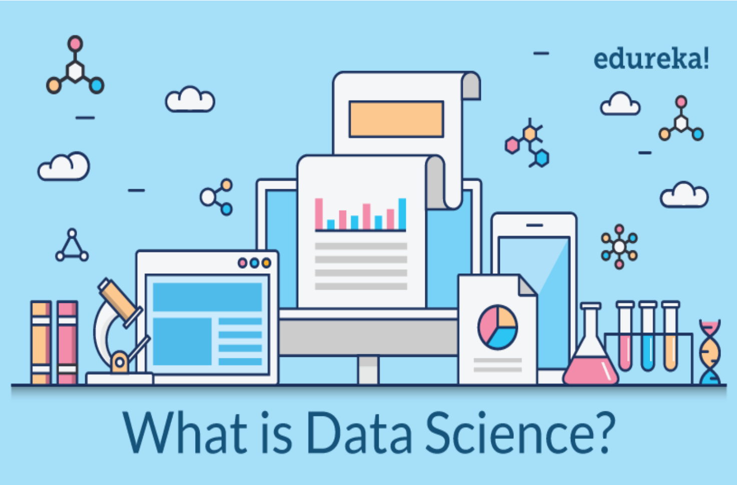
Lesson1. Introduction to Data Science

Data Science Explained



The investopedia.com explained that data science provides meaningful information based on large amounts of complex data or big data. It combines different fields of work in statistics and computation to interpret data for decision-making purposes. The continually increasing access to data is possible due to advancements in technology and collection techniques. Individuals buying patterns and behavior can be monitored and predictions made based on the information gathered.

Edureka.co further established the definition by stating that it is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data. But, according to investopedia.com, the ever-increasing data is unstructured and requires parsing for effective decision making. This process is complex and time-consuming for companies—hence, the emergence of data science.

Edureka.co also explained that data science, or data-driven science, uses big data and machine learning to interpret data for decision-making purposes.

According to oracle.com/ph, those who practice data science are called data scientists, and they combine a range of skills to analyze data collected from the web, smartphones, customers, sensors, and other sources. Furthermore, Oracle mentioned that companies are sitting on a treasure trove of data. As modern technology has enabled the creation and storage of increasing amounts of information, data volumes have exploded. It's estimated that 90 percent of the data in the world was created in the last two years. For example, Facebook users upload 10 million photos every hour.

But this data is often still just sitting in databases and data lakes, mostly untouched.

The wealth of data being collected and stored by these technologies can bring transformative benefits to organizations and societies around the world—but only if we can interpret it. That's where data science comes in.

Data science reveals trends and produces insights that businesses can use to make better decisions and create more innovative products and services. Perhaps most importantly, it enables machine learning (ML) models to learn from the vast amounts of data being fed to them rather than mainly relying upon business analysts to see what they can discover from the data.

Data is the bedrock of innovation, but its value comes from the information data scientists can glean from it, and then act upon.

In order for us to fully understand what data science is, let's first identify the some of the terms that is associated with data science.

- Oracle defined AI means getting a computer to mimic human behavior in some way, while builtin.com defined AI as a wide-ranging branch of computer science concerned with building smart machines capable of performing tasks that typically require human intelligence. AI is an interdisciplinary science with multiple approaches, but advancements in machine learning and deep learning are creating a paradigm shift in virtually every sector of the tech industry.

- Data science is a subset of AI, and it refers more to the overlapping areas of statistics, scientific methods, and data analysis—all of which are used to extract meaning and insights from data.

- Machine learning is another subset of AI, and it consists of the techniques that enable computers to figure things out from the data and deliver AI applications.

Dataversity.net explained that statistics, and the use of statistical models, are deeply rooted within the field of Data Science. Data Science started with statistics, and has evolved to include concepts/practices such as Artificial Intelligence, Machine Learning, and the Internet of Things, to name a few. As more and more data has become available, first by way of recorded shopping behaviors and trends, businesses have been collecting and storing it in ever greater amounts. With growth of the Internet, the Internet of Things, and the exponential growth of data volumes available to enterprises, there has been a flood of new information or Big Data. Once the doors were opened by businesses seeking to increase profits and drive better decision making, the use of Big Data started being applied to other fields, such as medicine, engineering, and social science.

A functional Data Scientist, as opposed to a general statistician, has a good understanding of software architecture and understands multiple programming languages. The Data Scientist defines the problem, identifies the key sources of information, and designs the framework for collecting and screening the needed data. Software is typically responsible for collecting, processing, and modeling the data. They use the principles of Data Science, and all the related sub-fields and practices encompassed within Data Science, to gain deeper insight into the data assets under review.

