KHAI PHÁ DỮ LIỆU

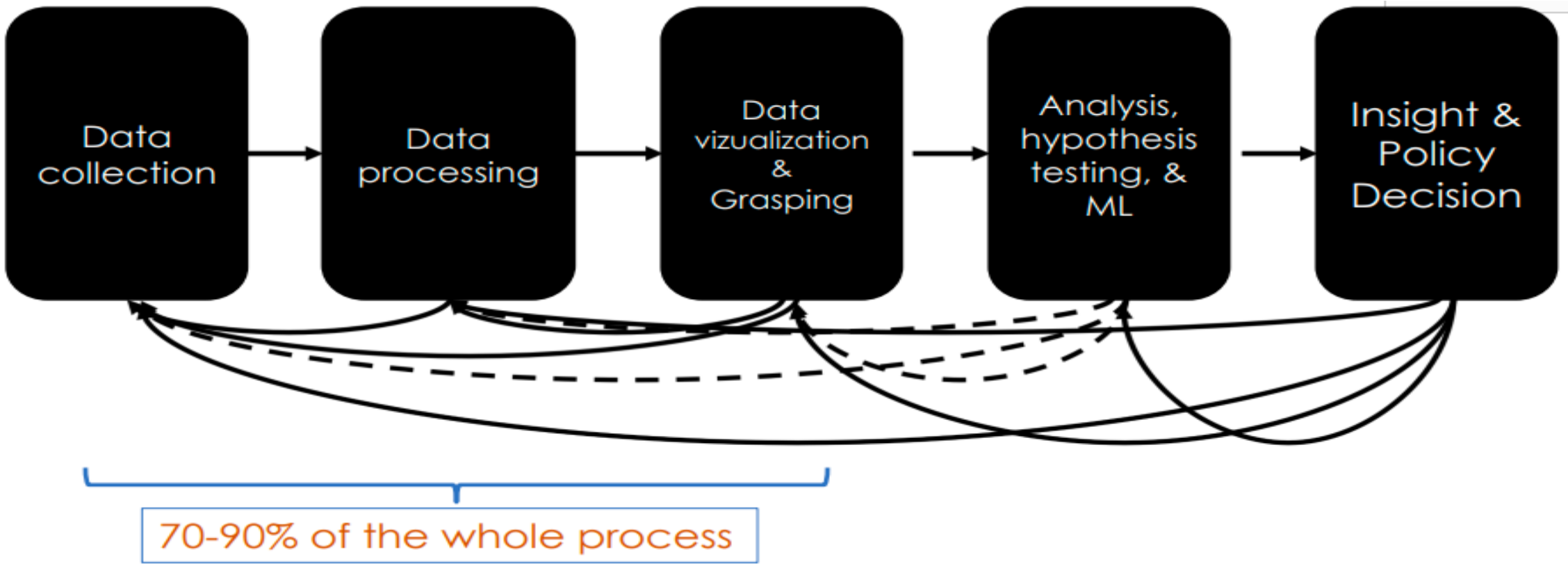
VO DUC QUANG – VINH UNIVERSITY

QUANGVD@VINHUNI.EDU.VN

Nội dung

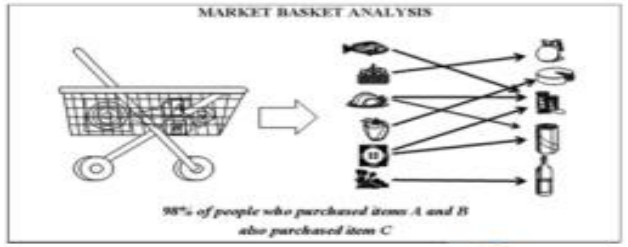
- Chương 1: Tổng quan về Data Mining
- **Chương 2: Dữ liệu và tiền xử lý dữ liệu**
- Chương 3: Bài toán phân lớp dữ liệu
- Chương 4: Bài toán phản cụm dữ liệu
- Chương 5: Khai phá luật kết hợp

Tiến trình khai phá dữ liệu

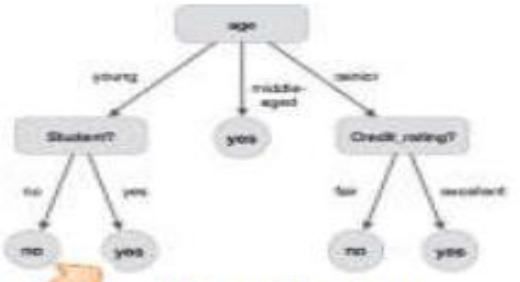


(John Dickerson, University of Maryland)

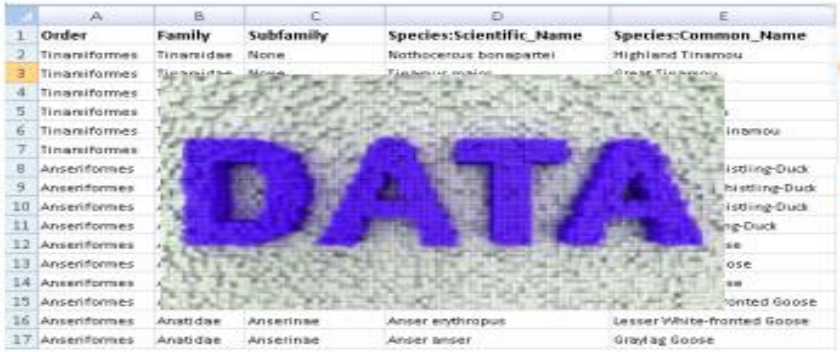
Một số bài toàn phổ biến

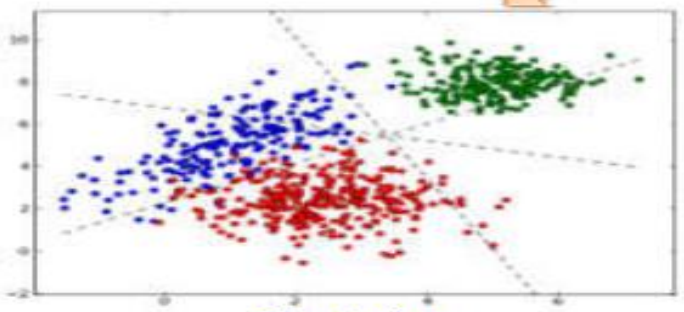


Association rules
(luật kết hợp)



