

**Khoa Khoa Học & Kỹ Thuật Máy Tính  
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh**

## **Chương 2: Các vấn đề tiền xử lý dữ liệu**

---

Cao Học Ngành Khoa Học Máy Tính

Giáo trình điện tử

Biên soạn bởi: TS. Võ Thị Ngọc Châu

(chauvtn@cse.hcmut.edu.vn)

# Tài liệu tham khảo

---

- ▣ [1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, 2006.
- ▣ [2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.
- ▣ [3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.
- ▣ [4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.
- ▣ [5] Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar, "Next Generation of Data Mining", Taylor & Francis Group, LLC, 2009.
- ▣ [6] Daniel T. Larose, "Data mining methods and models", John Wiley & Sons, Inc, 2006.
- ▣ [7] Ian H. Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", Second Edition, Elsevier Inc, 2005.
- ▣ [8] Florent Messegia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.
- ▣ [9] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", Second Edition, Springer Science + Business Media, LLC 2005, 2010.

# Nội dung

---

- Chương 1: Tổng quan về khai phá dữ liệu
- **Chương 2: Các vấn đề tiền xử lý dữ liệu**
- Chương 3: Hồi qui dữ liệu
- Chương 4: Phân loại dữ liệu
- Chương 5: Gom cụm dữ liệu
- Chương 6: Luật kết hợp
- Chương 7: Khai phá dữ liệu và công nghệ cơ sở dữ liệu
- Chương 8: Ứng dụng khai phá dữ liệu
- Chương 9: Các đề tài nghiên cứu trong khai phá dữ liệu
- Chương 10: Ôn tập

# Chương 2: Các vấn đề tiền xử lý dữ liệu

---

- ▣ 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu
- ▣ 2.2. Tóm tắt mô tả về dữ liệu
- ▣ 2.3. Làm sạch dữ liệu
- ▣ 2.4. Tích hợp dữ liệu
- ▣ 2.5. Biến đổi dữ liệu
- ▣ 2.6. Thu giảm dữ liệu
- ▣ 2.7. Rời rạc hóa dữ liệu
- ▣ 2.8. Tạo cây phân cấp ý niệm
- ▣ 2.9. Tóm tắt

## 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

---

### □ Giai đoạn tiền xử lý dữ liệu

- Quá trình xử lý dữ liệu thô/gốc (raw/original data) nhằm cải thiện chất lượng dữ liệu (quality of the data) và do đó, cải thiện chất lượng của kết quả khai phá.

#### □ Dữ liệu thô/gốc

- Có cấu trúc, bán cấu trúc, phi cấu trúc
- Được đưa vào từ các nguồn dữ liệu trong các hệ thống xử lý tập tin (file processing systems) và/hay các hệ thống cơ sở dữ liệu (database systems)

- Chất lượng dữ liệu (data quality): tính chính xác, tính hiện hành, tính toàn vẹn, tính nhất quán

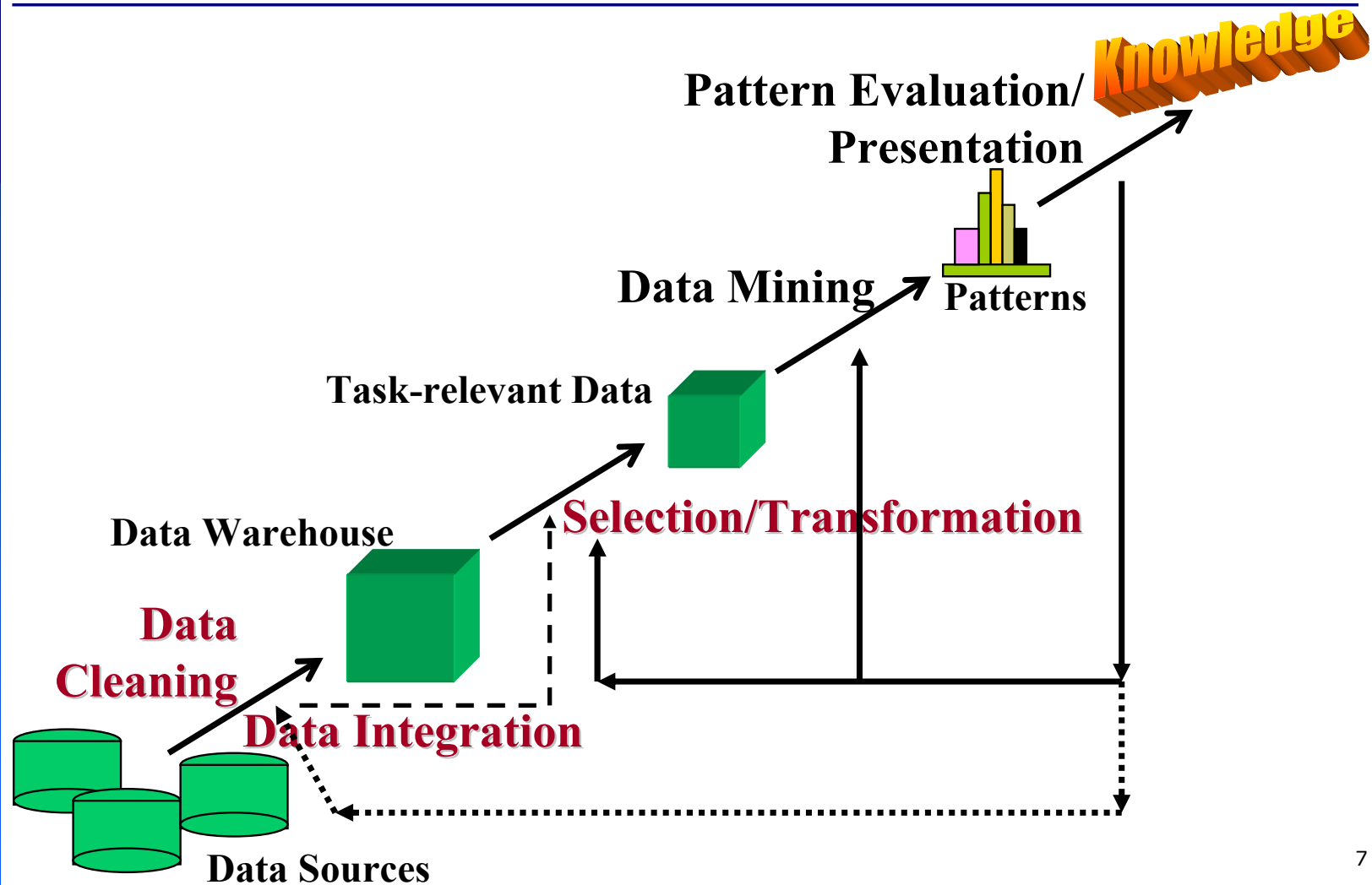
## 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

