Khai thác dữ liệu & Khai phá tri thức

Data Mining & Knowledge Discovery

Bài 3. **Tiền xử lý dữ liệu**

Mã MH: **505043**

TS. HOÀNG Anh

Nội dung

- Tiền xử lý dữ liệu: Tổng quan
  - Chất lượng dữ liệu/ Data Quality
  - Các nhiệm vụ cơ bản
- **Làm sạch dữ liệu/ Data Cleaning**
- **Tích hợp dữ liệu/ Data Integration**
- Giảm dữ liệu/ Data Reduction
- Chuyển đổi dữ liệu/ Data Transformation
- Tổng kết

Tại sao phải tiền xử lý dữ liệu?

- Đo lường chất lượng dữ liệu:
  - Độ chính xác/ Accuracy: correct or wrong, accurate or not
  - Độ hoàn thiện/ Completeness: not recorded, unavailable, …
  - Tính nhất quán/ Consistency: some modified but some not, dangling, …
  - Tính kịp thời/ Timeliness: timely update?
  - Độ tin cậy/ Believability: how trustable the data are correct?
  - Khả năng diễn giải/ Interpretability: how easily the data can be understood?

Nhiệm vụ cơ bản

- **Làm sạch dữ liệu/ Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Tích hợp dữ liệu/ Data integration**
  - Integration of multiple databases, data cubes, or files
- **Giảm dữ liệu/ Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Chuyển đổi dữ liệu/ Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

1. Làm sạch dữ liệu

- Dữ liệu thực tế thường "bẩn": Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - Chưa hoàn thiện/ incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=" " (missing data)
  - Nhiễu/ noisy: containing noise, errors, or outliers
    - e.g., *Salary*="−10" (an error)
  - Không nhất quán/ 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"
    - discrepancy between duplicate records
  - Cố ý/ Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

1.1 Dữ liệu chưa hoàn thiện

- Dữ liệu không có sẵn
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Lý do dữ liệu chưa hoàn thiện
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Có thể "suy luận" được dữ liệu bị thiếu/ missing data

1.1 Cách xử lý dữ liệu bị thiếu?

- Bỏ qua/ Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Điền vào thủ công: tedious + infeasible?
- Điền vào tự động:
  - 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

1.2 Dữ liệu nhiễu

- Nhiễu/ Noise: random error or variance in a measured variable
- Lý do dữ liệu bị nhiễu
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Các vấn đề khác
  - duplicate records
  - incomplete data
  - inconsistent data

1.2 Cách xử lý dữ liệu nhiễu?

- Phân hoạch/ 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.
- Hồi qui/ Regression
  - smooth by fitting the data into regression functions
- Gom cụm/ Clustering
  - detect and remove outliers
- Kết hợp kiểm tra máy tính và con người
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Làm sạch dữ liệu là một quá trình

- Phát hiện sai lệch dữ liệu/ 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)
- Di chuyển và tích hợp dữ liệu/ 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
- Tích hợp hai quá trình/ Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

2. Tích hợp dữ liệu

- **Tích hợp dữ liệu/ Data integration**:
  - Combines data from multiple sources into a coherent store
- **Lược đồ tích hợp**/ Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Vấn đề nhận dạng thực thể/ Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Phát hiện và giải quyết các giá trị dữ liệu xung đột
  - 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

Xử lý dữ liệu trùng lặp trong quá trình tích hợp

- Dữ liệu xảy ra trùng lặp khi được tích hợp từ nhiều nguồn
  - *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
- Tham số trùng lặp có thể được phát hiện bằng phân tích tương quan/ *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

2.1 Phân tích tương quan (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
- Tương quan không bao hàm quan hệ nhân quả
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-Square Calculation: Ví dụ

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

- 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

2.1 Phân tích tương quan (Numeric Data)

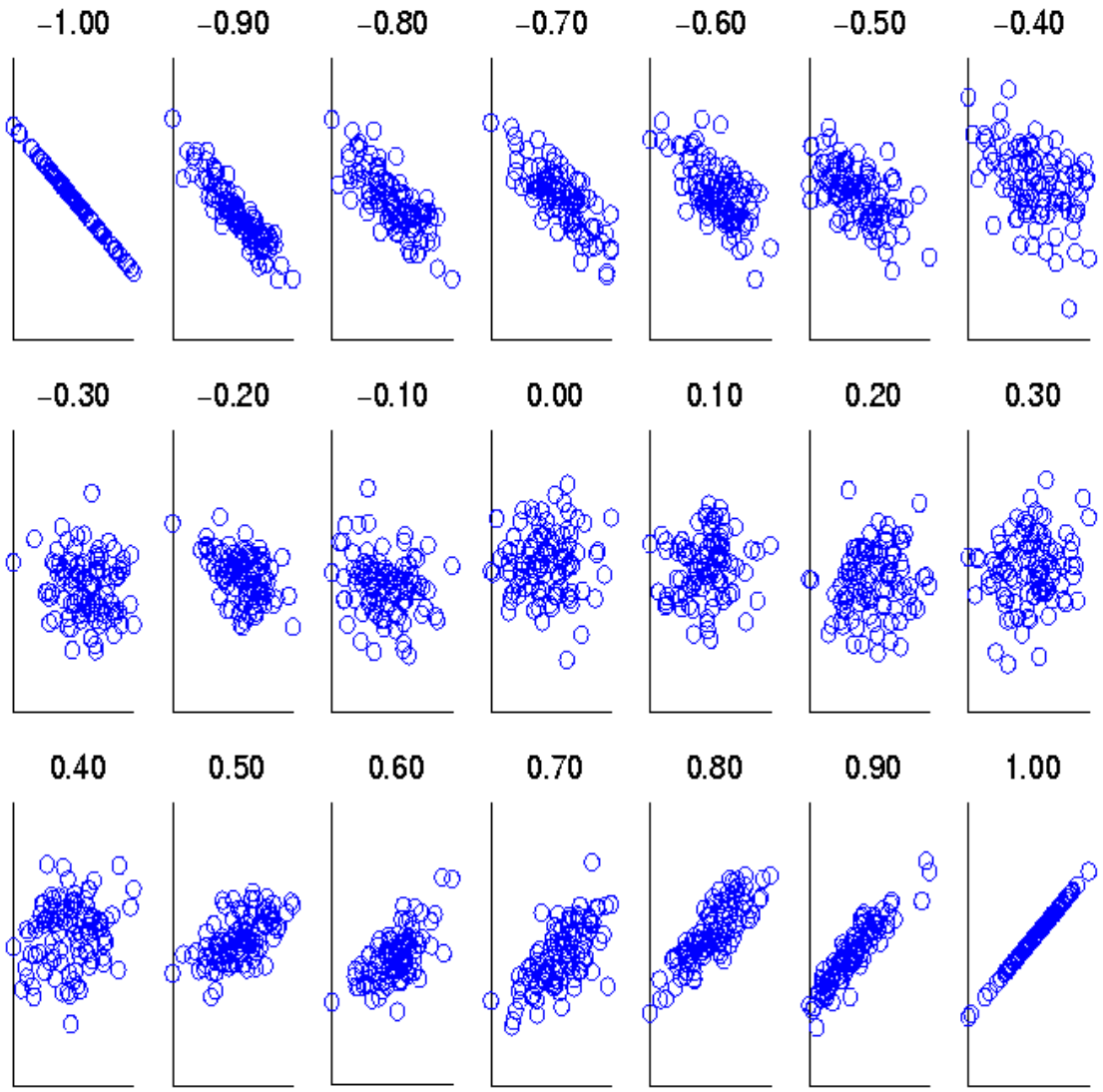
- Hệ số tương quan (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- 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; $r_{AB} < 0$: negatively correlated

Đánh giá trực quan mối tương quan



Scatter plots showing the similarity from –1 to 1.

