KHAI PHÁ DỮ LIỆU
(DATAMINING)

TỔNG QUAN

Mai Xuân Hùng

Khai phá dữ liệu

- Có sẵn khối dữ liệu lớn:
  - Các CSDL khổng lồ
  - Dữ liệu từ Internet

Khai phá dữ liệu là gì?

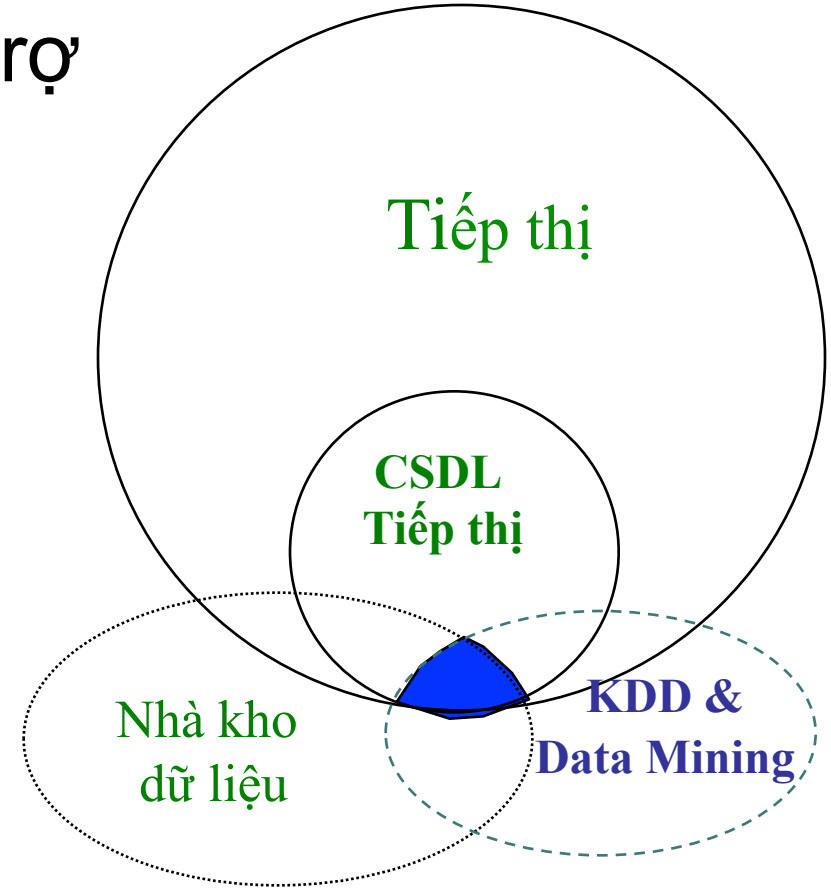
- Rút trích thông tin hữu ích, chưa biết, tiềm ẩn trong khối dữ liệu lớn
- Phân tích dữ liệu bán tự động
- Giải thích dữ liệu trên các tập dữ liệu lớn

Khai phá dữ liệu là gì?

- Thuật ngữ:
  - Khai phá dữ liệu - Data mining
    - KPDL là một bước của tiến trình KDD
  - Knowledge discovery in databases (KDD)
    - Thuật ngữ tổng quát gồm các bước:
      - Tiền xử lý
      - KPDL
      - Hậu xử lý

Khai phá dữ liệu có ích lợi gì ?

