IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 September 2025, accepted 29 September 2025, date of publication 6 October 2025, date of current version 15 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3618442

TOPICAL REVIEW

The Atlas of Data Science Research

SERGIO PICASCIA[1], (Member, IEEE), STEFANO MONTANELLI[1], SILVIA SALINI[2], AND STEFANO VERZILLO[3]

[1]Department of Computer Science, Università degli Studi di Milano, 20133 Milan, Italy
[2]Department of Economics, Management, and Quantitative Methods, Università degli Studi di Milano, 20122 Milan, Italy
[3]Joint Research Centre (JRC), European Commission, 21027 Ispra, Italy

Corresponding author: Stefano Verzillo (stefano.verzillo@ec.europa.eu)

ABSTRACT The origin and evolution of Data Science (DS) have been a subject of ongoing debate, with perspectives varying across disciplines. Understanding the development of this field requires a data-driven approach that systematically analyzes the scientific literature and provides a practical method for its exploration. In this paper, we present the "Atlas of Data Science Research" (DS-Atlas), an interactive visualization tool designed to study the landscape of the DS field. The DS-Atlas is built on a dataset of approximately 1.3 million scientific publications from the Elsevier Scopus database, leveraging Natural Language Processing, Large Language Models, and dimensionality reduction techniques to generate a semantic representation of the DS research. The DS-Atlas provides interactive operations to explore the dataset by allowing users to focus on specific areas, filter by keywords and/or time periods, and uncover thematic connections and research trends. Examples of concrete tasks that can be addressed by DS-Atlas are discussed to show how the proposed solution can support scholars in the data-driven analysis of the data science literature. As a further DS-Atlas contribution, the paper illustrates an analysis of the Data Science discipline in terms of geographical distribution of influential authors, institutions, and journal in the field. The DS-Atlas is publicly available online for exploration and testing.

INDEX TERMS Data science atlas, natural language processing, visual data exploration, empirical analysis.

I. INTRODUCTION

In 2015, David Donoho, a statistician at Stanford University, analyzed the "data science moment", and presented "*a vision of data science based on the activities of people who are 'learning from data', and described an academic field dedicated to improving that activity in an evidence-based manner*" [14]. In his work, Donoho provides a critical perspective on the evolution of Data Science and its relationship with statistics. He argues that Data Science is not merely an extension of traditional statistics, but a broader interdisciplinary field encompassing data collection, computing, machine learning, and real-world applications. In this view, the role of computational methods is presented as a key factor for (re-)shaping modern data science, with emphasis on large-scale data analysis and complex machine learning models as emerging research fields with respect to the past.

Origins and evolution of Data Science (DS) over time is still a lively discussion, and involve scholars and experts from multiple disciplines [5]. The analysis of the scientific DS literature through computational methods is sometimes proposed as a solution for investigating changes and trends in the field [2], [26]. This is the case for example of Nasution et al. (2020), that isolated a dataset of research papers on data science and applied text analysis techniques to propose possible applications and implications for the future of data scientists [19]. In such a kind of approach, we can say that *data science is used to analyze data science*, by stressing the benefits of relying on data-driven insights derived from combining statistical, computational, and mathematical methods to address a specific research question.

Data-driven approaches to data exploration are strongly influenced by the quality and reliability of the dataset in input.

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna D'Ulizia.

This is especially true in the case of data science, whose boundaries are differently perceived by people, continuously evolving, and variously interlaced across multiple disciplines. Along with the availability of a qualitative dataset, the reproducibility of data processing activities performed during the analysis must be ensured to enforce a transparent computational pipeline for possible comparison with solutions proposed by other scholars in a virtuous scientific exchange. In this context, we stress the crucial importance of interactive visualisation methods and tools in scientometric analysis, with the aim of enabling the final user to test and explore the results of the computations on the input dataset, so that anyone interested in the research can contribute with their own insights. Recent advancements in visualisation tools for scientometric analysis have been focusing on tracing the development of a research field through the exploitation of citation networks [7], [33] and metadata [8], [23], while text mining techniques have been employed for creating and visualizing maps of science [32].

In this paper, we present the "Atlas[1] of Data Science Research" (DS-Atlas), which aims at providing a comprehensive overview of the Data Science discipline, tracing its evolution in the last thirty years (from 1990 to 2023). The DS-Atlas has been built on top of a dataset of around $1.3M$ scientific publications taken from the Elsevier Scopus database.[2] The dataset is the result of a *thoughtful* selection process, where we had to decide what to consider as a data science paper and what to exclude, with the goal to provide - as much as possible - a reliable and shareable picture of the DS research in the considered time period. Inspired by the work of González-Márquez et al. [16], we release the DS-Atlas as an interactive, two-dimensional (2D) visual representation of DS research, that is generated by considering the main descriptive metadata of the dataset papers (i.e., title, abstracts, and author-assigned keywords), and by applying a computational pipeline based on Natural Language Processing (NLP) methods, transformer-based Large Language Models (LLMs), and dimensionality reduction techniques. The result is that the DS-Atlas provides a data-driven representation of the dataset contents, where the considered papers are projected in a common semantic space, in such a way that those discussing similar themes are more likely to be placed next to each other. The DS-Atlas offers multiple functionalities for exploring the underlying dataset. First, it is possible to focus on a specific area of the DS-Atlas to browse the papers therein contained. Then, it is possible to filter out the DS-Atlas to enlighten only publications characterized by specific keywords, and/or time lapses of interest. As a result, the DS-Atlas can be exploited to support several directions of analysis, like, for example, i) the thematic interconnections of field concepts, that are latently masked behind the variability of language and keywords, and ii) the research trajectories of the discipline, that are defined by the evolution of interests expressed by the scientific community over time.

The DS-Atlas is online for demo and testing at http://atlasds.islab.di.unimi.it.

The contribution of our work can be summarized as follows:

- a dataset of Data Science publications in the years from 1990 to 2023 based on descriptive metadata like title, abstract, and author-assigned keywords. The construction of the dataset is the result of a paper selection process that we performed according to our own idea of "data science discipline";
- a computational pipeline to build an Atlas of Data Science Research based on NLP methods, LLMs models, and dimensionality reduction techniques. The approach can be extended for application to further research domains given a dataset of scientific literature to consider;
- an analysis of Data Science research based on the DS-Atlas contents in terms of descriptive keywords, and geographical distribution of influential authors, institutions, and journal in the field.

The paper is organized as follows. In Section II, we provide a definition of Data Science, by also discussing our position with respect to previous work and definitions. Then, we introduce the dataset used for building the DS-Atlas (Section III), and we present the computational pipeline as well as the DS-Atlas finally generated with related descriptive keywords framed on the DS-Atlas for effective exploration (Section IV). Examples of possible use-cases based on DS-Atlas are discussed to show the benefits for scholars interested in analyzing the Data Science field (Section V). In Section VI, we illustrate a geo-based analysis of data science research. An empirical validation of our work is then proposed in Section VII. Limitations and concluding remarks are finally given in Sections VIII and IX, respectively.

II. DEFINING DATA SCIENCE

Data Science emerged as a research discipline in the late 20th century, particularly gaining momentum in the 1990s with the increasing availability of digital data and advancements in computing technology [24].

In the 1990s, the disciplinary areas mainly involved in Data Science included Statistics, Computer Science, and to some extent, Mathematics and domain-specific expertise [20]. Data mining and data analytics were key components of this early phase, focusing on extracting patterns and insights from structured and unstructured data [15].

Since that time, Data Science has evolved significantly. Today, it encompasses a broader range of disciplines, including Machine Learning, Artificial Intelligence, Big Data technologies, and Data Visualization [27]. The field has also

[1]The term "Atlas" usually refers to a collection of maps and geographical information, reflecting the legendary role of Atlas in the Greek mythology who was the guardian of the western lands and the far reaches of the Earth, so highlighting the connection between exploration, cartography, and the representation of the world (DS in our case).

[2]https://www.elsevier.com/products/scopus

expanded to include areas like Natural Language Processing (NLP), Deep Learning, and Predictive Analytics [26].

The impact of Data Science extends beyond its own disciplinary boundaries. It has revolutionized industries such as healthcare, finance, marketing, and transportation by enabling data-driven decision-making, personalized services, and predictive modelling [13]. In academia, Data Science has led to the emergence of new research areas and interdisciplinary collaborations, bridging the gap between theoretical knowledge and real-world applications.

However, Mathematics and Statistics still play a crucial role in DS by providing methods for data collection, summarization, analysis, and inference. The methodological contribution provided by these disciplines helps in understanding patterns, making predictions, and drawing conclusions from data. Similarly, Computer Science contributes by providing Algorithms, Programming Languages, and Computational Techniques for Processing and Managing large-scale data sets efficiently. Moreover, Knowledge Domain from fields like Business, Healthcare, Environmental Science, and Social Sciences provides context and insights into the data analysis process.

Before the term ''Data Science'' gained popularity, the field was often referred to as ''Data Mining'' or ''Data Analytics''. These terms emphasized the process of extracting insights and patterns from large datasets, typically using statistical and computational techniques [30]. Data Mining focused on discovering hidden patterns and relationships in data, while Data Analytics encompassed a broader range of activities including data exploration, visualization, and interpretation [25]. While these terms captured aspects of what would later be known as Data Science, they did not fully capture the interdisciplinary nature and breadth of the field as it is meant today.

John Chambers in 1993 made significant contributions to the field of Data Science through his work on statistical computing and software development [6]. Chambers was one of the principal developers of the *S* programming language, which later served as the foundation for the widely used statistical software *R*. His contributions laid the groundwork for modern data analysis and computational statistics, enabling researchers and practitioners to analyze data in a more effective way.

The real breakthrough year for Data Science was 2001. In 2001, Leo Breiman published a paper titled ''Statistical Modeling: The Two Cultures'', which sparked significant debate within the field of data analysis [4]. Breiman argued that there were two distinct approaches to data analysis: the statistical modeling culture, which focused on building complex models based on underlying assumptions, and the algorithmic modeling culture, which prioritized practical predictive accuracy through machine learning algorithms. This paper highlighted the evolving nature of data analysis and the need for interdisciplinary collaboration in fields such as Statistics and Computer Science.

