INTRODUCTION TO DATA SCIENCE

MODULE -1

LECTURER NOTES

by

MR. SITANSHU KAR

ASST. PROF. DEPT OF CSE

GIET UNIVERSITY, GUNUPUR

Introduction to Data Science

Data science enables businesses to process huge amounts of structured and unstructured big data to detect patterns. This in turn allows companies to increase efficiencies, manage costs, identify new market opportunities, and boost their market advantage. Asking a personal assistant like Alexa or Siri for a recommendation demands data science. So does operating a self-driving car, using a search engine that provides useful results, or talking to a chatbot for customer service. These are all real-life applications for data science. Data science is the practice of mining large data sets of raw data, both structured and unstructured, to identify patterns and extract actionable insight from them. This is an interdisciplinary field, and the foundations of data science include statistics, inference, computer science, predictive analytics, machine learning algorithm development, and new technologies to gain insights from big data. To define data science and improve data science project management, start with its life cycle. The first stage in the data science pipeline workflow involves capture: acquiring data, sometimes extracting it, and entering it into the system. The next stage is maintenance, which includes data warehousing, data cleansing, data processing, data staging, and data architecture.

Data processing follows, and constitutes one of the data science fundamentals. It is during data exploration and processing that data scientists stand apart from data engineers. This stage involves data mining, data classification and clustering, data modeling, and summarizing insights gleaned from the data—the processes that create effective data. Next comes data analysis, an equally critical stage. Here, data scientists conduct exploratory and confirmatory work, regression, predictive analysis, qualitative analysis, and text mining. This stage is why there is no such thing as cookie cutter data science—when it's done properly. During the final stage, the data scientist communicates insights. This involves data visualization, data reporting, the use of various business intelligence tools, and assisting businesses, policymakers, and others in smarter decision making.

By 2020, there will be around 40 zettabytes of data—that's 40 trillion gigabytes. The amount of data that exists grows exponentially. At any time, about 90 percent of this huge amount of data gets generated in the most recent two years, according to sources like IBM and SINTEF. In fact, internet users generate about 2.5 quintillion bytes of data every day. By 2020, every person on Earth will be generating about 146,880 GB of data every day, and by 2025, that will be 165 zettabytes every year. This means there is a huge amount of work in data science—much left to uncover. According to The Guardian, in 2012 only about 0.5 percent of all data was analyzed. Simple data analysis can interpret data from a single source, or a limited amount of data. However, data science tools are critical to understanding big data and data from multiple sources in a meaningful way. A look at some of the specific data science applications in business illustrate this point and provide a compelling introduction to data science.

Data Science vs Data Analytics

Although the work of data scientists and data analysts are sometimes conflated, these fields are not the same. The term data science analyst really just means one or the other. A data scientist comes in earlier in the game than a data analyst, exploring a massive data set, investigating its potential, identifying trends and insights, and visualizing them for others. A data analyst sees data at a later stage. They report on what it tells them, make prescriptions for better performance based on their analysis, and optimize any data related tools. The data analyst is likely to be analyzing a specific dataset of structured or numerical data using a given question or questions. A data scientist is more likely to tackle larger masses of both structured and unstructured data. They will also formulate, test, and assess the performance of data questions in the context of an overall strategy.

Data analytics has more to do with placing historical data in context and less to do with predictive modeling and machine learning. Data analysis isn't an open-minded search for the right question; it relies upon having the right questions in place from the start. Furthermore, unlike data scientists, data analysts typically do not create statistical models or train machine learning tools. Instead, data analysts focus on strategy for businesses, comparing data assets to various organizational hypotheses or plans. Data analysts are also more likely to work with localized data that has already been processed. In contrast, both technical and non-technical data science skills are essential to processing raw data as well as analyzing it. Of course, both roles demand mathematical, analytical, and statistical skills. Data analysts have less need for a broader business culture approach in their everyday work. Instead, they tend to adopt a more measured, nailed-down focus as they analyze pieces of data. Their scope and purpose will almost certainly be more limited than those of a data scientist. In summary, a data scientist is more likely to look ahead, predicting or forecasting as they look at data. The relationship between the data analyst and data is retrospective. A data analyst is more likely to focus on specific questions to answer digging into existing data sets that have already been processed for insights.

Big Data vs Data Science

Data comes from various sources, such as online purchases, multimedia forms, instruments, financial logs, sensors, text files, and others. Data might be unstructured, semi-structured, or structured. Unstructured data includes data from blogs, digital audio/video feeds, digital images, emails, mobile devices, sensors, social networks and tweets, web pages, and online sources. Semi-structured data includes data from system log files, XML files, and text files. Structured data which has already been processed in some way includes OLTP, RDBMS (databases), transaction data, and other formats. This is all "big data," and putting it to good use is a pressing job of the 21st century. It's simply not possible to process tremendous amounts of data from disparate sources with simple business intelligence tools, or even data analytics tools. Instead, data science presents businesses with advanced, complex algorithms and other tools for analyzing, cleansing, processing, and extracting meaningful insights from data. Data science is not one tool, skill, or method. Instead, it is a scientific approach that uses applied statistical and mathematical theory and computer tools to process big data.

The foundations of data science combine the interdisciplinary strengths of data cleansing, intelligent data capture techniques, and data mining and programming. The result is the data scientist's ability to capture, maintain, and prepare big data for intelligent analysis. This is one point that distinguishes the work of the data scientist from the data engineer, although sometimes the two roles are confused. The data engineer prepares data sets for the data scientist to work with and draw insights from, but the intelligent analysis work falls to data scientists, not "data science engineers." Big data is the raw material used in the field of data science. Characterized by its velocity, variety, and volume (the 3Vs), big data is the raw material for data science, which affords the techniques for analyzing the data.

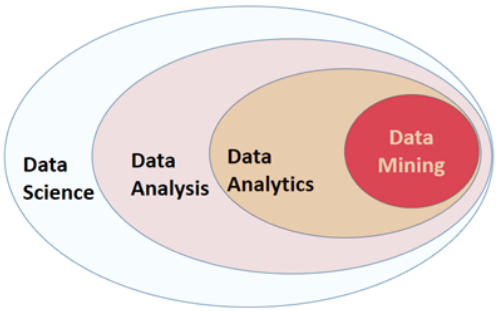
Data Science vs Statistics

Data science is a broad, interdisciplinary area that blends applied business management, computer science, economics, mathematics, programming, and software engineering along with statistics. Data science challenges require the collection, processing, management, analysis, and visualization of mass quantities of data, and data scientists use tools from various fields, including statistics, to achieve those goals. There is a close connection between data science and big data, and most big data exists in unstructured formats and includes some non-numeric data. Therefore, the task of processing data as a data scientist involves eliminating noise and extracting useful insights.

These statistical tasks demand specific design and implementation in four data areas: acquisition, architecture, analysis, and archiving. These "4As" of data science are unique to the field. Statistics is its own broad field demanding subject matter expertise. It does cope with the study of numerical and categorical data, and statistics is an applied area that sees use in numerous other verticals—including data science. For example, statistical theory and methods allow data scientists to gather data in more powerful ways, analyze and interpret them for specific uses, and draw conclusions to solve particular problems. Data scientists frequently employ statistical protocols as they design and conduct research to ensure their results are valid and consistent results. Statistical methods also ensure data scientists can thoroughly explore and describe data while fairly summarizing them. Finally, statistical protocols are essential to accurate prediction and insightful inferences.

Data Mining vs Data Science

Data mining is a technique used in business and data science both, while data science is an actual field of scientific study or discipline. Data mining's goal is to render data more usable for a specific business purpose. Data science, in contrast, aims to create data-driven products and outcomes—usually in a business context. Data mining deals mostly with structured data, as exploring huge amounts of raw, unprocessed data is within the bounds of data science. However, data mining is part of what a data scientist might do, and it's a skill that's part of the science.



Data Science vs Artificial Intelligence

The phrase “artificial intelligence” or (AI) just means simulated human brain function in computers. The traits that signal this kind of brain function include learning, logical reasoning, and self-correction. In other words, when a machine can learn, correct itself as it learns, and reason and draw inferences on its own, it is an AI. Artificial intelligence is either general or narrow. General AI refers to the types of intelligent computers we often see in movies. They can handle a wide range of activities almost like humans do, all of which demand reasoning, judgment, and thought. So far, this has not been achieved.

However, narrow AI involves using the same kinds of “thinking” skills, but on very specific tasks. For example, IBM's Watson is an AI that can interpret certain kinds of medical records for diagnostic purposes as well or better than humans under the right conditions. Scientists and engineers work to achieve artificial intelligence by creating artificial neural networks. But to teach machines to think like a human brain does, even for a very specific purpose, it takes an extraordinary amount of data. This is the intersection of data science, the field; artificial intelligence, the goal; and machine learning, the process.

Similarities Between Data Science and Machine Learning

AI, data science and machine learning all work in tandem. Machine learning is the field of data science that feeds computers huge amounts of data so they can learn to make insightful decisions similar to the way that humans do. For example, most humans learn as children what a flower is without thinking about it. However, the human brain achieves that learning through experience—by collecting data—on which specific features are associated with flowers.

A machine can do the same thing with human help. As humans feed the machine massive quantities of data, it can learn that various petals, stems, and other features are all connected to flowers. In other words, humans feed training data or raw data to the machine, so it can learn all of the data's associated features. Then, if the training was successful, testing with new data should reveal that the machine can distinguish the features it learned. If not, it needs more or better training.

Data Science vs Machine Learning

Data science is a natural extension of statistics. It evolved alongside computer science to handle massive amounts of data with the help of new technologies. In contrast, machine learning is part of data science, but it is more of a process. Machine learning allows computers to learn—and do so more effectively over time—without explicit programs for every bit of information. In machine learning, computers use algorithms to train themselves, but those algorithms rely on some source data. The machine uses that data as a training set, so it can improve its algorithm, tweaking and testing it, optimizing as it goes. It fine-tunes the various parameters of its data science algorithms this way using various statistical techniques, including naive Bayes, regression, and supervised clustering.

However, other techniques that require human input are also part of data science as we understand it today. For example, a machine can train another machine to detect data structures using unsupervised clustering to optimize a classification algorithm. But to completely finish the process, a human must still classify the structures the computer identifies—at least until it is fully trained. The scope of data science also goes far beyond machine learning, encompassing data that is generated not by any mechanical process, computer or machine. For example, data science also includes survey data, data from clinical trials, or really any other kind of data that exists—the full spectrum.

Data science also involves deploying data not just to train machines. Far from being limited to statistical data issues, the field of data science certainly includes automating machine learning and data-driven decisions. However, it also encompasses data integration, data engineering, and data visualization, along with distributed architecture, and the creation of dashboards and other business intelligence tools. In fact, any deployment of data in production mode is also within the scope of data science.

So, where a data scientist creates the insights they pull from data, a machine learns based on those insights that were already perceived by the data scientist. And while a machine may build its own insights on the existing algorithmic structure, the starting point relies on some kind of structured data. In short, a data scientist needs to understand machine learning, which uses many data science techniques. But “data” for a data scientist may or may not involve data from a mechanical process or machine.

Data Science vs Deep Learning

Deep learning is a function of AI that mimics how the human brain works as it processes data and generates patterns to use as it makes decisions. Deep learning is therefore a type of machine learning, focused on deep neural networks that can master

unstructured or unlabeled data without human assistance. This is also called deep neural learning. Deep learning uses hierarchical artificial neural networks to engage in the machine learning process. These artificial neural networks are like complex webs of neuron nodes, much like the human brain. Although traditional data analysis programs approach data in a linear fashion, the deep learning system's hierarchy of function enables a nonlinear approach to problems. Big data is typically unstructured, so deep learning is an important subset of data science research.

Data Science for Business

Data science and analytics come together when data science is applied in a business setting. Data science helps businesses better understand the specific needs customers have based on existing data. For example, with customer age, purchase history, past browsing history, income, and other demographics, a data scientist can more effectively train models for search and product recommendation.

Business Analytics vs Data Science

Both data science and business analytics focus on solving business problems, and both involve collecting data, modeling it, and then gleaning insights from the data. The main difference is that business analytics is specific to business-related problems such as profit and costs. In contrast, data science methods explore how a wide range of factors—anything from customer preferences to the weather—might affect a business. Data science combines data with technology and algorithm building to answer many questions. Business analytics is a narrower field, analyzing data from the business itself with statistical traditional theory to generate insights and business solutions.

Business Intelligence vs Data Science

Business intelligence, a subset of data analysis, analyzes existing data for insights into business trends. Business intelligence gathers data from internal and external sources, prepares and processes it for a specific use, and then creates dashboards with the data to answer business questions. For example, a business intelligence question is specific, such as, “What do we predict our quarterly revenue will be?” “What will our principal business problems be in the coming year?” Business intelligence tools can usually evaluate how certain, specific events might affect a company—at least in the near future. On the other hand, data science is a more exploratory, future-facing approach. Data science analyzes all relevant data, current or past, structured or unstructured—always with the goal of smarter, more informed decision making in mind. In this sense, data science questions are more open-ended, such as “what” events happen, and “how” or “why” they occur.

Data Science in Finance

Data science is a powerful tool for fraud detection and prevention, honing the ability of financial institutions to recognize problematic patterns in data faster. Data science can also help reduce non-performing assets, revealing downward trends sooner. For example, institutions that provide loans or other money on credit need to limit the probability of customers defaulting on payments. To do this, they might use data science to create a model that can perform predictive analytics on customer payment history data. This would allow the institution to predict whether future payments will occur in a timely way—or at all.

How Data Science is Transforming Policy Work

Government policymakers can use data science to better shape policies to meet the needs of their constituents, combat census undercount using big data and ML, and more. For example, policymakers might want to use geospatial data science and related data to drive decisions about when to evacuate an area based on historical weather patterns. The correlative or descriptive analysis of data sets can help make these kinds of decisions. Data scientists can collect and analyze data sets from aircrafts, ships, satellites, and radars to create models. These models can help forecast the weather more effectively every day. However, they also enable scientists to predict natural disasters with greater precision, improve vegetation management, prevent the next Paradise disaster, and help disaster response decision makers know when the optimal evacuation time might be. Learn more about HEAVY.AI's defense analytics and military analytics solutions for real-time defense and intelligence insights.