History of Data Science

Dataversity.net provide details on the history of data science.

There are many different dates and timelines that can be used to trace the slow growth of Data Science and its current impact on the Data Management industry, some of the more significant ones are outlined below.

In 1962, John Tukey wrote about a shift in the world of statistics, saying, "… as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt…I have come to feel that my central interest is in data analysis…" Tukey is referring to the merging of statistics and computers, at a time when statistical results were presented in hours, rather than the days or weeks it would take if done by hand.

In 1974, Peter Naur authored the *Concise Survey of Computer Methods*, using the term "Data Science," repeatedly. Naur presented his own convoluted definition of the new concept: "The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."

In 1977, The IASC, also known as the International Association for Statistical Computing was formed. The first phrase of their mission statement reads, "It is the mission of the IASC to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge."

In 1977, Tukey wrote a second paper, titled *Exploratory Data Analysis*, arguing the importance of using data in selecting "which" hypotheses to test, and that confirmatory data analysis and exploratory data analysis should work hand-in-hand.

In 1989, the Knowledge Discovery in Databases, which would mature into the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, organized its first workshop.

In 1994, Business Week ran the cover story, *Database Marketing*, revealing the ominous news companies had started gathering large amounts of personal information, with plans to start strange new marketing campaigns. The flood of data was, at best, confusing to company managers, who were trying to decide what to do with so much disconnected information.

In 1999, Jacob Zahavi pointed out the need for new tools to handle the massive amounts of information available to businesses, in Mining Data for Nuggets of Knowledge. He wrote:

> "Scalability is a huge issue in data mining… Conventional statistical methods work well with small data sets. Today's databases, however, can involve millions of rows and scores of columns of data… Another technical challenge is developing models that can do a better job analyzing data, detecting non-linear relationships and interaction between elements… Special data mining tools may have to be developed to address web-site decisions."

In 2001, Software-as-a-Service (SaaS) was created. This was the pre-cursor to using Cloud-based applications. Software as a service (or SaaS) is a way of delivering applications over the Internet—as a service. Instead of installing and maintaining software, you simply access it via the Internet, freeing yourself from complex software and hardware management.

In 2001, William S. Cleveland laid out plans for training Data Scientists to meet the needs of the future. He presented an action plan titled, *Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics.* It described how to increase the technical experience and range of data analysts and specified six areas of study for university departments. It promoted developing specific resources for research in each of the six areas. His plan also applies to government and corporate research.

In 2002, the International Council for Science: Committee on Data for Science and Technology began publishing the *Data Science Journal*, a publication focused on issues such as the description of data systems, their publication on the internet, applications and legal issues.

In 2006, Hadoop 0.1.0, an open-source, non-relational database, was released. Hadoop was based on Nutch, another open-source database.

In 2008, the title, "Data Scientist" became a buzzword, and eventually a part of the language. DJ Patil and Jeff Hammerbacher, of LinkedIn and Facebook, are given credit for initiating its use as a buzzword.

In 2009, the term NoSQL was reintroduced (a variation had been used since 1998) by Johan Oskarsson, when he organized a discussion on "open-source, non-relational databases".

In 2011, job listings for Data Scientists increased by 15,000%. There was also an increase in seminars and conferences devoted specifically to Data Science and Big Data. Data Science had proven itself to be a source of profits and had become a part of corporate culture.

In 2011, James Dixon, CTO of Pentaho promoted the concept of Data Lakes, rather than Data Warehouses. Dixon stated the difference between a Data Warehouse and a Data Lake is that the Data Warehouse pre-categorizes the data at the point of entry, wasting time and energy, while a Data Lake accepts the information using a non-relational database (NoSQL) and does not categorize the data, but simply stores it.

