数据分析实践

第3课. 数据预处理

兰 曼

计算机科学与技术学院

华东师范大学

内容

1 Why Data Preprocessing ?

- **Data in the real world is dirty**
  - **incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data**
    - **e.g., occupation=" "**
  - **noisy: containing errors or outliers**
    - **e.g., Salary="-10"**
  - **inconsistent: containing discrepancies in codes or names**
    - **e.g., Age="42" Birthday="03/07/1997"**
    - **e.g., Was rating "1,2,3", now rating "A, B, C"**
    - **e.g., discrepancy between duplicate records**
  - **Intentional (e.g., *disguised missing* data)**
    - **Jan. 1 as everyone's birthday?**

Why Is Data Dirty ?

- **Incomplete data may come from**
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- **Noisy data (incorrect values) may come from**
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- **Inconsistent data may come from**
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- **Duplicate records also need data cleaning**

Why Is Data Preprocessing Important ?

- **No quality data, no quality mining results!**
- **Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse**

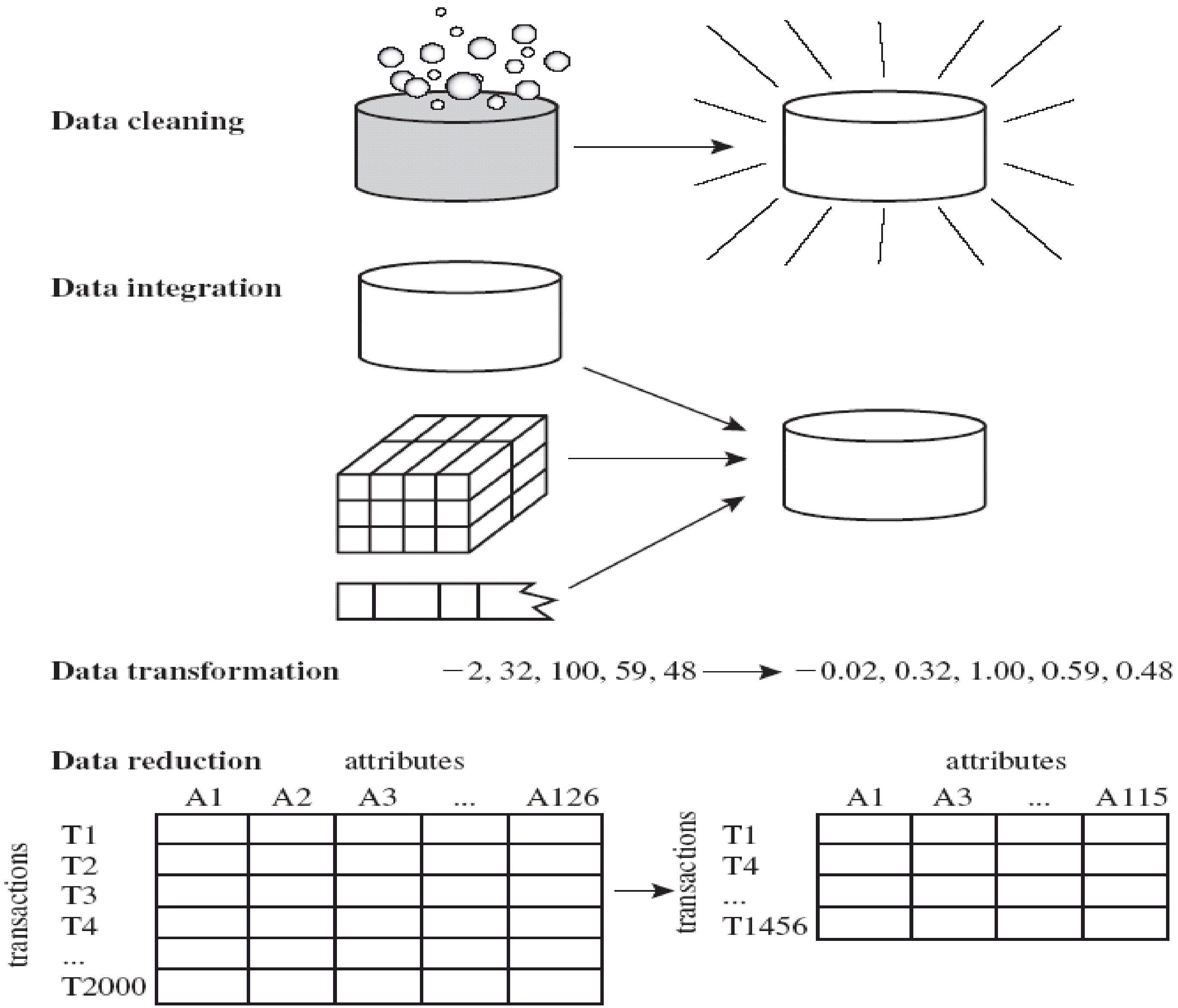
Multi-Dimensional Measure of Data Quality

- **Measures for data quality: A multidimensional view**
  - **Accuracy:** correct or wrong, accurate or not
  - **Completeness:** not recorded, unavailable, …
  - **Consistency:** some modified but some not, dangling, …
  - **Timeliness :** timely update?
  - **Believability:** how trustable the data are correct?
  - **Interpretability:** how easily the data can be understood?

