Scientometrics (2020) 122:1563–1581
https://doi.org/10.1007/s11192-020-03371-2

The evolution of data science and big data research: A bibliometric analysis

Daphne R. Raban[1] · Avishag Gordon[2]

Received: 8 August 2019 / Published online: 31 January 2020

Abstract

In this study the evolution of Big Data (BD) and Data Science (DS) literatures and the relationship between the two are analyzed by bibliometric indicators that help establish the course taken by publications on these research areas before and after forming concepts. We observe a surge in BD publications along a gradual increase in DS publications. Interestingly, a new publications course emerges combining the BD and DS concepts. We evaluate the three literature streams using various bibliometric indicators including research areas and their origin, central journals, the countries producing and funding research and startup organizations, citation dynamics, dispersion and author commitment. We find that BD and DS have differing academic origin and different leading publications. Of the two terms, BD is more salient, possibly catalyzed by the strong acceptance of the pre-coordinated term by the research community, intensive citation activity, and also, we observe, by generous funding from Chinese sources. Overall, DS literature serves as a theory-base for BD publications.

Keywords Big Data · Data Science · Evolution · Relationship · Bibliometric analysis

Introduction

Science research keeps expanding over the years and "new specialisms arise from old areas all the time" (Meadows 1998). The normal interdisciplinary trends of disciplines' creation in the past occurred when a new unifying concept brought together a wide range of knowledge (Ibid.:44). Meadows (1998) brings cybernetics as an example of a field arising from aggregation of a wide range of social science and engineering ideas. Furthermore, there are examples of fields that were split to several subfields before becoming a unified whole such as the case of terrorism studies that were dispersed before this research area became

✉ Daphne R. Raban
draban@univ.haifa.ac.il

Avishag Gordon
avishag@g.technion.ac.il

[1] School of Business Administration, University of Haifa, Haifa, Israel

[2] Information and Knowledge Management and Information and Library Studies, University of Haifa, Haifa, Israel

a cohesive one (Gordon 2004). Another evolutionary change may occur when a field that existed for many years was absorbed by a larger research field because of lack of researchers' commitment to the field (Creager 2010; Mullins 1972). Overall, disciplines exhibit dynamics from fragmentation to unification as time and necessities dictate (Balietti et al. 2015). Glänzel and Thijs (2012:196) set several criteria for detecting an emerging research area, including: the existence of a critical mass of publications to form a coherent cluster, emerging topic identification, and cognitive description of the new topic by analyzing articles' titles and/or keywords.

Observing recent development of new areas of research such as data science (DS) and big data (BD), we noticed that these study areas accumulated enough publications for a cognitive description, raising questions regarding their anticipated dynamics: will they develop in parallel into two distinct research fields or are they going to merge? What is the evolutionary tendency of these fields? The following literature review summarizes some of the devleopments already observed in these dynamic research fields, and sets the stage for the current research.

Literature review

We begin by providing working definition of BD and DS. From a business perspective, Laney (2001) referred to BD as data characterized by three V's: volume, velocity, and variety. The volume element of the definition refers to the massive amounts of data collected on people's actions and choices especially by the use of online applications; velocity refers to the speed of data generation; variety relates to the heterogeneity of the data. This definition was quoted in Sætra's paper (2018) who analyzed BD in terms of what this approach to science can or cannot do, especially in the area of psychological behaviorism.

A 2013 definition of DS reads: "At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" (Provost and Fawcett 2013: 52). This definition emphasizes the close relation of DS to data mining. BD definition by the same authors is as follows: "we will simply take big data to mean datasets that are too large for traditional data-processing systems and that therefore require new technologies" (Ibid.: 54). A recent publication indicates that the early definitions of these two concepts continue to be acceptable and, therefore, can be used in our study (Sanchez-Pinto et al. 2018).

The purpose of this study is to follow the evolution of the two research areas, DS and BD, using bibliometric measures, that is, by observing publications and citations in these areas over time, trying to determine where these fields are heading.

A study by Singh et al. (2015) mapped the area of BD, and noted the interdisciplinary nature of this field, and the growth rate in number of publications, authors, disciplines and countries involved in its development. A more recent study investigated the interdisciplinary collaboration in BD research and discovered that the main contributors to the BD research are Computer Science, Engineering, and Business and Economics, and that research communities on the subject are formed (Hu and Zhang 2017). Another recent study covering seven years of BD publications showed the strong dynamics of journal and even more so, of conference publications in the field, attributing much of the activity to the broad interest in BD from various research fields, as well as to the strong interest of large and powerful countries such as China and the U.S. (Gupta et al. 2019). A comment in Nature explained the complexity of BD, the need for new algorithms to aggregate the deluge of data, and the future contribution of BD to various areas of research (Mattmann 2013). This latter study does not differentiate between DS and BD, since the title says: “A vision for data science” but the content elaborates on big data.

