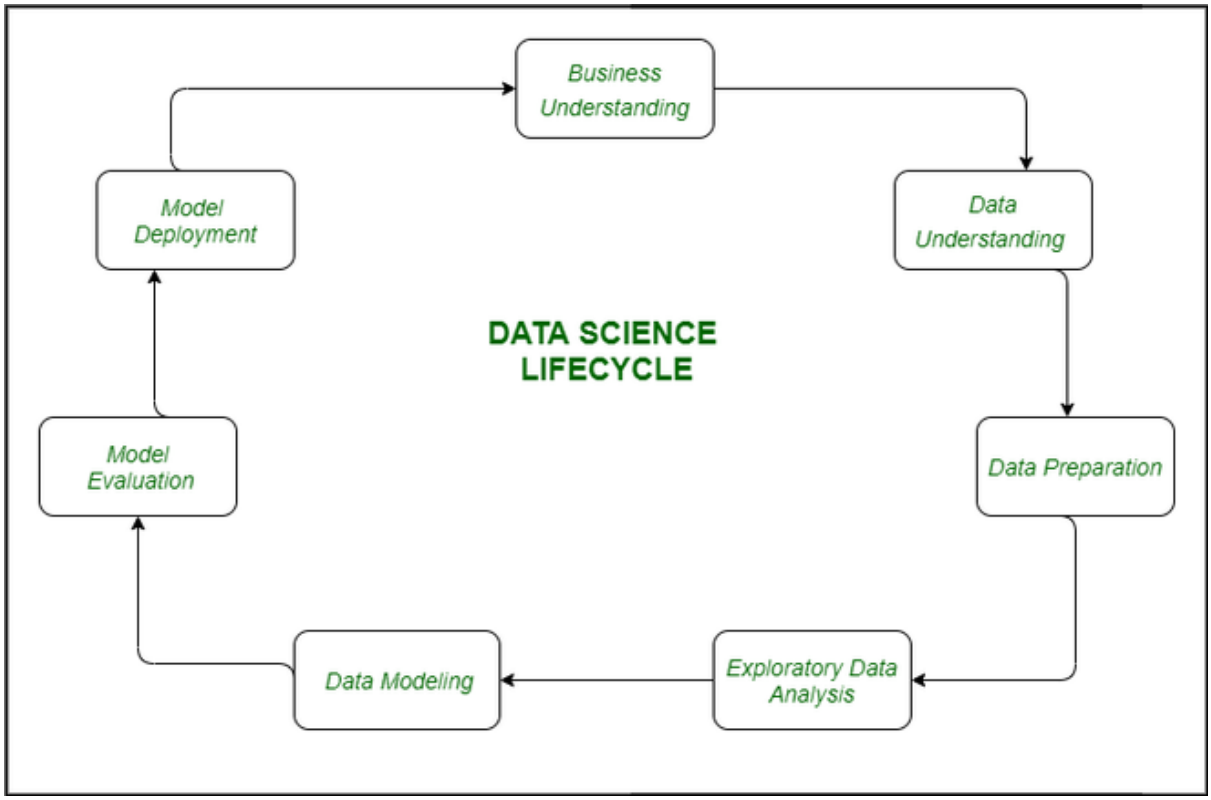
THE LIFECYCLE OF DATA SCIENCE



1. Business Understanding: The complete cycle revolves around the enterprise goal. What will you resolve if you do no longer have a specific problem? It is extraordinarily essential to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective. You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.

2. Data Understanding: After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

3. Preparation of Data: Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them. Constructing new data, derive new elements from present ones. Format the data into the preferred structure, eliminate undesirable columns and features. Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

4. Exploratory Data Analysis: This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps. Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

5. Data Modeling: Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

6. Model Evaluation: Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. Any data science

solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

7. Model Deployment: The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste. For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

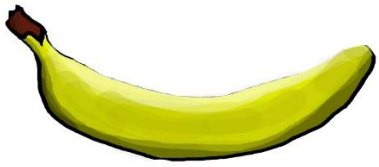
Supervised learning

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labelled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labelled data.

For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all the different fruits one by one like this:

- If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –**Apple**.
- If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –**Banana**.

Now suppose after training the data, you have given a new separate fruit, say Banana from the basket, and asked to identify it.



