Big Data Analytics in Chemical Engineering

Leo Chiang, Bo Lu, and Ivan Castillo

The Dow Chemical Company, Freeport, Texas 77541; email: hchiang@dow.com

Annu. Rev. Chem. Biomol. Eng. 2017. 8:63–85

First published online as a Review in Advance on February 27, 2017

The *Annual Review of Chemical and Biomolecular Engineering* is online at chembioeng.annualreviews.org

https://doi.org/10.1146/annurev-chembioeng-060816-101555

Keywords

big data analytics, Internet of things, data-driven modeling, Industry 4.0, machine learning

Abstract

Big data analytics is the journey to turn data into insights for more informed business and operational decisions. As the chemical engineering community is collecting more data (volume) from different sources (variety), this journey becomes more challenging in terms of using the right data and the right tools (analytics) to make the right decisions in real time (velocity). This article highlights recent big data advancements in five industries, including chemicals, energy, semiconductors, pharmaceuticals, and food, and then discusses technical, platform, and culture challenges. To reach the next milestone in multiplying successes to the enterprise level, government, academia, and industry need to collaboratively focus on workforce development and innovation.

INTRODUCTION

Big data is an emerging topic affecting many aspects of our lives. In 2012, President Obama launched his big data research and development initiative. In 2014, the resulting committee published a report (1) discussing big data potential and data policy, stating

> [w]e are living in the midst of a social, economic, and technological revolution. How we communicate, socialize, spend leisure time, and conduct business has moved onto the Internet. The Internet has in turn moved into our phones, into devices spreading around our homes and cities, and into the factories that power the industrial economy. The resulting explosion of data and discovery is changing our world.

The big data era is driven by the explosion of data in all fields in terms of new data generation (e.g., social media), new measurement capability (e.g., internet of things, smart digital devices), improved data storage power (e.g., cloud computing), and improved computing technology for analytics (e.g., machine learning, artificial intelligence, cognitive computing).

In a business context, big data topics are regularly covered in public media. Major business publishers, including *The Economist*, *Fortune*, *Forbes*, *Financial Times*, *Harvard Business Review*, *Newsweek*, *The New York Times*, *The Wall Street Journal*, and *The Washington Post*, have published articles on topics ranging from an overview of big data to enabling big data technologies, from business value and use cases to future prediction, from government policy to data security and privacy.

In a science and engineering context, there has been an exponential increase in big data publications (see **Figure 1**). Premier journals *Nature* and *Science* published editorials on the role of big data in scientific research and highlighted the challenges and opportunities surrounding big

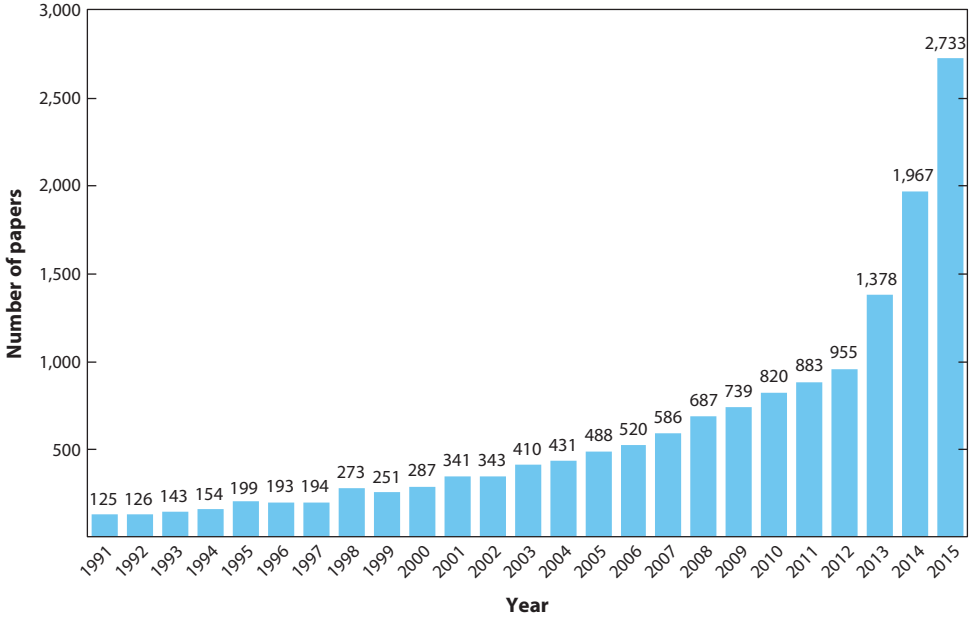

Figure 1

Number of publications containing the keyword "big data" since 1991 (through the Web of Science search engine).

data (2, 3). Perspectives of big data on specific fields such as neuroscience (4) and chemometrics (5) are available. The National Academy of Engineering discussed critical big data topics in two Frontiers of Engineering conferences and published its Winter 2014 report highlighting global perspectives on big data development (6).

In a chemical process industry context, Qin (7) commented that for well-understood chemical mechanisms, first-principles approaches can be used effectively to develop mechanistic models for process operations. For complex processes for which first principles are not well understood, process data analytics are valuable assets to provide insights on process improvements. Big data has room to grow into a new paradigm for process industries to enhance data-driven operations and control. In *Chemical Engineering Progress*'s March 2016 special issue on big data analytics, a four-article series outlines the Why (why you should care about big data) (8), the What (success stories in the process industries) (9), the How (getting started on the journey) (10), and the Future (challenges and future research directions) (11). Given that big data discussions are blooming in all fields, it is perhaps surprising to see that a literature search using keywords "big data" and "chemical engineering" does not result in a substantial number of references.

This article first provides a definition for big data and then explores recent advancements in data-driven approaches in five industries, including chemicals, energy, semiconductors, pharmaceuticals, and food. The goal of this article is twofold: (*a*) to educate the chemical engineering community about big data's ability to enable more possibilities to accelerate research and development (R&D) and to improve operational reliability and efficiency in various industry sectors and (*b*) to emphasize future big data research directions and how the community can collectively respond to challenges.

DEFINING BIG DATA FOR CHEMICAL ENGINEERS

Early discussions of big data date back to 2001, when an analyst of the META Group (currently Gartner) used the 3 V's to describe the characteristics of data growth (12):

- Volume: the ever-increasing amount of data generation and collection
- Velocity: the need for faster collection and processing speed to deal with large volumes of data
- Variety: the need to contextualize all types of data, including structured and unstructured data (such as texts, audio, video, webpages, and reports)

An additional V, veracity, is often added to indicate that not all data are created equal and that varied noise and uncertainty in data present a challenging aspect for data analysis. Data veracity is often a coupled challenge when the other three V's are present.

Since the term big data was coined, there have been several debates on its precise definition. As of September 2016, Wikipedia (13) defines big data as "data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time. Big data size is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data." The ARC advisory group points out the common misconception that big data is a thing (8), and that big data must be big and new hardware and software tools must be used. This is a likely reason why so few big data–themed articles were found in the chemical engineering literature. To clear up the misconception, ARC defines big data as "a journey toward more informed business and operational decisions" (8, p. 32). Simply put, a big data journey has started if volume, velocity, and variety characterize your data challenges and steps are taken to seize the opportunities. To emphasize more fully the potential of big data, a more preferable industry term, big data analytics, is used in this article. Big data analytics is the journey

to turn data into insights for more informed business and operational decisions. Although most papers cited in the next section do not reference the term big data, they use data-driven modeling approaches satisfying the original 3 V's:

- Volume: use an order of magnitude more data to improve decision making
- Velocity: speed up the decision making cycle between data generator and decision maker
- Variety: combine multiple sources of data to validate existing knowledge and to generate new ideas

ADVANCES IN APPLICATIONS

Chemical Process Industry

The chemical process industry spans a large scale from commodity chemicals, petrochemicals, refinery products, specialty chemicals, and life sciences to consumer products. The scale of manufacturing increases from small-scale, specialized products, such as life sciences or consumer products, to large-scale chemical and petrochemical production facilities. Modern petrochemical and chemical complexes are tightly integrated, with many manufacturing units geographically concentrated in one location. As a result, incremental improvements in energy efficiency, reliability, and safety would be amplified owing to the economy of scale. A recent survey by PricewaterhouseCoopers reveals that 88% of chemical industry executives acknowledge that data analytics will be crucial for retaining the competitive advantage in five years (14). As a result, big data analytics is expected to be a key growth area in the industry.

Historically, the chemical process industry is one of the earliest adopters of computer-based control. Safe and efficient plant operation requires constant monitoring of thousands of process variables. Collection of process data for monitoring and control became routine and provided a platform for exploration and development of data-driven methods and applications. Database servers designed specifically for storing process data, such as OSISoft PI™ and AspenTech IP21™, enabled easy and reliable access to process data by engineers and researchers. The chemical process industry has been a pioneer in adopting data-driven tools owing to this advantage (7). This section first highlights the ongoing data-driven analytics efforts that have already seen success in the industry and then discusses the emerging technologies that could potentially revolutionize analytics-driven decision making.

Advances in continuous processes with higher data volumes. In process monitoring, univariate control charts have been the de facto standard in ensuring processes are operating within safe limits (15). As plants become more heavily instrumented with thousands of sensors and actuators, the large volume of data can easily overwhelm plant personnel. As a result, many of the data streams are often not examined. Enterprise Manufacturing Intelligence (EMI) is a platform to contextualize critical plant key performance indicators into dashboards for real-time visualization. Based on process knowledge, a troubleshooting guideline can then be built into the EMI platform, facilitating data- and knowledge-driven decisions. The EMI platform has shown measurable successes at the Dow Chemical Company (referred to as Dow in later texts) (10, 16).