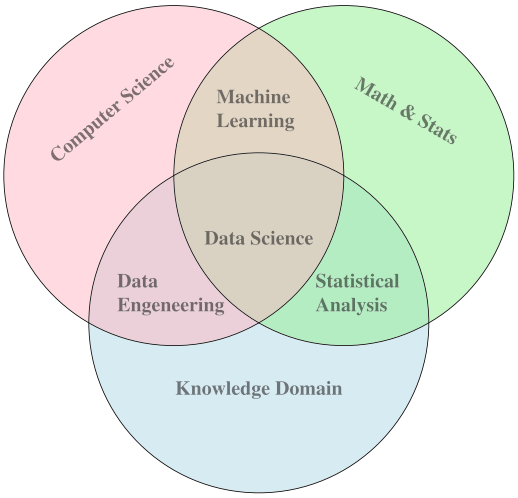

FIGURE 1. **"Data Science Venn Diagram", inspired by [11].**

Another influential paper published in 2001 was William S. Cleveland's ''Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics'' [9]. Cleveland proposed the term ''Data Science'' to encompass the expanding technical areas beyond traditional statistics. He advocated for a broader and more interdisciplinary approach to data analysis, emphasizing the importance of collaboration between statisticians, computer scientists, and domain experts. Cleveland's paper played a significant role in popularizing the term ''Data Science'' and shaping the direction of the field in the years to come.

A sort of synthetic view of the above positions is provided in the *Conway's Diagram of Data Science* that is also known as the *Data Science Venn Diagram* [11] (see Figure 1).

The diagram proposes an *ex-ante* perspective, and illustrates the theoretical/expected overlapping areas of expertise required in Data Science, namely Computer Science, Math and Statistics, and Knowledge Domain. The diagram confirms the positions expressed by the authors mentioned before, by supporting the idea that the integration of statistics, coding, and domain-specific knowledge is essential for a proper approach to Data Science, combining scientific rigor, computational skills, and understanding of the application context.

III. DS-ATLAS: DATASET DESCRIPTION

Our work is based on a dataset of Data Science publications collected from the Elsevier Scopus database. Our goal was to build a dataset of publications capable of covering the areas that characterize Conway's Diagram of Data Science.

The sources included in Scopus are either serial publications that have an ISSN (International Standard Serial Number), such as journals, book series, and conference series, or non-serial publications that have an ISBN (International Standard Book Number), such as monographs or one-off conference materials. The serial sources are classified using the ASJC (All Science Journal Classification) scheme.

TABLE 1. Composition of our dataset collected from Elsevier Scopus.

Scopus ASJC (code)	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19	2020-23	Total
Artificial Intelligence (1702)	15,023	23,552	30,277	49,271	73,246	101,125	140,705	433,199
General Decision Science (1800)	935	2,137	3,223	5,405	9,473	16,315	26,197	63,685
Decision Sciences (1801)	0	0	0	140	517	3,329	5,594	9,580
Information Systems and Management (1802)	4,221	5,684	8,972	15,976	27,104	35,148	39,195	136,300
Management Science and Op. Research (1803)	10,603	15,616	17,772	26,984	34,351	46,077	55,609	207,012
Statistics, Probability and Uncertainty (1804)	7,821	12,802	15,004	21,428	27,021	34,541	36,682	155,299
Statistics and Probability (2613)	21,796	27,388	30,803	49,348	62,130	87,659	81,486	360,610
Total	60,399	87,179	106,051	168,552	233,842	324,194	385,468	1,365,685

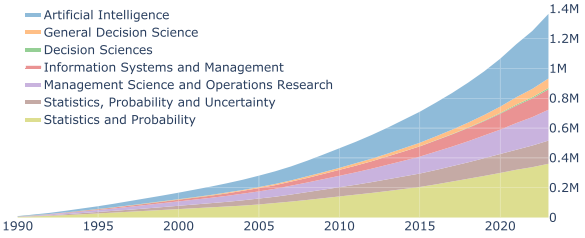

FIGURE 2. Dataset cumulative by subject area along time.

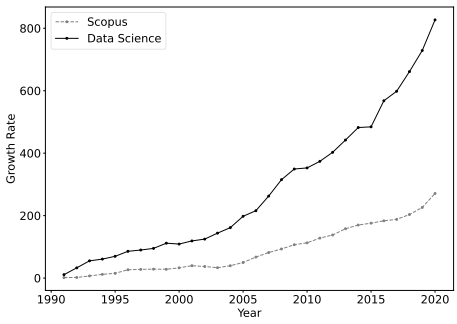

FIGURE 3. Growth rate of Scopus publications vs. Data Science publications using 1990 as basis. Source [29].

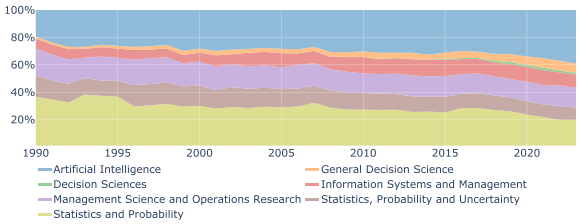

FIGURE 4. Weight of ASJC fields along time.

A source is associated with one or more ASJC codes by in-house experts of Scopus according to the scope, title, and content.

As such, we selected a subset of ASJC codes that we decided to consider as coherent with Conway's Diagram and pertinent to the Data Science field.[3] Then, we queried the Scopus database by collecting the metadata of publications belonging to all sources associated with the selected list of ASJC codes in the time interval spanning from 1990 to 2023. The collected metadata include **title**, **abstract**, **authors** with corresponding affiliations and countries, **year**, **source** (i.e., the journal/conference where the paper has been published), and **author-assigned keywords**.

The dataset has been acquired by employing the Elsevier API[4] (Application Programming Interface) that enables programmatic access to metadata and abstracts from scholarly journals and conferences, as indexed by the Scopus database. In particular, the publication metadata have been retrieved through the **scopus_search** interface by specifying a time interval as well as a list of subject areas of interest. In the download, we exploited the **complete view** modality of Scopus,[5] and only publications equipped with an abstract have been considered, meaning that around 4% of the collected publications have been excluded from the analysis.

A. DESCRIPTIVE EVIDENCE

Table 1 provides a summary view of the dataset contents according to the ASJC code and grouped into 5-year intervals.

[3]The complete list of ASJC codes is available at https://service.elsevier.com/app/answers/detail/a_id/15181/

[4]https://dev.elsevier.com/

[5]https://dev.elsevier.com/sc_search_views.html

The majority of papers are classified under ASJC fields such as Artificial Intelligence, Statistics and Probability, Management Science, and Operations Research. The relative shares within the DS domain of these fields have evolved dynamically over the 30-year period under study. Figure 2 offers additional insights on how the considered ASJC codes evolved in time according to the number of publications.

Overall, the total number of papers annually published in journals indexed in Scopus and belonging to the selected ASJC fields increased approximately by a factor of seven, from around 200, 000 in 1990 to 1.4 million in 2023. This significant growth is especially noteworthy when compared to the annual growth rate of the entire Scopus database (across all fields), as illustrated in Figure 3.

The growth rate of the considered ASJC fields has surpassed the growth rate of the entire Scopus database by approximately four times, revealing that data science as a whole is a distinct and rapidly expanding discipline over the past three decades. Moreover, Figure 4 highlights how the relative share of different sub-fields within DS has evolved over time.

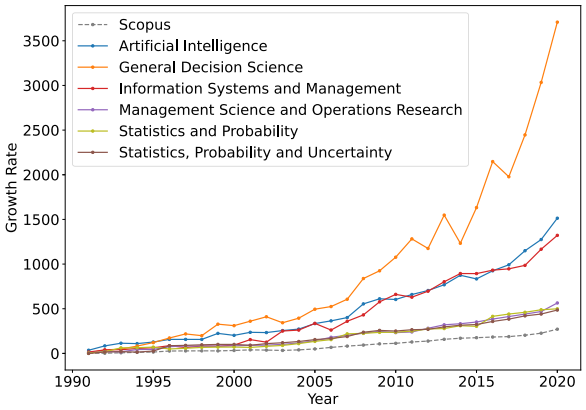

FIGURE 5. Growth rate of Scopus publications vs. the publications of each considered ASJC field using 1990 as basis.

TABLE 2. Distribution over ASJC codes for publications containing the "data science" bigram (publications NOT IN our dataset).

ASJC	Discipline	Papers	%
17	Computer Science	3,275	25,8
22	Engineering	1,552	12,2
26	Mathematics	1,218	9,6
27	Medicine	1,162	9,1
33	Social Sciences	957	7,5
13	Biochemistry, Genetics and Molecular Biology	529	4,2
31	Physics and Astronomy	518	4,1
25	Materials Science	505	4,0
23	Environmental Science	361	2,8
16	Chemistry	303	2,4
19	Earth and Planetary Sciences	275	2,2
15	Chemical Engineering	269	2,1
11	Agricultural and Biological Sciences	223	1,8
14	Business, Management and Accounting	220	1,7
	Others	1,337	10,5
	Total	12,704	

In the early 1990s, subfields related to Statistics, Probability, and Uncertainty accounted for roughly 50% of all published DS papers. However, this share has since declined to 30%, with a corresponding increase in the prominence of Artificial Intelligence and Decision Sciences. Figure 5 plots the growth rate of each considered ASJC field over the past three decades, revealing that Decision Sciences have experienced the steepest rise in the number of papers, particularly in the last 15 years. However, it is important to consider that the differences in growth rates between fields - such as Decision Sciences compared to AI and Information Science - may also be influenced by non-negligible differences in their relative sizes.

B. ASJC VALIDATION

A recent attempt to study the emergence of "Data Science" as a distinct scientific field has been made by Nasution et al. in [19]. This research is based on a dataset of Scopus publications where the bigram "data science" appears in the title, abstract, or keywords.

