Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set
  - 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)
    - Feature subset selection, feature creation
  - Numerosity reduction (some simply call it: Data Reduction)
    - Regression
    - Histograms, clustering, sampling
  - 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 and outlier analysis) becomes less meaningful
- **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

Wavelet Transformation

- Discrete wavelet transform (DWT)
  - For linear signal processing and multi-resolution analysis
- Compressed approximation
  - Store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better *lossy* compression

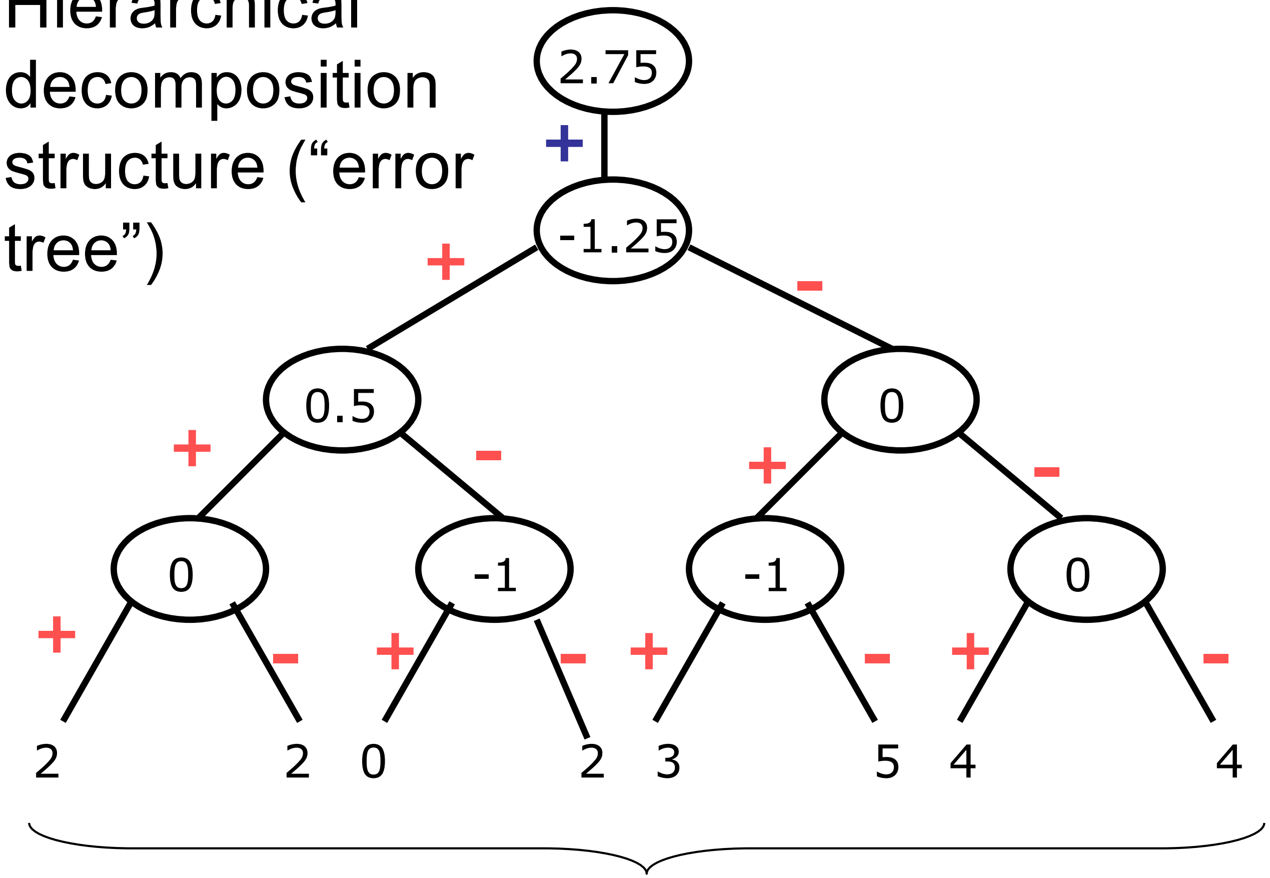
Wavelet Decomposition

- Wavelets: A math tool for space-efficient hierarchical decomposition of functions
- S = [2, 2, 0, 2, 3, 5, 4, 4] can be transformed to $S_\wedge$ = $[2^3/_4, -1^1/_4, {}^1/_2, 0, 0, -1, -1, 0]$
- Compression:
  - many small detail coefficients can be replaced by 0's
  - only the significant coefficients are retained