---

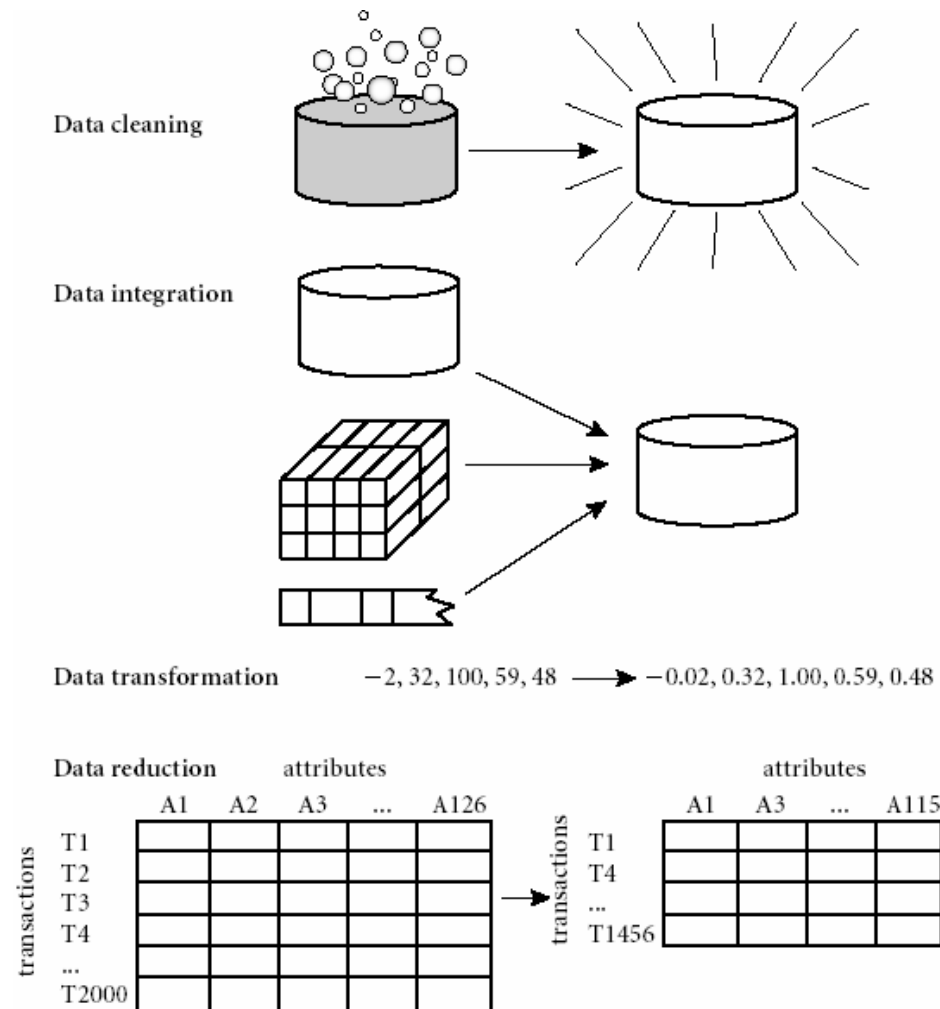
### ▣ Chất lượng dữ liệu (data quality)

- tính chính xác (accuracy): giá trị được ghi nhận đúng với giá trị thực.
- tính hiện hành (currency/timeliness): giá trị được ghi nhận không bị lỗi thời.
- tính toàn vẹn (completeness): tất cả các giá trị dành cho một biến/thuộc tính đều được ghi nhận.
- tính nhất quán (consistency): tất cả giá trị dữ liệu đều được biểu diễn như nhau trong tất cả các trường hợp.

## 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu



## 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu





## 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

---

### ▣ Các kỹ thuật tiền xử lý dữ liệu

- Làm sạch dữ liệu (data cleaning/cleansing): loại bỏ nhiễu (remove noise), hiệu chỉnh những phần dữ liệu không nhất quán (correct data inconsistencies)
- Tích hợp dữ liệu (data integration): trộn dữ liệu (merge data) từ nhiều nguồn khác nhau vào một kho dữ liệu
- Biến đổi dữ liệu (data transformation): chuẩn hoá dữ liệu (data normalization)
- Thu giảm dữ liệu (data reduction): thu giảm kích thước dữ liệu (nghĩa là giảm số phần tử) bằng kết hợp dữ liệu (data aggregation), loại bỏ các đặc điểm dư thừa (redundant features) (nghĩa là giảm số chiều/thuộc tính dữ liệu), gom cụm dữ liệu

## 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

---

### ▣ Các kỹ thuật tiền xử lý dữ liệu

#### ■ Làm sạch dữ liệu (data cleaning/cleansing)

- ▣ Tóm tắt hoá dữ liệu: nhận diện đặc điểm chung của dữ liệu và sự hiện diện của nhiễu hoặc các phần tử kì dị (outliers)
- ▣ Xử lý dữ liệu bị thiếu (missing data)
- ▣ Xử lý dữ liệu bị nhiễu (noisy data)

#### ■ Tích hợp dữ liệu (data integration)

- ▣ Tích hợp lược đồ (schema integration) và so trùng đối tượng (object matching)
- ▣ Vấn đề dư thừa (redundancy)
- ▣ Phát hiện và xử lý mâu thuẫn giá trị dữ liệu (detection and resolution of data value conflicts)

