1. Dimensionality Reduction

- **Concept:** Reduces the number of attributes (columns) in the dataset. Think of it as simplifying the "view" of your data by focusing on the most important features.
- **Methods:**
    - **(i) Wavelet Transform:**
        - **How it works:** Decomposes a signal (or data) into different frequency components. You can then keep only the most significant components, effectively compressing the data. Similar to how JPEG image compression works by discarding high-frequency details that the human eye is less sensitive to.
        - **Example:** Compressing a large audio file. You can remove very high and very low frequencies that are inaudible, significantly reducing the file size without noticeably changing the sound quality for most listeners.
        - Suitable for data cube, sparse and skewed data.
    - **(ii) Principal Component Analysis (PCA):**
        - **How it works:** Finds new, uncorrelated attributes (principal components) that capture the most variance in the data. You keep only the top few components, discarding the less important ones. Imagine rotating the data in a high-dimensional space to find the "best" viewing angles.
        - **Example:** Analyzing customer data with many attributes (age, income, purchase history, website clicks, etc.). PCA might reveal that two main components (e.g., "spending power" and "tech savviness") explain most of the differences between customers. You can then focus on these two components instead of the dozens of original attributes.
    - **(iii) Attribute Subset Selection:**
        - **How it works:** Identifies and removes irrelevant or redundant attributes. This is like decluttering your dataset by throwing away columns that don't contribute much useful information.
        - **Example:** In a dataset about predicting house prices, you might have "number of rooms" and "square footage." Since square footage often captures the information about the number of rooms, you might remove "number of rooms" as redundant. Or, a column like "date dataset was created" is likely irrelevant for predicting prices.
        - Methods are Stepwise Forward Selection, Stepwise Backward Elimination, Combination and Decision Tree induction.

2. Numerosity Reduction

- **Concept:** Reduces the number of data points (rows) in the dataset. Instead of looking at every single record, you work with a summarized or representative version.

- **Methods:**
    - **(i) Regression (Parametric):**
        - **How it works:** Fits a mathematical model (like a line or curve) to the data. You store the model's parameters instead of the raw data points. This works well if the data follows a clear pattern (e.g., linear, logarithmic).
        - **Example:** If you have data showing a clear linear relationship between advertising spending and sales, you could fit a linear regression model (Sales = a * Advertising + b). You store 'a' and 'b' instead of all the individual data points.
    - **(ii) Log-Linear Models (Parametric):**
        - **How it works:** Estimates the probability of data points in a multi-dimensional space using a logarithmic function. Useful for analyzing relationships between categorical variables.
        - **Example:** Analyzing the relationship between a customer's age, gender, and whether they purchased a product. A log-linear model could show how the probability of purchase changes based on combinations of age and gender.
    - **(iii) Histograms (Non-Parametric):**
      * **How It Works:** Groups the data into ranges (or bins), and we count how many of the data fall in each of the bins. It shows the shape of data, which can further be used in tasks like classification, etc.
      * **Example:** We can plot a histogram of the age column of customers, and see if the dataset is biased toward a certain age group or not.
    - **(iv) Clustering (Non-Parametric):**
      * **How It Works:** Groups similar data together, so we can analyze the group, instead of dealing with individual points. It is a very powerful technique for exploratory data analysis.
      * **Example:** In the iris flower dataset, we can create three clusters that closely resemble the three flower classes.
    - **(v) Sampling (Non-Parametric):**
        - **How it works:** Selects a representative subset of the data. This is like taking a poll to represent the opinions of a larger population.
        - **Example:** Instead of analyzing all customer transactions, you might randomly select 10% of the transactions to analyze purchasing patterns. Different sampling methods (random, stratified, cluster) are used depending on the data and goals.
    - **(vi) Data Cube Aggregation (Non-Parametric):**
        - **How it works:** Summarizes data at different levels of granularity. Think of it like rolling up daily sales data into monthly, quarterly, or yearly totals.

        - **Example:** In a retail database, you might have individual transaction details. Data cube aggregation allows you to quickly see total sales by product category, by region, by month, or by combinations of these factors.

    - **(vii) Data Compression:**
      * It employs modification, encoding or converting the structure of data in a way that consumes less space.
      * **Example:** Compressing large size files like videos, text and audio.

Key Considerations When Choosing a Data Reduction Method:

- **Data Type:** Are you working with numerical data, categorical data, or a mix?
- **Data Distribution:** Is the data normally distributed, skewed, or does it have outliers?
- **Goal:** Are you trying to improve model performance, reduce storage, speed up analysis, or gain insights?
- **Loss of Information:** Some methods (like PCA and lossy compression) involve some loss of information. Is this acceptable for your application?
- **Computational Cost:** Some methods (like PCA) can be computationally expensive, especially for very large datasets.

By understanding these methods and their trade-offs, you can effectively reduce the size and complexity of your data while preserving the crucial information needed for analysis and decision-making.