Khai phá dữ liệu (Data Mining)

Nguyễn Đình Thuân

UIT – VNU HCM

Nội dung môn học

1. Chương 1: Tổng quan về khai phá dữ liệu
2. Chương 2: Luật kết hợp
3. Chương 3: Dãy phổ biến
4. Chương 4: Phân lớp dữ liệu
5. Chương 5: Gom cụm dữ liệu
6. Giới thiệu 10 thuật toán hàng đầu của DataMining

Thông tin về môn học

Phương pháp đánh giá	*Trọng số[%]*
Chuyên cần, bài tập trên lớp	10%
Thực hành, thí nghiệm	15%
Kiểm tra giữa kỳ	15%
Tiểu luận, báo cáo trên lớp	20%
Thi cuối học kỳ	40%

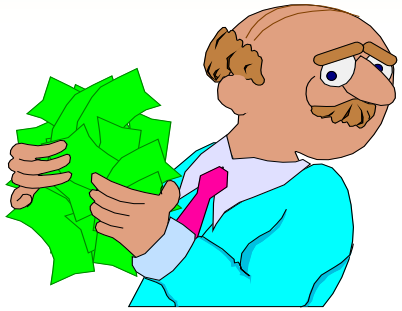
Tài liệu tham khảo

1. **Đỗ Phúc**, *Giáo trình + Slide Bài giảng Khai thác dữ liệu,* ĐHQG TPHCM, 2005.
2. **Hồ Tú Bảo**, *Introduction to knowledge discovery and data mining*, IOIT, 2001.
3. **Jiawei Han and Micheline Kamber**, *Data Mining Concepts and Techniques, University of Illinois, Morgan Kaufmann Publishers, 2006.*
4. **X. Wu, V. Kumar, J. Ross Quinlan, ...** *Top 10 Algorithms in Data Mining*, Chapman & Hall/CRC, Taylor & Francis Group, LLC, 2009.
5. **ZhaoHui Tang & Jamie MacLennan,** *Data Mining with SQL Server 2005,* Wiley Publishing, 2005.

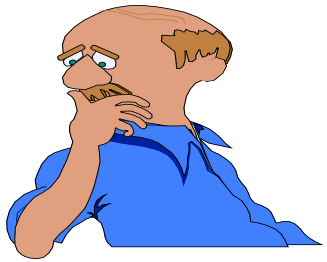
Chương 1 Tổng quan

Các khái niệm cơ bản

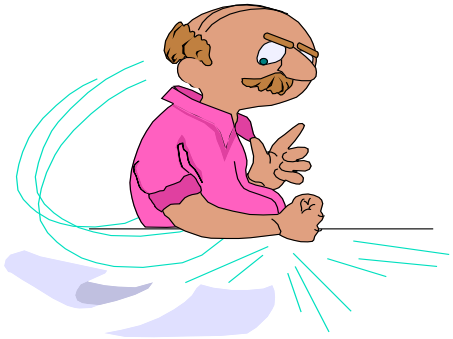
Dữ liệu (**Data**): có thể xem là chuỗi các bit, là số, ký tự…mà chúng ta thu thập hàng ngày trong công việc.



Thông tin (**Information**): là tập hợp của những dữ liệu đã được xử lý, dùng mô tả, giải thích đặc tính của một đối tượng nào đó.



Tri thức (**Knowledge**): là tập hợp những thông tin có liên hệ với nhau, được lập luận chặt chẽ hoặc được thực nghiệm kiểm chứng quan nhiều thế hệ. Tri thức thể hiện tư duy của con người về một vấn đề.



Các khái niệm cơ bản

Khám phá tri thức từ cơ sở dữ liệu:
(Knowledge Discovery in Databases – KDD)

- "KDD is the automatic extraction of non-obvious, hidden knowledge from large volumes of data." Fayyad, Platetsky-Shapiro, Smyth (1996)
- "Khám phá tri thức từ cơ sở dữ liệu là quy trình bao gồm nhiều công đoạn như: xác định vấn đề, tập hợp và chọn lọc dữ liệu, khai thác dữ liệu, đánh giá kết quả, giải thích dữ liệu, áp dụng tri thức vào thực tế
- http://www.kdnuggets.com/

Tại sao phải khai phá dữ liệu ?

