# Self-Service Analytics and the Processing of Hydrocarbons

Lim C. Siang[a,b], Shams Elnawawi[a], Darren Steele[c]

[a] *Department of Process Control Engineering, Burnaby Refinery, BC, Canada*
[b] *College of Computing, Georgia Institute of Technology, Atlanta, GA, United States*
[c] *Stiperstone Analytics Ltd. (on behalf of IT Vizion), Shrewsbury, United Kingdom*

**Abstract**

This paper describes the use of self-service analytics on time series process data for the troubleshooting and optimization of refinery operations in the context of data visualization principles and best practices. Refining-relevant examples are used to demonstrate how end-users can access real-time and historical process data and apply the following analytics operations across several refining functions, including (1) incident troubleshooting – identifying periods of interest and methods available to investigate related plant data, patterns, events and disturbances leading up to the incident, and (2) data cleansing – filtering sensor data to remove outliers and bad quality data, splicing and aligning data streams for more accurate analysis and to improve the confidence in the outputs of subsequent analysis, such as the outputs of multivariate, regression-based system identification. The paper also provides examples of how *ad hoc* analyses can be scaled up to plantwide analytics and evolve into routine, automated tasks. The importance of analytic provenance and collaboration in sharing new insights from data is also discussed. To address the human factors associated with self-service analytics innovation, the paper concludes with lessons learnt, observations and adaptations compared to the traditional "business-as-usual" approaches, best practices for data governance, and the implications for engineers that operate in a safety-critical environment.

*Keywords:* Hydrocarbon Processing, Self-Service Analytics, Data Visualization, Process Control

## 1. Introduction

Before refinery operations and process performance can be analysed, it must be measured. Nowadays, measured data is transmitted, stored, and readily available for refinery engineers to consume for real-time monitoring purposes and for basic analysis tasks using tools such as spreadsheets or vendor-supplied trending and charting tools. The analysis of refinery plant operating and process data is a necessity for all refinery engineers. However, conducting these analyses requires engineering knowledge and subject matter expertise that is not typically available from Information Technology (IT) departments or data science functions.

The task of Plant Performance Analysis (PPA) is described as using process measurements to develop models that are a mathematical representation a plants performance [1], and refinery performance monitoring and control can generate increased yields of higher margin products, increased capacity utilization, increased margins, increased quality and optimized energy costs etc. [2].

In addition to analysis of plant data for known, routine, areas of opportunity, there is also a need for *ad hoc* analysis of data for the troubleshooting of emerging operational issues and process disturbances. Therefore, timely access to decision-critical plant data and efficient analysis of the data can significantly reduce the impact of deviations from expected plant behaviour with the potential to reduce the risk to plant, profit, personnel, environment, and surrounding communities. With the prevalence of integrated datasets and increased data accessibility, the variety of analysis work is increasing with data science applications that also consume plant data, introducing further opportunities. Opportunity from an increased depth and variety of analysis is therefore, helping to foster continuous improvement in refinery performance and control.

Self-service analytics is an emerging form of data analysis enabling engineers to perform a wider variety of modelling and analysis tasks. Gartner defines self-service analytics as a form of business intelligence where

> "[...] line-of-business professionals are enabled and encouraged to perform queries and generate reports on their own, with nominal IT support." [3]

Within the context of this paper and hydrocarbon processing, the line-of business professionals are engineers. The types of analysis performed by the engineers, with examples presented herein, are more complex than querying and report generation. Moreover, the principal types of data used in business intelligence querying and reporting is relational and transactional data, however, self-service analytics for plant operations consumes near real-time, and historical sensor data. Therefore, in hydrocarbon processing, the description of self-service analytics as a form of Business Intelligence is questionable and we propose an alternative definition below:

> "Self-Service Analytics (SSA) is the detailed examination of data, performed by subject matter experts and line-of-business professionals with little

> or no input required by Information Technology functions. SSA tools offer ease of access to a variety of data sources and data streams, where the data is analysed with easy-to-use methods and functions to help solve problems and enhance human decision-making."

The usefulness of self-service analysis is determined by the functions and methods available in the software tools that engineers are equipped with, where spreadsheets are the most common tool. Spreadsheets are extremely accessible, require little training to use and spreadsheets enable many kinds of analysis to be performed on plant data. However, the use of spreadsheets should be questioned where data has not been cleansed or pre-analysed to remove bad-quality data, where datasets have measurement gaps, are subject to measurement noise, or where there is a need to remove data that would diminish the confidence in the output of any analysis. Consider a refinery with 50,000 sensors, storing 50,000 measurements each minute for the last five years: this refinery would generate 131 billion data points for the five-year period.

Spreadsheets and relational-data based dashboard applications may also be challenged to handle the quantity of data required for some types of sensor derived, time series data analysis, especially considering the amount of data generated by a refinery. Furthermore, self-service analytics often requires functions and methods that are beyond the original design intention of spreadsheet and dashboard applications such as,

- complex engineering calculations
- multi-variate based estimations and forecasts
- the generation of virtual sensors
- the cross referencing with plant events that have been derived from disparate data sources
- data cleansing operations
- scenario testing

The timeliness and variety of analysis, the need to include subject matter expertise, the quantity of data and the need to cleanse the data using engineering knowledge, all support the notion that self-service analytics will grow beyond the current toolsets and will require a dedicated branch of computing techniques.

In the literature, there are many competing definitions for the terms *'data'* and *'information'*. Rowley (2007) presented a comprehensive survey on the data-information-knowledge-wisdom (DIKW) model widely adopted in information science [4]. According to Rowley, the term 'data' is often defined by authors in terms of what it lacks; as a form of 'raw facts' obtained from measurements and operations, data lacks meaning and value without context. 'Information' is then defined in terms of data, as a higher-order form of raw data that has been formatted, processed or organized to impart meaning and value. As summarized by Rowley, several authors further argue that humans determine whether a message that they receive is data or information, based on their prior experiences and ability to assign meaning and significance to the message received.

The definitions of *data*, *information*, *knowledge* and *wisdom* have been a long-standing debate in many fields, and we do not seek to review all competing definitions nor argue about the semantics of higher-order structures like knowledge and wisdom. For the purposes of this paper, we subscribe to a simple understanding that information is a useful form of contextualized data that can help engineers with sense-making and decision-making. However, these basic distinctions are important, because chemical process data is commonly characterized as *data rich but information poor* [5]. In refining, big data is not necessarily good data. Despite the large volumes of data collected, extracting useful information from process historians is a non-trivial task. Furthermore, in the context of refinery process control, the plant is designed to operate at steady-state, a condition which typically provides little to no useful information for controller design, which is why step tests are often needed to excite the plant prior to process modeling.

In this paper, we present three case studies from the Burnaby Refinery describing real-world challenges in process control applications and demonstrating how self-service analytics could serve as an effective solution. Through these case studies, we analyze the utility of self-service analytics tools *vis-à-vis* classical methods by applying principles from the data visualization and human-computer interaction (HCI) literature, and explain how these tools could help translate raw data to value-added information for refinery engineers. We report qualitative feedback from practicing engineers on their experiences with self-service analytics and discuss our perspectives on the benefits and challenges of analytics and digital solutions in the context of refinery engineering applications.

## 2. Literature Review

Chemical engineers have been asked to work with larger and more complex datasets due to a proliferation of inexpensive instrumentation and wider availability of data in recent years [6]. However, chemical engineering education in general, has not kept pace with the skills required to manipulate, visualize and analyze these large datasets. Many chemical engineers often struggle with modern data-related tasks if their computing skills are limited to classical methods, such as manual data manipulation in spreadsheets, or simple univariate visualization of time series [6]. Likewise, several authors argued that the computational needs of practicing engineers have now expanded beyond simple engineering calculations, and mastery of advanced data manipulation skills is essential in the workplace [7, 8, 9]. Top-performing engineers are highly productive, motivated and skilled in integrating messy, disparate data sources to uncover engineering insights, and these engineers are functionally working as data scientists with deep engineering domain expertise [10]. However, these digitally-savvy individuals with a combination of strong domain knowledge, subject matter expertise and understanding of computational tools are rare in an organization [11].

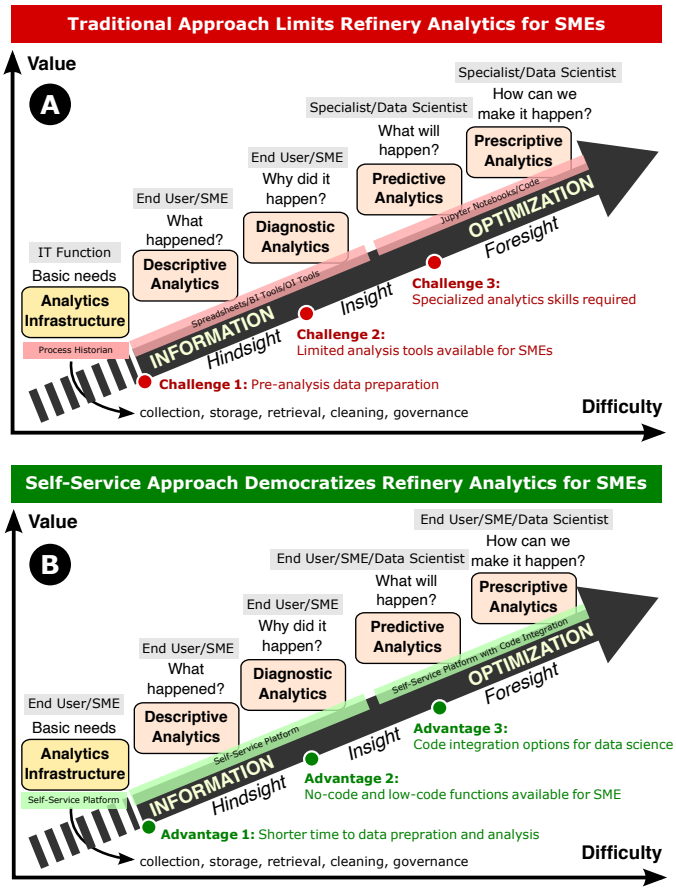


Figure 1: Four types of data analytics from the Gartner analytics ascendancy model [12] – descriptive, diagnostic, predictive and prescriptive, with increasing value and execution difficulty, starting with basic data infrastructure needs (data collection, storage, retrieval, cleaning, governance etc.) as the foundation for all higher level tasks. In top half of the figure (A), the traditional approach and challenges are illustrated in red, contrasted with the bottom half of the figure (B) where the self-service approach and its advantages are shown in green.

