Khoa Khoa Học & Kỹ Thuật Máy Tính
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh

Chương 1: Tổng quan về khai phá dữ liệu

Học kỳ 2 – 2019-2020

Nội dung

- 1.0. TÌNH HUỐNG
- 1.1. Quá trình khám phá tri thức
- 1.2. Các khái niệm
- 1.3. Ý nghĩa và vai trò của khai phá dữ liệu
- 1.4. Ứng dụng của khai phá dữ liệu
- 1.5. Tóm tắt

Tài liệu tham khảo

- **[1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, 2006.**
- **[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.**
- **[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.**
- [4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.
- [5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.
- [6] Oracle, "Data Mining Concepts", B28129-01, 2008.
- [7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.

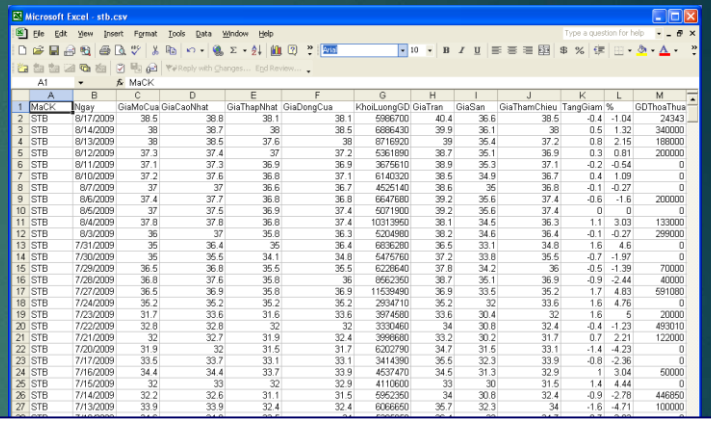
1.0. Tình huống 1

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

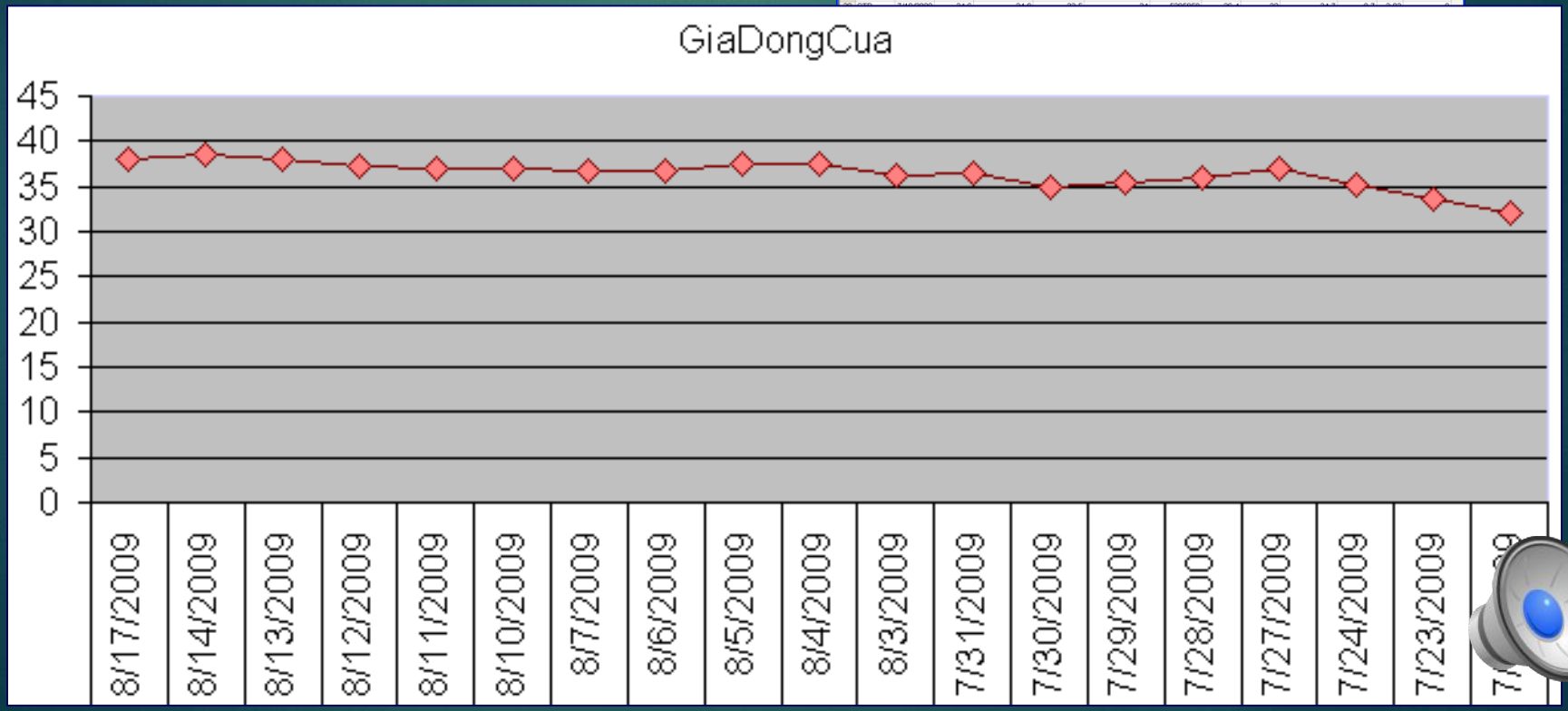
Ông A (Tid = 100) có khả năng trốn thuế???

1.0. Tình huống 2

Ngày mai cổ phiếu STB sẽ tăng???



	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
8	STB	8/7/2009	37	37	36.6	36.7	4525140	38.6	35	36.8	-0.1	-0.27	0
9	STB	8/6/2009	37.4	37.7	36.8	36.8	6647680	39.2	35.6	37.4	-0.6	-1.6	200000
10	STB	8/5/2009	37	37.5	36.9	37.4	5071900	39.2	35.6	37.4	0	0	0
11	STB	8/4/2009	37.8	37.8	36.8	37.4	10313950	38.1	34.5	36.3	1.1	3.03	133000
12	STB	8/3/2009	36	37	35.8	36.3	5204980	38.2	34.6	36.4	-0.1	-0.27	299000
13	STB	7/31/2009	35	36.4	35	36.4	6836280	36.5	33.1	34.8	1.6	4.6	0
14	STB	7/30/2009	35	35.5	34.1	34.8	5475760	37.2	33.8	35.5	-0.7	-1.97	0
15	STB	7/29/2009	36.5	36.8	35.5	35.5	6226640	37.8	34.2	36	-0.5	-1.39	70000
16	STB	7/28/2009	36.8	37.6	35.8	36	8662360	38.7	35.1	36.9	-0.9	-2.44	40000
17	STB	7/27/2009	36.5	36.9	35.8	36.9	11539490	36.9	33.5	35.2	1.7	4.83	591080
18	STB	7/24/2009	35.2	35.2	35.2	35.2	2934710	35.2	32	33.6	1.6	4.76	0
19	STB	7/23/2009	31.7	33.6	31.6	33.6	3974580	33.6	30.4	32	1.6	5	20000
20	STB	7/22/2009	32.8	32.8	32	32	3330460	34	30.8	32.4	-0.4	-1.23	493010
21	STB	7/21/2009	32	32.7	31.9	32.4	3698680	33.2	30.2	31.7	0.7	2.21	122000
22	STB	7/20/2009	31.9	32	31.5	31.7	6202790	34.7	31.5	33.1	-1.4	-4.23	0
23	STB	7/17/2009	33.5	33.7	33.1	33.1	3414390	35.5	32.3	33.9	-0.8	-2.36	0
24	STB	7/16/2009	34.4	34.4	33.7	33.9	4537470	34.5	31.3	32.9	1	3.04	50000
25	STB	7/15/2009	32	33	32	32.9	4110600	33	30	31.5	1.4	4.44	0
26	STB	7/14/2009	32.2	32.6	31.1	31.5	5952350	34	30.8	32.4	-0.9	-2.78	446850
27	STB	7/13/2009	33.9	33.9	32.4	32.4	6066650	35.7	32.3	34	-1.6	-4.71	100000

