Journal of Information Systems and Technology Management – Jistem USP
Vol. 19, 2022, e202219011
ISSN online: 1807-1775
DOI: 10.4301/S1807-1775202219011 Original Article

SOCIAL SCIENCES AND HUMANITIES ON BIG DATA: A BIBLIOMETRIC ANALYSIS

Gastón Becerra[1] https://orcid.org/0000-0001-9432-8848
Cristian Ratovicius[2] https://orcid.org/0000-0003-0420-8936

[1]Universidad de Buenos Aires, CONICET, Buenos Aires, Argentina
[2]Universidad Abierta Interamericana, Facultad de Tecnología Informática. Centro de Altos Estudios en Tecnología Informática. Buenos Aires, Argentina

ABSTRACT

The purpose of this paper is to provide a comprehensive bibliometric review of social science, psychology, and humanities literature focusing on big data. Methods: Production and authorship trends, topics and areas as well as citations were analyzed by means of conducting a bibliometric analysis of a corpus of 5,500 Scopus articles published from 2010 to 2020. Findings: Analysis revealed similarities and differences among social science, psychology, and humanities literature in terms of publication, framing, and referencing trends as compared with the general big data literature: both fields show a steady increase, although the increase rate slowed down as from 2015; text production of both specific and general fields is led by just a few countries, with the USA and China being on top of the ranking; single authorship has been decreasing in both fields; the specificity of big data framing, in social sciences and humanities, has been identified with a critical view that surpass the ethical considerations, to include the social construction of datasets, the political and ideological uses of big data, and the discussion of its philosophical and epistemological foundations. Value: To the best of our knowledge, this is the first study to provide a comprehensive view on social sciences and humanities big data bibliometrics while providing context to compare results.

Keywords: Big data, social sciences, humanities, bibliometric analysis, citation analysis.

Manuscript first received: 2021-05-31. Manuscript accepted: 2021-11-24.

Address for correspondence:
Gastón Becerra, Universidad de Buenos Aires, CONICET, Buenos Aires, Argentina. E-mail: gaston.becerra@gmail.com
Cristian Ratovicius, Universidad Abierta Interamericana, Facultad de Tecnología Informática. Centro de Altos Estudios en Tecnología Informática. Buenos Aires, Argentina. E-mail: cratovicius@yahoo.com

CC BY

INTRODUCTION

By using bibliometric analysis, this work explores scientific literature about big data in social sciences, psychology, and humanities. We aim to identify trends about authorship and collaboration, research topics, and the most influential works. Such objective is a specific step within a broader project aiming at comparing how big data is framed in different social systems, such as mass media, science or politics.

The term "big data" began being used in the late 1990's within the IT sector. It refers to the (technical) challenges of handling a vast amount of information. In a famous consultancy piece, Laney (2001) encapsulated these challenges by referencing 3 v's -volume, velocity, and variety-, a formula that expanded to include other v-words, such as visualization or value. In irony, "vexatious vagueness" is the implicit v-word, Halavais (2015) says. From here, big data has expanded to different areas of social life, wherein critical case studies are still required to understand the social meaning of big data (Beer, 2016).

At regards mass media discourse, in another work we've proposed that communications about big data usually include 2 highly debatable elements: a premise regarding the availability of huge volumes of data that can be exploited; a promise of actionable information that will reach all aspects of life. These elements are part of a widespread belief, rhetoric and mythology that relate big data with other socio-technical developments, such as artificial intelligence and algorithms, promising an objective, optimal, value-free and conflict-free social future (boyd & Crawford, 2012; Dijck, 2014; Sadin, 2018).

Scientific interest on big data has also been on the rise over the last decade. According to several studies (e.g., Belmonte et al., 2020; Liu et al., 2019), and drawing on different sources, scientific big data literature has increased at a X2 yearly rate for the 2010-2014 period; and, although this trend has slowed down in the last years, the number of papers per year has never decreased up to 2020. Other studies have estimated that 7-10% comes from social sciences and humanities (Kalantari et al., 2017; Liu et al., 2019). Facing big data, social sciences and humanities found a big challenge: to criticize and discuss the premise and the promise of its rhetoric and mythology, to dissect the social beliefs and ideologies around it, to illuminate the extent of social and ethical issues that it brings, and to re-shape big data from within to advance our understanding of social reality.

In this work we focus on literature from social sciences, psychology and humanities & arts. We do so in a bibliometric approach that aims at assessing and analyzing, in a quantitative manner, trends in the publications. Specifically, we are interested in 3 questions for which bibliometric analysis has proven to be an useful and valid methodology:

RQ1. Production and authorship trends: Which countries and institutions have contributed most in terms of paper count? How countries and institutions collaborate with each other? What is the average number of authors per publications?

RQ2. Topics and research areas: What are the key topics that are addressed in publications?

RQ3. Citation: What are the top cited publications? What are the top venues for the most cited publications?

CC BY

To answer these questions, this study works on a collection of 5,500 papers published within the 2010-2020 period and written in the social sciences, psychology, and humanities areas including "big data" in the title, abstract, and/or keywords. The rest of the article is organized as follows: #2 (Related works) presents a brief review of bibliometrics studies on big data not limited to the aforementioned disciplines; #3 (Methodology) presents the criteria for data extraction and data processing; #4 (Results) presents an in-depth analysis of our corpus; #5 (Conclusions) summarizes our findings and compares with those in the "general" big data literature; #6 (Discussion) discusses the value and limitation of our study.

RELATED WORKS

Several papers have used bibliometric analysis – with different goals and analysis levels – to map big data scientific works. In the remaining of this section, we will compare and summarize these works regarding authorship, topics, and citations in order to provide a benchmark for our analysis (Table 1).