In the following, we describe what we performed with the aim to validate our choice of ASJC fields as a proxy of DS. As done in [19], we downloaded from Scopus the publications with the bigram "data science" in the title, abstract, or keywords. We obtained a dataset of 9, 234 papers, of which 2, 188 belong to our ASJC fields (i.e., publications IN our dataset), while 7, 046 do not (i.e., publications NOT IN our dataset). Table 2 shows the distribution of the publications containing the "data science" bigram that are NOT IN our dataset according to the Scopus ASJC codes[6]

Interestingly, around 75% of papers in Table 2 are scattered across a wide range of macro fields, including Medicine, Chemistry, Environmental Sciences, Agricultural and Biological Sciences, and Social Sciences, among others. Most likely, these papers are about research findings where a data science method is applied to specific domains. This is indirect support for our choice to exclude these codes from the "core" ASJC fields that are representative of DS. By focusing on the 3, 275 papers of Computer Science (ASJC code 17), we note that most of them are concentrated in sub-fields related to Software (1712), Computer Science Applications (1706), and General Computer Science (1700). In particular, the DS papers in ASJC codes 1700, 1706, and 1712 are 1,260, 854, and 741, respectively. From 1990 to 2023, the overall number of papers in ASJC codes 1700, 1706, and 1712 are 1,521,950, 2,978,791, and 1,708,305, respectively. Including these additional sub-fields as part of our ASJC "core" list of codes characterizing DS would have implied the inclusion (and download) of two large sets of publications, with relatively few papers directly related to DS, and a significant amount of noise in the data.

Table 3 shows the distribution of the publications containing the "data science" bigram that are IN our dataset according to the Scopus ASJC codes.

TABLE 3. Distribution over ASJC codes for publications containing the "data science" bigram (publications IN our dataset).

ASJC	Discipline	Papers	%
18	Decision Sciences (1800, …, 1804)	1,557	32,9
17	Computer Science (1702)	1,040	22,0
26	Mathematics (2613)	902	19,1
14	Business, Management and Accounting	394	8,3
22	Engineering	302	6,4
33	Social Sciences	285	6,0
20	Economics, Econometrics and Finance	76	1,6
27	Medicine	39	0,8
12	Arts and Humanities	27	0,6
28	Neuroscience	20	0,4
31	Physics and Astronomy	18	0,4
23	Environmental Science	18	0,4
13	Biochemistry, Genetics and Molecular Biology	17	0,4
32	Psychology	12	0,3
30	Pharmacology, Toxicology and Pharmaceutics	7	0,1
36	Health Professions	6	0,1
11	Agricultural and Biological Sciences	5	0,1
19	Earth and Planetary Sciences	4	0,1
10	Multidisciplinary	2	0,0
	Total	4,731	

[6]It is worth noting that a journal can be associated with more than one ASJC code; therefore, the total number of papers in Tables 2 and 3 reflects the number of paper-field associations, rather than the number of unique papers retrieved from Scopus.

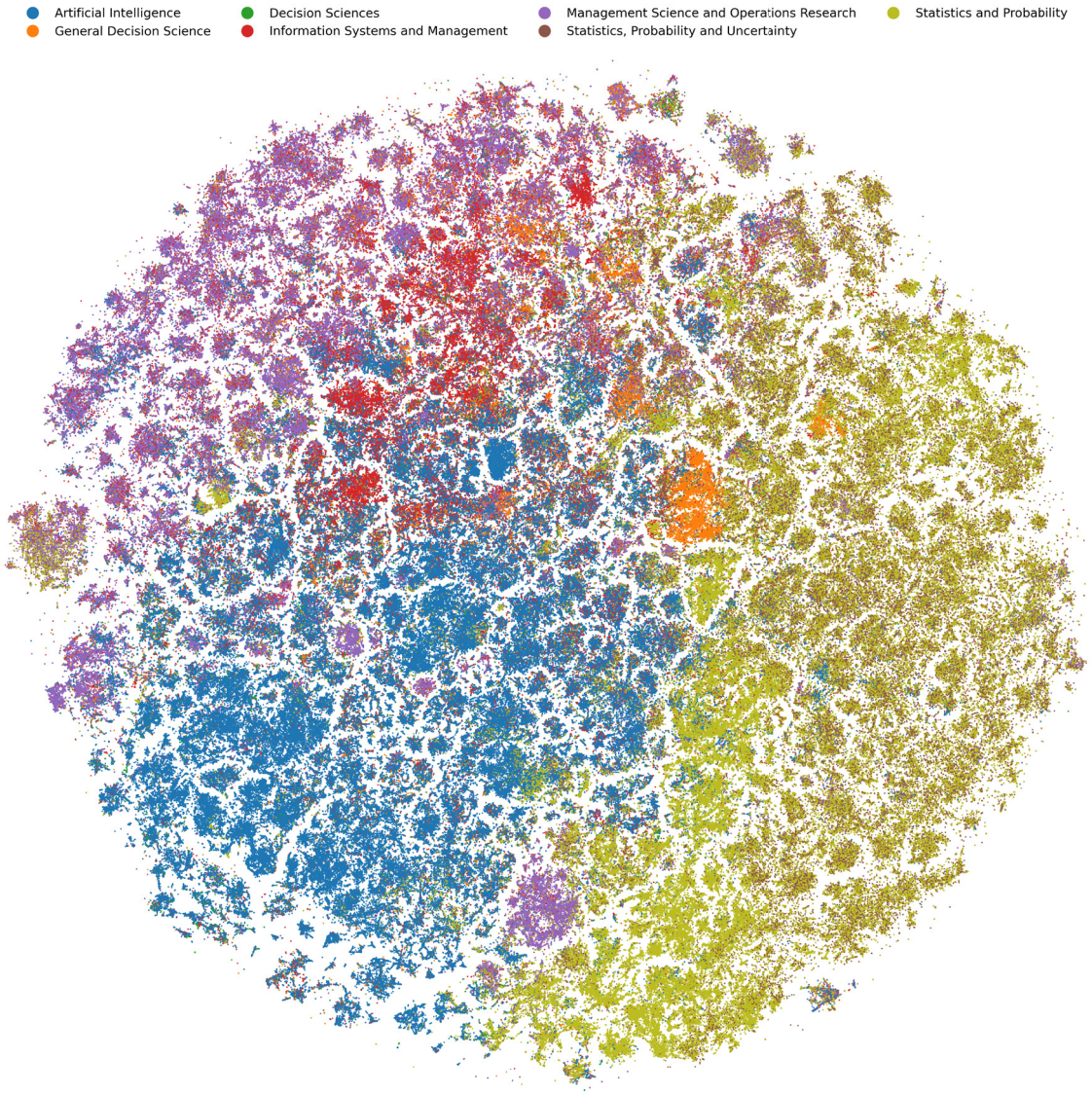

FIGURE 6. The DS-Atlas of DS publications from Elsevier Scopus with colour-coded ASJC fields.

Excluding the ASJC fields in our dataset (i.e., the first three rows in Table 3), it is interesting to note that only a few papers are associated with additional ASJC codes, such as Business, Management, and Accounting, Engineering, and Social Sciences. This result is consistent with Conway's Diagram of Figure 1, thus confirming the appropriateness of our choice of ASJC codes.

IV. DS-ATLAS: CONSTRUCTION AND EXPLORATION

The DS-Atlas is built on top of the dataset of Scopus publications described in Section III, and it is released as a 2D visualization tool where the position of each paper is determined by its abstract, so that papers with similar content are placed close to each other.

The DS-Atlas construction follows a two-stage processing pipeline based on i) *publication embedding*, and ii) *dimensionality reduction*.

A. PUBLICATION EMBEDDING

The first stage involves transforming each publication into a high-dimensional dense vector, i.e. an embedding. For embedding construction, we decided to consider the publication abstracts as an effective synthesis of the publication contents. The choice of using abstracts is also motivated by the need to have a roughly-equivalent length of text for each document. Embeddings have the role to project publications in a common semantic space, where distances between papers/vectors reflect the differences in abstract contents.

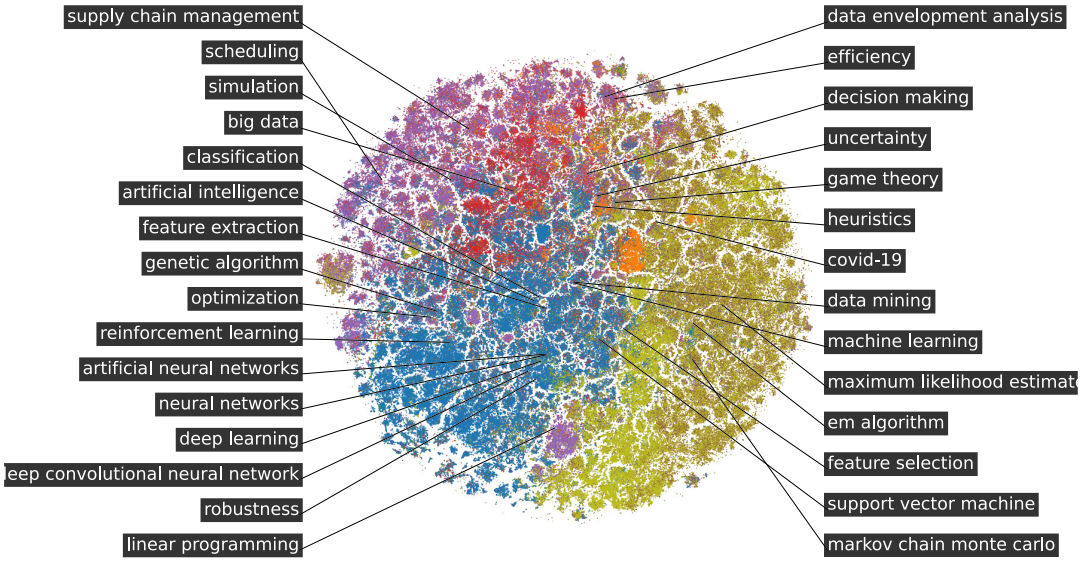

FIGURE 7. The DS-Atlas labeled with 30 most frequent keywords.

B. DIMENSIONALITY REDUCTION

The high-dimensional embeddings are not immediately suitable for a 2D visualization. Therefore, the second stage of our pipeline performs a dimensionality reduction of the given embeddings. As a result, each publication is denoted as a point on a two-dimensional semantic space that is convenient for the DS-Atlas representation.

The resulting DS-Atlas is shown in Figure 6.

