DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.1, FEB. 2022

# A study on the influencing factors of cited times and social attention score of data science papers

Chunli Liu[a]*, Nanhong Sheng[b]

a. Library, China Medical University, Shenyang, China

b. School of Health Management, China Medical University, Shenyang, China

**ABSTRACT**

Data science is an emerging interdisciplinary subject in the era of big data, integrating knowledge in many fields such as machine learning, statistics, and data visualization. By analyzing the output and basic characteristics of data science papers from 2015 to 2021, this paper examines the influence of author country, open access status, discipline category, literature type, publication year, and research hotspot on the number of citations and social attention score of data science papers. The results show that data science papers continue to increase annually, with the highest number in 2017. The authors are mainly from the United States, England, Germany, and China, and accordingly mainly from North America, Europe, and Asia. Article, Review and Editorial's material are the main types of papers. Open-access papers are nearly twice as likely as non-open-access papers; Statistical analysis further confirmed that publication age and literature type had significant influence on citation times. The age of the paper, the type of the paper, the country of the author, the state of open access, and the discipline category have a significant influence on the score of social concern. Then, the comparison of keyword co-occurrence clustering diagram between highly cited papers and papers with high social attention shows that there are similarities and differences between the research hotspots of highly cited papers and papers with high social attention. The similarities are that machine learning, big data visualization and big data analysis of electronic health records are common research hotspots. While the difference is that highly cited data science papers also focus on big data analysis of business competitive advantage and big data analysis of social media. Data science papers with high AAS scores focus on open science big data analysis, bioinformatics big data analysis, and reproducible research as well.

**KEYWORDS**

Data science; Citation; Altmetric attention score; Influencing factor

# 1 Introduction

## 1.1 Data Science

Data science is an interdisciplinary field that concentrates on extracting knowledge from large data sets. The term "data science"was pointed out by Peter Naur in 1974 firstly (Cao et

*Email: liuchunliliangxu@163.com

al., 2017). However, early studies or conferences revolved around statistics which has not drawn sufficient attention. In 1998, Hayashi Chikio gave his opinion that data science consisted of three components: data design, data collection, and data analysis (Fionn et al., 2018). In the 1990s, knowledge discovery and data mining are popular terms in data science (Fayyad & Stolorz, 1997; Neil, 1999). Since 2002, data science has become a widely used term (Butka et al., 2020; Fayyad, 2002; Polkowski, 2009; Shen, 2015). The major milestones are the creation of the periodical "Data Science Journal" by the Committee on Data for Science and Technology in 2002 and "The Journal of Data Science" launched by Columbia University in 2003 (Gil et al., 2020).

At present, from the perspective of knowledge system, data science is mainly based on the knowledge of statistics, machine learning and data visualization, and its main research content includes the basic theory of data science, data processing, data calculation, data management, data analysis and data product development. As big data brings changes to people's work, life and thinking mode, data science is gaining increasing attention both within and outside the discipline and academic community (Provost & Fawcett, 2013; Torous et al., 2015). In addition, the relevant research output of data science also has a profound impact on academic research and social attention.

### 1.2 Academic impact and social attention evaluation indicators

With the birth of Science Citation Index (SCI) in the 1960s, American information scientist Eugene Garfield applied the citation analysis method to this publication, and scientometric evaluation by citation analysis method became the mainstream (Wouters, 2017). Citation frequency has gradually been accepted by the academic community and has become the authoritative indicator of academic impact evaluation of articles and journals. However, some studies have found that citation times are affected by the subject field, publication time, document type and open access status of the article (Ioannidis et al., 2016; Li et al., 2013). For articles in the field of data science, further empirical tests are needed to determine whether citations are also affected by the above factors.

Since it was coined by Priem et al. in 2010, altmetrics is considered as a new indicator, different from the traditional citation-related indicators, which can measure the social attention that scientific outputs receive (Piwowar, 2013). The Altmetric Attention Score (AAS) that derived from the company Altmetric.com has been used by most scholars (Kwok, 2013). The AAS score represents a weighted approximation of nearly all the attention a scientific output has received. Specifically, the AAS score represents the weighted value of the attention score from different data sources. News, blogs, policy literature, patents, and attention from Wikipedia are given higher weight. In contrast, attention scores from Publons, Pubpeer peer review sites, F1000, Syllabi, LinkedIn, Twitter, Facebook, Reddit, etc., are given less weight (Bornmann, 2014; Zahedi et al., 2014). The social attention evaluation of articles in the field of data science is less, and the impact of general characteristics of articles on the social attention evaluation needs to be further tested.

## 2 Data source and Methods

### 2.1 Data source

From the core collection database of Web of Science, with the subject: (" Data Science ") as the retrieval strategy, relevant articles from 2015 to 2021 are selected as the research object.