John Naisbitt (www.naisbitt.com/) in 1982:

"We are drowning in data, but starving for knowlegde".

Dữ liệu được thu thập hàng ngày là rất lớn

- **Các CSDL khổng lồ**
- **Dữ liệu từ Internet**

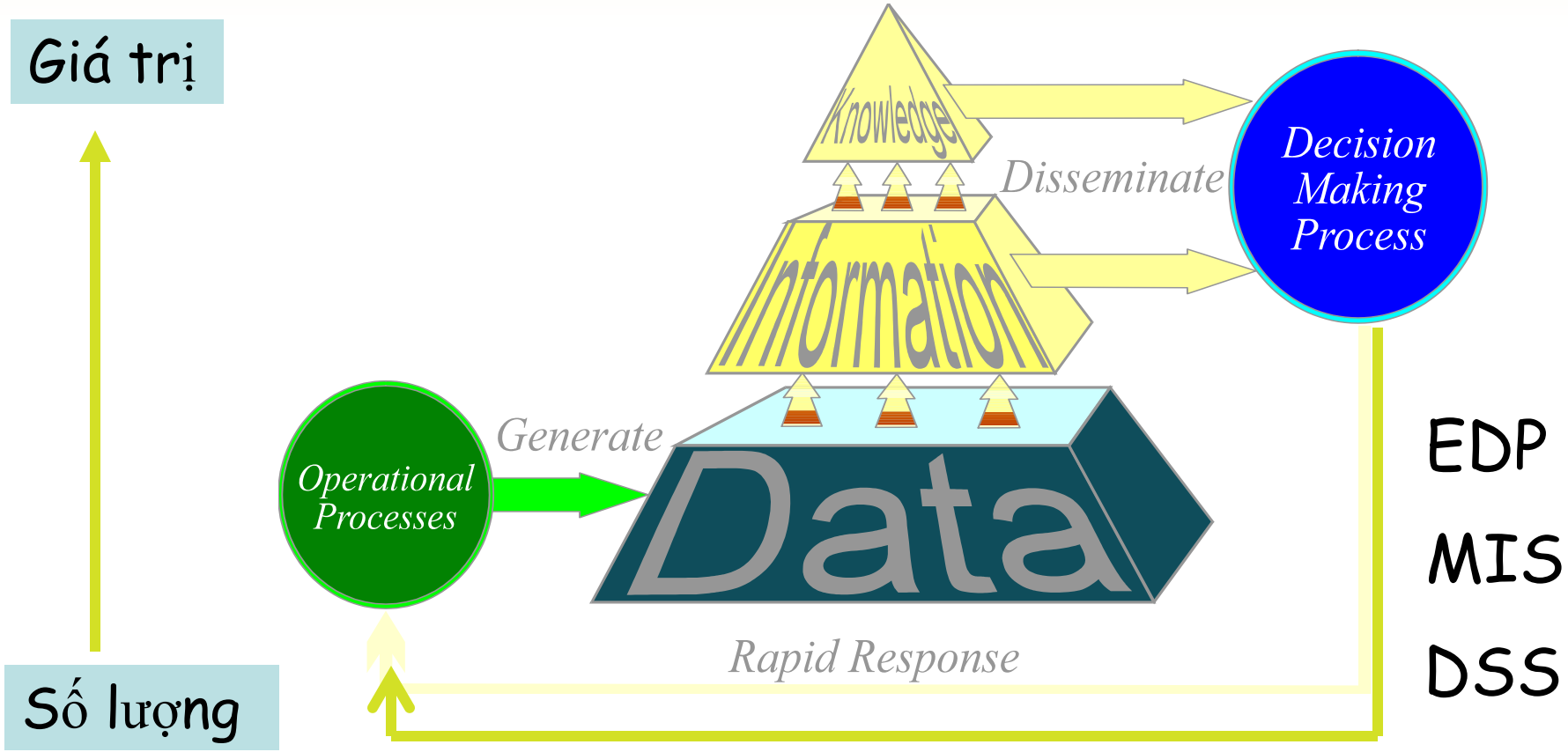
Theo các báo cáo của IBM, chỉ có 80% dữ liệu được khai thác, 20% còn lại ẩn trong các Database là những tri thức quý giá

Khai phá dữ liệu là gì?

