Perspective

Process Data Analytics in the Era of Big Data

S. Joe Qin
School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, 2001 Longxiang Blvd, Longgang, Shenzhen 518172 China

DOI 10.1002/aic.14523
Published online July 2, 2014 in Wiley Online Library (wileyonlinelibrary.com)

Keywords: big data, machine learning, optimization, process data analytics, process control, process improvement

Introduction

For engineering systems where processes, units, and equipment are designed with clear objectives and are usually operated under well-controlled circumstances as designed, mechanistic models and first principles are dependable. However, for emerging circumstances that are not factored into the design, data become indispensable assets for decision making in safe and efficient operations. In this Perspective article, we offer a brief introduction to the essence of big data, a description of how data have been effectively used in process operations and control, and new perspectives on how chemical process systems might evolve into a new paradigm of data-enhanced operations and control. The discussed perspectives include (1) the mining of time-series data with expanded depth in history and breadth in location for event discovery, decision making, and causality analysis; (2) the exploration of the power of new machine-learning techniques that have enjoyed tremendous development in nearly 2 decades; and (3) the anticipation of a system architecture shift towards a data-friendly information system to complement the current distributed control systems centric information system. In addition, high-level systems engineering tasks such as planning and scheduling[1] can also benefit from information extracted from big data since optimization and control have always relied on the interplay between models and data. We note that big data is not the answer to everything, but historical and real-time data are valuable for safe and efficient operations, especially for abnormal process behaviors or circumstances that are not considered in the design phase.

Roles of Data in Science and Engineering

With the recent development of the Internet, the Internet of Things, smart and wireless sensors, wireless communications, mobile devices, smart devices, e-commerce, and smart manufacturing,[2] the amount of data collected and stored has grown exponentially in a manner analogous to Moore's law of the increase of solid-state transistor density. The explosion of data size has made all sectors, including engineering, medicine, business, commerce, finance, and even science, endorse the concept and power of big data. Take bioinformatics and genomes, for example. The Human Genome Project as an international effort to sequence the entire human genome aimed to uncover the sequence of 3 billion base pairs of the haploid human genome.[3,4] Recently, BGI Institute of China launched the 3-Million Genomes Project,[5] which includes a million plant and animal genomes, a million human genomes, and a million micro-ecosystem genomes. The goal of this project is to establish a baseline for specific populations and connect the phenotypes of diseases and traits to genetic variations to understand the disease mechanisms. The massive amount of information-packed data, although it promises to lead to new scientific discoveries in this field, produces a pressing need for effective analytics by combining biological and information sciences.[4,6]

The massive amount of available data has prompted many disciplines and industries to reexamine their traditional roles and views, such as statistics, management science, econometrics, computer science, and engineering. As a result, a new discipline known as *data science,* or *informatics,* is forming to derive knowledge and information from massive data. Several examples have shown that the possession of a huge amount of data confers a tremendous advantage when it is combined with effective analytics and superior computing power capable of distilling the data into knowledge. Google's flu prediction is such an example;[7] it predicted the spread of the winter flu outbreak in 2009 in the United States down to the state level. Google took 50 million of the most common searches and compared them to the Centers for Disease Control (CDC) data on the spread of the winter

Correspondence concerning this article should be addressed to S. J. Qin at sqin@usc.edu.

S. Joe Qin is on leave from the Mork Family Dept. of Chemical Engineering and Materials Science, University of Southern California, Los Angeles , CA, 90089 flu from 2003 to 2008. Google's data-processing power screened through 150 million models to discover 45 features with a mathematical model that had a high correlation with the data from the CDC. In addition, Google could predict in nearly real time, whereas the CDC's data took weeks to compile. While this data analytic approach is entirely new to chemical engineers, the functionality of the models is known as *inferential sensors* and is practiced in process systems engineering.[8,9]

Needless to say, the relevance of big data to science, engineering, and commerce seems to be higher than one might think. A bestselling monograph by Mayer-Schönberger and Cukier[10] provided an account of the technology for the general public. It is argued that by collecting a complete set of data rather than a sample, we can now analyze the data set in its entirety. Data analytics in the big-data era will shift from statistical sampling and inference to a focus on "$N =$ all."

Big Data and Analytics

Big data is arguably a major focus in the next round of the transformation of information technology in industry. According to research by McKinsey Global Institute and McKinsey's Business Technology Office, the analysis of large data sets will become a key basis of competitiveness, productivity growth, and innovation.[11] This analysis depicts five ways of using big data to create values:

- Unlocking significant value by making information transparent and usable.
- Collecting more accurate and detailed performance information and, therefore, exposing variability and boosting performance.
- Precisely tailored products or services.
- Substantially improving decision making through sophisticated analytics.
- Finally, improving the development of the next generation of products and services through big data. For instance, manufacturers use data obtained from sensors embedded in products to create innovative aftersales service offerings, such as proactive maintenance.

Big data refers to the size and variety of data sets that challenge the ability of traditional software tools to capture, store, manage, and analyze. Increasingly massive data sets are gathered by equipment and process sensors, mobile and wireless devices, software logs, cameras, microphones, and wireless sensor networks. Often three *V*'s are used to characterize the essence of big data.[11]

- Volume: Enterprises have evergrowing data of all types and easily amass terabytes, or even petabytes, of data. For example, they can convert 350 billion annual meter readings to better predict consumption.
- Velocity: For time-sensitive processes such as catching fraudulent activity, big data must be used as it streams into the enterprise.
- Variety: Big data is composed of all types of data: structured and unstructured data such as texts, sensor data, audio, video, log files, and so on.

