KHAI PHÁ DỮ LIỆU

Bài 1. Tổng quan về khai phá dữ liệu

Giáo viên: TS. Trần Mạnh Tuấn

Bộ môn: Hệ thống thông tin

Khoa: Công nghệ thông tin

Email: tmtuan@tlu.edu.vn

Điện thoai: 0983.668.841

Nội dung

1. Giới thiệu chung
2. Khai phá dữ liệu là gì
3. Quá trình khai phá tri thức
4. Các kỹ thuật áp dụng trong KPDL
5. Ứng dụng khai phá dữ liệu

GIỚI THIỆU CHUNG

Tình huống 1

Người đang sử dụng thẻ ID = 1234 thật sự là chủ nhân của thẻ hay là một tên trộm?

GIỚI THIỆU CHUNG

Tình huống 2

Tid	Refund	Marital Status	Taxable Income	Evade
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

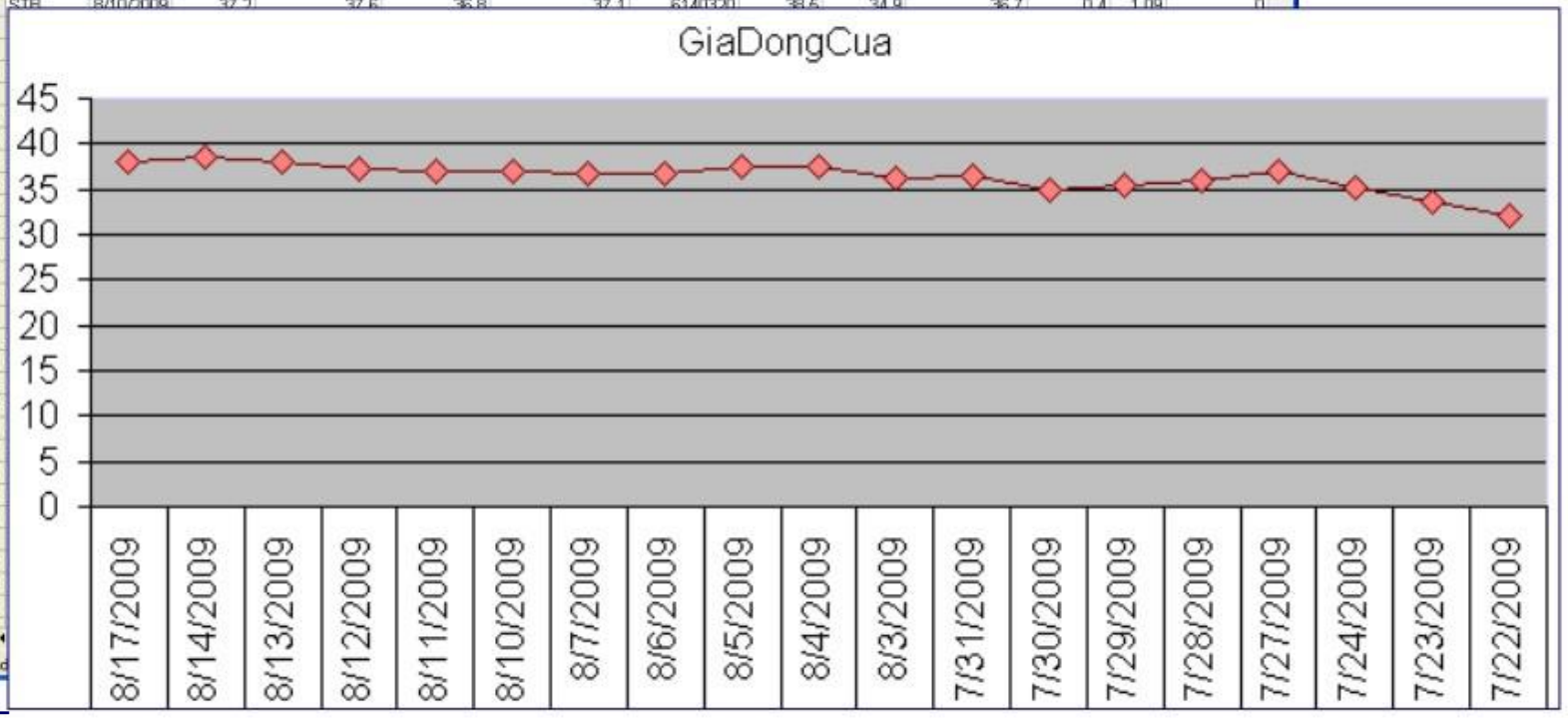
Ông A (Tid = 100) có khả năng trốn thuế???

