Data Pre-processing

Major Tasks in Data Pre-processing

- Major tasks in data pre-processing are data cleaning, data integration, data transformation, data reduction and data discretization.
- **Data cleaning** - Data Cleaning includes, filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
- **Data integration** - Data Integration includes integration of multiple databases, data cubes, or files.
- **Data transformation** - Data Transformation includes normalization and aggregation.
- **Data reduc**tion- Data reduction is achieved by obtaining reduced representation of data in volume but produces the same or similar analytical results.
- **Data Discretization** - Data Discretization is part of data reduction but with particular importance, especially for numerical data.

Data Cleaning

Real world data are incomplete, noisy, and inconsistent.

The data cleaning algorithm attempt to fill in missing values, smooth out noise data before applying data mining.

- Missing Data (Values)

Eg. Many tuples, rows or data has no value.

Roll Number	Name	Age	Class Label
101	Rahul	21	H
102	Karan	24	L
103	Abhishek	21	H
104	Mohan	23	L
105	Rohan	23.5	L

Methods of handling missing values

a). Ignore the tuple:

1) When the attribute with missing values does not contribute to any of the classes or has missing class label.

2) Effective only when more number of missing values are there for many attributes in the tuple.

3) Not effective when only few of the attribute values are missing in a tuple.

b). Fill in the missing value manually:

1) This method is time consuming

2) It is not efficient

3) The method is not feasible

c). Use of a Global constant to fill in the missing value:

1) This means filling with “Unknown” or “Infinity”

2) This method is simple

3) This is not recommended generally

d) Use the attribute mean to fill in the missing value:

That is, take the average of all existing income values and fill in the missing income value.

e) Use the attribute mean of all samples belonging to the same class as that of the given tuple:

Say, there is a class “Average income” and the tuple with the missing value belongs to this class and then the missing value is the mean of all the values in this class.

f) Use the most probable value to fill in the missing value:

This method uses inference based tools like Bayesian Formula, Decision tree etc.

- **Noisy Data**

Apply data smoothing techniques like the ones given below:

a) Binning Methods

Simple Discretization Methods: Binning

- **Equal-width (distance) partitioning** :

1. Divides the range into N intervals of equal size: uniform grid
2. if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
3. The most straightforward, but outliers may dominate presentation
4. Skewed data is not handled well

- **Equal-depth (frequency) partitioning:**

1. Divides the range into N intervals, each containing approximately same number of samples
2. Good data scaling
3. Managing categorical attributes can be tricky

- **Binning Method**

Smooth a sorted data value by consulting its "neighbourhood", i.e., the values around its (hence it is called local smoothing). The sorted values are distributed into a number of bins (or bucket).

Binning

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Binning

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by **bin median**:
  - Bin 1: 8.5, 8.5, 8.5, 8.5
  - Bin 2: 22.5, 22.5, 22.5, 22.5
  - Bin 3: 28.5, 28.5, 28.5, 28.5

Binning

- Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 28, 34
* Smoothing by **bin means**:
  - Bin 1: 9, 9, 9
  - Bin 2: 22, 22, 22
  - Bin 3: 29, 29, 29
* Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 25, 34

Binning

- Sorted data for price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 28, 34
* Smoothing by **bin median**:
  - Bin 1: 8, 8, 8
  - Bin 2: 21, 21, 21
  - Bin 3: 28, 28, 28

Noisy Data

- Apply data smoothing techniques like the ones given below:

a) Binning Methods

b) Regression

c) Clustering

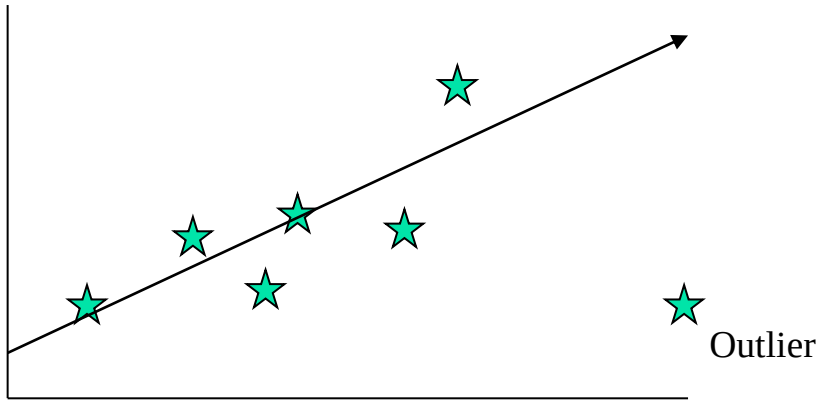
Regression

- Method of mapping the data into an linear mathematical equation and converging them to a single line and finding the outlying data
- Data mining technique which is used to fit an equation to a dataset