In addition to these basic characteristics, some have included value and veracity as additional *V*s for big data. Establishing trust in big data and conclusions based on them presents a challenge; this makes statistical learning theory an indispensable framework that brings science into the big-data picture.[12]

Both governments and private sectors are making serious efforts to embrace the opportunities in big data. In 2012, US President Barack Obama announced the Big Data Research and Development Initiative, which explored how big data could be used to address important problems facing the government. China considers big data, the Internet of Things, cloud computing, and smart cities as key technologies for the leap forward in the process of industrialization, digitalization, and urbanization. Traditional retail giant Walmart handles more than 1 million customer transactions every hour, and these are imported into databases that contain more than 2.5 petabytes of data, whereas e-commerce leaders like Amazon build their processes around data and derive significant market information from them.

It is interesting to note that the thrust of interest in big data does not only come from large information technology companies like Google, IBM, and Microsoft; it also comes from traditional companies such as GE and P&G. In 2011, GE announced a $1 billion investment in the building software and expertise for GE's version of big-data analytics.[13] Its objective is to build a global software center to power up data-science capabilities in GE's "Industrial Internet." The list goes on from search engines (eg, Google and Microsoft) and social networks (eg, Twitter, Facebook, and LinkedIn) to financial institutions, the health care industry, engineering companies, retail analytics, mobile analytics, marketing agencies, data-science vendors (eg, Teradata, SAS, and SPSS), utilities, and government. The opportunity for process systems engineering[14] is to integrate engineering, design, operation, and customer data to improve process operations, efficiency, product quality, and customer satisfaction with individualized specifications.

Massive and Diverse Multivariate Process Data

Manufacturing process operation databases are massive because of the use of process operation and control computers and information systems. The diversity of process measurement technologies from conventional process sensors to images, videos, and indirect measurement technologies has compounded the variety, volume, and complexity of process data. Multilevel and multiscale data in semiconductor manufacturing, for example, provides at least the following levels of massive data:

- Equipment-level process measurement data.
- Process-level metrology data.
- Add-on indirect quality measurement data.
- In-process wafer electrical property test data.
- Final wafer electrical property test data.

It is typical in a modern FAB that over 50,000 statistical process control charts are monitored to control the quality of over 300 manufacturing steps in the fabrication of the chip.[15] A recent *AIChE Journal* Perspective article[16] characterized the massiveness as "drowning in data."

Although process operations are rich in data, without effective analytical tools and efficient computing technology to derive information from data, it is often the case that data

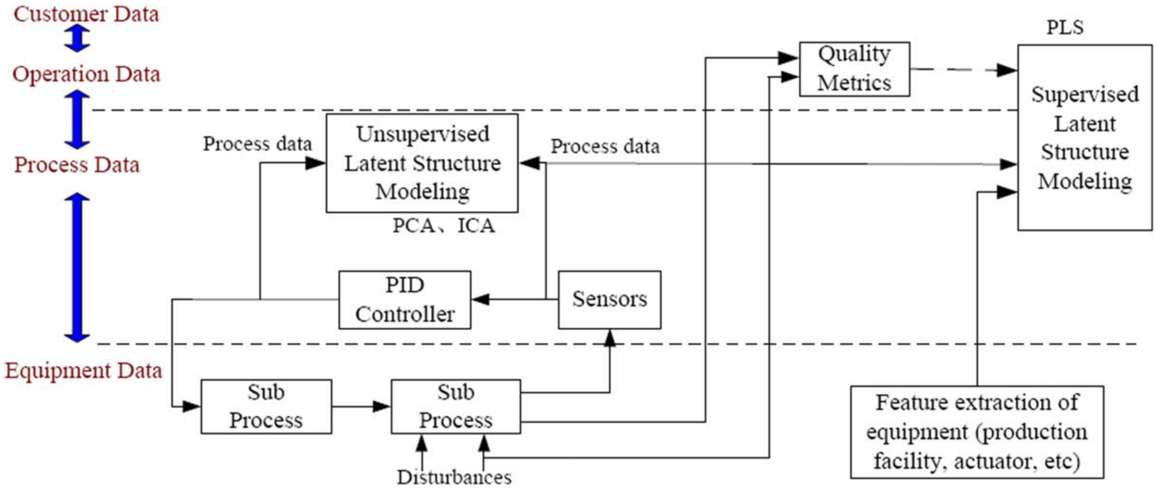


Figure 1. Process data hierarchy. ICA, independent component analysis; PID, proportional-integral-derivative.

are compressed and archived for record keeping and only retrieved for use in emergency analysis after the fact rather than being used in a routine manner in the decision-making process. Short-term archiving is typically adopted because the values from rare use of the historical data do not justify the expense of added hardware storage capacity. This current state in the chemical manufacturing industry is in sharp contrast to those of new data giants such as Amazon, Google, and eBay, who archive everything because they consider data as assets.