In the DS-Atlas, the papers of an ASJC field are denoted by a specific color. Colors allow to smartly observe the position of ASJC fields as a whole. Statistics and Probability (2613) and Statistics, Probability and Uncertainty (1804) are mostly on the right side of the diagram. On the left side, we have Artificial Intelligence (1702) in the bottom left, while Information Systems and Management (1802) and Management Science and Operations Research (1803) lie in the top-left part. Finally, General Decision Sciences (1800) and Decision Sciences (1801) are located in the top-centered area.

On top of the DS-Atlas, a *thematic layer* derived from the author-assigned keywords has been developed to enforce an effective and self-explanatory exploration of the DS-Atlas contents. The keywords are placed on the DS-Atlas by using the same semantic space of the publications. As such, a keyword should be descriptive/representative of the papers in its neighborhood on the atlas. Technical details on the DS-Atlas construction including the thematic layer are provided in Appendix A.

The DS-Atlas labeled with the 30 most frequent keywords is shown in Figure 7.

The DS-Atlas is designed to support the following exploration operations.

- **Publication filtering**. This operation has the goal to focus the exploration on selected points (i.e., publications) according to a given criterion like a keyword, an author, or a year of interest. Specifying a filtering condition, the DS-Atlas shows the matching points by gray-colouring the non-matching ones and by obtaining a tailored DS-Atlas visualization limited to the publications of interest.
- **Spacial insight**. This operation has the goal to focus the exploration on a specific region of the DS-AtlasṠelecting an area of interest (i.e., a rectangular space), the DS-Atlas shows two summary statistics for the publications belonging to the selected area, such as the percentage of publications belonging to the different subject areas (i.e., ASJC), together with the 10 most common keywords used to characterize those publications in the original dataset.

It is worth noting that both the operations can be combined with two additional functionalities. First, an additional criterion can be specified to limit the visualization to publications of selected subject areas (i.e. ASJC). This way, the result of a filter or insight can be further focused on exploring publications of specific subjects. Second, the punctual exploration of single publications is also possible. This way, by passing over a point, it is possible to view the metadata of the corresponding publication (i.e., Scopus identifier, title, authors, year of publication). Guidelines

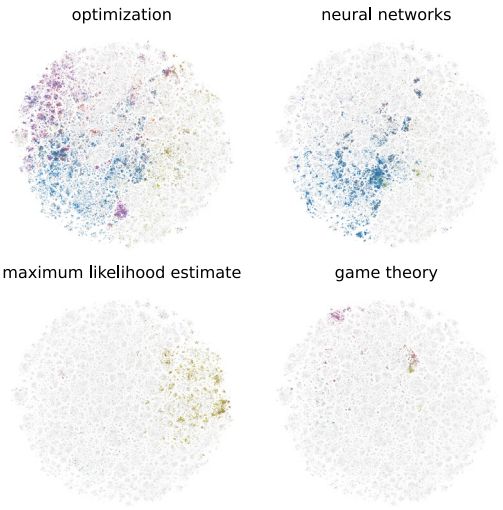

FIGURE 8. Examples of keyword distribution on the DS-Atlas.

on the practical use of filtering and insight operations are provided in Appendix C.

1) EXAMPLE ON FILTERING OPERATION (USING KEYWORDS)

Keywords can be used as a filtering operation over the DS-Atlas contents. For example, Figure 8 shows the distribution on the DS-Atlas of some featuring keywords of the DS field, such as 'optimisation', 'neural networks', 'maximum likelihood estimation', and 'game theory'. The combination of filters on keywords and ASJC fields con be used to highlight where keywords are more frequently used with respect to the subject areas of the publications. For instance, it is interesting to note that the distribution of the keyword 'maximum likelihood estimation' mostly overlaps with the ASJC field of *Statistics and Probability*, while the keyword 'optimisation' is more scattered across the various subject areas. As expected, the keyword 'neural networks' strongly overlaps with the ASJC field of *Artificial Intelligence* and *Decision Science*.

2) EXAMPLE ON FILTERING OPERATION (USING YEARS)

In this example, the filtering operation is employed to observe the publications of a specific year. In Figure 9, we show the papers of 2021 in General Decision Sciences (1800). This filtering highlights a peculiar distribution of papers influenced by their abstract similarity. Consider the orange cluster of points on the right-hand side of the diagram. This cluster represents a group of publications mostly focused on environmental monitoring. It is interesting to note that this cluster originated around 2012 (before it was not visible at all), then gained popularity, finally becoming so evident in 2021. As a further example, consider the papers in Figure 9b from 2021 in Management Science and Operations Research (1803). A small cluster of papers on the left part of the map

(a) (b)

FIGURE 9. General Decision Sciences (left), and Management Science and Operations Research (right) - 2021.

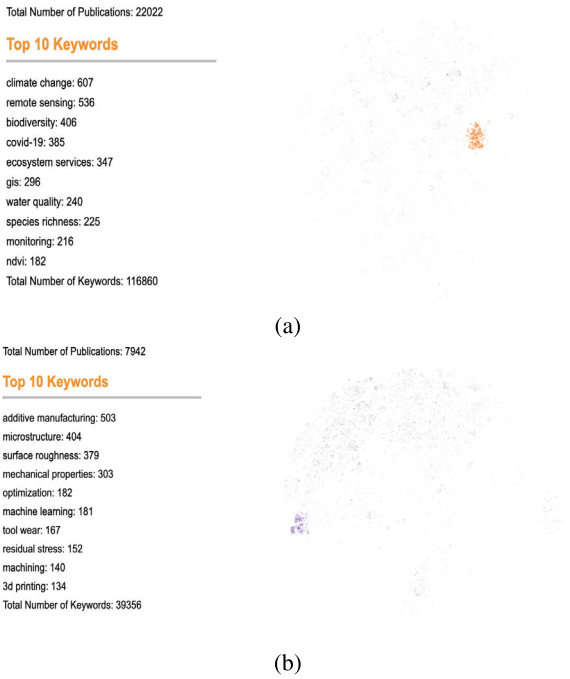

FIGURE 10. Examples of insight operations over the DS-Atlas.

is about additive manufacturing and it was absent in 2017 but grew in popularity in subsequent years.

3) EXAMPLE ON INSIGHT OPERATION

Consider the orange cluster of publications on environmental monitoring in the General Decision Sciences (1800) field (see Figure 9 of the previous example). The insight operation can be invoked to enforce a deep exploration of that region of the DS-Atlas. The insight reveals a coherent research area centered around environmental assessment and monitoring technologies (Figure 10). The most frequent keywords include "climate change", "remote sensing", "biodiversity", and "covid-19", indicating a strong focus on environmental monitoring applications that gained particular relevance in recent years. Similarly, when applying the insight operation to the cluster on additive manufacturing in Management Science and Operations Research (1803) (see

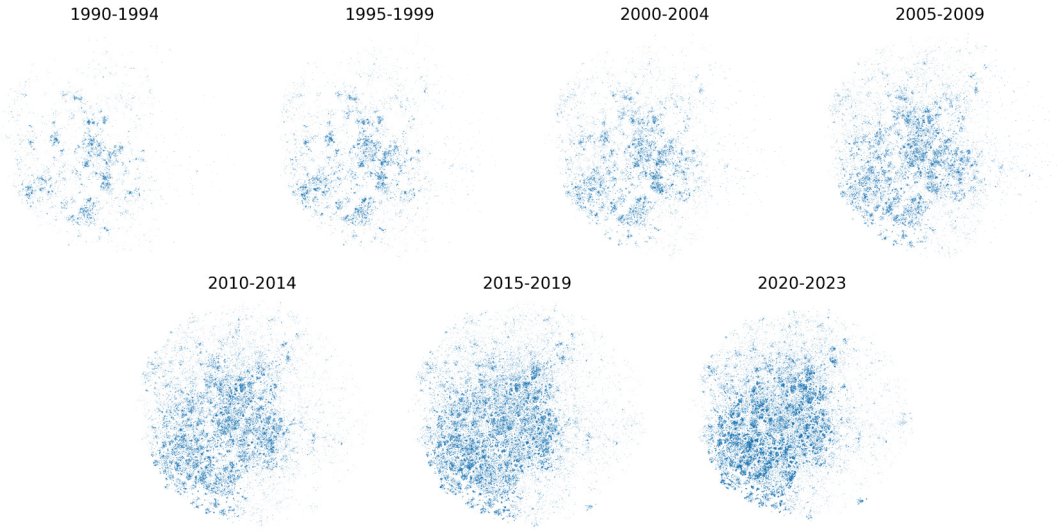

FIGURE 11. Papers in the Artificial Intelligence (1702) field over the considered seven lustres.

Figure 9b), the analysis reveals a highly specialized research domain (Figure 10b) with ''additive manufacturing'' as the dominant keyword, followed by technical aspects such as ''microstructure'', ''surface roughness'', and ''mechanical properties''.

The insight operation enables researchers to quickly assess the thematic density and specialization level of different regions, fostering the identification of emerging research trends and the discovery of interdisciplinary connections that might not be immediately apparent from traditional keyword-based searches. The last example on the insight operation also provides a sort of ''spot'' validation on the quality of results produced by our two-stage pipeline, thus demonstrating that the DS-Atlas is effectively capable to place publications with similar research themes close together. A systematic validation experiment on the quality of the semantic clustering performed for the DS-Atlas construction is presented in Appendix B.

V. WHAT DS-ATLAS CAN DO FOR SCHOLARS

The DS-Atlas can be exploited to visually explore and query the underlying dataset of DS publications with the ultimate goal to effectively analyze the size and magnitude of different subject areas and their mutual interactions. As a natural opportunity, the DS-Atlas can be used to enforce a **literature analysis** of a given topic of interest. Starting from a target publication or keyword, that are precise points on the DS-Atlas it is possible to explore similar and topic-related publications by viewing the points in the neighbourhood of the initial target. Both filtering and insight operations can be used to interactively refine the exploration. For example, filtering on years can refine the exploration considering publications of a specific interval of time. More articulated exploration modalities can be envisaged by combining filtering and insight operations supported by the DS-Atlas. In the following, two examples of possible use-case scenarios of DS-Atlas are discussed to show the benefits for scholars interested in analysing the data science domain.

