Introduction to Data Science

1. Introduction to Data Science

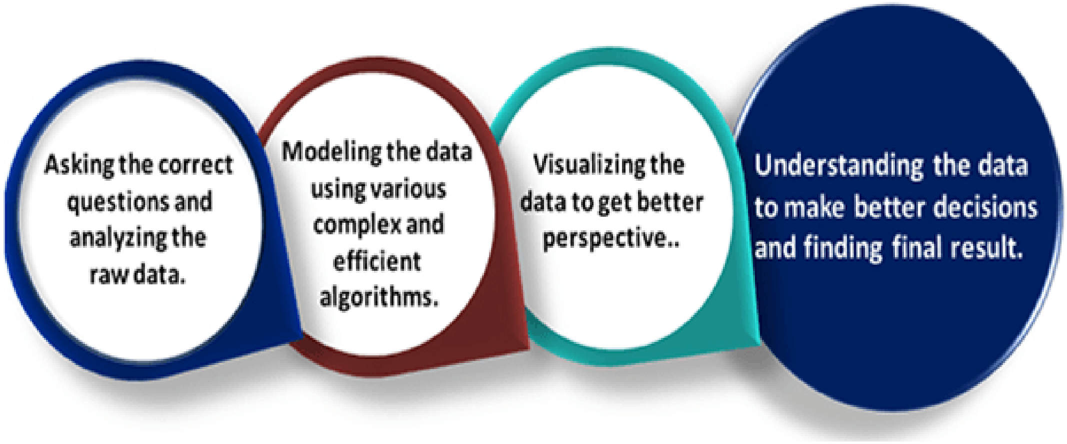
What is Data Science?

Data Science can be explained as the **entire process of gathering actionable insights from raw data** that involves various concepts that include statistical analysis, data analysis, machine learning algorithms, data modeling, preprocessing of data, etc.

Let's consider an example. A case study which also went to become a hollywood feature film "Moneyball". In the movie, they have shown how an underdog team went on to compete at the highest level of the baseball tournament by analyzing the statistical data points of each player and quantifying their performances to win the game. It can be aligned with how data science actually works.

How does Data Science Work?

- ✓ Asking the correct questions and analyzing the raw data.
- ✓ Modeling the data using various complex and efficient algorithms.
- ✓ Visualizing the data to get a better perspective.
- ✓ Understanding the data to make better decisions and finding the final result.


Asking the correct questions and analyzing the raw data.
Modeling the data using various complex and efficient algorithms.
Visualizing the data to get better perspective..
Understanding the data to make better decisions and finding final result.

Example:
Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Need for Data Science:
In today's world, data is becoming so vast, i.e., approximately **2.5 quintals bytes** of data is generating on every day, which led to data explosion. It is estimated as per researches, that by 2020, 1.7 MB of data will be created at every single second, by a single person on earth. Every Company requires data to work, grow, and improve their businesses.

Now, handling of such huge amount of data is a challenging task for every organization. So to handle, process, and analysis of this, we required some complex, powerful, and efficient algorithms and technology, and that technology came into existence as data Science. Following are some main reasons for using data science technology:

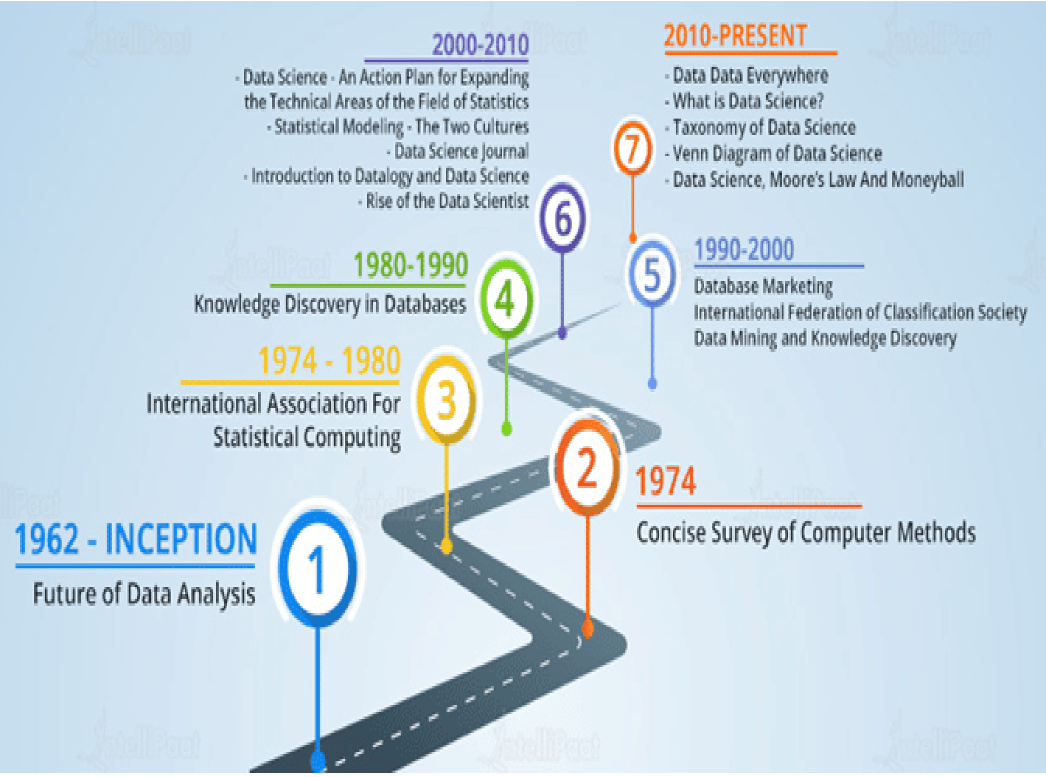
✓ With the help of data science technology, we can convert the massive amount of raw and unstructured data into meaningful insights.