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Forms of Data Preprocessing



2 Data Cleaning

- **data in the real world is dirty: lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error**
  - **incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = “ ” (missing data)
  - **noisy:** containing noise, errors, or outliers
    - e.g., *Salary* = “−10” (an error)
  - **inconsistent:** containing discrepancies in codes or names, e.g.,
    - *Age* = “42”, *Birthday* = “03/07/2010”
    - Was rating “1, 2, 3”, now rating “A, B, C”
  - **Intentional** (e.g., *disguised missing data*)
    - Jan. 1 as everyone’s birthday?

How to Handle Missing Data ?

- **Data is not always available, missing data may need to be inferred.**
- **Ignore the tuple: usually done when class label is missing (assuming the tasks in classification)—not effective when the percentage of missing values per attribute varies considerably.**
- **Fill in the missing value manually: tedious + infeasible?**
- **Fill in it automatically with:**
  - **a global constant : e.g., "unknown", a new class?!**
  - **the attribute mean**
  - **the attribute mean for all samples belonging to the same class: smarter**
  - **the most probable value: inference-based such as Bayesian formula or decision tree**

How to Handle Noisy Data ?

- **Noise:** random error or variance in a measured variable
- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

Data discrepancy detection

- Use metadata (e.g., domain, range, dependency, distribution)
- Check field overloading
- Check uniqueness rule, consecutive rule and null rule
- Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

Data migration and integration

- Data migration tools: allow transformations to be specified
- ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface

Integration of the two processes

- Iterative and interactive (e.g., Potter's Wheels)

- **Data integration:**
  - Combines data from multiple sources into a coherent store

- **Schema integration:** e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources

- **Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton

- **Detecting and resolving data value conflicts**
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification:* The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis and covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation (viewed as linear relationship)

- **Correlation measures the linear relationship between objects**
- **To compute correlation, we standardize data objects, A and B, and then take their dot product**

$$a'_k = (a_k - mean(A)) / std(A)$$

$$b'_k = (b_k - mean(B)) / std(B)$$

$$correlation(A, B) = A' \bullet B'$$

Correlation Analysis (Nominal Data)

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: An Example

2*2 Contingency table

	Play chess	Not play chess	Sum
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- **Expected = count(A=a_i) * count(B=b_j) / N**
- **X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)**

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- **It shows that *like_science_fiction* and *play_chess* are correlated in the group**

Correlation Analysis (Numerical Data)

- **Correlation coefficient** **(also called Pearson's product moment coefficient)**

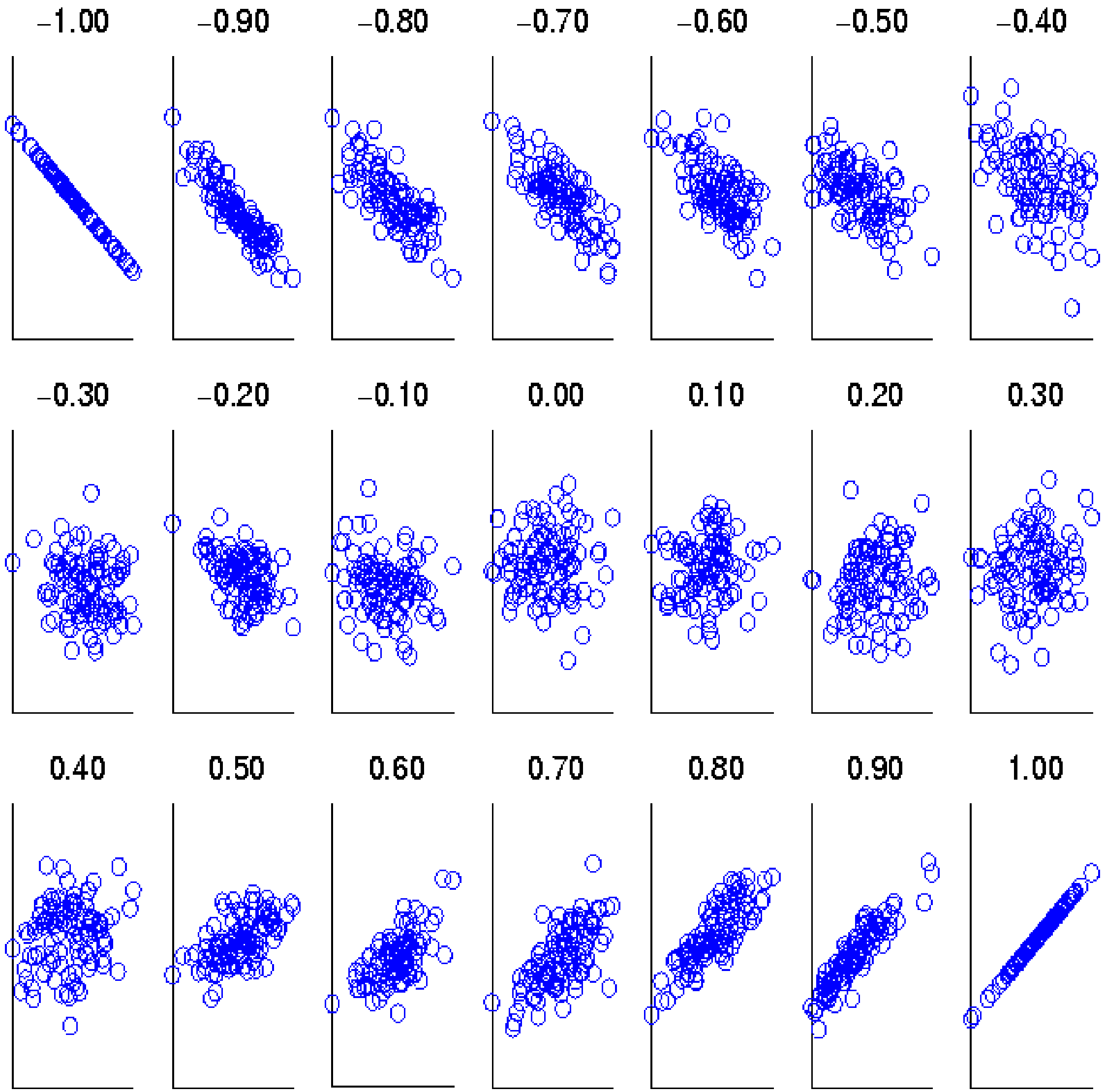
$$r_{A,B} = \frac{\sum (A-\bar{A})(B-\bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where *n* is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of **A** and **B**, $\boldsymbol{\sigma_A}$ and $\boldsymbol{\sigma_B}$ are the respective standard deviation of A and B, and Σ(AB) is the sum of the AB cross-product. $-1 \le r_{A,B} \le +1$