A. TREND ANALYSIS

The DS-Atlas can be exploited to observe trends over time for subsets of DS publications. By specifying an interval of years in the filter, the DS-Atlas allows to observe the year-by-year evolution of published papers on the available subject areas with possible focus on a subset of them. As an example, Figure 11 shows the evolution of Artificial Intelligence (1702 ASJC code) over the considered time period aggregated by lustres. The sequence of diagrams shows the expansion of AI along time. Initially, 'artificial intelligence' was a specific 'topic' in some disciplines (i.e., scattered distribution of points in the space), and it became a relevant research topic in the last three lustres when the intertwining with other disciplines increased as well (i.e., massive distribution of groups of points in the space).

B. KEYWORD ANALYSIS

The DS-Atlas can support the analysis of keywords used in the dataset and how keywords changed during time. This is useful to analyze how the research vocabulary evolved in a certain field. This kind of analysis can reveal how a keyword/topic has changed its importance. A keyword increases/decreases the number of occurrences from one year to another in a considered timeframe, denoting a possible gain/loss of relevance on that interval.

In Figure 12, we show how the author-assigned keywords have changed over time in the dataset. On the left, we have the top-20 keywords in papers of 1990-1995 (from top to bottom in terms of frequency); on the right, we have the top-20 keywords in papers of 2020-2023. In between the two extremes, we can observe the shift of keywords in terms of

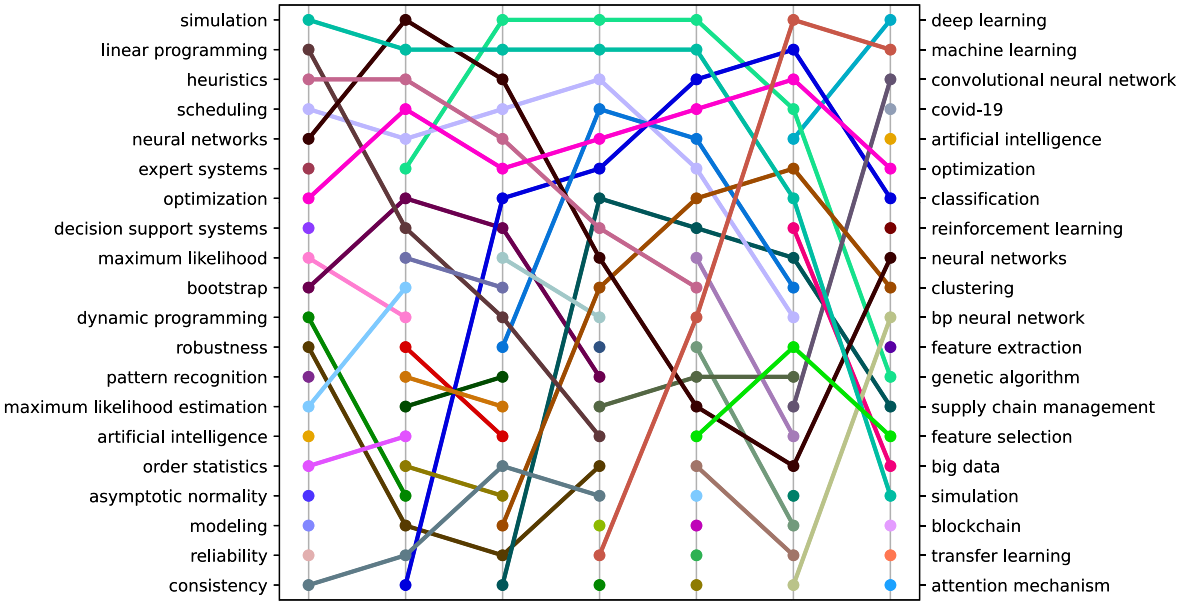

FIGURE 12. Top-20 keywords over time in the DS dataset. 1900-1995 on the left; 2020-2023 on the right.

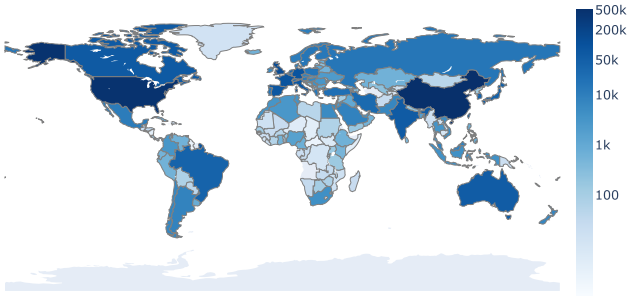

FIGURE 13. Number of Q1 publications per country in the Scopus dataset (log scale).

frequency in correspondence with the intermediate lustres (i.e., 1995-2000, 2000-2005, 2005-2010, 2010-2015, 2015-2020). The figure not only shows the general prevalence of keywords in the Data Science literature, but also provides insights into their temporal evolution, highlighting shifts in focus over the decades. In particular, 'deep learning', 'neural networks', and 'reinforcement learning' show rapid growth post-2012, coinciding with significant breakthroughs such as the introduction of deep convolutional networks and the widespread availability of specialized hardware to train complex models. For example, the term 'deep learning' saw minimal diffusion prior to 2012 but experienced exponential growth afterward, becoming one of the dominant topics in the late 2010s. Similarly, 'big data' emerged prominently after 2010, driven by advances in storage and collecting resources, cloud computing, and the growing volume of data from sources such as social networks and IoT devices. Another relevant temporal spike is observed for 'COVID-19', which entered the Data Science research landscape abruptly in 2020, reflecting the scientific community's willingness to offer their scientific skills to address urgent social challenges. In contrast, traditional optimization and heuristic techniques such as 'genetic algorithm', 'linear programming', and 'data envelopment analysis', demonstrate relatively stable or declining trends after peaking in the early 2000s. For example, the usage of 'genetic algorithm' plateaued post-2010, suggesting a shift towards more scalable and flexible machine learning-based methods. Similarly, statistical techniques like 'maximum likelihood estimation' and 'EM algorithm' show steady but less pronounced growth, overshadowed by the rising popularity of recently developed machine learning methodologies. These temporal patterns reflect the evolving priorities within the Data Science field: from classical optimization and statistical methods prevalent in the 1990s and early 2000s to modern AI-driven techniques dominating recent years. The figure thus captures not only the current research landscape but also the dynamic historical trajectories shaping the field.

VI. GEOGRAPHIC DISTRIBUTION OF INFLUENTIAL INSTITUTIONS, AUTHORS AND JOURNALS

The DS-Atlas and the underlying dataset of Scopus publications can support the systematic analysis of the Data Science field. In this section, we discuss our intuitions derived from the use of DS-Atlas by considering most influential authors, institutions with related countries, and journals.

Figure 13 illustrates the relative dominance of two leading countries in the DS discipline, namely the United States and China. The map presents the number of publications in the top-tier journals (Q1, representing the top 25% of journals by impact factor in each subfield), using a logarithmic scale. This diagram clearly shows that the US and China lead

TABLE 4. Top-30 institutions by number of publications in the Scopus dataset (Q1 journals).

Country	Institution	# Publications
CHN	Tsinghua University	10272
CHN	Harbin Institute Of Technology	9825
CHN	Xidian University	9060
CHN	Huazhong University Of Science And Technology	8539
CHN	University Of Chinese Academy Of Sciences	8182
CHN	Shanghai Jiao Tong University	7875
SGP	National University Of Singapore	7677
CHN	University Of Electronic Science And Technology Of China	7630
HKG	Hong Kong Polytechnic University	7514
CHN	University Of Science And Technology Of China	7054
CHN	Tianjin University	6899
CHN	Xi'An Jiaotong University	6896
CHN	Beihang University	6777
CHN	Northwestern Polytechnical University	6581
USA	University Of Michigan, Ann Arbor	6430
CHN	Dalian University Of Technology	6356
HKG	City University Of Hong Kong	6204
USA	Stanford University	6132
USA	University Of Washington	5803
CHE	Eth Zürich	5762
USA	Carnegie Mellon University	5687
USA	Texas A&M University	5479
USA	University Of California, Berkeley	5430
GBR	University Of Oxford	5406
CHN	Beijing Institute Of Technology	5405
GBR	Imperial College London	5364
CHN	Wuhan University	5361
BEL	Ku Leuven	5335
CHN	Chinese Academy Of Sciences	5330
ESP	Universidad De Granada	5181

by a substantial margin, reflecting their central roles in the global research landscape. They are followed by Australia, India, Japan, Brazil, Canada, and several European countries, including the UK, Belgium, and Switzerland. We observe that many South American and African countries are less productive in the DS discipline, highlighting a geographic disparity in research output within the DS domain.

A. ANALYSIS BY INSTITUTIONS

We now consider the institutions where the authors of DS publications in the dataset are affiliated. Table 4 presents the top-30 institutions by number of publications in Q1 journals. Twelve of the top fourteen institutions in the diagram are based in China. These institutions include Tsinghua University, Harbin Institute of Technology, Huazhong University of Science and Technology, University of the Chinese Academy of Sciences, Shanghai Jiao Tong University, and the University of Electronic Science and Technology of China. The only two institutions in the top fourteen not located in China are the National University of Singapore (ranked 7^{th}) and the Hong Kong Polytechnic University (ranked 9^{th}), both of which are still located in Asia. In contrast, leading universities from the United States, such as Stanford University, University of Washington, Carnegie Mellon University, Texas A&M University, and UC Berkeley, along with top European institutions like ETH Zurich, University of Oxford, Imperial College London, and KU Leuven, are positioned in the lower half of the top-30, with rankings ranging from 15^{th} to 30^{th}. This distribution reflects recent developments in the DS discipline, emphasizing the rapid expansion of research output in Asian institutions.

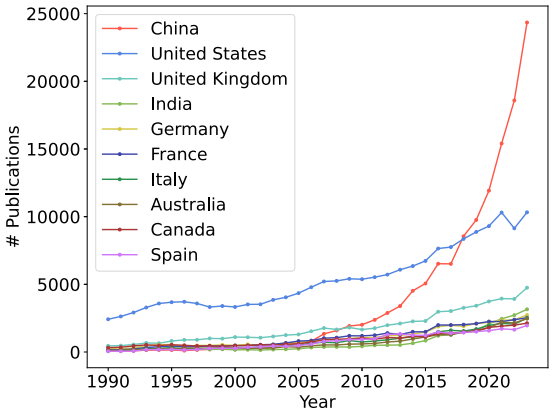

FIGURE 14. Top-10 countries in the number of publications by year (Q1 journals).