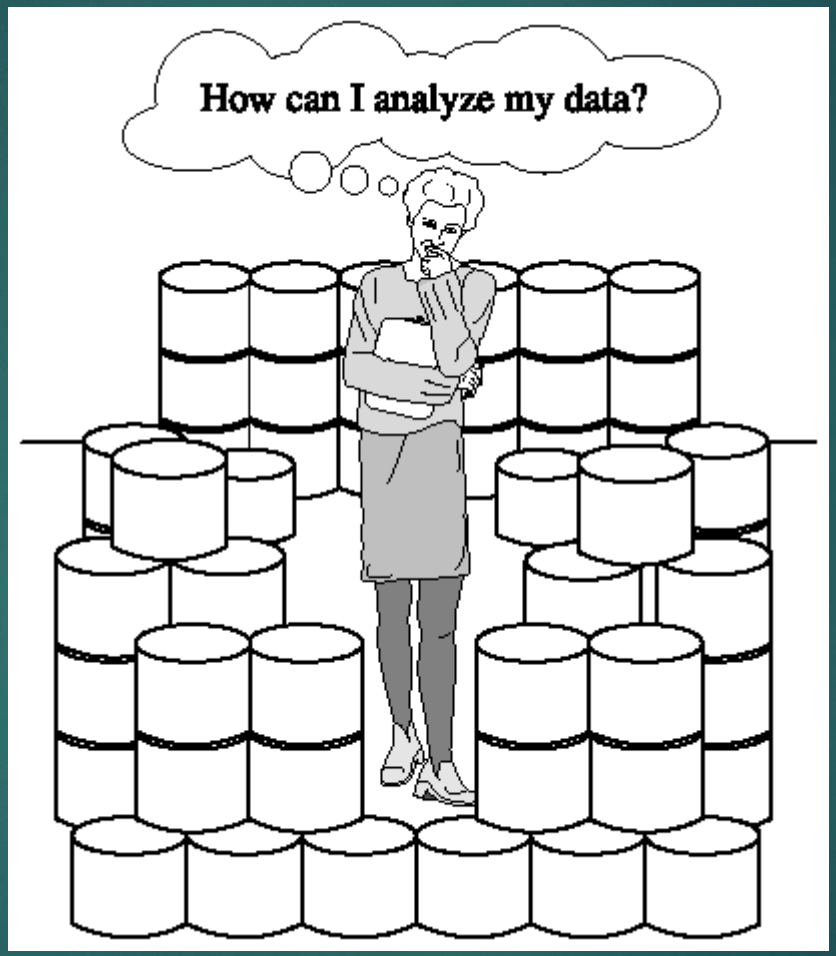


1.0. Tình huống 3

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

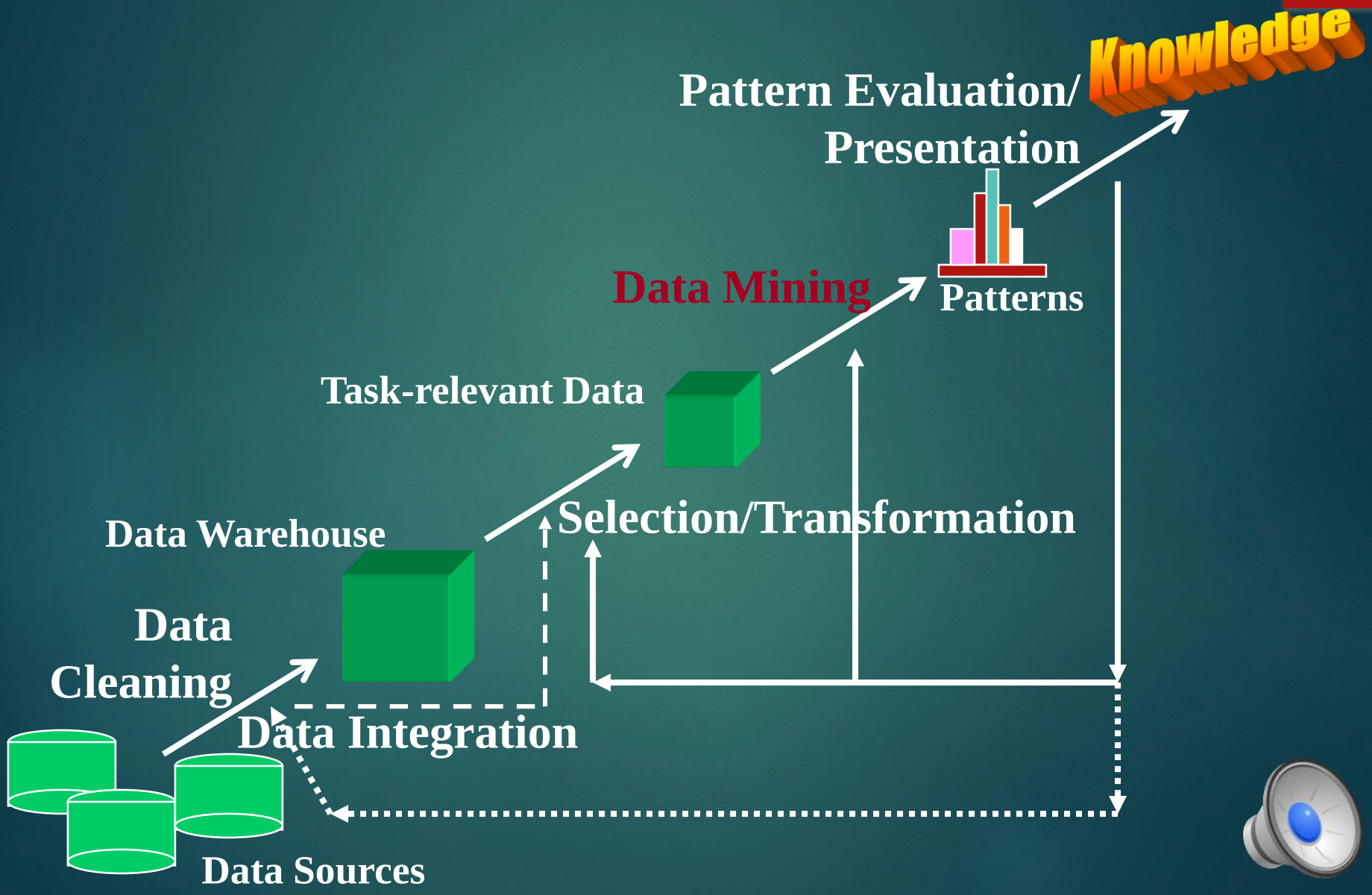
1.0. Tình huống …



We are data rich, but information poor.

“Necessity is the mother of invention”. - Plato

1.1. Quá trình khám phá tri thức



1.1. Quá trình khám phá tri thức

- "Knowledge discovery in **databases** is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable **patterns** in data."
  - Frawley, W. J et al. (1991). Knowledge discovery in databases: an overview.
- "Knowledge discovery from **databases** is the process of using the database along with any required selection, preprocessing, sub-sampling, and transformations of it; to apply data mining methods (algorithms) to enumerate **patterns** from it; and to evaluate the products of data mining to identify the subset of the enumerated patterns deemed **knowledge**."
  - Fayyad, U.M et al. (1996). Advances in Knowledge Discovery and Data Mining. MIT Press.

