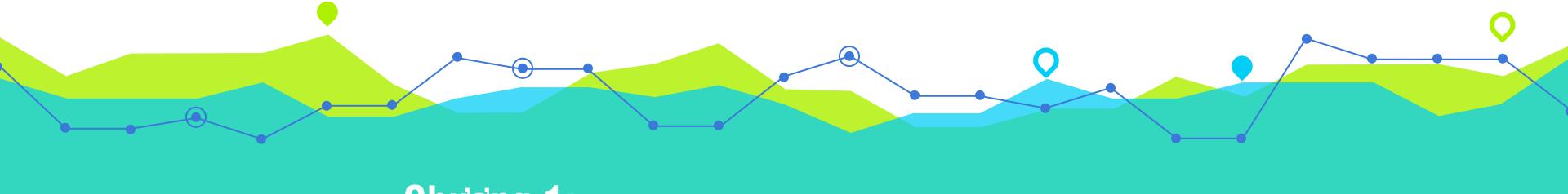


**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN  
KHOA HỆ THỐNG THÔNG TIN**

*Tài liệu bài giảng:*

# **KHAI THÁC DỮ LIỆU – IS252**



Chương 1:  
**TỔNG QUAN VỀ KTDL**

ThS. Dương Phi Long – Email: longdp@uit.edu.vn

# NỘI DUNG BÀI HỌC

01



Khám phá tri thức từ CSDL

02



Quá trình Khai thác Dữ liệu

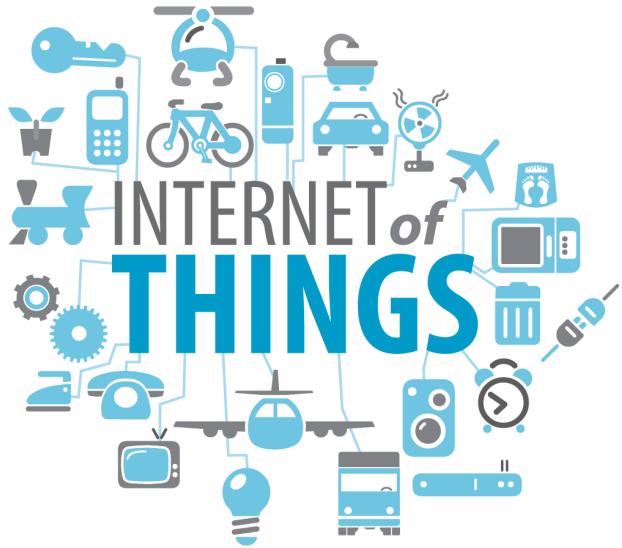
03



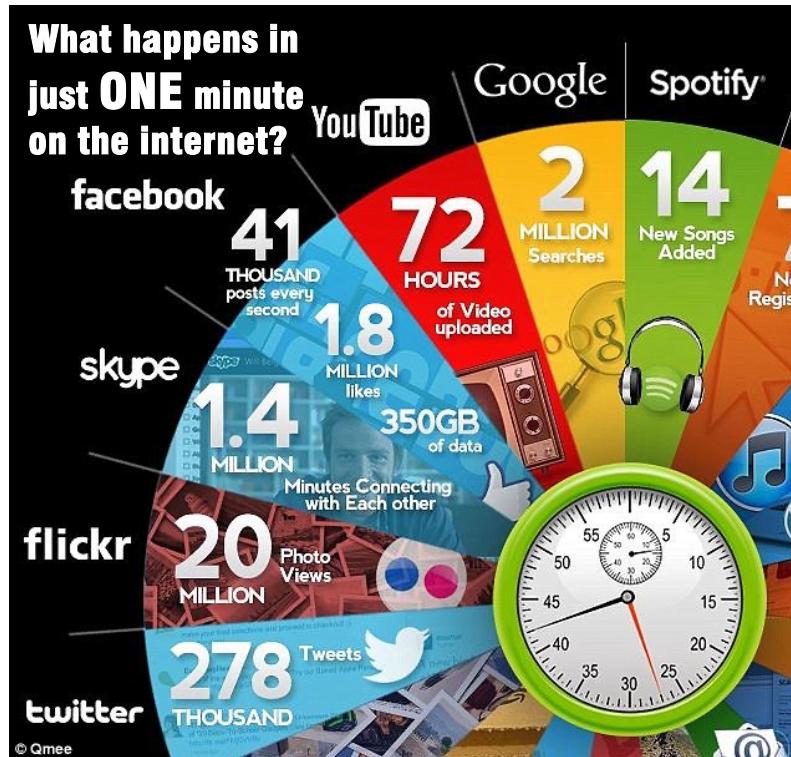
Ứng dụng và thách thức của KTDL



# Đặt vấn đề



THE EXPLOSIVE  
GROWTH OF DATA



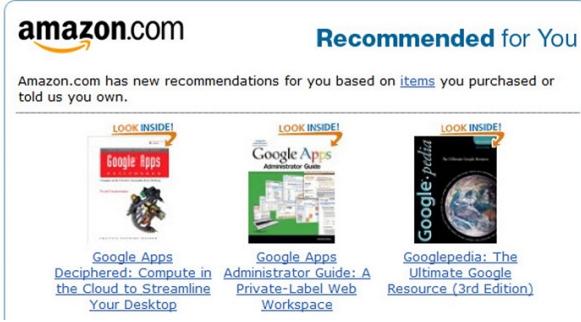
# Đặt vấn đề

# WHY DATA MINING?



THE EXPLOSIVE  
GROWTH OF DATA

WE ARE DROWNING IN DATA,  
BUT STARVING FOR KNOWLEDGE!



# **Customer Relationship Management (CRM)**



## Stock Price Prediction

# Đặt vấn đề

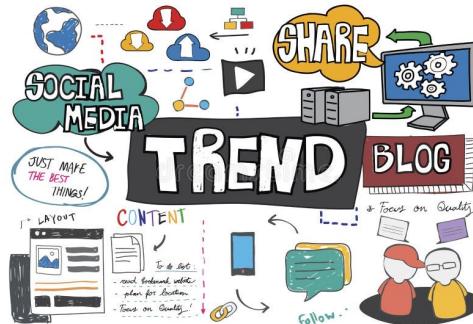


# Applications of **DATA MINING**



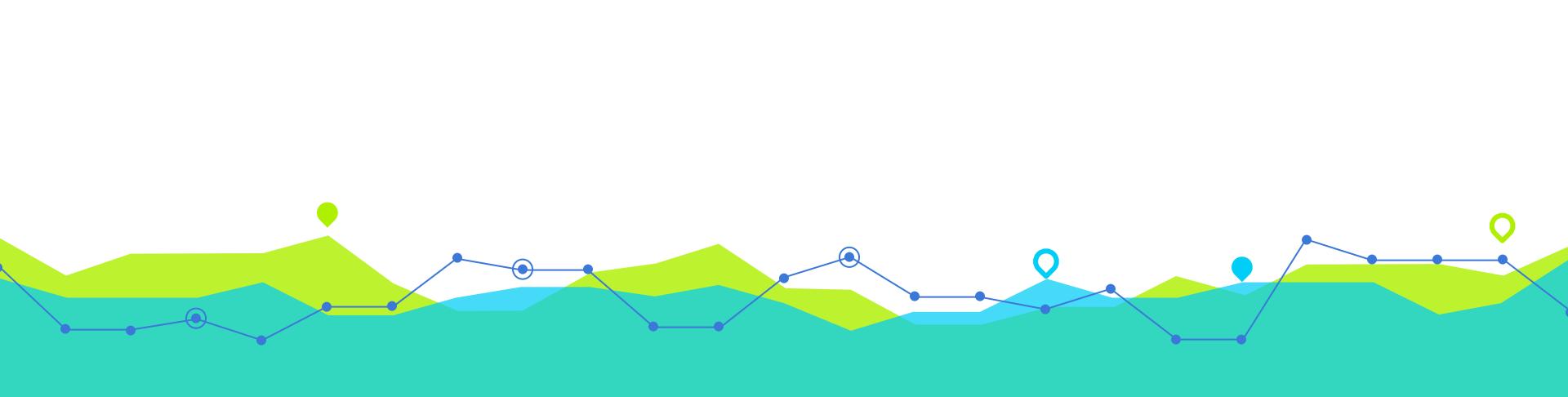
## **HEALTHCARE SECTOR**

## Healthcare



# 1

## Khám phá Tri thức từ CSDL

- 
1. Khái niệm
  2. Quá trình Khám phá Tri thức

# 1. Khái niệm

- Knowledge Discovery in Database (KDD)
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Quá trình không tâm thường để xác định các **mẫu tiềm ẩn**  
hợp lệ, mới lạ, hữu ích và có thể hiểu được bởi người dùng

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37-37.

# 1. Khái niệm

- The process of using the database along with any required selection, preprocessing, subsampling, and transformations of it; applying data-mining methods (algorithms) to enumerate patterns from it; and evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge.

Quá trình sử dụng CSDL cùng với kỹ thuật lựa chọn, tiền xử lý, lấy mẫu và chuyển đổi; áp dụng các phương pháp (thuật toán) KTDL để liệt kê **các mẫu từ CSDL** đó; và **đánh giá** các sản phẩm từ quá trình KTDL này và xác định tập hợp con các mẫu đã được liệt kê thể hiện **tri thức** hữu ích

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37-37.

