Data Analytics in Industry 4.0: A Survey

Lian Duan¹ · Li Da Xu²

Accepted: 17 August 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Industry 4.0 is the fourth industrial revolution for decentralized production through shared facilities to achieve on-demand manufacturing and resource efficiency. It evolves from Industry 3.0 which focuses on routine operation. Data analytics is the set of techniques focus on gain actionable insight to make smart decisions from a massive amount of data. As the performance of routine operation can be improved by smart decisions and smart decisions need the support from routine operation to collect relevant data, there is an increasing amount of research effort in the merge between Industry 4.0 and data analytics. To better understand current research efforts, hot topics, and tending topics on this critical intersection, the basic concepts in Industry 4.0 and data analytics are introduced first. Then the merge between them is decomposed into three components: industry sectors, cyber-physical systems, and analytic methods. Joint research efforts on different intersections with different components are studied and discussed. Finally, a systematic literature review on the interaction between Industry 4.0 and data analytics is conducted to understand the existing research focus and trend.

Keywords Industry 4.0 · Data analytics · Big data · Manufacturing · Cyber-physical system · Internet of things · Cloud computing · Digital twin · 5G · Blockchain

1 Introduction

In 2011, the German government supported an advanced computerization project in manufacturing (GFMER, 2011), where the term “Industry 4.0” was coined. Since then, Industry 4.0 is widely used for the fourth industrial revolution. It provides personalized products or services with flexible processes through cyber-physical systems which bridge the physical and digital world. After mechanization through water and steam power (Industry 1.0), mass production and assembly lines with electricity (Industry 2.0), and digitalization and automation of one set of devices in the same physical location (Industry 3.0), Industry 4.0 is the optimization on the computerization of Industry 3.0 to cooperate devices and facilities all over the world. It aims at decentralized production through shared facilities in an integrated global industrial system for

on-demand manufacturing to achieve personalization and resource efficiency enhanced by data analytics in smart and autonomous systems.

Industry 4.0 has a profound impact on the entire industry. In manufacturing, cyber-physical systems in Industry 4.0 connect physical devices with digital systems, which enable computers to automatically configure and dynamically adjust facilities to meet production plans and minimize human interventions in the entire production process. With cloud computing – one widely adopted technique nowadays, most companies do not need to maintain their physical IT infrastructure for data storage and computing anymore. Instead, they simply use computing services from cloud computing providers, such as Amazon, Microsoft, and Google, over the Internet to lower their operating costs and dynamically scale computing resources with their changing business needs. In Industry 4.0, a similar concept of cloud factories will be prevalent (Brettel et al., 2014; Fisher et al., 2018). Specialized companies, like the role of Amazon, Microsoft, and Google in “cloud computing”, will maintain a pool of manufacturing facilities, and typical manufacturers will pay these specialized companies based on their metered usage. It allows any manufacturer to use more facilities in its peak season and release unnecessary facilities to the cloud for other manufacturers to use in its off-season. Meanwhile, specialized companies offering cloud

✉ Lian Duan
Lian.Duan@hofstra.edu

¹ Department of Information Systems and Business Analytics, Hofstra University, Hempstead, NY, USA

² Department of Information Technology and Decision Sciences, Old Dominion University, Norfolk, VA, USA

factory services can hire technical teams to maintain their physical facilities more cost-effective due to their economies of scale. In addition, consumers can get more personalized products instead of choosing from several pre-defined models because Industry 4.0 allows dynamical reconfigurations in manufacturing systems based on customer needs (Ganschar et al., 2013). Industry 4.0 makes companies focus more on customer needs with less concern about whether having desired physical facilities to meet customer needs. It is particularly beneficial to small and medium size companies which only have very limited physical facilities nowadays.

Although most current research works in Industry 4.0 focus on improving adaptability, resiliency, scalability, and security of physical facilities through feedback loops with embedded sensors, processors, and actuators that are controlled and monitored by computers, there is an increasing interests and works on how to make factories more efficient and productive based on collected data, where data analytics play a very important role (Hunter et al., 2013; Kabugo et al., 2020; Tang et al., 2010). For example, with the tremendous monitoring data during the manufacturing process in Industry 4.0, it can be utilized to diagnose potential machine failures and optimize maintenance operations (Bagheri et al., 2011). Besides improving manufacturing processes, data analytics can also be used to improve the entire business cycle, including gathering natural resources, producing components, assembling products, delivering products to customers, and managing customer relationships (Xu, 2007). For example, whenever unexpected events occur, such as a shipment delay due to the Suez Canal blockage, connected supply chain systems in Industry 4.0 can proactively adjust manufacturing and delivery plans. Ultimately, the network of these collaborating facilities through cyber-physical systems together with dynamic smart decisions through data analytics will unleash the full power of Industry 4.0.

The rest of the paper is organized as follows. Section 2 introduces basic concepts in Industry 4.0 and data analytics. Then the interdisciplinary research between data analytics and industry 4.0 is discussed and analyzed with a systematic literature review in Section 3. Finally, the conclusion is drawn, and the future research direction is discussed in Section 4.

2 Related Concepts

In this section, the basic concepts related to Industry 4.0 and data analytics are introduced.

2.1 Industry 4.0 Techniques

The most critical component of Industry 4.0 is Cyber Physical System (CPS). The term “Cyber Physical System” was coined by Helen Gill at the NSF in the US to encourage research on

interconnected and integrated physical and computational components (Lee, 2015). Although CPS and Industry 4.0 are used interchangeably in many cases, CPS is the core component of Industry 4.0 to improve adaptability, resiliency, scalability, and security of physical facilities through feedback loops with embedded sensors, processors, and actuators that are controlled and monitored by computers, while Industry 4.0 is the entire business cycle of using CPS to dynamically adjust facilities and operation plans to meet customer needs and minimize human interventions (Xu et al., 2018). The hierarchy of related concepts in Industry 4.0 is presented in Fig. 1, and CPS and Industry 4.0 applications in different industry sectors are discussed in detail in the following.

2.1.1 Cyber-Physical Systems

Since Industry 4.0 is designed for the entire business cycle, including gathering natural resource, producing components, assembling products, delivering products to customers, and managing customer relationships, its CPS requires all the related components to be connected over the network for a higher level of automation, and allows dynamic connection and communication among components. Research efforts in CPS can be categorized into architecture, integration, and communication (Chen, 2017a). If a CPS is poorly designed, one component failure could lead to a cascading failure (Anand et al., 2006). The past research on system science has a crucial contribution to the design of the overwhelming complex systems in CPS (Xu, 2020). Therefore, there are unique requirements on the security, scalability, and reliability of CPSs to take into considerations when designing complex CPS.

In Industry 4.0, traditionally closed monitor and control systems in Industry 3.0 are modified to be adaptive and accessible over the Internet. Without an additional appropriate safeguard mechanism, a CPS can be attacked and malfunction. For example, a deliberate attack on an Australian sewage treatment system caused one million liters of untreated sewage to be released into local rivers over three months (Slay & Miller, 2007). To protect against different types of cybersecurity attacks, different methods are developed. For the attacks on the communication phase, such as packet injection or sniffing, cryptographic methods together with checksum mechanisms can protect our systems (Essa et al., 2018). Levitin et al. (2021) studied the balance between the task completion probability and the data theft success probability of co-resident attacks, where attackers can steal users' data through co-residing their virtual machines on the same physical server. Zhu and Basar (2011) studied a cross-layer security model for the trade-off between system security and accessibility.

A typical CPS has many sensors and actuators to collect data and is dynamically changing with its growing business,

Fig. 1 The hierarchy of related concepts in Industry 4.0

which undergoes scalability challenges. Sanislav et al. (2017) studied the integration between cloud computing and CPS. The entire architecture has three layers: the sensing and actuating layer, the network layer, and the processing and application layer. With different agents in the processing and application layer, including collector agents, manager agents, ontology agents, processing data agents, negotiation agents, and diagnostic agents, the proposed system can improve its scalability with such a decentralized architecture. Stojmenovic (2014) proposed a localized cooperative access stabilization method which allows machine-to-machine devices to collaborate with each other in addition to gateways. The system can be scaled to trillions of machine-to-machine devices without sacrificing its quality of services. Canizo et al. (2019) studied the integration among big data techniques, cloud computing, and CPSs for real-time monitoring. Apache Flume and Kafka are used for data collection, Apache Spark is used for data processing, and Apache Zookeeper is used for cloud resource management.