Then, all the records of 3512 related articles were exported and saved in the format of TXT file with TAB characters (Win). The data in TXT file was exported to a table file. After removing all articles without DOI number or without Pubmed ID number to ensure that sample articles were included by Altmetric.com, the full record data of 3369 articles provided by WOS database were obtained. Key fields of concern were obtained by filtering these variables: open access (OA), literature type (DT), language (LA), number of references (NR), cited frequency (TC), WOS category (WC), year of publication (PY), corresponding author address (RP), source journal (SO).

Altmetric.com tool was used to obtain altmetrics data in batches with DOI (Digital Object Unique Identifier). After removing the articles not included in altmetrics, a total of 1949 altmetric metrics were obtained. Then, the two tables obtained above and the tables downloaded from the Web of Science core collection database and downloaded from Altmetric.com were matched by the DOI numbers and merged into one data table.

### 2.2 Methods

Firstly, the general characteristics of the article are analyzed by descriptive statistics and frequency distributions. The effects of general characteristic variables on citations and Altmetric attention scores (AAS) of data science articles were analyzed by one-way ANOVA and T-test. Pearson correlation analysis was used to study the correlation between the general characteristic variables of articles and the number of citations and Altmetrics indicators. The differences in article topics between highly cited articles and articles with high Altmetrics scores were analyzed using the VOSviewer tool using the knowledge graph method of keyword co-occurrence.

## 3 Results

### 3.1 Description of basic features of data science papers

First, Table 1 summarizes the basic features of data science papers. Most data science papers are in English, accounting for 99.59%, and a small number of papers are in German (2), Japanese (1), and Spanish (5). In terms of publication years, 89.08% of papers were published in 2017, 25.45% in 2020, 21.55% in 2021, 19.39% in 2019, 14.83% in 2018, and only 9.7% in 2015 and 2016. Figure 1 shows the trend of the proportion of published data science papers from 2015 to 2021. The number of data science papers published has been on the rise year by year. Due to the statistical period ending on October 9, 2021, incomplete data statistics in 2021 led to a slight decrease in the number of papers published in 2021 compared with that in 2020. It is worth noting that from 2016 to 2017, there was a sudden growth of more than 9 times, and in 2018, it resumed the conventional growth range.

**Table 1** General feature distribution of data science papers

| General feature | Value | Percent (Number) |
|---|---|---|
| Language | English | 99.59% (1941) |
| | German | 0.10% (2) |
| | Japanese | <0.1% (1) |
| | Spanish | <0.26% (5) |
| Document types | Article | 68.75% (1340) |

| General feature | Value | Percent (Number) |
|---|---|---|
| | Article; Early Access | 2.31% (45) |
| | Article; Proceedings Paper | 1.80% (35) |
| | Article; Book Chapter | 0.56% (11) |
| | Article; Data Paper | 0.30% (6) |
| | Review | 15.85% (309) |
| | Review; Early Access | 0.56% (11) |
| | Review; Book Chapter | 0.41% (8) |
| | Editorial Material | 8.56% (167) |
| | Book Review | 0.26% (5) |
| | Letter | 0.26% (5) |
| | Correction | 0.21% (4) |
| | News Item | 0.10% (2) |
| | Software Review | <0.1% (1) |
| Publication year | 2015 | 3.08% (60) |
| | 2016 | 6.62% (129) |
| | 2017 | 89.08% (177) |
| | 2018 | 14.83% (289) |
| | 2019 | 19.39% (378) |
| | 2020 | 25.45% (496) |
| | 2021 | 21.55% (420) |
| Corresponding author´s country | United State | 43.92% (856) |
| | England | 9.49% (185) |
| | Germany | 5.54% (108) |
| | China | 4.26% (83) |
| | Spain | 3.49% (68) |
| | Netherlands | 2.98% (58) |
| | Italy | 2.51% (49) |
| | Japan | 2.15% (42) |
| | France | 2.15% (42) |
| | Australian | 2.05% (40) |
| | Others (Individual countries account for less than 2%) | 20.42% (398) |
| | The corresponding author´s address is missing. | 1.03% (20) |
| Corresponding author´s continent | North America | 46.13% (899) |
| | European | 36.94% (720) |
| | Asia | 11.44% (223) |
| | Oceania | 2.57% (50) |
| | South America | 1.69% (33) |
| | Africa | 0.21% (4) |
| | The corresponding author´s address is missing. | 1.03% (20) |
| Open Access Status | Open Access (OA) | 68.45% (1334) |
| | Non–Open Access (Non–OA) | 31.55% (615) |