Since the machine has already learned the things from previous data and this time has to use it wisely. It will first classify the fruit with its shape and color and would confirm the fruit name as BANANA and put it in the Banana category. Thus the machine learns the things from training data(basket containing fruits) and then applies the knowledge to test data(new fruit).

Supervised learning is classified into two categories of algorithms:

- **Classification**: A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Supervised learning deals with or learns with “labeled” data. This implies that some data is already tagged with the correct answer.

Types:-

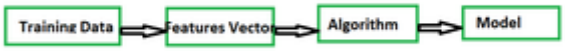
- Regression
- Logistic Regression
- Classification
- Naive Bayes Classifiers
- K-NN (k nearest neighbors)
- Decision Trees
- Support Vector Machine

Advantages:-

- Supervised learning allows collecting data and produces data output from previous experiences.
- Helps to optimize performance criteria with the help of experience.
- Supervised machine learning helps to solve various types of real-world computation problems.

Disadvantages:-

- Classifying big data can be challenging.
- Training for supervised learning needs a lot of computation time. So, it requires a lot of time.

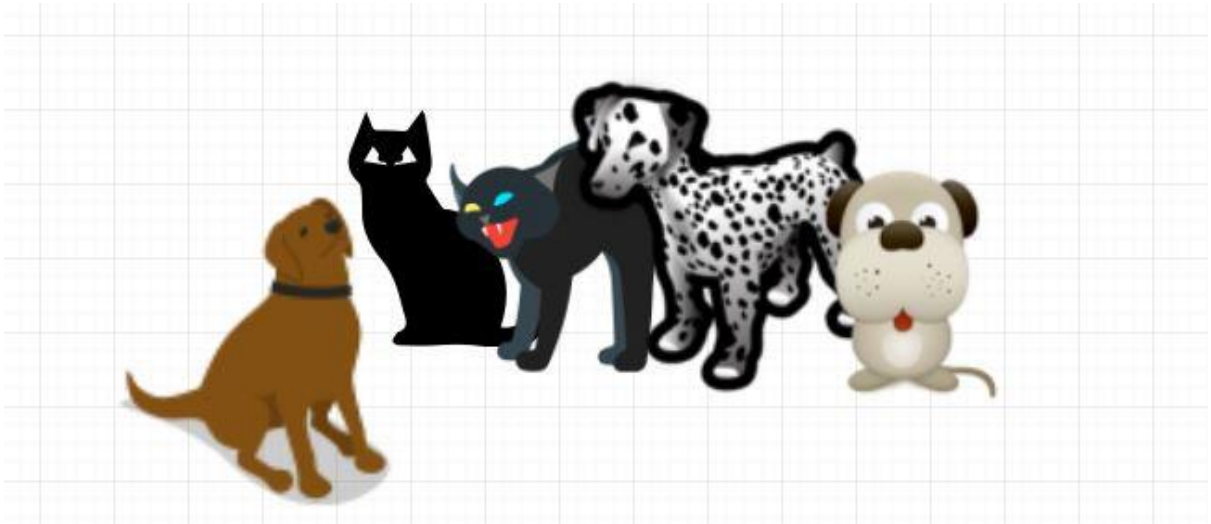


Steps

Unsupervised learning

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.
For instance, suppose it is given an image having both dogs and cats which it has never seen.



Thus the machine has no idea about the features of dogs and cats so we can't categorize it as 'dogs and cats '. But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having **dogs** in them and the second part may contain all pics having **cats** in them. Here you didn't learn anything before, which means no training data or examples.

It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.

Unsupervised learning is classified into two categories of algorithms:

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Types of Unsupervised Learning:-

Clustering

1. Exclusive (partitioning)
2. Agglomerative
3. Overlapping
4. Probabilistic

Clustering Types:-

1. Hierarchical clustering
2. K-means clustering
3. Principal Component Analysis
4. Singular Value Decomposition
5. Independent Component Analysis

Supervised Learning	Unsupervised Learning
Supervised learning algorithms are trained using labeled data.	Unsupervised learning algorithms are trained using unlabeled data.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning is not close to true artificial intelligence as in this, we first train the model for each data, and then only it can predict the correct output.	Unsupervised learning is more close to the true artificial intelligence as it learns similarly as a child learns daily routine things by his experiences.
It includes various algorithms such as linear regression, logistic regression, support vector machine, multi-class classification, decision tree, bayesian logic, etc.	It includes various algorithms such as clustering, knn, and apriori algorithm.