Multivariate analysis is another way to contextualize a large volume of data. Continuous processes result in dense (as opposed to sparse) and structured data streams. Multivariate analysis is an approach to exploit the naturally occurring correlation structures in these dense data sets owing to flow, mass transfer, energy transfer, and basic thermodynamics. These analysis methods can be applied in process monitoring to detect faults and ill conditions using process data with much higher dimensionality. Comprehensive overviews of the state-of-art fault-detection methodologies

are available (17–19). Among the data-driven modeling methods, projection-based methods, such as principal component analysis, partial least squares (PLS), independent component analysis, and canonical variate analysis, have dominated the literature because of their effectiveness in characterizing large, complex data sets (15). However, model degradation, multiple modal behavior, and process nonlinearity remain challenges to sustain model performances. Many efforts have been made to integrate new machine learning and statistical tools to address these issues, such as Gaussian mixture modeling (20), Gaussian process regression, multiple model systems (21), neural networks, support vector machines (22), and kernel-based extensions (23, 24). These methods have been demonstrated to be effective in simulation data sets, such as the widely used Tennessee-Eastman problem (25). The current research suggests that there is no magic bullet for every scenario; instead, the modelers need to create an appropriate combination of tools tailored for each application.

In addition to process monitoring, inferential sensors predict important variables using available process data. The important variables being predicted are often difficult or uneconomical to measure online. These predictions are often used for advanced control or quality monitoring and allow plants to react faster to process excursions to prevent off-grade products. Dow uses inferential sensors extensively for these purposes (26–28). The most common techniques used are PLS, multiple linear regression, artificial neural networks, support vector regression, and Gaussian process regression (29). Aside from challenges in nonlinearity, process drifts, and multiple modes, inferential sensors also need to consider the trade-off between model complexity and sensitivity. Through feature selection, model complexity can be minimized by selecting the relevant inputs to build a more robust model (27). A more comprehensive review of inferential sensor applications and future trends is available in References 29 and 30.

Advanced control and control loop monitoring also require process data on a much larger scale than the traditional hierarchical PID-based control structure. Lee & Lee (31) have provided an overview of recent advancements in advanced control, model predictive control, and plant-wide control. Plant-wide control loop performance monitoring approaches (32) are an important resource for assessing efficiency of plant operations. A controller performance assessment strategy that deals with 14,000 controllers coming from 40 plants located in 9 sites around the world is available (33). Starr et al. (34) provide another example of an ABB-developed industrial tool that monitors 600,000 control loops daily. Identifying control loops that have the most impact on the overall plant performance requires a significant amount of data and a hierarchical analysis process. Some examples that use this approach on the plant-wide scale can be found in References 35 and 36.

Advances in data variety. Outside of continuous processes, consistent and structured data sets are rare. In batch processes, scalar, vector, matrix, and sometimes tensor data can exist for each batch. The data quality is often poorer owing to different sampling rates, missing context, and limited instrumentation. In these scenarios, data-driven models require dimensionality reduction and feature extraction techniques to first establish a consistent structure before the classical approaches in continuous processes can be applied (37, 38). Unfolding, warping, and feature extraction are common techniques to preprocess the data prior to modeling (39, 40). For multistep or multistage batch processes, inputs can also be further grouped by phases or blocks to increase interpretability of the model results (41). Example applications in this area have used batch process data for fault detection of a commercial fed-batch fermentation system (42) and for product quality inferential sensors in a batch sulphite pulping process (43).

With the advancements in high-throughput experimental capabilities through laboratory automation, R&D in the chemical process industry has also benefited tremendously from the availability of large data sets (44). These large data sets usually contain data from high-throughput

imaging and online gas chromatography and spectroscopic analyzers. Using informatics methods, high-throughput experimentation greatly accelerates the pace of R&D in many areas. As an example, at Dow, high-throughput capabilities have been used to formulate new polyolefin catalysts (45), develop new carbon molecular sieves for gas separations (24), and improve formulation of drilling fluids additives (46).

In addition to instrumental data from the manufacturing environment, business data (order, sales, production, and customer demand) are also generated in huge quantities. Traditional hierarchies of enterprise resource planning usually decouple sales, supply chain, production, and process control into separate layers. With advances in numerical algorithms and faster computers, complex integrated scheduling and control problems that cross multiple layers of the enterprise resource planning hierarchy can now be solved in a reasonable time frame (47). In work by Nie et al. (48, 49), an integrated scheduling and control approach minimizes unproductive wait times between an upstream batch process and a downstream continuous process. The plant scheduling (future orders) data is included in the real-time optimization of downstream process units. The combination of scheduling, production forecast, and current operational data allows for smooth transitions without wait times. Information could also flow in the opposite direction, whereby uncertainties in production can be incorporated in the planning and scheduling problem via a stochastic formulation (50). Krumeich et al. (51) also propose integrating planning and control through big data approaches in the steel industry. Similar approaches to manage inventory and production can also be found in refinery operations (52–54). Maintenance data in terms of equipment failure rates and reliability data can also be used under an optimization framework to help optimize turnaround planning under uncertainty for large, complex chemical sites (55).

A common theme that emerges from surveyed works is that most of these applications are tied to fundamental engineering principles and physics. However, there is in addition a wealth of practical hands-on experience accumulated in the operational teams. As a result, many efforts have attempted to reconcile data and knowledge through gray-box modeling, expert systems, heuristic systems, and fuzzy logic systems. Chiang et al. (56) demonstrated that a causal map constructed from plant process flow diagrams could help in improving fault diagnosis performance. Reduced models can be created to optimize performance of units; for instance, Kumar et al. (57) used distributed temperature data to model the temperature distribution of a methane reforming furnace. The resulting data then led to a real-time optimization framework to help reduce temperature hotspots in the furnace.

Energy Industry

The use of data in the chemical process industry has demonstrated improvements in efficiency, reliability, and safety. The driving force in the energy industry is to fulfill energy consumption demands in a clean, low-cost, and sustainable way (58). Data-driven approaches are used to better estimate consumer demands, to optimize energy management, and to reduce environmental impact. On the energy supply side, optimizing electricity generation, anticipating plant outages, and predicting energy consumption are examples of the application of big data analytics. On the energy demand side, understanding consumer patterns reveals useful information to modify their behavior and, therefore, minimize consumption. This section focuses on how conventional and renewable generation, smart grids, and building energy management are benefiting from data-driven approaches.

Energy supply side. The main motivation of conventional generation is to generate clean electricity. This can be achieved by either retrofitting existing coal power plants or building new

natural gas combined cycle power plants. Policies such as the US Clean Power Plan (CPP) are inspiring efforts to reduce CO_2 generation (59). Big data analytics approaches are used to understand the impact of these policies. For instance, using monthly US electricity generation data from 2001 to 2014, the criteria to meet the specified CPP goals are tracked and the status of CO_2 emissions is evaluated. As a result, the impact of the CPP policy on CO_2 emissions can be quantified (59).

Renewable energy systems (including solar, wind, bioenergy, and hybrid) use energy models for energy demand forecasting, energy planning (to balance energy supply with demand), and electricity price estimation (60). To address the stochastic nature of renewable systems owing to meteorological conditions, data mining approaches have been used to better predict wind power, wind speed, and wind direction (61). These predictions can then be used to improve energy management.

To mitigate the intermittent generation of single renewable systems, hybrid renewable energy systems that combine wind, solar, and other energy generation and storage units (62) have been demonstrated to balance these fluctuations in power production. Large amounts of historical data are used to evaluate the impact on meeting electricity demand. For instance, the Wind Integration National Dataset toolkit combines meteorological, wind power production, and power forecast data sets for more than 126,000 locations across the United States for a seven-year period (2007–2013) (63). In a case study, Weitemeyer et al. (64) used hourly wind and solar power data for an eight-year period (2000–2007) in Germany; they found that storage capacity improved the ability to meet electricity demand by an additional 30% compared with the baseline case in which storage was not used.

Energy demand side. Big data analytics has been applied in smart grid management (65, 66), which uses forecasting (67, 68), real-time fault detection (69), load classification, and identification of energy consumption patterns. Smart grids facilitate information sharing among power generation, power transmission, power distribution, and demand management (68). Smart meters collect real-time data, such as device status and electricity consumption data (70, 71), at a higher resolution (i.e., 15- or 30-min intervals). Smart meters are capable of monitoring and controlling home appliances and can communicate with other meters, enabling customers to have more control over their energy use. Visualization plays an important role in identifying patterns and analyzing information from large amounts of data (72), where heat maps, 3D load graphs, and geographic information systems are useful for identifying issues in energy demand. For instance, the Mueller community in Austin, Texas, is participating in a smart grid demonstration project (70). The data collected in Austin are used to understand consumer patterns and then to improve energy efficiency in the community. Another example using a pilot smart grid evaluated the impact of appliances on energy consumption (73).

Weather is another important factor for consumer energy demands. Models to forecast electricity prices (67, 74) or household energy consumption patterns (73, 75) have benefited from incorporating weather data, resulting in superior performance (76).

