WIKIPEDIA

Data science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured,[1][2] similar to data mining.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge.[4][5]

When Harvard Business Review called it "The Sexiest Job of the 21st Century",[6] the term "data science" became a buzzword, and is now often applied to business analytics,[7] business intelligence, predictive modeling, or any arbitrary use of data, or used as a glamorized term for statistics.[8] In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness."[9] While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents.[7] To its discredit, however, many data science and big data projects fail to deliver useful results, often as a result of poor management and utilization of resources.[10][11][12][13]

History

The term "data science" has appeared in various contexts over the past thirty years but did not become an established term until recently. In an early usage it was used as a substitute for computer science by Peter Naur in 1960. Naur later introduced the term "datalogy".[14] In 1974, Naur published *Concise Survey of Computer Methods*, which freely used the term data science in its survey of the contemporary data processing methods that are used in a wide range of applications.

In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference ("Data Science, classification, and related methods"),[15] after the term was introduced in a roundtable discussion by Chikio Hayashi.[3]

In November 1997, C.F. Jeff Wu gave the inaugural lecture entitled "Statistics = Data Science?"[16] for his appointment to the H. C. Carver Professorship at the University of Michigan.[17] In this lecture, he characterized statistical work as a trilogy of data collection, data modeling and analysis, and decision making. In his conclusion, he initiated the modern, non-computer science, usage of the term "data science" and advocated that statistics be renamed data science and statisticians data scientists.[16] Later, he presented his lecture entitled "Statistics = Data Science?" as the first of his 1998 P.C. Mahalanobis Memorial Lectures.[18] These lectures honor Prasanta Chandra Mahalanobis, an Indian scientist and statistician and founder of the Indian Statistical Institute.

In 2001, William S. Cleveland introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data" in his article "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," which was published in Volume 69, No. 1, of the April 2001 edition of the International Statistical Review / Revue Internationale de Statistique.[19] In his report, Cleveland establishes six technical areas which he believed to encompass the field of data science: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory.

In April 2002, the International Council for Science (ICSU): Committee on Data for Science and Technology (CODATA)[20] started the *Data Science Journal*,[21] a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues.[22] Shortly thereafter, in January 2003, Columbia University began publishing *The Journal of Data Science*,[23] which provided a platform for all data workers to present their views and exchange ideas. The journal was largely devoted to the application of statistical methods and quantitative research. In 2005, The National Science Board published "Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century" defining data scientists as "the information and computer scientists, database and software and programmers, disciplinary experts, curators and expert annotators, librarians, archivists, and others, who are crucial to the successful management of a digital data collection" whose primary activity is to "conduct creative inquiry and analysis."[24]

Around 2007, Turing award winner Jim Gray envisioned "data-driven science" as a "fourth paradigm" of science that uses the computational analysis of large data as primary scientific method[4][5] and "to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other."[25]

In the 2012 Harvard Business Review article "Data Scientist: The Sexiest Job of the 21st Century",[6] DJ Patil claims to have coined this term in 2008 with Jeff Hammerbacher to define their jobs at LinkedIn and Facebook, respectively. He asserts that a data scientist is "a new breed", and that a "shortage of data scientists is becoming a serious constraint in some sectors", but describes a much more business-oriented role.

In 2013, the IEEE Task Force on Data Science and Advanced Analytics[26] was launched. In 2013, the first "European Conference on Data Analysis (ECDA)" was organised in Luxembourg, establishing the European Association for Data Science (EuADS) (http://euads.org). The first international conference: IEEE International Conference on Data Science and Advanced Analytics was launched in 2014.[27] In 2014, General Assembly launched student-paid bootcamp and The Data Incubator launched a competitive free data science fellowship.[28] In 2014, the American Statistical Association section on Statistical Learning and Data Mining renamed its journal to "Statistical Analysis and Data Mining: The ASA Data Science Journal" and in 2016 changed its section name to "Statistical Learning and Data Science".[29] In 2015, the International Journal on Data Science and Analytics[30] was launched by Springer to publish original work on data science and big data analytics. In September 2015 the Gesellschaft für Klassifikation (GfKl) (http://www.gfkl.org/welcome/) added to the name of the Society "Data Science Society" at the third ECDA conference at the University of Essex, Colchester, UK.