What is Marketing Data Science?

To understand how data science helps marketing, consider the levels of insight that big data can offer into consumer behavior. Companies can refine things like pricing and other marketing strategies using data science. For example, data science can drive pricing

for e-commerce companies. This allows them to refine their sense of what the market will bear for their product or service, and increase their profits. Data science also helps businesses more effectively develop and market their products, because it allows them to select target customers more efficiently.

Data science and data analytics offer insight into purchasing patterns. Businesses can use descriptive analytics to describe data sets surrounding how consumers buy under different conditions. They can also deploy correlative analysis to predict when relationships might exist between given data sets or variables. For example, data might reveal that a subset of consumers that purchase certain kinds of products are very likely to try products like their new offering. But data science goes further than this kind of analysis. It can also predict future patterns, identifying actions that could meaningfully affect overall business strategy. For instance, data scientists can uncover optimal price points, bids for programmatic advertising or ways to generate new customers in the future based on trends in existing data.

What are Data Science Ethics?

As with any scientific discipline, there is always the potential for bad behavior and abuse in data science. This is the reason data science ethics are so important. There are several basic ethical guidelines for data science to keep in mind. To protect users and the general public, businesses should aim to collect the data they need, but not more. They should protect relevant data with the best available technologies. Furthermore, companies should also promote transparency and guard privacy by keeping data aggregated. In other words, general trends in behavior should be sufficient for both answering business questions and protecting privacy. Learn more about public sector analytics.

Ethical best practices for the field of data science also include identifying and scrubbing sensitive data. This isn't just to protect users; it protects businesses, who can suffer serious reputation damage and customer loss when they fail to protect sensitive data. The ability to identify all sensitive data and secure it also demonstrates two important advantages of data science—value propositions the field offers. First, the business has the capacity to make smart use of big data. Second, it has the will and ability to guard user security despite the ongoing challenges of a dynamic security landscape. This in turn signals a company's ability to react quickly and professionally to data breaches—and the existing potential of data science for good. In this way, ethical best practices showcase data science as a service.

Will Data Science be Automated?

Why data science in an age of automation? The question will data science be automated is an ongoing debate. While many ask the question, “Will data science die,” the better query may be, “How will data science change with automation?”. Experts such as Bernard Marr believe that advances in data visualization and natural language processing (NLP) will mean that data will soon be processed automatically—essentially that many more people will be able to gather insights from data, thanks to augmented analytics and other data science technologies. A report from Gartner makes similar claims, and argues that by 2020, more than 40 percent of data science tasks will be automated. However, this doesn't mean that data science is disappearing—far from it.

With so much data being generated all the time, making data science products simpler for citizen data scientists to use merely improves the reach of businesses working in the space. The place for automation in data science is on manually intensive, repetitive, data science 101 tasks that do not demand deeper training and expertise. For now, the smart view on data science automation seems to be that simpler tasks can and will be automated—soon. However, human management of algorithms and analytics will remain important, because the ability to translate human needs into business questions and strategies is a long way off from being automated. The ability to glean actionable insights from complex data—which would require the automation of context-specific critical thinking—is even further away. Additionally, data scientists with deep business experience and notable industry acumen will continue to see high demand for their skills. Even as more routine, manual tasks related to data may be automated, smart, industry-savvy scientists with data analytical skills will be more in demand in the 21st century. Data science career paths are not going anywhere.

Can Data Science be Self Taught?

Can data science be self-taught, or is a data science specialization of some sort required in the field? In theory, data science can be self-taught, and in practice there are many people working in data science who call themselves self-taught. However, being a self-taught data scientist is a challenge. On the up side, it's relatively easy to find comprehensive lists of the skills and training you'll need to undergo. It’s not even that difficult to master the data science basics if you are motivated. However, it's much harder to develop

mastery in everything from statistical analysis to R or Python, plus niche business knowledge, without support and formal training. To start with, as a self-taught data scientist you'll need to acquire certain skills and training with these data science tools:

- Statistics training, including probability, inferential statistics, linear/vector algebra, and calculus;
- Python and/or R as your primary data science languages;
- Apache Spark;
- SQL;
- Training and experience on a range of data science platforms;
- Experience on Tableau and Snowflake or other visualization software;
- ML training, such as Google machine learning stack Tensorflow;
- NLP training; and
- Deep learning experience are all part of the data science toolkit.

Other programming languages, data science tutorials, statistical and mathematical training, expertise with data science software, and coursework at the "intro to data science" level are also useful. Gain experience and make connections in the field by joining data science associations, participating in hackathons, and solving data science problems in online forums.

R vs Python for Data Science

Data scientists need tools for data transformation, data cleaning, and data visualization. There is also a need to detect outliers, identify relationships between variables, and construct complete interpretive models inside a suitable environment. This is where data preparation and statistical analysis tools like R and Python come in. R was developed as a user-friendly language for statistics, data analysis, and graphical models. R has a large community of programmers that use and support it online, so there is no need to develop everything alone. R is particularly suited to data analysis tasks on individual servers that demand standalone analysis or computing. It's also excellent for exploratory work and ideal for data science visualization, working in tandem with visualization packages such as googleVis, ggvis, ggplot2, and rCharts. On the other hand, R may be too heavy and slow for your system. It also has difficult syntax, and comes with a learning curve that can be steep.

Python was developed as a more readable language for general uses, and it is simpler and more flexible to learn. Another key difference is that R exists mostly within the data science ecosystem, while Python is used in various verticals. The IPython Notebook system allows users to share notebooks with each other, enabling easier working without installations, dramatically reducing lost time. The easier learning curve also typically means shorter time before mastery, including writing and testing your own programs and code—including in other fields. The down side to Python for data science is less data visualization power. Python for data science works in many of the same ways and there is little need to learn them both. However, for some beginner users, Python may be easier to learn due to its simpler syntax. Conversely, for those with more statistical background or more statistical analysis demands, R for data science may be a better choice. Decide based on the data problems you will solve, your ability to learn and master the tool, how much data visualization you expect to do, and the current standards in your specific vertical.

How is Data Visualization Used in Data Science?

Data scientists represent data in the form of graphs, charts and other visualizations. These data visualizations allow users to "see" insights that are invisible in excel sheets of data. For example, you may want to depict how certain trends in data relate to each other, or how multiple factors coincide. Data visualization environments are a common mode of deploying the results of data science to a broader audience, for example, by using web-based tools that allow exploration and interaction with the resulting data. To support effective data visualization, a system must have access to the relevant data science outputs and have intuitive interaction capabilities. Visualizing the data in a scatter-plot or other graph can reveal patterns and relationships that are impossible to observe otherwise. It can also suggest further avenues for research, and new business strategies.

Data science tools

Data scientists rely on popular programming languages to conduct exploratory data analysis and statistical regression. These open source tools support pre-built statistical modeling, machine learning, and graphics capabilities. These languages include the following -

- R Studio: An open source programming language and environment for developing statistical computing and graphics.

- Python: It is a dynamic and flexible programming language. The Python includes numerous libraries, such as NumPy, Pandas, Matplotlib, for analyzing data quickly.

To facilitate sharing code and other information, data scientists may use GitHub and Jupyter notebooks. Some data scientists may prefer a user interface, and two common enterprise tools for statistical analysis include:

- SAS: A comprehensive tool suite, including visualizations and interactive dashboards, for analyzing, reporting, data mining, and predictive modeling.
- IBM SPSS: Offers advanced statistical analysis, a large library of machine learning algorithms, text analysis, open source extensibility, integration with big data, and seamless deployment into applications.

Data scientists also gain proficiency in using big data processing platforms, such as Apache Spark, the open source framework Apache Hadoop, and NoSQL databases. They are also skilled with a wide range of data visualization tools, including simple graphics tools included with business presentation and spreadsheet applications (like Microsoft Excel), built-for-purpose commercial visualization tools like Tableau and IBM Cognos, and open source tools like D3.js (a JavaScript library for creating interactive data visualizations) and RAW Graphs. For building machine learning models, data scientists frequently turn to several frameworks like PyTorch, TensorFlow, MXNet, and Spark MLib.

Given the steep learning curve in data science, many companies are seeking to accelerate their return on investment for AI projects; they often struggle to hire the talent needed to realize data science project's full potential. To address this gap, they are turning to multipersona data science and machine learning (DSML) platforms, giving rise to the role of "citizen data scientist." Multipersona DSML platforms use automation, self-service portals, and low-code/no-code user interfaces so that people with little or no background in digital technology or expert data science can create business value using data science and machine learning. These platforms also support expert data scientists by also offering a more technical interface. Using a multipersona DSML platform encourages collaboration across the enterprise.

Data science and cloud computing

Cloud computing scales data science by providing access to additional processing power, storage, and other tools required for data science projects. Since data science frequently leverages large data sets, tools that can scale with the size of the data is incredibly important, particularly for time-sensitive projects. Cloud storage solutions, such as data lakes, provide access to storage infrastructure, which are capable of ingesting and processing large volumes of data with ease. These storage systems provide flexibility to end users, allowing them to spin up large clusters as needed. They can also add incremental compute nodes to expedite data processing jobs, allowing the business to make short-term tradeoffs for a larger long-term outcome. Cloud platforms typically have different pricing models, such a per-use or subscriptions, to meet the needs of their end user—whether they are a large enterprise or a small startup.

Open source technologies are widely used in data science tool sets. When they're hosted in the cloud, teams don't need to install, configure, maintain, or update them locally. Several cloud providers, including IBM Cloud®, also offer prepackaged tool kits that enable data scientists to build models without coding, further democratizing access to technology innovations and data insights.

Use of Data Science

1. Data science may detect patterns in seemingly unstructured or unconnected data, allowing conclusions and predictions to be made.
2. Tech businesses that acquire user data can utilize strategies to transform that data into valuable or profitable information.
3. Data Science has also made inroads into the transportation industry, such as with driverless cars. It is simple to lower the number of accidents with the use of driverless cars. For example, with driverless cars, training data is supplied to the algorithm, and the data is examined using data Science approaches, such as the speed limit on the highway, busy streets, etc.
4. Data Science applications provide a better level of therapeutic customization through genetics and genomics research.

A brief history of Data Science

Interest in data science-related careers is at an all-time high and has exploded in popularity in the last few years. Data scientists today are from various backgrounds. If someone ran into you ask what data science is all about, what would you tell them? It is not an easy question to answer. Data science is one of the areas that everyone is talking about, but no one can define it well. Media has been hyping about "Data Science," "Big Data," and "Artificial Intelligence" over the past few years.

For outsiders, data science is whatever magic that can get useful information out of data. Everyone should have heard about big data. Data science trainees now need the skills to cope with such big data sets. What are those skills? You may hear about: Hadoop, a system using Map/Reduce to process large data sets distributed across a cluster of computers, or hear about Spark, a system builds atop Hadoop for speeding up the same by loading massive datasets into shared memory (RAM) across clusters with an additional suite of machine learning functions for big data. The new skills are for dealing with organizational artifacts of large data sets beyond a single computer's memory or hard disk and the large-scale cluster computing but not for better solving the real problem. A lot of data means more sophisticated tinkering with computers, especially a cluster of computers. The computing and programming skills to handle big data were the biggest hurdle for traditional analysis practitioners to be a successful data scientist. However, this hurdle is significantly reduced with the cloud computing revolution. After all, it isn't the size of the data that's

important. It's what you do with it. Your first reaction to all of this might be some combination of skepticism and confusion. We want to address this upfront that: we had that exact reaction. To declutter, let's start with a brief history of data science. If you hit up the Google Trends website, which shows search keyword information over time, and check the term "data science," you will find the history of data science goes back a little further than 2004. The way media describes it, you may feel that machine learning algorithms are new, and there was never "big" data before Google. That is not true. There are new and exciting developments in data science. But many of the techniques we are using are based on decades of work by statisticians, computer scientists, mathematicians, and scientists of many other fields.

In the early 19th century, when Legendre and Gauss came up with the least-squares method for linear regression , probably only physicists would use it to fit linear regression for their data. Now, nearly anyone can build linear regression using excel with just a little bit of self-guided online training. In 1936, Fisher came up with linear discriminant analysis. In the 1940s, we had another widely used model – logistic regression . In the 1970s, Nelder and Wedderburn formulated a "generalized linear mode (GLM) " which: By the end of the 1970s, there was a range of analytical models, and most of them were linear because computers were not powerful enough to fit non-linear models until the 1980s.

In 1984, Breiman introduced the classification and regression tree (CART) , one of the oldest and most utilized classification and regression techniques(Breiman et al., 1984). After that, Ross Quinlan came up with more tree algorithms such as ID3, C4.5, and C5.0. In the 1990s, ensemble techniques (methods that combine many models' predictions) began to appear. Bagging is a general approach that uses bootstrapping in conjunction with regression or classification model to construct an ensemble. Based on the ensemble idea, Breiman came up with the random forest model in 2001 (Breiman, 2001a). In the same year, Leo Breiman published a paper "Statistical Modeling: The Two Cultures1" (Breiman, 2001b) where he pointed out two cultures in the use of statistical modeling to get information from data:

- Data is from a given stochastic data model
- Data mechanism is unknown and people approach the data using algorithmic model

Most of the classical statistical models are the first type of stochastic data model. Black-box models, such as random forest, GMB, and deep learning , are algorithmic modeling. As Breiman pointed out, algorithmic models can be used on large complex data as a more accurate and informative alternative to stochastic data modeling on smaller data sets. Those algorithms have developed

rapidly with much-expanded applications in fields outside traditional statistics. That is one of the most important reasons that statisticians are not the mainstream of today's data science, both in theory and practice. We observe that Python is passing R as the most commonly used language in data science, mainly due to many data scientists' background. Since 2000, the approaches to getting information out of data have shifted from traditional statistical models to a more diverse toolbox that includes machine learning and deep learning models.

