



# **T**IỀN XỬ LÝ DỮ LIỆU

---

Giảng viên: Đặng Thị Thu Hiền, Nguyễn Tu Trung  
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2022

# Nội dung

---

- ❖ Tình huống KPDL
- ❖ Khái niệm tiền xử lý dữ liệu
- ❖ Các kỹ thuật tiền xử lý dữ liệu
- ❖ Xây dựng và đánh giá các mô hình KPDL

# Tình huống KPDL

---

❖ Dữ liệu điểm sinh viên: có giá trị NULL

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

# Tình huống KPDL

## ❖ Dữ liệu bị thiếu

	A	B	C	D	E
1	STT	Họ-tên	MSSV	Dữ liệu	Xử lý
2	1	Đỗ Duy Quốc	7140255	Thông tin sản phẩm	cập nhật dữ liệu, tìm kiếm dữ liệu, thống kê dữ liệu
3	2	Trương Thị Mỹ Ngọc	7140830	Thông tin khách hàng, sản phẩm	truy vấn dữ liệu
4	3	Nguyễn Thiện Khánh	7140241	Thông tin khách hàng, bất động sản	đưa dữ liệu có sẵn vào hệ thống hiện tại
5	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
6	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
7	6			Những thông tin của các đối tượng trong thế giới thực	Tạo, ghi, xem, xóa, thêm dữ liệu
8	7	Trần Văn Triết	7140262	Dữ liệu không cấu trúc trong thống kê về web, data warehouse	thống kê dữ liệu
9	8	Lê Nhật 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
10	9	Bùi Tiến Đức			Rút trích dữ liệu để tổng hợp, đánh giá. Từ đó, xây dựng biểu đồ cho sản phẩm
11	10	Trần Ngọc Như Quỳnh	7140256	Dữ liệu bán hàng online	Lọc dữ liệu, thống kê doanh thu, export dữ liệu, ...
12	11	Chu Xuân Tính	7140838	Dữ liệu về thu phí giao thông đường bộ	truy vấn dữ liệu, back up hệ thống dữ liệu, ...
13	12	Lê Nguyễn Dũng	7140224	Dữ liệu âm thanh	phân loại nhạc theo thể loại, dựa vào thông tin hiện trạng các application
14	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)	
15	14			Dữ liệu quản lý học sinh và giáo viên	Thêm, xóa, sửa, cập nhật, bổ sung, thống kê, ...
16	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
17	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 được	Làm sạch và xử lý nhiễu (loại bỏ sự gián đoạn), dự báo, ...
18	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 văn bản
19	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	
20	19	Ấu Mậu Dương	7140820	Dữ liệu về sinh viên, môn học	Tìm sinh viên, thống kê môn học, ...
21	20		13070269	Dữ liệu GIS, dữ liệu giao thông, big data lưu trong database MongoDB, dữ liệu thông tin quản lý bệnh	xác định đường đi ngắn nhất qua 2 điểm, tìm thông tin đối tượng xung quanh
22	21	Đặng Quốc Huỳnh	7140237		Giảm chiều-thu giảm kích thước dữ liệu, gom cụm dữ liệu, phân loại dữ liệu
23	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
24	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
25	24			Dữ liệu về web, logs trên web server	kiểm tra, thống kê, phân tích về thói quen người dùng dựa trên cách người dùng
26	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, ...	
27	26	Lê Minh Châu	7140818	Danh sách nhân viên	import file Excel vào database Oracle/MS SQL Server, thống kê báo cáo lưu trữ

# Tình huống KPDL

---

## ❖ Câu hỏi:

- ❖ \*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 KPDL giáo dục?
- ❖ Tất cả môn học đều được xem xét trong bài toán KPDL giáo dục?
- ❖ Ngoài kết quả điểm số môn học, đặc điểm gì của sinh viên có thể được xem xét trong bài toán KPDL giáo dục?
- ❖ Các giá trị bị thiếu cần xử lý như nào?
- ❖ .....

# Khái niệm tiền xử lý dữ liệu

---

- ❖ Là quá trình xử lý dữ liệu thô/gốc (raw/original data) nhằm cải thiện chất lượng dữ liệu (quality of the data) và chất lượng của kết quả KPD
- ❖ 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/hoặc các hệ thống cơ sở dữ liệu (database systems)
- ❖ Chất lượng dữ liệu (data quality):
  - ❖ Tính chính xác, tính toàn vẹn, tính nhất quán



# Các kỹ thuật tiền xử lý dữ liệu NB

---

- ❖ Tóm tắt mô tả về 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

# Tóm tắt mô tả về dữ liệu

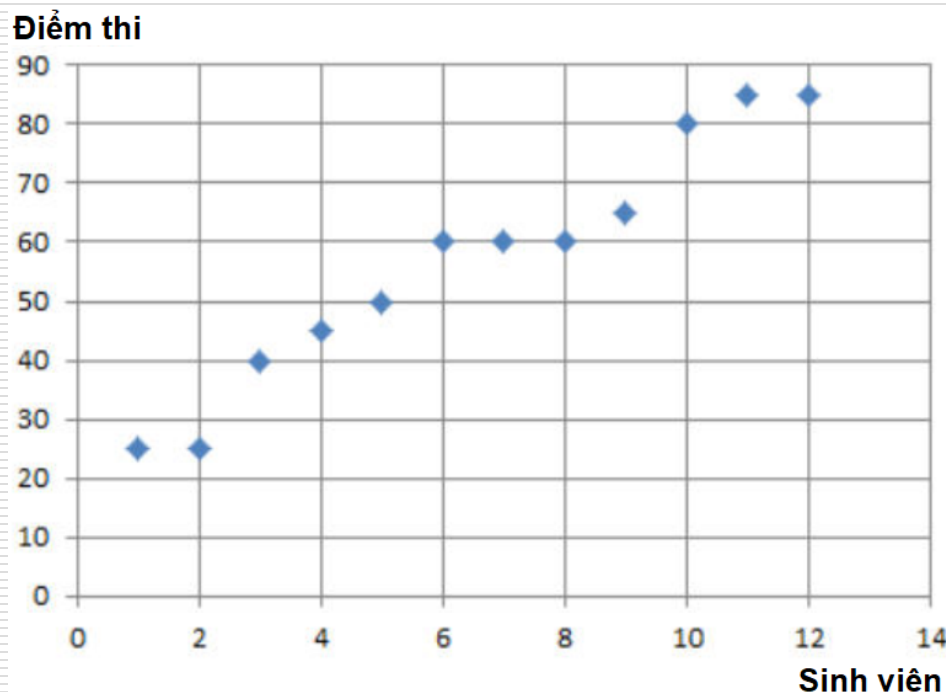
---

- ❖ Tình huống dữ liệu dữ liệu điểm sinh viên
- ❖ Định nghĩa về tóm tắt dữ liệu
- ❖ Các độ đo về xu hướng chính của dữ liệu
- ❖ Xu hướng chính của dữ liệu điểm sinh viên
- ❖ Các độ đo về sự phân tán dữ liệu
- ❖ Tóm tắt mô tả về sự phân bố dữ liệu



