**100 Machine Learning Interview Questions and Answers:** https://www.ilovephd.com/100-machine-learning-interview-questions-and-answers/

1. ***What is the process of Data Analysis?***
2. **Understanding the Problem:** Understand the business problem, define the organizational goals, and plan for a lucrative solution.
3. **Collecting Data:** Gather the right data from various sources and other information based on your priorities.
4. **Cleaning Data:** Clean the data to remove unwanted, redundant, and missing values, and make it ready for analysis.
5. **Exploring and Analyzing Data:** Use data visualization and business intelligence tools, data mining techniques, and predictive modeling to analyze data.
6. **Interpreting the Results:** Interpret the results to find out hidden patterns, future trends, and gain insights.
7. ***What do you understand by BI? Or what is BI?***

The term Business Intelligence refers to a collective meaning, including technologies, tools, applications, practices for the data collection, and providing those data to the users, especially to help in running the business or a part of it.

In other words, the business analyst’s reports generated and compiled using the Business Intelligence approaches are consumed by the higher management administrative and business executives to make better decisions for the overall maturity of the business.

1. ***Do you know of other BI platforms? How does Power BI compareto them?***

**The popular Business Intelligence (BI) tools used by Business Analysts are:**

1. Microsoft BI
2. Cognos
3. MicroStrategy
4. Tableau
5. SAS
6. Business Objects
7. OBIEE
8. Hyperion

## **Power BI vs Tableau: Key Differences**

**Power BI is easy to learn as compared to Tableau.**

* It allows you to create new columns, delete unnecessary columns, modify data, change formats, etc., but Tableau allows you to work on pre-processed data itself. Users can alter the data type and split columns, but they cannot remove unnecessary columns.
* Power BI focuses on the “flow” to have a smooth and consistent user experience. On the other hand, Tableau focuses on coding to provide you with new visual effects which third parties create.
* Power BI uses DAX for calculating and measuring columns, while Tableau deploys MDX for measures and dimensions.
* Power BI offers PowerPivot dots (user to perform efficient data analysis) more user-accessible, but Tableau does not provide this.
* <https://www.interviewbit.com/blog/power-bi-vs-tableau/>

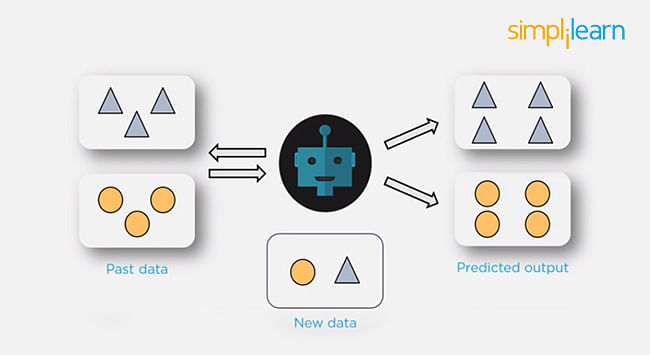
1. ***What is the difference between supervised and unsupervised learning?***

\*\*\* Supervised learning uses data that is completely labeled, whereas unsupervised learning uses no training data.

**Extra:** There are three types of machine learning:

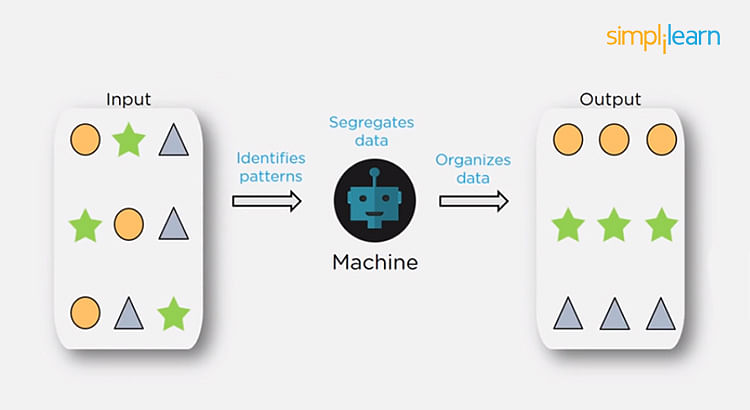
#### Supervised Learning

In supervised machine learning, a model makes predictions or decisions based on past or labeled data. Labeled data refers to sets of data that are given tags or labels, and thus made more meaningful.



