Literature review on using data mining in production planning and scheduling within the context of cyber physical systems

Paola Martins Seeger, Zakaria Yahouni, Gülgün Alpan

Literature review on using data mining in production planning and scheduling within the context of cyber physical systems

Paola Martins Seeger, Zakaria Yahouni* and Gülgün Alpan

University of Grenoble Alpes, CNRS UMR 5272, G-SCOP Laboratory (Science for Conception, Optimization and Production), Grenoble, France

*corresponding author: zakaria.yahouni@grenoble-inp.fr

Literature review on using data mining in production planning and scheduling within the context of cyber physical systems

***Abstract*— within the context of Industry 4.0, Cyber Physical Systems (CPS) are defined as technologies that can manage interconnected systems between its physical assets and computational capabilities. Successful integration of industry 4.0 principals requires a digital transformation in the production workshops, to interconnect the objects and equipment with the decision process. Due to this informatization, a big amount of data is generated and big data analysis must be employed in order to extract useful knowledge from the collected raw data. A literature review is conducted in this paper to identify data mining methods and technologies used in the context of production planning and scheduling. The results consist of a classification of papers according to their research methodology, CPS level implementation and technological optimization techniques. Finally, new research directions and some insights are discussed with regard to the implementation of a production shop floor 4.0.**

***Index Terms*— Industry 4.0; Cyber Physical System; Big Data; Data mining; Scheduling; Planning.**

I. INTRODUCTION

Industrial companies are currently preparing to undergo a major digital transition to meet numerous challenges of the next generation industry referred to as industry 4.0. Some examples of these challenges are related to resource usage efficiency, making production flexible, cost and time optimization, ensuring continuity in the digital production chain, among others. Consequently, various issues arise and some of them are related to Big Data, Internet of Things (IoT), Artificial Intelligence (AI), Augmented Reality (AR), robust scheduling, and man-machine cooperation (Lu 2017; Oztemel and Gursev 2018).

Today, in a production workshop, industrial companies' objects and equipment are increasingly connected using sensors that regularly generate hundreds of gigabytes of data. Nevertheless, few knowledge is generated from the collected raw data due to its complexity and heterogeneity. To obtain synthetic decision-aid information concerning the workshop, a rigorous analysis of these data must be carried out. Extracted knowledge is expected to have a positive impact in industries while increasing efficiency, anticipating breakdowns, and making manufacturing more competitive (Qi and Tao 2018).

The presented research consists of a literature review on the application of data-related-methods for optimization, scheduling and planning a production system. The objective is to analyze these methods and practices within the context of industry 4.0. In the current literature, there are few papers reviewing data techniques in the industry context. For example, Ismail et al. (2009) focus on the applications of data mining tasks in planning and scheduling. O'Donovan et al. (2015) provides a systematic mapping study of big data technologies in manufacturing. Particularly, some authors review data analytics in different areas of manufacturing such as fault detection, maintenance, and quality control (Choudhary, Harding, and Tiwari 2009; Harding et al. 2006). Cheng et al. (2018) review the development of data mining techniques in the big data era from a general perspective. They focus on its applications in production scheduling, quality improvement, defect analysis, and fault diagnosis. (Takeda-Berger et al. 2020) in a recent conference paper, has conducted a systematic literature review that analyzes 31 papers (published after 2011) that use machine learning techniques in scheduling. The work shows that the number of publication has grown constantly from 2015. The paper focus on scheduling aspect and identifies the used techniques without further details about the strategy of implementation.

The existing literature do not take into account the full implementation aspects of industry 4.0. "How to implement" and "in which context" are usually two main questions that lack a general answer. To have a more general and global overview about data techniques usage within the new challenges brought by the industry 4.0 era, our study focuses on the implementation aspects. As stated by Lee et al. (2015), Cyber Physical System is the conceptual model required for industry 4.0 implementation and it includes several phases; from data collection up to process automation. To our knowledge, there is a lack of studies addressing how data techniques are implemented within the different phases of CPS. The purpose of this question is to understand the different data techniques that a company may use in order to bring added value to its processes. We focus primarily on using data mining for planning/scheduling processes. Therefore, our main contribution consists of proposing a CPS-based classification of literature articles and the analyses of the existing case study applications.

This literature paper has three major contributions:

1. First, recent papers are analysed using a PRISMA systematic research methodology to identify the main literature contribution that considers data techniques in planning and scheduling. The results distinguish case studies from conceptual contributions.
2. As Cyber Physical System is an important aspect of industry 4.0, papers are analysed based on the CPS levels. The objective is to understand the advancement of industry 4.0 implementation from a CPS perspective.

3. The next contribution consists of analysing deeply the case studies and identify the main data techniques used in planning and scheduling.

The remaining of this paper is organized as follows. The main concepts related to data and industry 4.0 are introduced in Section 2, followed by the literature search methodology in Section 3. The results are presented and discussed in Section 4. This section consists of subsections exposing the work methodology for reviewing the literature and making the proposed classifications. Finally, future research directions and conclusions are presented in the last section.

II. MAIN CONCEPTS

This section addresses some concepts related to the data era such as Data Analysis, Data Mining, Big Data, etc. The aim is to clarify these concepts and the relationship between them. A brief overview of the concept of Cyber Physical System and its levels within the context of industry 4.0 is also given. These concepts are used for the proposed classification presented in the results section (section 4).

A. Data concepts

According to Van der Aalst (2016), "Data Science" is an interdisciplinary field aiming to turn data into real value. We can say that data science aims to extract useful knowledge from structured or unstructured data using different methods and algorithms. "Big Data" is one of the most relevant concepts related to data science. In the literature, many definitions are proposed. Elgendy and Elragal (2014) define big data as "a data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures, analytics, and tools in order to enable insights that unlock new sources of business value". In a similar way, Chen et al. (2014) define big data as "datasets that could not be perceived, acquired, managed, and processed by traditional IT and software/hardware tools within tolerable time".

The significant changes brought by big data era has been recognized not only by researchers, but also by almost every company that wants to create additional value for its business. The global consulting agency McKinsey & Company in its publication of "Big data: The next frontier for innovation, competition, and productivity" mentioned that big data is now part of every sector and function of the global economy (Manyika et al. 2011). Other definitions are proposed by Qi and Tao (2018); and, Shukla et al. (2019). Usually when speaking about big data, authors refer to four main features (4Vs) that characterize it: Volume, Variety, Velocity and Value. The volume represents the size and how enormous the data are; the variety means the various formats, types and kind of uses of data; the velocity is used for the rapid generation and the rate with which data are changing; and, the value refers to the huge value but very low density of the data (Chen et al. 2014; Elgendy and Elragal 2014; Laney 2001). Besides Volume, Variety, Velocity and Value, Babiceanu and Seker (2016) in their review mentioned that recent papers propose Veracity, Vision, Volatility, Verification, Validation, and, Variability. The veracity is the consistency and trustworthiness of the data; the vision addresses the likelihood of data generation process; the volatility refers to the limits of the data useful life; the erification and validation address the conformity of the data; and, the variability refers to the data uncertainty and impreciseness (Babiceanu and Seker 2016). In this context, another interesting definition of big data is proposed by Gantz and Reisel (2011). It is defined as "a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis."

Data science and big data involve many other concepts, e.g. data analysis, data analytics, data mining, artificial intelligence, machine learning, etc. According to the work of Qi and Tao (2018), which focus on the decision making process in smart manufacturing, big data analysis must be implemented to identify the behavior features and patterns, and have an insight into the potential trends. A method of analysis is referred to as "Data Analytics". Data analytics is the process of applying algorithms in order to analyze sets of data. Also, it is used to extract previously unknown, useful, valid, and hidden patterns and information from large data sets, as well as to detect important relationships among the stored variables (Adams 2010; Elgendy and Elragal 2014). Russom (2011) added that data analytics is where advanced analytic techniques operate on big data sets.

In order to extract patterns from data, one of the most relevant concepts of data analytics is "Data Mining". For years, there was no distinction between data mining and knowledge discovery in databases (KDD). According to Fayyad et al. (1996), KDD was introduced as the overall process of discovering useful knowledge from data. Data mining was introduced as a step in the KDD process. Likewise, Corne et al. (2012) say that data mining is at the heart of the KDD process. In other words, data mining is the application of specific algorithms for extracting patterns from data (Fayyad et al. 1996; Vazan et al. 2011). This area of data science contributes to the decision-making process in modern industries as far as knowledge can be obtained from the big amount of data generated and collected nowadays.