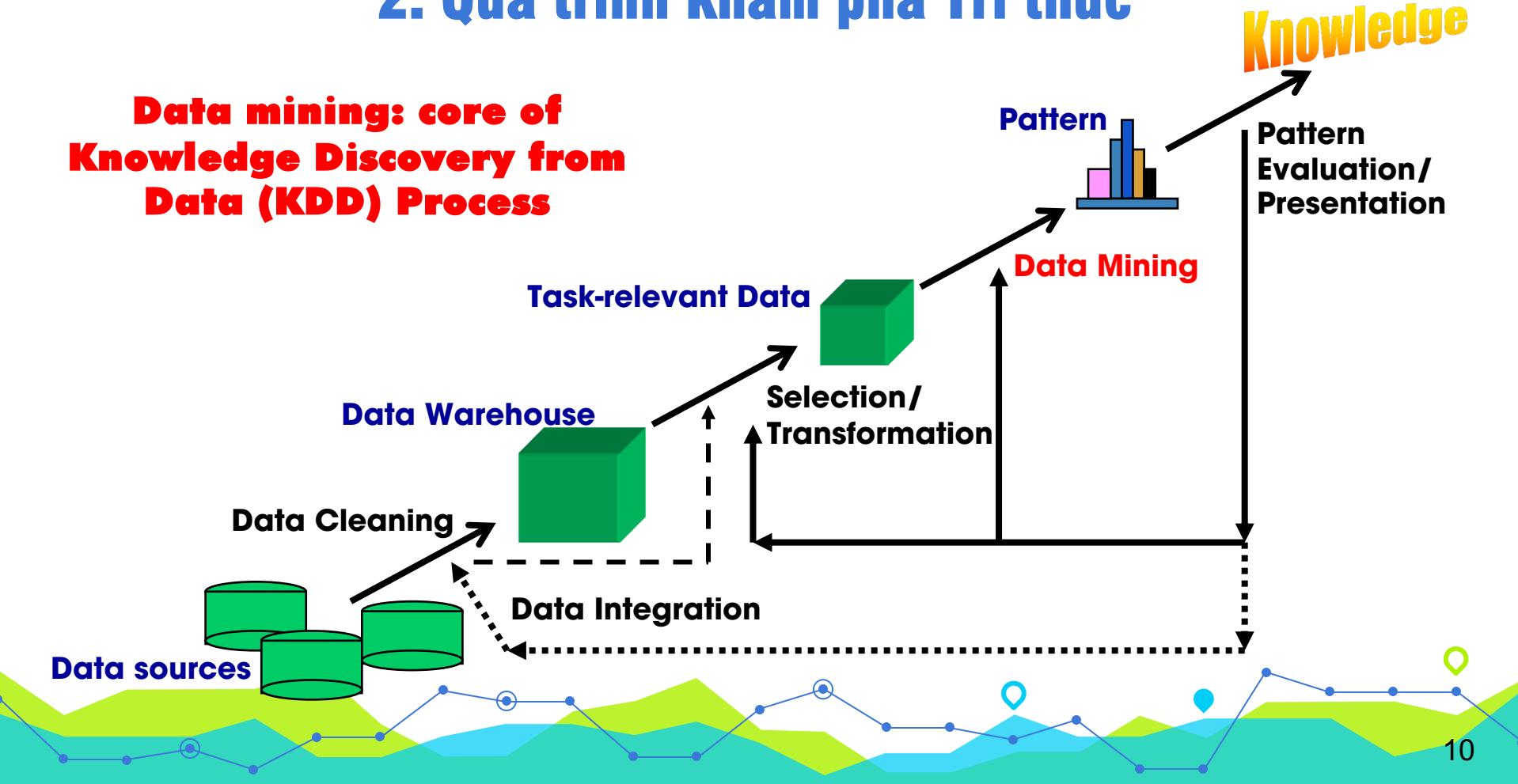
# 1. Khái niệm

## Mẫu tiềm ẩn:

- Mối quan hệ trong dữ liệu
- VD:
  - Người đàn ông mua quần tây thường sẽ mua thêm áo sơ mi
  - Những người có mức tín dụng tốt thì thường ít bị tai nạn
  - Người đàn ông trên 37 tuổi, thu nhập 50K-75\$ thường chi khoảng 25-50% cho việc đặt mua hàng qua catalog

## 2. Quá trình Khám phá Tri thức

**Data mining: core of  
Knowledge Discovery from  
Data (KDD) Process**



# Steps of a KDD Process

- Learning the application domain: relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation: Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining: summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation: visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

## 2. Quá trình Khám phá Tri thức

### 7 bước của Quá trình Khám phá tri thức:

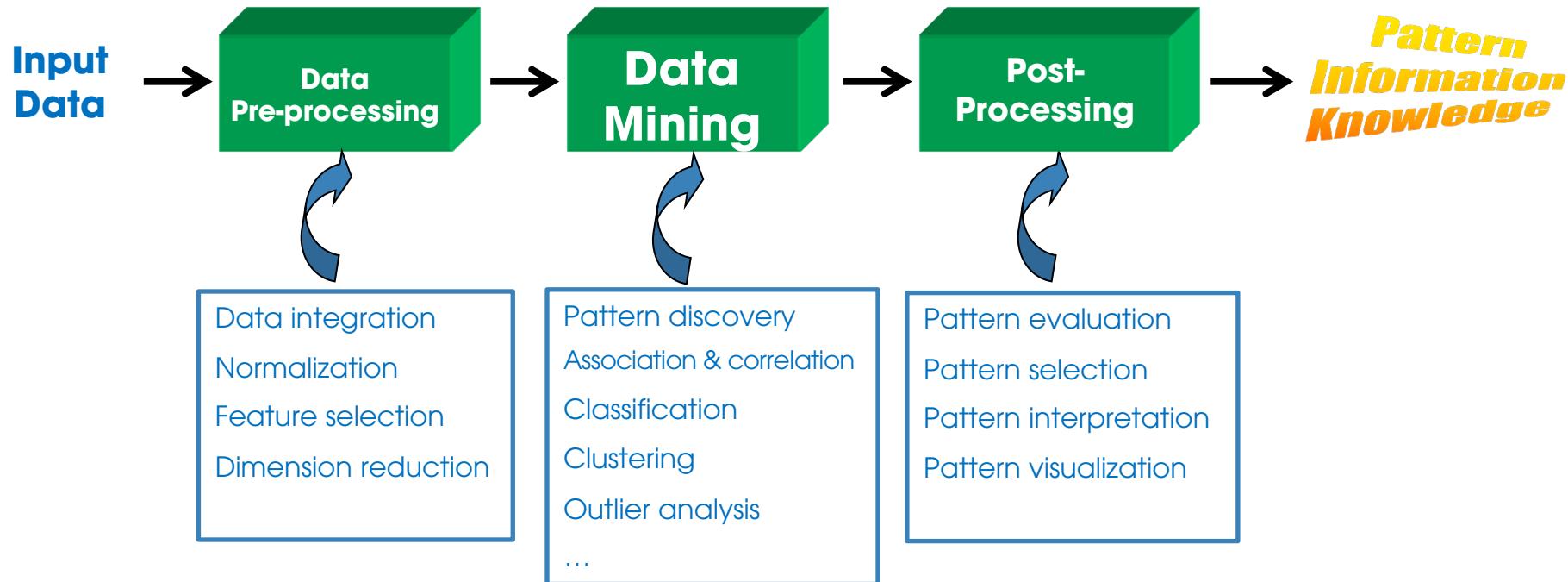
- Data cleaning: Làm sạch dữ liệu nhiễu, thiếu, ...
- Data integration: tích hợp dữ liệu từ nhiều nguồn (dữ liệu chồng lấn, dư thừa, định dạng khác nhau,...)
- Data selection: chọn lọc dữ liệu phù hợp (effectiveness và efficiency), loại bỏ thuộc tính (dư thừa, không ý nghĩa), ...
- Data transformation: chuyển đổi dữ liệu cùng định dạng, miền giá trị
- Data mining: rút trích các mô hình, các mẫu chứa tri thức.
- Pattern Evaluation: đánh giá tính hợp lệ, tính mới, tính hữu dụng, ... của các mẫu
- Pattern Presentation: biểu diễn, trực quan hóa các mẫu tri thức

## 2. Quá trình Khám phá Tri thức

### 5 thành phần chính của Quá trình Khám phá tri thức:

- Data sources: Dữ liệu ban đầu
  - Internal data source hoặc External data source
  - Nhiều định dạng khác nhau, các DBMS khác nhau
- Data Warehouse: Kho dữ liệu tích hợp dữ liệu từ nhiều nguồn khác nhau
- Task-relevant Data: Bộ dữ liệu sẵn sàng cho khai phá – Dữ liệu sau khi tiền xử lý (pre-processing: cleaning, integration, selection, transformation)
- Pattern: Các mẫu (kết quả) khám phá được
- Knowledge: Các mẫu (kết quả) đã được đánh giá, xử lý để biểu diễn dưới dạng tri thức có thể hiểu biết được và diễn giải rõ ràng được bởi người dùng, phục vụ việc ra quyết định

# KDD Process: A Typical View from Machine Learning and Statistics communities



# 2

## Quá trình Khai thác Dữ liệu



1. Khai thác Dữ liệu
2. Các loại dữ liệu
3. Các nhóm tri thức
4. Chức năng, nhiệm vụ chính
5. Một số kỹ thuật chính
6. Đánh giá tri thức đã khai thác
7. Mối liên hệ giữa KTDL và các lĩnh vực chuyên môn khác
8. Quy trình KTDL
9. Kiến trúc hệ thống KTDL

# 1. Khai thác Dữ liệu

## Data mining

- Quá trình trích xuất tri thức (extracting or mining knowledge) từ lượng lớn dữ liệu (1).
- Quá trình không dễ (non-trivial) để trích xuất thông tin ẩn (implicit), chưa được biết trước (previously unknown) và hữu ích (potentially useful) từ dữ liệu (2)
- Khai thác dữ liệu ~ Khám phá tri thức

(1) Frawley, William J., Gregory Piatetsky-Shapiro, and Christopher J. Matheus. "Knowledge discovery in databases: An overview." *AI magazine* 13.3 (1992): 57-57.

