Data Preprocessing

——Data Reduction——

徐华

清华大学 计算机系 智能技术与系统国家重点实验室

xuhua@tsinghua.edu.cn

Data Preprocessing

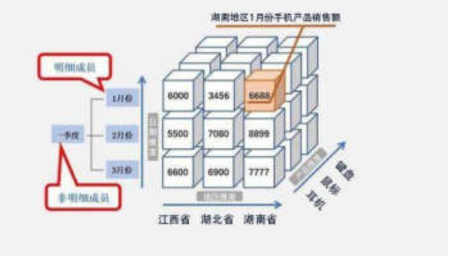
- About data
- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Reduction Strategy

- A data 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
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation (refer to chapter 4)
  - Dimensionality reduction—remove unimportant attributes
  - Data Compression—wavelet transform
  - Numerosity reduction—fit data into models
  - Discretization and concept hierarchy generation

Data Cube Aggregation

- The lowest level of a data cube (base cuboid)
  - The aggregated data for an individual entity of interest
  - E.g., a 3C selling 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
- Queries regarding aggregated information should be answered using data cube, when possible

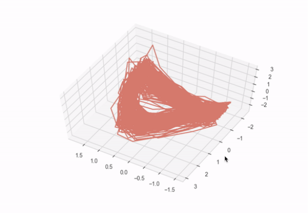


Dimensionality Reduction

- Feature selection（特征选择）

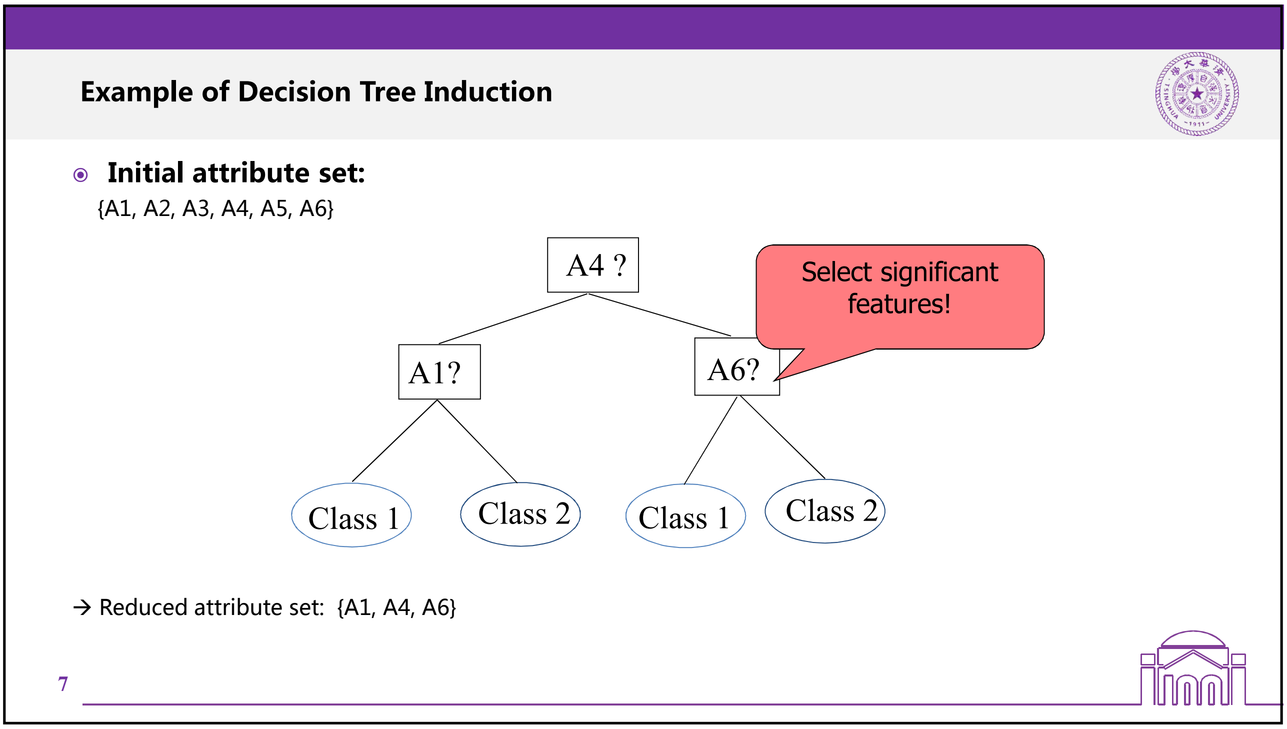
  Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features

