EDITORS
NORITA AHMAD American University of Sharjah; nahmad@aus.edu
PREETI CHAUHAN VP Technical Activities, IEEE Reliability Society; preeti.chauhan@ieee.org

Data Science: Hype and Reality

Norita Ahmad, Areeba Hamid, and Vian Ahmed, American University of Sharjah

Data science is considered a young field by many. This column shares the growing trends of data science as one of the most sought-after career options and as an emerging discipline in almost every industry in the world.

In 2012, the *Harvard Business Review* caused a stir by calling data scientist "the sexiest job of the 21st century."[1] The denomination "data scientist" refers to a profession that makes sense of the vast amount of big data. However, scientists, statisticians, computer scientists, librarians, and other professions have been analyzing and "making sense" of data for ages. As such, the term "data science" is not new and can be traced back to 1962 when John W. Tukey published a book titled *The Future of Data Analysis*.[2] In his book, Tukey, one of the most influential statisticians of the 20th century, suggested that statistics is "pure mathematics," but data analysis is "intrinsically an empirical science," and, therefore, it should take the characteristics of science rather than mathematics.[2] As such, Tukey acknowledged that the two are related but are separate disciplines, and to make progress in data analysis, it is important to focus on the tools and attitudes.

Today we're witnessing explosive growth in the field. This increase can be attributed to the growing amount of data generated by digital activities in our lives. According to the International Data Corporation, more than 59 zettabytes of data were captured in 2020, and the number is expected to increase with a five-year compound annual growth rate of 26%.[3] Data science enables companies not only to understand data from multiple sources but also to enhance decision making. As a result, data science is widely used in almost every industry, including health care, finance, marketing, banking, city planning, and more. With advances in technology, new approaches to the field, and people's positive attitudes toward data science, there is every reason to believe that the field will continue to grow in the future.

WHAT IS DATA SCIENCE?

The term *data science* is so widely used today that its definition has become blurry. Some associate it with computer science and some with statistics; most frequently, it is linked to machine learning (ML) and data mining.[4] ML deals with algorithms for extracting patterns from data, while data mining pertains to the analysis of structured

Digital Object Identifier 10.1109/MC.2021.3130365
Date of current version: 14 February 2022

data. Data science takes both of these tasks into account in addition to other data-related tasks such as capturing, cleaning, and transforming unstructured data; the use of big data technologies; and handling data ethics and regulation.[4] Besides these engineering-oriented fields, data science can also be business oriented in the form of business intelligence and analytics. Moreover, the data science umbrella consists of back-end and front-end data science.[5] The back-end part is often referred to as *data engineering* and deals with hardware, computing, and data storage infrastructure. The front end focuses more on data analysis and ML.

> Tukey acknowledged that the two are related but are separate disciplines, and to make progress in data analysis, it is important to focus on the tools and attitudes.

Perhaps a better way of describing this vast topic is by describing what it is not.[6] Data science is not all about using data for prediction or merely about data analysis. It is not a discipline confined to science, technology, engineering, and mathematics fields, and, most importantly, it is not a single discipline at all. Like other sciences, it is best understood as a "collection of disciplines with complementary foundations, perspectives, approaches, and aims, but with a shared grand mission," which is to use information and technologies to advance human society.[6]

THE EVOLUTION OF DATA SCIENCE

Tukey's work *The Future of Data Analysis* is known as a seminal moment in the history of data science. Tukey's impact on data science is immense. Besides Tukey, there are a few other prominent names and events that are worth examining.

For example, in 1974, Peter Naur, a well-known Danish computer scientist, published a book that was based on a survey of contemporary data processing methods in Sweden and the United States. Here, he defined data science as "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."[7] Following the success of his 1962 work, Tukey published a book in 1977 titled *Exploratory Data Analysis*, another important milestone in the field.[8] In the same year, the International Association for Statistical Computing was established as part of the International Statistical Institute, linking traditional statistical methods with computer technology in converting data into information and knowledge.

