Chương 1: TỔNG QUAN

Nguyễn Tấn Phú
ntanphu@ctuet.edu.vn

Bộ môn HTTT
Khoa CNTT - Đai học KT - CN Cần Thơ

Nội dung

1. Lĩnh vực nghiên cứu của nhà phân tích dữ liệu
2. Nguồn góc, tri thức, loại dữ liệu
3. Tiến trình phân tích dữ liệu
4. Phân tích định tính và định lượng

Lĩnh vực nghiên cứu của nhà phân tích dữ liệu

- Phân tích và xử lý dữ liệu (Data Analysis and Preprocessing)
- Khám phá dữ liệu (Data Exploration)
- Mô hình hóa dữ liệu (Data Modeling)
- Máy học và Học sâu (Machine Learning and Deep Learning)
- Khám phá tri thức (Knowledge Discovery)
- Trực quan hóa (Visualization)
- Phân tích định tính (Qualitative Analysis)

Lĩnh vực nghiên cứu của nhà phân tích dữ liệu (tt)

- Phát hiện thông tin quan trọng (Pattern Discovery)
- Dự báo và tối ưu hóa (Forecasting and Optimization)
- Phân tích định hướng (Sentiment Analysis)
- Phân tích xã hội học (Social Network Analysis)
- Phân tích dự trù (Forecasting Analysis)
- Phân tích thời gian (Time Series Analysis)

Phân tích và xử lý dữ liệu (Data Analysis and Preprocessing)

- **Giai đoạn tiền xử lý dữ liệu:** Quá trình xử lý dữ liệu thô/gốc (raw/original data) nhằm cải thiện chất lượng dữ liệu (quality of the data):
  - Làm sạch giữ liệu
  - Tích hợp dữ liệu
  - Biến đổi dữ liệu
  - Thu giảm dữ liệu
- **Phân tích dữ liệu:** Quá trình sử dụng các phương pháp, công cụ và kỹ thuật để khám phá, hiểu và trích xuất thông tin hữu ích từ tập dữ liệu. Đưa ra dự đoán, hỗ trợ quyết định thông qua việc tìm ra các mẫu, xu hướng và mối quan hệ trong dữ liệu.

Khám phá dữ liệu (Data Exploration)

- Quá trình sử dụng các phương pháp và công cụ thống kê và trực quan để khám phá và hiểu sâu hơn về tập dữ liệu trước khi tiến hành các phân tích chi tiết.
- Mục tiêu chính của Khám phá dữ liệu (Data Exploration): Tìm hiểu các đặc điểm, mẫu, xu hướng, và tương quan trong dữ liệu một cách sơ bộ để xác định các điểm quan trọng và định hướng cho việc phân tích và xử lý dữ liệu.
- Các bước trong quá trình khám phá dữ liệu bao gồm:
  - **Xem xét tổng quan:** Tập dữ liệu bằng cách xem thông tin cơ bản như số lượng dòng, số cột, loại dữ liệu của các cột, v.v.

Khám phá dữ liệu (Data Exploration) (tt)

- Các bước trong quá trình khám phá dữ liệu bao gồm:
  - **Tạo Biểu đồ và Đồ thị:** Sử dụng biểu đồ cột, biểu đồ tròn, biểu đồ đường và các đồ thị khác để hình dung dữ liệu và tìm kiếm sự biến đổi và xu hướng.
  - **Phân tích Thống kê:** Sử dụng các phép đo thống kê như mean, median, mode, phương sai, độ lệch chuẩn để hiểu về phân phối và tính chất của dữ liệu.
  - **Khám phá tương quan:** Xác định các mối quan hệ tương quan giữa các biến bằng cách sử dụng ma trận tương quan hoặc biểu đồ tương quan.
  - **Phát hiện dữ liệu thiếu:** Kiểm tra và xử lý các giá trị bị thiếu trong dữ liệu để không ảnh hưởng đến kết quả khám phá.