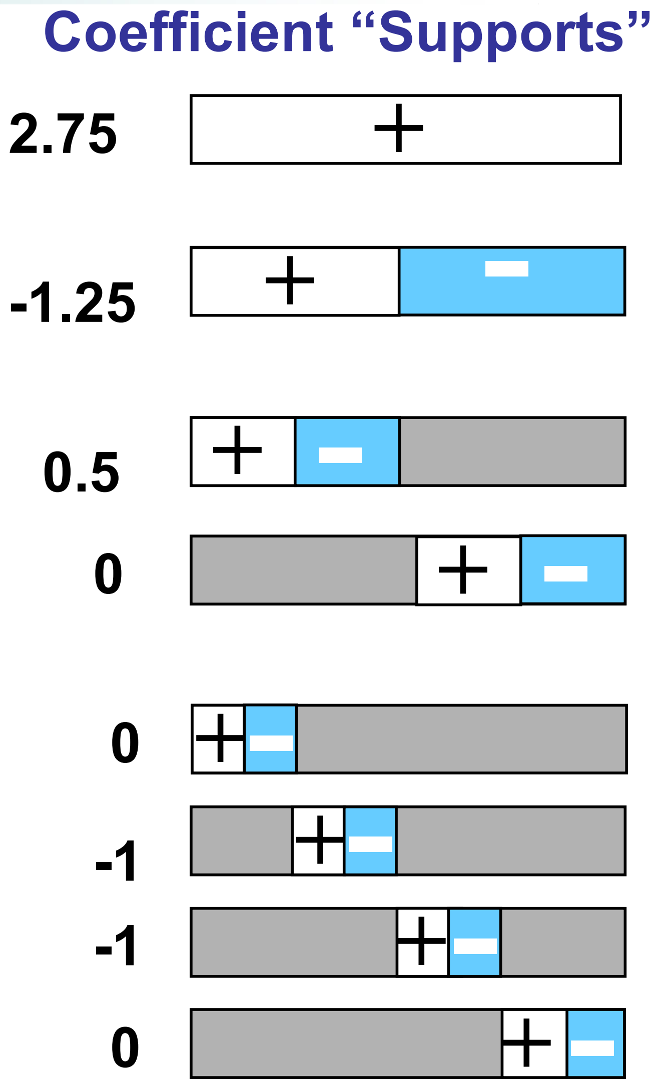
Resolution	Averages	Detail Coefficients
8	$[2, 2, 0, 2, 3, 5, 4, 4]$	
4	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
2	$[1\frac{1}{2}, 4]$	$[\frac{1}{2}, 0]$
1	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$

Haar Wavelet Coefficients

Hierarchical decomposition structure ("error tree")

