DTU Library

The Role of Big Data in Industrial (Bio)Chemical Process Operations

A. Udugama, Isuru; Gargalo, Carina L.; Yamashita, Yoshiyuki; Taube, Michael A.; Palazoglu, Ahmet; Young, Brent R.; Gernaey, Krist V.; Kulahci, Murat; Bayer, Christoph

Published in:
Industrial and Engineering Chemistry Research

Link to article, DOI:
10.1021/acs.iecr.0c01872

Publication date:
2020

Document Version
Peer reviewed version

The Role of Big Data in Industrial (Bio)Chemical Process Operations

Isuru A. Udugama[1], *Carina L. Gargalo*[1], *Yoshiyuki Yamashita*[2], *Michael A. Taube*[3], *Ahmet Palazoglu*[4], *Brent R. Young*[5], *Krist V. Gernaey*[1], *Murat Kulahci*[6,7], *Christoph Bayer*[8,*]

1 Process and Systems Engineering Center (PROSYS), Department of Chemical and Biochemical Engineering, Technical University of Denmark, Kgs. Lyngby, 2800, Denmark

2 Department of Chemical Engineering, Tokyo University of Agriculture and Technology, Koganei, Tokyo, 184-9599, Japan

3 S&D CONSULTING, INC., Houston, TX 77006, USA

4 Department of Chemical Engineering, University of California, Davis, CA 95616, USA

5 Industrial Information and Control Centre, Department of Chemical & Materials Engineering, The University of Auckland, Auckland, 1010, New Zealand

6 Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kgs. Lyngby, 2800, Denmark

7 Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, 97817, Sweden

8 Department of Process Engineering, TH Nürnberg, Nürnberg, 90489, Germany

* Corresponding Author: Christoph Bayer (email: christoph.bayer@th-nuernberg.de)

Keywords: Big Data applications, Chemical engineering, Closed-loop systems, Control engineering, Production management, Statistical analysis

Abstract

With the emergence of Industry 4.0 and Big Data initiatives there is a renewed interest in leveraging the vast amounts of data collected in (bio)chemical processes to improve their operations. The objective of this manuscript is to provide a perspective of the current status of Big Data-based process control methodologies and the most effective path to further embed these methodologies in the control of (bio)chemical processes. Therefore, this manuscript provides an overview of operational requirements, the availability and the nature of data, and the role of the control structure hierarchy in (bio)chemical processes and how they constrain this endeavor. The current state of the seemingly competing methodologies of Statistical Process Monitoring and (Engineering) Process Control is examined together with hybrid methodologies that are attempting to combine tools and techniques that belong to either camp. The technical and economic considerations of a deeper integration between the two approaches is then explored and a path forward is proposed.

1. Introduction

Recent breakthroughs in Big Data-based algorithms and analytics have enabled the solution of complex and multi-faceted real-time problems at an extraordinary level of performance. A recent “poster child” for this transformation is Google DeepMind, which is able to defeat an expert player in the two-player abstract strategy game Go [1]. Similar examples can also be found in video games where Big Data-based algorithms are now capable of excelling at complex multi-level games [2]. A clear comparison can be drawn between the level of reasoning and the nature of these video games and some of the operational requirements of (bio)chemical plants. For example, the objective in plant operation is to react to the current “state” of the plant with the aim of delivering safe operations (similar to surviving in a video game) and improving the overall economics (similar to reaching the next level). However, there are some key difficulties that differentiate a video game from (bio)chemical operations.

First of all, even in a popular video game such as Angry Birds, the large number of possible actions and the need to plan beyond the current move has caused the Big Data-driven Artificial Intelligence (AI) to fail in matching a good human performance [3]. This is despite the fact that over 60 expert teams used multiple combinations of Big Data-based concepts such as advanced simulation, reasoning, planning, heuristic search and various Machine Learning (ML) approaches (including deep learning).

Secondly, most of these algorithms are inherently “data hungry”, which is not a problem in a video game environment. The game engine can be used to generate millions of scenarios, which is infeasible in a (bio)chemical plant. Although, “digital models” of (bio)chemical plants could be deployed to generate large amounts of process data [4] for the implementation of "data hungry" algorithms.

Thirdly, unlike in a video game, the Big Data-based algorithms must not trip production facilities during operation, which means that they must ensure near 100% safety, while delivering optimal overall performance. This duality is one of the grand challenges in the deployment of AI [3] applications.

Even though such methods and algorithms cannot simply be reapplied to (bio)chemical operations, there is promise in Big Data-based concepts, ranging from clustering methods to regression methods and data processing techniques to the different ML methods. The recent rapid advances in the field of Big Data has created momentum in the chemical and biochemical industries towards implementation of these algorithms in various capacities with the primary motivation of using Big Data to get better insight from available plant data that can improve the overall economic bottom line.

However, it should be noted that a similar momentum in the past decades failed to materialize into tangible solutions due to a lack of organizational readiness to support and maintain these implementations [5]. Many corporations, even today, do not realize the significance the huge increase in digitalization and the interconnectivity of things may bring about. This has strong implications on future strategic pathways, which may be observed in the automotive industry that is struggling with the change from a production-centric industry to a business differentiated by software offerings. Since the (bio)chemical industry is not an end-consumer-facing industry, cyber-physical developments are not easy to spot and their significance may not be properly accounted for in managerial decisions. Therefore, it is important that one should neither get carried away by the exciting new developments outlined above nor be apprehensive of them.

Besides the algorithmic "auto-piloting" of a (bio)chemical plant, there are many other operational objectives that are of relevance to ensure smooth operations. These include:

- Planning and logistics in general that are required to ensure sufficient raw material, utilities and personnel are available to carry out process operations; the requirement to handle non-routine process upsets (for example, non-routine upsets of plant utilities that require a change in operational objectives [6]) and plant scheduling requirements that take into account the contractual obligations that must be fulfilled.
- Ensuring both asset utilization and equipment efficiencies are kept at high levels; as a part of this endeavor, the implementation of a robust equipment maintenance schedule that facilitates optimal equipment uptime and asset utilization and, to this end, the use of data-driven predictive maintenance concepts that allow for "just-in-time" scheduling.
- Safety and environmental factors that include minimizing waste, reducing plant trips, avoiding uncertain states and reducing storage of hazardous substances.

It should be noted that the above objectives (with the exception of predictive maintenance) currently rely on mathematical planning systems, personnel insights and ingenuity, handcrafted analyses or a combination thereof. The advances of Big Data-based methodologies can also supplement or replace some of these tasks, freeing people from repetitive jobs, with specifically trained operators/engineers only acting as application support.

Another aspect that must be addressed when discussing Big Data in process operations is the flow of data and the data infrastructure that must be established, in particular, to allow for "real-time" data gathering required to implement some of these Big Data-based concepts in an operational environment. Although the Industry 4.0 initiative [7] and, in particular, the Internet of Things (IoT) part, has been leading the facilitation of the "4 Vs" (volume, variety, velocity and veracity) of plant data [8,9], it can be seen that the process industries are lagging behind the automotive industry, for example, where these concepts have resulted in step changes in operations. This lag in the process industries is partially due to the lacking *value*, a 5th "V", of the data available. To this end, it is important to note that data flow from process plants needs to go beyond the traditional and well established 4/20 mA analog process signals toward digital readings and supporting infrastructure that can provide tangible improvements. Besides, new sensors that are capable of capturing information-rich data are required. Whereas the recent advances in Big Data-based algorithms together with the seemingly vast amount of "4 V" compatible data promise a seismic shift in the way process control and operations are handled in a (bio)chemical plant, a closer look at the current state of process control shows that these changes and their capabilities have yet to trickle down to the operational level.

Yet it is important to realize that this is not only a matter of technology. The different approaches put forward by statisticians and engineers, the former mainly in process monitoring and the latter in closed-loop process control, need to be synergistically combined to make Big Data work in a (bio)chemical plant. This, however, needs to start with a joint understanding of terminology. For instance, Box et al. [10] use Engineering *Process Control* (EPC) as a catch-all term that involves combinations of manipulated and controlled variables. Statistical *Process Control* (SPC) refers to the detection of faults, not explicitly covering the diagnosis of the cause. To distinguish between these, we will be using the term (Statistical) Process Monitoring (SPM) to encompass both detection and diagnosis of faults while the term Process Control (PC) will indicate the presence of various forms of closed-loop control structures as part of the operational assets.

Consequently, the objective of this manuscript is to provide a balanced perspective of Big Data capabilities and the requirements imposed by the plant operations. To this end, a systematic analysis will be performed to understand the plant requirements and how the developments in Big Data-based monitoring and control can fulfil them. Lastly, the focus of the manuscript will shift to understanding the barriers that must be overcome to implement these evolving developments.

2. Process Operations and Plant Data

2.1 Operation of (Bio)Chemical Plants

Prior to moving forward with the exploration of Big Data-based process monitoring and control, it is important to understand "the big picture" of a (bio)chemical plant in terms of its operational doctrine.

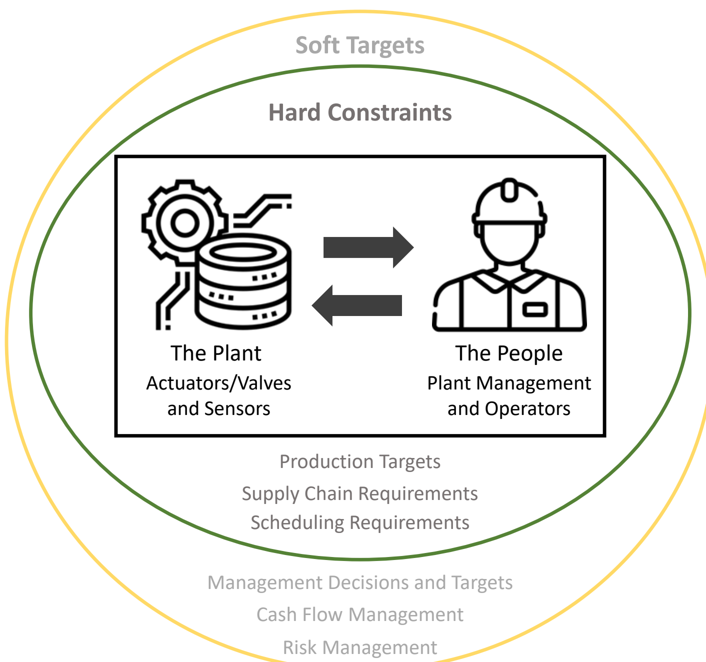

Figure 1. Operational requirements in a (bio)chemical plant.

As illustrated in Figure 1, the "system" consists of both the plant (the physical infrastructure) and the people, who operate the plant, support operations and assume managerial duties. The plant not only includes the unit operations and piping, but also the physical devices that facilitate process data gathering and control. This comprises control valves, sensors (including analysers) as well as the digital infrastructure such as the DCS (distributed control system) and software that enables the interaction with the physical plant and the people who operate the plant. Conceptually, this system may be further extended to account for procedural requirements imposed by the organization, affecting the execution of operational tasks and beyond. This set-up may be understood in terms of the P3 framework [11,12], which accounts for the interactions between different system element types as well as among elements of the same type. Together, the "system" works towards ensuring that all hard constraints are met. These constraints are, in general, contractual obligations such as promised production targets for both external customers as well as supply-chain requirements, or regulatory constraints such as safety related maintenance work. Then, the focus shifts to incorporating soft targets, which, in short, look at improving the overall economics of the plant based on top-level managerial decisions, cash flow optimization as well as risk management. Both process monitoring and process control work towards satisfying hard constraints and soft targets.