Khám phá dữ liệu (Data Exploration) (tt)

- Các bước trong quá trình khám phá dữ liệu bao gồm:
  - **Tìm kiếm mẫu và xu hướng:** Sử dụng các biểu đồ chuỗi thời gian, biểu đồ phân phối để tìm kiếm mẫu và xu hướng trong dữ liệu.
  - **Phân tích dữ liệu ngoại lai:** Xác định và xử lý với các giá trị ngoại lai có thể ảnh hưởng đến kết quả phân tích.
  - **Tạo báo cáo:** Tạo các báo cáo, biểu đồ và đồ thị để trình bày kết quả khám phá dữ liệu.

Mô hình hóa dữ liệu (Data Modeling)

- Quy trình tạo ra một mô hình dữ liệu cho một hệ thống thông tin bằng cách áp dụng một số kỹ thuật chính thức nhất định.
- Các loại mô hình dữ liệu:
  - **Mô hình dữ liệu khái niệm (Conceptual Data Models):** Là một mô hình trừu tượng của dữ liệu, tập trung vào việc biểu diễn cấu trúc dữ liệu, quan hệ giữa thực thể và các ràng buộc dữ liệu cơ bản.
  - ☞ **Thành phần của mô hình dữ liệu khái niệm:** Thực thể (Entities), Thuộc tính (Attributes), Mối quan hệ (Relationships), Ràng buộc (Constraints)

Mô hình hóa dữ liệu (Data Modeling) (tt)

❖ Các loại mô hình dữ liệu:

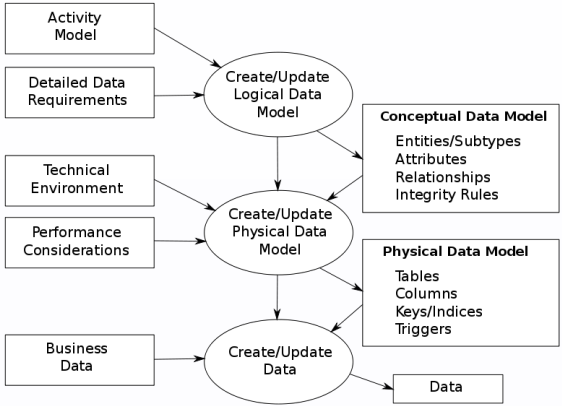
- **Mô hình dữ liệu logic (Logical Data Models):** Hiểu rõ hơn về cấu trúc dữ liệu và mối quan hệ giữa chúng một cách logic, dễ dàng triển khai cấu trúc dữ liệu và tạo các truy vấn hiệu quả.

☞ **Thành phần của mô hình dữ liệu logic:** Bảng (Tables), Cột (Columns), Khóa chính (Primary Keys), Khóa ngoại (Foreign Keys)

- **Mô hình dữ liệu vật lý (Physical Data Models):** Mô tả cách hệ thống sẽ được triển khai bằng cách sử dụng một hệ thống quản lý cơ sở dữ liệu cụ thể. Mô hình này thường được tạo bởi chuyên viên quản trị dữ liệu và các nhà phát triển. Mục đích là triển khai thực tế cơ sở dữ liệu.

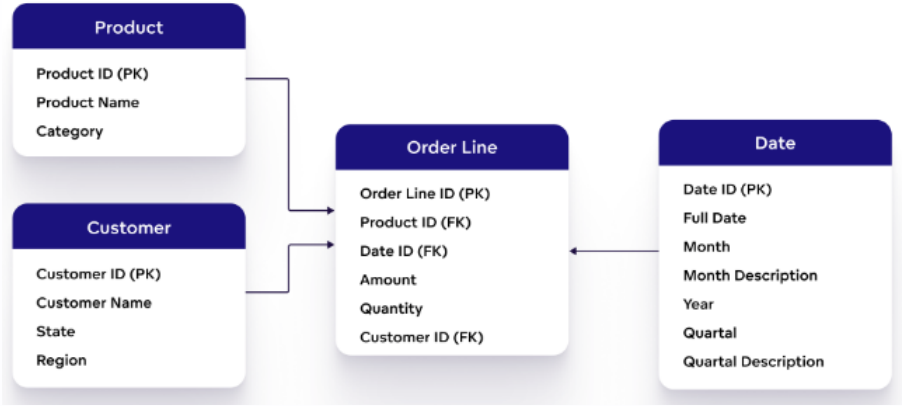
Mô hình hóa dữ liệu (Data Modeling) (tt)