As control systems in CPSs are more decentralized to meet the need for scalability, it is also critically important to ensure its reliability and resiliency. Zhang et al. (2016) studied a k-reliability model in different device connecting topologies to prevent cascading failures in the whole system better. Yang et al. (2021) evaluated CPS reliability in the event of communication failures. The communication failures are based on the instantaneous availability model using a Markov process with three system states, including working state, repairing state, and delay-repairing state. To improve reliability, the four-order Runge-Kutta method is used for optimal solutions. Bajaj et al. (2015) formulated the system reliability as an optimization problem of minimizing a cost function within the constraints of desired reliability. Two approaches, including integer-linear programming modulo reliability and integer-linear programming with approximate reliability, are studied for finding feasible solutions within a given amount of time.

2.1.2 Industry Sector

Most research in Industry 4.0 is related to the industry sector of manufacturing. However, there are also studies in the context of other industry sectors (Chen, 2017b; Li, 2020), including agriculture, healthcare, education, transportation, and energy. Frank et al. (2019) studied the adoption patterns of Industry 4.0 technologies in 92 manufacturing firms. Related Industry 4.0 technologies are categorized as front-end technologies and base technologies. The front-end technologies include smart manufacturing, smart products, smart supply chain, and smart working. The base technologies have Internet of Things (IoT), cloud services, and big data analytics. Its study shows smart manufacturing plays a central role in front-end technologies, and IoT and cloud services are more widely implemented than big data analytics in base technologies. Sharma et al. (2021) studied the integration of CPS, internet of things, cloud computing, and big data to improve agricultural supply chains and boost productivity. Lei et al. (2018) studied the interrelated work on security and reliability in electric power systems. Security issues focus on specific cyber intrusion mechanisms, while reliability issues are more associated with the intrinsic structure and topology of the electric power systems. Pace et al. (2018) used an edge-computing based method in the healthcare sector to support time-dependent applications. It consists of a mobile client module and a performing edge gateway supporting multi-radio, and multi-technology communication to collect and process data in different scenarios. It also exploits cloud platforms to achieve a higher flexible, robust, and adaptive service level.

2.2 Data Analytics Techniques

With the development of faster Internet, larger storage, and new sensor techniques, companies are collecting more and more data measured by Terabytes in 2005, Petabytes in 2010, Exabyte in 2015, and Zettabytes in 2020. There is an

increasing need to gain actionable insight from such a massive amount of data. Data analytics techniques can be categorized into system infrastructure and analytic methods. System infrastructure focuses on making data ready for analysis, while analytic methods focus on how to gain actionable insight from data. The hierarchy of related concepts in data analytics is presented in Fig. 2.

2.2.1 System Infrastructure

To make data ready for analysis, system infrastructure handles how to capture, transfer, store, and compute data.

Capture Different data capture techniques are developed when data sources are different. The most widely-used technique is graphical user interfaces (GUI), which allows users to interact with electronic devices through graphical icons, to collect data from humans (Palani, 2020). Besides the direct input from human, companies offering content services, such as Nasdaq, Twitter, and Google, might also provide their application programming interface (API) for other users to directly get their data on an agreed data format, such as CVS, JSON, and XML. Slightly different from getting data on an agreed data format from APIs, the World Wide Web (WWW) has abundant useful information, including reviews, tweets, news, and other social media information. Because that information is created for humans to read instead of computers, it takes extra effort to create specialized web crawlers (Kumar et al., 2017) to extract useful information from WWW and save them in an appropriate format for computers to read. In addition, sensors (Lopez et al., 2017) are a broad type of data capture technique to collect data from objects, such as chemical, light, temperature, sound, video, and GPS locations.

Network In many cases, the data capture devices and data storage devices are not in the same location. It requires appropriate network infrastructure to transfer data. There are two types of networks: wired network and wireless network. Wired networks offer faster speed but less flexible for connection. In a wired network (Hariharan et al., 2018), there are three types of physical cables: twisted pair, coaxial cable, and fiber optic. Twisted pairs offer the cheapest connection within a 100-m distance. It is widely used in our office and home setting and its current highest bandwidth is 10Gbps. Coaxial cables offer a longer length within 500 m than twisted pairs, but their available bandwidths are 100Mbps and 1Gbps. Fiber optic is the most expensive wired medium but offer the highest rate of data transmission up to 200 Gbps within a much longer distance up to 80 km. In a wireless network (Elhence et al., 2020), three common techniques are cellular, Wi-Fi, and Bluetooth. Cellular can transfer data over 20 km, Wi-Fi is only for 50 m, and Bluetooth is about 10 m.

Storage The most straightforward way of saving data is files. However, if data is dynamically changing, any modification causes the Input & Output (IO) operation on the entire file. To solve this issue, traditional relational database systems (Codd, 1970) were designed to break the entire data into many inter-related tables. In a tradition relation database, each table is corresponding to one self-contained entity, and tables are related to each other through primary key and foreign key pairs. Whenever a portion of data is changed, only the related tables instead of the entire database are updated. Whenever a portion of data is needed, only the related tables are joined together through primary key and foreign key pairs. Although traditional relational database systems overcome the problems of file systems in terms of data storage, they have their own limitation, such as limited storage capacity, single node

Fig. 2 The hierarchy of related concepts in data analytics

failure, low performance on join and summarized data, and difficulty in collaboration across organizations. Therefore, several new database techniques are developed to overcome some of the above issues. For example, Cassandra (Wahid & Kashyap, 2019) is designed to spread data over a cluster of computers to increase the storage capacity with a certain level of redundancy to handle single node failure issues. A data warehouse (Visscher et al., 2017) is designed to save historical data in a summarized cube with low granularity. Data warehouse can speed up the data aggregation in higher granularity but lose details at the transaction level. MongoDB (Jose & Abraham, 2017) is created to save tightly connected entities into one file instead of keeping them into separate tables. If tightly connected entities are always shown together, such as articles and related reviews, such integration can save many joint operations in the traditional relational database. All the above database techniques are developed for the collaboration within one company or organization. However, no company can just operate by itself. Whenever this is a collaboration across companies to save the collaboration data in a central database, all the companies want to control that database because each cannot fully trust the others. Although a third-party oversight can solve this issue, it adds additional costs. Blockchain (Zheng et al., 2018) is created to manage such a publicly distributed database system in a peer-to-peer network, where each peer collectively follows a protocol to communicate and validate new records.

Computing When the needed computing power is beyond the capacity of a single normal computer, high-performance computing techniques are developed to handle such a situation. Traditional supercomputers include many CPUs on a large shared-memory to coordinate computing-intensive tasks within one computing machine. However, such a design is not very scalable, and more and more applications start to use a cluster of cheap personal computers coordinated through a Hadoop or Spark system (Zhou et al., 2018). The Hadoop system is developed by Google in 2006 to handle expected hardware failure and workload rebalance in a computing cluster. The core operation is MapReduce which generates key-value pairs distributed in the cluster and then summarizes values with the same key. Each round of MapReduce involves reading data from hard drives and saving processed data into hard drives, which is time-consuming. To speed-up the Hadoop system, the sibling system Spark is developed to finish intermediate MapReduce operation in the memory to avoid the IO cost in hard drives. With the Spark system, only reading the raw data in the beginning and saving the final data in the end involve IO operation in hard drives for each application.

Besides the systems to connect computing nodes, computing units can also be specially designed for different tasks. In the normal setting, CPUs are created to handle complicated

logic operations. When generating output to a display device, the output is represented by a matrix, and the operation is mainly related to matrix calculations. To speed up matrix calculations, a graphics processing unit (GPU) is designed with way more arithmetic logic units (ALU) but fewer control units than CPU (Sahal et al., 2020). In data analytics, most data are shaped as tables and transformed into matrices to compute. Although GPU is originally designed for image processing, they are also widely used in data analytics since both image processing and data analytics involve intensive matrix calculations. Besides GPU, field-programmable gate arrays (FPGA) are another computing unit to speed up data analytics (Valente et al., 2019). FPGA is an integrated circuit that consists of internal hardware blocks with user-programmable interconnects to customize operation for a specific application. Different from GPU to excel at parallel processing for matrix, FPGAs excel in the application with low latency and small batch sizes, such as speech recognition and natural language processing, for customized hardware with integrated data analytic methods.

