Publication	Google Scholar	IEEE Xplore	Science Direct
2010	1880	5	27
2011	3220	15	45
2012	9510	239	157
2013	23400	1338	622
2014	33900	2464	1529
2015	38800	3366	3606

One reason big data research is so popular is owing to the fact that the data are generated from real-life environment or applications. For example, many smartphone applications can help track countless transactions for making business-related decisions [6]. Internet of Things can help collect data on a real-time basis literally without any physical boundary [8]. In this connection, the usages or applications of big data seem unbounded as well. The challenge, however, is that the size of such data generated nowadays is huge, making it meaningless to talk about how much data are generated every day, as the figure keeps changing. If a number is quoted here, it will be different after you finish reading this editorial! In this connection, handling data is difficult, and how many applications or companies can truly utilize the value hidden in those data is worth studying (e.g., [11]).

Above backdrop leads to the development of this Special Section, which focuses on the exploitation of data collected via industrial sensor networks (be it wireless or not) [7], Internet-based applications [12], or so, for industrial applications. For the sake of this Special Section, the guest editors label such data "industrial big data." In analyzing these industrial big data, we rely on the advances in technology to collect data. Nevertheless, latest development of the technology is out of the scope of this Special Section. This Special Section focuses more on the handling and analysis of the collected data. It is expected that analyzing the data can improve the reliability of industrial systems by predicting the occurrence of potential risks and then rectification can be made accordingly. Such risks are inevitably linked to uncertain factors hidden in the systems that can be revealed by the data analysis process. Fig. 1 helps depict this.

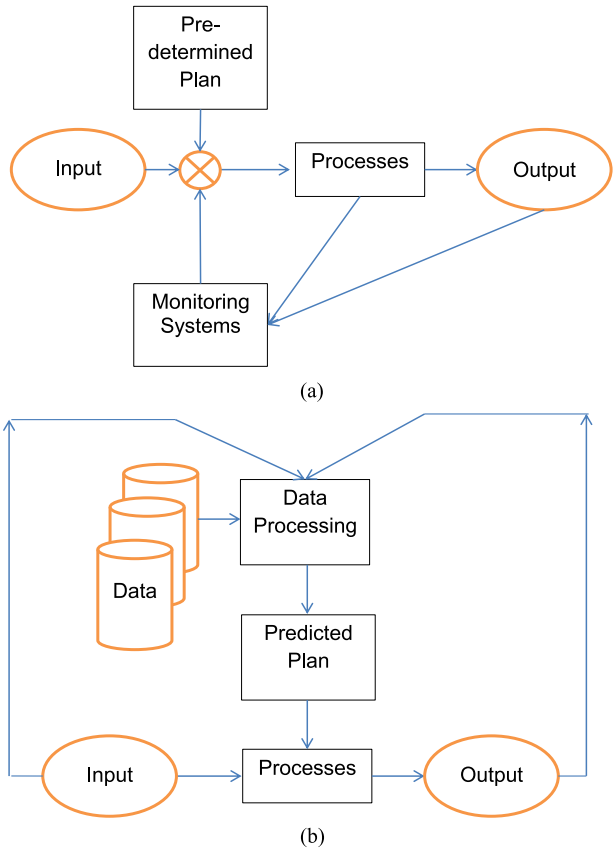

Fig. 1. Control systems. (a) Traditional industrial control systems. (b) Big data-based industrial control systems.

The major difference between Fig. 1(a) and (b) is the nature of data for the control mechanism. In Fig. 1(a), which illustrates traditional control systems, shows how the monitoring system records predefined and structured data. Then, structured decision can be made based on the designated control algorithm. In other words, the operations of the whole industrial system follow certain logic based on the sensory network in the system. Fig. 1(b) depicts how big data have changed the system. The

Digital Object Identifier 10.1109/TII.2016.2558477

predicted plan is not prepared based only on, for example, forecasting techniques but the big data. The data, which include both big data and process-related input and output data, are handled by the processing unit which makes prediction and then the operations of the whole system will follow the prediction. The major barrier is that the input data would be unstructured due to the nature of the sensory network; otherwise, the system could be classified as the traditional system in Fig. 1(a) with more inputs very easily. In this connection, the data processing unit (DPU) is the core engine of this big-data-based industrial system. In fact, this is the same as any other type of big-data-based systems, and the only difference is the final applications generated from the data. In summary, the traditional system in Fig. 1(a) is more reactive, whereas the big-data-based system in Fig. 1(b) is more predictive, if not proactive, in nature. This is also the reason why such systems sometimes are referred to as data-driven systems because a pure theoretical framework is not sufficient to explain the behavior of the systems.

