

Advances and opportunities in machine learning for process data analytics

S. Joe Qin^{a,b,*}, Leo H. Chiang^c

^a Mork Family Department of Chemical Engineering and Materials Science and Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA, 90089, USA

^b The Chinese University of Hong Kong, Shenzhen, 2001 Longxiang Ave, Longgang, Shenzhen, Guangdong, 518172, China

^c Chemometrics and Data Analytics, Data Services, The Dow Chemical Company, 332 SH 332 E, Lake Jackson, TX, 77566, USA

ARTICLE INFO

Article history:

Received 26 February 2019

Revised 2 April 2019

Accepted 6 April 2019

Available online 24 April 2019

Keywords:

Process data analytics

Machine learning

Neural networks

Latent variable methods

Industrial operations

Industry 4.0

Artificial intelligence

ABSTRACT

In this paper we introduce the current thrust of development in machine learning and artificial intelligence, fueled by advances in statistical learning theory over the last 20 years and commercial successes by leading big data companies. Then we discuss the characteristics of process manufacturing systems and briefly review the data analytics research and development in the last three decades. We give three attributes for process data analytics to make machine learning techniques applicable in the process industries. Next we provide a perspective on the currently active topics in machine learning that could be opportunities for process data analytics research and development. Finally we address the importance of a data analytics culture. Issues discussed range from technology development to workforce education and from government initiatives to curriculum enhancement.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

The availability of massive data has prompted many industries to reposition themselves to take advantage of the disruptive potentials of data analytics and machine learning, including management science, medicine, economics, computer science, materials science, and engineering. As a result, a new discipline known as data science is taking a center stage that aims to derive knowledge and information from massive data. Even the discipline of statistics is rethinking the impact of machine learning on the century-old discipline, as expressed by [Hastie et al. \(2009\)](#), *The field of Statistics is constantly challenged by the problems that science and industry brings to its door...With the advent of computers and the information age, statistical problems have exploded both in size and complexity*. The tasks of extracting important patterns and trends, and understanding “what the data says” are known as *learning from data*, which has led to a revolution in the statistical sciences.

A research report by the McKinsey Global Institute (MGI) in 2011 ([Chui et al., 2011](#)) points out that analyzing large data sets will become a key basis of competitiveness, productivity growth, and innovations, which indicates various ways of using big data to

create values. Many predictions in the report have happened over the last eight years, and in some areas have accelerated due to the AlphaGo demonstration by Google's DeepMind ([Silver et al., 2016](#)).

Although there is no precise definition of big data, it usually refers to the size and variety of datasets that challenge the ability of traditional processing and analytical tools to capture, store, manage, and analyze. It is commonly accepted that big data problems possess a combination of the following four V's.

- Volume: Enterprises have ever-growing data of all types, easily amassing petabytes and even exabytes.
- Velocity: For time-sensitive processes such as catching frauds, the timeliness of a detection and action must be guaranteed.
- Variety: Big data is composed of multiple types of data, such as text, sensor data, audio, video, and operator's log files.
- Veracity: Establish trust in big data and conclusions based on them, which makes statistical learning theory an indispensable framework ([National Research Council, 2013](#)).

The demonstration of AlphaGo by Google's DeepMind in early 2016 was one of the most shocking events to the world. Go is an old and complex game invented in ancient China and is one of the most popular board games in East Asia. On March 19, 2016, the world Go champion Mr. Lee Sedol played against DeepMind's artificial intelligence (AI) program, the AlphaGo. The result was a clear win by the AI program with a score of 4 to 1. The AlphaGo

* Corresponding author.

E-mail address: [sqjin@usc.edu](mailto:sjqin@usc.edu) (S.J. Qin).

algorithm uses a combination of tree search techniques and machine learning on data from human and computer plays. It used massive historical records of the game by average players on Google's website over the years, along with historical matches by professional players.

The power of AlphaGo did not stop there. A year later a younger world champion Mr. Ke Jie was set to challenge AlphaGo. At this time AlphaGo was improved and renamed as the 'Master'. The result is that the Master beats the best human player in the world with a score of 3 to 0. It also played against five world champion players as a team and gave no chance for the human players to win. Lastly, it teamed up with a human player and played doubles against another AlphaGo machine teamed with another human player. At this moment people wondered if AI will soon out-smarts the mankind.

The DeepMind's innovation continued. In December 2017 the Founder and CEO Dr. Demis Hassabis gave a talk at the 2017 NIPS Conference with a title *Learning from First Principles* (Hassabis, 2017). The talk revealed a new version of AlphaGo, named AlphaGo Zero, where Zero stands for the fact that no historical data was used at all. All data were generated by computer simulations. It took a few days of computer vs. computer play to generate enough 'data' for AlphaGo Zero to be trained to beat the AlphaGo with a score of 100 to 0.

The demonstrations of AlphaGo made at least three intriguing points. One point is that the training data do not need to be of high quality, but they need to be massive and contain relevant information. The second point is that it is possible for machine learning to be trained well on simulated data. Thirdly, it can work in a team.

The Go game is considered much more complex than chess. A chess game has typically 10^{123} possible moves, while Go has about 10^{360} possible moves. However complex it is, the game of Go has little in common with the complexity of real world industrial operation and control problems. Go enjoys the following ideal situations.

1. The model and rules are clearly known, i.e., it is a white box without any uncertainty about the past and present. There is no need to perform state estimation, for instance.
2. Disturbances (i.e., the opponent's moves) are perfectly measured with no errors or measurement delays.
3. The objective function is perfectly known and never changes from one game to another. In fact, the whole system is time-invariant.

Because of the above ideal situations it is possible for AlphaGo Zero to be trained on simulated data only. The massive data made it possible to utilize the power of deep neural networks.

On the other hand, the Google's flu trend (GFT) prediction has the typical characters of a real world industrial data analytics problem (Ginsberg et al., 2009). Google took 50 million most common web searches and compared them to the Center for Disease Control (CDC) data on the spread of winter flu from 2003 to 2008. They screened through 150 million models to discover 45 features that had high correlation to the data from CDC. Google could predict nearly in real time, while CDC's data took two weeks to compile. While the massive data size is unmatched in chemical process problems, its functionality is essentially to predict critical but hard to measure variables from other easily measured ones, which is known as inferential sensors in process systems engineering (Qin and McAvoy, 1992; Tham et al., 1991).