Relationship to statistics

The popularity of the term "data science" has exploded in business environments and academia, as indicated by a jump in job openings.[31] However, many critical academics and journalists see no distinction between data science and statistics. Writing in Forbes, Gil Press argues that data science is a buzzword without a clear definition and has simply replaced "business analytics" in contexts such as graduate degree programs.[7] In the question-and-answer section of his keynote address

at the Joint Statistical Meetings of American Statistical Association, noted applied statistician Nate Silver said, "I think data-scientist is a sexed up term for a statistician....Statistics is a branch of science. Data scientist is slightly redundant in some way and people shouldn't berate the term statistician."[8] Similarly, in business sector, multiple researchers and analysts state that data scientists alone are far from being sufficient in granting companies a real competitive advantage[32] and consider data scientists as only one of the four greater job families companies require to leverage big data effectively, namely: data analysts, data scientists, big data developers and big data engineers.[33]

On the other hand, responses to criticism are as numerous. In a 2014 Wall Street Journal article, Irving Wladawsky-Berger compares the data science enthusiasm with the dawn of computer science. He argues data science, like any other interdisciplinary field, employs methodologies and practices from across the academia and industry, but then it will morph them into a new discipline. He brings to attention the sharp criticisms computer science, now a well respected academic discipline, had to once face.[34] Likewise, NYU Stern's Vasant Dhar, as do many other academic proponents of data science,[34] argues more specifically in December 2013 that data science is different from the existing practice of data analysis across all disciplines, which focuses only on explaining data sets. Data science seeks actionable and consistent pattern for predictive uses.[1] This practical engineering goal takes data science beyond traditional analytics. Now the data in those disciplines and applied fields that lacked solid theories, like health science and social science, could be sought and utilized to generate powerful predictive models.[1]

In an effort similar to Dhar's, Stanford professor David Donoho, in September 2015, takes the proposition further by rejecting three simplistic and misleading definitions of data science in lieu of criticisms.[35] First, for Donoho, data science does not equate to big data, in that the size of the data set is not a criterion to distinguish data science and statistics.[35] Second, data science is not defined by the computing skills of sorting big data sets, in that these skills are already generally used for analyses across all disciplines.[35] Third, data science is a heavily applied field where academic programs right now do not sufficiently prepare data scientists for the jobs, in that many graduate programs misleadingly advertise their analytics and statistics training as the essence of a data science program.[35][36] As a statistician, Donoho, following many in his field, champions the broadening of learning scope in the form of data science,[35] like John Chambers who urges statisticians to adopt an inclusive concept of learning from data,[37] or like William Cleveland who urges to prioritize extracting from data applicable predictive tools over explanatory theories.[19] Together, these statisticians envision an increasingly inclusive applied field that grows out of traditional statistics and beyond.

For the future of data science, Donoho projects an ever-growing environment for open science where data sets used for academic publications are accessible to all researchers.[35] US National Institute of Health has already announced plans to enhance reproducibility and transparency of research data.[38] Other big journals are likewise following suit.[39][40] This way, the future of data science not only exceeds the boundary of statistical theories in scale and methodology, but data science will revolutionize current academia and research paradigms.[35] As Donoho concludes, "the scope and impact of data science will continue to expand enormously in coming decades as scientific data and data about science itself become ubiquitously available."[35]

References

1. Dhar, V. (2013). "Data science and prediction" (http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext). *Communications of the ACM*. **56** (12): 64. doi:10.1145/2500499 (https://doi.org/10.1145/2500499).

2. Jeff Leek (2013-12-12). "The key word in "Data Science" is not Data, it is Science" (http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/). Simply Statistics.

3. Hayashi, Chikio (1998-01-01). "What is Data Science? Fundamental Concepts and a Heuristic Example" (https://link.springer.com/chapter/10.1007/978-4-431-65950-1_3). In Hayashi, Chikio; Yajima, Keiji; Bock, Hans-Hermann; Ohsumi, Noboru; Tanaka, Yutaka; Baba, Yasumasa. *Data Science, Classification, and Related Methods* (https://www.springer.com/book/9784431702085). Studies in Classification, Data Analysis, and Knowledge Organization. Springer Japan. pp. 40–51. doi:10.1007/978-4-431-65950-1_3 (https://doi.org/10.1007/978-4-431-65950-1_3). ISBN 9784431702085.

4. Stewart Tansley; Kristin Michele Tolle (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery* (https://books.google.com/books?id=oGs_AQAAIAAJ). Microsoft Research. ISBN 978-0-9825442-0-4.

