KHAI THÁC DỮ LIỆU & KHAI PHÁ TRI THỨC

DATA MINING & KNOWLEDGE DISCOVERY

Bài 1. **Giới thiệu tổng quan**

Mã MH: 505043

TS. HOÀNG Anh

Nội dung

- **Tại sao** cần Khai thác dữ liệu/ Why Data Mining?
- **Khái niệm** Khai thác dữ liệu/ What is Data Mining?
- Các **loại dữ liệu** được khai thác/ Data types can be mined?
- Các **bài toán** khai thác dữ liệu/ DM functionalities
- Mẫu có ý nghĩa/ Interesting patterns?
- Phân loại hệ thống khai thác dữ liệu/ DM schemes
- **Thách thức** khi khai thác dữ liệu/ DM challenges

Dữ liệu lớn ở khắp mọi nơi!

- **Dữ liệu lớn** trong các CSDL thương mại và khoa học
- Thành quả từ sự phát triển khoa học công nghệ, tạo ra và thu thập dữ liệu tự động

- **Khả năng!**
  - Thu thập bất kỳ dữ liệu nào, vào bất kỳ khi nào, và ở bất kỳ nơi đâu bạn có thể.

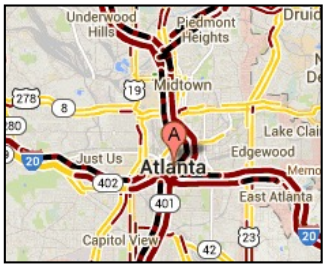
- **Kỳ vọng!**
  - Khai thác giá trị dữ liệu đã thu thập được
  - Hổ trợ ra quyết định

Cyber Security

An ninh mạng

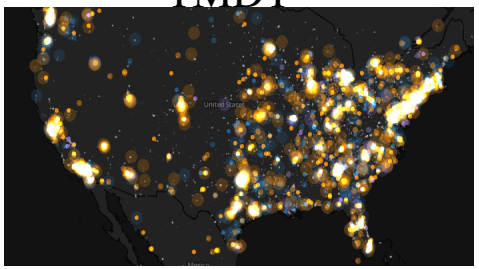
E-Commerce

TMĐT



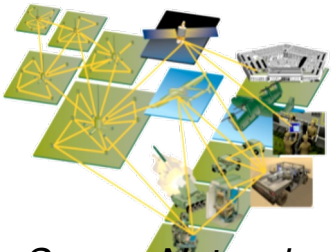
Traffic Patterns

Mạng lưới giao thông



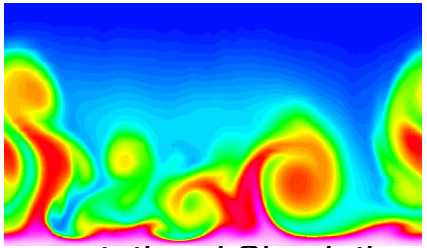
Social Networking: Twitter

Mạng xã hội



Sensor Networks

Mạng cảm biến



Computational Simulations

Mô phỏng tính toán

Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

1. Tại sao cần khai thác dữ liệu? Why

- Sự phát triển bùng nổ của dữ liệu: từ terabytes đến petabytes
  - Dữ liệu thu thập và dữ liệu khả dụng
    - Các công cụ thu thập dữ liệu tự động; Các hệ CSDL, Web
    - Xã hội số hóa
  - Các nguồn dữ liệu dồi dào
    - Kinh doanh: Web, TMĐT, các giao dịch, TT chứng khoán, …
    - Khoa học: Viễn thám, tin sinh học, khoa học mô phỏng, …
    - Xã hội và con người: tin tức, máy ảnh số,
- Chúng ta đang chìm đắm trong dữ liệu, nhưng đói khát kiến thức!

(We are drowning in **data**, but starving for **knowledge**!)

- **"Necessity is the mother of invention"**—Khai thác dữ liệu/**Data mining**—Phân tích tự động các tập dữ liệu lớn.

Từ góc nhìn thương mại

- Dữ liệu lớn được thu thập và lưu trữ
  - Dữ liệu Web
    - Google has Peta Bytes of web data
    - Facebook has billions of active users
  - Dữ liệu mua bán tại các cửa hàng/ Website
    - Amazon handles millions of visits/day
  - Các giao dịch ngân hàng

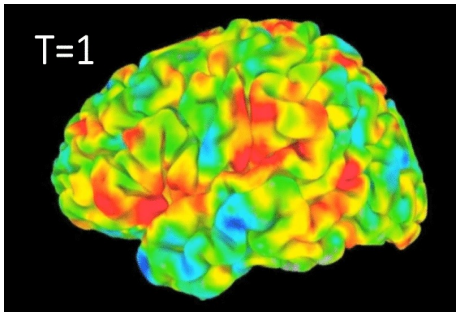
- **Máy tính/Computers** rẻ hơn, khả năng tính toán cao hơn

- Áp lực **cạnh tranh**
  - Cung cấp SP, dịch vụ tốt hơn/ nhanh hơn/ rẻ hơn,…Provide better, customized services for an edge (e.g in Customer Relationship Management)

Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

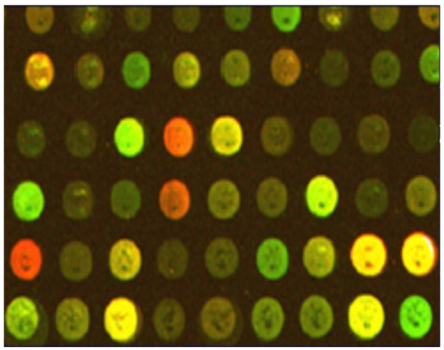
Từ góc nhìn khoa học

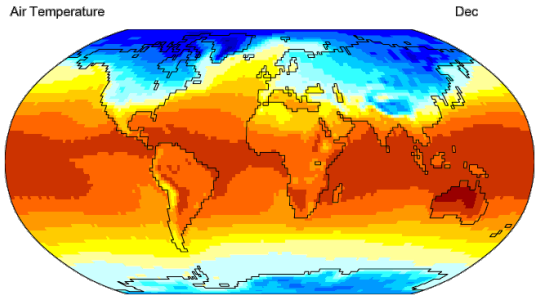
- Dữ liệu được thu thập và lưu trữ với tốc độ cao
  - Remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - Telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - Scientific simulations
    - Terabytes of data generated in a few hours
- **Khai thác dữ liệu** trợ giúp các nhà khoa học
  - In automated analysis of massive datasets
  - In hypothesis formation