The Gartner analytics ascendancy model [12] describes four major types of data analytics, corresponding to an organization's analytics maturity, as shown in Figure 1. *Descriptive analytics*, which answers questions around 'what happened' in hindsight, are commonly implemented as process trends, dashboards and KPIs (Key Performance Indicators). *Diagnostic analytics* attempt to uncover root causes and provide insights on why an event happened. *Predictive analytics*, such as forecasting and simulations, involve making predictions of what will happen in the future. In the process industries, predictive analytics using machine learning techniques have been gaining popularity in recent years [14]. Finally, *prescriptive analytics* help stakeholders with decision-making and prescribe actions to optimize the business. Each step provides increasingly valuable insights, but they are also progressively more difficult to execute accurately.

Maslow (1943) introduced the concept of a hierarchy of human needs, where lower-level physiological needs including food, water and shelter must first be satisfied, before higher-level needs like self-actualization can be achieved [15]. Data science practitioners and researchers have alluded to the idea of a data hierarchy that mirrors Maslow's insights [16, 17, 18, 19]. As illustrated in Figure 1, we can observe an increasing sophistication of data analytics needs, forming a *data-value* analogy of Maslow's hierarchy, where the engineer's basic data instructure and exploration needs must be adequately satisfied first, before advanced analytics can be successfully executed.

We can further distinguish between a traditional approach to refinery analytics and a self-service approach. The traditional approach has several challenges – (1) low-value but essential data preparation work must be done before value-added analysis can begin, (2) limited analysis functions are available to Subject Matter Experts (SMEs), typically in the form of spreadsheets, BI or visualization tools, (3) advanced analytics are performed in external tools by skilled specialists, and may not be directly integrated with lower-level analyses. The self-service approach provides a solution to these issues by integrating multiple data sources and analysis types in a single, unified platform that democratizes analytics to the wider organization. This approach equips SMEs with a broader spectrum of analysis functions, including those traditionally provided by IT functions, data scientists and third-party specialist applications. A self-service platform integrated with the process historian also provides data scientists and specialists with a mechanism to easily consume operational technology (OT) data using a no-code/low-code approach for data preparation work, resulting in a shorter time to value-added analysis for stakeholders.

An organization that is just starting their analytics journey would likely have many low-hanging fruits at the descriptive analytics and diagnostics analytics level that are yet to be exploited, potentially with significant business and productivity impact. After all, how can one manage machine learning projects effectively without the ability to easily access and visualize the data to begin with? Furthermore, it has been documented in the literature that the oil and gas sector is lagging behind many other industries on the analytics maturity curve [20] and the adoption of new technology [21].

Perrons et al. (2015) argued that the oil and gas industry, in general, still treats data as facts describing the state of an asset, whereas leading digital industries, such as software companies, understand that *the data itself* is invaluable for identifying complex patterns and hidden relationships [22]. This is a subtle and interesting point that requires further elaboration. As articulated by Perrons, the oil and gas industry generate massive datasets, many of which are only given a cursory glance, and much of it are simply archived away unless needed for specific scenarios, such as reporting, operational monitoring or engineering investigations. Leading digital organizations, however, understand that the value of big data analytics does not arise by monitoring known relationships or testing hypotheses on known variables. Rather, the true value lies in finding latent patterns and making predictions on complex relationships that were previously unknown, using the entirety of the organization's data, regardless of how disconnected and inconsequential the data might seem when it was initially collected.

Marshall and colleagues [23] investigated the connection between analytics and innovation in business organizations. Their study shows that, not only are leading organizations capable of

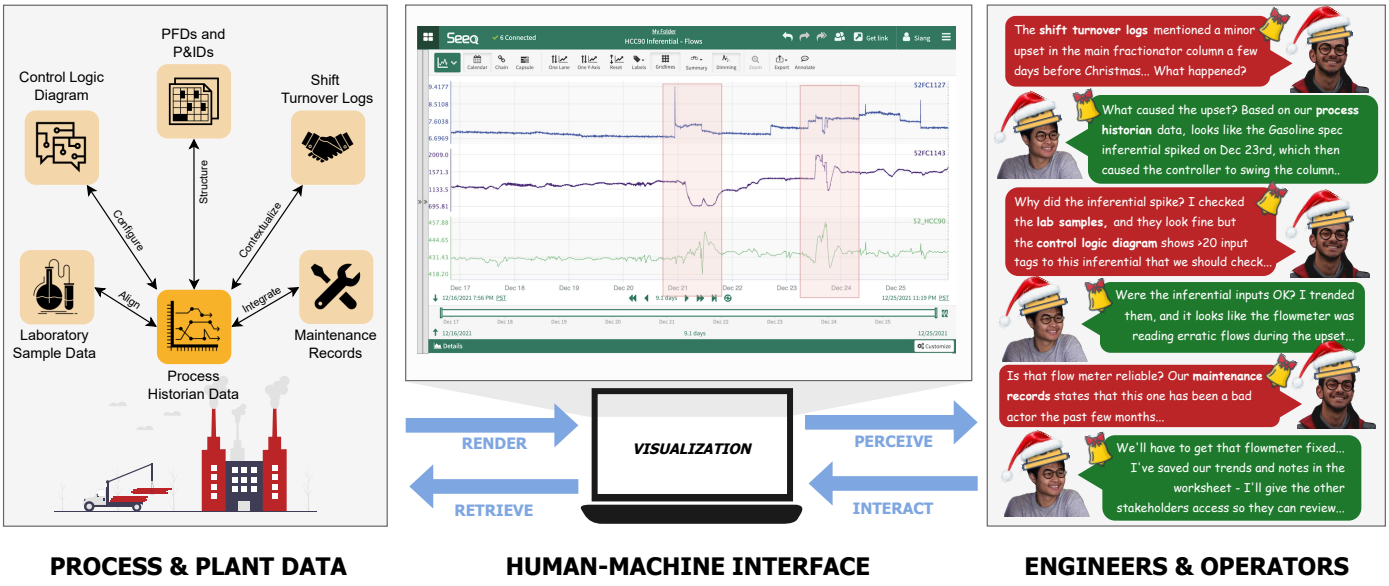


Figure 2: Engineers collaborate and interact with the underlying refinery data through visualization tools. As shown in the rightmost column by a representative conversation during a collaborative troubleshooting effort, engineering investigative tasks typically employ iterative interrogation techniques, such as the *5-Whys* methodology, and rely on the engineer's domain knowledge, reasoning skills and analytical capabilities. Moreover, refinery engineers often need to piece together information from disparate data sources throughout the investigative process, and these tasks can be supported by an effective data analytics and visualization platform to facilitate sensemaking and collaboration, with consideration of human factors [13].

extracting valuable information from different sources and performing deep analysis to obtain competitive insights, but they do so in a structured and targeted approach with 3 key strategies. These leaders (1) promote excellent data quality and accessibility, (2) make innovation integral to every role by investing in their employees' analytics training and (3) build a quantitative innovation culture to effectively measure their success. They also found that organizations struggling with analytics and innovation are typically more risk averse in nature, and innovation, if any, happens only in isolated pockets in the organization rather than in a strategic manner.

A 2015 article in Forbes noted that *early adopters* of technology are no longer common in the oil and gas industry, and the expectation of innovation has weakened compared to the industry's early days [24]. Business leaders and engineers in the oil and gas sector continue to accept the most basic of tools and make do with outdated technology, instead of pursuing better options and finding newer, more efficient ways of working. The article asks, given the pioneering nature of the oil and gas industry and its early innovations, ironically, why are these oil and gas companies now resistant to innovation? How did this 'spirit of modernization' get lost? Roberts et al. (2021) discussed the implications of these trends in the oil and gas industry and investigates the underlying reasons [21]. They identified 6 contributing factors, including, risk aversion, organizational culture and other psychological factors, as illustrated in Figure 3. The authors describe the oil and gas sector as an industry that exemplifies resistance to technological innovation, yet it is also an industry that must innovate to remain competitive and survive, which is indeed, an interesting paradox.

In the refining industry, driving innovation and gaining a competitive engineering advantage through data analytics may not necessarily involve complex models or cutting-edge machine learning techniques. Rather, having timely, high-quality information is important to aid engineers and subject matter experts with decision making to achieve better business outcomes such as increasing yields and reducing energy consumption. Similar views have been held by other multinational oil and gas companies [25], which we discuss later in this paper.

Engineering investigations are a routine analytics task for engineers in the downstream refining industry, where domain expertise is indispensable [14, 26]. These investigations involve hypothesis development, sensemaking activities and narrative building to construct a deep understanding of the process data; it is an activity that goes beyond straightforward information visualization tasks like identifying trends and outliers [27]. These investigative tasks are also cognitively challenging, requiring engineers to make sense of a large collection of data that relies heavily on their reasoning and analytical capabilities to find root causes [28], based on insights from data visualization, as illustrated by a representative troubleshooting conversation in Figure 2.