## 2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu

---

- Các kỹ thuật tiền xử lý dữ liệu
  - Biến đổi dữ liệu (data transformation)
    - Làm trơn dữ liệu (smoothing)
    - Kết hợp dữ liệu (aggregation)
    - Tổng quát hóa dữ liệu (generalization)
    - Chuẩn hóa dữ liệu (normalization)
    - Xây dựng thuộc tính (attribute/feature construction)
  - Thu giảm dữ liệu (data reduction)
    - Kết hợp khối dữ liệu (data cube aggregation)
    - Chọn tập con các thuộc tính (attribute subset selection)
    - Thu giảm chiều (dimensionality reduction)
    - Thu giảm lượng (numerosity reduction)
    - Tạo phân cấp ý niệm (concept hierarchy generation) và rời rạc hóa (discretization)

## 2.2. Tóm tắt mô tả về dữ liệu

---

- ▣ Xác định các thuộc tính (properties) tiêu biểu của dữ liệu về xu hướng chính (central tendency) và sự phân tán (dispersion) của dữ liệu
  - Các độ đo về xu hướng chính: mean, median, mode, midrange
  - Các độ đo về sự phân tán: quartiles, interquartile range (IQR), variance
- ▣ Làm nổi bật các giá trị dữ liệu nên được xem như nhiễu (noise) hoặc phần tử biên (outliers), cung cấp cái nhìn tổng quan về dữ liệu

## 2.2. Tóm tắt mô tả về dữ liệu

- Dữ liệu mẫu về đơn giá của các mặt hàng đã được bán

<i>Unit price (\$)</i>	<i>Count of items sold</i>
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350

## 2.2. Tóm tắt mô tả về dữ liệu

### ▣ Các độ đo về xu hướng chính của dữ liệu

- Mean 
$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

- Weighted arithmetic mean 
$$\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

- Median 
$$Median = \begin{cases} x_{\lceil N/2 \rceil} & \text{if } N \text{ odd} \\ (x_{N/2} + x_{N/2+1})/2 & \text{if } N \text{ even} \end{cases}$$

- Mode: giá trị xuất hiện thường xuyên nhất trong tập dữ liệu
- Midrange: giá trị trung bình của các giá trị lớn nhất và nhỏ nhất trong tập dữ liệu

## 2.2. Tóm tắt mô tả về dữ liệu

---

- ▣ Các độ đo về xu hướng chính của dữ liệu
  - Mean =  $\Sigma(\text{count}[i] * \text{price}[i]) / \Sigma(\text{count}[i])$
  - Weighted arithmetic mean
  - Median
  - Mode = price[i] nếu count[i] lớn nhất
  - Midrange =  $(\Sigma(\text{count}[i] * \text{price}[i]) + \Sigma(\text{count}[j] * \text{price}[j])) / (\Sigma(\text{count}[i]) + \Sigma(\text{count}[j]))$   
nếu price[i] lớn nhất và price[j] nhỏ nhất

## 2.2. Tóm tắt mô tả về dữ liệu

---

- ▣ Các độ đo về sự phân tán của dữ liệu

- Quartiles

- ▣ The first quartile (Q1): the 25<sup>th</sup> percentile
- ▣ The second quartile (Q2): the 50<sup>th</sup> percentile (median)
- ▣ The third quartile (Q3): the 75<sup>th</sup> percentile

- Interquartile Range (IQR) = Q3 – Q1

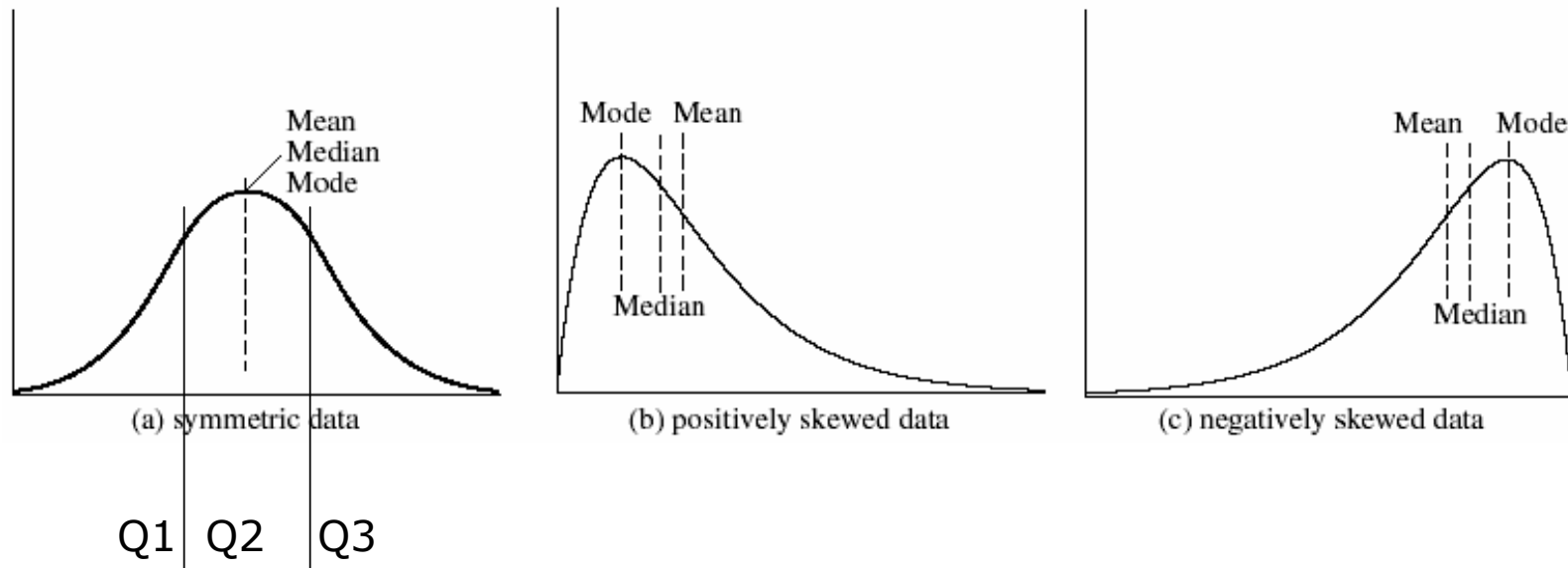
- ▣ Outliers (the most extreme observations): giá trị nằm cách trên Q3 hay dưới Q1 một khoảng 1.5xIQR

- Variance

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \frac{1}{N} \left[ \sum x_i^2 - \frac{1}{N} (\sum x_i)^2 \right]$$



## 2.2. Tóm tắt mô tả về dữ liệu



Tóm tắt mô tả về sự phân bố dữ liệu gồm năm trị số quan trọng: median, Q1, Q3, trị lớn nhất, và trị nhỏ nhất (theo thứ tự: Minimum, Q1, Median, Q3, Maximum).

