Data Science Publication: Thirty-Six Years Lesson of Scientometric Review

Agung Purnomo
Entrepreneurship Department,
BINUS Business School Undergraduate Program
Bina Nusantara University
Jakarta, Indonesia 11480
agung.purnomo@binus.ac.id

Elsa Rosyidah
Environmental Engineering Department
Universitas Nahdlatul Ulama Sidoarjo
Sidoarjo, Indonesia
elsarosyidah@unusida.ac.id

Mega Firdaus
English Education Department
Universitas Nahdlatul Ulama Sidoarjo
Sidoarjo, Indonesia
megafirdaus22@gmail.com

Nur Asitah
Primary Education Department
Universitas Nahdlatul Ulama Sidoarjo
Sidoarjo, Indonesia
nurasitah@gmail.com

Andre Septianto
Chemical Engineering Department
Universitas Nahdlatul Ulama Sidoarjo
Sidoarjo, Indonesia
andreseptianto@gmail.com

***Abstract*— Data science as part of technological development is growing and needed. It has been research yet the notion about data science review publication which showed the big picture using data from all countries. This research aims to study the position of the international publication map of data science indexed by Scopus using scientometric review. Scientometric methods and analyzed research data was used to analyze search results service from Scopus and the VOSviewer application. The research data of 5,202 documents published from 1983 to 2019 were obtained from the Scopus database. Most countries, subject areas, and type documents in data science publications were the United States, computer science; and conference paper. There were ten collaborative researchers' group patterns. This research proposes a convergence axis classification consisting of data science publication to characterize the body of knowledge generated from three decades of publication: Machine learning, Organism, Data mining, and Data analysis, abbreviated as MODD themes.**

Keywords—data science, publication mapping, scientometric, vosviewer.

I. INTRODUCTION

Technological progress has fundamentally changed the way we see a world full of data [1]. Automatic image acquisition, modern data, and machine learning-based data processing accelerate the process of finding statistical data [2]. The extreme multi-label classification that occurs in apps is different from the tag or ad prediction [3]. Various studies use advanced artificial intelligence to predict or predict problems that will occur in the real world [4]–[6]. Traditional data processing methods still have an important role in analyzing all types of data. However, this method has not been able to exploit large amounts of data or even process large amounts of data to find knowledge, describe complex relationships between data, and predict data behavior [7]. Data science is a study that synergizes programming skills, areas of expertise, and statistical and mathematical knowledge to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to text, numbers, images, audio, and video to produce artificial intelligence systems that perform tasks that replace human intelligence [8]. The need for increasing data analysis made the popularity of data science increase [9].

The development of data science is experiencing growth. The approach and methodology of data science are dynamic and have become so amazing all over the world [10], [11]. The field of data science development offers a path to a better understanding [12]. Data analysis is a branch of science including an algorithmic process that is used to analyze data sets which are then extracted to obtain information from these data [9]. Data are technological innovations and genetic solutions that provide solutions [12] that are needed for research. The rapid increase in the use of data science creates demand for professionals related to the data [13]. The availability of large-scale and exponential mobility data can trigger smart cities that can change our lives [14]. In-depth data learning can be a preparation for processing data better in the next generation [7].

The implementation of data science is increasingly needed. The world has transitioned from the environment with a scarcity of data to a data-rich environment. This happens with the use of digital technology such as cell phones in all directions [7], [15]. Every new era always provides opportunities and challenges [16]. Artificial intelligence simulates human intelligence and produces new intelligent machines that can process information with human consciousness, behavior, and thought [5]. The use of data science is not merely used in the world of technology but also in the area of health. Emmert-Streib, Dehmer, and Yli-Harja explained the model serves two cooperative stages that pair data generation and data analysis most beneficially from the patient's perspective to ensure quality standards of data analysis including reproducible research [16]. The best practices of data science spread across all fields of research. By providing new opportunities for research and education [17]. Data science strongly supports the development of business and entrepreneurship.

Big data analytics using data science can provide benefits for governments to facilitate services to citizens, as well as providing smart policies that are supported by strong data [14]. Not much research on data science providing a big picture that is visualized from year to year with data from all research publications at the global level. Also, it has been research yet specifically the notion about the relationship between authors, affiliations, and the impact of their research. Therefore, the research aims to study the position of the international publication map of data science indexed by Scopus using scientometric review. This research have

978-1-7281-7071-8/20/$31.00 ©2020 IEEE 13-14 August 2020
2020 International Conference on Information Management and Technology (ICIMTech)

been monitored the growth number of academic documents related to the topic of Data Science that have been published and indexed by Scopus from 1983 to December 2019.

