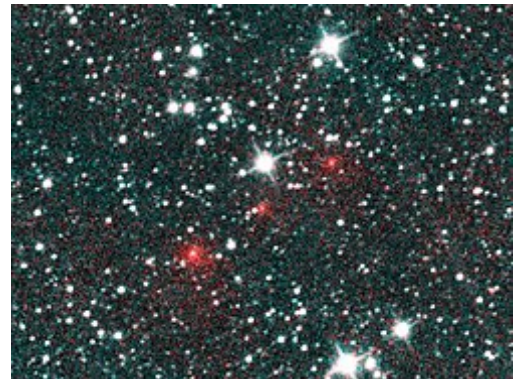


Data science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data,^{[1][2]} and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Data science is a "concept to unify statistics, data analysis, informatics, and their related methods" in order to "understand and analyze actual phenomena" with data.^[3] It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge.^[4] 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.^{[5][6]}

A **data scientist** is someone who creates programming code, and combines it with statistical knowledge to create insights from data.^[7]



The existence of Comet NEOWISE (here depicted as a series of red dots) was discovered by analyzing astronomical survey data acquired by a space telescope, the Wide-field Infrared Survey Explorer.

Contents

Foundations

Relationship to statistics

Etymology

Early usage

Modern usage

Technologies and techniques

See also

References

Foundations

Data science is an interdisciplinary field focused on extracting knowledge from data sets, which are typically large (see big data), and applying the knowledge and actionable insights from data to solve problems in a wide range of application domains.^[8] 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, information visualization, data sonification, data integration, graphic design, complex systems, communication and business.^{[9][10]} 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.^{[11][12]} 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.^[13]

Relationship to statistics

Many statisticians, including Nate Silver, have argued that data science is not a new field, but rather another name for statistics.^[14] Others argue that data science is distinct from statistics because it focuses on problems and techniques unique to digital data.^[15] Vasant Dhar writes that statistics emphasizes quantitative data and description. In contrast, data science deals with quantitative and qualitative data (e.g. images) and emphasizes prediction and action.^[16] Andrew Gelman of Columbia University has described statistics as a nonessential part of data science.^[17] 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.^[18] In summary, data science can be therefore described as an applied branch of statistics.

Etymology

Early usage

In 1962, John Tukey described a field he called "data analysis", which resembles modern data science.^[18] 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.^[19] 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.^{[20][21]}

The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science.^[22] In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic.^[22] However, the definition was still in flux. After the 1985 lecture in 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.^[23] In 1998, Hayashi Chikio argued for data science as a new, interdisciplinary concept, with three aspects: data design, collection, and analysis.^[21]

During the 1990s, popular terms for the process of finding patterns in datasets (which were increasingly large) included "knowledge discovery" and "data mining".^{[24][22]}

Modern usage

The modern conception of data science as an independent discipline is sometimes attributed to William S. Cleveland.^[25] 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.^[24] "Data science" became

more widely used in the next few years: in 2002, the Committee on Data for Science and Technology launched *Data Science Journal*. In 2003, Columbia University launched *The Journal of Data Science*.^[24] 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.^[26]

The professional title of "data scientist" has been attributed to DJ Patil and Jeff Hammerbacher in 2008.^[27] 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.^[28]

There is still no consensus on the definition of data science and it is considered by some to be a buzzword.^[29] Big data is a related marketing term.^[30] 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.^[31]

Technologies and techniques

There is a variety of different technologies and techniques that are used for data science which depend on the application.

- Linear regression
- Logistic regression
- Decision trees are used as prediction models for classification and data fitting. The decision tree structure can be used to generate rules able to classify or predict target/class/label variable based on the observation attributes.
- Support-vector machine (SVM)
- Cluster analysis is a technique used to group data together.
- Dimensionality reduction is used to reduce the complexity of data computation so that it can be performed more quickly.
- Machine learning is a technique used to perform tasks by inferencing patterns from data
- Naive Bayes classifiers are used to classify by applying the Bayes' theorem. They are mainly used in datasets with large amounts of data, and can aptly generate accurate results.