As explained above, it is difficult to define the barriers between these data aspects. To avoid any ambiguity, we use in this work the term data mining or data techniques. Precisely in Section 4.3 of this paper, data techniques are based on the overview of tasks and algorithms in data mining proposed by Corne et al. (2012). Fig. 1 illustrates these data mining tasks and some of the associated algorithms.

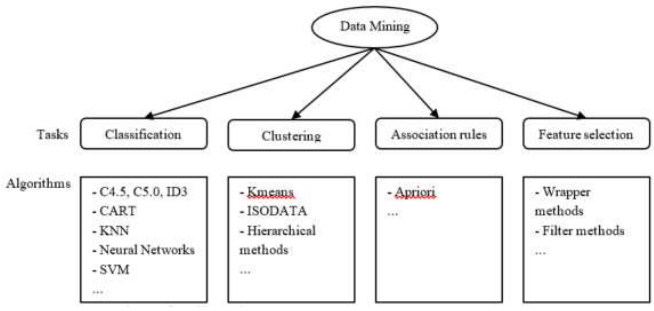

Figure 1. An overview of tasks and main algorithms in Data Mining (Corne et al. 2012)

Classification aims to build a model that predicts the value of one variable from the known values of other variables. Clustering is used to identify similarity between objects and to decompose a data set into groups. Association rules provide a very simple way to present correlations or other relationships among attributes. Finally, feature selection may be applied before one of the three previously presented tasks. It aims to find a good subset of features that form high quality clusters for a given number of clusters. In a manufacturing context, depending on the objectives that a company wants to achieve, some of these main four steps, if not all, are usually implemented to extract meaningful value from raw data.

Other concepts are related to data science and big data, such as artificial intelligence (AI) and machine learning concepts. According to Russell and Norvig (2010), AI is a set of techniques that enable machines to mimic human behavior. In computer science, AI characterizes systems and programs that can perform more complex tasks than direct programming (Mellit and Kalogirou 2008). One subfield of AI is machine learning. According to the International Business Machines (IBM 2019), machine learning enables a system to learn from data rather than through explicit programming. Applying machine learning techniques in manufacturing generates knowledge that may be used to help the decision makers or to improve the process directly (Wuest et al. 2016).

Both of these concepts (AI and machine learning) are not explored in depth in this review. The focus is mainly made on using data algorithms specifically data mining (such as presented in Fig. 1) for smart manufacturing. In Fig. 2 we propose a relationship summary between all aforementioned concepts. From a digitized transformation perspective, data science includes all concepts related to data. However, big data is a more general concept that may intersect with other concepts. The relation between data mining and machine learning is usually represented by some common algorithms, as shown in Fig. 2.

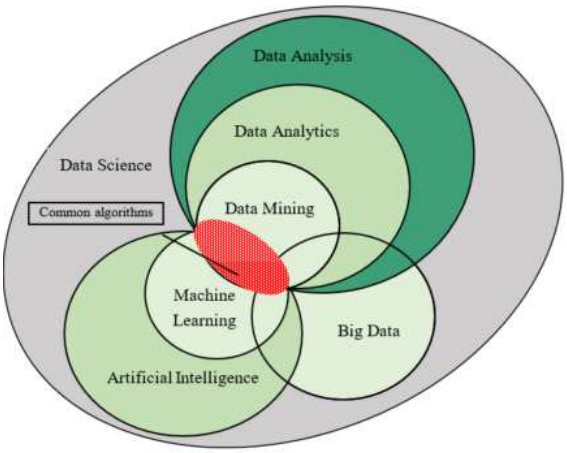

Figure 2. Relationship between Data concepts

B. Industry 4.0 and Cyber Physical System concepts

Besides the concepts related to data, there are two other important concepts for this study: Industry 4.0 and Cyber Physical System (CPS). The term “Industrie 4.0” was used for the first time in 2011 by a group of representatives from different fields in order to increase the German competitiveness in the manufacturing industry. Industry 4.0 defines a methodology to generate a transformation from machine dominant manufacturing to digital manufacturing based on automation and data exchange between machines (Lu 2017; Oztemel and Gursev 2018). Oztemel and Gursev (2018) emphasize the industry 4.0 components: Cyber Physical Systems (CPS), Cloud Systems, Machine to machine (M2M) communication, Smart factories, Big Data, etc. Some of these concepts and technologies have been initiated before 2011. Lee and Singh (2019) affirm that the Internet of Things initiated the fourth industrial revolution and propose a timeline of an evolution of disruptive technologies in manufacturing between 1999 and 2019.

A CPS is defined as transformative technologies for managing interconnected systems between its physical assets and computational capabilities (Lee et al. 2013, 2015). For Oztemel and Gursev (2018), CPS concerns the integration of computing and physical processes, which are essential components of industry 4.0 implementations. Baheti and Gill (2011) add that the ability to interact with, and expand the capabilities of, the physical world through computation, communication, and control is a key enabler for future technology developments. A benchmark on CPS architecture is proposed by Lee et al. (2015) and is regarded as a guideline for deploying industry 4.0. The proposed framework consists of five levels that span from data collection to decision-making:

- Level I: Smart Connection Level consists in collecting raw data from machines and their components using different resources such as sensors.
- Level II: Data-to-Information Conversion Level is where meaningful information is extracted from the collected data.
- Level III: Cyber Level is the central information in this architecture where the digital twin concept is incorporated.
- Level IV: Cognition Level generates a complete knowledge of the monitored system.
- Level V: Configuration Level refers to machines making decisions (self-configure, self-adjust and self-optimize).

The cyber level is regarded as the main contribution of the proposed CPS architecture. In this level, some advanced data techniques are used to mimic the behavior of the physical level, which is referred to as cyber twin (Lee et al. 2015). However, these levels are usually tackled from a conceptual point of view rather than a technical point of view. From an industry 4.0 implementation perspective, it is important to stress the different techniques that can be used within each level to achieve this CPS completeness. Therefore, our main research questions are to understand how-and-which data techniques to use for production planning/scheduling in a digitized industry. To answer these questions, we propose a

literature review about works addressing the problem of planning and scheduling from a data perspective and industry 4.0 point of view. In the following section, we start by explaining the literature search strategy/methodology for seeking such papers.

III. Research Methodology

In order to achieve the objectives of this research, the literature review process proposed by Cooper (1986) and a "Preferred Reporting Items for Systematic reviews and Meta-Analyses" (PRISMA) (Moher et al. 2009) are used. Fig. 3 shows the flow process of the bibliography research. First, in order to determine the relevant keywords that can be used for our searching process, some previous literature articles related to our main research question are explored (Chen et al. 2014; Elgendy and Elragal 2014; Fayyad et al. 1996; Lu 2017; Vazan et al. 2011; Chehbi-Gamoura et al. 2020). After a deep analysis, 17 keywords are extracted as shown in Table 1. From all the possible combinations, 59 relevant combinations are chosen. Each combination usually consists of one word related to data, one related to manufacturing/industry 4.0 and one related to scheduling/planning /optimization.

TABLE I
KEYWORDS AND DATABASES

Keywords			Databases
Category 1	Category 2	Category 3	
Internet of things	Manufactur*	Schedul*	Web of Science (WOS) Scopus IEEE
Data mining	Production	Optimi*	
Data analy*	Cyber Physical Production System	Plan*	
Big data Data	Cyber Physical System	Production control	
	Industr* 4.0	Product* plan*	
		Product* optimi*	
		Product* schedul*	

*Words truncated

Three major databases are used for the research as shown in Table 1 and the initial results represent 12903 papers (recorded as n1 in Fig. 3). The search is based on the title, abstract and keywords. 2591 are from Web of Science (WOS), 7339 from Scopus and 2973 from IEEE. It is important to highlight that no limitations are imposed for the initial phase, e.g. publication year, language, journals, conferences. Additionally, seven other papers are included (recorded as n2 in Fig. 3). Two of them are related to Cyber Physical System; the others are related to planning and scheduling in the context of industry 4.0 (Cheng et al. 2018; Fang et al. 2020, Gopalakrishman et al. 2020, Lee et al. 2013, 2015; Lu 2017; Trstenjak and Cosic 2017). The first filter of the systematic research consists of removing the duplicates among the three databases. The obtained results dropped to 5990. Then, only papers written in English are selected which reduces the number to 5202. In the same step, an analysis of the titles of these papers is conducted by combining two keywords, in order to focus on scheduling and planning issues in production/manufacturing. All combinations are illustrated in Table 2. After this process, the number of papers is narrowed down to 795.