Improving energy efficiency in buildings, either residential or commercial (73, 77, 78), is significant, as they consume approximately 40% of the total energy consumption in the world (79). Typically, smart buildings have a large amount of information available in real time, which can be used to identify the main energy consumers. The typical information available in commercial buildings includes (*a*) outside climate data; (*b*) temperature, humidity, and pressure inside the building, monitored for each room and floor; (*c*) physical building characteristics (such as construction materials or age); (*d*) number of occupants in each space of the building, which can be estimated by using timers and motion sensors; and (*e*) the overall energy consumed in the building from utilities and air-conditioning systems (80). When optimizing the energy consumed in the

building, occupant comfort level, typically defined by proper room temperature and lighting, is a key aspect that determines success.

In a case study applied to commercial buildings in Switzerland with high energy consumption (80), the energy consumption in the buildings is predicted by applying data-driven methods. This prediction is then used to understand consumption profiles to determine strategies to minimize them. Another example used three years of data to predict the steam load of heating, ventilating, and air-conditioning systems (78). Furthermore, residential appliance energy consumption data and weather data can be analyzed in tandem to arrive at optimal appliance operation strategies (73).

In addition to understanding consumption patterns, stochastic model predictive control has been demonstrated in simulation to account for weather uncertainties to ensure occupancy comfort while also reducing energy consumption (79). When designing buildings with renewable energy sources, machine learning approaches have been used to optimize design parameters of the building for thermal and visual comfort conditions (60).

Because of their exponential growth, the energy consumption in data centers has increased to 1.5% of total electricity consumption (81). Factors that impact data center energy consumption include computing resources (servers, storage devices, network hardware, and cooling systems) and physical resources (physical layout and facility location). Rong et al. (82) summarize multiple efforts to optimize a given data center's energy consumption. Goiri et al. (81) propose a system of data center workload management based on predicting the available amount of renewable energy and electricity prices.

Semiconductor Industry

Outside the chemical process industry, there is a similar need in semiconductor manufacturing to improve productivity and reduce cycle time and cost (83). The continuous reduction in the electronic component size of integrated circuits in a silicon wafer enables opportunities to improve performance. Another growing trend is the increase in the size of silicon wafer diameters from the current standard of 300 mm to 450 mm to produce more semiconductor devices. Advanced process control (APC) approaches play an important role in executing these manufacturing trends (84).

APC consists of the following technologies: (*a*) run-to-run control strategies (85), (*b*) fault detection and classification (FDC) systems (86), and (*c*) virtual metrology (VM) systems (87, 88). The enhancement of these APC systems (89) to satisfy semiconductor fabrication facility (fab) control (from the equipment level to overall product quality of the entire fab) is an example of the application of big data analytics in the semiconductor manufacturing industry. Additionally, there is an emerging trend shifting the current focus from analyzing fab data and developing offline models (90) to online solutions (91) for proactively correcting and improving the efficiency of the semiconductor fabs.

The measurements collected in a typical semiconductor fab (92) have different sampling periods that range from milliseconds to weeks. For instance, at the equipment level, temperature and voltage sensors are measured every second. As wafers are manufactured, they go through multiple processes, such as chemical mechanical polishing, lithography, etching, and chemical vapor deposition, among others. Quality properties, such as thickness and resistance, are measured periodically over a selected sample of wafers, with a typical sampling rate that ranges from hours to days depending on how the metrology stations are set up in the facility. The wafer acceptance test (either in-process or final) that ultimately provides information on the quality of the wafers manufactured can be stored daily to weekly (92). Challenges when working with semiconductor data sets involve addressing the variety and the volume characteristics.

The use of data to gain insights. Semiconductor manufacturing control involves the use of vast amounts of data, such as variables at the equipment level that reflect performance (83) or final electrical properties that reflect whether or not the wafer is defective. These data sets are transformed into insights in the following ways: (*a*) data mining for yield enhancement using neural networks, decision trees, and Bayesian networks (90, 93, 94) and (*b*) semiconductor fab-wide process understanding through modeling of the fab (a model for a material handling system can be found in Reference 95).

Real-time decision making. Multiple challenges arise when APC applications are used in real time, such as the need to improve query processing speeds and online model updating. Despite the current efforts to enhance existing storage systems, there is a need to promptly store and retrieve data so that it can be analyzed to generate corrective action in real time. An example of a Hadoop platform, illustrated in Reference 96, shows 4.5-times improvement for storing and querying in comparison with existing databases.

VM predicts a variety of low sampling rate measurements in real time by using higher sampling rate data collected in the previous process steps (87). FDC systems that are focused on monitoring the entire semiconductor fab in real time have been proposed (83, 97). Sophisticated FDC (86, 98, 99) systems that combine run-to-run controllers and VM are used for multiple semiconductor process units, such as the chemical mechanical polishing process (97).

Benefits of additional data sources and data quality improvements. The removal of noise from the data set is necessary to improve performance, accuracy, and prediction robustness. Approaches that use VM combined with FDC systems are capable of removing noise. An example applied to the photolithography process can be found in Reference 100.

Besides use of semiconductor fab data in APC applications, additional data sets are also being incorporated in analysis to improve performance. For instance, predictive maintenance, which estimates when equipment requires maintenance, reduces unscheduled downtime based on the usage, age, and performance of the equipment. Moyne et al. (96) propose a framework that integrates APC approaches and predictive maintenance. Macher et al. (101) provide another example of the use of different data sources in the evaluation of the impact of human resources on cycle time and yield enhancement.

Pharmaceutical Industry

The major themes in using data in the pharmaceutical industry focus on accelerating the pace of R&D and improving manufacturing quality of the end-products. The industry faces increasing pressure from rising costs of R&D and stagnant product pipelines. Competition from generic drugs and expectations from healthcare providers and consumers for lower-cost alternatives have driven a steady decline in the margins of existing drugs and products. Faced with these challenges, big data analytics is seen as a venue through which significant additional values could be realized. The McKinsey report on big data in pharmaceutical R&D estimates that big data–informed decision making could generate up to $100 billion in value across the US pharmaceutical industry alone (102).

Drug discovery from pharmaceutical data. Drug discovery has primarily been driven by the use of high-throughput screening techniques in the past. Although high-throughput screening has demonstrated considerable success in identifying effective compound-interaction pairings, the probability of success has seen diminishing returns in recent years (103). Computational drug

discovery and computer-aided drug design aim to reduce the search space by focusing on specific interactions using data-driven or mechanistic models. Kuhn et al. (104) gave a comprehensive review of advances in large-scale predictions of drug-target relationships using publicly accessible databases of drug-target interactions. Klipp et al. (105) focused on biochemical networks and pathways in which mathematical modeling can be used to aid in analysis of proteins for drug discovery. Alaimo et al. (106) developed a new network-based inference method that can predict drug-target interactions. The inference method is able to use existing domain-based knowledge on drug and target similarity. Cheng et al. (107) showed a similar application of network-based inference in their attempt to predict drug-target interactions specific to breast cancer cells. Koutsoukas et al. (108) reviewed the databases that are available currently for multitarget drug design, target predictions, and related applications. Lastly, Perlman et al. (109) applied logistic regression to use drug and gene similarity measures to make predictions of drug-target interactions.

Incorporating a variety of data in R&D. With the increased collaboration between health care providers, pharmaceutical corporations, and research institutions, a larger-than-ever variety of data sources are now accessible by R&D organizations. Using machine learning, researchers are attempting to combine and draw correlations across clinical, genetic, biochemical, and manufacturing data. Tothill et al. (110) demonstrated how to use k-means clustering to identify molecular subtypes of serous and endometrioid ovarian cancer that are linked to clinical outcomes. Ernst et al. (111) proposed an unsupervised machine learning method—a multivariate hidden Markov model that uses combinatorial patterns of chromatin marks to distinguish chromatin states of the human genome. Halpern et al. (112) developed an unsupervised learning framework that could predict clinical states from electronic health record data. They also showed that health records can be represented in a much lower dimension through the use of natural language processing techniques (113). Cheng & Gerstein (114) provided an example of a deep integrative approach for a regression task; they modeled the expression levels of genes in mouse embryonic stem cells from existing gene modifications.

Quality improvements in pharmaceutical manufacturing. Pharmaceutical manufacturing is another area in which data-driven approaches are used to improve the efficiency, reliability, and safety of manufacturing processes. This is particularly apparent in the Food and Drug Administration's Quality by Design initiative, which focuses on designing pharmaceutical manufacturing processes to be inherently robust and reliable (115). Robust and reliable processes require in-depth understanding not only from a fundamental level but also empirically on how the manufacturing data should behave in the production environment. Some work in this area reconciles existing knowledge with plant data. A first principles–based dynamic model is first estimated from the process data and then applied for control and optimization. For example, Boukouvala et al. (116) used a Kriging-based approach to identify the unknown parameters of a roller compactor process. The model was then used to optimize product quality and reduce defect rates. In cases where the exact structure of the model is not known, generative models from regression methods could be estimated and used for monitoring and local optimizations. Eberle et al. (117) showed that through iterative application of PLS and ANOVA analysis, the most frequent causes of yield loss could be identified and improved. To prevent drug product quality from being affected by undesired variability of incoming raw materials, Muteki et al. (118) showed that the effects of raw materials could be modeled using a latent-variable approach. A subsequent optimization can then eliminate such variabilities by strategically combining raw materials in later processing steps. Data from pharmaceutical processes are often heterogeneous (not in continuous time series) and highly

dimensional. To address these issues, Severson et al. (119) proposed using elastic net regression to impose regularization on the model inputs. The result is a sparse model that is more robust and easier to implement. The approach was demonstrated on an antibody purification process.

