# Data Mining and Preprocessing

# Data mining

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining involves the use of sophisticated data analysis tools to discover previously unknown valid patterns and relationships in large data set [1].
  - Data mining tools predict future trends and behaviors, helps organizations to take proactive knowledge-driven decision [2].
  - The questions that were traditionally tedious to settle can be settled by data mining tools

# Data Mining Functionalities

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks
- Data mining tasks can be classified into two categories: descriptive and predictive
  - Descriptive mining tasks characterize the general properties of the data in the database.
  - Predictive mining tasks perform inference on the current data in order to make predictions.

# Data Mining Functionalities

- Association and correlation analysis
  - Frequent patterns can be defined as a pattern (a set of items, subsequence, substructures, etc.) that appears intermittently in data
  - Intermittent item set is a set of data that occurs frequently together in a transaction data set for
  - example, a set of items, such as table and chair.
- Classification
  - Classification is used to builds models from data with predefined classes as the model is used to classify new instance whose classification is not known

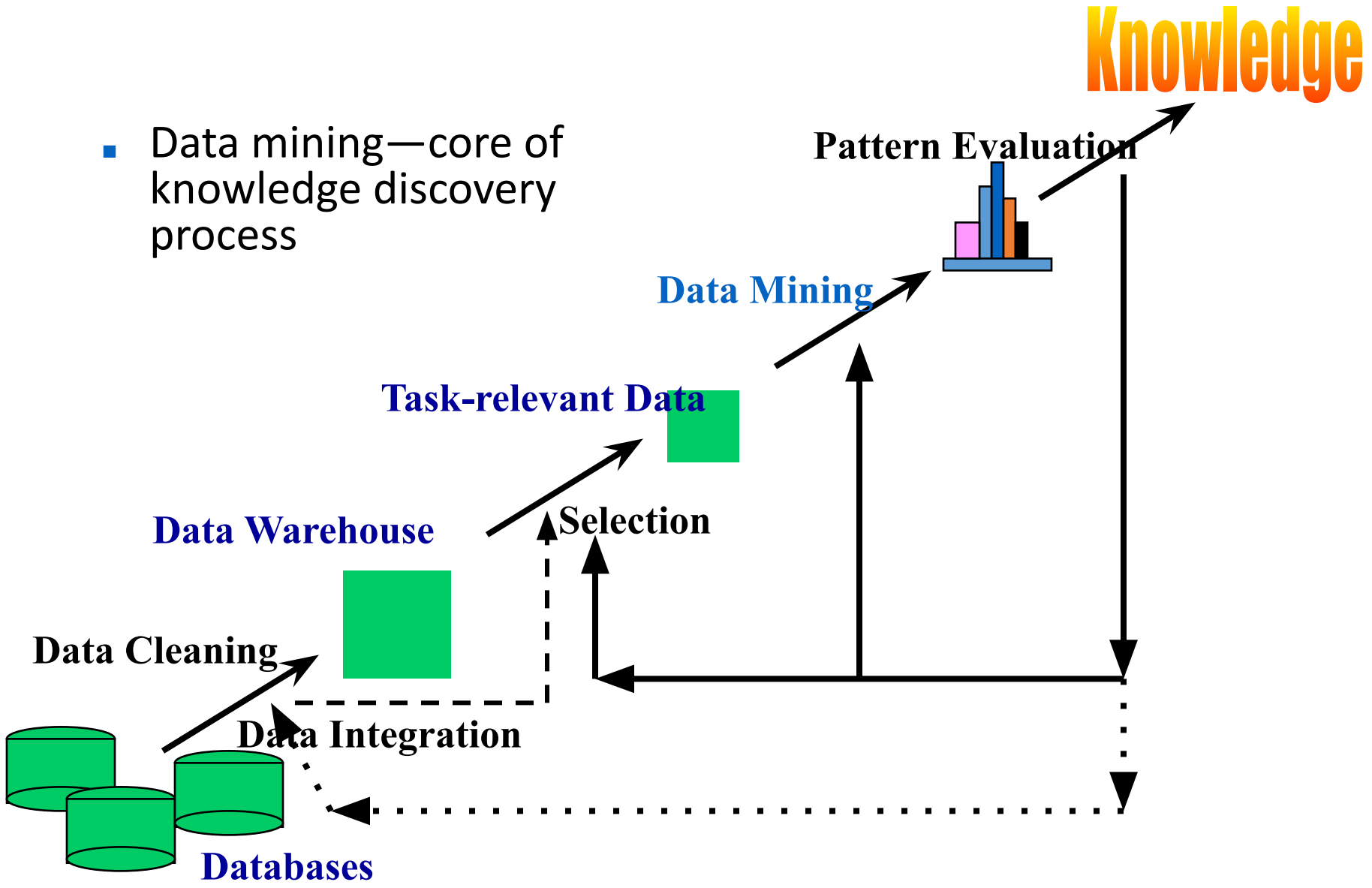
# Data Mining Functionalities

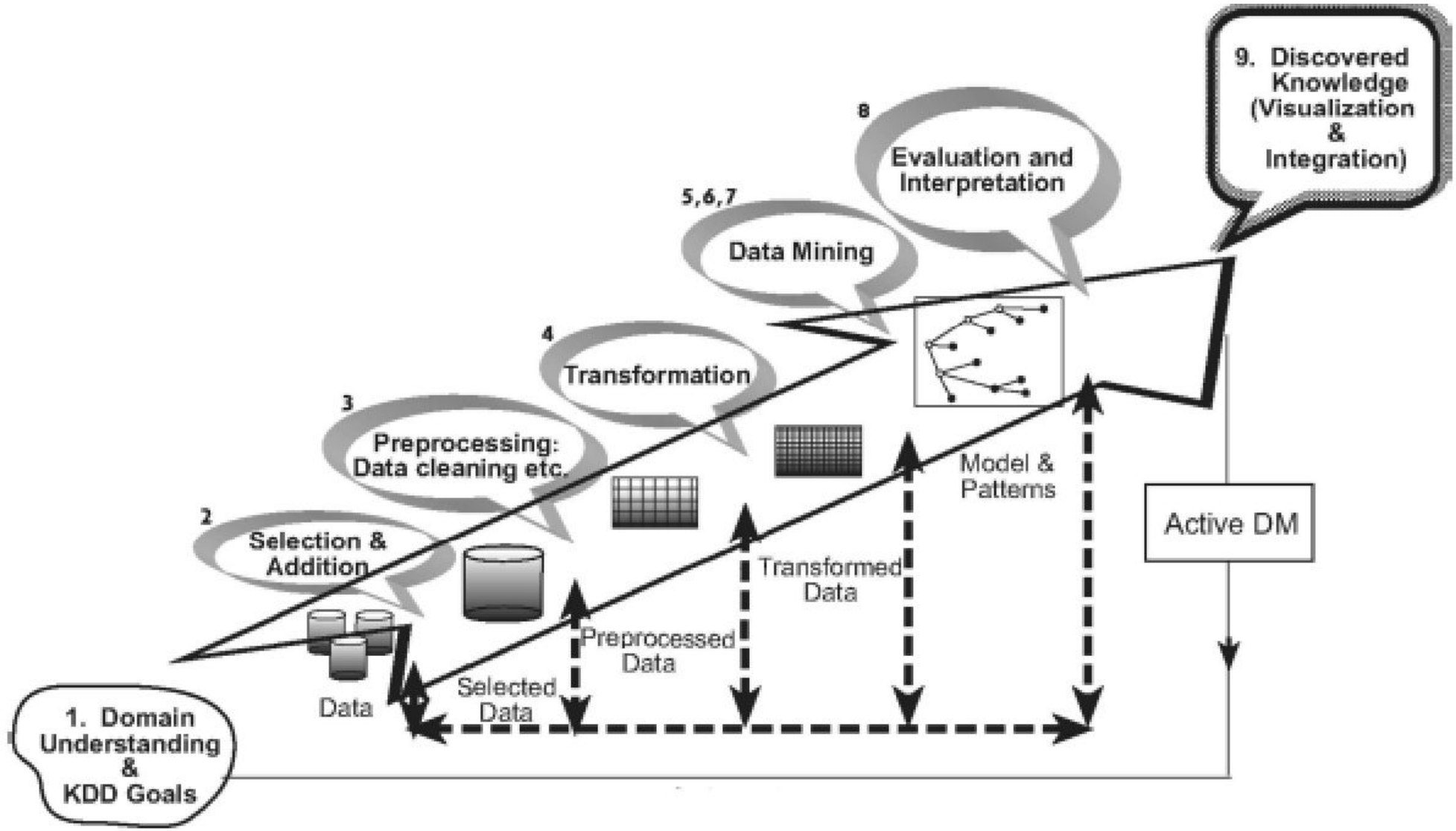
- Prediction
  - Predictive model determined the future outcome rather than present behavior
- Clustering
  - Clustering is the process of partitioning a set of object or data in a same group called a cluster
- Outlier analysis
  - Outer analysis is an object in database which is significantly different from the existing data

