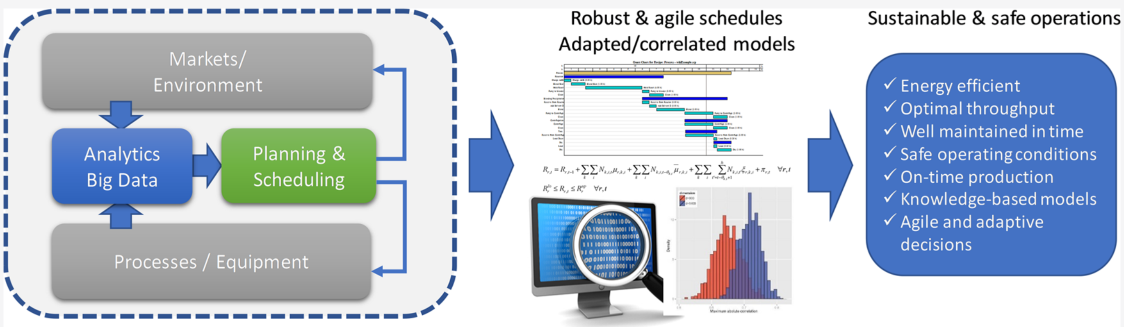
I&EC research
Industrial & Engineering Chemistry Research
pubs.acs.org/IECR

Article

Synergistic and Intelligent Process Optimization: First Results and Open Challenges

Iiro Harjunkoski,* Teemu Ikonen, Hossein Mostafaei, Tewodros Deneke, and Keijo Heljanko

Cite This: *Ind. Eng. Chem. Res.* 2020, 59, 16684−16694

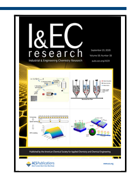


ABSTRACT: Data science has become an important research topic across scientific disciplines. In Process Systems Engineering, one attempt to create true value from process data is to use it proactively to improve the quality and accuracy of production planning as often a schedule based on statistical average data is outdated already when reaching the plant floor. Thus, due to the hierarchical planning structures, it is difficult to quickly adapt a schedule to changing conditions. This challenge has also been investigated in integration of scheduling and control studies (Touretzky et al. *AIChE J.* **2017**, *63* (66), 1959−1973). The project SINGPRO investigated the merging of big data platforms, machine learning, and data analytics with process planning and scheduling optimization. The goal was to create online, reactive, and anticipative tools for more sustainable and efficient operation. In this article, we discuss selected outcomes of the project and reflect the topic of combining optimization and data science in a broader scope.

1. INTRODUCTION

There is a lot of hype ongoing on big data analytics[2] and machine learning. Recently, many scientific meetings have been organized on the topic, e.g., FOPAM 2019 and several technical sessions in recent AIChE Annual Meetings, and for a good reason. Most companies collect continuously data from sensors that is stored for a certain time but never actually used, unless there is a need for post analytics as a part of troubleshooting.[3] The currently employed classical mathematical optimization models for scheduling[4] are typically based on fixed parameter sets, which are commonly maintained and updated offline by few domain experts and represent mainly statistical averages. Therefore, scheduling results are often criticized for being unrealistic, not dynamic enough and, thus, not reflecting the current production situation, partly due to which a significant body of research has investigated the integration of scheduling and control.[1,5−10] Nevertheless, such parameters could be estimated much more precisely in an online fashion using big data technologies, as already observed by many authors[11,12] and Santander et al.[13] using historical process data to learn statistical properties of process parameters. By creating collaboration interfaces between scheduling optimization, big data analytics, and machine learning, the process-related decision-making loop will become much more agile, self-aware, and flexible. Through this, also the negative effects of uncertainty and process variability can be expected to be reduced.[14] Rafiei and Ricardez-Sandoval[15] provide a thorough review on the integration of design and control, where data-driven methods are highlighted, e.g., to quantify uncertainty, provide surrogate models reducing the problem complexity, an extremely popular approach in the integrated optimization framework, and to narrow down the decision space through classification, e.g., in health-care applications. They see, as in the present paper, the potential of hybrid models combining knowledge-based and data-driven approaches in a systematic way, resulting in coordination between data and decision-supporting tools.

Received: April 23, 2020
Revised: July 19, 2020
Accepted: July 31, 2020
Published: July 31, 2020



https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

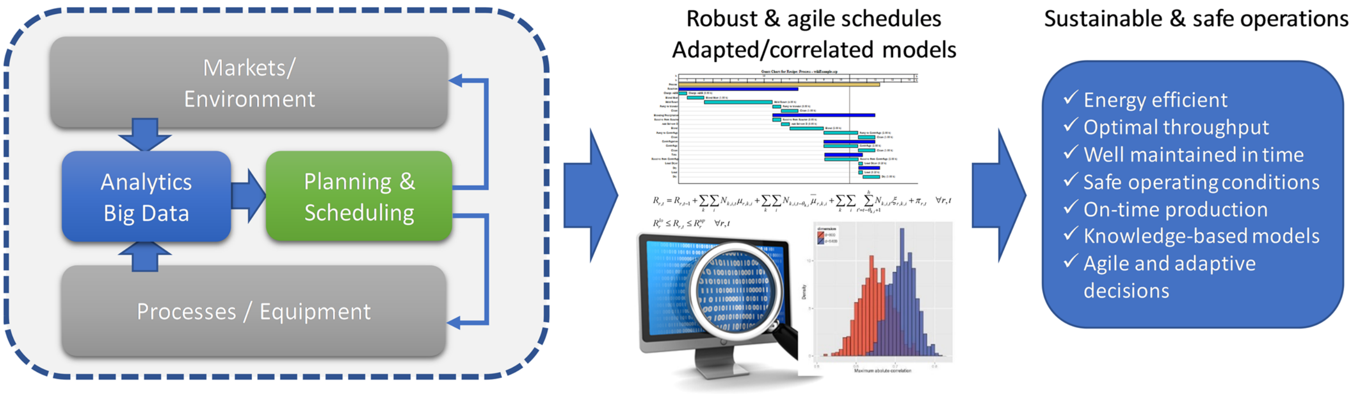


Figure 1. Use of data analytics and scheduling and its potential impact.

With the use of sophisticated data analytics methods, one can embed to the overall key performance indicators (KPI) also various information about the process, e.g., tracking abnormal situations (anomaly detection), individual process equipment performance degradations (predictive maintenance), anticipated process timings (prediction of future process behavior), and scenario simulation (e.g., artificial intelligence planning). Such an approach will help to select the best production strategies in order to maintain, e.g., production and energy efficiency as well as sustainability in rapidly changing market situations through data-driven self-adaptive scheduling models. Another research line focuses on the solvers, e.g., improving the Branch & Bound process.[16] However, here we will not discuss this research direction but rather focus on the modeling level.

The topic of data-driven models has already been investigated in other scientific domains, e.g., Operations Research and Data Science[17] and various tools that are already available for the process industry.[18] A good perspective on this topic is given by Venkatasubramanian.[19] It can also be expected that Industrial Internet of Things (IIoT) developments[20] provide the needed seamless connectivity, cloud computing infrastructure, and service-based business models to realize a closer connection between data and process and production optimization.