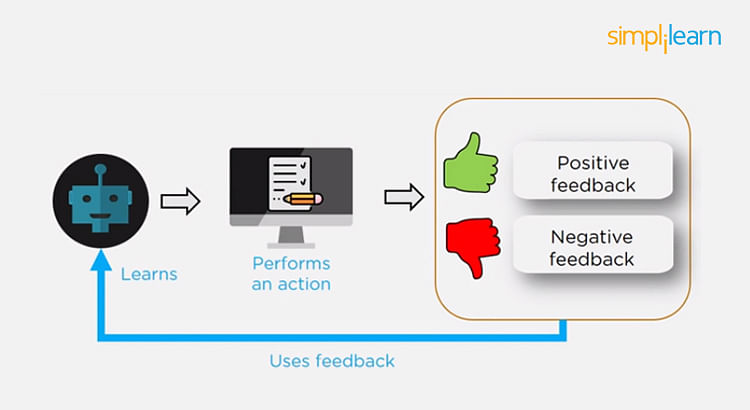
#### Unsupervised Learning

In unsupervised learning, we don't have labeled data. A model can identify patterns, anomalies, and relationships in the input data.



#### Reinforcement Learning

Using [reinforcement learning](https://www.simplilearn.com/tutorials/machine-learning-tutorial/reinforcement-learning), the model can learn based on the rewards it received for its previous action.



Consider an environment where an agent is working. The agent is given a target to achieve. Every time the agent takes some action toward the target, it is given positive feedback. And, if the action taken is going away from the goal, the agent is given negative feedback.

1. ***What is the central limit theorem?***

The Central Limit Theorem states that the sampling distribution of the mean of a large number of independent, identically distributed random samples will approach a normal distribution, regardless of the original population’s distribution.

(Central Limit Theorem states that when ***large samples usually greater than thirty are taken into consideration then***

***the distribution of sample arithmetic mean approaches the normal distribution irrespective of the fact that random***

***variables were originally distributed normally or not.)***

1. ***Tell me about yourself?***

"My name is Enejan, and I am a passionate data scientist with a strong foundation in mathematics. I hold a master's degree in physics and mathematics and have accumulated valuable experience in my field. My journey into the realm of data science began with a desire to broaden my knowledge of mathematics and integrate it with cutting-edge technology.

Driven by my passion for continuous learning and personal development, I discovered the field of data science. Over the past two years, I dedicated myself to studying data science, completing various training programs, and working on numerous projects. I further strengthened my skills through an internship at TechPro, where I had the opportunity to apply my knowledge to real-world data.

I am now seeking an environment where I can not only enhance my skills but also contribute my expertise and experience. I came across your job posting and am excited about the prospect of joining your team. I believe my background and passion for data science make me a valuable candidate who can significantly contribute to your organization.

Thank you for considering my application. I look forward to further discussing how my skills match your team's needs."

***2. Can you mention a few problems that data analysts usually encounter while***

***performing the analysis?***

1. **Biased Data :** Data could be biased due to the source from which it is collected. For instance, suppose you collect data to determine the winner of an electoral campaign, collecting from a specific region alone introduces one form of a bias, while collecting data from a specific income group introduces another form of bias.
2. **Duplicates in the data**: Data could have duplicates which may impact the result of analysis.
3. **Missing data:** All data points might not have the values for all attributes you are analyzing
4. **Noisy data:** The data could be noisy, usually a high value of variance indicates noise.
5. **Outliers in the data**: Points outside the expected range of data that introduce inconsistencies in the model.
6. **Difference in formats in various data sources** : Some data could be crawled and collected in html format, while other data might be collected from online reviews in text format. A third source of data might be structured data already in the database. A data analyst usually has to ingest several data sources to get richer data.
7. **Data Volume :** A large amount of data will require a different class of algorithms for processing to handle efficiently.
8. ***What do you understand by the term Normal Distribution?***

Normal distribution, also known as the **Gaussian distribution**, is a **bell-shaped** probability distribution characterized by **a symmetric and mound-shaped curve**. In a normal distribution, **mean, median, and mode are equal**, and the data is evenly distributed around the mean. The famous **68-95-99.7 rule** describes the percentage of data within **1, 2, and 3 standard deviations** from the mean.