See also

- International Journal of Population Data Science

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): 64–73. 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.

2. Jeff Leek (12 December 2013). "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. Archived (<https://web.archive.org/web/20140102194117/http://simplystatistics.org/2013/12/12/the-key-word-in-data-science-is-not-data-it-is-science/>) from the original on 2 January 2014. Retrieved 1 January 2014.
3. 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.
4. Cao, Longbing (29 June 2017). "Data Science: A Comprehensive Overview" (<https://doi.org/10.1145/3076253>). *ACM Computing Surveys*. **50** (3): 43:1–43:42. doi:10.1145/3076253 (<https://doi.org/10.1145%2F3076253>). ISSN 0360-0300 (<https://www.worldcat.org/issn/0360-0300>).
5. Tony Hey; Stewart Tansley; Kristin Michele Tolle (2009). *The Fourth Paradigm: Data-intensive Scientific Discovery* (https://books.google.com/books?id=oGs_AQAIAAJ). Microsoft Research. ISBN 978-0-9825442-0-4. Archived (https://web.archive.org/web/20170320193019/https://books.google.com/books?id=oGs_AQAIAAJ) from the original on 20 March 2017.
6. 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>).
7. 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): 70–6, 128. PMID 23074866 (<https://pubmed.ncbi.nlm.nih.gov/23074866>). Retrieved 18 January 2016.
8. "About Data Science | Data Science Association" (<http://www.datascienceassn.org/about-data-science>). *www.datascienceassn.org*. Retrieved 3 April 2020.
9. "1. Introduction: What Is Data Science? - Doing Data Science [Book]" (<https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html>). *www.oreilly.com*. Retrieved 3 April 2020.
10. "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.
11. 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.
12. "Basic Example" (<https://benfry.com/phd/dissertation/2.html>). *benfry.com*. Retrieved 3 April 2020.
13. "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.
14. "Nate Silver: What I need from statisticians - Statistics Views" (<https://www.statisticsviews.com/article/nate-silver-what-i-need-from-statisticians/>). *www.statisticsviews.com*. Retrieved 3 April 2020.
15. "What's the Difference Between Data Science and Statistics?" (<http://priceonomics.com/whats-the-difference-between-data-science-and/>). *Priceonomics*. Retrieved 3 April 2020.

16. DharVasant (1 December 2013). "Data science and prediction". *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>).
17. "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.
18. 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.
19. 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.
20. *Data science and its applications = La @science des données et ses applications*. Escoufier, Yves., Hayashi, Chikio (1918-....), Fichet, Bernard. Tokyo: Academic Press/Harcourt Brace. 1995. ISBN 0-12-241770-4. OCLC 489990740 (<https://www.worldcat.org/oclc/489990740>).
21. 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>).
22. CaoLongbing (29 June 2017). "Data Science" (<https://doi.org/10.1145%2F3076253>). *ACM Computing Surveys*. **50** (3): 1–42. arXiv:2007.03606 (<https://arxiv.org/abs/2007.03606>). doi:10.1145/3076253 (<https://doi.org/10.1145%2F3076253>).
23. Wu, C.F. Jeff. "Statistics=Data Science?" (<http://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf>) (PDF). Retrieved 2 April 2020.
24. 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.
25. Gupta, Shanti (11 December 2015). "William S Cleveland" (<https://www.stat.purdue.edu/~ws c/>). Retrieved 2 April 2020.
26. 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.
27. 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.
28. "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.
29. 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.
30. 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.
31. Martin, Sophia (20 September 2019). "How Data Science will Impact Future of Businesses?" (<https://towardsdatascience.com/how-data-science-will-impact-future-of-businesses-7f11f5699c4d>). *Medium*. Retrieved 3 April 2020.

This page was last edited on 2 March 2022, at 03:10 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License 3.0; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.