(2) Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy. "Advances in Knowledge Discovery and Data Mining". AAAI/MIT Press 1996, ISBN 0-262-56097-6



# 1. Khai thác Dữ liệu

**Is everything “Data mining”?**



vs



# 1. Khai thác Dữ liệu

- Tập hợp các công nghệ, quy trình và phương pháp phân tích được kết hợp với nhau để khám phá những hiểu biết về dữ liệu có thể được sử dụng để hỗ trợ đưa ra quyết định tốt hơn.
- Kết hợp số liệu thống kê, trí tuệ nhân tạo và máy học để tìm ra các mẫu, mối quan hệ và sự bất thường trong các tập dữ liệu lớn.
- Tìm các mối quan hệ và mẫu trong dữ liệu hiện tại, sau đó áp dụng chúng cho dữ liệu mới để dự đoán xu hướng trong tương lai hoặc phát hiện sự bất thường, chẳng hạn như gian lận.

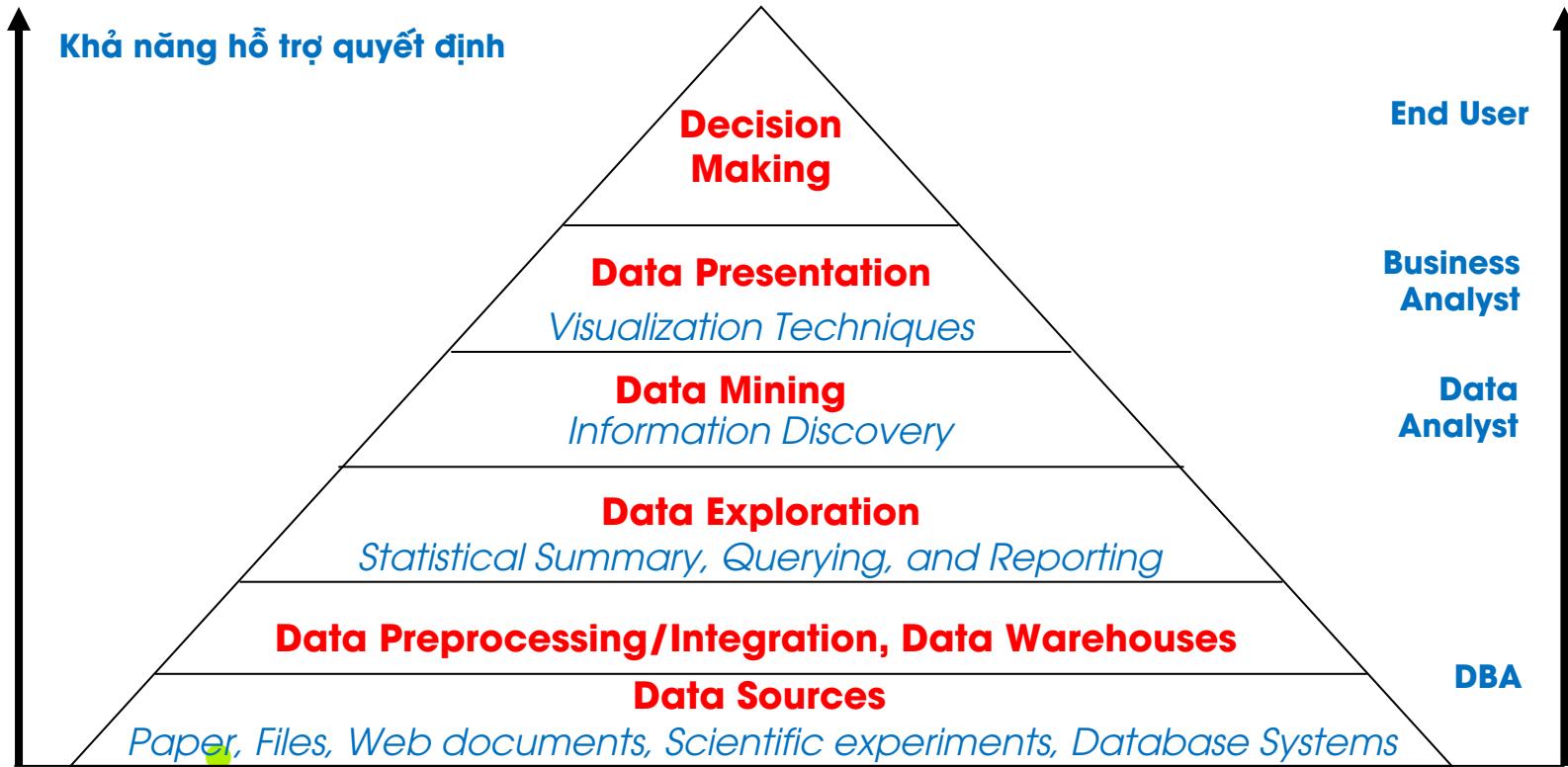
# Ví dụ: Health care & Medical data mining

Thường áp dụng quan điểm trong thống kê và máy học:

- Pre-processing data: feature extraction and dimension reduction
- Classification or/and clustering processes
- Post-processing for presentation



# Data Mining in Business Intelligence



## 2. Các loại dữ liệu trong KTDL

- Relational database, data warehouse, transactional database
- Advanced datasets
- Object-relational databases
- Data streams and sensor data
- Time-series data, temporal data, sequence data (include bio-sequences)
- Structure data, graphs, social networks and multi-linked data
- Multimedia database
- Text databases
- The World-Wide Web



### 3. Các nhóm tri thức từ Quá trình KSDL

- Description of data classes / concepts: các mô hình mô tả các lớp/ khái niệm (Đặc trưng hóa/ phân biên hóa)
- Prediction and Classification: mô hình dự đoán hoặc phân lớp các đối tượng
- Frequent patterns, association patterns: các mẫu phổ biến, khả năng kết hợp các phần tử trong các đối tượng.
- Clustering, outliers/ abnormality analysis: gom nhóm các đối tượng tương đồng, tìm các điểm ngoại biên, các điểm bất thường.
- Analysis of trends from data: mô hình thể hiện các xu hướng, khả năng thay đổi của các đối tượng theo thời gian

# 4. Các chức năng/ nhiệm vụ của KTDL

## Data Mining Functionalities:

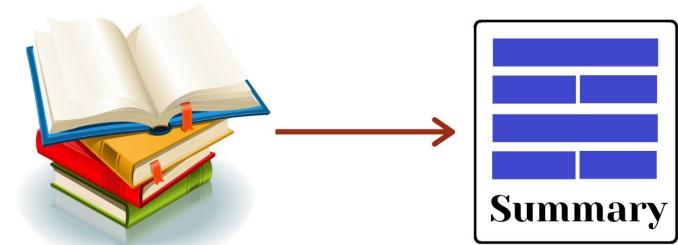
- Generalization – Khái quát dữ liệu
- Association and Correlation Analysis – Phân tích tương quan và luật kết hợp
- Classification and Prediction – Phân lớp và Dự báo
- Cluster Analysis – Gom cụm
- Outlier Analysis – Phân tích mẫu cá biệt, bất thường
- Trend and evolution analysis – Phân tích xu thế phát triển
- Structure and Network Analysis – Phân tích cấu trúc và mạng
- ...

## 4. Các chức năng/ nhiệm vụ của KTDL

### - Generalization – Khái quát dữ liệu

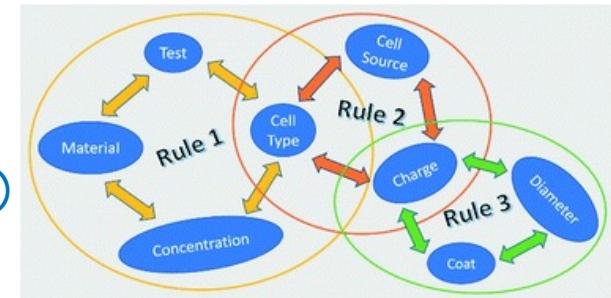
- Data characterization (Đặc trưng hóa dữ liệu): tóm tắt các đặc điểm chung của một lớp dữ liệu đối tượng. Đầu ra có thể được thể hiện dưới nhiều hình thức (data cube, ...).
- Data discrimination (Phân biệt dữ liệu): so sánh các đặc điểm chung giữa các đối tượng dữ liệu của lớp mục tiêu với các đối tượng từ một hoặc một tập hợp các lớp tương phản.

