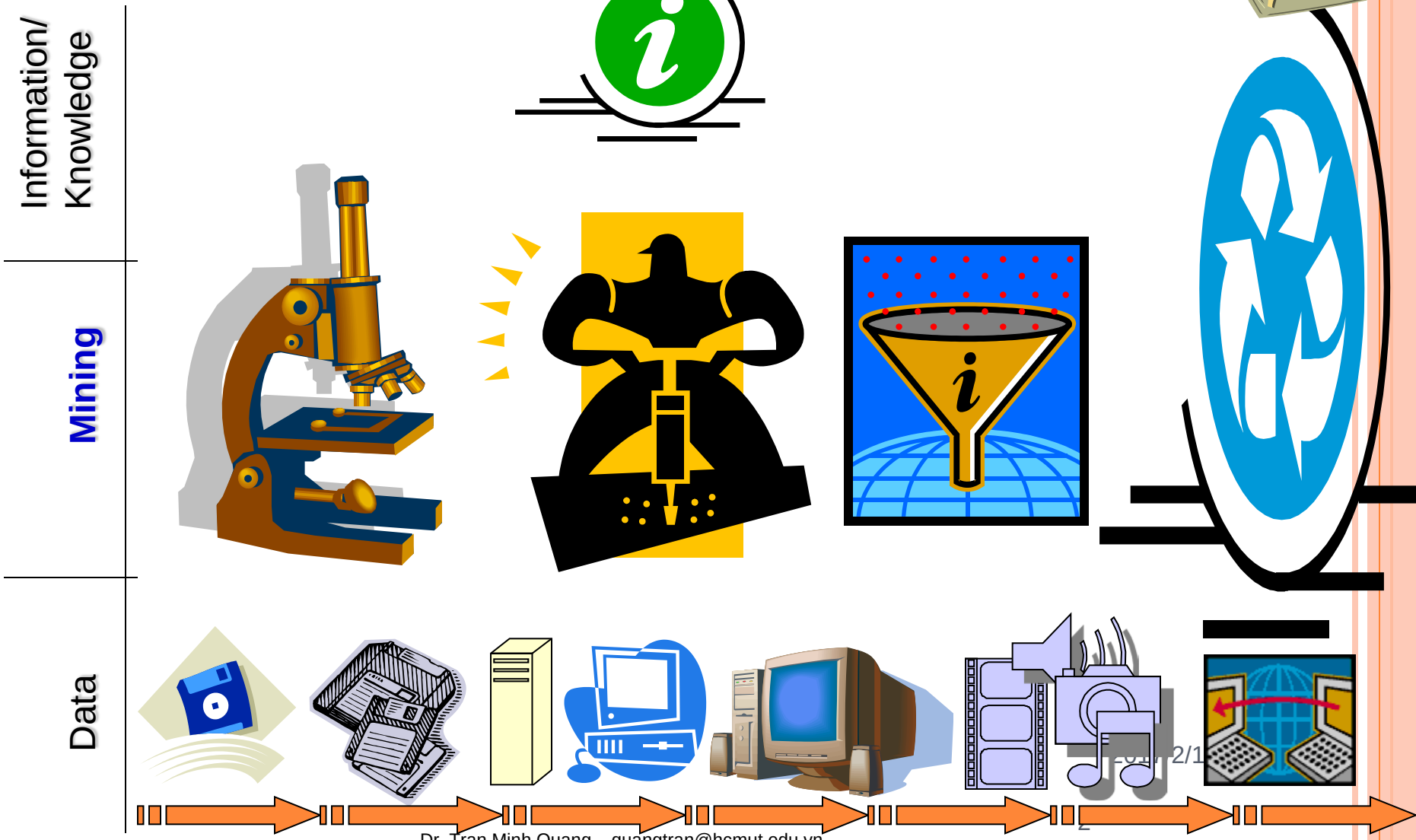
Khoa Khoa Học & Kỹ Thuật Máy Tính
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh

Chương 1
Tổng Quan về KPDL

TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://www.cse.hcmut.edu.vn/quangtran
http://researchmap.jp/quang

2017/2/10

KHAI PHÁ DỮ LIỆU (KPDL)



Nội Dung

1. Tình huống
2. Quá trình khám phá tri thức
3. Các khái niệm
4. Ý nghĩa và vai trò của KPDL
5. Ứng dụng
6. Tóm tắt

TÀI LIỆU THAM KHẢO

[1] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.

[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.

[8] Ian H.Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", 2nd Edition, Elsevier Inc, 2005.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

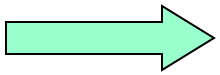
[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

1. TÌNH HUỐNG 1

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

1. 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**

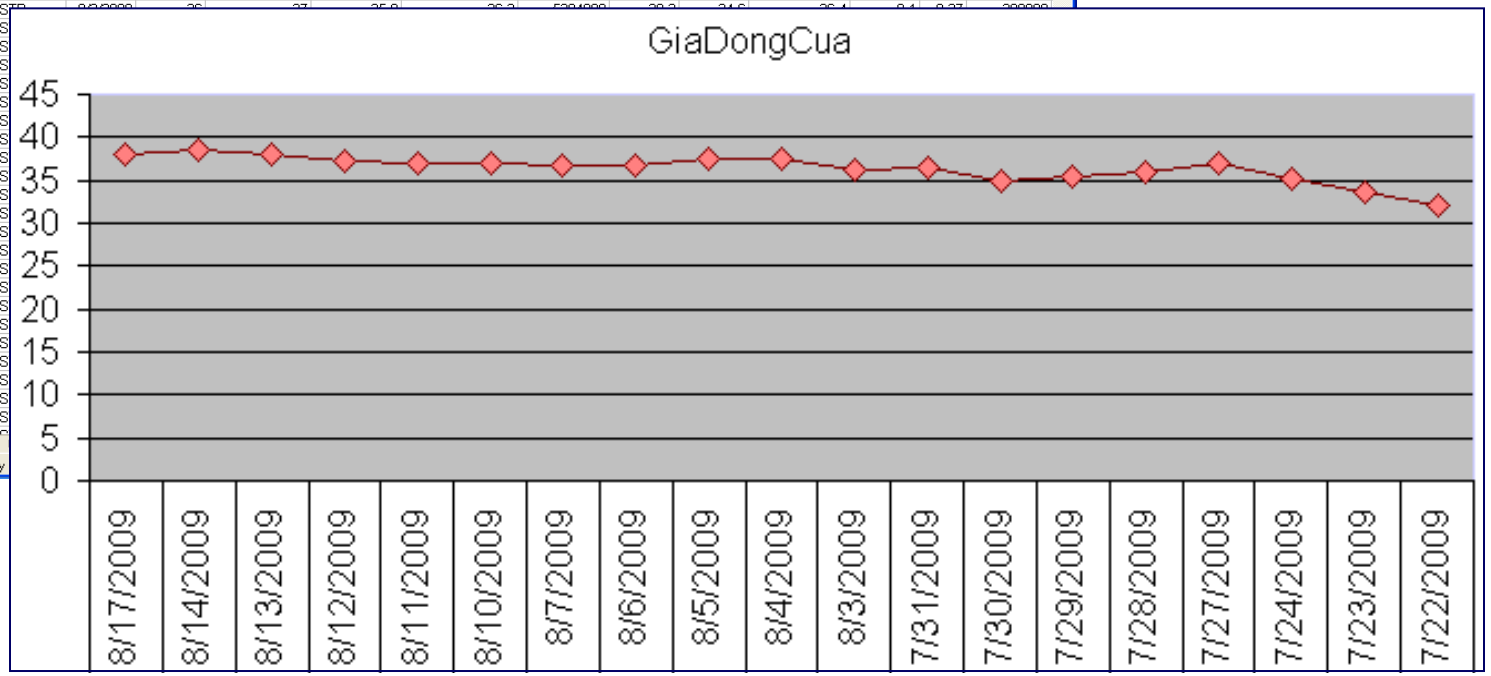


Ông A (Tid = 100) có khả năng trốn thuế?