2.2.2 Analytics Method

The data analytics methods are classified into three categories: (1) descriptive analytics, (2) predictive analytics, and (3) prescriptive analytics. This taxonomy is based on how and when these methods are used. Descriptive analytic methods are the first step focusing on summarizing historical patterns from data. With summarized historical patterns by descriptive analytic methods, predictive analytic methods can use these patterns to predict what will happen in the future based on the naïve assumption that what happened in the past will happen in a similar way in the future. With predictions made by predictive analytic methods, prescriptive analytic methods focus on making the optimal plan to meet the forecasted future needs.

Descriptive Analytics The most widely-used descriptive analytic methods are the traditional statistical measures, such as mean, median, standard deviation, and skewness. The major issue of these traditional statistical measures is the impractical assumption that an entire dataset is a homogeneous group. Therefore, association rules and clustering are developed to search for hidden interesting patterns in sub-populations.

Association rules search for meaningful connections among objects. Then their patterns can be used to select correct interventions to get desired outcomes. For example, technical analysis in finance searches for connections between signals calculated from historical data and their future stock price movement. Doctors diagnose patients' health conditions with the connection between diseases and symptoms on existing patient data. Traditional methods in this area have Pearson correlation coefficient (Pearson, 1895), Chi-square

statistics (Elderton, 1902), and regression analysis (Neter et al., 1996). However, these methods are designed to handle thousands of manually cleaned records with tens of preselected variables. It has severe challenges in big data with trillion of records and millions of variables. To solve the efficiency problem in big data, the classical Apriori method (Agrawal & Srikant, 1994) utilizes a downward-closed property of co-occurrence to prune the exponential search space. In this direction, other techniques like FP-Tree (Han et al., 2000) and ECLAT (Zaki, 2000) were proposed for faster counting. However, the main problem of this type is that co-occurrence is a sub-optimal measure for correlation. Therefore, there is a stream of work to combine monotonic (or downward-closed) properties with more effective correlation functions to speed up the search for more accurate patterns (Duan & Street, 2009; Xiong et al., 2006).

Because not every object has a correlation with others, discovered association rule patterns only cover a small portion of all the objects in a dataset. Different from association rules, clustering assigns each object to groups where objects have similarities. A cluster can be a group of customers with similar interests or products with similar functions. Then clustering results can be used to balance between one-size-fit-all (efficiency) and specialized treatment (effectiveness). Three typical clustering methods are centroid-based, density-based, and grid-based. The centroid-based clustering (MacQueen, 1967) generates a central vector and assigns objects to their nearest cluster center, where each cluster has a spherical shape. However, the spherical shape assumption is not true in some applications, such as land use, tumor shape, and customer purchase. To handle this issue, density-based clustering (Ester et al., 1996) is developed. It starts with a random object and checks its neighbors. If the neighborhood density of the current object is like the neighborhood density of one neighbor of the current object, the current object and that neighbor are merged into the same cluster until this cluster cannot be expanded anymore. As both centroid-based and density-based clustering needs to process each record one by one, grid-based clustering (Wu & Wilamowski, 2016) is proposed to divide the entire feature space into grids and merge objects in the same grid altogether instead of one by one.

Predictive Analytics Predictive analytic methods utilize historical patterns to forecast what will happen in the future based on the assumption that historical patterns repeat themselves in a similar way in the future. A typical predictive analytic method starts a historical dataset, where one (or more) attribute is identified as target attributes and the remaining attributes are classified as normal attributes. Target attributes are typically useful future information for current operation planning, such as sales in the next month and raw material prices in the next year. Although the value of target attributes for the current data is not available to us, the value of target attributes for

historical data can be collected. Normal attributes are typically related to target attributes and instantly available for even the current data, such as sales in this month, sales in the previous month, and the number of current active customers. Then a predictive analytic method summarizes the relationship between normal attributes and target attributes in historical data. With this relationship summarized from historical data and the value of normal attributes for the current data, the target attribute of the current data can be predicted. The six popular types of predictive analytic methods are (1) regression (Neter et al., 1996), (2) Bayesian statistics (Domingos & Pazzani, 1997), (3) decision tree (Quinlan, 1986), (4) neural network (Demuth et al., 2014), and (5) support vector machine (Suykens & Vandewalle, 1999). Regression methods search for a linear relationship between target attributes and normal attributes. If there is no linear relationship with raw data, different transformations might be applied in logistic regression, LOESS, and LOWESS. Bayesian statistics use the Bayes' theorem to build the relationship between target attributes and normal attributes. Naïve Bayesian is the most straightforward method assuming each attribute is independent of the other. A decision tree uses a utility function to divide current data into different branches iteratively to improve the degree of purity in each division. Neural network methods construct a network with input nodes, layers of hidden nodes, and output nodes to build a complex high-dimensional relationship among data. Support vector machine searches for a linear hyperplane to separate two classes.

Prescriptive Analytics With predictions made by predictive analytics, prescriptive analytics search for the optimal plan to meet the predicted future needs. This type of work is extensively studied in operation research, a sub-field of applied mathematics. In the real world, management sciences, decision sciences, and operation research are used interchangeably. It is typically related to maximizing (or minimizing) a meaningful positive (or negative) objective, such as profit, performance, loss, and cost, within a given set of constraints, such as budget, manpower, and time.

Techniques in this area can be classified into two types: (1) convex programming (Grant et al., 2006) and (2) heuristic search (Bonet & Geffner, 2001). Convex programming is related to the type of problems with a convex structure whose global optimal value can be approached in theory. It includes linear programming, second order cone programming, semidefinite programming, and geometric programming. As not all optimization problems have a convex structure, heuristic search is another type of method to search for a sub-optimal solution in this situation. Typical heuristic search methods are simulated annealing, genetic algorithm, and tabu search.

3 Current Interdisciplinary Research between Data Analytics and Industry 4.0

Industry 4.0 roots from Industry 3.0 which focuses on routine operations, while data analytics roots from statistics and computer sciences which focus on business insight and smart decision. On the one hand, making a smart decision based on useful business insight can help to optimize routine operations. On the other hand, searching for useful business insight needs support from routine operations to collect relevant data. Therefore, there is an increasing trend in merging between Industry 4.0 and data analytics. One example is the machine condition monitoring system for condition-based maintenance, which is very effective and cost-efficient to avoid catastrophic failure of machines. Bagheri et al. (2011) implemented an acoustic condition monitoring system for gear-boxes. Worn tooth face gear and broken tooth gear are two common faults in gear-sets. When it happens, gears will generate different acoustic signals. By adding an additional layer of sensors to collect acoustic signals and applying predictive analytics to classify acoustic signals into normal or abnormal types, Industry 4.0 provides data collection and computing devices for data analytics, and data analytics in return help to improve condition-based maintenance in Industry 4.0. Another example is online shopping recommender systems. Traditionally, online shopping companies like Amazon use association rules in descriptive analytics search for products purchased together from transaction data. When customers are purchasing items on online shopping websites, these co-purchase patterns can be used to recommend relevant products. However, transaction data only contain the patterns for complementary products like TV wall mount brackets for TVs, but not for substitute products like Samsung TVs for LG TVs, because customers will not buy substitute products together. To provide useful substitute product information, it takes extra effort for recommender systems to collect customers' historical web browsing data. Very often, customers will compare several substitute products before making their final purchase decisions. Therefore, a more effective recommender system needs its information system to collect extra web browsing data and then provide useful substitute product information displayed on its online shopping website. In addition, Kabugo et al. (2020) integrate the data analytics platform with industrial IoT platforms to enable data-driven soft sensors to predict syngas heating value and hot flue gas temperature in a waste-to-energy plant. Lee et al. (2019) conducted a survey study on the quality management ecosystem for predictive maintenance in Industry 4.0. The study shows an effective quality management ecosystem leverages big data analytics, smart sensors, and platform construction, and is the product of the organizational culture that nurtures collaborative efforts of all stakeholders, sharing of information, and co-creation of shared goals.

