Data science

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

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

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. It uses techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, information science, and domain knowledge. 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.

A data scientist is the professional who creates programming code and combines it with statistical knowledge to create insights from data.

Foundations

Data science is an interdisciplinary field 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. 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. 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.

Relationship to statistics

Many statisticians, including Nate Silver, have argued that data science is not a new field, but rather another name for statistics. Others argue that data science is distinct from statistics because it focuses on problems and techniques unique to digital data. 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 or customer information, etc.) and emphasizes prediction and action. Andrew Gelman of Columbia University has described statistics as a nonessential part of data science.

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.

Etymology

Early usage

In 1962, John Tukey described a field he called ‘data analysis’, which resembles modern data science. 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. 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.

The term ‘data science’ has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. 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. In 1998, Hayashi Chikio argued for data science as a new, interdisciplinary concept, with three aspects: data design, collection, and analysis.

During the 1990s, popular terms for the process of finding patterns in datasets (which were increasingly large) included ‘knowledge discovery’ and ‘data mining’.

Modern usage

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

The modern conception of data science as an independent discipline is sometimes attributed to William S. Cleveland. 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. ‘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. 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.

The professional title of ‘data scientist’ has been attributed to DJ Patil and Jeff Hammerbacher in 2008. 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.

There is still no consensus on the definition of data science, and it is considered by some to be a buzzword. Big data is a related marketing term. 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.