Data Preprocessing and Normalization

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names

- No quality data, no quality results!
  - Quality decisions must be based on quality data

Some Tasks in Data Preprocessing

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data transformation
  - Normalization (scaling to a specific range)
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
  - Data discretization: with particular importance, especially for numerical data
  - Data aggregation, dimensionality reduction, data compression,generalization

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not registering history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value manually: tedious + infeasible?
- Use a global constant to fill in the missing value: e.g., "unknown", a new class?!
- Use the attribute mean to fill in the missing value
- Use the attribute mean for all samples of the same class to fill in the missing value: smarter
- Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Noisy Data

- Q: What is noise?
- A: Random error in a measured variable.

- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention

- Other data problems which requires data cleaning
  - duplicate records
  - inconsistent data

How to Handle Noisy Data?

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Semi-automated method: combined computer and human inspection
  - detect suspicious values and check manually
- Regression (Curve fitting)
  - smooth by fitting the data into regression functions

Data Transformation: Normalization

Particularly useful for classification (Neural Networks, Distance measurements, Nearest Neighbour classification, etc)

- min-max normalization

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

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

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

Data Reduction

- Problem:

  Database may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set.

- Solution?
  - Data reduction…

Data Reduction

- Obtains a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation
  - Dimensionality reduction
  - Data compression
  - Numerosity reduction
  - Discretization and concept hierarchy generation

Dimensionality Reduction

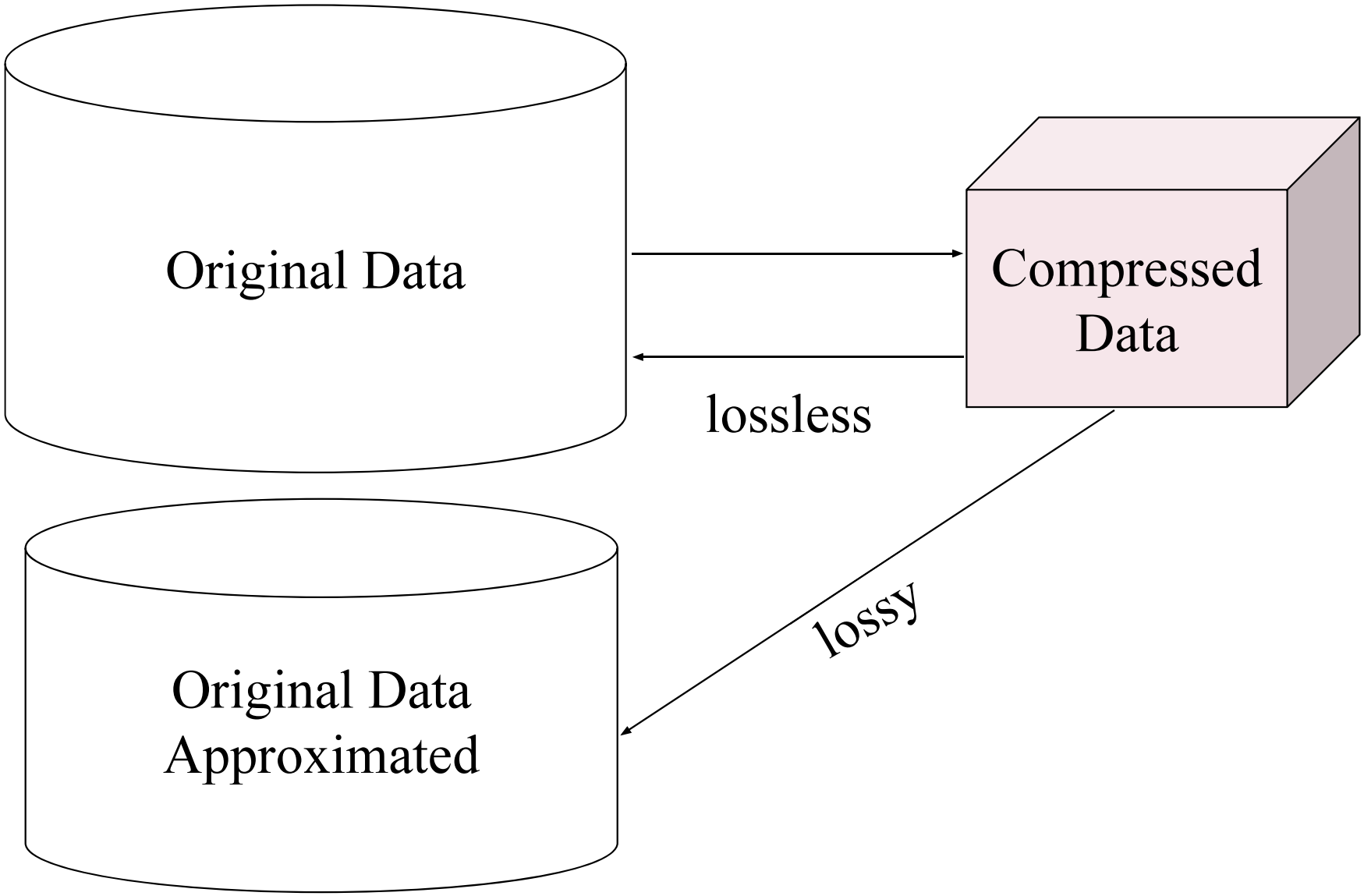
- Problem: Feature selection (i.e., attribute subset selection):
  - Select a minimum subset of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - Reduces # of attributes in the discovered patterns (which are now easier to understand)