Process operations are of tiered and hierarchical objectives; they make process data more or less structured rather than unstructured like those derived from social networks. In fact, a vast amount of informative data is collected in the form of time series with regular sampling periods and usually well-defined purposes to measure them. Some analytical and monitoring tools, such as multivariate statistics and neural networks, have been adopted by chemical manufacturers with varying degrees of success. These techniques often achieve the detection and root-cause diagnosis of quality-related operational faults sooner than actual product quality control and customer feedback. In other cases, they detect process degradation or changes that could lead to more serious process failure or unsafe incidents. Process operations benefit from statistical and data-driven methods for the following reasons:

- The complexity of process operations with multiple grades and varying raw materials.
- The control of complexity with multilevel objectives.
- Rich instrumentation and control feedback.
- Rich process data collection and archiving.

The process and data considered for the data-driven fault detection and diagnosis are illustrated in Figure 1, where the hierarchical data structure is shown. At the bottom level are the equipment sensor measurements that can happen in milliseconds. At the process level are regularly sampled process control data. The product quality measurements come in all forms and are often sampled irregularly. The top level is the customer feedback data, which can range from customer service channels to social network complaints. Compared to the expense and effort of rigorous process modeling, the advantages of data-driven latent structure modeling methods, such a principal component analysis (PCA) and projection to latent structures (PLS), are obvious; they can be used to detect abnormal changes in the process operations from real-time data because of their dimension-reduction capability, ease of visualization, and ease of real applications. The related fault diagnosis methods have been studied intensively and applied successfully in many industrial processes, for example, chemicals, iron and steel, polymers, and semiconductor manufacturing.

Process data are often categorized into process input and output data, quality output data, and indirect types of data (eg, vibration signals and images), as shown in Figure 1. The typical procedure of the current multivariate process data analytics includes

- The collection of (very clean) normal data with good coverage of the operating regions.
- Fault data cases, which can be useful but are not required.
- Latent variable (LV) methods (eg, PCA and PLS) to model the data.[17]
- Fault detection indices and control limits.
- Fault diagnosis and troubleshooting.

Although multivariate statistical approaches have been the favorite choice in the most recent 2 decades, it should be noted that data have been an integral part of process engineering solutions since time models have been used for process optimization and control. Data reconciliation[18] has been a necessary function in the deployment of real-time optimization with real-time data. Neural networks for inferential property modeling[8,9] are routinely used in industry now. Time-series trend analysis[19] was proposed to obtain real-time patterns from operating data for diagnosis and control. System identification and time-series modeling is a necessary tool for applying model-based control and control performance assessment. The Kalman filter,[20,21] developed half a century ago, provides an elegant framework for making a balanced use of both data and a mechanistic model. At the higher level of production planning and scheduling, industry has accounted for the typical uncertainty in, for example, product demand and price, using time-series analysis and data mining.[22]

However, these aforementioned data analytics and practice in process systems engineering have apparently not connected to the recent development in machine learning, data mining, and big-data analytics. They differ not only in terms of sizes but also in how and what data should be used in solving real operation problems. In some ways, process systems engineering solutions are confined to one set of principles that are believed to be sound, whereas the machine-learning and data-mining communities take the other way and achieve unexpected results and solutions that defy conventional wisdom. For example

- While multivariate data analytics tend to require a set of carefully collected clean data or the pretreatment of outliers and missing values,[23] the data-mining and machine-learning communities have developed robust methods that use imperfect data as the norm rather than as the exception.
- While neural networks are used for inferential modeling with a minimalist representation with as few layers and parameters as possible, deep learning techniques[24,25] outperform all of the alternatives by using many layers, or by being "greedy."
- While the time-series trend analysis[19] from operating data are only one of a few articles published in process systems engineering with little attention in many years, the time-series data-mining community has developed a whole set of techniques that are essentially trend analysis.[26,27]
- While most research work in process systems engineering has focused on the possession of a clean matrix of data with great regularity, with irregularly and indirectly measured data left largely unused, other industries have extracted valuable information from highly unstructured data.[7]

These gaps motivated me to provide several perspectives in the remainder of this article so as to inspire new ideas to enrich process-data-analytics methodologies in the era of big data. These include the following perspectives: (1) to increase variety, an analysis of heterogeneous sources of data during process operations and after products is made; (2) to increase volume, the mining of massive historical time-series data for event discovery, decision making, and causality analysis is considered; (3) to improve value and veracity, an embrace of the power of new machine-learning techniques developed over nearly 20 years is encouraged; and (4) to improve velocity, a potential shift in system architecture toward a data-friendly information system to complement the current control-centric system is explored.

