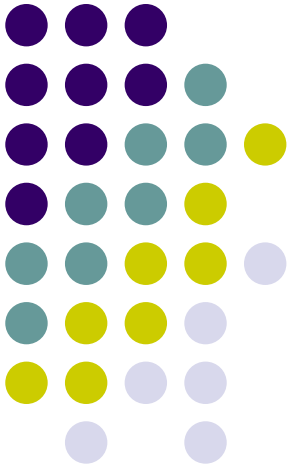
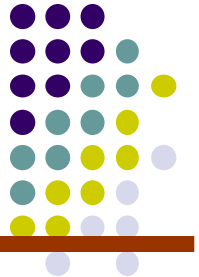
KHAI THÁC DỮ LIỆU & ỨNG DỤNG *(DATA MINING)*



GV : ThS.Lê Ngọc Thành

BÀI 1
TỔNG QUAN

NỘI DUNG

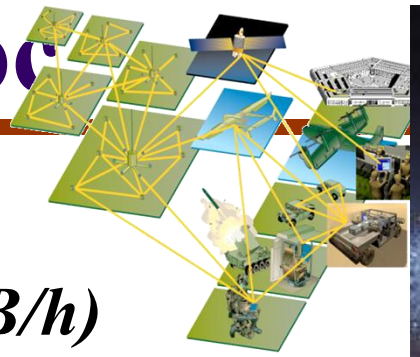


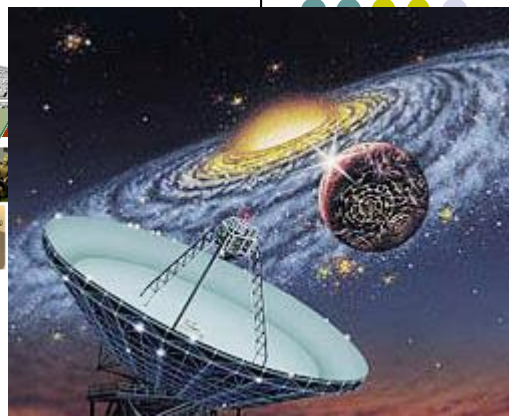
1. **Tại sao cần khai thác dữ liệu ?**
2. Khai thác dữ liệu (KTDL) là gì ?
3. Qui trình Khám phá tri thức (KDD)
4. Các nhiệm vụ chính của KTDL
5. Các kỹ thuật KTDL
6. Các thách thức của KTDL

SỰ CẦN THIẾT CỦA KTDL – Khía cạnh thương mại

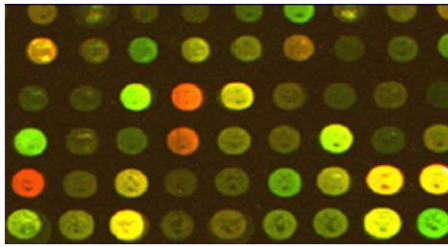
- ***Khối lượng lớn dữ liệu được thu thập và lưu trữ***
  - **Web data, e-commerce**
  - **Hóa đơn mua hàng tại siêu thị / trung tâm mua sắm**
  - **Giao dịch ngân hàng / thẻ tin dụng**
- ***Máy tính mạnh hơn, rẻ hơn***
- ***Áp lực cạnh tranh rất mạnh***
  - **Cung cấp các dịch vụ đa dạng, chất lượng tốt (CRM – Customer Relationship Management)**

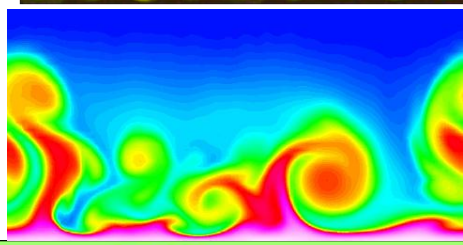
SỰ CẦN THIẾT CỦA KTDL – Khía cạnh Khoa học

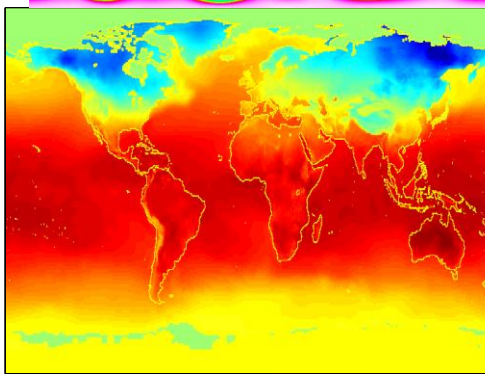




- ***Dữ liệu được thu thập và lưu trữ với tốc độ cao (GB/h)***
  - **Thiết bị remote sensor trên vệ tinh**
  - **Kính thiên văn quan sát bầu trời**
  - **Microarray tạo dữ liệu biểu diễn gien**
  - **Thử nghiệm khoa học tạo hàng TeraByte**
- ***Các kỹ thuật truyền thống không đủ khả năng làm việc với dữ liệu thô***
- ***KTDL có thể giúp các nhà khoa học***
  - **Phân loại và phân đoạn dữ liệu**
  - **Xây dựng giả thuyết**





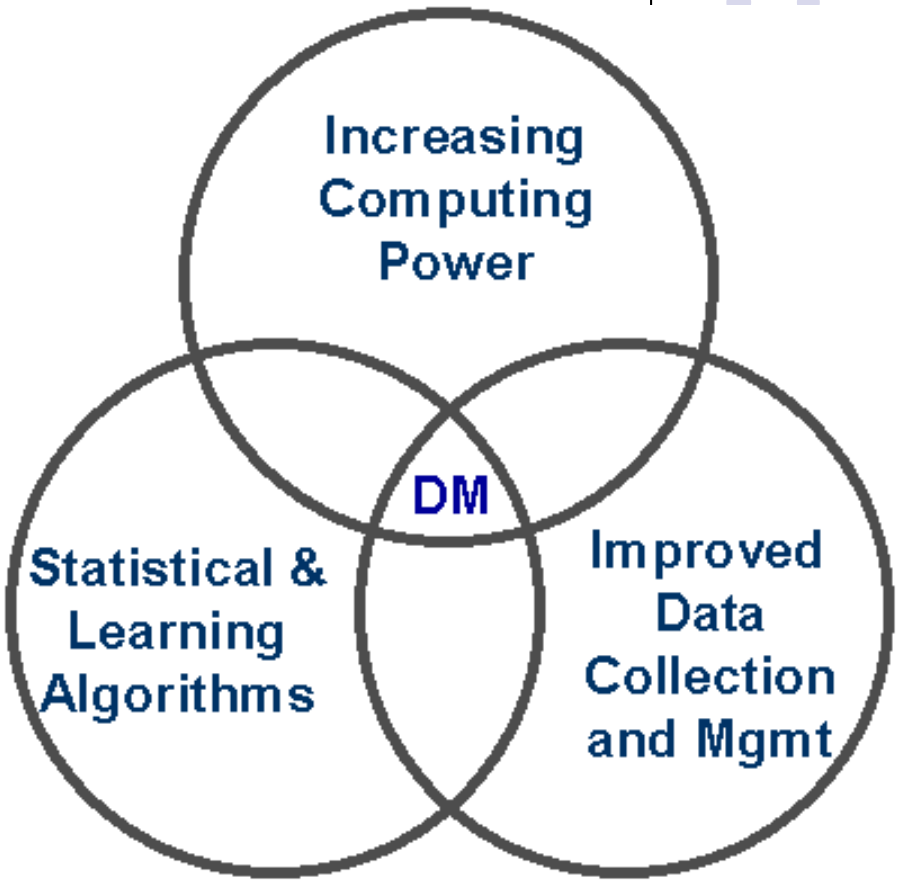


SỰ RA ĐỜI CỦA KTDL

- KTDL ra đời trong bối cảnh : GIÀU DL – NGHÈO TRI THỨC

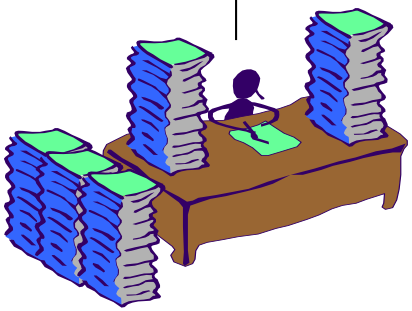
"We are drowning in data, but starving for knowledge!"