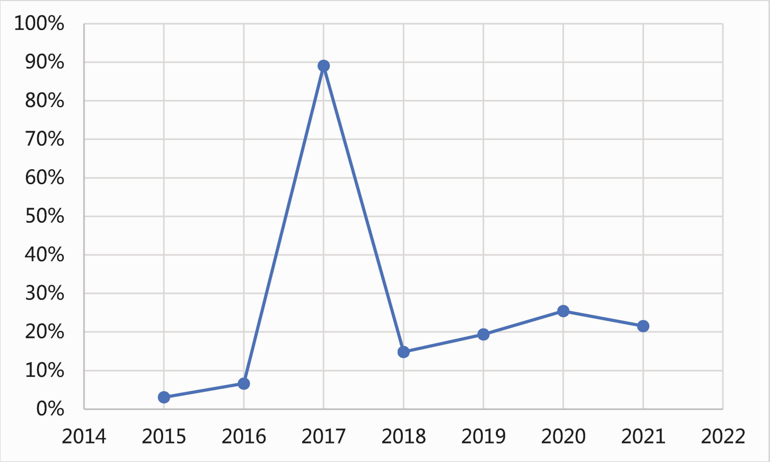

**Figure 1** Trends in the proportion of published data science papers from 2015 to 2021

In addition, the national distribution of corresponding authors' institutions (if the corresponding author belongs to two or more countries at the same time, the first country is used to calculate the author's attribution) was analyzed. As shown in Figure 2, the corresponding authors in the United States published 856 papers, accounting for 43.92%. Secondly, 185 papers were published by corresponding authors in the UK, accounting for 9.49%. German authors published 108 papers, accounting for 5.54%; Chinese scholars published 83 papers, accounting for 4.26%. Spanish authors published 68 papers, accounting for 3.49%; Dutch authors published 58 papers, accounting for 2.98%; Italian authors published 49 papers, accounting for 2.51%.

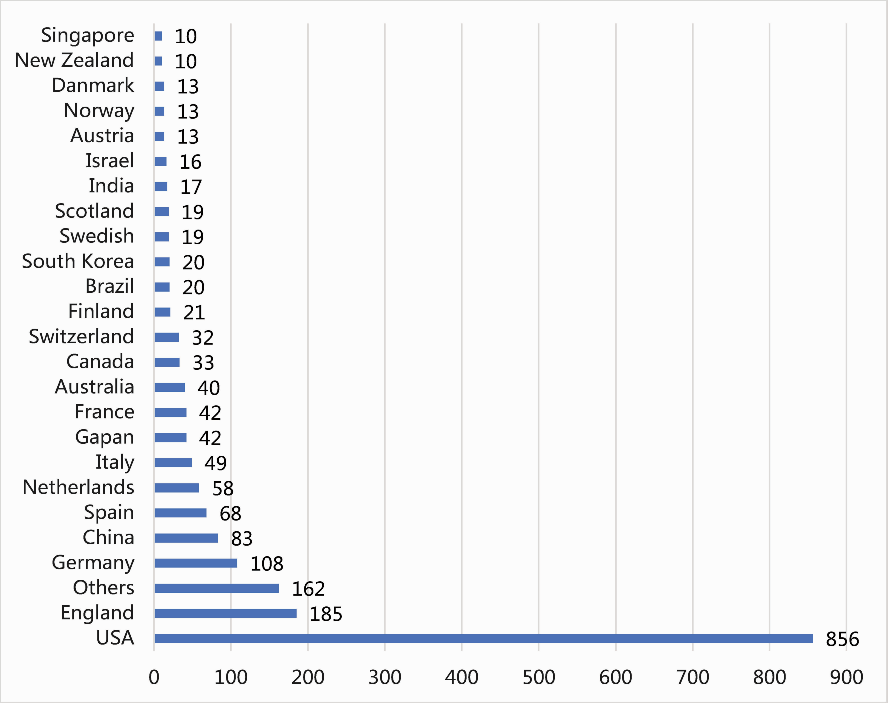

**Figure 2** Number of papers from different countries

According to the continents of corresponding authors, North America ranked first, with 899, accounting for 46.13%; Europe ranked second with 720, accounting for 36.94%; Asia ranked third with 223, accounting for 11.44%. Oceania ranked fourth with 50 articles, accounting for 2.57%; South America and Africa accounted for 1.69% and 0.21%, respectively (See Figure 3).