  Linear Regression: $y = b + mx$

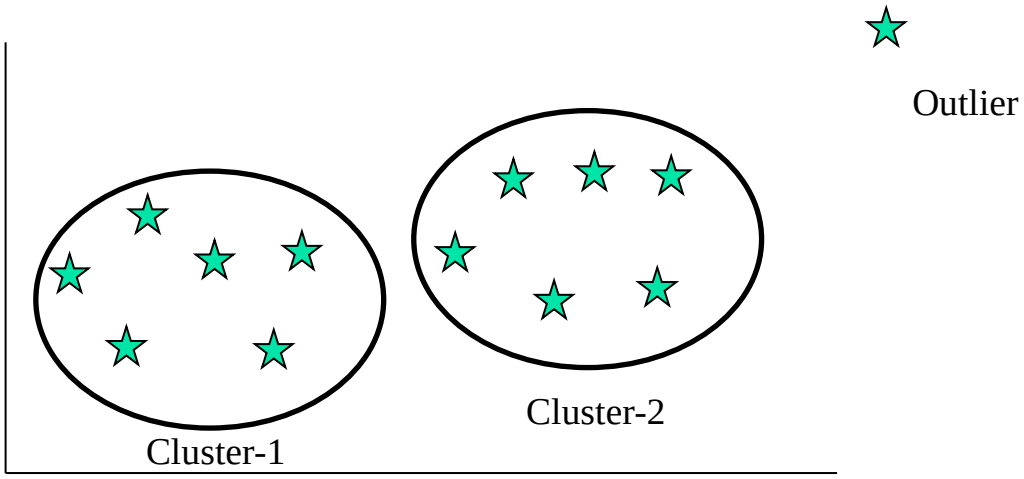
Where y is a predicted value, a is given value, m is slope, and b is a intercept.

- It is also known as data fitting function



Clustering

- Forming clusters and identifying the outlying data



Data Cleaning as a Process

- Handling missing data & noisy data is not all about data cleaning, but it is a big process.
- 1) First step in data cleaning is Discrepancy Detection

  a) Using Meta Data

  - Data about Data
  - Attribute Value Domain
  - Acceptable value for an attribute
  - Range of Attribute Values
  - Dependency between attributes
- To detect discrepancy use code or tool

Data Cleaning as a Process

b) Using Field Overloading

- Eg. 20004 entered instead of 2004

c) Using Unique Rule

d) Using Consecutive Rule

e) Using Null Rule

- How to fill in field values that have blanks, ?, special characters...

- Commercial tools to aid in discrepancy detection:

(i) Data Scrubbing Tool

- Uses simple domain knowledge to detect & correct errors

(ii) Data Auditing Tool

- Analyzes data to see if it satisfies rules & relationships
- Uses Statistical analysis to find correlations
- Uses clustering to find outliers 2

Data Cleaning as a Process

- 2) Second Step in Data Cleaning is Data Transformation
  - Method of correcting the identified data discrepancies.
- Commercial tools to aid in Data Transformation:

  (i) Data Migration Tool:
  - Simple Transformations
  - Eg. Replace “Gender” by “Sex”

  (ii) Extraction Transformation and Loading Tools (ETL):
  - Users specify transformations through GUI
- These two steps of Data Cleaning iterates until the user is satisfied.
- This process is error prone and time consuming
- Potter’s Wheel – Publicly available Data Cleaning Tool.

Data Integration

- Data integration:
  - combines data from multiple sources into a coherent store
- Schema integration
  - integrate metadata from different sources
  - Entity identification problem: identify real world entities from multiple data sources, e.g., A.cust-id ≡ B.cust-#
- Detecting and resolving data value conflicts
  - for the same real world entity, attribute values from different sources are different
  - Possible reasons can be - Different interpretations, different representation and different scaling

    - Eg. Sales amount represented in Dollars (USD) in one data source and as Pounds ($) in another data source.

Handling Redundant Data in Data Integration

- Redundant data occur often when integrating multiple DBs
  - The same attribute may have different names in different databases
  - One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant data may be able to be detected by correlational analysis

$$r_{A,B} = \frac{\Sigma(A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B}$$

- Careful integration can help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- It is a data preprocessing technique that transforms or consolidate the data into alternate forms appropriate for the data mining
- Involved Process:

1. Smoothing: Removing the noise from data
2. Aggregation: Summary or aggregate function - Constructing a data cube
3. Generalization: low level concepts are replaced with higher level
4. Normalization: Attribute values are normalized by scaling their value so that the fall in specified range

Normalization

- A) Min-Max Normalization:

$$v' = \frac{v - Min_x}{Max_x - Min_x}$$

Where, v' is a new value, v is a original value, Min_x is a minimum value of attribute (x), Max_x is a maximum value of attribute (x).