What is big data?

Big data is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

Big Data contains a large amount of data that is not being processed by traditional data storage or the processing unit. It is used by many **multinational companies** to **process** the data and business of many **organizations**. The data flow would exceed **150 exabytes** per day before replication.

Why is big data important?

Companies use big data in their systems to improve operations, provide better customer service, create personalized marketing campaigns and take other actions that, ultimately, can increase revenue and profits. Businesses that use it effectively hold a potential competitive advantage over those that don't because they're able to make faster and more informed business decisions.

For example, big data provides valuable insights into customers that companies can use to refine their marketing, advertising and promotions in order to increase customer engagement and conversion rates. Both historical and real-time data can be analyzed to assess the evolving preferences of consumers or corporate buyers, enabling businesses to become more responsive to customer wants and needs.

Big data is also used by medical researchers to identify disease signs and risk factors and by doctors to help diagnose illnesses and medical conditions in patients. In addition, a combination of data from electronic health records, social media sites, the web and other sources gives healthcare organizations and government agencies up-to-date information on infectious disease threats or outbreaks.

Here are some more examples of how big data is used by organizations:

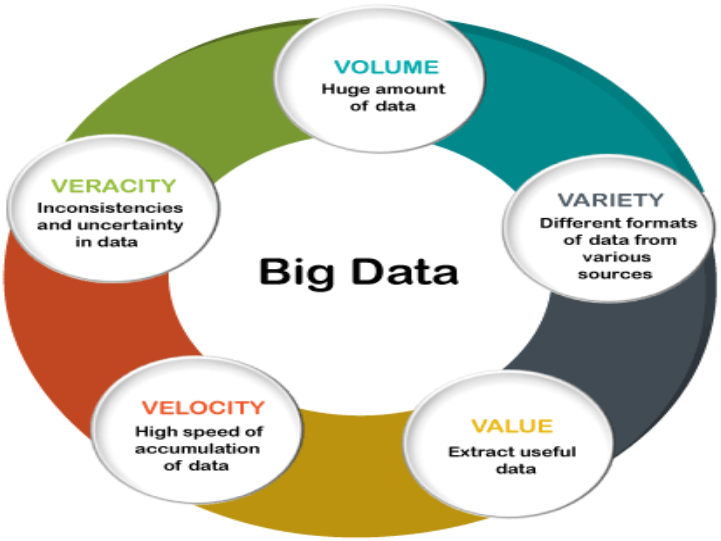
- In the energy industry, big data helps oil and gas companies identify potential drilling locations and monitor pipeline operations; likewise, utilities use it to track electrical grids.
- Financial services firms use big data systems for risk management and real-time analysis of market data.
- Manufacturers and transportation companies rely on big data to manage their supply chains and optimize delivery routes.
- Other government uses include emergency response, crime prevention and smart city initiatives.

Characteristics of Big Data

There are five v's of Big Data that explains the characteristics.

