Data science

Data science is an interdisciplinary academic field[1] that uses statistics, scientific computing, scientific methods, processes, algorithms and systems to extract or extrapolate knowledge and insights from potentially noisy, structured, or unstructured data.[2]

Data science also integrates domain knowledge from the underlying application domain (e.g., natural sciences, information technology, and medicine).[3] Data science is multifaceted and can be described as a science, a research paradigm, a research method, a discipline, a workflow, and a profession.[4]

Data science is "a concept to unify statistics, data analysis, informatics, and their related methods" to "understand and analyze actual phenomena" with data.[5] It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge.[6] However, data science is different from computer science and information 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.[7][8]

A **data scientist** is a professional who creates programming code and combines it with statistical knowledge to create insights from data.[9]

Foundations

Data science is an interdisciplinary field[10] focused on extracting knowledge from typically large data sets and applying the knowledge and insights from that data to solve problems in a wide range of application domains. The field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, and presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science, statistics, information science, mathematics, data visualization, information visualization, data sonification, data integration, graphic design, complex systems, communication and business.[11][12] Statistician Nathan Yau, drawing on Ben Fry, also links data science to human–computer interaction: users should be able to intuitively control and explore data.[13][14] In 2015, the American Statistical Association identified database management, statistics and machine learning, and distributed and parallel systems as the three emerging foundational professional communities.[15]

Relationship to statistics

Many statisticians, including Nate Silver, have argued that data science is not a new field, but rather another name for statistics.[16] Others argue that data science is distinct from statistics because it focuses on problems and techniques unique to digital data.[17] Vasant Dhar writes that statistics emphasizes quantitative data and description. In contrast, data science deals with quantitative and

qualitative data (e.g., from images, text, sensors, transactions, customer information, etc.) and emphasizes prediction and action.[18] Andrew Gelman of Columbia University has described statistics as a non-essential part of data science.[19]

Stanford professor David Donoho writes that data science is not distinguished from statistics by the size of datasets or use of computing and that many graduate programs misleadingly advertise their analytics and statistics training as the essence of a data-science program. He describes data science as an applied field growing out of traditional statistics.[20]

Etymology

Early usage

In 1962, John Tukey described a field he called "data analysis", which resembles modern data science.[20] In 1985, in a lecture given to the Chinese Academy of Sciences in Beijing, C. F. Jeff Wu used the term "data science" for the first time as an alternative name for statistics.[21] Later, attendees at a 1992 statistics symposium at the University of Montpellier II acknowledged the emergence of a new discipline focused on data of various origins and forms, combining established concepts and principles of statistics and data analysis with computing.[22][23]

The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name to computer science.[6] In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic.[6] However, the definition was still in flux. After the 1985 lecture at the Chinese Academy of Sciences in Beijing, in 1997 C. F. Jeff Wu again suggested that statistics should be renamed data science. He reasoned that a new name would help statistics shed inaccurate stereotypes, such as being synonymous with accounting or limited to describing data.[24] In 1998, Hayashi Chikio argued for data science as a new, interdisciplinary concept, with three aspects: data design, collection, and analysis.[23]

During the 1990s, popular terms for the process of finding patterns in datasets (which were increasingly large) included "knowledge discovery" and "data mining".[6][25]

Modern usage

In 2012, technologists Thomas H. Davenport and DJ Patil declared "Data Scientist: The Sexiest Job of the 21st Century",[26] a catchphrase that was picked up even by major-city newspapers like the New York Times[27] and the Boston Globe.[28] A decade later, they reaffirmed it, stating that "the job is more in demand than ever with employers".[29]

The modern conception of data science as an independent discipline is sometimes attributed to William S. Cleveland.[30] In a 2001 paper, he advocated an expansion of statistics beyond theory into technical areas; because this would significantly change the field, it warranted a new name.[25] "Data science" became more widely used in the next few years: in 2002, the Committee on Data for Science and Technology launched the *Data Science Journal*. In 2003, Columbia University launched *The*

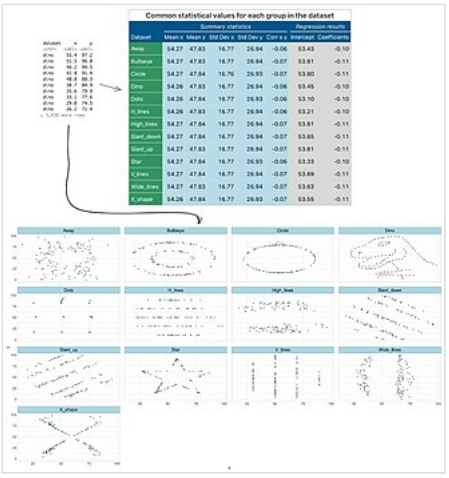
Journal of Data Science.[25] In 2014, the American Statistical Association's Section on Statistical Learning and Data Mining changed its name to the Section on Statistical Learning and Data Science, reflecting the ascendant popularity of data science.[31]