✓ Data science technology is opting by various companies, whether it is a big brand or a startup. Google, Amazon, Netflix, etc, which handle the huge amount of data, are using data science algorithms for better customer experience.

✓ Data science is working for automating transportation such as creating a self-driving car, which is the future of transportation.

✓ Data science can help in different predictions such as various survey, elections, flight ticket confirmation, etc.

2. A brief history of Data Science



2000-2010
- Data Science - An Action Plan for Expanding the Technical Areas of the Field of Statistics
- Statistical Modeling - The Two Cultures
- Data Science Journal
- Introduction to Datalogy and Data Science
- Rise of the Data Scientist
2010-PRESENT
- Data Data Everywhere
- What is Data Science?
- Taxonomy of Data Science
- Venn Diagram of Data Science
- Data Science, Moore's Law And Moneyball
7
6
1980-1990
Knowledge Discovery in Databases
4
5
1990-2000
Database Marketing
International Federation of Classification Society
Data Mining and Knowledge Discovery
1974 - 1980
International Association For Statistical Computing
3
2
1974
Concise Survey of Computer Methods
1962 - INCEPTION
Future of Data Analysis
1

1. 1962 – Inception

a. Future of Data Analysis – In 1962, John W Tukey wrote the "Future of Data Analysis" where he first mentioned the importance of data analysis with respect to science rather than mathematics.

2. 1974

a. Concise Survey of Computer Methods – In 1974, Peter Naur published the "Concise Survey of Computer methods that surveys the contemporary methods of data processing in various applications.

3. 1974 – 1980

a. International Association For Statistical Computing – In 1997, The committee was formed whose sole purpose is to link traditional statistical methodology with modern computer technology to extract useful information and knowledge from the data.

4. 1980-1990

a. Knowledge Discovery in Databases – In 1989, Gregory Piatetsky-Shapiro chaired the Knowledge Discovery in Databases that later went on to become the annual conference on knowledge discovery and data mining.

5. 1990-2000

a. Database Marketing – In 1994, BusinessWeek published a cover story that explains how big organizations are using the customer data to predict the likelihood of a customer buying a specific product or not. Kind of like how targeted ads work in the modern era for social media campaigns.

b. International Federation of Classification Society – For the first time in 1996, the term "Data Science" was used in a conference held in Japan.

6. 2000-2010

a. Data Science – An Action Plan for Expanding the Technical Areas of the Field of Statistics – In 2001, William S Cleveland published the action plan, that majorly focused on major areas of the technical work in the field of statistics and coined the term Data Science.

b. Statistical Modeling – The Two Cultures – In 2001, Leo Breiman wrote "There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown".

c. Data Science Journal – April 2002 saw the launch of a journal that focused on management of data and databases in science and technology.

7. 2010-Present

a. Data Everywhere – In February 2010, Kenneth Cukier wrote a special report for The Economist that said a new professional has arrived – a data scientist. Who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data.

b. What is Data Science? – In June 2010, Mike Loukides described data science as combining entrepreneurship with patience, the willingness to build data products incrementally, the ability to explore, and the ability to iterate over a solution.

3. Data science role and Skill tracks

Data science has three main skill tracks: engineering, analysis, and modeling.

1. Engineering

Data engineering is the foundation that makes everything else possible. It mainly involves in building the data pipeline infrastructure. 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.

(a) 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.

(b) 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.

(c) Production

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.

2. Analysis

Analysis turns raw information into insights in a fast and often exploratory way.

(a) Domain knowledge

Domain knowledge is the understanding of the organization or industry where you apply data science.

Some questions about the context are:

- What are the business questions?
- How to translate a business need to a data problem?

Domain knowledge helps you to deliver the results in an audience-friendly way with the right solution to the right problem.

(b) Exploratory analysis

This type of analysis is about exploration and discovery. It often involves different ways to slice and aggregate data.

(c) Storytelling

Storytelling with data is critical to deliver insights and drive better decision making. It usually requires data summarization, aggregation, and visualization. A business-friendly report or an interactive dashboard is the typical outcome of the analysis.

3. Modeling

Modeling is a process that dives deeper into the data to discover the pattern.

(a) Supervised learning

Supervised learning happens in the presence of a supervisor just like learning performed by a small child with the help of his teacher. As a child is trained to recognize fruits, colors, numbers under the supervision of a teacher this method is supervised learning. In this method, every step of the child is checked by the teacher and the child learns from the output that he has to produce.

(b) Unsupervised learning

Unsupervised learning happens without the help of a supervisor just like a fish learns to swim by itself. It is an independent learning process. In this model, as there is no output mapped with the input, the target values are unknown/ unlabeled. The system needs to learn by itself from the data input to it and detect the hidden patterns.

(c) Customized model development

A data scientist may need to develop new models to accommodate the features of the problem at hand.