Table 1. Related works

Paper	Corpus	Period	Source	Search.criteria
Ahmad, 2020	33,623	2008-2017	Scopus	"big data"; not limited to computer sciences
Liu, 2019	4,070	2013-2018	Scopus	"big data"
Kalantari, 2017	6,572	1980-2015	WoS	several keywords defined by expert surveying
Zhang, 2019	5,840	2000-2015	WoS	search keywords from technologies (e.g., "big data", "map reduce")
Belmonte, 2020	4,240	2010-2019	WoS	"big data" and "machine learning"
Raban, 2020	7,786	2010-2019	WoS	"big data" and "datascience"
Liang, 2018	10,637	1990-2017	SSCIE	"big data"

RQ1. Authorship and collaboration trends are the first topic we will analyze.

In terms of production by countries, People's Republic of China leads the ranking, followed by USA, sharing 50% of the different corpora. USA was leading the production up until 2015 approx, when China production exploded. These countries appear in inverse order only in Kalantari et al. (2017) and Belmonte et al. (2020) which includes a much a much broarder time span and set of search criteria. USA and China are the countries that collaborate the most. Comparatively, Canada, Australia, Switzerland, and Japan show less production than the aforementioned countries, but have a higher centrality in collaborations. South American, Middle-eastern, and African countries show the lowest production levels.

CC BY

Institutional affiliation has been analyzed in far less extent. Ahmad et al. (2020) reports that Chinese institutions not only rank on top of the production but also in the concentration of papers per institution; USA production seems to be less concentrated. Again, Belmonte et al. (2020) registered an inverse ranking, and Zhang et al. (2019) highlights that world-class universities from USA, although not that prominent as Chinese institutions, are much more interconnected in terms of collaboration. Liu et al. (2019) reported that only 2.8% of the papers were produced jointly between academia and corporations.

Most of the reviewed literature report collaborative authorship as the standard practice: only near 10-15% of the papers were written by a single author, more than 40% have at least 4 authors, and even there's a 1-2% of the papers that have more than 10 authors. Most papers also report that single authorship has been declining, and that there seems to be a trend toward including more authors. Liu et al. (2019) suggest that single authorship ranks last in terms of impact.

RQ2. Contents is the second issue to be explored. Most of the revised papers draw on keyword frequency and/or keyword correlation to identify topics and sub-areas of research.

After comparing the literature, we can identify 4 general topics: **(1)** The first topic include big data within a larger group of terms originated in computer science but cannot be univocally confined to that space anymore, such as, machine learning, data mining, cloud computing. Also there are references to tools and solutions for big data, such as Hadoop, MapReduce. This topic can be seen on most of the revised surveys. **(2)** The second topic is business intelligence and big data analytics, focusing on the application-oriented side. Keywords that most suggest this topic are decision making, commerce, information management, and (business sector) management. **(3)** On most surveys, there appears a third topic related to analysis notions, such as machine learning, model, statistics. Most of these topics have been around for decades and lost prominence by 2015, and regained attention thanks to the re-discovery of their potential with big data. This trend has been registered by Kalantari et al. (2017), Belmonte et al. (2020), and Zhang et al. (2019), the latter referring to some of these as "sleeping beauties". **(4)** Ahmad et al. (2020), Zhang et al. (2019) and Liang & Liu (2018) identified a final topic that deals with big data integration with novel technologies that provide new sources and streams of data, such as Internet of Things (IoT), bioinformatics and social big data (e.g., social networks and apps). These are all emerging topics, and are also the first to highlight social issues and critical concerns, e.g., about privacy protection and human/patient care.

Finally, Liang & Liu (2018) paper include a brief report on social science and humanities corpus, constructed by subsetting their big data and business intelligence corpus. Authors identified 10 topics by analyzing keywords and citation networks: the earliest incidence corresponds to medical-related issues, "where big data started in social sciences. It also echoes the fact that health care service is the most important field for big data applications"; the second topic, one that remains steady until 2015, renders big data as an emerging technology [On this topic, authors include a reference to Kitchin (2014), a work that both presented the first comprehensive sociological analysis of big data and a critical program for its analysis]; another important topic is research on agenda setting [Authors mention a reference to (probably) Lazer et al. (2014)], which spans between 2007-2009. "In 2014, there were still many papers referring to agenda setting … but almost none was cited after 2015. It may indicate that the studies involving agenda setting in big data had come to an end"; the rest of the topics they mention relate to business processes, and analysis techniques. However, this section is very brief and is not that clearly reported what they mean by the label of the topics, so it is very hard to engage in a discussion with them.

CC BY

RQ3. The last issue we are interested in is citation, as a mean to identify influential works and venues.

In terms of citation distribution, Ahmad et al. (2020) report that over 50% of publications are not cited, and that 80% of citations were received by approx. 15% of the publications. Authors highlight that this statistics is similar to those reported by Garounsi and Mäntylä (2016) for software engineering. Several of the papers reviewed report that the most cited works correspond to early survey papers, both on the general term of big data and on specific techniques, such as, "Big data: a survey" Chen et al. (2014), "Lexicon-based methods for sentiment analysis" Taboada et al. (2011), or "Data mining with big data" Xindong Wu et al. (2014). These are all, arguably, technical papers that map the state of the art, and the upcoming challenges in the field. There are also some critical analysis, reflections and/or experimentation drawing on big social data, such as the study on "emotional contagion" in Facebook (Kramer et al., 2014) or the uses of Google for health trends detection (Lazer et al., 2014). In these, controversy may play an important factor, because of the claims received about the lack of informed consent and the impossibility to opt-out of the experiment by the former, and the introduction about "big data hubris" in the latter. Another exception is "Critical Questions for Big Data" by danah boyd and Kate Crawford (2012), which poses concerns about big data mythology and possible missuses.

METHODOLOGY

We selected the Scopus database to construct our corpus because it offers complete metadata, abstracts, and references for the articles. Others databases, such as Jstor and Ebsco, were also analyzed but, although full-text content was available in some cases, the number of articles was far fewer and lacked some abstracts and references.