The shift from US-based to China-centric institutional dominance is illustrated by publication trends shown in Figure 14, which tracks annual DS publications in each country over time. In the first 20 years of the period under review (i.e., 1990-2010), US institutions consistently ranked among the most prolific in the world followed by UK universities. However, a significant change occurred after 2018, when Chinese institutions began to steadily close the gap and eventually surpass US institutions in terms of DS-related research outputs. By 2023, scholars based in Chinese universities were publishing almost 25,000 DS-related papers annually in Scopus-indexed journals, more than double the output of their US counterparts. This substantial increase reflects China's growing emphasis on DS scientific research and its impact on the global stage.

B. ANALYSIS BY AUTHORS

The rising popularity of DS research among Chinese institutions is also reflected in Table 5, which lists the top-10 most prolific authors according to the number of publications produced in the DS field. Seven out of ten have Chinese names and/or affiliations. The most prolific scholar in the field is Professor Witold Pedrycz who serves as a Professor and Canada Research Chair (CRC) in Computational Intelligence in the Department of Electrical and Computer Engineering at the University of Alberta, Edmonton, Canada. Since 1990, Professor Pedrycz has published 1,226 papers. His primary research interests encompass computational intelligence, fuzzy modeling and granular computing, knowledge discovery and data science, pattern recognition, knowledge-based neural networks, and control engineering. Ranking second is Professor Zeshui Xu, who holds a faculty position at Sichuan University in China. Professor Xu is renowned for his pioneering contributions to big data analysis and its applications. The third most prolific scholar, with 674 publications, is Professor Radko Mesiar. Since 1978, he has been affiliated with the Department of Mathematics at the Faculty of Civil Engineering, Slovak

TABLE 5. Top-10 authors by number of publications (Q1 journals).

Surname	Name	# Publications	Affiliation Countries
Pedrycz	Witold	1226	KOR, CHN, BRA, FRA, SAU, CAN, POL, ITA, TUR, GBR, USA
Xu	Zeshui	702	CHN, HKG
Mesiar	Radko	674	CHN, CZE, SVK, POL, AUT
Cao	Jinde	654	KOR, CHN, SAU, SVN, GBR
Cheng	T. C. Edwin	563	CHN, USA, HKG, CAN
Fujita	Hamido	554	TWN, CHN, VNM, JPN, ESP, MYS
Hall	Peter	537	KOR, BEL, GBR, DEU, USA, AUS
Herrera	Francisco	522	SAU, MEX, ESP
Laporte	Gilbert	459	BEL, DZA, FRA, CAN, GBR, NOR
Deng	Yong	450	CHN, CHE, USA, JPN

TABLE 6. Top-5 Q1 journals in the Scopus dataset by number of publications.

Journal	# Publications	CiteScore	Subject Area
Eur. J. Oper. Res.	36071	9.5	1802, 1803, 1804
Inform. Sciences	27940	12.1	1702, 1802
Physica A	25458	5.6	2613
Phys. Rev. E	19857	4.3	2613
Expert Syst. Appl.	18876	12.7	1702

University of Technology (STU) in Bratislava. His research interests cover a broad spectrum, including fuzzy sets, fuzzy logic, and intelligent computing. Closely following is Professor Jinde Cao, with 654 published papers. He holds an endowed Chair Professorship at Southeast University in Nanjing, China, and is widely recognized for his significant contributions to the analysis of neural networks, among the most popular predictive methods used in DS research. Notably, Professor Cao has been acknowledged as the most cited Chinese researcher since 2014, highlighting his substantial impact on the field. These four scholars, along with the other six researchers listed in Table 5, are undoubtedly pioneers of Data Science. It is worth noting here a further functionality of our DS-Atlas, which is the possibility to filter out publications by specifying an author of interest, thus enforcing a smart exploration of her/his contributions in the field by topic/keyword.

Overall, these figures and graphs highlight the rapidly evolving landscape of DS research, with China emerging over the past five to six years as the dominant scientific leader by quantity and impact of publications produced worldwide, hosting some of the most prolific authors in its own institutions. This geographic shift from the traditionally leading US universities to their Chinese counterparts is likely to have direct and significant implications for the future directions of DS research.

C. ANALYSIS BY JOURNALS

Now, we consider the venues where DS papers in the Scopus dataset are published. We can observe different rankings of scientific journals depending on the criteria used for evaluation. If we prioritize quantity, simply counting the number of published papers, Table 6 reveals that multidisciplinary journals, such as the *European Journal of Operations Research* and *Information Sciences*, are among the top venues for DS research. These journals span at least two/three ASJC fields, indicating their broad scope and appeal to DS scholars. These are closely followed by more specialized, methodological journals like *Physica A* and *Physica E*, which are focused more closely on the Statistics and Probability ASJC.

TABLE 7. Top-5 Q1 journals by CiteScore.

Journal	# Publications	CiteScore 2020	Subject Area
IEEE T. Pattern Anal.	7720	44.2	1702
Found. Trends Mach. Learn.	56	37.8	1702
Sci. Robot.	313	25.7	1702
J. Ind. Inf. Integr.	346	22.1	1802
Phys. Life Rev.	272	21.5	1702

On the opposite, when we focus on the quality of the journals where the DS papers are published, using the CiteScore 2020 ranking as a classification criterion, the landscape significantly changes (Table 7). Several Artificial Intelligence journals rise to the top of the list, including *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *Foundations and Trends in Machine Learning*, and *Science Robotics*. These journals are recognized for their high impact and influential nature of the research they publish in AI.

Regardless of the ranking criteria adopted for classification, whether based on the mere volume of contributions or the quality and impact as measured by the CiteScore 2020, these journals frequently publish research papers authored by the prolific scholars listed in Table 5. This correspondence between leading journals and prolific authors is not surprising (and partially expected), as these researchers are at the forefront of the DS field and consistently contribute to its most prestigious publications in the different sub-fields.

VII. EMPIRICAL VALIDATION OF THE CONWAY'S DIAGRAM

The Conway's Diagram introduced in Section II shows the multidisciplinary nature of the Data Science field in a visual and effective way, while also highlighting the expected interactions across the three main foundational areas of expertise that contribute to Data Science, namely Computer Science, Math, and Statistics, and Knowledge Domain. In this section, we discuss an experiment we performed to empirically validate Conway's Diagram of Data Science by using our Scopus dataset (see Section III).

Our DS dataset contains publications that belong to the following ASJC fields with corresponding mapping on the three main areas of Conway's Diagram:

- Artificial Intelligence (1702), Information Systems and Management (1802) → Computer Science
- Statistics, Probability and Uncertainty (1804), Statistics and Probability (2613) → Statistics
- General Decision Science (1800), Decision Sciences (1801), Management Science and Op. Research (1803) → Knowledge Domain

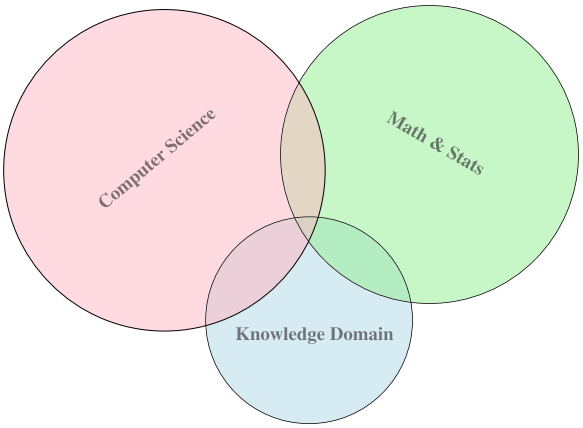

FIGURE 15. The empirical Conway's Diagram of Data Science built by using the publications in our Scopus dataset.

TABLE 8. Number of publications in our Scopus dataset for each overlapping area of the Conway's Diagram of Data Science.

Conway's Diagram Areas	# papers
Computer Science	517,492
Statistics	479,638
Knowledge Domain	230,745
Computer Science, Knowledge Domain	64,465
Computer Science, Statistics	39,598
Knowledge Domain, Statistics	33,423
Computer Science, Knowledge Domain, Statistics	525

The experiment consists of building a Conway's Diagram where each discipline is a bubble, and the size of a bubble is proportional to the number of publications belonging to that field. Since publications in Scopus can belong to multiple ASJC fields, we can also represent the intersections of two/three bubbles, whose size is proportional to the number of publications belonging to the involved areas of expertise.

The result of the experiment is the *data-driven* Conway's Diagram shown in Figure 15. We note that the diagram closely approximates the theoretical model proposed by Conway. The exact counts of publications for each discipline and related intersections are shown in Table 8. The largest bubble represents Computer Science, with 517, 492 publications, reflecting its substantial contribution to Data Science. Statistics and Mathematics, with 479, 638 publications, form the second-largest bubble, followed by the Knowledge Domain one with 230, 745 publications.

The intersections between these disciplines also provide insights into their overlap. For instance, the intersection of Computer Science and the Knowledge Domain contains 64, 465 publications, illustrating a significant collaboration between these fields. Similarly, the intersection of Computer Science and Statistics includes 39, 598 publications, showing the synergies between computational techniques and statistical methods in Data Science. The overlap between the Knowledge Domain and Statistics, with 33, 423 publications, further demonstrates the integration of domain-specific knowledge with statistical analysis, which is essential in any Data Science application. The empirical diagram also highlights the intersection of all three disciplines - Computer Science, Statistics, and the Knowledge Domain - with 525 publications. Although this intersection is relatively small, it represents the core of Data Science, where methodologies from all three fields converge.

The empirical version of Conway's Venn diagram of Figure 15, constructed using bibliometric data, successfully reproduces the theoretical intersections between Computer Science, Statistics, and the Knowledge Domain which were expected from a theoretical point of view. The proportionality of the bubbles and their intersections aligns well with the theoretical model, providing a robust approximation of the multidisciplinary nature of Data Science. This empirical validation not only reinforces the theoretical framework but also offers a data-driven visualization that enhances our understanding of the contributions from these disciplines to the field of Data Science.

