



# Chương 1

## KHÁI NIỆM CHUNG VỀ KHO DỮ LIỆU VÀ KHAI PHÁ DỮ LIỆU

1. Khái niệm về kho dữ liệu.
2. Khái niệm về khai phá dữ liệu.
3. Các loại dữ liệu và kiểu mẫu dùng để khai phá.
4. Các bài toán và phương pháp cơ bản trong khai phá dữ liệu.
5. Sự tích hợp của khai phá dữ liệu với một cơ sở dữ liệu hoặc với kho dữ liệu.
6. Ứng dụng của kho dữ liệu và khai phá dữ liệu.



# Khái niệm về kho dữ liệu

- Kho dữ liệu (**Data warehouse**) là kho lưu trữ dữ liệu lưu trữ bằng thiết bị điện tử của một tổ chức,
- Các kho dữ liệu được thiết kế để hỗ trợ việc phân tích dữ liệu và lập báo cáo.
- Kho dữ liệu có những đặc điểm:
  - ✓ Tích hợp (**Atomicity**): Từ nhiều nguồn khác nhau,
  - ✓ Theo chủ đề (**Consistency**): Có ích để khai thác,
  - ✓ Biến thời gian (**Isolation**): Dữ liệu không bị ảnh hưởng hoặc tác động lẫn nhau khi được truy suất,
  - ✓ Cố định (**Durable**): khi đã hoàn chỉnh thì không đổi.



# Khái niệm về kho dữ liệu

- ❖ Kho dữ liệu dung cho mục đích riêng biệt, lĩnh vực hẹp gọi là **Data Mart**.
- ❖ Một **Data warehouse** có thể hình thành nhiều **Data Mart**.
- ❖ Thuật ngữ **Data Warehousing**: Quá trình xây dựng và sử dụng một kho dữ liệu.

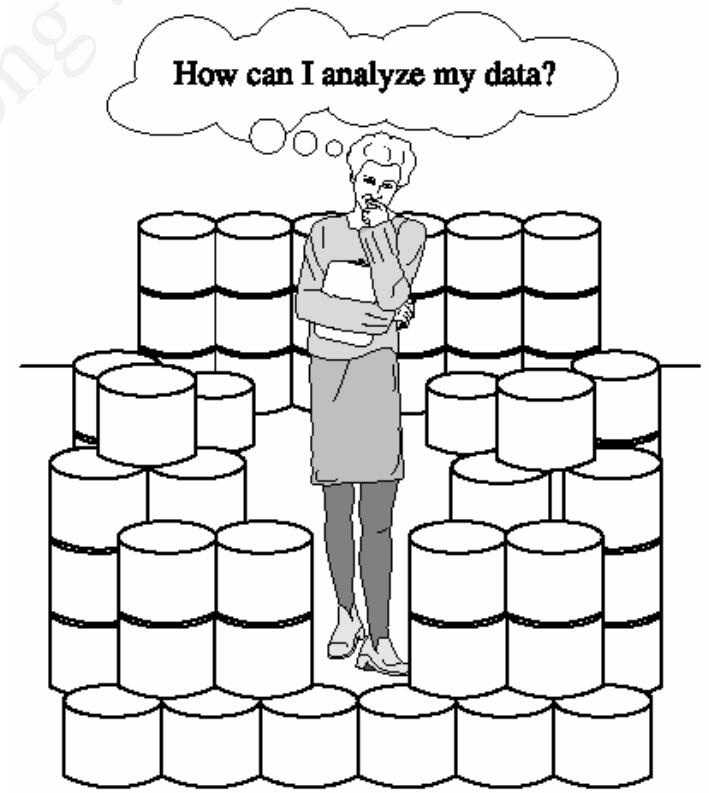


# Khái niệm về kho dữ liệu

- ❖ Công cụ **ETL** (**E**xtract – **T**ransform – **L**oad):
  - ✓ Rút trích (**E**xtract):
    - Rút trích thông tin từ những nguồn đã có,
    - Những phiên bản phụ thuộc thời gian của dữ liệu,
    - Chọn lựa dữ liệu.
  - ✓ Chuyển đổi (**T**ransform):
    - Chuyển đổi các định dạng khác nhau về định dạng cho trước.
  - ✓ Tải (**L**oad)
    - Sắp xếp, hợp nhất, lập chỉ mục, ... và phân hoạch.