Many of the experiences discussed in this paper were gained in the project SINGPRO[i] and focused on the vision of seamless collaboration between data analytics and planning and scheduling, as shown in Figure 1. The larger goal was to investigate how the existing and available data could be utilized for supporting process automation and improving decision-making, ultimately leading to more context-aware, holistic, and optimized operations. As there are many existing methodologies in data analytics and production planning and scheduling, one must partly rely on the intuition to select a good starting point for the research. Our research was also guided by a few industrial applications at hand and discussions with companies to get access to industrial data for the project. In this paper, we further elaborate on the vision of using available industrial data proactively to improve the quality and actuality of planning instead of relying on static data that does not reflect current operation conditions. Through the merging of data, this can also be seen as a way of supporting the integration of planning, scheduling, and control. However, instead of defining workflows between the levels, the target has been to provide more accurate estimates in advance in order to reduce the mismatch between planning and closer-to-online operations. We studied several cases of merging big data platforms, machine learning, and data analytics methods with process planning and scheduling optimization and will here discuss some of the most promising results. We will show results that can improve the estimation of processing times, leading to more robust schedules, and examples where using historical operational data allows us to exclude some decisions, resulting in scheduling problems with a smaller decision space. Finally, process data can also be used in estimating equipment conditions leading to better approaches that combine operational and maintenance scheduling optimization. The overall results show that it is possible to create more agile, self-aware, and flexible decision-making tools. In the present paper, we will not dive into the technical novelties published elsewhere but rather focus on discussing a few essential research questions presented in Section 2. These questions were carefully selected as they are industrially very relevant but cannot be answered without fundamental academic research. In the discussion, we combine various results and thus take a step back from the technicalities. The main project results are presented in Section 3, without going into depth. As it is our aim in this paper to give somewhat of a bird's eye view on this extremely rich and exciting research area, Section 4 summarizes the research findings and concludes by discussing open questions and possible next research steps.

2. RESEARCH QUESTIONS

Naturally, conducting research in such broad domains as production decision-making and data analytics may quickly turn into a mission impossible as the number of potential research questions is immense. Therefore, it is important to select a clear focus, not trying to drill down into each of the domains. Inspired by the results of the SINGPRO project, we pose here six main research questions, which to our knowledge have not been explicitly discussed in this extent in earlier publications:

1. Could I do better planning by knowing more about the process, i.e., by utilizing real-time data? One of the main motivations is the abovementioned common case where a production plan is "old" soon after being rolled out to the plant floor. This leads often to manual improvements, which basically undermine the optimization efforts performed to produce the "perfect" schedule.
2. Is it better to dynamically generate accurate statistics on process behavior every time I want to schedule? This is very closely related to the previous question but focuses

https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

more on the data that is used as a basis for the entire scheduling activity. In most cases, schedules are based on, e.g., average durations estimated by process experts and collected in tabular form. This has two major deficits. On the one hand, the data and tables may be old and the process expert who created them may have left the company and no one feels sufficiently qualified to question them. One the other hand, the situation where the data tables have been generated might deviate from today's operating environment or correspond only to a subset of it.

3. How many incidents can actually be predicted and avoided? Disturbances and breakdowns often come as a surprise and the topic of detecting these in advance is already widely investigated in asset-management related research. Nevertheless, making this information directly available to the scheduling engine might result both in automatic creation of maintenance jobs and in selecting the best operating modes to avoid serious production failures.
4. What information is actually relevant for root-cause analysis? Are there hidden relationships? Often, we focus on the most obvious data observations assuming a very simple form of causality. Understanding the processes and their data better might help avoiding problems in the first place or at least increase the accuracy of predictions and improve the taken decisions.
5. Are there decisions that can be excluded from the optimization scope based on what we know from the data? Many decisions in optimization add to the complexity and it is very common in modeling to systematically go through all possible decision options in production and to create the corresponding decision variables for each of them, some of which are continuous, while others are discrete 0-1 decisions. Especially being able to rule out discrete decisions can very efficiently reduce the size of the branch and bound search tree in solvers, leading to faster solutions.
6. What is the actual value of this data? Data is mostly collected and stored into databases only for trouble-shooting. Once a serious process incident or accident has occurred, usually the logged data around the incident is analyzed in order to find a likely root cause for the abnormal situation. While excellent methods for this already exist, there are still very few methods available that could be used online for improving the scheduling decisions a priori. Even on the conceptual level, there are lots of possibilities for new innovations.

In the next section, we discuss each of these questions providing a brief overview of the main results and, when possible, refer to other publications for more information. We will not aim to be too comprehensive but hope that the results presented can act as a motivation for the research community to develop and try out further ideas in order to bring this area forward and closer to real industrial implementations.

3. MAIN RESULTS AND REFLECTIONS ON THE RESEARCH QUESTIONS

In this section, we will return to the research questions discussed above referring to the main research results achieved to date. As each of the research questions could alone easily fill an article, we will only highlight the main findings and qualitatively estimate how well each question could be answered and what would be needed to generate a deeper knowledge around each of the topics. By doing so, we believe to provide the best contribution to the process systems engineering community and bring forward some novel aspects and ideas.

3.1. Research Question 1. In order to answer this question, we first took an experimental approach and investigated a grade change scheduling problem within the paper industry. Having access to about 2 TB of operational data, the first step was to analyze if a grade change can be automatically detected and distinguished from any other process disturbance. In the scheduling of paper production, the grade change is an important aspect because it may cause production breaks and/or produce off-spec materials, both of which reduce the productivity. Therefore, it is essential to try to meet, e.g., weekly production targets with as few grade changes as necessary. Another aspect is that some of the grade changes may be very cumbersome, requiring extensive amounts of labor or resulting in large amounts of non-sellable products.

The available operational data of course reflects the current practice of operating the plant. Figure 2 shows a table where

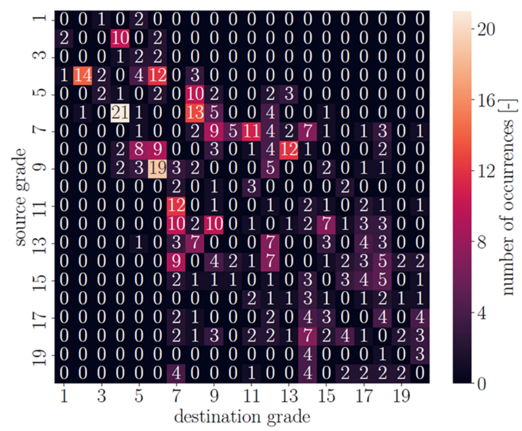


Figure 2. Past grade changes in a paper mill.

the grade changes between 20 different grades have been identified and counted applying data analytics on the historical process control data of the paper machine. The figure shows that only a small subset of possible grade changes were actually performed, some of which are very frequent. For instance, there are 21 grade changes from grade 6 to grade 4. This indicates that operationally the frequent grade changes are always preferred or the orders of the two grades have clearly dominated the production. Nevertheless, we assumed that the production is relatively stable over a year and applied a heuristic approach that uses this information to reduce the search space of an MILP-based model.

In addition to identifying the performed grade changes, the dataset can also be used to estimate the durations of the grade changes. As we use online data that accumulates continually, these durations can be defined based on the last production trends or relating them to other external factors, e.g., weather, season, or order patterns. Thus, it is possible to produce more accurate estimates than only the average or median duration.