- **If $r_{A,B} > 0$, A and B are positively correlated *(A's values increase as B's)*. The higher, the stronger correlation.**
- **$r_{A,B} = 0$: independent;**
- **$r_{A,B} < 0$: negatively correlated**

Visually Evaluating Correlation

华东师范大学计算机科学与技术学院
School of Computer Science and Technology



Scatter plots showing the similarity from –1 to 1.

4 Data Reduction

- **Data reduction：** Obtain a reduced representation of the data set that is much smaller in volume (thus more efficient) but yet produce the same (or almost the same) analytical results
- **Why data reduction?** -- A database/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 strategies**
  - Dimensionality reduction, e.g., remove unimportant features (attributes)
    - Feature (Attribute) subset selection, feature creation
    - Principal Components Analysis (PCA)
    - Wavelet transforms
  - Numerosity reduction (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression

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
- The possible combinations of subspaces will grow exponentially

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

Dimensionality reduction techniques

- Supervised and nonlinear techniques (e.g., feature selection)
- Principal Component Analysis
- Wavelet transforms

(A) Attribute Subset Selection (i.e. Feature selection)

- Reduce the data set size by **removing irrelevant or redundant attributes** (or dimensions).
- **Redundant attributes**
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- **Irrelevant attributes**
  - Contain no information that is useful for the data mining task at hand
  - E.g., students' ID is often irrelevant to the task of predicting students' GPA

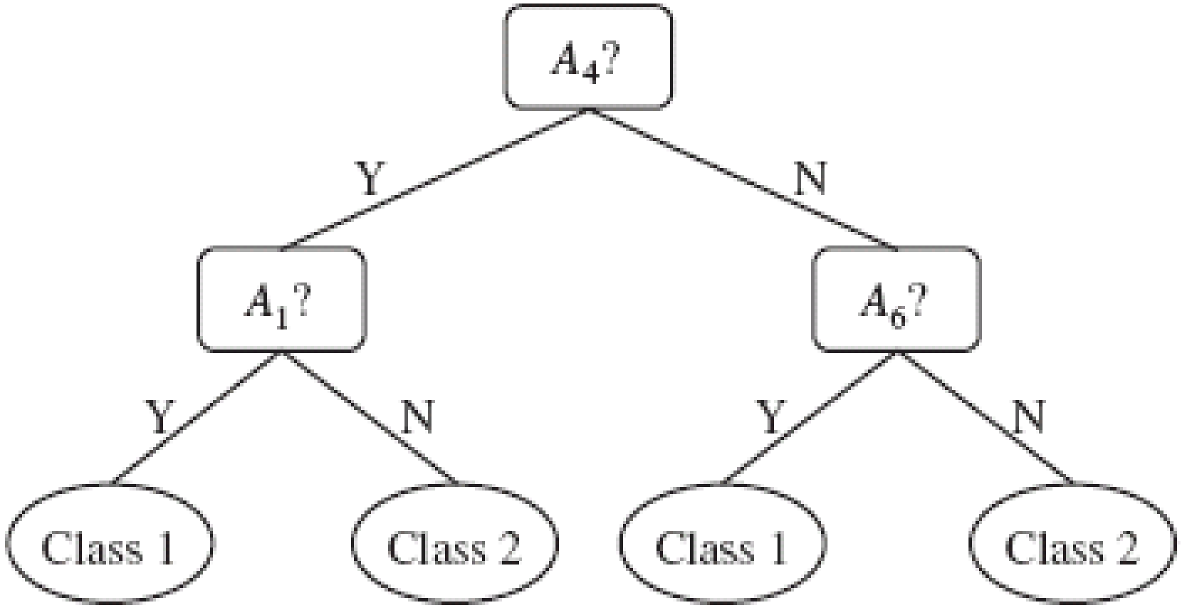
(A) Attribute Subset Selection (i.e. Feature selection)

