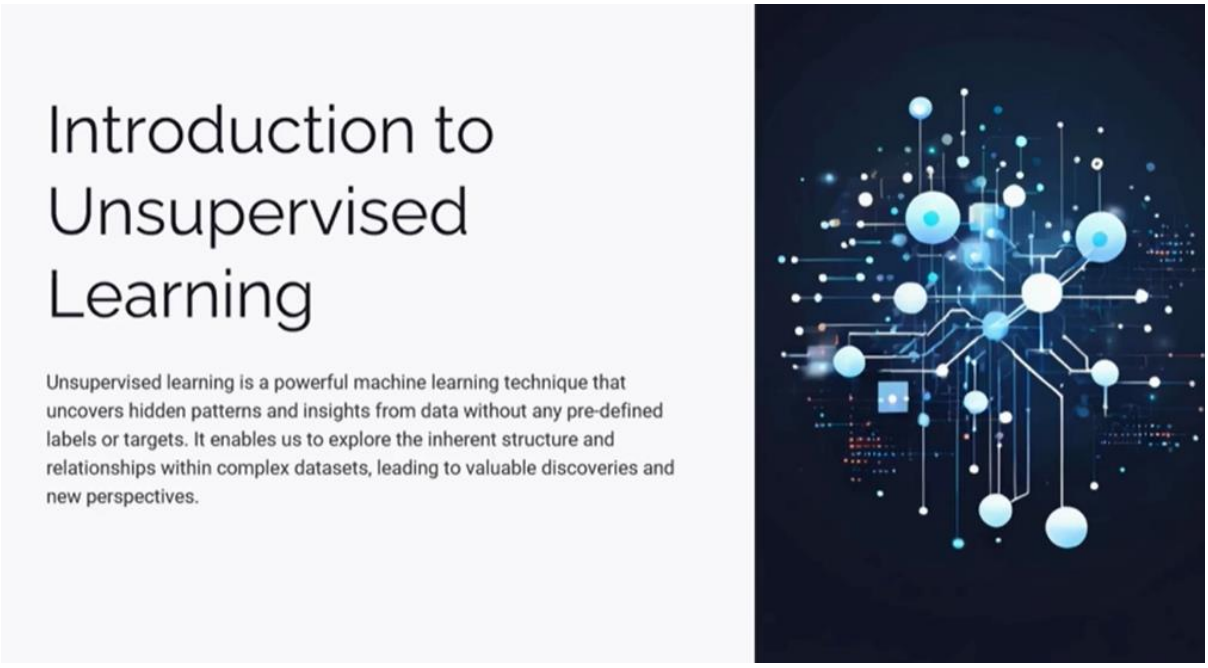
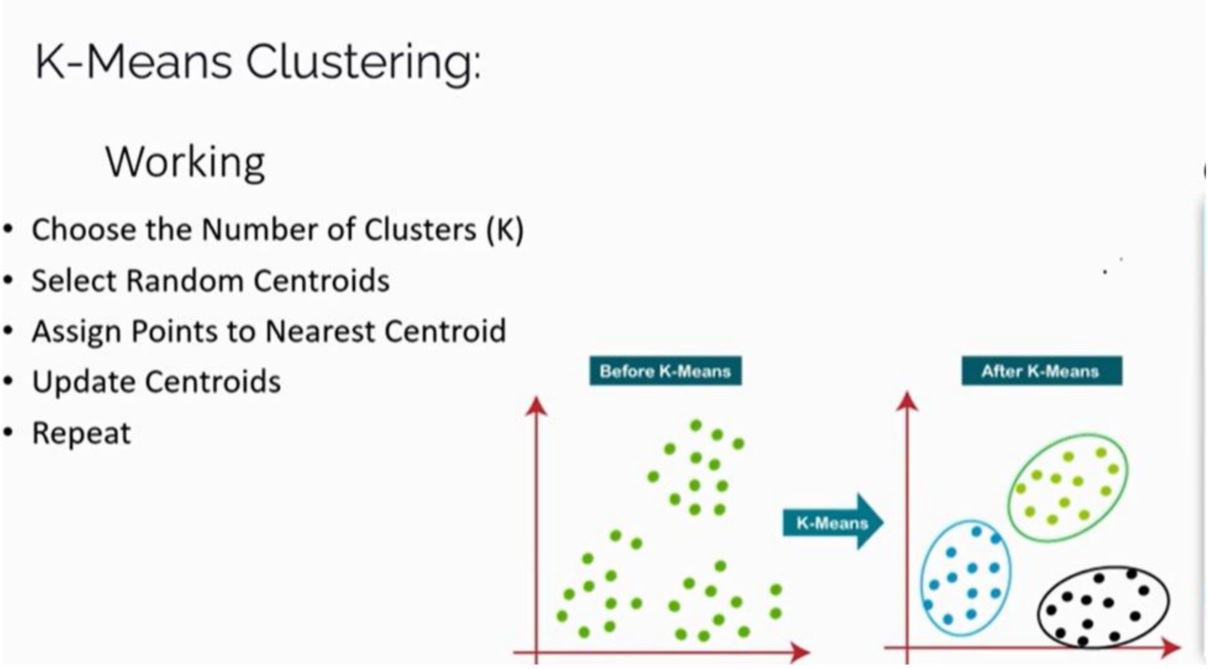