GIỚI THIỆU CHUNG

Tình huống 3

Microsoft Excel - stb.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	MaCK	Ngay	GiaMoCua	GiaCaoNhat	GiaThapNhat	GiaDongCua	KhoiLuongGD	GiaTran	GiaSan	GiaThamChieu	TangGiam	%	GDThoaThua
2	STB	8/17/2009	38.5	38.8	38.1	38.1	5986700	40.4	36.6	38.5	-0.4	-1.04	24343
3	STB	8/14/2009	38	38.7	38	38.5	6886430	39.9	36.1	38	0.5	1.32	340000
4	STB	8/13/2009	38	38.5	37.6	38	8716920	39	35.4	37.2	0.8	2.15	188000
5	STB	8/12/2009	37.3	37.4	37	37.2	5361890	38.7	35.1	36.9	0.3	0.81	200000
6	STB	8/11/2009	37.1	37.3	36.9	36.9	3675610	38.9	35.3	37.1	-0.2	-0.54	0
7	STB	8/10/2009	37.2	37.6	36.8	37.1	6140320	38.5	34.9	36.7	0.4	1.09	0



Ngày mai cổ phiếu STB sẽ tăng???

GIỚI THIỆU CHUNG

Tình huống 4

Khóa	MãSV	MônHọc1	MônHọc2	...	TốtNghiệp
2004	1	9.0	8.5	...	Có
2004	2	6.5	8.0	...	Có
2004	3	4.0	2.5	...	Không
2004	8	5.5	3.5	...	Không
2004	14	5.0	5.5	...	Có
...	...	...	...	...	
2005	90	7.0	6.0	...	Có (80%)
2006	24	9.5	7.5	...	Có (90%)
2007	82	5.5	4.5	...	Không (45%)
2008	47	2.0	3.0	...	Không (97%)
...	...	...	...	...	...

Làm sao xác định được khả năng tốt nghiệp của một sinh viên hiện tại?

GIỚI THIỆU CHUNG

- Những năm 60 bắt đầu sử dụng công cụ tin học để tổ chức khai thác các CSDL
- Khả năng thu thập, lưu trữ, xử lý, phân tích dữ liệu của các hệ thống thông tin không ngừng thay đổi
- Lượng thông tin ngày càng tăng lên
- Hướng tiếp cận mới về khai thác thông tin đưa ra các quyết định, tư vấn,…

KHAI PHÁ DỮ LIỆU

We are **data rich**, but **information poor**

KHAI PHÁ DỮ LIỆU

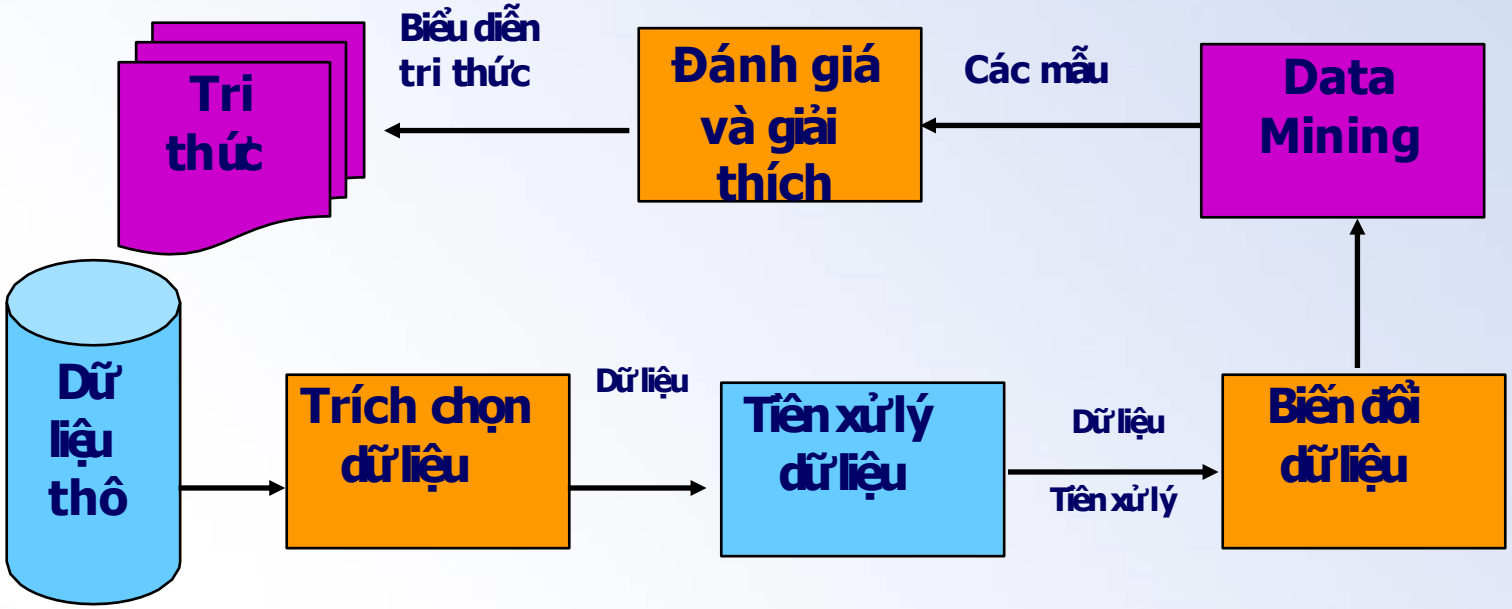
- Khai phá dữ liệu là một lĩnh vực nhằm tự động khai thác những thông tin tri thức đang tiềm ẩn trong dữ liệu.
- Khai phá dữ liệu là một lĩnh vực phát triển bền vững, mang lại nhiều lợi ích, triển vọng, ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống
- Các kỹ thuật được áp dụng dựa trên CSDL, học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê và tính toán hiệu năng cao.