Data Science as a concept preceded Big Data chronologically. In the 1960s only one paper that discussed BD but 52 papers dealt with DS, based on a search of Clarivate Analytics, which is part of the Web of Science database. Nevertheless, the DS concept mostly referred to the progression of social science and behavioral data and its uses, that is, data collation in the social sciences, and not in a sense of extracting knowledge from data as referred to this area today (Clarke 1975).

The term DS appeared in the 1960s in relation to social sciences data, or in the context of computer use and technology but not as a concept. The words “data” and “science” did not appear conjointly in the titles of scientific papers and were separated by several words. When keywords are not conjoint, they do not constitute a concept, or a name. An example of DS as a subject that is not yet formed into a field is in the following title: “Worldwide problem of numerical data for science and technology” (Rossini 1967). This was how papers focusing on data science were using the terms from the 1960s to 2001 when the titles of scientific papers started using the concept Data Science (Cleveland 2001).

The term Big Data appeared once as a concept in 1974 and then again in editorials in 2006 and 2007, and only in 2008 the use of “Big Data” as a concept started appearing regularly in scientific papers, but the actual implementation of the BD concept started in 2010 (Mervis 2012).

The interchangeable use of the BD and DS in the titles of publications was noted in Aronova et al. (2010) who used the concept “Big Science” that related to the 1960s Manhattan project and the American national space program.

Research question

What is the evolvement course shown by the literature on DS and BD? Are these fields merging, or are they developing in parallel routes that are not intended to meet?

Data

The data in this study was drawn from the database Clarivate Analytics (also known as the WoS, Web of Science) 2019 core collection. This is a selective index of good quality publications. The search was conducted on titles and abstracts of scientific peer-reviewed publications (N=41,961 for BD, N=244,695 for DS, N=3,552 for interchangeable use). Publications containing BD and DS as pre-coordinated concepts were retrieved from 2006 to March 2019, including publications that use these terms interchangeably (N = 7938 for BD, N = 2648 for DS, N=242 for interchangeable use).

The search for BD and DS was limited to the title field to enable the study of the core fields themselves rather than their applications in various scientific efforts. We assume that searching the title generates a reasonable, possibly representative sample, of the field of interest. We base this assumption on a long tradition of research on the nature and usefulness of article titles (Rons 2018).

After choosing the title as a search field, the search was narrowed even more by inserting quotation marks around the words DS and BD to extract the publications that relate to these two terms as concepts (N = 7299 for BD, N = 420 for DS, N = 67 for interchangeable use). The assumption behind this examination is that the more a term appears as a concept in a retrieval set, the closer we are to define the field by this concept. When referring to "publications" we mean all types of publications indexed by Clarivate Analytics: articles, reviews, whole books, book reviews, editorial items etc.

Methodology

The retrieved set of publications was analyzed to discover overall productivity, current research areas and their origin, central journals and citation patterns, the countries producing and funding research and startup organizations (Hartmann et al. 2016).

The dynamics of BD and DS over time was examined by bibliometric indicators including "highly cited" papers and the immediacy index. Highly cited papers are those that received a high number of citations, usually within the range of 10 recent years or less, depending on the discipline. The highly cited papers indicator was devised to bypass the high number of citations accumulated during a very long publications' history of researchers. Immediacy index is calculated by dividing citations by publications within the year of publication. The immediacy index indicates, to a large extent, the journal impact (Tomer 1986; Yue et al. 2004), and is also considered to be an indication of the "research front" of a science field (Meadows 1998: 61).

The immediacy index was complemented by the examination of the Price Index which measures the citations to publications in the last five years as compared to the total number of citations per topic, and examines the aging of the literature. "Ageing patterns can be characterized as a combination of phases of maturation and decline in citation processes" (Glänzel et al. 2016: 2169).

The three indicators were used to reveal which concept, BD or DS, is more in use. Intensive usage in a field indicates a dynamic and promising science field. The Price Index was measured in two years: in 2010 when all three literatures already existed, and in 2018, more recently, to observe the dynamic of the three trends.

Dispersion in the fields of BD and DS was calculated by comparing the number of publications yielded by searches by topic to the same searches by title. A high percent of dispersion indicates a field with a small cohesive literature core (Tal and Gordon 2017).

Another test was that of commitment of authors to the research field which is an indication of regularity and constancy by authors who are not "one- time visitors" to the field. Such authors could help in creating theories and paradigms in the research area and maintain continuity in the research field (González-Alcaide et al. 2016; Gordon 2007).

Results

Figure 1 shows that the number of publications on BD jumped from 56 during 2000-2009 to 7,603 in the following decade, 2010 to 2019. The growth rate of the number of publications on DS was more gradual, but also showed an exceptional growth during 2010-2019. While DS was more significant in early years for about five decades, BD made a leap during the recent decade. Interestingly, these results are in agreement with a previous study that retrieved results from Scopus for the years 2010-16 (Gupta et al. 2019).