For Variety: Multilevel Heterogeneous Data Analytics

Process data analytics for process monitoring should have the following desirable features:

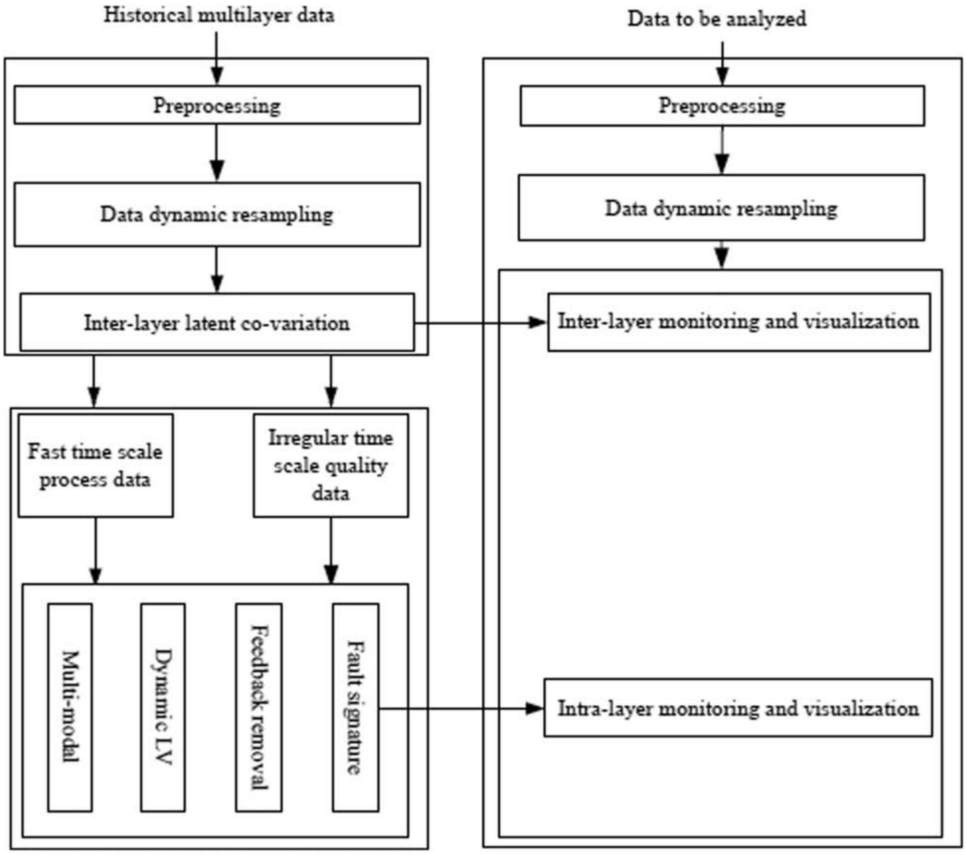
- They should be scalable (up to 100 000 variables).
- They should make use of all kinds of data, for example, process, spectra, vibration, and image data.[28]
- They should be relatively easy to apply to real processes compared to other methods.
- They should include online use for real-time operations and decision making.
- They should include offline troubleshooting as a valuable tool for continuous improvement.

To make use of all kinds of data from historical and real-time process operations, one obvious direction for process systems engineering to embrace the new techniques in big data is to increase the diversity of data used in the analysis. PCA has been a favorite tool for process analysis,[29] but it is only capable of analyzing data from a single level; that is, all variables are considered to be of the same importance in the exploration of cross-correlations or auto-associations. The machine-learning literature considers the PCA type of models to be single-layer modules.[30] A recent survey by Qin[31] and as the references therein gave an account of the available process data analytics and applications in process and quality monitoring. PLS offers one option for exploring multilevel data correlation and intralevel variations. Figure 2 depicts a scheme for two-level dynamic concurrent latent structure modeling and monitoring. Data from tiered operation objectives can be preprocessed and resampled to align features for latent structure modeling. The interlevel analysis and intralevel analysis in the concurrent PLS in Qin and Zheng[32] explained variations not only for the output level but also within the input level. The tiered analytic objectives are found in the semisupervised scheme of machine learning defined in Bengio et al.[30] as follows: "Semi-supervised learning: with inputs X and target Y to predict, a subset of the factors explaining X's distribution explain much of Y, given X."

In industrial processes, because of many active or passive operational changes, the normal operation conditions take multimodal distributions, making unimodal approaches like PCA inadequate. For process data with various operation conditions, multimode modeling methods should be developed. The existing work mainly includes multiple-model approaches, local-learning methods, and so on. For processes whose operation condition changes from one operational mode to another, multimodal methods are necessary. One type of multimode modeling method is local learning. In this method, the modeling construction is carried out online. Each time a sample is available, the first step is to search similar data samples in the historical data set. Then, on the basis of obtained similar data samples, an online model is developed for online monitoring. After that, the model is discarded, and the procedures are repeated for incoming data samples. Recently, the local-learning method was adopted by Ge et al.[33] for the online monitoring of nonlinear multimodal processes. In addition to multimode analytics, the ability to extract dynamic LVs and remove the effect of feedback in the data is necessary to deal with data collected under dynamic feedback control.

For Volume: Time-Series Data Representation

Existing process data analytics usually focus on well-defined data samples for carefully selected variables that bear the same conditions as the current process operation. There is virtually no effort to pool together data from a



Historical multilayer data
Preprocessing
Data dynamic resampling
Inter-layer latent co-variation
Fast time scale process data
Irregular time scale quality data
Multi-modal
Dynamic LV
Feedback removal
Fault signature
Data to be analyzed
Preprocessing
Data dynamic resampling
Inter-layer monitoring and visualization
Intra-layer monitoring and visualization

Figure 2. Multilevel dynamic concurrent latent structure modeling and monitoring.

diverse source of measurements, years of historical operations, and globally distributed plants that make the same products. Until these are explored, traditional process data analytics dwarf themselves compared to the big-data practice happening in other industrial sectors.[7] The capability of big-data analytics has the potential to knock down these somewhat artificial barriers to really enrich the volume and representativeness of the data. While the traditional data analysis tools emphasize the cleanness of the data to prevent potential misleading conclusions, big-data analytics consider data errors or messiness as unavoidable but use massive data to develop models and extract features that are robust to the imperfections in the data.