## 2.3. Làm sạch dữ liệu

---

- ▣ Xử lý dữ liệu bị thiếu (missing data)
- ▣ Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
- ▣ Xử lý dữ liệu không nhất quán (inconsistent data)

## 2.3. Làm sạch dữ liệu

---

- ▣ Xử lý dữ liệu bị thiếu (missing data)
  - Định nghĩa của dữ liệu bị thiếu
    - ▣ Dữ liệu không có sẵn khi cần được sử dụng
  - Nguyên nhân gây ra dữ liệu bị thiếu
    - ▣ Khách quan (không tồn tại lúc được nhập liệu, sự cố, ...)
    - ▣ Chủ quan (tác nhân con người)
  - Giải pháp cho dữ liệu bị thiếu
    - ▣ Bỏ qua
    - ▣ Xử lý tay (không tự động, bán tự động)
    - ▣ Dùng giá trị thay thế (tự động): hằng số toàn cục, trị phổ biến nhất, trung bình toàn cục, trung bình cục bộ, trị dự đoán, ...
    - ▣ Ngăn chặn dữ liệu bị thiếu: thiết kế tốt CSDL và các thủ tục nhập liệu (các ràng buộc dữ liệu)

## 2.3. Làm sạch dữ liệu

---

- ▣ Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Định nghĩa
    - ▣ Outliers: những dữ liệu (đối tượng) không tuân theo đặc tính/hành vi chung của tập dữ liệu (đối tượng).
    - ▣ Noisy data: outliers bị loại bỏ (rejected/discarded outliers) như là những trường hợp ngoại lệ (exceptions).
  - Nguyên nhân
    - ▣ Khách quan (công cụ thu thập dữ liệu, lỗi trên đường truyền, giới hạn công nghệ, ...)
    - ▣ Chủ quan (tác nhân con người)

## 2.3. Làm sạch dữ liệu

---

- ▣ Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp nhận diện phần tử biên
    - ▣ Dựa trên phân bố thống kê (statistical distribution-based)
    - ▣ Dựa trên khoảng cách (distance-based)
    - ▣ Dựa trên mật độ (density-based)
    - ▣ Dựa trên độ lệch (deviation-based)
  - Giải pháp giảm thiểu nhiễu
    - ▣ Binning
    - ▣ Hồi quy (regression)
    - ▣ Phân tích cụm (cluster analysis)

## 2.3. Làm sạch dữ liệu

---

### □ Giải pháp giảm thiểu nhiễu

- Binning (by bin means, bin median, bin boundaries)
  - Dữ liệu có thứ tự
  - Phân bố dữ liệu vào các bins (buckets)
  - Bin boundaries: trị min và trị max

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

Smoothing by bin boundaries:

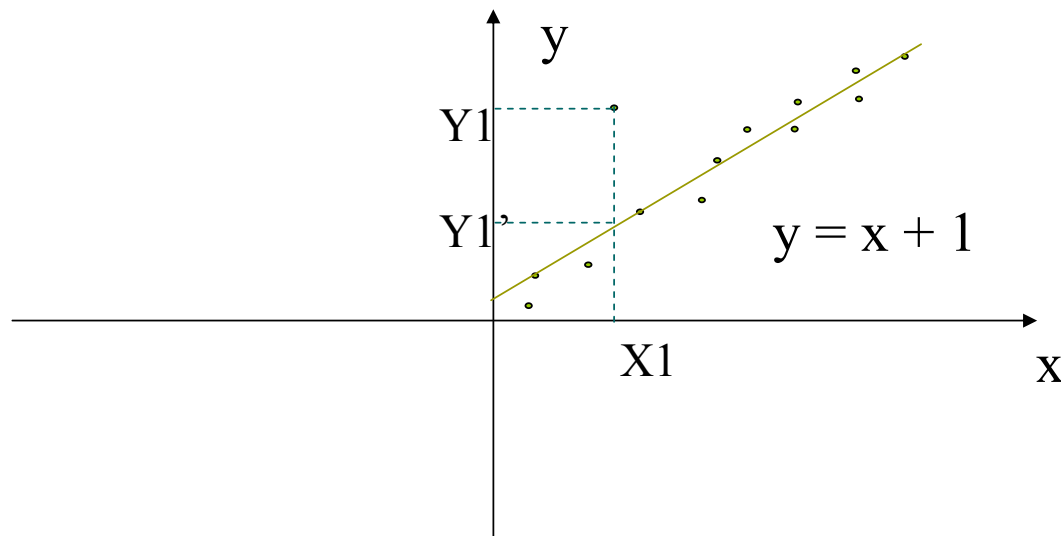
Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

## 2.3. Làm sạch dữ liệu

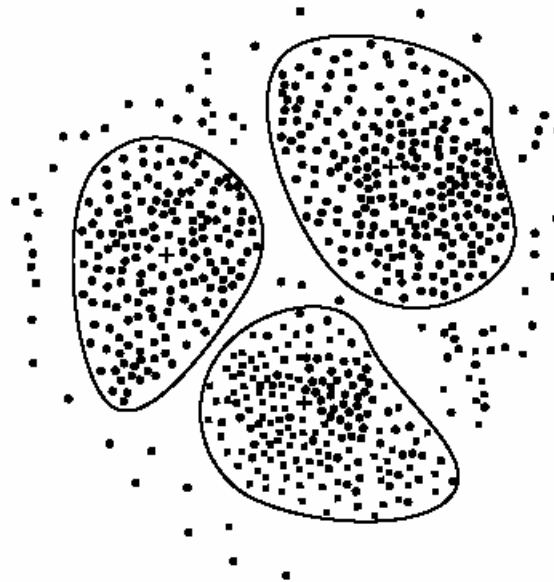
- ▣ Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp giảm thiểu nhiễu
    - ▣ Hồi quy (regression)



