TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

Giảng viên: Đặng Thị Thu Hiền, Nguyễn Tu Trung
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2022

Nội dung

- Một số tình huống
- Giới thiệu chung
- Data Mining là gì?
- Khám phá tri thức trong CSDL
- Các kỹ thuật áp dụng trong Data mining
- Cây phân cấp các kỹ thuật áp dụng KPDL
- Các dạng dữ liệu có thể khai phá
- Các tác vụ KPDL
- Các thành tố cơ bản đặc tả tác vụ KPDL
- Bốn thành phần cơ bản của giải thuật KPDL
- Quy trình KPDL
- Hệ thống KPDL
- Ý nghĩa và vai trò của KPDL
- Ứng dụng của KPDL

Một số tình huống

- Tình huống 1
- Tình huống 2
- Tình huống 3
- Tình huống 4

Tình huống 1

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

Tình huống 2

- Liệu ông A (Tid = 100) có khả năng trốn thuế???

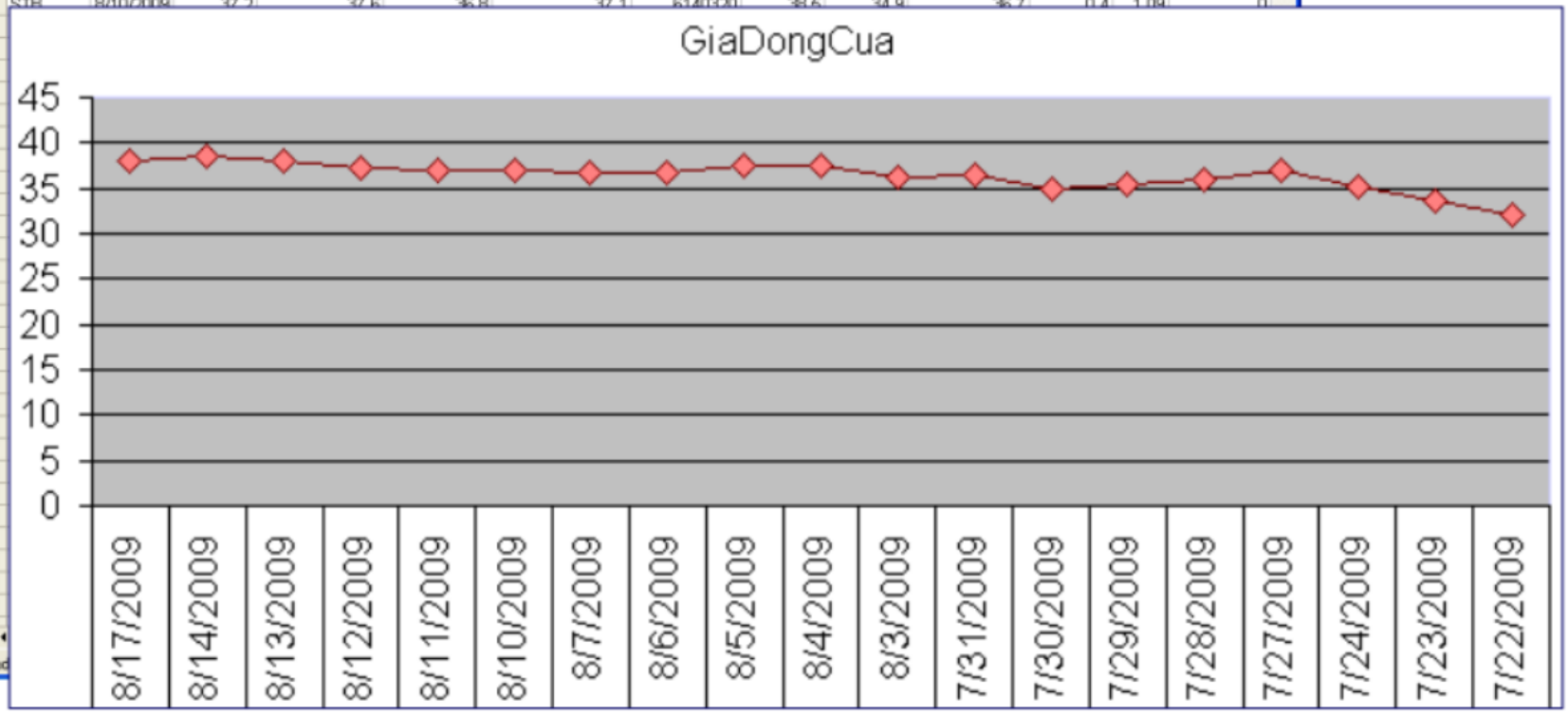
Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tình huống 3

- Ngày mai cổ phiếu STB sẽ tăng???

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	6140330	38.5	34.9	36.7	0.4	1.09	0



Tình huống 4

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

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%)
...	...	...	...	...	...

Giới thiệu chung

- Những năm 60 đã bắt đầu sử dụng các công cụ tin học để tổ chức và khai thác các CSDL
- Khả nẵng thu thập, lưu trữ và xử lý dữ liệu cho các hệ thống tin học không ngừng được nâng cao
- Thông tin được lưu trữ trên các thiết bị như đĩa, bằng từ, đĩa CD-ROM... cũng tăng lên
- Lượng thông tin trên các hệ thống tin học cứ sau 20 tháng lại tăng gấp đôi
- Cuối thập kỷ 80 sự phát triển rộng khắp của các CSDL ở mọi quy mồ đã tạo ra sự bùng nổ thông tin trên toàn cầu
- Đề cập đến khái niệm khủng hoảng phân tích dữ liệu tác nghiệp để cung cấp thông tin với yêu cầu chất lượng cao cho người làm quyết định trong các tổ chức tài chính, thương mại, khoa học,...

Giới thiệu chung

- Chúng ta giàu dữ liệu nhưng nghèo thông tin
- John Naisbett đã cảnh báo "Chúng ta đang chim ngập trong dữ liệu mà vân đói tri thức"

- Lượng dữ liệu khổng lồ này thực sự là một nguồn "tài nguyên" có nhiều giá trị bởi thông tin là yếu tố then chốt trong mọi hoạt động q quản lý, kinh doanh, phát triển sản xuất và dịch vụ

Data Mining là gì?

- Là một lĩnh vực
  - Nhằm tự động khai thác những thông tin, tri thức có tính tiềm ẩn, hữu ích từ những CSDL lớn cho các đơn vị, tổ chức, doanh nghiệp...
  - Phát triển bền vững, mang lại nhiều lợi ích, triển vọng, ưu thế hơn hẳn so với các công cụ phân tích dữ liệu
- Các kỹ thuật được áp dụng phần lớn thừa kế từ CSDL, học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê, và tính toán hiệu nẵng cao
- Có nhiều quan điểm khác nhau về Data Mining