1. TÌNH HUỐNG 3

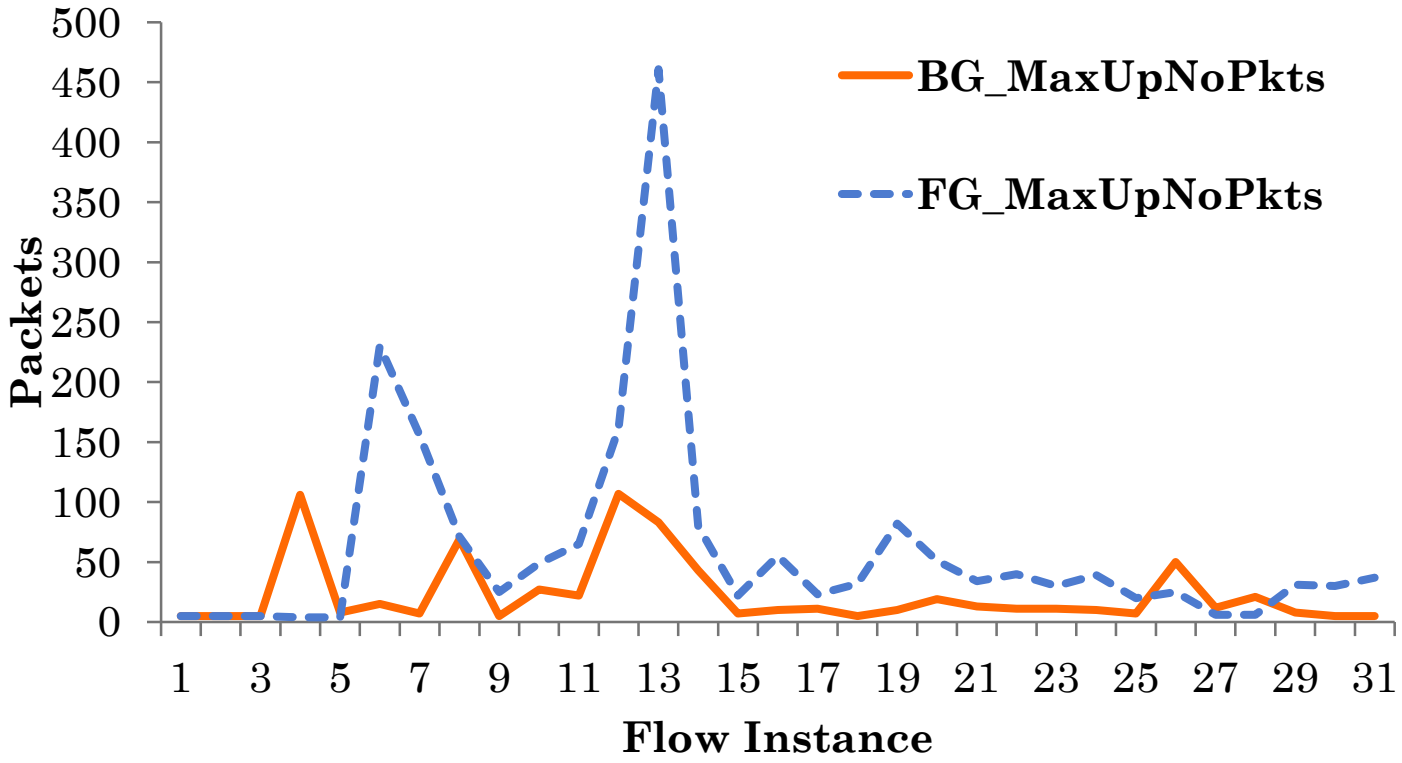
Microsoft Excel - stb.csv

	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
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



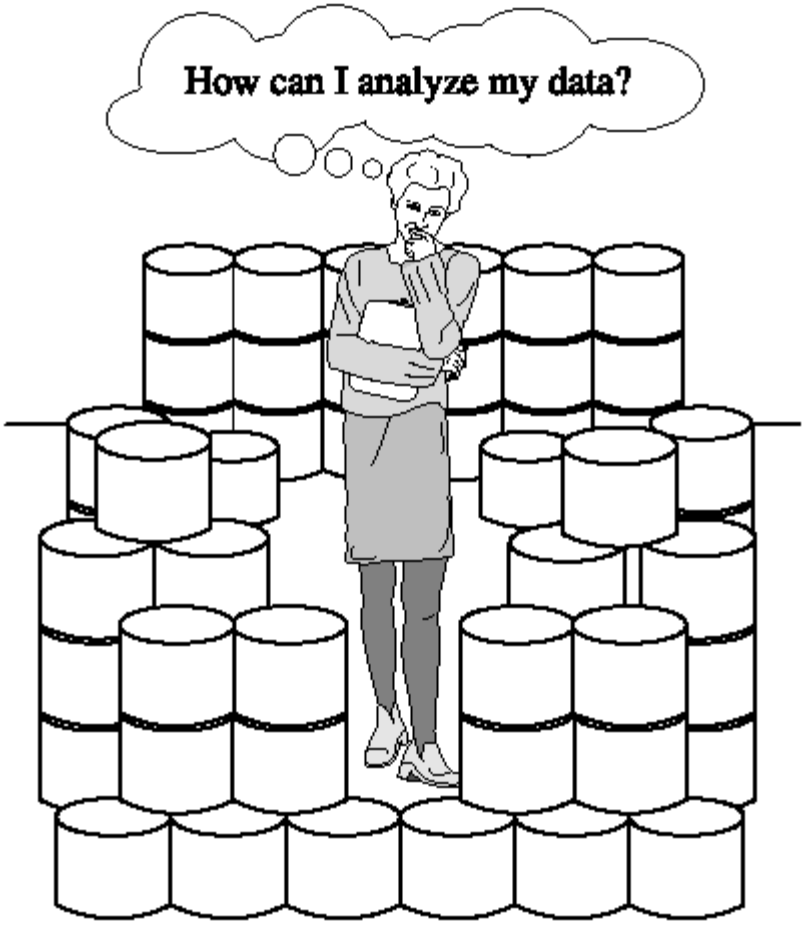
Ngày mai cổ phiếu STB sẽ tăng?

1. TÌNH HUỐNG 4



Khả năng mạng có đang bị tấn công hay không?

1. The Fact…



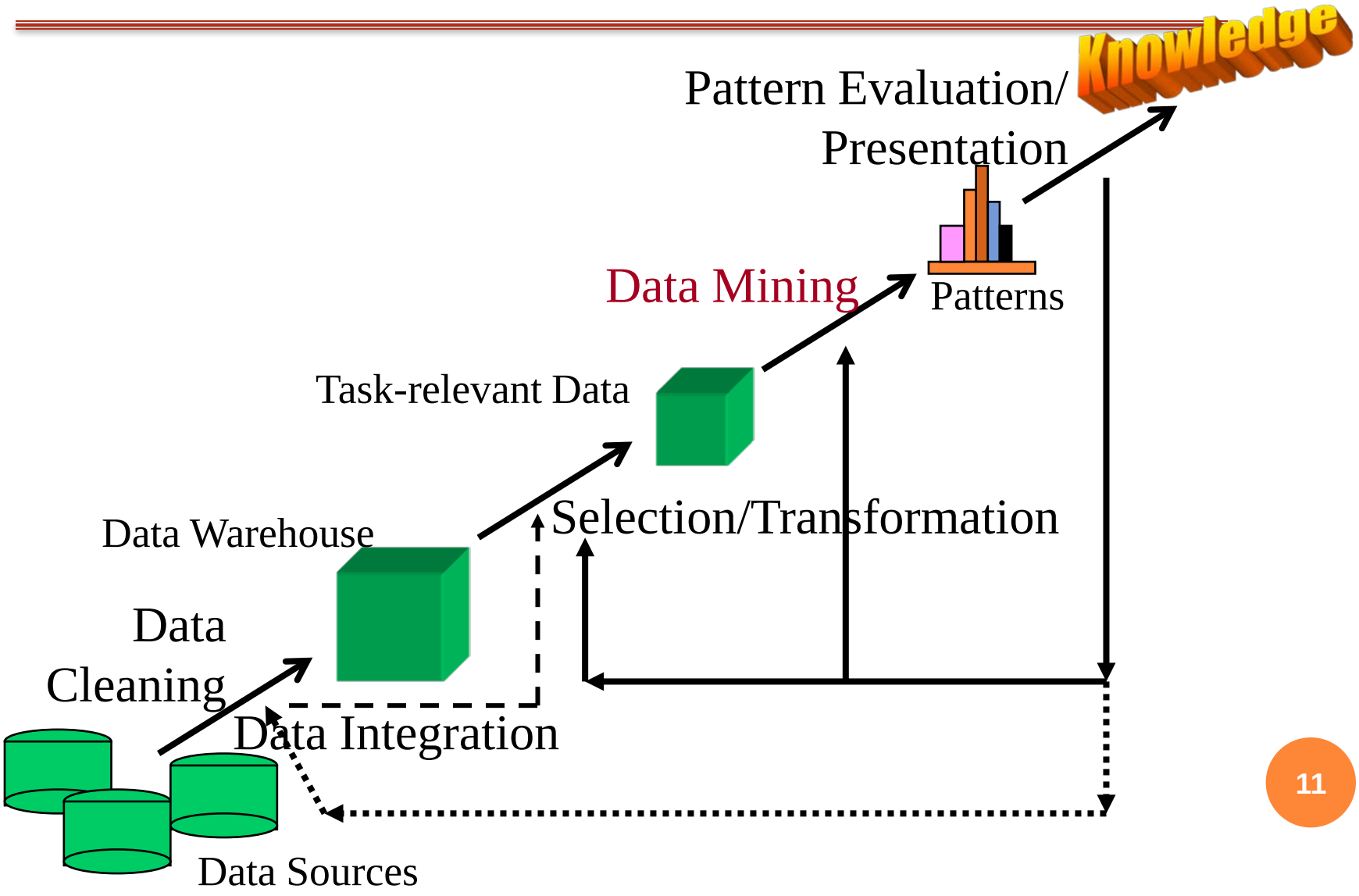
We are data rich, but information poor

“Necessity is the mother of invention” - Plato

2. Quá Trình Khai Phá Tri Thức

- "Knowledge discovery in **databases** is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable **patterns**"
  - 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.

