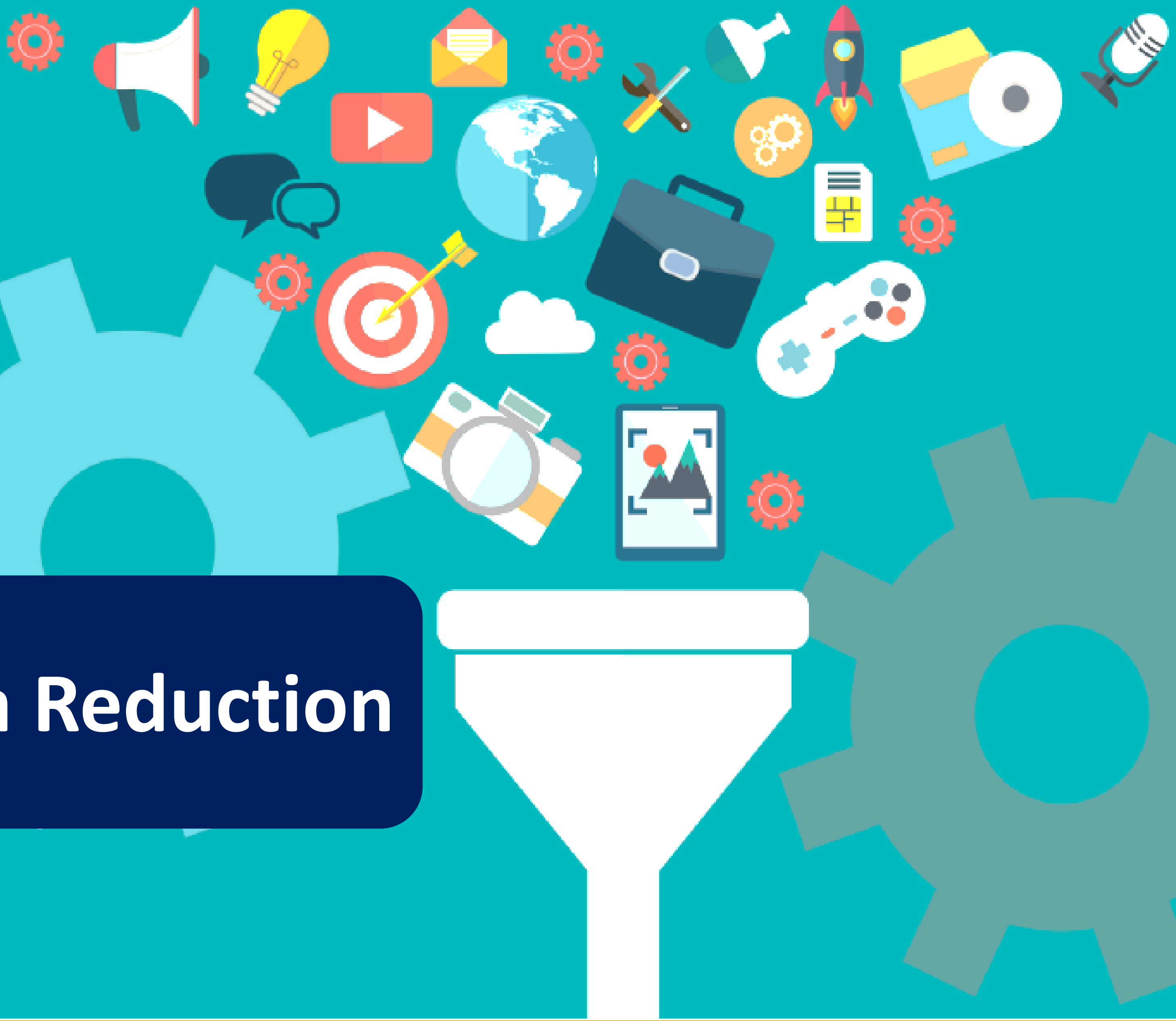
DATA MINING PRE-PROCESSING

Outline

1. Why data preprocessing?
2. Data cleaning
3. Data integration and transformation
4. Data reduction
5. Discretization and concept hierarchy generation
6. Summary