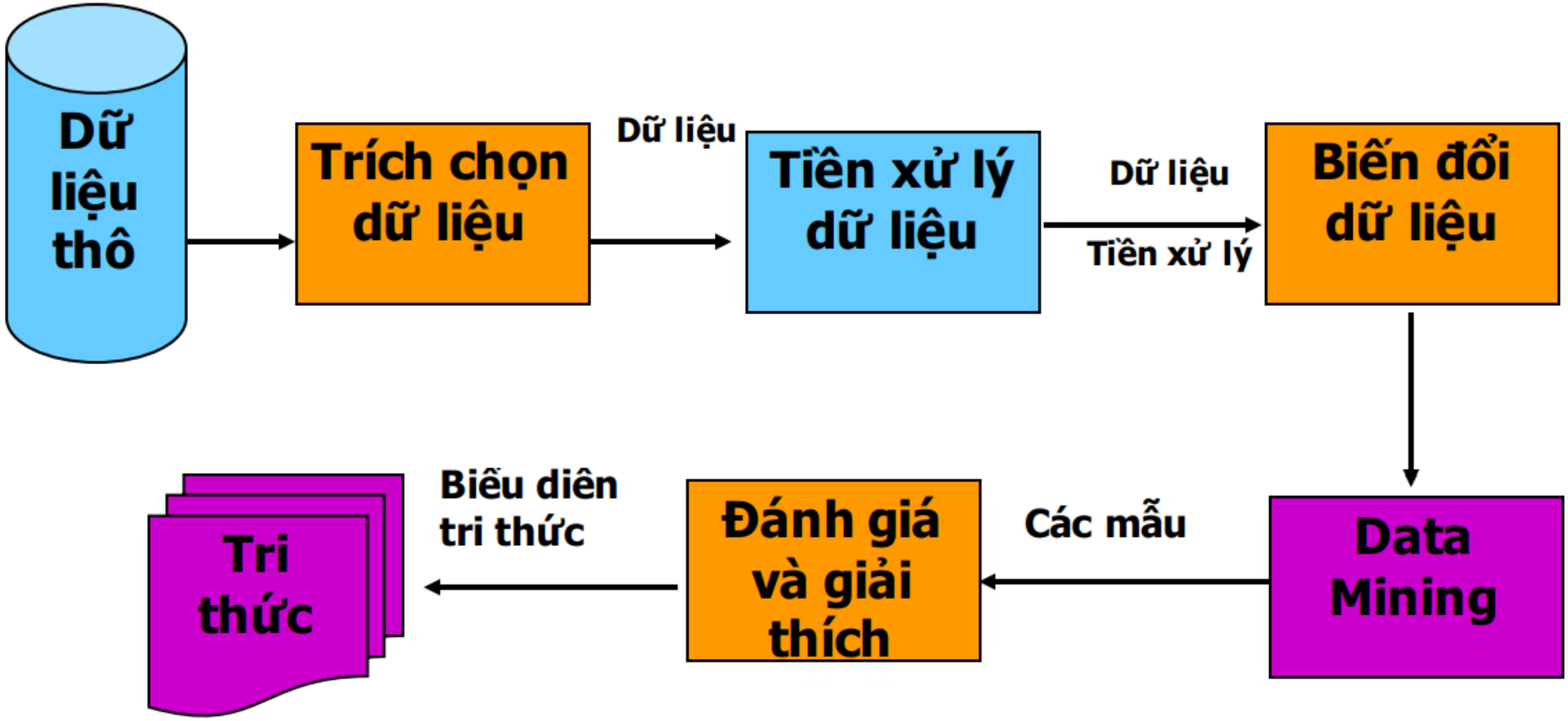
Data Mining là gì?

- Định nghĩa: DATA MINING là một quá trình tim kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong CSDL lớn
- Khám phá trí thức trong CSDL (Knowledge Discovery in Databases - KDD) là mục tiêu chính của Data mining
- Khái niệm Data Mining và KDD được các nhà khoa học xem là tương đương với nhau
- Xét một cách chi tiết thì Data Mining là một bước chính trong qua trinh KDD

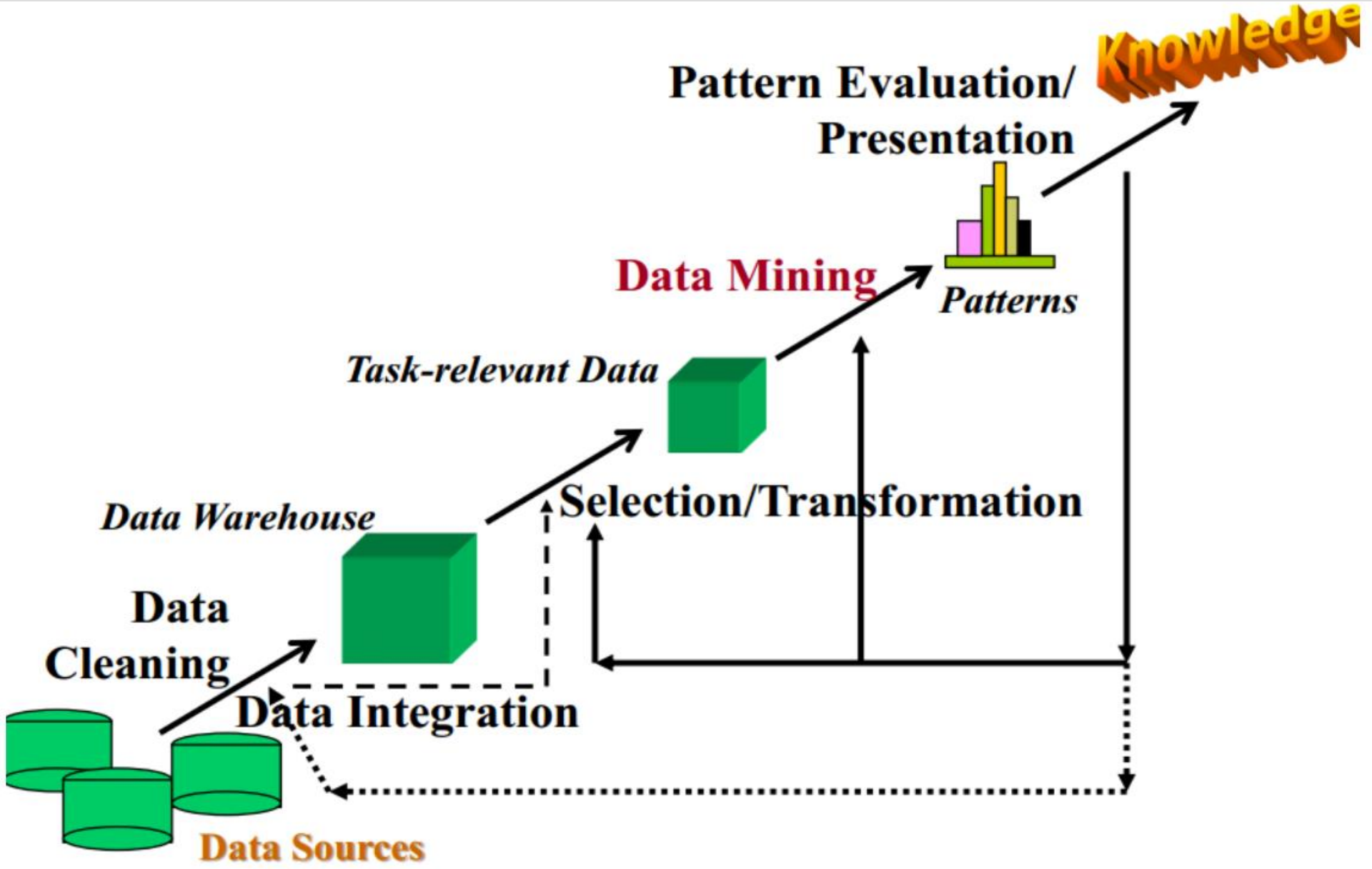
Khám phá tri thức trong CSDL

- Quy trình khám phá tri thức
- Quy trình lặp trong khám phá tri thức
- Các giai đoạn của khám phá tri thức
- Chuỗi các bước trong khai phá tri thức
- Tác động đến các đối tượng
- Tháp khám phá tri thức và ra quyết định

