Journal of Scientometric Res. 2022; 11(1):64-78
http://www.jscires.org

Research Article

'Big data' Research: A Bibliometric Analysis of the Scopus Database, 2009–2019

Henry Chavez[1,*], María Belén Albornoz[1], Fernando Martín[2]
[1]*CTS LAB – FLACSO, Quito, ECUADOR.*
[2]*FLACSO, Quito, ECUADOR.*

ABSTRACT
Scopus-database publications containing the keyword 'big data' have skyrocketed from 30 (2009) to almost 16,000 (2019). This trend reveals this field's importance across disciplines and contexts. Previous works have analysed the emergence and characteristics of scientific research on 'big data' but need updating. We undertook a bibliometric analysis of over 73,000 such 2009–2019 publications. This data helped to identify the primary trends, subjects, networks and institutions publishing on big data worldwide and explain the relations and differences between scientific communities working on this subject in central and peripheral countries. Furthermore, this research highlights Chinese researchers' and institutions' prominence in this field alongside the influence of American contributions, which are most frequently cited. The emergence of dynamic poles of scientific production in middle-income countries in Asia, Africa and South America are also studied. Despite the dynamism of the field, about 2% of the articles account for 40% of the field's citations, while 42% have no citations. Originating in computer science and engineering, big data research is increasingly becoming interdisciplinary. Keyword trends over time also show a shift from technical and prospective concerns towards (1) methodological and practical issues and (2) the development of AI and machine learning techniques. These indicators present differences between countries with varying geo-economic conditions. Collaboration networks have rapidly grown with the US and China as the main nodes and European countries as intermediaries in the circulation of this topic. Although still rare, there are some signs of South-South collaboration between Latin America, Africa and Asia.

Keywords: Big data, Bibliometrics, Scientific networks, Knowledge circulation.

Correspondence
Henry Chavez
CTS LAB – FLACSO, Quito-170201, ECUADOR.
Email id: henry.chavez@gmx.com
https://orcid.org/0000-0002-1834-3437

Received: 05-07-2021;
Revised: 24-12-2021;
Accepted: 06-02-2022.
DOI: 10.5530/jscires.11.1.7

INTRODUCTION

Production, management and processing of data have been constant concerns for modern States since their early days. Indeed, epidemics, population control, tax collection and warfare have stimulated innovations in the collection and processing of reliable data for several centuries in order to produce information useful for government decision-making. Technological innovations in the post-war period allowed the digitalisation of data, and with it, the States started systematically using digital platforms to organise and process the information collected on their citizens, territories, institutions and economies. As part of this process, concerns about processing large amounts of information started emerging during the 1970s and 1980s.[1]

However, the 'data problem' as we know it today started materialising with the introduction of the World Wide Web in the early 1990s. The exponential growth of data production and storage on digital media platforms encouraged several technological innovations to address both support (hardware) and processing (software) issues. Several publications from the 1990s and early 2000s account for this process.[2–4] Some of these publications have already explored the effects and potentialities of big data for economic and financial analysis.[5,6]

According to Gupta and Rani,[7] the term 'big data', in the sense it is currently used, comes from a number of different sources.[8–10] However, some authors[11–14] state that the term was coined around 2005 by Roger Magoulas from O'Reilly Media[15,16] to refer to an amount of data whose processing exceeded the capacity of a single machine and consequently required distribution over several computers.

This is the very principle followed in the *Hadoop Project* developed by *Yahoo* experts during the same time period. This project produced a series of open-source programs to facilitate the use of a network of many computers to solve problems involving massive amounts of data and operations. The *Hadoop Project* was based on two articles written at

Google regarding its own systems: *Google File System*[17] and *MapReduce*.[18] These seminal articles marked the north for this type of data processing. Owing to the great transformations produced by the application of this new technology, the notion of 'big data' began to spread from the computer engineering and data fields to other areas with immense potential such as health, business, finance and social sciences.

Previous studies[7,12,14,19,20] have pointed out that the number of publications on the subject started to increase from around 2008. However, the trends observed in the first publication are now very limited compared to the more recent volume of publications. In eight years alone (from 2011 to 2019) the annual publications on this topic (in all areas and disciplines) increased from 90 to 16,000. This paper aims to update the observations and findings of previous bibliometric works by extending their analysis to the entire scientific corpus in the last decade. Moreover, most of these publications have focused on undertaking an analysis of the general bibliometric performance of the field. However, little attention has been paid to the geographical and economical differences as well as the linkages and circulation of knowledge between countries from different economic regions. A second objective of this paper is to shed light on these differences and linkages by analysing not only the performance differences between countries but also the international collaboration networks behind this explosion of publications during this decade.

LITERATURE REVIEW

There is a handful of specific bibliometrics analysis on 'big data'. The first one, published as a blog post on *Research Trends* in 2012, highlighted the growing trend of publications since 2008 and the emergence of 'big data' as a research area.[12] It was followed by works by Singh *et al.*[20] from India, Tseng *et al.*[21] from Taiwan, Peng *et al.*[22] from China and Brazil, López-Robles *et al.*[23] from Spain and Mexico, Gupta and Rani[7] from India, Parlina *et al.*[14] from Indonesia and Raban and Gordon[19] from Israel. Table 1 summarises the main features and conclusions of these earlier works.

Table 1: Previous Bibliometric Analyses on 'Big Data'.

Year	Authors	Period of analysis	Data bases	Number of publications	Main variables	Main conclusions
2012	Halevi and Moed	2008-2012	Scopus	306	Timeline trends; subject areas; document types; geographical distribution and thematic characteristics	Explosion of publications / Growing interdisciplinary research / Leading role of the USA and China
2015	Singh *et al.*	2000-2015	WoS	8200	Total output; growth of output; authorship and country collaboration patterns; major contributors; top publication sources; thematic trends and emerging themes	Explosion of publications / The importance this emerging research area in different disciplines
2016	Tseng *et al.*	1983-2014	WoS	18000	Relations between "data mining" and "big data" scientific literature.	Common top country authorship and research areas / 2 of the top 10 journals and author organisations are the same / 1/3 of authors researching "big data" also published on "data mining"
2017	Peng *et al.*	2011-2015	WoS	4000	Collaboration networks on "big data" research	Collaboration exist, but networks are sparse / Small networks, low productivity and low interinstitutional and interdisciplinary connections / There are some better established networks with better interinstitutional connections and composed by researchers with high academic status
2018	López-Robles *et al.*	2012-2017	WoS	25000	Performance indicators; science mapping	Growing number of publications / Central role of the US and China / Engineering, Computer and Information sciences are the core disciplines / Four areas contain the motor, basic and transversal themes structuring the field of research during the studied period: Data Management, Decision Support, Privacy and Web and Social Networks

Continued...

Table 1: Cont'd.

Year	Authors	Period of analysis	Data bases	Number of publications	Main variables	Main conclusions
2018	Gupta and Rani	2000-2017	ACM, IEEE, SAGE, Science Direct, WoS	24000	Productivity indicators; document type; Publishers, Keywords; Open-source software systems; challenges	Growing trend of publications / Conference papers are the main source of publications / Main journals publishing on this field: IESS Access, Big Data & Society, International Journal of Distributed Sensor Networks / Most frequent keywords indicates the development and deployment of new technologies (cloud computing, distributed computing, security, MapReduce, Hadoop, etc.) and techniques (classification, clustering, data mining, machine learning, optimisation, visualization, etc.) / Overview on the main open-source software systems developed to process big data / Apache Spark stream processing platform has a better performance that MapReduce solution / Main challenges: data acquisition and metadata; quality; storage; sharing and transfer; scalability; analysis; querying and indexing; uncertainty; privacy, security and ethics and visualisation.
2020	Parlina, Ramli and Murfi	2009-2018	Scopus	7000	Core journals in Computer Sciences; most cited articles; top productive authors; countries and institutions; thematic clustering and evolution of the research	Growing trend of publication / Leading role of China and the US. / Dominance of some Chinese institutions / Absence of prominent US institutions / Research dominated by data analytics, tools and algorithms; infrastructure; security and privacy; applications and services and data related technologies
2020	Raban and Gordon	2006-2019	WoS	11000	Relation between "data science" and "big data"	Different paths of the scientific production ("data science": gradual vs. "big data": exponential / New trend of publications combining concepts from both corpus / The two fields have different academic origins and leading publications / Big data literature is more prominent, has intensive citation activity and bigger funding, particular from China / Data Science literature serve as a theory-base or a tool-box for big data publications.