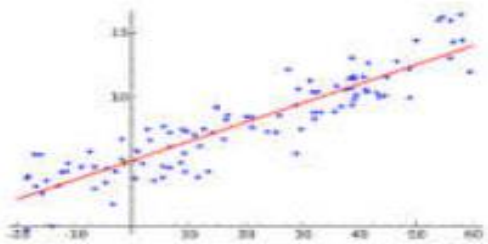
Classification
(phân lớp)



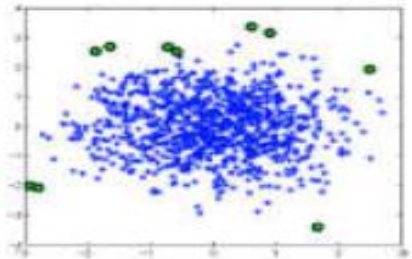


Clustering
(phân cụm)

Một số bài toán khác



Regression analysis
(phân tích hồi quy)



Outlier detection
(phát hiện ngoại lai)

Chương 2 – Dữ liệu: Thu thập và tiền xử lý

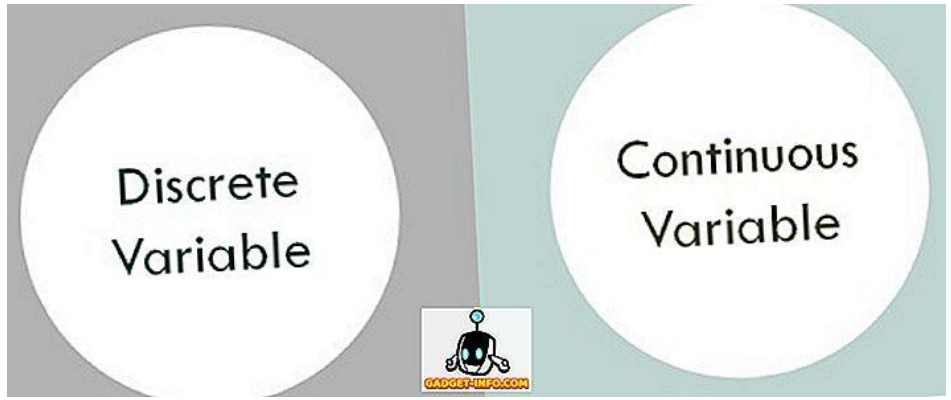
- Dữ liệu
- Thu thập dữ liệu (crawling)
- Xử lý dữ liệu
  - Làm sạch (Cleaning)
  - Tích hợp (Integrating)
  - Chuyển đổi (Transforming)

Dữ liệu

	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			
8	Uzbekistan	Europe	28541	28.9			
9	Uruguay	Americas	3395	22.05			

Một số khái niệm

- Nói chung, dữ liệu bao gồm những mệnh đề phản ánh thực tại, thường đo đạc hay quan sát về một đại lượng biến đổi. Các mệnh đề đó có thể bao gồm các số, từ hoặc hình ảnh
- Theo ‘điều 4 Luật Giao dịch điện tử’ ban hành ngày 29 tháng 11 năm 2005, Dữ liệu là thông tin dưới dạng ký hiệu,chữ viết, chữ số, hình ảnh, âm thanh hoặc dạng tương tự
- Các khái niệm
  - Nhị phân/Rời rạc/Liên tục
    - Nhị phân: 2 giá trị
    - Rời rạc: số lượng giá trị tối đa có thể đếm được
    - Liên tục: giá trị có thể là bất kỳ nằm trong một phạm vi



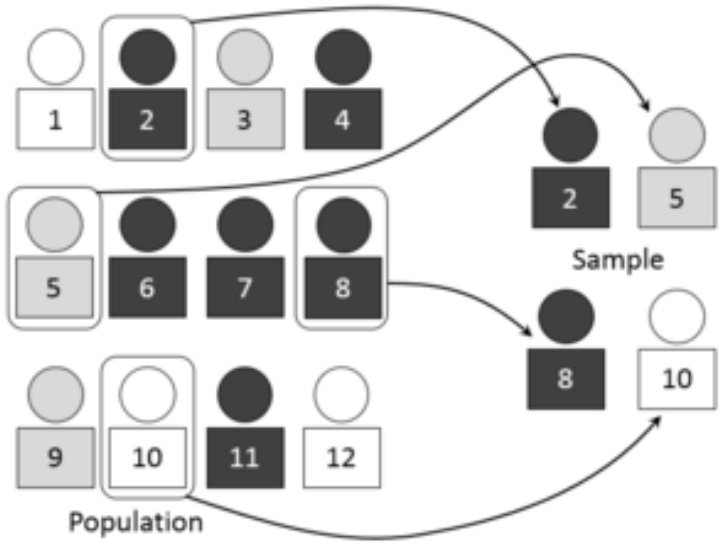
Hiểu về dữ liệu

- Định tính/Định lượng?
  - Dữ liệu định tính (thang đo danh nghĩa, thang đo thứ bậc): loại dữ liệu này liên quan đến mô tả phản ánh tính chất, ta **không tính toán được**
    - Ví dụ: giới tính: nam hay nữ; kết quả học tập của sinh viên: giỏi,khá, trung bình, yếu…
  - Dữ liệu định lượng (thang đo khoảng cách, thang đo tỉ lệ): loại dữ liệu thể hiện bằng con số thu thập được, các con số này có thể liên tục hay rời rạc, phản ánh mức độ, sự hơn kém và ta tính toán được.
  - Các phép toán cho dữ liệu định tính có những đặc điểm khác với phép toán dùng cho dữ liệu định lượng