TABLE 2
TITLE ANALYSIS

Keywords combination for the title analysis	Number of combinations
Schedul/Optimi/Plan/Data AND Manufactur/Production	8
Data AND Optimi/Schedul/Plan	3

In the eligibility step, all titles and some abstracts are manually analyzed in order to exclude some of the articles that are not related to our research question (the use of data techniques to improve the manufacturing shop floor). The selected exclusion keywords cover: additive manufacturing; grid computing; data stream; carbon; green; smart cities; data centers; agriculture; air traffic; data warehouse; project planning; public transportation; data/smart grid; data transfer; medical; wireless network; railway; micro grids. After the exclusion process, the results represent 513 papers.

Finally, a citation-based index is proposed to rank the remaining papers. The index considers the citation number of a paper with regard to its year of publication. For instance, a paper that has been cited 30 times and has been published in 2005 (14 years ago from 2019) has a proposed index of 2.14 (30/14). Papers are then ranked using this index and those with an index of at least 0.5 are analyzed manually based on the title and abstract. We note that the papers published in 2018-2020/2021 included are all considered in the eligibility phase regardless of the citation-based index. In the end, 60 papers are selected for the analysis.

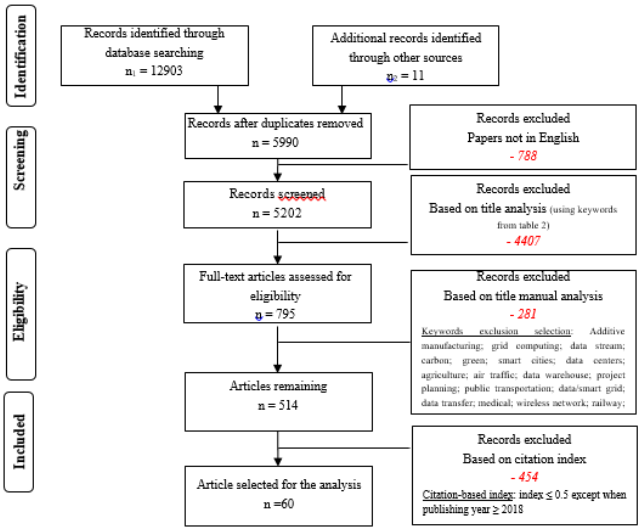

Figure 3. Systematic research based on PRISMA

IV. Analysis of the Selected Literature

This section presents the results and discussion about the 52

papers analyzed. Three cross analysis are performed. First, in Section 4.1 a work methodology classification is applied, followed by the CPS-based classification in Section 4.2. Finally, in Section 4.3, a classification based on data mining techniques regarding planning and scheduling is proposed.

A. Classification based on work methodology

This classification is based on whether the work is: a Review, Conceptual/Theoretical work, or a Case study/Application. Table 3 shows the result of the classification. Review papers are studied mainly to provide an overview on the topic and complete the discussion about the research question of this work. Papers are considered conceptual/theoretical when theories and/or conceptual-framework are discussed or the authors propose a new approach to solve a problem without further application. Case studies and applications concern articles with a practical implementation. This implementation may be a real case study (e.g. implementation in industry) or based on a standard literature problem (e.g. job shop scheduling problem). Some papers can be classified into more than one category such is the case for the second row of Table 3.

TABLE 3
Work methodology classification

Author	Review	Conceptual/ Theoretical	Case study/ Application	Number of papers
(Babiceanu and Seker 2016), (Ismail et al. 2009), (Harding et al. 2006), (Choudhary et al. 2009), (Qi and Tao 2018), (Rossit and Tohmé 2018), (Uhlmann and Frazzon 2018), (Shukla et al. 2019), (O'Donovan et al. 2015), (Singhal, Qi, and Ganeshan 2018), (Lu 2017), (Cheng et al. 2018), (Takeda-Berger et al. 2020).	x			13
(Bubeník and Horák 2014), (Schuh et al. 2017), (Harrath, Chebel-Morello, and Zerhouni 2002), (Altaf et al. 2018), (Zhong et al. 2014), (Balasundaram, Baskar, and Sankar 2012), (Wang et al. 2014), (Kück et al. 2016), (Shahzad and Mebarki 2012), (Zhu, Qiao, and Cao 2017), (Metan, Sabuncuoglu, and Pierreval 2010), (Woo, et al., 2018), (Gopalakrishman et al. 2020), (Leusin et al. 2018), (Morariu et al. 2020), (Ji, Yin, & Wang, 2019), (Ritou, et al., 2019).		x	x	17
(Lanza, Stricker, and Moser 2014), (Uhlemann et al. 2017), (Seitz and Nyhuis 2015), (Lee et al. 2015), (Lee et al. 2013), (Trstenjak and Cosic 2017), (Lee and Singh 2019), (Khan et al. 2015), (Bubenik, et al., 2014), (Li et al., 2015), (Blum, et al., 2017), (Fang et al. 2020),		x		12
(Wang 2007), (Ning, 2018), (Mao, et al., 2015), (Karthikeyan, et al., 2012), (Bergmann, et al., 2015), (Zhong, et al., 2015), (Makrymanolakis, et al., 2016), (Mavin, et al., 2018), (Subramaniyana, et al., 2018), (Zhong, 2018), (Kozjek, et al., 2018), (Küfner, et al., 2018), (Ning, et al., 2018), (Zahmani, et al., 2019), (Pimentel, et al., 2018), (Wauters, et al., 2011), (Dolgui, et al., 2018), (Yahouni, et al., 2021)			x	18
Total				60

According to Table 3, 13 review papers were studied to provide an overview about the topic and analyze what has been done until the end of 2018. Analysis of these papers shows that current literature work reviews already big data, industry 4.0 and related concepts, digital twin, etc. However, few information are given about considering data techniques such as data mining within the implementation aspects of the industry 4.0 in the context of planning and scheduling.

Lu (2017) presents an overview of the content, scope, and findings of Industry 4.0. Qi and Tao (2018) review the concepts and applications of big data and digital twins. In the same way of big data concepts, Singhal et al. (2018) and Shukla et al. (2019) summarize the concepts emerged with big data and the use of big data analytics in manufacturing and

service sectors. Differently, Babiceanu and Seker (2016) review the current status of virtualization and cloud-based services for manufacturing systems and the use of big data analytics from a general point of view. Uhlmann and Frazzon (2018) identify what has been studied about the production rescheduling process. Finally, Rossit and Tohmé (2018) investigate how the classical process of scheduling operations is affected by CPS. These papers address the impact of CPS and industry 4.0 elements on the decision/planning/scheduling process from a conceptual point of view without addressing in detail the potential data techniques and methods to emphasize the implementation aspects.

Examples of other studies reviewing data mining are those of Cheng et al. (2018), Ismail et al. (2009), O'Donovan et al. (2015), Choudhary et al. (2009) and Harding et al. (2006). Some authors review data mining in different areas of manufacturing such as fault detection, maintenance, and quality control (Choudhary et al. 2009; Harding et al. 2006). More general, O'Donovan et al. (2015) provides a systematic mapping study of big data technologies in manufacturing. Ismail et al. (2009) focus their discussion on the application of data mining tasks in planning and scheduling. Moreover, the work in Cheng et al. (2018) reviews the development of data mining techniques in the big data era by analyzing relevant papers from 2010. They focus on the applications of data mining techniques in production scheduling, quality improvement, defect analysis, and fault diagnosis. For the planning and scheduling problem, seven papers are identified which handle the problem from two perspectives; they either extract new scheduling rules using data mining or dynamically choosing the optimal schedule.