Continuing with the same line of analysis as these publications, this article aims to contribute, update and extend on their findings by adopting a double strategy. On one hand, we use a bigger and up-to-date dataset which will allow us to capture a broader picture of big data research's evolution until the end of 2019. On the other hand, we focus on some characteristics of this scientific production that were neglected until now in previous works. Through such an analysis, we expect to produce a more general picture of the international structure of scientific production on this topic in order to identify the weaknesses and opportunities for a more balanced global development of the field.

DATA AND METHODOLOGY

Most previous works have based their analysis on datasets from the Web of Science (WoS) delimited by the research area or type of documents. Gupta and Rani[7] and López-Robles *et al.*[23] Used the biggest datasets available for their analysis which were about 25,000 each. Aiming to obtain different insights, this paper uses a larger dataset from a less explored source. It has been built through a general query of documents containing the keyword 'big data' in the fields 'title, abstract or keywords' of the Scopus database. This query provided a result of 75.300 documents which were published between 1970 and 2019. The results were then exported as several 'csv' files containing 2,000 entries each with all available information on citation and bibliographic details, abstracts and keywords, funding and other details. These files were compiled and subsequently validated through a verification algorithm developed on python 3.7 to exclude duplicates and entries without even minimal information (author and affiliation) which were needed for further analysis. The resulting dataset

contained 73.230 entries. The raw data obtained in Scopus contained authorship and affiliation for each article in single cells. Therefore, additional splitting operations were required in order to obtain information on an individual basis. The split dataset of about 261.826 entries gives us individualised information on every author of these 73.230 publications. Even though several of these entries correspond to the same individuals, they provide specific information about the conditions under which these authors contributed to those publications (affiliations, funding, etc.).

This dataset was then analysed for three main aspects: scientific productivity and performance indicators; thematic maps, clusters and trends and collaboration networks. Scientific productivity and performance indicators include volume and growth rates of publications, major contributors, journals and articles disaggregated according to year, document type, institutions and countries. The thematic analysis considered the distribution of publications by research area, top research areas by year and country, keywords clustering and trend analysis over time. Finally, a network analysis was performed at three levels (authors, institutions and countries) to identify main, local and international collaboration clusters and trends. Special attention was given in each step to the differences and connections between countries, institutions and researchers from different geographical and economic conditions. We have used the location within continents and subcontinental regions as well as the World Bank Country and Lending Groups classification for 2020 as indicators of these differences: low-income economies (Gross National Income per capita (GNIpc) <= US$1,035); lower-middle-income economies (GNIpc between US$1,036 and US$4,045); upper-middle-income economies (GNIpc between $4,046 and $12,535) and high-income economies (GNIpc > $12,536).[24]

RESULTS AND DISCUSSION

Scientific Productivity and Performance Indicators

a) Scientific output

The very first recorded entry on the Scopus database containing the keyword 'big data' was published in 1970. Up to December 2019, 73,230 items containing this keyword or pre-coordinated concept[19] either in the title, the abstract or the keyword have been published. This literature is composed mainly of conference papers (58%) and journal articles (32%). Other contributions are less representative: Reviews (4%), book chapters (3%), books (1%) and others 3% (editorials, notes, articles in press, letters, short surveys, etc.). The production trend of this scientific output has not followed a steady trend; however, it follows an exponential one started a decade ago (Figure 1). Indeed, the scientific production on this subject advanced from around 30 articles published annually at the end of the 2000s to almost 16.000 by 2019. This shows an average growth rate of about 90% per year.

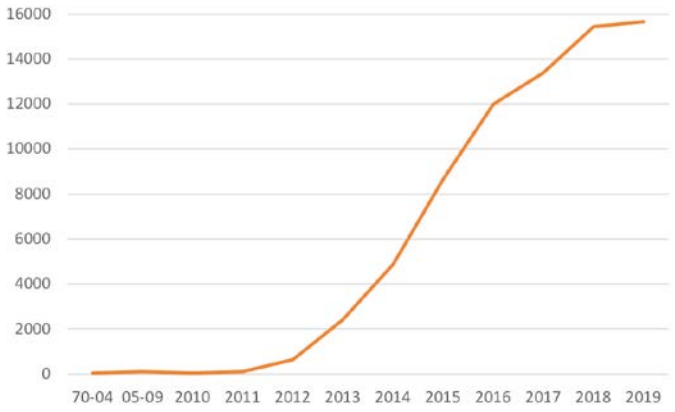

Figure 1: Total Number of Publications on Big Data by Year, 1970–2019.

Such scientific productions are not evenly distributed around the world. China (20,838) and the United States (16,696) have indisputable leadership in this field. However, it is worth noting that even though in the early years US-affiliated researchers led the field in terms of the number of publications, they were quickly overtaken by China-affiliated researchers. Since 2015, North American publications have stagnated while China have doubled the number of publications (Figure 2). This outperformance is most likely related to China's techno-nationalistic R&D policy that aims at the country's digital transformation and global leadership in the data and artificial intelligence industries.[25] India, with over 1,500, and Great Britain, with almost 900 publications, in 2019 also reflect a growing interest in the subject. During the last five years, the former has presented a higher average annual growth rate (31%) than the latter (15%). Other important contributors from Asia, Europe, Australia and Canada have annually produced 300–600 articles; this represents an average annual growth rate of about 14% during the last five years. Among newcomers, Russia is one of the fastest-growing contributors, with an average annual growth rate of about 36% since 2015.

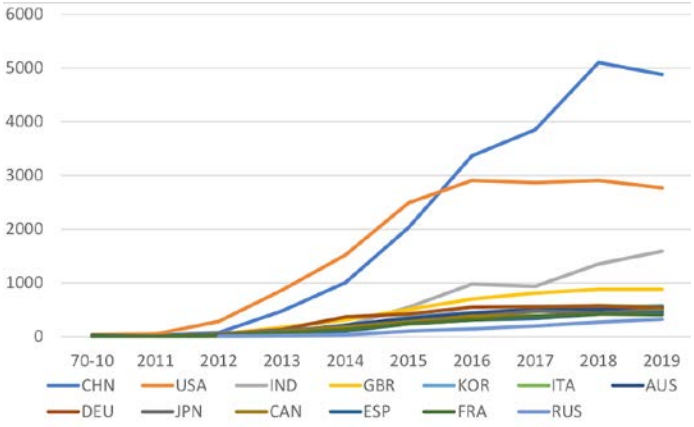

Figure 2: Publications on Big Data by Country and Year, 1970–2019.

Other minor contributors with increasing scientific output on 'big data' include Indonesia, Turkey, Norway, New Zealand, Israel, Mexico, Thailand and Chile with more than 30% of an increase per year since 2015; Pakistan, Iran, Morocco, South Africa, Egypt, Vietnam and Colombia with more than 40% and the Philippines, Mongolia, Ecuador and Myanmar with an average growth rate above 100% per year.

Grouped by continent and region, the primary Asian contributors (excluding China) fall in the southeast and eastern regions: India (5,680 publications), Korea (2,704), Japan (2,389) and Taiwan (1,408). Africa as a whole registers 1,907 publications, most of which were produced by North African countries such as Morocco (569) or Egypt (247), and also South Africa (412). Latin America produced about 1,770 articles; half of these articles were published by Brazil-affiliated researchers (813). Other contributors to this scientific production are based in Mexico (271), Colombia (184), Chile (132), Ecuador (109) and Argentina (103). The rest of the Latin American countries have fewer than 80 publications on the subject. Western European research is led by Great Britain (4,282), Germany (3,187), Italy (2,504), France (1,970) and Spain (1,917). Eastern European countries register a lower volume of publications than their Western neighbors but higher growth rates as Russia (1,703), Poland (541) or Romania (326). Lastly, in the Middle East, the main contributors are based in Saudi Arabia (551), Georgia (534), Turkey (525) and Israel (341). Most of these publications were produced in high (55%) and upper-middle-income countries (34%). Contributions from lower-middle- (11%) and low-income countries (0.1%) are less representative (Figure 3).

a) Authors and Institutions

Only 12% of these publications constituted of individual contributions. Most of these (75%) were produced by teams made of 2–5 researchers, 12% by teams of 6–10 and 1% by teams of more than 10 researchers. It is worth noting that 16 of the latter teams were the output of collaboration networks of 50 to 100 researchers and 8 of networks of more than 100 researchers from 28 countries. However, most of the collaboration networks behind these publications are nation-based. Only 19% are international, out of which 15% are binational, 3% include researchers from 3 countries, 1% from 4 countries and less than 1% were from 5 or more countries.