To make the search, we used the criteria "big data" in title, abstracts, and keywords. We limited the subject area to social sciences, psychology and humanities and arts, set the language to English, and the publication type to journal articles. The dataset was exported on the first of December of 2020. Although initially no date limits were applied, most of the publications were comprised between 2010-2020. From the original result of 5,610 records, we kept 5,500 that were included in this shorter time span, with abstracts.

Most of our research questions did not require more than descriptive statistics and visualization techniques. In order to answer them, we did the following preprocessing:

- For **RQ1** we parsed the metadata provided by Scopus on authors, and extracted their country and affiliation. To classify affiliation as academic (and to separate from corporation), we looked for "universi," "college," "school," "institu," "academ" and "politec." Collaboration between nations and types of institutions was analyzed using co-authorship as a proxy, a decision that has been discussed in the bibliometric literature (Ponomariov & Boardman, 2016). To express the author's collaboration we used Subramanyam Index (Subramanyam, 1983) which computes articles with +2 authors over the total of articles. In networks (such as collaborating countries or correlated keywords) we calculated communities using the Igraph implementation of an eigenvectors of matrices (Newman, 2006).

CC BY

- For **RQ2** we changed keywords and abstracts to lowercase, removed symbols and numbers, and replaced "big data" with "bigdata." We also stemmed words in abstracts. For keyword analysis we built an associative enriched network, that shows top 25 correlations to "big data" (1st degree), and the subsequent 100 correlation to those (2nd degree), creating a centered network with easily identifiable clusters. Also, to complement this analysis, we performed a structural topic modeling over the abstracts using the stm R package (Roberts et al., 2019), which posits different distributions of the corpus' vocabulary as topics, and calculates the proportional mix of them for each document, taking in consideration some metadata as co-variate (in our case: the year of publication). An important decision is the number of topics to be created, since this must be provided as a parameter (K). For this we performed several statistical test (held-out likelihood, semantic coherence) which resulted in K=50 topics, but since most of these were highly connected, we settled for a 25 topics solution, that we latter grouped in 5 general categories, after qualitative analysis.
- Although citation should not be taken as a proxy for quality of a publication, it is indeed a valuable metric to assess visibility and influence (Ebrahim et al., 2014). Thus, for **RQ3** we worked with the number of citation to the articles, and with the references provided by Scopus, which required extensive parsing, done by using the RubyGem version of Anystyle [https://anystyle.io/]. Although this provided excellent results for some fields such as authors and titles, it had inconsistent results for journals.

All data pre-processing (with the mentioned exception of the citation parsing), analysis and visualization was done using statistical software R (Team R Core, 2018).

RESULTS

According to paper count, the big data literature on social sciences and humanities has grown yearly in over a X3 rate for the 2010-2014 range, then this trend has slowed down near 2015, but up to 2020 the number of papers per year never decreased (Figure 1; Table 1).

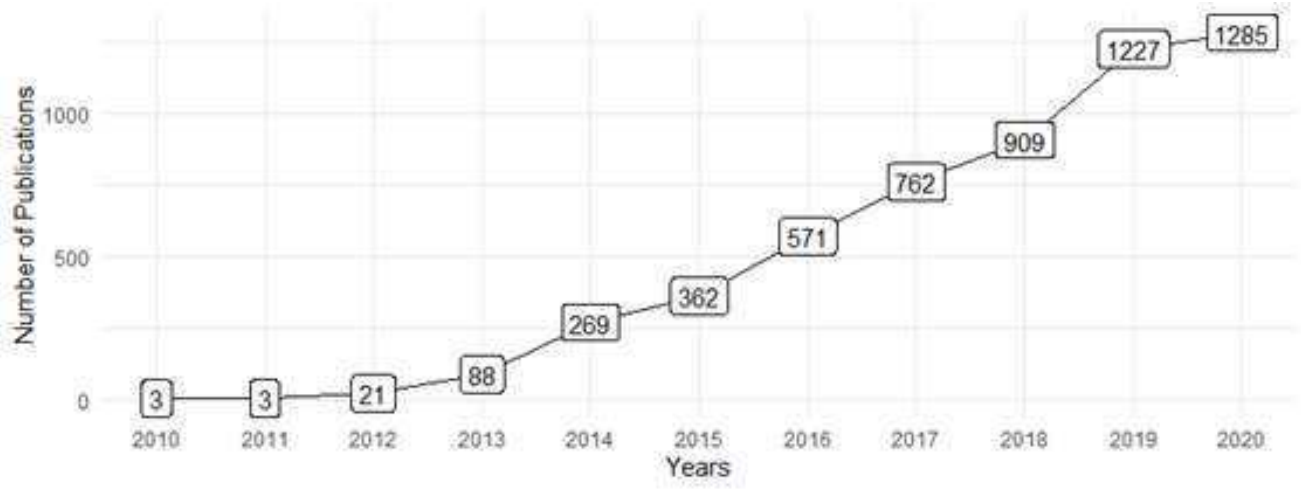


Figure 1. Evolution of scientific production in soc. sci. & humanities

CC BY

Table 2. % Yearly increase of scientific production in different corpora

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Own elaboration	–	0.00	600.00	319.05	205.68	34.57	57.73	33.45	19.29	34.98	4.73
Liu et. al., 2019	–	–	–	–	105.94	166.83	45.77	30.04	21.01	–	–
Belmonte et. al., 2020	–	0.00	1,400.00	286.67	203.45	100.00	55.40	43.51	46.75	1.22	–

RQ1. Production and authorship trends

Our results [Following Kalantari et al. (2017) we add each individual author's affiliation and country to achieve this result] shows USA on top of the ranking for scientific production regarding big data in the social sciences and humanities. This dominance has been undisputed for nearly the entire decade. China is in second place, although its production spiked up after 2018, surpassing USA in 2020. United Kingdom is in third place with a production that dates back to 2012. Australia, South Korea and Canada also show a steady increase in their production, while India and Spain have made a significant contribution in the last 2 years (Figure 2).