- **How can we find a 'good' subset of the original attributes?**
  - **For *n* attributes, there are 2^n possible subsets.**
- **Heuristic methods (due to exponential # of choices):**
  - **Best single features under the feature independence assumption: choose by significance tests**
  - **Step-wise forward selection**
    - **The best single-feature is picked first**
    - **Then next best feature condition to the first, ...**
  - **Step-wise backward elimination**
    - **Repeatedly eliminate the worst feature**
  - **Best combined forward selection and backward elimination**
  - **Optimal branch and bound:**
    - **Use attribute elimination and backtracking**

Feature Selection Functions

Function	Denoted by	Mathematical form
DIA association factor	$z(t_k, c_i)$	$P(c_i \mid t_k)$
Information gain	$IG(t_k, c_i)$	$\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \cdot \log \frac{P(t, c)}{P(t) \cdot P(c)}$
Mutual information	$MI(t_k, c_i)$	$\log \frac{P(t_k, c_i)}{P(t_k) \cdot P(c_i)}$
Chi-square	$\chi^2(t_k, c_i)$	$\frac{\lvert Tr \rvert \cdot [P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]^2}{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}$
NGL coefficient	$NGL(t_k, c_i)$	$\frac{\sqrt{\lvert Tr \rvert} \cdot [P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}}$
Relevancy score	$RS(t_k, c_i)$	$\log \frac{P(t_k \mid c_i) + d}{P(\bar{t}_k \mid \bar{c}_i) + d}$
Odds ratio	$OR(t_k, c_i)$	$\frac{P(t_k \mid c_i) \cdot (1 - P(t_k \mid \bar{c}_i))}{(1 - P(t_k \mid c_i)) \cdot P(t_k \mid \bar{c}_i)}$
GSS coefficient	$GSS(t_k, c_i)$	$P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i) \cdot P(\bar{t}_k, c_i)$

Heuristic methods for feature selection

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

Attribute Creation (Feature Generation)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space (see: data reduction)
    - E.g., Fourier transformation, wavelet transformation, manifold approaches (not covered)
  - Attribute construction
    - Combining features (see: discriminative frequent patterns in Chapter on "Advanced Classification")
    - Data discretization

Data Reduction 2: 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 discard the data (except possible outliers)
  - Example: 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