- B) Z-score Normalization: Zero mean normalization

$$v' = \frac{v - \bar{x}}{\sigma}$$

Where, v' is a new value, v is a original value, $\bar{x}$ is a mean of attribute (x), and σ is a standard deviation of attribute (x)

Data Reduction

- Problem:

  Data Warehouse 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

Data Cube Aggregation

- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation capable to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

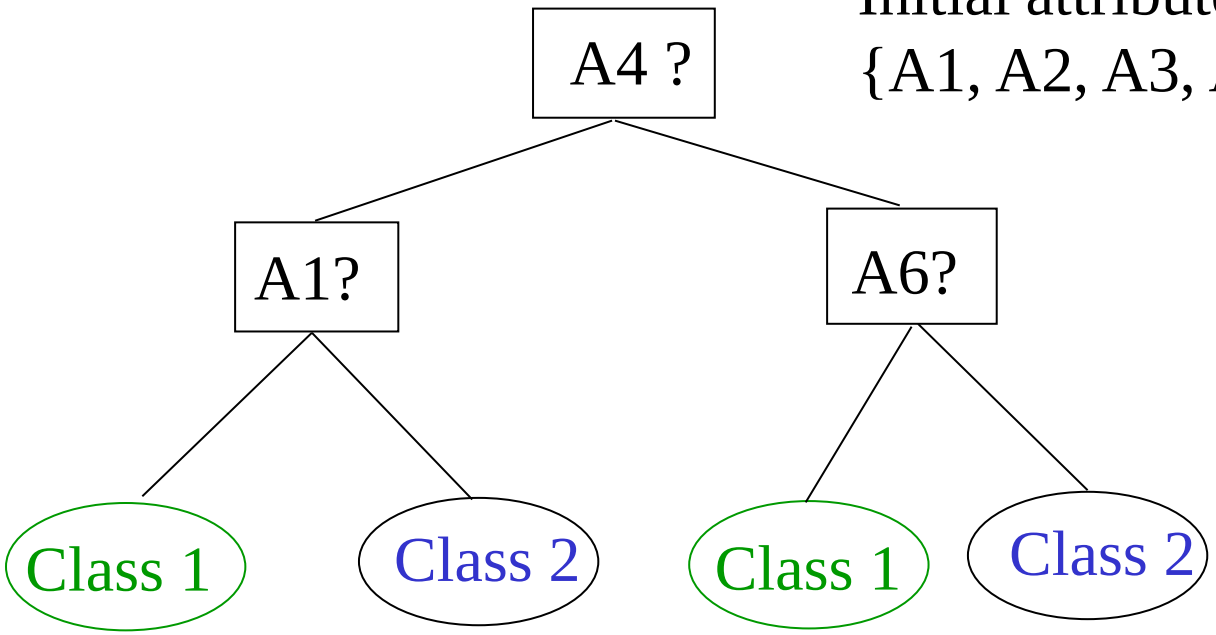
Dimensionality Reduction

- Problem: 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
  - Nice side-effect: reduces # of attributes in the discovered patterns (which are now easier to understand)
- Solution: Heuristic methods (due to exponential # of choices) usually greedy:
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction

Example of Decision Tree Induction

nonleaf nodes: tests
branches: outcomes of tests
leaf nodes: class prediction

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

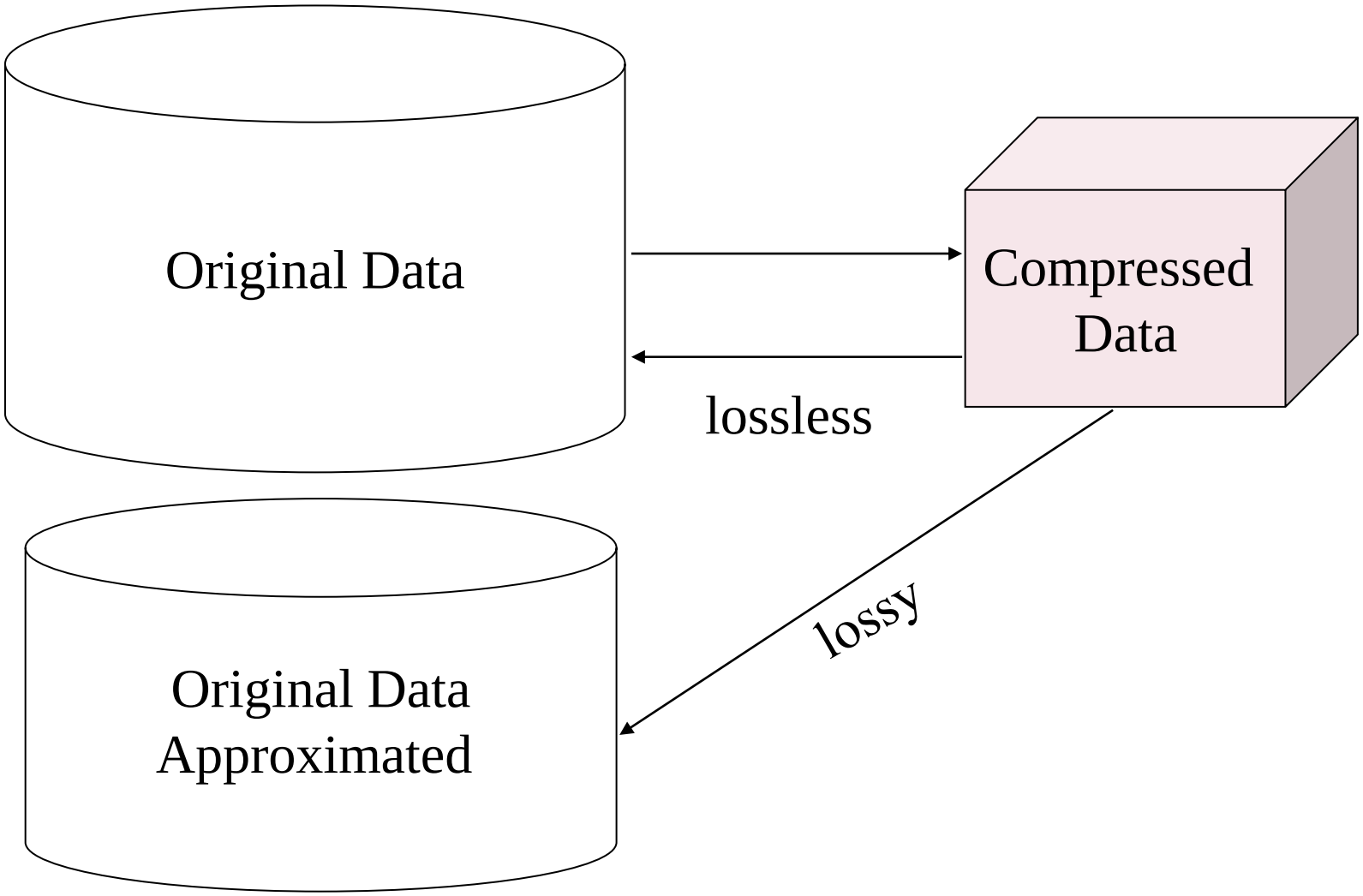


-------> Reduced attribute set: {A1, A4, A6}

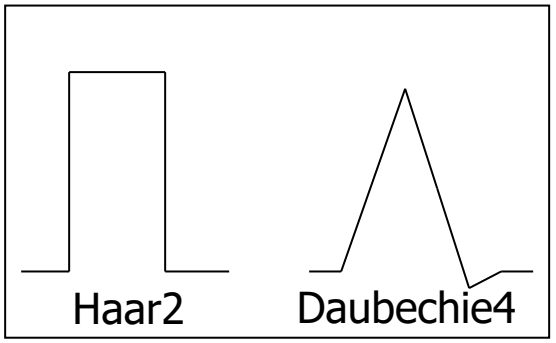
Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video, image 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

Data Compression



Wavelet Transforms



- Discrete wavelet transform (DWT):

  linear signal processing
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space (conserves local details)
- Method (hierarchical pyramid algorithm):
  - Length, L, must be an integer power of 2 (padding with 0s, when necessary)
  - Each transform has 2 functions:
    - smoothing (e.g., sum, weighted avg.), weighted difference
  - Applies to pairs of data, resulting in two sets of data of length L/2
  - Applies the two functions recursively, until reaches the desired length

Discrete Wavelet Transform (DWT)

- Wavelet transform is a widely used mathematical approach which allows to extract patterns information hidden in the original time series.
- By detecting time series in time frequency domain, the wavelet transform has the ability to analyse local characteristic and non-stationary of the data.
- The DWT method is a kind of wavelet transform and it is utilized to decompose the original series into several subseries with high or low frequency, which are relatively stationary compared to the original time series.

DWT Cont..

- In decomposition process, the original time series is decomposed by the high resolution filter and the low resolution filter. The corresponding decomposed subseries are detail component (high frequency) and approximation component (low frequency), and they are relatively stationary compared to the original time series.
- Then, the low frequency component is decomposed by two filters again while the high frequency component is preserved. Similar procedures are conducted until reaching pre-set decomposition levels.
- In reconstruction process, the ultimately obtained high and low frequency components are combined to produce reconstruction time series.
- Let J denote the decomposition levels, Di (i = 1, 2, · · · , J) denote the detail components and Ai (i = 1, 2, · · · , J) denote the approximation components.
- The reconstruction time series can be expressed in the equation

$$RTS = D_1 + A_1 = D_1 + D_2 + A_2 = D_1 + D_2 + D_3 + A_3$$

DWT Cont..