Even as data are pooled across spatiotemporal sources, the majority of process data are of time-series forms with regular sampling periods. This characteristic makes it appropriate to apply time-series data-mining and representation methods that have been studied extensively in other areas (Fu, 2011). The various tasks in time-series data-mining research include

- Indexing (query by content): Given a query time series and some similarity/dissimilarity measurement, it finds the nearest matching time series in the database.
- Clustering: It finds natural groupings of the time series in the database under some similarity/dissimilarity measure.
- Classification: Given an unlabeled time series, it assigned it to one of two or more predefined classes.
- Segmentation: Given a time series containing plenty of data points, it constructs a model with smaller piecewise segments such that the latter closely approximates the former. This representation makes the storage, transmission, and computation of the data more efficient. Specifically, in the context of data mining, piecewise linear representation is used.
- Dynamic time warping: This measures the distance between two time series after first aligning them in the time axis.

Segmentation may be performed simply to create a simpler representation of the time series that supports indexing, clustering, and classification[26] (Fu, 2011). It should be noted that the segmentation of the time series is similar to the trend analysis technique explored in Bakshi and Stephanopoulos[19] very early in process systems engineering. Pattern matching for disturbance mining was studied by Singhal and Seborg.[34] It is unfortunate that these efforts have not brought enough attention in process systems engineering, although a dynamic PCA approach[35] was used for the limited categories of disturbances simulated with the Tennessee Eastman Challenge problem. Recent work by Sun et al.[36] explored the use of historical time-series data to assess the disturbance model adequacy and model prediction accuracy in a model predictive control context.

With time-series historical data, another important analytic tool is Granger causality analysis.[37] Because of its simplicity, interpretability, and ease of implementation, Granger causality has found wide applications in economics. Recently, Granger causality has gained great attention in many other areas in the extraction of useful information and inner causal relationships, including the identification of root causes of important process features, such as the root cause of plantwide oscillations.[38] Granger causality builds a straightforward connection between causality and prediction and employs a statistical hypothesis test to determine whether one time series is helpful in forecasting another.

For Value and Veracity: Statistical Machine Learning

Machine learning and artificial neural networks have gone through several ups and downs since their inception. Initially showing promise in making machines learn from data in the 1960s, machine learning saw its downturn in the 1970s until it caught a new resurgence in the mid-1980s. This time, process systems researchers and practitioners joined efforts in applying machine learning to chemical engineering problems.[8,39,40] After the limitations and benefits of artificial neural networks were debated and explored, the whole domain of machine-learning techniques boiled down to the adoption of multilayer feed-forward networks for use as inferential sensors for chemical processes. This more or less wrapped up the process systems engineering community's interest in machine learning and neural networks. In the meantime, the machine-learning community also ran into its own bottleneck in the search for new ideas.

The golden period for statistical machine learning was between 1996 and 2006 with the development of support vector machines (SVMs) proposed by Vapnik[41] and the boosting and kernel methods discussed by Freund and Schapire.[42] Statistical machine learning has become a major branch in computer science and artificial intelligence.[12,43] In 2011, Judea Pearl won the Turing Award[44] for fundamental contributions to artificial intelligence through the development of a calculus for probabilistic and causal reasoning.

Statistical machine learning is considered to bridge computation and statistics, with ties to information theory, signal processing, algorithms, control theory, and optimization theory.[12] Effective techniques developed in this area include boosting and SVMs, kernel PCA, and multidimensional scaling. In addition, the work of Hinton and Salakhutdinov,[24] published in *Science,* focused on deep learning. Recent statistical machine-learning development possesses three attractive features: (1) it focuses on its ability to extract knowledge from data, whereas traditional ones focus on making the machine learn; (2) it establishes a theoretical basis in statistics as a discipline to control errors in inference; and (3) it is data driven and target driven and enjoys new contributions from information industry sectors.

In addition to kernel methods and SVM, the so-called latent data models are effective tools in analyzing multidimensional data. There are three attractive features for the LV specified in a latent data model: (1) they can potentially have useful physical meaning; (2) they are relatively independent, instead of having a strong correlation among the original variables; and (3) they can be designed with more efficient computation methods. Latent structure models, such as PLS and PCA, as applied to chemometrics and process monitoring were actually developed for the same purposes and reasons. It is, therefore, worthwhile to explore what latent data models can offer for process data analytics and their connections to latent structure models.