➢ **KTDL - giải pháp giúp phân tích tự động các núi DL và hỗ trợ ra quyết định .**



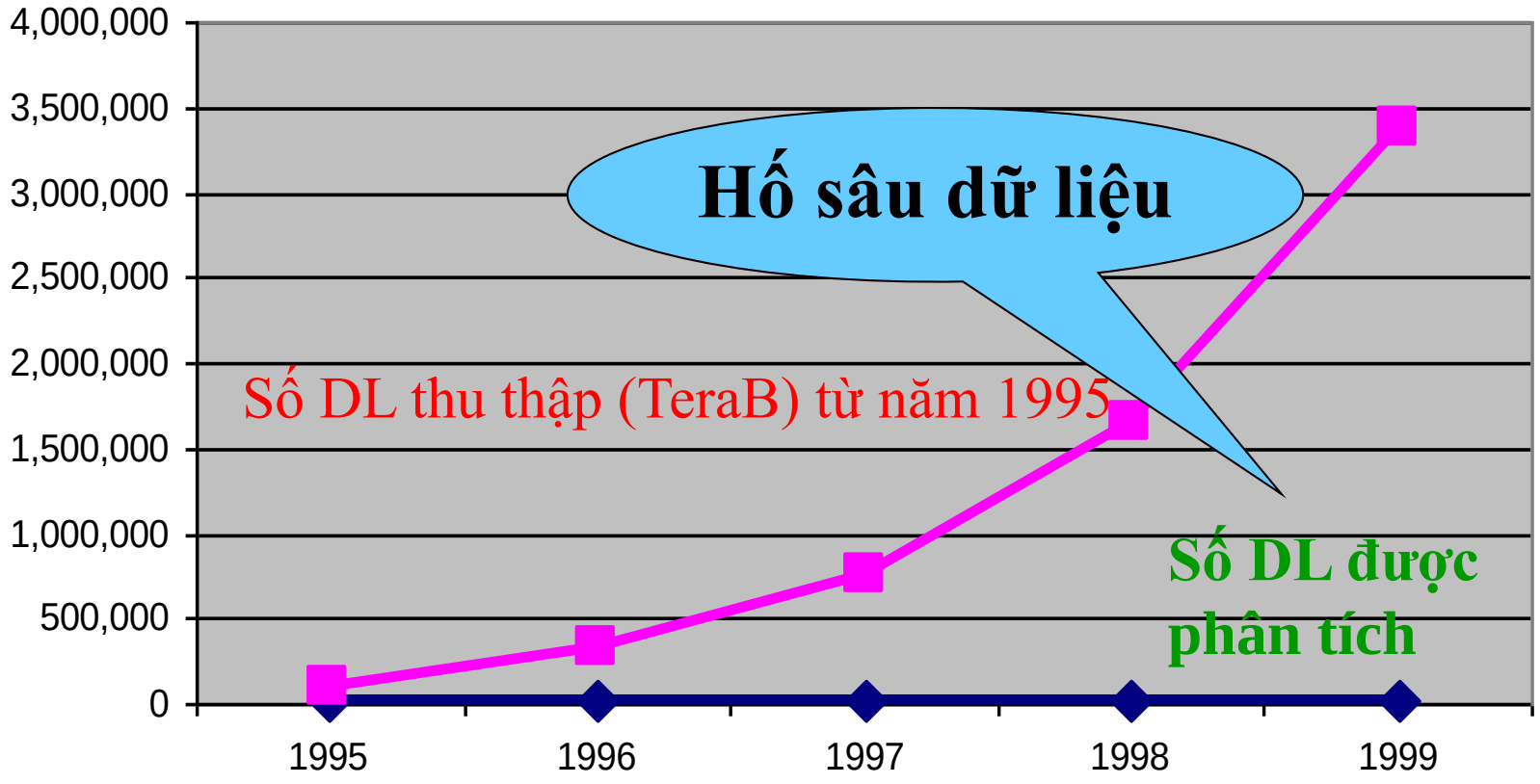
SỰ CẦN THIẾT CỦA KTDL

- *DL chứa rất nhiều thông tin giá trị, có lợi cho qui trình ra quyết định*
- Không thể phân tích DL = tay
  - Con người cần hàng tuần lễ để khám phá ra thông tin có ích
  - Phần lớn dữ liệu chưa bao giờ được phân tích cả
  - "Hố sâu giữa khả năng sinh ra DL và khả năng sử dụng DL" – Usama Fayyad



10^6-10^{12} bytes: Không bao giờ có thể nhìn thấy một cách đầy đủ tập dữ liệu hoặc đưa vào bộ nhớ của máy tính

SỰ CẦN THIẾT CỦA KTDL



SỰ DỤNG KTDL KHI NÀO?

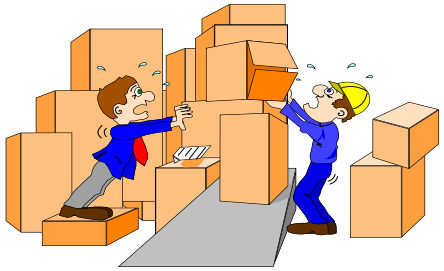
- Dữ liệu quá nhiều
- Dữ liệu lớn (chiều và kích thước)
  - Dữ liệu ảnh (kích thước)
  - Dữ liệu gene (số chiều)
- Có ít tri thức về dữ liệu

LĨNH VỰC ỨNG DỤNG KTDL

Thông tin thương mại

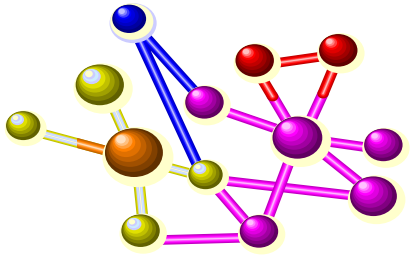
-Phân tích thị trường và mua bán
-Phân tích đầu tư
-Chấp thuận cho vay
-Phát hiện gian lận
…

Thông tin sản xuất



- Điều khiển và lên kế hoạch
- Quản trị mạng
- Phân tích các kết qủa thực nghiệm
…

Thông tin khoa học



- Thiên văn học
- Cơ sở dữ liệu sinh học
- Khoa học địa chất: bộ dò tìm động đất
…

Thông tin cá nhân

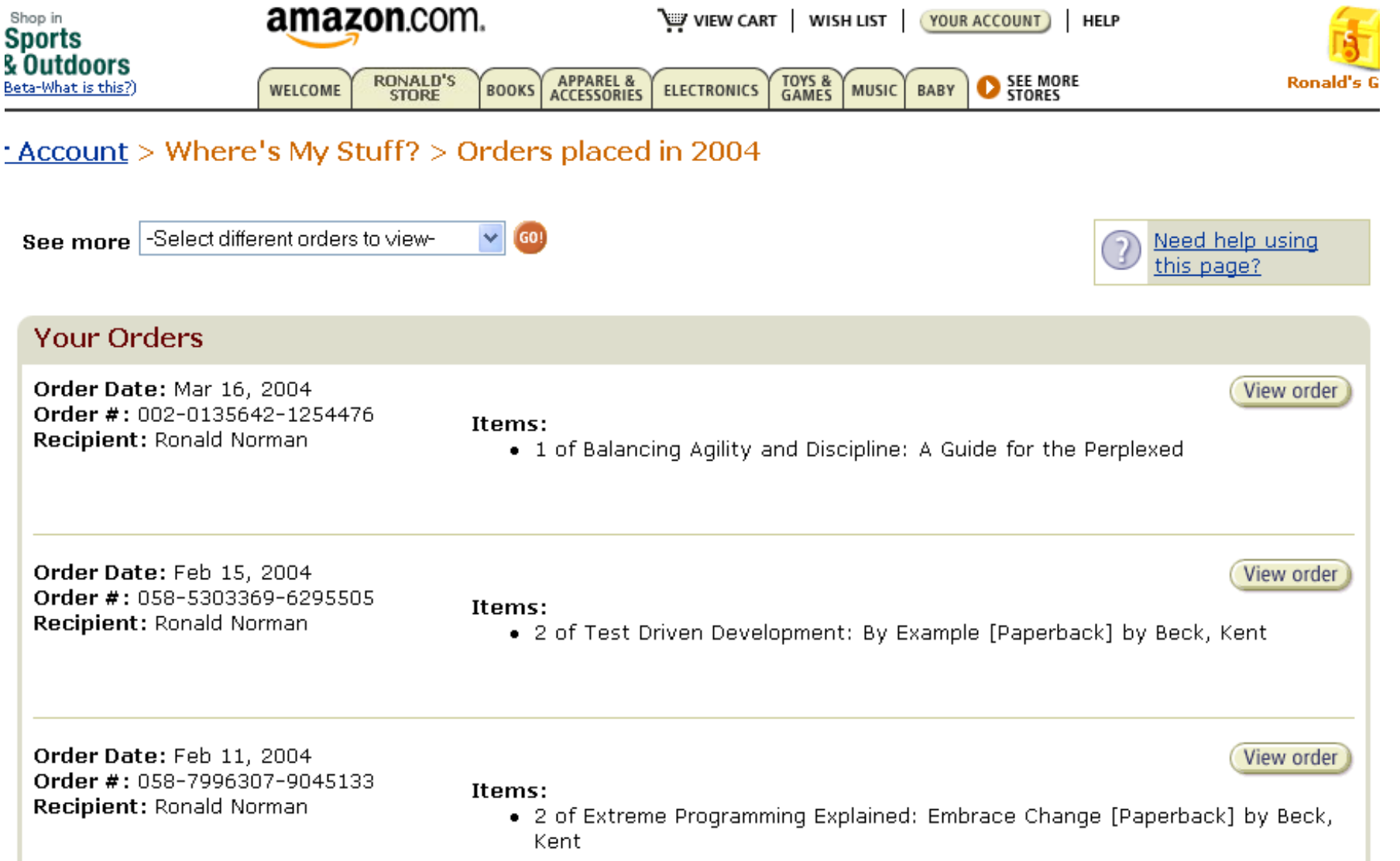
Customer Relationship Management (CRM)