# Introduction to Unsupervised Learning

Unsupervised learning is a powerful machine learning technique that uncovers hidden patterns and insights from data without any pre-defined labels or targets. It enables us to explore the inherent structure and relationships within complex datasets, leading to valuable discoveries and new perspectives.

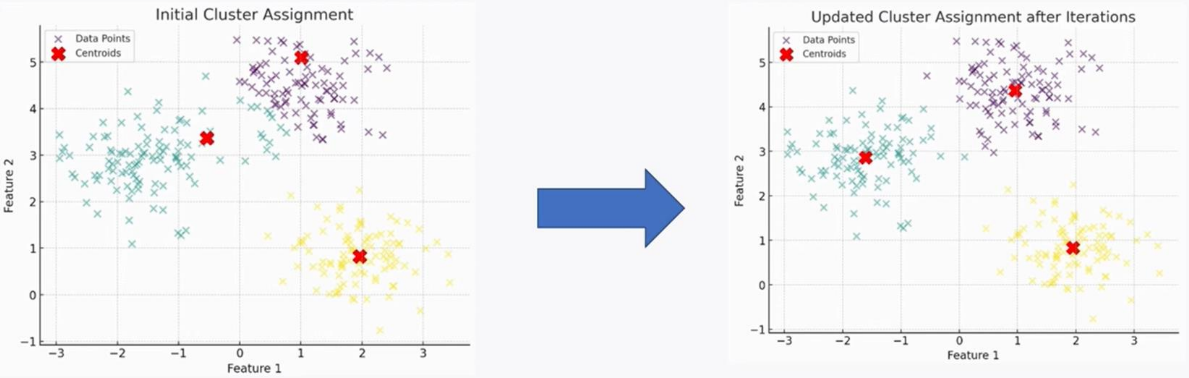


## K-Means Clustering:

### Working

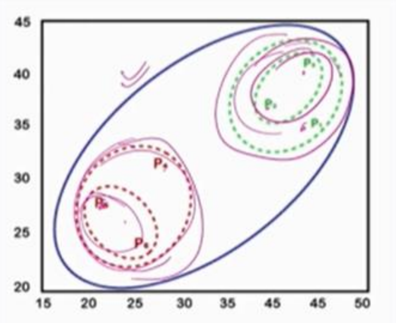
- Choose the Number of Clusters (K)
- Select Random Centroids
- Assign Points to Nearest Centroid
- Update Centroids
- Repeat