  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods （启发式方法）
  - step-wise forward selection
  - step-wise backward elimination
  - combining forward selection and backward elimination
  - decision-tree induction



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

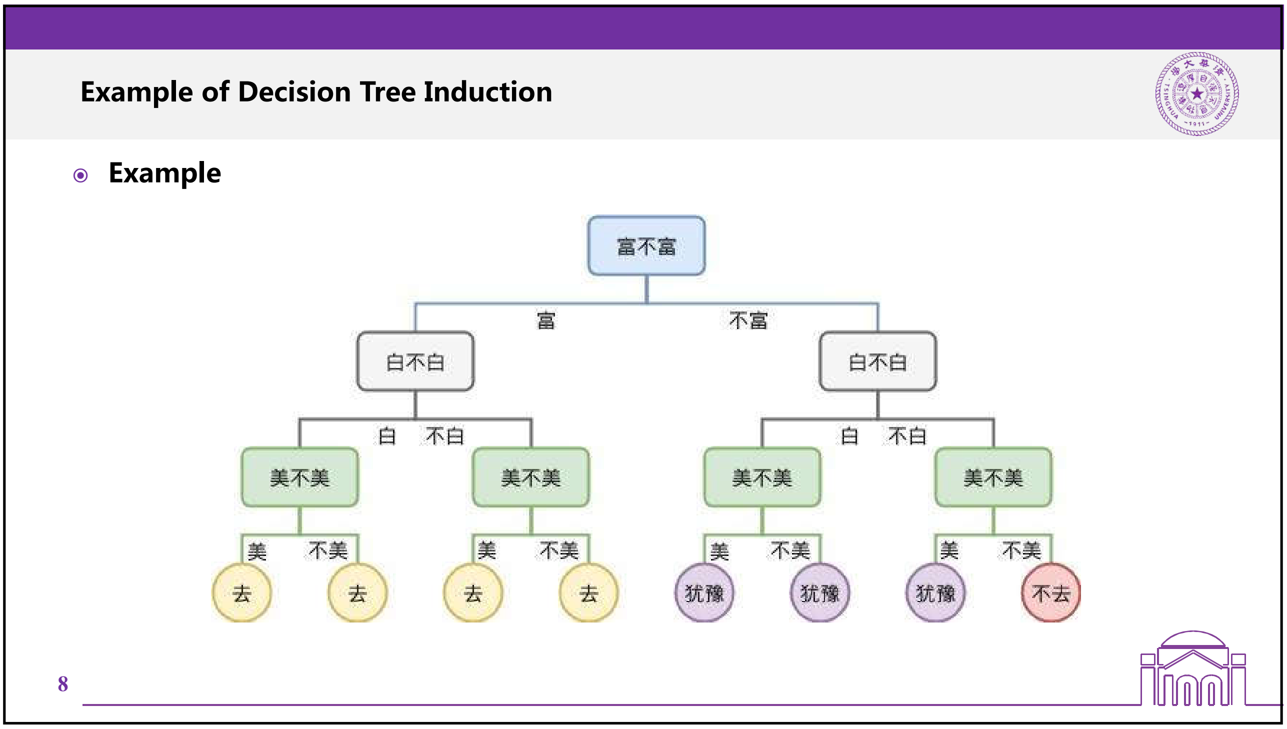


Example of Decision Tree Induction