- Cung cấp tri thức hỗ trợ ra quyết định
- Dự báo
- Khái quát dữ liệu



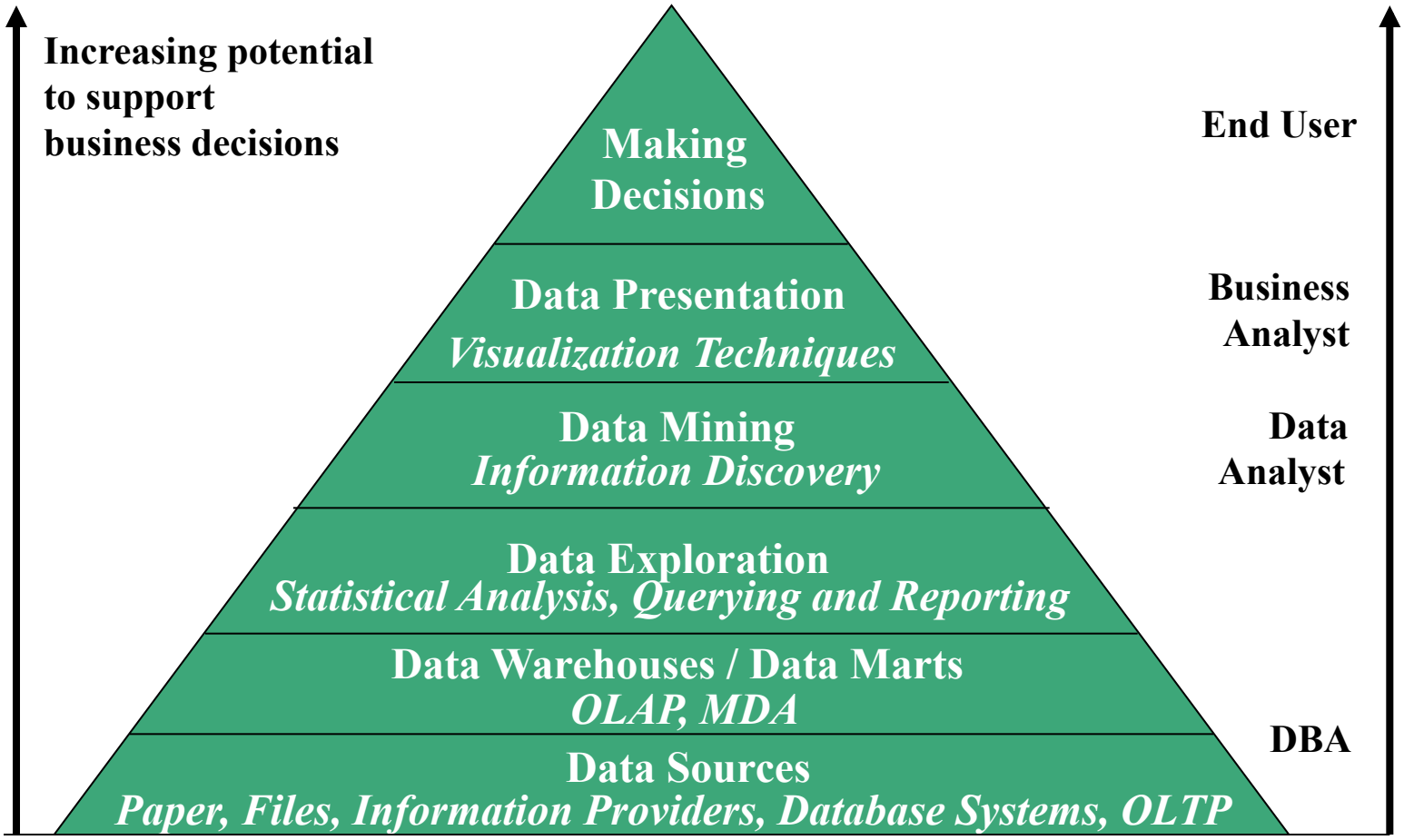
Các ứng dụng tiềm năng

- **Phân tích dữ liệu, hỗ trợ ra quyết định**
  - Phân tích và quản lý thị trường
  - Quản lý và phân tích rủi ro
  - Quản lý và phân tích các sai hỏng
- **Các ứng dụng khác:**
  - Khai thác Web
  - Khai thác văn bản (text mining)
  - etc.