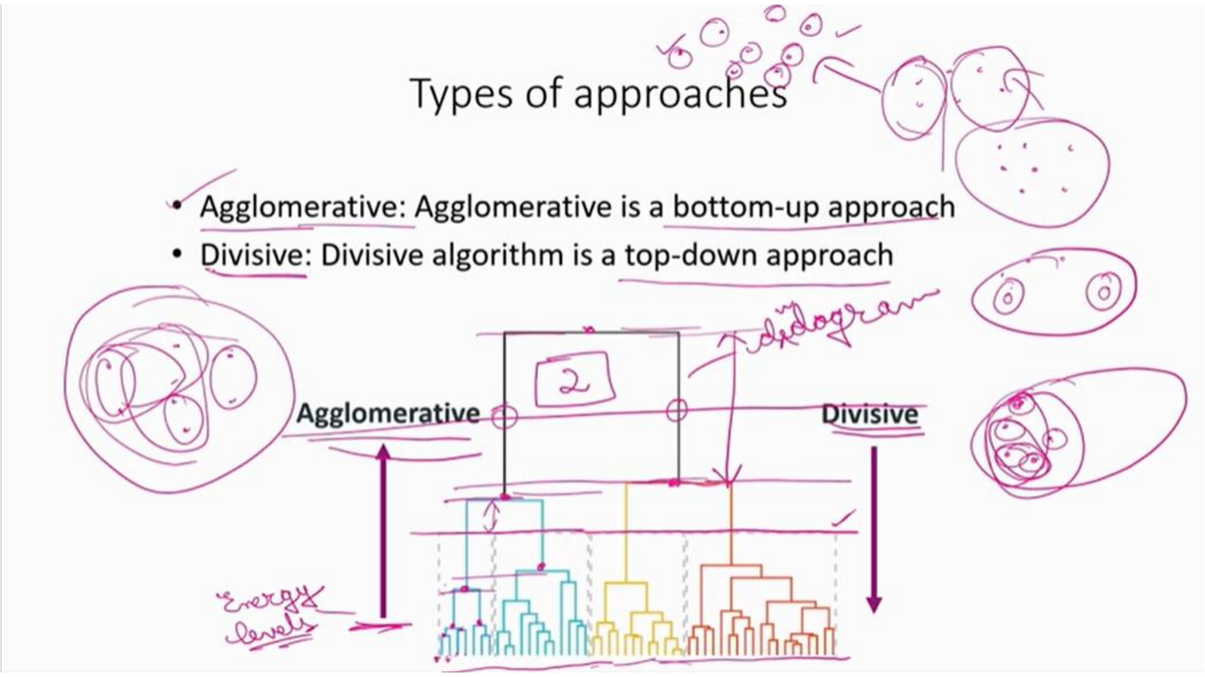
Before K-Means
K-Means
After K-Means



## Hierarchical Clustering:

- develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram





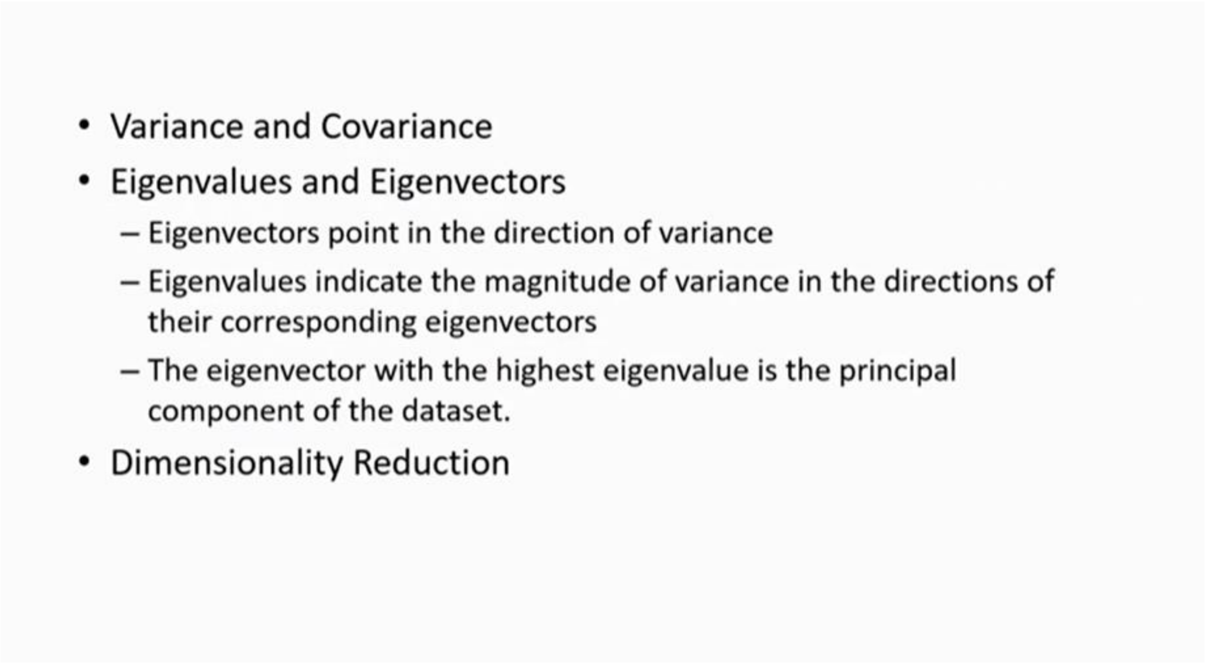
## Types of approaches

- Agglomerative: Agglomerative is a bottom-up approach
- Divisive: Divisive algorithm is a top-down approach