Customer Relationship Management (CRM)

Để xây dựng mối quan hệ với khách hàng, các công ty cần phải biết :

1. **Notice** – what its customers are doing
2. **Remember** – what it and its customers have done over time
3. **Learn** – from what it has remembered
4. **Act On** – what it has learned to make customers more profitable

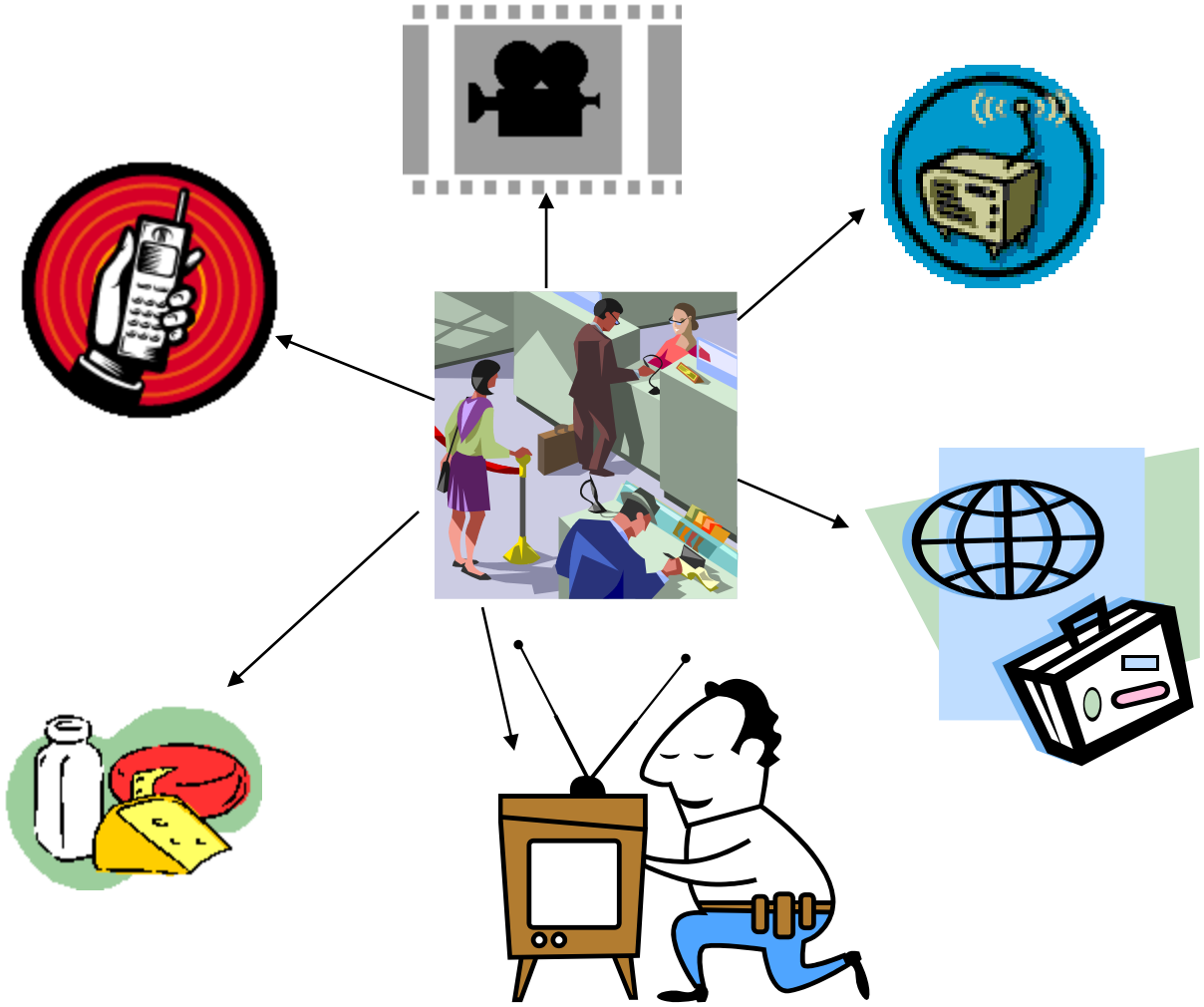
Dựa trên các dữ liệu giao dịch ("Transaction" Data)



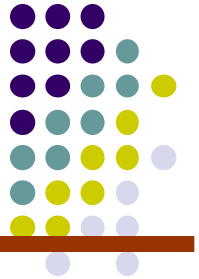
Dựa trên các dữ liệu giao dịch ("Transaction" Data)

	Date	Time	Rate	Minutes	Origination+		Phone number	Destination		Usage type	Call type	Airtime charges
1	05/10	12:12P	P	2	LA Mesa	CA	[redacted]	Voice Mail	CL	AR		Included
2	05/10	01:54P	P	1	LA Mesa	CA		San Diego	CA	MN		.00
3	05/10	01:55P	P	1	LA Mesa	CA		San Diego	CA	MN		.00
4	05/10	02:26P	P	8	Calexico	CA		Incoming	CL	MN		.00
5	05/10	02:59P	P	2	Calexico	CA		Mobile	CL	MN		.00
6	05/10	03:19P	P	1	Calexico	CA	[redacted]	Mobile	CL	MN		.00
7	05/10	04:07P	P	30	LA Mesa	CA		Incoming	CL	A		Included
8	05/11	11:06A	P	3	LA Mesa	CA		Incoming	CL	MN		.00
9	05/11	11:15A	P	1	San Diego	CA		LA Mesa	CA	A		Included
10	05/11	02:26P	P	1	Encinitas	CA		Voice Mail	CL	AR		Included
11	05/11	02:27P	P	2	Encinitas	CA	[redacted]	Chulavista	CA	A		Included
12	05/11	02:47P	P	3	San Diego	CA		Incoming	CL	MN		.00
13	05/11	08:31P	P	4	LA Mesa	CA		Rnchpnsqts	CA	A		Included
14	05/12	11:17A	P	8	LA Mesa	CA		Incoming	CL	MN		.00
15	05/12	11:33A	P	2	LA Mesa	CA		Mobile	CL	MN		.00

Phát hiện và nắm giữ mối quan hệ là chìa khoá của thành công



NỘI DUNG



1. Tại sao cần khai thác dữ liệu ?
2. **Khai thác dữ liệu là gì ?**
3. Qui trình KDD
4. Các nhiệm vụ chính của KTDL
5. Các kỹ thuật KTDL
6. Các thách thức của KTDL

THẾ NÀO LÀ KTDL

"**Khai thác dữ liệu** *là quá trình không tầm thường của việc xác định các* **mẫu tiềm ẩn** *có tính hợp lệ, mới lạ, có ích và có thể hiểu được tối đa trong CSDL"* – U.Fayyad, …(1996)

Đa xử lý

Quá trình không tầm thường

Chứng minh tính đúng
Của mẫu / Mô hình

Hợp lệ

Mới lạ

Không biết trước

Có ích

Có thể sử dụng được

Có thể hiểu được

Bởi con người và máy

KHAI THÁC DL …

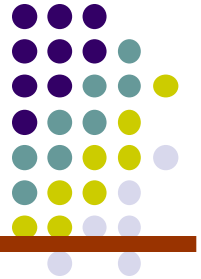
- Thế nào là **mẫu tiềm ẩn** ?
  - Là mối quan hệ trong dữ liệu ví dụ như :
  - Những người mua quần tây thường hay mua thêm áo sơ mi
  - *Những người có mức tín dụng tốt thì thường ít bị tai nạn.*
  - Đàn ông, 37+, thu nhập : 50K-75K, -> chi khoảng 25-50 cho đặt mua hàng qua catalog

KHAI THÁC DL ....