Tương quan (viewed as linear relationship)

- Tương quan đo lường mối quan hệ tuyến tính giữa các đối tượng.
- Để tính toán tương quan: 1) chuẩn hóa các đối tượng dữ liệu, A và B; 2) tính tích vô hướng

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

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

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

2.2 Hiệp phương sai/ covariance (Numeric Data)

- **Hiệp phương sai tương tự như tương quan**

$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

Hệ số tương quan $r_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance**: If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance**: If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence**: $Cov_{A,B} = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Co-Variance: Ví dụ

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 − 4 × 9.6 = 4
- Thus, A and B rise together since Cov(A, B) > 0.

3. Chiến lược "giảm" dữ liệu

- **Giảm dữ liệu/ Data reduction**: Biểu diễn rút gọn của tập dữ liệu gốc, nhưng vẫn đảm bảo kết quả phân tích tương đồng.

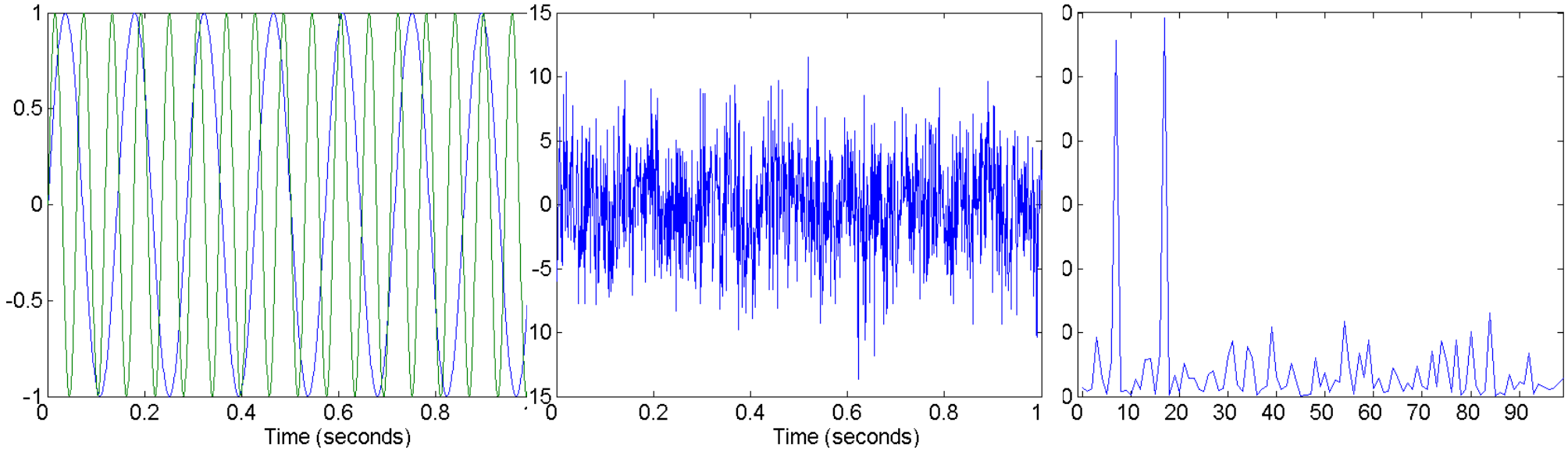
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- **Chiến lược "giảm" dữ liệu**
  - Giảm chiều/ Dimensionality reduction, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - Giảm số lượng/ Numerosity reduction (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Nén dữ liệu/ Data compression

3.1 Giảm chiều dữ liệu

- **Chiều dữ liệu**
  - 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
- **Giảm chiều dữ liệu**
  - Avoid the curse of dimensionality
  - Help eliminate irrelevant features and reduce noise
  - Reduce time and space required in data mining
  - Allow easier visualization
- **Kỹ thuật giảm chiều dữ liệu**
  - Wavelet transforms
  - Principal Component Analysis/ PCA
  - Supervised and nonlinear techniques (e.g., feature selection)

Ánh xạ dữ liệu sang không gian mới

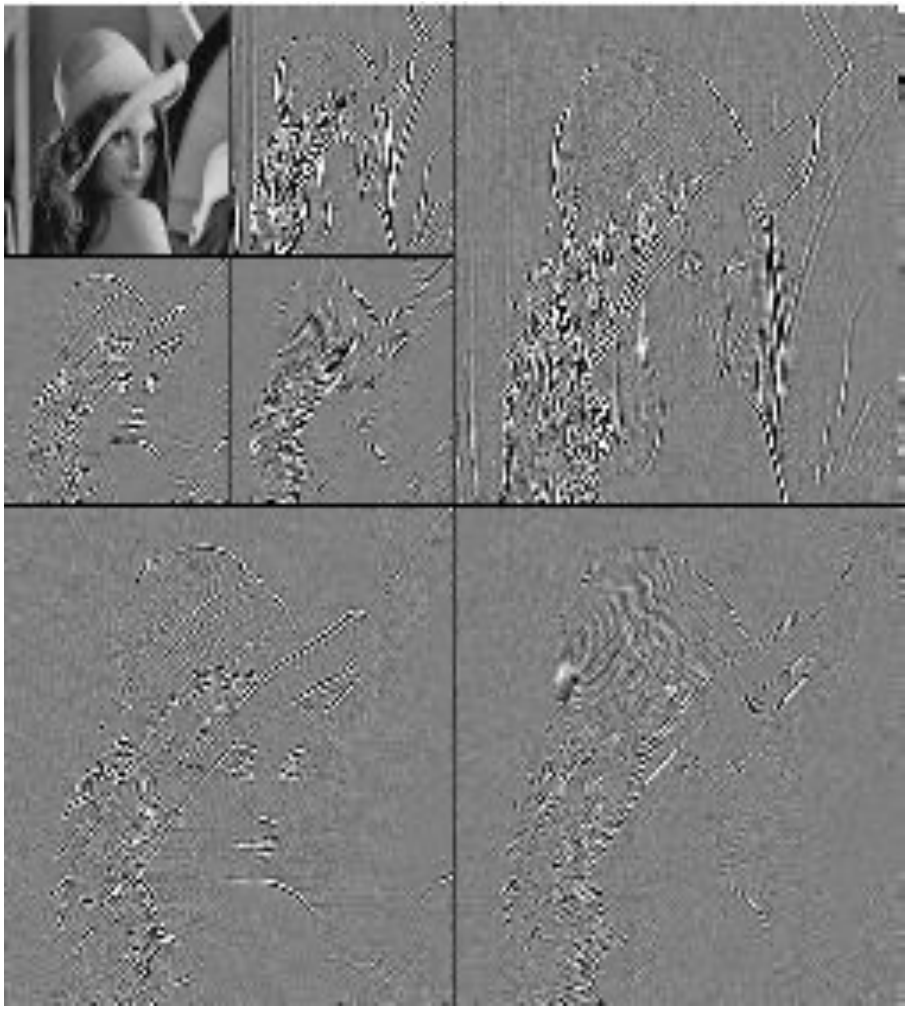
- **Fourier transform**
- **Wavelet transform**