- Figure displays the specific steps of three-level wavelet decomposition

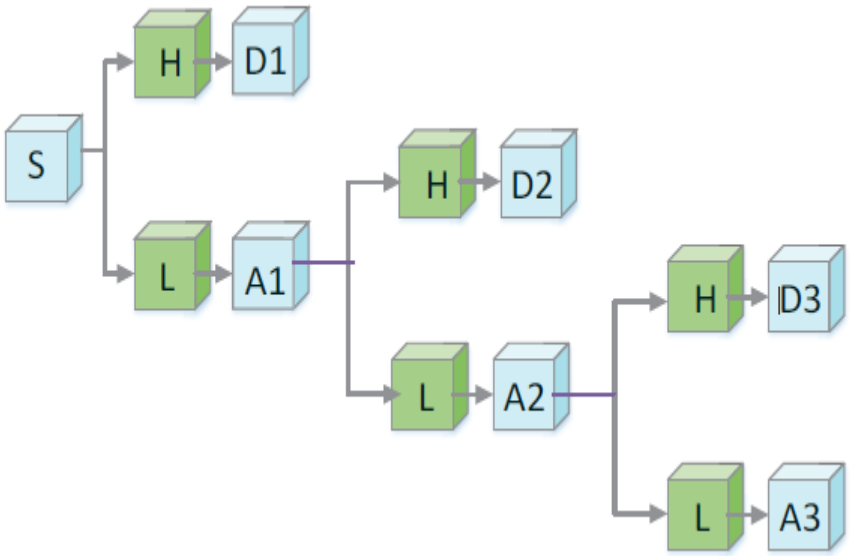
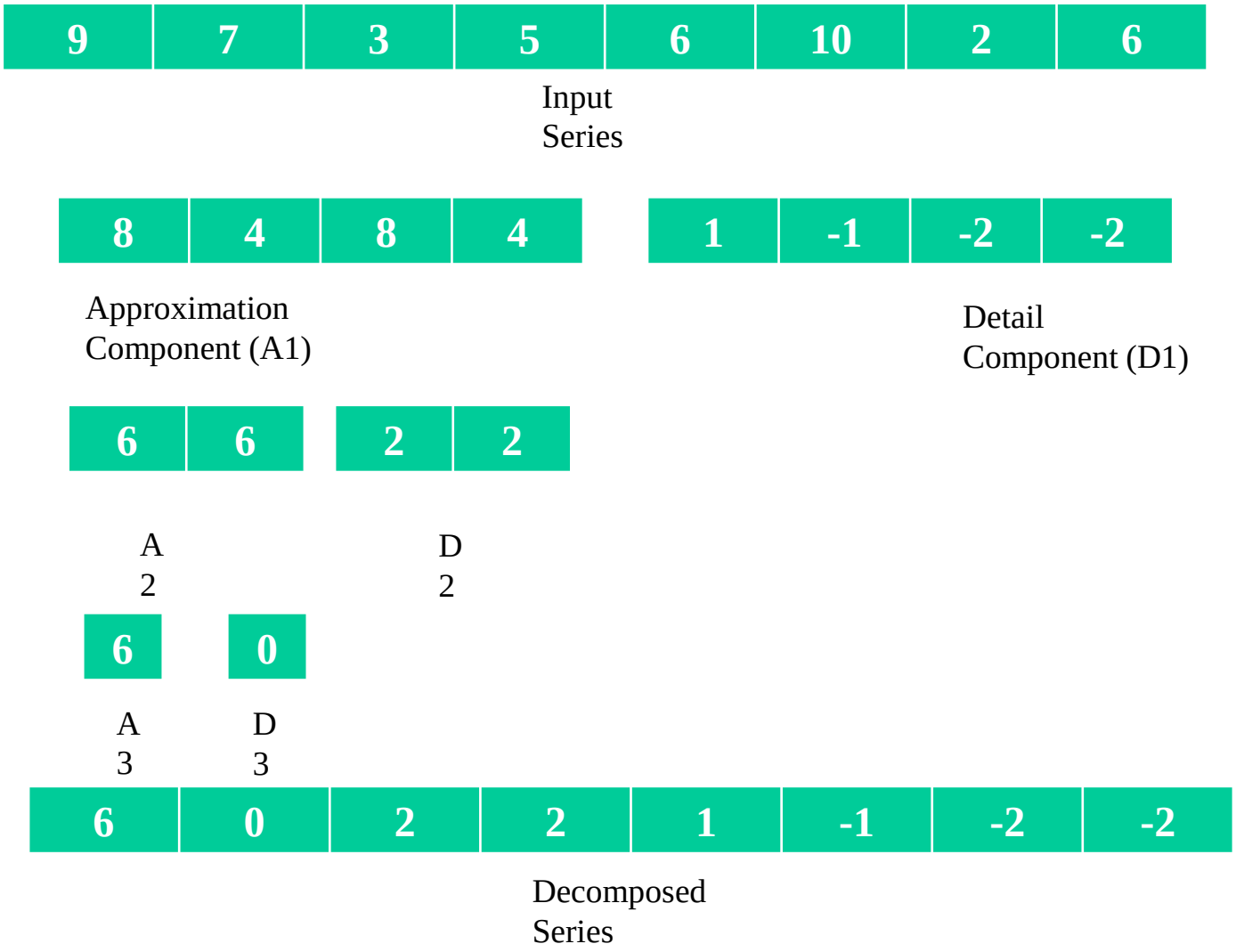