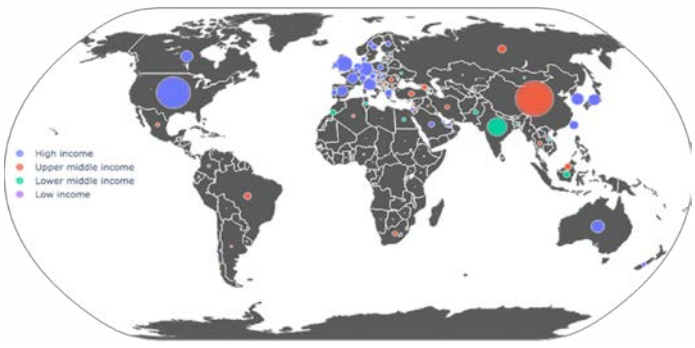

Figure 3: Publications on Big Data by Country and Income Group.

Overall, the scientific body researching 'big data' totals 168,463 researchers worldwide. Even though we have researchers from 158 countries, 25 of them concentrate 89% of the researchers, reflecting the unequal global distribution of knowledge and know-how in this field. Chinese (29%) and North American (19%) institutions are the primary hubs of such researchers. European countries collectively have around 17% and India has 6% of the researchers. Other countries such as Russia, Brazil, Indonesia, Malaysia have between 1,000 and 2,000 researchers each; Turkey, Morocco, Poland, Pakistan, Iran, South Africa or Israel have between 500 and 1,000 researchers and Mexico, Colombia, Egypt, Argentina, Vietnam, Chile or Ecuador have between 100 and 500 researchers.

The vast majority of these researchers (77%) are occasional contributors, while 22.4% have participated in 2–10 articles. However, the core of this scientific body, composed of authors with more than 10 publications, has only about 1,000 researchers. Among these, the top 10 authors have more than 50 publications (Table 2). They are based mainly in China, the US, the UK and India but also in Italy, Canada, Australia, Spain, Saudi Arabia, Korea and Portugal.

According to the country's income level, the main authors among upper-middle-income countries are mainly Chinese, South Africans and Colombians, with 35 to 50 publications. The leading researchers from lower-middle-income countries are predominantly affiliated with institutions in India and Morocco. These researchers have produced between 18 and 35 articles. Finally, researchers from low-income countries are mainly from Africa, the Middle East and Nepal. However, their productivity remains very low: 2 articles per author, except for Sun (12) and Maharjan (4) (Table 3).

Table 2: Top 10 Major Contributors to 'Big Data' Publications.

Author	Affiliations	Publications
Cuzzocrea A.	Italy, Canada, United Kingdom	119
Ranjan R.	United Kingdom, China, Australia	70
Chang V.	China, United Kingdom, Saudi Arabia, India	69
Yang L.T.	Canada, China	60
Yu P.S.	United States, China	55
Choo K.-K.R.	Australia, United States, China	55
Leung C.K.	Canada	52
Zomaya A.Y.	Australia	52
Herrera F.	Spain, Portugal, Saudi Arabia	52
Wang G.	China	51

Table 3: Top 5 Major Contributors by Income Level in Country.

Author	Publications	Country	Author	Publications	Country
High-income			Lower-middle-income		
Cuzzocrea A.	119	Italia, Canada, United Kingdom	Pandey M.	35	India
Ranjan R.	59	United Kingdom, China, Australia	Simmhan Y.	24	India, United States
Yang L.T.	53	Canada, China	Kumar N.	23	India
Leung C.K.	52	Canada	Erritali M.	21	Morocco
Herrera F.	52	Spain, Portugal, Saudi Arabia	Rautaray S.S.	20	India
Upper-middle-income			Low-income		
Wang G.	51	China	Sun Z.	12	Papua New Guinea, Australia
Chang V.	49	China, United Kingdom, Saudi Arabia, India	Maharjan R.	4	Nepal
Dou W.	45	China	Totohasina A.	2	Madagascar
Fong S.	45	China, South Africa	Qasem S.N.	2	Yemen
Jin H.	42	China	Masabo E.	2	Uganda

From the top 20 institutions which host these researchers, 18 are Chinese research institutes and universities (Nanjing University, Tsinghua University, Wuhan University, National University, Beijing University of Posts and Telecommunications, Peking University, Beihang University, University of Chinese Academy of Sciences, Shanghai Jiao Tong University, Zhejiang University, among others). On average, each one of these universities has published around 1,500 articles on 'big data'. Only in the 12th position with about 1,000 publications, we found the University of Delhi from India and in the 16th position, the Massachusetts Institute of Technology (MIT) from the US with around 800 articles. Split by country income level, the ten Chinese institutions mentioned above have indisputable leadership among upper-middle-income countries. The top ten institutions in lower income countries are in India, Indonesia and Morocco. On average, they have published about 200 papers, except for Delhi University, which has about 1,000 publications. High-income countries include other North American institutions (MIT, Purdue, Michigan, Southern California, Washington, Minnesota and Virginia Tech), the University of Tokyo (Japan), the Politecnico di Milano (Italy) and the University of Melbourne. On average, each of these institutions has approximately 500 publications each. Finally, among lower-income countries, we found mainly African universities from Rwanda, Uganda, Ethiopia, Niger and Madagascar but also from Nepal and Syria. These institutions have published 3 to 15 articles each (Table 4).

Table 4: Top 5 Institutions Researching on Big Data by Income Level Country.

Income group	Country	Institution	Number of publications
Upper-middle-income	China	Nanjing University	2 249
		Tsinghua University	2 001
		Wuhan University	1 780
		National University	1 706
		Beijing University of Posts and Telecommunications	1 498
Lower-middle-income	India	University of Delhi	1 024
		Amity University	496
	Morocco	Mohammed V University	274
	India	Anna University	234
		School of Information Technology	198
High-income	United States	Massachusetts Institute of Technology	811
		Purdue University	591
		University of Michigan	588
	Japan	University of Tokyo	572
	Italy	Politecnico di Milano	530
Low-income	Rwanda	University of Rwanda	15
	Uganda	Makerere University	14
	Syria	Higher Institute for Applied Sciences and Technology	12
	Ethiopia	Ambo University	6
		Addis Ababa University	6

Main Journals and Conferences

In order to complete this general overview of the scientific productivity and performance indicators of the global research on 'big data', it would be valuable to identify the main conferences and scientific journals which present and publish the results of these studies (Figure 4). With respect to the former, the proceedings published in the series *Lecture Notes in Computer Science* are the main agora for this kind of research. Other important publications for conference papers include the *ACM International Conference Proceeding*, the *Communications in Computer and Information Science*, the *Advances in Intelligent Systems and Computing* and the annual conferences of the IEEE on the topic of big data. The

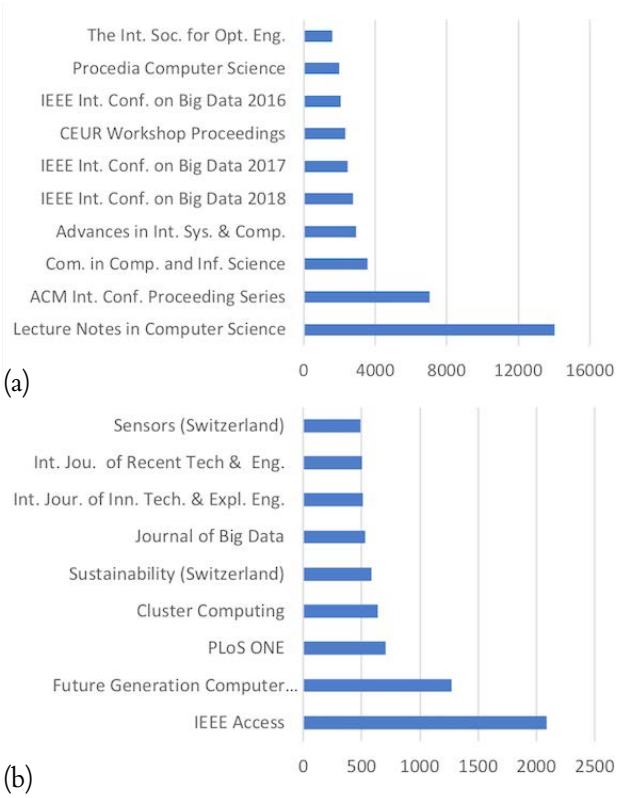

Figure 4: Main Conference Proceedings and Scientific Journals Publishing 'Big Data' Researches.

top 20 of these conferences represent 34% of the total scientific production on this subject. As previously stated, journal articles are fewer and more evenly distributed. *IEEE Access* is the leading journal with more than 470 published articles, followed by *Future Generation Computer Systems*, *Cluster Computing*, *International Journal of Innovative Technology and Exploring Engineering*, *International Journal of Recent Technology and Engineering* and the *Journal of Big Data* with an average of 200 papers each. Unlike conference publications, the top 20 journals represent only 14% of the total amount of this type of publication.

