# Data Preprocessing

# Overview

1. Data Reduction
    i. Dimensionality Reduction
    ii. Numerosity Reduction
    iii. Data compression

# Data Reduction

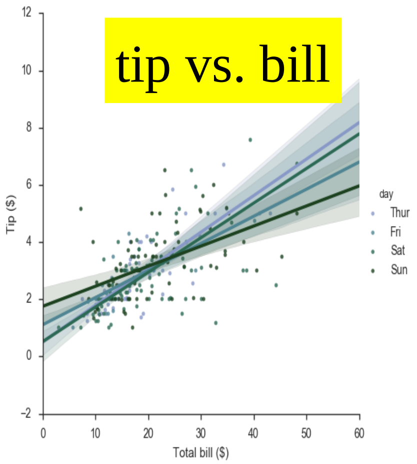
- **Data reduction**:
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- **Why data reduction?**
  - A database/data warehouse may store terabytes of data.
  - Complex data analysis may take a very long time to run on the complete data set.

# Data Reduction strategies

- **Dimensionality reduction:** Process of reducing the number of random variables or attributes
  - Wavelet transforms
  - Principal Components Analysis (PCA)
  - Feature subset selection, feature creation
- **Data compression:**
  - Transformations are applied to obtain a reduced or "compressed" representation of the original data.
  - **Lossless** : If the original data can be reconstructed from the compressed data without information loss.
  - **Lossy:** Reconstructing only approximation of the original data

# Data Reduction: **Parametric vs. Non-Parametric Methods**

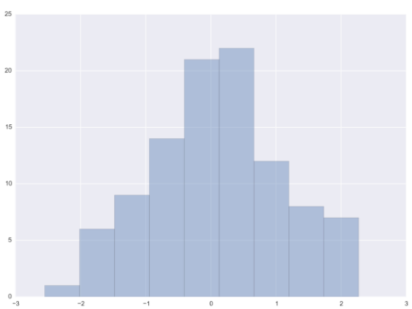
- **Numerosity Reduction:** Reduce data volume by choosing alternative, *smaller forms* of data representation



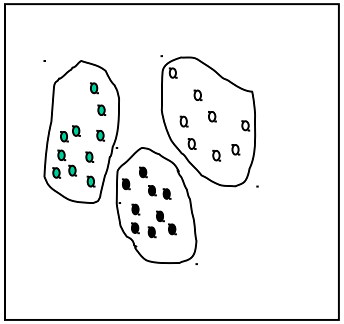

- **Parametric methods** (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and instead of actual data.
  - Ex.: Regression and Log-linear models

# Data Reduction: **Parametric vs. Non-Parametric Methods**

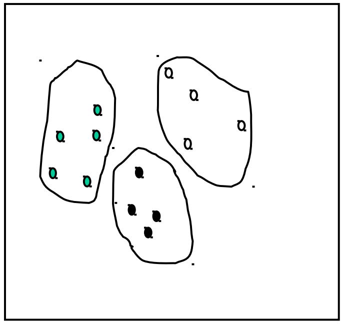
- **Non-parametric** methods :
  - Reduced representation of data
  - Do not assume models
  - Major families: histograms, clustering, sampling and data cube aggregration.



Histogram



Clustering on the Raw Data



Stratified Sampling

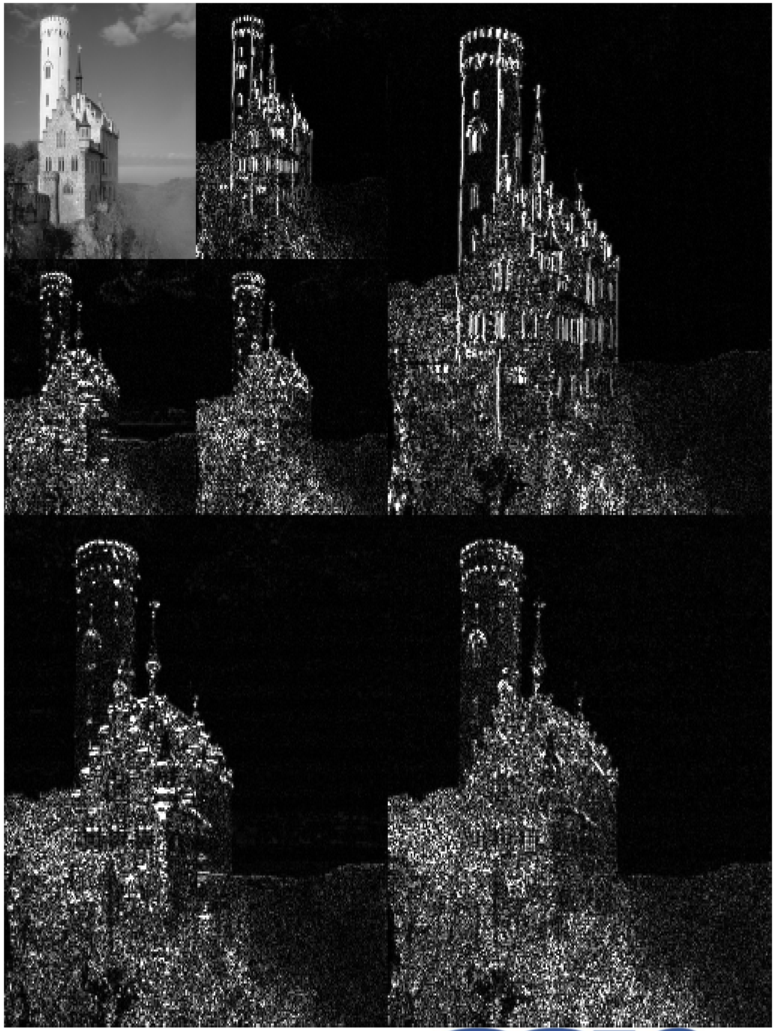
# Data Reduction 1: Dimensionality Reduction