# Tình huống tóm tắt dữ liệu điểm sinh viên

## ❖ Dữ liệu về điểm số của sinh viên



## ❖ Câu hỏi:

- ❖ Đặc điểm phân bố và xu hướng của dữ liệu ???
- ❖ Đặc điểm đặc biệt gì khác của dữ liệu ???

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

# Định nghĩa về tóm tắt dữ liệu

---

- ❖ Xác định các thuộc tính (properties) tiêu biểu:
  - ❖ Xu hướng chính (central tendency) của dữ liệu
    - ❖ Các độ đo về xu hướng chính: mean, median, mode, midrange
  - ❖ Sự phân tán (dispersion) của dữ liệu
    - ❖ 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)
  - ❖ Phân tử biên (outliers)
- ❖ => Cung cấp cái nhìn tổng quan về dữ liệu

# Các độ đo về xu hướng chính của dữ liệu

- ❖ Mean:  $\bar{x} = \frac{\sum_{i=1}^N x_i}{N}$
- ❖ Weighted mean:  $\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i}$
- ❖ Median: 
$$Median = \begin{cases} x \left[ \frac{N}{2} \right] & \text{if } N \text{ lẻ} \\ (x \left[ \frac{N}{2} \right] + x \left[ \frac{N}{2} + 1 \right]) & \text{if } N \text{ chẵn} \end{cases}$$
- ❖ 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

# Xu hướng chính của dữ liệu điểm 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

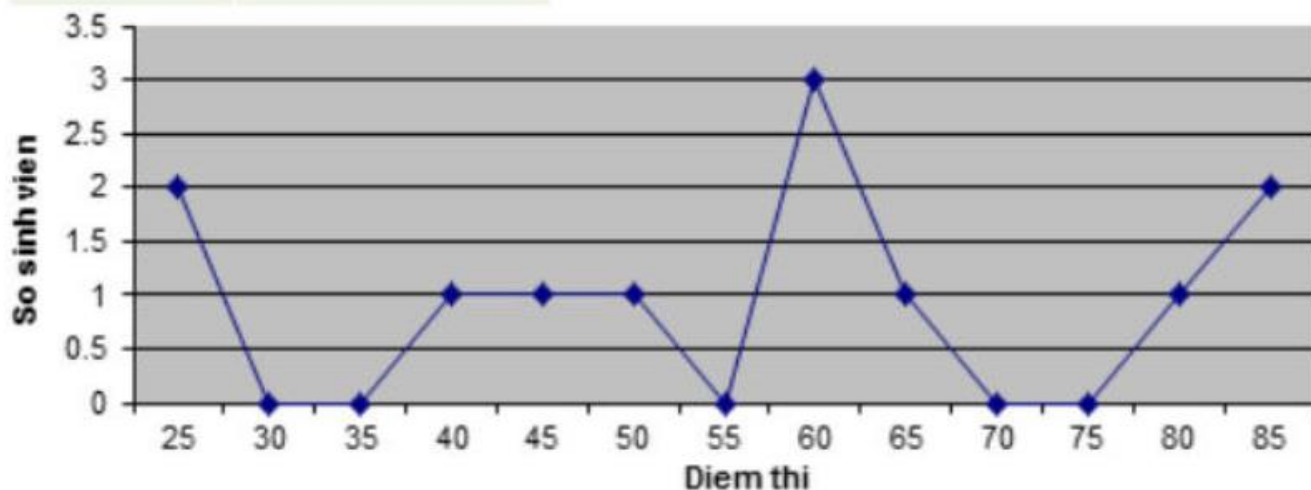
Điểm thi	Số sinh viên
25	2
30	0
35	0
40	1
45	1
50	1
55	0
60	3
65	1
70	0
75	0
80	1
85	2

Mean = 56.67

Median = 60

Mode = 60

Midrange = 55



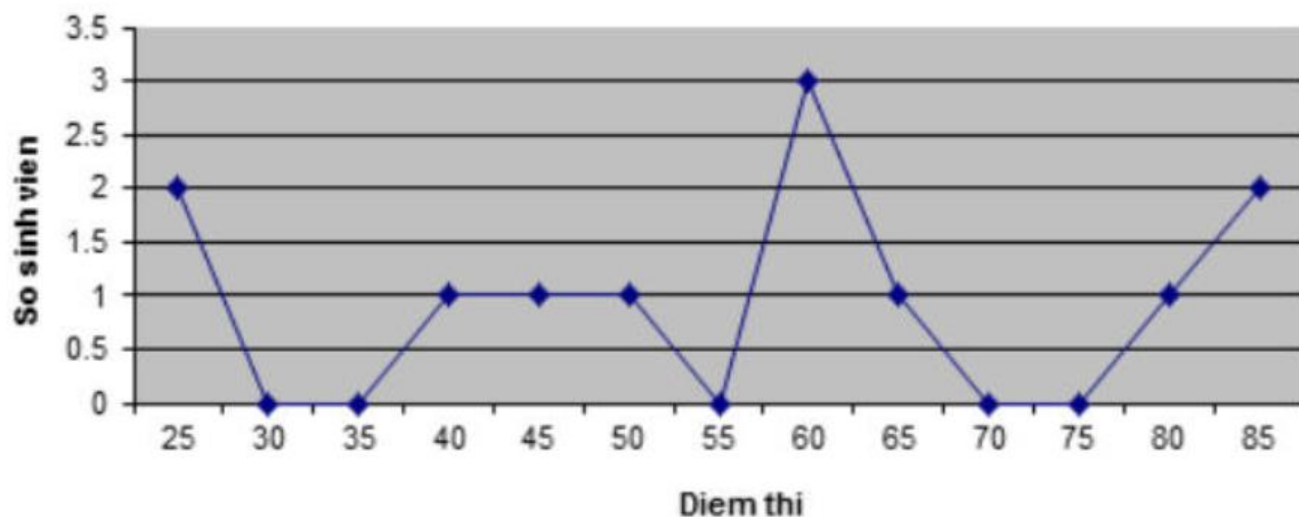
# Các độ đo về sự phân tán dữ liệu

---

- ❖ Các phần tư
  - ❖ Phần tư thứ nhất (Q1): the 25<sup>th</sup> percentile (median of first half)
  - ❖ Phần tư thứ hai (Q2): the 50<sup>th</sup> percentile (median)
  - ❖ Phần tư thứ ba (Q3): the 75<sup>th</sup> percentile (median of second half)
- ❖ Interquartile Range (IQR) = Q3 - Q1
  - ❖ Các yếu tố ngoại lai : giá trị nằm cách trên Q3 hay dưới Q1 một khoảng 1.5 x IQR
- ❖ Phương sai (Variance) 
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