Fig. Illustration of three-level wavelet decomposition

Wavelet Transformation



Principal Component Analysis (PCA) Karhunen-Loeve (K-L) method

- Given *N* data vectors from *k*-dimensions, find

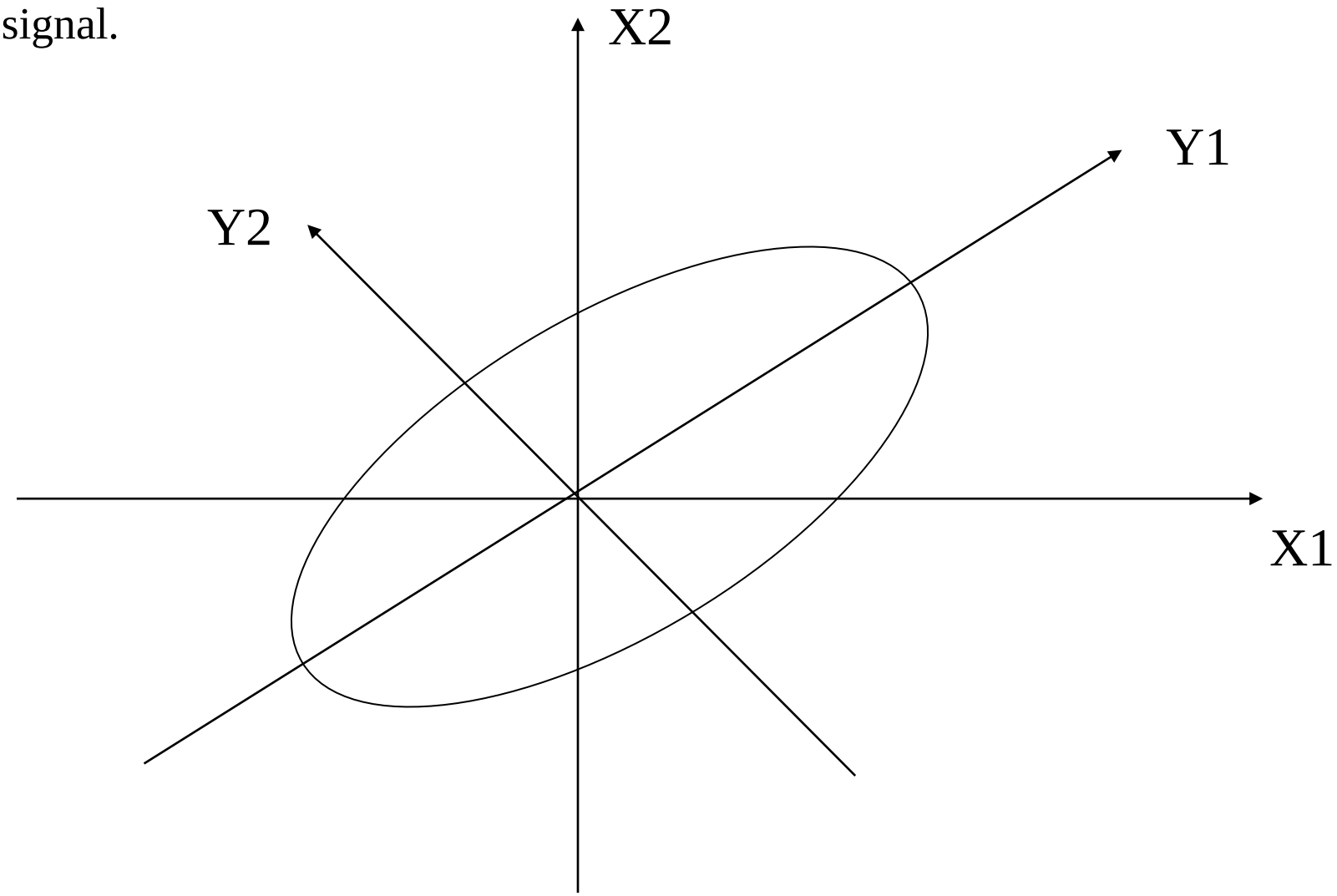
  c <= *k* orthogonal vectors that can be best used to represent data

  – The original data set is reduced (projected) to one consisting of N data vectors on c principal components (reduced dimensions)

- Each data vector is a linear combination of the *c* principal component vectors
- Works for ordered and unordered attributes
- Used when the number of dimensions is large

Principal Component Analysis

•The principal components (new set of axes) give important information about variance.
•Using the strongest components one can reconstruct a good approximation of the original signal.