Food Industry

Chemical engineering has played an important role in developing sterilization technologies (such as pasteurization, food packaging systems, preservatives, and irradiation) and reaction processes (brewing and fermentation). Unit operation (distillation, mixing, fluid and solid transfer) knowledge has also enabled scale-up of food production to an industrial scale. In addition, process automation and control have enabled efficient, large-scale, and high-quality production of food products and consumables. In the big data era, advances in high-throughput experimentation, new sensors, data-driven modeling, numerical solvers, and optimization algorithms are enabling a new wave of computer-aided developments in the food industry.

The types of data encountered in food processing industries have been as diverse as the industry itself. For instance, there exist scientific data, such as genomic information on agricultural seeds and crops; engineering data in manufacturing plants and quality laboratories; and business data from suppliers, consumers, and the market. This section focuses on scientific and engineering data-related applications.

Big data analytics at the laboratory scale. Similar to R&D in the chemical process industry, laboratory-scale R&D in the food industry is using data to develop better formulations, improve cause-and-effect understanding, and speed up the product development cycle.

In food formulation, big data analytics techniques allow for extracting interactions and useful correlations from large data sets. For example, Ahn et al. (120) developed a bipartite network to model the relationship between 381 ingredients commonly used in food recipes throughout the world and 1,021 compounds that are known to introduce flavor in the known ingredients. The bipartite network can be used to quantitatively describe differences in flavor and nutritional content between regional cuisines. Pinel & Varshney (121) later used this network model of food and recipes to computationally generate new recipes that satisfy existing preferences in nutrition and taste.

Big data analytics also play a role in the omics fields that are increasingly used for assessment of raw materials and final products and for development of new processes in food technology. Techniques such as 2D electrophoresis, hyperspectral imaging, mass spectrometry, and various tailored chromatography methods have generated abundant high-dimensional data sets that require machine learning and multivariate data analysis to be effective. In References 122 and 123, the importance of using multivariate statistical methods rather than conventional univariate methods is highlighted for high dimensionality of electrophoresis and spectrometry data. There are many applications of proteomics to address issues in the food industry (124–126).

Big data analytics at the industrial scale. Advancements in Fourier transform infrared (FTIR) spectroscopy have led to the development of many simple and nondestructive testing methods for many chemical and physical components. This work, combined with the use of multivariate/chemometrics models, has led to the adoption of online analyzers in many chemical industries (127, 128). The increases in FTIR measurement speeds have led to an abundance of large spectra data that are rich in process information (129). FTIR in food sciences allows for faster, more accurate, and better detection of contamination, adulteration, and food expiration at a large scale

that was previously uneconomical to perform. These advancements in sensors have improved the scale and speed of food quality monitoring and safety detection.

In the area of food authentication, spectroscopy-based sensors can be combined with classification algorithms to identify the true origin of the food or product. For example, in a study by Cozzolino et al. (130), 200 types of red and white wines from 13 regions in Australia were analyzed using spectroscopy. PLS discriminant analysis models were used to classify their origins. In a related study (131), detection of honey contamination by added sugar solutions was formulated as a classification problem. Milk contamination with tetracycline could also be detected using a similar approach (132). Similarly, classification and regression methods have been applied in detecting lard adulteration in chocolate (133), vegetable oil adulteration in extra-virgin olive oil (134), and the presence of high-density lipoprotein in hydrogenated products (135). The review by Rodriguez-Saona & Allendorf (129) provides a more comprehensive list of spectroscopy-based detection and quantification applications.

With the increasing availability of inexpensive storage and computing resources, image data have become more ubiquitous. This has led to the application of image processing and machine learning techniques in quality control and evaluation in the food industries. In addition to optical images, image processing techniques can also be applied to data from charge-coupled device cameras, ultrasound, magnetic resonance imaging, near infrared imaging, and electrical tomography (136). In one study (137), an image recognition system was deployed on an apple conveyer system that was able to sort apples into different grades based on detection of surface defects. The preprocessed image goes into a neural network–based classifier that classifies whether the defect on the apple is real or just part of the stem. Similar applications of defect detection in other species of apples have also been reported (138). Computer vision has also been applied in estimating quality of meat products, such as beef tenderness, pork color, and fat percentage in lamb. References 139–141 provide a comprehensive overview of many additional applications of image classifier or regression models that can aid in quantifying quality measures in many types of food processing systems.

CHALLENGES

Technical Challenges

Recall that big data in conjunction with analytics is the key to turn data into insights for more informed business and operational decisions. This section summarizes technical challenges outlined by Reis et al. (11). The most challenging aspect of volume is that not all data are created equal; users need analytics skill sets to distinguish whether the data are meaningful. For information-poor data sets, users must filter the noise to enhance the signal. Another critical analytics skill set is to realize when information is missing in the data sets and design of experiments is needed to generate the right data.

In terms of variety challenges, the chemical engineering community collects data in the usual scalar quantities (such as temperature, pressure, flow, and concentration), one-way arrays (such as spectrum, chromatogram, and particle-size distribution curves), two-way arrays (such as an image and gas chromatography with mass spectrometry), three-way and higher-order arrays (such as video and hyperspectral images), and text data (such as email, operator log books, lab notebooks, and social media discussions). To make the situation more challenging, all these data are stored in various sources, from process historians to application and business databases, websites, email memos, and handwritten notes. Combing all these data sources to make meaningful conclusions using analytics is far from trivial.

In terms of velocity challenges, massive data are collected in real time with different time resolutions from milliseconds to hours, days, or even months. A first challenge is to select the right real-time resolution for the analytics applications of interest. Because most chemical processes are dynamic by nature, a second challenge is to use real-time data to adapt the existing models to incorporate new information and knowledge.

One of the most common critiques of the big data era is that spurious patterns and correlations outnumber genuine discoveries. This is often true when big data analytics is applied to chemical engineering problems without context and domain knowledge. Chemical processes are governed by first principles. In fundamental modeling approaches, domain knowledge is used to develop a model, often in a detailed, dynamic, and nonlinear form. For complex industrial processes, the cost, time, and skill required to develop such fundamental models can be high. Data-driven models can complement domain knowledge to generate insights, but there have been limited results reported in the literature (56, 142–144). Integrating fundamental modeling and process knowledge with big data analytics tools to create enterprise-scale solutions remains another technical challenge.

Platform Challenges

Lack of an appropriate software platform is an important barrier to overcome to implement and sustain big data analytics applications. These applications require multiple data sources to gather data and also require a fairly complex computational engine to execute their algorithms. Whereas proof-of-concept demonstrations in an offline environment are relatively easy, implementing a tool online in a robust way to ensure industrial reliability is often more challenging. As a result, industry as a whole must carefully evaluate the trade-off between acquiring standard generic software that poses limitations on deploying advanced algorithms and developing custom-made applications that could present maintenance challenges in the longer term.

Culture Challenges

A common theme that emerged from the surveyed papers in the five industries is that there are multiple pockets of successes. Every industry has its own unique advantages and disadvantages in applying big data analytics. Take the chemical process industry and the energy industry, for example. The chemical process industry has a strong tradition of using process data for process control and monitoring. Building on this foundation, recent advances in process control, real-time optimization, and integrated scheduling further push the limits on efficiency and reliability. However, the chemical process industry is slow to respond to real-time customer feedback compared with the energy industry. For its part, the energy industry has shown considerable advances in estimating real-time electricity demands and consumer behavior. And the supply side has yet to use the information uncovered in an efficient manner. There is an opportunity for industries to leverage off each other and to collaborate at a systematic level.

A more fundamental reason for the observed pockets of success is the lack of common driving force within and across industries to realize successes at the enterprise level. As an example, there is a lack of standard benchmark big data analytics problems available to compare published works even in the same research area. This manifests into multiple approaches being published that show no significant differences in performance—reinventing the wheel, so to speak. The Netflix Prize (145), for which a benchmark problem was published and could be solved by any interested participant, is a good example of such a common driving force. This created an incentive and an environment to systematically benchmark the contributions of innovation in this area. The closest example within the chemical process industry would be the Tennessee-Eastman problem (25); however, this problem is becoming outdated for the chemical engineering community today.

DIRECTIONS

The overarching theme of the big data era is that data volume will continue to grow exponentially. A variety of data will arrive, and at a high velocity. To achieve exponential growth in insights, both workforce development and analytics innovation are needed in the chemical engineering community.

Workforce Development

As the chemical engineering community is collecting more data (volume) from different sources (variety), it is increasingly more challenging to use the right data and tools (analytics) to make the right decisions in real time (velocity). This will require additional skill sets outside of traditional chemical engineering education. The data scientist is a new breed, and the fastest growing career of the twenty-first century (146). A data scientist is a professional with technical skills (such as programming, statistics, mathematics, and model building) and curiosity to make unexpected discoveries in the big data era. Both the International Data Corporation (147) and McKinsey & Co. (148) predicted that by 2018, there will be a shortage of 140,000 to 190,000 data science–related positions in the United States alone.