With more and more data analytic components integrated into Industry 4.0 to improve routine operation, the CPS in Industry 4.0 needs to offer related computing infrastructure to data analytic components. Faheem et al. (2021) proposed a novel cross-layer data collection approach for active monitoring and control of manufacturing processes in Industry 4.0. Their method exploited the multi-channel and multi-radio architecture of sensor networks by dynamically switching among different frequency bands. It can handle the harsh nature of indoor industrial environments with high noise, signal fading, multipath effects, and heat and electromagnetic interference with guarantee QoS on high throughput, low packet loss, and low latency. Therefore, the CPS in Industry 4.0 and the system infrastructure in data analytics become one merged piece. In other words, the merging between Industry 4.0 and data analytics has three interwoven components: industry sector, cyber-physical system, and analytic method shown in Fig. 3. In the following, the interactions among different components are discussed. In addition, emerging hot techniques developed for handling data in more efficient or effective ways (Zhang & Chen, 2020), such as 5G network, big data, blockchain, cloud computing, deep learning, Internet of Things (IoT), and quantum computing, are also introduced together with the discussion on how they are related to the merging between Industry 4.0 and data analytics.

3.1 Different Intersection Regions

3.1.1 Industry Sector with Cyber-Physical System and Analytic Methods

CPSs root from Industry 3.0 designed for routine operations. In different industry sectors, their business workflows are different. Therefore, their corresponding CPSs must be very different. CPSs in healthcare (Dey et al., 2018) focus on integrating medical devices in hospitals over the network to improve overall healthcare quality, while CPSs in manufacturing (Adamson et al., 2017) focus on the integration of manufacturing facilities to reduce production costs. Although all the CPSs have common requirements on security, resiliency, and reliability, CPSs in different sectors have their own special requirements. For example, medical CPSs need an additional

Fig. 3 Three components between Industry 4.0 and data analytics

layer to protect patients' privacy data. In addition, depending on the application requirements, different system architectures are needed. For example, in the application that needs real-time with minimal latency with users from everywhere in the world like online shopping, cloud computing, and clouding manufacturing, it needs an additional load balancing layer (Li et al., 2017) among application, database, and cache servers. For special types of data that sensors can only collect on the field, such as temperature, wind speed, and sound, it takes communication distance, data quality, battery power, and all the other related factors in designing the sensor network topology and its communication and information aggregation strategy (Faheem & Gungor, 2018).

Besides the impact on CPSs, industry sectors also have an impact on data analytics, regarding feature selection and limitations of different analytic methods. Although many techniques are developed to handle so-called "big data", the so-called "big data" is still a marginal portion of all the data in the entire universe. Without utilizing domain knowledge in each industry sector, analytic components will waste valuable resources on irrelevant information. For example, to diagnose Covid-19, there is unnecessary to analyze all the possible chemical components in human blood. Domain knowledge can help us to focus on relevant but already big enough data for analysis. In addition, domain knowledge can also help to construct more relevant features to gain better analytic performance. In finance, technical indicators (Hu et al., 2021), such as relative strength index, Bollinger bands, and moving average convergence divergence, are calculated with historical stock price to forecast future stock price movements. In healthcare, body mass index (Ajala et al., 2017), calculated with weight and height, is used in applications where patients' health is affected by overweight conditions. In facial recognition, region features (Karczmarek et al., 2018), such as face oval, upper lip, lower lip, eyebrow, eye, cheek, nose bridge, and nose bottom, are used for better identification. Besides the above feature selection, different industry sectors might favor different analytic methods due to different limitations. For example, the healthcare sector favors those interpretable models, such as regression, decision tree, and Bayesian network, to understand how different demographic features, such as race, age, gender, and lifestyle, are related to different disease risks. On the contrary, facial identification and speech recognition favor more accurate deep learning methods whose models cannot be interpreted by humans.

3.1.2 Cyber-Physical System and Analytic Methods

There are two types of interactions between CPS and analytic methods. On the one hand, due to the limitation of current data analytic methods (or CPSs), the related CPS (or data analytic method) needs to be modified accordingly to make it work.

On the other hand, data analytics can help to improve the performance of CPS.

Modification for Current Method (or CPS) Limitations For some big data analytic applications like predictive maintenance for millions of facilities with deep learning methods (Canizo et al., 2017), they need a robust system to handle Petabyte's data with intensive computing power. A computing cluster equipped with GPUs, high-speed local network, and a Hadoop or Spark coordination system is needed to strike a balance between desired computing resources and costs. Mitra (2021) proposed a cellular-automata-based MapReduce model to facilitate big data process with low energy consumption in Industry 4.0. For some CPSs in multinational corporations, different portions of raw data are stored in each local branch for load balancing in its instant services. While it costs too much to aggregate raw data from local branches into a centralized place for data analysis, many analytical methods, including K-means, decision tree, and deep learning, have their specially designed distributed version to handle such a situation (Balcan et al., 2013; Ben-Nun & Hoefler, 2019; Bhaduri et al., 2008).

Intelligence Components in Cyber-Physical System Data analytics has been used to improve CPS's self-aware and self-maintained capabilities. Many existing manufacturing plans are made under the assumption of continuous facility readiness and consistent performance, which is often violated in practical manufacturing. To diagnose conditions on different machines, intrusive examination costs a lot on human labors and downtime on facilities. Alternatively, some indirect but related information, such as vibration, acoustic emission, lubrication oil, and particles, can be collected by different sensors in a non-intrusive way. This information can be used to predict potential failures (Canizo et al., 2017). Traditional CPSs make optimal operation plans with all the information in a central server. However, more and more CPSs are designed to be scalable to new devices with their growing business. Therefore, some self-optimizing autonomic strategies are desired for facilities to follow operation plans made based on local information but still globally optimal in dynamically changing environmental conditions and demands (Berger et al., 2019). For example, Wang et al. (2016) presented a smart factory framework to include network, cloud, supervisory controller, and smart shop-floor objects which is classified as various types of agents. Then a self-organized system under an intelligent negotiation mechanism is designed by leveraging big-data-based feedback and coordination for agents. Tang et al. (2017) studied a cloud-assisted self-organized architecture with smart agents and cloud to communicate and negotiate through networks. Decision-making agents use pre-defined knowledge base ontology for dynamic reconfiguration in a collaborative way to achieve agility and

flexibility. In addition, the agents' interactions are modeled to assign agents hierarchically to reduce the coordination complexity.

In addition to improving advanced self-aware and self-maintained capabilities of CPS, data analytics can also be applied to improve the fundamental requirements of CPS on its security, scalability, and reliability. Gokarn et al. (2017) applied an anomaly detection system using behavior analysis to enhance the security of its CPS. The anomaly detection system uses Kalman Filter to construct a system behavior model to predict estimated states. If the real time data is different from predicted states, it raises an alarm of abnormal events for further investigation. Qi et al. (2021) designed an AI-assisted lightweight authentication protocol for real-time access using a Chebyshev map in medical CPSs. The new system can overcome password guessing or smartcard lost attacks on traditional authentication schemes, such as password or smartcard, to validate users. Wang (2020) proposed a system to evaluate the reliability of CPS based on different analytic methods. First, clustering methods are used to classify the importance of nodes in complex networks. Second, a prediction model is constructed to evaluate CPS reliability based on the importance of different nodes. Finally, an online queuing algorithm is integrated to assess CPS reliability in real-time.

3.2 Relationship to Emerging Hot Techniques

In this section, emerging hot techniques in the alphabetic order are introduced and discussed on how they are related to the merging between Industry 4.0 and data analytics.

3.2.1 5G Network

5G is the fifth-generation technology standard for cellular networks. It is related to the network component. It is expected to have higher speed and lower latency compared to 4G technology in a cellular network. However, the discussion on 5G hype is exaggerated in social media for a "life-changing" technology in every aspect of lives. 5G is related to the cellular network, and cellular network is one type of wireless connection solution. After mobile devices communicate with cellular towers through 5G, cellular towers still need to pass information through a wired network. Therefore, advances in the wired network are also crucial to the performance of 5G. In addition, 5G only has advantages over Wi-Fi for the wireless communication over a longer distance. If applications only require short-distance wireless connection or wired connection, 5G has no impact on them at all. For example, there are some reports on remote surgery over the 5G network (Laaki et al., 2019). Such an application is only a test on 5G performance but does not really need a 5G network in most cases. Remote operating rooms can achieve better performance with

just wired connection as those remote operating rooms are not expected to move around very often. In the special events of a battlefield or natural disaster with temporary remote operating rooms, the assumption of normal function on 5G cellular towers is still not valid. Nevertheless, 5G has meaningful applications with high speed and low latency wireless connection requirements over a long distance according to a survey study (Li et al., 2018) on how 5G impacts industry 4.0 with IoT. For example, 5G is critically important for autonomous cars. In autonomous driving, cars need to share a large amount of sensing data and make instant decisions to avoid collisions. In addition, due to the long moving distance, the wired or Wi-Fi connection to cars is impractical, and 5G is the only solution for it. In addition, 5G can have an important impact on remote video games and augmented reality on mobile devices which do not have powerful enough processors to handle graphic data. In addition, to overcome the disadvantage of 5G limited by cellular towers, 6G empowered by satellites is also under the development to establish full coverage of the air-space-sea-land system (Lu & Ning, 2020).

