FoF Optimization

State-of-the-Art Analysis – version 3.8.0

Date: 08.3.2021

Contents

Glossary

AI	Artificial Intelligence
AS	Anomaly Score
AIDC	Automatic Identification and Data Capture
AIDC	Automatic Identification and Data Capture
AMQP	Advanced Message Queuing Protocol
AOA	Angle of Arrival
BLE	Bluetooth Low Energy
CLARA	Clustering LARge Applications
CPS	Cyber-Physical Systems
CNN	Convolutional Neural Net
CAT-M1	Category M1
CPS	Cyber-Physical Systems
CRM	Customer Relations Management
DIPC	Data Integration Platform Cloud
DSC	Dynamic Safety Case
DM	Distributed Manufacturing
DMP	Dynamic Movement Primitive
EAM	Enterprise Asset Management
EM	Expectation-Maximization
ERP	Enterprise Resource Planning
ETL	Extract-Transformation-Load
GATT	Generic Attribute
GSN	Goal Structuring Notation

GMM	Gaussian Mixture Model
GPS	Global Positioning System
HF	High Frequency
IIOT	Industrial Internet of Things
IOT	Internet of Things
JSON	Java Script Object Notation
K-S test	Kolmogorov-Smirnov test
LbD	Learning by Demonstration
LF	Low Frequency
LoRaWan	Long Range Wide Area Network
LoS	Line of Sight
LPWAN	Low Power Wide Area Network
LTE	Long Term Evolution
MDP	Markov Decision Process
M2M	Machine to Machine
MSC	Modular Safety Case
ML	Machine learning
MQTT	Message Queuing Telemetry Transport
MTU	Maximum Transmission Unit
NB-IOT	Narrow Band Internet of Things
OPC-UA	Open Platform Communications – Unified Architecture
PubSub	Publish-Subscribe
PAM	Partitioning Around Medoids
RATS	Real-time Asset Tracking System

RTLS	Real Time Location System
RNN	Recurrent Neural Network
RFID	Radio Frequency Identification
RSSI	Received Signal Strength Indicator
RTLS	Real Time Location System
RTT	Round Trip Time
SCADA	Supervisory Control and Data Acquisition
Sub-GHz	Sub GigaHertz
SVM	Support vector machine
TDOA	Time Distance of Arrival
TOA	Time of Arrival
TOF	Time of Flight
TSI	Time Series Insights
TWR	Two Way Ranging
UHF	Ultra High Frequency
UWB	Ultra Wide Band
WPS	Wifi Positioning System
XAI	Explainable Artificial Intelligence
6LoWPAN	IPv6 over Low power Wireless Personal Area Networks

1. Executive Summary

CyberFactory#1 aims at designing, developing, integrating and demonstrating a set of key enabling capabilities to foster optimization and resilience of the Factories of the Future (FoF). It will address the needs of pilots from Transportation, Automotive, Electronics and Machine manufacturing industries around use cases such as collaborative product design, autonomous machine reconfiguration, continuous product improvement, distributed manufacturing and real time situational awareness. It will also propose preventive and reactive capabilities to address cyber and physical threats and safety concerns to FoF.

In comparison with other Industry 4.0 related projects, the differentiating factors of our approach are threefold. First, the system considered is not a simple manufacturing asset, nor a sum of isolated assets, but a network of factories, which is considered in a System of Systems (SoS) approach. The challenge is to propose novel architectures, technologies and methodologies to optimize the level of efficiency and security of this SoS in a context where every step towards digitization exposes the manufacturing process to widening cyber-threats. Finally yet importantly, we intend to solve more than the technological challenges of Industry 4.0 in this project. Many studies have shown that demonstrating technical feasibility would not be enough to get the buy-in from workers, managers, entrepreneurs, decision makers and customers about novel manufacturing approaches. CyberFactory#1 will therefore embrace technical, economic, human and societal dimensions at once.

A first step is to deliver realistic digital models of FoF and their ecosystem, enabling to perform simulation-aided design, testing and validation of optimization and resilience components. A second output will be the development of key technology bricks for optimization of the manufacturing cycle, enabling real-time sensing and tracking of materials, humans and machines on the shop floor, optimization of human / machine collaboration, distributed manufacturing scenarios, data lake exploitation for process improvement and data-centric business creation. A third output will be to address the need for enhanced resilience of FoF, starting with human / machine access & trust management, human / machine behavior watch, robust machine learning and self-healing mechanisms. The key capabilities will be demonstrated in realistic environments, reflecting the variety of possible new factory types like user-centric plant or learning factories and taking into account business model shifts like turning products into services or developing data services on top of manufacturing activities.

2. Introduction

The system developed under CyberFactory#1 project largely fall into the category of SoS, which have special characteristics that impose several challenges on the architecture design of the solution. Initially the architecture modeling focused strongly on the business and operational level while newer approaches tried to include also architectural aspects of the involved systems. 1990's PERA (Purdue Enterprise Reference Architecture[1]) is an example of the first category, which aims at computer-integrated manufacturing. Architectures that are more recent have been created within Industry 4.0 initiative, with the models proposed by Platform Industry 4.0[2] and Industrial Internet Consortium (IIC), two of the largest organizations that research on topics related to industry and Industrial Internet, respectively, the most mentioned. Other prominent architectures are also ENISA Purdue Model, IBM Industry 4.0 and NIST Service-Oriented Architecture. These several reference models support the definition of CyberFactory#1 system architecture, defined in this document.

The CyberFactory #1 project will result in the validation of all 12 key capabilities listed in Figure 1 through testing and demonstration on a mix of real and simulated factory environments.

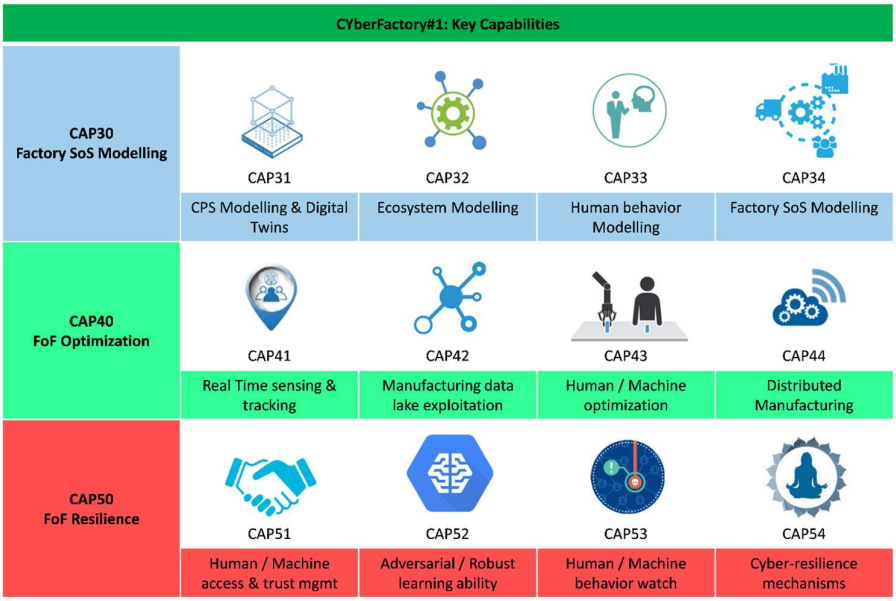

Figure 1 - Overview on key capabilities.

Each capability contributes to one of the 3 main Project outputs:

- **Modeling and simulation of Factory System of Systems**: a set of modular capabilities enabling to virtually replicate a factory, including machines, humans and its vital environment (supply chain, customers) to support decision making for process optimization and resilience.
- **Factory of the Future Optimization**: a set of modular capabilities enabling to optimize manufacturing and supply chain processes and develop data-centric business models generating new revenue streams, starting with improved management of factory shop floor through the use of the Industrial Internet of Things (IIoT), enhanced optimization of human and

1 *The Purdue enterprise reference architecture.* Williams, Theodore J. 1994, Computers in Industry.

2 Plattform Industrie 4.0. *Plattform Industrie 4.0.* https://www.plattform-i40.de/PI40/Navigation/EN/Home. [Accessed at March 13, 2020.]

robot collaboration, manufacturing load balancing techniques and data lake exploitation techniques.

- **Factory of the Future Resilience**: a set of modular capabilities enabling to enhance resilience of future factories, in particular by implementing permanent trust management techniques for physical and logical access control of critical assets, innovative techniques for anomaly / attack detection based on behavior analysis, offensive and defensive techniques for manipulation of artificial intelligence, and system self-healing mechanisms to ensure business continuity and recovery.

2.1. Purpose and Scope of this Document

Nowadays, sensors are getter cheaper and more efficient; these are often embedded in machines and devices, from a simple smartphone to a wristband. The amount of data generated every single day worldwide is astonishing; by 2025, approximately 463 exabytes of data will be created[3]. At the same time, communication speed as well as increased protocols make connectivity cheaper and more reliable, where cloud technology and mobile devices provide information any time and any place. Such large and fast volumes of data and the need to extract knowledge close-to-real-time from them are industrial challenges that must be addressed. Furthermore, the joint usage of manufacturing control and data lake analytics through the industrial internet will produce huge opportunities in all manufacturing areas. As are new devices becoming more intelligent and interconnected, the integration of these in a production line, will offer a new set of possibilities, such as: better scheduling and maintenance strategies, a better quality prediction (zero-defect manufacturing), more reliable safety management and mechanisms, and any kind of operation requiring some “step by step” process (installation, assembly, etc.) can be improved. Yet a number of limitations of state-of-the-art technologies restrain the potential for factory optimization and self-improvement.

The goal of this document is to first present state-of-art solutions for several constraints or limitations that exist nowadays on the transportation, automotive, electronics and manufacturing industries. Secondly, identify limitations on the current technologies or approaches and highlight how CyberFactory#1 will contribute with new solutions and approaches that might mitigate or eliminate them. Some of the topics discussed (but not limited) are:

- Real time situational awareness, asset location and tracking. The state-of-the-art indoor location technologies for LPWAN consist in having a tag listening to signals emitted by a nearby Bluetooth Low Energy (BLE) beacon, report the ID and the received power strength of the beacon to a location platform.
- Manufacturing data lake analytics and services. The data generated by real time situational awareness, asset location and tracking systems are going to populate a data lake, a large storage repository that holds a vast amount of raw data in its native format, without any connection, until it is needed. The goal now is to retrieve knowledge and patterns from these huge amounts of data, as the competitiveness and capacity for innovation depends on a company ability to manage and use knowledge. Using machine learning techniques, and large volumes of data with high quality and purity, several services such as anomaly

3 “How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read”https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=2034383660ba [Accessed at February 23, 2021], Forbes.

detection (measuring deviations from the normal behavior), condition monitoring in continuous processes, and other systems optimizations [4] [5] [6] have been proposed.

- A more collaborative human-machine interaction, where skills between humans and machines are shared. The current state of the art requires a robotic expert or a trained worker to program these cobots using teach-pendants. In the Factories of the Future (FoF), the collaborating worker should more intuitively program the cobots that they will be interacting with, based on their own preferences and experience to increase productivity and efficiency.
- Collaborative product design and distributed manufacturing. In the FoF, the manufacturing process is based on a network of geographically distributed, but connected, factories, suppliers, distributors, and end consumers. While digitalization and interlinkage of distributed manufacturing systems provide the basis to implement the FoF, these are heavily dependent and supported by machine learning algorithms, data lake or big data infrastructure and collaborative decision making techniques (highly dependent on a productive human-machine interaction), IoT and IioT platforms (highly dependent on cloud services) and asset location and tracking systems.

The CyberFactory#1 challenge is to propose novel architectures, technologies and methodologies to optimize the level of efficiency and security of this FoF in a context where every step towards digitization exposes the manufacturing process to widening cyber-threats. Using the abroad experience and knowledge of both industrial and academic partners, synergies are going to be created and established through regular meetings, where experiences are going to be shared. In the process business, opportunities are going to be identified, from the different and cross-domains.

2.2. Deliverable Structure

The document is divided into five main sections:

- **Section 3** is dedicated to the analysis of current technologies and applications concerning the real-time and near real-time sensing, tracking & supervision of materials and assets, both inside and outside of the factory.
- **Section 4** presents data lake and big data technologies applied to FoF. Furthermore, state-of-art artificial intelligence algorithms and their applications to Industry 4.0 are presented and studied.
- **Section 5** presents learning by demonstration state-of-art algorithms, mainly the limitation of the Dynamic Motion Primitives approaches and the propose solutions.
- **Section 6** focus on methodologies aiming to support the management of work-in-progress and manufacturing loads on the shop-floor, through a network of factories.
- **Conclusions** presents the limitations of the current factory architecture, and the novelty and contributions made in CyberFactory#1 project.

4 Lee, Jay, Behrad Bagheri, and Hung-An Kao. "A cyber-physical systems architecture for industry 4.0-based manufacturing systems." Manufacturing Letters 3 (2015): 18-23.

5 Maier, Alexander, Sebastian Schriegel, and Oliver Niggemann. "Big Data and Machine Learning for the Smart Factory—Solutions for Condition Monitoring, Diagnosis and Optimization." Industrial Internet of Things. Springer International Publishing, 2017. 473-485.

6 Lee, Jay, Hung-An Kao, and Shanhu Yang. "Service innovation and smart analytics for industry 4.0 and big data environment." Procedia Cirp 16 (2014): 3-8.

3. Real time asset sensing and tracking

The number of data sources on the shop floor has dramatically increased thanks to the fourth industrial revolution. Due to successive advancements, developments and innovations in sensor and networking technologies, their increased availability, affordability and the spread of industrial internet have made it possible to have real time/ near real time visibility into the shop floor assets and field operations. Digitized production processes offer controllability and traceability, in contrast to the traditional sub-processes, which are executed through human interactions[7]. The importance of digitization in production is well known and there is no sector in industry that has not recognized it yet. Broadly speaking, digitization here means being able to gather all relevant data about manufacturing processes / assets and use them to increase efficiency, productivity and control over them. According to 2018 Manufacturing Vision Study[8], manufacturers have already realized the real benefits of data connectivity: increased visibility into the entire manufacturing process; an accelerated pace in shipping and receiving; faster identification of points-of-failure; and deeper insights into the interworking of their operations. This study shows that the amount of data being captured but not connected to any system or staff to reduce from 21% to 11% through 2017 to 2022. In a connected plant floor, every physical asset has a digital profile. Manufacturers use these profiles to track real-time location, material allocation and condition of assets. The data can also be used to improve the overall manufacturing process, eliminate bottlenecks, communicate with suppliers and ensure quality. Although only 24% of those surveyed currently have technology-driven tracking capabilities in place, it's something manufacturers know they need. In five years, 63% of those surveyed plan to increase their tracking with more than 28% planning to adopt real-time monitoring [9]. In CyberFactory #1 project, CAP41 focuses on real time or near real time data collection from shop floor in the factories of the future. The data might be the position of a moving robot, presence of pallet inside a factory, materials in the process or operating machines in production lines. The operating machines are usually equipped with sensors to produce (a big amount of) data as basis for the use of analytical techniques and methods. Since a lot of applications in this area generate benefits by processing data from multiple machines and other sources together, it is first necessary to transfer the data over a network to a server to be able to apply analytical algorithms.

In this document, the state-of-the-art for RT or near-RT tracking technologies are examined in two chapters. The first one is real time location systems (RTLS) which locates anything such as manufacturing assets, people, products and anything that carries a RTLS tag on it inside a predetermined area. The second one is the industrial IoT platforms, which provides connectivity to shop floor and provides many additional features to users for facilitating production.

3.1. Real Time Location Systems (RTLS)

Real Time Locating Systems (RTLS) are wireless systems with the ability to locate the position of an item anywhere in a defined space (local/campus, wide area/regional, global) at a point in time that is, or is close to, real time. Position is derived by measurements of the physical properties of

---

7 Kerem Kayabay and others, 'Big Data for Industry 4.0: A Conceptual Framework', in *2016 International Conference on Computational Science and Computational Intelligence*, 2016, pp. 431–34.

8 "2018 Manufacturing Vision Study… The Road Ahead." [Online]. Available: https://cssi.com/2018/04/20/2018-manufacturing-vision-study-the-road-ahead/. [Accessed at May 27, 2020].

9 '2018 Manufacturing Vision Study… The Road Ahead' <https://cssi.com/2018/04/20/2018-manufacturing-vision-study-the-road-ahead/> [Accessed at May 27, 2020].'2018 Manufacturing Vision Study. '

the radio link [10]. With the usage of RTLS systems in industry, companies can minimize production slowdowns, identify areas to improve, and provide workers with added safety features by tracking manufacturing assets and workers movement in real-time [10].

3.1.1. Components of a Real Time Asset Tracking System

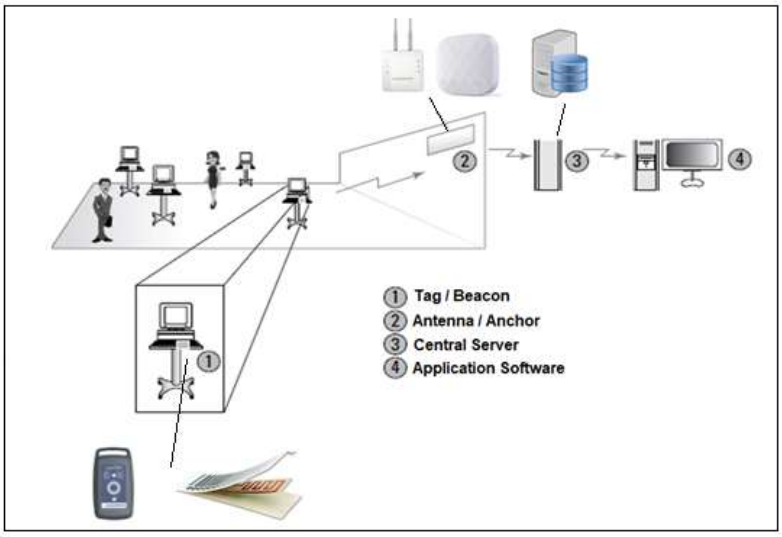

Figure 2 - Components of an RTLS system (extracted from[16]).

Every RTLS system incorporates a combination of hardware and software to create an enclosed positioning network. The infrastructure mainly consists of four main elements as shown in Figure 1.

Tag / Beacon: Small element, which needs to be attached to the target that is being tracked. They are miniature devices that are enabled with location technology. Depending on the tag's communication method with other parts of the RTLS, they are categorized as passive, semipassive or active.

- ***Passive tags*** are typically passive radio frequency identification (RFID) tags. The reader (interrogator) sends a radio signal that is received by the passive tags present in the RF field of the interrogator. Tags receive the signal via their antennas and then respond by transmitting their stored data. Passive RFID tags have no battery and obtain the operational power to transmit data from the RF field emitted by a corresponding interrogator[11].
- ***Semipassive tags*** (or battery assisted passive tags) are very similar to passive tags. They do not initiate any communication, need to be in the RF field of the interrogator to be read, and send data to the interrogator using the same backscatter technique as passive tags.

---

10 2014 International Organization for Standardization, 'ISO/IEC 24730-1:2014 Information Technology — Real-Time Locating Systems (RTLS) — Part 1: Application Programming Interface (API)', 2014.2014 International Organization for Standardization, 'ISO/IEC 24730-1:2014 Information Technology — Real-Time Locating Systems (RTLS) — Part 1: Application Programming Interface (API)', 2014.2014 International Organization for Standardization, 'ISO/IEC 24730-1:2014 Information Technology — Real-Time Locating Systems (RTLS) — Part 1: Application .

11 Rahul Bhattacharyya and Pavel Nikitin, 'Guest Editorial: Special Issue on IEEE RFID 2019 Conference', *IEEE Journal of Radio Frequency Identification*, 4.1 (2020), 1–2 <https://doi.org/10.1109/jrfid.2020.2973562>.

However, semipassive tags have a small battery[12]. The battery's main purpose is to either monitor environmental conditions or to offer greater range and reliability than passive tags. Note that the battery on semipassive tags is not used to generate RF energy.

- ***Active tags*** contain an onboard radio (transmitter or transceiver) and are typically powered by an internal battery[13]. Because they have onboard radio, they usually have a long range and can communicate without being prompted. For these reasons, these tags can be located in real time, say every second or any frequency needed by the application. However, the battery life becomes an important concern with increasing communication.
- ***Beacons*** are small always-on transmitters, which can use Wi-Fi, Bluetooth Low Energy (BLE), SubGHz NB-IoT, or other technology to broadcast signals to nearby portable devices, mainly BLE. They can be considered as a light house which repeatedly transmits a single signal that other devices can see[14]. The beacon sensors usually transmit data, but do not receive them. A device like a smartphone can see a beacon once it is in the device's range, and do what it is programmed to do when it sees the beacon.

Antenna / anchor: They are devices within an RTLS that typically have a known position and detect the location of tags. Location sensors locate tags by using a *physical parameter,* or a measurement, that exists between the sensors and the tags. The physical parameter can be something as simple as *visibility* of the tag to anchor or it can be more complex, such as measuring the time a signal takes to travel from the tag to the sensor. The number of location sensors needed in a facility usually depends on the technology, the application, and the desired accuracy. For example, if you are implementing an RTLS for recording and monitor where a production asset is, you might need location sensors covering all the facility. In another example, if you are trying to implement an RTLS to monitor is an asset is in the building or not, you just need to locate an anchor at the gates.

Central Server: It communicates with anchors and manages the antennas, calculates the locations (distance and position) of the tags, and makes those data available to be exploited to application software. It contains location engine that determines the location of the tags and the middleware, which resides among the pure RTLS technology components and the business applications as seen

12 Kasyap Suresh and others, 'A Comparative Survey on Silicon Based and Surface Acoustic Wave (SAW)-Based RFID Tags: Potentials, Challenges, and Future Directions', *IEEE Access*, 8 (2020), 91624–47 <https://doi.org/10.1109/ACCESS.2020.2976533>.

13 Arvind Lakshmanan and Vivek Maik, 'Active RFID Tag with Better Tracking Range for Automotive Applications', 2020, 165–69 <https://doi.org/10.1109/irce.2019.00040>.

14 Akshay Jayraj Suvarna, Avaneesh Pratap Singh, and H. K. Shashikala, 'Beacon Technology', *International Journal of Computer Science and Mobile Computing*, 8.6 (2019), 100–105.

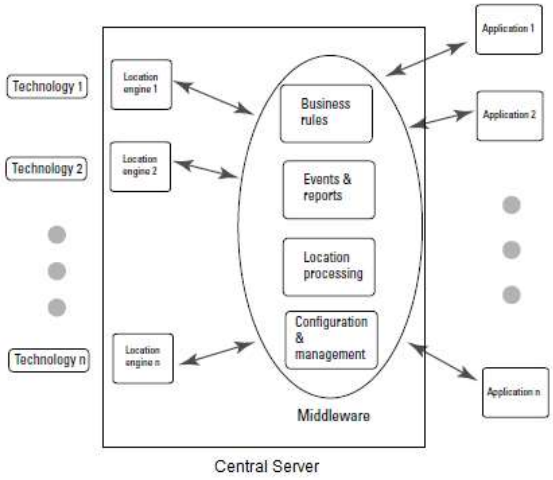

Figure 3 - General RTLS structure (extracted from[16]).

in Figure 2. In other words, the middleware resides among pure RTLS technology components and business applications.

Location estimation consist of ranging techniques to estimate distance between the tag and anchor, and position estimation techniques that derive the position of the tag.

Ranging techniques estimate distance, or range (usually in terms of feet or meters), between the tag and the anchor. The physical variables that are commonly used to determine the estimated distance are listed below [16]:

1. Time of Arrival (TOA): The time it takes a signal to travel from the location sensors to the tag or vice versa. This time to travel, also known as propagation delay, can be converted into distance between the tag and the location sensor by multiplying it by the signal's propagation speed.
2. Angle of Arrival (AOA): The angle between the propagation direction of signal and some reference direction, which is known as orientation. Here, direction sensitive antennas are used at anchor points to determine the direction and angle of the signal coming from a tag. Then, the tag position is estimated by finding the intersection of different signal propagation paths.
3. Time Distance of Arrival (TDOA): Similar to TOA but instead of exact TOA measurement TDOA measures the difference in transmission times between signals received from each of the transmitters to a tag or vice versa.
4. Time of flight (TOF): Uses measured elapsed time for a transmission between a tag and a location sensor based on the estimated propagation speed of a typical signal.
5. Round trip time (RTT): Uses the total time for a signal to start from the anchor and the acknowledgment to be received back.
6. Received signal strength indicator (RSSI): A measurement of the power present in a received radio signal. As a signal leaves its source, it attenuates, meaning that the power of the signal drops. The drop is logarithmic, and the signal attenuation in open space, as well as different medium, is well defined. Because the power levels at the start of transmission of a signal are known, RSSI can be used to estimate the distance the signal has traveled.

In general, no single variable can be used to provide accurate ranging estimation under all circumstances. Each of them has its own advantages and limitations in terms of location accuracy and generally, a combination of some of them is used.

Position estimation techniques derive the position of the tag. This includes making use of estimation algorithms on all the estimated tag distances from the location sensors and the actual position of all the anchors to the estimated tag position. Commonly used algorithms are[15]:

1. Trilateration: A technique in which you can estimate the position of something if you know the distance to three different locations.
2. Triangulation: A technique in which you can estimate the position of something if you know the line angle between that something and the three different locations with respect to a common reference line, such as a line pointing up.
3. Nearest neighbor: Simply neighbor relationships are used to estimate a position. A neighbor relationship is based on any of the ranging techniques, such as RSSI.

Many technologies are available that enable an RTLS with one of the variables and the algorithms described in previous paragraphs. Bluetooth, RFID, ultra wideband (UWB), Wi-Fi, ZigBee and etc are among those technologies which will be examined in detail later on in this document. Each of these positioning technologies has certain level of accuracy and a service area where it performs better. Then, you can't rely on any one technology to provide accurate location information in all environments. Each technology has its pros and cons as summarized in following chapters.

Application software[16]: The computer software interacts with the RTLS middleware to solve the problems challenging end users, such as enabling users to achieve the tasks that they wish to perform or solves problems for another application program. The application provides the end user value, the real time actionable business intelligence that is accessible to the relevant systems and people. It involves the establishment of alerts, alarms, actions, decisions, audit trails, and documentation.

3.1.2. Positioning Technologies

The asset tracking in production aims to control parts, elements or tools that take action in the production line, as well as the process itself, its disruptions and the data it produces. Nevertheless, the components of an RTLS and its capabilities may depend on the technology used for its implementation. In this document, focus is on the indoor positioning technologies. Therefore, previously mentioned technologies such as such as, RFID, BLE and UWB and their components will be reviewed in following parts:

Bluetooth Low Energy (BLE) Technology: It is a subset of the Bluetooth v4.0 standard (or emerging standards such as Bluetooth 5)[17]. It has a completely new protocol stack in reference to the OSI layer and oriented to simple connections in very low power applications. The technology is primarily used for mapping and location services using the RSSI (received signal strength indicator) estimate[18]. The broadcasted beacon signals can be captured by smart gadgets, like phones, to call ad-hoc actions as shown in Figure 4.

15 Pouria Zand and others, 'A High-Accuracy Concurrent Phase-Based Ranging for Large-Scale Dense BLE Network', *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, 2019-Septe (2019).

16 Ajay Malik, *RTLS for Dummies* (NJ: Wiley Publishing, 2009).

17 "BLE." [Online]. Available: https://www.elt.es/en/ble. [Accessed at April 30, 2020].

18 "Mark up the world using beacons." [Online]. Available: https://developers.google.com/beacons. [Accessed at April 30, 2020].

Figure 4 - BLE beacons (extracted from[19]).

Using a Bluetooth connection, the technology is cheaper than alternatives and easier to use and support. The main components of a RTLS based on BLE technology are the beacons and gateways:

***BLE Beacons**:* Beacons are small battery powered devices that emit a BLE signal that is picked up by the BLE GW. The BLE beacon hardware consists of a microcontroller with a Bluetooth LE radio chip and a battery. New radio chips are optimized for Bluetooth LE, whereas older versions were designed for Bluetooth Classic, which had higher power consumption. Coin cell batteries are the most popular choice for most beacons. These batteries are dense Lithium Ion cells and provide up-to 1,000 mAh of stored power in a very small form factor. Commonly available coin cell sizes are 240 mAh (CR2032, small size), 620 mAh (CR2450, medium size) and 1,000 mAh (CR2477, large size). Some beacons also use Alkaline AA batteries. A typical AA battery provides around 2,000 mAh power but at a significantly larger size than coin cell batteries. Lastly, some beacons are externally powered. They can be installed in a wall outlet or a USB outlet. These beacons do not need battery replacement and can be economical in certain situations. However, availability of a power outlet, without running new wiring, at exact spots where a beacon is required may not always be feasible.

Every beacon has a specific firmware, which is the logic (programmed code) that enables the beacon hardware to operate. The firmware can control several characteristics that impact the battery life:

- Transmit Power (tx power): Beacon devices transmit a signal with a fixed base power, known as the tx power. As the signal travels in air, the received signal strength decreases with distance from the beacon. Higher tx power means, the signal can travel longer distances. Lower tx power means, less battery consumption but also smaller range.
- Advertising Interval: It is the interval that a beacon emits a signal. An interval of 100ms means the signal is emitted every 100 milliseconds (or 10 times in a single second). A higher interval of 500 ms means the signal is emitted only twice per second, which means less battery drain for the beacon. As the advertising interval increases, the battery life of the beacon also increases, but the responsiveness of the phone decreases. There is no optimal choice of advertising intervals, and applications needing low latency should choose lower advertising intervals, and those needing higher battery life should increase the advertising interval.

Each beacon provides its own way of configuring the hardware and associated parameters (Tx power and advertising interval). Some beacons provide their own proprietary app to configure the beacons, normally to be installed in a smartphone with BLE interface available. Other beacons provide open interface via any GATT client. The main advantage of beacons supporting GATT method is that hundreds of beacons can be configured at once. Tx power and maximum coverage

19 'Apple Inc. (AAPL) iBeacon Technology Revolutionizes A New Vision For The Retail Industry' [online]. Available: https://dazeinfo.com/2014/03/25/apple-inc-aapl-ibeacon-technology-retail-industry-video/

distance of beacon, in direct line of sight, is shown Figure 5. The range decreases if there are any obstacles between the beacon and the receiver.

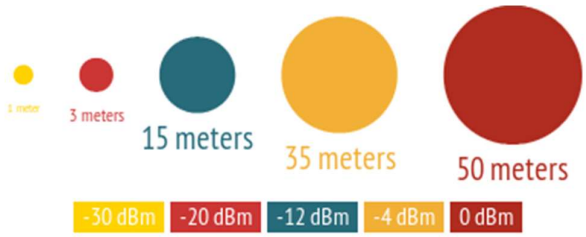

Figure 5 - Tx power and maximum coverage distance of a beacon (extracted from[20]).