However, unlike the success of AlphaGo, the GFT model failed terribly to predict the flu trend in the subsequent years, because it does not take into account of the time-varying and seasonal behavior of real-world problems. Several lessons have been learned from the failure of the original GFT model, as depicted in

Lazer et al. (2014). Later work by Yang et al. (2015) achieved accurate prediction results using autoregressive dynamic models inspired by hidden Markov models. Google has shared many possibilities of *hidden technical debt* or technical hindsight in deploying machine learning solutions (Sculley et al., 2015). One critical requirement is to succeed in the test of time.

AI applications today focus on partial intelligence, that is, to replace or enhance a specific aspect of human capabilities (Rich et al., 2009). Examples include: i) Self-driving cars; ii) face recognition for credit card transactions; iii) natural language processing (NLP); iv) multilingual translations; v) virtual and augmented reality; and vi) AI for Industry 4.0 or smart manufacturing. The focus of this paper is on process data analytics, that is, machine learning and AI techniques for the types of data and problems that arise from manufacturing processes.

The rest of the paper is organized as follows. Section 2 depicts the characteristics of process data and challenges to derive process intelligence from data. We give three attributes for machine learning techniques to be widely applicable in the process industries. Section 3 provides a retrospection of early machine learning and data analytics activities in the process systems engineering (PSE) community that are in line with the most recent advances in machine learning. Section 4 discusses the currently active topics in machine learning that could be opportunities for process data analytics. Section 5 addresses the importance of a data analytics culture. Issues discussed range from technology development to workforce education and from government initiatives to curriculum enhancement. The final section concludes the paper.

2. Process data analytics: Characteristics and attributes

2.1. Characteristics of process data

Process historical databases are massive due to the use of process control computers and manufacturing execution systems (MES). The variety of process measurement technology ranges from conventional process sensors to images, videos and other indirect measurement technologies such as spectroscopy and vibration measurements. In modern semiconductor chip manufacturing, for example, data come from the following levels at multi-scales (Qin et al., 2006).

- Level 1: Equipment measurement data, such as flow, pressure, and temperature;
- Level 2: Metrology data, such as film thickness and critical dimensions;
- Level 3: Indirect quality measurement data, such as optical emission spectroscopy;
- Level 4: In-line wafer electrical test data, an intermediate product quality measurement; and
- Level 5: Final wafer electrical test data.

It is common for modern fabs to have over 50,000 statistical process control (SPC) charts with each monitoring a specific quality variable.

A typical diagram for process control and data analytics schemes is depicted in Fig. 1. Process data often include process input, output and measured disturbances. These variables are often connected by the process flows and controller feedback, making them highly interactive and dynamic. Process data are measured at rather fast sampling rates, which can go from sub-seconds to minutes. Without additional quality or label data, one can apply principal component analysis (PCA) or other unsupervised learning techniques to analyze the process data. On the other hand, quality data and indirect quality data are often collected at a much lower sampling rate. These label data are often more critical to the manufacturing operation. Supervised learning techniques such as

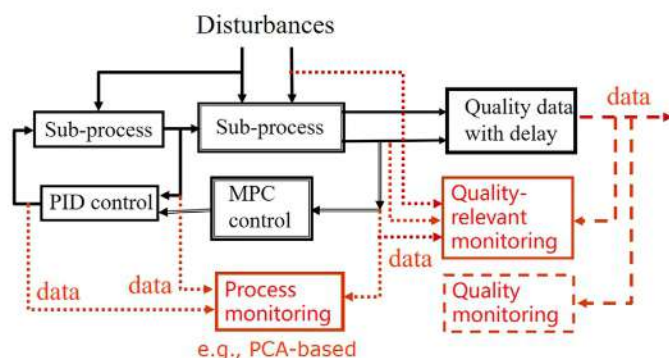


Fig. 1. Process and quality monitoring tasks considered in data-driven monitoring.

partial least squares (PLS), canonical correlation analysis (CCA), and neural networks can be applied to analyze process data and quality data jointly. Quality prediction with inferential sensors and quality-relevant fault diagnosis are examples of using supervised learning techniques.

It should be noted that, when computer control systems were conceived four decades ago (Williams, 1978), it was the control and operation functions that were the focuses of the design. Data collection and archiving were simply replacing strip chart paper recorders. The use of these recorded data has been varied and sparse. It is fortunate that the distributed control systems have gone through several generations of development in four decades, including Fieldbus for digital instrumentation and communication, independent data historians, along with three decades of data analytics practice. It is believed that process data analytics (PDA) will be a corner stone for Industry 4.0 where proactive data analytics are central for process operations and decision making. Infrastructure innovations are anticipated for the integrated control and data systems.

2.2. Unique attributes of process data analytics

What set process data analytics apart from other big data analytics, such as business intelligence? This question has to be answered in order to make an impact in practice. There are three important and unique attributes in order for machine learning to be widely applicable to process systems:

- Working compatibly with first principles models or process knowledge;
- dealing effectively with uncertainties that are usually time varying; and
- generating interpretable solutions.

First principles based on chemistry and physics are the foundation for how processes operate. Processes are designed with clear functions and principles, and very often, with sophisticated and reliable mathematical models. These models have been useful for optimization, simulations, and control. Having AI or machine learning models to replace them would be unwise if not impossible. AI and

machine learning should find its way to work with existing first principles.

Industrial process operations are like a coin with two sides. On one side it is the first principles models. Uncertainty is the other side of the coin. Unlike Go, where all the past state knowledge is perfectly known with no uncertainty at all, uncertainties are everywhere in industrial operations and control. Furthermore, the uncertainties are time-varying and dynamically changing in the mean and variance. For example, raw material compositions to a refinery process vary from time to time and are usually unknown until they result in product quality variations. As another example, catalyst deactivation happens over time. While it is difficult to model the deactivation process using first principles, its effects are reflected in the process data. Therefore, data contain information on the effect of raw material variations and the extent of catalyst deactivation, which is referred to as *situational knowledge*.