3.2.2 Big Data

Big data is the umbrella term for any technology to handle big data better. It covers every aspect of data analytics, including data capture, network, storage, computing, and analytic methods. The most-widely used characteristics of big data are the "3 Vs": volume, velocity, and variety. Any technique to handle a larger amount of data, process it faster, or robust to handle heterogeneous data is considered as a big data technique. Therefore, it covers all the emerging hot techniques discussed here. For example, 5G is one type of network technique to transfer a larger amount of data.

3.2.3 Blockchain

Blockchain is related to the data storage component, which is typically supported by database techniques. Blockchain is a growing list of records linked together through cryptographic hashes (Gorkhali et al., 2020). A peer-to-peer network manages it as a publicly distributed ledger, where peer nodes collectively adhere to a protocol with high Byzantine fault tolerance to communicate and validate new blocks. Each block contains a cryptographic hash of the previous block in a blockchain to prevent malicious insertion to change data. Once a block is created, it cannot be deleted. If the previous block updates information incorrectly, such a correction can only be finished by inserting a new block to undo the previous update instead of deleting the previous block. It allows share and update data across multi-peer organizations through a secured, authenticated, verifiable, and immutable mechanism without centralized administration. Whenever this is a collaboration across companies, if the collaboration data is saved in

a traditional database, all the companies want to control that database because each cannot fully trust the others. Blockchain offers the perfect solution in this situation to facilitate data sharing and collaboration over multi-organization. In return, blockchain impacts the merging between Industry 4.0 and data analytics in two ways. First, it offers essential data that is previously unavailable. Second, it improves data quality. New data and improved data quality can directly improve the daily operation of CPSs in Industry 4.0. In addition, new data and improved data quality can also potentially have new data analytical applications for new business gains. For example, Xu and Viriyasitavat (2019) applied blockchain with smart contract to establish the trust of process executions among IoT devices without intermediaries. Xu et al. (2021) conducted a survey study on how blockchain is used to improve IoT security in different layers, including sensor layers, network layers, and application layers, with different types of attacks, including DoS, equipment injection, falsifying, public block modifying, and time interval destruction.

3.2.4 Cloud Computing

Cloud computing is related to the data storage and computing components. It is the on-demand availability of data storage and computing power resources without direct active management by users. Its typical service models are infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). IaaS offers the low-level management of computer resources as virtual machines and allows the control of storage, network, security, location, and scaling. PaaS offers virtual machines with a pre-selected operating system and allows the installation of any software and the change of operating system settings. SaaS offers specific applications, such as an email server, analytic software, or database, to run. Cloud computing can achieve what companies want like any outsourced task. In the meanwhile, it is much cheaper due to the economics of scales and has better security and reliability with more specialized maintenance. However, privacy and confidentiality can be concerns in some situations as their downside.

3.2.5 Deep Learning

Deep learning is related to the analytic method component. It is a family of predictive analytical methods based on neural networks. It has different architectures, including graph neural network (Scarselli et al., 2008), recurrent neural network (Mikolov et al., 2010), and convolutional neural network (Lawrence et al., 1997), for applications in computer vision, speech recognition, and natural language processing with high-dimensional patterns inside. The word “deep” refers to the additional number of layers compared to traditional neural

networks. In deep learning, each layer learns to transform its input data into a slightly more abstract and composite representation to model more complex non-linear relationships in a high-dimensional space. Convolutional neural networks use a moving filter to segment images into many small images because one typical image might contain many objects. When two different images contain the same object, a convolutional neural network can map the right segments in different images to the same object. Recurrent neural networks are designed to model time (or sequence) dependent behaviors in language, financial market, and weather. Different from other neural networks to model each record independently, a recurrent neural network feeds back the output of the current record to the next record to model the sequential impact.

3.2.6 IoT

IoT is related to the data collection component. It refers to the object embedded with sensors, actuators, or software to connect and exchange data with other devices and systems over the Internet. The key technologies include identification, tracking, communication, network topology, and service management over four layers, including sensing layer, to achieve dynamical interaction among heterogeneous devices in a multitude of way (Xu et al., 2014). It roots in embedded systems in Industry 3.0 for the automation for a smaller scale among facilities in the same physical location. The traditional embedded systems are closed monitor and control systems within local area networks for facilities in the same physical location. They need to be redesigned to meet the need of Industry 4.0 for all the related components to be connected over the Internet for a higher level of automation and allows dynamic connection and communication among components. Therefore, it has more rigorous requirements on its security, scalability, and reliability. Li and Xu (2020) conducted a survey study on resource allocation on IoT. As IoT in Industry 4.0 is required to dynamically cooperate for complex goals, efficient allocation of IoT resources would improve the overall performance a lot. These resource allocations are under the constrain of energy, storage, spectrum, channel, service, bandwidth, computing power, and access and supported by different techniques, including M2M communication, edge computing, cache, wireless, and RFID.

3.2.7 Quantum Computing

Quantum computing is related to the data computing component. It refers to the exploitation of collective properties of quantum states to perform computation. As each quantum has two states at the same time, its computational power increase exponentially with the number of quanta while the computational power in our current classical computers increases linearly with the number of transistors. Therefore, it

can solve problems that need exponential time with our current classical computers. Since many of our current analytic methods are NP-complete problems for optimal solutions and NP-complete problems are a subset of exponential time problems, quantum computing can offer matching computing power to find optimal solutions for all the analytic methods with the matching number of quanta in theory.

3.3 A Systematic Literature Review

To find the research trend on the intersection between Industry 4.0 and data analytics, a systematic literature review is conducted. To understand details better, the data from Scopus (<https://www.scopus.com>) are intensively analyzed due to its high data quality with additional useful metadata. When selecting relevant literature from Scopus, the search is only limited to journal articles and conference papers, excluding reviews, books, notes and letters, because journal articles and conference papers focus on their ongoing projects and are more likely to be involved with peer reviews for quality control. During the search, documents are selected if both terms “industry 4.0” and “data analytics” are used within their titles, abstracts, or keywords. In total, 401 documents are selected. Our search was conducted on May 1st, 2021. When analyzing the trend, the data in 2021 are excluded as no complete data for 2021 are available at the time. The number of related articles in each year is shown in Fig. 4. Starting from 2015, there is a significant increasing trend on the intersection between Industry 4.0 and data analytics. At the same time, a similar search is conducted in the Google Scholar (<https://scholar.google.com>) if both terms “industry 4.0” and “data analytics” are used within their documents. Comparing to Scopus, Google Scholar has more comprehensive data but less quality control. Therefore, its

data are used as a complement. Such an increasing trend in Scopus is also supported by the Google Scholar data shown in Fig. 5. In the following, a more detailed analysis on keyword, title, and abstract is conducted.

3.3.1 Keyword Analysis

Besides the number of papers, keywords in documents are also analyzed. In the Scopus data, there are two types of keywords: author keywords and index keywords. Author keywords are chosen by the author(s) which best reflect the contents of their document in their opinion, while index keywords are chosen by content suppliers and are standardized based on publicly available vocabularies. When analyzing keywords, some keywords are merged into one for the same meaning with different spellings. For example, cyber-physical system, cyber-physical systems, cyber physical system, cyber physical systems, cyber-physical production systems, cyber-physical production system, cyber-physical, cyber-physical systems (cpss), and cyber-physical manufacturing system (cpms) are all considered as cyber-physical systems. Twenty-one keywords are manually selected based on the frequency on both author keywords and index keywords. In addition, 5G and blockchain are also selected for its future impact on this area although their frequency is low at the current moment. Among 401 collected documents, some do not have author keywords, while others do not have index keywords. To understand the difference between author keywords and index keywords, the percentage instead of frequency is presented in Table 1.