## 2.3. Làm sạch dữ liệu

---

- ▣ Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp giảm thiểu nhiễu
    - ▣ Phân tích cụm (cluster analysis)





## 2.3. Làm sạch dữ liệu

---

### ▣ Xử lý dữ liệu không nhất quán

#### ■ Định nghĩa của dữ liệu không nhất quán

- ▣ Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể → discrepancies from inconsistent data representations
  - 2004/12/25 và 25/12/2004
- ▣ Dữ liệu được ghi nhận không phản ánh đúng ngữ nghĩa cho các đối tượng/thực thể
  - Ràng buộc khóa ngoại

#### ■ Nguyên nhân

- ▣ Sự không nhất quán trong các quy ước đặt tên hay mã dữ liệu
- ▣ Định dạng không nhất quán của các vùng nhập liệu
- ▣ Thiết bị ghi nhận dữ liệu, ...

## 2.3. Làm sạch dữ liệu

---

- ▣ Xử lý dữ liệu không nhất quán (inconsistent data)
  - Giải pháp
    - ▣ Tận dụng siêu dữ liệu, ràng buộc dữ liệu, sự kiểm tra của nhà phân tích dữ liệu cho việc nhận diện
    - ▣ Điều chỉnh dữ liệu không nhất quán bằng tay
    - ▣ Các giải pháp biến đổi/chuẩn hóa dữ liệu tự động

## 2.4. Tích hợp dữ liệu

---

- ▣ Tích hợp dữ liệu: quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu
  - Vấn đề nhận dạng thực thể (entity identification problem)
    - ▣ Tích hợp lược đồ (schema integration)
    - ▣ So trùng đối tượng (object matching)
  - Vấn đề dư thừa (redundancy)
  - Vấn đề mâu thuẫn giá trị dữ liệu (data value conflicts)
- Liên quan đến cấu trúc và tính không thuần nhất (heterogeneity) về ngữ nghĩa (semantics) của dữ liệu
- Hỗ trợ việc giảm và tránh dư thừa và không nhất quán về dữ liệu → cải thiện tính chính xác và tốc độ quá trình khai phá dữ liệu

## 2.4. Tích hợp dữ liệu

---

- ▣ Vấn đề nhận dạng thực thể
    - Các thực thể (object/entity/attribute) đến từ nhiều nguồn dữ liệu.
    - Hai hay nhiều thực thể khác nhau diễn tả cùng một thực thể thực.
    - Ví dụ ở mức lược đồ (schema): `customer_id` trong nguồn S1 và `cust_number` trong nguồn S2.
    - Ví dụ ở mức thể hiện (instance): "R & D" trong nguồn S1 và "Research & Development" trong nguồn S2. "Male" và "Female" trong nguồn S1 và "Nam" và "Nữ" trong nguồn S2.
- Vai trò của siêu dữ liệu (metadata)

## 2.4. Tích hợp dữ liệu

---

### ▣ Vấn đề dư thừa

- Hiện tượng: giá trị của một thuộc tính có thể được dẫn ra/tính từ một/nhiều thuộc tính khác, vấn đề trùng lặp dữ liệu (duplication).
- Nguyên nhân: tổ chức dữ liệu kém, không nhất quán trong việc đặt tên chiều/thuộc tính.
- Phát hiện dư thừa: phân tích tương quan (correlation analysis)
  - ▣ Dựa trên dữ liệu hiện có, kiểm tra khả năng dẫn ra một thuộc tính B từ thuộc tính A.
  - ▣ Đối với các thuộc tính số (numerical attributes), đánh giá tương quan giữa hai thuộc tính với các hệ số tương quan (correlation coefficient, aka Pearson's product moment coefficient).
  - ▣ Đối với các thuộc tính rời rạc (categorical/discrete attributes), đánh giá tương quan giữa hai thuộc tính với phép kiểm thử chi-square ( $\chi^2$ ).

## 2.4. Tích hợp dữ liệu

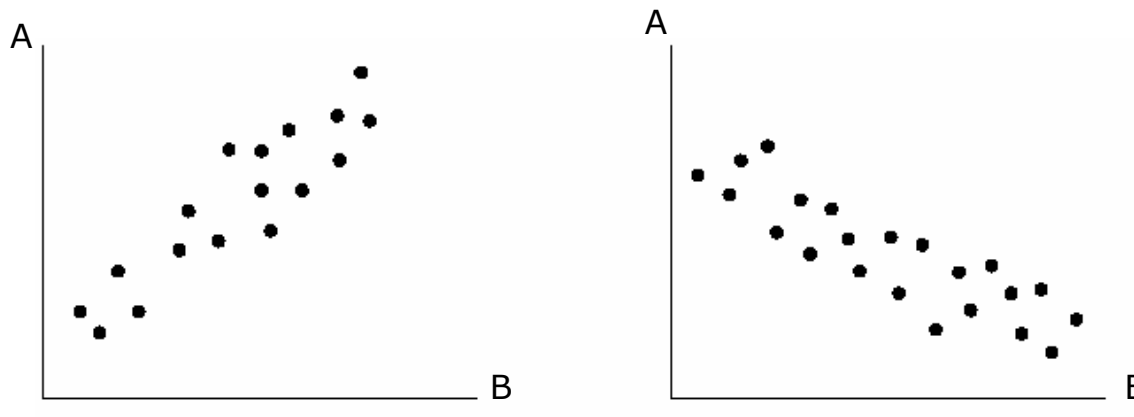
---

- Phân tích tương quan giữa hai thuộc tính số A và B
  - $r_{A,B} \in [-1, 1]$
  - $r_{A,B} > 0$ : A và B tương quan thuận với nhau, trị số của A tăng khi trị số của B tăng,  $r_{A,B}$  càng lớn thì mức độ tương quan càng cao, A hoặc B có thể được loại bỏ vì dư thừa.
  - $r_{A,B} = 0$ : A và B không tương quan với nhau (độc lập).
  - $r_{A,B} < 0$ : A và B tương quan nghịch với nhau, A và B loại trừ lẫn nhau.