Here is a list of questions that can help you decide the type of technique to use:

- Is your data labeled?
- Is your data easy to collect?

4. Some common skills

Data Preprocessing:

Data preprocessing is the process of converting raw data into clean data that is proper to use.

(a) Data preprocessing for data engineer

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.

(b) Data preprocessing for data analyst and scientist

A data analyst collects and stores data on sales numbers, market research, logistics, linguistics, or other behaviors. They bring technical expertise to ensure the quality and accuracy of that data, then process, design, and present it in ways to help people, businesses, and organizations make better decisions.

Data Scientist:

A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc.

4. Problem type

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. Questions of this kind are:

- What is the annual income distribution?
- What are the mean active days of different accounts?

2. Comparison

The first common modeling problem is to compare different groups.

Here are some examples:

- Are males more inclined to buy our products than females?
- Are there any differences in customer satisfaction in different business districts?

The commonly used statistical tests are chi-square test, t-test, and ANOVA.

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?

Clustering is an unsupervised learning mechanism. Unsupervised learning happens without the help of a supervisor just like a fish learns to swim by itself. It is an independent learning process. In this model, as there is no output mapped with the input, the target values are unknown/ unlabeled. The system needs to learn by itself from the data input to it and detect the hidden patterns.

What Is Unlabeled Dataset?

A dataset with unknown output values for all the input values is called an unlabeled dataset. **For Example,** while buying products online, if butter is put in the cart, then it suggests buying bread, cheese, etc. The unsupervised model looks at the data points and predicts the other attributes that are associated with the product.

The unsupervised learning algorithms include Clustering and Association Algorithms such as:

Apriori, K-means clustering and other association rule mining algorithms.

Clustering Algorithm: The methods of finding the similarities between data items such as the same shape, size, color, price, etc. and grouping them to form a cluster is cluster analysis.

Association Rule Mining: In this type of mining, it finds out the most frequently occurring itemsets or associations between elements. Associations such as “products often purchased together”, etc.

4. Classification

Here are some example questions:

- Will this customer likely to buy our product?
- Is it spam email or not?

Classification is a supervised learning mechanism. Supervised learning happens in the presence of a supervisor just like learning performed by a small child with the help of his teacher. As a child is trained to recognize fruits, colors, numbers under the supervision of a teacher this method is supervised learning.

In this method, every step of the child is checked by the teacher and the child learns from the output that he has to produce.

What Is a Labeled Dataset?

The dataset with outputs known for a given input is called a Labeled Dataset. **For example,** an image of fruit along with the fruit name is known. So when a new image of fruit is shown, it compares with the training set to predict the answer.

Supervised learning is a fast learning mechanism with high accuracy. The supervised learning problems include regression and classification problems.

Some of the supervised learning algorithms are:

Decision Trees, K-Nearest Neighbor, Linear Regression, and Neural Networks.

Classification: In these types of problems, we predict the response as specific classes, such as "yes" or "no". When only 2 classes are present, then it is called a Binary Classification. For more than 2 class values, it is called a Multi-class Classification. The predicted response values are discrete values.

For example, Is it the image of the sun or the moon? The classification algorithm separates the data into classes.

Difference Between Supervised Vs Unsupervised Learning

Supervised	**Unsupervised**
In supervised learning algorithms, the output for the given input is known.	In unsupervised learning algorithms, the output for the given input is unknown.
The algorithms learn from labeled set of data. This data helps in evaluating the accuracy on training data.	The algorithm is provided with unlabeled data where it tries to find patterns and associations in between the data items.
It is a Predictive Modeling technique which predicts the future outcomes accurately.	It is a Descriptive Modeling technique which explains the real relationship between the elements and history of the elements.
It includes classification and regression algorithms.	It includes clustering and association rules learning algorithms.
Some algorithms of supervised learning are Linear Regression, Naïve Bayes, and Neural Networks.	Some algorithms for unsupervised learning are k- means clustering, Apriori, etc.
It is more accurate than unsupervised learning as input data and corresponding output is well known, and the machine only needs to give predictions.	It has less accuracy as the input data is unlabeled. Thus the machine has to first understand and label the data and then give predictions.

5. Regression

Here are some example questions:

- What will be the temperature tomorrow?
- How much inventory should we have?

Regression: Regression problems predict the response as continuous values such as predicting a value that ranges from -infinity to infinity. It may take many values. **For example,** the linear regression algorithm that is applied, predicts the cost of the house based on many parameters such as location, nearby airport, size of the house, etc.

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.

Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It **predicts continuous/real values such as** temperature, age, salary, price, **etc.**

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?

5. List of potential data science careers

Data infrastructure engineer

Designing, building, and running the data infrastructure to support the Video organizations growing data needs.

Go, Python, AWS/Google Cloud/Azure, logstash, Kafka, and Hadoop

Data Engineer:

A data engineer works with massive amount of data and responsible for building and maintaining the data architecture of a data science project. Data engineer also works for the creation of data set processes used in modeling, mining, acquisition, and verification.

Skill required: Data engineer must have depth knowledge of **SQL, MongoDB, Cassandra, HBase, Apache Spark, Hive, MapReduce**, with language knowledge of **Python, C/C++, Java, Perl**, etc.
spark/scala, python, SQL, AWS/Google Cloud/Azure, Data modeling

BI engineer

Design, implement, and maintain systems used to collect and analyze business intelligence data. They create dashboards, databases, and other platforms that allow for efficient collection and evaluation of BI data.