# KDD Process

- Knowledge discovery in databases (KDD)
- Databases (or KDD) are frequently treated data mining and knowledge discovery as synonyms, data mining is actually part of knowledge discovery process
- The process starts with determining the KDD goals, and “ends” with the implementation of the discovered knowledge.

- Data mining—core of knowledge discovery process




9. Discovered Knowledge (Visualization & Integration)
8
Evaluation and Interpretation
5,6,7
Data Mining
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Model & Patterns
Active DM
Transformed Data
Preprocessed Data
Selected Data
Data
1. Domain Understanding & KDD Goals

# KDD Process

- **Developing an understanding of the application domain:**
  - Initial preparatory step.
  - prepares the scene for understanding what should be done with the many decisions
  - Need to understand and define the goals of the end-user and the environment
  - Revision of step (if required)
  - Data Preprocessing starts after understanding the KDD goals

# KDD Process

- **Selecting and creating a data set on which discovery will be performed.**
- Having defined the goals, the data that will be used for the knowledge discovery should be determined.
  - Finding out what data is available,
  - Obtaining additional necessary data,
  - Integrating data into one data set,
  - Including the attributes that will be considered for the process.

# KDD Process

- **Preprocessing and cleansing**
- Data reliability is enhanced
- Data Cleaning
  - Removal of noise or outliers
- Involve complex statistical methods or data mining algorithms are involved.
  - *example, if one suspects that a certain attribute is of insufficient reliability or has many missing data, then this attribute could become the goal of a data mining supervised algorithm. A prediction model for this attribute will be developed, and then missing data can be predicted.*

# KDD Process

- **Data transformation**
- generation of better data for the data mining
  - include dimension reduction
  - attribute transformation
- Crucial step for entire KDD process
- Project-specific.
  - **NOTE: However, even if we do not use the right transformation at the beginning, we may obtain a surprising effect that hints to us about the transformation needed (in the next iteration).**

# KDD Process

- **Choosing the appropriate Data Mining task**
- Which Data Mining to use?
  - classification, regression, or clustering?
  - Depends on the KDD goals and previous steps.
- There are two major goals in Data Mining:
  - Prediction (supervised)
  - Description (unsupervised)
- Inductive learning: model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples
- Strategy also takes into account the level of meta-learning for the particular set of available data

# KDD Process

- **Choosing the Data Mining algorithm**
- Selecting the specific method to be used for searching patterns
  - *Example: in considering precision versus understandability, the former is better with neural networks, while the latter is better with decision trees.*
- Meta-learning
- Attempts to understand the conditions under which a Data Mining algorithm is most appropriate

# KDD Process

- **Employing the Data Mining algorithm.**
  - Implementation of the Data Mining algorithm
  - This step we might need to employ the algorithm several times until a satisfied result is obtained
  - *Example: tuning the algorithm's control parameters, such as the minimum number of instances in a single leaf of a decision tree.*