Another interesting problem in statistical machine learning is low-rank matrix approximation. In theory and practical applications, it has been shown that a data matrix can usually be expressed as a low-rank matrix plus a sparse matrix. Robust PCA and matrix completion models have been proposed and applied effectively to video data and text data analysis. An important feature of the robust methods is that they are made insensitive to outliers or missing values through the use of zero norms or one norm rather than the typical two norms used in the standard PCA approach. For low-rank matrix approximation methods, it is important to design an appropriate computation method. The rich process systems expertise in optimization promises to make great contributions to the area of optimization for machine learning, as highlighted in Sra et al.[45]

It is inspiring to see that optimization and control researchers in process systems engineering have already made contributions to the interplay between optimization and machine learning.[46,47] Process-data-analytics techniques developed from big-data analytics could provide distilled data or data representation for the following systems engineering tasks:

- *Metamodeling:* The use of data-analytics techniques to generate surrogate or reduced-dimension process models that are accurate and computationally efficient.
- *Derivative-free optimization:* Optimization based on techniques that make use of data or their effective representation.
- *Optimization under uncertainty:* The characterization of uncertainty with data-driven machine-learning tools to enhance optimization techniques to work with uncertainty.

For Velocity: Real-Time, Data-Centric System Architecture

At present, process data historians are built around distributed control systems, where the central task is to perform feedback control and some form of operator user interface. Data historian systems, such as the PI system from OSIsoft, are real-time databases that replaced strip chart recorders. One limitation of such systems is that it is very difficult to even manually integrate them with other high-level operational databases. It is not realistic to expect that this type of system architecture would support the massive requirements in accessing a diverse source of historical and real-time data and be fast enough to build models on the fly.

Data-centric system architectures and computational frameworks are mainly built on large-scale distributed storage and processers. The storage structure and computing engine are constructed to be suitable for real-time data analysis, whereas knowledge is extracted from massive data with data analytics. The extracted knowledge can usually be delivered to end users by cloud services.

Hadoop, developed by Doug Cutting and Michael J. Cafarella in 2005 with support from Yahoo, is the most popular big-data-processing platform because of its attractive characteristics of high reliability, high efficiency, and high scalability.[48] It is based on the Google file system and the idea of MapReduce. Hadoop has become the de facto standard for big-data-processing frameworks. As a storage and processing platform of large distributed data-processing systems, Hadoop is composed of two main parts, a distributed file system to support massive data storage and a distributed computing framework to support MapReduce[49] data processing. The computational frameworks can be operated in clusters, whereas the computations and resources are managed and scheduled by the resource management system with efficient resource utilization. Computing frameworks can be categorized into three modes: the streaming mode, the batch mode, and the mixed mode.

Streaming mode

The streaming mode considers the data flowing in as a steady stream. When new data arrive, they are processed immediately. The processing is usually performed in memory because of the real-time requirement; this makes the processing rely on clever data structure in memory. The limited memory capacity can become a major bottleneck. The application of streaming models mainly focuses on real-time statistical analysis, online monitoring, and so on.

Batch mode

MapReduce as a typical batch framework was developed by Google, Inc.[49] It was originally designed to accomplish parallel processing of a large amount of data through large-scale, low-cost server clusters. MapReduce has advantages in simple interface and powerful data processing on large-scale parallel executions, fault tolerance, and load-balancing implementations. MapReduce has been widely used in data mining, machine learning, information retrieval, computer simulation, and scientific experiments.

Mixed mode

The streaming mode and batch mode can be integrated to form a mixed mode. The basic idea is to apply the MapReduce model for streaming processing. For example, Thusoo et al.[50] discussed the framework adjustment when applying MapReduce to streaming, single-pass processing. On the basis of this analysis, Mazur et al.[51] introduced the realization of a scalable platform of MapReduce for single-pass analysis. In the streaming MapReduce process discussed in Li et al.,[52] who considered event stream processing; the Mapper and Reducer in MapReduce were redefined to improve the processing ability for continuous data.

Process data analytics in the traditional data-processing architecture and the MapReduce-based big-data-processing architecture are faced with the following issues: (1) As MapReduce is a batch-oriented parallel computing model, it shows unsatisfying performances in real-time data analytics; 2) even common data-analytics (e.g., PCA) algorithms cannot be easily performed when the amount of data is too large; and 3) the existing big-data-processing platform usually performs simple data query but not in-depth data analytics.

It is desirable to establish a data-centric architecture for the real-time data analytics of multilevel heterogeneous data across spatial and temporal domains. The following directions are speculated:

- The integration of streaming processing algorithms with the MapReduce architecture, for example, incremental computing modeling and iterative incremental computing modeling.
- The establishment of a hardware-software architecture in which data resources and computing resources can be shared, allocated, and merged, for example, for the parallel operation of multiple computing platforms to increase the utilization and reduce maintenance costs.
- The development of efficient statistical machine-learning algorithms on the MapReduce (or Hadoop) framework to implement real-time data analysis, for example, for the feature extraction, mining, querying, clustering, and classification of time-series data. For process data analytics, the efficient querying of data segmentation can be achieved with the parallel computation of open-source software.

Summary

Process data analytics has been an indispensable tool in chemical process operations. However, in the era of big data and the development of advanced analytics in other sectors of industries and business operations, there appears to be much more room to grow. Physical and chemical sciences develop principles that are established for phenomena or processes with well-understood mechanisms; data, on the other hand, provide realistic information that reflects unknown changes in the operation of these processes and are the only reliable source of information for characterizing uncertain and emerging situations not considered in the process design phase. This Perspective article covers several recent developments in data analytics from other disciplines that are perceived to be relevant to chemical process operations. These new perceived ways of using massive amount of data might shift the balance between a mechanistic model and the data, much like what the Kalman filter enabled half a century ago. In addition, high-level optimization tasks such as planning and scheduling can also benefit from information extracted from massive data, since optimization has always been based on the interplay between models and data.