Engineering investigative questions often involve visualization of a conditional subset of large, multivariate time series data when the plant is operating at certain states or conditions. Several representative questions in a downstream refining context are presented in Table 1, and classified using MacEachren's task descriptions [29, 30]. These conditional filters can be set up using rudimentary SQL queries or Python scripts. However, the typical refinery engineer with a chemical engineering background may not have the expertise to write these queries effectively, and may resort to performing the analysis in a more

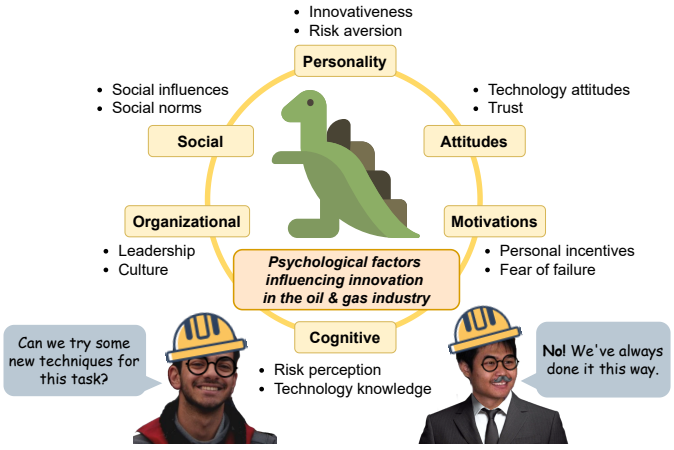


Figure 3: Corporate gatekeepers could either be facilitators or barriers to the adoption of new technologies in an organisation. Roberts et al. (2021) explored psychological factors influencing resistance to technological innovation in the oil and gas industry and identified 6 overarching categories – personality, attitudes, motivations, cognitive, organizational and social [21]. Gatekeepers and decision makers who were overly risk averse and uncomfortable with new technology were described as *dinosaurs*.

familiar environment like spreadsheets, which would be very inefficient, due to the laborious overhead of retrieving, loading and cleaning the data first prior to performing any value-added analysis. Furthermore, for data filtering tasks, separating the data retrieval task from the data visualization task would impose a heavy cognitive load on the user. Research has shown that the most effective implementations of filtering are those that immediately update the display, allowing users to quickly see how the filters affect the data representation [31].

Clearly, without the right tools to facilitate data wrangling and visualization, relatively simple tasks can be very tedious, as engineers will need to write custom, potentially complex *ad hoc* queries that may be beyond their expertise, or work with cumbersome configurations and data manipulations, or worse still, be dependent on someone else to query and clean data for them [9] before any productive and value-added work can begin. Any improvements to tools and methodologies at the descriptive and diagnostic levels would be helpful for engineers in their day-to-day tasks and lead to a competitive advantage, even before resorting to more powerful machine learning or advanced analytics techniques.

To tackle these challenges, self-service analytics has the potential to quickly empower engineers with a stronger ability to work with data and take advantage of their existing domain expertise, without expensive and time-consuming data upskilling programs, as well as alleviate the pressure and workload on strained IT departments to support additional analytics duties [9, 32].

Seeq Corporation's software product is one example of a self-service analytics system and is the SSA tool used to prepare the case studies herein. In a recent ARC Advisory Group webinar, Shell shared their digital transformation roadmap, which included self-service analytics solutions built in Seeq [25]. The ARC Advisory Group listed the following key components of

Table 1: A taxonomy of event-based refinery engineering investigation questions based on MacEachren's [29] temporal domain task descriptions, as presented by Aigner [30].

| Type | Representative Quote |
|---|---|
| **Existence:** Did an event occur? | Did the alarm annunciate during the upset? |
| **Temporal location:** When did an event occur? | When did we last operate at low temperatures? |
| **Temporal interval:** How long did an event last? | How long was last week's process upset? |
| **Temporal pattern:** How often does an event occur? | How often do we backflush the pump? |
| **Rate of change:** How fast is the event data changing | When do we expect the tank to reach capacity if we hold the current rates? |
| **Sequence:** In what order do the events occur? | Did we see any changes in the unit before the chloride concentration went up? |
| **Synchronization:** Are the events co-occuring? | Was the control loop in manual during feed-in? |

Seeq in their 'Industry Best Practice' report on Shell's SSA approach,

> "Using a combination of monitoring and descriptive, diagnostic, predictive, and prescriptive analytics, Seeq empowers users to determine what is happening now, why it happened, what happened in the past (and why), what will happen, and what should happen. This is allowing Shell to move from reactive to proactive operations." [25]

In this paper, we consider the usage of Seeq for self-service refinery analytics. Note that although Seeq is the tool used in this paper, our focus is not on a usability study of Seeq, nor a comparison of Seeq with other SSA tools, but rather a broader discussion of SSA concepts demonstrated with refinery engineering applications. We present three case studies from the Burnaby Refinery to illustrate the utility and advantages of self-service analytics compared to classical methods like spreadsheets. Applications of refinery analytics are discussed in the context of of data visualization principles and best practices through a user-centered, event-based visualization framework [30, 33].

## 3. Case Study 1: Conditional filtering of time series data

Our first case study illustrates how time series data filtering and visualization can be effectively performed using a no-code approach in self-service analytics tools.

Information visualization systems consist of two components: representation and interaction. Representation is related to how the data is rendered on the display, whereas interaction involves a discourse between the user and the system to query and explore the dataset in a goal-directed manner [34, 35]. These two components are not mutually exclusive, as user interaction with a system may result in changes to its representation [36]. Zooming and filtering are interactions that reduce

the complexity of the data representation by removing irrelevant information from the view [31]. Filtering can be achieved by user-defined conditions and ranges, such that only a subset of data satisfying the specified conditions are presented, and zooming can be considered a form of filtering by navigation [36].

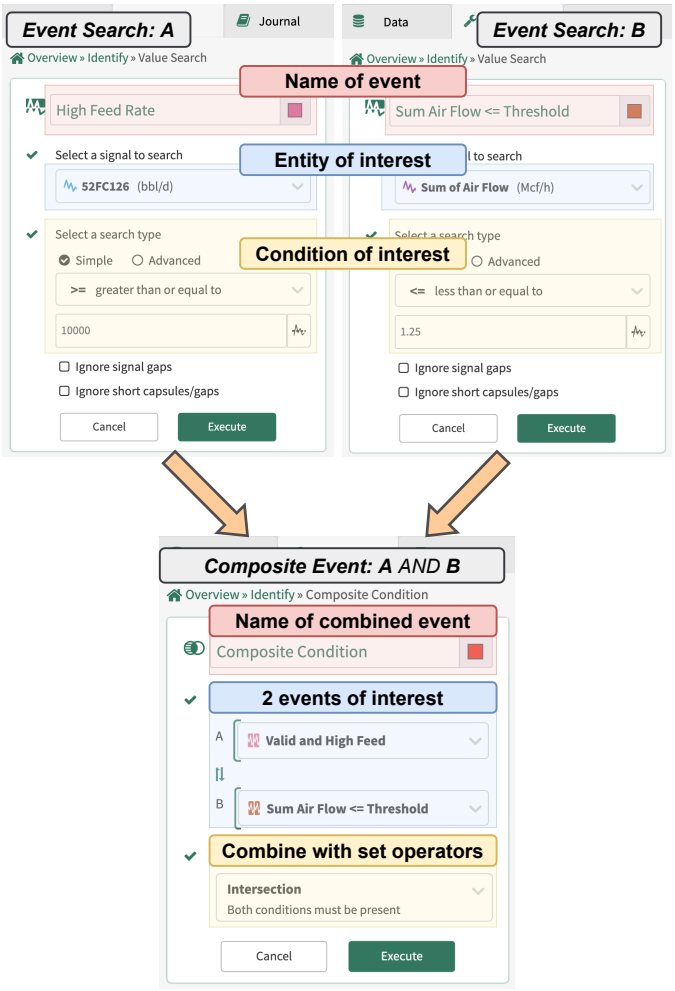


Figure 4: A visual event editor menu allows the user to quickly and easily express conditional filters as well as composite events in an intuitive, visual manner without writing code or textual queries.

**Case Study 1:** During a routine process audit, engineers at the Burnaby Refinery discovered that a low-flow trip setpoint on a compressor may be ineffective, and a higher trip set point was recommended by the team for safety reasons. However, the operations team raised concerns about the reliability of the compressor's flow meter readings, as they had observed sporadic dips in the measurements. A setpoint increase may lead to spurious trips and production losses due to faulty flow measurements. As part of their due diligence process, the engineering team wanted to identify time periods in the past 10 years with measurement anomalies to determine exactly how often these sporadic blips occur. The conditional filters of interest included (1) feed rate exceeding a threshold, to indicate normal operations, and (2) compressor air flow rates below a threshold, to indicate measurement anomalies. In other words, in the framework of *event-based visualization* [30], given a multivariate time series $x$, we are interested in finding a subset of $x$ such that $\{x \mid (x.\text{flow} < \varepsilon_1) \cap (x.\text{feed} > \varepsilon_2)\}$, where $\varepsilon_1$ and $\varepsilon_2$ are the user-defined thresholds.

**Challenges:** The anomalies may last only a few seconds, so high-resolution data from the process historian is needed for the analysis. An attempt at performing the analysis in spreadsheets was eventually determined to be infeasible due to the large, cumbersome dataset. The high-resolution data retrieval and conditional filtering formulas were time-consuming to configure and execute, as the entire dataset needed to be retrieved first and stored in the spreadsheet before the computations can be performed. More performant in-database queries, such as executing textual queries directly in the process historian were considered, which would've avoided the need to load the entire dataset into a spreadsheet first, but the team involved in the analysis did not have the familiarity and expertise to correctly express the conditional filters as historian queries. Studies have shown that textual event specifications, such as historian or SQL queries, are often intimidating to novice users [33], and the usage of visual editors can help users express their needs more easily, especially for defining complex, composite events.