Skill required: Tableau/looker/Mode, etc., data visualization, SQL, Python

Data Analyst:

Data analyst is an individual, who performs mining of huge amount of data, models the data, looks for patterns, relationship, trends, and so on. At the end of the day, he comes up with visualization and reporting for analyzing the data for decision making and problem-solving process.

Skill required: For becoming a data analyst, you must get a good background in **mathematics, business intelligence, data mining**, and basic knowledge of **statistics**. You should also be familiar with some computer languages and tools such as **MATLAB, Python, SQL, Hive, Pig, Excel, SAS, R, JS, Spark**, etc.

Data Scientist:

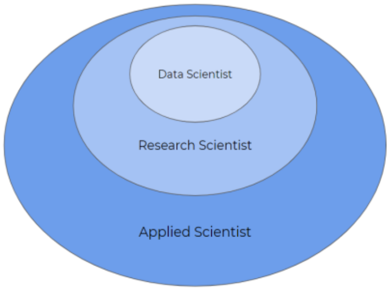
A data scientist is a professional who works with an enormous amount of data to come up with compelling business insights through the deployment of various tools, techniques, methodologies, algorithms, etc.

Skill required: To become a data scientist, one should have technical language skills such as **R, SAS, SQL, Python, Hive, Pig, Apache spark, MATLAB**. Data scientists must have an understanding of Statistics, Mathematics, visualization, and communication skills.

Research scientist

- ✓ Building research proposals
- ✓ Creating and conducting experiments
- ✓ Analysing results of the experiments
- ✓ Working with other researchers to use and develop end product
- ✓ Applying for grants to continue research

Skill required: R/Python, advanced statistics, experimental design, ML, research background, publications, conference contributions, algorithms



Applied scientist

An applied scientist is more interested in real-life applications.

An applied scientist does scientific research with a focus on applying the results of their studies to solving real-world problems. They use the scientific method to develop research questions and then conduct studies that lead to practical solutions.

Applied Scientists at Amazon, for example, focus on projects to enhance Amazon's customer experience like Amazon's Automatic Speech Recognition (ASR), Natural Language Understanding (NLU), Audio Signal Processing, text-to-speech (TTS)

Skill required: ML algorithm design, often with an expectation of fundamental software engineering skills

Machine Learning Engineer

Their tasks involve researching, building, and designing the artificial intelligence responsible for machine learning and maintaining and improving existing artificial intelligence systems.

Machine learning engineers design and create the AI algorithms capable of learning and making predictions that define machine learning (ML). An ML engineer typically works as part of a larger data science team and will communicate with data scientists, administrators, data analysts, data engineers and data architects. They should also have an understanding of various algorithms, problem-solving analytical skill, probability, and statistics.

Skill required: More advanced software engineering skillset, algorithms, machine learning algorithm design, system design

6. Life Cycle of Data Science (Stages of Data Science Project)

The Lifecycle of Data Science

The major steps in the life cycle of Data Science project are as follows:

1. Problem identification

Domain experts and Data Scientists are the key persons in the problem identification of problem. Domain expert has in depth knowledge of the application domain and exactly what is the problem to be solved. Data Scientist understands the domain and help in identification of problem and possible solutions to the problems.

2. Business Understanding

Understanding what customer exactly wants from the business perspective is nothing but Business Understanding. Whether customer wish to do predictions or want to improve sales or minimise the loss or optimise any particular process etc forms the business goals. During business understanding two important steps are followed:

- **KPI (Key Performance Indicator)**

For any data science project, key performance indicators define the performance or success of the project. There is a need to be an agreement between the customer and data science project team on Business related indicators and related data science project goals. Depending on the business need the business indicators are devised and then accordingly the data science project team decides the goals and indicators. To better understand this let us see an example. Suppose the business need is to optimise the overall spending of the company, then the data science goal will be to use the existing resources to manage double the clients. Defining the Key performance Indicators is very crucial for any data science projects as the cost of the solutions will be different for different goals.

- **SLA (Service Level Agreement)**

Once the performance indicators are set then finalizing the service level agreement is important. As per the business goals the service level agreement terms are decided. For example, for any airline reservation system simultaneous processing of say 1000 users is required. Then the product must satisfy this service requirement is the part of service level agreement.

Once the performance indicators are agreed and service level agreement is completed then the project proceeds to the next important step.

3. Collecting Data

Data Collection is the important step as it forms the important base to achieve targeted business goals.

The basic data collection can be done using the surveys. Generally, the data collected through surveys provide important insights. Much of the data is collected from the various processes followed in the enterprise. At various steps the data is recorded in various software systems used in the enterprise which is important to understand the process followed from the product development to deployment and delivery. The historical data available through archives is also important to better understand the business. Transactional data also plays a vital role as it is collected on a daily basis. Many statistical methods are applied to the data to extract the important information related to business. In data science project the major role is played by data and so proper data collection methods are important.

4. Pre-processing data

Large data is collected from archives, daily transactions and intermediate records. The data is available in various formats and in various forms. Some data may be available in hard copy formats also. The data is scattered at various places on various servers. All these data are extracted and converted into single format and then processed. As data warehouse is constructed where the Extract, Transform and Loading (ETL) process or operations are carried out. A data architect role is important in this stage who decides the structure of data warehouse and perform the steps of ETL operations.