2.2 The Process Control Hierarchy and its Purpose

The purpose of process monitoring and control is to ensure that both hard constraints and soft targets are met during daily operations. At the design phase of a plant, it is important to ensure that the variables that are required to calculate or infer these targets and constraints are actually included in the design. This is equally valid in the context of Big Data, as having the appropriate sensor allocation plan so that the data can truly be used to monitor the process and, ultimately, meet these constraints. To this end, the information needs to be interpreted and decisions need to be made such that proper corrective actions are taken. SPM often requires operators to complete this loop where they have to interpret the result, decide what the remedial action is and implement it on the plant (either through physical action or through the automated controls).

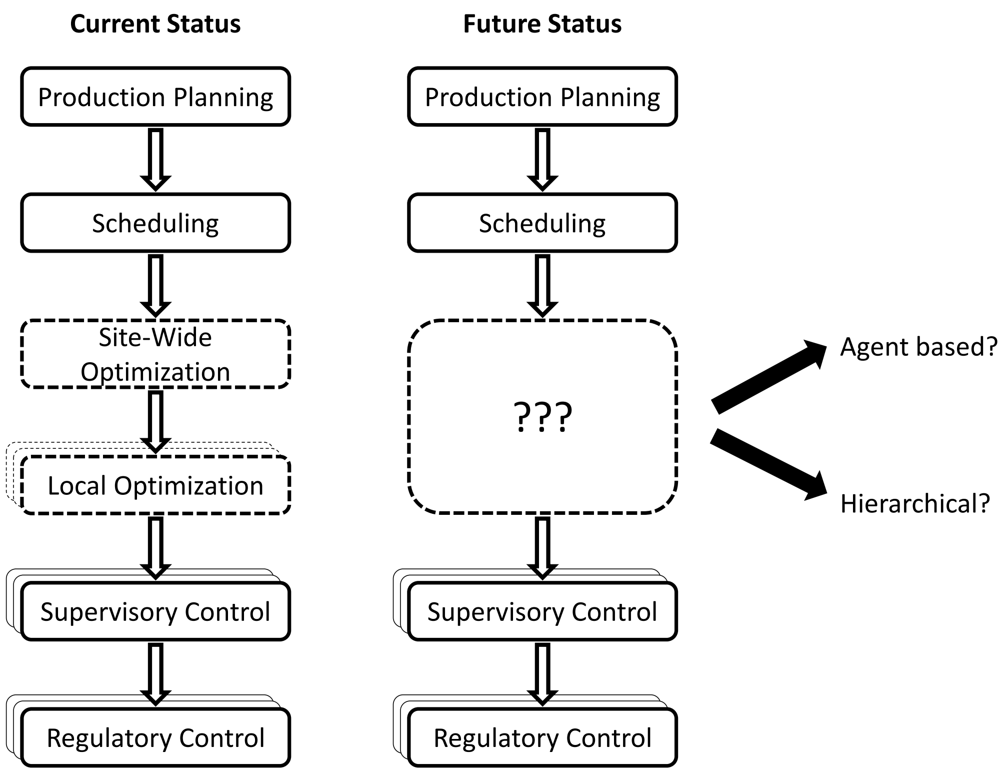

Figure 2. Process control hierarchy.

These targets and constraints are currently managed through a hierarchical approach as illustrated in Figure 2 [13]. Both are introduced at the long-term planning, the scheduling and the site-wide optimization levels and passed down to the plant through the layered control structure. However, in practice, the site-wide and local online optimization levels are not available in many industrial operations, which means that these constraints are either pushed up to the planning level or executed at the supervisory control layer. This disconnect also means that the information flow may need to be facilitated by a plant operator or an engineer.

Each layer (starting from the top) is progressively focusing on a subset, or a specific aspect, of the overall objective of managing targets / constraints and providing process targets (set-points) to the layer below. Thus, centralized planning decisions taken on time scales of months and years get distributed across the hierarchy down to “local” control loops that execute in the order of seconds or less. To this end, the lower levels of the hierarchy are increasingly focused on a smaller area (space) of the plant, but operate at a much higher frequency.

However, this hierarchical control structure might change with the current Industry 4.0 developments, which aim at greater agility in instrumentation and its ability to deliver flexibility in operations with a potentially fully distributed agent-based control structure [14]. These agent-based control systems can be visualized as a collection of “local” controls (having a scope similar to advanced regulatory controls), which perform their “local” control function, but are also interconnected with all other local controls through direct communications. As a result, an agent-based control system can be visualized as a mesh/network of controls acting collaboratively, similar to what the internet is to a computer network. Together, these local controls can communicate directly with each other and share relevant process information and co-ordinate their actions, which nullifies the need for a hierarchical control structure, as is common today.

In this context, the key to self-organization is the ability to automatically identify relevant connections amongst field devices. For instance, the concepts of controllability and observability could be put to use in finding relevant relationships between variables. Quantitative information from input / output data could be coupled to structural information, e.g., obtained from P&ID’s, to limit the range of (most) control actions to plant sectors, related fluid streams and equipment. This scope limitation will most certainly increase the robustness of the identification process, but also provide a different type of hierarchy, not in terms of decision-making, but in terms of logical composition of plant elements.

Without an actual implementation, it is difficult to assess benefits and drawbacks at scale. Nonetheless, the ability to continuously adapt to varying process conditions and operational requirements (as information flow to and from “local” controllers can be dynamically adjusted so that different sets of information can be prioritized) appears to be a highly beneficial objective. However, such a system does not exhibit natural means to prioritize control actions. All control actions from safety requirements, process regulation and optimization are all handled with equal levels of priority. Further, the relative complexity, especially for an outside engineer, requires specific system assessment techniques and different infrastructure requirements.

Even though the overall process control hierarchy might not be “ripped and replaced”, data-driven concepts will significantly influence the execution and hierarchical nature of site-wide optimization, local optimization and, to some extent, the supervisory control layer. This is due to the fact that these layers currently rely on fundamental process understanding that utilizes first-principles-based models [15], which have difficulty in accurately capturing secondary and tertiary phenomena, such as the generation of impurities as well as long-term effects in (bio)chemical processes (e.g., fouling and catalyst activity).

2.3 Process Monitoring vs. Process Control

In the context of this manuscript, it is important to clearly demarcate the boundary between statistical process monitoring (SPM) and process control (PC). While both schemes share the same goal of lowering the variation in the process and ensuring that soft target and hard process constraints are met (Figure 3), the way in which these two schemes achieve this outcome is different.

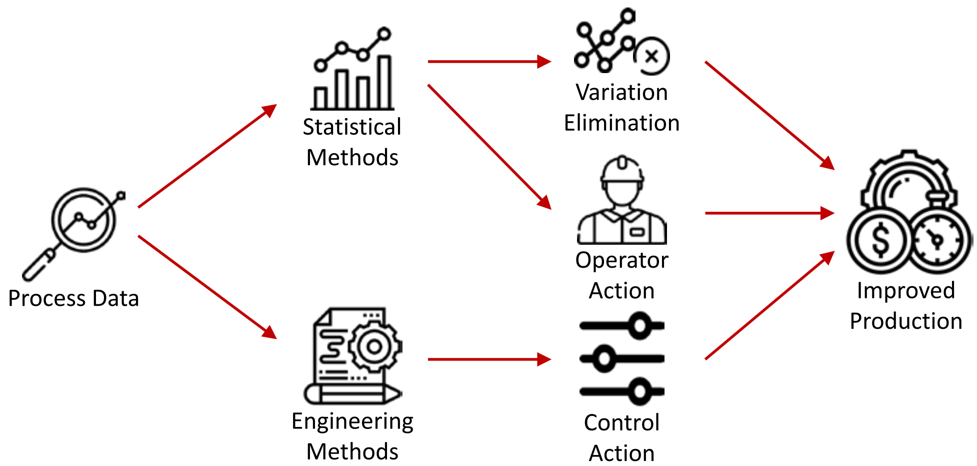

Figure 3. Routes from data to actions.

SPM refers to a broad term that encompasses Statistical Process Control (SPC) methodologies [16], where the objective is to monitor the progress of the process in real-time and to detect unusual process behavior as it develops. "Unusual behavior" refers to a change in the expected or predicted amount of variation in the process. While this has been the primary concern in process monitoring, increasing emphasis has been placed on diagnosis in recent applications. In the past, the diagnosis was expected to be accomplished by the operators/engineers who possess process understanding and knowledge. This was perfectly reasonable, particularly when process performance was monitored through few process variables or quality metrics. In the evolving era of Big Data, however, processes are more commonly monitored through a multitude of variables. To compensate for issues stemming from multiple testing and potential strong correlation among these variables, summary statistics are often used. This leads to complications in identifying the root cause of an unusual behavior detected by the SPM scheme. Commonly used approaches based on chemometrics methods such as Principal Component Analysis (PCA) and Partial Least Squares

(PLS) claim to achieve detection through summary statistics based on the principal components rather than the original variables and diagnosis through contributions of these variables to these statistics. However, the large contributions of certain variables to the monitored statistics may simply be a consequence and not the cause of the fault. Furthermore, it can be shown that while effective for simple faults, a diagnostic approach simply based on contributions may not be able to identify complex faults [17,18].

In that regard, there is a push towards incorporating data-driven methods for diagnosis on top of detection [19], which, in principle, provides the operators with an idea of how to fix the anomaly. There have been attempts to also use ML algorithms towards achieving this goal [19,20]. This can in theory still be aligned with the original context of SPC where following an alarm, the culprit, a.k.a. assignable/specific cause generating the unusual behavior, is expected to be identified and eliminated, rather than the alarm tempting the practitioner to 'tinker' with the process to compensate for its effect.

In comparison, (closed-loop) process control refers to a situation where data from sensors are used to calculate and then implement actions via the control valves and other actuators of a process in real-time. Thus, in real-time closed-loop control, the process is not only actively monitored, but corrective action is taken as the process progresses to ensure that targets / constraints are met. For example, Udugama et al. [20] investigated the reasons for off-specification product generation in a high-purity multi-component industrial methanol distillation column. The middle-boiling ethanol component that forms a bulge near the side-draw location was the culprit. A soft sensor, which combined three composition measurement signals (gathered from a multiplexed gas chromatograph) near the side-draw region, was then used to determine the ethanol bulge location and size. For normal operations, a PID controller performs the corrective control action by manipulating the reboiler duty based on the ethanol concentration at the side-draw. However, if the ethanol bulge was deemed out of the normal operation regime by the soft sensor, process variable conditioning was applied so that the PID controller would drive the energy balance to return the ethanol bulge to normal operation.

The different approaches of these two methodologies, process monitoring vs. closed-loop process control, often and unfortunately lead to an either-or situation for the practitioners. It certainly does not help that the two schemes belong to different camps, one led by statisticians and the other by engineers. Particularly, in many statistical process monitoring applications, the closed-loop nature of the process is simply ignored or overlooked. However, a few examples can be found that explicitly discuss this [10,21–26].

In the more generic sense, closed-loop control can be seen as a symptom reliever and becomes effective if the control action is easy and cheap to take. For more persistent disturbances, however, control action may be too costly. Potentially, this can be more effectively resolved through monitoring methods. Yet most monitoring methods are developed under open-loop assumptions and can lead to erroneous conclusions if closed-loop control is present but ignored. Moreover, if a specific cause is detected and identified, its elimination may require costly actions as well. In that sense, hybrid approaches that incorporate both schemes can be expected to exploit the synergy that would be more efficient and effective in variation reduction and optimization.