Khai phá dữ liệu (Datamining) là một bước trong quy trình khám phá tri thức, nhằm:

- 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 .

Lợi ích của khai phá dữ liệu



EDP: Electronic Data Processing
MIS: Management Information Systems
DSS: Decision Support Systems

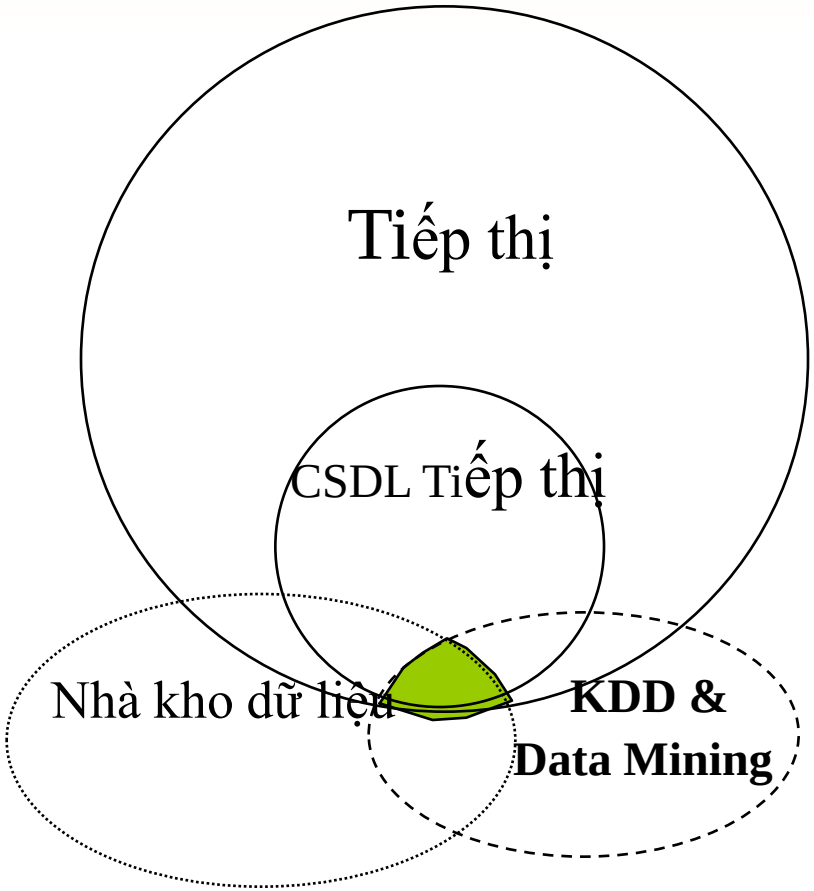
Khai phá dữ liệu là gì ?

Thuật ngữ:

- **Khai phá dữ liệu - Data mining**
  - **KPDL là một buớc của tiến trình KDD**
- **Knowledge discovery in databases (KDD)**
  - **Thuật ngữ tổng quát gồm các buớc như tiền xử lý, KPDL, hậu xử lý .**

Khai phá dữ liệu có ích lợi gì ?