5. Analyzing data

Now that the data is available and ready in the format required then next important step is to understand the data in depth. This understanding comes from analysis of data using various statistical tools available. A data engineer plays a vital role in analysis of data. This step is also called as Exploratory Data Analysis (EDA). Here the data is examined by formulating the various statistical functions and dependent and independent variables or features are identified. Careful analysis of data revels which data or features are important and what is the spread of data. Various plots are utilized to visualize the data for better understanding. The tools like Tableau, PowerBI etc are famous for performing Exploratory Data Analysis and Visualization. Knowledge of Data Science with Python and R is important for performing EDA on any type of data.

6. Data Modelling

Data modelling is the important next step once the data is analysed and visualized. The important components are retained in the dataset and thus data is further refined. Now the important is to decide how to model the data? What tasks are suitable for modelling? The tasks, like classification or regression, which is suitable is dependent upon what business value is required. In these tasks also many ways of modelling are available. The Machine Learning engineer applies various algorithms to the data and generates the output. While modelling the data many a times the models are first tested on dummy data similar to actual data.

7. Model Evaluation/ Monitoring

As there are various ways to model the data so it is important to decide which one is effective. For that model evaluation and monitoring phase is very crucial and important. The model is now tested with actual data. The data may be very few and in that case the output is monitored for improvement. There may be changes in data while model is being evaluated or tested and the output will drastically change depending on changes in data. So, while evaluating the model following two phases are important:

8. Model Training

Once the task and the model are finalised and data drift analysis modelling is finalized then the important step is to train the model. The training can be done is phases where the important parameters can be further fine tuned to get the required accurate output. The model is exposed to the actual data in production phase and output is monitored.

9. Model Deployment

Once the model is trained with the actual data and parameters are fine tuned then model is deployed. Now the model is exposed to real time data flowing into the system and output is generated. The model can be deployed as web service or as an embedded application in edge or mobile application. This is very important step as now model is exposed to real world.

10. Driving insights and generating BI reports

After model deployment in real world, next step is to find out how model is behaving in real world scenario. The model is used to get the insights which aid in strategic decisions related to business. The business goals are bound to these insights. Various reports are generated to see how business is driving. These reports help in finding out if key process indicators are achieved or not.

11. Taking a decision based on insight

For data science to make wonders, every step indicated above has to be done very carefully and accurately. When the steps are followed properly then the reports generated in above step helps in taking key decisions for the organization. The insights generated helps in taking strategic decisions like for example the organization can predict that there will be need of raw material in advance. The data science can be of great help in taking many important decisions related to business growth and better revenue generation.

7. Application of data Science in various Field

1. In Search Engines

The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly used Search engines like Google, Yahoo, Safari, Firefox, etc. So Data Science is used to get Searches faster.

For Example, When we search something suppose "Data Structure and algorithm courses" then at that time on the Internet Explorer we get the first link of GeeksforGeeks Courses. This happens because the GeeksforGeeks website is visited most in order to get information regarding Data Structure courses and Computer related subjects. So this analysis is done using Data Science, and we get the topmost visited Web Links.

2. In Transport

Data Science also entered into the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

For Example, In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what is the speed limit in Highway, Busy Streets, Narrow Roads and how to handle different situations while driving etc.

3. In Finance

Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

For Example, In Stock Market, Data Science is the main part. In the Stock Market, Data Science is used to examine past behavior with past data and their goal is to examine the future outcome. Data is analyzed in such a way that it makes it possible to predict future stock prices over a set timetable.

4. In E-Commerce

E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.

For Example, When we search for something on the E-commerce websites we get suggestions similar to choices according to our past data and also we get recommendations according to most buy the product, most rated, most searched, etc. This is all done with the help of Data Science.

5. In Health Care

Data Science is used for: Detecting Tumor, Drug discoveries, Medical Image Analysis, Virtual Medical Bots, Genetics and Genomics, Predictive Modeling for Diagnosis etc.

6. Image Recognition

Currently, Data Science is also used in Image Recognition. **For Example,** When we upload our image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science. When an Image is Recognized, the data analysis is done on one's Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.

7. Targeting Recommendation

Targeting Recommendation is the most important application of Data Science. Whatever the user searches on the Internet, he/she will see numerous posts everywhere. Suppose I want a mobile phone, so I just Google search it and after that, I changed my mind to buy offline. Data Science helps those companies who are paying for Advertisements for their mobile. So everywhere on the internet in the social media, in the websites, in the apps everywhere I will see the recommendation of that mobile phone which I searched for. So this will force me to buy online.

8. Airline Routing Planning

With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.

9. Data Science in Gaming

In most of the games where a user will play with an opponent i.e. a Computer Opponent, data science concepts are used with machine learning where with the help of past data the Computer will improve its performance. There are many games like Chess, EA Sports, etc. will use Data Science concepts.

10. Medicine and Drug Development

The process of creating medicine is very difficult and time-consuming and has to be done with full disciplined because it is a matter of Someone's life. Without Data Science, it takes lots of time, resources, and finance or developing new Medicine or drug but with the help of Data Science, it becomes easy because the prediction of success rate can be easily determined based on biological data or factors. The algorithms based on data science will forecast how this will react to the human body without lab experiments.

