# Chương 2 : Chuẩn bị dữ liệu và tiền xử lý dữ liệu

Trịnh Anh Phúc [1]

[1]Bộ môn Khoa Học Máy Tính, Viện CNTT & TT,
Trường Đại Học Bách Khoa Hà Nội

Ngày 2 tháng 5 năm 2014

# Giới thiệu

1. Tại sao phải tiền xử lý dữ liệu ?
   - Vấn đề bản thân dữ liệu
   - Quá trình tiền xử lý dữ liệu
2. Tổng hợp thông tin ban đầu dữ liệu
   - Đo sự tập trung (Measure of Central Tendency)
   - Đo sự phân tán (Measure of Dispersion)
   - Biểu diễn thông tin tổng hợp
3. Làm sạch dữ liệu
   - Điền lại dữ liệu bị mất
   - Chuốt dữ liệu để loại nhiễu
   - Kiểm tra và sửa tính không nhất quán
4. Tích hợp và chuyển đổi dữ liệu
5. Giảm chiều
   - Lựa chọn tập con các thuộc tính
   - Giảm chiều
   - Rời rạc hóa và trừu tượng khái niệm

# Tại sao phải tiền xử lý dữ liệu ?

## Vấn đề của bản thân dữ liệu

Dù muốn hay không dữ liệu thường luôn gặp những vấn đề như sau

- Không đầy đủ (Incomplete) : thiếu một vài giá trị thuộc trường dữ liêu hay thuộc tính ....

**Table I**
Illustration of hot deck imputation: incomplete data set

| Case | Item 1 | Item 2 | Item 3 | Item 4 |
|---|---|---|---|---|
| **1** | 10 | 2 | 3 | 5 |
| **2** | 13 | 10 | 3 | 13 |
| **3** | 5 | 10 | 3 | ??? |
| **4** | 2 | 5 | 10 | 2 |

# Tại sao phải tiền xử lý dữ liệu ?

## Nguyên nhân xảy ra không đầy đủ dữ liệu (Incomplete data)

có thể là ...

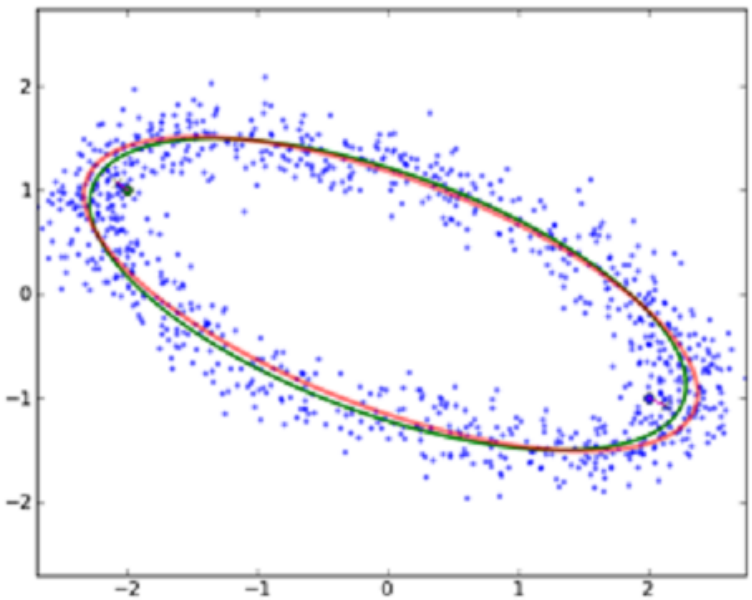
- trường dữ liệu quan tâm không có sẵn, e.g. lý do người dùng ko thích một sản phẩm
- dữ liệu không xuất hiện do đơn giản tại thời điểm thu thập dữ liệu nó không đc coi trọng
- dữ liệu liên quan không được ghi do lỗi thiết bị
- dữ liệu bị xóa do không nhất quan với các dữ liệu khác

$\Rightarrow$ đối với các giá trị bị mất, ta có thể suy diễn để điền lại đặc biệt với dữ liệu dạng bộ

# Tại sao phải tiền xử lý dữ liệu ?

## Vấn đề của bản thân dữ liệu

- Nhiễu (Noisy) : xuất hiện giá trị lỗi, lỗi chủ quan người nhập dữ liệu ....



# Tại sao phải tiền xử lý dữ liệu ?

## Nguyên nhân xảy ra nhiễu dữ liệu (Noisy data)

có thể là ...

- thiết bị ghi nhận bị hỏng
- lỗi khách quan và chủ quan khi có nhiều tác nhân thu thập dữ liệu
- lỗi do việc truyền dữ liệu
- lỗi công nghệ của thiết bị
- dữ liệu không chính xác cũng có thể do việc không nhất quan khi đặt giá trị