2. Quá Trình Khai Phá Tri Thức



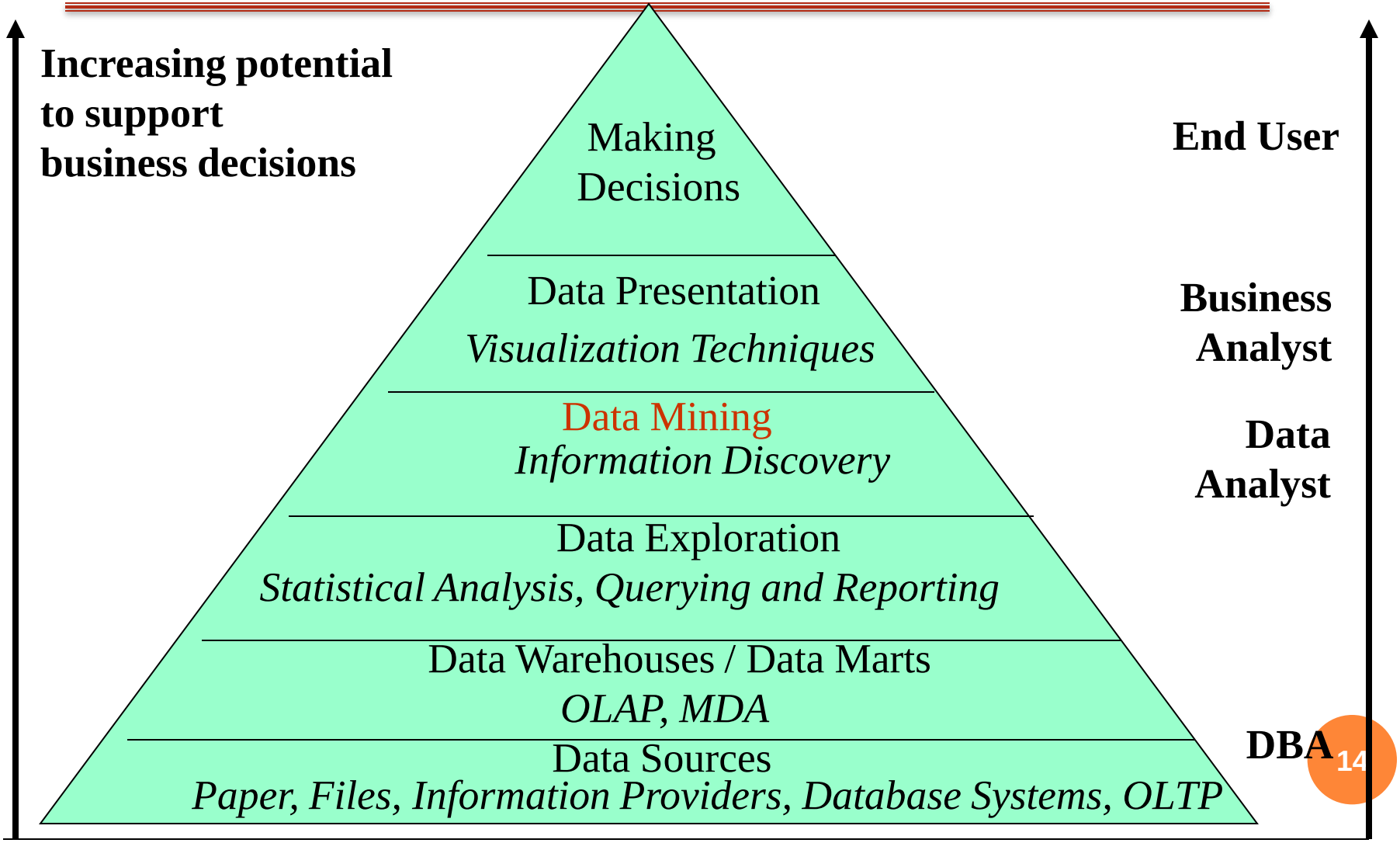
2. Quá Trình Khai Phá Tri Thức

- Quá trình khai 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)

2. Quá Trình Khai Phá Tri Thức

- Quá trình khai 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)

2. Quá Trình Khai Phá Tri Thức



3. Các Khái Niệm

- Khai phá dữ liệu (data mining)
- Các tác vụ KPDL (data mining tasks/functions)
- Các quy trình KPDL (data mining processes)
- Các hệ thống KPDL (data mining systems)

3.1. KPDL (DATA MINING)

- 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"

3.1. KPDL

- Thuật ngữ tương đương:
  - knowledge discovery/mining in data/databases (KDD)
  - knowledge extraction
  - data/pattern analysis
  - data archaeology, data dredging
  - information harvesting
  - business intelligence

3.1. 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 (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 gian - 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)

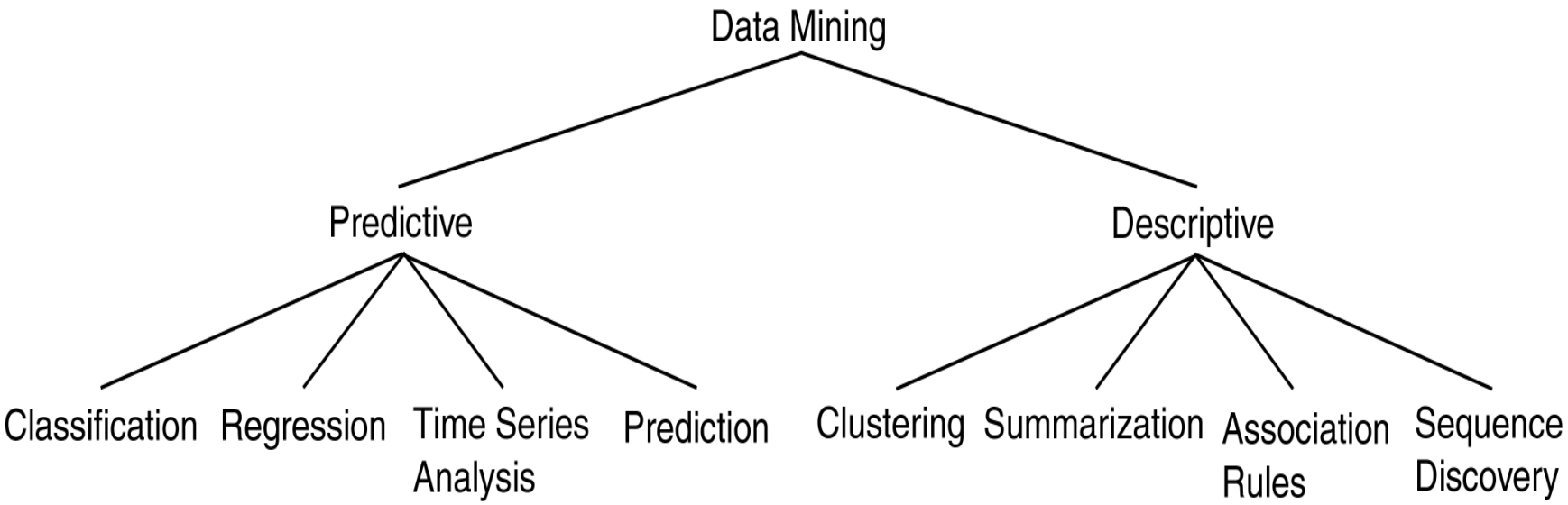
3.1. KPDL

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

3.1. KPDL

- Tri thức đạt được từ quá trình khai phá có thể là:
  - mô tả hay dự đoán
    - Mô tả (Descriptive): có khả năng đặc trưng hóa các thuộc tính chung của dữ liệu được khai phá (Tình huống 1)
    - Dự đoán (Predictive): có khả năng suy luận từ dữ liệu hiện có để dự đoán (Tình huống 2, 3, và 4)
  - có cấu trúc, bán cấu trúc, hoặc phi cấu trúc
  - được/không được người dùng quan tâm → các độ đo đánh giá tri thức đạt được
  - được dùng trong việc hỗ trợ ra quyết định, điều khiển quy trình, quản lý thông tin, xử lý truy vấn …