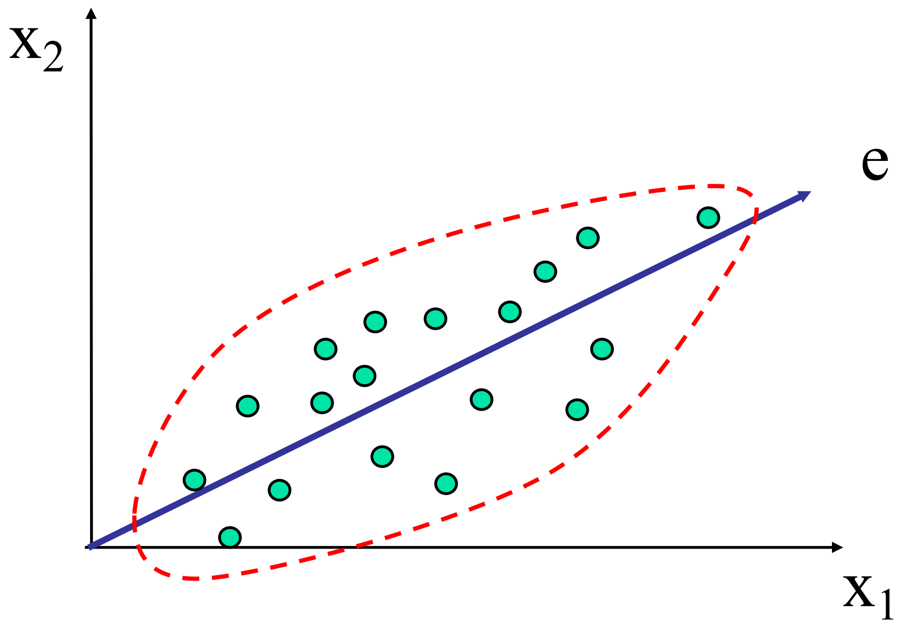


Original frequency distribution



Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- Original data are projected onto a much smaller space
  - Resulting in dimensionality reduction



Principal Component Analysis (Steps)

- Given *N* data vectors from *n*-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - The principal components are sorted in order of decreasing "significance" or strength
  - The size of the data can be reduced by eliminating the *weak components*, i.e., those with low strength
    - Using the strong principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only

Attribute Subset Selection

- Another way to reduce dimensionality of data
- *Redundant* attributes
  - 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

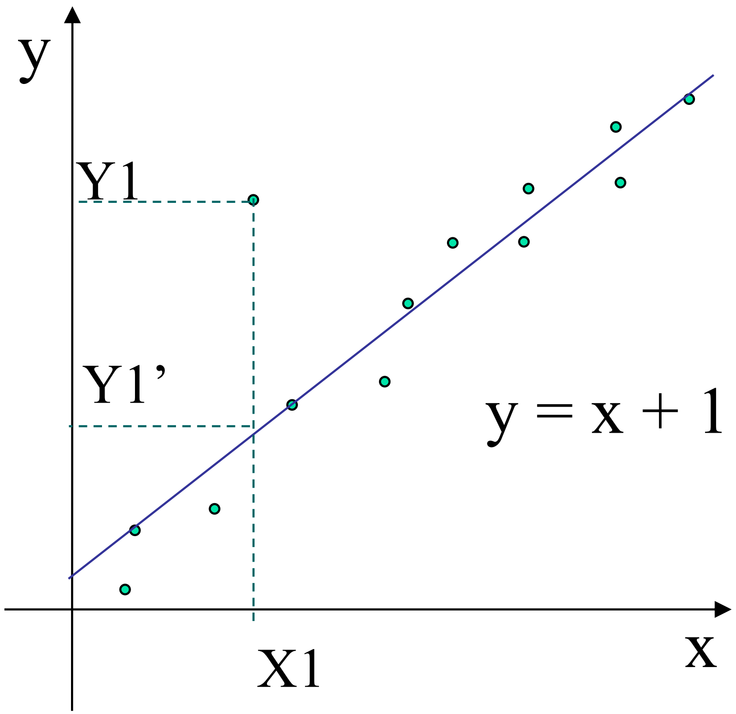
- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute

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
  - discard the data (except possible outliers)
- **Non-parametric** methods
  - Do not assume any models
  - Major families: histograms, clustering, and sampling

Regression Analysis

- Regression analysis
  - Modeling numerical data consisting of values of a ***dependent variable*** **(response variable)** and of one or more ***independent variables***
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly, the best fit is evaluated by using the ***least squares method***
  - But, other criteria have also been used
- Used for ***prediction*** (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships



Parametric Data Reduction: Regression

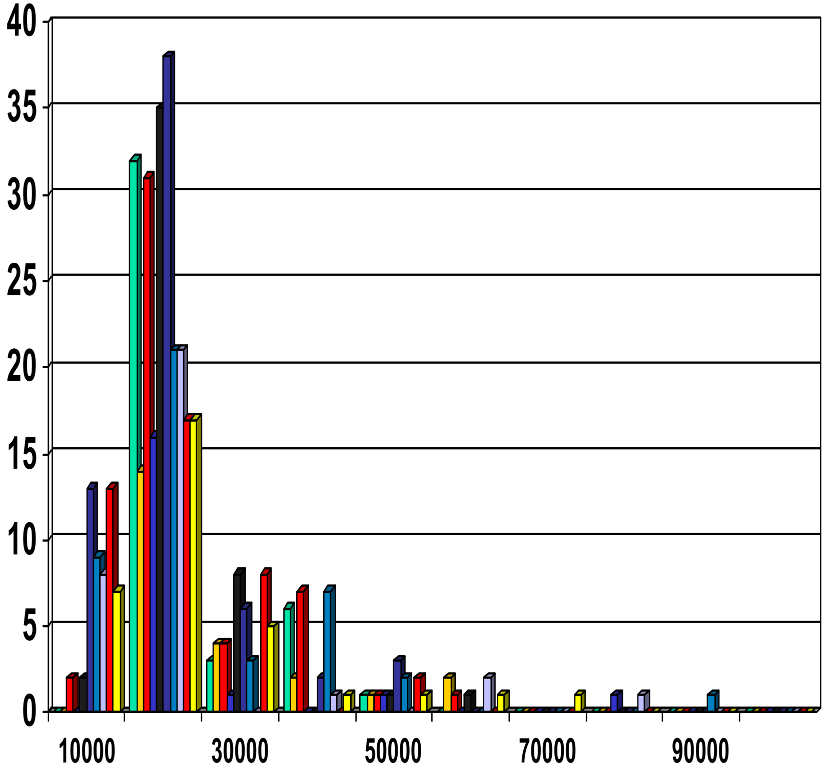
- **Linear regression**
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
- **Multiple regression**
  - Allows a dependent variable Y to be modeled as a linear function of two or more independent variables

Regression Analysis

- Linear regression: $Y = w X + b$
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Linear function involving more than one independent variables
  - Solved by SAS, SPSS, and S-Plus
- Nonlinear regression: $Y = b_0 + b_1 X + b_2 X^2$
  - Many nonlinear functions can be transformed into the above
  - By setting $X_1 = X$ and $X_2 = X^2$

Histogram Analysis

- Divide data into buckets and store *count* (or sum / average) for each bucket
- Partitioning rules:
  - Equal-width
    - Equal bucket range
  - Equal-frequency (or equal-depth)
    - Equal depth for buckets

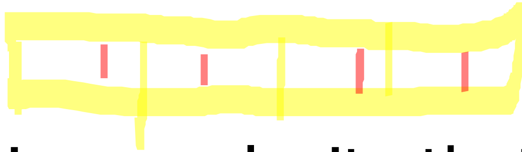


Clustering

- Partition data set into clusters based on similarity
- Then, 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
  - Cluster analysis will be studied in depth in Chapter 10

Sampling

- Sampling: obtaining a small set of samples *s* to represent the whole data set *N*