John Tukey identified **four forces driving data analysis** (there was no "data science" back to 1962):

1. The formal theories of math and statistics
2. Acceleration of developments in computers and display devices
3. The challenge, in many fields, of more and ever larger bodies of data
4. The emphasis on quantification in an ever-wider variety of disciplines

Tukey's 1962 list is surprisingly modern. Let's inspect those points in today's context. People usually develop theories way before they find potential applications. In the past 50 years, statisticians, mathematicians, and computer scientists have laid the theoretical groundwork for constructing "data science" today. The development of computers enables us to apply the algorithmic models (which can be very computationally expensive) and deliver results in a friendly and intuitive way. The striking transition to the internet and the internet of things generates vast amounts of commercial data. Industries have also sensed the value of exploiting that data. Data science seems sure to be a significant preoccupation of commercial life in the coming decades. All the four forces John identified exist today and have been driving data science. The toolbox and application have been expanding fast, benefiting from the increasing availability of digitized information and the possibility of distributing it through the internet. Today, people apply data science in many areas, including business, health, biology, social science, politics, etc. Now data science is everywhere.

Data science role and Skill tracks

There is a widely diffused Chinese parable about a group of blind men conceptualizing what the elephant is like by touching it. The first person, whose hand landed on the trunk, said: "This being is like a thick snake." For another one whose hand reached its ear, it seemed like a fan. Another person whose hand was upon its leg said the elephant is a pillar-like tree trunk. The blind man who placed his hand upon its side said: "elephant is a wall." Another who felt its tail described it as a rope. The last felt its tusk, stating the elephant is hard, smooth, and spear. Data science is the elephant. With the data science hype picking upstream, many professionals changed their titles to be "Data Scientist" without any necessary qualifications. Today's data scientists have vastly different backgrounds, yet each conceptualizes the elephant based on his/her professional training and application area. And to make matters worse, most of us are not even fully aware of our conceptualizations, much less the uniqueness of the experience from which they are derived.

It is annoying but true. So, the answer to the question "what is data science?" depends on who you are talking to. Who may you be talking to then? Data science has three main skill tracks: engineering, analysis, and modeling (and yes, the order matters!). There are some representative skills in each track. Different tracks and combinations of tracks will define different roles in data science. 2 When people talk about all the machine learning and AI algorithms, they often overlook the critical data engineering part that makes everything possible. Data engineering is the unseen iceberg under the water surface. Does your company need a data scientist? You are not ready for a data scientist if you don't have a data engineer yet. You need to have the ability to get data before making sense of it. If you only deal with small datasets with formatted data, you may be able to get by with plain text files such as CSV (i.e., comma-separated values) or even Excel Spreadsheet. As the data increasing in volume, variety, and velocity, data engineering becomes a sophisticated discipline in its own right.

Data Engineering

Data engineering is the foundation that makes everything else possible. It mainly involves in building the data pipeline infrastructure. In the (not that) old day, when data is stored on local servers, computers, or other devices, building the data infrastructure can be a massive IT project. It involves the software and the hardware used to store the data and perform data ETL (i.e., extract, transform, and load) process. As cloud service development, it becomes the new norm to store and compute data on the cloud. Data engineering today, at its core, is software engineering with data flow as the focus. The fundamental building block for automation is maintaining the data pipeline through modular, well-commented code and version control.

(1) Data environment - Designing and setting up the entire environment to support data science workflow is the prerequisite for data science projects. It may include setting up storage in the cloud, Kafka platform, Hadoop and Spark cluster, etc. Each company has a unique data condition and need. The data environment will be different depending on the size of the data, update frequency, the complexity of analytics, compatibility with the back-end infrastructure, and (of course) budget.

(2) Data management - Automated data collection is a common task that includes parsing the logs (depending on the stage of the company and the type of industry you are in), web scraping, API queries, and interrogating data streams. Determine and construct data schema to support analytical and modeling needs. Use tools, processes, guidelines to ensure data is correct, standardized, and documented.

(3) Production - If you want to integrate the model or analysis into the production system, you have to automate all data handling steps. It involves the whole pipeline from data access, preprocessing, modeling to final deployment. It is necessary to make the system work smoothly with all existing software stacks. So, it requires to monitor the system through some robust measures, such as rigorous error handling, fault tolerance, and graceful degradation to make sure the system is running smoothly and the users are happy.

Data Analysis

Analysis turns raw information into insights in a fast and often exploratory way. In general, an analyst needs to have decent domain knowledge, do exploratory analysis efficiently, and present the results using storytelling.

(1) **Domain knowledge** - Domain knowledge is the understanding of the organization or industry where you apply data science. You can't make sense of data without context. Some questions about the context are:

- What are the critical metrics for this kind of business?
- What are the business questions?
- What type of data do they have, and what does the data represent?
- How to translate a business need to a data problem?
- What has been tried before, and with what results?
- What are the accuracy-cost-time trade-offs?
- How can things fail?
- What are other factors not accounted for?
- What are the reasonable assumptions, and what are faulty?

In the end, domain knowledge helps you to deliver the results in an audience-friendly way with the right solution to the right problem.

(2) **Exploratory analysis** - This type of analysis is about exploration and discovery. Rigor conclusion is not a concern, which means the goal is to get insights driven by correlation, not causation. The latter one requires more advanced statistical skills and hence more time and resource expensive. Instead, this role will help your team look at as much data as possible so that the decision-makers can get a sense of what's worth further pursuing. It often involves different ways to slice and aggregate data. An important thing to note here is that you should be careful not to get a conclusion beyond the data. You don't need to write production-level robust codes to perform well in this role.

(3) **Story telling** – Story telling with data is critical to deliver insights and drive better decision making. It is the art of telling people what the numbers signify. It usually requires data summarization, aggregation, and visualization. It is crucial to answering the following questions before you begin down the path of creating a data story.

- Who is your audience?
- What do you want your audience to know or do?
- How can you use data to help make your point?

A business-friendly report or an interactive dashboard is the typical outcome of the analysis.

Data Modeling

Modeling is a process that dives deeper into the data to discover the pattern we don't readily see. A fancy machine learning model is the first thing that comes to people's minds when the general public thinks about data science. Unfortunately, fancy models only occupy a small part of a typical data scientist's day-to-day time. Nevertheless, many of those models are powerful tools.

(1) **Supervised learning** - In supervised learning, each sample corresponds to a response measurement. There are two flavors of supervised learning: regression and classification. In regression, the response is a real number, such as the total net sales in 2017 for a company or the yield of corn next year for a state. The goal for regression is to approximate the response measurement as much as possible. In classification, the response is a class label, such as a dichotomous response of yes/no. The response can also have more than two categories, such as four segments of customers. A supervised learning model is a function that maps some input variables (X) with corresponding parameters (beta) to a response (y). The modeling process is to adjust the value of parameters to make the mapping fit the given response. In other words, it is to minimize the discrepancy between given responses and the model output. When the response y is a real value number, it is intuitive to define discrepancy as the squared difference between model output and the response. When y is categorical, there are other ways to measure the difference, such as the area under the receiver operating characteristic curve (i.e., AUC) or information gain.

(2) **Unsupervised learning** - In unsupervised learning, there is no response variable. For a long time, the machine learning community overlooked unsupervised learning except for one called clustering. Moreover, many researchers thought that clustering was the only form of unsupervised learning. One reason is that it is hard to define the goal of unsupervised learning explicitly. Unsupervised learning can be used to do the following:

- Identify a good internal representation or pattern of the input that is useful for subsequent supervised or reinforcement learning, such as finding clusters;
- It is a dimension reduction tool that provides compact, low dimensional representations of the input, such as factor analysis.
- Provide a reduced number of uncorrelated learned features from original variables, such as principal component regression.

(3) **Customized model development** - In most cases, after a business problem is fully translated into a data science problem, a data scientist needs to use out of the box algorithms to solve the problem with the right data. But in some situations, there isn't enough data to use any machine learning model, or the question doesn't fit neatly in the specifications of existing tools, or the model needs to incorporate some prior domain knowledge. A data scientist may need to develop new models to accommodate the subtleties of the problem at hand. For example, people may use Bayesian models to include domain knowledge as the modeling process's prior distribution. Here is a list of questions that can help you decide the type of technique to use:

- Is your data labeled? It is straightforward since supervised learning needs labeled data.
- Do you want to deploy your model at scale? There is a fundamental difference between building and deploying models. It is like the difference between making bread and making bread machine. One is a baker who will mix and bake ingredients according to recipes to make a variety of bread. One is a machine builder who builds a machine to automate the process and produce bread at scale.
- Is your data easy to collect? One of the major sources of cost in deploying machine learning is collecting, preparing, and cleaning the data. Because model maintenance includes continuously collecting data to keep the model updated. If the data collection process requires too much human labor, the maintenance cost can be too high.
- Does your problem have a unique context? If so, you may not be able to find any off-the-shelf method that can directly apply to your question and need to customize the model.

Data Preprocessing

There are some common skills to have, regardless of the role people have in data science like Data Preprocessing - the process nobody wants to go through yet nobody can avoid No matter what role you hold in the data science team, you will have to do some data cleaning, which tends to be the least enjoyable part of anyone's job. Data preprocessing is the process of converting raw data into clean data that is proper to use.

(1) **Data preprocessing for data engineer -** Getting data from different sources and dumping them into a data lake, a dumping ground of amorphous data, is far from the data schema analyst and scientist would use. A data lake is a storage repository that stores a vast amount of raw data in its native format, including XML, JSON, CSV, Parquet, etc. It is a data cesspool rather than a data lake. The data engineer's job is to get a clean schema out of the data lake by transforming and formatting the data. Some common problems to resolve are:

- Enforce new tables' schema to be the desired one
- Repair broken records in newly inserted data
- Aggregate the data to form the tables with a proper granularity

(2) Data preprocessing for data analyst and scientist Not just for a data engineer, preprocessing also occupies a large portion of data analyst and scientist's working hours. A facility and a willingness to do these tasks are a prerequisite for a good data scientist. The data a data scientist gets can still be very rough even if it is from a nice and clean database that a data engineer sets up. For example, dates and times are notorious for having many representations and time zone ambiguity. You may also get market survey responses from your clients in an excel file where the table title could be multi-line, or the format does not meet the requirements, such as using 50% to represent the percentage rather than 0.5. In many cases, you need to set the data to be the right format before moving on to analysis. Even the data is in the right format. There are other issues to solve before or during analysis and modeling. For example, variables can have missing values. Knowledge about the data collection process and what it will be used for is necessary to decide a way to handle the missing. Also, different models have different requirements for the data. For example, some models may require a consistent scale; some may be susceptible to outliers or collinearity; some may not be able to handle categorical variables, and so on. The modeler has to preprocess the data to make it proper for the specific model.

Most of the people in data science today focus on one of the tracks. A small number of people are experts on two tracks. People who are proficient in all three? They are unicorns!

Problem type

What kind of questions can data science solve…

Many of the data science books classify various models from a technical point of view. Such as supervised vs. unsupervised models, linear vs. nonlinear models, parametric models vs. non-parametric models, and so on. Here we will continue on a "problem-oriented" track. We first introduce different groups of real-world problems and then present which models can answer the corresponding category of questions.

1. **Description -** The primary analytic problem is to summarize and explore a data set with descriptive statistics (mean, standard deviation, and so forth) and visualization methods. It is the most straightforward problem and yet the most crucial and common one. We will need to describe and explore the dataset before moving on to a more complex analysis. For problems such as customer segmentation, after we cluster the sample, the next step is to figure out each class's profile by comparing the descriptive statistics of various variables. Data description is often used to check data, find the appropriate data preprocessing method, and demonstrate the model results.
2. **Comparison –** The first common modeling problem is to compare different groups. Is A better in some way than B? Or more comparisons: Is there any difference among A, B, and C in a particular aspect? For those problems, it usually starts with some summary statistics and visualization by groups. After a preliminary visualization, you can test the differences between the treatment and control groups statistically. The commonly used statistical tests are chi-square test, t-test, and ANOVA. There are also methods using Bayesian methods. In the biology industry, such as new drug development, crop breeding, fixed/random/mixed effect models are standard techniques.
3. **Clustering -** Clustering is a widespread problem, and it can answer questions like: How many reasonable customer segments are there based on historical purchase patterns? Or How are the customer segments different from each other? Please note that clustering is unsupervised learning; there are no response variables. The most common clustering algorithms include K-Means and Hierarchical Clustering.
4. **Classification –** For classification problems, there are one or more label columns to define the ground truth of classes. We use other features of the training dataset as explanatory variables for model training. We can use the trained classifier to predict the labels of a new observation. Here are some example questions: such as Will this customer likely to buy our product? or Is the borrower going to pay us back? or Is it spam email or not? There are hundreds of different classifiers. In practice, we do not need to try all the models but several models that perform well generally. For example, the random forest algorithm is usually used as the baseline model to set model performance expectations.
5. **Regression –** In general, regression deals with a question like "how much is it?" and return a numerical answer. It is necessary to coerce the model results to be 0 or round it to the nearest integer in some cases. It is still the most common problem in the data science world. Here are some example questions: such as What will be the temperature tomorrow? or What is the projected net income for the next season? or How much inventory should we have?
6. **Optimization –** Optimization is another common type of problems in data science to find an optimal solution by tuning a few tune-able variables with other non-controllable environmental variables. It is an expansion of comparison problem and can solve problems such as: What is the best route to deliver the packages? or What is the optimal advertisement strategy to promote a new product?

List of potential data science careers

As companies learn about using data to help with the business, there is a continuous specialization of different data science roles. As a result, the old "data scientist" title is fading, and some other data science job titles are emerging. The misunderstanding of data science's fundamental work leads to confusing job postings and frustrations for both stakeholders and data scientists. Stakeholders are frustrated that they aren't getting what they expect, and data scientists are frustrated that their talent is not appreciated. We are glad to see that the change is underway. Here is a list of today's data science job titles. Some of them are relatively new, and the others have been around for some time but now are better defined.