Heuristic Feature Selection Methods

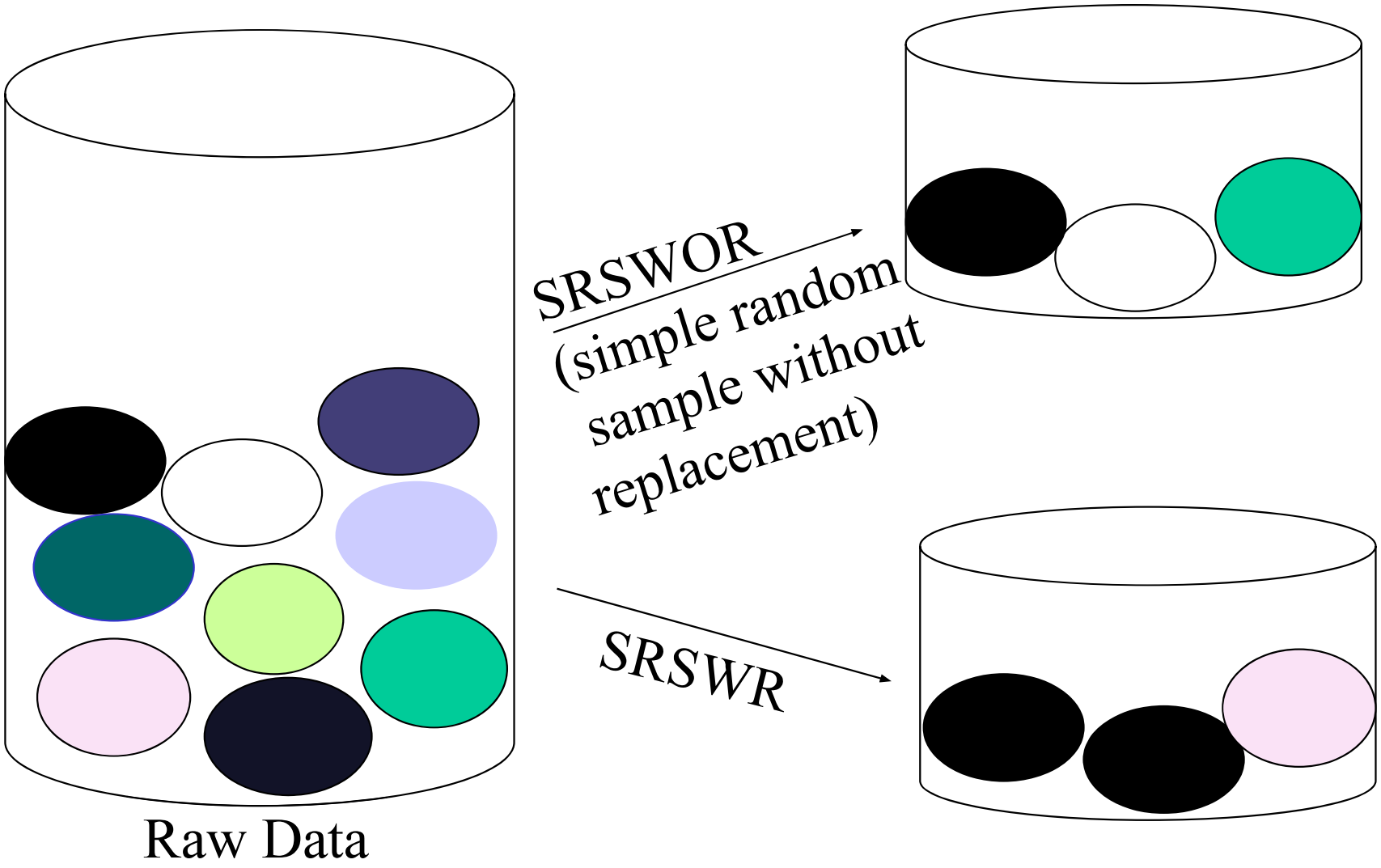
- There are 2^d possible subsets (or sub-features) of d features

- Several heuristic feature selection methods:
  - Step-wise feature selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination:
    - Repeatedly eliminate the worst feature
  - Combination

Data Compression



Sampling


SRSWOR
(simple random
sample without
replacement)
SRSWR
Raw Data