- **What is not Data Mining?**
  - Tìm số điện thoại trong danh bạ điện thoại
  - Tìm thông tin về "Amazon" trên serach engine

- **What is Data Mining?**
  - Các tên phổ biến tại khu vực xác định của Mỹ (O'Brien, O'Rurke, O'Reilly… ở vùng Boston)
  - Gom nhóm các tài liệu giống nhau thu được từ search engine dựa trên nội dung (VD: rừng nhiệt đới Amazon , Amazon.com)

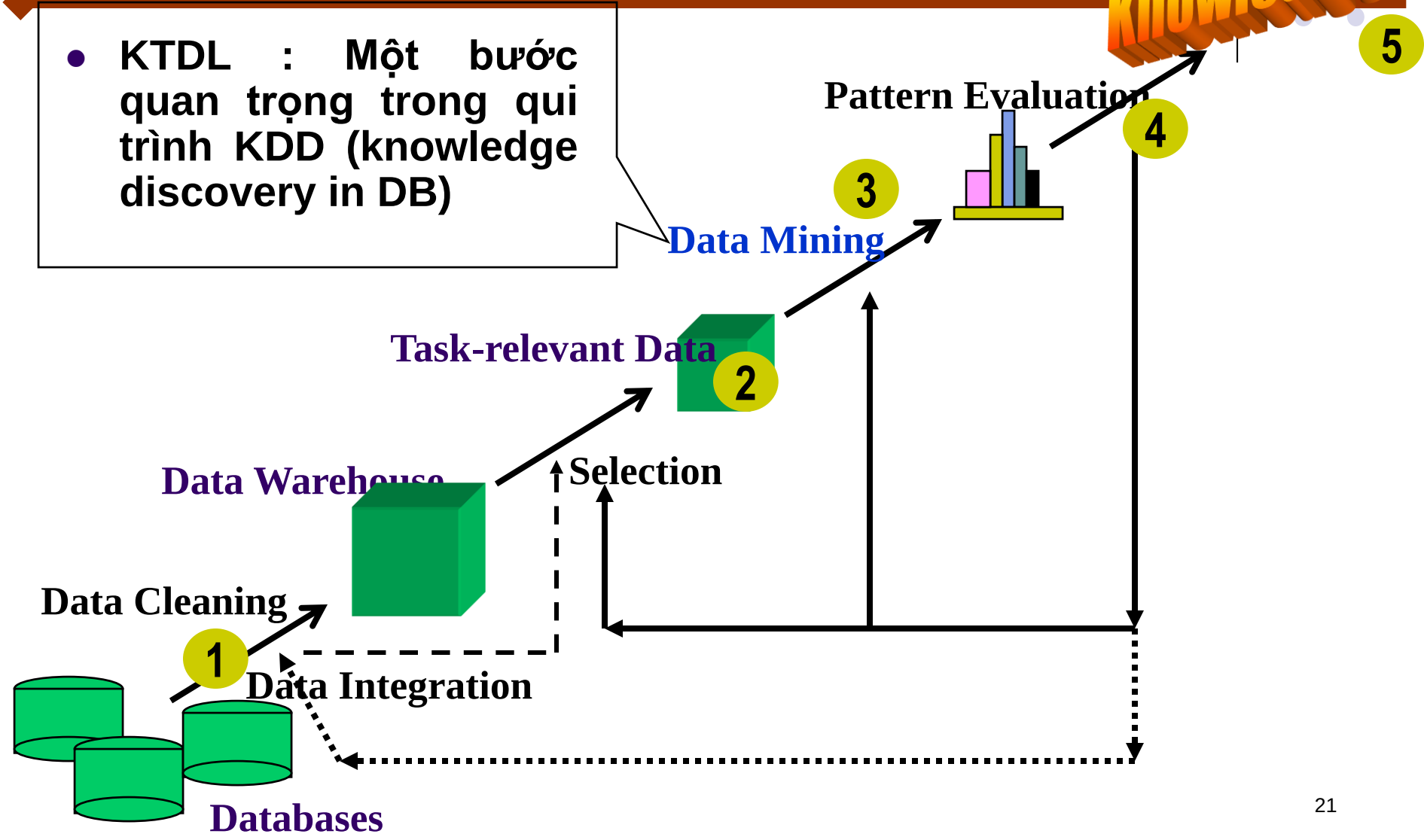
NỘI DUNG



1. Tại sao cần khai thác dữ liệu ?
2. Khai thác dữ liệu là gì ?
3. **Qui trình Khám phá tri thức (KDD)**
4. Các nhiệm vụ chính của KTDL
5. Các kỹ thuật KTDL
6. Các thách thức của KTDL

QUI TRÌNH KHÁM PHÁ TRI THỨC

- KTDL : Một bước quan trọng trong qui trình KDD (knowledge discovery in DB)



