

BIG DATA

Challenges and Future Research Directions

MARCO S. REIS

UNIV. OF COIMBRA, PORTUGAL

RICHARD D. BRAATZ

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEO H. CHIANG

THE DOW CHEMICAL CO.

The big data movement is creating opportunities for the chemical process industries to improve their operations. Challenges, however, lie ahead.

The big data movement is gaining momentum, with companies increasingly receptive to engaging in big data projects. Their expectations are that, with massive data and distributed computing, they will be able to answer all of their questions — from questions related to plant operations to those on market demand. With answers in hand, companies hope to pave new and innovative paths toward process improvements and economic growth.

An article in *Wired* magazine, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” (1), describes a new era in which abundant data and mathematics will replace theory. Massive data is making the hypothesize-model-test approach to science obsolete, the article states. In the past, scientists had to rely on sample testing and statistical analysis to understand a process. Today, computer scientists have access to the entire population and therefore do not need statistical tools or theoretical models. Why is theory needed if the entire “real thing” is now within reach?

Although big data is at the center of many success stories, unexpected failures can occur when a blind trust is placed in the sheer amount of data available — highlighting the importance of theory and fundamental understanding.

A classic example of such failures is actually quite dated. In 1936, renowned magazine *Literary Digest* conducted an extensive survey before the presidential election between Franklin D. Roosevelt and Alfred Landon, who was then

governor of Kansas. The magazine sent out 10 million postcards — considered a massive amount of data at that time — to gain insight into the voting tendencies of the populace. The *Digest* collected data from 2.4 million voters, and after triple-checking and verifying the data, forecast a Landon victory over Roosevelt by a margin of 57% to 43%. The final result, however, was a landslide victory by Roosevelt of 61% versus Landon’s 37% (the remaining votes were for a third candidate). Based on a much smaller sample of approximately 3,000 interviews, George Gallup correctly predicted a clear victory for Roosevelt.

Literary Digest learned the hard way that, when it comes to data, size is not the only thing that matters. Statistical theory shows that sample size affects sample error, and the error was indeed much lower in the *Digest* poll. But sample bias must also be considered — and this is especially critical in election polls. (The *Digest* sample was taken from lists of automobile registrations and telephone directories, creating a strong selection bias toward middle- and upper-class voters.)

Another example that demonstrates the danger of putting excessive confidence in the analysis of big data sets regards the mathematical models for predicting loan defaults developed by Lehman Brothers. Based on a very large database of historical data on past defaults, Lehman Brothers developed, and tested for several years, models for forecasting the probability of companies defaulting on

their loans. Yet those models built over such an extensive database were not able to predict the largest bankruptcy in history — Lehman Brothers' own.

These cases illustrate two common flaws that undermine big data analysis:

- the sample, no matter how big, may not accurately reflect the actual target population or process
- the population/process evolves in time (*i.e.*, it is nonstationary) and data collected over the years may not accurately reflect the current situation to which analytics are applied.

These two cases and other well-known blunders show that domain knowledge is, of course, needed to handle real problems even when massive data are available. Industrial big data can benefit from past experiences, but challenges lie ahead.

Like any new, promising field, big data must be viewed in terms of its capabilities as well as its limitations. Some of these limitations are merely challenges that can be addressed — enabling companies to make the most out of new opportunities created by data, technology, and analytics (Figure 1).

This article outlines ten critical challenges regarding big data in industrial contexts that need to be addressed, and suggests some emerging research paths related to them. The challenges are discussed in terms of the four Vs that define the context of big data: volume, variety, veracity, and velocity.

Volume challenges



Big data is, first of all, about handling massive amounts of data. However, in industrial processes, the first thing to realize is that not all data are created equal. Several

challenges arise from this point.

Meaningful data. Most industrial big data projects rely on happenstance data, *i.e.*, data passively collected from processes operating under normal operating conditions most of the time. Thus, a large amount of data is indeed available, but those data span a relatively narrow range of operating conditions encountered during regular production situations.