- **Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful

- **Dimensionality reduction**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization

# Wavelet Transform: A Data Compression Technique

- Wavelet Transform
  - Decomposes a signal into different frequency subbands
  - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable



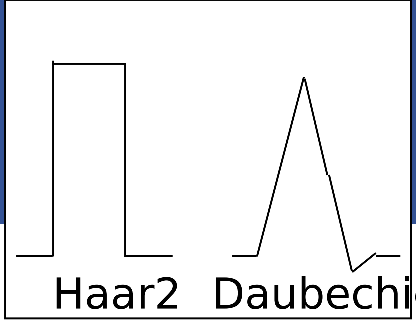
# Wavelet Transform: A Data Compression Technique

- Let X=(x1,x2,……xn) be a tuple of n-dimensional data vector depicting ‘n’ measurements made from ‘n’ database attributes.
- DWT When applied to data vector X, transforms it to numerically different vector X^n of wavelet coefficients
- Wavelet transformed data can be truncated.
- Compressed approximation:
  - Store only a small fraction of the strongest of the wavelet coefficients.
  - Other coefficient are set to zero.

# Wavelet Transform: Properties

- The resulting wavelet representation is very sparse.
- Computationally fast since performed in wavelet space.
- This technique  works to remove noise without smoothening out main features of the data.
- Given a set of coefficients an approximation of original  data can be constructed by applying the inverse of the DWT.
- Wavelet transform technique:
  - Good for sparse or skewed data
  - Data with ordered attributes and multidimensional
- Used in real world applications:
  - Compression of fingerprint images,Computer vision,
  - Analysis of time-series data and data cleaning

# Wavelet Transformation




- **Hierarchical Pyramid Algorithm:**
  - Length, L, must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform applies 2 functions: smoothing (sum or weighted average) and weighted difference (detailed feature).
  - Applies to pairs of data.
  - The result will contain low-frequency version of smoothed data and the high frequency content.
  - Applies two functions recursively, until reaches the desired length

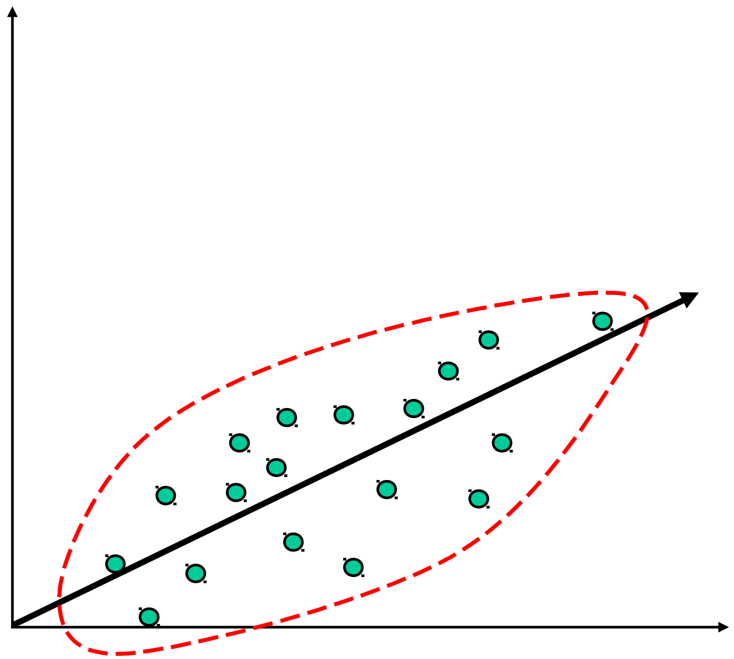
# Wavelet Decomposition

- **Wavelets:** A math tool for space-efficient hierarchical decomposition of functions
- S = [2, 2, 0, 2, 3, 5, 4, 4] can be transformed to S^ = [2 3/4, -1 1/4, 1/2, 0, 0, -1, -1, 0]
- Compression: many small detail coefficients can be replaced by 0's, and only the significant coefficients are retained