The professional title of "data scientist" has been attributed to DJ Patil and Jeff Hammerbacher in 2008.[32] Though it was used by the National Science Board in their 2005 report "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century", it referred broadly to any key role in managing a digital data collection.[33]

There is still no consensus on the definition of data science, and it is considered by some to be a buzzword.[34] Big data is a related marketing term.[35] Data scientists are responsible for breaking down big data into usable information and creating software and algorithms that help companies and organizations determine optimal operations.[36]

Data science and data analysis

Data science and data analysis are both important disciplines in the field of data management and analysis, but they differ in several key ways. While both fields involve working with data, data science is more of an interdisciplinary field that involves the application of statistical, computational, and machine learning methods to extract insights from data and make predictions, while data analysis is more focused on the examination and interpretation of data to identify patterns and trends.[37][38]



Example for the usefulness of exploratory data analysis as demonstrated using the Datasaurus dozen data set

Data analysis typically involves working with smaller, structured datasets to answer specific questions or solve specific problems. This can involve tasks such as data cleaning, data visualization, and exploratory data analysis to gain insights into the data and develop hypotheses about relationships between variables. Data analysts typically use statistical methods to test these hypotheses and draw conclusions from the data. For example, a data analyst might analyze sales data to identify trends in customer behavior and make recommendations for marketing strategies.[37]

Data science, on the other hand, is a more complex and iterative process that involves working with larger, more complex datasets that often require advanced computational and statistical methods to analyze. Data scientists often work with unstructured data such as text or images and use machine learning algorithms to build predictive models and make data-driven decisions. In addition to statistical analysis, data science often involves tasks such as data preprocessing, feature engineering, and model selection. For instance, a data scientist might develop a recommendation system for an e-commerce platform by analyzing user behavior patterns and using machine learning algorithms to predict user preferences.[38][39]

While data analysis focuses on extracting insights from existing data, data science goes beyond that by incorporating the development and implementation of predictive models to make informed decisions. Data scientists are often responsible for collecting and cleaning data, selecting appropriate analytical

techniques, and deploying models in real-world scenarios. They work at the intersection of mathematics, computer science, and domain expertise to solve complex problems and uncover hidden patterns in large datasets.[38]

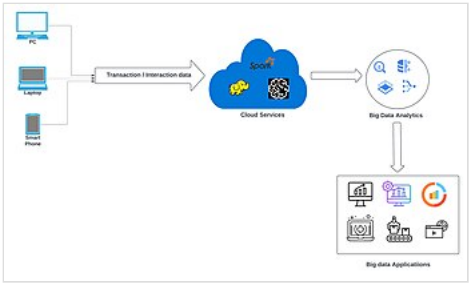
Despite these differences, data science and data analysis are closely related fields and often require similar skill sets. Both fields require a solid foundation in statistics, programming, and data visualization, as well as the ability to communicate findings effectively to both technical and non-technical audiences. Both fields benefit from critical thinking and domain knowledge, as understanding the context and nuances of the data is essential for accurate analysis and modeling.[37][38]

In summary, data analysis and data science are distinct yet interconnected disciplines within the broader field of data management and analysis. Data analysis focuses on extracting insights and drawing conclusions from structured data, while data science involves a more comprehensive approach that combines statistical analysis, computational methods, and machine learning to extract insights, build predictive models, and drive data-driven decision-making. Both fields use data to understand patterns, make informed decisions, and solve complex problems across various domains.

Cloud Computing for Data Science

Cloud computing can offer access to large amounts of computational power and storage.[40] In big data, where volumes of information are continually generated and processed, these platforms can be used to handle complex and resource-intensive analytical tasks.[41]

Some distributed computing frameworks are designed to handle big data workloads. These frameworks can enable data scientists to process and analyze large datasets in parallel, which can reducing processing times.[42]



A cloud-based architecture for enabling big data analytics. Data flows from various sources, such as personal computers, laptops, and smart phones, through cloud services for processing and analysis, finally leading to various big data applications.