Roughly speaking, first principles can describe better the known part of the processes, whereas it is more efficient to use data to reveal the unknown part of the processes via machine learning. The dual characters of industrial processes make them different from games like Go, where everything in the past is clearly known and the system is time-invariant for the future. The differences between the problem dealt with in AlphaGo and the industrial process problems are depicted in Fig. 2. Industrial process operations and control have to deal with unknown disturbances with certain known principles, which is like the *Yin & Yang* (i.e., black vs. white or unknown vs. known) fish in the ancient Chinese philosophy. On the other hand, the AlphaGo problem has the opponent/disturbance perfectly known, which is a Yang-Yang fish, or symmetric. Therefore, process data analytics must learn from data the unknown state of disturbances to make the best possible decisions. With sufficient learning and tracking of unknown disturbances, it is possible to characterize and reduce the uncertainties.

The third attribute to make machine learning and analytics applicable in process industries is that they need to be interpretable or understandable by the human operation personnel. Unless the sophisticated industrial operations can be made completely autopilot, human operators and decision makers have to be in the loop and take over whenever the AI or machine learning system fails. If the people do not understand what the machine learning is doing, they will not take the risk to trust it. In a modern manufacturing environment there are usually a hierarchy of people with clear responsibilities at each level. They must be convinced that the new solutions offer more values than risks to be adopted.

While interpretation and prediction accuracy are two major purposes for machine learning, interpretation is a necessity for industrial process operations. Comparing first principles models with machine learning models, industrial personnel tends to trust more on first principles models than data driven models not because they are more accurate, but because they are interpretable. Therefore, sophisticated machine learning methods including deep learning and deep reinforcement learning have to pass the test on interpretability before they can be widely adopted in industrial practice. This was also the case when model predictive control (MPC) was initially applied in process industries, where model



Fig. 2. Industrial process operations must deal with unknown disturbances, which is like the *Yin & Yang* fish. The AlphaGo problem has the opponent/disturbance perfectly known, which is *Yang & Yang* fish.

predictions have to be observed and trusted by the operation personnel for a period of time before the controller is commissioned to closed-loop (Qin and Badgwell, 2003).

In summary, process data analytics should have the following desirable features (Qin, 2014):

- Interpretable and scalable (up to thousands of variables)
- Making use of all kinds of data (e.g., process data, spectra, vibration, and image data)
- Relatively easy to apply to real processes comparing to other methods
- Online use for real-time monitoring and decision making
- Offline troubleshooting as a valuable tool for continuous improvement.

3. Retrospection of machine learning and AI in PSE

Data analytics and monitoring tools such as latent variable methods (MacGregor and Kourti, 1995) and neural networks have been adopted by chemical manufacturers with varying degrees of success. These techniques can achieve detection and diagnosis of process faults and abnormal changes that can lead to more serious failure or incidents. The typical procedure of process monitoring and data analytics is as follows.

1. Collection of clean, normal data with a good coverage of the normal operation regions. Extensive data cleaning and preprocessing have to be performed in practice (Famili et al., 1997).
2. Faulty data can be useful for fault diagnosis, but are not required. If plenty of faulty data are available, fault classification methods such as linear discrimination analysis (LDA) are applicable (Chiang et al., 2000; 2001).
3. Build analytics models such as PCA and PLS for the normal data, and establish fault detection indices and control limits.
4. Perform fault diagnosis and troubleshooting.

Process monitoring and inferential sensors are two of the most industrially adopted data-driven applications in process industries (Cinar et al., 2007; Nomikos and MacGregor, 1994; Piovoso et al., 1992; Qin, 1996; 2012). Whether the original contributors were aware of it or not, the multivariate latent variable methods are inherently statistical machine learning methods (Oja, 1982; 1992; Sun et al., 2013). The boundaries between them have been largely removed due to the recent development of statistical machine learning; therefore, we take a pragmatic view in that they are all in the class of statistical learning methods (Hastie et al., 2009).

While the chemical engineering contributions in process monitoring and inferential sensors are well known, there are a number of less well-known chemical engineering contributions in machine learning methods and applications. These contributions, with some of them depicted in recent papers (Qin, 2014; Venkatasubramanian, 2019), are very much in line with the most recent and exciting machine learning advances, such as deep neural networks, nonlinear principal components, regularization methods, and trend analysis.

The first notable contribution from chemical engineering is the auto-associative neural networks which was aimed to perform nonlinear PCA (Kramer, 1991). This work coincidentally is similar to the building blocks of deep neural networks for dimension reduction and feature extraction (Bengio, 2009). It should be noted that prior work on linear auto-association (Bourlard and Kamp, 1988) is available and subsequent work shows that nonlinear auto-association is not equivalent to nonlinear PCA, but is more general (Japkowicz et al., 2000). Subsequently, nonlinear PCA using neural networks inspired by principal curves (Dong and McAvoy, 1996) and input training neural networks for nonlinear PCA (Tan and Mayrovouniotis, 1995) have been developed by chemical engineers in mid-1990s.

The next notable contribution from chemical engineering is the integration of neural networks with statistical methods to produce a strip net (Bhat et al., 1992) that removes zero weights after applying a regularization penalty. This approach is analogous to the recently developed sparse learning methods such as the Lasso, where an l_1 norm penalty is used instead (Zou and Hastie, 2005). The regularization penalty was also integrated with neural network training to handle collinear inputs (Qin, 1993). However, the PSE community scaled back the interest in machine learning and did not provide a statistical understanding as profound as the Lasso and sparse learning methods (Hastie et al., 2009).

Another notable contribution from PSE is the time series trend analysis by Bakshi et al. (1994). Time series pattern matching for disturbance mining was studied by Singhal and Seborg (2002). It is unfortunate that these efforts have not brought enough attention in process systems engineering. In the recent years time series data mining, segmentation, and pattern matching have been very active in creating simpler representations of time series that perform indexing, clustering, and classification (Fu, 2011; Keogh and Kasetty, 2003).

4. Opportunities in process data analytics

Thousands of papers have been published in recent years in process monitoring and inferential sensors based on latent variable methods, neural networks, and other machine learning or statistical methods. The past works are recently reviewed in Qin (2012), Ge et al. (2013), Chiang et al. (2017) and Venkatasubramanian (2019). Here we provide a brief assessment of the opportunities and some recent developments.