To meet the growing demand for data scientists, over 70 universities in the United States are offering one- to two-year master's degree programs in areas such as analytics, data science, data analytics, data engineering, predictive analytics, business analytics, and applied computational science (149). These degree programs provide necessary but insufficient training for graduates to address big data analytics opportunities in chemical engineering. An interdisciplinary skill set is needed and should include not only the big data analytics approaches but also a traditional chemical engineering education, involving unit operations, thermodynamics, reaction kinetics, transport phenomena, and process control. To address implementation opportunities at the practitioner level, data scientists should be trained in a five-year bachelor's/master's degree program in chemical engineering, with the fifth year focusing on big data analytics topics. To address the technical challenges in big data analytics outlined in the previous section, researcher-level training from a chemical engineering PhD degree program is needed. In addition, on-the-job training in big data analytics is recommended for experienced professionals.

Big Data Analytics Innovation

Technology readiness level (TRL) is a commonly used scientific term to discuss maturity in technology. To propel big data analytics in the chemical engineering community, industry, academia, and government must collaborate at all TRL levels from fundamental research to commercialization in the innovation chain shown in **Figure 2**.

Government Contribution to Innovation

Government is the centerpiece to advance big data analytics innovation. At TRL 4–6, in 2014 the US government established the National Network for Manufacturing Innovation (NNMI) initiative to foster collaboration between industry, academia, and government. Among the nine NNMI institutions that were established, big data analytics and workforce development are key themes in the Digital Manufacturing and Design Innovation Institute and Clean Energy Smart Manufacturing Innovation Institute. These government-sponsored institutes unite the chemical engineering community to develop and demonstrate big data analytics technology to solve platform and technical challenges.

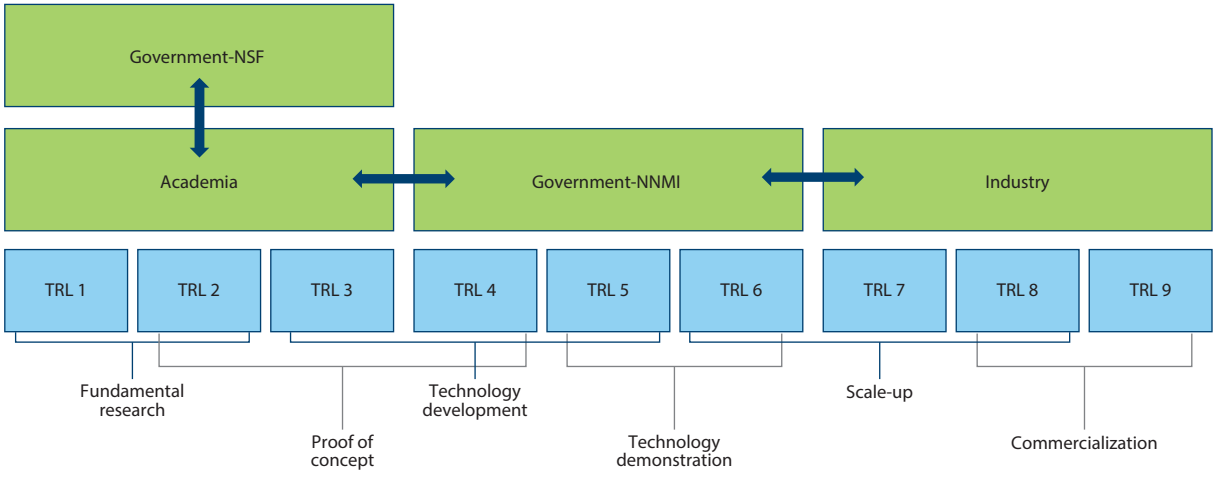

Figure 2

Big data analytics collaboration at all technology readiness levels (TRLs) between academia, industry, and government [National Science Foundation (NSF) and National Network for Manufacturing Innovation (NNMI)].

The US National Science Foundation (NSF) established the Big Data Science and Engineering funding program in 2012 to support fundamental research at TRL 1–3. As shown in **Figure 3**, NSF has ramped up funding to over $24 million to support 52 projects in 2016. The program seeks novel big data analytics approaches in computer science, statistics, computational science, and mathematics and innovative applications in social and behavioral sciences, geosciences, education, biology, physical sciences, and engineering (150). Out of the 174 funded proposals

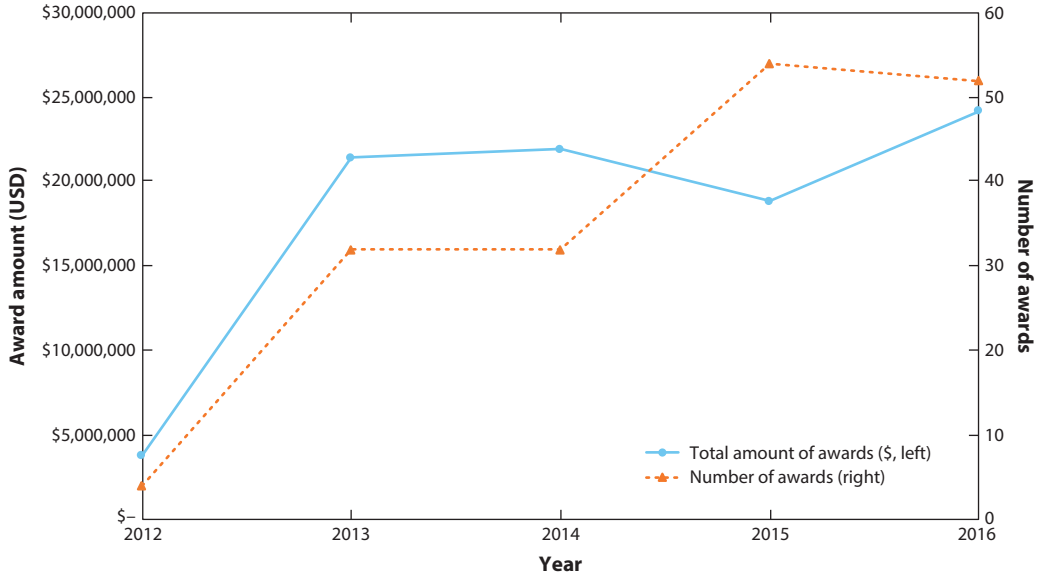

Figure 3

National Science Foundation Big Data Science and Engineering awards from 2012 to 2016.

from 2012 to 2016, none were from the chemical engineering community. Small pockets of big data–related research in chemical engineering are being supported through other NSF programs, such as cybermanufacturing systems under the Division of Civil, Mechanical and Manufacturing Innovation; computational and data-enabled science and engineering under the Division of Chemistry; or unsolicited proposals from the Division of Chemical, Bioengineering, Environmental and Transport Systems. Referring to the big data analytics culture challenge, NSF is in the position to provide a much-needed driving force and a common funding mechanism to advance big data analytics innovation germane to the chemical engineering community.

Academia Contribution to Innovation

The big data analytics technical challenge can be formulated into fundamental research problems in TRL 1–3. Chemical engineering faculty members in process systems engineering can expand their research focus to include PhD-level research programs in big data analytics. Industry is a good source for big data and funding. Big data analytics is also an active research area among the machine learning community in computer science departments. Multidisciplinary and industry collaborations are likely to bear fruits. PhD graduates from the interdisciplinary programs will become effective researchers to bring further impacts in the community. In the late 1990s and early 2000s, chemical engineering departments globally saw the need to change their names to highlight biochemical and biomolecular research; there is likewise a pressing need to integrate big data analytics research into these departments.

Industry Contribution to Innovation

It is a natural fit for industry to provide input to software vendors to develop sustainable big data analytics platforms at TRL 7–9. As discussed above, big data analytics shows pockets of success in R&D, manufacturing, supply chain, and maintenance in all major industry sectors. For companies that have already started the big data analytics journey and see successes in some functions, the next milestone is to establish a big data analytics culture to drive data-driven behavior across all functions. A big data analytics culture will nurture collaboration across all functions and provide a driving force for companies to take advantage of new and unarticulated enterprise opportunity. To meet these new business needs, companies must develop test beds to pilot innovation at TRL 4–6 and to provide feedback for academia and/or vendors to formulate research programs at TRL 1–3.

SUMMARY POINTS

1. Big data analytics is the journey to turn data into insights for more informed business and operational decisions. Based on this definition, the chemical engineering community started this journey decades ago using data-driven modeling approaches.
2. The chemical process industry has been a fertile ground for some of the foundational data-driven modeling work. Online instrumentation and centralized control systems allowed for aggregation and storage of process data for online monitoring and analysis. As a result, EMI, multivariate control, process monitoring, and inferential sensors have found early success in the industry. With recent advancements in both computer hardware and numerical algorithms, real-time model-based control, complex batch process monitoring, controller performance assessments, and integrated scheduling are recent examples that attempt to use data with more volume, variety, and velocity.

