

Predictive Analytics Applications for Oil and Gas Processing Facilities

by

Elias A. Machado Roberty

M.S. Petroleum Engineering
Universidad del Zulia, 2019

B.S. Chemical Engineering
Universidad Rafael Urdaneta, 2011

SUBMITTED TO THE SYSTEM DESIGN AND MANAGEMENT PROGRAM
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN ENGINEERING AND MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2021

©2021 Elias A. Machado Roberty. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature of Author.
Department of System Design and Management
August 6, 2021

Certified by.
Richard D. Braatz
Edwin R. Gilliland Professor in Chemical Engineering
Thesis Supervisor

Accepted by
Joan Rubin
Executive Director, System Design and Management Program

[Page intentionally left blank]

Predictive Analytics Applications for Oil and Gas Processing Facilities

by

Elias A. Machado Roberty

Submitted to the System Design and Management Program
on August 6, 2021, in partial fulfillment of the requirements for the
degree of
Master of Science in Engineering and Management

Abstract

The oil and gas industry faces profitability and sustainability challenges that demand companies to have more efficient, reliable, safe, and environmentally friendly operations. Furthermore, as oil and gas companies embark on the *Industry 4.0* journey, the pillar of big data becomes increasingly important in an industry that generates massive amounts of data with low to no value extracted from it. Data are generated across all value chain sectors—upstream, midstream, and downstream—starting at reservoirs up to the finished products delivered by the refining and petrochemical sectors. Processing facilities across the value chain, where physical and chemical unit operations convert raw products into intermediate and finished products, generate a wealth of data through their heavily instrumented automatic control systems, operational routines, and quality control systems. Analyzing process data can help companies develop models that predict key process-related parameters to correct potential process upsets timely. In addition, predictive models can also be incorporated into digital twins to emulate diverse operating scenarios for production optimization or facility design purposes. This thesis investigates and reviews the application of predictive analytics on process data, its potential untapped value, analytics as an enabler of digital twins, and big data analytics frameworks tailored for an oil and gas context. Use cases across all segments of the value chain are reviewed with their respective predictive methods. The value of predictive analytics in oil and gas is assessed by reviewing various sources, including a major oil company success case followed by the architectural integration of predictive analytics into the development of a digital twin employing a systems-oriented approach. The last chapter discusses the predictive component of a novel approach tailored for process data analytics: Smart Process Analytics. The advantages such a framework offers versus standard automated predictive model development processes are discussed. Lastly, big data architectures for SPA implementation at process plants are developed.

Thesis supervisor: Richard D. Braatz

Title: Edwin R. Gilliland Professor in Chemical Engineering

[Page intentionally left blank]

Acknowledgments

I want to thank God, who has always supported me along each step of my life. And my grandfather, who now watches over me from heaven.

I want to express my deepest gratitude to my thesis supervisor, Professor Richard D. Braatz, for his guidance and generosity in sharing knowledge critical for my research. Your advice allowed me to develop a new perspective and skills in a field that I am passionate about, data analytics. Thanks for caring about my work and guiding me through each research step.

I also want to thank Chevron for allowing me to pursue my studies at MIT; it has been one of the most incredible experiences of my life. I would like to extend special thanks to Keith Johnston, Chevron Digital Engineering Manager. Our conversations laid the foundations for my research thesis; thanks for sharing ideas and helping me connect my research interests with the company's business needs.

My gratitude also goes to the MIT System Design and Management program and MIT Energy Initiative faculty and staff for their support. Special thanks to Joan Rubin and Diane Rigos for their guidance.

I also want to thank my parents, who have always cared for my education. Thank you for supporting and motivating me to pursue my dreams. My gratitude also goes to my beloved wife, María Emilia, for her unwavering support and patience. And I want to thank my brother, Luis, for being my thought partner.

Last but not least, I would like to thank the other 15 Chevron scholars for their support and friendship.

[Page intentionally left blank]

Contents

Chapter 1	13
Introduction.....	13
1.1. Research Purpose	13
1.1.1. Motivation.....	13
1.1.1.1 Research Questions	17
1.2. Literature Review.....	17
1.2.1.1. Predictive Modelling Approaches.....	18
1.2.1.1.1. White-box models	19
1.2.1.1.2. Black-box models	20
1.2.1.1.3. Gray-box models.....	24
1.2.1.2. Oil and gas industry	26
1.2.1.3. Systems Thinking.....	28
1.2.1.4. Object Process Methodology (OPM).....	29
1.2.1.5. Industrial Internet of Things (IIoT).....	31
1.2.1.6. Cloud Computing.....	33
Chapter 2	35
Use Cases of Predictive Analytics for the Process Industries.....	35
2.1. Prediction of Distillation Products Composition	35
2.2. Prediction of H ₂ S/SO ₂ Tail Gas Ratio in a Sulfur Recovery Unit (SRU)	38
2.3. Catalyst Performance	39
2.4. Emissions prediction.....	41

2.5. Polymers	44
Chapter 3	47
Value of Predictive Analytics Applications in the Process Industries.....	47
3.1. Beyond reliability and efficiency gains	48
3.2. The business case of Artificial Intelligence and Big Data for Oil and Gas	49
3.3. Value creation of predictive analytics in oil and gas processing facilities	51
3.4. Predictive analytics as a digital twin enabler	52
3.5. Developing the modeling and analytics enablement layer	55
Chapter 4	61
A Framework for Predictive Analytics of Process Data	61
4.1. Automated Machine Learning (AutoML) for process data	61
4.2. Tailoring the analytics model development to process data.	64
4.3. Open-source architectures for implementing SPA for oil and gas processing facilities.....	73
Chapter 5	79
Conclusion	79
5.1. Summary	79
5.2. Future work	80
Bibliography	83

List of Figures

Figure 1. Historical annual WTI and Brent benchmark spot prices.	16
Figure 2. Machine Learning Methods employed in Process Data Analytics.....	23
Figure 3. Oil and gas value chain.....	27
Figure 4. Key OPM symbols with some OPL examples.	31
Figure 5. Environmental soft sensor development workflow	44
Figure 6. AI and Big Data Market Value.....	50
Figure 7. Oil and gas digital twin Object-Process Diagram (OPD).....	54
Figure 8. Process data workflow	56
Figure 9. Predictive analytics object-process diagram (OPD).....	58
Figure 10. Popular open-source and commercial (marked with an *) AutoML tools.	62
Figure 11. A high-level view of the AutoML process.	62
Figure 12. SPA predictive modeling workflow. Adapted from Sun & Braatz (2021)	67
Figure 13. 4-fold nested CV	72
Figure 14. 5-fold time-series CV (aka forward-chaining CV).....	73
Figure 15. Big data functional layers	76
Figure 16. Lambda architecture for data processing.....	77
Figure 17. Kappa architecture for data processing	78

[Page intentionally left blank]

List of Tables

Table 1. Data analytics success stories in oil and gas companies.....	15
Table 2. Architectural variants of hybrid models.	25
Table 3. Maintenance requirements per type of analyzer	42
Table 4. Value of predictive analytics in major oil and gas company - upstream.....	51
Table 5 Value of predictive analytics in of major oil and gas company - downstream....	52
Table 6. Predictive analytics architectural decisions	59
Table 7. Methods used in SPA.....	69

[Page intentionally left blank]

Chapter 1

Introduction

1.1. Research Purpose

This section presents the motivation for studying predictive analytics applications in the oil and gas industry and describes the objectives in the form of research questions that the thesis answers in the successive chapters.

1.1.1. Motivation

Key actors in the Process Industries sector hold great expectations in light of the transition to the utopic Industry 4.0. Companies envision Industry 4.0 as a concept in which cyber-physical systems in interknitted and complex networks drive smart processing facilities' operation. The pillars necessary to realize the benefits of embracing Industry 4.0 are Cyber-Physical Systems (CPS), the Internet of Things (IoT), and Big Data. Such pillars interact synchronously in industrial settings to allow machines and humans to make real-time decisions for maintaining or enhancing a production process's value creation.

Cyber-physical systems (CPS) “comprise interacting digital, analog, physical, and human components engineered for function through integrated physics and logic” (Thompson, 2014). Such systems control and monitor mechanisms through computer-based algorithms capable of real-time decision-making. The Process Industries are abundant in cyber-physical systems in the form of industrial control systems such as Supervisory Control and Data Acquisition (SCADA) systems, Distributed Control Systems (DCS), and Programmable Logic Controllers (PLC).

As an enabler of CPS, IoT systems aim to facilitate and enhance the interactions between physical and virtual/cyberworlds (Milenkovic, 2020); such systems have the potential to significantly improve efficiency and quality for the Process Industries. Both CPS and IoT are not new concepts to the Process Industry community who has used them for the last 50 years to monitor, control, and optimize complex processes such as oil & gas and pharmaceutical drug manufacturing. The Process Industries branch dedicated to such efforts is referred to as Process System Engineering (PSE).

The third pillar, Big Data, is a product of two elements. First is the industry's computational power explosion in the last decades, as projected by Moore's law (Moore, 1965). Second, the increased data availability boosts the ever-growing use of sensors and automation in all facets of production processes. For example, recent studies report that a single oil well can generate 10 Terabytes (TB) of data per day and that a conventional offshore drilling platform can have approximately 80,000 sensors which may generate 15 Petabytes (PB) (equivalent to 1.000.000 Gigabytes) during its lifecycle (Western Digital, 2021).

The massive amounts of data generated by industrial processes, such as encountered in the process industries, open the door to the application of a vast array of data analytics techniques in its different domains, such as descriptive, predictive, and prescriptive analytics. This research work focuses on predictive analytics, specifically in the application of predictive analytics to process system engineering problems encountered in Oil and Gas Systems. Table 1, although not a comprehensive list, shows some success cases about the application of predictive analytics in oil and gas processing facilities.

This research work motivation is two-pronged. On one side is the need for energy companies to adapt and survive to more challenging market conditions. On the other side is the vast amount of value from analyzing data that remains untapped.

Concerning the first motivation, energy markets have not fully recovered from the oil price slump suffered in 2014, as illustrated in Figure 1. After reaching a peak of \$107, the West Texas Intermediate (WTI) marker plummeted to \$44.08, creating ripple effects across the entire energy value chain, with the upstream sector (exploration and production) suffering

the majority of the blow. Since then, energy markets have not returned to the price levels observed pre-2014.

Table 1. Data analytics success stories in oil and gas companies

Case 1 AkerBP - Cognite	Case 2 Petrobras
<ul style="list-style-type: none"> • \$6 million in annual cost savings through production optimization and reduced environmental impact • Use of machine learning and physics to improve water contamination detection 	<ul style="list-style-type: none"> • \$100 million in revenue gains in 2020 through production optimization projects • Artificial intelligence technologies deployed in refineries • Use of data lakes and cloud-based technologies
Case 3 Shell – Microsoft – C3.ai	Case 4 Aramco
<ul style="list-style-type: none"> • \$2 million in maintenance cost savings at the Pernis oil refinery • Enterprise-wide deployment of AI platform • Machine-learning algorithms used to predict equipment malfunction and provide early warnings 	<ul style="list-style-type: none"> • 50% reduction in flare emissions since 2010 thanks to the use of Big Data • Use of convolutional neural networks to monitor and analyze flame characteristics • 18,000 sources of data used to monitor and forecast future flaring

Note. Data for case 1 from Aker (2021), for case 2 Petrobras (2020), for case 3 from Council (2019), and for case 4 from Aramco (2021).

2020 brought another blow to energy markets fueled by diminished crude oil demand and the outbreak of the SARS-CoV-2 pandemic, disrupting the conventional dynamics of energy commodities exchanges. For the first time in history, the price of the WTI (West Texas Intermediate) futures contracts dipped into negative territory down to a minimum of minus \$37.63 (occurred on April 20th, 2020), signaling the need for energy companies to adapt and survive to a new reality, commonly referred to as “lower for longer.” Major oil and gas companies reacted by starting to look for ways to reduce capital and operating

expenses in the form of capital project budget reductions, leaner operations, higher efficiency, reliability, and safety.

All these efforts require a crucial ingredient: *improved decision making*, which needs a good deal of backend data analytics to derive valuable insights.

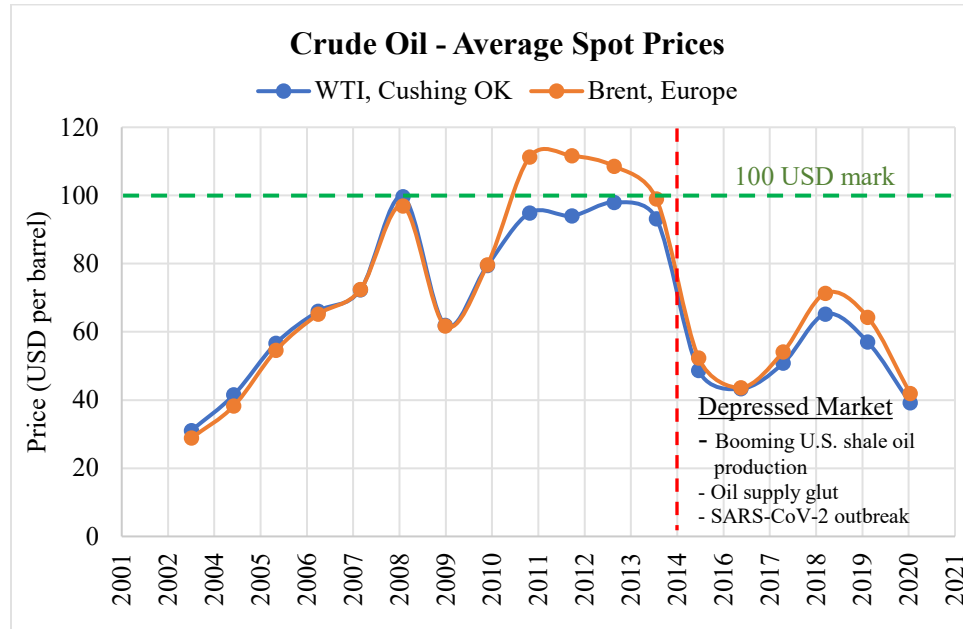


Figure 1. Historical annual WTI and Brent benchmark spot prices.

Here is where the second motivation of this research plays its part: the availability of a sea of data whose value remains untapped. Industry consultants and experts estimate that oil and gas companies use less than 1% of their operational data to aid the decision-making process (McKinsey & Company, 2020). Studies from leading companies in the IT sector, such as Google, support that number, estimating that energy companies only use around 5% of their data (Matthews, 2018). According to another report from Microsoft, one of the largest global mining companies uses only 1% of the data generated by approximately 6 million sensors (Microsoft, 2016). Given that the accuracy of models improves when trained with more data, oil and gas companies are leaving value on the table due to the massive amount of data that goes unanalyzed in such a heavily instrumented industry, ironically. Therefore, only when companies start treating data as an asset will the value creation possibilities be materialized.

In a nutshell, this thesis's motivation is to explore how the discipline of process system engineering can apply predictive analytics techniques to help boost the productivity and efficiency of oil and gas processing facilities.

1.1.1 Research Questions

This thesis seeks to answer several research questions about the application of predictive analytics to the process industries:

RQ1. What are the use cases of predictive analytics for the continuous processes commonly found in oil and gas processing facilities?

RQ2. What is the value created in employing predictive analytics to continuous processes?

RQ3. Are there any existing platforms or frameworks for implementing predictive analytics in oil and gas processing facilities? If so, which are best suited for the characteristics of the process data?

1.2. Literature Review

Data analytics span four different levels: descriptive, diagnostic, predictive, and prescriptive (van der Meulen & Rivera, 2014). Each level aims to answer questions of a different nature and demand a different amount and quality of data while delivering varying degrees of value. The data analytics levels are outlined below in order of increasing complexity and value creation:

1. Descriptive analytics operates at the lowest level of complexity and seeks to answer the question *What Happened?*. Its outputs require substantial human interpretation to transition from a data analysis phase to decision-making and, ultimately, execution phases.
2. Diagnostic analytics is concerned with the question *Why did it happen?*. It operates at a higher, more complex level than descriptive analytics, but with the potential of

- increased value delivery to the decision-making process, which reduced the amount of human input needed to interpret its outputs, make a decision, and execute it.
3. Predictive analytics is more complex than its diagnostic counterpart. However, it also offers more value and insight for the decision-making process since it allows the analyst to answer the question *What will happen?*. This level is more complex than diagnostic analytics; however, it allows the analyst to forecast the future and use that information in the decision-making process.
 4. Prescriptive analytics is the last and most complex level, and it aims to answer the question *What should I do?*. It is characterized by the addition of decision support and automation, therefore greatly reducing the amount of human input needed to transit from data analysis to decision making and ultimately action.

This research work is focused on the third level of analytics, predictive analytics, applied to the process industries, which is also a subset of the discipline known as Process Data Analytics which is defined as the application of data analytics tools to the problems commonly encountered in the process industries (Chiang and Braatz, 2020).

The application of predictive analytics to the process industries has gravitated mostly around inferential estimation to develop soft sensors, which soft sensors are predictive models that estimate a given process variable using other process measurements to facilitate process monitoring and control (Kadlec, Gabrys, & Strandt, 2009). It is also common to find other names for soft sensors in the literature, such as inferential sensors (Jordaan, Kordon, Chiang, & Smits, 2004; Qin, Yue, & Dunia, 1997) and virtual online analyzers (Han & Lee, 2002). The following sections dive deeper into the theoretical framework supporting the development of soft sensors.

1.2.1.1. Predictive Modelling Approaches

Predictive models are the cornerstone of predictive analytics as they provide a view into potential futures, known as outcomes, given a set of defined inputs. Such models are commonly derived by implementing supervised machine learning methods, which will be described in depth in the following sections. Models can be of three types, which in the

process industries are commonly referred to as white-, black-, and gray-box models.

1.2.1.1.1. White-box models

White-box models (aka mechanistic aka first-principles aka analytical models) are based on a well-defined and researched set of mathematical equations that describe the different states of a given system given a set of conditions (Chiang, Russell, & Braatz, 2001). Such models rely on heavily researched and proven theoretical laws and equations. In the specific case of chemical processes, white-box models describe the thermodynamics, kinetics, and transport phenomena (momentum, heat, and mass transfer mechanisms) of the unit operations.

These analytical models are not new to process systems practitioners, and have actually been used to varying degrees for more than 50 years. The literature abounds in examples of white-box approaches applied to process plants, for a wide variety of application use cases (Morari, Arkun, & Stephanopoulos, 1980). One use case of such analytical models is to estimate states, such as by use of a Kalman filter or Luenberger observer, from secondary measurements. The application of state estimation has been limited by the fact that typically the states for large process facilities are not observable from the measurements, which is a requirement for the state estimator to be implementable (Tham, Montague, Julian, & Lant, 1991). This requirement can be reduced by designing partial observers, which only estimate a subset of the states (Chang, Lin, & Georgakis, 2002). To design a partial observer, only the small subset of the states is required to be observable from the measurements. An example use case is where the unmeasured concentration of some key chemical needs to be controlled within a tight range. A partial observer to estimate only the concentration of the key chemical may be designable for processes in which some of the other concentrations are not observable.