Quy trình khám phá tri thức



Quy trình lặp trong khám phá tri thức



Các giai đoạn của khám phá tri thức

- Trích chọn dữ liệu:
  - Trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data warehouses, data repositories) ban đầu theo một số tiêu chí nhất định
- Tiền xử lý dữ liệu:
  - Làm sạch dữ liệu (xử lý với dl không đầy đủ, nhiễu, không nhất quán...)
  - Rút gọn dữ liệu (sử dụng hàm nhóm và tính tổng, các phương pháp nén dữ liệu , sử dụng histograms, lấy mẫu...)
  - Rời rạc hóa dữ liệu (rời rạc hóa dựa vào histograms, entropy, phân khoảng...)
  - => Sau bước này, di sẽ nhất quản, đầy đủ, được rút gọn, rời rạc hóa
- Biến đổi dữ liệu:
  - Là bước chuẩn hóa và làm mịn dl để đưa dữ liệu về dạng thuận lợi phục vụ cho các kỹ thuật khai phá ở bước sau

Khám phá tri thức trong CSDL

- Data mining:
  - Áp dụng những kỹ thuật phân tích (thường là các kỹ thuật của học máy) nhằm khai thác dl, trích chọn những mẫu thông tin, mối liên hệ đặc biệt trong dữ liệu
  - Đây là bước quan trọng và tốn nhiều thời gian nhất của toàn quá trình KDD
- Đánh giá và biểu diễn tri thức
  - Những mẫu thông tin và mối liên hệ trong dữ liệu đã được khám phá ở bước trên được chuyển dạng và biểu diễn ở một dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật...
  - Đánh giá những tri thức khám phá được theo những tiêu chí nhất định

Chuỗi các bước trong khai phá tri thức

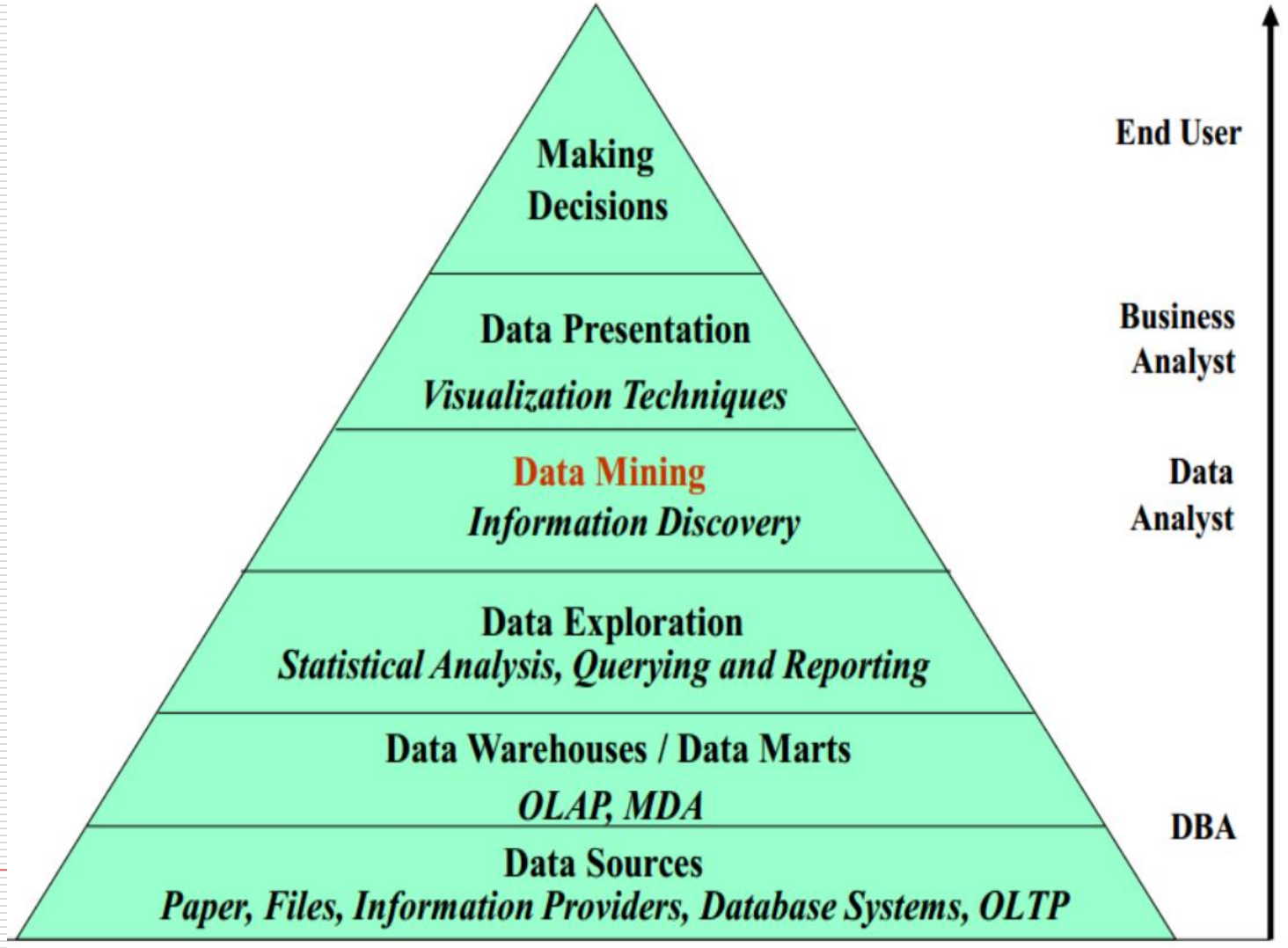
- Data cleaning (làm sạch dữ liệu)
- Data integration (tích hợp dữ liệu)
- Data selection (chọn lựa dữ liệu)
- Data transformation (biến đổi dữ liệu)
- Data mining (khai phá dữ liệu)
- Pattern evaluation (đánh giá mẫu)
- Knowledge presentation (biểu diễn tri thức)

Tác động đến các đối tượng

- Data sources (các nguồn dữ liệu)
- Data warehouse (kho dữ liệu)
- Task-relevant data (dữ liệu cụ thể sẽ được khai phá)
- Patterns (mẫu kết quả từ khai phá dữ liệu)
- Knowledge (tri thức đạt được)