There are many ways to design the DPU, and this is exactly the focus of this Special Section. Due to the unstructured nature of the data, intelligence is almost always a necessary element in this DPU. Otherwise, the DPU would not be able to react to the input data. Therefore, learning algorithms are still very big concerns in relevant research studies. Such learning algorithms analyze the data and then make intelligent decisions so that uncertainty and hence risk of the system can be kept at a minimal.

In general, the DPU aims to extract the relationships among data, which are normally grouped into clusters based on the similarity of the data. Then, using these clusters, the relationship between the inputs and the output is identified which helps yield correct actions or decisions. The input data sometimes are well structured already due to the system designs, but the interrelationships among factors or clusters are unknown, or keep changing. The rationale behind the grouping or clustering is straightforward: To breakdown the big dataset into smaller, manageable clusters (i.e., factors on some occasions) in order to simplify the problem. Computational power is also a contributing factor that requires such scaling down of the high volume of data.

After forming such clusters, regression, particularly linear regression or logistic regression, is the most basic technique to identify the relationships between the inputs and the output in order to investigate the relationships of the clusters. The shortcoming of this approach is that, no intelligence is included, and the only "reactive" element could be a dynamic algorithm to conduct regression iteratively. In addition, nonlinear regression would be difficult, if not impossible, to design. Obviously, time-series analysis can be added on top of such regression so that longitudinal data can be considered to improve the reliability and robustness of the algorithms.

Learning algorithms are very common in related studies to resolve the nonlinearity issue of regression. There are broadly two types of such learning algorithms: 1) supervised learning; and 2) unsupervised learning. Both types aim to achieve the same objective, which is to make use of training data (normally historical data) to determine the equations between the input and output variables. Supervised learning can handle the relatively structured pair of inputs and outputs. In other words, it is relatively easy to compare the actual outputs and the trained outputs. To be precise, the differences can be treated as the errors of the training algorithms. Therefore, most industrial systems can be configured under this category. Some examples of supervised learning algorithms are support vector machines (SVM) [9], which is designed for classification and regression analysis. Furthermore, nonlinear regression is possible with SVM. For instance, Garcia Nieto *et al*. [10] applied SVM in a nonlinear model to evaluate the remaining life and reliability of aircraft engines.

However, when big data are taken into account, unsupervised learning may be required due to their variety and unstructured nature. This learning approach reveals the hidden relationships from the input variables given the corresponding outputs. In this connection, clustering is normally a prerequisite in the data analysis process. *k*-means clustering is a commonly employed clustering method that is a centroid-based approach to allocate different inputs to *k* clusters so that the centroids among the clusters are balanced. For example, Bishnu and Bhattacherjee [4] applied the *k*-means clustering approach for faults prediction in software modules. Other clustering methods are more or less similar with an aim to balance the distribution of the clusters based on some parameters (e.g., distance, density, and so on). Readers are referred to a recent survey conducted by Fahad *et al*. [5] on clustering methods for big data analysis. Some approaches incorporate the clustering algorithms into the learning process, such as artificial neural network (ANN) [1]. The most famous applications of ANN are computer vision and pattern recognition [2], [3].

Despite the rich discussion on big data research on many applications, published studies are still very loosely coupled and real application of big data research is still at the exploratory stage. The former is probably due to the scope of big data research, which is diverse. The latter is probably due to the large scale of data involved and implementation of relevant systems takes time, let alone the complexity to classify and analyze the data. In this connection, this Special Section collects most recent research on a number of areas on risk and operations management for industrial applications based on big data.

With the analysis of a huge volume of transaction data, the marketplaces may exhibit many new properties under multiple risks. In the paper "European option pricing with a fast Fourier transform algorithm for big data analysis," Xiao *et al*. [13] applied the fast Fourier transform (FFT) algorithm to capture these newly developed features using the European options approach. The algorithm provides a quick solution in relation to high-speed data generation. The authors first used statistical analysis and stochastic differential equations to verify and model the new properties. They further derived the closed-form solution with the characteristic function method, and provided a FFT-based numerical algorithm. The study contributes to this Special Section by building an innovative stochastic model to depict newly emerged risk factors in the market and provide an accurate pricing tool. Furthermore, an efficient numerical algorithm is proposed to better address the corresponding big data.