Although white-box models have solid theoretical grounds, one of their main drawbacks is the failure to capture all the uncertainties that arise in real-world applications, which may generate inaccurate process variables prediction. Typically, first-principles models are based on assumptions that idealize the phenomena that they seek to describe so that the

mathematical equation can be solved for a given set of inputs. For example, typically, the fluid flow patterns are simplified rather than undergoing the high cost of formulating and running computational fluid dynamics software online. In practice, a white-box model may produce inaccurate predictions, especially when its inputs fall outside of the assumptions. Although white-box models can provide valuable information about a process, especially during equipment design, experts warn about the substantial amount of effort, time, and expertise required to develop and maintain such models over long-term industrial operation (Chiang et al., 2001; Fortuna, Graziani, Rizzo, & Xibilia, 2007).

Due to the limitations in using white-box models mentioned above, some other data-driven approaches like gray and black-box models are more widely used in online applications in the process industries.

1.2.1.1.2. Black-box models

Black-box models (aka data-driven aka empirical aka machine learning) are entirely constructed from process data. They are on the opposite spectrum of white-box models, with foundations that do not rely on physical or chemical relationships of the system but rather on computational algorithms and statistical relationships among input and output variables. Another key difference with white-box models is that they do not require a great deal of technical domain expertise to be constructed and trained. However, black-box models need datasets representative of the system states and modes of interest to generate consistent predictions.

Black-box models offer a rich set of methods that can be applied to use process data for inferring other key plant variables. However, “the proficiency of these methods is highly dependent on the quantity and quality of the process data” (Chiang et al., 2001, p. 7). In addition, the adjective “black-box” is attributed to the fact that these models are not as interpretable as white-box models. As their computational complexity increases, their interpretability decreases. One prime example is that of neural nets, which offer superior approximation capabilities at the expense of low-to-none interpretability. On the other hand, a linear regression model may offer greater interpretability at the expense of poor

predictive power. Obviously, this is just an example to illustrate the tradeoff between predictive power and interpretability typically encountered in black-box models. Furthermore, in the case of real-time applications such as inferential process control, these models must also be capable of operating at the same response rate as the other elements in the loop.

The theoretical background of black-box models is supported by machine learning and statistics. The machine learning domain is vast and constantly evolving, but can be decomposed into three main areas: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the algorithm “learns” or approximates function based on a map of inputs to outputs provided beforehand. In unsupervised learning, the task is not to approximate a function based on input-to-output relationships but rather to find patterns or associations between the observations provided. Lastly, in reinforcement learning, an agent is learning how to interact with its surrounding environment by executing actions and getting rewarded or penalized based on a policy; the intent is that agent the corresponding set of actions for maximizing the total reward for a specific policy (Alpaydin, 2014).

The research works developed by Sun and Braatz (2020) and Qin and Chiang (2019) provide an in-depth review of the most popular black-box models employed for process data analytics. The most popular methods used for predictive modeling rely on partial least squares regression (PLS-R), principal component regression (PCR), and neural networks, while some researchers have proposed the use of deep learning which is a much more sophisticated machine learning technique. Deep learning techniques must justify their lack of interpretation by providing greater predictive power; otherwise, the industry is reluctant to use them (Qin & Chiang, 2019). Other limitations of deep learning are that their training requires a very large quantity of data and deep learning is only better than standard neural networks at describing input-output relationships that are not smooth (Bresler & Nagaraj, 2021). Most input-output model relationships in which an online model would be developed are smooth, which is one of the reasons why neural networks are used to build black-box models for online applications by such companies as AspenTech and Rockwell Automation (Aspentech, 2021; Rockwell Automation, 2021). Black-box models applied to

the development of soft sensors in oil and gas processing facilities are the focus area of this study. These models have gained traction in the last 20 years, fueled by the ever-growing amount of data and computational power.

Figure 2 decomposes the three main machine learning areas into the methods mentioned by these authors.

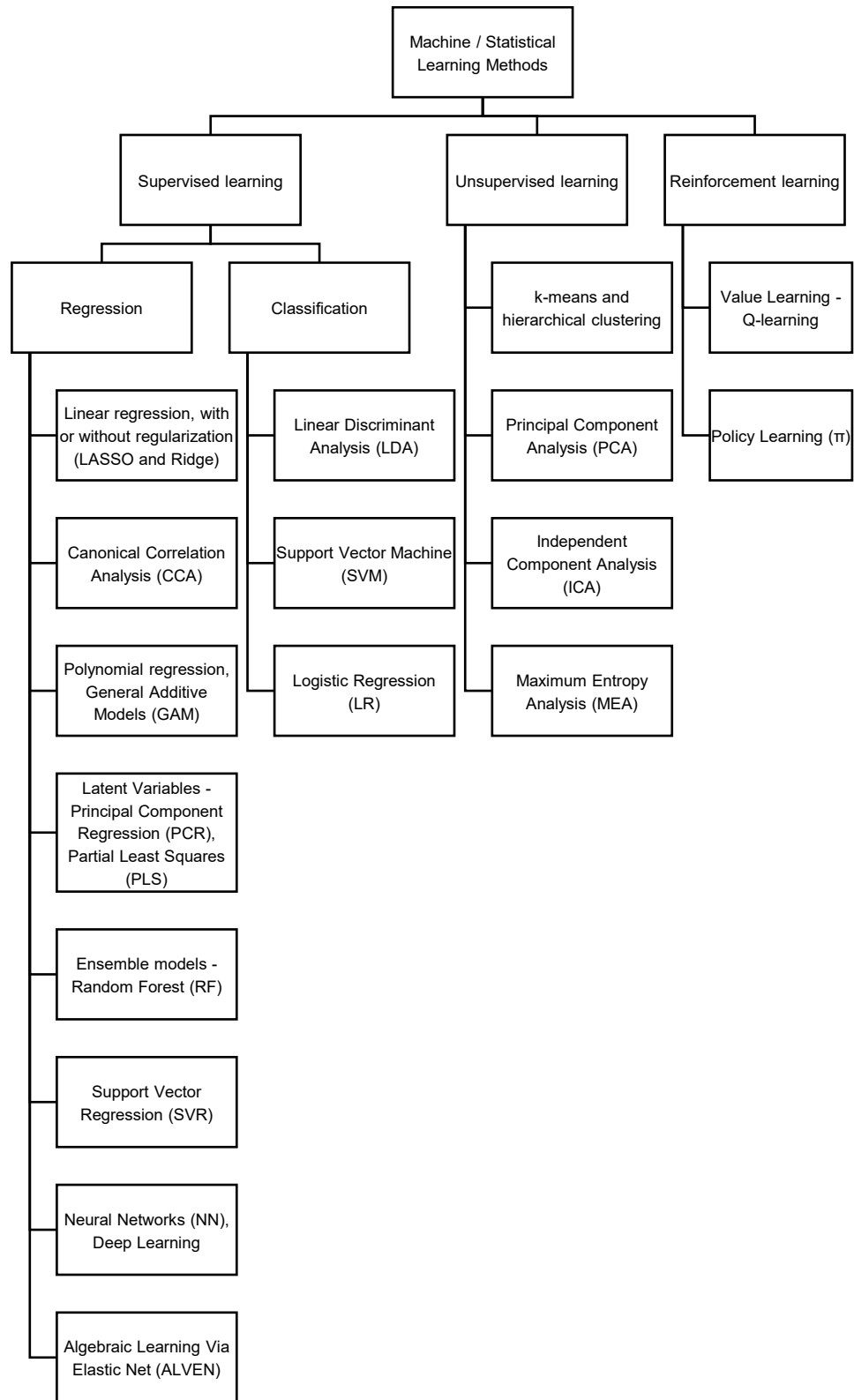


Figure 2. Machine Learning Methods employed in Process Data Analytics

1.2.1.1.3. Gray-box models

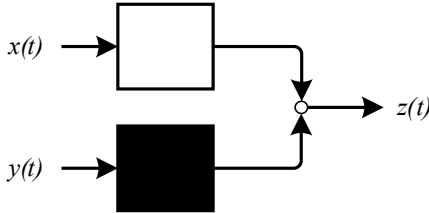
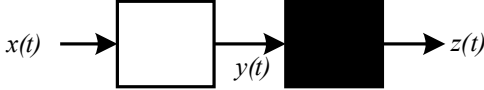
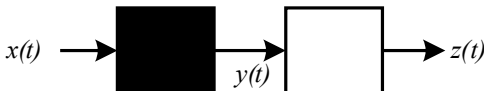


Gray-box models (aka hybrid models) are a combination of white- and black-box models. However, there is a deeper, more detailed mathematical reason why these models are considered a hybrid. von Stosch, Oliveira, Peres, and Feyo de Azevedo (2014) provided a holistic description of where hybrids fall within the distinct modeling techniques. They argue that model parametrization is the critical differentiator. In their view, there are two types of models: parametric and non-parametric. First-principles (or white-box) models fall within the parametric category because they have a fixed number of parameters, with a physical or chemical representation or meaning. On the other hand, data-driven (or black-box) models fall within the non-parametric category because “the number and nature of parameters are flexible and not fixed in advance by knowledge” (von Stosch et al., 2014). Under these criteria, hybrid models are categorized by the authors as hybrid semi-parametric since not all gray-box models are necessarily hybrid semi-parametric. With that description, the authors seek to differentiate hybrid semiparametric models from the broader gray-box family of models. The term “gray-box” can also encompass models that combine white-box, parametric submodels with empirical black-box, still parametric submodels. Such a model would still be an entirely parametric model in essence. Only hybrid semi-parametric models are considered in the development of this work.

The driving force behind the use of hybrid models is to provide an overall positive balance between predictive power and interpretability. Data-driven techniques are used to reduce the time-consuming task of developing entirely mechanistic models, by establishing the relationships that are challenging to model by first principles. First-principles models are used to describe those relationships that are well described by such models, to improve the model’s predictive power beyond the observations of the training and validation upon which data-driven models are developed.

With the possibility of combining parametric and non-parametric models also comes a variety of alternatives in which the two models can be combined. Research into potential ways to combine them is many decades old. For example, Agarwal (1997) studied several arrangements for combining neural networks with first-principles models, which he

referred as “prior models.” Agarwal proposed three arrangements: parallel, serial with the neural model before the prior model, and serial with neural model after the prior model. Such an arrangement would be further described in an article developed by von Stosch et al. (2014), especially the different ways to fuse the output of both models. Other experts have provided further details about the different schemas of sequential and hybrid/mixed models for predicting key variables (Aykol et al., 2021). Table 2 summarizes the architectural variants employed nowadays when developing hybrid models; further details can be found in the authors’ works.

Table 2. Architectural variants of hybrid models.

	<div style="display: flex; align-items: center;"> <div style="width: 20px; height: 20px; border: 1px solid black; margin-right: 5px;"></div> <div>White box model</div> </div> <div style="display: flex; align-items: center; margin-top: 5px;"> <div style="width: 20px; height: 20px; background-color: black; margin-right: 5px;"></div> <div>Black box model</div> </div>	Architectural variants
Parallel		<ul style="list-style-type: none"> • Superposition • Multiplication • Weighting of the predictions either by white- or black-box model
Sequential	 	<ul style="list-style-type: none"> • Residual or delta learning • Transfer learning¹ • Parameter learning
Hybrid/mixed	 	<ul style="list-style-type: none"> • Physics-constrained machine learning model • Machine-learning-accelerated physics-based model

¹ This term has a different meaning in the machine learning community.

1.2.1.2. Oil and gas industry

The oil and gas industry is comprised of three main segments: upstream, midstream, and downstream. These three sectors make up the oil and gas total value chain, starting with oil and gas exploration and ending with end-product sales from refineries and petrochemical process plants. The most critical activities along the value chain are exploration, extraction, processing (aka refining), and transportation of raw, intermediate, and finished products. The integrated oil and gas value chain is shown in Figure 3, with the most prominent activities at each segment.

In the upstream segment, exploration and production (E&P) activities take place. Typically, these activities include reservoir characterization, well location, drilling, completion, production, and preprocessing. These activities are highly specialized and deal with massive amounts of data necessary for efficient and profitable decision-making, given the uncertainty present in the reservoir production potential and the fluid properties. The preprocessing stage includes physical and chemical unit operations that aid in reducing contaminants and separating the water and gaseous phase from the incoming raw streams from the wells, which can be crude oil or gas, or both depending on the type of reservoir. Such processes include equipment like two and three-phase separators, electrostatic dehydrators, water knock-out or wash tanks, water treatment, and gas handling equipment.

Sequentially along the value chain, midstream lies in between upstream and downstream. The midstream segments cover the storage, transportation, and further processing of petroleum products, whether liquids or gases. Typical unit operations in midstream are networks of pipelines for either product distribution or gathering and their respective liquid-pumping or gas-compressing stations. Midstream assets connect the more remote production field areas with refining and petrochemical processing hubs and typically exhibit a wide geographical span because of the nature of their operations. Fluids processing facilities can also be encountered within the midstream segment, such as treatment plants and natural-gas-to-liquid plants (NGL). As expected, given its asset-heavy nature and complex operations topology, midstream generates seas of process data that can be further exploited for value creation.

The last segment in the sequence, downstream, encompasses the conversion of crude oil and gas into finished, more valuable products and their delivery to consumers worldwide. Its main constituents are the refining and petrochemical plants, where raw materials' conversion into finished products occurs. Typical refinery and petrochemical products are naphtha, gasoline, diesel, kerosene, jet fuel, propane, butane, liquified petroleum gas (LPG), among other fuel oils. Depending on the type of processing technologies employed at the refinery, specialty products and byproducts can also be produced, such as asphalt, aromatics, waxes, sulfur, pet coke, olefins, among others. Refineries are complex, highly interconnected facilities that intake the “raw” crude oil and split it into its constituent fractions through distillation as a first-pass operation. The various distillation intermediate streams are then sent to other highly specialized processing units for further conversion into finished products.

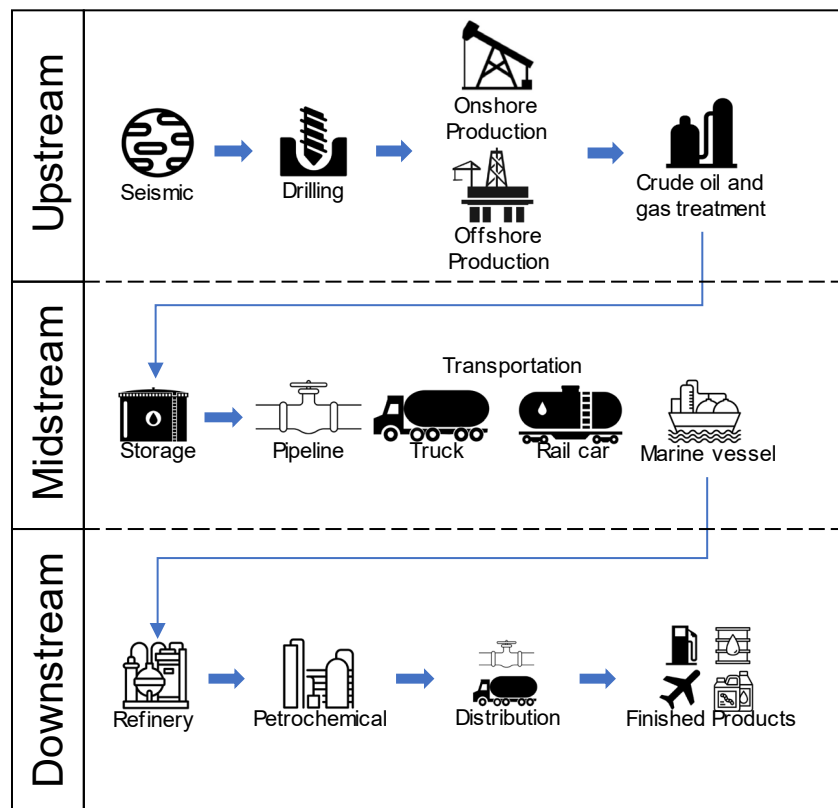


Figure 3. Oil and gas value chain

The finished products are then commercialized and distributed through the different sales

channels. Finished products such as gasoline, jet fuel, kerosene, and diesel are directly sold to consumers. In contrast, other products (or byproducts such as propylene and ethylene) from the refinery make their way to petrochemical plants to cover the demand for base chemicals (raw material for the specialty chemical sector) and plastics.

1.2.1.3. Systems Thinking

Crawley, Cameron, and Selva (2015) define systems thinking as “thinking about a question, circumstance, or problem explicitly as a system—a set of interrelated entities” (p. 8). The reason for introducing this line of thinking in this work is that the product or solution that arises from applying predictive analytics to manufacturing-related problems is in itself a complex system comprised of diverse interactions between software, hardware, and human entities. Therefore, it must be approached with the right set of principles, methods, and tools capable of holistically representing all the complex interactions between such entities.

System thinking encourages practitioners to approach the architectural synthesis and analysis of a system in four key steps outlined by Crawley et al. (2015):

1. “Identify the system, its form, and its function
2. Identify the entities of the system, their form and function, and the system boundary and context
3. Identify the relationships among the entities in the system and at the boundary, as well as their form and function
4. Identify the emergent properties of the system based on the function of the entities and their functional interactions” (p. 8).

It is outside the scope of this work to provide an in-depth explanation of systems thinking, principles, tools, and methods. Instead, this work considers the underlying foundation of the graphical and logical representations used in the following chapters to approach predictive analytics in a holistic system-oriented manner. The references provided in the previous paragraph are a good starting point for readers who wish to develop an in-depth

understanding of systems thinking.

1.2.1.4. Object Process Methodology (OPM)

Complex systems architectures contain vast amounts of information that must be adequately represented to prevent system users, designers, or architects from becoming overwhelmed and lose track of system interactions. When decomposing a complex system, the 7 ± 2 rule is employed. Each decomposition level is kept to a minimum of five and a maximum of nine entities. Such a rule is derived from the research of cognitive psychologist George A. Miller, who stated that the number of objects an average human can retain in short-term memory is 7 ± 2 (Miller, 1956). His work grew to be one of the most influential in the field of psychology, accounting for almost 34,000 citations per Google Scholar as of June 2021. System architects have therefore developed frameworks to represent systems with different views and projections. SysML (The Systems Modeling Language) and OPM (Object-Process Methodology) are two popular frameworks for representing and modeling complex systems.

The SysML Open Source Project defines SysML as a “general-purpose architecture modeling language for Systems Engineering applications” (2021, para. 1). SysML is a standard for Model-Based Systems Engineering (MBSE) that is an extension of a subset of UML (Unified Modeling Language). SysML provides practitioners with multiple views and diagrams to communicate and analyze different aspects of a system. Readers interested in knowing more about a SysML can refer to the SysML Open Source Project webpage or ISO 19514:2017(E) standard “Information technology – Object Management Group Systems Modeling Language (OMG SysML).”