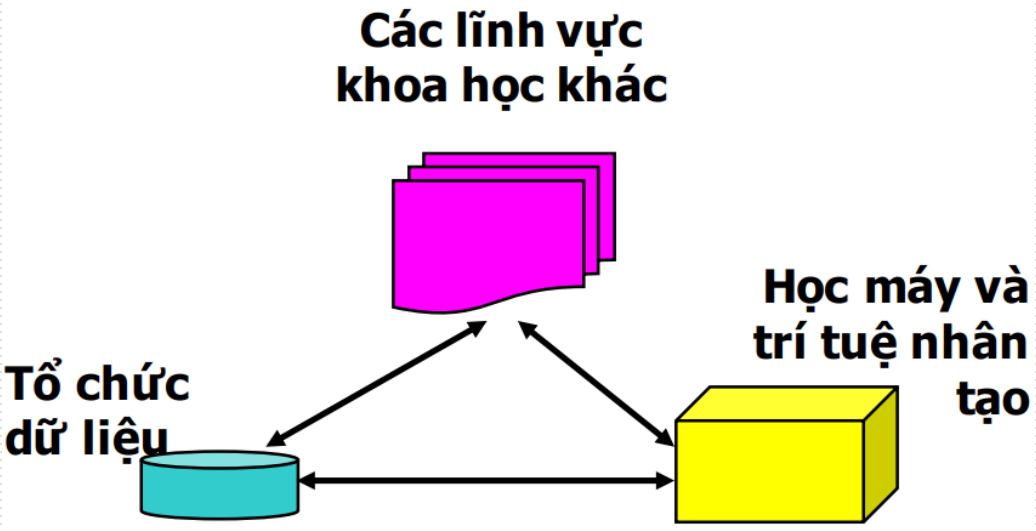
Tháp khám phá tri thức và ra quyết định

- Khám phá tri thức để hỗ trợ quyết định doanh nghiệp



Các kỹ thuật áp dụng trong Data Mining

- Khám phá tri thức trong CSDL là lĩnh vực liên ngành gồm:
  - Tổ chức dữ liệu
  - Học máy, trí tuệ nhân tạo
  - Các khoa học khác



- Phân loại các kỹ thuật dựa vào:
  - Quan điểm của học máy
  - Lớp các bài toán cần giải quyết

Phân loại dựa trên quan điểm của học máy

- 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 dữ liệu thành các lớp/cụm (clustering) dl tương tự nhau mà chưa biết trước các thông tin về lớp/tập các VDHL
- Học bán giám sát (Semi – Supervised learning)
  - Là quá trình phân chia một tập dữ liệu thành các lớp dựa trên một tập nhỏ các VDHL và một số các thông tin về một số nhãn lớp đã biết trước

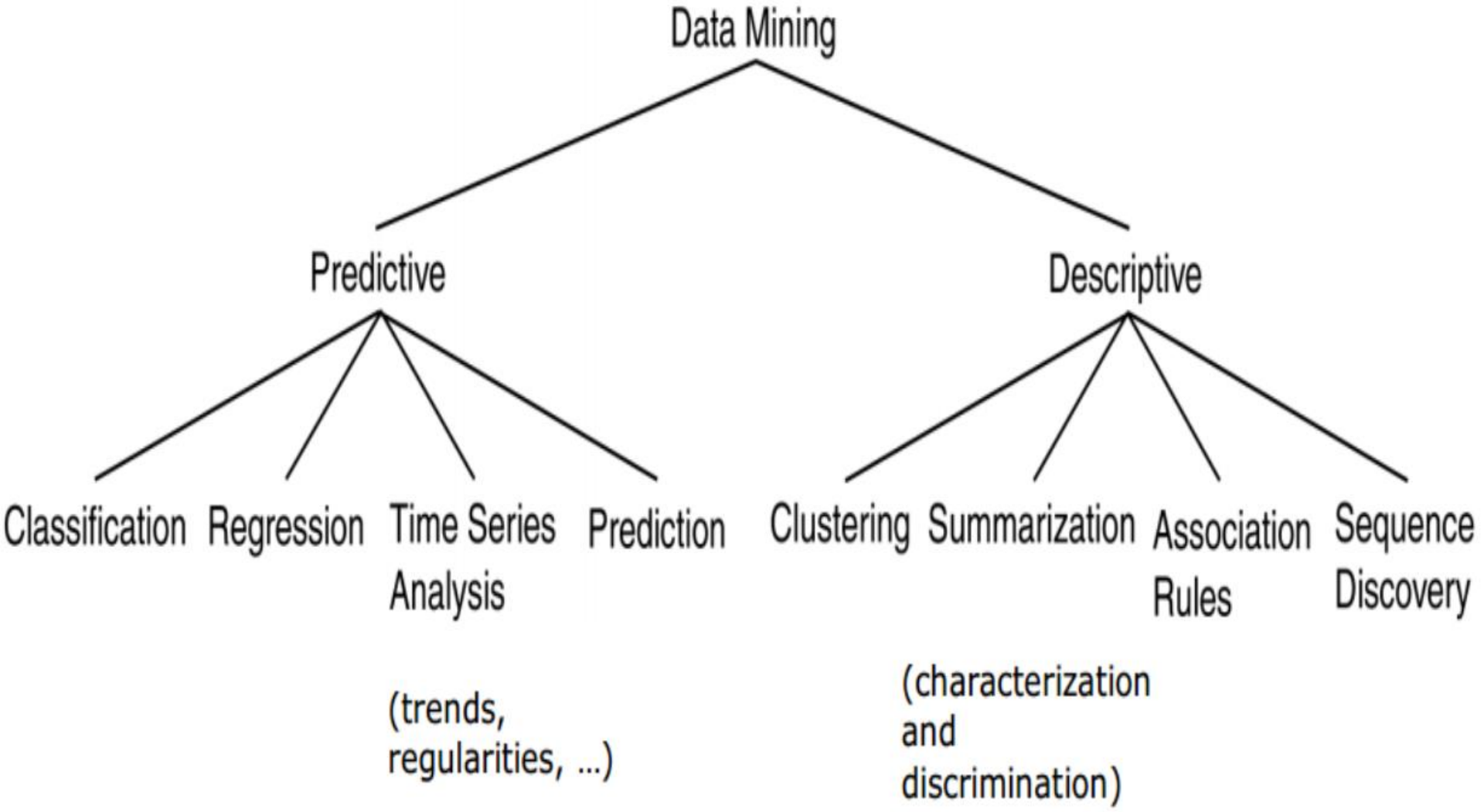
Phân loại dựa vào lớp các bài toán cần giải quyết

- Phân lớp và dự đoán (classification and prediction)
  - Xếp một đối tượng vào một trong những lớp đã biết trước
  - VD: phân lớp các bệnh nhân trong dữ liệu hồ sơ bệnh án
  - Thường sử dụng một số kỹ thuật của học máy như: cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network)...
  - Còn gọi là học có giám sát
- Luật kết hợp (association rules)
  - Là dạng luật biểu diễn tri thức ở dạng khá đơn giản
  - VD: "60 % nữ giới vào siêu thị nếu mua phấn thì có tới 80% trong số họ sẽ mua thêm son"
  - Ứng dụng nhiều trong kinh doanh, y học, tin-sinh, tài chính và thị trường chứng khoán, .v.v

Phân loại dựa vào lớp các bài toán cần giải quyết

- Phân tích chuỗi theo thời gian (sequential/ temporal patterns):
  - Tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian
  - Ứng dụng nhiều trong tài chính, thị trường chứng khoán vì nó có tính dự báo cao
- Phân cụm (clustering/ segmentation)
  - Xếp các đối tượng theo từng cụm dl tự nhiên
  - Còn gọi là học không có giám sát (unsupervised learning)
- Mô tả khái niệm (concept description and summarization):
  - Thiên về mô tả, tổng hợp và tóm tắt khái niệm
  - Ví dụ: tóm tắt văn bản