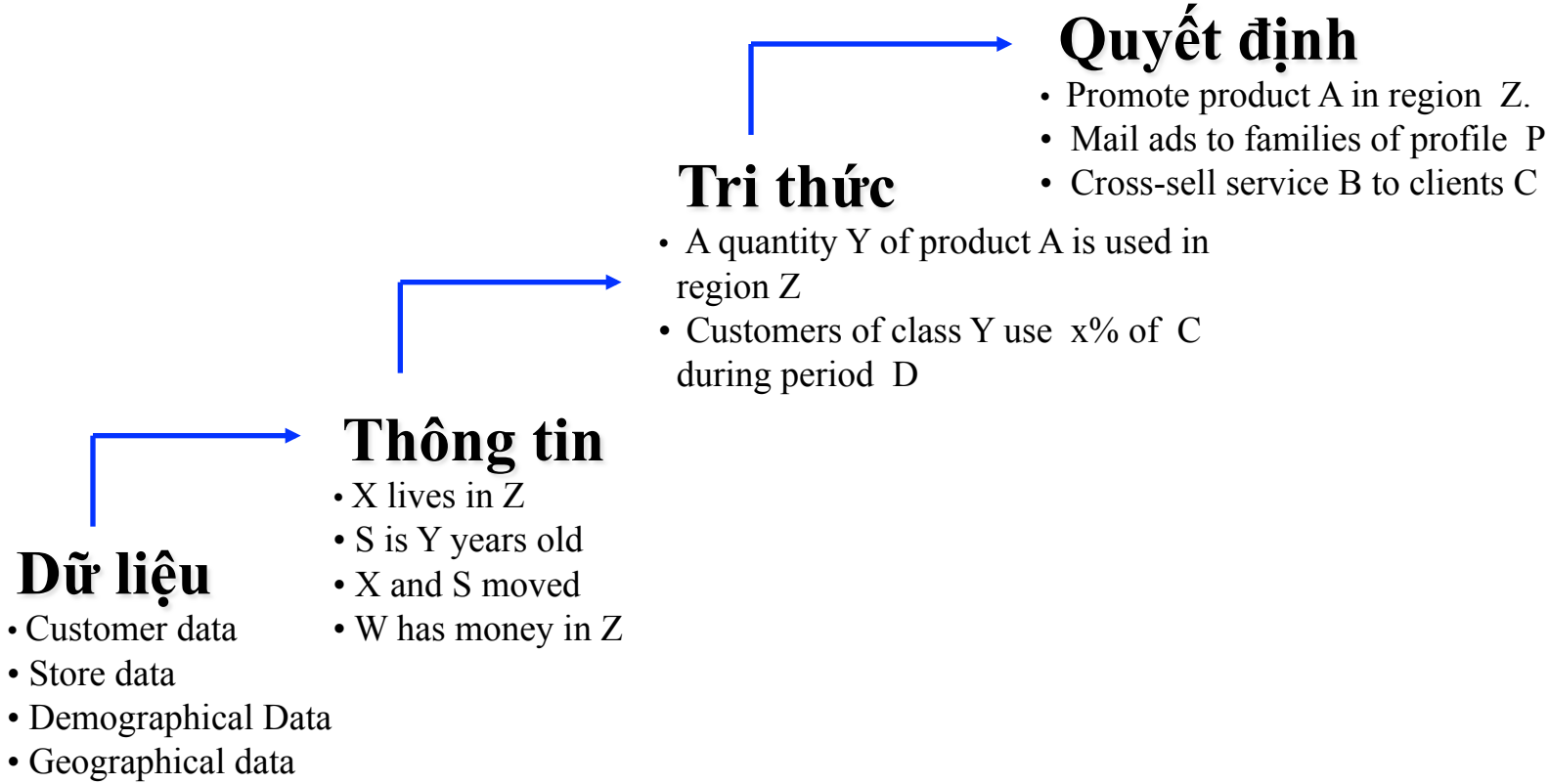
Tiến trình khai phá dữ liệu(1)

Tiến trình khai phá dữ liệu(2)

Khai phá dữ liệu



Từ dữ liệu đến quyết định



Giải thích

- Dữ liệu – thông tin – tri thức
  - **Dữ liệu**: Là sự diễn dịch những trường đơn lẽ ví dụ: Nguyễn Thị Hoa Mai, Sinh viên, ngành CNTT, môn CSDL.
  - **Thông tin**: Là mối liên hệ các thành phần của dữ liệu, Ví dụ: Nguyễn Thị Hoa Mai là sinh viên ngành công nghệ thông tin. Ngành công nghệ thông tin có môn CSDL.