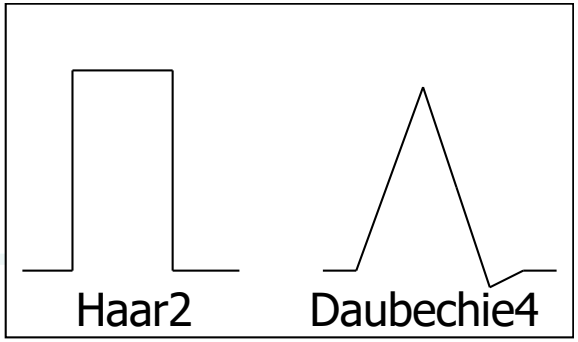
Two Sine Waves **Two Sine Waves + Noise** **Frequency**

Biến đổi Wavelet là gì?

- Phân tách dữ liệu thành các dãi băng tần con khác nhau
  - Applicable to n-dimensional signals
- Dữ liệu được chuyển đổi, duy trì khoảng cách giữa các đối tượng, ở các mức độ phân giải khác nhau
- Cho phép các cụm/nhóm dễ phân biệt hơn
- Được sử dụng để nén ảnh



Biến đổi Wavelet



- **Discrete wavelet transform** (DWT) for linear signal processing, multi-resolution 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
- Phương pháp:
  - ***Length, L, must be an integer power of 2 (padding with 0's, when necessary)***
  - ***Each transform has 2 functions: smoothing, difference***
  - ***Applies to pairs of data, resulting in two set of data of length L/2***
  - ***Applies two functions recursively, until reaches the desired length***

Wavelet Decomposition

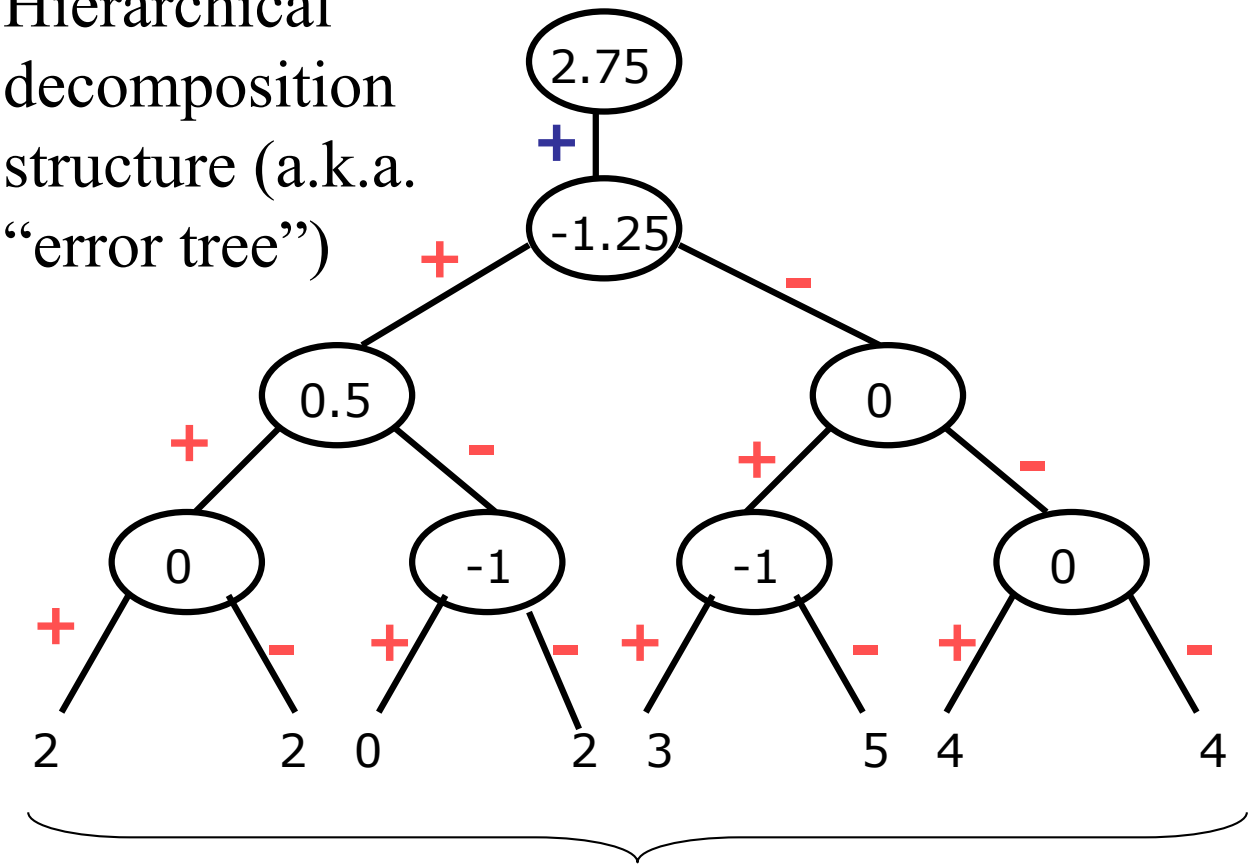
- Wavelets: A math tool for space-efficient hierarchical decomposition of functions
- **S = [2, 2, 0, 2, 3, 5, 4, 4]** can be transformed to

  $S_{\wedge}$ = [$2^{3}/_{4}$, $-1^{1}/_{4}$, $^{1}/_{2}$, 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	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$