***BLE Gateways**:* Normally BLE gateways are devices with two wireless interfaces, BLE and WiFi, that receives data via BLE and sends them via WiFi to the central server for further processing. In some cases the BLE can have additional intelligence so that it can process somehow the received signal, not just forwarding it. For instance, the GW can calculate the distance from the obtained RSSI from the beacon, or even smooth somehow the received RSSI so that to improve its quality. They can also provide additional functionalities such as device monitoring and security.

We can define technology by its defined OSI layers:

- Physical Layer: BLE technology is capable of using up to 40 2MHz channels in the 2.4 GHz ISM band.
- Link Layer: Layer that manages the connection (HCI Protocol) and the definition of roles (Advertiser, Scanner, Master and Slave) in communication.
- L2CAP Layer (Logic Link Control and Adaptation Protocol): Message format and encapsulation. MTU (27 Bytes).
- GAP & GATT Layers: Visibility and Interaction between two devices. GATT is based on the Attribute Protocol (ATT).

Radio Frequency Identification (RFID) Technology: It is the wireless non-contact technology use of radio frequency waves to transfer data, whereby digital data encoded in RFID tags are captured by a reader via radio waves. Tagging items with RFID tags allows users to automatically and uniquely identify and track inventory and assets. At a simple level, RFID systems consist of three components: an RFID tag, an RFID reader and an antenna. RFID readers are responsible for generating the waves that are emitted by the antennas towards the tags, and at the same time they also receive and decode the signal emitted by the tags that arrives through the antenna to the reader. The antenna and reader can be coupled into the same device or they can be two separate devices connected through a coaxial cable. This is why they are normally described together. The first option has the advantage of avoiding signal loss through the connection cable but, on the contrary, is more expensive as it normally implies to have always one reader per antenna. On the other hand, the second option has the advantage that multiple antennas can be added to the same reader (that is the most expensive device), but the signal loss through the coaxial cable must be considered and therefore lower power is emitted by the antenna and scope of the coverage is lower.

Among several specifications, the three most important to take into account when selecting a RFID reader and antenna are frequency, polarization and gain.

- Frequency: RFID operates at several frequencies bands [21]:

20 'The Hitchhikers Guide to iBeacon Hardware: A Comprehensive Report by Aislelabs' [online] available: https://www.aislelabs.com/reports/beacon-guide-2014/

21 Angell, I., Kietzmann, J. (2006). "RFID and the end of cash?"(PDF). *Communications of the ACM*. 49 (12): 90–96.

- Low Frequency (LF): This band covers 30 KHz to 300 KHz. Typically LF RFID systems operate at 125 KHz, although there are some that operate at 134 KHz. This frequency band provides a short read range of 10 cm, and has slower read speed than the higher frequencies, but is not very sensitive to radio wave interference. LF RFID applications include access control and livestock tracking.
- High Frequency (HF): This band ranges from 3 to 30 MHz. Most HF RFID systems operate at 13.56 MHz with read ranges between 10 cm and 1 m. HF systems experience moderate sensitivity to interference. HF RFID is commonly used for ticketing, payment, and data transfer applications.
- Ultra High Frequency (UHF): The UHF frequency band covers the range from 300 MHz to 3 GHz. Regarding UHF (Ultra High Frequency) RFID antennas, the most used frequencies are 902-928 MHz (US/FCC), 865-868 (EU/ETSI), 860-960 (Global). In Europe, there are two used bands, which are the low-band (865,6-867,6 MHz) and the high-band (915-921 MHz). Moreover, there is another important frequency, which is 433 MHz, that is widely used form large range vehicle identification, people/object control and security applications. The read range of passive UHF systems can be as long as 12 m, and UHF RFID has a faster data transfer rate than LF or HF. UHF RFID is the most sensitive to interference, but many UHF product manufacturers have found ways of designing tags, antennas, and readers to keep performance high even in difficult environments. Passive UHF tags are easier and cheaper to manufacture than LF and HF tags.

The selected frequency for an application depends on their requirements. The higher frequency, the higher the distance and the reading speed, but also the cost. When selecting an RFID antenna and reader, it must be assured that the selected frequency range is suitable for the region / country where it will be deployed.

- Gain: A higher gain implies a thinner electromagnetic beam. Therefore, a higher gain creates a narrower but longer coverage area, achieving a larger reading distance. This is, the antenna will be more directive. The ideal gain and beam width will depend on the application requirements. For example, using antennas of 8,5-10 dbi or higher will allow to emit a higher power and to obtain a higher reading rate.
- Polarization: Most of the antennas have circular or linear polarization. If all the tags will be read in the same orientation and height in a specific application, the linear polarization fits better. On the other hand circular polarization antennas have the advantage of fitting better on applications where is difficult to foresee the orientation or location of the tag. Therefore, the use of circular polarized antennas is wider due to this flexibility.

Taking all these aspects into consideration, the most suitable RFID deployment for a RTLS solution in an industrial environment should use UHF antennas and readers for European standards 865-868 MHz (EU/ETSI). It should also be a directional antenna with high gain (8,5 – 10 dbi or higher) for achieving larger reading distance. In addition, with circular polarization, for being independent of the tag orientation. Some of the aspects to consider when selecting RFID tags are the type (label, protected, etc.), operation frequency (LF, HF, UHF), environment (metal, indoor, outdoor, etc.), fastening method (adhesive, pinned, etc.), reading distance and rate, size and cost, among others.

- Operation Frequency: The tag frequency selection depends on aspects such as the required reading distance, material where the tag will be located and the reading rate. For example, if large distances are required (higher that 1-3 meters) UHF tags are required. When working with UHF tags, devices that are compliant with the well-known and widely used Gen2 standard are recommended.
- Tag type: The tag type selection depends on the reading distance, cost, size, weight and application type. In short, there are passive and active tags. Passive ones use tags with no internal power source and instead are powered by the electromagnetic energy transmitted from an RFID reader. The lower price point per tag makes employing passive RFID systems economical for many industries. Active tags are battery-powered that continuously broadcast their own signal. Active tags provide a much longer read range than passive tags, but they are also much more expensive.

- Environment: Regarding environment, it has to be considered that metal environments provoke signal to bounce and water environments absorb energy. There are solutions to the challenge of using RFID on metal and water, but specially designed tags are more expensive than generic tags that can be used on RF-friendly objects. For example, metal objects interfere with and disturb the tag functioning, especially on LF and HF. In this case, UHF tags are recommended and additionally they require a little separation between the tag and the metal surface.
- Orientation: The tag orientation regarding the antenna also affects the own tag performance. The best orientation is always when the tag and the antenna are in the parallel plane to each other. Like this, the tag can receive the whole energy from the antenna. While the tag turns, it presents a smaller effective area for the electromagnetic waves and, therefore, takes less amount of energy. The reading distance of the tag reduces as the obtained energy is smaller. This is why a small size tag will be readable to a shorter distance that a bigger tag.

Considering all these aspects, the most suitable RFID tag for a RTLS solution in an industrial environment should use UHF tags for higher reading distance and metal specific tags (taking into account that these ones cannot be passive or small size).

Ultra-Wide Band Technology: Indoor positioning with Ultra-Wide Band (UWB) has some significant advantages: The accuracy is 10-30 cm; latency time is very low (position request up to 100 times/second); height differences can be measured accurately. Anyway, the technique is a special solution which requires appropriate components that will summarized in this section. The base of the system is a set of anchors that are positioned in the area to be monitored, used as positioning reference. The other part of the system is one or multiple tags that are fixed to the object that are to be tracked. Both devices, anchor and tag, have very similar HW, they are usually equipped with a UWB module and power supply. In addition, they are then provisioned with tag or anchor firmware, depending on the role. For the UWB anchors, they are also provided with an additional interface (Ethernet, WiFi, LoRa, 3G, etc) for data transmission and configuration. Anchor is a referential device with a known position. In a RTLS context, anchors are electronic devices that detect UWB pulses emitted by UWB Tags and forward them to the location server for calculating tag positions. To cover the area with an indoor tracking system, a set of anchors needs to be installed above the area to create the location infrastructure, where tags are being located. An UWB RTLS Platform is fully scalable, allowing the unlimited expansion of monitored area just by adding extra anchors to the network. In a RTLS context, tags are small electronic devices that are attached to objects that need to be tracked. The tags send out blinks that are received by anchors and forwarded to the location server for calculating the tags' position. RTLS Tags are used for asset tracking, vehicle tracking, material flow analysis and employee location tracking for safety reasons.

Wi-Fi positioning systems (WPS): In WPS, multiple wireless access points measure the relative signal strength of assets to approximate their position inside a facility. Factories usually have a Wi-Fi network deployed and easy to make deployment of the system. GPS has a similar function on works outdoors.

ZigBee: It is another wireless standard operating in an unlicensed segment of the broadcast spectrum. Its range is slightly greater than Bluetooth, but this comes at the cost of some reliability problems. A wireless connectivity module implementing ZigBee standard is shown in Figure 6.

Advantages:

- Simple Technology—ZigBee is an older, relatively simple wireless technology that's easy to implement in any environment.
- Cost for tags and receivers are durable and inexpensive.

However, the signals can easily degrade in 'complicated' environments and have interferences.

Figure 6 - Example of Zigbee module (extracted from[22]).

Sensor and Micro-controllers: In commercial market, we have a multitude of sensors and communication elements for IoT. One example is NodeMCU which is the development board based on the ESP8266. It incorporates:

- a low-consumption 32-bit MCU (Tensilica L106)
- 2.4 GHz WiFi module
- RAM of about 50 kB
- 1 analog 10-bit input (ADC)
- 17 GPIO input and output pins (general purpose)

Outdoor Tracking Systems, possibilities offered by 5G technology: Location technologies will only get better with the advent of the new wireless standards within 5G. Location aware devices and components will boost up capabilities of industry 4.0, medicine, smart city applications. High positioning accuracy will be achieved with the help of large number of sensors, high bandwidth, low latency and new frequency ranges. Drones, AGVs, further location aware devices maybe combined with AR technologies to enrich logistics and asset tracking applications. Positioning accuracy in cellular networks is expected to increase from 20 meters to a few decimeters in upcoming new releases of 5G standards[23].

3.1.3. Data Collection

In this section, a short introduction to data collection and big data will be given. Then, data collection from RTLS technologies and other data sources will be described. Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Big data can be defined as data sets or combinations of data sets whose size (weight or volume), complexity (variability) and growth rate (speed and scalability) make it difficult to capture, manage, process or analyze them using conventional technologies and tools, as conventional statistical and relational databases or visualization packages, within the time necessary to be useful. The data set in normal use of technology generally varies from 50 terabytes to several petabytes.

The different types of data are also:

I. Types of unstructured data: documents, logs, social media videos, audios, etc.
II. Semi-structured data types: software, spreadsheets, reports.
III. Structured data types.

RTLS, as a two-way wireless communication system to share real time location data, is a reliable and efficient real-time big data source to establish communication between tags and readers and provides frequent updates.

---

22 'Digi XBee Zigbee Datasheet'[Online]. Available: https://www.digi.com/resources/library/data-sheets/ds_xbee_zigbee [Accessed at April 30, 2020]
23 "Expected Positioning Accuracy of 5G." [Online]. Available: https://www.iis.fraunhofer.de/en/ff/lv/lok/5g/accuracy.html [Accessed at April 30, 2020].

Nevertheless, data collection in a RTLS solution may depend on the technology used for its implementation. Therefore, data acquisition approaches of previously mentioned technologies will be summarized in this section. The central server of an RTLS system plays a crucial role in data collection. It is responsible for calculating the asset location based on the information received from the antennas and the proposed technologies.

Data Collection in RFID: RFID belongs to a group of technologies referred to as Automatic Identification and Data Capture (AIDC). AIDC methods automatically identify objects, collect data about them, and enter those data directly into computer systems with little or no human intervention. RFID tags contain an integrated circuit and an antenna, which are used to transmit data to the RFID reader. The reader then converts the radio waves to a more usable form of data. Information collected from the tags is then transferred through a communications interface to a host computer system, where the data can be stored in a database and further analyzed. Information collected from tags and stored in the repository, is normally a tag ID detected by an antenna and a timestamp. Thanks to this information, if each tag ID is linked to a monitored asset and each antenna is linked to a monitored area, we can establish that an asset is or is not in a certain area. Due to this fact, RFID based solutions are mostly used for identification purposes than for accurate location tracking, but depending on the requirements of the use case it can provide useful location information.

Data Collection in BLE: BLE beacons emit a BLE signal. The signal is picked up by the BLE gateway, and often transmitted to a central server. This central server processes the received signal strength indicator (RSSI) from each gateway to a same beacon, and based on this data it can estimate the distance from each gateway to the beacon. Then, thanks to trilateration techniques, at least the information from three gateways is needed, the beacon location is estimated. Due to the RSSI instability due to equipment and environment, the obtained location is not highly accurate, and therefore, BLE based solutions are more suitable for proximity evaluation.

Data Collection in UWB: UWB measures that Time Of Flight (ToF), time that the light signal takes to cover the distance between the anchor and the tag. It is recommended to have line of sight between the tag and the anchor. UWB anchors transmit a very wide pulse over a GHz of spectrum. The anchors then listen for chirps from the UWB tags. These tags have a spark-gap-style exciter that generates a little pulse within them, which creates a short, coded, very wide, nearly instantaneous burst. Each anchor then reports very accurate time measurements from the tag back to a central server, which processes and combines the received ranging results of the anchors to each tag and calculates the tag location. Nevertheless, there are several modes of functioning Tag – Anchor, such as:

- Two Way Ranging (TWR), a technique adopted for determining the distance between two devices. One device sends a packet to the second one, which responds immediately. Then, the first device is able to determine the time elapsed between the time instants related to the first transmission and the response reception (the so-called round-trip time). Since the speed of light is known, the distance can be directly estimated from the round-trip time measure. By combining several TWR measurements between anchors and the tag, a tag can be localized.
- Time Difference of Arrival (TdoA), a location technique according to which the tag position is estimated by comparing the time delay between the receptions of signals coming from different nodes (anchors), as done by GPS receivers with satellites signals. The TDoA methodology enables the tracking of an unlimited number of tags within the system and thousands of them within a single area for tracking assets, monitoring forklifts or employee location tracking.

Issues with Data Collection: the data collection process and their results are a challenging task and involve many issues that must be addressed before the data is collected and used.

The main issues in the process of data collection and utilization are:

- The raw collected data may not be in the usable form right away and may require additional efforts or changes to make it usable.
- User privacy issues: Privacy and security data politics. Not all data can be collected and processed. Sometimes you have to apply anonymity algorithms to collect the data.
- The amount of data can cause a system size problem. Data extraction can be extended in time by its quantity or by its sampling.
- Ethical norms about the data: Honesty, Objectivity, Integrity, Non-Discrimination …etc.

Monitoring Systems: Monitoring applications are systems for the massive analysis of data. One of them is ThingSpeak as shown in Figure 9. ThingSpeak™ is an IoT analytics platform service from MathWorks[24]. It can use any Internet-connected device with ThingSpeak. ThingSpeak provides instant visualizations of data posted by your devices or equipment. Execute MATLAB code in ThingSpeak, and perform online analysis and processing of the data as it comes in. ThingSpeak accelerates the development of proof-of-concept IoT systems, especially those that require analytics.

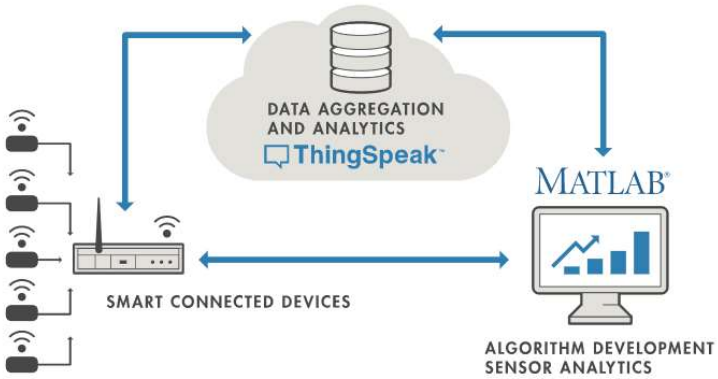

Figure 7 - ThingSpeak as an example of monitoring systems (extracted from[24]).

Connectivity Protocols: Modern connectivity protocols are very diverse depending on technical needs, the system you transmit, and your goal. A plethora of more traditional protocols is in use in industrial systems (field bus / wireless). In client server models, clients communicate directly with the known server. CoAP, MQTT, OPC, XMPP, AMQP, LwM2M, ZigBee, Bluetooth maybe listed as the examples of connectivity protocols. From this diverse set of connectivity protocols, MQTT and OPC are investigated further in this section. These protocols have had a separatist nature in their functions and mainly OPC is applied to the connectivity of machines in the industrial environment. And MQTT is presented as the protocol used for sensor communications in IIoT environments. An industrial environment is illustrated in the Figure 10:

24 “ThingSpeak for IoT.” [Online]. Available: https://thingspeak.com/pages/commercial_learn_more?EnergyMonitoring. [Accessed at June 12, 2020].

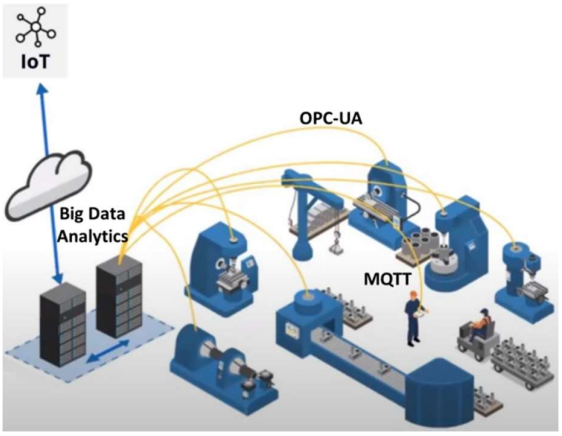

Figure 8 - OPC (machine) vs. MQTT (sensors) Protocols (extracted from[25]).

In which we can see the different technologies that are used: Big Data Analytics, Edge computing, MQTT, OPC-UA. The market presents solutions that unite the strengths of both prevailing connectivity technologies. The slogan that predominates in these implementations is "OPC UA + MQTT = A popular combination for IoT expansion", MQTT - OPC (in the case of the image, we see OPC -DA). We can see an architectural scheme in Figure 9.

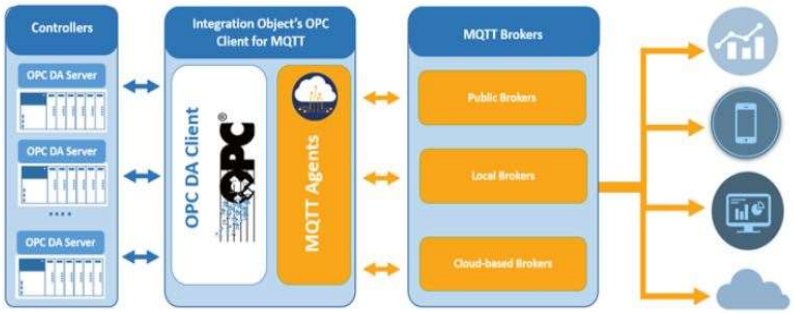

Figure 9- Architectural scheme joining MQTT – OPC (extracted from[20]).

With OPC UA or MQTT communication on the factory floor, it is possible to use a computer, server or edge gateway to relay information to IT systems and clouds [26], as it is shown in Figure 12.

[25] 'IoT Communication with MQTT, HiveMQ, and Azure IoT Edge' [Online], Available: https://www.hivemq.com/blog/azure-iot-edge-and-hivemq/

26 "Use Anybus CompactCom to connect to IoT protocols." [Online]. Available: https://www.anybus.com/products/embedded-index/embedded-features/opc-ua-mqtt/use-anybus-compactcom-to-connect-to-iot-protocols. [Accessed at June 10, 2020].

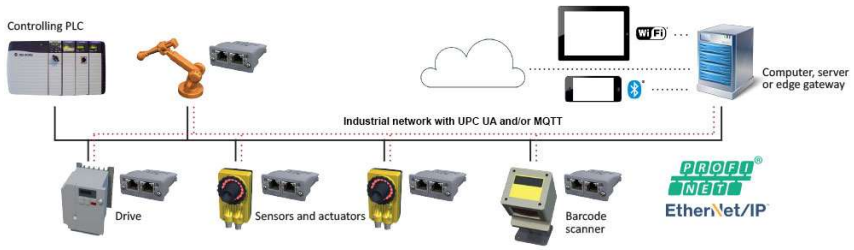

Figure 10 - Example of system (extracted from[26]).

Part 14 of the OPC UA Standard, "Publish-Subscribe" or "PubSub" was released, helping to encourage further interaction between OPC UA and communications protocols such as MQTT. According to the OPC Foundation, "PubSub enables the use of OPC UA directly over the Internet by utilizing popular data transports like MQTT (Message Queuing Telemetry Transport) and AMQP (Advanced Message Queuing Protocol) while retaining its key OPC UA end-to-end security and standardized data modelling advantages". The standard has also helped further communication abilities within IoT devices, such as gateways. IoT gateways provide a bridge between an on-premises communications network and a cloud-based communications network. Sometimes referred to as "Edge devices", IoT gateways provide data connectivity to sensors and end-devices, completely on-premises. Such systems are sometimes also called "intelligent gateways," due to the increase in processing power, memory and storage, as well as intelligent "middleware" encompassing key features such as protocol translators, data buffering, edge analytics, on-premise visualization and cloud enablement, which support emerging communication protocols including MQTT and AMQP for publishing data to the cloud. In the marketplace, companies as powerful as Intel have developed their own gateways. Intel's Internet of Things gateways offer businesses a key ingredient in enabling connectivity from legacy industrial devices and other systems to the Internet of Things. It integrates network technologies and protocols, embedded control, business security, and easy management of what specific application software you can run. Another design of architecture that combines the best of both protocols is given in Figure 11.

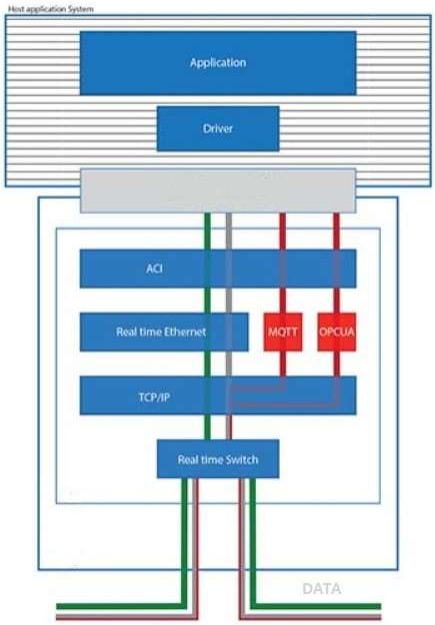

Figure 11 - Another Architecture combining MQTT – OPC (extracted from[26]).

Integrating OPC UA with AZURE: Microsoft is proposing an architecture where factory operation can be made more efficient and productive using OPC UA with Microsoft Azure tools like Time

Series Insights (TSI)[27]. This architecture consists of a series of modular components, which can include any or all of the following:

- An OPC UA Publisher Module to subscribe to OPC UA servers in your factory. As data is published, the Publisher module translates those OPC UA Attribute values to JSON and delivers them to the IoT hub.
- The IoT hub retains the data from the OPC UA servers and makes it available to the collection of data management, analytics and web services in Azure.
- Time Series Insights – In one application of the Connected Factory, IoT hub is an event source for the Time Series Insights service (TSI). Time Series Insights is an Azure service that provides analytics, visualization and data storage.
- In other applications of the connected factory, an OPC UA Client in the Cloud delivers commands to the OPC UA servers in the factory just as a local one would.
- An OPC UA Proxy Module to tunnel those OPC UA command and control messages from the cloud OPC UA Client back to the OPC UA servers using UA authorization and encryption.

Sensors: There are many commercial IoT sensors/systems with different technologies for data collection. For example Beacons, ESP32, ESP8266 and Raspberry PI. In Figure 12, we can see an example IoT topology and indoor positioning with several ESP32 microprocessors with temperature and humidity sensors as clients that connect to a referenced server with a Raspberry PI. It is possible to do a small demo application to use ESP32 modules as anchors/stations to do indoor positioning of iBeacon tags (tagged people, dogs, cats and objects) with trilateration. ESP32 modules will work as iBeacon monitoring stations, reporting all found Bluetooth beacons to the MQTT topic, with their MAC address and RSSI.

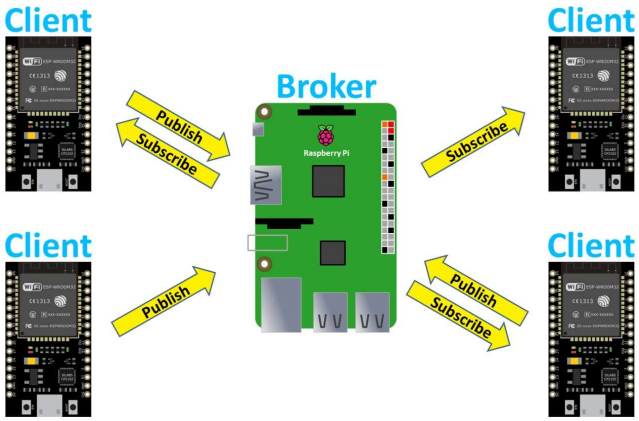

Figure 12 - MQTT sensors topology (extracted from[28]).

3.2. Industrial IoT Platforms

An important data source of industrial sites are integrated sensors to production machines and lines, processes and enterprise systems as well. An Industrial IoT platform is a rapidly growing segment of IoT technology comprising a collection of functions for edge device management, IoT data analytics, modern sensor technologies and connectivity solutions that enhance industrial equipment and industrial operations with remote monitoring, predictive maintenance, and extensive device data analytics. Thus, this section in part overlaps with section 4 Data lake exploitation,

27 “How Microsoft Is Leveraging OPC UA to Get an Irreplaceable Position in Your Factory.” [Online]. Available: https://www.automation.com/en-us/articles/2017/how-microsoft-is-leveraging-opc-ua-to-get-an-irrep. [Accessed at June 10, 2020].

28 ‘Raspberry Pi ESP32 MicroPython MQTT DHT22 Tutorial’ [Online]. Available: https://www.rototron.info/raspberry-pi-esp32-micropython-mqtt-dht22-tutorial/

especially considering real time and non-real time data analytic frameworks. In other words, common IIoT use cases in manufacturing include factory automation for operational efficiency; location tracking for locating tools, parts, and inventory; and predictive maintenance for maximizing uptime and disaster tolerance[29]. On a broader scale, an Industrial IoT platform is a key enabler of Industry 4.0, otherwise known as smart factory, which combines modern cloud computing, IIoT and AI to create intelligent, self-optimizing industrial equipment and production facilities.

Smart manufacturing in Industry 4.0 means not only developing and deploying new systems, but making legacy ones flexible and easily reconfigurable. In order to do that, one of the main objective of Industry 4.0 is connecting these systems into the same network, where all devices i.e. controllers, sensor and other peripherals are IP endpoints, which are addressable and accessible on the network directly. That is what called IoT Automation[30].

- Figure 13 shows the reference IIRA (Industrial Internet Reference Architecture) architecture for industrial IoT systems. Five functional domains of a typical IIoT system: control, operations, information, application, business domains and the flows between those domains are demonstrated. Green arrows show the data/information flows, grey/white arrows show decision flows and red arrows show command/request flows[31].
- Figure 14 shows the architecture that is used in Intel's smart factories. Intel uses gateways and sensors in its factories to collect and analyze that data is used to improve automation, such as defining the control limits that produce high-quality results within a process. Data from edge computing and IoT is extracted for real-time analysis on the factory floor as well as big data analysis in off-line databases[32].

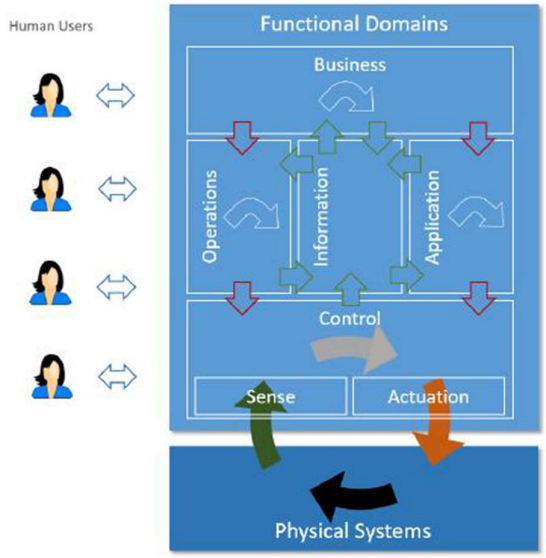

Figure 13 - Reference IIRA architecture for industrial IoT systems (extracted from[31]).

29 'IoT Data Management: The Rise of Industrial IoT and Machine Learning' <https://www.informatica.com/hk/resources/articles/iot-data-management-and-industrial-iot.html> [accessed 6 June 2020].

30 J. Delsing, *IOT Automation. Arrowhead Framework.* CRC Press, 2017.

31 Industrial Internet Consortium, *The Industrial Internet of Things Volume G1 : Reference Architecture*, *Industrial Internet Consortium White Paper*, 2019, VERSION 1.

32 Steve Chadwick and Steven J. Meyer, *Using Big Data in Manufacturing at Intel ' s Smart Factories*, 2016.

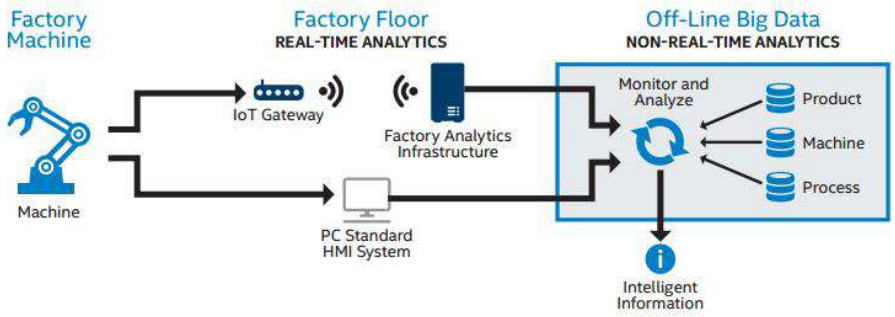

Figure 14 - Data collection and analysis at Intel's smart factories (extracted from[32]).

Different types of data sources in IIoT applications are[29]:

1. Industrial control systems: IoT makes it possible to leverage the data which is already available in SCADA system or historian.
2. Business applications: Data silos are still very common in industrial organizations. Data from applications like your CRM, ERP or EAM can provide context that goes beyond what is wrong with a machine.
3. Wearable: New wearables promise to make difficult and often dangerous jobs safer and easier.
4. Sensors and devices: Advances in sensor technology have made streaming real-time data easier than ever. Temperature, flow, pressure and humidity sensors have become big sources of industrial IoT data.
5. Media: Smartphones have made it possible to get real-time access to photos, videos and audio from the field. However, in industrial cases, we can go beyond using smartphones to upload a picture of a broken machine. One way to use media as a data source in oil and gas is to stream real-time infrared images when inspecting flare stacks. Flare systems need to be inspected regularly for fouling and corrosion.
6. Location: Location data could come from mobile devices, location beacons, GPS systems. The location technologies (RTLS systems) are described in previous sections in detail.

