INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

MESTRADO EM ENGENHARIA MECÂNICA

DATA SCIENCE FOR INDUSTRY 4.0 AND SUSTAINABILITY: A SURVEY AND ANALYSIS BASED ON OPEN DATA

FILIPE DIOGO SILVA COSTA

julho de 2022

POLITÉCNICO
DO PORTO

DATA SCIENCE FOR INDUSTRY 4.0 AND SUSTAINABILITY: A SURVEY AND ANALYSIS BASED ON OPEN DATA

Filipe Diogo Silva Costa

2022
ISEP – School of Engineering, Polytechnic of Porto
Department of Mechanical Engineering

POLITÉCNICO
DO PORTO

DATA SCIENCE FOR INDUSTRY 4.0 AND SUSTAINABILITY:

A SURVEY AND ANALYSIS BASED ON OPEN DATA

Filipe Diogo Silva Costa
1170371

Dissertation presented to ISEP – School of Engineering to fulfill the requirements necessary to obtain a Master's degree in Mechanical Engineering, carried out under the guidance of Dr. Hélio Cristiano Gomes Alves de Castro and Dr. Paulo António da Silva Ávila.

2022
ISEP – School of Engineering, Polytechnic of Porto
Department of Mechanical Engineering

JURY

President
António Manuel Pereira da Silva Amaral, Dr. Sci.
Adjunct Professor, School of Engineering, Polytechnic of Porto
Supervisor
Hélio Cristiano Gomes Alves de Castro, Dr. Sci.
Adjunct Professor, School of Engineering, Polytechnic of Porto
Second supervisor
Paulo António da Silva Ávila, Dr. Sci.
Coordinator Professor, School of Engineering, Polytechnic of Porto
Examiner
Cátia Filipa Veiga Alves, Dr. Sci.
Invited Assistant Professor, University of Minho

ACKNOWLEDGEMENTS

To my guidance mentor, Dr. Hélio Castro, for the valuable teachings, guidance and availability provided throughout the development of the dissertation.

To Ms. Sci. Tânia Ferreira, for the contribution and support dispended throughout the research.

To my friends and colleagues that helped me grow both academically and personally in this period of my life.

To my family, for the education, values and unconditional support given throughout all my academic journey.

KEYWORDS

Data Science; Industry 4.0; Sustainability; Open Data; Collaboration; Innovation

ABSTRACT

The last few years have been marked by the transition of companies and organizations to more efficient, productive and leaner practices in their processes and systems. In the spectrum of Industry and Engineering, the successful transition to Industry 4.0 is a clear goal for many Small and Medium Enterprises (SMEs) and bigger-sized companies. However, there are economic, social and environmental challenges for this transition that require innovative approaches to overcome them.

The starting point for the development of this dissertation is exploring the importance of Data as a crucial resource and Data-science as a tool for companies, organizations and even public institutions to achieve innovative solutions through collaboration. As it will be further explained, Data is essential in decision making, but in many cases, organizations can't access relevant information and tools because they are either proprietary or because there is a lack of collaboration between them and third parties. There is a common misconception that competition between companies within the same industry prohibits them from collaborating with each other. However, many times data-sharing and collaborative approaches can actually benefit both of them, increase the market they operate in, and accelerate innovation.

Even though the adoption of Industry 4.0 has been already underway, this transition cannot be considered successful unless it improves sustainability across the economic, social and environmental areas of society. Those three sustainable pillars should always be considered a priority in the research of industrial and engineering evolution. Today, more than ever before, information about those topics is widely available but there is still a lack of interest by scientists and scholars in studying some of them. The following research aims to study Industry 4.0 and Sustainability themes through Data Science by incorporating open data and leveraging open-source tools in order to achieve Sustainable Industry 4.0. For that, studying the trends and current state of Industry 4.0, Sustainability and open data in the world, as well as identifying the industries, regions, and enterprises that benefit the most from Industry 4.0 adoption, and understanding if openness of data has a positive impact on Social Sustainability are the main objectives of the study. For that are used methods such as SLR (Sistematic Literature Review) in the bibliographic review and quantitative analysis through open-source software such as *Python* and *R* in the development of the research.

The main results show a positive trend in Industry 4.0 adoption through sustainable practices, mainly on developed countries, and a growing trend of openness of data, which can be positive for transparency in both Industry and Sustainability.

PALAVRAS-CHAVE

Ciência de Dados; Indústria 4.0; Sustentabilidade; Dados Abertos; Colaboração; Inovação

RESUMO

Os últimos anos têm sido marcados pela transição por parte de empresas e organizações para práticas mais eficientes, produtivas e de menores desperdícios nos seus processos e sistemas. No espectro da Indústria e Engenharia, a transição bem sucedida para a Indústria 4.0 é um objetivo claro por várias Pequenas e Médias Empresas (PMEs) e também por empresas maiores. No entanto, existem desafios de cariz económico, social e ambiental para esta transição, que requerem abordagens inovadoras para que os mesmos sejam ultrapassados.

O ponto de partida para o desenvolvimento desta dissertação passou por explorar a importância de Dados como um recurso crucial e da Ciência de Dados como uma ferramenta para empresas, organizações e até mesmo instituições públicas atingirem soluções inovadoras através de colaboração. Como será explicado ao longo da dissertação, os dados são essenciais em tomadas de decisão, mas em muitos casos, as organizações não conseguem aceder a informação ou ferramentas relevantes porque ou são proprietárias, ou porque existe a falta de colaboração entre elas e terceiros. Existe também o conceito errado de que a competição entre empresas numa dada indústria as proíbe de colaborarem entre si. No entanto, muitas vezes a partilha de informação e abordagens colaborativas podem, na verdade, beneficiar ambas, expandindo o mercado onde operam e acelerando inovação.

Apesar da adoção da Indústria 4.0 estar em progresso, esta transição não pode ser considerada bem sucedida se não melhorar a sustentabilidade nas áreas económicas, sociais e ambientais da sociedade. Esses três pilares da sustentabilidade devem ser considerados uma prioridade no estudo da evolução industrial e da engenharia. Hoje, mais do que nunca, a informação acerca desses tópicos é facilmente acedida, mas continua a existir interesse por parte de cientistas e académicos no estudo de alguns deles. A presente pesquisa tenciona estudar a Indústria 4.0 e temas de Sustentabilidade através de Ciência de Dados, incorporando dados abertos e explorando ferramentas open-source, para contribuir para uma Indústria 4.0 Sustentável. Para tal, estudar a tendência e estado atual da Indústria 4.0, Sustentabilidade e abertura de dados no mundo, assim como identificar as indústrias, regiões e empresas que mais beneficiam desta adoção, e finalmente compreender se uma maior abertura de dados pode ter um impacto positivo na Sustentabilidade Social são os principais objetivos do estudo. Assim, são usados métodos como RSL (Revisão Sistemática da Literatura) na revisão bibliográfica e análise quantitativa através de software *open-source* como o *Python* e *R* nos capítulos de desenvolvimento. Os principais resultados mostram uma tendência positiva na adoção da Indústria 4.0 através de praticas sustentáveis, principalmente em países desenvolvidos, e uma tendência crescente na abertura de dados, que pode ser positiva para uma indústria mais sustentável e transparente.

LIST OF ABBREVIATIONS, UNITS AND SYMBOLS

List of Abbreviations

AI	Artificial Intelligence
ASEAN	Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam
BCG	Boston Consulting Group
BRIC	Brazil, Russia, India and China
CEO	Chief Executive Officer
CO_2	Carbon Dioxide
CNC	Computerized Numerical Control
DDDM	Data-Driven Decision Making
I4.0	Industry 4.0
IoT	Internet of Things
ISO	International Organization for Standardization
G7	United States, Canada, Japan, United Kingdom, France, Germany and Italy
LCA	Life Cycle Assessment
LCI	Life Cycle Inventory
MES	Manufacturing Executing Systems
n.d.	No date
R&D	Research and Development
SME	Small and Medium Enterprise
SLR	Systematic Literature Review
ISEP	Instituto Superior de Engenharia do Porto
ISO	International Organization for Standardization
UAE	United Arab Emirates
UK	United Kingdom
US	United States
USA	United States of America
WoS	Web of Science
WEF	World Economic Forum

List of Units

C	Celsius
cm	Centimeters
GB	Gigabyte
kB	Kilobyte
Kg	Kilogram
m	Metre

MB	Megabyte
mm	Millimetre
Pa	Pascal
TB	Terabyte
W	Watt

List of Symbols

$	Dollar
€	Euro
º	Grade
%	Percentage

GLOSSARY OF TERMS

Blockchain	Distributed database that is shared among the nodes of a computer network, storing information in a secure and decentralized way.
Cluster	Group of similar elements that occur together.
Industry 4.0	Term for “Fourth Industrial Revolution”, which represents the current trend of automation and data exchange in manufacturing and industrial processes.
Manufacturing	Process of transforming raw materials into finished good, through a combination of human labor and machining.
Open Design	Movement that involves the development of processes through the use of publicly shared design information, in which the final product is designed by the users.
Operator 4.0	Operator which takes part in Industry 4.0 systems or processes.
Software	Information program associated with the operation of a computer system.

FIGURES INDEX

TABLES INDEX

INDEX

INTRODUCTION

1.1 Contextualization

1.2 Objectives and Thesis Relevance

1.3 Research Methodologies

1.4 Thesis Structure

1 INTRODUCTION

This chapter contains the contextualization of the themes aborded in this dissertation, as well as its relevance and objectives. It is also presented a brief description of the thesis structure. The research methodologies for the Bibliographic review and Research Development are explained in detail in subchapter 2.1. and 4., respectively, so will not be included in the Introduction chapter.

1.1 Contextualization

In the last few years, the manufacturing, scientific and technologic fields have been subject to a revolution process of digitalization and technologic development called Industry 4.0 (Liao et al., 2017). This process is implementing changes that stimulates more competitive practices across many economic sectors. These changes are in great part supported by the growing acquisition and utilization of information and data, that can be exploited through big data technologies and data science.

Even though data is more accessible than ever before, the overwhelming majority of data is concentrated and centralized in private companies, organizations or institutions and inaccessible for scientific and academic research. This means that there is a wide range of limited solutions for economic, social and environmental challenges that can only be solved by those who own the data. The same can be said for the tools necessary to explore that data. Most data science platforms and tools developed in the past are proprietary and costly, which means that they are inaccessible for small businesses, individuals and scientists that can’t pay for the licenses for that software. Another limitation for that proprietary approach is that by being closed source, the development of those tools is limited by the developers of the organization that owns them, limiting the possible opportunities of collaborating with other developers to improve the tool itself.

The growing competition between organizations is a natural consequence that has been driven by technologic advancement. However, that competition can be often counterproductive for themselves, limiting the access of implementations that worked for competitors. Collaboration and cooperation between competitors in the same industry can often be favorable for both, resulting in expansion of markets instead of monopolistic practices.

The implementation of Industry 4.0 has proved to be successful mostly in the economic field of the sustainable framework of human development. However, social and environmental outlook of Industry 4.0 is still somehow in its early stages of development and should be explored in future research.

The bibliographic review of this dissertation explores the state-of-the-art for leveraging data and collaborative approaches in industry and engineering and the sustainable implementation of Industry 4.0 in economic, social and environmental fields.

The Research Development explores collected data from open databases, that is then organized and analyzed resulting in relevant visualizations that allow for taking conclusions relative to the themes identified.

1.2 Objectives and Thesis Relevance

The main objective of bibliographic review of the dissertation is to analyze the state-of-the-art and current available information regarding Industry 4.0, Data-Science, Innovation, Engineering and Sustainability and the relation between them.

The research also intends to identify industrial and engineering challenges and current solutions to those challenges regarding collaborative and open-sourced or proprietary approaches, as well as analyzing the sustainable Industry 4.0 framework.

The objectives for the development section of the dissertation is to analyze relevant themes related to Industry 4.0 and Sustainability through an Open Design Approach supported by principles of collaboration, open data and open source tools.

1.3 Research Methodology

Two methodologies were used throughout this dissertation. The first methodology was used in the bibliographic review and was based on two sequenced methods. The first method was the Systematic Literature Review (SLR) and the second method consisted in the use of VOSviewer software. The combined utilization of both methods is described in detail in subchapter 2.1 of the bibliographic review (Bibliometric Selection and Analysis).

The second methodology was used in the development part of the dissertation, in order to get results that can be analyzed to take conclusions. This methodology was designed within five categories: type, strategy, sampling, data collection methods and analysis techniques. This design was adequate considering the limitations and constraints of the research and resulted in a quantitative analysis of data through open-source software such as Python and R, which is described in detail in chapter 3.

1.4 Thesis Structure

The dissertation is organized in six main chapters.

The first chapter (INTRODUCTION) provides a contextualization for the theme approached highlighting its objectives and relevance and providing a structure for the dissertation.

The second chapter (BIBLIOGRAPHIC REVIEW) elaborates the review and analysis of the literature selected according to the defined source criteria. This chapter is divided into four subchapters. The first one (Bibliometric Selection and Analysis) presents the research methodology and source selection criteria, as well as the bibliometric analysis of six clusters correlated with the defined keywords of the bibliographic review. The second subchapter (The Role of Data in Industry 4.0) explores concepts of data-science and big data in the implementation of Industry 4.0. It also highlights the importance of open-source technologies and collaborative practices among companies and organizations to promote innovation and emergence of better solutions. The third subchapter (Industry and Engineering) explore the concept of "Open Innovation" and how collaborative practices presented in the former subchapter are implemented in industry and engineering. The final subchapter of the bibliographic review (Sustainable Industry 4.0) introduces sustainable Industry 4.0 in the economic, social and environmental areas of human development.

In the third chapter (RESEARCH METHODOLOGY) are represented in detail the research design methodology choices as well as its limitations. It is provided a summary to guide the reader in the end of that chapter.

The fourth chapter (RESEARCH MODEL) introduces all research themes related to Open Data for Industry 4.0, Open Data for Sustainability, and an Open Design approach for sustainable development, providing a consistent justification for why those themes should be analyzed.

Chapter five (RESULTS AND CRITICAL ANALYSIS) presents all results and visualization developed in the research as well as critical analysis for those results.

The sixth chapter (CONCLUSIONS) provides a summary for the conclusions taken in the previous chapter, as well as the contributions made from this research for literature. It also indicates limitations encountered in the research and future lines of investigations for other researchers.

BIBLIOGRAPHIC WORK

2.1 Bibliometric Selection and Analysis

2.2 The Role of Data in Industry 4.0

2.3 Collaborative approaches to Industry and Engineering

2.4 Sustainable Industry 4.0

2.5 Conclusions from the Bibliographic Work

2 BIBLIOGRAPHIC WORK

The bibliographic review chapter consists in the methodology of bibliometric selection and then the bibliometric analysis. The used method to develop knowledge about the study subject was the Systematic Literature Review (SLR) which is a process that enables researchers to answer a clearly formulated question (Xiao & Watson, 2019) by adopting a replicable, scientific and transparent process that differ from traditional narrative reviews (Tranfield et al., 2003).

The three sequential phases of SLR are represented in Figure 1.

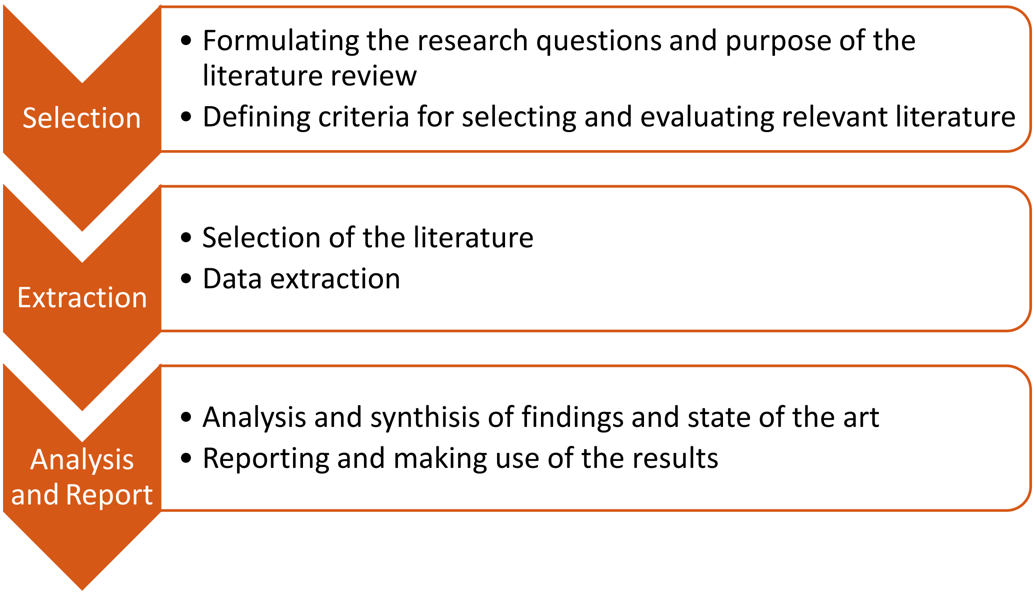

Selection
• Formulating the research questions and purpose of the literature review
• Defining criteria for selecting and evaluating relevant literature
Extraction
• Selection of the literature
• Data extraction
Analysis and Report
• Analysis and synthisis of findings and state of the art
• Reporting and making use of the results

Figure 1 - Research methodology according to systematic literature review adapted from (Denyer et al., 2008)

2.1 Bibliometric Selection and Analysis

It was essential defining the criteria for selection of sources of information since the beginning of the research. The process of database selection as well as the bibliometric analysis and visualization of identified clusters are represented in the following subchapters.

2.1.1 Research Methodology and Source selection criteria

The platform selected to initiate the bibliometric research was Web of Science (WoS) which is a rich database that can provide useful information of the literature written in English such as journals, countries, institutions and authors (Xu et al., 2021). The triage process was made by a sequence of 4 steps:

- Since the research subjects are recent and fast pacing evolving themes, the period defined for the information sources is between year 2000 and October 2021.
- The combination of keywords defined for the research were: “Industry 4.0 + Sustainability + Innovation”; “Data Science + Sustainability + Innovation”; “Industry 4.0 + Engineering + Innovation” and “Data Science + Engineering + Innovation”.
- The sample that resulted from the research criteria contained 862 available publications from a total number of 1897, that were collected and used in the next step.
- The publications that resulted from the previous step were complemented by publications obtained from other databases such as ScienceDirect and b-on. The keywords used for the research in those databases were the same that were used in WoS.

2.1.2 Bibliometric Analysis of identified Clusters

VOSviewer (version 1.6.17) is an open-source software tool that allows the construction and visualization of bibliometric networks (bibliometric mapping) (van Eck & Waltman, 2010). The selected publications according to the defined criteria were imported into the platform, which resulted in a bibliometric map (Figure 2) containing 6 different clusters.

Each cluster represents the interception of relevant scientific themes that are displayed in Table 1, according to the selected publications, resulting in a total of 87 relevant themes in the research.

Between the most cited themes are highlighted themes such “Industry 4.0”, “Innovation”, “Sustainability”, “Big Data analytics”, “Design”, “Machine learning”, “Supply Chain” and “Smart Factory”.

Between the lesser cited themes are highlighted themes such “Open innovation”, “Social Sustainability”, “SMEs”, “Collaboration” and “Sustainable development”, which require a deeper understanding and research in the future.

The bibliographic review of this research incorporates both types of themes converging into a deeper analysis of the lesser cited themes.

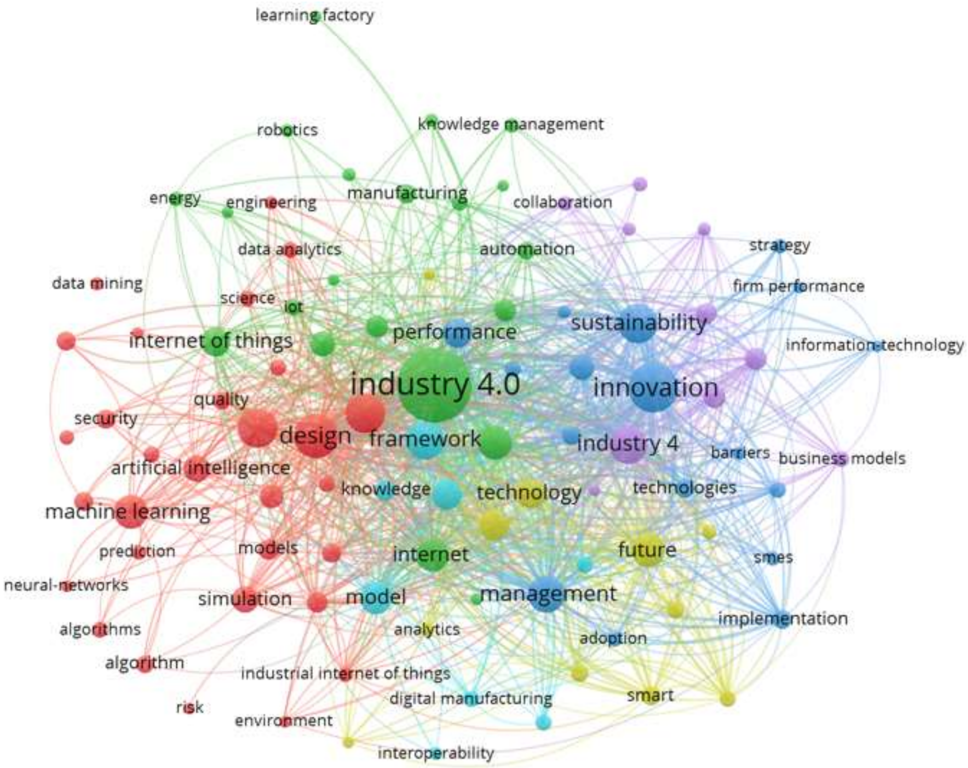


Figure 2 - Bibliometric map representing the most relevant research areas and networks correlating with the defined keywords

Table 1 - Identified Clusters in the bibliometric map

Cluster	Themes
Cluster 1 (27 Themes)	Big data, data science, data mining, design, engineering, environment, industrial internet of things, machine learning, artificial intelligence, algorithms, architecture, deep learning, networks, optimization, quality, privacy, risk, prediction, neural-networks, security, simulation, science, models, information, evolution, cloud computing, blockchain.
Cluster 2 (16 Themes)	Industry 4.0, manufacturing systems, smart manufacturing, maintenance, automation, cyber-physical systems, internet of things, learning factory, manufacturing, platform, robotics, tools, systems, efficiency, energy, digital twin.
Cluster 3 (15 Themes)	Supply-chain management, sustainability, SMEs, technologies, strategy, predictive analysis, logistics, performance, management, innovation, impact, information-technology, big data analytics, adoption, barriers
Cluster 4 (11 Themes)	Decision-making, research agenda, future, opportunities, smart, business model, analytics, challenges, digital transformation, fourth industrial revolution, technology

Cluster 5 (10 Themes)	Open innovation, collaboration, sustainable development, supply chain, product development, of-the-art, digitalization, industry 4, business models, circular economy
Cluster 6 (8 Themes)	Social sustainability, smart factory, integration, framework, interoperability, integration, knowledge, model

For the definition of subjects to develop in the research, the bibliometric networks allowed the identification of most relevant publications (Table 2), journals (Table 3), authors (Table 4) and geographies (Table 5), with the number of citations used as hierarchy factor.

Table 2 - Top 20 most cited publications correlating with the defined keywords.

Reference	Journal	Number of citations
(Boyes et al., 2018)	Computers in Industry	265
(Müller, Buliga, et al., 2018)	Technologic Forecasting and Social Change	260
(Müller, Kiel, et al., 2018)	Sustainability	220
(Mittal et al., 2018)	Journal of Manufacturing Systems	179
(A. C. Pereira & Romero, 2017)	Procedia Manufacturing	175
(Bonilla et al., 2018)	Sustainability	147
(Ghobakhloo, 2020)	Journal of Cleaner Production	144
(Machado et al., 2020)	International Journal of Production Research	137
(Piccarozzi et al., 2018)	Sustainability	130
(Roy et al., 2016)	CIRP Annals	127
(Witkowski, 2017)	Procedia Engineering	127
(Morrar et al., 2017)	Technology Innovation Management Review	96
(Müller & Voigt, 2018)	International Journal of Precision Engineering and Manufacturing-Green Technology	85

(Ghobakhloo & Fathi, 2019)	Journal of Manufacturing Technology Management	76
(Kerin & Pham, 2019)	Journal of Cleaner Production	71
(Bai et al., 2020)	International Journal of Production Economics	62
(Braccini & Margherita, 2018)	Sustainability	57
(Lin et al., 2017)	Sustainability	49
(Maresova et al., 2018)	Economies	42
(Savastano et al., 2019)	Sustainability	32

Table 3. Top 5 journals with the greatest number of cited publications

Journal	Number of citations
Sustainability	944
EPJ Data Science	457
Computers in Industry	430
Journal of Cleaner Production	379
Computers and Industrial Engineering	328

Table 4. Top 5 authors with the greatest number of cited publications

Author	Number of citations
Moat, Helen Susannah	302
Preis, Tobias	302
Boutros, Paul C.	277
Margolim, Adam A.	208
Stuart, Joshua M.	208

Table 5 - Top 5 geographies with the greatest number of cited publications

Geographic Location	Number of citations
England	6899
USA	4301
Italy	1734
Germany	2972
China	2310

2.2 The Role of Data in Industry 4.0

The term Industry 4.0 stands for the fourth industrial revolution, which is defined as a new level of organization and control over the entire value chain of the life cycle of products and it's geared towards the increment of individualized customer requirements (Vaidya et al., 2018). Industry 4.0 uses a series of enabling technologies that can be categorized into nine pillars (Figure 3), which will transform the production of isolated and optimized cells into a fully integrated, automated and optimized production flow (Vaidya et al., 2018).

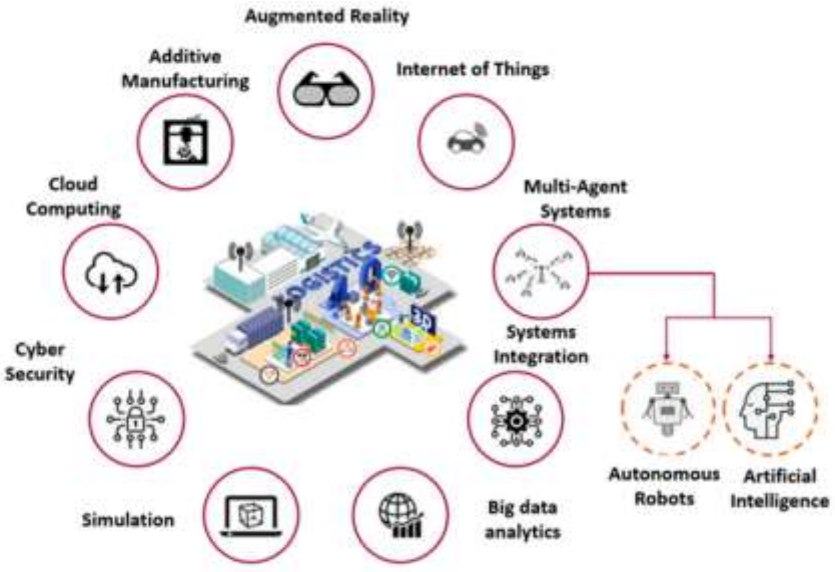


Figure 3 - Main pillars of Industry 4.0 adapted from (Benotsmane et al., 2019)

Industry 4.0 has gained increased adoption in recent years with its promise to use the power of data to revolutionize manufacturing. However, while the exploration of data has been a catalyst of business growth and efficiency gains, the manufacturing sector has been slow to adopt data-driven processes. According to Accenture, only 13% of manufacturing companies have implemented an Industry 4.0 approach (Tim Hall, 2020).

It is inevitable to data to become a cornerstone in decision-making of not only industrial processes, but also to the sustainability of economic, social and environmental approaches. Data-driven decision-making will be essential to the future of those areas and through Data Science and Big Data Analytics it can be implemented faster and more efficiently.

2.2.1 Data-Science and Big Data in data-driven decision making

As we live in a world that constantly produces and consumes data, it is a priority to understand the value that can be extracted from it. Mikalef et al. (2019) consider data science and the big data domains as the next frontier for both practitioners and researchers as they embody significant potentials in exploiting data to sustain competitive advantage.

Big data is the emerging field where innovative technology offers new ways of extracting value from new information. The ability to effectively manage information and extract knowledge is now seen as a key competitive advantage. Big data technology adoption within industrial sectors is an imperative need for most organizations to survive and gain competitive advantage (Cavanillas et al., 2016)

Data science is an interdisciplinary field that supports and guides the extraction of useful patterns from raw data by exploring advanced technologies, algorithms and processes (Provost & Fawcett, 2013a). The actual extraction of knowledge from data is defined as data mining, and it can be applied to a broad set of business areas such as marketing, customer relationship management, supply chain management or product optimization (Bilal et al., 2016).

As is shown in

Figure 4, there are a variety of fields that have a growing influence in decision making that correlate to each other and have the common source of information in data mining. The interception of all these fields can be represented by Data science (Figure 5).

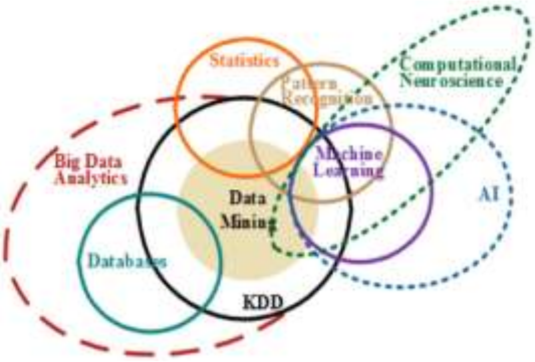


Figure 4 - Interception of data fields with data mining adapted from (Bilal et al., 2016)

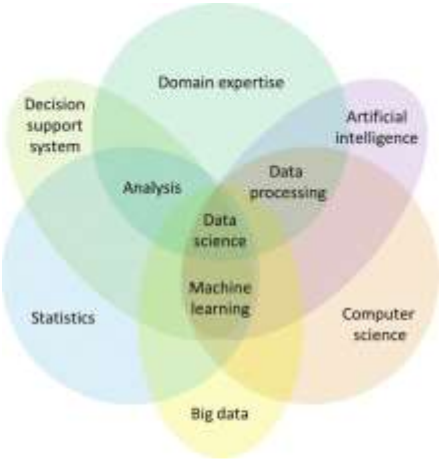


Figure 5 - Interception of data fields with Data Science adapted from (S. Lee et al., 2018)

Even though Data-Science and Big data are closely correlated, Data-Science should be seen as domain that originates from the emergence of big data technologies with data management skills and behavioral disciplines (Saritha et al., 2021).

In a business perspective, the goal in leveraging data-science and big data is usually improving decision making. Data-driven decision making (DDDM) refers to basing decisions on the analysis of data rather than purely on intuition and experience (Wang et al., 2019).

Provost & Fawcett (2013) represent in Figure 6 how the automation of decision making by computer systems in organizations is supported first, by the engineering and processing of data through big data analytics and second by reporting and visualizing that data through data science platforms.

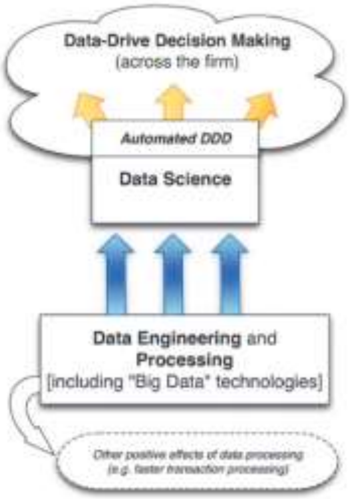

Data-Drive Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and Processing [including "Big Data" technologies]
Other positive effects of data processing (e.g. faster transaction processing)

Figure 6 - Representation how Data Science supports data-driven decision making, adapted from (Provost & Fawcett, 2013)

Brynjolfsson et al. (2011a) conducted a study of how DDDM affects firm performance. That study showed statistically that the more data-driven the firm is, the more productive it is, represented by a 4-6% increase in productivity.

DDDM is also correlated with higher return on assets, return on equity, asset utilization and market value (Provost & Fawcett, 2013).

2.2.2 Data-Sharing and Open-Source

Data-science and big data can be combined with co-creation and data-sharing technologies for organizations to leverage the creativity outside their own organizational boundaries (Runeson et al., 2021). Development and operation of software have become increasingly dependent on data (Gandomi & Haider, 2015) and this data can be more accessible to organizations and individuals through data-sharing and open-source technologies. Runeson (2019) highlights the need for the adoption of co-creation and collaboration principles to harness the innovation potential and to manage costs in the age of data.

Data as a Resource

For organizations, there is a steady increase in reliance on analytics that use enabling technologies such as sensors, the Internet of Things, robotics and ambient computing – all of which rely on huge amounts of data that stem from our many digital interactions (Hickin et al., 2021).

As of 2020, 2.5 quintillion bytes of data were produced every day worldwide (Bulao, 2021) and it is estimated that by 2025 that amount will increase nearly 200 times (Hickin et al., 2021). It is safe to assume that as the gap between the physical and the digital narrows, the data volume of connectivity will continue to grow steadily.

Data Ownership

Today, data volumes are exploding and not only is the rate of data generated per individual increasing, but so is the rate at which we share information. Lawmakers and organizations worldwide are trying to envision data's ownership future. Information remains largely centralized, but the trend is shifting toward a distributed and open model of data sharing (Hickin et al., 2021).

Hickin et al (2021) represent a possible transition from known technologies to future trends in which distributed approaches such open source, explainable AI and decentralized data ownership constitute a positive linear transition (Figure 7). However, if the future approaches to technologic advancements are closed source and proprietary that would mean a negative linear trend.

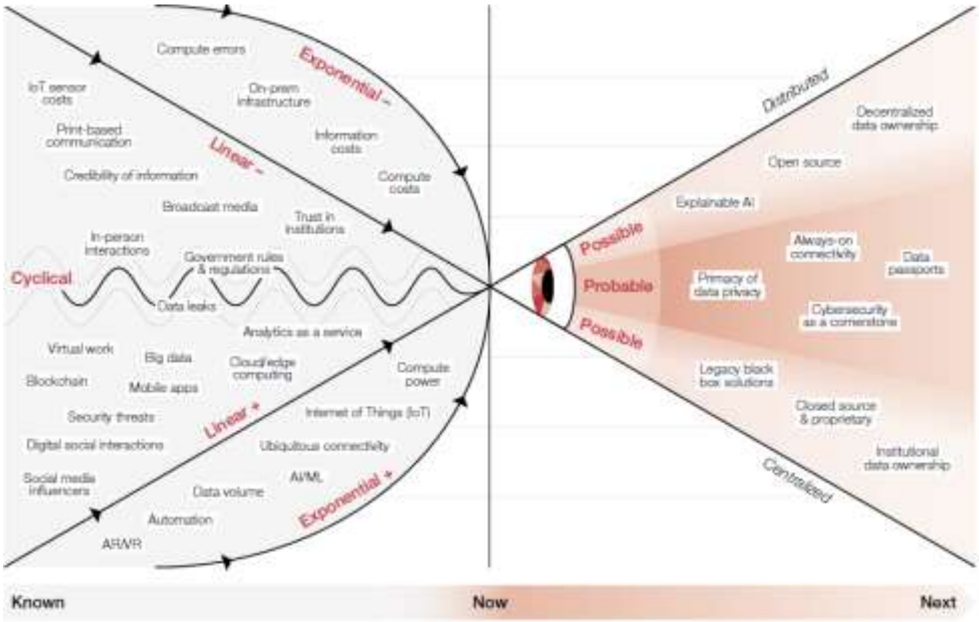


Figure 7 - Past and future trends for future approaches to technologic advancements, adapted from (Hickin et al., 2021)

According to literature, the approach to technologic advancements and future trends of how data and software is collected, stored, managed, modified and shared can be split into Proprietary and Open source (Castro, Putnik, Castro, & Fontana, 2019). Those differences are shown in Table 6 (Bamhdi, 2021; Boulanger, 2005; Caulkins et al., 2013; Hickin et al., 2021; Kilamo et al., 2012).

Table 6 - Approach to technologic advancements and future trends of data ownership

	Closed and Proprietary	**Open Source**
Data Ownership	**Institutional**	**Decentralized**
Approach to Technologic Advancements	Monetization of data by maintaining a closed-source approach that keeps intellectual property private and inaccessible to the end user	Developed and tested through open collaboration
	Software is owned solely by the individual or organization that developed it	Source code can be accessed, modified and redistributed by an open community of developers and programmers
	Limited market of developers and end users, influenced by costs and flexibility	Encourages innovation of SMEs and individual users by accessing useful open-source platforms with no costs
Future Trends	Several governmental organizations have been regulating the protection and privacy of data, giving consumers more control over personal information that businesses collect about them. With growing public awareness and discussion around data privacy and ownership, the future of closed and proprietary approaches to software and emerging technologies are likely to be more and more decentralized.	Recent shifts to open-source models are indicative of the increasingly collaborative nature of technology advancements, and of increased consumer interest in understanding how the technologies we use impact our lives. The major challenges to a wider adoption of open-source platforms are funding and security vulnerabilities but is likely that decentralized technologies and data ownership will play a bigger role in the future.

Considering several options of open-source platforms to use in Data-science projects (Castro, Putnik, Castro, & Bosco Fontana, 2019), Oliphant (2007) considers Python the best choice for scientists and engineers seeking a high-level language for writing scientific applications, since it provides unique features such as:

- An open-source license that permits the user use, sell, or distribute its Python-based applications
- Innumerous libraries modules developed and improved by its community
- Wide number of possible scientific areas in each it can be used
- The language's clean syntax yet powerful constructs
- The possibility to embed Python into existing applications, making the bridge between newer and older applications

Besides its powerful Standard Library, Oliphant (2007) indicates NumPy and SciPy as two useful libraries to use in data science applications. Table 7 represents the possible applications of some different Python libraries for scientific computing and graphical representation, including NumPy and SciPy, and practical studies in which those platforms were used.

Table 7 - Examples of Python libraries for scientific computing and graphical representation

Reference	Platform	Applications	Pratical Implementation
(Dash et al., 2022)	Python Scipy	- Optimization Algorithms - Multidimensional image operations - Solving differential equations and Fourier Transform - Linear Algebra	In this study, Scipy is used to develop graphic visualization of the impact of several defined socio-economic factors for sustainable and smart precision in the agriculture industry.
(Moon et al., 2021)	Python NumPy	- Data analysis - Similar functionalities of MATLAB in conjugation with other Python libraries	In this study, Python-based libraries such as TensorFlow, Keras, NumPy, Pandas, and Matplotlib were used to code the algorithms in the development of neural networks for optimizing chemical engineering processes such as conversion and yield of reactors and CO_2 reduction

(Lemenkova, 2019)	Python Pandas	- Data mining - Data cleaning - Numeric tables and time-series data	In this study Pandas was used in conjugation with Numpy and Scipy to process data to analyze and visualize the potential influence of various geological and tectonic factors in the shape of the Mariana Trench.
(Lou et al., 2013)	Python Matplotlib	- Correlation Analysis of variables - Confidence intervals - Visualizing distribution of data	The authors of this study developed a tool based on Matplotlib for studying and visualizing the Earth's three-dimensional seismic velocity variations through a method called "Seismic travel-time tomography".
(Waskom, 2021)	Python Seaborn	- Built on top of Matplotlib - Used for data visualization and exploratory data analysis - Works easily with data frames and Pandas library	This article provides a comparison of Seaborn to Matplotlib, stating that while the latter is highly flexible and well established, Seaborn offers an interface that permits rapid data exploration and prototyping of visualizations, while retaining much of the flexibility and stability that are necessary to produce publication-quality graphics.
(Abraham et al., 2014)	Python Scikit-Learn	- Clustering - Classification - Regression - Integrates a wide range of state-of-the-art machine learning algorithms	In this study, Scikit-Learn library was used for analysis in brain mapping by accepting objects and algorithms in the form of 2-dimmensional arrays originating from brain scans.

Beyond Python, there are several open-source platforms mentioned in literature that can be explored in industry, engineering and other areas of human development (Table 8).

Table 8 - Examples of open-source tools for data science projects.

Reference	Platform	Applications	Pratical Implementation
(Paradis et al., 2004)	R	- Statistical computation and graphics - Data analytics and visualization	The authors of this study developed a tool in R for reading and writing data and manipulating phylogenetic trees, as well as several advanced methods for phylogenetic and evolutionary analysis.
(Ragan-Kelley et al., 2014)	Jupyter	- Hosting code, data, notes, equations and other information in development environments - Computational notebook	This article describes Jupyter as an interface that can be used in research, education and a platform for hosting notebooks for a research group, supporting programming languages such as Python.
(Frank et al., 2004)	Weka	- Data mining tool	In this study, the Weka machine learning environment allowed the authors to automate data mining (classification, regression, clustering, feature selection) in bioinformatics research.
(Das et al., 2010)	Hadoop	- Storage and processing of big data on a distributed model	The authors built a scalable platform for deep analytics by integrating R statistical analysis systems with the Hadoop data management system.
(Meng et al., 2016)	Spark	- Analytics engine for big data	This article presents the core features of MLlib as a machine leaning library for Spark, adding to the existing libraries Spark SQL, Spark Streaming and Spark GraphX.
(P. Mazanetz et al., 2012)	KNIME	- Data mining - Visual workflows - Machine learning	The authors enumerate several applications of KNIME in drug discovery by combining chemistry and the visual assembly of data. According to the article, commercially data mining software is often prohibitively expensive and this open-source tool is gaining popularity among academia and industrial research.

2.2.3 Collaborative Decision-Making

The adaptation of companies to the exigencies of Industry 4.0 can be explored in other dimension beyond the main pillars described before. Collaborative decision-making, although not so often referred regarding Industry 4.0, is also expressing an increasing visibility and importance in problem solving through leveraging and sharing data (Sousa et al., 2021).

Sousa et al. (2021) split collaborative decision making into two main kinds of models. Mathematical models such as Multi-criteria methods and Game Theory stand out by analyzing different outcomes from a decision-making process. Mathematical models are usually applied in context of collaboratively solving industrial engineering and management challenges.

Between the Mathematical models, Sousa et al. (2021) indicate the most popular approaches (Table 9):

Table 9 - Collaborative Decision-making Mathematical Models

Model	Description of the Model
Analythic Hierarchical Process (AHP)	In this model, several variables or criteria are considered in the selection of only one alternative among the proposals. It allows to analyze, determine and decide the criteria that will influence decision-making by not only determining the best alternative but also justifying the choice in a consistent and coherent way. It can be useful even when two variables are incomparable, thus it can help to recognize which one of the criteria is more important.
The VIKOR method	This model stands for "Multi-criteria Optimization and Commitment Solution". It is helpful in solving decision-making problems with several criteria that are not expressed in the same unit, by focusing on the elaboration of a ranking of the criteria with the most proximity with the ideal solution.
The Shapley's values	Consists of a method of coalition game theory that allows to distribute value between resources.
The DEMATEL model	This model stands for "Decision Making Trial and Evaluation Laboratory" and allows to obtain quantitative relationships between multiple factors necessary to solve a problem by elaborating cause-effect correlations between the criteria and clusters through the creation of networks.
The TOPSIS	This model stands for "Technique for Order Preference by Similarity to Ideal Solution" and is used for evaluating the performance of alternatives through similarity with the ideal

	solution. This solution is the one that maximizes the benefit criteria and minimizes the cost criteria.
Nash Equilibrium	This model is a decision-making theorem within game theory that states that a player can achieve the desired outcome by not deviating from their initial strategy, considering that each player's strategy is optimal when considering the decision of other players.

The second and more recent collaborative decision-making method is an AI-based approach, which helped companies and organizations to solve complex problems through computers that by analyzing large volumes of data, generate intelligent recommendations that support decision making processes (Sousa et al., 2021). Table 10 represents some approaches to AI-based collaborative decision-making and meta-heuristics (Sousa et al., 2021):

Table 10 - Collaborative Decision-making AI-based Models

Model	Description of the Model
Probabilistic relational model (PRM)	This model has demonstrated an important role in the analysis of scientific data, Machine Learning, robotics, cognitive science and artificial intelligence. It provides a framework for understanding the mathematical language for the representation and manipulation of uncertainty. It has emerged as one of the main theoretical and practical approaches for designing machines that learn from the data acquired through experience.
Smart contracts	Smart contracts combine properties of AI and Blockchain allowing computers to learn from accessible data provided by collaborative inputs, which improves data reliability and automates decision-making processes.
Ant Colony Optimization (ACO)	This model is an interactive algorithm inspired by the natural behaviour of ant colonies, that leverage metaheuristics to solve combinatorial optimization problems through collaboration.
The Particle Swarm Optimization	This model is derived from an experimental algorithm that models social behaviour of a set of individuals within a certain population. It argues that the likelihood that a particular individual will make a specific decision will depend on past performance and the performance of certain neighbours.

The approaches to data-science and collaboration described in the subchapter 2.2 can have a key role in Industrial and Engineering future developments and in sustainable practices among economic, social and environmental areas, which are explored in the following subchapters.

2.3 Collaborative approaches to Industry and Engineering

Ever since the beginning of industrialization, technological and engineering advancements have led the shift in industrial practices, which we call “Industrial revolutions” (Lasi et al., 2014). The first industrial revolution was characterized by wide adoption of mechanization, the second by the intensive use of electric energy and the third by the widespread digitalization (Lasi et al., 2014). The fourth industrial revolution is the result of advances in nine fields represented in the subchapter 2.2, being all the result of advances in engineering of cyber-physical systems. It is a natural assumption that engineering innovation is crucial for improvements in efficiency and productivity in industrial production and that data management has an important role in how fast and effective those advancements are (J. Lee et al., 2015).

H. W. Chesbrough (2003) presents a model of open innovation for organizations to accelerate their innovative engineering processes and ways for expanding to new markets.

2.3.1 Open Innovation

Open innovation is defined by Chesbrough et al. (2008) by “the use of purposive inflows and outflows of knowledge to accelerate internal innovation and expand markets for the external use of innovation”. The model incentives companies to use external as well as internal ideas to create value from their technology.

Most of the open innovation model has been studied in large firms (H. Chesbrough & Crowther, 2006) but the recent adoption of the model in SMEs shows that it is worth studying the concept in that kind of firm.

Chabbouh & Boujelbene (2020) argue although SMEs are usually more flexible in decision making, less bureaucratic and take more risk, only a few of them would have sufficient capacity to support and manage the whole innovation process by themselves. Collaboration with other enterprises through data sharing can provide opportunities for innovation and growth that can’t be reached with only internal resources. Figure 8 illustrates a concept in which technologies and markets enablers and supporting tools create a bridge between problem owners and problem solvers, resulting in an open innovation platform that provides solutions as a product. That concept contrasts with the closed innovation model (Figure 9).

The Closed Innovation Model

According to H. W. Chesbrough (2003), in closed innovation, a company generates, develops and commercializes its own ideas. That model encourages leading industrial corporations to dominate and monopolize research and development (R&D) operations and has been the dominant R&D method for most of the 20th century.

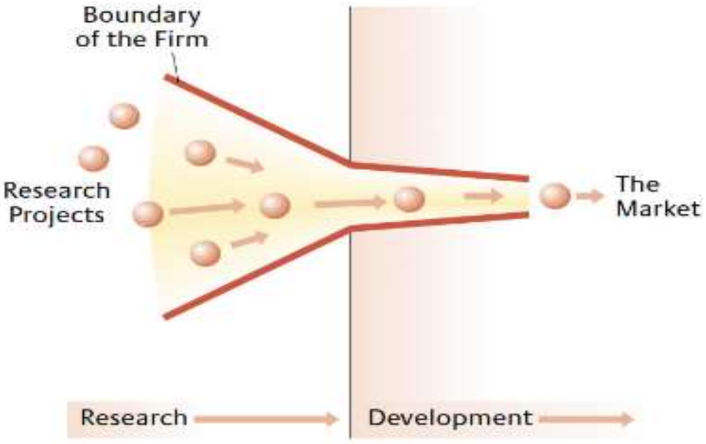


Figure 8 - Visual representation of the Closed Innovation Model, adapted from (H. W. Chesbrough, 2003)

The Open Innovation Model

In the open innovation model, an organization incorporates both its own ideas as well as ideas from other firms, seeking ways to incorporate internal and external innovations into solutions for established or new markets. The model idealizes each organization as a "porous" system, since it can be both the receiver or provider of solutions and innovations for other organizations (H. W. Chesbrough, 2003).