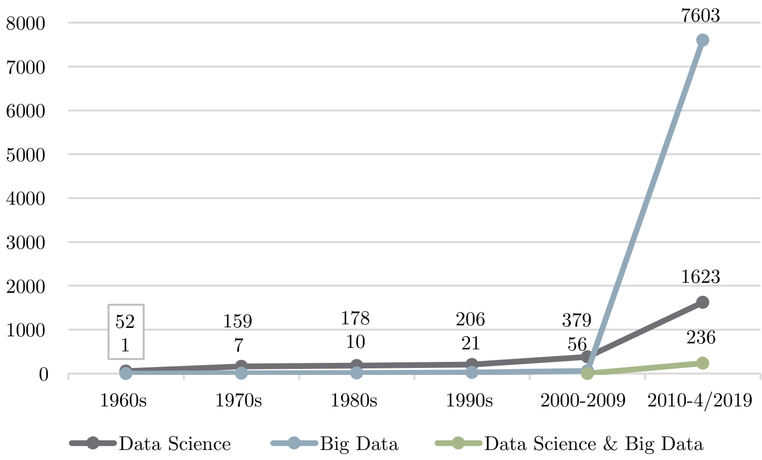

Fig. 1 Evolutionary trend in the number of publications covering data science and big data

Research areas related to DS and BD

Ranking the most covered research areas of BD and DS based on WoS Categories, displayed in Table 1, reveals differences and similarities in the coverage of these concepts, and the proximity or distance between them.

Table 1 shows that BD focuses on computer science, management, medical sciences, and engineering, while DS tends to be more disciplinary dispersed with some focus on computer science and environmental sciences. Similar dispersion is observed for publications using BD and DS interchangeably with some focus on computer science and medical sciences. The BD literature does not rely on inter and multidisciplinary areas for coverage as much as is the case with DS and the papers that are common to both concepts (Table 2).

Next, we examine the appearance of the search terms as pre-coordinated concepts. This is an indication of the formation of a terminological convention for the relevant research

Table 1 Research areas most covered by BD and DS 1965-2019. *Source*: Clarivate analytics, 2019

Research areas (WoS categories)	BD Rank and No. publications	DS Rank and No. Publications	Interchangeable BD and DS Rank and No. Publications
Computer science (and subfields)	**1** (2,529)	**2** (422)	**2** (52)
Management	**2** (1,450)	**8** (94)	**4** (20)
Medical sciences	**3** (1,263)	**4** (225)	**3** (48)
Engineering	**4** (620)	**6** (164)	**5** (20)
Telecommunication	**5** (609)	**10** (17)	N/A
Multi and interdisciplinary sciences	**6** (597)	**1** (463)	**1** (55)
Technology	**7** (480)	**7** (110)	**8** (4)
Environmental sciences	**8** (474)	**3** (324)	**6** (17)
Info and library Science	**9** (314)	**5** (251)	**7** (12)
Mathematics	**10** (249)	**9** (71)	**9** (3)

Table 2 The origin of coverage of BD and DS literatures in the 1960s. *Source*: Clarivate Analytics

Search terms	Year of first papers	No. of papers	Research areas covered
BD	1969	2	Mechanical Engineering Multidisciplinary science
DS	1965	6	Astronomy Astrophysics Physical chemistry Economics Business History and philosophy of sciences Management

community. High usage of the concept indicates that the community is cohesive and focused on a shared area of interest.

Table 3 shows that most BD publications appeared after the term became a concept. For the DS literature the trend is different: only 16% of these publications actually referred to this subject as a concept, and the rest appeared with the words "data" and "science" separated by several words, that is, not as a pre-coordinated concept. The interchangeable use literature shows a slow pace in terms of quantity of publications. Table 3 shows that computer science and its subfields are dominant in covering both BD and DS, but the number of BD publications surpasses that of DS and of the interchangeable use publications. The first part of the millennium showed a complete change in the DS coverage by the massive entrance of computer science papers to this field. This was more marked from 2010 to 2019 as indicated by continued coverage of DS in computer science publications, and strong growth in the number of DS papers. A large jump in the number of BD papers was noted too, as seen in Figure 1.

It is safe to establish that in this millennium BD and DS literatures became entangled. This is more noted in the appearance in 2010 of papers that use both concepts interchangeably.