A successful IOT solution consists of a combination of the following 6 capabilities:

- Connect to IoT endpoints
- Manage IoT endpoints/identities
- Ingest and process IoT data
- Visualize and analyze IoT data
- Build IoT applications
- Integrate IoT data into existing applications

Commercial IIoT platforms:

Siemens Mindsphere is the cloud-based IoT open operating system from Siemens. It connects products, plants, systems, and machines, enables to harness the wealth of the data with advanced analytics. In addition, it gives access to a growing number of apps and a dynamic development ecosystem. MindSphere serves software as a service for users[33]:

- Connect assets and upload data to the cloud
- Collect, monitor, and analyze data in real time

33 'MindSphere' <https://siemens.mindsphere.io/en> [Accessed at June 11, 2020].

- Gain insights that improve efficiency and profitability
- Add apps that increase the business value of your data

PTC Thingworx is a complete, end-to-end technology platform designed for the industrial Internet of Things (IIoT). It delivers tools and technologies that empower businesses to rapidly develop and deploy powerful applications and augmented reality (AR) experiences[34]. IBM Watson IoT is a complete solution form data collection from shop floor to unlock the power of data with AI and IoT to innovate asset management, optimize real estate and facilities, improve software and systems engineering, and advance digital transformation of factories[35]. Microsoft Azure is a service which builds new industry solutions, improve productivity, and reduce waste —and quickly process massive quantities of data from all kinds of IoT devices using AI and machine learning[36]. SAP Leonardo is a collection of tools and applications built on SAP Cloud Platform that can be used to create an administrative environment for managing and monitoring sensor data generated by technical objects that are part of the Internet of Things (IoT). In addition to the API services SAP Leonardo IoT offers a number of apps that are built upon the services and help you step inside the world of the Internet of Things[37]. GE Digital Predix Essentials is a complete solution for industrial data monitoring and event management, combining asset connectivity, edge-to-cloud analytics processing, and a feature-rich user console[38].

Proprietary platforms include typically repositories of enterprise-wide raw data, but combined with big data and search engines, a data lake (or enterprise data hub) can deliver impactful benefits. Data lakes bring together data from various sources and make it easily searchable, maximizing discovery, analytics, and reporting capabilities for end-users. Examples of benefits from these type of approaches include such as (Accenture):

- Data richness - as an example ability to store and process structured data (XML, JSON, audio, image, video) and unstructured data (text files, sensor data, surveys) from multiple sources and types.
- User productivity - search is a universal tool for finding information. End-users of the data can get the data they need quickly via a search engine, without SQL knowledge.
- Cost savings and scalability - example open source typically has zero licensing costs, allowing the system to scale as data grows.
- Complementary to existing data warehouses – data warehouse and data lake can work in conjunction for a more integrated data strategy.
- Expandability - data lake framework can be applied to a variety of use cases, from enterprise search to advanced analytics applications across industries

As an example, Bittium PROSE is based on a hybrid micro-service architecture with layered approach. Back-end functionality is split into logical business level micro-services, which form the base for all functionality. For the most part these business services do not implement pure micro-service principle though, as they share common data storage's behind the interfaces and services. This hybrid approach is possible due to relatively modest data traffic and simultaneous usages for

34 'ThingWorx Platform Product Brief' <https://www.ptc.com/en/resources/iiot/product-brief/thingworx-platform> [Accessed at June 10, 2020].

35 'The Internet of Things Delivers the Data' <https://www.ibm.com/internet-of-things?lnk=hpmpr_iot> [Accessed at June 12, 2020].

36 'Azure IoT' <https://azure.microsoft.com/en-us/overview/iot/#overview> [Accessed at June 12, 2020].

37 'Internet of Things (IoT)' <https://www.sap.com/products/intelligent-technologies/iot.html> [Accessed at June 10, 2020].

38 'Predix Platform' <https://www.ge.com/digital/iiot-platform> [Accessed at June 12, 2020].

most of the services. For the more heavily used part, of update packages distribution, pure micro-service architecture is applied. This will take the load off from the bigger part of services and still allow putting additional capacity to the services which are been used most heavily. Access to business services are restricted by authentication and authorization on information level and also controlled by different network segments so that for example the device updates service exposes only the limited information that is required for that purpose.

3.3. Conclusions and Limitations

Factories are complex environments that contain a variety of data sources coming from manufacturing assets, processes, products and so on. Real-time or near real-time asset tracking systems are used to gather data from the shop floor with different technologies. Real time location and industrial IoT systems are two of them, which fed data for developing optimization and resilience capabilities in CyberFactory#1 project.

When RTLS systems are considered, every solution has its own advantages and disadvantages. The selection of the technology strongly depends on the constraints of the operating environment and the requirements. Besides, organizations need to incorporate asset management capabilities that optimize not only their assets' availability, performance, and quality, but also their energy consumption in order to stay competitive in a rapidly changing market. Table 1 presents main advantages and disadvantages of each of the mentioned indoor RTLS technologies.

Table 1 Comparison of indoor RTLS technologies

Technology	Accuracy	Range	Battery Life	Cost	Strengths	Weaknesses
BLE	2-3 m	10-20 m	Good	Low	Ubiquity; Good balance between accuracy and total cost; Scalability	Shorter range than some alternatives (can be increased at the expense of battery)
RFID (active)	2-3 m	20-30 m	Good	Low	Low latency	Not many advantages over BLE
RFID (passive)	2-3 m	2-3 m	N.A.	Very low	Tags don't need battery and are very cheap	Provides only proximity
UWB	Sub-meter	20-30 m	Good	High	Has the best accuracy	High tag and infrastructure price
Wi-Fi	3-5 m	20-30 m	Poor	High	Can use existing infrastructure and devices	Worse than other technologies in almost every aspect
ZigBee	3-5 m	20-30 m	Good	Low	Can use a mesh topology	Not many advantages over BLE

BLE technology offers low cost and low power consumption with an average range in terms of RTLS. If enough number of fixed point devices are used, it also offers good accuracy. In addition, the new standard (version 5.1) has a special focus on indoor tracking by introducing direction estimation (by using AoA method) and improving the advertising mode. This will allow to further improve its accuracy (based on direction feature), scalability and reliability (based on the improved advertising).

UWB technology has proven to be effective in indoor positioning with highest accuracy. They have very good response time and can be more power efficient RF active tags, with similar range but higher accuracy. However, UWB can be susceptible to metal interference and requires complex installation.

Wi-Fi technology has the advantage that most buildings already have an existing WiFi infrastructure in place and reduces investment cost. However, software RTLS would be needed and the accuracy and range would depend on the distribution of the WiFi access points. Increasing the access points increases the coverage and accuracy but will increase the cost and the possibility for interference. Another drawback of Wi-Fi is its poor accuracy (3-5 m).

RFID technology has been around some decades and commonly used instead of barcodes which needs line-of-sight between reader and tag. The absence of battery in passive tags reduces the size and costs which in turn reduces the range to 2-3 m. If passive RFID is considered for RTLS, the limited range eliminates the possibility of proper tracking because it needs a prohibitive number of readers both in terms of cost and space. Another problem is the tag collision, which happens when a reader cannot process simultaneous responses from different tags. Active RFID tags have their own power source, which increases the range while increasing the size and cost of the tag. The major limitation of the active RFID technology is its average accuracy.

ZigBee is a low power RF technology used for RTLS. When compared to WiFi, it has lower power consumption and lower cost at the expense of a lower data rate. It usually uses a mesh topology, unlike WiFi using a star topology, meaning nodes in the network are connected to each other, allowing better scalability and reliability. The biggest drawback of ZigBee is the poor accuracy (3-5 m) and can be used only in room level tracking. IIoT platforms have gained importance due to their

functionality in production environments. They connect information systems, secure the data within the platform and of the interfaces, manage the processes and systems connected, analyze the data that is inside the platform to provide meaningful information, and build apps and agents to support user decision. Table 2 summarizes the strengths of leading IoT platforms.

Table 2 Comparison of IoT platforms [39]

	Connectivity	Integration	Analytics	Application	Security
Thingworx	x		X	X	
Watson IoT			X		X
Azure		X	X	X	
SAP Cloud			X	X	X
GE Predix			X		X

Azure manages to integrate enterprise systems, non-IP and IP capable devices by the help of its IoT hub and logic apps tools. Almost all platforms provides analytics tools. The SAP cloud platform shows advanced capabilities in application, analytics and security fields and also features an Intuitive user interface. The ThingWorx platform offers advanced solutions in analytics and application field. It has predictive and prescriptive analytics features with high support of augmented reality application. The Predix platform offers advanced capabilities in application and analytics fields. The analytics tool of Predix is capable of big data analysis, anomaly detection, identification of trends, visual recognition, analysis of texts, machine learning, and the application of many other analytical algorithms. The Watson IoT platform shows advanced proficiencies in analytics and security fields. The analytics capabilities of the platform are based on predictive, cognitive, real-time and con-textual methods.

This section of the document provided the state-of-the-art for RT and near RT location, tracking and sensing technologies/methods. Each technology has its pros and cons. Then, the selection for an application depends on the requirements of the system and the environment it will be implemented. In addition, the performance of the tracking system is very dependent on the application environment. The project has ten use cases that means very different manufacturing environments, from cheese manufacturing to aircraft part manufacturing, will be under consideration. Prominent real-time monitoring technologies will be deployed to different production environments and relevant development will be made in order to get valuable data from shop floor. Also, tailor made solutions will be developed for different industries. The obtained data is not limited to the geolocation data obtained from RTLS but also the data obtained from real-time / near-real-time tracking of manufacturing assets, materials and processes with IIoT platforms, which are in the heart of all kind of factories. The data obtained from RT tracking systems will be used for further simulation, optimization and security-related developments in other work packages of the project.

39 Jorg B. Hoffmann, Pit Heimes, and Semih Senel, 'IoT Platforms for the Internet of Production', *IEEE Internet of Things Journal*, 6.3 (2019), 4098–4105 <https://doi.org/10.1109/JIOT.2018.2875594>.

When it comes to the limitations of real time tracking and indoor positioning technologies, it may be argued that characteristics of these various techniques closely related to each other and only one aspect (accuracy, cost, precision, efficiency, performance, etc) can be optimized within one given implementation. Cyberfactory#1 project aims to tackle these drawbacks of the RTLS and IPS systems with characteristics better applying to specific usecases with improved optimization techniques such that low cost solutions are possible with acceptable accuracies while still offering good performance. Traditional on premise solutions for data collection and analytics that do not meet the requirements of the amount of data generated with advanced sensing and tracking techniques will be overcome through novel the data lake architectures that combine inputs from different types of data sources incorporated to the manufacturing processes.

New manufacturing machines can transmit data about production processes via wired or wireless connections thanks to IIoT. However, the majority of the machines in today's factories do not have such features we have just mentioned because the lifespan of the machines used in factories is decades and their replacement costs are very high. These machines either do not generate information about their production processes or cannot export the information they produce to an external database. This results in the lack of knowledge about production, limiting the possibilities to develop solutions for optimized and secure manufacturing.

With the real-time tracking and monitoring solutions developed in the Cyberfactory # 1 project, it will be ensured that information that could not be obtained from legacy machines and processes in factories before will be collected and stored. In the project, new platforms will be developed / integrated to capture and store data generated in the production machinery and related systems. In addition, new data sources will be created by integrating new sensors into legacy machines and systems. Thus, real-time or near real-time data will be obtained, transmitted and processed for new optimization and security services. For example, geolocation data of service robots used in production will be obtained in real time and security solutions will be developed (geofencing). The information obtained from all systems will also be accumulated in a data lake. This will allow to reveal hidden patterns in the data using big data processing methods (elastic search, machine learning etc) and to develop optimization, predictive and security services. In conclusion, the project will help to add new data sources from the shop floor in industrial sites, collect data in real-time or near real-time as much as possible and develop optimization and security services based on this new data sources.

4. Data lake exploitation

Data Lake is an umbrella term used to refer to a massively scalable storage repository, where data arrives at high velocity rate and in a native format (structured, unstructured or hybrid). These data storages are equipped with engines capable of querying and formatting data on the fly according to the users specificities and needs (schema-on-read). In the Industry 4.0 scenario[40], most of the data is context data generated by sensors, connector, controllers, manufacturing systems, people and many others.

Some of the data needs to be processed in real time (stream processing). In this case, the data is not fully stored, and the processing time is constrained to the incoming data velocity. It is usually used or implied on systems where decisions need to be taken in a critical time, for example in a robot navigation system, security systems, etc. Others store and process the data in batches with soft time constraints, the data is formatted and validated according to the objective function, for example predict the equipment maintenance, failures in the production, etc. The lakes must be dynamic in order to accommodate the highly mutable data[41]. Regardless of the data type or source,

40 Piccarozzi, Michela & Aquilani, Barbara & Gatti, Corrado. (2018). Industry 4.0 in Management Studies: A Systematic Literature Review. Sustainability. 10. 3821. 10.3390/su10103821.

41 Miloslavskaya, Natalia & Tolstoy, Alexander. (2016). Application of Big Data, Fast Data and Data Lake Concepts to Information Security Issues. 10.1109/W-Fi Cloud.2016.41.

one of the Industry 4.0 goals is to extract value from the data, in order to gain competitive advantage in the global market.

4.1. Data Lake Technologies

Kylo data lake, is an open-source data lake platform. It is built on Apache Hadoop and Spark. Kylo provides key services such as self-service data ingest data preparation and data discovery. Kylo is web application installed on a Linux "edge node" of a Spark and Hadoop cluster. Kylo uses Apache NiFi as its scheduler and orchestration engine, it offers flexibility to create more than 200 processors (data connectors and transforms). Kylo can be integrated with Apache Ranger or Sentry and CDH Navigator or Ambari for cluster monitoring[42].

Most of the Kylo UI depends on AngularJS and AngularJS material running in a Tomcat container. Furthermore, Kylo uses JBoss ModeShape and MySQL (or Postgres) to store metadata. It operates over RHEL, CentOs, SUSE, Ubuntu systems. Kylo is supported by Cloudera, Hortonworks, Map R, EMR, and vanilla Hadoop distributions. Kylo requires at least four core CPUs and a minimum of 16GB of RAM.

Elasticsearch[43] is a document-oriented database. The search engine is based on the Lucene library. It is open and free to the community. Elasticsearch runs on Amazon EC2, or on Amazon Elasticsearch Service. Elasticsearch manages data from different formats, such as structured or unstructured text, numerical data, categorical data or geospatial data. It is data-distributed oriented, and it offers analytic services in real-time. This is due to the fact, that instead of searching the key-words, it searches based on an index. Additionally, it supports full-text search, which is completely based on documents instead of tables or schemas.

Apache Kafka[44] is a real-time data stream, for storing, reading and analysing data stream. Apache Kafka organizes and categorizes information into topics, using the function Producer, which is an interface between applications and topics, Kafka Topic Log is a database for ordering and segment the data. Another important interface is Consumer, that allows topic logs to be read, and the information stored in them can be passed onto other applications.

Delta lake[45] is an open-source data storage layer that delivers reliability to the data lake, Delta lake implements ACID transactions, scalable metadata handling, and unifies the streaming and the batch processing. Delta lake leverages metadata through a Spark, thus it can handle petabyte-scale tables with billions of partitions and files. Delta lake keeps some sort of historical consistency, which means that data can be reverted to old formats or old experiments can be reproduced. Delta lake compresses data efficiently using an Apache Parquet format, and it can enforce schemas, this ensures that the data types are correct and required columns are presented, preventing bad data from causing data corruption. These schemas are not static, and they can be reconfigured any time.

---

42 https://kylo.readthedocs.io/en/v0.10.0/. [Accessed at March 4, 2021]

43 U. Thacker, M. Pandey and S. S. Rautaray, "Performance of elasticsearch in cloud environment with nGram and non-nGram indexing," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 3624-3628.

44 B. R. Hiraman, C. Viresh M. and K. Abhijeet C., "A Study of Apache Kafka in Big Data Stream Processing," 2018 International Conference on Information , Communication, Engineering and Technology (ICICET), Pune, 2018, pp. 1-3.

45 M. Armbrust, T. Das, L. Sun, B. Yavuz, S. Zhu, M. Murthy, J. Torres, H. van Hovell, A. Ionescu, A. Łuszczak, M. Świtakowski, M. Szafrański, X. Li, T. Ueshin, M. Mokhtar, P. Boncz, A. Ghodsi, S. Paranjpye, P. Senster, R. Xin, and M. Zaharia. 2020. "Delta lake: high-performance ACID table storage over cloud object stores". Proc. VLDB Endow. 13, 12 (August 2020), 3411–3424.

Hadoop[46] is an open source, Java based framework used for storing and processing big data in a distributed system. Several modules such as the HDFS, Thrift, Hive, MapReduce and Spark compose the Hadoop Ecosystem. These components can be used to create a data lake, for example, using Hadoop in a data lake architecture source data and curated data can co-exist together.

City Cloud[47] is a cloud service and it can be an important component of a data lake architecture. The City Cloud offers network services, hardware services, and monitor dashboard. It is compatible with the most of OpenStack APIs. Apache Nifi[48] is another tool that can be useful in a data lake architecture, Apache Nifi manages efficiently data flow between systems, it allows data visualization, data routing, data transformation and system mediation through intuitive graphs. It works as a Kylo scheduler and orchestration. NiFi is an open-source system, and it is executed inside of JVM hosted in the host operating system. The web server handles the HTTP-based command and controls the API, and Flow Controller is the core part, which is focus on managing the schedule and receiving resources to execute. The FlowFile Repository is where NiFi keeps track of the state of what it knows about a given FlowFile that is presently active in the flow[49]. The Content Repository is for storing the actual data content, and the Provenance Repository is for storing all provenance event data.

4.2. Data Lake Use Cases

There are several ways to use a data lake, it can be used to store big data[50], in another words, data that comes from equipment reading, telemetry data, logs, etc. It can be used to store IoT (Internet of Things) data types, it can store data stream used in near real-time analytics[51]. It can be useful for data scientists, for example, the data science lab is a platform for data scientist. It allows the processment of data, merging and combining data from different data sources for a large set of different task or applications such as financial, wealth, social, vendors, network, etc. The data science lab has a custom data lake infrastructure, which is based on object storage and the Apache Spark™ execution engine and related tools contained in Oracle Big Data Cloud. It uses Oracle Analytics Cloud for data visualization, data preparation and extract data relationships in the data lake. Furthermore, Oracle Database Cloud Service is used to manage metadata.

It an enterprise level, data lake can be used as a staging area of a data warehouse, this can be achieved using ETL (Extract-Transformation-Load) engines in order to transform non-structured data into highly structured data[52]. Depending on the level of transformation needed, offloading that

---

46 S. G. Manikandan and S. Ravi, "Big Data Analysis Using Apache Hadoop," 2014 International Conference on IT Convergence and Security (ICITCS), Beijing, 2014, pp. 1-4.

47 https://citycontrolpanel.com/apidoc/index.html. [Accessed at March 4, 2021].

48 https://nifi.apache.org/docs.html. [Accessed at March 4, 2021].

49 A. Pandya, et al., "Privacy preserving sentiment analysis on multiple edge data streams with Apache NiFi," in 2019 European Intelligence and Security Informatics Conference (EISIC), Oulu, Finland, 2019 pp. 130-133.

50 S. Munirathinam, S. Sun, J. Rosin, H. Sirigibathina and A. Chinthakindi, "Design and Implementation of Manufacturing Data Lake in Hadoop," 2019 IEEE International Conference on Smart Manufacturing, Industrial & Logistics Engineering (SMILE), Hangzhou, China, 2019, pp. 19-23, doi: 10.1109/SMILE45626.2019.8965302.

51 R. Liu, H. Isah and F. Zulkernine, "A Big Data Lake for Multilevel Streaming Analytics," 2020 1st International Conference on Big Data Analytics and Practices (IBDAP), Bangkok, Thailand, 2020, pp. 1-6, doi: 10.1109/IBDAP50342.2020.9245460.

52 Campbell, Chris. "Top Five Differences between Data Warehouses and Data Lakes". Blue-Granite.com. Retrieved 19 May 2017.

transformation processing to other platforms can both reduce the operational cost and free up data warehouse resources in order to focus on its primary role of serving data. Oracle's Data Integration Platform Cloud (DIPC) is an example of a engine used to extract, load and transform data for the data warehouse. The data lake can be used to augment the data warehouse. The data lake might be accessed via federated queries, which make its separation from the data warehouse transparent to end users through a data virtualization layer. Data Lake can be used also to store all the organizational data in order to support downstream reports and analytic activities[53]. Some organization use the data lake as a unique repository for all types of data. Usually, the goal is to store as much data as possible in order to support analytics that might yield valuable findings. Finally, from the application point-of-view, data lake can be a data source for a front-end application, it can also act as a publisher for a downstream application.

4.3. Big Data Technologies

In the literature, Big data, Data Lake and Fast data are concepts highly correlated and interrelated. Their differences mostly reside on user or application specificities. Big data technologies are built over five key elements: velocity, volume, value, variety, and veracity. The term velocity is related to the speed at which the data is generated, collected and analyzed, the ability to process data faster them competitors is an important competitive advantage, since it can help decision makers to make faster and more accurate business decisions. The amount of data that can be processed is also an important factor, this is due to the data dependency of machine learning algorithms, such as deep neural networks. These algorithms extrapolate more accurately when data is available in large amounts and variety. The data itself adds value to the company, it gives to the company competitive advantages with respect to competitors. Furthermore, useful insights and value can be derived and added from data aggregation and exploitation techniques, when using data from different cross-domains. The last key component is veracity or validity, and it reflects the data quality. Algorithms are expected to perform better when the data is clean and when the target variables are annotated accurately. This can be achieved by cleaning and filtering processes, and through the analysis and agreement of different raters or expert annotators.
To the best of our knowledge, mostly of the Big data technologies could be also used to create a data processment pipeline for a data lake infrastructure. Thus in this section, an overview of Big data technologies is presented, and their re-usability and migration to data lake infrastructure is explored. An extensive literature review of the big data technologies and their requirements on two case studies (railways and wind turbines) could be found in [32]. Some big data processing pipelines have already been proposed, these are divided in the following steps: distributed queuing, big data stream platforms, big data storage and stream SQL engines.

Distributed queuing systems are key tools, since they manage the messages from the consumers and the producers. The recent distributed system supports several consumer groups online and more importantly prevent data loss by utilizing persistent disks over replicated clusters. Some distributed management technologies are Kafka, RabbitMQ, Amazon Kinesis, Hubs and Google Pub/Sub. A full description of these technologies is presented in[54]. Following the rationale in [54] federated queues is the most important feature, since it provides a mechanism for balancing the load of a single queue across several nodes. It also supports migration among nodes without stopping the producers and the consumers. Currently the only technology that supports federated queues is RabbitMQ.

Big data platforms can be roughly divided in those specialized in batch processing or stream processing. Batch processing refers to the action of processing blocks of data that have already

53 Fang, Huang. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. 820-824. 10.1109/CYBER.2015.7288049.

54 Sahal, Radhya & Ali, Muhammad Intizar & Breslin, John. (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. 10.1016/j.jmsy.2019.11.004.

been stored in memory for a large periods of time, Hadoop MapReduce [55] is an example of a framework tailed for batch processing. On the other hand, stream processing requires analytic results in real-time. There are several open source stream processing platforms such as Apache Kafka, Apache Flink, Apache Storm, Apache Samza. The most important feature to explore is the presence of a fault tolerance mechanism. Some authors implement a check pointing RDDs, such as Spark, others use Checkpointing such as Samza, which seems to be a reasonable choice[56], since it processes a stream of data, it is scalable, it can use RabbitMQ as queuing management system, and it offers an interactive shell, although the learning curve of the programmer and maintenance might be an issue. In order to accommodate the large variability of data formats and requirements, several data storage technologies are at disposal. To handle data streams in a real-time fashion into a NoSQL and batch data: column-based storage (HBASE, Cassandra), document-based (MongoDB), CouchDB, DynamoDB, Riak, Redis, Neo4J are good choices. In [57], Cassandra is selected do to its compatibility with Presto but it is also chosen together with HBase for a large big data storage. To handle big data in a more historical perspective, Hadoop Distributed File System (HDFS) is in general a good option. The HDFS is a distributed file system that handles large data sets running on commodity hardware. The HDFS is a “master-slave” architecture i.e. a Master node that manage the file's namespace and adjust client access files and all the other nodes, and Slave nodes that are used to manage the storage. In Hive, a technology used for the Big data warehouse, built on HDFS is used to store and distribute large volumes of static data. From our current analysis, the data storage technology is highly dependent on the application requirements and the data type.

In order to be able to extract mining information from high velocity of data, some technologies have been proposed, such as Spark SQL, Table API, KSQL, PipelineDB, Squall, StreamCQL, SamzaSQL, StormSQL, Siddhi and Athemax. The most important feature is the windowing, knowledge is extracted from the streaming data, by applying a sliding window over the time series, inside this window a mathematical operation is performed. Afterwards, the outputs are aggregated and joined. There are five type of windows: Tumbling, Sliding, Hopping, Session and Snapshot window, a fully description can be found in[58] .

4.4. Proprietary Platforms

Proprietary platforms typically include repositories of enterprise-wide raw data, but combined with big data and search engines, a data lake (or enterprise data hub) can deliver impactful benefits. Data lakes bring together data from various sources and make it easily searchable, maximizing discovery, analytics, and reporting capabilities for end-users. Examples of benefits from these type of approaches include such as (Accenture):

Data richness- As an example ability to store and process structured and unstructured data from multiple sources and types, including XML, text, JSON, audio, image, video, etc.

User productivity- search is a universal tool for finding information. End-users of the data can get the data they need quickly via a search engine, without SQL knowledge.

---

55 Dean, Jeffrey & Ghemawat, Sanjay. (2004). MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM. 51. 137-150. 10.1145/1327452.1327492.

56 M. Pathirage, J. Hyde, Y. Pan and B. Plale, "SamzaSQL: Scalable Fast Data Management with Streaming SQL," in 2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Chicago, IL, USA, 2016 pp. 1627-1636.

57 Santos, Maribel & Sa, Jorge & Andrade, Carina & Lima, Francisca & Costa, Eduarda & Costa, Carlos & Martinho, Bruno & Galvão, João. (2017). A Big Data system supporting Bosch Braga Industry 4.0 strategy. International Journal of Information Management. 10.1016/j.ijinfomgt.2017.07.012.

58 Kleppmann, M. (2017), Designing Data-Intensive Applications , O'Reilly , Beijing .
Copy citation to your local clipboard.

Cost savings and scalability- Example open source typically has zero licensing costs, allowing the system to scale as data grows.

Complementary to existing data warehouses – data warehouse and data lake can work in conjunction for a more integrated data strategy.

Expandability- Data lake framework can be applied to a variety of use cases, from enterprise search to advanced analytics applications across industries

As an example, Bittium PROSE is based on a hybrid micro-service architecture with layered approach. Back-end functionality is split into logical business level micro-services which form the base for all functionalities. For the most part, these business services do not implement pure micro-service principle though, as they share common data storage's behind the interfaces and services. This hybrid approach is possible due to relatively modest data traffic and simultaneous usages for most of the services. For the more heavily used part, of update packages distribution, pure micro-service architecture is applied. This decreases the load on the bigger part of services and still allows putting additional capacity to the more services heavily used. Access to business services are restricted by authentication and authorization on information level and also controlled by different network segments so that for example the device updates service exposes only the limited information that is required for that purpose.

4.5. Intelligent Techniques

The gradual implementation of big data and the Internet of Things in a modern work environment is revolutionizing the business sector. The concept of the factory of the future refers to a new way of organizing production materials and responding to the impact of globalization on our economic existence in the form of the fourth industrial revolution. Industry 4.0 implements new methods of data processing, intelligent software, and sensors to fully digitize the value chain so that suppliers and consumers can monitor what is happening inside the company in real time. By deploying the right combination of AI technologies, namely statistic methods, machine learning, deep learning, producers can increase efficiency, increase flexibility, accelerate processes, and even achieve self-optimized operations. Using AI can reduce the conversion cost of the manufacturer[59]. Producers can use AI to develop and produce innovative products that target specific customers and deliver these products with shorter delivery times, resulting in more sales. Therefore, AI is an indispensable part of the factory of the future, and its technology will enhance the flexibility of the factory structure and process. AI enables computers and machines to perform tasks in an intelligent manner. It can help producers to determine the best sequence of operations to achieve their goals and enable them to manage operations remotely in real time. AI is widely used, since many industry leaders expect AI to transform processes end to end in the value chain, including engineering, procurement, supply chain management, industrial operations (production and related functions), marketing, sales, and customer service. AI enhances rather than replaces existing levers that producers use to continuously increase productivity. It is one of the main technical building blocks of Industry 4.0. In addition, manufacturers can use AI to enhance traditional efficiency levers, such as automation and lean management. For example, by identifying the root cause of quality problems and thereby helping to eliminate defects, artificial intelligence supports lean management efforts to reduce waste. The use of AI will greatly change the composition of the workforce and reduce conversion costs because it reduces the need for manual activities in the production process. For example, tasks related to quality control now require a high degree of human involvement, will be highly automated, and have extensive AI support. AI represents a paradigm shift in factories. Today's factories use rule-based methods to automate processes and machinery, and today's robot programming solves a fixed set of situations. In contrast, factories in the future will use AI support to automate processes and machines to respond to unfamiliar or unexpected situations by making informed decisions. As a result, the technical system will be more flexible and adaptable. For example, under the rule-based method, the robot cannot identify and select the required parts from the unclassified parts box because it lacks the detailed programming necessary to cope with the various possible directions of the parts. In contrast, AI-enabled robots can pick the required parts

59 M. Keen, "Successful applications of AI in manufacturing industry," IEE Colloquium on Industrial Applications of AI (Artificial Intelligence), (Digest No.014), London, UK, 1992, pp. 5/1-5/4.

from unsorted materials, regardless of their direction. AI-techniques can be divided into two categories: statistical methods and machine leaning algorithms. Deep learning is a subsection of machine learning method.