Even that these papers are somehow related to our work, there is still a missing link between the three aspects: planning/scheduling, data techniques (data mining) and industry 4.0 implementation. The concept of CPS has emerged recently and stresses the implementation issue of industry 4.0. Moreover, the number of papers considering data analysis in the last decade has significantly increased and new techniques are proposed. Therefore, the aim of our work is to understand the relation between planning/scheduling, data techniques and CPS implementation. In the next section, we present an overview about the work done within each level of the CPS, so as to provide some insights about how to implement a complete digital production chain.

B. CPS-based classification

CPS concerns the management of interconnected systems. According to Lu (2017), the main roles of a CPS are to fulfill the agile and dynamic requirements of production, and to improve the effectiveness and efficiency of the entire industry. Oztemel and Gursev (2018) complement that sensors and CPS facilitate easy communication capability between machines. In Table 4, a CPS-based classification of the conceptual/theoretical and case study/application articles is proposed based on the five levels introduced in Lee et al. (2015).

TABLE 4
CPS-based classification

Work methodology	Author	Level I	Level II	Level III	Level IV	Level V	Number of papers
Case study	(Mao, et al., 2015), (Küfner, et al., 2018)		x				2
	(Karthikeyan, et al., 2012), (Pimentel, et al., 2018)				x		2
	(Zhong, 2018)	x	x				1
	(Zhong, et al., 2015)	x			x		1
	(Zahmani, et al., 2019), (Ning & You, 2018, p. b), (Kozjek, et al., 2018), (Makrymanolakis, et al., 2016), (Fang et al. 2020), (Dolgui, et al., 2018).		x		x		6
	(Ning & You, 2018, p. a)			x	x		1
	(Wang 2007), (Wauters, et al., 2011), (Yahouni, et al., 2021)	x	x		x		2
	(Mavin, et al., 2018), (Subramaniyana, et al., 2018)	x		x	x		2
	(Bergmann, et al., 2015)		x	x	x		1

Conceptual/ Theoretical	(Khan et al. 2015), (Bubenik, et al., 2014)		x				2
	(Uhlemann et al. 2017), (Blum, et al., 2017)			x			2
	(Lanza et al. 2014)				x		1
	(Li et al., 2015)	x			x		1
	(Seitz and Nyhuis 2015), (Lee et al. 2013), (Trstenjak and Cosic 2017)	x	x	x	x		3
	(Lee et al. 2015), (Lee and Singh 2019)	x	x	x	x	x	2
Conceptual/ Theoretical and Case study/ Application	(Harrath et al. 2002),		x				1
	(Balasundaram et al. 2012), (Ji, Yin, & Wang, 2019)		x		x		2
	(Kück et al. 2016), (Shahzad and Mebarki 2012)			x	x		2
	(Zhu et al. 2017)	x	x	x			1
	(Schuh et al. 2017), (Wang et al. 2014), (Gopalakrishman et al. 2020), (Morariu et al. 2020), (Ritou, et al., 2019)	x	x		x		4
	(Bubeník and Horák 2014)		x	x	x		1
	(Altaf et al. 2018), (Zhong et al. 2014), (Metan et al. 2010), (Woo, et al., 2018)	x	x	x	x		4
	(Leusin et al. 2018)	x		x	x		1
	Total of papers for each level	24	34	20	38	2	
	Total of papers						47

In the above table, papers are classified in the Smart Connection Level (level I) when they address data collection. In the Data-to-Information Conversion Level (level II), we seek for the transformation of the raw data into information using some techniques, among which we can cite the data mining techniques, for instance. The Cyber Level (level III) is related to the digital twin. Articles are classified in this level, if they present some information generated from the connected machines; specific analytics must be used. Articles are classified in the Cognition Level (level IV) when data are synthetized and decisions are taken by the expert users. Finally, the Configuration Level (level V) allows closing the loop by interconnecting the fifth level to the first. For Lee et al. (2015; 2019), in this level, machines are able to self-configure, self-adjust and self-optimize, without human interaction. For such self-sufficiency, the machines must be capable of performing all levels of CPS concept, starting from level 1.

As mentioned before, an implementation of an industry 4.0 consists of implementing a CPS within the five levels perspectives. We observe from Table 4 that only two papers consider all five levels and are rather theoretical. Most of the papers tackle only specific parts of the CPS digital chain. Therefore, in the following, we try to understand the practical

implementation of each level so to understand the link between levels and overall the practical implementation of a CPS.

- Level I and Level II

Thirty-seven out of the forty-seven selected papers consider at least one of these two levels. Eighteen papers address the first level and stress the issue of data collection. Acquiring accurate and reliable data from machines then transforming it into useful information is often a challenge. Usually, data is captured by sensors, controllers or information systems (Lee et al. 2015). (Seitz and Nyhuis 2015) discussed the challenges of combining CPS with logistic models for improving production monitoring and control. One of the most used technologies is radio frequency identification (RFID), mentioned in Altaf et al. (2018), Zhong et al. (2014) and Fang et al. (2020). RFID is widely used in manufacturing for collecting shop floor data. Cutting force, vibration, acoustic emission, temperature, pressure, task starting/finishing time, are some of the different types of signals that can be captured from the workshop. Data can be transmitted through a Manufacturing Execution System such as in (Wauters, et al., 2011).

The captured data can be enormous and full of inaccurate, incomplete and missing records (Zhong et al. 2014). For Schuh et al. (2017) the quality of the decision-making is highly dependent on the quality and integrity of the used data. In this context, three prerequisites are mentioned and must be fulfilled for excellent data integrity: the data need to be complete, consistent and correct (Schuh et al. 2017).

However, in most cases raw data are barely useful, even if it is free of noise, and several further steps are required to extract value from it (Qi and Tao 2018). This concerns the second level of CPS: Data is transformed into synthetic information using data analysis techniques such as data mining. Fig. 4 shows the intersection between data mining and CPS levels. Examples of papers that apply data mining algorithms are: Balasundatam et al (2012), Bubeník and Horák (2014), Harrath et al. (2002), and (Khan et al. 2015). In this second level, different data algorithms may be applied such as decision tree, hierarchical ascending clustering, etc. Moreover, in a manufacturing context, data can be difficult to process due to its complexity, diversity and time-scale heterogeneity. The processing step can be very specific to the manufacturing context/expertise.

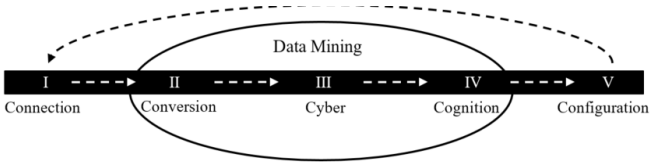

Figure 4. Relationship between CPS levels and data mining

- Level III

This level characterizes the key aspect of the cyber physical architecture, where the physical word communicates with the virtual world. It consists of transforming the information generated in level II to a more synthetic information for the decision making process (level IV). It is also the intermediate phase of data mining application where knowledge is generated.

In this level, the concept of digital twin comes out and different definitions are proposed in the literature. According to Schroeder et al. (2016), it is a virtual representation of a real process in the context of Cyber Physical System. For Gabor et al. (2016) digital twin is the simulation of the physical object itself to predict future states of the system. Other definitions can be found in Negri et al. (2017). In a manufacturing context, a cyber-twin is responsible for capturing time machine records and synthesizing future steps to provide self-awareness and self-prediction as stated by (Lee et al. 2015). Also, recently it was pointed out that data-driven modeling can enable manufacturing companies to obtain useful information and integrate them with other technologies for improving productivity and innovation (Lee and Singh 2019).

Most of the analyzed papers mention simulation when it comes to the digital twin. In (Zhu et al. 2017) for example, a data-based scheduling discrete event simulation model is proposed to mimic the behavior of the physical world. Collected production data are fed back to the model to update the parameters of the scheduling model and keep it effective/accurate with respect to the physical system. The idea is to predict a key performance indicator at time t using simulation. Then, compare it with the real value of the manufacturing system at time t. If a deviation is found, dispatching rules are adjusted at time t+1. In (Mavin, et al., 2018), simulation is used with optimization for the scheduling problem of a construction product manufacturing facility. The idea is to evaluate the feasibility of a pre-calculated schedule through simulation. This will help schedulers evaluate what-if scenarios either during the scheduling or the rescheduling phase if a disturbance occurs.