TABLE 9. Weighted F_1 scores for classifying articles into Scopus subject areas using representations from different embedding models.

Model	F_1
TF-IDF + SVD	0.528
BERT	0.534
SciBERT	0.585
SPECTER	0.594
SciNCL	0.599
SBERT	**0.604**

VIII. LIMITATIONS

Although the DS-Atlas allows to visualize the landscape of Data Science research, limitations must be recognized that may affect the interpretation and generalizability of our findings.

A. CHOICE OF ADOPTING SCOPUS AS PRIMARY DATABASE

As a general remark, we note that the literature demonstrates that country-level indicators of scientific output and citations are stable and largely independent of the database used - particularly when comparing ISI Web of Science and Scopus [1]. However, we recognize that journal coverage may differ from repository to repository. Notably, Web of Science coverage in Social Sciences and Arts and Humanities remains relatively low, with these disciplines being under-represented compared to Scopus coverage [18]. This makes Scopus more suitable for our interdisciplinary analysis of the Data Science field. Therefore, we are confident that similar results would be obtained using alternative, well-recognized bibliometric sources. However, the exclusive reliance on Scopus may still introduce certain coverage biases, particularly regarding publications in non-English languages, regional journals not

indexed by Scopus, or emerging publication venues such as preprint servers and conference proceedings that are not comprehensively covered.

B. SHAREABILITY OF THE "DATA SCIENCE DEFINITION"

In this paper, we propose a definition of Data Science with grounding on relevant authors and publications in the field. This definition has been used to outline the boundaries of the Data Science discipline, and to determine the contents of the considered dataset of publications. The categorization of publications in ASJC fields is inherently subjective and depends on the editorial choices of Elsevier Scopus as well as our own perspectives as field experts with particular viewpoints and disciplinary background. Additionally, the choice to focus on certain sub-disciplines while excluding others may introduce systematic bias, potentially overlooking relevant intersections or emerging areas that do not fit neatly within traditional disciplinary boundaries. This challenge is particularly acute given Data Science's inherently interdisciplinary nature, where relevant research may be published across diverse venues that span multiple traditional academic domains.

C. TECHNIQUES USED FOR DATA PROCESSING

The pipeline used for the DS-Atlas construction and the related state-of-the-art techniques introduce their own limitations. The semantic representation generated through LLMs may not capture all nuances of scientific discourse, particularly domain-specific terminology, methodological innovations, or subtle theoretical distinctions. Moreover, the dimensionality reduction process necessarily involves information loss.

All these factors collectively limit the accuracy and comprehensiveness of the empirical validation we provide, particularly regarding our proposed conceptual framework of Data Science. However, we argue that this validation represents a valuable first attempt to provide concrete, data-driven support for understanding the field's structure and evolution.

IX. CONCLUSION

In this paper, we presented the Atlas of Data Science Research (DS-Atlas), generated by considering the main descriptive metadata of a selected dataset of scientific publications in the field, and a computational pipeline based on NLP methods, LLMs, and dimensionality reduction techniques. The analysis of data science research is presented in terms of descriptive keywords and geographical distribution of influential authors, institutions, and journals in the field.

The DS-Atlas represents a significant contribution to understanding the evolving landscape of Data Science research through a data-driven approach. By analyzing approximately 1.3 million scientific publications from the Scopus database, we have created an interactive visualization tool that provides unprecedented insights into the structure and evolution of Data Science scholarship. Our approach demonstrates how modern computational techniques can be leveraged to make sense of large-scale bibliometric data, offering researchers and practitioners a practical method for exploring the complex interdisciplinary nature of the field. An extension of the DS-Atlas to further disciplines of interest can be considered by involving appropriate domain experts in the selection of the publication dataset.

APPENDIX A
DS-ATLAS CONSTRUCTION

In this section, technical details about the DS-Atlas construction are discussed. In particular, we focus on the two stages of the pipeline, i.e., *publication embedding* and *dimensionality reduction*, as well as on the *thematic layer* based on author-assigned keywords that is built on top of publications.

A. PUBLICATION EMBEDDING

For embedding the publication abstracts, we employ a Large Language Model (LLM); in particular we use Sentence-BERT (SBERT) [28], that is a modification of the pre-trained BERT model [12]. The reason for choosing SBERT over the standard BERT model, is that BERT achieves excellent performance on token-level tasks but struggles to create meaningful sentence-level embeddings. Indeed, the model outputs token-level representations that must be aggregated using pooling strategies, such as mean or max pooling, to get a sentence-level embedding. However, these pooling methods are not optimized during BERT's pretraining phase, limiting the quality of the resulting embeddings. An alternative could be the use of the CLS token provided by the model, which represents the overall meaning of the sentence. However, this token is designed primarily for downstream classification tasks rather than to capture nuanced semantic relationships between sentences or documents.

On the other hand, SBERT employs siamese and triplet network structures during its fine-tuning phase to learn sentence representations that are directly comparable using distance metrics, e.g. cosine similarity. The resulting embeddings are highly effective for semantic similarity search, clustering, and content-based information retrieval. This specialized training makes SBERT particularly suited for capturing the meaning of long text sequences, like abstracts, which allows to accurately group topically related papers.

Specifically, we employ the `all-mpnet-base-v2`[7] model from the HuggingFace library, which outputs a 768-dimensional vector embedding for each processed abstract.

B. DIMENSIONALITY REDUCTION

For transforming the 768-dimensional SBERT embeddings into a 2-dimensional representation, we employ FIt-SNE [17], an accelerated implementation of the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [31].

[7] Sentence-transformers/all-mpnet-base-v2.

t-SNE is a non-linear dimensionality reduction technique renowned for its ability to preserve the local structure of high-dimensional data, generating meaningful low-dimensional visual representations that may reveal clusters and latent relationships in the underlying data. However, traditional t-SNE can be computationally intensive for large datasets. For this reason, we employ FIt-SNE, which significantly accelerates the dimensionality reduction process through Fourier Transform-based interpolation.

Furthermore, we leverage the `openTSNE` [22] implementation, which not only provides FIt-SNE capabilities, but also offers the functionality of mapping new, unseen data points onto an existing, pre-computed embedding space. This feature not only enabled the placement of keywords in the DS-Atlas, as depicted in Figure 7, but also allows for a continuous update of the visualization with new publications, without the need to recompute from scratch a new embedding space.

C. THEMATIC LAYER

This layer is derived from author-assigned keywords that need to be placed in the same semantic space used for the publications. Each keyword must be associated with a representative text (i.e., an abstract, like we did for publications) to use in embedding and dimensionality reduction. To this end, for each keyword, we collected the abstract of the corresponding Wikipedia page, namely the page with a title matching the keyword. The Wikipedia abstract is passed to SBERT for embedding and subsequent dimensionality reduction. Finally, the keyword is plotted in the thematic layer.

Only keywords associated with at least 50 publications have been considered and used in this layer to avoid labels that are poorly representative of the dataset. All the considered keywords had a matching Wikipedia page for abstract extraction.

APPENDIX B
DS-ATLAS EVALUATION

To provide a quantitative basis for our choice of embedding model, we conducted a comparative evaluation of several leading solutions, following the evaluation methodology provided in González-Márquez et al. [16], the article inspiring this work. The goal of this evaluation was to determine which model could produce the most semantically meaningful representations of our corpus, as measured by the ability of a classifier to exploit these representations in order to assign articles to their correct subject areas.

We considered different models for computing the publications' embedding features:

- TF-IDF + SVD: a classic, non-neural approach. To create a dense vector comparable to the other models, we applied Truncated Singular Value Decomposition (SVD) on the resulting sparse matrix of the Term Frequency-Inverse Document Frequency (TF-IDF) representation of the abstract, reducing its dimensionality to 300;
- BERT[8] [12]: the original transformer model trained on a general corpus;
- SciBERT[9] [3]: a BERT model trained from scratch on a large corpus of scientific papers from computer science and biomedicine;
- SPECTER[10] [10]: a SciBERT-based model fine-tuned using a contrastive objective on the scientific citation graph, designed to embed papers such that citing papers are close to cited papers;
- SciNCL[11] [21]: another SciBERT-based model that uses a contrastive learning framework based on citation links to produce improved scientific paper embeddings;
- SBERT[12] [28]: a sentence-transformer model fine-tuned on a massive and diverse dataset of sentence pairs.

We framed the evaluation as a multi-class classification task. For each model, we generated the high-dimensional embeddings for the abstract of every labeled paper, and then reduced them to two dimensions with t-SNE. We then trained a standard k-Nearest Neighbour classifier, with $k = 10$, on a random 80% of the labeled data to predict the Scopus subject area from the reduced embeddings. Performance was measured using the weighted F_1 score on the remaining 20% test set. The results of the comparison are presented in Table 9.

Ultimately, SBERT delivered the highest classification performance: we attribute its success to its specific fine-tuning objective, which is designed to produce embeddings where semantic similarity directly translates to vector proximity across a wide range of domains. While models like SPECTER and SciNCL are highly optimized for citation-based tasks, the broad training of SBERT on diverse sentence-pairing tasks appears to generate more versatile and effective representations for general topic classification. Based on this quantitative evidence, we selected SBERT for generating the embeddings for our final analysis and visualization, confident in its superior ability to represent the thematic content of our article collection.

APPENDIX C
GUIDELINES ON THE USE OF THE DS-ATLAS WEBAPP

In the default visualization, the web application displays the complete DS-Atlas with all the publications colored according to their respective subject areas. Keywords are shown in overlay on the DS-Atlas based on their frequency in the dataset. Initially, top-frequent keywords are shown, while keywords with lower frequency appear when the user zooms into a specific region. Users can modify the color palette and toggle keyword visibility using the control buttons located in the bottom right corner of the interface.

[8] Google-bert/bert-base-uncased.
[9] Allenai/scibert_scivocab_uncased.
[10] Allenai/specter.
[11] Malteos/scincl.
[12] Sentence-transformers/all-mpnet-base-v2.

By moving over any point/publication, the DS-Atlas displays a tooltip containing the corresponding metadata including the Scopus identifier, publication year, subject area, title, authors, and journal information. By clicking on a publication point, the user is redirected to the corresponding online publication page for detailed exploration.