OPM is another methodology to apply MBSE that was invented by Professor Dov Dori, who is a faculty member of both the Israel Institute of Technology and a Visiting Professor at the Massachusetts Institute of Technology (MIT). The first applications of OPM date back to 1993 in the automation of transforming hand-made drawings to CAD (Computer-aided Design) models. Since then, OPM would grow to be one of the leading

methodologies for the application of MBSE, recognized in 2008 by the International Council on Systems Engineering (INCOSE), and adopted as an ISO standard in 2015 under the name “ISO 19450—Automation systems and integration—Object-Process Methodology.” OPM is a methodology focused on three key elements: objects, processes, and their relationships. Such elements are used to represent the form and function of a system and its entities. Objects can be elements of form, such as instruments that enable a specific function (process) or operands, which convey information about the flow of things or state changes within the system architecture; objects communicate what the system is. Processes convey information about the system functions or what the system does.

In contrast to SysML, which consists of 13 diagrams or views, OPM provides a unified, single, integrated model. OPM provides both a graphical representation and a standardized language as part of the integrated model. For an in-depth explanation of OPM, see Dori (2016).

The graphical representation of OPM is the Object-Process Diagram OPD, and its language companion is the Object-Process Language (OPL). OPM is the methodology used in subsequent sections of this research to represent the predictive analytics architecture tailored for oil and gas processing facilities. Figure 4 illustrates key OPM symbols used in the development of OPDs with some examples.

In OPM, objects are considered “stateful,” each object can have states that may be affected by a system process (function). For example, the “pumping” function enabled by a centrifugal pump can change the state of a given fluid from “still” to “flowing;” in this case, the centrifugal pump is an object considered an instrument (it enables the “pumping” function). The fluid is another object (operand) upon which the “pumping” function is exerted.

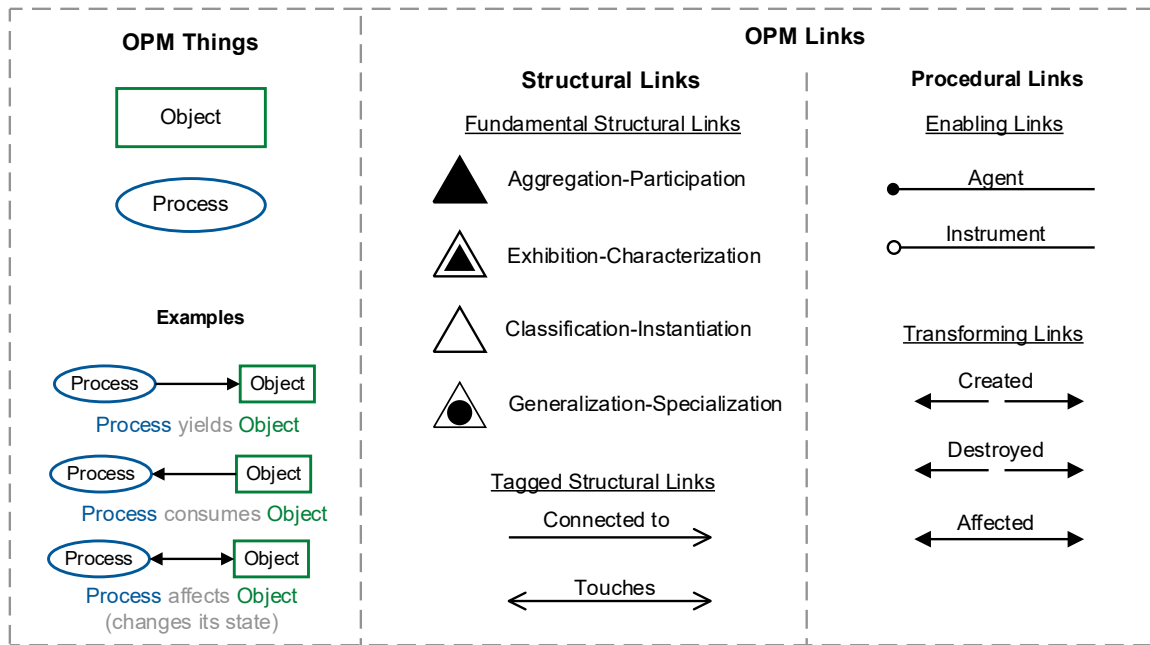


Figure 4. Key OPM symbols with some OPL examples.

1.2.1.5. Industrial Internet of Things (IIoT)

The branch dedicated to large-scale IoT concepts is commonly known as the Industrial Internet of Things and has seen advances in manufacturing and agricultural sectors fueled by the need to increase efficiency, lower cost, and improve safety. IIoT aims to bridge the gap between hardware (i.e., sensors, actuators, among others) scattered around the field and decision-makers (typically human actors) through networks connected to the internet. IIoT is essentially the same concept as a smart home, for example, but with the scalability requirements of industrial processes.

Process data are at the heart of any endeavor related to process monitoring and optimization (aka in some companies as PMO), predictive maintenance, process control, production planning, and scheduling, and, more recently, digital transformation initiatives such as digital twins. As the oil and gas companies embark on the Industry 4.0 journey, they need to morph and adapt their current data management frameworks and technologies to cope with Big Data's 3Vs: volume, velocity, and variety. Process plant personnel no longer only rely on structured (aka tabular) time-series data from a process historian, to name an example. Modern plants also generate other types of unstructured data such as video and

images (e.g., thermal imaging, flaring monitoring) to communicate process equipment conditions. IIoT offers end-to-end data management capabilities, all of the way from field data generation to remote decision-support centers. Operating companies, services contractors, and technology vendors are gearing their organizations, capabilities, products, and services towards a data-centric strategy. It is now becoming prevalent to see entire oil fields and offshore platforms being monitored and operated from remote decision centers, thereby reducing personnel exposure to hazardous areas and harsh weather conditions while preventing potential process downtime.

However, the concepts embedded within the IIoT paradigm are not new to the oil and gas industry, which has relied on automatic control systems connected to field sensors and actuators for decades. Some of the most popular systems have been Supervisory Control and Data Acquisition (SCADA) and Distributed Control Systems (DCS). More recently, around the 2000 decade, the industry has also exhibited an increased interest in wireless field sensors, which have seen some difficulties for full deployment given concerns about cybersecurity, data transmission, and reliability (especially around safety-critical functions). In 2020, the International Organization for Standardization (ISO), in conjunction with the International Electromechanical Commission (IEC), published the standard TR 30166:2020 “Internet of things (IoT) – Industrial IoT,” which lays the foundations for implementing IIoT by covering its technical aspects along with its architectural elements. Such standards have the vital role of providing a signpost for companies (such as oil and gas ones) companies to deploy IIoT with a lifecycle approach in mind.

In closing, IIoT is a critical enablement layer for predictive analytics, by facilitating the generation, collection, transmission, storing, and processing of data. Furthermore, IIoT increases access to data by connecting the production process to the internet. All of these steps are requisites to create value from big data. Analytics cannot be fully deployed at an industrial scale without the IIoT foundational layer. By bringing the data to the experts instead of vice versa, personnel can focus on using precious time to extract value from data instead of finding themselves wrangling among different data management systems (maintenance databases, process historians, among others).

1.2.1.6. Cloud Computing

Cloud computing refers to the execution of computational routines and services over the internet. Microsoft (2021) defines cloud computing as “the delivery of computing services—including servers, storage, databases, networking, software, analytics, and intelligence—over the Internet (“the cloud”) to offer faster innovation, flexible resources, and economies of scale.”

Like IIoT, cloud computing is another pillar necessary to deploy predictive analytics at an industrial scale. The advantage provided by cloud computing lies in its distributed architecture which handles data and computations in different clusters throughout the network. This pooling of computational resources enables more efficient data processing and predictive model development on massive amounts of data.

Leading companies in the IT arena are now offering cloud computing platforms. Some of the most popular are Microsoft’s Azure, Amazon Web Services, and Google Cloud. Although the cloud service models are very diverse, there are three that have become prominent and work incrementally:

- Infrastructure-as-a-service (IaaS): a foundational resource layer with its scope oriented towards providing data storage, security, and computing resources over the internet.
- Platform-as-a-service (PaaS): the services provided by IaaS plus resources oriented at web applications developed over the internet.
- Software-as-a-service (SaaS): the services provided by PaaS plus licensing, distribution, and management of software applications over the internet.

The different cloud service models offer flexible levels of responsibility between the service provider and a client. SaaS is the model that puts most of the responsibility for the operation and maintenance of cloud resources over the service provider. A key differentiator of cloud services is that they are charged on a *pay-as-you-go* basis.

Both IIoT and cloud computing are concepts in constant change and still have much room

for improvement—especially around cybersecurity. Recent studies state that cybersecurity threats affect the following cloud computing system ilities: availability, reliability, integrity, and confidentiality, and that the use of cloud computing does, in fact, increases the likelihood of attacks happening (Sasubilli & R, 2021). Some popular examples of cyberattacks are distributed denial of service (DDoS), data theft, malicious insider, ransomware, phishing, nefarious use of cloud computing, the man in the middle (MITM), and flooding attacks (a type of denial of service attack). Each of the aforementioned cloud service models exhibits more or fewer security issues, making them vulnerable to diverse types of cyberattacks. Sasubilli & R (2021) provide an in-depth development of the security issues, threats, and countermeasures for each type of cloud service model.

In the context of the oil and gas industry, compromising the safety of operating personnel and the integrity of process equipment due to cyberattacks is plain and simple not acceptable. Process plants deal with massive amounts of energy and hazardous materials that, if uncontrolled, can lead to catastrophic accidents affecting the personnel directly involved in the facility operation and any surrounding community and the environment. In addition, the process industries as a whole are plagued with nefarious process safety incidents capable of having continent-wide ripple effects, such as the Chernobyl Nuclear Plant for an example.

Cyberattacks have already hit the oil and gas industry. For example, in May 2021, a group of hackers known as “DarkSide” executed what has been considered one of the most harmful ransomware attacks on energy infrastructure. The group of hackers shut off the 5500-mile Colonial Pipeline, the largest fuel pipeline in the United States, cutting the flow of around 2.5 million barrels per day of fuel along the East Coast and sending gasoline prices to \$3 per gallon due to supply shortages (Tidy, 2021). In closing, until security concerns around IIoT and cloud computing are fully addressed in the near future, oil and gas companies will not fully develop trust in deploying big data technologies in processing facilities.

Chapter 2

Use Cases of Predictive Analytics for the Process Industries

This chapter presents the extent of the applications of predictive analytics to continuous processes and the evolution and refinement of the various machine learning methods used in chronological order. The information presented in this chapter serves to provide order and structure to the sea of literature about machine learning by categorizing it into each specific use and can be utilized as a framework to investigate which predictive technique may be more suitable to a specific unit operation. Most of the use cases of soft sensors (aka inferential models) gravitate towards estimating variables subject to large measurement delays or high sample times; for such cases, the operator's decision-making process would greatly benefit from more frequent estimation.

2.1. Prediction of Distillation Products Composition

The documented application of soft sensors to distillation units is ample and rich in the diversity of techniques applied. Distillation is one of the most widely known and studied chemical unit operations. It is present in all oil and gas industry sectors, without exception, and has served as the backbone of value creation by transforming and refining a wide array of feedstocks into products such as gasoline, jet fuel, diesel, gas oil, to name a few. Given its critical role in many industrial processes, several experts have developed and implemented soft sensors for estimating key quality parameters such as distillate and residue purity using other easy-to-measure process variables provided by hard sensors.

One of the most popular applications of soft sensors is to use principal component regression (PCR) and/or partial least squares (PLS) estimators to predict the product

composition of a multicomponent distillation column from temperature measurements (Mejdell and Skogestad, 1991). Nonlinearities plague the distillation data employed to train the model in the column. The nonlinear dependencies between product compositions and temperatures are addressed by applying logarithmic transformations to the data before building the estimators. For the case study in of Mejdell and Skogestad (1991), the calibration/training dataset was developed using a nonlinear simulator of the distillation column, and the techniques were applied both to binary and multicomponent cases. A 7-fold cross-validation procedure was employed using the mean squared error of prediction (MSEP), also known as the mean squared error (MSE), as the performance evaluation criterion. The hyperparameter used in their procedure was the number of k principal components in the models. Limiting the number of principal components reduces the sensitivity to errors or noise in the data.

In a related research study, Tham et al. (1991) employed two different approaches to develop soft sensors, which they called “adaptive estimators.” The first approach consisted of an adaptive estimation using an input-output process representation, while the second used a state-space representation. Both techniques were demonstrated over the inferential control of a distillation column leading to satisfactory results.

Kresta (1994) also observed good predictive power of PLS regression in another distillation case study with a dataset of highly correlated variables. An interesting aspect of this study is the emphasis given to training the model using a representative dataset that contains the typical noise levels likely to be found in the final inferential control scheme. In this case, the PLS model also performed well under missing data and sensor failure, which are problems likely to be encountered in the industry.

Kano, Miyazaki, Hasebe, and Hashimoto (2000) revisited the static PLS estimator work performed by Mejdell and Skogestad in 1991, mainly to develop ways to cope with the deterioration of the PCR estimator under changes in the feed composition, which indicated the need for a dynamic regression estimator considering time-series data. Like their predecessors, they also relied on cross-validation to select the optimal number of latent variables, using the MSEP and explained variance of prediction (EVP) as guiding criteria.

Kano et al. developed a dynamic PLS estimator using time series data plus other variables such as flow rates, heat duties, and pressures to predict top and bottom product compositions. Their model generated successful results when deployed under a cascade inferential control scheme, using temperature for the inner loops and product composition on the outer loops.

Another alternative approach was the employment of a locally weighted regression (LWR) combined with PCA and PLS for dimensionality reduction of the input variables (Park & Han, 2000). The LWR was used to deal with even stronger nonlinearities than those studied by Mejdell and Skogestad in 1991 and Kresta in 1994. Once more, cross-validation was employed to determine the optimal number of principal components. A key distinction from other regression methods is that LCR seeks to generate a smooth function of the predictor variables by fitting a linear function using weighted least-squares (WLS). The weights are calculated by considering the distances of the Q nearest neighbors of a given predictor variable (x_i). For the detailed procedure, see the author's paper. Their method proved to be a good alternative when the dataset contains nonlinearities and collinear variables.

Another application of soft sensors to distillation columns was provided by Fortuna et. al. (2007), where a nonlinear moving average (NMA) model was applied to estimate the stabilized gasoline in the overhead product of a debutanizer column and the butane content in the bottoms residue. Fortuna provides a rich perspective on applying several regression methods to select the number of optimal input variables for the model and regressor estimation. Methods for input variable selection, that were considered are simple correlation, partial correlation, linear Mallow statistics, nonlinear Mallow statistics, PLS, and NPLS. For regressor estimation, they employed partial-least squares (PLS), nonlinear PLS (NPLS), and a neural model of multilayer perceptrons (MLP). The neural models using latent variables as inputs resulted in the highest correlation coefficients or the highest predictive power, followed by the NPLS model using latent variables as inputs.

In the last decade, many other researchers have demonstrated the capabilities of different structures and types of neural nets for soft sensor development. Their application comes as

no surprise due to the popularity and funding levels that Machine Learning-related science has acquired after the so-called “second wave” of AI referred by DARPA (Launchbury, 2020). Such applications include rather sophisticated machine learning methods; for example, deep learning has been demonstrated for the prediction of heavy diesel 95% cut point of a crude distillation unit (CDU) (Shang, Yang, Huang, & Lyu, 2014).

2.2. Prediction of H₂S/SO₂ Tail Gas Ratio in a Sulfur Recovery Unit (SRU)

Applications of predictive analytics to Sulfur Recovery Units (SRU) have also been documented in the available literature (Fortuna et al., 2007). For an SRU, Fortuna et al. estimated the concentrations of H₂S and SO₂ in the unit’s tail gas compositions employed a nonlinear moving average (NMA) model. The authors employed four alternatives for estimating the regressors of the NMA model: radial basis function neural networks (RFBFNN), multi-layer perceptron (MLP), neuro-fuzzy networks (NF), and nonlinear least square fitting (NLSQ). The resulting model was required for an online control implementation, so the computational burden of each alternative was critical in model selection. An SRU is an environment-critical process plant whose function is to remove hydrogen sulfide (H₂S) and ammonia (NH₃) from gas streams from refinery process units or gas plants. H₂S is removed from the gas by conversion to molten Sulphur (S₂), while NH₃ is destroyed by conversion to N₂ and O₂. The ratio of H₂S and SO₂ in the tail gas is critical to control, and indicates how efficiently combustion stoichiometry is being maintained in the reaction furnace (also known as the Claus furnace). Details about the Claus sulfur recovery process can be found in many sources throughout the internet (Sulfur Recovery Engineering, 2021). Fortuna et al. (2007) found that the MLP model provided the best performance during its online verification phase. Also, the RBF network provided satisfactory results but was considered too computational demanding for real-time application.

2.3. Catalyst Performance

Data-driven models have also been developed to predict catalyst performance. Kadlec and Gabris (2011) developed a local learning-based soft sensor to predict catalysts' activity in polymerization reactors. Their purpose was two-fold: (1) to demonstrate that it is possible to cope with the degradation of the sensor model by incorporating adaptive capabilities, and (2) to help a manufacturing company better predict catalyst activity for enhanced process operation. The authors proposed a new adaptive algorithm named the Incremental Local Learning Soft Sensing Algorithm (ILLSA). The idea behind ILLSA is to overcome the drawbacks of the recursive PLS (RPLS) method, such as its susceptibility to the selection of parameters like forgetting factors and adaption windows length, which can severely affect the soft sensor performance. In the ILLSA, statistically different "receptive fields" are developed from historical process data, then a "local expert" (a model) is trained for each field. In this case, the authors used PLS, given the collinearity present between the process variables. Each local experts (predictor) generates predictions of process variables; such predictions are then combined into a final prediction employing a weighted sum under a Bayesian framework. The ILLSA provided better predictions than the RPLSE technique using the Mean Squared Error (MSE) as an indicator for a dataset with nonlinear relationships and collinearity.

Another example of catalyst activity prediction is that provided by Gharehbaghi and Sadeghi (2016), who developed a gray-box model to estimate the deactivation of zeolite catalysts used in the petrochemical industry for disproportionation (DP) of toluene and transalkylation (TA) of C₉ aromatics and toluene to C₈ aromatics in fixed-bed reactors. The authors refer to their model as data-based mechanistic (DBM) with state-dependent parameters (SDP). The specific type of SDP model employed is the state-dependent dynamic auto-regressive with exogenous (SDARX) model, which allows to estimate parameters that vary with time and vary with other system variables. The performance of their model was assessed on process data from a petrochemical complex, obtaining satisfactory results.

More recently, Steurtewagen and Van den Poel (2020) applied machine learning

techniques for estimating catalyst saturation levels of a Fluid Catalytic Cracking (FCC) Unit. Their model served as a soft sensor used for optimizing the processing unit leading to increased performance, specifically a higher yield and less catalyst consumption. The goal of an FCC unit is to recover higher valued, lighter hydrocarbon from heavier hydrocarbon refinery streams such as the vacuum residue from a distillation unit; in doing so, a catalyst is used to promote the cracking reactions of the heavier molecules into lighter ones. In that process, the catalyst experiences regeneration cycles to burn the carbon deposits generated during the cracking reactions; however, it is not possible to burn all the carbon deposits depending on the operating scheme of the FCC. For this specific case, the authors developed a data-driven soft sensor following the methodology proposed by Kadlec et al. (2009). Their catalyst saturation prediction would ultimately be used as an input to an optimization model for prescriptive analytics purposes, demonstrating the combination of the two analytics domains can be combined to squeeze more value from process data.