Ethical consideration in Data Science

Data science involve collecting, processing, and analyzing data which often including personal and sensitive information. Ethical concerns include potential privacy violations, bias perpetuation, and negative societal impacts [43][44]

Machine learning models can amplify existing biases present in training data, leading to discriminatory or unfair outcomes.[45][46]

See also

- Open Data Science Conference
- Scientific Data

- Women in Data
- Python (programming language)
- R (programming language)
- Data engineering
- Big data
- Machine learning

References

1. Donoho, David (2017). "50 Years of Data Science" (https://doi.org/10.1080%2F10618600.2017.1384734). *Journal of Computational and Graphical Statistics*. **26** (4): 745–766. doi:10.1080/10618600.2017.1384734 (https://doi.org/10.1080%2F10618600.2017.1384734). S2CID 114558008 (https://api.semanticscholar.org/CorpusID:114558008).
2. 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): 64–73. doi:10.1145/2500499 (https://doi.org/10.1145%2F2500499). S2CID 6107147 (https://api.semanticscholar.org/CorpusID:6107147). Archived (https://web.archive.org/web/20141109113411/http://cacm.acm.org/magazines/2013/12/169933-data-science-and-prediction/fulltext) from the original on 9 November 2014. Retrieved 2 September 2015.
3. Danyluk, A.; Leidig, P. (2021). Computing Competencies for Undergraduate Data Science Curricula (https://dstf.acm.org/DSTF_Final_Report.pdf) (PDF). *ACM Data Science Task Force Final Report* (Report).
4. Mike, Koby; Hazzan, Orit (20 January 2023). "What is Data Science?" (https://doi.org/10.1145%2F3575663). *Communications of the ACM*. **66** (2): 12–13. doi:10.1145/3575663 (https://doi.org/10.1145%2F3575663). ISSN 0001-0782 (https://www.worldcat.org/issn/0001-0782).
5. Hayashi, Chikio (1 January 1998). "What is Data Science ? Fundamental Concepts and a Heuristic Example" (https://www.springer.com/book/9784431702085). In Hayashi, Chikio; Yajima, Keiji; Bock, Hans-Hermann; Ohsumi, Noboru; Tanaka, Yutaka; Baba, Yasumasa (eds.). *Data Science, Classification, and Related Methods*. 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%2F978-4-431-65950-1_3). ISBN 9784431702085.
6. Cao, Longbing (29 June 2017). "Data Science: A Comprehensive Overview" (https://doi.org/10.1145%2F3076253). *ACM Computing Surveys*. **50** (3): 43:1–43:42. arXiv:2007.03606 (https://arxiv.org/abs/2007.03606). doi:10.1145/3076253 (https://doi.org/10.1145%2F3076253). ISSN 0360-0300 (https://www.worldcat.org/issn/0360-0300). S2CID 207595944 (https://api.semanticscholar.org/CorpusID:207595944).
7. Tony Hey; 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. Archived (https://web.archive.org/web/20170320193019/https://books.google.com/books?id=oGs_AQAAIAAJ) from the original on 20 March 2017.
8. 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%2Fscience.1170411). ISSN 0036-8075 (https://www.worldcat.org/issn/0036-8075). PMID 19265007 (https://pubmed.ncbi.nlm.nih.gov/19265007). S2CID 9743327 (https://api.semanticscholar.org/CorpusID:9743327).
9. Davenport, Thomas H.; Patil, D. J. (October 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*. **90** (10): 70–76, 128. PMID 23074866 (https://pubmed.ncbi.nlm.nih.gov/23074866). Retrieved 18 January 2016.