# Tại sao phải tiền xử lý dữ liệu ?

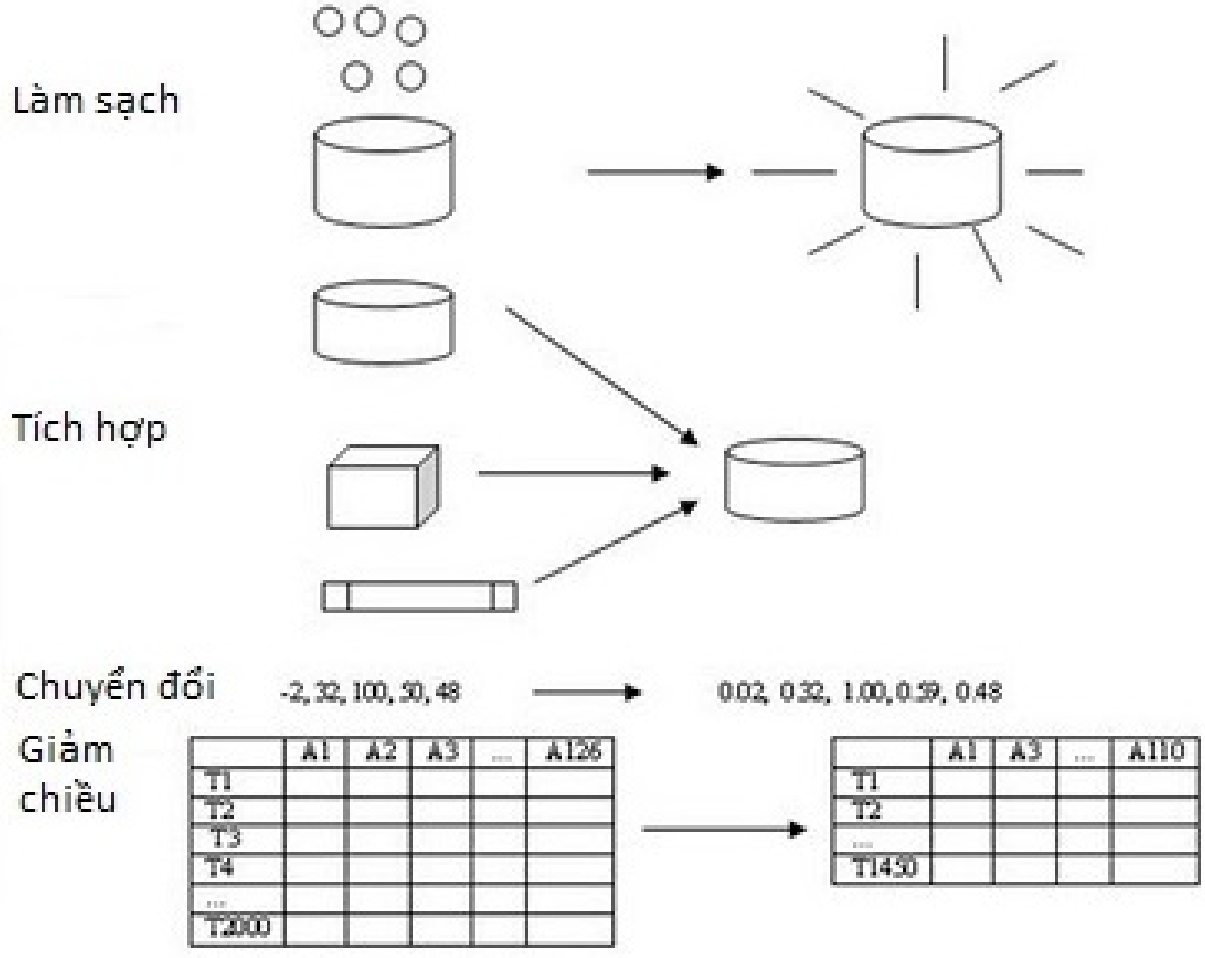
## Vấn đề của bản thân dữ liệu

- Không nhất quán (Inconsistent) : sự khác biệt trong cách phân loại, phân biệt hay logic của dữ liệu ...

| Time | TX 1 | TX 2 | TX 3 |
|---|---|---|---|
| T0 | Begin TX | Begin TX | Begin TX |
| T1 | Update A | Read A | Update B |
| T2 | | Read B | |
| T3 | Commit | Use values of A and B to update C | Commit |
| T4 | | Commit | |

# Tại sao phải tiền xử lý dữ liệu ?

## Qui trình tiền xử lý dữ liệu



# Tại sao phải tiền xử lý dữ liệu ?

## Quá trình tiền xử lý dữ liệu gồm

- Làm sạch : Loại bỏ các giá trị sai, kiểm tra tính nhất quan của dữ liệu
- Tích hợp : Dữ liệu có nhiều nguồn nên cần lưu theo một cách thức thống nhất
- Chuyển đổi : Chuẩn hóa và tập hợp dữ liệu
- Giảm chiều : Mô tả dữ liệu trong kích thước nhỏ nhưng không làm mất kết quả cần kiết xuất

# Tại sao phải tiền xử lý dữ liệu ?

## Dữ liệu xuất hiện sau tiền xử lý cần có thuộc tính

- Đúng đắn (Accuracy)
- Đầy đủ (Completeness)
- Nhất quán (Consistency)
- Trình tự (Timeliness)
- Giải thích (Interpretability)
- Truy cập (Accessibility)
- Khung cảnh (Contextual)
- Mô tả (Representational)

# Dành cho trả lời câu hỏi

# Tổng hợp thông tin ban đầu dữ liệu

## Ý nghĩa việc tổng hợp

- Cần một bức tranh toàn cảnh về dữ liệu hiện tại
- Xác định các thuộc tính sẵn có của dữ liệu đồng thời nổi bật các thành phần nhiễu hay không đáng quan tâm

# Tổng hợp thông tin ban đầu dữ liệu

## Hai hướng tổng hợp chính

- Tập trung (Central Tendency) : xác định giá trị trung tâm của dữ liệu
- Phân tán (Dispersion) : xác định sự dao động của dữ liệu quanh vị trí trung tâm

$\Rightarrow$ Sử dụng các giá trị mô tả trong thống kê

- Tập trung : trung bình (mean), trung vị (median), mode (xuất hiện nhiều nhất), chính giữa (midrange)
- Phân tán : phương sai (variance), phạm vi (range), tứ phân vị (quatiles), độ trải giữa (InterQuartile Range - IQR)

# Tổng hợp thông tin ban đầu dữ liệu

## Đo sự tập trung

Cho tập $x_1, x_2, ..., x_N$ là tập $N$ giá trị quan sát được, ví du như của một thuộc tính dữ liệu

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + ... + x_N}{N} \quad (1)$$

Công thức (1) tính giá trị trung bình cho tập giá trị $x_i, i = \overline{1, N}$

## Ví dụ

Cho tập các số {1, 2, 2, 3, 4, 7, 9} thì giá trị trung bình là 4

# Tổng hợp thông tin ban đầu dữ liệu

## Đo sự tập trung (tiếp)

Đôi khi có trọng số $w_i$ đi kèm $x_i, i = \overline{1, N}$ thì công thức trở thành

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i w_i}{\sum_{i=1}^{N} w_i} = \frac{x_1 w_1 + x_2 w_2 + ... + x_N w_N}{w_1 + w_2 + ... + w_N} \quad (2)$$

Công thức (2) tính giá trị trung bình có trọng số cho tập giá trị $x_i, i = \overline{1, N}$

# Tổng hợp thông tin ban đầu dữ liệu

## Đo sự tập trung (tiếp)

Số *trung vị* của danh sách $x_1, x_2, ..., x_N$, ta xếp tăng dần tất cả các giá trị quan sát, rồi lấy $x_{N/2}$ nằm giữa danh sách. Nếu $N$ là số chẵn, người ta thường lấy trung bình của hai giá trị nằm giữa.

## Ví dụ

Cho tập các số {1, 2, 2, 3, 4, 7, 9} thì số trung vị là 3

# Tổng hợp thông tin ban đầu dữ liệu

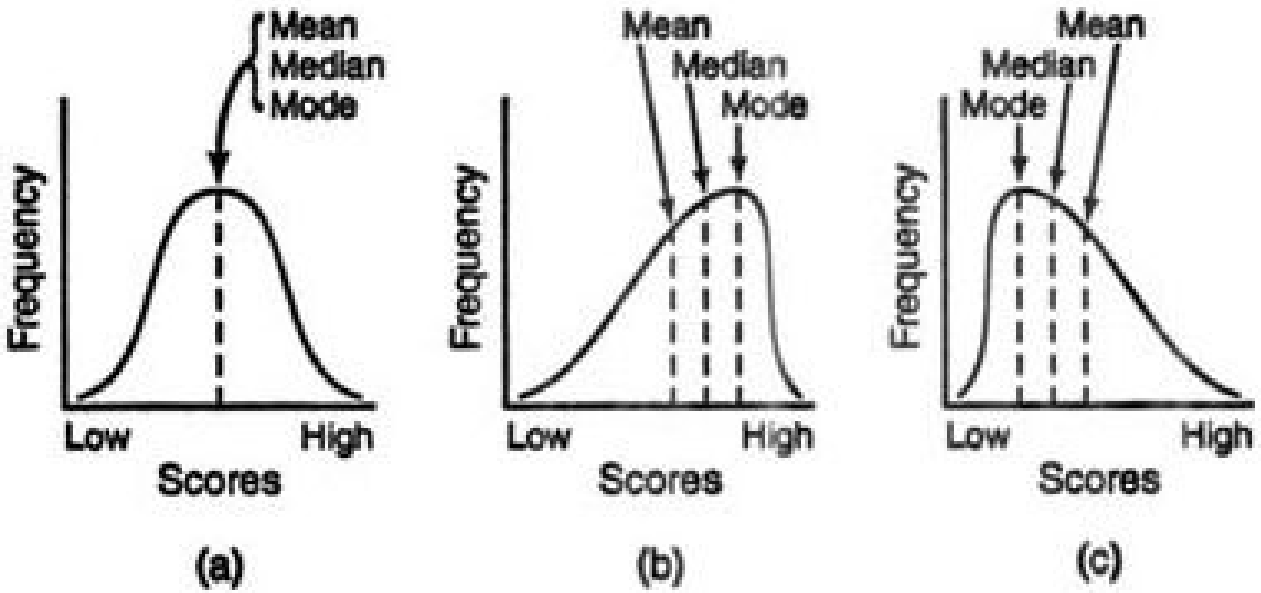
## Đo sự tập trung (tiếp)