4

Data Reduction

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

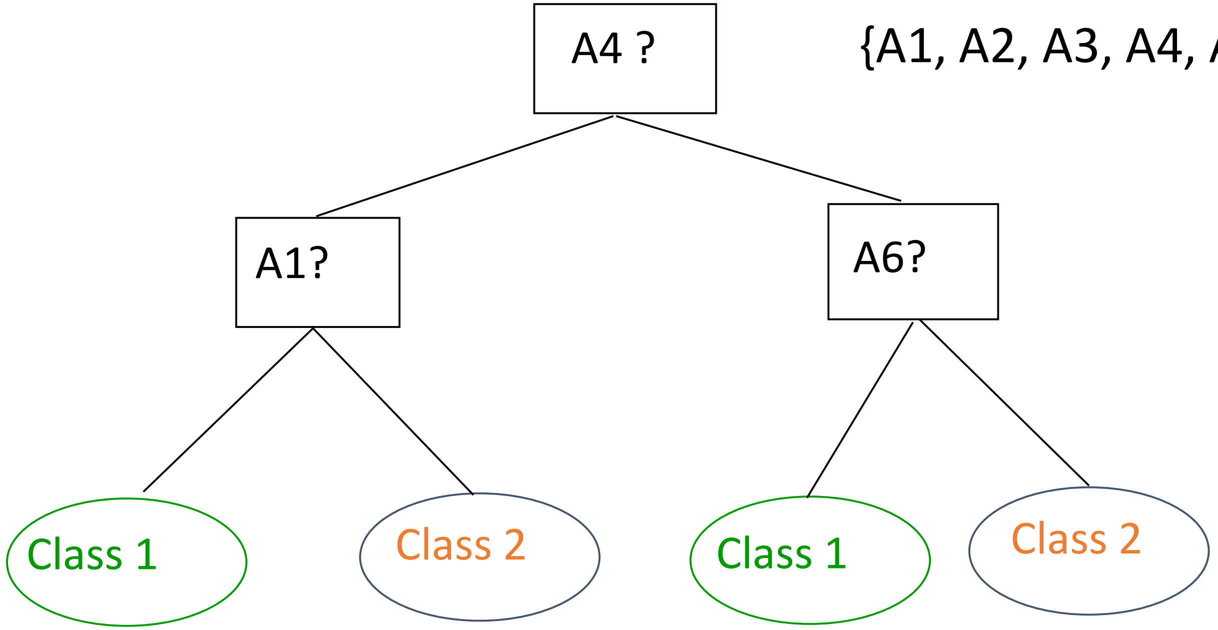
Dimensionality Reduction

- 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