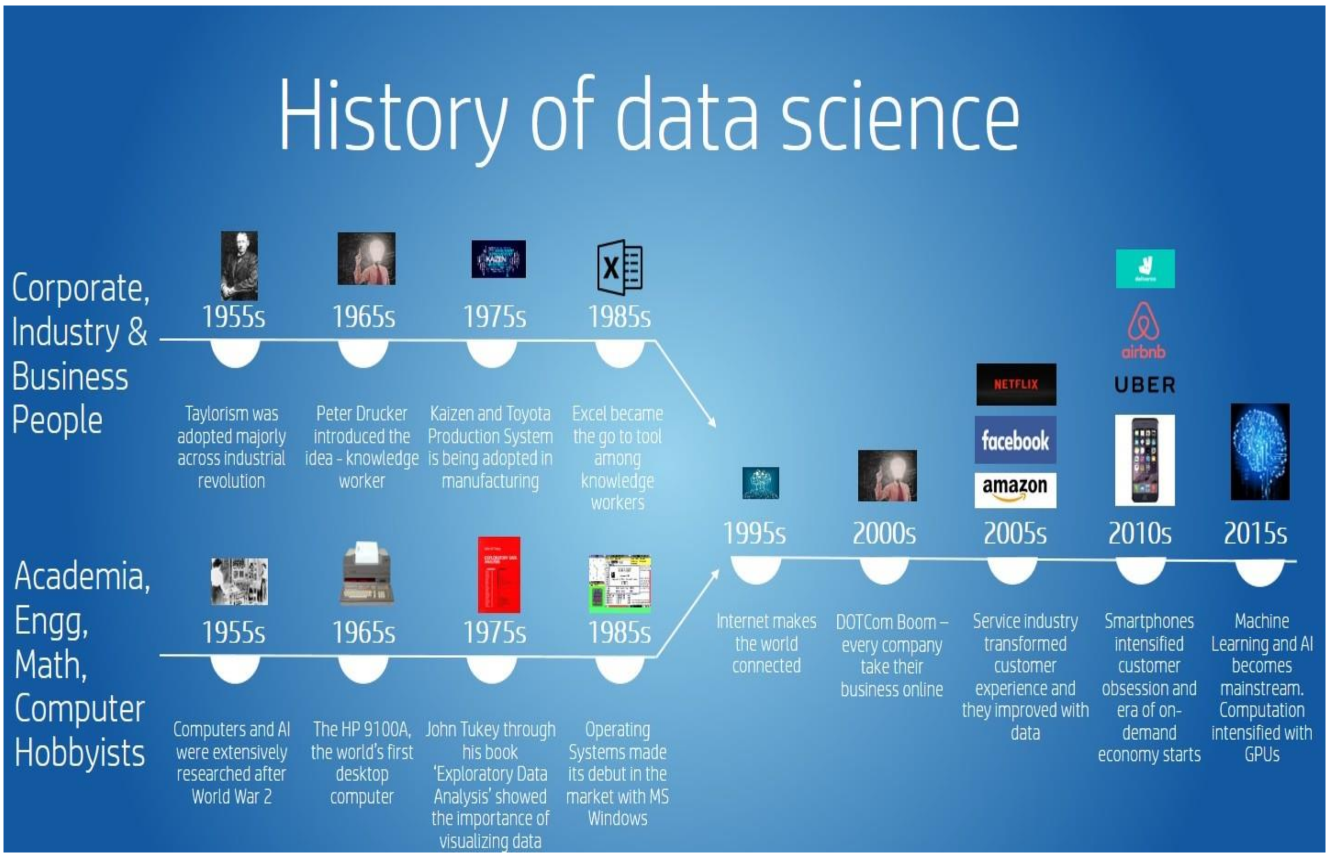
In 2013, IBM shared statistics showing 90% of the data in the world had been created within the last two years.

In 2015, using Deep Learning techniques, Google's speech recognition, Google Voice, experienced a dramatic performance jump of 49 percent.

In 2015, Bloomberg's Jack Clark, wrote that it had been a landmark year for Artificial Intelligence (AI). Within Google, the total of software projects using AI increased from "sporadic usage" to more than 2,700 projects over the year.

In the past ten years, Data Science has quietly grown to include businesses and organizations world-wide. It is now being used by governments, geneticists, engineers, and even astronomers. During its evolution, Data Science's use of Big Data was not simply a "scaling up" of the data, but included shifting to new systems for processing data and the ways data gets studied and analyzed. Data Science has become an important part of business and academic research. Technically, this includes machine translation, robotics, speech recognition, the digital economy, and search engines. In terms of research areas, Data Science has expanded to include the biological sciences, health care, medical informatics, the humanities, and social sciences. Data Science now influences economics, governments, and business and finance.

One result of the Data Science revolution has been a gradual shift to writing more and more conservative programming. It has been discovered Data Scientists can put too much time and energy into developing unnecessarily complex algorithms, when simpler ones work more effectively. As a consequence, dramatic "innovative" changes happen less and less often. Many Data Scientists now think wholesale revisions are simply too risky, and instead try to break ideas into smaller parts. Each part gets tested, and is then cautiously phased into the data flow.