**TEXT SUMMARIZATION** 



## 4. Các chức năng/ nhiệm vụ của KTDL

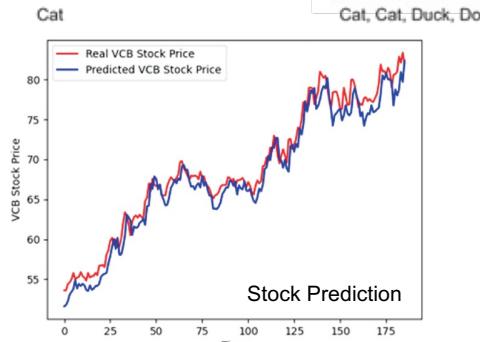
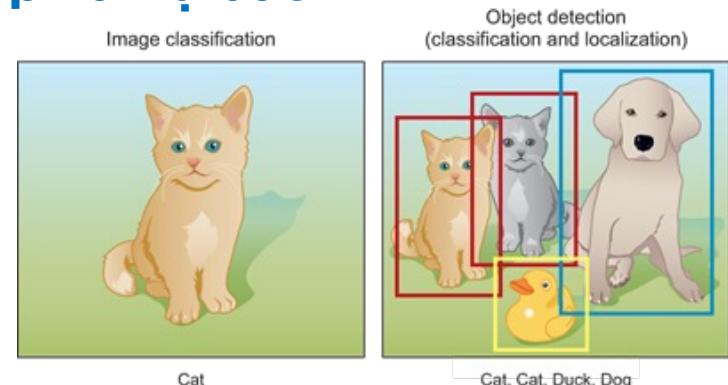
- **Association and Correlation Analysis – Phân tích tương quan và luật kết hợp:** Phân tích tập hợp các items thường xuất hiện cùng nhau trong tập dữ liệu
  - Các mẫu phổ biến hoặc tập phổ biến
    - Những mặt hàng nào thường được mua cùng nhau?
  - Các liên kết, quan hệ tương quan, nhân quả
  - Luật kết hợp điển hình
    - Tā → Bia (0,5%, 75%) (support, confidence)



# 4. Các chức năng/ nhiệm vụ của KTDL

## - Classification and Prediction – Phân lớp và Dự báo

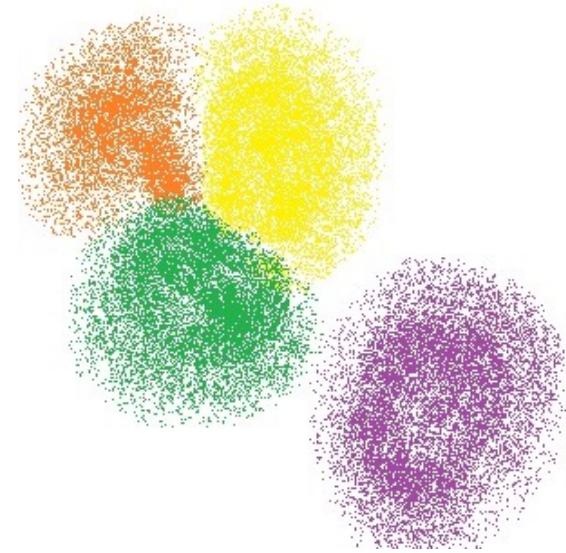
- Xây dựng các mô hình (hàm) mô tả dựa trên phân tích tập dữ liệu huấn luyện.
- Phân biệt các lớp hoặc khái niệm để phân lớp một đối tượng mới.
- Các bài toán dự đoán những giá trị chưa biết hoặc bị thiếu.



# 4. Các chức năng/ nhiệm vụ của KTDL

## - Cluster Analysis – Gom cụm

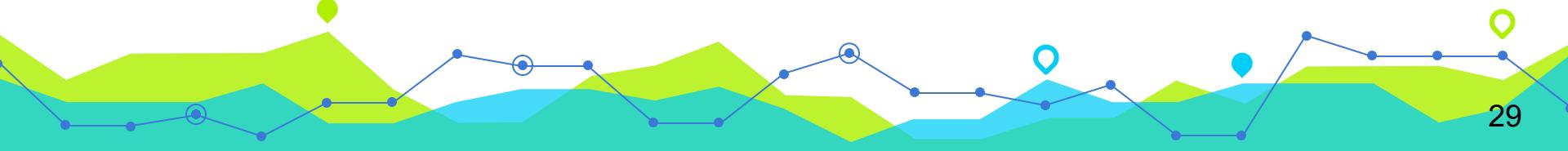
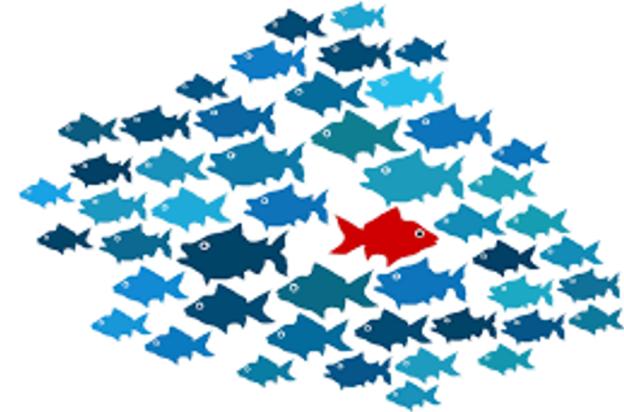
- Unsupervised learning (học không giám sát), nhãn lớp không xác định.
- Phân chia các phần tử dữ liệu thành các cụm (cluster) theo một tiêu chí nào đó.
- Nguyên tắc: Tối đa hóa mức độ tương đồng (similarity) giữa các phần tử trong cụm và tối thiểu hóa mức độ tương đồng giữa phần tử khác cụm



## 4. Các chức năng/ nhiệm vụ của KTDL

### - Outlier Analysis – Phân tích mẫu cá biệt, bất thường

- Mẫu cá biệt, bất thường: đối tượng dữ liệu không tuân theo các quy luật chung của các đối tượng dữ liệu trong cùng một không gian
- Là đối tượng nghiên cứu trong việc phát hiện gian lận, phân tích các sự kiện hiếm gặp, điều tra tội phạm,...
- Phương pháp: clustering, regression,...



## 4. Các chức năng/ nhiệm vụ của KTDL

### - Trend and evolution analysis – Phân tích xu thế phát triển

- Phân tích xu thế, chuỗi thời gian (time-series).

VD: hồi quy, dự đoán giá trị, ...

- Tìm mẫu tuần tự (Sequential pattern mining).

VD: mua máy ảnh, sau đó sẽ mua thẻ nhớ

- Phân tích dự báo (Periodicity analysis)

- Phân tích dựa trên sự tương đồng (Similarity-based analysis)



## 4. Các chức năng/ nhiệm vụ của KTDL

### - Structure and Network Analysis – Phân tích cấu trúc và mạng

- Khai thác đồ thị (Graph mining)

VD: Tìm các đồ thị con phổ biến, cây (XML), ...

- Information network analysis

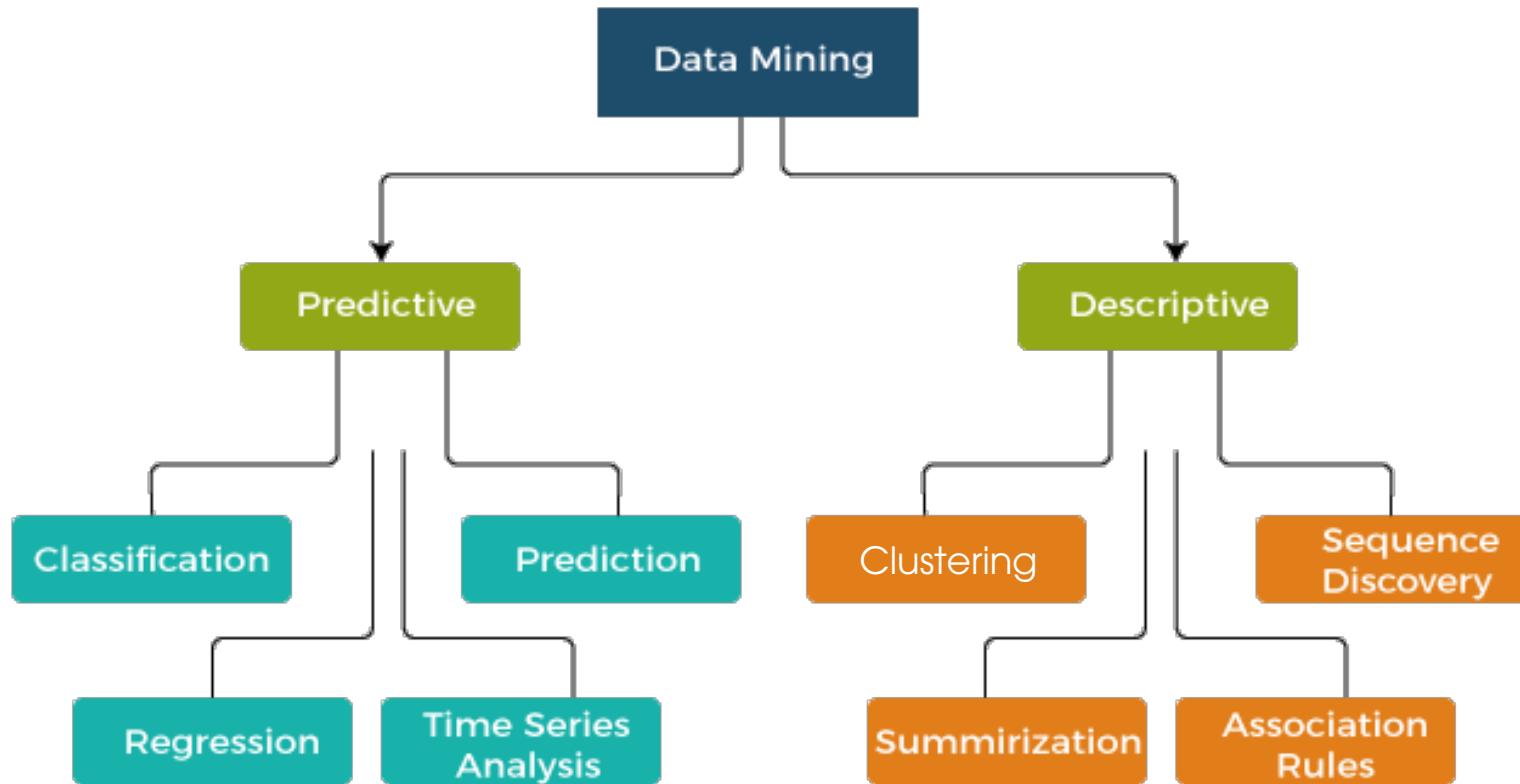
VD: Mạng xã hội: tác nhân (đối tượng, nút) và mối quan hệ (cạnh)

- Web mining

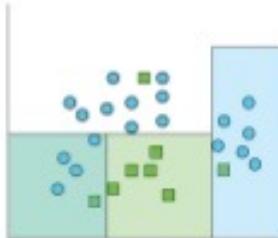
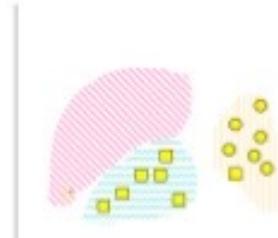
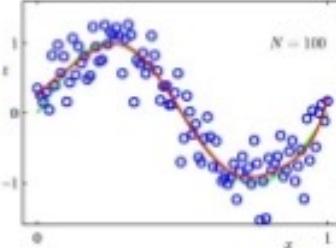
VD: Web community discovery, opinion mining, ...