**Solution:** Using a visual editor, the team was able to rapidly configure the conditional filters of interest using an intuitive graphical interface as illustrated in Figure 4. These 'conditions' are also known as *'event types'* in the framework of event-based visualization [30]. A combination of event types can be chained together using set operators (AND, OR, NOT etc.) or *temporal predicates* (`before`, `after`, `overlaps` etc.) [33] to form composite events in the visual editor. In Seeq, the subset of data that matches the condition or event type is marked as a collection of *capsules* at the top of the time series trends. Capsules are encoded with different colors, representing periods of interest defined by the user to add context to time series data. Capsules can be considered a form of visual representation of an *event instance* [33].

To drill down into the relevant data, a *chain view* feature modifies the display window to hide a subset of data that do not match the conditional filters and isolate the time periods of interest, as shown in Figure 5. This chain view concept aligns with Shneiderman's widely-held data visualization principle or information-seeking mantra of *"overview first, zoom and filter, then details on demand"* [37] and is recognized to be effective for managing visualization needs for large, complex datasets. The built-in capsule summary table in Seeq further provides a count for exactly how many times these measurement anomalies occurred in the past 10 years.

**Results:** Using these new insights, and comparing it to the trip logic in the DCS (Distributed Control System), the engineering team was able to verify that the new setpoints would have a low risk of spurious trips due to redundancy in the flow measurements and the existing logic configuration. This exercise increased confidence in proceeding with the setpoint change to improve process safety. In contrast with a classical spreadsheet-based method, which took several engineers at least 40 hours performing low-value work such as tweaking historian settings,

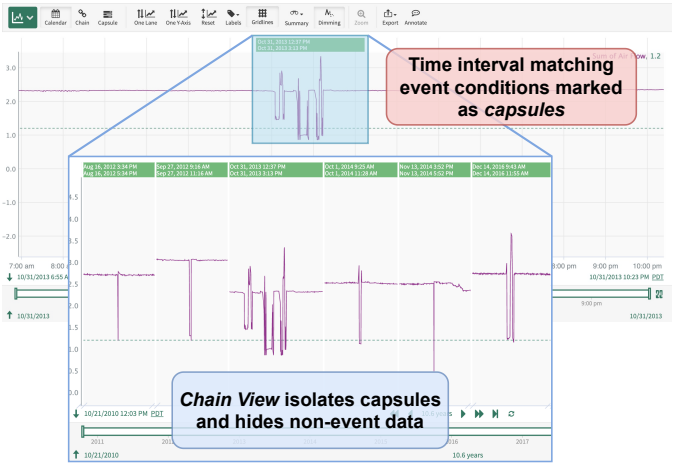


Figure 5: Shneiderman's information-seeking mantra of 'overview first, then zoom and filter' implemented as a *chain view* feature in visualization tools. Chain view stitches together subsets of the data that matches the active conditional filters to reduce visual noise and clutter. By hiding irrelevant data, visualization tools can help isolate time periods of interest for the user, allowing users to drill down and observe relevant capsules for deeper insights.

waiting for data retrieval, and manually inspecting spreadsheet calculations, without much confidence in the final results, the visual interface allowed the team to rapidly and efficiently obtain the right answers in less than an hour using an intuitive visual editor for specifying complex events and querying a large dataset in a no-code manner.

## 4. Case Study 2: Inferential performance assessment

Our second case study illustrates how self-service analytics can be used for time series data cleaning and computations with a low-code approach in the context of inferential performance assessment. Inferentials are also known as soft sensors, and are widely used in the refining industry to estimate process quality variables such as product compositions and boiling points in the absence of expensive online measurements. The true values are measured offline using lab equipment, which offers a limited number of samples with significant time delay, typically in the order of hours or even days. These lab values are then used to bias or correct the inferential model predictions, using a technique typically referred to as sensor fusion in the literature [38].

**Case Study 2:** Due to operational, equipment or process changes, inferential accuracy will degrade over time and periodic model updates are needed to maintain its performance [39]. A common method of measuring inferential performance is the usage of residual-based KPIs, where the inferential predictions are compared to the 'true' lab values. Due to the time delay in obtaining lab values, the 'true' lab value in the residual calculations must be compared with the inferential prediction at the sampling time, and not the time when the lab results are available [39, 40]. In other words, we wish to compute the absolute residual, $e_i = |\hat{y}_i - y_j|$ where $y$ is the lab value sampled at time index $t_i$ with results obtained at a later time, $t_j$, and $\hat{y}$ is the prediction at time index $t_i$. Due to the time delay, the lab sample collected at $t_i$ will not have results available until a later time index $j > i$, typically in the order of hours or days, as shown in Figure 6.

**Challenges:** At the Burnaby Refinery, the DCS logic for inferentials is configured such that an indicator signal, $\mathbf{1}_A$ is set to 1 when $A$ is true, where $A$ is the event that a sample is taken but the lab results are unavailable yet, and 0 otherwise. Due to the time delay between the sample and results as described earlier, there is a need to track the sampling time and align the time stamps of the lab results with the inferential predictions for correct computation of the residuals. This can be achieved shifting the irregularly sampled $y$ backwards based on the indicator signal, $\mathbf{1}_A$. An early attempt to setup the residual calculation as a textual historian query or manually in spreadsheets proved to be too cumbersome and tedious to align the timestamps.

**Solutions:** Converting the indicator signals to capsules is the first step in using self-service analytics for this case study. The start of the capsule at time index $i$ would correspond to the prediction $\hat{y}$ and the end of the capsule at time index $j$ would correspond to the lab results $y$ collected at time $i$. The correct residual calculation is then simply the absolute difference between the first $\hat{y}$ value at the start of the capsule and the last $y$ value at the end of the capsule, as illustrated in Figure 6. The usage of an indicator function helps provide a straightforward method to compute the residuals, since the time delays between sampling and results are always different for each sample depending on the lab turnaround time. By aligning the data using the indicator function and capsule length, users can avoid configuring messy time-varying, time-shift formulas to align the signals, as the correct computations would be implicitly handled based on the start and end times of the capsules.

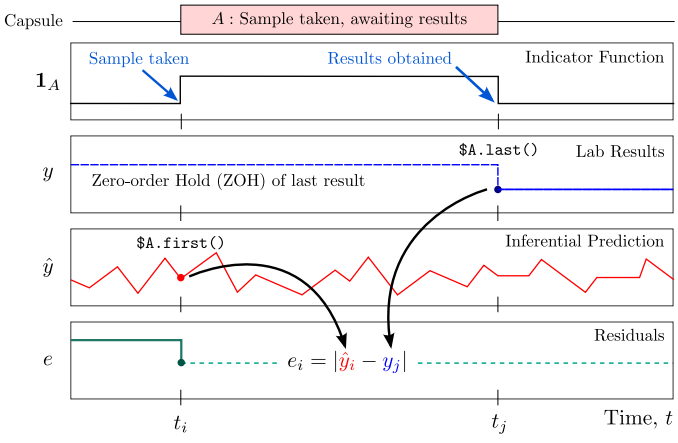


Figure 6: Process quality variables are typically measured offline using lab equipment with a time delay in the order of hours or days. The lab results are used to bias or correct inferential model predictions. For accurate computation of the residuals, i.e. the difference between the inferential and lab results, the lab result timestamp must be aligned with when the sample was taken, and not when the result was available. This re-alignment step can be facilitated using an indicator function, $\mathbf{1}_A$, where $A$ denotes an event that a sample was taken but results are not yet available. The points being aligned are marked with `$A.first()` denoting the start of the capsule, and `$A.last()` denoting the end of the capsule.

**Results:** The engineering team at the refinery was able to

easily configure the correct inferential monitoring calculations using self-service analytics methods, compared to previous attempts in spreadsheets with manual timestamp alignments. The analytics platform used establishes a direct connection to the process historian, avoiding the need to manually retrieve data or determine the correct historian configuration and data resolution settings. Calculations are cached, allowing the user to perform interactions like zooming, filtering and navigating between different time periods without expensive, repeated computations. Hierarchical 'asset trees' are used to scale up the calculations from a single inferential to multiple inferentials refinery-wide without repeating the analysis or copy-pasting formulas. Reporting dashboards are configured to share results with stakeholders. These topics, and the concept of *analytic provenance* [41], are discussed later in this paper.

## 5. Case Study 3: System identification

Our final case study focuses on the development of new process models using system identification for the Diesel Hydrotreating (DHT) unit at the Burnaby Refinery. The DHT unit is primarily responsible for using hydrogen to remove sulfur from diesel products, taking in untreated diesel, jet fuel, and similar components from other upstream units as feeds. The feeds are combined, preheated, hydrotreated, and the different desulfurized fractions are separated [42]. For controlling such processes, the APC (Advanced Process Control) model can be represented by a set of 'process response curves' for each input-output pairing in the system, and each of the feed products is treated as a separate input due to differing physicochemical properties. A new initiative at the Burnaby Refinery involves co-processing canola oil as a renewable feedstock in the DHT unit, as described in the organization's quarterly filings, and annual Sustainability Report [43]. To improve DHT control, the canola feed rate must be in incorporated into the APC model as a feedforward variable so that the controller can maintain product specifications by adjusting DHT reactor temperatures in anticipation of varying canola feed rates by using historical closed-loop data. The engineering team's goals when running this identification on canola oil feed rate was meant to be a 'sanity check' on engineering judgment, aiming to obtain a rough measurement of the canola-sulfur gain and see if the result was in line with the actual sulfur adjustments made through manual control in the Burnaby Refinery, before performing an actual step test that would disrupt production. The workflow involved in system identification is summarized in Figure 7.