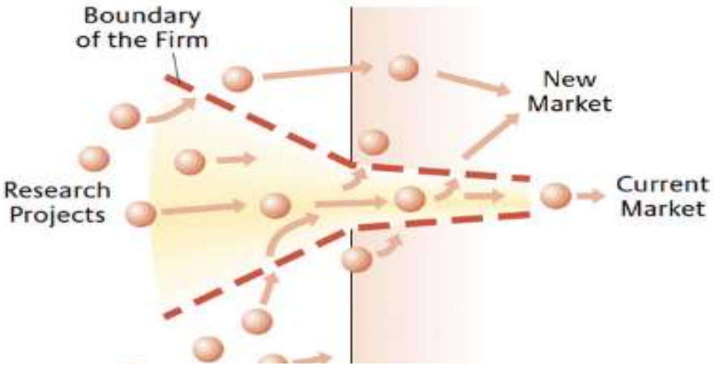


Figure 9 - Visual representation of the Open Innovation Model, adapted from (H. W. Chesbrough, 2003)

Data-sharing is a policy of management of data closely correlated with open-innovation, as it can be an approach for sharing solutions for problems shared across sectors, companies, industries, or regions (Almirall et al., 2014).

Betti et al. (2020) share a BCG survey in 2020 among 996 manufacturing managers, that found that the total value that companies can create in five key areas of data sharing is estimated to be more than $100 billion, focusing on operational improvements alone. On an important note, almost three quarters of them consider sharing data with other manufacturers to improve operations (Figure 10).

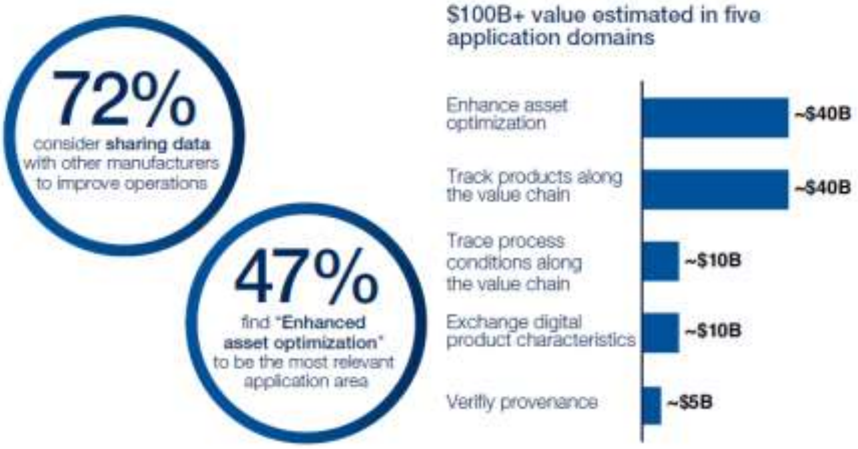


Figure 10. Results of the BCG survey, adapted from (Betti et al., 2020)

Between the five areas highlighted in the survey, almost half of managers find "Asset Optimization" to be the most relevant application area between the five, which improvements represent roughly $40 billion of value.

In the subchapters below, it is explored how SME collaboration through data-sharing and open-source technologies can improve asset optimization and predictive maintenance in industrial environments.

2.3.2 Asset optimization and Predictive Maintenance

In the context of this dissertation, asset optimization addresses how advanced analytics and AI can improve predictive maintenance, by predicting machine failures and improve quality performance.

The main challenge to predict machine failures is collecting data about them. Since unexpected failures are rare, so is the data collected about them and as a result, most manufacturers do not have enough data to build efficient predictive maintenance and SMEs are affected the most (Betti et al., 2020).

To build predictive algorithms it is necessary to collect and combine data from various sources and sensors in the production line, which can also be challenging for many SMEs (Betti et al., 2020). Jain et al. (2020) illustrate this challenge in the prognostics of failures, definition of efficient cutting paths or choosing the most efficient cutting tool in CNC processes. Traditional methods include expensive or complex sensors that require big investments from SMEs.

Open innovation might represent a viable path to solving those challenges through data-sharing and open-source technologies.

Since it might be inevitable sharing data between competitors in each industry, the most trustworthy solution provider would be a third-party supplier or service provider that would gather information from different companies and provide innovative solutions that could work to both. For the open innovation model to work, it is critical to all parties to address issues such as data ownership and security and to share the benefits of the collaboration (Betti et al., 2020).

J. Lee (2003) includes gathering data for predictive maintenance as one of the main objectives of Manufacturing Execution Systems (MES). A platform is needed to serve as a transfer function between the manufacturing data acquisition system and the MES. An open innovation approach for this need could be a platform or provider that could aggregate data shared from failures of different companies in order to build algorithms for predictive maintenance that could work for both companies (Betti et al., 2020).

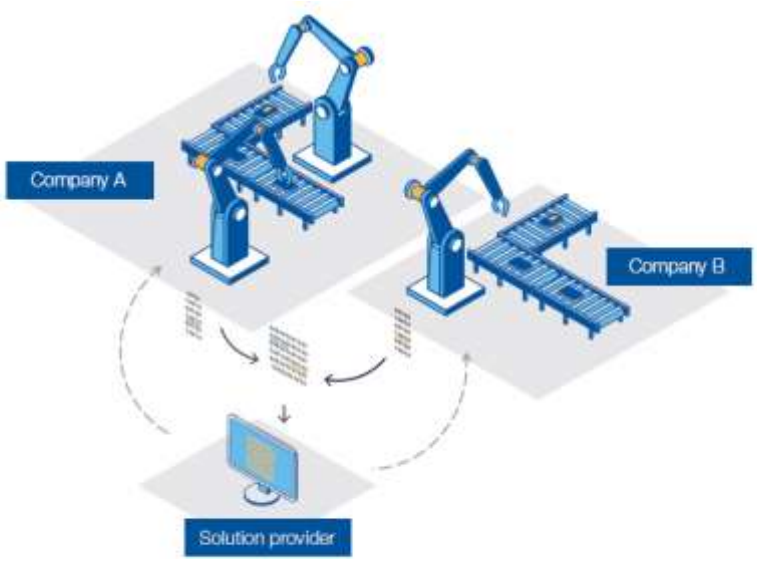


Figure 11 - Visual representation of data-sharing collaboration between different companies, adapted from (Betti et al., 2020)

Below, in Table 11, are described practical implementations of some concepts explored in this subchapter.

Table 11 - Practical implementation of the concepts explored in the previews chapter.

Bibliographic Reference	Pratical Implementations
(Han & Trimi, 2022)	The article explores how the collaboration between SMEs can increase organizational competitiveness and become valuable to its larger partners. With the objective of implementing I4.0 in smaller enterprises and eliminate big data challenges in I4.0 adoption, it is developed a Data science platform for systemizing big data to extract solutions for collaboration between SMEs. The developed database framework was implemented in a Greek SME concluding that with the wide adoption of data based I4.0 models is expected an improvement in collaborative creation of value in the value chain.
(Jain et al., 2020)	In this study it is proposed the utilization of an open-source technology - Auto-WEKA – as a low-cost solution to implement product quality prognostics (prediction of the remain useful life of an asset based on its current condition) in SMEs. It is common for the most used methods being expensive or require specialist workers to do these prognostics. The case study consisted in verifying the suitability and reliability of the software in the prognostics of CNC milling cutters, which is usually made by installing a dynamometer and doing complex force analysis (Beruvides, 2019). The results showed that the proposed open-source solution scored 82% for Suitability, 68% for Reliability and approximately 100% for Quality and Applicability, with a lower cost between 5 and 16 times compared traditional sensor-based solutions.

2.4 Sustainable Industry 4.0

Industry 4.0 technologies have the potential to not only reducing operating costs and increase productivity, but also to reduce production waste, overproduction, energy consumption and facilitating sustainable practices in economic, social and environmental areas (Kamble et al., 2018). Even though the positive economic impacts of Industry 4.0 have been noted with relative quickness, significant improvements in standards of health and safety of the workforce and other social issues are still scarce (Luthra & Mangla, 2018). Advances in digitalization and cyber-physical systems of Industry 4.0 have not successfully solved major environmental challenges either, so this is also an important subject that lacks research and should be further studied (Luthra & Mangla, 2018). Kamble et al. (2018) consider the issue of sustainability has received very little attention in the industry 4.0 literature so more research in this direction is required to evaluate the various solutions of Industry 4.0 for economic, social and environmental challenges.

Kamble et al. (2018) also propose a Sustainable Industry 4.0 framework comprising of three main components guided by principles such as interoperability and decentralization (Figure 12):

- Industry 4.0 technologies – pillars of industry 4.0
- Process integration – human-machine collaboration and equipment integration
- Sustainable outcome – economic, social and environmental sustainability

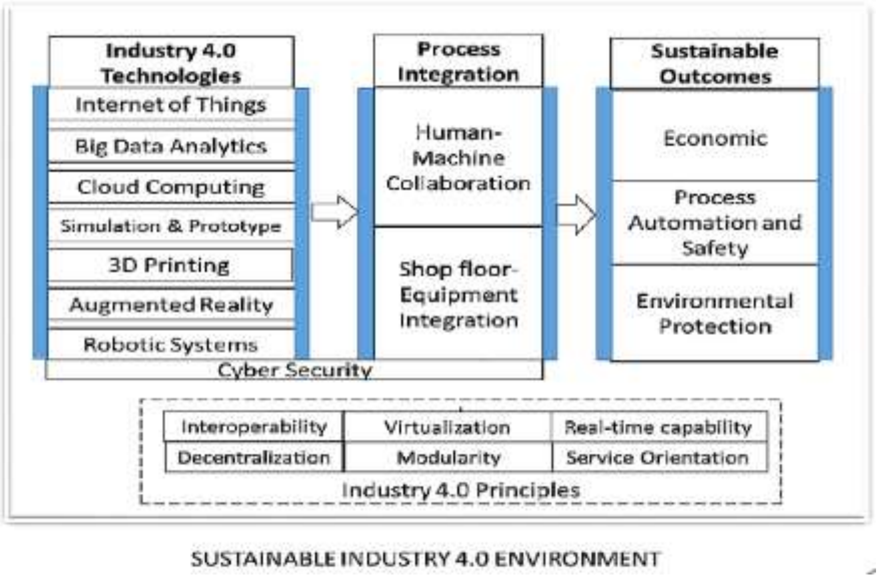


Figure 12 - Sustainable Industry 4.0 framework, adapted from (Kamble et al., 2018)

The following subchapters address each of the three sustainability areas individually.

2.4.1 Economic Sustainability

Available literature supports the idea of Industry 4.0 leading to reduced costs in manufacturing and maintenance, reduce times of production, improve supply-demand forecasting and increase productivity overall, which lead to improved economic performance (Kamble et al., 2018; Ramadan et al., 2017; Schuh et al., 2014).

In the next five years, more than 80% of European companies will digitalize their value chain and increase efficiency by 18% (M. T. Pereira et al., 2019). As SMEs account approximately to 90% of the world enterprises (Inyang, 2013), it is crucial for this type of firms to accelerate innovation and digitalization to stay competitive in a global scale. Pivoto et al. (2021) point that to do so, manufacturing companies need to integrate science capabilities vertically and horizontally across the organization and shift towards data-driven manufacturing. From a quantitative perspective, data-driven organizations have demonstrated 6% higher productivity and efficiency than similar organizations that have not adopted data-driven processes and with further implementation of Industry 4.0 this number is set to increase (Brynjolfsson et al., 2011b). Gökalp et al. (2021) address a study by McKinsey (Bughin, 2018) that expects non-adopters of data science in their processes will experience a 20% decrease in their cashflows by 2030.

Wee et al. (2015) represent the positive impact of Industry 4.0 in increase of productivity and reduction of unproductive times and costs in eight economic value drivers (Figure 13). Those improvements are largely owed to increased accuracy in supply/demand forecasting through big data analytics (Enyoghasi & Badurdeen, 2021).

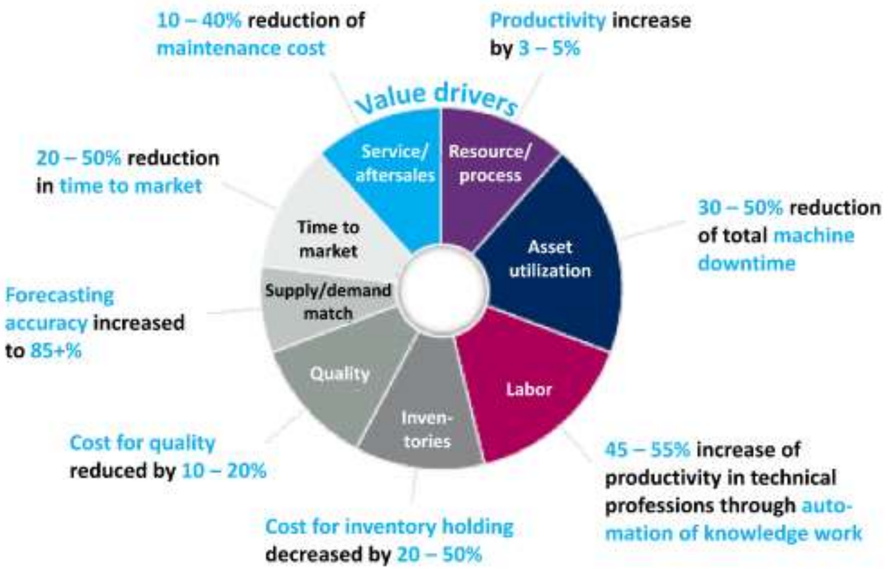

10 – 40% reduction of maintenance cost
Productivity increase by 3 – 5%
Value drivers
Service/ aftersales
Resource/ process
20 – 50% reduction in time to market
Time to market
Asset utilization
30 – 50% reduction of total machine downtime
Supply/demand match
Forecasting accuracy increased to 85+%
Quality
Labor
Inven-tories
Cost for quality reduced by 10 – 20%
45 – 55% increase of productivity in technical professions through auto-mation of knowledge work
Cost for inventory holding decreased by 20 – 50%

Figure 13 - Economic Value drivers of Industry 4.0, adapted from (Wee et al., 2015)

Enyoghasi & Badurdeen (2021) cite some literature in which is discussed the opportunities offered by Industry 4.0 to implement sustainable economic practices in manufacturing and the potential for a Circular Economy. According to this author, circular economy and sustainable manufacturing are often used interchangeably in the literature, referring to eliminating wastes through improved resource utilization. However, Circular Economy differs from sustainable manufacturing since it is a business model focusing in all economic sectors (like food, governmental, transportation, services, manufacturing, etc) and the later focus solely in manufacturing (Enyoghasi & Badurdeen, 2021).

Economic sustainability is a huge focus for companies, governments and institutions in its operation, and also relevant for their social and environmentally sustainable progress (Epstein et al., 2018).

2.4.2 Social Sustainability

Even though Social Sustainability is arguably the most relevant topic for the sustainable development of the human future, it is the one with the scarcest literature and bibliometric available resources. Comparatively with economic and environmental sustainability goals, it appears to exist a lack of interest and research about how scientists, engineers, companies, institutions and communities should approach social issues and challenges in the future. That finding is evidenced by many publications that approach economic and social sustainability as a whole instead of digging deeper into the social side of the equation. For that reason, available literature reveals a profound need for research on social data.

The review of the literature lead to the identification of three main social issues regarding Industry 4.0:

- Automation and the future of work
- Safety of workers
- Human-Machine collaboration

Automation and the future of work

One of the main issues regarding the relation between the adoption of Industry 4.0 and the future of work is job shortages. The increasing digitalization and automation of business and service tasks often lead to worries about permanent replacement of human labor force by machines. However, literature shows that that can be a misconception of the future of work. Shet & Pereira (2021) argue that Industry 4.0 can actually generate job prospects by creating new employment opportunities in emerging domains, like Science, Technology, Engineering and Mathematics. While technologic advancements and automation tend to minimize employment prospects in some sectors, it also brings about the simultaneous emergence of new business and services linked with economic growth and new markets, which leads to the rise of new job opportunities (Shet & Pereira, 2021). However, Shet & Pereira (2021) also warns that those jobs created by digitalization and automation also require a high level of skill, knowledge, competence and specialization that is not required by traditional jobs, leaving unskilled workers more vulnerable to the gradual increase in demand of qualified workforce.

The World Economic Forum conducted a survey in 2020 among a wide number of companies that indicate that 55% of them are looking to transform the composition of their value chain, 43% will introduce further automation and reducing the current workforce. On the other way, the same survey showed that 34% of them will expand their workforce as a result of deeper technologic integration and 41% are looking into expanding their use of contractors for task-specialized work (WEF, 2020).

Below, in Table 12, are represented two studies that evaluate current and future trends on the future of work.

Table 12 - Trends on the future of work.

Bibliographic Reference	Trends on the future of work
(WEF, 2020)	The World Economic Forum predicts the jobs with higher risk of automation are computer operators, secretaries and assistants, typists, machine feeders, telemarketers, among others. The estimated share of workers at risk of unemployment by industry is 47% in accommodation and food services, 15% in wholesale and retail, 15% in transportation, 15% in education, 15% in construction, 15% in manufacturing and 14% in health care. The industries with the lower risk of unemployed workers duo to automation are mining, agriculture and utilities with 4%, 3% and 2% of share, respectively.
(Nagaraj, 2020)	The author identifies the most disruptive technologies that will drive business and industry in the near future, as well as require workforce with skills in those technologies. The list includes data science, artificial intelligence, cloud computing, nano computing, quantum computing, augmented and virtual reality, robotics, machine learning, bioinformatics, among others.

Safety of workers

Literature shows work-related accidents and injuries are one of the main worries of companies regarding human resources. As Industry 4.0 introduces new technologic components and machines to several economic sectors, it is mandatory to guarantee that those technologies provide stability and safe manufacturing environments to the workmen (Kamble et al., 2018). According to Kamble et al. (2018), new Industry 4.0 environments will lead to a revolution in safety management practices provided by data analytics on automated systems, improved equipment maintenance and innovative protection and ergonomics for workmen.

Human-Machine collaboration

(Kamble et al. (2018) cite Qian et al. (2017) in the identification of cooperative control and optimization of production processes through human-cyber-physical interaction as an element for implementing smart and optimal manufacturing processes. The efficacy, acceptance, adaptability and overall performance of human-machine systems and human-system interaction are dependent on how technologies are implemented and how workers program and operate those technologies (Kamble et al., 2018; Quintas et al., 2017).

On a manufacturing level, the integration of those technologies with skilled workforce led to the emergence of Smart Factories.

A Smart Factory is a complex system that integrates the main elements of Industry 4.0 as well as Smart Devices, humans (employees and customers) and Smart Products to make production more competitive, efficient, flexible and sustainable (Benotsmane et al., 2019).
S. Wang et al. (2016) identify some characteristic and advantages of Smart Factories in comparison to traditional Factories (Table 13).

Table 13 - Characteristic and advantages of Smart Factories

Decentralized control and monitorization of production processes.
Humans, smart devices and smart workpieces communicate and collaborate with each other continuously.
More efficient resource utilization, due to self-organizing, self-regulating and self-adapting operations.
Part of the workforce can be reduced and replaced by intelligent devices and a highly skilled workforce is needed for the programming and operation of the intelligent devices.

Benotsmane et al. (2019) consider understanding requirements and the impact of Smart Factories in both Economic and Social Sustainability is key to a better implementation of Industry 4.0. In order to improve quality in processes and delivered products or services, it is necessary to seamlessly integrate sustainable economic, social and environmental processes, which correlate cyclically with each other (Figure 14) (Benotsmane et al., 2019).

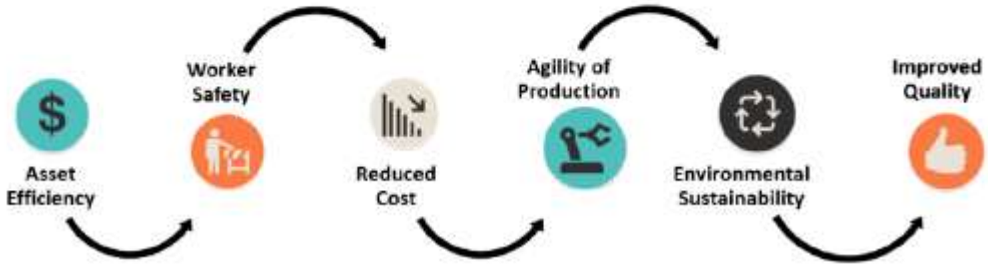


Figure 14 - Positive Economic and Social impacts of Smart Factories, adapted from (Benotsmane et al., 2019)

Below, in Table 14, are described practical implementations of some concepts explored in this subchapter.

Table 14 - Practical implementation of the concepts explored in the previews chapter.

Bibliographic Reference	Pratical Implementations
(Hermann et al., 2015)	Based on literature review, the authors of the article identify six principles to implementing Industry 4.0. Decentralization and Interoperability are two principles that support the concept of exploring open-data technologies and collaboration between enterprises. Workers are described as spectators in correlation with intelligent machines in smart factories, so that human skills can be focused on social decision making.
(Peruzzini et al., 2020)	The study explores the concept and viability of Operator 4.0 as an organizational structure representative of human and social factors in Industry 4.0. The case study consisted in implementing, tracking and monitoring physiologic data such as performance and reaction in order to improve ergonomic conditions and safety between humans and machines.

2.4.3 Environmental Sustainability

In the environmental context, sustainable Industry 4.0 promotes efficient resource allocation like energy, water, raw materials and other products, based on real-time data analysis and other technologies, resulting in sustainable green practices (Kamble et al., 2018; Stock & Seliger, 2016).

Klaus Schwab, Founder and Executive Chairman of the WEF (World Economic Forum), has addressed climate change as “the single greatest threat there has ever been to our planet and livelihoods” and that companies all over the world have a huge opportunity for positive climate impact through decarbonizing supply chains (WEF & BCG, 2021).

Supply-chain Decarbonization

According to WEF & BCG (2021), addressing supply-chain emissions alone enables many companies to impact a volume of emissions several times higher than they could if they were to focus on decarbonizing their operations and power consumption alone. Figure 15 represents the share of carbon emissions by different industries through their own operations (Scope 1), consumed power (Scope 2) and supply chain (Scope 3). Even though the share of emissions of the three scopes are fairly balanced in raw materials industries, in the end products industries the carbon emissions of supply-chain operations is far larger than the sum of the other two scopes combined, accounting to almost 90% of emissions.

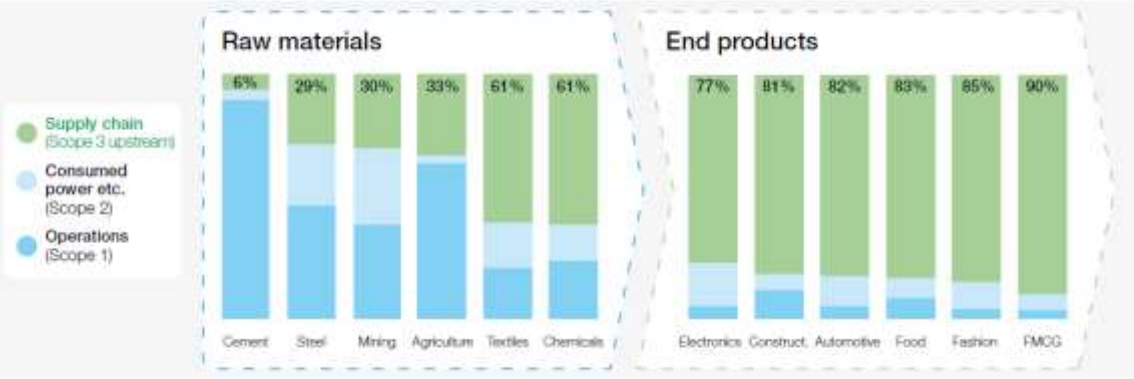


Figure 15 - Share of carbon emissions by industry, adapted from (WEF & BCG, 2021)

Azevedo et al. (2021) propose a conceptual model consisting of advanced technologies to support sustainable and highly efficient supply-chains. This model considers that digitalization and collaboration between businesses will lead to more intelligent decision making with less carbon emissions.

It is important to consider that for a wide adoption of decarbonizing practices in supply chains it is necessary to guarantee sustainable economic solutions both for the companies and the end consumers. According to (WEF & BCG, 2021) around 40% of emissions in supply-chains in several economic sectors could be eliminated with affordable costs (Figure 16) resulting in a marginal impact on end-product costs. Taking in consideration only zero supply chain emissions transition, end consumer costs would go up by 4% at the most in the medium term (WEF & BCG, 2021).

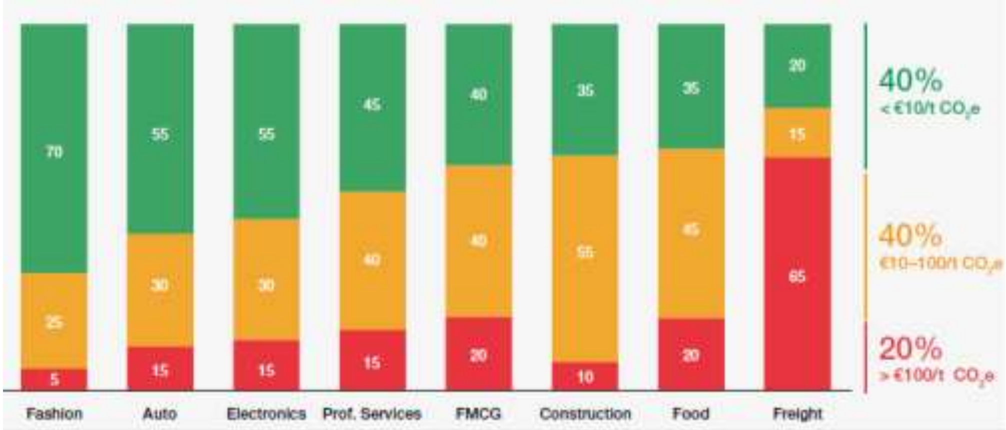


Figure 16 - Share of abatement lever cost by value chain (%), adapted from (WEF & BCG, 2021)

Several studies suggest that when given the option, more than 50% of consumers prefer green products and are willing to pay more compared to the non-sustainable option (Biswas & Roy, 2015; McGoldrick & Freestone, 2008).

Even though decarbonizing supply-chains can bring positive impacts for the environment, companies and consumers, there are still some challenges to bigger efforts toward net-zero supply chains.

Challenges and possible solutions

WEF & BCG (2021) identify three main barriers for a broader supply chain: lack of transparency and data-sharing, financial and engineering challenges and limited support by institutions.

From a data gathering and transparency perspective, open innovation and collaboration models presented throughout this dissertation could present solutions for companies set clear targets and standards that worked for other companies and suppliers.

Chakraborty & Helling (2021) present "Life Cycle Assessment" (LCA) as data-science-based solution for sustainable supply chains that brings environmental insights into decisions, supplementing consideration of cost, performance and social impact. This tool is built on data and models that use materials and energy as inputs and obtains emissions, wastes and products as outputs, which must be known for every step in a product lifecycle.

This model defines four stages of an LCA study (Figure 17) (Chakraborty & Helling, 2021):

1. Goal and scope definition: comparing the potential environmental impacts of two or more choices so that more comprehensive decisions can be made;
2. Life Cycle Inventory (LCI): complete, company-specific, database of every product it makes followed by analysis on the mass and energy flows to and from the nature for a product life cycle;
3. Impact Assessment: calculation of the potential environmental impacts using the LCI results;
4. Lyfe cicle Interpretation: establishes the relation between the previous three phases of the study in order to provide insightful insights to decision makers.

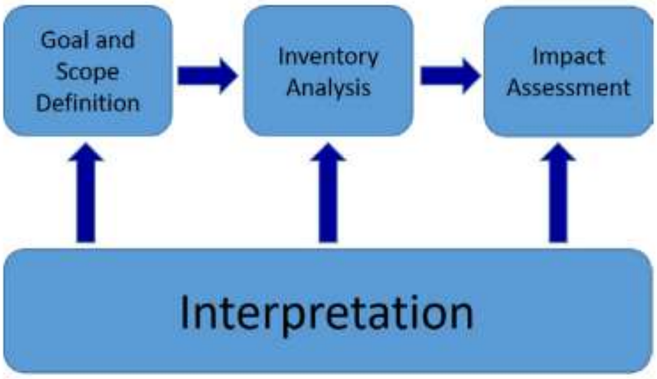


Figure 17 - Four stages of an LCA study, adapted from (Chakraborty & Helling, 2021)

Brenner & Hartl (2021) cite several authors that correlate digitalization of businesses with sustainable cities, circular economies and supply chain management. Digitalization and data analytics will improve resource efficiency and accelerate innovation in logistics and transportation of products, improving economic and environmental sustainability (Brenner & Hartl, 2021).

From a general point of view, literature shows there are still some challenges regarding the adoption of Industry 4.0 and sustainable economic, social and environmental practices. However, there a wide variety of possible solutions and incentives for that implementation, that will lead to newer and greater development possibilities for the humankind.

2.5 Conclusions from the Bibliographic Work

The main conclusions drawn from the bibliographic review are the following:

1. The topics in study that received more attention from scientific and academic research were the broad themes of “Sustainability” and “Industry 4.0”;
2. The Industry 4.0 pillar of Big data analytics and Data science technologies have a growing importance in industry and engineering, and even though they have already shaped many economic sectors, they can be considered to be in their infancy.
3. Proprietary technology is still the main approach to innovation, but open source is a growing method for developing software. Smaller enterprises, scientists and academics are the main beneficiaries of the adoption of collaborative decision-making and data-sharing.
4. Open innovation doesn’t embrace collaboration at its fullest since most companies are eager to share data from their challenges and difficulties but often resist to share their solutions to third parties;
5. Social Sustainability was the pillar that received less attention from researchers, sometimes being combined with economic issues and with most studies focused on ergonomic conditions and safety of workers. There are many publications about Economic and Environmental Sustainability but there is a lot of room for improvement and optimization in those fields.

The main limitations in these topics that were identified in literature are concentrated in open approaches to innovation and social sustainability. As it is explained in subchapter 2.4.2, social issues reveal a profound need for further research on how they can be identified and solved to be in equilibrium with economic and environmental areas.

Understanding these topics might require further lines of investigation and exploration, which require gathering of data and information. This data will allow to assess all relevant factors involved, define what is impeding solutions and hopefully reveal which actions can be implemented.

RESEARCH METHODOLOGY

3 RESEARCH METHODOLOGY

This chapter consists in the definition of the Research Methodology for the dissertation considering the methods gathered in the bibliographic review (chapter 2) and the Research Themes that are detailed in chapter 4. The dissertation aim is to study Industry 4.0 and Sustainability themes through Data Science by incorporating open data and leveraging open-source tools, so the methodology for the development part should be adequate to this objective.

The Methodology chapter purpose is to detail all research design choices that were made and for exploiting that, this chapter is split into two sections. The first section, Research Design, presents and justifies all research design choices. In the second section, Methodology Summary, are referred the limitations for the selected methodology and a final methodology summary to guide the reader in the next chapters of the research. Both sections are structured in Table 15.

Table 15 - Research Methodology

Methodology Section	Contents
Research Design	Research Type
	Research Strategy
	Sampling Strategy
	Data Collection Methods
	Data Analysis Techniques
Methodology Summary	Methodologic Limitations
	Methodology Summary

3.1 Research Design

The Research Design section aims to present the dissertation researcher design to the reader. In this section, all key design choices are detailed and justified logically, according to the dissertation theme.

3.1.1 Research Type

According to J. S. Lee et al. (2011), the approach to studying a research theme can be either inductive or deductive (Table 16). With inductive research, theory is generated from the collected data (from the ground up), allowing for conclusions to be taken after analyzing that data. For that reason, inductive studies tend to be exploratory by nature. On the other hand, deductive research starts with established theories or hypothesis, seeking confirmation in collected data. Thus, these studies tend to be confirmatory in approach.

Table 16 - Methodology research types

Research Type	Approach	Description
Inductive	Exploratory	Data analysis before taking conclusions
Deductive	Confirmatory	Hypothesis formulation before collecting data

The adequate Research Type for this dissertation is the Inductive type since it aims to explore relevant themes and afterwards take adequate conclusions and contributions, instead of pre-establishing hypothesis or theories about those subjects.

Another sensible aspect to the research type is the approach to collecting data: whether the study adopts a qualitative, quantitative or mixed methods methodology (Table 17). Qualitative research focuses on collecting and analyzing textual data or subjective data points such as body language or visual elements, whereas quantitative research uses measurement, visualization and testing through numerical data. Logically, the mixed methods methodology attempts to combine both qualitative and quantitative methodologies to create an integrated perspective (Venable et al., 2016).

Table 17 - Quantitative and qualitative methodology types

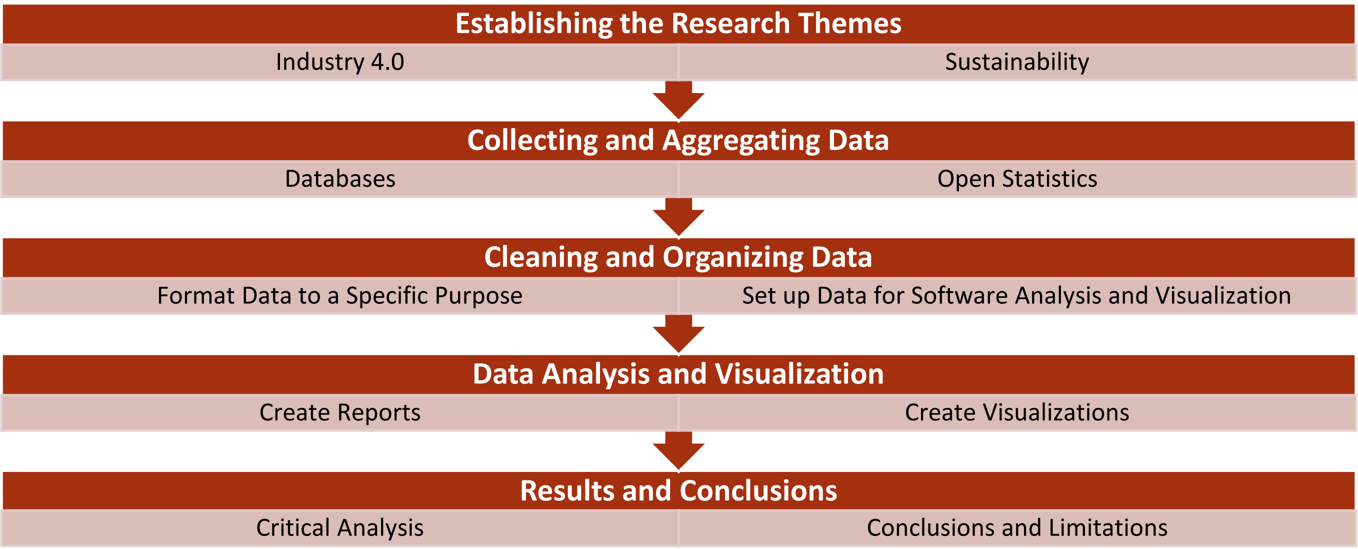
Research Type	Data Type
Quantitative	Numerical Data
Qualitative	Textual and Subjective Data
Mixed Methods	Numerical, Textual and Subjective Data

The preferred research type for this dissertation is Quantitative. Each research theme, weather is related to Industry 4.0, Sustainability or Open Design is supported by quantitative data that is used in an exploratory inductively to then make conclusions through the respective graphs and visualizations.

So, now it is possible to establish that the research type methodology is Inductive and Quantitative.

3.1.2 Research Strategy

This research design category represents the process and line of action for conducting the research based on the aims of the study. This strategy should consider the research type selected (Inductive and Quantitative) in order to establish a sequence of steps from the definition of the research themes to the results and conclusions (Hevner et al., 2004). Below, in Figure 18, is represented the Research Strategy for this dissertation, with indication of the objective for each step.



Establishing the Research Themes
Industry 4.0 | Sustainability

Collecting and Aggregating Data
Databases | Open Statistics

Cleaning and Organizing Data
Format Data to a Specific Purpose | Set up Data for Software Analysis and Visualization

Data Analysis and Visualization
Create Reports | Create Visualizations

Results and Conclusions
Critical Analysis | Conclusions and Limitations

Figure 18 - Research Strategy

3.1.3 Sampling Strategy

The sampling strategy establishes the desired type of sample from which data will be collected from. Even though Peffers et al. (2007) suggest many sample options depending on the research purpose, the more adequate categories of sampling design for this study are either probability sampling or non-probability sampling. Both sampling strategies are summarized below, in Table 18.

Table 18 - Methodology Sampling Strategies

Sampling Strategy	Type of Sample	Charateristics	Approaches
Probability Sampling	Random Group or Population	The results of the study can be generalizable within a population,	Collecting data from General Databases or National Statistics

		country, industry or group	
Non-Probability Sampling	Specific Group	The results are true and unique only for the specific sample analyzed	Interviews and Surveys

In accordance with the research type and strategy and goals of the dissertation, the most adequate Sampling Strategy is Probability Sampling. In the context of the study, results gathered from databases can be generalized within groups such as countries, industries or enterprises size.

3.1.4 Data Collection Methods

The choice of which data collection method to use depends on the dissertation overall research aims and objectives, as well as practicalities and resource constraints (Peffers et al., 2007). These constraints are usually influenced by the type and sampling of the research or the accessibility of available data.

Qualitative research is usually done through collection methods such as interviews, focus groups or participant observations. Quantitative Research, the preferred type for this dissertation, usually relies on surveys, data generated by lab equipment, analytics software or existing datasets (Table 19). Each method has its own advantages, disadvantages and barriers to be used, so selecting the most adequate data collection method is crucial for conducting a correct analysis.

Table 19 - Data Collection Methods

Research Type	**Data Collection Methods**
Qualitative	Interviews
	Focus Groups
	Participant Observations
Quantitative	Surveys
	Datasets
	Lab Equipment
	Analytics and Software

The preferred data collection method for this research is by collecting and analyzing data from existing datasets. Those datasets, however, can only be useful if their content is fully open for being downloaded, modified and published by its providers, which is one of the prevalent characteristics of Open Design. For that reason, for each dataset collected and analyzed, it is assured that there is a license that assures the open accessibility of its data, as well as rights for modeling and publishing eventual results.

To analyze the research themes, it was gathered data for the time period of September 2021 to May 2022 that was compiled into different datasets. Table 20 represents a resume of the datasets used in the Results and Critical Analysis chapter.

Table 20 - Datasets Collected

Dataset	Time Horizon	Source
Gross Domestic Product (GDP)	2015 - 2020	The World Bank (Open Source)
Manufacturing value added as percentage of GDP	1960 - 2020	The World Bank (Open Source)
Manufacturing Share of total Employment	2000 - 2019	The World Bank (Open Source)
Manufacturing Value added from High-Tech	2000 - 2018	The World Bank (Open Source)
Smart City and Smart Factory Index	2020	Data World (Open Source)
R&D expenditure as percentage of GDP	1996 - 2014	The World Bank (Open Source)
Researchers in R&D per million people	1996 - 2017	The World Bank (Open Source)
Small-scale industries % to total industry value added	2004 - 2019	World in Data (Open Source)
Proportion of small-scale industries with a loan or line of credit (%)	2019	The World Bank (Open Source)
Environmental, Social and Governmental Data	1960 - 2020	The World Bank (Open Source)
Sustainability Requirements from Enterprises	2020	World in Data (Open Source)
Skill Migration	2015 - 2019	The World Bank (Open Source)
Workforce Skills Requirements	2015 - 2019	The World Bank (Open Source)
CO_2 emissions per unit value added	1960 - 2018	World in Data (Open Source)

3.1.5 Data Analysis Techniques

The final research design method that needs to be addressed is the data analysis technique. This refers to the methodology for analyzing the collected that in order to extract results and conclusions that are the most adequate for the research.

Selecting the adequate techniques largely depends on the type, sample and data that were previously identified. For quantitative studies, the most frequently used techniques are descriptive statistics and inferential statistics (such correlation and regression analysis) (Hevner et al., 2004).
The most prevalent techniques across the study will be frequency graphs and visualizations for inferential statistics that analyze correlations from selected variables. Another important aspect of the data analysis techniques is to use non-proprietary, open-source tools and software. This comes in harmony with the bibliographic work done about Open Source software (subchapter 2.2.2) and the general theme of the dissertation.

The Open Source software tools used to analyze the data, both referenced in subchapter 2.2.2. of the bibliographic review, are Python and R.

R is a free open-source programming language that provides an analytics computer environment. R provides a variety of statistical and graphical techniques that can be used by importing useful packages. These techniques can be used to handle raw data and retrieve information in order to have a sense on how the data is distributed or patterns that are masked (R Core Team, 2022). The R packages used and its utilities are represented in Table 21.

Table 21 - R packages used

Tool	Packages	Application
R	*arules*	Rule Association
	arulesViz	Rule Association
	RQDA	Quantitative Analysis

Python is currently the fastest growing programming language in the world, thanks to its open accessibility, ease-of-use, fast learning curve and its numerous high quality packages for data science and machine-learning. Together with R, Python provides great utility for identifying correlations between variables and creating powerful visualizations such as graphs, matrixes, plots or maps (Vallat, 2018). The main Python libraries used are shown in Table 22.

Table 22 - Python Libraries used

Tool	Libraries	Application
Python	*Matplotlib*	Data visualization and exploratory data analysis
	Numpy	Correlation Analysis of variables
	Seaborn	Clustering and Visualization

3.2 Methodologic Limitations and Summary

Even with the research design outlined, there are always differences between the ideal design from what is practical a viable to conduct the research. Methodology limitations can vary depending on constraints such as time, budget, sample and other (Peffers et al., 2007).
In the case of this dissertations there are two main limitations to implementing the ideal research methodology (Table 23): the different time horizons between databases and lack of available data about relevant research themes.

Table 23 - Methodologic Limitations

Methodologic Limitation	Justification
Different Time Horizon between Datasets	Since there are approached themes that, even though correlated, are different in provider and objective, the time horizon between different datasets is usually different. That means that it is not possible to establish a constant time horizon throughout all analysis resulting in a variable, yet relevant, time horizon depending on the datasets.
Lack of available data about relevant research themes	As it is referred in the data collection methods subchapter, the preferred method for this research is collecting data from existing open datasets. While this method is very favorable in terms of variety and complementarity of subjects, it is challenging to find data that is open to being downloaded and modeled, regarding all research themes.

Finally, in Table 24, is represented the Methodology Summary to guide the reader throughout the next chapters.

Table 24 - Research Methodology Summary

Research Design	Method
Research Type	Inductive and Quantitative
Research Strategy	1. Establishing the research themes 2. Collecting and Aggregating Data 3. Cleaning and Organizing Data 4. Data Analysis and Visualization 5.Results and Conclusions

Sampling Strategy	Probability Sampling within groups such as regions, countries, industries and enterprise size
Data Collection Methods	Datasets
Data Analysis Techniques	Programming through Open Source software tools such as Python and R

RESEARCH MODEL

4.1 Research Themes

4.2 Conceptual Model

4 RESEARCH MODEL

Subchapter 4 presents the Investigation Model for the development of the dissertation. This model includes the Research Themes (subchapter 4.1.) and the Conceptual Model (subchapter 4.2.) that will form the basis for the remaining study.

4.1 Research Themes

The bibliographic work represents how Data Science and Open Data are being leveraged for collaborative and innovative applications to Industry 4.0 considering the three main pillars of Sustainability. To take a step forward, the research developed in the following chapters aims to reach relevant questions regarding those themes by finding data that can be treated and analyzed to support conclusions. So, the first step is to explore which issues should be addressed and which questions should be asked, considering the information that is, or is not, currently available in previous studies.

4.1.1 Open Data for Industry 4.0

As is described by Tim Hall (2020) in subchapter 2.2., one of the key drivers for the adoption of Industry 4.0 across the globe is the ability to use the power of data to revolutionize manufacturing. However, the manufacturing sector has been slow to benefit from these drivers evenly across different industries, enterprise sizes and geographies. Since most of Industry 4.0 technologies require substantial investments to be successfully implemented, the Economic factor is undeniably crucial for this adoption. Therefore, the differences in economic contexts of enterprises and countries can be immediately associated with the speed and rate of success of Industry 4.0 adoption but it cannot be considered the only one driver for it (Varela et al., 2019). In subchapter 2.3. J. Lee (2003) considers that the quality of the platforms and data that are used by those organizations are possibly the most critical factor for the success of Manufacturing Executions Systems in industry.

Smart Factories and Smart Cities are another relevant study theme as technologic advancements and digitalization are changing how companies operate their business and organizations reshape communities. All those changes and advancements require big R&D investments and qualified researchers and workers.

Since there are many economic challenges and difficulties to recruit the most qualified workers, the adoption of those technologies might be slow unoptimized for SMEs, which need to adapt to technologic changes in order to grow and compete.

To further investigate the Open Data for Industry 4.0 theme, below are specified the most relevant issues that will be addressed.

4.1.1.1 Manufacturing

The World Bank (2022) considers that ideally, industrial output should be measured through regular censuses and surveys of firms. But in most developing countries and smaller enterprises such surveys are infrequent, so survey results must be extrapolated using appropriate indicators. The technologic power of sensors also influences the quality of manufacturing data.

Moreover, much smaller industrial production is organized in unincorporated ventures that are not captured by surveys aimed at the formal sector. Even in large industries, where regular surveys are more likely, monetary and fiscal influences tend to minimize the actual value added of its operations (Varela et al., 2022). Beyond economic factors, digitalization and automation in manufacturing can have resistance from social factors such fear of unemployment from workers or environmental ones, like added carbon emissions from supply-chains. These factors influence not only the enterprises but also countries' government support for manufacturing growth and digitalization. Analyzing the value added by the adoption of technology in manufacturing is also key to understanding the need for Industry 4.0 and its implications (Putnik & Ávila, 2021).

Data Science can leverage Open Data to answer manufacturing questions related to Industry 4.0 such as:

I. What are the global manufacturing drivers?
II. Why is the manufacturing growth diverse around the world?
III. How valuable is manufacturing growth and digitalization for the development of a country?

For this research topic, it was collected open data that is described in depth in the Data Collection Methods subchapter (3.1.4) that include information about gross domestic product (GDP), manufacturing value added as percentage of GDP, manufacturing employment share of total employment and manufacturing value added from high-tech in industry.

4.1.1.2 Smart Cities and Smart Factories

As is referred in subchapter 2.4.2., the integration of Industry 4.0 technologies with skilled workforce is the driver for the emergence of Smart Factories. Since a Smart Factory is a complex system that has the human-machine relationship as its foundation, it is mandatory that these systems are implemented in a sustainable manner, based on concrete data.

The fourth industrial revolution includes technologies that not only change manufacturing and industries but also establishes changes to public dynamics in society. Merritt et al. (2021) mention that there are many policy areas of Industry 4.0 that are present in both Smart Cities and Smart Factories such as equity, inclusivity and social impact, security and resilience, privacy and transparency, openness and sustainability. As new technologies and approaches to industrial operation are evolving, there is also an urgent need for cities to meet policy benchmarks for technology and smart city development. Only by addressing the gaps between Industry and Society approaches to digitalization is possible to be confident that citizens' long-term interests are protected as new technologies are deployed (Merritt et al., 2021).

For this research topic, it was collected open data based on mobility, environment, government, economy and people that result on a Smart City Index further explained in subchapter 3.1.4. The indices utilized to create these insights were developed exclusively from Open Datasets.

4.1.1.3 R&D Efforts for Innovation

Another factor that can be assumed as premise for innovation in Industry 4.0 technologies and Industry 4.0 adoption is Research and Development expenditures of companies and governments. R&D is usually considered to have a high reward profile but with some risks and costs associated (The World Bank, 2022b). Typically, R&D involves big monetary efforts and high skilled researchers which is a barrier to SMEs and developing countries. These researchers are professionals who conduct research and improve or develop concepts, theories, models, techniques and software of operational models (The World Bank, 2022b).

The analysis of this theme considers open data that is further described in subchapter 3.1.4. regarding R&D Expenditures as share of GDP and number of researchers engaged in R&D per million people, across different regions and industries.

4.1.1.4 SME Growth and Adaptability

SMEs are one of the main beneficiaries of the adoption of Industry 4.0 through low-cost methods like Open-Source Software and Open Data for analytics and decision-making. Since the adoption of Industry 4.0 is still so low in this kind of enterprises, understanding which barriers are preventing these companies from growing and how they are overcoming those barriers is relevant to this research. The geographic location of those enterprises is also a key factor in growth and adaptability, since SMEs account for a larger share of total enterprises in developing countries than in developed ones. Formulating policies to help SMEs overcoming growth constraints by leveraging Open Design could support Industry 4.0 adoption and help small businesses and industries add more value to its countries as a share of GDP.