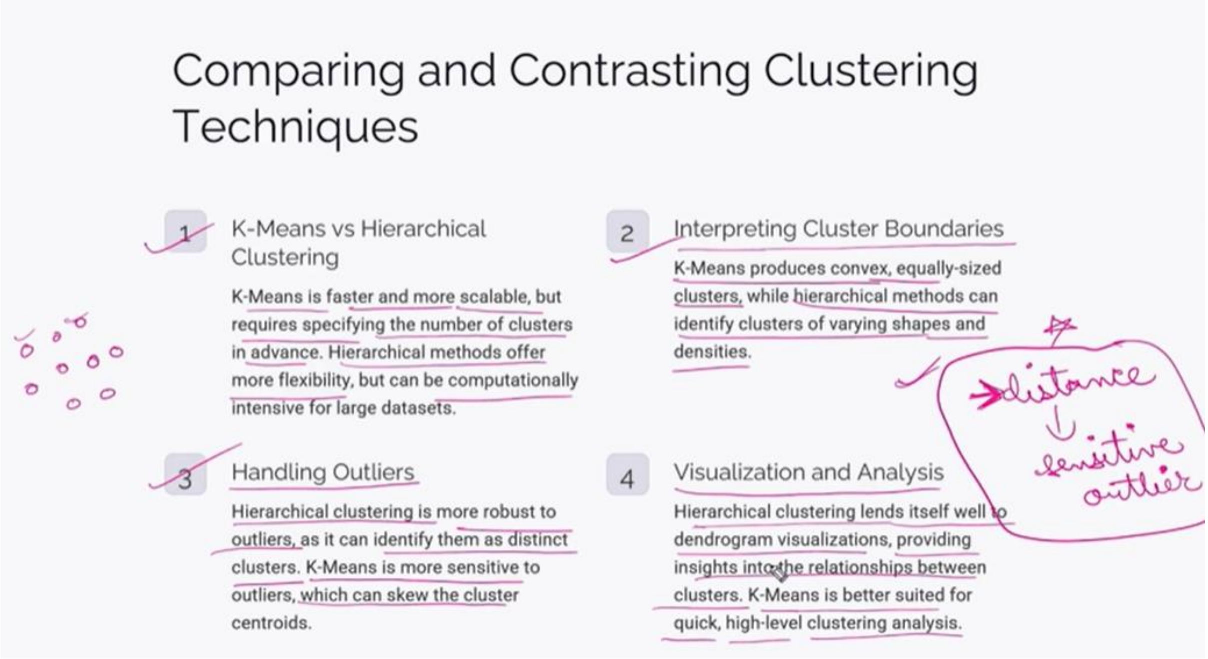
## Principal Component Analysis (PCA): Dimensionality Reduction

PCA is a powerful technique for dimensionality reduction, allowing complex high-dimensional data to be projected onto a lower-dimensional subspace while preserving the maximum amount of variance.