Cây phân cấp các kỹ thuật áp dụng KPDL



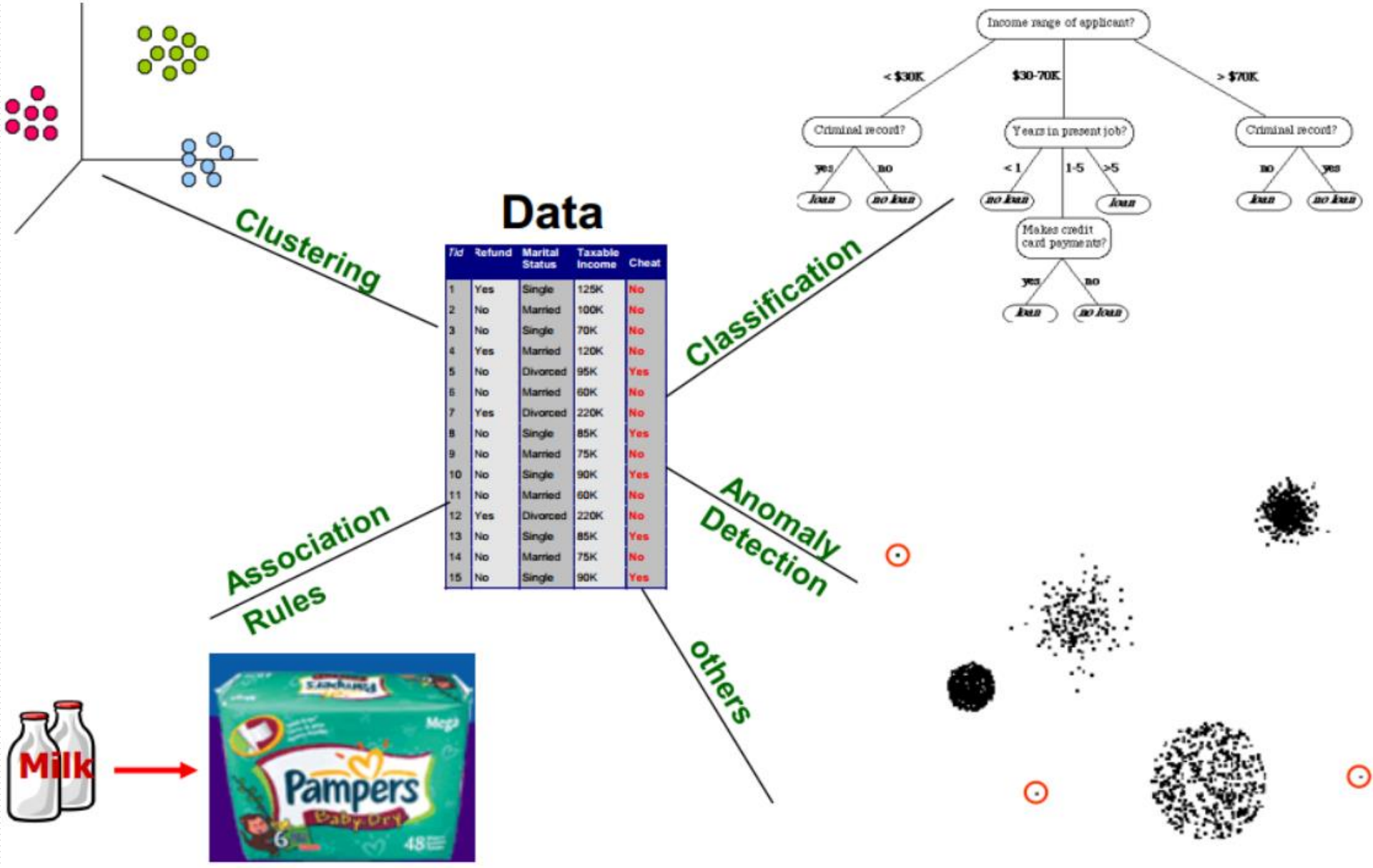
Các dạng dữ liệu có thể khai phá

- 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 tác vụ KPDL

- Khai phá mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa DL)
- Khai phá luật kết hợp/tương quan
- Phân loại dữ liệu
- Dự đoán
- Gom cụm dữ liệu
- Phân tích xu hướng
- Phân tích độ lệch và phần tử biên
- Phân tích độ tương tự

Các tác vụ KPDL



Các thành tố cơ bản đặc tả tác vụ KPDL

- Dữ liệu cụ thể sẽ được khai phá (task-relevant data)
- Loại tri thức sẽ đạt được (kind of knowledge)
- Tri thức nền (background knowledge)
- Các độ đo (interestingness measures)
- Các kỹ thuật biểu diễn tri thức/trực quan hóa mẫu (pattern visualization and knowledge presentation)

Dữ liệu cụ thể sẽ được khai phá

- Phần dữ liệu từ các nguồn dữ liệu quan tâm
- Tương ứng với các thuộc tính hay chiều dữ liệu được quan tâm
- Bao gồm:
  - Tên kho dữ liệu/CSDL
  - Các bảng hay các khối DL
  - Các điều kiện chọn DL
  - Các thuộc tính hay chiều DL
  - Các tiêu chí gom nhóm/phân cụm DL

Loại tri thức sẽ đạt được

- Bao gồm:
  - Đặc trưng hóa
  - Phân biệt hóa dữ liệu
  - Mô hình phân tích kết hợp hay tương quan
  - Mô hình phân lớp, dự báo, phân cụm, phân tích phần tử biên, phân tích tiến hóa…
- Tương ứng với tác vụ khai phá dữ liệu cụ thể sẽ được thực thi

Tri thức nền

- Tương ứng với lĩnh vực cụ thể sẽ được khai phá
- Hướng dẫn quá trình khám phá tri thức
  - Hỗ trợ khai phá dữ liệu ở nhiều mức trừu tượng khác nhau
- Đánh giá các mẫu được tìm thấy
- Bao gồm: các phân cấp ý niệm, niềm tin của người sử dụng về các mối quan hệ của dữ liệu

Các độ đo

- Thường đi kèm với các ngưỡng giá trị (threshold)
- Dẫn đường cho quá trình khai phá hoặc đánh giá các mẫu được tìm thấy
- Tương ứng với loại tri thức sẽ đạt được và do đó, tương ứng với tác vụ khai phá dữ liệu cụ thể sẽ được thực thi
- Kiểm tra:
  - Tính đơn giản (simplicity)
  - Tính chắc chắn (certainty)
  - Tính hữu dụng (utility)
  - Tính mới (novelty)

Các kỹ thuật biểu diễn tri thức/trực quan hóa mẫu

- Xác định dạng các mẫu/tri thức được tìm thấy để thể hiện đến người sử dụng
- Bao gồm:
  - Luật (rules)
  - Bảng (tables)
  - Báo cáo (reports)
  - Biểu đồ (charts)
  - Đồ thị (graphs)
  - Cây (trees)
  - Khối (cubes)

Bốn thành phần cơ bản của giải thuật KPDL

- 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)

