

# KHAI PHÁ DỮ LIỆU

## Bài 2. Tiền xử lý dữ liệu

Giáo viên: TS. Trần Mạnh Tuấn

Bộ môn: Hệ thống thông tin

Khoa: Công nghệ thông tin

Email: [tmtuan@tlu.edu.vn](mailto:tmtuan@tlu.edu.vn)

Điện thoại: 0983.668.841

# Nội dung

1 Tổng quan về giai đoạn tiền xử lý dữ liệu

2 Tóm tắt mô tả về dữ liệu

3 Làm sạch dữ liệu

4 Tích hợp dữ liệu

5 Biến đổi dữ liệu

6 Thu giảm dữ liệu

7 Rời rạc hóa dữ liệu

8 Tạo cây phân cấp ý niệm

# Tổng quan về giai đoạn tiền xử lý dữ liệu

## Tình huống KPDL giáo dục

\* Dự đoán khả năng tốt nghiệp đúng hạn của sinh viên đại học chính quy

MSSV	Mã MH	Năm học	Học kỳ	Điểm giữa kỳ	Điểm cuối kỳ
50503660	001001	2005	1	6	5.5
50503660	004010	2005	1	NULL	8
50503660	004009	2005	1	NULL	7
50503660	006004	2005	1	3.5	13
50503660	007005	2005	1	NULL	4
50501879	007005	2005	1	5	10
50501879	006001	2005	1	4	13

"NULL" nên được diễn dịch theo những nghĩa nào?

Miền trị của điểm số: [0, 1]; [0, 10]; {yếu, kém, trung bình, trung bình khá, khá, giỏi, xuất sắc}

Tất cả sinh viên đều được xem xét trong bài toán khai phá dữ liệu giáo dục?

Tất cả môn học đều được xem xét trong bài toán khai phá dữ liệu giáo dục?

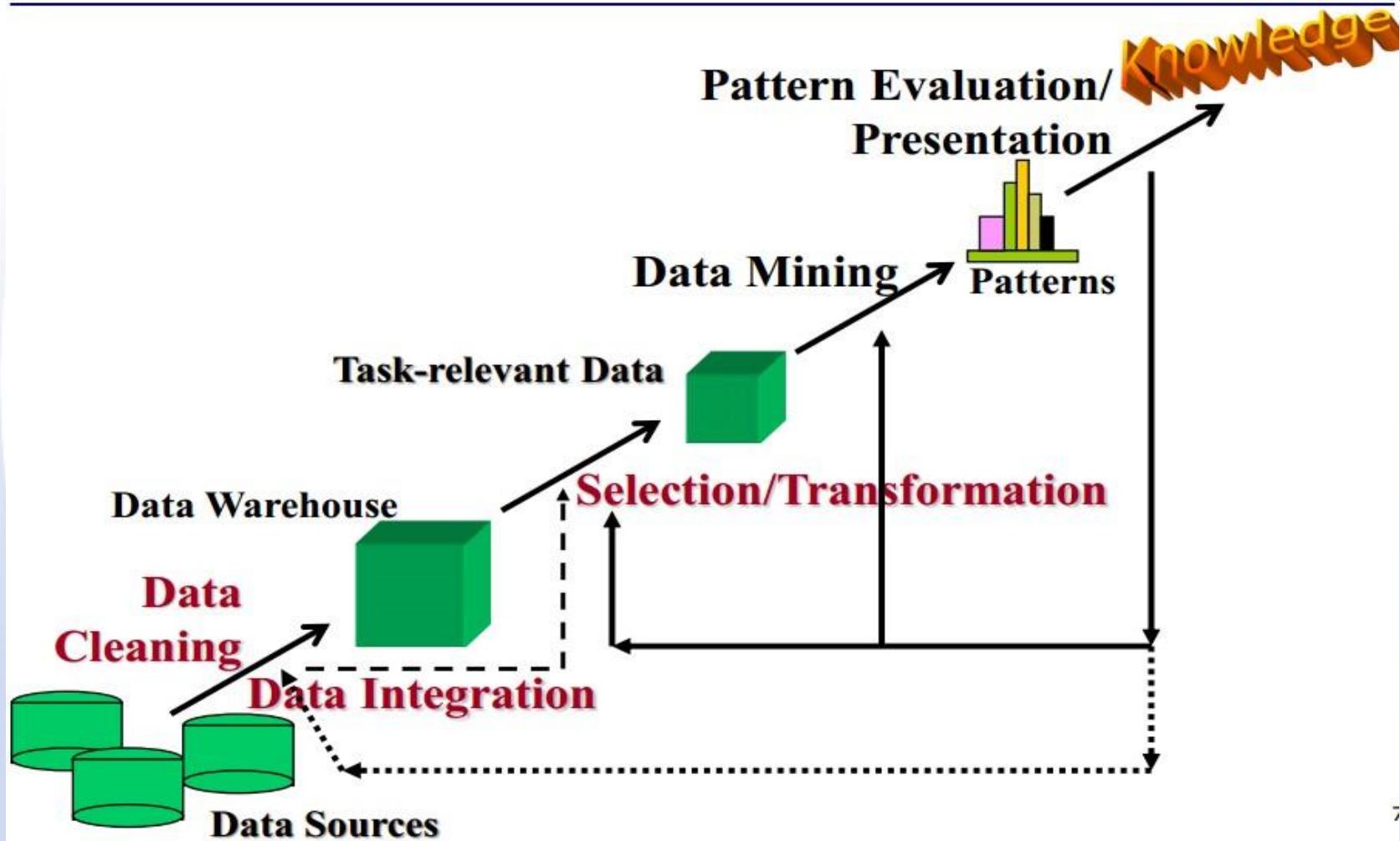
Ngoài kết quả điểm số môn học, đặc điểm gì của sinh viên có thể được xem xét trong bài toán khai phá dữ liệu giáo dục? .....