**Case Study 3:** In APC projects, to avoid MV correlations that affect model quality, engineers typically design step tests such that only one manipulated variable (MV) is stepped while keeping all other MVs steady. The responses in all controlled variables (CVs) are measured once they reach steady state; the MV moves and CV response data are then used for system identification, typically using regression-based methods. Performing system identification using historical closed-loop data follows a similar procedure, but MV inputs will not be intentionally held constant during the measurement periods. To put

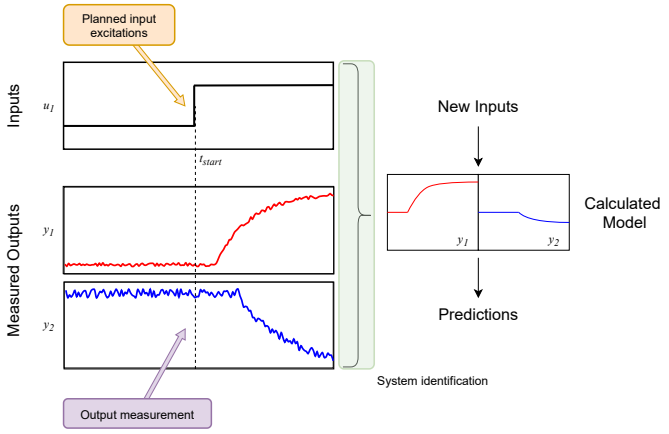


Figure 7: Simplified workflow of the system identification proces. Existing process data (left) are correlated to develop a mathematical model, shown as a series of process response curves (right). The model then takes in new inputs and generates predictions.

this in more specific terms, given historical process data consisting of a single input variable $u_1$, with a set of output variables, $Y = y_1, y_2, \cdots, y_n$, we wish to identify the relationship between $u_1$ and $Y$ by filtering for regions where $\frac{\mathrm{d}u_1}{\mathrm{d}t} > r_{\min}$. The threshold, $r_{min}$, describes the smallest derivative of $u_1$ that may constitute a 'step', such that the move size in $u_1$ is big enough to register a reasonable response in the output variables. Data capsules meeting this criterion are filtered out, but not all of these capsules will be usable for identification, largely due to the presence of excessive movement in other input variables. The objective here is to decide which capsules are useful for validating engineering judgment, based on a 'snapshot' of variable behaviours in that time. This selection operation is referred to as 'brushing' in the realm of event-based visualization [33, 44].

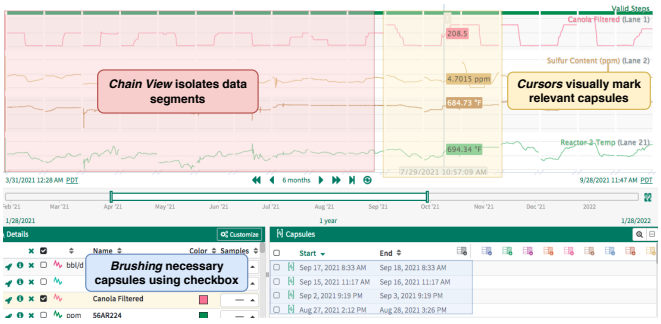


Figure 8: Due to the necessity of manual data selection rather than automatic filtering, viewing capsules side-by-side provides an important advantage. Capsules can be marked using visual markers for a more apparent visual filtering, and are selected via the checkbox list. Brushing by such selection allows users to easily add/remove the data they need to use.

This case study aims to run a simple system identification exercise to obtain an estimate of the steady-state gain to confirm engineering judgment and operational observations – the goal is not to run a full system identification to find accurate process dynamics, but to just estimate the process gains.

**Challenges**: Data cleaning is important due to noisy process data. Using a simple derivative of $u_1$, we can identify

potentially useful data capsules, but the other MV inputs still fluctuate significantly within the capsules, which will affect the gain calculations. The capsules required some additional filtration in order to isolate the ones useful for identification. However, finding quantitative parameters that filtered the data so that there was a sufficient number of useful capsules proved to be a difficult task, because any threshold values chosen for simple filtering were arbitrary. An example of this is filtering capsules by the variances of the MVs that are not being tested. Heuristic values, known as 'typical moves', can be used as benchmarks for normal variable movements, but imposing an exact filter based on the typical move is a difficult decision to justify. As such, obtaining a useful final dataset of the data segments was a judgment call for the control engineers involved, which meant that the task required strong process knowledge. Manual selection of useful capsules must be done by control engineers and can be facilitated by a brushing interaction, where users identify useful capsules one-by-one and select them.

**Solutions:** As illustrated in Figure 8, the resulting capsules are shown in Chain View. Each capsule corresponds to a 'step' that can be used for identification. The steps are isolated by imposing a threshold on the derivative of canola feed rate and 'padding' those points with two hours before and four hours after them; this is to ensure the effects of co-processing on the DHT are captured in their entirety based on process knowledge of the time to steady state. The visualization in Figure 8 is based on a few key variables, but variables can be added on-the-go based on what the control engineers decide they should look out for, to certify the validity of each step. Usable capsules are brushed by selecting them in a checkbox list, such that the user and collaborators can visually identify them. Steady-state gain is calculated for each capsule using the following:

$$K_{ij}\Big|_c = \frac{\Delta y_i'\Big|_c}{\Delta u_j'\Big|_c} \tag{1}$$

Where $K_{ij,c}\Big|_c$ is the steady-state gain between output $i$ and input $j$ over the capsule $c$, and $\Delta y_i'\Big|_c$ and $\Delta u_j'\Big|_c$ are the steady-state changes in the deviations of output $i$ and input $j$ over the capsule $c$, respectively. The refinery's APC system identification tool is also used to perform this calculation, and the gain estimates are compared between both methods using the Seeq method as a sanity check.

**Results:** With respect to the initial goal of this case study, this experiment was successful in using self-service analytics to run a simple form of system identification for quickly validating existing process knowledge without resorting to full-featured APC tools or actual step tests. Steady-state gains obtained using simple self-service analytics methods are found to be within about 10% to those from the relevant APC tools. Additionally, such open-source tools can improve system identification in self-service analytics by making it highly extensible and supported by the controls community, for example the Seeq SysID toolbox [45]. By providing an extensible framework through a software development kit, self-service analytics tools can allow users to extend their analyses beyond the platform's base functionalities.

Aside from the main goal of system identification, this case study establishes a broader framework for *ad hoc* self-service analytics, giving engineers at the Burnaby Refinery more self-service options and familiarity in testing process hypotheses and validating engineering intuition before making actual changes to the APC system. Using these toolboxes, engineers at the refinery can quickly test theories on hypothetical APC model changes on-the-go without having to go through additional time-consuming barriers like putting in formal requests to schedule step tests and deploy APC changes.

## 6. Business outcomes and advantages

### *6.1. Business Impact*

To qualitatively assess the time-in-motion savings and productivity improvements from the case studies presented and other engineering use cases, we've conducted verbal, semi-structured interviews with engineers at the Burnaby Refinery involved in the trial of self-service analytics tools to understand their experiences. This open-ended, semi-structured interview approach is inspired by the work of Kandel and colleagues [9]. We present relevant quotes and responses below, edited for clarity:

> "I was impressed by this tool and how we were able to quickly gather information. In just minutes, we determined that the new trip setpoint would have resulted in possibly one spurious trip in the past ten years. Analyzing that amount of data with spreadsheets would take us so much longer."
>
> (Engineer 1, on data volume)

> "I like how we are able to quickly filter data and reduce the time folks spend building trends for troubleshooting. I'm supportive of finding ways to help folks work more efficiently, and this looks like it would. We all know our biggest challenge is not having enough time in the work day to get everything we want done, so this should help."
>
> (Engineer 2, on productivity)

> "I now realize how powerful it is when we connect it to Python, there are just so many opportunities and possibilities, rather than just waiting for our vendors to develop new features, which is nice, but sometimes quite limited."
>
> (Engineer 3, on code integration)

> "I feel comfortable using it after 2 short training sessions because it is similar to our existing tools, and appears to be designed by chemical engineers for chemical engineers, unlike our more complex unit monitoring tools, which took me hours to add a variable to it due to the steep learning curve (and we had 2-3 full days of training for that!)."
>
> (Engineer 4, on user-friendliness)

> "Past incidents and investigations would have benefited from this tool. I often build PI trends for these investigations but it is tedious and limited to time-based graphs."
>
> (Engineer 5, on troubleshooting)

The Burnaby Refinery's peers and other oil and gas organizations have also reported similar insights and benefits from their SSA approach. We summarize findings by multinational oil and gas and petrochemical companies like Shell, Chevron, Covestro and Sinopec below.

As reported by Shell, their organization has been transitioning away from traditional spreadsheet approaches for data analysis to Seeq to 'better leverage its vast data store' [25]. They raise an interesting point that self-service analytics bring significant value to problems that 'do not initially appear to be overly complex'. As elaborated further in a ARC Digital Transformation Council webinar [46], Shell explained that many engineering problems they investigate and solve are relatively simple in terms of data analytics needs and algorithmic techniques, especially if they already know beforehand the relevant physics and first-principles theories to explore. The value of self-service analytics, however, is that SSA tools make it significantly easier for their engineers to test and verify multiple competing theories. In the webinar, they provided case studies to elaborate on these comments, noting that without their deployment of SSA tools, their subject matter experts would most likely have made incorrect, costly assumptions in root cause analyses due to the inability to test and verify their theories in a timely and accessible manner.

The case studies herein demonstrate *ad hoc* and routine types of SSA with some diagnostic troubleshooting and optimization opportunity identification analysis. SSA also provides predictive and prescriptive types of analysis, as demonstrated by Covestro in an application to predict heat exchanger fouling and anticipate when equipment will need to be cleaned or serviced so they can schedule downtime and mitigate disruption [47].