Role	Skills
Data infrastructure engineer	Go, Python, AWS/Google Cloud/Azure, logstash, Kafka, and Hadoop

Role	Skills
Data engineer	spark/scala, python, SQL, AWS/Google Cloud/Azure, Data modeling
BI engineer	Tableau/looker/Mode, etc., data visualization, SQL, Python
Data analyst	SQL, basic statistics, data visualization
Data scientist	R/Python, SQL, basic + applied statistics, data visualization, experimental design.
Research scientist	R/Python, advanced statistics + experimental design, ML, research background, publications, conference contributions, algorithms
Applied scientist	ML algorithm design, often with an expectation of fundamental software engineering skills
Machine Learning	Engineer More advanced software engineering skillset, algorithms, machine learning algorithm design, system design

	Business Knowledge	Production	Data Frequency	Engineering Skill	Math/Stat	(Un)Str Data
Data infrastructure engineer	Very Low	Yes	Very High	High	Very Low	Both
Data engineer	Low	Yes	High	High	Low	Both
BI engineer	Mid	Depends	Mid	Mid	Low	Str
Data analyst	Very High	No	Mid	Very Low	Mid	Str
Data scientist	High	Mostly No	Mid	Low/Mid	High/Very High	Mostly Str
Research scientist	High	No	Mid	Low/Mid	High/Very High	Str
Applied scientist	High	Depends	Mid/High	Mid	Mid/High	Mostly Str
Machine Learning Engineer	Low	Yes	High	High	Mid	Both

Not only is data science a cutting-edge field that allows you to make an important impact, both with within your company and on a global scale, it's also one that's growing at an astounding rate. As an exponentially expanding number of industries see the benefit of using analytical data to improve business practices, big data and data science career opportunities are exploding. In fact, employment for statisticians in data science related careers is projected to grow 33.8% from 2016 to 2026, according to the Bureau of Labor Statistics (BLS), described as the fastest occupational growth in the mathematical area of the industry.

Data science related occupations are likely to enjoy excellent job prospects, as many companies report difficulties finding highly skilled workers. That means there's more demand for data science professionals than there is supply, which is good news for data science students and professionals. As a result of this shortage, you'll find that there is a wealth of different avenues that a data science career can take. While it's always good to have options, it can sometimes be difficult to understand how these careers differ and what kinds of skillsets and educational backgrounds are required for each. This can present a challenge for those just starting out in the world of data science.

Life Cycle Phases of Data Analytics

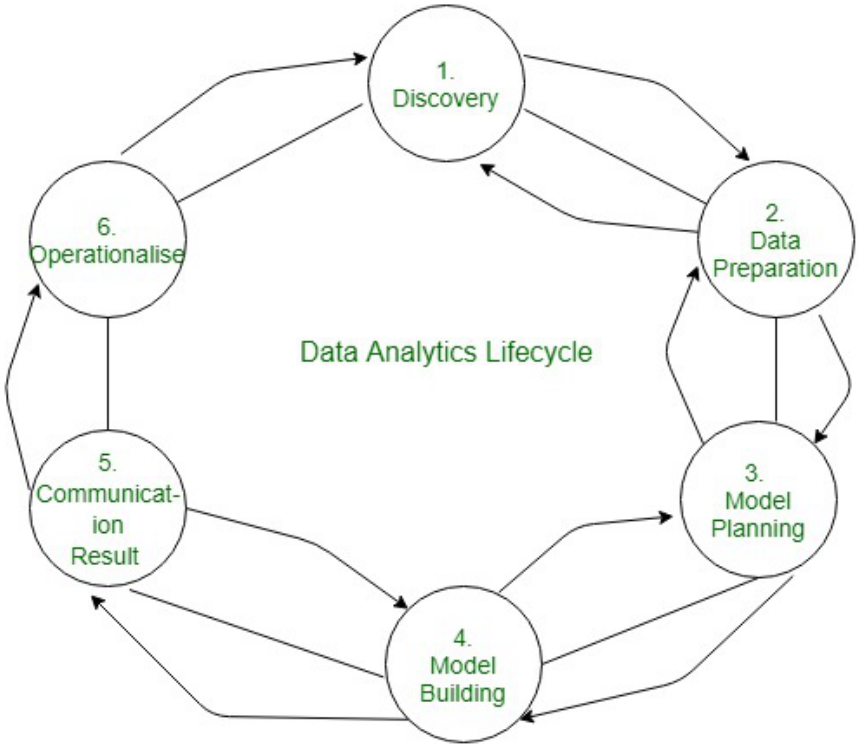
In today's digital-first world, data is of immense importance. It undergoes various stages throughout its life, during its creation, testing, processing, consumption, and reuse. Data Analytics Lifecycle maps out these stages for professionals working on data analytics projects. These phases are arranged in a circular structure that forms a Data Analytics Lifecycle. Each step has its significance and characteristics. The Data Analytics Lifecycle is designed to be used with significant big data projects. It is used to portray the actual project correctly; the cycle is iterative. A step-by-step technique is needed to arrange the actions and tasks involved in gathering, processing, analyzing, and reusing data to explore the various needs for assessing the information on big data. Data analysis is modifying, processing, and cleaning raw data to obtain useful, significant information that supports business decision-making.

Data Analytics Lifecycle defines the roadmap of how data is generated, collected, processed, used, and analyzed to achieve business goals. It offers a systematic way to manage data for converting it into information that can be used to fulfill organizational and project goals. The process provides the direction and methods to extract information from the data and proceed in the right direction to accomplish business goals. Professionals use the lifecycle's circular form to proceed with data analytics in either forward

or backward direction. Based on the newly received insights, they can decide whether to proceed with their existing research or scrap it and redo the complete analysis. The Data Analytics lifecycle guides them throughout this process.

DATA ANALYTICS LIFECYCLE PHASES

Data is extremely important in today's digital-first world, as it has always been. Throughout its life cycle, it goes through a number of stages, including creation, testing, processing, consumption, and repurposing. The Data Analytics Lifecycle is a diagram that depicts these steps for professionals that are involved in data analytics projects. The phases of the Data Analytics Lifecycle are organized in a circular framework, which is referred to as the Data Analytics Lifecycle. Each stage has its own significance as well as its own peculiarities. The phases that are fundamental to each data analytics process. Hence, they are more likely to be present in most data analytics projects' lifecycle. The Data Analytics lifecycle primarily consists of 6 phases.



Phase 1: Discovery – This phase is all about defining the data's purpose and how to achieve it by the end of the data analytics lifecycle. The stage consists of identifying critical objectives a business is trying to discover by mapping out the data. During this process, the team learns about the business domain and checks whether the business unit or organization has worked on similar projects to refer to any learnings. In this phase, the team also evaluates technology, people, data, and time. For example, while dealing with a small dataset, the team can use Excel. However, heftier tasks demand more rigid tools for data preparation and exploration. In such scenarios, the team will need to use Python, R, Tableau Desktop or Tableau Prep, and other data cleaning tools. This phase's critical activities include framing the business problem, formulating initial hypotheses to test, and beginning data learning. The key aspects to be considered in this phase are : -

- The data science team learn and investigate the problem.
- Develop context and understanding.
- Come to know about data sources needed and available for the project.
- The team formulates initial hypothesis that can be later tested with data.

Phase 2: Data Preparation – In this phase, the experts' focus shifts from business requirements to information requirements. One of the essential aspects of this phase is ensuring data availability for processing. During this phase's initial stage, the team gathers valuable information and proceeds with the business ecosystem's lifecycle. Various data collection methods are used for this purpose, such as

a) Data Entry – Collecting recent data using manual data entry techniques or digital systems within the organization
b) Data Acquisition – Gathering data from external sources
c) Signal Reception – Capturing data from digital devices, including the Internet of Things and control systems.

The stage encompasses the collection, processing, and cleansing of the accumulated data. The key aspects to be considered in this phase are : -

- Steps to explore, preprocess, and condition data prior to modeling and analysis.
- It requires the presence of an analytic sandbox, the team execute, load, and transform, to get data into the sandbox.
- Data preparation tasks are likely to be performed multiple times and not in predefined order.
- Several tools commonly used for this phase are – Hadoop, Alpine Miner, Open Refine, etc.

Phase 3: Model Planning – This phase needs the availability of an analytic sandbox for the team to work with data and perform analytics throughout the project duration. The team can load data in several ways. The team identifies variables for categorizing data, identifies and amends data errors. Data errors can be anything, including missing data, illogical values, duplicates, and spelling errors. For example, the team imputes the average data score for categories for missing values. It enables more efficient data processing without skewing the data.

After cleaning the data, the team determines the techniques, methods, and workflow for building a model in the next phase. The team explores the data, identifies relations between data points to select the key variables, and eventually devises a suitable model. The key aspects to be considered in this phase are : -

- Team explores data to learn about relationships between variables and subsequently, selects key variables and the most suitable models.
- In this phase, data science team develop data sets for training, testing, and production purposes.
- Team builds and executes models based on the work done in the model planning phase.
- Several tools commonly used for this phase are – Matlab, STASTICA.

Phase 4: Model Building – In this phase, the team develops testing, training, and production datasets. Further, the team builds and executes models meticulously as planned during the model planning phase. They test data and try to find out answers to the given objectives. They use various statistical modeling methods such as regression techniques, decision trees, random forest modeling, and neural networks and perform a trial run to determine whether it corresponds to the datasets. The key aspects to be considered in this phase are : -

- Team develops datasets for testing, training, and production purposes.
- Team also considers whether its existing tools will suffice for running the models or if they need more robust environment for executing models.
- Free or open-source tools – Rand PL/R, Octave, WEKA.
- Commercial tools – Matlab , STASTICA.

Phase 5: Communication Results – This phase aims to determine whether the project results are a success or failure and start collaborating with significant stakeholders. The team identifies the vital findings of their analysis, measures the associated business value, and creates a summarized narrative to convey the stakeholders' results. key aspects to be considered in this phase are : -

- After executing model team need to compare outcomes of modeling to criteria established for success and failure.
- Team considers how best to articulate findings and outcomes to various team members and stakeholders, taking into account warning, assumptions.
- Team should identify key findings, quantify business value, and develop narrative to summarize and convey findings to stakeholders.

Phase 6: Operationalize – In this final phase, the team presents an in-depth report with coding, briefing, key findings, and technical documents and papers to the stakeholders. Besides this, the data is moved to a live environment and monitored to measure the analysis's effectiveness. If the findings are in line with the objective, the results and reports are finalized. On the other hand, if they deviate from the set intent, the team moves backward in the lifecycle to any previous phase to change the input and get a different outcome. key aspects to be considered in this phase are : -

- The team communicates benefits of project more broadly and sets up pilot project to deploy work in controlled way before broadening the work to full enterprise of users.
- This approach enables team to learn about performance and related constraints of the model in production environment on small scale , and make adjustments before full deployment.
- The team delivers final reports, briefings, codes.

- Free or open source tools – Octave, WEKA, SQL, MADlib.

DATA ANALYTICS LIFECYCLE EXAMPLE

Consider an example of a retail store chain that wants to optimize its products' prices for boosting its revenue. The store chain has thousands of products over hundreds of outlets, making it a highly complex scenario. Once you identify the store chain's objective, you find the data you need, prepare it, and go through the Data Analytics lifecycle process. You observe different types of customers, such as ordinary customers and customers like contractors who buy in bulk. According to you, treating various types of customers differently can give you the solution. However, you don't have enough information about it and need to discuss this with the client team. In this case, you need to get the definition, find data, and conduct the hypothesis testing to check whether various customer types impact the model results and get the right output. Once you are convinced with the model results, you can deploy the model, integrate it into the business, and you are all set to deploy the prices you think are the most optimal across the outlets of the store.

The Data Analytics lifecycle's circular process consists of 6 primary stages that dictate how information is created, collected, processed, used, and analyzed. Mapping out business objectives and striving towards achieving them will guide you through the rest of the stages.

Application of data Science in various Field

The role of Data Science Applications hasn't evolved overnight. Thanks to faster computing and cheaper storage, we can now predict outcomes in minutes, which could take several human hours to process. The top 10 applications that build upon the concepts of Data Science, exploring various domains such as the following:

1. Fraud and Risk Detection
2. Healthcare
3. Internet Search
4. Targeted Advertising
5. Website Recommendations
6. Advanced Image Recognition
7. Speech Recognition
8. Airline Route Planning
9. Gaming
10. Augmented Reality

1. Fraud and Risk Detection

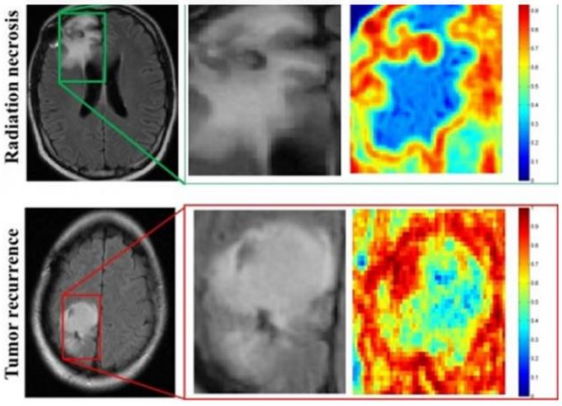
The earliest applications of data science were in Finance. Companies were fed up of bad debts and losses every year. However, they had a lot of data which use to get collected during the initial paperwork while sanctioning loans. They decided to bring in data scientists in order to rescue them from losses. Over the years, banking companies learned to divide and conquer data via customer profiling, past expenditures, and other essential variables to analyze the probabilities of risk and default. Moreover, it also helped them to push their banking products based on customer's purchasing power.