# Tổng quan về giai đoạn tiền xử lý dữ liệu

## Tình huống KPDL giáo dục

A	B	C	D	E
STT	Họ-tên	MSSV	Dữ liệu	Xử lý
1	Đỗ Duy Quốc	7140255	Thông tin sản phẩm	cập nhật dữ liệu, tìm kiếm dữ liệu, thống kê dữ liệu
2	Trương Thị Mỹ Ngọc	7140830	Thông tin khách hàng, sản phẩm	truy vấn dữ liệu
3	Nguyễn Thiên Khanh	7140241	Thông tin khách hàng, bất động sản	đưa dữ liệu có sẵn vào hệ thống hiện tại
4	Đoàn Dũ	7140223	Dữ liệu khách hàng	So sánh doanh số, trực quan dữ liệu với chart, dự đoán doanh số trong tương lai
5			Dữ liệu là thông tin, có thể khai thác từ 1 tập rộng lớn	Thêm, xóa, sửa dữ liệu, truy vấn dữ liệu, tạo thống kê, báo cáo
6			Những thông tin của các đối tượng trong thế giới thực	Tạo, ghi, xem, xóa, thêm dữ liệu
7	Trần Văn Triết	7140262	Dữ liệu không cấu trúc trong thống kê về web, data warehouse	thống kê dữ liệu
8	Lê Nhựt Trường	7140263	Dữ liệu về quản lý học vụ và quản lý sinh viên Đại học Cần Thơ, dữ liệu về quản lý vật tư và chi phí,	Tạo mẫu báo cáo dữ liệu và kế hoạch
9	Bùi Tiến Đức			Rút trích dữ liệu để tổng hợp, đánh giá. Từ đó, xây dựng biểu đồ cho sản phẩm
10	Trần Ngọc Như Quỳnh	7140256	Dữ liệu bán hàng online	Lọc dữ liệu, thống kê doanh thu, export dữ liệu, ...
11	Chu Xuân Tinh	7140838	Dữ liệu về thu phí giao thông đường bộ	truy vấn dữ liệu, back up hệ thống dữ liệu, ...
12	Lê Nguyên Dũng	7140224	Dữ liệu âm thanh	phân loại nhạc theo thể loại, dựa vào thông tin hiện trạng các application
13	Lê Nguyễn Khánh Duy	7140226	Phân loại nhạc theo thể loại, phân tích email có là spam hay không (dùng phương pháp thống kê)	
14			Dữ liệu quản lý học sinh và giáo viên	Thêm, xóa, sửa, cập nhật, bổ sung, thống kê, ...
15			Dữ liệu/thông tin về trạng thái của thiết bị mạng	truy vấn, thống kê, tìm lỗi của hệ thống thông qua dữ liệu
16	Bùi Đức Hiếu	7140231	Dữ liệu là thông tin được lưu trữ lại và dựa vào những dữ liệu này, chúng ta có thể khai thác ra điều gì	Làm sạch và xử lý nhiễu (loại bỏ sự gián đoạn), dự báo, ...
17			text, video, ảnh văn bản, thông tin về việc sử dụng đất	xử lý ảnh văn bản về dạng text, lập chỉ mục, thêm, sửa xóa các dữ liệu về...
18	Lê Văn	7141249	Dữ liệu là những thông tin được sắp xếp và sàng lọc theo một nội dung hay trình tự nào đó	
19	Âu Mậu Dương	7140820	Dữ liệu về sinh viên, môn học	Tìm sinh viên, thống kê môn học, ...
20		13070269	Dữ liệu GIS, dữ liệu giao thông, big data lưu trong database MongoDB, dữ liệu thông tin quản lý bể	xác định đường đi ngắn nhất qua 2 điểm, tìm thông tin đối tượng xung quanh
21	Đặng Quốc Huynh	7140237		Giảm chiều-thu giảm kích thước dữ liệu, gom cụm dữ liệu, phân loại dữ liệu
22	Trần Nhật Hoàng Anh	7140218	Tập hợp các thông tin được lưu trữ trên hệ thống/máy tính để có thể xử lý, thao tác được	Tìm kiếm, thu thập, nhập dữ liệu, xây dựng cơ sở dữ liệu, truy vấn dữ liệu
23	Nguyễn Phương Nhung	7140251	Tập hợp các thông tin được tổ chức lại và lưu trữ trên các phương tiện để xử lý: dữ liệu của một cửa hàng	Viết phần mềm quản lý cho việc bán điện thoại của cửa hàng đó
24			Dữ liệu về web, logs trên web server	kiểm tra, thống kê, phân tích về thói quen người dùng dựa trên cách người
25	Nguyễn Khắc Trung	7140839	Dữ liệu về dự án, khách hàng, chuỗi dây chuyền sản xuất sản phẩm thông qua các ứng dụng và các hệ quản trị cơ sở dữ liệu, ...	
26	Lê Minh Châu	7140818	Danh sách nhân viên	import file Excel vào database Oracle/MS SQL Server, thống kê báo cáo lưu