The workflow employed by Steurtewagen and Van den Poel (2020) consisted of data gathering, cleaning, and treatment for outliers. After that, and given the high correlation among the process variables in their dataset, they decided to apply a dimensionality reduction technique previous to regression. The dimensionality reduction technique employed was PCA, which allowed them to capture 71% of the variance with only 12 variable groups, down from 136 original variables in a dataset containing 300 million sensor readings approximately. As is commonly the case with process data, their dataset had measurements of variables sampled at different frequencies; therefore, the authors had to employ a multi-rate aggregation technique to bring all variables to a standard polling rate. Concerning the regression part, the authors tested three algorithms: linear regression (as baseline), Random Forest (Breiman, 2001), and XGBoost (Chen & Guestrin, 2016). The authors compared the three models on the basis of the coefficient of determination (aka R^2) and the root mean squared error (RMSE) using a cross-validation procedure. The optimal hyperparameters of these models were found through a grid search algorithm in the cross-validation step. The Random Forest model gave the best results in terms of R^2 and RMSE. The online deployment of the model resulted in satisfactory results that

allowed the plant operator to optimize the feedstock flows to the FCC to increase overall unit utilization. Furthermore, the model mitigated the reliance on manual catalyst sampling and analysis procedures highly influenced by the operating crew performing the task, thus reducing operator-induced bias.

2.4. Emissions prediction

Prediction of greenhouse gas emissions from oil and gas processing facilities has been another area of active research for predictive analytics. Greenhouse gas emissions are an area of concern for all oil and gas processing facilities, given the increased pressures from regulators and the community for emissions-free operations. The growing concerns of global warming led a coalition of countries to sign the Paris agreement on November 4th, 2016, to limit global warming to 1.5 degrees Celsius below pre-industrial levels. Such an agreement signals regulators' increased focus on the private and public companies to curb emissions levels (United Nations Framework Convention on Climate Change, 1992).

Most of the emissions in oil and gas processing facilities come from process gas flaring during process upsets and startup or shutdown operations. The emissions sources are typically stationary ones, such as furnaces, boilers, and flare stacks. Greenhouse gases (GHG) like carbon dioxide (CO₂), sulfur oxides (SO_x), and nitrogen oxides (NO_x) are present in the streams released into the atmosphere. Oil and gas companies employ continuous emission monitoring systems (CEMS) to detect, measure, and act upon such emissions on time. However, the CEMS depends on complex analytical sensors (aka hard sensors), which require intensive maintenance and re-calibration. These sensors are also prone to improper installation procedures that may compromise the accuracy of the reading. Recent studies from industry experts, such as AMETEK Process Instruments, suggest that CEMS are among the most complex type of analyzers commonly encountered, requiring a dedicated organization just for analyzers maintenance purposes. Table 3 summarizes their findings, and it can be observed that CEMS, which fall under the environmental category, require on average 2.5 man-hours per week per analyzer.

Moreover, other researchers have found the total initial cost of a CEMS (direct plus indirect

cost) to be in the range of \$195,000 to \$366,000 (Chien et al., 2003). In the United States, soft sensors developed to serve as mirrors of the CEMS also have to comply with the Environmental Protection Agency (EPA) technical requirements, mainly that emissions monitoring systems must be online for at least 95% of the time, demanding high reliability during the soft sensor lifecycle (Qin et al., 1997). Researchers have applied predictive analytics to develop soft sensors that provide faster measurements that allow operators to make timely adjustments for all the reasons stated above.

Table 3. Maintenance requirements per type of analyzer

Complexity factor	Category	Type of analyzer	Estimated man-hours per week
1~5	Simple	pH, conductivity, gas detection, O ₂	2
6~8	Physical property	Boiling point, flash point, freeze point, RVP, and viscosity	3
9	Environmental	CEMS, SO ₂ , CO, H ₂ S, and opacity	2.5
10~15	Complex	Tail gas, gas chromatography, mass spectrometry, NIR, and FTIR	4

Note. Adapted from Misfer et al. (2009).

The data-driven models used for predicting GHG emissions are commonly known as predictive emission monitoring systems (PEMS). They have an established operational performance in places like the US and Europe. EPA owns PEMS standards in the US, and European Committee for Standardization (CEN) owns such standards in Europe. The history of PEMS in the US dates back to 1990, and recent studies estimate that there are more than 300 PEMS installed throughout the country as a backup system to their CEMS counterparts (Si, Tarnoczi, Wiens, & Du, 2019). To mention a few examples, PREMS haven been deployed to predict SO₂, CO, NO, O₂, flue gas flow, and particulate emissions in refineries, petrochemical, and power generation plants. A wide array of predictive models has also been studied in PREMS development, from purely first-principles techniques to data-driven, including their combination as hybrid models. The next paragraphs describe some of the data-driven approaches to the development of PREMS.

Dong, McAvoy, and Chang (1995) predicted the NO_x content emitted by an industrial heater from a dataset with five days of operational history and 45 variables sampled every minute. The authors fully exploited the predictive power of techniques like PLS, PCA, and neural networks (NN) by combining them in a way where each technique complemented the shortcomings of the others. A neural network PLS (NNPLS) model was employed to model the nonlinearities associated with the process variables, while a nonlinear PCA model was used to detect a faulty sensor (input data) and reconstruct its missing values. The idea behind an NNPLS technique is that the PLS part deals with variables collinearity, and the neural network captures the nonlinear relationships.

Dong et al. (1995) followed four main steps to develop their predictive emissions model: data pre-treatment, variable selection, sensor validation with missing sensor detection, and model building. The workflow the authors followed, along with the modeling techniques, is illustrated in Figure 5. The authors demonstrated that a self-validating soft sensor performed well in detecting erroneous data, reconstructing it, and ultimately predicting the NO_x emission of the industrial heater.

The work by Dong et al. (1995) was key in advancing the development of predictive emissions monitoring systems (PEMS), which are a particular type of soft sensors for predicting GHG emissions that must comply with the strict reliability requirements mentioned above. In 1997, Qin et al. expanded Dong's research by refining the hard sensor validation and missing sensor detection outlined in Figure 5. Their work consisted of developing a self-validating inferential sensors (SVIS) framework based on PCA for detecting, identifying, and reconstructing faulty input sensors. Their framework was tested with both single and multiple (sequential faults) input sensor failures, giving satisfactory results. However, Qin et al. (1997) arrived at the conclusion that the optimal prediction scheme is to have a hard analyzer backing up several inferential sensors, just in case model extrapolation occurs, compromising the accuracy of the predicted results, in this way, maintenance resources can still be reduced by avoiding the installation of several analyzers throughout the plant. In their prediction model, a neural network PCR (NNPCR) provided better predictivity than a linear regression model given the nonlinear relationships present in the data. Complimentarily, collinearity was simultaneously addressed by the PCR part

of the neural network.

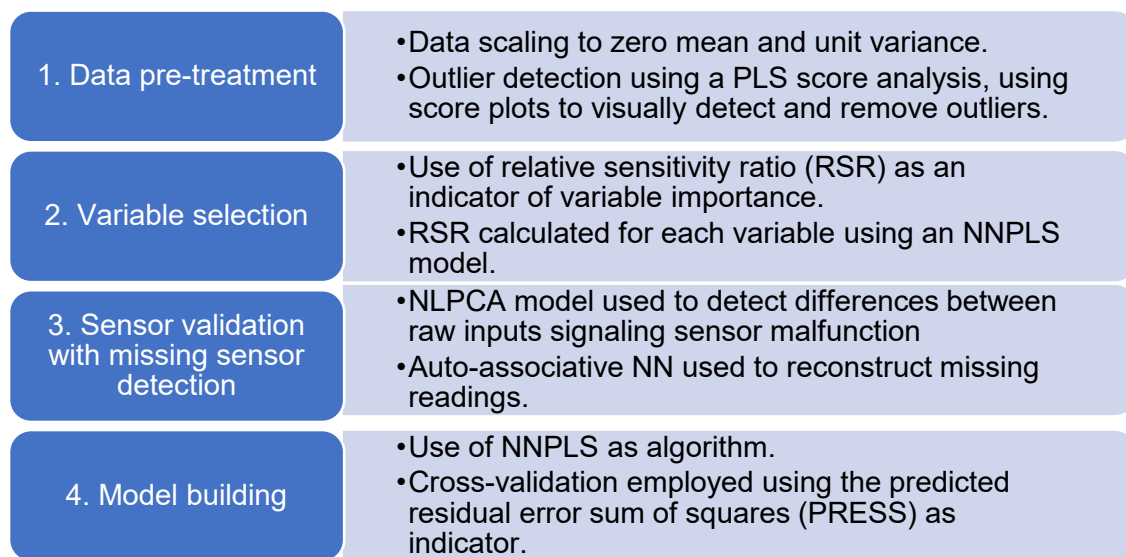


Figure 5. Environmental soft sensor development workflow

More recently, Si et al. (2019) developed a PEMS for a cogeneration unit employing the open-source machine learning library Keras using its Python and R interfaces. The authors randomly selected five feedforward neural networks, which were trained using both the Adam and Nadam gradient descent optimization algorithms. All their models resulted compliant with EPA standards. The authors considered the 32-64-64-64 neural network trained with the Nadam algorithm to be the best model in terms of its MAE and correlation with the real values measured by an existing CEMS.

2.5. Polymers

The polymer segment typically falls under the chemical industry's grasp, which consumes products from the oil and gas value chain. However, some companies, depending on their market strategy, size, technical and commercial capabilities, include polymer production within their petrochemical segment scope. Some examples are integrated oil and gas companies like Saudi Aramco, Chevron, Shell, and Total. Therefore, this section will touch on some applications of predictive modeling in the area of polymers production.

Predictive modeling has been employed for the development of soft sensors that can

accurately predict quality variables like melt index, polymerization, and conversion rates (Kordon, 2020). One example was demonstrated by Schley et al. (2000), who employed an extended Kalman filter (KFC) incorporated in a nonlinear quadratic dynamic matrix control (NL-QDMC) algorithm to control for the number average molecular weight (NAMW) and polydispersity of a styrene polymerization reactor. Their approach, however, is heavily dependent on an existing process model that can accurately represent the system states. In other words, their approach is not data-driven but rather first-principles-driven.

Polypropylene and polyethylene production processes have been the target of soft sensor applications in the quest for higher profitability, reliability, and efficiency. The work of Ge et al. (2011) exemplifies the application of data-driven approaches for predicting quality variables in polymerization processes, specifically the melt index in polypropylene production. The authors employed a type of regression not commonly seen in the development of soft sensors, Gaussian Process Regression (GPR), on the grounds of its suitability to model the process nonlinearities. The method is named combined local GPR (CLGPR) since the standard GPR method is extended to model multiple operation modes. A key advantage of using GPR is that it can provide uncertainty ranges for its predictions and model parameters given its Bayesian basis. The authors also employed PCA to address dimensionality reduction and multicollinearity. The method yielded satisfactory results when applied using both data from the process plant and laboratory analyses (for the melt index), and yielded better results than PLS, SVR, and ANN, using RMSE as the accuracy indicator for that specific dataset. However, the model was not tested online.

Another development in the propylene production arena is that of Diaz et al. (2011), who developed a soft sensor to control the polymerization reactor and its product properties. In their work, the predicted properties are the melt index (MI) and the xylene extractable fraction (XS). The data-driven model employed a particle filter (PF) and an artificial neural network (ANN) in a two-step sequential fashion. The PF is first employed to address uncertainty and generate the latent variable that will later serve as input to the ANN. Secondly, the ANN estimates the MI and XS properties. The polymerization reactor is of the continuous stirred tank reactor type used for bulk polymerization at high pressures. The author's model yielded satisfactory results reporting R-squared values higher than 0.9 for

both the validation and test sets after employing cross-validation (CV) for the training phase of the ANN. However, R-squared values do not represent the predictive power of a model entirely. In addition, the simple hold-out cross-validation procedure may not yield the best results when the amounts of sample points or observations are small, even when the model yields high (>0.9) R-squared values in the test set.

Kaneko and Funatsu (2014) present a comprehensive framework for the development of sensors by combining Support Vector Regression (SVR) with time-series modeling methods. The authors' objective is to counter the predictive power decay that soft sensors experience once operating conditions or process modes experience sudden changes. The proposed method is named Online Support Vector Regression (OSVR). To demonstrate its effectiveness, OSVR was compared with other popular methods like PLS embedded in moving window, time differences and, distance-based, correlation-based, and locally weighted just-in-time methods. The OSVR method was assessed in terms of R-squared and RMSE. The authors employed the radial basis function as the kernel function for their SVM model. The model was applied to a polymerization reactor in which the predicted variables were the melt flow rate (MFR) and density. The set of predictor variables encompassed 38 process variables representative of temperatures, pressures, and monomer and comonomer concentrations. The OSVR model was reported to be the most accurate among the techniques mentioned above and was able to handle time-varying nonlinearities involved in the process.

Chapter 3

Value of Predictive Analytics Applications in the Process Industries

Although the development of soft sensors has proliferated in the process industries, the documentation of the value created by its implementation is not widely captured in the literature. Only some specific examples for some use cases are provided. This section, therefore, seeks to investigate in depth the potential value creation from implementing predictive analytics techniques in the process industries beyond just the apparent benefit of estimating process quality variables. This research hypothesizes that there is much more value from the implementation of predictive analytics at a plant-wide systems level than just to specific use cases.

Value creation in the process industries may manifest by improving and optimizing several system ilities such as efficiency, reliability, safety, and sustainability. Nowadays, sustainability has taken an important place in the process industries, especially those processes dealing with the transformation and processing of fossil fuels, due to the increased societal concerns about the greenhouse gas (GHG) emissions generated by such processes along their value chain (upstream, midstream, and downstream). Therefore, sustainability should be weighted when incorporating predictive analytics techniques into existing manufacturing processes.

Another aspect of value creation that will be covered is the intersection of predictive analytics and digital twins. One of the most prominent aspects of smart manufacturing, as posed by Industry 4.0, is the adoption of digital twins that can recreate a wide array of plant operating modes and conditions to allow for accelerated decision-making. Soft sensors hold a privileged position in enabling the successful generation of digital twins for modeling cyber-physical systems.

3.1. Beyond reliability and efficiency gains

Some authors have quantified the value of implementing soft sensors by relating their capital and maintenance costs to their corresponding hard sensors. For example, the utilization of soft sensors for estimating GHG emissions from process plants is normally performed by special analyzers whose acquisition cost ranges from \$100,000 to \$200,000, with a maintenance cost of approximately \$15,000/year (Dong et al., 1995). However, for the case of continuous emissions monitoring systems (CEMS), this research argues that there is far more value untapped. For example, the prevention of costly environmental fines from regulatory agencies because of process excursions beyond the allowable thresholds due to an analyzer failure. Since 2000 the US Environmental Protection Agency (EPA) has entered into 37 settlements with U.S. energy companies to reduce annual emissions of nitrogen oxides (NO_x) and sulfur dioxides (SO_2) by 95,000 and 260,000 tons by year, respectively. Those settlements represent 95% of the US refining capacity and translate into expenses of approximately \$7 billion in control technologies (United States Environmental Protection Agency (EPA), 2013). Control technologies imply tighter control schemes and systems, of which soft sensors are therefore a critical component.

Furthermore, some authors have captured the benefits of soft sensors by measuring the reduction of operating costs (OPEX) after implementing inferential control strategies. One such example is that of Kim et al. (2013), who presented a 0.6% OPEX reduction in a cracked gasoline fractionation unit (CGL) used in ethylene production after successfully implementing a soft sensor to predict the concentration of aromatics in the cracked gasoline overhead product (CGL). The savings came from optimizing the energy consumption (heat duty) required by the cracking furnace to successfully keep the concentration of the aromatics in the CGL product within contract quality ranges. The concentration of aromatics would be typically analyzed once a day in the laboratory. In turn, the implementation of a locally-weighted-PLS (LW-PLS) model allowed for real-time control of the cracking furnace coil outlet temperature (COT). Although a 0.6% OPEX reduction for a distillation unit may seem low initially, it is necessary to consider that distillation units can represent 50% of the plant-wide operating costs of fractionation facilities (Lee, Son, Lee, & Won, 2019). Assuming that the ethylene production process where the soft

sensor was deployed uses ethane cracking technology, other studies estimate production costs in the range of \$400 to \$750 per ton of ethylene produced (Ren, Patel, & Blok, 2008). By making those connections, the compounding effect of a successful implementation can translate into savings of \$1.2 to \$2.2 per ton (assuming that distillation accounts for 50% of the production cost) on the overall production cost of the processing facility for this specific use case. The next section expands the evaluation of the potential value of predictive analytics for oil and gas using a market-wide approach.

3.2. The business case of Artificial Intelligence and Big Data for Oil and Gas

Assigning a dollar value to the application of predictive analytics to oil and gas processing facilities is a significant endeavor for which specific data is very scarce, especially around soft sensor implementation. However, a signpost of such potential value can be provided by looking at the overall market value of Artificial Intelligence (AI) and Big Data in the oil and gas industry. The information presented in this section aims to communicate the order of magnitude of such value instead of a precise, absolute number.

As illustrated in Figure 6, there is an increasing trend for the AI and Big Data market value for the oil and gas industries. The worldwide value for AI covers elements such as machine learning, data science, robotic controls, among others. The Big Data component covers data acquisition, processing, management, and potential insights from analyzing such data.

Per Figure 6, the AI market for oil and gas is projected to reach \$2000 million by 2024 at a compound annual growth rate (CAGR) of 10% approximately. On the other hand, the Big Data market is forecasted to be around \$3900 million by 2024, with a CAGR of 15%. The surveyors of such reports factored in the value of analytics both in AI and Big Data markets. Although these numbers cannot all be attributed to the value of analytics exclusively, they signal the ever-growing importance of improving data analytics in the oil and gas landscape, signaled by the increased financial commitment of major companies across the entire spectrum of the value chain. It is important to point out that, although the market data was gathered from different sources and analyzed by different companies, they

all project a sustained increase in the value of AI and Big Data for oil and gas.

After examining the value of analytics at a high level, more market-oriented perspective, the following section will zoom into more specific value creation cases for oil and gas companies.