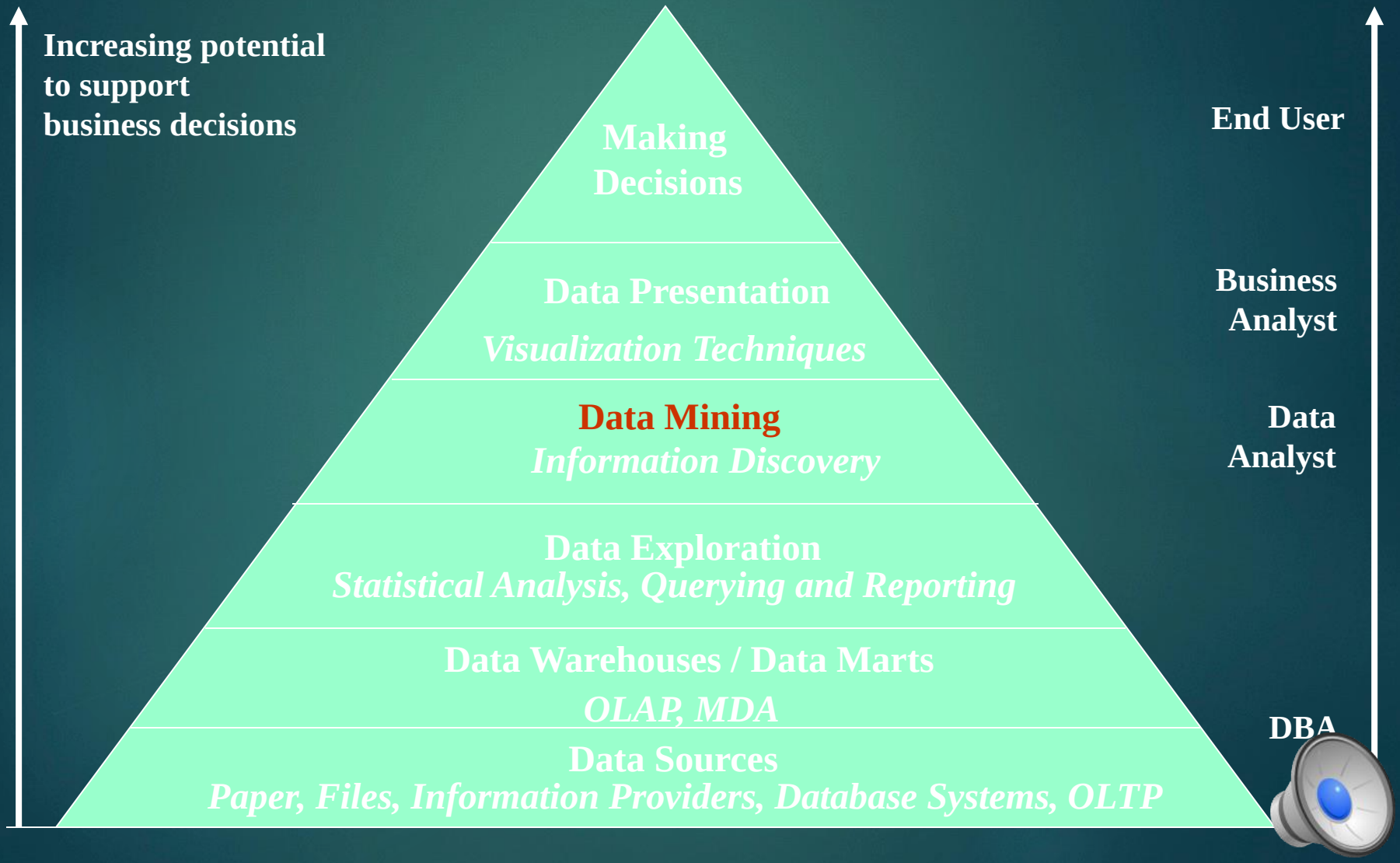
1.1. Quá trình khám phá tri thức

- 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)

1.1. Quá trình khám phá tri thức

- Quá trình khám phá tri thức là một chuỗi lặp gồm các bước được thực thi với:
  - 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)

1.1. Quá trình khám phá tri thức



1.2. Các khái niệm

- 1.2.1. Khai phá dữ liệu (data mining)
- 1.2.2. Các tác vụ khai phá dữ liệu (data mining tasks/functions)
- 1.2.3. Các quy trình khai phá dữ liệu (data mining processes)
- 1.2.4. Các hệ thống khai phá dữ liệu (data mining systems)

1.2.1. Khai phá dữ liệu

- Khai phá dữ liệu
  - một quá trình trích xuất tri thức từ lượng lớn dữ liệu
    - "extracting or mining knowledge from large amounts of data"
    - "knowledge mining from data"
  - một quá trình không dễ trích xuất thông tin ẩn, hữu ích, chưa được biết trước từ dữ liệu
    - "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data"
- 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

1.2.1. Khai phá dữ liệu

- 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 (flat files)
    - Các cơ sở dữ liệu quan hệ (relational databases) hay quan hệ đối tượng (object relational databases)
    - Các cơ sở dữ liệu giao tác (transactional databases) hay kho dữ liệu (data warehouses)
    - Các cơ sở dữ liệu hướng ứng dụng: cơ sở dữ liệu không gian (spatial databases), cơ sở dữ liệu thời gian (temporal databases), cơ sở dữ liệu không thời gian (spatio-temporal databases), cơ sở dữ liệu chuỗi thời gian (time series databases), cơ sở dữ liệu văn bản (text databases), cơ sở dữ liệu đa phương tiện (multimedia databases), …
    - Các kho thông tin: the World Wide Web, …
  - Dữ liệu tạm thời: các dòng dữ liệu (data streams)

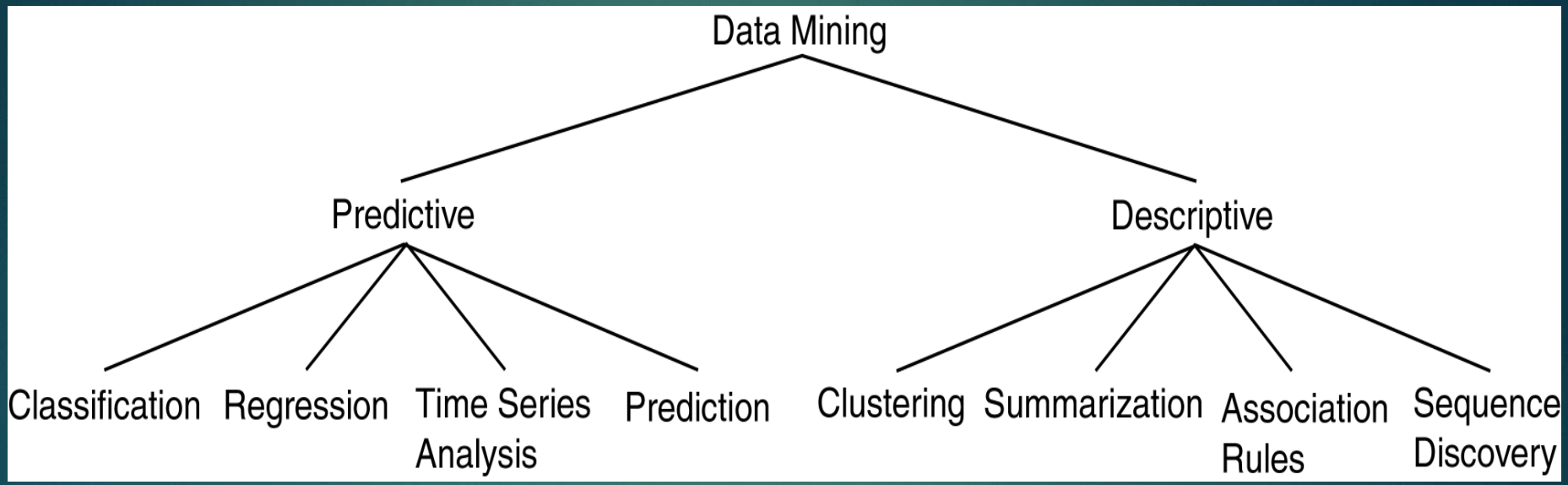
1.2.1. Khai phá dữ liệu

- 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
  - …

1.2.1. Khai phá dữ liệu

- 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 dữ liệu 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 …