In 1996, the term data science was included for the first time in the title of the biennial conference of the International Federation of Classification Societies.[9] A year later, in his inaugural lecture for the H.C. Carver Chair in Statistics at the University of Michigan, Prof. Jeff Wu called for statistics to be renamed data science and the statisticians to be called data scientists.[10] In 2001, Prof. William Cleveland, a computer scientist and professor of statistics at Purdue University, published an article titled "Data Science: An Action Plan for Expanding the Technical Areas of the Fields of Statistics." Just like Tukey, Cleveland was promoting the idea of merging various fields such as computer science and statistics.[11]

One of the most significant events in the field of data science was the launch of the *Data Science Journal* by the Committee on Data for Science and Technology of the International Council for Science in 2002.[12] The journal serves as a platform for everyone interested in data to present their work and exchange ideas. In 2005, the National Science Board published a report titled "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," which defines data scientists as "the information and computer scientists, database and software engineers and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection."[13]

In January 2009, in an interview with *McKinsey Quarterly*, Hal Varian, Google's chief economist, said that the occupation of statistician would be a sexy job in the next 10 years.[14] Then, in 2011, D.J. Patil, a renowned mathematician and computer scientist, wrote an article, "Building Data Science Teams," which many perceived to be the start of data science as the distinct professional specification known as data scientist today.[15] From there on, data science has been covered at length and started to gain more attention as businesses recognized the importance of data.

THE EVOLUTION OF THE DATA SCIENTIST

We can see that data science is not an entirely new field. It was derived from already existing topics. Statistics and computer science are the most closely related to what is now called *data science*. In fact, computer science is the top related query for data science on Google Trends as of 2021. Over the past five years, the popularity of both fields has also risen similarly (see Figure 1). Computer science is also most relevant for data science software experts and developers. In early data science academic programs, most higher education institutions offered data science programs using existing courses with faculty from academic areas such as computer science and statistics, with

some shifts in teaching and professional development.[5,16]

However, with the growing popularity and diversity of data science, institutions have created dedicated data science programs. In addition, many other fields also infuse data science across their disciplines to introduce data literacy and data analytics as important skills among their students.[4,5] One such discipline is business. The science of business statistics was virtually nonexistent until Silicon Valley tech giants started using big data.[17] The field of business analytics and intelligence has vastly grown ever since businesses started using data in the form of *insights data science* for strategic decision making, *product data science* for product testing and optimization, and *engineering data science* for business. Engineering data science includes ML engineers, data engineers, and analytical data engineers. These engineers are differentiated from other data scientists because they produce, rather than synthesize, data.[18]

DATA SCIENCE EDUCATION AND CAREERS

Upon analyzing data from 465 U.S. data science programs,[19] the following conclusions can be drawn. Master's programs, at 63.6%, are the most common higher education programs for data science (see Figure 2). This conclusion also coincides with worldwide Google Trends for data science programs (see Figure 3), where a master's degree program is more popular than any other type. Similarly, results from datascienceprograms.org, which analyzed more than 500 universities in the United States, also found master's degrees to be the most popular in the field.[20] The high demand for master's programs could be credited to the need for skills that other programs lack, the fact that master's students already have relevant experience, and the growing popularity of data science professionals within the areas of business.[5,21]

Popularity within the business field can be seen through the high number of business intelligence and analytics programs offered, which is the most common program category in Rawlings-Goss's data (21.9%), as well as master of business administration programs (7.1%), most of which also have concentrations in business analytics (see Figure 4).[19]

Another emerging field in data science education, as seen through these data, is health-care/biomedical informatics and information management. This is one of the fastest growing fields in the health-care sector today, requiring the acquisition, storage, retrieval, and use of medical data.[22] The data of Rawlings-Goss also show that fields such

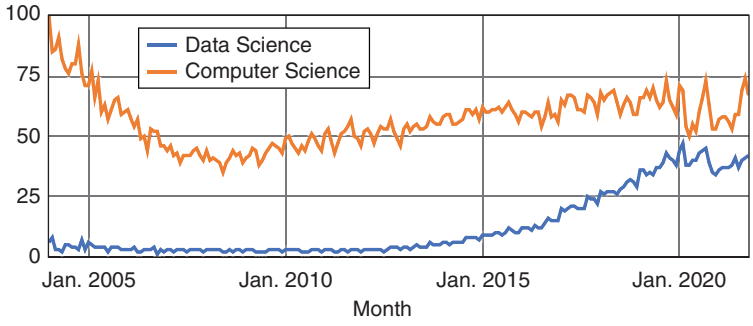

FIGURE 1. Google Trends for worldwide interest in data science versus computer science over time as of 29 October 2021. 100 designates the peak popularity of the term.