# Tổng quan về giai đoạn tiền xử lý dữ liệu

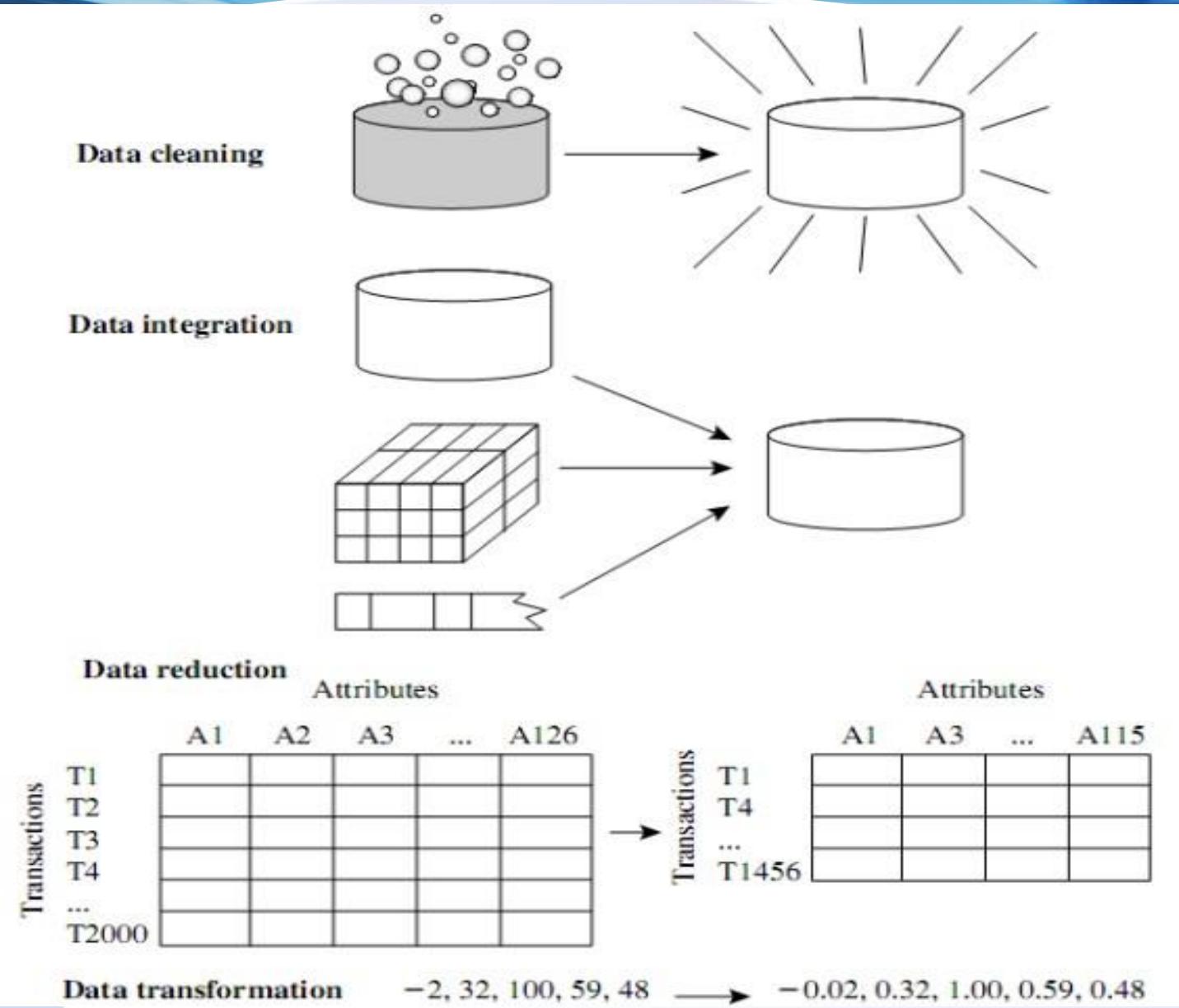


# Tổng quan về giai đoạn tiền xử lý dữ liệu

## 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

# Tổng quan về giai đoạn tiền xử lý dữ liệu



# 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

# 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 hóa 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)

# Tổng quan về giai đoạn tiền xử lý dữ liệu

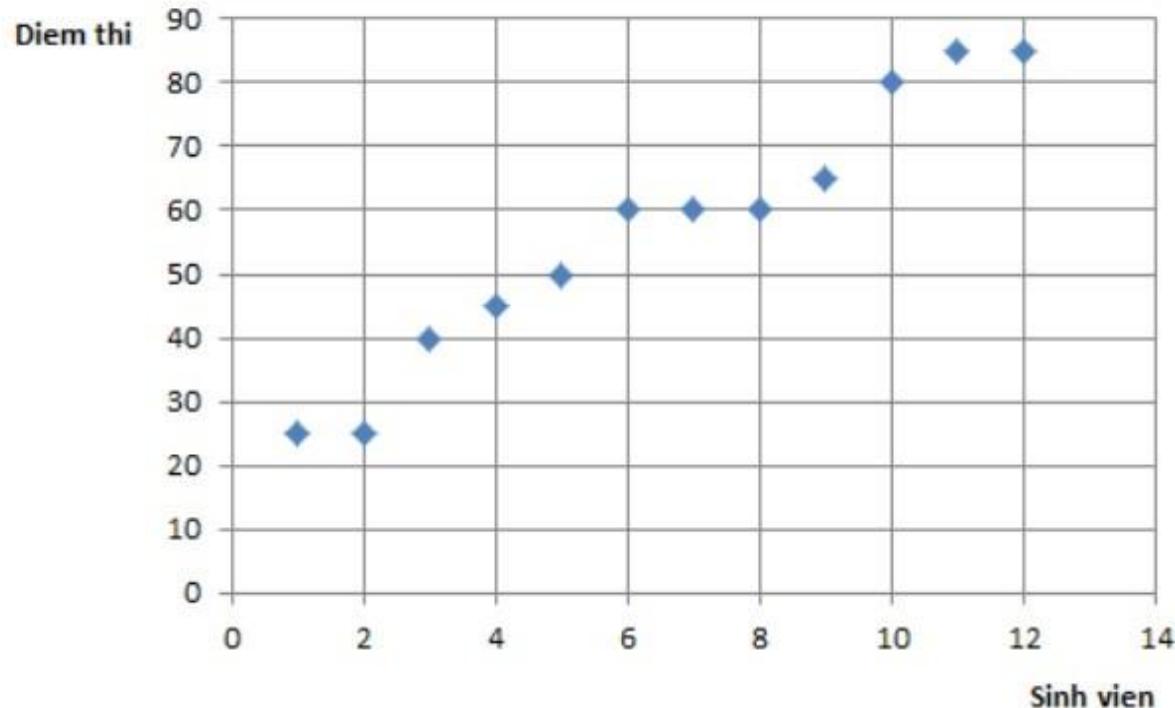
## Các kỹ thuật tiền xử lý dữ liệu

- 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)

# Tóm tắt mô tả về dữ liệu

## Dữ liệu về điểm số của các sinh viên

Sinh viên	Điểm thi
1	25
2	25
3	40
4	45
5	50
6	60
7	60
8	60
9	65
10	80
11	85
12	85



Đặc điểm phân bố và xu hướng của dữ liệu ???

Đặc điểm “đặc biệt” gì khác của dữ liệu ???

# 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

# Tóm tắt mô tả về dữ liệu

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

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

# Tóm tắt mô tả về dữ liệu

## Dữ liệu về điểm số của các sinh viên

Sinh viên	Điểm thi
1	25
2	25
3	40
4	45
5	50
6	60
7	60
8	60
9	65
10	80
11	85
12	85

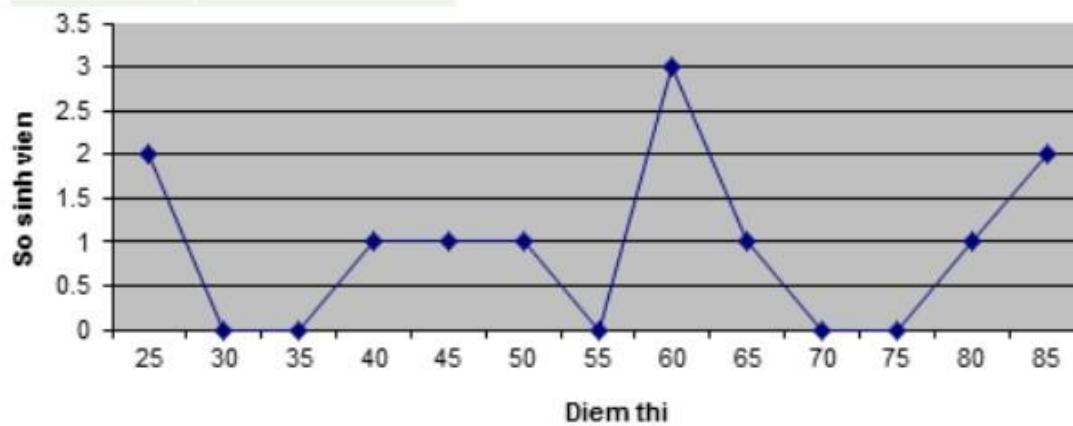
Điểm thi	Số sinh viên
25	2
30	0
35	0
40	1
45	1
50	1

Mean = 56.67

Median = 60

Mode = 60

Midrange = 55



# 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  $1.5 \times \text{IQR}$

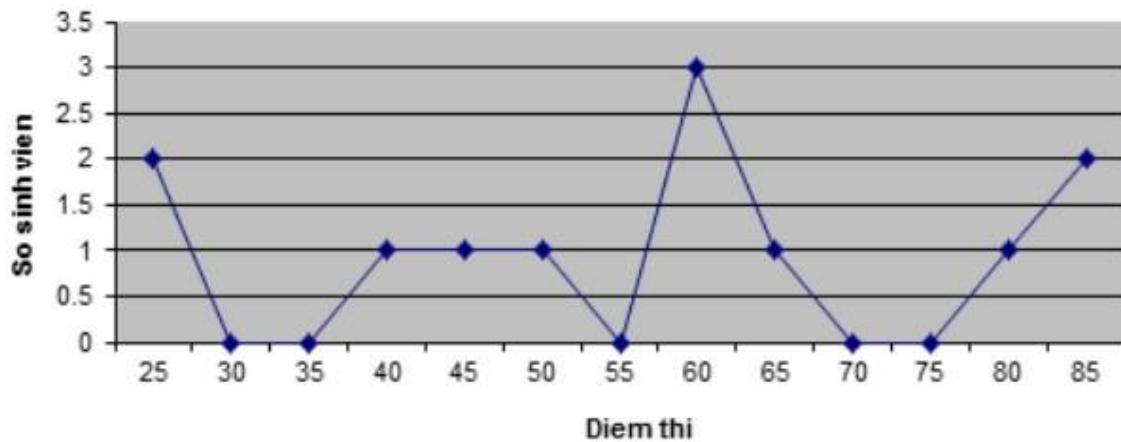
### ■ Variance

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

# Tóm tắt mô tả về dữ liệu

## Dữ liệu về điểm số của các sinh viên

Sinh viên	Điểm thi	Độ lệch
1	25	-31.66667
2	25	-31.66667
3	40	-16.66667
4	45	-11.66667
5	50	-6.666667
6	60	3.333333
7	60	3.333333
8	60	3.333333
9	65	8.333333
10	80	23.33333
11	85	28.33333
12	85	28.33333