II. Research Methods

This research has studied the position of map publications in the global level of data science and indexed Scopus. Research data were obtained from the Scopus database using the document search service in April 2020 [18]. This study has carried out scientometric methods and the data were analyzed using the analyze search results service from Scopus and the VOSviewer application [19], [20]. VOSViewer tools were used to build and visualize scientometric networks, namely the number of studies, researchers, academic affiliations, countries, fields, keywords, and author collaboration [21]. This survey has been carried out by identifying keywords related to data science to find and identify related articles from publications with the Scopus database for 5,202 academic documents published from 1983 to December 2019 at the global level. The data collection was limited to 2019 without looking at 2020 (exclude 2020) thus the annual data obtained was the illustration to the condition of the study in one whole year from January to December. The query command that is applied when mining data on Scopus was TITLE-ABS-KEY ("data science") AND PUBYEAR <2020.

The research analyzed co-authorship author unit analysis and full counting methods using VOSViewer to get the author's collaboration network. The study carried out an analysis of theme or co-occurrence along with keywords analysis and a full calculation method using VOSViewer to obtain a network of keywords.

III. Result and Discussion

Publications about data science have increased every year. The highest publication peak was in 2019, 2000 documents. Research on data science itself has begun since 1983.

A. Most Common Country of Data Science Publications

The top research country in data science publications was the United States with 2.109 documents. Then, followed by the United Kingdom, India, Germany, China, Italy, Canada, Australia, Netherlands, and France.

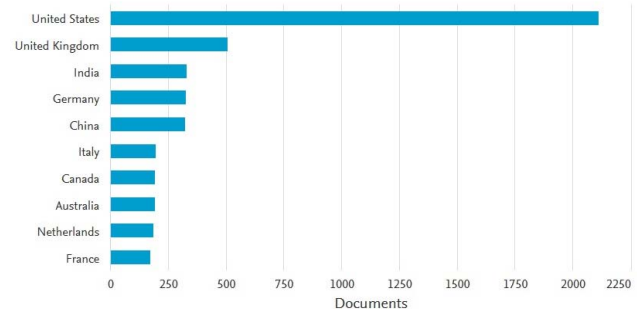

Fig. 1. Country Number of Data Science in Year

B. Most Common Institutional Affiliations of Data Science Publications

The top 10 research institutions in data science publications was Stanford University. Then, followed by University of Washington, Seattle, Georgia Institute of Technology, Massachusetts Institute of Technology, University of California, Berkeley, Columbia University in the City of New York, the University of Illinois at Urbana-Champaign, Carnegie Mellon University, CNRS Centre National de la Recherche Scientifique, New York University, University of Michigan, Ann Arbor, and the University of Southern California.

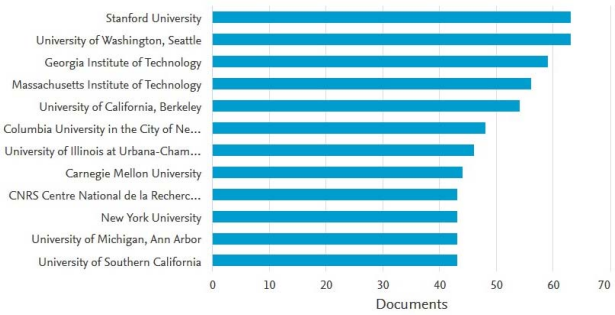

Fig. 2. Affiliation Number of Data Science in Year

C. Most Individual Authors in Data Science Publications

The author with the most publications in the field of Data Science was Leung, C.K. with 25 documents. Followed by Lötsch, J. with 24 documents, Kalidindi, S.R. with 18 documents, Emmert-Streib, F. and Saltz, J.S. with 15 documents, Dehmer, M. with 14 documents, Cuzzocrea, A. and Demchenko, Y. with 12 documents, Ultsch, A. with 11 documents, Delaney, C.W., Howe, B., Pappalardo, L., and Wiktorski, T. with 10 documents.