Decision support for refinery engineers requires more than time-series sensor data alone. Operational intelligence systems must therefore integrate a variety of data types from a variety of data sources in order to provide more holistic decision support, i.e. sensor data from process historians, Laboratory Information Management Systems (LIMS) data, maintenance work order data from Enterprise Asset Management (EAM) systems, planning and scheduling data, and operator shift logs etc. One company that has implemented an operational intelligence system has reported that the system contributed to significant benefits in Chevron's El Segundo refinery [48], as shown in Table 2.

Table 2: Benefits reported for Chevron's El Segundo Refinery following implantation of real-time operational intelligence system [48].

| **Business Improvement Category** | **Multi-Year Average** |
|---|---|
| Reduced Operating Expenses | 8% |
| Increased Facility Utilization | 8.5% |
| Increased Operational Availability | 2.5% |
| Increased High Value Product Production | 10.5% |
| Reduced Environmental Incidents | 18% |
| Reduced OSHA Recordable Injury Rate | 39% |

Despite the lower relative percentage improvement for availability in Table 2, the absolute monetary weighted benefit is significantly higher since the majority of operating expenditure for a refinery is maintenance-related costs. The Sinopec Qingdao refinery makes use of predictive analytics to improve availability by preventing unplanned downtime with anomaly detection [49]. Given the significance of maintenance costs and the cost of unplanned outages, it is rational that the initial focus of refinery analytics is placed upon anomaly detection with high-cost rotating equipment assets, such as gas turbine generators and compressors. However, it should be noted that issues with the plant's material balance can also cause unplanned shutdowns and loss of production.

The second highest operational expenditure for a refinery is energy cost. However, results from a survey of refiners in 2018 revealed that energy management was only ranked as the 5$^{th}$ highest area with respect to benefiting from digital technologies [50]. One possible explanation for the low ranking of energy-related opportunities with digital technologies is that identification of use cases for refinery plant data analysis is overly influenced by the hierarchical nature of the data infrastructures, as described herein; energy-related analysis generally requires a non-hierarchical, cross-equipment, cross-unit perspective. Another potential explanation is that energy reduction and energy losses are often a by-product of equipment related analysis. This could be addressed by reporting the results of analysis wins in terms of monetary gain and reduction in lost energy opportunity. With the growing importance of energy efficiency, perhaps analysis use case selection should be weighted more in terms of lost energy opportunity.

### *6.2. Analytic Provenance*

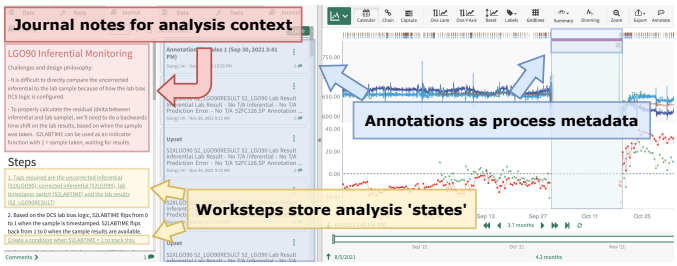


Figure 9: Analytic provenance and multi-user collaboration is facilitated by a *journal* or note-taking feature, allowing users to document their analytical reasoning and thought process within the same analysis window. Data-aware annotations can be marked using *annotation capsules* to highlight interesting and relevant observations in the time series data. *Application bookmarking* is achieved by recording analysis states as *worksteps* in the form of a clickable hyperlink.

The concept of *provenance* in visual analytics [41] involves recording the directly observable aspects of an analysis, includ-

ing the history of data flow, visualization states and user interactions, and also the non-directly observable aspects such as user thoughts, insights and analytical reasoning throughout the analysis lifecycle.

Provenance supports *recall*, which is the awareness and understanding of current and previous analysis states [41]. The investigative questions faced by plant engineers in their day-to-day tasks may be ambiguous and iterative in nature. In exploratory data analysis, it is common to venture into pathways that do not validate any hypotheses nor yield any meaningful results. As described elegantly by Pike and colleagues,

> "In many analysis tasks, goals are unstable, and a straightforward progression down a path of discovery is impossible." [34]

For example, the question *"Why did the reactor feed rate drop?"* is a very high level inquiry that begins by inspecting process data with a wide range of possible findings (e.g. instrumentation malfunctions, operational constraints, process upsets, control issues etc.) that would determine subsequent analysis steps and investigative avenues, as previously illustrated in Figure 2.

Since the engineer's analysis goals may change as more information is discovered during the investigation process, it is important that they keep detailed records of the investigation workflow. Encountering certain analysis dead ends, most engineers may intentionally discard the entire analysis and any intermediate outputs, deeming them as *throwaway* artifacts, because the end result was not meaningful [9]. Without proper documentation, engineers will struggle to keep track of past analyses, and their collaborators will also be confused, leading to wasted efforts and confusion over how certain results and conclusions were obtained. To build upon existing work, knowledge of past explorations — what has been attempted and justification for choosing certain approaches is important [51, 52, 53].

To record provenance of the engineer's insights, many self-service analytics platforms, including Seeq, provide a note-taking or *'Journal'* feature that allows users to input textual notes to document their analysis steps, integrating documentation, *data-aware annotations* [54] and comments directly into the main visualization window, as illustrated in Figure 9.

Provenance also supports collaborative communication and presentation of insights to other stakeholders [41]. Real-world engineering analysis is a social process that involves multiple stakeholders in discussion, interpretation and dissemination of results. Transferring findings from analysis to business actions requires successful communication between technical analysts and non-technical stakeholders [53]. In many organizations, these engineering findings are, unfortunately, often communicated using static screenshots sent through emails, memos or PowerPoint slides, and not directly integrated with the analysis platform.

To support these important social and collaborative interactions, visualization tools should empower users to record work steps and capture the internal state of an analysis as clickable hyperlinks. These links can be sent to collaborators so that they can directly observe the same analysis. This technique is also known as view sharing via *application bookmarking* [54]. Unlike static screenshots, application bookmarking provides more flexibility, allowing stakeholders to pick up an exploration where their collaborators left off, or even navigate to other views of interest that may not have been considered in the original analysis.

### *6.3. Automation and Analytics at Scale*

Instead of working with individual, decontextualized process tags, a refinery's data and assets can be better represented using hierarchical, contextualized models [55] known as *asset trees* in Seeq or *AF (Asset Framework) objects* in OSISoft PI. The integration of contextual metadata and structure in asset trees can help engineers understand the individual data points that comprise a specific asset [56]. Another key advantage of implementing a hierarchical, class-like knowledge representation is the support for polymorphism across assets, which facilitates analytics tasks and allows the engineer to reuse calculations and trends as templates across different assets. These concepts are related to *objects* in object-oriented programming (OOP) or *frames* in artificial intelligence knowledge representation [57],

As a more concrete example, we could organize the Burnaby Refinery's inferential models into an asset tree as shown in Figure 10. The parent process units can have multiple child inferential models, and the leaf node of each inferential contains the same three-tag structure – predictions, lab results and indicator function. The inferential monitoring calculations configured for one particular inferential in Case Study 2 can then be re-used and reapplied to all other inferential models in the refinery since they have a similar structure. The usage of hierarchical models and asset trees allows an engineer to conveniently scale up calculations from a single asset to plantwide analytics.

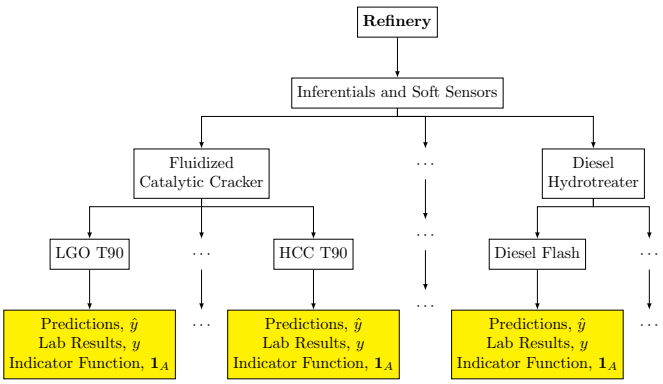


Figure 10: Plantwide asset tree of refinery inferentials with two exemplary process units, FCC (Fluidized Catalytic Cracker) and DHT (Diesel Hydrotreater Unit), as parent nodes. Each parent node can contain one or more child inferential nodes (e.g. LGO T90 or Light Gas Oil endpoint, HCC or Heavy Cat Cracked endpoint). The leaf nodes in each inferential all have the same structure, and contain the relevant process data tags associated with that particular inferential.

## 7. Limitations

Analytics tools are not one-size-fits-all. In general, there exists a trade-off between more advanced visualization capabilities and a shorter learning curve in self-service analytics tools. For example, certain platforms provide a *trellis* or *small multiples* [54] feature to quickly split a dataset into several subsets based on one or more user-defined categories, which can be very helpful for engineering investigations such as splitting histograms of temperatures by month to visualize slow process drifts over time. Likewise, more advanced visualizations like interactive heatmaps, which have applications in visualizing high-dimensional biological data [58] and diagnosing process control issues [59], can be difficult to configure natively in time-series oriented tools. Users must find a delicate balance between the tool's complexity and utility for their intended task.

An interesting study by Bessen [60] explored the link between technology usage and *industry concentration*, which is a measure of the extent of domination of firms in a particular market. Bessen found a strong relationship between the strategic usage of technology and higher organization revenues and productivity, even more so than mergers and acquisitions or other variables. The study also noted that productivity gains are not shared equally amongst all firms that adopt new technology. It's not just about purchasing the 'best' tools — top-performing firms actually utilize new technology productively, whereas their lesser competitors are not able to, and are therefore disadvantaged. Given the importance of data and digital skills for sustaining organizational competitiveness, what are the roadblocks in embracing modern, self-service analytics in an engineering organization? Our literature review uncovered several reasons.