KHAI PHÁ DỮ LIỆU

- Có nhiều quan điểm khác nhau về Khai phá dữ liệu.
- Khai phá tri thức trong CSDL (Knowledge Discovery in Databases - KDD) là mục tiêu chính của Khai phá dữ liệu.
- Khai phá dữ liệu là một bước chính trong khai phá tri thức.

Quá trình khám phá tri thức

Quy trình khám phá tri thức



Các giai đoạn khai phá tri thức

- Trích chọn dữ liệu: trích chọn những tập dữ liệu cần khai phá từ các tập dữ liệu khác nhau theo một tiêu chí nhất định.
- Tiền xử lý dữ liệu:
  - Làm sạch dữ liệu
  - Rút gọn dữ liệu
  - Rời rạc hoá dữ liệu

Sau bước này dữ liệu sẽ được nhất quán và đồng nhất

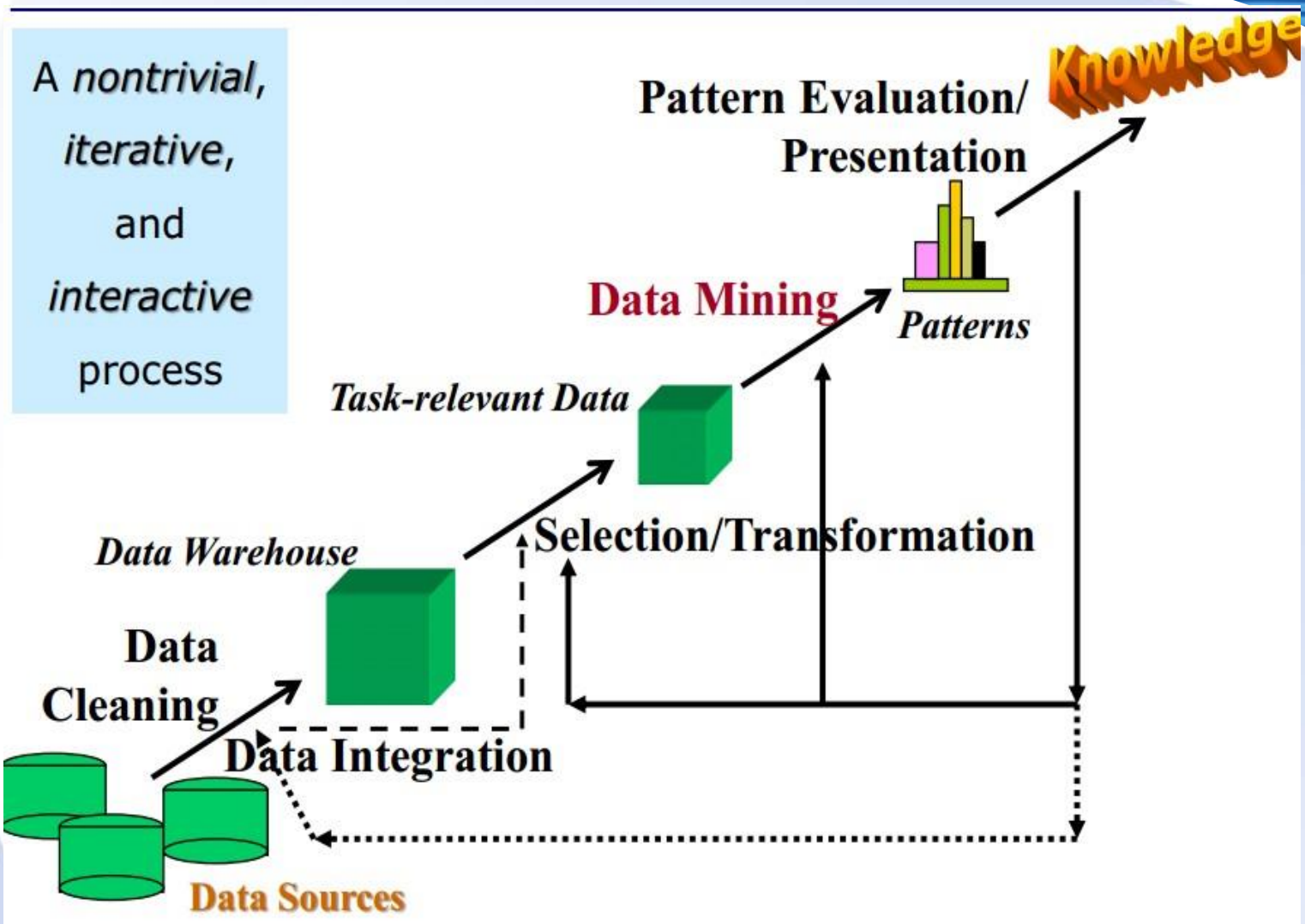
Các giai đoạn khai phá tri thức

- Biến đổi dữ liệu: là bước chuẩn hoá và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi phục vụ cho các kỹ thuật khai phá ở bước sau.
- Khai phá dữ liệu: áp dụng các kỹ thuật phân tích (thường là các kỹ thuật của học máy) nhằm:
  - Khai thác dữ liệu
  - Trích chọn mẫu thông tin
  - Xây dựng tri thức

Các giai đoạn khai phá tri thức

- Đánh giá và biểu diễn tri thức:
  - Những mẫu thông tin và mã liên hệ trong dữ liệu đã được khám phá ở bước trên được chuyển về biểu diễn ở một dạng gần với thế giới thực của người sử dụng như: đồ thị, cây, bảng biểu, luật,…
  - Đánh giá những tri thức khám phá được theo những tiêu chí nhất định.