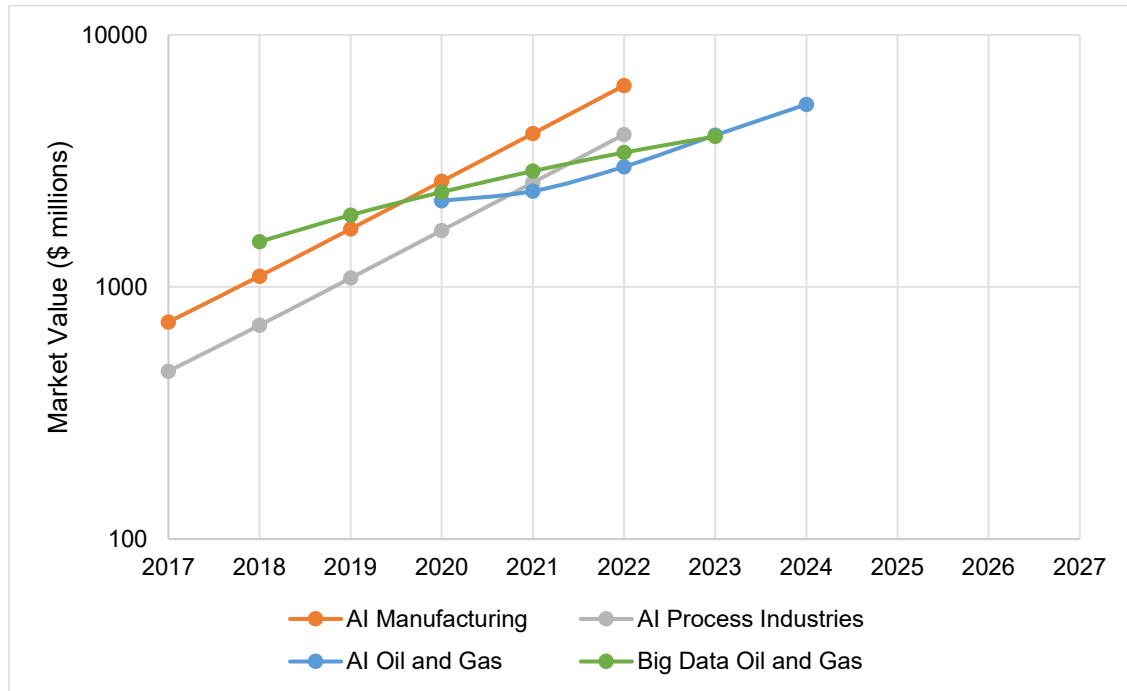


Figure 6. AI and Big Data Market Value. Note. This figure illustrates the market value on a log scale in the vertical axis. Data for AI Manufacturing and AI Process Industries from Technavio (2018), for AI Oil and Gas from GlobalData (2021), and for Big Data Oil and Gas from Technavio (2019).

Furthermore, market struggles for the oil and gas industry also affect a non-negotiable, core requisite of processing facilities: safety. Surveys from industry experts present a correlation between periods with lower oil prices and periods where a series of catastrophic events occur, such as those experienced around 1986, 2008, and 2014. A survey conducted in 2015 revealed that the 100 largest losses between the period 1974 to 2015 in the oil and industry accounted for more than \$33 billion in damages, with upstream, downstream (refineries), and petrochemical representing 87% of the lost value; and the most common even type being explosions (Marsh, 2016). The same survey shows a serial correlation

between low oil prices and accident occurrence. It is widely known that correlation does not imply causation (much less in safety where accidents are outcomes of complex interactions among system entities besides purely financial causes). However, such a trend may suggest relationships between cost-saving initiatives triggered by depressed market conditions and compromised safety safeguards (hardware, software, and human actors). Besides use cases tied to environmental and reliability aspects, data-driven predictive approaches can also help prevent the occurrence or escalation of hazardous initiating events in process plants by alerting the process operator on time about potential process upsets. Process monitoring is an area of predictive analytics which specialized in detecting such abnormal patterns in process states.

3.3. Value creation of predictive analytics in oil and gas processing facilities

In some specific examples of value creation through the application of predictive analytics in the oil and gas sector, several companies have reported the reduction of operating costs thanks to improved efficiency and reliability in their processing facilities. In recent years, one of the major integrated oil and gas companies consulted by the author reported that predictive analytics generated value across the entire value chain. Table 4 presents the value generated in the upstream segment, and Table 5 presents the same information for downstream.

Table 4. Value of predictive analytics in major oil and gas company - upstream

Year	Value Created (\$ million)
2019	23
2020	15

Table 5 Value of predictive analytics in of major oil and gas company - downstream

Facility	Number of inferential sensors in operation	Number of optimizers in operation	Value created ^a (\$ million / year)
Refinery 1	100+	40+	+40
Refinery 2	110+	45+	+42

^aAverage value generated in the years 2018, 2019, and 2020

The financial benefits presented in Table 4 originate from process system engineering opportunities where analytics techniques were employed to optimize production or predict equipment breakdown proactively. These values originate from tangible projects deployed at major oil and gas company's processing facilities distributed around the world.

The numbers presented in Table 5 represent the financial benefits generated from projects implemented in refineries of a major oil and gas company in the United States. Such projects generate value through production optimization employing soft sensors embedded in model predictive controls (MPC) that push the process plant as close as possible to the product specifications that provide the highest value. Production optimization is achieved through various channels: throughput maximization (capacity utilization), improved product yields, increased recovery of more valuable products, and reduced energy consumption. Some of the product properties predicted by these soft sensors are flash, Reid vapor pressure (RVP), and distillation cut specifications. The predictive models are developed using other process variables or measurements such as flows, temperatures, and pressures. Such models are periodically adjusted online using lab results or analyzer to correct any model decay or deterioration.

3.4. Predictive analytics as a digital twin enabler

With the increased interest from the private sector in adopting digital twins for enhanced decision-making, predictive analytics has taken a spotlight in the digital transformation journey of the process industries. The so-called digital twins can be thought of as regression models on steroids, that is, an entire complex system modeled using first-principles and

data-driven techniques fed by the massive volumes of big data generated at processing plants. Relevant to this discussion is some recent research from the Massachusetts Institute of Technology (MIT) on the digital twin system concept (LeBlanc, 2020). The author proposed a digital twin concept tailored for oil and gas facilities and operations, employing the taxonomy laid out in the International Organization for Standardization (ISO) 14224:2016 standard (2016). The author's digital twin concept was decomposed into four main entities:

1. Enablement layer
2. Modeling & analytics
3. Data foundations
4. Physical entities

Dori (2016) defines a sociotechnical system as “a system that integrates technology, people, and services, combining perspectives from engineering, management, and social sciences.” Digital twins and their respective entities are inherently sociotechnical systems, given the interactions between human operators (or decision-makers), hardware, and software. A systems-oriented approach can be employed to thoroughly characterize, develop and analyze a predictive analytics system concept embedded within the overall digital twins framework, or in other words, to approach a predictive analytics concept as the so-called “system of systems.” This section focuses on the modeling & analytics layer and proposes a predictive analytics concept for oil and gas processing facilities. In doing so, the digital twin architecture developed by LeBlanc (2020) and illustrated in Figure 7 is used as a governing framework to ensure a holistic, interface-compatible, and value-oriented design. From Figure 7, it can be observed that the “Modeling and Analytics” is central to the operation of a Digital Twin; therefore, the following sections propose an architectural design suitable for the requirements of the big data generated at oil and gas processing facilities. The carrying out of this task starts with a high-level operational view of the entire data system that allows further development of its elements of form and its associated functions. The elements of form and function are necessary for developing the predictive analytics system architecture.

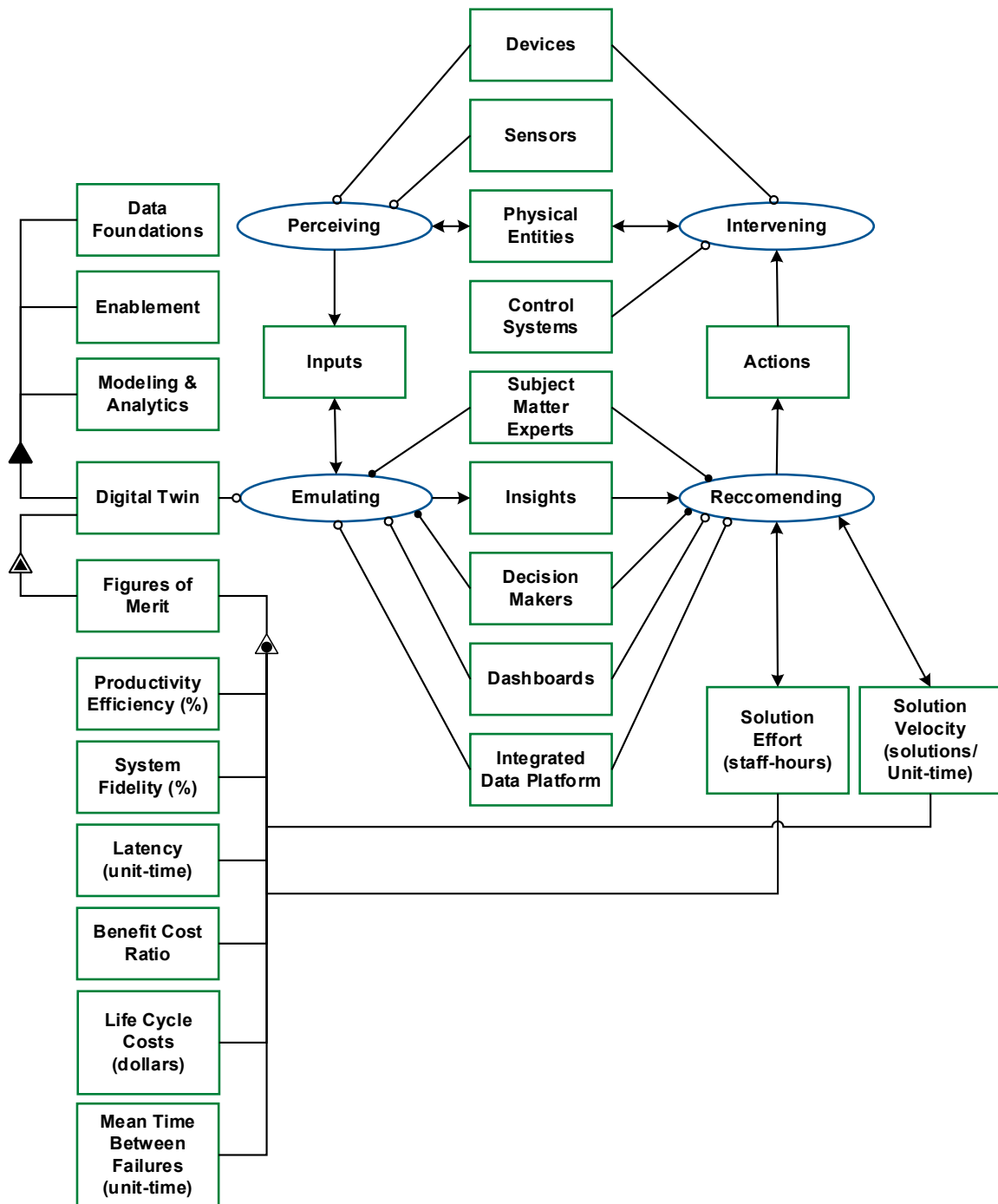


Figure 7. Oil and gas digital twin Object-Process Diagram (OPD). The figure illustrates the object-process diagram (OPD) of a digital twin system functional architecture. Adapted from LeBlanc (2020).

3.5. Developing the modeling and analytics enablement layer

Typically, field instrumentation is installed on process equipment to sense, monitor, and control process variables at processing facilities. Process variables are a representation of process states generated by thermodynamic interactions in each unit operation. Field instrumentation senses such process states through different mechanisms, such as pneumatic, ultrasonic, and electric devices. Such measurements are communicated to control systems through different communication protocols and transmission devices. Commonly found control systems in the oil and gas facilities are Supervisory Control and Data Acquisition (SCADA) or Distributed Control System (DCS).

Control systems convey information to human operators through specialized human-machine interfaces (HMIs). Such interfaces allow the console operators to remotely supervise the process facilities and intervene whenever automatic controllers cannot correctly address process upsets. Console operators typically monitor and control the process from remote control rooms located in safe areas of the process plant. Control systems also communicate with process historians that collect, register, and store process data. Process historians are essential for process engineers to monitor and optimize process plant operations. They allow engineers to analyze process variables trends and run some basic-level statistics to detect, study and prevent potential process disruptions that could lead to downtime. However, process data exhibit all the traits of big data: volume, velocity, and variety. An engineer can quickly become overwhelmed with the massive amounts of information if “artisanal” data manipulation procedures are employed. Even spreadsheet-based monitoring tools (such as Excel) can quickly become sluggish in performance as more data from the process historian is incorporated. Furthermore, holistic, data-driven insights may never be obtained by only looking at a subset of process variables for a limited operational time window.

Advances in statistical or machine learning approaches provide means to shift from a reactive to a proactive strategy by harnessing the power of historical data to predict future system states. Additionally, advances in computational power, cloud computing, storage capacity, and network speed and connectivity push the boundaries of data processing

capabilities, thereby reducing the time to insight. Predictive analytics comes in as a solution that integrates data and models while considering the entire value path of data from the moment the data are generated in the field until actionable insights are produced. Furthermore, such integration must occur in a harmonious manner that factors in stakeholder needs (typically decision-makers, like engineers and operations and maintenance personnel) in order for a predictive analytics solution to fit into the overall digital twin strategy and components smoothly.

Figure 8 provides a path or signpost to move from a reactive, ad-hoc approach to process data analysis to a more predictive approach. It illustrates the high-level workflow proposed to handle and extract value from process data generated in oil and gas processing facilities employing predictive analysis. However, such a high-level view abstracts the complex relationships between software, hardware, and human actors. Therefore, OPM is used to represent such relationships both in a graphical and language-based manner.

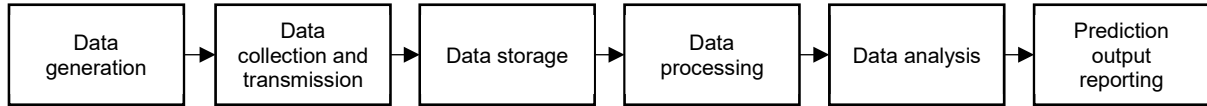


Figure 8. Process data workflow

Figure 9 illustrates the object-process diagram (OPD) for the proposed system architecture of a modeling and analytics enablement layer of the digital twin concept presented in Figure 7. It conveys both elements of form and function and their interactions and is also accompanied by its respective object-process language (OPL).

Figure 9 shows that the inputs to the proposed system are the plant's process variables generated by field instruments or IoT sensors and external data such as weather-related or market-related data. The system's primary externally delivered value-related function is the "reporting predicted process variables;" this is how the system delivers value at its boundary. Its principal internal function (using the OPM construct where functions always end at "ing") are: sensing, transmitting, storing, processing, analyzing, and reporting. Its

external function is generating. Data are the value-related operand in the architecture; as the data move through the different functions, its attributes and states change. In this case, data exhibit the following attributes (or states): internal or external type, storage size, transmission state, processing state, predicted and reported states. Each instrument object is mapped to the corresponding system functions it enables; they are shown on the right-most part of the OPD. There are also agents, in this case, human actors handling some internal functions; they are also shown on the right-most part of the OPD.

Object-Process language of
Figure 9:

Generating requires **Process equipment** and **Laboratory analysis**
Generating yields **Process Variables**
Sensing requires **IoT devices** and **Field instruments**
Sensing yields **Internal data**
Sensing consumes **Process variables**
Transmitting requires **Wired network, Wireless network, RTU and SCADA or DCS**
Transmitting consumes **Internal data** and **External data**
Transmitting yields **Transmitted data**
Storing requires **Storage devices, Relational database (SQL), NoSQL database, Data warehouse and Data lake**
Storing consumes **Transmitted data**
Storing yields **Unprocessed data**
Engineer handles **Processing**
Processing requires **Data policies, Data curation and Queries**
Processing consumes **Unprocessed data**
Processing yields **Processed data**
Engineer handles **Analyzing**
Analyzing requires **Data analytics engine and Statistical learning model**
Analyzing yields **Predictions**
Engineer handles **Reporting**
Reporting requires **Visualization platform, Asset performance tool and SCADA or DCS**
Reporting yields **Process Interventions**

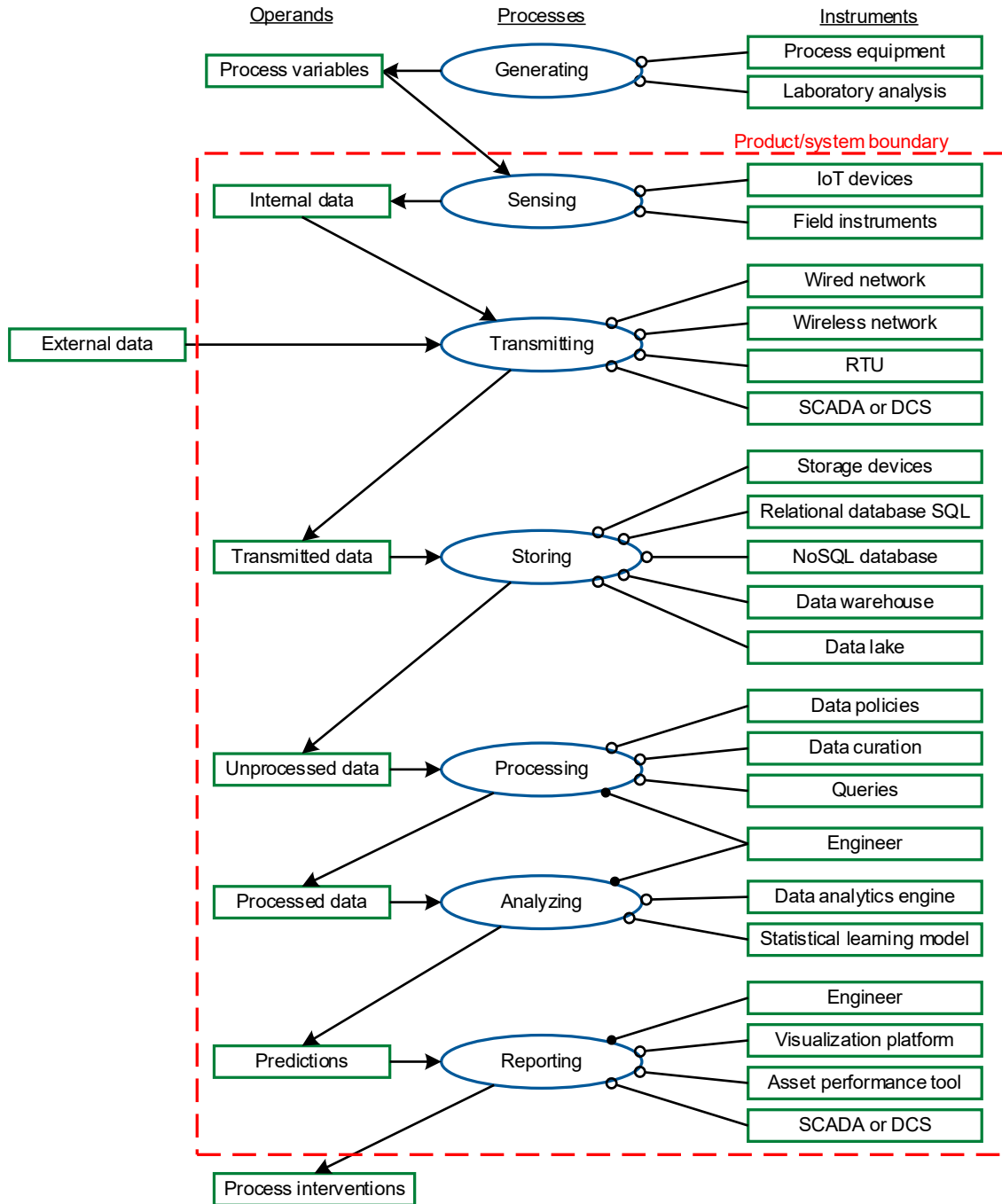


Figure 9. Predictive analytics object-process diagram (OPD)

One of the advantages of using OPM is that it provides a tool- and vendor-agnostic view of the system architecture, allowing the system architects to develop solutions tailored to specific stakeholder needs. In systems thinking, solutions are fixed to user needs instead of users being fixed to tools or off-the-shelf products. OPM, therefore, conveys a solution-

neutral representation of potential system concepts.