# KDD Process

- **Evaluation**
  - evaluate and interpret the mined patterns w.r.t the goals defined in the first step.
  - Re-consider the preprocessing steps with respect to their effect on the Data Mining algorithm results
  - focuses on the comprehensibility and usefulness of the induced model
  - discovered knowledge is also documented for further usage

# KDD Process

- **Using the discovered knowledge**
- Incorporating the knowledge gained into the system
  - changes to the system and measure the effects.
- Challenges
  - Change in data structure
  - Change in domain
- Success of this step determines the effectiveness of the entire KDD process

# Data Cleaning

- Missing Values
  - lacking attribute values
- Noisy data
  - containing errors or outliers
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data Transformation
  - Normalization and aggregation

# Data Reduction

- 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
  - Dimensionality reduction
  - Discretization and concept hierarchy generation

# Data Cube Aggregation

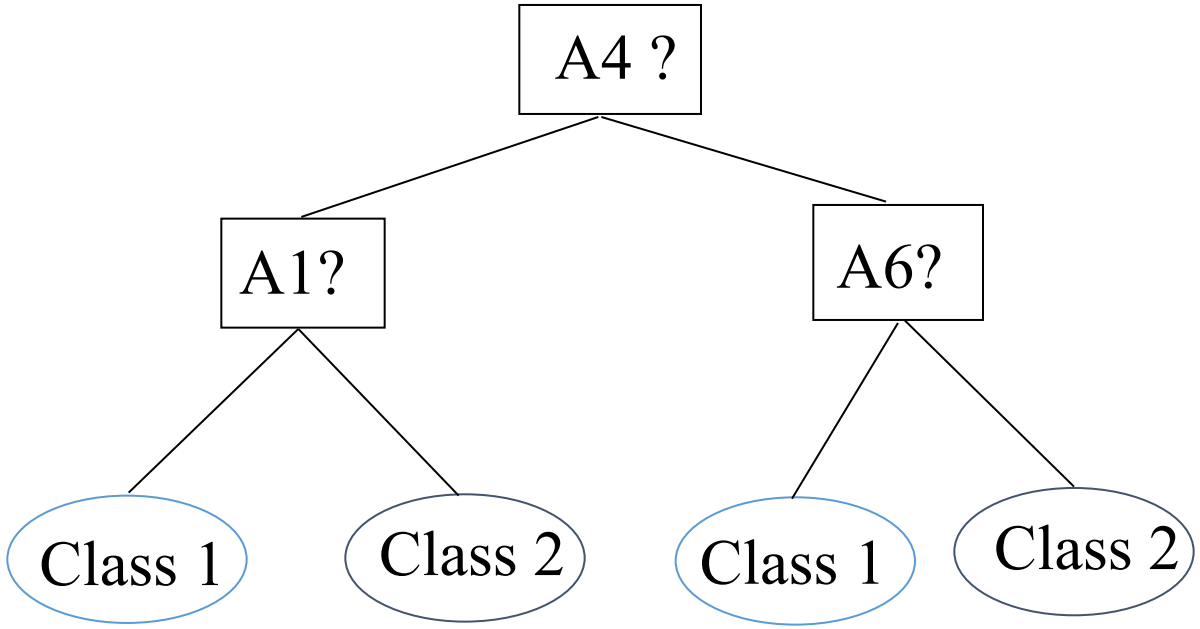
- The lowest level of a data cube
  - the aggregated data for an individual entity of interest
  - e.g., a customer in a phone calling data warehouse.
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation 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
  - there are often too many factors on the basis of which the final classification is done.
  - These factors are basically variables called features.
  - The higher the number of features, the harder it gets to visualize the training set and then work on it.
  - Sometimes, most of these features are correlated, and hence redundant.
  - This is where dimensionality reduction algorithms come into play.