Cấu trúc mẫu hay cấu trúc mô hình

- Mô hình là mô tả của tập dữ liệu, mang tính toàn cục ở mức cao
- Mẫu là đặc điểm (đặc trưng) của dữ liệu, mang tính cục bộ, chỉ cho một vài bản ghi/đối tượng hay vài biến.
- Cấu trúc biểu diễn các dạng chức năng chung với các thông số chưa được xác định trị
- Cấu trúc mô hình là một tóm tắt toàn cục về dữ liệu
  - Ví dụ: $Y = aX + b$ là một cấu trúc mô hình và $Y = 3X + 2$ là một mô hình cụ thể được định nghĩa dựa trên cấu trúc này
- Cấu trúc mẫu là những cấu trúc liên quan một phần tương đối nhỏ của dữ liệu hay của không gian dữ liệu
  - Ví dụ: $p(Y> y1|X>x1) = p1$ là một cấu trúc mẫu và $p(Y>5|X>10) = 0.5$ là một mẫu được xác định dựa trên cấu trúc này

Hàm tỉ số

- Là hàm xác định một cấu trúc mô hình/mẫu đáp ứng tập dữ liệu đã cho tốt ở mức độ nào đó
- Cho biết liệu một mô hình có tốt hơn các mô hình khác hay không
- Không nên phụ thuộc nhiều vào tập dữ liệu, không nên chiếm nhiều thời gian tính toán
- Một vài hàm tỉ số thông dụng:
  - Likelihood
  - Sum of Squared errors
  - Misclassification rate...

Phương pháp tìm kiếm và tối ưu hóa

- Mục tiêu của phương pháp tìm kiếm và tối ưu hóa là xác định cấu trúc và giá trị các thông số đáp ứng tốt nhất hàm tỉ số từ dữ liệu sẵn có
- Tìm kiếm các mẫu và mô hình
- Không gian trạng thái: tập rời rạc các trạng thái
  - Bài toán tìm kiếm:
    - Bắt đầu tại một node (trang thái) cụ thể
    - Di chuyển qua không gian trạng thái để tìm thấy node tương ứng với trạng thái đáp ứng tốt nhất hàm tỉ số
  - Phương pháp tìm kiếm: chiến lược tham lam, có dùng heuristics, chiến lược nhánh-cận
- Tối ưu hóa thông số

Chiến lược quản lý dữ liệu

- Dữ liệu được khai phá
  - Ít, toàn bộ được xử lý đồng thời trong bộ nhớ chính
  - Nhiều, trên đĩa, một phần được xử lý đồng thời trong bộ nhớ chính
- Chiến lược quản lý dữ liệu hỗ trợ cách dữ liệu được lưu trữ, đánh chỉ mục, và truy xuất
  - Giải thuật khai phá dữ liệu hiệu quả (efficiency) và có tính co giãn (scalability) với dữ liệu được khai phá
  - Công nghệ cơ sở dữ liệu

Quy trình khai phá dữ liệu

- Khái niệm quy trình KPDL
- Sự cần thiết của quy trình KPDL
- Quy trình KPDL CRISP-DM

Khái niệm quy trình KPDL

- Là một chuỗi lặp (iterative) và tương tác (interactive) gồm các bước (giai đoạn) bắt đầu với dữ liệu thô (raw data) và kết thúc với tri thức (knowledge of interest) đáp ứng được sự quan tâm của người sử dụng
- 2 quy trình KPDL:
  - CRISP-DM (Cross Industry Standard Process for Data Mining, www.crisp-dm.org)
  - SEMMA (Sample, Explore, Modify, Model, Assess) at the SAS Institute

Sự cần thiết của quy trình KPDL

- Cách thức tiến hành (hoạch định và quản lý) dự án khai phá dữ liệu có hệ thống
- Đảm bảo nỗ lực dành cho một dự án khai phá dữ liệu được tối ưu hóa
- Việc đánh giá và cập nhật các mô hình trong dự án được diễn ra liên tục

Quy trình KPDL CRISP-DM

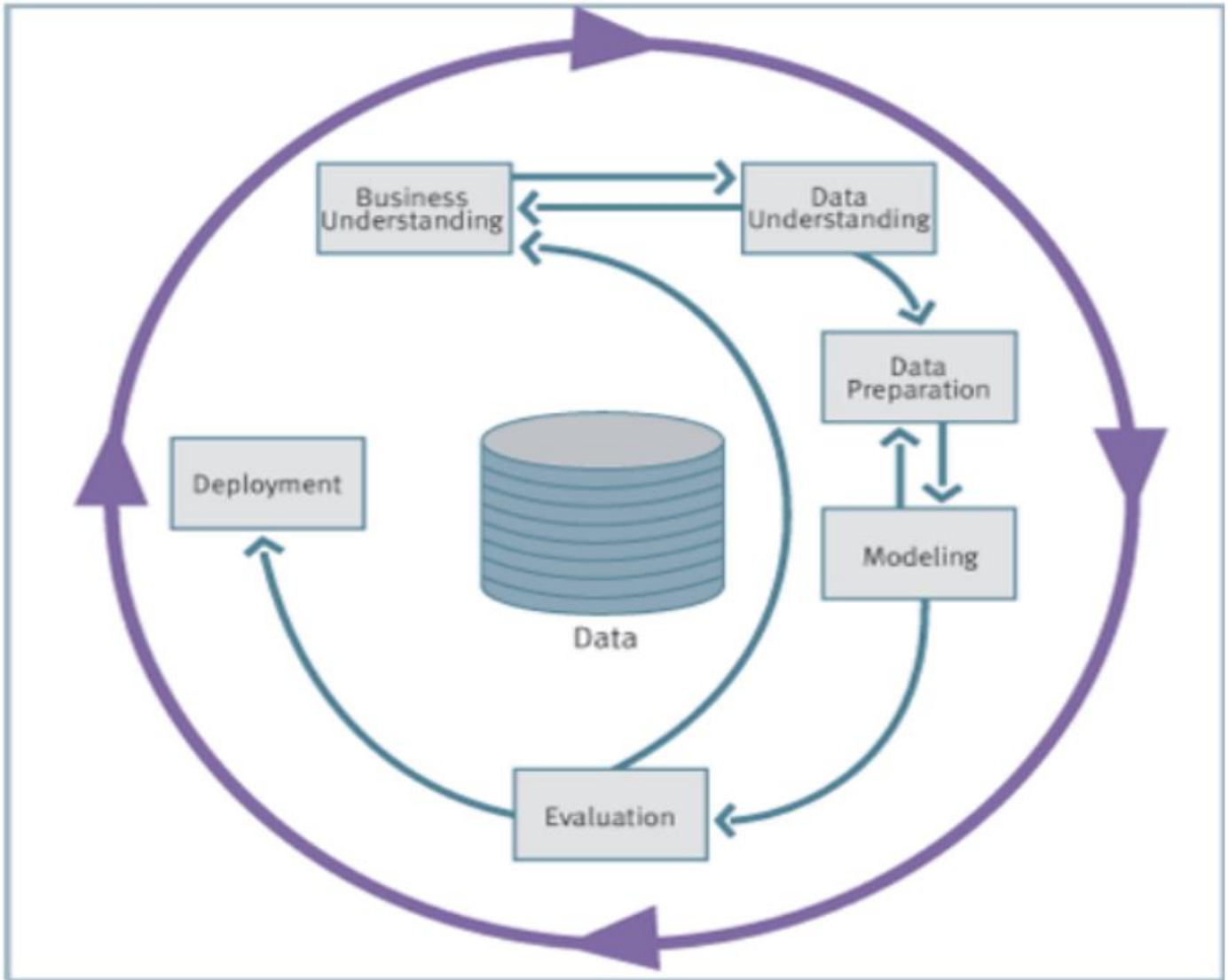
- Chuẩn quy trình công nghiệp
- Sơ đồ quy trình CRISP-DM
- Các giai đoạn của quy trình CRISP-DM