2. Healthcare

The healthcare sector, especially, receives great benefits from data science applications.

a. Medical Image Analysis

Procedures such as detecting tumors, artery stenosis, organ delineation employ various different methods and frameworks like MapReduce to find optimal parameters for tasks like lung texture classification. It applies machine learning methods, support vector machines (SVM), content-based medical image indexing, and wavelet analysis for solid texture classification.



b) Genetics & Genomics

Data Science applications also enable an advanced level of treatment personalization through research in genetics and genomics. The goal is to understand the impact of the DNA on our health and find individual biological connections between genetics, diseases, and drug response. Data science techniques allow integration of different kinds of data with genomic data in the disease research, which provides a deeper understanding of genetic issues in reactions to particular drugs and diseases. As soon as we acquire reliable personal genome data, we will achieve a deeper understanding of the human DNA. The advanced genetic risk prediction will be a major step towards more individual care.

c) Drug Development

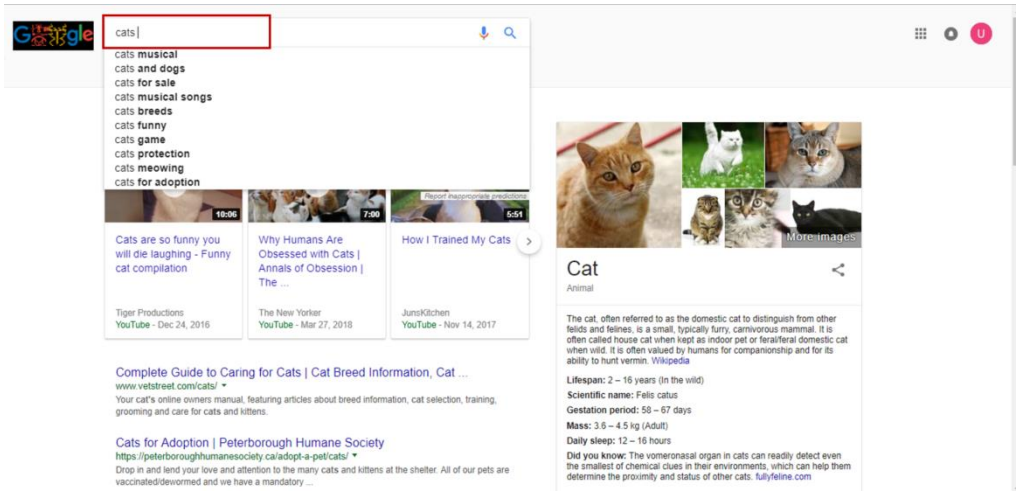
The drug discovery process is highly complicated and involves many disciplines. The greatest ideas are often bounded by billions of testing, huge financial and time expenditure. On average, it takes twelve years to make an official submission. Data science applications and machine learning algorithms simplify and shorten this process, adding a perspective to each step from the initial screening of drug compounds to the prediction of the success rate based on the biological factors. Such algorithms can forecast how the compound will act in the body using advanced mathematical modeling and simulations instead of the “lab experiments”. The idea behind the computational drug discovery is to create computer model simulations as a biologically relevant network simplifying the prediction of future outcomes with high accuracy.

d) Virtual assistance for patients and customer support

Optimization of the clinical process builds upon the concept that for many cases it is not actually necessary for patients to visit doctors in person. A mobile application can give a more effective solution by bringing the doctor to the patient instead. The AI-powered mobile apps can provide basic healthcare support, usually chatbots. You simply describe your symptoms, or ask questions, and then receive key information about your medical condition derived from a wide network linking symptoms to causes. Apps can remind you to take your medicine on time, and if necessary, assign an appointment with a doctor. This approach promotes a healthy lifestyle by encouraging patients to make healthy decisions, saves their time waiting in line for an appointment, and allows doctors to focus on more critical cases. The most popular applications nowadays are Your.MD and Ada.

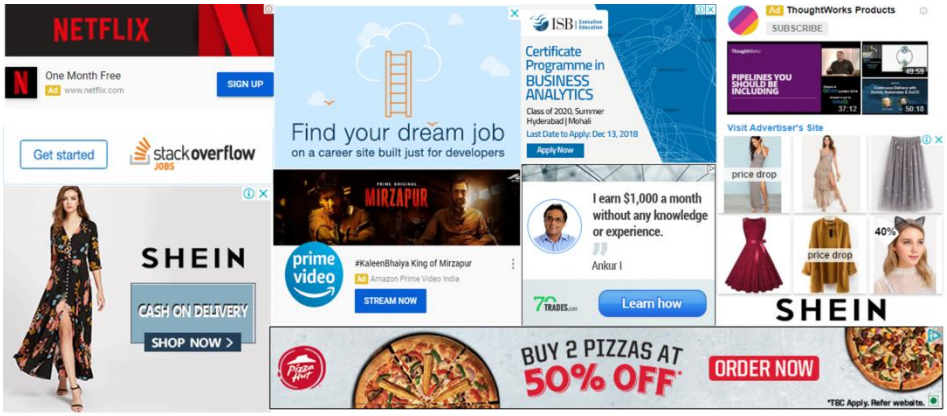
3. Internet Search

Now, this is probably the first thing that strikes your mind when you think Data Science Applications. When we speak of search, we think ‘Google’. Right? But there are many other search engines like Yahoo, Bing, Ask, AOL, and so on. All these search engines (including Google) make use of data science algorithms to deliver the best result for our searched query in a fraction of seconds. Considering the fact that, Google processes more than 20 petabytes of data every day. Had there been no data science, Google wouldn’t have been the ‘Google’ we know today.



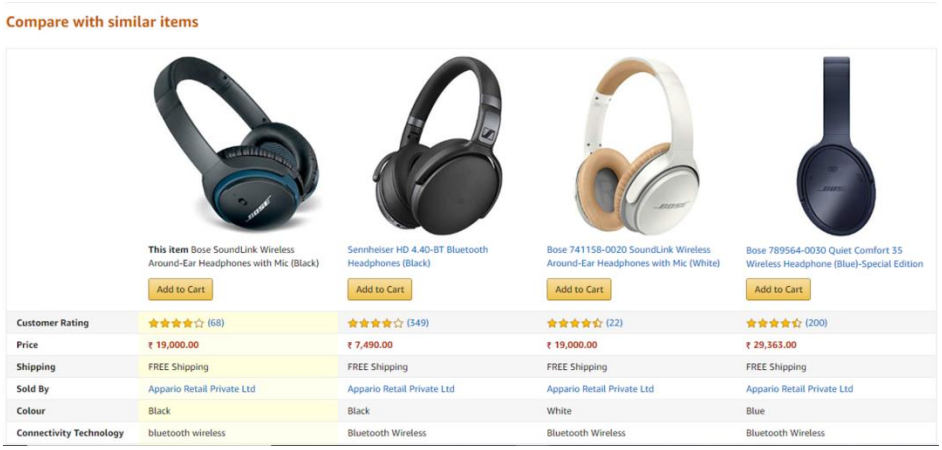
4. Targeted Advertising

If you thought Search would have been the biggest of all data science applications, here is a challenger – the entire digital marketing spectrum. Starting from the display banners on various websites to the digital billboards at the airports – almost all of them are decided by using data science algorithms. This is the reason why digital ads have been able to get a lot higher CTR (Call-Through Rate) than traditional advertisements. They can be targeted based on a user's past behavior. This is the reason why you might see ads of Data Science Training Programs while I see an ad of apparels in the same place at the same time.



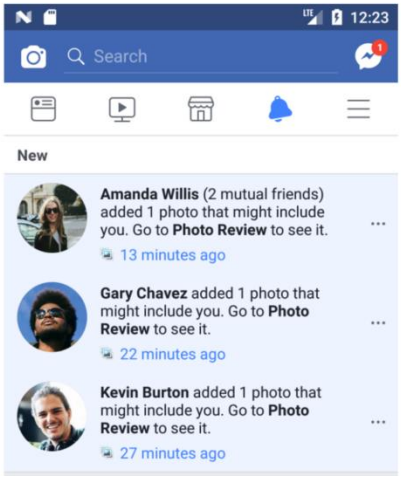
5. Website Recommendations

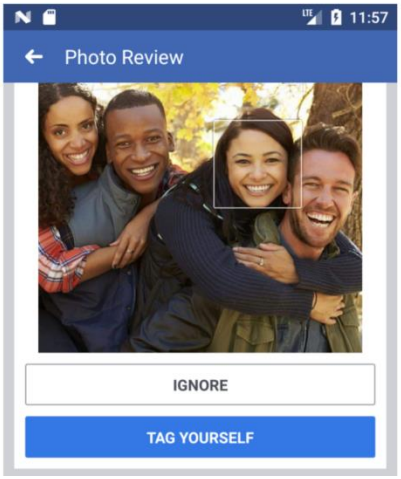
Aren't we all used to the suggestions about similar products on Amazon? They not only help you find relevant products from billions of products available with them but also add a lot to the user experience. A lot of companies have fervidly used this engine to promote their products in accordance with user's interest and relevance of information. Internet giants like Amazon, Twitter, Google Play, Netflix, Linkedin, IMDb, and much more use this system to improve the user experience. The recommendations are made based on previous search results for a user.



6. Advanced Image Recognition

You upload your image with friends on Facebook and you start getting suggestions to tag your friends. This automatic tag suggestion feature uses face recognition algorithm. In their latest update, Facebook has outlined the additional progress they've made in this area, making specific note of their advances in image recognition accuracy and capacity."We've witnessed massive advances in image classification (what is in the image?) as well as object detection (where are the objects?), but this is just the beginning of understanding the most relevant visual content of any image or video. Recently we've been designing techniques that identify and segment each and every object in an image, a key capability that will enable entirely new applications." In addition, Google provides you with the option to search for images by uploading them. It uses image recognition and provides related search results.





7. Speech Recognition

Some of the best examples of speech recognition products are Google Voice, Siri, Cortana etc. Using the speech-recognition feature, even if you aren't in a position to type a message, your life wouldn't stop. Simply speak out the message and it will be converted to text. However, at times, you would realize, speech recognition doesn't perform accurately.

8. Airline Route Planning

Airline Industry across the world is known to bear heavy losses. Except for a few airline service providers, companies are struggling to maintain their occupancy ratio and operating profits. With high rise in air-fuel prices and need to offer heavy discounts to customers has further made the situation worse. It wasn't for long when airlines companies started using data science to identify the strategic areas of improvements. Now using data science, the airline companies can:

- Predict flight delay
- Decide which class of airplanes to buy
- Whether to directly land at the destination or take a halt in between (For example, A flight can have a direct route from New Delhi to New York. Alternatively, it can also choose to halt in any country.)
- Effectively drive customer loyalty programs

Southwest Airlines, Alaska Airlines are among the top companies who've embraced data science to bring changes in their way of working.

9. Gaming

Games are now designed using machine learning algorithms that improve/upgrade themselves as the player moves up to a higher level. In motion gaming also, your opponent (computer) analyzes your previous moves and accordingly shapes up its game. EA Sports, Zynga, Sony, Nintendo, Activision-Blizzard have led the gaming experience to the next level using data science.

10. Augmented Reality

This is the final of the data science applications which seem most exciting in the future. Augmented reality.Data Science and Virtual Reality do have a relationship, considering a VR headset contains computing knowledge, algorithms and data to provide you with the best viewing experience. A very small step towards this is the high-trending game of Pokemon GO. The ability to walk around things

and look at Pokemon on walls, streets, things that aren't really there. The creators of this game used the data from Ingress, the last app from the same company, to choose the locations of the Pokemon and gyms

However, Data Science makes more sense once VR economy becomes accessible in terms of pricing, and consumers use it often like other apps. Though, not much has been revealed about them except the prototypes, and neither do we know when they would be available for a common man's disposal.

Data Security Issues

What is Data Security?

Data security is the process of protecting corporate data and preventing data loss through unauthorized access. This includes protecting your data from attacks that can encrypt or destroy data, such as ransomware, as well as attacks that can modify or corrupt your data. Data security also ensures data is available to anyone in the organization who has access to it.Some industries require a high level of data security to comply with data protection regulations. For example, organizations that process payment card information must use and store payment card data securely, and healthcare organizations in the USA must secure private health information (PHI) in line with the HIPAA standard.But even if your organization is not subject to a regulation or compliance standard, the survival of a modern business depends on data security, which can impact both the organization's key assets and private data belonging to its customers.

Why is Data Security Important?

The Ponemon Institute's Cost of Data Breach Study found that on average, the damage caused by a data breach in the USA was $8 million. 25,575 user accounts were impacted in the average data incident, which means that beyond financial losses, most incidents lead to loss of customer trust and damage to reputation.

Average cost of data breaches is the highest in the USA.Lawsuits, settlements, and fines related to data breaches are also on the rise, with many governments introducing more stringent regulations around data privacy. Consumers have much more extensive rights, especially in the EU, California, and Australia, with the introduction of GDPR, CCPA, APP, and CSP234.Companies operating in regulated industries are affected by additional standards, such as HIPAA for healthcare organizations in the USA, and PCI/DSS for organizations processing credit card data.

In the past decade, social engineering, ransomware and advanced persistent threats (APTs) are on the rise. These are threats that are difficult to defend against and can cause catastrophic damage to an organization's data.There is no simple solution to data security—just adding another security solution won't solve the problem. IT and information security teams must actively and creatively consider their data protection challenges and cooperate to improve their security posture. It is also critical to evaluate the cost of current security measures, their contribution to data security, and the expected return on investment from additional investments.

Data Security vs Data Privacy

Data privacy is the distinction between data in a computer system that can be shared with third parties (non-private data), and data that cannot be shared with third parties (private data). There are two main aspects to enforcing data privacy:

- **Access control**—ensuring that anyone who tries to access the data is authenticated to confirm their identity, and authorized to access only the data they are allowed to access.
- **Data protection**—ensuring that even if unauthorized parties manage to access the data, they cannot view it or cause damage to it. Data protection methods ensure encryption, which prevents anyone from viewing data if they do not have a private encryption key, and data loss prevention mechanisms which prevent users from transferring sensitive data outside the organization.

Data security has many overlaps with data privacy. The same mechanisms used to ensure data privacy are also part of an organization's data security strategy.The primary difference is that data privacy mainly focuses on keeping data confidential, while data security mainly focuses on protecting from malicious activity. For example, encryption could be a sufficient measure to protect privacy, but may not be sufficient as a data security measure. Attackers could still cause damage by erasing the data or double-encrypting it to prevent access by authorized parties.

Data Security Risks

Below are several common issues faced by organizations of all sizes as they attempt to secure sensitive data.

Accidental Exposure

A large percentage of data breaches are not the result of a malicious attack but are caused by negligent or accidental exposure of sensitive data. It is common for an organization's employees to share, grant access to, lose, or mishandle valuable data, either by accident or because they are not aware of security policies.This major problem can be addressed by employee training, but also by other measures, such as data loss prevention (DLP) technology and improved access controls.

Phishing and Other Social Engineering Attacks

Social engineering attacks are a primary vector used by attackers to access sensitive data. They involve manipulating or tricking individuals into providing private information or access to privileged accounts.Phishing is a common form of social engineering. It involves messages that appear to be from a trusted source, but in fact are sent by an attacker. When victims comply, for example by providing private information or clicking a malicious link, attackers can compromise their device or gain access to a corporate network.