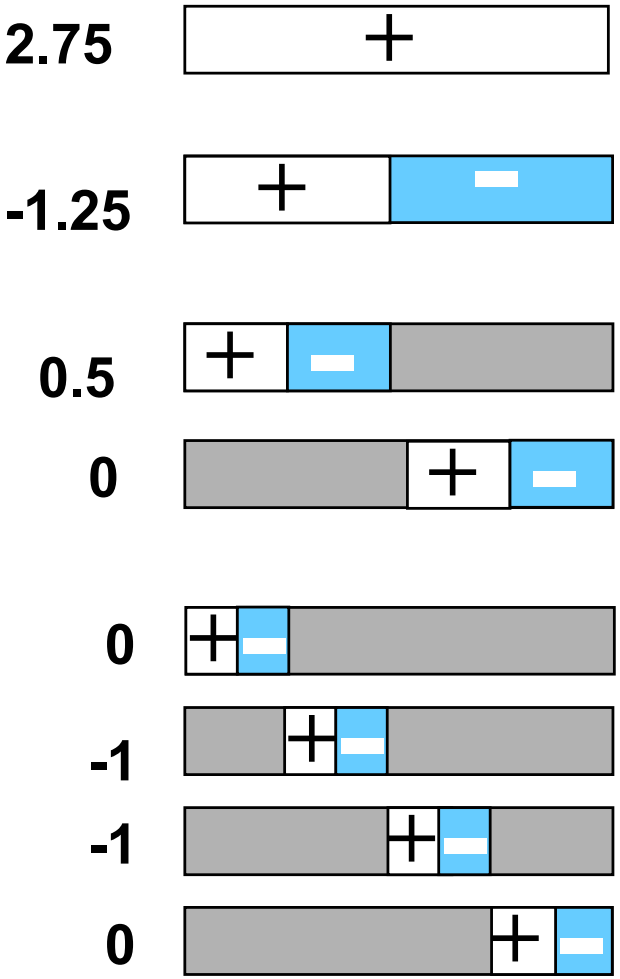
Hệ số Haar Wavelet

Hierarchical decomposition structure (a.k.a. "error tree")



Original frequency distribution

Coefficient "Supports"

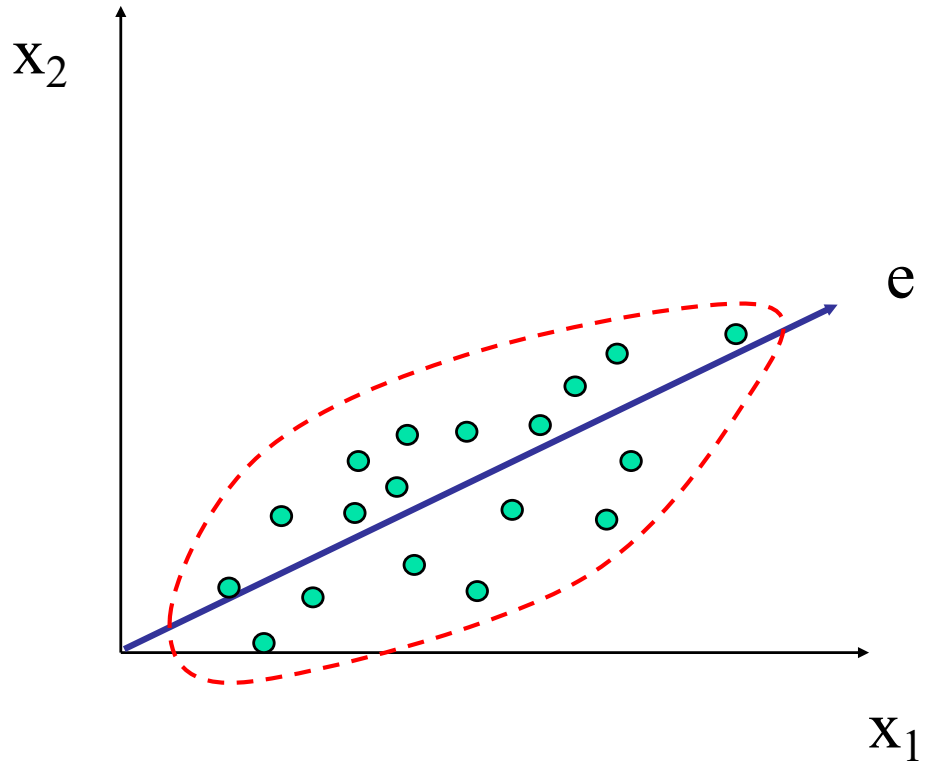


Lý do cần biến đổi Wavelet?

- Use hat-shape filters
  - Emphasize region where points cluster
  - Suppress weaker information in their boundaries
- Effective removal of outliers
  - Insensitive to noise, insensitive to input order
- Multi-resolution
  - Detect arbitrary shaped clusters at different scales
- Efficient
  - Complexity O(N)
- Only applicable to low dimensional data

Phân tích thành phần chính (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. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Phân tích thành phần chính (Các bước)

- Given N data vectors from n-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  - 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

Lựa chọn tập con các thuộc tính

- Another way to reduce dimensionality of data
- 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

Tìm kiếm theo kinh nghiệm trong lựa chọn thuộc tính

- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - Step-wise attribute elimination:
    - Repeatedly eliminate the worst attribute
  - Best combined attribute selection and elimination
  - Optimal branch and bound:
    - Use attribute elimination and backtracking

Tạo thuộc tính mới (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 7)
    - Data discretization

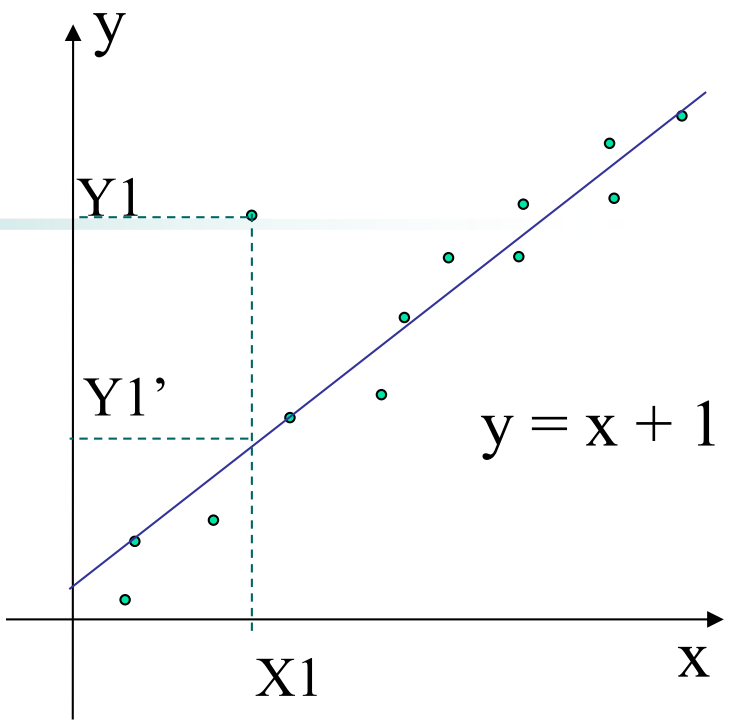
3.2 Giảm số lượng dữ liệu

- 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)
  - Ex.: 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, …