Since the economic factor is probably the most constraining to small sized enterprises, another relevant topic is how easy is for this companies to get lines of credit both from government and private lenders. Getting capital leverage is usually critical for implementing new technologies but it is also risky for SMEs that operate with large amounts of debt if they can't increase revenue and earnings fast enough to pay down that debt in the future. There are also risks for lenders and investors in this kind of enterprises, especially in developing countries, since if the company does not grow according to expectations, investors may loose the capital invested.

For this research topic, it was used a large dataset covering a large number of countries to investigate what determines firm size and if the economic support given to these companies is enough, since understanding these determinants is key in formulating policies to develop the SME sector. This data is described in depth in the Database Collection subchapter (3.1.4) that include information about Small-Scale Industries value added as a share of GDP and Proportion of small-scale industries with a loan or line of credit (%).

4.1.2 Open Data for Sustainability

The bibliographic work done in subchapter Sustainable Industry 4.0 (2.4) reiterated that there is the need for deeper research about Sustainability in Industry 4.0, since it has received very little attention from academics and researchers. In Kamble et al. (2018) framework of sustainability in Industry 4.0, the three sustainable outcomes that should be ideally accomplished from Industry 4.0 Technologies and Process Integration are economic, process automation and safety and environmental protection. Other models include open innovation and collaboration as guiding principles for sustainability in Industry. Social Sustainability is present in themes such as employment and automation, safety of workers, human-machine collaboration and gender equality.

Considering that these themes change in a fast pace and influence each other to some degree, it is crucial to make the research based on available open data to analyze and take concrete conclusions.

4.1.2.1 Collaboration for Sustainable Development Goals

The Sustainable Development Goals are a collection of 17 global goals established by United Nations (UN) in 2015 that are tracked and evaluated till 2030. These goals are integrated across the three dimensions of sustainable development: economic, social and environmental (United Nations, 2015). The Goals and targets intend to stimulate action over the next years in areas of critical importance for society and planet, bringing prosperity, peace and collaboration.

All UN members (countries and stakeholders) pledge to implement actions through collaborative partnerships to accomplish prosperity in fields like industry, innovation and infrastructure, work and economic growth, sustainable cities and communities, gender equality, climate, energy and others.

In this research, analyzing the progress towards accomplishing those goals through open data available is considered an overall evaluation of Sustainability across the three pillars. Since these are broad goals established not only in countries but also organizations and companies, a successful progress towards accomplishing these goals is also positive to accomplish Sustainability in Industry 4.0.

It is important for UN members to collaborate across all established goals. Even more so because the Goal 17 itself – Partnerships for the Goals – focuses in evaluating member's progress towards Economic, Social and Environmental collaboration between them. For that reason, it is reasonable to assume that progressing in Goal 17 is essential to accomplish successful collaboration in the remaining goals. This goal is the focus for this theme – Collaboration for Sustainable Development Goals.

The details for the open data used for this theme is detailed in subchapter 3.1.4.

4.1.2.2 Sustainability Requirements from Enterprises

Even though there are universal sustainable goals established by international organizations, individual enterprises also define their own requirements for sustainable practices and outcomes for their own operations, which are usually unique for the specific context they operate in. Even though those contexts are unique, they can be grouped into common categories called Minimum and Advanced Requirements. Those include:

I. Stakeholder Engagement
II. Assessing impacts beyond the company boundaries and along the supply chain
III. Supplier and consumer engagement on sustainability issues
IV. Procurement and Sourcing practices
V. Environmental performance information in the form of intensity values over time, such as consumption of energy per unit of profit

Of course, the ideal outcome is all those categories have positive developments over time, but it can be unpractical to assess all that information across the large number of enterprises that operate in a given country. However, accessing the number of enterprises that do publish reports about sustainable requirements and the trends on that number over the years can be a leading indicator of their efforts to having sustainable operations.

For the research of this theme, it will be analyzed data about the number of companies publishing reports about advanced sustainability requirements around the world, which is detailed in subchapter 3.1.4.

4.1.2.3 Skill Migration

As it was discussed in the social factor of Sustainability in bibliographic review, one of the main social issues regarding the digitalization and automation of Industry is how employment and skill requirements will be affected. The common sense regarding this issue is that automation eliminates the need for human workers, which will bring unemployment and social unsatisfaction. However, researchers such as Shet & Pereira (2021) actually believe that Industry 4.0 generates new job prospects in emerging domains of Science, Technology and Engineering. Those domains usually require a high level of skill and specialization than traditional jobs that leaves unskilled workers more vulnerable to the gradual increase in demand of qualified workers.

The survey conducted by the World Economic Forum (2020) mentioned in subchapter 2.4.2. also indicates the trend in the skills demanded by companies in the near future since 34% of them are expanding their workforce in technological fields.

This need for qualified workers in the digital age drives the migration of which skills are the most demanded by employers, taught by academic institutions and searched by students and workers.

In this research, studying the skill migration trends of Industry 4.0 and how diverse are these skill demands around the globe is a relevant approach for understanding the social implications of the digitalization and automation that Industry 4.0 brings. While it is presumable that most of developed countries are migrating workers from traditional skills to technological skills, only data can support such premises, since each country is influenced by its own political and economic context.

For that reason, this theme will be supported by Open Data about Skill Migration and Workforce Skill Requirements that is further detailed in subchapter 3.1.4.

4.1.2.4 Carbon Emissions

Environmental Sustainability is arguably the sustainability pillar that has been gaining more attention from researchers in recent years, along with economic concerns. Combating climate change and its impacts in society and communities is urgent and necessary. These actions include conserve and sustainably use the oceans, protect and restore terrestrial ecosystems such as forests and ensure the access to renewable energy in the near-term future. In an Industrial perspective, reduce drastically the carbon emissions is probably the most relevant sustainable issue since industrial and economic operations overall account for the majority of carbon emissions for the atmosphere (Bin & Dowlatabadi, 2005).

As was discussed in subchapter 2.4.3., between the three scopes of emissions that are mentioned by WEF & BCG (2021), the supply-chain currently accounts for around 90% of emissions from companies in their operations. It is clear that in order to accomplish Environmental Sustainability it is necessary to analyze and formulate conclusions for reduction of carbon emissions that result from supply chains. For that reason, the carbon emissions theme also considers open data regarding emissions from supply chains that is detailed in subchapter 3.1.4., as well as open data about CO_2 emissions per unit value added in traded goods around the globe.

4.1.3 Open Design for Sustainable Development

The main objective of this section aims to demonstrate how open-source data and technology can be used and correlated with current trends in Industry and Sustainability. The bibliographic work done in the previous subchapters demonstrates that the openness of data and technology is dependent on many different elements that are exploited differently by industry, geography or enterprise size.

For that reason, the first step for starting the quantitative research and study is to identify the main variable of the study, which must be representative of the overall openness of country, industry or enterprise, and relevant to economic, social and environmental sustainability innovation.

By evaluating data published on official National Statistical Offices (NSOs), the Open Data Inventory (ODIN) 2020/21 provides an assessment of the coverage and openness of official statistics in 187 countries, monitors the progress of open data that are relevant to the economic, social, and environmental development of a country. The available statistics are grouped into each one of those three sustainable pillars as is represented in Table 25.

Table 25 - Categories and Indicators for Open Data Scoring System

	Category	Representative Indicators
Economic Sustainability	National Accounts	Production by industry; expenditure by government and households
	Labor Statistics	Employment; unemployment; child labor
	Price Indexes	Consumer price index; Producers price index
	Central Government Finance	Actual revenues; actual expenditures
	Money and Banking	Money supply
	International Trade	Exports and imports
	Balance of Payments	Exports and imports of goods and services; foreign investment
Social Sustainability	Population and Vital Statistics	Population by 5-year age groups; crude birth rate; crude death rate
	Education Facilities	Number of schools and classrooms; teaching staff; annual budget
	Education Outcomes	Enrollment and completion rates; literacy rates and/or competency exam results
	Health Facilities	Core operational statistics of health system
	Health Preventive Care	Immunization rates; incidence and prevalence major communicable diseases
	Reproductive Health	Maternal mortality ratio; infant mortality rate; under-5 mortality rate; fertility rate; contraceptive prevalence rate
	Food and Nutrition	
	Gender Statistics	Specialized studies of the status and condition of women

	Crime and Justice	
	Poverty Statistics	Number and percentage of poor at national poverty line; distribution of income
Environmental Sustainability	Land Use	Land area
	Resource Use	Fishery harvests; forests coverage and deforestation; major mining activities; water supply & use
	Energy Use	Consumption of electricity, coal, oil, and renewables
	Pollution	Emissions of air and water pollutants; CO2 and other GHG; toxic substances
	Built Environment	Access to drinking water; access to sanitation; housing quality

For each one of the 22 categories, there is a preferred disaggregation that should be made by the NSOs that is also scored. This includes disaggregation by sex, age groups and employment by industry in labor statistics for example. Such disaggregations greatly increase the analytical value of the data (ODW, 2021).

Now that the categories are identified, it is necessary to understand the criteria behind the classification of the openness scores. This methodology will be referred as *Scoring System*.

4.1.3.1 Scoring System

There are two main dimensions of each data category that are assessed by ODIN: coverage and openness.
For data coverage, ODIN considers 5 elements that are quantified with one point if the criterion is satisfied, one-half point if the criterion is partly satisfied and zero if the criterion is not satisfied. Table 26 represents the elements of the coverage criteria scoring.

Table 26 - Elements for Coverage criteria scoring

Time Coverage		**Geographic**		**Disaggregation**
Data available in last 5 years (cs_1)	Data available in last 10 years (cs_2)	First admin level (cs_3)	Second admin level (cs_4)	Recommended disaggregations (cs_5)
Complete: 1 Some: 0.5 None: 0		Yes: 1 No: 0		All:1 Some: 0.5 None:0

For each one of the 22 categories, the coverage score is given by:

$$CS = cs_1 + cs_2 + cs_3 + cs_4 + cs_5 = \sum_{1}^{5} cs$$

There are also five elements to the data openness dimension, which are classified the same way as the coverage criteria (Table 26). According to ODIN, these elements are a representation of standards for open data, such as the Open Definition (Open Knowledge Foundation, 2022). These elements are representative of the ability to select, access and share data.

Table 26 - Elements for Openness criteria scoring

Download Format			**Metadata Available**	**Licensing Terms**
Machine Readable (os_1)	Non-proprietary (os_2)	User selection /API or bulk download (os_3)	Metadata available (os_4)	Terms of use (ToU) stated/ CC BY 4.0 (os_5)
Yes: 1 No: 0		User selected: 0.5 API option: plus 0.5	Specific to indicator/data set: 1 Non-specific:0.5 No: 0	ToU: 0.5 CC BY: plus 0.5

For each category, the openness score is given by:

$$OS = os_1 + os_2 + os_3 + os_4 + os_5 = \sum_{1}^{5} os$$

The Category Scores are obtained for each one of the 22 categories by the average of the 10 scores obtained from the elements of the coverage and openness criteria (5 elements each):

$$Category\ Score = \frac{CS + OS}{10} = \frac{\sum_1^5 cs + \sum_1^5 os}{10}$$

Economic Score is the average of its 7 category scores, Social Score is the average of its 10 Category Scores and the Environmental Score is the average of its 5 category scores.

For this dissertation, it was considered an equal weighting between the three sustainability elements, that corresponds to one third of the overall score classification.

The Overall Score, value between 0 and 1 (or 0% and 100%), represents the final score represented by an equal weighting of each sustainable pillar:

$$Overall\ Score = {}^{1}/_{3}\ Economic\ Score + {}^{1}/_{3}\ Social\ Score + {}^{1}/_{3}\ Environmental\ Score$$

The result of the valuation of those categories is a robust dataset of scores of the openness of data. The disaggregation of these indicators by country and industry will be considered throughout the analysis of correlation of the scores with other quantitative data regarding Industry 4.0 and Sustainability.

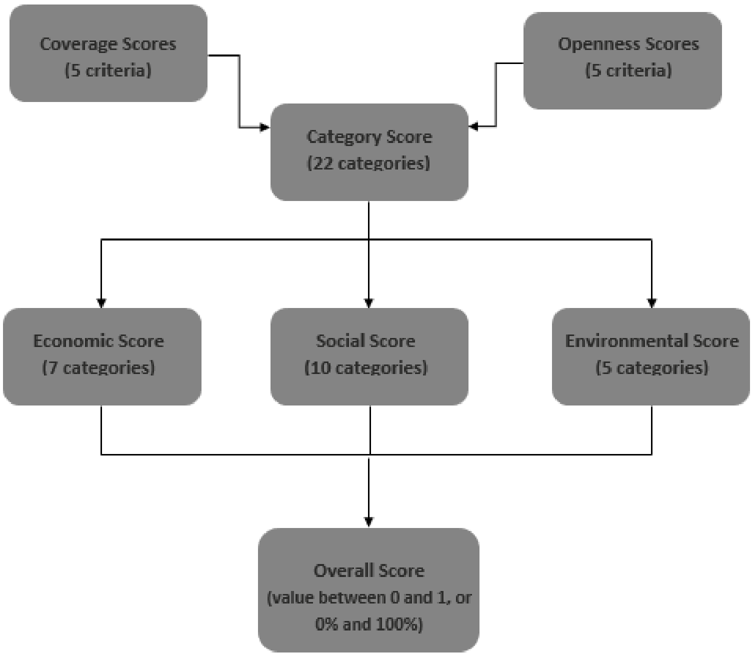


Figure 19 - Schematic representation Openness Scoring System

The nonproprietary dataset with the openness scores is extracted from the *Open Data Inventory* and include the complete amount of data categories as well as data elements, considering each Sustainability Category (Economic, Social and Environmental) to contribute one third of the overall score. The timeframe for the study is from year 2015 to 2020, excluding year 2019 (no data available). The overall score is obtained to every one of the 187 countries which have all the data available.

4.2 Conceptual Model

The conceptual model for this dissertation intends to establish a framework of Data Science for Industry 4.0 and Sustainability moderated by an Open Design Approach that is supported by open concepts found in literature, such as Collaboration, Open Data and Non-Proprietary Tools.

Figure 20 represents the conceptual model of Data Science for Industry 4.0 and Sustainability based on an Open Design Approach.

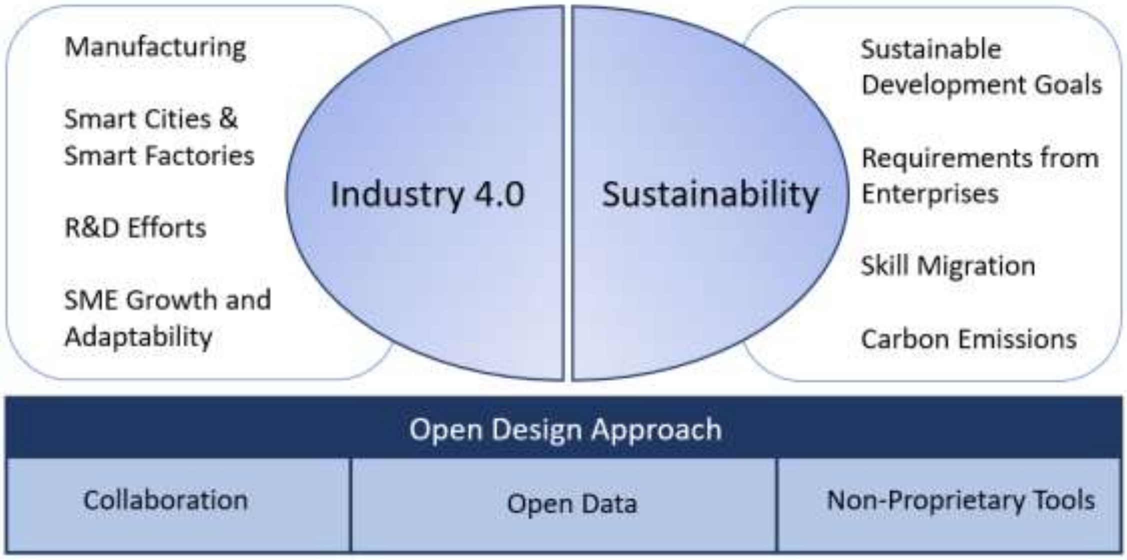


Figure 20 - Conceptual Model for Data Science for Industry 4.0 and Sustainability based on Open Design Approach

Industry 4.0 considers the research themes referenced in subchapter 4.1.1., which are Manufacturing value to GDP, Smart Cities and Smart Factories, R&D Efforts for Innovation and SME growth and adaptability.

Sustainability considers economic, social and environmental themes referenced in subchapter 4.1.2., such as Collaboration for Sustainable Development Goals, Sustainability Requirements from Enterprises, Skill Migration and Carbon Emissions.

Those themes are moderated by an Open Design Approach that is based on three concepts that should be ideally common across the research.

- Availability of Open Data for Decision-Making: based on bibliographic information analyzed in subchapter 2.2.
- Collaboration between organizations, countries and enterprises: based on bibliographic information analyzed in subchapter 2.3.
- Non-Proprietary and Open Source Tools: based on bibliographic information analyzed in subchapter 2.2.2.

The research themes of this model are grouped in each concept in Table 27.

Table 27 - Research Themes

Concept	Themes
Data Science for Industry 4.0	Open Data for Industry 4.0, Open Data for Sustainability
Industry 4.0	Manufacturing value to GDP, Smart Cities and Smart Factories, R&D Efforts for Innovation, SME Growth and Adaptability

Sustainability	Collaboration for Sustainable Development Goals, Sustainability Requirements from Enterprises, Skill Migration and Carbon Emissions
Open Design Approach	Based on Open Data, Collaboration and Non-Proprietary Tools

RESULTS AND CRITICAL ANALYSIS

5.1. Open Data for Industry 4.0

5.2. Open Data for Sustainability

5.3. Open Design for Sustainable Development

5 RESULTS AND CRITICAL ANALYSIS

This chapter presents the results from the data treatment from the selected datasets. For each research theme identified in subchapter 4.1., are represented several relevant visualizations and its respective critical analysis in the context of the Thesis.

5.1 Open Data for Industry 4.0

The first part of results and analysis obtain from the data treatment are representative of the Open Data for Industry 4.0 themes described in subchapter 4.1.1. As it is referred previously, this subchapter approaches Industry 4.0 themes such as Manufacturing, Smart Cities and Smart Factories, R&D efforts for innovation and SME growth and adaptability, which were supported as relevant themes in the bibliographic work.

5.1.1 Manufacturing Value to GDP

Manufacturing is one of the main sectors of Industry around the world and also one of the main adopters of Industry 4.0 (Thames & Schaefer, 2017). Because of that manufacturing has great value to this research, by analyzing available open data and use it alongside with other relevant variables that measure development such as a country's GDP, this research intends to give a brighter perspective on the issues presented in subchapter 4.1.1.1.

To study the manufacturing landscape around the world let's start to quantify the Manufacturing value added to GDP for the two biggest nations in the world by GDP, US and China (Figure 21).

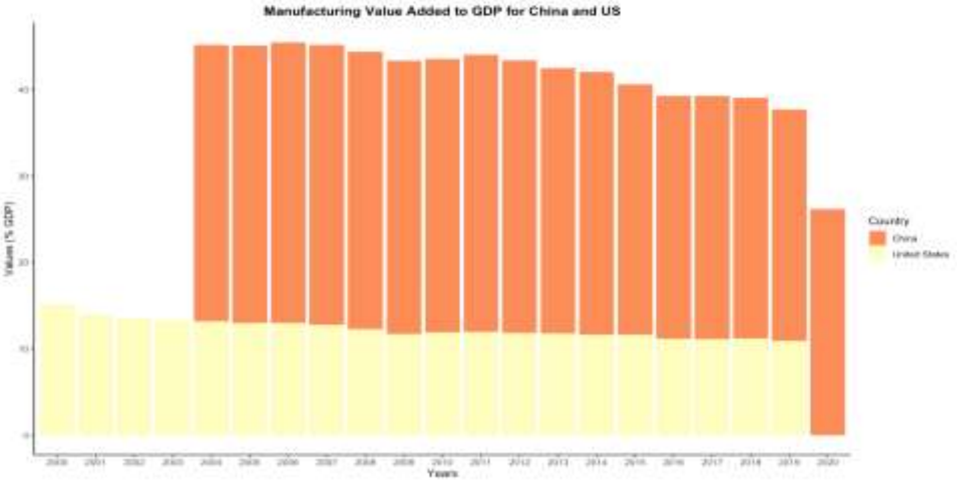


Figure 21 - Manufacturing value added to GDP in US and China

Data is available from 2000 to 2019 for the US and from 2004 to 2020 for China. Manufacturing value added to GDP is much superior for China comparing to the US, averaging for almost 35% of GDP, while in the US averages around 12% of GDP. The trend is descending in both countries, so it is safe to assume that manufacturing has been losing importance for both countries GDP from along the years.

While manufacturing value is so different for those countries, it is interesting to understand how it differs from developed countries to developing countries. For that matter, the following visualizations consider the G7 (United States, Canada, Germany, France, Italy, Japan and United Kingdom) as a sample of developed countries and the BRIC (Brazil, Russia, India and China) as a sample of developing countries. Manufacturing value added to GDP for G7 is represented in Figure 22 and Figure 23, respectively.

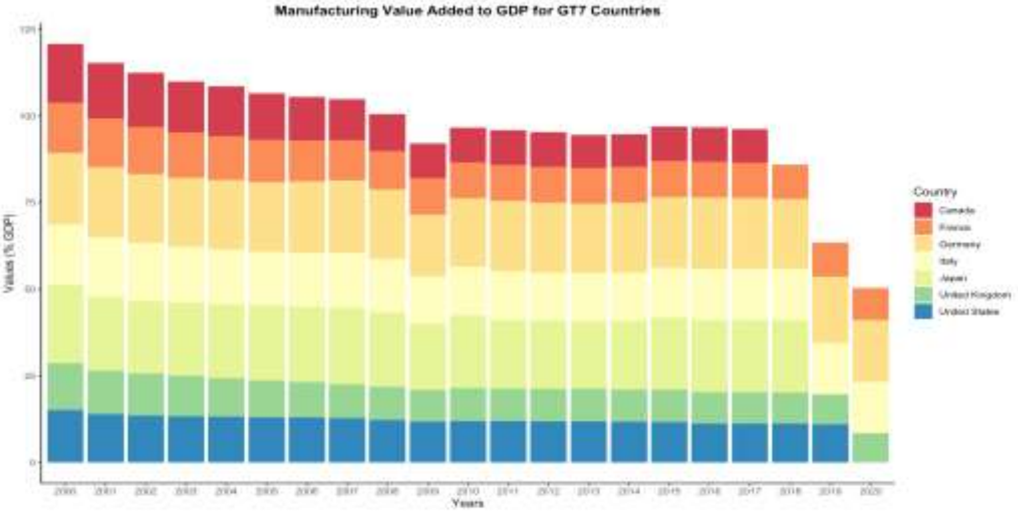


Figure 22 - Manufacturing value added to GDP in the G7 countries

Data is unavailable for Canada in 2018 and 2020, US in 2020 and Japan in 2019 and 2020 for this visualization. The overall trend in manufacturing value added to GDP for G7 is descending, since 2000 marks the year with higher values and 2009 the year with lowest values (considering all countries). Japan and Germany are the countries that have the highest values while the United Kingdom have the lowest. While France and Canada present a clear descending pattern, Italy and the US appear to maintain the same manufacturing value added to GDP throughout the years.

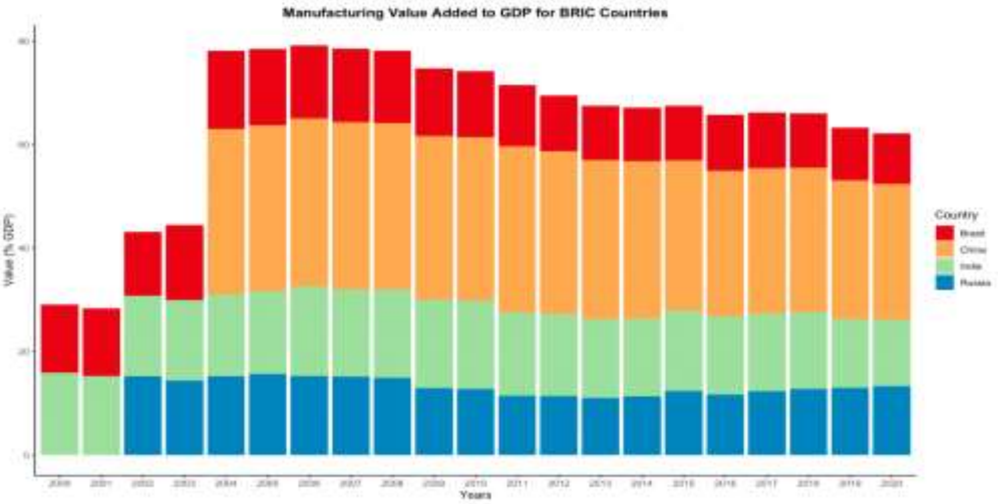


Figure 23 - Manufacturing value added to GDP in the BRIC countries

For BRIC countries, data is fully available from 2004 to 2020. Similarly to the G7 countries, BRIC countries seem to have a descending trend in manufacturing value added to GDP with a few exceptions. Russia has an increase in those values from 2013 to 2020, even though its peak was in 2006. Brazil and India have the steepest decrease in share of manufacturing value in GDP from 2004 to 2020. It is plausible that other technological sectors and services have been increasing in share of GDP for both G7 and BRIC over time.

Finally, for the year 2020 it is represented a geographic visualization of Manufacturing value added to GDP around the world (Figure 24), that was adapted from the UN Conference on Trade and Development databases.

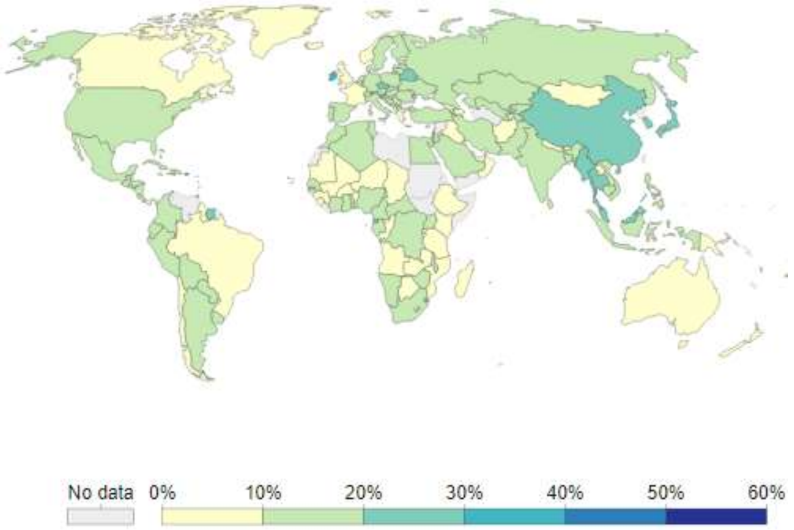


Figure 24 - Global Manufacturing Value added to GDP in 2020, adapted from (UN, 2022a)

By this analysis is clear that China is one of the countries in the world in each a big share of its GDP is allocated in Manufacturing at around 40%. The majority of countries appear to have between 10% and 20% of manufacturing value added to GDP. The continents with larger share of countries that have less than 10% of their GDP value added from manufacturing are Africa and Oceania, while in Europe, North America and South America and Asia there few countries with less than 10% manufacturing value added to GDP. It appears that no country on Earth has more than 50% of its GDP value allocated to manufacturing.

5.1.2 Smart Cities and Smart Factories

A Smart City uses information and technology to improve operational efficiency, share information and provide better quality of life to its citizens and workers (Angelidou, 2014). Implementing Smart technologies and processes within factories and services also intends to promote economic growth, social integrity and environmental sustainability in industrial sectors through Industry 4.0 adoption, creating new jobs in the high-tech and creative industries (Angelidou, 2014).

The dataset evaluates cities across six Smart Categories: Mobility, Environment, Government, Economy, People and Living. The conjunction of those scores translates to the Smart City Index of a city. This database only evaluates cities that score the minimum score in at least one of the six categories.

The software tool used for this analysis is Python, and the used libraries are pandas, Matplotlib and Seaborn.

To begin the analysis of Smart Cities and Smart Factories as an indicator of Industry 4.0 across the world, it is represented in Figure 25 how many cities per country are considered Smart Cities in the database for the year 2020. The frequency of each country will show a first rough estimate of not only if Industry 4.0 is being adopted in that particular country but also how diversified is this adoption across different cities.

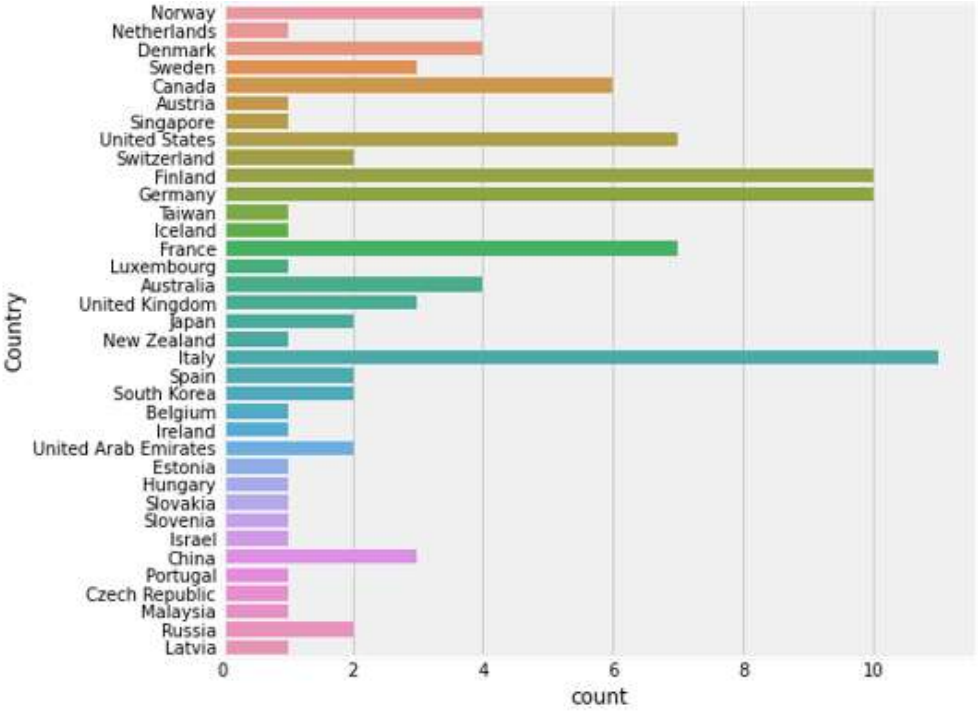


Figure 25 - Number of Smart Cities per country in 2020

At a first glance, in 2020 there are 36 countries with at least one Smart City and a total of 102 Smart Cities across the globe. Of those 36 countries, half of them (18) have only one Smart City, and only 6 countries have more than four Smart Cities.

The country with the highest number of Smart Cities is Italy with 11 cities, followed by Finland and Germany with 10, the United States and France with 7, and Canada with 6 cities.

As it was expected, there is a high prevalence of developed countries in this count, since only 5 of the 36 coutnries are developing countries (International Monetary Fund, 2018) (United Arab Emirates, Hungary, China, Malasya and Russia). In aggregation, the total number of Smart Cities in developing countries is 9, which accounts for only 9% of the total number of Smart Cities in this database.

The only continents without Smart Cities are Africa and South America. While it appears that Europe and North America are dominant in the number of Smart Cities, it is necessary to take a deeper look at how those cities score in comparison with each other.

It is important to take in consideration that high number of Smart Cities in a particular country doesn't necessarily mean that the country scores highly in the Smart City scores. It only means that many cities in that country scored the minimum amount to be considered Smart Cities in the database.

So, to take conclusions about the highest ranked countries in the Smart City Index, analyzing the cities that score highly in the Smart City Index and in each one of the six categories is more appropriate to evaluate smart countries.

Below (Figure 26) are shown the top 10 cities that scored the highest values in the overall Smart City Index.

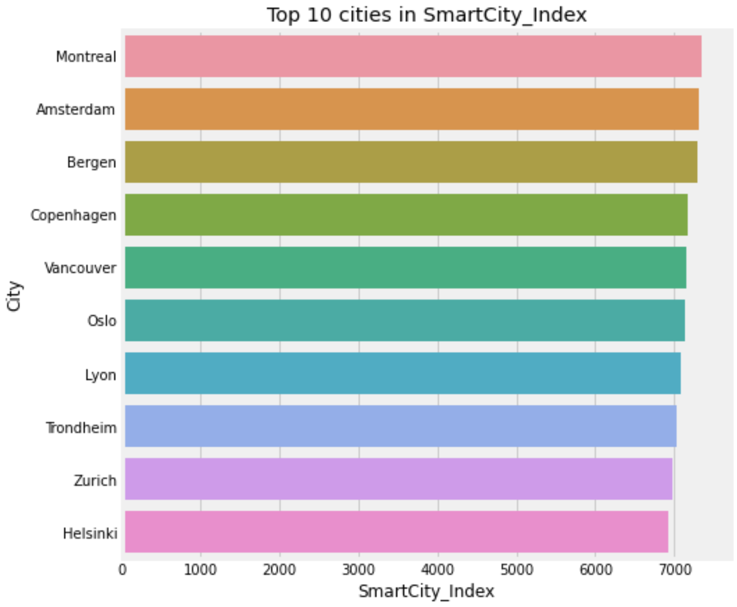


Figure 26 - Top 10 Smart Cities

The city that scored the highest value is Montreal, in Canada. Canada is the only country outside Europe that has at least one city in the top 10 ranking, also placing Vancouver in this ranking. Event thought Europe is clearly dominant in this ranking with 8 cities out of the first 10, the highest scoring city in North American. Since there aren't many conclusions that can be taken by ranking cities that belong to the same continent or region, let's compare how countries score instead of cities.

Again, analyzing each category individually adds a deeper understanding of the previous analysis and provides more data visualization that are useful to make relevant conclusions. The next step in the analysis is visualizing how countries scored in each one of the six Smart Categories (Figure 27).

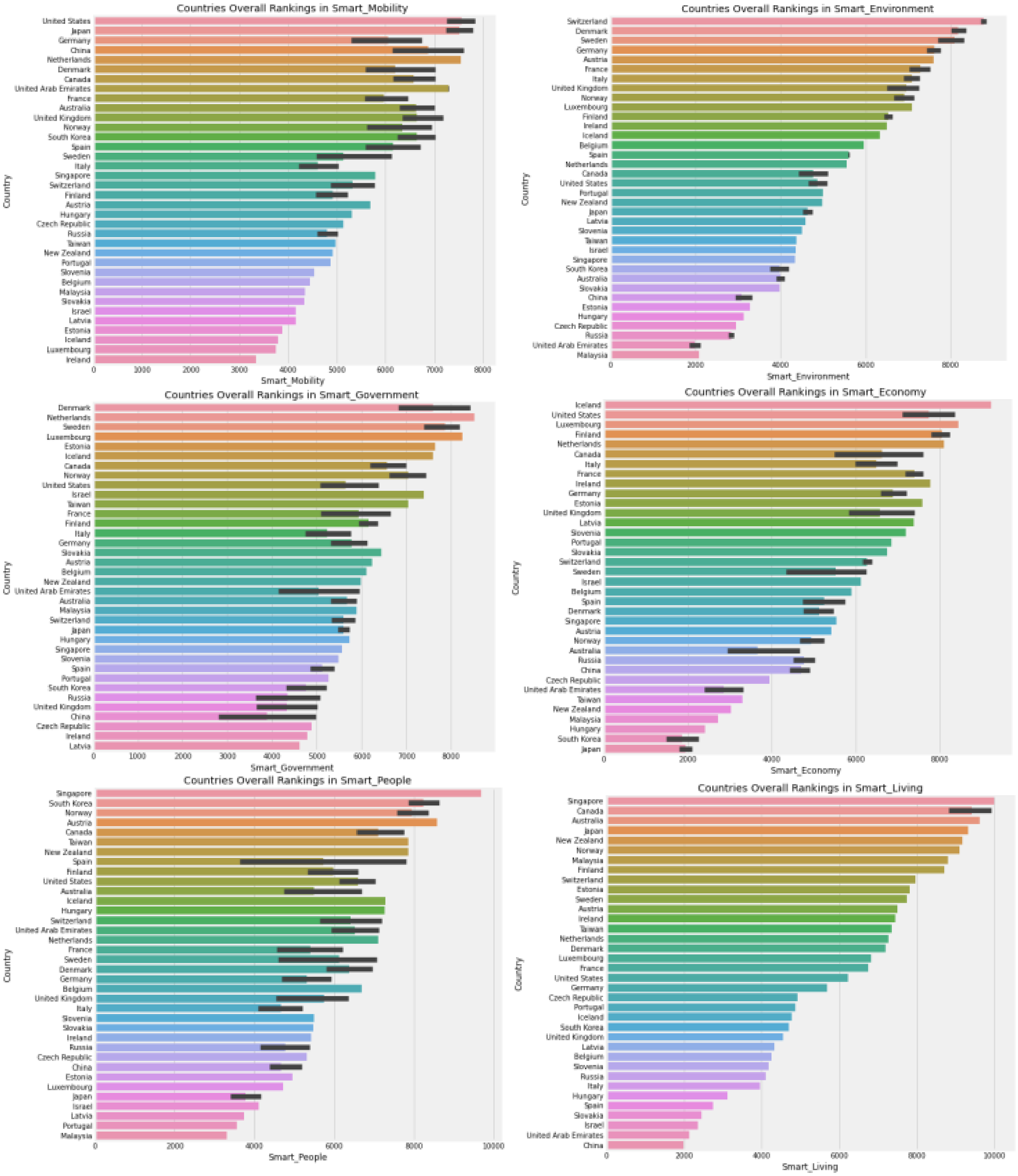


Figure 27 - Smart City scores by category

Each column represents the average score value of all smart cities of the country, for each category. The countries that have more than one smart city in the category also have a grey bar overlapped that represents the amplitude of the lowest scored city and highest scored city for that country.

The visualizations add new information to the analysis since an overall score is not representative of all sectors of a country. While some countries always occupy the top positions in all categories, like Canada, Denmark, Switzerland and Netherlands, and others always occupy the bottom positions, like Hungary, Slovakia and Russia, many countries can both have very high scores in one category and low scores in other categories. Examples of this are Japan with a very high Smart Mobility score and the lowest Smart Economy score, the United Arab Emirates with a high Mobility Score and low Smart Environmental and Smart Living Scores, the United Kingdom, with high Smart Environment score and low Smart Government Score, or China with a high Mobility Score and low Government and Living scores.

The only country with the highest score in more than one category is Singapore (People and Living), and Malaysia is the only country that scores the lowest in more than one category (People and Environment), both Asian countries. The continent with highest scored countries is Europe with Switzerland in Environment, Iceland in Economy and Denmark in Government. However European countries did not score well in Mobility.

The United States score the highest in Smart Mobility and is second in Smart Economy. China seems to perform poorly overall although it has a high Mobility score.

Now that is known the frequency of smart cities in each country, and the highest scored countries in each category, we can analyze the countries' overall rankings (Figure 28).

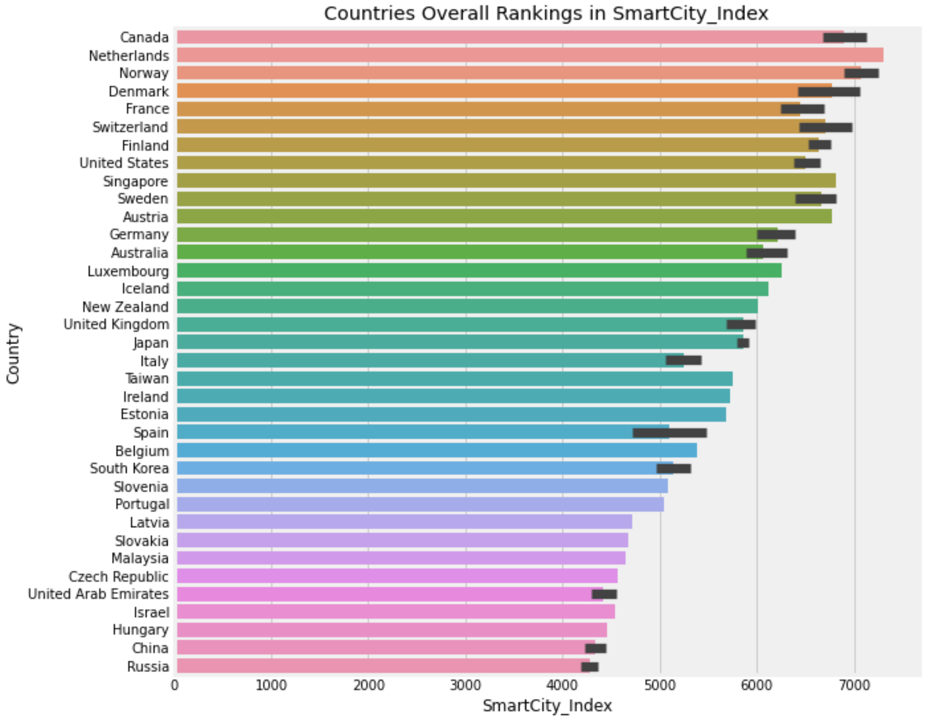


Figure 28 - Top 10 countries by overall Smart City Index

The overall rankings consider the scores of all categories, averaging into a Smart City Index. The countries with the highest overall scores are Canada, Netherlands, Norway, Denmark and France. The lowest scored countries are Russia, China, Hungary, Israel and the United Arab Emirates. Before it was highlighted that a country having a high number of smart cities doesn't necessarily mean that the country itself has a high Smart score, which is confirmed by this ranking. Italy, which has the highest number of smart cities (Figure 25) only ranks 19 in the overall Smart City Index. On the other hand, the top 10 cities with the highest score (Figure 26) correlate almost perfectly with the top 10 countries. Below, in Figure 29, is represented a map with the countries overall scores visualized, which also confirms the dominance of Europe and North America in the Smart City Index.

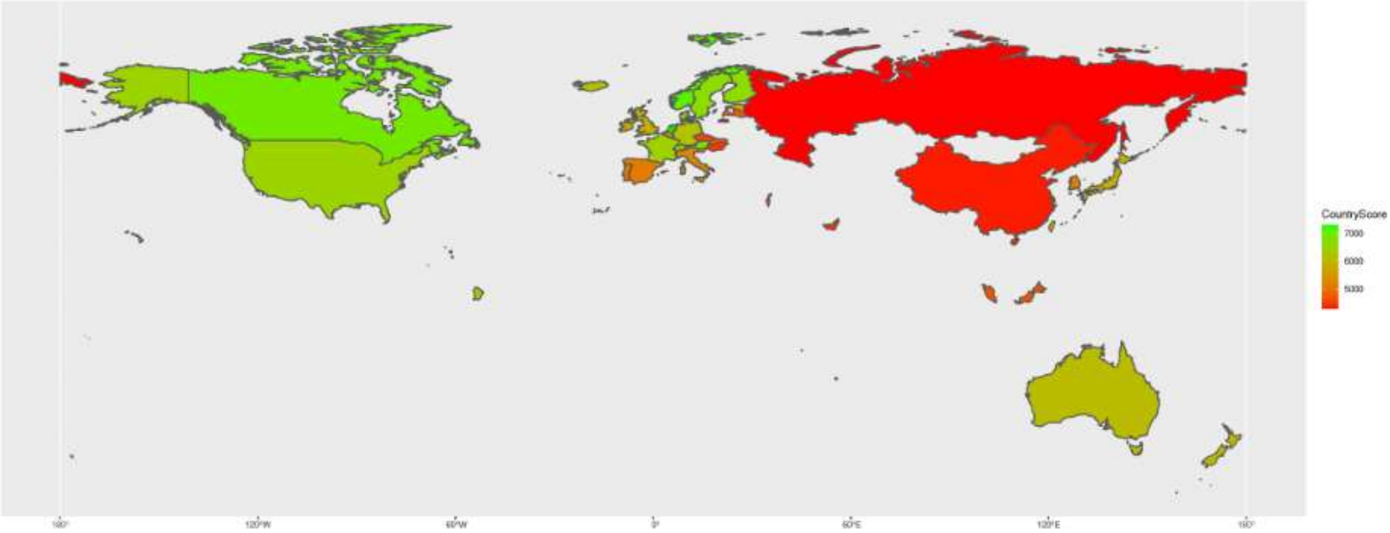


Figure 29 - Geographic representations of Country scores for Smart City Index

The final analysis in this theme is a pair plot that studies the relation between each category with each other. Figure 30 represents a grid of Axes such that all categories are represented in both the y and x-axis where the plots represent the countries' coordinates.

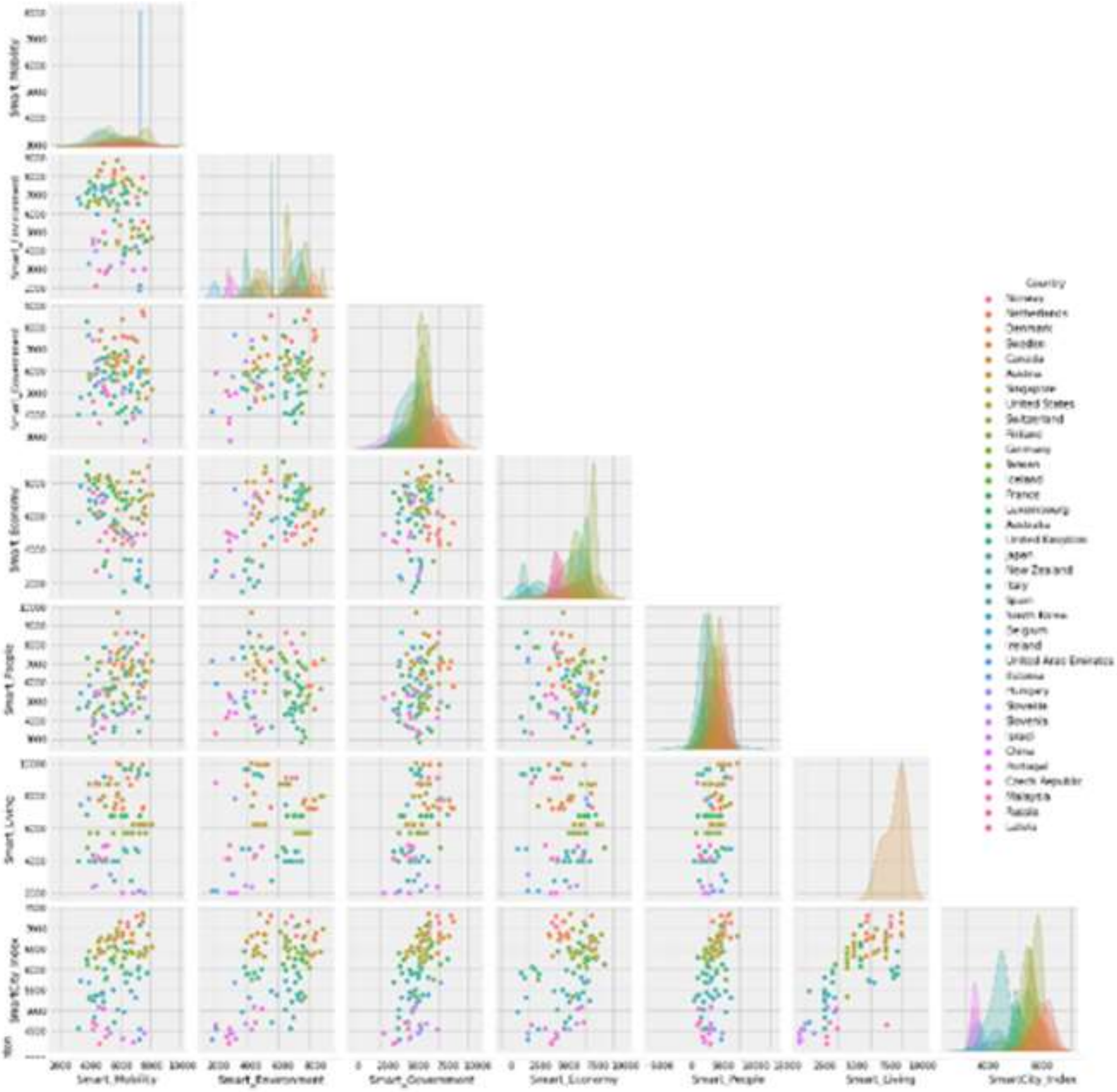


Figure 30 - Correlation plot between all categories and Smart CIty Index

By analyzing each individual category with the overall Smart City Index, it looks like the key factors that seem to correlate more strongly with the overall index are Smart Living and Smart Economy. Since Industry 4.0 is such a big driver for digitalization and automation in the global economy, it makes sense to accelerate the transition to a Smart Economy and Smart way of Living in developed and developing nations that seek develop their cities in technologic and sustainable way.