# Sự phân tán dữ liệu điểm 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



$$Q1 = 42.5$$

$$IQR = Q3 - Q1 = 30$$

$$Q2 = \text{median} = 60$$

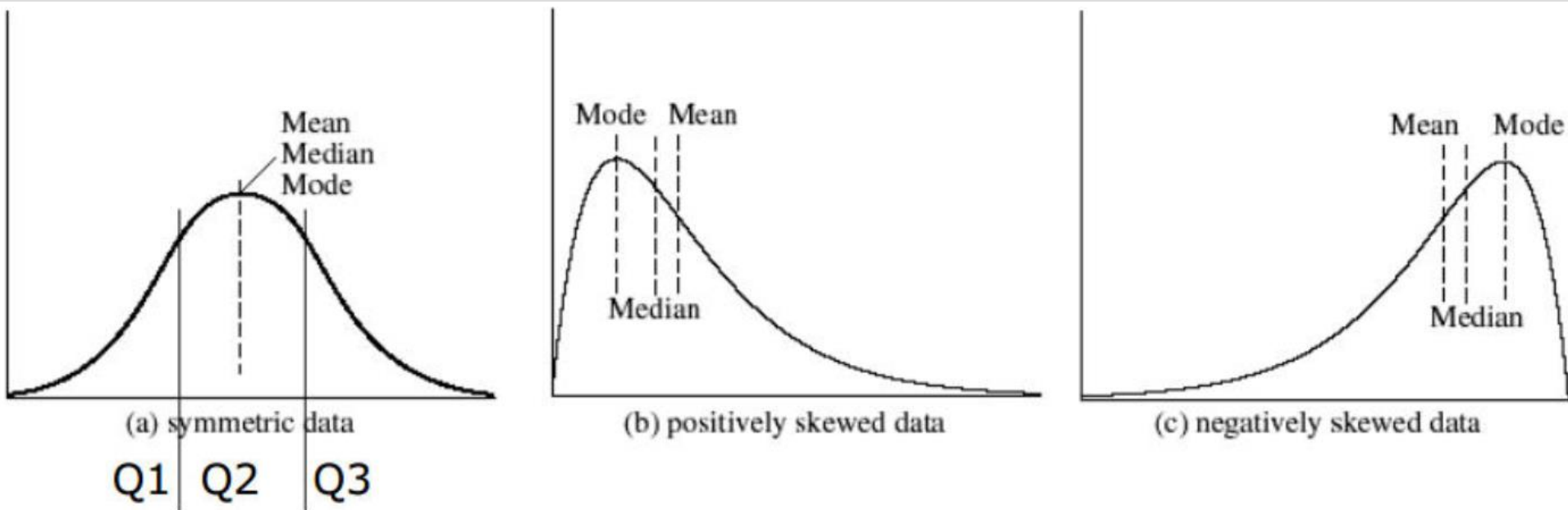
→ Outliers = ???

$$Q3 = 72.5$$

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

$$\sigma = 19.83$$

# Tóm tắt mô tả về sự phân bố dữ liệu



- ❖ a: đối xứng; b,c: nghiêng
- ❖ 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



# 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)

# Xử lý dữ liệu bị thiếu

---

- ❖ Dữ liệu bị thiếu là 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)

# **Nhận diện phần tử biên và giảm thiểu nhiễu**

---

- ❖ Định nghĩa và nguyên nhân
- ❖ Giải pháp nhận diện phần tử biên
- ❖ Giải pháp giảm thiểu nhiễu

# Định nghĩa và nguyên nhân

---

## ❖ Định nghĩa

- ❖ Phần tử biên (Outliers): những dữ liệu (đối tượng) không tuân theo đặc tính/hành vi chung của tập dữ liệu (đối tượng)
- ❖ Nhiễu (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)

# 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)

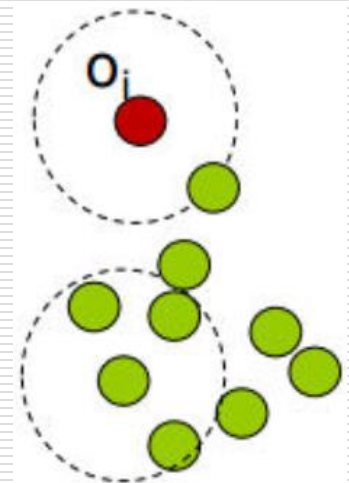
# Nhận diện phần tử biên dựa trên thống kê

---

- ❖ Thủ tục khối (block):
  - ❖ Tất cả các đối tượng tình nghi là outliers hoặc không
- ❖ Thủ tục lần lượt/tuần tự (consecutive/sequential):
  - ❖ Giả sử tập dữ liệu tuân theo một mô hình phân bố  $F$  cho trước (phân bố chuẩn, phân bố Poisson...)
  - ❖ 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

# Nhận diện phần tử biên dựa trên khoảng cách

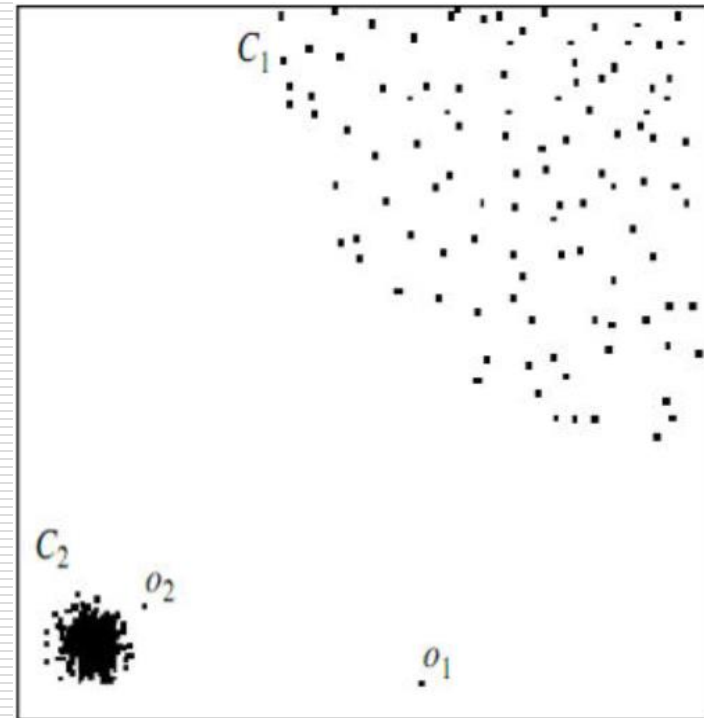
- ❖ Xem xét khoảng cách giữa các đối tượng đến đối tượng tình nghi  $o_i$ :
  - ❖ Nếu ít nhất một lượng đối tượng pct cách đối tượng  $o_i$  xa hơn một khoảng cách  $d_{min}$  thì  $o_i$  là outlier
- ❖ Outlier là những đối tượng không có đủ láng giềng trong khu vực được xác định bởi một khoảng cách cho trước
- ❖ Xác định giá trị pct và  $d_{min}$ 
  - ❖ pct: tỉ lệ số đối tượng không là láng giềng của outliers với tổng đối tượng
  - ❖  $d_{min}$ : minimum distance dùng xác định vùng láng giềng của mỗi đối tượng
- ❖ Ví dụ:  $O_i$  là outlier với pct = 0.8 và  $d_{min} = 1$





# Nhận diện phần tử biên dựa trên mật độ

- ❖ Dựa trên mật độ của vùng láng giềng của mỗi đối tượng
- ❖  $o_1$  và  $o_2$  là phần tử biên dựa trên mật độ
- ❖ Mức độ của outlierness được xác định qua LOF (local outlier factor) của mỗi đối tượng  $p$ 
  - ❖ Phụ thuộc vào mức độ cách ly của  $p$  đối với vùng láng giềng
  - ❖  $k$ -distance của  $p$
  - ❖  $k$ -distance neighborhood của  $p$
  - ❖ Khoảng cách tiếp cận của  $p$  đối với  $o$
  - ❖ Khoảng cách tiếp cận của  $p$
  - ❖  $\Rightarrow \text{LOF}(p)$  càng cao,  $p$  càng được xem là một local outlier



- ❖ Tham khảo: LOF: Identifying Density-Based Local Outliers

# Nhận diện phần tử biên dựa trên độ lệch

---

- ❖ 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
- ❖ Kỹ thuật ngoại lệ tuần tự
  - ❖ Mô phỏng cách con người phân biệt những đối tượng khác biệt (ngoại lệ) khỏi nhóm các đối tượng tương tự nhau
  - ❖ Tập ngoại lệ là tập:
    - ❖ Tập gồm các outliers
    - ❖ Tập con nhỏ nhất mà việc loại bỏ tập con này dẫn đến việc giảm đi nhiều nhất sự khác

# Giải pháp giảm thiểu nhiễu

---

- ❖ Chia ngăn, thùng (Binning)
- ❖ Sử dụng hồi quy (regression)
- ❖ Phân tích cụm (cluster analysis)

# Chia ngăn, thùng (Binning)

- ❖ Đầu vào: Dữ liệu có thứ tự
- ❖ Phân bố dữ liệu vào các bins (buckets) dựa vào số phần tử của bin
- ❖ Làm trơn bin dựa vào
  - ❖ bin means
  - ❖ bin median
  - ❖ bin boundaries: giá trị min hay 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:

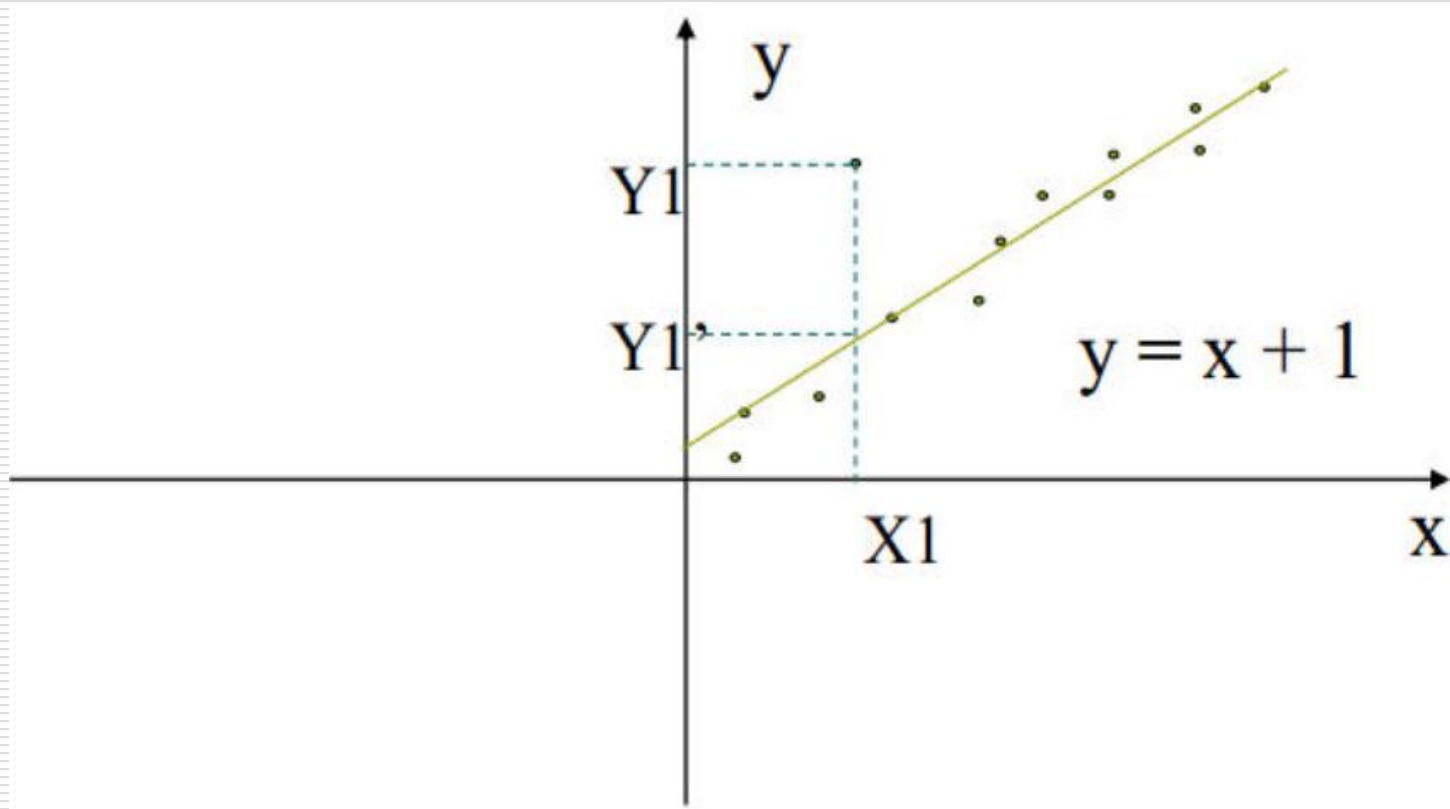
Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

# Sử dụng hồi quy (regression)

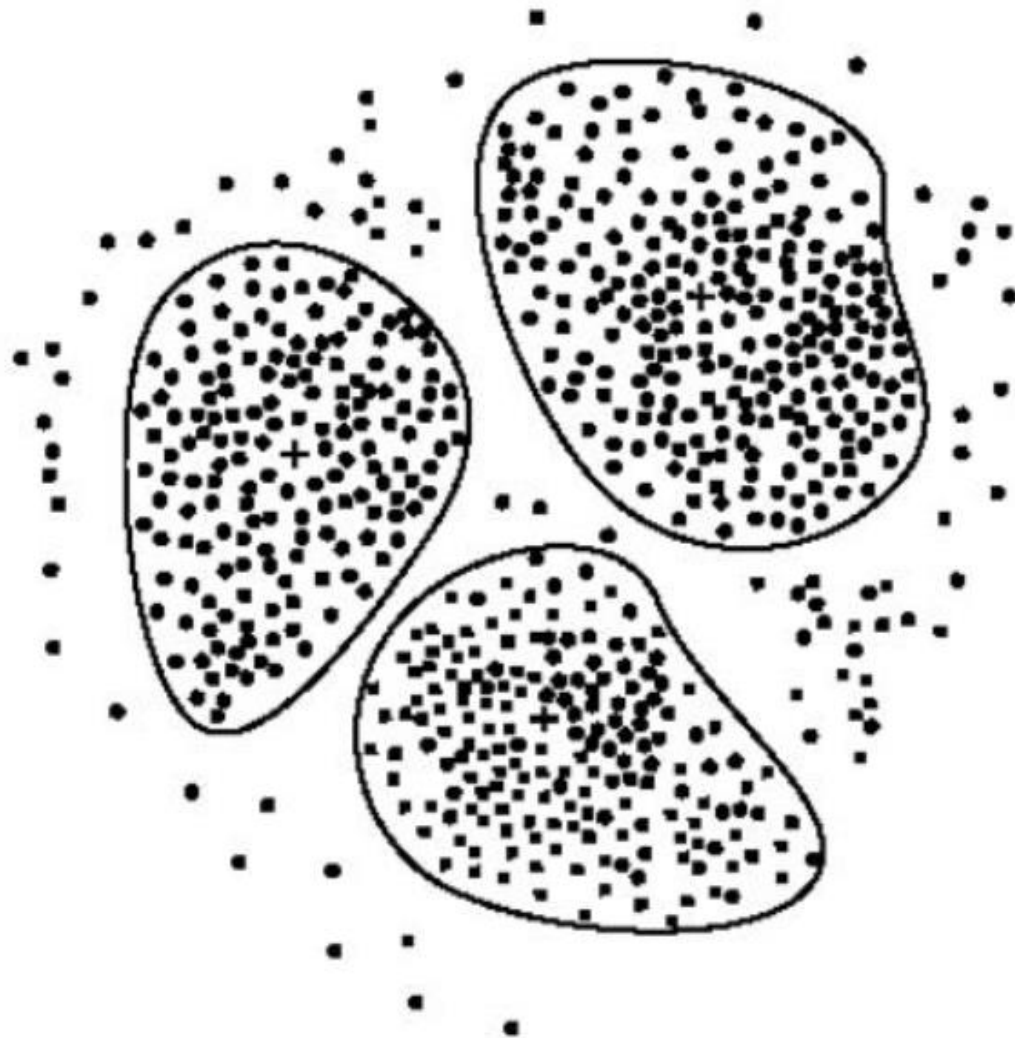
- ❖ Y1 sẽ được hiệu chỉnh lại thành Y1' để phù hợp với phương trình hồi quy tìm được



# Phân tích cụm

---

- ❖ Phần tử ngoại lai nằm ngoài phạm vi cụm



# Xử lý dữ liệu không nhất quán

---

- ❖ Định nghĩa dữ liệu không nhất quán
  - ❖ Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể
    - ❖ Vd1: Định dạng ngày/tháng/năm: 2004/12/25 và 25/12/2004
    - ❖ Vd2: Tên môn học: KPD L, Khai phá dữ liệu, Data mining
  - ❖ Dữ liệu được ghi nhận không phản ánh đúng ngữ nghĩa cho các đối tượng/thực thể
    - ❖ Vd3: Không đúng ràng buộc khóa ngoại



# Xử lý dữ liệu không nhất quán

---

## ❖ 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

## ❖ 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

---

- ❖ Là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu
- ❖ 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
- ❖ Các vấn đề cần giải quyết:
  - ❖ Nhận dạng thực thể (entity identification problem)
    - ❖ Tích hợp lược đồ (schema integration)
    - ❖ So trùng đối tượng (object matching)
  - ❖ Dư thừa (redundancy)
  - ❖ Mâu thuẫn giá trị dữ liệu (data value conflicts)

# 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):
    - ❖ Vd: `customer_id` trong nguồn S1 và `cust_number` trong nguồn S2
  - ❖ Ở mức thể hiện (instance):
    - ❖ Vd: “R & D” trong nguồn S1 và “Research & Development” trong nguồn S2
    - ❖ “Male” và “Female” trong nguồn S1 và “Nam” và “Nữ” trong nguồn S2

# 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 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
  - ❖ Đối với các thuộc tính rời rạc (categorical/discrete attributes): đánh giá tương quan giữa hai thuộc tính với phép kiểm thử chi-square ( $\chi^2$ )

# Phân tích tương quan hai thuộc tính số A,B

- ❖ Công thức 
$$r_{A,B} = \frac{\sum_{i=1}^N (a_i - \bar{A})(b_i - \bar{B})}{N \sigma_A \sigma_B}$$
- ❖  $r_{A,B} \in [-1,1]$
- ❖  $r_{A,B} > 0$ :
  - ❖ 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

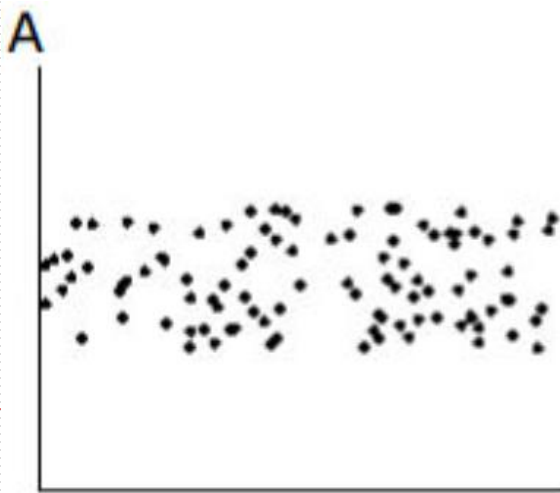
$$\sigma_A = \sqrt{\sigma_A^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

# Phân tích tương quan hai thuộc tính số A,B

❖ Tương quan thuận và nghịch



❖ Không quan sát được tương quan



# Phân tích tương quan hai thuộc tính rời rạc

---

- ❖ Tính giá trị  $X^2$  thực tế
- ❖ Bảng Chi-Square
- ❖ Tra cứu  $X^2$  từ bảng Chi-Square
- ❖ Ví dụ phân tích tương quan 2 thuộc tính rời rạc



# Tính giá trị $X^2$ thực tế

- ❖ 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 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  $(A_i, B_j)$
- ❖  $count(A = a_i)$ : số lượng đối tượng có trị thuộc tính A là  $a_i$
- ❖  $count(B = b_j)$ : số lượng đối tượng có trị thuộc tính B là  $b_j$
- ❖ N: Tổng số đối tượng

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

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

# Bảng Chi-Square

- ❖ Tham khảo: <https://nghiencuugiaoduc.com.vn/bang-phan-phoi-chi-binh-phuong-chi-square-distribution/>
- ❖ Dùng để tra cứu  $X^2$  theo phân phối Chi-Square theo bậc tự do df (degree of freedom) và mức ý nghĩa  $\alpha$  (significance level)

	0.10	0.05	0.02	0.01	0.001
df					
1	2.706	3.841	5.412	6.635	10.828
2	4.605	5.991	7.824	9.210	13.816
3	6.251	7.815	9.837	11.345	16.266
4	7.779	9.488	11.668	13.277	18.467
5	9.236	11.070	13.388	15.086	20.515
6	10.645	12.592	15.033	16.812	22.458
7	12.017	14.067	16.622	18.475	24.322
8	13.362	15.507	18.168	20.090	26.124
9	14.684	16.919	19.679	21.666	27.877
10	15.987	18.307	21.161	23.209	29.588
11	17.275	19.675	22.618	24.725	31.264

# Tra cứu $X^2$ từ bảng Chi-Square

---

- ❖ Phép kiểm thống kê chi-square kiểm tra giả thuyết liệu A và B có độc lập với nhau dựa trên một mức ý nghĩa với độ tự do
- ❖ Nếu giả thuyết bị loại bỏ một trong hai thuộc tính thì A và B có sự liên hệ với nhau dựa trên thống kê
- ❖ Độ tự do:  $df = (r-1)*(c-1)$
- ❖ Tra bảng phân bố chi-square để xác định giá trị  $x^2$
- ❖ Nếu giá trị tính toán được lớn hơn hay bằng giá trị tra bảng được thì hai thuộc tính A và B tương quan với nhau (giả thuyết sai)

# Vd phân tích tương quan 2 thuộc tính rời rạc

- ❖ 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  $X^2$  sẽ kiểm tra giả thuyết liệu gender và preferred\_reading có độc lập với nhau không

# Vd phân tích tương quan 2 thuộc tính rời rạc

- ❖  $o_{11} = 250; o_{12} = 200; o_{21} = 50; o_{22} = 1000$
- ❖  $e_{11} = (\text{count}(\text{male}) * \text{count}(\text{fiction})) / N = (300 * 450) / 1500 = 90$
- ❖  $e_{12} = (\text{count}(\text{female}) * \text{count}(\text{fiction})) / N = (1200 * 450) / 1500 = 360$
- ❖  $e_{21} = (\text{count}(\text{male}) * \text{count}(\text{non\_fiction})) / N = (300 * 1050) / 1500 = 210$
- ❖  $e_{22} = (\text{count}(\text{female}) * \text{count}(\text{non\_fiction})) / N = (1200 * 1050) / 1500 = 840$
- ❖ 
$$X^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 284.44 + 121.90 + 71.11 + 30.48 = 507.93$$
- ❖  $df = (2-1)*(2-1) = 1; \text{sl} = 0.001$
- ❖ Tra bảng:  $X^2 = 10.828 \lll X^2$  tính được từ tập dữ liệu (507.93)  
 $\Rightarrow$  bác bỏ giả thuyết độc lập, gender và preferred\_reading có tương quan với nhau

# Vấn đề mâu thuẫn giá trị dữ liệu

---

- ❖ Cùng một thực thể, các giá trị thuộc tính đến từ các nguồn dữ liệu khác nhau có thể khác nhau về:
  - ❖ Cách biểu diễn (representation)
    - ❖ Ví dụ: "2004/12/25" với "25/12/2004"
  - ❖ Đo lường (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
  - ❖ Mã hóa (encoding)
    - ❖ Ví dụ: "yes" và "no" với "1" và "0"

# Biến đổi dữ liệu

---

- ❖ Định nghĩa:
  - ❖ Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu
- ❖ Bao gồm:
  - ❖ 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)
  - ❖ Thu giảm dữ liệu

# Làm trơn dữ liệu

---

- ❖ 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

---

- ❖ 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 hoá

---

- ❖ Chuyển đổi dữ liệu:
  - ❖ Từ 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 hoá

---

- ❖ Chuẩn hóa min-max
- ❖ Chuẩn hóa z-score
- ❖ Chuẩn hóa bằng chia tỉ lệ thập phân (decimal scaling)
- ❖ => Các giá trị thuộc tính được chuyển đổi vào một miền trị nhất định được định nghĩa trước

# Chuẩn hoá min-max

- ❖ 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 dải 0-4.0 sang dải 0-10.0
- ❖ Công thức:
$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A$$
- ❖ Ví dụ:
  - ❖ Giả sử giá trị nhỏ nhất và lớn nhất cho thuộc tính “thu nhập bình quân” là 500.000 và 4.500.000
  - ❖ Chúng ta muốn ánh xạ giá trị 2.500.000 về khoảng [0.0, 1.0] sử dụng chuẩn hóa min-max
  - ❖ Giá trị mới thu được là:

$$v' = \frac{2.500.000 - 500.000}{4.500.000 - 500.000} (1.0 - 0) + 0 = \frac{2.000.000}{4.000.000} = 0.5$$

# Chuẩn hoá z-score

❖ Giá trị cũ:  $v$  tương ứng với trung bình (mean)  $\bar{A}$  và độ lệch chuẩn (standard deviation)  $\sigma_A$  với thuộc tính A

❖ Giá trị mới:

$$v' = \frac{v - \bar{A}}{\sigma_A} \quad \sigma_A = \sqrt{\sigma_A^2} = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

❖ Ví dụ:

❖ Giả sử thu nhập bình quân có trung bình và độ lệch tiêu chuẩn là: 1.000.000 và 500.000

❖ Sử dụng phương pháp z-score thì giá trị 2.500.000 được ánh xạ thành:

$$v' = \frac{2.500.000 - 1.000.000}{500.000} = \frac{1.500.000}{500.000} = 3$$

# Chuẩn hoá chia tỉ lệ thập phân

---

❖ Giá trị cũ:  $v$

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

❖ Giá trị mới:  $v'$

❖  $j$  là số nguyên nhỏ nhất sao cho  $\text{Max}(|v'|) < 1$

❖ Ví dụ:

❖ Giả sử rằng các giá trị của thuộc tính A được ghi nhận nằm trong khoảng -968 đến 917  $\Rightarrow$  Giá trị tuyệt đối lớn nhất của miền là 968

❖ Chia các giá trị chia cho 1.000 ( $j = 3$ )  $\Rightarrow$  giá trị -968 đổi thành -0.968 và 917 đổi thành 0.917

# Xây dựng thuộc tính/đặc tính

---

- ❖ 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

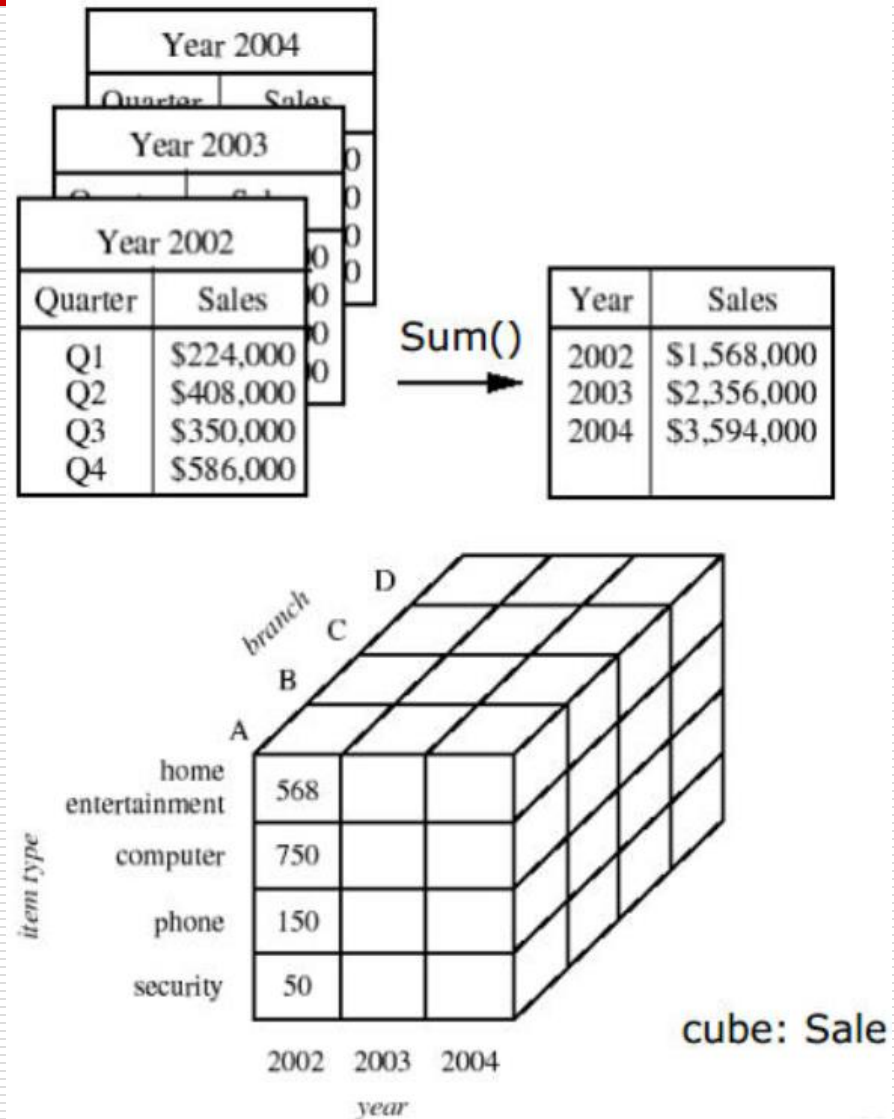
---

- ❖ 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: không mất (lossless) và mất thông tin (lossy)



# Kết hợp khối dữ liệu

- ❖ 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



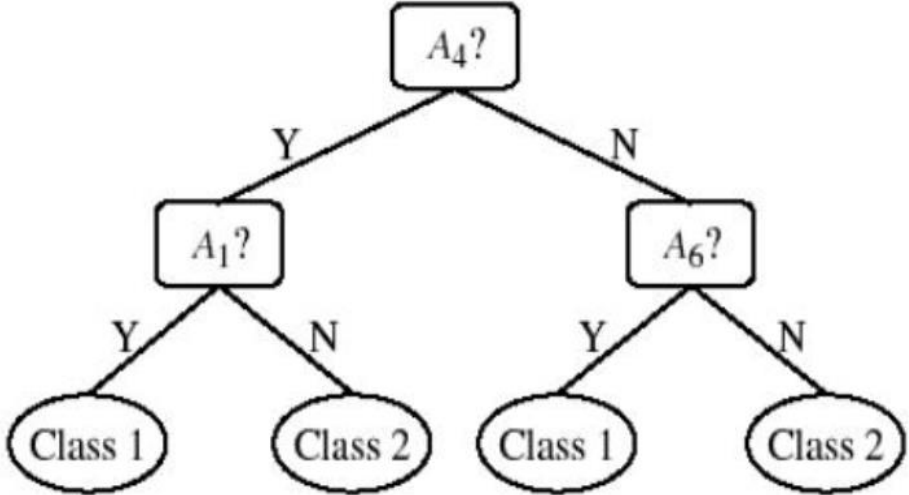
# Chọn một số thuộc tính

---

- ❖ 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 các thuộc tính ít 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

# Chọn một số thuộc tính

## ❖ Một số cách thức:

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>

# Thu giảm chiều

---

- ❖ Biến đổi wavelet (wavelet transforms)
- ❖ Phân tích nhân tố chính (principal component analysis)

# Thu giảm lượng

---

- ❖ Bằng các dạng biểu diễn dữ liệu thay thế
- ❖ Các phương pháp có tham số (parametric):
  - ❖ Mô hình ước lượng dữ liệu => các thông số được lưu trữ thay cho dữ liệu thật
  - ❖ Ví dụ: Hồi quy
- ❖ Các phương pháp phi tham số (nonparametric):
  - ❖ Lưu trữ các biểu diễn thu giảm của dữ liệu
  - ❖ Ví dụ phương pháp:
    - ❖ Histogram, Clustering, Sampling
    - ❖ Simple random sample without replacement (SRSWOR)
    - ❖ Simple random sample with replacement (SRSWR)
    - ❖ Cluster sample
    - ❖ Stratified sample

# Thu giảm lượng

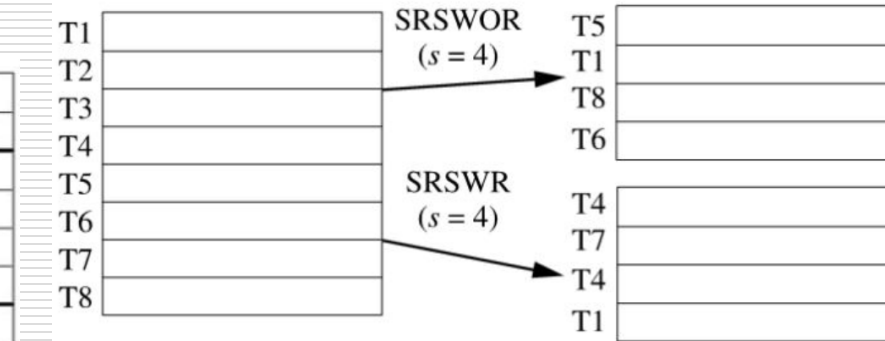
❖ Các phương pháp phi tham số (nonparametric):

❖ Minh họa

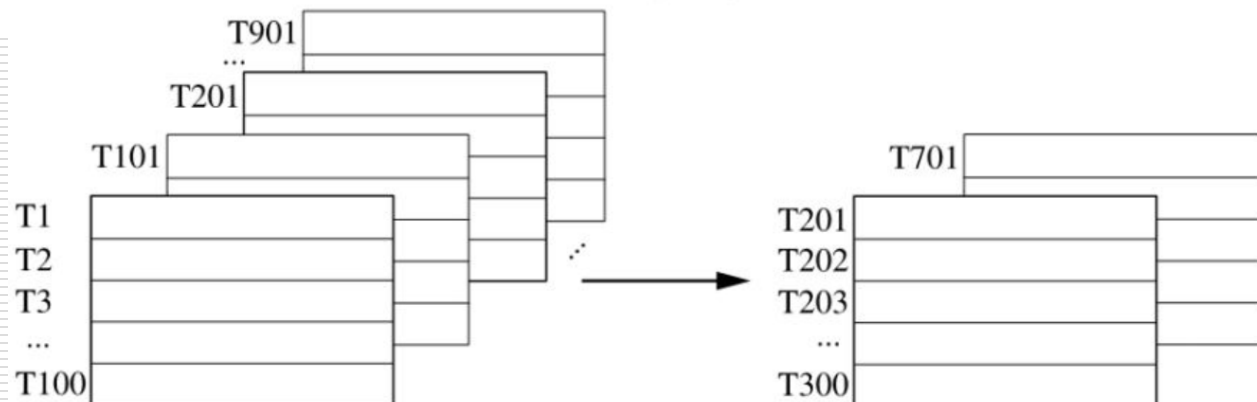
T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

Stratified sample  
(according to age)

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



Cluster sample  
( $s = 2$ )



# Rời rạc hóa dữ liệu

---

- ❖ Nguyên tắc rời rạc hóa dữ liệu
- ❖ Rời rạc hóa dữ liệu thuộc tính số (numeric attributes)
- ❖ Các phương pháp rời rạc hóa dữ liệu thuộc tính số

# Nguyên tắc rời rạc hóa 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 thuộc tính số

---

- ❖ 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
- ❖ Chỉ 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

# Các phương pháp rời rạc hóa dữ liệu thuộc tính số

---

- ❖ Binning
- ❖ Histogram analysis
- ❖ Interval merging (Hợp khoảng) by phân tích tương quan  $x^2$
- ❖ Cluster analysis
- ❖ Entropy-based discretization
- ❖ Discretization by natural/intuitive partitioning (phân vùng tự nhiên/Trực giác)

# Tạo cây phân cấp ý niệm

---

- ❖ Dữ liệu phân loại (categorical data)
  - ❖ Là 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 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
  - ❖ 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

# Xây dựng và đánh giá các mô hình KPD L

---

- ❖ Xây dựng mô hình KPD L 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....
- ❖ Xây dựng mô hình KPD L đòi hỏi phải được kiểm thử chặt chẽ nhằm đảm bảo tính chính xác và hiệu quả
- ❖ Quá trình kiểm thử yêu cầu dữ liệu phải được chia làm hai phần
  - ❖ Phần đầu để xây dựng mô hình
  - ❖ Phần sau để kiểm thử