Thu thập dữ liệu

- là quá trình thu thập và đo lường thông tin về các biến được nhắm mục tiêu trong một hệ thống đã được thiết lập, sau đó cho phép một người trả lời các câu hỏi có liên quan và đánh giá kết quả

- Các vấn đề liên quan
  - Phương pháp: khảo sát, phỏng vấn, quan sát,…
  - Chất lượng và tính toàn vẹn dữ liệu: thiếu, thừa, sai lệch,..,
  - Các kỹ thuật
    - Lấy mẫu (sampling)
    - Trong dữ liệu web: Crawling, logging, scraping

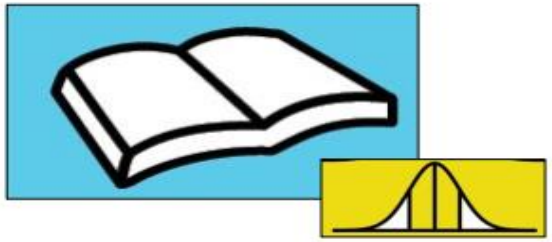


Thu thập dữ liệu

- Lấy mẫu
  - WHAT? lấy tập mẫu nhỏ, phổ biến để đại diện cho lĩnh vực cần "học"
  - WHY? không thể học toàn bộ. Giới hạn về thời gian và khả năng tính toàn
  - HOW?
    - **Variety** – tập mẫu thu được đủ đa dạng để phủ hết các ngữ cảnh của lĩnh vực
    - **Bias** – dữ liệu cần tổng quát, không bị sai lệch, thiên vị về 1 bộ phận nhỏ nào đó của lĩnh vực

Input

Vấn đề cần giải quyết



Output

Mẫu dữ liệu

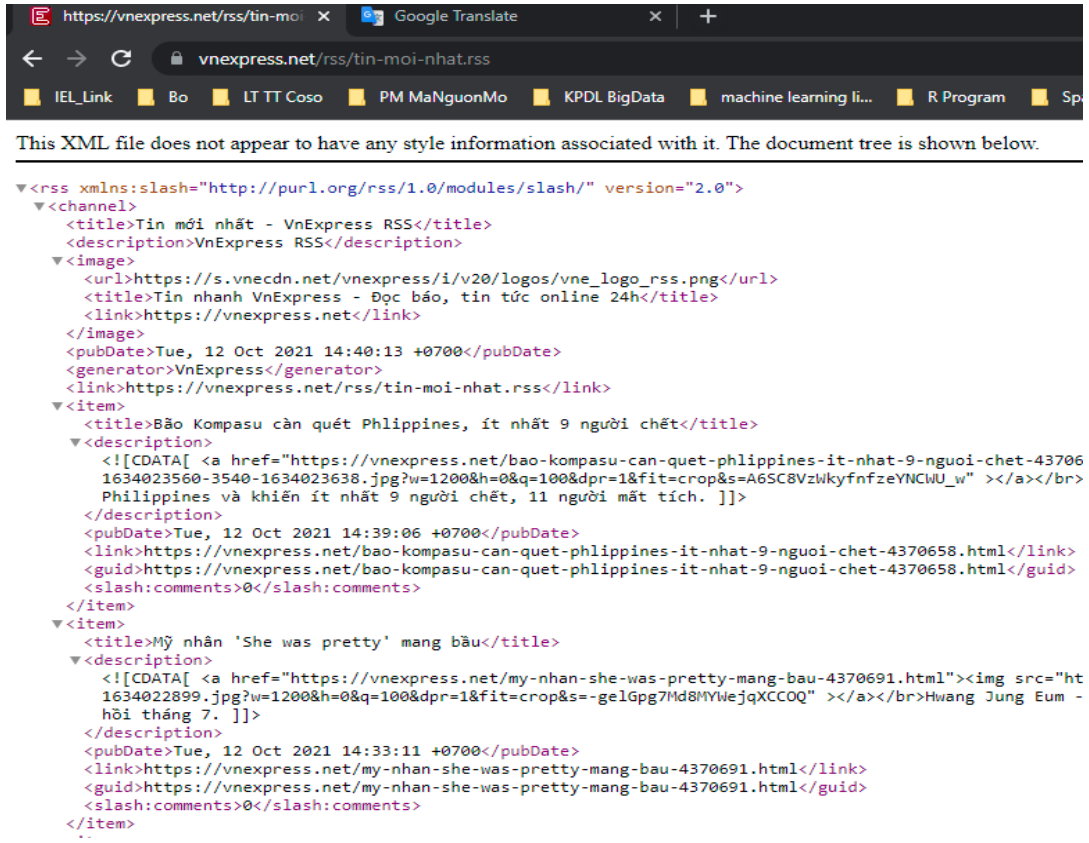
	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu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