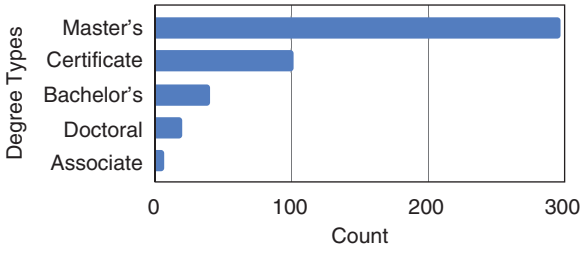

FIGURE 2. Counts of data science program degrees.

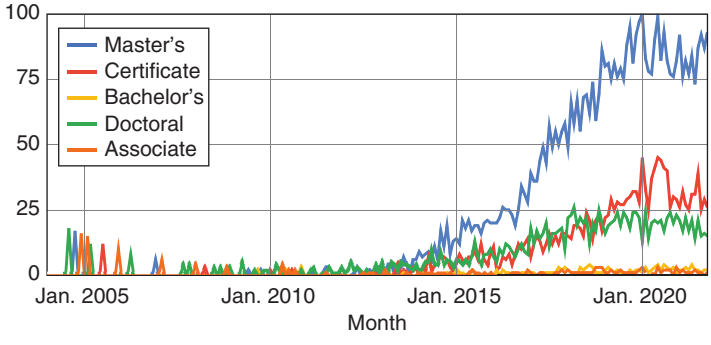

FIGURE 3. Google Trends for worldwide interest in data science programs over time as of 28 October 2021. 100 designates the peak popularity of the term.

as computer science and statistics, from which data science heavily derives, continue to offer data science programs as well.[19]

Degrees in data science often target areas such as data wrangling and mining, data visualization, ML, programming (with R and Python being the most popular languages), and probability and statistics, whether offered as courses/certifications or dedicated degrees in these topic areas.[20] It should also be noted that, even though a variety of data science degrees is offered, there is no single organization that accredits data science programs. However, some organizations accredit related program areas such as the Accreditation Board for Engineering and Technology for computer technology, analysis, and engineering and the Association to Advance Collegiate Schools of Business for business analytics.[20] The top five universities for data science by region, based on research performance in data science from 14,160 universities in 183 countries, are listed in Table 1.[23,24]

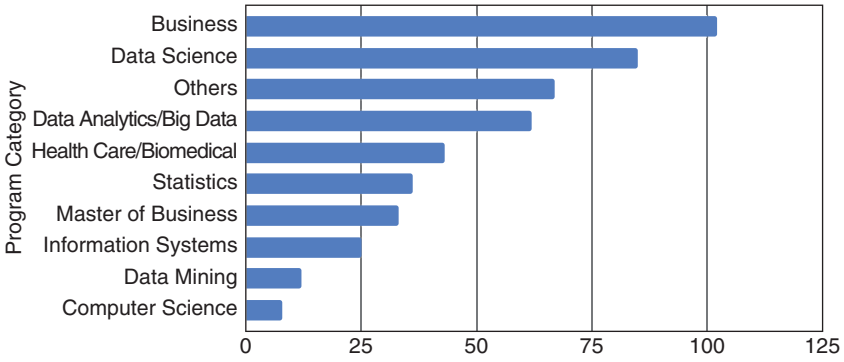

FIGURE 4. Counts of categories of programs offered in data science in the United States.

TABLE 1. The world's top undergraduate and graduate programs for data science, according to EduRank.[23,24]

Region	Top Five Universities
North America	1. Stanford University, United States 2. Harvard University, United States 3. Massachusetts Institute of Technology, United States 4. University of Illinois at Urbana-Champaign, United States 5. University of Washington, Seattle, United States
Latin America	1. University of Sao Paulo, Brazil 2. Federal University of Minas Gerais, Brazil 3. National Polytechnic Institute, Mexico 4. State University of Campinas, Brazil 5. National Autonomous University of Mexico, Mexico
Europe	1. University of Oxford, United Kingdom 2. University College London, United Kingdom 3. University of Cambridge, United Kingdom 4. Imperial College London, United Kingdom 5. University of Manchester, United Kingdom
Asia	1. Tsinghua University, China 2. National University of Singapore, Singapore 3. Nanyang Technological University, Singapore 4. Peking University, China 5. Wuhan University, China
Africa	1. University of Pretoria, South Africa 2. Ain Shams University, Egypt 3. University of Stellenbosch, South Africa 4. University of Cape Town, South Africa 5. Mohammed V University, Morocco
Oceania	1. University of Melbourne, Australia 2. University of New South Wales, Australia 3. University of Sydney, Australia 4. University of Technology Sydney, Australia 5. University of Queensland, Australia