Cho tập $x_1, x_2, ..., x_N$ là tập $N$ giá trị quan sát được, số mode là giá trị xuất hiện nhiều nhất

## Ví dụ

Cho tập các số {1, 2, 2, 3, 4, 7, 9} thì số mode là 2

# Tổng hợp thông tin ban đầu dữ liệu




Measures of Central Tendency

# Tổng hợp thông tin ban đầu dữ liệu

## Đo sự phân tán

Cho tập $x_1, x_2, ..., x_N$ là tập $N$ giá trị quan sát được

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 = \frac{1}{N}\left(x_i - \frac{1}{N}\sum_{i=1}^{N}x_i\right)^2 \quad (3)$$

Công thức (3) tính giá trị phương sai hay độ lệch chuẩn

## Ví dụ

Cho tập các số {1, 2, 2, 3, 4, 7, 9} thì giá trị phương sai $\sigma^2 = 7.42857$

# Tổng hợp thông tin ban đầu dữ liệu

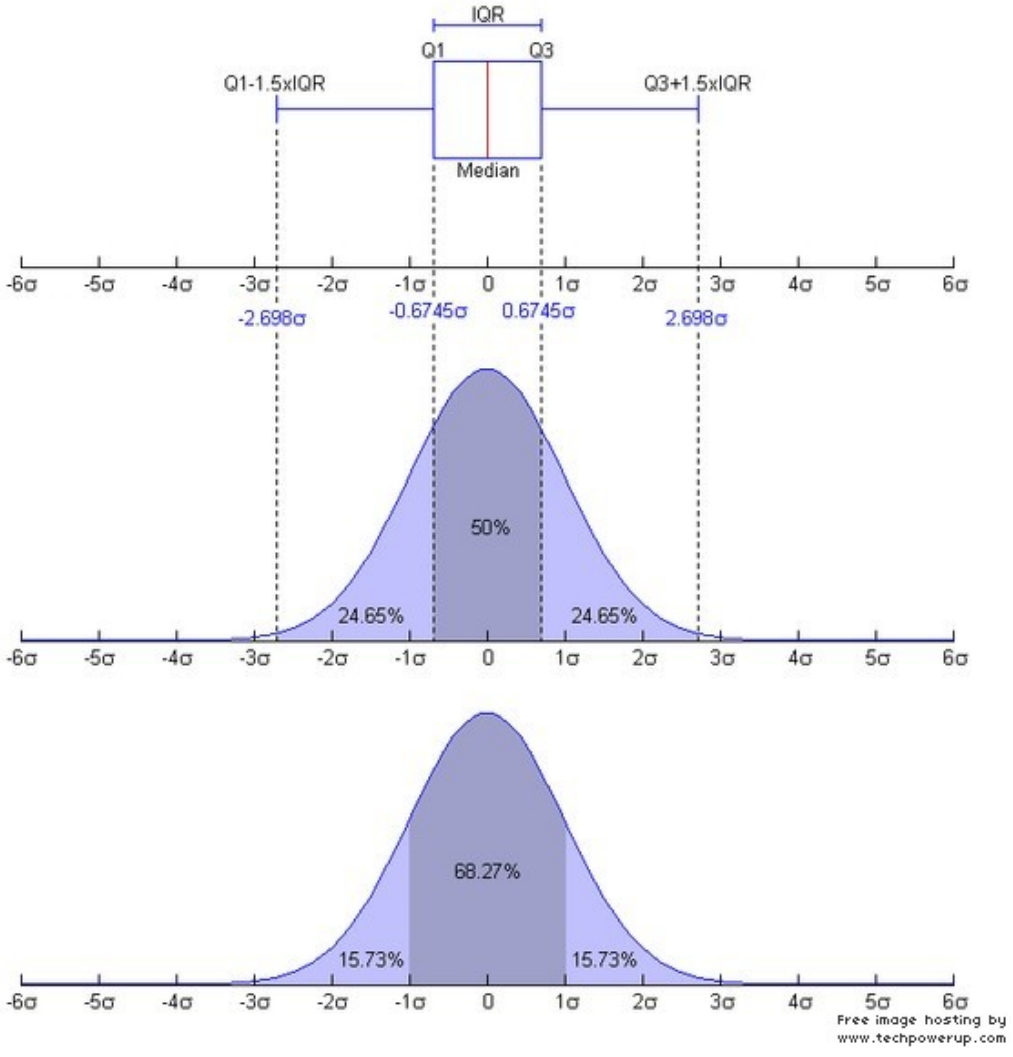
## Đo sự phân tán (tiếp)

Cho tập $x_1, x_2, ..., x_N$ là tập $N$ giá trị quan sát được, ta còn có các độ đo khác như

- Phạm vi (Range) : khoảng giữa giá trị nhỏ nhất $x_{min}$ và lớn nhất $x_{max}$ trong tập
- Tứ phân vị (Quatiles) : gồm có 3 giá trị là thứ nhất $Q_1$, thứ nhì $Q_2$ và thứ ba $Q_3$. Ba giá trị này chia một tập hợp dữ liệu (đã sắp xếp dữ liệu theo trật từ từ bé đến lớn) thành 4 phần có số lượng quan sát đều nhau.
- Độ trải giữa (Interquartile range) : Đại lượng này được tính ra bằng cách lấy giá trị tứ phân vị thứ ba trừ đi giá trị tứ phân vị thứ nhất