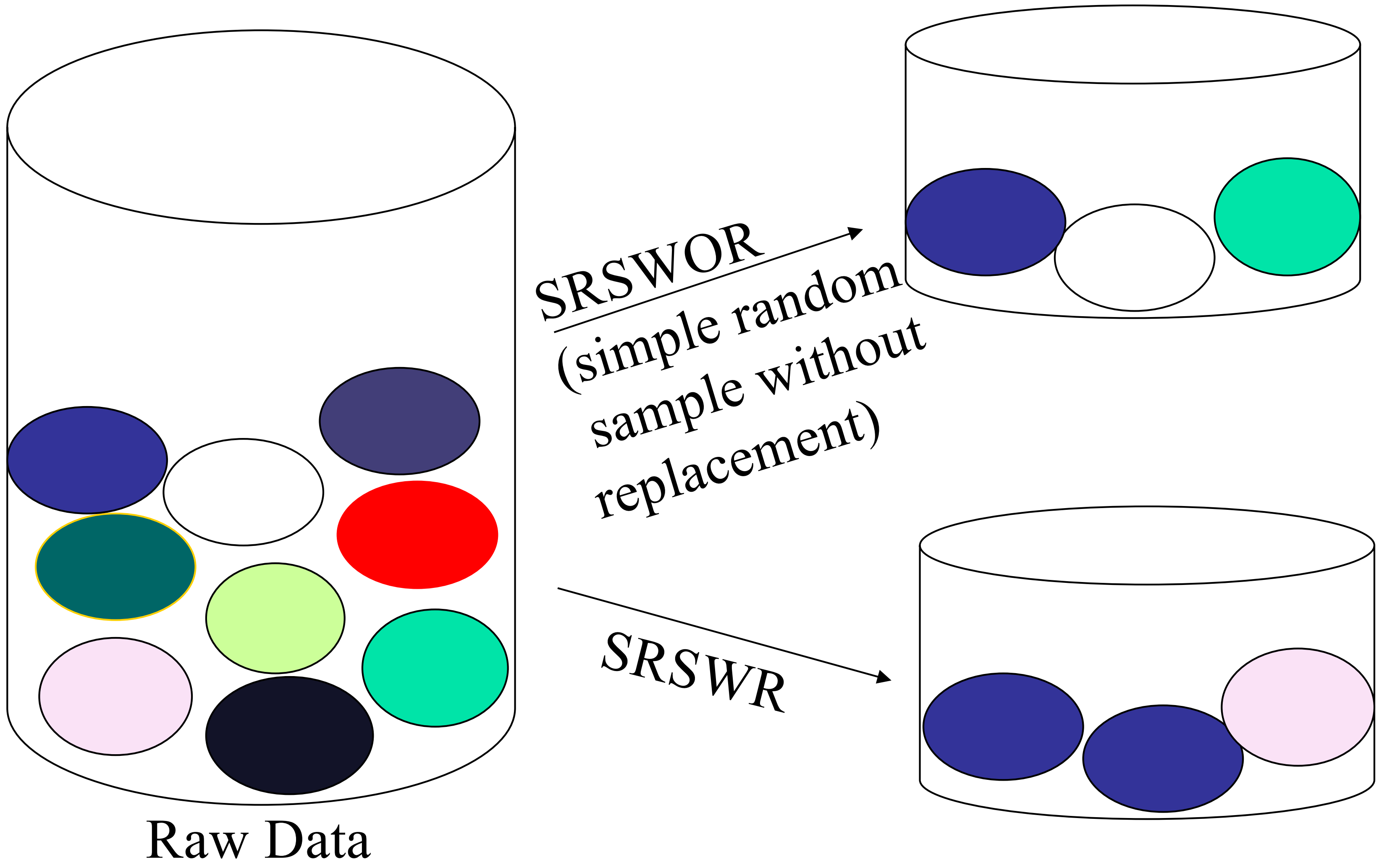


- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Key principle: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling
- Note: Sampling may not reduce database I/Os (page at a time)

Types of Sampling

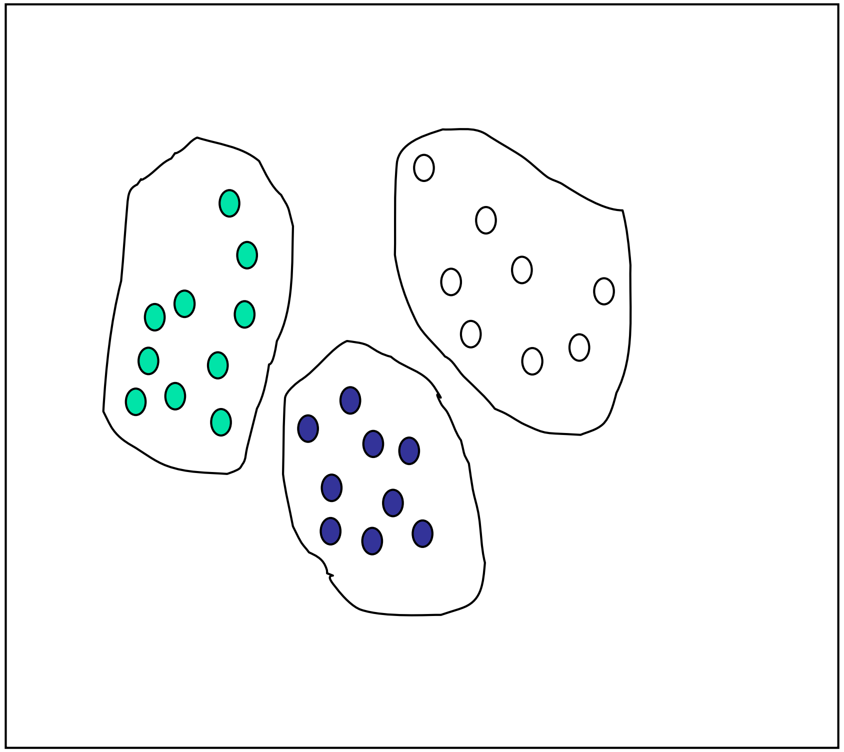
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is *removed* from the population
- **Sampling with replacement**
  - A selected object is *not removed* from the population
- **Stratified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