These papers consider other levels as well, such as the previous ones or the next one (level IV) where simulation is used for decision-making. In a manufacturing context, simulation models are highly dependent on the pre-processed collected data (level I and II). Models are used for adjusting the parameters of the simulation model or for evaluating a KPI of a possible solution. The challenge is to create a simulation model that mimics the exact dynamic behavior of a system, which is difficult due to the changes that may arise in the real system. However, simulation can be very crucial for the decision process and especially for evaluating a what-if scenario (level IV).

- Level IV and Level V

According to Lee and Singh (2019), the up-to-date real time information would be a waste if the results cannot be communicated clearly and effectively to decision makers. In the fourth level of a CPS, through graphs, charts and reports, humans make decisions and improve the system. Using data mining analysis, the behavior features and patterns can be identified in order to provide some insights about the trends, helping users making decisions (Qi and Tao 2018). This process is considered as the last step of data mining (Fig. 4).

The decision may be classified in short, medium or long-term. In business forecasting, short-term usually refers to under three months; medium-term, three months to two years; and long-term, greater than two years (Ismail et al. 2009). In a production workshop we could use the same reference time,

e.g. if a machine-tool degrades the quality of a production part, deciding whether to stop the process or not could be considered as short term. Moreover, planning the production or investing in new equipment could be considered tactical and strategic, respectively, and therefore medium-long term.

From Table 4, it can be shown that thirty-eight papers consider level 4. Different works has been proposed combining data mining and decision making/aiding. For example, in (Karthikeyan, et al., 2012), authors use a combination of particle swarm optimization algorithm (PSO) and data mining for solving flexible job shop scheduling problems. PSO is used to propose a set of solutions and generated data is mapped into different classes to find the relationship between operations' characteristics and their order. In (Zhong & Xu, 2015), a data-driven model for the job shop problem is proposed. This model uses real-time data as feedback to correct the scheduling parameters and improve decision-making. In (Wang 2007), author discusses about the use and implementation of data mining for product design and manufacturing. Two practical cases are studied: manufacturing maintenance decision-making and predicting assembly quality. Decision tree is used for both cases. Author stresses the limits of using datamining techniques such as: time consumption for data cleaning/preparation and need of good collaboration between users and experts to have a clear understanding of the manufacturing problem and the provided solution. Moreover, in (Subramaniyana, et al., 2018), a predictive algorithm is applied for detecting potential machines bottlenecks using machines active periods. ARIMA (Autoregressive Integrated Moving Average) statistical method is applied on real data captured from the manufacturing execution system to forecast active periods of machines and detect future bottlenecks.

In the work of (Pimentel, et al., 2018), simulation is combined with data-driven optimization for solving a flexible job shop problem. In this case study, simulation is used as a decision making tool. The proposed scenario consists of computing the order between tasks using a genetic algorithm, then simulation is used for estimating the completion time of the schedule. Simulation is used in this case to implement the probabilistic processing times of operations. If a disturbance occurs such as a machine breakdown, a change in the processing time, etc., the genetic algorithm computes another schedule based on the new data. The management team uses the simulation results to make decisions and cope with disturbances.

The last level of CPS usually refers to the automation of the processes based on the collected feedback and decisions made in the cognition level (Lee et al. (2015; 2019)). The loop of the CPS is then completed and the whole process is supposed to function automatically without too much human intervention. In the studies presented in Lee et al. (2015; 2019), automation guarantees a worry free, near zero downtime production and production planning optimization.

From the table above, it can be seen that there are no case studies yet considering level IV and level V at the same time. This may be because the majority of connected devices in manufacturing are not that smart yet and not able to make all decisions without human intervention (Lee and Singh 2019). Furthermore, due to the perturbations that may occur on manufacturing shop floors, human intervention during the decision process is crucial. Especially for medium and long-term decisions, such as in planning. Therefore, the configuration level can be characterized by a partial automation depending on the problem nature. For a short time decision, such as stopping a machine, data analytics can provide KPI about failures or quality defects that cannot be manually or rapidly detected and need to be urgently taken care of.

C. Classification based on data techniques

CPS combined with logistic models has the potential of improving production planning and control. To understand how CPS can be implemented in practice, we focus on only case studies. Then, only papers describing data mining techniques in planning or scheduling are analyzed. Table 5 summarizes the analysis of the twenty-six selected papers. The objective is first, to highlight the most used data techniques (clustering, prediction, etc.) and seconds, to understand how optimization algorithms are used with these techniques in the context of planning and scheduling.

Table 5 shows that most of the papers that use data mining focus on the classification process using the decision tree technique. Besides that, C4.5 program is mostly used for generating the decision tree/rules. Decision tree is a supervised machine learning method for constructing prediction models from data (Balasundaram et al. 2012). It is one of the most popular techniques due to its easier use and its simplicity to be understood and interpreted (Balasundaram et al. 2012; Corne et al. 2012).

Furthermore, the most used optimization scheduling techniques are simple dispatching rules such as FIFO (First In First Out), EDD (Earliest Due Date), SPT (Shortest Processing Time) and CR (critical ratio). The fact that they are often used can be explained by their simplicity. Dispatching rules are characterized by a low effort of implementation as explained in (Kück et al. 2016). In some cases, metaheuristics such as Simulated Annealing, Swarm particle optimization or Genetic algorithms are used.

Different approaches are considered for coupling these data techniques with optimization and simulation. Three interesting approaches are highlighted in the following:

1) Many authors exploit historical data using data mining techniques to discover or select dispatching rules. See for instance the work by (Shahzad and Mebarki 2012), (Bergmann, et al., 2015), (Zhu et al. 2017), (Zahmani, et al., 2019), (Metan et al. 2010) and (Kozjek, et al., 2018). In the work of Zhu et al. (2017) for example, a discrete event simulation model is proposed to mimic the behavior of the physical world. Then, scheduling strategies are extracted from the simulation and applied to the production line of wafers in semi-conductors industry.

TABLE 5 CLASSIFICATION OF DATA MINING AND OPTIMIZATION TECHNIQUES WITH REGARD TO PLANNING AND SCHEDULING

Planning/Scheduling	Author	Data mining techniques	Optimization techniques (Heuristic/Metaheuristic)
Planning and Scheduling	(Kozjek, et al., 2018)	Random forest algorithms and regression for prediction	Dispatching rule, simulation
	(Subramaniyana, et al., 2018)	Autoregressive Integrated Moving Average (ARIMA) statistical method	
	(Ning & You 2018a)	Parameter correlation (kernel density estimation)	
	(Altaf et al. 2018)	RANSAC	Simulated Annealing, Particle Swarm Optimization
	(Zhong et al. 2014)	Decision tree; C4.5	Dispatching rule
Planning	(Bubeník and Horák 2014)	Decision tree; C4.5	
	(Mavin, et al., 2018)	Simple preprocessing of data	Hybrid simulated annealing with genetic algorithm, Simulation
	(Zhong, 2018)	Data cleasing and K-means for data clusturing	
	(Küfner, et al., 2018)	Artificial Neural Network for prediction	
	(Ning & You, 2018b)	Likelihood estimation for the Multinoulli distribution,	
	(Woo, et al., 2018), (Ji, Yin, & Wang, 2019)	Artificial Neural Network and Regression (for Woo et al. 2018)	
	(Schuh et al. 2017)	Adapted Association rules induction	
Scheduling	(Wang 2007)	Decision tree; C4.5	
	(Harrath et al. 2002)	ChiMerge; Decision tree (C4.5 and See5)	Genetic algorithm
	(Balasundaram et al. 2012)	Decision tree	Dispatching rule
	(Wang et al. 2014), (Metan et al. 2010)	Decision tree; C4.5	Dispatching rule
	(Shahzad and Mebarki 2012)	Decision tree	Dispatching rule; Tabu search
	(Karthikeyan, et al., 2012)	Attribute-oriented mining for classification & Concept hierarchy for clustering	Particle Swarm Optimization
	(Zahmani, et al., 2019)	Decision tree with multi-label classification	Dispatching rules, Simulation
	(Makrymanolakis, et al., 2016)	Decision tree	Threshold acceptance (deterministic version of simulation anealing)
	(Fang, et al., 2020)	Deep Neural Network for prediction and	
	(Bergmann, et al., 2015)	K-Nearest Neighbors, decision trees, Naive Bayes Classifier and support vector machines	
	(Morariu et al. 2020)	One class Support Vector Machine	
	(Wauters, et al., 2011)	Neural Network, Regression trees, model trees and K-nearest neighbor	LSP scheduling and local search

2) Other works use data techniques for estimating the processing time or finding the probability distribution of operations such as (Pimentel, et al., 2018), (Zhong et al. 2014), (Ning & you, 2018a), (Wauters, et al., 2011). The estimated values are then used in the simulation and/or optimization algorithms.
3) More rarely, Data mining is also used for parameter setting of optimization/simulation tools. For example, (Makrymanolakis, et al., 2016) estimate the three parameters (number of solutions, number of repetition and number of sets) of the threshold accepting metaheuristics using a decision tree.