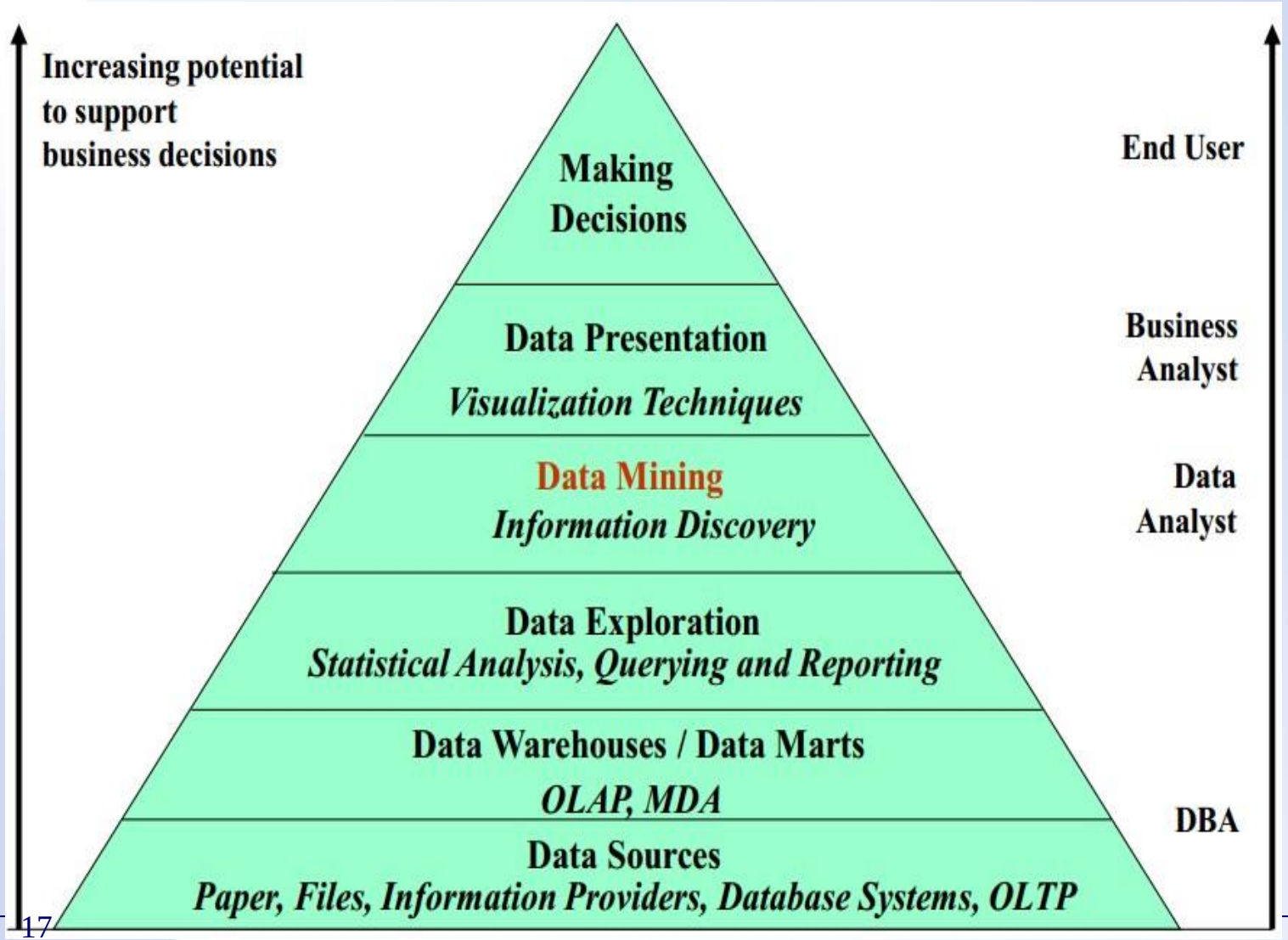
Quá trình khám phá tri thức



Quá trình khám phá tri thức

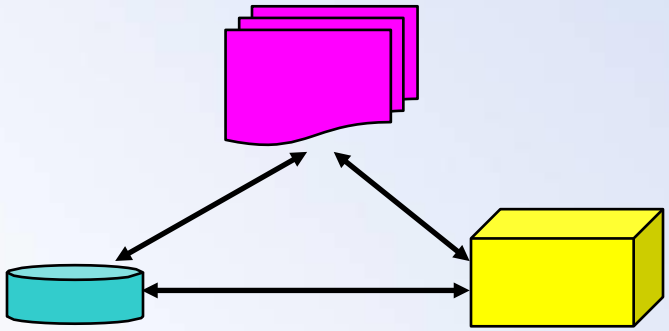
- Quá trình khám phá tri thức là một chuỗi lặp gồm các bước:
  - Data cleaning (làm sạch dữ liệu)
  - Data integration (tích hợp dữ liệu)
  - Data selection (chọn lựa dữ liệu)
  - Data transformation (biến đổi dữ liệu)
  - Data mining (khai phá dữ liệu)
  - Pattern evaluation (đánh giá mẫu)
  - Knowledge presentation (biểu diễn tri thức)

Quá trình khám phá tri thức



Các kỹ thuật áp dụng trong KPDL

Các lĩnh vực khoa học khác



Tổ chức dữ liệu

Học máy và trí tuệ nhân tạo

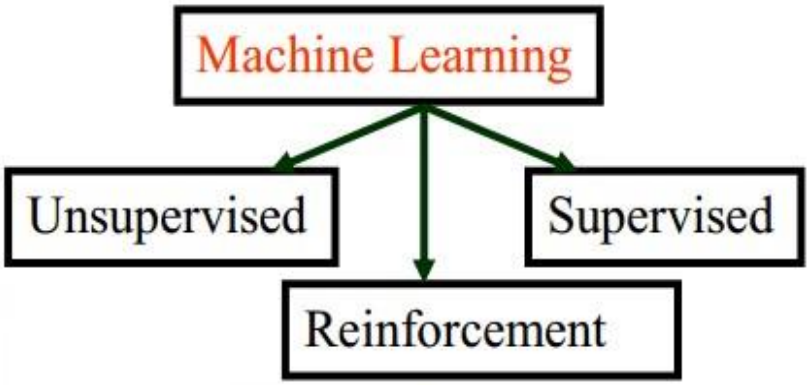
Các lĩnh vực liên quan đến khai phá tri thức