5 V's of Big Data

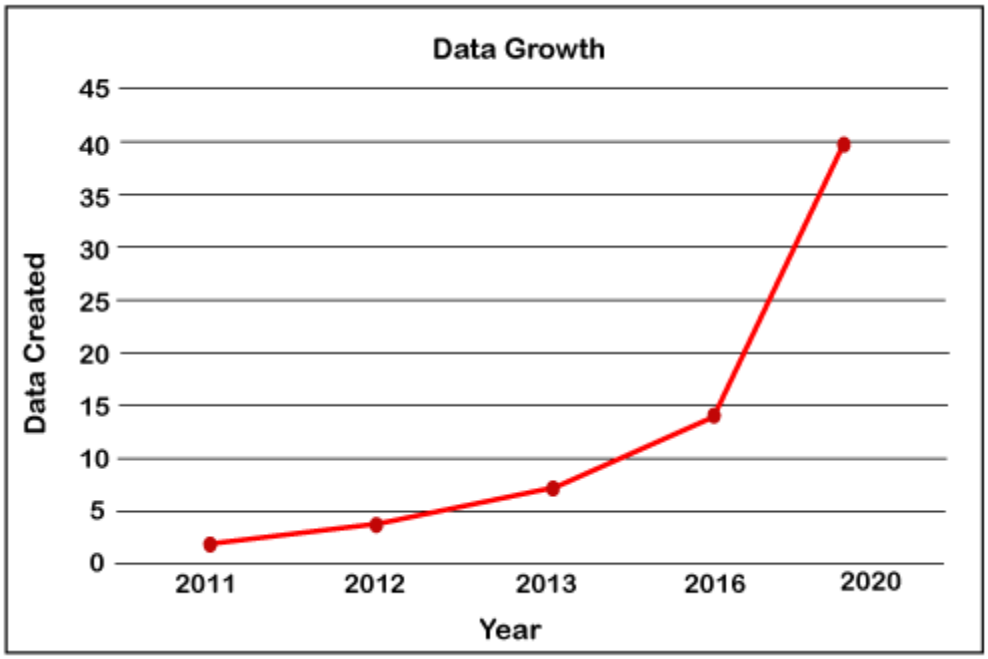
- **Volume**
- **Veracity**
- **Variety**
- **Value**
- **Velocity**



Volume

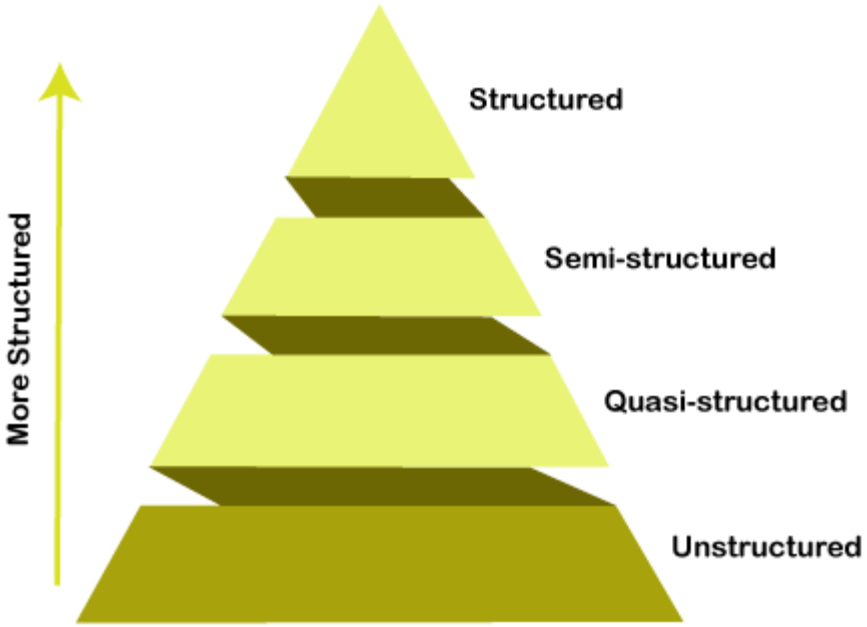
The name Big Data itself is related to an enormous size. Big Data is a vast 'volumes' of data generated from many sources daily, such as **business processes, machines, social media platforms, networks, human interactions,** and many more.

Facebook can generate approximately a **billion** messages, **4.5 billion** times that the "**Like**" button is recorded, and more than **350 million** new posts are uploaded each day. Big data technologies can handle large amounts of data.



Variety

Big Data can be **structured, unstructured, and semi-structured** that are being collected from different sources. Data will only be collected from **databases** and **sheets** in the past, But these days the data will comes in array forms, that are **PDFs, Emails, audios, SM posts, photos, videos,** etc.



The data is categorized as below:

a. **Structured data:** In Structured schema, along with all the required columns. It is in a tabular form. Structured Data is stored in the relational database management system.

b. **Semi-structured:** In Semi-structured, the schema is not appropriately defined, e.g., **JSON, XML, CSV, TSV**, and **email**. OLTP (**Online Transaction Processing**) systems are built to work with semi-structured data. It is stored in relations, i.e., **tables**.

c. **Unstructured Data**: All the **unstructured files, log files, audio files**, and **image** files are included in the unstructured data. Some organizations have much data available, but they did not know how to **derive** the value of data since the data is raw.

d. **Quasi-structured Data:**The data format contains textual data with inconsistent data formats that are formatted with effort and time with some tools.