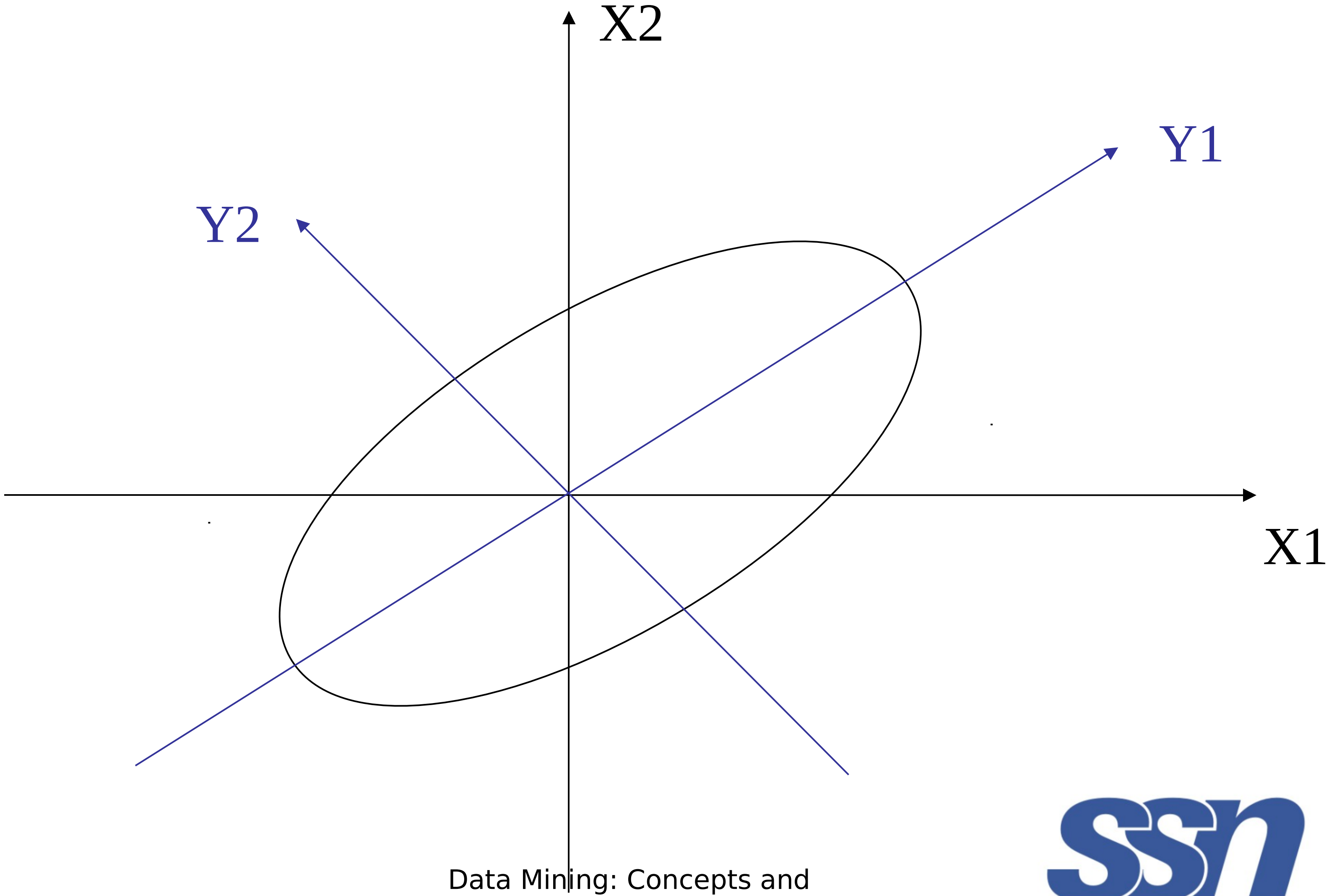
| Resolution | Averages | Detail Coefficients |
|---|---|---|
| 8 | $[2, 2, 0, 2, 3, 5, 4, 4]$ | |
| 4 | $[2, 1, 4, 4]$ | $[0, -1, -1, 0]$ |
| 2 | $[1\frac{1}{2}, 4]$ | $[\frac{1}{2}, 0]$ |
| 1 | $[2\frac{3}{4}]$ | $[-1\frac{1}{4}]$ |

# Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data.
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.
- The eigenvectors of the covariance matrix is obtained and these eigenvectors define the new space.