Các kỹ thuật áp dụng trong KPDL

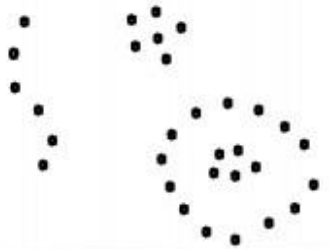
- ✓ **Đứng trên quan điểm của học máy (Machine Learning), các kỹ thuật trong Data Mining gồm:**
- Học có giám sát (Supervised learning): Quá trình gán nhãn lớp cho các phần tử trong CSDL dựa trên một tập các VDHL và các thông tin về nhãn lớp đã biết.
- Học không có giám sát (Unsupervised learning): Quá trình phân chia một tập dl thành các lớp/cụm (clustering) dl tương tự nhau mà chưa biết trước các thông tin về lớp/tập các VDHL.
- Học nửa giám sát (Semi - Supervised learning): Là quá trình phân chia một tập dl thành các lớp dựa trên một tập nhỏ các VDHL và một số các thông tin về một số nhãn lớp đã biết trước.

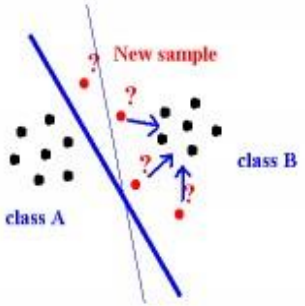
Các kỹ thuật áp dụng trong KPDL

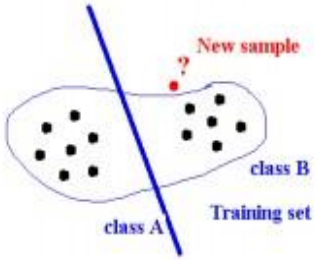
- Khai phá dữ liệu và học máy

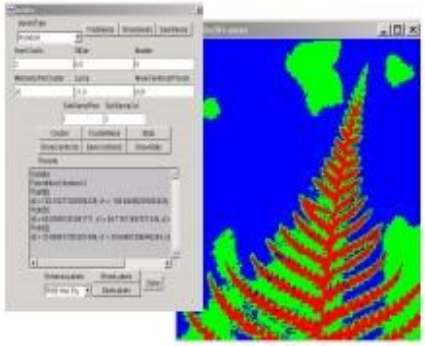


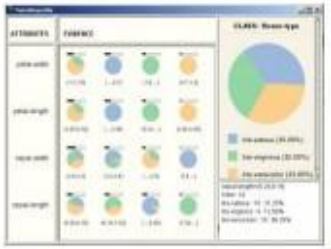
"Natural groupings"