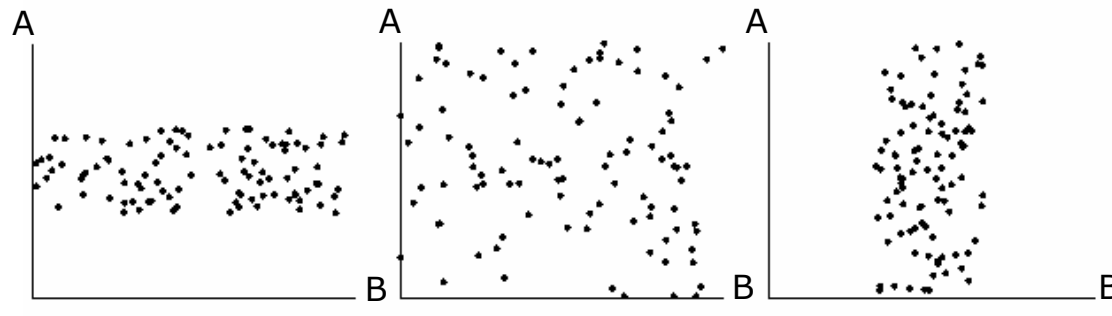
$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^N (a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

## 2.4. Tích hợp dữ liệu

- Phân tích tương quan giữa hai thuộc tính số A và B



Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.



Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

## 2.4. Tích hợp dữ liệu

- Phân tích tương quan giữa hai thuộc tính rời rạc A và B
  - A có  $c$  giá trị phân biệt,  $a_1, a_2, \dots, a_c$ .
  - B có  $r$  giá trị phân biệt,  $b_1, b_2, \dots, b_r$ .
  - $o_{ij}$ : số lượng đối tượng (tuples) có trị thuộc tính A là  $a_i$  và trị thuộc tính B là  $b_j$ .
  - $\text{count}(A=a_i)$ : số lượng đối tượng có trị thuộc tính A là  $a_i$ .
  - $\text{count}(B=b_j)$ : số lượng đối tượng có trị thuộc tính B là  $b_j$ .

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$



## 2.4. Tích hợp dữ liệu

---

- ▣ Phân tích tương quan giữa hai thuộc tính rời rạc A và B

- Phép kiểm thống kê chi-square kiểm tra giả thuyết liệu A và B có độc lập với nhau dựa trên một mức significance (significance level) với độ tự do (degree of freedom).

- ▣ Nếu giả thuyết bị loại bỏ thì A và B có sự liên hệ với nhau dựa trên thống kê.

- Độ tự do (degree of freedom):  $(r-1)*(c-1)$

- ▣ Tra bảng phân bố chi-square để xác định giá trị  $\chi^2$ .

- ▣ Nếu giá trị tính toán được lớn hơn hay bằng trị tra bảng được thì hai thuộc tính A và B không độc lập nhau (giả thuyết sai)

## 2.4. Tích hợp dữ liệu

---

- ▣ Vấn đề mâu thuẫn giá trị dữ liệu
  - Cho cùng một thực thể thật, các giá trị thuộc tính đến từ các nguồn dữ liệu khác nhau có thể khác nhau về cách biểu diễn (representation), đo lường (scaling), và mã hóa (encoding).
    - ▣ Representation: "2004/12/25" với "25/12/2004".
    - ▣ Scaling: thuộc tính *weight* trong các hệ thống đo khác nhau với các đơn vị đo khác nhau, thuộc tính *price* trong các hệ thống tiền tệ khác nhau với các đơn vị tiền tệ khác nhau.
    - ▣ Encoding: "yes" và "no" với "1" và "0".

## 2.5. Biến đổi dữ liệu

---

- ▣ Biến đổi dữ liệu: quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu
  - Làm trơn dữ liệu (smoothing)
  - Kết hợp dữ liệu (aggregation)
  - Tổng quát hoá (generalization)
  - Chuẩn hoá (normalization)
  - Xây dựng thuộc tính/đặc tính (attribute/feature construction)

## 2.5. Biến đổi dữ liệu

---

### ▣ Làm trơn dữ liệu (smoothing)

- Các phương pháp binning (bin means, bin medians, bin boundaries)
- Hồi quy
- Các kỹ thuật gom cụm (phân tích phần tử biên)
- Các phương pháp rời rạc hóa dữ liệu (các phân cấp ý niệm)

→ Loại bỏ/giảm thiểu nhiễu khỏi dữ liệu.

## 2.5. Biến đổi dữ liệu

---

### ▣ Kết hợp dữ liệu (aggregation)

- Các tác vụ kết hợp/tóm tắt dữ liệu
- Chuyển dữ liệu ở mức chi tiết này sang dữ liệu ở mức kém chi tiết hơn
- Hỗ trợ việc phân tích dữ liệu ở nhiều độ mịn thời gian khác nhau

→ Thu giảm dữ liệu (data reduction)

## 2.5. Biến đổi dữ liệu

---

- ▣ Tổng quát hóa (generalization)
  - Chuyển đổi dữ liệu cấp thấp/nguyên tố/thô sang các khái niệm ở mức cao hơn thông qua các phân cấp ý niệm
- Thu giảm dữ liệu (data reduction)

## 2.5. Biến đổi dữ liệu

---

- ▣ Chuẩn hóa (normalization)
  - min-max normalization
  - z-score normalization
  - Normalization by decimal scaling
- Các giá trị thuộc tính được chuyển đổi vào một miền trị nhất định được định nghĩa trước.

## 2.5. Biến đổi dữ liệu

---

### ▣ Chuẩn hóa (normalization)

#### ■ min-max normalization

▣ Giá trị cũ:  $v \in [\min_A, \max_A]$

▣ Giá trị mới:  $v' \in [\text{new\_min}_A, \text{new\_max}_A]$

→ Ví dụ: chuẩn hóa điểm số từ 0-4.0 sang 0-10.0.

→ Đặc điểm của phép chuẩn hóa min-max?

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$



## 2.5. Biến đổi dữ liệu

---

### ▣ Chuẩn hóa (normalization)

#### ■ z-score normalization

▣ Giá trị cũ:  $v$  tương ứng với mean  $\bar{A}$  và standard deviation  $\sigma_A$

▣ Giá trị mới:  $v'$

→ Đặc điểm của chuẩn hóa z-score?