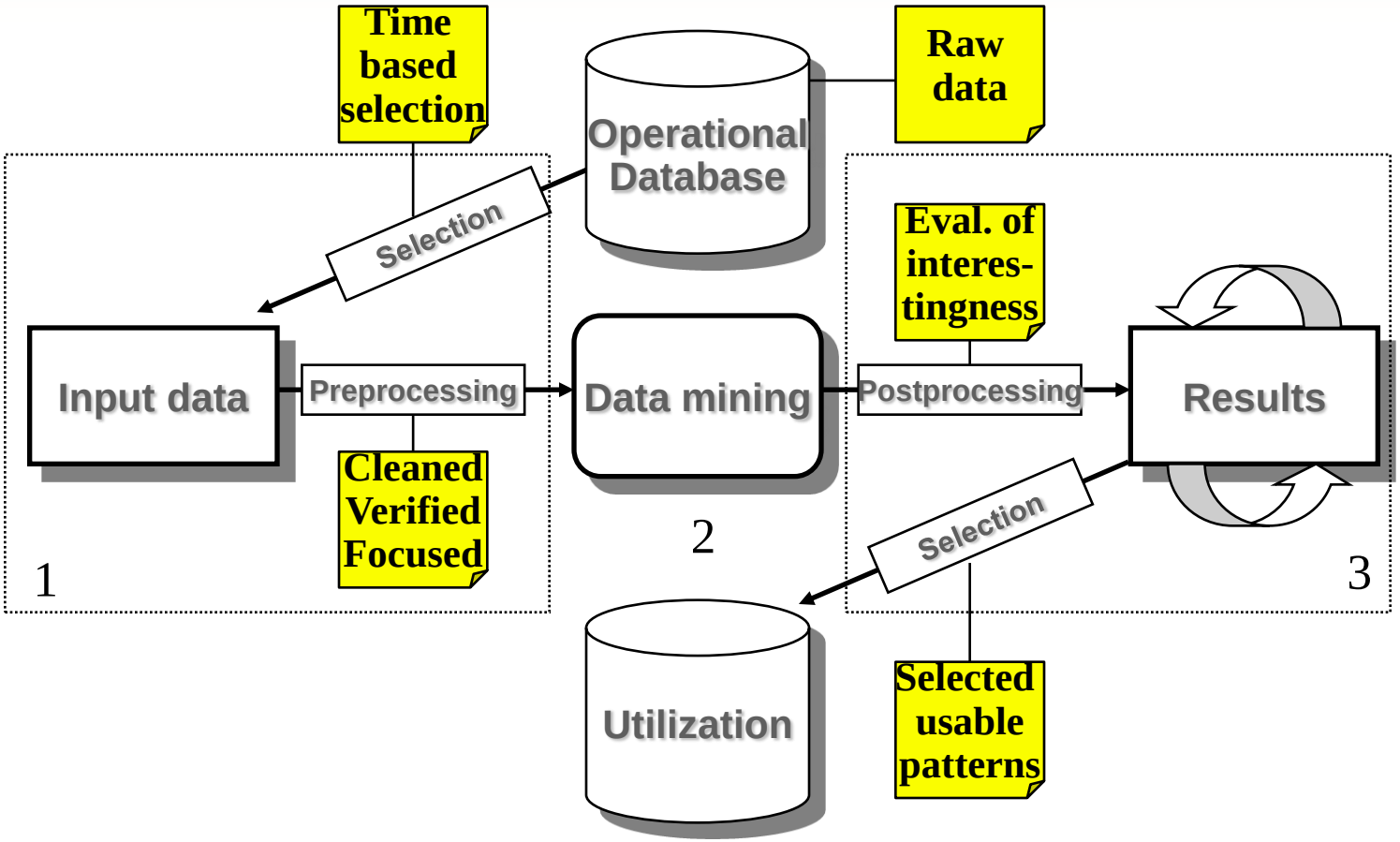
- Cung cấp tri thức hỗ trợ ra quyết định
- Dự báo
- Khái quát dữ liệu