First, the top-5 terms are Industry 4.0, data analytics, big data, manufacturing, and Internet of Things in both author keywords and index keywords. Industry 4.0 and data analytics are expected to be the top-2 keywords as they are the terms used to search for relevant documents. Big data and

Fig. 4 The related articles from Scopus

Fig. 5 The related articles from Google Scholar

manufacturing are at the ranking 3rd or 4th to reveal the different focuses. Big data is related to data analytics to emphasize on the techniques to handle a large amount of data, while

Table 1 Author keywords and index keywords

Terms	Author Keywords		Index Keywords	
	Percentage	Ranking	Percentage	Ranking
5G	1.09%	20	2.01%	21
Advanced Analytics	0%	22	17.06%	9
Artificial Intelligence	3.28%	15	11.37%	14
Big Data	30.87%	3	32.78%	4
Blockchain	1.37%	19	0.33%	23
Cloud Computing	5.19%	12	6.69%	17
Cyber-physical Systems	17.76%	7	18.06%	8
Data Analytics	43.17%	2	77.59%	1
Data Mining	1.91%	17	6.35%	18
Decision Making	4.64%	14	21.74%	6
Deep Learning	2.73%	16	3.34%	20
Digital Twin	5.46%	11	6.02%	19
Embedded Systems	0%	22	19.73%	7
Industry 4.0	63.93%	1	58.86%	2
Internet of Things	23.77%	5	29.1%	5
Learning Systems	0.55%	21	11.71%	13
Machine Learning	9.29%	8	8.7%	15
Maintenance	7.92%	9	14.72%	11
Manufacturing	28.69%	4	47.83%	3
Predictive Analytics	7.1%	10	15.72%	10
Prescriptive Analytics	1.37%	18	1%	22
Smart Factory	18.03%	6	14.38%	12
Supply Chain	4.92%	13	8.7%	16

manufacturing is related to Industry 4.0 to reveal that the most existing research in Industry 4.0 is in the industry sector of manufacturing. Internet of things in the 5th ranking place is related to the intersection between cyber-physical systems in Industry 4.0 and system infrastructure in data analytics. IoT is the core component to bridge cyberspace and physical space in Industry 4.0 and collects relevant data for future analysis in data analytics.

Second, most ranking in author keywords and index keywords are consistent. However, advanced analytics, decision making, embedded systems, and learning systems are more popular in index keywords, while digital twin and smart factory are more popular in author keywords. Such differences only reflect the different preferences between authors and publishers. For example, embedded systems are ranked at the 7th place in index keywords but never selected by authors. It is a more legacy term for the same thing – “Internet of Things”, which emphasize more on their connectivity to the Internet.

Third, the trend on different keywords is also analyzed in Fig. 6. The 2015 data is removed, because there are only two related documents, which is not enough for a reliable estimation of their related percentage. For the top-5 terms – Industry 4.0, data analytics, big data, manufacturing, and Internet of Things, they are some fluctuations but generally flat over the years. Artificial intelligence starts to gain more popularity in author keywords. Still, it only reflects authors’ preference instead of the rise of research on this area because artificial intelligence, data mining, deep learning, machine learning, and predictive analytics are often used interchangeably. The term “digital twin” starts to gain its popularity since 2018, while the term “smart factory” loses its popularity over the years. Digital twin refers to real-time digital representation of a physical object or process, and it is closely related to

Fig. 6 Trend on different keywords

IoT. For example, Jiang et al. (2021) proposed a unified architecture for IoT to support an internal extension of digital twins and multi-digital twin connection, which isolates the direct access of business, strengthens the division and cooperation between local and cloud, and promotes the integration and synchronization of virtual and real. In addition, 5G and blockchain start to get more attentions since 2018.

3.3.2 Title and Abstract Analysis

To complement the analysis on keywords, the analysis on titles and abstracts are also conducted. Each time a given word appears once in a title, it gains a weight of 3. Similarly, each time a given word appears once in an abstract, it gains a weight of 1. The assignment of different weights is purely

empirical because titles highlight more relevant information than abstracts. The patterns in each year are consistent with each other and are also consistent with those in keywords in Fig. 7. Highlighted patterns are data analytics, Industry 4.0, and manufacturing. In addition, general terms like system, process, production, and technologies are also highlighted.

4 Conclusion and Future Research Direction

This paper discussed the basic concepts in Industry 4.0 and data analytics. Then the benefits of the merging on Industry 4.0 and data analytics are discussed and studied. Industry 4.0 evolves from Industry 3.0 which focuses on routine operations, while data analytics roots from statistics and computer

Fig. 6 (continued)

sciences which focuses on smart decisions. As the performance of routine operation can be improved by smart decisions and smart decisions need the support from routine operation to collect relevant data, Industry 4.0 and data analytics are perfect complements of each other. The merge between Industry 4.0 and data analytics has three interwoven components: industry sectors, CPSs, and analytic methods. Joint research on different intersections with different components are studied and discussed. In addition, a systematic literature review on the interaction between Industry 4.0 and data analytics is conducted to understand the existing research focus and trend. Besides the terms Industry 4.0 and data analytics used for literature search, big data, manufacturing, and Internet of

Things are identified as top-5 keywords among relevant research works. Big data is corresponding to the component of analytic methods, manufacturing is corresponding to the component of industry sectors, and Internet of Things are corresponding to the component of CPSs. Although most research focus on big data, manufacturing, and Internet of Things, there are still some trending and meaningful sub-areas in each component. Regarding industry sectors, agriculture, energy, healthcare, and others also benefit from Industry 4.0 besides manufacturing. Regarding CPSs, digital twin is on the uptrend as a virtual representation of the real-time digital counterpart of a physical object to facilitate data analytics on CPSs. Regarding analytic methods, research supported by 5G and

Fig. 7 Trend on titles and abstracts

blockchain is on the uptrend. The different intersections among industry sectors, CPSs, and analytic methods offers a wide range of research opportunities to investigate. In addition, new breakthroughs like digital twin, 5G and blockchain opens new area of work. However, like any historical technical advances, such as deep learning and Internet of Things, new breakthroughs are only part of the entire solution, and should be integrated into industry sectors, CPSs, and analytic methods properly considering their advantages and disadvantages. For example, 5G is good for applications with high speed and low latency wireless connection requirement over a long distance. It is more meaningful for autonomous driving than remote surgery because most remote surgery can be better served by wired connection. Blockchain is good for the collaboration part across multi-organizations to solve trust-worthy concerns but is not a solution for bigger data storage or faster information retrieval. Therefore, blockchain is meaningful for collaborative applications across multi-organizations rather than applications within one company. In all, the most important issue is to step out of the narrow focus of each technique and start with a meaningful application with the appropriate integration among industry sectors, CPSs, and analytic methods. Doing so enables selecting the most desirable technique to achieve the goal and to realize its overall impact.