- Quy trình mô hình hóa dữ liệu, Hình ảnh minh họa cách mô hình dữ liệu được phát triển và sử dụng ngày nay:



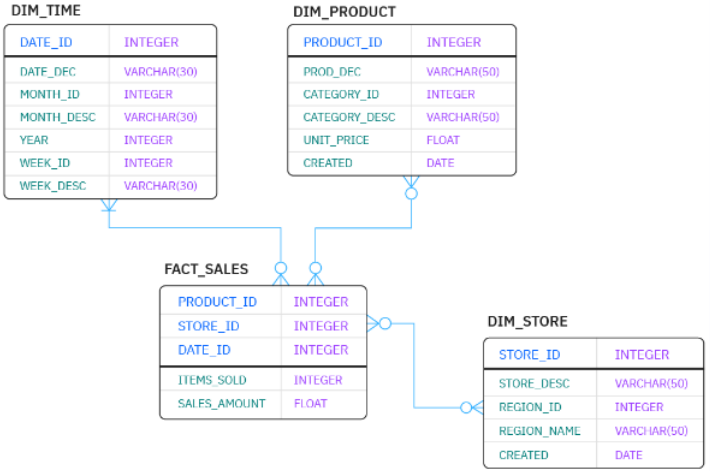
Mô hình hóa dữ liệu (Data Modeling) (tt)

- Mô hình dữ liệu logic (Logical Data Models):



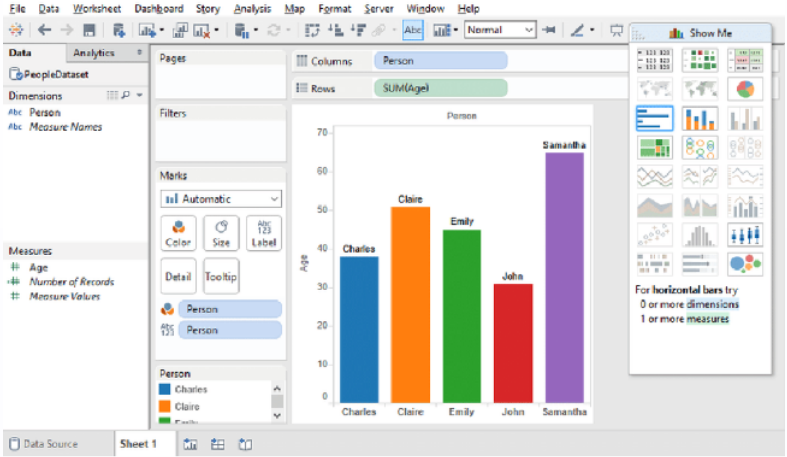
Mô hình hóa dữ liệu (Data Modeling) (tt)

- Mô hình dữ liệu vật lý (Physical Data Models):