Sampling: With or without Replacement

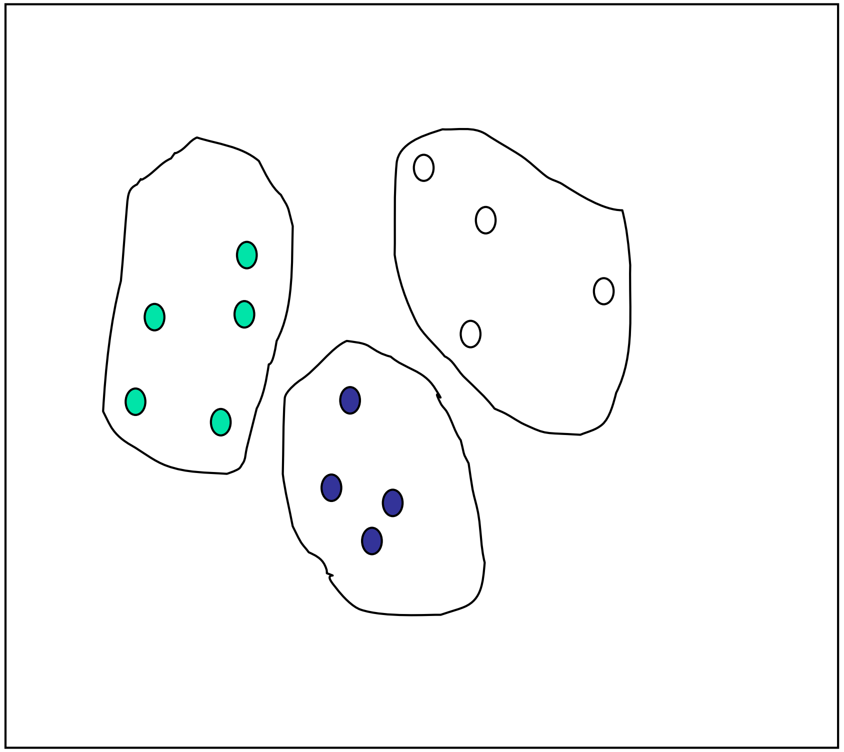


Sampling: Cluster or Stratified Sampling

Raw Data



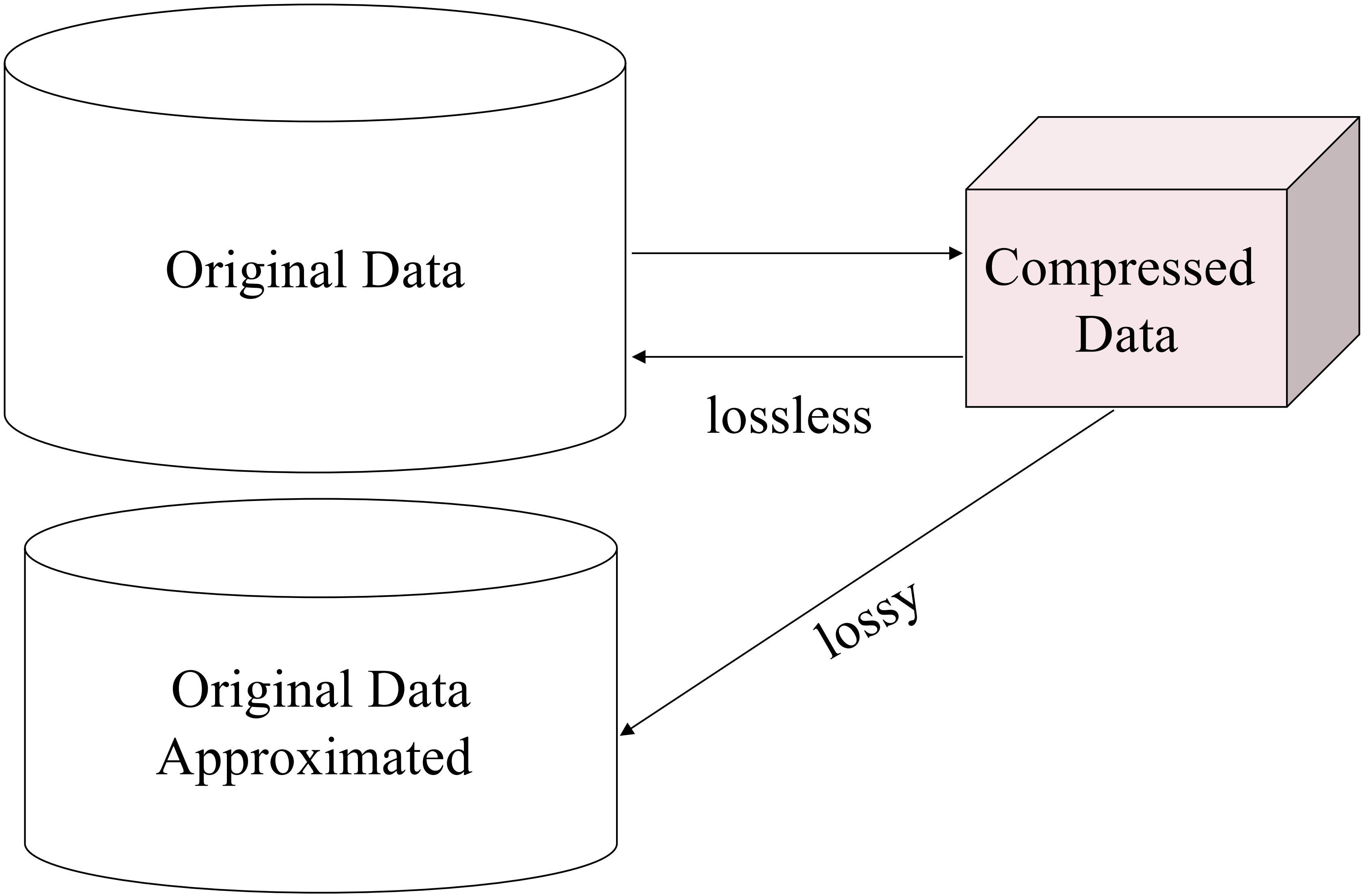
Stratified Sample



Data Reduction 3: Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
- Audio/video compression
  - Typically lossy compression, with progressive refinement
- Time sequence
  - Typically short and vary slowly with time
- Dimensionality and numerosity reduction may also be considered as forms of data compression

Data Compression



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Transformation

- Maps the entire set of values of a given attribute to a new set of replacement values
  - Each old value needs to be identified with one of the new values
- Methods
  - *Normalization*: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - *Discretization*: Concept hierarchy climbing

Normalization

- **Min-max normalization**: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$
- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$
- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|ν'|) < 1

Discretization

- Three types of attributes
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- Discretization: divide the range of a continuous attribute into intervals
  - Labels are assigned to intervals to replace actual data values
  - Effect of discretization
    - Data size is reduced
    - Similar values become identical
  - Used for further analysis, e.g., classification

Simple Discretization: Binning

- Equal-width (distance) partitioning
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward
  - Problems
    - Outliers may dominate presentation
    - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into N intervals, each containing approximately same number of samples
  - Good data scaling