4.6. Concepts and techniques for machine learning and prediction of production dynamics

The field of pattern recognition is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories[60]. Pattern recognition has its origins in statistics, while the modern approaches to pattern recognition include the use of machine learning, due to the increased availability of big data. Statistical pattern recognition has been used successfully to design the recognition systems. In statistical pattern recognition, a pattern is represented by a set of *d* features, which is viewed as a d-dimensional feature vector. The goal is to choose those features that allow pattern vectors belonging to different categories to occupy compact and disjoint regions in a d-dimensional feature space. The effectiveness of the representation is determined by how well patterns from different groups can be separated. In the statistical decision theoretic approach, the decision boundaries are determined by the probability distributions of the patterns belonging to each class, which must either be specified or learned. Well-known concepts from statistical decision theory are utilized to establish decision boundaries between patterns. The recognition system is operated in two modes: training and testing. The role of the preprocessing module is to segment the pattern of interest from the background, remove noise, normalize the pattern, and any other operation which will contribute in defining a compact representation of the pattern. In the training mode, the feature selection module finds the appropriate features for representing the input patterns and the classifier is trained to partition the feature space. The feedback path allows a designer to optimize the preprocessing and feature selection strategies[61]. The analytic results based on current data, will help to predict the future behavior of the system and hence to plan or suggest actions to take for optimal outcomes. In order to perform the predictions, an appropriate model is required. In the following a model for an exemplarily factory context is described followed by an introduction to machine learning techniques for predictive modelling analysis.

Statistics and Data Mining Approaches:

It is not easy to distinguish data mining from statistical approaches, since there is not a concrete definition of those terms, and it depends on the background and the point of view of the reader. According to[62], “Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. It shares common ground with Artificial Intelligence, Machine Learning, Pattern Recognition and Data Visualization. It is an interdisciplinary subfield of computer science and statistics; it evolves database and data management aspects, such as data pre-processing, model creation and visualization, inference, etc. In the data mining, the extraction of patterns and knowledge from a group of data records is entitled cluster analysis, in an unusual records is entitled anomaly detection, or in time series data (sequential pattern mining). Data mining techniques are often used to support decisions. Perhaps some of the most well-known examples

60 Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

61 A. K. Jain, R. P. W. Duin and Jianchang Mao, "Statistical pattern recognition: a review," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

62 Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery: an overview. Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, USA, 1–34.

are E-commerce websites; many of them use Data mining to offer cross-sells and up-sells through their websites.

On the other hand, statistic is a more general term, and concerns with the collection, organization, analysis, interpretation and presentation of the data. According to[63], statistics should be defined as set of tools and topics/problems, such as probabilistic theory, real analysis, decision theory, Markov chains, ergotic theory, etc. Statistics focus on the data, including the planning of the data collection in terms of design a survey or/and an experiment. There are two main methods in statistics: descriptive statistic, which summarizes the data from a dataset (e.g. mean and standard deviation) and inferential statistics, which extract and validate hypothesis from the data that a subject to a random data perturbation. Statistical analysis are very often in the Industry, manage and measure the quality level of a production line, measure the employee efficiency, forecast future demands through the analysis of the past statistical trends.

4.6.1. Mathematical and statistical modeling

A model is a representation of a system and it is made of the composition of concepts. The model helps to describe the system and also to study different components effects, and to make predictions about future behavior of system. Mathematical modeling is the art of translating problems from an application area into tractable mathematical formulations whose theoretical and numerical analysis provides insight, answers, and guidance useful for the originating application[64]. It is successful in many applications by providing precision and direction for problem solution. Mathematical models provide a thorough understanding of the system modeled which will enable to improve the system design or control of it.

A statistical model as a mathematical model, embodies set of statistical assumptions in regard to the accessible historical data[65]. A statistical model represents the data generating process. What distinguishes a statistical model from other mathematical models is that it can be non-deterministic. In a statistical model, which is specified with mathematical equations, all the variables do not necessarily have specific values, and some of variables have probability distributions instead; i.e. some of the variables are stochastic.

Factory Model Setting:

Considering a factory set up with a fleet of transport robots and a set of machines and stores (both are just sources and sinks), where the robot fleet fulfills the task of transporting goods between machines and stores. The information about all the machines and stores in addition to the information about observed data on previous transport jobs, e.g. a list of all transport jobs of the last month, is accessible. In the setup factory, the first goal is to find an appropriate model that best represents the task generating data, which can be performed by applying data processing and pattern recognition techniques upon information from historical runs. The more data is gathered by simulation, the better accuracy of matching configured patterns to data and more accurate

[63] BROADFOOT, L. (2001). Graphical Analysis of Multiresponse Data: Illustrated with a Plant breeding Trial, by K. E. BASFORD & amp; J. W. TUKEY. xvii 587 pp. Boca Raton, Florida: Chapman & Hall/CRC (1999). The Journal of Agricultural Science.

64 Neumaier A. (2004) Mathematical Model Building. In: Kallrath J. (eds) Modeling Languages in Mathematical Optimization. Applied Optimization, vol 88. Springer, Boston, MA.

65 Adèr, H.J. (2008), "Modelling", in Adèr, H.J.; Mellenbergh, G.J., Advising on Research Methods: a consultant's companion, Huizen, The Netherlands: Johannes van Kessel Publishing, pp. 271–304.

prediction of future behavior of machines in generating tasks. Each of the transport tasks is illustrated by the following elements:

- m: Owner of the task, the machine that generated the transport request. $m \in M$, while M is set of all machines that generate tasks t: when the transport task was generated, i.e. when the fleet of robots first knew about the task.
- t: When the transport task was generated, i.e. when the fleet of robots first knew about the task.
- a: Source location, where the goods during current task should be picked up from, it can be location of either a machine or a store. $a \in A$, while A is set of all possible source locations.
- b: Sink location, where the goods during current task should be transported to, it can be location of either a machine or a store. $b \in B$, while B is set of all possible sink locations.
- t_{pick}: Time that the transport goods were picked up from the source location to fulfill the current task.
- t_{reach}: Time that the transport goods were reached at the sink location to fulfill the current task.

Therefore, each task will be referred as a vector,

$$task_i = \begin{bmatrix} m \\ t \\ a \\ b \\ t_{pick} \\ t_{re} \end{bmatrix}$$

$i \in I$, I is set of all simulation runs.

The set $TASK$ includes all n number of generated tasks during a simulation run:

$$TASK = \{task_i \mid i \in [1, n]\} \qquad \textit{Equation 1 .}$$

Predictive Model Analysis:

Predictive analytics is the process of using data analytics to make predictions based on historical data. This process is applicable by using data analysis, data mining, statistics, machine learning, and deep learning techniques to create a quantitative predictive model, which will help to forecast future events, as shown in Figure 15. Typically, the data from system behavior in the past is used to build a mathematical model, where this model captures the more important trends. It can be used on current data of systems behavior to let predict what will happen next or to find the actions to optimize the outcomes. The supervised machine learning techniques are used to predict a future value (how long this machine can run before requiring maintenance) or to estimate a probability (how likely it is this customer to default on a loan)[66].

The enhancement of predictive web analytics calculates statistical probabilities of future events in an online manner. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences and exploiting them to predict (see Figure 15)[67]. One common usage of predictive models is for weather and energy load forecasting to predict energy demands in future. The energy producers need accurate forecasts of energy load to make decisions for managing loads in the electric grid. Vast amounts of data are available, and using predictive analytics, grid operators can turn this information into actionable insights.

66 https://www.mathworks.com/discovery/predictive-analytics.html (Accessed:15.07.2020)

67 Baierl, R., Behrens, J., & Brem, A. (2019). *Digital Entrepreneurship*. Springer.

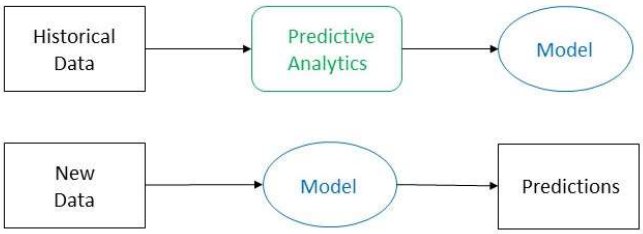

Figure 15 - Prediction Analytics diagram.

In order to use historical complex data to make an actionable information or decision, the predictive analytics should proceed the following steps and find the answers to the corresponding question:

- Descriptive, i.e. what happened.
- Diagnostics, i.e. why did it happen.
- Predictive, i.e. what will happen.
- Prescriptive, i.e. what should be done.

Having a prediction of the task generation in the setup of the factory described above will help to plan the transport robots in a more efficient manner. Especially a prediction of transport job requests in a given time frame and a prediction of phases with really low or high number of transport needs are important. A transport plan may include the charge schedules, parking location assignments, and maintenance dates. The workflow for a predictive analytics application can be defined by the following basic steps:

- Access to and import historical data.
- Clean and preprocess the data, e.g. remove outliers, or identify data spikes and possible missing data points.
- Merge data.
- Perform time-series modeling to extract important predictors.
- Develop an accurate predictive model based on the aggregated data using statistics, curve fitting tools or machine learning. When the training is complete, the model should be tried against new data to see how well it performs.
- Integrate the model into a factory task generation system.

Another example of Predictive Analytics application is in retail trade. Different retail companies always need to improve their sale. One of the most prevalent examples is in their recommendations, which is based on prediction about customers' requirements. When the customer makes a purchase from the website of one retail companies, the company will offer a list of other relevant items.

Shmueli and Koppius[68] highlighted the need to integrate predictive analytics into information systems research. They describe six roles for predictive analytics: new theory generation, measurement development, comparison of competing theories, improvement of existing models, relevance assessment, and assessment of the predictability of empirical phenomena. Since explanatory power does not imply predictive power and thus predictive analytics are necessary for assessing predictive power and for building empirical models that predict well.

Predictive Model Building

To extract the insights holding by historical data, an accurate predictive model is needed. Predictive modeling uses mathematical and computational methods to predict an event or outcome. These models forecast an outcome at some future state or time, based upon changes to the model inputs.

68 Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly, 35*(3), 553-572. doi:10.2307/23042796.

Using an iterative process, the model will be developed by using a training data set and then test and validate it to determine the accuracy of predictions[69]. Examples include time-series regression models for predicting traffic flow or predicting fuel efficiency based on a linear regression model of engine speed versus load.

The predictive analytics will help to predict demands and foresee if a change will help the setup factory to reduce risks and to improve operations. The main objective is to answer the question, what is most likely to happen based on the current historical data and what can change the performance. The model should learn from previous runs data from factory. For example, the model might look at historical data of a specific machine task generation. By establishing the right controls and algorithms, the model can be trained to look at how many tasks were generated on a certain time or were performed from specific source to specific sink location and correlate that data into predictions about future task generation by that machine. In another word, to steady one of the elements in task vector, and analyze the pattern in other elements. The predictive analytics model should be able to identify patterns about the flow of generating tasks in factory and their details. Also, it is critical to regularly retrain the learning model. It is always possible that the actual patterns change due to factors like the time of year or changes in the factory production.

A schematic of the model building steps in explanatory and predictive modeling is shown in Figure 16. Although the main steps are the same, within each step a predictive model dictates different operations and criteria[70].

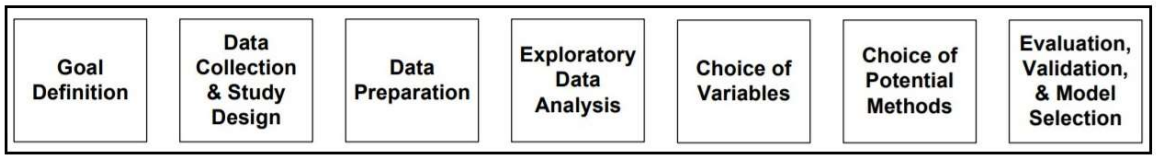

Figure 16 - Schematic of the Steps in Building an Empirical Model (Predictive or Explanatory) (extracted from[71])

Predictive modelers use a variety of tools to explore and analyze the data. Most analytical tools offer some exploratory capabilities. Basic tools enable analysts to compile descriptive statistics of various fields (e.g., min/max values and standard deviation), while others incorporate more powerful data profiling tools that analyze the characteristics of data fields and identify relationships between columns within a single table and across tables.

The basic process of creating analytic models involves running one or more algorithms against a data set with known values for the dependent variable that are interested to predict. Then, the data set is cut in half, one set for creating a training model and the other set for testing the training model[72]. For the setup factory, an algorithm, which trained by the tasks of the last few month, can be developed to predict the task upcoming in a specific time frame. Then, the resulting training model is applied against the other part of the database to see how well it predicts which tasks actually have been generated. Last, the model should be validated in real time by testing it against live data.

The process of training, testing, and validation is iterative. Meanwhile, different combinations of variables should be identified and tested to see which ones have the most impact. In order to start

---

69 https://www.mathworks.com/discovery/predictive-analytics.html [Accessed at July 16, 2020].

70 Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly, 35*(3), 553-572. doi:10.2307/23042796.

71 Shmueli, G., & Koppius, O. (2011). Predictive Analytics in Information Systems Research. MIS Quarterly, 35(3), 553-572. doi:10.2307/23042796

72 Eckerson, W. W. (2007). Predictive analytics. Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report, 1, 1-36.

the process, using statistical tools can help with identifying significant trends in data, e.g. time of the day. Consulting with expert people in the same field, and studying the previous researches may lead to find the appropriate and important variables too. These are the variables that should be implemented in the result model. In next steps, a variety of algorithms should be tested to see which one works best on training data.

The approaches for conducting predictive analytics can be classified into machine learning techniques and regression techniques. Machine learning techniques are popular in conducting predictive analytics due to their outstanding performance in handling large datasets.

Predictive Modeling using Machine Learning

Predictive models often rely on nonparametric data mining algorithms, e.g. classification trees, neural networks, and k-nearest-neighbors and nonparametric smoothing methods, e.g. moving average forecasters. The flexibility of such methods enables them to capture complex relationships in the data without making restricting statistical assumptions.

During the modelling phase data is searched for useful patterns. Within the machine learning process, dataset needs to pass through the modeling process to identify the patterns from the datasets. Predictive modeling is the general concept of building a model that is able to make predictions. Such a model includes a machine learning algorithm that learns certain properties from a training dataset to make the predictions. Predictive modeling can be divided further into two sub areas: regression and pattern classification. Regression models are based on the analysis of relationships between variables in order to make predictions about continuous variables, e.g. the prediction of the maximum number for the upcoming days in weather forecasting. Pattern classification assigns discrete class labels to particular observations as outcomes of a prediction. To go back to the above example: A pattern classification task in weather forecasting could be the prediction of a sunny, rainy, or snowy day[73].

Clustering is the division of data into groups of similar objects. Clustering, as one common data analysis technique, models the data by its clusters to get an intuition to find either hidden patterns or groups. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup are very similar and data points in different clusters are very different. From a machine learning perspective, clusters correspond to hidden patterns, the search for clusters is unsupervised learning and the resulting model represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval, marketing, medical diagnostics and many others. While representing the data by fewer clusters necessarily loses certain fine details, but it achieves simplification instead[74]. The goal of clustering is to investigate the structure of data by grouping the data points into distinct subgroups, e.g. clustering the data from factory runs in different groups based on day of week. Different clustering analysis methods are applicable with different software, such as R, SPSS, MATLAB, Python or even Excel, through number of toolboxes, packages and libraries.

Data clustering algorithms can be divided into two main types: hierarchical and partition. Hierarchical algorithms find successive clusters using previously established clusters, whereas partition algorithms determine all clusters at time. Hierarchical clustering builds a cluster hierarchy or, i.e. a tree of clusters. Every cluster node contains child clusters and sibling clusters. This approach allows exploring data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down). An agglomerative clustering starts with one point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively

73 R Raschka, S. (2014). Predictive modeling, Supervised machine learning, and pattern classification. *Machine Learning*.

74 Berkhin P. (2006) A Survey of Clustering Data Mining Techniques. In: Kogan J., Nicholas C., Teboulle M. (eds) Grouping Multidimensional Data. Springer, Berlin, Heidelberg.

splits the most appropriate cluster. The process continues until a stopping criterion (frequently, the requested number k of clusters) is achieved. Advantages of hierarchical clustering include:

- Embedded flexibility regarding the level of granularity.
- Ease of handling of any forms of similarity.
- Applicability to any attribute types.

In addition, the disadvantages of hierarchical clustering are:

- Vagueness of termination criteria.
- Most hierarchical algorithms do not revisit once constructed (intermediate) clusters with the purpose of their improvement.

The partitioning algorithms divide data into several subsets. Because checking all possible subset systems is computationally infeasible, certain greedy heuristics are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found.

More specifically, probabilistic models assume that the data comes from a mixture of several populations whose distributions and priors we want to find. One clear advantage of probabilistic methods is the interpretability of the constructed clusters. Having concise cluster representation also allows inexpensive computation of intra-clusters measures of fit that give rise to a global objective. Another approach starts with the definition of objective function depending on a partition. Depending on how representatives are constructed, iterative optimization partitioning algorithms are subdivided into k-medoids and k-means methods. k-means clustering and partitioning around medoids are well known techniques for performing non-hierarchical clustering.

The k-means algorithm assigns each point to the cluster whose center, centroid, is nearest. The center is the mean of all the points in the cluster that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The algorithm passes through the following steps:

I. Select a number of classes/groups to use and randomly initialize their respective center points. To figure out the number of classes to use, data is pre-examined quickly to identify any distinguishable groups. The center points are vectors of the same length as each data point vector.
II. Classify each data point by computing the distance between that point and each group center and assign the point to that group whose center is closest to it.
III. Based on these classified points, recompute the group center by taking the mean of all the vectors in the group.
IV. Repeat these steps for a set number of iterations or until the group centers don't change much between iterations. It is also possible to randomly initialize the group centers a few times, and then select the run that looks like it provided the best results.

k-means tries to cluster the dataset into k numbers of predefined distinct non-overlapping groups where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. k-means clustering minimizes within-cluster variances (squared Euclidean distances). Better Euclidean solutions can be found using k-medians and k-medoids.

k-means has the advantage that it is pretty fast and does a very good job when the clusters have a kind of spherical shapes, but has a couple of disadvantages. It does not learn the number of clusters from the data and requires pre-definition of possible clusters. k-means also starts with a random choice of cluster centers and therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency. Other cluster methods

are more consistent. Moreover, it suffers as the geometric shapes of clusters deviates from spherical shapes.

K-means clustering is known to be sensitive to the outliers although it is quite efficient in terms of the computational time. For this reason, k-medoids clustering are sometimes used, where representative objects called medoids are considered instead of centroids.

Since k-medoids clustering is based on the most centrally located object in a cluster, it is less sensitive to outliers and noises. Instead of using the mean point as the center of a cluster, k-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with a minimum sum of distances to other points[75]. Representation by k-medoids has two advantages. First, it presents no limitations on attributes types, second, the choice of medoids is dictated by the location of a predominant fraction of points inside a cluster and therefore, it is less sensitive to the presence of outliers. In k-means case a cluster is represented by its centroid, which is a mean (usually weighted average) of points within a cluster[76].

Two early versions of k-medoid methods are the algorithm PAM (Partitioning Around Medoids) and the algorithm CLARA (Clustering LARge Applications)[77]. PAM, known to be most powerful among algorithms for k-medoids clustering, is iterative optimization that combines relocation of points between perspective clusters with re-nominating the points as potential medoids. The guiding principle for the process is the effect on an objective function, which is obviously a costly strategy. CLARA uses several samples with thousands points each, which are each subjected to PAM. The whole dataset is assigned to the resulting medoids, the objective function is computed and the best system of medoids is retained. However, both methods have the drawback that they are working inefficiently for a large data set due to time complexities[78]. There have been some efforts in developing new algorithms for k-medoids clustering.

It can be proved that the k-means and k-medoids algorithms will always terminate, but it does not necessarily find the best set of clusters, corresponding to minimizing the value of the objective function. The initial selection of centroids can significantly affect the result. To overcome this, the algorithm can be run several times for a given value of k, each time with a different choice of the initial k centroids, the set of clusters with the smallest value of the objective function then being taken. The most obvious drawback of this method of clustering is that there is no principled way to know what the value of k ought to be[79].

One of the major drawbacks of k-means is its naive use of the mean value for the cluster center. One difference between k-means and Gaussian mixture models is that while k-means performs hard classification, Gaussian Mixture Model (GMM) performs soft classification, i.e. k-means tells us what data point belong to which cluster but will not provide us with the probabilities that a given data point belongs to each of the possible clusters. The other difference is that k-means does not account for variance. GMMs give more flexibility than k-means, by assuming the data points are Gaussian distributed; this is a less restrictive assumption than saying they are circular by around the mean. The clusters can take any ellipse shape, rather than being restricted to circles.

---

75 Jin, X.; Han, J. (2010), “K-Medoids Clustering”

76 Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert systems with applications*, *36*(2), 3336-3341.

77 Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

78 Han, J. (2001). Spatial clustering methods in data mining: A survey. *Geographic data mining and knowledge discovery*, 188-217.

79 Bramer M. (2016), “Principles of data mining”, Berlin: Springer; DOI: https://doi.org/10.1007/978-1-4471-7307-6

In two dimensions, variance determines the shape of the distribution. One way to think about the k -means model is that it places a circle, or a hyper-sphere in higher dimensions, at the center of each cluster, with a radius defined by the most distant point in the cluster. This method works properly when data is circular. When data have different shapes, GMM can help more where models can handle even oblong clusters. There are two parameters to describe the shape of the clusters: the mean and the standard deviation. To find the parameters of the Gaussian for each cluster, optimization algorithm called Expectation-Maximization (EM) is used. The Gaussians distributions should be fitted to the clusters. The procedure is described as follows:

I. Select the number of clusters and randomly initialize the Gaussian distribution parameters for each cluster. Taking a look at the data or heuristic methods can provide a good guess for the initial parameters. The Gaussians can initiate randomly but quickly being optimized.
II. Given the Gaussian distributions for each cluster, compute probability that each data point belongs to a particular cluster. The closer a point is to the Gaussian's center, the more likely it belongs to that cluster.
III. Based on the probabilities, EM computes new set of parameters for the Gaussian distributions to maximize the probabilities of data points within the clusters. One way is to compute the new parameters by using a weighted sum of the data point positions, where the weights are the probabilities of the data point belonging in that particular cluster.
IV. Steps 2 and 3 are repeated iteratively until convergence, where the distributions don't change much from iteration to iteration.

Due to the standard deviation parameter, k-means is actually a special case of GMM in which each cluster's covariance along all dimensions approaches. Since GMMs use probabilities, they can have multiple clusters for one data point. So if a data point is in the middle of two overlapping clusters, it is simply defined to belong with X-percent to class 1 and with Y-percent to class 2, i.e. GMMs support mixed membership[80].

Two pioneering methods are the Random Decision Forests system[81], and the Random Forests system[82]. Both use the approach of generating a large number of decision trees in a way that has a substantial random element, measuring their performance and then selecting the best trees for the ensemble. Ho argues that traditional trees often cannot be grown over a certain level of complexity without risking a loss of generalization caused by overfitting the training data. Ho proposes inducing multiple trees in randomly selected subsets of the feature space. He claims that the combined classification will improve, as the individual trees will generalize better on the classification for their subset of the feature space [83].

Ho's work introduced the idea of making a random selection of the attributes in order to use when each classifier is generated. Additionally Bierman introduced a technique known as bagging for generating multiple different but related training sets from a single set of training data, with the aim to reduce overfitting and improving classification accuracy[84]. Naturally, this is computationally expensive to do. Random Forests are powerful algorithm, which take care of missing values, outliers and other non-linearities in the data set. Random forest or random decision forest is a learning

80 https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 [Accessed at July 15, 2020].

81 Ho, Tin Kam. "Random decision forests." *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, 1995.

82 Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

83 Stahl, F., & Bramer, M. (2011, December). Random prism: An alternative to random forests. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 5-18). Springer, London.

84 Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

method for classification, regression and other tasks that can be operated by constructing different decision trees during training time. The outcome is either the class in classification field or mean of prediction in regression of the individual trees. It is simply a collection of classification trees, which was the reason to name the method as 'forest'. It uses bagging and feature randomness when building each individual tree to try a creation of an uncorrelated forest of trees whose prediction by group is more accurate than any individual tree decision. Random forests are an effective tool in prediction, and because of the Law of Large Numbers they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors.

A forest of trees is inexplicable as far as simple interpretations of its mechanism go. In some applications, e.g. in analysis of medical experiments, it is required to understand the interaction of variables. A start on this problem is made by using internal out-of-bag estimates, and then verify them by reruns over selected variables. The forests use the random selection of features at each node to determine how to split. An important question is how many features to select at each node. For guidance, internal estimates of the generalization error, classifier strength and dependence are computed. These are called out-of-bag estimates.

4.6.2. Online Learning

Most of the machine learning algorithms assume the availability of whole historical data before the start of learning from them. However, sometimes it is required to learn from data at the same time as gathering them. Time-critical and memory-critical applications require real time learning, and massive data clustering with limited memory respectively. For example, business transaction of a large retail company require massive data clustering with limited memory. The data stream clustering problem is defined as to maintain a consistently good clustering of sequence observed so far, using a small amount of memory and time[85]. There are several modes in which the data may be available; in batch mode, a finite dataset is available from the beginning. This is the most commonly analyzed setting. Streaming mode refers to a setting where there is a finite amount of data to be processed but it arrives one data point at a time, and the entire dataset cannot be stored in memory. The online mode departs from the previous two settings in that there is an infinite amount of data to be analyzed and it arrives one data point at a time. Since there is an infinite amount of data, it is impossible to store all the data in memory[86]. In the online setting, it is much less clear how to analyze an algorithm's performance.

The recent data is considered as all the data available in the memory from the data stream, the historical data is the data observed in the data stream so far, which include the recent data. Except for the portion of recent data, historical data are not available in memory. The unprocessed recent data is called newly arrived data. If the entire historical data were available in memory, Gaussian mixture model would have been effectively estimated using the EM algorithm. For a data without complete historical records, the EM or any of its known variations is not applicable. One can adapt probability density based clustering algorithms to solve data stream clustering problems much more efficiently than applying the EM on the entire historical data, by applying standard EM algorithm only on newly arrived data. The resulted incremental GMM estimation algorithm merges Gaussian components that are statistically equivalent. The sufficient statistics of mean and covariance for a multivariate normal distribution make it possible to perform the tests and merging without resort to historical data. The idea is to merge density components rather than data, a deterministic and randomized incremental clustering algorithm with a provably good performance. It works in phases, each consisting of a merging and an updating stage. In the merging stage, the algorithm reduces the number of clusters by combining selected pairs; in the updating stage, the algorithm accepts new updates and tries to maintain specific number of clusters without increasing the clusters boundaries or violating certain constraints. It is undoubtedly important that algorithm to analyze generating task data operate in the online learning setting. This setting is applicable for forecasting,

85 Guha, S., Mishra, N., Motwani, R., & o'Callaghan, L. (2000, November). Clustering data streams. In *Proceedings 41st Annual Symposium on Foundations of Computer Science* (pp. 359-366). IEEE.

86 King, A. (2012). Online k-means clustering of nonstationary data. *Prediction Project Report*, 1-9.

real-time decision making, and resource-constrained learning, e.g. in case of failure occurrence in machines, transferring robots or driving paths.

- X_T: Random data point (vector) in $\mathbb{R}^d$ which is observed at time T (superscript T is omitted when the value of T is not specified).
- d: Number of dimensions in data.
- $g^T: \mathbb{R}^T \to \mathbb{R}$, an estimator of probability density function, $p_0(x)$, based on data points observed from beginning time 1 to time T.
- $g^N(x)$: An estimator of $p_0(x)$ based on historical data $X_1, \dots, X_N$.
- $g^{N+M}(x)$: An estimator of $p_0(x)$ based on both historical data $X_1, \dots, X_N$ and newly arrived data $X_{N+1}, \dots, X_{N+M}$; M newly arrived data sample.

So, the data stream clustering problem will address: Obtain $g^{N+M}(x)$ from $g^N(x)$. Here we assume that the cluster of each data point can be uniquely determined by $g(x)$.

- Let $g(x)$ be $g^N(x)$.
- Let $q(x)$ be an estimator of $p_0(x)$, based on $X_{L+1}, \dots, X_{N+M}$ data.

Estimators $g(x)$ and $q(x)$ are very likely different but obviously related to each other because they share the data $X_{L+1}, \dots, X_N$.

- Let $t(x)$ be an estimator of $p_0(x)$ based on $X_1, X_2, \dots, X_L$ (the data points removed from consideration of $q(x)$).
- Let $a(x)$ an estimator of $p_0(x)$ based on $X_{N+1}, X_{N+2}, \dots, X_{N+M}$ (the newly arrived data points included in $q(x)$).

Under the assumption that the data points are independently and identically distributed (i.i.d.), a theorem is established that dictates the way $q(x)$ is updated from $g(x)$ without exact knowledge of the overlapped data $X_{L+1}, \dots, X_N$. When $L, M << N$, the theorem to be established will have direct impact on significantly improving the efficiency of the density estimation.

Estimator Updating Theorem: for the estimators $t(x)$, $g(x)$, $a(x)$ and $q(x)$ defined on data points in four time intervals within $[1, N + M]$,

$$q(x) = \frac{Ng(x) - Lt(x) + Ma(x)}{N - L + M}. \quad \text{Equation 2}$$

This theorem is inspired by a relation among the empirical cumulative distribution functions (c.d.f.s). The empirical c.d.f.s $T(x)$, $Q(x)$, $G(x)$, $A(x)$, corresponding to $t(x)$, $q(x)$, $g(x)$ *and* $a(x)$, respectively, are

$$Q(x) = \frac{|\{x_n \le x | n = L+1, L+2, \dots, N+M\}|}{N - L + M} \quad \text{Equation 3}$$

$$G(x) = \frac{|\{x_n \le x | n = 1, 2, \dots, N\}|}{N} \quad \text{Equation 4}$$

$$T(x) = \frac{|\{x_n \le x | n = 1, 2, \dots, L\}|}{L} \quad \text{Equation 5}$$

$$A(x) = \frac{|\{x_n \le x | n = N+1, N+2, \dots, N+M\}|}{M}. \quad \text{Equation 6}$$

From the four definitions above, we have evidently

$$LT(x) + (N - L + M)Q(x) = NG(x) + MA(x). \quad \text{Equation 7}$$

Thus, we obtain

$$Q(x) = \frac{NG(x) - LT(x) + MA(x)}{N - L + M}. \quad \text{Equation 8}$$

The algorithm for online data stream clustering by merging components in GMM, exploits the statistical structures of data using only newly arrived data, instead of keeping all historical data. The algorithm does show a tendency to produce more clusters than the standard EM algorithm. Sometimes two clusters really belong to one Gaussian component cannot be merged, because they do have different density individually.

Online Clustering with Experts[87]

Where the raw data is not yet labeled for any classification task, an unsupervised learning like clustering techniques can be used to summarize large quantities of data, but the outputs can be hard to evaluate. A domain expert may be useful in judging the quality of resulting clusters, but having a human in the loop is not always desirable. One way to analyze clustering algorithms is to formulate objective functions, and then to prove that the clustering algorithm either optimizes it or is an approximation algorithm. In case of success in defining reasonable objectives, approximation guarantees will fulfill the requirements, namely Online k-means Approximation.