Dạng luật kết hợp

- **Tri thức**: Là mối liên hệ của các thành phần thông tin, có hai cấp độ.
  - Chỉ giới hạn một nhóm nhỏ thông tin. **Ví dụ**: Nguyễn Thị Hoa Mai là sinh viên ngành công nghệ thông tin nên phải học môn CSDL.
  - Là những thông tin mang tính quy luật phổ biến. Ví dụ: Nếu X là sinh viên ngành CNTT thì X phải học môn CSDL.

ví dụ

- Dữ liệu khổng lồ từ: Internet, từ nhiều lĩnh vực trong đời sống xã hội, quản lý kinh tế, khoa học kỹ thuật …Ví dụ: CSDL dân cư Thành Phố HCM có hơn 50 triệu dân khẩu, CSDL tuyển sinh đại học hơn 1 triệu
- Từ khối dữ liệu này =>rút trích những thông tin hữu ích, chưa biết tiềm ẩn trong khối dữ liệu hỗ trợ tiến trình ra quyết định, dự báo, các nhà nghiên cứu đã phát triển các phương pháp, kỹ thuật và phần mềm mới hỗ trợ tiến trình khám phá, phân tích tổng hợp thông tin.

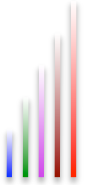
Ví dụ

- Khai thác thông tin truyền thống : 80 % thông tin từ CSDL, còn lại 20% thông tin nhưng chứa đựng thông tin quan trọng.
- Khai thác dữ liệu-Data Mining (KTDL) là tiến trình khám phá tri thức tiềm ẩn trong các CSDL. Cụ thể hơn, đó là tiến trình trích lọc, sản sinh những tri thức hoặc các mẫu tiềm ẩn, chưa biết nhưng hữu ích từ các CSDL lớn.

Hình thức KTDL

- KTDL theo hướng kiểm tra: Đề xuất giả thiết và hệ thống kiểm tra tính đúng đắn của giả thuyết, KTDL theo hướng kiểm tra gồm: truy vấn, báo cáo, phân tích thống kê.
- KTDL theo hướng khám phá: Tìm kiếm những tri thức tiềm ẩn trong CSDL.

Ứng dụng của khai thác dữ liệu



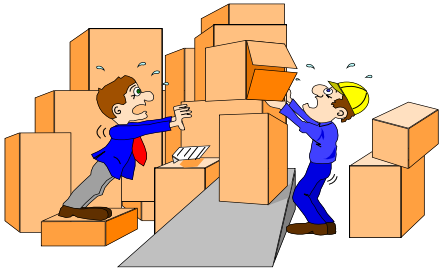
- Trong ngân hàng: Dự đoán rủi ro tính dụng
- Trong thương mại điện tử : Web, bán hàng qua mạng
- Công nghệ sinh học và dược phẩm : Phân tích các dữ liệu di truyền.
- Nhân sự: Chọn ứng cử viên khi tuyển dụng

Các ứng dụng

Kinh doanh

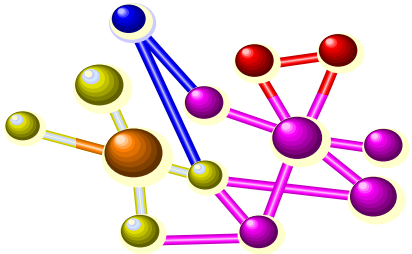
- Phân tích dữ liệu bán hàng và tiếp thị
- Phân tích đầu tư
- Chứng khoán
- Xác định gian lận

Sản xuất



- Điều khiển và lập lịch
- Quản trị mạng lưới
- Phân tích kết quả thử nghiệm

Khoa học

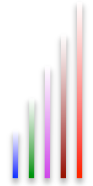


- Không gian
- Sinh học
- Địa lý
- etc.