Although we consider multivariate latent variable methods as a subset of the statistical learning methods, the field of machine learning has grown tremendously in the last two decades. Recent advances include the development of support vector machines (SVM) proposed by Cortes and Vapnik (1995) and kernel methods (Hastie et al., 2009). Statistical machine learning becomes a major framework for artificial intelligence (National Research Council, 2013), which bridges computations and statistics, with ties to information theory, signal processing, control theory and optimization theory. Furthermore, the work of Hinton and Salakhutdinov (2006) is largely credited for the extremely high interest in deep learning.

All machine learning methods can be divided into supervised learning and unsupervised learning. The supervised learning uses input data X to predict or interpret target or label data Y , while unsupervised learning analyzes X without supervision from Y . A summary of many machine learning methods is given in Fig. 3. Depending on the characteristics of data and models, such as linear, collinear, or nonlinear data, there are respective supervised and unsupervised learning methods to choose from. To deal with collinearity in the data, for example, PCA is unsupervised learning and PLS is supervised learning. For categorical data such as modes of operations, grades of products, and classes of faults, the corresponding analytics are classification for supervised learning and clustering for unsupervised learning. Deep neural networks can be used for supervised learning or unsupervised learning, depending on the availability of target or label data. Reinforcement learning (Badgwell et al., 2018; Hoskins and Himmelblau, 1992; Sutton et al., 1992), with the goal to learn optimal actions based on data, also performs supervised learning.

4.1. Straightforward applications to process analytics

Given the plethora of supervised and unsupervised learning methods developed in the machine learning field, there are

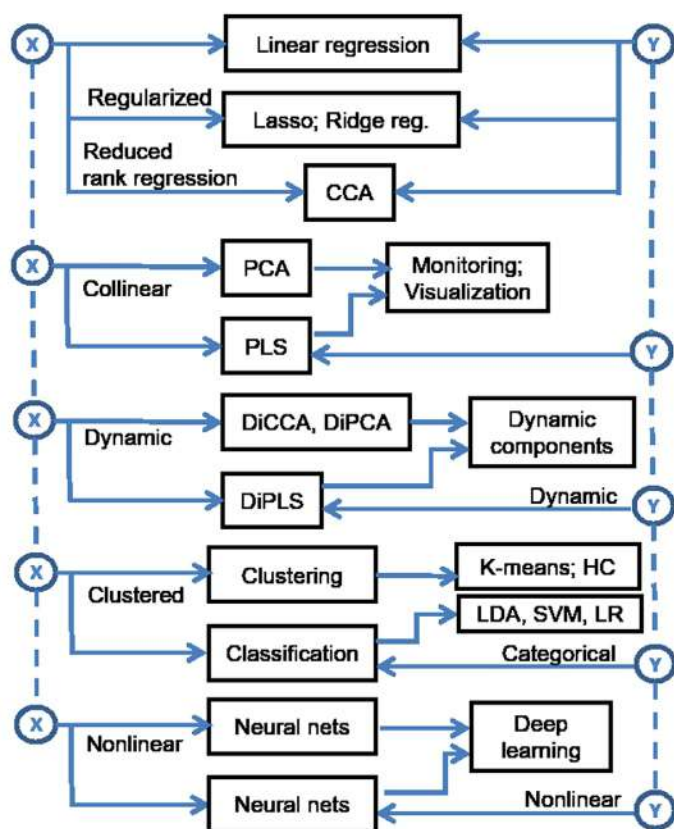


Fig. 3. A summary chart for supervised and unsupervised methods in statistical machine learning.

straightforward applications of these methods to process monitoring, fault diagnosis, and inferential sensors. For example, any supervised learning methods such as deep learning can be used as inferential sensors. Unsupervised learning methods, especially those which are nonlinear extensions of PCA, can be used for process monitoring and fault diagnosis. There are already numerous papers published along these lines.

For straightforward applications of machine learning methods to process monitoring, fault diagnosis, and inferential sensors, it is critical to demonstrate why the chosen method shows advantages for the specific application. On the other hand, it can be fruitful to work on new applications where few data analytics have been applied. In this direction there are still opportunities to show how machine learning methods should be applied to various types of use cases.

4.2. Integrating first principles with data

With the presence of sustained variabilities and abnormal situations in process operations, it is unrealistic to require that first principles models to include all possible scenarios of uncertainties. In these situations data provide situational knowledge of operations with the aid of proper data analytics. Therefore, model and data must interplay for intelligent operations and decision making, with models useful for expected situations and data useful for unexpected situations.

Referring to Fig. 2, data can be used to explore the 'dark' side of the process, or behaviors not included by the first-principles models. One possible scheme is to use neural networks or PCA to learn the residuals of the first principles models (Psichogios and Ungar, 1992; Wachs and Lewin, 1998). In principle, the well known Kalman filter is a systematic framework to use first principles

models for the deterministic subsystem and use data to model the residuals (Odelson et al., 2006).

4.3. Troubleshooting using data with possible anomalies

Process monitoring requires that a set of normal or 'good' data are available. Data analytics are used to provide a descriptive model for the normal situation. This requirement, however, is not always easy to fulfill. Usually one is faced with the following two situations: i) data contain normal records as the majority, but there are many outliers or faulty values mixed in them; and ii) the very task is to figure out if the current operation is normal or not, by troubleshooting the data to uncover undesirable patterns or variabilities.

When data are contaminated with outliers or bad data, the robust low-rank matrix approximation in developed machine learning is applicable (Liu et al., 2013). In theory and practice, it is shown that a data matrix can be expressed as a low-rank matrix plus a sparse matrix of outlying values. Robust PCA and matrix completion models are 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 by using the l_1 -norm, rather than the typical two-norm used in standard PCA.

To troubleshoot for undesirable behaviors or dynamics in the operation, Dong and Qin (2018a,b) developed dynamic data analytics that are efficient in extracting predictable dynamics in the data. These dynamic data analytics can efficiently extract oscillations and drifts in the data, since they are most predictable and easily visualizable. The dynamic components can be traced back to the contributing variables by analyzing the loadings. In addition, the dynamic analytics can reduce the uncertainty in the data, since a portion of the uncertainty is predictable by the dynamic component models.

