Data Preprocessing

Content

- What & Why preprocess the data?
- Data cleaning
- Data integration
- Data transformation
- Data reduction

Data Preprocessing

It is a data mining technique that involves transforming raw data into an understandable format.

Why preprocess the data?

Data Preprocessing

- Data in the real world is:
  - incomplete: lacking values, certain attributes of interest, etc.
  - noisy: containing errors or outliers
  - inconsistent: lack of compatibility or similarity between two or more facts.
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Measure of Data Quality

- Accuracy
- Completeness
- Consistency
- Timeliness
- Believability
- Value added
- Interpretability
- Accessibility

Data preprocessing techniques

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

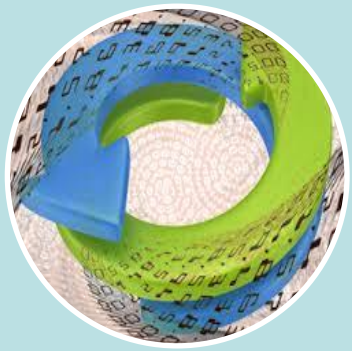
Major Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results

Data Preprocessing

Data Cleaning

Data Integration



Data Transformation

Data Reduction

Data Cleaning

Data Cleaning

"Data Cleaning attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the real world data."

- Missing Values
- Noisy Data
- Inconsistent Data

Data Cleaning - Missing Values

- Ignore the tuple
- Fill in the missing value manually
- Use a global constant
- Use attribute mean
- Use the most probable value

(decision tree, Bayesian Formalism)

Data Cleaning - Noisy Data

- Binning
- Clustering
- Combined computer and human inspection
- Regression

Data Cleaning - Inconsistent Data

- Manually, using external references
- Knowledge engineering tools

Few Important Terms

- Discrepancy Detection
  - Human Error
  - Data Decay
  - Deliberate Errors
- Metadata
- Unique Rules
- Null Rules

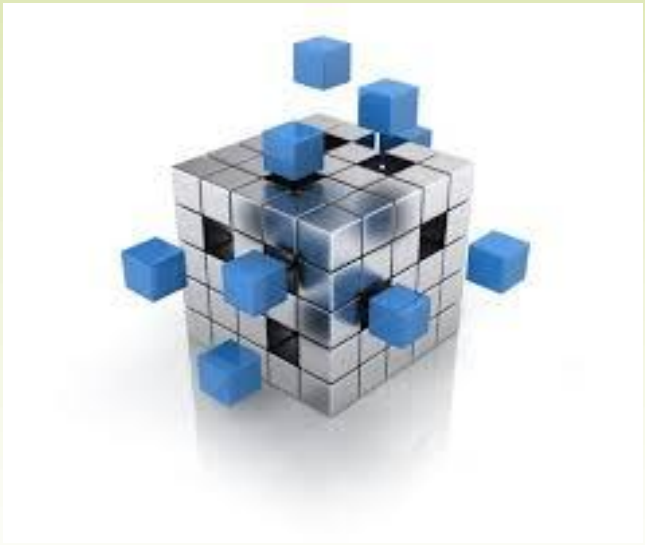
Data Integration

Data Integration

“Data Integration implies combining of data from multiple sources into a coherent data store(data warehouse). ”

Data Integration - Issues

- Entity identification problem
- Redundancy
- Tuple Duplication
- Detecting data value conflicts



Data Transformation

Data Transformation

"Transforming or consolidating data into mining suitable form is known as Data Transformation."

- Smoothing
- Aggregation
- Generalization
- Normalization
- Attribute construction

Handling Redundant Data in Data Integration

- Redundant data occur often when integration of multiple databases
  - The same attribute may have different names in different databases
  - One attribute may be a “derived” attribute in another table, e.g., annual revenue

Handling Redundant Data in Data Integration

- Redundant data may be able to be detected by correlation analysis
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing

Data Reduction

Data Reduction

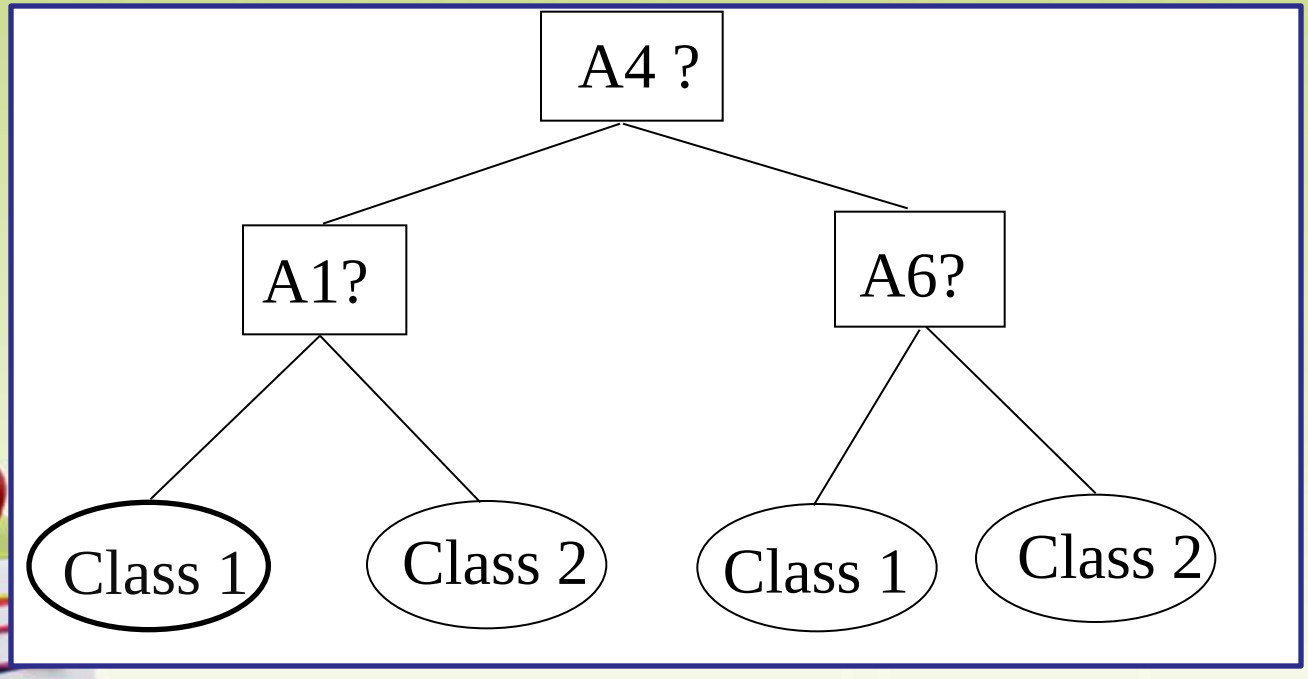
"Data reduction techniques are applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of base data."

Data Reduction - Strategies

- Data cube aggregation
- Dimension Reduction
- Data Compression
- Numerosity Reduction
- Discretization and concept hierarchy generation

Example of Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}

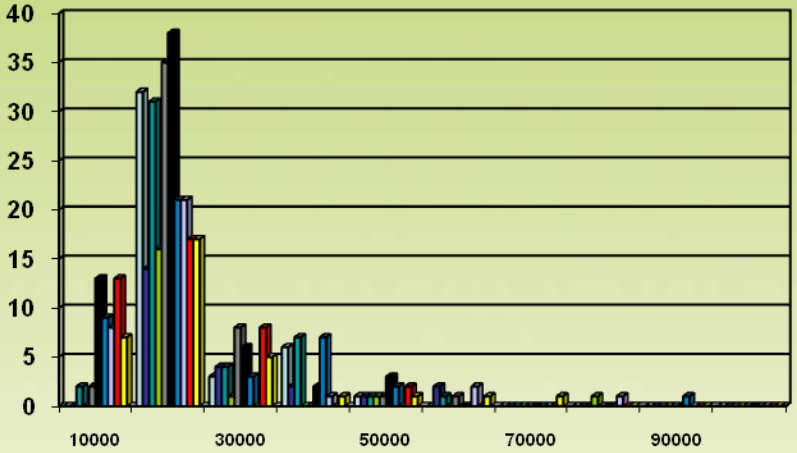


---------->

Reduced attribute set: {A1, A4, A6}

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension.
- Related to quantization problems.