Chuẩn quy trình công nghiệp

- Được khởi xướng từ 09/1996 và được hỗ trợ bởi hơn 200 thành viên
- Chuẩn mở
- Hỗ trợ công nghiệp/ứng dụng và công cụ khai phá dữ liệu hiện có
- Tập trung vào các vấn đề nghiệp vụ cũng như phân tích kỹ thuật
- Tạo ra một khung thức hướng dẫn qui trình khai phá dữ liệu
- Có nên tảng kinh nghiệm từ các lĩnh vực ứng dụng

Sơ đồ quy trình CRISP-DM

- Là một quy trình lặp, có khả năng quay lui (backtracking)



Các giai đoạn của quy trình CRISP-DM

- Gồm 6 giai đoạn:
  - Tìm hiểu nghiệp vụ (Business understanding)
  - Tìm hiểu dữ liệu (Data understanding)
  - Chuẩn bị dữ liệu (Data preparation)
  - Mô hình hoá (Modeling)
  - Đánh giá (Evaluation)
  - Triển khai (Deployment)

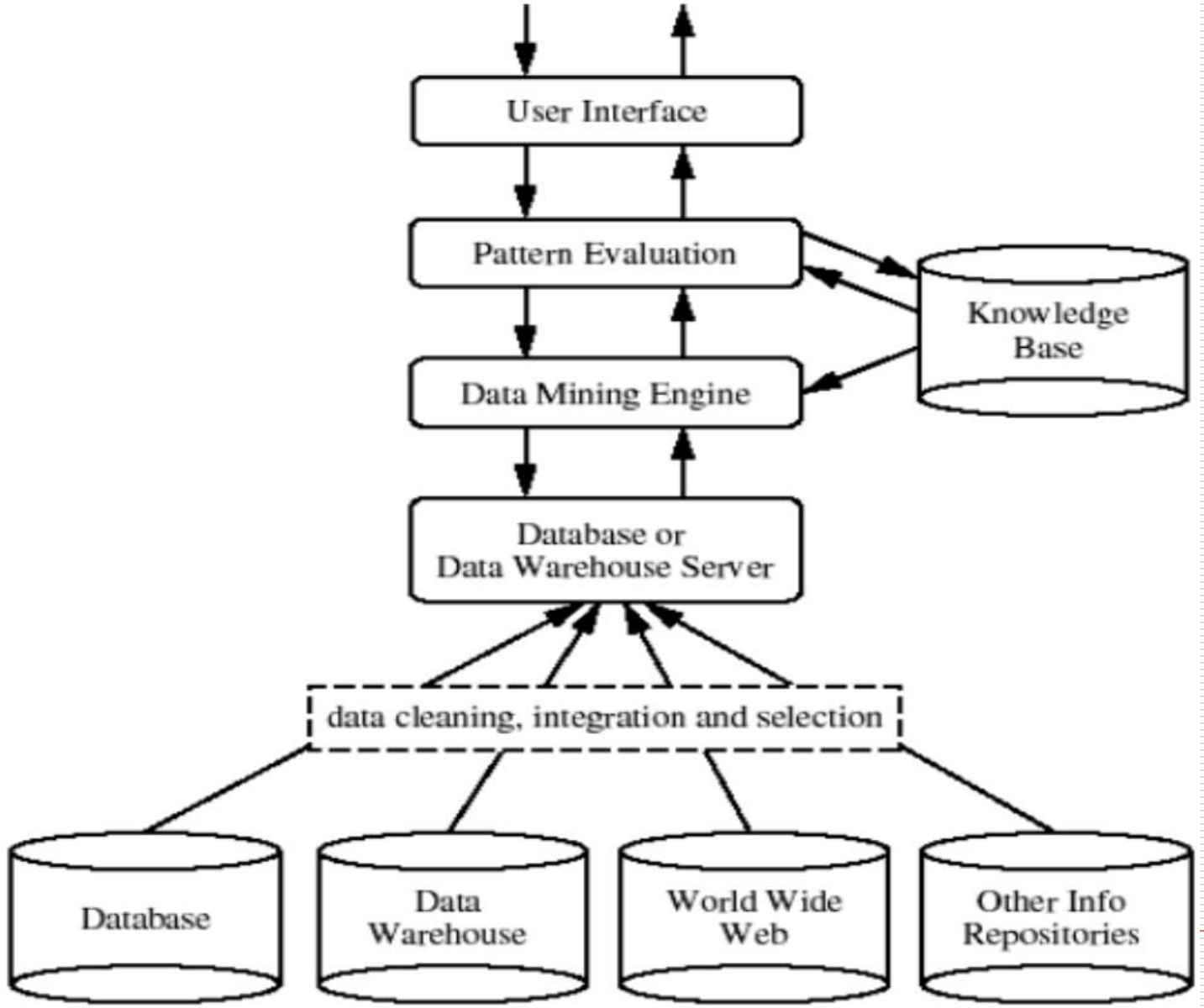
Hệ thống khai phá dữ liệu

- Khái niệm hệ thống KPDL
- Kiến trúc hệ thống KPDL
- Các thành phần chính có thể có
- Các đặc điểm dùng để khảo sát hệ thống KPDL
- Một số hệ thống khai phá dữ liệu

Khái niệm hệ thống KPDL

- Hệ thống khai phá dữ liệu được phát triển dựa trên khái niệm rộng của khai phá dữ liệu
- Khai phá dữ liệu là một quá trình khám phá tri thức được quan tâm từ lượng lớn dữ liệu trong các cơ sở dữ liệu, kho dữ liệu, hay các kho thông tin khác.
- Các thành phần chính có thể có
  - Database, data warehouse, World Wide Web, và information repositories
  - Database hay data warehouse server
  - Knowledge base
  - Data mining engine
  - Pattern evaluation module
  - User interface

Kiến trúc hệ thống KPDL



Các thành phần chính có thể có

- Database, data warehouse, World Wide Web, và information repositories
  - Thành phần này là các nguồn dữ liệu/thông tin sẽ được khai phá
  - Trong những tình huống cụ thể, thành phần này là nguôn nhập (input) của các kỹ thuật tích hợp và làm sạch dữ liệu
- Database hay data warehouse server
  - Thành phần chịu trách nhiệm chuẩn bị dữ liệu thích hợp cho các yêu cầu khai phá dữ liệu
- Knowledge base
  - Thành phần chứa tri thức miền, được dùng để hướng dân quá trình tìm kiếm, đánh giá các mâu kết quả được tìm thấy
  - Tri thức miền có thể là các phân cấp khái niệm, niêm tin của người sử dụng, các ràng buộc hay các ngưỡng giá trị, siêu dữ liệu...
- Data mining engine
  - Thành phần chứa các khối chức năng thực hiện các tác vụ khai phá dữ liệu