History of data science

Corporate, Industry & Business People

1955s: Taylorism was adopted majorly across industrial revolution

1965s: Peter Drucker introduced the idea - knowledge worker

1975s: Kaizen and Toyota Production System is being adopted in manufacturing

1985s: Excel became the go to tool among knowledge workers

Academia, Engg, Math, Computer Hobbyists

1955s: Computers and AI were extensively researched after World War 2

1965s: The HP 9100A, the world's first desktop computer

1975s: John Tukey through his book 'Exploratory Data Analysis' showed the importance of visualizing data

1985s: Operating Systems made its debut in the market with MS Windows

1995s: Internet makes the world connected

2000s: DOTCom Boom – every company take their business online

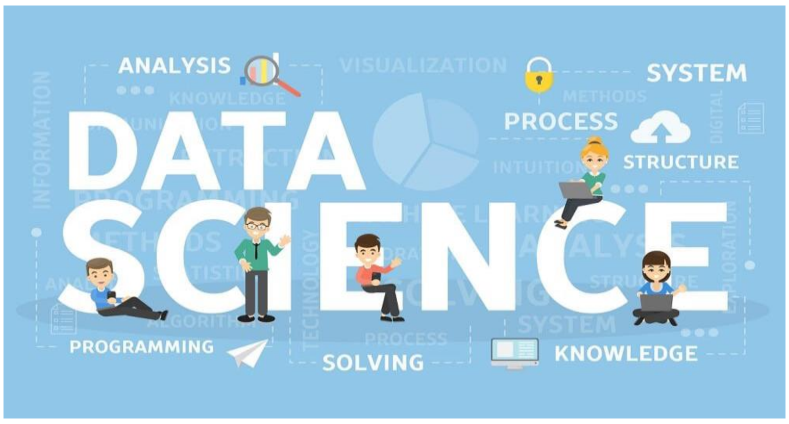
2005s: Service industry transformed customer experience and they improved with data

2010s: Smartphones intensified customer obsession and era of on-demand economy starts

2015s: Machine Learning and AI becomes mainstream. Computation intensified with GPUs

Lesson2. What are the types of Cloud Computing?

The need for Data Science



Simplilearn.com explained that data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

The data used for analysis can be from multiple sources and present in various formats.

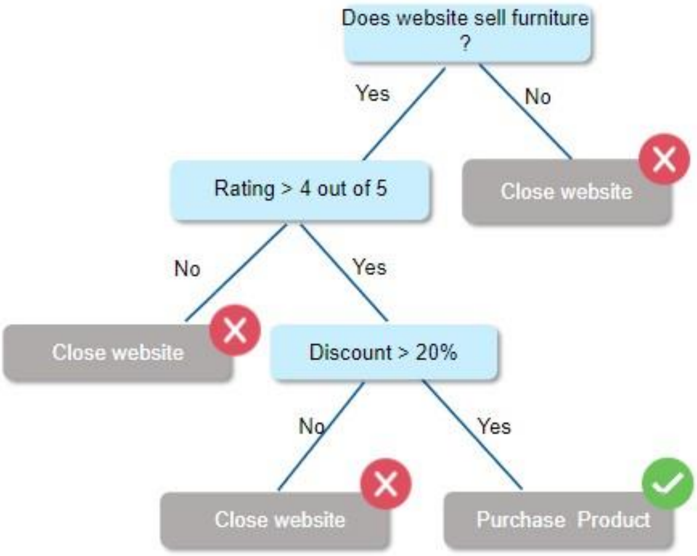
Data science or data-driven science enables better decision making, predictive analysis, and pattern discovery. It lets you:

- Find the leading cause of a problem by asking the right questions
- Perform exploratory study on the data
- Model the data using various algorithms
- Communicate and visualize the results via graphs, dashboards, etc.

In practice, data science is already helping the airline industry predict disruptions in travel to alleviate the pain for both airlines and passengers. With the help of data science, airlines can optimize operations in many ways, including:

- Plan routes and decide whether to schedule direct or connecting flights
- Build predictive analytics models to forecast flight delays
- Offer personalized promotional offers based on customers booking patterns
- Decide which class of planes to purchase for better overall performance

In another example, let's say you want to buy new furniture for your office. When looking online for the best option and deal, you should answer some critical questions before making your decision.



