# Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA) –project original data in smaller space
    - Attribute subset selection – irrelevant/weakly relevant/ redundant attributes are detected and removed
  - Numerosity reduction - replace the original data volume by alternative, smaller forms of data representation.
    - Parametric -Regression and log-linear models
    - Non-parametric -histograms, clustering, sampling, and data cube aggregation
  - Data compression

# Data Reduction 1: Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

# Principal Component Analysis (PCA)

## Introduction

Principal component analysis (PCA) is a standard tool in modern data analysis - in diverse fields from neuroscience to computer graphics.

It is very useful method for extracting relevant information from confusing data sets.

# Definition

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

The number of principal components is less than or equal to the number of original variables.

# Goals

- The main goal of a PCA analysis is to identify patterns in data
- PCA aims to detect the correlation between variables.
- It attempts to reduce the dimensionality.

# Dimensionality Reduction

It reduces the dimensions of a d-dimensional dataset by projecting it onto a (k)-dimensional subspace (where k<d) in order to increase the computational efficiency while retaining most of the information.

# Transformation

This transformation is defined in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the next highest possible variance.

# PCA Approach

- Standardize the data.
- Perform Singular Vector Decomposition to get the Eigenvectors and Eigenvalues.
- Sort eigenvalues in descending order and choose the k- eigenvectors
- Construct the projection matrix from the selected k- eigenvectors.
- Transform the original dataset via projection matrix to obtain a k-dimensional feature subspace.

# Applications of PCA :

- Interest Rate Derivatives Portfolios
- Neuroscience

# Linear Discriminant Analysis (LDA)

## Introduction

Linear Discriminant Analysis (LDA) is used to solve dimensionality reduction for data with higher attributes

- Pre-processing step for pattern-classification and machine learning applications.
- Used for feature extraction.
- Linear transformation that maximize the separation between multiple classes.
- “Supervised” - Prediction agent

# Feature Subspace :

To reduce the dimensions of a d-dimensional data set by projecting it onto a (k)-dimensional subspace

(where k < d)

Feature space data is well represented?

- Compute eigen vectors from dataset
- Collect them in scatter matrix
- Generate *k*-dimensional data from d-dimensional dataset.