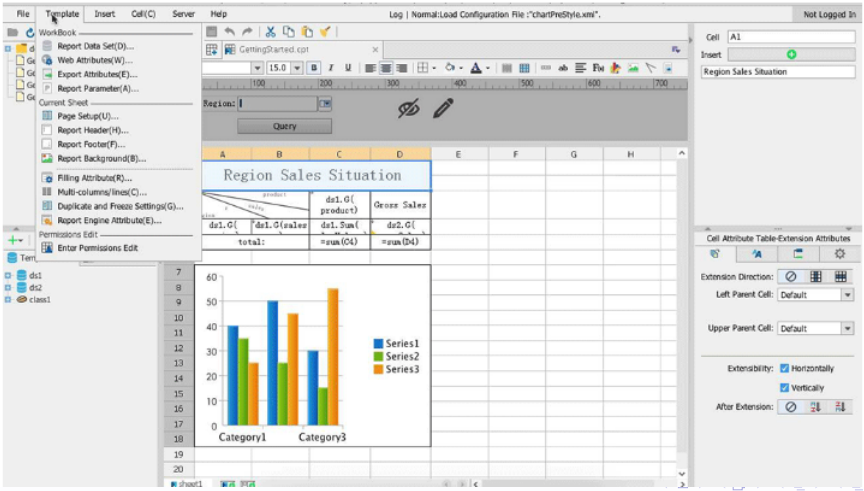
Công cụ ứng dụng cho Data Model

- Tableau:



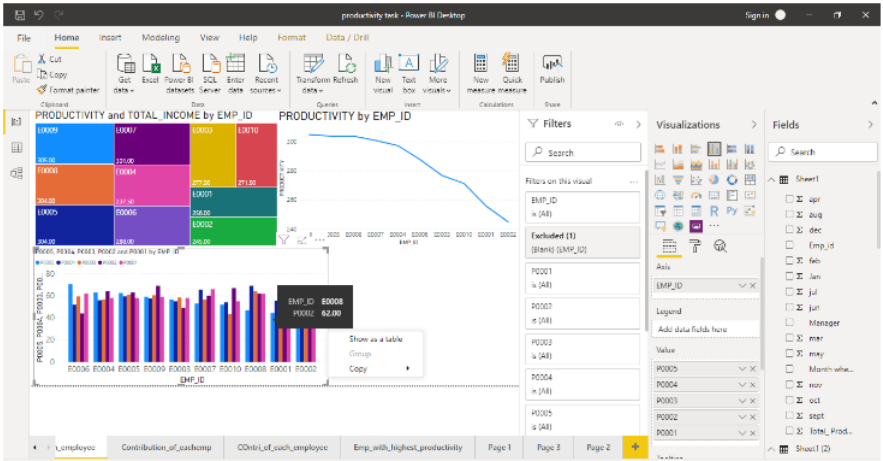
Công cụ ứng dụng cho Data Model

- FineReport:



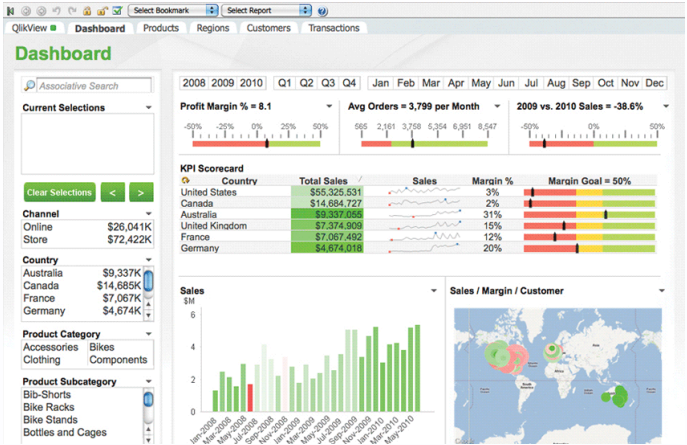
Công cụ ứng dụng cho Data Model

- Power BI:



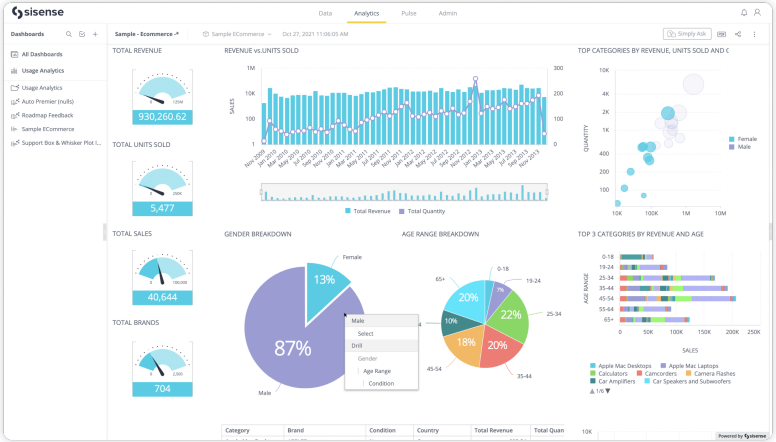
Công cụ ứng dụng cho Data Model

- QlikView:



Công cụ ứng dụng cho Data Model

- Sisense:



Máy học và Học sâu (Machine Learning and Deep Learning)

- **Học máy (Machine Learning – ML):** Là một lĩnh vực nghiên cứu của Trí tuệ nhân tạo (Artificial Intelligence – AI):
  - Học có giám sát (Supervised Learning)
  - Học không giám sát (Unsupervised Learning)
  - Học bán giám sát (Semi-Supervised Learning)
- **Học sâu (Deep Learning):** Là một nhánh con của máy học, tập trung vào việc xây dựng và huấn luyện các mạng neuron nhân tạo để mô hình hóa các đặc trưng phức tạp từ dữ liệu.

Khám phá tri thức (Knowledge Discovery)

- Là quá trình tìm ra thông tin hữu ích, tri thức mới và có giá trị từ các nguồn dữ liệu lớn và phức tạp.
- Quá trình này bao gồm nhiều bước khác nhau (thu thập dữ liệu, xử lý dữ liệu, phân tích và khám phá thông tin tiềm ẩn).
- Có thể áp dụng trong nhiều lĩnh vực như khoa học dữ liệu, kinh doanh thông minh, y tế, và nhiều ngành khác.

Trực quan hóa (Visualization)

- **Trực quan hóa dữ liệu (đồ thị hoá dữ liệu/hình ảnh hoá dữ liệu) là quá trình sử dụng các yếu tố hình ảnh như:**
  - Đồ thị
  - Biểu đồ hoặc bản đồ để trình bày dữ liệu thô
  - Biến các con số rời rạc, đơn lẻ thành một bức tranh tổng có thể bóc tách thông tin nhanh.

Nguồn gốc dữ liệu

- **Tệp – Files:**
  - Excel; Text/CSV; XML; JSON
- **Cơ sở dữ liệu – Database:**
  - SQL Server Database; Access Database; SQL Server Analysis Services Database; Oracle Database; IBM DB2 Database; IBM Informix database Beta; IBM Netezza Beta; MySQL Database; PostgreSQL Database; Sybase Database; Teradata Database; SAP HANA Database; SAP Business Warehouse server; Amazon Redshift Impala; Google BigQuery Beta; Snowflake.

Nguồn gốc dữ liệu

- **Dịch vụ đám mây - Cloud Services:**
  - SharePoint Online List; Microsoft Exchange Online; Dynamics 365 online; Common Data Service; Salesforce Objects; Salesforce Reports; Google Analytics; appFigures Beta; comScore Digital Analytix; Dynamics 365 for Customer Insights; Facebook; GitHub Beta.

Một số loại dữ liệu

- **Categorical (Kiểu liệt kê):**
  - Dữ liệu là một tập xác định các giá trị.
  - **VD**: màu mắt (xanh, nâu, đen); xếp hạng (tốt, trung bình, xấu); chiều cao (cao, trung bình, thấp).
  - Nhị phân - Binary (Yes/No, True/False).
- **Numeric – kiểu số:**
  - Dữ liệu nhiệt độ, thời gian (ngày, giờ, năm) chiều dài, chiều cao.
  - Rời rạc - Discrete (7 ngày trong tuần).
  - Liên tục - Continuous ($37^0, 37.5^0, 37.3^0, \ldots$).
  - Nhị phân - Binary (0/1).