Figure 3 shows the significant reduction of the number of grade changes when selecting from the superset only those that were identified in the dataset. As in most cases, the grade

https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

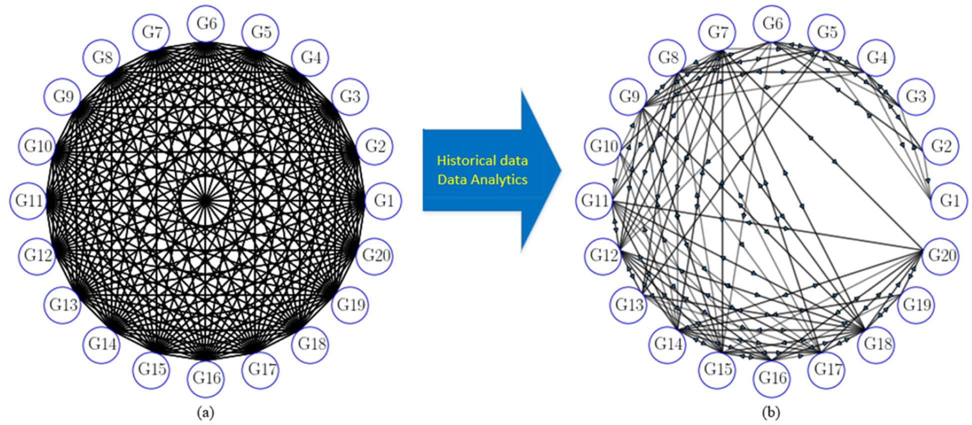


Figure 3. Superset with all grade changes vs subset based on the data analytics results.

changes must be modeled using binary variables, which take the value true if grade i is directly followed by grade i'. Since such decision variables can be treated as 0-1 continuous variables in our optimization model, the number of discrete variables remains the same for both the full-space and data-driven models in Table 1.

Table 1. Computational Results with the Objective of Minimizing the Grade Transition Cost

	full-space		data-driven	
# production runs	16	17	16	17
CPU(s)[a]	18,000	18,000	6215.5	9081.4
# discrete variables	672	714	672	714
# total variables	6550	6977	3700	3937
# constraints	8748	9316	5313	5652
objective ($)	58536.1	50613.8	53988.8	50338.8
relative gap (%)	76.08	72.14	0	0

[a]GAMS/CPLEX 12.7.1 (Intel i5-7300U, 2.60 GHz, 8 GB of RAM, Windows 10, 64 bit).

A comparison of the full-space model (Figure 3a) and the reduced model (Figure 3b) is shown in Table 1. The production of 2 weeks was considered with altogether 20 different paper grades. The solution procedure of the full-space model was terminated at a CPU limit of 18,000 s (i.e., 5 h) with an integer gap of more than 70%. The reduced model (data-driven model) could solve the problems to optimality within 2.5 h and reached a better solution than the full-space model despite the use of only a subset of possible grade changes. This example shows that reducing the search space, even when possibly compromising the theoretical global optimum, can in practice lead to better solutions. More details can be found in Mostafaei et al.[21] These results are also relevant to the research questions 2 and 5, which will be discussed later in this section.

Another theme that we investigated was whether the decision-making process itself could be improved using reinforcement learning (RL),[22] i.e., a branch of machine learning. In RL, an active agent interacts with a passive environment and seeks actions that maximize its reward. During the recent years, the field of RL has taken major leaps forward. A key ingredient in the development has been in the use of deep neural networks, enabling the definition of high-dimensional state and action spaces.[23] RL agents have been trained to make explicit scheduling decisions in process systems engineering[24] and in vehicle operation.[25,26] Recently, in the context of process control, Shin et al.[27] highlighted the development of hierarchical structures of RL (the higher level) and mathematical optimization (the lower level) as an important future research direction.

We proposed a hierarchical way of utilizing RL in process scheduling – the main idea is not to focus on the optimization model itself but on the meta-level of optimization. In particular, the following decisions are important: (i) when to trigger a new rescheduling procedure, (ii) how much computing resource to allocate, (iii) how far ahead to schedule (i.e., the length of the scheduling horizon), and (iv) whether to use a mathematical programming model or a heuristic algorithm.[28] In the proposed framework, we define the RL environment to consist of both the process and an optimizer, the latter of which makes the explicit scheduling decisions (Figure 4). The actions of the RL agent only affect the optimizer, and therefore, the agent interacts with the process only implicitly. The state space of the agent describes the changes in the process environment with respect to the time of the previous rescheduling action and the status of the optimizer.

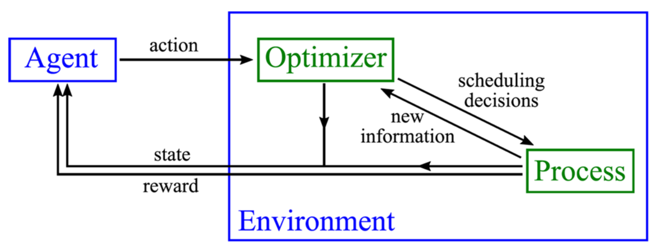


Figure 4. Using a reinforcement learning algorithm to improve the decision-making on rescheduling procedures.

We studied the first two (i.e., the rescheduling timing and computing time allocation) of the abovementioned four rescheduling decisions on simple routing problems, using Neuroevolution of Augmenting Topologies (NEAT)[29] as the RL algorithm. The reason why we chose NEAT is because the algorithm, by design, yields optimized neural networks with low topological complexity, which allows us to interpret structure and behavior of the trained RL agents. For a review of neuroevolution based RL, the reader is referred to Stanley et al.[30]

https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

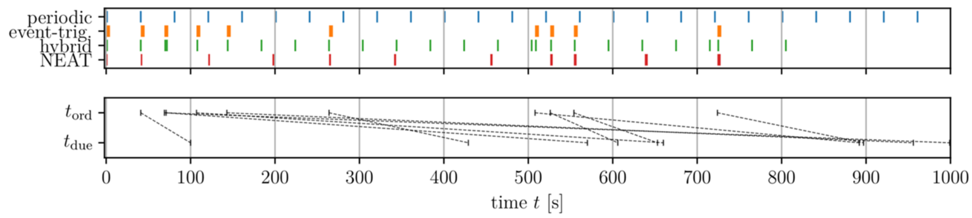


Figure 5. Rescheduling procedures by the NEAT algorithm and conventional rescheduling methods.

We studied the approach on four test cases with varying rates of incoming new orders and compared the results to those obtained by conventional rescheduling methods (i.e., periodic and event-triggered rescheduling as well as a hybrid method combining the two). The NEAT algorithm yielded, on average, better closed-loop schedules than the conventional rescheduling methods in three out of four test cases. The upper pane of Figure 5 shows a comparison of the rescheduling timings by the NEAT agent and the conventional rescheduling methods. In the lower pane, one can see the order dates, t_{ord} (time when an order becomes visible to the optimizer), and the due dates of the orders, t_{due}. The steeper the line, the more urgent is the order. It can be seen that the NEAT algorithm mainly triggers a scheduling when a new urgent order appears.

The approach can be trained for real processing environments, where the response time is often critical. If we, e.g., know that in a certain case the scheduling activity should take place within 120 s, the approach can automatically select to use an algorithm that is likely to give a good answer within the given computing time budget. More results can be seen in Ikonen et al.[31]

Thus, based on two simple test cases, the answer to the first research question is a clear yes. We can significantly improve the efficiency and/or the quality of the planning by utilizing the existing real-time data as well as advanced methodologies to support a more agile decision-making. Future work should investigate the approaches on more complex and elaborated test cases, in which possibly even higher benefits can be obtained.

3.2. Research Question 2.

As earlier stated, often process timings are based on some static tables that are seldom or never refined. The use of data analytics on recent process data can yield more accurate predictions of, e.g., process timings based on the current environment and processing conditions. This could be based on weather forecasts, equipment conditions, upcoming maintenance breaks, or composition of a working shift.