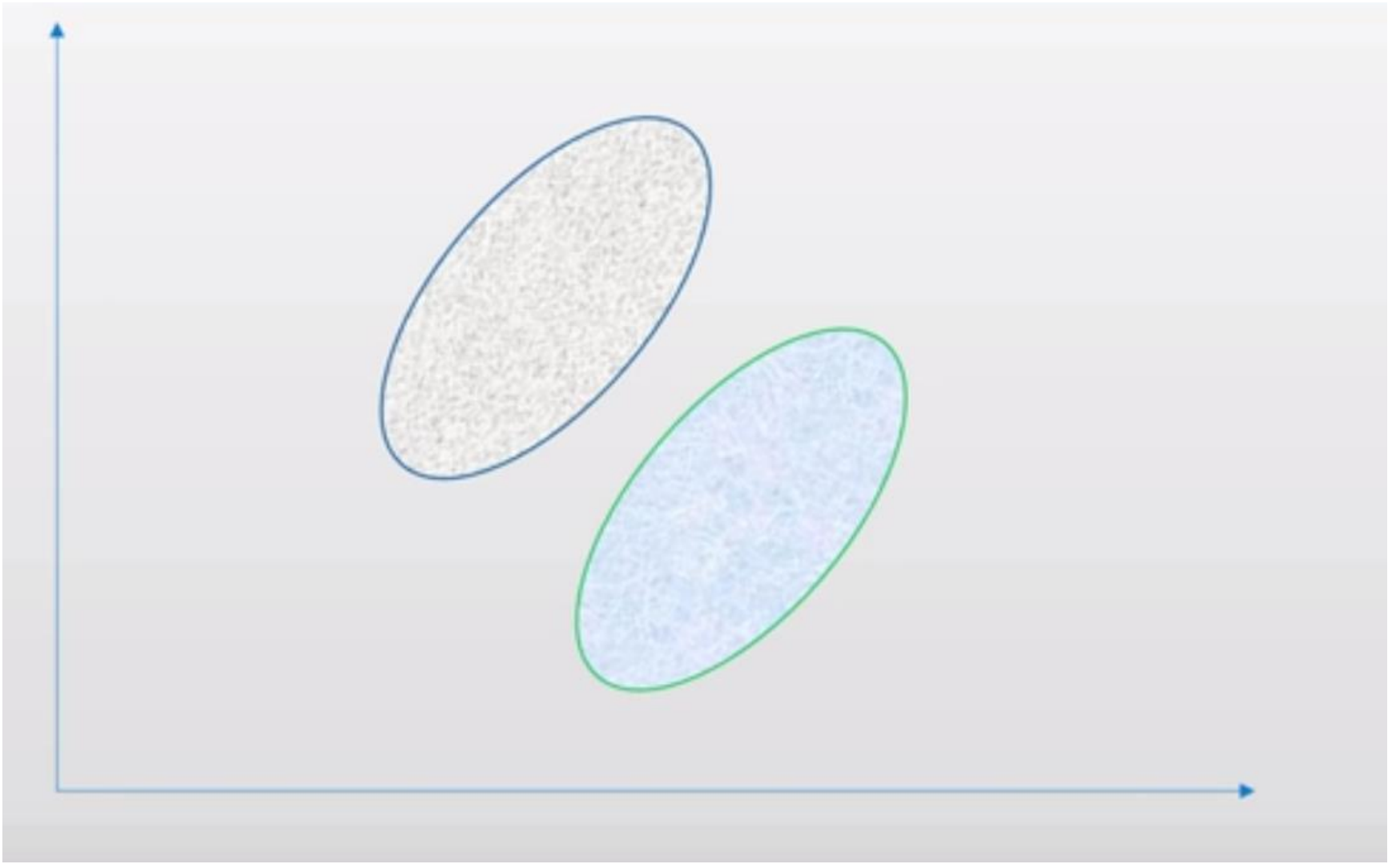
## Scatter Matrix:

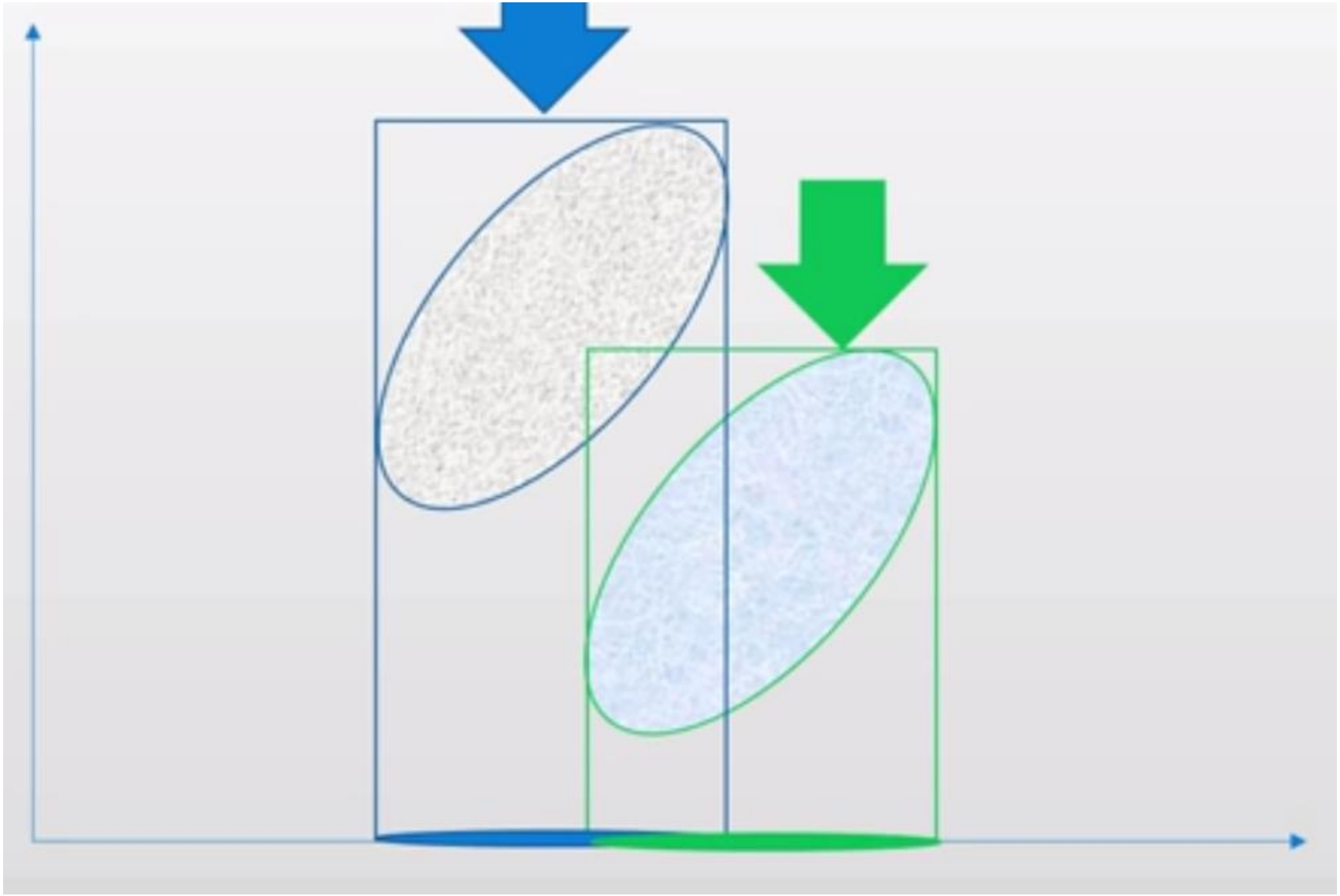
- Within class scatter matrix
- In between class scatter matrix