Một số loại dữ liệu

❖ Dữ liệu văn bản:

- Mỗi văn bản là một vector chứa các từ/ thuật ngữ ("term"), giá trị của mỗi thuộc tính chính là số lần xuất hiện của các từ/thuật ngữ ("term") trong văn bản.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Một số loại dữ liệu

❖ Dữ liệu giao dịch:

- Mỗi giao dịch chứa tập hợp các mục dữ liệu (item)).

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

- Tập các mục dữ liệu (items) được biểu diễn dưới dạng **binary vector**.

TID	Bread	Coke	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Một số loại dữ liệu

❖ Dữ liệu có thứ tự (Ordered Data):

- Dữ liệu gen (Genomic sequence data): Dữ liệu là chuỗi có thứ tự các ký tự.

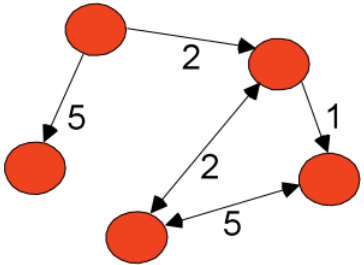
```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

- Dữ liệu chuỗi thời gian (Time series).

Một số loại dữ liệu

❖ **Dữ liệu đồ thị:**

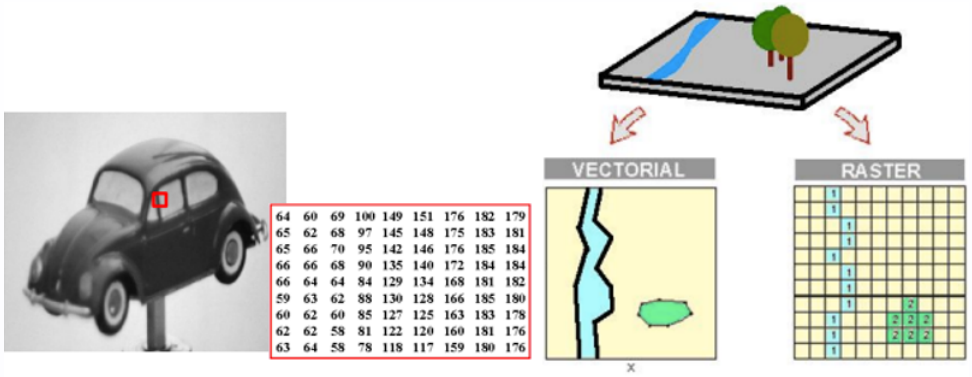
- Web graph and HTML Links.



 Data Mining
  Graph Partitioning
  Parallel Solution of Sparse Linear System of Equations
 <a href="papers/papers.html#ffff"> N-Body Computation and Dense Linear System Solvers

Một số loại dữ liệu

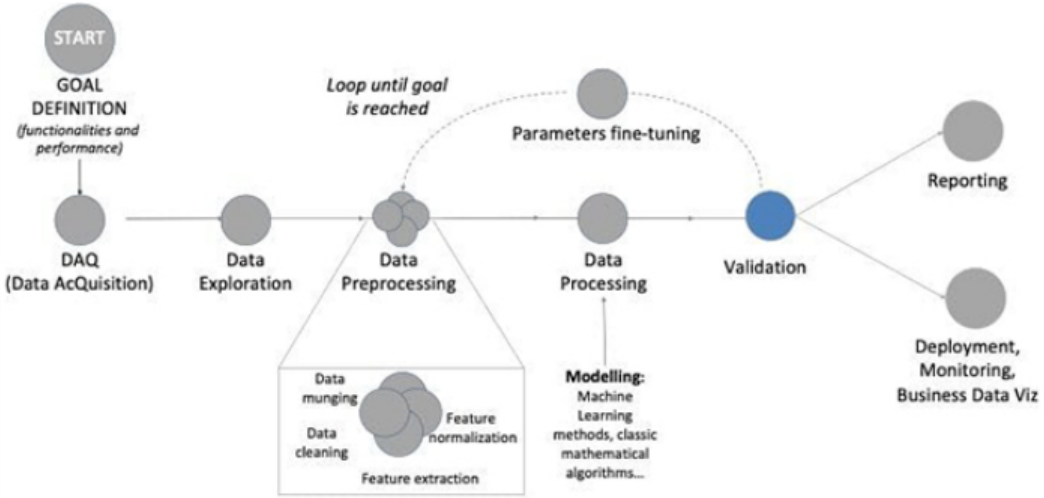
- **Không gian, ảnh và đa phương tiện:**
  - Dữ liệu không gian: bản đồ GIS (raster, vector)
  - Dữ liệu ảnh, Dữ liệu Video, Audio, ...



Tri thức

- **Tri thức đat được từ quá trình khai phá:**
  - Tri thức đạt được có thể có tính mô tả hay dự đoán tùy thuộc vào quá trình khai phá cụ thể.
    - Mô tả (Descriptive): có khả năng đặc trưng hóa các thuộc tính chung của dữ liệu được khai phá.
    - Dự đoán (Predictive): có khả năng suy luận từ dữ liệu hiện có để dự đoán.
  - Tri thức đạt được có thể có cấu trúc, bán cấu trúc, hoặc phi cấu trúc