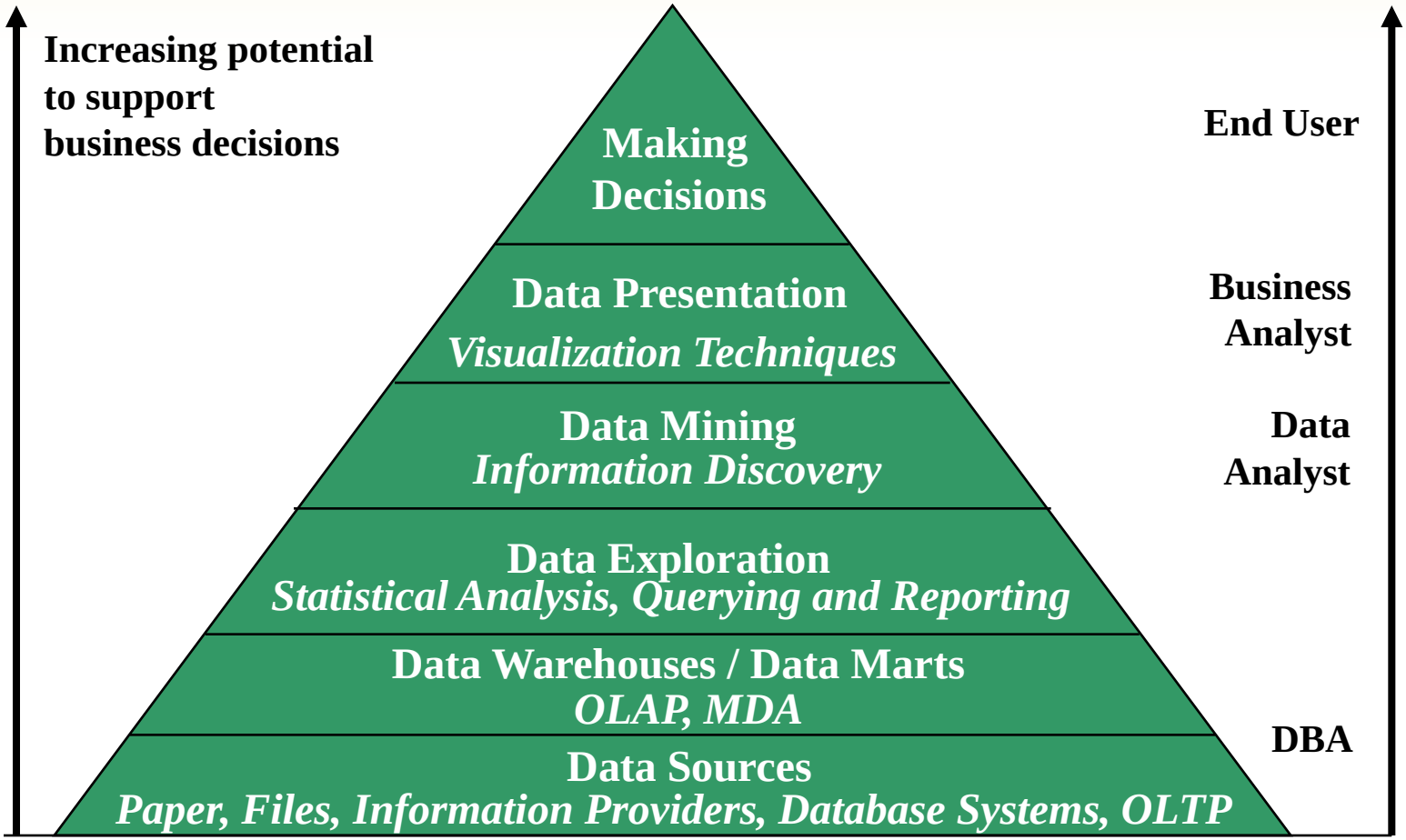
Tiến trình khai phá dữ liệu(1)

Tiến trình khai phá dữ liệu(2)