By identifying the principal components - the directions of greatest variance in the data - PCA can significantly reduce the number of features, simplifying data analysis and visualization.



- Variance and Covariance
- Eigenvalues and Eigenvectors
  - Eigenvectors point in the direction of variance
  - Eigenvalues indicate the magnitude of variance in the directions of their corresponding eigenvectors
  - The eigenvector with the highest eigenvalue is the principal component of the dataset.
- Dimensionality Reduction



# Comparing and Contrasting Clustering Techniques

1 K-Means vs Hierarchical Clustering

K-Means is faster and more scalable, but requires specifying the number of clusters in advance. Hierarchical methods offer more flexibility, but can be computationally intensive for large datasets.

2 Interpreting Cluster Boundaries

K-Means produces convex, equally-sized clusters, while hierarchical methods can identify clusters of varying shapes and densities.

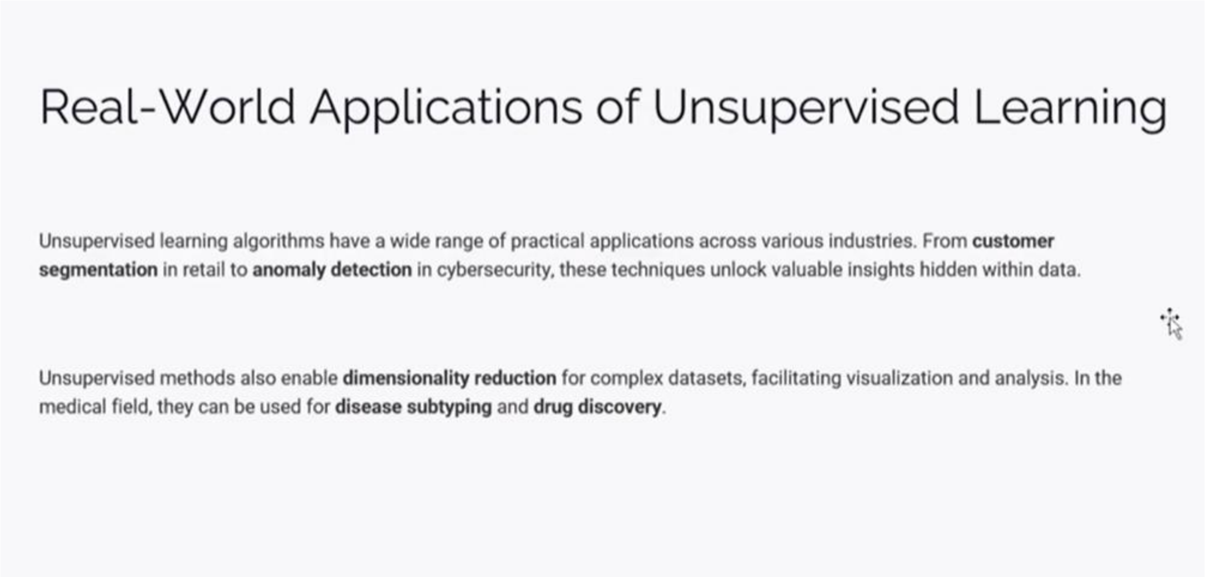
3 Handling Outliers

Hierarchical clustering is more robust to outliers, as it can identify them as distinct clusters. K-Means is more sensitive to outliers, which can skew the cluster centroids.

4 Visualization and Analysis

Hierarchical clustering lends itself well to dendrogram visualizations, providing insights into the relationships between clusters. K-Means is better suited for quick, high-level clustering analysis.