Data sets collected under those circumstances may be suitable for process monitoring and fault detection activities (2), which rely on a good description of the normal operating conditions (NOC) as a reference to detect any assignable or significant deviation from such behavior. However, their value is limited for predictive activities, and even more so for control and optimization tasks. Prediction can only be carried out under the same conditions found in the data used to construct the models. As a corollary, only when all the NOC correlations linking the input variables are respected can the model be used for prediction.

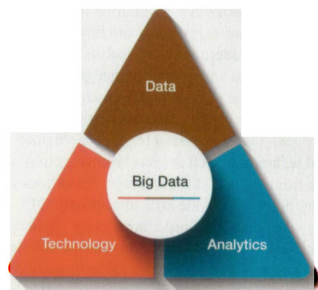
For process control and optimization activities, the

process description must capture the actual influence of each manipulated input variable on the process outputs. Its construction requires experimentation — *i.e.*, the active collection of process data via a design of experiments (DOE) program for process optimization or via system identification (SI) experiments for process control.

Future research is needed to determine ways to use DOE in the context of big data to complement the information already available and increase the data's value for predictive, control, and optimization activities. This will likely require methods to selectively remove data with very little informative value. The presence of such data is not only unnecessary for developing models, but also detrimental, as their presence induces a bias in the models toward highly sampled regions of the operational space. The modern theory of optimal DOE may provide a suitable framework to begin addressing this challenge.

Information-poor data sets. Despite the sheer volume of industrial data, the relevant or interesting information may happen on only a few, dispersed occasions. Examples include batches with abnormally excellent quality or runs that experience several types of process upsets.

Current data mining and knowledge discovery tools (3, 4) can handle very large volumes of data that are rich in information. Such tools include methodologies such as partial least-squares regression, least-absolute-shrinkage and selection operator (LASSO) regression, and ensemble methods (*e.g.*, random forests and gradient boosting), among others. However, by design, those methods are not suited to analyze information-poor data sets, in which the interesting information is rare and scattered. And, traditional data visualization tools — which are recommended for any data analysis activity, especially to identify potentially interesting outlying data points — may not always be



▲ Figure 1. The big data movement stems from the availability of data, high-power computer technology, and analytics to handle data characterized by the four Vs — volume, variety, veracity, and velocity.

effective when applied to big data. For example, creating a classical plot from big data might produce what looks like a black cloud of data points that is not useful.

An engineer who is not able to rely on visualization might be tempted to perform some sort of massive statistical testing to pinpoint abnormal situations or to extract potentially interesting correlations, only to find a very large number of such situations (or correlations). That is a consequence of the extreme power of the tests, induced by the massive number of observations used. The significant events detected may not (and most of the time will not) have any practical relevance because of their small impact.

The situation can be even worse when an engineer cleans the data using an algorithm that automatically removes outlying observations from data sets prior to analysis. Such algorithms often incorporate standard rules of an empirical nature that eliminate the data embedded with the rare gems of information.

Future research should focus on the development of analytical methods applicable to information-poor data, including visualization tools that can condense large amounts of data while being sensitive to abnormal observations, and sound ways of detecting outlying (but interesting) observations (and variable associations), namely by incorporating the available domain knowledge.

Variety challenges



Big data is also characterized by its complexity. The complexity of industrial data can arise from different sources, and is usually related to the variety of objects to be analyzed. Different challenges arise depending on the origin of the complexity.

Multiple data structures. In addition to the usual scalar quantities (temperature, pressure, and flow measurements), data collected in modern industrial settings also include other data structures arranged as higher-order tensors, such as one-way arrays (e.g., spectra, chromatograms, nuclear magnetic resonance [NMR] spectra, particle-size distribution curves), two-way arrays (e.g., data obtained from analytical techniques such as gas chromatography with mass spectrometry [GC-MS] and high-performance liquid chromatography with diode array detection [HPLC-DAD]), and three-way and higher-order arrays (e.g., hyperspectral images, color videos, hyphenated instruments). These data structures are examples of profiles (5), abstractly defined as any data array, indexed by time and/or space, that characterize a product or process.