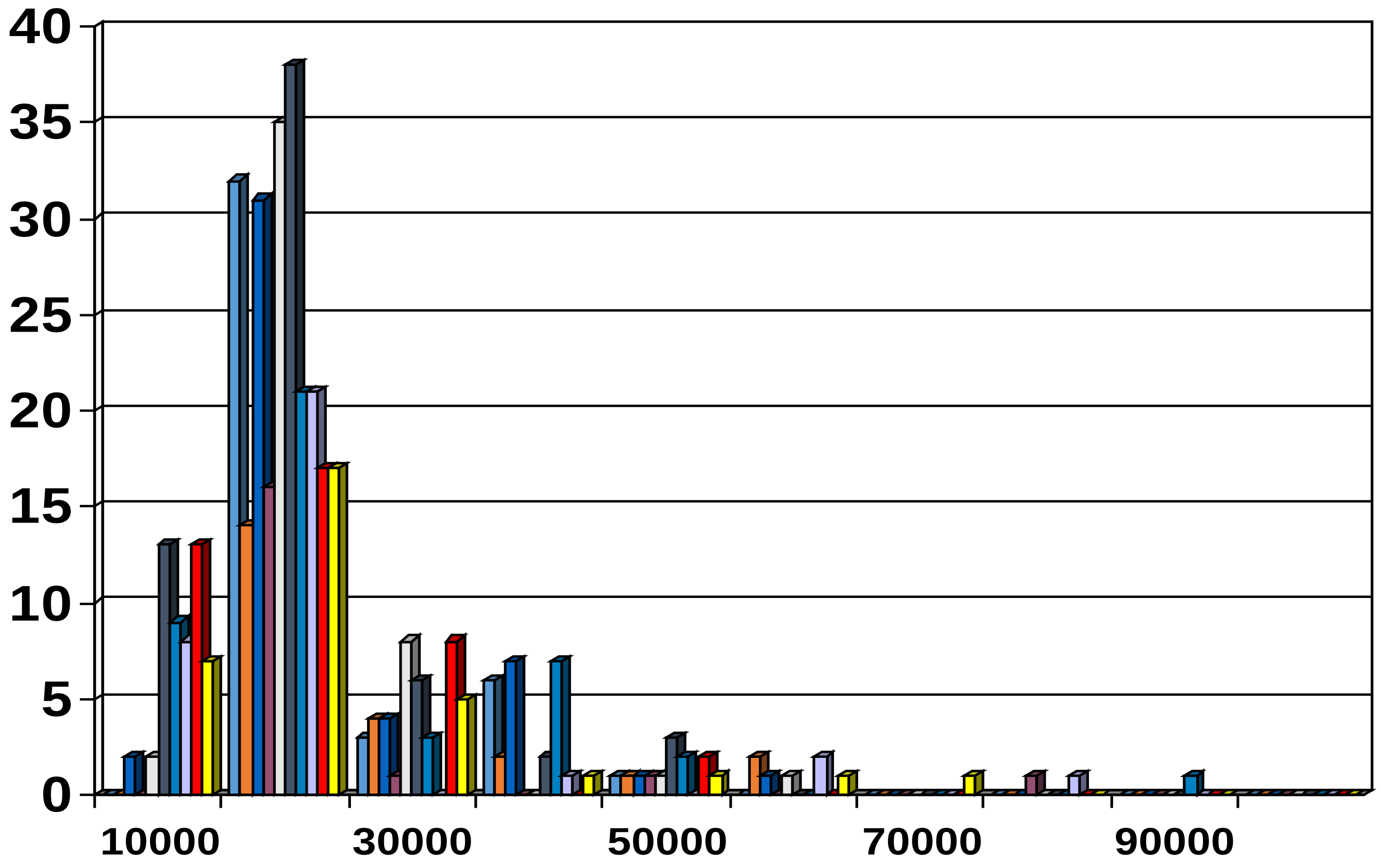
Regression and Log-Linear Models

- **Linear regression:** $Y = w\,X + b$ (*w*: slope, *b*: Y-intercept)
  - Data modeled to fit a straight line
  - Two regression coefficients, w and b, specify the parameters of model
  - Often uses the least squares criterion to fit the line based on the known values of

  $Y_1, Y_2, \ldots, X_1, X_2, \ldots.$
- **Multiple regression:** $Y = b_0 + b_1 X_1 + b_2 X_2$.
  - allows a response variable Y to be modeled as a linear function of multidimensional feature vector
  - Many nonlinear functions can be transformed into the above
- **Log-linear model:**
  - approximates discrete multidimensional probability distributions
  - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
  - Useful for dimensionality reduction and data smoothing

Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket



- **Partitioning rules:**
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)