11. In Delivery Logistics

Various Logistics companies like DHL, FedEx, etc. make use of Data Science. Data Science helps these companies to find the best route for the Shipment of their Products, the best time suited for delivery, the best mode of transport to reach the destination, etc.

12. Autocomplete

AutoComplete feature is an important part of Data Science where the user will get the facility to just type a few letters or words, and he will get the feature of auto-completing the line. In Google Mail, when we are writing formal mail to someone so at that time data science concept of Autocomplete feature is used where he/she is an efficient choice to auto-complete the whole line. Also in Search Engines in social media, in various apps, AutoComplete feature is widely used.

13. 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.

8. 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.

Why is Data Security important?

Data is a valuable asset that generates, acquires, saves, and exchanges for any company. Protecting it from internal or external corruption and illegal access protects a company from financial loss, reputational harm, consumer trust degradation, and brand erosion.

Main elements of Data Security

- ✓ **Confidentiality:** Ensures that only authorized users, with appropriate credentials, have access to data.
- ✓ **Integrity:** Ensures that all data is accurate, trustworthy, and not prone to unjustified changes.
- ✓ **Availability:** Ensures that data is accessible and available for ongoing business needs in a timely and secure manner.

Data Privacy:

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.

The primary difference is that data privacy mainly focuses on keeping data confidential, while data security mainly focuses on protecting from malicious activity.

Differentiate between Data Privacy and Data Security

Data Privacy	**Data Security**
Data Privacy is all about the reflection of what data is important and why.	Data Security is all about the reflection of how those policies got enforced.
Data privacy sets about proper usage, collection, retention, deletion, and storage of data.	Data security sets the policies, methods, and means to secure personal data.
It offers to block websites, internet browsers, cable companies, and internet service providers from tracking your information and your browser history.	It offers to protect you from other people accessing your personal information and other data.
Data privacy tools include browser extensions and add-on, password managers, private browsers and email services, encrypted messaging, private search engines, web proxies, file encryption software, and ad and tracker blockers.	Data Security tools involve with identity and access management, data loss prevention, anti-malware, anti-virus, event management and data masking software.
For e.g. The European Union's General Data Protection Regulation is a type of international standard for protecting the privacy of EU citizens.	For e.g. The Payment Card Industry Data Security Standard is a set of rules which protect the sensitive payment card information and cardholder data.

Data Security Risks

1. 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.

A data breach or data leak is the release of sensitive, confidential or protected data to an untrusted environment.

2. 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.

3. 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.

4. 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.

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.

5. 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.

6. 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.

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. Authentication and Authorization

Organizations must put in place strong authentication methods, such as OAuth 2.0 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.

Authentication verifies the identity of a user or service, and authorization determines their access rights.

Comparing these processes to a real-world example, when you go through security in an airport, you show your ID to authenticate your identity. Then, when you arrive at the gate, you present your boarding pass to the flight attendant, so they can authorize you to board your flight and allow access to the plane.

9. 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.

10. 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. 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.

11. 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.

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.

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.

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:

- ✓ 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.

- ✓ Use strong passwords to protect access to search clusters

- ✓ Encrypt all communications using SSL/TLS SSL (Secure Socket Layer) and TLS (Transport Layer Security)
- ✓ Use IP (Internet Protocol) filtering for client access
- ✓ Turn on auditing and monitor logs on an regular basis

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 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 sensitive because it can expose an organization’s most valuable asset—customer relationships.

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.

9. Data collection strategies

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.

Why is Data Collection important?

- **The trustworthiness of The Research –** A critical purpose behind data collection via quantitative or qualitative techniques is to guarantee that the research question's honesty is kept up without a doubt.
- **Diminish the probability of blunders or errors –** The right utilization of suitable data collection strategies decreases the probability of blunders during different research processes.
- **Effective and accurate decision making –** To limit the danger of blunders or errors in decision making, it is significant that precise data is gathered, so the specialists do not settle on clueless choices.
- **Save Cost and Time –** Data collection plays a significant role in saving time and money that can otherwise be squandered without more profound comprehension of the point or topic.
- **Empowers a new idea or change –** To demonstrate the requirement for an adjustment or new change, it is critical to collect data and information as proof to help these cases.

Depending on the type of data, the data collection method is divided into two categories namely,

- Primary Data Collection methods
- Secondary Data Collection methods

Primary Data Collection Methods

Primary data or raw data is a type of information that is obtained directly from the first-hand source through experiments, surveys or observations. The primary data collection method is further classified into two types. They are

- Quantitative Data Collection Methods
- Qualitative Data Collection Methods

Quantitative Data Collection Methods

It is based on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures. This method is cheaper than qualitative data collection methods and it can be applied in a short duration of time.

Qualitative Data Collection Methods

It does not involve any mathematical calculations. This method is closely associated with elements that are not quantifiable. This qualitative data collection method includes interviews, questionnaires, observations, case studies, etc. There are several methods to collect this type of data. They are

Observation Method

Observation method is used when the study relates to behavioural science. This method is planned systematically. It is subject to many controls and checks. The different types of observations are:

- Structured and unstructured observation
- Controlled and uncontrolled observation
- Participant, non-participant and disguised observation

Interview Method

The method of collecting data in terms of verbal responses. It is achieved in two ways, such as

- Personal Interview – In this method, a person known as an interviewer is required to ask questions face to face to the other person. The personal interview can be structured or unstructured, direct investigation, focused conversation, etc.
- Telephonic Interview – In this method, an interviewer obtains information by contacting people on the telephone to ask the questions or views, verbally.