Crawley et al. (2015) support the view of “system architecting as a decision-making process” (p. 312). Thus, the process of architectural synthesis implies and should be accompanied by the identification and evaluation of architectural decisions. Not all decisions are architectural; architectural decisions differentiate themselves because they *significantly* impact the system's emergence and performance. Table 6 outlines a list of potential architectural decisions for the development of a predictive analytics system. Table 6 by no means presents a set of architectural decisions that are *mutually exclusively and collectively exhaustive* in the trade space of potential combinations. Big data and analytics are evolving at such an accelerated pace that new methods, theories, and technologies quickly become outdated or form hybrids solutions amongst themselves.

Table 6. Predictive analytics architectural decisions

Architectural Decision	Alternatives
Computing services delivery	<ul style="list-style-type: none"> • On-premises • Private cloud • Public Cloud • Edge computing
Data processing scheme	<ul style="list-style-type: none"> • Stream processing • Batch processing • Hybrid (stream and batch) • Graph processing
Data storage	<ul style="list-style-type: none"> • Distributed file system • Relational database (standard) • NoSQL • NewSQL
Analytics model development	<ul style="list-style-type: none"> • Smart process analytics for predictive modeling • AutoML (automated machine learning) • Manual

Cloud service model

- Infrastructure as a Service (IaaS)
 - Platform as a Service (Paas)
 - Software as a Service (Saas)
-

The analytics model development decision is further studied in the following section as it is considered one of the critical aspects of this research work. The rest of the architectural decisions are equally crucial for a successful analytics solution deployment but thoroughly discussing and developing their theoretical groundworks constitutes an entirely different work on its own.

Chapter 4

A Framework for Predictive Analytics of Process Data

This chapter touches upon diverse supporting platforms typically used for predictive model development. It incorporates the challenges encountered when working with a process data set (aka manufacturing data set) and explores the suitability of popular automated model development schemes. Recognizing that process data have marked differences from other data sets prevalent in the Machine Learning community, a recommended framework customized for process data is reviewed. Lastly, an open-source architecture is proposed for the deployment of analytics at an industrial scale.

4.1. Automated Machine Learning (AutoML) for process data

The predictive model development process is a sophisticated one, requiring specialized skills in mathematics, statistics, and computational science, as a minimum. These skills, however, can only take a data scientist so far in developing a robust, accurate, and interpretable model. One key ingredient—often overlooked—is technical domain knowledge since it bridges the purely data-driven with the specific challenges of each data set. Therefore, it is no surprise that technical professionals with data analytics skills have become a precious asset in the industrial ecosystem. At the same time, the lack of such professionals has also become an obstacle in infusing data analytics into legacy productive processes, such as oil and gas. Additionally, depending on the characteristics of the data set, the model development process may take a considerable amount of time and computational power, even for experienced data scientists.

In response to the challenges above, automated model development processes, known as AutoML, have been developed to provide low to no-code solutions for less experienced

professionals and increase the efficiency of the more experienced ones. AutoML processes are still expected to handle complex data set characteristics (i.e., nonlinear relationships, multicollinearity among predictor variables, among others) in a proper manner to yield acceptable models. AutoML tools have been around for some years now; Figure 10 illustrates recent historical developments for some open-source and commercial (aka AutoML as a Service—AutoMLaaS) tools.

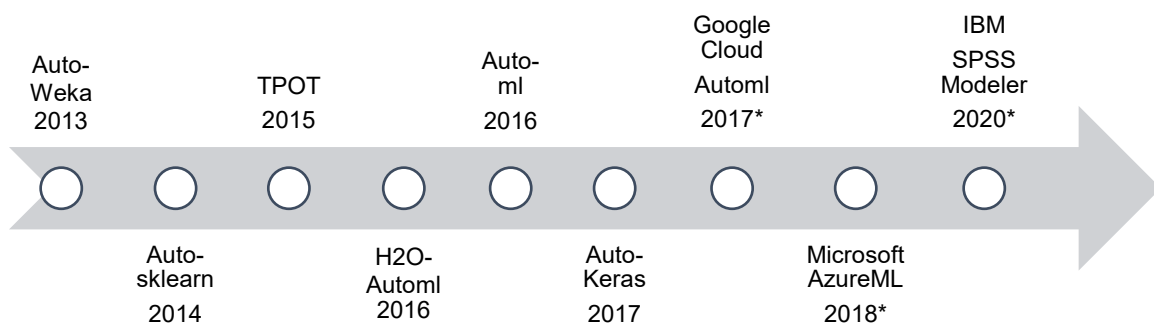


Figure 10. Popular open-source and commercial (marked with an *) AutoML tools.

The majority of AutoML tools follow the process (aka pipeline in the Machine Learning community) shown in Figure 11.

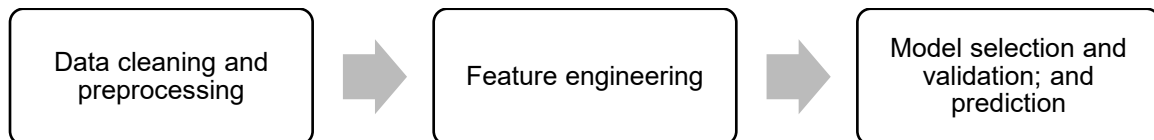


Figure 11. A high-level view of the AutoML process. The middle box typically includes both feature engineering and nonlinear regression.

Each tool shown in Figure 10 has unique attributes when it comes to:

- Supported data sources: structured (i.e., tabular data) or unstructured data (i.e., videos, images, and music)
- Supported data types: numerical, time series, categorical, other
- Type of ML techniques: supervised, unsupervised or both
- Feature engineering: handling of missing data, feature selection, and reduction
- Hyperparameter optimization: grid search, Bayesian search, random search, among others.

AutoML tools rely on searching the space of potential candidate models by considering many of them through different methods (i.e., OLS and RF) and then using a scheme as cross-validation to compare the different models under the exact figure of merit and selecting the optimal one. Furthermore, such tools were not created thinking of process data. Instead, a more general approach was taken to cater to a wide range of use cases that can benefit from machine learning (beyond regression-related problems), such as machine vision and natural language processing (NLP).

Furthermore, although in theory, the name AutoML itself emphasizes the word *automation* as a value-adding differentiator, in reality, that is not entirely accurate. Experts in the field of machine learning have pointed out that there are steps—such as data cleaning and preprocessing—that are still highly dependable on human input for producing an acceptable model (Truong et al., 2019). Other authors have acknowledged that understanding the attribute of domain-specific datasets is still a lagging factor for data scientists and domain experts, and only when this process has properly been completed is when AutoML step may take place (Santu et al., 2021).

Because of the limitations mentioned above of AutoML, analytics practitioners who work with process data should be wary of implementing automatic computational routines for predictive model development for two reasons. First, research studies have demonstrated that the higher and more complex the candidate models, the higher the degrees of freedom, which also increases the likelihood of overfitting (in this situation, the model does not generalize well). This observation motivates the application of a data interrogation step to process data to select the type of model (Sun & Braatz, 2021). Second, given the

characteristic of process data (aka manufacturing data), the conventional or default cross-validation procedures may not be appropriate to select the best model (Sun & Braatz, 2021).

Also, process data analytics typically requires both data-science-related skills and domain knowledge due to the challenges of missing values, sensor measurement error, measurement delays, varying sampling rates, multicollinearity, outliers, nonlinear relationships, and drifted data (Kadlec et al., 2009).

In closing, the development of a suboptimal model compromises the value realization and the reliability derived from data-driven techniques in the eyes of end-users, such as process engineers and plant operators. Also, the use of poorly performing models can lead to safety and economic consequences should a human actor or a computer make decisions based on erroneous predictions. Therefore, each predictive model pipeline in the case of process data calls for frameworks that customize the analytics and machine learning toolkit to the problem at hand—a systematic approach.

4.2. Tailoring the analytics model development to process data.

When it comes to time-series process data, conventional or standard analytics methods must be carefully assessed before implementation. Mainstream or popular methods such as neural networks and conventional k-fold cross-validation procedures (just to name some examples) have provided superior performance for specific use cases in the Machine Learning community. However, it does not necessarily mean they will yield the same performance when used on process data (aka manufacturing data).

Process data pose unique challenges such as high diversity in data quality and quantity. Therefore, the model developer must carefully balance its familiarity and knowledge of analytics methods with the goals of the intended model, which implies the need for technical domain expertise besides a pure analytics-based approach. Sun and Braatz (2021) have developed a framework called Smart Process Analytics (SPA) tailored to manufacturing data that enables the user to focus on goals rather than methods. SPA is a

tool- and vendor-agnostic framework that can be deployed in open-source programming languages such as Python and R. SPA can also be embedded into any existing big data framework to automate the predictive model development process.

The SPA framework considers the specific data characteristics and domain knowledge while employing specialized cross-validation schemes. The authors argue that two sequential steps must be followed for successful predictive modeling:

1. Data interrogation for investigating data characteristics
2. Method selection considering data characteristics and domain knowledge

The main driver for data interrogation is the need to select a model suited for training data characteristics. As a starting point and baseline, Sun and Braatz (2021) highlight that the Gauss-Markov theorem states that an Ordinary Least Squares (OLS) estimator is suitable for model development only if all these conditions hold:

1. Residuals are uncorrelated, have equal variances (homoscedastic), and have zero mean.
2. The relationship between the input variables and the output variable should be linear.
3. The training predictor matrix $\mathbf{X} \in \mathbb{R}^{N \times m_x}$ in the OLS Equation

$$\hat{\mathbf{w}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (1)$$

has full rank.

However, it is widely known that time-series process data very rarely exhibit these ideal conditions. The SPA's authors argue that other alternatives should be considered when process data exhibits any of the following conditions:

1. Nonlinear relationships between predictors and output variable
2. Multicollinearity among predictors
3. Serially correlated model residuals or errors
4. Heteroscedasticity of model residuals

The objective of SPA is to provide a robust and automated predictive model systematically. For doing so, SPA considers the four potential data characteristics mentioned above during the data interrogation step before selecting a predictive method or algorithm.

Figure 12 illustrates the model development process, highlighting how the above challenges are addressed throughout the process.

Firstly, nonlinear relationships refer to any other type of mathematical relationship that deviates from a straight line between two variables of importance (aka a bivariate).

Secondly, multicollinearity is the presence of high absolute values of correlations between two or more predictor variables, denoting a strong linear relationship among them. Under this situation, multiple linear regression may yield erroneous coefficient estimates, which can vary drastically with small changes in the training data, therefore leading to a poorly constructed model.

Thirdly, serial correlations among model residuals signal time-varying relationships among the output and input variables, yielding misspecified models if such relationships are not adequately considered.

Lastly, heteroscedasticity refers to the volatility of the errors in a regression model. A heteroscedastic model is when error terms have non-constant variance across the different values of either the input or the output variables. Not accounting for heteroscedasticity can also lead to poorly specified models.

4.1.1. Data interrogation step

In SPA, the data interrogation process occurs in the following manner:

- Nonlinearity check
 - a. Pairwise linear correlations are first evaluated using Pearson's correlation coefficient. The Pearson's correlation coefficient varies between -1 and 1 , with a higher absolute value indicating a strong linear relationship
 - b. Two linearity tests are also conducted as a second step: the quadratic test and

maximal correlation test.

- c. A bilinear test can also be conducted to investigate a strong correlation between the output variable (aka response variable) and interactions of predictor variables.
- d. Decision rule: “If both the linear correlation and the maximal correlation are close to zero or one, it means that the variables are uncorrelated or linearly correlated, respectively. If the linear correlation is close to zero, but the maximal correlation is close to 1, it indicates the variables are nonlinearly correlated” (Sun & Braatz, 2021, p. 5).

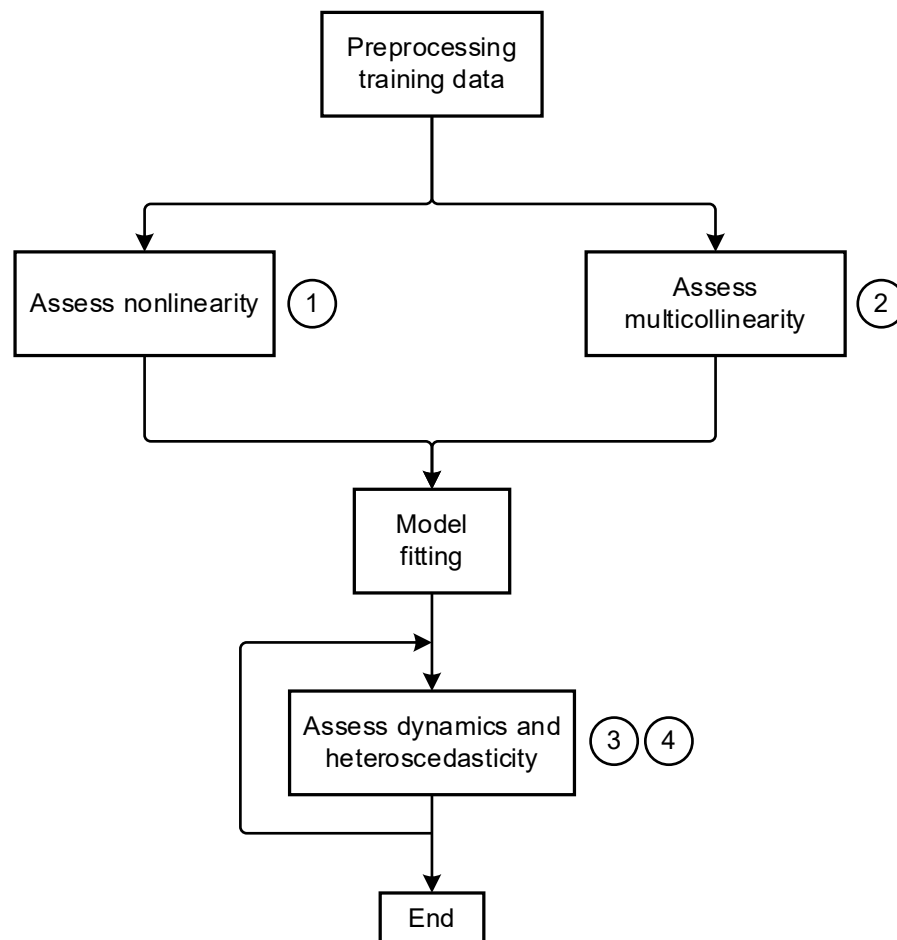


Figure 12. SPA predictive modeling workflow. Adapted from Sun & Braatz (2021)

- Multicollinearity check
 - a. The variable inflation factor (VIF) is employed to check for the presence of multicollinearity among input (aka predictor) variables. The VIF quantifies how much the variance of a regression coefficient is impacted due to multicollinearity. Each predictor is regressed against the rest of the predictors, and the resulting R-squared value is used to compute the VIF:

$$VIF_k = \frac{1}{1 - R_k^2} \quad (2)$$

- b. Decision rule: “VIF greater than 10 is a sign of significant multicollinearity” (Sun & Braatz, 2021, p. 5).

- Dynamics check
 - a. Dynamics are assessed by first inspecting the static model residuals, that is the regression residuals (ε) when time is not factored into the model. The reason for starting with a static model first is that not all time-series regressions strictly require the use of a dynamic model. If the model residuals reveal a dynamic behavior that the static model does not capture, then a dynamic model (i.e., DALVEN) is needed. The autocorrelation function (Kočenda & Černý, 2015),

$$\text{corr}(\varepsilon_t, \varepsilon_{t-1}) \quad (3)$$

is employed in SPA. The autocorrelation function yields the level of correlation between a time-series variable and its lagged values.

- b. Decision rule: “A dynamic model is needed when the autocorrelation in the model residual is significant” (Sun & Braatz, 2021, p. 5).

- Heteroscedasticity check
 - a. There is no single recipe for evaluating heteroscedasticity within the SPA framework; instead, several methods are considered, such as the Goldfeld-Quandt test, the Breusch-Pagan test, and the White test. More information about these methods can be found in the works of Breusch and Pagan (1979) and Goldfeld and

Quandt (2021). The SPA authors recommend using such a test accompanied by a visual inspection of model errors to arrive at plausibly valid conclusions about errors heteroscedasticity.

4.1.2. Model fitting step

Regarding the modeling techniques, SPA offers a diverse selection of methods suited for specific data characteristics, which are illustrated in

Table 7. Dense and sparse methods are considered.

Table 7. Methods used in SPA

Dataset characteristic	Suitable methods
Nonlinearity only	ALVEN, SVR, RF
Dynamics only	CVA, MOESP, SSARX
Multicollinearity only	RR, Elastic net PLS, Sparse PLS
Nonlinearity and dynamics	RNN, DALVEN
Nonlinearity and multicollinearity	ALVEN, SVR, RF
Multicollinearity and dynamics	CVA, MOESP, SSARX
Nonlinearity, multicollinearity, and dynamics	RNN, DALVEN

Note. ALVEN, algebraic learning via elastic net; DALVEN, dynamic ALVEN; SVR, support vector regression; RF, random forest; CVA, canonical variate analysis; MOESP, multivariable output error state space; RR, Ridge regression; PLS, partial least squares; RNN, recurrent neural network. For more information about each method, see Sun (2020).

The methods presented in Table 7 have hyperparameters that must be properly tuned to

procure acceptable model generalization. Cross-validation (CV) is the most popular method use for hyperparameter tuning and selection. Cross-validation offers a wide variety of sub-methods (aka schemes), each tailored for specific characteristics of the dataset. SPA employs such sub-methods in a fit-for-purpose manner. For picking the suitable CV scheme, SPA considers:

1. The total number of samples (aka observations) and predictors
2. The presence of replicate samples and predictors
3. The specific objectives of the analysis
4. The time available to perform CV
5. The ordering of the samples in dataset
6. Consequences of flawed results, either overly optimistic or pessimistic.

The CV schemes considered are:

1. Simple held-out validation set, which is the simplest form of CV. The dataset is split into a single training and validation set. The model is trained on the training set, and the predictive power is assessed on the validation set. The model that yields the best performance on the validation set is selected and trained on the entire dataset.
2. k -fold cross-validation, where the dataset is split into k groups, $k - 1$ folds are used for model training, and 1 group is held out for model performance evaluation. The procedure is repeated k times where each unique fold must be held out as a validation set. A final score is calculated by averaging the model's performance across all the k validation sets.
3. Repeated k -fold: this procedure is very similar to standard k -fold CV. Under this scheme, the dataset is split into k subsets of validation and training sets, with the difference that the procedure is repeated N times, and model performance is averaged across all validation sets.
4. Monte Carlo-based cross-validation, which is carried out in a similar fashion like

repeated k -fold CV. The difference is that the dataset set is randomly shuffled and split into a training and validation set for each k fold. Model performance is also averaged across all validation sets.