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

## 2.5. Biến đổi dữ liệu

---

- ▣ Chuẩn hóa (normalization)

- Normalization by decimal scaling

- ▣ Giá trị cũ:  $v$

- ▣ Giá trị mới:  $v'$  với  $j$  là số nguyên nhỏ nhất sao cho  $\text{Max}(|v'|) < 1$

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

## 2.5. Biến đổi dữ liệu

---

- ▣ Xây dựng thuộc tính/đặc tính (attribute/feature construction)
    - Các thuộc tính mới được xây dựng và thêm vào từ tập các thuộc tính sẵn có.
    - Hỗ trợ kiểm tra tính chính xác và giúp hiểu cấu trúc của dữ liệu nhiều chiều.
    - Hỗ trợ phát hiện thông tin thiếu sót về các mối quan hệ giữa các thuộc tính dữ liệu.
- Các thuộc tính dẫn xuất

## 2.6. Thu giảm dữ liệu

---

- ❑ Tập dữ liệu được biến đổi đảm bảo các toàn vẹn, nhưng nhỏ/ít hơn nhiều về số lượng so với ban đầu.
  - ❑ Các chiến lược thu giảm
    - Kết hợp khối dữ liệu (data cube aggregation)
    - Chọn một số thuộc tính (attribute subset selection)
    - Thu giảm chiều (dimensionality reduction)
    - Thu giảm lượng (numerosity reduction)
    - Rời rạc hóa (discretization)
    - Tạo phân cấp ý niệm (concept hierarchy generation)
- Thu giảm dữ liệu: lossless và lossy

## 2.6. Thu giảm dữ liệu

- Kết hợp khối dữ liệu (data cube aggregation)
  - Dạng dữ liệu: additive, semi-additive (numerical)
  - Kết hợp dữ liệu bằng các hàm nhóm: average, min, max, sum, count, ...
- Dữ liệu ở các mức trừu tượng khác nhau.
- Mức trừu tượng càng cao giúp thu giảm lượng dữ liệu càng nhiều.

Year 2004	
Quarter	Sales
Q1	0
Q2	0
Q3	0
Q4	0

Year 2003	
Quarter	Sales
Q1	0
Q2	0
Q3	0
Q4	0

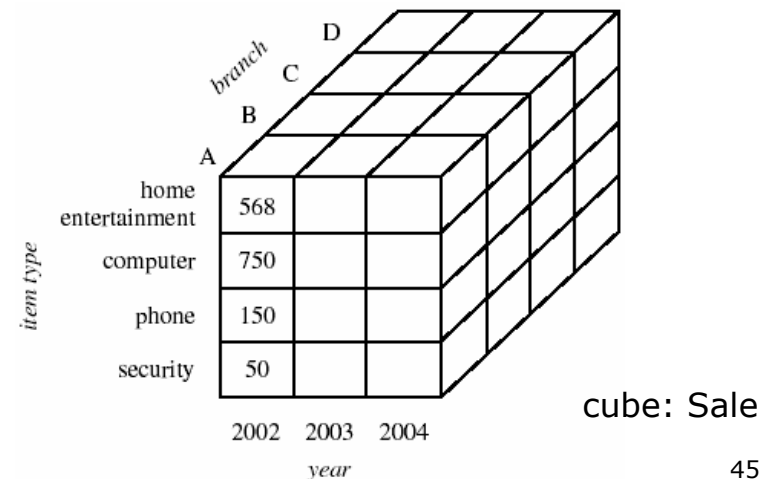
  

Year 2002	
Quarter	Sales
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

Sum() →

Year	Sales
2002	\$1,568,000
2003	\$2,356,000
2004	\$3,594,000



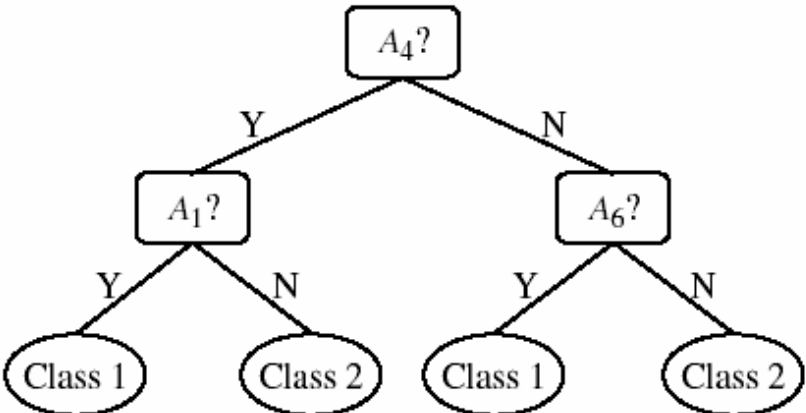
## 2.6. Thu giảm dữ liệu

---

- Chọn một số thuộc tính (attribute subset selection)
    - Giảm kích thước tập dữ liệu bằng việc loại bỏ những thuộc tính/chiều/đặc trưng (attribute/dimension/feature) dư thừa/không thích hợp (redundant/irrelevant)
    - Mục tiêu: tập ít các thuộc tính nhất vẫn đảm bảo phân bố xác suất (probability distribution) của các lớp dữ liệu đạt được gần với phân bố xác suất ban đầu với tất cả các thuộc tính
- Bài toán tối ưu hóa: vận dụng heuristics

## 2.6. Thu giảm dữ liệu

### ▣ Chọn một số thuộc tính (attribute subset selection)

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set:  <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p> <p>Initial reduced set:  <math>\{\}</math>  <math>\Rightarrow \{A_1\}</math>  <math>\Rightarrow \{A_1, A_4\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set:  <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math>  <math>\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}</math>  <math>\Rightarrow \{A_1, A_4, A_5, A_6\}</math>  <math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>	<p>Initial attribute set:  <math>\{A_1, A_2, A_3, A_4, A_5, A_6\}</math></p>  <pre> graph TD     A4["A4?"] -- Y --&gt; A1["A1?"]     A4 -- N --&gt; A6["A6?"]     A1 -- Y --&gt; C1_1((Class 1))     A1 -- N --&gt; C2_1((Class 2))     A6 -- Y --&gt; C1_2((Class 1))     A6 -- N --&gt; C2_2((Class 2))     </pre> <p><math>\Rightarrow</math> Reduced attribute set:  <math>\{A_1, A_4, A_6\}</math></p>

## 2.6. Thu giảm dữ liệu

---

- Thu giảm chiều (dimensionality reduction)
    - Biến đổi wavelet (wavelet transforms)
    - Phân tích nhân tố chính (principal component analysis)
- đặc điểm và ứng dụng?