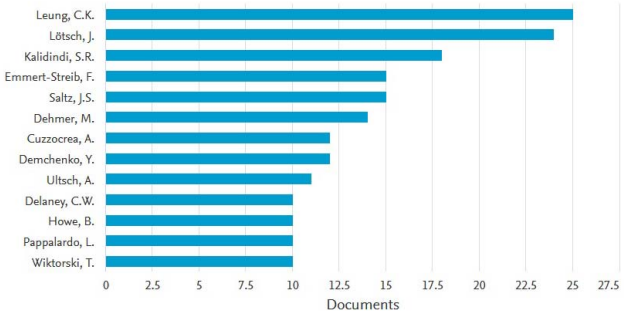

Fig. 3. Most Individual Authors of Data Science

D. Most Frequency of Data Science by Subject Area

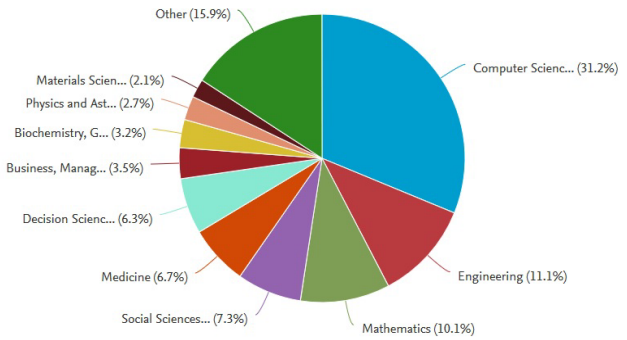

Fig. 4. Most Frequency of Data Science Publications by Subject Area

The most subject area in data science publications was Computer Science with 3.004 documents (31.2%). Followed by Engineering with 1.070 documents (11.1%), Mathematics with 976 documents (10.1%), Social Sciences with 703 documents (7.3%), Medicine with 647 documents (6.7%), Decision Sciences with 603 documents (6.3%), Business, Management and Accounting with 333 documents (3.5%), Biochemistry, Genetics and Molecular Biology with 304 documents (3.2%), Physics and Astronomy with 260 documents (2.7%), and Materials Science with 201 documents (2.1%).

E. Most Frequent Type Document of Data Science Publications

The most frequent type document in data science publication was Conference Paper (40.8%) with 2.125 documents. Then Article (36.9%) with 1.918 documents, Review (6.5%) with 340 documents, Book Chapter (4.3%) with 224 documents, Conference Review (4.2%) with 217 documents, Editorial (2.5%) with 128 documents, Note (1.7%) with 91 documents, Book (1.4%) with 73 documents, Erratum (0.6%) with 31 documents, and Short Survey (0.5%) with 24 documents.