5. Grouped cross-validation, which comes in handy when there is some sort of group structure in the dataset. In manufacturing data, group structure could be due to measurements being carried out by different sensors over time. For further details about this scheme, refer to Sun (2020).
6. Nested cross-validation, which is a super scheme that works as an outer iteration layer for any of the CV schemes mentioned above. On a series of repetitions, the original dataset is split into a training and a test set. The training set is then split into training and validation set for performing any of the CV schemes. The test set of the outer loop will only be used for evaluating the model performance and not for hyperparameter tuning. Holding a separate test set just for performance evaluation provides a more realistic generalization error than using the validation set to both hyperparameter tuning and model selection. This method can also yield a more stable and robust model. However, the SPA authors warn about the computational power and thus the time required to perform this procedure. Figure 13 illustrates an example of this scheme.
7. Cross-validation for dynamic models, which applies to time-series data with temporal dependencies among samples, in which conventional CV procedures are not adequate. Splitting the data arbitrarily or randomly when there are temporal relationships causes data leakage because the temporal relationships are lost. Therefore, the splitting of the dataset into training and validation sets must occur in sequential chronological order. SPA refers to this scheme as time-series CV, although it is also known as forward-chaining or rolling basis CV in other fields, such as econometrics, where time-series financial data is used to generate forecasts. Figure 14 illustrates an example of this procedure.

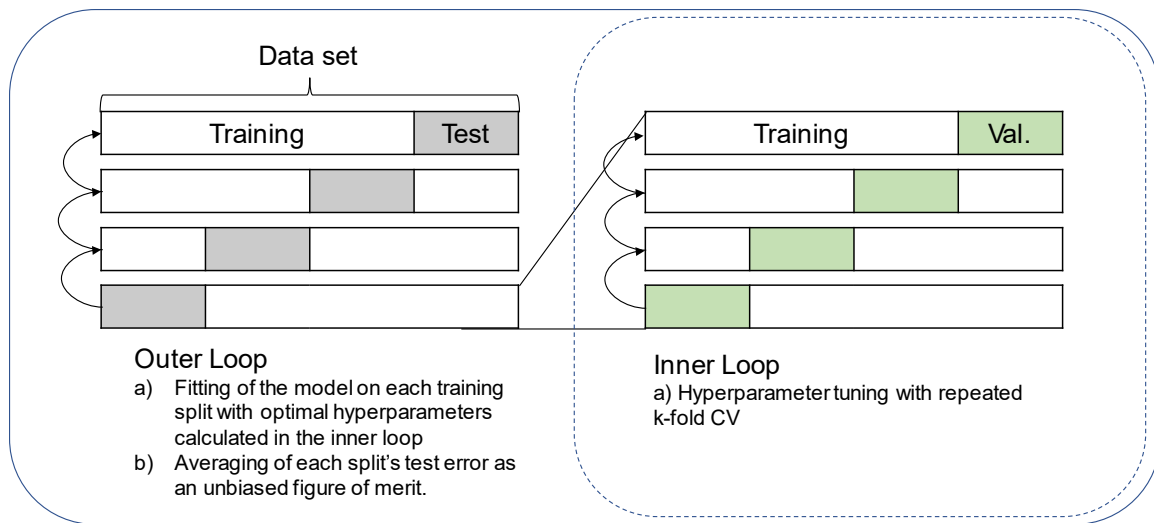


Figure 13. 4-fold nested CV.

For cases where a dynamic model is needed, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are also part of the SPA model evaluation workflow (Akaike, 1974; Schwarz, 1978).

In order to balance the model's complexity, interpretability and prediction power, SPA uses the one-standard-error rule to pick the final model. The one-standard-error rule “selects the most parsimonious model whose error is smaller than one standard deviation above the error of the best model... The parsimonious model not only has better interpretability but is also less prone to overfitting” (Sun & Braatz, 2021, p. 12).

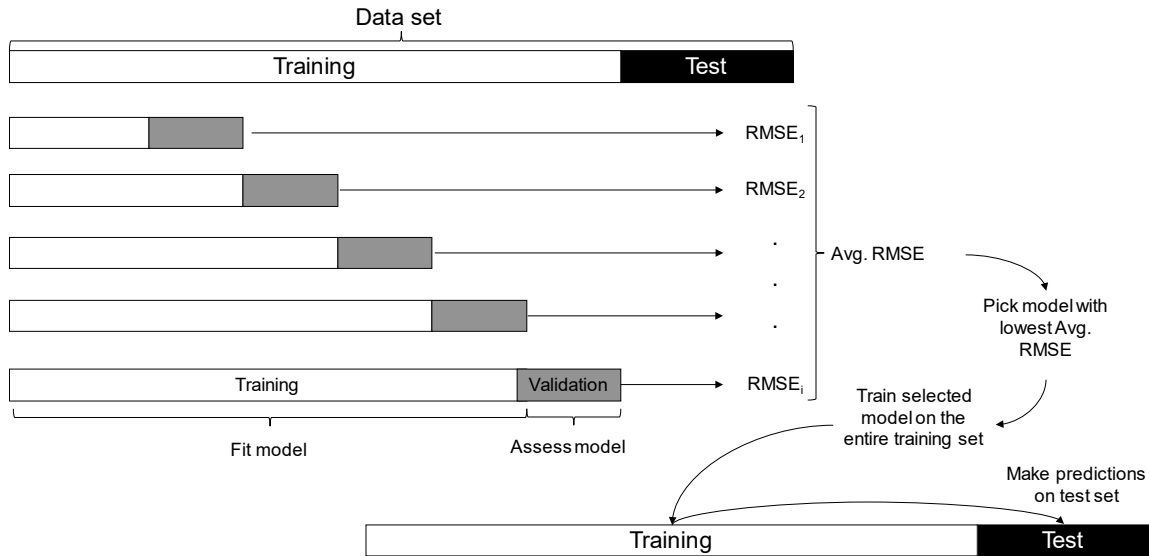


Figure 14. 5-fold time-series CV (aka forward-chaining CV). Note. RMSE is only shown as an example of a figure of merit, others can be used (i.e. MAE, MSE).

4.3. Open-source architectures for implementing SPA for oil and gas processing facilities

Regarding big data analytics architectures, there are many combinations possible given the prolific R&D ecosystem both in the academic and private sectors. This section aims to discuss the most popular concepts when it comes to developing big data solutions for oil and gas processing facilities, given recent advances in the areas of cloud computing, IIoT, and open-source resources.

Figure 9 outlines the steps along the data analytics value path: data generation, collection, storage, processing, analysis, visualization, and reporting. Essentially all data big data architectures facilitate these steps. However, there are differences in how they do it. As a result, diverse frameworks have been developed to organize and classify analytics-related architectural alternatives based on: the underlying data management platforms, the network layer where computations are carried out, and the type of data processing (aka data analytics system) to be performed (Marhani & Nasaruddin, 2017; Nguyen, Gosine, &

Warrian, 2020).

About the data management platform, two broad categories are found:

- Big data on a single server (aka vertically-scalable platform), which includes carrying computations on
 - High-performance computing clusters (HPC)
 - Multicore CPU and graphics processing units (GPU)
 - Specialized hardware units such as field-programmable gate arrays (FPGA) and application-specific integrated circuits (ASIC).
- Big data on multiple servers (aka horizontally-scalable platform), where computations are carried out in
 - Peer-to-peer networks (TCP/IP)
 - Apache Hadoop
 - Apache Spark
 - Other platforms. Only some of the most popular are mentioned.

Regarding the cloud network layers at which computations occur, three types have become increasingly popular:

- Cloud computing, which is the highest level layer, where computations (i.e., data processing) occur at the datacenter level (Nguyen et al., 2020).
- Fog computing, which is a middle where computations are carried out in distributed nodes located between the cloud and the edge of the network.
- Edge computing, which is the level closest to the sources of data generation; the idea behind edge computing is to reduce response time by bringing computations as close as possible to data generation devices (e.g., smartphones, sensors, wearable devices).

Concerning the type of data processing, three main categories exist:

- Batch analytics: this processing model is also known as off-line analytics, and is

employed when low response times are not required (Marhani & Nasaruddin, 2017). It can be deployed executed on distributed clusters to exploit the power of parallel processing. The MapReduce algorithm developed by Google is typically used under this approach to carry out complex parallel computations on massive data sets (Wang, Yang, Wang, Sherratt, & Zhang, 2020).

- Streaming (aka real-time) analytics: this type of analytics is performed on constantly changing and or flowing data and when low response times are needed. Some examples include processing data from field sensors or instruments; it can be deployed through distributed clusters or memory-based computing platforms (Marhani & Nasaruddin, 2017).
- Hybrid analytics: is a combination of the previous two options; there are situations where both batch and stream analytics are employed. A prominent example is the Lambda architecture to be explored later in this section.

Many combinations can be performed among the architectural alternatives mentioned above, depending on the use case requirements at hand. However, for predicting process variables in oil and gas processing facilities, the following requirements must be considered when architecting a solution:

- Low latency: Typically, predictive models will be embedded within process control schemes, requiring a quick response time to keep control loop dead time as low as possible.
- Fault-tolerant systems: architectures that operate in oil and gas settings must provide enough redundancy to cope with unplanned failures (Nguyen et al., 2020).
- Highly scalable: the platform must be able to scale as more data is generated, and therefore, more resources are demanded. It is estimated that a large refinery can generate 1 TB of data per day (Bechtold, 2018).

Big data platforms are architecture using a layered approach illustrated in Figure 15, which has become the go-to reference for many analytics-related companies. It was formally adopted in 2020 as the reference architecture by the standard ISO/IEC 20547-3:2020

Information technology — Big data reference architecture. Other authors also refer to the platform layer as the collecting and storage layer (Wang et al., 2020), which is the layer where databases for structured or unstructured data reside; this layer is also closest to the field sensors and instruments (aka infrastructure layer). The processing layer is where analytics-related routines are executed, whether for batch, stream, or hybrid analytics. The applications layer serves as an interface between the big data platforms and their end-users; this is the layer where application programming interfaces (API) are utilized to execute functions that allow the user to interact with the operating environment.

Most of the architectural concepts used for big data are developed following the platform approach laid out in Figure 15; the differences arise in the arrangement of functions within each layer and how they interact and connect with the rest of the layers. Since the oil and gas industry possesses use cases that demand real-time analytics, two of the most popular architectures employed in the processing layer will be discussed: Lambda and Kappa.

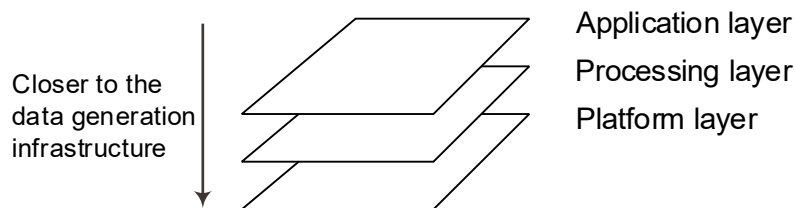


Figure 15. Big data functional layers

4.1.1. Lambda architecture

Nathan Marz introduced the Lambda architecture concept in 2011 in a blog post titled “How to beat the CAP theorem.” Since then, it has become one of the most popular architectures to facilitate real-time big data analytics. Major companies like Microsoft and Amazon offer it in their platforms Azure and AWS, respectively. The driver behind the Lambda architecture is that for massive data sets, it becomes increasingly challenging to run complex queries at or near real-time because of the introduction of latency, even when

algorithms such as MapReduce are employed. Typically MapReduce would split the data to conduct parallel processing across the splits and then combine the results for the user to execute the required queries; however, this process may take several hours, yielding late results for the decision-making process for big data. The Lambda architecture overcomes the introduction of latency through the use of a batch layer and speed layer that work in parallel. The speed layer is subject to low-latency requirements and therefore yields faster results and real-time views but at the expense of prediction accuracy.

On the other hand, the batch layer is not subject to latency requirements and can yield more accurate results at the expense of higher response time. The speed and batch layer results converge in the service layer (aka merging layer), which enables the querying process for such results (Sanla & Numnonda, 2019). Figure 16 illustrates the Lambda architecture. One drawback of the Lambda architecture is its complexity; having two different layers to carry out computation increases the coding overhead and overall maintenance (Forgeat, 2015). Diverse open-source technologies are available to implement the Lambda architecture, such as Apache Storm, Apache Spark, Apache Kafka, and Apache Streaming. Most of these technologies support multiple programming languages such as Python, R, Java, and Scala.

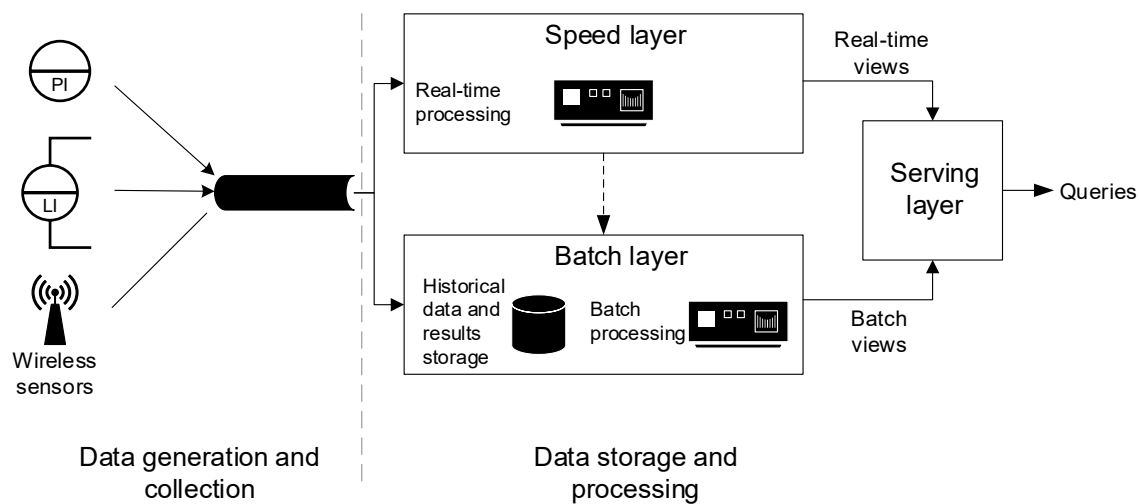


Figure 16. Lambda architecture for data processing

4.1.2. Kappa architecture

Jay Kreps introduced the Kappa architecture in 2014 (Sanla & Numnonda, 2019) to overcome the complexity of the Lambda architecture. The main difference is that there is no longer a batch layer, only a speed, and a serving layer. In other words, all the data flows along a single path, the real-time path. Under this scheme, typically, a single stream of data is processed. The reprocessing of data is only performed when part of the stream processing code needs to be modified or updated; when this happens, another processing job is run (Forgeat, 2015). Under this architecture, data are constantly replaced by incoming new data. Similar to the Lambda architecture, the serving layer enables the querying process. Figure 17 illustrates the Kappa architecture. Like the Lambda case, many open-source technologies are available to implement this architecture, especially from the Apache software foundation. Some of the most popular are Apache Spark, Drill, Streaming, Cassandra, among many others. Some studies address which tools are suitable for each use case and step of the data management process; for more information, see the works of Sanla and Numnonda (2019), Marhani and Nasaruddin (2017), and Wang et al. (2020).

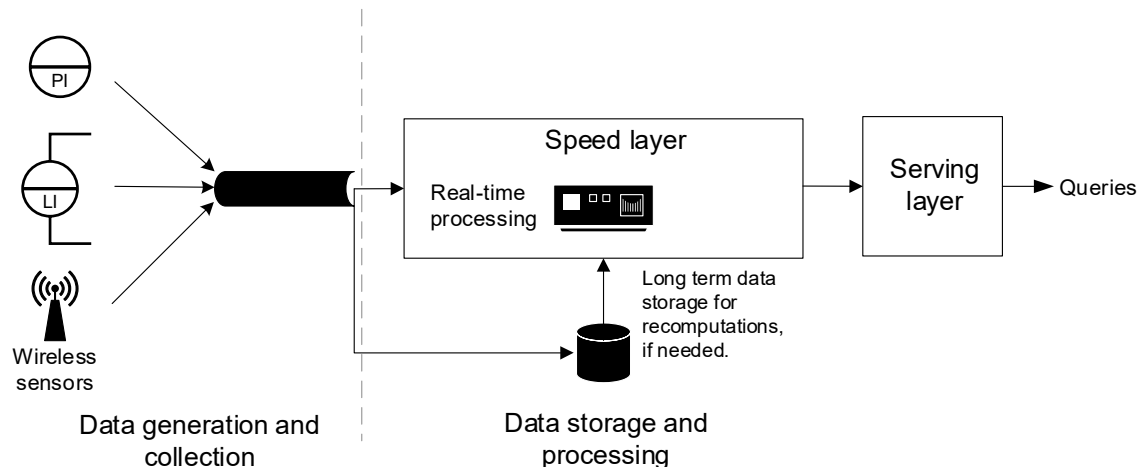


Figure 17. Kappa architecture for data processing

Chapter 5

Conclusion

This section presents key findings from each chapter, along with recommendations and future work.

5.1. Summary

This thesis presented a review and proposal of potential predictive analytics applications to oil and gas processing facilities in a systems-oriented approach.

Chapter 1 outlined the value case and reasons why oil and gas companies should implement predictive analytics into their existing workflows to reduce the time-to-insight in their decision-making processes. Predictive analytics is a value enabler within the Big Data pillar in the overall Industry 4.0 paradigm. A literature review presents data analytics methods and system representation methodologies as a theoretical foundation to develop the rest of the chapters.

Chapter 2 presented use cases for oil and processing facilities covering the value chain from the upstream to the downstream sector. A review of diverse analytics methods applied for such use cases was presented. PCR and PLS are the most prevalent methods for developing predictive models, such as the soft sensors commonly found in inferential control strategies. Use cases in which neural networks were employed are also discussed; however, these methods do not always generalize well depending on the specific type of neural network.

The potential value from the use of predictive analytics was investigated in Chapter 3 through the review of secondary sources. The big data sector in oil and gas, in predictive analytics, falls, is expected to grow at a sustained pace over the time horizon between 2021

and 2024. The role of analytics as a system of systems within the overall digital twin framework was also evaluated. Analytics is a value enabler in the digital twin architecture, as it allows a data-driven representation of system states, and it is a foundational layer for the data management process.

Lastly, Chapter 4 introduces Smart Process Analytics (SPA) as a data analytics framework tailored for process data (aka manufacturing data), such as that generated at oil and gas facilities. In contrast to automated model development processes (aka AutoML), SPA follows rigorous data interrogation and model selection methods. As a result, SPA intends to build a predictive model that balances robustness and complexity. Potential architectures to implement a data analytics system capable of handling big data are also presented. Such architectures are presented in a vendor-agnostic manner leveraging open-source frameworks and tools. The Lambda and Kappa architectures present themselves as tangible options to implement data analytics for massive data sets generated rapidly, such as the case of process data in oil and gas facilities.