## 2.6. Thu giảm dữ liệu

---

- Thu giảm lượng (numerosity reduction)
  - Các kỹ thuật giảm lượng dữ liệu bằng các dạng biểu diễn dữ liệu thay thế.
  - Các phương pháp có thông số (parametric): mô hình ước lượng dữ liệu → các thông số được lưu trữ thay cho dữ liệu thật
    - Hồi quy
  - Các phương pháp phi thông số (nonparametric): lưu trữ các biểu diễn thu giảm của dữ liệu
    - Histogram, Clustering, Sampling

## 2.7. Rời rạc hóa dữ liệu

---

- ❑ Giảm số lượng giá trị của một thuộc tính liên tục (continuous attribute) bằng các chia miền trị thuộc tính thành các khoảng (intervals)
- ❑ Các nhãn (labels) được gán cho các khoảng (intervals) này và được dùng thay giá trị thực của thuộc tính
- ❑ Các trị thuộc tính có thể được phân hoạch theo một phân cấp (hierarchical) hay ở nhiều mức phân giải khác nhau (multiresolution)

## 2.7. Rời rạc hóa dữ liệu

---

- Rời rạc hóa dữ liệu cho các thuộc tính số (numeric attributes)
  - Các phân cấp ý niệm được dùng để thu giảm dữ liệu bằng việc thu thập và thay thế các ý niệm cấp thấp bởi các ý niệm cấp cao.
  - Các phân cấp ý niệm được xây dựng tự động dựa trên việc phân tích phân bố dữ liệu.
  - Chi tiết của thuộc tính sẽ bị mất.
  - Dữ liệu đạt được có ý nghĩa và dễ được diễn dịch hơn, đòi hỏi ít không gian lưu trữ hơn.

## 2.7. Rời rạc hóa dữ liệu

---

- ▣ Các phương pháp rời rạc hóa dữ liệu cho các thuộc tính số
  - Binning
  - Histogram analysis
  - Interval merging by  $\chi^2$  analysis
  - Cluster analysis
  - Entropy-based discretization
  - Discretization by “natural/intuitive partitioning”

## 2.8. Tạo cây phân cấp ý niệm

---

- ▣ Dữ liệu phân loại (categorical data)

- Dữ liệu rời rạc (discrete data)

- Miền trị thuộc tính phân loại (categorical attribute)

- ▣ Số giá trị phân biệt hữu hạn

- ▣ Không có thứ tự giữa các giá trị

→ Tạo phân cấp ý niệm cho dữ liệu rời rạc

## 2.8. Tạo cây phân cấp ý niệm

---

- ▣ Các phương pháp tạo phân cấp ý niệm cho dữ liệu rời rạc (categorical/discrete data)
  - Đặc tả thứ tự riêng phần (partial ordering)/thứ tự toàn phần (total ordering) của các thuộc tính tương minh ở mức lược đồ bởi người sử dụng hoặc chuyên gia
  - Đặc tả một phần phân cấp bằng cách nhóm dữ liệu tương minh

## 2.8. Tạo cây phân cấp ý niệm

---

- ▣ Các phương pháp tạo phân cấp ý niệm cho dữ liệu rời rạc (categorical/discrete data)
  - Đặc tả một tập các thuộc tính, nhưng không bao gồm thứ tự riêng phần của chúng
  - Đặc tả chỉ một tập riêng phần các thuộc tính (partial set of attributes)
  - Tạo phân cấp ý niệm bằng cách dùng các kết nối ngữ nghĩa được chỉ định trước

## 2.9. Tóm tắt

---

- ❑ Dữ liệu thực tế: không đầy đủ (incomplete/missing), nhiễu (noisy), không nhất quán (inconsistent)
- ❑ Quá trình tiền xử lý dữ liệu
  - làm sạch dữ liệu: xử lý dữ liệu bị thiếu, làm trơn dữ liệu nhiễu, nhận dạng các phần tử biên, hiệu chỉnh dữ liệu không nhất quán
  - tích hợp dữ liệu: vấn đề nhận dạng thực thể, vấn đề dư thừa, vấn đề mâu thuẫn giá trị dữ liệu
  - biến đổi dữ liệu: làm trơn dữ liệu, kết hợp dữ liệu, tổng quát hóa, chuẩn hóa, xây dựng thuộc tính/đặc tính
  - thu giảm dữ liệu: kết hợp khối dữ liệu, chọn một số thuộc tính, thu giảm chiều, rời rạc hóa và tạo phân cấp ý niệm



## 2.9. Tóm tắt

---

### ❑ Rời rạc hóa dữ liệu

- Thu giảm số trị của một thuộc tính liên tục (continuous attribute) bằng cách chia miền trị thành các khoảng (interval) có dán nhãn. Các nhãn này được dùng thay cho các giá trị thực.
- Tiến hành theo hai cách: trên xuống (top down) và dưới lên (bottom up), có giám sát (supervised) và không có giám sát (unsupervised).
- Tạo phân hoạch phân cấp/đa phân giải (multiresolution) trên các trị thuộc tính → phân cấp ý niệm cho thuộc tính số (numerical attribute)

### ❑ Tạo cây phân cấp ý niệm

- Hỗ trợ khai phá dữ liệu ở nhiều mức trừu tượng
- Cho thuộc tính số (numerical attributes): binning, histogram analysis, entropy-based discretization,  $\chi^2$ -merging, cluster analysis, discretization by intuitive partitioning
- Cho thuộc tính phân loại/rời rạc (categorical/discrete attributes): chỉ định tường minh bởi người sử dụng hay chuyên gia, nhóm dữ liệu tường minh, dựa trên số lượng trị phân biệt (khác nhau) của mỗi thuộc tính

# Hỏi & Đáp ...

---