Numerosity Reduction

- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - E.g.: 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

Regression and Log-Linear Models

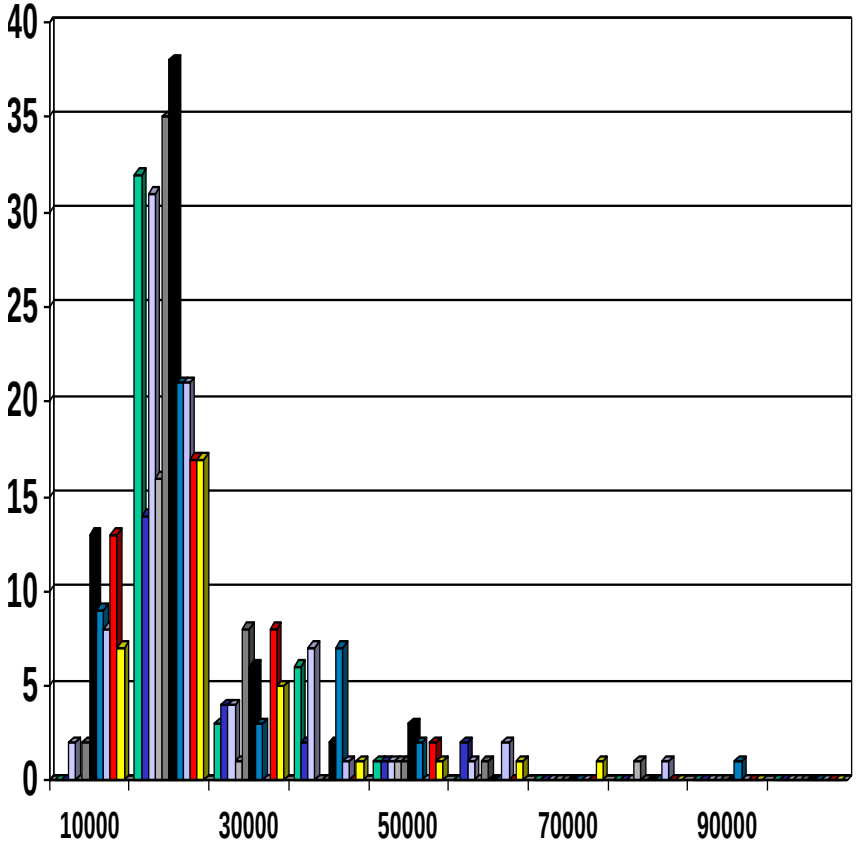
- Linear regression: Data are modeled to fit a straight line:
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable y to be modeled as a linear function of multidimensional feature vector (predictor variables)
- Log-linear model: approximates discrete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

Regression 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

- Approximate data distributions
- Divide data into buckets and store average (sum) for each bucket
- A bucket represents an attribute-value/frequency pair
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.