The k-means objective is a simple, intuitive, and widely-cited clustering objective. For a finite set S of n points in $\mathbb{R}^d$ and a fixed positive integer k the k-means objective is to choose a set of k cluster centers C in $\mathbb{R}^d$ to minimize:

$$\Phi_X(C) = \sum_{x \in S} \min_{c \in C} \|x - c\|^2, \qquad \text{Equation 9}$$

which is referred to as the "k-means cost" of C on X. This objective formalizes an intuitive measure of goodness for a clustering of points in Euclidean space. Optimizing the k-means objective is known to be NP-hard, even for $k = 2$. Therefore, the goal is to design approximation algorithms[88].

However, few algorithms provably approximate it, even in the batch setting. One flexible framework is required in which the algorithm take a set of candidate clustering algorithms as experts and track the performance of the "best" or best sequence of experts. The algorithms compute an approximation to the current value of the k-means objective obtained by each expert instead of computing prediction errors and re-weighting the experts. As an analog, it is inspired by the evaluation framework to regret[89]. It proposed two possible methods for evaluating the performance of an online clustering scheme. The first is an approximation algorithm style bound: If at time t an algorithm sees a new data point x_t and then outputs a set of k clusters C_t, then an approximation algorithm would find some $\alpha \geq 1$ such that $Cost(C_t) \leq \alpha\, OPT$, where OPT is the cost for the best k clusters for $x_1, \ldots, x_t$. Another type of performance metric is regret, which is a common framework in online learning. Under a regret framework, an algorithm would announce a set of clusters C_t at time t, then recieve a new data point x_t and incur a loss for that point equal to the squared distance from x_t to its closest cluster in C_t.

The regret framework for the analysis of supervised online learning algorithms evaluates algorithms with respect to their additional prediction loss. With the goal of analyzing online clustering algorithms, one can bound the difference between the cumulative clustering loss since the first observation:

87 Choromanska, A., & Monteleoni, C. (2012, March). Online clustering with experts. In *Artificial Intelligence and Statistics* (pp. 227-235).

88 Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine learning*, *75*(2), 245-248.

89 Dasgupta, S. (2008), Course notes, CSE 291: Topics in unsupervised learning. Lecture 6: Clustering in an online/streaming setting. Section 6.2.3. In http://cseweb.ucsd.edu/~dasgupta/291-unsup/lec6.pdf.

$$L_T(alg) = \sum_{t \le T} \min_{c \in C_t} \|x_t - c\|^2 \qquad \text{Equation 10}$$

Where the algorithm outputs a clustering C_t before observing the current point x_t and the optimal k-means cost on the points seen so far. It provides clustering variants of predictors with expert advice and analyze them by first bounding this quantity in terms of regret with respect to the cumulative clustering loss of the best expert or best sequence of experts.

Online k-means Clustering of Nonstationary Data[90]

Considering a motivation example of a t-shirt retailer that receives online data about their sales. The retailer sells women, men, girls' and boys' t-shirts. Fashion across these groups is different; the colors, styles, and prices that are popular in women's t-shirts are quite different from those that are popular in boys' t-shirts. As fashion changes over time, so do the characteristics of each demographic. At every point in time, the retailer would like to be able to segment its market data into correctly identified clusters and find the appropriate current cluster average. This is a problem of online clustering of non-stationary data. More formally, consider dimensional data points, which arrive online and need to be clustered. While the cluster should remain intact, its center essentially shifts in d-space over time. The entire data set to be processed may not be available when one needs to begin learning; the data in the t-shirt example is not even finite.

The standard k-means cost objective is unsuitable in the case of non-stationary data, therefore an alternative cost objective should be used. The objective function for traditional k-means clustering for available historical data is determined by

$$cost(C_1, \dots, C_k, z_1, \dots, z_k) = \sum_k \sum_{i:x_i \in C_k} ||x_i - z_k||_2^2, \qquad \text{Equation 11}$$

$C_1, \dots, C_k$, are the k clusters and

$z_1, \dots, z_k$, are the k cluster centers

In the online setting, the cost at any point in time is calculated by

$$cost_t(C_1^t, \dots, C_k^t, z_1^t, \dots, z_k^t) = \sum_k \sum_{i:x_i \in C_k^t} ||x_i - z_k^t||_2^2 \qquad \text{Equation 12}$$

This objective implicitly assumes that the goal is to find the best online clustering for all of the points seen so far, and at time t, all points $x_1, \dots, x_t$ contribute equally. In the non-stationary setting old data points should not be worth as much as new data points; the data may have shifted significantly over time.

In order to consider this, the data from different times should be weighted differently.

$$cost_t(C_1^t, \dots, C_k^t, z_1^t, \dots, z_k^t) = \sum_k \sum_{i:x_i \in C_k^t} \delta^{t-i} \left|\left|x_i - z_k^t\right|\right|_2^2, \delta \in (0.1) \qquad \text{Equation 13}$$

The last objective function will result in k clusters, which minimize the weighted squared distance from each point to its cluster center, where the weight is exponentially decreasing in the age of the point.

90 King, A. (2012). Online k-means clustering of nonstationary data. *Prediction Project Report*, 1-9.

4.6.3. Statistical Model of Time Patterns

The statistical model of time pattern on the task generating flow can be built with mathematical statistics analysis and distribution fitting. The construction of statistical model of flow time pattern mainly includes three steps: data processing, distribution fitting and goodness-of-fit test[91].

Data Processing

Data processing should be done on the observed data from simulations. The following factors should be determined:

- The number of generated tasks in time unit, e.g. one hour, as a sample set X in a total of n days simulation, expressed as $x_1, \dots, x_n$.
- Maximum, minimum, and mean value as well as variance over each time unit segment over different days simulations.
- Frequency of generating task number per unit time.
- Histogram of measured frequencies.

Distribution Fitting

Distribution Fitting is the comparison probability distribution frequency histograms, Poisson distributions and normal distributions of generated tasks in unit time by using SPSS or MATLAB tools. The distribution Goodness-of-fit indicates whether the fitting distribution curve is reasonable. The method for the time pattern of flow, e.g. traffic flow, is usually a chi-square test, regression method or Kolmogorov-Smirnov test (K-S test). The K-S test is a nonparametric method to check the consistency between sample observation values and the given theoretical distribution and to determine if the sample observations from the given theoretical distribution based on the distribution analysis of the two differences. The basic idea is to

- Set $S_n(x)$ as a cumulative probability distribution function with a random sample observation value of n time observations, called empirical distribution function.
- $F_0(x)$ is a specific cumulative probability distribution function, called theoretical distribution function.

The sample empirical distribution function can be obtained directly from a sample and reflect the sample distribution situation directly, set to

$$S_n(x) = \begin{cases} 0 & \text{if } x < x_{(1)} \\ \frac{k}{n} & \text{if } x_{(k)} \le x \le x_{(k+1)} \\ 1 & \text{if } x \ge x_{(n)} \end{cases}. \qquad \textit{Equation 14}$$

Absolute difference will be

$$D_n = |S_n(x) - F_0(x)|. \qquad \textit{Equation 15}$$

For every x value, if $S_n(x)$ is very close to $F_0(x)$ and the difference is very small, we can conclude that the fitting degree between the empirical distribution function and specific distribution function is very high and there is reason to think that the sample data come from the overall with the theoretical distribution.

91 Liu, Z., Liu, J., Liu, Y., & Wang, K. (2013). Model of prediction and statistics on time pattern of traffic flow. In *ICTIS 2013: Improving Multimodal Transportation Systems-Information, Safety, and Integration* (pp. 2341-2350).

4.6.4. Tools and Platforms for Machine Learning in Industry

Intelligent techniques like machine learning methods for analyzing big data are promising tools for multiple applications in manufacturing like quality control[92], predictive maintenance[93] and resource management[94]. Applying such a technique requires processing of big data: data collection, data analysis, querying and storage. In the last years, multiple open-source tools for those processing steps were developed. We are referring to Sahal et al. for a comparison of those tools as decision support on which tool fits the best to the own use case[95]. The compared technologies comprise data collection, analysis, storing and querying. Especially the data analysis part of data processing can strongly benefit from machine learning techniques. For a comprehensive overview on challenges, advantages and applications on machine learning for manufacturing, we refer to[96,97] and the following tools:

TensorFlow

TensorFlow is a comprehensive open source platform for machine learning, which allows Data preprocessing, model training and evaluation as well as data visualization[98]. Due to the simple modelling tools and easy to perform tutorials, it is a suitable tool for beginners in AI technologies. TensorFlow can be used directly in browser via google colab as well as API for Phyton, Javascript and IOS devices. As highlight for manufacturing, there is the TensorFlow Extended platform to compile a pipeline of components for production related ML implementations.

Microsoft Azure

Microsoft Azure is a cloud platform, which offers single services for database solutions, development, networking and computation. Each service can be composed to an overall system for customer's individual solutions.[99] For predictive analysis and machine learning purposes there is the Microsoft Azure "Machine Learning Studio" tool to interactively design your machine learning environment with components for data importing and preprocessing, model training and evaluation and comparison, predictive experimenting and web publishing. Due to a reduced deployment time for predictive models, the "Machine Learning Studio" tool allows the application of machine learning to production and manufacturing.[100]

---

92 Escobar, Carlos A., and Ruben Morales-Menendez. "Machine learning techniques for quality control in high conformance manufacturing environment." *Advances in Mechanical Engineering* 10.2 (2018): 1687814018755519.

93 Zonta, Tiago, et al. "Predictive maintenance in the Industry 4.0: A systematic literature review." *Computers & Industrial Engineering* (2020): 106889.

94 Cadavid, Juan Pablo Usuga, et al. "Machine learning applied in production planning and control: a state-of-the-art in the era of industry 4.0." *Journal of Intelligent Manufacturing* (2020): 1-28.

95 Sahal, Radhya, John G. Breslin, and Muhammad Intizar Ali. "Big data and stream processing platforms for Industry 4.0 requirements mapping for a predictive maintenance use case." *Journal of manufacturing systems* 54 (2020): 138-151.

96 Wuest, Thorsten, et al. "Machine learning in manufacturing: advantages, challenges, and applications." *Production & Manufacturing Research* 4.1 (2016): 23-45.

97 Wang, Jinjiang, et al. "Deep learning for smart manufacturing: Methods and applications." Journal of Manufacturing Systems 48 (2018): 144-156.

98 https://www.tensorflow.org/. [Accessed at March 4, 2021]

99 https://www.computerweekly.com/de/definition/Microsoft-Azure. [Accessed at March 4, 2021]

100 Barga, R., Fontama, V., Tok, W. H., & Cabrera-Cordon, L. (2015). Predictive analytics with Microsoft Azure machine learning (pp. 21-43). Berkely, CA: Apress.

IBM SPSS

IMBSS SPSS is a commercial Statistic software set to perform an overall data analytic process including data preparation, analysis, report and implementation. Originally developed for researchers, the software is now one of the leading data analysis products for various companies, public authorities and pollsters. Different modules for e.g. model testing by bootstrapping, regressions (linear, normal, logistic regression), descriptive statistics, neural networks, decision trees and marketing analysis cover a wide range of applications including the manufacturing domain.

SAS analytics

SAS analytics[101] is also one of the leading products for advanced data analytics. The platform offers a comprehensive toolchain for model building, categorical, multivariate and cluster data analysis and graphical result representation. [102] With simulation and project scheduling capabilities SAS analytics supports the decision making and the evaluation of alternatives in manufacturing and production management.

H20

“Democratize AI for everyone” - With this slogan H2O.ai offers his leading open source AI platform for companies, data scientists & factories.[103] It is an interactive and easy to use tool for data importing, analyzing and graphical presentation. For manufacturing purposes, H2O.ai is suitable to predict and forecast supply chains, detect machine failures and faults, manage product in- and exports and perform optimization with a Security Cyberlake[104].

4.7. Machine learning applied to Industry 4.0

The major goal of the Industry 4.0 is to enable the regeneration of the industrial sector towards a more reliable, efficient, resilient and competitive in the global market. In order to do so, it is necessary to converge information and technology[59], thus creating a net system between the real and the physical world. Using such a net, advanced services can be provided from the exploration of the data, thus enabling systems to detect and identify anomalies, prevent downtimes, adjust schedules and better plan future action. In this sense, big data analytics and machine learning techniques are emerging as essential tools to extract mining from the data and also they can give some important insights about future events (e.g. predict a failure in the production line), support actions/decisions (e.g. adjust the production schedule) or even to give the most likely explanation of past events. On the other hand, it is not easy to extract value from the digital information: the amount of data is huge, possibly very heterogeneous (data from different sensors such as numeric, audio and video), in different formats (unformatted text, reports or other source), different processing time requirements (online or batch data processing) and thus different computation requirements. All this data needs to be integrated in a common pipeline, in order to analyze and optimize production processes. Big data challenges can be schematized into four major dependent and interconnected components: Descriptive Analytics; Predictive Analytics; Prescriptive Analytics; Automated Analytics[105]. In the next sections, these concepts are going to be explored more deeply.

---

101 https://www.sas.com/en_us/home.html. [Accessed at March 4, 2021].

102 https://reviews.financesonline.com/p/sas-advanced-analytics/. [Accessed at March 4, 2021]

103 Candel, A., Parmar, V., LeDell, E., & Arora, A. (2016). Deep learning with H2O. H2O. ai Inc.

104 https://www.h2o.ai/democratizing-ai/. [Accessed at March 4, 2021]

105 "Guest Editorial: Data Science Challenges in Industry 4.0," in *IEEE Transactions on Industrial Informatics*, vol. 16, no. 9, pp. 5924-5928, Sept. 2020, doi: 10.1109/TII.2020.2984061.

Prescriptive Analytics:

The prescriptive analytic algorithms aim is to suggest to the user data-driven decisions based on predictions of models. These decision makers track deviations on the data distribution over the time (i.e. a time instance, where the object probability distribution has changed significantly) and are able to learn or re-learn the parameters of the distribution through fitting algorithms such as EM algorithm. In[106], a weighted one-class support vector machine (SVM) can adapt its decision boundary on online data, in this way, the algorithm learns new patterns and forget old ones. Other approaches[107], use an ensemble of classifiers (naïve Bayes or Linear Four Rates method[108]) in balance and unbalance datasets or more recently deep recurrent neural nets[109]. Deep learning models are the state-of-art approach in many domains and applications, due to their innate capability to approximate high complex probability distribution functions. On the other hand, these models suffer from huge data demand, a potential limitation concerning the amount of data available and its variability. The need to retrain or adapt the deep neural net to novel learning tasks and perhaps more importantly the lack of model comprehension. To recommend an action, it is (almost) mandatory to explain why such an action is the optimal one and on what facts such an action is supported. Based on these facts, some authors choose algorithms that can incrementally learn and adapt their internal configuration through data driven mechanisms.

Predictive Analytics: Forecasting algorithms

Another class of algorithms try to predict events, e.g. predict the machine speed in order to adjust processes, thus optimize the production throughput, and at the same time minimize the energy consumption. This class of algorithms use past observations to forecast a fixed length sequence of future values, for e.g. in [110] the author tries to predict the internal speed of a metal can bodymaker machine. These algorithms are divided into traditional and advanced machine learning approaches. Some examples of traditional approaches are autoregressive integrated moving average[111], SVMs

106 Krawczyk, Bartosz & Wozniak, Michal. (2014). One-class classifiers with incremental learning and forgetting for data streams with concept drift. Soft Computing. 10.1007/s00500-014-1492-5.

107 Krawczyk, Bartosz. (2017). Active and Adaptive Ensemble Learning for Online Activity Recognition from Data Streams. Knowledge-Based Systems. 10.1016/j.knosys.2017.09.032.

108 C. Lin, D. Deng, C. Kuo and L. Chen, "Concept Drift Detection and Adaption in Big Imbalance Industrial IoT Data Using an Ensemble Learning Method of Offline Classifiers," in IEEE Access, vol. 7, pp. 56198-56207, 2019, doi: 10.1109/ACCESS.2019.2912631.

109 L. Cao, Z. Qian, H. Zareipour, Z. Huang and F. Zhang, "Fault Diagnosis of Wind Turbine Gearbox Based on Deep Bi-Directional Long Short-Term Memory Under Time-Varying Non-Stationary Operating Conditions," in IEEE Access, vol. 7, pp. 155219-155228, 2019, doi: 10.1109/ACCESS.2019.2947501.

110 A. Essien and C. Giannetti, "A Deep Learning Model for Smart Manufacturing Using Convolutional LSTM Neural Network Autoencoders," in IEEE Transactions on Industrial Informatics, vol. 16, no. 9, pp. 6069-6078, Sept. 2020, doi: 10.1109/TII.2020.2967556.

111 M. Alipour, B. Mohammadi-Ivatloo and K. Zare, "Stochastic Scheduling of Renewable and CHP-Based Microgrids," in IEEE Transactions on Industrial Informatics, vol. 11, no. 5, pp. 1049-1058, Oct. 2015, doi: 10.1109/TII.2015.2462296.

[112] and regression trees[113]. On the other hand, Recurrent Neural Network (RNN)[114], convolutional neural nets (CNNs)[115], deep belief nets[116], autoencoders [117] or even a mix of RNN and CNN layers are examples of advanced methods. Deep learning algorithms are considered state-of-art approaches, because they can explore very efficiently the temporal dependencies on the data, thought ingenious mechanisms such as gates in LSTMs or receptive fields in CNNs. Furthermore, these algorithms can handle data with a high dimensionality and generate a compact latent representation of the input data, using specialized encoding layers. On the other hand, traditional approaches do not handle efficiently data with a high dimensionality and in general dimensionality reduction techniques such as Principal Component Analysis (PCA) are applied. The PCA is considered an unsupervised dimensionality reduction technique, the goal is to select a projection onto some subspace where the data variability is not reduced significantly. Regarding this issue, Manifold Learning methods have been also used to compact the information of the original feature space into a lower dimensional space. One of such examples, is the so-called Self-Organizing Maps (SOM), an unsupervised learning technique based on a grid of neurons that learns a discrete representation of the input signal and keeps the topological properties of the original signal. SOM has been extensively used in feature reduction, novelty detection [118] and fault prognostic[119]. A linear discriminant analysis is a supervised technique that attempts to maximize the linear separation between data points belonging to different classes[120].

---

112 Chun-Hsin Wu, Jan-Ming Ho and D. T. Lee, "Travel-time prediction with support vector regression," in IEEE Transactions on Intelligent Transportation Systems, vol. 5, no. 4, pp. 276-281, Dec. 2004, doi: 10.1109/TITS.2004.837813.

113 Cinzia Giannetti, Rajesh S. Ransing, Risk based uncertainty quantification to improve robustness of manufacturing operations, Computers & Industrial Engineering, Volume 101, 2016, Pages 70-80, ISSN 0360-8352, https://doi.org/10.1016/j.cie.2016.08.002.

114 Zaytar, Mohamed Akram & El Amrani, Chaker. (2016). Sequence to Sequence Weather Forecasting with Long Short-Term Memory Recurrent Neural Networks. International Journal of Computer Applications. 143. 7-11. 10.5120/ijca2016910497.

115 X. Dong, L. Qian and L. Huang, "A CNN based bagging learning approach to short-term load forecasting in smart grid," *2017 IEEE SmartWorld*, San Francisco, CA, 2017, pp. 1-6.

116 J. Zhang, R. Xue, X. Gao, F. Chen, Y. Chen and J. Yan, "Medium and Long Term Electricity Demand Forecasting of Different Industries Based on Deep Belief Network," *2018 2nd IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Beijing, 2018, pp. 1-6.

117 C. Liu, L. Tang and J. Liu, "A Stacked Autoencoder With Sparse Bayesian Regression for End-Point Prediction Problems in Steelmaking Process," in *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 2, pp. 550-561, April 2020.

118 Saucedo-Dorantes, Juan & Delgado-Prieto, Miguel & Romero-Troncoso, Rene & Osornio-Rios, Roque. (2019). Multiple-fault detection and identification scheme based on hierarchical self-organizing maps applied to an electric machine. Applied Soft Computing. 81. 105497. 10.1016/j.asoc.2019.105497.

119 X. Jin, Z. Que, Y. Sun, Y. Guo and W. Qiao, "A Data-Driven Approach for Bearing Fault Prognostics," in IEEE Transactions on Industry Applications, vol. 55, no. 4, pp. 3394-3401, July-Aug. 2019.

120 Abdi, Hervé & Williams, Lynne. (2010). Principal Component Analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2. 433 - 459. 10.1002/wics.101.

Predictive Analytics: Maintenance Algorithms

Another class of algorithms try to monitor the machine state, i.e. working under nominal (healthy) or faulty conditions. In detecting abnormalities in the machinery, industrial processes and schedules can be adjusted in order to avoid or minimize disruptions in the production line. The main limitation of these approaches is related to the number of unlikely events such as fault diagnosis in the databases. In Industrial applications, such conditions are likely to happen over undesired operating conditions, therefore maintenance support of classical approaches are limited. Furthermore, many of these outlier conditions are not properly annotated. In this sense, the algorithms task is to detect patterns that differ significantly from those available in the training set (novelty detection) and if possible, to predict a known fault scenario based on the training data (fault detection). The class of algorithms or machine learning frameworks fitted to this scenario are known as open set recognition problems. In this setting, a fraction of known classes is present in the training set, but perhaps novel or unseen classes might be present in the testing data. One standard approach is to apply a multi-class SVM as in[121], where for each class there a one-class classifier associated. This approach tries to associate an observation to a single classifier, in the case of two or more classes are likely to explain the data, similarity analysis is performed in order to assign a final class. In case that the data does not fit properly any of the classes, it is considered a novelty. Other studies do follow a slightly different approach, the algorithm starts to measure the similarity of the input data with samples from the training set. In[122] a density analysis of the data space is expressed by a Cauchy function. In case of dissimilarity, the data is classified as coming from a novel unknown class. If there is a degree of similarity, a multi-class classification algorithm is performed, and an observation is associated to one of the known classes. Another set of approaches use the so-called Anomaly Score (AS), a quantitative that measures the degree of "outlier" of an observation. AS can be used in an industrial environment to monitor the nominal condition of a system and select the optimal policy on the run.

Automation analysis: Flexible manufacturing

Flexibility is one of the key features of the Industry 4.0, the goal is to boost performance, reduce product life-cost and save resources using optimal consumption policies and at the same time reply and satisfy customer demands. In order to do so, efficiency and flexibility will heavily depend on the dynamic cooperation of large-scale machines. To manage and coordinate a large-scale of autonomous machines on accomplishing a customization task, it is necessary to find a high-dimensional coordination decision-making policy[123]. A cluster of autonomous machines can coordinate their action if for example a large-scale machine coordination system is implemented and solved by a large-scale Markov Decision Process (MDP). Mostly of the policy learning algorithms require a large amount of trials and simulations in order to achieve high performances. The amount of experiments and perhaps their complexity might become problematic for a high dimensional family of coordination control policies. The most known approaches to solve large-scale learning and decision problems for flexible manufacturing are: multiagent coordination control (based on a decentralized MDPs, where each agent dynamically adapts to the new coordination

121 Hao, Pei-Yi & Chiang, Jung-Hsien & Hsiu-Lin, Yen. (2009). A new maximal-margin spherical-structured multi-class support vector machine. Appl. Intell.. 30. 98-111. 10.1007/s10489-007-0101-z.

122 Costa, Bruno & Angelov, Plamen & Guedes, Luiz Affonso. (2014). Fully unsupervised fault detection and identification based on recursive density estimation and self-evolving cloud-based classifier. Neurocomputing. 150. 10.1016/j.neucom.2014.05.086.

123 J. Wang, Y. Sun, W. Zhang, I. Thomas, S. Duan and Y. Shi, "Large-Scale Online Multitask Learning and Decision Making for Flexible Manufacturing," in IEEE Transactions on Industrial Informatics, vol. 12, no. 6, pp. 2139-2147, Dec. 2016.

environment) [124], multitask policy gradient learning[125] (the model is described by a set of states, actions and rewards, the optimal policy is attained by applying a gradient descendent method on a likelihood policy dependent function) and large-scale decision making [126] (focus on off-policy batch multi-task learning). One approach based on multitask policy gradient learning is policy reuse [127], in this case previous learned tasks are reused to bias the learning of new tasks with a certain associated probability.

Descriptive Analytics: Self-Explainable algorithms

Currently, deep learning algorithms are often seen as “black-box” models, i.e. a system which is simply viewed in terms of its input and output computations. The computations on the nodes happen and are driven by not completely known reasons. Furthermore, these nets can be easily tricked through adversarial attacks, i.e. input data that is intentionally designed to break or disrupt the learning process of a net, or even in the worst-case scenario to completely neutralize the net and therefore its designed functionality. These limitations open space and opportunities to improve not only the AI safety but also AI transparency in Industry 4.0. Since the FoF are going to be built on AI agents spread around the shop floor. These AI agents are going to be responsible to manage automatic industrial processes (sensors, actuators, etc.) or even perhaps safety mechanism. Therefore, it is mandatory to protect these systems from external entities, which may have malicious intentions, and make them more robust against adversarial attacks, by using redundant or self-explainable mechanisms. If the net can provide an explanation of its decision in a way that resembles a rational followed by a human been, its decisions are more likely to be accepted and understandable by a human operator. On the other hand, unusual system behaviors can be more easily track by a human operator, simply by inspecting the consistency of the model’s explanation over the time. The challenges of the explainable artificial intelligence (XAI) algorithms include defining model explainability, formulating explainable tasks, and finally designing measurements that allow the evaluation of the model[128]. Attention mechanism is a self-explainable algorithm that tries to mimic the cortex of a human brain[129]. When our brain tries to describe an image, it does not look to the entire scene equally, but instead it focuses on specific shapes, colors, regions, object distribution, etc. These Regions of Interest are only used during the classification process and they are highlighted in the original image, in doing so the human users can understand more clearly the rational followed by XAI algorithms. Currently, there are also some case studies of XAI algorithms and their application to Industrial scenarios, for example in search and recommendation and ranking

124 A. Bratukhin and T. Sauter, "Functional Analysis of Manufacturing Execution System Distribution," in IEEE Transactions on Industrial Informatics, vol. 7, no. 4, pp. 740-749, Nov. 2011,

125 Wilson, Aaron & Fern, Alan & Ray, Soumya & Tadepalli, Prasad. (2007). Multi-task reinforcement learning: a hierarchical Bayesian approach, 1015-1022. 10.1145/1273496.1273624.

126 Li, Hui & Liao, Xuejun & Carin, Lawrence. (2009). Multi-task Reinforcement Learning in Partially Observable Stochastic Environments. Journal of Machine Learning Research. 10. 1131-1186. 10.1145/1577069.1577109.

127 Fernández, Fernando & Veloso, Manuela. (2012). Learning domain structure through probabilistic policy reuse in reinforcement learning. Progress in Artificial Intelligence. 2. 10.1007/s13748-012-0026-6.

128 Gade, Krishna & Geyik, Sahin & Kenthapadi, Krishnaram & Mithal, Varun & Taly, Ankur. (2019). Explainable AI in Industry. 3203-3204. 10.1145/3292500.3332281.

129 M. Carletti, C. Masiero, A. Beghi and G. A. Susto, "Explainable Machine Learning in Industry 4.0: Evaluating Feature Importance in Anomaly Detection to Enable Root Cause Analysis," *2019 IEEE International Conferen*

algorithms[130]. Understanding the decisions made by an AI agent[131] or understanding why the content of an email or web-page is fraudulent [132] or even understanding the sales predictions in terms of customer up-sell/churn [133] are some examples of resilient systems based on XAI.

4.8. Conclusions and Limitations

From the use cases, several sources of data are going to populate the lake. The amount and diversity of data, that is going to populate the lake is linked to their corresponding use case. For example, in textile industries, it is important to monitor the environmental humidity and temperature in order to produce high quality rubber. In cheese industry, the precise measurement of raw materials and products is crucial for a quality control. Regardless of its source, the data is going to be stored in the proposed data lake architecture. The proposed design will simplify the access to the data and the detection of patterns. It will allow deriving insights about existing and future operations in close-to-real-time. The lake must be dynamic in order to accommodate the high and mutable data. From the lake itself, a set of disruptive intelligent services based on ML approaches will foster a sustainable, secure, distributed and efficient set of data knowledge services, including: services for aggregation of information from the real-time monitoring; asset tracking, source code analytics, analytic services for the efficiency and efficacy of data usage, based on hybrid methodologies; intelligent decision support, able to transform the available data into knowledge that support decision making; cyber-security services from prevention, protection, detection and response to incidents on IT (critical) infrastructures. From the interaction itself between humans and machines, unexplored data sources like mouse and keyboard interaction events; speech and image data extracted from microphones and cameras are going to populate the lake. From such a data, health services are going to be provided, for example to measure stress, fatigue levels on the collaborators. This way, we'll bring a holistic view of FoF assets contextual information, paving the way for disruptive intelligent services.

Regarding the data lake and its efficiency, since the data lake is a repository to storage data in its natural format, it requires skills and a large effort to use it, especially for data scientists since machine learning algorithms usually have hard constraints about the data format and structure. Furthermore, explore the data might be challenging, since it might not be easy to understand correlations if the data is not on the same format, the same rational for queering the data or perform some advance analytics. However, in our opinion, the biggest risk of data lakes is related to security and access control. Since data can be placed into a lake without any oversight, as some of the data may have privacy and regulatory need. On the other hand, Data Lake is easy to storage data, with less up-front time investment, because the data is stored in its original native format with no structure required initially.

Smart factories need accurate forecasting techniques in order to make management decisions. Currently there is a big amount of data available, either from real factories or for by simulation runs. This information can be turned into actionable insights by hiring the data analysis and building a predictive model. The result insight will help to predict the demands and foresee the alterations in the setup factory to reduce risks and to improve operations. The data from system past behavior is used to build a

---

130 Yongfeng Zhang; Xu Chen, "Explainable Recommendation: A Survey and New Perspectives," in *Explainable Recommendation: A Survey and New Perspectives,* now, 2020.

131 Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, et al.. Interpretable Credit Application Predictions With Counterfactual Explanations. *NIPS 2018 - Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy*, Dec 2018, Montreal, Canada.

132 Z. Liu and A. Lu, "Explainable Visualization for Interactive Exploration of CNN on Wikipedia Vandal Detection," 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 2019, pp. 2354-2363.

133 K. Lin and J. J. P. Tsai, "A Deep Learning-Based Customer Forecasting Tool," 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), Taipei, 2016, pp. 198-205.

representative model, where this model captures the more important features. Hence it can be used on current data of system behavior to let us predict what will happen next or to find the actions to optimize the outcomes.

Towards understanding of the factory behavior, the studies in terms of recognizing patterns in how it either behaved in past or behaves in real time becomes essential. Pattern recognition is applicable by using data analysis, data mining, machine learning, and deep learning techniques to create quantitative predictive model. Predictive modeling is the concept of building a model that is able to make predictions. Such a model includes a machine learning algorithm that learns certain properties from a training dataset to make the predictions. The core of predictive analytics relies on first to find and exploit two types of variables in system behavior, namely explanatory and predicted variables, and second to capture the relationships between them. Clustering as one common data analysis technique, models data and finds hidden patterns or groups, to get an intuition about the data. It identifies clusters such that data points in the same subgroups are very similar and data points in different clusters are very different. From machine learning perspective, the search for clusters is unsupervised learning. The goal of clustering is to investigate the structure of data by grouping the data points into distinct subgroups, e.g. clustering the factory runs data into different groups based on workload distributed over different days of week. Subsequently, the predictive analysis will formalize the relation between days and workload and the result model can be employed to achieve the prediction goal.