To test the hypothesis, we took an openly available dataset of taxi trips in New York City (NYC)[ii] and divided a subset of the trips (around 40k trips) into a training set and a test set. The subset includes the trips that start or end within 500 m from Wall Street (Figure 6). The purpose is to predict the trip duration based on the pick-up and drop-off coordinates and use the predictions in a devised scheduling problem. This was a good option as real process scheduling data is very difficult to obtain, and the duration of a taxi trip from place A to B can be expected to reflect similar uncertainty as in a processing task. The sources of uncertainty are, for example, traffic conditions, varying skill levels of the drivers, and their driving style. This stochastic behavior can also be seen in a process where operators work differently, and a processing step can be delayed by lacking upstream input or downstream capacity.

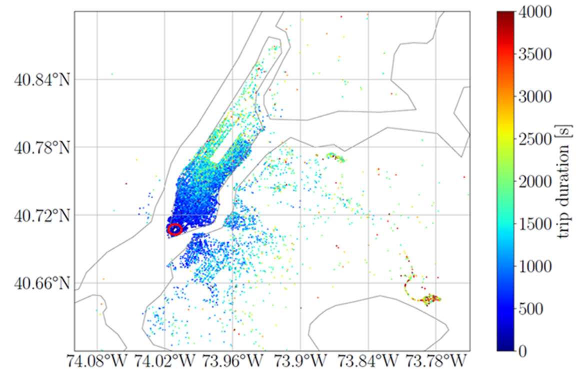


Figure 6. Durations of taxi trips in New York City, starting from or ending at Wall Street.

The scheduling problem is to minimize the makespan of performing six surveys in remote locations in NYC, subject to limited number of vehicles and surveying teams. Each survey includes three tasks: (i) the outbound trip, (ii) the actual survey, and (iii) the inbound trip (Figure 7). The durations of the trips are predicted, whereas the actual surveys have a deterministic duration. In chemical processing, this can correspond to a three-stage batch process, where stages one and three are subject to uncertainties. These could be, e.g., reaction processes or processing steps, the duration and/or the progress of which must be determined based on samples. To reflect the reality, i.e., there are also batch processes with nearly no variation in their durations, stage two was assumed to be well predictable and a stable process, e.g., an automated packaging stage.

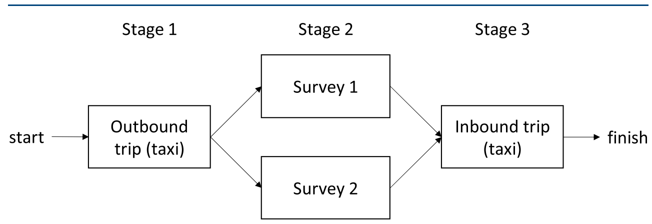


Figure 7. Three-stage batch process derived from the taxi-trip data.

The six remote locations were randomly chosen in such a way that a representative trip can be found in the test set. The scheduling optimization was performed by a slightly modified version of the model by Shaik and Floudas.[32] The trip durations were predicted by three prediction models of different fidelity that were trained on the training data. After the scheduling, these durations were updated by those in the test set in order to obtain the realized schedule.

We discovered that a very common approach of using the average value of the past durations was not very accurate but rather using more advanced methods could significantly

https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

improve the accuracy of the input data. Table 2 compares three instances of deriving the data used for scheduling. The average

Table 2. Comparison of Three Approaches to Estimate the "Processing" Times

prediction model	RMSE	r^2
average	701.8	-2.37×10^{-4}
rate	556.5	0.371
GP	438.4	0.610

prediction model, having the lowest fidelity, calculates the average duration of all trips that start from or end at Wall Street. The rate prediction model first determines the average speed of all trips in the training data and then predicts the duration of a trip based on the average speed and the geographical distance between its start and end points. The GP prediction model, having the highest fidelity, uses Gaussian Process regression[33] with two features, which are the latitude and longitude of the remote location. We chose Gaussian process regression, instead of other supervised learning methods (e.g., neural networks or random forests) because the method yields also an estimate of the prediction uncertainty. This is helpful when assessing the results and could also facilitate the use of stochastic optimization approaches in the future. As earlier mentioned, the data was divided into two sets; we train the models on the training set and then test the trained models on the test set. Table 2 shows the prediction accuracy on the test set based on the root mean square error (RMSE) and the coefficient of determination (r^2). Both measures indicate that the prediction accuracy improves when increasing the fidelity of the prediction model.

Next, we used these predicted durations as parameters in the scheduling model where the objective was to minimize the makespan. Figure 8 shows the results on 30 optimization problems with different remote locations. The optimized makespan is normalized with respect to an ideal prediction model, having an RMSE of 0 and r^2 of 1.0. It was derived by "cheating", i.e. using the actual trip durations in the test set. Based on the results, improving the accuracy of the parameters also yields, on average, better solutions. The reader is referred to Ikonen and Harjunkoski[34] for further details on the scheduling problem and the prediction models.

To summarize, the use of more data in a dynamic manner and applying sophisticated methodologies to predict a process duration can result in significantly better parameter accuracy, which further has an impact on the quality of the scheduling solutions. One task that we did not investigate is how large an improvement could be gained by further considering the trip data per hour of a day and using it as an additional feature in the prediction model. The scheduling model for this would be more complex but this could result in further improvements in the parameter accuracy and the quality of the scheduling solutions.

3.3. Research Question 3. This question is by itself too vast to be answered holistically. In the SINGPRO project, we focused on a paper machine and its very large 2 TB data set collected for 1 year from a process control history database. The target was to try to identify signals that would early enough indicate a paper break and thus enable to take measures to prohibit such incidents. Figure 9 shows a typical

Figure 9. Machine learning steps for creating a prediction model for paper breaks.

process how to derive a prediction model from raw data. The main idea is to use the existing data set in an efficient way to build a model of a phenomenon (here, the paper break) such that the model can also be used to well interpret unseen data. This is very challenging as the paper manufacturing process is quite mature and well analyzed using conventional techniques. In addition, there are normal deviations in the process and as the same machine is producing many different products, certain products may have a different behavior resulting in product-specific measurement signal data.

The first step was to identify from the data when a paper break actually took place, and this is part of the data preparation process. By itself, this might not seem not very valuable since once the data indicates a paper break it has already happened, and it could not find the root cause of this event. Nevertheless, this was a necessary first step in starting to look at the prediction of the break as it could differentiate between normal deviations and values that in fact led to a problem situation. The data preparation process also included decompressing the row data, resampling it at a common sampling rate and segmenting it to break trajectories, which are the bases for a break prediction modeling and are often done using big data frameworks such as Spark.[35,36] The next step was to select the important set of variables from the dataset that may have a causal relationship and thus can be relevant for predicting the break. The selected variables can further be consolidated and denoised through dimensionality reduction techniques at this stage. Finally, in the training phase and once the data could be matched with an actual break, it was possible to try out various machine learning technologies to identify possible warning signals that are preceding the break.