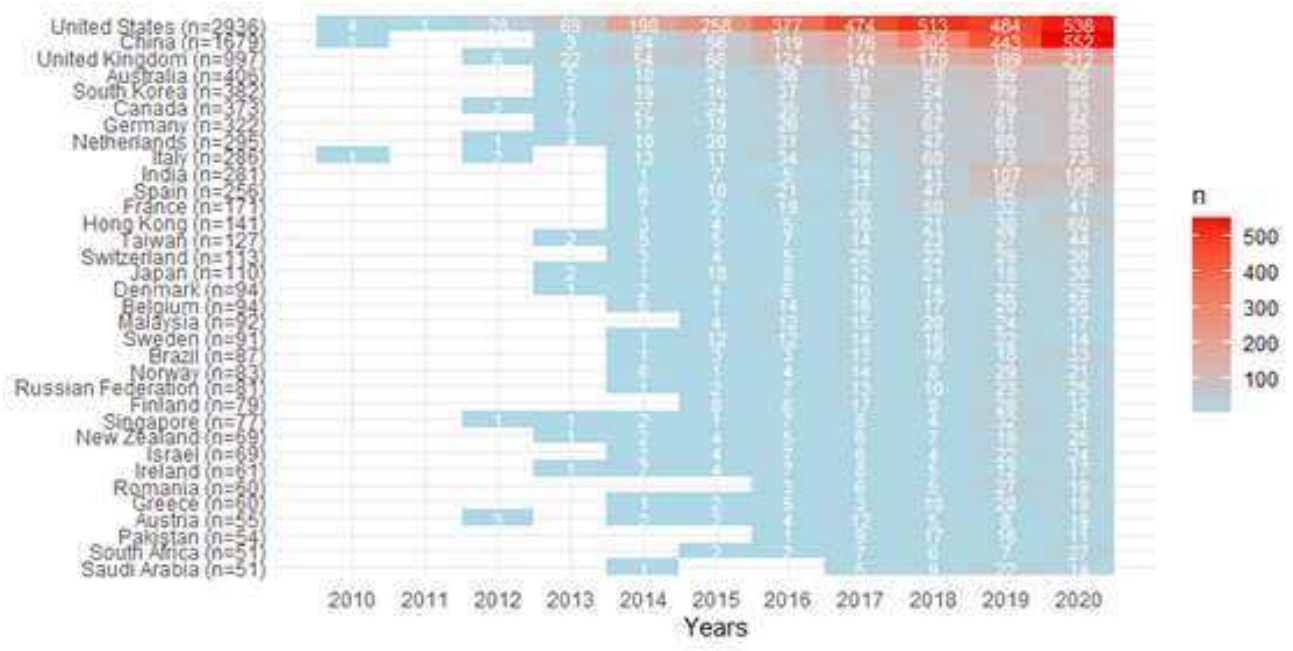


Figure 2. Production by country and year (countries with n > 50)

Regarding institutional production, the collaboration between academia and corporate institutions (Figure 3) seems to be growing slowly within social sciences and humanities. Top institutions, in terms of affiliation of authors (Figure 4), reflect the dominance of the USA, China and the UK. Chinese production by institutions (both academic and corporate) are far more concentrated than USA's.

CC BY

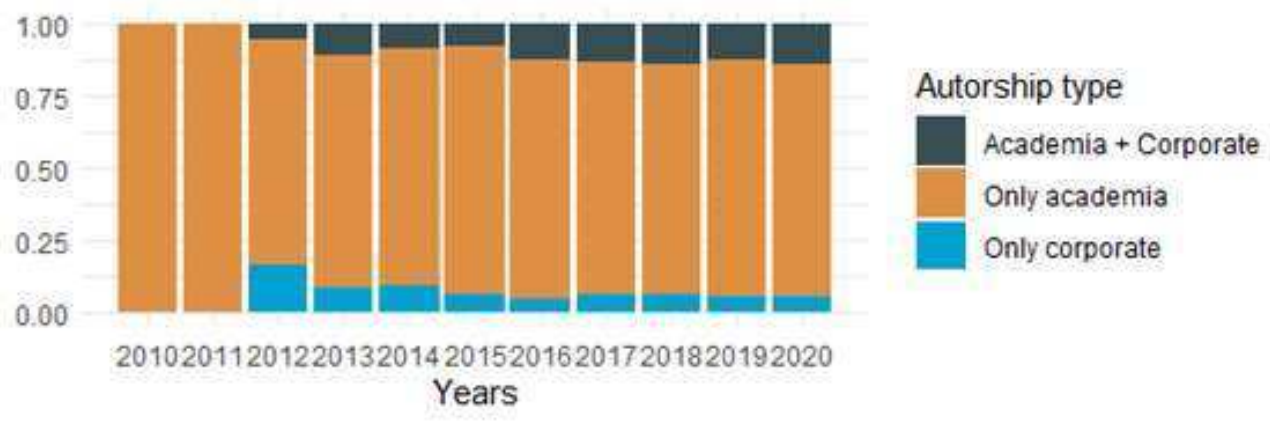


Figure 3. Distribution of authorship type

In terms of collaboration (based on co-joint authorship), we can see a hub of collaboration between USA -as the most central country, with the highest number of collaborations- China, the United Kingdom, Australia, and South Korea (Figure 4). The USA is also highly connected to Canada and to European countries like Germany, Italy and Spain, whose production has ramped up in the last years.



Figure 4. Collaboration by country (countries with n > 50)

Regarding authorship practices, we noticed 2 facts in our corpus [for the following remarks we are not considering the 6 papers from 2010 to 2011]. First, papers with only 1 author is the largest group, although it has been decreasing fast, from about 50% in 2013 to near 25% in 2018; as from 2018, there are more papers with 2 or 3 authors than with only 1. Second, there is clear tendency toward more collaborative papers: the Subramanyam Index (SI, Table 3) that shows papers with +2 authors has been spiking up since 2012, from 0.52 in 2012, and reaching 0.79 by 2020.