Journals

Journals (as well as books) are archetypical examples of formal communication (Meadows 1998:7). Throughout time, journals became the markers of research fields. Today, when a researcher considers sending an article to a journal, s/he must make sure firstly that the theme of his writings and the methodology match the journal's aims and scope. Therefore, the knowledge division between BD and DS (extracted from the WoS core collection), and the interchangeable use of both, could be observed by examining the core journals of each concept. Table 4 presents the five most productive journals for each concept:

Citations and self-citations

Table 5 shows that the use of self-citations in BD studies is considerably more extensive than that of DS or that of the mixed terminology literature. It could be argued that since the number of publications of BD is more than double that of DS, it is conceivable that the number of self-citations will also be much higher for BD, but the number of citations per year, which reduces the effect of accumulation of citations over time, does not support this argument. The accumulation period and "per year" citations are two distinct

Table 3 Publications and coverage since BD and DS concepts were formed

Area no. publications since 1960s (search by Title)	No. publications since forming a concept**	Percent of publications after forming concept (%)	Subjects most covered since forming a concept
BD (7938)	7299 (since 2006)	92	CS* Information systems (1134) CS theory and methods (724) Engineering electrical, electronic (680) Telecommunication (580) CS software Engineering (435)
DS (2648)	420 (since 1997)	16	CS Interdisciplinary applications (46) CS information systems (38) Statistics and probability (35) CS theory and methods (34) Ta Information science library science (28)
Interchangeable BD and DS (242)	67 (since 2010)	28	CS Interdisciplinary applications (10) Nursing (10) CS theory methods (8) Management (7) CS artificial intelligence (6)

* CS refers to computer science

** The numbers of publications where BD, DS and the interchangeable use appearing as concepts

Table 4 Five most productive journals in BD, DS and the combined terms 1965–2019

BD journal titles and number of papers	DS journal titles and Number of papers	Interchangeable BD and DS journal titles and number of papers
IEEE Access (160)	Abstract Papers of the American Chemical Society (58)	Big Data and Society (6)
International Journal of eScience (108)	Science (47)	Annals of the American academy of Political and Social Science (5)
Big Data (62)	Monthly Notices of the Royal Astronomical Society (21)	Big Data (5)
Big Data & Society (58)	PlosOne (20)	Current Opinion in Behavioral Sciences (5)
Cluster computing, The journal of Networks Software Tools and Applications (49)	Scientometrics (20)	Big Data Research (3)

Table 5 Citations, self-citations and citations per year of the BD, DS and both literatures

Literature	No. of years since appearing as concept	No. of citations	No. of self-citations	% of self-citations	Citations per year
BD (since 2006)	13	66,932	14,875	**22.22**	1487.4
DS (since 1997)	22	29,354	835	2.84	1334.3
Interchangeable BD and DS (since 2010)	9	3,007	62	2.06	334.11

phases. The “per year” measure enables viewing clearer the citations’ dynamic of a field independently of the start- time when the field emerged.

Research funding

Table 6 shows that BD papers are the most funded of the three literatures, although the percent of funded papers is similar to the DS funding. There is a difference in the number of funding countries, and the years the funding started. The ranking of funding countries shows that the BD literature is funded mostly by China, and the USA is ranked second in this respect. The funding trend of DS and the combined terms is more “classical”, and the USA and western countries occupy the first places among the funding countries. The newer, mixed terminology literature shows a relatively, high percent of funded papers.

The dynamics of DS and BD

Table 7 shows that although publications on DS started earlier, the impact of the BD publications is higher than that of the DS publications. The difference in the dynamics of the two concepts is shown by the larger proportion of highly cited papers of BD, and the higher immediacy index of BD papers in 2018. These results are added to the exceptional jump in the number of BD publications from 2010 to 2019.

The papers with interchangeable use of the two concepts show a higher “highly cited” percentage, which means that when the two terms, BD and DS, are discussed in the same paper, it has a good chance of becoming a highly cited paper. The Price index indicates that the most dynamic trend is that of BD, although the mixed term use trend shows a promising dynamic too.

Table 6 Funding agencies of BD and DS publications, 2003–2019

Categories	Number of countries funding	Number of papers funded	Percent of funded papers of total	Years funding started	Five leading funding countries
BD (N=7,603)	67	1572	20.67	2003	**China (902**) USA (592) UK (130) Canada (79) Australia (780)
DS (n=1,623)	45	285	17.56	2007	**USA (216)** UK (100) Germany (57) Australia (47) France (44)
Interchangeable (n=236)	27	71	30.08	2013	**USA (42)** Germany (13) UK (8) China (8) Canada (8)

Table 7 Dynamics indicators of DS and BD publications 1965–2019 (*Source*: Clarivate Analytics, 2019)

Indicators	BD (N=7,603)	DS (N=1623)	Interchangeable use (N=236)
Highly cited	180 (2.32% of total)	26 (0.99% of total)	8 (3.37% of total)
Start publications in database (1960s)	1	51	1 (Starting year 2010)
Immediacy index (2018). Citations/publications	1.537 (3,364/2188)	0.895 (291/325)	0.88 (52/59)
Price Index in 2010 (measuring 2006-2010)	1,736/2,534= 68.50%	6,860/18,437= 37.20%	NA
Price Index (measuring 2014-2018)	54,671/68,527 =79.78%	7,958/33,769 =23.56%	2,164/3,063 =70.64%

Countries

Assuming a relation between BD and DS publications and the actual use of BD and DS in various types of startup organizations in countries (Papadopoulos 2019; Simon and Leker 2016), a Spearman's rho correlation was performed for the data presented in Table 8. The 34 countries chosen for the analysis were those that have the highest number of startup organizations in 2019 (Papadopoulos 2019).