# Example of Decision Tree Induction

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




Reduced attribute set: {A1, A4, A6}

# Heuristic Feature Selection Methods

- There are $2^d$ possible sub-features of $d$ features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests.
  - Best 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
  - Best combined feature selection and elimination:
  - Optimal branch and bound:
    - Use feature elimination and backtracking

# Regression and Log-Linear Models

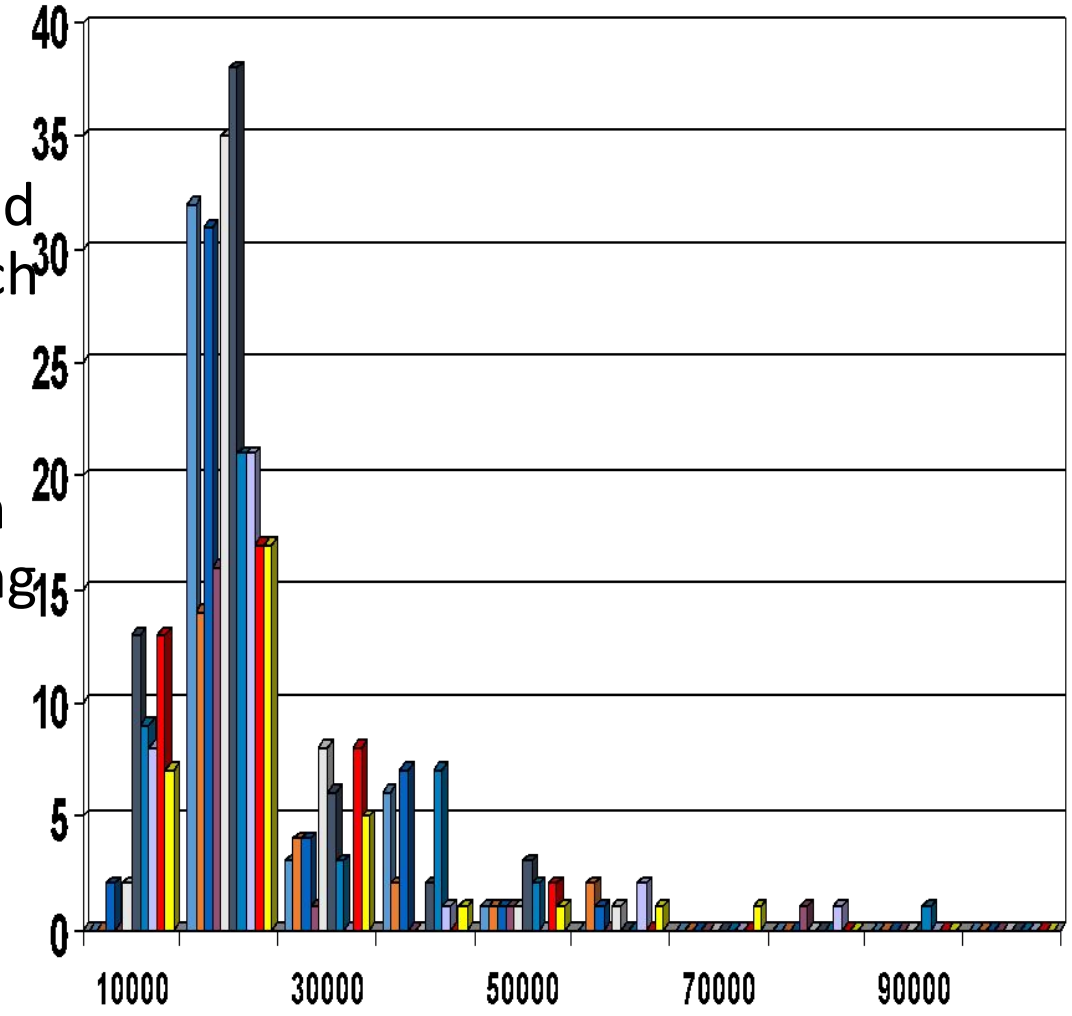
- Linear regression: Data are modeled to fit a straight line
    - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability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