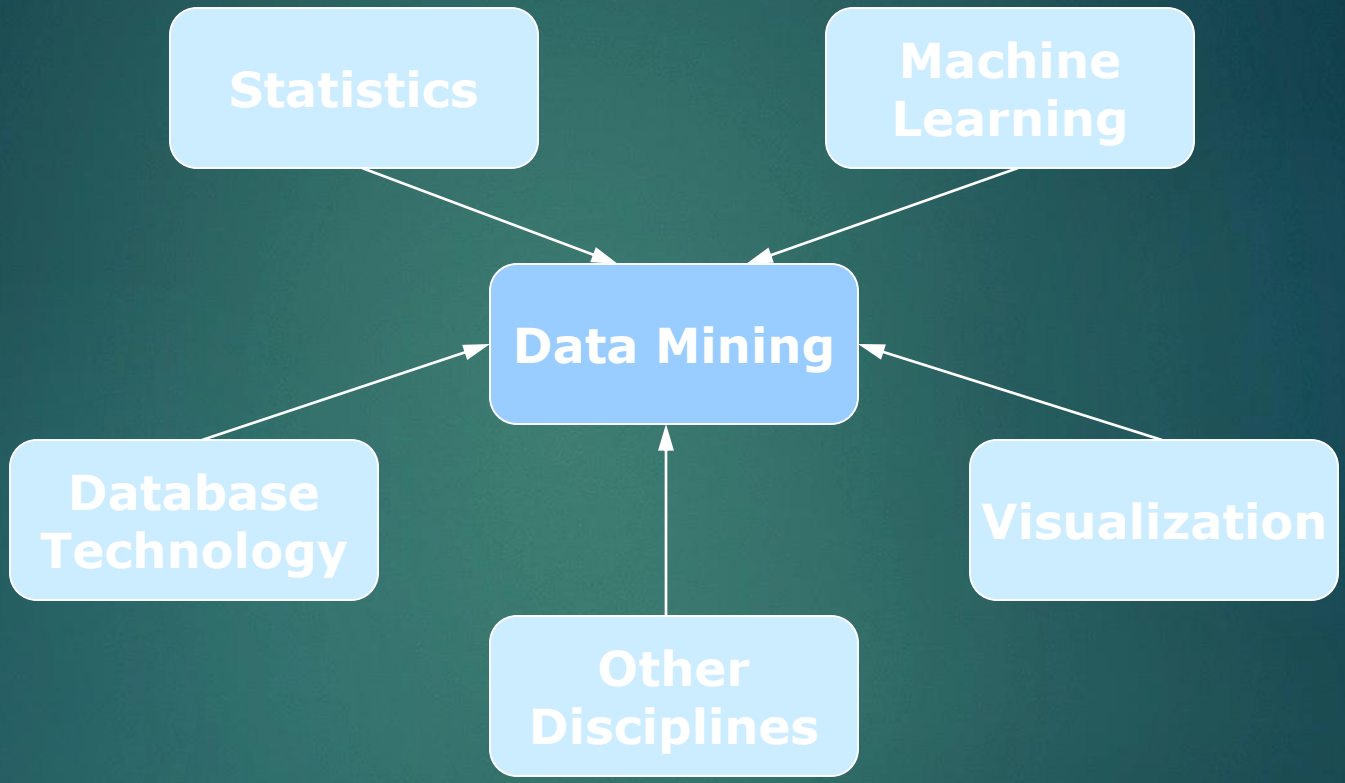
1.2.1. Khai phá dữ liệu



(trends, regularities, …)

(characterization and discrimination)

1.2.1. Khai phá dữ liệu



- Khai phá dữ liệu là một lĩnh vực liên ngành, nơi hội tụ của nhiều học thuyết và công nghệ.
  - "Data mining as a confluence of multiple disciplines"

1.2.1. Khai phá dữ liệu

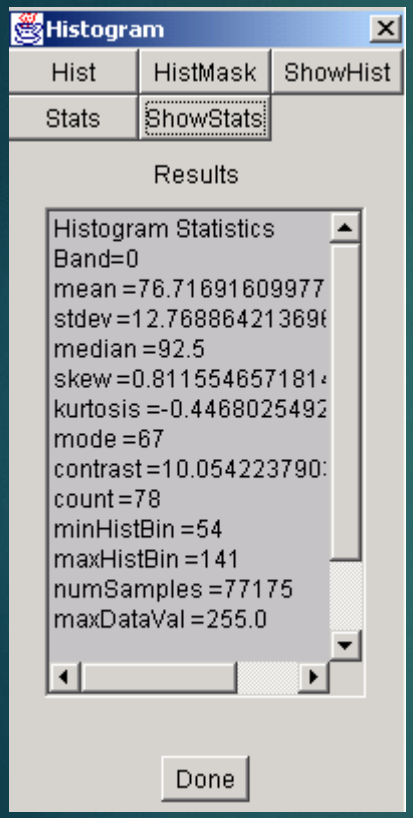
- Khai phá dữ liệu và công nghệ cơ sở dữ liệu
  - Khả năng đóng góp của công nghệ cơ sở dữ liệu
    - Công nghệ cơ sở dữ liệu cho việc quản lý dữ liệu được khai phá.
      - Dữ liệu rất lớn, có thể vượt quá khả năng của bộ nhớ chính (main memory).
      - Dữ liệu được thu thập theo thời gian.
    - Các hệ cơ sở dữ liệu có khả năng xử lý hiệu quả lượng lớn dữ liệu với các cơ chế phân trang (paging) và hoán chuyển (swapping) dữ liệu vào/ra bộ nhớ chính.
    - Các hệ cơ sở dữ liệu hiện đại có khả năng xử lý nhiều loại dữ liệu phức tạp (spatial, temporal, spatiotemporal, multimedia, text, Web, …).
    - Các chức năng khác (xử lý đồng thời, bảo mật, hiệu năng, tối ưu hóa, …) của các hệ cơ sở dữ liệu đã được phát triển tốt.

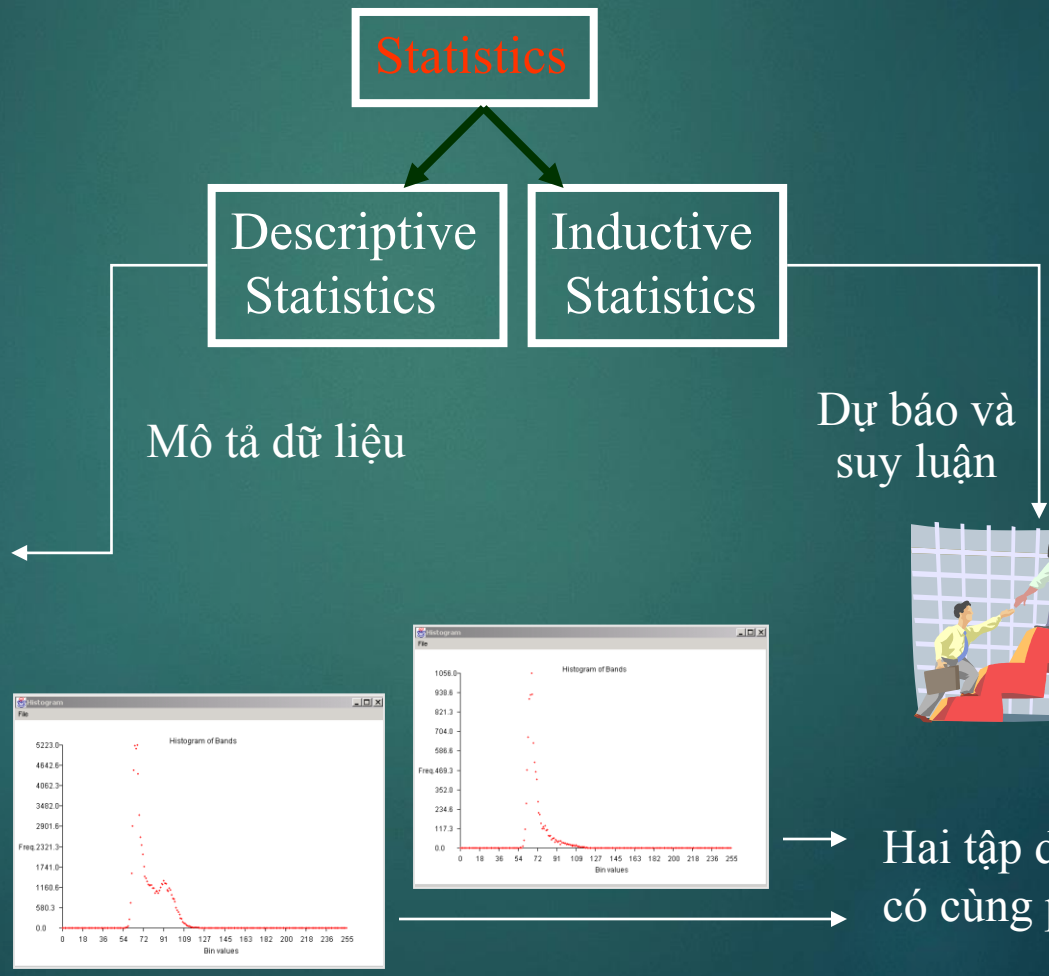
1.2.1. Khai phá dữ liệu

- Khai phá dữ liệu và công nghệ cơ sở dữ liệu
  - Thực trạng đóng góp của công nghệ cơ sở dữ liệu
    - Các hệ quản trị cơ sở dữ liệu (DBMS) hỗ trợ khai phá dữ liệu.
      - Oracle Data Mining (Oracle 9i, 10g, 11g)
      - Các công cụ khai phá dữ liệu của Microsoft (MS SQL Server 2000, 2005, 2008)
      - Intelligent Miner (IBM)

1.2.1. Khai phá dữ liệu

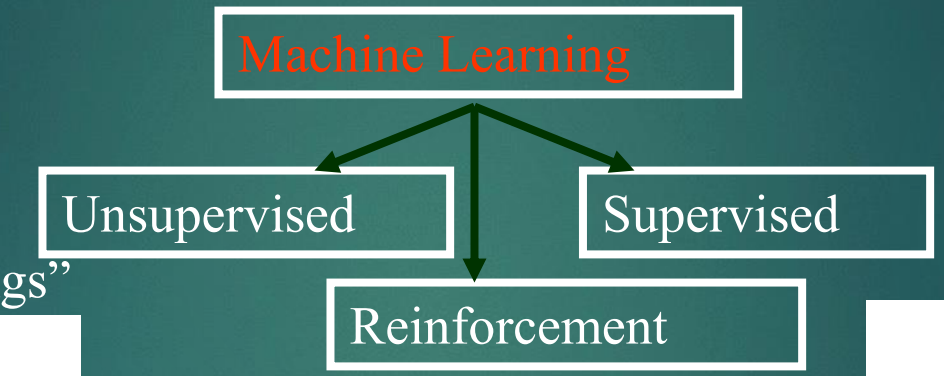
- Khai phá dữ liệu và lý thuyết thống kê