2.4 The Origin of Data in (Bio)Chemical Plants

Prior to defining Big Data for the purpose of process monitoring and control, it is important to understand the potential sources of relevant data available in (bio)chemical manufacturing. The data originate mainly from the following sources as illustrated in Figure 4.

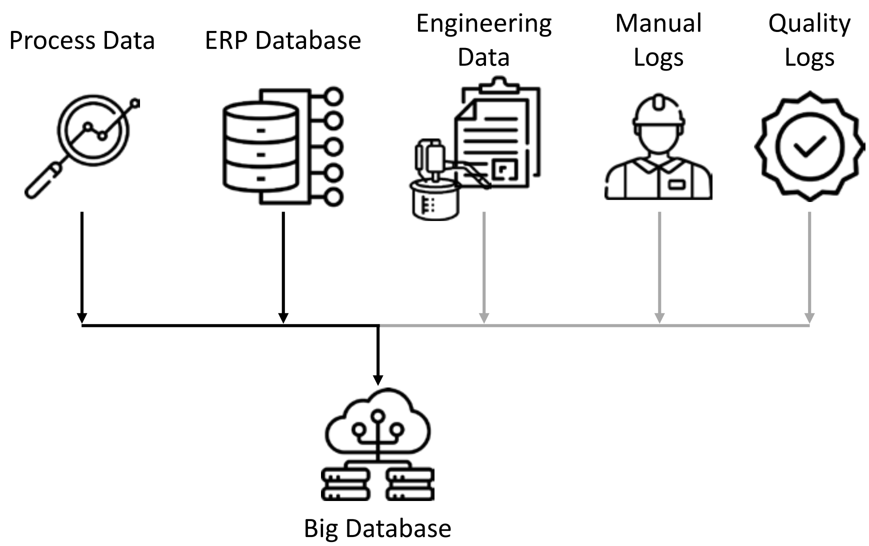

Figure 4. Data repositories in a (bio)chemical plant.

2.4.1 Process Data

Process data comprises time-stamped information gathered from sensors installed in the plant. In general, the historical process data consist of all process variables that are available for plant operators as well as data related to actions taken by the operators over time. In addition, the historical process data may also contain calculated process values such as energy usage or the efficiency values of a key unit operation, start-up and shut-down data related to unit operations (generated at a PLC level). If available, variables that indicate the state of the process (e.g., normal operation, start-up / shut-down, plant trip), the status of control loops (e.g., automatic, manual or cascade) as well as other sensor data (e.g., valve positioner, vibrational measurements of rotating machinery) are tracked.

Historically, for the purpose of storage, these data were mostly compressed, as hard disk space was expensive. Data compression is still used today, but perhaps not to the same extent. Setting standards on data compression and how to deal with inherent irregularities between sampling rates (temperature sensor vs. online composition analyzer) is of importance in ensuring that the recorded data can be easily accessed for Big Data-based development purposes as opposed to the significant efforts that are needed in data pre-processing currently. Similarly, having a uniform data time-stamping protocol between the other sources of data (in particular, quality and Enterprise Resource Planning (ERP) data) is crucial in ensuring smooth data reconciliation efforts. SPM and PC activities (including model building [27]) rely on the data that is continuously generated. Hence, it is important to recognize that the data must comprise unambiguous information about the structure of the process and any changes to it. For instance, a seemingly insignificant change (control on / off) may induce nonlinearities and changes the input / output structure.

2.4.2 Enterprise Resource Planning (ERP) Database

They contain time-stamped information related to what has actually been processed in a plant as well as information related to equipment usage and breakdowns, resembling a digital logbook. In particular, day-to-day, week-to-week, month-to-month operations that have a direct effect on the economic success of a company are documented here. This includes information such as data on raw materials and other inputs used in production, as well as information related to what has been processed, in particular, plant recipes and plant capacities. In addition, the ERP system carries information related to the supply-chain and its constraints. In the context of process control, ERP data are critical for tracking and tracing equipment breakdowns or changes to raw materials or production specifications, which inherently affect plant operations. Process historical data and ERP data can be jointly used to identify stable and relevant process operating periods.

2.4.3 Engineering Data

During plant design, a wealth of data is created. Steady-state process simulation work generally starts with the development of property data packages and their validation against experimental data and plants that serve as reference. From that foundation, the process layout (flowsheet) is developed, accompanied with material and energy balances. This essentially is the starting point of the plant's life cycle.

These data are subsequently used to engineer the plant in greater detail, spanning diverse domains from equipment design to piping, electrical and instrumentation to civil engineering. As every domain prefers specialized software, there is generally not one (e.g., simulation, calculation or 3D) model with a single database, but several. A case in point is the equipment design, which uses calculation tools with different scenarios (e.g., normal operation or water runs), so that property data packages tailored for process design are often not used. Instead, the rather limited property data needed for such design tasks is often individually specified by the engineers, whenever required. Every domain requires specialized data that is mostly of limited value to other domains. This is due to some extent that the veracity of data remains uncertain. Furthermore, even within the suite of tools of a single software provider, it is not always conveniently possible to exchange and reutilize data. The large engineering contractors and in-house engineering divisions appear to have better aligned tools and databases, but even they rely to some extent on abundant labor to make up for the limitations that arise during data interchange. Although joint database systems (e.g., Smartplant Reference Data), data interchange protocols (ISO 15926) and ontologies [28] exist that permit and facilitate rapid data transfer in principle, many companies lack work process consistency (a.k.a. enforced work standards and interface ownership) to leverage such technologies. Besides, the data that is available often gets "lost" during commissioning, because of the direct and indirect cost of data handling and integration from the contractor. Since such data sets are not required to operate a plant and incur upfront costs without tangible benefits in sight, it remains a strategic decision to adopt a Big Data mindset [9,29].

Consequently, P&ID data, equipment datasheets, HAZOP studies, operator handbooks and similar documents are the sole memory of the engineering process. As such, they contain the plant layout together with all available sensors and control structures. The equipment specifications encompass physical dimensions of the equipment including practical constraints, such as operating pressure and temperature, metallurgy and pipe wall thicknesses. This information can be used to cluster relevant process historical data to given unit operations and to define operating envelopes. Such data are usually stored in separate repositories, either digital or analogue, and the relevant data are mostly unavailable for automatic analysis. In the context of process control, a modified form of the P&ID diagrams is used as the Human Machine Interface (HMI) in many DCS applications.

2.4.4 Manual Logs

Considering that many (bio)chemical plants were commissioned in an era prior to Big Data, there are many data sets that are still recorded manually and are often not available digitally, though, a shift towards digital logs can be observed. This information can include maintenance logs, observations, raw material additions or changes to actuators that are carried out by the operators.

2.4.5 Quality Logs

Quality logs are a type of manual logs that are generally attributed to the laboratory analysis of samples. However, these logs may also be attributed to in-line and at-line quality measurements that are performed by plant operators on specialized equipment and are not necessarily carried out at a laboratory. These data are generated by scheduled off-line sampling and analysis recorded by operators as well as quality assurance personnel.

2.4.6 Alternative Sources

In addition to the five areas of data repositories that represent the current state of data utilization, there is the possibility of also exploiting other sources of data. For example, validated dynamic process simulations of key unit operations in plants connected to the operation (also referred to as digital twins) are increasingly being developed in the industry. These simulations, unlike the "physical" process, can easily be used to explore the boundaries of the process envelope without placing the actual process at risk and generate large quantities of process data to augment plant data sets.

Similarly, emerging technologies such as free-floating sensors, tracer experiments and, for that matter, Computational Fluid Dynamics (CFD) models can be used to gather spatio-temporal data that is often missing from the traditional data repositories. Other data sources such as videos and other peripheral sensors may also be beneficial in tracking actions (in particular by operators) when such actions are not recorded.

3. Big Data in the Context of (Bio)Chemical Plants

In (bio)chemical plants the aforementioned data have been recorded for decades, but not automatically compiled into a joint database that can be efficiently analyzed and exploited. Consequently, significant efforts by process control engineers have been required to pre-process and pre-structure the different data sources. Hence, analyzing these data sets has been a challenge in (bio)chemical manufacturing even before the Big Data revolution.

Prior to moving forward with a detailed look at the use of data in SPM and PC applications in (bio)chemical processes, it is necessary to address the value proposition and nature of "big data".

The value proposition of Big Data is somewhat intertwined with the advancement of data-driven process monitoring methodologies. Noticeable groundwork has, in fact, been carried out using data-driven methods to extract meaningful correlations out of Big Data, and it is the foundation of SPC activities [30], illustrating that Big Data is capable of associating process variables with factors such as quality and the status of unit operations. For example, SPC techniques were employed to predict the harvest time in a fermenter [30], which for all practical purposes is impossible to accomplish without the existence of Big Data sets. This is also the case even in hybrid approaches that rely on mechanistic models, since Big Data sets are required for both parameter and state estimation both in model development as well as real-time calibrations [31]. Data-driven models have been used to predict future behavior of critical quality parameters in seemingly chaotic systems [32], where historical and current behaviors of the process are linked to the expected future behavior. In short, the value of "big data" in (bio)chemical operations is its ability to support the operation of complex mechanistic and data-driven tools which can transform this data into critical actionable information for plant operators (engineers) and process control systems. However, it should be noted that the process control engineers as well as field engineers find these tools somewhat complicated to use as these are not off-the-shelf application solutions.

To understand the nature of big data and thus its inherent value, the "4 Vs" concept [8] can be interpreted as follows for (bio)chemical operations:

- Volume: What makes Big Data "big" is the volume of data that is available. This is reflected in the number of variables that are recorded at a given frequency over a long-enough time span, thus providing a large number of data points. Acquisition of current data may be complemented with "static" and "quasi-static" data from other data repositories, e.g., design data or internal transfer prices. By their nature, static or quasi-static data usually represent many more variables, but fewer data points. It should be mentioned that not all large data sets are necessarily "big". In particular, if a phenomenon of interest occurs quite rarely, a Big Data set should contain enough occurrences as to provide sufficient insight. A similar point can be made about the breadth of data, or rather the lack of it.
- Velocity: In the context of (bio)chemical operations, one may find two extremes of data in terms of the speed with which such data are generated, whereas a faster update rate often seems beneficial. Process data originating from process sensors generally have a high velocity of data generation. It can be argued that the optimum velocity of data generation is a function of the smallest meaningful time constant of the process. As such, for process data it might even be beneficial to sacrifice some of the velocity for improved accuracy. At the other end of the spectrum, quality logs and manual logs are often recorded at a fairly low velocity due to time consuming tests that need to be performed and the fact that operators are busy with managing the plant. In these cases, automating

sampling and testing efforts can significantly improve the speed at which such data become available. Process analytical technologies (PAT) tackle this need through the development of soft sensors and inferential sensors that provide operators with actionable information [33–35].