# Principal Component Analysis



# Principal Component Analysis (Steps)

- Given N data vectors from n-dimensions, find $k \le n$ orthogonal vectors (principal components) that can be best used to represent data.
- PCA combines the essence of attributes by creating alternate small set of variables.
- Basic procedure for PCA:
  - Normalize input data: Each attribute falls within the same range
  - Compute k orthonormal (unit) vectors, i.e., principal components.
  - Each input data (vector) is a linear combination of the k principal component vectors

# Principal Component Analysis (Steps)

- The principal components are sorted in order of decreasing "significance" or strength.
- The principal components act as new set of axes for the data providing information about variance.
- The size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance
- Using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only (ordered,unordered attributes, sparse and skewed data).

# Attribute Subset Selection

- **Feature selection** (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes
  - Best and worst attributes are determined using statistical significance or information gain measure

# Attribute Subset Selection

Heuristic search:

- Methods are greedy in nature
- Make a locally optimal choice that will lead to globally optimal solution
- Best and worst attributes determined by statistical significance (information gain measure)

Heuristic methods (due to exponential # of choices):

- Step-wise forward selection
- Step-wise backward elimination
- Combining forward selection and backward elimination
- Decision-tree induction

# Attribute Subset Selection

- **Step-wise forward selection:**
  - Starts with empty set of attributes.
  - The best of the attributes is determined and added
  - At each iteration the best of the remaining is added.
- **Step-wise backward elimination:**
  - Repeatedly eliminate the worst feature from the original in each iteration
- **Best combined feature selection and elimination**
- **Decision tree induction:**
  - Use feature elimination and backtracking

# Attribute Subset Selection

- **Decision tree induction:**
  - Constructs flowchart like structure
  - Internal node denotes a test on an attribute
  - Branch corresponds to an outcome of the test
  - External leaf corresponds to class predicition.
  - Attributes that don’t appear in the tree are irrelevant.

# Attribute Subset Selection

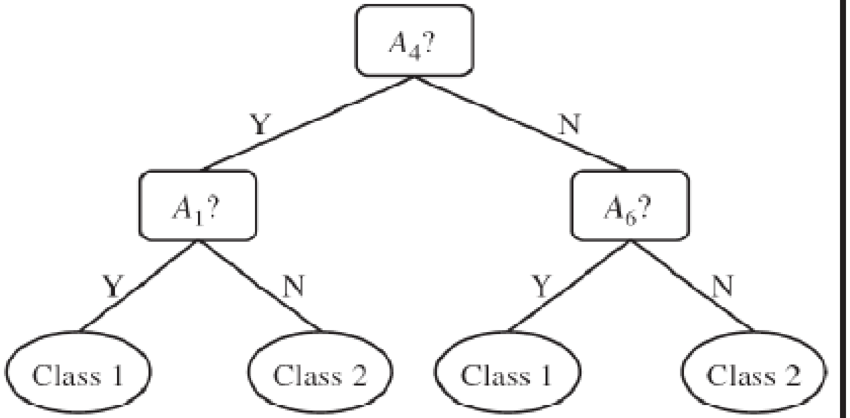
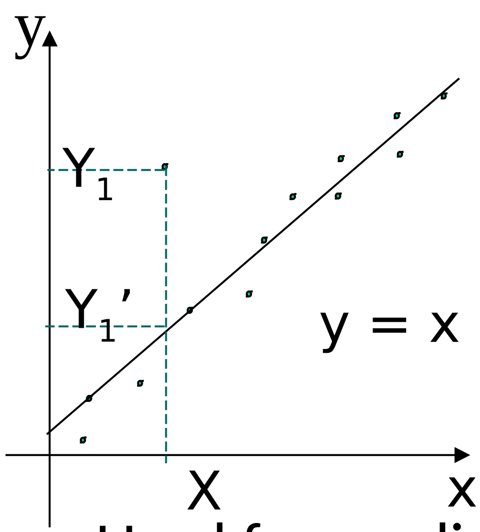
| Forward selection | Backward elimination | Decision tree induction |
|---|---|---|
| Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>Initial reduced set:<br>$\{\}$<br>=> $\{A_1\}$<br>=> $\{A_1, A_4\}$<br>=> Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ | Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>=> $\{A_1, A_3, A_4, A_5, A_6\}$<br>=> $\{A_1, A_4, A_5, A_6\}$<br>=> Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ | Initial attribute set:<br>$\{A_1, A_2, A_3, A_4, A_5, A_6\}$<br><br>=> Reduced attribute set:<br>$\{A_1, A_4, A_6\}$ |