QUI TRÌNH KDD

Dữ liệu được tổ chức theo chức năng

Tạo ra/chọn lọc CSDL đích

Data warehousing

Chọn lựa kỹ thuật điển hình và dữ liệu mẫu

1

Thay thế những giá trị thiếu

Khử nhiễu Dữ liệu

2

Chuẩn hoá giá trị

Biến đổi giá trị

Tạo các thuộc Tính dẫn xuất

Tìm thuộc tính quan trọng &Miền giá trị

3

4

Lựa chọn nhiệm vụ DM

Lựa chọn phương pháp DM

Trích xuất Tri thức

Kiểm tra tri thức

Tính chế Tri thức

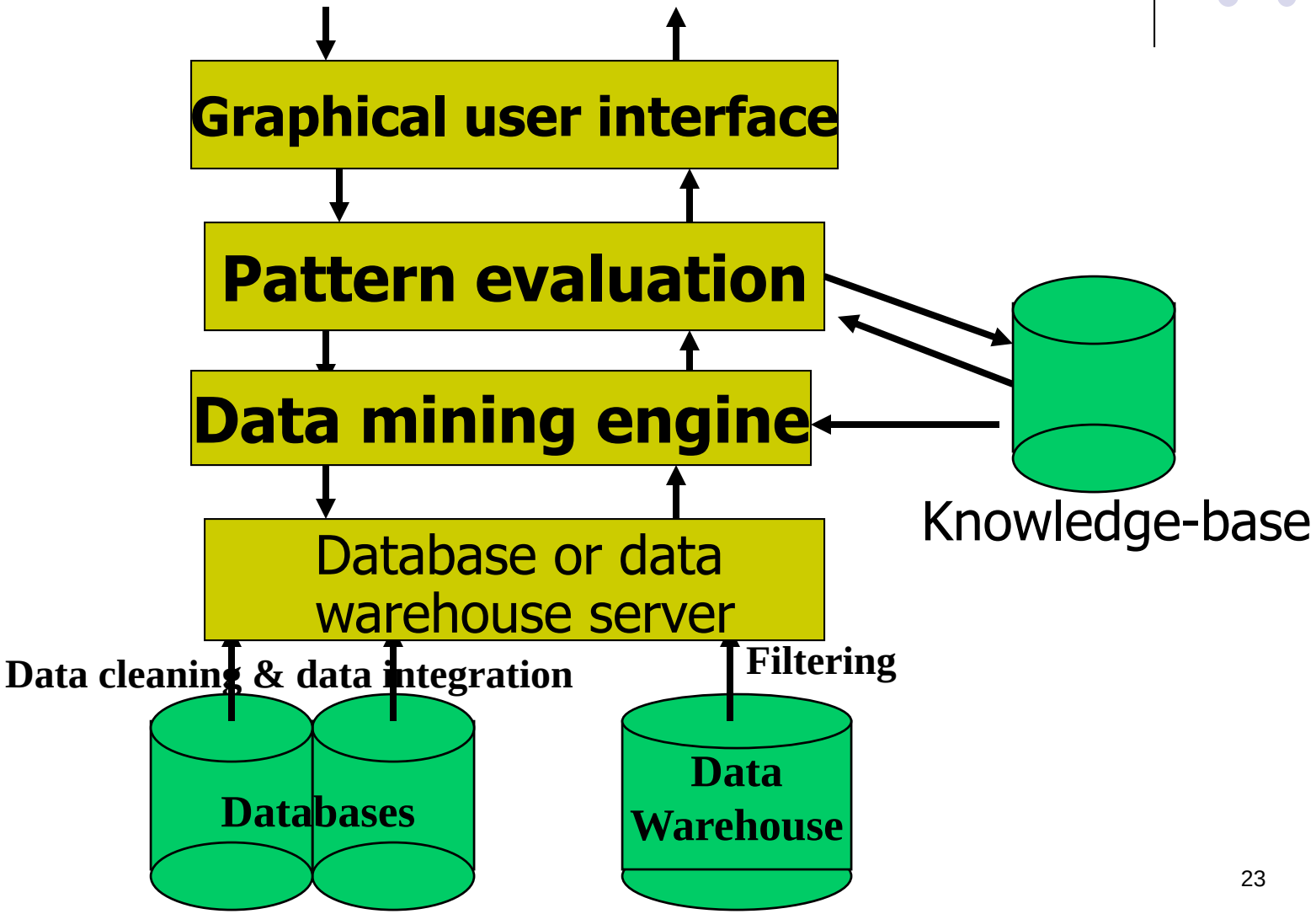
Phát sinh ra câu hỏi và báo cáo

Các phương pháp cải tiến kiểu kết hợp và lập dãy

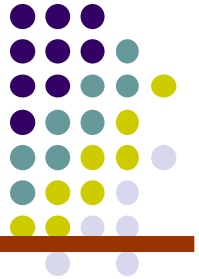
5

Biến đổi qua biểu diễn khác

KIẾN TRÚC HỆ THỐNG KTDL TIÊU BIỂU

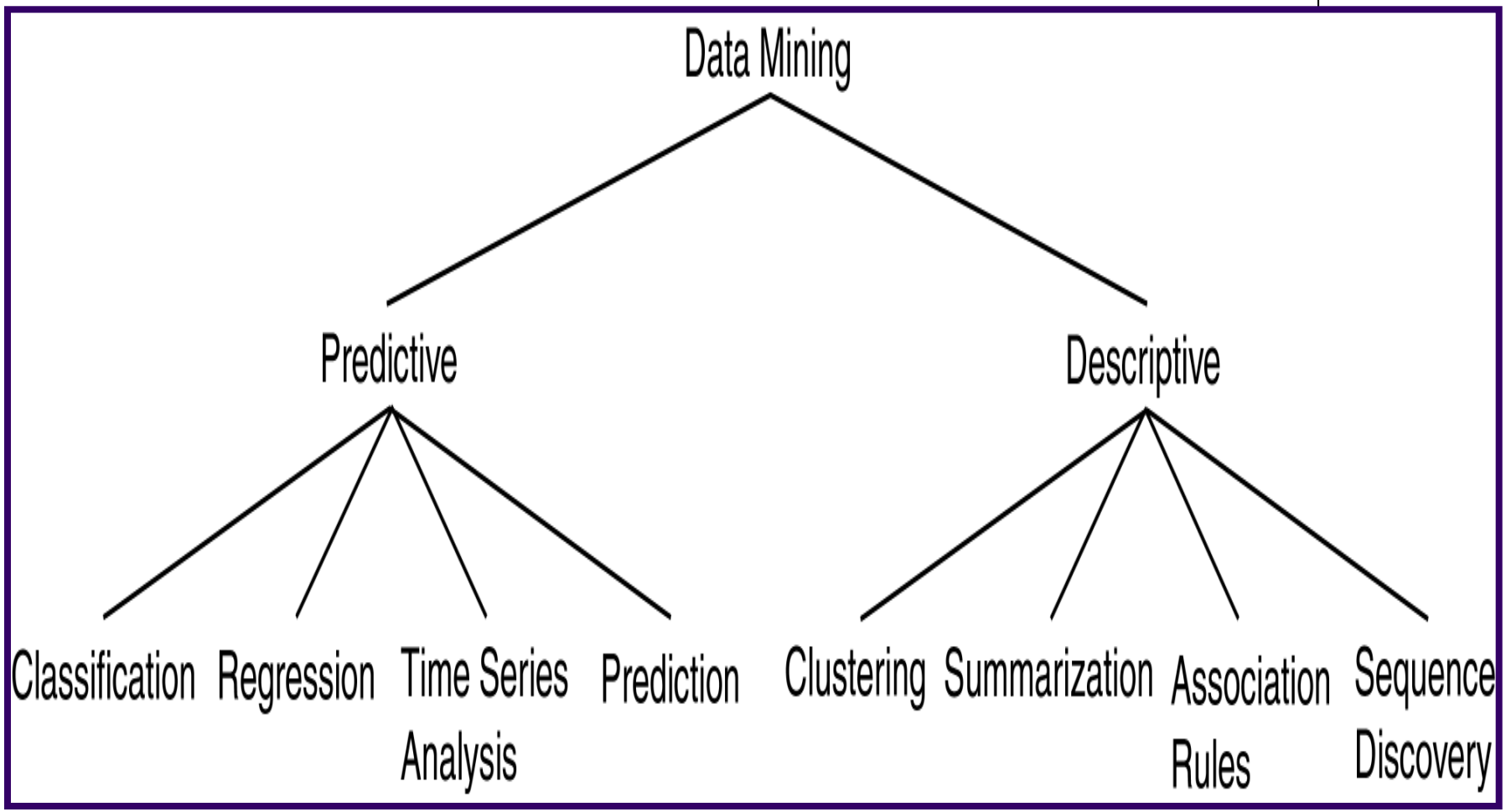


NỘI DUNG



1. Tại sao cần khai thác dữ liệu ?
2. Khai thác dữ liệu là gì ?
3. Qui trình khám phá tri thức (KDD)
4. **Các nhiệm vụ chính của KTDL**
5. Các kỹ thuật KTDL
6. Các thách thức của KTDL

CÁC NHIỆM VỤ CHÍNH CỦA KTDL



CÁC NHIỆM VỤ CHÍNH CỦA KTDL

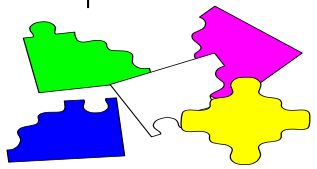
- **Dự đoán (Predictive) :**
  - *Sử dụng một vài biến để dự báo giá trị chưa biết hoặc giá trị tương lai của các biến khác*
    - *Phân lớp*
    - *Hồi qui*
    - *Phát hiện sự thay đổi /lạc hướng*
- **Mô tả (Descriptive) :**
  - *Xác định các mẫu mô tả DL mà con người có thể hiểu được*
    - *Gom cụm*
    - *Tóm tắt*
    - *Mô hình hóa phụ thuộc*

CÁC NHIỆM VỤ CHÍNH CỦA KTDL

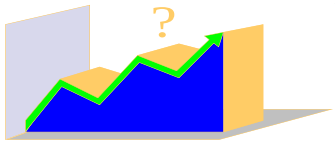
Phát hiện ra mô tả của một vài lớp đã được xác định và phân loại dữ liệu vào một trong các lớp đó.

Phân lớp

Tìm ra một tập xác định Các nhóm hay các cụm để mô tả dữ liệu



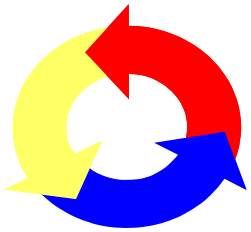
Gom cụm



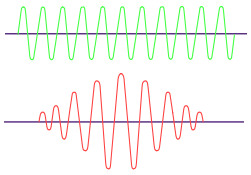
Ánh xạ từ một mẫu dữ liệu thành một biến dự đoán trước có giá trị thực .