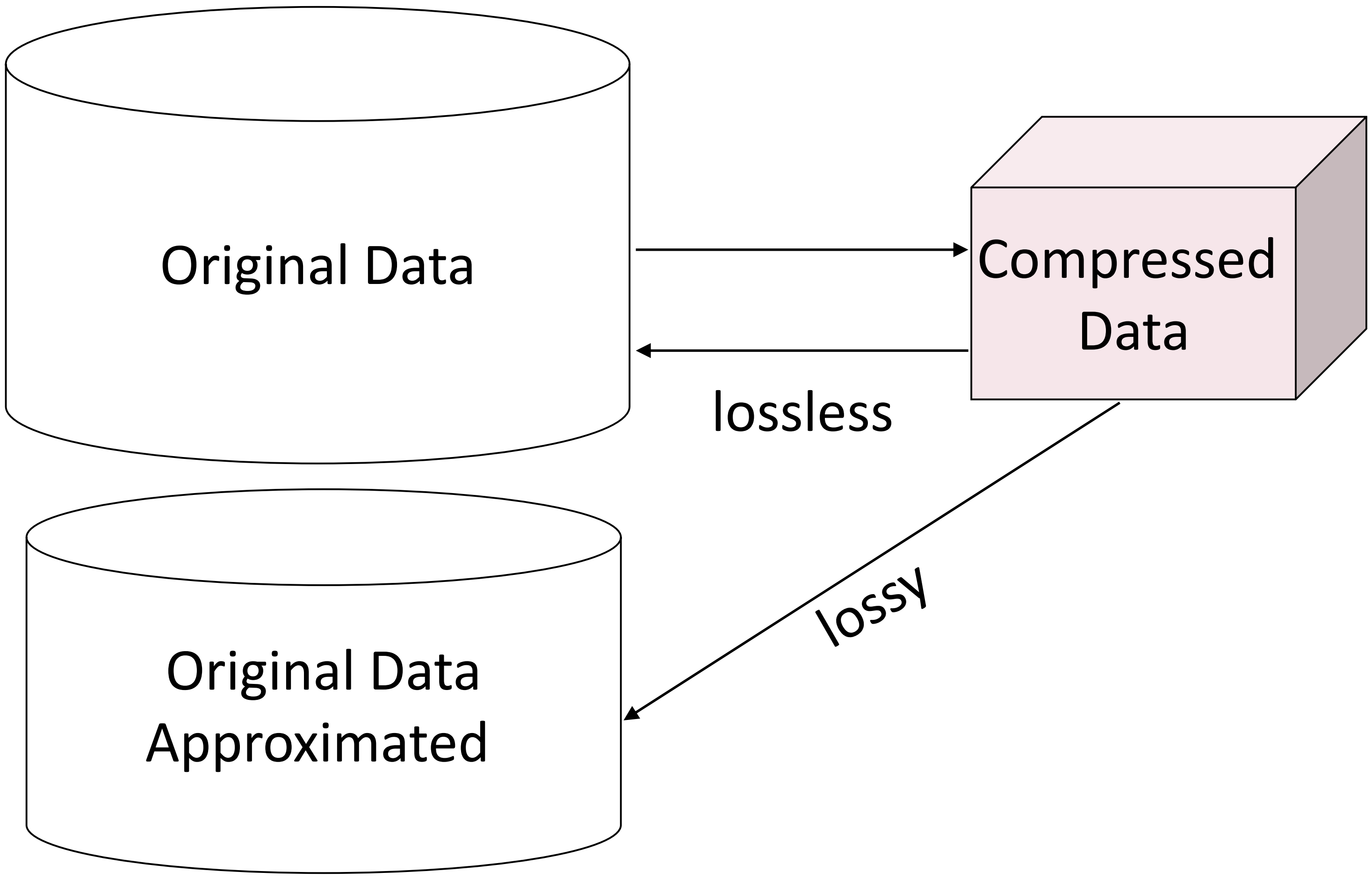
Audio/video, image compression

- 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



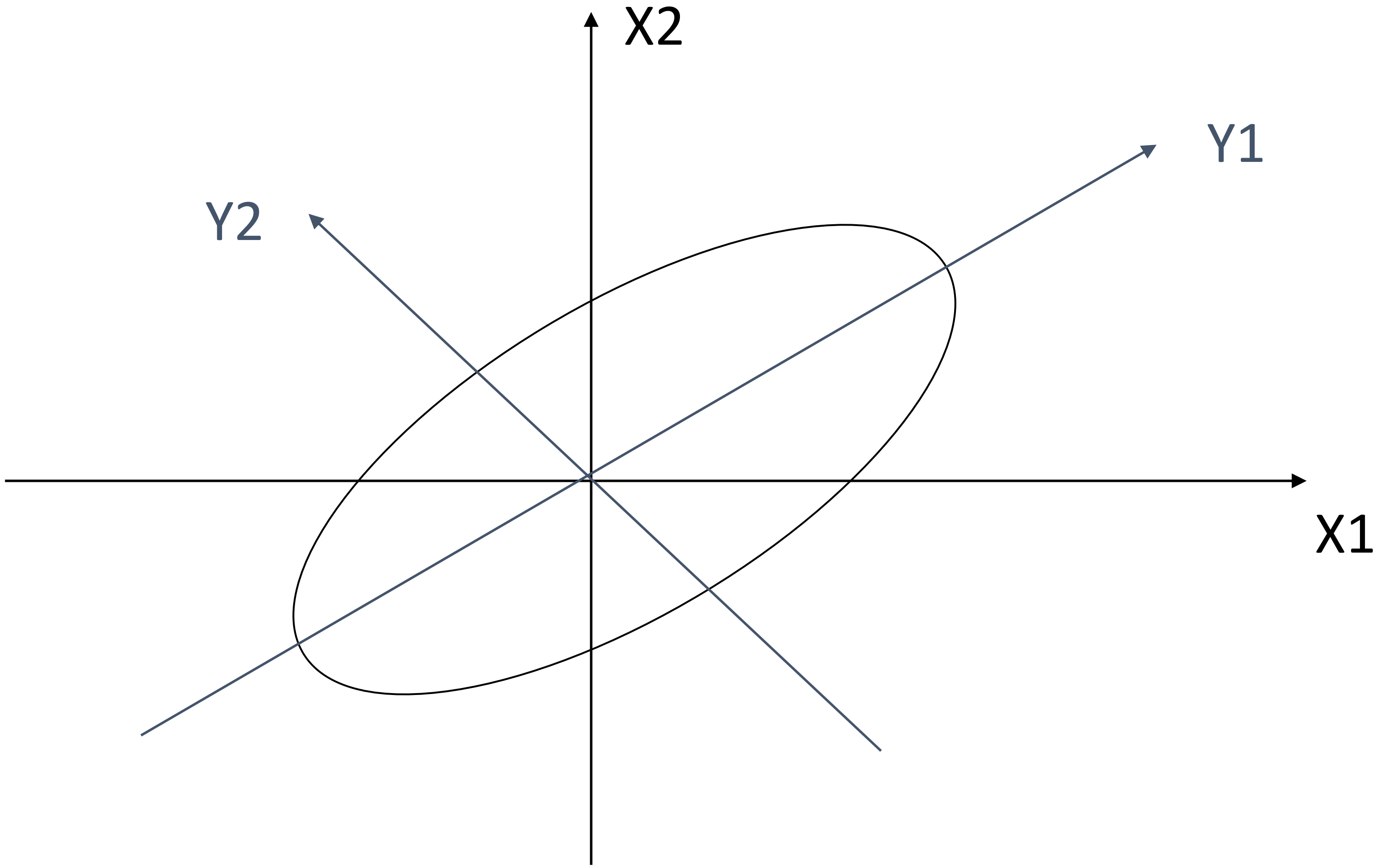
Principal Component Analysis (PCA) Karhunen-Loeve (K-L) method