Authors who are affiliated with different countries follow different publication strategies. For example, researchers from high- and middle-income countries publish in top journals such as *IEEE Access*, *Future Generation Computer System* or *Sustainability*. However, the remaining 7 of the top ten for each income category are different. In lower-middle countries, only *IEEE Access* occupies a position in the list of the top ten journals while in low-income countries, the *Journal of Big Data* is the principal publication used to present the results of this kind of research.

All other journals used by researchers from lower-middle- and low-income countries to publish their results are different (Table 5). These numbers suggest that despite the availability of some journals which articulate the research and debates

Table 5: Main Scientific Journals Publishing 'Big Data' Related Articles by Authors' Income Country Group.

Journals	ISSN	Articles	Journals	ISSN	Articles
High-income			**Upper-middle-income**		
IEEE Access	2169-3536	357	IEEE Access	2169-3536	321
Future Generation Computer Systems	0167-739x	304	Journal of Advanced Oxidation Technologies	1203-8407	171
PLoS ONE	1932-6203	143	Boletín Técnico	0376-723x	154
Journal of Big Data	2196-1115	126	Future Generation Computer Systems	0167-739x	140
Big Data and Society	2053-9517	124	Cluster Computing	1386-7857	112
Lower-middle-income			**Low-income**		
International Journal of Innovative Technology and Exploring Engineering	2278-3075	163	Journal of Big Data	2196-1115	9
International Journal of Recent Technology and Engineering	2277-3878	162	Journal of Computer Information Systems	0887-4417	3
International Journal of Applied Engineering Research	0973-4562	112	Journal of Advanced Research in Dynamical and Control Systems	1943-023x	2
Journal of Advanced Research in Dynamical and Control Systems	1943-023x	110	Agronomy	2073-4395	1
International Journal of Engineering and Advanced Technology	2249-8958	85	American Behavioral Scientist	0002-7642	1

from different latitudes, scientific communities from different geo-economic regions reproduce themselves in different paths.

However, papers presented in conferences depict a different picture. *Lecture Notes in Computer Science*, *ACM International Conference* and the IEEE conferences are the main agoras for scientists from countries of all income categories. Other conference-based publications such as *Advances in Intelligent Systems and Computing* and *Communications in Computer and Information Sciences* appear to be more frequented by researchers from high and

Table 6: Main Scientific Conferences Where 'Big Data' Related Papers Have Been Presented by Authors' Income Country Group.

Conference	Papers	Conference	Papers
High-income		**Upper-middle-income**	
Lecture Notes in Computer Science	2508	Lecture Notes in Computer Science	1855
ACM International Conference Proceeding Series	1093	ACM International Conference Proceeding Series	912
Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018	726	Communications in Computer and Information Science	550
Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017	670	Advances in Intelligent Systems and Computing	293
CEUR Workshop Proceedings	572	Proceedings - 2018 IEEE	260
Lower-middle-income		**Low-income**	
Advances in Intelligent Systems and Computing	438	Lecture Notes in Computer Science	7
ACM International Conference Proceeding Series	411	ACM International Conference Proceeding Series	4
Lecture Notes in Computer Science	249	Proceedings - International Conference on Software Engineering	3
Procedia Computer Science	228	2018 International Conference on Advances in Big Data, Computing and Data Communication Systems, icABCD 2018	2
Communications in Computer and Information Science	175	23rd Italian Symposium on Advanced Database Systems, SEBD 2015	2

upper-middle-income countries. Similarly, conferences such as *ICSITech 2017*, *Confluence 2016* and *ICSITech 2017* attracted more researchers from lower-middle-income countries while *icABCD 2018*, *SEBD 2015* and *OBD 2015* gained attention from low-income countries (Table 6).

Citations

Last but not least, besides the volume of scientific production, the number of citations of an article denotes its relevance or influence in the field. The dataset used for this study shows that the biggest part of this scientific production (42%) has no citations at all. This means that their results and insights have not yet found an echo in their respective scientific communities. Another 38% has between one to five citations, 8% between 5 and 10 and 9% up to 50 citations. 1.8% has more than 50 citations, which represents 41% of total citations in the field. Among the latter, about 30 articles have been cited more than 500 times and about 10 articles more than 1,000 times. Undoubtedly, these articles are the seminal articles in the field (Table 7). Subjects, domains and sources of publications of this literature are very heterogeneous. The domains include bioinformatics, urbanism, computer sciences, artificial intelligence (AI), business, health and psychology, among others.

It is worth noting that despite the leadership of the Chinese production in this field, the influence of researchers affiliated to American institutions on these seminal publications is more dominant. German, British and Spanish researchers have also contributed to this core literature. Moreover, some of these publications also disclose the existence of collaboration networks between the Chinese, British and North American authors of these articles. If we look at the top 20 authors with more than one publication, it is evident that half of them are researchers affiliated to American institutions. The other half consists of Malaysia-, China-, UK-, South Africa- and Georgia-affiliated researchers (Table 8). Guizani appears to be the most influential author in the field with 25 publications and more than 2,600 citations. However, 85% of these citations refer to only one of his articles: 'Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications'. Other authors, such as those from the Spanish team behind the top article (Table 7), have been excluded from this list since they have only made one contribution to this domain of research. Most of this research is interdisciplinary. This suggests that many citations of these articles are not related to 'big data' and originate in other scientific communities. The dataset used for this study doesn't contain sufficient information to only measure the citations within the same corpus. This issue could be addressed by further research.

If we split this data according to the income categories of the countries, we find that the top ten influential authors in high-income countries are all North American affiliated researchers. On average, every author has approximately 10 publications each and between 1,700 and 2,600 citations. In

Table 7: Top 10 Most Cited Articles Containing the Keyword 'Big Data'.

Title	Journal or Conference	Authors	Year	Country	Citations
DnaSP, DNA polymorphism analyses by the coalescent and other methods	Bioinformatics	Rozas J., Sánchez-DelBarrio J.C., Messeguer X., Rozas R.	2003	Spain	4863
Rethinking the Inception Architecture for Computer Vision	Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition	Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z.	2016	United States, United KIngdom	2898
Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications	IEEE Communications Surveys and Tutorials	Al-Fuqaha A., Guizani M., Mohammadi M., Aledhari M., Ayyash M.	2015	United States, Qatar	2277
Business intelligence and analytics: From big data to big impact	MIS Quarterly: Management Information Systems	Chen H., Chiang R.H.L., Storey V.C.	2012	United States	2158
Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon	Information Communication and Society	Boyd D., Crawford K.	2012	United States	1914
Internet of things in industries: A survey	IEEE Transactions on Industrial Informatics	Xu L.D., He W., Li S.	2014	United States, China, United Kingdom	1654
Data mining with big data	IEEE Transactions on Knowledge and Data Engineering	Wu X., Zhu X., Wu G.-Q., Ding W.	2014	China, United States	1315
Data-intensive applications, challenges, techniques and technologies: A survey on Big Data	Information Sciences	Philip Chen C.L., Zhang C.-Y.	2014	China	1266
Big data: A survey	Mobile Networks and Applications	Chen M., Mao S., Liu Y.	2014	China, United States	1230
Beyond the hype: Big data concepts, methods, and analytics	International Journal of Information Management	Gandomi A., Haider M.	2015	Canada	1081

Table 8: Top 10 Most Cited Authors With More Than One Publication on 'Big Data'.