fMRI Data from Brain


Sky Survey Data


Gene Expression Data



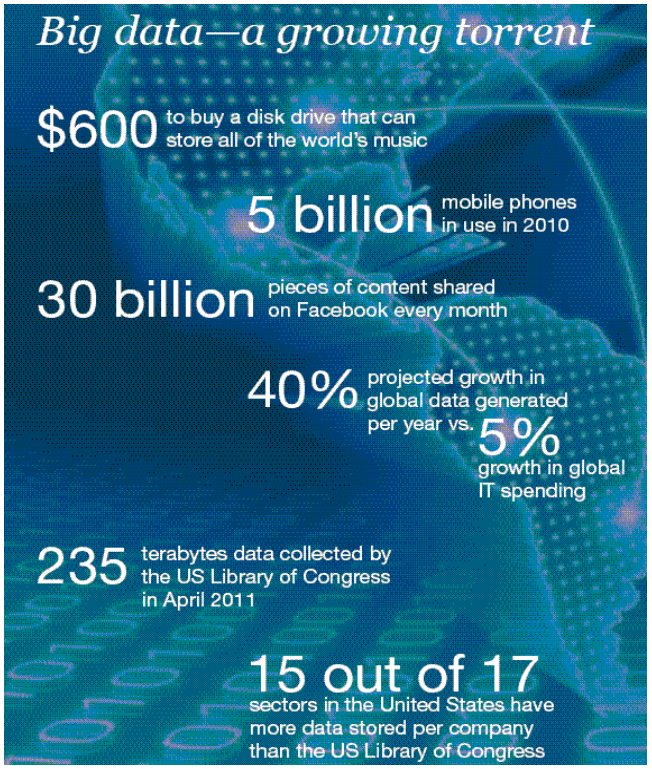
Surface Temperature of Earth

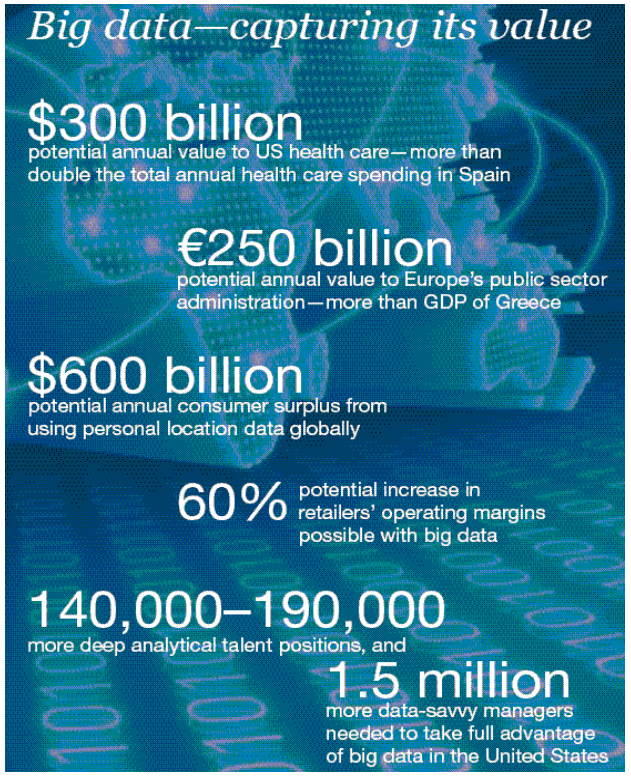
Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Tạo ra nhiều cơ hội lớn để phát triển sản phẩm

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

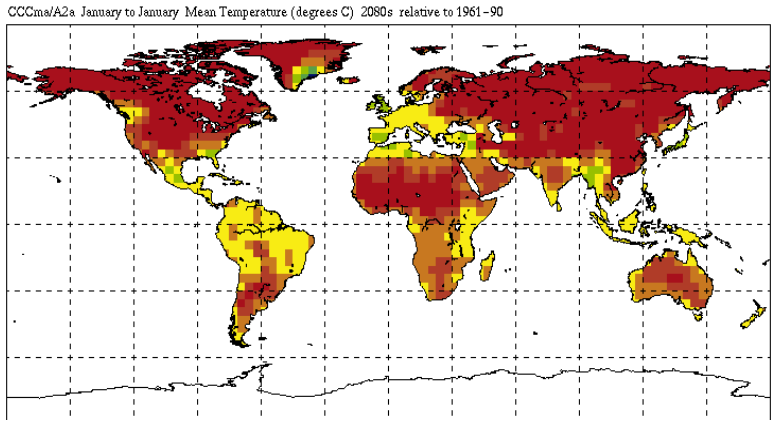




Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Giải quyết các vấn đề lớn của xã hội

Improving health care and reducing costs



Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Lịch sử phát triển của CSDL

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s:
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

Các ứng dụng tiềm năng/ Applications

- **Phân tích dữ liệu/** Data analysis**, và Hệ hỗ trợ ra quyết định/** decision support, decision making
  - Phân tích, quản lý thị trường (Market analysis and management)
    - Target marketing, customer relationship management (CRM), market basket analysis, market segmentation
  - Phân tích, quản lý rủi ro (Risk analysis and management)
    - Forecasting, customer retention, quality control, competitive analysis
  - Phát hiện gian lận (Fraud detection and detection of unusual patterns (outliers))

Các ứng dụng tiềm năng khác

- Một số ứng dụng khác
  - Khai thác văn bản/ Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Phân tích dữ liệu sinh học/ Bioinformatics and bio-data analysis
  - …

Phân tích, quản lý thị trường

- Dữ liệu đến từ đâu?/ Where does the data come from?
  - Credit card transactions, discount coupons, customer complaint calls

- Marketing mục tiêu
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time

Phân tích, quản lý thị trường

- **Phân tích Cross-market**
  - Associations/co-relations between product sales, & prediction based on such association
- **Hồ sơ khách hàng/ Customer profiling**
  - What types of customers buy what products
- **Phân tích nhu cầu khách hàng/ Customer requirement analysis**
  - Identifying the best products for different customers
  - Predict what factors will attract new customers

Phát hiện gian lận

- Phương pháp: Clustering & model construction for frauds, outlier analysis

- Ứng dụng: Health care, retail, credit card service, telecom.
  - Bảo hiểm y tế/ Medical insurance
    - Professional patients, and ring of doctors
    - Unnecessary or correlated screening tests
  - Viễn thông/ Telecommunications:
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Bán lẽ/ Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees

Một số ứng dụng khác

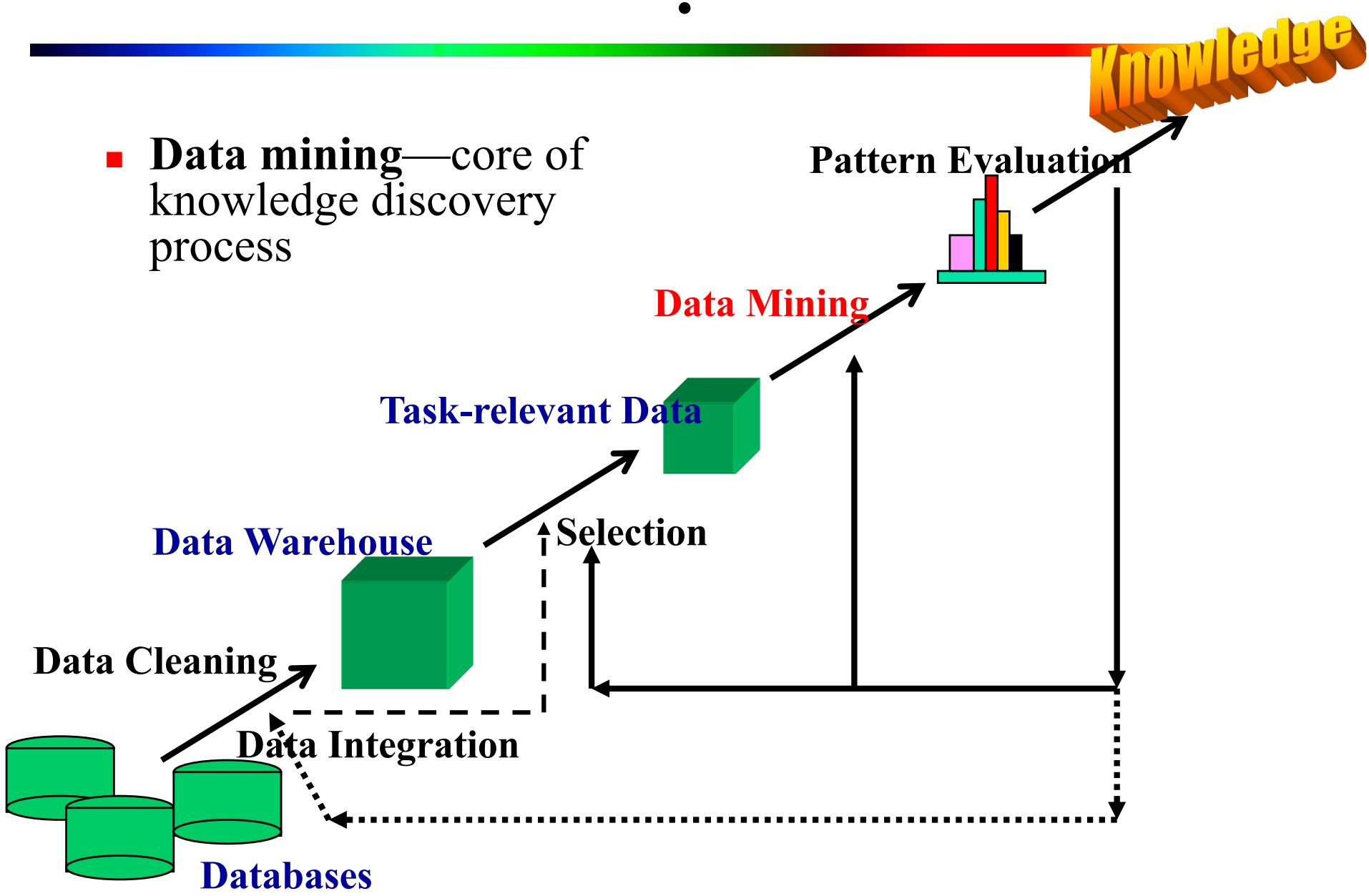
- Hổ trợ lướt Web/ Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.
  - …

2. Khai thác dữ liệu là gì? What

- Khai thác dữ liệu (**knowledge discovery** from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) **patterns or knowledge** from **huge amount of data**
- Tên gọi khác
  - Khai phá tri thức từ các CSDL/ Knowledge discovery in databases (**KDD**)
- Watch out: Is everything "data mining"?
  - Query processing
  - Expert systems/ Machine learning
  - Statistical programs