# 5. Một số kỹ thuật chính trong KTDL

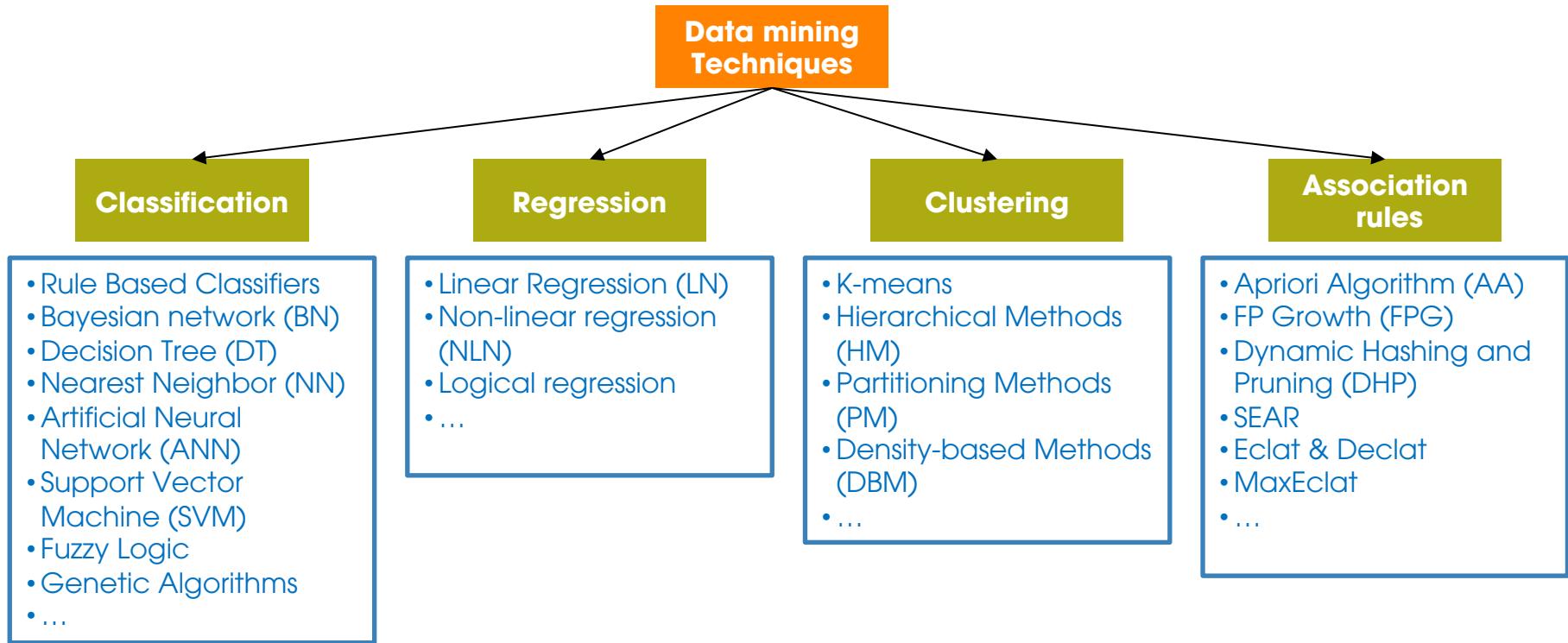


# 5. Một số kỹ thuật chính trong KTDL

Predictive methods	Descriptive methods
<b>Classification</b>  <p>Learns a method for predicting the instance class from pre-labeled (classified) instances</p>	<b>Clustering</b>  <p>Finds "natural" grouping of instances given un-labeled data</p>
<b>Regression</b>  <p>An attempt to predict a continuous attribute</p>	<b>Association Rules</b>  <p>Method for discovering interesting relations between variables in large DBs</p>



# 5. Một số kỹ thuật chính trong KTDL



## 6. Đánh giá Tri thức đã khai thác

- Khai thác dữ liệu có thể tạo ra hàng ngàn mẫu: Không phải tất cả chúng đều tốt (interesting pattern). Một số có thể chỉ phù hợp với không gian kích thước nhất định (thời gian, địa điểm,...)
- Interesting pattern: dễ hiểu đối với con người, hợp lệ trên dữ liệu mới hoặc thử nghiệm với một mức độ chắc chắn.



## 6. Đánh giá Tri thức đã khai thác

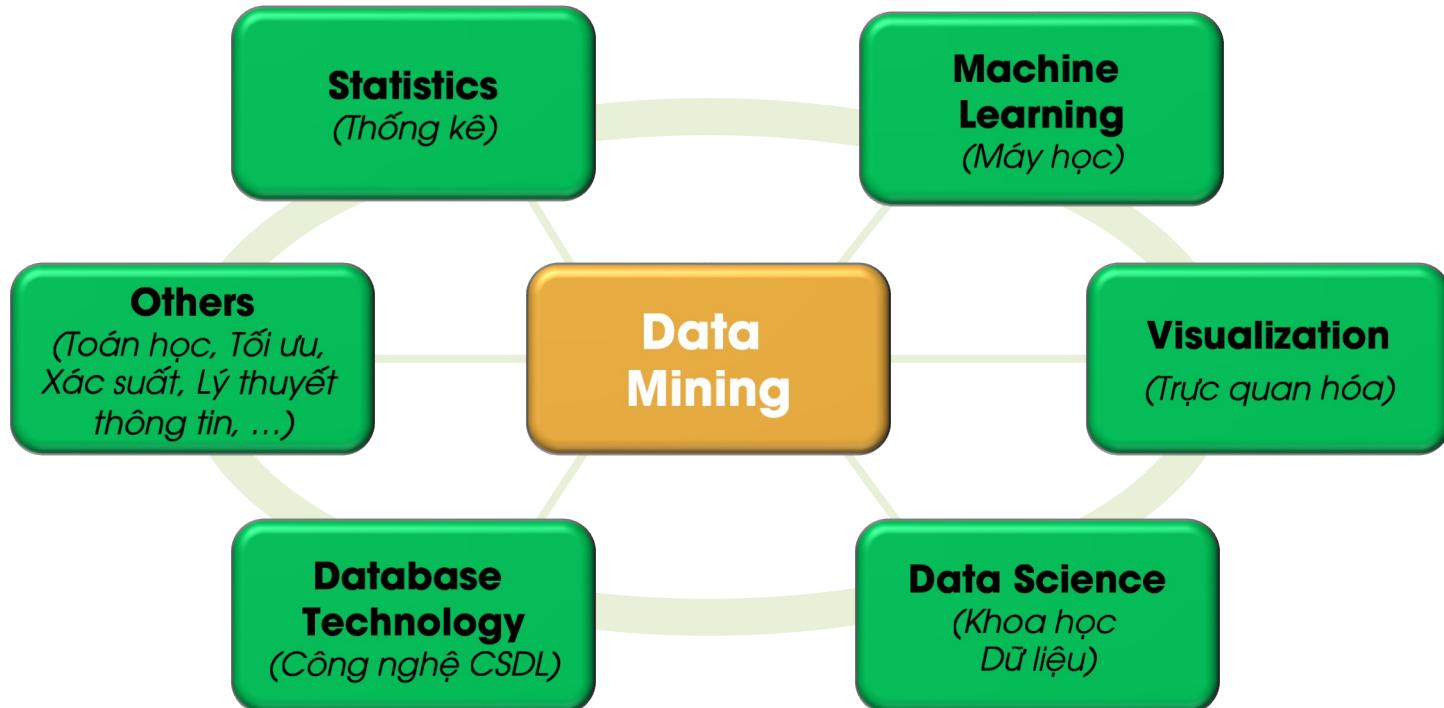
- Đánh giá Tri thức đã khai thác (Evaluation of mined knowledge)
  - Khách quan: dựa trên số liệu thống kê và cấu trúc của các mẫu

VD: độ phổ biến (support), độ tin cậy (confidence), độ chính xác (accuracy), ....

- Chủ quan: dựa trên niềm tin của người dùng vào dữ liệu

VD: tính bất ngờ, tính mới, .....