- Variety: As detailed before, there is a large variety of process data available in a (bio)chemical plant, with different structural properties in separate repositories. For example, process data is stored with corresponding time stamps and other meta-information, while ERP data often contain start and end times or datasets sorted according to a batch number. Alternatively, ERP data may provide time spans for which settings, recipes, etc. are valid, whereas they do not necessarily indicate whether such settings are used or to which extent. Generally, ERP data come in many different flavors and colors, often dependent on organizational hierarchies (business units, geography, etc.). They often require extensive pre-processing and validation in addition to the creation of custom-extractors to assemble the data sets for further use, e.g., for BASF's Verbund-Simulation [36]. Similarly, the quality and manual logs are usually sorted with batch numbers, while P&ID diagrams contain "static" structural and to some extent spatial information of the process. As a consequence, the variety of data and the disjointed and uncoordinated nature of the data structures require significant pre-processing. By contrast, the difference in velocity within a single type of data (in particular in process data) necessitates careful reconciliation and pre-processing to align the data sets.
- Veracity: Data sets inherently differ in accuracy, precision and the level of trust that can be placed in them, which is covered by the term veracity. Data volume and data veracity

can complement each other to some degree, but it remains a challenge to infer insights and decisions from data of poor quality. Maintenance of sensors in industrial-scale chemical facilities is generally dependent on their relevance for operation and even more so if the operating license depends on proper servicing and calibration. Safety Instrumented Systems (SIS) that ensure a high Safety Integrity Level (SIL) are well documented, regularly checked and calibrated. By contrast, conventional installations often do not draw much attention, so that their reliability deteriorates over time. Hence, readings from conventional sensors are often low quality and permit a qualitative view into the process, but their quantitative value is limited. However, validated mechanistic models could be used to track the quality of sensors through data reconciliation and to infer true values. Similarly, the veracity of manually recorded data, e.g., in ERP systems, depends on the organization's culture. Only if the value of correct data is fully ingrained in the management and operational layers of an organization, such data can be harvested and insights be generated that are indeed generalizable.

The "4 Vs" discussed above provide a framework for the assessment of the value of data. Whenever the data available is sufficient to generate insight into the process, it is clearly of value. To be of value, it must be available for analysis in sufficient volume and velocity, covering a sufficiently broad variety of relevant factors and be trustworthy (veracity). Then, the data is deemed rich in "information". Information richness is of great interest in particular for fermentation-based processes, where the typical pressure, flow and temperature sensor data do not contain sufficient "information" to predict states of interest such as product formation rates and inhibition phenomena. In this instance, prior to unleashing Big Data tools, the focus needs to be shifted to data gathering and augmentation that is, in aggregate, information rich, for example, by making use of image-based data sources [37]. Another example is the use of a non-linear infrared spectrometer to gather vital state information for a second-generation bioethanol process [31].

4. State-of-the-art in Process Monitoring and Process Control

As discussed before, Big Data relates to many data of differing characteristics, whose value arises from its exploitability. Big Data and associated methods are of great value if they provide actionable information either to operator teams or to automatic controls. In particular, results from Big Data analytics should provide guidance and insight to make better decisions. Accordingly, information must be presented in a comprehensible and coherent manner to well-trained operations staff and domain experts alike to elicit informed decisions.

To provide an answer to the question whether Big Data-based process monitoring and control currently exists, the recent developments in these fields and relevant Big Data-based methodologies employed in other domains are analyzed.

4.1 Process Monitoring

The current state-of-the-art in process monitoring falls into three main classes: model-based, knowledge-based and data-driven methods [38].

The term *model* is used broadly and applies equally to first-principles models derived from physico-chemical insights into the system as well as qualitative or quantitative relations between sets of variables. A model-based method in this context often relies on first-principles models. If such a model is able to represent the true plant state, any deviation between the model and the plant can be interpreted as an indicator of the health of the process, for which the “digital twin” concept has gained increased attention [31]. However, the development and validation of high fidelity models can be expensive and difficult due to the complexity of current industrial processes and the size of the operating envelope.

The knowledge-based methods take advantage of expert domain knowledge, which may be represented as graphs and trees or as connected objects that represent structural information or functions [38]. Although the results tend to be more intuitive, gathering all the required knowledge is a lengthy and costly process.

Both approaches are of importance in the domain of process monitoring and employ process data for the purpose of validation and parameter estimation. However, these methods do not generally utilize large amounts of data and instead focus on specific information (i.e., selected process measurements). For that reason, neither model-based nor knowledge-based methods will receive any attention here. Nonetheless, they might prove invaluable in the hybridization of data-driven methods to compensate for the lack of *a priori*-structure.

4.1.1 SPC and PCA

Data-driven process monitoring methods have become quite popular, especially when dealing with complex industrial processes where models and expert knowledge are hard to obtain [39]. Moreover, the increasing availability of sensors and data gathering systems have subsequently motivated the development and evolution of data-driven industrial process monitoring methods. In fact, these methods enabled process control engineers to concatenate large data sets into comprehensible features to make predictions about crucial process variables. Statistical Process Control (SPC) involves schemes for continuous monitoring of a production process so as to identify situations exhibiting more than the expected amount of variation [40]. To this end, SPC control charts have been effective over the years to maintain statistical control over critical manufacturing outputs and other complex systems where the variation originates from a broad range of sources. Taking into consideration that many industrial processes require simultaneous monitoring and control of several variables, multivariate extensions of the above-mentioned SPC methods have also been developed [41].

However, another important issue when dealing with multiple variables at once is the potentially high degree of correlation among them. Standard approaches to multivariate SPC as in the case of Hotelling's T^2 chart are highly impaired under such conditions and an accepted practice to tackle this issue is to first transform the physical variable space to a space of uncorrelated latent variables, followed by the development of a low-order model, which captures most of the variance of the original dataset.

PCA is the most popular latent structure method for process monitoring due to its ability to successfully handle high-dimensional, noisy and highly correlated data [42]. PCA (and PLS) models are built on historical datasets, which are assumed to come from a period of "in control" (normal) operation; then, the future observation/sample is checked to see if it follows the model [41]. PCA-based multivariate monitoring methods have been widely applied for fault detection in chemical processes to improve quality and productivity. These methods have the benefit that they can handle both process and product quality variables and, even though they have been used primarily in the area of chemometrics, they are appealing in any multivariate process setting.

However, in industry a great number of processes exhibit nonlinear behavior, and thus, the use of PCA or PLS in cases where the data has nonlinear characteristics leads to poor monitoring performance. If nonlinear systems are investigated by applying PCA or PLS, nonlinear effects can become partly visible in the higher latent variables. This may hinder the identification of the optimal number of significant latent variables. To tackle this, several approaches have been developed along the years, such as the nonlinear PCA method based on auto-associative neural networks [43] and principal curves modelled with neural networks [44], which aim at performing a nonlinear data reduction analogous to PCA. Further, the general function approximation capabilities of neural networks have been exploited to expand PLS regression to nonlinear problems [45–47].

Several examples of the application of PCA to real chemical data from DuPont and other chemical companies have been discussed and published [42]. Likewise, there have been reported cases of the application of PLS models in SPC as in monitoring of a distillation unit and a fluidized bed reactor [48]. It is also important to mention Controller Performance Assessment (CPA) as a branch of these data-driven methodologies that focus on monitoring and diagnosis of the performance of process control loops, in particular at the regulatory level [49].

4.1.2 t-SNE, LargeVis and UMAP

t-SNE [50], LargeVis [51] and UMAP [52] belong to a group of manifold learning techniques that aim at data reduction. Similar to PCA, such techniques map vectors to a lower dimensional space while preserving the intrinsic structure of the data. These methods scale to large data sets with millions of data points, while capturing nonlinearities, in contrast to PCA. UMAP, which represents the current state of the art, has recently been applied to the visualization of industrial chemical process data by Dow Inc. [53]. In principle, UMAP and similar techniques may be applied to any monitoring task.

4.1.3 Neural Networks

One of the significant commercial applications of (artificial) neural networks (ANNs) is in image classification. Generally, layers of Convolutional Neural Networks (CNNs) are combined with pooling layers to extract features from 3-D data sets, i.e., images. Additional layers provide the ability to extract more complex features. In the context of process monitoring, these type of networks are employed in analyzing time-series of variables that can be “color-coded” either for direct identification of characteristic patterns or Fourier-transformed. An example is given by Wu and Zhao [54], who use CNNs to identify faults in the Tennessee Eastman process. CNNs have also been proposed and employed in the classification of hyperspectral imaging [55], which has potential applications in the bio(chemical industries), such as the “real-time” functional property identification [56]. It should be noted that these applications are generally employed to compare a given time window of operations (be it classical process data or data from novel sensors/methods) with overall process operations to identify if there is a departure from the normal operation and to which group this time window of data belongs. As such, these approaches belong to grouping and clustering methods, which are primarily used in process monitoring.

On the other hand, regression-based methodologies such as Recurrent Neural Networks (RNNs) attempt to act as a model identification mechanism, which can ultimately be used in closed-loop control. RNNs feed outputs back to the network and thereby realize a storage (memory) mechanism. This mechanism captures temporal dependencies between input and output data, conceptually similar to a system of discretized ordinary differential equations (ARX-type models). Because of the difficulties that arise from training deep networks (e.g., vanishing gradients), plain RNNs cannot cope with effects that span more than a few time steps, rendering them unsuitable to handle widely differing time constants. For process control and monitoring applications, they cannot account for the long time-scale associated with process dynamics that must be captured by a dynamic model.

This has led to the development of new architectures such as the Long-Short Term Memory (LSTM) [57] or the Gated Recurrent Unit (GRU) [58], a simplified variant, that enable the training of deep networks and thus capturing long-range trends in parallel to short-term effects. In the context of (bio)chemical processing, training of such models on input-output data is an act of model identification and the resulting model can, in principle, be used for every task that requires dynamic models. As stated before, if such a model is able to replicate the true plant state, any deviation between the model and the plant can be interpreted as an indicator of the health of the process.

The pre-processing of data prior to being used in an RNN type model identification is of practical importance when dealing with industrial plant data. Auto-encoders learn structural information, e.g., correlations among variables, by setting the targets to be the input values. Recreating the input through a constrained network helps in the identification of features, similar to PCA [59]. This is valuable in high dimensional process data to observe trends. Besides, an auto-encoder could be a useful pre-processor to transform input data to latent variables, thus easing the training of the network segment, e.g., an RNN-type network attached to the auto-encoder.

However, the training of ANNs does not just require a lot of data, but a lot of information-rich data. This is in particular an issue in continuous production processes such as in chemical and fine chemical operations where time-series are often stable and tightly regulated for most of the time, as such processes do not provide much dynamic information. Similarly, semi-continuous processes such as fermentations often follow similar trends during batch operations, which effectively limits the information content of the data obtained. To this end, repurposing of pre-trained models, i.e. transfer learning [60,61], may soften this requirement and widen the range of successful applications.

4.1.4 Soft Sensors

Soft sensors are advanced inferential on-line calculations, which combine different types of variables to predict the desired output. This output can then be employed for the purpose of control. Often these sensors are based on data-driven methodologies and are used as input process variables to feedback controllers. They generally rely on models, be they derived from data (e.g., PLS, ANNs or SVM), first principles or part of a Kalman-type filter. In general, such soft sensors are not exclusively a true process monitoring or control concept, but rather an enabling data-driven technology. Luttmann et al.[62] provide an overview of soft sensor techniques applied to bio-based production processes, mainly focusing on mechanistic models and purely data-driven approaches. In recent years, soft sensing techniques using hybrid models have been researched intensively as well [63].

In some publications, spectroscopic measurements have also been labelled as ‘soft sensors’, which can be explained by the fact that spectra (e.g., near-infrared, Raman) are usually pre-processed followed by a chemometric modelling step, where the chemometric model (in most cases PLS, the industry standard) produces information similar to hardware sensors. It is the authors’ experience that attempts made to employ ANNs in general has not added further value yet to practical applications.