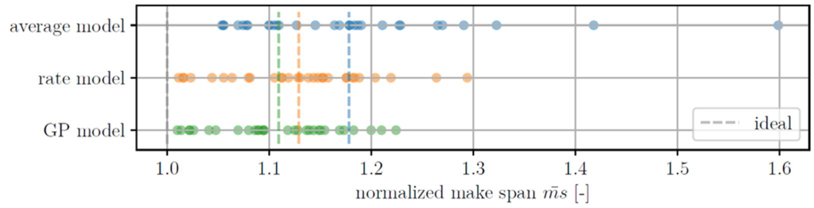


Figure 8. Normalized makespans based on the three prediction models. The green, yellow, and blue dashed lines indicate the average values.

https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

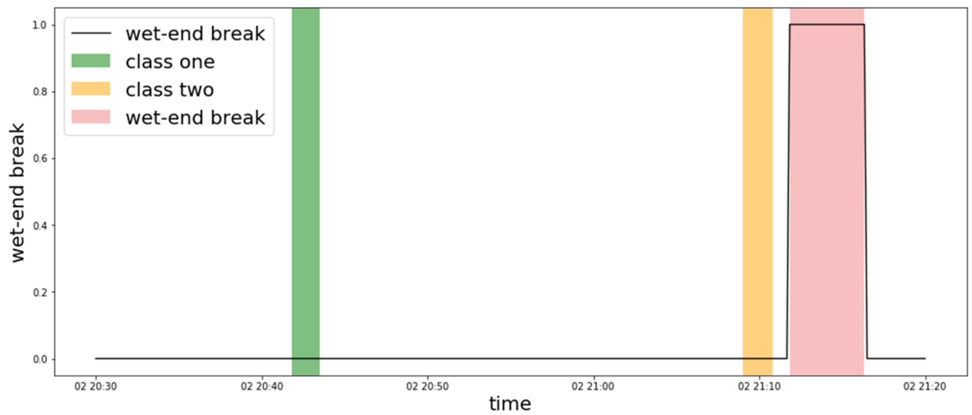


Figure 10. Classification of the investigated data points.

Among others, prediction models that classify data samples as near or far from a possible break can be obtained as well as models predicting the remaining time until the next break can be generated. We experimented with the random forest and long short-term memory (LSTM) learning algorithms to fit our prediction model. LSTMs are types of recurrent neural nets that have memory and feedback that allows them to inherently capture temporal trends making them suitable for learning sequence to sequence and temporal prediction problems. Random forests, on the other hand, construct multiple decision trees (often in parallel) based on the training data and combine their results for producing the final prediction and a variable importance score.

Figure 10 illustrates the identified classification problem: We created a model that can predict whether a data point belongs to a class one, which indicates that the sample is further away from a possible wet-end break, or class two, which implies that the sample is near a possible wet-end break. For training the model, we first extracted break trajectories that have at least an hour of normal operation followed by at least a 5 min break. Doing this is crucial as many breaks appear in clusters, and we are only interested in prediction of the first break in the sequence of breaks together with its root causes. We further filtered out break trajectories that belong to a controlled intended clean up breaks, a class of breaks that was not labeled in the dataset but was identified from the data analytics results when different break categories were automatically identified. Features that showed a high correlation with the target break signal were also filtered out. We then constructed our training set by extracting two classes of data samples consisting near break samples (just over 0.5 min away from the break) and far break samples (more than 15 min away from the break). To balance the number of samples from each class, we only maintained a 3 min window worth of samples from each. A PCA was then applied for dimensionality reduction before fitting prediction models using random forest and LSTM. Since random forest did not have an inherent mechanism to model the temporal aspect of the break trajectories, we added lagged versions for each feature in the training set, allowing also random forest methods to be used for the prediction task. The resulting models have a classification accuracy of 64.07% for the LSTM and 58.39% for the random forest.

In summary, it is possible to some extent to automatically predict the paper breaks, which can provide enough time for precautionary actions. The main contribution here is the approach to modeling the problem and cleaning up the dataset in order to predict and find root causes of the first wet end break in a long production run. The results have shown some possibilities of identifying whether a data sample belongs to normal operation or not. More than that, the prediction models have paved the way for identifying signals that are possibly the root causes of breaks, which will be presented in detail in the next section. Further improvement can be done to improve the accuracy of the models through improved data pre-processing such as identifying the various break trajectory categories and modeling them separately. We also believe that an improved feature selection mechanism, either automated or based on domain expert recommendation, is crucial for improving the accuracy.

3.4. Research Question 4. Very much related to the previous research questions is the relevancy of information that is determined by automated means. Sometimes the most obvious data relations are too trivial and not knowing the physical background of a signal, a large-effort data analysis may for instance result in a causality between the control variable (e.g., a set point) and the output variable, something that is evident and could have initially been excluded from the entire scope as the main purpose of a working control system is to keep these equal. Thus, the main challenges are to identify the true possible root-cause signals, identify delays between these root-cause signals and the break, and identify and remove highly correlated signals (with no delay) to the break. Lucke et al.[37] and other similar previous works focus mostly on detecting the occurrence of faults and identifying their categories. However, it is important is to understand a possible root cause for an identified fault. These challenges particularly become important in the absence of a domain expert prefiltering variables to be used as features for machine learning. In this work, such variable selection was achieved through correlation analysis and permutation-based feature importance scoring. Correlation analysis is often used to statistically evaluate the strength of the relationship between two variables. A high correlation at a given lag indicates that two or more variables are strongly related to each other at that lag, while a weak correlation indicates hardly any relation.

The approach to finding the signals that are potential root causes was developed and differs from earlier methods presented in the literature: Besides using correlation analysis to filter out the signals that have a very high correlation with the break signal at lag zero (most such signals are basically caused by the break), we used such analysis to calculate the correlation of all signals to wet-end break at various

https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

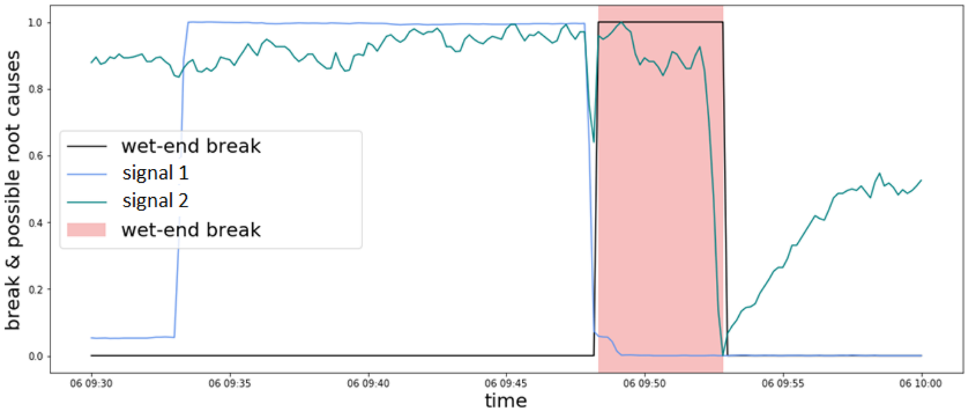


Figure 11. Machine learning steps for creating a prediction model for paper breaks.

consecutive lags. We then calculated a cumulative sum over the differences in correlation results of each signal against the break at consecutive lags. Based on such correlation analysis, we selected signals that show a growing correlation trend with the target at the latter lags than near-zero lags. The selected signals have therefore a higher possibility of a causal relationship with the target. We then fit predictive models based on the selected signals using LSTM and random forest. Further root-cause analysis was done through the use of a variable importance measure obtained from the fitted prediction models. The importance of a variable to the overall prediction is calculated relative to its contribution toward the overall prediction of the model. In this work, we have employed a new permutation-based variable importance measure to calculate the importance of a feature to the prediction and the model. More specifically, the permutation-based feature importance is calculated as the decrease in a model score when the samples related to the feature are randomly shuffled along the time axis. Such a procedure certainly breaks the relationship between the feature and the target, leading to an expected drop in the model score. The relative amount of this drop in the prediction score is indicative of how much the model depends on the feature and can be used as a generic feature importance measure in a fully machine learning algorithm agnostic manner.

Figure 11 shows two signals that were identified as having a possible causal relationship through our correlation analysis and a relatively higher feature importance score. In the figure, a particular signal (signal 1) is shown to have spiked starting around 12 min earlier than a wet-end break.