5. Bell, G.; Hey, T.; Szalay, A. (2009). "COMPUTER SCIENCE: Beyond the Data Deluge". *Science*. **323** (5919): 1297–1298. doi:10.1126/science.1170411 (https://doi.org/10.1126/science.1170411). ISSN 0036-8075 (https://www.worldcat.org/issn/0036-8075).

6. Davenport, Thomas H.; Patil, DJ (Oct 2012), *Data Scientist: The Sexiest Job of the 21st Century* (https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/), Harvard Business Review

7. "Data Science: What's The Half-Life Of A Buzzword?" (https://www.forbes.com/sites/gilpress/2013/08/19/data-science-whats-the-half-life-of-a-buzzword/). Forbes. 2013-08-19.

8. "Nate Silver: What I need from statisticians" (http://www.statisticsviews.com/details/feature/5133141/Nate-Silver-What-I-need-from-statisticians.html). 23 Aug 2013.

9. Warden, Pete (2011-05-09). "Why the term "data science" is flawed but useful" (http://radar.oreilly.com/2011/05/data-science-terminology.html). *O'Reilly Radar*. Retrieved 2018-05-20.

10. "Are You Setting Your Data Scientists Up to Fail?" (https://hbr.org/2018/01/are-you-setting-your-data-scientists-up-to-fail). *Harvard Business Review*. 2018-01-25. Retrieved 2018-05-26.

11. "70% of Big Data projects in UK fail to realise full potential" (https://www.consultancy.uk/news/16839/70-of-big-data-projects-in-uk-fail-to-realise-full-potential). *www.consultancy.uk*. Retrieved 2018-05-26.

12. "The Data Economy: Why do so many analytics projects fail? - Analytics Magazine" (http://analytics-magazine.org/the-data-economy-why-do-so-many-analytics-projects-fail/). *Analytics Magazine*. 2014-07-07. Retrieved 2018-05-26.

13. "Data Science: 4 Reasons Why Most Are Failing to Deliver" (https://www.kdnuggets.com/2018/05/data-science-4-reasons-failing-deliver.html). *www.kdnuggets.com*. Retrieved 2018-05-26.

14. Naur, Peter (1 July 1966). "The science of datalogy". *Communications of the ACM*. **9** (7): 485. doi:10.1145/365719.366510 (https://doi.org/10.1145/365719.366510).

15. Press, Gil. "A Very Short History Of Data Science" (https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/).

16. Wu, C. F. J. (1997). "Statistics = Data Science?" (http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf) (PDF). Retrieved 9 October 2014.

17. "Identity of statistics in science examined" (http://ur.umich.edu/9899/Nov09_98/4.htm). The University Records, 9 November 1997, The University of Michigan. Retrieved 12 August 2013.

18. "P.C. Mahalanobis Memorial Lectures, 7th series" (https://web.archive.org/web/20131029191813/http://www.isical.ac.in/~statmath/html/pcm/pcm_recent.html). P.C. Mahalanobis Memorial Lectures, Indian Statistical Institute. Archived from the original (http://www.isical.ac.in/~statmath/html/pcm/pcm_recent.html) on 26 Feb 2017. Retrieved 18 Jul 2017.

19. Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics (https://pdfs.semanticscholar.org/915c/d8e2b39eb02723553913d592b2237d4d9960.pdf). International Statistical Review / Revue Internationale de Statistique, 21–26

20. International Council for Science : Committee on Data for Science and Technology. (2012, April). CODATA, The Committee on Data for Science and Technology. Retrieved from International Council for Science : Committee on Data for Science and Technology: http://www.codata.org/

21. Data Science Journal. (2012, April). Available Volumes. Retrieved from Japan Science and Technology Information Aggregator, Electronic: http://www.jstage.jst.go.jp/browse/dsj/_vols

22. Data Science Journal. (2002, April). Contents of Volume 1, Issue 1, April 2002. Retrieved from Japan Science and Technology Information Aggregator, Electronic: http://www.jstage.jst.go.jp/browse/dsj/1/0/_contents

23. The Journal of Data Science. (2003, January). Contents of Volume 1, Issue 1, January 2003. Retrieved from http://www.jds-online.com/v1-1

24. National Science Board. "Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century" (http://www.nsf.gov/pubs/2005/nsb0540/). National Science Foundation. Retrieved 30 June 2013.

25. Markoff, John (2009-12-14). "Essays Inspired by Microsoft's Jim Gray, Who Saw Science Paradigm Shift" (https://www.nytimes.com/2009/12/15/science/15books.html). *The New York Times*. ISSN 0362-4331 (https://www.worldcat.org/issn/0362-4331). Retrieved 2018-04-26.