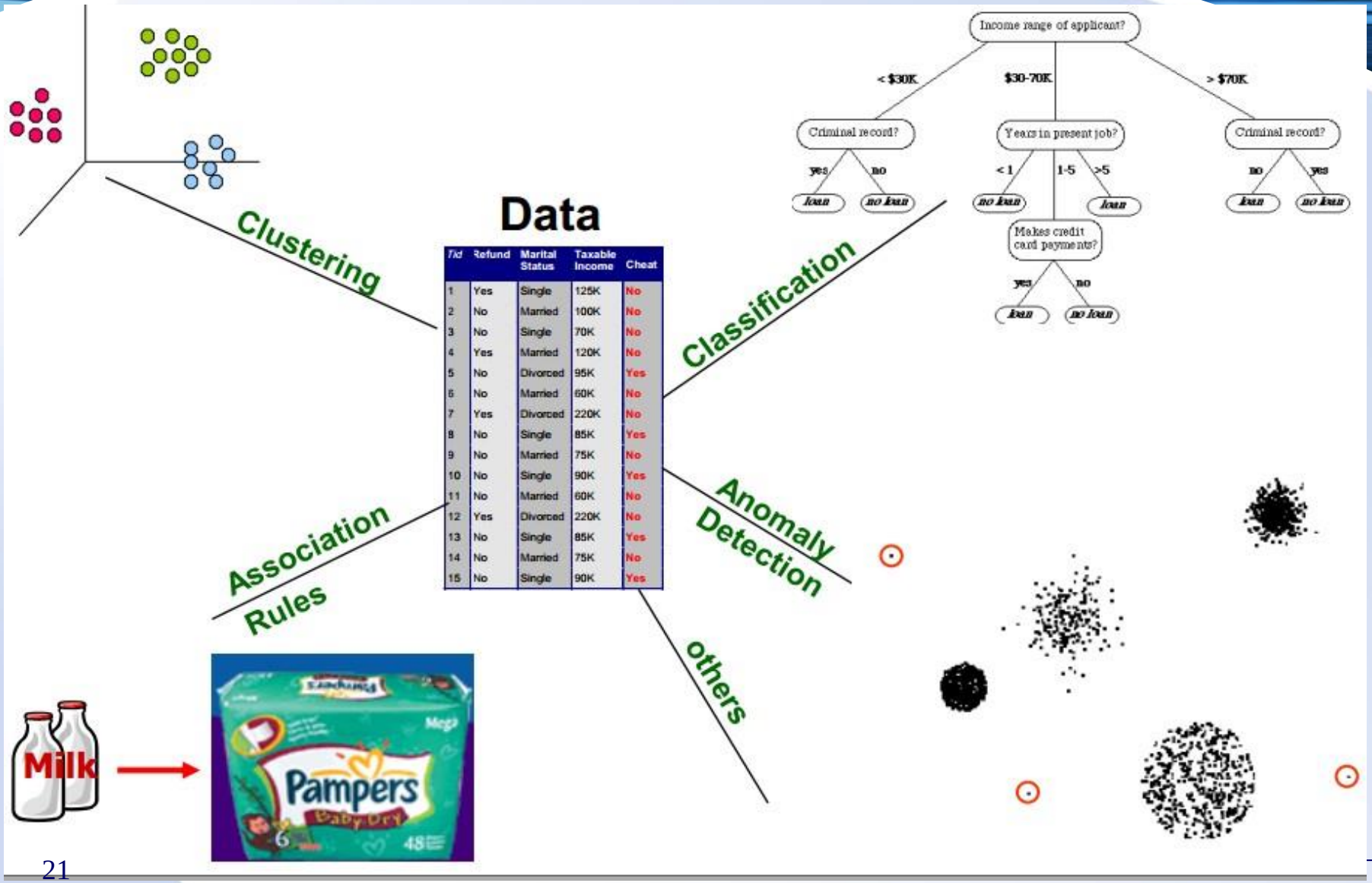








Các kỹ thuật áp dụng trong KPDL



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
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

Các kỹ thuật áp dụng trong KPDL

Bốn thành phần cơ bản của một giải thuật khai phá dữ liệu

- Cấu trúc mẫu hay cấu trúc mô hình (model or pattern structure)
- Hàm tỉ số (score function)
- Phương pháp tìm kiếm và tối ưu hóa (optimization and search method)
- Chiến lược quản lý dữ liệu (data management strategy)

Các kỹ thuật áp dụng trong KPDL

- Một quá trình trích xuất tri thức từ lượng lớn DL
- Một quá trình không dễ trích xuất thông tin ẩn, hữu ích, chưa được biết trước từ dữ liệu
- Các thuật ngữ thường được dùng tương đương: knowledge discovery/mining in data/databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence

Các kỹ thuật áp dụng trong KPDL

- Tri thức đạt được từ quá trình khai phá
- Mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa)
- Mẫu thường xuyên, các mối quan hệ kết hợp/tương quan
- Mô hình phân loại và dự đoán