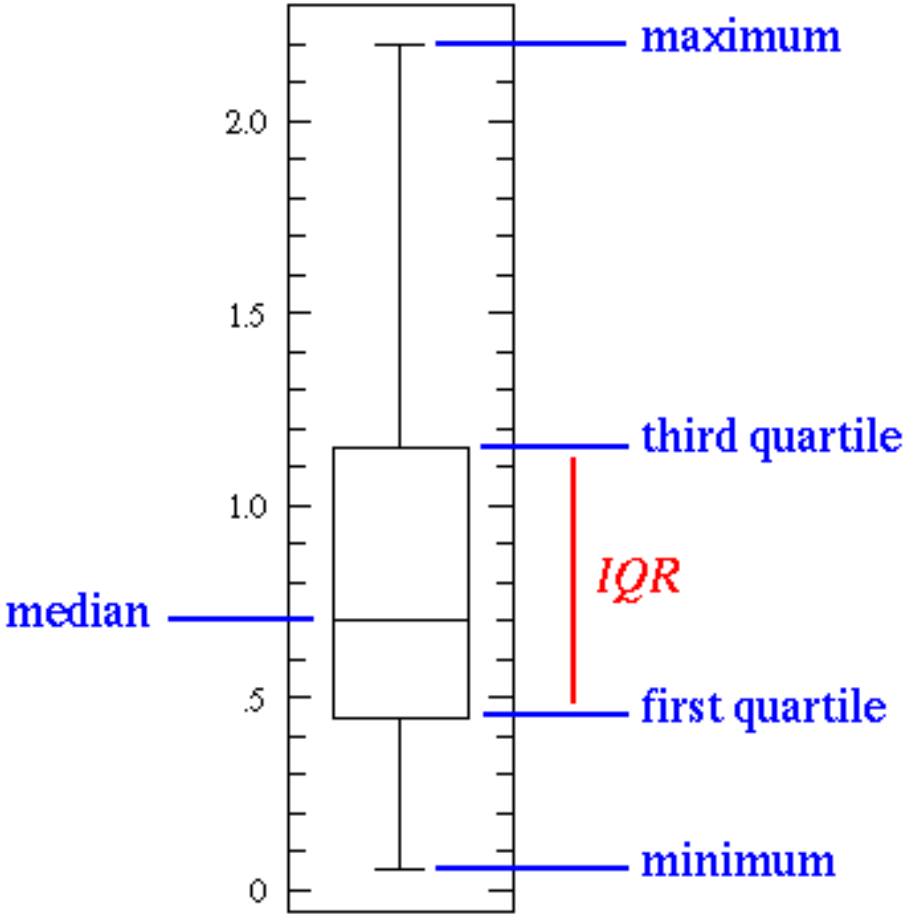
$$IQR = Q_3 - Q_1$$

# Tổng hợp thông tin ban đầu dữ liệu



# Tổng hợp thông tin ban đầu dữ liệu

## Biểu diễn phân bố dư liệu bằng Boxplots



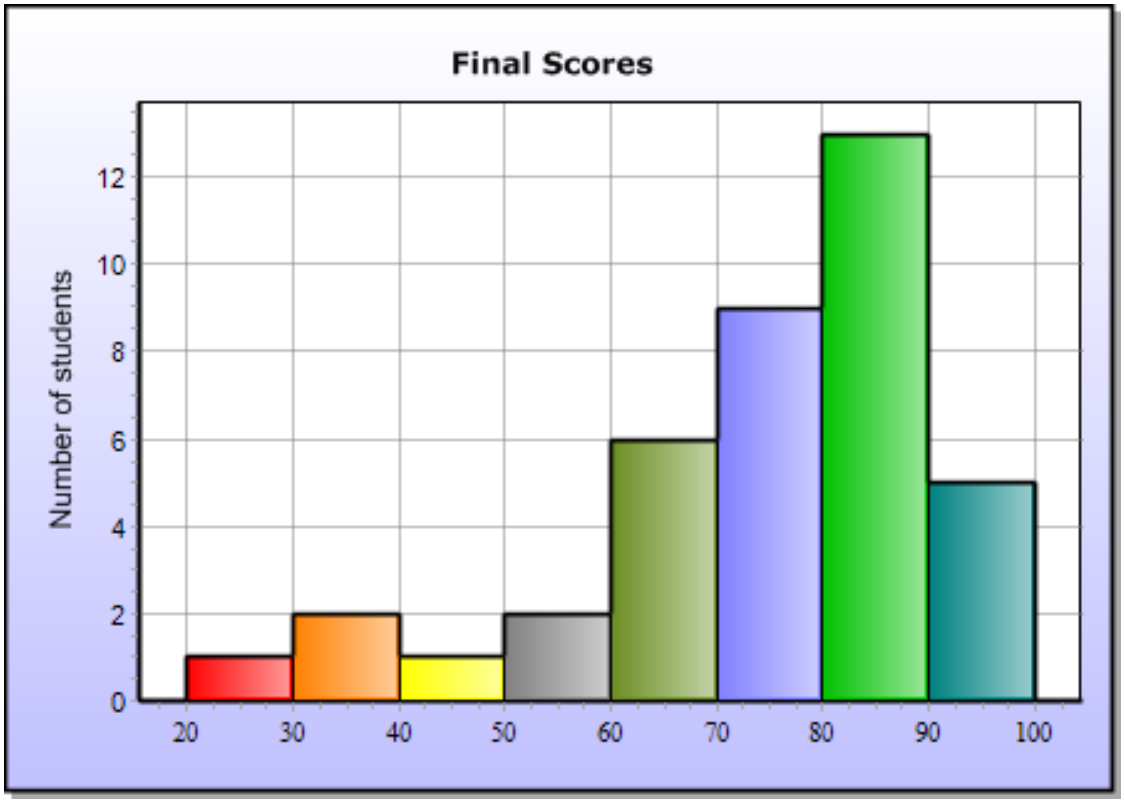

Gồm năm giá trị

- $x_{max}$ và $x_{min}$ tương ứng phạm vi
- $Q_3$ và $Q_1$ tương ứng độ trải giữa
- trung vị

# Tổng hợp thông tin ban đầu dữ liệu

## Biểu đồ tần xuất (Histogram)

Gồm các cột hình chữ nhật, chiều cao mỗi cột là số lần xuất hiện của dữ liệu quan sát được



# Tổng hợp thông tin ban đầu dữ liệu

## Đồ thị phân vị (Quantile plot)

Cho tập $x_1, x_2, ..., x_N$ là tập $N$ giá trị được sắp xếp từ nhỏ đến lớn mỗi số được ghép với $f_i, i = \overline{1, N}$ là số phần trăm xuất hiện của các giá trị nhỏ hơn bằng $x_i$.

$$f_i = \frac{i - 0.5}{N} \tag{4}$$

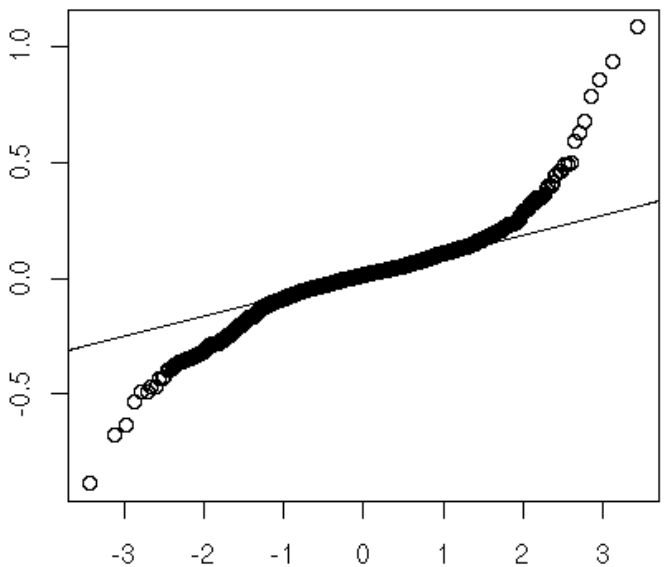
Đồ thị vẽ quan hệ giá trị $(x_i, f_i)$ đc gọi là đồ thị phân vị