**Figure 2.15.** Greedy (heuristic) methods for attribute subset selection

# Parametric Data Reduction: Regression Analysis

- **Regression analysis:** A collective name for techniques with 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 (also known as **explanatory variables or predictors**)



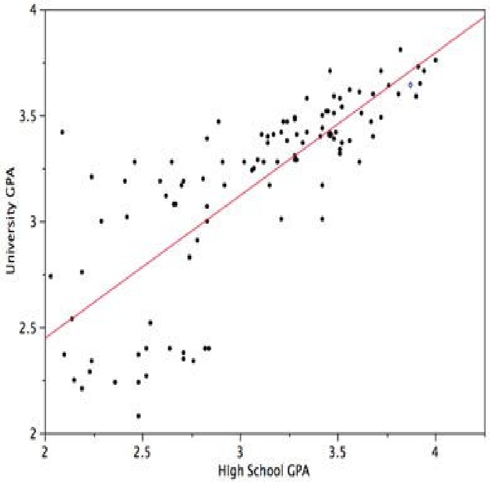

- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

# Parametric Data Reduction: Regression Analysis

- 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

# Linear and Nonlinear Regression

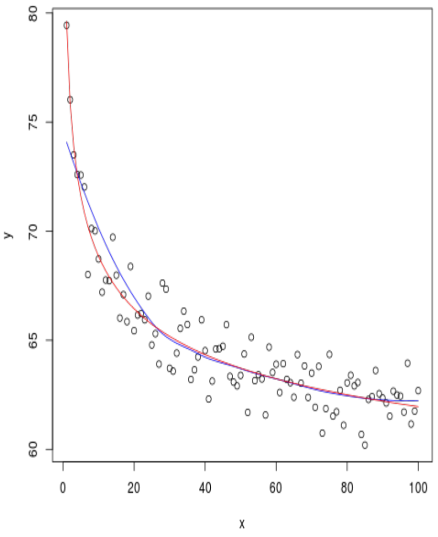
- **<u>Linear regression</u>:** $Y = w X + b$ Data modeled to fit a straight line




  - Often uses the least-square method to fit the line
  - X and Y are numeric database attributes
  - Two regression coefficients, *w* and *b*, specify the line and are to be estimated by using the least squares criterion.
  - *The coefficients minimizes the error between the actual line and estimated line*

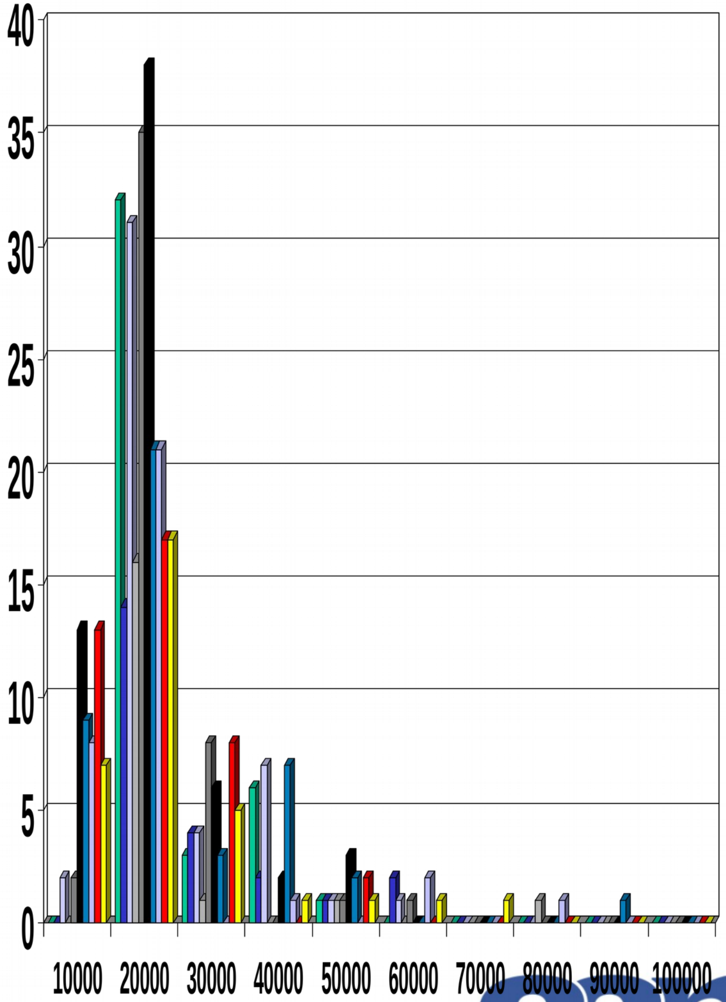
# Linear and Nonlinear Regression

- **Nonlinear regression**:
  - Data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables
  - The data are fitted by a method of successive approximations