V. DISCUSSION

It can be seen from the CPS-based classification (table 4) that there is a missing link between all levels. This link should ensure the digital chain continuity between the physical world (machines and equipment) and the higher-level services in a company such as decision-making. This classification (table 4) still does not explain the practical implementation of a CPS such as "which data techniques are used in these levels". Nevertheless, the proposed data mining classification (table 5) helps understanding which algorithms are mostly used in planning and scheduling and how optimization algorithms can be enhanced using data mining.

Overall, different optimization approaches are used with data mining in the context of planning and scheduling as shown in table 5. The aim of such techniques is to deploy optimal/near-optimal scheduling rules, but different challenges have to be considered. The scheduling problem is NP-hard. Furthermore, it is often a non-deterministic problem in the industrial context even though most optimization methods treat it as a deterministic one. The problem has been studied for decades and robust optimization where flexibility is introduced seems to be a more adaptive solution for real case studies. To handle the uncertainties and cope with the different disturbances that may occur in a shop floor, data mining techniques can be used for anticipating or reacting to these changes (i.e. constructing a flexible robust schedule for example).

In this context, simulation can be a very useful tool either for providing KPIs about what-if-scenario or for validating new models based on historical data. The idea behind this second point is to replay a historical scenario using a simulation model and check whether the simulation output matches the expected results. The challenge however, is constructing a robust simulation model that mimics exactly the behavior of the physical world. Then, test data can be generated from this simulation model and processed to anticipate a machine breakdown for example. In this scenario point of view, the digital twin represented by simulation should link the gap of what happened in the past and what is expected to happen in the future.

VI. CONCLUSION AND FUTURE RESEARCH DIRECTION

This paper proposes a literature review analysis about the usage of data techniques for optimizing the shop floor in the context of industry 4.0. The focus is mainly made on using data mining for planning and scheduling. By analyzing the selected papers from the literature, we observe that the concepts mainly related to data might be confusing. Therefore, we introduce the relationship between data concepts such as big data, data mining, data analytics, etc.

In view of the existing papers, we start by classifying the papers based on their work methodology, then based on the CPS levels. Finally, we investigate the applications of data mining with regard to the planning and scheduling problem. From the previous analysis, three main approaches are identified for optimizing the production workshop using data mining. The application of a data mining algorithm depends on the problem and objective context. However, without a doubt, these techniques can provide additional value within the digitalization era.

Studies in this area are still arising and different challenges are still to be addressed. These challenges are summarized in the following points and may be considered as future research directions:

- First, as said above research around this topic has significantly increased in the last few years. The year-on-year publication growth for big data in manufacturing and supply chain management from 2000 to 2015 is impressive (Lamba and Singh 2017) and the tendency is still growing (Takeda-Berger et al. 2020). Many papers were not included in this review due to the large number of articles. It is suggested for further analysis to use text-mining techniques to find patterns among papers and to automate the process of analyzing the articles. The technique used in Choudhary et al. (2009) can be applied for future research.
- Concerning the phases of data analytics and particularly data mining, the data collection phase is a crucial process, on which the final decision depends. Capturing data from the workshop is nowadays less challenging. However, processing these data in real time for quick decision-making is still problematic. Most of the studies use a posteriori analysis based on historical data. Thus, exploring cases where data are collected, cleaned, corrected and analyzed in short/real time may be an interesting topic.
- As seen in Section 4.2, the digital twin represents the main feature of a CPS for its integration and communication between the physical and virtual world. However, different definitions appear for digital twin and the Cyber level implementation in the industry is still at its development phase. For some authors, the definition of digital twin is related to simulation, for others, it is more general and is about virtual representation (Negri et al. 2017). Future perspective studies may focus on the CPS third level, analyzing the possible digital twin implementations within different contexts of application, e.g. production workshop.

- The proposed CPS-based classification in this work relies on the five levels proposed by Lee et al. (2015). As shown from the analysis, research addressing the whole connection between the five levels is still recent and rather conceptual. Moreover, most of the case studies are conference papers and are under development. Furthermore, the papers do not explore in depth the process of CPS automation/Configuration Level (level V). In this level, the machine for instance, is supposed to be smart and should be able to self-configure, self-adjust and self-optimize. This means that the results of data mining must be fed back to the machine in an automatic way. The advances in artificial intelligence might be an opportunity and open up new research possibilities at level V.
- Connection, communication and decision-making are some of the central themes of industry 4.0. However, future research could explore the position of humans in the loop of the future industry automation. Some important questions may arise concerning the Level of Automation. This question is not new and different levels from fully manual to fully automatic are proposed by researchers for decision-making and cognitive fields. However, the question is still open within the context of the new technologies brought by the era of industry 4.0 such as Automated Integrated vehicles, Cobots, etc.

AKNOWLEDGEMENT

This work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003).

REFERENCES

Adams, Niall M. 2010. “Perspectives on Data Mining.” *International Journal of Market Research* 52(1):11–19.

Altaf, Mohammed Sadiq, Ahmed Bouferguene, Hexu Liu, Mohamed Al-Hussein, and Haitao Yu. 2018. “Integrated Production Planning and Control System for a Panelized Home Prefabrication Facility Using Simulation and RFID.” *Automation in Construction* 85(February 2017):369–83.

Babiceanu, Radu F. and Remzi Seker. 2016. “Big Data and Virtualization for Manufacturing Cyber-Physical Systems: A Survey of the Current Status and Future Outlook.” *Computers in Industry* 81(2015):128–37.

Baheti, Radhakisan and Helen Gill. 2011. “Cyber Physical Systems.” *The Impact of Control Technology* 12(1):161–66.

Balasundaram, R., N. Baskar, and R. Siva Sankar. 2012. “A New Approach to Generate Dispatching Rules for Two Machine Flow Shop Scheduling Using Data Mining.” *Procedia Engineering* 38:238–45.

Bergmann S., Feldkamp N. et Strassburge S. 2015. "Approximation of dispatching rules for manufacturing simulation using data mining methods". *Winter Simulation Conference* (WSC). - Huntington Beach : [s.n.]. pp. 2329-2340.

Blum Matthias and Schuh Gùnther. 2017. "Towards a data-oriented optimization of manufacturing processes a real-time architecture for the order processing as a basis for data analytics methods" *ICEIS proceeding*.

Bubeník, Peter and Filip Horák. 2014. “Knowledge-Based Systems To Support Production Planning.” *Tehnicki Vjesnik-Technical Gazette* 21(3):505–9.

Chehbi-Gamoura, S. Derrouiche, R., Damand, D. Barth, M. 2020. “Insights from big Data Analytics in supply chain management: an all-inclusive literature review using the SCOR model.” *Production Planning and Control* 31 (5):355–382.

Cheng, Ying, Ken Chen, Hemeng Sun, Yongping Zhang, and Fei Tao. 2018. “Data and Knowledge Mining with Big Data towards Smart Production.” *Journal of Industrial Information Integration* 9:1–13.

Choudhary, A. K., J. A. Harding, and M. K. Tiwari. 2009. “Data Mining in Manufacturing: A Review Based on the Kind of Knowledge.” *Journal of Intelligent Manufacturing* 20(5):501–21.

Cooper, Harris M. 1986. “Organizing Knowledge Syntheses : A Taxonomy of Literature Reviews.”

Corne, David, Clarisse Dhaenens, and Laetitia Jourdan. 2012. “Synergies between Operations Research and Data Mining: The Emerging Use of Multi-Objective Approaches.” *European Journal of Operational Research* 221(3):469–79.

