DATA ANALYTICS

Unit 1:Data Reduction

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:Data Reduction

Mamatha H R
Department of Computer Science and Engineering

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, …

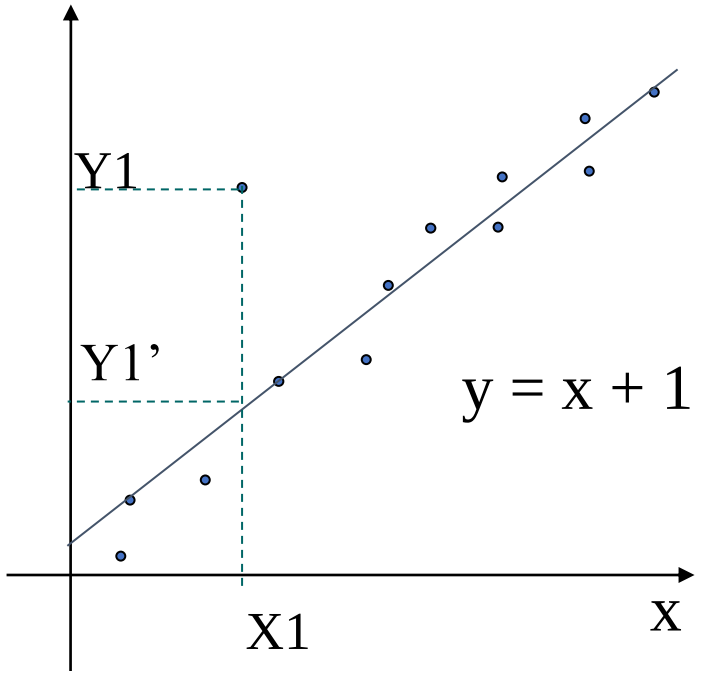
Parametric Data Reduction: Regression and Log-Linear Models

- **Linear regression**
  - Data modeled to fit a straight line
  - Often uses the least-square method to fit the line
- **Multiple regression**
  - Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- **Log-linear model**
  - Approximates discrete multidimensional probability distributions

Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (aka. ***explanatory variables*** or ***predictors***)
- 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

Regression Analysis



- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regression Analysis and Log-Linear Models

- **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 Y1, Y2, …, X1, X2, ….

- **Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$**

- Many nonlinear functions can be transformed into the above

Regression Analysis and Log-Linear Models

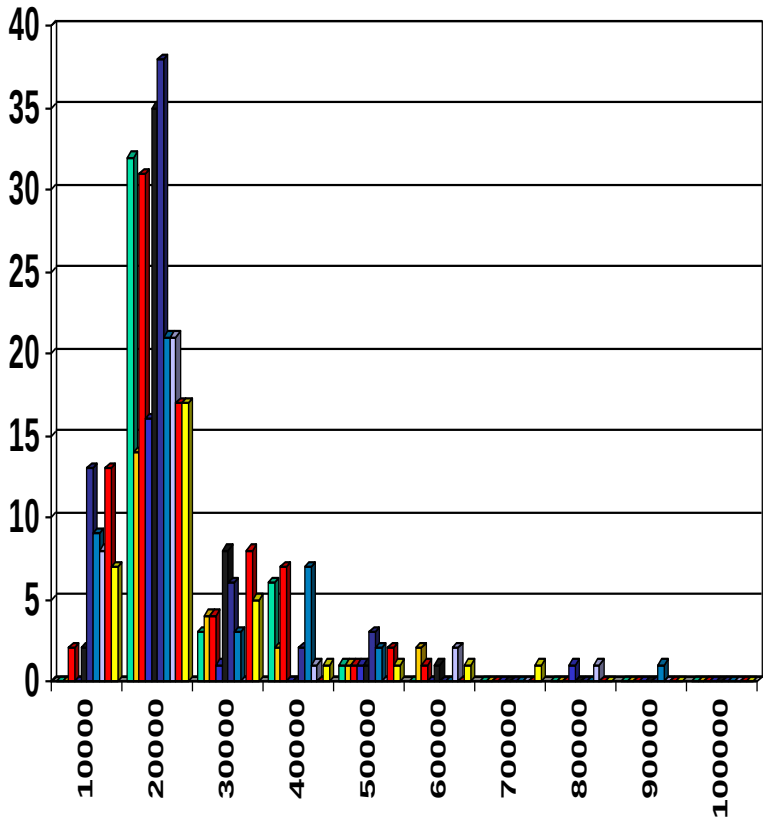
Log-linear models:

- Approximate discrete multidimensional probability distributions
- Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
- Useful for dimensionality reduction and data smoothing

DATA ANALYTICS

Histogram Analysis

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



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

The process of identifying a subset from a population of elements (aka observations or cases) is called sampling process or simply sampling

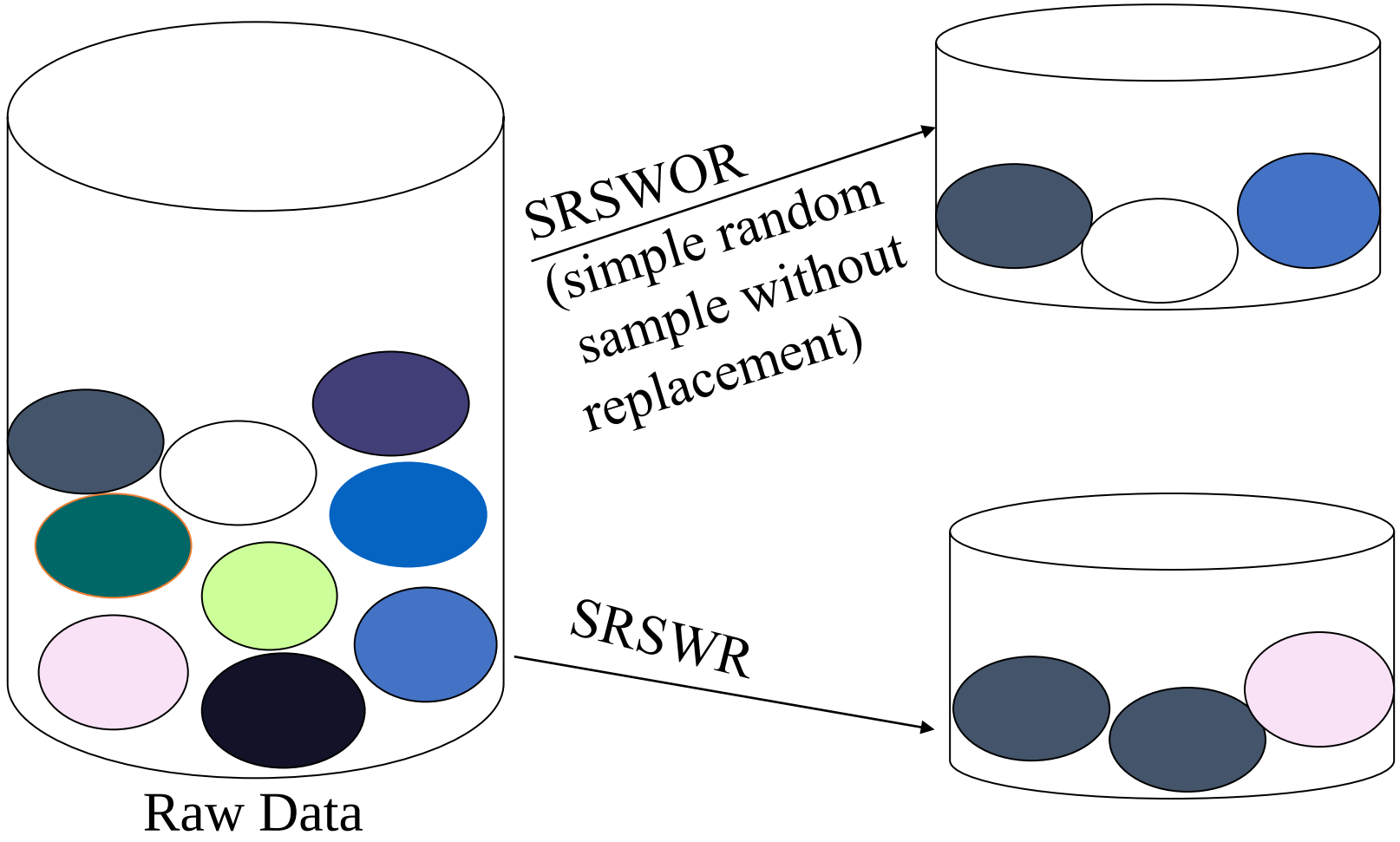
Steps used in any Sampling process:

- Identification of target population that is important for a given problem under study
- Decide the sampling frame.
- Determine the sample size
- Sampling method