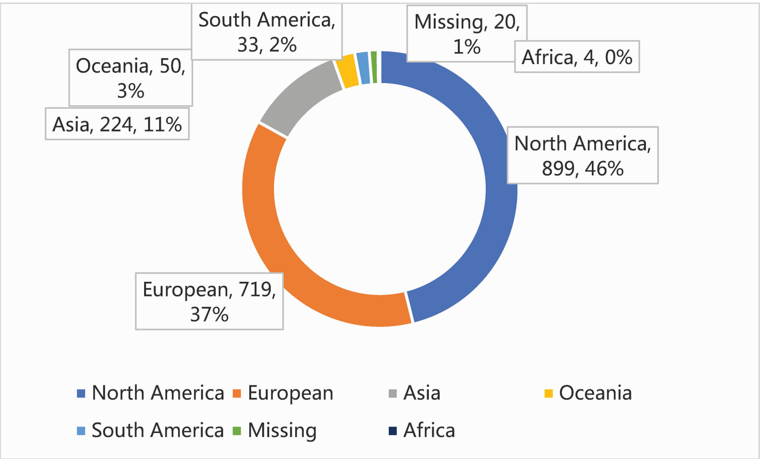

**Figure 3** Number of papers from different continents

In terms of document types, there are eight types of papers on data science, including Article (1437), Review (328), Editorial Material (167), Book Review (5), Letter (5), Correction (4), News Item (2), and Software Review (1) (See Figure 4). Among them, Article accounted for 73.73%, Review accounted for 16.83%, and Editorial material accounted for 8.57%. The Article category also includes general Article papers (1340), Proceedings papers (35), Book chapter papers (11), Data papers (6), and Early Access papers (45), according to Web of Science. The Review category includes general Review papers (309), Early access papers (11), and Book chapter papers (8).

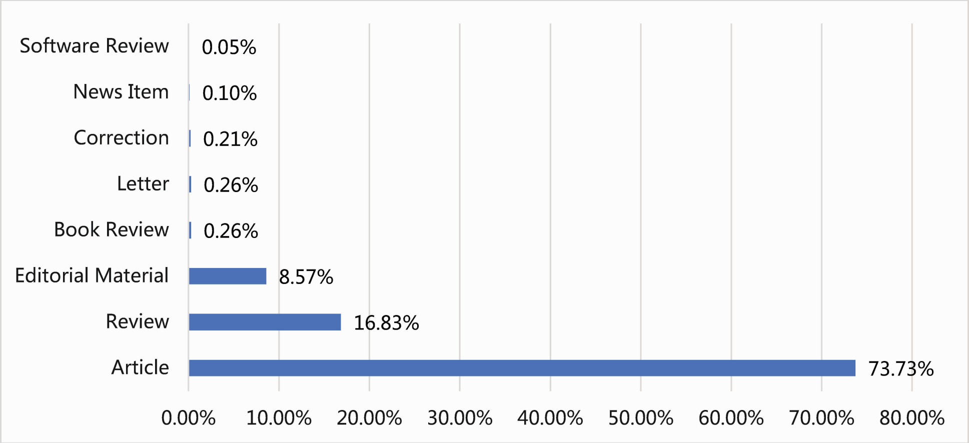

**Figure 4** Proportion of document types

Table 2 reveals the subject classification of data science papers and the number of publications, percentage, and cumulative percentage. Due to limited space, only 28 disciplines and the number of papers in the top 60.80% are listed. In data science papers, there are more papers belonging to Multidisciplinary Sciences (124); the next most popular subject is computer science and its sub-subjects (Information Systems (111), Interdisciplinary Applications (86), Artificial Intelligence (79)), Statistics & Probability, Biochemical Research Methods (54), et al. While there are 22 papers belonging to medical informatics and 21 papers belonging to Information Science & Library Science.

**Table 2** Distribution of Web of Science Category of data science papers

| | Web of Science Category | Numbers | Percentage | Cumulative Percentage |
|---|---|---|---|---|
| 1 | Multidisciplinary Sciences | 124 | 6.36% | 6.36% |
| 2 | Computer Science, Information Systems | 111 | 5.70% | 12.06% |
| 3 | Computer Science, Interdisciplinary Applications | 86 | 4.41% | 16.47% |
| 4 | Computer Science, Artificial Intelligence | 79 | 4.05% | 20.52% |
| 5 | Statistics & Probability | 54 | 2.77% | 23.29% |
| 6 | Biochemical Research Methods | 51 | 2.62% | 25.91% |
| 7 | Chemistry, Multidisciplinary | 51 | 2.62% | 28.53% |
| 8 | Chemistry, Physical | 47 | 2.41% | 30.94% |
| 9 | Environmental Sciences | 45 | 2.31% | 33.25% |
| 10 | Health Care Sciences & Services | 41 | 2.10% | 35.35% |
| 11 | Education & Educational Research | 39 | 2.00% | 37.35% |
| 12 | Public, Environmental & Occupational Health | 38 | 1.95% | 39.30% |
| 13 | Business | 35 | 1.80% | 41.10% |
| 14 | Biochemistry & Molecular Biology | 33 | 1.69% | 42.79% |
| 15 | Oncology | 33 | 1.69% | 44.48% |
| 16 | Nursing | 31 | 1.59% | 46.07% |
| 17 | Medicine, General & Internal | 30 | 1.54% | 47.61% |
| 18 | Materials Science, Multidisciplinary | 27 | 1.39% | 49.00% |
| 19 | Radiology, Nuclear Medicine & Medical Imaging | 26 | 1.33% | 50.33% |
| 20 | Management | 25 | 1.28% | 51.61% |
| 21 | Biology | 24 | 1.23% | 52.85% |
| 22 | Biotechnology & Applied Microbiology | 23 | 1.18% | 54.03% |
| 23 | Cardiac & Cardiovascular Systems | 23 | 1.18% | 55.21% |
| 24 | Engineering, Biomedical | 23 | 1.18% | 56.39% |
| 25 | Green & Sustainable Science & Technology | 22 | 1.13% | 57.51% |
| 26 | Medical Informatics | 22 | 1.13% | 58.64% |
| 27 | Information Science & Library Science | 21 | 1.08% | 59.72% |
| 28 | Pharmacology & Pharmacy | 21 | 1.08% | 60.80% |