3. In the energy industry, smart meters, weather forecasts, and other data sources are improving forecasting capabilities to better understand energy consumption behaviors of large populations and local neighborhoods. Electricity suppliers are also tapping into these data sources to anticipate demand surges, detect faults and outages, and improve system reliability and efficiency.
4. The semiconductor industry is reducing costs through the use of existing data collected in automated systems. Real-time APC applications are attempting to identify faults, improve yields, and reduce product variability.
5. Pharmaceutical R&D is incorporating machine learning, dimensionality reduction, and visualization methods to analyze complex data sets such as gene expression, protein interactions, and drug discovery data. In addition, pharmaceutical manufacturing is improving quality through model-based control, real-time optimization, and real-time process monitoring.
6. In the food industry, diverse data sources are incorporated in modeling tastes, nutritional content, and recipe formulation for food products. The industry is shifting toward using multivariate methods to identify correlations from bigger data sets. In food manufacturing, spectrum analyzers and imaging sensors are used to detect product quality issues. Lastly, model-based control and optimization enable safer and more efficient processes.
7. The overarching theme of the big data era is that data volume will continue to grow exponentially. A variety of data will arrive at high velocity. Technical, platform, and culture challenges lie ahead in the chemical engineering community. To address the opportunities, academia, industry, and government must collaborate on workforce development and analytics innovation. Early adopters need to establish a culture to continuously explore new opportunities and to motivate the rest of the community to start the big data journey.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors would like to gratefully acknowledge feedback from Lloyd Colegrove, Mary Beth Seasholtz, Mike de Poortere, Billy Bardin, and JD Tate.

LITERATURE CITED

1. Podesta J, Pritzker P, Moniz EJ, Holdren J, Zients J. 2014. *Big data: seizing opportunities, preserving values*. White House Rep., Washington, DC
2. Marx V. 2013. Biology: the big challenges of big data. *Nature* 498(7453):255–60
3. Sci. Staff. 2011. Introduction: dealing with data: challenges and opportunities. *Science* 331(6018):692–93
4. Nat. Neurosci. Staff. 2014. Focus on big data. *Nat. Neurosci.* 17(11):1429
5. Martens H. 2015. Quantitative big data: where chemometrics can contribute. *J. Chemom.* 29(11):563–81
6. Shi Y. 2014. Editor's note: a global view of big data. *Bridge* 44:6–12
7. Qin SJ. 2014. Process data analytics in the era of big data. *AIChE J.* 60(9):3092–100

8. White D. 2016. Big data: What is it? *CEP Magazine*, March, pp. 33–35
9. García-Muñoz S, MacGregor JF. 2016. Big data: success stories in the process industries. *CEP Magazine*, March, pp. 36–40
10. Colegrove LF, Seasholtz MB, Khare C. 2016. Big data: getting started on the journey. *CEP Magazine*, March, pp. 41–45
11. Reis MS, Braatz RD, Chiang LH. 2016. Big data: challenges and future research directions. *CEP Magazine*, March, pp. 46–50
12. Laney D. 2001. 3D data management: controlling data volume, velocity, and variety. *META Delta*, Feb. 6. **http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf**
13. Wikipedia.org. 2016. *Big Data*. **https://en.wikipedia.org/wiki/Big_data**
14. Westerman A, Morawietz M, Geissbauer R, Vedso J, Schrauf S. 2016. *Industry 4.0: building the digital enterprise*. PWC Glob. Ind. 4.0 Surv. **https://www.pwc.com/gx/en/industries/industry-4.0.html**
15. Qin SJ. 2012. Survey on data-driven industrial process monitoring and diagnosis. *Annu. Rev. Control* 36:220–34
16. Colegrove L. 2015. Data initiative improves insights. *Chemical Processing*, March 12
17. Chiang LH, Russell EL, Braatz RD. 2012. *Fault Detection and Diagnosis in Industrial Systems*. London: Springer
18. Qin SJ. 2012. Survey on data-driven industrial process monitoring and diagnosis. *Annu. Rev. Control* 36(2):220–34
19. Severson K, Chaiwatanodom P, Braatz RD. 2015. Perspectives on process monitoring of industrial systems. *IFAC-PapersOnLine* 48(21):931–39
20. Yu J, Qin SJ. 2008. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE J.* 54(7):1811–29
21. Liu J. 2007. On-line soft sensor for polyethylene process with multiple production grades. *Control Eng. Pract.* 15(7):769–78
22. Liu Y, Zhang Z, Chen J. 2015. Ensemble local kernel learning for online prediction of distributed product outputs in chemical processes. *Chem. Eng. Sci.* 137:140–51
23. Jin H, Chen X, Yang J, Wu L. 2014. Adaptive soft sensor modeling framework based on just-in-time learning and kernel partial least squares regression for nonlinear multiphase batch processes. *Comput. Chem. Eng.* 71:77–93
24. Liu J, Han C, McAdon M, Goss J, Andrews K. 2015. High throughput development of one carbon molecular sieve for many gas separations. *Microporous Mesoporous Mater.* 206:207–16
25. Downs JJ, Vogel EF. 1993. A plant-wide industrial process control problem. *Comput. Chem. Eng.* 17(3):245–55
26. Chiang LH, Colegrove LF. 2007. Industrial implementation of on-line multivariate quality control. *Chemom. Intell. Lab. Syst.* 88(2):143–53
27. Lu B, Castillo I, Chiang L, Edgar TF. 2014. Industrial PLS model variable selection using moving window variable importance in projection. *Chemom. Intell. Lab. Syst.* 135:90–109
28. Kordon A, Chiang L, Stefanov Z, Castillo I. 2014. Consider robust inferential sensors. *Chemical Processing*, Oct. 2
29. Kadlec P, Gabrys B, Strandt S. 2009. Data-driven soft sensors in the process industry. *Comput. Chem. Eng.* 33(4):795–814
30. Kano M, Fujiwara K. 2013. Virtual sensing technology in process industries: trends and challenges revealed by recent industrial applications. *J. Chem. Eng. Jpn.* 46(1):1–17
31. Lee JH, Lee JM. 2014. Progress and challenges in control of chemical processes. *Annu. Rev. Chem. Biomol. Eng.* 5:383–404
32. Bauer M, Horch A, Xie L, Jelali M, Thornhill N. 2016. The current state of control loop performance monitoring—a survey of application in industry. *J. Process Control* 38:1–10
33. Paulonis MA, Cox JW. 2003. A practical approach for large-scale controller performance assessment, diagnosis, and improvement. *J. Process Control* 13(2):155–68
34. Starr KD, Petersen H, Bauer M. 2016. Control loop performance monitoring—ABB's experience over two decades. *IFAC-PapersOnLine* 49(7):526–32

35. Chioua M, Bauer M, Chen S-L, Schlake JC, Sand G, et al. 2016. Plant-wide root cause identification using plant key performance indicators (KPIs) with application to a paper machine. *Control Eng. Pract.* 49:149–58
36. Yuan T, Qin SJ. 2014. Root cause diagnosis of plant-wide oscillations using Granger causality. *J. Process Control* 24(2):450–59
37. Rato TJ, Rendall R, Gomes V, Chin S-T, Chiang LH, et al. 2016. A systematic methodology for comparing batch process monitoring methods: part I—assessing detection strength. *Ind. Eng. Chem. Res.* 55(18):5342–58
38. Chiang LH, Leardi R, Pell RJ, Seasholtz MB. 2006. Industrial experiences with multivariate statistical analysis of batch process data. *Chemom. Intell. Lab. Syst.* 81(2):109–19
39. Kourti T. 2003. Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions. *J. Chemom.* 17(1):93–109
40. Kassidas A, MacGregor JF, Taylor PA. 1998. Synchronization of batch trajectories using dynamic time warping. *AIChE J.* 44(4):864–75
41. Yao Y, Gao F. 2009. A survey on multistage/multiphase statistical modeling methods for batch processes. *Annu. Rev. Control* 33(2):172–83
42. Lennox B, Hiden HG, Montague GA, Kornfeld G, Goulding PR. 2000. Application of multivariate statistical process control to batch operations. *Comput. Chem. Eng.* 24(2–7):291–96
43. Rao M, Corbin J, Wang Q. 1993. Soft sensors for quality prediction in batch chemical pulping processes. *Proc. Int. Symp. Intell. Control*, pp. 150–55. New York: IEEE
44. Peil KP, Neithamer DR, Patrick DW, Wilson BE, Tucker CJ. 2004. Applications of high throughput research at the Dow Chemical Company. *Macromol. Rapid Commun.* 25(1):119–26
45. Boussie TR, Diamond GM, Goh C, Hall KA, LaPointe AM, et al. 2003. A fully integrated high-throughput screening methodology for the discovery of new polyolefin catalysts: discovery of a new class of high temperature single-site group (IV) copolymerization catalysts. *J. Am. Chem. Soc.* 125(14):4306–17
46. Mohler CE, Kuhlman RL, Witham CA, Poindexter MK. 2011. *Development of high-performance drilling fluids using high-throughput methods*. Presented at AADE Natl. Tech. Conf. Exhib., April 12–14, Houston, TX
47. Wassick JM, Agarwal A, Akiya N, Ferrio J, Bury S, You F. 2012. Addressing the operational challenges in the development, manufacture, and supply of advanced materials and performance products. *Comput. Chem. Eng.* 47:157–69
48. Nie Y, Biegler LT, Villa CM, Wassick JM. 2015. Discrete time formulation for the integration of scheduling and dynamic optimization. *Ind. Eng. Chem. Res.* 54(16):4303–15
49. Nie Y, Biegler LT, Wassick JM. 2012. Integrated scheduling and dynamic optimization of batch processes using state equipment networks. *AIChE J.* 58(11):3416–32
50. Chu Y, You F, Wassick JM, Agarwal A. 2015. Integrated planning and scheduling under production uncertainties: bi-level model formulation and hybrid solution method. *Comput. Chem. Eng.* 72:255–72
51. Krumeich J, Werth D, Loos P, Jacobi S. 2014. Advanced planning and control of manufacturing processes in steel industry through big data analytics case study and architecture proposal. *Proc. 2nd IEEE Int. Conf. Big Data, Oct. 27–30, Washington, DC*, pp. 16–24. New York: IEEE
52. Chang A-F, Pashikanti K, Liu YA. 2013. *Refinery Engineering: Integrated Process Modeling and Optimization*. Hoboken, NJ: John Wiley & Sons
53. Karuppiah R, Furman KC, Grossmann IE. 2008. Global optimization for scheduling refinery crude oil operations. *Comput. Chem. Eng.* 32(11):2745–66
54. Méndez CA, Grossmann IE, Harjunkoski I, Kaboré P. 2006. A simultaneous optimization approach for off-line blending and scheduling of oil-refinery operations. *Comput. Chem. Eng.* 30(4):614–34
55. Amaran S, Zhang T, Sahinidis NV, Sharda B, Bury SJ. 2016. Medium-term maintenance turnaround planning under uncertainty for integrated chemical sites. *Comput. Chem. Eng.* 84:422–33
56. Chiang LH, Jiang B, Zhu X, Huang D, Braatz RD. 2015. Diagnosis of multiple and unknown faults using the causal map and multivariate statistics. *J. Process Control* 28:27–39
57. Kumar A, Baldea M, Edgar TF. 2016. Real-time optimization of an industrial steam-methane reformer under distributed sensing. *Control Eng. Pract.* 54:140–53