5.1.3 R&D Efforts for Innovation

One of the main drivers of innovation, particularly in the technologic and industrial fields, is the financing of Research and Development (R&D) by enterprises, academic researchers and scientists (Mansfield & Lee, 1996). However, because of the uncertainty of the level of return and the payback period, this kind of investment is not equally accessible to different countries, industries and size of enterprises. Accessing which countries benefit the most from R&D investments from their enterprises and which industries allocate more expenditures to R&D might be a representation of the efforts to implement Industry 4.0, since this implementation needs a high level of technological advancements, as it is explicated in subchapter 2.2.

The following Figure 31 considers the 2000 companies with the highest percentage of R&D expenditure, as a share of its country GDP, and the company's location.

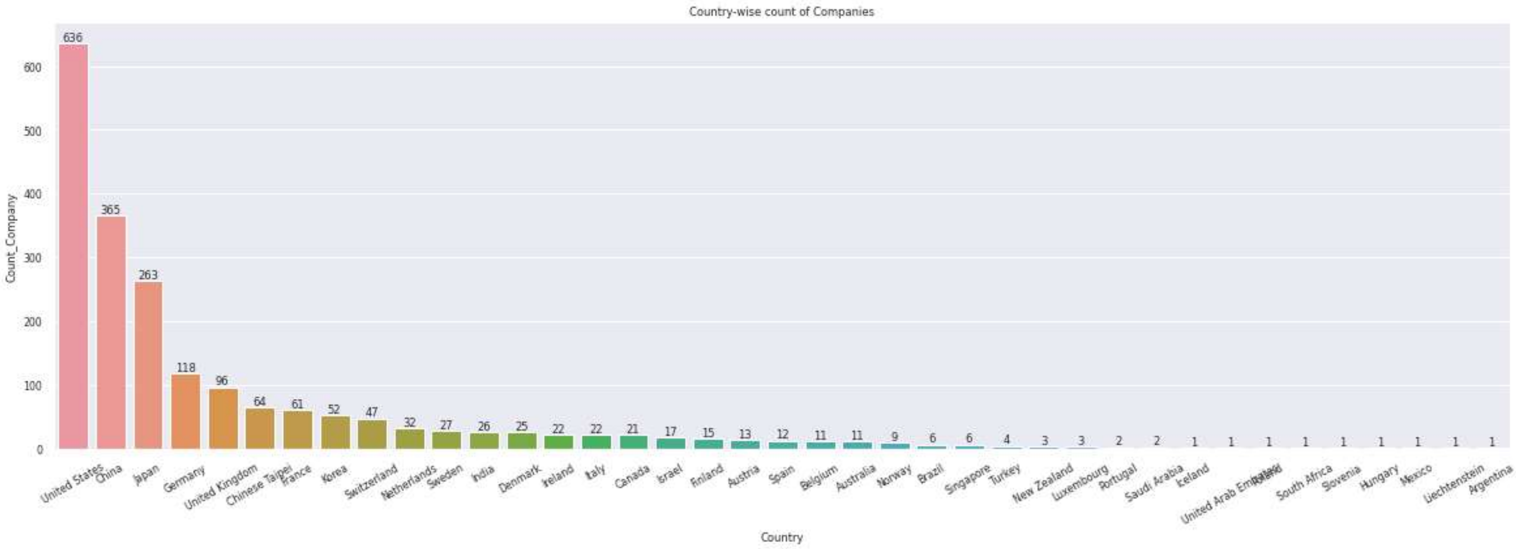


Figure 31 - Number of companies with high R&D expenditure by country, between the 2000 with highest expenditure in the world

We can see that the United States far exceeds the other countries in the number of companies with a high expenditure in R&D as a percentage of GDP. Since currently the US is the country with the highest GDP in the world (The World Bank, 2022a), the difference of investment in R&D is even more pronounced in absolute terms. Of the 2000 companies, 636 of them are American, which is almost a third of the total number and almost double the amount of the second country, which is China with 365 companies. The third country is Japan, with 263 companies, the fourth is Germany with 118 companies and the fifth is the United Kingdom with 64. Together, the top 5 countries with the largest number of R&D companies account for 1215 companies, more than half of the total number and more than the rest 34 countries combined. This discrepancy clearly shows that the economic power of the countries of these companies have a huge influence on how many companies can invest heavily in R&D.

Another aspect of relevance that can be added to the geographic location of these companies is what industry and sector they operate in (Figure 32). It is expected that more technological sectors require more R&D expenditure than traditional ones.

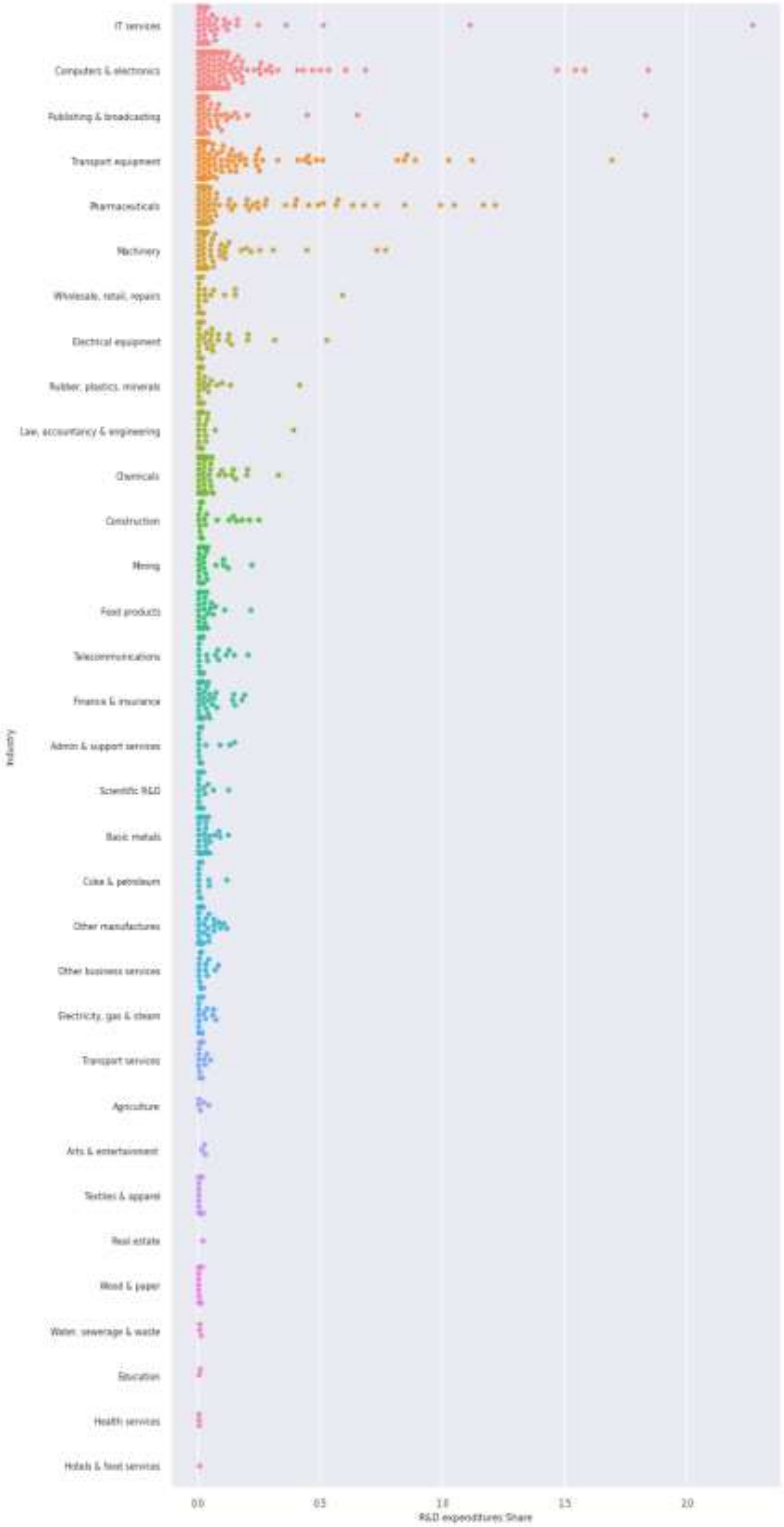


Figure 32 - Share of total R&D expenditure by company, by industry

Most companies, independent of the sector, account for no more than 0.5% of the total R&D expenditure of the 2000 companies combined. However, the main conclusions can be taken from the outliers that account for more than 0.5%, 1%, 1.5% or even 2% of that share. The industries that are represented by those outliers are IT Services, Computers and Electronics, Publishing and Broadcasting, Transport and Equipment and Machinery. Generally speaking, those sectors require a high level of innovation in a fast-changing technological landscape and are representative of the efforts to implements Industry 4.0. It must be noted that this graph represents the companies individually but the sum of the share of all companies in each industry is probably more representative of the total R&D investment of each industry (see next graph).

Now that is known what countries and industries have high R&D expenditures, let's analyze the relation of that expenditure to actual innovation, which is represented by the Patents share in the visualization in Figure 33. This graph represents the total share of each industry, which is referenced in the previous paragraph. R&D total expenditure share by industry is represented by the blue bars and accounts for the collective share of the companies in that industry from the previous graph, and Patents share is represented by the red dots.

It is expected that industries with high R&D expenditure share should also have high Patents share.

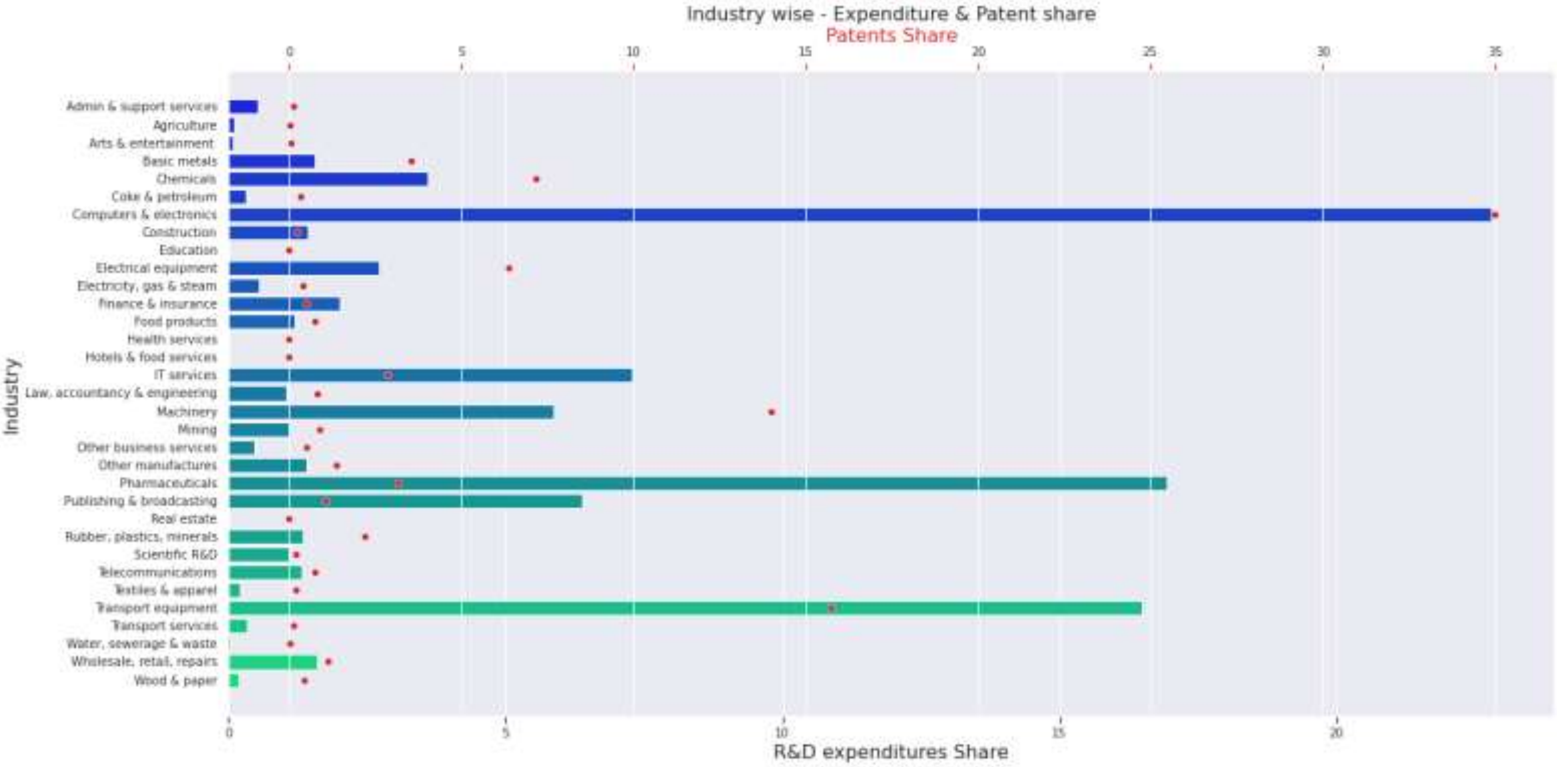


Figure 33 - R&D Expenditure and Patents share by industry

Now it is possible to identify Computer and Electronics as the industry with the highest R&D expenditure share (close to 25%). As expected, it is also the industry with the highest patents share (35%). Pharmaceuticals now appears as the second industry with the highest R&D expenditure with around 17% of share, followed by Transport Equipment, IT Services and Publishing and Broadcasting with 16,5%, 7,5% and 6% share respectively. Surprisingly, the patent share doesn't follow that distribution so closely in those industries. The Pharmaceutical sector is only the seventh sector in Patent share even though is the second in R&D expenditure share. This might be caused by other factors such as regulation and difficulty in innovating the existing solutions. IT Services also issues a low Patent share compared to R&D expenditure share.

Transport Equipment is another sector that has a much higher R&D expenditure share compared to Patent share.

In the other hand, Machinery is the third sector with the highest patent share with almost 15%, even though it occupies the sixth position in R&D expenditure. Electrical Equipment, Chemicals and Basic metals are other sectors with much larger Patent share compared with R&D expenditure share.

The final graph (Figure 34) adds Trademarks Share to R&D expenditure share and Patent share to compare how different countries perform in each category against each other. In the business perspective, trademarks are key company characteristics that are unique and legally differentiate them from other companies or products of its kind. This includes brand names, company logos, slogans or product names.

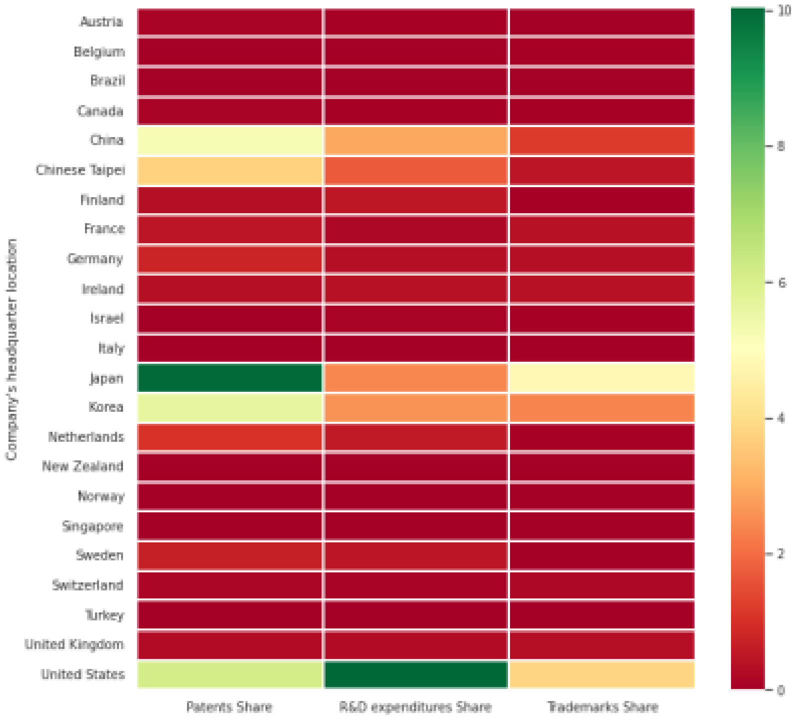


Figure 34 - R&D expenditure, Patents and Trademarks share by country

The countries that are highlighted with relatively high shares in those three categories are the United States, Japan, South Korea and China. The US has the highest R&D expenditure share but surprisingly is Japan that accounts for the highest Patent Share and Trademark Share. From the 4 highlighted countries, 3 of them are Asian and one is

North American. The rest of the countries don't have more than 4% share in any category, which emphasizes the innovating power of the 4 countries.

5.1.4 SME Growth and Adaptability

SMEs represent a group of enterprises that are considered in this research to be one of the main beneficiaries of Industry 4.0 adoption through Open Data and tools. As it was covered in across the bibliographic review subchapters, the elevated costs for implementing technologically advanced tools, the difficulty of accessing relevant decision-making data, and competition from big enterprises are pressing threats for those kinds of companies to grow and adapt.

However, understanding the value of SMEs to a country's economy comparing to high-tech industries and how that affects employment should be taken into consideration for implementing supporting measures.

The visualizations for this theme are adapted from the UN Conference on Trade and Development databases.

To start this analysis, Figure 35 represents SMEs as a share of total value added to industry from 2005 to 2019 in developed countries such as the US, Germany, France, United Kingdom, Italy and Japan, and in developing countries such as Brazil and India.

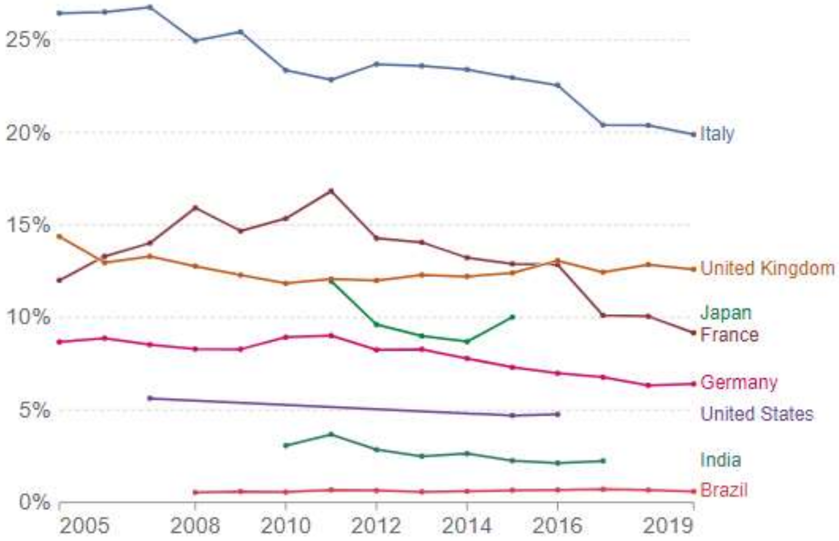


Figure 35 - SMEs as a share of total value added to industry, adapted from (UN, 2022a)

The country that started that period with the highest share in SME value added to industry is Italy with more than 25%, followed by the UK and France with around 15% and Germany with 10%. The countries with the lowest initial values are the developing countries Brazil and India. All countries with no exception have declining or stagnant values of SME value added to industry, which is be representative of the industrial value being concentrated in bigger or more technologically advanced companies. This premise can be observed in the share of high-tech company's value added to industry between 2000 and 2019, for the G7 and BRIC countries (Figure 36).

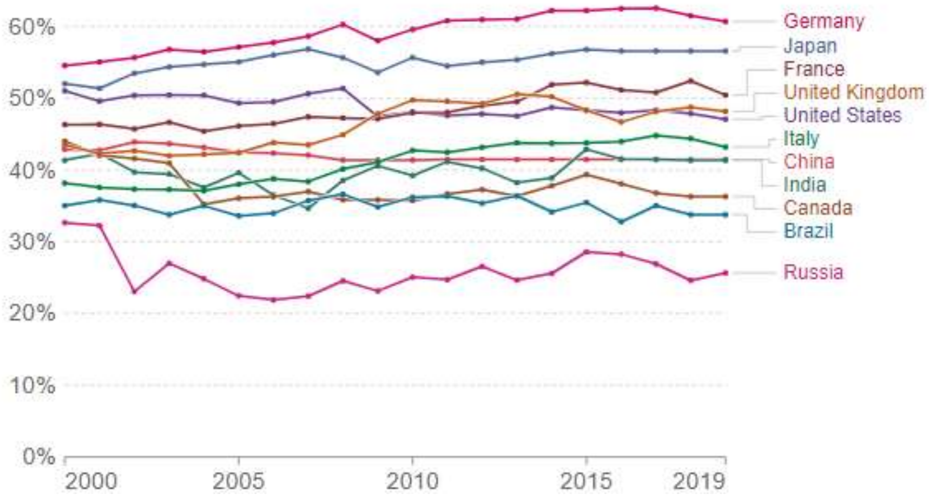


Figure 36 - High-tech companies value added to industry, adapted from (UN, 2022a)

Contrary to SMEs, high-tech companies value added to industry have been increasing its share in the total manufacturing value added in most of the countries in the study. By gaining more share of the markets' value, those companies increase their dominance over the smaller ones, which brings risks to their growth and adaptability. Another interesting finding is that G7 already have more than 50% of manufacturing value coming from high-tech industries and companies, while the BRIC countries are struggling to keep up with this change in the industrial landscape. This is shows that industries and companies that adopt high-tech technologies characteristic from Industry 4.0 have competitive advantages compared to SMEs that struggle to do the same.

5.2 Open Data for Sustainability

The second part of results and analysis obtain from the data treatment are representative of the Open Data for Sustainability themes described in subchapter 4.1.2. As it is referred previously, this subchapter approaches Economic, Social and Environmental Sustainability themes such as Collaboration for Sustainable Development Goals, Sustainability Requirements from Enterprises, Skill Migration and Carbon Emissions, which were identified as relevant themes in the bibliographic work done for Sustainable Industry 4.0 (subchapter 2.4.).

5.2.1 Collaboration for Sustainable Development Goals

Collaboration is one of the three aspects of the Open Design framework developed in this research. It is also part of the Sustainable Development Goals established by the UN. Through Goal 17 – "Partnerships for the Goals" – the UN 2022) explain that a "successful sustainable development agenda requires partnerships between governments, the private sector and civil society. These inclusive partnerships built upon principles and values, a shared vision, and shared goals that place people and the planet at the center are needed at the global, regional, national and local level".

The visualizations for this theme adapted from the UN Conference on Trade and Development databases.

The framework for this research considers collaboration essential for leveraging Industry 4.0 technologies and Sustainability efforts in order to accomplish mutual goals for human development.

This collaboration is even more important for SMEs and developing countries since they can't access tools and resources that big enterprises and rich countries can to invest in growth. For that reason, the first graph of this section (Figure 37) accesses the trend in the number of countries with bilateral investment treaties in both developing and least developed countries, from 1959 to present.

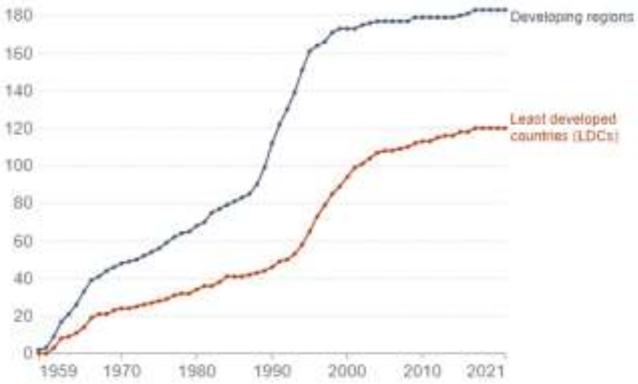


Figure 37 - Trend in the number of countries with bilateral investment treaties in both developing and least developed countries, adapted from (UN, 2022c)

Both developing and least developed countries show a growing trend in number of bilateral investment treaties with an exponential growth in the 90's decade. Since the beginning of the 21st century this trend as the tendency to decelerate for both regions.

The next graph (Figure 38) combines technology with sustainability, by showing the amount of approved funding for countries to promote the development of environmentally sound technologies in less developed countries. In this analysis are considered the G7 and BRIC.

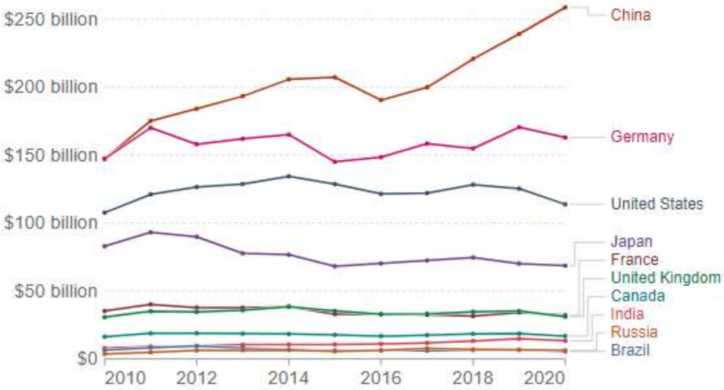


Figure 38 - amount of approved funding for countries to promote the development of environmentally sound technologies in less developed countries, adapted from (UN, 2022c)

While the G7 countries started the 2010's decade with high investments in environmental found technologies in least developed countries (cases of Germany, US and Japan), that funding has been stagnant or slightly decreasing towards 2020. On the other hand, BRIC (except China), have much lower starting funding but is also stagnant across the decade. China is the country with a big effort in collaborating with less the developed countries in order to develop environmentally sound technologies. In 2010 China was already the country that allocated more funding towards that goal, together with Germany. However, comparing to the other countries, it had rapid acceleration in that founding throughout the years, culminating in more than $250 billion approved in 2020.

Now that the countries that fund those investments are identified, we now identify which nations are receiving that support. Figure 39 is a geographic visualization of the dollar value of financial and technical assistance committed to developing countries in 2019.

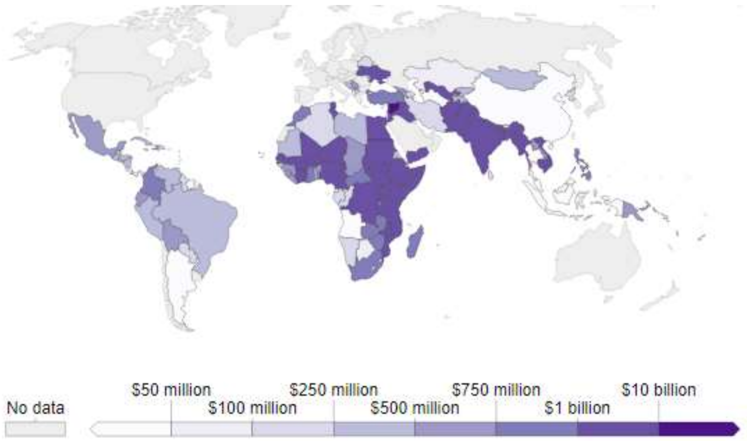


Figure 39 - Geographic visualization of the dollar value of financial and technical assistance committed to developing countries in 2019, adapted from (UN, 2022c)

From a general point of view, African nations seem to receive the majority of that founding. South American and Asian Countries also receive a sizable portion of those investments. Curiously, two of BRIC nations, Brazil and India are also big beneficiaries of that collaboration, with the later receiving between $1 and $10 billion of funding in 2019.

For this theme it is also important to track the rate of success in Sustainable Development Goals implementation across the world. For that there are multi-stakeholder monitoring frameworks that track that implementation. Figure 40 shows the number of coordinators (countries) tracking the implementation of SDG commitments for the year 2018, in which a country can identify as both a provider and a recipient of development coordination.

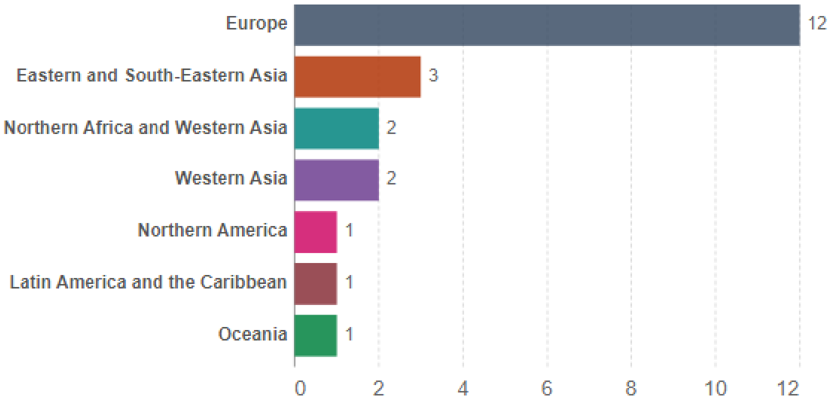


Figure 40 - Number of coordinators (countries) tracking the implementation of SDG commitments for the year 2018, adapted from (UN, 2022c)

Europe has the majority of collaborators, 4 times the number of collaborators of the second region, which is Eastern and South-Eastern Asia. Being a governmental union, it makes sense for Europe to promote collaboration between its nations as the development of each one positively or negatively impacts the rest of them to a certain extent.

The collaboration of organizations is important for countries at a public and private level. Partnerships between companies is already widely present though producer-supplier partnerships for example. However, arguably more important than accelarate that kind of collaboration is to invest in public-private and social partnerships. Those partnerships are usually done to service the population and society of a country or union, in which the private party bears significant management responsibility to providing a public asset or service (van Ham & Koppenjan, 2001). In this research, that kind of relationship is measured in Figure 41 through the dollar amount commited to public-private partnerships for infrastructure between 2000 and 2020 considering the BRIC countries.

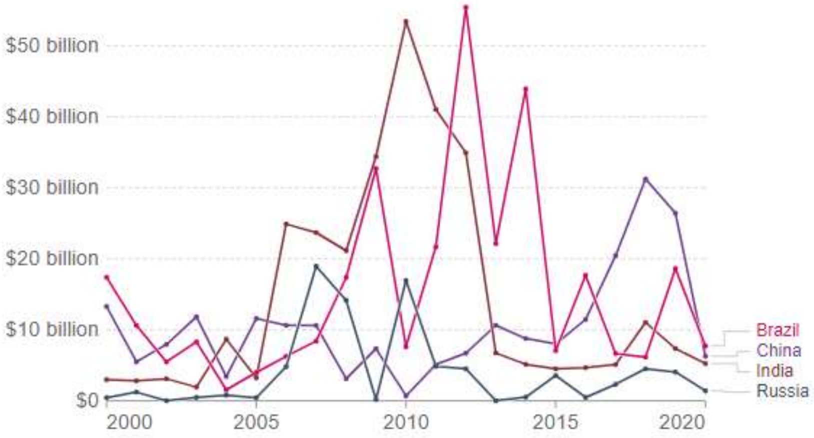


Figure 41 - Dollar amount committed to public-private partnerships for infrastructure between 2000 and 2020 considering the BRIC countries, adapted from (UN, 2022c)

At a first glance, all BRIC countries have great volatility in the amount of funding for those partnerships. All of them comit around the same values in 2020 that commited in 2000 (between $0 and $20 billion), with big spikes along the way. Brazil and India commited the highest values between the BRIC countries, which reached around 450 billion at the beginning of te 2010's decade. That amount rapidily declined in the following years. More recently, China increased that amount of funding, but it reached only a fraction of Brazil and India highest values, and rapidly decrased.

5.2.2 Sustainability Requirements from Enterprises

The previous theme is focused on the overall collaboration for Sustainable Development Goals particularly from governments and public organizations. However, understanding the efforts and requirements of enterprises regarding Sustainability is crucial for a joint effort from public and private organizations for a Sustainable future. Since analyzing this may seem subjective, one of the most objective ways of verifying the endeavor of companies for implementing sustainable operations is to quantify and verify their Sustainability Reports.

Those Sustainability reports are split into two categories by the UN: Reports regarding minimum requirements and advanced requirements. Minimum requirements cover the company's governance practices towards its economic, social and environmental impacts, whereas advanced requirements include more complex KPI's such as stakeholder engagement, impacts beyond the company boundaries along the supply chain, supplier and consumer engagement on sustainability issues, procurement and sourcing practices and environmental performance (UN, 2022b).

The visualizations for this theme are adapted from the UN Conference on Trade and Development databases. Below (Figure 42) is quantified the number of companies meeting the minimum sustainability requirements across the G7 and BRIC countries, for the year 2020. Figure 43 represents the same values globally in a geographic visualization.

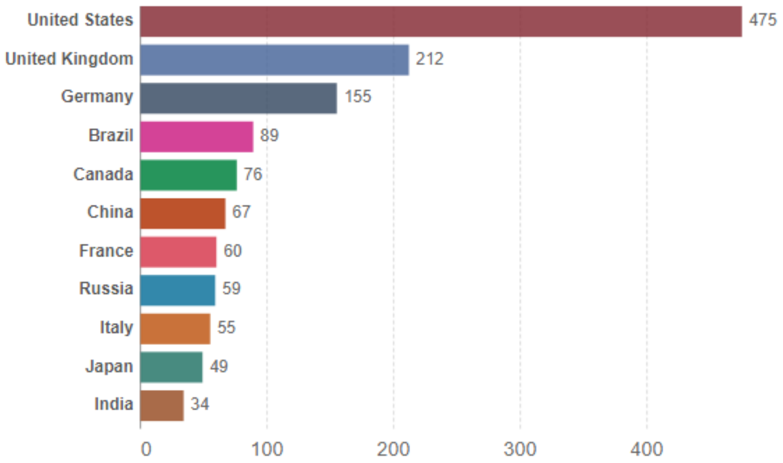


Figure 42 - Number of companies meeting the minimum sustainability requirements across the G7 and BRIC countries, for the year 2020, adapted from (UN, 2022b)

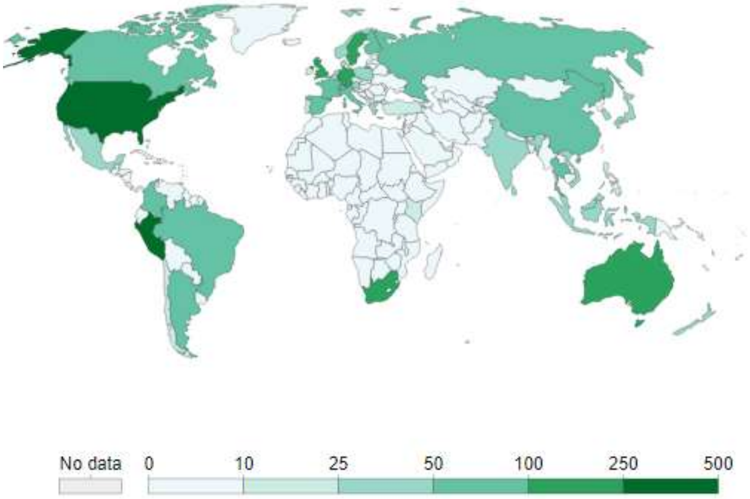


Figure 43 - Number of companies meeting the minimum sustainability requirements globally, for the year 2020, adapted from (UN, 2022b)

The US is the clear global leader in terms of corporate sustainability efforts with 475 companies meeting minimum requirements, more than double the second country, the UK with 212. Germany closes the top 3, which means that the G7 companies have big efforts towards corporate sustainability. Brazil is the leader of BRIC countries is this matter, being fourth out of the 11 countries. India closes the ranking with 34 companies.

The continents that appear to have a higher share of sustainable companies are North America, South America, Europe and Oceania, whereas Asia and Africa either have lower share or lack data for the analysis. Figures 44 and 45 now represent the number of companies meeting advanced sustainability requirements.

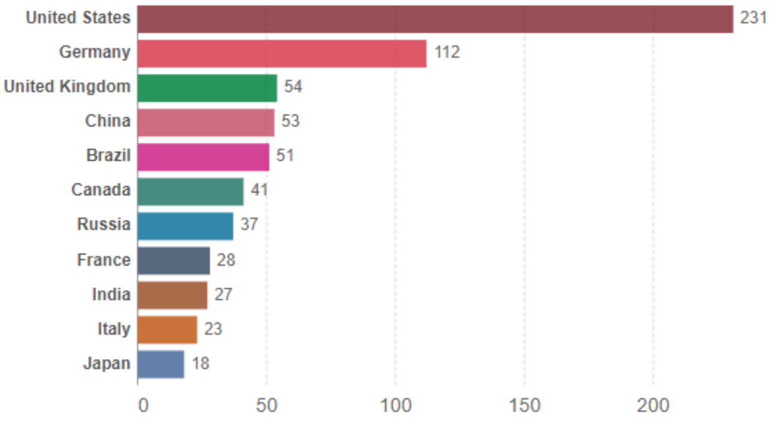


Figure 44 - Number of companies meeting advanced sustainability requirements across the G7 and BRIC countries, for the year 2020, adapted from (UN, 2022b)

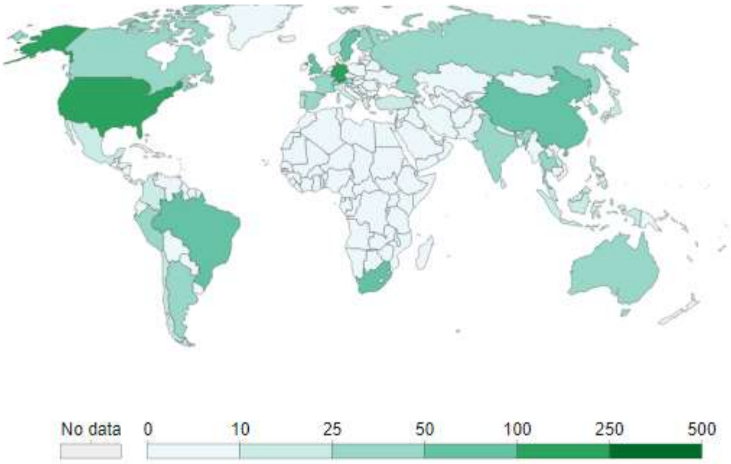


Figure 45 - Number of companies meeting advanced sustainability requirements globally, for the year 2020, adapted from (UN, 2022b)

The outstanding leader is also the US with a whopping 231 companies, more than double the second country Germany, and almost five times the third, United Kingdom. China now stands in fourth position and Brazil in fifth. BRIC countries rank higher in advanced requirements in comparison to minimum requirements, however the G7 is still the group with most sustainable companies.

5.2.3 Skill Migration

Skill Migration can be defined as the trends in both supply and demand for professional skills throughout the years (World Bank, 2022). As economies and labor markets change, much because of the evolution of consumer behavior and the adoption of new technologies, so do the skills that are demanded from enterprises and public services. Better education also means better qualified workers that migrate from traditional industries to more technologic and digitalized ones (Kerr et al., 2016). This dynamic is also accelerated from Industry 4.0 adoption. However, since not all countries are equal in economic growth, technology adoption and industry digitalization, naturally Skill Migration varies not only between industries but also in geography.

Because of that, governments and researchers understand that rapidly evolving labor markets require skills, occupations and industries to be analyzed and studied with available data (Kerr et al., 2016). The research for this theme gives a data-driven perspective to Skill Migration across the world through different categories of industry skills.

The first skill category incorporates specialized industry skills, such as cybersecurity, social services, politics, environmental engineering, law, machining and retail. This category can indicate the overall migration of workers from less skilled industries, such as agriculture and fishing, to jobs that required specialized knowledge, in both public and private sectors. Figures 46, 47 and 48 compare that migration in the two biggest economies in the world, US and China, the developed G7 countries and the developing BIC countries (Brazil, India and China), respectively.

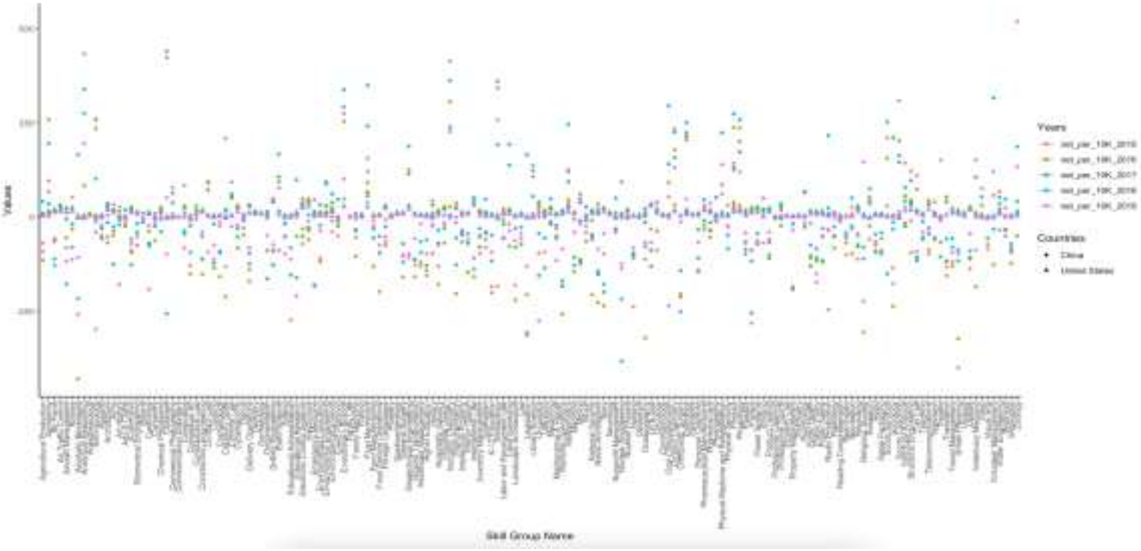


Figure 46 - Skill Migration in Spatialized Industry in US and China

The skill migration for Specialized Industry in US and China looks mixed overall, depending on the specific industry.

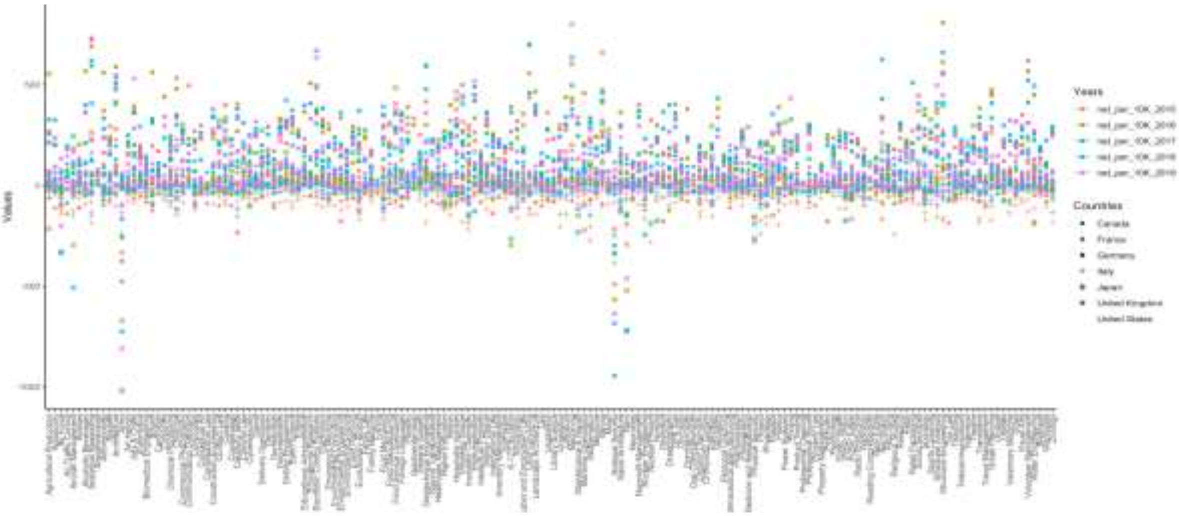


Figure 47 - Skill Migration in Spatialized Industry in the G7 countries

The G7 countries demonstrate a clear positive migration trend for specialized skills, with very few exceptions. From those developed countries, Japan and Germany are highlighted in National Security, Army and Navy for been losing skilled workers in those categories over the years.

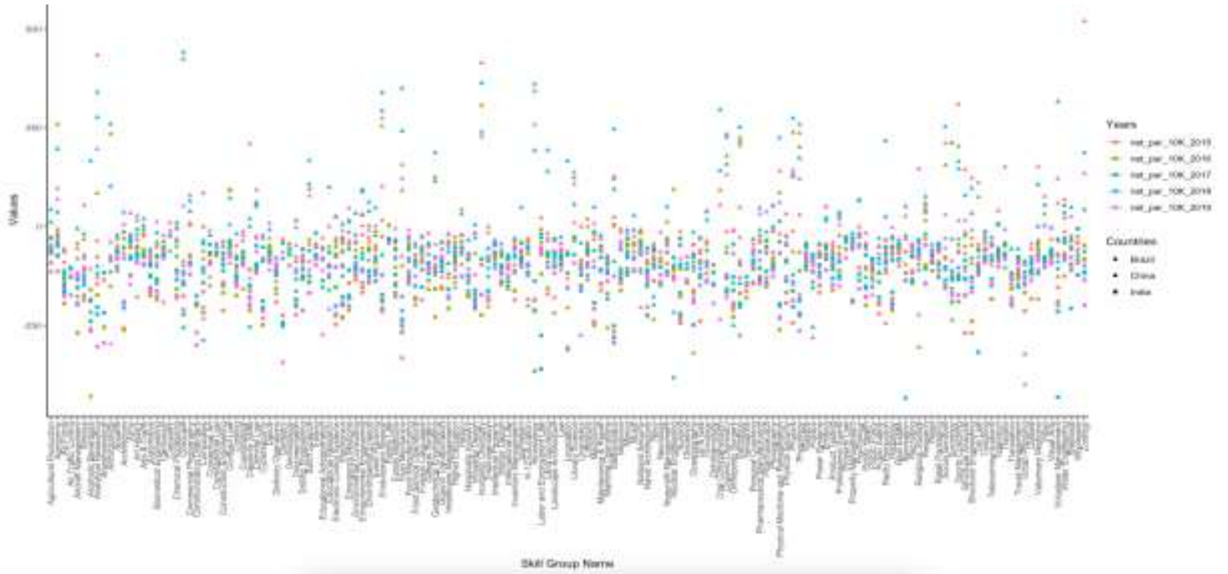


Figure 48 - Skill Migration in Spatialized Industry in the BIC countries

On the other hand, the developing economies of Brazil, India and China demonstrate an overall negative skill migration trend in specialized industries, which contrasts with the positive trend of G7 countries.

The second type of skill in this study is Business Skills. This category incorporates skills that correlate more with corporate and business jobs, like Economics, Administrative Work, Human Resources, Bookkeeping, Corporate Communications, Manufacturing Operations, Advertising and Project Management. Figures 49, 50 and 51 should be representative of the skill migration trend in business and enterprises for US and China, G7 and BIC countries, respectively.

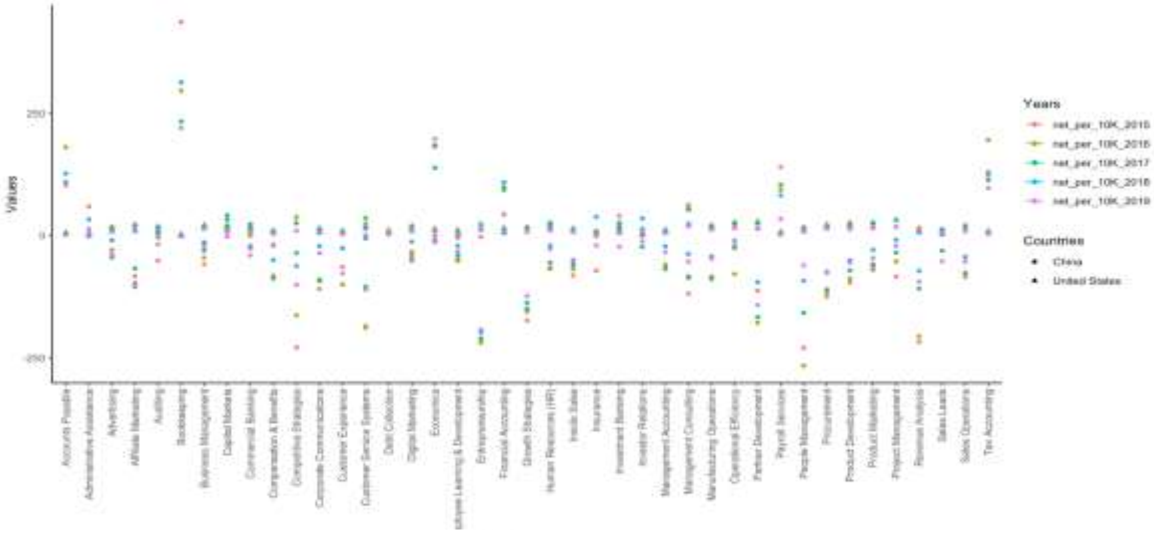


Figure 49 - Skill Migration in Business in US and China

The US and China appear to have negative flows of skilled workers in business sectors, even though China had positive skill migration in bookkeeping, economics, financial accounting, payment services, accounts payable and tax accounting in every year between 2015 and 2019. This demand for financial business skills in China is growing, while it is stagnant in the US.