Ví dụ

- Crawling, Scrapping

Ví dụ

- Thu thập dữ liệu

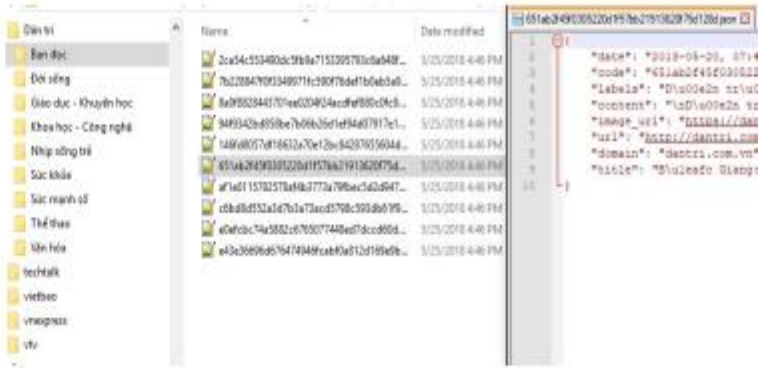


Input

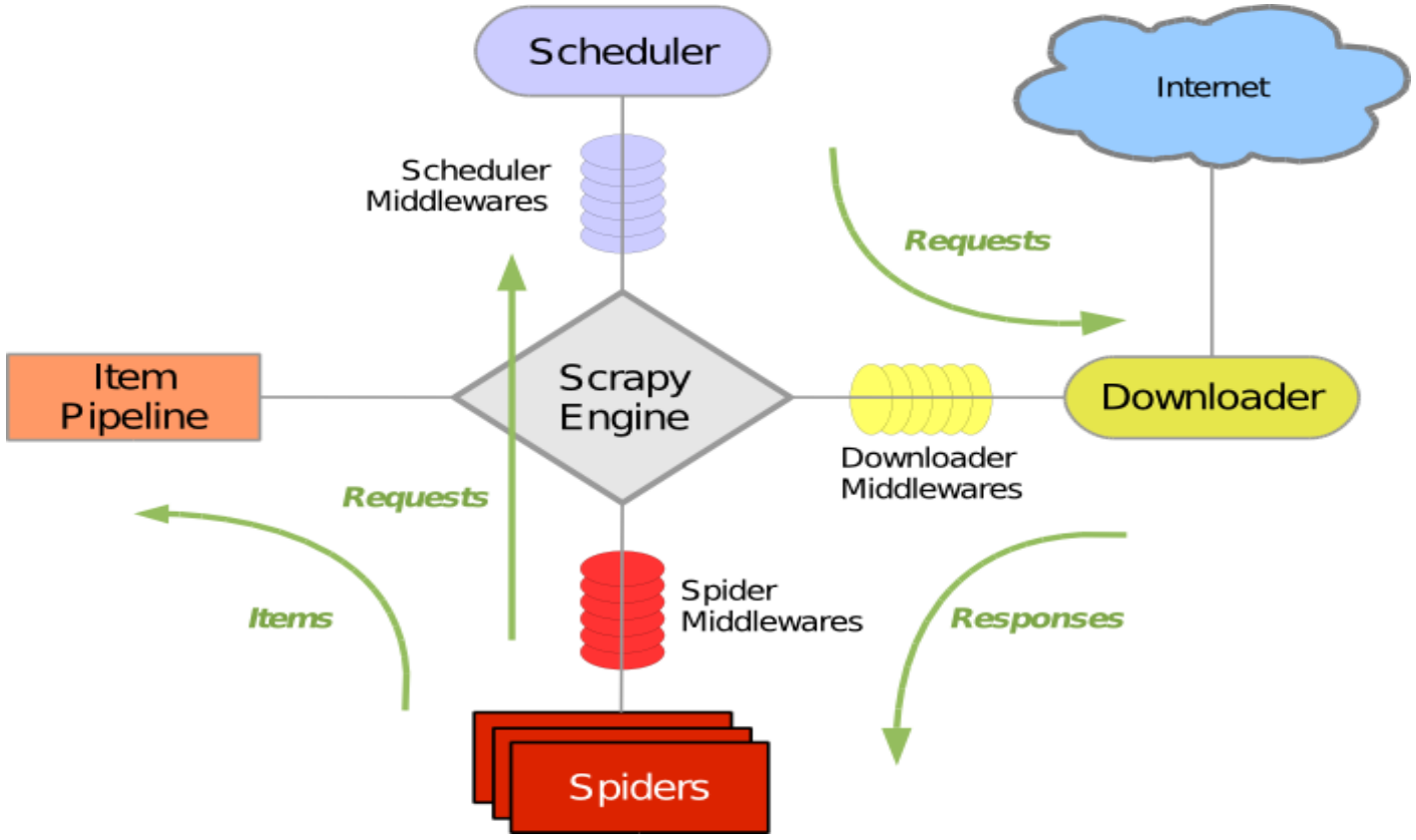
Vấn đề: phân loại văn bản báo chí

Output

Mẫu dữ liệu: báo chí và nhãn tương ứng



Ví dụ


Scheduler
Internet
Scheduler
Middlewares
Requests
Item
Pipeline
Scrapy
Engine
Downloader
Downloader
Middlewares
Requests
Items
Spider
Middlewares
Responses
Spiders

Xử lý dữ liệu

- Dữ liệu thô
  - Tính đầy đủ? (Completeness)
  - Tính trung thực? (Integrity)
    - Nguồn chính thống, chính xác
  - Tính đồng nhất? (Homogeneity)
    - Ví dụ: ngày sinh, vote
  - Tính cấu trúc? (Structures)
    - Chuẩn hóa
- **XỬ LÝ**
  - ✓Làm sạch (Cleaning)
  - ✓Tích hợp (Integrating)
  - ✓Chuyển đổi (Transforming)

Input

Mẫu dữ liệu thô

(text, ảnh, audio, ...)