Example: Web server logs, i.e., the log file is created and maintained by some server that contains a list of **activities**.

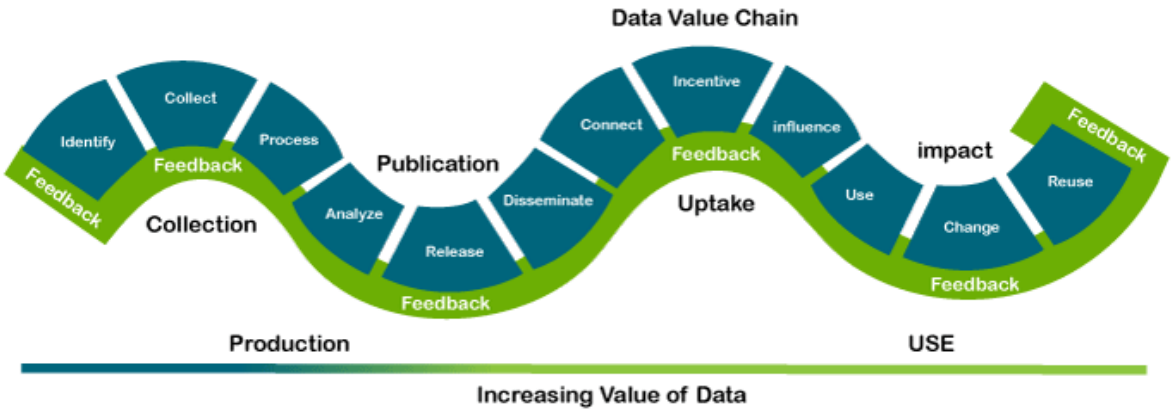
Veracity

Veracity means how much the data is reliable. It has many ways to filter or translate the data. Veracity is the process of being able to handle and manage data efficiently. Big Data is also essential in business development.

For example, **Facebook posts** with hashtags.

Value

Value is an essential characteristic of big data. It is not the data that we process or store. It is **valuable** and **reliable** data that we **store, process**, and also **analyze**.



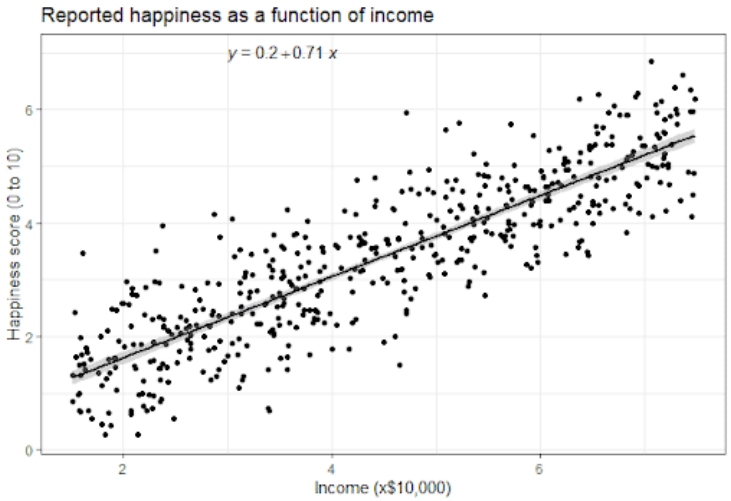
Velocity

Velocity plays an important role compared to others. Velocity creates the speed by which the data is created in **real-time**. It contains the linking of incoming **data sets speeds, rate of change**, and **activity bursts**. The primary aspect of Big Data is to provide demanding data rapidly.

Big data velocity deals with the speed at the data flows from sources like **application logs, business processes, networks, and social media sites, sensors, mobile devices,** etc.

Statistical Modeling

Statistical modeling is the use of mathematical models and statistical assumptions to generate sample data and make predictions about the real world. A statistical model is a collection of probability distributions on a set of all possible outcomes of an experiment.



What is Statistical Modeling?