10. Emmert-Streib, Frank; Dehmer, Matthias (2018). "Defining data science by a data-driven quantification of the community" (https://doi.org/10.3390%2Fmake1010015). *Machine Learning and Knowledge Extraction*. **1**: 235–251. doi:10.3390/make1010015 (https://doi.org/10.3390%2Fmake1010015).
11. "1. Introduction: What Is Data Science?" (https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html). *Doing Data Science [Book]*. O'Reilly. Retrieved 3 April 2020.
12. "the three sexy skills of data geeks" (https://medriscoll.com/post/4740157098/the-three-sexy-skills-of-data-geeks). *m.e.driscoll: data utopian*. 27 May 2009. Retrieved 3 April 2020.
13. Yau, Nathan (4 June 2009). "Rise of the Data Scientist" (https://flowingdata.com/2009/06/04/rise-of-the-data-scientist/). *FlowingData*. Retrieved 3 April 2020.
14. "Basic Example" (https://benfry.com/phd/dissertation/2.html). *benfry.com*. Retrieved 3 April 2020.
15. "ASA Statement on the Role of Statistics in Data Science" (https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/). *AmStatNews*. American Statistical Association. 1 October 2015. Archived (https://web.archive.org/web/20190620184935/https://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science/) from the original on 20 June 2019. Retrieved 29 May 2019.
16. "Nate Silver: What I need from statisticians" (https://www.statisticsviews.com/article/nate-silver-what-i-need-from-statisticians/). *Statistics Views*. 23 August 2013. Retrieved 3 April 2020.
17. "What's the Difference Between Data Science and Statistics?" (http://priceonomics.com/whats-the-difference-between-data-science-and/). *Priceonomics*. 13 October 2015. Retrieved 3 April 2020.
18. Vasant Dhar (1 December 2013). "Data science and prediction" (http://archive.nyu.edu/handle/2451/31553). *Communications of the ACM*. **56** (12): 64–73. doi:10.1145/2500499 (https://doi.org/10.1145%2F2500499). S2CID 6107147 (https://api.semanticscholar.org/CorpusID:6107147).
19. "Statistics is the least important part of data science « Statistical Modeling, Causal Inference, and Social Science" (https://statmodeling.stat.columbia.edu/2013/11/14/statistics-least-important-part-data-science/). *statmodeling.stat.columbia.edu*. Retrieved 3 April 2020.
20. Donoho, David (18 September 2015). "50 years of Data Science" (http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf) (PDF). Retrieved 2 April 2020.
21. Wu, C. F. Jeff (1986). "Future directions of statistical research in China: a historical perspective" (https://www2.isye.gatech.edu/~jeffwu/publications/fazhan.pdf) (PDF). *Application of Statistics and Management*. **1**: 1–7. Retrieved 29 November 2020.
22. Escoufier, Yves; Hayashi, Chikio; Fichet, Bernard, eds. (1995). *Data science and its applications*. Tokyo: Academic Press/Harcourt Brace. ISBN 0-12-241770-4. OCLC 489990740 (https://www.worldcat.org/oclc/489990740).
23. Murtagh, Fionn; Devlin, Keith (2018). "The Development of Data Science: Implications for Education, Employment, Research, and the Data Revolution for Sustainable Development" (https://doi.org/10.3390%2Fbdcc2020014). *Big Data and Cognitive Computing*. **2** (2): 14. doi:10.3390/bdcc2020014 (https://doi.org/10.3390%2Fbdcc2020014).
24. Wu, C. F. Jeff. "Statistics=Data Science?" (http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf) (PDF). Retrieved 2 April 2020.
25. 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/). *Forbes*. Retrieved 3 April 2020.
26. Davenport, Thomas (1 October 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*. Retrieved 10 October 2022.
27. Miller, Claire (4 April 2013). "Data Science: The Numbers of Our Lives" (https://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html). *New York Times*. New York City. Retrieved 10 October 2022.