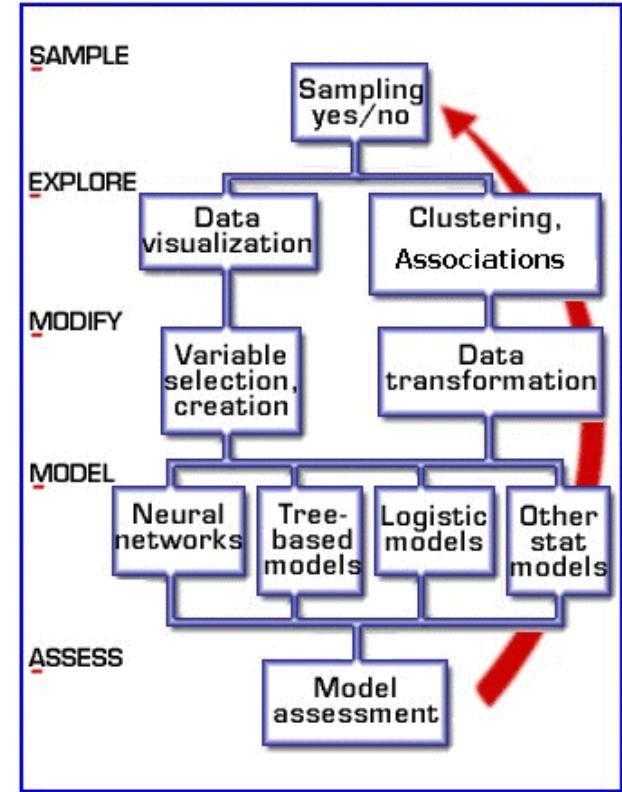
## 7. Mối liên hệ giữa KTDL và các lĩnh vực chuyên môn khác



# 8. Các quy trình Khai thác Dữ liệu

## Quy trình SEMMA của SAS Institute:

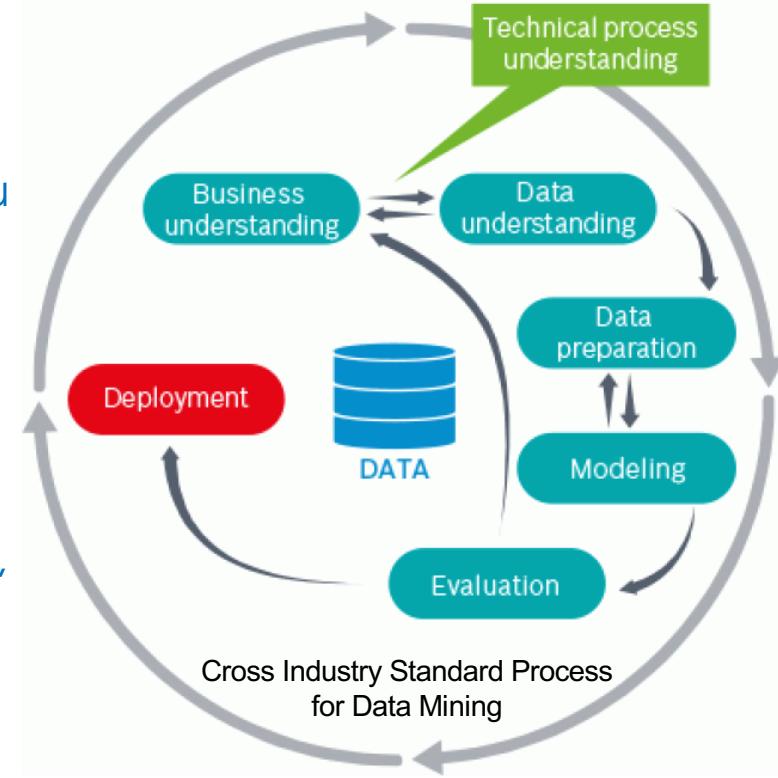
- **Sample:** Lấy mẫu (đủ lớn, đủ nhỏ)
- **Explore:** Khám phá ban đầu, khảo sát dữ liệu
- **Modify:** Sửa đổi dữ liệu
- **Model:** Lựa chọn, mô hình hóa
- **Assess:** Đánh giá độ tin cậy, tính hữu ích. Có thể triển khai áp dụng thực tế (nếu tốt) hoặc lặp lại quá trình (nếu chưa tốt)



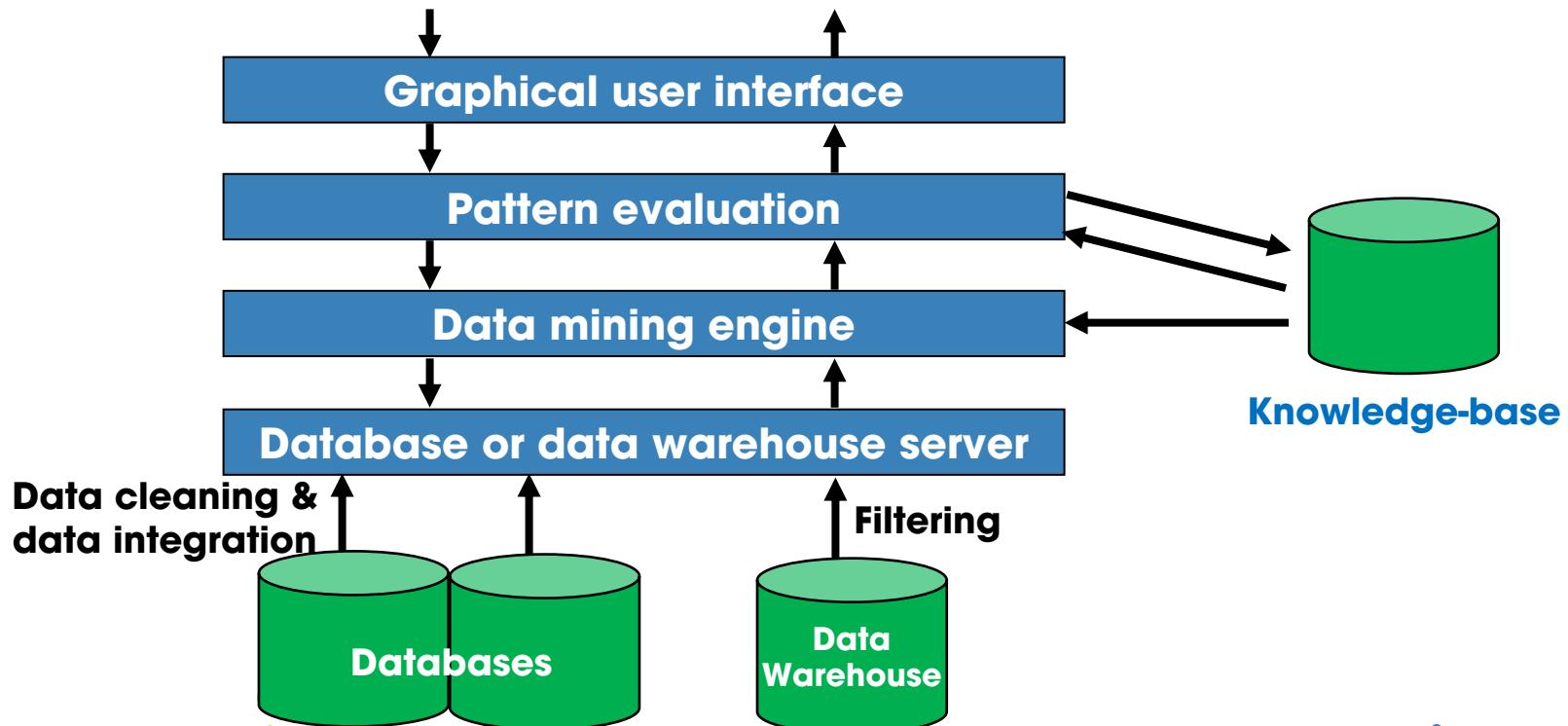
# 8. Các quy trình Khai thác Dữ liệu

## Quy trình CRISP chuẩn công nghiệp:

- Business Understanding: Tìm hiểu nghiệp vụ để xác định những tri thức cần đạt được
- Data Understanding: Khảo sát dữ liệu
- Data Preparation: Data pre-processing
- Modeling
- Evaluation: Đánh giá effectiveness, efficiency, scalability, ...
- Deployment

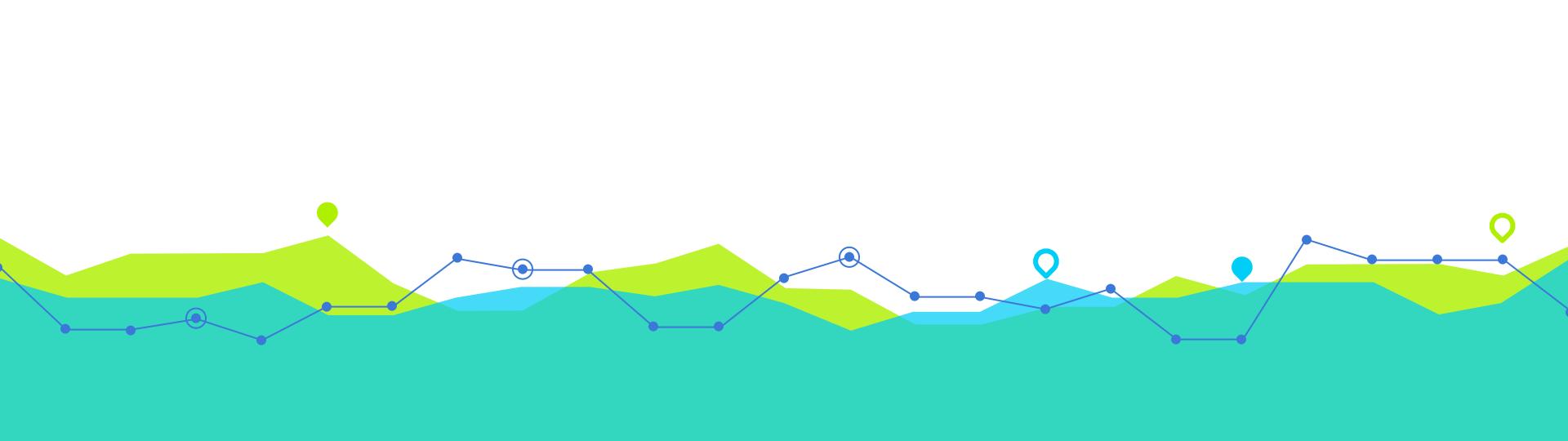


## 9. Architecture: Typical Data Mining System



# 3

## Ứng dụng và thách thức của KTDL

- 
1. Ứng dụng của KTDL
  2. Một số lĩnh vực ứng dụng và ví dụ
  3. Thách thức của KTDL