In general, despite the large number of publications on development of soft sensors for bio-based production processes, most developments have been and remain academic with relatively few well-documented applications industrially. This situation is only changing gradually, and is often driven by the fact that there is much easier access to process data in modern bio-based production plants.

4.2 Process Control

The domain of closed-loop process control, in contrast to process monitoring, exhibits very few purely data-driven methodologies and concepts that are currently applied in industrial settings. Even though there are examples of data-driven closed-loop control concepts in academia, most are hybrid approaches that augment the capabilities of classical mechanistic process control concepts. These examples are not necessarily Big Data-type applications, but rather methods that act on limited data to account for process nonlinearities.

4.2.1 Self-Optimizing Control (SOC)

An established and long-standing example of using data in closed-loop process control is in improving controller performance through self-optimizing or self-tuning controls. In general, self-optimizing control adjusts parameters of a controller (e.g., PID), to improve closed-loop performance. These controllers can adapt in real-time to deal with process nonlinearities and changes in operational requirements. One method of achieving this is to first identify the dynamics of the plant (online model identification) and subsequent decision making based on this model. This is then used to modify the controller parameters in accordance with closed-loop performance measures [64–66]. Alternatively, an ANN can carry out simultaneously both the model identification and the decision-making. For instance, Kumar et al. [67] completely replaced the PID structure with an ANN-based controller with three hidden nodes that acts similar to a PID controller, where a recursive least square training method is used to optimize the performance in real-time.

In general, the concept of a self-optimizing PID controller can be considered as a hybrid approach, where both mechanistic and data-driven concepts are merged to provide a feasible solution. However, with the relatively sparse use of data (often a single process variable and manipulated variable value) an argument can be made that SOC applications at a regulatory control level do not truly utilize Big Data.

4.2.2 Model Predictive Control (MPC)

Model predictive control (MPC) is a control concept used primarily for constraint optimization of unit operations. In particular, most commercial implementations of MPC structures rely on step-testing data to derive the input and output relationships (model identification) required to set up an MPC, be it a linear or a nonlinear model [15,68]. Consequently, even the classic MPC implementations rely on data in the development of the models used. The use of ANNs for the purpose of model identification has been an area of both industrial and academic interest since the 90's with Pavilion Technologies leading the industrial implementation drive. In industry, these model identification tasks are generally performed offline during the set-up process of an MPC. This is mainly because the online ("real-time") model identification is seen as a technically complex activity, which requires a high degree of robustness and well-trained operational support that can be often missing.

One key technical issue that arises when implementing "real-time" model identification is the need to extract a model in closed-loop as opposed to the open-loop step tests that can be performed in an "offline" identification situation. For this reason, MPC concepts suitable for online closed-loop model re-identification have been explored, in particular for polymerization reactions [69,70]. A similar concept has also been applied to the control of an artificial pancreas [71]. These MPCs employ a combination of model accuracy monitoring together with closed-loop model re-identification to adapt "online" to inherently nonlinear process conditions. In general, these models employ some form of regression, such as recursive least squares techniques in the model identification part, while concepts such as Kalman filters are used to identify the mismatch between the model employed and the current state of the process together with performance criteria to decide when model identification should be carried out. However, the structure of the process model is constrained by the initial user specification, hence, these attempts are best categorized as hybrid approaches.

In contrast, Akpan and Hassapis [72] have developed a completely ANN-based implementation of adaptive MPC, representing an almost purely data-driven control attempt. In this instance, an adaptive recursive least square (ARLS) method is employed for initial model identification, which is then used to derive a controller design and tuning for control purposes. This process is repeated at each point in time whenever new data are available [73].

4.2.3 Machine Learning-based Methodologies

Machine Learning-based algorithms can in general be classified into the three main areas of supervised and unsupervised learning as well as Reinforcement Learning (RL). Supervised and unsupervised learning algorithms attempt to find mappings of variables that can be employed to explain the movement of variables of interest. Unsupervised learning thereby relies on matching similarities in the data, whereas supervised learning relates input-output data. RL is different as it aims to derive a control policy by interacting with the system. Random deviations from a policy that is acting on the system improve the learned policy, which improves the policy over time. The general versatility of ML-based methodologies allows these concepts to be applied to a multitude of tasks in aiding or performing closed-loop process control.

A prime example of ML concepts applied in an actual closed-loop setting is the ANN-based implementation of an adaptive MPC as discussed in the previous section [72]. In this instance, the

MPC is used to compute the controller response, but all other aspects of the controller rely on a ML concept. Further, concepts such as predictive representation of states [74,75] may be combined with a standard MPC in making a decision about whether or not to compensate for a given process disturbance. The benefit of such a combined MPC strategy is that it reduces the frequency of control moves a MPC will make, which can have ramifications for some unit operations where, for example, temperature and pressure swings are undesirable.

Gautier et al. [76] implemented a model-free strategy based on genetic algorithms (GA) to control fluid flow in aerodynamics. This controller performs equally or better than the “state-of-the-art” mechanistic benchmark. The real value of this approach comes with the identification of a new control policy. RL utilizes agents that learn a strategy to maximize rewards. They are particularly well suited to play relatively complicated video games such as Super Mario [77], whereas more recent video games with less deterministic gaming engines, such as Angry Birds, have been rather challenging [78]. These algorithms are somewhat similar to classic closed-loop process control concepts as they are trained to meet targets and to maximize an objective. They represent another promising area of research that has been exploited in a limited manner in closed-loop control [79].

Mnih et al. [80] trained a deep Q-network to play Atari games with 38 days of game experience. Considering that only discrete actions are possible and rapid responses from the game engine may be observed, a direct application to a (bio)chemical plant appears feasible only if the network is trained on a simulation. Besides, one has to keep in mind that many control actions are continuous and long-range effects are difficult to pick up and thus to learn. As with the approach by Gautier et al. [76], the real value would arise from the identification of new operation policies.

Further, ML can augment the capabilities of other closed-loop process control concepts, and as such establish a pillar in hybrid closed-loop applications. For instance, high-dimensional data such as camera pictures (including hyperspectral imaging [56]) can be used to extract features of interest, such as functional properties. These features can then be combined with traditional mechanistic control concepts such as Fuzzy Logic [81] in control applications. As previously discussed, data pre-processors such as auto-encoders extract the relevant features from input data. By omitting the unneeded data fraction, they reduce noise in the data. This advanced noise filter would allow controllers to be tuned more aggressively while maintaining robustness.

Similarly, online monitoring of sensor performance is another avenue in which ML concepts can add value to mechanistic control [82,83], especially when sensors lack the desired level of robustness. Process classification and benchmarking with local models [84] can further be used to adjust operational objectives, e.g., in the course of catalyst deactivation. This also applies to the controller performance assessment (CPA) domain, which focuses on issues such as valve stiction, deteriorating measurements and control performance [49]. Beyond the assessment on the individual controller level, plant-wide clustering methods may help to identify external influences that cause individual controller performance issues. It should be noted that the content in MPC, SOC and ML subtopics in this section seem to be somewhat intertwined. From a "helicopter" perspective, both SOC and MPC controllers were originally developed based on "classical" mechanistic process control philosophies. However, these concepts have increasingly moved towards ML and statistical methodologies, in particular when contemplating real-time implementations and diagnosis, as such becoming hybrid in nature. In contrast, the practical ML methodologies have also borrowed elements or concepts from mechanistic process control philosophies. Hence, one key difference among MPC, SOC and ML is the proportion of mechanistic and data-driven methodologies that are employed.

5. Future Directions

An analysis of the current state of process monitoring and control applications shows that a key step will be the development of methods that merge closed-loop process control with additional functionality. Such functionality may comprise methods for troubleshooting and diagnosing as well as (closed-loop) model identification and the triggering of model updates. This is going to be a natural extension of the established SPM and ML concepts. In this context, it might even be more appropriate to talk about Big Data-*assisted* methods instead of Big Data-*based* methods, to distinguish between added functionality vs. hitherto unseen approaches to closed-loop control.

By contrast, monitoring methods must become aware of implications that arise from the presence of feedback control loops, which are part of the system under investigation and may be turned on and off or modified to account for changing process characteristics. Clearly, monitoring of the system's integrity and performance is of value, but functions that provide quick diagnosing of the underlying issues yield significant additional benefits, especially in the face of retirement of experienced personnel and other organizational restructuring initiatives. Supplementing quantitative monitoring methods with structural information, i.e., engineering data such as P&IDs or knowledge-based methods [38], may provide the needed reasoning capability.

There are three main drivers that are currently pushing the development of Big Data-based closed-loop control in the (bio)chemical industry. First is the economic value proposition for Big Data-based methods as opposed to the classical mechanistic process control concepts, while second is the availability of ever improving SPM and ML algorithms. The third and final driver comprises the improvements in the digital infrastructure including wireless sensing and cloud computing capabilities.

Any process control concept that promises to reduce the expert time requirement has an inherent economic drive. Thus, the economic value of SPM and ML methods originates from minimum engineering and/or expert time for development. However, in practice, these methods can take similar development and deployment times and expert involvement as classical concepts. Second, the field of information technology, i.e., computer science and AI, continues to develop advanced algorithms for a multitude of purposes including facial recognition and natural language processing (NLP), with increasing capabilities and sophistication. From a process engineering and (bio)chemical industry perspective, these developments are improving the capabilities of what SPM and ML algorithms may be able to achieve with limited investment. This relatively 'free' technology development is a natural driver to adopt Big Data-based process monitoring and control. From an implementation point of view, the effort for adopting suitable technologies must be reasonable, but more importantly, should become well defined over time so that risks and benefits may be properly assessed ahead of a project's start. In addition, the assessment of risks and benefits should become easier over time as the investments in digital infrastructure facilitate "bolt-on" software implementations and access to additional data sources (data warehousing), for instance, at the ERP and top layer of the operational hierarchy.

5.1 Added Value in the Control Hierarchy

From a fundamental point of view, one can argue that data-driven methods, as opposed to mechanistic methods, have a greater value proposition when processes appear to be complex and domain knowledge is lacking.

At the *regulatory control* level, it is unlikely that Big Data-based SPM concepts will be able to add further value to the established PID controllers individually. In the authors' experience, Big

Data and/or ML tools, including Performance Monitoring tools, may be able to identify inappropriate, undesirable and/or inadequate PID loop behavior, but, at least to-date, they are unable to provide causal insights for the same (i.e., interaction with other loops, nonlinear process behavior or just poor tuning). Thus, while PID control is, in principle, simple and, for many applications, straightforward to develop for experts and novices alike, domain knowledge (e.g., process understanding) is still required to trouble-shoot and resolve unacceptable loop performance. However, SPM methods that benchmark the performance of control loops [49] have a distinct value proposition when considering the large number of industrial applications that must be kept in check by fewer domain-knowledgeable personnel. Furthermore, the value at the regulatory level arises most likely from the ability of data-driven methodologies to identify relevant operational issues that may be otherwise difficult to detect on-line, for example, long-term structural or operational changes at a plant-wide level that affect certain clusters of regulatory control loops.