References

- Adamson, G., Wang, L., & Moore, P. (2017). Feature-based control and information framework for adaptive and distributed manufacturing in cyber physical systems. *Journal of Manufacturing Systems*, 43, 305–315.
- Agrawal, R., & Srikant, R. (1994). "Fast algorithms for mining association rules." In: *Proceedings of 20th International Conference Very Large Data Bases, VLDB*, September 12–15, (vol. 1215, pp. 487–499). Morgan Kaufmann Publishers Inc.
- Ajala, O., Mold, F., Boughton, C., Cooke, D., & Whyte, M. (2017). Childhood predictors of cardiovascular disease in adulthood. A systematic review and meta-analysis. *Obesity Reviews*, 18(9), 1061–1070.
- Anand, M., E. Cronin, M. Sherr, M. Blaze, Z. Ives, and I. Lee. 2006. "Security Challenges in Next Generation Cyber Physical Systems". In: B. Krogh, M. Ilic, and S. S. Sastry (eds.), *Beyond SCADA: Cyber Physical Systems Meeting* (HCSS-NEC4CPS), November 8 & 9, 2006, Pittsburgh, Pennsylvania.
- Bagheri, B., Ahmadi, H., & Labbafi, R. (2011). Implementing discrete wavelet transform and artificial neural networks for acoustic condition monitoring of gearbox. *Elixir Mechanical Engineering*, 35, 2909–2911.
- Bajaj, N., Nuzzo, P., Masin, M., & Sangiovanni-Vincentelli, A. (2015, March). Optimized selection of reliable and cost-effective cyber-physical system architectures. In: 2015 design, Automation & Test in Europe Conference & Exhibition (DATE) (pp. 561–566). IEEE.
- Balkan, M. F. F., Ehrlich, S., & Liang, Y. (2013). Distributed k -means and k -median clustering on general topologies. *Advances in Neural Information Processing Systems*, 26, 1995–2003.
- Ben-Nun, T., & Hoefler, T. (2019). Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *ACM Computing Surveys (CSUR)*, 52(4), 1–43.
- Berger, S., Häckel, B., & Häfner, L. (2019). Organizing self-organizing systems: A terminology, taxonomy, and reference model for entities in cyber-physical production systems. *Information Systems Frontiers*, 1–24.
- Bhaduri, K., Wolff, R., Giannella, C., & Kargupta, H. (2008). Distributed decision-tree induction in peer-to-peer systems. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 1(2), 85–103.
- Bonet, B., & Geffner, H. (2001). Planning as heuristic search. *Artificial Intelligence*, 129(1–2), 5–33.
- Brettel, M., Friederichsen, N., Keller, M., & Rosenberg, M. (2014). How virtualization, decentralization and network building change the manufacturing landscape: An industry 4.0 perspective. *International Journal of Mechanical, Industrial Science and Engineering*, 8(1), 37–44.
- Canizo, M., Onieva, E., Conde, A., Charramendieta, S., & Trujillo, S. (2017, June). Real-time predictive maintenance for wind turbines using big data frameworks. In: 2017 IEEE International Conference on Prognostics and Health Management (ICPHM) (pp. 70–77). IEEE.
- Canizo, M., Conde, A., Charramendieta, S., Minon, R., Cid-Fuentes, R. G., & Onieva, E. (2019). Implementation of a large-scale platform for cyber-physical system real-time monitoring. *IEEE Access*, 7, 52455–52466.
- Chen, H. (2017a). Theoretical foundations for cyber-physical systems: A literature review. *Journal of Industrial Integration and Management*, 2(03), 1750013.
- Chen, H. (2017b). Applications of cyber-physical system: A literature review. *Journal of Industrial Integration and Management*, 2(03), 1750012.
- Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377–387. <https://doi.org/10.1145/362384.362685>
- Demuth, H. B., Beale, M. H., De Jess, O., & Hagan, M. T. (2014). *Neural network design*. Martin Hagan.
- Dey, N., Ashour, A. S., Shi, F., Fong, S. J., & Tavares, J. M. R. (2018). Medical cyber-physical systems: A survey. *Journal of Medical Systems*, 42(4), 1–13.
- Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2), 103–130. <https://doi.org/10.1023/A:1007413511361>
- Duan, L., and W. N. Street. (2009). Finding maximal fully-correlated Itemsets in large databases. In: ICDM, vol. 9, pp. 770–775.
- Elderton, W. P. (1902). Tables for testing the goodness of fit of theory to observation. *Biometrika*, 1(2), 155–163.
- Elhence, A., Chamola, V., & Guizani, M. (2020). Notice of retraction: Electromagnetic radiation due to cellular, Wi-fi and Bluetooth technologies: How safe are we? *IEEE Access*, 8, 42980–43000.
- Essa, A., Al-Shoura, T., Al Nabulsi, A., Al-Ali, A. R., & Aloul, F. (2018, August). Cyber physical sensors system security: Threats, vulnerabilities, and solutions. In: 2018 2nd international conference on smart grid and smart cities (ICSGSC) (pp. 62–67). IEEE.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd (Vol. 96, no. 34, pp. 226–231).
- Faheem, M., & Gungor, V. C. (2018). Energy efficient and QoS-aware routing protocol for wireless sensor network-based smart grid applications in the context of industry 4.0. *Applied Soft Computing*, 68, 910–922.
- Faheem, M., Butt, R. A., Ali, R., Raza, B., Ngadi, M. A., & Gungor, V. C. (2021). CBI4. 0: A cross-layer approach for big data gathering for

- active monitoring and maintenance in the manufacturing industry 4.0. *Journal of Industrial Information Integration*, 24, 100236.
- Fisher, O., Watson, N., Porcu, L., Bacon, D., Rigley, M., & Gomes, R. L. (2018). Cloud manufacturing as a sustainable process manufacturing route. *Journal of Manufacturing Systems*, 47, 53–68.
- Frank, A. G., Dalenogare, L. S., & Ayala, N. F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics*, 210, 15–26.
- Ganschar, O., Gerlach, S., Hämmerle, M., Krause, T., & Schlund, S. (2013). Arbeit der Zukunft – Mensch und Automatisierung. In D. Spath (Ed.), *Produktionsarbeit Der Zukunft-Industrie 4.0* (pp. 50–56). Fraunhofer Verlag.
- German Federal Ministry of Education and Research. (2011). Industrie 4.0 - bmbf. <https://www.bmbf.de/de/zukunftsprojekt-industrie-4-0-848.html>. Accessed 1 May 2021.
- Gokarn, V., Kulkarni, V., & Singh, P. (2017, March). Enhancing cyber physical system security via anomaly detection using behaviour analysis. In: 2017 international conference on wireless communications, signal processing and networking (WiSPNET) (pp. 944–948). IEEE.
- Gorkhali, A., Li, L., & Shrestha, A. (2020). Blockchain: A literature review. *Journal of Management Analytics*, 7(3), 321–343.
- Grant, M., Boyd, S., & Ye, Y. (2006). Disciplined convex programming. In L. Liberti & N. Maculan (Eds.), *Global optimization: From theory to implementation* (pp. 155–210). Springer.
- Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. *ACM SIGMOD Record*, 29(2), 1–12 ACM.
- Hariharan, S., Loeffelholz, T., & Lumanog, G. (2018, October). Powering outdoor small cells over twisted pair or coax cables. In: 2018 IEEE international telecommunications energy conference (INTELEC) (pp. 1–6). IEEE.
- Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, 4(1), 9.
- Hunter, T., Das, T., Zaharia, M., Abbeel, P., & Bayen, A. M. (2013). Large-scale estimation in cyberphysical systems using streaming data: A case study with arterial traffic estimation. *IEEE Transactions on Automation Science and Engineering*, 10(4), 884–898.
- Jiang, Z., Guo, Y., & Wang, Z. (2021). Digital twin to improve the virtual-real integration of industrial IoT. *Journal of Industrial Information Integration*, 22, 100196.
- Jose, B., & Abraham, S. (2017). “Exploring the merits of NOSQL: A study based on MongoDB”. In: 2017 international conference on Networks & Advances in computational technologies (NetACT), IEEE, 266–271.
- Kabugo, J. C., Jämsä-Jounela, S. L., Schiemann, R., & Binder, C. (2020). Industry 4.0 based process data analytics platform: A waste-to-energy plant case study. *International Journal of Electrical Power & Energy Systems*, 115, 105508.
- Karczmarek, P., Kiersztyn, A., Pedrycz, W., & Dolecki, M. (2018). Linguistic descriptors in face recognition. *International Journal of Fuzzy Systems*, 20(8), 2668–2676.
- Kumar, M., Bhatia, R., & Rattan, D. (2017). A survey of web crawlers for information retrieval. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6), e1218.
- Laaki, H., Miche, Y., & Tammi, K. (2019). Prototyping a digital twin for real time remote control over mobile networks: Application of remote surgery. *IEEE Access*, 7, 20325–20336.
- Lawrence, S., Lee Giles, C., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 8(1), 98–113.
- Lee, E. A. (2015). The past, present and future of cyber-physical systems: A focus on models. *Sensors*, 15(3), 4837–4869. <https://doi.org/10.3390/s150304837>
- Lee, S. M., Lee, D., & Kim, Y. S. (2019). The quality management ecosystem for predictive maintenance in the industry 4.0 era. *International Journal of Quality Innovation*, 5(1), 1–11.
- Lei, H., Chen, B., Butler-Purry, K. L., & Singh, C. (2018, May). Security and reliability perspectives in cyber-physical smart grids. In: 2018 IEEE innovative smart grid technologies-Asia (ISGT Asia) (pp. 42–47). IEEE.
- Levitin, G., Xing, L., & Xiang, Y. (2021). Minimization of expected user losses considering co-resident attacks in cloud system with task replication and cancellation. *Reliability Engineering & System Safety*, 214, 107705.
- Li, L. (2020). Education supply chain in the era of industry 4.0. *Systems Research and Behavioral Science*, 37(4), 579–592.
- Li, X., & Xu, L. (2020). A review of internet of things—Resource allocation. *IEEE Internet of Things Journal*, 8(11), 8657–8666.
- Li, D., Tang, H., Wang, S., & Liu, C. (2017). A big data enabled load-balancing control for smart manufacturing of industry 4.0. *Cluster Computing*, 20(2), 1855–1864.
- Li, S., Da Xu, L., & Zhao, S. (2018). 5G internet of things: A survey. *Journal of Industrial Information Integration*, 10, 1–9.
- Lopez, J., Rios, R., Bao, F., & Wang, G. (2017). Evolving privacy: From sensors to the internet of things. *Future Generation Computer Systems*, 75, 46–57.
- Lu, Y., & Ning, X. (2020). A vision of 6G-5G's successor. *Journal of Management Analytics*, 7(3), 301–320.
- MacQueen, J. (1967). “Some methods for classification and analysis of multivariate observations.” In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, 281–297.
- Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). “Recurrent neural network based language model”. In: *Eleventh Annual Conference of the International Speech Communication Association*.
- Mitra, A. (2021). On the capabilities of cellular automata-based MapReduce model in industry 4.0. *Journal of Industrial Information Integration*, 21, 100195.
- Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. 1996. *Applied linear statistical models*, 318. Vol. 4. Irwin.
- Pace, P., Aloï, G., Gravina, R., Caliciuri, G., Fortino, G., & Liotta, A. (2018). An edge-based architecture to support efficient applications for healthcare industry 4.0. *IEEE Transactions on Industrial Informatics*, 15(1), 481–489.
- Palani, N. (2020). ONE-GUI designing for medical devices & IoT introduction. *Trends in Development of Medical Devices*, 17–34.
- Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58, 240–242. <https://doi.org/10.1098/rspl.1895.0041>
- Qi, R., Ji, S., Shen, J., Vijayakumar, P., & Kumar, N. (2021). Security preservation in industrial medical CPS using Chebyshev map: An AI approach. *Future Generation Computer Systems*, 122, 52–62.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1), 81–106. <https://doi.org/10.1007/BF00116251>
- Sahal, R., Breslin, J. G., & Ali, M. I. (2020). Big data and stream processing platforms for industry 4.0 requirements mapping for a predictive maintenance use case. *Journal of Manufacturing Systems*, 54, 138–151.
- Sanislav, T., Zeadally, S., & Mois, G. D. (2017). A cloud-integrated, multilayered, agent-based cyber-physical system architecture. *Computer*, 50(4), 27–37.
- Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1), 61–80.
- Sharma, R., Parhi, S., & Shishodia, A. (2021). Industry 4.0 applications in agriculture: Cyber-physical agricultural systems (CPASs). In: *Advances in mechanical engineering* (pp. 807–813). Springer.