- 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

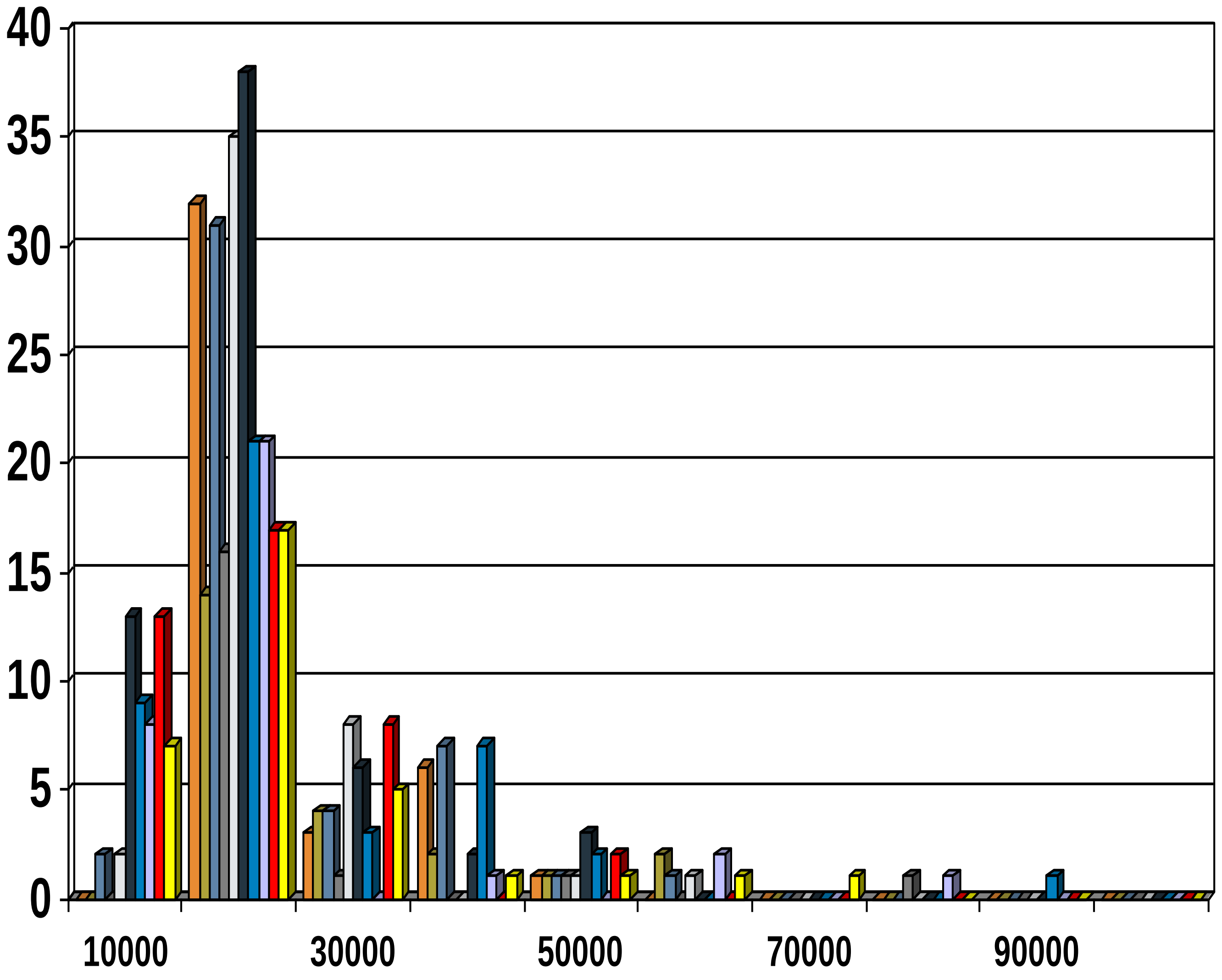
- Linear regression: Data are modeled to fit a straight line:
  - ✓Often uses the least-square method to fit the line
- Multiple regression: allows a response variable y to be modeled as a linear function of multidimensional feature vector (predictor variables)
- Log-linear model: approximates discrete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