Khai thác dữ liệu: KDD Process

- **Data mining**—core of knowledge discovery process



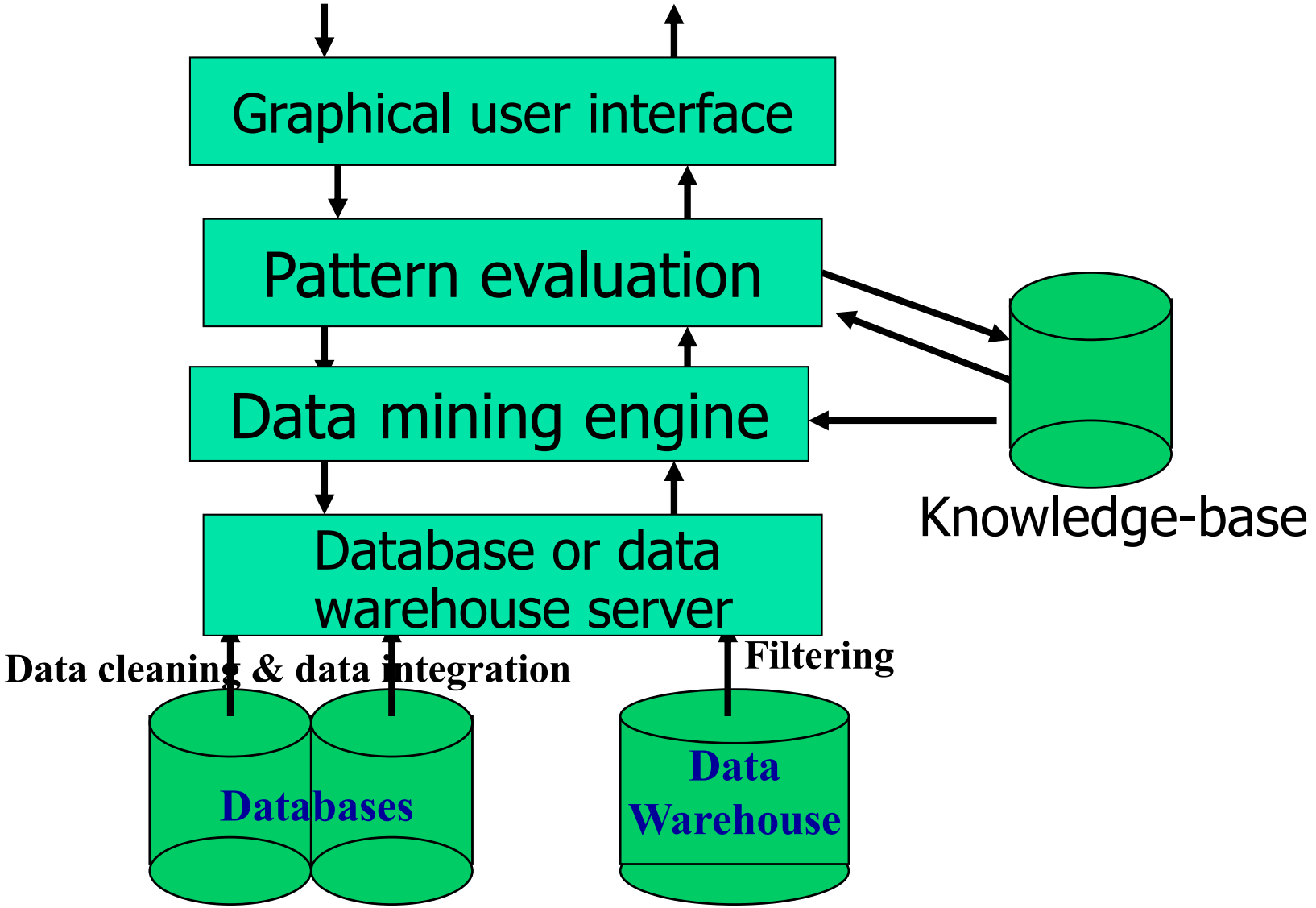
Qui trình KDD

- Lĩnh vực ứng dung/ Learning the **application domain**
  - Relevant prior knowledge and goals of application
- Thu thập dữ liệu/ Creating **a target data set**: data selection
- Tiền xử lý dữ liệu/ Data **cleaning and preprocessing**:

(*may take 60% - 80% of effort!*)

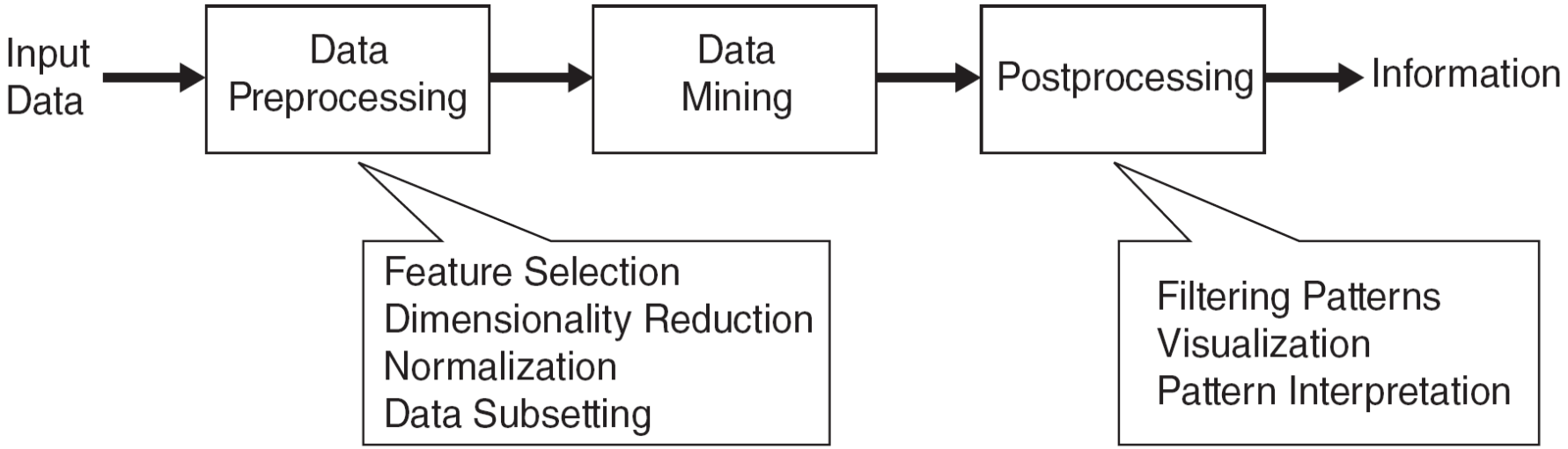
- Giảm chiều dữ liệu và chuyển đổi/ Data **reduction and transformation**
  - Find useful features, dimensionality/variable reduction.
- Lựa chọn chức năng/ Choosing **functions** of data mining
  - Summarization, classification, regression, association, clustering.
- Lựa chọn thuật toán/ Choosing the **mining algorithm(s)**
- Khai thác/ Data mining: search for patterns of interest
- **Đánh giá mẫu/ Pattern evaluation and knowledge presentation**
  - Visualization, transformation, removing redundant patterns, etc.
- Sử dung tri thức được khám phá/ Use of discovered **knowledge**