Inference using massive data with controlled error tolerance is of major interest, where the goal is to turn data into knowledge and to support effective decision making and optimization. While it was possible to require clean and accurate data in small data samples, we might have to live with messiness of the data and contain the errors with massive data. Robust methods in statistical machine learning are effective ways to handle messy data, although some level of preprocessing is always helpful.

To make use of machine learning to extract knowledge from big data (sometimes all data but not small data), practitioners should familiarize themselves with data science tools that are relevant to but different from those in information science and statistics. Scientists and engineers are encouraged to possess (1) the ability to develop or use the big-data-processing architecture (MapReduce or Hadoop), (2) the ability to develop or appropriately use algorithms, and (3) the ability to model data and judge the validity of the knowledge discovered from the analysis of the data.

As process data analytics could be routinely used for process operations and decision making, the issue of cybersecurity in process operations and control becomes important because of the increased reliance on data and sensors. Although no computer systems are immune to this potential risk, process operations and control often require even higher security than other systems such as social networks. Data analytics itself with validity checking and intrusion detection could provide useful solutions. Another related issue, although somewhat deliberate, is the perpetual question of the confidentiality of accessing data and the ownership of the process or equipment data. Industries, including vendors and end users, must agree to engage in a new relationship that allows for a win-win situation to maximize the benefit of big-data analytics.

Finally, it should be noted that big data is not likely the answer to everything. Because of Silicon Valley's typical overreaction to new ideas, it would not surprise anyone if big-data excitement will settle down after some time. However, often, the tide will come back again with breakthroughs happening somewhere. The up-and-down progressions in machine-learning and time-series data modeling over half a century has provided enough food for thought in that innovations require an open mind and persistent effort. It takes patience and insight to bring a new technology to process systems operations where traditional procedures and solutions are in place.

Acknowledgments

The author is grateful to Qiang Liu, Quan Xu, Lei Zhang, Zhihua Zhang, and Gang Li for discussions in organizing this article and to anonymous reviewers for insightful comments. He would like to acknowledge funding from the Natural Science Foundation of China under Grant 61020106003, the 111 Project of Ministry of Education of China under Grant B08015, and SAPI Fundamental Research Funds (2013ZCX04), CiSoft between Chevron and the University of Southern California, and members of the Texas-Wisconsin-California Control Consortium.

Literature Cited

1. Grossmann IE. Advances in mathematical programming models for enterprise-wide optimization. *Comput Chem Eng*. 2012;47:2–18.
2. Davis J, Edgar T, Porter J, Bernaden J, Sarli M. Smart manufacturing, manufacturing intelligence and demand-dynamic performance. *Comput Chem Eng*. 2012;47:145–56.
3. Collins FS, Morgan M, Patrinos A. The Human Genome Project: lessons from large-scale biology. *Science*. 2003; 300(5617):286–90.
4. Bickel PJ, Brown JB, Huang H, Li Q. An overview of recent developments in genomics and associated statistical methods. *Philos Trans A Math Phys Eng Sci*. 2009; 367:4313–37.
5. 3M project. BGI. http://www.genomics.cn/en/navigation/show_navigation?nid=5656. Accessed on May 10, 2014.
6. Waterman MS. *Introduction to Computational Biology: Maps, Sequences and Genomes*. Boca Raton, FL: CRC; 1995.
7. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. *Nature*. 2009;457:1012–4.
8. Tham M. Soft-sensors for process estimation and inferential control. *J Process Control*. 1991;1(1):3–14.
9. Qin SJ, McAvoy TJ. A data-based process modeling approach and its applications. In Proceedings of the 3rd IFAC DYCORD Symposium. College Park, MD: 1992: 321–6.
10. Mayer-Schönberger V, Cukier K. Big Data: A Revolution that Will Transform How We Live, Work, and Think. Boston, MA: Houghton Mifflin Harcourt; 2013.
11. Manyika J, Chui M, Brown B, et al. Big data: the next frontier for innovation, competition, and productivity. McKinsey & Company. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Published May 2011. Accessed on May 10, 2014.
12. National Research Council. *Frontiers in Massive Data Analysis*. Washington, DC: National Academies Press; 2013.
13. Tata Consultancy Services. The Emerging Big Returns on Big Data: A TCS 2013 Global Trend Study. Tata Consultancy Services: 2013. http://www.tcs.com/big-data-study/Pages/download-report.aspx. Accessed on May 10, 2014.
14. Stephanopoulos G, Reklaitis GV. Process systems engineering: from Solvay to modern bio- and nanotechnology. A history of development, successes and prospects for the future. *Chem Eng Sci*. 2011;66:4272–306.
15. Qin SJ, Cherry G, Good R, Wang J, Harrison CA. Semiconductor manufacturing process control and monitoring: a Fab-wide framework. *J Process Control* 2006;16: 179–91.
16. Venkatasubramanian V. Drowning in data: informatics and modeling challenges in a data-rich networked world. *AIChE J*. 2009;55(1):2–8.
17. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal Chim Acta*. 1986;185:1–17.
18. Mah RS, Stanley GM, Downing DM. Reconciliation and rectification of process flow and inventory data. *Ind Eng Chem Process Des Dev*. 1976, 15(1), 175–83.
19. Bakshi BR, Stephanopoulos G. Representation of process trends—IV. Induction of real-time patterns from operating data for diagnosis and supervisory control. *Comput Chem Eng*. 1994;18(4):303–32.
20. Kalman RE. A new approach to linear filtering and prediction problems. *J Basic Eng*. 1960;82(1):35–45.
21. Gibbs BP. Advanced Kalman Filtering, Least-Squares and Modeling: A Practical Handbook. New York, NY: Wiley; 2011.
22. Klatt KU, Marquardt W. Perspectives for process systems engineering—personal views from academia and industry. *Comput Chem Eng*. 2009;33:536–50.
23. Nelson PR, Taylor PA, MacGregor JF. Missing data methods in PCA and PLS: score calculations with incomplete observations. *Chemometr Intell Lab Syst*. 1996;35:45–65.
24. Hinton GE, Salakhutdinov R. Reducing the dimensionality of data with neural networks. *Science*. 2006; 313(5786):504–7.