Future research should focus on developing analytical platforms that can effectively incorporate and fuse all

of these heterogeneous sources of information found in industrial processes, for instance, through the development of more flexible multiblock methodologies. Such methodologies incorporate the natural block-wise structure of data, where each block may carry information about distinct aspects of the problem and present a characteristic structure and dimensionality.

Heterogeneous data. Variety does not originate only from the presence of different data structures to be handled simultaneously. Another source of variety is the presence of data in the same data set that were collected when the process underwent meaningful changes, including in its structure (e.g., new equipment was added, procedures were changed). By not taking such changes into account during the analysis of the entire data set, you may fall into the trap of mixing apples with oranges — an issue that also raises concerns of data quality, which is discussed in the veracity section of this article. Overlooking heterogeneity in time is detrimental for analytical tasks such as process monitoring and quality prediction.


A future research path to address this challenge is developing methods to detect and handle these issues, as well as to deal with the time-varying nature of processes, namely through evolutionary and adaptive schemes (6). Such schemes can adapt to complex and/or changing conditions by continuously seeking the optimal operational settings or by periodically retuning the models (through re-estimation or recursive updating approaches).

Multiple data-management systems. Data are also collected from a variety of sources across the company's value chain, from raw materials, plant operations, and quality laboratories, to the commercial marketplace. Each stage usually has its own data-management system, and each records data in a different way.

Future efforts should be directed toward the development of integrated platforms that link all of the different sources of data in the value chain. Market data, in particular, have not been included in conventional models used in the chemical process industries (CPI). Data-driven methods — which incorporate the time-delayed structure of the processes and use different types of data aggregation — should be developed to make this integration effective.

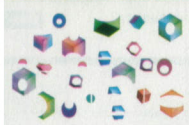
A priori knowledge. Some knowledge about the main sources of variety affecting a massive data set is usually available. However, making use of it in conventional industrial analytics is not straightforward. Big data methods tend to be of a black box type, lacking the flexibility to incorporate *a priori* knowledge about the processes under analysis.

Incorporating information about the structure of the processes in data-driven analysis is an important research path for the future, especially in the fields of fault diagnosis and predictive modeling (7–9). Fault diagnosis requires informa-



tion about the causal structure of the systems, which conventional data-driven monitoring methods cannot provide. Predictive modeling also requires this type of knowledge, in particular for process control and optimization applications. Bayesian approaches (10, 11) and data transformation based on network inference, together with hybrid gray-box modeling frameworks, are potential ways to introduce *a priori* knowledge into data-driven modeling.

Veracity challenges



A major concern in the analysis of massive data sets has to do with the quality of data, *i.e.*, their veracity. As previously mentioned, quantity does not imply quality. On the contrary,

quantity creates more opportunities for problems to occur. To make matters worse, the detection of bad observations in massive data sets through visualization techniques is more challenging and automatic-cleaning algorithms cannot be relied on either. Data quality also depends on the way the data are collected (bias issues may emerge that are very difficult to detect), on whether the information is updated or no longer makes sense (due to time-varying changes in the system), and on the signal-to-noise ratio (measurement uncertainty), among other factors.

Uncertainty data. In addition to the collected data, information associated with uncertainty is also available. Measurement uncertainty is defined as a parameter associated with the result of a measurement that characterizes the dispersion of the values that could reasonably be attributed to the quantity to be measured (12). Combining uncertainty data with the raw measurements can improve data analysis, empirical modeling, and subsequent decision-making (13, 14).

Specification of measurement uncertainty in big data contexts and developing methods that take advantage of knowledge about uncertainty should be explored in more depth.

Unstructured variability. Process improvement activities require a careful assessment of the multiple sources of variability of the process, which are typically modeled using suitable mathematical equations (ranging from first-principles models to purely data-driven approaches). The analysis should involve both the deterministic backbone of the process behavior, as well as the unstructured aspects of the process arising from stochastic sources of variability, including disturbances, sample randomness, measurement noise, operators' variation, and machine drifting. Jumping into the analysis of massive data sets while overlooking the main sources of unstructured variability is ill-advised, and is contrary to a reliable statistical engineering approach to addressing process improvement activities.