- The reality is that spreadsheets are still ubiquitous in the process industries and appeal to the general engineering workforce [61]. In the literature, commercial BI tools are purported to be excellent for data exploration and visualization, but lack the flexibility for manipulating data and performing computations in a transparent manner [62], aspects which are important in engineering tasks and have traditionally been the strength of spreadsheets. Thus, the engineering community's reliance on spreadsheets poses significant friction in the adoption of modern analytics tools and platforms.

- Furthermore, despite the limitations of spreadsheets, an obvious spreadsheet-alternative for analytics tasks that would be acceptable for the general engineering workforce is notably absent [63]. Practicing engineers have varying computer literacy skills and they require tools that are appropriate for their tasks and abilities. Their diverse needs are unlikely to be fulfilled by a single software vendor [64], as casual users may find advanced tools too complicated to learn, and power users will be frustrated by the limitations of simpler tools [65]. Forcing both casual and power users on the same platform will likely lead to failure, as these tools are not one size fits all.

- Academic papers on process data analytics primarily focus on reporting novel algorithmic developments and applications of more advanced techniques like artificial intelligence (AI) and machine learning in the process industries [14], which typically falls under predictive and prescriptive categories. However, the average refinery engineer may not be aware of these new developments in advanced analytics, nor have the necessary technical skills and software platform to implement them effectively [63]. Furthermore, these research papers target an academic audience, and there is usually insufficient guidance for the practitioner on how these advanced techniques may be applicable to their day to day work, and how they can integrate them within existing processes and systems. Notably, this academic-industry gap has been discussed by process control researchers for decades [66, 67, 68]. In the adjacent field of machine learning, researchers have also pointed out community-wide scholarship issues, where many authors are hyper-focused on trying to one-up each other on contrived benchmark datasets, instead of measuring concrete impact on real-world problems [69, 70].

- Human factors challenges in embracing modern data analytics include a lack of perceived need or interest by organizational leadership [23, 64], as well as other psychological factors [21]. Senior leaders serious about analytics strategy cannot just pay lip service to digital transformation by purchasing the latest trends in technology, nor can they just participate passively by delegating work down to middle management [11]. To drive a culture of innovation for data analytics, organizations must develop a coherent digitalization vision and invest in equipping engineers with the ability to translate business opportunities into meaningful analytics projects [7]. As we can gather from Bessen's study [60], buying new technology will not help if organizational leadership does not know how to deploy and use it productively.

## 8. Future work and conclusions

Self-service analytics provides clear benefits for refiners, with time-in-motion savings and facilitation of reductions in operating expenditure and lost opportunity. There are also opportunities to improve quality and yields. Therefore, it is vital that refinery engineers are equipped with professional, purpose built, SSA tools in addition to self-service functions provided by existing spreadsheets, visualization and BI applications.

The following topics should be researched further as the use of self-service analytics grows throughout the refining industry and beyond,

1. The use of ungoverned spreadsheets with embedded code in the refining industry should be researched further. Indeed, ungoverned analysis in spreadsheets or any application, has the potential to lead to questionable outputs, thereby having the potential to impact operational decision making. One important subject to be considered is

the basic human tendency to trust a seemingly intelligent computer [71]. Programming code, whether embedded inside spreadsheets or developed in data science applications, is becoming a more common skill amongst engineers. However, Kletz et al. (1995) provided stark warning of the risks that software code poses,

> Software errors are systemic, they will always occur when the same conditions arise. As these conditions may arise infrequently, software errors can lie in wait like time bombs. [71]

Therefore, further research must be conducted to investigate the risk of ungoverned, bespoke, code and the extent to which the code has been tested. The European Spreadsheet Risks Interest Group (EuSpRIG) was founded in 1999 to systematically investigate the topic of spreadsheet integrity [72].

2. The process for understanding how data governance applies to the data feeds into the self-service systems has not been fully explored and users may not be competent in understanding the application of the tool to validate their business rules [73]. The subsequent cleansing of data also requires further research to ensure data cleansing operations such as filtering are fully understood, for instance, applying the wrong filter function to data from a particular type of plant sensor may unintentionally remove valid data instead of signal noise. Given the high-risk environment of hydrocarbon processing and the high potential for catastrophic incidents, the topic of data governance for SSA must be researched further.
3. The topic of data governance should also be extended to include the use of artificial intelligence and the extent to which AI will replace engineering judgement and the subject matter expertise required to perform refinery plant analysis. It is imperative that we understand if the use of AI could manifest as a form of ungoverned data analytics.
4. Different types of analytics are described from a timeliness perspective, however, most of these types of analysis can be grouped into categories such as 'routine' and '*ad hoc*'. Data access patterns and user access patterns for these two categories of analysis are different. Although routine analysis lends itself more to automation, self-service applications can be used to build the initial, reusable analysis that can then be scheduled. However, the extent to which the triggering and build of the analysis can be automated as part of a digital work flow requires further research.
5. Since the majority of a refineries operating expenditure costs is attributed to maintenance, it is unsurprising that predictive analytics has been widely adopted by refinery maintenance and reliability departments to help reduce unplanned outages. However, the second highest operating expenditure cost is energy and the use of analytics to help reduce energy consumption and energy losses must be researched further.

As cautioned by other industry practitioners [26], the successful deployment of AI and analytics in the refining and chemical industries is not dependent on just tools and algorithms. Human factors like training and culture are also critical for success, as well as a strong understanding and domain knowledge of refining. In this paper, we have highlighted several case studies illustrating challenges faced by refinery engineers and solutions using SSA tools. We have illustrated how relatively simple analytics tasks like data collection and cleaning can be a surprisingly time-consuming affair for engineers without effective workflows and tools. With this paper, we hope to encourage stronger collaboration between the refining industry, analytics providers and academia to work on tackling these real-world issues that will lead to a measurable and concrete impact on refinery operations. In particular, we urge further research at the intersection of chemical engineering, data visualization and human-computer interaction, as improvements in these areas can have a significant impact on engineering productivity and utility of self-service analytics tools.

## 9. Acknowledgements

The authors would like to gratefully acknowledge technical assistance from IT Vizion: Máté Haragovics and Sri Rahul Midde, as well as support from the Burnaby Refinery team and in particular, staff members in the Process Control department: Jin Li, Amy Chiu, Paul Herar, Hao Zhou, David Beaudoin and Eric Loong for the development of the analytics methodologies, workflows and case studies presented in this work.

## References

[1] C. H. Alhéritière, Cost-benefit analysis of refinery process data in the evaluation of plant performance, University of London, University College London (United Kingdom), 1999.

[2] O. Bascur, J. Kennedy, Measuring, managing and maximizing refinery performance, Hydrocarbon Processing 75 (1) (1996).

[3] Gartner, Inc., Definition of Self-Service Analytics (2021).
URL https://web.archive.org/web/20220206193348/https://www.gartner.com/en/information-technology/glossary/self-service-analytics

[4] J. Rowley, The wisdom hierarchy: representations of the dikw hierarchy, Journal of information science 33 (2) (2007) 163–180.

[5] D. Dong, T. McAvoy, Nonlinear principal component analysis—based on principal curves and neural networks, Computers & Chemical Engineering 20 (1) (1996) 65–78. doi:https://doi.org/10.1016/0098-1354(95)00003-K.
URL https://www.sciencedirect.com/science/article/pii/009813549500003K

[6] D. A. Beck, J. M. Carothers, V. R. Subramanian, J. Pfaendtner, Data science: Accelerating innovation and discovery in chemical engineering (2016).

[7] P. M. Piccione, Realistic interplays between data science and chemical engineering in the first quarter of the 21st century: Facts and a vision, Chemical Engineering Research and Design 147 (2019) 668–675.

[8] P. Mikalef, M. N. Giannakos, I. O. Pappas, J. Krogstie, The human side of big data: Understanding the skills of the data scientist in education and industry, in: 2018 IEEE global engineering education conference (EDUCON), IEEE, 2018, pp. 503–512.

[9] S. Kandel, A. Paepcke, J. M. Hellerstein, J. Heer, Enterprise data analysis and visualization: An interview study, IEEE Transactions on Visualization and Computer Graphics 18 (12) (2012) 2917–2926.

[10] K. Parker, How pipeline engineering gets done today, Oil & Gas Engineering 73 (6) (2019) 5–6.

[11] P. M. Piccione, Realistic interplays between data science and chemical engineering in the first quarter of the 21st century, part 2: Dos and don'ts, Chemical Engineering Research and Design 169 (2021) 308–318.

[12] D. Laney, L. Kart, Emerging role of the data scientist and the art of data science, Gartner Group. White paper (2012).

[13] M. Tory, T. Moller, Human factors in visualization research, IEEE transactions on visualization and computer graphics 10 (1) (2004) 72–84.

[14] V. Venkatasubramanian, The promise of artificial intelligence in chemical engineering: Is it here, finally, AIChE J 65 (2) (2019) 466–478.

[15] A. H. Maslow, A theory of human motivation., Psychological review 50 (4) (1943) 370.

[16] A. Parameswaran, Enabling data science for the majority, Proceedings of the VLDB Endowment 12 (12) (2019) 2309–2322.

[17] R. Jurney, Agile Data Science 2.0: Building Full-Stack Data Analytics Applications with Spark, O'Reilly Media, Incorporated.
URL https://books.google.ca/books?id=ZI_ujwEACAAJ

[18] M. Rogati, The AI Hierarchy of Needs (2017).
URL https://web.archive.org/web/20220210045737/https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

[19] J. Glassman, B. Shao, R. St Louis, Don't get the cart before the horse: There are no shortcuts to prescriptive analytics, in: Proc. of Hawaii International Conference on System Sciences, 2019.