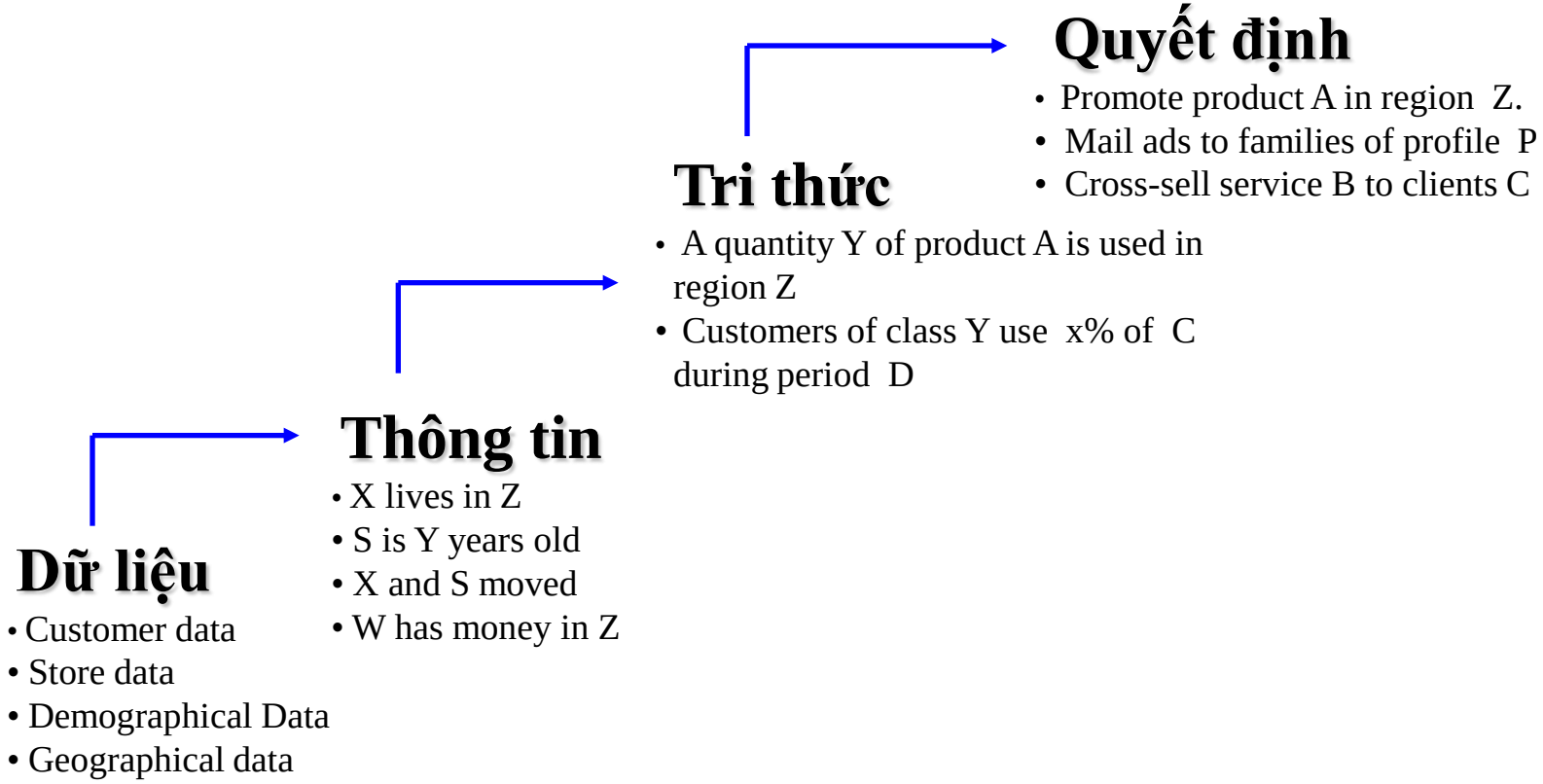
Tiến trình KDD tiêu biểu



Khai phá dữ liệu



Từ dữ liệu đến quyết định



Các quan niệm về KPDL

Các tiếp cận tổng quan:

- KPDL mô tả :
  - Cho biết điều gì là hữu ích có thể tìm thấy được trong dữ liệu
  - Giải thích dữ liệu đó
- KPDL dự báo:
  - Dựa trên dữ liệu quá khứ, dự báo tương lai
  - Xu thế phát triển!

Các quan niệm về KTDL

Quan niệm dựa trên …

- CSDL để khai thác
- Tri thức được khám phá
- Các kỹ thuật được sử dụng
- Các ứng dụng

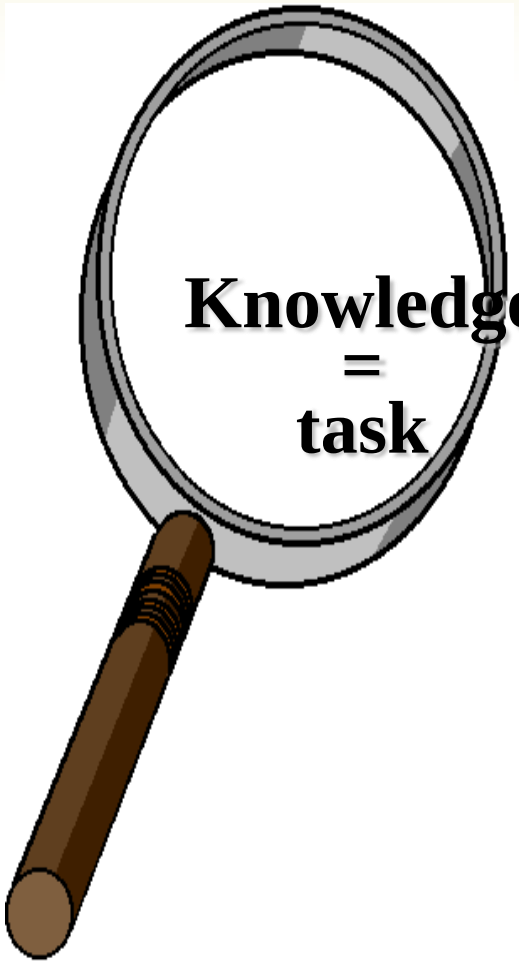
Các quan niệm về KPDL

CSDL cần khai thác

Quan hệ
Giao tác
Hướng đối tượng
Hướng đối tượng, quan hệ
Active
Không gian
Thời gian

Text, XML
Multi-media
Heterogeneous
Legacy
Inductive
WWW
etc.

Các quan niệm về KPDL



Tác vụ khai thác

Đặc trưng
Phân biệt
Kết hợp
Phân lớp
Gom cụm
Xu thế

Phân tích độ lệch
Phân tích hiếm
etc.

Các quan niệm KPDL

Các kỹ thuật đã sử dụng

- CSDL
- Nhà kho dữ liệu (OLAP)
- Máy học
- Thống kê
- Trực quan hóa
- Mạng nơron và thuật giải GA
- . . . .

Các quan niệm về KPDL

Các ứng dụng

Bán lẻ, siêu thị

Ngân hàng

Khai thác gen

Phân tích cổ phiếu

KTDL Web

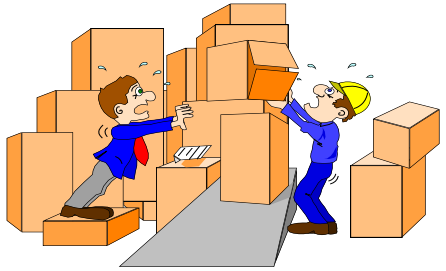
Phân tích dữ liệu

Các ứng dụng

Kinh doanh

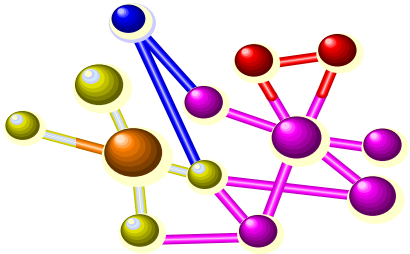
- Phân tích dữ liệu bán hàng và tiếp thị
- Phân tích đầu tư
- Chứng khoán
- Xác định gian lận

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

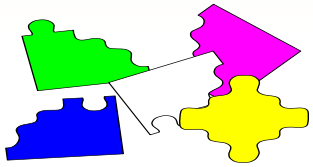
- Bệnh lý
- Sinh học

Các kỹ thuật sử dụng

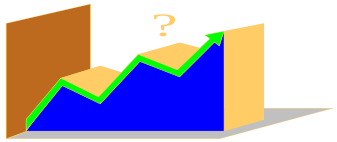
Tìm các đặc trưng của lớp các đối tượng và sử dụng để phân lớp dữ liệu mới.

Phân lớp

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



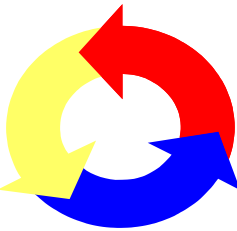
Gom cụm



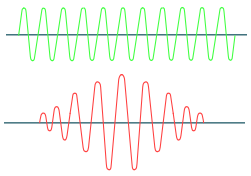
Dữ đoán dữ liệu tương lai dựa trên dữ liệu quá khứ.

Dự đoán

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



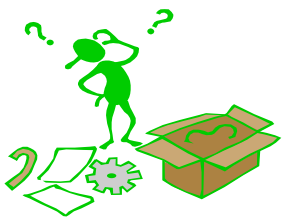
Luật kết hợp



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

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 …)