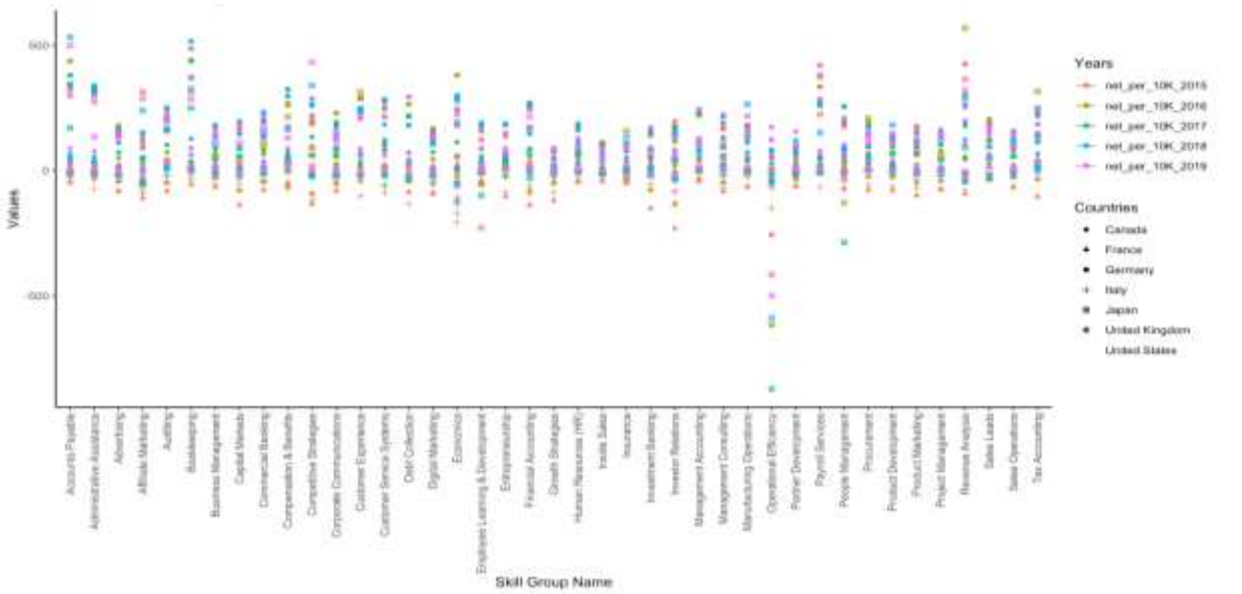


Figure 50 - Skill Migration in Business in the G7 countries

The G7 countries have positive migration and high demand for business skills between 2015 and 2019. The only category that seems to be an outlier is Operational Efficiency. Even though the most logical conclusion is that Operational Efficiency have been losing importance for the G7 countries, based on the analysis of the previous themes of this research, automation and digitalization caused by Industry 4.0 adoption might be the driver for business demanding less skilled workers in that category, since they can be now replaced by more efficient technologies such as Artificial Intelligence and Big Data Analytics.

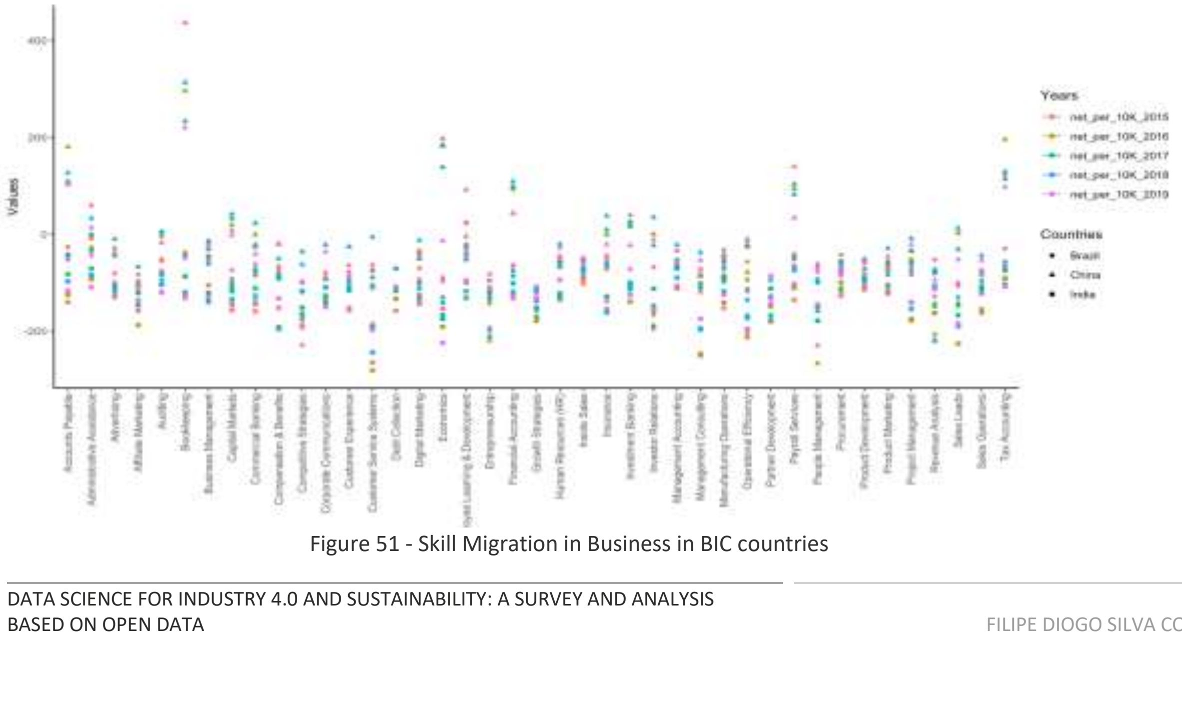


Figure 51 - Skill Migration in Business in BIC countries

BIC countries appear to have negative migration for the majority of business skills, with China being the exception by having high demand for financial business skills, as it was discussed previously.

The third category, Soft Skills, include important social skills such as problem solving, leadership, teamwork, communication, time management, persuasion and negotiation, which are essential skills for workers independent of location or industry sector. The analysis for the same three groups (US and China, G7 and BIC) are represented in Figures 52, 53 and 54, respectively.

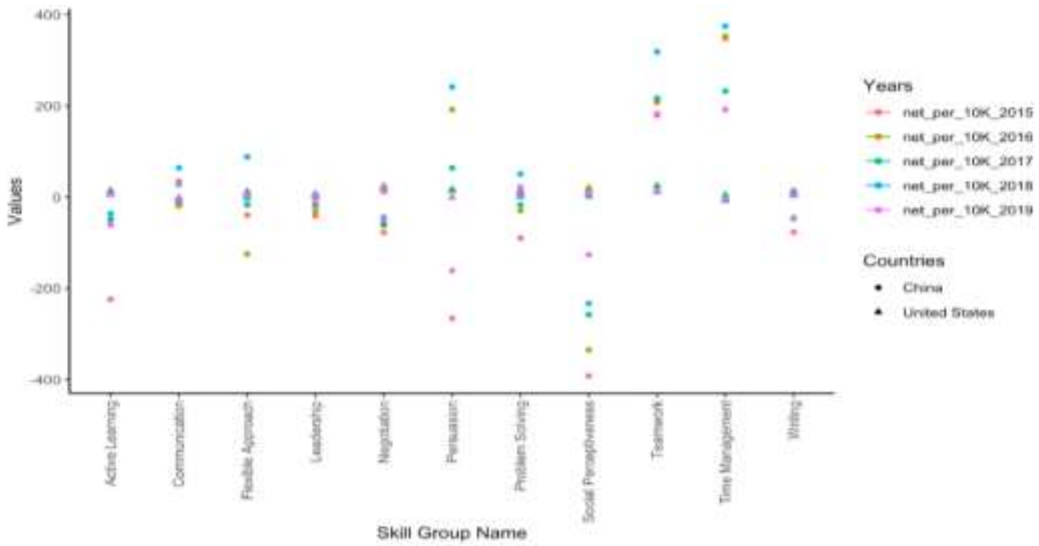
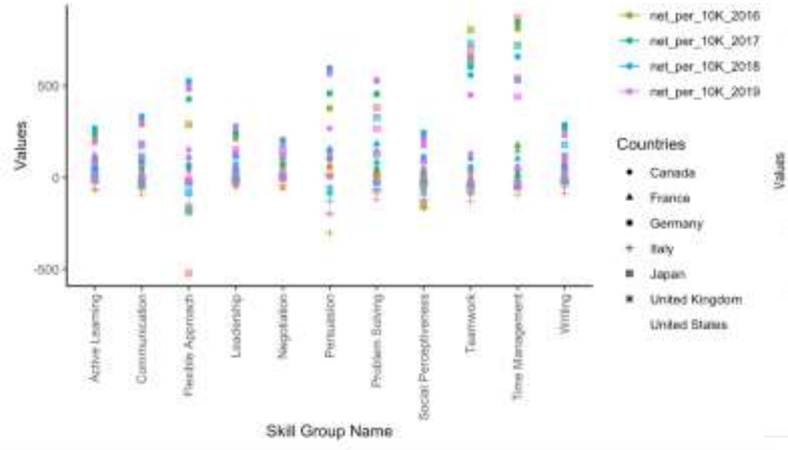


Figure 52 - Soft Skill Migration in US and China

This visualization is highly relevant for this research by highlighting one of the main hypotheses of this research, which is that Social Sustainability is being neglected by countries and companies in comparison with the Economic and Environmental spectrums. As it is represented above, while Teamwork and Time Managements Skills are highly demanded and valued by the US and China, Social Perceptiveness have been having huge outflows of skilled workers in that field in both countries.

Contrarily to Operational Efficiency, the business skill mentioned above that may have been replaced by Industry 4.0 technologies, social skills such as perceptiveness are probably the most difficult skills to replace by technology, which accentuates the need for attention and valorization for social skills and issues.



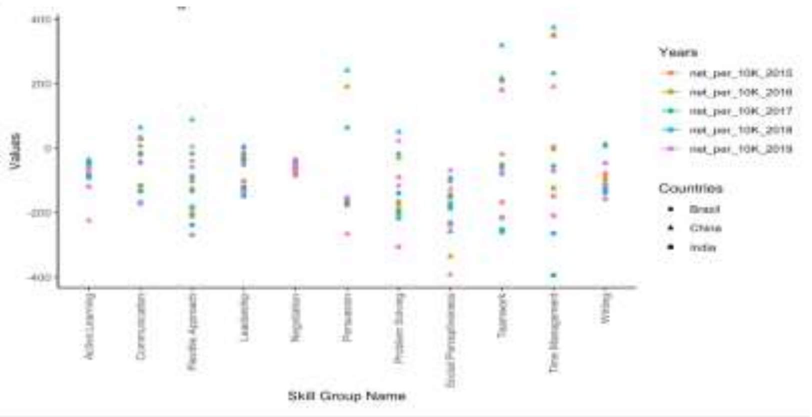


Figure 54 - Soft Skill Migration in G7 countries

Figure 53 - Soft Skill Migration in BIC countries

When comparing the G7 countries to BIC countries, G7 still have overall positive demand for soft skills, while BIC have a negative migration trend, mainly from India and Brazil. Social perceptiveness is still the most neglected soft skill across those developing countries which is worrying in terms of future Social Sustainability.

The fourth and final category studied in this database is Disruptive Tech Skill Migration. If the assumptions reiterated for Skill Migration in the Research Model subchapter are correct, Industry 4.0 adoption shouldn't be considered a barrier to obtaining Social Sustainability, particularly because it would generate new high skilled jobs that would compensate for the ones replaced by automation and digitalization. Those kinds of jobs are represented by the following visualizations in Figures 55, 56 and 57.

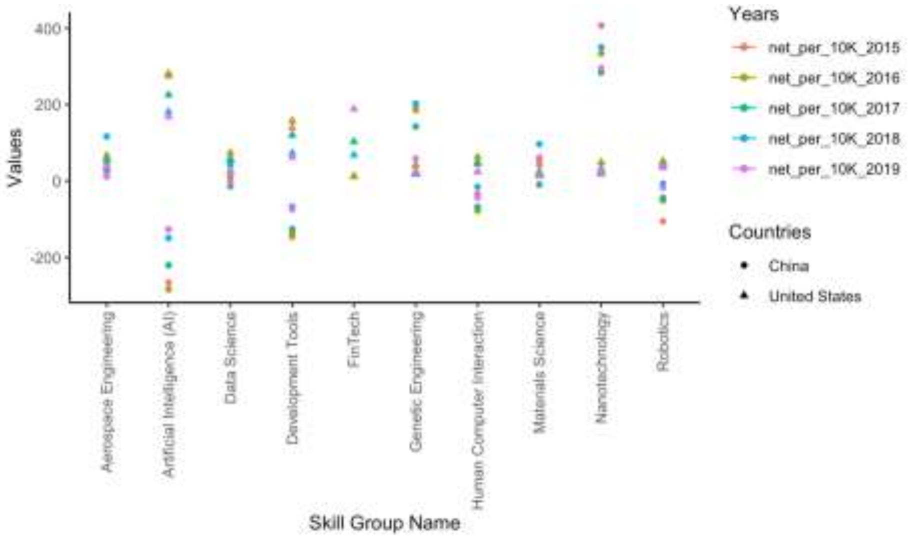


Figure 55 - Skill Migration in disruptive tech in US and China

The overall trend for Disruptive Tech Skills Migration is positive for both China and US. Industry 4.0 skills such as Data Science, Human-Computer Interaction and Robotics have been receiving a lot of skilled workers between the year 2015 and 2019, which meets the assumption cited above. Artificial Intelligence is a surprising outlier. AI is one of the Industry 4.0 technologies that are expected to have high demand as Industry 4.0 is implemented in both developed and developing countries. However, for that particular skill group, China has been losing demand for AI skilled workers, while the US have been recruiting more workers in all High-tech fields, including AI.

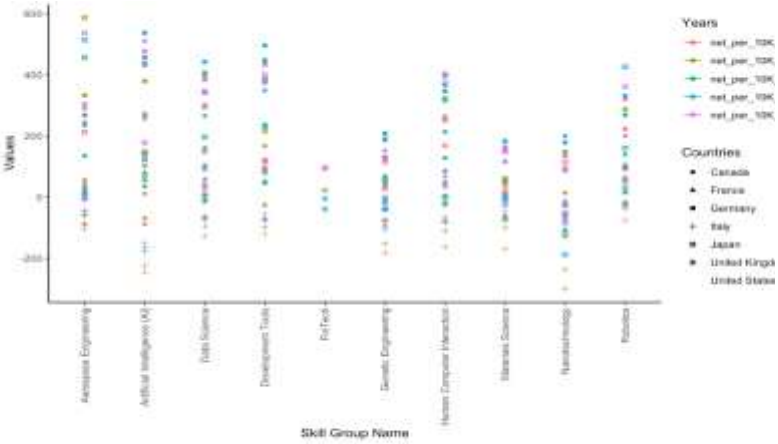


Figure 57 - Skill Migration in disruptive tech in the G7

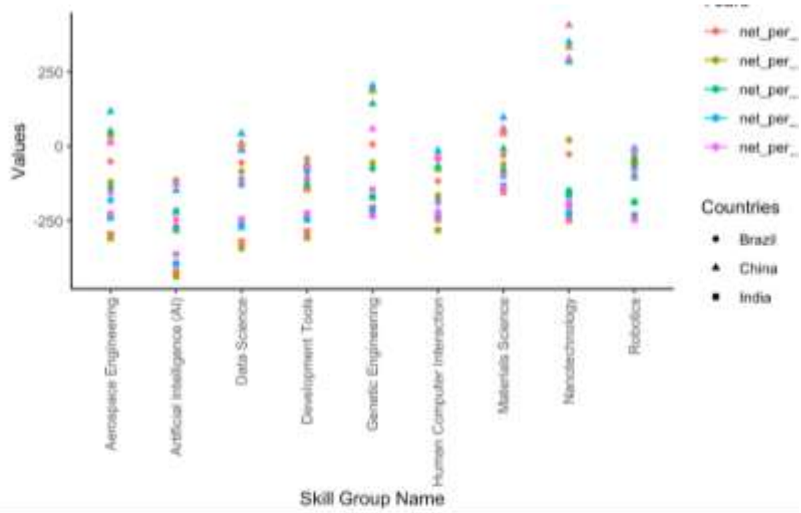


Figure 56 - Skill Migration in disruptive tech in the BIC countries

When comparing the G7 countries with the BIC countries, similarly to the other skill categories, the G7 have very high values of positive migration for these high-tech skills, while the BIC post mixed results, tending more to the negative side, particularly in Brazil and India, and to the positive side in China.

It is not possible to reiterate a definite and global conclusion about Industry 4.0 being able to compensate the loss of jobs in automated and digitalized fields by creating High-tech skilled ones. However, the evidence shows that at least in richer countries such as the G7 and China, as Industry 4.0 is implemented with higher efficiency and quality, it is clear that the demand for workers will shift more and more for high skilled tech-focused ones.

5.2.4 Carbon Emissions

Considering the bibliographic review done in chapter 2.4. regarding the Environmental pillar of Sustainability, the CO_2 emissions are undoubtedly one of the main issues to consider and study this research.

In this research theme the objective is to use analysis and visualizations to answer environmental questions such as which countries pollute the most in terms of CO_2 emissions. It is expected that more developed countries, with higher GDP and industrial output, also pollute the most. However, understanding what their emissions trend over time, is highly relevant since many countries are putting big efforts into greener economies with less carbon emissions.

Carbon emissions are very different depending on the country but differ mostly across industries. Understanding which sectors contribute more heavily to carbon emissions might give a direction to which sectors should receive more attention from governments and researchers in order to decarbonize production and supply-chain.

The following chart (Figure 58) represents the CO_2 emissions in metric tons per capita since 1960 until 2016, from the G7 countries and BRIC countries (except Russia), which are representative of highly developed and fast developing countries, respectively.

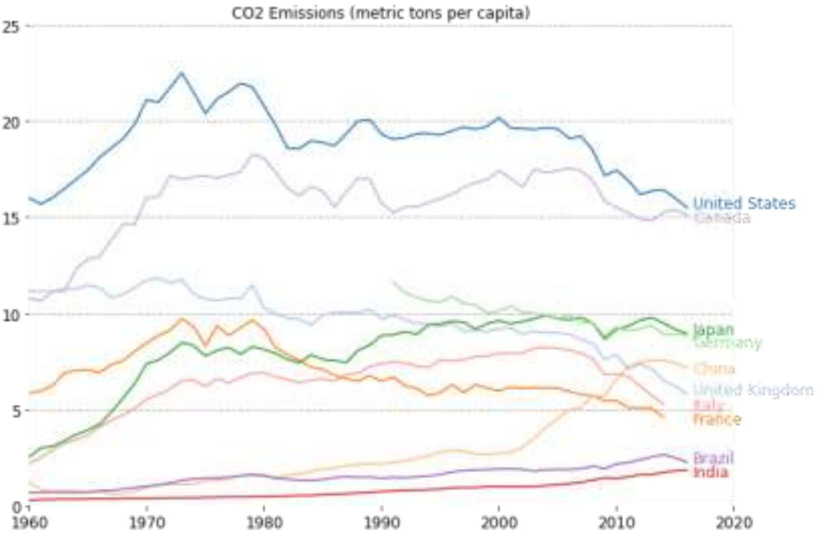


Figure 58 - CO2 emissions in metric tons per capita since 1960 until 2016, from G7 and BRICs

As it was expected, the biggest country of the world by GDP, the United States, is also the biggest polluter from 1960 to 2016. It is followed by Canada, which has a very similar trend over the years. Just like France, the US and Canada peaked their emissions in the 1970's and have been slowly reducing those emissions since then. From the G7 countries, the UK has been reducing those emissions drastically since 1960 and Japan is the only one that has been increasing CO_2 emissions in that period.

All developing countries have been increasing their CO_2 emissions. While this increase is almost linear from Brazil and India, China had an exponential increase in emissions since 2000.

Overall, if we look to the trends of every country, we can see that there are three distinct periods of carbon emissions. The first period, from 1960 to 1970, marks fast growth in carbon emissions across all countries. The majority of those emissions peak in that decade. The second period, between 1970 and around 2006, represents a stable CO_2 emissions pattern. Then, with a few exceptions, CO_2 emissions appear to decline at a faster rate from 2006 to 2016.

Considering the second and third periods mentioned above, the following charts (Figure 59) compare the trend in CO_2 emissions from 1970 to 2016 and from 2006 to 2016 in which the circles represent the top 10 polluting countries.

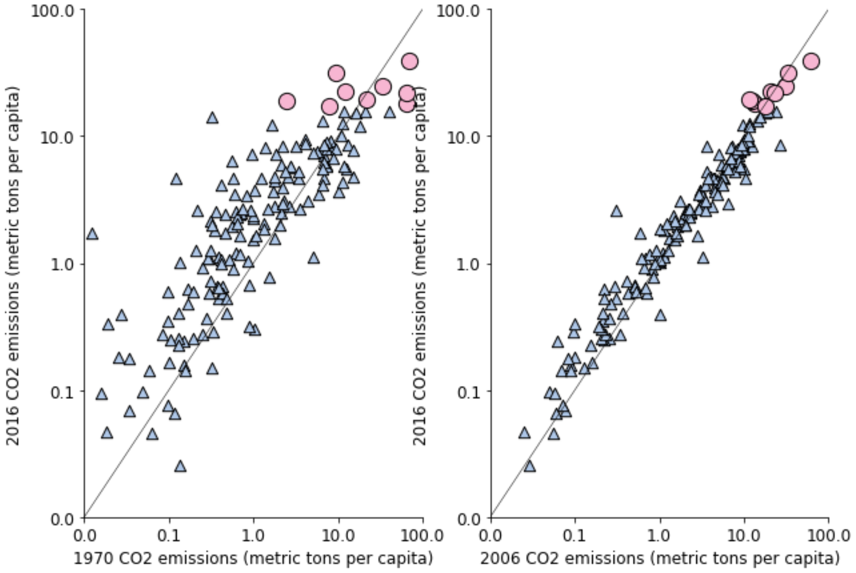


Figure 59 - Trend in CO2 emissions from 1970 to 2016 and from 2006 to 2016

The top 10 polluting countries in terms of CO_2 emissions in metric tons per capita identified in both charts are Qatar, Kuwait, Trinidad and Tobago, Saudi Arabia, United Arab Emirates, Bahrain, New Caledonia, Gibraltar and Brunei.

The trend line separates the countries that increased their CO_2 emissions (above the line) from the ones that decreased those emissions (bellow the line). While around 70% of countries increased their CO_2 emissions per capita from 1970 to 2016, only 58% of them increased those emissions from 2006 to 2016. The latest period also shows that the countries are much closer to the trend line than the first period, which represents a common behavior from nations around the world in the later years. However, 58% is still a very elevated number considering the efforts that are required to achieve global Environmental Sustainability.

Metric tons per capita is a good metric for measuring relative improvements between countries with different populations sizes, however absolute CO_2 emissions is more relevant when considering which countries have the most impact on global emissions (Dong et al., 2019). Below are represented the top 10 polluting countries but now regarding absolute CO_2 emissions in kilotons (Figure 60).

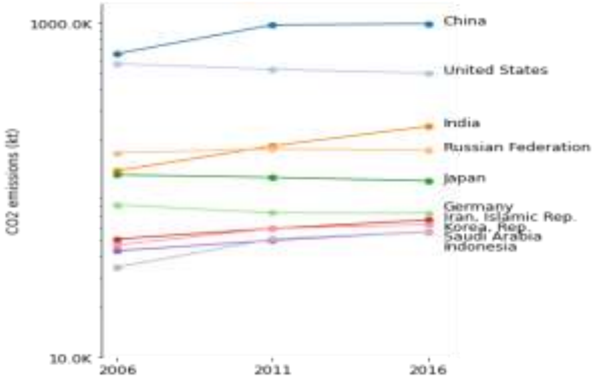


Figure 60 - top 10 polluting countries in absolute CO2 emissions

The top 10 polluting countries in absolute terms are China, United States, India, Russia, Japan, Germany, Iran, South Korea, Saudi Arabia and Indonesia. China is the biggest contributor to CO2 emissions since 2006 and has been increasing them, alongside with India, Iran, Korea, Saudi Arabia and Indonesia. United States, Japan and Germany are the biggest polluters that have been decreasing those emissions.

In terms of value added to GDP, the following visualization adapted from the UN Conference on Trade and Development, offers a geographic perspective how well are carbon emissions actually determinant in increasing GDP in 2018 (Figure 61).

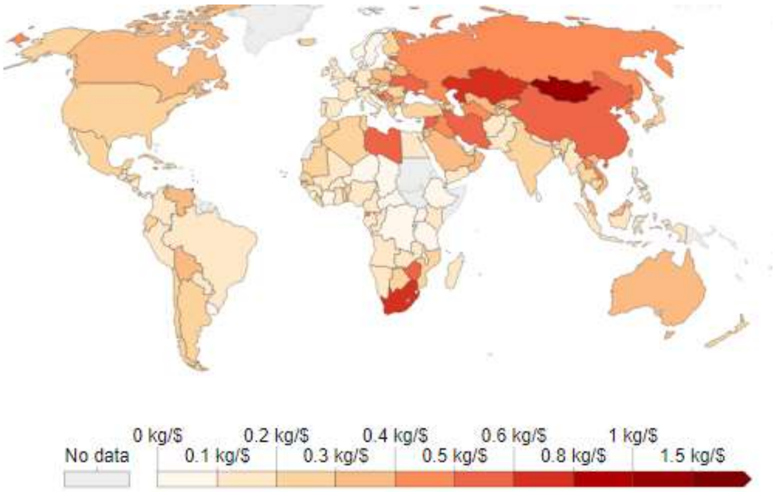


Figure 61 - Geographic representation of carbon emission value added to GDP in 2018, adapted from (UN, 2022b)

While occidental countries seem to have relatively low CO2 emissions value added to GDP in the order of between 0 kg/$ and 0,4 kg/$, oriental countries such as China have both high CO2 emissions with low value to its GDP, at around 1 kg/$.

Now that we can identify the CO2 emission trend across geographies, it is necessary to identify which economic sectors contribute the most to those emissions. In Figure 62 are represented the CO2 emissions share of the 4 sectors that account for the majority of emissions in high polluting countries.

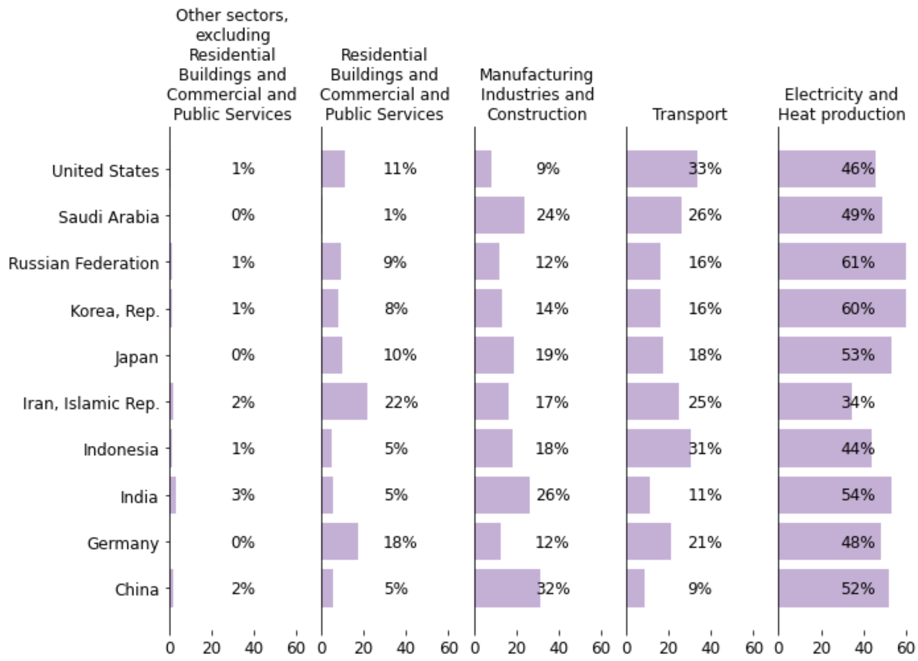


Figure 62 - CO2 emissions share of the 4 sectors that account for the majority of emissions of high polluting countries

The four sectors that account for the majority of CO_2 emissions are Electricity and Heat Production, Transportation, Manufacturing and Construction and Residential, Commercial and Public buildings. Eletricity and Heat production is the sector with the biggest share of emissions in every single country, accounting for roughly 50% across all of them. Transportation is the second most polluting industry in the United States, Saudi Arabia, Russia, Korea, Iran, Indonesia and Germany, while Manufacturing and Construction is the second sector for China, Japan and India. Those four sectors combined account for almost 100% of CO_2 emissions across all the mentioned countries.

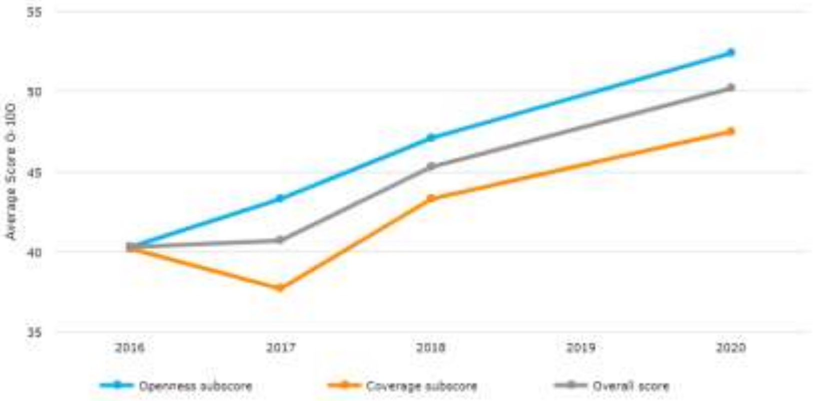
5.3 Open Design for Sustainable Development

The goal of the research for this subchapter is to correlate the Open Data Scores referenced and explained on subchapter 4.1.3. to other variables that can are representative of Economic, Social and Environmental Sustainability. First the trend of Open Data across the world are identified an analysis, with the second part focusing on the three sustainable pillars particularly on the US and China. Finally, this study gives a special emphasis on Open Data for Social Sustainability, since it was identified previously as the Sustainable pillar that receives less effort from researchers, policymakers and enterprises and this research aims to contribute meaningfully to that cause.

5.3.1 Open Data Trends across the World

To better understand the trend of openness between the years 2016, 2017, 2018 and 2020, this theme research starts by analyzing some visualizations sourced in ODIN's Open Data Inventory Annual Report 2020/2021 (ODW, 2021), which present an overall view of openness trends across different geographic regions.

In Figure 63 is shown the trend of average scores for the openness and coverage subscores, as well as the average overall score, between 2016 and 2020. At the right side (Figure 64) are also represented the overall scores for openness in the year 2020.



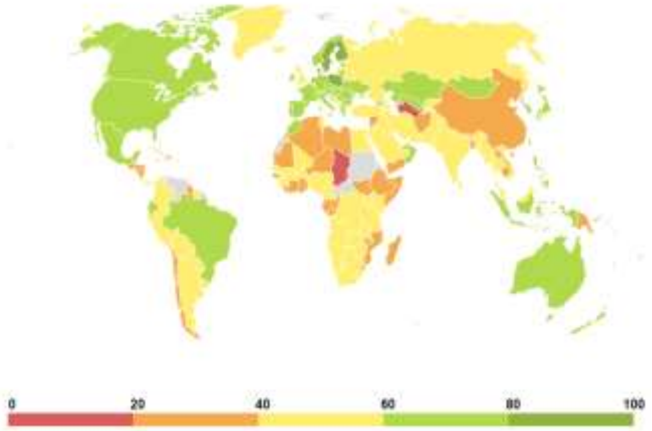


Figure 64 - Trend of average scores for openness, coverage and overall scores, adapted from (ODW, 2021)

Figure 63 - Global Overall Openness scores in 2020, adapted from (ODW, 2021)

The trend is clearly positive, with exception of the year 2017, that had a decrease in the average coverage subscore in that year. It rapidly gained positive momentum in the following years, but it is still the lower value between the three in 2020.

It is possible to see that the most open regions are Europe, North America and Oceania, with Africa and Asia scoring lower than those continents.

According to ODW (2021), the median score for 2020 is 48.8, which is an increase of 7 points compared to the 2018 value, representing a 17% increase in a two year period.

Figure 65 represents the trend in the average overall score grouped by class, from low income to high income.

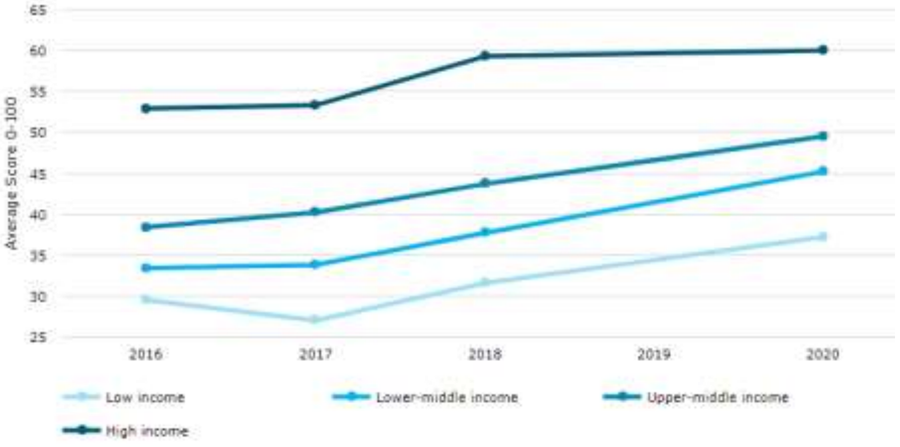


Figure 65 - Trend in the average overall score grouped by class, from low income to high income, adapted from (ODW, 2021)

While the high-income countries had an initial overall score much higher than the other three, it had an increase in that value between 2017 and 2018 and stagnated from that year to 2020. This means that the 17% increase in global openness between 2018 and 2020 was derived from lower income regions. That is clearly represented by the low and lower-middle income average scores started low compared to the high income but had a rapid acceleration in openness that continues till 2020.

In terms of regional relative change in the overall score between 2018 and 2020, Figures 66 and 67 give a numerical and geographic representation.

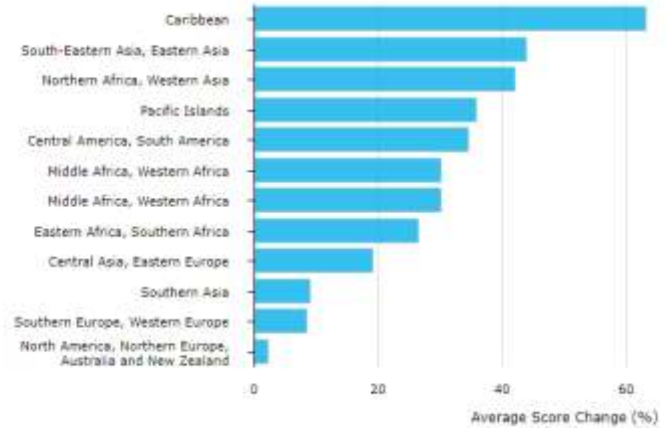


Figure 66 - Regional relative change in the overall score between 2018 and 2020, adapted from (ODW, 2021)

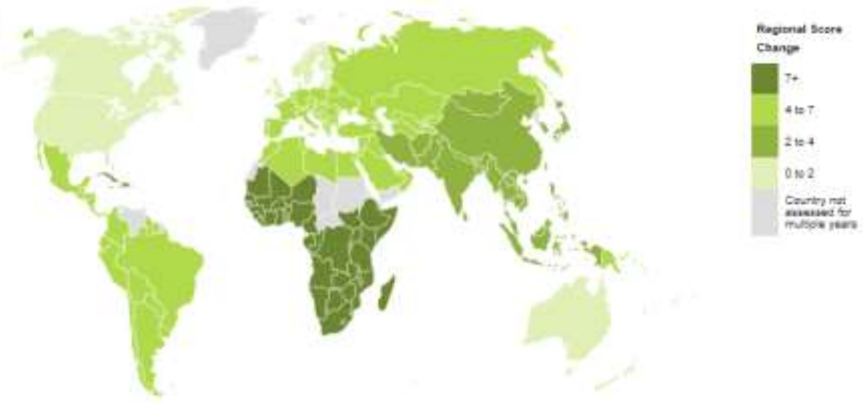


Figure 67 – Geographic visualization of regional relative change in the overall score between 2018 and 2020, adapted from (ODW, 2021)

The conclusion of the previous analysis is also clear in those images, since lower-income regions from developing continents such as Africa, Asia, Central and South American had the biggest score increase for that period. On the other hand, the most developed regions such as North America and Oceania had the lowest relative increases. Europe, while being a developed region, had mixed changes, which probably depend on the specific policies and income of each country.

5.3.2 Open Data for Sustainability in the US and China

The second part of the analysis of the overall relation between Openness and Development focuses on the correlation between the Openness Scores with Economic, Social and Environmental categories of two countries that have a determinant influence on the future of global development and are also the two biggest countries in world by GDP: the US and China.

Table 28 represents the group of categories with the highest score from year 2015 to 2020 in the US and China. For each country are shown the 6 categories with the highest score and its value, by year.

Table 28 - Highest ranked categories from USA and China from 2015 to 2020

		2015		2016		2017		2018		2020	
Categories		CHIN A	US*	CHIN A	US	CHIN A	US	CHIN A	US	CHIN A	US
Economic	National Accounts	72					89		94		94
	Central Government Finance	72		61		72		67	89	61	94
	International Trade	69		63	88		94		100		100
	Balance of Payments	69			88	69	94	69	100	56	100
	Labor	65			95		90				95
	Money and Banking			63		69		63		50	
	Price Indexes							72			
Social	Population and Vital Statistics	65		60	90		95		100		
	Health Facilities			60		65				55	
	Education Facilities			55	85						
	Education Outcomes				85						
Environmenta l	Resource Use					65		67		61	
	Energy Use					55	80	61	89	63	88
OVERALL SCORE		55,15		44,10	74,80	41,50	68,60	44,40	73,69	35,10	70,40

(*) – Data not available.

The Overall Scores in the US indicate 2016 to be the year with the highest openness followed by 2018. The year 2017 brought a pullback in the overall openness and 2020 was the year with the less openness in the study. China's scores show a descending trend of openness from 2015 to 2020 with 2018 being an outlier. By comparing the scores of US and China, it is clear that the US has a higher openness across the Economic, Social and Environmental categories, with the highest scores being International Trade and Balance of Payments in 2018 and 2020 and Population and Vital Statistics in 2018. China's highest score was 72 in National Accounts in 2015, Central Government Finance in 2015 and 2017 and Price Indexes in 2018, all of which are lower than the lowest American score between the represented categories, which is 85 for Education Facilities and Education Outcomes in 2016. We can also conclude that the Economic pillar had the highest influence on the openness for both countries, since it accounts for 7 of the 13 categories with the highest scores and it is followed by the Social pillar with 4 categories. The Environmental pillar accounts for only 2 of the 13 pillars but it has gained influence on the overall openness mainly in China, appearing between its 6 highest ranked categories since 2017.

The categories with the highest openness scores are International Trade, Balance of Payments, Population and Vital Statistics, Central Government Finance and Labor. However, to analyze the importance of the categories into the openness of all geographic regions, it is necessary to understand not only the score value but also the co-occurrence of the category between the highest ranked ones for each year. This can be obtained and analyzed by using a visualization of association rules (Figure 68). The *support* indicates how frequently a set of items appear, the *confidence* how often a *support-rule* is true, and the *lift value* is the ratio between the confidence and the expected confidence of that rule. If the value is higher than 1, they are positively correlated, if it lower than 1, they are negatively correlated, and if it equal to 1 they are independent.

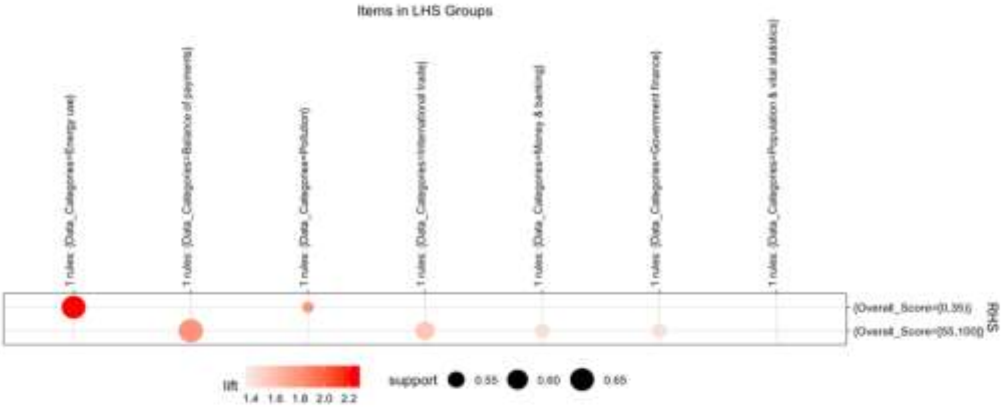


Figure 68 - Association rules visualization

7 rules were generated, assuming a minimum support value equal to 0.02 and a confidence value equal to 0.5. All rules have a *lift value* greater than 1, which mean the 7 represented categories not only have high openness values, but also have high co-occurrences between the highest scored categories for a large number of countries evaluated in the dataset.

According to this analysis, International Trade, Balance of Payments and Central Government Finance are important categories for openness in Economic Sustainability, Population and Vital Statistics in Social Sustainability and Pollution and Energy Use in Environmental Sustainability.

5.3.3 Open Data for Social Sustainability

As it was mentioned before, Social Sustainability is the arguably the pillar that gets less attention from researchers and organizations. One of the main objectives of this research is to contribute for the social cause exploring the concepts developed throughout the study. For that reason, it is important to understand if by leveraging information, technology, and tools, Open Data friendly countries can establish happier sustainable societies and serve as an example of social success for the rest of the globe.

To study this theme, it is used again the Open Data Scoring dataset that evaluates openness across different countries with scores from 0 to 100, considering the values for the year 2020. For the social sustainability perspective, it is used considered the World Happiness Report from 2020 and its respective dataset, which evaluates social happiness across different countries in a score from 0 to 10. This report is a survey published by the UN that ranks 156 countries reviewing their state of social happiness.

The following Figure 69 is a clustering representation of the openness and social happiness scores from different regions for 2020 and the respective trendline.

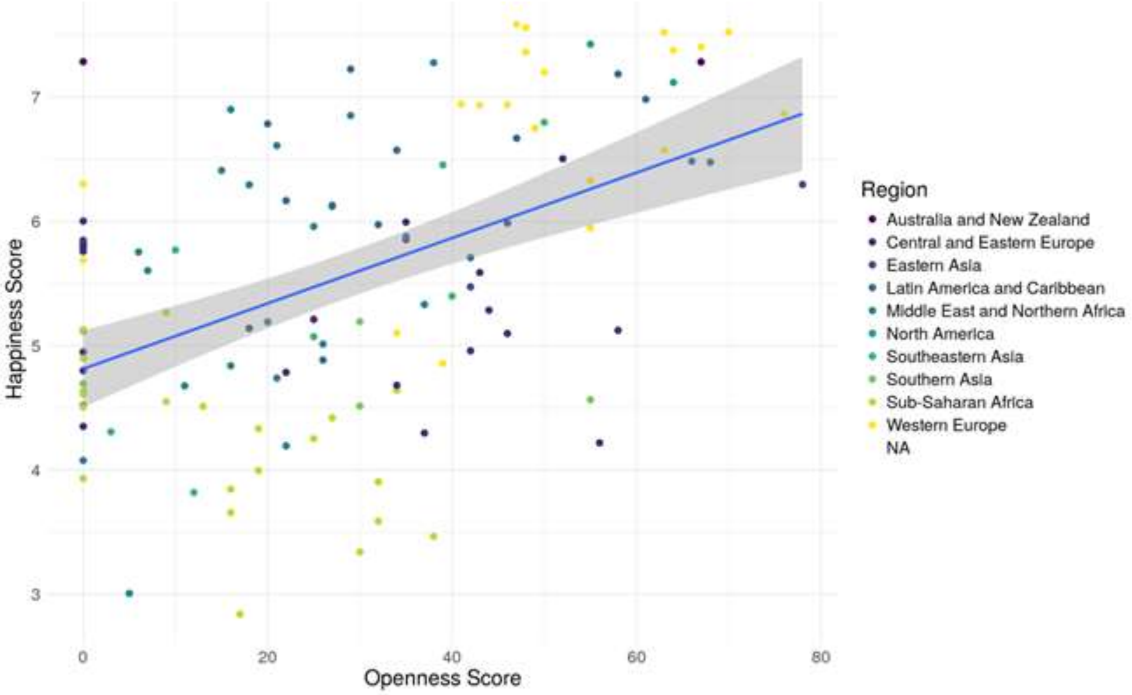


Figure 69 - Openness and Social Happiness clustering and correlation in different regions

By observing the trend line, it is possible to affirm that overall, a higher openness score is positively correlate with a higher happiness score.

From a general perspective, European and Northern American countries tend to cluster above the trend line, with higher openness and happiness scores, while African and Asian countries tend to group on the bottom left, which means a low openness and happiness scores. The Southeastern Asia is an interesting region for having an apparent diversity of clusters, some on the bottom left and others on the top right corner.

To better understand how different groups of countries behave in this correlation, below (Figure 70) it is represented the same plot but now grouping and clustering them into three established associations: the G7 (US, UK, Canada, France, Italy, Germany and Japan), the BRICs (Brazil, Russia, India and China) and the southeastern ASEAN (Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam).

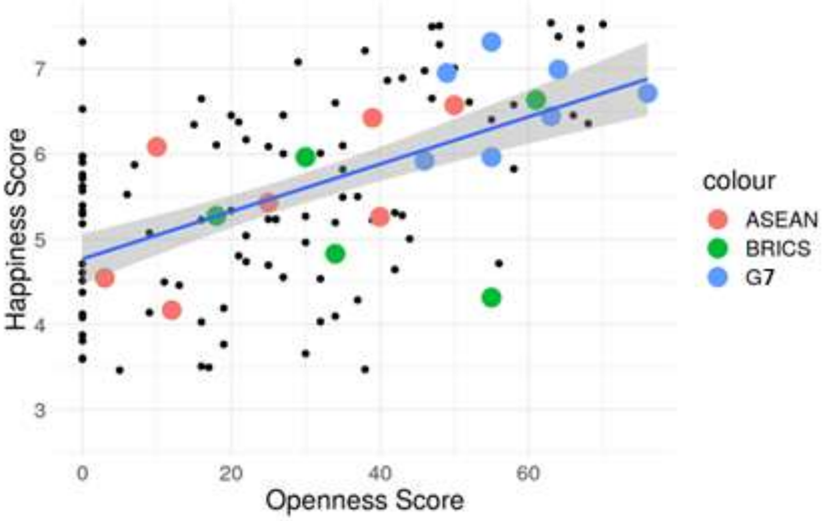


Figure 70 - Openness and Social Happiness clustering and correlation in the G7, ASEAN and BRIC countries

This confirms the previous point that G7 countries have much higher openness in their data policies, as well as bigger indices of social happiness than the BRIC and ASEAN countries, which places G7 clusters in the upper-right corner of the plot. The BRICs have a somewhat contradictory behavior since the cluster with the second highest openness score is also the one with the lowest happiness score, while the second lowest in openness is the second highest in happiness. Finally, as referred previously, the ASEAN countries can have clusters in the bottom-left corner, as well as clusters closer to the upper-right corner. Similarly to the G7, this group closely matches the trend line, which means that countries in this group with high openness also tend to have high social happiness.