Autor	Publications	Citations	Country
Guizani M.	25	2616	United States
Chiang R.H.L.	5	2483	United States
Chen H.	28	2434	United States
Al-Fuqaha A.	7	2394	United States
Mohammadi M.	3	2387	United States
Chen M.	29	2343	China
Gani A.	24	2338	Malaysia
Aledhari M.	4	2285	United States
Storey V.C.	8	2243	United States
Crawford K.	8	2230	United States

upper-middle-income countries, researchers with Malaysian and Chinese affiliations are the most influential; on average, they have more publications than their North American peers (19) and between 1,300 and 2,300 citations each. The top 10 researchers from lower middle income are mostly from India. On average, they have published 14 papers and have been cited between 190 and 1000 times each. Finally, among the researchers from lower-income countries, Sun, affiliated with the PNG University of Technology at Papua New Guinea, is the most prolific author with about 14 papers and over 90 citations. Other researchers from these countries include researchers from Syria, Ethiopia, Yemen, Benin, Madagascar and Nepal. All of them have about 2 publications and 3–9 citations each (Table 9).

Thematic Analysis

Distribution of Publications by Research Area

Most of these publications are contributions to computer science and engineering (81%). Contributions to medicine and social sciences are only 4% each and mathematics, biology, business, physics, earth, decision and environmental science are 1% each. The remaining contributions of 4% are shared by the other 17 fields of the Scopus classification (Figure 5). However, these numbers can be deceptive since several of these publications are situated at the intersection of different fields. In fact, 48% of these publications are categorised in at least two different research areas, 11% in three and 4% in four or more areas, which highlights the interdisciplinary character of this research field. If we disregard the computer science and engineering fields in the publications categorised under at least two areas, the consequent distribution is more heterogeneous. Mathematics represents about 18% of these contributions, those in decision science 11%, social science 7% and medicine 5%. The rest of the fields, which collectively represent 18%, double their participation (Table 8). Finally, if considered individually, the publications in every field in relation to the total number of publications in decision science rise up to

Table 9: Top 5 Most Cited Authors on 'Big Data' per Income Group Country.

Countries	Autor	Publications	Citations	Countries	Autor	Publications	Citations
High-income				Upper-middle-income			
United States	Guizani M.	25	2616	China	Chen M.	29	2343
	Chiang R.H.L.	5	2483	Malaysia	Gani A.	24	2338
	Chen H.	28	2434	United States	Storey V.C.	8	2243
	Al-Fuqaha A.	7	2394	Malaysia	Yaqoob I.	15	1990
	Mohammadi M.	3	2387		Hashem I.A.T.	18	1877
Lower-middle-income				Low-income			
India	Dubey R.	23	1075	Papua New Guinea,	Sun Z.	14	93
	Manogaran G.	21	478		Pambel F.	2	9
	Simmhan Y.	28	403	Syria	Aljoumaa K.	2	9
	Lopez D.	16	359		Jafar A.	2	9
	Goudar R.H.	6	330	Yemen	Qasem S.N.	2	6

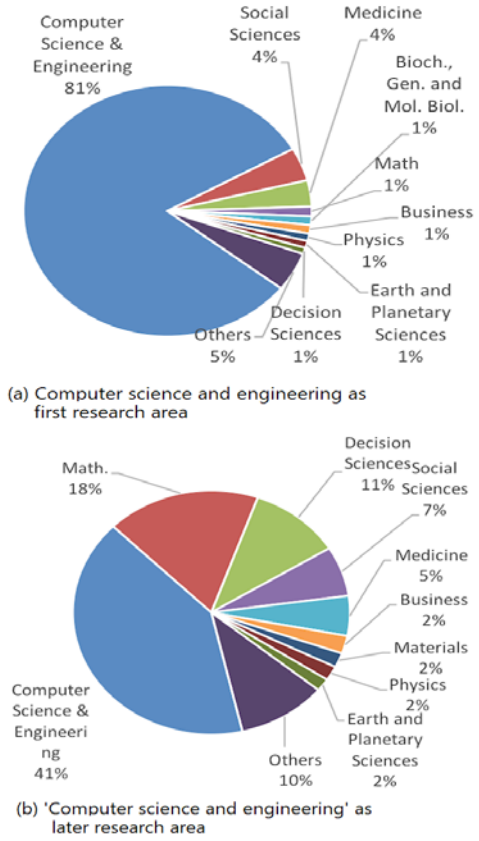

Figure 5: Publications on Big Data by Research Area.

14%, social sciences to 11%, medicine to 7% and business to 6%. This reveals that approximately 15% of these publications are at the intersection between the social sciences, decision sciences and engineering and computer science fields.

Research Areas by Year

Big data research started as a computer science subject but has rapidly spread in other research areas. The publications before 2013 were mainly contributions to the computer science and engineering field (58%). The share of this field has, however, been reduced to 35% between 2017 and 2019. Thus, this reduction reflects the diversification of approaches and development of interdisciplinary projects with scientists from other areas, particularly mathematics, social and decision sciences, medicine and business (Figure 6).

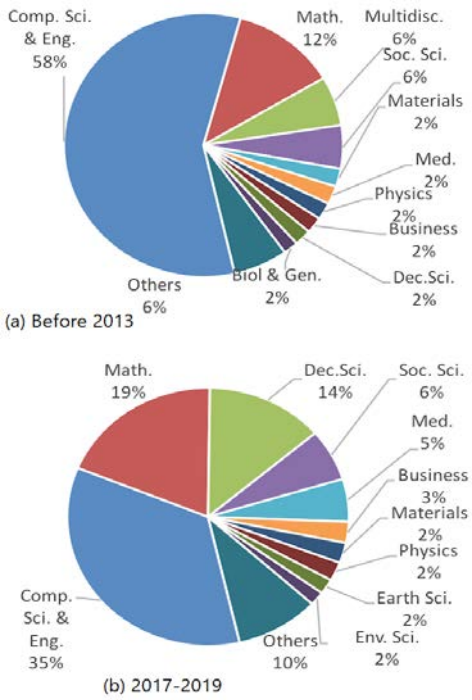

Figure 6: Publications on Big Data by Area and Period of Time.

Research Areas by Country

By country income group, we observe slight differences between high and upper-middle-income countries with relatively more contributions from computer sciences, engineering, social sciences, medicine and business fields in

the high-income countries and more mathematics, decision sciences and materials in the upper-middle-income countries. Lower middle and low-income countries showcase an increased number of differences for this indicator. The contributions from the former are relatively more concentrated in computer sciences and less in social science and medicine than all the other categories. The latter has relatively fewer contributions in computer sciences and engineering (33%) and relatively more in the decision and environmental sciences (Figure 7).

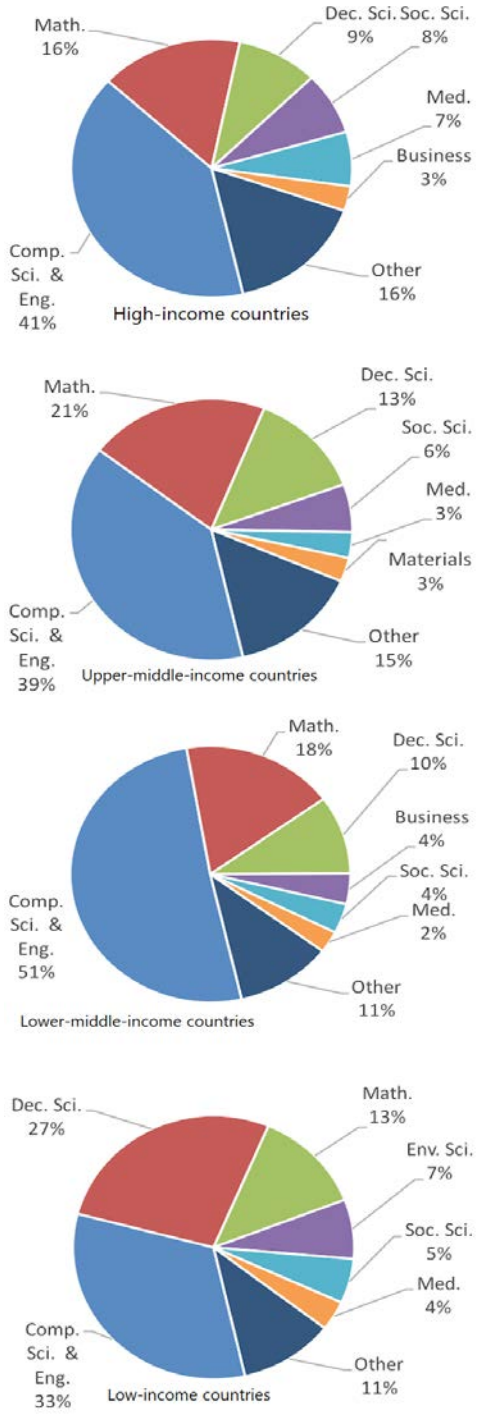

Figure 7: Publications on Big Data by Area and Income Group.