4.4. Sparse learning methods

Inspired from shrinkage regression analysis such as ridge regression, a plethora of sparse learning methods have been developed after the invention of the Lasso (Zou and Hastie, 2005). The Lasso uses the l_1 norm to penalize the regression coefficients and therefore, achieves exactly zero coefficients for irrelevant variables. This result leads to a natural way for variable selection. The method has been extended to unsupervised learning, leading to sparse PCA algorithms (Shen et al., 2008; Zou et al., 2006). The ability to zero out irrelevant variables can provide better interpretation of the analytics models. Sparse PCA has also been applied to process monitoring to replace the standard PCA model (Gajjar et al., 2018).

4.5. Supervised monitoring and diagnosis

Most work in process monitoring are focused on PCA and unsupervised learning methods. Although PLS has been used for process monitoring, the quality-relevant diagnosis has only been recently developed (Li et al., 2010; Qin and Zheng, 2013; Zhu et al., 2017). Unlike diagnosis based on unsupervised learning models where one can at most analyze abnormal changes in process variables, quality-relevant diagnosis refers to the use of supervised learning models for the diagnosis of a quality anomaly due to relevant abnormal process variations. Since this approach aims to monitor and diagnose quality anomalies before they are measured and confirmed, Zhu and Qin refer to it as *supervised monitoring and diagnosis* (Zhu and Qin, 2019).

Supervised process monitoring has received increased attention since the early work on Total PLS and current PLS (Li et al.,

2010; 2009; Qin and Zheng, 2013). Liu et al. (2014) and Ge and Chen (2016) developed quality-relevant monitoring and diagnosis by taking into account of dynamics in the data. Huang and Yan (2017) proposed a two-block monitoring scheme based on mutual information (MI) and kernel PCA. Ma et al. (2018) put forth a new robust Gaussian mixture model for multi-mode processes. Qin and Zhao (2018) designed a monitoring method based on quality information and process dynamics for closed-loop manufacturing processes. Supervised monitoring and diagnosis deserves much more attention since it is predictive and can reduce quality-irrelevant alarms significantly.

4.6. Deep learning: Accuracy vs. interpretability

There are two essential aspects of data analytics, prediction and interpretation. To achieve accurate prediction, often complex models such as deep neural networks are more capable. On the other hand, in order to achieve interpretability, simpler models are preferred. As depicted in Section 2 of this paper, interpretability is critical for industrial adoption, since operators and decision makers must establish a level of trust in the analytics algorithms.

Deep learning has been applied to derive inferential sensors from data (Shang et al., 2014). Although the main goal for inferential sensors is prediction accuracy, it is necessary to provide a trade-off between model complexity and maintainability. The increased complexity of deep neural networks needs to be justified against the incremental accuracy gain comparing to other simpler models. The Google's GFT model for flu prediction is a classic example where the complex deep model fails the test of time. The alternative autoregressive dynamic models perform much better (Yang et al., 2015).

Deep neural networks have also been applied for image classification (Zhu et al., 2019) for a combustion furnace. This is directly applying image analysis methods to process operations.

In the field of image classification, researchers have recently questioned about what part of deep networks is effective in outperforming other methods. Deep networks are usually trained using a stochastic gradient descent (SGD) method. However, training such a network critically depends on the modeler's expertise in parameter tuning.

As a challenge or benchmark to deep neural networks, Chan et al. (2015) developed PCA-net, which uses PCA only to form a deep structure. The authors demonstrate that the PCA-net beats or meets the performance of deep neural networks on all major testbeds for image analysis and recognition.

To avoid the stochastic gradient training in deep convolutional neural networks, Kuo et al. (2018) propose an interpretable feed-forward design for the deep learning structure using PCA with added bias terms. The method is also tested on image classification benchmarks with favorable results. This work provides another evidence and idea to develop interpretable deep learning models.

5. Data analytics and culture changes in industries

Industry 4.0 refers to the fourth industrial revolution using cyber-physical systems to produce breakthrough improvement in manufacturing productivity, safety, and competitiveness. It is driven by a number of emerging technologies including data analytics, robotics, augmented reality, digital twins, internet of things, cybersecurity, and artificial intelligence. Of the numerous success stories reported across all industries (Bardin, 2017; Wikipedia.org, 2019a), process data analytics serves a common tool to improve the decision making process by getting the right data to the right people at the right time so that the right decision can be made. As a direct benefit of digital transformation in the Industry 4.0 journey, decision making processes are proactive instead of reactive.

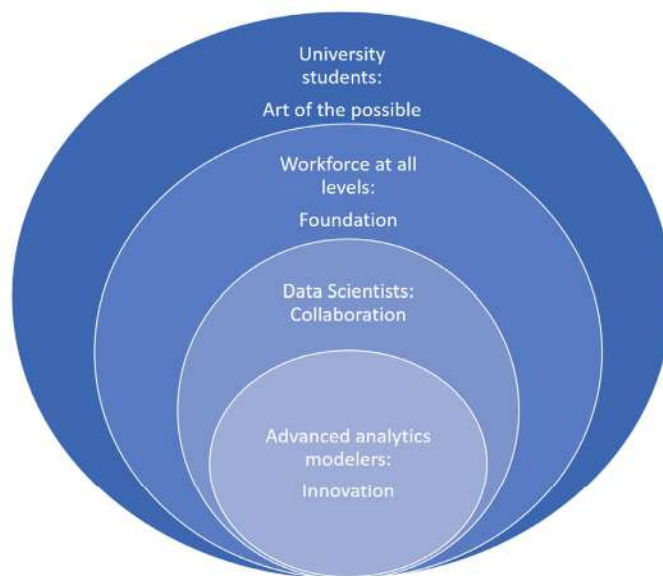


Fig. 4. Elements of a process data analytics culture.

For companies that have already started the process data analytics journey, as outlined in Colegrove et al. (2016), the next milestone is to scale the successes across all functions in the company by establishing a data analytics culture. Fig. 4 illustrates four elements of a process data analytics culture. Three elements (innovation, collaboration, and foundation) require company commitments at all levels while the fourth element (art of the possible) require breakthrough in workforce development by universities.