[20] T. Nguyen, R. G. Gosine, P. Warrian, A systematic review of big data analytics for oil and gas industry 4.0, IEEE Access 8 (2020) 61183–61201.

[21] R. Roberts, R. Flin, D. Millar, L. Corradi, Psychological factors influencing technology adoption: A case study from the oil and gas industry, Technovation 102 (2021) 102219. doi:https://doi.org/10.1016/j.technovation.2020.102219.
URL https://www.sciencedirect.com/science/article/pii/S0166497220300912

[22] R. K. Perrons, J. W. Jensen, Data as an asset: What the oil and gas sector can learn from other industries about "big data", Energy Policy 81 (2015) 117–121.

[23] A. Marshall, S. Mueck, R. Shockley, How leading organizations use big data and analytics to innovate, Strategy & Leadership (2015).

[24] M. Aviles, Technology innovation and adoption in oil & gas industry–why did it slow? forbes. 14 july 2015 (2015).

[25] J. Abel, Shell Employs Analytics for Enterprise-wide Sustainability and Business Continuity (2020).
URL https://web.archive.org/web/20220219224115/https://www.arcweb.com/industry-best-practices/shell-employs-analytics-enterprise-wide-sustainability-business-continuity

[26] L. Colegrove, Artificial intelligence in the chemical industry—why my industry puzzles over our vendors' struggles, Journal of Advanced Manufacturing and Processing 2 (3) (2020) e10052.

[27] Y.-a. Kang, C. Gorg, J. Stasko, Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study, in: 2009 IEEE Symposium on Visual Analytics Science and Technology, IEEE, 2009, pp. 139–146.

[28] J. Stasko, C. Görg, Z. Liu, Jigsaw: supporting investigative analysis through interactive visualization, Information visualization 7 (2) (2008) 118–132.

[29] A. M. MacEachren, How maps work: representation, visualization, and design, Guilford Press, 2004.

[30] W. Aigner, S. Miksch, W. Müller, H. Schumann, C. Tominski, Visual methods for analyzing time-oriented data, IEEE transactions on visualization and computer graphics 14 (1) (2007) 47–60.

[31] B. Craft, P. Cairns, Beyond guidelines: what can we learn from the visual information seeking mantra?, in: Ninth International Conference on Information Visualisation (IV'05), IEEE, 2005, pp. 110–118.

[32] E. M. Shepherd, A multicase study of critical success factors of self-service business intelligence initiatives, Ph.D. thesis, Walden University (2021).

[33] C. Tominski, Event-based concepts for user-driven visualization, Information Visualization 10 (1) (2011) 65–81.

[34] W. A. Pike, J. Stasko, R. Chang, T. A. O'connell, The science of interaction, Information visualization 8 (4) (2009) 263–274.

[35] E. Dimara, C. Perin, What is interaction for data visualization?, IEEE transactions on visualization and computer graphics 26 (1) (2019) 119–129.

[36] J. S. Yi, Y. ah Kang, J. Stasko, J. A. Jacko, Toward a deeper understanding of the role of interaction in information visualization, IEEE transactions on visualization and computer graphics 13 (6) (2007) 1224–1231.

[37] B. Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, in: The craft of information visualization, Elsevier, 2003, pp. 364–371.

[38] J. Sansana, R. Rendall, Z. Wang, L. H. Chiang, M. S. Reis, Sensor fusion with irregular sampling and varying measurement delays, Industrial & Engineering Chemistry Research 59 (6) (2020) 2328–2340.

[39] B. Lu, L. Chiang, Semi-supervised online soft sensor maintenance experiences in the chemical industry, Journal of Process Control 67 (2018) 23–34.

[40] Z. Wang, L. Chiang, Monitoring chemical processes using judicious fusion of multi-rate sensor data, Sensors 19 (10) (2019) 2240.

[41] E. D. Ragan, A. Endert, J. Sanyal, J. Chen, Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes, IEEE transactions on visualization and computer graphics 22 (1) (2015) 31–40.

[42] M. I. Ahmad, N. Zhang, M. Jobson, Integrated design of diesel hydrotreating processes 89 (7) 1025–1036. doi:10.1016/j.cherd.2010.11.021.

[43] Parkland Corporation, Drive to Zero - Parkland Sustainability Report (2021).
URL https://web.archive.org/web/20220106192407/https://www.parkland.ca/en/sustainability/sustainability-report

[44] Z. Liu, J. Heer, The effects of interactive latency on exploratory visual analysis, IEEE Transactions on Visualization and Computer Graphics 20 (12) (2014) 2122–2131. doi:10.1109/TVCG.2014.2346452.

[45] J. Hedengren, Byu-prism seeq sysid toolbox, https://github.com/BYU-PRISM/Seeq (2021).

[46] P. van den Heuvel, R. Kroes, Shell's Advanced Analytics Journey in the Real-Time Data Domain (2020).
URL https://web.archive.org/web/20220220053330/https://www.arcweb.com/events/shells-advanced-analytics-journey-real-time-data-domain

[47] A. W. S. (AWS), Covestro Improves Chemical Manufacturing Process Running Seeq on AWS (2021).
URL https://web.archive.org/web/20220219224308/https://aws.amazon.com/partners/success/covestro-seeq/

[48] G. Simanjuntak, An integrated field operations to support hydrocarbon transportation case study at pt.cpi, in: 4th International Conference on Technology and Operations Management (ICTOM04), 2014.

[49] Siemens AG, Refinery improves efficiency using smart data (2021).
URL https://web.archive.org/web/20220106191850/https://new.siemens.com/global/en/company/stories/industry/refinery-improves-efficiency-using-smart-data.html

[50] T. Countryman, R. Holsman, A. Coward, E. Lemaitre, J. Adams, Accenture 2018 Digital Refining Survey: The Intelligent Refinery (2019).
URL https://www.accenture.com/_acnmedia/pdf-79/accenture-2018-refining-research.pdf

[51] M. B. Kery, M. Radensky, M. Arya, B. E. John, B. A. Myers, The story in the notebook: Exploratory data science using a literate programming tool, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–11.

[52] A. Rule, A. Tabard, J. D. Hollan, Exploration and explanation in computational notebooks, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–12.

[53] A. Y. Wang, A. Mittal, C. Brooks, S. Oney, How data scientists use computational notebooks for real-time collaboration, Proceedings of the ACM on Human-Computer Interaction 3 (CSCW) (2019) 1–30.

[54] J. Heer, B. Shneiderman, Interactive dynamics for visual analysis, Communications of the ACM 55 (4) (2012) 45–54.

[55] E. Kharlamov, F. Martin-Recuerda, B. Perry, D. Cameron, R. Fjellheim, A. Waaler, Towards semantically enhanced digital twins, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018, pp. 4189–4193.

[56] O. A. Bascur, J. O'Rourke, Measuring, managing, and transforming data

for operational insights, in: Smart Manufacturing, Elsevier, 2020, pp. 129–165.
[57] M. Minsky, A framework for representing knowledge. reprinted in the psychology of computer vision, p. winston (1975).
[58] N. F. Fernandez, G. W. Gundersen, A. Rahman, M. L. Grimes, K. Rikova, P. Hornbeck, A. Ma'ayan, Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data, Scientific data 4 (1) (2017) 1–12.
[59] S. Elnawawi, L. C. Siang, D. L. O'Connor, R. B. Gopaluni, Interactive visualization for diagnosis of industrial model predictive controllers with steady-state optimizers, Control Engineering Practice 121 (2022) 105056. doi:https://doi.org/10.1016/j.conengprac.2021.105056.
[60] J. Bessen, Industry concentration and information technology, The Journal of Law and Economics 63 (3) (2020) 531–555.
[61] C. Srivastava, Z. Yang, R. K. Jain, Understanding the adoption and usage of data analytics and simulation among building energy management professionals: A nationwide survey, Building and Environment 157 (2019) 139–164.
[62] D. Birch, D. Lyford-Smith, Y. Guo, The future of spreadsheets in the big data era (2018). arXiv:1801.10231.
[63] S. J. Qin, L. H. Chiang, Advances and opportunities in machine learning for process data analytics, Computers and Chemical Engineering 126 (2019) 465–473.
[64] T. W. Dinsmore, Self-service analytics, in: Disruptive Analytics, Springer, 2016, pp. 199–230.
[65] C. Lennerholt, J. van Laere, E. Söderström, Implementation challenges of Self Service Business Intelligence : A literature review, in: 51st Hawaii International Conference on System Sciences, Hilton Waikoloa Village, Hawaii, USA, January 3-6, 2018, Vol. 51, IEEE Computer Society, 2018, pp. 5055–5063.
[66] B. A. Ogunnaike, A contemporary industrial perspective on process control theory and practice, Annual Reviews in Control 20 (1996) 1–8.
[67] F. G. Shinskey, Process control: as taught vs as practiced, Industrial & Engineering Chemistry Research 41 (16) (2002) 3745–3750.
[68] B. W. Bequette, Process control practice and education: Past, present and future, Computers & Chemical Engineering 128 (2019) 538–556.
[69] Z. C. Lipton, J. Steinhardt, Troubling Trends in Machine Learning Scholarship: Some ML Papers Suffer from Flaws That Could Mislead the Public and Stymie Future Research, Queue 17 (1) (2019) 45–77. doi:10.1145/3317287.3328534.
[70] K. Wagstaff, Machine learning that matters (2012). arXiv:1206.4656.
[71] T. A. Kletz, P. W. Chung, C. Shen-Orr, Computer control and human error, Institution of Chemical Engineers, 1995.
[72] D. Chadwick, Stop that subversive spreadsheet!, in: Working Conference on Integrity and Internal Control in Information Systems, Springer, 2002, pp. 205–211.
[73] P. Clarke, G. Tyrrell, T. Nagle, Governing self service analytics, Journal of Decision systems 25 (2016) 145–159.