Since "Happiness" is a broad and relative concept it is also interesting to visualize how Openness correlates with the social variables considered in the dataset to attribute the final Happiness Score. Those variables are: GDP, Family, Health, Freedom, Trust and Generosity.

It is possible to see how those variables correlate to each other in the following matrix, with a heatmap representing a numeral scale between -1 and 1, being 1 the maximum positive correlation, -1 the maximum negative correlation and 0 a neutral correlation (Figure 71).

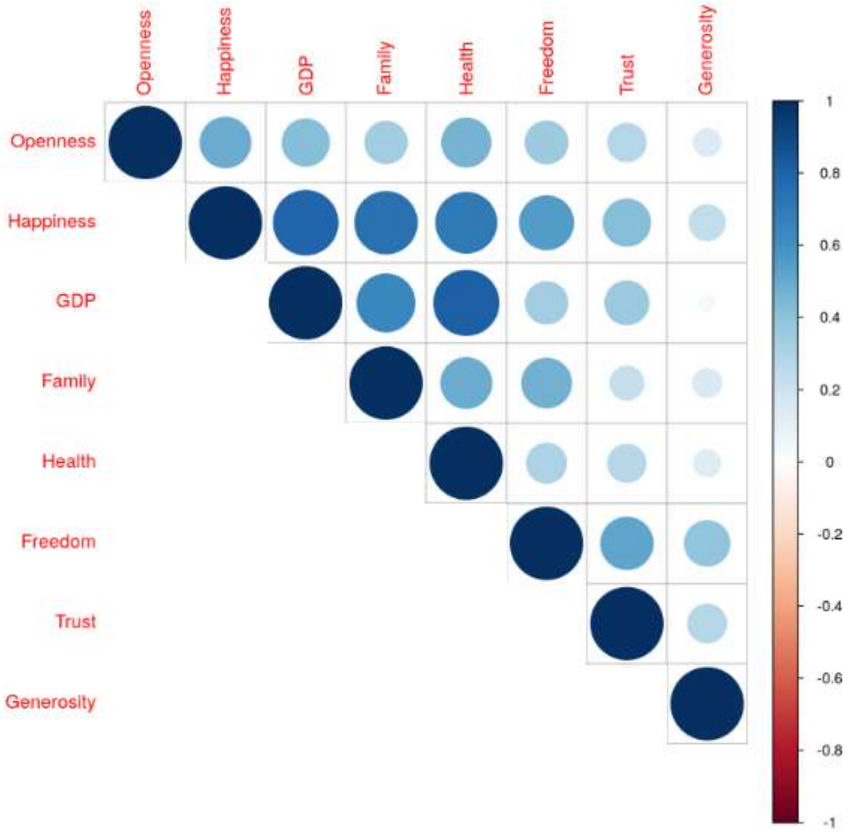


Figure 71 - Openness and SOcial happiness correlation matrix

The variables that openness influences positively the most are Happiness, GDP, Health and Freedom. The variables that correlate the most with Happiness are GDP, Family, Health and Freedom. Both Openness and Happiness correlate the less with Generosity. Other variables that correlate highly with each other are Family, Health and Freedom with GDP, Health and Freedom with Family, and Trust with Freedom. All those correlations variables do have a meaningful impact on Social Sustainability and are positively influenced by Openness, for the majority of the global countries.

This analysis concludes the Results and Critical Analysis chapter and gives support to the research conclusions in the following chapter.

5.4 Key Findings Summary

This section concludes the study, summarizing the key research findings in relation to the research aims and themes.

The findings of the study are summarized below, in Table 29, for each key subject of the Research Model.

Table 29 - Conclusions and key findings for each research theme

Research Theme	Key Findings
Open Data for Industry 4.0	Manufacturing is an industry with decreasing value added to GDP for G7 and BRIC countries
	China is one of the countries with highest manufacturing value added to its GDP
	Europe and North America are the regions with highest Industry 4.0 adoption in terms of Smart Cities
	Smart Economy and Smart Living scores have the highest correlation with Smart Cities overall scores
	The United States is the country with the highest R&D efforts for innovation, in terms of expenditure as percentage of GDP and in number of companies, while Japan has the highest share of patents in the world
	Computer Electronics, Pharmaceuticals and Transport equipment are the industries with highest R&D expenditure
	Computer Electronics is the leading R&D industry in terms of R&D expenditure and share of emitted patents
	SME's value added to Industry has been declining in both developed and developing countries, while High-tech value to industry has been increasing
	Industries and companies that adopt high-tech technologies characteristic from Industry 4.0 have competitive advantages compared to SMEs that struggle to do the same
Open Data for Sustainability	China is the country that invests the most in partnerships with developing countries
	Africa is the region that receives the highest amounts of funding from collaborating partners
	The US is the country with the highest number of companies that comply with minimum and advanced sustainability requirements
	There is a general skill migration from traditional industries to technological and specialized ones, mainly on the G7 countries
	Social Sustainability is being neglected by countries and companies in comparison with the Economic and Environmental spectrums

	The United States and China are the biggest polluters in absolute CO_2 emissions and Qatar, Saudi Arabia and United Arab Emirates are the biggest polluters in CO_2 emissions per capita
	Electricity and Heat generation are the economic sectors that are responsible for the majority of CO_2 emissions around the globe
Open Design for Sustainable Development	Data Openness has been increasing across both low and high-income countries
	High income countries still have much higher openness in comparison to low-income ones
	Openness in International Trade, Balance of Payments and Central Government Finance are important categories for Economic Sustainability, Population and Vital Statistics in Social Sustainability and Pollution and Energy Use in Environmental Sustainability
	Higher openness scores are usually positively correlate with a higher social happiness score
	The G7 countries have high openness and social happiness scores, the ASEAN countries the low openness and social happiness scores and the BRIC countries have mixed openness and social happiness scores
	The studied variables that openness influences positively the most are Happiness, GDP, Health and Freedom
	The studied variables that correlate the most with Happiness are GDP, Family, Health and Freedom
	Data Openness can be considered a positive factor for Social Sustainability

CONCLUSIONS

6.1 Conclusions and Contributions

6.2 Limitations and future lines of Investigation

6 CONCLUSION

The Conclusions chapter discusses the main conclusions and contributions of the results and critical analysis of the dissertation, as well as the limitations throughout the research and provides future lines of investigation about the themes aborded in the dissertation.

6.1 Conclusions

One of the most significant trends in society is the sustainability concern and the openness of data should support sustainability awareness and mechanisms within Industry 4.0.

Considering this and the initial objectives of the research, it is possible to conclude that while Industry 4.0 adoption is still its initial stage, there is a positive trend in broad adoption. The same can be said about Sustainability awareness as a whole, even though there is still some negligence of the social aspect.

In terms of geographic exposure, the regions that seem to be adopting Industry 4.0 successfully and implementing sustainable practices are the US, China, G7 and developed countries. The industries that seem to be exploiting technology the most are computer electronics, pharmaceuticals, and other technologic sectors. In terms of enterprise size, bigger corporations still have much more resources and capacity to adopt technology faster and with more efficiency. On the other hand, SMEs have many growth constraints mainly by inability to invest as much in technology as big corporations.

In terms of openness of data, developed countries have much more openness of that currently. However, data openness is growing faster in developing countries. Either way, there is still room for increasing transparency and collaboration through increasing openness globally.

Finally, by evaluating the results of open data for sustainable development, it was possible to conclude that openness can be considered positive for Social Sustainability, mainly in G7 and ASEAN countries, regions that showed high correlation between openness of data and social happiness.

6.2 Limitations and future lines of Investigation

While the main objectives of the research were accomplished, there are still limitations to the developed work. The purpose of identifying those limitations is not to undermine the research but to show the reader what difficulties can be encountered by researchers on similar works and to point out how future lines of investigation can improve this dissertation's results.

As it was referred on subchapter 3.2., the limitations encountered in the research methodology were the different time horizons between datasets and the lack of available data about relevant research themes.

In terms of the research analysis, the fact that some datasets have uncomplete or omitted data for a determined country or year is also a limitation that damages the best output possible.

In this research, the focused groups tended to be countries that share similar characteristics, such as stage of development. That grouping worked because it nearly impossible to find, model, analyze and compare data of all countries on earth with each other. *Ideally* the used groups should be as complete as possible but for this research was considered that adding all countries would contribute little to the overall objective of the study.

The same limitation applies other variables used, such as industries, since it is impractical to list and analyze all industries in existence.

Since the subjects approached in this research are recent in existence and fast-evolving, there are many interesting topics that future lines of research can build upon this work. As it was discussed, an Open Design approach to intellectual property is still in its early stages of development and adoption, and while there are already some useful open technologies, a good roadmap and plan for broad adoption is still lacking.

Another important topic are SMEs. Data shows that those enterprises are losing importance for economies around the world in comparison to big high-tech enterprises. Studying how SMEs can be supported and partnered with bigger companies is an interesting investigation path.

Finally, while it was as big focus in this research, it is clear that Social Sustainability should get more attention from researchers in future works and other methods beyond data science and data openness can and should be leveraged to promote this cause.

6.3 Contributions

The main contributions from this work to research were:

1. Two Conference Papers (Table 30) - Attachments A and B - submitted and accepted for publication *Procedia Computer Science*.

A) *"Data Science for Industry 4.0: A Literature Review on Open Design Approach"* within the *International Conference iSCSi 2022*, in press.

B) *"An Analysis of Open Data Scoring System towards Data Science for Sustainability in Industry 4.0"* within the International Conference *CENTERIS – International Conference on ENTERprise Information Systems 2022, accepted.*

2. One Journal Paper (Table 30) – Attachment C– *"Data Science for Industry 4.0 and Sustainability: A Survey and Analysis based on Open Data"* – at *Computers*, submitted.
3. Reviewed current available literature regarding highly relevant themes for Engineering such as Data Science, Industry 4.0, Open Data and Sustainability.
4. Evaluated trends across the identified themes for Industry 4.0, Sustainability and Open Data.
5. Provided an Open Design Approach for Industry 4.0 and Sustainability supported by the principles of Collaboration, Open Data and Open Source Tools.
6. Introduced and analyzed the correlation between Data Openness and the often neglected pillar of Social Sustainability

Table 30 - Scientific Papers published throughout the research period

Paper	Status	Conference	Journal
"Data Science for Industry 4.0: A Literature Review on Open Design Approach"	In press (Attachment A)	*International Conference iSCSi 2022*	*Procedia Computer Science*
"An Analysis of Open Data Scoring System towards Data Science for Sustainability in Industry 4.0"	Accepted (Attachment B)	*CENTERIS – International Conference on ENTERprise Information Systems 2022*	*Procedia Computer Science*
"Data Science for Industry 4.0 and Sustainability: A Survey and Analysis based on Open Data"	Reviewing (Attachment C)	-	*Computers*

REFERENCES AND OTHER SOURCES OF INFORMATION

7 REFERENCES AND OTHER SOURCES OF INFORMATION

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00014

Almirall, E., Lee, M., & Majchrzak, A. (2014). Open innovation requires integrated competition-community ecosystems: Lessons learned from civic open innovation. *Business Horizons*, *57*(3). https://doi.org/10.1016/j.bushor.2013.12.009

Angelidou, M. (2014). Smart city policies: A spatial approach. *Cities*, *41*, S3–S11. https://doi.org/10.1016/j.cities.2014.06.007

Azevedo, S. G., Pimentel, C. M. O., Alves, A. C., & Matias, J. C. O. (2021). Support of Advanced Technologies in Supply Chain Processes and Sustainability Impact. *Applied Sciences*, *11*(7). https://doi.org/10.3390/app11073026

Bai, C., Dallasega, P., Orzes, G., & Sarkis, J. (2020). Industry 4.0 technologies assessment: A sustainability perspective. *International Journal of Production Economics*, *229*. https://doi.org/10.1016/j.ijpe.2020.107776

Bamhdi, A. (2021). Requirements capture and comparative analysis of open source versus proprietary service oriented architecture. *Computer Standards & Interfaces*, *74*. https://doi.org/10.1016/j.csi.2020.103468

Benotsmane, R., Kovács, & Dudás, L. (2019a). Economic, Social Impacts and Operation of Smart Factories in Industry 4.0 Focusing on Simulation and Artificial Intelligence of Collaborating Robots. *Social Sciences*, *8*, 143. https://doi.org/10.3390/socsci8050143

Benotsmane, R., Kovács, G., & Dudás, L. (2019b). Economic, Social Impacts and Operation of Smart Factories in Industry 4.0 Focusing on Simulation and Artificial Intelligence of Collaborating Robots. *Social Sciences*, *8*(5). https://doi.org/10.3390/socsci8050143

Beruvides, G. (2019). *Modeling Techniques for Micromachining Processes*. https://doi.org/10.1007/978-3-030-03949-3_2

Betti, F., Bezamat, F., Fendri, M., Fernadez, B., Küpper, D., & Okur, A. (2020). Share to Gain: Unlocking Data Value in Manufacturing. *Http://Www3. Weforum. Org/Docs/WEF_Share_to_Gain_Report. Pdf [Stand: 27.04. 2020]*.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, *30*(3), 500–521. https://doi.org/https://doi.org/10.1016/j.aei.2016.07.001

Bin, S., & Dowlatabadi, H. (2005). Consumer lifestyle approach to US energy use and the related CO2 emissions. *Energy Policy*, *33*(2), 197–208. https://doi.org/10.1016/S0301-4215(03)00210-6

Biswas, A., & Roy, M. (2015). Leveraging factors for sustained green consumption behavior based on consumption value perceptions: testing the structural model. *Journal of Cleaner Production*, *95*. https://doi.org/10.1016/j.jclepro.2015.02.042

Bonilla, S., Silva, H., Terra da Silva, M., Franco Gonçalves, R., & Sacomano, J. (2018). Industry 4.0 and Sustainability Implications: A Scenario-Based Analysis of the Impacts and Challenges. *Sustainability*, *10*(10). https://doi.org/10.3390/su10103740

Boulanger, A. (2005). Open-source versus proprietary software: Is one more reliable and secure than the other? *IBM Systems Journal*, *44*(2). https://doi.org/10.1147/sj.442.0239

Boyes, H., Hallaq, B., Cunningham, J., & Watson, T. (2018). The industrial internet of things (IIoT): An analysis framework. *Computers in Industry*, *101*. https://doi.org/10.1016/j.compind.2018.04.015

Braccini, A., & Margherita, E. (2018). Exploring Organizational Sustainability of Industry 4.0 under the Triple Bottom Line: The Case of a Manufacturing Company. *Sustainability*, *11*(1). https://doi.org/10.3390/su11010036

Brenner, B., & Hartl, B. (2021). The perceived relationship between digitalization and ecological, economic, and social sustainability. *Journal of Cleaner Production*, *315*. https://doi.org/10.1016/j.jclepro.2021.128128

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011a). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1819486

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011b). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1819486

Bughin, J. (2018, October 12). *Marrying artificial intelligence and the sustainable development goals: The global economic impact of AI*. https://www.mckinsey.com/mgi/overview/in-the-news/marrying-artificial-intelligence-and-the-sustainable

Bulao, J. (2021, December 7). *How Much Data Is Created Every Day in 2021?* TechJury. http://techjury.net/blog/how-much-data-is-created-every-day/

Castro, H., Pinto, N., Pereira, F., Ferreira, L., Ávila, P., Bastos, J., Putnik, G. D., & Cruz-Cunha, M. (2021). Cyber-Physical Systems using Open Design: An approach towards an Open Science Lab for Manufacturing. *Procedia Computer Science*, *196*(2021), 381–388. https://doi.org/10.1016/j.procs.2021.12.027

Castro, H., Pinto, N., Pereira, F., Ferreira, L., Avila, P., Putnik, G., Felgueiras, C., Bastos, J., & Cunha, M. (2021). Open Science Laboratory for Manufacturing: An education tool to contribute to sustainability. *ACM International Conference Proceeding Series*, 819–823. https://doi.org/10.1145/3486011.3486564

Castro, H., Putnik, G., Castro, A., & Bosco Fontana, R. D. (2019). Open Design initiatives: an evaluation of CAD Open Source Software. *Procedia CIRP*, *84*, 1116–1119. https://doi.org/10.1016/j.procir.2019.08.001

Castro, H., Putnik, G., Castro, A., & Fontana, R. D. B. (2019). Could Open Design learn from Wikipedia? *Procedia CIRP*, *84*, 1112–1115. https://doi.org/10.1016/j.procir.2019.07.001

Caulkins, J. P., Feichtinger, G., Grass, D., Hartl, R. F., Kort, P. M., & Seidl, A. (2013). When to make proprietary software open source. *Journal of Economic Dynamics and Control*, *37*(6). https://doi.org/10.1016/j.jedc.2013.02.009

Cavanillas, J. M., Curry, E., & Wahlster, W. (Eds.). (2016). *New Horizons for a Data-Driven Economy*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3

Chabbouh, H., & Boujelbene, Y. (2020). Open innovation in SMEs: The mediating role between human capital and firm performance. *The Journal of High Technology Management Research*, *31*(2), 100391. https://doi.org/https://doi.org/10.1016/j.hitech.2020.100391

Chakraborty, D., & Helling, R. K. (2021). Industry sustainable supply chain management with data science. In *Data Science Applied to Sustainability Analysis*. Elsevier. https://doi.org/10.1016/B978-0-12-817976-5.00010-3

Chesbrough, H., & Crowther, A. K. (2006). Beyond high tech: early adopters of open innovation in other industries. *R&D Management*, *36*(3), 229–236. https://doi.org/https://doi.org/10.1111/j.1467-9310.2006.00428.x

Chesbrough, H., Vanhaverbeke, W., & West, J. (2008). *Open Innovation: Researching A New Paradigm*.

Chesbrough, H. W. (2003). The Era of Open Innovation. *MIT Sloan Management Review*.

Das, S., Sismanis, Y., Beyer, K. S., Gemulla, R., Haas, P. J., & McPherson, J. (2010, June 6). Ricardo: integrating R and Hadoop. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/1807167.1807275

Dash, P. B., Naik, B., Nayak, J., & Vimal, S. (2022). Socio-economic factor analysis for sustainable and smart precision agriculture: An ensemble learning approach. *Computer Communications*, *182*. https://doi.org/10.1016/j.comcom.2021.11.002

Denyer, D., Tranfield, D., & van Aken, J. E. (2008). Developing Design Propositions through Research Synthesis. *Organization Studies*, *29*(3). https://doi.org/10.1177/0170840607088020

Dong, K., Jiang, H., Sun, R., & Dong, X. (2019). Driving forces and mitigation potential of global CO2 emissions from 1980 through 2030: Evidence from countries with different income levels. *Science of The Total Environment*, *649*, 335–343. https://doi.org/10.1016/j.scitotenv.2018.08.326

Enyoghasi, C., & Badurdeen, F. (2021). Industry 4.0 for sustainable manufacturing: Opportunities at the product, process, and system levels. *Resources, Conservation and Recycling*, *166*. https://doi.org/10.1016/j.resconrec.2020.105362

Epstein, M. J., Elkington, J., & Leonard, H. B. "Dutch." (2018). *Making Sustainability Work*. Routledge. https://doi.org/10.4324/9781351280129

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, *20*(15). https://doi.org/10.1093/bioinformatics/bth261

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2). https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Ghobakhloo, M. (2020). Industry 4.0, digitization, and opportunities for sustainability. *Journal of Cleaner Production*, *252*. https://doi.org/10.1016/j.jclepro.2019.119869

Ghobakhloo, M., & Fathi, M. (2019). Corporate survival in Industry 4.0 era: the enabling role of lean-digitized manufacturing. *Journal of Manufacturing Technology Management*, *31*(1). https://doi.org/10.1108/JMTM-11-2018-0417

Gökalp, M. O., Gökalp, E., Kayabay, K., Koçyiğit, A., & Eren, P. E. (2021). Data-driven manufacturing: An assessment model for data science maturity. *Journal of Manufacturing Systems*, *60*, 527–546. https://doi.org/https://doi.org/10.1016/j.jmsy.2021.07.011

Hall, T. (2020, May 20). *The Role of Data in Industry 4.0*. https://industrytoday.com/the-role-of-data-in-industry-4-0/

Han, H., & Trimi, S. (2022). Towards a data science platform for improving SME collaboration through Industry 4.0 technologies. *Technological Forecasting and Social Change*, *174*, 121242. https://doi.org/https://doi.org/10.1016/j.techfore.2021.121242

Hermann, M., Pentek, T., & Otto, B. (2015). *Design Principles for Industrie 4.0 Scenarios: A Literature Review*. https://doi.org/10.13140/RG.2.2.29269.22248

Hevner, March, Park, & Ram. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625

Hickin, R., Bechtel, M., Golem, A., Erb, L., & Buscalno, R. (2021). *Technology Futures: Projecting the Possible, Navigating What's Next*. https://www3.weforum.org/docs/WEF_Technology_Futures_GTGS_2021.pdf

International Monetary Fund. (2018). World Economic Outlook: Challenges to Steady Growth. *World Economic and Finantial Surveys*.

Inyang, B. J. (2013). Defining the Role Engagement of Small and Medium-Sized Enterprises (SMEs) in Corporate Social Responsibility (CSR). *International Business Research*, *6*(5). https://doi.org/10.5539/ibr.v6n5p123

Jain, A. K., Dhada, M., Parlikad, A. K., & Lad, B. K. (2020). Product Quality Driven Auto-Prognostics: Low-Cost Digital Solution for SMEs. *IFAC-PapersOnLine*, *53*(3), 78–83. https://doi.org/https://doi.org/10.1016/j.ifacol.2020.11.012

Kamble, S. S., Gunasekaran, A., & Gawankar, S. A. (2018). Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. *Process Safety and Environmental Protection*, *117*. https://doi.org/10.1016/j.psep.2018.05.009

Kerin, M., & Pham, D. T. (2019). A review of emerging industry 4.0 technologies in remanufacturing. *Journal of Cleaner Production*, *237*. https://doi.org/10.1016/j.jclepro.2019.117805

Kerr, S. P., Kerr, W., Özden, Ç., & Parsons, C. (2016). *High-Skilled Migration and Agglomeration*. https://doi.org/10.3386/w22926

Kilamo, T., Hammouda, I., Mikkonen, T., & Aaltonen, T. (2012). From proprietary to open source—Growing an open source ecosystem. *Journal of Systems and Software*, *85*(7). https://doi.org/10.1016/j.jss.2011.06.071

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, *6*(4). https://doi.org/10.1007/s12599-014-0334-4

Lee, J. (2003). E-manufacturing—fundamental, tools, and transformation. *Robotics and Computer-Integrated Manufacturing*, *19*(6). https://doi.org/10.1016/S0736-5845(03)00060-7

Lee, J., Ardakani, H. D., Yang, S., & Bagheri, B. (2015). Industrial Big Data Analytics and Cyber-physical Systems for Future Maintenance &amp; Service Innovation. *Procedia CIRP*, *38*. https://doi.org/10.1016/j.procir.2015.08.026

Lee, J. S., Pries-Heje, J., & Baskerville, R. (2011). *Theorizing in Design Science Research* (pp. 1–16). https://doi.org/10.1007/978-3-642-20633-7_1

Lee, S., Ju, E., Choi, S., Lee, H., Shim, J., Chang, K., Kim, K., & Kim, C. (2018). *Prediction of Cancer Patient Outcomes Based on Artificial Intelligence*. https://doi.org/10.5772/intechopen.81872

Lemenkova, P. (2019). PROCESSING OCEANOGRAPHIC DATA BY PYTHON LIBRARIES NUMPY, SCIPY AND PANDAS. *Aquatic Research*. https://doi.org/10.3153/AR19009

Liao, Y., Deschamps, F., Loures, E. de F. R., & Ramos, L. F. P. (2017). Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal. *International Journal of Production Research*, *55*(12). https://doi.org/10.1080/00207543.2017.1308576

Lin, K., Shyu, J., & Ding, K. (2017). A Cross-Strait Comparison of Innovation Policy under Industry 4.0 and Sustainability Development Transition. *Sustainability*, *9*(5). https://doi.org/10.3390/su9050786

Lou, X., van der Lee, S., & Lloyd, S. (2013). AIMBAT: A Python/Matplotlib Tool for Measuring Teleseismic Arrival Times. *Seismological Research Letters*, *84*(1). https://doi.org/10.1785/0220120033

Luthra, S., & Mangla, S. K. (2018). Evaluating challenges to Industry 4.0 initiatives for supply chain sustainability in emerging economies. *Process Safety and Environmental Protection*, *117*. https://doi.org/10.1016/j.psep.2018.04.018

Machado, C. G., Winroth, M. P., & Ribeiro da Silva, E. H. D. (2020). Sustainable manufacturing in Industry 4.0: an emerging research agenda. *International Journal of Production Research*, *58*(5). https://doi.org/10.1080/00207543.2019.1652777

Mansfield, E., & Lee, J.-Y. (1996). The modern university: contributor to industrial innovation and recipient of industrial R&amp;D support. *Research Policy*, *25*(7), 1047–1058. https://doi.org/10.1016/S0048-7333(96)00893-1

Maresova, P., Soukal, I., Svobodova, L., Hedvicakova, M., Javanmardi, E., Selamat, A., & Krejcar, O. (2018). Consequences of Industry 4.0 in Business and Economics. *Economies*, *6*(3). https://doi.org/10.3390/economies6030046

McGoldrick, P. J., & Freestone, O. M. (2008). Ethical product premiums: antecedents and extent of consumers' willingness to pay. *The International Review of Retail, Distribution and Consumer Research*, *18*(2). https://doi.org/10.1080/09593960701868431

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D. B., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R., Zaharia, M., & Talwalkar, A. (2016). MLlib: Machine Learning in Apache Spark. *J. Mach. Learn. Res.*, *17*(1), 1235–1241.

Merritt, J., Antunes, M., & Tanaka, Y. (2021). Governing Smart Cities: Policy Benchmarks for Ethical and Responsible Smart City Development. *World Economic Forum*. https://www3.weforum.org/docs/WEF_Governing_Smart_Cities_2021.pdf

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment. *British Journal of Management*, *30*(2), 272–298. https://doi.org/https://doi.org/10.1111/1467-8551.12343

Mittal, S., Khan, M. A., Romero, D., & Wuest, T. (2018). A critical review of smart manufacturing &amp; Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs). *Journal of Manufacturing Systems*, *49*. https://doi.org/10.1016/j.jmsy.2018.10.005

Moon, J., Gbadago, D. Q., Hwang, G., Lee, D., & Hwang, S. (2021). Software platform for high-fidelity-data-based artificial neural network modeling and process optimization in chemical engineering. *Computers & Chemical Engineering*. https://doi.org/10.1016/j.compchemeng.2021.107637

Morrar, R., Arman, H., & Mousa, S. (2017). The Fourth Industrial Revolution (Industry 4.0): A Social Innovation Perspective. *Technology Innovation Management Review*, *7*, 12–20. https://doi.org/http://doi.org/10.22215/timreview/1117

Müller, J. M., Buliga, O., & Voigt, K.-I. (2018). Fortune favors the prepared: How SMEs approach business model innovations in Industry 4.0. *Technological Forecasting and Social Change*, *132*. https://doi.org/10.1016/j.techfore.2017.12.019

Müller, J. M., Kiel, D., & Voigt, K.-I. (2018). What Drives the Implementation of Industry 4.0? The Role of Opportunities and Challenges in the Context of Sustainability. *Sustainability*, *10*(1). https://doi.org/10.3390/su10010247

Müller, J. M., & Voigt, K.-I. (2018). Sustainable Industrial Value Creation in SMEs: A Comparison between Industry 4.0 and Made in China 2025. *International Journal of Precision Engineering and Manufacturing-Green Technology*, *5*(5). https://doi.org/10.1007/s40684-018-0056-z

Nagaraj, S. V. (2020). Disruptive technologies that are likely to shape future jobs. *Procedia Computer Science*, *172*. https://doi.org/10.1016/j.procs.2020.05.164

ODW. (2021). *Open Data Inventory 2020/21 Annual Report*. https://opendatawatch.com/publications/open-data-inventory/

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, *9*(3). https://doi.org/10.1109/MCSE.2007.58

Open Knowledge Foundation. (n.d.). *The Open Definition*. Retrieved May 30, 2022, from https://opendefinition.org/

P. Mazanetz, M., J. Marmon, R., B. T. Reisser, C., & Morao, I. (2012). Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Current Topics in Medicinal Chemistry*, *12*(18). https://doi.org/10.2174/156802612804910331

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, *20*(2). https://doi.org/10.1093/bioinformatics/btg412

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pereira, A. C., & Romero, F. (2017). A review of the meanings and the implications of the Industry 4.0 concept. *Procedia Manufacturing*, *13*. https://doi.org/10.1016/j.promfg.2017.09.032

Pereira, M. T., Silva, A., Ferreira, L. P., Sá, J. C., & Silva, F. J. G. (2019). A DMS to Support Industrial Process Decision-Making: a contribution under Industry 4.0. *Procedia Manufacturing*, *38*, 613–620. https://doi.org/https://doi.org/10.1016/j.promfg.2020.01.079

Peruzzini, M., Grandi, F., & Pellicciari, M. (2020). Exploring the potential of Operator 4.0 interface and monitoring. *Computers & Industrial Engineering*, *139*, 105600. https://doi.org/https://doi.org/10.1016/j.cie.2018.12.047

Piccarozzi, M., Aquilani, B., & Gatti, C. (2018). Industry 4.0 in Management Studies: A Systematic Literature Review. *Sustainability*, *10*(10). https://doi.org/10.3390/su10103821

Pinheiro, P., Putnik, G. D., Castro, A., Castro, H., Fontana, R. D. B., & Romero, F. (2019). Industry 4.0 and industrial revolutions: An assessment based on complexity. *FME Transactions*, *47*(4), 831–840. https://doi.org/10.5937/fmet1904831P

Pivoto, D. G. S., de Almeida, L. F. F., da Rosa Righi, R., Rodrigues, J. J. P. C., Lugli, A. B., & Alberti, A. M. (2021). Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review. *Journal of Manufacturing Systems*, *58*, 176–192. https://doi.org/https://doi.org/10.1016/j.jmsy.2020.11.017

Provost, F., & Fawcett, T. (2013a). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Provost, F., & Fawcett, T. (2013b). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Provost, F., & Fawcett, T. (2013c). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1). https://doi.org/10.1089/big.2013.1508

Putnik, G., & Ávila, P. (2021). Manufacturing system and enterprise management for Industry 4.0: Guest editorial. *FME Transactions*, *49*(4), 769–772. https://doi.org/10.5937/fme2104769P

Qian, F., Zhong, W., & Du, W. (2017). Fundamental Theories and Key Technologies for Smart and Optimal Manufacturing in the Process Industry. *Engineering*, *3*(2). https://doi.org/10.1016/J.ENG.2017.02.011

Quintas, J., Menezes, P., & Dias, J. (2017). Information Model and Architecture Specification for Context Awareness Interaction Decision Support in Cyber-Physical Human–Machine Systems. *IEEE Transactions on Human-Machine Systems*, *47*(3). https://doi.org/10.1109/THMS.2016.2634923

R Core Team. (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/.

Ragan-Kelley, M., Perez, F., Granger, B., Kluyver, T., Ivanov, P., Frederic, J., & Bussonnier, M. (2014). The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. *AGU Fall Meeting Abstracts*, *2014*, H44D-07.

Ramadan, M., Al-Maimani, H., & Noche, B. (2017). RFID-enabled smart real-time manufacturing cost tracking system. *The International Journal of Advanced Manufacturing Technology*, *89*(1–4). https://doi.org/10.1007/s00170-016-9131-1

Roy, R., Stark, R., Tracht, K., Takata, S., & Mori, M. (2016). Continuous maintenance and the future – Foundations and technological challenges. *CIRP Annals*, *65*(2). https://doi.org/10.1016/j.cirp.2016.06.006

Runeson, P. (2019, May). Open Collaborative Data - using OSS Principles to Share Data in SW Engineering. *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. https://doi.org/10.1109/ICSE-NIER.2019.00015

Runeson, P., Olsson, T., & Linåker, J. (2021). Open Data Ecosystems — An empirical investigation into an emerging industry collaboration concept. *Journal of Systems and Software*, *182*. https://doi.org/10.1016/j.jss.2021.111088

Saritha, B., Bonagiri, R., & Deepika, R. (2021). Open source technologies in data science and big data analytics. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.01.610

Savastano, M., Amendola, C., Bellini, F., & D'Ascenzo, F. (2019). Contextual Impacts on Industrial Processes Brought by the Digital Transformation of Manufacturing: A Systematic Review. *Sustainability*, *11*(3). https://doi.org/10.3390/su11030891

Schuh, G., Potente, T., Varandani, R., & Schmitz, T. (2014). Global Footprint Design based on genetic algorithms – An “Industry 4.0” perspective. *CIRP Annals*, *63*(1). https://doi.org/10.1016/j.cirp.2014.03.121

Shet, S. v., & Pereira, V. (2021). Proposed managerial competencies for Industry 4.0 – Implications for social sustainability. *Technological Forecasting and Social Change*, *173*. https://doi.org/10.1016/j.techfore.2021.121080

Sousa, A. C., Bertachini, A. F., Cunha, C., Chaves, R., & Varela, M. L. R. (2021). Literature review and discussion on collaborative decision making approaches in Industry 4.0. *FME Transactions*, *49*(4). https://doi.org/10.5937/fme2104817S

Stock, T., & Seliger, G. (2016). Opportunities of Sustainable Manufacturing in Industry 4.0. *Procedia CIRP*, *40*. https://doi.org/10.1016/j.procir.2016.01.129

Thames, L., & Schaefer, D. (2017). *Industry 4.0: An Overview of Key Benefits, Technologies, and Challenges* (pp. 1–33). https://doi.org/10.1007/978-3-319-50660-9_1

The World Bank. (2022a). GDP (Current US$). *World Bank*.

The World Bank. (2022b). Researchers in R&D (per million people). *The World Bank*. https://databank.worldbank.org/metadataglossary/jobs/series/SP.POP.SCIE.RD.P6

Toledo, R. F. de, Farias Filho, J. R. de, Castro, H. C. G. A. de, Putnik, G. D., & Silva, L. E. da. (2021). Is the incorporation of sustainability issues and Sustainable Development Goals in project management a catalyst for sustainable project delivery? *International Journal of Sustainable Development and World Ecology*, *28*(8), 733–743. https://doi.org/10.1080/13504509.2021.1888816

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, *14*(3). https://doi.org/10.1111/1467-8551.00375

UN. (2022a). Goal 9 - Industry, Innovation and Infrastructure. *Sustainable Development Goals*.

UN. (2022b). Goal 12 - Responsible Comsumption and Production. *Sustainable Development Goals*.

UN. (2022c). Goal 17 - Partnerships for the Goals. *Sustainable Development Goals*.

United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. *Department of Economic and Social Affairs and Sustainable Development*. https://sdgs.un.org/2030agenda

Vaidya, S., Ambad, P., & Bhosle, S. (2018). Industry 4.0 – A Glimpse. *Procedia Manufacturing*, *20*, 233–238. https://doi.org/https://doi.org/10.1016/j.promfg.2018.02.034

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2). https://doi.org/10.1007/s11192-009-0146-3 van Ham, H., & Koppenjan, J. (2001). BUILDING PUBLIC-PRIVATE PARTNERSHIPS: Assessing and managing risks in port development. *Public Management Review*, *3*(4), 593–616. https://doi.org/10.1080/14616670110070622

Varela, L., Araújo, A., Ávila, P., Castro, H., & Putnik, G. (2019). Evaluation of the Relation between Lean Manufacturing, Industry 4.0, and Sustainability. *Sustainability*, *11*(5), 1439. https://doi.org/10.3390/su11051439

Varela, L., Ávila, P., Castro, H., Putnik, G. D., Fonseca, L. M. C., & Ferreira, L. (2022). Manufacturing and Management Paradigms, Methods and Tools for Sustainable Industry 4.0-Oriented Manufacturing Systems. *Sustainability*, *14*(3), 1574. https://doi.org/10.3390/su14031574

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, *25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Wang, B., Wu, C., Huang, L., & Kang, L. (2019). Using data-driven safety decision-making to realize smart safety management in the era of big data: A theoretical perspective on basic questions and their answers. *Journal of Cleaner Production*, *210*, 1595–1604. https://doi.org/https://doi.org/10.1016/j.jclepro.2018.11.181

Wang, S., Wan, J., Li, D., & Zhang, C. (2016). Implementing Smart Factory of Industrie 4.0: An Outlook. *International Journal of Distributed Sensor Networks*, *12*(1). https://doi.org/10.1155/2016/3159805

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60). https://doi.org/10.21105/joss.03021

Wee, D., Kelly, R., Cattel, J., & Breuning, M. (2015). Industry 4.0-how to navigate digitization of the manufacturing sector. *Mckinsey & Company*, *58*.

WEF. (2020). The Future of Jobs Report 2020. *WEF*.

WEF, & BCG. (2021). *Net-Zero Challenge: The supply chain opportunity*.

Witkowski, K. (2017). Internet of Things, Big Data, Industry 4.0 – Innovative Solutions in Logistics and Supply Chains Management. *Procedia Engineering*, *182*. https://doi.org/10.1016/j.proeng.2017.03.197

World Bank. (2022a). *Industry (including construction), value added (% of GDP)*. World Bank.

World Bank. (2022b). Skills | LinkedIn Data. *Data Catalog*.

Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research*, *39*(1). https://doi.org/10.1177/0739456X17723971

Xu, Z., Zhu, Y., Hu, Y., Huang, M., Xu, F., & Wang, J. (2021). Bibliometric and visualized analysis of Neuropathic pain based on Web of Science and CiteSpace over the last 20 years. *World Neurosurgery*. https://doi.org/10.1016/j.wneu.2021.12.025

ATTACHMENTS

8 ATTACHMENTS

8.1 Attachment A – Paper *"Data Science for Industry 4.0: A Literature Review on Open Design Approach"*

International Conference on Industry Sciences and Computer Science Innovation

Data Science for Industry 4.0: A Literature Review on Open Design Approach

Hélio Castro[a, b]1, Filipe Costa[a], Luís Ferreira[c], Paulo Ávila[a, b], Goran D. Putnik[d], Manuela Cruz-Cunha[c]

[a]School of Engineering, Polytechnic Institute of Porto, Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal
[b]INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[c]2Ai - Applied Artificial Intelligence Lab, School of Technology, IPCA, Barcelos, Portugal
[d]University of Minho, School of Engineering, Campus de Azurém, 4800-058 Guimarães, Portugal

Abstract

Data Science is a tool for organizations to accelerate the development of innovative solutions through collaboration among Small and Medium Enterprises (SMEs), leading to the Industry 4.0 approach. Data is essential in decision making. Many organizations can't access timely relevant information because they don't own it or there is a lack of collaboration among third parties. Often, data-sharing and collaborative approaches can benefit both, increase the market they operate in, and accelerate innovation. This paper identifies and analyzes the current utility of Data Science for Industry 4.0 from an Open Design perspective to accelerate innovation across different industries.

Peer-review under responsibility of the scientific committee of the International Conference on Industry Sciences and Computer Sciences Innovation

Keywords: Data Science; Industry 4.0; Open-Data; Collaboration; Innovation; Sustainability

1. Introduction

In the last few years, the manufacturing, scientific and technologic fields have been subject to a revolution process of digitalization and technologic development called Industry 4.0 (Liao et al., 2017).

* Corresponding author. Tel.: +351 +351 22 83 40 500; fax: +351 22 83 21 159.
E-mail address: hcc@isep.ipp.pt

This process is implementing changes that stimulate more competitive practices across many economic sectors. These changes are in great part supported by the growing acquisition and utilization of information and data, that can be exploited through big data technologies and data science. Even though data is more accessible than ever before, the overwhelming majority of data is concentrated and centralized in private companies, organizations, or institutions and inaccessible for scientific and academic research. This means that there is a wide range of limited solutions for economic, social, and environmental challenges that can only be solved by those who own the data. The same can be said for the tools necessary to explore that data. Most data science platforms and tools developed in the past are proprietary and costly, which means that they are inaccessible for small businesses, individuals, and scientists that can't pay for the licenses for that software. Another limitation for that proprietary approach is that by being closed source, the development of those tools is limited by the developers of the organization that owns them, limiting the possible opportunities of collaborating with other developers to improve the tool itself. This paper presents a literature review of Data Science for Industry 4.0, and how *Open Design* approaches compare to existing alternatives in industry and engineering.

2. Research Methodology

The used method to extract information about the study subject was the *Systematic Literature Review* (SLR) which is a process that enables researchers to answer a formulated question (Xiao & Watson, 2019) by adopting a replicable, scientific, and transparent process that differs from traditional narrative reviews (Tranfield et al., 2003).

The triage process was made by a sequence of 4 steps:

1. Since the research subjects are recent and fast pacing evolving themes, the period defined for the information sources is between the year 2000 and October 2021.
2. The combination of keywords defined for the research were: “Industry 4.0 + Sustainability + Innovation”; “Data Science + Sustainability + Innovation”; “Industry 4.0 + Engineering + Innovation” and “Data Science + Engineering + Innovation”.
3. The sample that resulted from the research criteria contained 862 available publications from a total number of 1897, that were collected and used in the next step.
4. The publications that resulted from the previous step were complemented by publications obtained from other databases such as *ScienceDirect* and *b-on*. The keywords used for the research in those databases were the same that were used in *WoS*.

The selected publications were imported into the VOSviewer platform (version 1.6.17), which compiled a network of themes into a bibliometric map containing 6 different clusters (Figure 1). Each cluster represents the interception of relevant scientific themes, according to the selected publications, resulting in a total of 87 relevant themes in the research.

Among the most cited themes are highlighted themes such as “Industry 4.0”, “Innovation”, “Sustainability”, “Big Data analytics”, “Design”, “Machine learning”, “Supply Chain” and “Smart Factory”.

Between the lesser cited themes are highlighted themes such as “Open innovation”, “Social Sustainability”, “SMEs”, “Collaboration” and “Sustainable development”, which require a deeper understanding and research in the future.

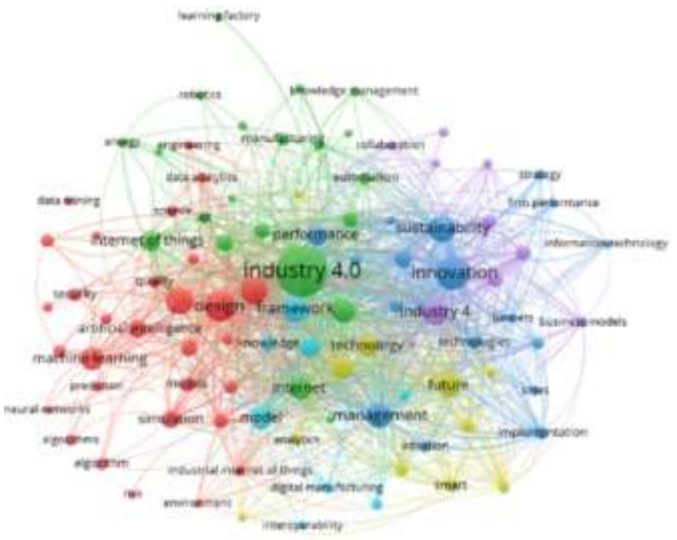


Fig. 1. Bibliometric map representing the most relevant research areas and networks correlating with the defined keywords

The bibliographic review of this research incorporates both types of themes converging into a deeper analysis of the lesser cited themes.

3. Bibliographic Review

This section presents the main results of the literature review that was structured one main subchapter: “The Role of Data in Industry 4.0”. In this subchapter are explored concepts such as Data-driven decision making, Data ownership which represent highly relevant concepts to the study.

3.1. The Role of Data in Industry 4.0

Industry 4.0 has gained increased adoption in recent years with its promise to use the power of data to revolutionize manufacturing. However, while the exploration of data has been a catalyst of business growth and efficiency gains, the manufacturing sector has been slow to adopt data-driven processes. According to Accenture, only 13% of manufacturing companies have implemented an Industry 4.0 approach (Hall, 2020). It is inevitable for data to become a cornerstone in the decision-making of not only industrial processes, but also the sustainability of economic, social, and environmental approaches. Data-driven decision-making will be essential to the future of those areas and through Data Science and Big Data Analytics it can be implemented faster and more efficiently.

3.2. Data Science and Big Data in Data-driven decision making

As we live in a world that constantly produces and consumes data, it is a priority to understand the value that can be extracted from it. (Mikalef et al., 2019) consider data science and the big data domains as the next frontier for both practitioners and researchers as they embody significant potentials in exploiting data to sustain competitive advantage.

Big data is the emerging field where innovative technology offers new ways of extracting value from new information. The ability to effectively manage information and extract knowledge is now seen as a key competitive advantage. Big data technology adoption within industrial sectors is an imperative need for most organizations to survive and gain a competitive advantage (Cavanillas et al., 2016).

Data science is an interdisciplinary field that supports and guides the extraction of useful patterns from raw data by exploring advanced technologies, algorithms, and processes (Provost & Fawcett, 2013a). The actual extraction of knowledge from data is defined as data mining, and it can be applied to a broad set of business areas such as marketing, customer relationship management, supply chain management, or product optimization (Bilal et al., 2016).

As is shown in Figure 2, there are a variety of fields that have a growing influence in decision making that correlate to each other and have a common source of information in data mining. The interception of all these fields can be represented by Data Science.

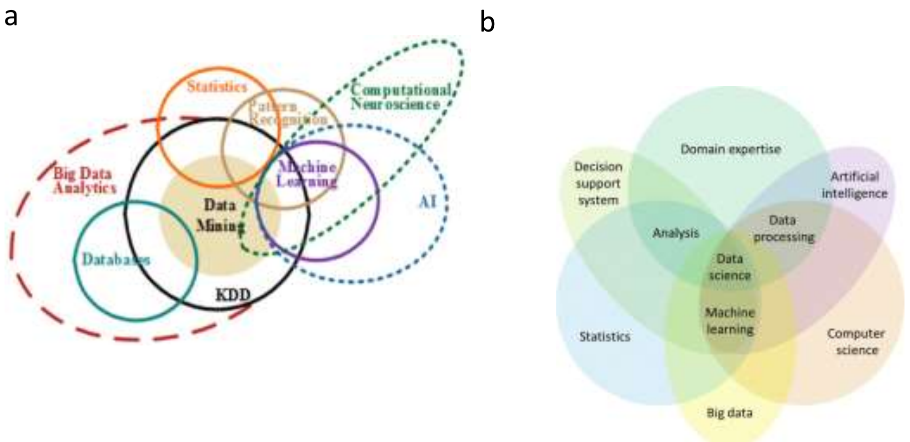


Fig. 2. (a) Interception of data fields with data mining adapted from (Bilal et al., 2016); (b) Interception of data fields with Data Science adapted from (S. Lee et al., 2018)

Even though *Data Science* and *Big Data* are closely correlated, Data Science should be seen as a domain that originates from the emergence of Big Data technologies, data management skills, and behavioral disciplines (Saritha et al., 2021). From a business perspective, the goal in leveraging data science and big data is usually improving decision making. *Data-driven decision-making* (DDDM) refers to basing decisions on the analysis of data rather than purely on intuition and experience (B. Wang et al., 2019). F. Provost and T. Fawcett (Provost & Fawcett, 2013b) represent how the automation of decision making by computer systems is supported first, by the processing of data through Big Data analytics and second, by visualizing that data through Data Science platforms (Figure 3).

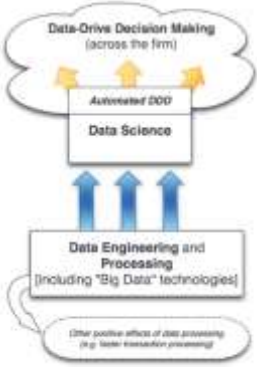

Data-Drive Decision Making
(across the firm)
Automated DDD
Data Science
Data Engineering and Processing
[including "Big Data" technologies]
(Other positive effects of data processing
(e.g. faster transaction processing))

Fig. 3. Representation of how Data Science supports data-driven decision making, adapted from [12]

[13] conducted a study of how DDDM affects firm performance. That study showed statistically that the more data-driven the firm is, the more productive it is, represented by a 4-6% increase in productivity. DDDM is also correlated with a higher return on assets, return on equity, asset utilization, and market value.