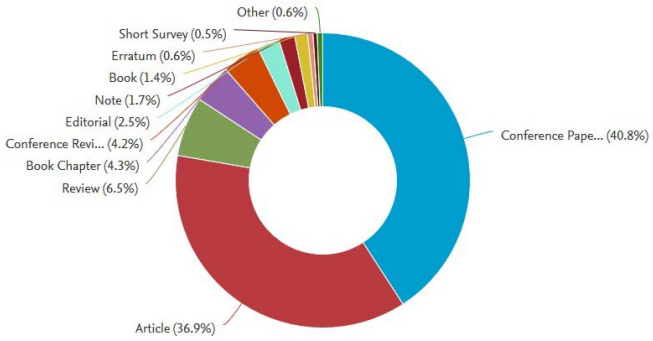

Fig. 5. Most Frequent Type Document of Data Science Publications

F. Year Documents Based on Sources of the Data Science

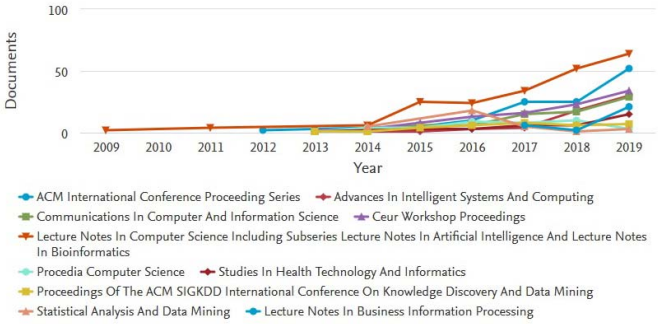

Fig. 6. Number of Documents Each Year Based on Sources of the Data Science

The number of documents each year based on sources in international publications in the Data Science publications was Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics with 211 documents. Followed by the ACM International Conference Proceeding Series with 126 documents, Ceur Workshop Proceedings with 99 documents, Communications In Computer And Information Science with 76 documents, Advances In Intelligent Systems And Computing with 57 documents, Procedia Computer Science with 39 documents, Studies In Health Technology And Informatics with 34 documents, Proceedings Of The ACM SIGKDD International Conference On Knowledge Discovery And Data Mining with 33 documents, Statistical Analysis And Data Mining with 32 documents, and Lecture Notes In Business Information Processing with 29 documents.

G. Document Years in Data Science

International publications on data science have been started since 1983. The number of international publications on the topic of data science has shown an increasing trend every year. This can be seen in Figure 7, the highest publication peak in 2019 with 2,000 documents.

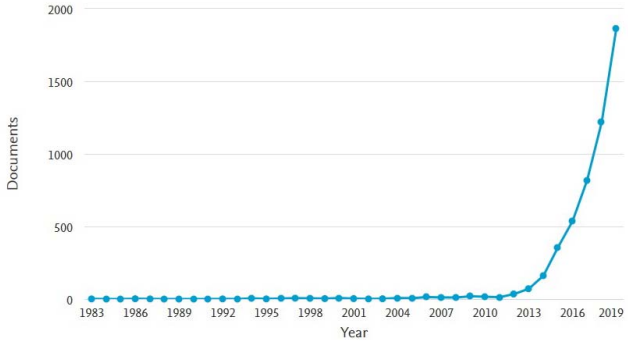

Fig. 7. Number of Documents Per Year of Data Science Publications

H. The document cited of the Data Science

The topmost cited publications were the work of Jordan, M.I., Mitchell, T.M. in 2015 entitled “Machine Learning: Trends, Perspectives, And Prospects” in Science journal with cited by 862 documents.

I. *Documents by Funding Sponsor of the Data Science*

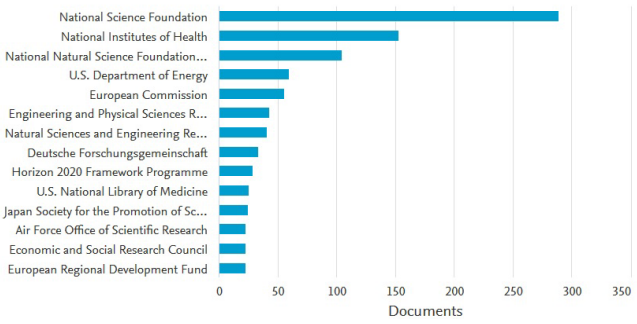

Fig. 8. Number of Documents by Funding Sponsor of the Data Science

The number of documents by funding sponsors from the Data Science was the National Science Foundation with 288 documents. Followed by the National Institutes of Health with 152 documents, National Natural Science Foundation of China with 104 documents, U.S. Department of Energy with 59 documents, European Commission with 55 documents, Engineering and Physical Sciences Research Council with 42 documents, Natural Sciences and Engineering Research Council of Canada with 40 documents, Deutsche Forschungsgemeinschaft with 33 documents, Horizon 2020 Framework Programme with 28 documents, U.S. National Library of Medicine with 25 documents, Japan Society for the Promotion of Science with 24 documents, and Air Force Office of Scientific Research, Economic and Social Research Council, European Regional Development Fund with 22 documents.

J. *Map of Publication Theme*

Construction on the data science keyword network for the publication theme map was built with the VOSViewer application. The criterion for a minimum number of documents related to keywords was thirty repetitions. Thus, from 25,677 keywords, 237 keywords met the thresholds.