The rather strong relation between the two variables shown in Table 8 was disrupted by the difference in rank of Israel: ranked fourth in the number of startups existing in this country yet with relatively few BD publications (52), Israel is in the 33rd place. So, the relation between the two variables remains strong. The relation of countries' startups and DS was strong, Rs =0.515. However, when the search terms combined both BD and DS, the picture changed completely, and the correlation yielded Rs = -0.053. The possible explanation for these results is that the mixed publication trend of BD and DS is relatively new, since it started around 2010, and many of the countries that provided publications on BD and DS publications did not yet enter the newer combined trend. Alternatively, it may be concluded that the interchangeable use of the terms is unrelated to the startup scene and remains an academic interest only.

Subject dispersion

The subject dispersion of each category (results of search by topic minus results of search by title, divided by search by topic) demonstrates the cohesiveness of each. Table 9 shows that BD is the least dispersed category, although 81.35% dispersion rate cannot be described as cohesive. The DS is the most dispersed category of the three. A cohesiveness threshold was not established, but in this study, the lower the amount of dispersion the higher is the subject's cohesiveness.

In the course of the search in WoS databases, the analysis of the retrieval set showed a certain percent of the search results that could not be classified to any subject area. This indicator was added to the dispersion measure, as another dispersion value. The lower the number of papers that were not classified, the lower the dispersion rate and the higher the subject area's cohesion. The results show that in the BD retrieval set only 2.98% of the data were not classified, while 12.01% of the papers in the DS set were not classified, and 35.44% of the combined concepts' set were not classified to any subject area. Thus, the BD literature shows a higher cohesion than that of DS or the combined set.

Author commitment

Table 10 shows that BD authors are highly committed. 20% of BD authors wrote more than one paper on the topic, and up to 65 papers. DS authors are also committed as 22% of them wrote more than one paper and up to 27 papers. The use of both terms in one publication seems to attract less commitment as almost all authors wrote only once, and those who wrote more than once, wrote a fairly limited number of publications, up to six.

Table 8 The relation between startup organizations in 34 countries and the number of publications on BD and DS in these countries. *Sources*: CEOWORLD magazine, Jan. 02, 2019, and Clarivate Analytics, 2019)

Countries' startups' rank		Publications' on BD, rank	Publications on DS, rank	Publications on interchangeable BD and DS, rank
United States	1	1	1	1
United Kingdom	2	3	2	3
Canada	3	5	7	5
Israel	4	33	34	27
India	5	11	14	20
Germany	6	6	3	2
Poland	7	28	30	44
Malaysia	8	26	43	40
Sweden	9	16	23	46
Denmark	10	24	21	26
Switzerland	11	14	11	14
France	12	10	5	10
Singapore	13	17	39	13
Australia	14	4	4	7
China	15	2	10	4
Estonia	16	35	53	36
Ireland	17	32	28	0
Russia	18	29	17	31
South Korea	19	7	25	23
Spain	20	8	6	9
Finland	21	27	29	0
Netherlands	22	13	8	8
Japan	23	12	12	15
Lithuania	24	54	81	0
Austria	25	31	22	16
Portugal	26	30	32	29
Italy	27	9	9	6
Czech Republic	28	49	37	35
Belgium	29	21	16	17
Romania	30	40	47	30
United Arab Emirates	31	41	62	32
Greece	32	23	33	18
Indonesia	33	42	46	0
Slovakia	34	68	0	0
Spearman's correlation		Rs1 =0.585 p<.05	Rs2= 0.515 p<.05	Rs3 = −0.053 NS

Table 9 Dispersion in BD and DS publications as a ratio between search by topic and search by title

Category	Search by topic	Search by title	Percent of dispersion of search by topic
BD	41,961	7826	81.35
DS	244,695	2620	98.93
Interchangeable BD and DS	3552	237	93.33

Table 10 Minimal and maximal number of papers written by all authors in BD, DS and combined literatures from 2010 to 2018

Category	Minimal No. of papers per author	Maximum papers per author	Number of authors
BD	1 (80%)	65	19,727
DS	1 (78%)	27	2500
Interchangeable BD and DS	1 (93%)	6	1740

Discussion

Given the accelerated development of innovative computational methods in various fields of research and in industry, we set out to study the evolvement of literature on BD and DS. After tracing and discussing the origins of the concepts DS and BD, publication dynamics are compared. A brief discussion of standard bibliometrics follows (core journals and citations), and then we discuss funding in a geographic context. The discussion concludes with a broad synthesis of the findings.