25. Bengio Y, Lamblin P, Popovici D, Larochelle H. Greedy layer-wise training of deep networks. *Adv Neural Inf Process Syst.* 2007;19:153–60.
26. Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002:102–11.
27. Fu TC. A review on time series data mining. *Eng Appl Artif Intell.* 2011;24:164–81.
28. Yu H, MacGregor JF. Multivariate image analysis and regression for prediction of coating content and distribution in the production of snack foods. *Chemometr Intell Lab Syst.* 2003;67:125–44.
29. Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometr Intell Lab Syst.* 1995;28(1):3–21.
30. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell.* 2013;35(8):1798–828.
31. Qin SJ. Survey on data-driven industrial process monitoring and diagnosis. *Annu Rev Control.* 2012;36: 220–34.
32. Qin SJ, Zheng YY. Quality-relevant and process-relevant fault monitoring with concurrent projection to latent structures. *AIChE J.* 2013;59:496–504.
33. Ge ZQ, Yang CJ, Song ZH, Wang HQ. Robust online monitoring for multimode processes based on nonlinear external analysis. *Ind Eng Chem Res.* 2008;47:4775–83.
34. Singhal A, Seborg DE. Evaluation of a pattern matching method for the Tennessee Eastman challenge process. *J Process Control.* 2006;16:601–13.
35. Ku W, Storer R, Georgakis C. Disturbance detection and isolation by dynamic principal component analysis. *Chemometr Intell Lab Syst.* 1995;30(1):179–96.
36. Sun ZL, Qin SJ, Singhal A, Megan L. Performance monitoring of model-based controllers via model residual assessment. *J Process Control.* 2013;23:473–82.
37. Granger C. Some recent development in a concept of causality. *J Econometr.* 1988;39:199–211.
38. Yuan T, Qin SJ. Root cause diagnosis of plant-wide oscillations using granger causality. *J Process Control.* 2014;24:450–9.
39. Hoskins JC, Himmelblau DM. Artificial neural network models of knowledge representation in chemical engineering. *Comput Chem Eng.* 1988;12:881–90.
40. Ydstie BE. Forecasting and control using adaptive connectionist networks. *Comput Chem Eng.* 1990;14(4): 583–99.
41. Vapnik V. The Nature of Statistical Learning Theory. New York, NY: Springer-Verlag; 1999.
42. Freund Y, Schapire RE. A short introduction to boosting. *J Jpn Soc Artif Intell.* 1999;14(5):771–80.
43. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found Trends Mach Learn.* 2011;3(1):1–122.
44. Russell SJ. Judea Pearl. A. M. Turning Award. http://amturing.acm.org/award_winners/pearl_2658896.cfm. Accessed on May 10, 2014.
45. Sra S, Nowozin S, Wright SJ. Optimization for Machine Learning. Cambridge, MA: MIT Press; 2011.
46. Cozad A, Sahinidis NV, Miller DC. Learning surrogate models for simulation-based optimization. *AIChE J.* 2014;60(6):2211–27.
47. Choi J, Realff MJ, Lee JH. Dynamic programming in a heuristically confined state space: a stochastic resource-constrained project scheduling application. *Comput Chem Eng.* 2004;28(6):1039–58.
48. Vance A. Hadoop, a free software program, finds uses beyond search, The New York Times. March 16, 2009. http://www.nytimes.com/2009/03/17/technology/business-computing/17cloud.html?_r=0. Accessed on May 10, 2014.
49. Dean J, Ghemawat S. MapReduce: simplified data processing on large clusters. *Commun ACM.* 2008;51(1): 107–13.
50. Thusoo A, Sarma JS, Jain N, et al. Hive—a petabyte scale data warehouse using Hadoop. IEEE 26th International Conference on Data Engineering. 2010; 996–1005.
51. Mazur E, Li B, Diao Y, Shenoy P. Towards scalable one-pass analytics using MapReduce. In IPDPSW ‘11 Proceedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum. Washington, DC: IEEE Computer Society; 2011:1102–11.
52. Li B, Mazur E, Diao Y, McGregor A, Shenoy P. A platform for scalable one-pass analytics using MapReduce. In SIGMOD ‘11 Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. New York, NY: Association for Computing Machinery; 2011:985–996.