In various industrial fields, big data need to be gathered on a real-time basis for analyzing the risks of industrial operations. Nowadays, indoor wireless sensor network (WSN) has become the key technology to gather the real-time big data in a complex indoor industrial environment. However, a major challenge for indoor WSN is to ensure that big data can be transmitted to the data center when the industrial environment is changing. To solve this problem in their paper "A real-time big data gathering algorithm based on indoor WSNs for risk analysis of industrial operations," Ding *et al*. [14] developed an energy-balanced real-time big data gathering (RTBDG) algorithm based on indoor WSN. The received signal strength indicator values are used to determine whether two sensors nodes can communicate with each other instead of the distance between two sensor nodes. The RTBDG algorithm has an adaptive capacity in a complex indoor environment, and it not only efficiently uses the limited energy of network nodes but also balances the energy consumption of all nodes.

In the context of big data analytics, soft sensors are arising as an important technology for estimating quantities that are costly or impossible to measure. Soft sensors are generally based on machine learning techniques that are becoming more and more promising owing to their versatility to treat multiple data sources and formats. In order to exploit time-series data for predictive modeling, it is necessary to summarize the information they contain as a set of features (to use as model regressors). Typically, this is done in an unsupervised fashion using simple techniques, such as computing statistical moments, principal components, or wavelet decompositions, often leading to significant information loss, and hence suboptimal predictive models. In the paper "Supervised aggregative feature extraction for big data time series regression," Susto *et al*. [15] introduced a methodology called SAFE (supervise aggregative feature extraction), which exploits a functional learning paradigm in a supervised fashion to derive continuous smooth estimates of time-series data, while simultaneously estimating a continuous shape function yielding optimal predictions. Using simulation studies and a practical semiconductor manufacturing case study, the strengths of the SAFE are shown with respect to standard feature extraction approaches.

Failure mode and effect analysis (FMEA) is a well-established method in risk management. In the paper "Data-driven system reliability and failure behavior modelling using FMECA," Khorshidi *et al*. [16] examined a data-driven system to evaluate reliability of industrial systems using FMEA. An algorithm is proposed using soft computing techniques for risk management of complex systems based on qualitative data. In their study, a hierarchical structure from failure mode level to system level is introduced. Based on this structure and qualitative data of FMEA's parameters, an overall failure index (OFI) is proposed to aggregate the failure measures through the whole system. In addition, OFI can be used to prioritize the improvement actions. Subsequently, two optimization models, a linear and a nonlinear, are developed to find the optimal actions subject to budget constraint. The algorithm was verified with a case study.

In the paper "Wind turbine modeling with data-driven methods and radially uniform designs," Tan and Zhang [17] proposed a radially uniform design to rapidly sample informative data points from a big dataset for facilitating data-driven analyses. A large volume of industrial wind turbine data is applied to evaluate the effectiveness of the proposed radially uniform design. Data mining algorithms including ANN, multivariate adaptive regression splines, SVM, k nearest neighbors, and linear regression are utilized to develop data-driven models based on data points sampled with the radially uniform sampler and the complete dataset. The comparative analysis reports that the modeling results based on sampled data points and a complete dataset are close. The radially uniform sampler is also compared with the random sampler and maximin sampler. Extensive computational experiments are performed to prove the advantage of the radially uniform algorithm in sampling informative data points. Computational results also show that the radially uniform sampler is more effective in sampling data points for building nonlinear data-driven models.

In the paper "A big data clustering algorithm for mitigating the risk of customer churn," Bi *et al*. [18] proposed a new big data clustering algorithm called the semantic-driven subtractive clustering method (SDSCM), and successfully implement it through a Hadoop MapReduce framework in the industrial application of China Telecom. Results show that SDSCM has a stronger semantic strength and the parallel one enjoys a fast running speed when dealing with telco big data. The SDSCM improves the accuracy of the subtractive clustering method and decreases the risk of imprecise operations management by using axiomatic fuzzy sets. Additionally, the industrial application of parallel SDSCM to China Telecom offers novel insights for managers to raise the level of customer churn risk management in the big data context.