Output

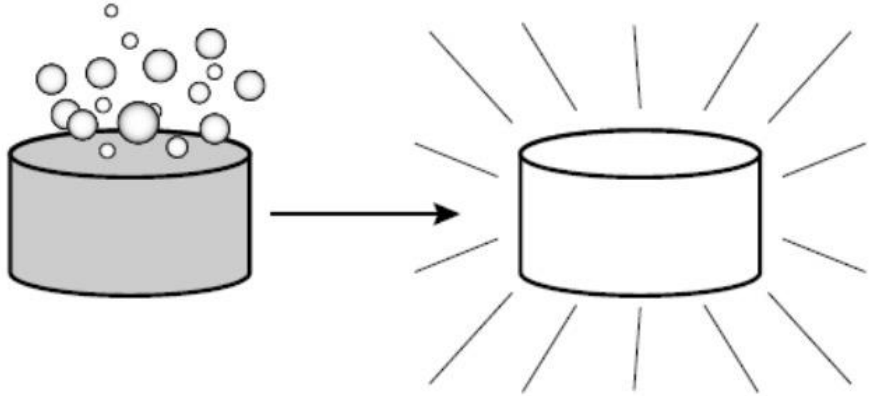
Dữ liệu số theo từng ML/AI model(s)

$$x^{(n)} = \begin{matrix} -0.0920 \\ 3.4931 \\ -1.8493 \\ \dots \\ \dots \\ -0.2010 \\ -1.3079 \end{matrix} \qquad \mathcal{D} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(n)} \end{bmatrix}$$

Làm sạch dữ liệu (Cleaning)

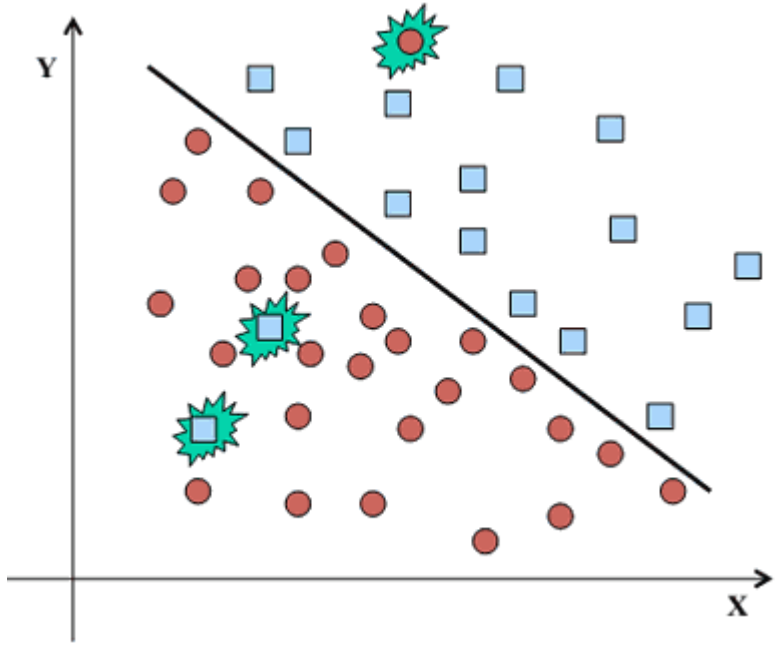
Đáp ứng:

- ✓Tính đầy đủ, trung thực
- ✓ Điền giá trị thiếu
- ✓ Tính đồng nhất



Cleaning: Đầy đủ trung thực

- Mẫu dữ liệu cần được thu thập từ các nguồn **đáng tin cậy**. Phản ánh vấn đề cần giải quyết
- Loại bỏ **nhiễu** (ngoại lai): bỏ vài mẫu dữ liệu mà có khác biệt lớn với các mẫu khác
- Một mẫu dữ liệu có thể bị trống (thiếu, chưa đầy đủ), cần có chiến lược phù hợp:
  - ✓Bỏ qua, không đưa vào phân tích
  - ✓Bổ sung các trường còn thiếu cho mẫu?



Cleaning: Điền các giá trị thiếu

- Điền lại giá trị bằng tay
- Gán cho giá trị nhãn đặc biệt hay ngoài khoảng biểu diễn
- Gán giá trị trung bình cho nó
- Gán giá trị trung bình của các mẫu khác thuộc cùng lớp đó
- Tìm giá trị có xác suất lớn nhất điền vào chỗ bị mất
  - (hồi quy, suy diễn Bayes,…)

bank.sav [DataSet9] - IBM SPSS Statisti

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 : jtype 1

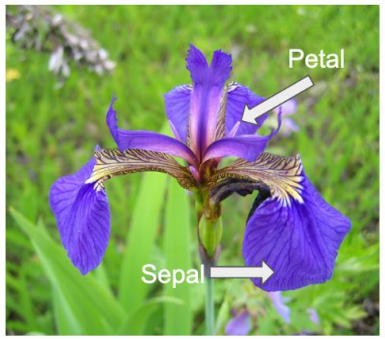
	educ	marit	start	jtype	whours	salary
1	.	2	07-May-2016	1	28.25	$1,6
2	4	1	27-Oct-2026	1	.	$1,7
3	5			1	22.75	$1,5
4	1			.	27.25	$1,9
5	3			1	.	$1,3
6	6	2	08-Dec-2016	2	43.75	$3.5

System missing values are indicated by dots.