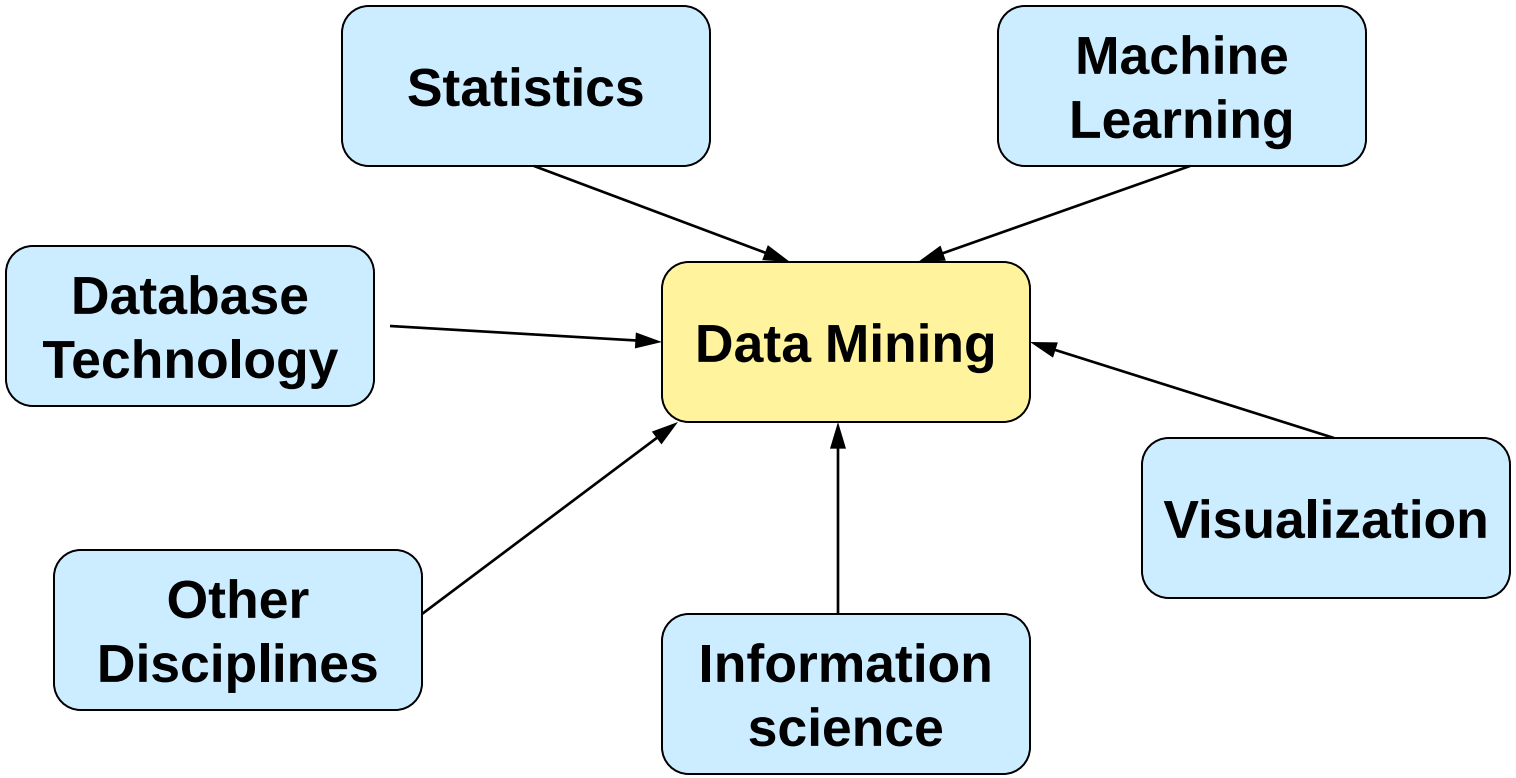
3.1. KPDL



(trends, regularities, …)

(characterization and discrimination)

3.1. KPDL



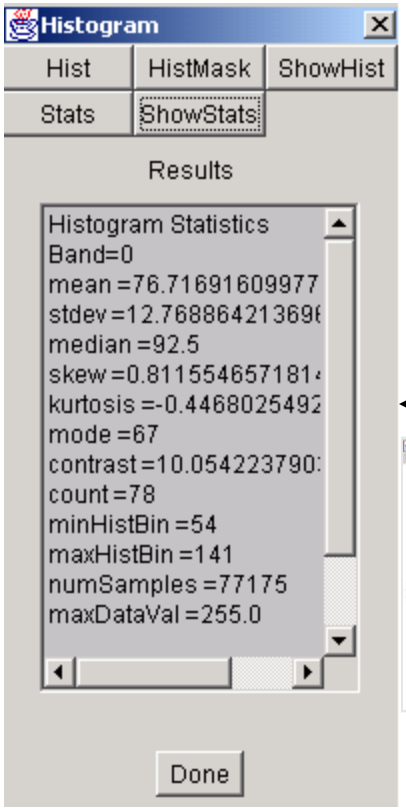
"Data mining as a confluence of multiple disciplines"

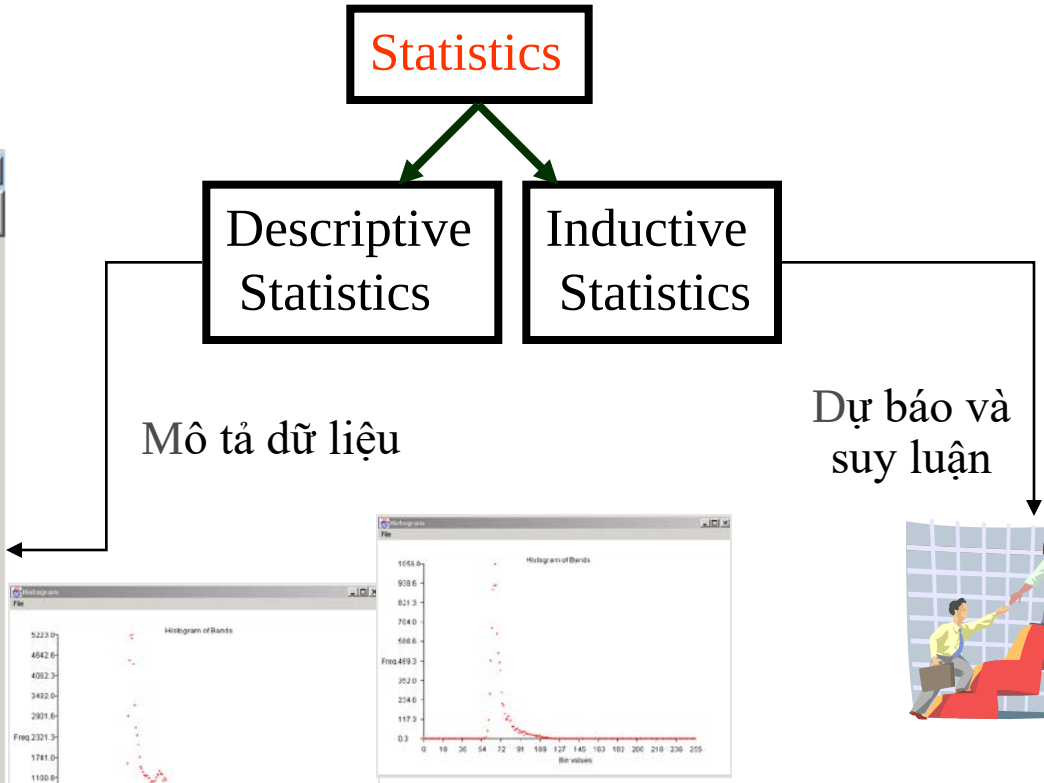
3.1. KPDL

- KPDL và công nghệ cơ sở dữ liệu
  - Quản lý hiệu quả dữ liệu được khai phá
    - Dữ liệu rất lớn: phân trang (paging), hoán chuyển (swapping) dữ liệu vào/ra bộ nhớ chính
    - Dữ liệu được thu thập theo thời gian
  - Có khả năng xử lý nhiều loại dữ liệu phức tạp (spatial, temporal, spatiotemporal, multimedia, text, Web, …)
  - Hỗ trợ xử lý đồng thời, bảo mật, hiệu năng, tối ưu hóa...
  - Hỗ trợ các tính năng/công cụ khai phá dữ liệu
    - Oracle Data Mining (Oracle 9i, 10g, 11g)
    - Các công cụ khai phá dữ liệu của SQL Server
    - Intelligent Miner (IBM)
    - Các hệ CSDL qui nạp (inductive DB): hỗ trợ KPTT
    - Chuẩn SWL/MM 6: Data Mining của ISO/IEC 13249-6:2006 hỗ trợ KPDL