CC BY

Table 3. % number of authors by year

	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**
1	0.33	0.67	0.48	0.44	0.40	0.39	0.33	0.30	0.25	0.25	0.21
2	0.33	–	0.29	0.31	0.24	0.26	0.24	0.26	0.25	0.26	0.26
3	0.33	–	0.10	0.17	0.18	0.15	0.18	0.16	0.19	0.20	0.21
4	–	–	0.05	0.03	0.10	0.09	0.11	0.12	0.15	0.14	0.15
5	–	–	–	0.02	0.04	0.06	0.05	0.07	0.07	0.07	0.08
6	–	–	–	0.01	0.01	0.03	0.03	0.04	0.04	0.04	0.04
7	–	0.33	–	0.01	0.00	0.01	0.02	0.02	0.01	0.02	0.03
8	–	–	–	–	0.00	0.01	0.01	0.01	0.01	0.01	0.01
9	–	–	0.05	–	0.00	0.00	0.01	0.01	0.01	0.00	0.01
10	–	–	–	–	0.01	0.01	0.00	0.01	0.00	0.00	0.00
+10	0.00	0.00	0.05	0.00	0.01	0.01	0.01	0.01	0.01	0.00	0.01
S.I.	**0.67**	**0.33**	**0.52**	**0.56**	**0.60**	**0.61**	**0.67**	**0.70**	**0.75**	**0.75**	**0.79**

S.I.: Subramanyam Index

RQ2. Topics and areas

To explore topics and sub-areas of big data research, we first focused on keywords and their correlations. We built enriched association networks that show the top 25 correlated keywords to "big data," and the 100 correlated thereof (Figure 5). We can identify 5 clusters:

1. the first cluster links big data to research and experiments with humans, with keywords that let us infer the challenges for research design and procedures that could prove risky;
2. a second cluster, closely related to the previous group, renders big data as a phenomenon related to ethics, surveillance and data privacy;
3. a third cluster links big data to more general and data-related notions, like science, analytics, algorithms, and a few notions related to the smart city project;
4. the fourth cluster links big data to machine learning and artificial intelligence, with learning systems in the center of this hub;
5. the final cluster is related to technologies that enable the handling of big volume of data, such as cloud computing and Hadoop.

CC BY



Figure 5. Keyword correlation to big data, in 1st degree (top25) and 2nd degree (top100)

Although keyword frequency and correlation are a widely used metrics to analyze topics, we can also use abstracts through a text-classification task. For this purpose, we modeled latent topics and generated 25 non-exclusive categories (documents can belong to a mix of categories in different proportions), which were later re-grouped in 5 general categories (see supplementary files for topic model table) [Not all inferred topics are interpreted: the last 3 topics had a mix of subjects with no clear grouping]:

1. *Methodology, mix methods and techniques*: includes articles that deal with methodological innovations in big (social) data research, such as, the integration of qualitative analysis and text-mining techniques, the use of modeling and machine learning for different purposes, and the technical and infrastructure requirements. Good examples that rank highest in this topic are papers published in the Quality and Quantity Journal, such as Davidson et al. (2019), or in the International Journal of Qualitative Methods, such as Brower et al. (2019), both dealing with the irruption of "big qual."
2. *Philosophy, epistemology, ethics, theory*: comprises articles that deal with challenges to science, both from a epistemological or philosophical point of view (e.g., critiques to the claims of dataism, the history of big data), the ethics of research with big data (e.g., discussing personal data rights), and also some papers that survey and analyze these changes in social sciences' practices. Good examples of the former are those that deal with the invisibilization of women as a result of applying gender-biased datasets in machine learning scenarios, or those that focus on minorities that have long been analyzed through qualitative and small data, such as Gieseking (2018), or the very interesting work by Hill et al. (2016) that proposes that cultural criteria for assessing visualizations of big data are gender-stereotyped. We've cited a few of the latter already in section #2.
3. *Policy, politics, smart city*: papers that rank on top of these categories include discussions about data policy, the role of governments, and of human-computer relations; we've also included in this group the topic of urban design and smart city. Examples of these are

CC BY

Mahrenbach et al. (2018) that analyze the treatment of big data by Southern (Brazil, India and China) political actors, or the survey papers in Cities Journal, such as Lim et al. (2018).

4. *Business and industry*: these articles deal with management in different sectors that have been impacted by the 4.0 revolution, spanning from customer relations to transport or supply chain. A few examples of these are studies that analyze big data adoption strategies and the challenge to transform data insights into value, such as Medeiros et al. (2020)
5. *Social issues*: finally, we've grouped a few specific topics that refer to social issues, such as, ecological sustainability, climate change, or educations.

RQ3. Citation

According to the data provided by Scopus, near 26% articles have not received any citations (these include almost 1,000 papers from 2019-2020).

The most cited article in our corpus were published in 2012-2015. These are excellent works that introduced big data in a critical light, and assessed its impact on social sciences, while setting (or surveying) the agenda for research (see supplementary files). The most cited papers is boyd & Crawford (2012), wherein big data is presented as a technological, scholarly and cultural phenomenon (with its own mythology and beliefs), and that warns about both the utopian and dystopian rhetoric that it triggers. Then, authors go on to debate some epistemological claims about the "promise" of big data, asserting that "big Data reframes key questions about the constitution of knowledge," that "claims to objectivity and accuracy are misleading," and even shed warning about its "premises" by reminding that "limited access to Big Data creates new digital divides." Such line of criticism is also present at Kitchin (2014) and Dijck (2014) that discuss the "new forms of empiricism" or the "ideology of dataism" that lies behind some of big data claims. Gandomi & Haider (2015) warns that analytical methods created for structured data need to be revised in the big data realm. Ostrom et al. (2015) surveyed how service research is transforming because of big data. Colleoni et al. (2014) dwell on big social data and machine learning techniques to analyze the dynamics of politics in Twitter, and end up raising concerns about research that treated social-networks as a virtual scenario, enclosed and separate from the social practices. The "smart city" is also a big topic discussed in Kitchin (2014), Batty (2013), and Hashem et al. (2016). These articles both detail how big data is changing the management of the urban space, and the promises and risks involved. Surveillance is one of these risks, one that Zuboff (2015) render as the core component of a new logic of capitalist accumulation (Table 4).