A. PUBLICATION FILTERING

This operation allows to focus the exploration on specific publications, according to selected criteria.

- *Keywords*: the tool enables keyword-based filtering through direct interaction with the keywords in overlay. Users can click on any keyword to filter the DS-Atlas, and to visualize only publications that have the selected keyword in the author-assigned ones. The lower left corner provides a reset option to clear the visualization and return to the complete DS-Atlas view.
- *Subject areas*: the top right corner features a subject area filter that allows users to select or deselect entire subject areas. This filter enables focused exploration of specific research domains by including or excluding publications from particular ASJC fields according to user preferences.
- *Metadata*: the top left section provides a comprehensive search functionality based on three metadata fields: publication year, title, and author. Users can enter search strings to visualize the publications containing the specified text within the selected metadata field. This feature fosters targeted searches based on authors, publication periods, or titles of interest.

B. SPACIAL INSIGHT

The left panel contains the ''Area Search'' functionality, which implements the insight operation. This feature allows users to select a rectangular region of the DS-Atlas and analyze the distribution of publications and keywords within that area. To activate this functionality, the user has to select two points on the DS-Atlas, each one corresponding to a point/publication (use the select button to choose the target publications to delimit the rectangle of interest). By clicking on the search button, the DS-Atlas zooms-in the selected area and displays statistics on the publications in the region, including publication counts and keyword frequencies.

ACKNOWLEDGMENT

The information and views set out in this article are those of the authors and do not reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

REFERENCES

[1] É. Archambault, D. Campbell, Y. Gingras, and V. Larivière, ''Comparing bibliometric statistics obtained from the web of science and scopus,'' *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 7, pp. 1320–1326, Jul. 2009.

[2] M. Aria, M. Misuraca, and M. Spano, ''Mapping the evolution of social research and data science on 30 years of social indicators research,'' *Social Indicators Res.*, vol. 149, no. 3, pp. 803–831, Jun. 2020.

[3] I. Beltagy, K. Lo, and A. Cohan, ''SciBERT: A pretrained language model for scientific text,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, Nov. 2019, pp. 3615–3620.

[4] L. Breiman, ''Statistical modeling: The two cultures (with comments and a rejoinder by the author),'' *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, Aug. 2001.

[5] L. Cao, ''Data science: A comprehensive overview,'' *ACM Comput. Surv.*, vol. 50, no. 3, p. 43, Jun. 2017.

[6] J. M. Chambers, ''Greater or lesser statistics: A choice for future research,'' *Statist. Comput.*, vol. 3, no. 4, pp. 182–184, Dec. 1993.

[7] C. Chen, ''CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature,'' *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 359–377, Feb. 2006.

[8] C. Chen and M. Song, ''Visualizing a field of research: A methodology of systematic scientometric reviews,'' *PLoS ONE*, vol. 14, no. 10, Oct. 2019, Art. no. e0223994.

[9] W. S. Cleveland, ''Data science: An action plan for expanding the technical areas of the field of statistics,'' *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 7, no. 6, pp. 414–417, Dec. 2014.

[10] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, ''SPECTER: Document-level representation learning using citation-informed transformers,'' in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 2270–2282.

[11] Drew Conway. *The Data Science Venn Diagram*. Accessed: Apr. 3, 2025. [Online]. Available: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2018, pp. 4171–4186.

[13] V. Dhar, ''Data science and prediction,'' *Commun. ACM*, vol. 56, no. 12, pp. 64–73, Dec. 2013.

[14] D. Donoho, ''50 years of data science,'' *J. Comput. Graph. Statist.*, vol. 26, no. 4, pp. 745–766, Oct. 2017.

[15] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, ''From data mining to knowledge discovery in databases,'' *AI Mag.*, vol. 17, no. 3, pp. 37–54, Mar. 1996.

[16] R. González-Márquez, L. Schmidt, B. M. Schmidt, P. Berens, and D. Kobak, ''The landscape of biomedical research,'' *bioRxiv*, vol. 5, no. 6, Apr. 2024, Art. no. 100968.

[17] G. C. Linderman, M. Rachh, J. G. Hoskins, S. Steinerberger, and Y. Kluger, ''Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data,'' *Nature Methods*, vol. 16, no. 3, pp. 243–245, Mar. 2019.

[18] P. Mongeon and A. Paul-Hus, ''The journal coverage of web of science and Scopus: A comparative analysis,'' *Scientometrics*, vol. 106, no. 1, pp. 213–228, Jan. 2016.

[19] M. K. M. Nasution, O. S. Sitompul, E. B. Nababan, E. S. M. Nababan, and E. Sinulingga, ''Data science around the indexed literature perspective,'' in *Softw. Eng. Perspect. Intell. Syst., 4th Comput. Methods Syst. Softw.*, 2020, pp. 1051–1065.

[20] P. Naur, *Concise Survey of Computer Methods*. New York, NY, USA: Petrocelli/Charter, 1974.

[21] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, and G. Rehm, ''Neighborhood contrastive learning for scientific document representations with citation embeddings,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 11670–11688.

[22] P. G. Policar, M. Strazar, and B. Zupan, ''OpenTSNE: A modular Python library for t-SNE dimensionality reduction and embedding,'' *J. Stat. Softw.*, vol. 109, pp. 1–30, 2024.

[23] A. Ponsard, F. Escalona, and T. Munzner, ''PaperQuest: A visualization tool to support literature review,'' in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2016, pp. 2264–2271.

[24] Gil Press. (2013). *A Very Short History of Data Science*. Accessed: Mar. 4, 2025. [Online]. Available: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-o%f-data-science/

[25] F. Provost and T. Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*, 1st ed., Sebastopol, CA, USA: O'Reilly Media, 2013.

[26] A. Purnomo, E. Rosyidah, M. Firdaus, N. Asitah, and A. Septianto, "Data science publication: Thirty-six years lesson of scientometric review," in *Proc. Int. Conf. Inf. Manage. Technol. (ICIMTech)*, Aug. 2020, pp. 893–898.
[27] D. R. Raban and A. Gordon, "The evolution of data science and big data research: A bibliometric analysis," *Scientometrics*, vol. 122, no. 3, pp. 1563–1581, Mar. 2020.
[28] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019, pp. 3982–3992.
[29] M. Thelwall and P. Sud, "Scopus 1900-2020: Growth in articles, abstracts, countries, fields, and journals," *Quantum Sci. Stud.*, vol. 3, no. 1, pp. 37–50, 2021.
[30] W. M. P. van der Aalst, "Data science in action," in *Process Mining*. Berlin, Germany: Springer, 2016, pp. 3–23.
[31] L. J. P. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[32] N. Jan van Eck and L. Waltman, "Text mining and visualization using VOSviewer," 2011, *arXiv:1109.2058*.
[33] N. J. van Eck and L. Waltman, "CitNetExplorer: A new software tool for analyzing and visualizing citation networks," *J. Informetrics*, vol. 8, no. 4, pp. 802–823, Oct. 2014.

SERGIO PICASCIA (Member, IEEE) received the B.Sc. degree in economics from the Università degli Studi di Napoli, Naples, Italy, in 2019, and the M.Sc. degree in data science and economics from Università degli Studi di Milano, Milan, Italy, in 2021, where he is currently pursuing the Ph.D. degree in computer science.

He has been a Research Fellow with the Department of Computer Science, Università degli Studi di Milano, being involved in the NextGenerationUPP project, which focused on the application of artificial intelligence and advanced information management techniques for the digital transformation of Italian legal processes and digital justice. He has published in international journals and conferences, focusing his research interests on natural language processing and its applications to the humanities and social sciences.

STEFANO MONTANELLI received the Ph.D. degree in informatics from the Department of Computer Science "Giovanni Degli Antoni," Università degli Studi di Milano (UNIMI), in 2007. He is currently a Full Professor with the Department of Computer Science "Giovanni Degli Antoni," UNIMI. He is a member of the Data Science Research Center, UNIMI. Since 2024, he has been the Head of the Interdepartmental Master's Degree Program in Data Science for Economics, UNIMI. He participates in the activities of the Information Systems and Knowledge Management (ISLab) research group. He is the author of several publications in international journals and conference proceedings. He actively participates in national and international research projects, collaborating with academic and industrial institutions on multidisciplinary applications of computer science with a focus on the humanities, social sciences, and law. His main research interests include data modeling, semantic web, data matching, data classification, data science, NLP, and human-in-the-loop data management.

SILVIA SALINI received the degree in statistical sciences from Università Cattolica del Sacro Cuore, Milan, in 1999, and the Ph.D. degree in statistics from Università degli Studi di Milano Bicocca, in 2002.

From 2004 to 2015, she was a Researcher, and from 2015 to 2024, she was an Associate Professor with Università degli Studi di Milano. Since 2024, she has been a Full Professor of statistics with the Department of Economics, Management, and Quantitative Methods, Università degli Studi di Milano. Her research focuses on data science, robust statistics, data analysis for the social and economic sciences, scientometrics, and the ethical, legal, and social implications of artificial intelligence. She coordinated the Master's Degree Program in Data Science for Economics, Università degli Studi di Milano, for six years. She teaches various courses in statistics, data analysis, and statistical learning at both undergraduate and graduate levels, and is a member of the faculty board of the Ph.D. program in public health sciences at Università degli Studi di Milano. Her work includes numerous scientific publications in international journals. In her academic activity, she combines methodological rigor with practical applications, promoting the use of statistics and data science to understand complex phenomena and to support informed decision-making in economic, social, and healthcare contexts. She has participated in numerous national and international research projects focusing on complex data analysis.

STEFANO VERZILLO is currently a Senior Researcher with the Joint Research Centre (JRC), European Commission, Competence Centre on Microeconomic Evaluation (CC-ME). Since 2012, he has also been an Adjunct Professor with Università degli Studi di Milano Bicocca, where he teaches statistical modeling and machine learning in the Ph.D. program in economics, statistics, and data science, and in the master's program AI and data analytics for business (AIDA). He also serves on the board of the Ph.D. program in public health, epidemiology, statistics, and economics. He has published on these topics in several international journals. His research interests include applied microeconometrics and counterfactual impact evaluation in education, labor, and health economics.