3.3. Data Sharing and Open-Source approaches

Data Science and Big Data can be combined with co-creation and data-sharing technologies for organizations to leverage creativity outside their organizational boundaries [14]. The development and operation of software have become increasingly dependent on data [15] and this data can be more accessible to organizations and individuals through data-sharing and open-source technologies. [16] highlights the need for the adoption of co-creation and collaboration principles to harness the innovation potential and to manage costs in the age of data.

For organizations, there is a steady increase in reliance on analytics that uses enabling technologies such as sensors, the Internet of Things, robotics, and ambient computing – all of which rely on huge amounts of data that stem from our many digital interactions [17].

As of 2020, 2.5 quintillion bytes of data were produced every day worldwide [18] and it is estimated that by 2025 that amount will increase nearly 200 times. It is safe to assume that as the gap between the physical and the digital narrows, the data volume of connectivity will continue to grow steadily.

Today, data volumes are exploding, and not only is the rate of data generated per individual increasing but so is the rate at which we share information. Lawmakers and organizations worldwide are trying to envision data's ownership future. Information remains largely centralized but the trend is shifting toward a distributed and open model of data sharing (Hickin et al., 2021). The same authors represent a possible transition from known technologies to future trends in which distributed approaches such as open-source, explainable AI and decentralized data ownership constitute a positive linear transition. However, if the future approaches to technologic advancements are closed source and proprietary that would mean a negative linear trend [18].

According to literature, the approach to technologic advancements and future trends of how data and software are collected, stored, managed, modified, and shared can be split into Proprietary and Open Source. Those differences are shown in Table 1 [18], [19] – [22]:

Table 1. Approach to technologic advancements and future trends of data ownership

	Closed and Proprietary	**Open Source**
Data Ownership	**Institutional**	**Decentralized**
Approach to Technologic Advancements	Monetization of data by maintaining a closed-source approach that keeps intellectual property private and inaccessible to the end-user	Developed and tested through open collaboration
	Software is owned solely by the individual or organization that developed it	Source code can be accessed, modified, and redistributed by an open community of developers and programmers
	The limited market of developers and end-users, influenced by costs and flexibility	Encourages innovation of SMEs and individual users by accessing useful open-source platforms with no costs
Future Trends	Several governmental organizations have been regulating the protection and privacy of data, giving consumers more control over the personal information that businesses collect about them. With growing public awareness and discussion around data privacy and ownership, the future of closed and proprietary	Recent shifts to open-source models are indicative of the increasingly collaborative nature of technology advancements, and of increased consumer interest in understanding how the technologies we use impact our lives. The major challenges to the wider adoption of open-source platforms are funding and security vulnerabilities but are likely that

	approaches to software and emerging technologies are likely to be more and more decentralized.	decentralized technologies and data ownership will play a bigger role in the future.

Upon reviewing several options of Open-Source platforms to use in Data Science projects, [23] considers *Python* the best choice for scientists and engineers seeking a high-level language for writing scientific applications since it provides unique features such as:

- An open-source license that permits the user to use, sell, or distribute its Python-based applications;
- Innumerous libraries modules developed and improved by its community;
- Wide number of possible scientific areas in each it can be used;
- The language's clean syntax yet powerful constructs;
- The possibility to embed Python into existing applications, making the bridge between newer and older applications.

Besides its powerful standard library, there are many useful Python libraries such as *Scipy* [24], *Numpy* [25], *Pandas* [26], *Matplotlib* [27], *Seaborn* [28], and *Scikit-learn* [29]. Other open-source platforms mentioned in the literature that can be explored in industry, engineering, and other areas of human development include *R* [30], *Jupyter* [31], *Weka* [32], *Hadoop* [33], *Spark* [34], and *KNIME* [35].

4. Conclusions and Future Lines of Research

This paper presents the results of the literature review of Data Science for Industry 4.0 and Open Design approaches to innovation and engineering. Drawn from the bibliographic review presented, "Sustainability" and "Industry 4.0" are the topics that received more attention from the research community. Within the Industry 4.0 pillars, Big data analytics and Data science technologies have a growing importance in industry and engineering, and even though they have already shaped many economic sectors, they can be considered to be in an early stage of development. Proprietary technology is still the main approach to innovation, but open source is a growing method for development. SMEs, scientists, and academics are the main beneficiaries of the adoption of collaborative decision-making and data-sharing. Open innovation still does not embrace collaboration at its fullest since most companies are eager to share data from their challenges and difficulties but often resist sharing their solutions with third parties. The main limitations in these topics that were identified in the literature are concentrated in open approaches to innovation and social sustainability. Understanding these topics might require further lines of investigation and exploration, which require the gathering of data and information. This data will allow us to assess all relevant factors involved, define what is impeding solutions, and hopefully, reveal which actions can be implemented.

5. Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00014

Almirall, E., Lee, M., & Majchrzak, A. (2014). Open innovation requires integrated competition-community ecosystems: Lessons learned from civic open innovation. *Business Horizons*, *57*(3). https://doi.org/10.1016/j.bushor.2013.12.009

Angelidou, M. (2014). Smart city policies: A spatial approach. *Cities*, *41*, S3–S11. https://doi.org/10.1016/j.cities.2014.06.007

Azevedo, S. G., Pimentel, C. M. O., Alves, A. C., & Matias, J. C. O. (2021). Support of Advanced Technologies in Supply Chain Processes and Sustainability Impact. *Applied Sciences*, *11*(7). https://doi.org/10.3390/app11073026

Bai, C., Dallasega, P., Orzes, G., & Sarkis, J. (2020). Industry 4.0 technologies assessment: A sustainability perspective. *International Journal of Production Economics*, *229*. https://doi.org/10.1016/j.ijpe.2020.107776

Bamhdi, A. (2021). Requirements capture and comparative analysis of open source versus proprietary service oriented architecture. *Computer Standards & Interfaces*, *74*. https://doi.org/10.1016/j.csi.2020.103468

Benotsmane, R., Kovács, & Dudás, L. (2019a). Economic, Social Impacts and Operation of Smart Factories in Industry 4.0 Focusing on Simulation and Artificial Intelligence of Collaborating Robots. *Social Sciences*, *8*, 143. https://doi.org/10.3390/socsci8050143

Benotsmane, R., Kovács, G., & Dudás, L. (2019b). Economic, Social Impacts and Operation of Smart Factories in Industry 4.0 Focusing on Simulation and Artificial Intelligence of Collaborating Robots. *Social Sciences*, *8*(5). https://doi.org/10.3390/socsci8050143

Beruvides, G. (2019). *Modeling Techniques for Micromachining Processes*. https://doi.org/10.1007/978-3-030-03949-3_2

Betti, F., Bezamat, F., Fendri, M., Fernadez, B., Küpper, D., & Okur, A. (2020). Share to Gain: Unlocking Data Value in Manufacturing. *Http://Www3. Weforum. Org/Docs/WEF_Share_to_Gain_Report. Pdf [Stand: 27.04. 2020]*.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, *30*(3), 500–521. https://doi.org/https://doi.org/10.1016/j.aei.2016.07.001

Bin, S., & Dowlatabadi, H. (2005). Consumer lifestyle approach to US energy use and the related CO2 emissions. *Energy Policy*, *33*(2), 197–208. https://doi.org/10.1016/S0301-4215(03)00210-6

Biswas, A., & Roy, M. (2015). Leveraging factors for sustained green consumption behavior based on consumption value perceptions: testing the structural model. *Journal of Cleaner Production*, *95*. https://doi.org/10.1016/j.jclepro.2015.02.042

Bonilla, S., Silva, H., Terra da Silva, M., Franco Gonçalves, R., & Sacomano, J. (2018). Industry 4.0 and Sustainability Implications: A Scenario-Based Analysis of the Impacts and Challenges. *Sustainability*, *10*(10). https://doi.org/10.3390/su10103740

Boulanger, A. (2005). Open-source versus proprietary software: Is one more reliable and secure than the other? *IBM Systems Journal*, *44*(2). https://doi.org/10.1147/sj.442.0239

Boyes, H., Hallaq, B., Cunningham, J., & Watson, T. (2018). The industrial internet of things (IIoT): An analysis framework. *Computers in Industry*, *101*. https://doi.org/10.1016/j.compind.2018.04.015

Braccini, A., & Margherita, E. (2018). Exploring Organizational Sustainability of Industry 4.0 under the Triple Bottom Line: The Case of a Manufacturing Company. *Sustainability*, *11*(1). https://doi.org/10.3390/su11010036

Brenner, B., & Hartl, B. (2021). The perceived relationship between digitalization and ecological, economic, and social sustainability. *Journal of Cleaner Production*, *315*. https://doi.org/10.1016/j.jclepro.2021.128128

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011a). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1819486

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011b). Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1819486

Bughin, J. (2018, October 12). *Marrying artificial intelligence and the sustainable development goals: The global economic impact of AI*. https://www.mckinsey.com/mgi/overview/in-the-news/marrying-artificial-intelligence-and-the-sustainable

Bulao, J. (2021, December 7). *How Much Data Is Created Every Day in 2021?* TechJury. http://techjury.net/blog/how-much-data-is-created-every-day/

Castro, H., Pinto, N., Pereira, F., Ferreira, L., Ávila, P., Bastos, J., Putnik, G. D., & Cruz-Cunha, M. (2021). Cyber-Physical Systems using Open Design: An approach towards an Open Science Lab for Manufacturing. *Procedia Computer Science*, *196*(2021), 381–388. https://doi.org/10.1016/j.procs.2021.12.027

Castro, H., Pinto, N., Pereira, F., Ferreira, L., Avila, P., Putnik, G., Felgueiras, C., Bastos, J., & Cunha, M. (2021). Open Science Laboratory for Manufacturing: An education tool to contribute to sustainability. *ACM International Conference Proceeding Series*, 819–823. https://doi.org/10.1145/3486011.3486564

Castro, H., Putnik, G., Castro, A., & Bosco Fontana, R. D. (2019). Open Design initiatives: an evaluation of CAD Open Source Software. *Procedia CIRP*, *84*, 1116–1119. https://doi.org/10.1016/j.procir.2019.08.001

Castro, H., Putnik, G., Castro, A., & Fontana, R. D. B. (2019). Could Open Design learn from Wikipedia? *Procedia CIRP*, *84*, 1112–1115. https://doi.org/10.1016/j.procir.2019.07.001

Caulkins, J. P., Feichtinger, G., Grass, D., Hartl, R. F., Kort, P. M., & Seidl, A. (2013). When to make proprietary software open source. *Journal of Economic Dynamics and Control*, *37*(6). https://doi.org/10.1016/j.jedc.2013.02.009

Cavanillas, J. M., Curry, E., & Wahlster, W. (Eds.). (2016). *New Horizons for a Data-Driven Economy*. Springer International Publishing. https://doi.org/10.1007/978-3-319-21569-3

Chabbouh, H., & Boujelbene, Y. (2020). Open innovation in SMEs: The mediating role between human capital and firm performance. *The Journal of High Technology Management Research*, *31*(2), 100391. https://doi.org/https://doi.org/10.1016/j.hitech.2020.100391

Chakraborty, D., & Helling, R. K. (2021). Industry sustainable supply chain management with data science. In *Data Science Applied to Sustainability Analysis*. Elsevier. https://doi.org/10.1016/B978-0-12-817976-5.00010-3

Chesbrough, H., & Crowther, A. K. (2006). Beyond high tech: early adopters of open innovation in other industries. *R&D Management*, *36*(3), 229–236. https://doi.org/https://doi.org/10.1111/j.1467-9310.2006.00428.x

Chesbrough, H., Vanhaverbeke, W., & West, J. (2008). *Open Innovation: Researching A New Paradigm*.

Chesbrough, H. W. (2003). The Era of Open Innovation. *MIT Sloan Management Review*.

Das, S., Sismanis, Y., Beyer, K. S., Gemulla, R., Haas, P. J., & McPherson, J. (2010, June 6). Ricardo: integrating R and Hadoop. *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/1807167.1807275

Dash, P. B., Naik, B., Nayak, J., & Vimal, S. (2022). Socio-economic factor analysis for sustainable and smart precision agriculture: An ensemble learning approach. *Computer Communications*, *182*. https://doi.org/10.1016/j.comcom.2021.11.002

Denyer, D., Tranfield, D., & van Aken, J. E. (2008). Developing Design Propositions through Research Synthesis. *Organization Studies*, *29*(3). https://doi.org/10.1177/0170840607088020

Dong, K., Jiang, H., Sun, R., & Dong, X. (2019). Driving forces and mitigation potential of global CO2 emissions from 1980 through 2030: Evidence from countries with different income levels. *Science of The Total Environment*, *649*, 335–343. https://doi.org/10.1016/j.scitotenv.2018.08.326

Enyoghasi, C., & Badurdeen, F. (2021). Industry 4.0 for sustainable manufacturing: Opportunities at the product, process, and system levels. *Resources, Conservation and Recycling*, *166*. https://doi.org/10.1016/j.resconrec.2020.105362

Epstein, M. J., Elkington, J., & Leonard, H. B. “Dutch.” (2018). *Making Sustainability Work*. Routledge. https://doi.org/10.4324/9781351280129

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, *20*(15). https://doi.org/10.1093/bioinformatics/bth261

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2). https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Ghobakhloo, M. (2020). Industry 4.0, digitization, and opportunities for sustainability. *Journal of Cleaner Production*, *252*. https://doi.org/10.1016/j.jclepro.2019.119869

Ghobakhloo, M., & Fathi, M. (2019). Corporate survival in Industry 4.0 era: the enabling role of lean-digitized manufacturing. *Journal of Manufacturing Technology Management*, *31*(1). https://doi.org/10.1108/JMTM-11-2018-0417

Gökalp, M. O., Gökalp, E., Kayabay, K., Koçyiğit, A., & Eren, P. E. (2021). Data-driven manufacturing: An assessment model for data science maturity. *Journal of Manufacturing Systems*, *60*, 527–546. https://doi.org/https://doi.org/10.1016/j.jmsy.2021.07.011

Hall, T. (2020, May 20). *The Role of Data in Industry 4.0*. https://industrytoday.com/the-role-of-data-in-industry-4-0/

Han, H., & Trimi, S. (2022). Towards a data science platform for improving SME collaboration through Industry 4.0 technologies. *Technological Forecasting and Social Change*, *174*, 121242. https://doi.org/https://doi.org/10.1016/j.techfore.2021.121242

Hermann, M., Pentek, T., & Otto, B. (2015). *Design Principles for Industrie 4.0 Scenarios: A Literature Review*. https://doi.org/10.13140/RG.2.2.29269.22248

Hevner, March, Park, & Ram. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75. https://doi.org/10.2307/25148625

Hickin, R., Bechtel, M., Golem, A., Erb, L., & Buscalno, R. (2021). *Technology Futures: Projecting the Possible, Navigating What's Next*. https://www3.weforum.org/docs/WEF_Technology_Futures_GTGS_2021.pdf

International Monetary Fund. (2018). World Economic Outlook: Challenges to Steady Growth. *World Economic and Finantial Surveys*.

Inyang, B. J. (2013). Defining the Role Engagement of Small and Medium-Sized Enterprises (SMEs) in Corporate Social Responsibility (CSR). *International Business Research*, *6*(5). https://doi.org/10.5539/ibr.v6n5p123

Jain, A. K., Dhada, M., Parlikad, A. K., & Lad, B. K. (2020). Product Quality Driven Auto-Prognostics: Low-Cost Digital Solution for SMEs. *IFAC-PapersOnLine*, *53*(3), 78–83. https://doi.org/https://doi.org/10.1016/j.ifacol.2020.11.012

Kamble, S. S., Gunasekaran, A., & Gawankar, S. A. (2018). Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives. *Process Safety and Environmental Protection*, *117*. https://doi.org/10.1016/j.psep.2018.05.009

Kerin, M., & Pham, D. T. (2019). A review of emerging industry 4.0 technologies in remanufacturing. *Journal of Cleaner Production*, *237*. https://doi.org/10.1016/j.jclepro.2019.117805

Kerr, S. P., Kerr, W., Özden, Ç., & Parsons, C. (2016). *High-Skilled Migration and Agglomeration*. https://doi.org/10.3386/w22926

Kilamo, T., Hammouda, I., Mikkonen, T., & Aaltonen, T. (2012). From proprietary to open source—Growing an open source ecosystem. *Journal of Systems and Software*, *85*(7). https://doi.org/10.1016/j.jss.2011.06.071

Lasi, H., Fettke, P., Kemper, H.-G., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business & Information Systems Engineering*, *6*(4). https://doi.org/10.1007/s12599-014-0334-4

Lee, J. (2003). E-manufacturing—fundamental, tools, and transformation. *Robotics and Computer-Integrated Manufacturing*, *19*(6). https://doi.org/10.1016/S0736-5845(03)00060-7

Lee, J., Ardakani, H. D., Yang, S., & Bagheri, B. (2015). Industrial Big Data Analytics and Cyber-physical Systems for Future Maintenance &amp; Service Innovation. *Procedia CIRP*, *38*. https://doi.org/10.1016/j.procir.2015.08.026

Lee, J. S., Pries-Heje, J., & Baskerville, R. (2011). *Theorizing in Design Science Research* (pp. 1–16). https://doi.org/10.1007/978-3-642-20633-7_1

Lee, S., Ju, E., Choi, S., Lee, H., Shim, J., Chang, K., Kim, K., & Kim, C. (2018). *Prediction of Cancer Patient Outcomes Based on Artificial Intelligence*. https://doi.org/10.5772/intechopen.81872

Lemenkova, P. (2019). PROCESSING OCEANOGRAPHIC DATA BY PYTHON LIBRARIES NUMPY, SCIPY AND PANDAS. *Aquatic Research*. https://doi.org/10.3153/AR19009

Liao, Y., Deschamps, F., Loures, E. de F. R., & Ramos, L. F. P. (2017). Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal. *International Journal of Production Research*, *55*(12). https://doi.org/10.1080/00207543.2017.1308576

Lin, K., Shyu, J., & Ding, K. (2017). A Cross-Strait Comparison of Innovation Policy under Industry 4.0 and Sustainability Development Transition. *Sustainability*, *9*(5). https://doi.org/10.3390/su9050786

Lou, X., van der Lee, S., & Lloyd, S. (2013). AIMBAT: A Python/Matplotlib Tool for Measuring Teleseismic Arrival Times. *Seismological Research Letters*, *84*(1). https://doi.org/10.1785/0220120033

Luthra, S., & Mangla, S. K. (2018). Evaluating challenges to Industry 4.0 initiatives for supply chain sustainability in emerging economies. *Process Safety and Environmental Protection*, *117*. https://doi.org/10.1016/j.psep.2018.04.018

Machado, C. G., Winroth, M. P., & Ribeiro da Silva, E. H. D. (2020). Sustainable manufacturing in Industry 4.0: an emerging research agenda. *International Journal of Production Research*, *58*(5). https://doi.org/10.1080/00207543.2019.1652777

Mansfield, E., & Lee, J.-Y. (1996). The modern university: contributor to industrial innovation and recipient of industrial R&amp;D support. *Research Policy*, *25*(7), 1047–1058. https://doi.org/10.1016/S0048-7333(96)00893-1

Maresova, P., Soukal, I., Svobodova, L., Hedvicakova, M., Javanmardi, E., Selamat, A., & Krejcar, O. (2018). Consequences of Industry 4.0 in Business and Economics. *Economies*, *6*(3). https://doi.org/10.3390/economies6030046

McGoldrick, P. J., & Freestone, O. M. (2008). Ethical product premiums: antecedents and extent of consumers' willingness to pay. *The International Review of Retail, Distribution and Consumer Research*, *18*(2). https://doi.org/10.1080/09593960701868431

Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D. B., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M. J., Zadeh, R., Zaharia, M., & Talwalkar, A. (2016). MLlib: Machine Learning in Apache Spark. *J. Mach. Learn. Res.*, *17*(1), 1235–1241.

Merritt, J., Antunes, M., & Tanaka, Y. (2021). Governing Smart Cities: Policy Benchmarks for Ethical and Responsible Smart City Development. *World Economic Forum*. https://www3.weforum.org/docs/WEF_Governing_Smart_Cities_2021.pdf

Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment. *British Journal of Management*, *30*(2), 272–298. https://doi.org/https://doi.org/10.1111/1467-8551.12343

Mittal, S., Khan, M. A., Romero, D., & Wuest, T. (2018). A critical review of smart manufacturing &amp; Industry 4.0 maturity models: Implications for small and medium-sized enterprises (SMEs). *Journal of Manufacturing Systems*, *49*. https://doi.org/10.1016/j.jmsy.2018.10.005

Moon, J., Gbadago, D. Q., Hwang, G., Lee, D., & Hwang, S. (2021). Software platform for high-fidelity-data-based artificial neural network modeling and process optimization in chemical engineering. *Computers & Chemical Engineering*. https://doi.org/10.1016/j.compchemeng.2021.107637

Morrar, R., Arman, H., & Mousa, S. (2017). The Fourth Industrial Revolution (Industry 4.0): A Social Innovation Perspective. *Technology Innovation Management Review*, *7*, 12–20. https://doi.org/http://doi.org/10.22215/timreview/1117

Müller, J. M., Buliga, O., & Voigt, K.-I. (2018). Fortune favors the prepared: How SMEs approach business model innovations in Industry 4.0. *Technological Forecasting and Social Change*, *132*. https://doi.org/10.1016/j.techfore.2017.12.019

Müller, J. M., Kiel, D., & Voigt, K.-I. (2018). What Drives the Implementation of Industry 4.0? The Role of Opportunities and Challenges in the Context of Sustainability. *Sustainability*, *10*(1). https://doi.org/10.3390/su10010247

Müller, J. M., & Voigt, K.-I. (2018). Sustainable Industrial Value Creation in SMEs: A Comparison between Industry 4.0 and Made in China 2025. *International Journal of Precision Engineering and Manufacturing-Green Technology*, *5*(5). https://doi.org/10.1007/s40684-018-0056-z

Nagaraj, S. V. (2020). Disruptive technologies that are likely to shape future jobs. *Procedia Computer Science*, *172*. https://doi.org/10.1016/j.procs.2020.05.164

ODW. (2021). *Open Data Inventory 2020/21 Annual Report*. https://opendatawatch.com/publications/open-data-inventory/

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, *9*(3). https://doi.org/10.1109/MCSE.2007.58

Open Knowledge Foundation. (n.d.). *The Open Definition*. Retrieved May 30, 2022, from https://opendefinition.org/

P. Mazanetz, M., J. Marmon, R., B. T. Reisser, C., & Morao, I. (2012). Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Current Topics in Medicinal Chemistry*, *12*(18). https://doi.org/10.2174/156802612804910331

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, *20*(2). https://doi.org/10.1093/bioinformatics/btg412

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pereira, A. C., & Romero, F. (2017). A review of the meanings and the implications of the Industry 4.0 concept. *Procedia Manufacturing*, *13*. https://doi.org/10.1016/j.promfg.2017.09.032

Pereira, M. T., Silva, A., Ferreira, L. P., Sá, J. C., & Silva, F. J. G. (2019). A DMS to Support Industrial Process Decision-Making: a contribution under Industry 4.0. *Procedia Manufacturing*, *38*, 613–620. https://doi.org/https://doi.org/10.1016/j.promfg.2020.01.079

Peruzzini, M., Grandi, F., & Pellicciari, M. (2020). Exploring the potential of Operator 4.0 interface and monitoring. *Computers & Industrial Engineering*, *139*, 105600. https://doi.org/https://doi.org/10.1016/j.cie.2018.12.047

Piccarozzi, M., Aquilani, B., & Gatti, C. (2018). Industry 4.0 in Management Studies: A Systematic Literature Review. *Sustainability*, *10*(10). https://doi.org/10.3390/su10103821

Pinheiro, P., Putnik, G. D., Castro, A., Castro, H., Fontana, R. D. B., & Romero, F. (2019). Industry 4.0 and industrial revolutions: An assessment based on complexity. *FME Transactions*, *47*(4), 831–840. https://doi.org/10.5937/fmet1904831P

Pivoto, D. G. S., de Almeida, L. F. F., da Rosa Righi, R., Rodrigues, J. J. P. C., Lugli, A. B., & Alberti, A. M. (2021). Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review. *Journal of Manufacturing Systems*, *58*, 176–192. https://doi.org/https://doi.org/10.1016/j.jmsy.2020.11.017

Provost, F., & Fawcett, T. (2013a). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Provost, F., & Fawcett, T. (2013b). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1), 51–59. https://doi.org/10.1089/big.2013.1508

Provost, F., & Fawcett, T. (2013c). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, *1*(1). https://doi.org/10.1089/big.2013.1508

Putnik, G., & Ávila, P. (2021). Manufacturing system and enterprise management for Industry 4.0: Guest editorial. *FME Transactions*, *49*(4), 769–772. https://doi.org/10.5937/fme2104769P

Qian, F., Zhong, W., & Du, W. (2017). Fundamental Theories and Key Technologies for Smart and Optimal Manufacturing in the Process Industry. *Engineering*, *3*(2). https://doi.org/10.1016/J.ENG.2017.02.011

Quintas, J., Menezes, P., & Dias, J. (2017). Information Model and Architecture Specification for Context Awareness Interaction Decision Support in Cyber-Physical Human–Machine Systems. *IEEE Transactions on Human-Machine Systems*, *47*(3). https://doi.org/10.1109/THMS.2016.2634923

R Core Team. (2022). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. https://www.R-project.org/.

Ragan-Kelley, M., Perez, F., Granger, B., Kluyver, T., Ivanov, P., Frederic, J., & Bussonnier, M. (2014). The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. *AGU Fall Meeting Abstracts*, *2014*, H44D-07.

Ramadan, M., Al-Maimani, H., & Noche, B. (2017). RFID-enabled smart real-time manufacturing cost tracking system. *The International Journal of Advanced Manufacturing Technology*, *89*(1–4). https://doi.org/10.1007/s00170-016-9131-1

Roy, R., Stark, R., Tracht, K., Takata, S., & Mori, M. (2016). Continuous maintenance and the future – Foundations and technological challenges. *CIRP Annals*, *65*(2). https://doi.org/10.1016/j.cirp.2016.06.006

Runeson, P. (2019, May). Open Collaborative Data - using OSS Principles to Share Data in SW Engineering. *2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER)*. https://doi.org/10.1109/ICSE-NIER.2019.00015

Runeson, P., Olsson, T., & Linåker, J. (2021). Open Data Ecosystems — An empirical investigation into an emerging industry collaboration concept. *Journal of Systems and Software*, *182*. https://doi.org/10.1016/j.jss.2021.111088

Saritha, B., Bonagiri, R., & Deepika, R. (2021). Open source technologies in data science and big data analytics. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.01.610

Savastano, M., Amendola, C., Bellini, F., & D'Ascenzo, F. (2019). Contextual Impacts on Industrial Processes Brought by the Digital Transformation of

Manufacturing: A Systematic Review. *Sustainability*, *11*(3). https://doi.org/10.3390/su11030891

Schuh, G., Potente, T., Varandani, R., & Schmitz, T. (2014). Global Footprint Design based on genetic algorithms – An “Industry 4.0” perspective. *CIRP Annals*, *63*(1). https://doi.org/10.1016/j.cirp.2014.03.121

Shet, S. v., & Pereira, V. (2021). Proposed managerial competencies for Industry 4.0 – Implications for social sustainability. *Technological Forecasting and Social Change*, *173*. https://doi.org/10.1016/j.techfore.2021.121080

Sousa, A. C., Bertachini, A. F., Cunha, C., Chaves, R., & Varela, M. L. R. (2021). Literature review and discussion on collaborative decision making approaches in Industry 4.0. *FME Transactions*, *49*(4). https://doi.org/10.5937/fme2104817S

Stock, T., & Seliger, G. (2016). Opportunities of Sustainable Manufacturing in Industry 4.0. *Procedia CIRP*, *40*. https://doi.org/10.1016/j.procir.2016.01.129

Thames, L., & Schaefer, D. (2017). *Industry 4.0: An Overview of Key Benefits, Technologies, and Challenges* (pp. 1–33). https://doi.org/10.1007/978-3-319-50660-9_1

The World Bank. (2022a). GDP (Current US$). *World Bank*.

The World Bank. (2022b). Researchers in R&D (per million people). *The World Bank*. https://databank.worldbank.org/metadataglossary/jobs/series/SP.POP.SCIE.RD.P6

Toledo, R. F. de, Farias Filho, J. R. de, Castro, H. C. G. A. de, Putnik, G. D., & Silva, L. E. da. (2021). Is the incorporation of sustainability issues and Sustainable Development Goals in project management a catalyst for sustainable project delivery? *International Journal of Sustainable Development and World Ecology*, *28*(8), 733–743. https://doi.org/10.1080/13504509.2021.1888816

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, *14*(3). https://doi.org/10.1111/1467-8551.00375

UN. (2022a). Goal 9 - Industry, Innovation and Infrastructure. *Sustainable Development Goals*.

UN. (2022b). Goal 12 - Responsible Comsumption and Production. *Sustainable Development Goals*.

UN. (2022c). Goal 17 - Partnerships for the Goals. *Sustainable Development Goals*.

United Nations. (2015). Transforming our world: the 2030 Agenda for Sustainable Development. *Department of Economic and Social Affairs and Sustainable Development*. https://sdgs.un.org/2030agenda

Vaidya, S., Ambad, P., & Bhosle, S. (2018). Industry 4.0 – A Glimpse. *Procedia Manufacturing*, *20*, 233–238. https://doi.org/https://doi.org/10.1016/j.promfg.2018.02.034

Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2). https://doi.org/10.1007/s11192-009-0146-3

van Ham, H., & Koppenjan, J. (2001). BUILDING PUBLIC-PRIVATE PARTNERSHIPS: Assessing and managing risks in port development. *Public Management Review*, *3*(4), 593–616. https://doi.org/10.1080/14616670110070622

Varela, L., Araújo, A., Ávila, P., Castro, H., & Putnik, G. (2019). Evaluation of the Relation between Lean Manufacturing, Industry 4.0, and Sustainability. *Sustainability*, *11*(5), 1439. https://doi.org/10.3390/su11051439

Varela, L., Ávila, P., Castro, H., Putnik, G. D., Fonseca, L. M. C., & Ferreira, L. (2022). Manufacturing and Management Paradigms, Methods and Tools for Sustainable Industry 4.0-Oriented Manufacturing Systems. *Sustainability*, *14*(3), 1574. https://doi.org/10.3390/su14031574

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a Framework for Evaluation in Design Science Research. *European Journal of Information Systems*, *25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Wang, B., Wu, C., Huang, L., & Kang, L. (2019). Using data-driven safety decision-making to realize smart safety management in the era of big data: A theoretical perspective on basic questions and their answers. *Journal of Cleaner Production*, *210*, 1595–1604. https://doi.org/https://doi.org/10.1016/j.jclepro.2018.11.181

Wang, S., Wan, J., Li, D., & Zhang, C. (2016). Implementing Smart Factory of Industrie 4.0: An Outlook. *International Journal of Distributed Sensor Networks*, *12*(1). https://doi.org/10.1155/2016/3159805

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60). https://doi.org/10.21105/joss.03021

Wee, D., Kelly, R., Cattel, J., & Breuning, M. (2015). Industry 4.0-how to navigate digitization of the manufacturing sector. *Mckinsey & Company*, *58*.

WEF. (2020). The Future of Jobs Report 2020. *WEF*.

WEF, & BCG. (2021). *Net-Zero Challenge: The supply chain opportunity*.

Witkowski, K. (2017). Internet of Things, Big Data, Industry 4.0 – Innovative Solutions in Logistics and Supply Chains Management. *Procedia Engineering*, *182*. https://doi.org/10.1016/j.proeng.2017.03.197

World Bank. (2022a). *Industry (including construction), value added (% of GDP)*. World Bank.

World Bank. (2022b). Skills | LinkedIn Data. *Data Catalog*.

Xiao, Y., & Watson, M. (2019). Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research*, *39*(1). https://doi.org/10.1177/0739456X17723971

Xu, Z., Zhu, Y., Hu, Y., Huang, M., Xu, F., & Wang, J. (2021). Bibliometric and visualized analysis of Neuropathic pain based on Web of Science and

CiteSpace over the last 20 years. *World Neurosurgery*. https://doi.org/10.1016/j.wneu.2021.12.025

8.2 Attachment B – Paper *"An Analysis of Open Data Scoring System towards Data Science for Sustainability in Industry 4.0"*

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

An analysis of Open Data Scoring System towards Data Science for Sustainability in Industry 4.0

Hélio Castro[a, b]2, Filipe Costa[a], Tânia Ferreira[c], Paulo Ávila[a, b], Manuela Cruz-Cunha[d], Luís Ferreira[d], Goran D. Putnik[e], João Bastos[a, b]

[a]*School of Engineering, Polytechnic of Porto, Rua Dr. António Bernardino de Almeida, 431, 4249-015 Porto, Portugal*
[b]*INESC TEC, Rua Dr. Roberto Frias, 4200-465 Porto, Portugal*
[c]*Data CoLAB, Av. de Cabo Verde, 1, 4900-568 Viana do Castelo, Portugal*
[d]*2Ai - Applied Artificial Intelligence Lab, School of Technology, IPCA, Barcelos, Portugal*
[e]*University of Minho, School of Engineering, Campus de Azurém, 4800-058 Guimarães, Portugal*

Abstract

In a society based on data-driven, data inclusion and data access play a significant role in societal development. A called democratization of data through open access, Open Data, must be nurtured by countries to empower their citizens, entrepreneurs, companies, industries, academics, and organizations, in general. Open Data Scoring System is an evaluation system that ranks countries in 22 categories of openness in data, divided into the 3 pillars of sustainability. In this paper, we will present the importance of Industry 4.0 and its relation to sustainability and the role of Data Science in Industry 4.0 assuming an Open Design approach. Then, an analysis is made considering the Gross Domestic Product (GDP) of the most relevant countries worldwide, the USA and China, concerning the six (6) higher ranked categories of openness data of these countries, supported by the Open Data Scoring System from 2015 to 2020. Our findings reveal that in the USA and China the main categories are seven (7), five (5), and 2 (two) categories of economic, social, and environmental sustainability, respectively. Through a correlations and co-occurrences analysis of the open data scoring worldwide reveals that the most significant categories are four (4) economic, one (1) social, and two (2) environmental.

* Corresponding author. Tel.: +351 +351 22 83 40 500; fax: +351 22 83 21 159.
E-mail address: hcc@isep.ipp.pt

Peer-review under responsibility of the scientific committee of the CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2022
Keywords: Industry 4.0; Open Design; Sustainabilty; Data Science; Open Data Scoring System.

1. Introduction

The manufacturing, scientific and technological fields have been subject to a revolution process of digitalization and technological development called Industry 4.0 (Liao et al., 2017). This process is implementing changes that stimulate more competitive practices across many economic sectors. These changes are in great part supported by the growing acquisition and utilization of information and data, that can be exploited through Big Data technologies and Data Science. The implementation of Industry 4.0 has proved to be successful mostly in the economic field of the framework in sustainability. This urges the social and environmental mindset of Industry 4.0 and it is still in its early stages of development and approaches to these pillars of sustainability. Industry 4.0 is still an undergoing vision (Pinheiro et al., 2019) and sustainability is still a required topic of development, even in other fields such as project management (Toledo et al., 2021).

To leverage many organizations, namely Universities, Research Institutes, and SMEs (small and medium enterprises), the Open Design concept is being adopted to develop software and hardware solution (Castro, Pinto, Pereira, Ferreira, Ávila, et al., 2021), even tools for educational purposes in sustainability (Castro, Pinto, Pereira, Ferreira, Avila, et al., 2021). Open Data, as an approach of Open Design to data, is a possible path to pull organizations regarding the implementation of Big Data technologies and Data Science bason in open-source and low-cost within digital community-based platforms. The second section frames the concept of Industry 4.0 relating to Science Data and Open Design, and the three (3) pillars of Sustainability (Economic, Social, and Environmental). The third section introduces the research methodology by presenting the Open Data Scoring System and the data analysis tool. Results from GPD data and openness scores and corresponding critical analysis are performed in the fourth section.

2. Literature Review

This section presents the main findings of the literature review. It is structured in four subchapters: "Data Science in Industry 4.0 and Open Design" – presents relevant information found about uses of data science and ownership in Industry 4.0 – "Economic Sustainability in Industry 4.0", "Social Sustainability in Industry 4.0" and "Environmental Sustainability in Industry 4.0" which explore the current issues and trends of each one of the three Sustainability pillars in the context of Industry 4.0.

2.1. Data Science in Industry 4.0 and Open Design

Industry 4.0 has gained increased adoption in recent years with its promise to use the power of data to revolutionize manufacturing. However, while the exploration of data has been a catalyst for business growth and efficiency gains, the manufacturing sector has been slow to adopt data-driven processes. According to Accenture, only 13% of manufacturing companies have implemented an Industry 4.0 approach (Hall, 2020). It is inevitable to data to become a cornerstone in the decision-making of not only in industrial processes, but also in the sustainability of economic, social, and environmental approaches. As we live in a world that constantly produces and consumes data, it is a priority to understand the value that can be extracted from it. Mikalef et al. (Mikalef et al., 2019) consider data science and the big data domains as the next frontier for both practitioners and researchers as they embody significant potential in exploiting data to sustain competitive advantage. Data science is an interdisciplinary field that supports and guides the extraction of useful patterns from raw data by exploring advanced technologies, algorithms, and processes (Provost & Fawcett, 2013a). Information remains largely centralized, but the trend is shifting toward a distributed and open model of data sharing (Hickin et al., 2021). Hickin et al (Hickin et al., 2021) represent a possible transition from known technologies to future trends in which distributed approaches such as open-source explainable AI and decentralized data ownership constitute a positive linear transition, which envisions an Open Design architecture.

2.2. Economic Sustainability in Industry 4.0

Available literature supports the idea of Industry 4.0 leading to reduced costs in manufacturing and maintenance, reduce times of production, improve supply-demand forecasting and increase productivity overall, which leads to improved economic performance [10-12]. In the next five years, more than 80% of European companies will digitalize their value chain and increase efficiency by 18% (M. T. Pereira et al., 2019). As SMEs account for approximately 90% of the world's enterprises (Inyang, 2013), it is crucial for this type of firm to accelerate innovation and digitalization to stay competitive on a global scale. Pivoto et al. [15] point out that to do so, manufacturing companies need to integrate science capabilities vertically and horizontally across the organization and shift toward data-driven manufacturing. From a quantitative perspective, data-driven organizations have demonstrated 6% higher productivity and efficiency than similar organizations that have not adopted data-driven processes and with further implementation of Industry 4.0, this number is set to increase (Brynjolfsson et al., 2011b). Gökalp et al. (Gökalp et al., 2021) address a study by McKinsey (Bughin, 2018) that expects non-adopters of data science in their processes will experience a 20% decrease in their cashflows by 2030. Wee et al. Wee et al. [19] represent the positive impact of Industry 4.0 in an increase of productivity and reduction of unproductive times and costs in eight economic value drivers: Resources, Asset Utilization, Labor, Inventories, Quality, Supply/Demand match, Time to market and Services. Those improvements are largely owed to increased accuracy in supply/demand forecasting through big data analytics (Enyoghasi & Badurdeen, 2021). Economic sustainability is a huge focus for companies, governments, and institutions in its operation, and is relevant for their social and environmentally sustainable progress (Epstein et al., 2018).

2.3. Social Sustainability in Industry 4.0

Even though Social Sustainability is arguably the most relevant topic for the sustainable development of the human future, it is the one with the scarcest literature and bibliometric available resources. For that reason, available literature reveals a profound need for research on social data. One of the main issues regarding the relationship between the adoption of Industry 4.0 and the future of work is job shortages. The increasing digitalization and automation of business and service tasks often lead to worries about the permanent replacement of the human labor force by machines. However, literature shows that that can be a misconception about the future of work. Shet & Pereira (Shet & Pereira, 2021) argue that Industry 4.0 can generate job prospects by creating new employment opportunities in emerging domains, like Science, Technology, Engineering, and Mathematics. The World Economic Forum conducted a survey in 2020 among a wide number of companies that indicate that 55% of them are looking to transform the composition of their value chain, and 43% will introduce further automation and reduce the current workforce. On the other way, the same survey showed that 34% of them will expand their workforce because of deeper technological integration and 41% are looking into expanding their use of contractors for task-specialized work (WEF, 2020).

2.4. Environmental Sustainability in Industry 4.0

In the environmental context, sustainable Industry 4.0 promotes efficient resource allocation like energy, water, raw materials, and other products, based on real-time data analysis and other technologies, resulting in sustainable green practices (Kamble et al., 2018; Stock & Seliger, 2016). According to WEF & BCG (WEF & BCG, 2021), addressing supply-chain emissions alone enables many companies to impact a volume of emissions several times higher than they could if they were to focus on decarbonizing their operations and power consumption alone. The share of carbon emissions by different industries can be split into three different scopes: own operations (Scope 1), consumed power (Scope 2), and supply chain (Scope 3). Even though the share of emissions of the three scopes are balanced in raw materials industries, in the end, in products industries the carbon emissions of supply-chain operations are far larger than the sum of the other two scopes combined, accounting for almost 90% of emissions (WEF & BCG, 2021). It is important to consider that for wide adoption of decarbonizing practices in supply chains it is necessary to guarantee sustainable economic solutions both for the companies and the end consumers. According to (WEF & BCG, 2021) around 40% of emissions in supply-chains in several economic sectors could be eliminated with affordable costs, resulting in a marginal impact on end-product costs. From a general point of view, literature shows there are still some challenges regarding the adoption of Industry 4.0 and sustainable economic, social, and environmental practices. However, there is a wide variety of possible solutions and incentives for that implementation, that will lead to newer and greater development possibilities for humankind.

3. Research Methodology

The methodology for the research and analysis of this Data Science for Sustainability in Industry 4.0 is based on an Open Data Scoring System which is used in conjunction with the statistical and graphical programming language R.

3.1. Open Data Scoring System

The first step for starting the research is to identify the main variable of the study, which must be representative of the overall openness of the country and relevant to economic, social, and environmental sustainability. By evaluating data published on official National Statistical Offices (NSOs), the Open Data Inventory (ODIN) 2020/21 provides an assessment of the coverage and openness of official statistics in 187 countries, monitors the progress of open data that are relevant to the economic, social, and environmental development of a country. The available statistics are grouped into each one of those three groups as is represented in Table 1.

Table 1. Economic, Social and Environmental categories for Open Data Scoring

	Category	**Representative Indicators**
Economic Statistics	National Accounts	Production by industry; expenditure by government and households
	Labor Statistics	Employment; unemployment; child labor
	Price Indexes	Consumer price index; Producers price index
	Central Government Finance	Actual revenues; actual expenditures
	Money and Banking	Money supply
	International Trade	Exports and imports
	Balance of Payments	Exports and imports of goods and services; foreign investment
Social Statistics	Population and Vital Statistics	Population by 5-year age groups; crude birth rate; crude death rate
	Education Facilities	Number of schools and classrooms; teaching staff; annual budget
	Education Outcomes	Enrollment and completion rates; literacy rates and/or competency exam results
	Health Facilities	Core operational statistics of health system
	Health Preventive Care	Immunization rates; incidence and prevalence major communicable diseases
	Reproductive Health	Maternal mortality ratio; infant mortality rate; under-5 mortality rate; fertility rate; contraceptive prevalence rate
	Food and Nutrition	Quantity and quality of food; Access to food and nutrition
	Gender Statistics	Specialized studies of the status and condition of women
	Crime and Justice	Number of crimes and justice
	Poverty Statistics	Number and percentage of poor at national poverty line; distribution of income
Environmental Statistics	Land Use	Land area
	Resource Use	Fishery harvests; forests coverage and deforestation; major mining activities; water supply & use
	Energy Use	Consumption of electricity, coal, oil, and renewables
	Pollution	Emissions of air and water pollutants; CO2 and other GHG; toxic substances
	Built Environment	Access to drinking water; access to sanitation; housing quality

For each one of the 22 categories, there is a preferred disaggregation that should be made by the NSOs that is also scored. This includes disaggregation by sex, age groups, and employment by industry in labor statistics, for example. Such disaggregations greatly increase the analytical value of the data (ODW, 2021). For the analysis of this paper, it will be added one more category "All categories" considers all the categories mentioned above.

Now that the categories are identified, it is necessary to understand the criteria behind the classification of the openness scores. This methodology will be referred to as Scoring System. There are two main dimensions of each data category that are assessed by ODIN: coverage and openness. For data coverage, ODIN considers 5 elements that are quantified with one point if the criterion is satisfied, one-half point if the criterion is partly satisfied, and zero if the criterion is not satisfied. Table 2 represents the elements of the coverage criteria scoring.

Table 2. Coverage criteria scoring

Time Coverage		**Geographic**		**Disaggregation**
Data available in the last 5 years (cs1)	Data available in the last 10 years (cs2)	First admin level (cs3)	Second admin level (cs4)	Recommended disaggregations as described in attachment x (cs5)
Complete: 1 Some: 0.5 None: 0		Yes: 1 No: 0		All:1 Some: 0.5 None:0

For each one of the 22 categories, the coverage score is given by the sum of the score of each element: cs1, cs2, cs3, cs4, and cs5. There are also five elements to the data openness dimension, which are classified the same way as the coverage criteria. According to ODIN, these elements are a representation of standards for open data, such as the Open Definition (Open Knowledge Foundation, n.d.). These elements are representative of the ability to select, access, and share data.

Table 3. Data openness criteria scoring

Download Format			**Metadata Available**	**Licensing Terms**
Machine Readable (os1)	Non-proprietary (os2)	User selection /API or bulk download (os3)	Metadata available (os4)	Terms of use (ToU) stated/ CC BY 4.0 (os5)
Yes: 1 No: 0		User selected: 0.5 API option: plus 0.5	Specific to indicator/dataset: 1 Non-specific:0.5 No: 0	ToU: 0.5 CC BY: plus 0.5

For each category, the openness score is given by the sum of the score of each element: os1, os2, os3, os4, and os5. The Category Scores are obtained for each one of the 22 categories by the average of the 10 scores obtained from the elements of the coverage and openness criteria (5 elements each). Economic Score is the average of its 7 category scores, Social Score is the average of its 10 Category Scores and the Environmental Score is the average of its 5 category scores. For this analysis, it was considered an equal weighting between the three sustainability elements, which corresponds to one-third of the overall score classification. The Overall Score, value between 0 and 1 (or 0% and 100%), represents the final score represented by an equal weighting of each sustainable pillar:

$$Overall\ Score = {}^{1}/_{3}\ Economic\ Score + {}^{1}/_{3}\ Social\ Score + {}^{1}/_{3}\ Environmental\ Score$$

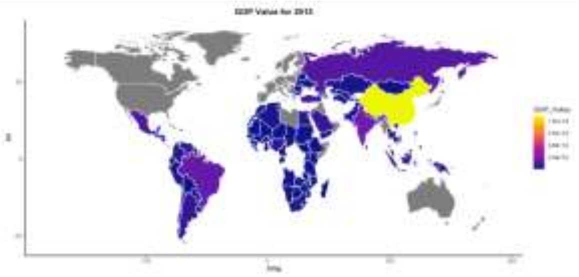
The result of the valuation of those categories is a robust dataset of scores of the openness of data that will be considered throughout the analysis of the correlation of the scores with other quantitative data regarding Industry 4.0 and Sustainability.

3.2. Data Analysis Tool

For the analysis described below, it is used the R programming language. R is a free open-source programming language and provides a computer environment for statistical and graphical techniques that can be used by importing libraries. These techniques can be used to handle raw data and retrieve information to have a sense of how the data is distributed or patterns that are masked (R Core Team, 2022). The R packages used were *arules* and *arulesViz* for the rule association.

4. Results and Critical Analysis

Our data set is composed by six (6) different variables: Year, Region, Country, Data Categories, Overall Score, and GDP ($). This data set was obtained by joining the values of GDP for all years under analysis and the information for the data categories and respective overall scores by country. The sample consists of 5 years of sampling (2015 to 2018, 2020). Although it was possible to obtain the value of GDP for all countries and years between 2015 and 2020, it was not possible to obtain the evaluation of open data by category for 2019. For this reason, the year 2019 was not considered in the analysis, since its focus is to understand how the value of GDP is distributed by countries over the years and how the type of data influences or not the degree of opening these same data. The first part of the analysis is focused on understanding how the value of GDP is distributed by country over the different years. For this purpose, we use a plot that presents a map with a scale that represents the range of values that were registered throughout the years for all the countries under analysis. In the graph for the year 2015, it is possible to see that China is the country that reaches the highest value. The following countries that stand out the most, although not so distinctly, are Brazil, India, Mexico, and Russia. The remaining countries always present values within the same spectrum.