At the *supervisory and local optimization level*, the established technology is Model Predictive Control. MPC has a long track record of handling complex process dynamics. For the more complex situations (i.e., nonlinearities and state multiplicities), “workarounds” have been created and extensions implemented to enable the MPC to tackle a wide range of challenges. Thus, the value proposition of introducing SPM and ML at the “middle layer” is significant, if it improves dealing with complex process behaviors. In addition, the local optimization layer deals with hard/soft constraints, which requires on-going development. Nonetheless, even if purely data-driven controls overcome these deficiencies, they must outperform the current MPC controller if they are to be adopted in an industrial setting or provide the benefit of user-friendliness, because a changeover from the current MPC framework requires support staff to be retrained and the new set-up to be integrated into the control structure hierarchy. The value proposition of this change must outweigh the associated burden. A more practical and intermediate step is the hybridization of process control concepts of either the prediction (i.e., the replacement of the optimizer with an alternative technique) or the implementation part, that is the identification of unstructured process models, which could provide lasting value.

In comparison, at the *site-wide optimization and scheduling level*, the value proposition of SPM methods seems to be limited, as the decisions made at these levels are mostly well-defined and discrete. The current optimization algorithms (typically based on Linear Programming) are well suited for scheduling and optimization tasks, sometimes even considering the stochastic nature of the process. The value proposition of data-driven methodologies at these top layers therefore relies on their ability to pre-detect situations and scenarios that would have otherwise been deemed impossible to detect, or to incorporate unstructured information (e.g., contractual information) into the decision-making process. One relatively recent development is the reformulation of MPC as Economic MPC (EMPC) to integrate the real-time optimization hierarchy [85]. It would be important to monitor the progress of this technology as a viable alternative for higher-level decisions.

5.2 Added Value in the Development of the Control Structure

If automated and on-line model identification becomes feasible at all levels of the process control hierarchy, SPM and ML methods will minimize the requirement for data treatment and model identification that is currently carried out manually by process control engineers. Thus, such methodologies may reduce the time and cost required to develop and implement process control solutions. Furthermore, the process control structures that are developed from the regulatory control to optimization layers could benefit from concepts such as real-time model adaptation and tuning, which, in principle, would allow for better performance over time while providing the ability to handle a wider operating envelope. However, this could also be carried out as a monitoring task, so that a process control engineer can modify the models and tuning parameters as necessary.

Performance monitoring and diagnosis of process control structures throughout the hierarchy is another area that may benefit from SPM and ML methods. These methods can be used to effectively identify and "flag" controller structures that are not performing optimally and additionally identify time periods during which the process or a key process variable is out of the normal operating range.

Based on these observations, it can be seen that the future direction of integrating SPM and ML into closed-loop process control should focus on the development of complementary hybrid solutions where monitoring and control can be combined synergistically. This combined approach will have benefits at all levels of the process control hierarchy, but, in particular, at the supervisory and local optimization layers, both in terms of performance and effort required for implementation and on-going maintenance.

6. Practical Considerations

Apart from the technical hurdles that need to be overcome, other practical limitations may hinder widespread adoption and implementation (Figure 5). Therefore, this section sheds light on issues that must be resolved in order to implement Big Data-based closed-loop control concepts in practice.

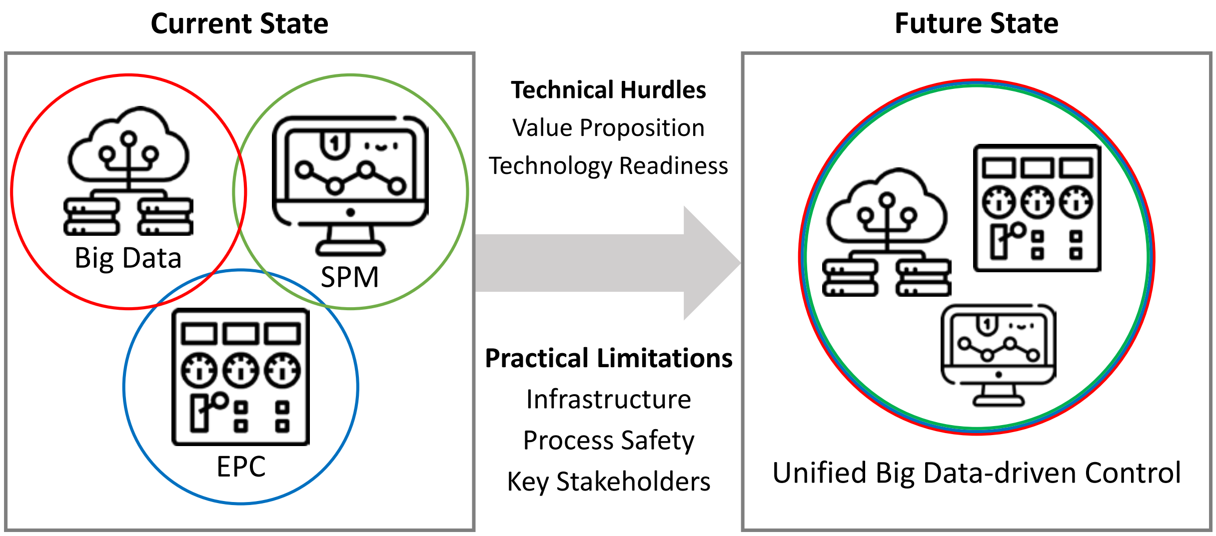

Figure 5. Practical limitations and hurdles in the implementation of data-driven closed-loop process control methods.

6.1 Operational Safety and Deterministic Behavior

A key "sticking point" that hinders the implementation of data-driven concepts is the aversion of process engineers to trust a concept they cannot inspect from the inside. This is because data-driven methods by nature represent grey-box or black-box models.

Both process control engineers and operators have very little understanding as to why a specific control action may be suggested by a data-driven methodology. In general, this is an issue when introducing any novel control concept. For example, the early MPC implementations were often abandoned as the operators neither understood nor agreed with the control moves made. The MPC's mechanistic underpinnings were too complex for the field control engineer and the operator to easily understand.

One way to convince and educate both engineers and operators of the benefits and reliability of new data-driven control concepts is to develop platforms, i.e., digital twins, on which their capabilities are demonstrated without putting the actual process at risk [86]. There is also a need for new curricula and courses to educate and provide continuing education of students, engineers and operators in Big Data tools and approaches.

6.2 Adding SPM Methods to Closed-loop Controlled Systems

The day-to-day operations of many (bio)chemical plants have a significant number of closed-loop controls at all levels of the control hierarchy. However, it has also been an industrial practice for some local optimizing control loops to operate in pseudo-closed-loop where the operators make changes based on information received from the process models. In order to capture the full benefit of SPM in (bio)chemical process operations, there is a clear need to identify and develop SPM methods that can operate within this framework. Both sparse PCA [16], where the SPM method operates on top of the regulatory and advanced regulatory control layer to provide valuable information, and controller performance monitoring concepts [49] are designed to complement the existing closed-loop controls. In both of these examples, the classic engineering based closed-loop controllers are augmented with the capabilities of the statistical methods. This type of combined thinking shows how a synergistic approach where statistical control concepts add value to the process control concepts is better than attempting to replace the existing engineering controls.

6.3 Big Data Readiness in the (Bio)Chemical Industries

In comparison to many other industries, the (bio)chemical industries have a proven track record of having sensors embedded in its processes and collecting large amounts of data. However, the data collected requires further structuring and pre-processing. Moreover, the use of data compression reduces the quality and the quantity of available process data, which diminishes their ability to identify relationships between phenomena of interest and potential process variables behavior, since it can destroy potential multivariate features in the data [87,88]. As a result of these deficiencies, a significant amount of pre-processing is required before these datasets can be used for real-time data-driven process monitoring.

This problem is further exacerbated by recipe data as well as other key pieces of information, which are sometimes only available as scanned documents or in paper form, which require an additional digitalization step. This lack of digital data has been an area of active focus in the (bio)chemical industry, where noticeable resources and personnel are devoted to continual data reviews and internal programs focused on educating key stakeholders about the benefits of having digital data. Furthermore, a few companies in the process industries are investing heavily in making process data accessible in the cloud, such that process chemists are given the opportunity to easily compare data across batches within one production site, but also across production sites. However, many corporations still prefer to keep process data in dedicated internal historian servers due to perceived cybersecurity concerns. This can, as a first benefit, yield a direct optimization potential by comparing and adjusting batch or fed-batch process performance across sites. Moreover, it finally results in a situation where ML and AI tools can be readily applied to the collected process data. Undoubtedly, the next step should then be closing the loop, and adjusting process set-points based on real-time information about the state of the process inferred from the data.

6.4 Convincing Key Stakeholders

In order for Big Data-based closed-loop controls to be implemented at any part of the control hierarchy, there is the need to convince multiple stakeholders within an organization that the solution does not just add value, but also provides an equal level of process safety as conventional solutions do. From a plant operations perspective, the key stakeholders in a control changeover are management, (technical and field) engineers and operators with often conflicting expectations and requirements.

In today's economic environment, plant managers can be convinced to implement Big Data-based closed-loop control as there is support for Industry 4.0 and Big Data topics from corporate management. However, plant management is generally risk averse and might opt to invest time and resources on other (Big Data) applications, e.g., predictive maintenance, where risk can be better managed.

Technical and field engineers must be convinced that they can reasonably fix and alter the Big Data-based closed-loop controllers if needed, as their primary role is to act as operations support and ensure an implemented solution is kept in operation over time. One key benefit of Big Data-based concepts is their adaptability to changing process conditions, e.g., due to catalyst deactivation. This ability to cyclically re-identify or modify models promises a better representation of the plant for longer periods of time without intervention from operations support. Hence, it may be argued that Big Data-based control loops will require less operations support than classical ones. Although, whenever operational support is needed, the current technical engineers might require additional help in the form of either a dedicated Big Data control support engineer or IT support. An additional requirement is further support over the solution's lifetime. Finally, development of training and continuing education [86] in Big Data concepts and tools is also a necessary requirement to support such approaches.

One should keep in mind that the primary task of the operator is to ensure that the plant operates within the safety limits and Big Data control concepts must not distract the operators from this primary responsibility.

7. Conclusions

Even though researchers and engineers have been working on data-driven techniques to improve process monitoring and control for decades, the increase in computing power and recent algorithmic developments in other business areas make for exciting possibilities in the Big Data domain. Methods like predictive maintenance are on the verge of large-scale deployment to real world (bio)chemical facilities, which illustrates the underlying potential. However, methods aiming at closing the loop still require development work as well as education/training to ensure workforce acceptance and the high degree of automation that is required to add sufficient value to warrant the technical and organizational changes.

Further, not all methods (e.g., in SPC) are equally suited to handle complex time-dependent closed-loop systems, requiring targeted development work. The process industries exhibit specific characteristics that need to be dealt with. (Bio)Chemical processes continuously generate lots of data, but they are information-poor. Most data points are serially correlated; *a priori*-knowledge and first principles have to be incorporated as to leverage these vast amounts of data, which underlines the need for hybrid approaches. Otherwise, it is difficult to imagine that purely data-driven techniques can fully compensate for the low information-value of process data and instill confidence among operations and management.

Nevertheless, the greatest value addition is expected in the middle layers of the control structure hierarchy, where model development is time consuming and expensive and softly-constrained optimization requires further progress. Machine Learning techniques may overcome these drawbacks and may excel in slowly and continuously changing environments. Moreover, the creation and implementation of easy-to-use capabilities, such as processing of datasets of *any* size and integrated non-linear model fitting, may also provide lasting value. Finally, for Big Data techniques to materialize in an organization and to reap the benefits, a key element is the creation of a data-centric organizational mindset. Only data of known value that is readily accessible will provide the basis for successful and long-living Big Data applications throughout an organization.