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

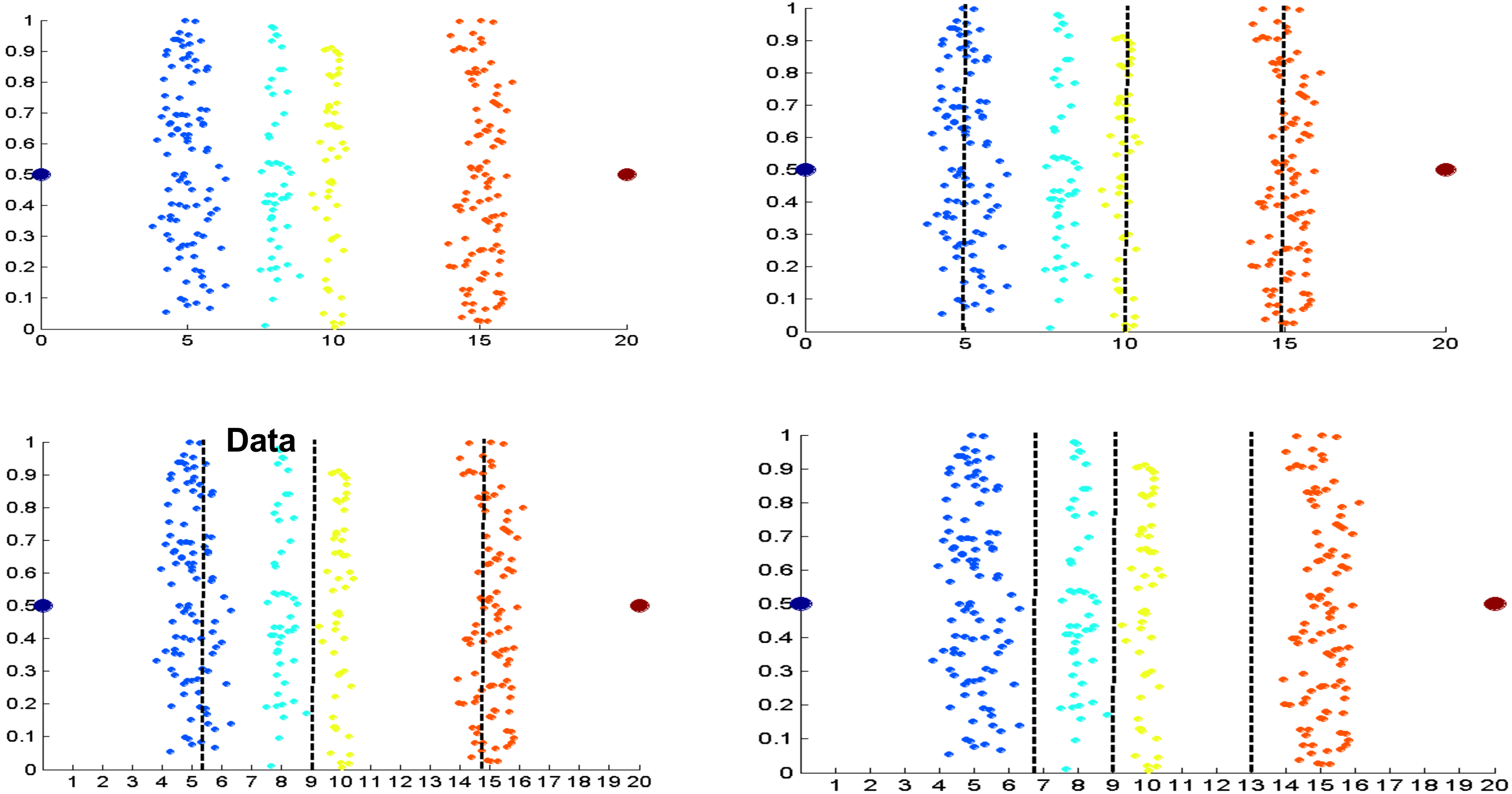
* Smoothing by **bin means**:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

* Smoothing by **bin boundaries**:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Discretization Without Using Class Labels (Binning vs. Clustering)



Equal frequency (binning)

K-means clustering leads to better results

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Summary

- **Data quality**: accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning**: e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization