Data Extraction, Cleanup & Transformation Tools/Data Processing Tools

Define-Data Preprocessing

- Data preprocessing is a data mining technique that involves transforming raw data into an understandable format.
- **Data pre-processing** is an important step in the data mining process.
- The product of data pre-processing is the final training set.

Why Data Preprocessing?

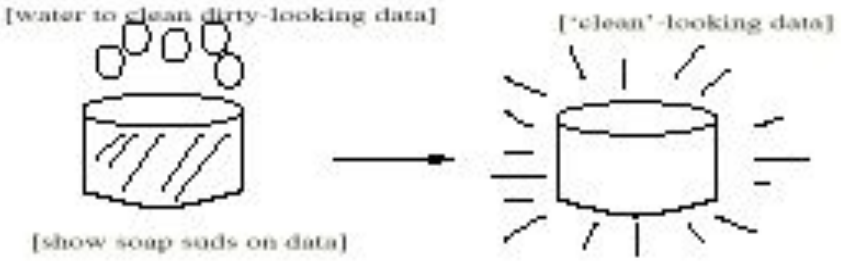
- Data in the real world is dirty.
  - noisy: containing errors or outliers.
  - Incomplete: Missing Values, Lacking attribute values.
  - Inconsistent Data
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Major Tasks in Data Preprocessing

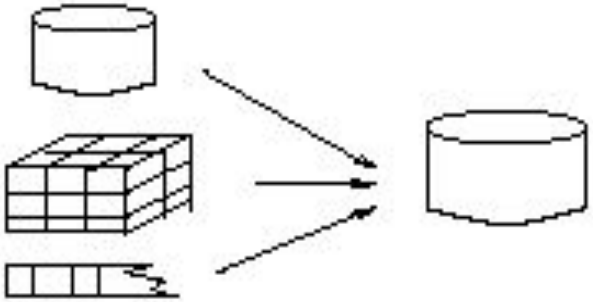
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation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

Forms of data preprocessing

Data Cleaning



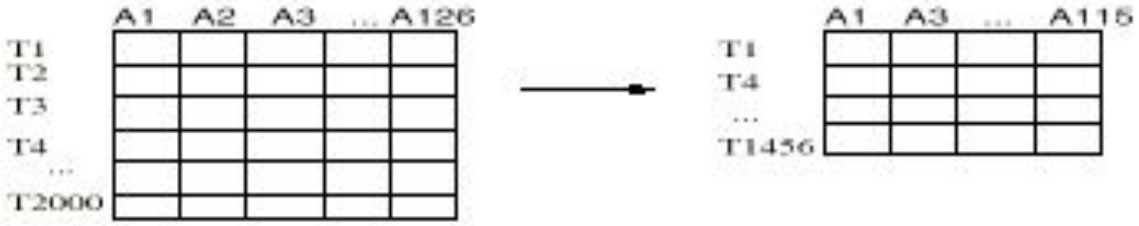
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



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 register history or changes of the data

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing
  - (Can be applicable for large data set)
- Fill in the missing value manually: tedious + infeasible for large database?
- Use a global constant to fill in the missing value
- Use the attribute mean to fill in the missing value
- Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree

Noisy Data/Outlier

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - inconsistency in naming convention
  - duplicate records
  - incomplete data
  - inconsistent data

OUTLIER

- A Data object or observations that do not comply with the general behavior or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers.
- A data object that deviates significantly from the normal objects as if it were generated by a different mechanism.
- Ex:

How to Handle Noisy Data? (Not Now)

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

Data integration and transformation

Data Integration

- Data integration:
  - combines data from multiple sources into a coherent store

Three Problems involved in data integration

- Schema integration
- Detecting and resolving data value conflicts.
- Redundant data occur often when integration of multiple databases

Data Transformation

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

Normalization: scaled to fall within a small, specified range

  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

DATA REDUCTION

Data Reduction Strategies

- Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set
- 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
  - Numerosity reduction
  - concept hierarchy generation

Data Cube Aggregation

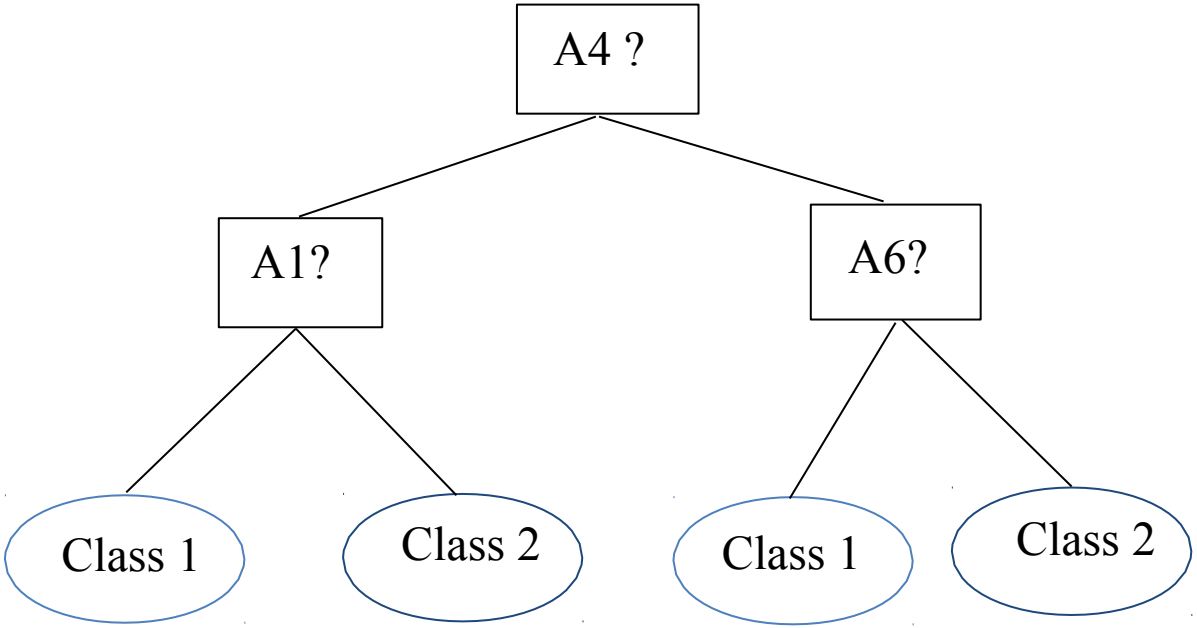
- The lowest level of a data cube
  - the aggregated data for an individual entity of interest
  - e.g., a customer in a phone calling data warehouse.
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task

Dimensionality Reduction

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set 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
  - reduce # of patterns in the patterns, easier to understand

Example of Decision Tree Induction

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



-------> Reduced attribute set: {A1, A4, A6}

Regression and Log-Linear Models

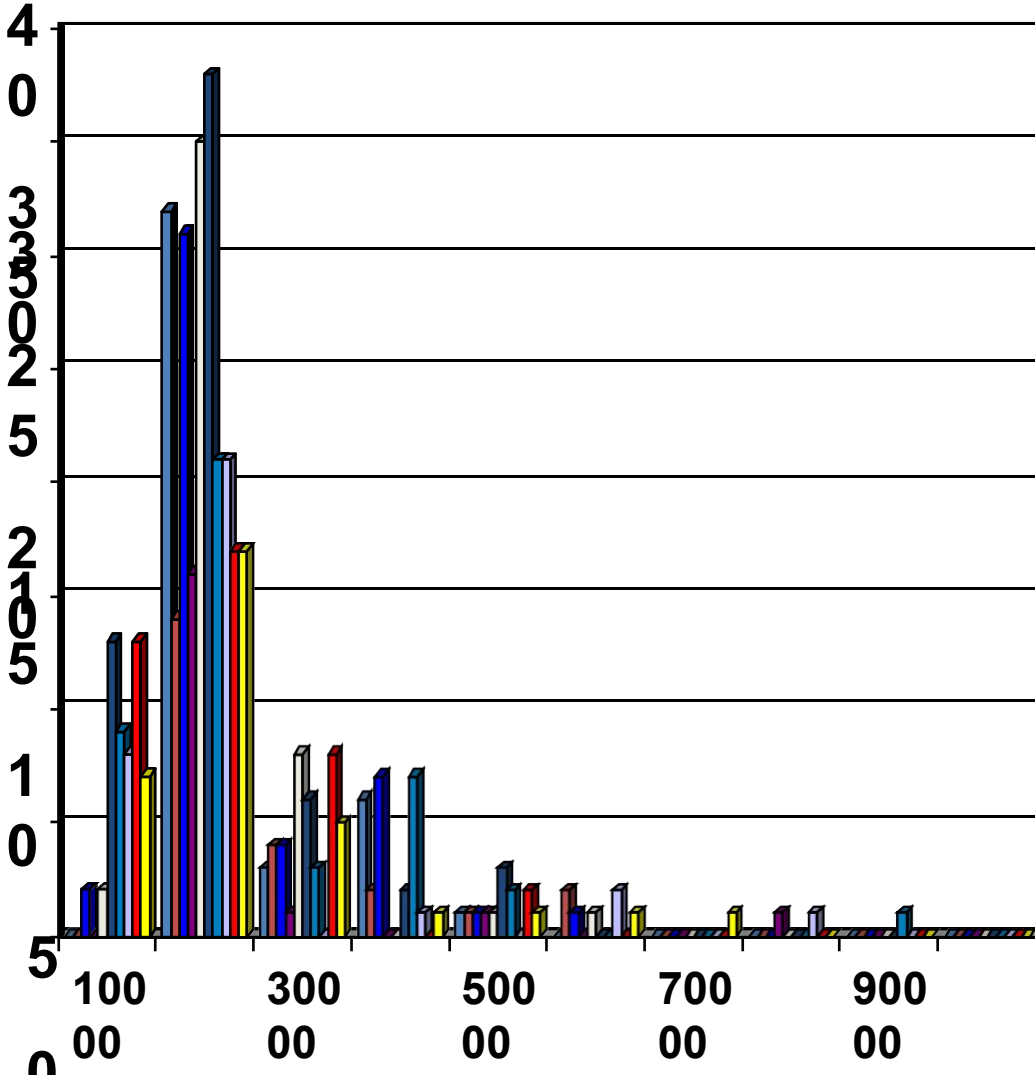
- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability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

Regress Analysis and Log- Linear Models

- Linear regression: $Y = \alpha + \beta X$
  - Two parameters , α and β 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, …, X_1, X_2, ….
- Multiple regression: $Y = b0 + b1\ X1 + b2\ X2.$
  - Many nonlinear functions can be transformed into the above.
- Log-linear models:
  - The multi-way table of joint probabilities is approximated by a product of lower-order tables.
  - Probability: $p(a, b, c, d) = \alpha_{ab}\ \beta_{ac} \chi_{ad}\ \delta_{bcd}$

Histograms

- A popular data reduction technique
- Divide data into buckets and store average (sum) for each bucket
- Can be constructed optimally in one dimension using dynamic programming
- 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 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, further detailed in Chapter 8

Samplin g

- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the
- data

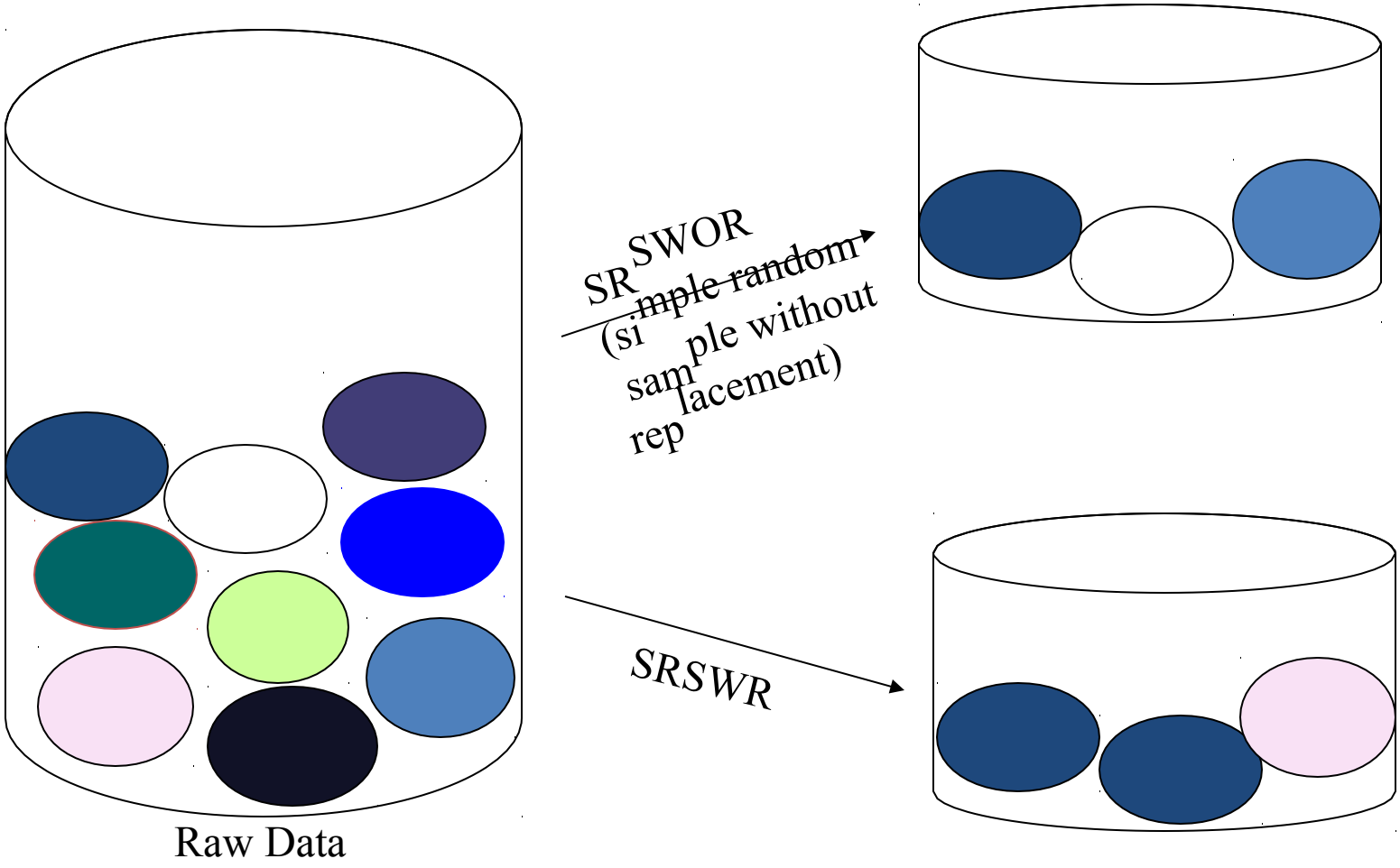
  Choose a representative subset of the data

  – Simple random sampling may have very poor
- performance in the presence of skew

  Develop adaptive sampling methods

  – Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data

Sampling


SRSWOR
(simple random
sample without
replacement)
SRSWR
Raw Data

Concept hierarchy

- Arrangement of concepts such as time , location.
  - reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

EXTRACT

- Process of extract the relevant data from the operational database before bringing into the data warehouse.
- Many commercial tools are available to help with the extraction process.
- **Data Junction** is one of the commercial products.

LOADING

- Loading often implies physical movement of the data from the computer(s) storing the source database(s) to that which will store the data warehouse database.
- This takes place immediately after the extraction. The most common channel for data movement is a high-speed communication link.
- Ex: Oracle Warehouse Builder is the API from Oracle, which provides the features to perform the ETL task on Oracle Data Warehouse

ETL tools

- A large number of commercial tools support the ETL process for data warehouses in a comprehensive way.

- COPYMANAGER (InformationBuilders)
- DATASTAGE (Informix/Ardent)
- POWERMART (Informatica)
- DECISIONBASE (CA/Platinum)
- DATATRANSFORMATIONSERVICE (Microsoft)
- METASUITE (Minerva/Carleton)
- SAGENTSOLUTIONPLATFORM (Sagent)
- WAREHOUSEADMINISTRATOR (SAS)