

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

TRAN MINH QUANG

quangtran@hcmut.edu.vn

<http://www.cse.hcmut.edu.vn/staff/Staff/quangtran>

<http://researchmap.jp/quang>

1

NỘI DUNG

1. Tổng quan về tiền xử lý dữ liệu
2. Tóm tắt mô tả 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
9. Tóm tắt

TÀI LIỆU THAM KHẢO

- [1] Jiawei Han, Micheline Kamber, and Jian Pei, “Data Mining: Concepts and Techniques”, 3rd Edition, Morgan Kaufmann Publishers, 2012.
- [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] ZhaoHui Tang, Jamie MacLennan, “Data Mining with SQL Server 2005”, Wiley Publishing, 2005.
- [6] Oracle, “Data Mining Concepts”, B28129-01, 2008.
- [7] Oracle, “Data Mining Application Developer’s Guide”, B28131-01, 2008.
- [8] Ian H.Witten, Eibe Frank, “Data mining : practical machine learning tools and techniques”, 2nd Edition, Elsevier Inc, 2005.
- [9] Florent Messegli, Pascal Poncelet & Maguelonne Teisseire, “Successes and new directions in data mining”, IGI Global, 2008.
- [10] Oded Maimon, Lior Rokach, “Data Mining and Knowledge Discovery Handbook”, 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL

- Ví dụ dẫn nhập: thực hiện việc gom cụm sv

MSSV	Mã MH	Năm	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

- Vấn đề về dữ liệu:

- “NULL” nên được diễn dịch như thế nào?
- Miền trị của điểm số là gì? $[0,1]$; $[0,10]$, $\{Y, K, TB, \dots\}$?
- Mọi sv, mọi môn học đều được xem xét?
- Ngoài điểm số, đặc điểm gì của sv cần được xem xét?

1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL

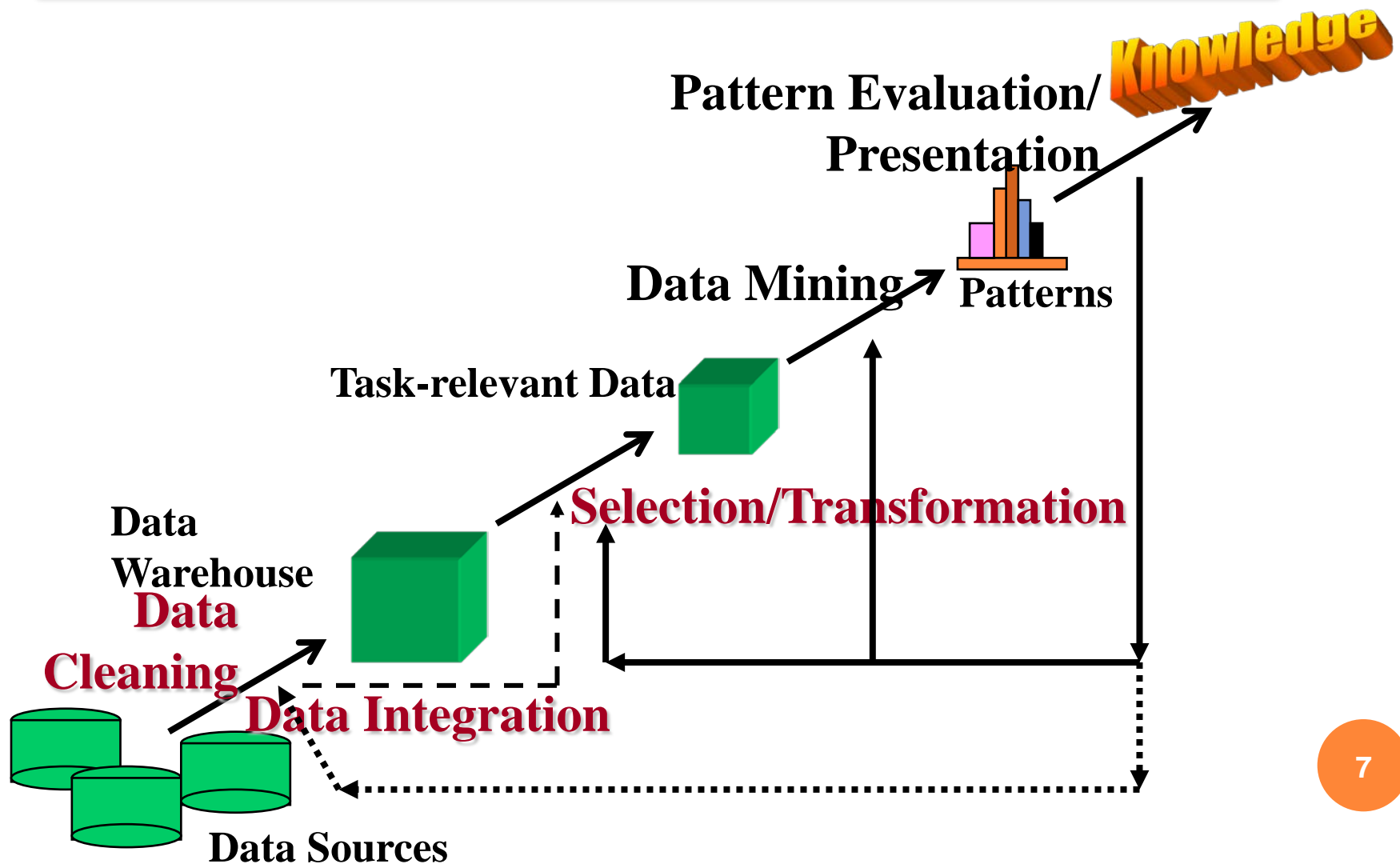
○ Giai đoạn tiền xử lý dữ liệu

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

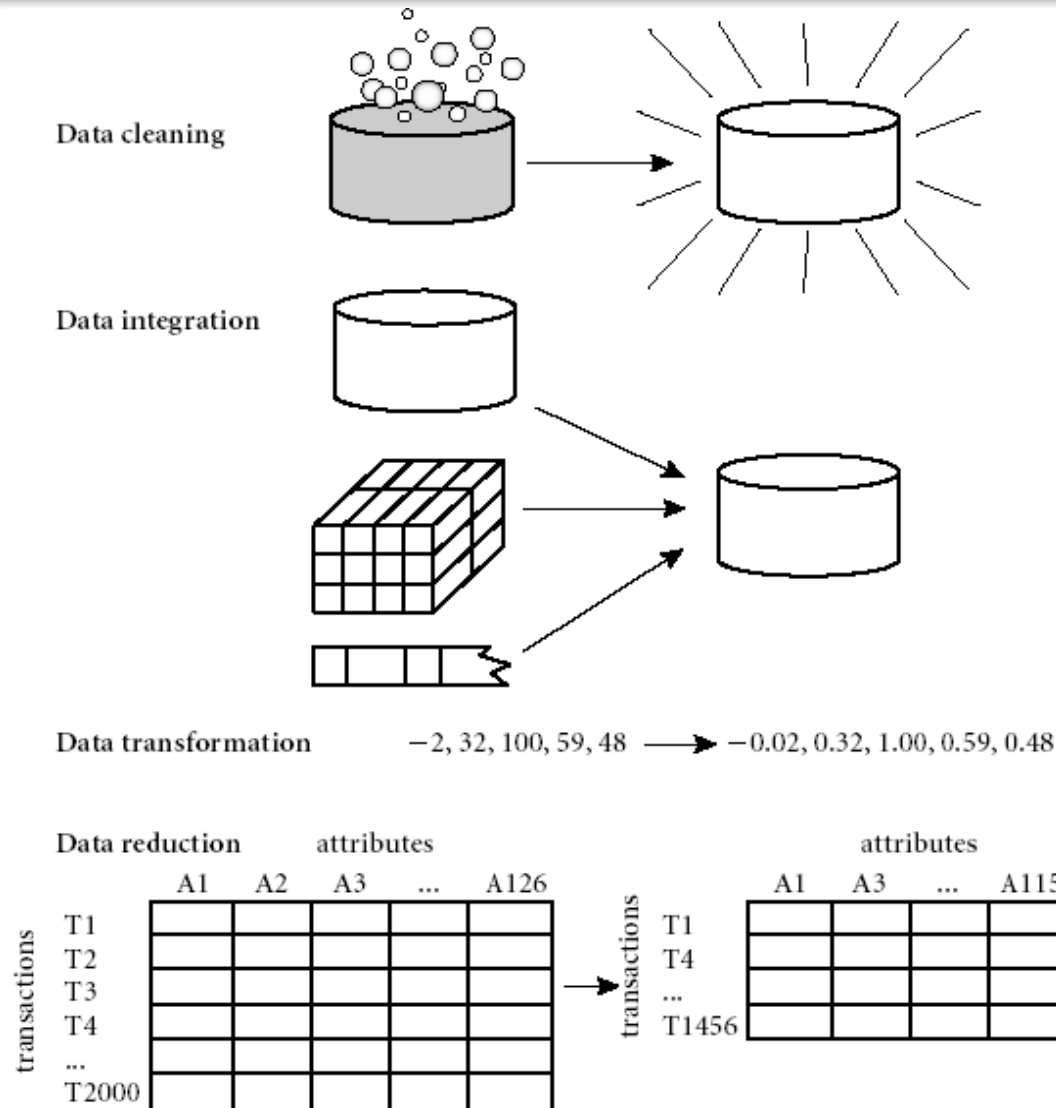
1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL

- 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

1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL



1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL



1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL

○ 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): 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), gom cụm dữ liệu
 - ✓ loại bỏ các đặc điểm dư thừa (redundant features) (nghĩa là giảm số chiều/thuộc tính dữ liệu)

1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL

- Các kỹ thuật tiền xử lý dữ liệu (con't)
 - 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) và 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)

1. TỔNG QUAN VỀ TIỀN XỬ LÝ DL

- Các kỹ thuật tiền xử lý dữ liệu (Con't)
 - 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)

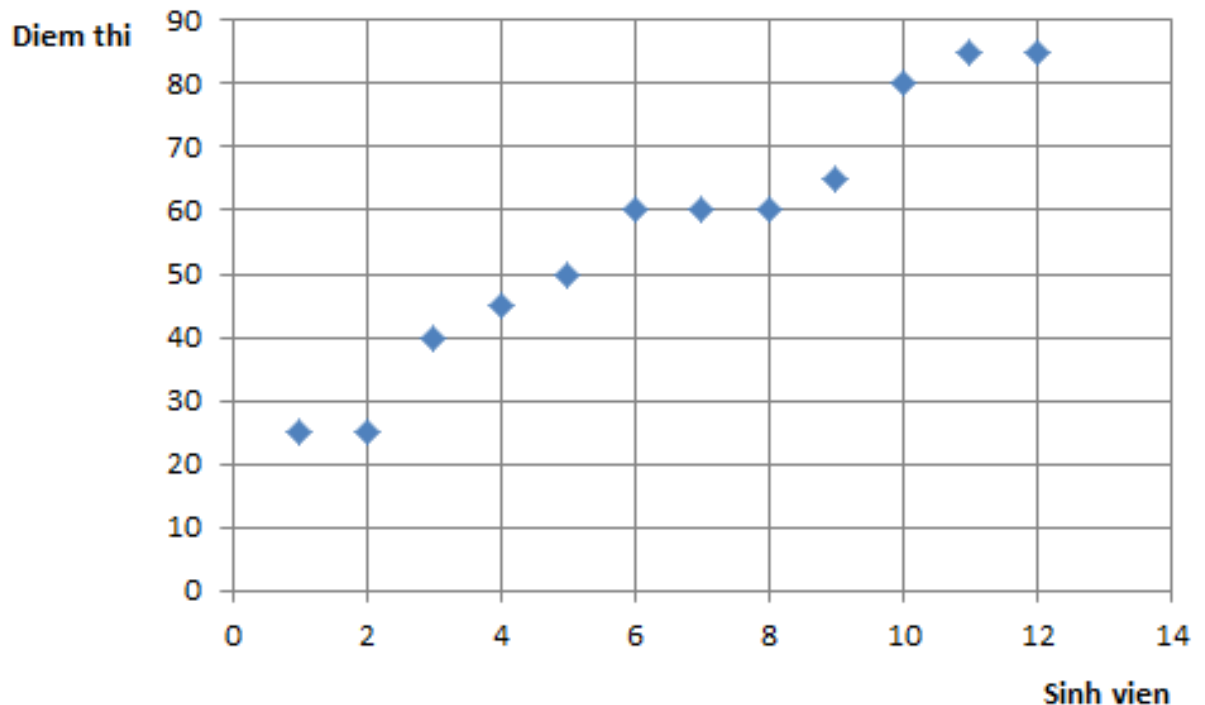
2. TÓM TẮT MÔ TẢ DL

- 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. TÓM TẮT MÔ TẢ DL

- 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?

2. TÓM TẮT MÔ TẢ DL

○ 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 (dữ liệu có thứ tự) $Median = \begin{cases} x_{\lceil N/2 \rceil} & \text{if } N \text{ odd} \\ (x_{N/2} + x_{N/2+1})/2 & \text{if } N \text{ even} \end{cases}$
- 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. TÓM TẮT MÔ TẢ DL

○ Dữ liệu về điểm số của 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

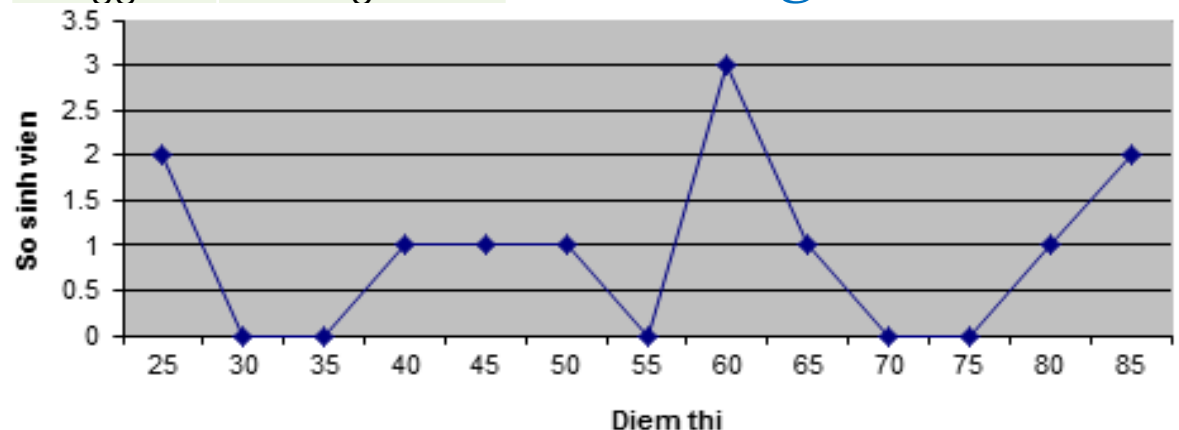
Mean = 56.67

Median = 60

Mode = 60

Midrange = 55

Histogram



2. TÓM TẮT MÔ TẢ DL

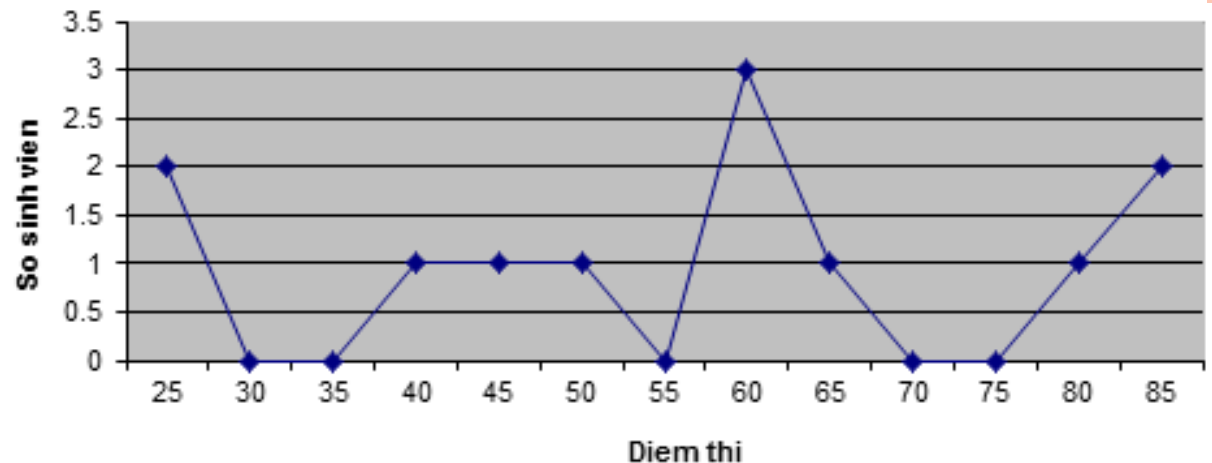
- Các độ đo về sự phân tán của dữ liệu
 - Quartiles
 - ✓ 1st quartile (Q1): the 25th percentile
 - ✓ 2nd quartile (Q2): the 50th percentile (median)
 - ✓ 3rd quartile (Q3): the 75th percentile
 - Interquartile Range (IQR) = Q3 – Q1
 - ✓ Outliers: $\geq Q3 + 1.5 \times IQR$ hay $\leq Q1 - 1.5 \times IQR$
 - ✓ Extreme: $\geq Q3 + 3 \times IQR$ hay $\leq Q1 - 3 \times IQR$
 - Variance

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

2. TÓM TẮT MÔ TẢ DL

- Dữ liệu về điểm số của các sinh viên

Sinh viên	Điểm thi	Độ lệch
1	25	-31.6667
2	25	-31.6667
3	40	-16.6667
4	45	-11.6667
5	50	-6.66667
6	60	3.33333
7	60	3.33333
8	60	3.33333
9	65	8.33333
10	80	23.3333
11	85	28.3333
12	85	28.3333



$$Q1 = 42.5$$

$$IQR = Q3 - Q1 = 30$$

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

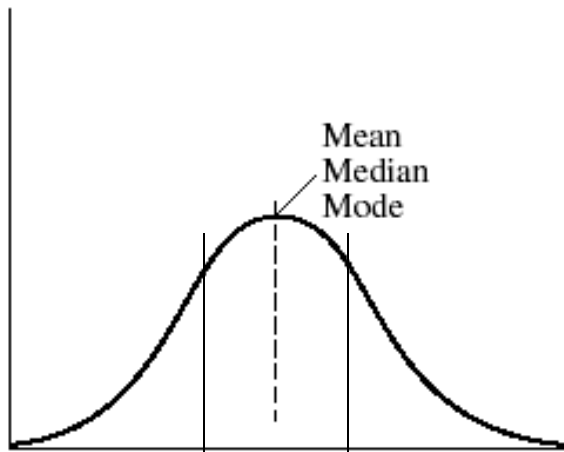
→ Outliers = ???

$$Q3 = 72.5$$

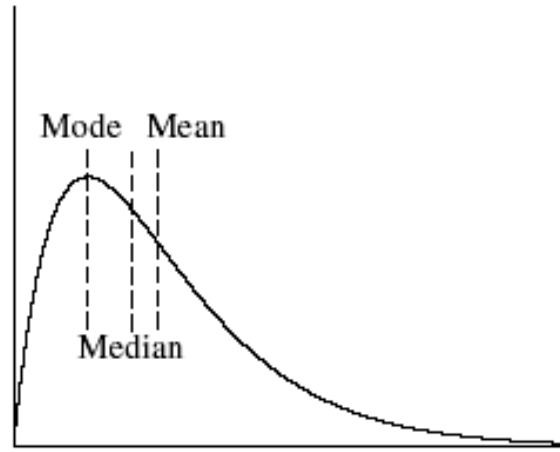
$$\text{Variance} = \sigma^2 = 4310.56$$

$$\sigma = 65.65$$

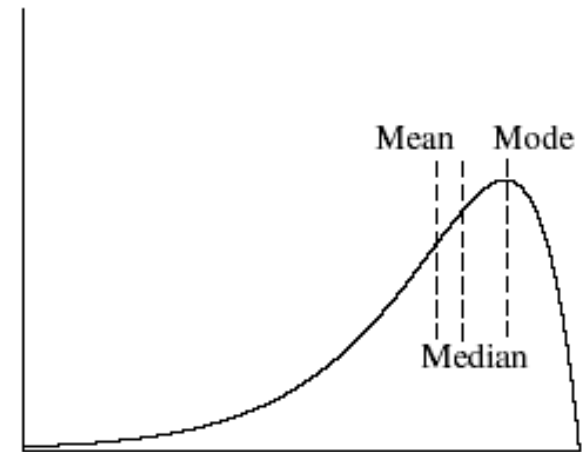
2. TÓM TẮT MÔ TẢ DL



(a) symmetric data



(b) positively skewed data



(c) negatively skewed data

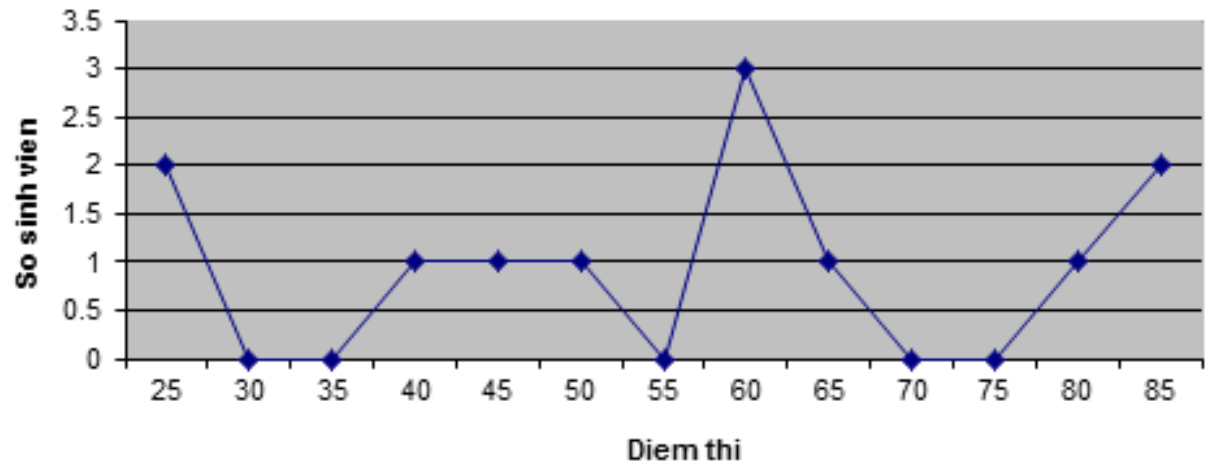
Q1 | Q2 | Q3

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. TÓM TẮT MÔ TẢ DL

○ Dữ liệu về điểm số của 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

3. LÀM SẠCH DỮ LIỆU

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

3.1 XỬ LÝ DỮ BỊ THIẾU

- Dữ liệu bị thiếu (missing data): DL không có sẵn khi cần sử dụng
- Nguyên nhân:
 - ✓ Khách quan: không tồn tại lúc nhập liệu, sự cố,...
 - ✓ Chủ quan: tác nhân con người
- Giải pháp:
 - ✓ Bỏ qua
 - ✓ Cập nhật bằng tay
 - ✓ Dùng giá trị thay thế (tự động): hằng số toàn cục, trị phổ biến, trung bình (toàn cục, cục bộ), trị dự đoán,...
 - ✓ Ngăn chặn từ ban đầu: thiết kế tốt CSDL và các thủ tục nhập liệu (các ràng buộc dữ liệu)

3.2 NHẬN DIỆN PT BIÊN & GIẢM THIỂU NHIỀU

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
- *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)

3.2 NHẬN DIỆN PT BIÊN & GIẢM THIỂU NHIỀU

- 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)

3.2 NHẬN DIỆN PT BIÊN & GIẢM THIỂU NHIỀU

- 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:

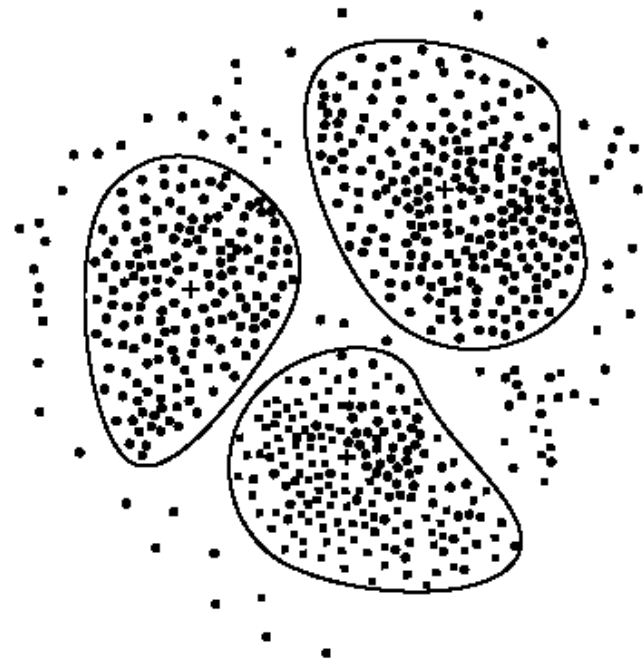
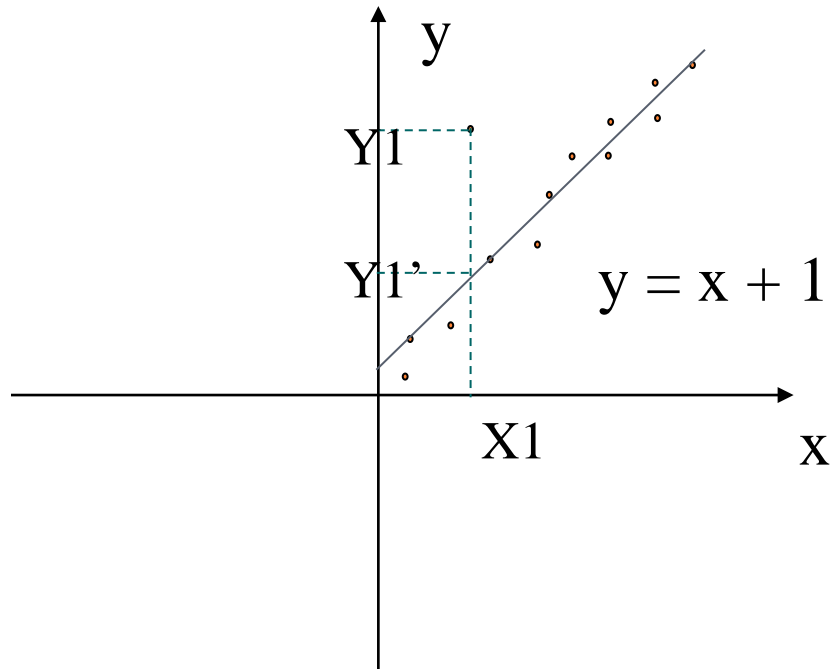
Bin 1: 4, 4, 15

Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

3.2 NHẬN DIỆN PT BIÊN & GIẢM THIỂU NHIỀU

- Giải pháp giảm thiểu nhiễu
 - Hồi quy (regression) và phân tích cụm (cluster analysis)



3.3 XỬ LÝ DL KHÔNG NHẤT QUÁN

○ Khái niệm:

- ✓ Dữ liệu được ghi nhận khác nhau cho cùng một đối tượng/thực thể (discrepancies from inconsistent data representations)

-> Ex. 2004/12/25 và 25/12/2004

- ✓ Dữ liệu được ghi nhận không phản ánh đúng ngữ nghĩa cho các đối tượng/thực thể

-> Ex. Ràng buộc khóa ngoại

○ Nguyên nhân

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

3.3 XỬ LÝ DL KHÔNG NHẤT QUÁN

○ 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

4. TÍCH HỢP DL (DATA INTEGRATION)

- Tích hợp dữ liệu: quá trình tập hợp dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình KPD L
 - 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 đồng nhất (heterogeneity) về ngữ nghĩa (semantics) của dữ liệu
- Hỗ trợ việc giảm, tránh dư thừa và không nhất quán về dữ liệu → cải thiện tính chính xác và tốc độ quá trình KPD L

4. TÍCH HỢP DL (DATA INTEGRATION)

- 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)

4. TÍCH HỢP DL (DATA INTEGRATION)

○ Vấn đề dư thừa

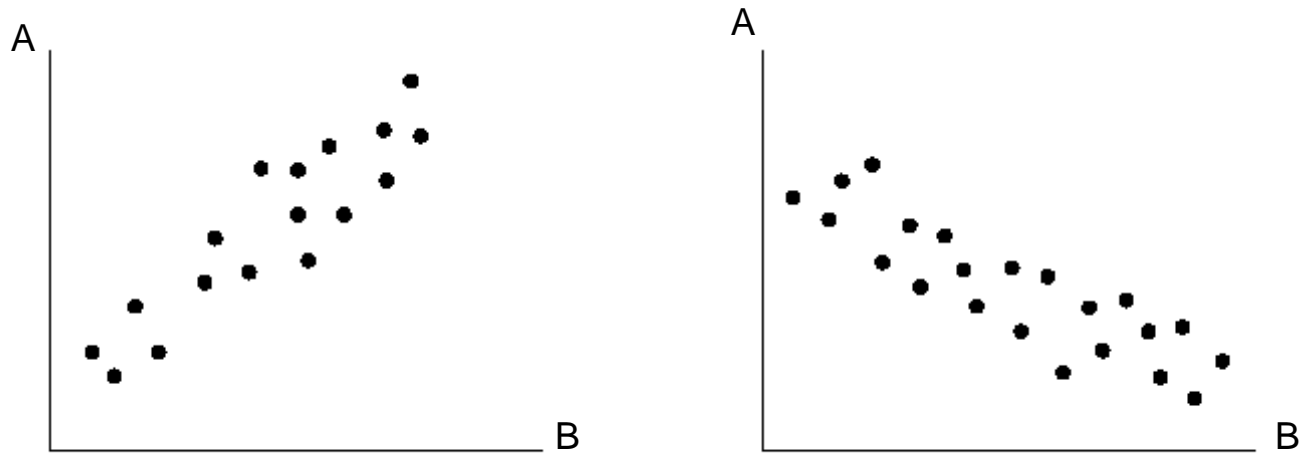
- Hiện tượng: giá trị của một thuộc tính có thể được dẫn xuất từ một/nhiều thuộc tính khác (-> data duplication)
- Nguyên nhân: tổ chức dữ liệu kém, không nhất quán trong việc đặt tên chiều/thuộc tính
- Phát hiện dư thừa: phân tích tương quan (correlation analysis)
 - ✓ Dựa trên dữ liệu hiện có, kiểm 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 attributes), đánh giá tương quan giữa hai thuộc tính với phép kiểm thử chi-square (χ^2)

4. TÍCH HỢP DL (DATA INTEGRATION)

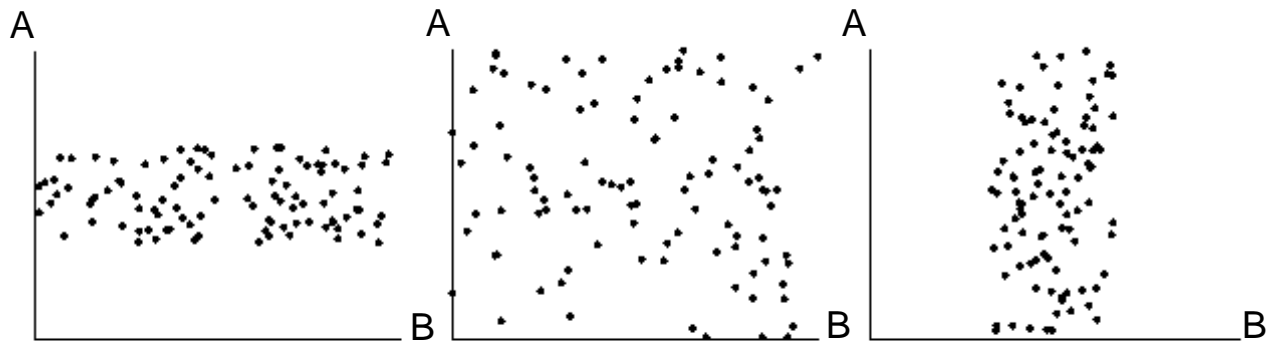
- Phân tích tương quan giữa hai thuộc tính số A và B
 - $r_{A,B} \in [-1, 1]$
 - $r_{A,B} > 0$: A và B tương quan thuận với nhau, trị số của A tăng khi trị số của B tăng, $r_{A,B}$ càng lớn thì mức độ tương quan càng cao, A hoặc B có thể được loại bỏ
 - $r_{A,B} = 0$: A và B không tương quan với nhau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

4. TÍCH HỢP DL (DATA INTEGRATION)

- 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.

4. TÍCH HỢP DL (DATA INTEGRATION)

- 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 .

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{N}$$

4. TÍCH HỢP DL (DATA INTEGRATION)

- 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 χ^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)

4. TÍCH HỢP DL (DATA INTEGRATION)

- 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 χ^2 sẽ kiểm tra giả thuyết liệu *gender* và *preferred_reading* có độc lập với nhau không

4. TÍCH HỢP DL (DATA INTEGRATION)

	<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

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

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

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

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

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

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

Degree of freedom = $(2-1)*(2-1) = 1$; Significance level = 0.001

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

→ bác bỏ giả thuyết độc lập: gender và preferred_reading có tương quan với nhau

4. TÍCH HỢP DL (DATA INTEGRATION)

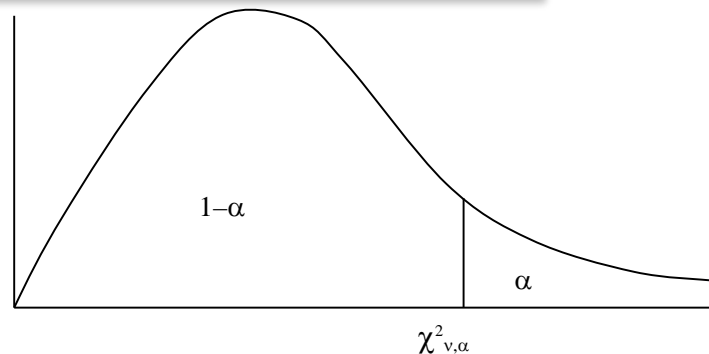
BẢNG TRA PHÂN PHỐI CHI-SQUARE

(Cho $\alpha=0.1$, bậc tự do =6, giá trị Chi-square_{alpha} sẽ là 10.64.

Mật độ XS

Ý nghĩa: $P(\text{Chi-square} > \text{Chi-square}_{\alpha}) = \alpha$

=CHIINV(v,α)



Bậc tự do v	Chi-Square Alpha									
	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	3.93E-05	1.57E-04	9.82E-04	3.93E-03	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.1026	0.2107	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.610	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55
7	0.989	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40

0.001
10.8276

4. TÍCH HỢP DL (DATA INTEGRATION)

- 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: *GPA* : [0, 4] hay [0, 10]; *Price* trong các hệ thống tiền tệ khác nhau
 - ✓ Encoding: “yes” và “no” với “1” và “0”

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

- Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình KPDL
 - 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)

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

- Làm trơn dữ liệu (smoothing)
 - Binning (bin means, bin medians, bin boundaries)
 - Hồi quy
 - Các kỹ thuật gom cụm (phân tích phần tử biên)
 - 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
 - DL chi tiết -> dl tổng hợp (min, max, average, sum,...)
 - Phân tích dữ liệu ở nhiều độ mịn thời gian khác nhau
 - Thu giảm dữ liệu (data reduction)

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

- 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.

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

○ 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.

$$v' = \frac{v - \min A}{\max A - \min A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

○ Chuẩn hóa (normalization)

- z-score normalization

- ✓ Giá trị cũ: v tương ứng với mean \bar{A} và standard deviation σ_A
- ✓ Giá trị mới: v'

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

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

○ Chuẩn hóa (normalization)

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

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

5. BIẾN ĐỔI DL (DATA TRANSFORMATION)

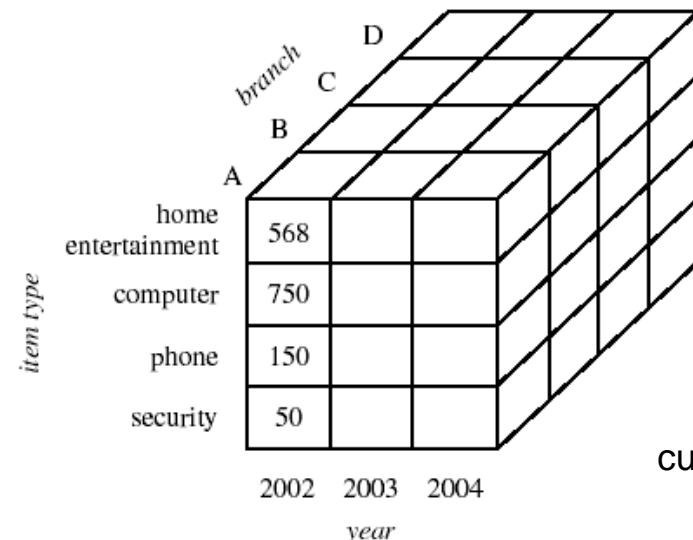
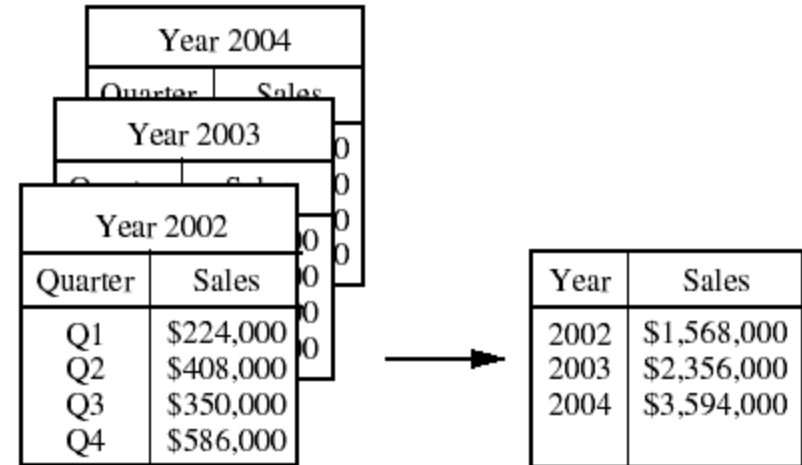
- Xây dựng thuộc tính/đặc tính (attribute/feature construction)
 - Các thuộc tính mới được xây dựng và thêm vào từ tập các thuộc tính sẵn có
 - Hỗ trợ kiểm tra tính chính xác và giúp hiểu cấu trúc của dữ liệu nhiều chiều
 - Hỗ trợ phát hiện thông tin thiếu sót về các mối quan hệ giữa các thuộc tính dữ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

6. THU GIẢM DL

- 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

6. THU GIẢM DL

- Kết hợp khối dữ liệu (data cube aggregation)
 - Dạng dữ liệu: additive, semi-additive (numerical)
 - Data aggregation: average, min, max, sum, count, ...
 - Các mức trừu tượng: mức trừu tượng càng cao giúp thu giảm lượng dữ liệu càng nhiều



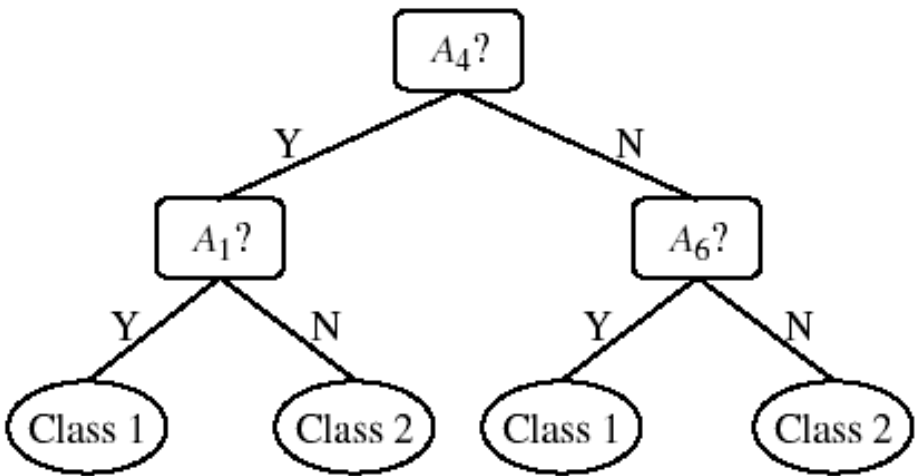
cube: Sale

6. THU GIẢM DL

- Chọn thuộc tính (attribute subset selection)
 - 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

6. THU GIẢM DL

○ Chọn thuộc tính (attribute subset selection)

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

6. THU GIẢM DL

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

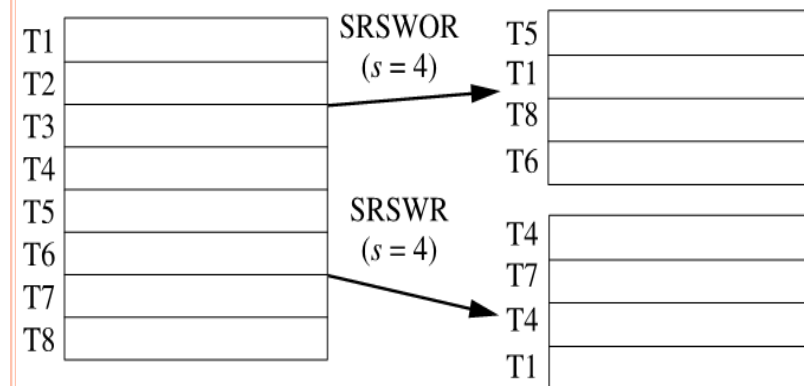
6. THU GIẢM DL

○ Thu giảm lượng (numerosity reduction)

- Các kỹ thuật giảm lượng dl bằng các dạng biểu diễn dl 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**
 - ✓ Simple random sample without replacement (SRSWOR)
 - ✓ Simple random sample with replacement (SRSWR)
 - ✓ Cluster sample
 - ✓ Stratified sample

6. THU GIẢM DL

○ Các phương pháp lấy mẫu (sampling)

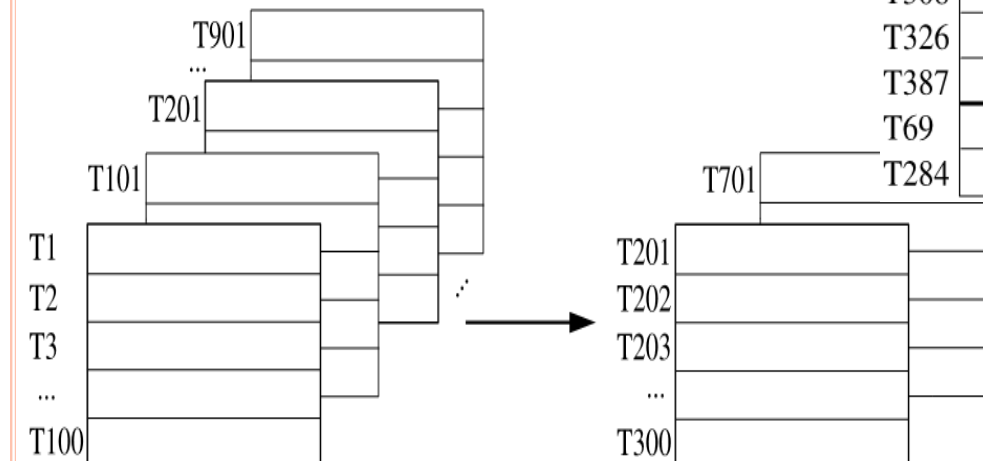


Stratified sample
(according to age)

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

Cluster sample
($s = 2$)



7. RỜI RẠC HÓA DL

- 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 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)

7. RỜI RẠC HÓA DL

- 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

7. RỜI RẠC HÓA DL

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

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

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 phân 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

5. 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

5. TÓM TẮT

○ Rời rạc hóa dữ liệu

- Chia khoảng và gán nhãn cho các thuộc tính có trị liên tục (continuous values)
- 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, χ^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

Q&A

quangtran@hcmut.edu.vn