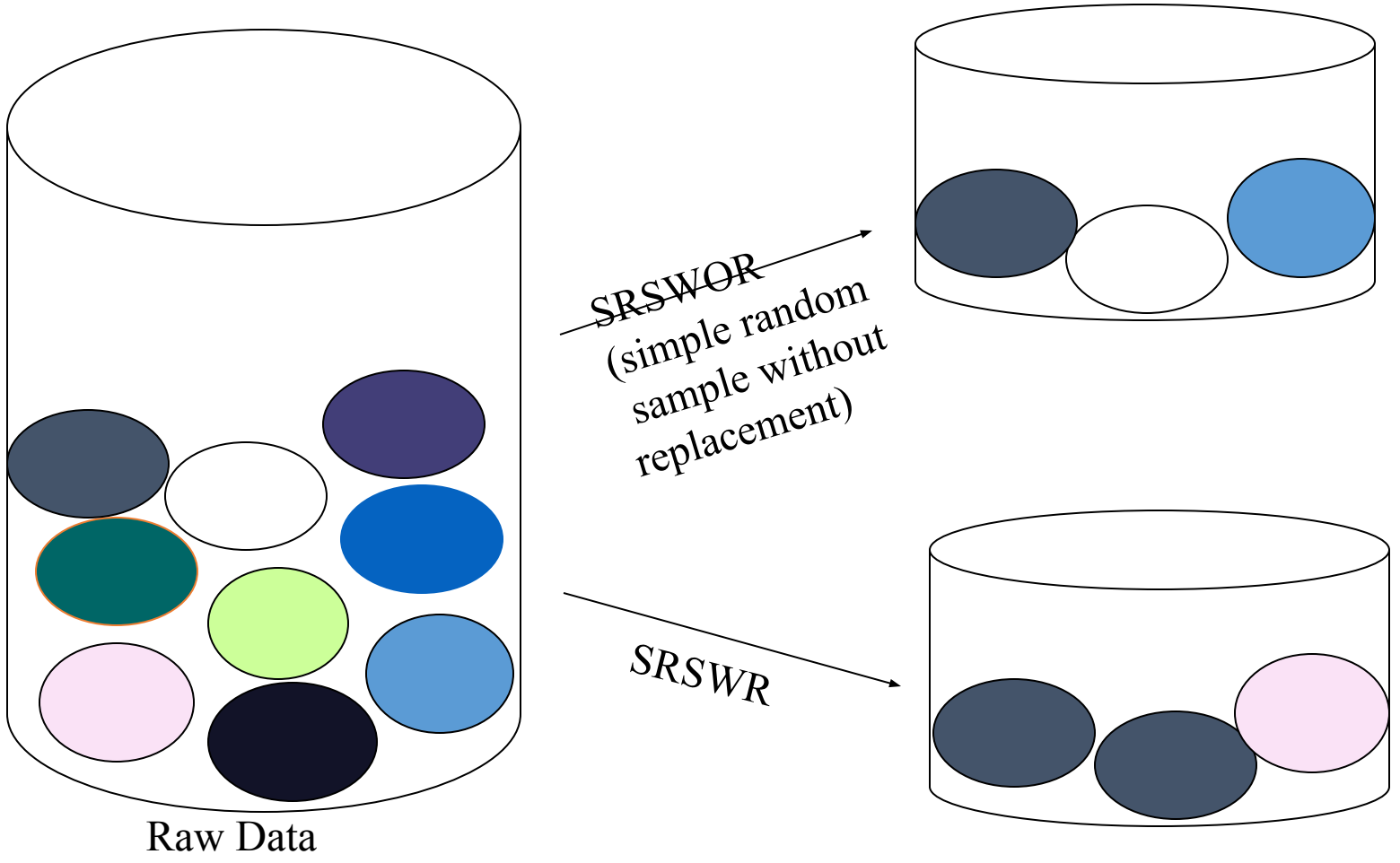
# Clustering

- 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

# Sampling

- 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

# Discretization

- Three types of attributes:
  - Nominal — values from an unordered set
  - Ordinal — values from an ordered set
  - Continuous — real numbers
- Discretization:
  - divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

# Discretization and Concept hierachy

- Discretization
  - reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.
- Concept hierarchies
  - 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).