Parametric Data Reduction: Regression and Log-Linear Models

- **Linear regression**
  - Data 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
- **Log-linear model**
  - Approximates discrete multidimensional probability distributions

Phân tích hồi qui



- Regression analysis: A collective name for techniques for the 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* (aka. ***explanatory variables*** or ***predictors***)
- 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

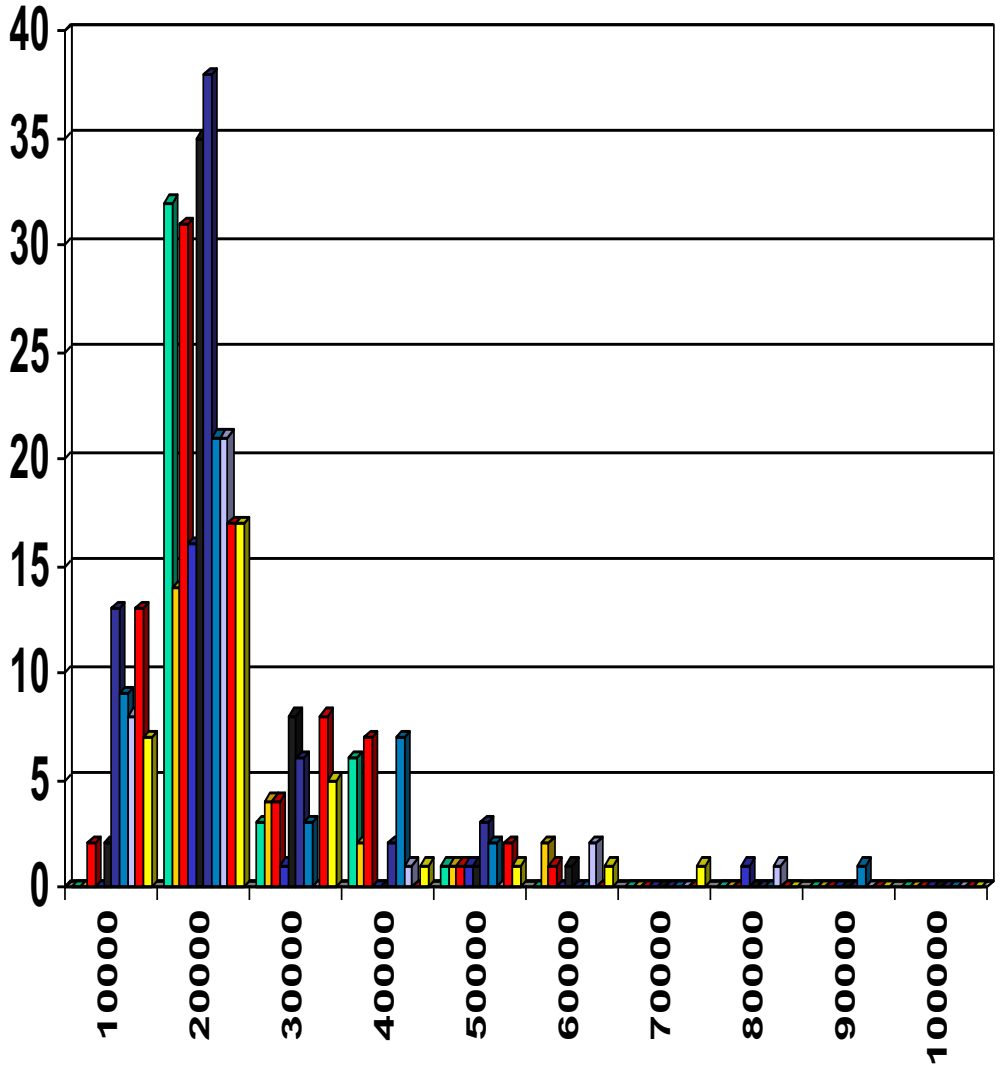
- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Regress Analysis 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: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - Approximate 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