There is a variety of educational approaches to data science, but many argue that theoretical education is not enough. To become a successful data scientist, one requires not only a knowledge of theory and tools but also sufficient "real-world" experience to know how to reach, and when to trust, results.[25]

In general, data science jobs can be divided into either data generalists or specialists (see Figure 5 for details). Data generalists are commonly referred to as *data scientists* because they are broadly trained and "wear many hats" in the field of data science. While many organizations still look for generalists, the trend toward data specialists is now increasing as companies use data science more specifically.[19] LinkedIn's "Jobs on the Rise" reports from different regions around the world indicate that data analytics professionals, data analysts, and artificial intelligence (AI) professionals are the most in-demand roles.[26] Figure 5 also shows countries that listed each type of data science role among their top jobs of 2021. By 2026, it was reported that 11.5 million new data science jobs will be created in the United States alone.[27] The trends indicate that jobs are expected to continue to grow in data science.

THE FUTURE OF DATA SCIENCE AND DATA SCIENTISTS

According to the U.S. Bureau of Labor Statistics, the average annual salary for a data scientist in 2021 was US$103,930 in the United States.[27] A 2020 Burtch Works study also shows that the median salary for data scientists is US$95,000 at entry level, US$130,000 at midlevel (US$195,000 for managerial positions), and US$165,000 for experienced data scientists (US$250,000 for managerial positions).[28] Glassdoor also reports the average salary of a data scientist to be US$117,212. This is higher than the average pay of US$74,239 for programmers or US$88,989 for statisticians, for example.[29–31] The high salary amount is attributed to the supply of data professionals still not catching up with the demand. Data scientists also require a high level of expertise and often acquire advanced degrees, as noted earlier with the popularity of master's degrees.

Given the hype around data science, the reality is that most companies still fail to use much of the data they collect and store during business activities. Gartner defines these unused data as "dark data."[32] According to Splunk's "State of Dark Data Report," 55% of an organization's data are dark.[33] Yet, for a Fortune 1000 company, just a 10% increase in data accessibility will lead to more than US$65 million additional net income.[34] The importance of understanding and utilizing data and the science behind it is becoming increasingly visible. Big data are now not an asset exclusive to big companies. The trend is heading to the "democratization of data," which essentially means the involvement of everyone, data scientist or not.[35] This will eventually lead to many trends in the future. An example includes the use of "small data" for analysis of the most vital data in situations where time, bandwidth, and energy expenditure are limited.[36] A closely linked concept is TinyML, which refers to ML algorithms that are designed to take little space and run on low-powered hardware, in embedded systems/the Internet of Things (IoT). Data democratization will also be driven by AutoML, which will make ML accessible to everyone.[36]

Given the hype around data science, the reality is that most companies still fail to use much of the data they collect and store during business activities.

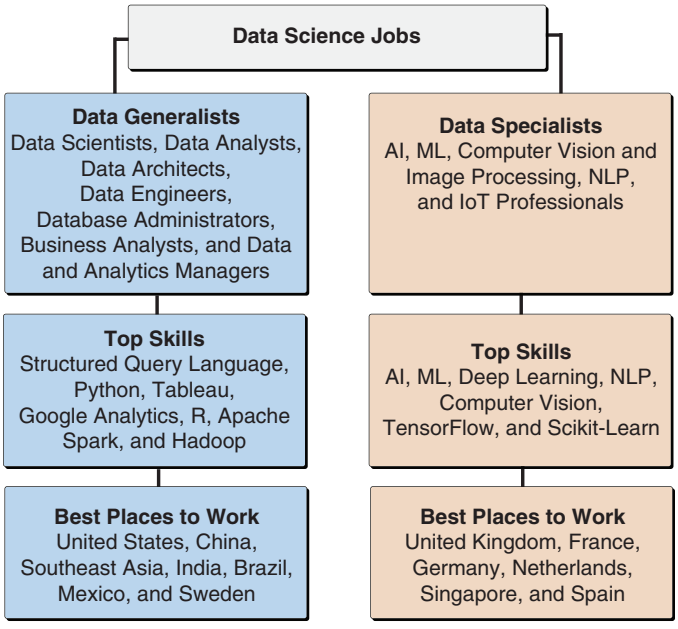

FIGURE 5. Common job titles, top skills/areas of expertise, and best countries to work as data professionals. AI: artificial intelligence; NLP: natural language processing; IoT: Internet of Things.