5.1. Innovation

Qin (2014), Reis et al. (2016), Chiang et al. (2017), and Venkatasubramanian (2019) provided comprehensive reviews of process data analytics and discussed a wide range of technical challenges. These articles reflected on a critique that spurious patterns and correlations outnumber genuine discoveries especially when process data analytics is applied to chemical engineering problems without context and domain knowledge.

Facing these technical challenges and critique, the process data analytics community is at a crossroad especially when breakthrough successes in machine learning and AI are routinely reported outside of the community (see Section 1 for examples). There is a consensus among these authors that innovation is needed to integrate data analytics tools with fundamental knowledge to create a robust and scalable solution for industrial processes. To drive innovation in process industries, a company must invest to create a critical mass of chemical engineers with technical skills in statistics, mathematics, modeling, optimization, process control, visualization, simulation, and programming. This group of formally trained and purposely built advanced analytics modelers shares common attributes of curiosity, creativity, and self-efficacy to translate business opportunities into projects that can be solved using the right selection of process data analytics techniques.

Of the machine learning and AI techniques mentioned in the preceding sections of this paper and in Venkatasubramanian (2019), it is clear that data analytics has been evolutionary in process industries. Much recent research work in applying machine learning to process problems is straightforward extensions of the multivariate statistical process applications over three decades (Cinar and Undey, 1999; MacGregor and Koutoudi, 1995; Qin, 2003). Recent work reported in process industries include deep learning in image classification (Rendall et al., 2018; Zhu et al.,

2019)), batch process applications (Wang et al., 2018), reinforcement learning in supply chain scheduling (Hubbs et al., 2018), and process control (Badgwell et al., 2018).

Advanced analytics modelers often face a platform challenge, as discussed in Chiang et al. (2017), to quickly test and deploy analytics models. Open source codes require a lot of customization for industrial use and vendor analytics solutions are often not ready for plug-and-play as advertised. This is where the digital office concept comes to life. Digital office, common in many industries, is designed to facilitate contributions from solution providers, software architects, and advanced analytics modelers to quickly test drive new ideas and deploy solutions.

As the process data analytics community is growing in size and impact, it is encouraging to see that traditional conferences such as Process Systems Engineering (PSE), Dynamics and Control of Process Systems (DYCOPS), Advanced Control of Chemical Processes (ADCHEM), and Advanced Control of Industrial Processes (ADCONIP) created new session topics around process data analytics. American Institute of Chemical Engineers (AIChE) took on additional efforts to revamp its spring meeting to include Big Data Analytics topical conference in 2015 and Industry 4.0 topical conference in 2018. Journal of Advanced Manufacturing and Processing is a new peer-reviewed journal launched in 2018 by AIChE to highlight cutting-edge manufacturing techniques, such as process data analytics, to enhance manufacturing (Aiche.org, 2019). The *Journal of Process Control* and an open journal *Processes* recently published special issues on process data analytics to bring together recent advances which can add to our understanding of the field (Mdpi.com, 2017; Qin, 2018). Computer Aids for Chemical Engineering (CACHE) sponsors a new conference, Foundations of Process Analytics and Machine learning (FOPAM), in 2019. These conferences and journals serve as important platform for advanced data analytics modelers to share latest results and exchange ideas, and importantly, to form new collaborations between industry, academia, and government.

While the advanced analytics modelers continue to have a bright career prospect in process industries, they also have a lot more alternative career opportunities, especially in machine learning and data science companies. The working environment and highly competitive salary at these companies are often more attractive to the young generation. For example, they do not need to argue whether data analytics will bring in values, since the revenues of the companies are generated mainly from big data and data analytics. This challenge will exist until the hirings at these companies reach a point of saturation.

5.2. Collaboration

As discussed in Section 3, there is a rich history of process data analytics applications based on process data collected in manufacturing processes. While manufacturing plants continue to collect even more data, other functions including R&D, supply chain, maintenance, process safety, sale & marketing, business finances, and human resources, also collect data for their domain specific needs. Where data are collected, there is a desire to generate actionable insights. Demand for data scientist roles in different functions has increased commensurate with this interest. It is perhaps not surprising to see that part of the workforce is willing to take on these new roles through leveraged technical skills and self-learning. Based on inherent networking, a community of practice is organically formed.

It is interesting to note the community of practice is typically an order of magnitude larger than the group of advanced analytics modelers in any company. While the smaller group of advanced analytics modelers serve an important role to provide technical mentorship to the data scientists in projects and leadership

to grow this informal network, the corporate executive leadership is in a prime position to provide a structure to foster collaborations among different functions and to provide resources to enrich data scientist analytics skill sets. The term “data acumen” has been coined to describe what technical skills are necessary for data scientists to be successful. These skills include general purpose or domain specific analytics software for basic analytics needs and computer programming for advanced analytics needs. There are a number of opportunities to grow this community. First and foremost, timely company-sponsored training is much needed and appreciated in this community. Other events such as company symposium, formatted data challenges, exchange or rotational programs offer a dual purpose of networking and technical problem solving.

The National Academies of Sciences, Engineering, and Medicine, in collaboration with the National Academies Board on Mathematical Sciences and Analytics, held a workshop to explore opportunities to use data science to transform chemical sciences and engineering (NASEM, 2018a; 2018b). There is a general consensus that it is much easier to train chemical engineers on data analytics topics rather than to train data scientists on chemical engineering topics. As process industries employ a large number of chemical engineers, this community of practice is a fertile hunting ground to send motivated chemical engineers to pursue an advanced degree in data analytics and related disciplines.

5.3. Foundation

As the advanced analytics modelers become tightly connected to the data scientists, the broader analytics community will be making bigger impacts to the company. At the foundation of the analytics culture, a new core competency, data literacy, will be required for the workforce at all levels in the company.