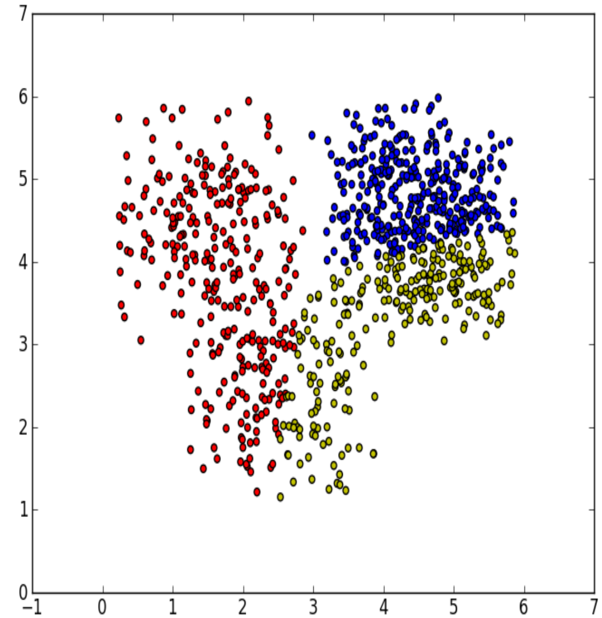
# Histogram Analysis

- Uses binning for data approximation
- Partition the data distribution of A into disjoint subsets referred as buckets
- Partitioning rules:
  - Equal-width: equal bucket range (width of $10 for price)
  - Equal-frequency (or equal-depth) Each bucket contains the same of contiguous data samples



# Clustering

- Partition data set into clusters so that the object within cluster are “similar” and “dissimilar” to objects in other clusters
- 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