While various methods have been proposed for the predictive analysis, there is a big challenge to study on which methods are more suitable for a given data. In addition, the sensitivity of the methods with regard to their parameter configuration should be evaluated and optimized in order to fulfill the requirements of pattern recognition and predicting the future behavior of the factory.

The enhancement of predictive analytics projects the system behavior model and predict the future events in an online manner. Most of the machine learning algorithms assume whole data are available before start learning; however, to make real time decisions, it is also required to learn from data at the same time as gathering them. The online learning will help to maintain a consistently good clustering of observed data so far, and simultaneously build and update representative model of system behavior using a small amount of memory and time.

5. H/M collaboration and optimization

Cobots, or collaborative robots, are robots intended for direct human robot interaction within a shared space, or where humans and robots are in close proximity. Cobot applications contrast with traditional industrial robot applications in which robots are isolated from human contact. The aim of the H/M collaboration is to make the cobots more adaptive to the workers. The cobot capabilities shall be more easily extended to new situations, even by users without programming or robotics background. The common ground between all developments is the learning by demonstration methods. Learning from Demonstration or Learning by Demonstration (LbD) is a paradigm for enabling robots to autonomously perform new tasks. Rather than requiring roboticists to manually program a desired task, work in Learning by Demonstration takes the view that an a cobot behavior can be generated from observations of a human's own performance.

5.1. Learning by demonstration

The main principle of cobot LbD is that users can teach cobots new tasks without programming. Consider, for example, an industrial cobot in a shop floor that the worker wishes to perform a pick-and-place operation with objects with different sizes. The task itself may involve different steps, such as detecting the different object shapes, adjusting the configuration to different form factors, and finally, applying the right pressure to handle the object. Furthermore, each time this task is performed, the robot will need to track the location of the objects over time. In a traditional programming scenario, a cobot developer would have to think in advance and code a robot controller that can respond to any situation the robot may face. This process may involve breaking down the task into multiple different steps, and thoroughly testing each step. If errors or new

circumstances arise after the robot is deployed, the entire costly process may need to be repeated. In contrast, LbD allows the user to perform a new action simply by showing it how to perform the task, no programming is required. Then, when failures occur, the user needs only to provide more demonstrations, rather than calling for development support. LbD hence embed robots with the ability to learn a task by generalizing from observing several demonstrations.

5.2. Definition

Robot Learning by Demonstration started in the 1980s[134]. Then, and now, robots had to be programmed for every task they performed. Learning by Demonstration (LbD)[135] seeks to minimize or eliminate, this difficult step by letting users train their cobot to fit their needs. The expectation is that the methods of LbD, being user-friendly, will allow cobots to be utilized in day-to-day interactions with non-specialist humans. Research on LbD has grown in importance, since that period and several surveys have been published. Most of the work on LbD follows a more machine learning approach.[136] Surveys of works in this area include Billard et al 2013 work[137]. At the core, however, LbD is inspired by the way humans learn from being guided by experts, from infancy through adulthood. Nehaniv & Dautenhahn[138] phrased the problems faced by Learning by Demonstration in a set of basic questions:

- What to imitate?
- How to imitate?
- Who to imitate?
- When to imitate?

What to imitate, relates to the problem of determining which aspects of the demonstration should be imitated. For a given task, certain properties of the environment may be irrelevant and safely ignored. Key to determining what is and is not important is understanding the metric by which the robot's behavior is being evaluated. Teaching what is and is not important can be done in multiple ways. The simplest approach is to take a statistical perspective and consider as relevant the parts of the data which are consistent across all demonstrations. If the dimension of the data is too high, such an approach may require too many demonstrations to gather enough statistics. An alternative is then to have the teacher help, the cobot determine what is relevant by pointing out parts of the task that are most important[139]. The issue of what to imitate take its root in developmental psychology. A fundamental step in child development occurs when children acquire the ability to

134 Hirzinger G, Heindl J. 1983. Sensor programming, a new way for teaching a robot paths and forces torques simultaneously. In Intelligent Robots: Conference on Robot Vision and Sensory Controls,Cambridge, Massachusetts/USA.

135 Billard, A. (2002). Imitation. Handbook of Brain Theory and Neural Networks: MIT Press.

136 Schaal, S., Ijspeert, A. and Billard, A. (2003). Computational approaches to motor learning by imitation, Philosophical Transactions: Biological Sciences (The Royal Society).

137 Billard, A, Calinon, S, and Dillmann, R. (2013). Learning from Human Demonstration. Handbook of Robotics: MIT Press.

138 Nehaniv, C.L. (2007), Nine Billion Correspondence Problems. In C. L. Nehaniv & K. Dautenhahn (Eds.), Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions, Cambridge University Press, 2007.

139 Lockerd, A. and Breazeal, C. (2004), Tutelage and Socially Guided Robot Learning, IEEE Int. Conf. on Robotics and Intelligent Systems, IROS.

perform discriminative imitation[140], that is, they move from imitating everything to imitating only the goal of the actions.

How to imitate consists in determining how the cobot will perform the learned behaviors to maximize the metric found when solving what to imitate problem. Often, a robot cannot act the same way as a human does, due the physical differences. This issue is related to that of the '''Correspondence Problem'''[141]. Robots and humans, while sharing the same space and interacting with the same object perceive and interact with the world in different ways. To evaluate the similarity between the human and robot behaviors, we must first deal with the fact that the human and the robot may occupy different state spaces. We identify two different ways in which states of demonstrator and imitator can be said to correspond:

- Perceptual equivalence: due to differences between human and robot sensory capabilities, the same scene may appear vastly different to each. For instance, while a human may identify humans and gestures visually, a robot may use depth measurements to observe the same scene. As the same data may therefore not be available to both humans and robots, successfully teaching a robot may require a good understanding of the robot's sensors and their limitations.
- Physical equivalence: due to differences between human and robot representation, humans and robots may perform different actions to accomplish the same physical effect. For instance, even when performing the same task, humans and robots may interact with the environment in different ways. Solving this discrepancy in motor capabilities is akin to solving the how to imitate problem and is the focus of the Learning by Demonstration. For example, a robot may compute a path (in Cartesian space) for its end-effector that is close to the path followed by the human, while relying on inverse kinematics to find the appropriate motion to compensate the fact that the two bodies are different.

Taken together, these two equivalences deal with discrepancies in how robots and humans are embodied. The perceptual equivalence deals with the way the agents perceive the world and makes sure that the information necessary to perform the task is available to both. Physical equivalence deals with the way agents affect and interact with the world and makes sure that the task is performable by both.

5.3. Approaches

The interface used to provide demonstrations plays a role in the way the information is captured and transmitted. We distinguish three major trends:

- **Directly recording human motions**. When interested only in the kinematics of the motions, one may use any of various existing motion tracking systems, based on vision, exoskeleton or other wearable motion sensors. This method is based on the visual sensors of the robot, skeleton tracking systems or external motion sensors the demonstrator must wear. The robot will track the human movements with its sensors, registering all the human body or only the skeleton. Another form of teaching can be performed with body's joints, given by motion tracking sensors worn by the demonstrator. This type of recording requires a robust tracking system, but it has the advantage to be external, thus giving a precise measurement of the movement. The movement first must be extracted from the environment (for the vision and tracking) and then to be adapted to the robot.

This approach is suitable for application with high degree of freedom robots or non-anthropomorphic robots, where kinesthetic teaching is difficult. However, the machine learning problem is complicated by the need to map movements from the human's actions to those executables by the

140 Gergely, G, Bekkering, H. and Király, I. (2002), Rational imitation in preverbal infants, Nature, 415(6873) p. 755-755.

141 Nehaniv, C.L. (2007), Nine Billion Correspondence Problems. In C. L. Nehaniv & K. Dautenhahn (Eds.), Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions, Cambridge University Press, 2007.

robot. Occlusions during recordings, rapid movement, and sensor noise in the observations makes this approach not applied in the industry although it has been reported in furniture assemblies[142].

- **Kinesthetic teaching**, where the robot is physically guided through the task by the humans. With this approach, no explicit physical correspondence is needed, as the user demonstrates the skill with the cobot's structure. It also provides a natural teaching interface to correct a skill reproduced by the cobot. This method has no correspondence problem, there is no need to adapt the recorded solution to the robot. One drawback of kinesthetic teaching is that the human must often use more of their own degrees of freedom to move the robot than the number of degrees of freedom they are trying to control. A typically task that would require synchronization between multiple limbs are difficult to teach kinesthetically. The human can have some difficulties to execute the movement. If a task requires the use of several parts of the robot, the human may experience some troubles to move them simultaneously, especially to reach the desired orientation. To achieve a good precision, the demonstrator will most of the time use more degrees of freedom than the robot for the motion. As a result, complex tasks cannot be done with this method, except if they are sequenced into smaller simplest ones and then combined.

The most common approach for introducing demonstrations in manufacturing applications is through kinesthetic teaching. On lightweight industrial robots this approach has been applied extensively to manipulators[143] due its intuitive approach and minimal training requirements. It relies completely on the robot itself as it does not need additional sensors, interfaces or inputs. Finally, because it is using the robot integrated sensors eliminates the correspondence problem[144]. On the other hand, the kinesthetic teaching depends on the dexterity of the human user as it often requires post-processing smoothing techniques.

- **Teleoperation scenarios**, where a human operator is limited to using the cobot's own sensors and effectors to perform the task. Going further than kinesthetics teaching, which limits the user to the robot's own body, teleoperation seeks to also limit the user's perception to those of the robot. This method is based on teleoperation systems, such as haptic devices or others remote control artifacts, allowing to remotely control the robot. Some external devices can be used to collect demonstration data, for instance joystick with feedback from the user can be used to record effort information.

Teleoperation is advantageous in that it not only entirely solves the correspondence problem, but also allows for the training of robots from a distance. As the teacher no longer needs to be near the robot, it is well suited for teaching navigation and locomotion patterns. The disadvantages of this method are the understanding of the remote-control solution and its configuration to map the kinematic model of the robot. This is the correspondence problem of this method. In industrial environments, it requires additional effort to develop an input interface and a longer user training process bur it can be applied to complex systems such as robotic hands[145] and underwater robots[146].

142 Hayes B, Scassellati B. 2014. Discovering task constraints through observation and active learning. In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE.

143 Pervez A, Lee D. 2018. Learning task-parameterized dynamic movement primitives using mixture of GMMs. Intelligent Service Robotics 11:61-78.

144 Argall BD, Chernova S, Veloso M, Browning B. 2009. A survey of robot learning from demonstration. Robotics and Autonomous Systems 57:469-483.

145 Aleotti J, Caselli S. 2011. Part-based robot grasp planning from human demonstration. In 2011 IEEE International Conference on Robotics and Automation. IEEE.

146 Havoutis I, Calinon S. 2018. Learning from demonstration for semi-autonomous teleoperation. Autonomous Robots 43:1-14.

5.4. Current methods

Current approaches to encoding skills through LbD can be divided between two trends: a low-level representation of the skill, taking the form of a non-linear mapping between sensory and motor information, and, a high-level representation of the skill that splits the skill in a sequence of action units.

5.4.1. Low level learning of individual motions

Individual motions could be taught separately instead of all at once. The human teacher would then provide one or more examples of each sub-motion apart from the others. If learning proceeds from the observation of a single instance of the motion, one calls this one-shot learning[147]. Examples can be found in the literature[148] for learning locomotion patterns. Different from simple record and play, here the controller is provided with prior knowledge in the form of primitive motion patterns and learns parameters for these patterns from the demonstration. Multi-shot learning can be performed in batch after recording several demonstrations, or incrementally as new demonstrations are performed. Learning generally performs inference from statistical analysis of the data from the demonstrations. Popular methods include Gaussian Process, GMMs and SVMs[149].

5.4.2. Teaching Force-Control Tasks

While most LbD work has focused on learning kinematic motions of end-effectors or other joints, one work has investigated extracting force-based signals from human demonstration[150]. Transmitting information about force is difficult for humans and for robots alike since force can be sensed only when performing the task ourselves. This line of work is fostered by advances in the design of haptic devices and tactile sensing, and on the development of torque and variable impedance actuated systems to teach force-control tasks through human demonstration.

5.4.3. Learning high-level action composition

Learning complex tasks, composed of a set of individual motions, is the goal of LbD. A common approach is to first learn models of all of the individual motions, using demonstrations of each of these actions individually[151], and then learn the right combination in a second stage either by

147 Wu, Y and Demiris, Y (2010), Towards One Shot Learning by imitation for humanoid robots, IEEE-RAS Int. Conf. on Robotics and Automation (ICRA).

148 Nakanishi, J.; Morimoto, J.; Endo, G.; Cheng, G.; Schaal, S.; Kawato, M. (2004). Learning from demonstration and adaptation of biped locomotion, Robotics and Autonomous Systems, 47, 2-3, pp.79-91.

149 Abbeel, P. & Ng, A. (2004), Apprenticeship Learning via Inverse Reinforcement Learning, International Conference on Machine Learning, ICML04.

150 Rozo, L, Jimenez, P. and Torras (2011), C. "Robot Learning from Demonstration of Force-based Tasks with Multiple Solution Trajectories," in 15th International Conference on Advanced Robotics, ICAR'11.

151 Daniel, C., Neumann, G. and Peters, J., Learning concurrent motor skills in versatile solution spaces, In proceedings of the IEEE International Conference on Robotics and Intelligent Systems (IROS'2012), p. 3591 – 3597.

observing a human performing the whole task[152] or through reinforcement learning[153]. An alternative is to watch the human perform the complete task and to automatically split the task to extract the primitive actions. The main advantage is that both the primitive actions and the way they should be combined are learned in one pass. One issue that arises is that the number of primitive tasks is often unknown, and there could be multiple possible segmentations, which must be considered[154].

5.5. Motion planning approaches

After recording the taught movement, it is necessary to obtain the mathematical representation of the movement. The resulting model must be an adaptation of the trajectory, which is robust against perturbations. There are to main techniques to achieve this representation: the Gaussian model and the Dynamic Motion Primitives.

5.5.1. Gaussian Mixture Model

This method extracts a set of primitive behaviors or actions from the given task and classifies them. Then it learns how to reproduce the movement from the individual behaviors and how to generalize it to new situations. This approach can be combined with a trajectory encoding which builds models operating in continuous spaces. For instance, it encodes the angular position of the robot's joints or the Cartesian position, speed or torque of an end-effector by mapping the sensory inputs to motor outputs and velocities[155].

5.5.2. Dynamic Movement Primitives

Dynamic movement primitives (DMPs) are a method of trajectory control / planning from Stefan Schaal's lab. They were presented back in 2002 and then updated in 2013[156]. This work was motivated by the desire to find a way to represent complex motor actions that can be flexibly adjusted without manual parameter tuning or having to worry about instability.

This control policy leads to DMPs which are autonomous nonlinear differential equations involving the positions, velocities and accelerations of the joints. This dynamical system encodes a trajectory from its initial state to its final state. This method does not require time-indexing and is robust against perturbations, thanks to the characteristics of the differential equations[157]. Finally, forcing terms are added to this model, allowing the learning of complex movements. DMPs are also simple

---

152 Skoglund, A., Iliev, B., Kadmiry, B. and Palm, R. Programming by Demonstration of Pick-and-Place Tasks for Industrial Manipulators using Task Primitives. International Symposium on Computational Intelligence in Robotics and Automation, 2007. CIRA 2007.

153 Mülling, K., Kober, J., Krömer, O., Peters, J. (2013). Learning to Select and Generalize Striking Movements in Robot Table Tennis, International Journal of Robotics Re- search, 32(3), pp. 280–298.

154 Grollman, D and Jenkins, O.C (2010), Incremental learning of subtasks from unsegmented demonstration, In International Conference on Intelligent Robots and Systems, Taipei, Taiwan, October 2010.

155 Ijspeert A., Nakanishi J., Hoffmann H., Pastor P., Schaal S. Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. Neural Computation, 25(2):328–373, February 2013.

156 Ijspeert A., Nakanishi J., Schaal S. Movement imitation with nonlinear dynamical systems in humanoid robots. In IEEE International Conference on Robotics and Automation (ICRA2002, pages 1398–1403, 2002.

157 Ijspeert A., Nakanishi J., Hoffmann H., Pastor P., Schaal S. Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. Neural Computation, 25(2):328–373, February 2013.

to learn by a robot because the weights of the forcing terms are learned separately and independently of each other to reduce the state space. Therefore, learning DMPs can be done quickly and efficiently even if a movement involves multiple degrees of freedom. The idea behind the DMPs is to use simple formulations of equations to encode the basic behavioral patterns and then to use statistical learning to adjust the obtained system to the task. This learning is obtained with the forcing terms. Due to the mathematical formulation, the DMPs have several properties including stability, invariance and robustness. There are several methods for both statistical and dynamical encodings of a skill, but not many studies have been conducted to comparing the DMP mechanisms. The only comprehensive comparison is presented by Silvain[158], where several methods are compared.

5.6. Commercial solutions

For industrial ready-to-market products, the most widespread approach is the kinesthetic teaching. This approach has been included in some commercialized product as shown in Figure 17 from several manufacturers: (a) LBR iiwa arm from Kuka, (b) Yumi bi-manual manipulator from ABB, (c) Sawyer cobot from Rethink Robotics, and (d) SYB from Isybot.

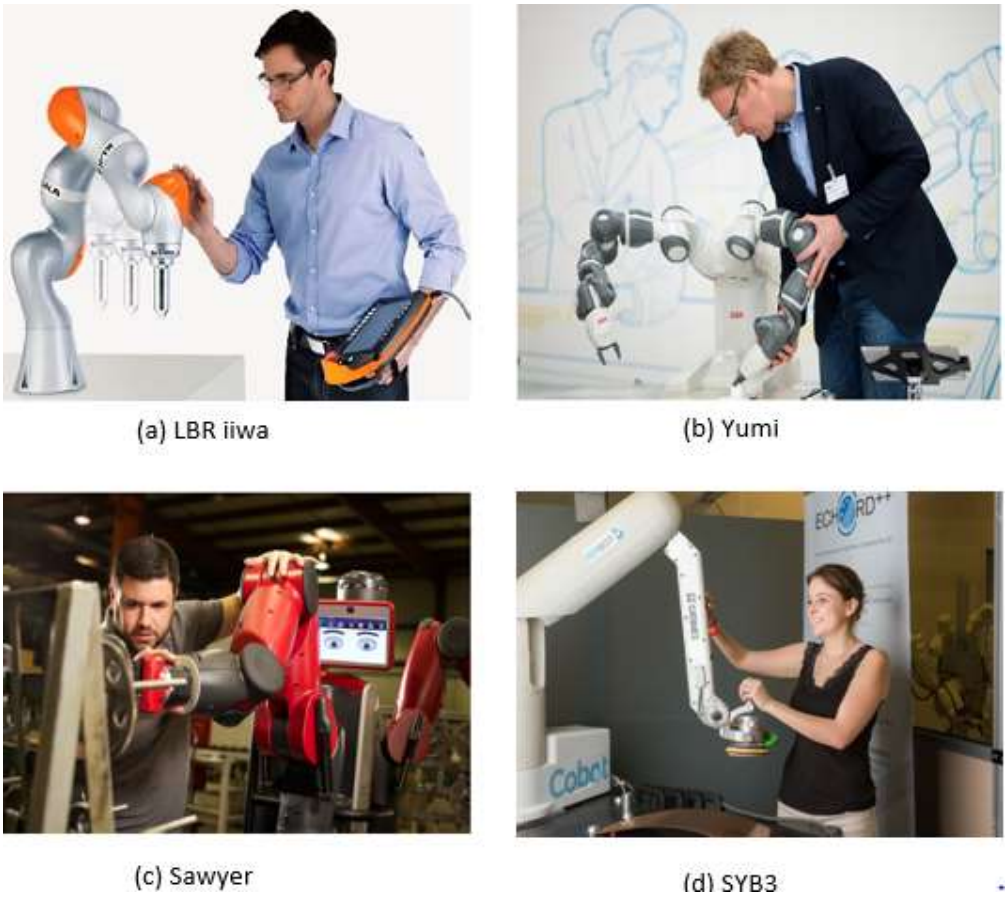

Figure 17 – Kinesthetic teaching with a commercialized robot, (Sanchez Restrepo, Susana. (2018). Intuitive, iterative and assisted virtual guides programming for human-robot co-manipulation. 10.13140/RG.2.2.11845.55520.).

158 Sylvain Calinon, Florent D'halluin, Eric L Sauser, Darwin G Caldwell, and Aude G Billard. Learning and reproduction of gestures by imitation. Robotics & Automation Magazine, IEEE, 17(2):44–54, 2010.

Teleoperation teaching has been used in industrial contexts. These devices are an evolution of the widespread teach-pendant devices found in most of the industrial robots with more advanced sensors such as haptic, force, magnetic, and inertia. Haption Designs, manufacturers and sells solution based on force-feedback. The Virtuoso 6D haptic device from Haption provides a solution for teleoperated robotic and industrial applications.

5.7. Limitations of current approaches and solutions

The DMP method does not consider the correlation changes between the movement variables and the variations observed among multiple demonstrations. Research in LbD is progressing rapidly, pushing back limits and posing new questions all the time. However, there are long-standing limitations and open questions. Generally, work in LbD assumes a fixed, given form for the robot's control policy, and learns appropriate parameters. To date, there are several different forms of policies in common usage, and there is no dominant technique. Furthermore, it is possible that a system could be provided with multiple possible representations of controllers and select which is most appropriate. The combination of reinforcement learning and imitation learning has been shown effective in addressing the acquisition of skills that require fine tuning of the robot's dynamics[159]. Likewise, more interactive learning techniques have proven successful in allowing for collaborative improvement of the learnt policy by switching between human-guided and robot-initiated learning. But there do not yet exist protocols to determine when it is best to switch between the various learning modes available. The answer may in fact be task dependent. In work to date, teaching is usually done by a single teacher, or teachers with an explicit concept of the task to teach. More work needs to be done to address issues related to conflicting demonstrations across teachers with different styles. Experiments in LbD have mostly focused on a single task (or set of closely related tasks) and each experiment starts from scratch. No previous tasks are considered. As learning of complex tasks progresses, means to store and reuse prior knowledge at a large scale will have to be discovered. Learning stages, similar perhaps to those found in child development, may be required. There will need to be a formalism to allow the robot to select information, to reduce redundant information to select features, and to store efficiently new data. More recent investigations on LbD point out as main limitation the generalization problem[160]. Generalization is the ability to respond to unseen circumstances which allows dealing with the real-world variability[161]. The ability to generalize differentiates systems that learn from systems that simply mimic a task. The generalization is at the core of the machine learning but some of the assumptions do not hold in robotic applications. The supervised algorithms assume that training and testing data is independent and identically distributed. However, in co-bot demonstrators some of the parts of the problem space are not covered[162]. DMPs formulate a nonlinear differential equation and produce the observed movement from a demonstration that generalizes appropriately changes in the initial and destination goals of the trajectory. However, it is hard to generate new behaviors from using DMPs. New investigations introduce task-oriented regression algorithms with a cost function that

159 Guenter, F., Hersch, M., Calinon, S. and Billard, A. (2007) Reinforcement Learning for Imitating Constrained Reaching Movements. RSJ Advanced Robotics, Vol. 21, No. 13, pp. 1521-1544.

160 Ravichandar, H., Polydoros, A., Chernova, S. and Billard, A. (2020) Recent Advances in Robot Learning from Demonstration. Annual Review of Control, Robotics, and Autonomous Systems, Vol. 3, No. 1, pp 297-330.

161 Shepard RN. 1987. Toward a universal law of generalization for psychological science. Science 237:1317-1323.

162 Bagnell JA. 2015. An invitation to imitation. Tech. rep., CARNEGIE-MELLON UNIV PITTSBURGH PA ROBOTICS INST.

takes task constraints into consideration to improve the DMPs generalization[163]. The implementation of Ijspeert et at[164] DMP algorithm may lead to robot singularities when deployed onto a real cobot arm during the reproduction of the demonstrated movements. The singularities are the configurations in which the robot end-effector becomes blocked in certain directions. Overcoming this problem might be achieved by improving the algorithm that calculates the variable joint parameters to place the end effector of the cobot in a given position and orientation relative to the start of the chain.

6. Distributed Manufacturing

The industrial scenario is undergoing a profound transformation. The emergence of digital technologies led to the analysis and review of the entire production process. The Germans even define this phenomenon as Industry 4.0, the Americans as "smart manufacturing" and in France, people speak of "industry of the future"[165]. Regardless of the term adopted, the idea is the same: the industry is on the verge of a real revolution.

In a world where communication between machines (with other devices or with the environment) is permanent, the need for each industrial organization to develop its management, organization and transformation of data flow into excellent ideas is growing. Is necessary to improve its industrial performance through the connection between operators, machines and processes, to significantly improve performance. Additionally, with an estimated volume of data of 2.5 Exabyte per day and with increased connectivity across all manufacturing processes[166], the ability to organize and transform dispersed data into business ideas becomes a competitive advantage for companies. With the recurring entry and exit of data, it becomes necessary to identify new ways of interacting with the machines and the industrial environment, through the creation of interfaces to facilitate the daily lives of employees and improve their productivity.

With the growth in demand and with increasingly demanding standards[1], advanced technologies are becoming the best solution for obtaining quantitative and qualitative improvements in production processes (e.g. increase productivity and flexibility with new generation technologies: collaborative robotics, additive manufacturing, advanced materials).Typically, companies use products of other companies (or extract them from nature) to be used as components or raw materials to be incorporated in their products. In this logic, a company specializes in a finite group of specific products and does not produce a final product (available to the customer or final consumer) from raw materials, which draws from nature, because it has all the necessary resources to perform all tasks in all phases necessary to obtain the raw material and the final finished product. In the past, manufacturing has meant centralized, large-scale processes with long lead times.

---

163 Y. Zhou and T. Asfour, 2017. Task-oriented generalization of dynamic movement primitive, IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, 2017, pp. 3202-3209.

[164] Ijspeert A., Nakanishi J., Hoffmann H., Pastor P., Schaal S. Dynamical Movement Primitives: Learning Attractor Models for Motor Behaviors. Neural Computation, 25(2):328–373, February 2013

165 McKinsey&Company. (2017). The great re-make: Manufacturing for modern times. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Business Functions/Operations/Our Insights/The great remake Manufacturing for modern times/The-great-remake-Manufacturing-for-modern-times-full-compenium-October-2017-final.ashx.

166 Desjardins, J. (2019). How much data is generated each day? | World Economic Forum. Retrieved May 12, 2020, from World Economic Forum website: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/.

Distributed manufacturing (DM) is changing the paradigm of manufacturing entirely. It includes the concepts of Industry 4.0 and smart manufacturing but adds decentralization to the concept. The new concept is enabled by the emergence of cloud-based digital platforms and driven by the need of highly flexible, customized and material efficient production[167]. Distributed manufacturing means manufacturing processes that are distributed in several geographically separate but digitally coupled locations.

The concept relies on digital platforms managing the production to meet supply and demand in nearly real time and on sensors and data from production systems to manage the processes and manufacturing techniques such as 3D printing to create the products[168].Distributed manufacturing scenarios have the following key characteristics[169]:

- **Digitalization** – Digitalization is essential in Distributed Manufacturing. It permits a product to exist in a virtual form, ready to be physically rendered at any time. This means that it can be potentially produced anywhere given the local availability of resources and access to the new production technologies. New production technologies, because they can operate at a small scale and possess the agility that implies, permit a proliferation in the number of production sites, as well as less restrictions on where they might be located.
- **Personalization** – Customization and personalization are direct consequences of digitalization, which facilitates the modification, both subtle and extensive, of physical products.
- **Localization** – the greater number of dispersed locations opens up new business models and logistics challenges. Together with cloud manufacturing services, it is a precursor of resilience, as manufacturing can still take place even if problems occur in particular locations or even globally. The emergence of distributed manufacturing is examined as a new form of localized production, distinct from previous manifestations of multi-domestic and indigenous production.
- **Enhanced user and producer participation** – the localized and digital nature of distributed manufacturing allows mass customization of products. On the other hand, mixed global and local manufacturer supply chains. Also known as collaborative manufacturing, defined as a new business paradigm where some companies work together and use their own expert competencies to manufacture a product in order to achieve the overall business network performances. There are many kinds of collaboration in manufacturing that have been implemented by companies for competitiveness (e.g. project planning and scheduling, product design and development, forecasting, production systems and supply chains).

It can be understood as: "*technology, systems and strategies that change the economics and organization of manufacturing, particularly with regard to location and scale*". Manufacturing components in different physical locations and then managing the supply chain to deliver them together for final assembly of a product is also considered a form of Distributed Manufacturing. In terms of processes and technologies, a core component of a DM system is a focus on advanced technological developments (e.g. automation and robotics, additive manufacturing) that could potentially enable a much more integrated manufacturing system to be created.

---

167 https://systemsinnovation.io/distributed-manufacturing-how-it-works/ [Accessed at March 4, 2021]

168 Sells, Ed, Zach Smith, Sebastien Bailard, Adrian Bowyer, and Vik Olliver. "Reprap: the replicating rapid prototyper: maximizing customizability by breeding the means of production." Handbook of research in mass customization and personalization (2010).

169 Srai, Jagjit & Kumar, Mukesh & Graham, Gary & Phillips, Wendy & Tooze, James & Ford, Simon & Beecher, Paul & Raj, Baldev & Gregory, Mike & Tiwari, Manoj & Ravi, B. & Neely, Andy & Shankar, Ravi & Charnley, Fiona & Tiwari, Ashutosh. (2016). Distributed Manufacturing: scope, challenges and opportunities. International Journal of Production Research.

6.1. Production Scheduling

The factory of the future (FoF) is the epicenter of a distribution chain that combines customers, suppliers, distributors and partners with advanced analytical systems. This can lead to "perfect production" with minimal downtime, neglect, waste and inefficiency. Heizer & Render[170], speak not only of the importance of innovation but also of the importance of a customer-oriented production system, that is, of basing production planning on orders and not on forecasts. For these authors, the best way to proceed is to divide the production planning process into three distinct levels, and make the best possible decision for each of these levels: strategic planning, tactical planning and operational planning. The level of strategic planning concerns long-term decisions that provide the company's overall direction. Therefore, the current situation of the company, the goals to be achieved and their strategies to be followed must be characterized. Examples of strategic level decisions include product design, choice of suppliers, location of facilities, among others[171] These types of options are reviewed, usually, either between quarters or once a year. These are more specific plans than the previous ones, which work in each functional area of the company, serving as a kind of link between strategic and operational plans. Examples of such plans include purchasing and production decisions, inventory policies or transport policies. The last of the three levels of planning is the level of operational planning, which serves as a regulator of day-to-day decisions, thus having a short-term planning horizon. Scheduling (scheduling), routing (route generation) and truck loading (product loading) decisions are examples of decisions that need to be made when a company's operational planning is carried out.