As far as open access status is concerned, the Web of Science Platform provides data obtained from the OurResearch Unpaywall Database. Of the 1949 papers, 1334 were open access papers, accounting for 68.45%, while 615 were non-open access papers, accounting for 31.55%. Open-access papers account for twice as many as non-open-access papers and more than two-thirds of all papers. According to the definition of Unpaywall, Open access status includes legal Gold or Bronze (free content at a publisher's website) and Green (e.g., Author self-archived in a repository) OA versions. Gold includes gold and hybrid, and Green includes Green published, Green accepted, and Green Submitted. For a paper, it may belong to one of these categories, or it may belong to more than two of these categories simultaneously. For example, a paper belongs to Bronze, Green Published, Green Submitted; Some papers belong to gold, Green Published, Green Submitted, and Green Accepted. We counted the frequency of each category separately and found that in the open access state, The frequencies from high to low are Green Published (64), Green Submitted (60), Green accepted (57), gold (32), hybrid (28) and bronze (22).

## 3.2 Associations of the features of papers with TC and AAS

We analyzed the effect of the features of papers on TC and AAS by one-way ANOVA and T-test. Table 3 has shown the results. Firstly, publication age and document types significantly influenced TC and AAS ($p < 0.001$). The older the publication age, the higher the citation times. Similarly, the older the publication age, the higher the Altmetric attention score ($p < 0.001$). Document type only significantly affected the citation times of papers ($p < 0.001$), but the influence on AAS was not statistically significant ($p > 0.05$). For TC, the top five document types with the most citations are Letter, Review, Software Review, Article, and Correction, respectively; for AAS, the top five ones with the most AAS are News Item, Letter, Editorial Material, Article, Review, and Correction respectively. Although Country, Open access, and Web of Science Subject Category did not significantly affect citation times, their effect on AAS was statistically significant. Data science papers from the UK and the US received the highest AAS scores, while those from China had relatively low AAS scores. Open-access data science papers received higher AAS scores, nearly double the AAS scores of non-open-access data science papers. Statistically, it was found that the Web of Science Subject categories of papers significantly affected the average AAS score of papers. Ranking the Web of Science subject categories by the number of papers, papers from the top ten subjects have the highest average AAS scores, followed by papers from the 11-20 subjects, and papers from the 21st and later subjects have a relatively lower average AAS scores.

**Table 3** Mean scores of symptoms of TC and AAS according to the features of papers.

| Variables | TC | F/t value | P value | AAS | F/t value | P value |
|---|---|---|---|---|---|---|
| Publication Age (year) | | 20.842 | 0.000 | | 9.457 | 0.000 |
| 4–6 | 38.58±100.51 | | | 31.92±98.96 | | |
| 2–3 | 22.27±94.92 | | | 18.92±62.76 | | |
| 1 | 7.04±12.61 | | | 14.40±29.36 | | |
| 0 | 2.69±27.56 | | | 10.79±21.04 | | |
| Document types | | 32.057 | 0.000 | | 5.159 | 0.000 |
| Article | 14.54±38.40 | | | 18.13±59.07 | | |
| Review | 27.36±92.33 | | | 17.17±64.33 | | |
| Editorial Material | 7.81±17.90 | | | 20.53±38.40 | | |
| Book Review | 0.80±0.83 | | | 2.80±3.03 | | |
| Letter | 468.40±1045.13 | | | 61.60±106.65 | | |
| Correction | 11.25±20.54 | | | 5.25±7.93 | | |

| Variables | TC | F/t value | P value | AAS | F/t value | P value |
|---|---|---|---|---|---|---|
| News Item | 5.00±2.82 | | | 256.50±358.50 | | |
| Software Review | 26 | | | 3 | | |
| Country | | 0.786 | 0.534 | | 3.243 | 0.012 |
| USA | 20.32±103.50 | | | 22.25±72.78 | | |
| England | 18.32±49.04 | | | 23.58±42.84 | | |
| Germany | 15.33±38.19 | | | 8.80±19.87 | | |
| China | 16.39±36.06 | | | 2.61±6.13 | | |
| Others | 14.00±31.59 | | | 15.62±51.38 | | |
| Open Access | | 0.528 | 0.467 | | 11.048 | 0.001 |
| Yes | 18.05±76.54 | | | 21.49±61.96 | | |
| No | 15.47±65.00 | | | 11.87±53.43 | | |
| Web of Science subject category | | 0.904 | 0.405 | | 10.457 | 0.000 |
| Top ten subjects | 22.62±98.09 | | | 30.67±88.36 | | |
| Top 11–20 subjects | 18.51±46.84 | | | 27.89±92.74 | | |
| Rank 21 and beyond | 16.13±70.82 | | | 15.07±45.75 | | |

TC=Times of being cited.
AAS=Altmetric Attention Scores.