Regression Analysis and Log-Linear Models

- Linear regression: $Y = b_0 + b_1 X_1$
  - 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 $Y1, Y2, \ldots, X1, X2, \ldots$.
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$.
  - 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)

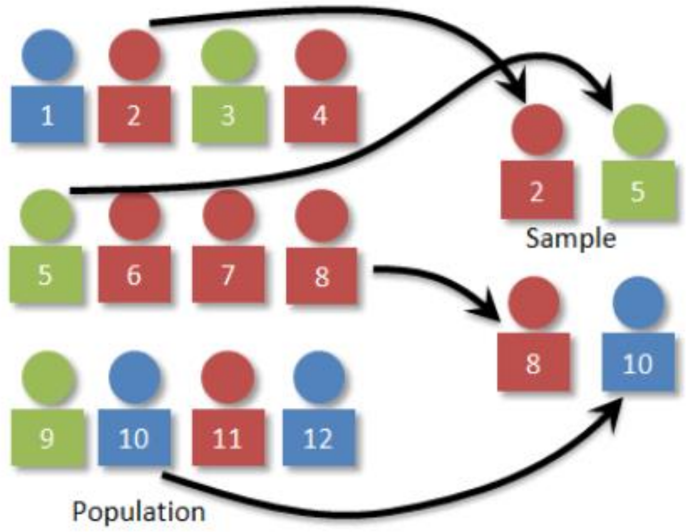
Sampling

- 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 Techniques

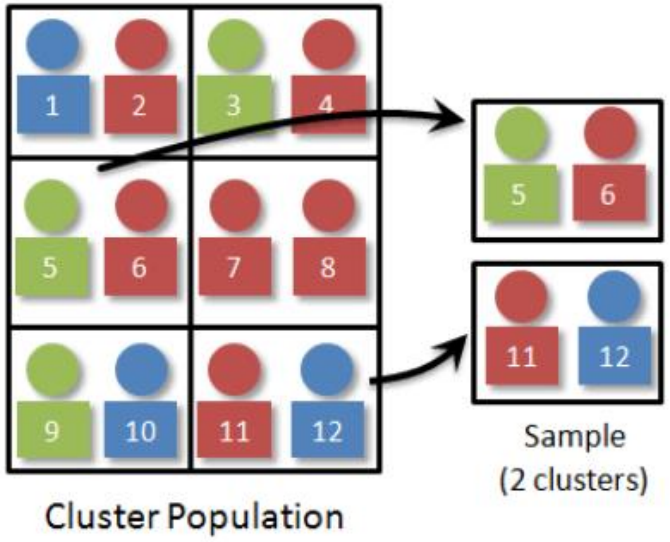
- ✓ Random samples : Selected using chance method or random methods
- ✓ Systematic samples : Numbering each subject of the populations & select every k-th number
- ✓ Stratified samples : Dividing the population into groups according some characteristic that is important to the study, then sampling from each group
- ✓ Cluster samples : Dividing the population into sections/clusters, then randomly select some of those cluster & then chose all members from those selected cluster