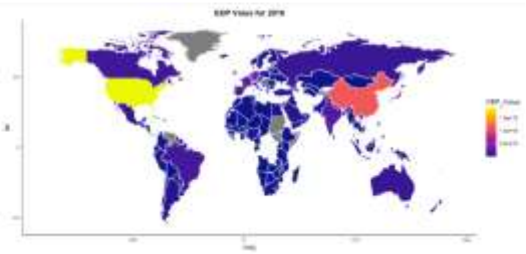


Fig. 1. Analysis for the year 2015

Fig. 2. Analysis for the year 2016

In 2016 China continues to be highlighted. However, the United States is the country with the highest GDP value. As for the remaining countries, there are two more that stand out from the rest and they are Japan and Germany. For the years 2017, 2018, and 2020 the graphs are practically the same. In a general context, we can see that the United States is the country that always obtains the highest GDP values, and the only country that follows it is China. It is also possible to see that China has shown an increase in its GDP value over the years, reaching its maximum value in 2020 (Figure 3), like the United States. Germany reaches its highest value in 2018 and Japan in 2020. India is one of the countries that stands out from the rest and reaches its maximum value in 2018. Brazil was more unstable and registered its lowest value in 2020.

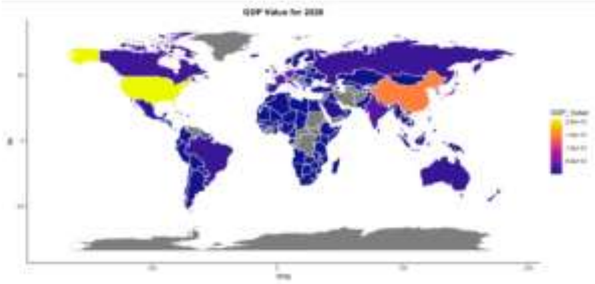


Fig. 3. Analysis for the year 2020.

Below is shown the graph, in Fig.4, that represents the sum of the GDP values for each region, that is, for each of the years and regions, the GDP value of each country per year was added.

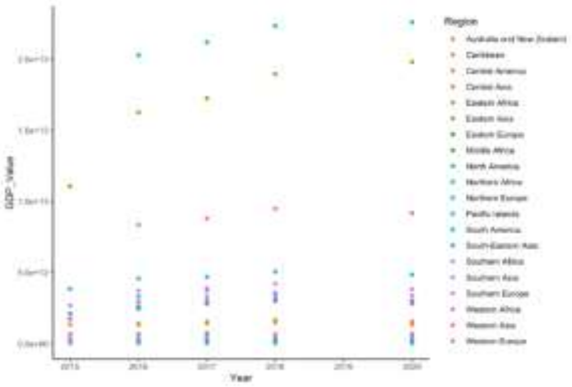


Fig. 5. Analysis of GDP by region per year

To better understand the influence of the Economic, Social, and Environmental categories on GDP, the research focus now on the two biggest countries in the world by GDP: USA and China. Table 3 represents the group of categories with the highest score from the year 2015 to 2020 in the USA and China. For each country is shown the 6 categories with the highest score and its value, by year. The Overall Scores in the USA indicate 2016 to be the year with the highest openness followed by 2018. The year 2017 brought a pullback in the overall openness and 2020 was the year with the less openness in the study. China's scores show a descending trend of openness from 2015 to 2020 with 2018 being an outlier. By comparing the scores of the USA and China, the USA has a higher openness across the Economic, Social and Environmental categories, with the highest scores being International Trade and Balance of Payments in 2018 and 2020 and Population and Vital Statistics in 2018. China's highest score was 72 in National Accounts in 2015 and Central Government Finance in 2015 and 2017 and Price Indexes in 2018, all of which are lower than the lowest score of the USA between the represented categories, which is 85 for Education Facilities and Education Outcomes in 2016. We can also conclude that the Economic pillar had the highest influence on the openness of both countries since it accounts for 7 of the 13 categories with the highest scores and it is followed by the Social pillar with 4 categories. The Environmental pillar accounts for only 2 of the 13 pillars but it has gained influence on the overall openness mainly in China, appearing among its 6 highest ranked categories since 2017.

Table 3. Highest ranked categories from USA and China from 2015 to 2020

		2015		2016		2017		2018		2020	
Categories		CHINA	USA *	CHINA	USA	CHINA	USA	CHINA	USA	CHINA	USA
Economic	National Accounts	72					89		94		94
	Central Government Finance	72		61		72		67	89	61	94
	International Trade	69		63	88		94		100		100
	Balance of Payments	69			88	69	94	69	100	56	100
	Labor	65			95		90				95
	Money and Banking			63		69		63		50	
	Price Indexes							72			
Social	Population and Vital Statistics	65		60	90		95		100		
	Health Facilities			60		65				55	
	Education Facilities			55	85						
	Education Outcomes				85						

Environmenta l	Resource Use					65		67		61	
	Energy Use					55	80	61	89	63	88
OVERALL SCORE		55,15		44,10	74,80	41,50	68,60	44,40	73,69	35,10	70,40

(*) – Data is not available.

The categories with the highest openness scores are International Trade, Balance of Payments, Population and Vital Statistics, Central Government Finance, and Labor. However, to analyze the importance of the categories in the openness of all geographic regions, it is necessary to understand not only the score value but also the co-occurrence of the category between the highest-ranked ones for each year. To explore the part of the data related to the categories and the degree of openness of the data, we used association rules. Association rules are used to discover correlations and co-occurrences between the data set. The support indicates how frequently a set of items appear. The confidence demonstrates how often the support rule is true. The lift value is the ratio between the confidence of the rule and the expected confidence of that rule. If the value is higher than one (1) then they are positively correlated, if it is less than one (1), they are negatively correlated and if it is equal to one (1) they are independent.
Considering the dataset, seven (7) rules were generated, assuming a minimum support value equal to 0.02 and a confidence value equal to 0.5. The generated rules are exemplified in Figure 5.

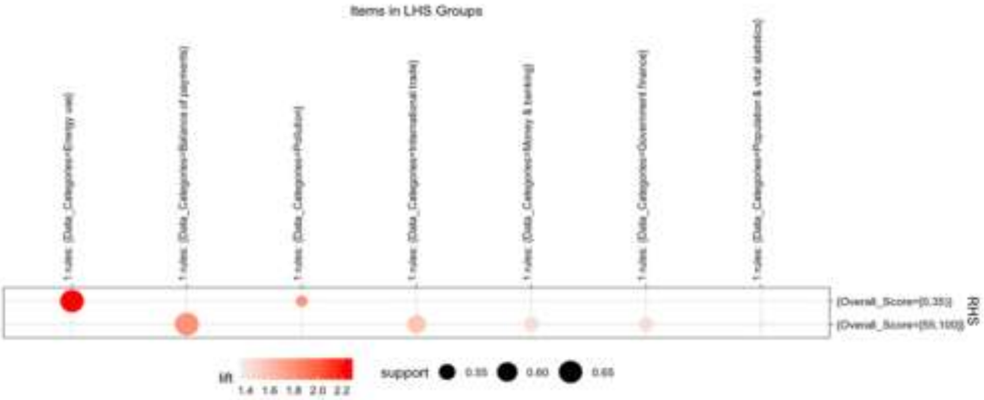


Fig. 5. Representation of the rules used for the correlations and co-occurrences of the data set

All rules have a lift value greater than 1 which means the rules appear together more often than expected. The rule with higher lift and confidence is Energy Use with an Overall Score of 0,35 followed by Balance of Payments, Pollution, International Trade, Money and Banking, Central Government Finance, and Population and Vital Statistics. This analysis demonstrates that the categories which have both a high openness score and high co-occurrence between the data set are International Trade, Balance of Payments, and Central Government Finance in the Economic spectrum and Population and Vital Statistics in the Social spectrum.

5. Conclusions

One of the most significant trends in society is the sustainability concern and the openness of data should support sustainability awareness and mechanisms within Industry 4.0. Regarding the three (3) pillars of sustainability (economic, social, and environmental), social and environmental sustainability pillars are relegated to secondary plane due to the strong emphasis on the economic sustainability pillar, although most recently its importance is growing in academia and lawmakers' communities. To understand the impact of Open Data in society, we related the most significant countries in the world in the economic aspect, considering the GDP, in this case, the USA and China, and an open data inventory scoring. One of the first conclusions is that the USA has the highest openness scores in all categories than China, and China has a negative trend in the open score in the time series of this study. Considering the six (6) higher scores in the years 2015, 2016, 2017, 2018, and 2020 of open data in each country, the economic categories have more influence on the openness with seven (7) categories (National Accounts, Central Government

Finance, International Trade, Balance of Payments, Labor, Money and Banking, and Price Indexes), followed social categories with five (5) (Population and Vital Statistics, Health Facilities, Education Facilities, and Education Outcomes), and environmental categories with 2 (two) (Resource Use and Energy Use).

A correlations and co-occurrences analysis of the open data scoring worldwide reveals that Energy Use, Balance of Payments, Pollution, International Trade, Money and Banking, Central Government Finance, and Population and Vital Statistics are the most significant categories (four (4) economic, one (1) social, and two (2) environmental).

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. This paper was funded by national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the projects UIDB/05549/2020 and UIDP/05549/2020.

References

[1] Y. Liao, F. Deschamps, E. de F. R. Loures, and L. F. P. Ramos, “Past, present and future of Industry 4.0 - a systematic literature review and research agenda proposal,” *International Journal of Production Research*, vol. 55, no. 12, Jun. 2017, doi: 10.1080/00207543.2017.1308576.

[2] P. Pinheiro, G. D. Putnik, A. Castro, H. Castro, R. D. B. Fontana, and F. Romero, “Industry 4.0 and industrial revolutions: An assessment based on complexity,” *FME Transactions*, vol. 47, no. 4, pp. 831–840, 2019, doi: 10.5937/fmet1904831P.

[3] R. F. de Toledo, J. R. de Farias Filho, H. C. G. A. de Castro, G. D. Putnik, and L. E. da Silva, “Is the incorporation of sustainability issues and Sustainable Development Goals in project management a catalyst for sustainable project delivery?,” *International Journal of Sustainable Development and World Ecology*, vol. 28, no. 8, pp. 733–743, 2021, doi: 10.1080/13504509.2021.1888816.

[4] H. Castro *et al.*, “Cyber-Physical Systems using Open Design: An approach towards an Open Science Lab for Manufacturing,” *Procedia Computer Science*, vol. 196, no. 2021, pp. 381–388, 2021, doi: 10.1016/j.procs.2021.12.027.

[5] H. Castro *et al.*, “Open Science Laboratory for Manufacturing: An education tool to contribute to sustainability,” *ACM International Conference Proceeding Series*, pp. 819–823, 2021, doi: 10.1145/3486011.3486564.

[6] Tim Hall, “The Role of Data in Industry 4.0,” May 20, 2020.

[7] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, “Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment,” *British Journal of Management*, vol. 30, no. 2, pp. 272–298, 2019, doi: https://doi.org/10.1111/1467-8551.12343.

[8] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Big Data*, vol. 1, no. 1, pp. 51–59, 2013, doi: 10.1089/big.2013.1508.

[9] R. Hickin, M. Bechtel, A. Golem, L. Erb, and R. Buscalno, “Technology Futures: Projecting the Possible, Navigating What’s Next,” Apr. 2021. Accessed: Dec. 26, 2021. [Online]. Available: https://www3.weforum.org/docs/WEF_Technology_Futures_GTGS_2021.pdf

[10] G. Schuh, T. Potente, R. Varandani, and T. Schmitz, “Global Footprint Design based on genetic algorithms – An ‘Industry 4.0’ perspective,” *CIRP Annals*, vol. 63, no. 1, 2014, doi: 10.1016/j.cirp.2014.03.121.

[11] M. Ramadan, H. Al-Maimani, and B. Noche, “RFID-enabled smart real-time manufacturing cost tracking system,” *The International Journal of Advanced Manufacturing Technology*, vol. 89, no. 1–4, Mar. 2017, doi: 10.1007/s00170-016-9131-1.

[12] S. S. Kamble, A. Gunasekaran, and S. A. Gawankar, “Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives,” *Process Safety and Environmental Protection*, vol. 117, Jul. 2018, doi: 10.1016/j.psep.2018.05.009.

[13] M. T. Pereira, A. Silva, L. P. Ferreira, J. C. Sá, and F. J. G. Silva, “A DMS to Support Industrial Process Decision-Making: a contribution under Industry 4.0,” *Procedia Manufacturing*, vol. 38, pp. 613–620, 2019, doi: https://doi.org/10.1016/j.promfg.2020.01.079.

[14] B. J. Inyang, “Defining the Role Engagement of Small and Medium-Sized Enterprises (SMEs) in Corporate Social Responsibility (CSR),” *International Business Research*, vol. 6, no. 5, Apr. 2013, doi: 10.5539/ibr.v6n5p123.

[15] D. G. S. Pivoto, L. F. F. de Almeida, R. da Rosa Righi, J. J. P. C. Rodrigues, A. B. Lugli, and A. M. Alberti, “Cyber-physical systems architectures for industrial internet of things applications in Industry 4.0: A literature review,” *Journal of Manufacturing Systems*, vol. 58, pp. 176–192, 2021, doi: https://doi.org/10.1016/j.jmsy.2020.11.017.

[16] E. Brynjolfsson, L. M. Hitt, and H. H. Kim, “Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?,” *SSRN Electronic Journal*, 2011, doi: 10.2139/ssrn.1819486.

[17] M. O. Gökalp, E. Gökalp, K. Kayabay, A. Koçyiğit, and P. E. Eren, “Data-driven manufacturing: An assessment model for data science maturity,” *Journal of Manufacturing Systems*, vol. 60, pp. 527–546, 2021, doi: https://doi.org/10.1016/j.jmsy.2021.07.011.

[18] J. Bughin, “Marrying artificial intelligence and the sustainable development goals: The global economic impact of AI,” Oct. 12, 2018. Accessed: Dec. 16, 2021. [Online]. Available: https://www.mckinsey.com/mgi/overview/in-the-news/marrying-artificial-intelligence-and-the-sustainable

[19] D. Wee, R. Kelly, J. Cattel, and M. Breuning, “Industry 4.0-how to navigate digitization of the manufacturing sector,” *Mckinsey & Company*, vol. 58, 2015.

[20] C. Enyoghasi and F. Badurdeen, “Industry 4.0 for sustainable manufacturing: Opportunities at the product, process, and system levels,” *Resources, Conservation and Recycling*, vol. 166, Mar. 2021, doi: 10.1016/j.resconrec.2020.105362.

[21] M. J. Epstein, J. Elkington, and H. B. “Dutch” Leonard, *Making Sustainability Work*. Routledge, 2018. doi: 10.4324/9781351280129.

[22] S. v. Shet and V. Pereira, “Proposed managerial competencies for Industry 4.0 – Implications for social sustainability,” *Technological Forecasting and Social Change*, vol. 173, Dec. 2021, doi: 10.1016/j.techfore.2021.121080.

[23] WEF, “The Future of Jobs Report 2020,” *WEF*, Oct. 2020.

[24] T. Stock and G. Seliger, “Opportunities of Sustainable Manufacturing in Industry 4.0,” *Procedia CIRP*, vol. 40, 2016, doi: 10.1016/j.procir.2016.01.129.

[25] WEF and BCG, “Net-Zero Challenge: The supply chain opportunity,” Jan. 2021.

[26] ODW, “Open Data Inventory 2020/21 Annual Report,” Feb. 2021. Accessed: May 30, 2022. [Online]. Available: https://opendatawatch.com/publications/open-data-inventory/

[27] Open Knowledge Foundation, “The Open Definition.” Accessed: May 30, 2022. [Online]. Available: https://opendefinition.org/

[28] R Core Team, “R: A language and environment for statistical computing,” *R Foundation for Statistical Computing*, 2022, Accessed: May 30, 2022. [Online]. Available: https://www.R-project.org/.

8.3 Attachment C – Paper *"Data Science for Industry 4.0 and Sustainability: A Survey and Analysis based on Open Data"*

Article

Data Science for Industry 4.0 and Sustainability: a Survey and Analysis based on Open Data

Hélio Castro [1], Filipe Costa [2], Tânia Ferreira, Paulo Ávila, Manuela Cruz-Cunha, Luís Ferreira, Goran D. Putnik, João Bastos [2,*]

[1] School of Engineering, Polytechnic of Porto; e-mail@e-mail.com
[2] Affiliation 2; e-mail@e-mail.com
* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials)

Abstract

In the last few years, the industrial, scientific and technologic fields have been subject to a revolution process of digitalization and automation called Industry 4.0. Its implementation has been successful mostly in the economic field of Sustainability, while the environmental field has been gaining more attention from researchers and recently. However, the social scope of Industry 4.0 is still somewhat neglected by researchers and organizations. This research aims to study Industry 4.0 and Sustainability themes through Data Science, by incorporating open data and open source tools to achieve Sustainable Industry 4.0. For that, is used a quantitative analysis through open source software such as Python and R to study the trends of Industry 4.0, Sustainability and open data in the world. The mains results show a positive trend in Industry 4.0 adoption through sustainable practices, mainly on developed countries, and a growing trend of openness of data, which can be positive for transparency of both industry and sustainability.

Keywords: Industry 4.0; Data Science; Sustainability; Social Sustainability Open Data; Open-Source

1. Introduction

This chapter provides an overview of the three main themes (Data Science and Open Data, Industry 4.0 and Sustainability) that are explored across the research, and a brief bibliographic review of each of those themes.

1.1. Data Science and Open Data

As we live in a world that constantly produces and consumes data, it is a priority to understand the value that can be extracted from it. Mikalef et al. (2019) consider data science and the big data domains as the next frontier for both practitioners and researchers as they embody significant potentials in exploiting data to sustain competitive advantage. Data science is an interdisciplinary field that supports and guides the extraction of useful patterns from raw data by exploring advanced technologies, algorithms and processes (Provost & Fawcett, 2013a). The actual extraction of knowledge from data is defined as data mining, and it can be applied to a broad set of business areas such as marketing, customer relationship management, supply chain management or product optimization (Bilal et al., 2016). Data-Science should be seen as domain that originates from the emergence of big data technologies with data management skills and behavioral disciplines (Saritha et al., 2021). Data-science and big data can be combined with co-creation and data-sharing technologies for organizations to leverage the creativity outside their own organizational boundaries (Runeson et al., 2021). Development and operation of software have become increasingly dependent on data (Gandomi & Haider, 2015) and this data can be more accessible to organizations and individuals through data-sharing and open-source technologies. Runeson (2019) highlight the need for the adoption of co-creation and collaboration principles to harness the innovation potential and to manage costs in the age of data.

Today, data volumes are exploding and not only is the rate of data generated per individual increasing, but so is the rate at which we share information. Lawmakers and organizations worldwide are trying to envision data's ownership future. Information remains largely centralized, but the trend is shifting toward a distributed and open model of data sharing (Hickin et al., 2021).

1.2. Open Data for Industry 4.0

As is described by Tim Hall (2020), one of the key drivers for the adoption of Industry 4.0 across the globe is the ability to use the power of data to revolutionize manufacturing. However, the manufacturing sector has been slow to benefit from these drivers evenly across different industries, enterprise sizes and geographies. Since most of Industry 4.0 technologies require substantial investments to be successfully implemented, the Economic factor is undeniably crucial for this adoption. Therefore, the differences in economic contexts of enterprises and countries can be immediately associated with the speed and rate of success of Industry 4.0 adoption but it cannot be considered the only one driver for it. Smart Factories and Smart Cities are another relevant study theme as technologic advancements and digitalization are changing how companies operate their business and organizations reshape communities. All those changes and advancements require big R&D investments and qualified researchers and workers. Since there are many economic challenges and difficulties to recruit the most qualified workers, the adoption of those technologies might be slow unoptimized for SMEs, which need to adapt to technologic changes in order to grow and compete.

1.3. Open Data for Sustainability

Wee et al. (2015) reiterate that there is the need for deeper research about Sustainability in Industry 4.0, since it has received very little attention from academics and researchers. In Kamble et al. (2018) framework of sustainability in Industry 4.0, the three sustainable outcomes that should be ideally accomplished from Industry 4.0 Technologies and Process Integration are economic, process automation and safety and environmental protection. Other models include open innovation and collaboration as guiding principles for sustainability in Industry. In this research, analyzing the progress towards accomplishing those goals through open data available is considered an overall evaluation of Sustainability across the three pillars. Since these are broad goals established not only in countries but also organizations and companies, a successful progress towards accomplishing these goals is also positive to accomplish Sustainability in Industry 4.0. It is important for U**N** members to collaborate across all established goals. Even more so because the Goal 17 itself – Partnerships for the Goals – focuses in evaluating member's progress towards Economic, Social and Environmental collaboration between them. For that reason, it is reasonable to assume that progressing in Goal 17 is essential to accomplish successful collaboration in the remaining goals. For the social factor of Sustainability, one of the main social issues regarding the digitalization and automation of Industry is how employment and skill requirements will be affected. The common sense regarding this issue is that automation eliminates the need for human workers, which will bring unemployment and social unsatisfaction. However, researchers such as Shet & Pereira (2021) actually believe that Industry 4.0 generates new job prospects in emerging domains of Science, Technology and Engineering. Those domains usually require a high level of skill and specialization than traditional jobs that leaves unskilled workers more vulnerable to the gradual increase in demand of qualified workers.

2. Research Methodology

The Research Design section aims to present the dissertation researcher design to the reader. In this section, all key design choices are detailed and justified logically, according to the dissertation theme.

Table 1 - Research Methodology

Research Design	**Method**

Research Type	Inductive and Quantitative
Research Strategy	1. Establishing the research themes 2. Collecting and Aggregating Data 3. Cleaning and Organizing Data 4. Data Analysis and Visualization 5.Results and Conclusions
Sampling Strategy	Probability Sampling within groups such as regions, countries, industries and enterprise size
Data Collection Methods	Datasets
Data Analysis Techniques	Programming through Open Source software tools such as Python and R

The adequate Research Type for this dissertation is the Inductive Quantitative type since it aims to explore quantitative data relevant to the research themes and afterwards take adequate conclusions and contributions, instead of pre-establishing hypothesis or theories about those subjects. The preferred data collection method for this research is by collecting and analyzing data from existing datasets. Those datasets, however, can only be useful if their content is fully open for being downloaded, modified and published by its providers, which is one of the prevalent characteristics of Open Design. To analyze the research themes, it was gathered data for the time period of September 2021 to May 2020 that was compiled into different datasets. Selecting the adequate techniques largely depends on the type, sample and data that were previously identified. For quantitative studies, the most frequently used techniques are descriptive statistics and inferential statistics (such correlation and regression analysis) (Hevner et al., 2004). The most prevalent techniques across the study will be frequency graphs and visualizations for inferential statistics that analyze correlations from selected variables. Another important aspect of the data analysis techniques is to use non-proprietary, open-source tools and software. The Open Source software tools used to analyze the data, both referenced in subchapter 2.2.2. of the bibliographic review, are Python and R. R is a free open-source programming language that provides an analytics computer environment. R provides a variety of statistical and graphical techniques that can be used by importing useful packages. These techniques can be used to handle raw data and retrieve information in order to have a sense on how the data is distributed or patterns that are masked (R Core Team, 2022). The R packages used were *arules, arulesViz* for rule association and RQDA for quantitative analysis. Python is currently the fastest growing programming language in the world, thanks to its open accessibility, ease-of-use, fast learning curve and its numerous high quality packages for data science and machine-learning. Together with R, Python provides great utility for identifying correlations between variables and creating powerful visualizations such as graphs, matrixes, plots or maps (Vallat, 2018). The main Python libraries used were *Matplotlib*, *Numpy*, and *Seaborn.*

3. Research Model

The conceptual model for this research intends to establish a framework of Data Science for Industry 4.0 and Sustainability moderated by an Open Design Approach that is supported by open concepts found in literature, such as Collaboration, Open Data and Non-Proprietary Tools. Figure 1 represents the conceptual model of Data Science for Industry 4.0 and Sustainability based on an Open Design Approach.

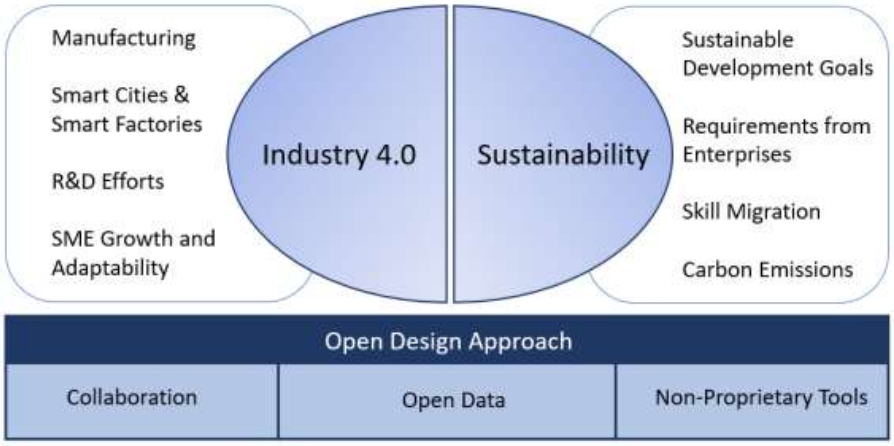


Figure 1 - Research Model

Industry 4.0 considers the research themes such as Manufacturing value to GDP, Smart Cities and Smart Factories, R&D Efforts for Innovation and SME growth and adaptability.
Sustainability considers economic, social and environmental themes, such as Collaboration for Sustainable Development Goals, Sustainability Requirements from Enterprises, Skill Migration and Carbon Emissions. Those themes are moderated by an Open Design Approach that is based on three concepts that should be ideally common across the research.

- Availability of Open Data for Decision-Making
- Collaboration between organizations, countries and enterprises
- Non-Proprietary and Open Source Tools

4. Results and Critical Analysis

This chapter presents the results from the data treatment from the selected datasets. For each research theme identified in the previous subchapter, are represented several relevant visualizations and its respective critical analysis in the context of the research.

4.1. Open Data for Industry 4.0

The first part of results and analysis obtain from the data treatment are representative of the Open Data for Industry 4.0 themes. As it is referred previously, this subchapter approaches Industry 4.0 themes such as Manufacturing, Smart Cities and Smart Factories, R&D efforts for innovation and SME growth and adaptability.

Manufacturing is one of the main sectors of Industry around the world and also one of the main adopters of Industry 4.0 (Thames & Schaefer, 2017). By analyzing available open data and using it alongside with other relevant variables that measure development such as a country’s GDP, this research intends to give a brighter perspective on the issues of the Research Model. The following map (Figure 2) represents the manufacturing value added to GDP around the world in 2020.

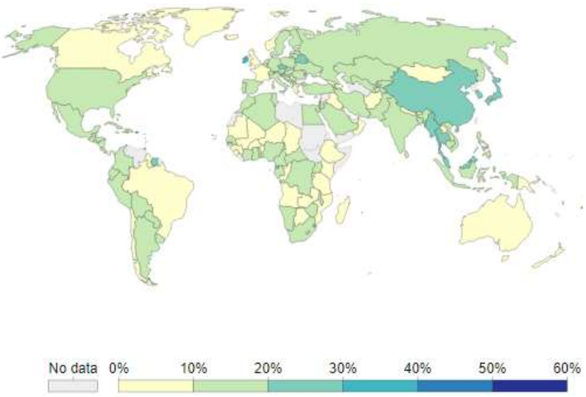


Figure 2 - Global Manufacturing value added to GDP in 2020, adapted from UN (2022)

China is one of the countries in the world in each a big share of its GDP is allocated in Manufacturing at around 40%. The majority of countries appear to have between 10% and 20% of manufacturing value added to GDP. The continents with larger share of countries that have less than 10% of their GDP value added from manufacturing are Africa and Oceania, while in Europe, North America and South America and Asia there few countries with less than 10% manufacturing value added to GDP. It appears that no country on Earth has more than 50% of its GDP value allocated to manufacturing.

A Smart City uses information and technology to improve operational efficiency, share information and provide better quality of life to its citizens and workers (Angelidou, 2014). Implementing Smart technologies and processes within factories and services also intends to promote economic growth, social integrity and environmental sustainability in industrial sectors through Industry 4.0 adoption, creating new jobs in the high-tech and creative industries (Angelidou, 2014). The dataset evaluates cities across six Smart Categories: Mobility, Environment, Government, Economy, People and Living. The conjunction of those scores translates to the Smart City Index of a city. Below, Figure 4 (a) ranks the countries by overall smart city scores and (b) presents a plot for correlations between all 6 variables with the overall score.

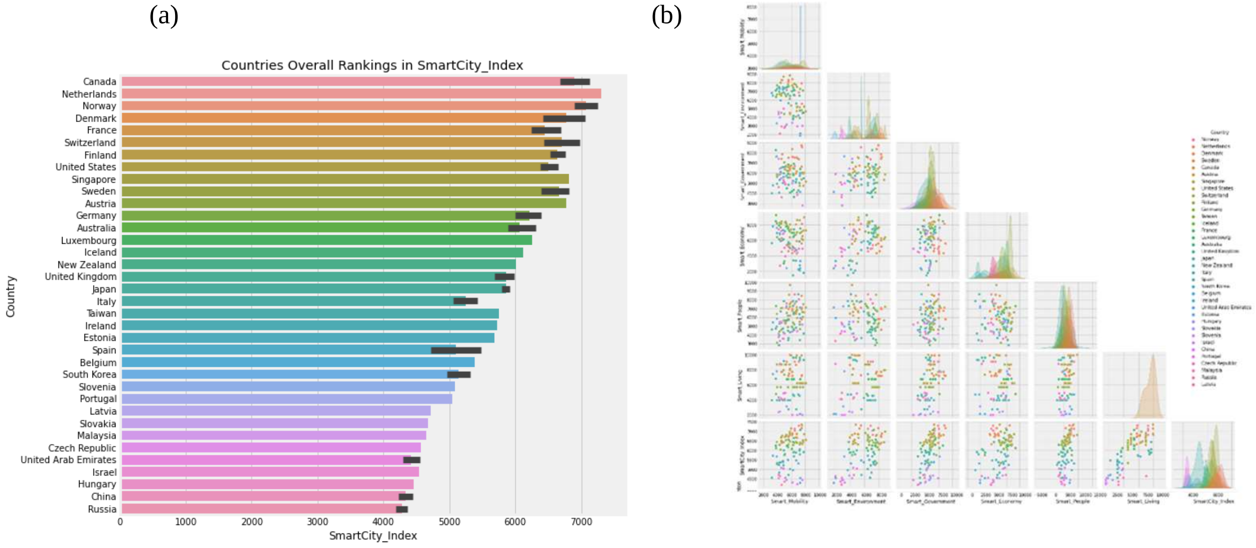


Figure 3 - (a) Countries Overall rankings in Smart City Index (b) Smart City categories correlations with the overall Smart City score

The countries with the highest overall scores are Canada, Netherlands, Norway, Denmark and France. The lowest scored countries are Russia, China, Hungary, Israel and the United Arab Emirates. The fact that a country has a high number of smart cities doesn't necessarily mean that the country itself has a high Smart score.

By analyzing each individual category with the overall Smart City Index, it looks like the key factors that seem to correlate more strongly with the overall index are Smart Living and Smart Economy. Since

Industry 4.0 is such a big driver for digitalization and automation in the global economy, it makes sense to accelerate the transition to a Smart Economy and Smart way of Living in developed and developing nations that seek develop their cities in technologic and sustainable way.

One of the main drivers of innovation, particularly in the technologic and industrial fields, is the financing of Research and Development (R&D) by enterprises, academic researchers and scientists (Mansfield & Lee, 1996). However, because of the uncertainty of the level of return and the payback period, this kind of investment is not equally accessible to different countries, industries and size of enterprises. Accessing which countries benefit the most from R&D investments from their enterprises and which industries allocate more expenditures to R&D might be a representation of the efforts to implement Industry 4.0

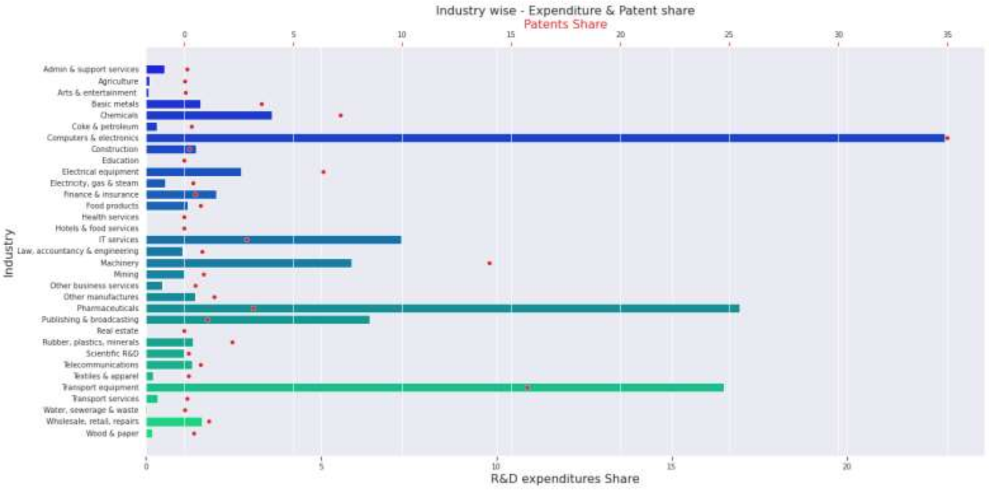


Figure 4 - Industry expenditure in R&D and patent share

It is possible to identify Computer and Electronics as the industry with the highest R&D expenditure share (close to 25%). As expected, it is also the industry with the highest patents share (35%). Pharmaceuticals now appears as the second industry with the highest R&D expenditure with around 17% of share, followed by Transport Equipment, IT Services and Publishing and Broadcasting with 16,5%, 7,5% and 6% share respectively. Surprisingly, the patent share doesn't follow that distribution so closely in those industries. The Pharmaceutical sector is only the seventh sector in Patent share even though is the second in R&D expenditure share. This might be caused by other factors such as regulation and difficulty in innovating the existing solutions. IT Services also issues a low Patent share compared to R&D expenditure share. Transport Equipment is another sector that has a much higher R&D expenditure share compared to Patent share. In the other hand, Machinery is the third sector with the highest patent share with almost 15%, even though it occupies the sixth position in R&D expenditure. Electrical Equipment, Chemicals and Basic metals are other sectors with much larger Patent share compared with R&D expenditure share.

4.2. Open Data for Sustainability

This subchapter approaches mainly the Social pillar of Sustainability since, as it is evidenced in the bibliographic work done in chapter 1, is arguably one of the most pressing sustainability issues and yet the one that receives less attention from researchers. One of the main issues regarding the relation between the adoption of Industry 4.0 and the future of work is job shortages. The increasing digitalization and automation of business and service tasks often lead to worries about permanent replacement of human labor force by machines. However, literature shows that that can be a misconception of the future of work. Shet & Pereira (2021) argue that Industry 4.0 can actually generate job prospects by creating new employment opportunities in emerging domains, like Science, Technology, Engineering and Mathematics. While technologic advancements and automation tend to minimize employment prospects in some sectors, it also brings about the simultaneous emergence of new business and services linked with economic growth and new markets, which leads to the rise of new job opportunities (Shet & Pereira, 2021). However, Shet & Pereira (2021) also warns that those jobs created by digitalization and automation also require a high level of skill, knowledge, competence and specialization that is not required by traditional jobs, leaving unskilled workers more vulnerable to the gradual increase in demand of qualified workforce.

To study this Social Sustainability, this research considered data from Skill Migration across different countries and industries, to compare supply and demand trends for skilled workers.

The first studied category, Soft Skills, includes important social skills such as problem solving, leadership, teamwork, communication, time management, persuasion and negotiation, which are essential skills for workers independent of location or industry sector. The analysis for the same three groups (US and China, G7 and BIC) are represented in Figures 6 (a), (b) and (c), respectively.

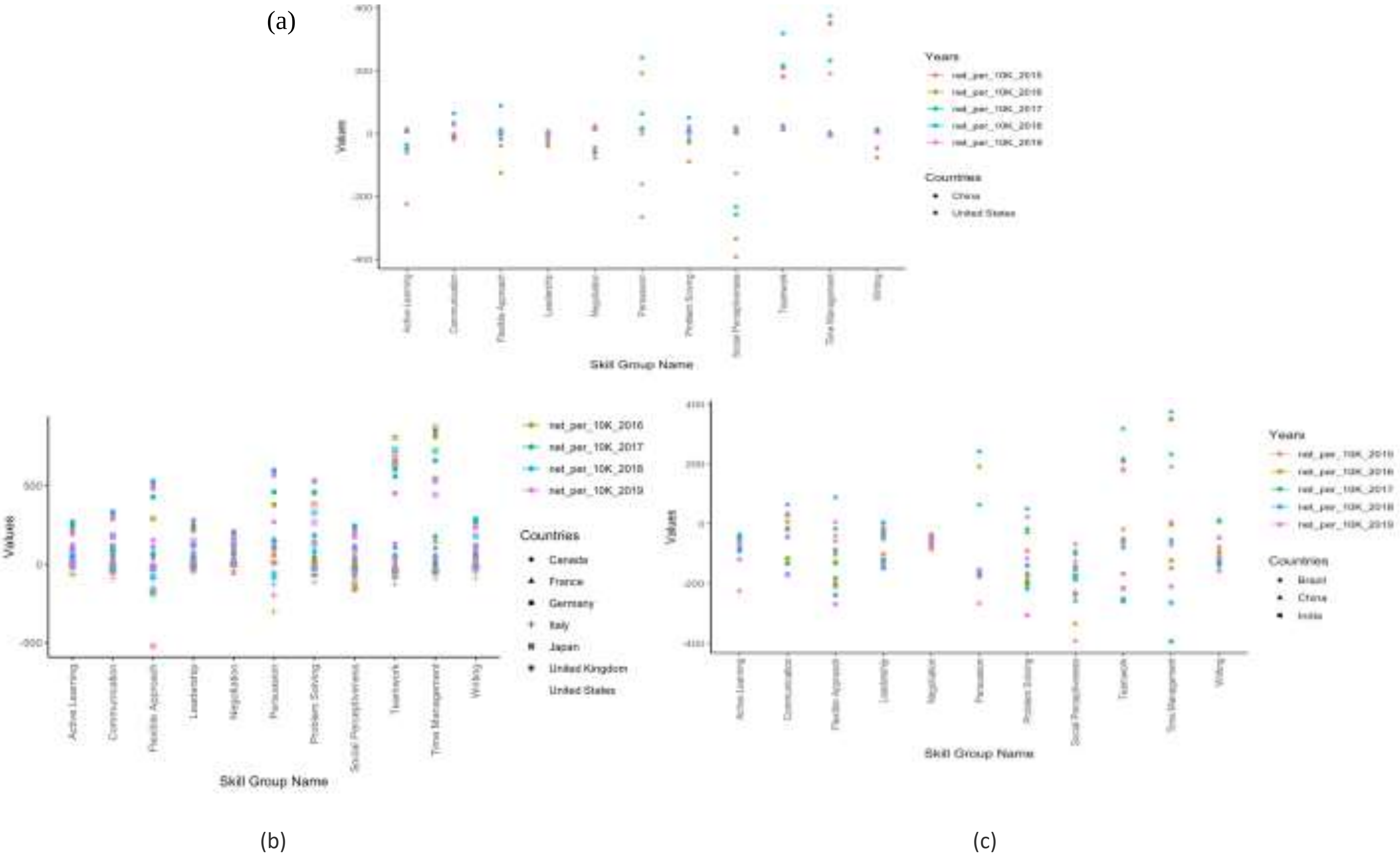


Figure 5 – Soft Skill Migration in (a) US and China (b) G7 countries (c) BIC countries

This visualization is highly relevant for this research by highlighting one of the main hypotheses of this research, which is that Social Sustainability is being neglected by countries and companies in comparison with the Economic and Environmental spectrums. As it is represented above, while Teamwork and Time Managements Skills are highly demanded and valued by the US and China, Social Perceptiveness have been having huge outflows of skilled workers in that field in both countries. Social skills such as perceptiveness are probably the most difficult skills to replace by technology, which accentuates the need for attention and valorization for social skills and issues.

When comparing the G7 countries to BIC countries, G7 still have overall positive demand for soft skills, while BIC have a negative migration trend, mainly from India and Brazil. Social perceptiveness is still the most neglected soft skill across those developing countries which is worrying in terms of future Social Sustainability

4.3. Open Design for Social Sustainability

As it was mentioned before, Social Sustainability is the arguably the pillar that gets less attention from researchers and organizations. One of the main objectives of this research is to contribute for the social cause exploring the concepts developed throughout the study. For that reason, it is important to understand if by leveraging information, technology, and tools, Open Data friendly countries can establish happier sustainable societies and serve as an example of social success for the rest of the globe. To study this theme, it is used the Open Data Scoring from ODIN (ODW, 2021) dataset that evaluates openness across different countries with scores from 0 to 100, considering the values for the year 2020. For the social sustainability perspective, it is used considered the World Happiness Report from 2020 and its respective dataset, which evaluates social happiness across different countries in a score from 0 to 10. This report is a survey published by the UN that ranks 156 countries reviewing their state of social happiness. To better understand how different groups of countries behave in this correlation, below (Figure 7) it is represented the correlation of Openness with Social Happiness, grouped into three established associations: the G7 (US, UK, Canada, France, Italy, Germany and Japan), the BRICs (Brazil, Russia, India and China) and the southeastern ASEAN (Brunei, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand and Vietnam).

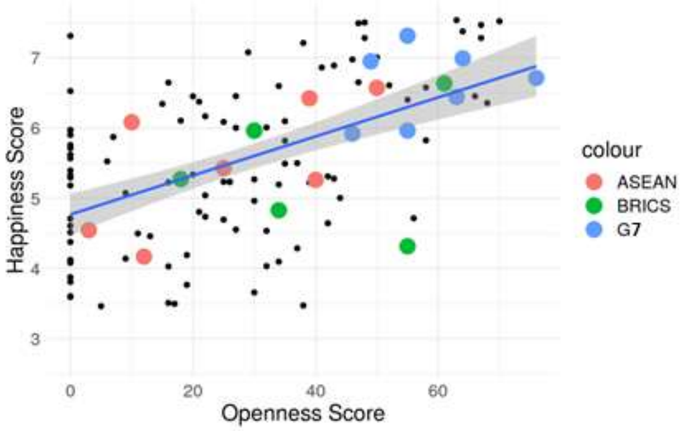


Figure 6 – Clustering for the correlation between Openness and Happiness in ASEAN, BRIC and G7 countries

G7 countries have much higher openness in their data policies, as well as bigger indices of social happiness than the BRIC and ASEAN countries, which places G7 clusters in the upper-right corner of the plot. The BRICs have a somewhat contradictory behavior since the cluster with the second highest openness score is also the one with the lowest happiness score, while the second lowest in openness is the second highest in happiness. Finally, as referred previously, the ASEAN countries can have clusters in the bottom-left corner, as well as clusters closer to the upper-right corner. Similarly to the G7, this group closely matches the trend line, which means that countries in this group with high openness also tend to have high social happiness.

5. Conclusions

One of the most significant trends in society is the sustainability concern and the openness of data should support sustainability awareness and mechanisms within Industry 4.0.

Considering this and the initial objectives of the research, it is possible to conclude that while Industry 4.0 adoption is still its initial stage, there is a positive trend in broad adoption. The same can be said about Sustainability awareness as a whole, even though there is still some negligence of the social aspect.

In terms of geographic exposure, the regions that seem to be adopting Industry 4.0 successfully and implementing sustainable practices are the US, China, G7 and developed countries. The industries that seem to be exploiting technology the most are computer electronics, pharmaceuticals, and other technologic sectors. In terms of enterprise size, bigger corporations still have much more resources and capacity to adopt technology faster and with more efficiency. On the other hand, SMEs have many growth constraints mainly by inability to invest as much in technology as big corporations.

In terms of openness of data, developed countries have much more openness of that currently. However, data openness is growing faster in developing countries. Either way, there is still room for increasing transparency and collaboration through increasing openness globally. Finally, by evaluating the results of open data for sustainable development, it was possible to conclude that openness can be considered positive for Social Sustainability, mainly in G7 and ASEAN countries, regions that showed high correlation between openness of data and social happiness.

6. Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

References

[1] P. Mikalef, M. Boura, G. Lekakos, and J. Krogstie, “Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment,” *British Journal of Management*, vol. 30, no. 2, pp. 272–298, 2019, doi: https://doi.org/10.1111/1467-8551.12343.

[2] F. Provost and T. Fawcett, “Data Science and its Relationship to Big Data and Data-Driven Decision Making,” *Big Data*, vol. 1, no. 1, pp. 51–59, 2013, doi: 10.1089/big.2013.1508.

[3] M. Bilal *et al.*, “Big Data in the construction industry: A review of present status, opportunities, and future trends,” *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 500–521, 2016, doi: https://doi.org/10.1016/j.aei.2016.07.001.

[4] B. Saritha, R. Bonagiri, and R. Deepika, “Open source technologies in data science and big data analytics,” *Materials Today: Proceedings*, Mar. 2021, doi: 10.1016/j.matpr.2021.01.610.

[5] P. Runeson, T. Olsson, and J. Linåker, “Open Data Ecosystems — An empirical investigation into an emerging industry collaboration concept,” *Journal of Systems and Software*, vol. 182, Dec. 2021, doi: 10.1016/j.jss.2021.111088.

[6] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” *International Journal of Information Management*, vol. 35, no. 2, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[7] P. Runeson, “Open Collaborative Data - using OSS Principles to Share Data in SW Engineering,” May 2019. doi: 10.1109/ICSE-NIER.2019.00015.

[8] R. Hickin, M. Bechtel, A. Golem, L. Erb, and R. Buscalno, “Technology Futures: Projecting the Possible, Navigating What’s Next,” Apr. 2021. Accessed: Dec. 26, 2021. [Online]. Available: https://www3.weforum.org/docs/WEF_Technology_Futures_GTGS_2021.pdf

[9] T. Hall, “The Role of Data in Industry 4.0,” May 20, 2020. https://industrytoday.com/the-role-of-data-in-industry-4-0/ (accessed Dec. 16, 2021).

[10] D. Wee, R. Kelly, J. Cattel, and M. Breuning, “Industry 4.0-how to navigate digitization of the manufacturing sector,” *Mckinsey & Company*, vol. 58, 2015.

[11] S. S. Kamble, A. Gunasekaran, and S. A. Gawankar, “Sustainable Industry 4.0 framework: A systematic literature review identifying the current trends and future perspectives,” *Process Safety and Environmental Protection*, vol. 117, Jul. 2018, doi: 10.1016/j.psep.2018.05.009.

[12] S. v. Shet and V. Pereira, “Proposed managerial competencies for Industry 4.0 – Implications for social sustainability,” *Technological Forecasting and Social Change*, vol. 173, Dec. 2021, doi: 10.1016/j.techfore.2021.121080.

[13] Hevner, March, Park, and Ram, “Design Science in Information Systems Research,” *MIS Quarterly*, vol. 28, no. 1, p. 75, 2004, doi: 10.2307/25148625.

[14] R Core Team, “R: A language and environment for statistical computing,” *R Foundation for Statistical Computing*, 2022, Accessed: May 30, 2022. [Online]. Available: https://www.R-project.org/.

[15] R. Vallat, “Pingouin: statistics in Python,” *Journal of Open Source Software*, vol. 3, no. 31, p. 1026, Nov. 2018, doi: 10.21105/joss.01026.

[16] L. Thames and D. Schaefer, “Industry 4.0: An Overview of Key Benefits, Technologies, and Challenges,” 2017, pp. 1–33. doi: 10.1007/978-3-319-50660-9_1.

[17] M. Angelidou, “Smart city policies: A spatial approach,” *Cities*, vol. 41, pp. S3–S11, Jul. 2014, doi: 10.1016/j.cities.2014.06.007.

[18] E. Mansfield and J.-Y. Lee, “The modern university: contributor to industrial innovation and recipient of industrial R&amp;D support,” *Research Policy*, vol. 25, no. 7, pp. 1047–1058, Oct. 1996, doi: 10.1016/S0048-7333(96)00893-1.

[19] ODW, “Open Data Inventory 2020/21 Annual Report,” Feb. 2021. Accessed: May 30, 2022. [Online]. Available: https://opendatawatch.com/publications/open-data-inventory/