Dolgui, A., Bakhtadze, N., Pyatetsky, V., Sabitov, R., Smirnova, G., Elpashev, D., & Zakharov, E. (2018). Data Mining-Based Prediction of Manufacturing Situations. IFAC-PapersOnLine, 51(11), 316-321.

Elgendy, Nada and Ahmed Elragal. 2014. “Big Data Analytics: A Literature Review Paper.” *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8557 LNAI:214–27.

Fang, W., Guo, Y., Liao, W., Ramani, K., & Huang, S. 2020. "Big data driven jobs remaining time prediction in discrete manufacturing system: a deep learning-based". *International Journal of Production Research*, 58(9), 2751-2766.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. “From Data Mining to Knowledge Discovery in Databases.” *AI Magazine* 17(3):37–54.

Gabor, Thomas, Lenz Belzner, Marie Kiermeier, Michael Till Beck, and Alexander Neitz. 2016. “A Simulation-Based Architecture for Smart Cyber-Physical Systems.” *Proceedings - 2016 IEEE International Conference on Autonomic Computing, ICAC 2016* 374–79.

Gantz, John and David Reinsel. 2011. “Extracting Value from Chaos State of the Universe.” *IDC IView* (June):1–12.

Gopalakrishnan, M., Subramaniyan, M., & Skoogh, A. 2020. "Data-driven machine criticality assessment – maintenance decision support for increased productivity". Production Planning and Control, 1-19.

Harding, J. A., M. Shahbaz, Srinivas, and A. Kusiak. 2006. “Data Mining in Manufacturing: A Review.” *Journal of Manufacturing Science and Engineering* 128(4):969.

Harrath, Youssef, Brigitte Chebel-Morello, and Noureddine Zerhouni. 2002. “A Genetic Algorithm and Data Mining Based Meta-Heuristic for Job Shop Scheduling Problem.” *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* 7:280–85.

IBM. 2019. "Data Science and Machine Learning." Accessed May 10, 2019. https://www.ibm.com/analytics/machine-learning

Ismail, Ruhaizan, Zalinda Othman, and Azuraliza Abu Bakar. 2009. “Data Mining in Production Planning and Scheduling: A Review.” *2009 2nd Conference on Data Mining and Optimization, DMO 2009* (October):154–59.

Ji, W., Yin, S., & Wang, L. (2019). A big data analytics based machining optimisation approach. Journal of Intelligent Manufacturing, 30, 1483–1495.

Karthikeyan, S.; Asokan, P.; Nickolas, S.; Page, Tom. 2012. "Solving flexible job-shop scheduling problem using hybrid particle swarm optimisation algorithm and data mining" *International Journal of Manufacturing Technology and Management (IJMTM)*. Inderscience, 1-4 : Vol. 26. - pp. 81-103.

Khan, Abdul Rauf, Henrik Schiøler, Torben Knudsen, and Murat Kulahci. 2015. “Statistical Data Mining for Efficient Quality Control in Manufacturing.” *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA* 2015–October:1–4.

Kozjek, Dominic; Vrabic, Rok; Rihtarsic, Borut; Butala, Peter. 2018. "Big data analytics for operations management in engineer-to-order manufacturing" *Procedia CIRP*. Vol. 72. pp. 209-214.

Kück, Mirko, Jens Ehm, Michael Freitag, Enzo M. Frazzon, and Ricardo Pimentel. 2016. “Potential of Data-Driven Simulation-Based Optimisation Approach for Adaptive Scheduling and Control of Dynamic Manufacturing Systems.” *Advanced Materials Research* 1140:449–56.

Küfner Thomas, Uhlemann Thomas H.-J. et Ziegler Bastian. 2018. "Lean Data in Manufacturing Systems: Using Artificial Intelligence for Decentralized Data Reduction and Information Extraction" *Procedia CIRP*. Vol. 72. pp. 219-224.

Lamba, K. and Singh, S.P. 2017. " Big data in operations and supply chain management: current trends and future perspectives" *Production*

Planning and Control. 28(11-12):877-890.

Laney, Doug. 2001. “3-d Data Management: Controlling Data Volume, Velocity and Variety.” *META Delta* 949(February 2001):4.

Lanza, G., N. Stricker, and R. Moser. 2014. “Concept of an Intelligent Production Control for Global Manufacturing in Dynamic Environments Based on Rescheduling.” *IEEE International Conference on Industrial Engineering and Engineering Management* 315–19.

Lee, Jay, Behrad Bagheri, and Hung An Kao. 2015. “A Cyber-Physical Systems Architecture for Industry 4.0-Based Manufacturing Systems.” *Manufacturing Letters* 3:18–23.

Lee, Jay, Edzel Lapira, Behrad Bagheri, and Hung an Kao. 2013. “Recent Advances and Trends in Predictive Manufacturing Systems in Big Data Environment.” *Manufacturing Letters* 1(1):38–41.

Lee, Jay and Jaskaran Singh. 2019. “Industrial AI : Is It Manufacturing’s Guiding Light ?” (April).

Leusin, M. E., Frazzon, E. M., Maldonado, M. U., Kück, M., & Freitag, M. (2018). Solving the Job-Shop Scheduling Problem in the Industry 4.0 Era. Technologies, 6(4).

Li, Wenxiang; Pi, Chunchun; Han, Mei; Ran, Chong; Chen, Wei; Ke, Peng. 2015. "A scheduling method for IOT-aided packaging and printing manufacturing system" *11th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QSHINE)*. Taipei : [s.n.], pp. 335-340.

Lu, Yang. 2017. “Industry 4.0: A Survey on Technologies, Applications and Open Research Issues.” *Journal of Industrial Information Integration* 6:1–10.

Makrymanolakis N., Marinaki M. et Marinakis Y. 2016. "Data mining parameters' selection procedure applied to a multi-start local search algorithm for the permutation flow shop scheduling problem" *IEEE Symposium Series on Computational Intelligence (SSCI)*. - Athens. pp. 1-8.

Manyika, J., Brown Chui, M., J. B., Bughin, R. Dobbs, C. Roxburgh, and A. Hung Byers. 2011. “Big Data: The next Frontier for Innovation, Competition and Productivity.” *McKinsey Global Institute* (June).

Mao N. et Tan J. 2015. "Complex Event Processing on uncertain data streams in product manufacturing process" *International Conference on Advanced Mechatronic Systems (ICAMechS)*. Beijing : [s.n.], 2015. pp. 583-588.

Mellit, Adel and Soteris A. Kalogirou. 2008. “Artificial Intelligence Techniques for Photovoltaic Applications: A Review.” *Progress in Energy and Combustion Science* 34(5):574–632.

Metan, Gokhan, Ihsan Sabuncuoglu, and Henri Pierreval. 2010. “Real Time Selection of Scheduling Rules and Knowledge Extraction via Dynamically Controlled Data Mining.” *International Journal of Production Research* 48(23):6909–38.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, Doug Altman, Gerd Antes, David Atkins, Virginia Barbour, Nick Barrowman, Jesse A. Berlin, Jocalyn Clark, Mike Clarke, Deborah Cook, Roberto D’Amico, Jonathan J. Deeks, P. J. Devereaux, Kay Dickersin, Matthias Egger, Edzard Ernst, Peter C. Gøtzsche, Jeremy Grimshaw, Gordon Guyatt, Julian Higgins, John P. A. Ioannidis, Jos Kleijnen, Tom Lang, Nicola Magrini, David McNamee, Lorenzo Moja, Cynthia Mulrow, Maryann Napoli, Andy Oxman, Ba’ Pham, Drummond Rennie, Margaret Sampson, Kenneth F. Schulz, Paul G. Shekelle, David Tovey, and Peter Tugwell. 2009. “Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement.” *Journal of Chinese Integrative Medicine* 7(9):889–96.

Morariu, C., Morariu, O., Răileanu, S., Borangiu, & Theodor. (2020). Machine learning for predictive scheduling and resource allocation in large scale manufacturing systems. Computers in Industry, 2020, 103244.

Nahhas, Abdulrahman; Lang, Sebastian; Bosse, Sascha; Turowski, Klaus. 2018. "Toward Adaptive Manufacturing: Scheduling Problems in the Context of Industry 4.0". *Sixth International Conference on Enterprise Systems*. pp. 108-115.