According to Wikipedia (Wikipedia.org, 2019b), data literacy is the ability to read, understand, create and communicate data as information. Much like literacy as a general concept, data literacy focuses on the competencies involved in working with data. A common goal of process data analytics is “getting the right data to the right people at the right time so that the right decision can be made.” In a typical use case, advanced analytics modelers or data scientists create analytics models to provide the right critical piece of data. The right people to make the right decision at the right time are typically plant operators and process engineers, not the modelers themselves. Another closely related term, data acuity, is coined to describe the need for general workforce to master basic statistics skills to interpret data and to ask the right questions to discern signal from noise (Colegrove, 2017). New behaviors are needed from both the modelers and the end users. While it is critical for the end users to master data acuity, it is more critical for the advanced analytics modelers and data scientists to earn their trust and credibility.

In the 90s, long before the term process data analytics was coined, the now ubiquitous Excel spreadsheet represented an innovation to a pocket calculator. Excel not only provides all the functions that a calculator can provide, but also additional functions (such as table, graph, and VBA programming language) that a calculator could not provide. More than two decades later, as we move towards the Industry 4.0 era, it is regretful to see that Excel is sometimes overly relied on as a primary tool at the base level for analytics, statistics, visualization, database, and knowledge retention. While it is clear that data scientists need to master general purpose and domain specific analytics software as well as computer programming, it is less clear as to what is the “Excel-alternative” analytics software for the general base-level workforce to master. This is a pressing question that remains unanswered today.

5.4. Art of the possible

Because industries need a robust pipeline of data analytics talent from universities, process data analytics culture cannot be sustained without breakthrough educational initiatives at universities. Data literacy/acuity should become a core skill for all students, not just science and engineering students. For example, Thomas Ewing, a History professor at Virginia Tech, incorporated data analytics elements in one of his senior level history classes. He used statistics and visualization to tell a story of his students' journey to understand and evaluate a monument to the City of Washington D.C.'s war dead from the first world war. He inspired students to appreciate the complex interactions and interpretation of human behavior as presented in numbers, statistics, and data (NASEM, 2018a).

For chemical engineering students to be successful in process industries, they need to go beyond data literacy/acuity to understand what the data means, understand how statistics and mathematics matter in context, master basic programming skill and algorithmic thinking, and, most importantly, how to use fundamental engineering principles and data to make sound decisions.

There is no one-size-fit-all solution to introduce data analytics education in chemical engineering curriculum. National Academies' report Data Science for Undergraduates: Opportunities and Options (NASEM, 2018a) is a solid reference point. The National Academies' data science workshop (NASEM, 2018b) offered additional insights on the question, *What training is needed to enable the future generation to increase and improve integration of data science into the chemical sciences and engineering*. A multi-pronged effort will be needed in three areas; namely domain specific data analytics training, integration of data analytics elements into existing classes, and general data analytics training.

Introducing a new data analytics course in chemical engineering is challenging, especially when the current curriculum at undergraduate and graduate level are intensive already. Understanding this burning platform, NSF took on a leadership role to provide a funding mechanism under the Innovations in Graduate Education (IGE) Program (Nsf.gov, 2019). The Department of Chemical Engineering and Materials Science at the University of Southern California has been offering a new graduate course in Process Data Analytics and Machine Learning for three consecutive years and made it a permanent course in 2019, where many new analytics and learning methods are introduced systematically, such as those discussed in Venkatasubramanian (2019). While some engineering departments are starting to explore the possibility to replace outdated course contents with new data analytics topics, there is still much progress to be made.

There are tremendous opportunities for chemical engineering departments to add data analytics elements in a number of existing classes. The CACHE website offers a helpful collection of instructional materials and ideas to consider. As compared to a new domain specific data analytics training development, this route has a lower barrier to success.

Data science/analytics institutes are forming in many university campuses. These institutes provide formal training in foundational data analytics concepts and exposure to emerging analytics topics. They also provide informal networking for students to engage in data analytics activities such as data challenge competition. As stated in the preceding sections of the paper, machine learning, AI, and process data analytics communities are rapidly developing breakthrough algorithms. However, the adoption of these emerging techniques are lagging in academic research and process industries. These institutes serve an important role to expose these techniques to students and researchers early in their career.

To summarize, the entire chemical engineering community has an obligation to advocate for changes in data analytics education today. Using process control in the chemical engineering curricu-

lum as an analogy, chemical engineering students have not been taught the 'PID control' equivalent of data analytics. Industry must communicate data analytics gaps observed in workforce and how this is impacting U.S. scientific leadership and manufacturing competitiveness. Academia must embrace changes to develop a multi-pronged approach to providing data analytics education to all students. Government must provide meaningful funding to grow data analytics programs in universities.

6. Conclusions

Artificial intelligence and machine learning are seen as cornerstones for the Industry 4.0 and smart manufacturing. To make these techniques widely applicable in the process industries, they need to have certain attributes in terms of working with first principles knowledge, dealing with uncertainties and lack of critical data, and being interpretable. The characteristics of process data are unique in that the processes are designed with engineering principles but operated with various kinds of uncertainties and disturbances. While there are massive process data available, critical quality or label data are sparse and expensive.

The current advances in machine learning and AI provide opportunities for process data analytics, but it requires innovative research to make them widely applicable and easily maintainable in the process industries. To promote the technology transition to Industry 4.0, academic institutions, industrial players, and government agencies must work collaboratively. It takes not only technology innovations, but also workforce education and curriculum enhancement.