Another emerging trend will be data-driven customer experience. There has been an increase in investment and innovation in online retail technology in recent years.[35] Interactions with businesses are also becoming more digital. So, it is realistic to expect less hassle in e-commerce, more user-friendly interfaces and front ends, quicker and smoother customer service, and greater levels of personalization in goods and services. Moreover, deep fakes, generative AI, and synthetic data are also expected to rise beyond the arts and entertainment industries.[36] Last, an increasing amount of data science will take place at the intersection of transformative technologies such as AI, the IoT, cloud computing, and superfast networks like 5G. These technologies will make new types of data transfer commonplace and increase automation to create smart homes, factories, and cities.[36] With data science's footprint in practically every aspect of our everyday lives, there is always something new to learn. Technological change is perceived as fast, but data science, in particular, has been rising vigorously.

ACKNOWLEDGMENTS

We would like to thank Prof. Keith Miller and Prof. Joanna DeFranco for their valuable comments on previous versions.

REFERENCES

1. T. H. Davenport and D. J. Patil, "Data scientist: The sexiest job of the 21st century," *Harvard Bus. Rev.*, vol. 90, no. 10, pp. 70–76, 2012.
2. J. W. Tukey, "The future of data analysis," *Ann. Math. Statist.*, vol. 33, no. 1, pp. 1–67, 1962, doi: 10.1214/aoms/1177704711.
3. M. Shirer and J. Rydning, "IDC's global datasphere forecast shows continued steady growth in the creation and consumption of data," IDC, Needham, MA, USA, May 8, 2020. [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=prUS46286020
4. J. D. Kelleher and B. Tierney, *Data Science*. Cambridge, MA, USA: MIT Press, 2018.
5. R. A. Irizarry, "The role of academia in data science education," *Harvard Data Sci. Rev.*, vol. 2, no. 1, 2020, doi: 10.1162/99608f92.dd363929.
6. X.-L. Meng, "Data science: An artificial ecosystem," *Harvard Data Sci. Rev.*, vol. 1, no. 1, 2019, doi: 10.1162/99608f92.ba20f892.
7. P. Naur, *Concise Survey of Computer Methods*. New York, NY, USA: Petrocelli Books, 1974.
8. J. W. Tukey, *Exploratory Data Analysis*. vol. 2, 1977, pp. 131–160.
9. C. Weihs and K. Ickstadt, "Data science: The impact of statistics," *Int. J. Data Sci. Anal.*, vol. 6, no. 3, pp. 189–194, 2018, doi: 10.1007/s41060-018-0102-5.
10. D. Donoho, "50 years of data science," *J. Comput. Graphical Statist.*, vol. 26, no. 4, pp. 745–766, 2017, doi: 10.1080/10618600.2017.1384734.
11. W. S. Cleveland, "Data science: An action plan for expanding the technical areas of the field of statistics," *Int. Statist. Rev.*, vol. 69, no. 1, pp. 21–26, 2001. doi: 10.1111/j.1751-5823.2001.tb00477.x.
12. "About focus and scope," in *Data Science Journal*. London, U.K.: Ubiquity Press, 2021. [Online]. Available: https://datascience.codata.org/about/
13. D. Simberloff *et al.*, "NSB-05-40, long-lived digital data collections enabling research and education in the 21st century," National Science Foundation, Alexandria, VA, USA, 2005. [Online]. Available: https://www.nsf.gov/pubs/2005/nsb0540/
14. "Hal Varian on how the web challenges managers," McKinsey & Company, New York, NY, USA, 2009. [Online]. Available: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers
15. D. J. Patil, *Building Data Science Teams*. Sebastopol, CA, USA: O'Reilly Media, Inc, 2011.
16. B. P. Ricca, B. E. Blaine, K. Donovan, and A. Geraci, "Changing paradigms: Faculty moving to data science from other disciplines," Mathematical and Computing Sciences Faculty/Staff Publications, May 17, 2019. [Online]. Available: https://fisherpub.sjfc.edu/math_facpub/20/
17. T. H. Davenport and J. G. Harris, *Competing on Analytics: Updated, with a New Introduction: The New Science of Winning*. Cambridge, MA, USA: Harvard Univ. Press, 2017.
18. G. Silvera, "Insights vs product vs engineering data science, and how each provides value to your business," Linkedin, 2021. [Online]. Available: https://www.linkedin.com/pulse/insights-vs-product-engineering-data-science-how-each-gordon-silvera/
19. R. Rawlings-Goss, *Data Science Careers, Training, and Hiring*. Cham, Switzerland: Springer International Publishing, 2019.
20. "Guide for students looking for a degree in data science," Data Science Programs, 2021. [Online]. Available: https://www.datascienceprograms.org/
21. "Why employers are looking for applicants with a master's in data science," Cabrini University, Radnor, PA, USA. Accessed: Nov. 15, 2021. [Online]. Available: https://www.cabrini.edu/graduate-degrees/programs/data-science/why-employers-are-looking-for-applicants-with-a-masters-in-data-science
22. T. Stobierski, "What is health informatics?" Northeastern Univ., Boston, MA, USA, 2021. [Online]. Available: https://www.northeastern.edu/graduate/blog/what-is-health-informatics/
23. "World's 100+ best data science universities [2021 Rankings]," EduRank, 2021. [Online]. Available: https://edurank.org/cs/data-science/
24. "EduRank's university ranking methodology," EduRank. Accessed: Nov. 2, 2021. [Online]. Available: https://edurank.org/methodology/
25. M. R. Berthold, "What does it take to be a successful data scientist?" *Harvard Data Sci. Rev.*, vol. 1, no. 2, 2019, doi: 10.1162/99608f92.e0eaabfc.
26. "Jobs on the rise reports: The fastest-growing jobs in the world," Linkedin. Accessed: Nov. 2, 2021. [Online]. Available: https://business.linkedin.com/talent-solutions/emerging-jobs-report#all
27. "Occupational employment and wage statistics," U.S. Bureau of Labor Statistics, Washington, DC, USA, 2021. [Online]. Available: https://www.bls.gov/oes/current/oes152098.htm
28. L. Burtch, "The Burtch works study salaries of data scientists & predictive analytics professionals," Burtch Works, 2020. [Online]. Available: https://www.burtchworks.com/wp-content/uploads/2020/08/Burtch-Works-Study_DS-PAP-2020.pdf
29. "Data scientist salaries," Glassdoor, 2021. [Online]. Available: https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm
30. "Programmer salaries," Glassdoor, 2021. [Online]. Available: https://www.glassdoor.com/Salaries/programmer-salary-SRCH_KO0,10.htm
31. "Statistician salaries," Glassdoor, 2021. [Online]. Available: https://