Concept origins

The concepts of BD and DS have gone through three stages of change. They started as unrelated terms, such as “big” or “science” and “data” separated by several words. Later, they became pre-coordinated concepts. The third stage took place when the literature on BD and DS became entangled and publications included both concepts. The BD literature presents clear signs of an emerging research field as mentioned by Glänzel and Thijs (2012), and by Jaric et al. (2014): rapid generation of knowledge, a huge increase in the number of papers produced, especially since 2010, and a large number of recent publications.

The relation between the emergence of the two concepts goes against the expected developmental trend. According to the definition, “data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data” assumes the existence of data from which information and knowledge could be extracted. That is, one requires a certain amount of data first to subsequently create ways to extract knowledge from it. Yet, the developmental trend showed that the coverage of

BD in publications (before taking shape as a pre-coordinated concept) followed DS rather than precede it. Another explanation to this order of things is that DS dealt initially with limited amount of data, and the BD trend pulled DS to another level. The analysis of the DS literature as compared to the BD publications, as discussed below, helps to clarify this developmental trend.

Hu and Zhang (2017) have already elaborated on the interdisciplinary nature of publications on BD. In this study we trace the evolvements of BD and DS and observe the disciplinary changes these two concepts have gone through, and the time the literature of these two trends were intermingled (traced by a search that used the operator AND that combined the two concepts, "BD" and "DS"). A developmental and historical observation on the emergence of BD and DS shows that the disciplinary origin of the two concepts is not the same; The first papers indexed in Clarivate Analytics in the 1960s on the subject of BD are from mechanical engineering and multidisciplinary sciences, and those on the subject of DS originate from Astronomy, physical chemistry, business/economics/management and history of sciences, as shown in Table 2.

An interesting characteristic of the DS literature is that since the concept was formed, only around 16% of the literature written used it, and the rest of the DS literature appeared as before 1997, that is, with the words "Data" and "science" separated by several words. This is compared with the 91.95% publications of BD literature that appeared as a concept since the it was formed (Table 3). It seems that adoption of a pre-coordinated term (BD) by the community is an important factor in scientific community development, continuity and cohesion.

Publication dynamics

Differences appear in the publication dynamic of the DS literature and that of BD. The BD publications show an enormous jump in the number of publications, while the DS literature shows a more gradual growth rate. The BD literature is less dispersed, that is, it is more cohesive than that of DS. Growth and field dynamic indicators show that BD is ahead of DS and the combined concepts' literature in all measures. Nevertheless, the interchangeable concepts' publications trend demonstrates good results in all measures, despite the relative novelty of this literature. The immediacy index shows a higher impact of the BD literature than that of DS, and a relatively high impact of the combined concepts, considering its short time of existence. The Price index shows that the BD and the combined trend literatures are maturing, and keep growing in size, but the DS, when comparing the years 2006-2010 to 2014 -2018 is showing an aging trend.

Standard bibliometrics

The core journals covering BD and DS belong to different research orientations: computer science and engineering are more typical of BD, while interdisciplinary sciences are more typical of DS. The combined usage of BD and DS shows that three of the five core journals are those of the BD research area (Big Data and Society, Big Data, Big Data Research). Nevertheless, the combined usage included journals from the social sciences that characterize the DS research.

Citations to BD and DS literature are a function of the amount of literature published, mainly since these terms became concepts. Self-citations, though, show an inclination to increase the impact of BD studies, since it exceeds the acceptable percent of acceptable self-citations for the sciences (Pandita and Singh 2017).

Research funding

The funded publications analysis, shown in Table 8, confirms several aforementioned observations: that in the recent decade the BD publications became dominant, and this is shown also by the year the funding of this concept's publications started. It also confirms the emergence of a literature that combines both concepts, and its funding that started in 2013.

The high percent of funded papers among the publications of the BD category, as the the relatively high number of publications of this category after forming a concept (27.68%), and the high presence of BD literature, indicate that it could be the new future publication course of BD and DS, as the interchangeable use of BD and DS indicates. The analysis of the funding countries shows the massive investment of China in BD studies, while the pattern of publications of the DS literature resembles that of other subjects, that is, the USA and the western countries are in the lead of funding research in this area.

The analysis by country shows that the amount of BD and DS literatures appearing in the databases is related to the actual use of BD and DS by startups in various countries. However, the literature that combines both terms is not yet related to the actual use of BD and DS in the countries, probably because the literatures from most countries still relate to the these concepts as standing alone rather than combined, and the literature that combines both concepts is relatively new. These results help to confirm the relation between subject publications and actual usage of BD and DS in the economic context of countries.