58. Davis J, Edgar T, Graybill R, Korambath P, Schott B, et al. 2015. Smart manufacturing. *Annu. Rev. Chem. Biomol. Eng.* 6:141–60
59. Davis C, Bollinger LA, Dijkema GPJ. 2016. The state of the states: data-driven analysis of the US clean power plan. *Renew. Sustain. Energy Rev.* 60:631–52
60. Suganthi L, Iniyan S, Samuel AA. 2015. Applications of fuzzy logic in renewable energy systems—a review. *Renew. Sustain. Energy Rev.* 48:585–607
61. Colak I, Sagiroglu S, Yesilbudak M. 2012. Data mining and wind power prediction: a literature review. *Renew. Energy* 46:241–47
62. Wang X, Palazoglu A, El-Farra NH. 2015. Operational optimization and demand response of hybrid renewable energy systems. *Appl. Energy* 143:324–35
63. Draxl C, Clifton A, Hodge B-M, McCaa J. 2015. The Wind Integration National Dataset (WIND) Toolkit. *Appl. Energy* 151:355–66
64. Weitemeyer S, Kleinhans D, Vogt T, Agert C. 2015. Integration of renewable energy sources in future power systems: the role of storage. *Renew. Energy* 75:14–20
65. Zhou K, Fu C, Yang S. 2016. Big data driven smart energy management: from big data to big insights. *Renew. Sustain. Energy Rev.* 56:215–25
66. Diamantoulakis PD, Kapinas VM, Karagiannidis GK. 2015. Big data analytics for dynamic energy management in smart grids. *Big Data Res.* 2(3):94–101
67. Ludwig N, Feuerriegel S, Neumann D. 2015. Putting big data analytics to work: feature selection for forecasting electricity prices using the lasso and random forests. *J. Decis. Syst.* 24(1):19–36
68. Javed F, Arshad N, Wallin F, Vassileva I, Dahlquist E. 2012. Forecasting for demand response in smart grids: an analysis on use of anthropologic and structural data and short term multiple loads forecasting. *Appl. Energy* 96:150–60
69. Hare J, Shi X, Gupta S, Bazzi A. 2016. Fault diagnostics in smart micro-grids: a survey. *Renew. Sustain. Energy Rev.* 60:1114–24
70. Rhodes JD, Upshaw CR, Harris CB, Meehan CM, Walling DA, et al. 2014. Experimental and data collection methods for a large-scale smart grid deployment: methods and first results. *Energy* 65:462–71
71. Aslam W, Soban M, Akhtar F, Zaffar NA. 2015. Smart meters for industrial energy conservation and efficiency optimization in Pakistan: scope, technology and applications. *Renew. Sustain. Energy Rev.* 44:933–43
72. Hansen TM, Suryanarayanan S, Maciejewski AA, Siegel HJ, Modali AV. 2015. A visualization aid for demand response studies in the smart grid. *Electr. J.* 28(3):100–11
73. Kavousian A, Rajagopal R, Fischer M. 2015. Ranking appliance energy efficiency in households: utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings. *Energy Build.* 99:220–30
74. Weron R. 2014. Electricity price forecasting: a review of the state-of-the-art with a look into the future. *Int. J. Forecast.* 30(4):1030–81
75. Zhou K, Yang S. 2016. Understanding household energy consumption behavior: the contribution of energy big data analytics. *Renew. Sustain. Energy Rev.* 56:810–19
76. Leon-Garcia A. 2010. Price prediction in real-time electricity. *IEEE Trans. Smart Grid* 1(2):120–33
77. Chou J-S, Ngo N-T. 2016. Smart grid data analytics framework for increasing energy savings in residential buildings. *Autom. Constr.* 72(3):247–57
78. Kusiak A, Li M, Zhang Z. 2010. A data-driven approach for steam load prediction in buildings. *Appl. Energy* 87(3):925–33
79. Oldewurtel F, Parisio A, Jones CN, Gyalistras D, Gwerder M, et al. 2012. Use of model predictive control and weather forecasts for energy efficient building climate control. *Energy Build.* 45:15–27
80. Moreno MV, Dufour L, Skarmeta AF, Jara AJ, Genoud D, et al. 2016. Big data: the key to energy efficiency in smart buildings. *Soft Comput.* 20(5):1749–62
81. Goiri Í, Haque ME, Le K, Beauchea R, Nguyen TD, et al. 2015. Matching renewable energy supply and demand in green datacenters. *Ad Hoc Netw.* 25:520–34
82. Rong H, Zhang H, Xiao S, Li C, Hu C. 2016. Optimizing energy consumption for data centers. *Renew. Sustain. Energy Rev.* 58:674–91

83. Qin SJ, Cherry G, Good R, Wang J, Harrison CA. 2006. Semiconductor manufacturing process control and monitoring: a fab-wide framework. *J. Process Control* 16:179–91
84. May GS, Spanos CJ. 2006. *Fundamentals of Semiconductor Manufacturing and Process Control*. Hoboken, NJ: Wiley-Intersci. 305 pp.
85. Wang Y, Gao F, Doyle FJ. 2009. Survey on iterative learning control, repetitive control, and run-to-run control. *J. Process Control* 19(10):1589–600
86. Chien CF, Chuang SC. 2014. A framework for root cause detection of sub-batch processing system for semiconductor manufacturing big data analytics. *IEEE Trans. Semicond. Manuf.* 27(4):475–88
87. Lu B, Stuber J, Edgar TF. 2014. Integrated online virtual metrology and fault detection in plasma etch tools. *Ind. Eng. Chem. Res.* 53(13):5172–81
88. Khan AA, Moyne JR, Tilbury DM. 2008. Virtual metrology and feedback control for semiconductor manufacturing processes using recursive partial least squares. *J. Process Control* 18(10):961–74
89. Moyne J. 2004. Making the move to fab-wide apc. *Solid State Technol.* 47(9):47–52
90. Chien C-F, Wang W-C, Cheng J-C. 2007. Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Syst. Appl.* 33:192–98
91. Hsu SC, Chien CF. 2007. Hybrid data mining approach for pattern extraction from wafer bin map to improve yield in semiconductor manufacturing. *Int. J. Prod. Econ.* 107(1):88–103
92. Su AJ, Yu CC, Ogunnaike BA. 2008. On the interaction between measurement strategy and control performance in semiconductor manufacturing. *J. Process Control* 18(3–4):266–76
93. Chien CF, Hsu CY, Chen PN. 2013. Semiconductor fault detection and classification for yield enhancement and manufacturing intelligence. *Flex. Serv. Manuf. J.* 25(3):367–88
94. Moyne J, Schulze B. 2010. Yield management enhanced advanced process control system (YMeAPC)—part I: description and case study of feedback for optimized multiprocess control. *IEEE Trans. Semicond. Manuf.* 23(2):221–35
95. Kuo Y, Yang T, Peters BA, Chang I. 2007. Simulation metamodel development using uniform design and neural networks for automated material handling systems in semiconductor wafer fabrication. *Simul. Model. Pract. Theory* 15(8):1002–15
96. Moyne J, Samantaray J, Armacost M. 2016. Big data capabilities applied to semiconductor manufacturing advanced process control. *IEEE Trans. Semicond. Manuf.* 29(4):283–91
97. Tsuda T, Inoue S, Kayahara A, Imai S, Tanaka T, et al. 2015. Advanced semiconductor manufacturing using big data. *IEEE Trans. Semicond. Manuf.* 28(3):229–35
98. Jianbo Y. 2012. Semiconductor manufacturing process monitoring using Gaussian mixture model and Bayesian method with local and nonlocal information. *Semicond. Manuf. IEEE Trans.* 25(3):480–93
99. Fan S-KS, Chang Y-J. 2013. An integrated advanced process control framework using run-to-run control, virtual metrology and fault detection. *J. Process Control* 23(7):933–42
100. Kim D, Kang P, Lee S-k, Kang S, Doh S, Cho S. 2015. Improvement of virtual metrology performance by removing metrology noises in a training dataset. *Pattern Anal. Appl.* 18(1):173–89
101. Macher JT, Mowery DC. 2003. "Managing" learning by doing: an empirical study in semiconductor manufacturing. *J. Prod. Innov. Manag.* 20(5):391–410
102. Cattell J, Chilukuri S, Levy M. 2013. *How big data can revolutionize pharmaceutical R & D*. White Pap., McKinsey & Co., New York. **http://www.mckinsey.com/industries/pharmaceuticals-and-medical-products/our-insights/how-big-data-can-revolutionize-pharmaceutical-r-and-d**
103. Sliwoski G, Kothiwale S, Meiler J, Lowe EW. 2014. Computational methods in drug discovery. *Pharmacol. Rev.* 66(1):334–95
104. Kuhn M, Campillos M, González P, Jensen LJ, Bork P. 2008. Large-scale prediction of drug-target relationships. *FEBS Lett.* 582(8):1283–90
105. Klipp E, Wade RC, Kummer U. 2010. Biochemical network-based drug-target prediction. *Curr. Opin. Biotechnol.* 21(4):511–16
106. Alaimo S, Pulvirenti A, Giugno R, Ferro A. 2013. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 29(16):2004–8
107. Cheng F, Liu C, Jiang J, Lu W, Li W, et al. 2012. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLOS Comput. Biol.* 8(5):e1002503