Kiến trúc hệ thống khai thác dữ liệu



2. Khai thác dữ liệu là gì? What

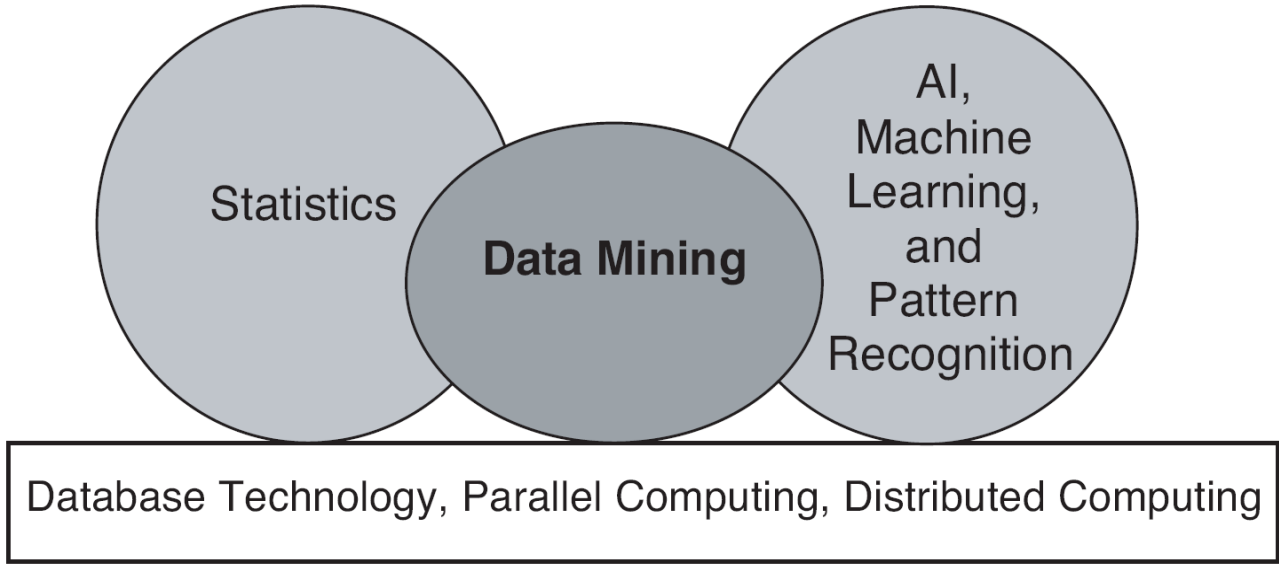
- Một số khái niệm:
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns



Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Nguồn gốc của khai thác dữ liệu

- Các ý tưởng đến từ **học máy, trí tuệ nhân tạo, nhận dạng mẫu, thống kê, và các hệ CSDL** (Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems)
- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed



- A key component of the **emerging** field of data science and data-driven discovery

Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

3. Những loại dữ liệu nào có thể khai thác?

- **CSDL/ Database** data (RDBMs)
- Kho dữ liệu/ Data **warehouse**
- Dữ liệu giao dịch/ **Transactional** data
- Một số khác/ Other types of data:
  - Sequence data, data streams (cont.), spatial data (maps), engineering design data, hypertext, multimedia, web data, etc.

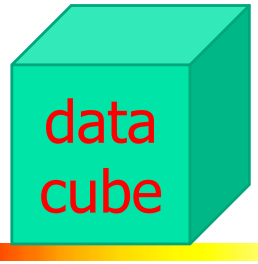
- Advanced database and information repository
  - Spatial and temporal data
  - Time-series data
  - Stream data
  - Multimedia database
  - Text databases & WWW

CSDL (RDBMs): Relational -> tables

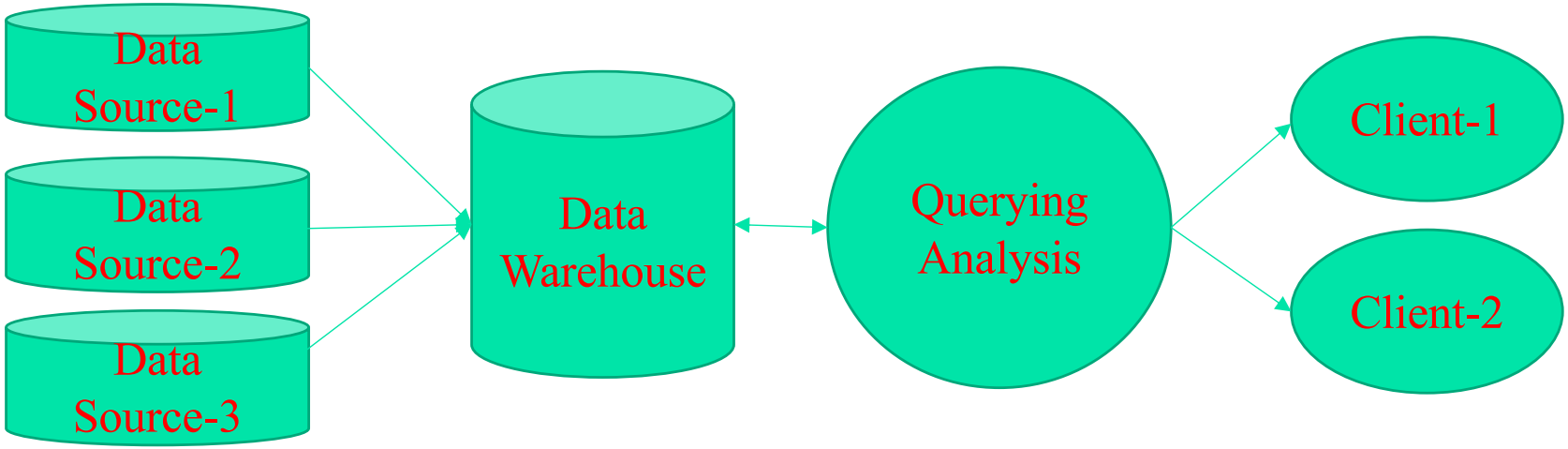
- **RDBMs: Bảng**
  - Set of tables – has rows (tuples) and columns (attributes)
  - While mining databases, we can search for trends or data pattern

- **Ví dụ**:
  - Analysing customer data to predict the credit risks of new customers (based on previous data)
  - Analysing sales data - (any deviations)

Kho dữ liệu/ Data warehouse



- Dữ liệu thu thập được từ nhiều nguồn bằng các truy vấn và ra quyết định trên dữ liệu.
- In data warehouse, data is stored in multidimensional structure (datacube) where each dimension is each attribute



Dữ liệu giao dịch/ Transactional data

- Mỗi bản ghi được gọi là một giao dịch/ transaction
  - sales,
  - flight booking,
  - user clicks on web page
- Mỗi giao dịch sẽ có ID, và danh sách các mục tạo nên giao dịch đó/ Transaction has transaction ID, list of other items making transaction
- Có thể khai thác các mẫu thường xuyên/ frequent patterns
- Các loại dữ liệu khác:
  - Sequence data, data streams (cont.), spatial data (maps), engineering design data, hypertext, multimedia, web data, etc.

4. Bài toán khai thác dữ liệu/ Tasks

- Dữ liệu luôn được kết hợp với các mô tả về lớp/ khái niệm**:**
  - Đặc tính của dữ liệu/ Data **characterisation**:
    - Refers to the summary of the class/ concept
    - Output -> General overview
  - Phân loại dữ liệu/ Data **discrimination**:
    - Compares the common features of the classes
    - Output -> barcharts, curves, etc.
- Khai thác mẫu thường xuyên, sự kết hợp, và tương quan/ **frequent patterns, Association**, and **Correlations**
  - Mẫu thường xuyên/ Frequent patterns:
    - Things which are found most commonly in data
      - Frequent itemsets (data items/ data objects)
      - Frequent subsequence
      - Frequent substructure
  - Phân tích luật kết hợp/ Association analysis: (relationship)
    - It is a way identifying the relation between various items
    - Example: used to determine sales of items that are frequently purchased together

4. Bài toán khai thác dữ liệu

- Phân tích tương quan/ C**orrelation** analysis:
  - Mathematical technique
  - Shows how strongly pair of attributes are related together
  - Example: tall peope tend to have more weight

- **Phân lớp và Hồi qui/ Classification** and **Regression** for predictive analysis
  - **Phân lớp/ Classsification:**
    - Process of finding a model that distinguishes data items
      - Decision tree is used for classification
  - **Hồi qui/ Regression:**
    - Statistical methodology that is used for numeric prediction (done based on previous data) of missing data

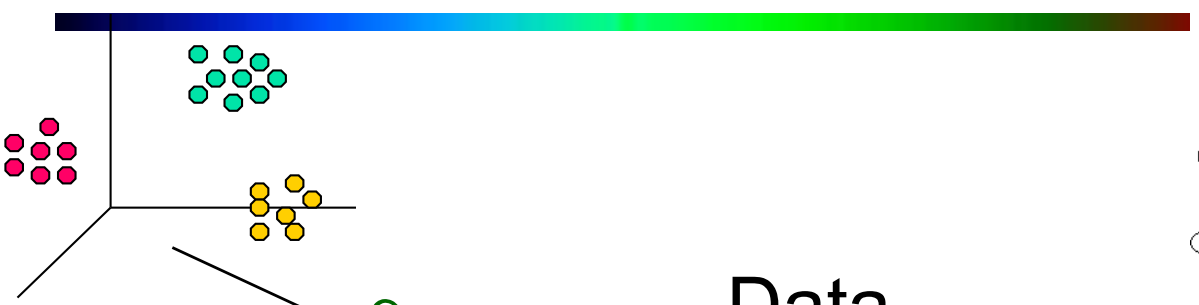
4. Bài toán khai thác dữ liệu

- **Gom cụm/ Cluster** analysis (Group)
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity

- **Phân tích Outlier**
  - Outlier: a data object that does not comply with the general behavior of the data
  - Useful in fraud detection, rare events analysis

- **Phân tích xu hướng/ Trend** and **evolution** analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis

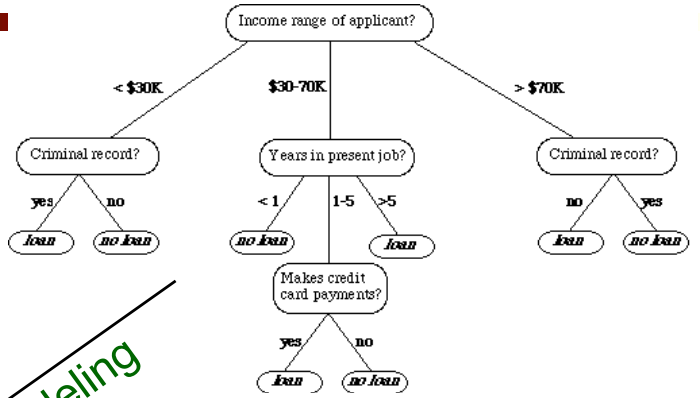
Data Mining Tasks …



Clustering

Data

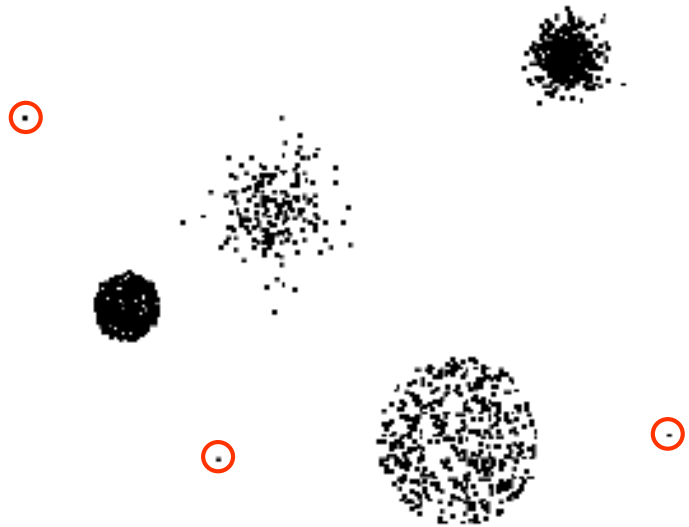
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No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



Predictive Modeling

Association Rules

Anomaly Detection



Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

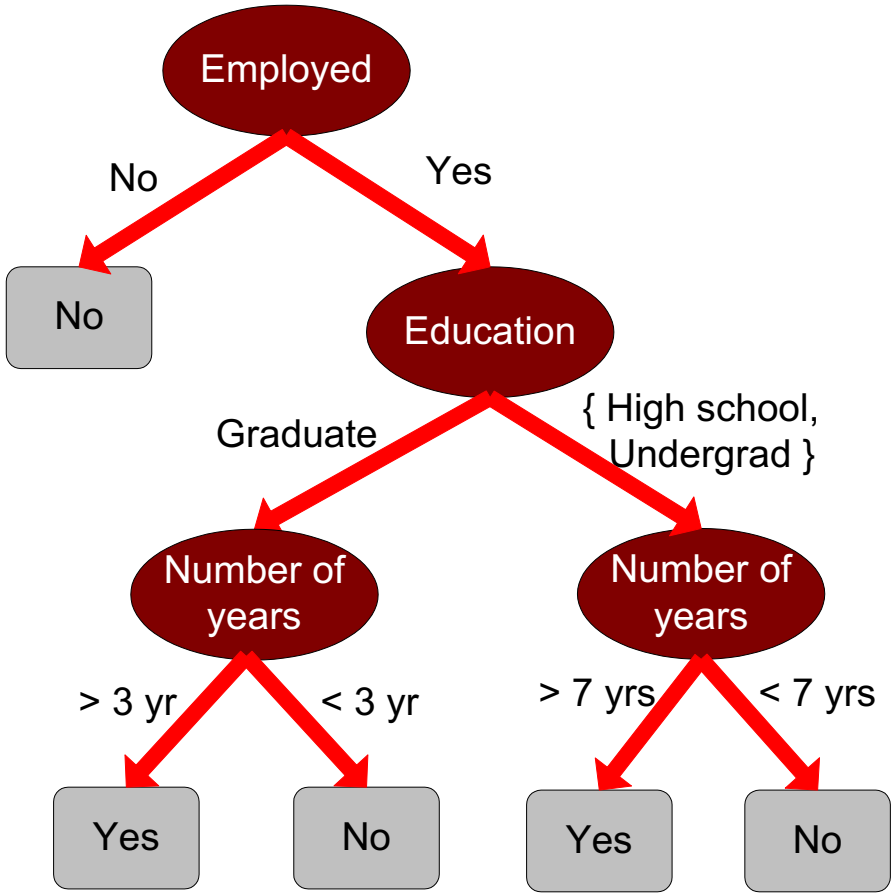
Mô hình dự báo: Phân lớp/ Classification

- Tìm kiếm mô hình phân lớp là hàm của các thuộc tính dữ liệu đầu vào.

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness



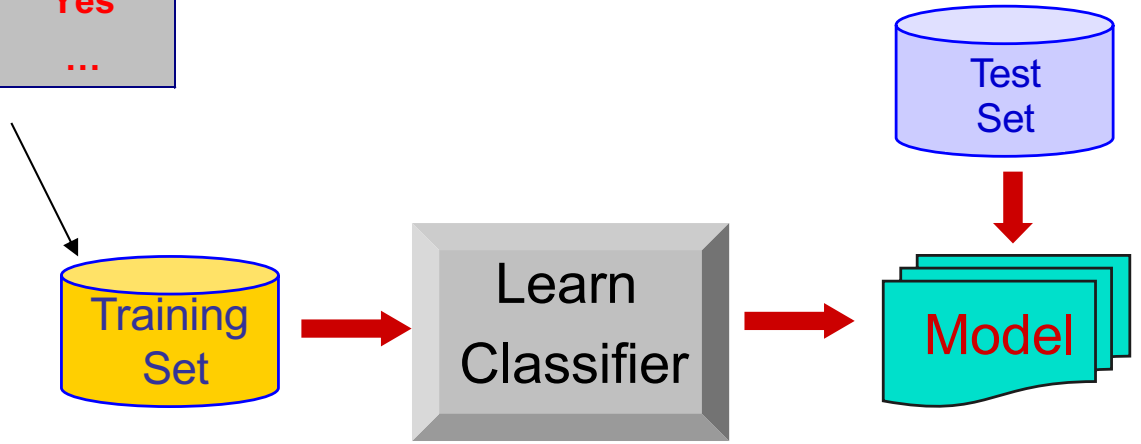
Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Bài toán phân lớp

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

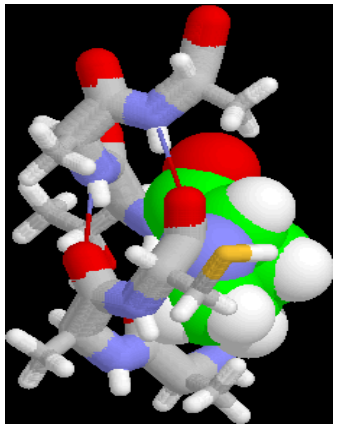
Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Bài toán phân lớp

- Classifying **credit card transactions** as legitimate or fraudulent
- Classifying **land covers** (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as **finance, weather, entertainment, sports**, etc
- Identifying intruders in the **cyberspace**
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of **protein** as alpha-helix, beta-sheet, or random coil



Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Phân lớp: Ứng dụng 1

- **Phát hiện gian lận/ Fraud** Detection
  - **Mục đích:** Dự báo các trường hợp gian lận trong giao dịch thẻ tín dụng.
  - **Phương pháp:**
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Phân lớp: Ứng dụng 2

- **Churn** prediction for telephone customers
  - **Mục đích:** To predict whether a customer is likely to be lost to a competitor.
  - **Phương pháp:**
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