If we look at the leading countries by research area, we can see that they more or less follow the same pattern as the global analysis: China and the US lead almost every field and particularly the most important ones. However, there are some areas where the US maintains leadership over Chinese research. It is evident in the case of social sciences, agriculture, arts and humanities, biochemistry, economics, business, health, immunology, medicine, psychology, pharmacology and neurosciences. It is equally interesting to note that apart from the main players, not all countries have produced research in all areas, and its development is uneven. This suggests a certain degree of specialisation in some countries (Figure 8).

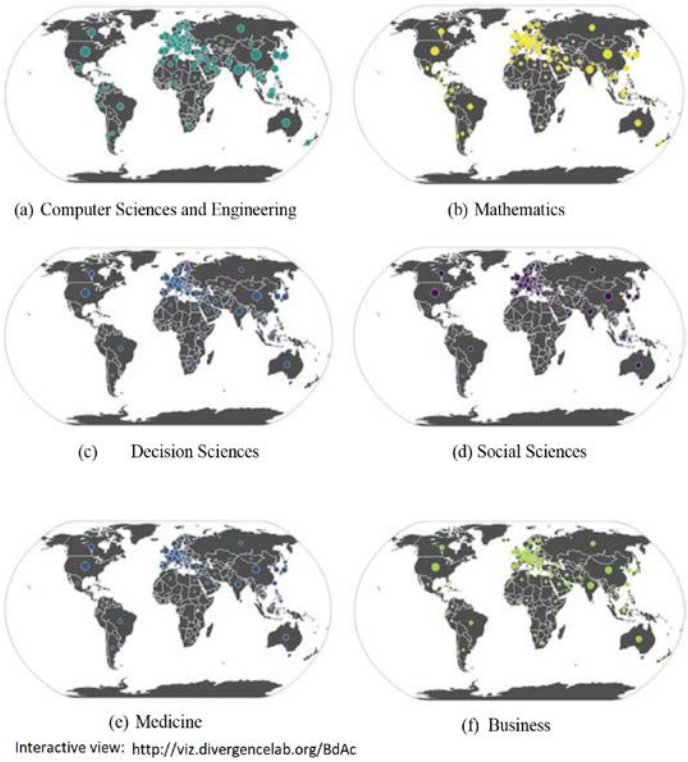

Figure 8: Publications on Big Data by Area and Country.

Keywords Trends over Time

Beyond the classification by research areas, keywords used to index these publications offer us a more specific picture of the main topics studied within big data literature. They also reveal the predominant trends and evolution of this research over time. The word clouds in Figure 9 show three different phases within this literature. At first, the publications were more diverse, but they were also more general and mainly oriented towards technical challenges, impacts and potential applications. By 2015, the research focused on technical and methodological issues. By the end of the decade, the focus of the research seems to have shifted from the application of the accumulated knowledge and techniques to the development

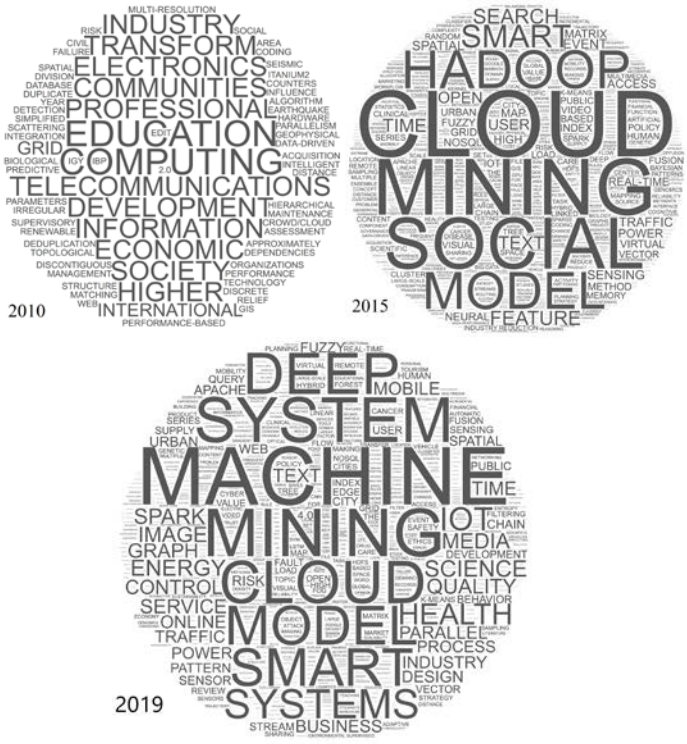

Figure 9: Keywords Evolution in Publications 2010, 2015 and 2019.

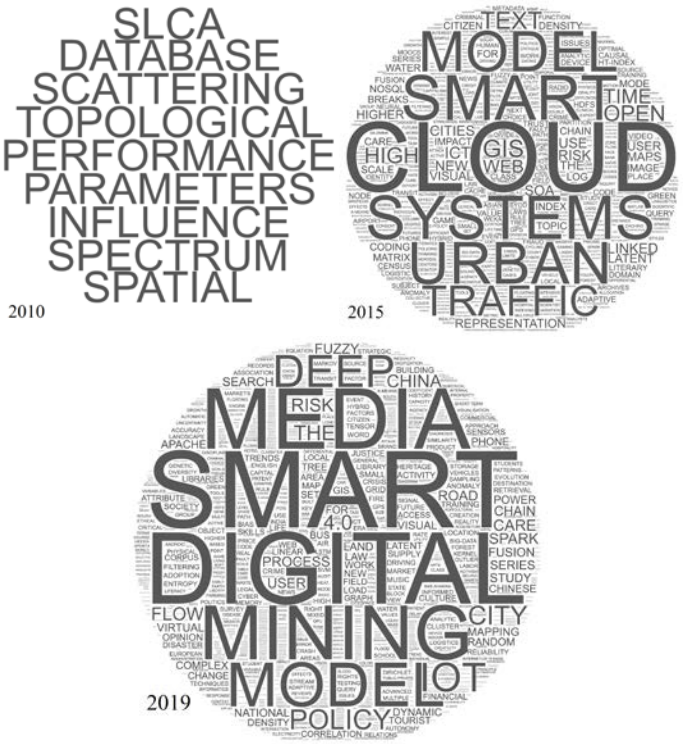

Figure 10: Keywords Evolution in Social Science Publications 2010, 2015 and 2019.

of artificial intelligence and machine learning techniques in order to study other specific problems in different fields.

A review of the subjects of the publications within the field of social sciences shows that despite the importance which big data has acquired in recent years, there is a deficit of academic production in the role and effects that it has on the State's decision-making processes, design of public policy instruments and its relations with other sectors of society.[26–28] Figure 10 shows word cloud with the main keywords of the articles published each year. Words' sizes reflect the relative frequency of each keyword within each year. As we can see, initially, the few articles on social sciences using the concept were related to technical problems. However, by 2015, a great deal of these publications seem to have focused on urban problems such as traffic or smart cities along with technical and methodological issues. By 2019, digital media, artificial intelligence, smart cities and systems seem to have become the main concerns of social scientists using this concept (Figure 10).

The main topic picture also changes if we look at the differences among group income countries. Publications from high-income countries are more focused on the application of this knowledge to machine learning, analytics, computing, management and to a lesser extent to social and smart systems problems. Upper middle-income countries follow a similar trend, but they seem to be more advanced in the application of this knowledge to the development of artificial intelligence systems and techniques. On the other hand, lower-middle and low-income countries present quite a different picture. The latter countries are focused on the research of technical issues related to big data such as mining, the Internet of things, smart systems, etc. Even if they are less prolific, the latter countries are more diverse and centred on leveraging this knowledge in learning, predicting, detecting and, to some extent, in social problems (Figure 11).