If we look at (outbound) references, we can observer a lot of repetition with those mentioned above. Again, boyd & Crawford (2012), Kitchin (2014) -and also his book "The data revolution"- and Gandomi & Haider (2015) are within the most influential and discussed works. However, we can also see references to books and pieces outside our corpus, like Mayer-Schonberger & Cukier (2013) and Anderson (2008), which have been criticized by some social sciences papers as being promoters of the first ideas on big data, including the epistemological claims criticized above. There are also a excellent papers that draw on, or discuss, big data driven research, such as Lazer et al. (2014). Overall, these references show a particular framing for big data focusing on its importance for the social sciences. It offers a critical discussion about its meaning and foundations, and it looks for an integration between data-driven analysis and a rich theoretical awareness of such disciplines.

CC BY

Table 4. Top referenced works

Reference	n
Boyd, D, Crawford, K – 2012 – Critical Questions For Bi…	407
Mayerschonberger, V, Cukier, K – 2013 – Big Data A Revolution Tha…	298
Kitchin, R – 2014 – Big Data New Epistemologi…	118
Laney, D – 2001 – D Data Management Control…	93
Kitchin, R – 2014 – The Data Revolution Big D…	79
Lazer, D, Kennedy, R, King, G, Vespignani, A – 2014 – The Parable Of Google Flu…	76
Pasquale, F – 2015 – The Black Box Society The…	64
Gandomi, A, Haider, M – 2015 – Beyond The Hype Big Data …	50
Ginsberg, J, Mohebbi, Mh, Patel, Rs, Brammer, L, Smolinski, Ms, Brilliant, L – 2009 – Detecting Influenza Epide…	49
Anderson, C – 2008 – The End Of Theory The Dat…	48

Yet, if we look at the top journals cited in our corpus, this framing is not that clear. Top transdisciplinary journals like Science, Nature or Plos One are within the most cited, even across several of the topics we constructed [It should be noted that highly influential papers by Lazer (Lazer et al., 2014, 2020) have been published by Science]. Then, particular social science and social issues journals, such as Big data and Society, or Scientometrics, or Communication & Society rank among the most cited within specific groups of articles. Yet, the only journal that is clearly focused on engineering or informatics is IEEE. However, as reported in the methodology section, parsing this information proved to be difficult and these results should be taken with caution.

CONCLUSION

In this section we summarize our results and provide a comparison with the results reported by surveys on big data literature that are not limited to the disciplines we focused.

According to the revised surveys, big data literature has grown over a X2 rate yearly in the 2010-2014 period, then this trend slowed down in 2015, and the number of papers per year has never decreased by 2020. Our dataset, focused on social sciences, psychology and humanities, follows such general trend.

CC BY

RQ1. Regarding production, we've noted the predominance of USA's and China's production. This also resembles the data reported by general surveys, although in an inverse order for these 2 countries. It would seems like Chinese production on big data is much more focused on technical aspects than on its social significance. The rest of the countries (and also institutions) that rank on top coincide in our dataset and in the reports of the general big data field. The relative absence of works from South America, Middle East, and Africa suggest that there is a big data divide (McCarthy, 2016) in the publication arena too, probably because of costs and unequal access to data.

Collaboration between academia and corporate provided near 12.5% of our corpus, a value much higher than the 2% reported by Liu et al. (2019) for the general big data field. Data scientist are key players in the big data research, and their presence and professional demand are reshaping the profile of social scientists. Collaboration between these two players may help close the gap in skill and access to data, big elements in the big data divide. Finally, we've reported a much larger presence of single-authorship articles than in the other surveys, for which this practice was the larger group only in the early years of publication. However, collaborative authorship is clearly growing in big data research in the social sciences and is likely to be the norm in the future.

RQ2. As expected, the framing of big data in publications from social sciences, psychology and humanities is very different than in the general big data literature. By counting and correlating keywords we identified 5 topics: (k1) ethical challenges for research, (k2) surveillance, (k3) data handling, (k4) modeling and machine learning, (k5) big data technology. By classifying abstract, through topic modeling, we identified 5 topics: (tm1) mix methodology, (tm2) philosophy and theory of big data, (tm3) policy and smart cities, (tm4) business and industry applications, (tm5) social issues. Drawing on the classifications provided by the surveys in the general big data field, we identified 4 recurring topics: (g1) data handling and big data technology, (g2) business intelligence, (g3) machine learning and analysis, (g4) novel sources of data (including medical and big social data, and thus ethics concerns).

It is interesting to compare the topics identified by the general big data literature and in our corpus, in order to identify where are potential links for multi-disciplinary collaboration, and where the particular focus of the social sciences and humanities is set. Regarding coincidences, there is a shared interest in data handling, data analysis and machine learning, but also in exploring the possible business and industry applications of big data. Ethical considerations, which in the general literature is mostly connected to the medical field, is also present. These interests could be considered the "common ground" for all big data sub-fields and disciplines. In social sciences and humanities, these interests are re-framed in the discussions on mix-methodology and in particular social issues. Regarding the differences, social sciences and humanities drive the considerations on the critical view of big data, which spans across subjects that exceed the ethical considerations -which are present in most fields- to include the the social nature of (the construction of) datasets, the political and ideological uses of big data, and the discussion of its philosophical and epistemological foundations.