- Mô hình gom cụm
- Các phần tử biên
- Xu hướng hay mức độ thường xuyên của các đối tượng có hành vi thay đổi theo thời gian

Các kỹ thuật áp dụng trong KPDL

Một số dạng dữ liệu:

- CSDL quan hệ.
- CSDL đa chiều (multidimensional structures, data warehouses).
- CSDL dạng giao dịch.
- CSDL quan hệ - hướng đối tượng.
- Dữ liệu không gian và thời gian.
- Dữ liệu chuỗi thời gian.
- CSDL đa phương tiện.
- Dữ liệu Text và Web, …

Các kỹ thuật áp dụng trong KPDL

Lượng lớn dữ liệu sẵn có để khai phá

- Bất kỳ loại dữ liệu được lưu trữ hay tạm thời, có cấu trúc hay bán cấu trúc hay phi cấu trúc
- Dữ liệu được lưu trữ
- Các tập tin truyền thống
- Các cơ sở dữ liệu quan hệ hay quan hệ đối tượng
- Các cơ sở dữ liệu giao tác hay kho dữ liệu
- Các cơ sở dữ liệu hướng ứng dụng: cơ sở dữ liệu không gian, cơ sở dữ liệu thời gian, cơ sở dữ liệu không thời gian, cơ sở dữ liệu chuỗi thời gian, cơ sở dữ liệu văn bản, cơ sở dữ liệu đa phương tiện, …
- Các kho thông tin: the World Wide Web, …
- Dữ liệu tạm thời: các dòng dữ liệu