The scheduling of production can be considered, nowadays, a very studied topic, being focused on by several authors[172,173]. It can be defined a:

> *"(…) decision-making process that is used regularly in many manufacturing and service industries. It is about allocating resources (machines, processing units, crew, and others, depending on the system they are in) to tasks (operations performed through the use of resources), in a given time, to optimize one or more objectives, which can also take different forms, such as, for example, minimizing the total execution time or minimizing the number of tasks that are completed after the scheduled date*"[174].

Le Pape[175], states that given a set of resources with known capabilities, a set of activities with known processing times and resource requirements and a set of time constraints between activities, a "pure" scheduling problem consists of deciding when to perform each activity, considering the

170 Heizer, J. H., & Render, B. (2006). Operations management. Pearson Prentice Hall.

171 Framinan, Jose & Leisten, Rainer & Ruiz, Rubén. (2014). The Context of Manufacturing Scheduling. 10.1007/978-1-4471-6272-8_2.

172 Toptal, A., Sabuncuoglu, I., & Eguï Toptal, A. (2010). Distributed scheduling: a review of concepts and applications. International Journal of Production Research, 48(18), 5235–5262. https://doi.org/10.1080/00207540903121065.

173 Turner, C., Moreno, M., Mondini, L., Salonitis, K., Charnley, F., Tiwari, A., & Hutabarat, W. (2019). Sustainable production in a circular economy: A business model for re-distributed manufacturing. Sustainability (Switzerland), 11(16). https://doi.org/10.3390/su11164291.

174 Pinedo, M. L. (2012). Scheduling: Theory, algorithms, and systems: Fourth edition. In Scheduling: Theory, Algorithms, and Systems: Fourth Edition (Vol. 9781461423). https://doi.org/10.1007/978-1-4614-2361-4.

175 Artigues, C., Demassey, S., & Néron, E. (2008). Resource-Constrained Project Scheduling: Models, Algorithms, Extensions and applications. In F. Sourd (Ed.), CAM - Control Systems, Robotics and Manufacturing Series.

satisfaction of time and resource constraints. Kendall[176] mentioned some of the most common objectives in scheduling problems: minimize makespan, which represents the total execution time (of the entire process). It is a criterion normally used to measure the level of machines uptime. It can be measured by the difference between the instant when the last task finishes processing and the instant when the first one starts.

- Minimize the waiting time for each task (production sequence operations), that is, the time it takes since a product is ready to start a task and it starts.
- Minimize the costs of carrying out the different activities.

There is the possibility of making certain simplifications, such as, for example, there are no restrictions on the availability of components, which are available in zero time. It can also be assumed that the machines are always available for the programming period considered, assuming that there is only one route available for each product, among others. This makes solving a problem quicker and simpler, however, in general, it leads to more modest quality solutions[177].

6.1.1. Key concepts in scheduling

In all scheduling problems[178], both the number of tasks and the number of machines is considered finite, with the number of tasks denoted by N and the number of machines by the letter M. The indices i and j refer to specific machines and tasks, respectively, with the pair (i, j) referring to the processing stage of the task j (j = 1,…, N) on the machine i (i = 1,…, M).The processing time, $P_{i,j}$, represents the time to perform the task j on the machine i. The machine index can be omitted if the job's processing time does not depend on the machine on which it is performed, or if the task can only be assigned to a single machine. Associated with each task, the release and due date, denoted by R_j and for D_j respectively. The release date, R_j, corresponds to the time when the task j can be started. The due date, D_j, corresponds to the instant when the task must be finished. The latter, if not met, may lead to penalties for the operators responsible for carrying out the task in question. The weight, W_j, of a task is a priority factor, that is, it is a measure that denotes the importance of the task j in relation to the other tasks of the system.

As far as the concept of parallel machines is concerned, we can have several types of machines operating in a given factory. It is considered that the machines operate in parallel, when they are identical, any one of them can perform the same operation required by a task. However, even between identical machines in parallel, which are designed to perform the same task, more efficient machines can coexist than others (with shorter production times). This is an important factor to consider when scheduling in that work section.

Since scheduling problems with parallel machines occur in many real cases, this is a target area for many studies. The authors say they are embarking on a sub-area that is not well studied, this being the area of production scheduling problems with setup times dependent on the sequence, in an environment characterized by machines in parallel and that are not related (that is, not necessarily carry out the same operation).

The precedence restrictions can appear in any factory environment. Makarychev & Panigrahi[179], state that scheduling tasks using precedence restrictions for a set of identical machines, with the aim of minimizing total processing time, is a fundamental problem of combinatorial optimization.

---

176 Kendall, G. (2005). Multidisciplinary Scheduling: Theory and Applications: 1st International Conference, MISTA ’03 Nottingham, UK, 13-15 August 2003. Selected Papers.

177 Bandyopadhyay, S. (2020). Production and operations analysis: traditional, latest, and smart views. CRC Press.

178 Pinedo, M. L. (2012). Scheduling: Theory, algorithms, and systems: Fourth edition. In Scheduling: Theory, Algorithms, and Systems: Fourth Edition (Vol. 9781461423).

179 Makarychev, K., & Panigrahi, D. (2014). Precedence-constrained Scheduling of Malleable Jobs with Preemption. Retrieved from http://arxiv.org/abs/1404.6850.

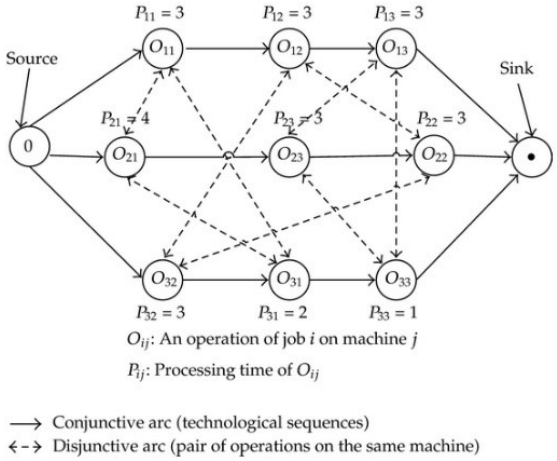

Figure 18 - A disjunctive graph problem, (source: https://www.hindawi.com/journals/jam/2012/651310/fig2/, accessed at April 28, 2020).

These require that one or more tasks have to be completed before successor tasks can be started (see Figure 18). The author defines three types of precedence restrictions, namely: chain; *intree* type; *outtree* type. The first, in a chain, identify situations in which each task has at most one predecessor and one successor. If each task has at most one successor, the *intree* type precedence restrictions are defined. Finally, if each task has at most one predecessor, we are faced with *outtree* restrictions. The last relevant concept concerns recirculation.

Recirculation is a term that comes up often related to job-shop scheduling, and is therefore important[180]. Recirculation can occur in industries that use the normal or flexible job-shop strategy, when a product passes through the same machine more than once during its production process. In the case of a flexible job-shop strategy, there is recirculation if a task visits the same work center more than once along its route.

The possible representation of a Job Shop problem could be done through a Gantt chart or through a Network representation. Gantt (1916) created innovative charts for visualizing planned and actual production. According to Cox, a Gantt chart is, “the earliest and best-known type of control chart especially designed to show graphically the relationship between planned performance and actual performance”. It measures activities by the amount of time needed to complete them and use the space on the chart to represent the amount of the activity that should have been done in that time.

A Network representation was first introduced by Roy and Sussman[181]. The representation is based on “disjunctive graph model”. This representation starts from the concept that a feasible and optimal solution of Job shopping plan can originate from a permutation of task’s order. Tasks are defined in a network representation through a probabilistic model, observing the precedence constraints, characterized in a machine occupation matrix M and considering the processing time of each tasks, defined in a time occupation matrix T. In order to obtain a scheduling solution and to evaluate makespan, we have to collect all feasible permutations of tasks to transform the undirected arcs in directed ones in such a way that there are no cycles (see Figure 18).

180 Pinedo, M. L. (2005). Machine Scheduling and Job Shop Scheduling. In Planning and Scheduling in Manufacturing and Services (pp. 81–113).

181 Marcello Fera, Fabio Fruggiero, Alfredo Lambiase, Giada Martino and Maria Elena Nenni (2013). Production Scheduling Approaches for Operations Management, Operations Management, Massimiliano M. Schiraldi, IntechOpen.

6.1.2. Approaches

Pinedo defines several scheduling classes motivated, in turn, by three concepts, namely: the sequence, the plan and the scheduling policy. Sequence, normally, corresponds to an ordering between a certain number of tasks or else the order in which the tasks must be processed on a given machine. The plan usually corresponds to the allocation of tasks to a given set of machines, at certain times of time. Finally, the scheduling policy appears in scenarios of stochastic configurations, and aims to "prescribe" the most appropriate action for all states in which the system can be found. Deterministic models only consider the sequence and the plan. Thus, the first scheduling class, the Non-Delay Schedule, appears when there are free machines whenever an operation needs to be processed (see Figure 19).This scheduling class prohibits unforced idleness. The second class is the active scheduling that is imposed when it is not possible to build another scheduling, through changes in the order of the processes on the machines, which lead to at least one operation to finish earlier and none to finish later than in the start scheduling. The third and last class is that of semi-active scheduling. This class identifies escalations in which no operation can be completed earlier without changing the processing order on any of the machines.

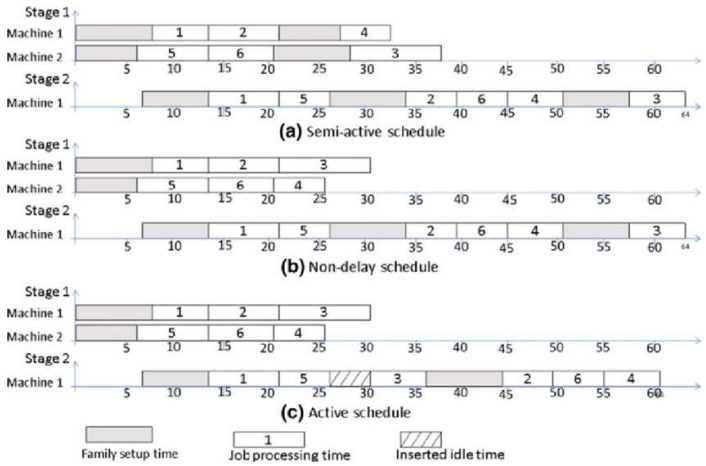

Figure 19 - Gantt example chart of semi-active non-delay and active schedule (Luo, Hao 2013).

Different strategies can be used to schedule production activities, and their application leads to processes with different structures and characteristics. These structures are instigated by the choice of strategy, and range from the project to the flow-shop, through the job-shop and batch, for example. The two main strategies that can be used, which are at opposite points in terms of differences in features and structure, are the job-shop and the flow-shop. In several production and assembly facilities, each task has to subject to a series of operations, and often, these individual operations have to perform in the same order of execution[12]. In this case, it can be said that all tasks have to follow the same route. It is then assumed that the machines are placed in series, and the surrounding work environment is referred to as flow-shop. Graham Kendall[8], with a similar opinion, defined the flow-shop strategy as a linear structure in which the different machines are organized in series of operation / task.

Thus, there is an initial machine where the operations begin, then going through several intermediate machines until they reach the last one, where the process is completed. The fact that the structure is linear does not imply that operations have to use all available machines, that is, an operation can start by using the first machine and then going to the third without going through the second. The main characteristics of this type of strategy: *(a) High standardization and speed; (b) Reduced material handling; (c) Flows with short periods of time; (d) Reduced unit processing costs; (e) High investment costs and mass production needs; (f) Specialized equipment and low skilled workers; and (g) Prevents flexibility*. The authors generalize the flow-shop strategy, adding this type of scheduling to an environment characterized by parallel machines. Here, instead of series machines, phases / stages are defined in series, and each of these steps requires a certain number of machines in parallel. The steps must respect a pre-defined order, and each task is performed by only one of the machines belonging to each step, and all machines that are within the same phase can perform the task for which this phase is intended.

When the routes are fixed, but are not necessarily the same for each task, the model is called job-shop. This model has the advantage of admitting a phenomenon, common in the real world, known as recirculation, that is, when a task has to pass through the same machine more than once before the process is completed. This strategy requires defining the order of the set of operations for each machine, that is, finding precedence conditions between operations[182]. For the authors, the job-shop strategy does not have the same flow restrictions as the previous, flow-shop strategy. In this case, operations can be performed on the machines in any order.

Also, we can highlight, the main characteristics of this strategy: *(a) Wide variety of customized products; (b) Flexible features; (c) specialized human resources; (d) "Mixed" flows; (e) High material handling; (f) High inventories; (g) High flow times; (h) Highly structured information systems; and (i) High costs per unit of product but investment is low*. The Figure 20 compares through a graphical representation, a flow-shop strategy with a job-shop strategy. In this it can be seen that while in a flow-shop strategy all products execute the same type of route (in the same direction), and may not just go through all machines, in the job-shop strategy there is a route that has the opposite direction, or that is, routes are customized to products.

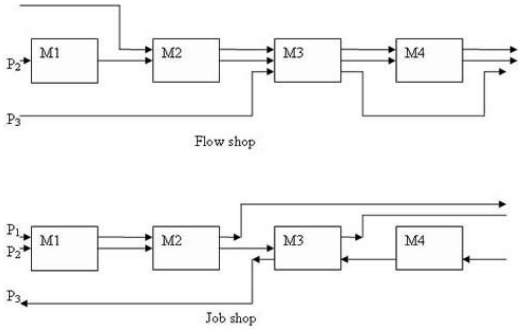

Figure 20 - Illustration of two modes of processing: flow shop and job shop (source: http://www.rvholon.cimr.pub.ro/distrib_prod.html, accessed April 28, 2020).

In the flexible job-shop strategy, there is a mix between the job-shop strategy and environments with parallel machine. Instead of serial machines, there are a number of work centers, each containing a number of identical machines in parallel. In this case, each task has an associated route and has to be processed in all work centers, but in only one of its machines, and all machines in a given center can perform the task to which the product must be subjected.

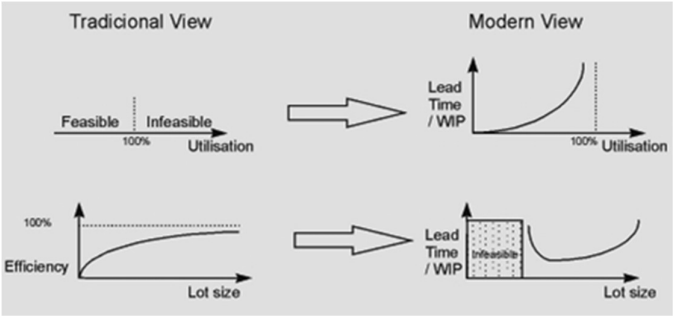

Figure 21 - - The new "vision" of productive capacity (source: X-CITTIC, 1997).

A third possible strategy, is the open-shop. In this strategy, each task has to be processed on each of the available machines (in which *n* jobs must be executed once at each of the c ≥ 2 stages (or

182 Yamada, T., & Nakano, R. (1997). Job-shop scheduling. In A. M. S. Zalzala & P. J. Fleming (Eds.), Genetic algorithms in engineering systems. The Institution of Electrical Engineers.

machine centers without interruption)[183]. A stage consists of a number of parallel machines, and at least one of these stages includes more than one machine. A job has to be processed at each stage by using only one of the machines. The sequence of each stage that processes jobs and the route of each job passing through the stages can be chosen arbitrarily. The objective is to find a schedule that simultaneously determines machine processing orders and job visiting routes to optimize some criteria, such as makespan or total completion time (see Figure 21).

> *"In an open shop, a set of jobs has to be processed on **m** machines. Every job consists of **n** operations, each of which must be processed on a different machine for a given processing time. The operations of each job can be processed in any order. At any time, at most one operation can be processed on each machine, and at most one operation of each job can be processed. Moreover, every job has a release date, only after which the operation of that job can be processed."*[14]

In the various production environments, mentioned previously, not only the number of resources and their disposition can vary. There are other additional features that can make a problem more complex, as described on Table 3.

Table 3 production scheduling constraint additional features

Additional feature	Description
Failure Interruption	The process of executing a job can be interrupted and it can be finished later and even on another machine or equivalent production resource. In the case of parallel machines, the same job can even be performed on two different machines.
Machines Availability	There may be several reasons why a machine is found unavailable, and such situations are required to consider when production scheduling. Not only can the machine be unavailable due to the allocation of other jobs, there may be a breakdown or maintenance period. Periodic maintenance, despite creating periods of machine downtime, reduces unexpected downtime caused by breakdowns.
Associated Priorities or Penalties	Associated with each job there may be a value that can mean a priority (urgency with which the job must be done), or a penalty (lost value if its execution does not correspond to the delivery time). The penalty may also Associated with advancing the performance of a work a certain amount, which may be the associated cost (for example, costs associated with storage) or symbolic, but it leads to the work being carried out as close as possible to the date of delivery. The arrival of high priority jobs in the system can lead to a particular machine being interrupted, being an unexpected situation that, like malfunctions, lead to the reorganization of the scheduling plan.
Precedence Constraints	The processing of a job may depend on the execution of another one, that is, a job can only be executed when the other one finishes its execution. When there is a dependency relationship between jobs, they are considered dependent jobs. This dependency may be due to a more complex job being able to be divided into several tasks or activities and these have to be carried out in a certain order[184].

183 Bai, D., & Tang, L. (2013). Open shop scheduling problem to minimize makespan with release dates. Applied Mathematical Modelling, 37(4), 2008–2015.

184 Kozik, A. (2017). Handling precedence constraints in scheduling problems by the sequence pair representation. Journal of Combinatorial Optimization, 33(2), 445–472.

Machine Setup Time	The time to prepare the machines (setups) is important to consider, as this can be dependent on the sequence of jobs, that is, the setup between jobs can vary depending on the jobs in question. The existence of setup times for machines (setups), can also vary depending on the resource, in the case of uniform parallel machines or unrelated parallel machines, also being a factor to be considered in the scheduling.
Auxiliary Resources	The auxiliary resources can be the most varied, ranging from tools, fixing devices and manipulators to human operators who assist the main resource. There may also be a need for these to be considered, for example in the allocation of human operators or subcontracting period.

At this point, two concepts are distinguished that often appear in approaches to scheduling problems: simulation and optimization. Other important concepts are also relevant, such as dynamic and stochastic programming, and discrete and continuous time models.

Simulation can be seen as a methodology, which has been widely studied, and which emerges with the aim of developing new tools and improving existing ones, so that they can represent and solve various problems more effectively and efficiently, as is the case of production scheduling[185].

One of the main problems of this methodology is related to the simplifying hypotheses introduced by the users in order to reduce the simulation time and the complexity of the system. This procedure can lead to results that are not as accurate as they should be. As already mentioned, another method of dealing with the complexity of the system is to work on a higher level of abstraction which, in turn, can also have negative impacts. Often, the lower level details of the system can be overestimated or even forgotten, which compromises information and makes the system unable to be implemented in reality.

Optimization is one of the main tools used to solve scheduling problems[186],[187]. Operational Research is applied to problems that aim to conduct and coordinate the operations that take place within an organization. For these authors, optimization appears as a characteristic of Operational Research, since it often tries to find an optimal solution (not "the optimal solution" since there may be several solutions considered alternatives) for the target problem of the study.

In other words, the objective is to identify an action plan, among a set that includes the best possible action plans. We can assume that optimization is defined as a mathematical discipline that involves finding the minimum or maximum of functions that are subject to restrictions, therefore involving mathematical formulation. Today, optimization contains a wide variety of Operational Research, artificial intelligence and computer science techniques, and is used to improve business processes in virtually all industries[188].

6.2. Distributed Production Scheduling

In the course of globalization, many enterprises change their strategies and are coupled in partnerships with suppliers, subcontractors and customers. This coupling forms supply chains

185 Hoover, S. V. . P. R. F. (1989). Simulation: A Problem-Solving Approach (1st edition). Prentice Hall.

186 Georgiadis, G. P., Elekidis, A. P., & Georgiadis, M. C. (2019). Optimization-based scheduling for the process industries: From theory to real-life industrial applications. Processes, 7(7).

187 Davis, W. J., & Jones, A. T. (1989). Techniques in Real-Time Production Scheduling.

188 Usuga Cadavid, J. P., Lamouri, S., Grabot, B., & Fortin, A. (2019). Machine learning in production planning and control: A review of empirical literature. IFAC-PapersOnLine, 52(13), 385–390.

comprising several geographically distributed production facilities. Production planning in a supply chain is a complicated and difficult task, as it has to be optimal both for the local manufacturing units and for the whole supply chain network[189]. In the production and distribution environment, there are companies that have physical resources used to produce, transport and store products. In this sense, a company is an autonomous entity (with decision-making capacity and own actions), which owns and exploits resources, with the objective of making specific products available to other companies, or to the final consumer. In this form, the company will depend on others corporate customers, who consume their products and suppliers, which supplies the company, the products it consumes.

This interrelationship between enterprises is an important aspect to be considered. Another equally important aspect is the geographic decentralization. By diminishing restrictions to international trade and compliance the standards and laws, any product can be sold widespread. The same can be stated with production/manufacturing since the means of production are increasingly accessible anywhere in the globe: it is common practice, products to be developed in one site/country, engineering of the production process carried out in another, and production in yet another.

6.2.1. Definition

For the production and distribution of most of the tangible goods it is necessary to have a network of coordinated activities. Traditionally, there are two basic models, according to which these activities can be classified: the market and the hierarchy[190]. In a market model, there are groups of experts who negotiate with each other in an open market. They agree on prices and establish contracts with each other, which normally limit mutual obligations for a certain period of time. After this period, they are free to associate with other specialists. On the other hand, the hierarchy model, the network of activities, is contained in a hierarchical command structure. the flow of products is planned centrally, and there is no competition or redundancy in the structure.

With the globalization of markets and the pressure of competition, mainly small and medium-sized companies are forced to specialize, cooperate, plan and control their activities in a coordinated way, organizing themselves in supply chains. This leads to a decrease in response times, given the consumer's needs. The trend is towards greater coordination and integration of the activities of these players in the value chain, both in terms of configuration and control[191], which are increasingly supported by ICT.

The evolution of these support technologies helps in the transformation of products, processes, companies and competition between companies. Currently, most of the production processes are performed by a network of multiple companies. previously limited typically to particular domains of business, as is the case of the automotive industry, the organization in a cooperative network, is today a trend that has become widespread and extended to other domains, covering production, distribution and services. However, it is noted that concerning the coordinated planning of logistics activities in the production and distribution network, there are interesting studies from the middle of the 20th century[192]. Even so, only more recently, in the 90s, attention has been given to the problem

189 Saharidis, Georgios K.; Dallery, Yves; Karaesmen, Fikri. Centralized versus decentralized production planning. RAIRO - Operations Research - Recherche Opérationnelle, Volume 40 (2006) no. 2, pp. 113-128.

190 World Economic Forum. (2017). Technology and Innovation for the Future of Production: Accelerating Value Creation. Retrieved from www.weforum.org.

191 Baines, T. S., Lightfoot, H. W., Benedettini, O., & Kay, J. M. (2009). The servitization of manufacturing: A review of literature and reflection on future challenges. In Journal of Manufacturing Technology Management (Vol. 20, pp. 547–567).

192 Bandyopadhyay, S. (2020). Production and operations analysis: traditional, latest, and smart views. CRC Press.

of escalation in decentralized contexts. Toptal[193] defines distributed scheduling as "*an approach in which smaller parts of a scheduling problem are solved by local decision makers who possibly have conflicting objectives and who coordinate their sub solutions through certain communication mechanisms to achieve overall system objectives*".

Cooperative production and distribution network presuppose a group of companies linked by customer-supplier relationships, which perform production and distribution activities (acting as suppliers, producers, distributors and retailers), cooperating in placing a product or final products on the market. The cooperative network nodes correspond to specialized companies, which perform various types of tasks, phases or stages (of production, storage, transport) complementary to the production and distribution process. By specializing their skills and organizing themselves in cooperative networks, companies can obtain competitive advantages over the so-called monolithic competition, as they can share technical and commercial information, act in a coordinated way in the introduction of new products, reduce costs that arise from the need security in the face of their uncertainty, and reduce response time in the face of variable demand. the objectives of the network can be generically translated by the maxim of logistics: placing the right product and quantity at the right time and place[194].

The organization of companies in cooperative networks combines the advantages of both models of organization, market and hierarchy[195]. On the one hand, the cooperative network is a relatively durable, less volatile and transient organization than contracts negotiated at frequent intervals between participants in a market model. It, therefore, allows for greater planning and control, but avoids the inertia and rigidity of this model (e.g. Figure 22).

It facilitates competition and price transparency, which characterizes the market model, avoiding adversity and reduced vision originating in the market context. This combination, together with the support of ICT, is at the origin of management paradigms such as Rapid Response, Accurate Response, Integrated Supply Chain Management, Agile Production, Virtual Company, and Extended Company[196].

193 Toptal, A., Sabuncuoglu, I., & Eguï Toptal, A. (2010). Distributed scheduling: a review of concepts and applications. International Journal of Production Research, 48(18), 5235–5262.

194 Benoit Lung, & Laszlo Monostori. (2007). Manufacturing scheduling and control in the extended enterprise. In Marco Taisch, Klaus-Dieter Thoben, & Marco Montorio (Eds.), Advanced Manufacturing. An ICT and Systems Perspective.

195 Víctor Fernández. (2012). Modelling And Optimization Of Flexible Manufacturing Systems. Bragança.

196 Eyob, E., & Tetteh, E. G. (2012). Customer-oriented global supply chains: Concepts for effective management. In Customer-Oriented Global Supply Chains: Concepts for Effective Management.

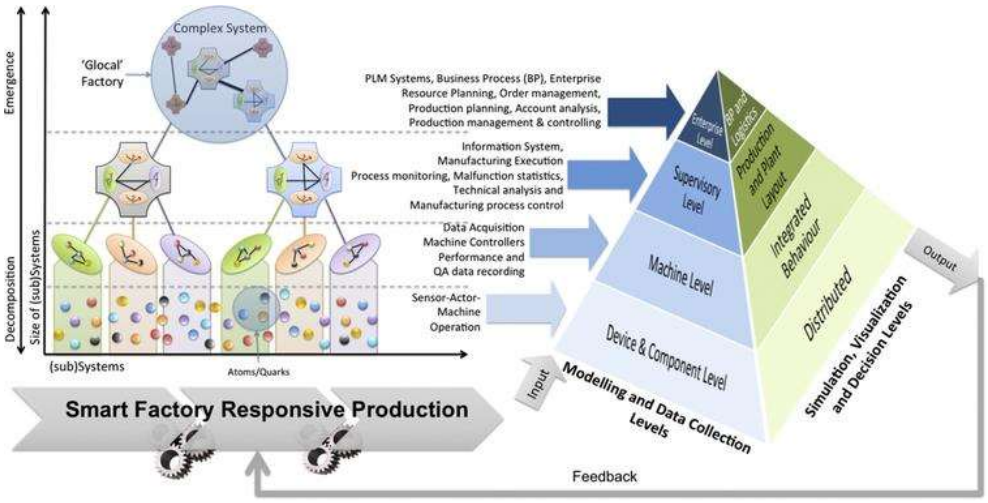

Figure 22 – example schema of decentralized decision support for intelligent manufacturing in Industry 4.0[197].

To be successful, a company must consider the reduction of response times, in addition to increasing the quality and flexibility of its products/services. This concern must be linked to all iterations of the business process, and not only within their organizational boundaries but in the industry's supply chain. There must also be a measure, not only the coordinated production and distribution of products but also the cooperative development of new products (e.g. Extended Company, or Virtual Company models)[198,199].

The extended company is defined as a configurable and transversal fusion of the business processes of different business units, to form a system that places certain products on the market. This system is a network whose nodes are specialized companies, acting simultaneously as customers and suppliers, where coordination is ensured through inter-company communication, supported by electronic networks and ICT resources. This concept is applied to production or service business units, which are geographically distributed, capable of operating as a controlled and planned system for making products/services available on the market. According to this model, the coordination of tasks must be ensured by a supervisory unit, connected to all nodes in the network (it is a team where each company involved is represented).

It acts as a virtual node or network coordinating agent that integrates the individual perspectives of the nodes, from a global perspective of the network. It presupposes an infrastructure that enables fast communications and the sharing of common data, necessary for coordination. In the context of the project, it can be assumed that a scaling environment is one, in which there is a multi-product production and distribution network (with multiple end products), whose activities are coordinated by an extended company. Each of the nodes in the production and distribution network can deliver, or make available, a finite set of products, at the expense of the consumption of certain materials or components, performing production and distribution tasks (production, storage or transportation). nodes are macro-resources that correspond to factories, warehouses or transport units, whose

197 Marques, Maria et al. 'Decentralized Decision Support for Intelligent Manufacturing in Industry 4.0'. 1 Jan. 2017: 299 – 313.

198 Benoit Lung, & Laszlo Monostori. (2007). Manufacturing scheduling and control in the extended enterprise. In Marco Taisch, Klaus-Dieter Thoben, & Marco Montorio (Eds.), Advanced Manufacturing. An ICT and Systems Perspective.

199 Ed Davis, & Rober Spekman. (2003). Extended Enterprise: gaining competitive advantage through collaborative supply chains (T. Moore, ed.). Prentice Hall PTR.

capacity is managed in a coordinated manner, by the participating companies of the extended company.

These entities may be different departments in a company, cells in a flexible manufacturing system, multiprocessors in a communication network, or companies in a supply chain[200]. Decision making in such a distributed manner increases system responsiveness, which may be especially important for scheduling. Scheduling as a short-term decision-making process requires up-to-date and timely information to generate feasible schedules in practice.

6.2.2. Approaches

The proliferation of cyber-physical systems introduces the fourth stage of industrialization, commonly known as Industry 4.0. The vertical integration of various components inside a factory to implement a flexible and reconfigurable manufacturing system, i.e., smart factory, is one of the key features of Industry 4.0[201].Smart factories are being shaped into smart shop-floors (i.e. powered by context-awareness), with autonomous agents (i.e. smart objects) and the scheduling activity of logistical tasks is seen from the perspective of the Multi-Agent paradigm, of Distributed Artificial Intelligence[202].