Collaboration networks

Figure 12 depicts the evolution of international collaboration networks in the last decade. The scientific production of big data began in various national-centred clusters (the US, China and some European countries) with few relations between them. This feature quickly changed and by 2014, different collaboration networks were established all around the world with the US and China as the main nodes and European countries as the main intermediaries. This basic structure has densified during the following years, thereby increasing connections and reaching new countries and scientific communities. An interesting fact regarding the structure and evolution of this global network of collaborations is that European institutions played a major role in the development and circulation of this knowledge towards middle- and lower-middle-income countries. In Latin America, 48% of the 1,768 registered external links between 2010 and 2019 were established with researchers affiliated to European institutions

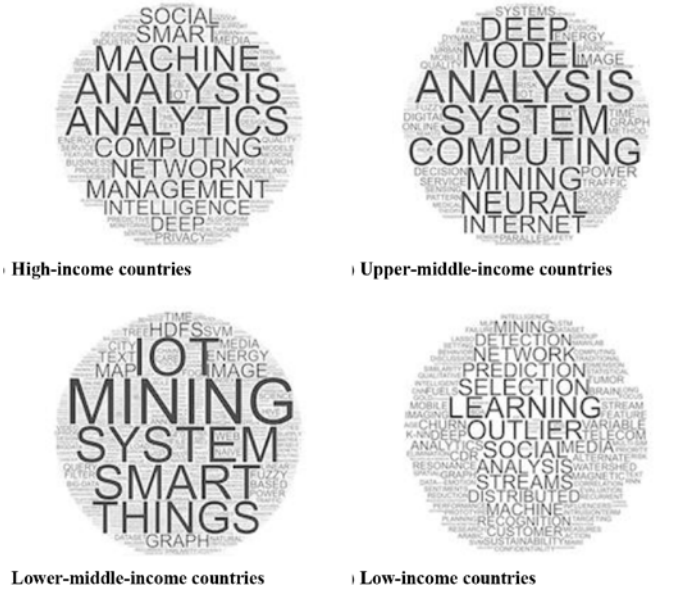

Figure 11: Keywords in Publications from 2010, 2015 and 2019 by Country Income Group.

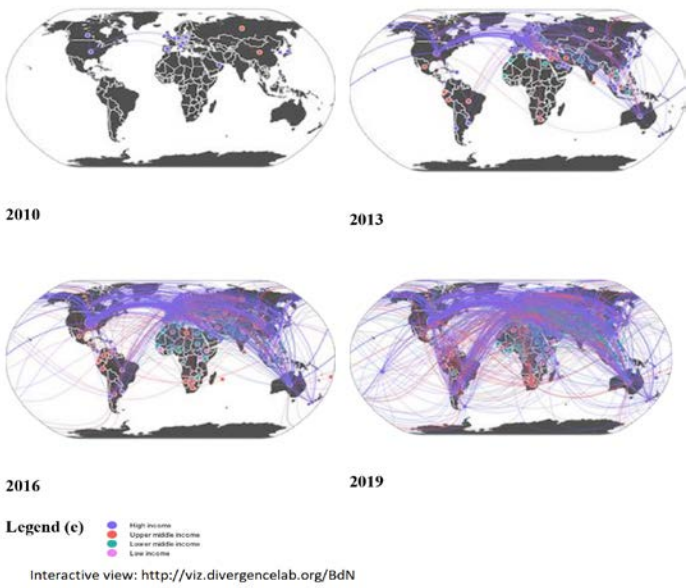

Figure 12: Co-authoring networks evolution 2010-2019.

while North Americans represented only 15%. This was also the case with Africa, where approximately 40% of 1,300 external collaborations were with European countries. Asian networks seem to be more self-centred with more than 36% of the 16,000 links within the continent, but they also have strong connections with North America (27%) and Europe (25%).

If we analyses international collaborations by author, we get a relatively scattered landscape where about 168,000 individuals mostly collaborate in small independent networks: 62% have only one link, 13% have more than five links and only 5% have more than 10 links. Figure 13 shows the core network of the scientific community working on big data which connects about 1,000 researchers with more than 10 publications with approximately 9,000 collaborators in their countries and abroad: 60% collaborate within the same country and only 18% collaborate with authors from countries with different income categories. Most of these collaborations are between Chinese, American and European researchers. There are collaborations between South American researchers and researchers in North America as well.

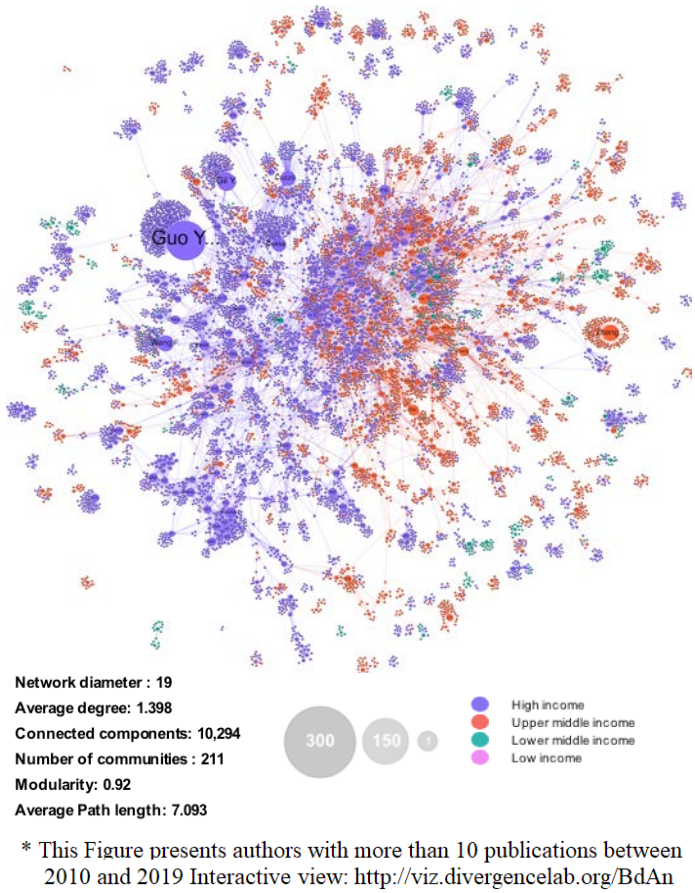

Figure 13: International collaboration networks by authors*.

Finally, if we take a look at the collaboration networks by affiliation institution, we get a more structured global network of about 44,000 actors organised around two hubs: one formed by Chinese institutions (17%) and the other formed by North American (18%) and European institutions (30%), especially universities (42%). These two hubs directly interact through a great variety of smaller actors, including institutions in different continents, regions and geo-economic categories. The average institution has approximately 100 links with other institutions and the biggest ones (<1%) have more than 1,000 links. Figure 14 shows the network of affiliation institutions collaborating on research regarding big data with institutions

in the same country (47%) and abroad (53%). Among the latter, only 17% work with institutions in different income country categories. European institutions have links with African, Latin American and Asian institutions. The data also shows some south–south collaborations among countries in Latin America, Africa and Asia.

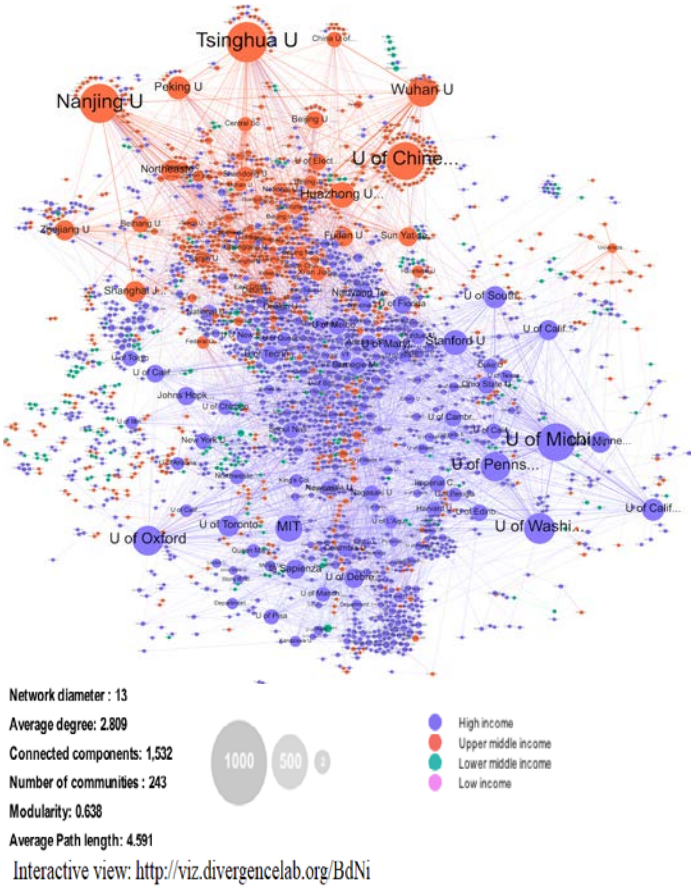

Figure 14: International collaboration networks by institutions.