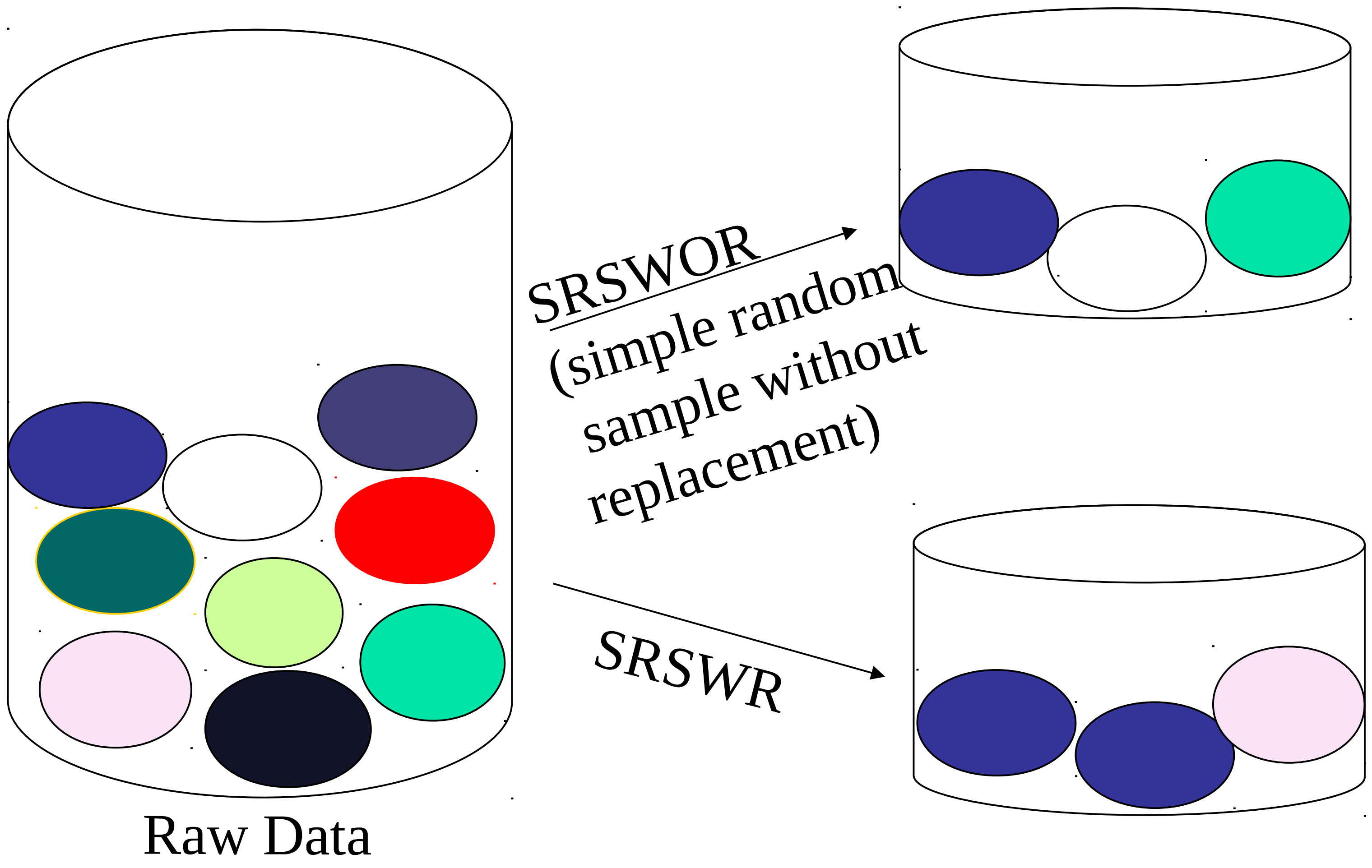
# Sampling

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

# Types of Sampling

- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

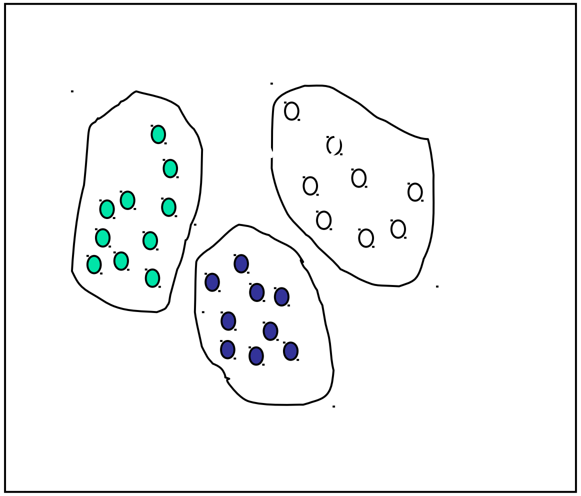
# Sampling: With or without Replacement

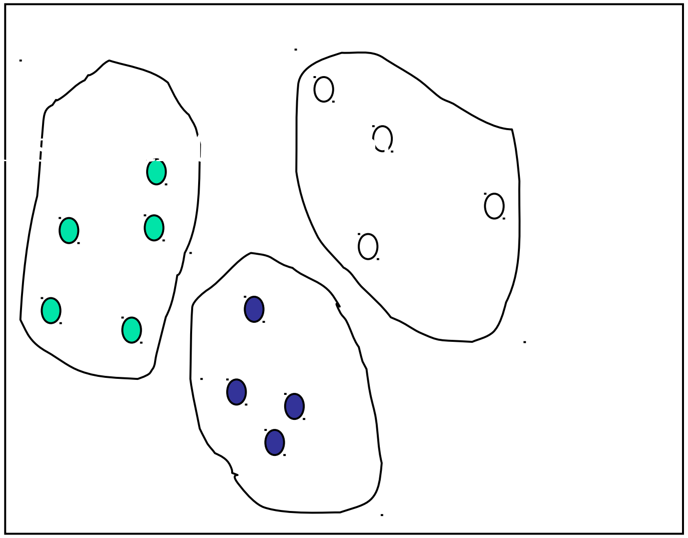


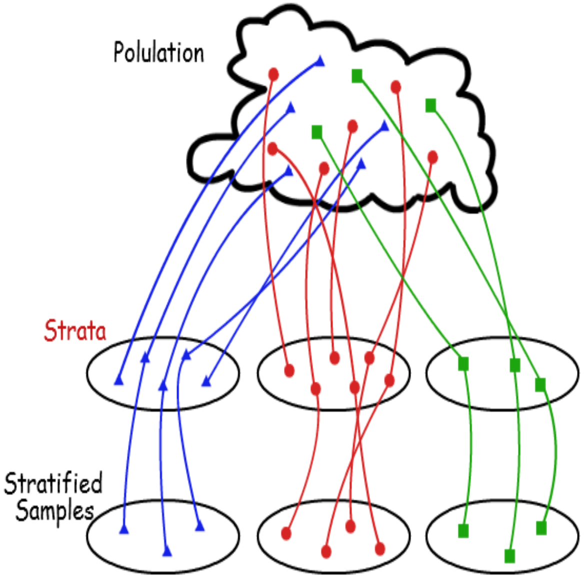
# Types of Sampling

- **Cluster Sample:**
  - Tuples in data set D are grouped into disjoint "clusters" M.
  - Obtained  s clusters where s<M
- **Stratified sampling:**
  - Divide D into mutually disjoint parts called strata.
  - Draw stratified samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