$$Q_1 = 42.5$$

$$IQR = Q_3 - Q_1 = 30$$

$$Q_2 = \text{median} = 60$$

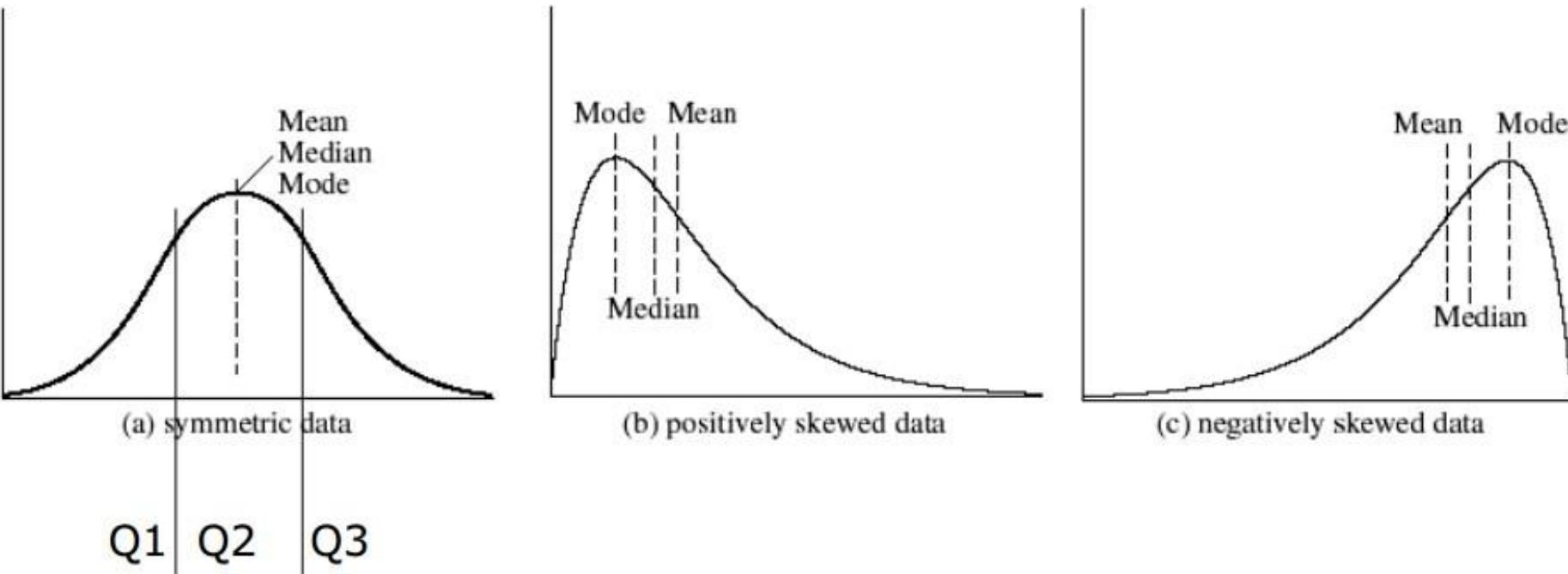
$\rightarrow$  Outliers = ???

$$Q_3 = 72.5$$

$$\text{Variance} = \sigma^2 = 393.06$$

$$\sigma = 19.83$$

# Tóm tắt mô tả về dữ liệu

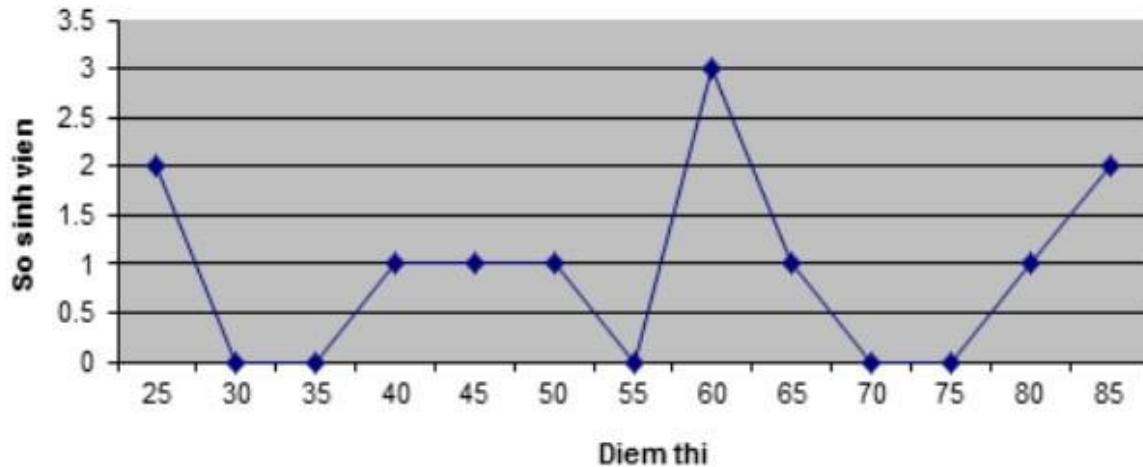


Tóm tắt mô tả về sự phân bố dữ liệu gồm năm trị số quan trọng:  
median, Q1, Q3, trị lớn nhất, và trị nhỏ nhất (theo thứ tự:  
Minimum, Q1, Median, Q3, Maximum).

# Tóm tắt mô tả về dữ liệu

## Dữ liệu về điểm số của các sinh viên

Sinh viên	Điểm thi
1	25
2	25
3	40
4	45
5	50
6	60
7	60
8	60
9	65
10	80
11	85
12	85



Mean = 56.67 < Mode = Median = 60

→ Negatively skewed data

Minimum, Q1, Median, Q3, Maximum

25, 42.5, 60, 72.5, 85

# 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)

# 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)

# 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)

# Làm sạch dữ liệu

## Giải pháp nhận diện phần tử biên

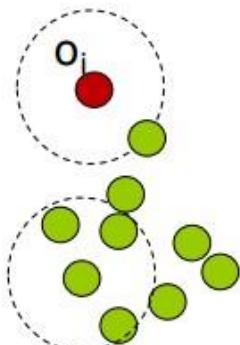
- Dựa trên phân bố thống kê (statistical distribution-based)
  - Thủ tục khối (block): tất cả các đối tượng tình nghi là outliers hoặc không
  - Thủ tục lần lượt/tuần tự (consecutive/sequential): đối tượng tình nghi nhất là outlier thì những đối tượng cực trị hơn cũng là outlier; nếu không thì đối tượng tình nghi kế sẽ được kiểm tra
    - Giả sử tập dữ liệu tuân theo một mô hình phân bố  $F$  cho trước (phân bố chuẩn, phân bố Poisson, ...), xác định các đối tượng  $o_i$  là outlier đối với mô hình phân bố này dùng phép thử discordancy nhằm kiểm tra 2 hypotheses:
      - *Working hypothesis*: với  $F$ , nếu significance probability SP(giá trị thống kê  $v_i$  của  $o_i$ ) =  $\text{Prob}(T > v_i)$  đủ nhỏ,  $o_i$  được xem là khác biệt (discordant) và working hypothesis không được chấp nhận. Nếu không thì tất cả đối tượng tuân theo  $F$ .
      - *Alternative hypothesis*: xác định xác suất mà working hypothesis không được chấp nhận khi  $o_i$  thật sự là outlier. Khi này,  $o_i$  tuân theo một mô hình phân bố  $G$ .