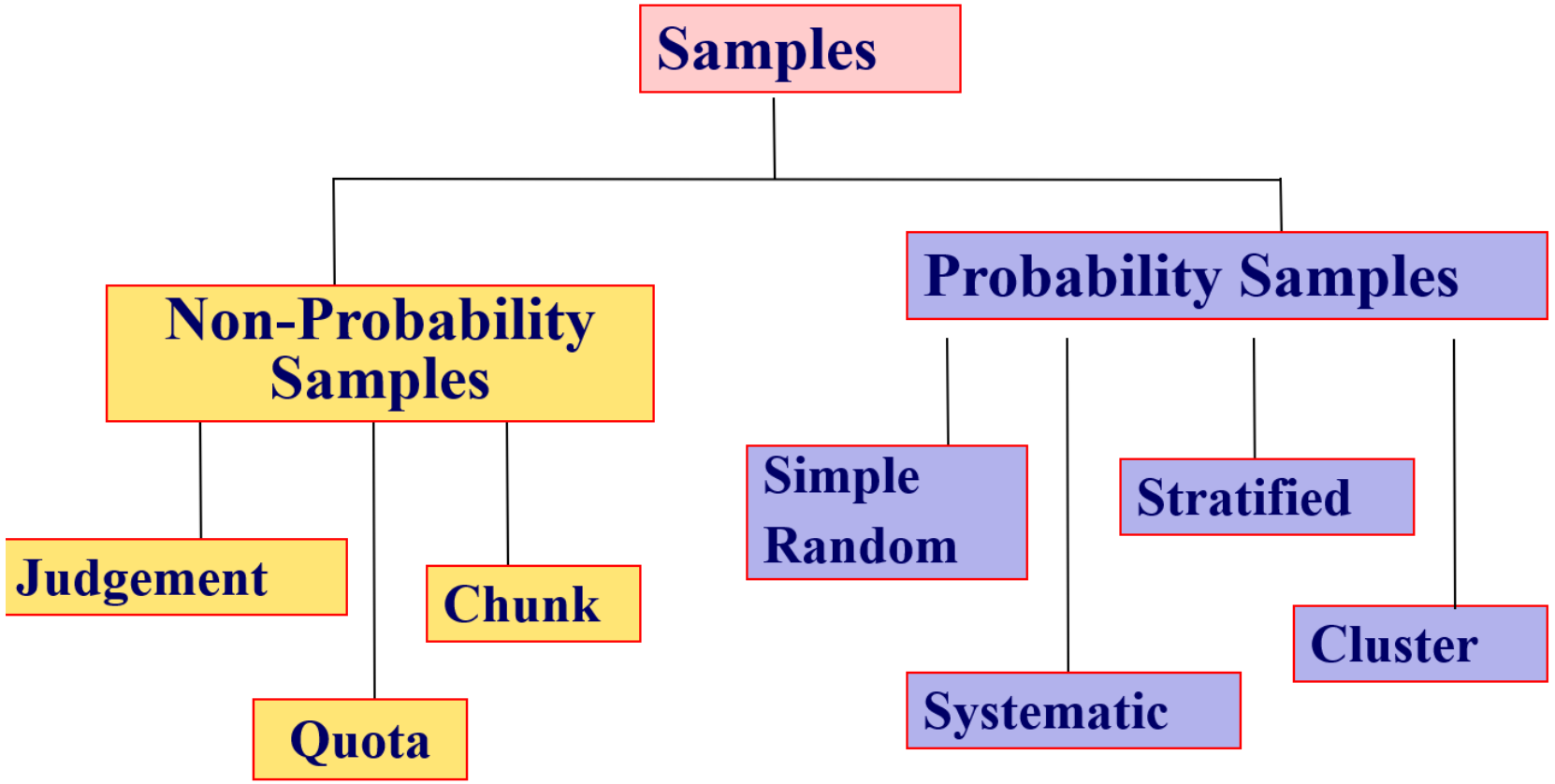
Sampling

- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow an analytics 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)

Sampling: With or without Replacement



Types of Sampling Methods



Data Cube Aggregation

- The lowest level of a data cube (base cuboid)
  - The aggregated data for an individual entity of interest
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