Hồi qui

Phát hiện ra một mô hình mà mô tả phụ thuộc quan trọng nhất giữa các biến



Mô hình hóa phụ thuộc



Phát hiện ra những thay đổi quan trọng nhất trong dữ liệu

Phát hiện sự thay đổi/lạc hướng

Phát hiện ra một mô tả tóm tắt cho một tập con dữ liệu

Tóm tắt

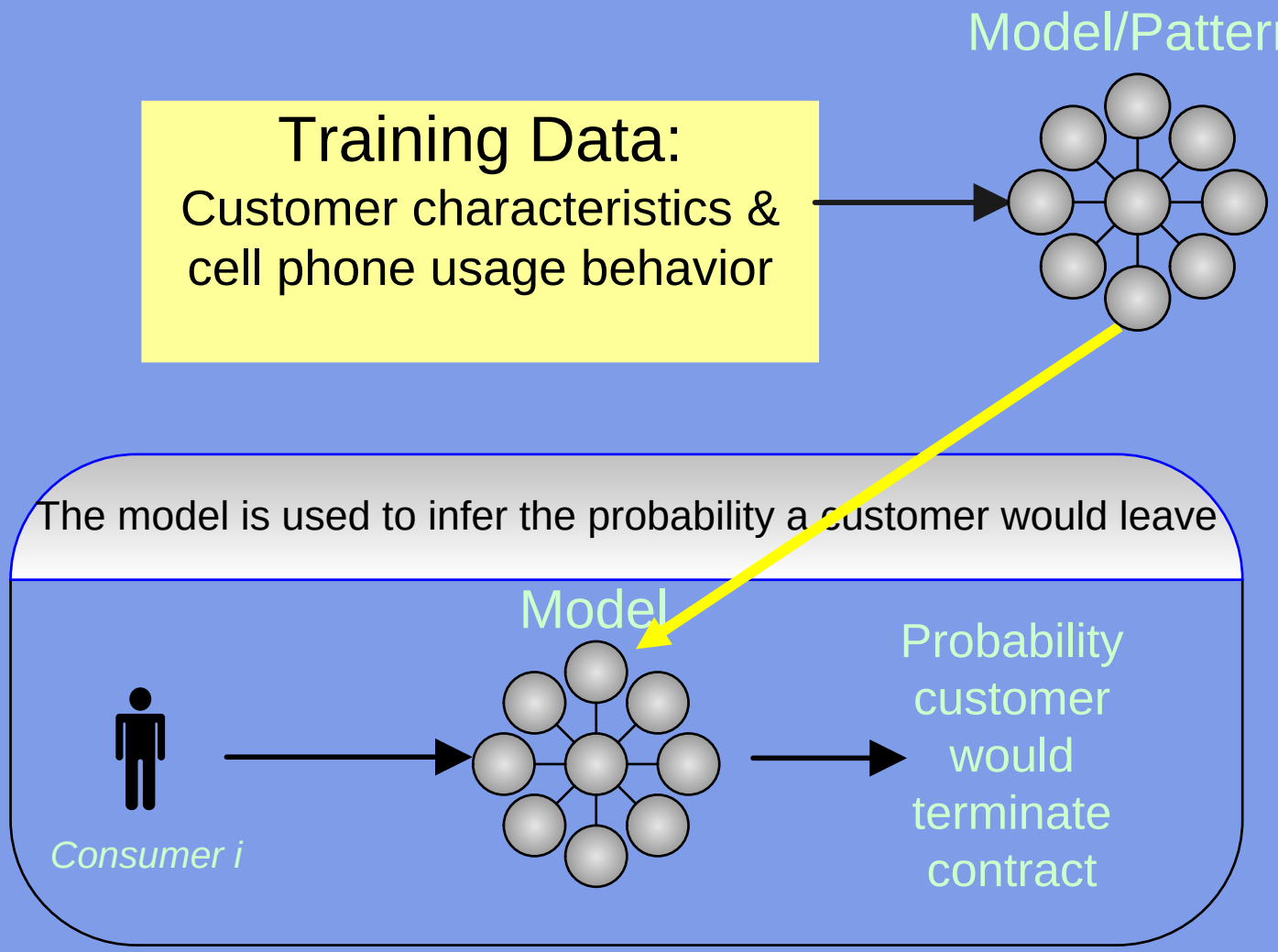
VÍ DỤ PHÂN LỚP

- **Công ty Verizon Wireless** :
  - Công ty cung cấp thiết bị, dịch vụ không dây lớn nhất ở Mỹ. www.verizonwireless.com
  - *Số lượng khách hàng : 65.7 triệu (cuối năm 2007)*
  - Thu nhập hằng năm: 43.9 tỷ $
- **Vấn đề :**
  - Tỷ lệ khách hàng bị mất cao : 2%/tháng (1,300,000 khách hàng rời bỏ/tháng)
  - *Chi phí thay thế : hàng trăm triệu $/năm*
  - Chi phí trung bình cho mỗi khách hàng mới : 320$

VÍ DỤ PHÂN LỚP

- **Giải pháp thông thường :**
  - Chào mời, khuyến mãi ***tất cả khách hàng*** trước khi hết hợp đồng
    - Chí phí quá tốn kém, lãng phí
- **Giải pháp của KTDL :**
  - **Xây dựng mô hình dự đoán**
    - Dùng mô hình dự đoán để xác định các khách hàng có khả năng rời bỏ
  - **Sau đó :**
    - Khuyến mãi, chào mời (VD: một điện thoại mới) cho những khách hàng có nhiều khả năng rời bỏ nhất
    - Phát triển kế hoạch mới nhằm đáp ứng nhu cầu của khách hàng
  - **Kết quả : giảm tỷ lệ mất khách hàng dưới 1.5 %/ tháng**

VÍ DỤ PHÂN LỚP



PHÂN LỚP: ỨNG DỤNG 1

- **Phát hiện gian lận :**
  - ***Mục đích*** : Dự đoán các trường hợp gian lận trong giao dịch thẻ tín dụng
  - ***Hướng giải quyết*** :
    - Dùng các giao dịch thẻ tín dụng và thông tin của chủ thẻ như thuộc tính
      - *Khách hàng mua cái gì, lúc nào, số lần dùng thẻ*
    - Gán nhãn giao dịch cũ là gian lận hay hợp lý, đúng - tạo thành thuộc tính lớp
    - Xây dựng mô hình cho lớp các giao dịch
    - *Dùng mô hình để khám phá gian lận trên các giao dịch thẻ tín dụng*

PHÂN LỚP: ỨNG DỤNG 2

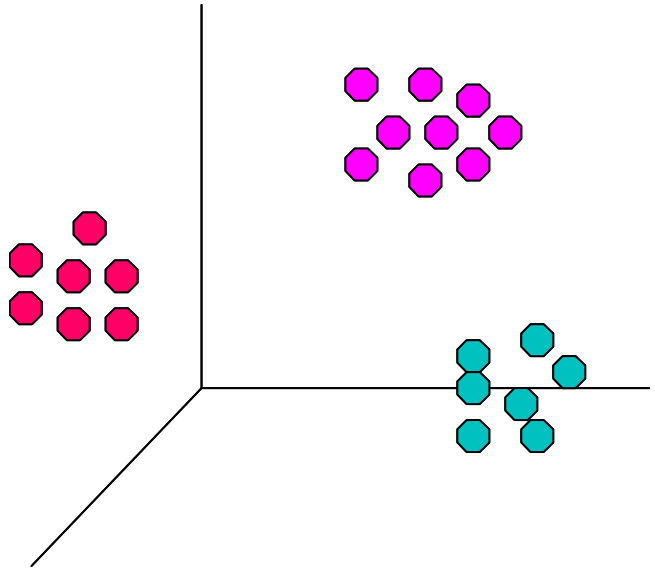
- **Quảng cáo :**
  - ***Mục đích*** : Giảm chí phí thư tín bằng cách tập trung vào nhóm khách hàng có nhiều khả năng mua sản phẩm điện thoại di động mới
  - ***Hướng giải quyết*** :
    - Sử dụng dữ liệu cho sản phẩm tương tự trước đây
    - Dùng quyết định {mua, không mua} làm thuộc tính lớp
    - Thu thập thông tin cá nhân, cách sống và quan hệ của tất cả các khách hàng
    - Dùng các thông tin trên như là dữ liệu đầu vào để xây dựng mô hình phân lớp

GOM CỤM : Minh họa

- Gom cụm dựa trên khoảng cách Euclide trong không gian 3-D

Intracluster distances are minimized

Intercluster distances are maximized



GOM CỤM : ỨNG DỤNG 1

- **Gom nhóm khách hàng :**
  - ***Mục đích*** : Chia khách hàng thành các nhóm/cụm riêng biệt để có thể áp dụng các biện pháp quảng cáo khác nhau
  - ***Hướng giải quyết :***
    - Thu thập thông tin cá nhân, cách sống của tất cả các khách hàng
    - *Xác định các cụm/nhóm khách hàng giống nhau*
    - Kiểm tra chất lượng của các cụm thông qua việc quan sát đặc trưng mua hàng của khách hàng trong cùng một cụm so với khách hàng khác cụm

GOM CỤM : ỨNG DỤNG 2

- **Gom cụm tài liệu :**
  - ***Mục đích*** : Tìm nhóm tài liệu giống nhau dựa trên các từ quan trọng
  - ***Hướng giải quyết*** :
    - Xác định độ phổ biến của từ trong tài liệu. Xây dựng độ đo tương tự dựa trên độ phổ biến của các từ để gom cụm.
    - ***Lợi ích*** : *Trong lĩnh vực truy vấn thông tin (IR), có thể dùng các cụm để liên kết tài liệu mới với các tài liệu đã gom cụm*