3.1. KPDL

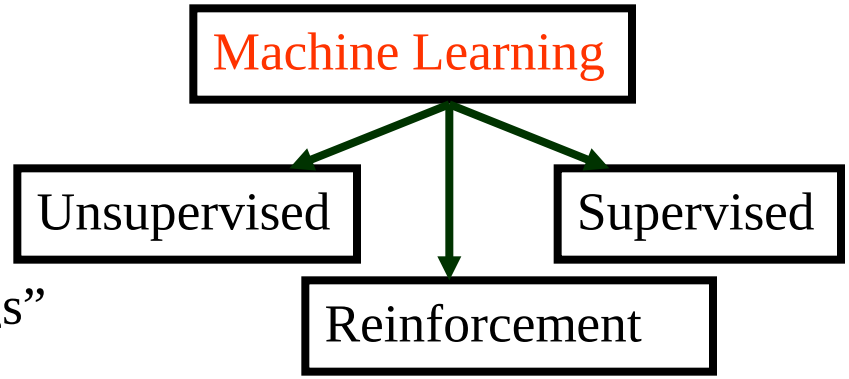
- KPDL và lý thuyết thống kê



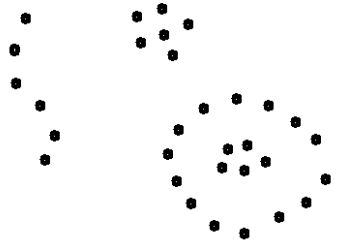


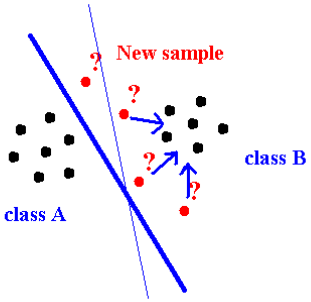
3.1. KPDL

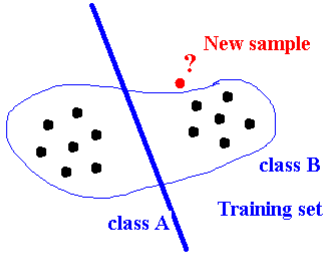
- KPDL và học máy

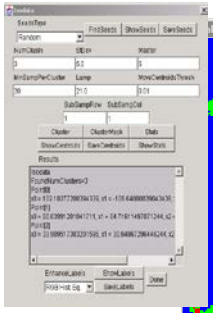


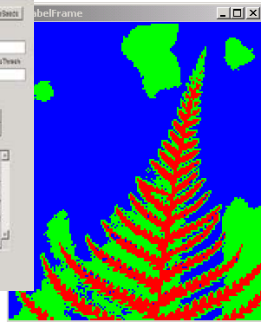
"Natural groupings"

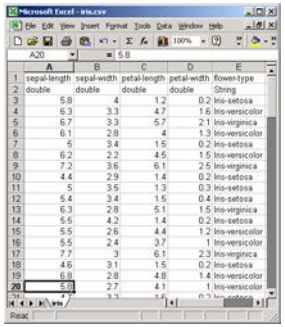


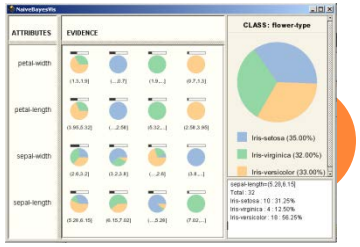






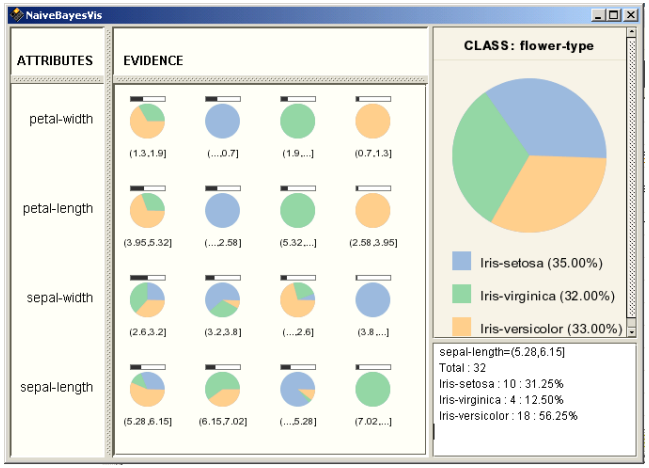




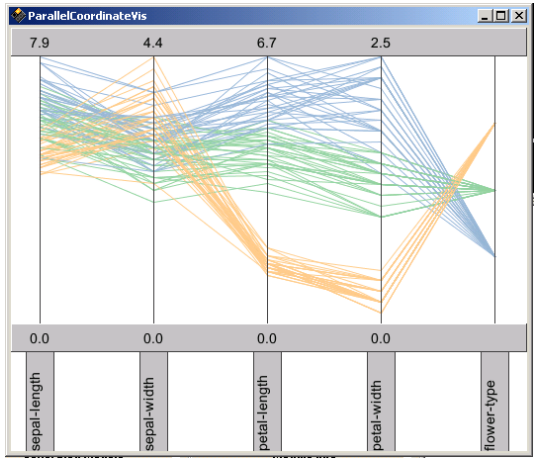


3.1. KPDL

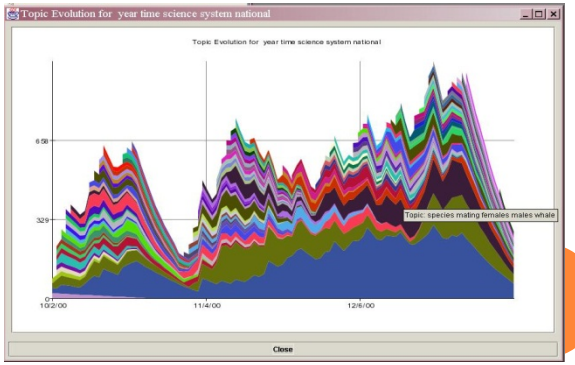
- KPDL và trực quan hóa
  - Dữ liệu: 3D cubes, distribution charts, curves, surfaces, link graphs, image frames and movies, parallel coordinates
  - Kết quả (tri thức): pie charts, scatter plots, box plots, association rules, parallel coordinates, temporal