The sources of variability are actually the core of many improvement activities, in particular those aimed at reducing process variation and increasing product quality and consistency. Big data cannot replace the need to understand how data are acquired and the underlying mechanisms that generate variability, and statistical engineering principles should be brought to the analysis of big data sets in the future (15).

Velocity challenges



In big data scenarios, large quantities of data are collected at high speed. This creates several challenges in the implementation of online collection techniques and in defining the

appropriate granularity to adopt for data analysis.

Data with a high time resolution. The high speed at which data are collected in modern chemical plants produces information with very fine time granularity, *i.e.*, the data have, by default, a high time resolution (on the order of minutes, or even seconds). This default is a conditioning factor for all the subsequent stages of data analysis, as the usual practice is to avoid throwing out potentially valuable data. Consequently, the analysis is prone to producing over-parameterized models.

It is important to select the most effective resolution (16) for your particular data analysis. A default resolution selected by a third party with no knowledge of your specific data will probably not be appropriate.

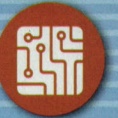
Future research should develop sound ways for selecting the proper resolution, including the possibility of using multiple time resolutions (17) that take into account the variables' dynamic and noise features.

Adaptive fault detection and diagnosis. The high speed of data collection provides the potential for fast detection and diagnosis of faults, failures, and other abnormal conditions. Many effective methods for fault detection and identification of associated variables are available, including techniques that account for dynamics (18–21).

A limitation of the standard data-based fault diagnosis methods is that they rely on historical data that were collected, analyzed, and labeled during past abnormal conditions (22, 23). One way around this requirement is to incorporate causal information from the process flowsheet (24).

Drawing on ideas from the machine learning community (25), a more effective solution could be to treat fault diagnosis as an online learning problem. Adaptive learning methods could generate fault diagnosis systems that become increasingly effective over time, with the objective of moving toward prognostics (*i.e.*, the early prediction of future operational problems) instead of learning about abnormal conditions after a catastrophic incident.

Article continues on next page



Final thoughts

Big data creates new possibilities to drive operational and business performance to higher levels. However, gaining access to such potential is far from trivial. New strategies, processes, mindsets, and skills that are not yet in place are necessary. In addition, challenges emerge when big data problems are considered in industrial contexts.

This article has summarized ten such challenges to be addressed in the future — to make this journey an insightful learning experience and a successful business opportunity for companies. We also believe the dominating ideas and premises of big data need to evolve and mature.

As we have discussed, big data by itself will not answer all of your questions. Processes evolve over time, under quite restrictive operating conditions, and data just reflect this reality. We cannot expect data to tell us more than the information contained in the data. But big data and domain knowledge can be used synergistically to move forward and answer important questions, to design better experiments, or to determine additional sensors needed to address those questions.

Big data offers new opportunities for managing our operations, improving processes at all levels, and even adapting the companies' business models. So the important question is: Can we afford not to enter the big data era? **CEP**

