Data science

Data science is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms and systems to extract or extrapolate knowledge from potentially noisy, structured, or 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" 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 a professional who creates programming code and combines it with statistical knowledge to summarize data.

== Foundations ==
Data science is an interdisciplinary field focused on extracting knowledge from typically large data sets and applying the knowledge from that data to solve problems in other application domains. The field encompasses preparing data for analysis, formulating data science problems, analyzing data, and summarizing these findings. As such, it incorporates skills from computer science, mathematics, data visualization, graphic design, communication, and business.
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. Andrew Gelman of Columbia University has described statistics as a non-essential 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 to 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.

=== Modern usage ===
In 2012, technologists Thomas H. Davenport and DJ Patil declared "Data Scientist: The Sexiest Job of the 21st Century", a catchphrase 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 that "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 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.

== Data science and data analysis ==

In data science, data analysis is the process of inspecting, cleaning, transforming, and modelling data to discover useful information, draw conclusions, and support decision-making. It includes exploratory data analysis (EDA), which uses graphics and descriptive statistics to explore patterns and generate hypotheses, and confirmatory data analysis, which applies statistical inference to test hypotheses and quantify uncertainty.
Typical activities comprise:

data collection and integration;
data cleaning and preparation (handling missing values, outliers, encoding, normalisation);
feature engineering and selection;
visualisation and descriptive statistics;
fitting and evaluating statistical or machine-learning models;
communicating results and ensuring reproducibility (e.g., reports, notebooks, and dashboards).
Lifecycle frameworks such as CRISP-DM describe these steps from business understanding through deployment and monitoring.
Data science involves working with larger 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. Data science often uses statistical analysis, data preprocessing, and supervised learning.

== Cloud computing for data science ==

Cloud computing can offer access to large amounts of computational power and storage. 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.
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 reduce processing times.

== Ethical consideration in data science ==

Data science involves collecting, processing, and analyzing data which often includes personal and sensitive information. Ethical concerns include potential privacy violations, bias perpetuation, and negative societal impacts.
Machine learning models can amplify existing biases present in training data, leading to discriminatory or unfair outcomes.

== See also ==
Python (programming language)
R (programming language)
Data engineering
Big data
Machine learning
Artificial intelligence
Bioinformatics
Astroinformatics
Topological data analysis
List of data science journals
List of open-source data science software
Data science notebook software

== References ==