Insider Threats

Insider threats are employees who inadvertently or intentionally threaten the security of an organization's data. There are three types of insider threats:

- **Non-malicious insider**—these are users that can cause harm accidentally, via negligence, or because they are unaware of security procedures.
- **Malicious insider**—these are users who actively attempt to steal data or cause harm to the organization for personal gain.
- **Compromised insider**—these are users who are not aware that their accounts or credentials were compromised by an external attacker. The attacker can then perform malicious activity, pretending to be a legitimate user.

Ransomware

Ransomware is a major threat to data in companies of all sizes. Ransomware is malware that infects corporate devices and encrypts data, making it useless without the decryption key. Attackers display a ransom message asking for payment to release the key, but in many cases, even paying the ransom is ineffective and the data is lost.Many types of ransomware can spread rapidly, and infect large parts of a corporate network. If an organization does not maintain regular backups, or if the ransomware manages to infect the backup servers, there may be no way to recover.

Data Loss in the Cloud

Many organizations are moving data to the cloud to facilitate easier sharing and collaboration. However, when data moves to the cloud, it is more difficult to control and prevent data loss. Users access data from personal devices and over unsecured networks. It is all too easy to share a file with unauthorized parties, either accidentally or maliciously.

SQL Injection

SQL injection (SQLi) is a common technique used by attackers to gain illicit access to databases, steal data, and perform unwanted operations. It works by adding malicious code to a seemingly innocent database query.SQL injection manipulates SQL code by adding special characters to a user input that change the context of the query. The database expects to process a user input, but instead starts processing malicious code that advances the attacker's goals. SQL injection can expose customer data, intellectual property, or give attackers administrative access to a database, which can have severe consequences.SQL injection vulnerabilities are typically the result of insecure coding practices. It is relatively easy to prevent SQL injection if coders use secure mechanisms for accepting user inputs, which are available in all modern database systems.

Common Data Security Solutions and Techniques

There are several technologies and practices that can improve data security. No one technique can solve the problem, but by combining several of the techniques below, organizations can significantly improve their security posture.

Data Discovery and Classification

Modern IT environments store data on servers, endpoints, and cloud systems. Visibility over data flows is an important first step in understanding what data is at risk of being stolen or misused. To properly protect your data, you need to know the type of data, where it is, and what it is used for. Data discovery and classification tools can help.Data detection is the basis for knowing what data you have. Data classification allows you to create scalable security solutions, by identifying which data is sensitive and needs to be secured. Data detection and classification solutions enable tagging files on endpoints, file servers, and cloud storage systems, letting you visualize data across the enterprise, to apply the appropriate security policies.

Data Masking

Data masking lets you create a synthetic version of your organizational data, which you can use for software testing, training, and other purposes that don't require the real data. The goal is to protect data while providing a functional alternative when needed.Data masking retains the data type, but changes the values. Data can be modified in a number of ways, including encryption, character shuffling, and character or word substitution. Whichever method you choose; you must change the values in a way that cannot be reverse-engineered.

Identity Access Management

Identity and Access Management (IAM) is a business process, strategy, and technical framework that enables organizations to manage digital identities. IAM solutions allow IT administrators to control user access to sensitive information within an organization.Systems used for IAM include single sign-on systems, two-factor authentication, multi-factor authentication, and privileged access management. These technologies enable the organization to securely store identity and profile data, and support governance, ensuring that the appropriate access policies are applied to each part of the infrastructure.

Data Encryption

Data encryption is a method of converting data from a readable format (plaintext) to an unreadable encoded format (ciphertext). Only after decrypting the encrypted data using the decryption key, the data can be read or processed.In public-key cryptography techniques, there is no need to share the decryption key – the sender and recipient each have their own key, which are combined to perform the encryption operation. This is inherently more secure.Data encryption can prevent hackers from accessing sensitive information. It is essential for most security strategies and is explicitly required by many compliance standards.

Data Loss Prevention (DLP)

To prevent data loss, organizations can use a number of safeguards, including backing up data to another location. Physical redundancy can help protect data from natural disasters, outages, or attacks on local servers. Redundancy can be performed within a local data center, or by replicating data to a remote site or cloud environment.Beyond basic measures like backup, DLP software solutions can help protect organizational data. DLP software automatically analyzes content to identify sensitive data, enabling central control and enforcement of data protection policies, and alerting in real-time when it detects anomalous use of sensitive data, for example, large quantities of data copied outside the corporate network.

Governance, Risk, and Compliance (GRC)

GRC is a methodology that can help improve data security and compliance:

- Governance creates controls and policies enforced throughout an organization to ensure compliance and data protection.
- Risk involves assessing potential cybersecurity threats and ensuring the organization is prepared for them.
- Compliance ensures organizational practices are in line with regulatory and industry standards when processing, accessing, and using data.

Password Hygiene

One of the simplest best practices for data security is ensuring users have unique, strong passwords. Without central management and enforcement, many users will use easily guessable passwords or use the same password for many different services. Password spraying and other brute force attacks can easily compromise accounts with weak passwords.A simple measure is enforcing longer passwords and asking users to change passwords frequently. However, these measures are not enough, and organizations should consider multi-factor authentication (MFA) solutions that require users to identify themselves with a token or device they own, or via

biometric means.Another complementary solution is an enterprise password manager that stores employee passwords in encrypted form, reducing the burden of remembering passwords for multiple corporate systems, and making it easier to use stronger passwords. However, the password manager itself becomes a security vulnerability for the organization.

Authentication and Authorization

Organizations must put in place strong authentication methods, such as OAuth for web-based systems. It is highly recommended to enforce multi-factor authentication when any user, whether internal or external, requests sensitive or personal data.In addition, organizations must have a clear authorization framework in place, which ensures that each user has exactly the access rights they need to perform a function or consume a service, and no more. Periodic reviews and automated tools should be used to clean up permissions and remove authorization for users who no longer need them.

Data Security Audits

The organization should perform security audits at least every few months. This identifies gaps and vulnerabilities across the organizations' security posture. It is a good idea to perform the audit via a third-party expert, for example in a penetration testing model. However, it is also possible to perform a security audit in house. Most importantly, when the audit exposes security issues, the organization must devote time and resources to address and remediate them.

Anti-Malware, Antivirus, and Endpoint Protection

Malware is the most common vector of modern cyberattacks, so organizations must ensure that endpoints like employee workstations, mobile devices, servers, and cloud systems, have appropriate protection. The basic measure is antivirus software, but this is no longer enough to address new threats like file-less attacks and unknown zero-day malware.Endpoint protection platforms (EPP) take a more comprehensive approach to endpoint security. They combine antivirus with a machine-learning-based analysis of anomalous behavior on the device, which can help detect unknown attacks. Most platforms also provide endpoint detection and response (EDR) capabilities, which help security teams identify breaches on endpoints as they happen, investigate them, and respond by locking down and reimaging affected endpoints.

Zero Trust

Zero trust is a security model introduced by Forrester analyst John Kindervag, which has been adopted by the US government, several technical standards bodies, and many of the world's largest technology companies. The basic principle of zero trust is that no entity on a network should be trusted, regardless of whether it is outside or inside the network perimeter.Zero trust has a special focus on data security, because data is the primary asset attackers are interested in. A zero trust architecture aims to protect data against insider and outside threats by continuously verifying all access attempts, and denying access by default.Zero trust security mechanisms build multiple security layers around sensitive data—for example, they use micro-segmentation to ensure sensitive assets on the network are isolated from other assets. In a true zero trust network, attackers have very limited access to sensitive data, and there are controls that can help detect and respond to any anomalous access to data.

Database Security

Database security involves protecting database management systems such as Oracle, SQL Server, or MySQL, from unauthorized use and malicious cyberattacks. The main elements protected by database security are:

- The database management system (DBMS).
- Data stored in the database.
- Applications associated with the DBMS.
- The physical or virtual database server and any underlying hardware.
- Any computing and network infrastructure used to access the database.

A database security strategy involves tools, processes, and methodologies to securely configure and maintain security inside a database environment and protect databases from intrusion, misuse, and damage.

Big Data Security

Big data security involves practices and tools used to protect large datasets and data analysis processes. Big data commonly takes the form of financial logs, healthcare data, data lakes, archives, and business intelligence datasets. Within the big data perimeter there are three primary scenarios that require protection: inbound data transfers, outbound data transfers, and data at rest.Big data security aims to prevent accidental and intentional breaches, leaks, losses, and exfiltration of large amounts of data. Let's review popular big data services and see the main strategies for securing them.

AWS Big Data

AWS offers analytics solutions for big data implementations. There are various services AWS offers to automate data analysis, manipulate datasets, and derive insights, including Amazon Simple Storage Service (S3), Amazon Kinesis, Amazon Elastic Map/Reduce (EMR), and Amazon Glue.AWS big data security best practices include:

- **Access policy options**—use access policy options to manage access to your S3 resources.
- **Data encryption policy**—use Amazon S3 and AWS KMS for encryption management.
- **Manage data with object tagging**—categorize and manage S3 data assets using tags, and apply tags indicating sensitive data that requires special security measures.

Azure Big Data

Microsoft Azure cloud offers big data and analytics services that can process a high volume of structured and unstructured data. The platform offers elastic storage using Azure storage services, real-time analytics, database services, as well as machine learning and data engineering solutions.Azure big data security best practices include:

- Monitor as many processes as possible.
- Leverage Azure Monitor and Log Analytics to gain visibility over data flows.
- Define and enforce a security and privacy policy.
- Leverage Azure services for backup, restore, and disaster recovery.

Google Cloud Big Data

The Google Cloud Platform offers multiple services that support big data storage and analysis. BigQuery is a high-performance SQL-compatible engine, which can perform analysis on large data volumes in seconds. Additional services include Dataflow, Dataproc, and Data Fusion.Google Cloud big data security best practices include:

- Define BigQuery access controls according to the least privilege principle.
- Use policy tags or type-based classification to identify sensitive data.
- Leverage column-level security to check if a user has the right to view specific data at query time.

Snowflake

Snowflake is a cloud data warehouse for enterprises, built for high performance big data analytics. The architecture of Snowflake physically separates compute and storage, while integrating them logically. Snowflake offers full relational database support and can work with structured and semi-structured data.Snowflake security best practices include:

- Define network and site access through IP allow/block lists.
- Use SCIM to manage user identities and groups.
- Leverage key pair authentication and rotation to improve client authentication security.
- Enable multi-factor authentication.

Elasticsearch

Elasticsearch is an open-source full-text search and analytics engine that is highly scalable, allowing search and analytics on big data in real-time. It powers applications with complex search requirements. Elasticsearch provides a distributed system on top of Lucene StandardAnalyzer for indexing and automatic type prediction, and utilizes a JSON-based REST API to Lucene features.Elasticsearch security best practices include:

- Use strong passwords to protect access to search clusters
- Encrypt all communications using SSL/TLS

- Leverage role-based access control (RBAC)
- Use IP filtering for client access
- Turn on auditing and monitor logs on an regular basis

Splunk

Splunk is a software platform that indexes machine data, makes it searchable and turns it into actionable intelligence. It pulls log files from applications, servers, mobile devices, and websites, aggregates them, and provides rich analysis features.Splunk security best practices include:

- Preventing unauthorized access by defining RBAC, data encryption, and obfuscation of credentials.
- Using SSL/TLS encryption for data ingestion and internal Splunk communications.
- Hardening Splunk instances by ensuring they are physically secure and do not store secrets in plaintext.
- Using audit events to track any changes to Splunk system configuration.

Securing Data in Enterprise Applications

Enterprise applications power mission critical operations in organizations of all sizes. Enterprise application security aims to protect enterprise applications from external attacks, abuse of authority, and data theft.

Email Security

Email security is the process of ensuring the availability, integrity, and reliability of email communications by protecting them from cyber threats.Technical standards bodies have recommended email security protocols including SSL/TLS, Sender Policy Framework (SPF), and DomainKeys Identified Mail (DKIM). These protocols are implemented by email clients and servers, including Microsoft Exchange and Google G Suite, to ensure secure delivery of emails. A secure email gateway helps organizations and individuals protect their email from a variety of threats, in addition to implementing security protocols.

ERP Security

Enterprise Resource Planning (ERP) is software designed to manage and integrate the functions of core business processes such as finance, human resources, supply chain, and inventory management into one system. ERP systems store highly sensitive information and are, by definition, a mission critical system.ERP security is a broad set of measures designed to protect an ERP system from unauthorized access and ensure the accessibility and integrity of system data. The Information Systems Audit and Control Association (ISACA) recommends regularly performing security assessments of ERP systems, including software vulnerabilities, misconfigurations, separation of duties (SoD) conflicts, and compliance with vendor security recommendations.

DAM Security

Digital Asset Management (DAM) is a technology platform and business process for organizing, storing, and acquiring rich media and managing digital rights and licenses. Rich media assets include photos, music, videos, animations, podcasts, and other multimedia content. Data stored in DAM systems is sensitive because it often represents company IP, and is used in critical processes like sales, marketing, and delivery of media to viewers and web visitors.Security best practices for DAM include:

- Implement the principle of least privilege.
- Use an allowlist for file destinations.
- Use multi-factor authentication to control access by third parties.
- Regularly review automation scripts, limit privileges of commands used, and control the automation process through logging and alerting.

CRM Security

Customer Relationship Management (CRM) is a combination of practices, strategies, and technologies that businesses use to manage and analyze customer interactions and data throughout the customer lifecycle. CRM data is highly sensitivebecause it can expose an organization’s most valuable asset—customer relationships. CRM data is also personally identifiable information (PII) and is subject to data privacy regulations.Security best practices for CRM include:

- Perform period IT risk assessment audits for CRM systems.
- Perform CRM activity monitoring to identify unusual or suspicious usage.
- Encourage CRM administrators to follow security best practices.
- Educate CRM users on security best practices.
- If you operate CRM as SaaS, perform due diligence of the SaaS provider's security practices.

Data collection strategies

The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data Collection. Knowledge is power, information is knowledge, and data is information in digitized form, at least as defined in IT. Hence, data is power. But before you can leverage that data into a successful strategy for your organization or business, you need to gather it. That's your first step. So, to help you get the process started, we shine a spotlight on data collection. What exactly is it? Believe it or not, it's more than just doing a Google search! Furthermore, what are the different types of data collection? And what kinds of data collection tools and data collection techniques exist?