Random Sampling

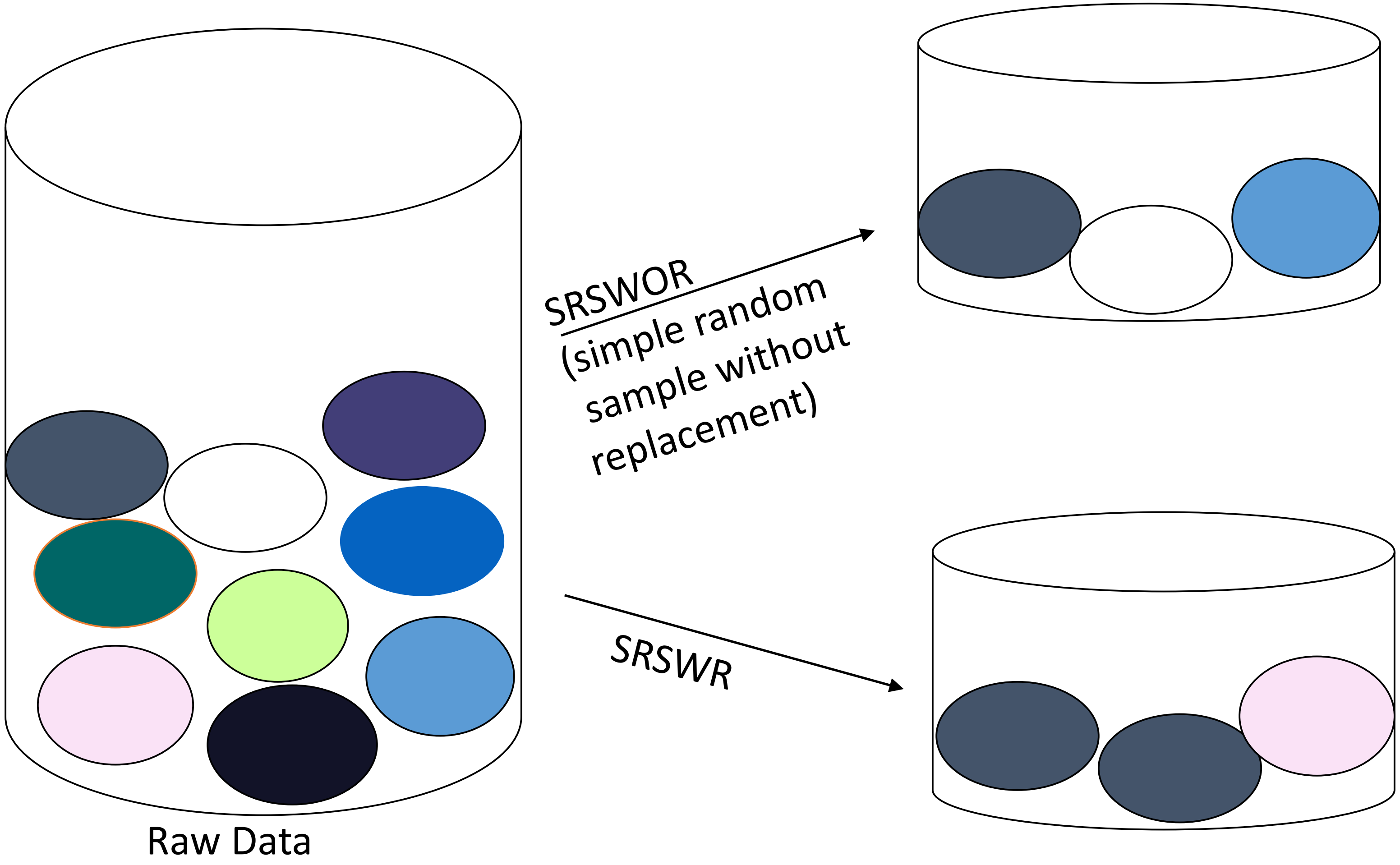


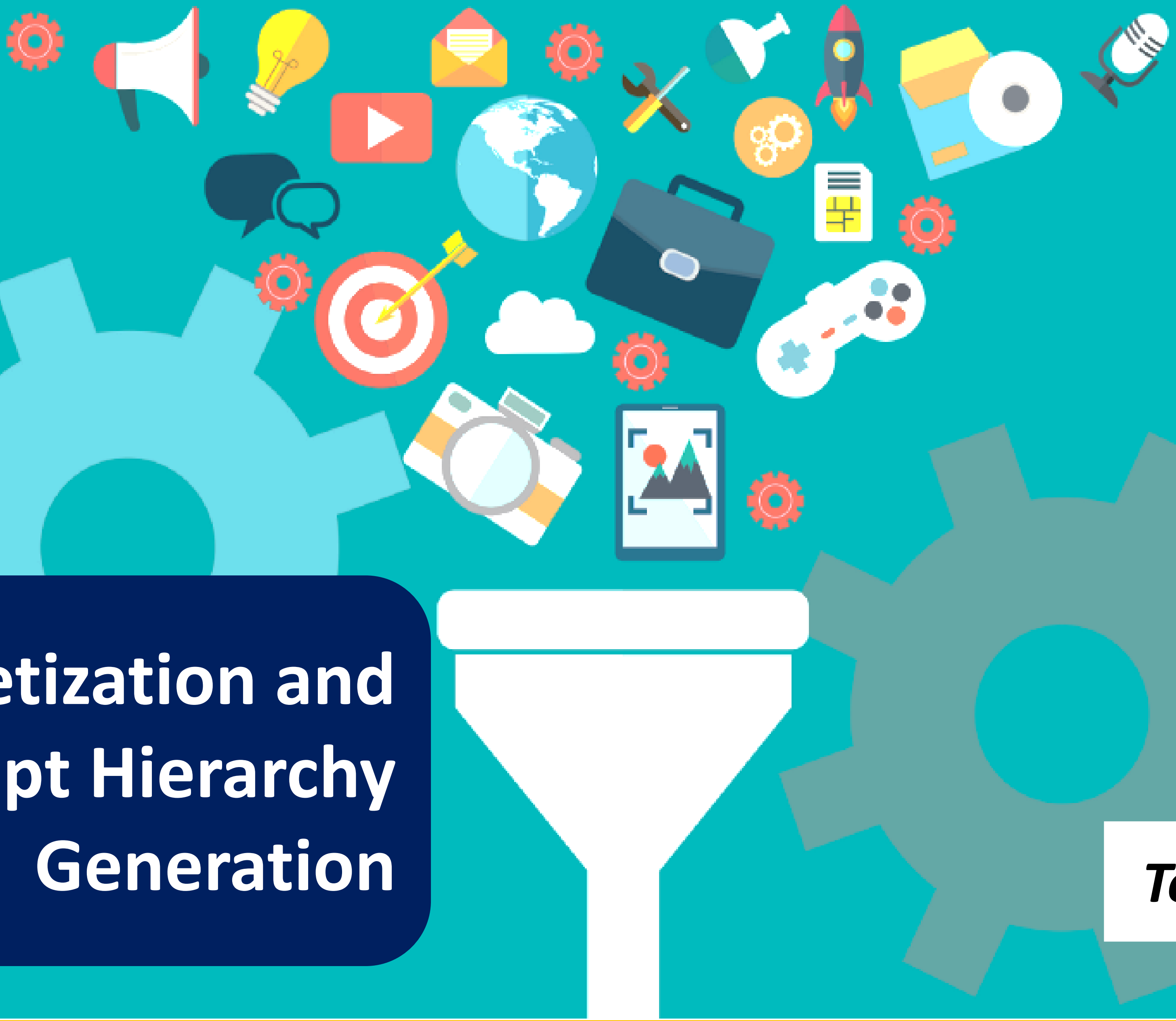
Stratified Sampling

Cluster Sampling



Sampling


SRSWOR
(simple random sample without replacement)
SRSWR
Raw Data



5 Discretization and Concept Hierarchy Generation

To be continued