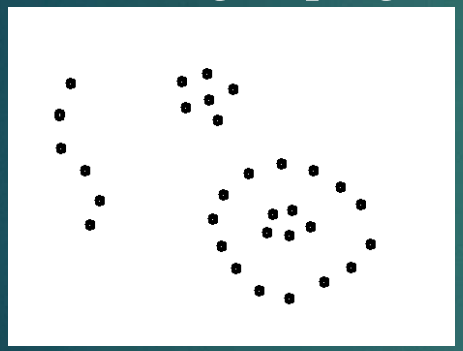


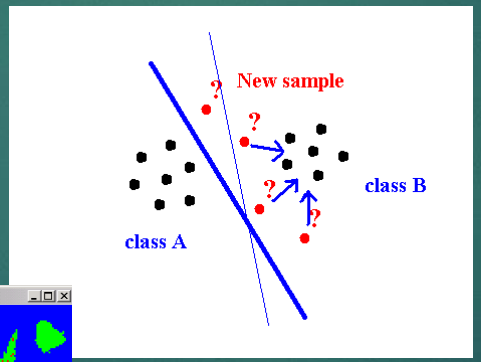
1.2.1. Khai phá dữ liệu

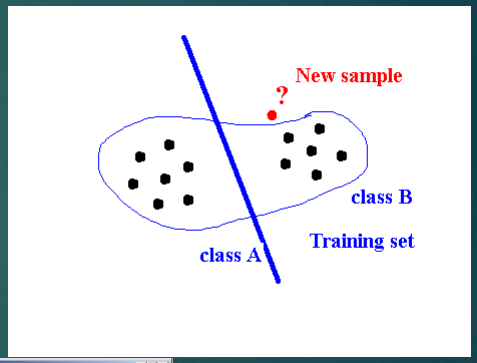
- Khai phá dữ liệu và học máy

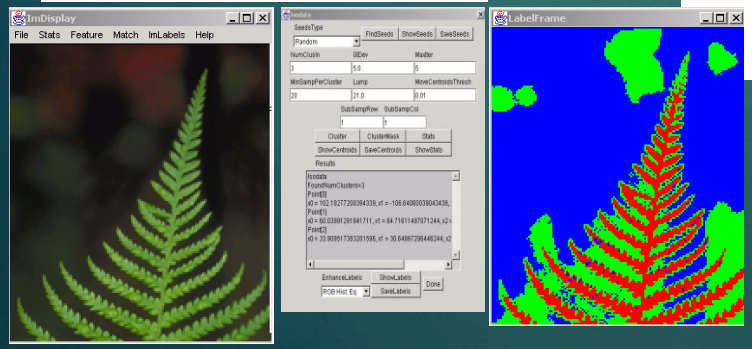


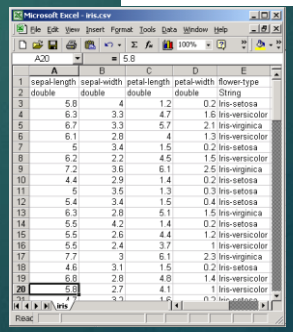
"Natural groupings"





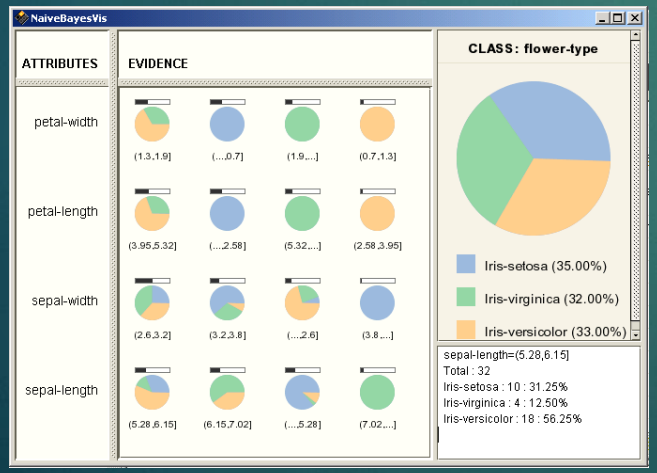




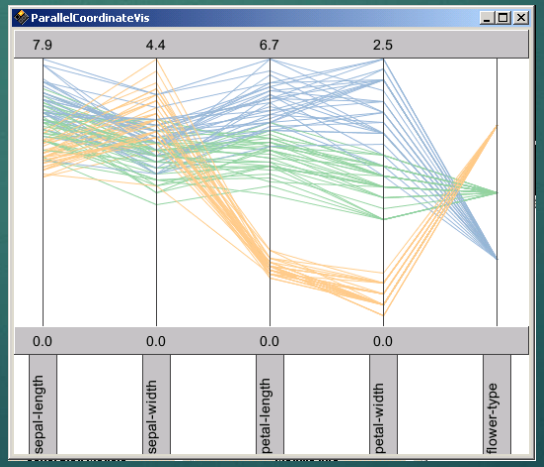


1.2.1. Khai phá dữ liệu

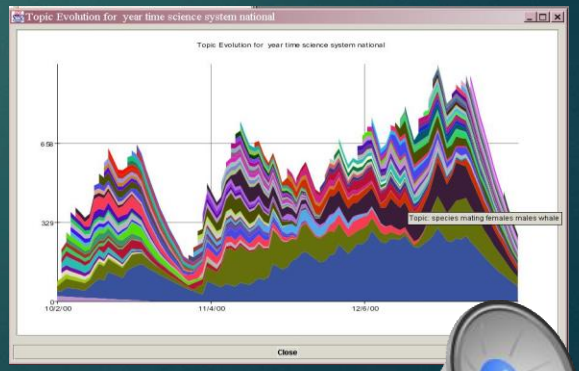
- Khai phá dữ liệu và trực quan hóa
  - Dữ liệu: 3D cubes,distribution charts, curves, surfaces, link graphs, image frames and movies, parallel coordinates
  - Kết quả (tri thức): pie charts, scatter plots, box plots, association rules, parallel coordinates, dendograms, temporal evolution



Pie chart



Parallel coordinates

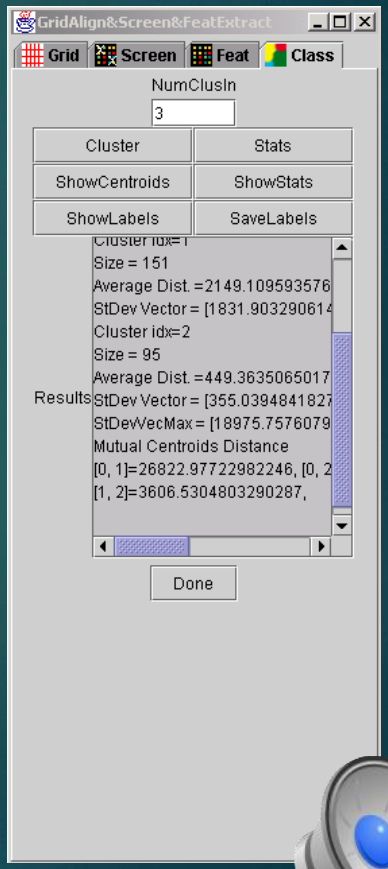


Temporal evolu

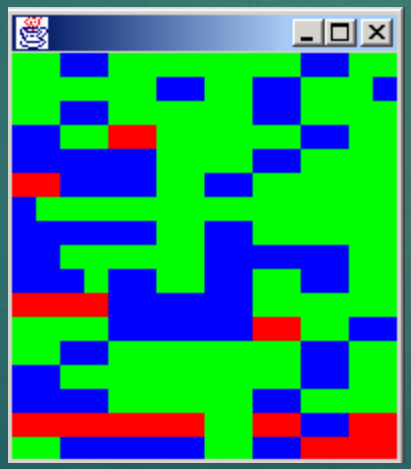
1.2.1. Khai phá dữ liệu

- Khai phá dữ liệu và trực quan hóa
  - Gán nhãn các lớp

Isodata (K-means) Clustering



Mean Feature Image

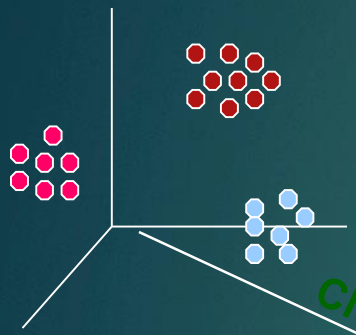


Label Image

1.2.2. Các tác vụ khai phá dữ liệu

- Khai phá mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa dữ liệu)
- 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
- …