Phân lớp: Ứng dụng 3

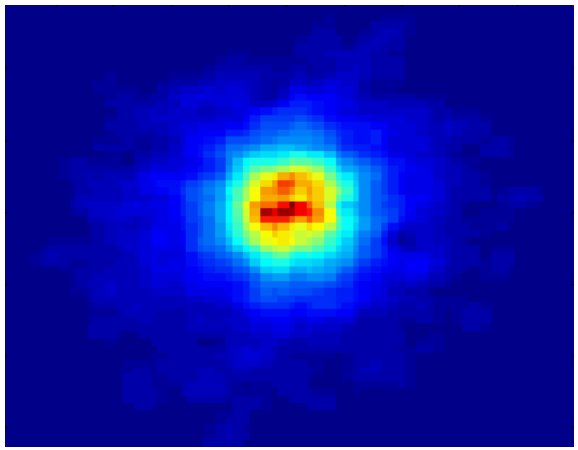
- Sky Survey Cataloging
  - **Mục đích:** To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Phương pháp:**
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Phân loại thiên hà/ Classifying Galaxies

Courtesy: http://aps.umn.edu

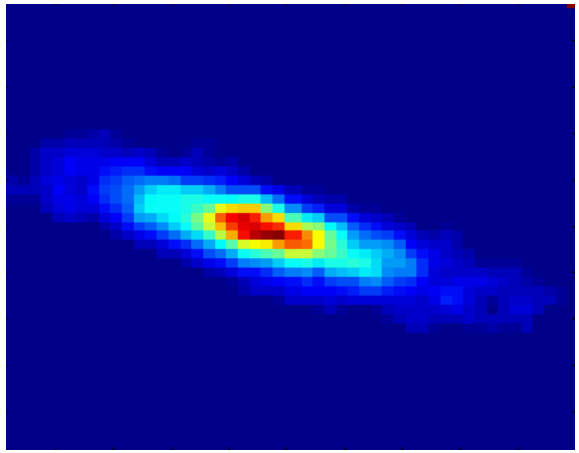
Early



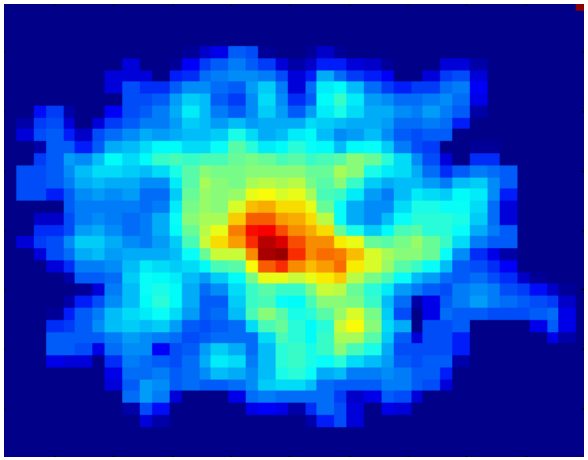
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

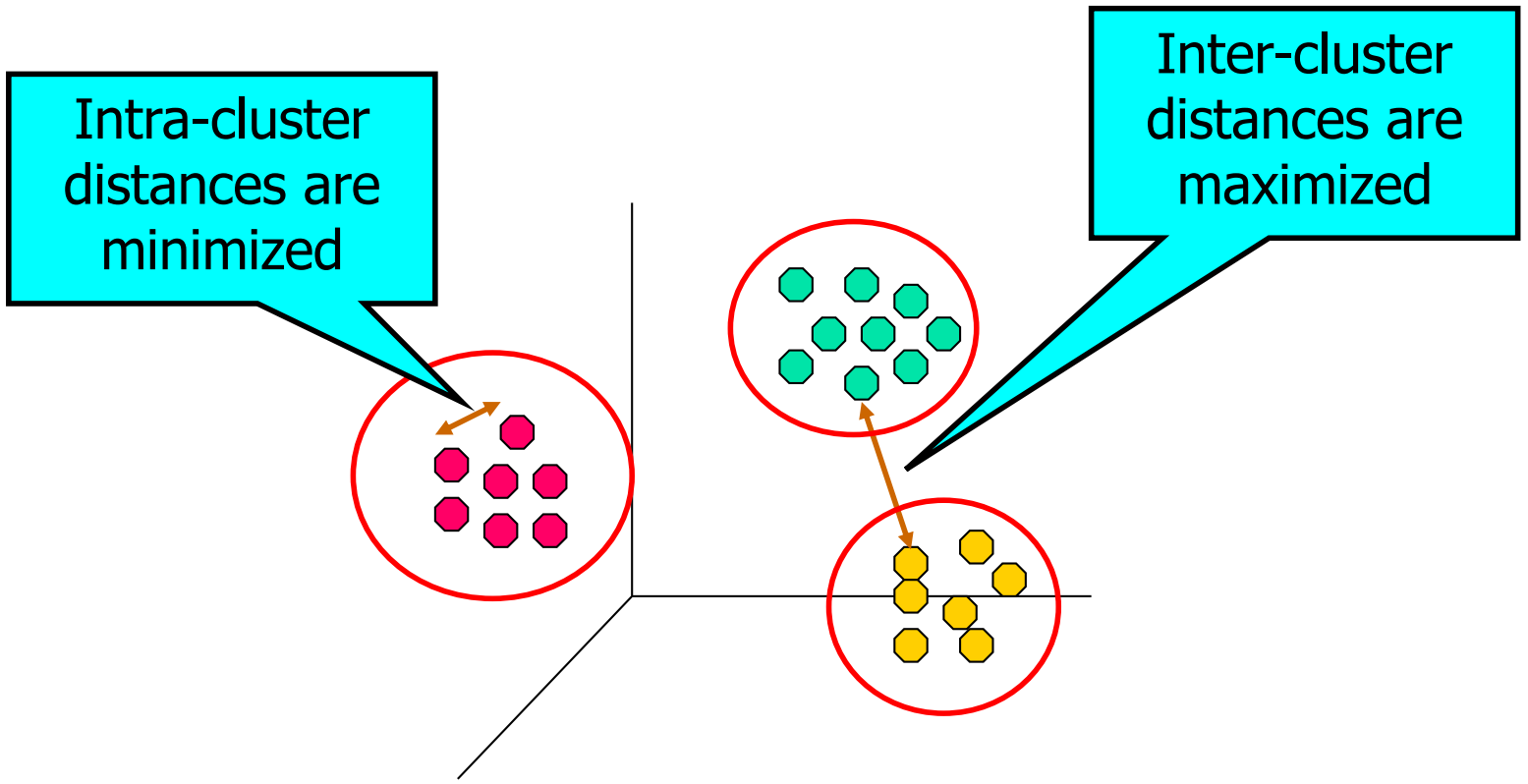
Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Bài toán Hồi qui/ Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Bài toán Gom cụm/ Clustering

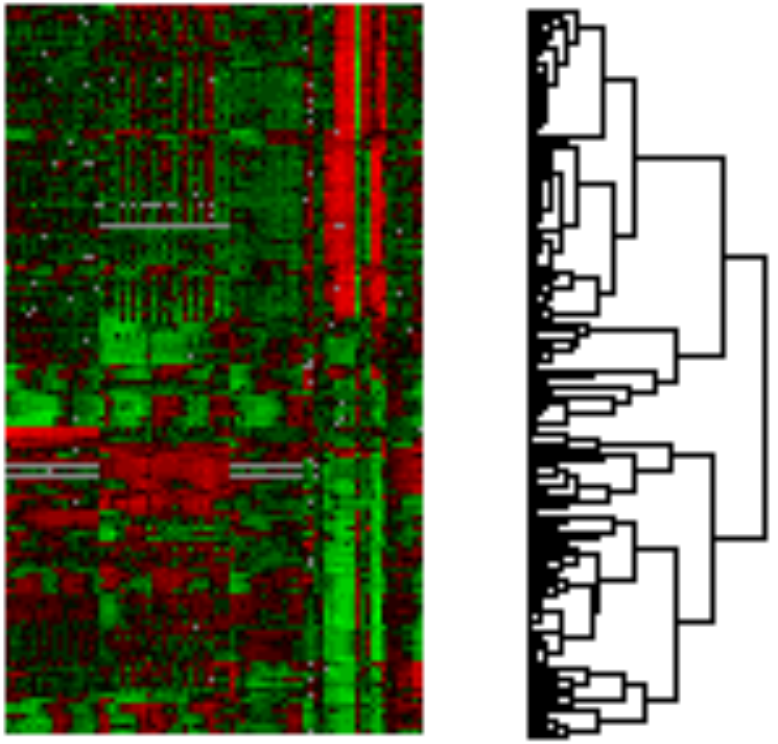
- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