1. ***What are database and relational database management systems (RDMBS)?***

A **database** is a set of data stored in a computer. This data is usually structured in a way that makes the data easily accessible.

A**relational database**is a type of database. It uses a structure that allows us to identify and access data in relation to another piece of data in the database. Often, data in a relational database is organized into tables.

In **RDBMS**, data storage is in the form of tables, and each row within the table is a record with a unique ID called a key. The table's columns hold the data attributes, and the record usually contains the value of each attribute, to establish the relationship among data points quickly.

The following **structures** exist in relational databases

#### Database

A database is a logical grouping of data. It contains the data associated with one application or with a group of related applications. It includes a collection of related table spaces and index spaces.

#### Table

A table is a logical structure made of columns and rows. Rows have no specific order, but columns follow a fixed order to retrieve the data.

#### Indexes

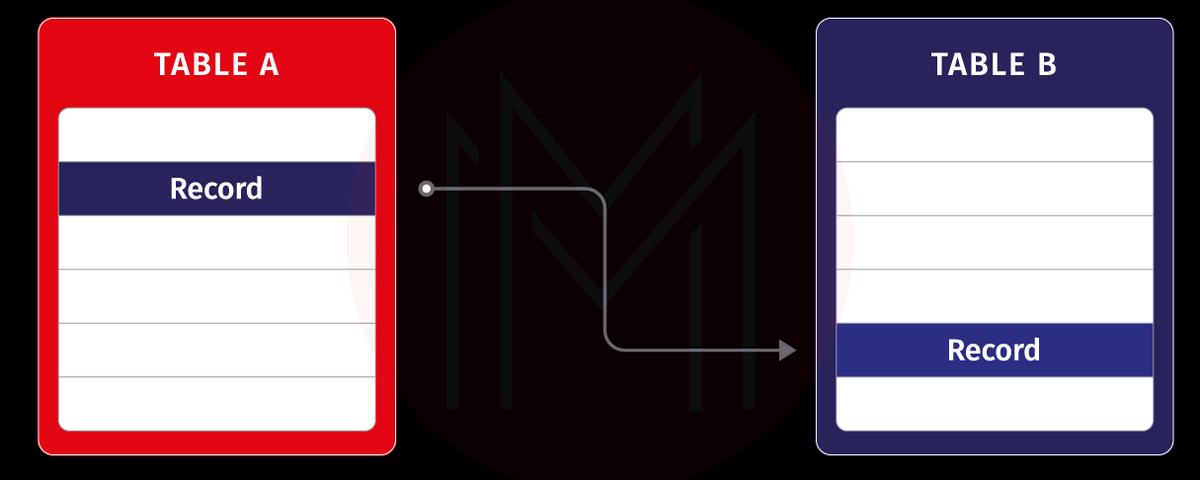
An index is nothing more than an ordered set of pointers to rows of a table. Unlike the rows of the table, the rows in the indexes maintain a specific order to retrieve data.

#### Keys

A key is one or more columns specified as keys when a table, index, or referential integrity is defined. The various types of keys in RDBMS include primary key, unique key, and foreign key.

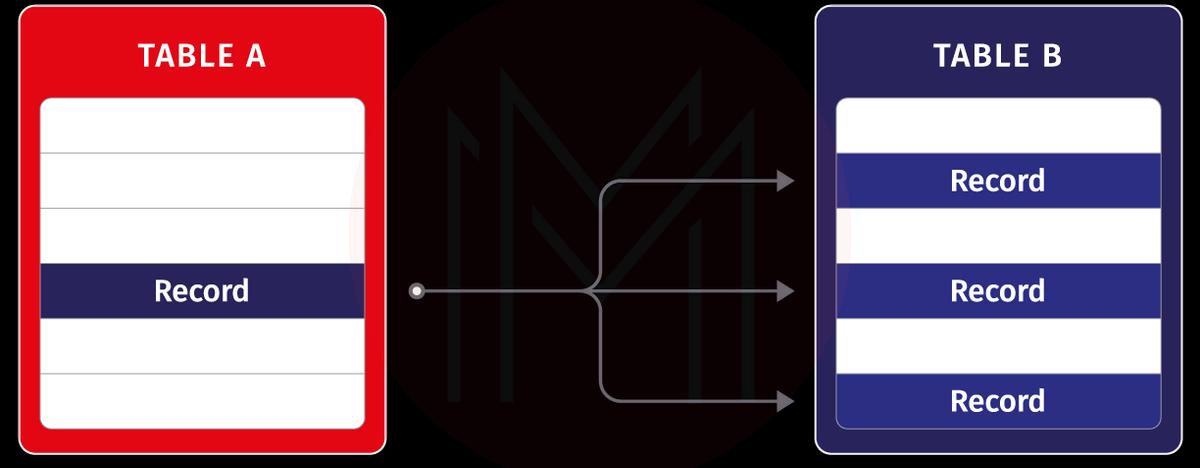
There are **three main types of relationships** between the tables to ensure the absolute flexibility of the relational database model.