# Khái niệm về khai phá dữ liệu

- Các các nhân, tổ chức **ngập trong dữ liệu** nhưng **đói thông tin**.



- Giải pháp: **Kho dữ liệu** và **Khai phá dữ liệu**



# Khái niệm về khai phá dữ liệu

- Khai phá dữ liệu (**Data mining**) là quá trình phát hiện và trích xuất tri thức từ lượng dữ liệu lớn,
- Lượng dữ liệu lớn dùng cho khai phá gồm:
  - ✓ Có cấu trúc,
  - ✓ Bán cấu trúc,
  - ✓ Phi cấu trúc,
  - ✓ Được lưu trữ tạm thời hay ổn định.
- Các thuật ngữ: knowledge discovery/mining in data/knowledge extraction/data archeology, ...



# Khái niệm về khai phá dữ liệu

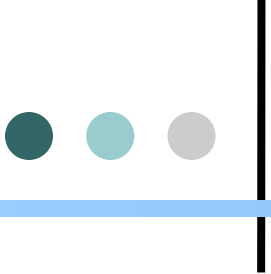
- Tri thức đạt được từ quá trình khai phá:
  - ✓ Mô hình phân loại và dự đoán,
  - ✓ Mô hình gom cụm,
  - ✓ Mẫu thường xuyên, các mối qua hệ, tương quan,
  - ✓ Mô tả lớp/khái niệm,
  - ✓ Có cấu trúc, bán cấu trúc hoặc phi cấu trúc,
  - ✓ Có thể dùng trong điều khiển quy trình, ra quyết định, ...
  - ✓ ...





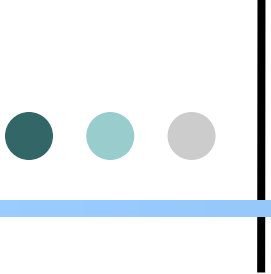
# Khái niệm về khai phá dữ liệu

- Ý nghĩa và vai trò:
  - ✓ Ứng dụng được trong mọi lĩnh vực có dữ liệu,
  - ✓ Hỗ trợ nhiều đối tượng khác nhau:
    - Doanh nghiệp,
    - Khách hàng,
    - Nhà khoa học,
    - Giáo dục học, ...



# Các loại dữ liệu và kiểu mẫu dùng để khai phá

- ❖ Dữ liệu **hướng chủ thể**:
  - ✓ Dữ liệu hướng theo từng nhóm đối tượng: khách hàng, bệnh nhân, sản phẩm, ...
  - ✓ Tập trung vào việc mô hình hóa và phân tích các dữ liệu cho các nhà sản xuất quyết định
  - ✓ Chuyển từ hướng ứng dụng sang hướng hỗ trợ quyết định.
  - ✓ Không dùng cho các hoạt động hàng ngày hoặc xử lý giao dịch.



# Các loại dữ liệu và kiểu mẫu dùng để khai phá

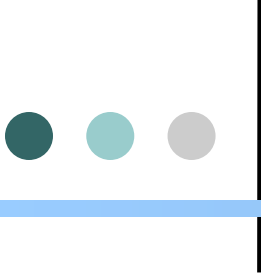
## ❖ Tính tích hợp:

- ✓ Dữ liệu được tập hợp từ nhiều nguồn: có thể khác kiểu, khác cấu trúc, ...
- ✓ Các nguồn: cơ sở dữ liệu quan hệ, tập tin có cấu trúc, tập tin phẳng, ...
- ✓ Cần được chuẩn hóa để đảm bảo tính nhất quán trong quy ước đặt tên, ...
- ✓ Việc chuẩn hóa cần thực hiện trước khi tích hợp.