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

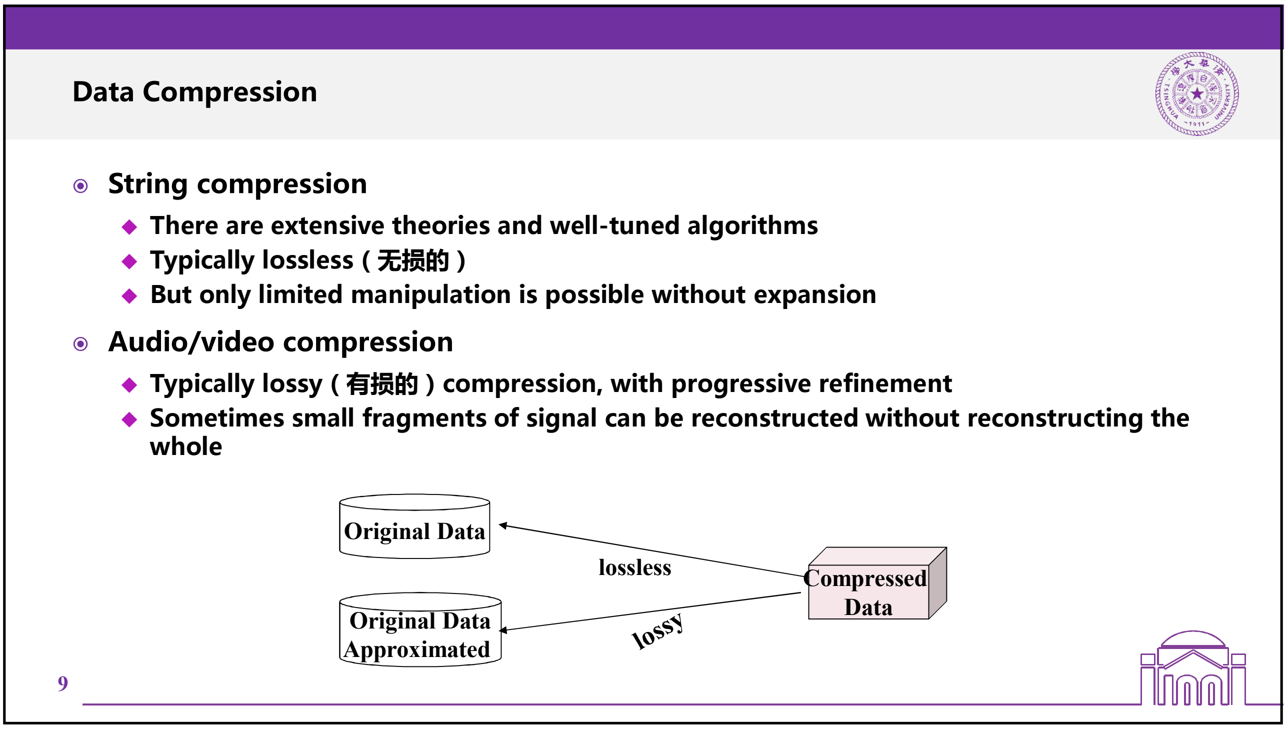
Select significant features!

→ Reduced attribute set: {A1, A4, A6}



Example of Decision Tree Induction

- **Example**



Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless（无损的）
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy（有损的）compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