### 3.3 Comparison of the research hotspots for high TC papers and the research hotspots for high AAS papers

By counting the frequency of the occurrence of key words in the same document, a co-word network composed of these word pairs can be formed. The co-occurrence of high-frequency keywords can reveal hot topics in the research field and analyze the research hot spots and frontiers in a certain field. By comparing the co-word network differences between highly cited papers and papers with high social attention, the differences of research hotspots between the two data sets can be revealed, and the influence of research hotspots on TC and AAS can be further reflected. We took the top 20% of TC papers as highly cited papers, and there were 395 papers in total. The top 20% of AAS papers are highly concerned by the society, 390 papers in total. We use keyword co-occurrence method to analyze the topic distribution of highly cited papers and papers with high AAS, and the threshold is set at 7.

Figure 5 and Figure 6 show the keyword co-occurrence network of high TC papers and high AAS papers, respectively. We used VOSviewer software for keyword co-occurrence analysis and clustering. Before clustering, we cleaned the data and merged synonyms, singular and plural words, abbreviations, and other different expressions of the same word.

A total of 2271 keywords were screened out from the keyword co-occurrence network of highly cited papers (See Figure 5). Finally, 63 keywords with word frequency greater than or equal to 7 were selected for analysis, and 4 categories were obtained. In the keyword co-occurrence cluster graph, the label size represents the frequency of keywords, and the larger the label, the higher the frequency of keywords. We summarized the topic names of each cluster according to the high-frequency keywords, and revealed the research hot spots of each cluster as follows:

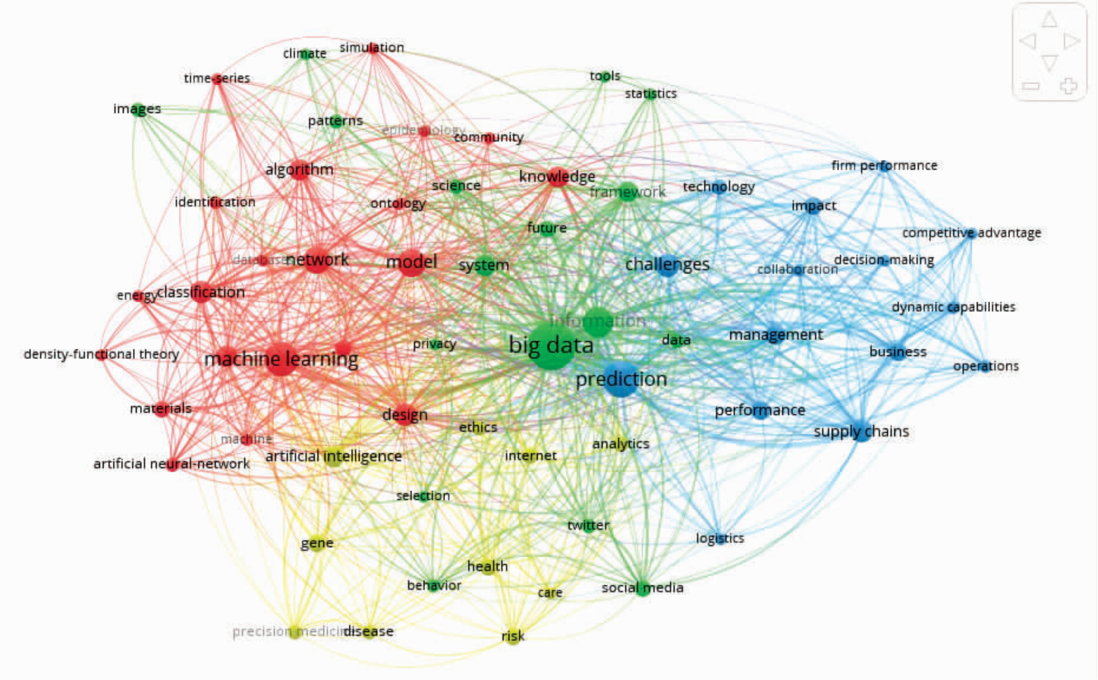

**Figure 5** Keyword co-occurrence network of high TC papers of data science

Cluster 1: machine learning algorithm including artificial neural-network simulation, deep learning, density-functional theory, classification of cancer images and prognostic prediction using machine learning methods.

Cluster 2: big data analysis on social media includes big data analysis on social media usage behaviors, patterns, tools, systems, climate, and other aspects.

Cluster 3: big data analysis of the business competitive advantage based on supply chain management, including big data from logistics, supply chains, firm performance, impact, management, dynamic capabilities, prediction, operations.

Cluster 4: big data research on artificial intelligence in the field of health care, including precision medicine, disease risk, health analytics, health care, gene analytics.