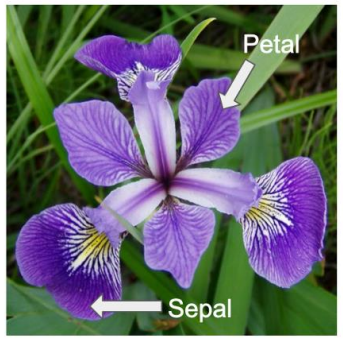
Cleaning: Tính đồng nhất

- Các mẫu dữ liệu cần có tính đồng nhất về cách biểu diễn, ký hiệu
- Ví dụ không đồng nhất:
  - *Rating "1, 2, 3" & "A, B, C";*
  - *Age = 42 & Birthday = 03/08/2020*

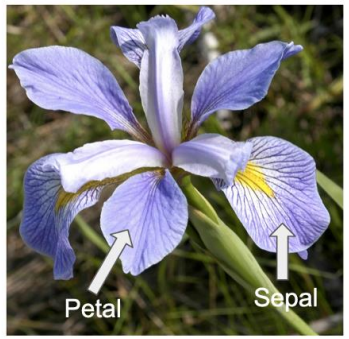
Iris setosa



Iris versicolor



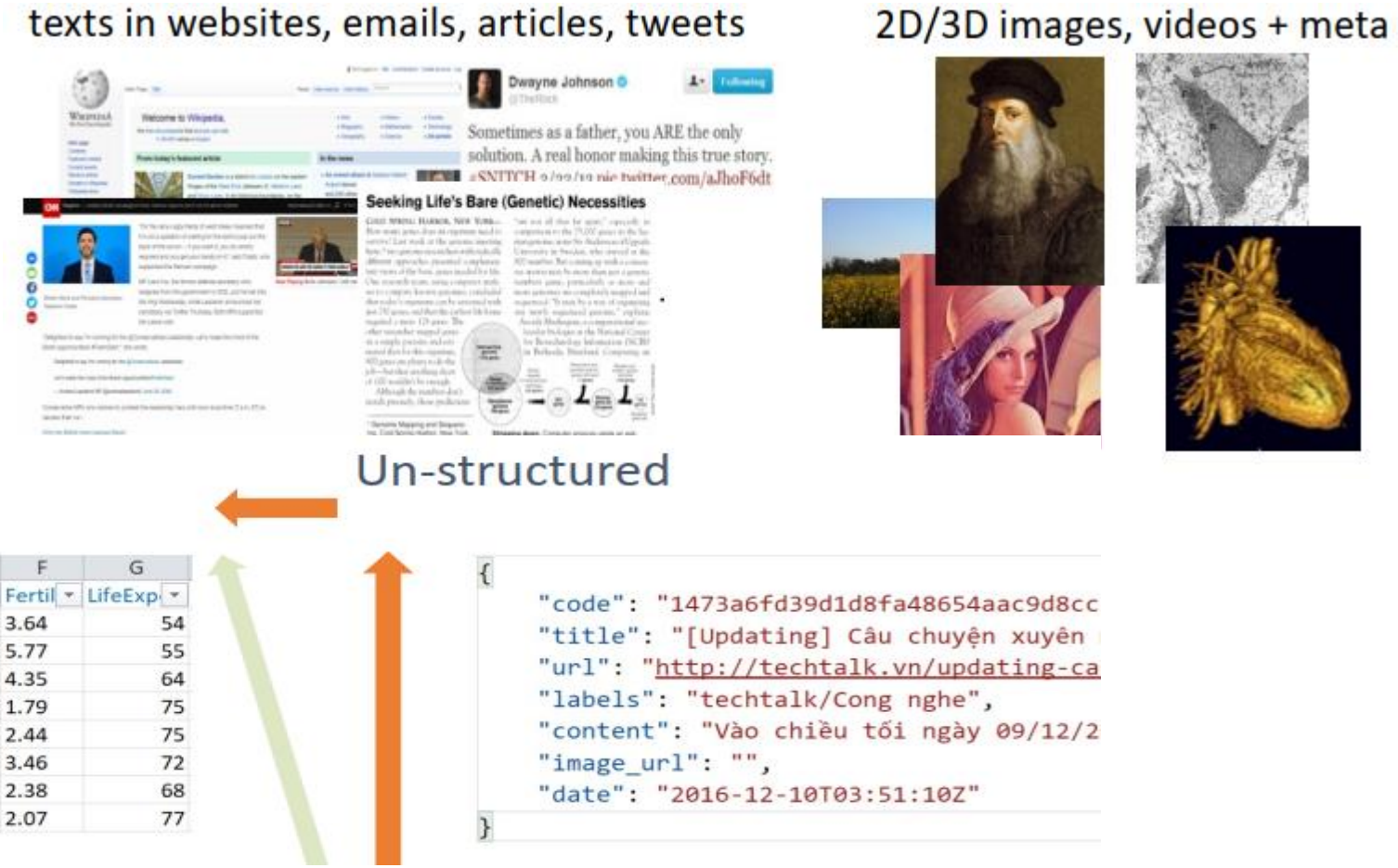
Iris virginica



	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
5	6	5.4	3.9	1.7	0.4	Iris-setosa
6	7	4.6	3.4	1.4	0.3	Iris-setosa
7	8	5.0	3.4	1.5	0.2	Iris-setosa
8	9	4.4	2.9	1.4	0.2	Iris-setosa
9	10	4.9	3.1	1.5	0.1	Iris-setosa

Tích hợp (Integrating)