- Slay, J., & Miller, M. (2007). "Lessons learned from the Maroochy water breach." *Critical infrastructure protection* (pp. 73–82). Springer.
- Stojmenovic, I. (2014). Machine-to-machine communications with in-network data aggregation, processing, and actuation for large-scale cyber-physical systems. *IEEE Internet of Things Journal*, 1(2), 122–128.
- Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3), 293–300. <https://doi.org/10.1023/A:1018628609742>
- Tang, L. A., Yu, X., Kim, S., Han, J., Hung, C. C., & Peng, W. C. (2010). "Tru-Alarm: Trustworthiness Analysis of Sensor Networks in Cyber-Physical Systems." In: Data Mining (ICDM), 2010 IEEE 10th international conference on, 1079–1084. IEEE.
- Tang, H., Li, D., Wang, S., & Dong, Z. (2017). CASOA: An architecture for agent-based manufacturing system in the context of industry 4.0. *IEEE Access*, 6, 12746–12754.
- Valente, G., Muttillio, V., Muttillio, M., Barile, G., Leoni, A., Tiberti, W., & Pomante, L. (2019). SPOF—Slave Powerlink on FPGA for smart sensors and actuators interfacing for industry 4.0 applications. *Energies*, 12(9), 1633.
- Visscher, S. L., Naessens, J. M., Yawn, B. P., Reinalda, M. S., Anderson, S. S., & Borah, B. J. (2017). Developing a standardized healthcare cost data warehouse. *BMC Health Services Research*, 17(1), 396.
- Wahid, A., & Kashyap, K. (2019). Cassandra—A distributed database system: An overview. In A. Abraham, P. Dutta, J. K. Mandal, & A. Battacharya (Eds.), *Emerging technologies in data mining and information security* (pp. 519–526). Springer.
- Wang, H. (2020). Research on real-time reliability evaluation of CPS system based on machine learning. *Computer Communications*, 157, 336–342.
- Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for industry 4.0: A self-organized multi-agent system with big data based feedback and coordination. *Computer Networks*, 101, 158–168.
- Wu, B., & Wilamowski, B. M. (2016). A fast density and grid based clustering method for data with arbitrary shapes and noise. *IEEE Transactions on Industrial Informatics*, 13(4), 1620–1628. <https://doi.org/10.1109/TII.2016.2628747>
- Xiong, H., Shekhar, S., Tan, P. M., & Kumar, V. (2006). TAPER: A two-step approach for all-strong-pairs correlation query in large databases. *IEEE Transactions on Knowledge and Data Engineering*, 18(4), 493–508. <https://doi.org/10.1109/TKDE.2006.1599388>
- Xu, L. (2007). Editorial: Inaugural Issue. *Enterprise Information Systems*, 1(1), 1–2. <https://doi.org/10.1080/17517570712331393320>
- Xu, L. D. (2020). The contribution of systems science to industry 4.0. *Systems Research and Behavioral Science*, 37(4), 618–631.
- Xu, L., & Viriyasitavat, W. (2019). Application of blockchain in collaborative internet-of-things services. *IEEE Transactions on Computational Social Systems*, 6(6), 1295–1305.
- Xu, L., He, W., & Li, S. (2014). Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, 10(4), 2233–2243.
- Xu, L., Xu, E. L., & Li, L. (2018). Industry 4.0: State of the art and future trends. *International Journal of Production Research*, 56(8), 2941–2962.
- Xu, L., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. *IEEE Internet of Things Journal*, 8(13), 10452–10473.
- Yang, Y., Wang, S., Wen, M., & Xu, W. (2021). Reliability modeling and evaluation of cyber-physical system (CPS) considering communication failures. *Journal of the Franklin Institute*, 358(1), 1–16.
- Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3), 372–390. <https://doi.org/10.1109/69.846291>
- Zhang, C., & Chen, Y. (2020). A review of research relevant to the emerging industry trends: Industry 4.0, IoT, blockchain, and business analytics. *Journal of Industrial Integration and Management*, 5(01), 165–180.
- Zhang, Z., An, W., & Shao, F. (2016). Cascading failures on reliability in cyber-physical system. *IEEE Transactions on Reliability*, 65(4), 1745–1754.
- Zheng, Z., Xie, S., Dai, H. N., Chen, X., & Wang, H. (2018). Blockchain challenges and opportunities: A survey. *International Journal of Web and Grid Services*, 14(4), 352–375.
- Zhou, B., Li, J., Wang, X., Gu, Y., Xu, L., Hu, Y., & Zhu, L. (2018). Online internet traffic monitoring system using spark streaming. *Big Data Mining and Analytics*, 1(1), 47–56.
- Zhu, Q., & Basar, T. (2011, April). Towards a unifying security framework for cyber-physical systems. In: Proceedings of the workshop on the foundations of dependable and secure cyber-physical systems (FDSCPS-11) (pp. 47–50).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Lian Duan received the PhD degree in management sciences from the University of Iowa and the PhD degree in computer sciences from the Chinese Academy of Sciences. He is an associate professor in the Department of Information Systems and Business Analytics with Hofstra University. His research interests include correlation analysis, industry 4.0, health informatics, and social networks. He has published in the ACM Transactions on Knowledge Discovery from Data, KDD, ICDM, Information Systems, the Annals of Operations Research, etc.

Li Da Xu (Fellow, IEEE) received the B.S. and M.S. degrees in information science and engineering from the University of Science and Technology of China, Hefei, China, in 1978 and 1981, respectively, and the Ph.D. degree in systems science and engineering from Portland State University, Portland, OR, USA, in 1986. He is an Academician of the European Academy of Sciences, the Russian Academy of Engineering (formerly, USSR Academy of Engineering), and the Armenian Academy of Engineering. He is a 2016–2020 Highly Cited Researcher in the field of engineering named by Clarivate Analytics (formerly, Thomson Reuters Intellectual Property & Science).