Pie chart



Parallel coordinates

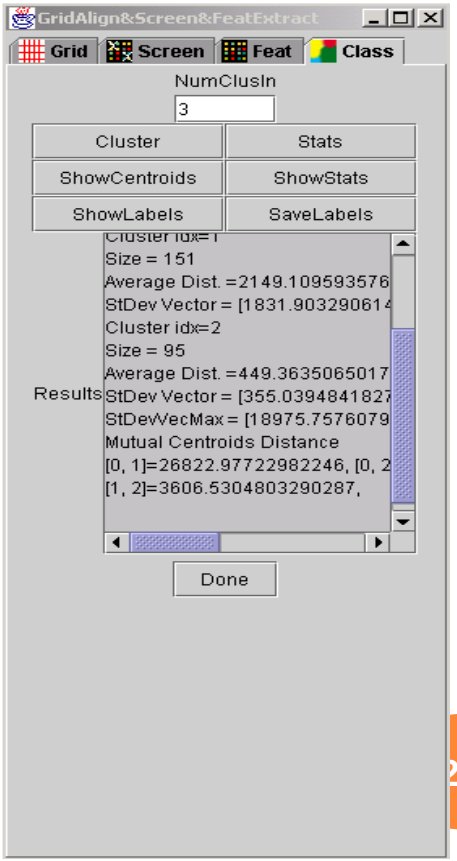


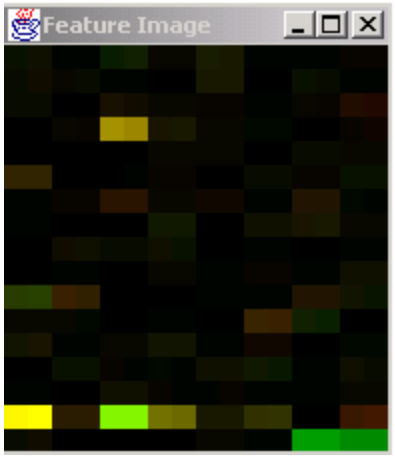
Temporal evolution

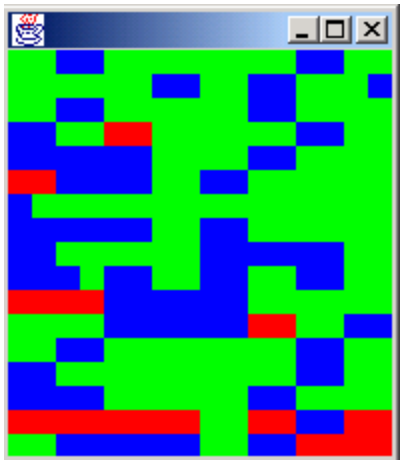
3.1. KPDL

- KPDL và trực quan hóa
  - Gán nhãn các lớp

Isodata (K-means) Clustering







Mean Feature Image

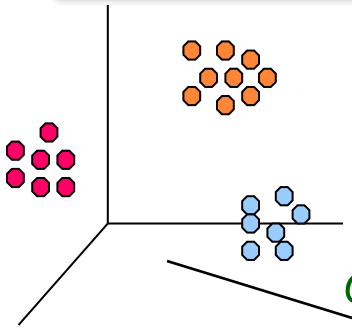
Label Image

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

3.2. Các Tác Vụ KPDL

- Khai phá mô tả lớp/khái niệm (đặc trưng hóa và phân biệt hóa dữ liệu) - Description
- Khai phá luật kết hợp/tương quan - Association rule
- Phân loại/lớp dữ liệu - Classification
- Dự đoán - Prediction
- Gom cụm dữ liệu - Clustering
- Phân tích xu hướng – Trend analysis
- Phân tích độ lệch và phần tử biên - Outlier
- Phân tích độ tương tự - Similarity analysis
- …

3.2. CÁC TÁC VỤ KPDL

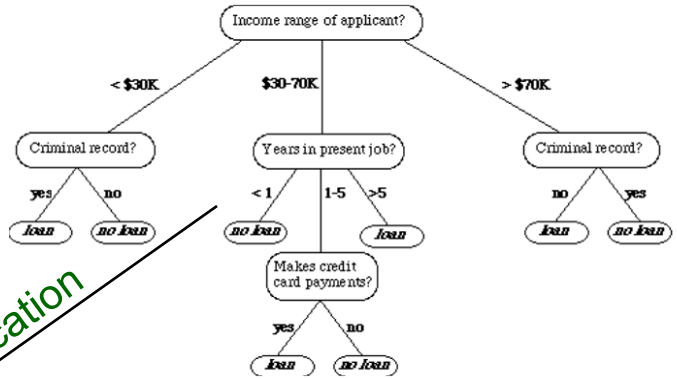


Clustering

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

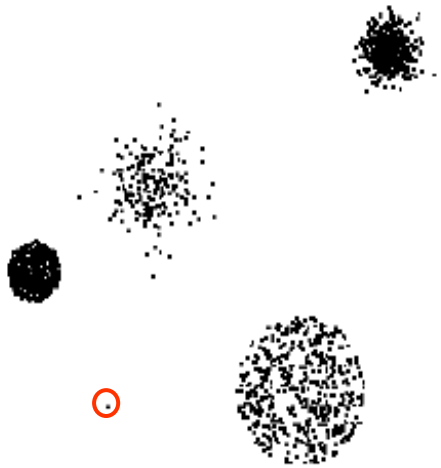
Classification



Association Rules

Milk

Anomaly Detection



others

3.2. Các Tác Vụ KPDL

- Năm thành tố cơ bản để đặc tả một 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)

3.2. Các Tác Vụ KPDL

- Dữ liệu cụ thể sẽ được khai phá (task-relevant data)
  - Dữ liệu từ các dữ liệu nguồn được quan tâm
  - Tương ứng với các thuộc tính (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 quan tâm, các tiêu chí gom nhóm dữ liệu

3.2. Các Tác Vụ KPDL

- 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

3.2. Các Tác Vụ KPDL

- 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

3.2. Các Tác Vụ KPDL

- 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 KPDL cụ thể
  - Kiểm tra: tính đơn giản (simplicity), tính chắc chắn (certainty), tính hữu dụng (utility), tính mới (novelty)

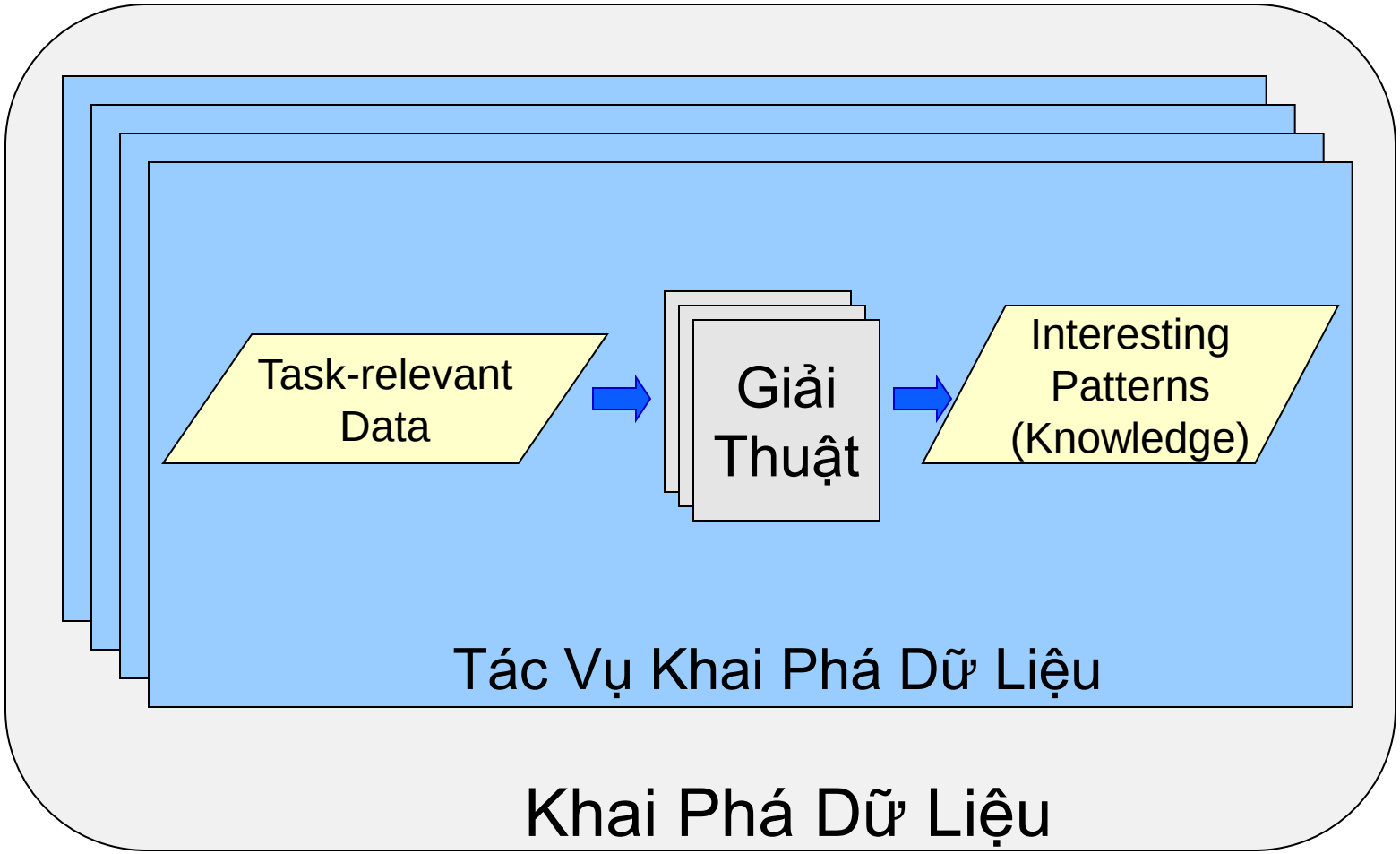
3.2. Các Tác Vụ KPDL

- 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)

3.2. Các Tác Vụ KPDL

- Các tác vụ KPDL thường gặp
  - Phân loại dữ liệu (Data classification)
    - Logistic regression, Bayes Network, Decision tree, ANN,…
  - Gom cụm dữ liệu (Clustering)
    - Hierachical, partition methods
    - K-means like algorithms,…
  - Khai phá luật kết hợp (Association rules)
    - Giải thuật Apriori, FP-Grownth,..
  - Outlier/Abnormality detection
  - ….