RQ3. We used citations to explore highly influential works and top venues for publication. Regarding top cited works in our corpus, we identified important works that presented a critical research agenda. Out of the top 10 cited papers, 5 were classified as "Policy, politics, smart city," 3 as "Philosophy, epistemology, ethics, theory," and 2 as "Methodology, mix methods and techniques." Interestingly, no papers classified as "Business and industry" or "Social issues" -topics that we have seen in other surveys from the general big data literature- made this rank, which could be indicative of the auto-referential dynamic of the big data research in social science and humanities fields. Yet,

CC BY

some of these papers are mentioned between the most influential in the general big data literature. Regarding the top venues (in terms of outbound citation count), the predominance of computing and engineering papers reported by the surveys of the general big data literature was not seen in our corpus, in which prevailed top transdisciplinar journals like Science or Nature, among social science focused journals, such as Big data and Society, or Scientometrics, or Communication & Society.

DISCUSSION

Publication trends on big data in social sciences, psychology, and humanities and arts show an evolving research field. These critical sciences have been warning about the risks of big (social) social data research -e.g., data collection through Facebook apps and active campaigns-, pinpointing issues on privacy and social reactivity in times when research/experimentation seem to be fading and regulatory definitions demand revisions (Leonelli, 2016; Metcalf et al., 2016; Metcalf & Crawford, 2016; Weinhardt, 2020); also they have been raising awareness that big data is embedded with limitations and conditioning from data gathering/creation settings, and that also the data curation process required for algorithmical analysis is a heavy decision task that challenges the neutrality and objectivity of some data science claims (Gitelman, 2013; Mützel, 2015). Lastly, the social sciences, psychology and humanities have been discussing methodological integrations for mixed social research in a way that could allow guiding data-driven analysis through the rich theoretical awareness of these disciplines (Burrows & Savage, 2014; Halford & Savage, 2017). In Halavais (2015) words regarding sociology:

> "Big data does provide a challenge to the social sciences, but not a particularly new one. It is, in fact, the core challenge of sociology: connecting the micro-connections between individuals to the vast social structures that shape us (and are shaped by us) as a society. Mills famously suggested that this ability to both connect and disconnect the personal with the social was at the core of what he called the 'sociological imagination.'"

In this study we have advanced one step in a larger project aiming at comparing how big data is portrayed, framed and treated in differentiated areas of social communication. Based on a corpus of 5,500 Scopus articles published between 2010-2020, and comparing the results reported on several big data bibliometric studies -without limiting to social sciences, psychology, or humanities and arts-, we have been able to show how these highly intertwined fields share some trends with big data literature, while also highlighting its particular features. We have not only drawn on bibliometric variables but also have included other descriptive statistical techniques like unsupervised document classification that allowed us to differentiate further analysis. To the best of our knowledge, this is the first study to provide a comprehensive view on big data research focusing on the social sciences, psychology and humanities, and to provide a comparative view with global production across disciplines.

Nevertheless, this work presents several limitations. First, as any case study, our analysis is dependent on the selected database. Other databases -e.g., one with more research in other languages than English, or with more open-access and no-processing-charges- could improve the presence of underrepresented production contexts. Second, limiting the search criteria to "big data" is a very debatable decision to gather a research field. Big data is not only an academic concept but a designation introduced by the IT industry which rapidly converted into a business buzzword. In

CC BY

social communications, big data comes along with other phenomena, like artificial intelligence of data science, and neologisms like "big quals." Third, parsing different parts of dataset -such as journal names in references- proved to be a challenging task; while other analysis and visualizations -like enriched associative networks- are highly dependent on the decisions we made (and reported). All of these limitations could be surpassed in future investigations. Also, in-depth qualitative analysis and other systematic reviews are required to deepen our understanding of how social sciences and humanities create a particular contextual and framing account of big data, something signaled by literature as a way to create a conceptual history of big data.

REFERENCES

Ahmad, I., Ahmed, G., Shah, S. A. A., & Ahmed, E. (2020). A decade of big data literature: analysis of trends in light of bibliometrics. *The Journal of Supercomputing*, *76*(5), 3555–3571. https://doi.org/10.1007/s11227-018-2714-x

Anderson, C. (2008). The end of theory. The data deluge makes the scientific method obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Batty, M. (2013). Big data, smart cities and city planning. *Dialogues in Human Geography*, *3*(3), 274–279. https://doi.org/10.1177/2043820613513390

Becerra, G. (2018). Interpelaciones entre el Big data y la Teoría de los sistemas sociales. Propuestas para un programa de investigación. *Hipertextos*, *6*(9), 41–62. http://revistahipertextos.org/ediciones/hipertextos-no-9/

Beer, D. (2016). How should we do the history of Big Data? *Big Data & Society*, *3*(1), 205395171664613. https://doi.org/10.1177/2053951716646135

Belmonte, J. L., Segura-Robles, A., Moreno-Guerrero, A. J., & Parra-González, M. E. (2020). Machine learning and big data in the impact literature. A bibliometric review with scientific mapping in web of science. *Symmetry*, *12*(4). https://doi.org/10.3390/SYM12040495

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679.

Brower, R. L., Jones, T. B., Osborne-Lampkin, L., Hu, S., & Park-Gaghan, T. J. (2019). Big Qual: Defining and Debating Qualitative Inquiry for Large Data Sets. *International Journal of Qualitative Methods*, *18*, 1–10. https://doi.org/10.1177/1609406919880692

Burrows, R., & Savage, M. (2014). After the crisis? Big Data and the methodological challenges of empirical sociology. *Big Data & Society*, *1*(1), 205395171454028. https://doi.org/10.1177/2053951714540280

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, *64*(2), 317–332. https://doi.org/10.1111/jcom.12084

Davidson, E., Edwards, R., Jamieson, L., & Weller, S. (2019). Big data, qualitative style: a breadth-and-depth method for working with large amounts of secondary qualitative data. *Quality and Quantity*, *53*(1), 363–376. https://doi.org/10.1007/s11135-018-0757-y

Dijck, J. van. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance and Society*, *12*(2), 197–208.