Original Data

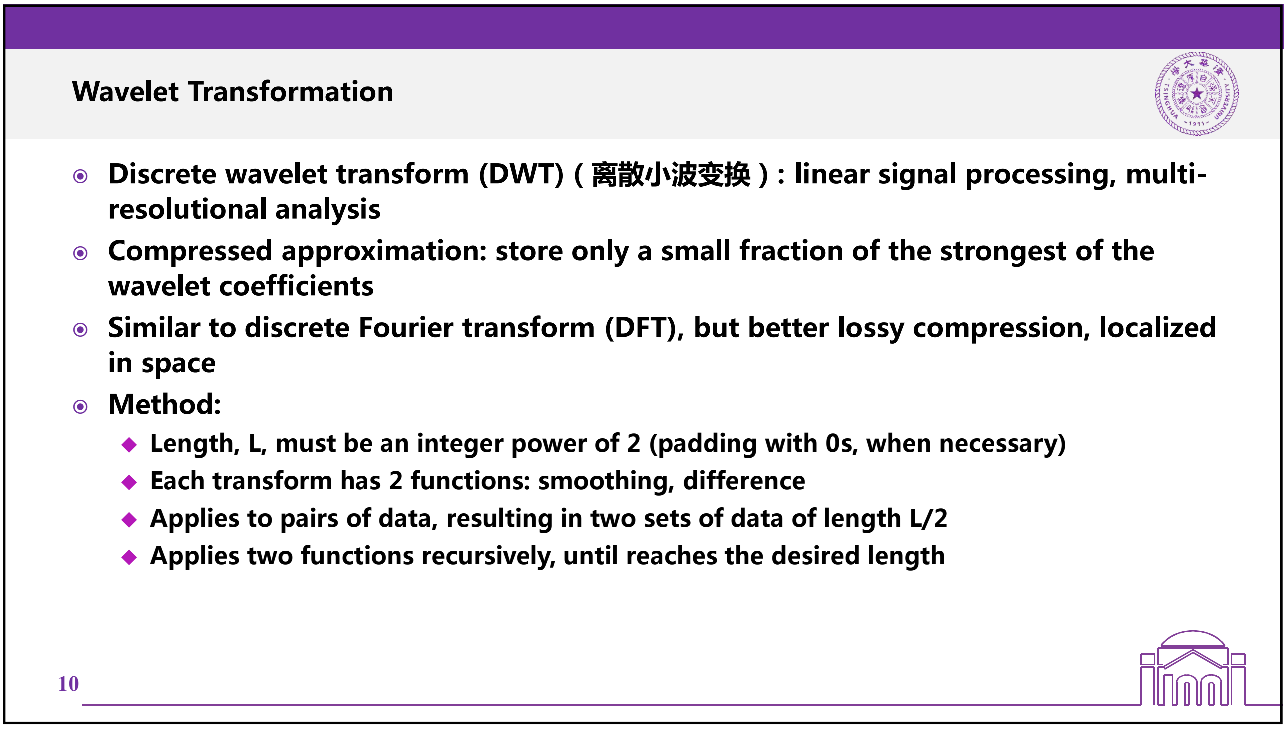
lossless

Compressed Data

Original Data Approximated

lossy

9

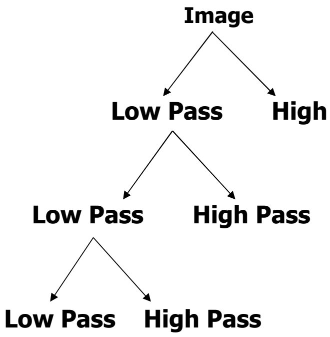


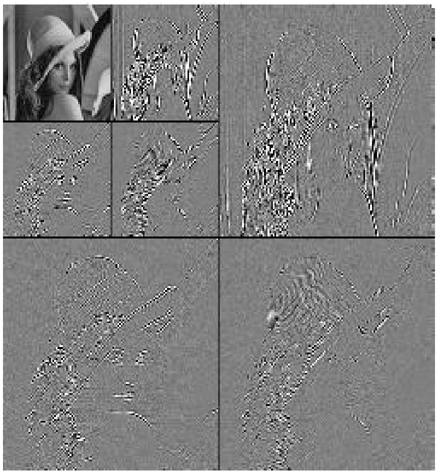
Wavelet Transformation

- Discrete wavelet transform (DWT)（离散小波变换）: linear signal processing, multi-resolutional analysis
- 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
- Method:
  - Length, L, must be an integer power of 2 (padding with 0s, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two sets of data of length L/2
  - Applies two functions recursively, until reaches the desired length

10

DWT for Image Compression





11

Principal Component Analysis

- Given *N* data vectors from *n*-dimensions, find *k* ≤ *n* orthogonal vectors (*principal components*) that can be best used to represent data
- Steps
  - 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
  - The principal components are sorted in order of decreasing “significance” or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only
- Used when the number of dimensions is large

12

Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
  - Assume the data fit some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Example: Log-linear models（对数线性模型）—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

Data Reduction Method (1): Regression and Log-Linear Models

- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method（最小二乘法） 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

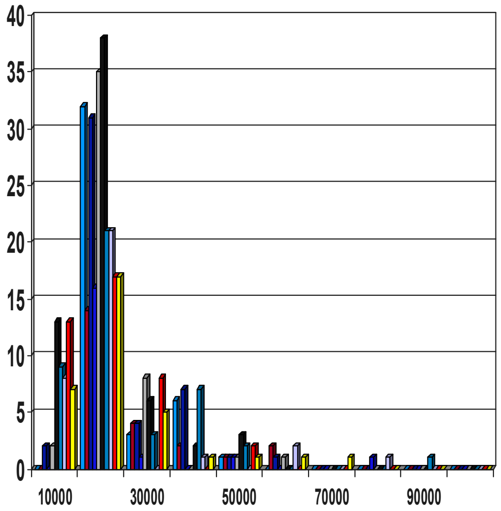
Regression and Log-Linear Models

- **Linear regression**: $Y = w X + b$
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$.
- **Multiple regression**: $Y = b0 + b1\ X1 + b2\ X2.$
  - Many nonlinear functions can be transformed into the above
- **Log-linear models（对数线性模型）**:
  - The multi-way table of joint probabilities is approximated by a product of lower-order tables
  - Probability: $p(a, b, c, d) = \alpha ab\ \beta ac \chi ad\ \delta bcd$