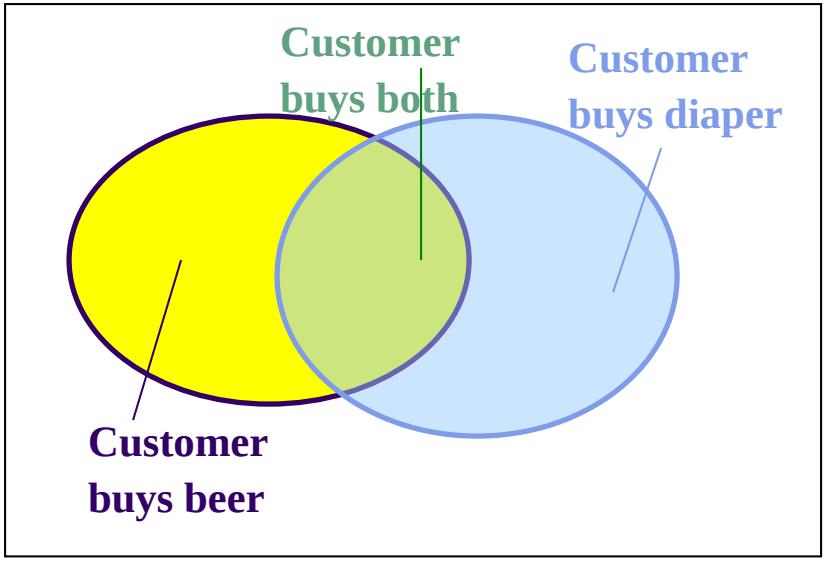
Gom cụm DL cổ phiếu S&P 500

- Quan sát sự biến động của giá cổ phiếu hàng ngày
- *Dữ liệu : Cổ phiếu – {UP/DOWN}*
- Độ đo tương tự : các sự kiện thường giống nhau trong cùng một ngày

	Discovered Clusters	***Industry Group***
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

KHAI THÁC LUẬT KẾT HỢP

Transaction-id	Items bought
10	A, B, C
20	A, C
30	A, D
40	B, E, F

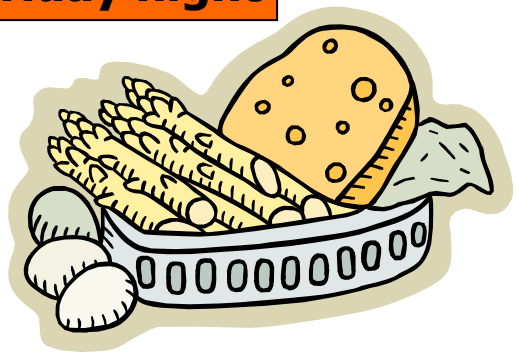


- Itemset $X=\{x_1, \ldots, x_k\}$
- Tìm mối quan hệ giữa các thuộc tính thường xuất hiện đồng thời

$A \rightarrow C$ (50%, 66.7%)

$C \rightarrow A$ (50%, 100%)

Buy diapers on Friday night → **Then** → **Buy beer**



Khai thác LKH : ỨNG DỤNG 1

- **Quản lý quầy hàng siêu thị:**
  - ***Mục đích*** : Xác định những mặt hàng được nhiều khách hàng mua chung
  - ***Hướng giải quyết*** :
    - Xử lý dữ liệu bán hàng để tìm mối liên hệ giữa các mặt hàng
    - Luật cổ điển : Nếu khách hàng mua tã giấy và sữa thì có khả năng mua bia.

Khai thác LKH : ỨNG DỤNG 2

- **Quản lý hàng hóa:**
  - ***Mục đích*** : Công ty bảo trì thiết bị tiêu dùng muốn đoán trước nguyên nhân sửa chữa các sản phẩm tiêu dùng và trang bị các xe bảo trì các bộ phận cần thiết để giảm thiểu số lần đến nhà khách hàng
  - ***Hướng giải quyết*** :
    - Xử lý dữ liệu trên các dụng cụ và bộ phận đã yêu cầu trong các lần sửa trước để tìm các mẫu đồng xuất hiện.

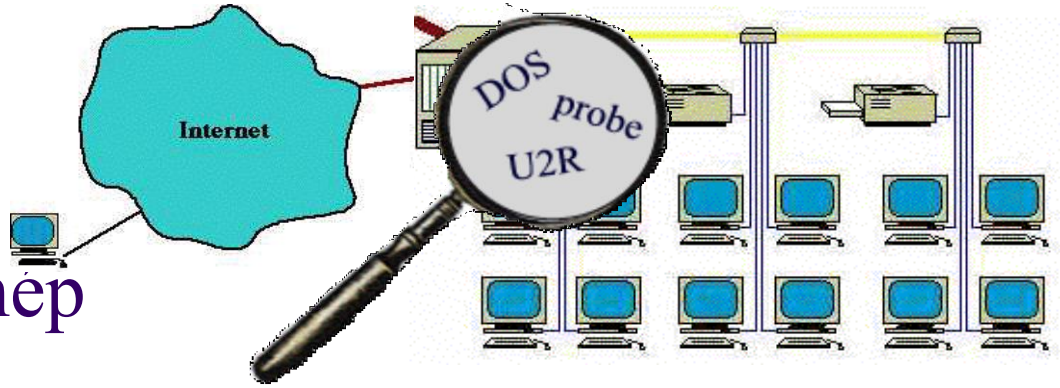
HỒI QUI

- Dự đoán giá trị của biến *dựa trên giá trị của các biến khác*
- **Ví dụ :**
  - Dự báo khối lượng bán hàng của sản phẩm mới dựa trên chi phí quảng cáo
  - *Dự đoán tốc độ gió như một hàm của nhiệt độ, độ ẩm, áp suất không khí, ...*
  - Dự đoán chỉ số thị trường chứng khoán

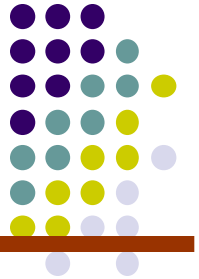
Phát hiện sự Lạc hướng/ Bất bình thường

- Xác định sự lệch hướng rõ rệt so với hành vi thông thường
- **Ứng dụng :**
  - Phát hiện gian lận thẻ tín dụng