What is Data Collection: A Definition

Data collection is the process of gathering, measuring, and analyzing accurate data from a variety of relevant sources to find answers to research problems, answer questions, evaluate outcomes, and forecast trends and probabilities. Our society is highly dependent on data, which underscores the importance of collecting it. Accurate data collection is necessary to make informed business decisions, ensure quality assurance, and keep research integrity. During data collection, the researchers must identify the data types, the sources of data, and what methods are being used. We will soon see that there are many different data collection methods. There is heavy reliance on data collection in research, commercial, and government fields. Before an analyst begins collecting data, they must answer three questions first:

- What's the goal or purpose of this research?
- What kinds of data are they planning on gathering?
- What methods and procedures will be used to collect, store, and process the information?

Additionally, we can break up data into qualitative and quantitative types. Qualitative data covers descriptions such as color, size, quality, and appearance. Quantitative data, unsurprisingly, deals with numbers, such as statistics, poll numbers, percentages, etc.

Why Do We Need Data Collection?

Before a judge makes a ruling in a court case or a general creates a plan of attack, they must have as many relevant facts as possible. The best courses of action come from informed decisions, and information and data are synonymous. The concept of data collection isn't a new one, as we'll see later, but the world has changed. There is far more data available today, and it exists in forms that were unheard of a century ago. The data collection process has had to change and grow with the times, keeping pace with technology. Whether you're in the world of academia, trying to conduct research, or part of the commercial sector, thinking of how to promote a new product, you need data collection to help you make better choices.

What Are the Different Methods of Data Collection?

Now that you know what is data collection, let's take a look at the different methods of data collection. While the phrase "data collection" may sound all high-tech and digital, it doesn't necessarily entail things like computers, big data, and the internet. Data collection could mean a telephone survey, a mail-in comment card, or even some guy with a clipboard asking passersby some questions. But let's see if we can sort the different data collection methods into a semblance of organized categories.

Data collection breaks down into two methods. As a side note, many terms, such as techniques, methods, and types, are interchangeable and depending on who uses them. One source may call data collection techniques "methods," for instance. But whatever labels we use, the general concepts and breakdowns apply across the board whether we're talking about marketing analysis or a scientific research project.

The two methods are:

1. **Primary.**

As the name implies, this is original, first-hand data collected by the data researchers. This process is the initial information gathering step, performed before anyone carries out any further or related research. Primary data results are highly accurate provided the researcher collects the information. However, there's a downside, as first-hand research is potentially time-consuming and expensive.

2. **Secondary.**
   Secondary data is second-hand data collected by other parties and already having undergone statistical analysis. This data is either information that the researcher has tasked other people to collect or information the researcher has looked up. Simply put, it's second-hand information. Although it's easier and cheaper to obtain than primary information, secondary information raises concerns regarding accuracy and authenticity. Quantitative data makes up a majority of secondary data.

Primary Data Collection Techniques

I. **Interviews. -** The researcher asks questions of a large sampling of people, either by direct interviews or means of mass communication such as by phone or mail. This method is by far the most common means of data gathering.
II. **Projective Technique. -** Projective data gathering is an indirect interview, used when potential respondents know why they're being asked questions and hesitate to answer. For instance, someone may be reluctant to answer questions about their phone service if a cell phone carrier representative poses the questions. With projective data gathering, the interviewees get an incomplete question, and they must fill in the rest, using their opinions, feelings, and attitudes.
III. **Delphi Technique. -** The Oracle at Delphi, according to Greek mythology, was the high priestess of Apollo's temple, who gave advice, prophecies, and counsel. In the realm of data collection, researchers use the Delphi technique by gathering information from a panel of experts. Each expert answers questions in their field of specialty, and the replies are consolidated into a single opinion.
IV. **Focus Groups. -** Focus groups, like interviews, are a commonly used technique. The group consists of anywhere from a half-dozen to a dozen people, led by a moderator, brought together to discuss the issue.
V. **Questionnaires. -** Questionnaires are a simple, straightforward data collection method. Respondents get a series of questions, either open or close-ended, related to the matter at hand.

Secondary Data Collection Techniques

Unlike primary data collection, there are no specific collection methods. Instead, since the information has already been collected, the researcher consults various data sources, such as:

I. Financial Statements
II. Sales Reports
III. Retailer/Distributor/Deal Feedback
IV. Customer Personal Information (e.g., name, address, age, contact info)
V. Business Journals
VI. Government Records (e.g., census, tax records, Social Security info)
VII. Trade/Business Magazines
VIII. The internet

Data Collection Tools

Now that we've explained the various techniques, let's narrow our focus even further by looking at some specific tools. For example, we mentioned interviews as a technique, but we can further break that down into different interview types (or "tools").

- **Word Association. -** The researcher gives the respondent a set of words and asks them what comes to mind when they hear each word.
- **Sentence Completion. -** Researchers use sentence completion to understand what kind of ideas the respondent has. This tool involves giving an incomplete sentence and seeing how the interviewee finishes it.
- **Role-Playing. -** Respondents are presented with an imaginary situation and asked how they would act or react if it was real.
- **In-Person Surveys. -** The researcher asks questions in person.
- **Online/Web Surveys. -** These surveys are easy to accomplish, but some users may be unwilling to answer truthfully, if at all.

- **Mobile Surveys. -** These surveys take advantage of the increasing proliferation of mobile technology. Mobile collection surveys rely on mobile devices like tablets or smartphones to conduct surveys via SMS or mobile apps.
- **Phone Surveys. -** No researcher can call thousands of people at once, so they need a third party to handle the chore. However, many people have call screening and won't answer.
- **Observation. -** Sometimes, the simplest method is the best. Researchers who make direct observations collect data quickly and easily, with little intrusion or third-party bias. Naturally, it's only effective in small-scale situations.

The Importance of Ensuring Accurate and Appropriate Data Collection

Accurate data collecting is crucial to preserving the integrity of research, regardless of the subject of study or preferred method for defining data (quantitative, qualitative). Errors are less likely to occur when the right data gathering tools are used (whether they are brand-new ones, updated versions of them, or already available). Among the effects of data collection done incorrectly, include the following -

- Erroneous conclusions that squander resources
- Decisions that compromise public policy
- Incapacity to correctly respond to research inquiries
- Bringing harm to participants who are humans or animals
- Deceiving other researchers into pursuing futile research avenues
- The study's inability to be replicated and validated

When these study findings are used to support recommendations for public policy, there is the potential to result in disproportionate harm, even if the degree of influence from flawed data collecting may vary by discipline and the type of investigation.

Issues Related to Maintaining the Integrity of Data Collection

In order to assist the errors detection process in the data gathering process, whether they were done purposefully (deliberate falsifications) or not, maintaining data integrity is the main justification (systematic or random errors). Quality assurance and quality control are two strategies that help protect data integrity and guarantee the scientific validity of study results. Each strategy is used at various stages of the research timeline:

- Quality control - tasks that are performed both after and during data collecting
- Quality assurance - events that happen before data gathering starts

Quality Assurance

As data collecting comes before quality assurance, its primary goal is "prevention" (i.e., forestalling problems with data collection). The best way to protect the accuracy of data collection is through prevention. The uniformity of protocol created in the thorough and exhaustive procedures manual for data collecting serves as the best example of this proactive step. The likelihood of failing to spot issues and mistakes early in the research attempt increases when guides are written poorly. There are several ways to show these shortcomings:

- Failure to determine the precise subjects and methods for retraining or training staff employees in data collecting
- List of goods to be collected, in part
- There isn't a system in place to track modifications to processes that may occur as the investigation continues.
- Instead of detailed, step-by-step instructions on how to deliver tests, there is a vague description of the data gathering tools that will be employed.
- Uncertainty regarding the date, procedure, and identity of the person or people in charge of examining the data
- Incomprehensible guidelines for using, adjusting, and calibrating the data collection equipment.

Quality Control

Despite the fact that quality control actions (detection/monitoring and intervention) take place both after and during data collection, the specifics should be meticulously detailed in the procedures manual. Establishing monitoring systems requires a specific communication structure, which is a prerequisite. Following the discovery of data collection problems, there should be no ambiguity

regarding the information flow between the primary investigators and staff personnel. A poorly designed communication system promotes slack oversight and reduces opportunities for error detection.

Direct staff observation conference calls, during site visits, or frequent or routine assessments of data reports to spot discrepancies, excessive numbers, or invalid codes can all be used as forms of detection or monitoring. Site visits might not be appropriate for all disciplines. Still, without routine auditing of records, whether qualitative or quantitative, it will be challenging for investigators to confirm that data gathering is taking place in accordance with the manual's defined methods. Additionally, quality control determines the appropriate solutions, or "actions," to fix flawed data gathering procedures and reduce recurrences. Problems with data collection, for instance, that call for immediate action include:

- Fraud or misbehavior
- Systematic mistakes, procedure violations
- Individual data items with errors
- Issues with certain staff members or a site's performance

Researchers are trained to include one or more secondary measures that can be used to verify the quality of information being obtained from the human subject in the social and behavioral sciences where primary data collection entails using human subjects. For instance, a researcher conducting a survey would be interested in learning more about the prevalence of risky behaviors among young adults as well as the social factors that influence these risky behaviors' propensity for and frequency.

What are Common Challenges in Data Collection?

There are some prevalent challenges faced while collecting data, let us explore a few of them to understand them better and avoid them.

- **Data Quality Issues -** The main threat to the broad and successful application of machine learning is poor data quality. Data quality must be your top priority if you want to make technologies like machine learning work for you. Let's talk about some of the most prevalent data quality problems in this blog article and how to fix them.
- **Inconsistent Data -** When working with various data sources, it's conceivable that the same information will have discrepancies between sources. The differences could be in formats, units, or occasionally spellings. The introduction of inconsistent data might also occur during firm mergers or relocations. Inconsistencies in data have a tendency to accumulate and reduce the value of data if they are not continually resolved. Organizations that have heavily focused on data consistency do so because they only want reliable data to support their analytics.
- **Data Downtime -** Data is the driving force behind the decisions and operations of data-driven businesses. However, there may be brief periods when their data is unreliable or not prepared. Customer complaints and subpar analytical outcomes are only two ways that this data unavailability can have a significant impact on businesses. A data engineer spends about 80% of their time updating, maintaining, and guaranteeing the integrity of the data pipeline. In order to ask the next business question, there is a high marginal cost due to the lengthy operational lead time from data capture to insight. Schema modifications and migration problems are just two examples of the causes of data downtime. Data pipelines can be difficult due to their size and complexity. Data downtime must be continuously monitored, and it must be reduced through automation.
- **Ambiguous Data -** Even with thorough oversight, some errors can still occur in massive databases or data lakes. For data streaming at a fast speed, the issue becomes more overwhelming. Spelling mistakes can go unnoticed, formatting difficulties can occur, and column heads might be deceptive. This unclear data might cause a number of problems for reporting and analytics.
- **Duplicate Data -** Streaming data, local databases, and cloud data lakes are just a few of the sources of data that modern enterprises must contend with. They might also have application and system silos. These sources are likely to duplicate and overlap each other quite a bit. For instance, duplicate contact information has a substantial impact on customer experience. If certain prospects are ignored while others are engaged repeatedly, marketing campaigns suffer. The likelihood of biased analytical outcomes increases when duplicate data are present. It can also result in ML models with biased training data.
- **Too Much Data -** While we emphasize data-driven analytics and its advantages, a data quality problem with excessive data exists. There is a risk of getting lost in an abundance of data when searching for information pertinent to your analytical efforts. Data scientists, data analysts, and business users devote 80% of their work to finding and organizing the appropriate

data. With an increase in data volume, other problems with data quality become more serious, particularly when dealing with streaming data and big files or databases.

- **Inaccurate Data -** For highly regulated businesses like healthcare, data accuracy is crucial. Given the current experience, it is more important than ever to increase the data quality for COVID-19 and later pandemics. Inaccurate information does not provide you with a true picture of the situation and cannot be used to plan the best course of action. Personalized customer experiences and marketing strategies underperform if your customer data is inaccurate. Data inaccuracies can be attributed to a number of things, including data degradation, human mistake, and data drift. Worldwide data decay occurs at a rate of about 3% per month, which is quite concerning. Data integrity can be compromised while being transferred between different systems, and data quality might deteriorate with time.
- **Hidden Data -** The majority of businesses only utilize a portion of their data, with the remainder sometimes being lost in data silos or discarded in data graveyards. For instance, the customer service team might not receive client data from sales, missing an opportunity to build more precise and comprehensive customer profiles. Missing out on possibilities to develop novel products, enhance services, and streamline procedures is caused by hidden data.
- **Finding Relevant Data -** Finding relevant data is not so easy. There are several factors that we need to consider while trying to find relevant data, which include -
    - Relevant Domain
    - Relevant demographics
    - Relevant Time period and so many more factors that we need to consider while trying to find relevant data.

  Data that is not relevant to our study in any of the factors render it obsolete and we cannot effectively proceed with its analysis. This could lead to incomplete research or analysis, re-collecting data again and again, or shutting down the study.

- **Deciding the Data to Collect -** Determining what data to collect is one of the most important factors while collecting data and should be one of the first factors while collecting data. We must choose the subjects the data will cover, the sources we will be used to gather it, and the quantity of information we will require. Our responses to these queries will depend on our aims, or what we expect to achieve utilizing your data. As an illustration, we may choose to gather information on the categories of articles that website visitors between the ages of 20 and 50 most frequently access. We can also decide to compile data on the typical age of all the clients who made a purchase from your business over the previous month. Not addressing this could lead to double work and collection of irrelevant data or ruining your study as a whole.
- **Dealing With Big Data -** Big data refers to exceedingly massive data sets with more intricate and diversified structures. These traits typically result in increased challenges while storing, analyzing, and using additional methods of extracting results. Big data refers especially to data sets that are quite enormous or intricate that conventional data processing tools are insufficient. The overwhelming amount of data, both unstructured and structured, that a business faces on a daily basis. The amount of data produced by healthcare applications, the internet, social networking sites social, sensor networks, and many other businesses are rapidly growing as a result of recent technological advancements. Big data refers to the vast volume of data created from numerous sources in a variety of formats at extremely fast rates. Dealing with this kind of data is one of the many challenges of Data Collection and is a crucial step toward collecting effective data.
- **Low Response and Other Research Issues -** Poor design and low response rates were shown to be two issues with data collecting, particularly in health surveys that used questionnaires. This might lead to an insufficient or inadequate supply of data for the study. Creating an incentivized data collection program might be beneficial in this case to get more responses.