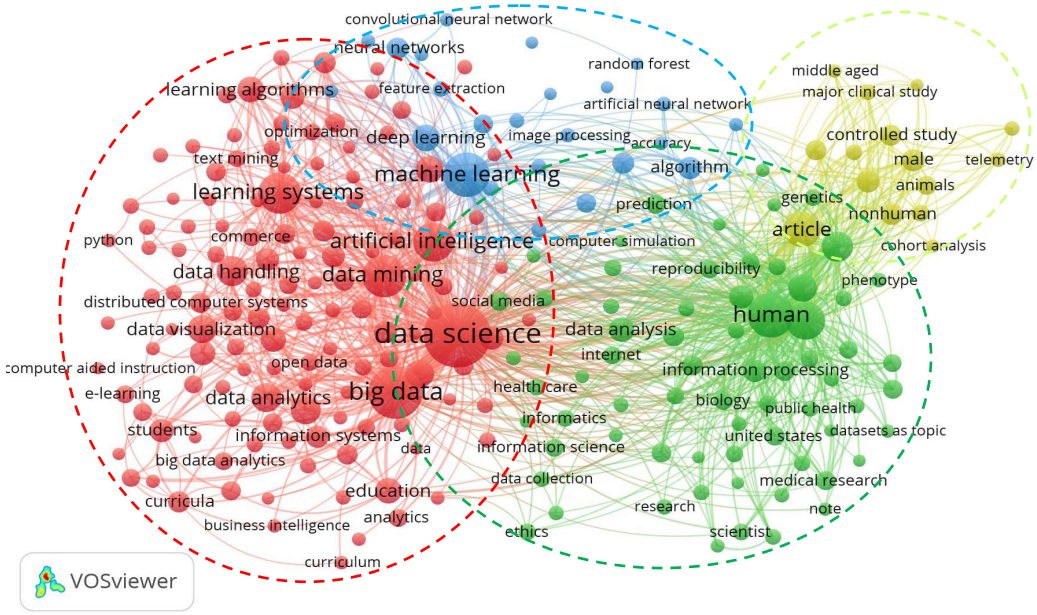

Fig. 9. Keyword Network

Fig. 9. Shows that there were four groups of publication themes based on research keywords related to data science publication, abbreviated as MODD themes.

1. Machine learning (blue). This cluster dominated by the keywords machine learning, deep learning, algorithm, neural networks, and image processing. Most of these keywords relate to machine learning themes.
2. Organism cluster (yellow). In this cluster, we can find organism themes. This cluster was related by the keywords nonhuman, animals, animal experiment, female, male, telemetry, and controlled study.
3. Data mining cluster (red). In this cluster, we can find data mining themes. This cluster was related by the keywords data mining, data, big data, data handling, artificial intelligence, learning system, digital storage, and data visualization.
4. Data analysis cluster (green). This cluster dominated by the keyword data analysis, data processing, information processing, computer simulation, human, internet, informatics, statistical model, data quality, and data set. Most of these keywords relate to data analysis themes.

K. *Authorship Network Map*

The criteria for the minimum number of documents per author were seven documents. Thus, from 14,125 authors, 102 authors were found who met the thresholds. There were ten group collaboration network between researchers in the data science publications as seen in Fig. 10. Authorship Network Map.

1. Red cluster: They domined from United States affiliation which contains Zhao, Z., Zhang, X., Zhang, H., Huang, Y.;
2. Green cluster: They domined with Medical major researcher which contains Li, Z., Zhang, Y., Wang, C., Chen, L.;
3. Yellow cluster: They domined from USA affiliation which contains Li, L., Xu, H., Hripcsak, G., Ryan, P.B., Dumontier, M., Huser, V.;
4. Purple cluster: They domined from United States affiliation awhich contains Ryan, P.B., Hripcsak, G., Xu, H., Chen, J., Li, L.;
5. Brown cluster: They domined from Canada affiliation which contains Pazdor, A.G.M., Jiang, F., Vuzzocrea, A., Leung, C.K.
6. Yellow cluster: They domined from Netherlands affiliation which contains Huser, V., Dumontier, M., Ryan, P.B., Hrup.;
7. Pink cluster: They domined from Case Western Reserve University which contains Xu, Y., Sun, J., Bruchman, L.S., French, R.H.;
8. Orange cluster: They domined from University of Michigan which contains Dehmer, M., Emmert-strei, F., Zhu, Y., Chen, Z., Shi, Y.;
9. Blue cluster: They domined from New York which contains Li, X., Xhang, J., Jee, S., Wang, L., Liu, J., Chen, Y., Li, Y., Shekhar, S., Kumar, V., Saltz, J., Shamshurin, I., Saltz, J.S.;
10. Light blue cluster: They domined from United States of America affiliation which contains Wang, J, Bakken, S., Howe, B., Kraska, T.