Synthesis of findings

The relatively slow rise in the number of publications of the DS literature compared with that of BD, and the decline in the number of recent citations of DS demonstrated by the Price index of 2006-2010 compared with 2014-2018, suggests that the DS literature serves more as a theory-base or a tool-box for the BD publications, and this was indicated also by the aforementioned definitions of the DS, "At a high level, data science is a set of fundamental principles that support and guide the principled extraction of information and knowledge from data" (Provost and Fawcett 2013: 52). As suggested in this analysis, the explanation to this order of things is that DS dealt initially with a limited amount of data, and the BD trend pulled DS to another level by which DS literature serves as guide to information extraction from the Big Data retrieved.

The combined concepts literature shows a puzzling course of evolution. On one hand it has a promising Price Index recency results, but on the other hand it has a large percentage (93%) of "one time visitors", authors who contributed only one paper to this research area in the last eight years. It is difficult to establish, by the results reached, if this is a new, growing trend that marks the future of both concepts, or is it a dwindling trend that will be swollen by the BD publications' trend, and will disappear with time.

Limitations

The main limitation of this study is that it relies on WoS data which is a scholarly, selective database, but this limitation also assures the reliability of the result from an academic-bibliometric point of view. While WoS covers selected conference proceedings, coverage is partial for this type of document, thereby limiting our analysis mostly to peer-reviewed journal papers with minor coverage of conference proceedings. Observing the results of an earlier study, we conclude that the trajectory of conference publications is steeper than that of journal papers (Gupta et al. 2019). The implication for our research is to say that our findings are conservative and that including more conference proceedings would result in stronger effects in the same direction. In addition, limiting our retrieval set to items that included the search terms in their titles may have omitted some relevant items, however, searching by keywords is not supported by WoS.

Conclusions

The quantity of publications on BD since 2010 is showing an evolving research field. The emergence of the DS literature preceded that of BD, but is showing a slower trend in recent years. A new, combined literature of BD and DS appeared in the last five years. The dynamic of BD and DS literatures could be interpreted as a theory and practice relationships, as explained by Kantarovich (1993: 27) “the initial version of a theory, accompanied by a dynamic process by which a theory is adjusted to the data, and further elaborated.”

Big data can expose regularities hidden in a data, and help create stronger generalities, as well as strengthen existing theories, and contribute to advancing knowledge, so that in the future BD and DS are supposed to become concepts that are nourishing each other.

Jones (2002) explained the difference between the attitude towards data, before big data was utilized, writing that it used to be, that if data seemed to conflict with the belief one already has, the results was very often that the data in question was rejected. This state of affairs is rarely possible today, and it explains where BD research is presently heading.

References

Application Delivery Strategies. (2001). Retrieved from https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Aronova, E., Baker, K. S., & Oreskes, N. (2010). Big Science and Big Data in Biology: From the International Geophysical Year through the International Biological Program to the Long Term Ecological Research (LTER) Network, 1957–Present. *Historical Studies in the Natural Sciences, 40*(2), 183–224. https://doi.org/10.1525/hsns.2010.40.2.183.

Balietti, S., Mäs, M., & Helbing, D. (2015). On Disciplinary Fragmentation and Scientific Progress. *PLOS ONE, 10*(3), e0118747. https://doi.org/10.1371/journal.pone.0118747.

Clarke, D. A. (1975). A new guide to social science data. *Higher Education Review*, *7*(2), 11. Retrieved from https://search.proquest.com/openview/faee6f199b4f42f4d3f51feda759493d/1?pq-origsite=gscholar&cbl=1820949

Cleveland, W. S. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review / Revue Internationale de Statistique, 69*(1), 21. https://doi.org/10.2307/1403527.

Creager, A. N. H. (2010). The paradox of the phage group: Essay review. *Journal of the History of Biology, 43*(1), 183–193. https://doi.org/10.1007/s10739-010-9226-8.

Glänzel, W., & Thijs, B. (2012). Using 'core documents' for detecting and labelling new emerging topics. *Scientometrics, 91*(2), 399–416. https://doi.org/10.1007/s11192-011-0591-7.

Glänzel, W., Thijs, B., & Chi, P.-S. (2016). The challenges to expand bibliometric studies from periodical literature to monographic literature with a new data source: the book citation index. *Scientometrics, 109*(3), 2165–2179. https://doi.org/10.1007/s11192-016-2046-7.

González-Alcaide, G., Llorente, P., & Ramos, J. M. (2016). Bibliometric indicators to identify emerging research fields: publications on mass gatherings. *Scientometrics, 109*(2), 1283–1298. https://doi.org/10.1007/s11192-016-2083-2.

Gordon, A. (2004). *The status of terrorism in the academy: The comparative aspects and the role of periodicals*. Israel: University of Haifa.

Gordon, Avishag. (2007). Transient and continuant authors in a research field: The case of terrorism. *Scientometrics, 72*(2), 213–224. https://doi.org/10.1007/s11192-007-1714-z.