Acknowledgments

We kindly thank Dr. Ian Cameron and other anonymous reviewers for their supportive and insightful comments that significantly improved this manuscript.

References

(1) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529* (7587), 484–489. https://doi.org/10.1038/nature16961.

(2) Arzate Cruz, C.; Ramirez Uresti, J. A. Player-Centered Game AI from a Flow Perspective: Towards a Better Understanding of Past Trends and Future Directions. *Entertain. Comput.* **2017**, *20*, 11–24. https://doi.org/10.1016/j.entcom.2017.02.003.

(3) Renz, J.; Ge, X.; Stephenson, M.; Zhang, P. AI Meets Angry Birds. *Nat. Mach. Intell.* **2019**, *1* (7), 328–328. https://doi.org/10.1038/s42256-019-0072-x.

(4) Andersen, E. B.; Udugama, I. A.; Gernaey, K. V.; Bayer, C.; Kulahci, M. Big Data Generation for Time Dependent Processes: The Tennessee Eastman Process for Generating Large Quantities of Process Data. In *PROCEEDINGS OF THE 30th European Symposium on Computer Aided Process Engineering (ESCAPE30)*; 2020.

(5) Venkatasubramanian, V. The Promise of Artificial Intelligence in Chemical Engineering: Is It Here, Finally? *AIChE J.* **2019**, *65* (2), 466–478. https://doi.org/10.1002/aic.16489.

(6) Udugama, I. A.; Taube, M. A.; Mansouri, S. S.; Kirkpatrick, R.; Gernaey, K. V.; Yu, W.; Young, B. R. A Systematic Methodology for Comprehensive Economic Assessment of Process Control Structures. *Ind. Eng. Chem. Res.* **2018**, *57* (39), 13116–13130. https://doi.org/10.1021/acs.iecr.8b01883.

(7) Kagermann, H.; Wahlster, W.; Helbig, J. *Industrie 4.0*; 2013. https://doi.org/10.13140/RG.2.1.1205.8966.

(8) Reis, M. S.; Braatz, R. D.; Chiang, L. H. Big Data: Challenges and Future Research Directions. *Chem. Eng. Prog.* **2016**, *3*, 46–50.

(9) Qin, S. J.; Chiang, L. H. Advances and Opportunities in Machine Learning for Process Data Analytics. *Comput. Chem. Eng.* **2019**, *126*, 465–473. https://doi.org/https://doi.org/10.1016/j.compchemeng.2019.04.003.

(10) Box, G. E. P.; Luceño, A.; Paniagua-Quiñones, M. D. C. *Statistical Control by Monitoring and Feedback Adjustment*, 2nd ed.; Wiley, 2009.

(11) Cameron, I. T.; Seligmann, B.; Hangos, K. M.; Lakner, R.; Németh, E. The P3 Formalism: A Basis for Improved Diagnosis in Complex Systems. In *CHEMECA 2007: the 37th Australasian Chemical Engineering Conference*; Melbourne, Australia, 2007; pp 12–23.

(12) Lee, J.; Cameron, I.; Hassall, M. Improving Process Safety: What Roles for Digitalization and Industry 4.0? *Process Saf. Environ. Prot.* **2019**, *132*, 325–339. https://doi.org/10.1016/j.psep.2019.10.021.

(13) Svrcek, W. Y.; Mahoney, D. P.; Young, B. R. *A Real-Time Approach to Process Control, 3rd Edition*; Wiley, 2014.

(14) Kolberg, D.; Zühlke, D. Lean Automation Enabled by Industry 4.0 Technologies. In *IFAC-PapersOnLine*; 2015; Vol. 28, pp 1870–1875. https://doi.org/10.1016/j.ifacol.2015.06.359.

(15) Udugama, I. A.; Wolfenstetter, F.; Kirkpatrick, R.; Yu, W.; Young, B. R. A Comparison of a Novel Robust Decentralised Control Strategy and MPC for Industrial High Purity, High Recovery, Multicomponent Distillation. *ISA Trans.* **2017**, *69*, 222–233. https://doi.org/10.1016/J.ISATRA.2017.04.008.

(16) Gajjar, S.; Kulahci, M.; Palazoglu, A. Real-Time Fault Detection and Diagnosis Using Sparse Principal Component Analysis. *J. Process Control* **2018**, *67*, 112–128. https://doi.org/10.1016/j.jprocont.2017.03.005.

(17) Qin, J. S. Statistical Process Monitoring: Basics and Beyond. *J. Chemom.* **2003**, *17* (8–9), 480–502. https://doi.org/10.1002/cem.800.

(18) Yoon, S.; MacGregor, J. F. Fault Diagnosis with Multivariate Statistical Models Part I: Using Steady State Fault Signatures. *J. Process Control* **2001**, *11* (4), 387–400. https://doi.org/10.1016/S0959-1524(00)00008-1.

(19) Du, Y.; Budman, H.; Duever, T. A.; Du, D. Fault Detection and Classification for Nonlinear Chemical Processes Using Lasso and Gaussian Process. *Ind. Eng. Chem. Res.* **2018**, *57* (27), 8962–8977. https://doi.org/10.1021/acs.iecr.8b01110.

(20) A. Udugama, I.; Munir, M. T.; Kirkpatrick, R.; Young, B. R.; Yu, W. Side Draw Control Design for a High Purity Multi-Component Distillation Column. *ISA Trans.* **2018**, *76*, 167–177. https://doi.org/10.1016/j.isatra.2018.03.006.

(21) MacGregor, J. F. On-Line Statistical Process Control. *Chem. Eng. Prog.* **1988**, *84* (10), 21–

31.

(22) Vander Wiel, S. A.; Tucker, W. T.; Faltin, F. W.; Doganaksoy, N. Algorithmic Statistical Process Control: Concepts and an Application. *Technometrics* **1992**, *34* (3), 286–297. https://doi.org/10.1080/00401706.1992.10485278.

(23) MacGregor, J. F. [Statistical Process Monitoring and Feedback Adjustment]: Discussion. *Technometrics* **1992**, *34* (3), 273–275. https://doi.org/10.2307/1270030.

(24) Tucker, W. T.; Faltin, F. W.; Vander Wiel, S. A. Algorithmic Statistical Process Control: An Elaboration. *Technometrics* **1993**, *35* (4), 363–375. https://doi.org/10.1080/00401706.1993.10485351.

(25) Vining, G. Technical Advice: Statistical Process Control and Automatic/Engineering Process Control. *Qual. Eng.* **2010**, *22* (3), 222–224. https://doi.org/10.1080/08982111003800547.

(26) Capaci, F.; Vanhatalo, E.; Palazoglu, A.; Bergquist, B.; Kulahci, M. On Monitoring Industrial Processes Under Feedback Control. *Qual. Reliab. Eng. Int.* **2020**, *in press*. https://doi.org/10.1002/qre.2676.

(27) Van den Hof, P. M. J.; Dankers, A. G.; Weerts, H. H. M. Identification in Dynamic Networks. *Comput. Chem. Eng.* **2018**, *109*, 23–29. https://doi.org/https://doi.org/10.1016/j.compchemeng.2017.10.005.

(28) Morbach, J.; Yang, A.; Marquardt, W. OntoCAPE-A Large-Scale Ontology for Chemical Process Engineering. *Eng. Appl. Artif. Intell.* **2007**, *20* (2), 147–161. https://doi.org/10.1016/j.engappai.2006.06.010.

(29) Colegrove, L. F.; Seasholtz, M. B.; Khare, C. Big Data - Getting Started on the Journey. *Chem. Eng. Prog.* **2016**, *3*, 41–45.

(30) Spooner, M.; Kold, D.; Kulahci, M. Harvest Time Prediction for Batch Processes. *Comput. Chem. Eng.* **2018**, *117*, 32–41. https://doi.org/10.1016/j.compchemeng.2018.05.019.

(31) Lopez, P. C.; Udugama, I. A.; Thomsen, S. T.; Roslander, C.; Junicke, H.; Mauricio-Iglesias, M.; Gernaey, K. V. Towards a Digital Twin: A Hybrid Data-Driven and Mechanistic Digital Shadow to Forecast the Evolution of Lignocellulosic Fermentation. *Biofuels, Bioprod. Biorefining* **2020**, bbb.2108. https://doi.org/10.1002/bbb.2108.

(32) Montes, F.; Gernaey, K. V.; Sin, G. Implementation of a Radial Basis Function Control Strategy for the Crystallization of Ibuprofen under Uncertainty. In *Computer Aided Chemical Engineering*; 2018; Vol. 44, pp 565–570. https://doi.org/10.1016/B978-0-444-64241-7.50089-6.

(33) Esmonde-White, K. A.; Cuellar, M.; Uerpmann, C.; Lenain, B.; Lewis, I. R. Raman Spectroscopy as a Process Analytical Technology for Pharmaceutical Manufacturing and Bioprocessing. *Anal. Bioanal. Chem.* **2017**, *409* (3), 637–649. https://doi.org/10.1007/s00216-016-9824-1.

(34) Karimi, M.; Jahanmiri, A.; Azarmi, M. Inferential Cascade Control of Multi-Effect Falling-Film Evaporator. *Food Control* **2007**, *18* (9), 1036–1042. https://doi.org/10.1016/j.foodcont.2006.06.009.

(35) A. Udugama, I.; Alvarez Camps, M.; Taube, M. A.; Thawita, C.; Anantpinijwatna, A.; Mansouri, S. S.; Young, B. R.; Yu, W. Novel Soft Sensor for Measuring and Controlling

Product Recovery in a High-Purity, Multicomponent, Side-Draw Distillation Column. *Ind. Eng. Chem. Res.* **2019**, *58* (43), 20026–20035. https://doi.org/10.1021/acs.iecr.9b04594.

(36) Brüggemann, S.; Bauer, N.; Fuchs, E.; Polt, A.; Wagner, B.; Wulkow, M. Support of Strategic Business Decisions at BASF's Largest Integrated Production Site Based on Site-Wide Verbund Simulation. *Comput. Aided Chem. Eng.* **2008**, *25*, 925–930. https://doi.org/10.1016/S1570-7946(08)80160-5.

(37) Pontius, K. Monitoring of Bioprocesses Opportunities and Challenges Monitoring of Bioprocesses Opportunities and Challenges, Technical University of Denmark, 2019.

(38) Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S. N. A Review of Process Fault Detection and Diagnosis. *Comput. Chem. Eng.* **2003**, *27* (3), 313–326. https://doi.org/10.1016/S0098-1354(02)00161-8.

(39) Ge, Z.; Song, Z.; Gao, F. Review of Recent Research on Data-Based Process Monitoring. *Ind. Eng. Chem. Res.* **2013**, *52* (10), 3543–3562. https://doi.org/10.1021/ie302069q.

(40) Psarakis, S. A Review of Statistical Process Control and Engineering Process Control. In *Quality Control: Developments, Methods and Applications*; Nova Science Publishers, Inc., 2012; pp 115–139.

(41) Koutras, M. V.; Bersimis, S.; Maravelakis, P. E. Statistical Process Control Using Shewhart Control Charts with Supplementary Runs Rules. *Methodol. Comput. Appl. Probab.* **2007**, *9* (2), 207–224. https://doi.org/10.1007/s11009-007-9016-8.

(42) Wise, B. M.; Gallagher, N. B. The Process Chemometrics Approach to Process Monitoring and Fault Detection. *J. Process Control* **1996**, *6* (6), 329–348. https://doi.org/10.1016/0959-

1524(96)00009-1.