IV. Conclusion

The results demonstrated that there were maps and visual trends in increasing the number of publications on data science study at the international level. The country that has the largest contribution in making publications in data science studies was the United States with 2,109 documents. The most productive research institution in the publication of the data science studies was Stanford University with 63 documents. The individual researcher with the most publications in the data science study was Leung, C.K with 25 documents. The most document types published are Computer Science with 3,004 documents (31.2%). Most documents per year by the source in international publications in the Data science studies were the Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics with 211 documents. The most sponsoring institution that helped in the study of the publication of Data Science was the National Science Foundation with 288 documents. The highest publication of international academic documents in data science studies was in 2019 with 2000 documents. There were ten collaboration groups on research related to the data science publication.

Future research is to analyze contributions and explain the impact of data science publication based on a combination of data obtained from Scopus and Web of Science.

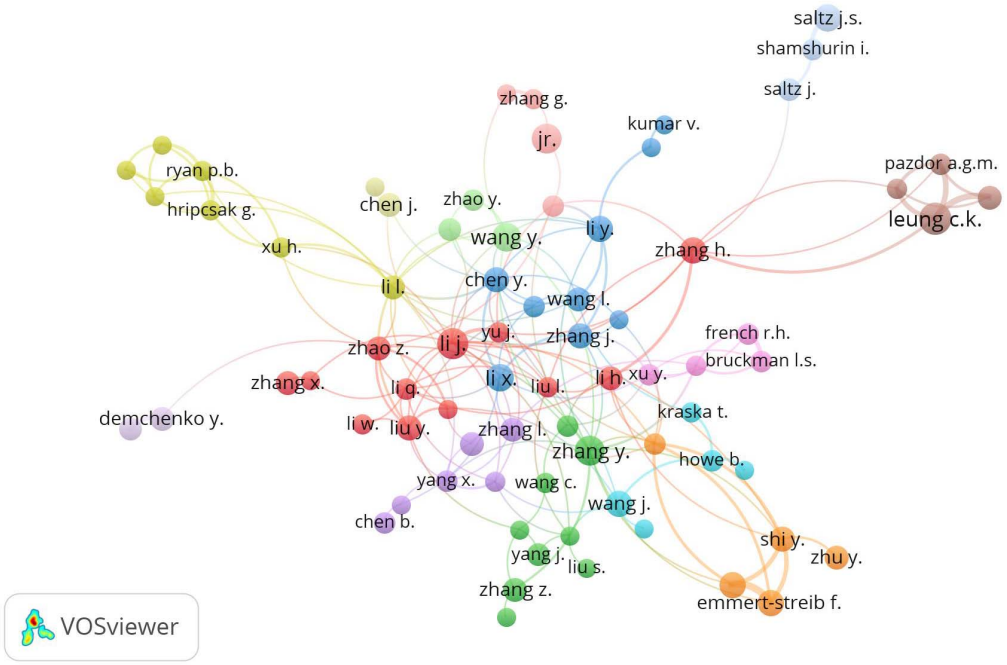

Fig. 10. Authorship Network Map

In terms of contributing implications to knowledge, this research proposes a convergence axis classification consisting of data science studies to characterize the body of knowledge generated from three decades of publication: Machine learning, Organism, Data mining, and Data

analysis, abbreviated as MODD themes. As implications for practical, identifying key themes in the data science sector leads to understanding the development of studies to understand common topics and contexts, as well as the research gaps. With all of this, new studies can be led to address a lack of study and advance knowledge in the areas. The themes most researched also demonstrate the data science contribution to computer science.

REFERENCES

[1] B. Resch and M. Szell, "Human-Centric Data Science for Urban Studies," *Int. J. Geo-Information*, vol. 8, no. 584, pp. 1–9, 2019, doi: 10.3390/ijgi8120584.

[2] A. J. Bevan, "Machine Learning Techniques for Detecting Topological Avatars of New Physics," *Philos. Trans. A Res. Soc.*, vol. 377, no. 20190392, pp. 1–10, 2019, doi: 10.1098/rsta.2019.0392.

[3] L. Feremans, B. Cule, C. Vens, and B. Goethals, "Combining Instance and Feature Neighbours for Extreme Multi-Label Classification," in *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 2017, pp. 109–118, doi: 10.1109/DSAA.2017.70.