  - Phát hiện xâm nhập mạng trái phép



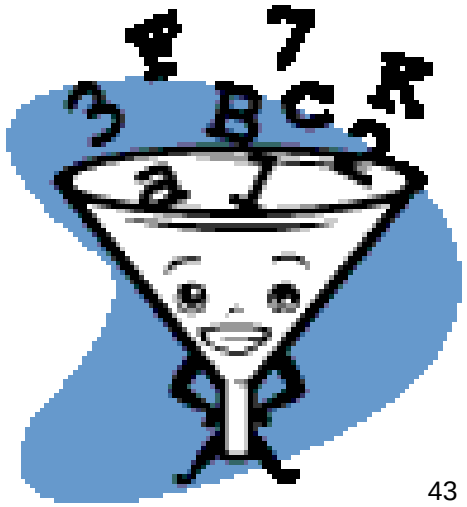
NỘI DUNG



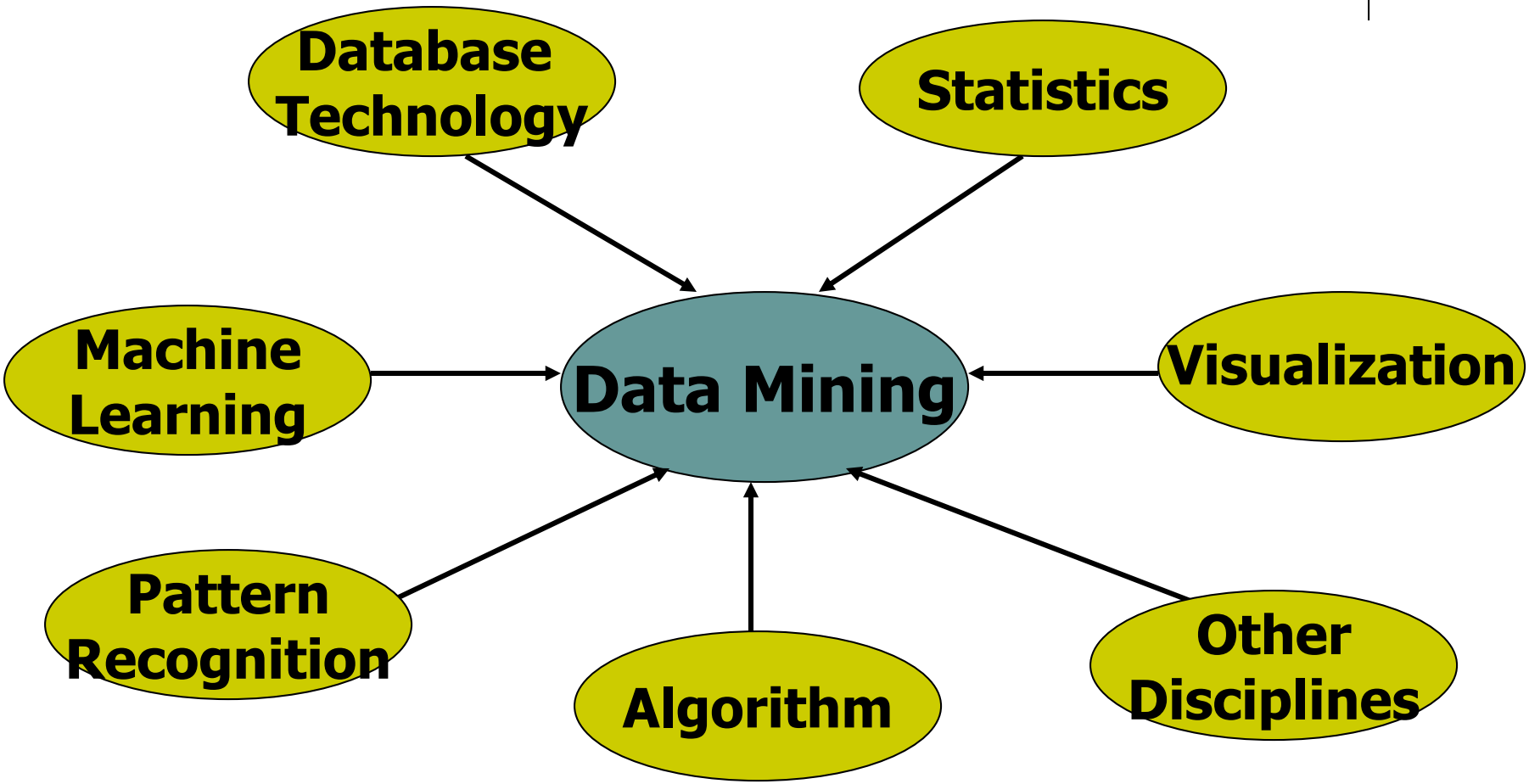
1. Tại sao cần khai thác dữ liệu ?
2. Khai thác dữ liệu là gì ?
3. Qui trình Khám phá tri thức (KDD)
4. Các nhiệm vụ chính của KTDL
5. **Các kỹ thuật KTDL**
6. Các thách thức của KTDL

CÁC KỸ THUẬT KTDL

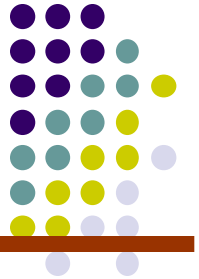
- KTDL lấy ý tưởng từ các lĩnh vực như máy học, thống kê, nhận dạng, hệ thống DL…
- Các kỹ thuật truyền thống có thể không phù hợp do :
  - Kích thước lớn của DL
  - Số chiều DL lớn
  - Bản chất DL không đồng nhất



KTDL – KẾT HỢP CÁC PHƯƠNG PHÁP

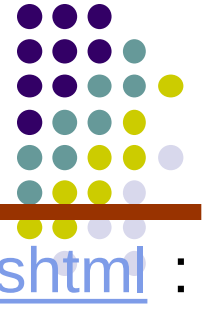


NỘI DUNG



1. Tại sao cần khai thác dữ liệu (DM) ?
2. DM là gì ?
3. Qui trình KDD
4. Các nhiệm vụ chính của KTDL
5. Các kỹ thuật KTDL
6. **Các thách thức của KTDL**

CÁC THÁCH THỨC CỦA KTDL



Nguồn : http://www.cs.uvm.edu/~icdm/10Problems/index.shtml : 2005-2006 của ICDM

- Developing a Unifying Theory of Data Mining
- Scaling Up for High Dimensional Data and High Speed Data Streams
- Mining Sequence Data and Time Series Data
- Mining Complex Knowledge from Complex Data
- Data Mining in a Network Setting
- Distributed Data Mining and Mining Multi-agent Data
- Data Mining for Biological and Environmental Problems
- Data-Mining-Process Related Problems
- Security, Privacy and Data Integrity
- Dealing with Non-static, Unbalanced and Cost-sensitive Data

TẠI SAO CẦN NGHIÊN CỨU KTDL

Các nhóm thảo luận và tự đưa ra câu trả lời.

TÓM TẮT

- Khám phá mẫu có ích, chưa biết từ khối lượng lớn DL
- Qui trình khám phá tri thức (KDD)
  - Thu thập và tiền xử lý DL -> KTDL -> Đánh giá mẫu -> Biểu diễn tri thức
- Khai thác trên nhiều loại DL, thông tin
- Các loại mẫu cần khai thác
  - Luật kết hợp, mẫu tuần tự, phân lớp, gom nhóm, mẫu hiếm, mẫu cá biệt, sai lệch

Sự phát triển của KTDL

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences từ 1998 và SIGKDD Explorations
- Nhiều hội nghị khác về KTDL
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), …
- ACM Transactions on KDD từ 2007

Bài tập theo nhóm số 1

- **Thời gian thảo luận : 15’**
  - Thảo luận tình huống KTDL trong nhóm và 01 người đại diện cho nhóm trình bày.
  - *Thời gian trình bày : tối đa 3’ .*
    - *Trình bày tình huống*
    - *Hướng giải quyết và lợi ích*
  - **Tình huống 1 : Thị trường bán lẻ (ví dụ cần tăng doanh thu bán hàng)**
  - ***Nhóm :***
  - **Gợi ý :**
    - **Dạng DL nào được thu thập . Sử dụng nhiệm vụ nào của KTDL ?**
    - **Các thông tin nào ta cần biết về khách hàng**
    - **Có cần biết khách hàng mua các mặt hàng gì?**
    - **Có cần phân loại khách hàng ?,…**

Bài tập theo nhóm số 1

- **Thời gian : 15’**
  - Thảo luận tình huống KTDL trong nhóm và 01 người đại diện cho nhóm trình bày
  - *Thời gian trình bày : tối đa 3’*
    - *Trình bày tình huống*
    - *Hướng giải quyết và lợi ích*
  - **Tình huống 2 : Quảng cáo sản phẩm (ví dụ chọn lựa hình thức, đối tượng quảng cáo để giảm chi phí, tăng lợi nhuận)**
  - ***Nhóm :***
  - Gợi ý :
    - DL cần thu thập là gì. Sử dụng nhiệm vụ nào của KTDL ?
    - Có cần thiết gửi tờ quảng cáo sản phẩm đến tất cả các khách hàng Hay chỉ gửi cho 1 nhóm có chọn lọc.
    - Có thể dự kiến khả năng phản hồi của khách hàng so với chi phí gửi quảng cáo ?

CÁC CÔNG VIỆC CẦN LÀM

1. ***Post bài tập nhóm số 1***
   - ***Tất cả các nhóm*** sẽ post kết quả thảo luận nhóm lên website môn học (trong mục diễn đàn thảo luận)
   - **Hạn chót post : 23h00 – 1/08/2011**
2. ***Chuẩn bị bài 2 : Qui trình chuẩn bị DL***
   - **Xem nội dung bài tập nhóm chương 2: các vấn đề khi làm việc với DL thực tế.**
   - ***Cách thực hiện :***
     - ***Nghiên cứu slide, xem ví dụ.***
     - ***Tham khảo trên Internet và tài liệu tham khảo***

TÀI LIỆU THAM KHẢO

- G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. *From data mining to knowledge discovery: An overview*. U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996
- http://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u : bách khoa toàn thư mở wikipedia
- J.Han, M.Kamber, Chương 1 – Data mining : Concepts and Techniques
- P.-N. Tan, M. Steinbach, V. Kumar, Chương 1 - Introduction to Data Mining

BÀI TẬP

1. Thế nào là khai thác dữ liệu ? Cho ví dụ minh họa.
2. Các kiểu dữ liệu, thông tin nào có khả năng được sử dụng trong qui trình KDD?
3. Cho ví dụ thực tế về việc áp dụng KTDL đem đến thành công trong kinh doanh (*ngoài các ví dụ có trong bài giảng*).
   - **Gợi ý** : Bài toán tăng doanh thu của thị trường bán lẻ. Bài toán xây dựng kế hoạch quảng cáo và khuyến mãi
   - Loại DL nào được thu thập ? Loại nhiệm vụ nào của KTDL được sử dụng ? Có thể thay bằng phương pháp truy vấn DL hay phân tích thống kê đơn giản không ?