- là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu có sẵn.
- Cần tránh:
  - Dư thừa dữ liệu
  - Mâu thuẫn dữ liệu
  - Trùng lặp dữ liệu



Chuyển đổi (Transforming)

- Mục tiêu: Trích xuất đặc trưng và chuẩn hóa
  - Rời rạc hóa đặc trưng: một số thuộc tính tỏ ra hiệu quả hơn khi được gom nhóm các giá trị
  - Chuẩn hóa đặc trưng: chuẩn hóa giá trị thuộc tính, về cùng một miền giá trị, dễ dàng trong tính toàn
- Giảm kích cỡ:
  - Giúp giảm kích thước của dữ liệu và đồng thời giữ được ngữ nghĩa cốt lõi của dữ liệu
  - Giúp tăng tốc quá trình học hoặc khai phá tri thức
- Một số chiến lược
  - Lựa chọn đặc trưng (feature selection): các thuộc tính không liên quan, dư thừa hoặc các chiều cũng có thể xóa hay loại bỏ
  - Giảm chiều (dimension reduction): dùng một số thuật toán (ví dụ PCA, ICA, LDA,…) để biến đổi dữ liệu ban đầu về không gian có ít chiều hơn

Chuẩn hóa dữ liệu

- Chuẩn hóa min-max
- Chuẩn hóa z-score
- Chuẩn hóa decimal scaling

Chuẩn hóa min-max

- Chuẩn hoá *min-max*
  - Thực hiện chuyển đổi tuyến tính dựa trên dữ liệu gốc
  - $v \in [min_A, max_A]$
    - giá trị nhỏ nhất và lớn nhất của thuộc tính *A*
  - $v' \in [new\_min_A, new\_max_A]$
    - giá trị nhỏ nhất và lớn nhất chuyển đổi tương ứng

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

Min-max normalization. Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range [0.0, 1.0]. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$. ■

Chuẩn hóa z-score

❑ Chuẩn hoá z-score

- Các giá trị ứng với thuộc tính *A* được chuẩn hoá dựa trên giá trị trung bình (mean(A) hay $\bar{A}$) và độ lệch chuẩn của *A* (σ_A).
- Giá trị *v* của *A* sẽ được chuẩn hoá tương ứng với giá trị *v'* thông qua công thức

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

- Ví dụ

z-score normalization. Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$. ■

Chuẩn hóa decimal scaling

❑ Chuẩn hoá decimal scaling

- Thay đổi giá trị của thuộc tính *A* theo hệ số 10. Mức độ thay đổi phụ thuộc vào giá trị tuyệt đối lớn nhất của *A*.
- Cách chuyển:

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

- Trong đó: *j* là giá trị nguyên nhỏ nhất sao cho $max(|v'|) < 1$
- Ví dụ:

Decimal scaling. Suppose that the recorded values of *A* range from −986 to 917. The maximum absolute value of *A* is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that −986 normalizes to −0.986 and 917 normalizes to 0.917. ■

Một số kiến thức đo lường cơ bản

- Giá trị trung bình (Mean)
  - Xét dãy gồm N giá trị $\{x_1, x_2, \ldots, x_N\}$
  - Giá trị trung bình (mean) được xác định như sau, ký hiệu $\bar{x}$ hoặc $mean(x)$

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

Mean. Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110 , we have

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is $58,000. ■

Một số kiến thức đo lường cơ bản

- Trung bình (Mean)
  - Nếu mỗi giá trị x_i có một trọng số w_i đi kèm thì giá trị trung bình được gọi là trung bình dựa trên trọng số (weighted average)

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

Một số kiến thức đo lường cơ bản

- Trung vị (Median)
  - Xét dãy gồm N giá trị $\{x_1, x_2, \ldots, x_N\}$ được sắp có thứ tự
    - Nếu N lẻ: $median(x) = x_{\left[\frac{N}{2}\right]+1}$ (phần tử chính giữa)
    - Nếu N chẵn: $median(x) = (x_{\left[\frac{N}{2}\right]} + x_{\left[\frac{N}{2}\right]+1})/2$ (trung bình cộng của 2 phần tử giữa dãy)
- Ví dụ:
  - Tìm giá trị trùng bình, trung vị của dãy: **2,3,4,7,8,9,89,200 ???**

Một số kiến thức đo lường cơ bản

- Số trội (mode): là giá trị có tần suất xuất hiện nhiều nhất trong tập dữ liệu đang xem xét

$$mode(x) = x_i \text{ nếu } count(x_i) \text{ là } MAX$$

- Ví dụ: Tìm mode của tập giá trị: **2,3,8,5,3,12,3,5,2,32,15**
- Khoảng trung bình (midrange): xác định độ tập trung của dữ liệu, là giá trị trung bình cộng của giá trị lớn nhất và nhỏ nhất trong tập dữ liệu

$$midrange(x) = mean(max(x), min(x))$$

Một số kiến thức đo lường cơ bản

- Phương sai (Variance)

Phương sai của N giá trị $x_1, x_2, \ldots, x_N$ được xác định:

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

- Độ lệch chuẩn (Standard deriation): Bằng căn bậc 2 của phương sai

Độ đo khoảng cách

- Dữ liệu đo lường tuyến tính (interval-scaled)
  - Dữ liệu số biểu diễn các thuộc tính như: trọng lượng (weight), chiều cao (height), toạ độ (latitude, longitude), nhiệt độ,…
  - Thay đổi đơn vị đo sẽ làm thay đổi khoảng cách
  - Khoảng cách Euclidean (i.e., straight line)

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

  - Khoảng cách Manhattan (i.e., city block)