[4] J. Zheng, M. Liang, A. Ekstrom, L. Ge, W. Yu, and F. Hsieh, "On Association Study of Scalp EEG Data Channels Under Different Circumstances," in *International Conference on Wireless Algorithms, Systems, and Applications*, 2018, pp. 683–695, doi: 10.1007/978-3-319-94268-1_56.

[5] J. Liu *et al.*, "Applications of Deep Learning to MRI Images: A Survey," *Big Data Min. Anal.*, vol. 1, no. 1, pp. 1–18, 2018, doi: 10.26599/BDMA.2018.9020001.

[6] L. Liu, M. Han, Y. Zhou, and Y. Wang, "LSTM Recurrent Neural Networks for Influenza Trends Prediction," in *Bioinformatics Research and Applications*, 2018, vol. 10847, pp. 259–264, doi: 10.1007/978-3-319-94968-0_25.

[7] H. Han and W. Liu, "The Coming Era Of Artificial Intelligence In Biological Data Science," *BMC Bioinformatics*, vol. 20, no. 22, p. 712, 2019, doi: 10.1186/s12859-019-3225-3.

[8] Data Robot, "Data Science," *DataRobot, Inc*, 2020. .

[9] R. Subramanian and H. Zhang, "Automatic Code Parallelization for Data-Intensive Computing in Multicore Systems," *J. Phys. Conf. Ser.*, vol. 1411, no. 012014, 2019, doi: 10.1088/1742-6596/1411/1/012014.

[10] H. Li *et al.*, "Portfolio Analysis of Research Grants in Data Science Funded by the National Heart, Lung, and Blood Institute," *Circ. Genomic Precis. Med.*, vol. 12, no. e002746, pp. 575–583, 2019, doi: 10.1161/CIRCGEN.119.002746.

[11] K. Gopal, N. R. Salim, and A. F. M. Ayub, "Fuzzy Conjoint Modelin Describing Malaysian Undergraduates' Perceptions of Statistics Classroom Engagement," in *AIP Conference Proceedings 2184*, 2019, p. 030002, doi: 10.1063/1.5136370.

[12] K. Abbasi, "From Data Science to Drinking Water: Love of The Old and New," *J. R. Soc. Med.*, vol. 112, no. 12, p. 491, 2019, doi: 10.1177/0141076819893209.

[13] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. J. Patil, and D. Barton, "Big Data: The Management Revolution," *Harv. Bus. Rev.*, vol. 90, no. 10, pp. 61–67, 2012.

[14] G. Andrienko *et al.*, "(So) Big Data and The Transformation of The City," *Int. J. Data Sci. Anal.*, vol. 10, 2020, doi: 10.1007/s41060-020-00207-3.

[15] H. J. Miller and M. F. Goodchild, "Data-driven Geography," *GeoJournal*, vol. 80, pp. 449–461, 2015, doi: 10.1007/s10708-014-9602-6.

[16] F. Emmert-Streib, M. Dehmer, and O. Yli-Harj, "Ensuring Quality Standards and Reproducible Research for Data Analysis Services in Oncology: A Cooperative Service Model," *Perspective*, vol. 7, no. 349, pp. 1–5, 2019, doi: 10.3389/fcell.2019.00349.

[17] S. Ashetty and L. Lahti, "Microbiome Data Science," *J. Biosci.*, vol. 44, no. 115, pp. 1–6, 2019, doi: 10.1007/s12038-019-9930-2.

[18] A. Purnomo, "Data Science Publication (1983-2019)," *Mendeley Data*, 2020. [Online]. Available: https://data.mendeley.com/datasets/4c3mpmwk74/1.

[19] N. J. van Eck and L. Waltman, "Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, 2010.

[20] I. Setyawati, A. Purnomo, D. E. Irawan, M. Tamyiz, and D. U. Sutiksno, "A Visual Trend of Literature on Ecopreneurship Research Overviewed within the Last Two Decades," *J. Entrep. Educ.*, vol. 21, no. 4, 2018.

[21] B. Ranjbar-Sahraei and R. R. Negenborn, *Research Positioning & Trend Identification: a Data-Analytics Toolbox*, Version 2. Walanda: Delft University of Technology, 2017.