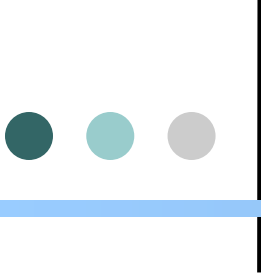
# Các loại dữ liệu và kiểu mẫu dùng để khai phá

- ❖ Dữ liệu **biến thời gian**.
  - ✓ Thông tin về quá khứ, hiện tại,
  - ✓ So sánh dữ liệu theo chiều thời gian,
  - ✓ Hỗ trợ quyết định cho tương lai.
  - ✓ Thành phần thời gian có thể tương minh hoặc ngầm định.
- ❖ Dữ liệu mang tính **bền vững, chỉ đọc** (non volatile):
  - ✓ Có thể thêm vào, nhưng không thay thế,
  - ✓ Phục vụ việc nghiên cứu, phân tích



# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

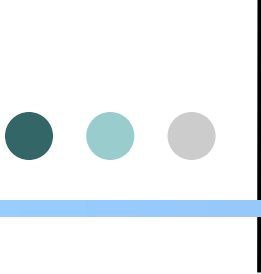
- ❖ Khai phá dữ liệu nhằm mục đích:
  - ✓ Mô tả được một số khía cạnh của tập dữ liệu lớn,
  - ✓ Dự báo về những giá trị chưa biết hoặc sẽ có của các biến.



# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

❖ Một số bài toán cơ bản:

1. Mô tả khái niệm,
2. Quan hệ kết hợp,
3. Gom cụm,
4. Phân lớp,
5. Hồi quy,
6. Mô hình phụ thuộc,
7. Phát hiện thay đổi và độ lệch.



# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

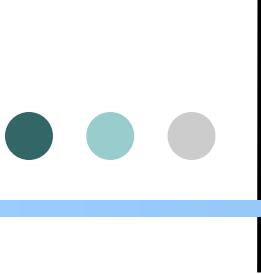
## 1. Bài toán mô tả khái niệm:

- ✓ Tìm ra các đặc trưng và tính chất của khái niệm,
- ✓ Tổng quát hóa, tóm tắt, ... để tìm ra các đặc trưng của dữ liệu.

# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

2. Bài toán tìm quan hệ kết hợp (Association Rule):
- ✓ Phát hiện mối quan hệ kết hợp giữa các tập thuộc tính trong kho dữ liệu.
  - ✓ Bài toán **khai phá luật kết hợp** là một bài toán tiêu biểu
  - ✓ Ví dụ:
    - {**Tóc đen**, **Da vàng**} → {người Châu á},
    - {**Mật ong**, **Đường**} → {Ngọt}