In the paper "Dynamic pricing and risk analytics under competition and stochastic reference price effects," Wu and Wu [19] employ a mobile phone case to determine the optimal pricing employed in a competitive market. To deal with the uncertainty, stochastic modeling is applied to take dynamic customer preference and reference prices into consideration. Their study illustrates that such industrial big data can help overcome the uncertain factors in making decisions.

The Guest Editors are glad that they received overwhelming responses to the call for papers, and they hope that you enjoy reading this Special Section as much as they did editing it.

ACKNOWLEDGMENT

The Guest Editors would like to express their deep gratitude to all the authors who have submitted their valuable contributions, and to the numerous and highly qualified anonymous reviewers. They would like to thank Prof. K. Man, Editor-in-Chief of the IEEE TRANSACTIONS IN INDUSTRIAL INFORMATICS (TII), for giving them the opportunity to organize this Special Section and for all the encouragement, help, and support given throughout the process, and L. Patillo, TII staff, for her professional support and assistance during the whole preparation of this Special Section.

HING KAI CHAN, *Guest Editor*
Nottingham University Business School China
University of Nottingham–Ningbo
Ningbo, China
e-mail: hingkai.chan@nottingham.edu.cn
TSAN-MING CHOI, *Guest Editor*
Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
e-mail: jason.choi@polyu.edu.hk

XIAOHANG YUE, *Guest Editor*
University of Wisconsin–Milwaukee
Milwaukee, WI USA
e-mail: xyue@uwm.edu

REFERENCES

[1] B. C. Pijanowski, A. Tayyebi, J. Doucette, B. K. Pekin, D. Braun, and J. Plourde, "A big data urban growth simulation at a national scale: Configuring the GIS and neural network based land transformation model to run in a high performance computing (HPC) environment," *Environ. Modell. Softw.*, vol. 51, pp. 250–268, 2014.
[2] J. M. Alonso-Weber, M. P. Sesmero, and A. Sanchis, "Combining additive input noise annealing and pattern transformations for improved handwritten character recognition," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8180–8188, 2014.
[3] D. C. Cireşan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Deep, big, simple neural nets for handwritten digit recognition," *Neural Comput.*, vol. 22, no. 12, pp. 3207–3220, 2010.
[4] P. S. Bishnu and V. Bhattacherjee, "Software fault prediction using quad tree-based k-means clustering algorithm," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1146–1150, Jun. 2012.
[5] A. Fahad *et al.*, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014.
[6] H. K. Chan, X. Wang, E. Lacka, and M. Zhang, "A mixed-method approach to extracting the value of social media data," *Prod. Oper. Manage.*, vol. 25, no. 3, pp. 568–583, 2016.
[7] C. Alippi, M. Roveri, and F. Trovó, "A self-building and cluster-based cognitive fault diagnosis system for sensor networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 6, pp. 1021–1032, Jun. 2014.
[8] M. Díaz, C. Martín, and B. Rubio, "State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing," *J. Netw. Comput. Appl.*, vol. 16, pp. 99–117, 2016, doi: 10.1016/j.jnca.2016.01.010.
[9] P. Li, L. Dong, H. Xiao, and M. Xu, "A cloud image detection method based on SVM vector machine," *Neurocomputing*, vol. 169, no. 2, pp. 34–42, 2015.
[10] P. J. García Nieto, E. García-Gonzalo, F. Sánchez Lasheras, and F. J. de Cos Juez, "Hybrid PSO–SVM-based method for forecasting of the remaining useful life for aircraft engines and evaluation of its reliability," *Rel. Eng. Syst. Safety*, vol. 138, pp. 219–231, 2015.
[11] K. Zhou, C. Fu, and S. Yang, "Big data driven smart energy management: From big data to big insights," *Renew. Sustain. Energy Rev.*, vol. 56, pp. 215–225, 2016.
[12] A. Bousdekis, B. Magoutas , D. Apostolou, and G. Mentzas, "A proactive decision making framework for condition-based maintenance," *Ind. Manage. Data Syst.*, vol. 115, no. 7, pp. 1225–1250, 2015.
[13] S. Xiao, S.-H. Ma, G. Li, and S. K. Mukhopadhyay, "European option pricing with a fast Fourier transform algorithm for big data analysis," *IEEE Trans. Ind. Informat.*, to be published
[14] X. Ding, Y. Tian, and Y. Yu, "A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations," *IEEE Trans. Ind. Informat.*, to be published.
[15] G. A. Susto, A. Schirru, S. Pampuri, and S. McLoone, "Supervised aggregative feature extraction for big data time series regression," *IEEE Trans. Ind. Informat.*, to be published.
[16] H. A. Khorshidi, I. Gunawan, and M. Y. Ibrahim, "Data-driven system reliability and failure behavior modelling using FMECA," *IEEE Trans. Ind. Informat.*, to be published.
[17] M. Tan and Z. Zhang, "Wind turbine modeling with data-driven methods and radially uniform designs," *IEEE Trans. Ind. Informat.*, to be published.
[18] W. Bi, M. Cai, M. Liu, and G. Li, "A big data clustering algorithm for mitigating the risk of customer churn," *IEEE Trans. Ind. Informat.*, to be published.
[19] L. -L. B. Wu and D. Wu, "Dynamic pricing and risk analytics under competition and stochastic reference price effects," *IEEE Trans. Ind. Informat.*, to be published.