(43) Kramer, M. A. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE J.* **1991**, *37* (2), 233–243. https://doi.org/10.1002/aic.690370209.

(44) Dong, D.; McAvoy, T. J. Nonlinear Principal Component Analysis - Based on Principal Curves and Neural Networks. *Comput. Chem. Eng.* **1996**, *20* (1), 65–78. https://doi.org/10.1016/0098-1354(95)00003-K.

(45) Qin, S. J.; McAvoy, T. J. Nonlinear PLS Modeling Using Neural Networks. *Comput. Chem. Eng.* **1992**, *16* (4), 379–391. https://doi.org/10.1016/0098-1354(92)80055-E.

(46) Holcomb, T. R.; Morari, M. PLS/Neural Networks. *Comput. Chem. Eng.* **1992**, *16* (4), 393–411. https://doi.org/10.1016/0098-1354(92)80056-F.

(47) Doymaz, F.; Palazoglu, A.; Romagnoli, J. A. Orthogonal Nonlinear Partial Least-Squares Regression. *Ind. Eng. Chem. Res.* **2003**, *42* (23), 5836–5849. https://doi.org/10.1021/ie0109051.

(48) Kruger, U.; Dimitriadis, G. Diagnosis of Process Faults in Chemical Systems Using a Local Partial Least Squares Approach. *AIChE J.* **2008**, *54* (10), 2581–2596. https://doi.org/10.1002/aic.11576.

(49) Harris, T. J.; Yu, W. Controller Assessment for a Class of Non-Linear Systems. *J. Process Control* **2007**, *17* (7), 607–619. https://doi.org/10.1016/j.jprocont.2007.01.011.

(50) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2625.

(51) Tang, J.; Liu, J.; Zhang, M.; Mei, Q. Visualizing Large-Scale and High-Dimensional Data. In *25th International World Wide Web Conference, WWW 2016*; 2016; pp 287–297. https://doi.org/10.1145/2872427.288304l.

(52) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3* (29), 861. https://doi.org/10.21105/joss.00861.

(53) Joswiak, M.; Peng, Y.; Castillo, I.; Chiang, L. H. Dimensionality Reduction for Visualizing Industrial Chemical Process Data. *Control Eng. Pract.* **2019**, *93*, 104189. https://doi.org/10.1016/j.conengprac.2019.104189.

(54) Wu, H.; Zhao, J. Deep Convolutional Neural Network Model Based Chemical Process Fault Diagnosis. *Comput. Chem. Eng.* **2018**, *115*, 185–197. https://doi.org/10.1016/j.compchemeng.2018.04.009.

(55) Paoletti, M. E.; Haut, J. M.; Plaza, J.; Plaza, A. A New Deep Convolutional Neural Network for Fast Hyperspectral Image Classification. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 120–147. https://doi.org/10.1016/j.isprsjprs.2017.11.021.

(56) Munir, M. T.; Wilson, D. I.; Yu, W.; Young, B. R. An Evaluation of Hyperspectral Imaging for Characterising Milk Powders. *J. Food Eng.* **2018**, *221*, 1–10. https://doi.org/10.1016/j.jfoodeng.2017.10.001.

(57) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9* (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

(58) Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.;

Bengio, Y. Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*; 2014; pp 1724–1734. https://doi.org/10.3115/v1/d14-1179.

(59) Bengio, Y. Deep Learning of Representations for Unsupervised and Transfer Learning. In *JMLR: Workshop and Conference Proceedings*; 2011; Vol. 7, pp 1–20.

(60) Li, W.; Gu, S.; Zhang, X.; Chen, T. Transfer Learning for Process Fault Diagnosis: Knowledge Transfer from Simulation to Physical Processes. *Comput. Chem. Eng.* **2020**, *139*, 106904. https://doi.org/https://doi.org/10.1016/j.compchemeng.2020.106904.

(61) Wu, H.; Zhao, J. Fault Detection and Diagnosis Based on Transfer Learning for Multimode Chemical Processes. *Comput. Chem. Eng.* **2020**, *135*, 106731. https://doi.org/https://doi.org/10.1016/j.compchemeng.2020.106731.

(62) Luttmann, R.; Bracewell, D. G.; Cornelissen, G.; Gernaey, K. V.; Glassey, J.; Hass, V. C.; Kaiser, C.; Preusse, C.; Striedner, G.; Mandenius, C.-F. Soft Sensors in Bioprocessing: A Status Report and Recommendations. *Biotechnol. J.* **2012**, *7* (8), 1040–1048. https://doi.org/10.1002/biot.201100506.

(63) von Stosch, M.; Davy, S.; Francois, K.; Galvanauskas, V.; Hamelink, J.-M.; Luebbert, A.; Mayer, M.; Oliveira, R.; O'Kennedy, R.; Rice, P.; Glassey, J. Hybrid Modeling for Quality by Design and PAT-Benefits and Challenges of Applications in Biopharmaceutical Industry. *Biotechnol. J.* **2014**, *9* (6), 719–726. https://doi.org/10.1002/biot.201300385.

(64) Mosaad, M. I.; Salem, F. LFC Based Adaptive PID Controller Using ANN and ANFIS

Techniques. *J. Electr. Syst. Inf. Technol.* **2014**, *1* (3), 212–222. https://doi.org/10.1016/j.jesit.2014.12.004.

(65) Chang, W.-D.; Hwang, R.-C.; Hsieh, J.-G. A Multivariable On-Line Adaptive PID Controller Using Auto-Tuning Neurons. *Eng. Appl. Artif. Intell.* **2003**, *16* (1), 57–63. https://doi.org/10.1016/S0952-1976(03)00023-X.

(66) Moura, J. P.; Neto, J. V. F.; Ferreira, E. F. M.; Filho, E. M. A. On the Design and Analysis of Structured-ANN for Online PID-Tuning to Bulk Resumption Process in Ore Mining System. *Neurocomputing* **2020**, *402*, 266–282. https://doi.org/10.1016/j.neucom.2020.03.074.

(67) Kumar, V.; Gaur, P.; Mittal, A. P. ANN Based Self Tuned PID like Adaptive Controller Design for High Performance PMSM Position Control. *Expert Syst. Appl.* **2014**, *41* (17), 7995–8002. https://doi.org/10.1016/j.eswa.2014.06.040.

(68) Ricker, N. L. Decentralized Control of the Tennessee Eastman Challenge Process. *J. Process Control* **1996**, *6* (4), 205–221. https://doi.org/10.1016/0959-1524(96)00031-5.

(69) Bustos, G. A.; Ferramosca, A.; Godoy, J. L.; González, A. H. Application of Model Predictive Control Suitable for Closed-Loop Re-Identification to a Polymerization Reactor. *J. Process Control* **2016**, *44*, 1–13. https://doi.org/10.1016/j.jprocont.2016.04.011.

(70) Kheradmandi, M.; Mhaskar, P. Model Predictive Control with Closed-Loop Re-Identification. *Comput. Chem. Eng.* **2018**, *109*, 249–260. https://doi.org/10.1016/j.compchemeng.2017.11.016.

(71) Hajizadeh, I.; Rashid, M.; Sevil, M.; Brandt, R.; Samadi, S.; Hobbs, N.; Cinar, A. Adaptive

Model Predictive Control for Nonlinearity in Biomedical Applications. *IFAC-PapersOnLine* **2018**, *51* (20), 368–373. https://doi.org/10.1016/j.ifacol.2018.11.061.

(72) Akpan, V. A.; Hassapis, G. D. Nonlinear Model Identification and Adaptive Model Predictive Control Using Neural Networks. *ISA Trans.* **2011**, *50* (2), 177–194. https://doi.org/10.1016/j.isatra.2010.12.007.

(73) Prokhorov, D. V. Training Recurrent Neurocontrollers for Real-Time Applications. *IEEE Trans. Neural Networks* **2007**, *18* (4), 1003–1015. https://doi.org/10.1109/TNN.2007.899521.

(74) Littman, M. L.; Sutton, R. S.; Singh, S. Predictive Representations of State. In *Advances in Neural Information Processing Systems*; 2002; pp 1555–1561.

(75) Modayil, J.; White, A.; Sutton, R. S. Multi-Timescale Nexting in a Reinforcement Learning Robot. *Adapt. Behav.* **2014**, *22* (2), 146–160. https://doi.org/10.1177/1059712313511648.

(76) Gautier, N.; Aider, J.-L.; Duriez, T.; Noack, B. R.; Segond, M.; Abel, M. Closed-Loop Separation Control Using Machine Learning. *J. Fluid Mech.* **2015**, *770*, 442–457. https://doi.org/10.1017/jfm.2015.95.

(77) Aloupis, G.; Demaine, E. D.; Guo, A.; Viglietta, G. Classic Nintendo Games Are (Computationally) Hard. *Theor. Comput. Sci.* **2015**, *586*, 135–160. https://doi.org/10.1016/j.tcs.2015.02.037.

(78) Stephenson, M.; Renz, J.; Ge, X. The Computational Complexity of Angry Birds. *Artif. Intell.* **2020**, *280*, 103232. https://doi.org/10.1016/j.artint.2019.103232.

(79) Wang, Y.; Velswamy, K.; Huang, B. A Novel Approach to Feedback Control with Deep

Reinforcement Learning. *IFAC-PapersOnLine* **2018**, *51* (18), 31–36. https://doi.org/10.1016/j.ifacol.2018.09.241.

(80) Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; Hassabis, D. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518* (7540), 529–533. https://doi.org/10.1038/nature14236.

(81) Boiocchi, R.; Gernaey, K. V.; Sin, G. A Novel Fuzzy-Logic Control Strategy Minimizing N2O Emissions. *Water Res.* **2017**, *123*, 479–494. https://doi.org/10.1016/j.watres.2017.06.074.

(82) Kano, M.; Ogawa, M. The State of the Art in Advanced Chemical Process Control in Japan. In *IFAC Proceedings Volumes (IFAC-PapersOnline)*; 2009; Vol. 42, pp 10–25. https://doi.org/10.3182/20090712-4-tr-2008.00005.

(83) Kano, M.; Fujiwara, K. Virtual Sensing Technology in Process Industries: Trends and Challenges Revealed by Recent Industrial Applications. *J. Chem. Eng. Japan* **2013**, *46* (1), 1–17. https://doi.org/10.1252/jcej.12we167.

(84) Fujiwara, K.; Kano, M.; Hasebe, S.; Takinami, A. Soft-Sensor Development Using Correlation-Based Just-in-Time Modeling. *AIChE J.* **2009**, *55* (7), 1754–1765. https://doi.org/10.1002/aic.11791.

(85) Ellis, M.; Liu, J.; Christofides, P. D. *Economic Model Predictive Control*; Springer, 2017.

(86) Udugama, I. A.; Gernaey, K. V.; Taube, M. A.; Bayer, C. A Novel Use for an Old Problem:

The Tennessee Eastman Challenge Process as an Activating Teaching Tool. *Educ. Chem. Eng.* **2020**, *30*, 20–31. https://doi.org/10.1016/j.ece.2019.09.002.

(87) Klimkiewicz, A. Multivariate Statistical Process Optimization in the Industrial Production of Enzymes, Faculty of Science, University of Copenhagen, 2016.

(88) MacGregor, J. F.; Kourti, T. Statistical Process Control of Multivariate Processes. *Control Eng. Pract.* **1995**, *3* (3), 403–414. https://doi.org/10.1016/0967-0661(95)00014-L.

Table of Contents graphic