References

- Aiche.org. 2019. <https://www.aiche.org/publications/journals/journal-of-advanced-manufacturing-and-processing>.
- Badgwell, T.A., Lee, J.H., Liu, K.H., 2018. Reinforcement learning overview of recent progress and implications for process control. In: Eden, M.R., Ierapetritou, M.G., Towler, G.P. (Eds.), 13th international symposium on process systems engineering (PSE 2018), vol. 44 of computer aided chemical engineering. Elsevier, pp. 71–85.
- Bakshi, B., Stephanopoulos, G., 303–332, 1994. Representation of process trends IV. induction of real-time patterns from operating data for diagnosis and supervisory control. *Comput. Chem. Eng.* 18 (4).
- Bardin, B., 2017. Industry 4.0 challenge and opportunity for the chemical industry. San Antonio, TX, 2017. AIChE Spring Meeting.
- Bengio, Y., 2009. Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2 (1), 1–127.
- Bhat, N., McAvoy, T., 271–281, 1992. Determining model structure for neural models by network stripping. *Comput. Chem. Eng.* 16 (4).
- Bourlard, H., Kamp, Y., 1988. Auto-association by multilayer perceptrons and singular value decomposition. *Biol. Cybern.* 59 (4), 291–294.
- Chan, T.H., Jia, K., Gao, S., Lu, J., Zeng, Z., Ma, Y., 2015. Pcanet: a simple deep learning baseline for image classification? *IEEE Trans. Image Process.* 24, 5017–5032.
- Chiang, L., Lu, B., Castillo, I., 2017. Big data analytics in chemical engineering. *Annu. Rev. Chem. Biomolecular Eng.* 8, 411–423.
- Chiang, L., Russell, E., Braatz, R., 2000. Fault diagnosis and fisher discriminant analysis, discriminant partial least squares, and principal component analysis. *Chemom. Intell. Lab. Syst.* 50, 243–252.
- Chiang, L., Russell, E., Braatz, R., 2001. Fault detection and diagnosis in industrial systems. *Advanced textbooks in control and signal processing*. Springer-Verlag, London, Great Britain.
- Chui, M.J.M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A., 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, Seattle Tech. rep.
- Cinar, A., Palazoglu, A., Kayihan, F., 2007. Chemical Process Performance Evaluation. Taylor & Francis CRC Press, Boca Raton, FL.
- Cinar, A., Undey, C., 1999. Statistical process and controller performance monitoring. In: a tutorial on current methods and future directions, in: American control conf., san diego, pp. 2625–2639.
- Colegrove, L., 2017. Big data analytics skills a revolution lacking in revolutionaries. AIChE southwest process technology conference, galveston, TX.
- Colegrove, L., Seasholtz, M., Khare, C., 2016. Big data - getting started on the journey. *Chem. Eng. Prog.* 41–45.
- Cortes, C., Vapnik, V., 1995. Support-vector networks. *Mach. Learn.* 20 (3), 273–297.
- Council, N.R., 2013. Frontiers in Massive Data Analysis. The National Academies Press, Washington, DC doi:10.17226/18374.
- Dong, D., McAvoy, T., 1996. Nonlinear principal component analysis based on principal curves and neural networks. *Comput. Chem. Eng.* 20, 65–78.

Iniciado em terça-feira, 5 mar. 2024, 19:06

Estado Finalizada

Concluída em terça-feira, 5 mar. 2024, 19:06

**Tempo
empregado** 12 segundos

Questão 1

Não respondido

Vale 1,00 ponto(s).

v8 (mais recente)

Qual das alternativas apresenta o tema central do artigo de Qin e Chiang (2019)?

Escolha uma opção:

- ☐ A ciência de dados que busca extrair informação e conhecimento de uma quantidade massiva de dados.
- ☐ A análise de grandes conjuntos de dados que é a base para a competitividade, crescimento de produtividade e inovação.
- ☐ A definição de *big data* que inclui uma combinação dos seguintes elementos: volume, velocidade, variedade e veracidade.
- ☐ A análise de dados de processo que considera *machine learning* e inteligência artificial para os tipos de dados e problemas que surgem a partir de processos de manufatura.

Questão 2

Não respondido

Vale 1,25 ponto(s).

v8 (mais recente)



Qin e Chiang (2019) apresentam características dos dados de processo, desafios para obter inteligência a partir dos dados e atributos para aplicação de técnicas de machine learning nos setores de processo. Nesse enfoque, analise as assertivas a seguir:

- (1) A inteligência artificial e *machine learning* podem substituir de forma muito inteligente os princípios físicos, químicos e matemáticos que regem processos industriais.
- (2) O uso de dados, por meio de *machine learning*, pode ser eficiente para revelar a parte desconhecida dos processos industriais, como variações na composição da matéria-prima.
- (3) *Machine learning* e análise de dados devem ser interpretáveis pela equipe que opera o processo nas indústrias, para eles possam ser convencidos de que as novas soluções oferecem mais valores do que risco.
- (4) Há três características importantes para que *machine learning* possa ser aplicado a sistemas de processos: machine learning precisa atuar de forma compatível com princípios fundamentais dos processos, *machine learning* deve lidar com incertezas de forma efetiva sem considerar variações no tempo, e *machine learning* deve gerar soluções interpretáveis.

Selecione a alternativa que apresenta a avaliação correta das afirmativas anteriores em Verdadeiras (V) ou Falsas (F):

Escolha uma opção:

- ☐ F – V – F – V
- ☐ V – F – V – V
- ☐ F – V – V – F
- ☐ V – V – F – F

Questão 3

Não respondido

Vale 1,75 ponto(s).

v13 (mais recente)

Assinale a alternativa **correta**:

Escolha uma opção:

- ☐ Os usuários finais das análises de dados precisam de uma matriz sobre a acuidade dos dados e os cientistas de dados e desenvolvedores de modelos avançados de análise precisam ter credibilidade.
- ☐ Monitoramento e diagnósticos supervisionados são preditivos e podem reduzir, de forma significativa, os alarmes de qualidade irrelevantes.
- ☐ Há quatro elementos da cultura de análise de dados de processos que são dependentes do compromisso da empresa/indústria em todos os seus níveis.
- ☐ A análise de dados tem dois aspectos essenciais: a previsão e a interpretação, sendo que a primeira é geralmente conseguida por meio de modelos mais simples e a segunda, por meio de modelos mais complexos.

Questão 4

Não respondido

Vale 3,00 ponto(s).

v7 (mais recente)

Os autores Qin e Chiang comparam dois contextos de *machine learning*, utilizando os símbolos Yin-Yang e Yang-Yang. Explique os dois contextos e as razões apresentadas pelos autores para sustentarem tal comparação.



Questão 5

Não respondido

Vale 3,00 ponto(s).

v7 (mais recente)

Ao discutirem a cultura da análise de dados, os autores comentam sobre o elemento 'arte do possível'.

- Defina o elemento 'arte do possível' e discorra sobre **uma** das três áreas em que o referido elemento pode ser desenvolvido no currículo dos cursos de engenharia química, de acordo com o texto.
- Observando a Figura 4, selecione **um** dos outros três elementos da cultura da análise de dados e explique como ele se relaciona com o elemento 'arte do possível'.