Questionnaire Method

In this method, the set of questions are mailed to the respondent. They should read, reply and subsequently return the questionnaire. The questions are printed in the definite order on the form. A good survey should have the following features:

- Short and simple, Should follow a logical sequence
- Provide adequate space for answers, Avoid technical terms

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.

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.

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.

Schedules

This method is similar to the questionnaire method with a slight difference. The enumerations are specially appointed for the purpose of filling the schedules. It explains the aims and objects of the investigation and may remove misunderstandings, if any have come up. Enumerators should be trained to perform their job with hard work and patience.

Secondary Data Collection Methods

Secondary data is data collected by someone other than the actual user. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

Published data are available in various resources including

- Government publications, Public records, Historical and statistical documents
- Business documents, Technical and trade journals

Unpublished data includes

- Diaries, Letters, Unpublished biographies, etc.

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").

1. Word Association

The researcher gives the respondent a set of words and asks them what comes to mind when they hear each word.

2. 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.

3. Role-Playing

Respondents are presented with an imaginary situation and asked how they would act or react if it was real.

4. In-Person Surveys

The researcher asks questions in person.

5. Online/Web Surveys

These surveys are easy to accomplish, but some users may be unwilling to answer truthfully, if at all.

6. 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.

7. 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.

8. 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.

9. Photography and video: Photographs and videos show still or moving images. Photographs can be used on their own, but are more often accompanied by written captions, providing additional information. Videos are often accompanied by a commentary.

10. Focus group discussions: Focus group discussions (FGDs) are facilitated discussions, held with a small group of people who have specialist knowledge or interest in a particular topic. They are used to find out the perceptions and attitudes of a defined group of people. FGDs are typically carried out with around 6-12 people, and are based around a short list of guiding questions, designed to probe for in-depth information.

11. Case studies and stories of change: A case study is not a data collection tool in itself. It is a descriptive piece of work that can provide in-depth information on a topic. It is often based on information acquired through one or more of the other tools described in this paper, such as interviews or observation. Case studies are usually written, but can also be presented as photographs, films or videos. Case studies often focus on people (individuals, households, communities). But they can also focus on any other unit of analysis such as locations, organisations, policies or the environment. Stories of change are similar to case studies.

12. Surveys and questionnaires: These are designed to collect and record information from many people, groups or organisations in a consistent way. A questionnaire is a form containing questions. It may be a printed form or one designed to be filled in online. Questionnaires may be administered in many different ways. A survey, by contrast, is normally a large, formal exercise. It typically consists of three different aspects: an approved sampling method designed to ensure the survey is representative of a wider population; a standard questionnaire that ensures information is collected and recorded consistently; and a set of analysis methods that allow results and findings to be generated.

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.

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. 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.

What are the Key Steps in the Data Collection Process?

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.

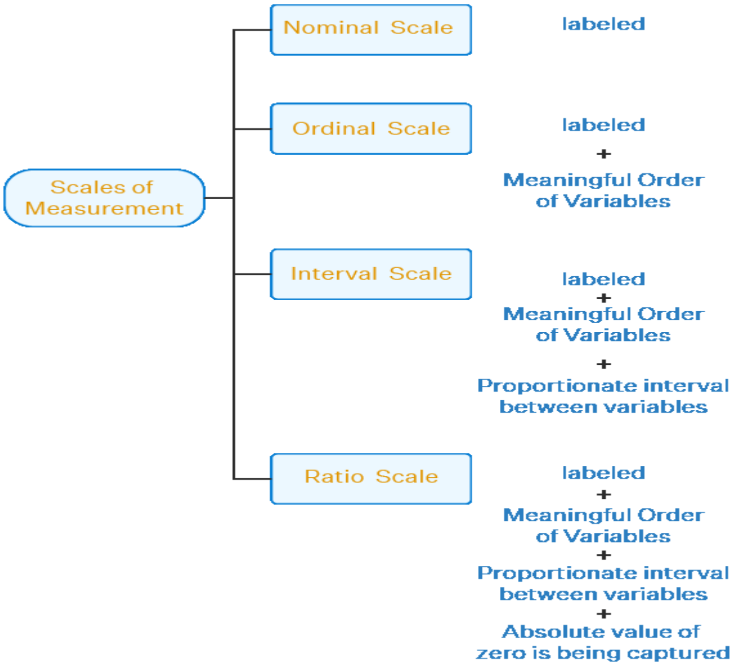
4. Gather Information

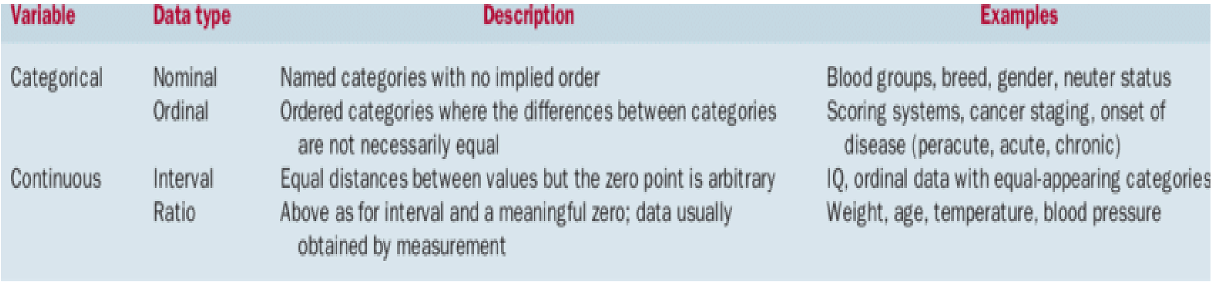
Once our plan is complete, we can put our data collection plan into action and begin gathering data. In our DMP, we can store and arrange our data. We need to be careful to follow our plan and keep an eye on how it's doing. Especially if we are collecting data regularly, setting up a timetable for when 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.