1.2.2. Các tác vụ khai phá dữ liệu

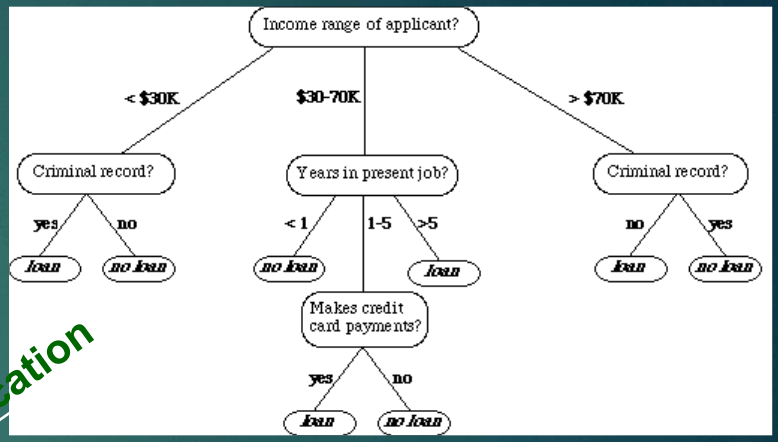


Clustering

Data

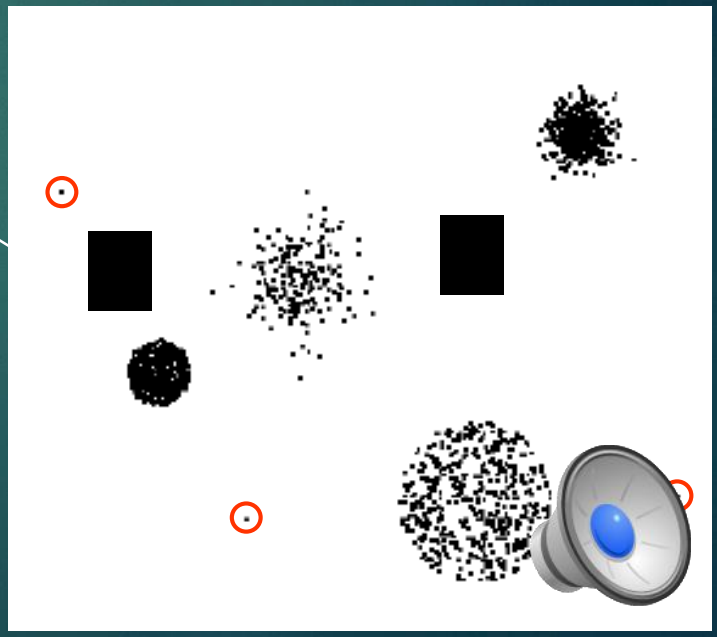
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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

Classification



Association Rules

Anomaly Detection



others

1.2.2. Các thành phần tác vụ khai phá dữ liệu

- Năm thành tố cơ bản để đặc tả một tác vụ khai phá dữ liệu
  - 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)

1.2.2. Các tác vụ khai phá dữ liệu

- Dữ liệu cụ thể sẽ được khai phá (task-relevant data)
  - Phần dữ liệu từ các dữ liệu nguồn được quan tâm
  - Tương ứng với các thuộc tính hay chiều dữ liệu được quan tâm
  - Bao gồm: tên kho dữ liệu/cơ sở dữ liệu, các bảng dữ liệu hay các khối dữ liệu, các điều kiện chọn dữ liệu, các thuộc tính hay chiều dữ liệu được tâm, các tiêu chí gom nhóm dữ liệu

1.2.2. Các tác vụ khai phá dữ liệu

- Loại tri thức sẽ đạt được (kind of knowledge)
  - Bao gồm: đặc trưng hóa dữ liệu, phân biệt hóa dữ liệu, mô hình phân tích kết hợp hay tương quan, mô hình phân lớp, mô hình dự đoán, mô hình gom cụm, mô hình phân tích phần tử biên, mô hình phân tích tiến hóa
  - Tương ứng với tác vụ khai phá dữ liệu cụ thể sẽ được thực thi

1.2.2. Các tác vụ khai phá dữ liệu

- Tri thức nền (background knowledge)
  - 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

1.2.2. Các tác vụ khai phá dữ liệu

- Các độ đo (interestingness measures)
  - 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)

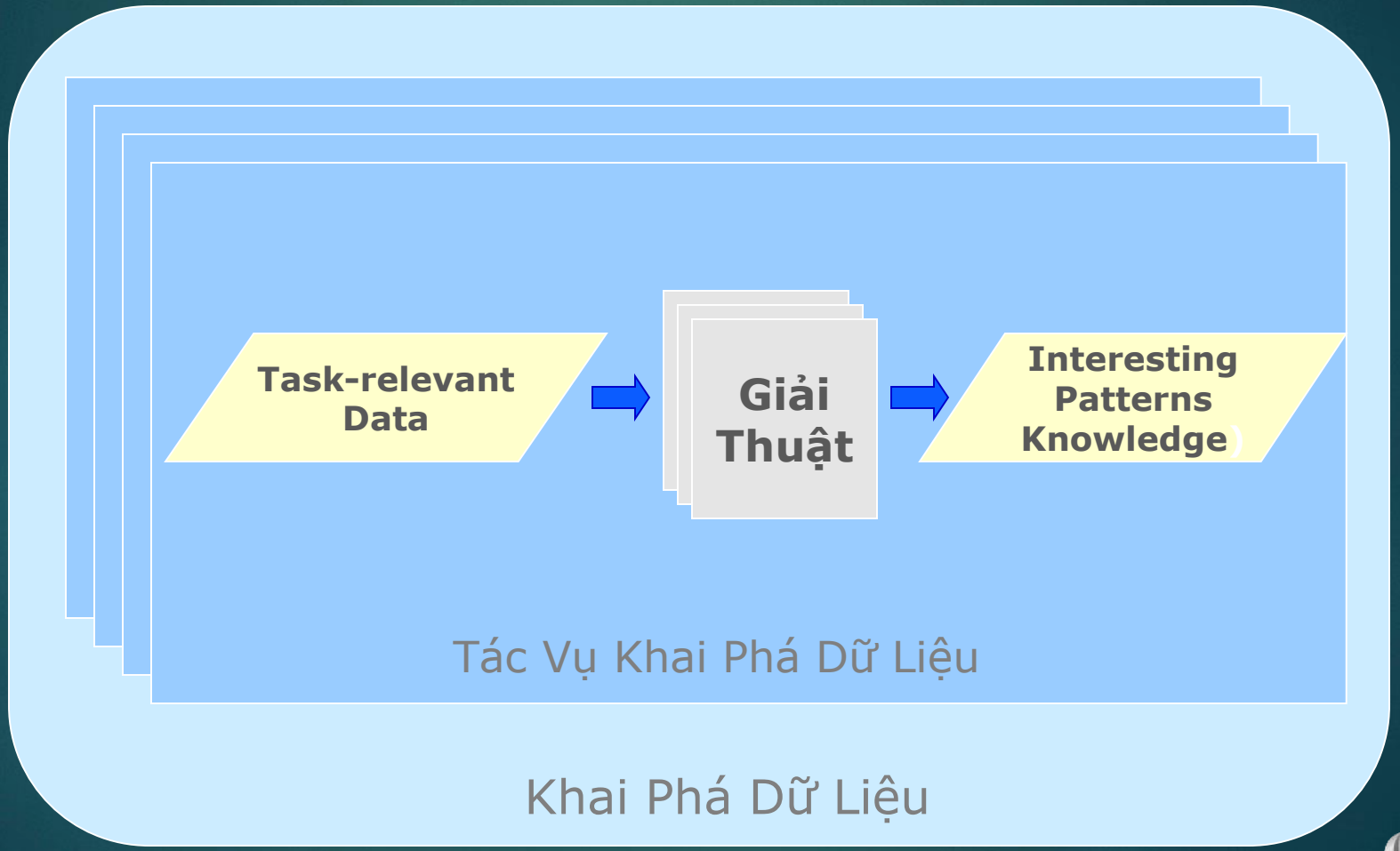
1.2.2. Các tác vụ khai phá dữ liệu

- Các kỹ thuật biểu diễn tri thức/trực quan hóa mẫu (pattern visualization and knowledge presentation)
  - 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), và khối (cubes)