15

Data Reduction Method (2): Histograms

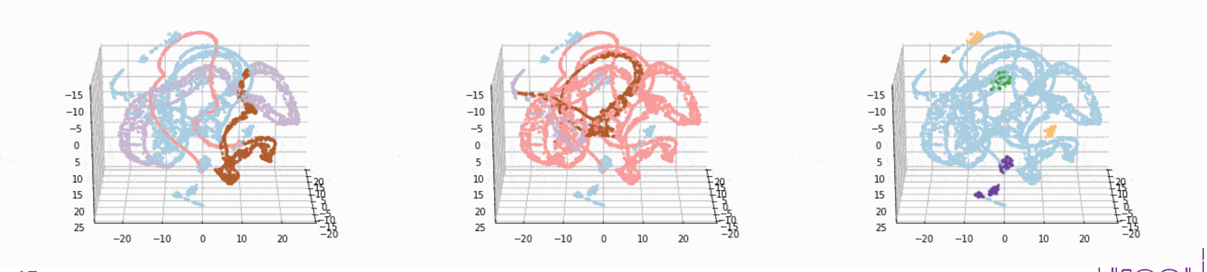
- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)
  - V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)
  - MaxDiff: set bucket boundary between each pair for pairs have the β–1 largest differences



16

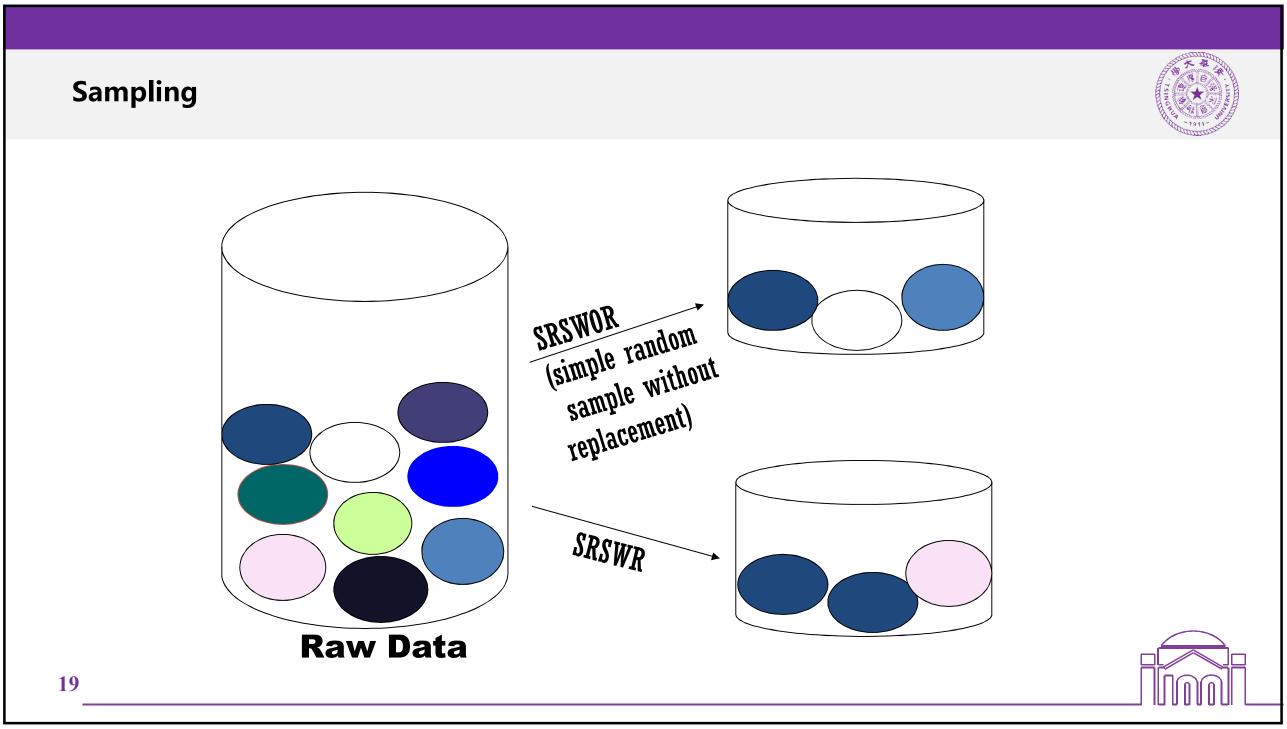
Data Reduction Method (3): Clustering

- Partition data set into clusters based on similarity, and 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
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth in the latter Chapter.



Data Reduction Method (4): 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
- SRSWOR, SRSWR, Cluster sampling
- Simple random sampling without replacement (SRSWOR)
- Develop adaptive sampling methods
  - Stratified sampling (分层采样) :
    - 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

19

Thanks !

20