Đồ thị Histogram

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



Gom cụm

- 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 Chapter 10

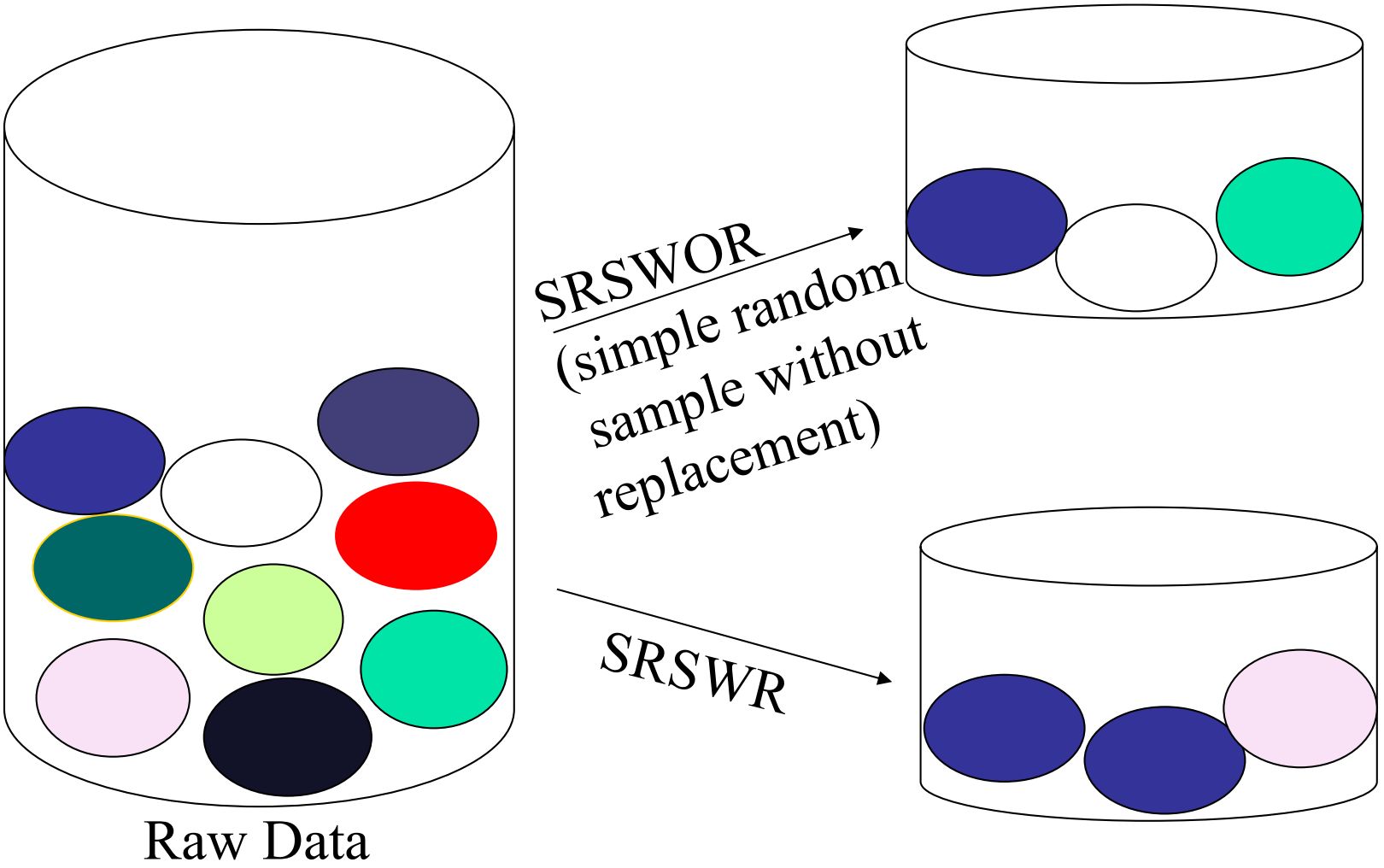
Lấy mẫu/ Sampling

- 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)

Các kiểu lấy mẫu

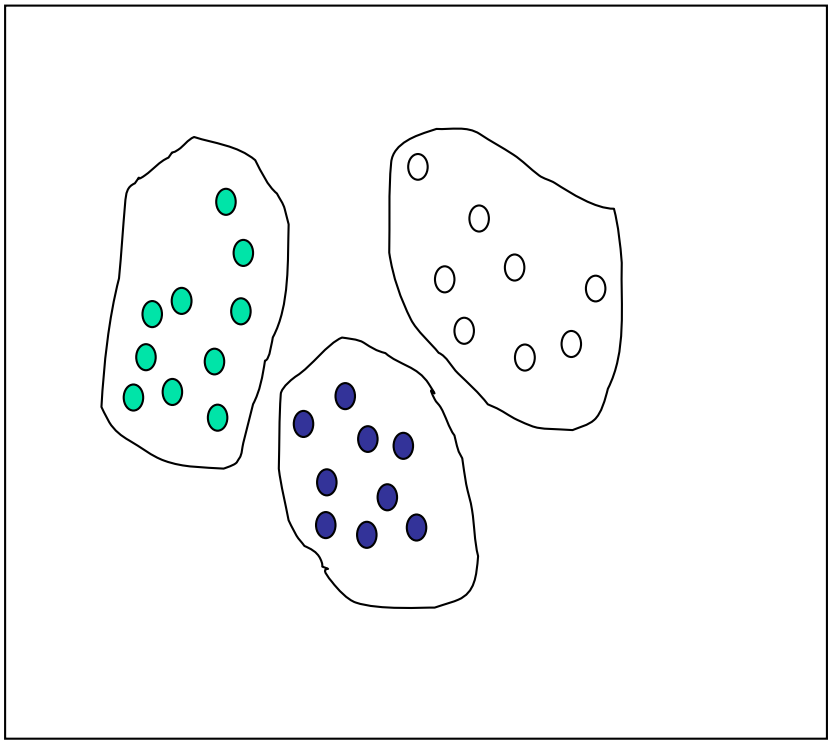
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified 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