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 later

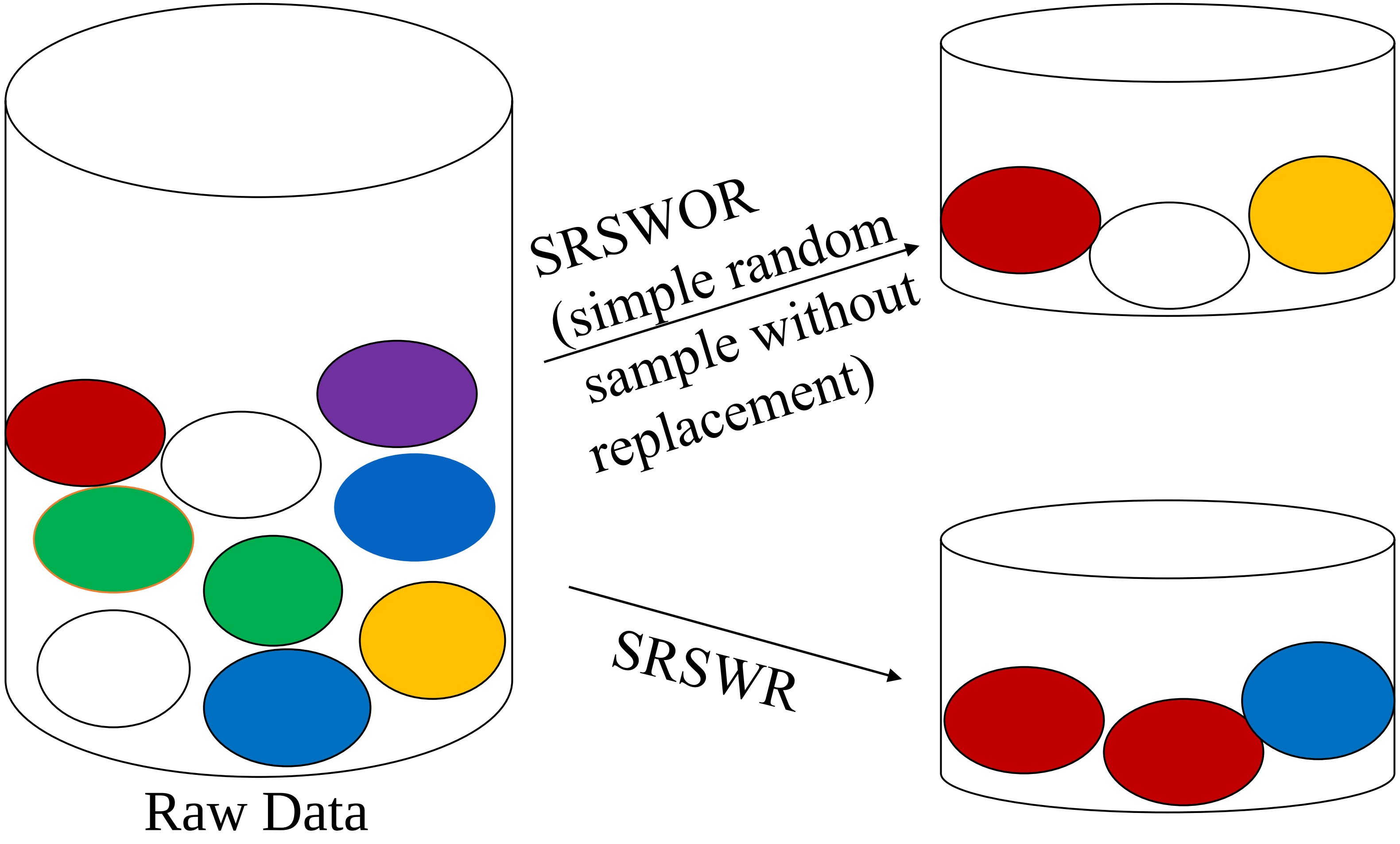
Sampling

- Sampling: obtaining a small sample s to represent the whole dataset 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)

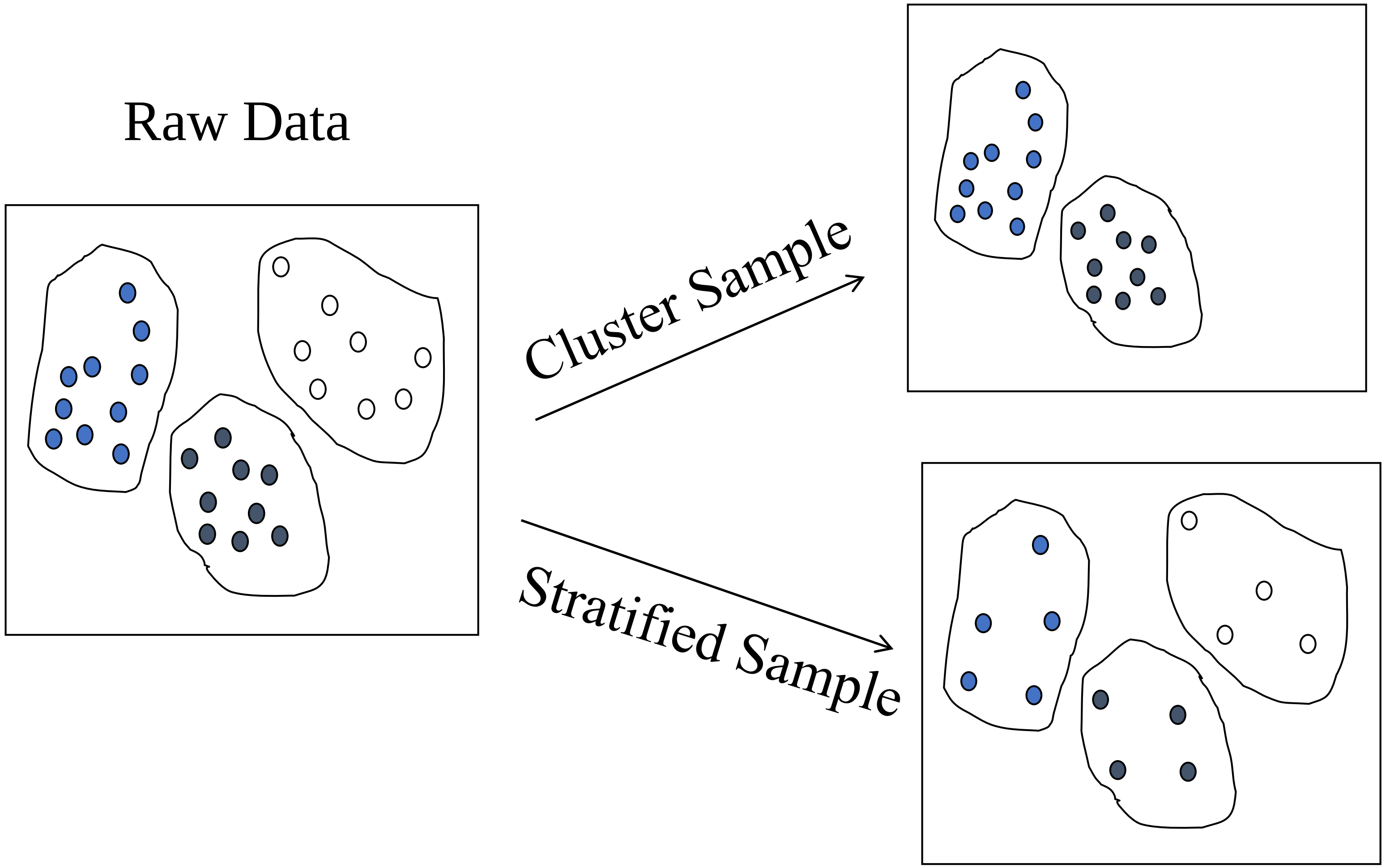
Types of Sampling

- **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
- **Cluster sampling**
  - Group data into clusters and draw samples from each cluster
- **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