Steps in the Data Collection Process

In the Data Collection Process, there are 5 key steps. They are explained briefly below -

1. Decide What Data You Want to Gather - The first thing that we need to do is decide what information we want to gather. We must choose the subjects the data will cover, the sources we will use to gather it, and the quantity of information that we would require. For instance, we may choose to gather information on the categories of products that an average e-commerce website visitor between the ages of 30 and 45 most frequently searches for.

2. Establish a Deadline for Data Collection - The process of creating a strategy for data collection can now begin. We should set a deadline for our data collection at the outset of our planning phase. Some forms of data we might want to continuously collect. We might want to build up a technique for tracking transactional data and website visitor statistics over the long term, for instance.

However, we will track the data throughout a certain time frame if we are tracking it for a particular campaign. In these situations, we will have a schedule for when we will begin and finish gathering data.

3. Select a Data Collection Approach - We will select the data collection technique that will serve as the foundation of our data gathering plan at this stage. We must take into account the type of information that we wish to gather, the time period during which we will receive it, and the other factors we decide on to choose the best gathering strategy.

4. Gather Information - Once our plan is complete, we can put our data collection plan into action and begin gathering data. In our DMP, we can store and arrange our data. We need to be careful to follow our plan and keep an eye on how it's doing. Especially if we are collecting data regularly, setting up a timetable for when we will be checking in on how our data gathering is going may be helpful. As circumstances alter and we learn new details, we might need to amend our plan.

5. Examine the Information and Apply Your Findings - It's time to examine our data and arrange our findings after we have gathered all of our information. The analysis stage is essential because it transforms unprocessed data into insightful knowledge that can be applied to better our marketing plans, goods, and business judgments. The analytics tools included in our DMP can be used to assist with this phase. We can put the discoveries to use to enhance our business once we have discovered the patterns and insights in our data.

Data Collection Considerations and Best Practices

We must carefully plan before spending time and money traveling to the field to gather data. While saving time and resources, effective data collection strategies can help us collect richer, more accurate, and richer data. Below, we will be discussing some of the best practices that we can follow for the best results -

1. Take Into Account the Price of Each Extra Data Point - Once we have decided on the data we want to gather, we need to make sure to take the expense of doing so into account. Our surveyors and respondents will incur additional costs for each additional data point or survey question.

2. Plan How to Gather Each Data Piece - There is a dearth of freely accessible data. Sometimes the data is there, but we may not have access to it. For instance, unless we have a compelling cause, we cannot openly view another person's medical information. It could be challenging to measure several types of information. Consider how time-consuming and difficult it will be to gather each piece of information while deciding what data to acquire.

3. Think About Your Choices for Data Collecting Using Mobile Devices - Mobile-based data collecting can be divided into three categories -

- IVRS (interactive voice response technology) - Will call the respondents and ask them questions that have already been recorded.
- SMS data collection - Will send a text message to the respondent, who can then respond to questions by text on their phone.
- Field surveyors - Can directly enter data into an interactive questionnaire while speaking to each respondent, thanks to smartphone apps.

We need to make sure to select the appropriate tool for our survey and responders because each one has its own disadvantages and advantages.

4. Carefully Consider the Data You Need to Gather - It's all too easy to get information about anything and everything, but it's crucial to only gather the information that we require. It is helpful to consider these 3 questions:

- What details will be helpful?
- What details are available?
- What specific details do you require?

5. Remember to Consider Identifiers - Identifiers, or details describing the context and source of a survey response, are just as crucial as the information about the subject or program that we are actually researching. In general, adding more identifiers will enable us to pinpoint our program's successes and failures with greater accuracy, but moderation is the key.

6. Data Collecting Through Mobile Devices is the Way to Go - Although collecting data on paper is still common, modern technology relies heavily on mobile devices. They enable us to gather many various types of data at relatively lower prices and are accurate as well as quick. There aren't many reasons not to pick mobile-based data collecting with the boom of low-cost Android devices that are available nowadays.

Data Categororization

Types of Data in Statistics

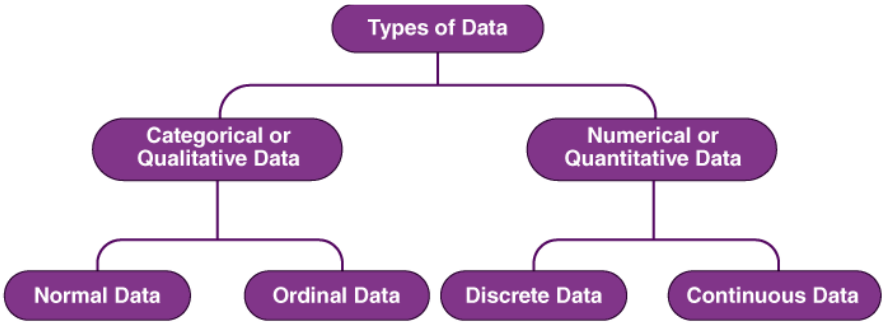
There are different types of data in Statistics, that are collected, analyzed, interpreted and presented. The data are the individual pieces of factual information recorded, and it is used for the purpose of the analysis process. The two processes of data analysis are interpretation and presentation. Statistics are the result of data analysis. Data classification and data handling are important processes as it involves a multitude of tags and labels to define the data, its integrity and confidentiality. The different types of data in statistics are

What are Types of Data in Statistics?

The data is classified into majorly four categories:

- Nominal data
- Ordinal data
- Discrete data
- Continuous data

Further, we can classify these data as follows:



Let us discuss the different types of data in Statistics herewith examples.

Qualitative or Categorical Data

Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person's gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have a mathematical sense. Examples of the categorical data are birthdate, favourite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning.

Nominal Data

- Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.
- The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.

Ordinal Data

- Ordinal data/variable is a type of data that follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on.
- The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualization tools. The information may be expressed using tables in which each row in the table shows the distinct category.

Quantitative or Numerical Data

Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on. The quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.

Discrete Data

- Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in whole numbers.
- Example: Number of students in the class

Continuous Data

- Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range.
- Example: Temperature range

Data in Data Analytics

- Entity: A particular thing is called entity or object.
- Attribute. An attribute is a measurable or observable property of an entity.
- Data. A measurement of an attribute is called data.

Note

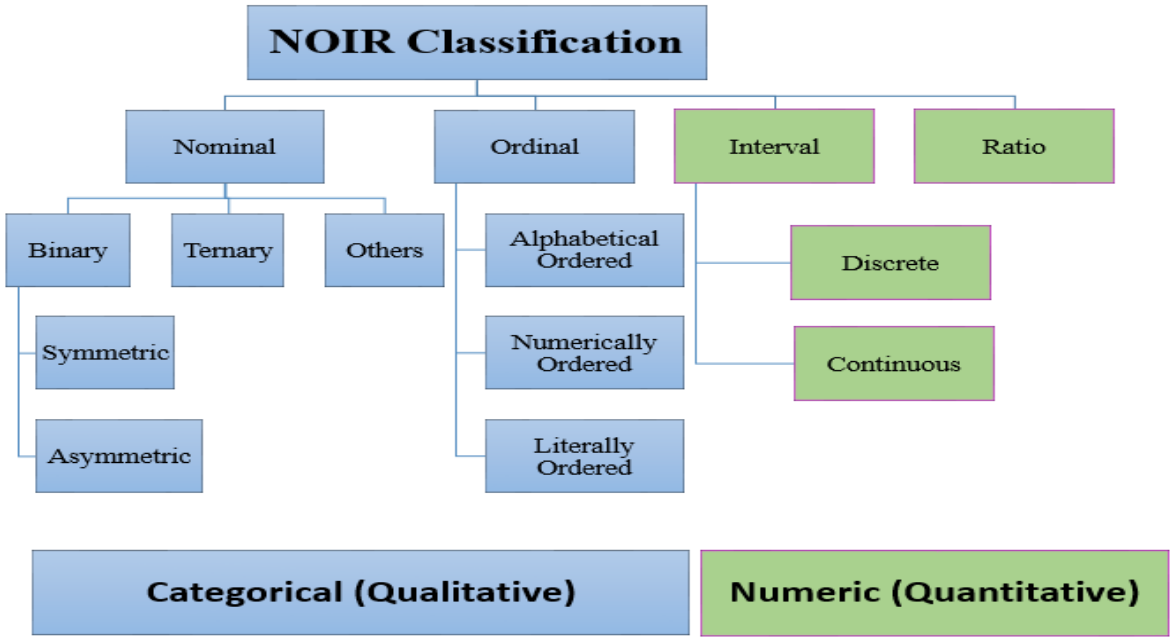
I. Data defines an entity.
II. Computer can manage all type of data (e.g., audio, video, text, etc.).

In general, there are many types of data that can be used to measure the properties of an entity. A good understanding of data scales (also called scales of measurement) is important. Depending the scales of measurement, different technique is followed to derive hitherto unknown knowledge in the form of patterns, associations, anomalies or similarities from a volume of data.

NOIR classification

The NOIR scale is the fundamental building block on which the extended data types are built. The mostly recommended scales of measurement are

N: Nominal **O:** Ordinal **I:** Interval **R:** Ratio



Properties of data

Following FOUR properties (operations) of data are pertinent.

#	Property	Operation	Type
1.	Distinctiveness	= and ≠	Categorical (Qualitative)
2.	Order	< , ≤ , > , ≥	
3.	Addition	+ and -	Numerical (Quantitative)
4.	Multiplication	* and /	

- Nominal (with distinctiveness property only)
- Ordinal (with distinctive and order property only)
- Interval (with additive property + property of Ordinal data)
- Ratio (with multiplicative property + property of Interval data)

nominal and ordinal are collectively referred to as categorical or qualitative data. Whereas, interval and ratio data are collectively referred to as quantitative or numeric data.

Nominal scale : A variable that takes a value among a set of mutually exclusive codes that have no logical order is known as a nominal variable.

- The nominal scale is used to label data categorization using a consistent naming convention.
- The labels can be numbers, letters, strings, enumerated constants or other keyboard symbols.
- Nominal data thus makes "category" of a set of data.
- The number of categories should be two (binary) or more (ternary, etc.), but countably finite.
- A nominal data may be numerical in form, but the numerical values have no mathematical interpretation. For example, 10 prisoners are 100, 101, … 110, but; 100 + 110 = 210 is meaningless. They are simply labels.
- Two labels may be identical (=) or dissimilar (≠). These labels do not have any ordering among themselves. For example, we cannot say blood group B is better or worse than group A.
- Labels (from two different attributes) can be combined to give another nominal variable. For example, blood group with Rh factor (A+ , A- , AB+, etc.)

Binary scale :-

- A nominal variable with exactly two mutually exclusive categories that have no logical order is known as binary variable. Examples Switch: {ON, OFF} Attendance: {True, False} Entry: {Yes, No}.
- A Binary variable is a special case of a nominal variable that takes only two possible values.
- Different binary variables may have unequal importance. If two choices of a binary variable have equal importance, then it is called symmetric binary variable. Example: Gender = {male , female} - usually of equal probability.
- If the two choices of a binary variable have unequal importance, it is called asymmetric binary variable. Example: Food preference = {V , NV}
- Summary statistics applicable to nominal data are mode, contingency correlation, etc. Arithmetic (+,-,*and/) and logical operations (<,>,≠ etc.) are not permitted.
- The allowed operations are : accessing (read, check, etc.) and re-coding (into another non-overlapping symbol set, that is, one-to-one mapping) etc.
- Nominal data can be visualized using line charts, bar charts or pie charts etc.
- Two or more nominal variables can be combined to generate other nominal variable. Example: Gender (M,F) × Marital status (S, M, D, W)

Ordinal scale

- Ordered nominal data are known as ordinal data and the variable that generates it is called ordinal variable. Example: Shirt size = { S, M, L, XL, XXL}
- The values assumed by an ordinal variable can be ordered among themselves as each pair of values can be compared literally or using relational operators (< , ≤ , > , ≥).
- Usually relational operators can be used on ordinal data. Summary measures mode and median can be used on ordinal data.

- Ordinal data can be ranked (numerically, alphabetically, etc.) Hence, we can find any of the percentiles measures of ordinal data.
- Calculations based on order are permitted (such as count, min, max, etc.). Spearman's R can be used as a measure of the strength of association between two sets of ordinal data.
- Numerical variable can be transformed into ordinal variable and vice-versa, but with a loss of information. For example, Age [1, … 100] = [young, middle-aged, old]

Interval scale

- Interval-scale variables are continuous measurements of a roughly linear scale. Example: weight, height, latitude, longitude, weather, temperature, calendar dates, etc.
- Interval data are with well-defined interval. Interval data are measured on a numeric scale (with +ve, 0 (zero), and –ve values).
- Interval data has a zero point on origin. However, the origin does not imply a true absence of the measured characteristics. For example, temperature in Celsius and Fahrenheit; 0^o does not mean absence of temperature, that is, no heat!
- We can add to or from interval data. For example: date1 + x-days = date2
- Subtraction can also be performed. For example: current date – date of birth = age
- Negation (changing the sign) and multiplication by a constant are permitted.
- All operations on ordinal data defined are also valid here.
- Linear (e.g. cx + d) or Affine transformations are permissible.
- Other one-to-one non-linear transformation (e.g., log, exp, sin, etc.) can also be applied.
- Interval data can be transformed to nominal or ordinal scale, but with loss of information.
- Interval data can be graphed using histogram, frequency polygon, etc.

Ratio scale

- Interval data with a clear definition of "zero" are called ratio data. Example: Temperature in Kelvin scale, Intensity of earth-quake on Richter scale, Sound intensity in Decibel, cost of an article, population of a country, etc.
- All ratio data are interval data but the reverse is not true. In ratio scale, both differences between data values and ratios (of non-zero) data pairs are meaningful.
- Ratio data may be in linear or non-linear scale. Both interval and ratio data can be stored in same data type (i.e., integer, float, double, etc.)
- All arithmetic operations on interval data are applicable to ratio data. In addition, multiplication, division, etc. are allowed. Any linear transformation of the form (ax + b)/c are known.