Gupta, V., Singh, V. K., Ghose, U., & Mukhija, P. (2019). A quantitative and text-based characterization of big data research. *Journal of Intelligent and Fuzzy Systems, 36*(5), 4659–4675. https://doi.org/10.3233/JIFS-179016.

Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing value from big data – a taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management, 36*(10), 1382–1406. https://doi.org/10.1108/IJOPM-02-2014-0098.

Hu, J., & Zhang, Y. (2017). Discovering the interdisciplinary nature of Big Data research through social network analysis and visualization. *Scientometrics, 112*(1), 91–109. https://doi.org/10.1007/s11192-017-2383-1.

Jones, M. (2002). The Concept of Prematurity and the Philosophy of Science. In E. B. Hook (Ed.), *Prematurity in Scientific Discovery: On Resistance and Neglect* (p. 306). Retrieved from https://books.google.co.il/books?id=SgCSC2P1IToC&lpg=PA306&ots=5WEOq0lbAG&dq=the concept of prematurity and the philosophy of science&lr&pg=PA306#v=onepage&q=the concept of prematurity and the philosophy of science&f=false

Kantarovich, A. (1993). *Scientific Discovery: Logic and Tinkering - Aharon Kantorovich - Google Books*. Retrieved from https://books.google.co.il/books?hl=en&lr=&id=vMFc43w0FfEC&oi=fnd&pg=PR11&dq=Scientific+discoveries+,+logic+and+tinkering&ots=Zi_qRXPpgM&sig=obyfIV07i9CU2qEIsU_lPXi4GHQ&redir_esc=y#v=onepage&q=Scientific discoveries %2C logic and tinkering&f=fal

Mattmann, C. A. (2013). A vision for data science. *Nature, 493*(7433), 473–475. https://doi.org/10.1038/493473a.

Meadows, A. J. (1998). *Communicating Science*. San Diego: Academic Press.

Mervis, J. (2012). US science policy. Agencies rally to tackle big data. *Science, 336*(6077), 22.

Mullins, N. C. (1972). The development of a scientific specialty: The phage group and the origins of molecular biology. *Minerva, 10*(1), 51–82. https://doi.org/10.1007/BF01881390.

Pandita, R., & Singh, S. (2017). Self-citations, a trend prevalent across subject disciplines at the global level: an overview. *Collection Building, 36*(3), 115–126. https://doi.org/10.1108/CB-03-2017-0008.

Papadopoulos, A. (2019, January). Most Startup Friendly Countries In The World, 2019 | CEOWORLD magazine. *CEO World*. Retrieved from https://ceoworld.biz/2019/01/02/most-startup-friendly-countries-in-the-world-2019/

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

Rons, N. (2018). Bibliometric approximation of a scientific specialty by combining key sources, title words, authors and references. *Journal of Informetrics, 12*(1), 113–132. https://doi.org/10.1016/j.joi.2017.12.003.

Rossini, F. D. (1967). The World-Wide Problem of Numerical Data for Science and Technology. *Research Management, 10*(2), 107–115. https://doi.org/10.1080/00345334.1967.11755849.

Sætra, H. S. (2018). Science as a Vocation in the Era of Big Data: the Philosophy of Science behind Big Data and humanity's Continued Part in Science. *Integrative Psychological and Behavioral Science, 52*(4), 508–522. https://doi.org/10.1007/s12124-018-9447-5.

Sanchez-Pinto, L. N., Luo, Y., & Churpek, M. M. (2018). Big Data and Data Science in Critical Care. *Chest, 154*(5), 1239–1248. https://doi.org/10.1016/J.CHEST.2018.04.037.

Simon, H., & Leker, J. (2016). USING STARTUP COMMUNICATION FOR OPPORTUNITY RECOGNITION — AN APPROACH TO IDENTIFY FUTURE PRODUCT TRENDS. *International Journal of Innovation Management, 20*(08), 1640016. https://doi.org/10.1142/s1363919616400168.

Singh, V. K., Banshal, S. K., Singhal, K., & Uddin, A. (2015). Scientometric mapping of research on 'Big Data'. *Scientometrics, 105*(2), 727–741. https://doi.org/10.1007/s11192-015-1729-9.

Tal, D., & Gordon, A. (2017). Publication attributes of leadership: what do they mean? *Scientometrics, 112*(3), 1391–1402. https://doi.org/10.1007/s11192-017-2425-8.

Tomer, C. (1986). A statistical assessment of two measures of citation: The impact factor and the immediacy index. *Information Processing & Management, 22*(3), 251–258. https://doi.org/10.1016/0306-4573(86)90057-9.

Yue, W., Wilson, C., & Rousseau, R. (2004). The immediacy index and the journal impact factor: Two highly correlated derived measures. *Canadian Journal of Information and Library Science*, *28*(1), 33–48. Retrieved from https://lirias.kuleuven.be/1110637?limo=0