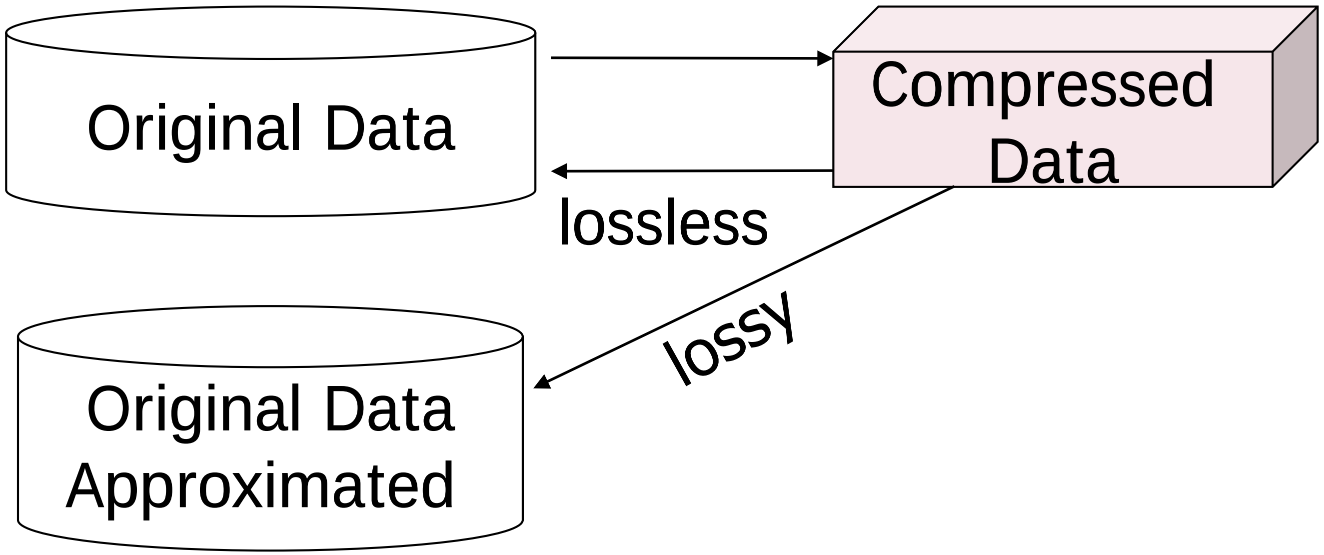


Sampling: Cluster or Stratified Sampling



Data Reduction 3: Data Compression

- Data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data.
  - **Lossless:** If the original data can be reconstructed from the compressed data without any loss of information.
  - **Lossy:** we can reconstruct only an approximation of the original data.



Data Reduction 3: 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**
- **Time sequence is not audio**
  - **Typically short and vary slowly with time**
- **Dimensionality and numerosity reduction may also be considered as forms of data compression**

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values.
- **Smoothing:** Remove noise from data
- **Attribute/feature construction**
  - New attributes constructed from the given ones
- **Aggregation:** Summarization, data cube construction
- **Normalization:** scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Discretization:** concept hierarchy climbing

Normalization

- **Min-max normalization: to [new_min$_A$, new_max$_A$]**

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - **Ex. Let *income* range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to**

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- **Z-score normalization (zero-mean)** $v' = \frac{v - \mu_A}{\sigma_A}$

(μ: mean, σ: standard deviation):

  - **Ex. Let μ = 54,000, σ = 16,000. Then** $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j}$$

Where *j* is the smallest integer such that Max(|ν'|) < 1

Discretization

- **Three types of attributes**
  - Nominal—values from an unordered set, e.g., color, profession
  - Ordinal—values from an ordered set, e.g., military or academic rank
  - Numeric—real numbers, e.g., integer or real numbers
- **Discretization: Divide the range of a continuous attribute into intervals**
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Data Discretization Methods

- **Typical methods: All the methods can be applied recursively**
  - **Binning**
    - **Top-down split, unsupervised**
  - **Histogram analysis**
    - **Top-down split, unsupervised**
  - **Clustering analysis (unsupervised, top-down split or bottom-up merge)**
  - **Decision-tree analysis (supervised, top-down split)**
  - **Correlation (e.g., χ2) analysis (unsupervised, bottom-up merge)**

Simple Discretization Methods: Binning

Equal-width (distance) partitioning

- Divides the range into N intervals of equal size: uniform grid
- if A and B are the lowest and highest values of the attribute, the width of intervals will be:

  $W = (B - A)/N$.
- The most straightforward, but outliers may dominate presentation
- Skewed data is not handled well

Equal-depth (frequency) partitioning

- Divides the range into N intervals, each containing approximately same number of samples (equal number)
- Good data scaling
- Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

➤ Step1: Sort & Partition

* Sorted data for *price* (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into equal-frequency (equal-depth) bins:
  4 data per bin