LITERATURE CITED

1. Anderson, C., "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, www.wired.com/2008/06/pb-theory/ (June 23, 2008).
2. Chiang, L. H., et al., "Fault Detection and Diagnosis in Industrial Systems," Springer-Verlag London (2001).
3. Han, J., and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, San Francisco, CA (2001).
4. Wang, X. Z., "Data Mining and Knowledge Discovery for Process Monitoring and Control," Springer-Verlag London (1999).
5. Reis, M. S., and P. M. Saraiva, "Prediction of Profiles in the Process Industries," *Industrial and Engineering Chemistry Research*, 51 (11), pp. 4524–4266 (Feb. 27, 2012).
6. Rato, T. J., et al., "A Systematic Comparison of PCA-Based Statistical Process Monitoring Methods for High-Dimensional, Time-Dependent Processes," *AIChE Journal*, 62 (1), pp. 127–142 (Jan. 2016).
7. Reis, M. S., et al., "Challenges in the Specification and Integration of Measurement Uncertainty in the Development of Data-Driven Models for the Chemical Processing Industry," *Industrial and Engineering Chemistry Research*, 54 (37), pp. 9159–9177 (Aug. 31, 2015).
8. Reis, M. S., and P. M. Saraiva, "Integration of Data Uncertainty in Linear Regression and Process Optimization," *AIChE Journal*, 51 (11), pp. 3007–3019 (Nov. 2005).
9. Chiang, L. H., and R. D. Braatz, "Process Monitoring Using the Causal Map and Multivariate Statistics: Fault Detection and Identification," *Chemometrics and Intelligent Laboratory Systems*, 65 (2), pp. 159–178 (Feb. 28, 2003).
10. Bakshi, B. R., et al., "Multiscale Bayesian Rectification of Data from Linear Steady-State and Dynamic Systems without Accurate Models," *Industrial and Engineering Chemistry Research*, 40 (1), pp. 261–274 (Dec. 6, 2000).
11. Yu, J., and M. M. Rashid, "A Novel Dynamic Bayesian Network-Based Networked Process Monitoring Approach for Fault Detection, Propagation, Identification, and Root Cause Diagnosis," *AIChE Journal*, 59 (7), pp. 2348–2365 (July 2013).
12. Joint Committee for Guides in Metrology, "Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement," JCGM 100:2008, JCGM, Paris, France, p. 134 (Sept. 2008).
13. Reis, M. S., et al., "Challenges in the Specification and Integration of Measurement Uncertainty in the Development of Data-Driven Models for the Chemical Processing Industry," *Industrial and Engineering Chemistry Research*, 54 (37), pp. 9159–9177 (Aug. 31, 2015).
14. Reis, M. S., and P. M. Saraiva, "Integration of Data Uncertainty in Linear Regression and Process Optimization," *AIChE Journal*, 51 (11), pp. 3007–3019 (Nov. 2005).
15. Hoerl, R., and R. D. Snee, "Statistical Thinking: Improving Business Performance," Duxbury Press, Pacific Grove, CA (2001).
16. Reis, M. S., and P. M. Saraiva, "Generalized Multiresolution Decomposition Frameworks for the Analysis of Industrial Data with Uncertainty and Missing Values," *Industrial and Engineering Chemistry Research*, 45 (18), pp. 6330–6338 (Aug. 9, 2006).
17. Reis, M. S., and P. M. Saraiva, "Multiscale Statistical Process Control with Multiresolution Data," *AIChE Journal*, 52 (6), pp. 2107–2119 (June 2006).
18. Russell, E. L., et al., "Fault Detection in Industrial Processes Using Canonical Variate Analysis and Dynamic Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, 51, pp. 81–93 (2000).
19. Zhu, X., and R. D. Braatz, "Two-Dimensional Contribution Map for Fault Detection," *IEEE Control Systems*, 34 (5), pp. 72–77 (Oct. 2014).
20. Jiang, B., et al., "Canonical Variate Analysis-Based Contributions for Fault Identification," *Journal of Process Control*, 26, pp. 17–25 (Feb. 2015).
21. Jiang, B., et al., "Canonical Variate Analysis-Based Monitoring of Process Correlation Structure Using Causal Feature Representation," *Journal of Process Control*, 32, pp. 109–116 (Aug. 2015).
22. Chiang, L. H., et al., "Fault Diagnosis in Chemical Processes Using Fisher Discriminant Analysis, Discriminant Partial Least Squares and Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, 50, pp. 240–252 (2000).
23. Jiang, B., et al., "A Combined Canonical Variate Analysis and Fisher Discriminant Analysis (CVA-FDA) Approach for Fault Diagnosis," *Computers and Chemical Engineering*, 77, pp. 1–9 (June 9, 2015).
24. Chiang, L. H., et al., "Diagnosis of Multiple and Unknown Faults Using the Causal Map and Multivariate Statistics," *Journal of Process Control*, 28, pp. 27–39 (April 2015).
25. Severson, K., et al., "Perspectives on Process Monitoring of Industrial Systems," in Proceedings of the 9th IFAC Symposium on Fault Detection, Supervision, and Safety for Technical Processes, Paris, France (Sept. 2–4, 2015).