One lesson learned from this task was that nothing can replace an experienced domain expert. Blindly looking for relationships without having any process knowledge is very resource-consuming and can result in none or very poor assumptions and machine learning models that are not usable in the operational environment. Therefore, it is important that domain experts also get to know the main machine learning principles and can actively support the machine learning process as part of an iterative process to refine the prediction models and identify potential root causes. A follow-up research question is how the domain knowledge could be automatically embedded into machine learning and whether this is possible with minimal human intervention.

3.5. Research Question 5.

In the literature, process data has been used to quantify the uncertainty from data for stochastic optimization problems. Calfa et al.[38] use kernel smoothing to determine the constraints from historical data for change-constrained optimization.[39] Shang et al.[40] and Bertsimas et al.[41] propose data-driven methods to quantify the uncertainty sets from data for robust optimization,[42] in order to reduce the conservatism of the optimization approach. However, depending on the process, a historical data set may also contain information that helps to explicitly exclude, or preassign, some of the decisions in the optimization problem.

We investigated this research question pragmatically by studying two optimization problems identified in the industry. The first was the grade change scheduling on a paper machine, introduced and discussed already in Section 3.1. In this study, we used historical data to identify those grade transitions that were performed during the last year of operation. Relying on the past operational practice, this allowed the exclusion of most grade change options, which, depending on the mathematical formulation, can significantly reduce the number of variables (continuous or binary) and constraints, leading to faster solution times. However, while this approach significantly reduces the computing time, the global optimum may be excluded from the decision space if the optimal solution involves a grade change that has not been performed during the recording of the historical dataset.

The second optimization problem, identified by another industrial company, was to decide which maintenance actions (repair or replacement) should be performed for the components of a large-scale engineering system during a planned maintenance shutdown. The components are arranged in a serial-parallel system (see Figure 12 as an example); the system is functioning if at least one component at each serial stage is functioning. In the literature, this problem is referred to as selective maintenance optimization and was first studied by Rice et al.[43] Recently, elaborate selective maintenance optimization models have been published in the literature,

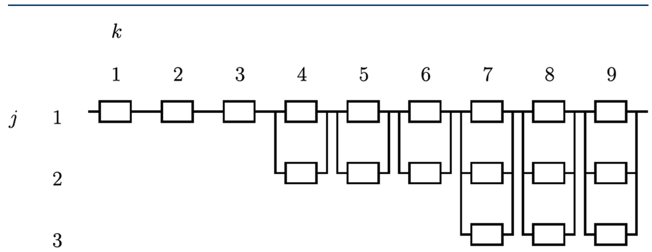


Figure 12. Example of a component arrangement in a serial-parallel system.

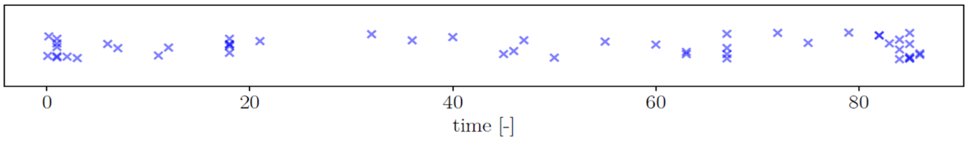


Figure 13. Failure times of a given component type.[49]

including, e.g., joint selective maintenance optimization and repair personnel assignment,[44] systems that have a serial n-out-of-k reliability structure,[45] and the consideration of break and/or mission durations with uncertainty.[46,47] However, in selective maintenance optimization, the component lifetimes are typically modeled to have either an exponential or a Weibull distribution, and the linking of failure data to the optimization models has not been discussed.[48] The novelty of our study was especially in considering bathtub-shaped failure rates, caused by the infant mortality and degradation of the components, in a planning model, and in linking the failure rates to component lifetime data.

In the optimization problem described by the industrial company, the components were electronic devices, the online monitoring of which is extremely difficult. The components typically work until they break without a warning. Thus, we used two relevant failure time datasets from the literature,[49,50] the failure rate of which is bathtub-shaped (see Figure 13 for an example of such data). From the figure, it can be seen that, e.g., during the time period of 20−40 time units, only a very few failures occur. This time period is after the infant mortality period of the component but before any significant degradation.

We fitted failure models[51] into the lifetime datasets by maximizing the log likelihood of a failure model by SLSQP (sequential least squares programming).[52] We then derived the change in reliability, if the component is replaced, as a function of the age and the length of the next operation window (Figure 14). The figure has a region (indicated by white) where the change in the reliability is negative. Replacing such a component would be non-sensible as it would reduce the overall reliability of the system. Thus, we defined a pre-assignment, which precludes the replacement of such components from the maintenance actions. The pre-assignment reduces the combinatorial complexity of the problem

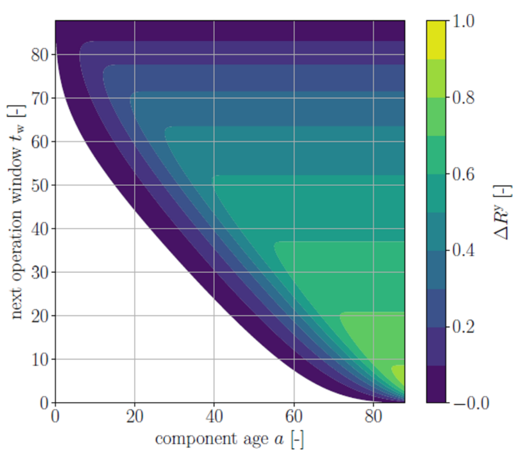


Figure 14. Change in the component-specific reliability if the component is replaced. The contour plot is determined based on a bathtub-shaped failure rate that is fitted to lifetime data.

and, in our experiments, reduced the solution time of the optimization problem by roughly an order of magnitude. In contrary to the grade change scheduling case, this pre-assignment does not have a risk of removing the global optimum from the decision space.

We also improved the efficiency of solving the problems by extending the convexification, originally proposed by Ye et al.[53] For more details on the data analytics method and the convexification, the reader is referred to Ikonen et al.[54]

The finding of preassigning non-sensible component replacements based on data showed that the optimization effort for even slightly more complex cases can be efficiently reduced by deploying data analytics methods. Similar to the research question in Section 3.1, where we discussed the scheduling of grade changes in a paper machine, here we can efficiently use data to reduce the size of the search space. This indicates that using machine learning or advanced data analytics as a pre-processing step for optimization can have a significant positive effect on the performance while the negative impact on the solution quality is negligible or small.

3.6. Research Question 6. As discussed earlier, process data is mostly collected and stored in databases for a limited time (often at least 1 year), depending on the amount of data and available storage capacity. Using the data online and in a predictive/preemptive manner has lots of potential for improving operations of many industrial sites and systems. In our research, we could clearly identify many qualitative improvements with also a strong financial potential. However, without having access to the business figures nor the possibility to make well-organized and comprehensive "before/after" comparisons, it is not possible to quantify the actual benefits or value of the data used[iii].

Nevertheless, many industries are looking into this question[55] as they want to understand how much they should invest in data science, machine learning, and other related activities. Also, data analytics companies try to promote these to their best in order to improve their marketing position, which has already resulted in a big hype. From the engineering point of view, one can try to achieve maximum stability, safety, and profitability of a given process, but the value of the data (added profit) must be done in collaboration across various disciplines and an equally relevant and complex research question is how to do this in a best way. It is evident that the various approaches discussed in this paper contribute to the value creation of the data by using it more actively.

4. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

Many of the results discussed in this paper were generated during the SINGPRO project, which was only a 2-year activity but provided us a very good starting point to dive into the world of combining data analytics with optimization. We could see clear and tangible benefits from using more advanced methods to process historical/online data through the application of AI/ML methods. We showed the ability to

https://dx.doi.org/10.1021/acs.iecr.0c02032
Ind. Eng. Chem. Res. 2020, 59, 16684−16694

significantly improve the accuracy of scheduling, enhance the predictability of processes, and largely reduce the search domain of large-scale problems, which all can lead to better and more optimized operations resulting in higher profits. However, the data-related work is still very problem-specific and a generic "cookbook" is missing to reduce the efforts. Also, domain knowledge turned out to be often invaluable in these types of exercises.

We were also able to identify many new research questions, e.g., in how to even further improve the data analytics and machine learning results by better clustering (e.g., the time-dependency in the taxi-trip data), how to efficiently distinguish and filter out unnecessary data that shows correlation but not direct causality, and how to try to automate the domain knowledge in handling various process-related data.

The question that we could not really give a satisfactory answer to was related to the value of the data. It is, nevertheless, clear that the process data has value, which is potentially even very significant, but defining the value quantitatively is something that the industry should investigate with its access to the exact business figures. Nevertheless, it is clear that there are many opportunities for advanced data analytics to support scheduling, e.g., by enabling model simplification, partial model building, and reducing the search space. For instance, some traditional mathematical programming models for scheduling, e.g., assume generic sequencing variables that in principle allow any sequence to occur (over-definition), which is then limited by inefficient big-M constraints. A more efficient way is to only define those sequences that are relevant and thus focus on the core sequencing decisions. The key question is how to ensure that similar opportunities are identified and utilized in a correct fashion? One way for engineers to support this is by involving the process systems engineering community in a seamless and true collaboration with other disciplines (e.g., computer science, mathematics, and statistics). Based on our learnings from the multi-disciplinary SINGPRO research project, we see this as a prerequisite for success. The journey of deploying synergistic methods that efficiently combine data analytics with optimization has only begun, and we can surely expect very exiting novel research contributions in the near future.

■ AUTHOR INFORMATION

Corresponding Author

Iiro Harjunkoski − *Aalto University, Department of Chemical and Metallurgical Engineering, 00076 Aalto, Finland; Hitachi ABB Power Grids Research, 68309 Mannheim, Germany;* orcid.org/0000-0003-0428-0751; Phone: +49 171 3322306; Email: iiro.harjunkoski@aalto.fi, iiro.harjunkoski@hitachi-powergrids.com

Authors

Teemu Ikonen − *Aalto University, Department of Chemical and Metallurgical Engineering, 00076 Aalto, Finland*

Hossein Mostafaei − *Aalto University, Department of Chemical and Metallurgical Engineering, 00076 Aalto, Finland;* orcid.org/0000-0002-4761-3037

Tewodros Deneke − *University of Helsinki, Department of Computer Science, 00014 Helsinki, Finland*

Keijo Heljanko − *University of Helsinki, Department of Computer Science, 00014 Helsinki, Finland; Helsinki Institute for Information Technology (HIIT), 00014 Helsinki, Finland*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.0c02032

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Financial support is gratefully acknowledged from the Academy of Finland project "SINGPRO", Decision No. 313466 and No. 313469. We also thank all the industrial and academic collaborators for their valuable contributions.

■ ADDITIONAL NOTES

[i]https://singpro.github.io/
[ii]https://www.kaggle.com/c/nyc-taxi-trip-duration
[iii]https://www.techrepublic.com/article/how-to-find-the-real-value-in-operational-big-data/

■ REFERENCES

(1) Touretzky, C. R.; Harjunkoski, I.; Baldea, M. Dynamic models and fault diagnosis-based triggers for closed-loop scheduling. *AIChE J.* **2017**, *63*, 1959−1973.

(2) Qin, S. J. Process data analytics in the era of big data. *AIChE J.* **2014**, *60*, 3092−3100.

(3) Shu, Y.; Ming, L.; Cheng, F.; Zhang, Z.; Zhao, J. Abnormal situation management: Challenges and opportunities in the big data era. *Comput. Chem. Eng.* **2016**, *91*, 104−113.

(4) Harjunkoski, I.; Maravelias, C. T.; Bongers, P.; Castro, P. M.; Engell, S.; Grossmann, I. E.; et al. Scope for industrial applications of production scheduling models and solution methods. *Comput. Chem. Eng.* **2014**, *62*, 161−193.

(5) Dias, L. S.; Ierapetritou, M. G. 110th anniversary: Integration of scheduling and robust model predictive control. *Ind. Eng. Chem. Res.* **2020**, *59*, 265−280.

(6) Charitopoulos, V. M.; Papageorgiou, L. G.; Dua, V. Closed-loop integration of planning, scheduling and multi-parametric nonlinear control. *Comput. Chem. Eng.* **2019**, *122*, 172−192.

(7) Biegler, L. T. Advanced optimization strategies for integrated dynamic process operations. *Comput. Chem. Eng.* **2018**, *114*, 3−13.

(8) Pistikopoulos, E. N.; Diangelakis, N. A. Towards the integration of process design, control and scheduling: Are we getting closer? *Comput. Chem. Eng.* **2016**, *91*, 85−92.

(9) Nie, Y.; Biegler, L. T.; Villa, C. M.; Wassick, J. M. Discrete time formulation for the integration of scheduling and dynamic optimization. *Ind. Eng. Chem. Res.* **2015**, *54*, 4303−4315.

(10) Park, J.; Du, J.; Harjunkoski, I.; Baldea, M. Integration of scheduling and control using internal coupling models. *Comput.-Aided Chem. Eng.* **2014**, *33*, 529−534.

(11) Dias, L. S.; Ierapetritou, M. G. Integration of planning, scheduling and control problems using data-driven feasibility analysis and surrogate models. *Comput. Chem. Eng.* **2020**, *134*, 106714.

(12) Valdez-Navarro, Y. I.; Ricardez-Sandoval, L. A. A novel back-off algorithm for integration of scheduling and control of batch processes under uncertainty. *Ind. Eng. Chem. Res.* **2019**, *58*, 22064−22083.

(13) Santander, O.; Betts, C. L.; Archer, E. E.; Baldea, M. On the interaction and integration of production planning and (advanced) process control. *Comput. Chem. Eng.* **2020**, 106627.

(14) Gupta, D.; Maravelias, C. T. Framework for studying online production scheduling under endogenous uncertainty. *Comput. Chem. Eng.* **2020**, 106670.

(15) Rafiei, M.; Ricardez-Sandoval, L. A. New frontiers, challenges, and opportunities in integration of design and control for enterprise-wide sustainability. *Comput. Chem. Eng.* **2020**, 106610.

(16) Zarpellon, G.; Jo, J.; Lodi, A.; Bengio, Y. Parameterizing Branch-and-Bound Search Trees to Learn Branching Policies. *arXiv* **2020**, *051*, 2020.