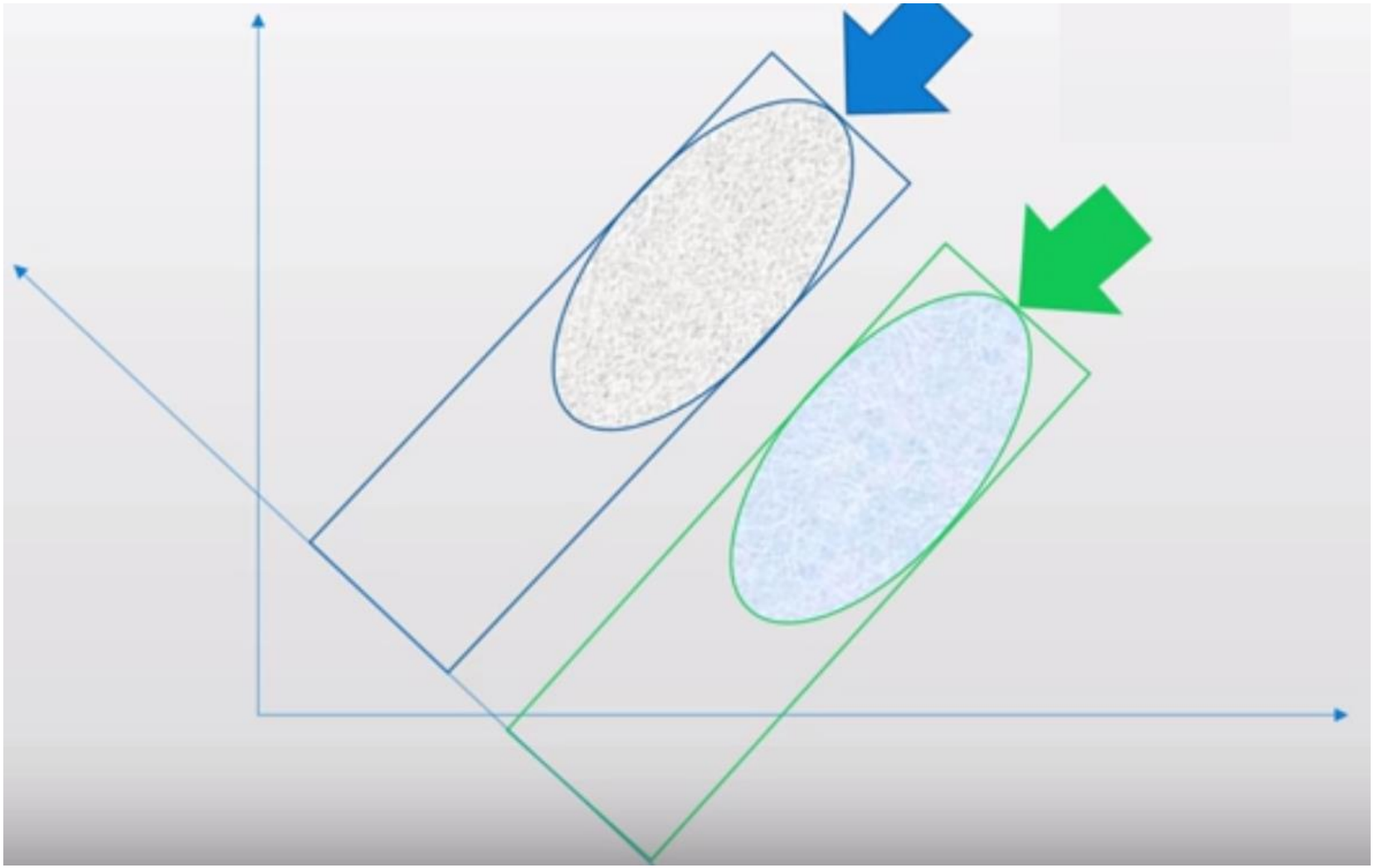
$$S_W = \sum_{i=1}^{c} S_i$$

$$S_B = \sum_{i=1}^{c} N_i (\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

Maximize the between class measure & minimize the within class measure.







# LDA steps:

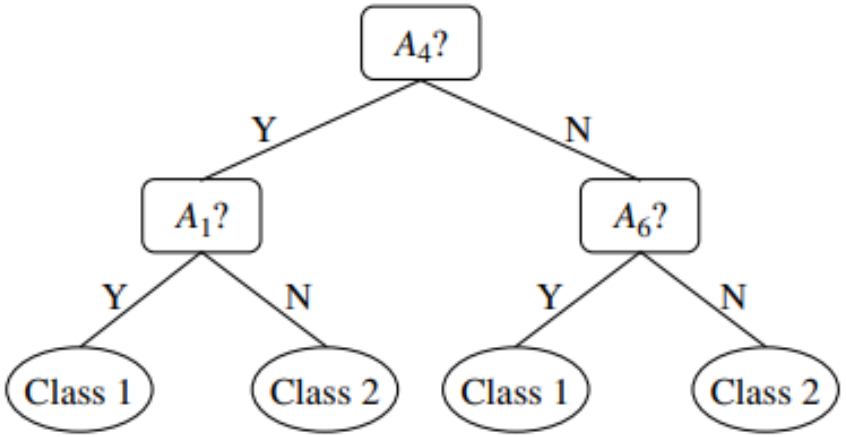
1. Compute the d-dimensional mean vectors.
2. Compute the scatter matrices
3. Compute the eigenvectors and corresponding eigenvalues for the scatter matrices.
4. Sort the eigenvalues and choose those with the largest eigenvalues to form a d×k dimensional matrix
5. Transform the samples onto the new subspace.