Negri, Elisa, Luca Fumagalli, and Marco Macchi. 2017. “A Review of the Roles of Digital Twin in CPS-Based Production Systems.” *Procedia Manufacturing* 11(June):939–48.

Ning Chao and You, Fengqi 2018a. "Data-driven decision making under uncertainty integrating robust optimization with principal component analysis and kernel smoothing methods" *Computers & Chemical Engineering*. Vol. 112. pp. 190-210.

Ning Chao et You Fengqi. 2018b. "data-driven stochastic robust optimization: general computational framework and algorithm leveraging machine learning for optimization under uncertainty in the big data era". *Computers & Chemical Engineering*. Vol. 111. pp. 115-133.

O’Donovan, Peter, Kevin Leahy, Ken Bruton, and Dominic T. J. O’Sullivan. 2015. “Big Data in Manufacturing: A Systematic Mapping Study.” *Journal of Big Data* 2(1).

Oztemel, Ercan. and Samet. Gursev. 2018. “Literature Review of Industry 4.0 and Related Technologies.” *Journal of Intelligent Manufacturing* 1–56.

Qi, Qinglin and Fei Tao. 2018. “Digital Twin and Big Data towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison.” *IEEE Access* 6:3585–93.

Pimentel, Ricardo; Santos, Pedro P. P.; Danielli, Apolo M. Carreirão; Frazzon, Enzo M.; Pires, Matheus C. 2018. "Towards an Adaptive Simulation-Based Optimization Framework for the Production Scheduling of Digital Industries" *International Conference on Dynamics in Logistics*. pp. 257-263.

Ritou, M., Belkadi, F., Yahouni, Z., Da Cunha, C., Laroche, F., & Furet, B. (2019). Knowledge-based multi-level aggregation for decision aid in the machining industry. CIRP Annals, 68(1), 475-478.

Rossit, Daniel and Fernando Tohmé. 2018. “Scheduling Research Contributions to Smart Manufacturing.” *Manufacturing Letters* 15(November 2018):111–14.

Russell, Stuart and Peter Norvig. 2010. *Artificial Intelligence A Modern Approach Third Edition*.

Russom, Philip. 2011. “Big Data Analytics.” *TDWI Best Practices Report* 1–40.

Sabine, Mavin; Kayode, Owa; Dirk, Steinhauer; Elkin, Castro; Graham, Herries; Robert, John; Svetan, Ratchev. 2018. "Optimised – developing a state of the art system for production planning for industry 4.0 in the construction industry using simulation-based optimisation" T*he 25th International Conference on Transdisciplinary Engineering (TE2018)*. Modena.

Schroeder, Greyce N., Charles Steinmetz, Carlos E. Pereira, and Danubia B. Espindola. 2016. “Digital Twin Data Modeling with AutomationML and a Communication Methodology for Data Exchange.” *IFAC-PapersOnLine* 49(30):12–17.

Schuh, Günther, Christina Reuter, Jan Philipp Prote, Felix Brambring, and Julian Ays. 2017. “Increasing Data Integrity for Improving Decision Making in Production Planning and Control.” *CIRP Annals - Manufacturing Technology* 66(1):425–28.

Seitz, Kai Frederic and Peter Nyhuis. 2015. “Cyber-Physical Production Systems Combined with Logistic Models-a Learning Factory Concept for an Improved Production Planning and Control.” *Procedia CIRP* 32(Clf):92–97.

Shahzad, Atif and Nasser Mebarki. 2012. “Data Mining Based Job Dispatching Using Hybrid Simulation-Optimization Approach for Shop Scheduling Problem.” *Engineering Applications of Artificial Intelligence* 25(6):1173–81.

Shukla, Nagesh, Manoj Kumar Tiwari, and Ghassan Beydoun. 2019. “Next Generation Smart Manufacturing and Service Systems Using Big Data Analytics.” *Computers and Industrial Engineering* 128(December 2018):905–10.

Singhal, Kalyan, Feng Qi, and Ram Ganeshan. 2018. “Special Issue on Perspectives on Big Data.” *Production and Operations Management* 27(9).

Subramaniyana, Mukund; Skoogha, Anders; Salomonssonb, Hans; Bangaloreb, Pramod; Bokrantza, Jon. 2018. "A data-driven algorithm to predict throughput bottlenecks in a production system based on active periods of the machines" *Computers and Industrial Engineering*. Vol. 125. pp. 533-544.

Takeda-Berger, S. L., Frazzon, E. M., Broda, E., & Freitag, M. (2020). Machine learning in production scheduling: an overview of the academic literature. International Conference on Dynamics in Logistics (pp. 409-419). Springer.

Trstenjak, Maja and Predrag Cosic. 2017. “Process Planning in Industry 4.0 Environment.” Procedia Manufacturing 11(June):1744–50.

Uhlemann, Thomas H. J., Christoph Schock, Christian Lehmann, Stefan Freiberger, and Rolf Steinhilper. 2017. “The Digital Twin: Demonstrating the Potential of Real Time Data Acquisition in Production Systems.” *Procedia Manufacturing* 9:113–20.

Uhlmann, Iracyanne Retto and Enzo Morosini Frazzon. 2018. “Production Rescheduling Review: Opportunities for Industrial Integration and Practical Applications.” *Journal of Manufacturing Systems* 49(September):186–93.

Van der Aalst, Wil. 2016. “Process Mining: Data Science in Action.” *Process Mining: Data Science in Action* (April 2014):1–467.

Vazan, Pavel, Pavol Tanuska, and Michal Kebisek. 2011. “The Data Mining Usage in Production System Management.” *International Journal of Mechanical, Aeroespace, Industrial, Mechatronic and Manufacturing Engineering* 5(5):688–92.

Wang, Kesheng. 2007. “Applying Data Mining to Manufacturing: The Nature and Implications.” *Journal of Intelligent Manufacturing* 18(4):487–95.

Wang, Yan-hong, Ye-hong Zhang, Yi-hao Yu, and Cong-yi Zhang. 2014. “Data Mining Based Approach for Jobshop Scheduling.” *Proceedings of 2013 4th International Asia Conference on Industrial Engineering and Management Innovation (IEMI2013)* 761–71.

Wauters, T., Verbeeck, K., Verstraete, P., Berghe, G. V., Causmaecker, & D., P. (2011). Real-world production scheduling for the food industry: An integrated approach. Eng. Appl. Artif. Intell., 25, 222-228.

Woo, Jungyub; Shin, Seung-Jun; Seo, Wonchul; Meilanitasari, Prita. 2018. "Developing a big data analytics platform for manufacturing systems: architecture, method, and implementation" *The International Journal of Advanced Manufacturing Technology*. 9-12 : Vol. 99. pp. 2193–2217.

Wuest, Thorsten, Daniel Weimer, Christopher Irgens, and Klaus Dieter Thoben. 2016. “Machine Learning in Manufacturing: Advantages, Challenges, and Applications.” *Production and Manufacturing Research* 4(1):23–45.

Yahouni, Z., Ladj, A., Belkadi, F., Meski, O., & Ritou, M. (2021). A smart reporting framework as an application of multi-agent system in machining industry. International Journal of Computer Integrated Manufacturing , 34(5).

Zahmani M. Habib et Atmani B. 2019. "A Data Mining Based Dispatching Rules Selection System for the Job Shop Scheduling Problem" *Journal of Advanced Manufacturing Systems*. 01 : Vol. 18. pp. 35-56.

Zhong, Ray Y., George Q. Huang, Q. Y. Dai, and T. Zhang. 2014. “Mining SOTs and Dispatching Rules from RFID-Enabled Real-Time Shopfloor Production Data.” *Journal of Intelligent Manufacturing* 25(4):825–43.

Zhong R. Y. et Xu C. 2015. "A job-shop scheduling model with real-time feedback for physical internet-based manufacturing shopfloor" *IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*. Beijing. pp. 638-641.

Zhong Ray Y. 2018. "Analysis of RFID datasets for smart manufacturing shop floors" *IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*. Zhuhai. pp. 1-4.

Zhu, Xuechu, Fei Qiao, and Qiushi Cao. 2017. “Industrial Big Data-Based Scheduling Modeling Framework for Complex Manufacturing System.” *Advances in Mechanical Engineering* 9(8):1–12.