A total of 1811 keywords were selected from the keyword co-occurrence network of papers with high social concern (See Figure 6). Finally, 49 keywords with word frequency greater than or equal to 7 were selected for analysis, and a total of 5 categories were obtained. We summarized the research hot spots of each cluster as follows:

Cluster 1: Machine learning algorithm including artificial neural-network simulation, deep learning, density-functional theory, classification of cancer images, and prognostic prediction using machine learning methods.

Cluster 2: Big data analysis of electronic health records, including analyzing the mortality risk, quality of life, genetic risk, and trends in health and exercise benefits from private data

Cluster 3: Big data visualization of the framework, model, system, design, challenges, future, climate, knowledge.

Cluster 4: big data analytics on bioinformatics, the community, education, open science, reproducibility and reproducible research, statistics, the tools.

Cluster 5: big data analytics on health and care, ethics, information, internet, COVID-19,

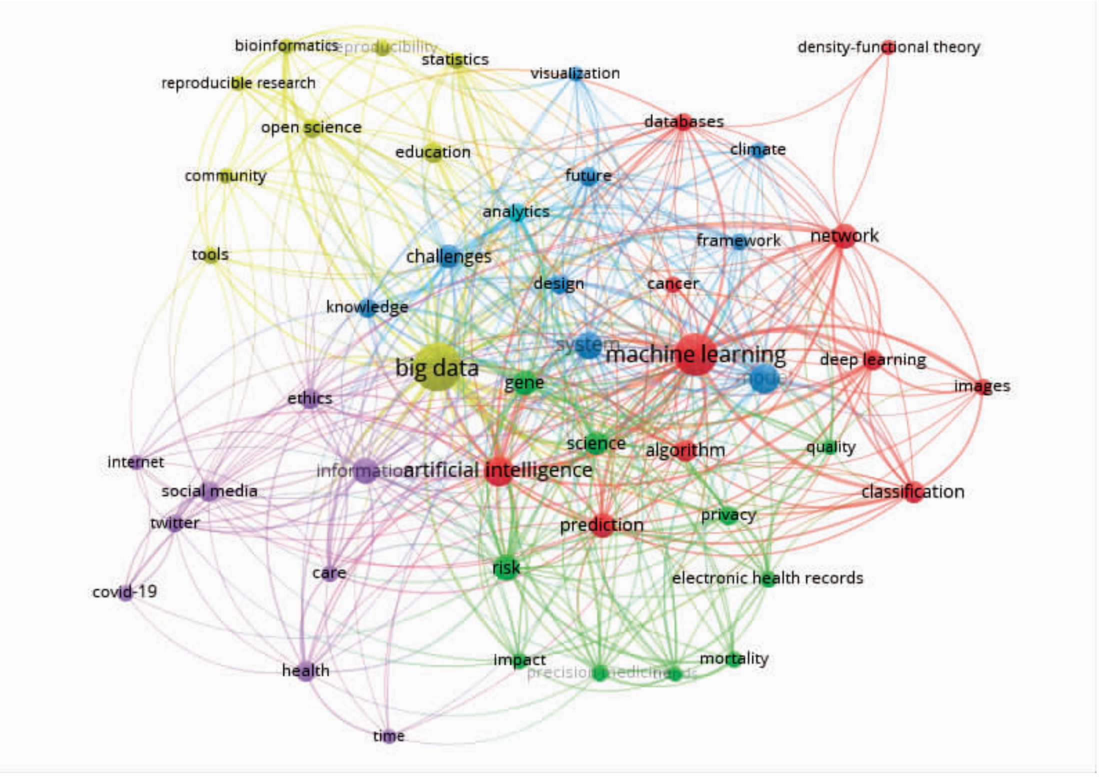

**Figure 6** Keyword co-occurrence network of high AAS papers of data science

and social media such like twitter.

Though comparing the research hotspots for high TC papers and the research hotspots for high AAS papers, we found the machine learning algorithm, big data analytics of social media, big data analytics of electronic health records deriving from the artificial intelligence equipment are common research hotspots of both high TC papers and high AAS papers. The difference is that high TC papers emphasize on the big data analysis of the business competitive advantage. While high AAS papers focus on bioinformatics, open science, reproducible research, community research. Therefore, the content of the study undoubtedly influenced the number of citations and social attention scores.

## 4 Conclusions

In this paper, we descriptively analyzed the basic features of data science articles such as language, publication year, open access status, document types, country of the corresponding author, web of science subject category. We have found that English is the main language of data science papers; 2017 was the year with the highest number of data science papers published in nearly seven years, and it was the year that stood out. The United States, The United Kingdom, Germany, and China are the top four countries with the largest number of data science research papers. Accordingly, North America, Europe, and Asia accounted for the highest proportion of data science papers. Article and review are the two largest publication types of data science papers. Editorial material ranked third. The company Clarivate explains the document type "editorial material" as an article that gives the opinions of a person, group, or organization. Researchers have pointed out that editorial materials should be considered valuable outputs and should not be excluded from acceptable outputs (Leeuwen et al., 2013). Open-access papers are nearly twice as likely as non-open-access papers.