# Tổng hợp thông tin ban đầu dữ liệu

## Đồ thị phân vị-phân vị (Quantile-quantile plot)

Cho hai tập dữ liệu $x_1, ..., x_N$ và $y_1, ..., y_M$ đều được sắp xếp theo giá trị nhỏ đến lớn

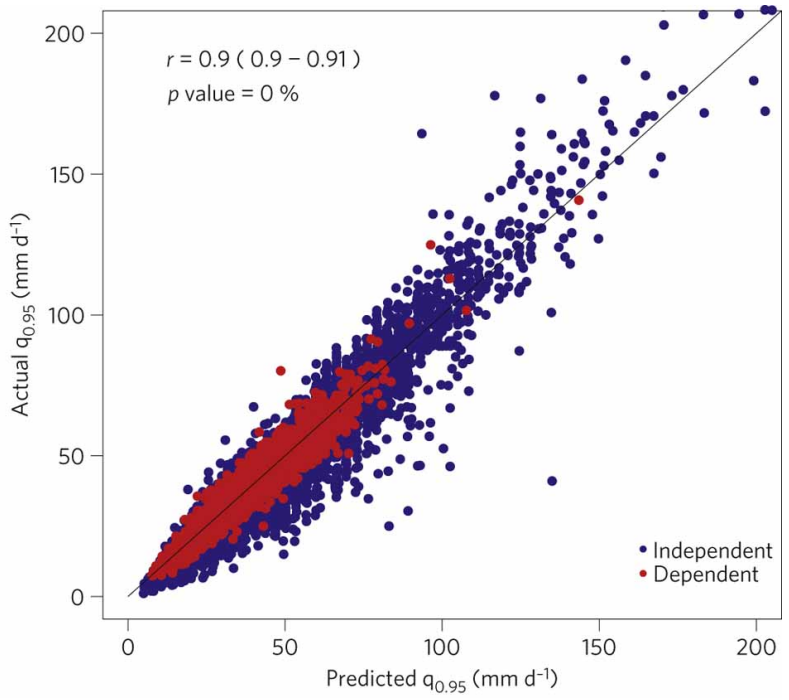
- $N = M$ ta chỉ cần vẽ đồ thị $(x_i, y_i)$ với $i = \overline{1, N}$
- $M < N$ ta vẽ $M$ điểm $(x_i, y_i)$ với $f_i = (i - 0.5)/M$ và $i = \overline{1, M}$



# Tổng hợp thông tin ban đầu dữ liệu

## Đồ thị phân tán (scatter plot)

Giống đồ thị phân vị-phân vị nêu quan hệ giữa hai thuộc tính thuộc bảng dữ liệu



# Dành cho trả lời câu hỏi

# Làm sạch dữ liệu

## Mục đích

Đây là thủ tục quan trọng gồm ba phần chính

- điền đầy các giá trị bị mất
- chuốt dữ liệu để loại nhiễu
- kiểm tra và sửa tính không nhất quán

# Làm sạch dữ liệu

**Bước 1** : Điền đầy các giá trị bị mất

# Làm sạch dữ liệu

## Điền đầy các giá trị bị mất

Sau đây liệt kê 6 phương pháp điền giá trị dữ liệu

1. Bỏ không xét đến bộ dữ liệu bị mất giá trị
2. Điền lại giá trị bằng tay
3. Gán cho giá trị nhãn đặc biệt hay ngoài khoảng biểu diễn
4. Gán giá trị trung bình cho nó
5. Gán giá trị trung bình của các mẫu khác thuộc cùng lớp đó
6. Tìm giá trị có xác suất lớn nhất điền vào chỗ bị mất (hồi quy, suy diễn Bayes, cây quyết định qui nạp)

# Làm sạch dữ liệu

**Bước 2** : Nhiễu là lỗi ngẫu nhiên tạo nên phương sai giá trị cần đo. Các phương pháp chống nhiễu phổ biến gồm

1. Hồi quy (Regression) : sẽ dành chương riêng
2. Phân cụm (Cluster) : sẽ dánh chương riêng
3. Giá trị lô (Binning method) : xem chi tiết slice sau

# Làm sạch dữ liệu

## Giá trị lô (Binning method)

Để chuốt dãy giá trị đã được sắp xếp. Ta nhóm chúng theo lô, mỗi lô là một tập cố định các số trong dãy. Tìm giá trị đặc trưng lô

- giá trị trung bình : gán lại cho tất cả các giá trị
- giá trị trung vị : gán lại cho tất cả các giá trị
- giá trị biên : có hai biên trên-dưới, giá trị mới sẽ được gán cho biên gần nhất

# Làm sạch dữ liệu

Cho dãy giá trị được sắp xếp theo thứ tự
{4, 8, 15, 21, 21, 24, 25, 28, 34}

Chia thành 3 lô

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

Gán giá trị trung bình

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

Giá giá trị biên

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

# Làm sạch dữ liệu

**Bước 3 :** kiểm tra và sửa tính không nhất quán trong dữ liệu. Trước hết ta cần phát hiện sự không nhất quán dữ liệu, nguyên nhân xuất hiện không nhất quan dữ liệu có thể ...

- thiết kế biểu mẫu nhập liệu, có quá nhiều phần phải điền
- lỗi con người khi nhập liệu, e.g. ngày "2014/31/2"
- lỗi cố ý, e.g. những thông tin nhạy cảm người dùng
- lỗi cập nhật, e.g. dùng địa chỉ cũ, số điện thoại cũ
- lỗi thiết bị ghi, lỗi hệ thống e.g. mất điện, ổ cứng hỏng
- dữ liệu có dạng không tương thích yêu cầu, e.g. giá trị gây tràn số

# Làm sạch dữ liệu

## Phát hiện sự bất thường

Đề phát hiện sự bất thường trong giá trị dữ liệu, ta tuân thủ ba luật sau

- Luật giá trị duy nhất (unique rule) : mỗi giá trị của một thuộc tính sẽ khác biệt với các giá trị khác thuộc cùng thuộc tính
- Luật liên tục (consecutive rule) : không có giá trị bị mất giữa giá trị lớn nhất và nhỏ nhất tương ứng một thuộc tính
- Luật giá trị rỗng (null rule) : xác định trước các ký hiệu hay cách đánh dấu giá trị rỗng, e.g. "?" hay "don't know"