# Sampling: Cluster or Stratified Sampling



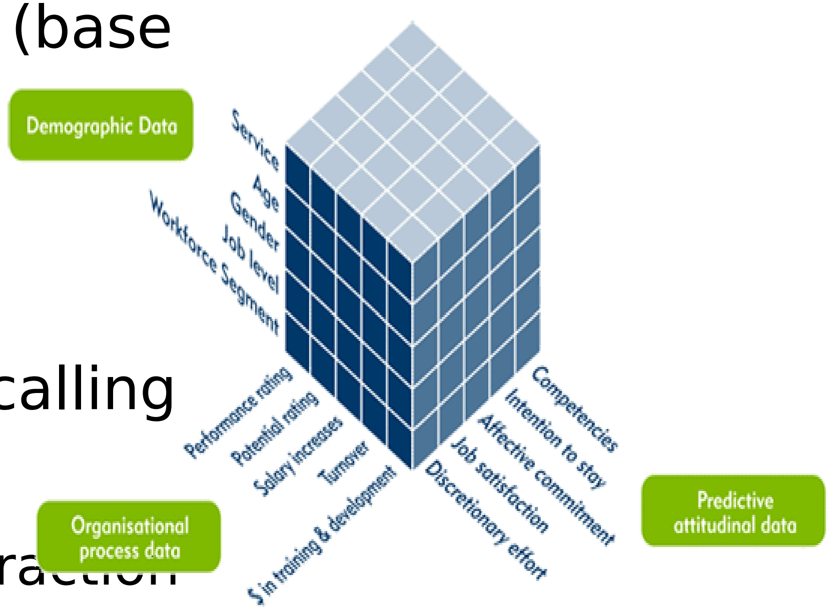





The stratified samples should in proportion to strata

# 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
- A cube at the highest level of abstraction is the (apex cuboid)
- Multiple levels of aggregation in data cubes
  - Each higher abstraction level further reduces the resulting data size

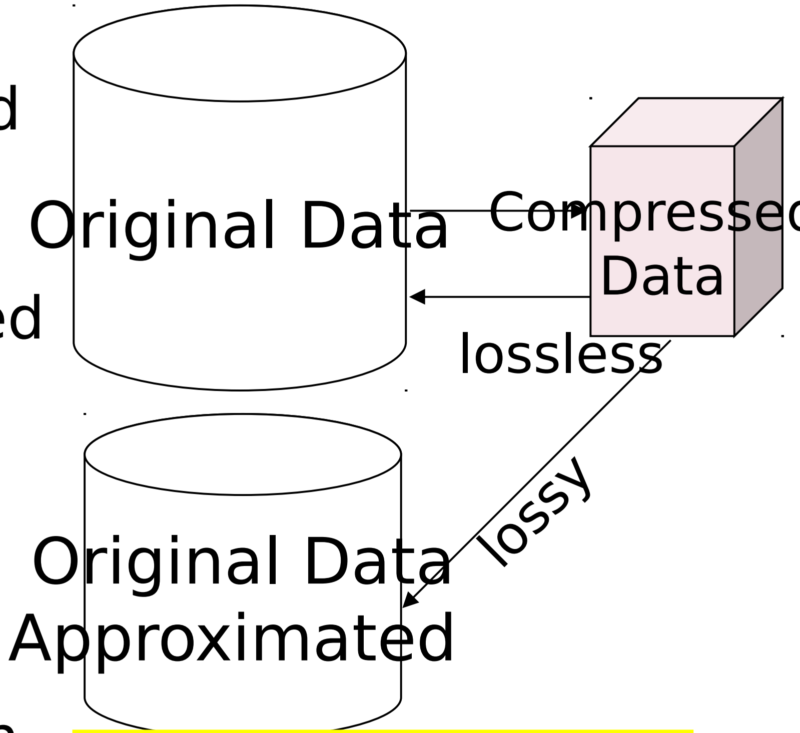


# Data Cube Aggregation

- 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
- Concept hierarchies may exist for each attribute allows analysis of data at multiple abstraction levels.

# 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




Lossy vs. lossless compression

# Data Compression

- **Time sequence is not audio**
  - Typically short and vary slowly with time
- Data reduction and dimensionality reduction may also be considered as forms of data compression