As for the papers on data science, most of them belong to the category of multidisciplinary, computer science and statistics. This is followed by biochemistry, environmental science, health care, education, public, environmental and occupational health, and business. The results of one-way ANOVA showed that publication year and literature type significantly affected citation times and AAS score. In addition, the author's country of residence, the open access status of the paper and the discipline category also significantly affected the AAS score of the paper. By comparing the research hotspots of data science papers with high citation scores and high AAS scores, it is found that there are similarities and differences between them. The results show that data science papers focusing on machine learning, big data visualization, and big data analysis of electronic health records have higher citation times and AAS scores. However, the difference is that highly cited data science papers also include big data analysis of business competitive advantage and big data analysis of social media. Data science papers with high AAS scores also include open science big data analysis, bioinformatics big data analysis, and reproducible research.

## Acknowledgments

This work was supported by one of the research results of "Evaluation and Prediction of academic influence, Social Attention, and clinical transformation potential of COVID-19 Scientific Research Achievements", 2020 Scientific Research Fund Project of Education Department of Liaoning Province - Young Scientific and Technological Talents "Seedling" project (Project No. QNRW2020004).

## References

Bornmann, L. (2014) . Validity of altmetrics data for measuring societal impact: A study using data from Altmet– ric and F1000Prime. *Journal of Informetrics, 8* (4) , 935–950.

Butka, P., Bednár, P., & Ivanáková, J. (2020) . Methodologies for Knowledge Discovery Processes in Context of AstroGeoInformatics. In P. Škoda & F. Adam (Eds.) , *Knowledge Discovery in Big Data from Astronomy and Earth Observation* (pp.1–20) . Elsevier.

Cao, L. (2017) . Data Science: A comprehensive overview. *ACM Computing Surveys, 50* (3) , 1–42.

Fayyad, U. (2002) . Data mining and knowledge discovery in databases: Implications for scientific databases. *Proceedings of International Conference on Scientific & Statistical Database Management*. IEEE.

Fayyad, U., & Stolorz, P. (1997) . Data mining and KDD: Promise and challenges. *Future Generation Computer Systems, 13* (2–3) , 99–115.

Ioannidis, J., Kevin, B., & Wouters, P. F. (2016) . Citation metrics: A primer on how (not) to normalize. *PLOS Biology, 14* (9) , e1002542.

Kwok, R. (2013) . Research impact: Altmetrics make their mark. *Nature, 500* (7463) , 491.

Leeuwen, T., Costas, R., Calero–Medina, C., et al. (2013) . The role of editorial material in bibliometric re– search performance assessments. *Scientometrics, 95* (2) , 817–828.

Li, Y., Radicchi, F., Castellano, C., & Ruiz–Castillo, J. (2013) . Quantitative evaluation of alternative field nor– malization procedures. *Journal of Informetrics, 7* (3) , 746–755.

Murtagh, F., & Devlin, K. (2018) . The development of data science: Implications for education, employment, research, and the data revolution for sustainable development. *Big Data and Cognitive Computing, 2* (2) , 14.

Neil, J. R., & Korb, K. B. (1999) . The evolution of causal models: A comparison of Bayesian metrics and structure priors. *Pacific –Asia Conference on Methodologies for Knowledge Discovery & Data Mining.*

Springer–Verlag.

Piwowar, H. (2013) . Altmetrics: Value all research products. *Nature, 493* (7431) ,159.

Polkowski, L. (2009) . Data–mining and knowledge discovery: Case–based reasoning, nearest neighbor and rough sets. In R. Meyers (ed.) , *Encyclopedia of Complexity and Systems Science* (pp. 1789–1810) . Springer.

Press, G. (2013) . A Very Short History of Data Science. Forbes. https://www.forbes.com/sites/gilpress/2013/05/28/a–very–short–history–of–data–science

Provost, F., & Fawcett, T. (2013) . Data science and its relationship to big data and data–driven decision making. *Big Data, 1* (1) , 51–59.

Shen, B. (2015) . Universal knowledge discovery from big data using combined dual–cycle. *International Journal of Machine Learning & Cybernetics, 9*, 133–144.

Torous, J., & Baker, J. T. (2015) . Why psychiatry needs data science and data science needs psychiatry. *JAMA Psychiatry, 73* (1) , 3–4.

Wouters, P. (2017) . Eugene Garfield (1925–2017) . *Nature, 543*, 492. https://doi.org/10.1038/543492a

Zahedi, Z., Costas, R., & Wouters, P. (2014) . How well developed are altmetrics? A cross–disciplinary analysis of the presence of ´alternative metrics´ in scientific publications. *Scientometrics, 101* (2) , 1491–1513.