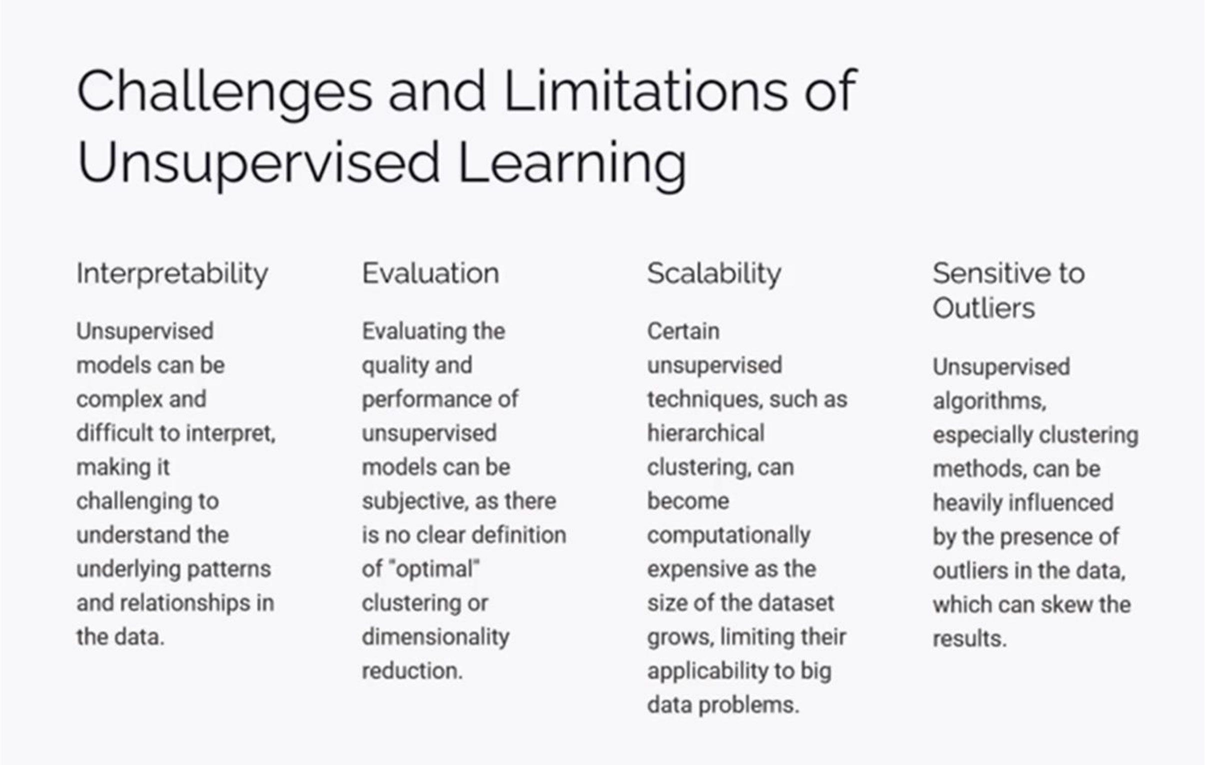
→distance
↓
sensitive
outlier



# Real-World Applications of Unsupervised Learning

Unsupervised learning algorithms have a wide range of practical applications across various industries. From **customer segmentation** in retail to **anomaly detection** in cybersecurity, these techniques unlock valuable insights hidden within data.

Unsupervised methods also enable **dimensionality reduction** for complex datasets, facilitating visualization and analysis. In the medical field, they can be used for **disease subtyping** and **drug discovery**.



# Challenges and Limitations of Unsupervised Learning

### Interpretability

Unsupervised models can be complex and difficult to interpret, making it challenging to understand the underlying patterns and relationships in the data.

### Evaluation

Evaluating the quality and performance of unsupervised models can be subjective, as there is no clear definition of "optimal" clustering or dimensionality reduction.

### Scalability

Certain unsupervised techniques, such as hierarchical clustering, can become computationally expensive as the size of the dataset grows, limiting their applicability to big data problems.

### Sensitive to Outliers

Unsupervised algorithms, especially clustering methods, can be heavily influenced by the presence of outliers in the data, which can skew the results.