How data science is transforming business

Oracle.com mentioned that organizations are using data science to turn data into a competitive advantage by refining products and services. Data science and machine learning use cases include:

- Determine customer churn by analyzing data collected from call centers, so marketing can take action to retain them
- Improve efficiency by analyzing traffic patterns, weather conditions, and other factors so logistics companies can improve delivery speeds and reduce costs
- Improve patient diagnoses by analyzing medical test data and reported symptoms so doctors can diagnose diseases earlier and treat them more effectively
- Optimize the supply chain by predicting when equipment will break down
- Detect fraud in financial services by recognizing suspicious behaviors and anomalous actions
- Improve sales by creating recommendations for customers based upon previous purchases

Many companies have made data science a priority and are investing in it heavily. In Gartner's recent survey of more than 3,000 CIOs, respondents ranked analytics and business intelligence as the top differentiating technology for their organizations. The CIOs surveyed see these technologies as the most strategic for their companies, and are investing accordingly.

Pre-requisites for Data Science

1. Machine Learning

Machine learning is the backbone of data science. Data Scientists need to have a solid grasp on ML in addition to basic knowledge of statistics.

2. Modeling

Mathematical models enable you to make quick calculations and predictions based on what you already know about the data. Modeling is also a part of ML and involves identifying which algorithm is the most suitable to solve a given problem and how to train these models.

3. Statistics

Statistics are at the core of data science. A sturdy handle on statistics can help you extract more intelligence and obtain more meaningful results.

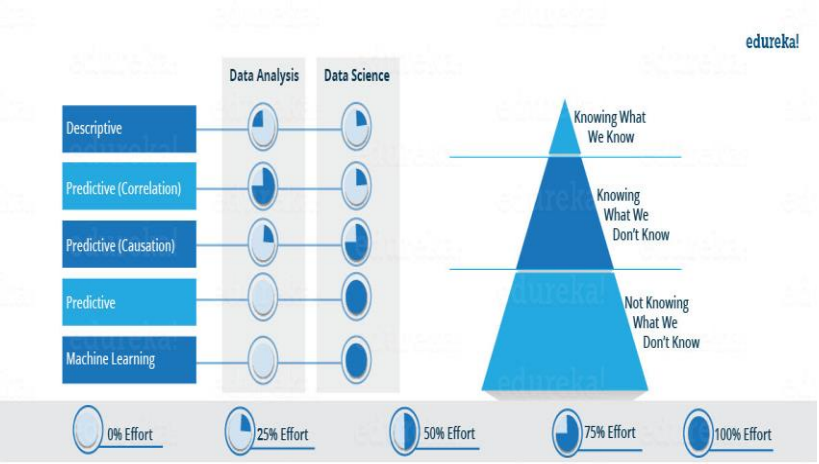
4. Programming

Some level of programming is required to execute a successful data science project. The most common programming languages are Python, and R. Python is especially popular because it's easy to learn, and it supports multiple libraries for data science and ML.

5. Databases

A capable data scientist, you need to understand how databases work, how to manage them, and how to extract data from them.

Understanding Data Science Approach



Edureka.co explained that data Science is primarily used to make decisions and predictions making use of predictive causal analytics, prescriptive analytics (predictive plus decision science) and machine learning.

- ✓ **Predictive causal analytics – ** a model that can predict the possibilities of a particular event in the future, you need to apply predictive causal analytics. For example, if you are providing money on credit, then the probability of customers making future credit payments on time is a matter of concern for you. Here, you can build a model that can perform predictive analytics on the payment history of the customer to predict if the future payments will be on time or not.

- ✓ **Prescriptive analytics:** a model that has the intelligence of taking its own decisions and the ability to modify it with dynamic parameters. This relatively new field is all about providing advice. In other terms, it not only predicts but suggests a range of prescribed actions and associated outcomes. The best example for this is Google's self-driving car the data gathered by vehicles can be used to train self-driving cars. You can run algorithms on this data to bring intelligence to it. This will enable your car to take decisions like when to turn, which path to take, when to slow down or speed up.

- ✓ **Machine learning for making predictions** — a model to determine the future trend, then machine learning algorithms are the best bet. This falls under the paradigm of supervised learning. It is called supervised because you already have the data based on which you can train your machines. For example, a fraud detection model can be trained using a historical record of fraudulent purchases.

- ✓ **Machine learning for pattern discovery** — a model that used to find out the hidden patterns within the dataset to be able to make meaningful predictions. This is nothing but the unsupervised model as you don't have any predefined labels for grouping. The most common algorithm used for pattern discovery is Clustering.
  Let's say you are working in a telephone company, and you need to establish a network by putting towers in a region. Then, you can use the clustering technique to find those tower locations which will ensure that all the users receive optimum signal strength.