CC BY

Ebrahim, N. A., Salehi, H., Embi, M. A., Tanha, F. H., Gholizadeh, H., & Motahar, S. M. (2014). Visibility and citation impact. *International Education Studies*, *7*(4), 120–125. https://doi.org/10.5539/ies.v7n4p120

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gieseking, J. J. (2018). Size Matters to Lesbians, Too: Queer Feminist Interventions into the Scale of Big Data. *Professional Geographer*, *70*(1), 150–156. https://doi.org/10.1080/00330124.2017.1326084

Gitelman, L. (2013). *"Raw Data" Is an Oxymoron*. The MIT Press. https://doi.org/10.1080/1369118X.2014.920042

Halavais, A. (2015). Bigger sociological imaginations: framing big social data theory and methods. *Information Communication and Society*, *18*(5), 583–594. https://doi.org/10.1080/1369118X.2015.1008543

Halford, S., & Savage, M. (2017). Speaking Sociologically with Big Data: Symphonic Social Science and the Future for Big Data Research. *Sociology*, *51*(6), 1132–1148. https://doi.org/10.1177/0038038517698639

Hashem, I. A. T., Chang, V., Anuar, N. B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E., & Chiroma, H. (2016). The role of big data in smart city. *International Journal of Information Management*, *36*(5), 748–758. https://doi.org/10.1016/j.ijinfomgt.2016.05.002

Hill, R. L., Kennedy, H., & Gerrard, Y. (2016). Visualizing Junk:Big Data Visualizations and the Need for Feminist Data Studies. *Journal of Communication Inquiry*, *40*(4), 331–350. https://doi.org/10.1177/0196859916666041

Kalantari, A., Kamsin, A., Kamaruddin, H. S., Ale Ebrahim, N., Gani, A., Ebrahimi, A., & Shamshirband, S. (2017). A bibliometric approach to tracking big data research trends. *Journal of Big Data*, *4*(1), 1–18. https://doi.org/10.1186/s40537-017-0088-1

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, *1*(1). http://journals.sagepub.com/doi/10.1177/2053951714528481

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety* (No. 2001; Vol. 949). META Group. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Fly: Traps in Big Data Analysis. *Science*, *343*(March), 1203–1205. http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

Lazer, D., Pentland, A., Watts, D. J., Aral, S., Contractor, N., Freelon, D., Gonzalez-bailon, S., & King, G. (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 13–16. https://doi.org/10.1126/science.aaz8170

Leonelli, S. (2016). Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2083). https://doi.org/10.1098/rsta.2016.0122

Liang, T. P., & Liu, Y. H. (2018). Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study. *Expert Systems with Applications*, *111*(128), 2–10. https://doi.org/10.1016/j.eswa.2018.05.018

Lim, C., Kim, K. J., & Maglio, P. P. (2018). Smart cities with big data: Reference models, challenges, and considerations. *Cities*, *82*(April), 86–99. https://doi.org/10.1016/j.cities.2018.04.011

Liu, X., Sun, R., Wang, S., & Wu, Y. J. (2019). The research landscape of big data: a bibliometric analysis. *Library Hi Tech*, *38*(2), 367–384. https://doi.org/10.1108/LHT-01-2019-0024

Mahrenbach, L. C., Mayer, K., & Pfeffer, J. (2018). Policy visions of big data: views from the Global South. *Third World Quarterly*, *39*(10), 1861–1882. https://doi.org/10.1080/01436597.2018.1509700

Mayer-Schonberger, V., & Cukier, K. (2013). *Big data. A revolution that whill transform how we live, work, and think*. Eamon Dolan/Houghton Mifflin Harcourt.

McCarthy, M. T. (2016). The big data divide and its consequences. *Sociology Compass*, *10*(12), 1131–1140. https://doi.org/10.1111/soc4.12436

Medeiros, M. M. de, Maçada, A. C. G., & Freitas Junior, J. C. da S. (2020). The effect of data strategy on competitive advantage. *Bottom Line*, *33*(2), 201–216. https://doi.org/10.1108/BL-12-2019-0131

Metcalf, J., & Crawford, K. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *Big Data and Society*, *3*(1), 1–14. https://doi.org/10.1177/2053951716650211

Metcalf, J., Keller, E. F., & Boyd, D. (2016). *Perspectives on Big Data, Ethics, and Society*. https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/

Mützel, S. (2015). Facing Big Data: Making sociology relevant. *Big Data & Society*, *2*(2), 205395171559917. https://doi.org/10.1177/2053951715599179

Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *74*(3), 1–19. https://doi.org/10.1103/PhysRevE.74.036104

Ostrom, A. L., Parasuraman, A., Bowen, D. E., Patrício, L., & Voss, C. A. (2015). Service Research Priorities in a Rapidly Changing Context. *Journal of Service Research*, *18*(2), 127–159. https://doi.org/10.1177/1094670515576315

Ponomariov, B., & Boardman, C. (2016). What is co-authorship? *Scientometrics*, *109*(3), 1939–1963. https://doi.org/10.1007/s11192-016-2127-7

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, *91*. https://doi.org/10.18637/jss.v091.i02

Sadin, É. (2018). *La inteligencia artificial o el desafío del siglo. Anatomía de un antihumanismo radical*. Caja Negra.

Subramanyam, K. (1983). Bibliometric studies of research collaboration: A review. *Journal of Information Science*, *6*(1), 33–38. https://doi.org/10.1177/016555158300600105

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Team R Core. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Weinhardt, M. (2020). Ethical issues in the use of big data for social research. *Historical Social Research*, *45*, 342–368. https://doi.org/10.12759/hsr.45.2020.3.342-368

Xindong Wu, Xingquan Zhu, Gong-Qing Wu, & Wei Ding. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Zhang, Y., Huang, Y., Porter, A. L., Zhang, G., & Lu, J. (2019). Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. *Technological Forecasting and Social Change*, *146*(April), 795–807. https://doi.org/10.1016/j.techfore.2018.06.007

Zuboff, S. (2015). Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75–89. https://doi.org/10.1057/jit.2015.5

CC BY