Statistical modeling refers to the data science process of applying statistical analysis to datasets. A statistical model is a mathematical relationship between one or more random variables and other non-random variables. The application of statistical modeling to raw data helps data scientists approach data analysis in a strategic manner, providing intuitive visualizations that aid in identifying relationships between variables and making predictions.

Common data sets for statistical analysis include Internet of Things (IoT) sensors, census data, public health data, social media data, imagery data, and other public sector data that benefit from real-world predictions.

Statistical Modeling Techniques

The first step in developing a statistical model is gathering data, which may be sourced from spreadsheets, databases, data lakes, or the cloud. The most common statistical modeling methods for analyzing this data are categorized

as either supervised learning or unsupervised learning. Some popular statistical model examples include logistic regression, time-series, clustering, and decision trees.

Supervised learning techniques include regression models and classification models:

- Regression model: a type of predictive statistical model that analyzes the relationship between a dependent and an independent variable. Common regression models include logistic, polynomial, and linear regression models. Use cases include forecasting, time series modeling, and discovering the causal effect relationship between variables.
- Classification model: a type of machine learning in which an algorithm analyzes an existing, large and complex set of known data points as a means of understanding and then appropriately classifying the data; common models include models include decision trees, Naive Bayes, nearest neighbor, random forests, and neural networking models, which are typically used in Artificial Intelligence.

Unsupervised learning techniques include clustering algorithms and association rules:

- K-means clustering: aggregates a specified number of data points into a specific number of groupings based on certain similarities.
- Reinforcement learning: an area of deep learning that concerns models iterating over many attempts, rewarding moves that produce favorable outcomes and penalizing steps that produce undesired outcomes, therefore training the algorithm to learn the optimal process.

There are three main types of statistical models: parametric, nonparametric, and semiparametric:

- Parametric: a family of probability distributions that has a finite number of parameters.
- Nonparametric: models in which the number and nature of the parameters are flexible and not fixed in advance.

- Semiparametric: the parameter has both a finite-dimensional component (parametric) and an infinite-dimensional component (nonparametric).

How to Build Statistical Models

The first step in building a statistical model is knowing how to choose a statistical model. Choosing the best statistical model is dependent upon several different variables. Is the purpose of the analysis to answer a very specific question, or solely to make predictions from a set of variables? How many explanatory and dependent variables are there? What is the shape of the relationships between dependent and explanatory variables? How many parameters will be included in the model? Once these questions are answered, the appropriate model can be selected.

Once a statistical model is selected, it must be built. Best practices for how to make a statistical model include:

- Start with univariate descriptives and graphs. Visualizing the data helps with identifying errors, understanding the variables you're working with, how they look, how they are behaving and why.
- Build predictors in theoretically distinct sets first in order to observe how related variables work together, and then the outcome once the sets are combined.
- Next, run bivariate descriptives with graphs in order to visualize and understand how each potential predictor relates individually to every other predictor and to the outcome.
- Frequently record, compare and interpret results from models run with and without control variables.
- Eliminate non-significant interactions first; any variable involved in a significant interaction must be included in the model by itself.
- While identifying the many existing relationships between variables, and categorizing and testing every possible predictor, be sure not to lose sight of the research question.

Web Scrapping:

https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/

Analysis V/S Reporting:

https://www.orbitanalytics.com/understanding-the-difference-between-reporting-and-analytics/

AI and Data Science:

https://intellipaat.com/blog/data-science-vs-artificial-intelligence-difference/

Myths of Data Science:

https://www.analyticsvidhya.com/blog/2020/09/11-data-science-myths/

Matplotlib:

https://jakevdp.github.io/PythonDataScienceHandbook/04.00-introduction-to-matplotlib.html

Numpy:

https://www.ggpsbokaro.org/images/download1/5155.pdf

Scikit-Learn:

https://jakevdp.github.io/PythonDataScienceHandbook/05.02-introducing-scikit-learn.html

NLTK:

https://realpython.com/nltk-nlp-python/

Visualizing Data:

https://hevodata.com/learn/data-science-visualization/

Working with files:

https://realpython.com/working-with-files-in-python/

Clustering:

https://www.geeksforgeeks.org/clustering-in-machine-learning/