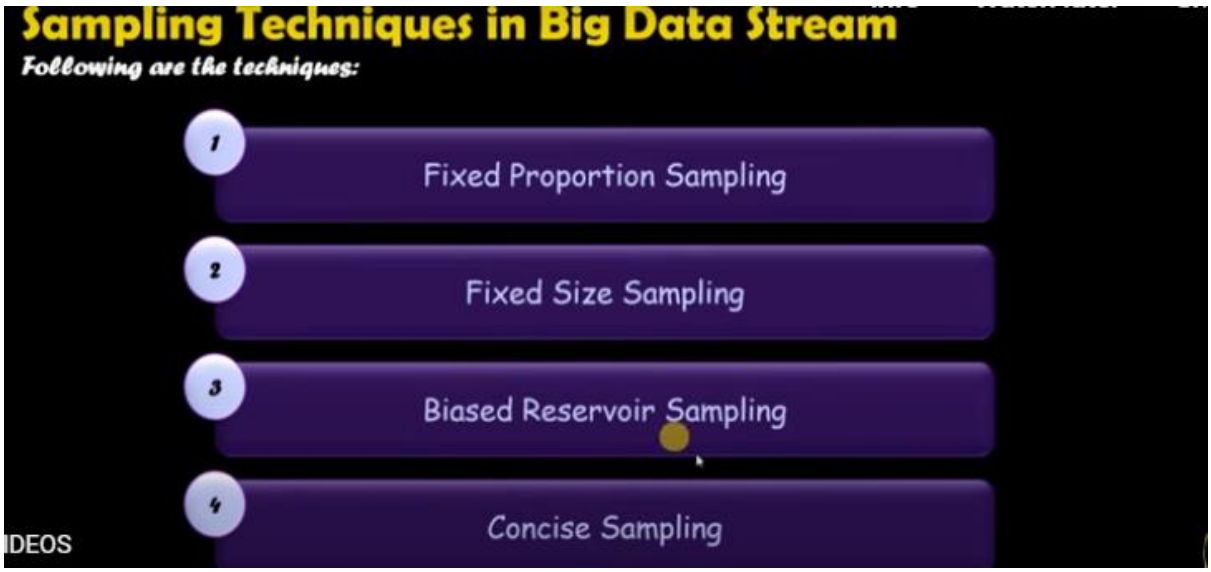
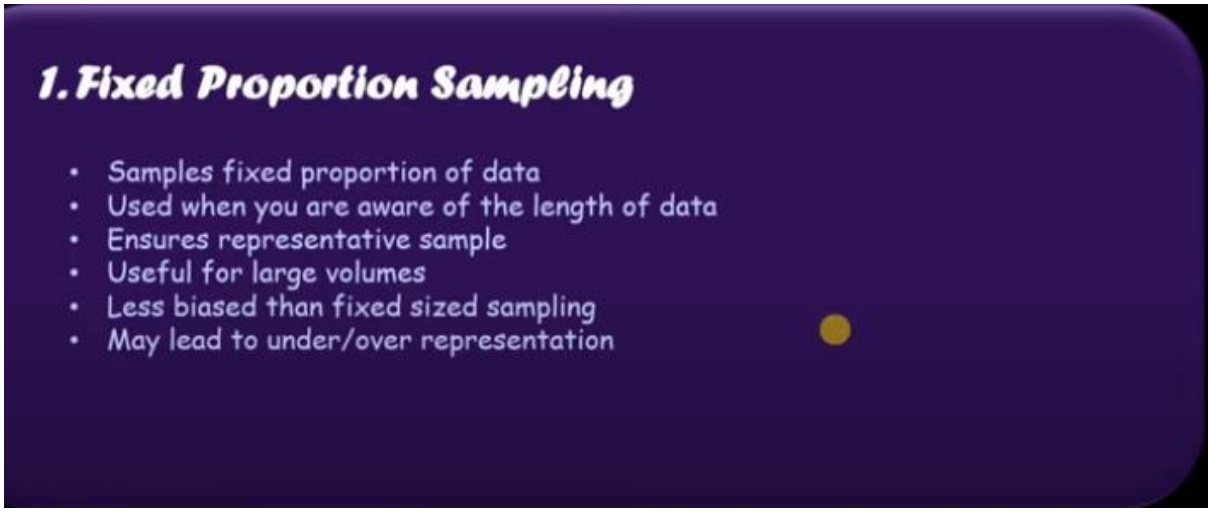
Explain different sampling techniques.





1. Fixed Proportion Sampling

Example

A social media platform wants to analyze the sentiments of its users towards a topic. They receive millions of tweets per day and use fixed proportion sampling to select a representative sample. They randomly select 1% of the tweets received each hour, ensuring a representative sample for statistical analysis of user sentiments towards the topic.

2. Fixed Size Sampling

- Samples fixed number of data points.
- Does not guarantee representative sample.
- Useful for reducing data volume.
- Can be biased if data is not randomly distributed
- Less effective when data size increases

2. Fixed Size Sampling

Example

Suppose we have a data stream of customer orders for an online store, with 10,000 orders coming in every hour. Using fixed size sampling, we randomly select 1,000 orders from each hour's data stream for analysis, thus reducing the total number of data points to process from 10,000 to 1,000 per hour.

3. Biased Reservoir Sampling

- Used in streams to select a subset of the data in a way that is not uniformly random.
- Can lead to a biased sample that may not be representative of the full dataset.
- The selection of elements is based on a predetermined probability distribution that may be weighted towards certain elements or groups of elements.
- The probability distribution used for biased reservoir sampling may be based on various factors, such as the frequency of occurrence of certain types of data or the importance of certain data points.
- Used when there are constraints on the resources available for sampling, such as limited memory or computational power.
- It is important to carefully consider the potential biases introduced by this sampling technique and adjust the analysis accordingly.

3. Biased Reservoir Sampling

Example

Suppose we have a data stream of product ratings, and we want to select a sample of ratings to estimate the average rating of a product. However, we know that some users tend to give higher ratings than others. Using biased reservoir sampling, we can assign a higher probability of selection to ratings from users who tend to give more accurate ratings. This way, our sample is more likely to represent the true average rating of the product.

4. Concise Sampling

- Goal is to maintain a small reservoir of a fixed size while still achieving representative sampling of the data stream
- Number of samples that can be stored in memory at a given time is limited, which can be a challenge when dealing with large data streams.
- Size of the sample may need to be adjusted based on the amount of memory available to store the data.
- Instead of selecting samples randomly, the sampling algorithm may prioritize choosing samples with unique or representative values of a particular attribute in the data stream

4. Concise Sampling

Example

- A bank wants to analyze customer spending habits from a stream of transactions.
- They use concise sampling to choose distinct customer IDs as their attribute.
- The size of the reservoir is limited to 1000 customers.
- They adjust the sample size based on available memory.
- This allows for efficient analysis while maintaining accuracy.