108. Koutsoukas A, Simms B, Kirchmair J, Bond PJ, Whitmore AV, et al. 2011. From in silico target prediction to multi-target drug design: current databases, methods and applications. *J. Proteom.* 74:2554–74
109. Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R. 2011. Combining drug and gene similarity measures for drug-target elucidation. *J. Comput. Biol.* 18(2):133–45
110. Tothill RW, Tinker AV, George J, Brown R, Fox SB, et al. 2008. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* 14(16):5198–208
111. Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9(3):215–16
112. Halpern Y, Choi Y, Horng S, Sontag D. 2014. Using anchors to estimate clinical state without labeled data. *AMIA Annu. Symp. Proc.* 2014:606–15
113. Halpern Y, Horng S, Choi Y, Sontag D. 2016. Electronic medical record phenotyping using the anchor and learn framework. *J. Am. Med. Inform. Assoc.* 23:731–40
114. Cheng C, Gerstein M. 2012. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.* 40(2):553–68
115. Yu LX. 2008. Pharmaceutical quality by design: product and process development, understanding, and control. *Pharm. Res.* 25(4):781–91
116. Boukouvala F, Muzzio FJ, Ierapetritou MG. 2011. Dynamic data-driven modeling of pharmaceutical processes. *Ind. Eng. Chem. Res.* 50(11):6743–54
117. Eberle L, Sugiyama H, Papadokonstantakis S, Graser A, Schmidt R, Hungerbühler K. 2016. Data-driven tiered procedure for enhancing yield in drug product manufacturing. *Comput. Chem. Eng.* 87:82–94
118. Muteki K, Yamamoto K, Reid GL, Krishnan M. 2011. De-risking scale-up of a high shear wet granulation process using latent variable modeling and near-infrared spectroscopy. *J. Pharm. Innov.* 6(3):142–56
119. Severson K, VanAntwerp JG, Natarajan V, Antoniou C, Thömmes J, Braatz RD. 2015. Elastic net with Monte Carlo sampling for data-based modeling in biopharmaceutical manufacturing facilities. *Comput. Chem. Eng.* 80:30–36
120. Ahn Y-Y, Ahnert SE, Bagrow JP, Barabási A-L. 2011. Flavor network and the principles of food pairing. *Sci. Rep.* 1:196
121. Pinel F, Varshney LR. 2014. Computational creativity for culinary recipes. *Proc. Ext. Abstr. 32nd Annu. ACM Conf. Hum. Factors Comput. Syst.*, pp. 439–42. New York: Assoc. Comput. Mach.
122. Nordén B, Broberg P, Lindberg C, Plymoth A. 2005. Analysis and understanding of high-dimensionality data by means of multivariate data analysis. *Chem. Biodivers.* 2(11):1487–94
123. Karp NA, Spencer M, Lindsay H, O'Dell K, Lilley KS. 2005. Impact of replicate types on proteomic expression analysis. *J. Proteome Res.* 4(5):1867–71
124. Pedreschi R, Hertog M, Lilley KS, Nicolai B. 2010. Proteomics for the food industry: opportunities and challenges. *Crit. Rev. Food Sci. Nutr.* 50(7):680–92
125. Piras C, Roncada P, Rodrigues PM, Bonizzi L, Soggiu A. 2016. Proteomics in food: quality, safety, microbes, and allergens. *Proteomics* 16(5):799–815
126. Dajana Gaso-Soka, Spomenka Kova DJ. 2010. Application of proteomics in food technology and food biotechnology: process development, quality control and product safety. *Food Technol. Biotechnol.* 48(3):284–95
127. Griffiths PR. 2006. Introduction to vibrational spectroscopy. In *Handbook of Vibrational Spectroscopy*, ed. JM Chalmers, PR Griffiths. Hoboken, NJ: Wiley. 4000 pp.
128. Guillen MD, Cabo N. 1997. Infrared spectroscopy in the study of edible oils and fats. *J. Sci. Food Agric.* 75:1
129. Rodriguez-Saona LE, Allendorf ME. 2011. Use of FTIR for rapid authentication and detection of adulteration of food. *Annu. Rev. Food Sci. Technol.* 2(1):467–83
130. Cozzolino D, Holdstock M, Dambergs RG, Cynkar WU, Smith PA. 2009. Mid infrared spectroscopy and multivariate analysis: a tool to discriminate between organic and non-organic wines grown in Australia. *Food Chem.* 116(3):761–65
131. Sivakesava S, Irudayaraj J. 2002. Classification of simple and complex sugar adulterants in honey by mid-infrared spectroscopy. *Int. J. Food Sci. Technol.* 37(4):351–60

132. Sivakesava S, Irudayaraj J. 2002. Rapid determination of tetracycline in milk by FT-MIR and FT-NIR spectroscopy. *J. Dairy Sci.* 85(3):487–93
133. Che Man YB, Syahariza ZA, Mirghani MES, Jinap S, Bakar J. 2005. Analysis of potential lard adulteration in chocolate and chocolate products using Fourier transform infrared spectroscopy. *Food Chem*. 90:815–19
134. Gurdeniz G, Ozen B. 2009. Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data. *Food Chem*. 116(2):519–25
135. McCarthy J, Barr D, Sinclair A. 2008. Determination of trans fatty acid levels by FTIR in processed foods in Australia. *Asia Pac. J. Clin. Nutr.* 17(3):391–96
136. Du C-J, Sun D-W. 2004. Recent developments in the applications of image processing techniques for food quality evaluation. *Trends Food Sci. Technol.* 15(5):230–49
137. Li Q, Wang M, Gu W. 2002. Computer vision based system for apple surface defect detection. *Comput. Electron. Agric.* 36:215–23
138. Leemans V, Magein H, Destain MF. 1999. Defect segmentation on "jonagold" apples using colour vision and a Bayesian classification method. *Comput. Electron. Agric.* 23(1):43–53
139. Brøndum J, Egebo M, Agerskov C, Busk H. 1998. On-line pork carcass grading with the Autofom ultrasound system. *J. Anim. Sci.* 76(7):1859–68
140. Fernández C, Gallego L, Quintanilla A. 1997. Lamb fat thickness and longissimus muscle area measured by a computerized ultrasonic system. *Small Rumin. Res.* 26(3):277–82
141. Du CJ, Sun DW. 2006. Learning techniques used in computer vision for food quality evaluation: a review. *J. Food Eng.* 72(1):39–55
142. Kumar A, Baldea M, Edgar TF, Ezekoye OA. 2015. Smart manufacturing approach for efficient operation of industrial steam-methane reformers. *Ind. Eng. Chem. Res.* 54(16):4360–70
143. Boukouvala F, Muzzio FJ, Ierapetritou MG. 2010. Predictive modeling of pharmaceutical processes with missing and noisy data. *AIChE J.* 56(11):2860–72
144. Ng CW, Hussain MA. 2004. Hybrid neural network—prior knowledge model in temperature control of a semi-batch polymerization process. *Chem. Eng. Process. Process Intensif.* 43(4):559–70
145. Bennett J, Lanning S. 2007. *The Netflix Prize*. **https://www.cs.uic.edu/~liub/KDD-cup-2007/NetflixPrize-description.pdf**
146. Davenport TH, Patil DJ. 2012. Data scientist: the sexiest job of the 21st century. *Harv. Bus. Rev.* 90(10):71–76
147. Vesset D, Olofson CW, Nadkarni A, Zaidi A, McDonough B, et al. 2015. *IDC FutureScape: Worldwide big data and analytics 2016 predictions*. White Pap., IDC Research, Inc., Framingham, MA. **https://www.idc.com/getdoc.jsp?containerId=259835**
148. McKinsey & Co. 2011. *Big data: the next frontier for innovation, competition, and productivity*. Rep. 156, McKinsey Glob. Inst. **http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation**
149. Data-Informed.com. 2016. *Map of University Programs in Big Data Analytics*. **http://data-informed.com/bigdata_university_map/**
150. Natl. Sci. Found. *Critical techniques, technologies and methodologies for advancing foundations and applications of big data sciences and engineering (BIGDATA)*. Accessed on February 1, 2017. **http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504767**

Annual Review of Chemical and Biomolecular Engineering

Volume 8, 2017

Contents