www.glassdoor.com/Salaries/statistician-salary-SRCH_KO0,12.htm
32. "Dark data," Gartner. Accessed: Nov 15, 2021. [Online]. Available: https://www.gartner.com/en/information-technology/glossary/dark-data
33. "Press release: Dark data research reveals widespread complacency in driving business results and career growth," Splunk, Apr. 30, 2019. [Online]. Available: https://www.splunk.com/en_us/newsroom/press-releases/2019/dark-data-research-reveals-widespread-complacency-in-driving-business-results-and-career-growth.html
34. L. Myler, "Better data quality equals higher marketing ROI," *Forbes*, 2017. [Online]. Available: https://www.forbes.com/sites/larrymyler/2017/07/11/better-data-quality-equals-higher-marketing-roi/?sh=750bc78c7b68
35. B. Marr, "What is data democratisation and why it is a business game-changer?" Bernard Marr & Co., 2021. [Online]. Available: https://bernardmarr.com/what-is-data-democratisation-and-why-it-is-a-business-game-changer/
36. B. Marr, "The 5 biggest data science trends in 2022," *Forbes*, 2021. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2021/10/04/the-5-biggest-data-science-trends-in-2022/?sh=28d8692440d3

NORITA AHMAD is a professor of management information systems at the American University of Sharjah, United Arab Emirates. She is a Member of IEEE. Contact her at nahmad@aus.edu.

AREEBA HAMID is a master of business administration student at the American University of Sharjah, United Arab Emirates. Contact her at g00074541@aus.edu.

VIAN AHMED is a professor of industrial engineering at the American University of Sharjah, United Arab Emirates. Contact her at vahmed@aus.edu.

Digital Object Identifier 10.1109/MC.2022.3142911