Các kỹ thuật áp dụng trong KPDL

Tri thức đạt được từ quá trình khai phá

- Tri thức đạt được có thể có tính mô tả hay dự đoán tùy thuộc vào quá trình khai phá cụ thể.
- Mô tả (Descriptive): có khả năng đặc trưng hóa các thuộc tính chung của DL được khai phá
- Dự đoán (Predictive): có khả năng suy luận từ dữ liệu hiện có để dự đoán.
- Tri thức đạt được có thể có cấu trúc, bán cấu trúc, hoặc phi cấu trúc.
- Tri thức đạt được có thể được/không được người dùng quan tâm -> các độ đo đánh giá tri thức đạt được.
- Tri thức đạt được có thể được dùng trong việc hỗ trợ ra quyết định, điều khiển quy trình quản lý thông tin, xử lý truy vấn

Các kỹ thuật áp dụng trong KPDL

Các đặc điểm được dùng để khảo sát một hệ thống khai phá dữ liệu

- Kiểu dữ liệu
- Các vấn đề hệ thống
- Nguồn dữ liệu
- Các tác vụ và phương pháp luận khai phá dữ liệu
- Vấn đề gắn kết với các hệ thống kho dữ liệu/cơ sở dữ liệu
- Khả năng co giãn dữ liệu
- Các công cụ trực quan hóa
- Ngôn ngữ truy vấn khai phá dữ liệu và giao diện đồ họa cho người dùng

Các kỹ thuật áp dụng trong KPDL

Phân biệt các hệ thống khai phá dữ liệu với

- Các hệ thống phân tích dữ liệu thống kê (statistical data analysis systems)
- Các hệ thống học máy (machine learning systems)
- Các hệ thống truy hồi thông tin (information retrieval systems)
- Các hệ cơ sở dữ liệu diễn dịch (deductive database systems)
- Các hệ cơ sở dữ liệu (database systems)
- ...

Các kỹ thuật áp dụng trong KPDL

Công nghệ hiện đại trong lĩnh vực quản lý thông tin

- Hiện diện khắp nơi (ubiquitous) và có tính ẩn (invisible) trong nhiều khía cạnh của đời sống hằng ngày
  - Làm việc, mua sắm, tìm kiếm thông tin, nghỉ ngơi, …
- Được áp dụng trong nhiều ứng dụng thuộc nhiều lĩnh vực khác nhau
- Hỗ trợ các nhà khoa học, giáo dục học, kinh tế học, doanh nghiệp, khách hàng, …

Ứng dụng trong KPDL

Là một lĩnh vực được quan tâm và ứng dụng rộng rãi:

- Phân tích dữ liệu và hỗ trợ quyết định
- Điều trị y học.
- Text mining & Web mining
- Tin-sinh (bio-informatics).
- Tài chính và thị trường chứng khoán.
- Bảo hiểm (insurance), .v.v.

Ứng dụng trong KPDL

- Trong thiên văn Hệ thống SKICAT dùng phân tích ảnh, phân loại và xếp nhóm các vật thể không gian từ các ảnh quan sát vũ trụ.
- Dùng để xử lý 3 terabytes dữ liệu ảnh từ Đài thiên văn Palomar, với khoảng 1 tỉ vật thể không gian phát hiện được.
- SKICAT có thể làm được những công việc tính toán cực lớn trong việc phân loại các ảnh vật thể không rõ ràng.

Ứng dụng trong KPDL

- Trong kinh doanh: các UD trong tiếp thị, tài chính (đặc biệt là đầu tư), phát hiện gian lận, sản xuất, viễn thông và các Internet agent (tác tử).
- Tiếp thị: UD trong hệ thống CSDL tiếp thị, phân tích các DL khách hàng để phân loại các nhóm khách hàng khác nhau và dự báo về sở thích của họ.
- Đầu tư: LBS Capital Management dùng để quản lý danh mục vốn đầu tư.
- Phát hiện gian lận:
  - Hệ thống HNC Falcon and Nestor PRISM dùng để theo dõi các gian lận thẻ tín dụng.
  - Hệ thống FAIS dùng để thẩm định các giao dịch thương mại gồm cả việc chuyển tiền bất hợp pháp

Trao đổi, câu hỏi?