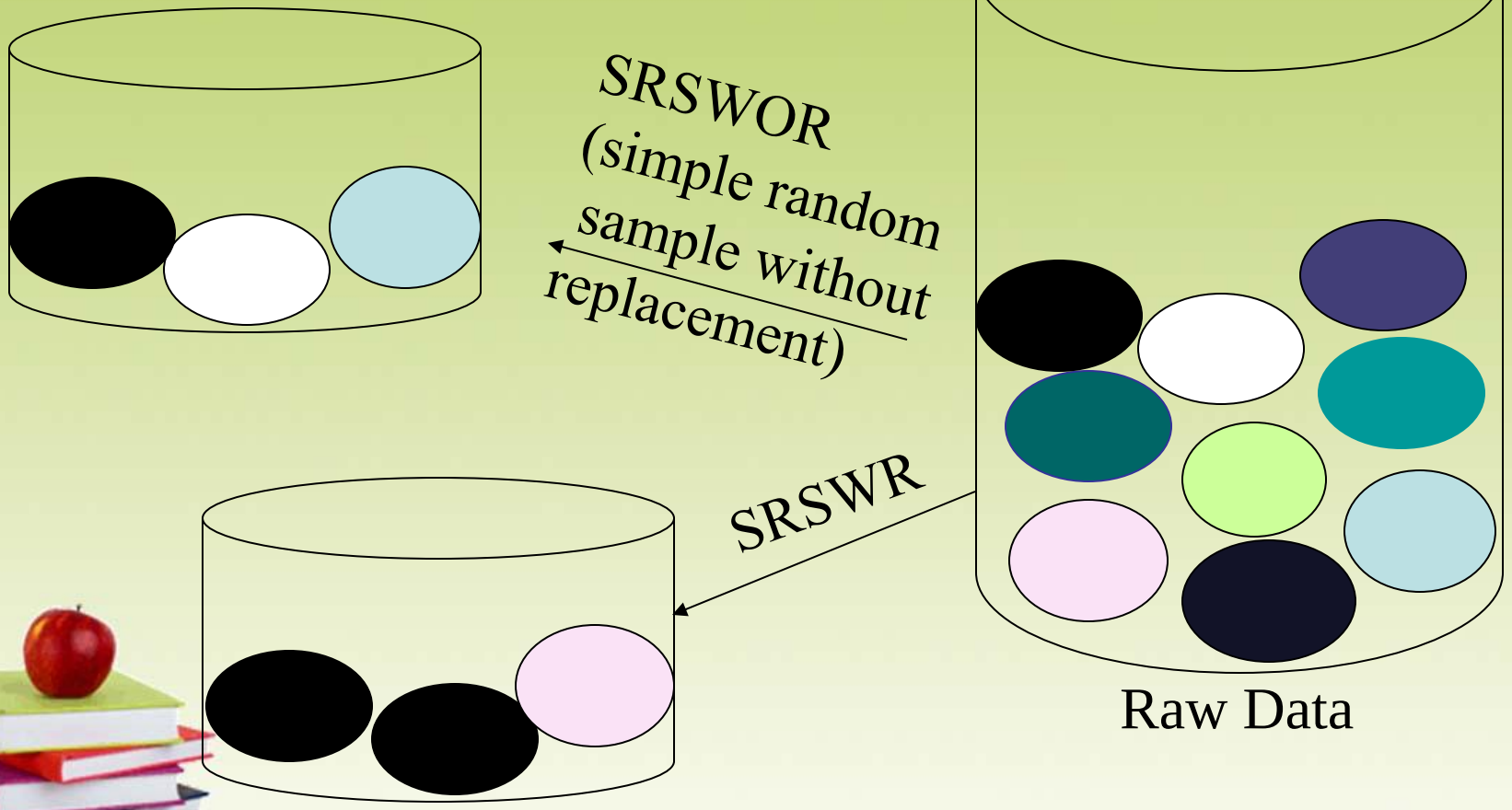
Clustering

- Partition data set into clusters, and one can store cluster representation only
- Can be very effective if data is clustered.
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures.