Các thành phần chính có thể có

- Pattern evaluation module
  - Thành phần này làm việc với các độ đo (và các ngưỡng giá trị) hỗ trợ tìm kiếm và đánh giả các mâu sao cho các mầu được tìm thấy là những mẫu được quan tâm bởi người sử dụng
  - Thành phần này có thể được tích hợp vào thành phần Data mining engine
- User interface
  - Thành phần hỗ trợ sự tương tác giữa người sử dụng và hệ thống khai phá dữ liệu
  - Người sử dụng có thể chỉ định câu truy vấn hay tác vụ khai phá dữ liệu
  - Người sử dụng có thể được cung cấp thông tin hỗ trợ việc tìm kiếm, thực hiện khai phá dữ liệu sâu hơn thông qua các kết quả khai phá trung gian
  - Người sử dụng cũng có thể xem các lược đồ cơ sở dữ liệu/kho dữ liệu, các cấu trúc dữ liệu; đánh giá các mẫu khai phá được; trực quan hóa các mẫu này ở các dạng khác nhau

Đặc điểm dùng để khảo sát hệ thống KPDL

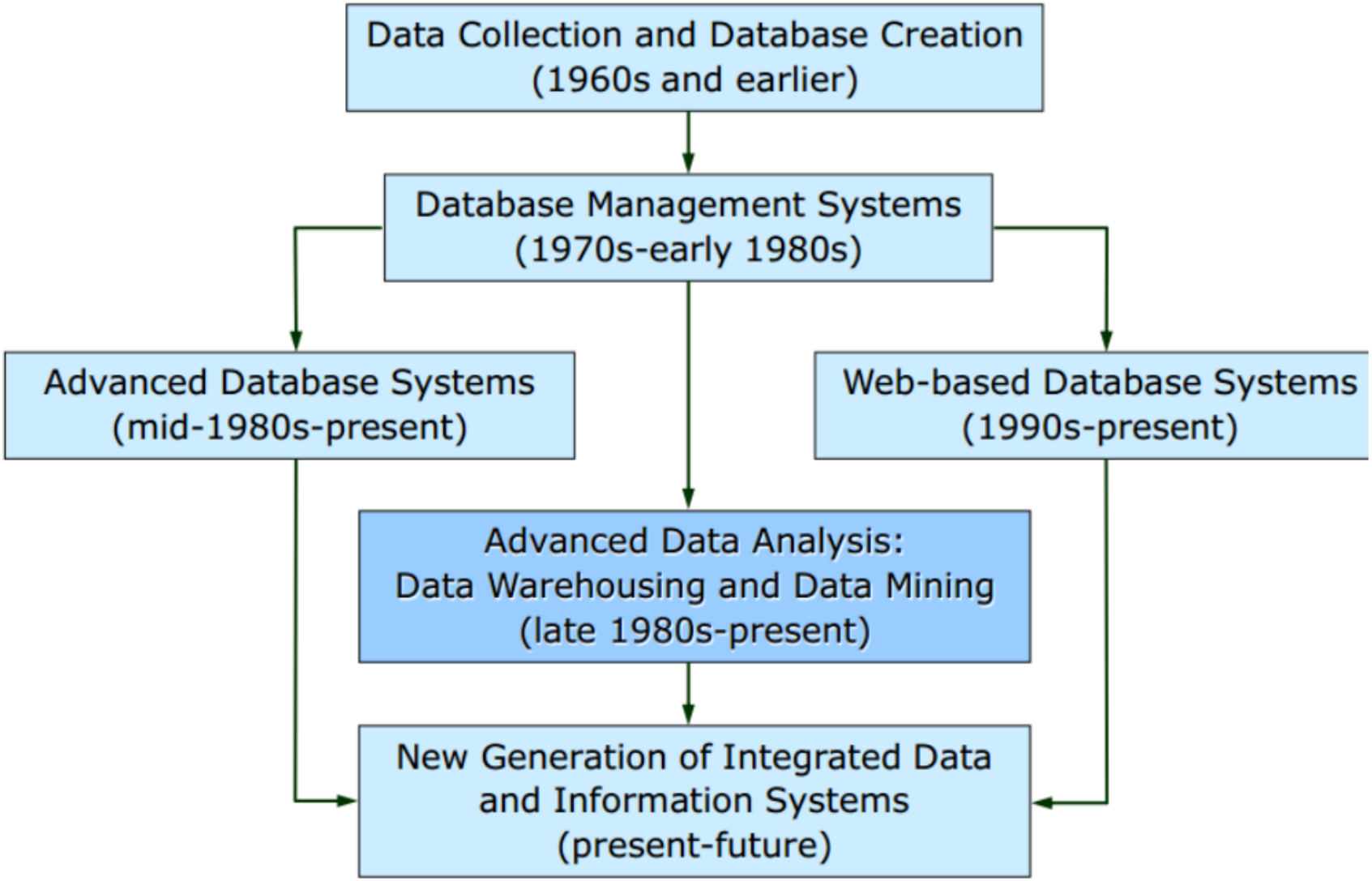
- Kiểu dữ liệu
- Các vấn đê hệ thống
- Nguồn dữ liệu
- Các tác vụ và phương pháp luận khai phá dữ liệu
- Vấn đề gắn kết với các hệ thống kho dữ liệu/cơ sở dữ liệu
- Khả năng co giãn dữ liệu
- Các công cụ trực quan hóa
- Ngôn ngữ truy vấn khai phá dữ liệu và giao diện đồ họa cho người dùng

Một số hệ thống khai phá dữ liệu

- Intelligent Miner (IBM)
- Microsoft data mining tools (Microsoft SQL Server 2000/2005/2008)
- Oracle Data Mining (Oracle 9i/10g/11g)
- Enterprise Miner (SAS Institute)
- Weka (the University of Waikato, New Zealand, WWW.Cs.Waikato.ac.nz/mil/weka)

Ý nghĩa và vai trò của KPDL

- Theo quy trình tiến hóa của công nghệ hệ CSDL



Ý nghĩa và vai trò của KPDL

- Công nghệ hiện đại trong lĩnh vực quản lý thông tin
  - Hiện diện khắp nơi (ubiquitous) và có tính ẩn (invisible) trong nhiêu khía cạnh của đời sống hằng ngày
    - Làm việc, mua sắm, tìm kiếm thông tin, nghỉ ngơi, ...
  - Được áp dụng trong nhiêu 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 của KPDL

- Là lĩnh vực được quan tâm và ứng dụng rộng rãi:
  - Phân tích dữ liệu và hỗ trợ quyết định
  - Điêu trị y học
  - Text mining & W/eb mining
  - Tin-sinh (blo-informatics)
  - Tài chính va thị trường chứng khoán
  - Bảo hiểm (insurance), .V.V
- 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 của KPDL

- Trong kinh doanh:
  - Các ứng dụng trong tiếp thị, tài chính (đặc biệt là đầu tư), phát hiện gian lận, sản xuất, viên thông và các Internet agent (tác tử)
- Tiếp thị:
  - Ứng dụng trong hệ thống CSDL tiếp thị, phần tích các dữ liệu khách hàng để phân loại các nhóm khách hàng khác nhau và dự báo về sở thích của họ
- 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

Ứng dụng của KPDL

- Đầu tư:
  - LBS Capital Management dùng để quản lý danh mục vốn đầu tư
- Sản xuất:
  - Hệ thống xử lý sự cố CASSIOPEE được sử dụng để phát hiện và tiên đoán các sự cố của máy bay Boeing
- Viễn thông:
  - Hệ thống TASA dùng để phân tích các lỗi báo động trên đường truyền
- Các tác tử thông minh:
  - Dùng để duyệt qua một môi trường nhiều thông tin như Internet


THE END