10. Data Categororization: NOIR


Scales of Measurement
Nominal Scale
labeled
Ordinal Scale
labeled
+
Meaningful Order of Variables
Interval Scale
labeled
+
Meaningful Order of Variables
+
Proportionate interval between variables
Ratio Scale
labeled
+
Meaningful Order of Variables
+
Proportionate interval between variables
+
Absolute value of zero is being captured



Variable	Data type	Description	Examples
Categorical	Nominal	Named categories with no implied order	Blood groups, breed, gender, neuter status
	Ordinal	Ordered categories where the differences between categories are not necessarily equal	Scoring systems, cancer staging, onset of disease (peracute, acute, chronic)
Continuous	Interval	Equal distances between values but the zero point is arbitrary	IQ, ordinal data with equal-appearing categories
	Ratio	Above as for interval and a meaningful zero; data usually obtained by measurement	Weight, age, temperature, blood pressure

Nominal Scale of Measurement

A nominal scale of measurement is used for qualitative data. It does not give any numerical meaning to the data. Using the nominal scale of measurement, the data can be classified but cannot be added, subtracted, multiplied, or divided. It can cover a wide variety of qualitative data. Some of the situations where nominal measurement scale can be used are given below:

- ✓ Study to find the country of birth of people in a town
- ✓ In collecting data on the eye color of people
- ✓ Classifying people into categories like male/female, working-class population/unemployed, vaccinated/unvaccinated people, etc.

Some of the properties of the nominal scale of measurement are given below:

- ✓ It can categorize variables but does not put them in any order.
- ✓ It does not show any numerical value.
- ✓ It is used for qualitative data.

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.

Symmetric and Asymmetric Binary Scale

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}

Ordinal Scale of Measurement

The ordinal scale of measurement groups the data into order or rank. It contains the property of nominal scale as well, which is to classify data variables into specific labels. And in addition to that, it organizes data into groups though it does not have any numerical value. For example, the study of people's satisfaction with a company's product on a scale of #1 - Very happy, #2 - satisfactory, #3 - neutral, #4 - unhappy, and #5 - extremely dissatisfied. This is an example of an ordinal scale of measurement. This measurement scale can be used for the following purposes:

- ✓ Ranks of players in a race.
- ✓ Data collection on variables such as hottest to coldest, richest to poorest, etc.
- ✓ Data on people's satisfaction with any product, person, or government.

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}

Some of the properties of the ordinal measurement scale are listed below:

- ✓ It displays the order or rating of the variables.
- ✓ It does not give any numerical value to the data. So, it is also used for qualitative data as similar to nominal measurement scale.
- ✓ It contains variables that can be placed in order like heaviest to lightest, ranks of players or students, etc.

Interval Scale of Measurement

The interval scale of measurement includes those values that can be measured in a specific interval, for example, time, temperature, etc. It shows the order of variables with a meaning proportion or difference between them. For example, on a temperature scale, the difference between 20 °C and 30 °C is the same as the difference between 50°C ad 60°C. It is an example of an interval measurement scale. On the other hand, the difference between the scores of the first two rankers in a race and the two runner-ups will be different, which is an example of an ordinal scale.

Some of the properties of the interval scale of measurement are listed below:

- ✓ It includes the properties of both nominal and ordinal scales.
- ✓ It shows meaningful divisions between variables.
- ✓ The difference between the variables can be presented in numerical terms.
- ✓ It includes variables that can be added or subtracted from each other.
- ✓ It gives a meaning to 'Zero" which was not possible in the above two scales. For example, zero degrees of temperature.

Ratio Scale of Measurement

The ratio scale is the most comprehensive scale among others. It includes the properties of all the above three scales of measurement. The unique feature of the ratio scale of measurement is that it considers the absolute value of zero, which was not the case in the interval scale. When we measure the height of the people, 0 inches or 0 cm means that the person does not exist. On the interval scale, there are values possible on both sides of 0, for example, temperature could be negative as well. While the ratio scale does not include negative numbers because of its feature of showing absolute zero. An example of the ratio measurement scale is determining the weight of people from the following options: less than 20 kgs, 20 - 40 kgs, 40 - 60 kgs, 60 - 80 kgs, and more than 80 kgs.

Some of the properties of the ratio scale of measurement are listed below:

- ✓ It is used for quantitative data.
- ✓ It shows the absolute value of zero which means if the value is 0, it's nothing.
- ✓ The variables can be added, subtracted, multiplied, or divided. In addition to these, calculation of mean, median, and mode is also possible with this scale.
- ✓ it doesn't include negative numbers because of the feature of true zero value.

Types of Data:

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.

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