1.2.2. Các tác vụ khai phá dữ liệu

- Khai phá dữ liệu
  - Phân loại dữ liệu
    - Giải thuật phân loại với cây quyết định
    - Giải thuật phân loại với mạng Bayes
    - …
  - Gom cụm dữ liệu
    - Giải thuật gom cụm k-means
    - Giải thuật gom cụm phân cấp nhóm
    - …
  - Khai phá luật kết hợp
    - Giải thuật Apriori
    - …
  - …

1.2.2. Các tác vụ khai phá dữ liệu



1.2.3. Các quy trình khai phá dữ liệu

- Quy trình khai phá dữ liệu 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.
  - Cross Industry Standard Process for Data Mining (CRISP-DM at www.crisp-dm.org)
  - SEMMA (***S***ample, ***E***xplore, ***M***odify, ***M***odel, ***A***ssess) at the SAS Institute

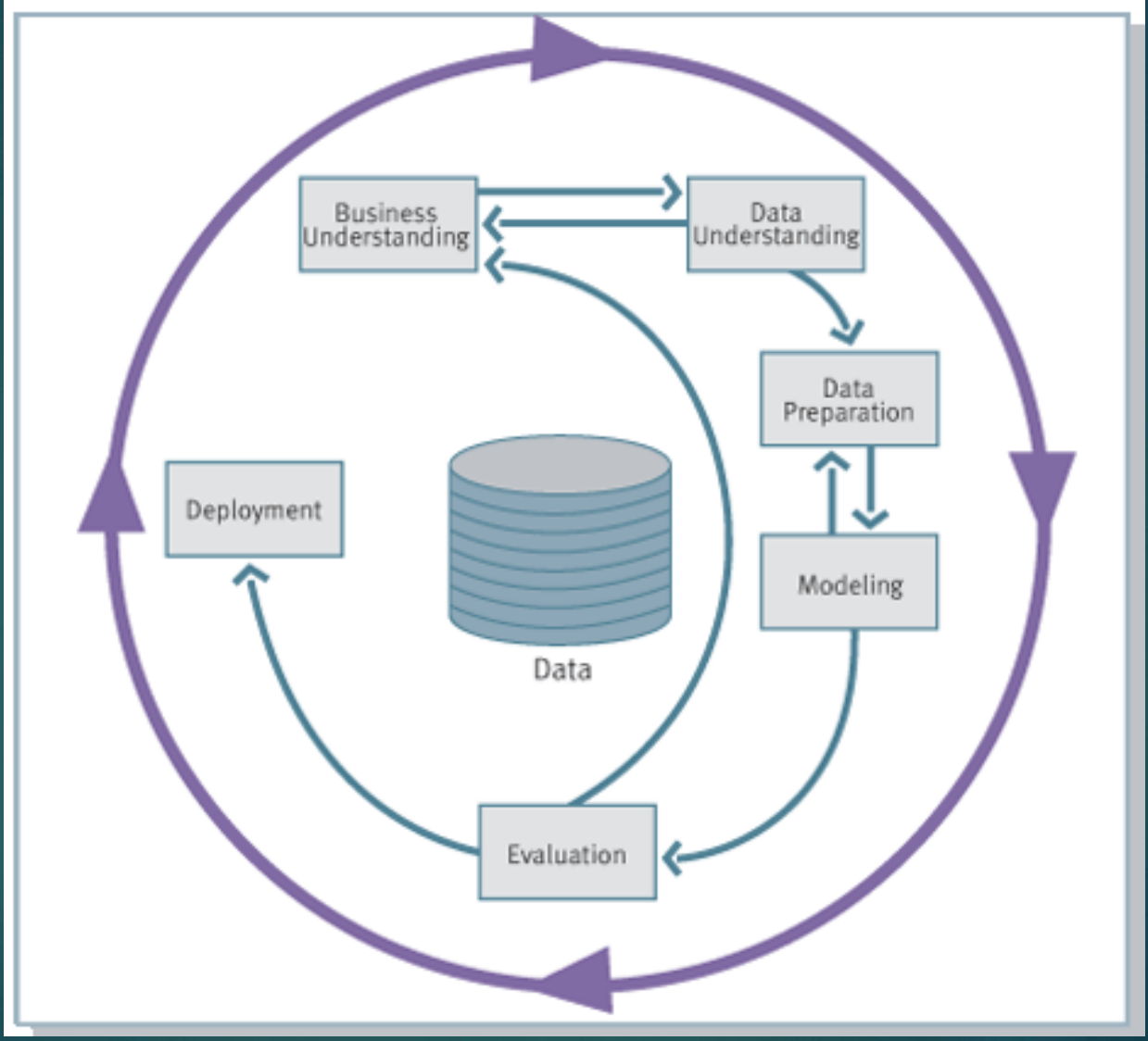
1.2.3. Các quy trình khai phá dữ liệu

- Sự cần thiết của một quy trình khai phá dữ liệu
  - 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.

1.2.3. 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 nền tảng kinh nghiệm từ các lĩnh vực ứng dụng

1.2.3. Quy trình CRISP-DM



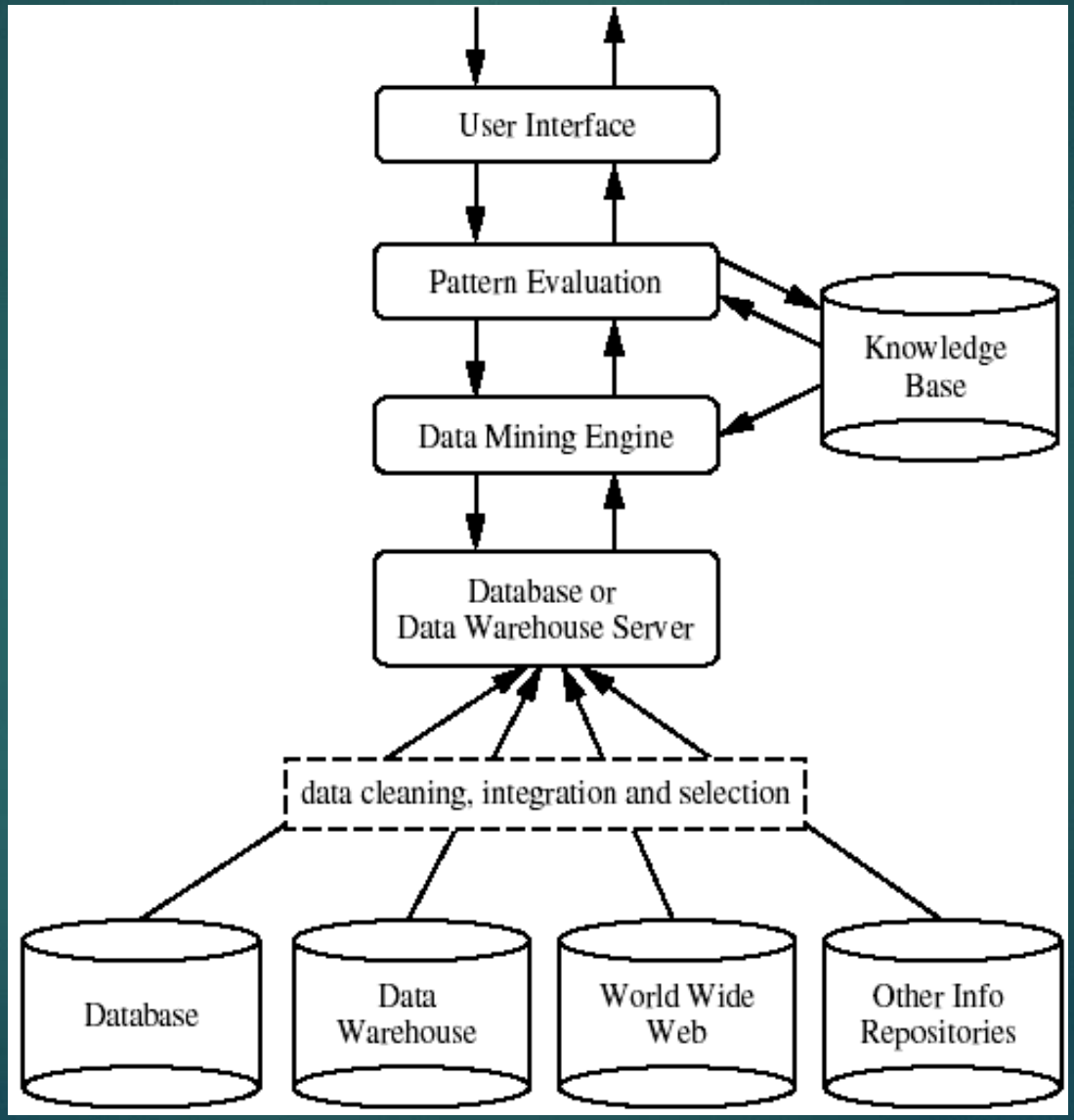
1.2.3. Quy trình CRISP-DM

- Quy trình CRISP-DM là một quy trình lặp, có khả năng quay lui (backtracking) 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)

1.2.4. Các hệ thống khai phá dữ liệu

- 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

1.2.4. Kiến trúc của một hệ thống khai phá dữ liệu



1.2.4. Các hệ thống khai phá dữ liệu

- 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 nguồn 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.