(17) Van der Aalst, W.; Weijters, T.; Maruster, L. Workflow mining: Discovering process models from event logs. *IEEE Trans. Knowl. Data Eng.* **2004**, *16*, 1128−1142.
(18) Wilson, Z. T.; Sahinidis, N. V. The ALAMO approach to machine learning. *Comput. Chem. Eng* **2017**, *106*, 785−795.
(19) Venkatasubramanian, V. The promise of artificial intelligence in chemical engineering: Is it here, finally? *AIChE J.* **2019**, *65*, 466−478.
(20) Isaksson, A. J.; Harjunkoski, I.; Sand, G. The impact of digitalization on the future of control and operations. *Comput. Chem. Eng.* **2018**, *114*, 122−129.
(21) Mostafaei, H.; Ikonen, T.; Kramb, J.; Deneke, T.; Heljanko, K.; Harjunkoski, I. Data-Driven Approach to Grade Change Scheduling Optimization in a Paper Machine. *Ind. Eng. Chem. Res.* **2020**, *59*, 8281−8294.
(22) Sutton, R. S.; Barto, A. G. *Reinforcement learning: An introduction*; MIT press: Cambridge, MA, USA, 2018.
(23) Arulkumaran, K.; Deisenroth, M. P.; Brundage, M.; Bharath, A. A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26−38.
(24) Hubbs, C. D.; Li, C.; Sahinidis, N. V.; Grossmann, I. E.; Wassick, J. M. A Deep Reinforcement Learning Approach for Chemical Production Scheduling. *Comput. Chem. Eng.* **2020**, 106982.
(25) Šemrov, D.; Marsetič, R.; Žura, M.; Todorovski, L.; Srdic, A. Reinforcement learning approach for train rescheduling on a single-track railway. *Transp. Res. Part B: Methodol.* **2016**, *86*, 250−267.
(26) Atallah, R. F.; Assi, C. M.; Khabbaz, M. J. Scheduling the operation of a connected vehicular network using deep reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 1669−1682.
(27) Shin, J.; Badgwell, T. A.; Liu, K.-H.; Lee, J. H. Reinforcement learning−overview of recent progress and implications for process control. *Comput. Chem. Eng.* **2019**, *127*, 282−294.
(28) Ikonen, T. J.; Harjunkoski, I. Decision-making of online rescheduling procedures using neuroevolution of augmenting topologies. *Comput.-Aided Chem. Eng.* **2019**, *46*, 1177−1182.
(29) Stanley, K. O.; Miikkulainen, R. Evolving neural networks through augmenting topologies. *Evol. Comput.* **2002**, *10*, 99−127.
(30) Stanley, K. O.; Clune, J.; Lehman, J.; Miikkulainen, R. Designing neural networks through neuroevolution. *Nat. Mach. Intell.* **2019**, *1*, 24−35.
(31) Ikonen, T. J.; Heljanko, K.; Harjunkoski, I. Reinforcement learning of adaptive online rescheduling timing and computing time allocation. *Comput. Chem. Eng.* **2020**, 106994.
(32) Shaik, M. A.; Floudas, C. A. Novel unified modeling approach for short-term scheduling. *Ind. Eng. Chem. Res.* **2009**, *48*, 2947−2964.
(33) Rasmussen, C.E.; Williams, C.K.I. *Gaussian processes for machine learning*; MIT Press: Cambridge, MA, USA. 2006.
(34) Ikonen, T. J.; Harjunkoski, I. Incorporation of parameter prediction models of different fidelity into job shop scheduling. *IFAC-PapersOnLine* **2019**, *52*, 142−147.
(35) Borrison, R.; Klöpper, B.; Chioua, M.; Dix, M.; Sprick, B. Reusable Big Data System for Industrial Data Mining - A Case Study on Anomaly Detection in Chemical Plants. In *International Conference on Intelligent Data Engineering and Automated Learning*; Springer: Cham, 2018, 611, DOI: 10.1007/978-3-030-03493-1_64.
(36) Ming, L.; Zhao, J. Review on chemical process fault detection and diagnosis. In *2017 6th International Symposium on Advanced Control of Industrial Processes* IEEE: 2017, DOI: 10.1109/ADCONIP.2017.7983824.
(37) Lucke, M.; Stief, A.; Chioua, M.; Ottewill, J. R.; Thornhill, N. F. Fault detection and identification combining process measurements and statistical alarms. *Control Eng. Pract.* **2020**, 104195.
(38) Calfa, B. A.; Grossmann, I. E.; Agarwal, A.; Bury, S. J.; Wassick, J. M. Data-driven individual and joint chance-constrained optimization via kernel smoothing. *Comput. Chem. Eng.* **2015**, *78*, 51−69.
(39) Charnes, A.; Cooper, W. W. Chance-constrained programming. *Manage. Sci.* **1959**, *6*, 73−79.
(40) Shang, C.; Huang, X.; You, F. Data-driven robust optimization based on kernel learning. *Comput. Chem. Eng.* **2017**, *106*, 464−479.
(41) Bertsimas, D.; Gupta, V.; Kallus, N. Data-driven robust optimization. *Math. Program.* **2018**, *167*, 235−292.
(42) Ben-Tal, A.; El Ghaoui, L.; Nemirovski, A. *Robust optimization (Vol. 28)*; Princeton University Press: Princeton, NJ, USA, 2009.
(43) Rice, W. F.; Cassady, C. R.; Nachlas, J. A. Optimal maintenance plans under limited maintenance time. In *Proceedings of the seventh industrial engineering research conference*; 1998, pp. 1−3.
(44) Khatab, A.; Diallo, C.; Venkatadri, U.; Liu, Z.; Aghezzaf, E.-H. Optimization of the joint selective maintenance and repairperson assignment problem under imperfect maintenance. *Comput. Ind. Eng.* **2018**, *125*, 413−422.
(45) Diallo, C.; Venkatadri, U.; Khatab, A.; Liu, Z. Optimal selective maintenance decisions for large serial k-out-of-n: g systems under imperfect maintenance. *Reliab. Eng. Syst. Saf.* **2018**, *175*, 234−245.
(46) Khatab, A.; Aghezzaf, E. H.; Djelloul, I.; Sari, Z. Selective maintenance optimization for systems operating missions and scheduled breaks with stochastic durations. *J. Manuf. Syst.* **2017**, *43*, 168−177.
(47) Liu, Y.; Chen, Y.; Jiang, T. On sequence planning for selective maintenance of multi-state systems under stochastic maintenance durations. *Eur. J. Oper. Res.* **2018**, *268*, 113−127.
(48) Cao, W.; Jia, X.; Hu, Q.; Zhao, J.; Wu, Y. A literature review on selective maintenance for multi-unit systems. *Qual. Reliab. Eng. Int.* **2018**, *34*, 824−845.
(49) Aarset, M. V. How to identify a bathtub hazard rate. *IEEE Trans. Reliab.* **1987**, *R-36*, 106−108.
(50) Meeker, W. Q.; Escobar, L. A. *Statistical methods for reliability data*; John Wiley & Sons: 2014.
(51) Sarhan, A. M.; Apaloo, J. Exponentiated modified Weibull extension distribution. *Reliab. Eng. Syst. Saf.* **2013**, *112*, 137−144.
(52) Kraft, D. *A software package for sequential quadratic programming*; Technical Report DFVLR-FB 1988, *88−28*.
(53) Ye, Y.; Grossmann, I. E.; Pinto, J. M. Mixed-integer nonlinear programming models for optimal design of reliable chemical plants. *Comput. Chem. Eng.* **2018**, *116*, 3−16.
(54) Ikonen, T. J.; Mostafaei, H.; Ye, Y.; Bernal, D. E.; Grossmann, I. E.; Harjunkoski, I. Large-scale selective maintenance optimization using bathtub-shaped failure rates. *Comput. Chem. Eng.* **2020**, 106876.
(55) Ugray, Z.; Paper, D.; Johnson, J. How Business Value Is Extracted from Operational Data: A Case Study. In *Digital Business Models*; Aagaard, A. (eds)Palgrave Macmillan: Cham, 2019.