#### One-to-One Relationship



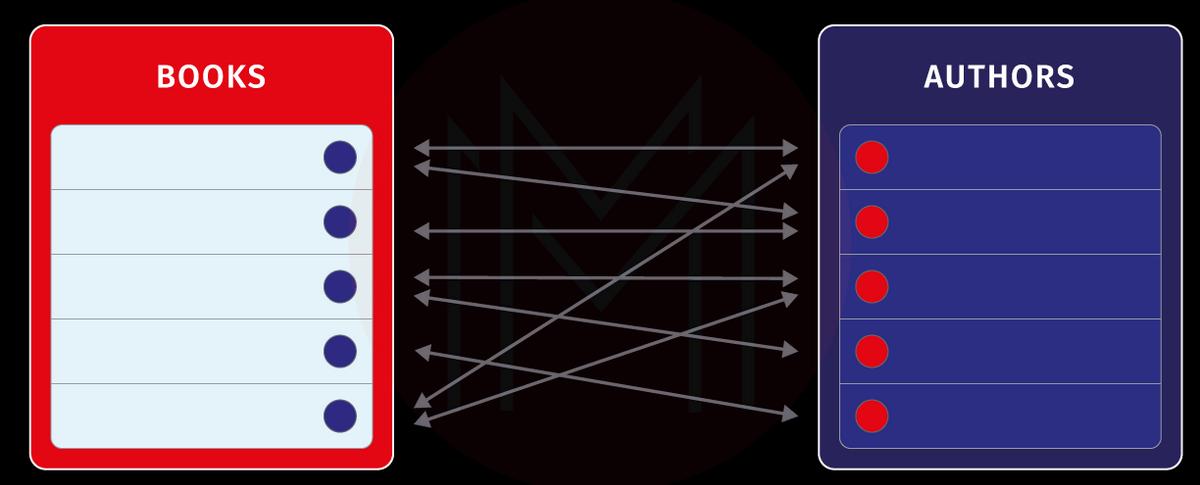
Suppose a single record of table A is related to a single record of table B. In that case, it is a one-to-one relationship.

#### One-to-Many Relationship



Suppose a single record of table A is related to multiple records of table B. In that case, it is a one-to-many relationship.

#### Many-to-Many Relationship



Suppose multiple records of table A are related to multiple records of table B. In that case, it is a many-to-many relationship.

***5. Explain the concept of regularization in machine learning.***

**Regularization** is a technique in machine learning used to prevent **overfitting** ***by adding a penalty to the model’s objective function***. It encourages, simpler models and better generalization. Regularization can be achieved through methods like L1 and L2 regularization.

**L1 Regularization (Lasso Regularization):**  
Penalty term is proportional to the absolute value of the coefficients.  
Encourages sparsity by driving some coefficients to exactly zero, effectively performing feature selection.  
Suitable for situations where some features are less relevant, leading to a more interpretable model.

**L2 Regularization (Ridge Regularization):**  
Penalty term is proportional to the square of the coefficients.  
Shrinks the coefficients towards zero but rarely makes them exactly zero.  
Helps to mitigate the impact of multicollinearity and is generally more stable in the presence of highly correlated features.

#### Explain the concept of ridge regression and its role in regularization.

Ridge regression, also known as L2 regularization, is a linear regression technique that addresses the problem of multicollinearity (high correlation between predictor variables) and overfitting in a regression model. It does so by adding a penalty term to the linear regression cost function, encouraging the model to keep the coefficient values of the predictor variables small. This helps to stabilize the model and reduces the sensitivity to small changes in the input data. In traditional linear regression, the goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of squared errors between the predicted values and the actual target values. In ridge regression, the cost function is augmented with a regularization term, which is proportional to the sum of the squares of the coefficients.