28. Borchers, Callum (11 November 2015). "Behind the scenes of the 'sexiest job of the 21st century' " (https://www.bostonglobe.com/business/2015/11/11/behind-scenes-sexiest-job-century/Kc1cvXIu31DfHhVmyRQeIJ/story.html). *Boston Globe*. Boston. Retrieved 10 October 2022.
29. Davenport, Thomas (15 July 2022). "Is Data Scientist Still the Sexiest Job of the 21st Century?" (https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century). *Harvard Business Review*. Retrieved 10 October 2022.
30. Gupta, Shanti (11 December 2015). "William S. Cleveland" (https://www.stat.purdue.edu/~wsc/). Retrieved 2 April 2020.
31. Talley, Jill (1 June 2016). "ASA Expands Scope, Outreach to Foster Growth, Collaboration in Data Science" (https://magazine.amstat.org/blog/2016/06/01/datascience-2/). *Amstat News*. American Statistical Association.. In 2013 the first European Conference on Data Analysis (ECDA2013) started in Luxembourg the process which founded the European Association for Data Science (EuADS) www.euads.org in Luxembourg in 2015.
32. Davenport, Thomas H.; Patil, D. J. (1 October 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*. No. October 2012. ISSN 0017-8012 (https://www.worldcat.org/issn/0017-8012). Retrieved 3 April 2020.
33. "US NSF – NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century" (https://www.nsf.gov/pubs/2005/nsb0540/). *www.nsf.gov*. Retrieved 3 April 2020.
34. Press, Gil. "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*. Retrieved 3 April 2020.
35. Pham, Peter. "The Impacts of Big Data That You May Not Have Heard Of" (https://www.forbes.com/sites/peterpham/2015/08/28/the-impacts-of-big-data-that-you-may-not-have-heard-of/). *Forbes*. Retrieved 3 April 2020.
36. Martin, Sophia (20 September 2019). "How Data Science will Impact Future of Businesses?" (https://static1.squarespace.com/static/5ff2adbe3fe4fe33db902812/t/6009dd9fa7bc363aa822d2c7/1611259312432/ISLR+Seventh+Printing.pdf) (PDF). *Medium*. Retrieved 3 April 2020.
37. Gareth, Hastie; Witten, Tibshira (29 September 2017). "An Introduction to Statistical Learning: with Applications in R." (https://www.researchgate.net/publication/354866394) *Springer*.
38. Provost, Foster; Tom Fawcett (1 August 2013). "Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking" (https://www.researchgate.net/publication/256438799). *O'Reilly Media, Inc*.
39. Han, Kamber; Pei (2011). *Data Mining: Concepts and Techniques* (https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques). ISBN 9780123814791.
40. Hashem, Ibrahim Abaker Targio; Yaqoob, Ibrar; Anuar, Nor Badrul; Mokhtar, Salimah; Gani, Abdullah; Ullah Khan, Samee (2015). "The rise of "big data" on cloud computing: Review and open research issues" (https://linkinghub.elsevier.com/retrieve/pii/S0306437914001288). *Information Systems*. **47**: 98–115. doi:10.1016/j.is.2014.07.006 (https://doi.org/10.1016%2Fj.is.2014.07.006).
41. Qiu, Junfei; Wu, Qihui; Ding, Guoru; Xu, Yuhua; Feng, Shuo (2016). "A survey of machine learning for big data processing" (https://doi.org/10.1186%2Fs13634-016-0355-x). *EURASIP Journal on Advances in Signal Processing*. **2016** (1). doi:10.1186/s13634-016-0355-x (https://doi.org/10.1186%2Fs13634-016-0355-x). ISSN 1687-6180 (https://www.worldcat.org/issn/1687-6180).
42. Armbrust, Michael; Xin, Reynold S.; Lian, Cheng; Huai, Yin; Liu, Davies; Bradley, Joseph K.; Meng, Xiangrui; Kaftan, Tomer; Franklin, Michael J.; Ghodsi, Ali; Zaharia, Matei (27 May 2015). "Spark SQL: Relational Data Processing in Spark" (https://dl.acm.org/doi/10.1145/2723372.2742797). *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM. pp. 1383–1394. doi:10.1145/2723372.2742797 (https://doi.org/10.1145%2F2723372.2742797). ISBN 978-1-4503-2758-9.

43. Floridi, Luciano; Taddeo, Mariarosaria (28 December 2016). "What is data ethics?" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124072). *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. **374** (2083): 20160360. Bibcode:2016RSPTA.37460360F (https://ui.adsabs.harvard.edu/abs/2016RSPTA.37460360F). doi:10.1098/rsta.2016.0360 (https://doi.org/10.1098%2Frsta.2016.0360). ISSN 1364-503X (https://www.worldcat.org/issn/1364-503X). PMC 5124072 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5124072). PMID 28336805 (https://pubmed.ncbi.nlm.nih.gov/28336805).
44. Mittelstadt, Brent Daniel; Floridi, Luciano (2016). "The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts" (http://link.springer.com/10.1007/s11948-015-9652-2). *Science and Engineering Ethics*. **22** (2): 303–341. doi:10.1007/s11948-015-9652-2 (https://doi.org/10.1007%2Fs11948-015-9652-2). ISSN 1353-3452 (https://www.worldcat.org/issn/1353-3452). PMID 26002496 (https://pubmed.ncbi.nlm.nih.gov/26002496).
45. Barocas, Solon; Selbst, Andrew D (2016). "Big Data's Disparate Impact" (https://lawcat.berkeley.edu/record/1127463). *California Law Review*. doi:10.15779/Z38BG31 (https://doi.org/10.15779%2FZ38BG31).
46. Caliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind (14 April 2017). "Semantics derived automatically from language corpora contain human-like biases" (https://www.science.org/doi/10.1126/science.aal4230). *Science*. **356** (6334): 183–186. arXiv:1608.07187 (https://arxiv.org/abs/1608.07187). Bibcode:2017Sci...356..183C (https://ui.adsabs.harvard.edu/abs/2017Sci...356..183C). doi:10.1126/science.aal4230 (https://doi.org/10.1126%2Fscience.aal4230). ISSN 0036-8075 (https://www.worldcat.org/issn/0036-8075).