This activity involves a multi-agent coordination system, in which the agents cooperate to achieve joint objectives and meet the final demand over time, respecting capacity restrictions and scheduling schedules. However, there is space for competitive scheduling activity: agents will be able to compete in meeting individual scheduling goals or preferences. This philosophy is more similar to human professional relationships, in which each employee has their own goals, but knows that to achieve them, they will have to work collaboratively. In this way, the company's transversal objectives are met in the best possible way, with little supervision effort (e.g. in cases of inter-agent conflicts). Two levels can be classified in this context:

- **The level of physical resources**, gear by the production and distribution network. This network contains nodes (the macro resources), interdependent due to customer-supplier relations and specialized in logistical tasks of production, storage and transportation. To satisfy orders, or orders, tasks are created from abroad, to be staggered at the network nodes. A group network process is a group of tasks linked by time precedence links, which come from customer-supplier relationships between nodes. the scheduling of a task on a node occurs by affecting the capacity of the node to the task in a given time interval. Each node has a limited capacity (production, storage or transport) to affect tasks. At this level, physical product flows occur.
- **The decision level**, made up of cooperating agents. These agents represent the companies participating in the Extended Company and generate the capacity of the nodes in the production and distribution network. Each agent is in charge of managing the capacity of a network node. Customer-supplier relationships between nodes are extended to agents: an agent is both a supplier and a customer of other agents, with whom he can communicate. Each agent can receive orders from client agents. To satisfy requests, the agent schedules the appropriate tasks on the node he manages. In addition, the agent sends the appropriate

200 Víctor Fernández. (2012). Modelling And Optimization Of Flexible Manufacturing Systems. Bragança.

201 Mourinho, J., Leiras, F., Correia, R., Fernandes, M., Conceição, L., Praça, I., & Marreiros, G. (2018). A Vertical Predictive Maintenance Approach for Manufacturing 4.0. In J. T. Farinha & D. Galar (Eds.), Proceedings of Maintenance Performance Measurement and Management (MPMM) Conference (pp. 2–9). FCTUC-DEM.

202 Shoham, Y., & Leyton-Brown, K. (2008). Multiagent systems: Algorithmic, Game-Theoretic, and logical foundations. Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations, 9780521899, 1–483.

orders to supplying agents, to ensure the necessary supplies to perform the tasks. At this level, information flows occur.

In this concept of distributed scheduling, it is a generalization of the problem of conventional scheduling, with decentralized scheduling activity (represented by agents) and inter-agent communication. Therefore, decisions are not attributed to a single central agent, but distributed to a set of agents with an inference / decision capacity, each with the ability to decide on a single resource. additionally, and to avoid deadlocks[203], they communicate with each other to coordinate decisions.

6.2.3. Testing data integrity

Distributed production scheduling is dependent on accessible data networks with high level of data integrity to ensure the scheduling is not hindered by technical problems in the underlying network structure. Current network trends show a tremendous increase in network speeds due to the growing number of connected devices and the increase of data and video traffic. At the same time, the percentage of encrypted traffic grows and is up to 85% in 2020. Encryption increases the confidentiality level of the data transmitted over the networks, however, at the same time it increases the complexity of inspecting the traffic for security risks. Combined with the high network speeds, the encrypted data traffic puts a high strain on security tools. Due to this, only a handful of security tools can manage the task of inspecting encrypted traffic in high data speeds[204].

The standard IEEE 802.1ae that defines MACSec was initially released in June 2006 and has been updated several times; latest update is currently from 2018[205]. MACsec protocol header format is shown in Figure 23[204].

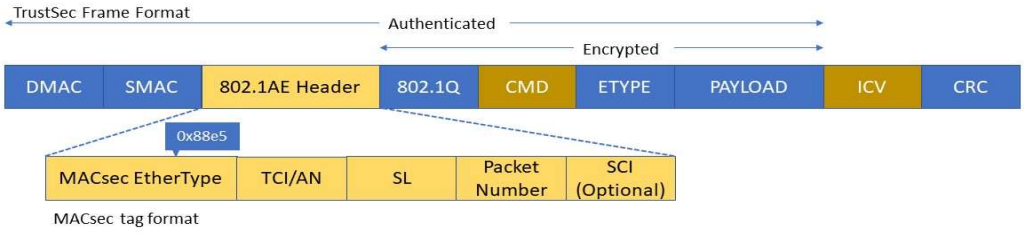

Figure 23 - MACsec protocol header format [206].

MACsec is used to provide point-to-point security on Ethernet links and it is defined by IEEE standard 802.1AE. MACSec operates by performing link layer encryption for each hop through a network. To provide end-to end security, it is often used in connection to other security protocols, such as IP security IPsec and Secure Socket Layer. What makes MACsec important, is its capability to identify and prevent most security threats. It secures an Ethernet link for more traffic than what most other security solutions are able to, including frames from Link Layer Discovery Protocol, Dynamic Host Configuration Protocol, Address Resolution Protocol, and others protocols[207]. An

203 Toptal, A., Sabuncuoglu, I., & Eguï Toptal, A. (2010). Distributed scheduling: a review of concepts and applications. International Journal of Production Research, 48(18), 5235–5262.

204 https://www.fortinet.com/blog/industry-trends/keeping-up-with-performance-demands-of-encrypted-web-traffic [Accessed at February 14, 2021].

205 https://tools.ietf.org/html/rfc2544 [Accessed at February 8, 2021].

207 https://www.juniper.net/documentation/en_US/junos/topics/concept/macsec.html [Accessed at February 14, 2021].

important advantage of MACsec over other security protocols, namely IPsec, is its minimal impact on network performance. The overall throughput of a network router is limited by the performance of the IPsec engine, while a required link speed can be realized when using MACsec encryption[208]. Another driving factor between the shift from IPsec to MACsec is the change in traffic patterns to any-to-any model, which is dictated by cloud, machine-to-machine (M2M) communications and the internet of Things (IoT).

Due to these network trends and risks, it is necessary to test the network functionality and the capability of the installed security elements to ensure data integrity. At the same time, due to the high strain of CU usage posed by the encryption/decryption, MACsec is only now emerging in network testing solutions aimed at high-speed networks. These solutions are required for network engineers fully test networks for the MACsec design and implementation to ensure quality and performance of the network deployment[209,210].

6.3. Market Solutions

In order to manage a large-scale production operation, control software is needed. ERP (Enterprise Resource Planning) systems are used to manage the overall running of the company, and they are linked to MES (Manufacturing Execution System) systems to support manufacturing scheduling. The available MES systems provide a variety of features to support setting up, managing, execution and costing of production operations. There are MES systems available that support discrete manufacturing and options for process manufacturing.

There are several widely used solutions on the market, provided by well-known established companies catering to the needs of multiple types of manufacturing operations[117]. As distributed manufacturing is a new and upcoming concept, the needs for it may vary from those of traditional manufacturing technologies. Traditionally, the overall manufacturing process happens within a single factory, although components and sub-assemblies may be sourced from suppliers. The overall process is oriented for mass production. However, distributed manufacturing is defined by decentralizing the manufacturing process, often shorter production cycles and more customized products[118].

Table 4 Available MES software's and their corresponding websites

Product	Website	Industry 4.0 mentioned
Mastercontrol	mastercontrol.com	No
Aegis FactoryLogix	aiscorp.com	No
Infor Cloudsuite Industrial	infor.com/products/cloudsuite-industrial	Yes
Simio	simio.com	Yes
Parsec TrackSys	parsek-corp.com	Yes
Pinpoint	pinpointinfo.com	Yes

208https://www.cisco.com/c/dam/en/us/td/docs/solutions/Enterprise/Security/MACsec/WP-High-Speed-WAN-Encrypt-MACsec.pdf [Accessed at February 14, 2021].

209 https://www.keysight.com/fi/en/assets/3120-1442/data-sheets/IxNetwork-MACsec-Test-Solution.pdf [Accessed at February 12, 2021].

210https://assets.ctfassets.net/wcxs9ap8i19s/2YiuYFjn66PWMfptZBLauU/296f7ed827098cd8a6bb986469a0dc3b/Landslide_C100-M4_datasheet.pdf [Accessed at February 10, 2021].

A key requirement for smoothly coordinating a distributed, decentralized manufacturing system is detailed real-time view of the production processes, in-process data and resulting quality measures. The data from across the production system must be accessible from all locations, requiring cloud-based access to the data.

The system needs to be flexible to enable co-operation between sites and accommodate for plan adjustments to suit the changing conditions and requirements. The system must readily support geographically distributed multiple plants and production locations and it must be compatible to be integrated with other existing manufacturing and business systems. The system must also support shorter production cycles[119]. According to a study[120], most current systems do not support a distributed manufacturing system[15]. The table 4 lists a number of key MES suppliers and their capabilities in regards to the key requirements of distributed manufacturing.

6.4. Safety assurance for adaptive SoS

6.4.1. Definition

Adaptive SoS are getting more popular due the rapid changes in the industry and their application domains. This is particularly due to their capability for reconfiguration at run time, facilitating more autonomy and reacting more flexible to the changes either in the SoS or in its context. Examples for such adaptive SoS are networks of collaborating machines and transport robots or AGVs in factories of the future. Adaptive SoS are able to react onto internal and external changes, adapting their member systems and reconfiguring the relations between these. Reconfiguration in such example can be applied with the goal of optimization techniques aiming to improve the efficiency of FoF.

Ensuring continued safety for adaptive SoS is challenging, because either the multitude of relevant configurations must be assessed at design time, or assessment must done dynamically at run time.

Adaptive SoS are usually consist of autonomous systems which can decide to be part of the SoS and to build a dynamic connectivity as well as to benefit from cooperation within the SoS, in order to achieve greater fulfillment of their own goals, along with the higher SoS goals[211].

Each individual system may be seen as a Cyber-Physical System and their integration as a Cyber-Physical System of Systems[212]. In such systems, a reconfiguration is the process of changing an already developed and operatively used system in order to adapt it to new requirements, to extend its functionality, to eliminate errors, or to improve quality characteristics. [213]

Reconfiguration is divided in two methods, which we could define as programmed and ad-hoc reconfiguration. Programmed reconfiguration is predefined at design time. In contrast, ad-hoc reconfiguration creates higher flexibility, because the system generates possible configurations at run time. Yet, from an assurance point of view, this method is far more challenging[214].

---

211 Boardman, J., Sauser, B.: System of Systems - the meaning of of. In: 2006 IEEE/SMC International Conference on System of Systems Engineering, Los Angeles, California, USA, 24–26 April 2006. IEEE (2006).

212 Ferrer, B.R., et al.: Towards the adoption of cyber-physical systems of systems paradigm in smart manufacturing environments. In: 2018 IEEE 16th International Conference on Industrial Informatics (INDIN), Porto (2018).

213 Matevska, J.: RekonFiguretion komponentenbasierter Softwaresysteme zur Laufzeit. Vieweg+Teubner Verlag/Springer Fachmedien Wiesbaden, Wiesbaden,Wiesbaden, Wissenschaft (2010).

214 Batista, T., Joolia, A., Coulson, G.: Managing dynamic reconFiguretion in component-based systems. In: Morrison, R., Oquendo, F. (eds.) EWSA 2005. LNCS, vol. 3527, pp. 1–17. Springer, Heidelberg (2005).

There are three levels of flexibility regarding the selection of new configurations in adaptive SoS[215].

1. Predefined Selection: Once a dynamic change is initiated, the system chooses a configuration based on a predefined selection made at design time.
2. Constrained Selection from a predefined set: Suitable configurations are defined at design time in relation to given situations or system states. Once a dynamic change is initiated, the system will select the most appropriate configuration from the set of configurations that matches the current situation.
3. Unconstrained Selection: Once a dynamic change is initiated, the system may choose freely from a multitude of possible configurations, or even generate new configurations at run time.

From the safety perspective, the development of adaptive SoS with programmed reconfiguration is well supported by design-time assurance methods. However, the complexity of covering all potential configurations of the SoS at design time poses a major challenge to these methods, and unconstrained selection cannot be handled purely with design-time assurance methods.
In particular, the openness of adaptive SoS and the dynamic nature of the relations between their members make it impossible to foresee and assess all relevant configurations, unless one strongly restricts the selection flexibility. Alternatively, safety assurance at run time appears as a suitable means to master this challenge, because it could help to restrict the assurance effort to assessing only those configurations which are relevant in a certain reconfiguration situation.[216].

6.4.2. Approaches

Safety Cases are one important assurance approach to be considered in this context. The UK Defence Standard 00-5[217] defines a safety case as “a structured argument, supported by a body of evidence, which provides a compelling, comprehensible and valid case that a system is safe for a given application in a given environment”.

Originally, the safety case methodology was developed to support safety assurance during design time. In order to be truly applicable for adaptive systems of systems - that emerge and change structurally at run-time, and those constituents change at run-time, too - safety case smust be modular and adaptive, too. Different approaches have been developed throughout the years to address modularity and dynamicity for safety cases. The approaches described in literature support several aspects of what is required for safe reconfiguration in adaptive systems of systems, but none of them seems to cover all required elements[218].

The key elements of safety cases based on this definition are:

- **Claims**, defining properties of the system.
- **Evidences**, used as the basis of the safety argument, being facts, assumptions or sub-claims.

---

215 Bradbury, J.S., Cordy, J.R., Wermelingerb, M.: A Survey of Self-Management in Dynamic Software Architecture Specifications. ACM, New York, NY (2004).

216 Mirzaei E., Thomas C., Conrad M. (2020) Safety Cases for Adaptive Systems of Systems: State of the Art and Current Challenges. In: Bernardi S. et al. (eds) Dependable Computing - EDCC 2020 Workshops. EDCC 2020. Communications in Computer and Information Science, vol 1279. Springer, Cham.

217 Ministry of Defence: Defence Standard 00-56: Safety Management Requirements for Defence Systems (2007).

218 Mirzaei E., Thomas C., Conrad M. (2020) Safety Cases for Adaptive Systems of Systems: State of the Art and Current Challenges. In: Bernardi S. et al. (eds) Dependable Computing - EDCC 2020 Workshops. EDCC 2020. Communications in Computer and Information Science, vol 1279. Springer, Cham.

- **Arguments**, linking evidences to claims.
- **Context**, being the environment in which all these safety analysis and arguments are valid.

Traditionally, most people were using textual notations to define safety cases. As structuring of the arguments for complex system became challenging, graphical notations were developed supporting the safety case development. In general, there are different notations supporting structuring the arguments to associate the evidences to the claims. Two of the most popular ones are:

- Goal Structuring Notation (GSN)[219].
- Claim-Arguments-Evidence[220].

GSN well supports capturing the underlying rationale of arguments. This helps to scope areas affected by a particular change and thus helps developers to propagate the change mechanically through the goal structure. However, GSN do not tell if the suspect elements of the argument in question are still valid. Hence, using GSN does not directly help to maintain the argument after a change, but it helps to more easily determine the questions to be asked to do[221].

The basic idea in all these structuring approaches is very simple: In many cases, in order to make a claim about a property of an object, we need to investigate whether the object has this property by evaluating its components. To do this, we need to clarify what the property is, what the rules are regarding how to view the object as being composed of components, and how the properties of these components can be combined (i.e., how reliability properties of components are combined when the degree of independence is not known)[222]. In general, understanding, developing, evaluating, and maintaining safety cases is a non-trivial task due to the volume and diversity of information that a typical system safety case must aggregate when best engineering practice is followed [223].

6.4.3. Limitation of current approaches

The complexity and flexibility of these SoS necessitates the development of new approaches to analyse and maintain safety. One approach to address this challenge could be identifying safety-related system variables at design time, monitoring them at run time, and analyzing their variation for prospective configurations during the configuration selection process. However, this is not yet sufficient to cover completely unconstrained selection, since this approach focuses on known variables of known systems. In a truly unconstrained selection, we will not only see parametric changes to existing variables, but also structural changes. For this, we need a dynamic approach that also can dynamically compose and assess safety cases at run time.

The basic safety case concept has evolved over the years, and has led to amendments and extensions such as the modular safety case (MSC) approach, and the dynamic safety case (DSC) approach. MSC, in particular was introduced to cope with the system complexity by breaking down safety arguments into modules, in order to reduce cost and impact of changes during the system

---

219 Kelly, T.P., Weaver, R.: The goal structuring notation-a safety argument notation. In: Proceedings of the Dependable Systems and Networks 2004 Workshop on Assurance Cases (2004).

220 Bishop P., Bloomfield R.: A methodology for safety case development. In: Redmill F., Anderson T. (eds.) Industrial Perspectives of Safety-critical Systems. Springer, London.

221 Denney, E., Pai, G., Habli, I.: Dynamic safety cases for through-life safety assurance. In: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering, Florence, Italy (2015).

222 Bloomfield, R., Netkachova, K.: Building blocks for assurance cases. In: 2014 IEEE International Symposium on Software Reliability Engineering Workshops, Naples, Italy (2014).

223 Denney, E., Pai, G., Whiteside, I.: Proceedings of the Formal Foundations for Hierarchical Safety Cases, HASE 2015. IEEE, Piscataway, NJ (2015).

life-cycle, whilst DSC were developed to bridge the gap between assumptions taken during design time and properties of the realized system becoming apparent during run time.

There are many methods supporting DSC and many approaches for MSC. However, there is a lack of focus in combining the two approaches addressing the above-mentioned challenges as well as benefiting of both method's assets. A combination of both approaches in the sense of "Dynamic Modular Safety Cases" could be a good solution.

We explored the current state of the art of MSC and DSC and discussed the application of these approaches for dynamic reconfiguration of SoS. We examined their application in two different life-cycle phases at design time and at run time to identify the unexploited features and properties of these approaches in both phases. The limitations of the MSC and DSC approaches for dynamic reconfiguration of adaptive SoS necessitate combining the two methods into an approach called Dynamic Modular Safety Cases to cope with the challenge of unconstrained selection during reconfiguration at run time and benefiting from both method's assets[224].

6.5. Limitations of Current Approaches and Solutions

Production scheduling is the base of distributed manufacturing. Distributed Production scheduling approaches can be broadly applied to several other problems of distributed manufacturing, for example, to the use of different types of resources beyond production equipment's, such as automatic guide vehicles (AGVs) for the transportation of goods or community-based production[225]. In all the cases, however, there is a set of common characteristics of distributed manufacturing: personalization, localization, new enabling technologies and enhanced user and producer participation.

Limitations of current approaches on production scheduling are related to its inability to extract the maximum value of the distributed manufacturing characteristics. Current production scheduling systems are based in the traditional operations research-derived approach, based on an optimization functions and a set of constraints. These constraints must be known upfront. It is based on certain assumptions or forecasts about the availability of assets or materials or customer orders. In highly dynamic and responsive SoS systems such as distributed manufacturing systems, the information changes frequently. Currently, systems try to cope with system dynamics through sensitivity analysis (providing optimal intervals for optimality). Nevertheless, highly dynamic systems can go out of the bounds of the sensitivity analysis and render the scheduling plan suboptimal. Comprehensive production scheduling is often a highly dimensional problem and the search space for optimal solutions can be a problem as it takes a considerable processing time to achieve optimal solutions.

This becomes a more complex problem, if mass customization needs are addressed, where individual orders and preferences are integrated into the problem, potentially exploding the search space, due to the highly combinatory nature of the problem. The known approaches are also rigid, producing rigid solutions, and last-minute or unpredicted changes are difficult to accommodate. For example, if one AGV or equipment fails, the scheduling process needs to be restarted, being time consuming. Additionally, the scheduling process is a top-down, centralized decision-making process. This is not adequate to modern distributed manufacturing organizations which function as a SoS, where decentralized decision making is needed to ensure resilience.

Current approaches are unable to support dynamic real-time negotiation between systems nor decentralized decision making. Overcoming the problems mentioned implies evolving the current

224 Denney, E., Pai, G., Whiteside, I.: Proceedings of the Formal Foundations for Hierarchical Safety Cases, HASE 2015. IEEE, Piscataway, NJ (2015).

225 Srai, Jagjit & Kumar, Mukesh & Graham, Gary & Phillips, Wendy & Tooze, James & Ford, Simon & Beecher, Paul & Raj, Baldev & Gregory, Mike & Tiwari, Manoj & Ravi, B. & Neely, Andy & Shankar, Ravi & Charnley, Fiona & Tiwari, Ashutosh. (2016). Distributed Manufacturing: scope, challenges and opportunities. International Journal of Production Research. 10.1080/00207543.2016.1192302.

scheduling models to fully support the aforementioned characteristics of distributed manufacturing. With such dynamic environments, the constraints to the production scheduling are not fully known upfront. Therefore, there is the need to feed improved information quality to reduce uncertainty to the scheduling models. This will be achieved by making use of machine learning and cognitive analytics to enhance prediction, increase the efficiency of the scheduling models and to derive the best decision.

As an example, AGV fleet requests can be forecasted using regression models and other machine learning techniques, so the AGV fleet can be prepared to deal with potential demand peaks, thus reducing bottlenecks and improving overall efficiency. The behavior of the system components – the equipment availability, the human factors, the goods availability and quality can be forecasted and fed into the scheduling models, increasing their effectiveness and decision quality. The state of the art also proposes several solutions to decrease the search space, either with branch-and-bound techniques, tree pruning, metaheuristics (e.g. GRASP, Tabu Search).

The potential search space explosion can also be prevented by designing new type of algorithms that instead of trying to produce a complete scheduling plan, produce several incremental plans. In highly dynamic environments such as multi-location manufacturing organizations with complex product mixes, the information available changes fast. This means that uncertainty is higher in the beginning of the scheduling plan processing. Therefore, the produced solution may be quickly rendered suboptimal due to the inherent dynamics of the system.

A new type of algorithms that are real time, lighter processing and have a distributed nature are required. They should not provide a complete scheduling plan but deliver incremental results and can process information arriving in real time is needed instead. Current scheduling algorithms are also highly centralized and “top-down”.

Cloud manufacturing, big data and collaborative manufacturing processes are used as tools or resources through platforms to attain the objectives of improved efficiency, reduced lead time [2] and reduced total cost, along with maximum mutual profit precisely and easily in a networked manufacturing system[226].

Cloud services in manufacturing (CMfg) represent an evolution of networked and service-oriented manufacturing models that comprise a set of reconfigurable and interchangeable items on the factory floor, and can access a shared set of computing devices according to cloud computing (CC) principles, in which case CC is part of the CMfg model at the IaaS (Infrastructure as a Service) abstraction level[227].

To overcome traditional manufacturing problems (i.e., the monolithic approach has its advantages, but it is not sufficient in today's dynamic manufacturing environment), it is necessary to integrate both functions and means to achieve better system performance. Subsequently, the need to integrate both activities of these issues has found its basis in the context of the networked and collaborative manufacturing environment. However, no conventional shop floor control system based on a centralized or hierarchical control architecture can handle the necessary adaptive and autonomous control of the manufacturing system.

In distributed manufacturing, however, the several locations and assets have often localization specificities (ex: mixed local and global supply chains, different regulations, different contexts) and are subjected to different cybersecurity realities. The distributed manufacturing organization as a whole is highly based on digitalization for manufacturing and coordination. The robustness and resilience of the organization in face of cyberthreats is, therefore, of maximum importance.

In this context, production scheduling needs to work in a decentralized way, with every location and asset/group of assets being able to negotiate production but also to be able to work independently

[226] [1] T. Borangiu et al., “Digital transformation of manufacturing. Industry of the Future with Cyber-Physical Production Systems,” 2020.
[227] Kubler S., Holmstrom J., Fr ¨ Amling K., Turkama P., ¨ Technological Theory of Cloud Manufacturing, Service Orientation in Holonic and Multi-Agent Manufacturing, Studies in Computational Intelligence 640, pp. 267–276, 2016.

(at least temporarily) from the other parts. Decentralized scheduling also benefits global efficiency through load sharing and economics optimization (e.g. production costs, fluctuation in local conditions) and scalability[228]. However, the biggest advantage of decentralized production scheduling is the ability to support dynamic system reconfiguration in face of the internal or external system dynamics, allowing, for example, faster recovery from resource or equipment unavailability and cyberattacks, allowing the systems to negotiate, reconfigure their relations, data flows and production. Distributed manufacturing organizations should also be able to fully exploit the advantages of digitalization. Therefore, the control architecture is gradually being shifted to the distributed, decentralized and autonomous control (DDAC) architecture. Since DDAC shop floor control system may have complete local autonomy, governing the reconfigurability, scalability as well as fault tolerance, it is suitable for a dynamically changing environment.

To achieve the successful information and knowledge exchange between different facilities, there is a need for internet and communication technology IoT (internet of things) through which it can be possible to link all of them. Some of the key literature reviews for planning and scheduling and their integration, regarding the application of artificial intelligence-based approaches, multi-agent-based simulation, cloud manufacturing, internet of things, big data and digitalization.

In traditional manufacturing, the machines associated with the jobs are located and conditioned in a single workshop or company. However, networked manufacturing jobs and machines are distributed in different workshops or companies located globally at greater distances. Thus, it can be inferred that a networked manufacturing situation is similar to that found in a flexible manufacturing system, where many possible machines, operations are feasible but possibly not on the same shop floor.

Current ERP-solutions lack extended enterprise support and a shared cloud-based approach. On the other hand, current MES solutions can operate the manufacturing process, but not for distributed manufacturing[229]. The answer lies in a cloud-based manufacturing, which can present rich optimization and scheduling services in a decentralized way. It can also leverage the extraction of value from data through new business models based on “as-a-service and data-driven paradigms (ex: AI-as-a-Service).

Intelligent production is a term that encompasses industry 4.0 and smart factories, and is characterized by all cyber-physical systems that allow equipment, resources and people to communicate in an interconnected way and in real time, to exchange information and to optimize processes and resources throughout the manufacturing process. According to a study by Capgemini[230], smart factories could add between 500 billion to 1.5 billion dollars to the global economy by 2022, allowing companies to enjoy an efficiency rate 7 times higher than the growth.

Furthermore, it can be said that in networked manufacturing, generating an optimal process plan for each job in the presence of various dynamic constraints, such as the current state of machines, tools and fixtures, at a given manufacturing site is posing a real challenge.

228 Hedberg, Thomas, Helu, Moneer, and Sprock, Timothy. "A Standards and Technology Roadmap for Scalable Distributed Manufacturing Systems." Proceedings of the ASME 2018 13th International Manufacturing Science and Engineering Conference. Volume 3: Manufacturing Equipment and Systems. College Station, Texas, USA. June 18–22, 2018. V003T02A019. ASME.

229 Petri Helo, Mikko Suorsa, Yuqiuge Hao, Pornthep Anussornnitisarn, Toward a cloud-based manufacturing execution system for distributed manufacturing, Computers in Industry, Volume 65, Issue 4, 2014, Pages 646-656, ISSN 0166-3615.

230 Eyob, E., & Tetteh, E. G. (2012). Customer-oriented global supply chains: Concepts for effective management. In Customer-Oriented Global Supply Chains: Concepts for Effective Management.

7. Conclusions

In this document, optimization techniques aiming to improve the effectiveness, resilience and robustness of the FoF are presented. In doing so, the project aims to boost several capabilities of FoF by improving the shop-floor management (chapter III), improve efficiency, security, safety, and resilience of FOF through machine learning algorithms (chapter IV). Furthermore, from the exploitation of the different data sources in the data lake, data-driven business models are going to be identified and exploit (chapter IV), such models aim to continuously improve both processes and products. Enhance the interaction between humans and machines (chapter V) and improve the management of manufacturing load through a network of factories (chapter VI)

Our first goal, the improvement of the shop-floor management is attained by exploring sensing and tracking technologies, which might be useful on the shop-floor.

In an indoor tracking system, the goal is to get asset location in terms of X Y Z coordinates. We can use this information in different ways, such as asset visualization, georeference and trigger alerts when an asset gets into or out of a concrete area, etc.

The choice for a RTLS are strongly dependent on operating environment constraints and requirements. In Section III, advantages and disadvantages of technologies such as BLE, UWB, Wi-Fi and ZigBee are discussed, with a special highlight on issues such as battery, range, tag collision and security. The chosen technology is UWB for an indoor system, because of its high accuracy (30 cm approximately) and because it operates in industrial (metal) environments.

Furthermore, IIoT platforms are discussed in terms of their connectivity, integration, analytics, application and security. If all platforms offer analytic tools, not all services are at disposal in a single platform. Therefore, the user choice should be based on the features needed, for example, SAP cloud platform is a great choice regarding security but lacks on connectivity and integration modules. The data extracted by a RTLS is going to populate the data lake, using such a data, novel services and management decisions could be better performed. For example, in a robot-fleet scenario, a RTLS might be used to attribute a task to the robot that is closest to a good.

In Section IV, data lake technologies are described and explored. From Big data technologies, some components can be reused in order to create a feasible and efficient data lake pipeline, namely distributed queuing, big data stream platforms, big data storage and stream SQL engines. The proposed data lake architecture extends from a shop-floor network to a telecommunication network. The architecture is divided into three domains: a control domain, where data is going to be processed by ingestion systems. Operation and information domain, where data is storage on platforms like Hadoop, that allow machine learning algorithms to be able to train or infer class distributions from the data. Finally, application and business are the last domain, where several services are provided from data visualization to analytic reports.

The lake is going to be populated using data from different sources: machine, processes and product data, control system (product real-time tracking system), environmental and contextual data, ERP, logs events, mouse and keyboard interaction events, IP traffic logs, temperature data, geolocation, electronic tomography, machine vision and audio data. The existence of a single repository for such large and diverse collection of data is a unique opportunity to merge and extract even more value from the lake. Such a value comes not only from each single data source but also from the result of combining different type of source data. Furthermore, from such data multi-model systems can be indeed created, offering a unique perspective of the complex events that happen on the shop-floor.

One of the key expected results of the project is the creation of a new set of services and applications such as analytical services, intelligent support decision service, prediction, maintenance, safety and security applications.

In chapter V the human and machine interaction is considered, namely the process of teaching a robot to perform simple and complex tasks by a visual human demonstration. One of such techniques is LbD, where a control policy needs to be learn by the robot. Currently, there are several different forms of policies, and there is no dominant technique. On the other hand, an approach based on reinforcement learning and imitation learning has been shown effective in addressing the acquisition of skills. In general, a single teacher usually does teaching at each time and there might be conflicting demonstrations across teachers with different styles. Furthermore,

each experiment is independent, i.e. no previous tasks are considered in the learning process. In this project, prior knowledge are going to be reused in order to solve large scale complex task. In doing so, the robot is going to select its own features and previous acquired knowledge but at the same time, redundancy is going to be deleted for a more storage and processing efficiency.

In chapter VI, distributed manufacturing schemes and strategies are discussed. The system aims to manage different type of distributed resources and processes, gather and report analytical results, and create a scalable service oriented for distributed manufacturing. Such a system, should be able to derive the best operational conditions, optimize resources and control, perform decisions on real-time or near real-time and at the same time be aware of the global manufacturing environment. However, a scalable distributed manufacturing system is a complex problem, since the search space increases exponentially with the number of inputs. If the model is too simple, it might be worthless for real cases, but on the other hand, if the model is too complex, it will take a long time to derive a solution and thus not suitable for real- time or near real-time requirements. In this project, an innovative architecture is going to be proposed, which is based on a distributed processing and collaborative intelligent agents. Such architecture is going to be build based on an interoperable web service that will accept several types of inputs, such as customers, orders, factories, transportation costs, human resources, etc. Using the inputs, an intelligent algorithm is going to distribute the workload across different factories and within each factory the operation sequence and assignments, in such way that minimizes or maximizes a customized cost or profit function of the entire global manufacturing process respectively.