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

  - Khoảng cách Minkowski
  - …

Độ đo khoảng cách

Dữ liệu phân loại (nominal/categorical)

- Khoảng cách giữa đối tượng *i* và *j*

$$d(i, j) = \frac{p - m}{p}$$

- Trong đó:
  - *p*: số thuộc tính phân loại quan tâm
  - *m*: thuộc tính mà *i*, *j* có giá trị giống nhau
- Ví dụ:
  - Chỉ quan tâm đến thuộc tính *test*-1
  - *d*(1, 2) = (1-0)/1 = 1
  - *d*(4, 1) = (1-1)/1 = 0

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Dữ liệu

Là tập hợp các đối tượng (Objects) và các thuộc tính của chúng (Attributes)

Thuộc tính là một tính chất riêng biệt của đối tượng hay đặc tính mô tả đối tượng

- Ví dụ: màu mắt của một người, nhiệt độ, …

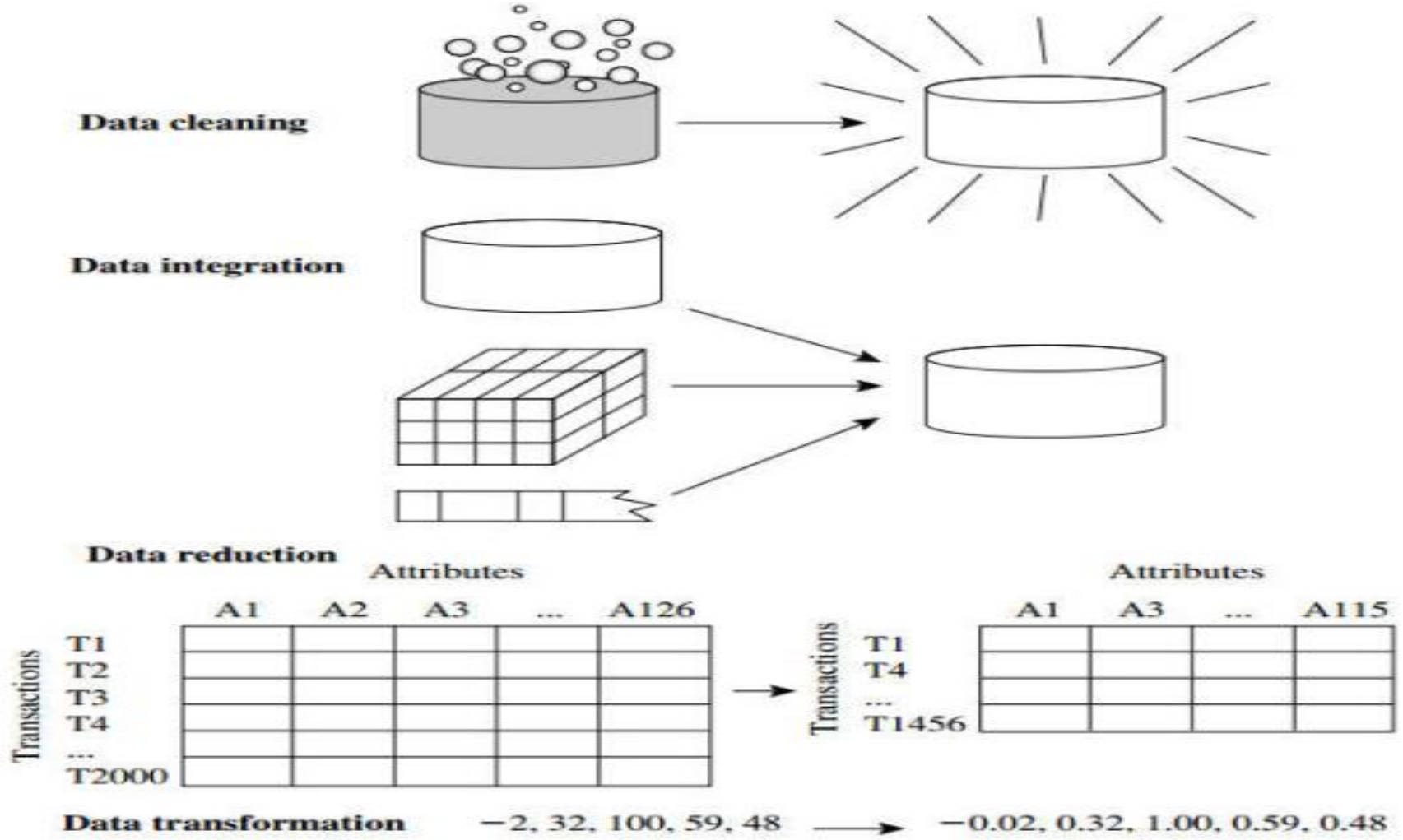
Một tập các thuộc tính dùng để mô tả một đối tượng

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
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

Tiền xử lý dữ liệu



Tóm tắt chương

- Dữ liệu trong một lĩnh vực trước khi vào hệ thống học máy phải được thu thập và biểu diễn thành dạng cấu trúc với một số đặc tính: đầy đủ, ít nhiễu, nhất quán, có cấu trúc xác định
- Dữ liệu thu thập cho quá trình học là tập nhỏ, tuy vậy cần phản ánh đầy đủ các mặt vấn đề cần giải quyết
- Dữ liệu thô sau khi thu thập và tiền xử lý phải giữ được sự đầy đủ các đặc trưng ngữ nghĩa – các đặc trưng ảnh hưởng đến khả năng giải quyết vấn đề
- Cần thiết phải: sử dụng được công cụ lập trình (thư viện)+nắm vững các kiến thức cơ bản