# Làm sạch dữ liệu

## Giải pháp nhận diện phần tử biên

### ■ Dựa trên khoảng cách (distance-based)

- Xem xét khoảng cách giữa các đối tượng đến đối tượng tình nghi  $o_i$ , nếu ít nhất một lượng đối tượng  $pct$  cách đối tượng  $o_i$  xa hơn một khoảng cách  $dmin$  thì  $o_i$  là outlier.
  - Outlier là những đối tượng không có đủ láng giềng trong khu vực được xác định bởi một khoảng cách cho trước.
  - Xác định giá trị  $pct$  và  $dmin$  cần dùng trial-and-error.



$o_i$  là outlier với  $pct = 0.8$  và  $dmin = 1$ .

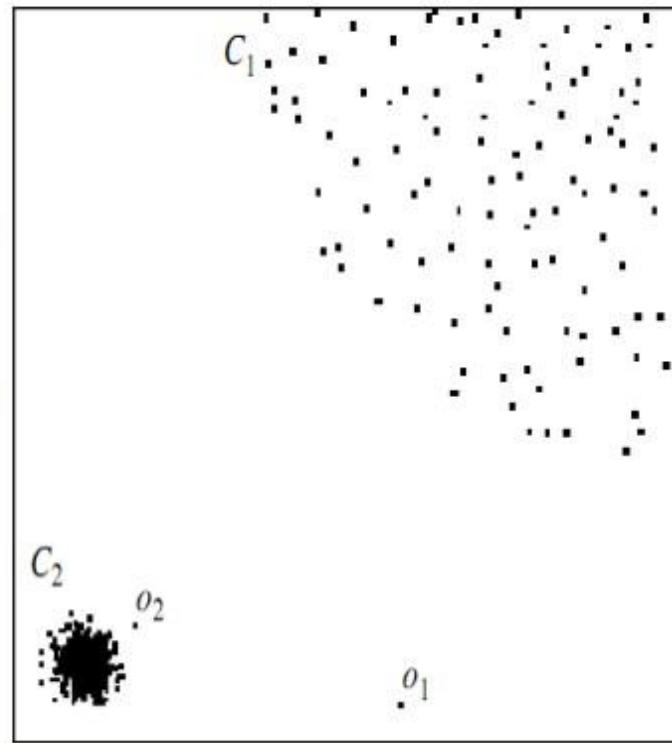
$pct$ : tỉ lệ số đối tượng không là láng giềng của outliers

$dmin$ : minimum distance dùng xác định vùng láng giềng của mỗi đối tượng

# Làm sạch dữ liệu

## Giải pháp nhận diện phần tử biên

- Dựa trên mật độ (density-based)
  - Dựa trên mật độ của vùng láng giềng của mỗi đối tượng
  - Mức độ của outlierness được xác định qua LOF (local outlier factor) của mỗi đối tượng p
    - Mức độ phụ thuộc vào mức độ cách ly của đối tượng đó đối với vùng láng giềng
    - k-distance của p
    - k-distance neighborhood của p
    - Reachability distance của p đối với o
    - Local reachability density của p
  - LOF(p) càng cao, p càng được xem là một local outlier.



$o_1$  và  $o_2$  là density-based outliers.

# Làm sạch dữ liệu

## Giải pháp nhận diện phần tử biên

- Dựa trên độ lệch (deviation-based)
  - Dựa trên việc kiểm tra các đặc điểm chính của các đối tượng trong một nhóm
    - Outliers là những đối tượng lệch khỏi các đối tượng khác dựa trên những đặc điểm chính
  - Sequential exception technique
    - Mô phỏng cách human phân biệt những đối tượng khác biệt khỏi chuỗi các đối tượng giống nhau; sử dụng dư thừa dữ liệu ngầm định
    - Xác định tập ngoại lệ (exception set): cho mỗi tập con trong chuỗi các tập con được tạo ra từ tập dữ liệu ban đầu, nếu tập con được bỏ khỏi tập dữ liệu và sự khác biệt giữa các đối tượng được giảm đi thì mức độ làm trơn (smoothing factor) của tập đó được xác định. Tập con có giá trị này lớn nhất là tập ngoại lệ
      - Tập ngoại lệ: tập gồm các outliers là tập con nhỏ nhất mà việc loại bỏ tập con này dẫn đến việc giảm đi nhiều nhất sự khác

# Làm sạch dữ liệu

## Giải pháp giảm thiểu nhiễu

- Binning (by bin means, bin median, bin boundaries)
  - Dữ liệu có thứ tự
  - Phân bổ dữ liệu vào các bins (buckets)
    - Equal-frequency
      - Số phần tử
    - Equal-width
      - Miền trị
  - 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

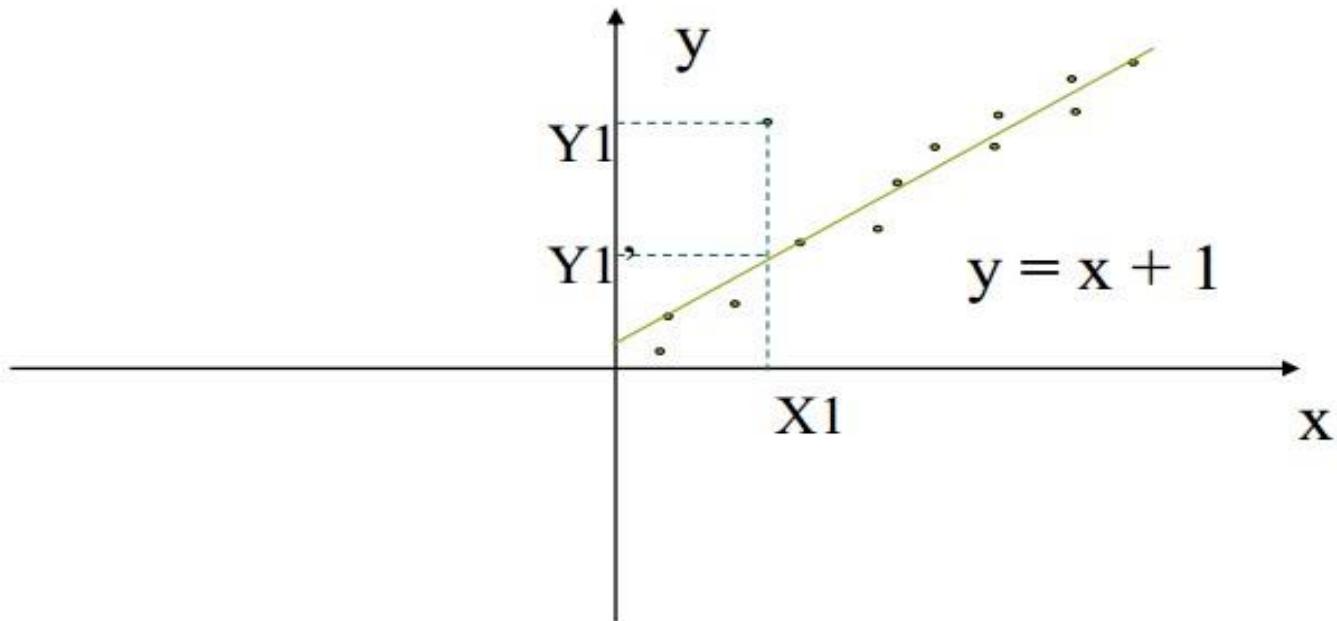
Smoothing by bin boundaries:

Bin 1: 4, 4, 15  
Bin 2: 21, 21, 24  
Bin 3: 25, 25, 34

# Làm sạch dữ liệu

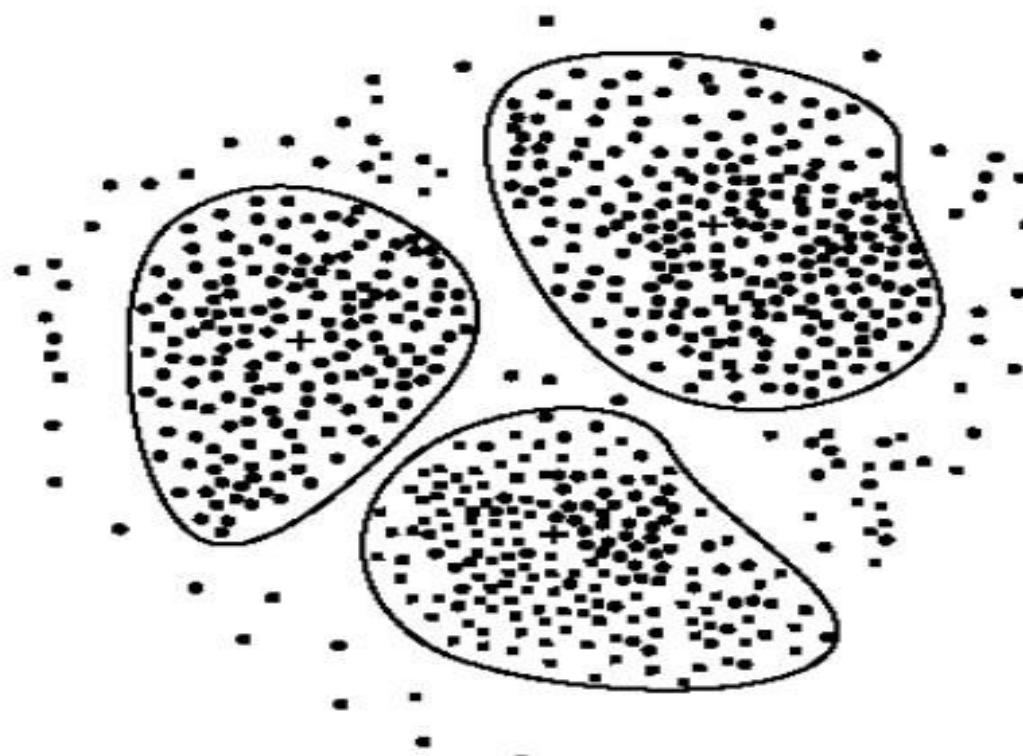
## Giải pháp giảm thiểu nhiễu

- Hồi quy (regression)



## Giải pháp giảm thiểu nhiễu

- Phân tích cụm (cluster analysis)



# 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
    - Định dạng ngày/tháng/năm: 2004/12/25 và 25/12/2004
    - Tên môn học: KPDЛ, Khai phá dữ liệu, Data mining
  - 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 qui ước đặt tên hay mã dữ liệu
  - Định dạng không nhất quán của các vùng nhập liệu
  - Thiết bị ghi nhận dữ liệu hay hệ thống bị lỗi

## 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

# 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

# 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 thật.
    - Ở mức lược đồ (schema):
      - customer\_id trong nguồn S1 và cust\_number trong nguồn S2
    - 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)

# 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 lắp 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 ( $\chi^2$ ).

# Tích hợp dữ liệu

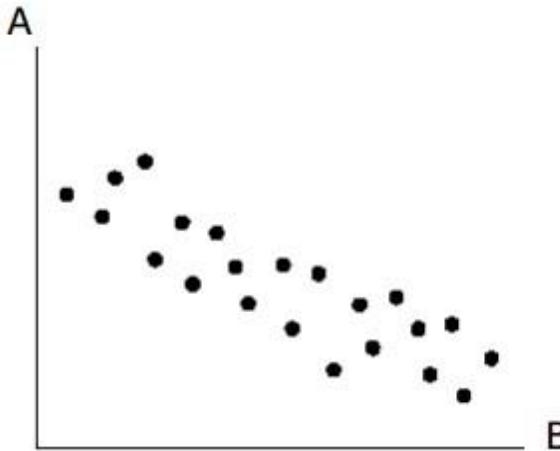
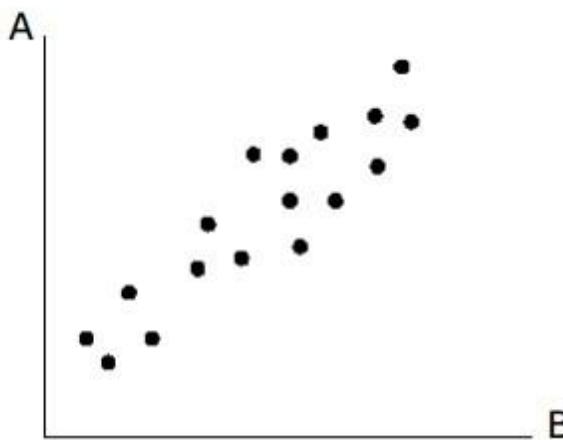
## Phân tích tương quan giữa hai thuộc tính số A và B

- $r_{A,B} \in [-1, 1]$
- $r_{A,B} > 0$ : 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.

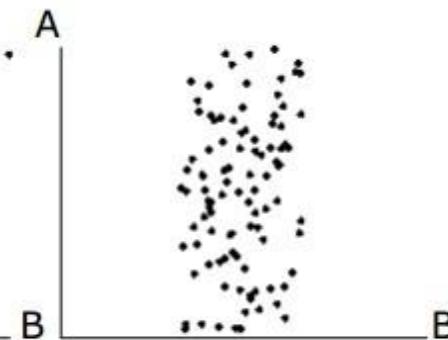
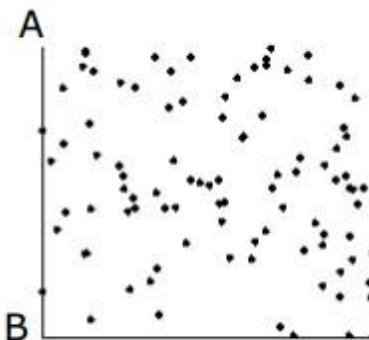
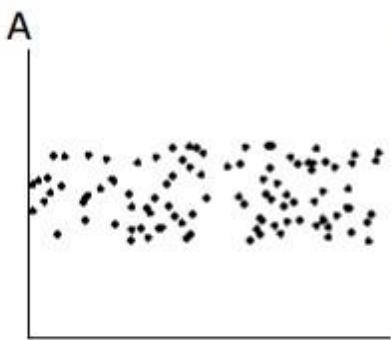
$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^N (a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

# 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.

# 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  $b_j$ .
- $e_{ij}$ : tần số kỳ vọng (expected frequency) của  $(A_i, B_j)$ .
- $\text{count}(A=a_i)$ : 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  $b_j$ .

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

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

# 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 tương quan với nhau (giả thuyết sai).

# Tích hợp dữ liệu

- Phân tích tương quan giữa hai thuộc tính rời rạc A và B
  - Giả sử khảo sát 1500 người với 2 thuộc tính *gender* và *preferred\_reading*

	<i>male</i>	<i>female</i>	Total
<i>fiction</i>	250 (90)	200 (360)	450
<i>non-fiction</i>	50 (210)	1000 (840)	1050
Total	300	1200	1500

Kiểm tra: *gender* và *preferred\_reading* có tương quan với nhau không

→ Phép kiểm thống kê  $\chi^2$  sẽ kiểm tra giả thuyết liệu *gender* và *preferred\_reading* có độc lập với nhau không

# Tích hợp dữ liệu

	male	female	Total
fiction	250 (90)	200 (360)	450
non-fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Kiểm tra: *gender* và *preferred\_reading* có tương quan với nhau không

→ Phép kiểm thống kê  $\chi^2$  sẽ kiểm tra giả thuyết liệu *gender* và *preferred\_reading* có độc lập với nhau không

$$o_{11} = 250; o_{12} = 200; o_{21} = 50; o_{22} = 1000$$

$$e_{11} = (\text{count(male)} * \text{count(fiction)}) / N = (300 * 450) / 1500 = 90$$

$$e_{12} = (\text{count(female)} * \text{count(fiction)}) / N = (1200 * 450) / 1500 = 360$$

$$e_{21} = (\text{count(male)} * \text{count(non_fiction)}) / N = (300 * 1050) / 1500 = 210$$

$$e_{22} = (\text{count(female)} * \text{count(non_fiction)}) / N = (1200 * 1050) / 1500 = 840$$

$$\begin{aligned}\chi^2 &= \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.\end{aligned}$$

$$\text{Degree of freedom} = (2-1) * (2-1) = 1; \text{Significance level} = 0.001$$

Tra bảng:  $\chi^2 = 10.828 << \chi^2$  tính được từ tập dữ liệu (507.93)

→ bác bỏ giả thuyết độc lập: *gender* và *preferred\_reading* có tương quan với nhau.

## 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”.

# 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 hóa (generalization)
- Chuẩn hóa (normalization)
- Xây dựng thuộc tính/đặc tính (attribute/feature construction)

## 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.

## 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)

## 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)

## 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.

## Chuẩn hóa (normalization)

### ■ min-max normalization

- Giá trị cũ:  $v \in [minA, maxA]$
- Giá trị mới:  $v' \in [new\_minA, new\_maxA]$ 
  - Ví dụ: chuẩn hóa điểm số từ 0-4.0 sang 0-10.0.
  - Đặc điểm của phép chuẩn hóa min-max?

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

## 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?

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

## Chuẩn hóa (normalization)

- Normalization by decimal scaling

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

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

# 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

# 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

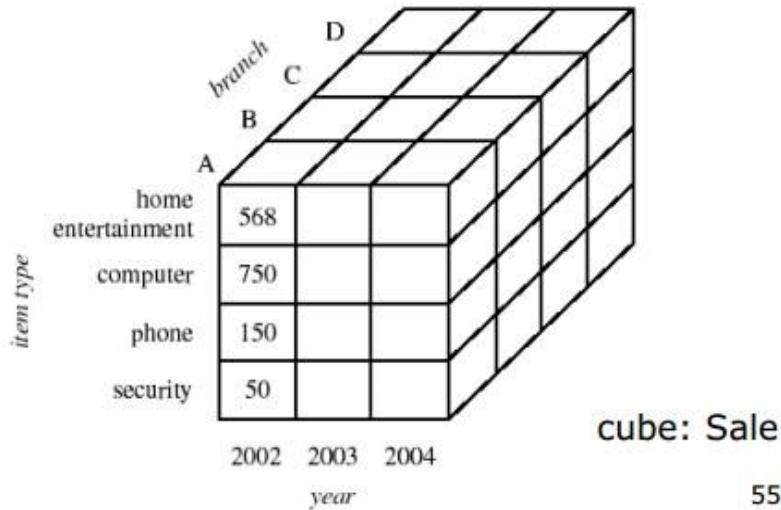
# Thu giảm dữ liệu

## 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.

The diagram illustrates the process of data cube aggregation. On the left, three separate tables represent data for the years 2002, 2003, and 2004. Each table has columns for Quarter and Sales. An arrow labeled "Sum()" points from these three tables to a final summary table on the right. This summary table has a single column for Year and a single column for Sales, showing the total sales for each year.

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



# 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

# Thu giảm dữ liệu

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

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$
Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$		<pre>graph TD; A4[A4?] -- Y --&gt; A1[A1?]; A4 -- N --&gt; A6[A6?]; A1 -- Y --&gt; Class1_1((Class 1)); A1 -- N --&gt; Class2_1((Class 2)); A6 -- Y --&gt; Class1_2((Class 1)); A6 -- N --&gt; Class2_2((Class 2))</pre> <p><math>\Rightarrow</math> Reduced attribute set: <math>\{A_1, A_4, A_6\}</math></p>

# 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?

# 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

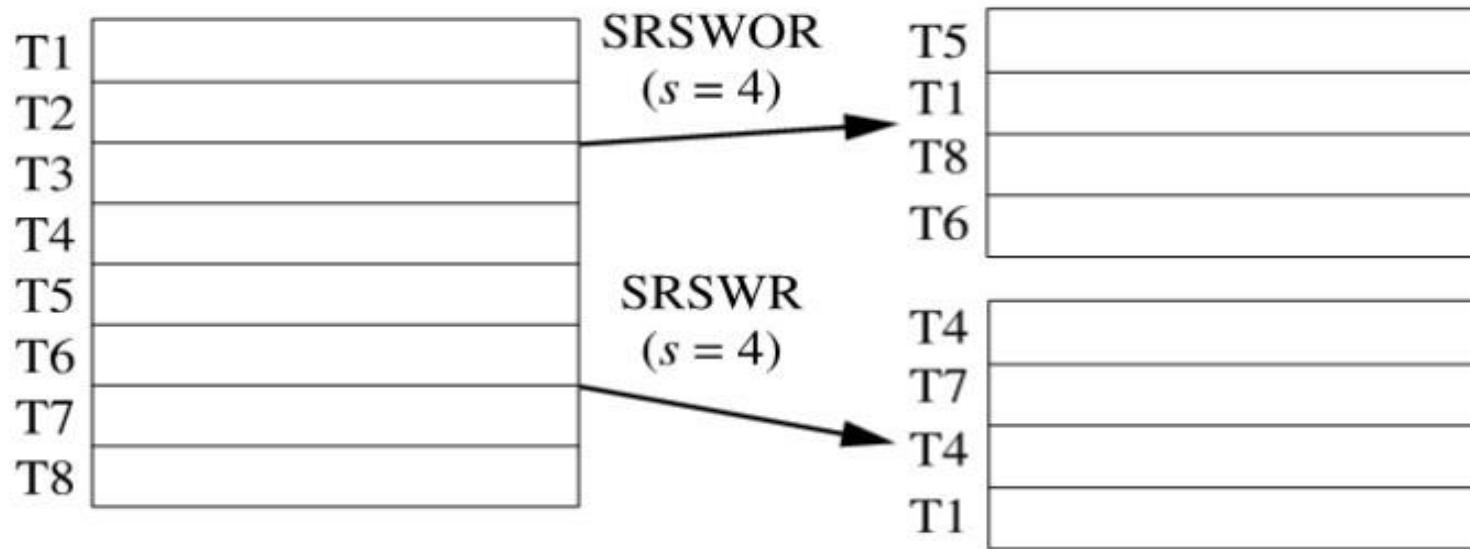
## Thu giảm lượng (numerosity reduction)

- Các phương pháp phi thông số (nonparametric):  
Sampling
  - Simple random sample without replacement (SRSWOR)
  - Simple random sample with replacement (SRSWR)
  - Cluster sample
  - Stratified sample

# Thu giảm dữ liệu

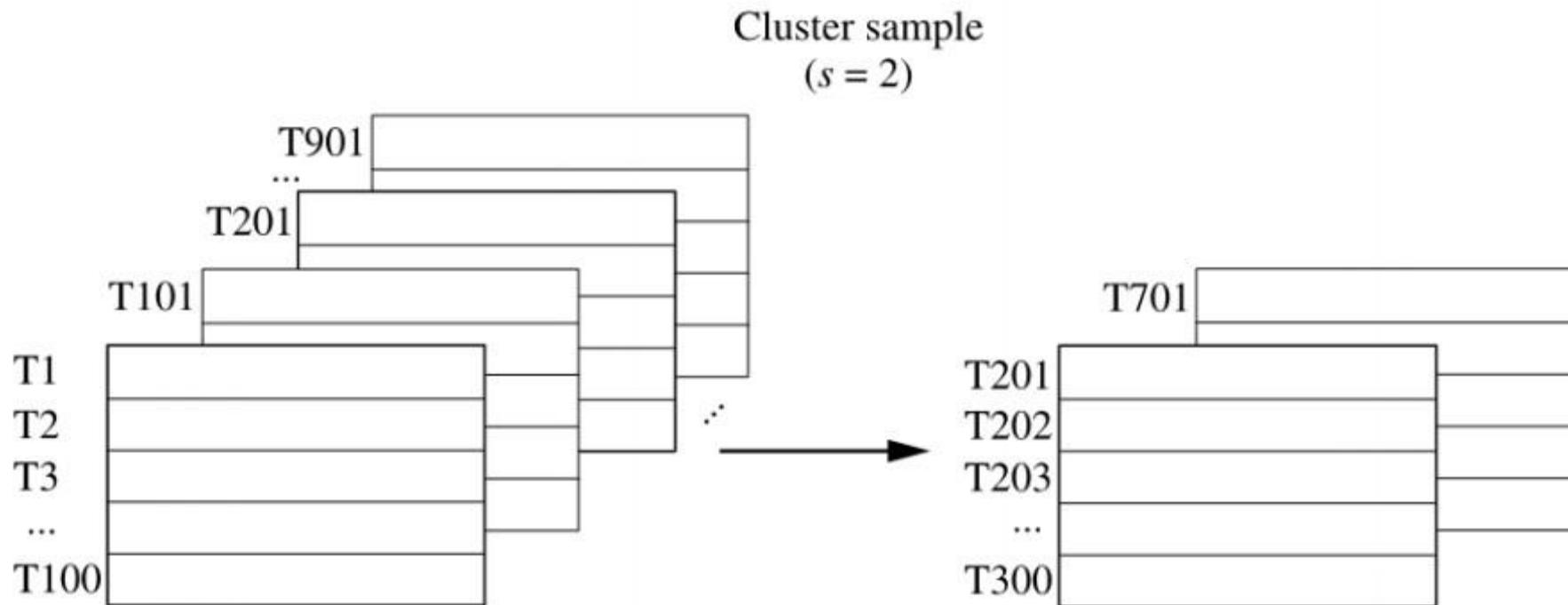
## Thu giảm lượng (numerosity reduction)

- Các phương pháp phi thông số (nonparametric): Sampling



# Thu giảm dữ liệu

- ❑ Thu giảm lượng ( numerosity reduction )
  - Các phương pháp phi thông số (nonparametric): Sampling



# Thu giảm dữ liệu

## Thu giảm lượng ( numerosity reduction)

- Các phương pháp phi thông số (nonparametric):  
Sampling

Stratified sample  
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

## Rời rạc hoá dữ liệu

Giảm số lượng giá trị của một thuộc tính liên tục (continuous attribute) bằng cách 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)

# 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.

# 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”

# Tạo cây 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

# Tạo cây 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 tường minh ở mức lược đồ bởi người sử dụng hoặc chuyên gia
- Đặc tả một phần phân cấp bằng cách nhóm dữ liệu tường minh

# Tạo cây 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

# Tạo cây 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

# Tổng kết

## Xây dựng và đánh giá các mô hình KPDL

- XD mô hình KPDL là một quá trình lặp.
- Cần phải khảo sát nhiều mô hình khác nhau để tìm ra mô hình thích hợp.
- Mô hình có thể là cây quyết định, mạng nơ ron ...
- Việc lựa chọn mô hình sẽ ảnh hưởng đến giai đoạn chuẩn bị dữ liệu.
- VD: mạng nơ ron yêu cầu các giá trị rõ ràng,....

# Tổng kết

## Xây dựng và đánh giá các mô hình KPDL

- XD mô hình KPDL đòi hỏi phải được kiểm thử chặt chẽ nhằm đảm bảo tính chính xác và hiệu quả.
- Quá trình kiểm thử yêu cầu DL phải được chia làm hai phần, phần đầu để XD mô hình, phần sau để kiểm thử.

# Tổng kết

## Triển khai mô hình và thu thập kết quả

- Dùng mô hình để tìm ra các mẫu có ý nghĩa dưới dạng biểu diễn tương ứng với các ý nghĩa đó.
- Các mẫu này phải có khả năng sử dụng tiềm tàng, tức là sau khi xử lý phải dẫn đến những hành động có ích nào đó, được đánh giá bởi một hàm lợi ích.
  - VD: trong dữ liệu các khoản vay, hàm lợi ích đánh giá khả năng tăng lợi nhuận từ các khoản vay. Mẫu khai thác được phải có giá trị với các DL mới với độ chính xác nào đó.

# Tổng kết

## Triển khai mô hình và thu thập kết quả

- Với các giải thuật và các nhiệm vụ của KPDL rất khác nhau, các mẫu chiết xuất được cũng rất đa dạng.
- Mẫu chiết xuất được có thể là một mô tả xu hướng, một hành động.
- Các mẫu có thể liên quan đến các giá trị của các trường trong cùng một bản ghi, VD: Nếu độ ẩm 85% thì dự báo= trời mưa.
- Các mẫu cũng có thể liên quan đến các giá trị tổng hợp từ một nhóm các bản ghi. VD như các khách hàng lớn tuổi thường thích mua quần áo màu xám

# **Trao đổi, câu hỏi?**