# Làm sạch dữ liệu

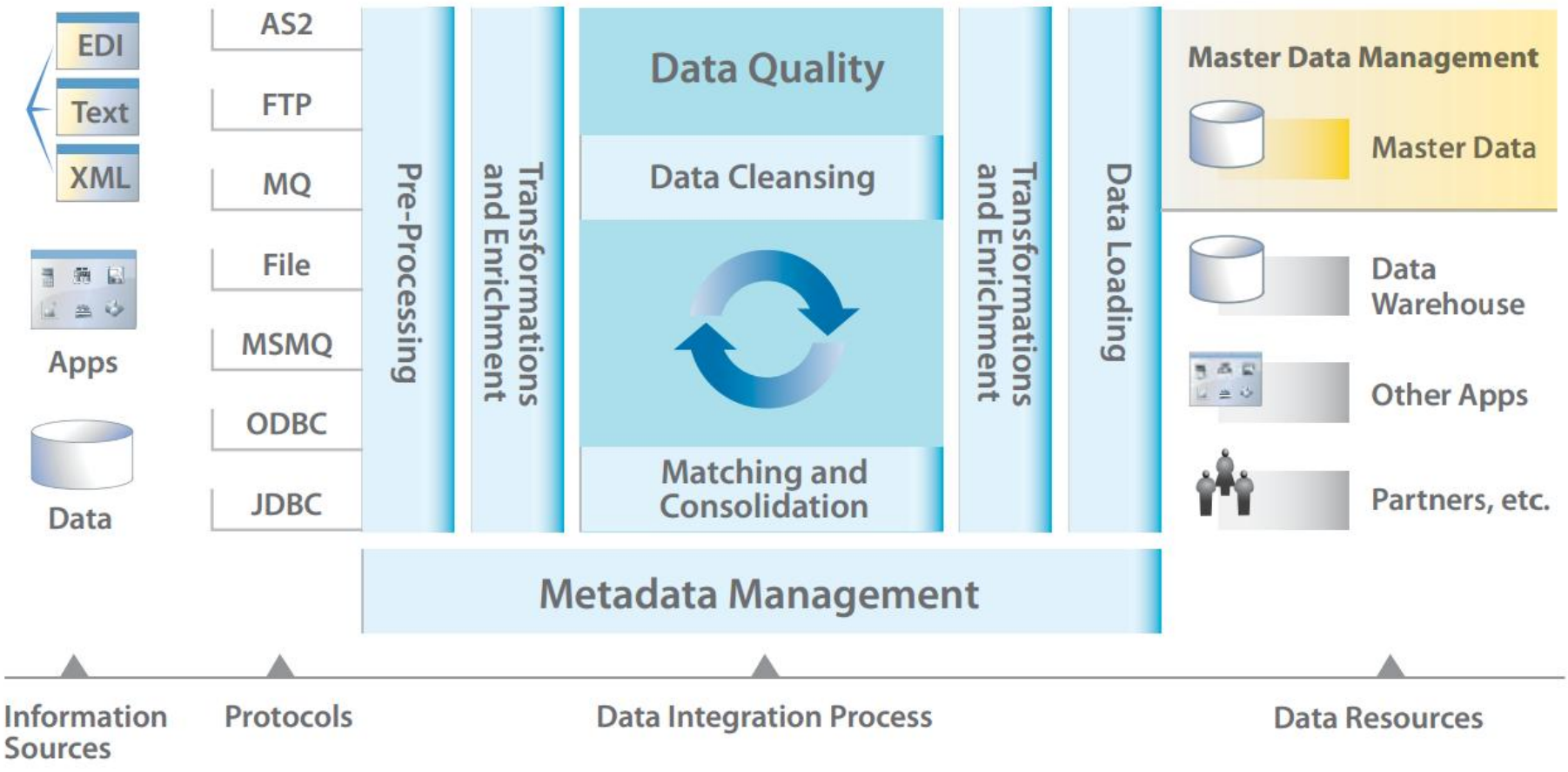
## Các công cụ

Dùng để phát hiện sự không nhất quán dữ liệu

- Công cụ chà dữ liệu (Data scrubbing tools) : dùng cho một lĩnh vực cụ thể
- Công cụ kiểm toán dữ liệu (Data auditing tools) : dùng cho việc phân tích dữ liệu, xác định quan hệ, xác định các luật. Sau đó tự động sữa chữa các giá trị không tương thích

# Dành cho trả lời câu hỏi

# Tích hợp và chuyển đổi dữ liệu



# Tích hợp và chuyển đổi dữ liệu

## Ý nghĩa của chuyển đổi dữ liệu

Dữ liệu biến đổi hoặc cũng cố dưới dạng chuẩn để tiến hành khai phá. Ta có các bước sau để chuyển đổi dữ liệu

- Tập hợp : các giá trị dữ liệu tạo thành bộ hay khối, e.g. giá trị dữ liệu theo ngày, tháng, năm
- Tổng quát hóa (generalization) : các dữ liệu "thô" được thay bằng các khái niệm đã chuẩn hóa, e.g. mức thu nhập dưới nghèo, trung lưu hay giầu có
- Chuẩn hóa (normalization) : nếu phạm vi dữ liệu lớn thì đưa nó về phạm vi chuẩn, e.g. **min-max normalization**
- Xây dựng thuộc tính (attribute construction) : thuộc tính mới thêm vào giúp quá trinh khai phá dữ liệu

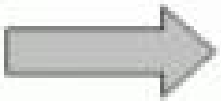
# Dành cho trả lời câu hỏi

# Giảm chiều

## Ý nghĩa của giảm kích thước dữ liệu

Việc giảm kích thước của dữ liệu cần đồng thoải giữ được tính phân tích dữ liệu

| m/z | Intensity |
|---|---|
| 31 | 1200 |
| 32 | 1250 |
| 33 | 1300 |
| 34 | 1350 |
| 35 | 200 |
| 36 | 1225 |
| 37 | 12300 |
| 38 | 100 |
| 39 | 40 |
| 40 | 700 |
| 41 | 500 |
| 42 | 35000 |
| 43 | 300 |
| 44 | 4000 |
| 45 | 550 |



| m/z | Intensity |
|---|---|
| 33 | 1300 |
| 34 | 1350 |
| 37 | 12300 |
| 42 | 35000 |
| 44 | 4000 |

# Giảm chiều

Các chiến lược để giảm kích thước dữ liệu

- **Lựa chọn tập con các thuộc tính** : trong đó các thuộc tính không liên quan, dư thừa hoặc các chiều cũng có thể xóa hay loại bỏ
- **Giảm chiều** : trong đó cơ chế mã hóa được sử dụng để giảm kích cỡ tập dữ liệu
- **Rời rạc hóa và trừu tượng khái niệm** : trong đó các giá trị dữ liệu thô được thay thế bằng các khái niệm trừu tượng đã rời rạc hóa.

# Giảm chiều

## Lựa chọn tập con các thuộc tính

Mục đích của lựa chọn tập con các thuộc tính là cực tiểu hóa tập các thuộc tính sao cho kết quả phân bố xác suất của lớp dữ liệu là gần phân bố gốc gồm tất cả các thuộc tính.