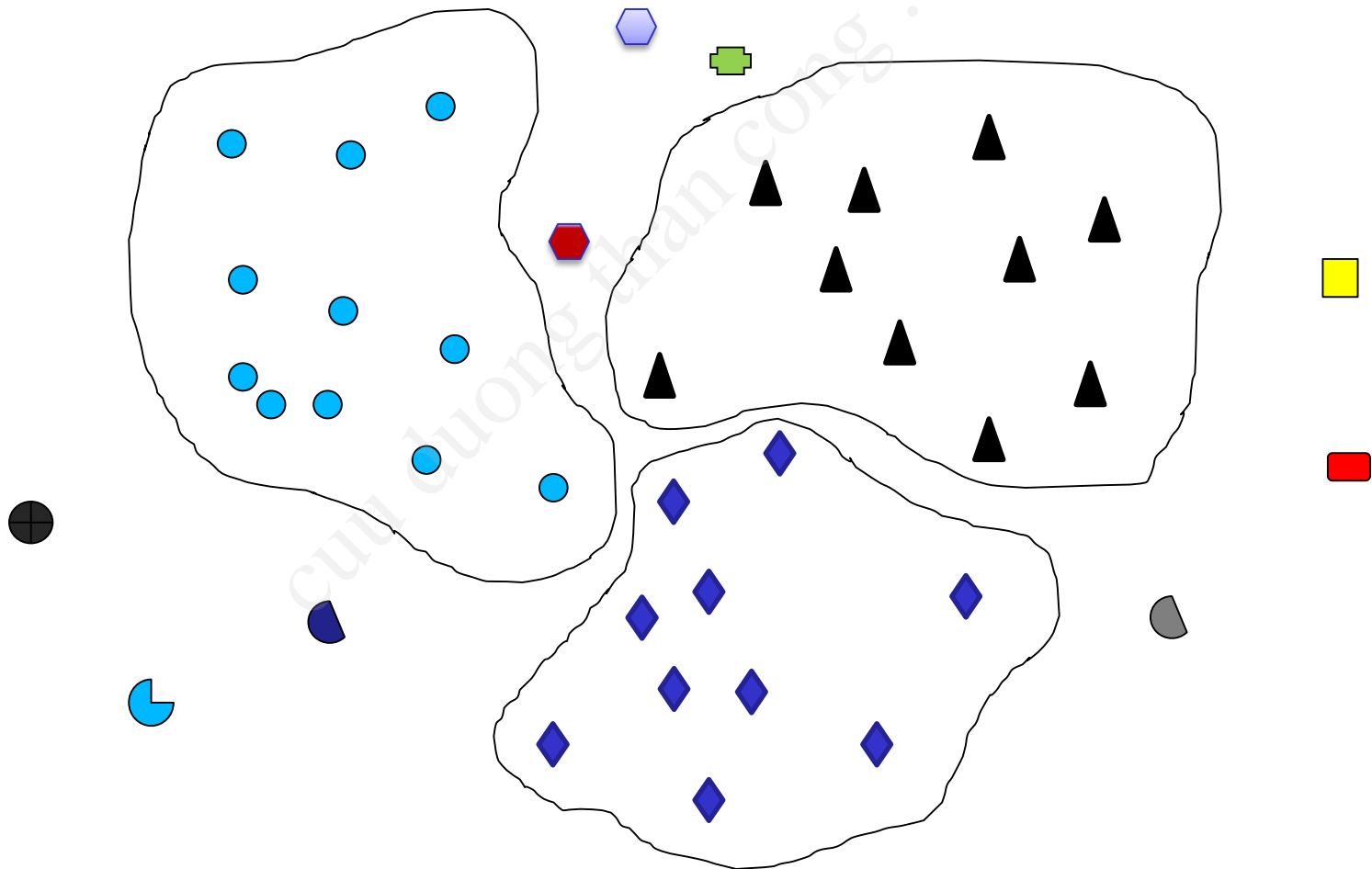
# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

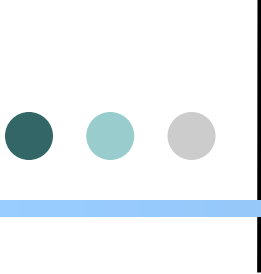
## 3. Bài toán gom cụm dữ liệu (**clustering**):

- ✓ Gom các dữ liệu có **độ tương đồng cao** thành các “cụm” để có thể phát hiện được đặc trưng của các thuộc tính trong miền ứng dụng.
- ✓ Mục tiêu: cực đại hóa tính tương đồng giữa các phần tử trong cùng cụm, và cực tiểu hóa tính tương đồng giữa các phần tử khác cụm.
- ✓ Phân cụm còn được gọi là bài toán “học máy không có giám sát” (**unsupervised learning**).

# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

## 3. Bài toán gom cụm dữ liệu (clustering):



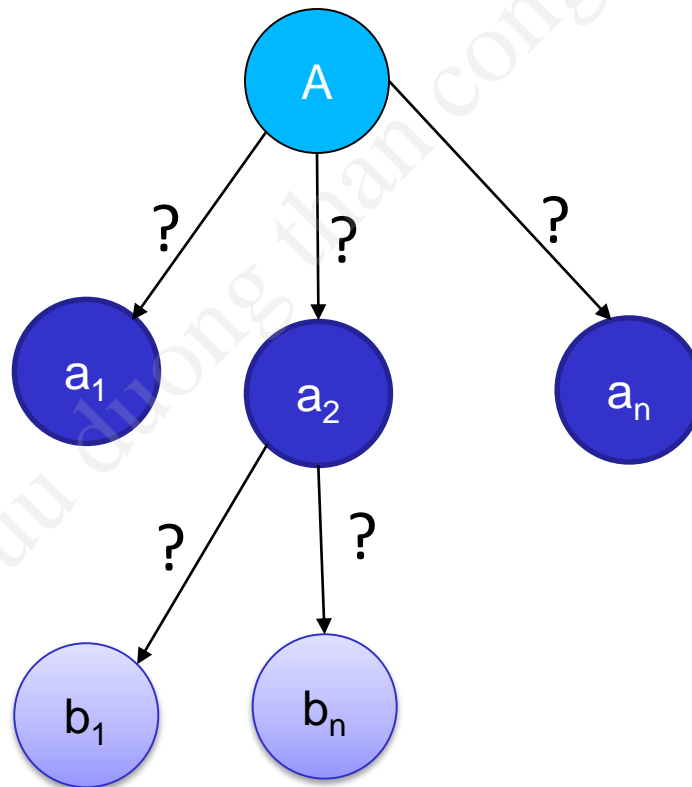


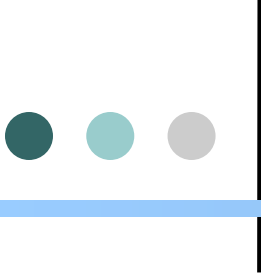
# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

4. Bài toán phân lớp (**classification**):
- ✓ Xây dựng (mô tả) các mô hình (hàm) nhằm đặc tả, phát hiện đặc trưng các lớp hoặc khái niệm để dự báo cho các dữ liệu tiếp theo.
  - ✓ Số lớp (nhóm) được xác định trước.
  - ✓ Một số phương pháp: **cây quyết định, mạng Bayes, mạng neuron,...**
  - ✓ Phân lớp thuộc nhóm bài toán “học máy có giám sát” (**supervised learning**).

# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

## 4. Bài toán phân lớp (classification):

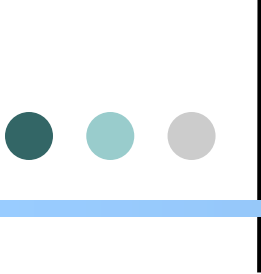




# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

## 5. Bài toán hồi quy:

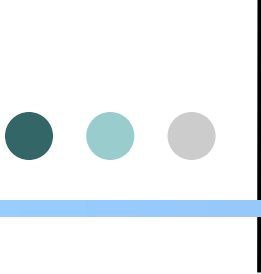
- ✓ Diễn hình trong phân tích thống kê và dự báo.
- ✓ Dự đoán các giá trị của một hoặc một số biến phụ thuộc vào giá trị của một tập hợp các biến độc lập.
- ✓ Có thể quy về việc học một hàm ánh xạ dữ liệu nhằm xác định giá trị thực của một biến theo một số biến khác.



# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

## 6. Bài toán tìm mô hình phụ thuộc:

- ✓ Tìm ra một mô hình mô tả sự phụ thuộc có ý nghĩa giữa các biến.
- ✓ Bao gồm 2 mức:
  - **Mức cấu trúc của mô hình:** thường biểu diễn dạng đồ thị để phát hiện sự phụ thuộc bộ giữa các biến.
  - **Mức định lượng của mô hình:** Phát hiện độ mạnh của tính phụ thuộc dựa trên trọng số của các thuộc tính.



# Các bài toán và phương pháp cơ bản trong khai phá dữ liệu

7. Bài toán phát hiện thay đổi và độ lệch:
- ✓ Tập trung phát hiện sự thay đổi có ý nghĩa dưới dạng độ đo đã biết trước hoặc giá trị chuẩn,
  - ✓ Cung cấp những tri thức về sự biến đổi và độ lệch cho người dùng.
  - ✓ Thường được ứng dụng trong bước tiền xử lý.