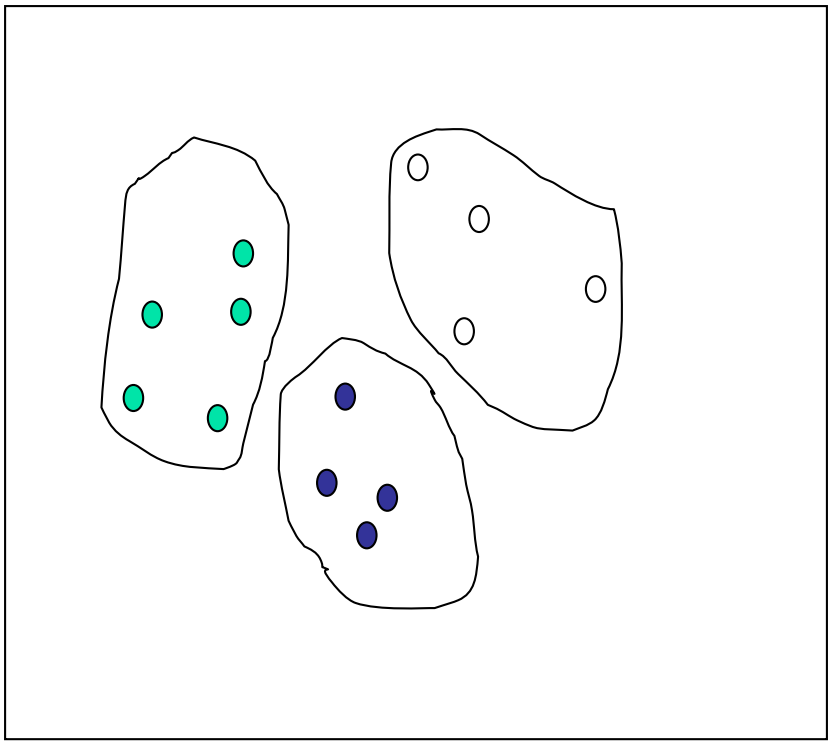


Lấy mẫu: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



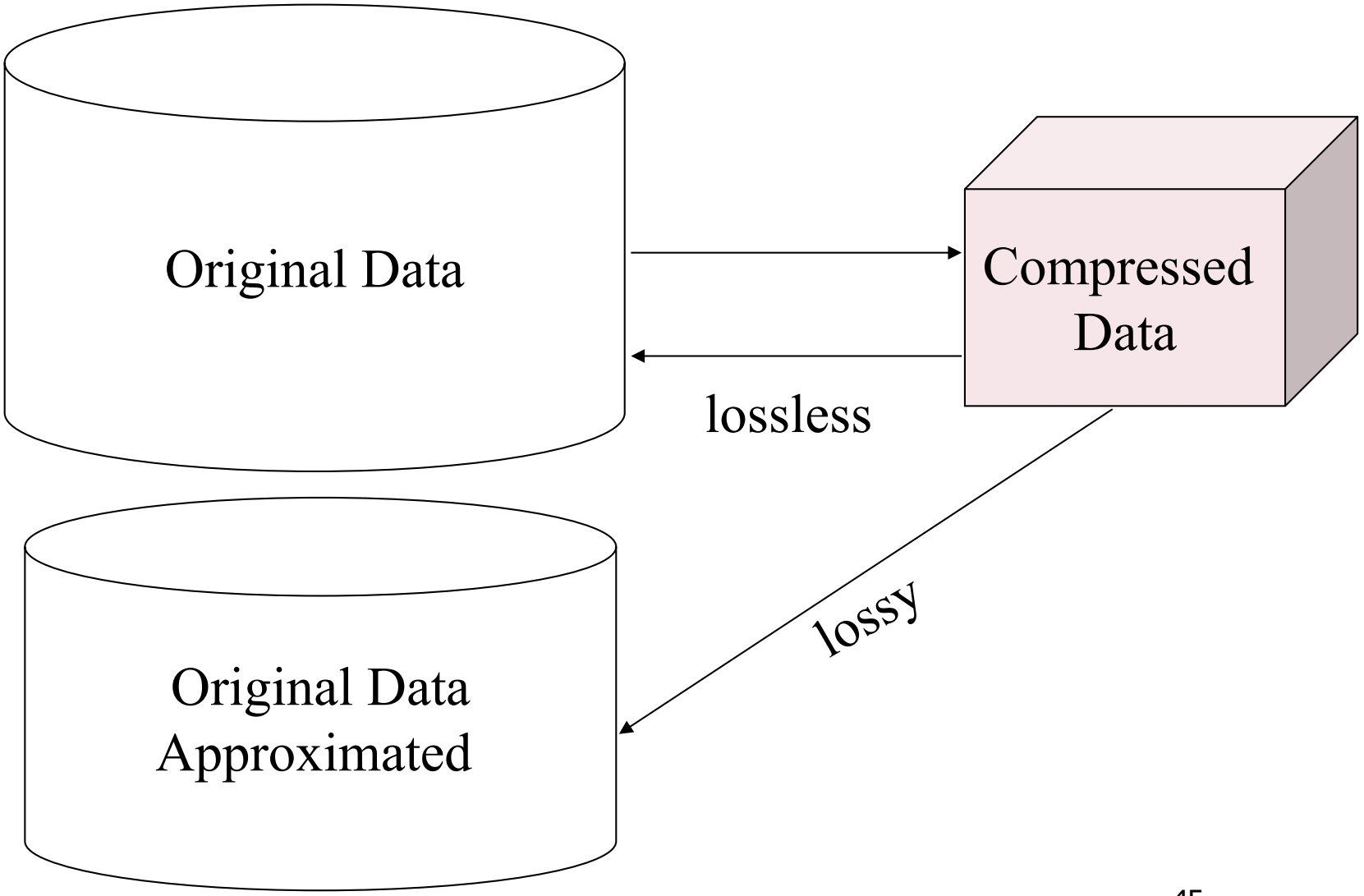
Tổng hợp dữ liệu khối

- 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
- 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

3.3 Nén dữ liệu

- 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

Nén dữ liệu/ Data Compression



4. Chuyển đổi dữ liệu

- Hàm ánh xạ toàn bộ tập giá trị của một thuộc tính đã cho (không gian củ) sang tập giá trị thay thế mới (không gian mới).
- Phương pháp/ Methods
  - Làm mịn/ Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Tổng hợp/ Aggregation: Summarization, data cube construction
  - Chuẩn hóa/ Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling
  - Rời rạc hóa/ Discretization: Concept hierarchy climbing

Chuẩn hóa/ 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** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - 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(|v'|) < 1

Rời rạc hóa

- 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

Phương pháp rời rạc hóa dữ liệu

- Tất cả các phương pháp có thể được áp dụng đệ qui
  - 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., χ^2) analysis (unsupervised, bottom-up merge)

Rời rạc đơn giản: Binning

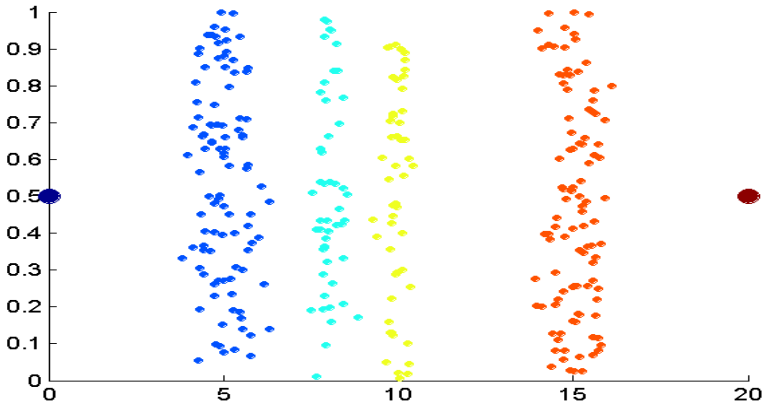
- Equal-width (distance) partitioning
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- Equal-depth (frequency) partitioning
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Phương pháp Binning làm mịn dữ liệu

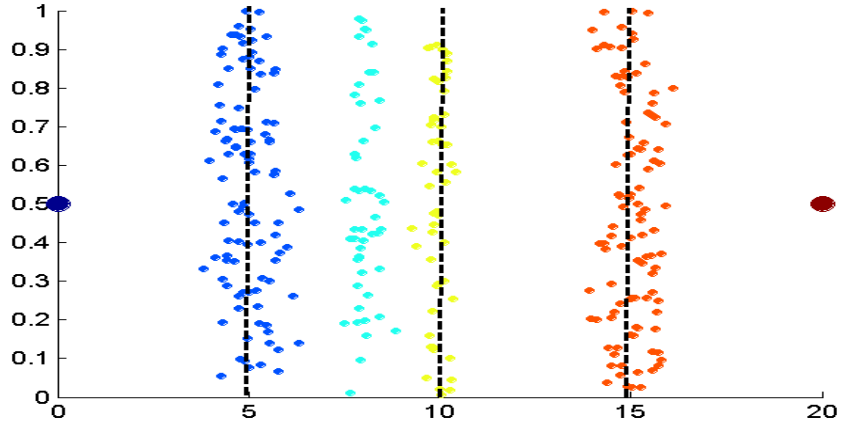
- Sắp xếp giá (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

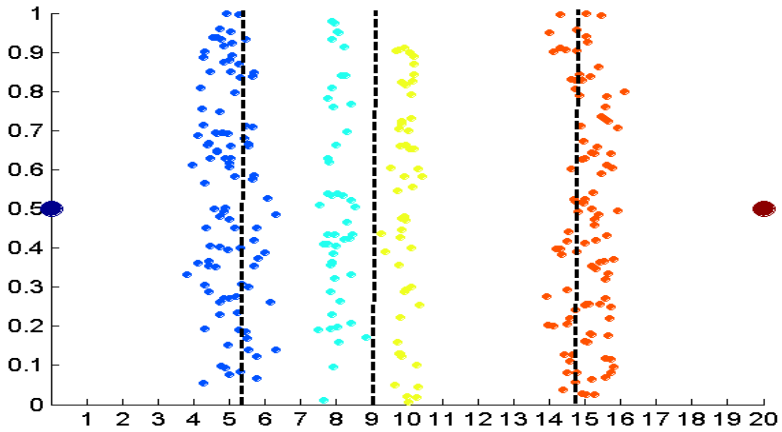
* Partition into equal-frequency (**equi-depth**) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by **bin means**:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by **bin boundaries**:
  - Bin 1: 4, 4, 4, 15
  - Bin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