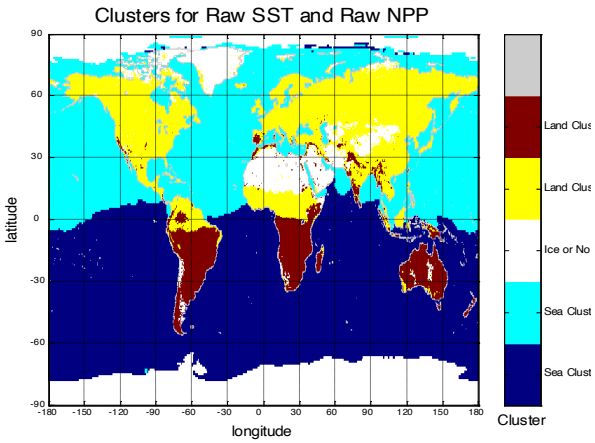
Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Ứng dụng bài toán gom cụm

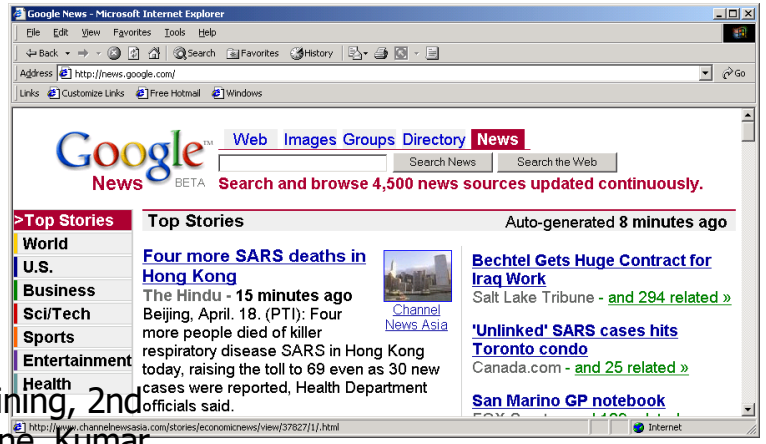
- **Understanding**
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- **Summarization**
  - Reduce the size of large data sets



Courtesy: Michael Eisen



Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.



Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

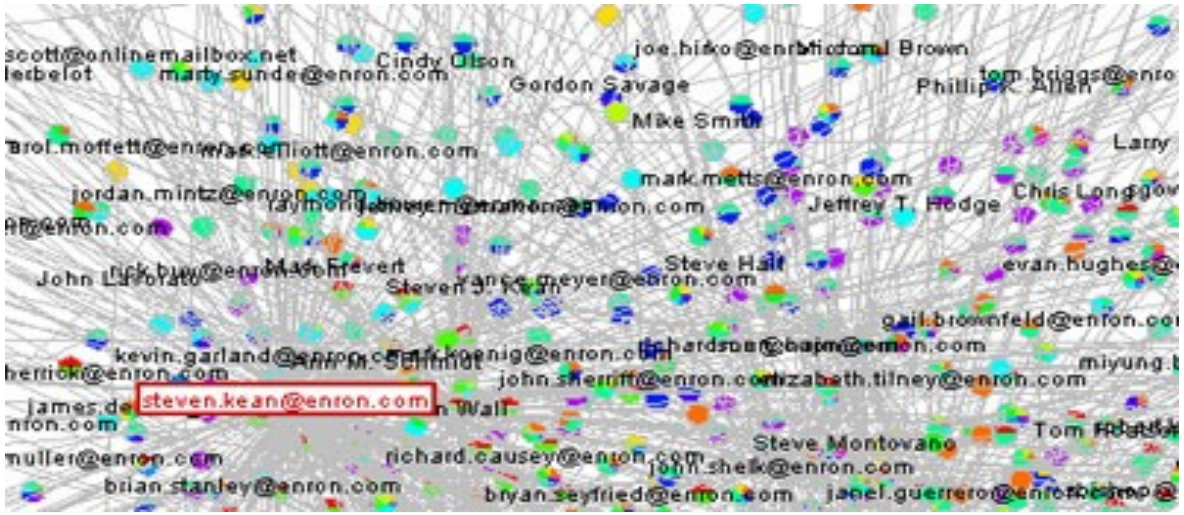
Gom cụm: Ứng dụng 1

- Phân khúc thị trường/ Market Segmentation:
  - **Mục đích:** subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - **Phương pháp:**
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Gom cụm: Ứng dụng 2

- Gom cụm tài liệu/ Document Clustering:
  - **Mục đích:** To find groups of documents that are similar to each other based on the important terms appearing in them.
  - **Phương pháp:** To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset



Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Bài toán Khai thác luật kết hợp

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Khai thác luật kết hợp

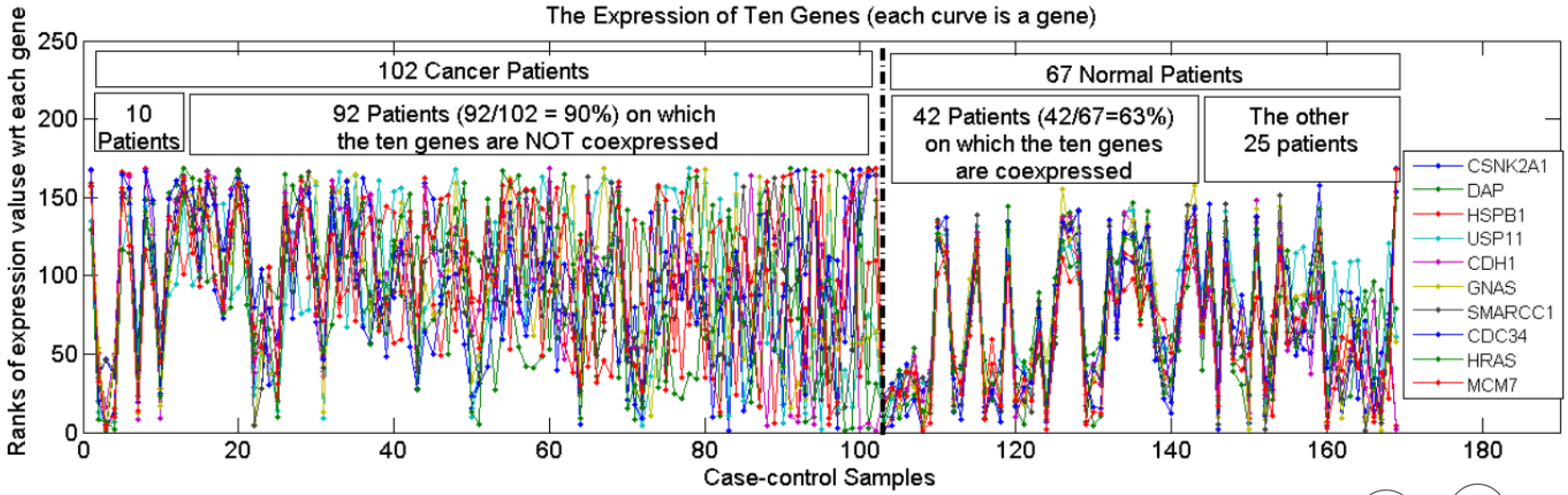
- **Phân tích rổ thị trường**/ Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- **Phát hiện báo động viễn thông**/ Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- **Tin học y tế**/ Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

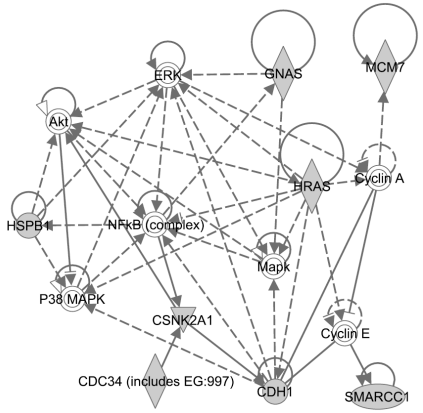
Luật kết hợp: Ứng dụng

- An Example Subspace Differential Co-expression Pattern from lung cancer dataset

Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]



Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

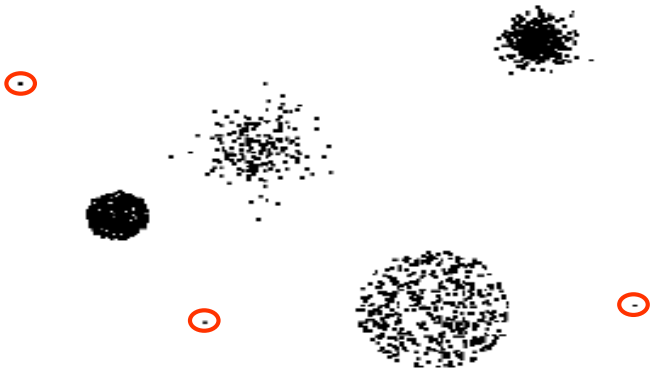


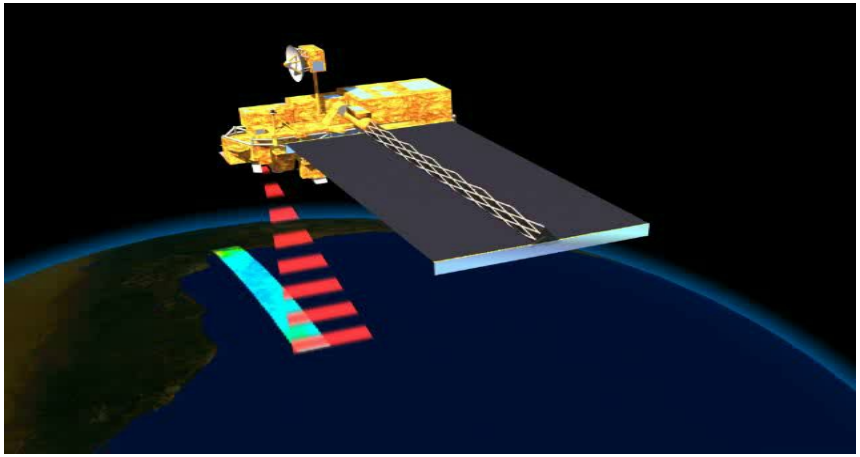
[Fang et al PSB 2010]

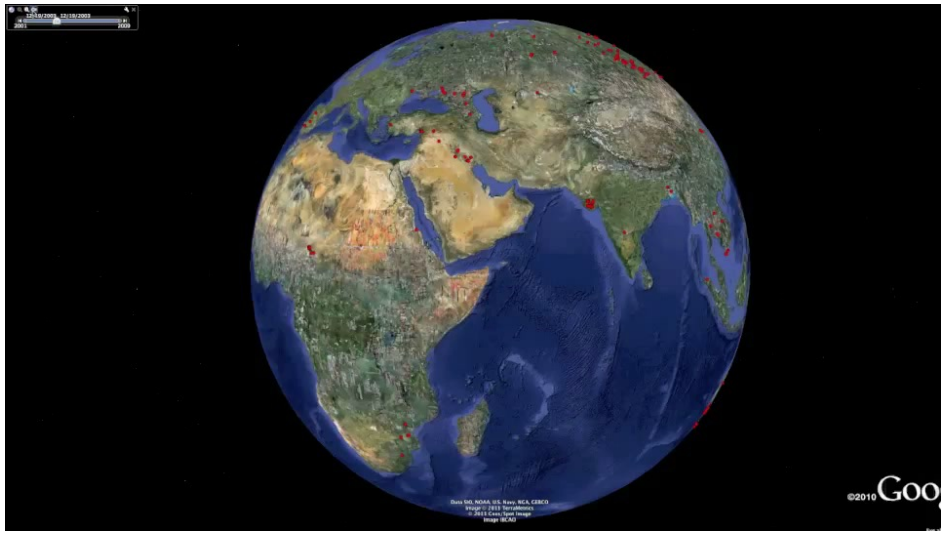
Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.







Nguồn: Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar

5. Tất cả các mẫu được khám phá có "thú vị" không?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty.

6. Khai thác dữ liệu: Classification Schemes

- Different views, different classifications
  - Kinds of data to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

Góc nhìn đa chiều về Khai thác dữ liệu

- Dữ liệu được khai thác/ Data to be mined
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, WWW

- Tri thức được khai thác/ Knowledge to be mined
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels

Góc nhìn đa chiều về Khai thác dữ liệu

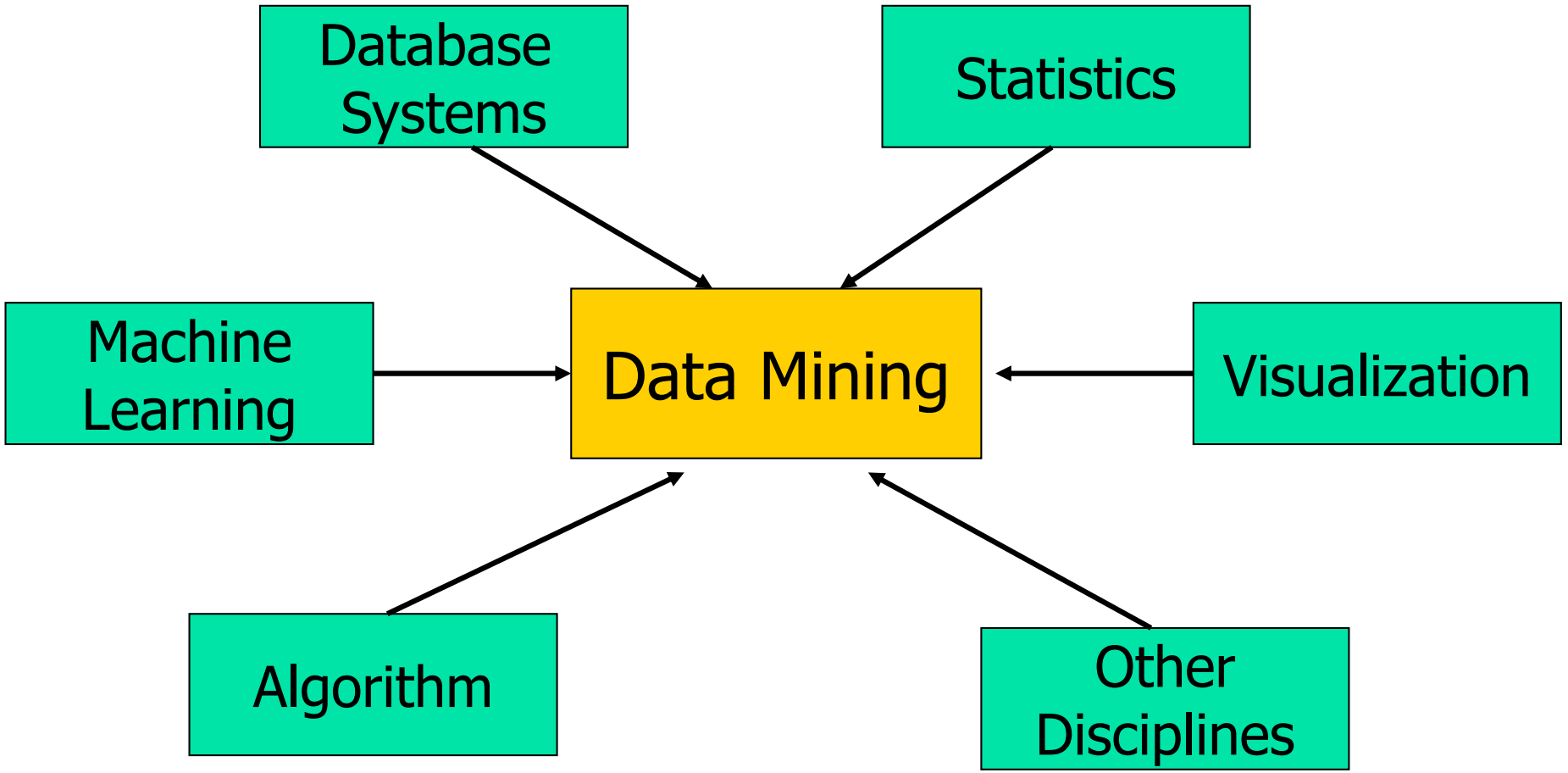
- Kỹ thuật được sử dung/ Techniques utilized
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.

- Ứng dung được cập nhật/ Applications adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, Web mining, etc.

OLAP Mining: Integration of Data Mining and Data Warehousing

- Data mining systems, DBMS, Data warehouse systems coupling
- On-line analytical mining data
  - Integration of mining and OLAP technologies
- Interactive mining multi-level knowledge
  - Necessity of mining knowledge and patterns at different levels of abstraction.
- Integration of multiple mining functions
  - Characterized classification, first clustering and then association

Data Mining: Confluence of Multiple Disciplines



7. Một số thách thức/ challenges

- Phương thức khai thác/ Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion

7. Một số thách thức

- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction

- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

Tổng kết

- Data mining: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures
- Major issues in data mining

Tham khảo/ References?

- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- Data mining and KDD
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations
- Database systems
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: ACM-TODS, IEEE-TKDE, JIIS, J. ACM, etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.