  - Tri thức đạt được có thể được/không được người dùng quan tâm $\Rightarrow$ các độ đo đánh giá tri thức đạt được.
  - Tri thức đạt được có thể được dùng trong việc hỗ trợ ra quyết định, điều khiển quy trình, quản lý thông tin, xử lý truy vấn ,...

Tiến trình phân tích dữ liệu



Phân tích định tính và định lượng

❖ **Phân tích định tính (Qualitative Analysis):**

- Phân tích định tính tập trung vào việc nghiên cứu và hiểu các tính chất, đặc điểm, và ngữ cảnh của dữ liệu không đo lường được hoặc dựa vào mô tả chất lượng.
- Dữ liệu định tính thường bao gồm mô tả về các biến số không đo lường bằng các phần tử chữ (như mô tả, loại hình, phân loại).
- Các phương pháp phân tích định tính bao gồm phân tích nội dung (content analysis), phân tích nội dung văn bản, và phân tích thematic (phân tích chủ đề).

Phân tích định tính và định lượng

❖ **Phân tích định lượng (Quantitative Analysis):**

- Phân tích định lượng liên quan đến việc thu thập và xử lý dữ liệu có tính chất đo lường hoặc số lượng.
- Dữ liệu định lượng thường là các con số và có thể được thực hiện các phép tính số học và thống kê trên chúng.
- Các phương pháp phân tích định lượng bao gồm kiểm định giả thuyết, phân tích biến động (variance analysis), hồi quy (regression analysis), và các kỹ thuật học máy.

Phân tích định tính và định lượng

☞ Khi nghiên cứu, quyết định sử dụng phân tích định tính hoặc định lượng thường phụ thuộc vào bản chất của dữ liệu và mục tiêu nghiên cứu:

- Phân tích định tính thường được sử dụng khi muốn hiểu sâu về các yếu tố không đo lường hoặc nghiên cứu các quan điểm, tình cảm, hoặc chất lượng.
- Phân tích định lượng thường được sử dụng khi muốn xác định mối quan hệ giữa các biến số, dự đoán và kiểm tra giả thuyết, và thực hiện phân tích số liệu thống kê.

Question and Answer