Rời rạc không sử dụng nhãn lớp (Binning vs. Clustering)

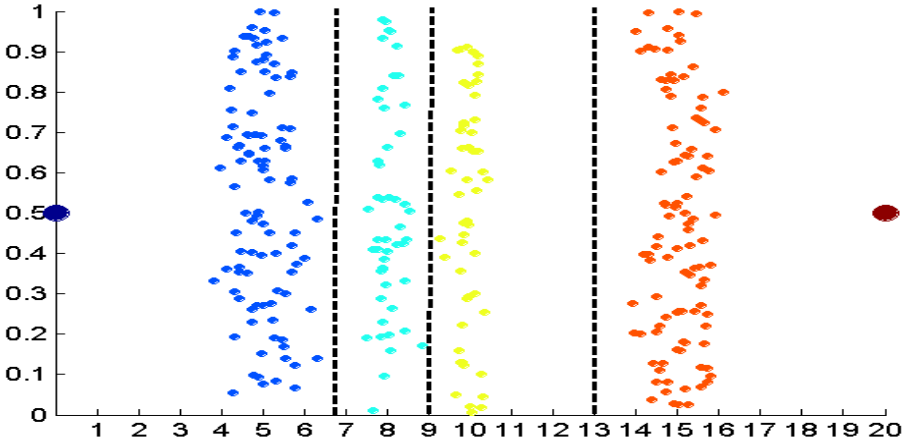


Data





Equal frequency (binning)



K-means clustering leads to better results

Rời rạc dựa theo phân loại và phân tích tương quan

- Phân loại (e.g., cây quyết định/ decision tree analysis)
  - Supervised: Given class labels, e.g., cancerous vs. benign
  - Using *entropy* to determine split point (discretization point)
  - Top-down, recursive split
  - Details to be covered in Chapter 7
- Phân tích tương quan (e.g., Chi-merge: χ^2-based discretization)
  - Supervised: use class information
  - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
  - Merge performed recursively, until a predefined stopping condition

Tạo phân cấp khái niệm

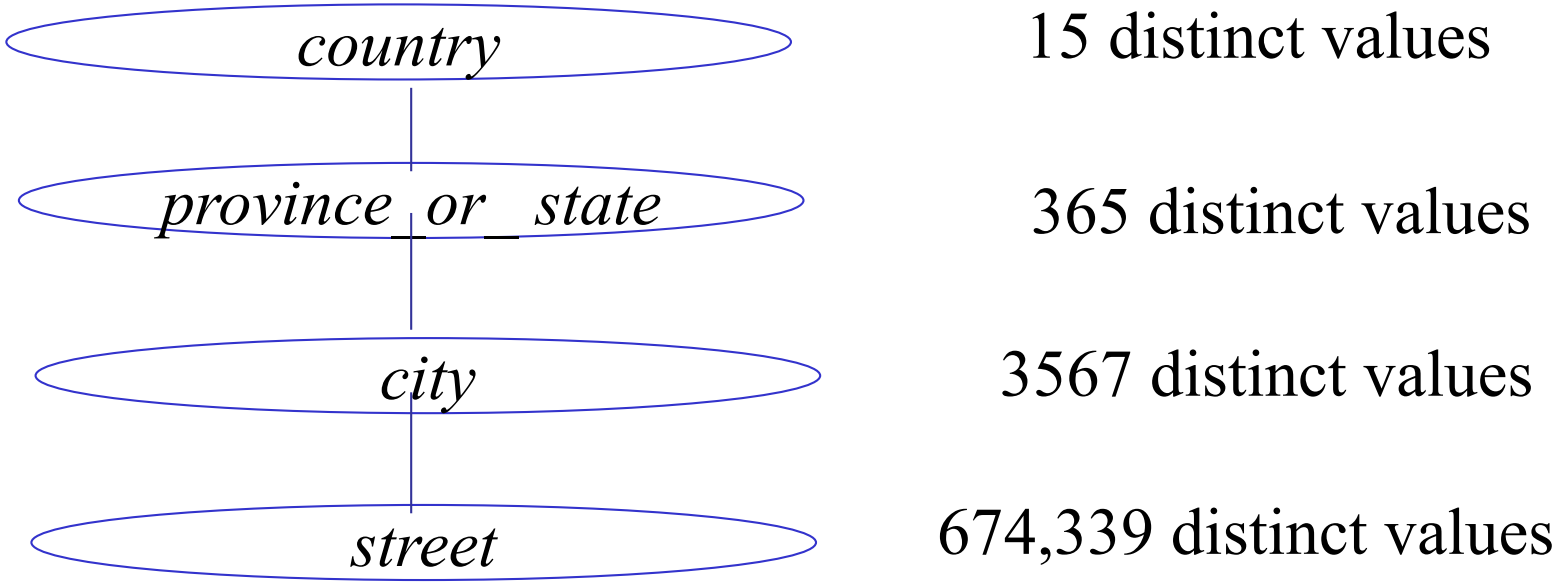
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively 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.

Tạo khái niệm phân cấp với dữ liệu định danh

- Đặc tả thứ tự một phần/toàn bộ các thuộc tính một cách rỏ ràng ở cấp lược đồ (scheme) bởi người dùng hoặc chuyên gia
  - *street < city < state < country*
- Đặc tả cấu trúc phân cấp cho một tập hợp các giá trị bằng cách nhóm dữ liệu rỏ ràng
  - {Urbana, Champaign, Chicago} < Illinois
- Đặc tả chỉ một phần đặc trưng
  - E.g., only *street < city*, not others
- Tự động tạo phân cấp dựa trên phân tích số giá trị khác nhau trên mỗi thuộc tính
  - E.g., for a set of attributes: {*street, city, state, country*}

Tự động tạo khái niệm phân cấp

- Một số phân cấp được tạo tự động dựa trên phân tích số các giá trị khác nhau trên mỗi đặc tính trong dữ liệu
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



Tổng kết

- **Chất lượng dữ liệu?** accuracy, completeness, consistency, timeliness, believability, interpretability
- **Làm sạch dữ liệu**: missing/noisy values, outliers
- **Tích hợp dữ liệu** từ nhiều nguồn khác nhau:
  - Entity identification problem
  - Remove redundancies
  - Detect inconsistencies
- **Giảm dữ liệu**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Chuyển đổi và rời rạc hóa dữ liệu**
  - Normalization
  - Concept hierarchy generation

Tham khảo/ References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- A. Bruce, D. Donoho, and H.-Y. Gao. Wavelet analysis. *IEEE Spectrum*, Oct 1996
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data*. Duxbury Press, 1997.
- H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. *VLDB'01*
- M. Hua and J. Pei. Cleaning disguised missing data: A heuristic approach. *KDD'07*
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), Dec. 1997
- H. Liu and H. Motoda (eds.). *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer Academic, 1998
- J. E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. *Data Quality: The Field Guide*. Digital Press (Elsevier), 2001
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995