## Vấn đề gặp phải với bài toán tìm tập con

Giả sử ta có $n$ thuộc tính, dẫn đến $2^n$ các tập con thuộc tính. Một giải thuật tìm kiếm vét cạn sẽ có thời gian lâu, ta cần dùng một *giải thuật heuristic* có thời gian chạy chấp nhận được.

# Giảm chiều

| Lựa chọn tiến | Khử ngược |
|---|---|
| Tập thuộc tính ban đầu | Tập thuộc tính ban đầu |
| $\{A_1, A_2, A_3, A_4.A_5, A_6\}$ | $\{A_1, A_2, A_3, A_4.A_5, A_6\}$ |
| Tập thuộc tính khởi tạo | Tập thuộc tính khởi tạo |
| $\{\}$ | $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ |
| $\Rightarrow \{A_1\}$ | $\Rightarrow \{A_1, A_4, A_5, A_6\}$ |
| $\Rightarrow \{A_1, A_4\}$ | $\Rightarrow \{A_1, A_4, A_6\}$ |
| $\Rightarrow \{A_1, A_4, A_6\}$ | |

- Lựa chọn tiến : khởi tạo tập rỗng ban đầu, ta lựa chọn thuộc tính "tốt nhất" vào từ tập còn lại. Tiếp tục đến khi có tập con được lựa chọn
- Khử ngược : khởi tạo tập đầy đủ ban đầu, ta lựa chọn thuộc tính "tồi nhất" và loại ra khỏi tập khởi tạo. Tiếp tục đến khi có tập con được lựa chọn

# Giảm chiều

## Thông tin thêm (Information gain)

Bài toán phân loại, e.g. hai lớp $c_1$ và $c_2$

$$IG = \underbrace{\{entropy(c_1) + entropy(c_2)\}}_{\text{các thuộc tính}} - \underbrace{\{p(c_1|A)entropy(c_1) + p(c_2|A)entropy(c_2)\}}_{\text{loại bỏ/thêm vào thuộc tính A}}$$

trong đó $entropy(c_i) = -p(c_i)log_2 p(c_i)$ với $i = 1, 2$

- Lựa chọn tiến : Mỗi bước ta chọn thêm thuộc tính A sao cho *IG nhỏ nhất* có thể
- Khử ngược : Mỗi bước ta chọn loại bỏ thuộc tính A sao cho *IG lớn nhất* có thể

# Giảm chiều

## Ý nghĩa của giảm chiều

Dữ liệu sẽ được mã hóa hay chuyển đổi sang một cách biểu diễn thu gọn hay "nén" so với dữ liệu gốc trước đó. Hai cách làm giảm chiều được đề cập trong phần này

- Biến đổi wavelet (Wavelet Transforms)
- Phân tích thành phần chính (Principal Components Analysis)

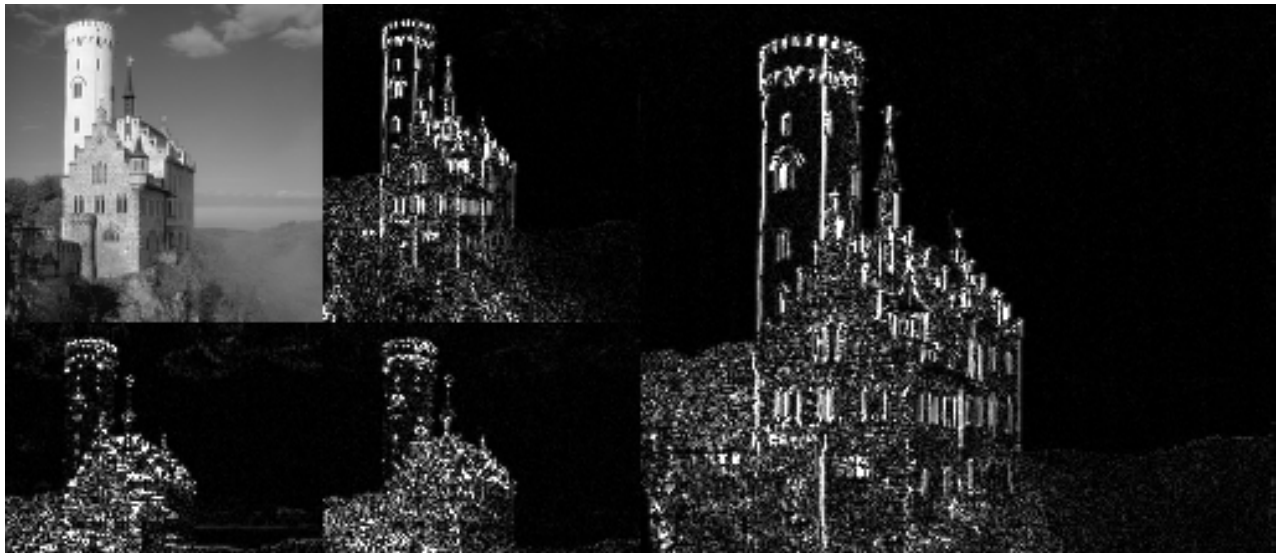
# Giảm chiều

## Biến đổi Wavelet rời rạc (Discrete Wavelet Transform - DWT)

Là một kỹ thuật biến đổi tuyến tính dữ liệu cho phép giữ lại các hệ số quan trọng trong phép biến đôi. Luôn có hai chiều

- Xuôi : từ dữ liệu gốc sang dữ liệu thu gọn
- Ngược : từ dữ liệu thu gọn chuyển ngược dữ liệu gốc

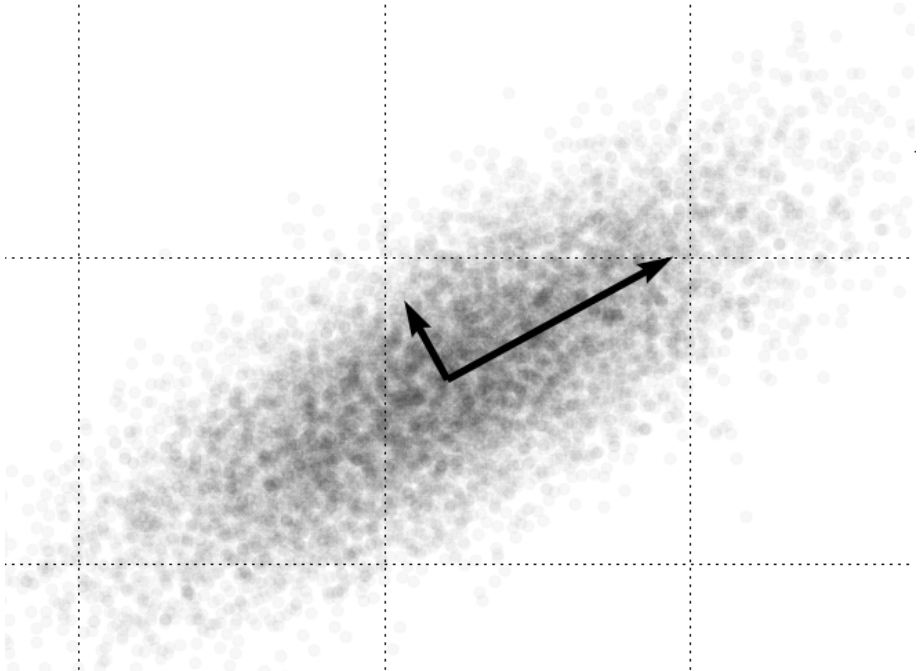
Ví dụ: Haar-2, Daubechies-4 và Daubechies-6