# 1. Ứng dụng của KTDL

- Phân tích trang web: từ phân lớp trang web, gom cụm đến thuật toán PageRank & HITS
- Phân tích cộng tác & hệ thống khuyến nghị
- Phân tích dữ liệu giỏ hàng để tiếp thị mục tiêu
- Phân tích dữ liệu sinh học và y tế: phân lớp, gom, phân tích trình tự sinh học, phân tích mạng lưới sinh học
- Khai thác dữ liệu và công nghệ phần mềm
- Từ các hệ thống/công cụ khai thác dữ liệu chuyên dụng chính (VD: SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) đến khai thác dữ liệu ẩn (invisible data)

## 2. Một số lĩnh vực ứng dụng KTDL

- **Lĩnh vực ngân hàng**

- Xây dựng mô hình dự báo rủi ro tài chính
- Tìm kiếm tri thức, quy luật của thị trường chứng khoán và đầu tư bất động sản

- **Thương mại điện tử**

- Công cụ tìm hiểu, định hướng, thúc đẩy giao tiếp với khách hàng (chatbot, ChatGPT)
- Phân tích khách hàng duyệt web
- Phân tích hành vi mua sắm online và cho biết thông tin tiếp thị phù hợp với loại khách hàng

## 2. Một số lĩnh vực ứng dụng KTDL

- **Công nghệ sinh học, y học và dược phẩm**
  - Xây dựng công cụ KTDL trực quan cho phép phát hiện sự hiện diện của dược chất
  - Phân tích dữ liệu di truyền
  - Mô hình phát hiện, phân lớp, chẩn đoán bệnh
- **Các lĩnh vực khác**
  - Tuyển dụng ứng viên phù hợp với nhu cầu công ty
  - Phát hiện giả mạo, gian lận thẻ tín dụng
  - Phát hiện tấn công, xâm nhập mạng trái phép

# Classification – Ví dụ 1

## - Công ty Verizon Wireless

- Cung cấp thiết bị, dịch vụ không dây ở Mỹ
- Số lượng khách hàng: 65.7 triệu (2007)
- Thu nhập: 43.9 tỷ \$/ năm



## - Vấn đề:

- Khách hàng rời bỏ cao: 1,300,000/ tháng (2%/ tháng)
- Chi phí cho khách hàng mới: 320\$/ người
- Chi phí thay thế: trăm triệu \$/ năm



# Classification – Ví dụ 1

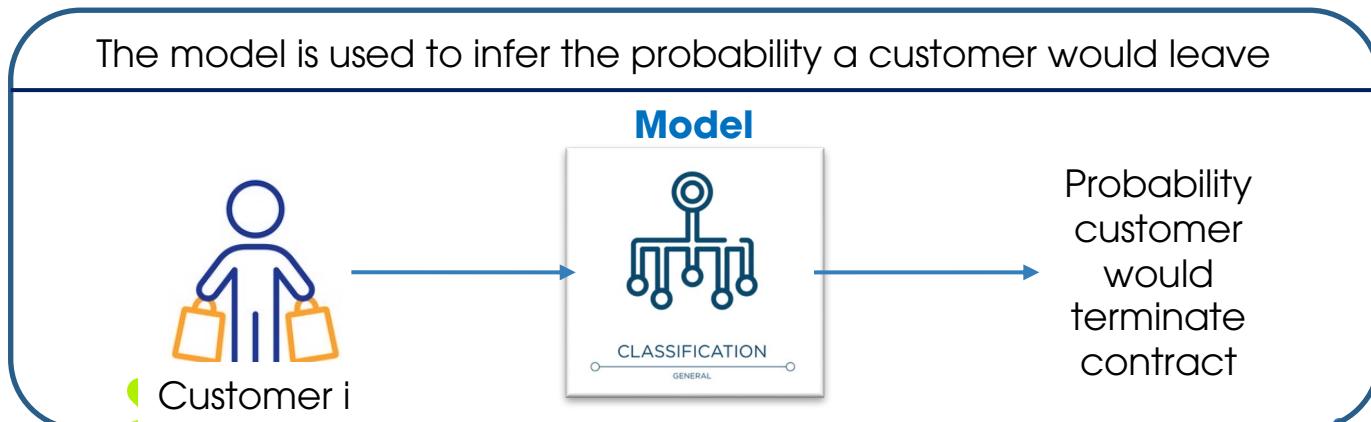
## - Giải pháp thông thường:

- Khuyến mãi, chào mới **tất cả** khách hàng trước khi hết hợp đồng
- Chi phí tốn kém, lãng phí



# Classification – Ví dụ 1

## - Giải pháp của KTDL:



# Classification – Ví dụ 1

## - Giải pháp của KTDL:

- Xây dựng mô hình dự đoán: xác định khách hàng có khả năng rời bỏ
- Khuyến mãi, chào mới cho những khách hàng có nhiều khả năng rời bỏ nhất
- Phát triển kế hoạch mới nhằm đáp ứng nhu cầu khách hàng  
→ Giảm tỷ lệ khách hàng rời bỏ: <1.5% / tháng



# Classification – Ví dụ 2

## Phát hiện gian lận trong giao dịch thẻ tín dụng

- Mục đích: Dự đoán các trường hợp gian lận trong giao dịch
- Hướng giải quyết:
  - Dữ liệu: các giao dịch thẻ tín dụng, thông tin của chủ thẻ.

VD: Số lần dùng thẻ, thời gian sử dụng, mua mặt hàng gì, ...



- Gán nhãn dữ liệu cũ (phân lớp): giao dịch gian lận hay hợp lệ
- Xây dựng mô hình cho lớp các giao dịch
- Sử dụng mô hình để phát hiện gian lận trong giao dịch thẻ tín dụng



# Clustering – Ví dụ 1

## Gom cụm khách hàng



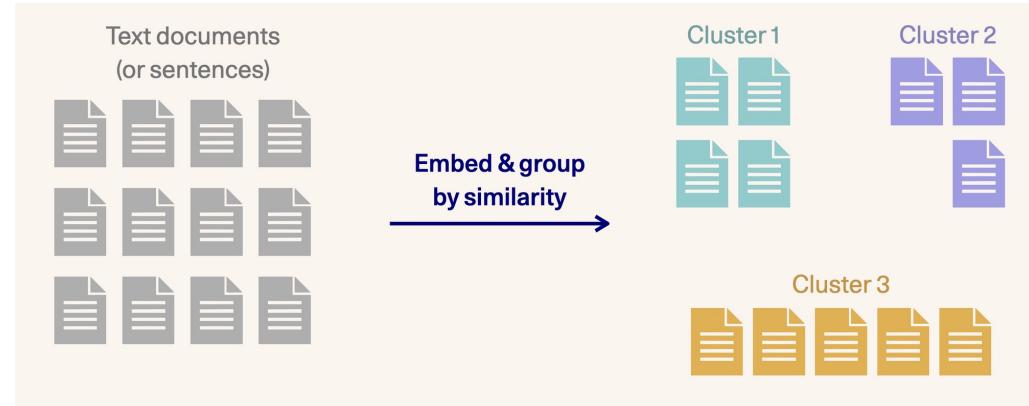
- Mục đích: Chia khách hàng thành các nhóm/ cụm riêng biệt để áp dụng các chiến dịch quảng cáo khác nhau
- Hướng giải quyết:
  - Dữ liệu: Thu thập thông tin cá nhân, cách sống, quan hệ của khách hàng
  - Xác định các cụm/ nhóm khách hàng giống nhau
  - Kiểm tra chất lượng các cụm thông qua việc quan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



# Clustering – Ví dụ 2

## Gom cụm tài liệu

- Mục đích: Tìm nhóm tài liệu giống nhau dựa trên các từ khóa
- Hướng giải quyết:
  - Xác định độ phổ biến của từ trong tài liệu
  - Xây dựng độ đo tương đồng dựa trên độ phổ biến của các từ để gom cụm
- Lợi ích: Trong lĩnh truy vấn thông tin tài liệu, có thể sử dụng các cụm để liên kết tài liệu mới với các tài liệu đã gom cụm.



# Clustering – Ví dụ 3

## Gom cụm cổ phiếu S&P 500

- Mục đích: gom các cụm cổ phiếu và sự kiện tăng/ giảm
- Hướng giải quyết:
  - Dữ liệu: sự biến động của giá cổ phiếu. Cổ phiếu – {UP/DOWN}
  - Độ đo tương đồng: các sự kiện thường giống nhau trong cùng 1 ngày

	<i>Discovered Clusters</i>	<i>Industry Group</i>
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



# Association rules – Ví dụ 1

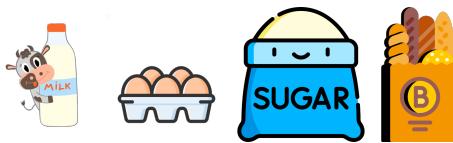
## Quản lý quầy sản phẩm trong siêu thị

- 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



# Association rules – Ví dụ 1

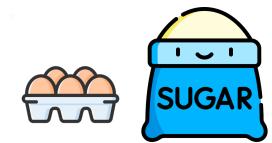
Quản lý quầy sản phẩm trong siêu thị



Customer 1



Customer 2



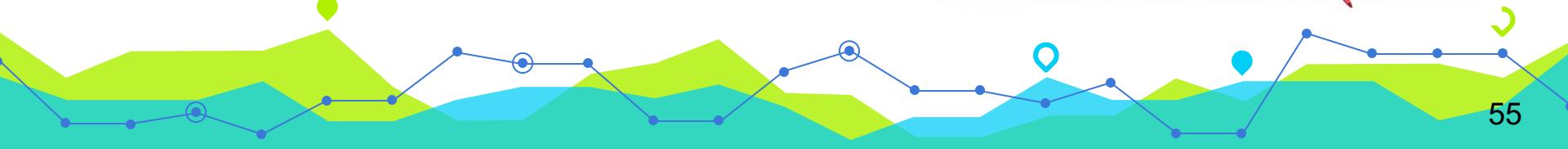
Customer 3

Tìm các Luật kết hợp?