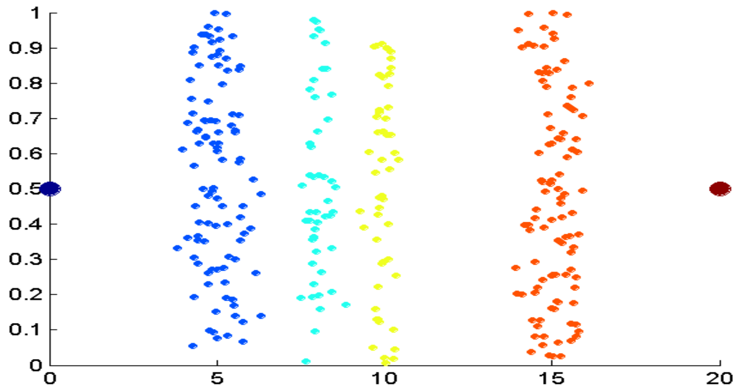
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

➤ Step 2: Smooth

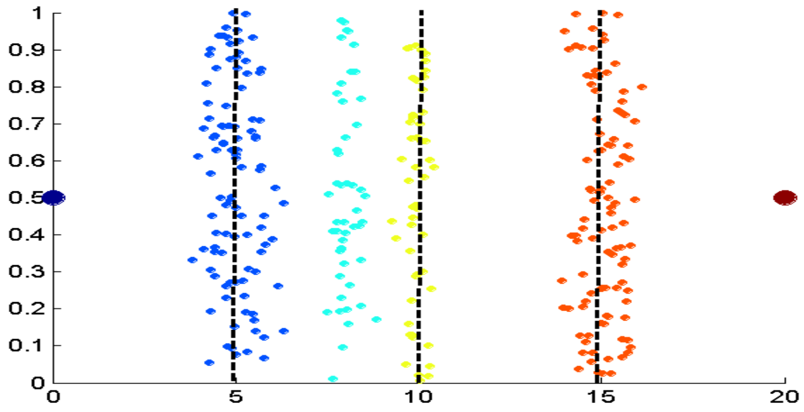
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Discretization Without Using Class Labels

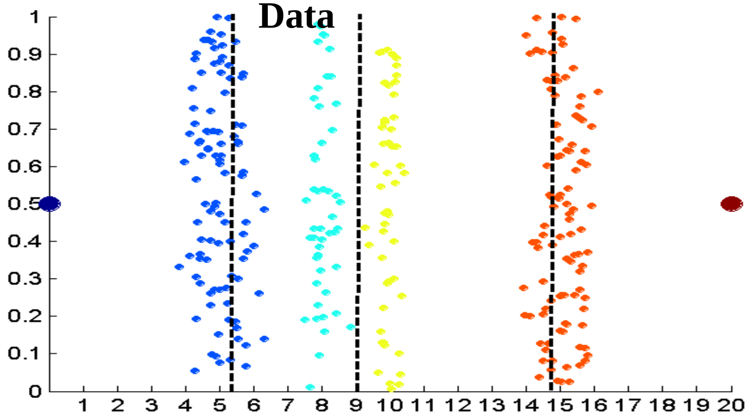
➢(Binning vs. Clustering)



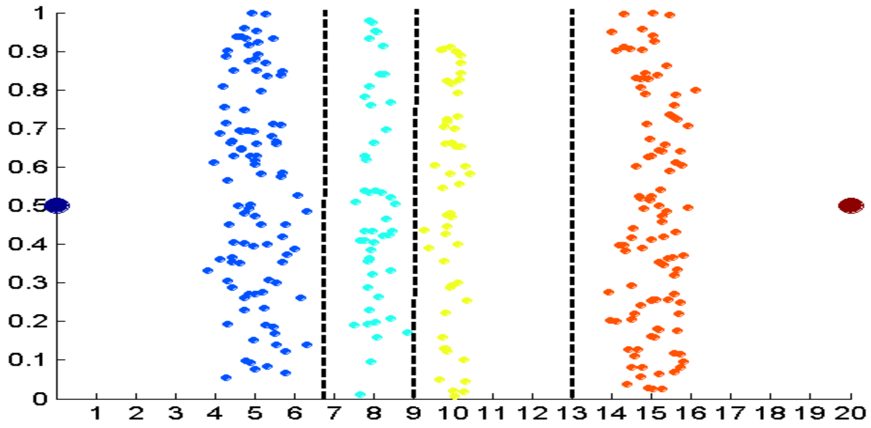
Data



Equal width (binning)



Equal frequency (binning)



K-means clustering leads to better results

Concept Hierarchy Generation

- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate **drilling and rolling** in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth, adult, or senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data—For numeric data, use discretization methods shown

Concept Hierarchy Generation for Nominal Data

- **Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts**
  - ***street* < *city* < *state* < *country***
- **Specification of a hierarchy for a set of values by explicit data grouping**
  - **{Urbana, Champaign, Chicago} < Illinois**
- **Specification of only a partial set of attributes**
  - **E.g., only *street* < *city*, not others**
- **Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values**
  - **E.g., for a set of attributes: *{street, city, state, country}***

Automatic Concept Hierarchy Generation

- **Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set**
  - **The attribute with the most distinct values is placed at the lowest level of the hierarchy**
  - **Exceptions, e.g., *weekday, month, quarter, year***