DATA ANALYTICS

Data Cube Aggregation

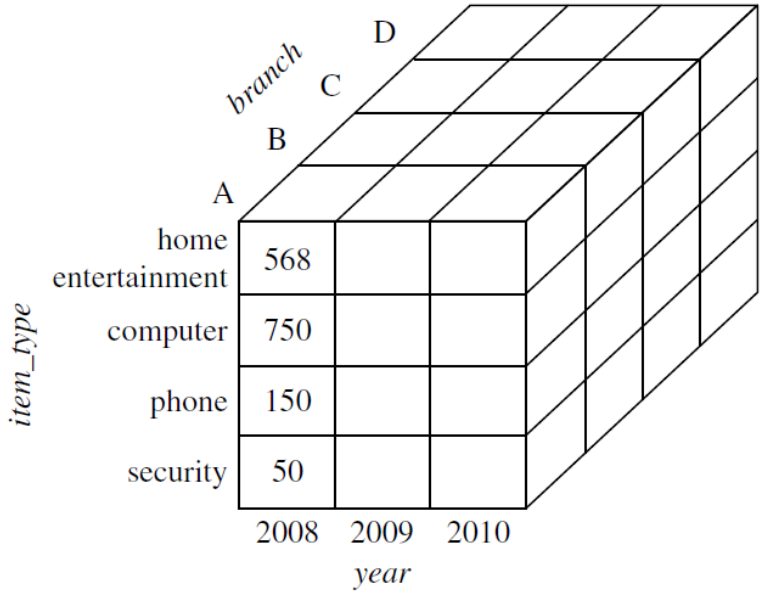
Year 2010

Year 2009

Year 2008

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

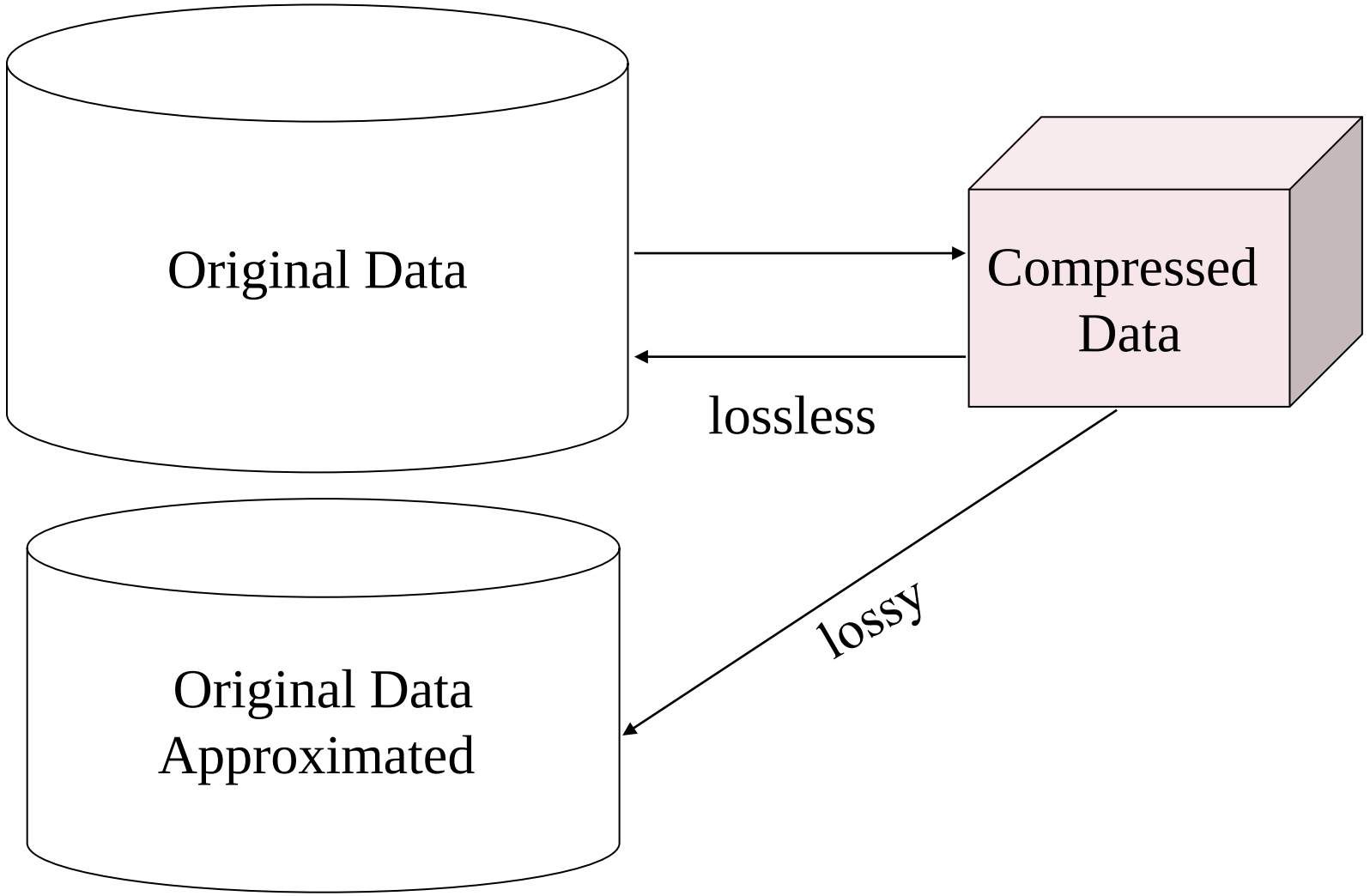
Year	Sales
2008	$1,568,000
2009	$2,356,000
2010	$3,594,000



Data Reduction 3: Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless, but only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time
- Dimensionality and numerosity reduction may also be considered as forms of data compression

Data Compression



Exercise

- Mention and explain the different parametric and non parametric methods used in data reduction.
- Compare and contrast the probability and non probability sampling methods.

References

Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834