# Attribute Subset Selection

- Another way to reduce dimensionality of data
- Redundant attributes
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g. purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
  - Contain no information that is useful for the data mining task at hand
  - E.g. students' ID is often irrelevant to the task of predicting students' GPA

# Heuristic Search in Attribute Selection

- There are $2^d$ possible attribute subsets of $d$ attributes
- Exhaustive search for the optimal subset -expensive
- Typical heuristic attribute selection methods:(greedy)
  - Step-wise forward selection:
    - starts with an empty set of attributes
    - The best of the original attributes is determined and added to the reduced set.
  - Stepwise backward elimination:
    - Starts with the full set of attributes
    - At each step, it removes the worst attribute remaining in the set
  - Combination of forward selection and backward elimination
  - Decision tree induction

| Forward selection | Backward elimination | Decision tree induction |
|---|---|---|
| Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>Initial reduced set:<br>$\{\}$<br>$\Rightarrow \{A_1\}$<br>$\Rightarrow \{A_1, A_4\}$<br>$\Rightarrow$ Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ | Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$<br>$\Rightarrow \{A_1, A_4, A_5, A_6\}$<br>$\Rightarrow$ Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ | Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br>  <br><br>$\Rightarrow$ Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ |

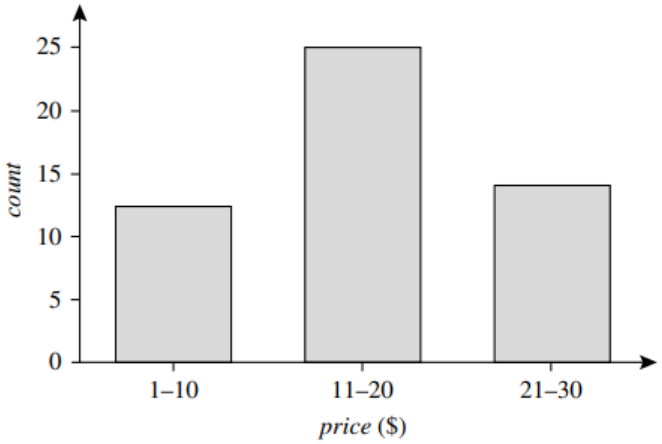
Greedy (heuristic) methods for attribute subset selection.

# Data Reduction 2: Numerosity Reduction

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods** (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Ex.: Log-linear models—obtain value at a point in *m*-D space as the product on appropriate marginal subspaces
- **Non-parametric** methods
  - Do not assume models
  - Major families: histograms, clustering, sampling, …

# Histogram Analysis

- Histograms use binning to approximate data distributions
- A histogram for attribute, A, partitions into disjoint subsets, called buckets or bins
- If each bucket represents only a single attribute– singleton buckets
- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)




An equal-width histogram for *price*, where values are aggregated so that each bucket has a uniform width of $10.

# Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

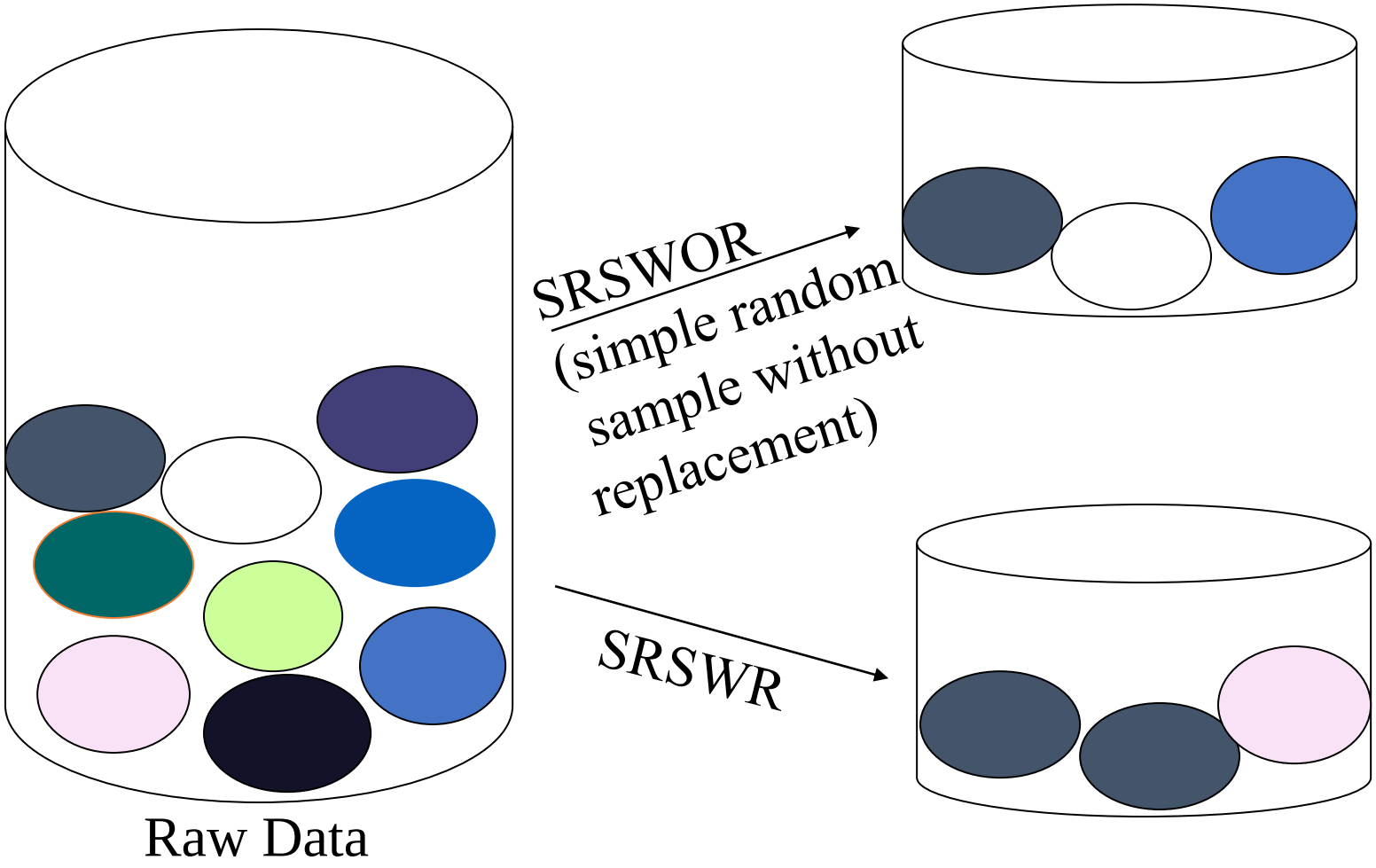
# Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling

# Types of Sampling

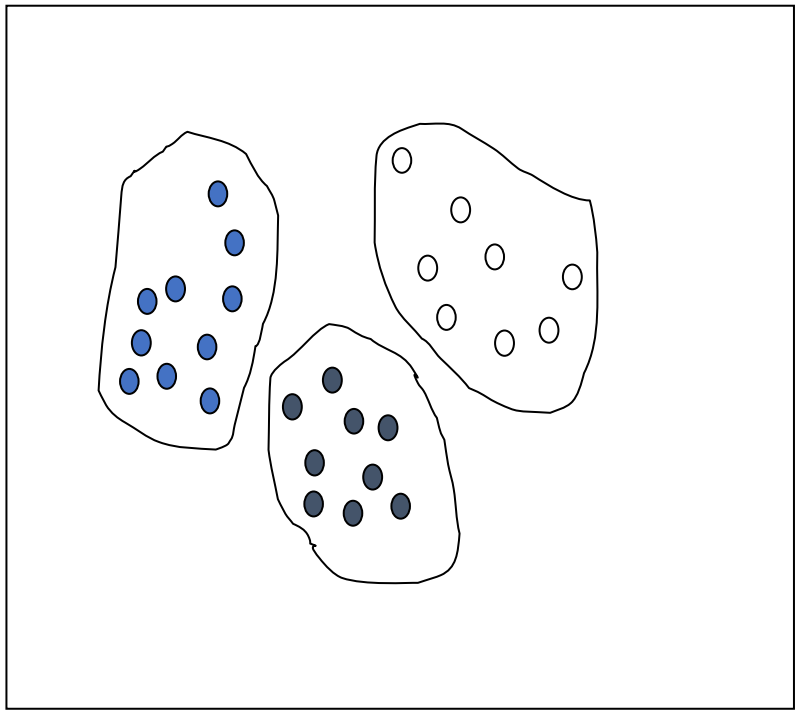
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

# Sampling: With or without Replacement



# Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample