

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN**

Tài liệu bài giảng:

KHAI THÁC DỮ LIỆU – IS252

Chương 2:

TIỀN XỬ LÝ DỮ LIỆU

ThS. Dương Phi Long – Email: longdp@uit.edu.vn

NỘI DUNG BÀI HỌC

01

Giới thiệu

02

Làm sạch dữ liệu (Data cleaning)

03

Tích hợp dữ liệu (Data integration)

04

Rút gọn dữ liệu (Data reduction)

05

Biến đổi, mã hóa dữ liệu (Data transformation)

1

Giới thiệu



1. Các dạng bộ dữ liệu
2. Đối tượng dữ liệu
3. Thuộc tính
4. Thu thập dữ liệu
5. Chất lượng của dữ liệu
6. Tiền xử lý dữ liệu
7. Các kỹ thuật tiền xử lý dữ liệu

Dữ liệu

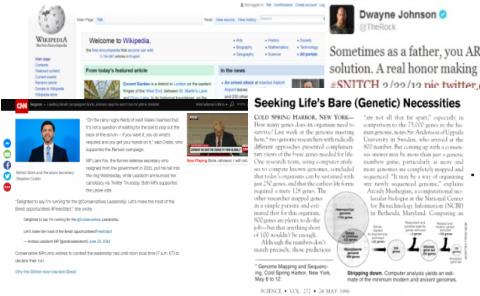
Structured – relational (table-like)

A	B	C	D	E	F	G	
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo Americas)		29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247	37.37	6.02	3.46	72
8	Uzbekistan	Europe	28541	28.9	6.38	2.38	68
9	Uruguay	Americas	3395	22.05	18.59	2.07	77

Un-structured

```
{  
  "code": "1473a6fd39d1d8fa48654aac9d8cc2754232",  
  "title": "[Updating] Câu chuyện xuyên mưa về :  
  "url": "http://techtalk.vn/updating-cau-chuyen-xuyen-mua",  
  "labels": "techtalk/Cong nghe",  
  "content": "Vào chiều tối ngày 09/12/2016 vừa  
  "image_url": "",  
  "date": "2016-12-10T03:51:10Z"  
}
```

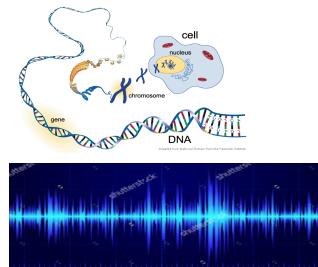
texts in websites, emails, articles, tweets



2D/3D images, videos + meta



spectrograms, DNAs, ...



1. Các dạng bộ dữ liệu (Type of Data sets)

- Record

- Relational records
- Data matrix: numerical matrix, crosstabs
- Document data: text documents term-frequency vector

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

a) Record

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

c) Transaction data

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

b) Data matrix

team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0
Document 2	0	7	0	2	1	0	0	3	0
Document 3	0	1	0	0	1	2	2	0	3

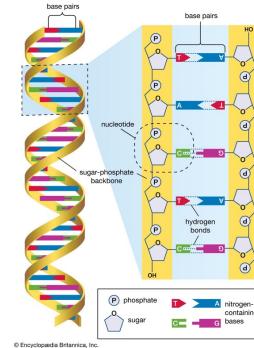
d) Document-term matrix



1. Các dạng bộ dữ liệu (Type of Data sets)

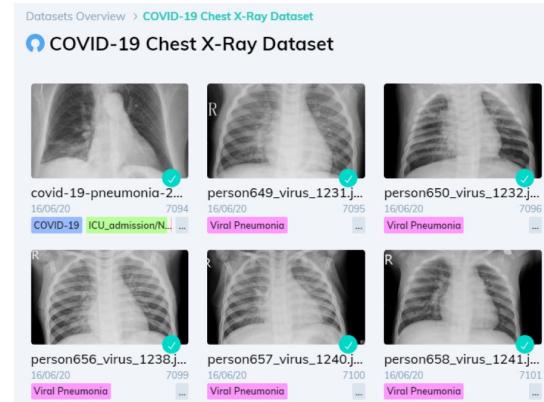
- Ordered

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data



dna_sequence	dna_label
CCGAGGGCTATGGTTTGGAAAGTTAGAACCCCTGGGGCTTCGCGGACACC	0
GAGTTATATGCGCGAGCCCTAGGTTTTGTACTGTTGCGCTCG	0
GATCAGTAGGGAAACAAACAGAGGCCAGCACATCTAGCAGGTAGCCT	0
GTCCACGACCGAACTCCACCTTGAACGCAGAGGTACCCAGAGCCCTG	1
GGCGACCGAACTCCAACTAGAACCTGCATAACTGGCCTGGAGATATGGT	1
AGACATGTCAGAACTTGTGCGCCACTGAGCGACCCGAACCTGGAC	1
CCCGCGAAGGCTGACGAATCCTCGACCCAACTCCAGTGAAGCCAACCGG	1
AGGCAGGTGGTCGTAATGTGTTCAAGAGATAGGGGGCCAGAGCCCTC	0
TACTGCCTATAGCGAAGAGCGCAGAGGTATATCGAAGAATACCGAGCA	0
CGTATCTCGTGTGCTCTCTTTAGAACCTGCATCTCTAGAGTCAGAGGAG	0

e) Genetic sequence data



f) Image data



2. Đối tượng dữ liệu (Data object)

- Data set được tạo thành từ các data object.
- Một data object đại diện cho một thực thể (entity).
- **VD:**
 - CSDL bán hàng: khách hàng, mặt hàng lưu trữ, bán hàng
 - CSDL dữ liệu y tế: bệnh nhân, phương pháp điều trị
 - CSDL trường đại học: sinh viên, giáo sư, khóa học
- Còn được gọi là mẫu (sample, example), thể hiện (instance), điểm dữ liệu (data point), đối tượng (object), bộ giá trị (tuple).
- Các data object được mô tả bởi các thuộc tính (attribute).
- Row ~ data object; column ~ attribute

3. Thuộc tính (Attribute)

- Attribute (hoặc dimension, feature, variable): đặc điểm hoặc đặc trưng của một data object.
- **VD:** Customer_ID, Name, Address
- **Các loại thuộc tính:**
 - Nominal
 - Binary
 - Ordinal
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled



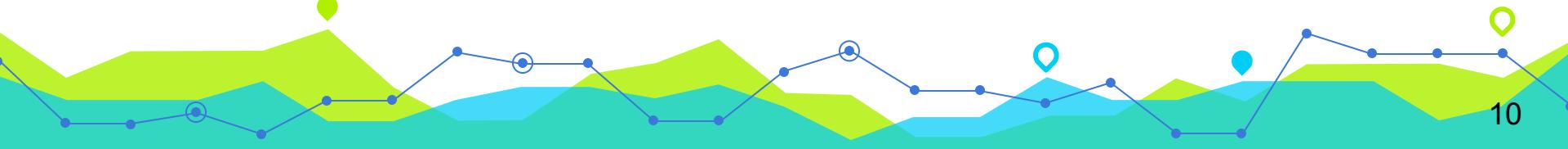
3.1. Các loại thuộc tính

- **Nominal:** danh mục, trạng thái hoặc “names of things”
 - **VD:** Màu tóc = {màu nâu vàng, đen, vàng, nâu, xám, đỏ, trắng}
 - Tình trạng hôn nhân, nghề nghiệp, số ID, zip code
 - **Binary**
 - Thuộc tính nominal chỉ có 2 trạng thái (0 và 1)
 - Nhị phân đối xứng (Symmetric binary): cả hai đều quan trọng như nhau. **VD:** giới tính
 - Nhị phân bất đối xứng (Asymmetric binary): không quan trọng như nhau. **VD:** kết quả xét nghiệm y tế (dương tính, âm tính)
- Quy ước (Convention): gán 1 cho kết quả quan trọng hơn. **VD:** dương tính với HIV

3.1. Các loại thuộc tính

- **Ordinal**

- Các giá trị có thứ tự có ý nghĩa (xếp hạng) nhưng độ lớn giữa các giá trị liên tiếp không được biết.
- **VD:** Kích thước = {nhỏ, trung bình, lớn}, Điểm = {A, B, C, D, E, F}, cấp bậc, quân hàm.



3.1. Các loại thuộc tính

- **Numeric:** định lượng, có thể đo được, giá trị nguyên hoặc thực. 2 loại:
 - Interval-scaled (*Thuộc tính tỷ lệ khoảng thời gian*)
 - Loại dữ liệu được đo theo thang đo, khoảng cách giữa hai điểm được chuẩn hóa và bằng nhau. Không có điểm 0 thực sự
 - Dữ liệu không thể nhân hoặc chia, nhưng có thể trừ hoặc cộng
 - VD: Độ C, Độ F, ngày trong tháng
 - Ratio-scaled (*Thuộc tính tỷ lệ*)
 - Có điểm 0 thực sự
 - Có thể nói một dữ liệu là bội (hoặc tỷ lệ) với một dữ liệu khác
 - VD: number_of_words, số lượng tiền tệ, độ dài, số năm kinh nghiệm

3.2. Thuộc tính rời rạc và Thuộc tính liên tục

- Thuộc tính rời rạc (Discrete Attribute)

- Chỉ có một tập giá trị hữu hạn hoặc vô hạn đếm được
- VD: zip code, nghề nghiệp, tập hợp các từ trong bộ sưu tập tài liệu
- Được biểu diễn dưới dạng số nguyên, số thực
- Thuộc tính binary là trường hợp đặc biệt của thuộc tính rời rạc

- Thuộc tính liên tục (Continuous Attribute)

- Giá trị thuộc tính: số thực
- VD: nhiệt độ, chiều cao hoặc cân nặng
- Trên thực tế, các giá trị thực chỉ có thể được đo lường và biểu diễn bằng số lượng chữ số hữu hạn
- Thường được biểu diễn dưới dạng các biến dấu phẩy động



4. Thu thập dữ liệu

Input
Vấn đề cần giải quyết



Output
Mẫu dữ liệu

A screenshot of a Wikipedia page featuring a table of data and a photograph of a woman wearing a hat.

A	B	C	D	E	F	G
Country	Region	Population	Under15	Over60	Fertil	LifeExp
Zimbabwe	Africa	13724	40.24	5.68	3.64	54
Zambia	Africa	14075	46.73	3.95	5.77	55
Yemen	Eastern M	23852	40.72	4.54	4.35	64
Viet Nam	Western P	90796	22.87	9.32	1.79	75
Venezuela (Bo Americas)		29955	28.84	9.17	2.44	75
Vanuatu	Western P	247				
Uzbekistan	Europe	28541				
Uruguay	Americas					



4. Thu thập dữ liệu

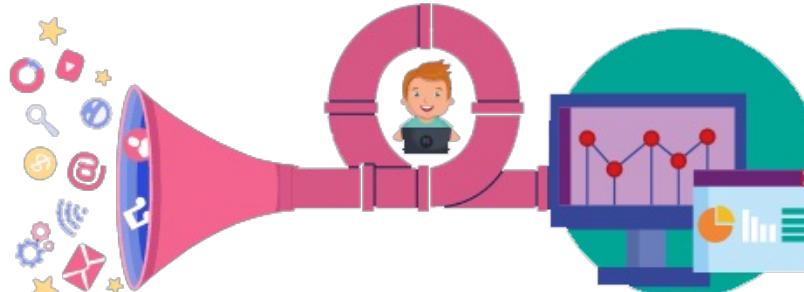
Nguyên tắc lấy mẫu (Sampling):

- **WHAT:** Lấy tập mẫu phổ biến, đại diện cho lĩnh vực cần học, khai thác.
- **WHY:** Không thể học, khai thác toàn bộ. Giới hạn về thời gian và khả năng tính toán.
- **HOW:** Thu thập các mẫu từ thực tế, hoặc từ các nguồn chứa dữ liệu (web, database, ...)

4. Thu thập dữ liệu

Lấy mẫu như thế nào?

- **Variety:** Tập thu được đủ đa dạng để phủ hết các ngữ cảnh, đặc trưng của lĩnh vực.
- **Bias:** Dữ liệu cần tổng quát, cân bằng, không bị sai lệnh, thiên vị về 1 bộ phận nhỏ nào đó của lĩnh vực.

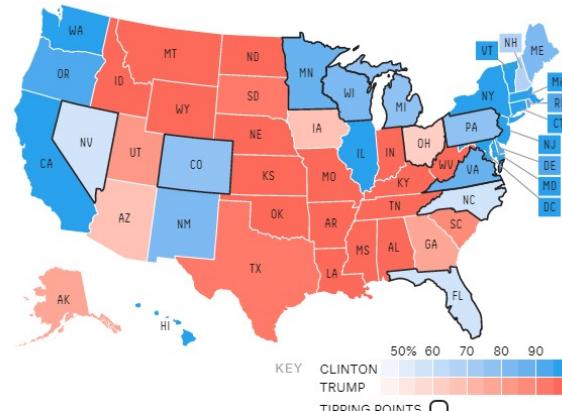


4. Thu thập dữ liệu

Variety & Bias

Dữ liệu đa dạng,
cân bằng để phản
 ánh khách quan?

Chance of winning

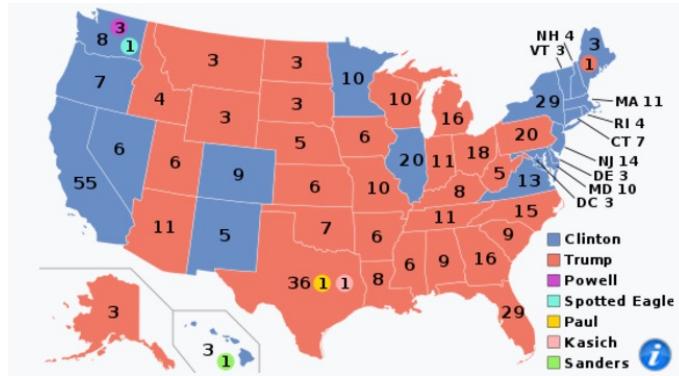


Electoral votes

■ Hillary Clinton	302.2
■ Donald Trump	235.0

Popular vote

■ Hillary Clinton	48.5%
■ Donald Trump	44.9%



<https://projects.fivethirtyeight.com/2016-election-forecast/>
<http://edition.cnn.com/election/results/president>
Image credit: Wikipedia, FiveThirtyEight

4. Thu thập dữ liệu

Các kỹ thuật thu thập dữ liệu:

- **Crow-sourcing:** Survey – các khảo sát.
- **Logging:** Lưu lại lịch sử tương tác, truy cập của người dùng, ...
- **Scraping:** Lưu lại dữ liệu từ các website



SURVEYS



17

4. Thu thập dữ liệu

Ví dụ - Scrapping:

- **Mục tiêu:** Dữ liệu cho bài toán phân lớp văn bản (dữ liệu báo chí).
- **Hướng giải quyết:** Hệ thống thu thập (crawl) dữ liệu báo

Input

Vấn đề: phân lớp văn bản báo chí



Output

Mẫu dữ liệu: Bài báo và nhãn tương ứng

Name	Date modified
1 Dân trí	
2 Ban đọc	5/25/2018 4:46 PM
3 Đời sống	5/25/2018 4:46 PM
4 Giáo dục - Khuyển học	5/25/2018 4:46 PM
5 Khoa học - Công nghệ	5/25/2018 4:46 PM
6 Nhịp sống trê	5/25/2018 4:46 PM
7 Sức khỏe	5/25/2018 4:46 PM
8 Sức mạnh sô	5/25/2018 4:46 PM
9 Thể thao	5/25/2018 4:46 PM
10 Văn hóa	5/25/2018 4:46 PM
11 techtalk	
12 vietbao	
13 vnexpress	
14 vtv	

1	{	"date": "2018-05-20, 07:44:",
2	"code":	"65labzf5f0305220d
3	"labels":	"D\u000e2n tr\u000e1v\u000e0"
4	"content":	
5	"image_url":	"http://dantri.com.v
6	"url":	"http://dantri.com.vn"
7	"domain":	"dantri.com.vn"
8	"title":	"B\u00f9uleafc Giang: \\"
9		
10		



5. Chất lượng của dữ liệu

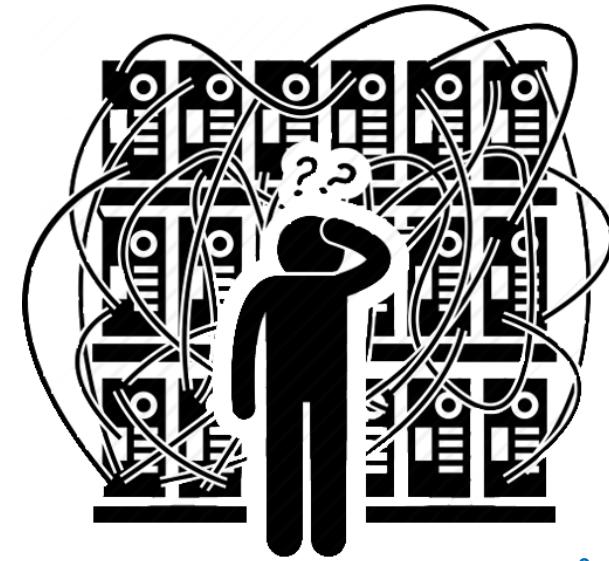
Các yếu tố đánh giá:

- **Tính chính xác (Accuracy):**

- Có các giá trị không chính xác
- Có thể là lỗi của con người hoặc máy tính

- **Tính đầy đủ, toàn vẹn (Completeness):**

- Dữ liệu không đầy đủ có thể xảy ra
- VD: thông tin khách hàng cho dữ liệu bán hàng & giao dịch có thể không phải lúc nào cũng có sẵn.



5. Chất lượng của dữ liệu

Các yếu tố đánh giá:

- **Tính nhất quán (Consistency):**
 - Có thể do quy ước đặt tên, đặt mã, định dạng không nhất quán
 - Các bộ dữ liệu trùng lặp.
- **Tính kịp thời (Currency/ Timeliness):** Dữ liệu được cập nhật đầy đủ và kịp thời?
- **Độ tin cậy (Believability):** Mức độ người dùng tin tưởng vào dữ liệu
- **Khả năng diễn giải (Interpretability):** Mức độ dễ hiểu của người dùng đối với dữ liệu.



5. Chất lượng của dữ liệu

Trên thực tế, dữ liệu có chất lượng không tốt

- **Dữ liệu thiếu, không đầy đủ:** Thiếu giá trị của thuộc tính, thiếu các thuộc tính quan tâm hoặc chỉ chứa dữ liệu tích hợp
VD: Tuổi, cân nặng = "".
- **Dữ liệu bị tạp, nhiễu (noise):** lỗi hoặc sai
VD: Lương = -100.000
- **Dữ liệu mâu thuẫn, không thống nhất**
VD: Tuổi = 23, Ngày sinh 01/01/2023; US = USA?; Rating “1, 2, 3” “A, B, C”



6. Tiền xử lý dữ liệu

- Giúp ích trong việc lưu trữ, truy vấn
- Các mô hình học máy thường làm việc với dữ liệu có cấu trúc: ma trận, vector, chuỗi, ...
- Học máy thường làm việc hiệu quả nếu **biểu diễn dữ liệu phù hợp**

Input

Mẫu dữ liệu thô (text, image, audio,...)

A	B	C	D	E	F	G	
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247				
8	Uzbekistan	Europe	20544				
9	Uruguay	Americas					



Output

Dữ liệu số theo từng model (AI/ ML)

$$x^{(n)} = \begin{matrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{matrix}$$

$$\mathcal{D} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(n)} \end{bmatrix}$$



Tiền xử lý dữ liệu – Cần lưu ý

Completeness (Đầy đủ)

Từng mẫu thu thập nên đầy đủ thông tin các trường thuộc tính cần thiết

Homogeneity (Đồng nhất)

Thu thập dữ liệu từ nhiều nguồn

Rating “1,2,3” & “A,B,C”
Age = “23” & Birthday = “01/01/2023”

Integrity (Trung thực)

Nguồn thu thập chính thống, đảm bảo mẫu thu được chứa giá trị chính xác trên thực tế.

Structures (Cấu trúc)

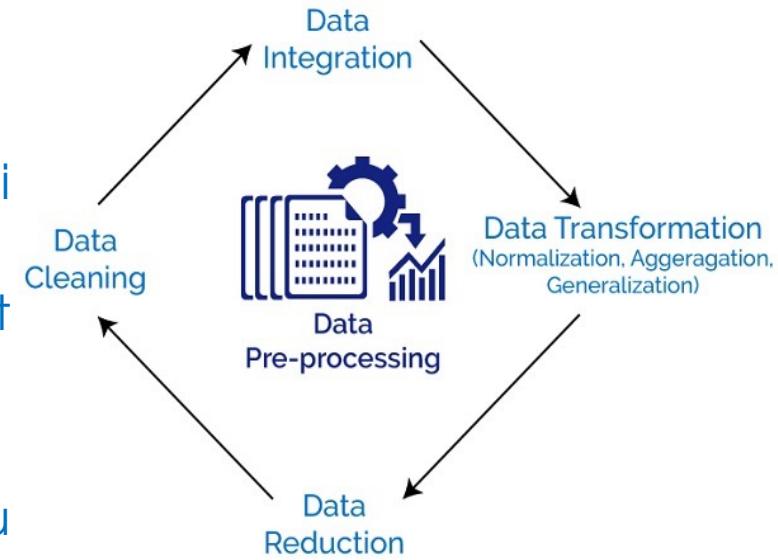
C	D	E	F	G
13724	40.24	5.68	2.64	54
14075	46.73	3.95	5.77	55
23852	40.72	4.54	4.35	64
90796	22.87	9.32	1.79	75
29955	28.84	9.17	2.44	75
247	37.37	6.02	3.46	72
28541	28.9	6.38	2.38	68
3395	22.05	18.59	2.07	77

	X ₁	X ₂	X ₃	X ₄	X ₅	class
x ₁	5.5	3.0	4.2	1.5		Iris-versicolor
x ₂	6.9	3.1	4.9	1.5		Iris-versicolor
x ₃	6.6	2.9	4.6	1.3		Iris-versicolor
x ₄	4.6	3.2	1.4	0.2		Iris-setosa
x ₅	6.0	2.2	4.0	1.0		Iris-versicolor
x ₆	4.7	3.2	1.3	0.2		Iris-setosa
x ₇	6.5	3.0	5.8	2.2		Iris-virginica
x ₈	5.8	2.7	5.1	1.9		Iris-virginica
x ₉	7.7	3.8	6.7	2.2		Iris-virginica
x ₁₀	5.1	3.4	1.5	0.2		Iris-setosa



6. Các kỹ thuật tiền xử lý dữ liệu

- **Làm sạch dữ liệu (Data cleaning)**
 - Điền vào các giá trị còn thiếu
 - Khử dữ liệu nhiễu
 - Xác định hoặc loại bỏ các giá trị ngoại lệ, sai lệch
 - Giải quyết các dữ liệu không nhất quán, mâu thuẫn
- **Tích hợp dữ liệu (Data integration)**:
 - Tổng hợp, tích hợp dữ liệu từ nhiều CSDL, khối dữ liệu hoặc tập tin

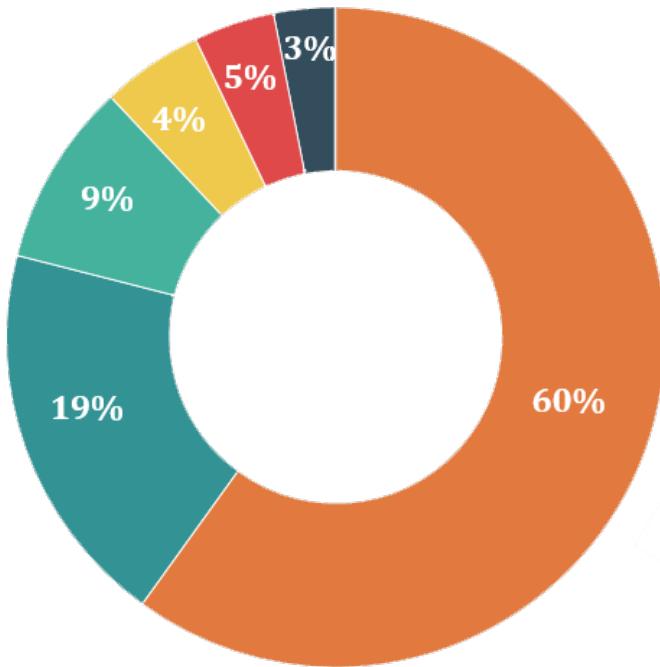


6. Các kỹ thuật tiền xử lý dữ liệu

- **Rút gọn dữ liệu (Data reduction)**
 - Giảm chiều dữ liệu (Dimensionality reduction)
 - Giảm số lượng (Numerosity reduction)
 - Nén dữ liệu (Data compression)
- **Biến đổi và rời rạc hóa dữ liệu (Data transformation, data discretization):**
 - Chuẩn hóa dữ liệu (Normalization)
 - Hệ thống khái niệm phân cấp (Concept hierarchy generation)



Thời gian dành cho phân tích dữ liệu ra sao?

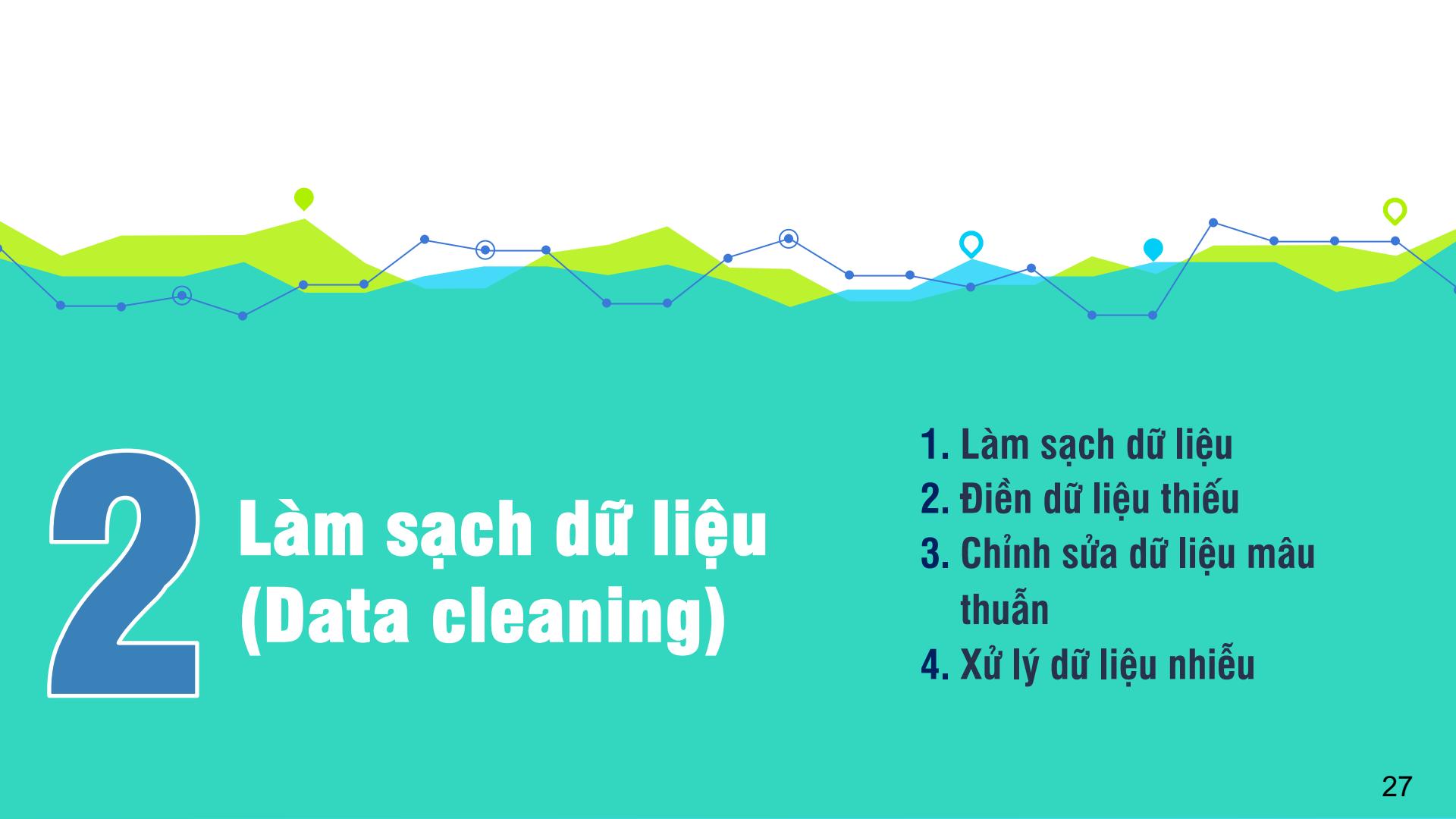


- Thu thập dữ liệu: 19%
- Xử lý và làm sạch dữ liệu: 60%
- Tạo tập dữ liệu huấn luyện: 3%
- Khai thác dữ liệu: 9%
- Cải thiện thuật toán: 4%
- Khác: 5%

Data science report, CrowdFlower, 2016

2

Làm sạch dữ liệu (Data cleaning)

- 
1. Làm sạch dữ liệu
 2. Điền dữ liệu thiếu
 3. Chỉnh sửa dữ liệu mâu thuẫn
 4. Xử lý dữ liệu nhiễu

1. Làm sạch dữ liệu (Data cleaning)

- Dữ liệu trong thế giới thực có thể bị “bẩn”. Rất nhiều dữ liệu có khả năng không chính xác.
 - VD: thiết bị bị lỗi, lỗi của con người hoặc máy tính, lỗi đường truyền
 - Không đầy đủ (incomplete): thiếu giá trị một vài thuộc tính
 - Nhiễu (noisy): giá trị không chính xác, trùng lặp,
 - Không nhất quán (inconsistent)
 - Cố ý: ngụy tạo dữ liệu bị thiếu (intentional: disguised missing data)

Mẫu dữ liệu cần thu thập từ các nguồn đáng tin cậy.
Phản ánh vấn đề cần giải quyết



1. Làm sạch dữ liệu (Data cleaning)

color	director_name	duration	gross	movie_title	language	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall Street	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful Eight	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	Margaret	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of Smaug	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of Justice	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected Journey	English	USA	180000000	2012	7.9
Color	Edward Hall	180		Restless	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The Avengers	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud Atlas	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon Tattoo	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of God	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Desolation of Smaug	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les Misérables	English	USA	61000000	2012	7.6

Dirty / Unclean Data set

1. Làm sạch dữ liệu (Data cleaning)

- Là vấn đề quan trọng bậc nhất
- Các nhiệm vụ của công đoạn làm sạch
 - Dữ liệu mâu thuẫn: Chỉnh sửa
 - Dữ liệu thiếu, chưa đầy đủ: Cần có chiến lược phù hợp
 - Bỏ qua, không đưa vào phân tích?
 - Bổ sung, điền vào dữ liệu thiếu?
 - Dữ liệu nhiễu, ngoại lai: Loại bỏ



Data Cleaning

30

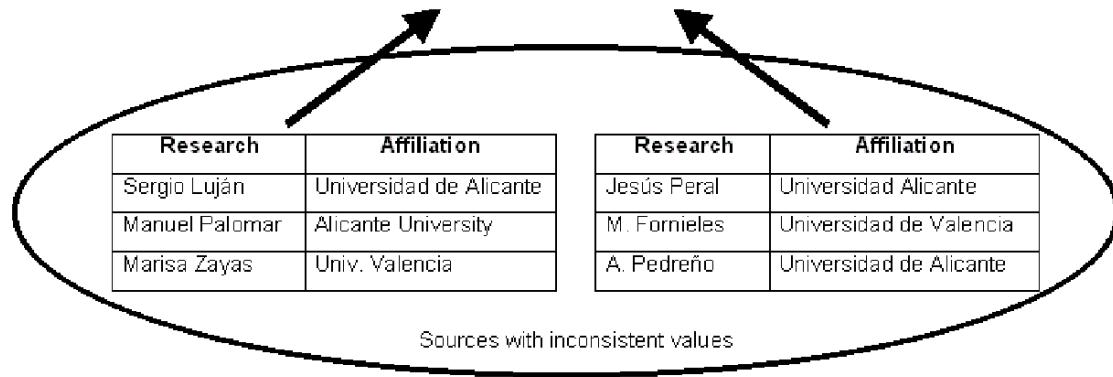
2. Chỉnh sửa dữ liệu mẫu thuẫn

Common repository with consistent information

Các mẫu dữ liệu cần nhất quán
về cách biểu diễn, ký hiệu

VD: Rating ={1,2,3,4,5}

Research	Affiliation
Sergio Luján	Universidad de Alicante
Manuel Palomar	Universidad de Alicante
Marisa Zayas	Universidad de Valencia
Jesús Peral	Universidad de Alicante
M. Fornieles	Universidad de Valencia
A. Pedreño	Universidad de Alicante



3. Điền dữ liệu thiếu

- **Điền giá trị thiếu bằng tay**
- **Điền giá trị thiếu tự động**
 - Bằng hằng số chung, 1 lớp mới. VD: unknown
 - Bằng giá trị trung bình của thuộc tính
 - Bằng giá trị trung bình của các mẫu thuộc cùng lớp đó
 - Bằng giá trị có xác suất, khả năng lớn nhất: suy ra từ Hồi quy, Cây quyết định, Công thức suy diễn Bayesian, Giải thuật EM (Expectation Maximization),

A1	A2	A3	A4	A5	A6	A7	A8	y
?	3.683	?	-0.634	1	0.409	7	30	5
?	?	60	1.573	0	0.639	7	30	5
?	3.096	67	0.249	0	0.089	?	80	3
2.887	3.870	68	-1.347	?	1.276	?	60	5
2.731	3.945	79	1.967	1	2.487	?	100	4

3. Điền dữ liệu thiếu

BT01

Bài tập tình huống:

Thu thập dữ liệu về sinh viên thuộc ĐHQG-HCM để phân tích mức sống của SV

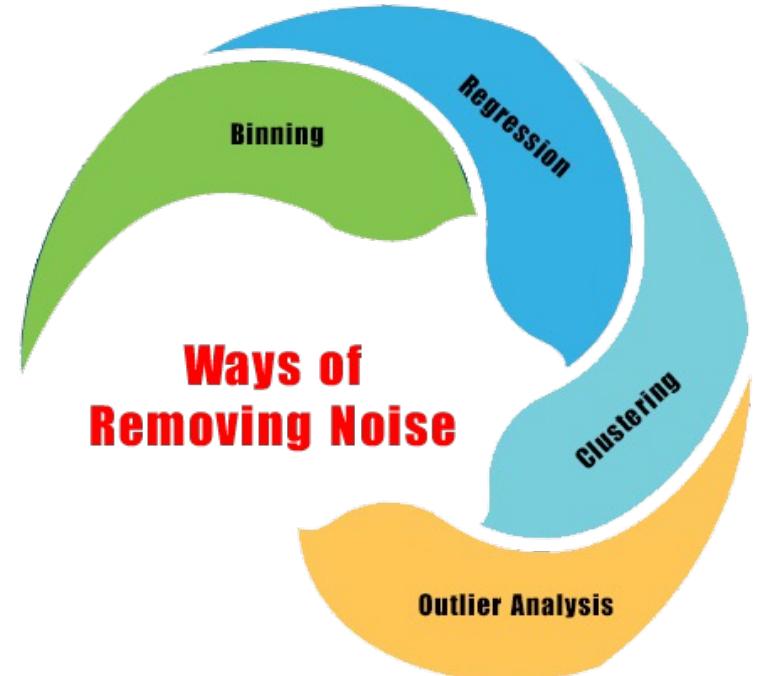
1. Liệt kê các thuộc tính có thể có?
2. Giả sử thuộc tính bị thiếu giá trị là thuộc tính Tiền thuê nhà. Cách giải quyết?

Hình thức thực hiện: Thảo luận nhóm (4 sinh viên) trong 07 phút, chọn và trình bày 01 giải pháp tốt nhất.

4. Xử lý dữ liệu nhiễu

- Chia giỏ (Binning):

- B1: Sắp xếp dữ liệu và phân vùng vào các giỏ theo độ rộng (Equal-width) hoặc độ sâu (Equal-depth)
- B2: Làm trơn (smooth), khử nhiễu bằng:
 - Giá trị trung bình (bin mean)
 - Giá trị trung vị (bin median)
 - Giá trị biên giỏ (bin boundary).



4. Xử lý dữ liệu nhiễu

- **Hồi quy (Regression)**

- Làm trơn bằng cách đưa dữ liệu vào các hàm hồi quy

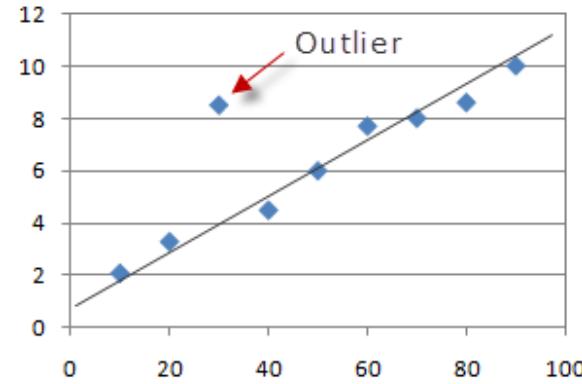
- **Phân cụm (Clustering)**

- Phát hiện và loại bỏ các ngoại lệ

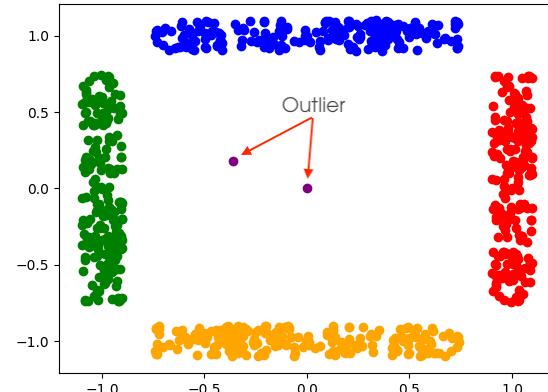
- **Kết hợp kiểm duyệt máy tính và con người**

- Phát hiện các giá trị đáng ngờ và kiểm duyệt bởi con người.

VD: xử lý các ngoại lệ có thể xảy ra)



a) Hồi quy



b) Phân cụm



Phương pháp chia giỏ - Binning

- Phương pháp rời rạc hóa
- **Bước 1 - Chia giỏ:** 2 cách:
 - **Chia theo độ rộng (Equal-width – khoảng cách)**
 - Chia vùng giá trị thành N khoảng cùng kích thước
 - Độ rộng của từng khoảng = $\frac{\text{Giá trị lớn nhất} - \text{Giá trị nhỏ nhất}}{N}$
 - **Chia theo độ sâu (Equal-depth – tần suất)**
 - Chia vùng giá trị thành N khoảng, mỗi khoảng có chứa gần như cùng số lượng mẫu
- **Bước 2 – Làm trơn (smooth):** Giá trị trung bình (bin mean), Giá trị trung vị (bin median), Giá trị biên giới (bin boundary).



Phương pháp chia giỏ - Binning

Chia giỏ theo độ rộng (Equal-width – khoảng cách)

Giá trị nhiệt độ: 64 65 68 69 70 71 72 72 75 75 80 81 83 85

Số giỏ = 7

Đếm \Rightarrow Độ rộng của từng khoảng = $\frac{85 - 64}{7} = 3$

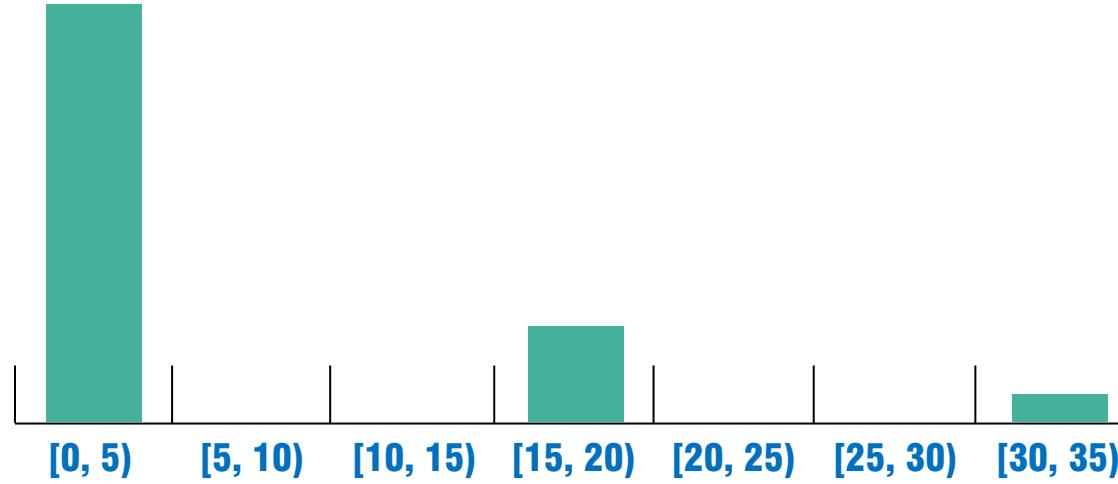


Biên trái giỏ \leq giá trị $<$ biên phải giỏ



Phương pháp chia giỏ - Binning

Chia giỏ theo độ rộng (*Equal-width – khoảng cách*): không tốt khi dữ liệu bị lệch



Mức lương trong công ty (triệu đồng)

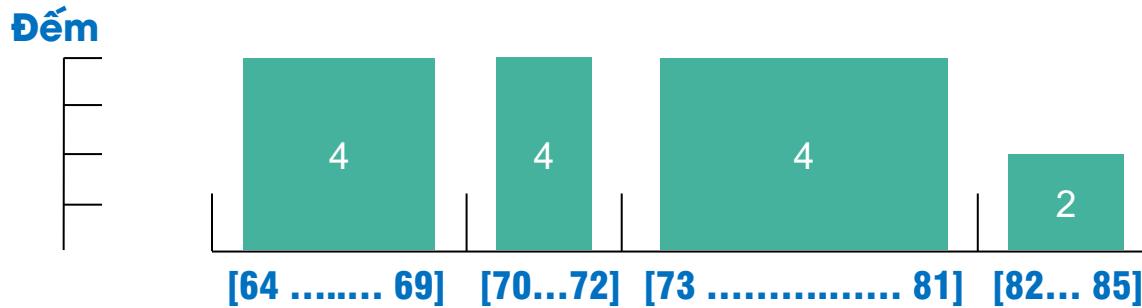


Phương pháp chia giỏ - Binning

Chia giỏ theo độ sâu (Equal-depth – tần suất)

Giá trị nhiệt độ: 64 65 68 69 70 71 72 72 75 75 80 81 83 85

$$\text{Số giỏ} = 4 \Rightarrow \text{Độ sâu} = 14/4 \approx 4$$



Biên trái giỏ \leq giá trị \leq biên phải giỏ



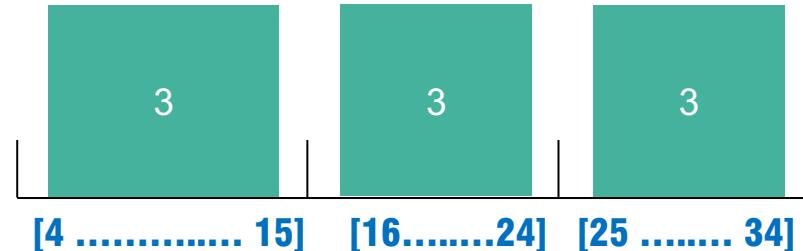
Phương pháp chia giỏ - Binning

VD1: Dữ liệu giá sản phẩm (\$): 15, 21, 8, 4, 24, 21, 25, 34, 28

- Sắp xếp: 4, 8, 15, 21, 21, 24, 25, 28, 34

- Chia giỏ theo độ sâu = 3

- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34



- Làm tròn:

• **Bằng trung bình giỏ**

- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29, 29

• **Bằng trung vị giỏ**

- Bin 1: 8, 8, 8
- Bin 2: 21, 21, 21
- Bin 3: 28, 28, 28

• **Bằng biên giỏ (trái)**

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



Phương pháp chia giỏ - Binning

BTO2

Bài tập:

Cho dữ liệu giá của cổ phiếu CKH (1000đ)

19, 25, 17, 15, 31, 33, 29, 41, 45, 52, 47, 64, 52, 45, 42,

Số giỏ cần chia: 4

Sử dụng phương pháp chia giỏ theo độ rộng và độ sâu

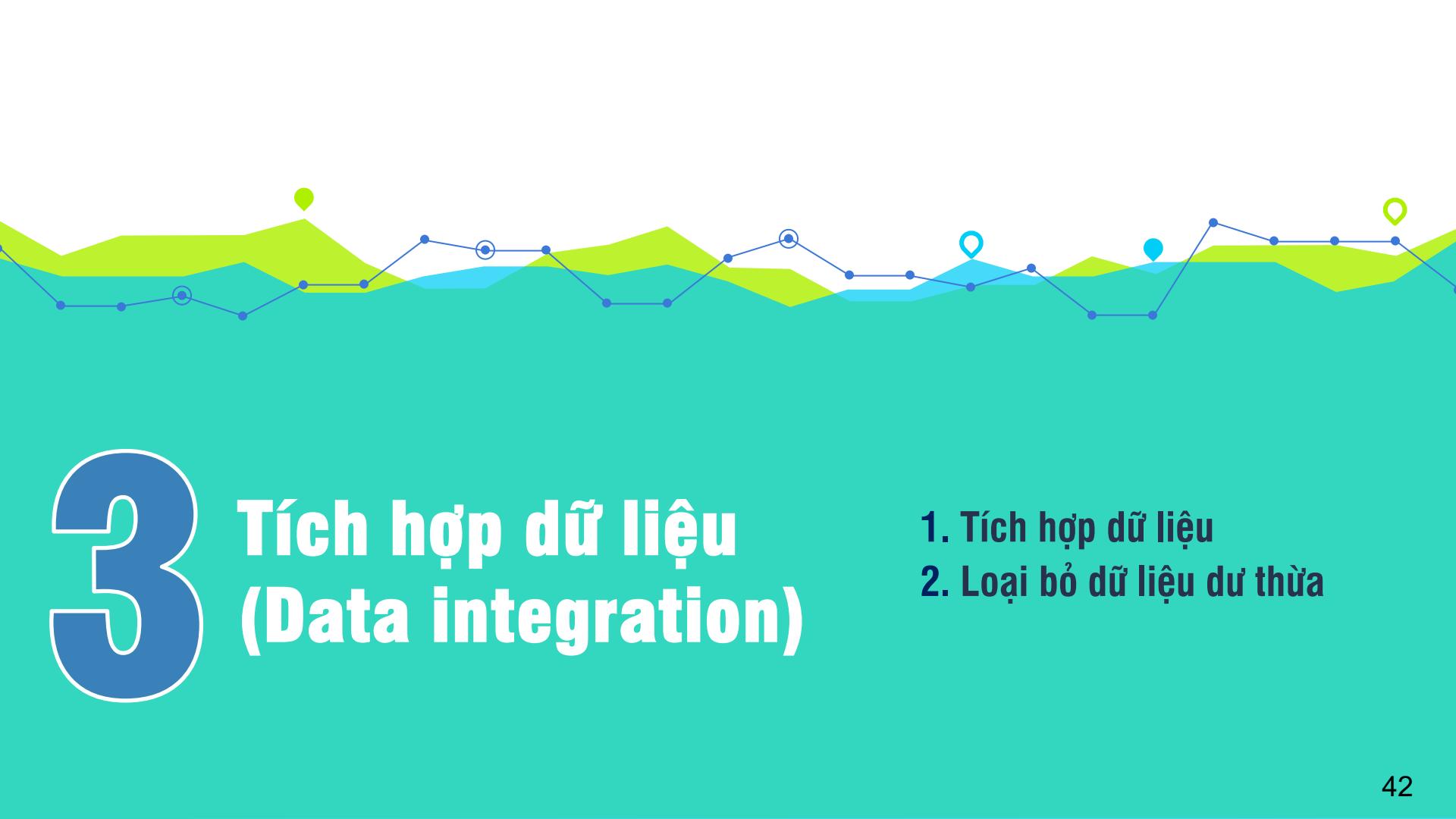
1. Làm tròn theo trung vị
2. Làm tròn theo biên giỏ (trái, phải)

Hình thức thực hiện: Nhóm (4 sinh viên) trong 10 phút

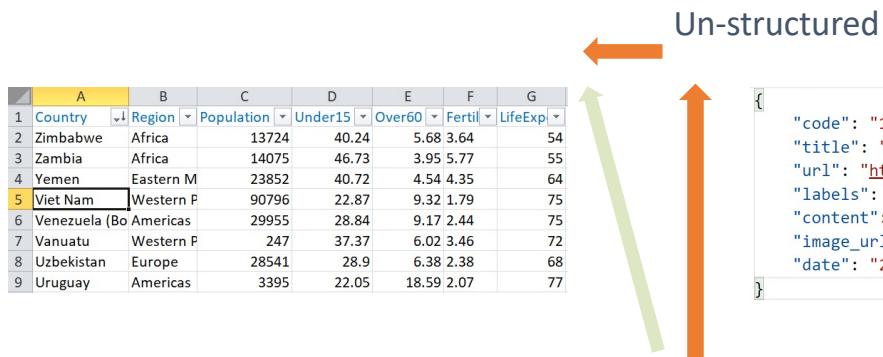


3

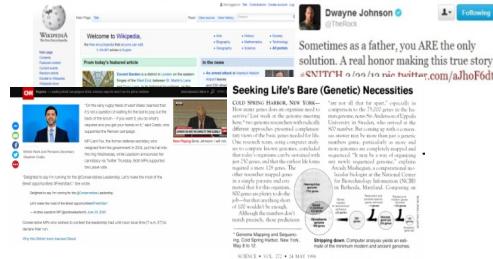
Tích hợp dữ liệu (Data integration)

- 
- 1. Tích hợp dữ liệu
 - 2. Loại bỏ dữ liệu dư thừa

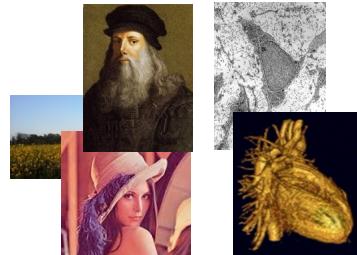
1. Tích hợp dữ liệu (Data integration)



texts in websites, emails, articles, tweets



2D/3D images, videos + meta



spectrograms, DNAs, ...

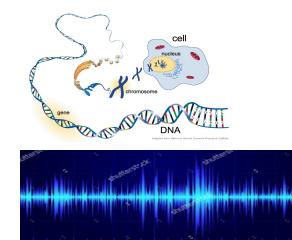


image credits: wikipedia, shutterstock, CNN



1. Tích hợp dữ liệu

Tập hợp dữ liệu từ nhiều nguồn khác nhau vào trong một CSDL

- Có thể dư thừa dữ liệu
- Có thể trùng lặp dữ liệu
- Chỉ chọn những dữ liệu cần thiết để KTDL

Tích hợp cẩn thận dữ liệu từ nhiều nguồn:
Giảm/tránh dư thừa và không nhất quán
Cải thiện tốc độ và chất lượng khai thác dữ liệu



1. Tích hợp dữ liệu (Data integration)

- Vấn đề nhận diện thực thể

- Xác định các thực thể trong thế giới thực từ nhiều nguồn dữ liệu
- VD: Customer_ID = cust_number; Bill Clinton = William Clinton

- Phát hiện và giải quyết mâu thuẫn dữ liệu

- Đối với cùng một thực thể trong thế giới thực, các giá trị thuộc tính từ các nguồn khác nhau là khác nhau
- Lý do có thể: cách biểu diễn khác nhau, tỷ lệ khác nhau
- VD: Kilometer vs Mile; Pound vs Kilogram

- Loại bỏ dữ liệu dư thừa và trùng lặp



2. Loại bỏ dữ liệu dư thừa

- Dư thừa dữ liệu thường xảy ra khi tích hợp nhiều cơ sở dữ liệu
 - Cùng một thuộc tính hoặc đối tượng có thể có các tên khác nhau trong các CSDL khác nhau
 - Dữ liệu có thể dẫn xuất: Một thuộc tính có thể là thuộc tính “có nguồn gốc” trong một bảng khác. VD: doanh thu hàng năm, trị giá hóa đơn
 - Một số mẫu bị lặp lại
- Các thuộc tính dư thừa có thể được phát hiện bằng phân tích tương quan (correlation analysis), hiệp phương sai (covariance analysis)

2.1. Phân tích tương quan (nominal data)

- χ^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \quad (1)$$

- χ^2 càng lớn thì khả năng các thuộc tính có liên quan với nhau càng cao
- Các giá trị cell đóng góp nhiều nhất vào χ^2 là những ô có số lượng thực tế (Observed) rất khác so với số lượng kỳ vọng (Expected)
- Mỗi tương quan không bao hàm quan hệ nhân quả
 - #bệnhviện và #vụtrộmxe trong thành phố có mối tương quan với nhau
 - Cả hai đều có quan hệ nhân quả với thuộc tính thứ ba: #dânsố

2.1. Phân tích tương quan (nominal data)

VD2:

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

(...): Expected

- Phép tính X^2 (chi-square)

$$X^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

- Kết luận: like_science_fiction và play_chess có tương quan với nhau

2.2. Phân tích tương quan (numeric data)

- Hệ số tương quan Pearson (Pearson Correlation coefficient)

$$r_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n \sigma_X \sigma_Y} \quad (2)$$

- n : số bộ giá trị
- \bar{X}, \bar{Y} : giá trị trung bình của X, Y
- $\sigma_X \sigma_Y$: độ lệch chuẩn của X, Y

$$\sigma_X = \sqrt{\frac{1}{n} \sum (x_i - \bar{X})^2} \text{ và } \sigma_Y = \sqrt{\frac{1}{n} \sum (y_i - \bar{Y})^2}$$



2.2. Phân tích tương quan (numeric data)

- **Hệ số tương quan Pearson (Pearson Correlation coefficient)**

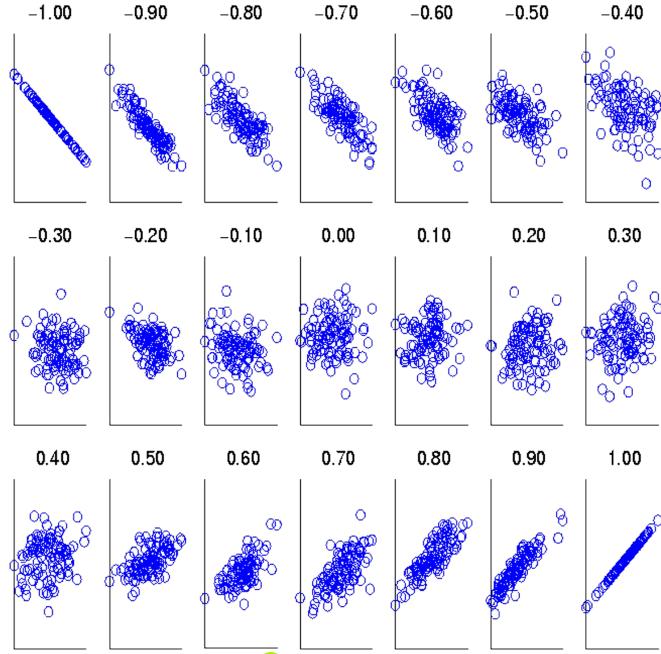
$$r_{X,Y} = \frac{\sum[(x_i - \bar{X})(y_i - \bar{Y})]}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}} \quad (3)$$

$$r_{X,Y} = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2] [n \sum y^2 - (\sum y)^2]}} \quad (4)$$



2.2. Phân tích tương quan (numeric data)

- Hệ số tương quan Pearson (Pearson Correlation coefficient)

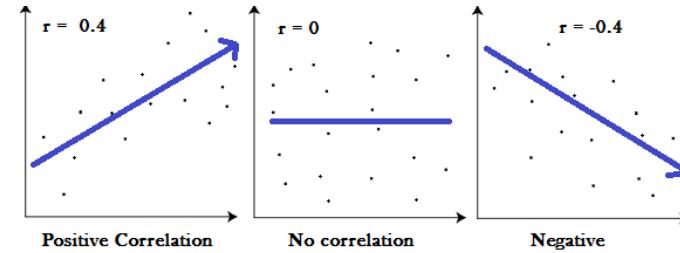


$$r_{X,Y} \in [-1, 1]$$

$r_{X,Y} = 0$: X, Y không tương quan

$r_{X,Y} > 0$: X, Y tương quan thuận ($X \uparrow \leftrightarrow Y \uparrow$)

$r_{X,Y} < 0$: X, Y tương quan nghịch ($X \downarrow \leftrightarrow Y \uparrow$)



2.3. Phân tích hiệp phương sai

- Hệ số hiệp phương sai (Covariance coefficient)

$$cov(X, Y) = E((X - \bar{X})(Y - \bar{Y})) = \frac{\sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})}{n} \quad (5)$$

$$cov(X, Y) = E(X \cdot Y) - E(X) E(Y) \quad (6)$$

$$r_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad (7)$$

- $E(A)$: kỳ vọng $E(A) = \sum_i p_i a_i$ với p_i : xác suất a_i xuất hiện
- n : số bộ giá trị
- \bar{X}, \bar{Y} : giá trị trung bình của X, Y
- $\sigma_X \sigma_Y$: độ lệch chuẩn của X, Y

$cov(X, Y)$ càng lớn thì khả năng các thuộc tính phụ thuộc nhau càng cao

2.3. Phân tích hiệp phương sai

VD3: Giả sử hai cổ phiếu A và B có các giá trị sau trong một tuần:

Day of week	Stock A	Stock B
Mon	2	5
Tue	3	8
Wed	5	10
Thu	4	11
Fri	6	14

Câu hỏi: Nếu các cổ phiếu bị ảnh hưởng bởi cùng một xu hướng ngành, giá của chúng sẽ tăng hay giảm cùng nhau?



2.3. Phân tích hiệp phương sai

VD3:

Day of week	Stock A	Stock B
Mon	2	5
Tue	3	8
Wed	5	10
Thu	4	11
Fri	6	14

$$E(A) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$$

$$E(B) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9,6$$

$$\text{Cov}(A, B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9,6 = 4$$

Do đó, A và B cùng tăng vì $\text{Cov}(A, B) > 0$.



Phân tích tương quan

Bài tập: Tìm hệ số tương quan Pearson, hệ số hiệp phương sai và kết luận sự tương quan giữa 2 thuộc tính:

Subject	Age (X)	Glucose Level (Y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Hình thức thực hiện: Nhóm (4 sinh viên) trong 10 phút



4 Rút gọn dữ liệu (Data Reduction)

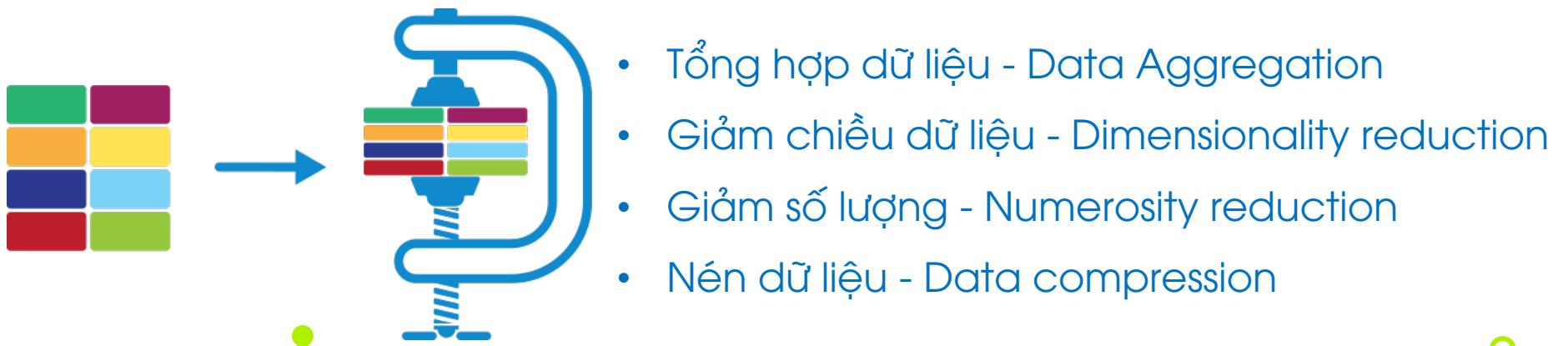


1. Rút gọn dữ liệu
2. Tổng hợp và tổng quát hóa
3. Giảm chiều dữ liệu
4. Giảm số lượng
5. Nén dữ liệu

1. Rút gọn dữ liệu (Data reduction)

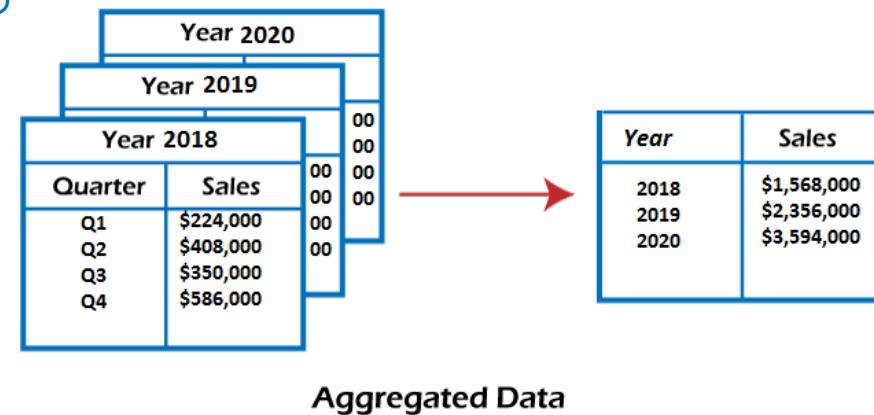
Rút gọn dữ liệu: Dữ liệu có khối lượng nhỏ hơn nhiều nhưng vẫn tạo ra kết quả phân tích giống nhau (hoặc gần như giống nhau)

Tại sao rút gọn dữ liệu? - Dữ liệu có thể quá lớn. Phân tích dữ liệu mất nhiều thời gian.



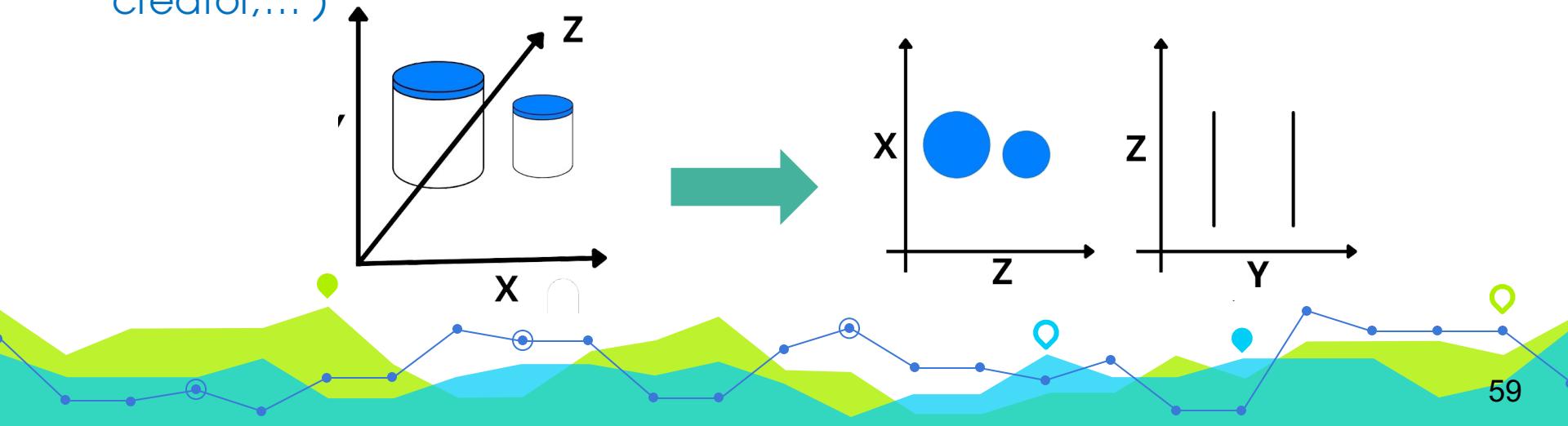
2. Tổng hợp và tổng quát hóa (Data Aggregation)

- Tổng hợp dữ liệu ở dạng đơn giản hơn.
- Tổ hợp từ 2 thuộc tính (đối tượng) trở lên thành 1 thuộc tính (đối tượng)
 - VD: các thành phố tổng hợp vào vùng, khu vực, nước, ...
- Tổng hợp/ tổng quát dữ liệu cấp thấp vào dữ liệu cấp cao:
 - Giảm số thuộc tính
 - Tăng tính lý thú của mẫu



3. Giảm chiều dữ liệu (Dimensionality reduction)

- Loại bỏ các thuộc tính không quan trọng, dư thừa.
- Các phương pháp:
 - Wavelet transforms
 - Principal Components Analysis (PCA)
 - Supervised and nonlinear techniques (VD: feature selection, feature creator,...)

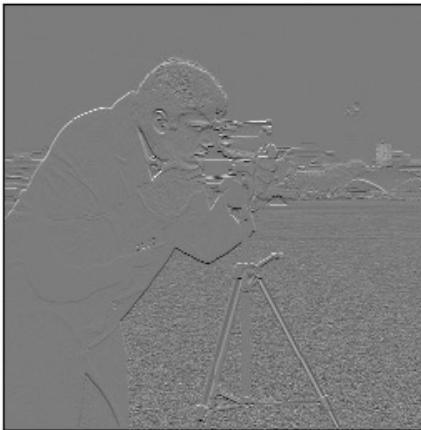


3.1. Wavelet transforms (*)

Approximation



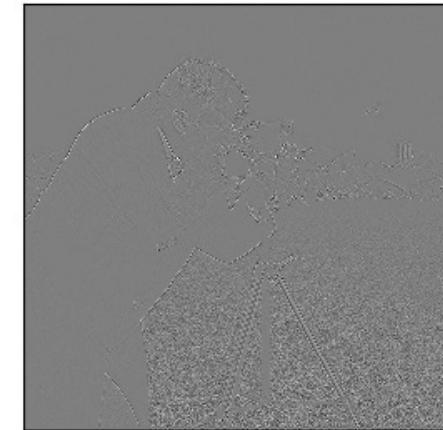
Horizontal detail



Vertical detail



Diagonal detail

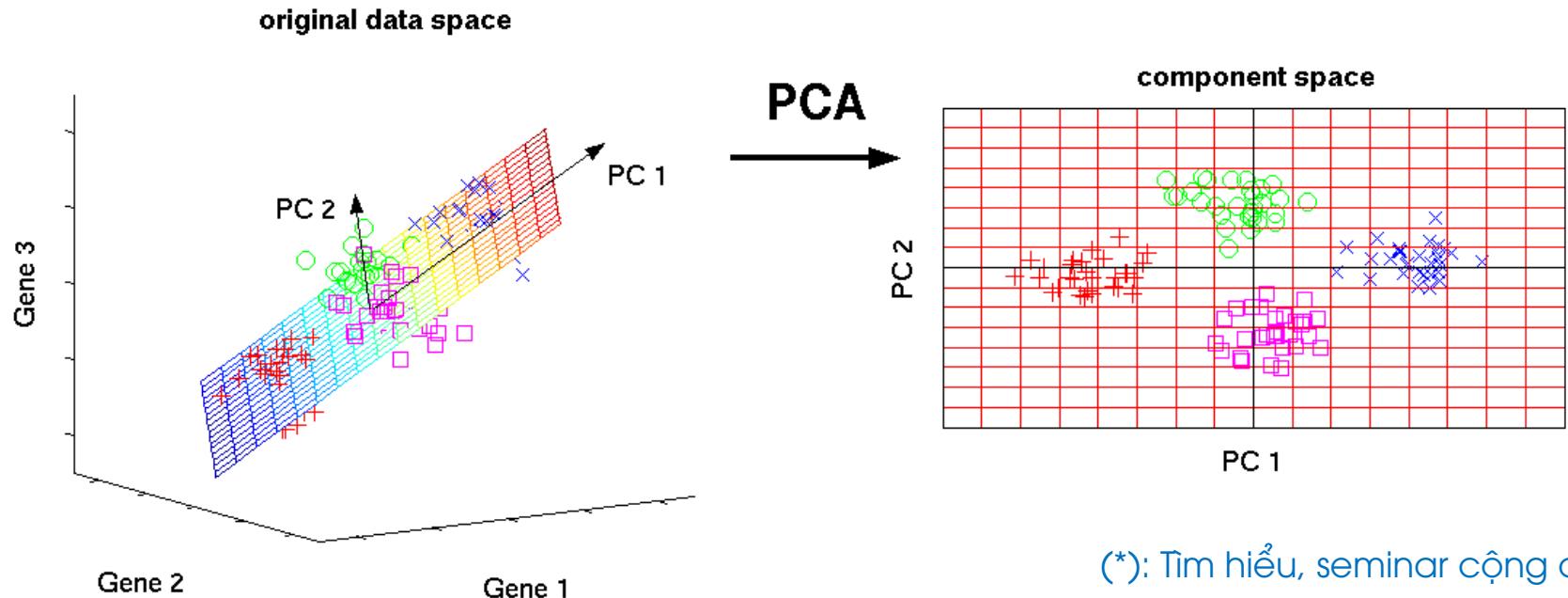


(*): Tìm hiểu, seminar cộng điểm

FARGE, Marie. Wavelet transforms and their applications to turbulence. *Annual review of fluid mechanics*, 1992, 24.1: 395-458



3.2. Principal Components Analysis (PCA) (*)



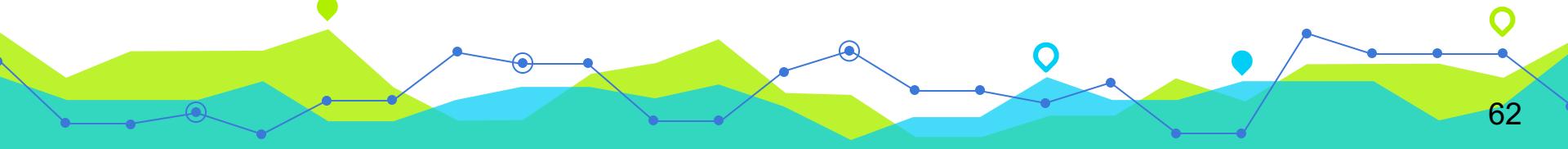
(*): Tìm hiểu, seminar cộng điểm

SMITH, Lindsay I. A tutorial on principal components analysis. 2002



3.3. Chọn lựa đặc trưng (Feature selection)

- Chọn m thuộc tính từ n thuộc tính ($m < n$)
- Loại bỏ các thuộc tính không liên quan, dư thừa
 - VD: MSSV không liên quan đến dự đoán điểm trung bình của sinh viên
 - VD: giá bán sản phẩm và tiền thuế đã bao gồm?



3.3. Chọn lựa đặc trưng (Feature selection)

Có 2^n tập con thuộc tính của n thuộc tính. Chọn đặc trưng?

- **Vết cạn:** Độ phức tạp tính toán quá cao
- **Phương pháp Heuristic:**
 - Step-wise forward Heuristic
 - Step-wise backward Heuristic
 - Kết hợp forward & backward
 - Cây quyết định quy nạp



3.3. Chọn lựa đặc trưng (Feature selection)

- Phương pháp Heuristic - Step-wise forward

- Chọn thuộc tính đơn tốt nhất
- Chọn thuộc tính đơn tốt nhất trong số còn lại. Tiếp tục lặp lại

VD: Tập thuộc tính ban đầu {A1, A2, A3, A4, A5, A6}

- Tập rút gọn ban đầu = {}
- B1 = {A1}
- B2 = {A1, A4}
- B3 = {A1, A4, A6}



3.3. Chọn lựa đặc trưng (Feature selection)

- Phương pháp Heuristic - Step-wise backward

- Loại thuộc tính đơn xấu nhất
- Loại thuộc tính đơn xấu nhất trong số còn lại. Tiếp tục lặp lại

VD: Tập thuộc tính ban đầu {A1, A2, A3, A4, A5, A6}

- Tập rút gọn ban đầu = {A1, A2, A3, A4, A5, A6}
- B1 = {A1, A3, A4, A5, A6}
- B2 = {A1, A4, A5, A6}
- B3 = {A1, A4, A6}



3.3. Chọn lựa đặc trưng (Feature selection)

- Phương pháp Heuristic – Kết hợp

- Chọn thuộc tính đơn tốt nhất và loại xấu nhất
- Chọn thuộc tính đơn tốt nhất và loại xấu nhất trong số còn lại. Tiếp tục lặp lại

VD: Tập thuộc tính ban đầu {A1, A2, A3, A4, A5, A6}

- Tập rút gọn ban đầu = {A1, A2, A3, A4, A5, A6}
- B1 = {A1, A3, A4, A5, A6}
- B2 = {A1, A4, A5, A6}
- B3 = {A1, A4, A6}



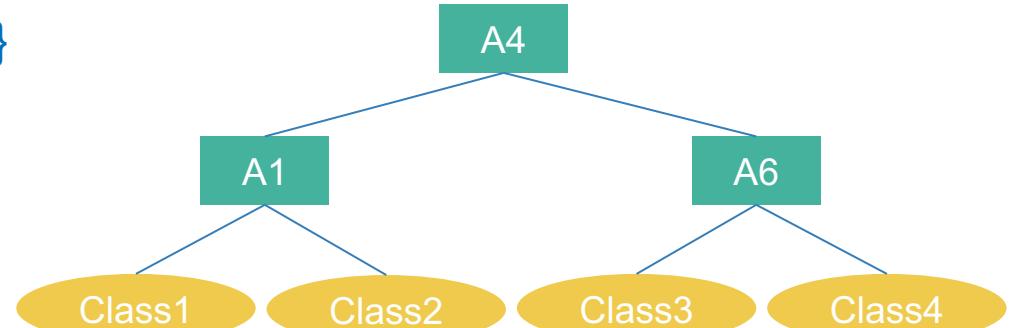
3.3. Chọn lựa đặc trưng (Feature selection)

- Phương pháp Cây quyết định quy nạp

- Xây dựng cây quyết định
- Loại bỏ những thuộc tính không xuất hiện trên cây

VD: Tập thuộc tính ban đầu {A1, A2, A3, A4, A5, A6}

Tập rút gọn = {A1, A4, A6}



3.4. Tạo đặc trưng mới (Feature creator/construction)

- Tạo các thuộc tính (đặc trưng) mới có thể chứa thông tin quan trọng, hiệu quả hơn các thuộc tính ban đầu.
 - Các phương pháp chung:
 - Rút trích, trích xuất thuộc tính - Attribute extraction
 - Ánh xạ dữ liệu sang không gian mới
- VD: Fourier transformation, Wavelet transformation
- Xây dựng thuộc tính: Kết hợp các đặc trưng,



4. Giảm số lượng (Numerosity reduction)

- Chọn dạng biểu diễn dữ liệu thay thế, nhỏ hơn

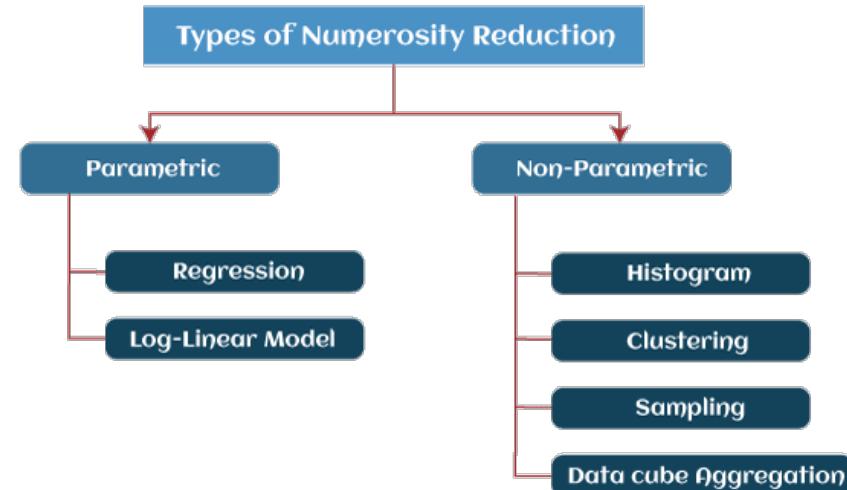
- **Các phương pháp:**

- **Phương pháp tham số:**

- Sử dụng mô hình toán học để tính toán và lưu giữ các tham số

- **Phương pháp phi tham số:**

- Không sử dụng mô hình toán học, lưu giữ biểu diễn rút gọn của dữ liệu



4.1. Các phương pháp tham số

- **Linear regression (Hồi quy tuyến tính)**

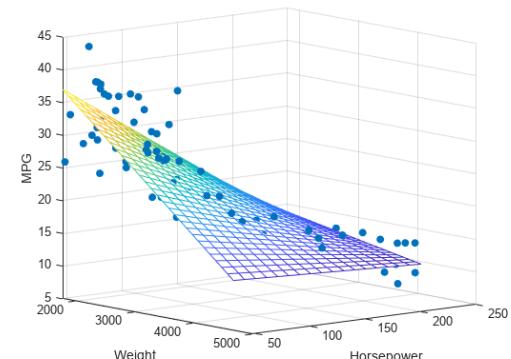
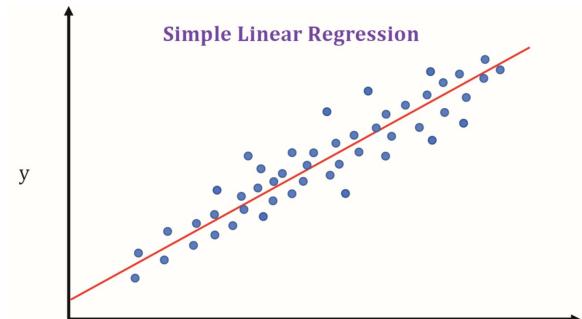
$$Y = \alpha + \beta X$$

- **Multiple regression (Hồi quy bội)**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

- **Log-linear model (Mô hình log tuyến tính)**

$$\log y_i = \sum_j \lambda_j x_{ij}$$



4.2. Các phương pháp phi tham số

Phân tích Histogram

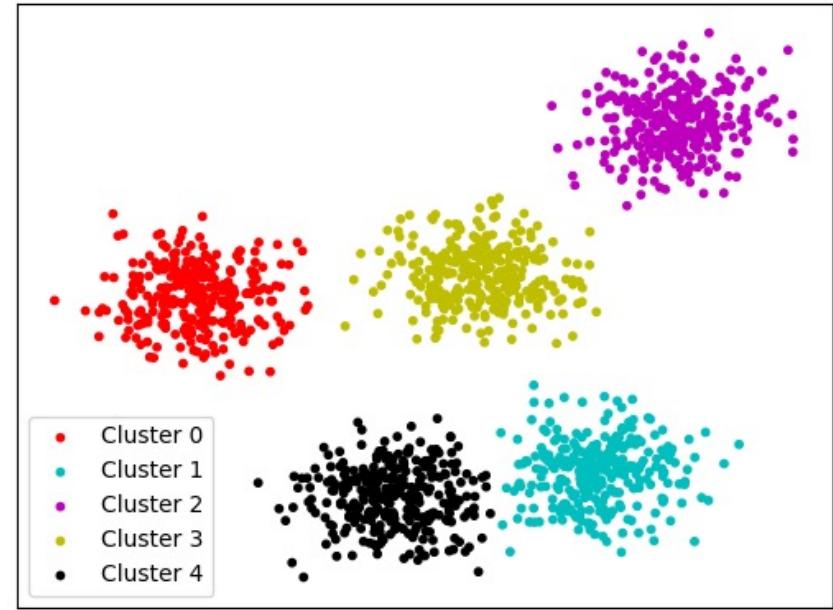
- Thông dụng để rút gọn dữ liệu
- Phân chia dữ liệu vào các giỏ. Chiều cao của cột là số đối tượng mỗi giỏ
- Chỉ lưu giá trị trung bình của mỗi giỏ
- Hình dạng biểu đồ tùy thuộc số lượng giỏ
- Quy tắc chia giỏ:
 - Equal-width: phạm vi giỏ bằng nhau
 - Equal-depth: tần số bằng nhau



4.2. Các phương pháp phi tham số

Gom cụm (Clustering)

- Dữ liệu phân thành các cụm dựa trên độ tương đồng và chỉ lưu trữ biểu diễn cụm (VD: trọng tâm và đường kính)
- Rất hiệu quả nếu dữ liệu tập trung thành cụm, ngược lại khi dữ liệu rải rác
- Rất nhiều thuật toán phân cụm



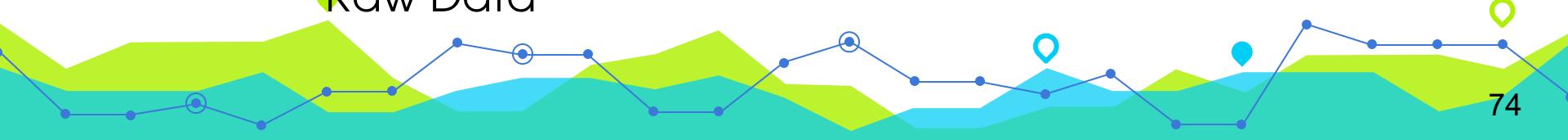
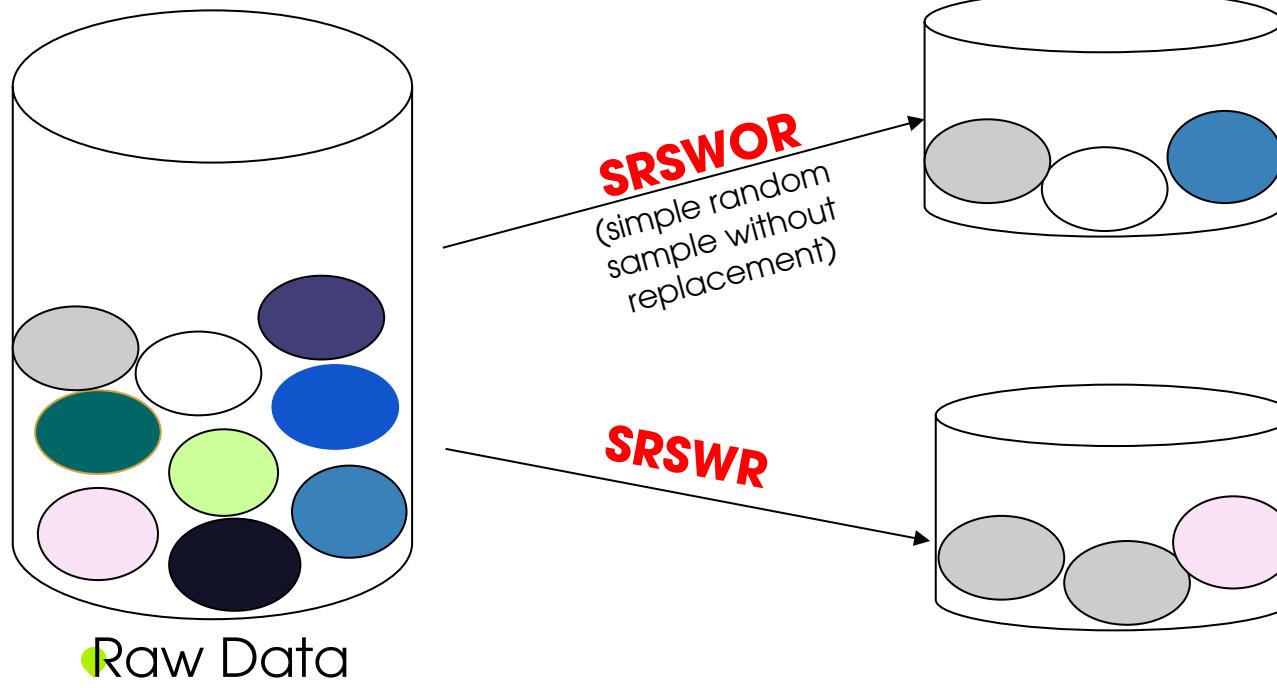
4.2. Các phương pháp phi tham số

Lấy mẫu (Sampling)

- Dùng tập mẫu ngẫu nhiên nhỏ hơn để thay thế tập dữ liệu lớn
- **Các phương pháp:**
 - **Lấy mẫu ngẫu nhiên đơn giản** (Simple random sampling): Xác suất bằng nhau khi chọn bất kỳ mẫu nào
 - **Lấy mẫu không thay thế** (Sampling without replacement - SRSWOR): Khi một mẫu được chọn, nó sẽ bị xóa khỏi tập dữ liệu
 - **Lấy mẫu có thay thế** (Sampling with replacement - SRSWR): Một mẫu được chọn không bị xóa khỏi tập
 - **Lấy mẫu phân tầng** (Stratified sampling): Phân vùng tập dữ liệu và lấy các mẫu từ mỗi phân vùng (theo tỷ lệ phần trăm của dữ liệu)

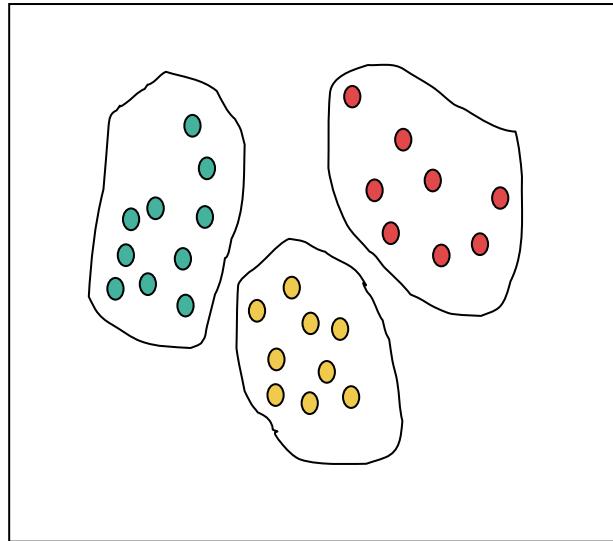
4.2. Các phương pháp phi tham số

Lấy mẫu (Sampling)

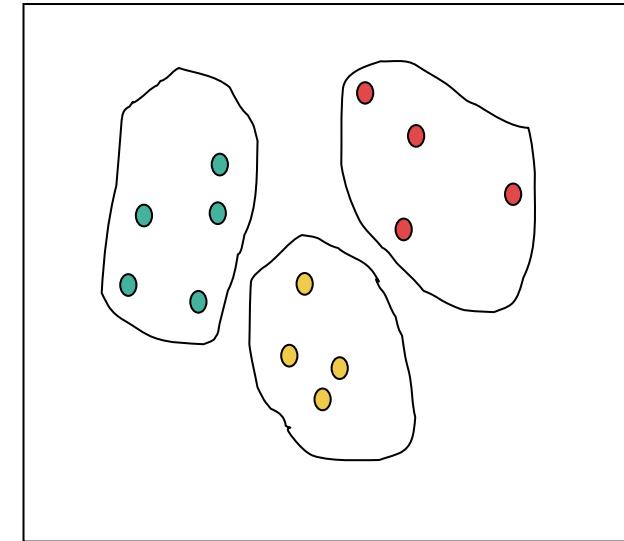


Clustering or Stratified Sampling

Raw Data



Cluster/Stratified Sample

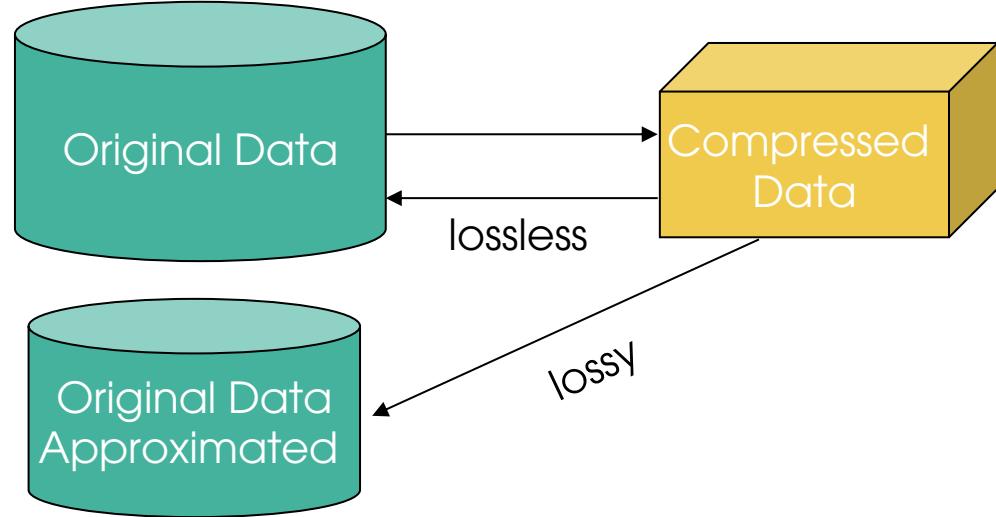


Chia dữ liệu thành nhiều phân vùng; sau đó rút các mẫu ngẫu nhiên từ mỗi phân vùng



5. Nén dữ liệu (Data compression)

- Mã hóa, biến đổi dữ liệu
- Nén không mất dữ liệu (lossless).
VD: string compression
- Nén mất dữ liệu (lossy).
VD: audio/ video compression
- Một số phương pháp:
 - Wavelet transform,
 - PCA,
 -



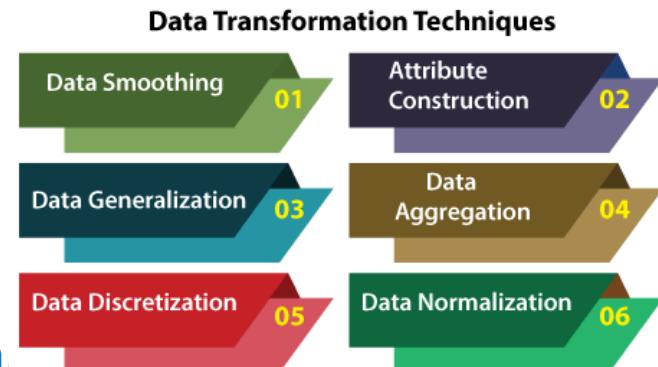
5

Biến đổi, mã hóa dữ liệu (Data Transformation)

- 
1. Biến đổi, mã hóa dữ liệu
 2. Tổng hợp và tổng quát hóa
 3. Giảm chiều dữ liệu
 4. Giảm số lượng
 5. Nén dữ liệu

1. Biến đổi dữ liệu (Data transformation)

- Ánh xạ toàn bộ tập giá trị của một thuộc tính sang tập giá trị mới.
- **Các phương pháp:**
 - Smoothing: Làm trơn, loại bỏ nhiễu
 - Attribute/feature construction: Xây dựng thuộc tính mới
 - Aggregation: Tổng hợp, xây dựng data cube
 - Normalization: Biến đổi về miền giá trị nhỏ hơn
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
 - Discretization: Rời rạc hóa, phân cấp khái niệm



2. Chuẩn hóa (Normalization)

- **Min-max normalization:** chuyển về $[new_min_A, new_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

VD: Giả sử Salary thuộc khoảng $[\$12,000, \$98,000]$, chuyển về $[0.0, 1.0]$

Giá trị \$73,600 chuyển thành:

$$\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$$



2. Chuẩn hóa (Normalization)

- **z-score normalization:** (μ : trung bình, σ : độ lệch chuẩn)

$$v' = \frac{v - \mu}{\sigma}$$

VD: Với $\mu = 54,000, \sigma = 16,000$.

Khi đó, $v' = \frac{73,600 - 54,000}{16,000} = 1,225$

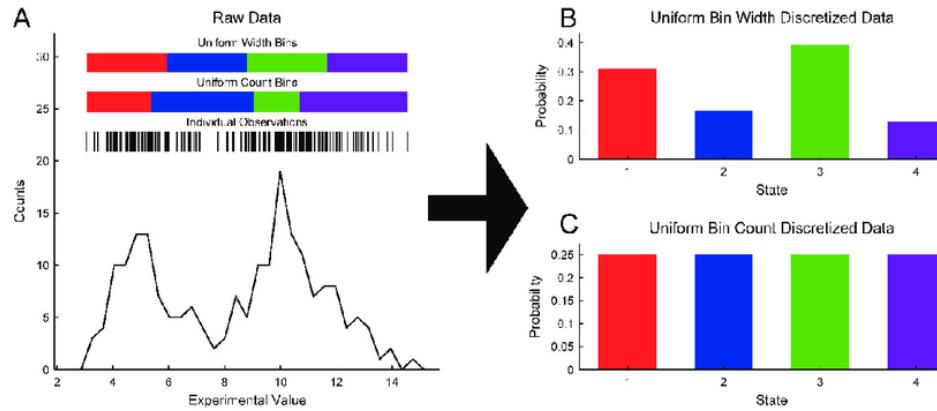
- **normalization by decimal scaling:**

$$v' = \frac{v}{10^j} \text{ với } j: \text{số nguyên nhỏ nhất sao cho } \text{Max}(|v'|) < 1$$



3. Rời rạc hóa (Discretization)

- Kiểu dữ liệu: nominal, ordinal, numeric
- Chia phạm vi của một thuộc tính liên tục thành các khoảng
- Có thể được thực hiện đệ quy trên một thuộc tính
- Chuẩn bị cho phân tích sâu hơn. VD: phân lớp



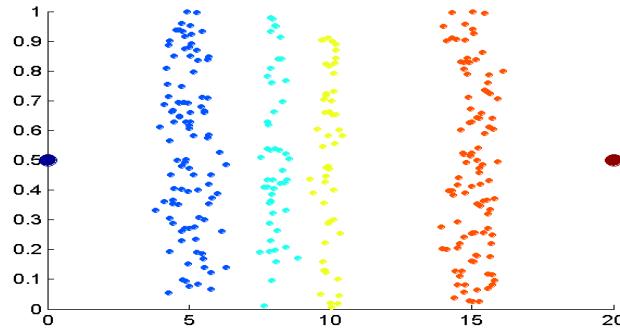
3. Rời rạc hóa (Discretization)

- Các phương pháp:

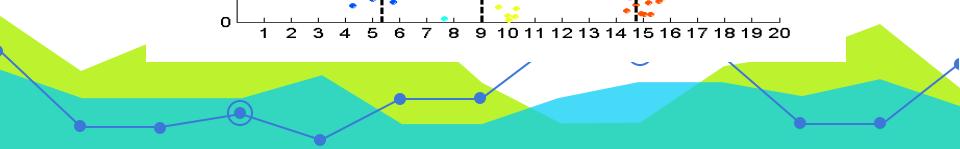
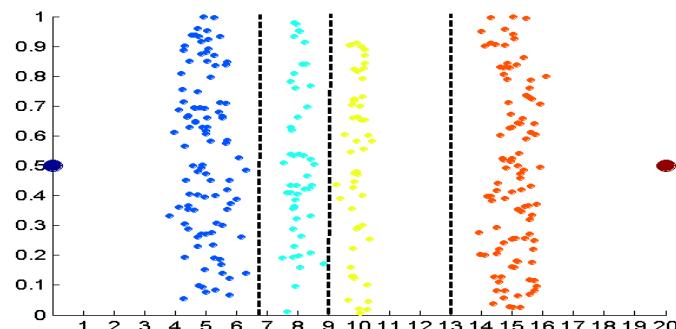
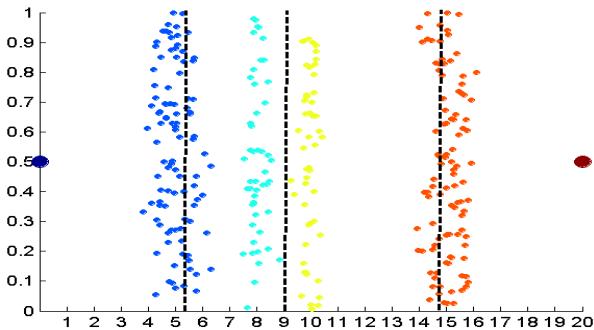
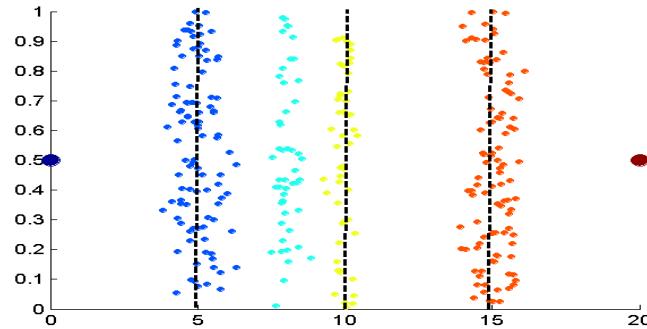
- Chia giỏ (Binning)
 - Top-down split, unsupervised
- Phân tích Histogram
 - Top-down split, unsupervised
- Gom cụm (Clustering)
 - Top-down split or bottom-up merge, unsupervised,
- Cây quyết định (Decision-tree)
 - Top-down split, supervised
- Phân tích tương quan (Correlation analysis). VD: X^2, r
 - Bottom-up merge, unsupervised

Rời rạc hóa không sử dụng nhãn lớp (Binning so với Clustering)

Equal depth (binning)



K-means clustering leads to better results



Phân cấp khái niệm (Concept Hierarchy Generation)

- Tổ chức các khái niệm (giá trị thuộc tính) theo thứ bậc và thường được liên kết với từng dimension trong dữ liệu.
- Tạo phân cấp khái niệm: thu thập và thay thế các khái niệm cấp thấp bằng các khái niệm cấp cao hơn.
- Hệ thống phân cấp khái niệm có thể được tạo tự động cho numeric data (các phương pháp trước đó) và nominal data.
- VD: Age= {Child, Young, Mature, Old} thay vì lưu trữ số tuổi

Attribute	Age	Age	Age	Age
	1,5,4,9,7	11,14,17,13,18,19	31,33,36,42,44,46	70,74,77,78
After Discretization	Child	Young	Mature	Old

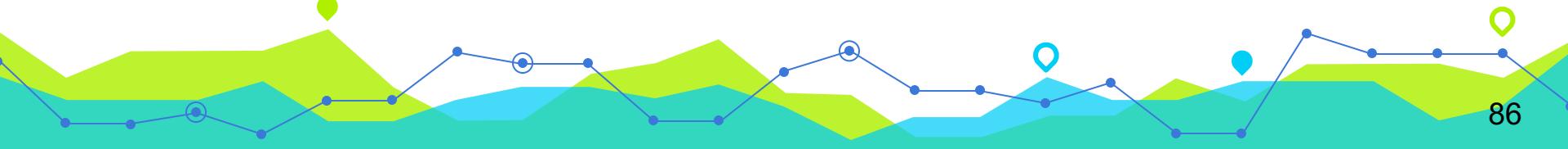
Phân cấp khái niệm (Nominal data)

- Đặc tả thứ tự một phần/toàn bộ các thuộc tính một cách rõ ràng ở cấp lược đồ
VD: street < city < state < country
- Đặc tả cấu trúc phân cấp cho một tập hợp các giá trị bằng cách nhóm dữ liệu rõ ràng
VD: {Urbana, Champaign, Chicago} < Illinois
- Đặc tả chỉ một phần tập thuộc tính
VD: chỉ street < city, không phải những street khác



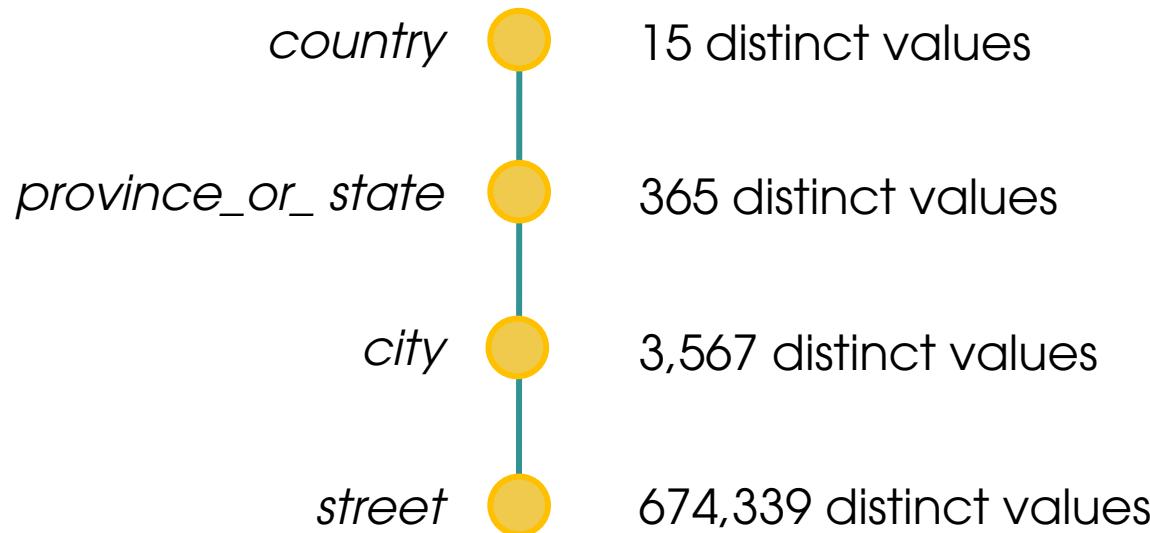
Phân cấp khái niệm (Nominal data)

- Tự động tạo hệ thống phân cấp bằng cách phân tích số lượng giá trị riêng biệt của từng thuộc tính
 - Thuộc tính có giá trị riêng biệt nhiều nhất được đặt ở mức thấp nhất của hệ thống phân cấp
 - Ngoại lệ: day of week, month, quarter, year



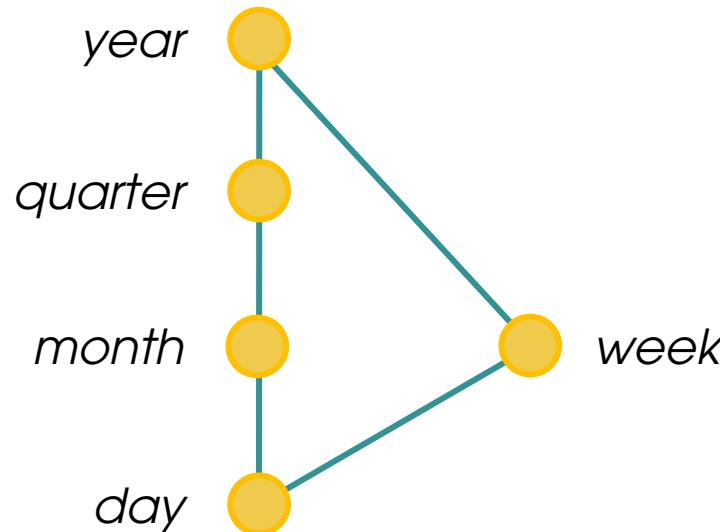
Phân cấp khái niệm (Nominal data)

VD: Tập hợp các thuộc tính: {street, city, state, country}



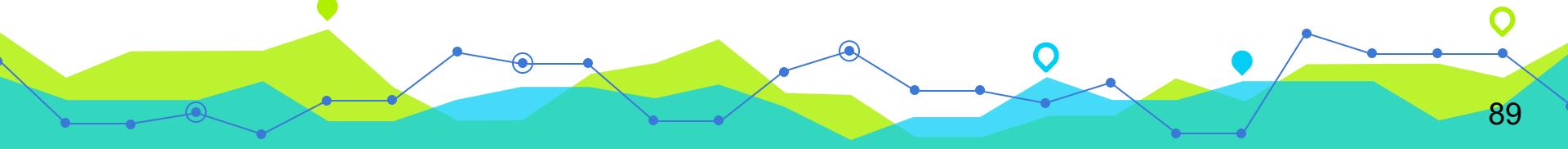
Phân cấp khái niệm (Nominal data)

VD: Tập hợp các thuộc tính: {day, week, month, quarter, year}



Biến đổi dữ liệu – Ví dụ

- Chuyển đổi giá trị logic thành 1, 0
- Chuyển đổi giá trị ngày tháng thành số
- Chuyển đổi các cột có giá trị số lớn thành tập các giá trị trong vùng nhỏ hơn (chia chúng cho hệ số nào đó)
- Nhóm các giá trị có cùng nghĩa: Nam bộ, Trung bộ, Tây nguyên, Bắc bộ,...
- Thay thế giá trị tuổi = {Child, Young, Mature, Old}



Tổng kết chương



Giới thiệu

- 1.Các dạng bộ dữ liệu
- 2.Đối tượng dữ liệu
- 3.Thuộc tính
- 4.Thu thập dữ liệu
- 5.Chất lượng của dữ liệu
- 6.Tiền xử lý dữ liệu
- 7.Các kỹ thuật tiền xử lý dữ liệu



Làm sạch dữ liệu (Data cleaning)

- 1.Làm sạch dữ liệu
- 2.Điền dữ liệu thiếu
- 3.Chỉnh sửa dữ liệu mâu thuẫn
- 4.Xử lý dữ liệu nhiễu



Tích hợp dữ liệu (Data integration)

- 1.Tích hợp dữ liệu
- 2.Loại bỏ dữ liệu dư thừa



Tổng kết chương



Rút gọn dữ liệu (Data reduction)

- 1.Rút gọn dữ liệu
- 2.Tổng hợp và tổng quát hóa
- 3.Giảm chiều dữ liệu
- 4.Giảm số lượng
- 5.Nén dữ liệu



Biến đổi, mã hóa dữ liệu (Data transformation)

- 1.Biến đổi, mã hóa dữ liệu
- 2.Tổng hợp và tổng quát hóa
- 3.Giảm chiều dữ liệu
- 4.Giảm số lượng
- 5.Nén dữ liệu



Tóm tắt

- **Chất lượng dữ liệu:** chính xác, đầy đủ, nhất quán, kịp thời, đáng tin cậy, có thể hiểu.
- **Thực tế dữ liệu:** thiếu, mâu thuẫn, nhiễu, nhiều chiều
- **Tiền xử lý dữ liệu:** là vấn đề quan trọng của KTDL
 - Làm sạch dữ liệu và tích hợp: thiếu, nhiễu, dư thừa, ...
 - Rút gọn dữ liệu: giảm chiều, giảm số lượng, nén dữ liệu, ...
 - Biến đổi, rời rạc hóa dữ liệu: chuẩn hóa, phân cấp khái niệm, ...
- Dữ liệu tốt là chìa khóa tạo ra các mô hình giá trị, đáng tin cậy.
- Lĩnh vực nghiên cứu còn nhiều thách thức

Bài tập chương 2

2.1. Giả sử CSDL có thuộc tính Tuổi với các giá trị trong các mẫu tin (tăng dần):

13, 15, 16, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35,
36, 40, 45, 46, 52, 70

- a. Sử dụng phương pháp chia giỏ theo độ sâu với số giỏ bằng 4, khử nhiễu dữ liệu trên bằng giá trị trung bình của giỏ
- b. Sử dụng dữ liệu trên vẽ biểu đồ cùng chiều rộng (equal-width histogram) với độ rộng = 10



THANKS!

Any questions?