1.2.4. Các hệ thống khai phá dữ liệu

- Knowledge base
  - Thành phần chứa tri thức miền, được dùng để hướng dẫn quá trình tìm kiếm, đánh giá các mẫu kết quả được tìm thấy.
  - Tri thức miền có thể là các phân cấp khái niệm, niềm 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.

1.2.4. Các hệ thống khai phá dữ liệu

- Pattern evaluation module
  - Thành phần này làm việc với các độ đo (và các ngưỡng giá trị) hỗ trợ tìm kiếm và đánh giá các mẫu sao cho các mẫu 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.

1.2.4. Các hệ thống khai phá dữ liệu

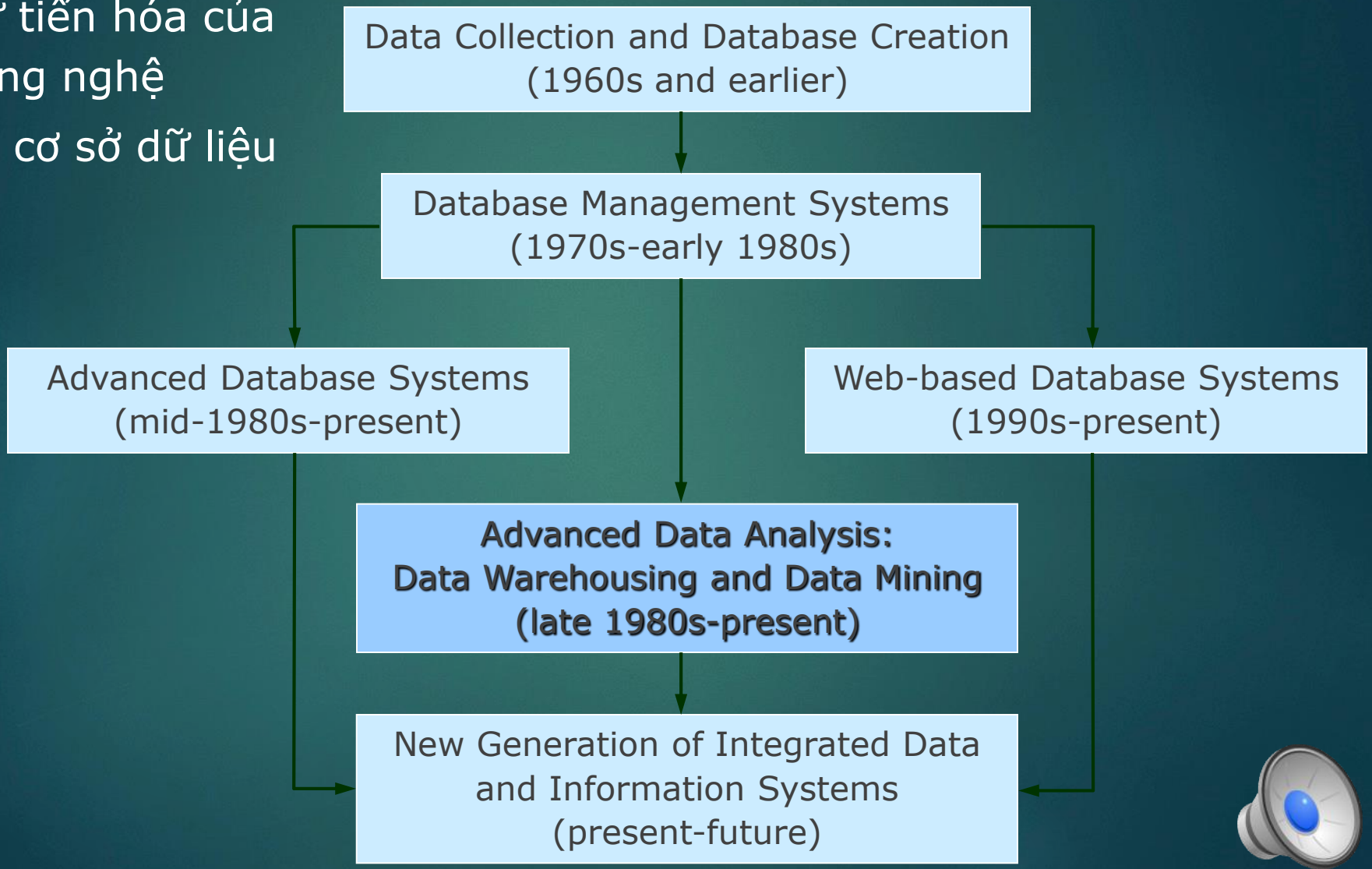
- User interface
  - Thành phần hỗ trợ sự tương tác giữa người sử dụng và hệ thống khai phá dữ liệu.
    - Người sử dụng có thể chỉ định câu truy vấn hay tác vụ khai phá dữ liệu.
    - Người sử dụng có thể được cung cấp thông tin hỗ trợ việc tìm kiếm, thực hiện khai phá dữ liệu sâu hơn thông qua các 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.

1.2.4. Các hệ thống khai phá dữ liệu

- 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/ml/weka)
  - …

1.3. Ý nghĩa và vai trò của khai phá dữ liệu

Sự tiến hóa của công nghệ hệ cơ sở dữ liệu



1.3. Ý nghĩa và vai trò của khai phá dữ liệu

- 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, …

1.4. Ứng dụng của khai phá dữ liệu

- Trong kinh doanh (business)
- Trong tài chính (finance) và tiếp thị bán hàng (sales marketing)
- Trong thương mại (commerce) và ngân hàng (bank)
- Trong bảo hiểm (insurance)
- Trong khoa học (science) và y sinh học (biomedicine)
- Trong điều khiển (control) và viễn thông (telecommunication)
- …

1.5. Tóm tắt

- Khai phá dữ liệu là quá trình khám phá ra các mẫu được quan tâm từ lượng lớn dữ liệu.
  - Mẫu kết quả khai phá được là những mẫu thể hiện tri thức nếu chúng dễ hiểu, hợp lệ với một mức độ chắc chắn, hữu dụng, và mới đối với người dùng.
  - Lượng lớn dữ liệu từ các cơ sở dữ liệu truyền thống/hiện đại, kho dữ liệu, hay từ các nguồn thông tin khác (spatial, time series, text, multimedia, web, …).
  - Các tác vụ khai phá dữ liệu bao gồm khai phá mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa dữ liệu), khai phá luật kết hợp/tương quan, phân lớp, dự đoán, gom cụm, phân tích xu hướng, phân tích độ lệch và phần tử biên, phân tích độ tương tự, …
    - Năm thành tố cơ bản để đặc tả một tác vụ khai phá dữ liệu: dữ liệu cụ thể sẽ được khai phá, loại tri thức sẽ đạt được, tri thức nền, các độ đo, và các kỹ thuật biểu diễn/trực quan hóa tri thức.
    - Bốn thành phần cơ bản của một giải thuật khai phá dữ liệu: cấu trúc mẫu hay mô hình, hàm tỉ số, phương pháp tìm kiếm và tối ưu hóa, chiến lược quản lý dữ liệu.

1.5. Tóm tắt

- Khai phá dữ liệu được xem như là một phần của quá trình khám phá tri thức.
- Quá trình khám phá tri thức là một chuỗi lặp gồm các bước: làm sạch dữ liệu, tích hợp dữ liệu, chọn lựa dữ liệu, biến đổi dữ liệu, khai phá dữ liệu, đánh giá mẫu, và biểu diễn tri thức.
- Nhiều lĩnh vực khác nhau có liên quan với khai phá dữ liệu: công nghệ cơ sở dữ liệu, lý thuyết thống kê, học máy, khoa học thông tin, trực quan hóa, …
- Các vấn đề liên quan: phương pháp luận khai phá dữ liệu, vấn đề tương tác người dùng, khả năng co giãn dữ liệu và hiệu suất, vấn đề xử lý lượng lớn các kiểu dữ liệu khác nhau, vấn đề khai thác các ứng dụng khai phá dữ liệu cũng như sự ảnh hưởng xã hội của chúng.