5.2. Future work

Although the SPA workflow has been tested on the Tennessee Eastman simulation dataset, it can also be employed on another dataset with other unit operations representative of other sectors of the oil and gas value chain. For example, on datasets from upstream facilities such as crude oil and gas dehydration plants.

This thesis considered the predictive modeling component of SPA; there is also a process monitoring component under development for the early detection of process upsets that may lead to process downtime so that promptly corrections can be performed. Future work can be directed at exploring the use cases for process monitoring and embedding its tools and methods within the overall analytics architecture.

Although Chapter 1 discusses some aspects of cybersecurity associated to cloud computing, it is an extensive area identified as an obstacle in the digitalization of the oil and gas industry. As a result, future works can be focused on investigating and proposing

safeguards that prevent or mitigate the risk from cyberattacks.

Lastly, it would be beneficial to frame the deployment of the data analytics system concept proposed under one of the project workflows commonly adopted in several industries, such as the Cross Industry Standard Process Data Mining (CRISP-DM) and the Job Task Analysis (JTA) developed by the Institute for Operations Research and the Management Sciences (INFORMS).

[Page intentionally left blank]

Bibliography

- Agarwal, M. (1997). Combining neural and conventional paradigms for modelling, prediction and control. *International Journal of Systems Science*, 28(1), 65–81. <https://doi.org/10.1080/00207729708929364>
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Aker. (2021). DataOps in action: Optimizing oil and gas production with hybrid machine learning. Retrieved June 27, 2021, from <https://www.cognite.com/customers/dataops-oil-gas-hybrid-machine-learning>
- Alpaydin, E. (2014). Reinforcement learning. In *Introduction to Machine Learning* (3rd ed., pp. 517–545). Cambridge, MA: MIT Press.
- Aramco. (2021). Big data, big insights. Retrieved June 27, 2021, from <https://americas.aramco.com/en/magazine/elements/2020/big-data-big-insights>
- Aspentech. (2021). Aspen inferential qualities. Retrieved July 12, 2021, from Reduce off-spec product and maintain tighter quality control website: <https://www.aspentech.com/en/products/msc/aspen-inferential-qualities>
- Aykol, M., Gopal, C. B., Anapolsky, A., Herring, P. K., van Vlijmen, B., Berliner, M. D., ... Storey, B. D. (2021). Perspective—Combining physics and machine learning to predict battery lifetime. *Journal of The Electrochemical Society*, 168(3), 030525. <https://doi.org/10.1149/1945-7111/abec55>
- Bechtold, B. (2018). Beyond the barrel: How data and analytics will become the new currency in Oil and Gas. Retrieved June 17, 2021, from <https://gblogs.cisco.com/ca/2018/06/07/beyond-the-barrel-how-data-and-analytics-will-become-the-new-currency-in-oil-and-gas/>
- Breiman, L. (2001). Random Forest. *Machine Learning*, 45, 5–32. <https://doi-org.libproxy.mit.edu/10.1023/A:1010933404324>
- Bresler, G., & Nagaraj, D. (2021). Sharp representation theorems for ReLU networks with precise dependence on depth. *ArXiv:2006.04048 [Cs, Stat]*. Retrieved from <http://arxiv.org/abs/2006.04048>
- Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287. <https://doi.org/10.2307/1911963>
- Chang, C.-T., Lin, Y.-S., & Georgakis, C. (2002). A simple graphic approach for observer

- decomposition. *Journal of Process Control*, 12(8), 857–873. [https://doi.org/10.1016/S0959-1524\(02\)00020-3](https://doi.org/10.1016/S0959-1524(02)00020-3)
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. San Francisco California USA: ACM. <https://doi.org/10.1145/2939672.2939785>
- Chiang, L. H., & Braatz, R. D. (2020). Process data analytics. *Processes*. Retrieved from https://www.mdpi.com/journal/processes/special_issues/data_analytics
- Chiang, L. H., Russell, E. L., & Braatz, R. D. (2001). *Fault Detection and Diagnosis in Industrial Systems*. London: London. <https://doi.org/10.1007/978-1-4471-0347-9>
- Chien, T. W., Chu, H., Hsu, W. C., Tseng, T. K., Hsu, C. H., & Chen, K. Y. (2003). A feasibility study on the predictive emission monitoring system applied to the Hsinta power plant of Taiwan Power Company. *Journal of the Air & Waste Management Association*, 53(8), 1022–1028. <https://doi.org/10.1080/10473289.2003.10466241>
- Council, J. (2019, November 26). Shell’s companywide AI effort shows early returns. *Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/shells-companywide-ai-effort-shows-early-returns-11574764204>
- Crawley, E., Cameron, B., & Selva, D. (2015). *System Architecture* (1st ed.). Hoboken, NJ: Pearson.
- Dong, D., McAvoy, T. J., & Chang, L. J. (1995). Emission monitoring using multivariate soft sensors. *Proceedings of the American Control Conference*, 5. Seattle, WA.
- Dori, D. (2016). *Model-based Systems Engineering with OPM and SysML*. New York, NY: Springer. <https://doi.org/10.1007/978-1-4939-3295-5>
- Forgeat, J. (2015). Data processing architectures – Lambda and Kappa. Retrieved June 18, 2021, from <https://www.ericsson.com/en/blog/2015/11/data-processing-architectures--lambda-and-kappa>
- Fortuna, L., Graziani, S., Rizzo, A., & Xibilia, M. (2007). *Soft Sensors for Monitoring and Control of Industrial Processes*. London: Springer.
- Ge, Z., Chen, T., & Song, Z. (2011). Quality prediction for polypropylene production process based on CLGPR model. *Control Engineering Practice*, 19(5), 423–432. <https://doi.org/10.1016/j.conengprac.2011.01.002>
- Gharehbaghi, H., & Sadeghi, J. (2016). A novel approach for prediction of industrial catalyst deactivation using soft sensor modeling. *Catalysts*, 6(7), 93. <https://doi.org/10.3390/catal6070093>
- GlobalData PLC. (2021). *AI in oil and gas*. United Kingdom. Retrieved from <https://oilgas->

globaldata-com.libproxy.mit.edu/Analysis/details/thematic-research-ai-in-oil-gas-2021

- Goldfeld, S. M., & Quandt, R. E. (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association*, 60(310), 10.
- Han, C., & Lee, Y.-H. (2002). Intelligent integrated plant operation system for Six Sigma. *Annual Reviews in Control*, 26(1), 27–43. [https://doi.org/10.1016/S1367-5788\(02\)80008-6](https://doi.org/10.1016/S1367-5788(02)80008-6)
- International Organization for Standardization. (2016). *ISO 14224:2016 Petroleum, Petrochemical and Natural Gas Industries—Collection and Exchange of Reliability and Maintenance Data for Equipment*. International Organization for Standardization. Retrieved from <https://www.iso.org/standard/64076.html>
- Jordaan, E., Kordon, A., Chiang, L. H., & Smits, G. (2004). Robust inferential sensors based on ensemble of predictors generated by genetic programming. *Parallel Problem Solving from Nature - PPSN VIII*, 3242, 522–531. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-30217-9_53
- Kadlec, P., & Gabrys, B. (2011). Local learning-based adaptive soft sensor for catalyst activation prediction. *AIChE Journal*, 57(5), 1288–1301. <https://doi.org/10.1002/aic.12346>
- Kadlec, P., Gabrys, B., & Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering*, 33(4), 795–814. <https://doi.org/10.1016/j.compchemeng.2008.12.012>
- Kaneko, H., & Funatsu, K. (2014). Application of online support vector regression for soft sensors. *AIChE Journal*, 60(2), 600–612. <https://doi.org/10.1002/aic.14299>
- Kano, M., Miyazaki, K., Hasebe, S., & Hashimoto, I. (2000). Inferential control system of distillation compositions using dynamic partial least squares regression. *Journal of Process Control*, 10(3), 157–166. [https://doi.org/10.1016/S0959-1524\(99\)00027-X](https://doi.org/10.1016/S0959-1524(99)00027-X)
- Kim, S., Kano, M., Hasebe, S., Takinami, A., & Seki, T. (2013). Long-term industrial applications of inferential control based on just-in-time soft-sensors: Economical impact and challenges. *Industrial & Engineering Chemistry Research*, 52(35), 12346–12356. <https://doi.org/10.1021/ie303488m>
- Kočenda, E., & Černý, A. (2015). *Elements of Time Series Econometrics: An Applied Approach*. Prague: Karolinum Press.
- Kordon, A. K. (2020). *Applying Data Science: How to Create Value with Artificial Intelligence*. Cham, Switzerland: Springer International Publishing. <https://doi.org/10.1007/978-3-030-36375-8>

- Kresta, J. V., Marlin, T. E., & MacGregor, J. F. (1994). Development of inferential process models using PLS. *Computers & Chemical Engineering*, 18(7), 597–611. [https://doi.org/10.1016/0098-1354\(93\)E0006-U](https://doi.org/10.1016/0098-1354(93)E0006-U)
- Launchbury, J. (2020). A DARPA perspective on artificial intelligence. Retrieved March 19, 2021, from <https://www.darpa.mil/about-us/darpa-perspective-on-ai>
- LeBlanc, M. (2020). *Digital Twin Technology for Enhanced Upstream Capability in Oil and Gas* (Master's thesis). Massachusetts Institute of Technology, Cambridge, MA.
- Lee, J., Son, Y., Lee, K., & Won, W. (2019). Economic analysis and environmental impact assessment of heat pump-assisted distillation in a gas fractionation unit. *Energies*, 12(5), 852. <https://doi.org/10.3390/en12050852>
- Marhani, M., & Nasaruddin, F. (2017). Big IoT data analytics: Architecture, opportunities, and open research challenges. *IEEE Access*, 5, 5247–5261. <https://doi.org/10.1109/ACCESS.2017.2689040>
- Marsh. (2016). *The 100 largest losses 1974-2015*. Marsh & McLennan Companies. Retrieved from Marsh & McLennan Companies website: <https://www.marsh.com/us/insights/research/the-100-largest-losses-in-the-hydrocarbon-industry-1974-2015.html>
- Marz, N. (2011). How to beat the CAP theorem. Retrieved June 18, 2021, from Thoughts from the red planet website: <http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html>
- Matthews, C. M. (2018, July 24). Silicon Valley to big oil: We can manage your data better than you. *Wall Street Journal*. Retrieved from <https://www.wsj.com/articles/silicon-valley-courts-a-wary-oil-patch-1532424600>
- McKinsey & Company. (2020). Digital & advanced analytics. Retrieved February 15, 2021, from <https://www.mckinsey.com/industries/oil-and-gas/how-we-help-clients/digital-and-advanced-analytics>
- Mejdell, T., & Skogestad, S. (1991). Estimation of distillation compositions from multiple temperature measurements using partial-least-squares regression. *Industrial & Engineering Chemistry Research*, 30(12), 2543–2555. <https://doi.org/10.1021/ie00060a007>
- Microsoft. (2016). *Empowering the oil & gas and mining industry*. Retrieved from <https://info.microsoft.com/rs/157-GQE-382/images/Whitepaper%20-%20Oil%20Gas%20Mining%20Strategy%20v2.pdf?aliId=300490449>
- Microsoft. (2021). What is cloud computing? A beginner's guide. Retrieved June 14, 2021, from <https://azure.microsoft.com/en-us/overview/what-is-cloud-computing/>

- Milenkovic, M. (2020). *Internet of Things: Concepts and System Design*. Cham, Switzerland: Springer International Publishing. <https://doi.org/10.1007/978-3-030-41346-0>
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- Misfer, A. S., Vedula, S., Hauer, R., & Juddy, Z. (2009). *Process analyzer best practices for sulfur recovery, enhanced Claus and tail gas treating applications*. Presented at the Sour Oil & Gas Advanced Technology (SOGAT), Abu Dhabi, UAE.
- Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics*, 38(8), 4.
- Morari, M., Arkun, Y., & Stephanopoulos, G. (1980). Studies in the synthesis of control structures for chemical processes: Part I: formulation of the problem. Process decomposition and the classification of the control tasks. Analysis of the optimizing control structures. *AIChE Journal*, 26(2), 220–232. <https://doi.org/10.1002/aic.690260205>
- Nguyen, T., Gosine, R. G., & Warrian, P. (2020). A systematic review of big data analytics for oil and gas Industry 4.0. *IEEE Access*, 8, 61183–61201. <https://doi.org/10.1109/ACCESS.2020.2979678>
- Park, S., & Han, C. (2000). A nonlinear soft sensor based on multivariate smoothing procedure for quality estimation in distillation columns. *Computers & Chemical Engineering*, 24(2–7), 871–877. [https://doi.org/10.1016/S0098-1354\(00\)00343-4](https://doi.org/10.1016/S0098-1354(00)00343-4)
- Petrobras. (2020). Petrobras on refining and natural gas. Retrieved June 27, 2021, from https://www.agenciapetrobras.com.br/Materia/ExibirMateria?p_materia=983055
- Qin, S. J., & Chiang, L. H. (2019). Advances and opportunities in machine learning for process data analytics. *Computers & Chemical Engineering*, 126, 465–473. <https://doi.org/10.1016/j.compchemeng.2019.04.003>
- Qin, S. J., Yue, H., & Dunia, R. (1997). Self-validating inferential sensors with application to air emission monitoring. *Industrial & Engineering Chemistry Research*, 36(5), 1675–1685. <https://doi.org/10.1021/ie960615y>
- Ren, T., Patel, M., & Blok, K. (2008). Steam cracking and methane to olefins: Energy use, CO₂ emissions and production costs. *Energy*, 33(5), 817–833. <https://doi.org/10.1016/j.energy.2008.01.002>
- Rockwell Automation. (2021). Pavilion8 model predictive control. Retrieved July 12, 2021, from Optimize manufacturing processes to maximize plant yield, improve quality and minimize risk website: <https://www.rockwellautomation.com/en-us/products/software/factorytalk/operationsuite/pavilion8.html>

- Sanla, A., & Numnonda, T. (2019). A comparative performance of real-time big data analytic architectures. *2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 1–5. Beijing, China: IEEE. <https://doi.org/10.1109/ICEIEC.2019.8784580>
- Santu, S. K. K., Hassan, M. M., Smith, M. J., Xu, L., Zhai, C., & Veeramachaneni, K. (2021). AutoML to date and beyond: Challenges and opportunities. *ArXiv:2010.10777 [Cs]*. Retrieved from <http://arxiv.org/abs/2010.10777>
- Sasubilli, M. K., & R, V. (2021). Cloud computing security challenges, threats and vulnerabilities. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 476–480. Coimbatore, India: IEEE. <https://doi.org/10.1109/ICICT50816.2021.9358709>
- Schley, M., Prasad, V., Russo, L. P., & Bequette, B. W. (2000). Nonlinear model predictive control of a styrene polymerization reactor. In F. Allgöwer & A. Zheng (Eds.), *Nonlinear Model Predictive Control* (pp. 403–417). Basel: Birkhäuser. https://doi.org/10.1007/978-3-0348-8407-5_23
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Shang, C., Yang, F., Huang, D., & Lyu, W. (2014). Data-driven soft sensor development based on deep learning technique. *Journal of Process Control*, 24(3), 223–233. <https://doi.org/10.1016/j.jprocont.2014.01.012>
- Si, M., Tarnoczi, T. J., Wiens, B. M., & Du, K. (2019). Development of predictive emissions monitoring system using open source machine learning library – Keras: A case study on a cogeneration unit. *IEEE Access*, 7, 113463–113475. <https://doi.org/10.1109/ACCESS.2019.2930555>
- Steurtewagen, B., & Van den Poel, D. (2020). Machine learning refinery sensor data to predict catalyst saturation levels. *Computers & Chemical Engineering*, 134, 106722. <https://doi.org/10.1016/j.compchemeng.2020.106722>
- Sulfur Recovery Engineering. (2021). The claus sulfur recovery process. Retrieved March 21, 2021, from Sulfur Recovery Engineering Inc. website: <https://www.sulfurrecovery.com/sulfur-recovery-process>
- Sun, W. (2020). *Advanced process data analytics* (Doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.
- Sun, W., & Braatz, R. D. (2021). Smart process analytics for predictive modeling. *Computers & Chemical Engineering*, 144, 107134. <https://doi.org/10.1016/j.compchemeng.2020.107134>
- SysML. (2021). SysML open source project - What is SysML? Who created it? Retrieved June 2, 2021, from SysML.org website: <https://sysml.org/index.html>

- Technavio. (2018). *Global artificial intelligence market in the industrial sector 2018-2022*. Retrieved from <https://insights-technavio-com.libproxy.mit.edu/report/global-artificial-intelligence-market-in-the-industrial-sector-analysis-share-2018>
- Technavio. (2019). *Global big data market in the oil and gas sector 2019-2023*. Retrieved from <https://insights-technavio-com.libproxy.mit.edu/report/global-big-data-market-in-the-oil-and-gas-sector-industry-analysis>
- Tham, M. T., Montague, G. A., Julian, M., & Lant, P. A. (1991). Soft-sensors for process estimation and inferential control. *Journal of Process Control*, 1(1), 3–14. [https://doi.org/10.1016/0959-1524\(91\)87002-F](https://doi.org/10.1016/0959-1524(91)87002-F)
- Thompson, K. (2014, June 20). Cyber-physical systems. Retrieved February 9, 2021, from NIST website: <https://www.nist.gov/el/cyber-physical-systems>
- Tidy, J. (2021, May 10). Colonial hack: How did cyber-attackers shut off pipeline? *BBC News*. Retrieved from <https://www.bbc.com/news/technology-57063636>
- Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., & Farivar, R. (2019). Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, 1471–1479. <https://doi.org/10.1109/ICTAI.2019.00209>
- United Nations Framework Convention on Climate Change. (1992). The Paris Agreement | UNFCCC. Retrieved April 16, 2021, from <https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement>
- United States Environmental Protection Agency (EPA). (2013, May 3). Petroleum Refinery National Case Results. Retrieved March 22, 2021, from US EPA website: <https://www.epa.gov/enforcement/petroleum-refinery-national-case-results>
- van der Meulen, R., & Rivera, J. (2014). Gartner says advanced analytics is a top business priority. Retrieved March 10, 2021, from Gartner website: <https://www.gartner.com/en/newsroom/press-releases/2014-10-21-gartner-says-advanced-analytics-is-a-top-business-priority>
- von Stosch, M., Oliveira, R., Peres, J., & Feyer de Azevedo, S. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering*, 60, 86–101. <https://doi.org/10.1016/j.compchemeng.2013.08.008>
- Wang, J., Yang, Y., Wang, T., Sherratt, R. S., & Zhang, J. (2020). Big data service architecture: A survey. *Journal of Internet Technology*, 21(2), 393–405.
- Western Digital. (2021). Data storage for oil & gas companies. Retrieved February 9, 2021, from Western Digital website: <https://www.westerndigital.com/solutions/oil-gas>