# Association rules – Ví dụ 2

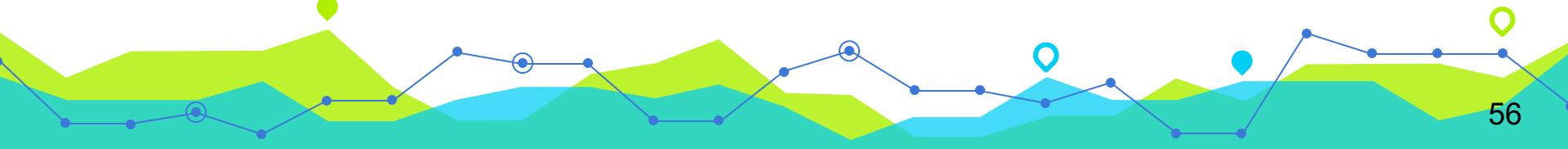
## Quản lý thiết bị sửa chữa

- Mục đích: Công ty bảo trì thiết bị tiêu dùng muốn đoán trước nguyên nhân sửa chữa các thiết bị và trang bị các xe bảo trì, các bộ phận thay thế cần thiết để giảm thiểu số lần đến nhà khác hàng, chi phí.
- Hướng giải quyết:
  - Xử lý dữ liệu trên các thiết bị và bộ phận đã được yêu cầu trong các lần sửa chữa trước để tìm các mẫu đồng xuất hiện.



# Regression – Ví dụ

- Dự đoán giá trị của biến dựa trên giá trị của các biến khác.
- VD:
  - 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



# Outlier analysis – Ví dụ

- Xác định sự lệch hướng rõ rệt so với hành vi thông thường.
- VD:
  - Phát hiện gian lận thẻ tín dụng
  - Phát hiện xâm nhập, tấn công mạng trái phép



### 3. Thách thức của KTDL

#### - Phương pháp khai thác

- Khai thác các tri thức mới từ các loại dữ liệu đa dạng, phức tạp.  
VD: bio, Web, stream, ...
- Hiệu suất (Performance): efficiency, effectiveness, scalability
- Đánh giá mẫu: interesting pattern
- Tích hợp kiến thức nền
- Xử lý nhiễu và dữ liệu không đầy đủ
- Phương pháp khai thác song song (parallel), phân tán (distributed) và  
gia tăng (incremental)
- Hợp nhất kiến thức: Tích hợp tri thức đã khám phá với tri thức hiện có

### 3. Thách thức của KTDL

- **Tương tác người dùng**

- Ngôn ngữ truy vấn khai thác dữ liệu và khai thác ad-hoc
- Biểu diễn và trực quan hóa kết quả khai thác dữ liệu
- Khai thác tương tác tri thức ở nhiều mức độ trừu tượng

- **Ứng dụng KTDL và tác động xã hội**

- Khai thác dữ liệu theo miền cụ thể & khai thác dữ liệu ẩn (invisible data)
- Bảo vệ an ninh dữ liệu, tính toàn vẹn và quyền riêng tư

# Tổng kết chương



## Khám phá tri thức từ CSDL

1. Khái niệm
2. Quá trình Khám phá Tri thức



## Quá trình Khai thác Dữ liệu

1. Khai thác Dữ liệu
2. Các loại dữ liệu
3. Các nhóm tri thức
4. Chức năng, nhiệm vụ chính
5. Một số kỹ thuật chính
6. Đánh giá tri thức đã khai thác
7. Mối liên hệ giữa KTDL và các lĩnh vực chuyên môn khác
8. Quy trình KTDL
9. Kiến trúc hệ thống KTDL

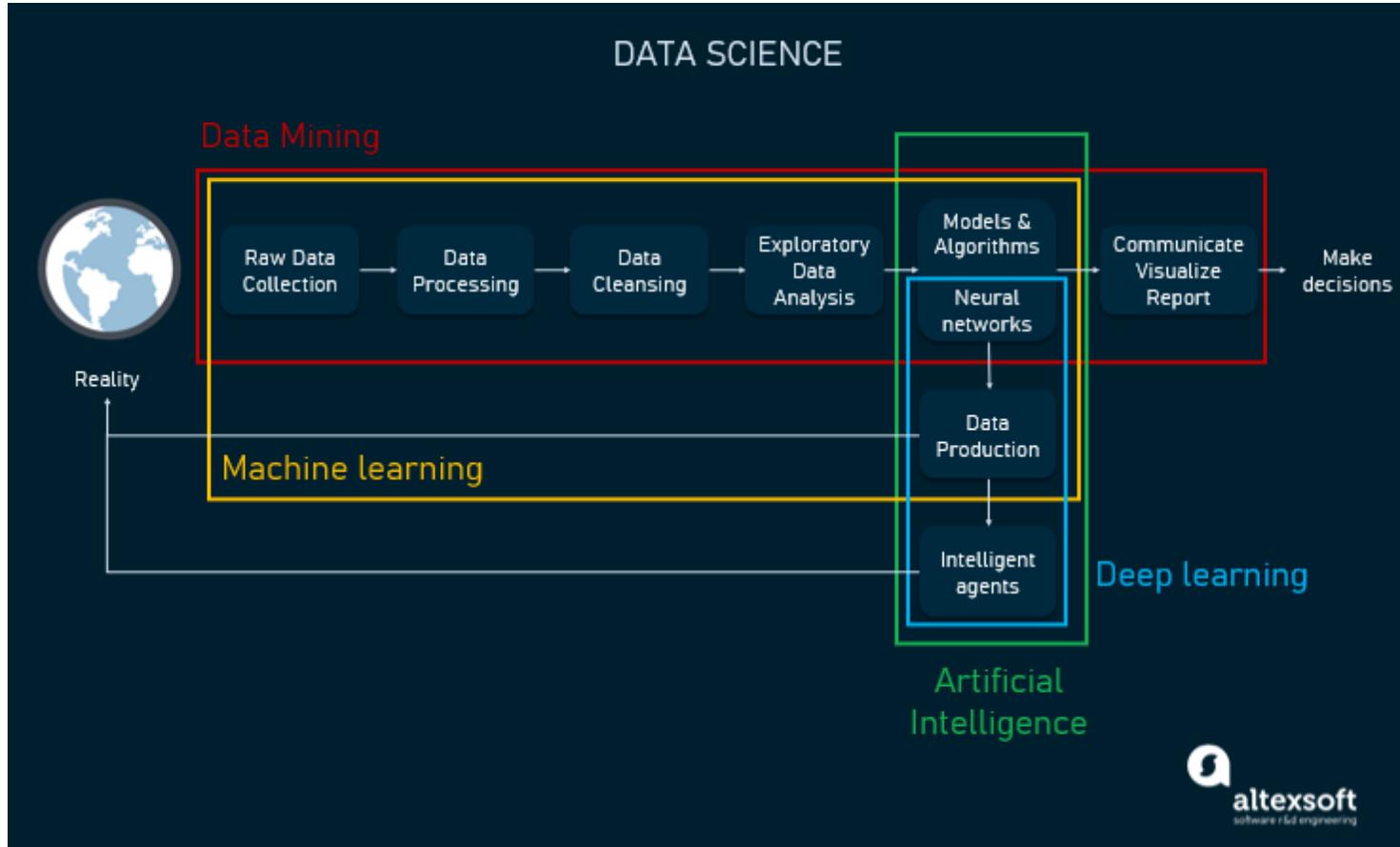


## Ứng dụng và thách thức của KTDL

1. Ứng dụng của KTDL
2. Một số lĩnh vực ứng dụng và ví dụ
3. Thách thức của KTDL



# DATA SCIENCE



# Nhìn đa chiều về KTDL

## - **Dữ liệu:**

- Dữ liệu quan hệ, hướng đối tượng, không đồng nhất, kế thừa
- Kho dữ liệu, dữ liệu giao dịch, stream, dữ liệu không gian, dữ liệu thời gian, chuỗi thời gian, văn bản, web, đa phương tiện, đồ thị, mạng xã hội



# Nhìn đa chiều về KTDL

## - **Tri thức:**

- Đặc trưng hóa dữ liệu (Characterization)
- Phân biệt dữ liệu (discrimination)
- Kết hợp và phân tích tương quan (association and correlation analysis)
- Phân lớp và dự đoán (classification and prediction)
- Gom cụm (clustering)
- Phân tích xu thế (trend and evolution analysis)
- Phân tích mẫu cá biệt, bất thường (outlier analysis)
- ...



# Nhìn đa chiều về KTDL

- **Mô tả (Descriptive) vs Dự đoán (predictive)**
  - Tích hợp nhiều chức năng, khai thác ở nhiều cấp độ
- **Kỹ thuật sử dụng**
  - Chuyên sâu về dữ liệu, kho dữ liệu (OLAP), máy học, thống kê, nhận dạng mẫu, trực quan hóa, hiệu suất cao, ....
- **Ứng dụng**
  - Bán lẻ, viễn thông, ngân hàng, phân tích gian lận, thị trường chứng khoán, khai thác dữ liệu sinh học, khai thác văn bản, Web, ...



# THANKS!

Any questions?