Sampling

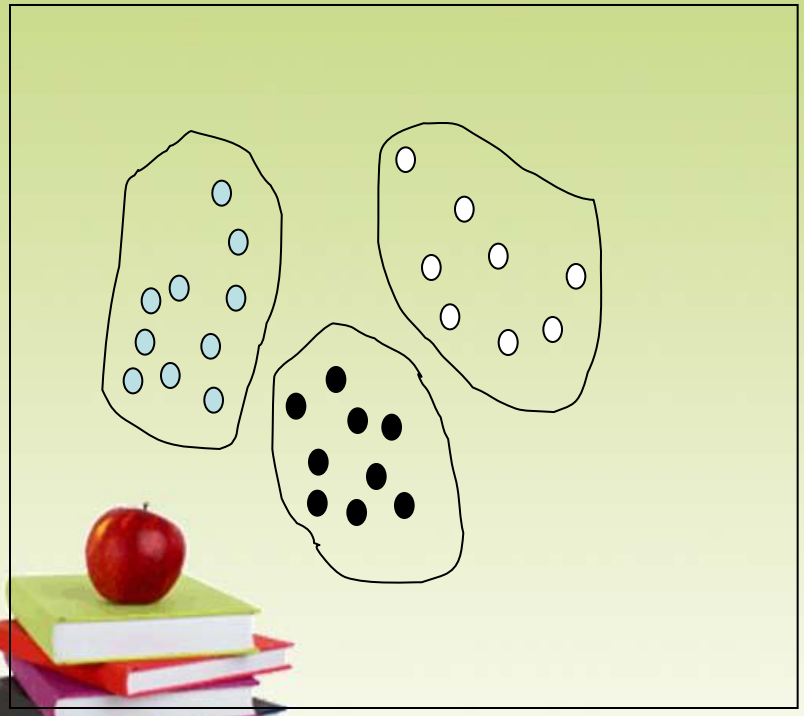
- Allows a large data set to be represented by a much smaller of the data.
- Let a large data set D, contains N tuples.
- Methods to reduce data set D:
  - Simple random sample without replacement (SRSWOR)
  - Simple random sample with replacement (SRSWR)
  - Cluster sample
  - Stright sample

Sampling

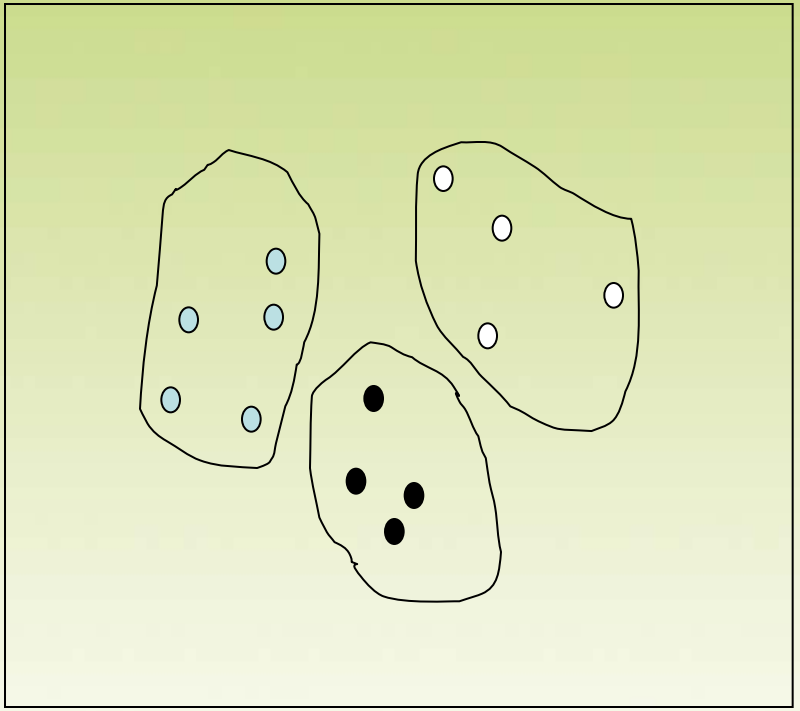

SRSWOR
(simple random
sample without
replacement)
SRSWR
Raw Data

Sampling

Raw Data



Cluster/Stratified Sample




Thank You