3.2. Các Tác Vụ KPDL



3.2. Các Tác Vụ KPDL

- Bốn thành phần cơ bản của một 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)

3.2. Các Tác Vụ KPDL

- Cấu trúc mẫu hay cấu trúc mô hình
  - Mô hình (model) là mô tả của tập dữ liệu, mang tính toàn cục ở mức cao
  - Mẫu (pattern) là đặc điểm của dữ liệu, mang tính cục bộ, chỉ cho một vài bản ghi/đối tượng hay vài biến
  - Cấu trúc (structure) 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

3.2. Các Tác Vụ KPDL

- Hàm tỉ số (score function)
  - Hàm xác định một cấu trúc mô hình/mẫu đáp ứng tập dữ liệu đã cho **tốt** ở mức độ nào đó
  - Dùng để so sánh các mô hình
  - Hàm tỉ số 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,…

3.2. Các Tác Vụ KPDL

- Phương pháp tìm kiếm và tối ưu hóa (optimization and search method)
  - Mục tiêu: 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 (trạng 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

3.2. Các Tác Vụ KPDL

- Chiến lược quản lý dữ liệu (data management strategy)
  - 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

3.3. Các Quy Trình KPDL

- 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 đáp ứng được sự quan tâm của người sử dụng (knowledge of interest)
  - 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

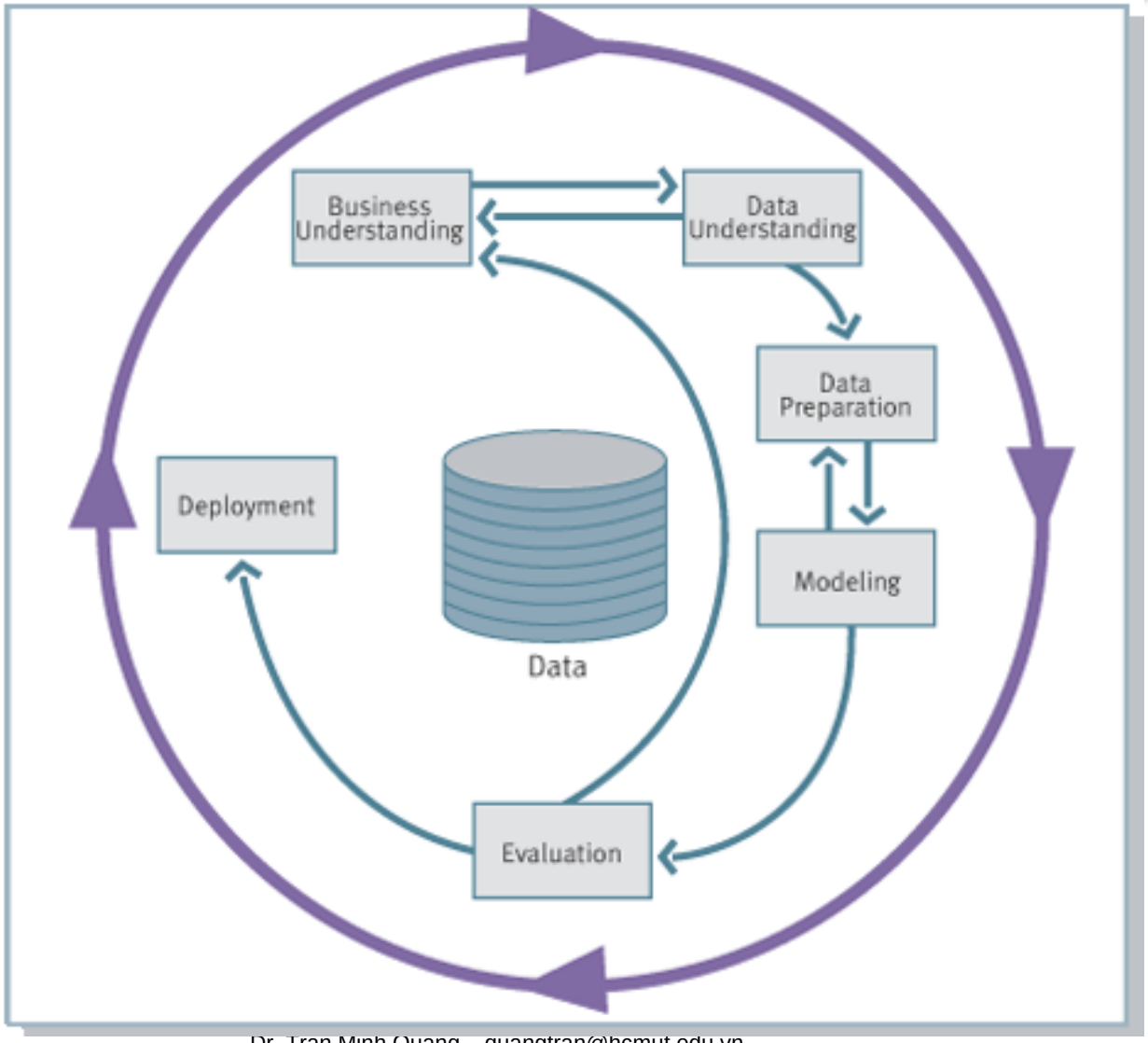
3.3. Các Quy Trình KPDL

- Sự cần thiết của một quy trình KPDL
  - Cách thức tiến hành (hoạch định và quản lý) dự án KPDL có hệ thống
  - Đảm bảo nỗ lực dành cho một dự án KPDL được tối ưu hóa
  - Việc đánh giá và cập nhật các mô hình trong dự án được diễn ra liên tục

3.3. Các Quy Trình KPDL

- Chuẩn quy trình công nghiệp
  - Khởi xướng từ 09/1996 với hơn 200 thành viên
  - Chuẩn mở
  - Hỗ trợ công nghiệp/ứng dụng và công cụ KPDL hiện có
  - Tập trung vào các vấn đề nghiệp vụ cũng như phân tích kỹ thuật
  - Tạo ra một khung thức hướng dẫn quy trình KPDL
  - Có nền tảng kinh nghiệm từ các lĩnh vực ứng dụng

3.3. Các Quy Trình KPDL

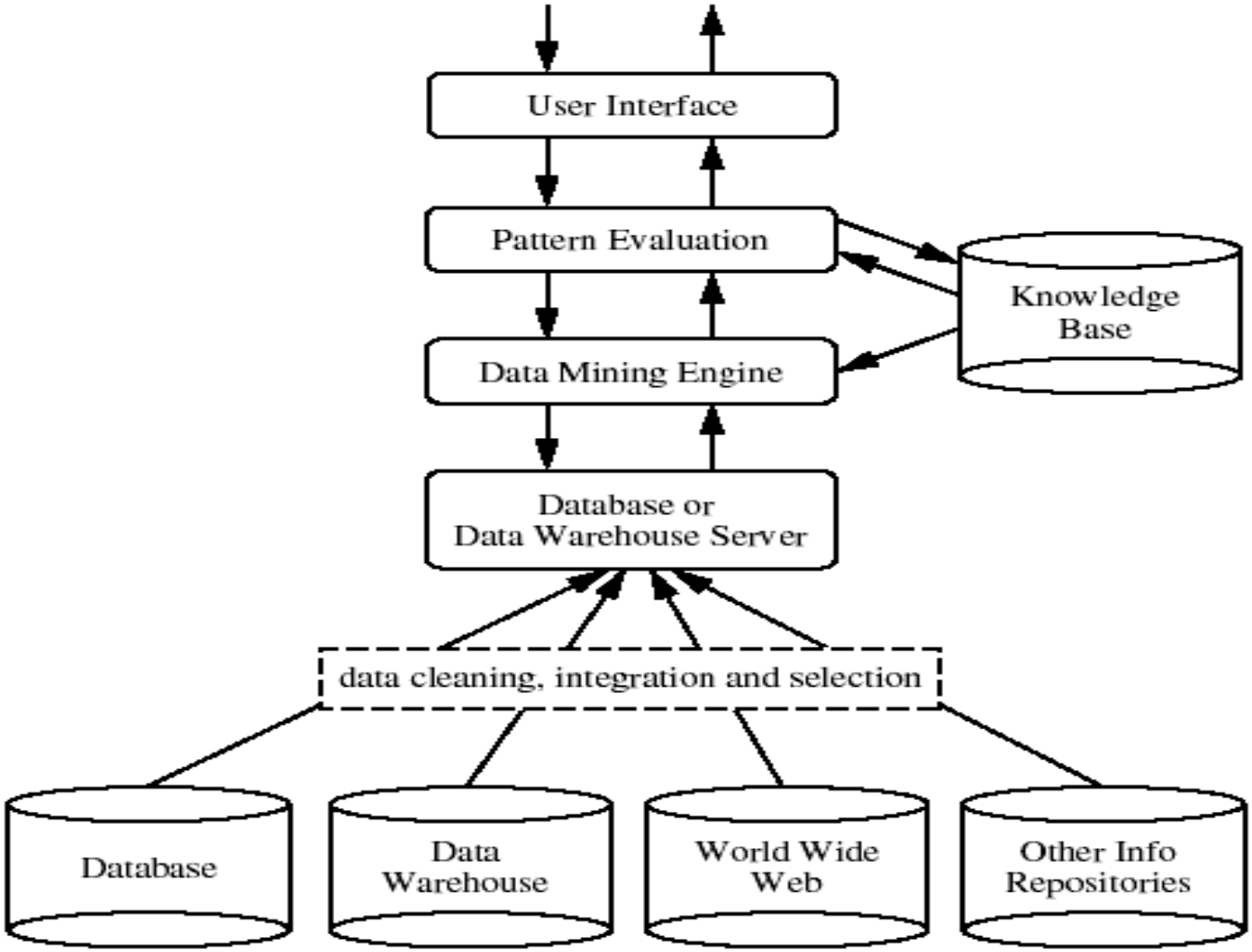


3.4. Các Hệ Thống KPDL

- Hệ thống KPDL được phát triển dựa trên khái niệm rộng của KPDL
  - KPDL 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