# Sự tích hợp của khai phá dữ liệu với một cơ sở dữ liệu hoặc với kho dữ liệu

## 1. Tích hợp dữ liệu:

- ✓ Cần có một lượng dữ liệu đủ lớn để phân tích và khai phá.
- ✓ Dữ liệu có thể thu thập từ nhiều nguồn: không thống nhất,
- ✓ Dữ liệu từ các nguồn khác nhau có thể là:
  - Có cấu trúc: cơ sở dữ liệu quan hệ, ...
  - Phi cấu trúc: Tập tin phẳng (flat file),
  - Được lưu trữ tạm thời hoặc ổn định, ...



# Sự tích hợp của khai phá dữ liệu với một cơ sở dữ liệu hoặc với kho dữ liệu

## 1. Tích hợp dữ liệu:

- ✓ Hợp nhất các nguồn có thể dẫn đến:
  - ✓ Cùng một thuộc tính nhưng có thể không tương đương nhau về ý nghĩa,
  - ✓ Không tương đồng về mặt giá trị,
  - ✓ Dư thừa dữ liệu,
  - ✓ ...

# Sự tích hợp của khai phá dữ liệu với một cơ sở dữ liệu hoặc với kho dữ liệu

2. Biến đổi dữ liệu: Tạo tính tương thích giữa dữ liệu của nhiều nguồn khác nhau.
  - ✓ **Làm mịn:** loại bỏ trường hợp nhiễu.
  - ✓ **Tổng hợp:** Rút gọn dữ liệu và tạo khối dữ liệu cho việc phân tích.
  - ✓ **Khái quát hóa:** Chuyển dữ liệu mức thấp sang mức cao.
  - ✓ **Chuẩn hóa:** Chuyển khoảng giá trị rộng thành khoảng giá trị nhỏ hơn ( $[10..1.000] \rightarrow [0.0..1.0]$ )
  - ✓ **Xác định thêm thuộc tính.**

# Sự tích hợp của khai phá dữ liệu với một cơ sở dữ liệu hoặc với kho dữ liệu

## 2. Biến đổi dữ liệu:

❖ Một số phương pháp biến đổi:

✓ Min-Max:

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new\_max}_A - \text{new\_min}_B) + \text{new\_min}_A$$

- $\min_A, \max_A$ : giá trị lớn nhất và nhỏ nhất của thuộc tính A
- $\text{New\_min}_A, \text{new\_max}_A$ : miền giá trị mới.

# Sự tích hợp của khai phá dữ liệu với một cơ sở dữ liệu hoặc với kho dữ liệu

## 2. Biến đổi dữ liệu:

### ❖ Một số phương pháp biến đổi:

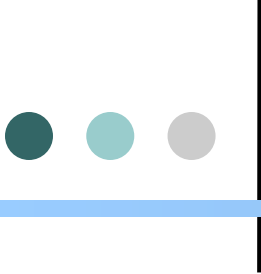
#### ✓ Z-score:

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

- $\bar{A}$ : giá trị trung bình của thuộc tính A,
- $\sigma_A$ : độ lệch chuẩn.

#### ✓ Thay đổi tỷ lệ.

#### ✓ Lựa chọn tập thuộc tính con



# Ứng dụng của kho dữ liệu và khai phá dữ liệu

- ❖ Kinh doanh (**Business**),
- ❖ Tài chính (**finance**),
- ❖ Tiếp thị (**sales marketing**),
- ❖ Thương mại (**commerce**),
- ❖ Bảo hiểm (**insurance**),
- ❖ Khoa học (**science**),
- ❖ Điều khiển (**control**),
- ❖ ...

- ❖ Câu 1.2 ([1] – trang 34)
- ❖ Câu 1.5 ([1] – trang 35): Giải thích sự khác biệt và tương đồng giữa các **phân biệt** và **phân loại**, giữa tả và **gom cụm**, và giữa các **phân lớp** và **hỏi quy?**
- ❖ Câu 1.8 ([1] – trang 35).
- ❖ Câu 1.9 ([1] – trang 35).