Y học

- Bệnh lý
- Sinh học

Các kỹ thuật khai thác dữ liệu



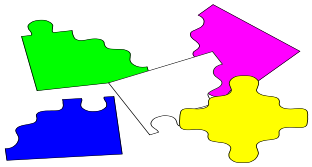
- Tập phổ biến và luật kết hợp
- Khai thác mẫu tuần tự
- Tập thô (reduct)
- Phân lớp dữ liệu
- Gom cụm (Clustering)

Các kỹ thuật khai thác dữ liệu

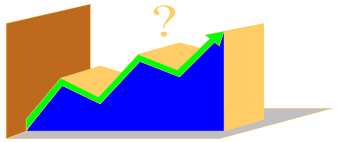
Tìm các đặc trưng của lớp các đối tượng và sử dụng để phân lớp dữ liệu mới.

Phân lớp

Xác định các cụm tiềm ẩn trong các tập đối tượng chưa được xếp lớp.



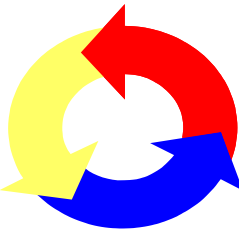
Gom cụm



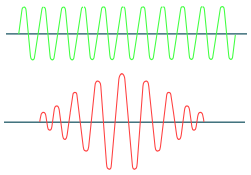
Dữ đoán dữ liệu tương lai dựa trên dữ liệu quá khứ.

Dự đoán

Tìm các mẫu phổ biến từ dữ liệu và mối quan hệ của các đối tượng dữ liệu.



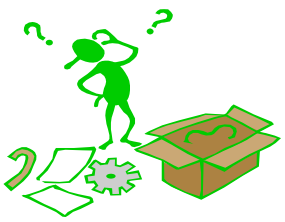
Luật kết hợp



Khám phá các mẫu tín hiệu phổ biến nhất từ dữ liệu các sự kiện

Mẫu tuần tự

Xác định trật tự dữ liệu, cấu trúc lưu trữ phù hợp với tác vụ khai phá



Nhà kho- OLAP

Tập phổ biến và luật kết hợp

- Tìm các thuột tính xuất hiện phổ biến của các đối tượng dữ liệu. Từ tập phổ biến này ta tiến hành tạo ra các luật kết hợp nhằm phát hiện khả năng xuất hiện đồng thời của các thuộc tính trong tập các đối tượng.
- Nếu mua X thì sẽ mua Y. (có 66.6% khách hàng mua Bia thì sẽ mua mực)

Khai thác mẫu tuần tự

- Khai thác các mẫu tuần tự phổ biến phản ánh mối quan hệ giữa các biến cố trong CSDL hướng thời gian
- $X \rightarrow Y$ sự xuất hiện biến cố X sẽ dẫn đến sự xuất hiện của biến cố Y
- 80% khách hàng gởi tiền tiết kiệm trên 80 triệu thì 3 tháng sau gởi thêm 20 triều nữa
- Dùng để khám phá xu thế phát triển của đối tượng

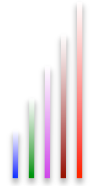
Tập thô (reduct)

- Dùng để rút gọn chiều trong bài táon phân lớp dữ liệu

Phân lớp dữ liệu

- Khám phá các luật phân loại cho tập dữ liệu.
- Ví dụ: Những bệnh nhân có các triệu chứng ho, lạnh, nhức đầu thì được phân lớp vào bệnh sốt rét.

Gom cụm (Clustering)



- Phân lớp dữ liệu là tiến trình phân các đối tượng thành các cụm đối tượng.
- Sao cho:
  - Các đối tượng trong cùng một cụm có mực độ tương đồng càng cao
  - Các đối tượng khác cụm có mức độ tương đồng thấp

Kết luận

- KPDL: tiến trình khám phá bán tự động các thông tin, mẫu có ích từ CSDL lớn
- Các bước của KDD
  - Tiền xử lý
  - KTDL(data mining tasks)
  - Hậu xử lý
- Các quan niệm, khía cạnh …
  - CSDL (quan hệ, hướng đối tượng, không gian, WWW, …)
  - Tri thức (đặc trưng, gom cụm, kết hợp, …)
  - Kỹ thuật (máy học, thống kê, trực quan hóa, …)
  - Ứng dụng (bán lẻ, điện thoại, khai thác Web …)