3.4. CÁC HỆ THỐNG KPDL

- Kiến trúc của một hệ thống KPDL



3.4. Các Hệ Thống KPDL

- Database, data warehouse, World Wide Web, và information repositories:
  - 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 KPDL

3.4. Các Hệ Thống KPDL

- Knowledge base
  - Chứa tri thức miền (Domain knowledge), 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
  - Chứa các khối chức năng thực hiện các tác vụ KPDL

3.4. Các Hệ Thống KPDL

- Pattern evaluation module
  - Dùng các độ đo (interestingness measure) và các ngưỡng giá trị (threshold) 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
  - Có thể được tích hợp vào thành phần Data mining engine

3.4. Các Hệ Thống KPDL

- User interface: Hỗ trợ sự tương tác giữa user và hệ thống KPDL:
  - Chỉ định câu truy vấn hay tác vụ KPDL
  - Được cung cấp thông tin hỗ trợ việc tìm kiếm, thực hiện KPDL sâu hơn thông qua các kết quả khai phá trung gian
  - 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.

3.4. Các Hệ Thống KPDL

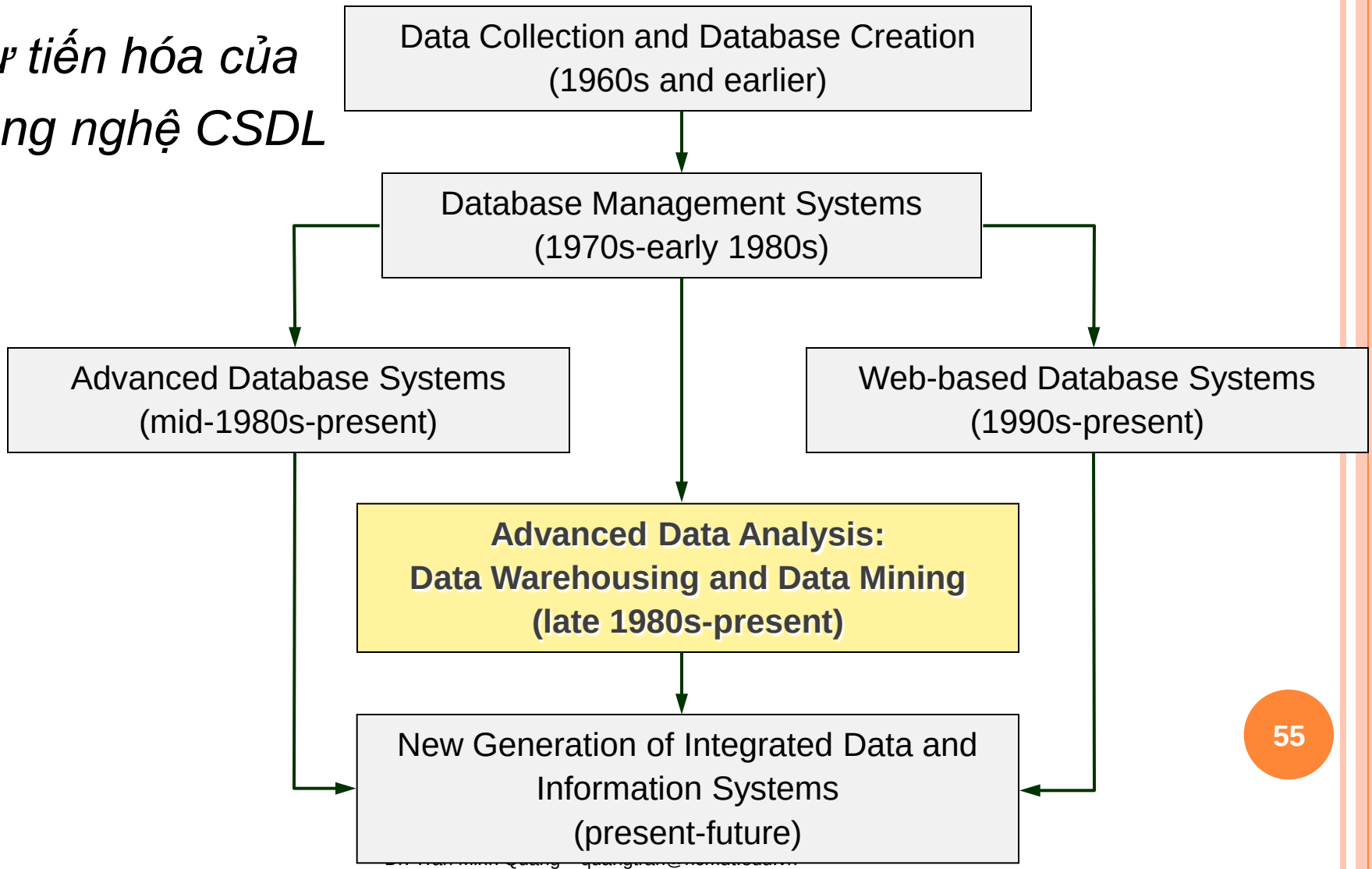
- Các đặc điểm được dùng để khảo sát một hệ thống khai phá dữ liệu
  - Kiểu dữ liệu
  - Nguồn dữ liệu
  - Các tác vụ và phương pháp luận khai phá dữ liệu
  - Vấn đề gắn kết với các hệ thống kho dữ liệu/cơ sở dữ liệu
  - Khả năng co giãn dữ liệu
  - Các công cụ trực quan hóa
  - Ngôn ngữ truy vấn khai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

3.4. Các Hệ Thống 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)
  - …

4. Ý NGHĨA VÀ VAI TRÒ CỦA KPDL

Sự tiến hóa của công nghệ CSDL



4. 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 nhiều khía cạnh của đời sống
    - 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, …

5. Ứng Dụng KPDL

- Ứng dụng trong nhiều lĩnh vực như:
  - kinh doanh (business), quản lý (management),…
  - tài chính (finance) và tiếp thị bán hàng (sales marketing)
  - thương mại (commerce) và ngân hàng (bank)
  - bảo hiểm (insurance)
  - khoa học (science) và y sinh học (biomedicine)
  - điều khiển (control) và viễn thông (telecommunication)
  - …

5. Tóm Tắt

- KPDL: khám phá các mẫu được quan tâm từ lượng dữ liệu lớn
- Tri thức khai phá được phải dễ hiểu, hữu dụng, hợp lý với một độ chắc chắn nhất định (đo/đánh giá được)
- Dữ liệu có thể từ nhiều nguồn (transactional DB, kho dữ liệu, web, các hệ thống điều khiển,…) với đa dạng các kiểu dữ liệu (text, multimedia, streaming data, có kèm thông tin về không gian và thời gian,…)

5. Tóm Tắt

- Các tác vụ KPDL thông thường: đặc trưng hóa/phân biệt hóa dữ liệu, phân lớp, dự báo, gom cụm, các luật kết hợp, phân tích sự bất thường, phân tích độ tương quan,…
- Năm thành tố đặc tả một tác vụ KPDL: dữ liệu được khai phá, loại tri thức cần khai phá, tri thức nền, các độ đo, các kỹ thuật biểu diễn tri thức
- Bốn thành phần cơ bản của một giải thuật KPDL: cấu trúc mẫu/mô hình, hàm tỉ số, phương pháp tìm kiếm và tối ưu, chiến lược quản lý dữ liệu

5. Tóm Tắt

- 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
- KPDL là một phần của quá trình khám phá tri thức
- Các lĩnh vực liên quan: công nghệ cơ sở dữ liệu, thống kê, học máy, khoa học thông tin, trực quan hóa,…
- Các vấn đề liên quan: phương pháp luận KPDL, tương tác người dùng, khả năng co giãn 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

Q&A

quangtran@hcmut.edu.vn