# Giảm chiều

## Phân tích thành phần chính (Principal Components Analysis - PCA)

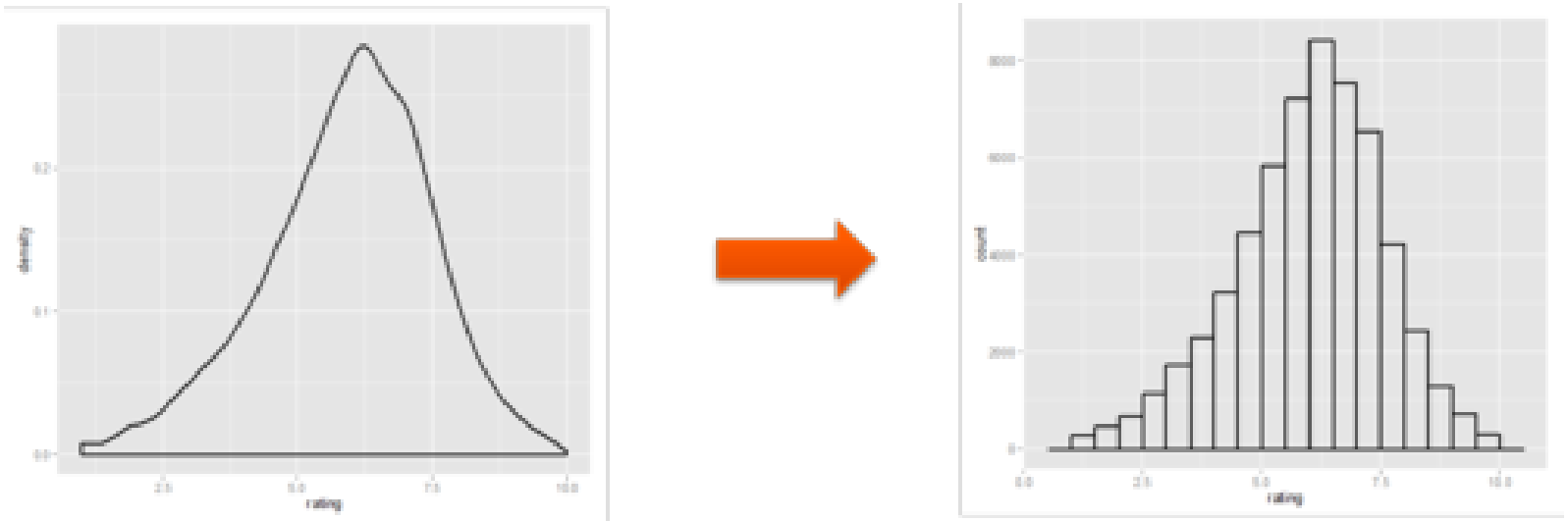
Phép phân tích ánh xạ tập dữ liệu $n$ chiều sang một không gian vuông góc trong đó các chiều chỉ đến hướng dữ liệu từ dao động nhất đến ít dao động nhất.



# Giảm chiều

## Ý nghĩa của rời rạc hóa

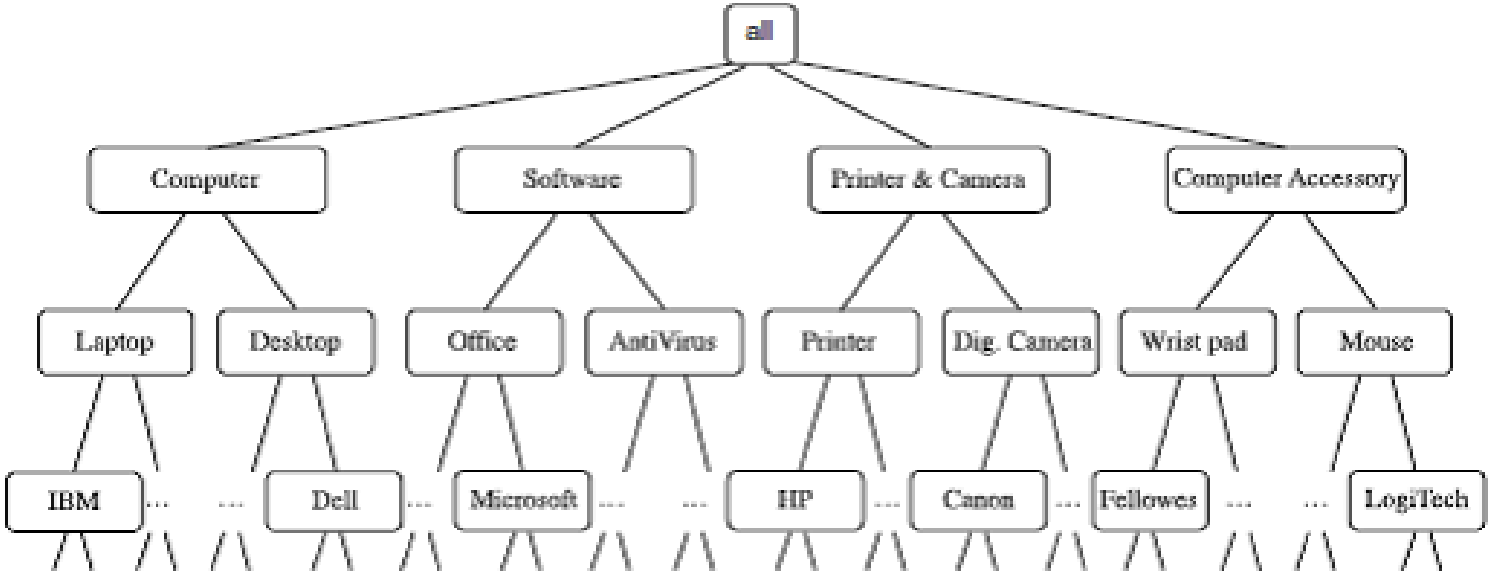
Rời rạc hóa thuộc tính giá trị liên tục thành các khoảng, nỗi khoảng được gán một nhãn rời rạc. Những nhãn thay thế này giảm kích thước khi so với biểu diễn giá trị liên tục.



# Giảm chiều

## Ý nghĩa của trừu tượng khái niệm

Các nhãn sau khi tập hợp lại sau đó được trừu tượng hóa thông qua một cây biểu diễn các khái niệm



# Chương trình Weka

## Yêu cầu

- Tải tập dữ liệu glass.arff
- Chạy tiền xử lý (preprocess)
- Xem và giải thích ý nghĩa từng cửa sổ
- Xem các giá trị tiền xử lý từng cửa sổ