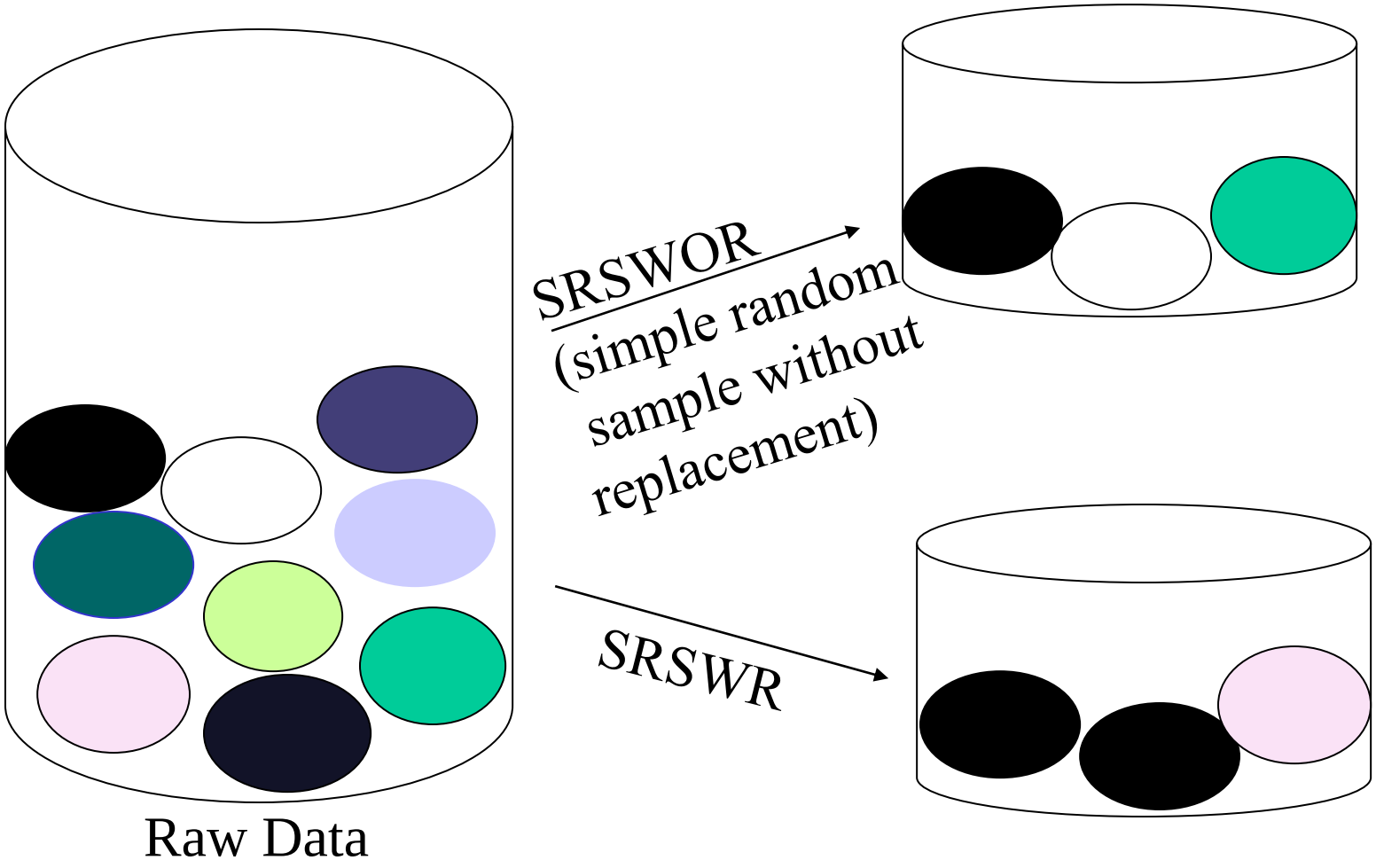
Clustering

- Partition data set into clusters, and store cluster representation only
- Quality of clusters measured by their diameter (max distance between any two objects in the cluster) or centroid distance (avg. distance of each cluster object from its centroid)
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering (possibly stored in multi-dimensional index tree structures (B+-tree, R-tree, quad-tree, etc))
- There are many choices of clustering definitions and clustering algorithms (further details later)

Sampling

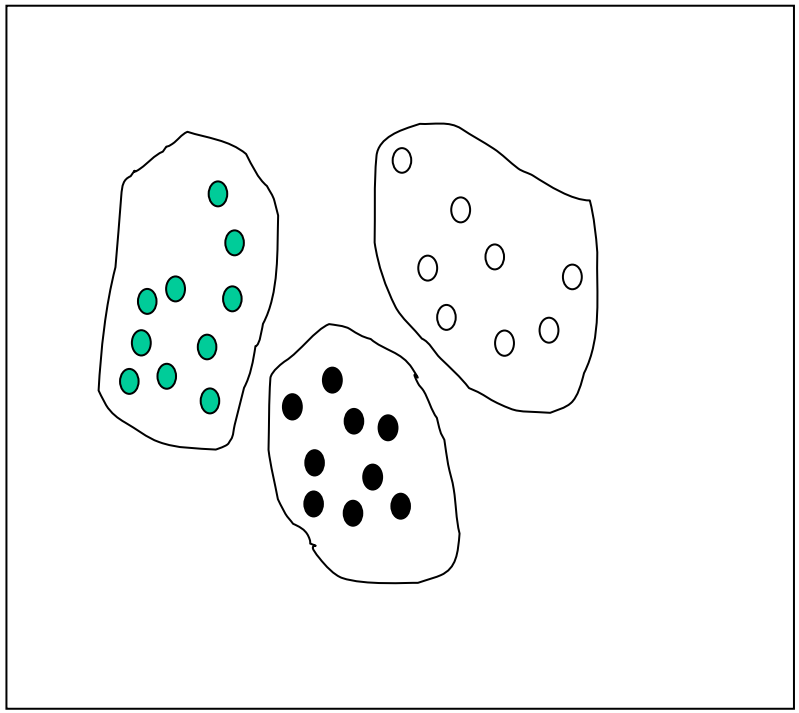
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Cost of sampling: proportional to the size of the sample, increases linearly with the number of dimensions
- Choose a representative subset of the data
  - Simple random sampling may have very poor 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 may not reduce database I/Os (page at a time).
- Sampling: natural choice for progressive refinement of a reduced data set.

Sampling


SRSWOR
(simple random sample without replacement)
SRSWR
Raw Data

Sampling

Raw Data



Cluster/Stratified Sample