26. "IEEE Task Force on Data Science and Advanced Analytics" (http://www.dsaa.co).

27. "2014 IEEE International Conference on Data Science and Advanced Analytics" (http://datamining.it.uts.edu.au/conferences/dsaa14/).

28. "NY gets new bootcamp for data scientists: It's free, but harder to get into than Harvard" (https://venturebeat.com/2014/04/15/ny-gets-new-bootcamp-for-data-scientists-its-free-but-harder-to-get-into-than-harvard/). *Venture Beat*. Retrieved 2016-02-22.

29. Talley, Jill (2016-06-01). "ASA Expands Scope, Outreach to Foster Growth, Collaboration in Data Science" (http://magazine.amstat.org/blog/2016/06/01/datascience-2/). *AMSTATNEWS*. American Statistical Association. Retrieved 2017-02-04.

30. "Journal on Data Science and Analytics" (https://www.springer.com/41060).

31. Darrow, Barb (May 21, 2015). "Data science is still white hot, but nothing lasts forever" (http://fortune.com/2015/05/21/data-science-white-hot/). *Fortune*. Retrieved November 20, 2017.

32. Miller, Steven (2014-04-10). "Collaborative Approaches Needed to Close the Big Data Skills Gap" (http://www.jorgdesign.net/article/view/9823). *Journal of Organization Design*. **3** (1): 26–30. doi:10.7146/jod.9823 (https://doi.org/10.7146/jod.9823). ISSN 2245-408X (https://www.worldcat.org/issn/2245-408X).

33. De Mauro, Andrea; Greco, Marco; Grimaldi, Michele; Ritala, Paavo. "Human resources for Big Data professions: A systematic classification of job roles and required skill sets" (http://linkinghub.elsevier.com/retrieve/pii/S0306457317300018). *Information Processing & Management*. doi:10.1016/j.ipm.2017.05.004 (https://doi.org/10.1016/j.ipm.2017.05.004).

34. Wladawsky-Berger, Irving (May 2, 2014). "Why Do We Need Data Science When We've Had Statistics for Centuries?" (https://blogs.wsj.com/cio/2014/05/02/why-do-we-need-data-science-when-weve-had-statistics-for-centuries/). *The Wall Street Journal*. Retrieved November 20, 2017.

35. Donoho, David (September 2015). "50 Years of Data Science" (http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf) (PDF). *Based on a talk at Tukey Centennial workshop, Princeton NJ Sept 18 2015*.

36. Barlow, Mike (2013). *The Culture of Big Data*. O'Reilly Media, Inc.

37. Chambers, John M. (1993-12-01). "Greater or lesser statistics: a choice for future research" (https://link.springer.com/article/10.1007/BF00141776). *Statistics and Computing*. **3** (4): 182–184. doi:10.1007/BF00141776 (https://doi.org/10.1007/BF00141776). ISSN 0960-3174 (https://www.worldcat.org/issn/0960-3174).

38. Collins, Francis S.; Tabak, Lawrence A. (2014-01-30). "NIH plans to enhance reproducibility" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4058759). *Nature*. **505** (7485): 612–613. doi:10.1038/505612a (https://doi.org/10.1038/505612a). ISSN 0028-0836 (https://www.worldcat.org/issn/0028-0836). PMC 4058759 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4058759). PMID 24482835 (https://www.ncbi.nlm.nih.gov/pubmed/24482835).

39. McNutt, Marcia (2014-01-17). "Reproducibility" (http://science.sciencemag.org/content/343/6168/229). *Science*. **343** (6168): 229–229. doi:10.1126/science.1250475 (https://doi.org/10.1126/science.1250475). ISSN 0036-8075 (https://www.worldcat.org/issn/0036-8075). PMID 24436391 (https://www.ncbi.nlm.nih.gov/pubmed/24436391).

40. Peng, Roger D. (2009-07-01). "Reproducible research and Biostatistics" (https://academic.oup.com/biostatistics/article/10/3/405/293660). *Biostatistics*. **10** (3): 405–408. doi:10.1093/biostatistics/kxp014 (https://doi.org/10.1093/biostatistics/kxp014). ISSN 1465-4644 (https://www.worldcat.org/issn/1465-4644).

Retrieved from "https://en.wikipedia.org/w/index.php?title=Data_science&oldid=850062413"

This page was last edited on 13 July 2018, at 10:14 (UTC).