Hing Kai Chan (M'98–SM'04) received the B.Eng. degree (Hons.) in electrical and electronic engineering, the M.Sc. (Eng.) degree in industrial engineering and industrial management, and the Ph.D. degree with a focus on supply chain management from the University of Hong Kong, Hong Kong, in 1995, 2000, and 2007, respectively, and the B.Sc. degree (Hons.) in economics and management from the London School of Economics and Political Science in 2006.

He is an Associate Professor of Operations Management with Nottingham University Business School China, University of Nottingham Ningbo, Ningbo, China. He has published more than 100 academic articles and (co-)edited several special issues for reputable international journals. His publications appear in various IEEE TRANSACTIONS, *Production and Operations Management, Decision Support Systems*, the *European Journal of Operational Research*, the *International Journal of Production Economics*, among others.

Prof. Chan serves as an Editorial Board Member (or similar) in a number of journals, such as *Transportation Research Part E: Logistics and Transportation Review,* and *Online Information Review*. He is a fellow of the Institution of Engineering and Technology (FIET) and the Higher Education Academy (FHEA). He is also a senior member of the Institute of Electrical and Electronics Engineers, and a member of the Chartered Institute of Marketing and the Chartered Institute of Logistics and Transport. He is a Chartered Engineer and a Chartered Marketer.

Tsan-Ming Choi (S'00–M'01) received the B.Eng. (Hons.), M.Phil., and Ph.D. degrees from the Department of Systems Engineering and Engineering Management, Faculty of Engineering, Chinese University of Hong Kong, Hong Kong, in 1997, 1999, and 2002, respectively.

He is currently a Professor of Fashion Business with the Hong Kong Polytechnic University, Hung Hom, Hong Kong. Before joining his current department in 2004, he was an Assistant Professor at the Chinese University of Hong Kong. He has published extensively in leading journals such as the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, the IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, *Automatica, Production and Operations Management, Naval Research Logistics, INFORMS Service Science*, and various other IEEE TRANSACTIONS.

Prof. Choi is currently a Senior Editor of *Production and Operations Management*, and *Decision Support Systems*, and an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—SYSTEMS AND INFORMATION SCIENCES. He received the President's Award for Excellent Achievements from the Hong Kong Polytechnic University in 2008. He was honored by being selected as the recipient of the Best Associate Editor Award from the IEEE Systems, Man, and Cybernetics Society in 2013 and 2014. Most recently, he and his collaborators received the Second Prize of the Natural Science Award of the Scientific Research Excellence (Science and Technology) in Colleges and Universities, by the Ministry of Education of China in 2016.

Xiaohang Yue received the Ph.D. degree in operations and supply chain management from the University of Texas at Dallas, Richardson, TX, USA, in 2002.

He is currently an Associate Professor with the School of Business, University of Wisconsin–Milwaukee, Milwaukee, WI, USA. He mainly engages in research on supply chain and logistics systems management, and industrial manufacturing systems management. He has published extensively in leading journals such as the IEEE TRANSACTIONS, *Decision Sciences, IIE Transactions, Naval Research Logistics, Production and Operations Management*, the *European Journal of Operational Research, Omega*, etc.

Dr. Yue has served as an Editorial Board Member of *Production and Operations Management*. He is a member of the American Association of Supply Chain Management and INFORMS.