CONCLUSION

We have presented an analytical mapping of research on big data from a set of more than 73,000 entries of the Scopus database which were published in the last two decades. We evaluated this corpus for three main aspects: (1) scientific productivity and performance indicators; (2) main research areas and thematic trends and (3) collaboration networks between authors, research institutions and affiliation countries. We directed special attention towards the main relations and differences between scientific communities working on this subject in countries situated in different regions and economic conditions. Our main findings show that scientific productivity has exponentially increased since 2010 with an average growth rate of approximately 90% per year. China and the United States lead the scientific production on big data. According to country income level, Chinese, South Africans and Colombians lead upper-middle countries, while India and Morocco lead lower-middle-income countries. Of the top 20 institutions hosting big data researchers, 18 are Chinese, followed by the University of Delhi and MIT. *Lectures Notes in Computer Science* is the most important agora on big data, and *IEEE Access* is the principal academic journal on this topic. Regarding citations, besides the higher productivity of Chinese researchers and institutions in this field, American contributions remain the most influential and the most cited articles and authors come from the US. Despite the dynamism of the field, about 2% of the articles concentrate 40% of the citations of the field, while 42% of these publications have no citations whatsoever. The main scientific field publishing on big data is computer science and engineering (81% of the publications in the entire period) followed by medicine and social sciences. However, since 2017, the relative importance of this area has reduced to 35% due to the development of interdisciplinary projects in mathematics, social and decision sciences, medicine and business. The keywords trend over time shows that, by 2010, literature was mainly oriented towards technical challenges, impacts and possible applications of these technologies; by 2015, it focused on technical and methodological issues; and in the past few years, it has shifted towards the development of AI and machine learning techniques. Lastly, we observe an important scientific collaboration activity: only 12% of the publications are individual contributions. However, most collaboration networks are nationality-based; only 19% belong to international networks. This has changed over the last five years from national-centred clusters to a more international network, where the US and China are the main nodes. European countries appear to be the main intermediaries in the circulation and development of the knowledge in this field with countries from Africa and South America. Although to a lesser extent, we have also detected some south–south collaborations between Latin America, Africa and Asia. Thus, we have presented a detailed characterisation and a comprehensive analysis of the big data global research over the last decade. This research updates previous bibliometric works by extending the previously analysed corpus and exploring an unresearched database. Our most important contribution to bibliometric analysis is the insights we provide on the differences in scientific productivity, research areas and trend topics, as well as the collaboration networks between countries from different geo-economic conditions. These differences highlight the uneven distribution and circulation of big data knowledge behind the growth in publications over the last decade. Further research could provide a deeper and more detailed characterisation of research in this field in specific regions and countries, as well as in specific research areas and topics.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Undersea technology. BOMEX becomes BOMAP for Big Data handling job. Undersea Technology. 1970;11(2):28-9.
2. Dwinnell W. Tools that deal with big data: Modeling and analysis. PC AI. 2001;15(5):42.
3. Kenwright D, Banks D, Bryson S, Haimes R, van Liere R, Uselton S. Automation or interaction: What's best for big data? Proceedings of the IEEE visualization conference. 1999;1:491-5.
4. Tremblay M, Grohoski G, Burgess B, Killian E, Colwell R, Rubinfeld PI. Challenges and trends in processor design. Computer. 1998;31(1):39-48. doi: 10.1109/2.641976.
5. Diebold FX. "Big Data" dynamic factor models for macroeconomic measurement and forecasting: A discussion of the papers by Lucrezia Reichlin and by Mark W Watson. In: Dewatripont M, Hansen LP, Turnovsky SJE, editors. Advances in economics and econometrics: Theory and applications, Eighth World Congress. Cambridge University Press; 2003;115-22.
6. Kriksciuniene D, Liutvinavicius M, Sakalauskas V, Tamasauskas D. Research of customer behavior anomalies in big financial data. Kuwait: Institute of Electrical and Electronic Engineers (Institute of Electrical and Electronics Engineers); 2003. p. 91-6.
7. Gupta D, Rani R. A study of big data evolution and research challenges. Journal of Information Science. 2019;45(3):322-40. doi: 10.1177/0165551518789880.
8. Laney D. 3D data management: controlling data volume, velocity and variety [internet]. p. Gatner2001 [cited Jul 20 2020]. Available from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
9. Mashey J. Big data and the next wave of InfraStress problems, solutions, opportunities [internet]. Monterey, CA: Usenix Association; 1999 [cited Jul 20 2020]. Available from: https://www.usenix.org/legacy/publications/library/proceedings/usenix99/invited_talks/mashey.pdf.
10. Weiss SM, Indurkhya N. Predictive data mining: A practical guide. 1st ed. San Francisco: Morgan Kaufmann Publishers; 1997.
11. Chaorasiya V, Shrivastava A. A survey on Big Data: Techniques and Technologies. International Journal of Research and Development in Applied Science and Engineering. 2015;8(1):4.
12. Halevi G, Moed H. The Evolution of Big Data as a research and scientific topic: Overview of the literature [internet]; 2012. Research Trends [cited Jun 23 2019]. Available from: https://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/.
13. Emani CK, Cullot N, Nicolle C. Understandable Big Data: A survey. Computer Science Review. 2015;17:70-81. doi: 10.1016/j.cosrev.2015.05.002.
14. Parlina A, Ramli K, Murfi H. Theme mapping and bibliometrics analysis of one decade of big data research in the Scopus database. Information. 2020;11(2):1-26. doi: 10.3390/info11020069.
15. Montes L. Datos a explotar antes de la llegada de una auténtica 'superinteligencia' [internet]; 2017. El Mundo [cited Jun 23 2019]. Available from: https://www.elmundo.es/economia/2017/02/02/5893527c268e3eb04b8b46f7.html.
16. Ross JM. Roger Magoulas on big data [internet]. O'Reilly Radar2010 [cited Jun 23 2019]. Available from: http://radar.oreilly.com/2010/01/roger-magoulas-on-big-data.html.
17. Ghemawat S, Gobioff H, Leung S-T. The google file system [internet]. SIGOPS Oper Syst Rev. Proceedings of the 19th ACM symposium on Operating Systems Principles. Bolton, Landing, NY. 2003;37(5):29-43. doi: 10.1145/1165389.945450.
18. Dean J, MapReduce GS. Simplified data processing on large clusters. In: OSDI: Sixth symposium on operating system design and implementation. San Francisco: 2004; '04. p. 137-50.
19. Raban DR, Gordon A. The evolution of data science and big data research: A bibliometric analysis. Scientometrics. 2020;122(3):1563-81. doi: 10.1007/s11192-020-03371-2.
20. Singh VK, Banshal SK, Singhal K, Uddin A. Scientometric mapping of research on 'Big Data'. Scientometrics. 2015;105(2):727-41. doi: 10.1007/s11192-015-1729-9.
21. Tseng SF, Won YL, Yang JM. A bibliometric analysis on data mining and big data. International Journal of Electronic Business. 2016;13(1):38-69. doi: 10.1504/IJEB.2016.075333.
22. Peng Y, Shi J, Fantinato M, Chen J. A study on the author collaboration network in big data. Information Systems Frontiers. 2017;19(6):1329-42. doi: 10.1007/s10796-017-9771-1.
23. López-Robles JR, Otegi-Olaso JR, Porto Gomez I, Gamboa-Rosales NK, Gamboa-Rosales H, Robles-Berumen H. Bibliometric network analysis to identify the intellectual structure and evolution of the big data research field. Lecture notes in Computer Science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). 2018;11315 LNCS:113-20.
24. World Bank country and lending groups [internet]; 2020 [cited May 7 2021]. Available from: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.
25. Fabre G. China's digital transformation. Why is artificial intelligence a priority for Chinese R&D? FMSH - Working Papers [internet]. Vol. 136; 2018 [cited Mar 5 2021]. Available from: https://wpfmsh.hypotheses.org/994.
26. Fredriksson C. Big data creating new knowledge as support in decision-making: Practical examples of big data use and consequences of using big data as decision support. Journal of Decision Systems. 2018;27(1):1-18. doi: 10.1080/12460125.2018.1459068.
27. Gamage P. New development: Leveraging 'big data' analytics in the public sector. Money and Management. 2016;36(5):385-90. doi: 10.1080/09540962.2016.1194087.
28. Mahrenbach LC, Mayer K, Pfeffer J. Policy visions of big data: views from the Global South. Third World Quarterly. 2018;39(10):1861-82. doi: 10.1080/01436597.2018.1509700.