TỔNG QUAN

Biên soạn: **ThS. Nguyễn Thị Anh Thư**

Email: thunta@uit.edu.vn

NỘI DUNG

1. **Khái niệm**
2. Ứng dụng
3. Data mining và Big Data
4. Data mining và Data Science
5. Quy trình khám phá tri thức
6. Thách thức
7. Các loại mạng

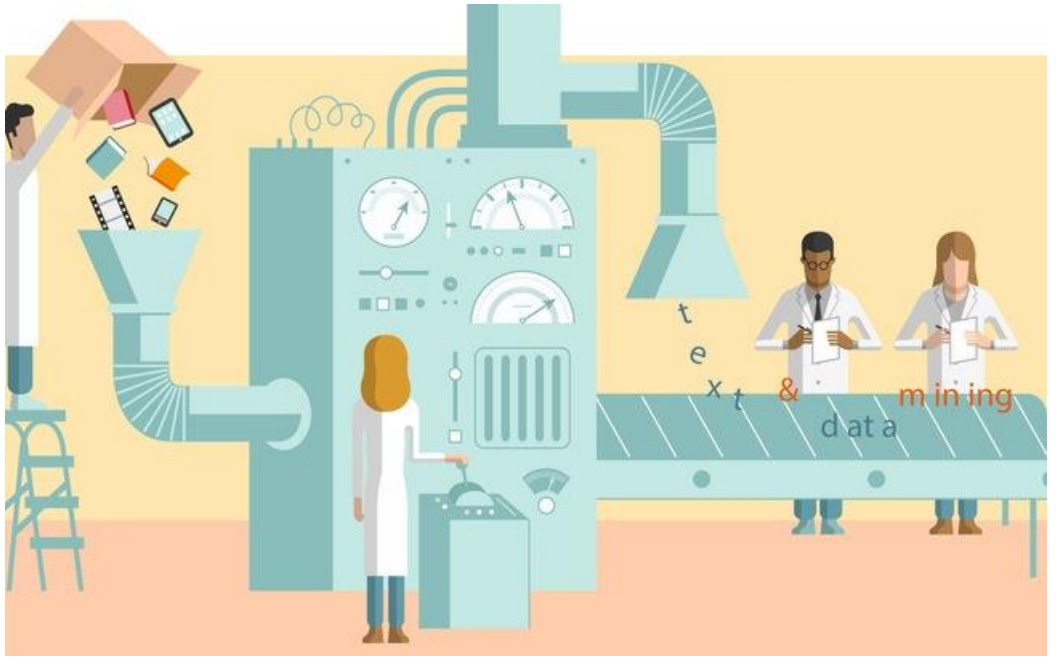
1. KHÁI NIỆM

- Khai thác dữ liệu
- Đồ thị
- Mạng
- Mạng xã hội

KHAI THÁC DỮ LIỆU

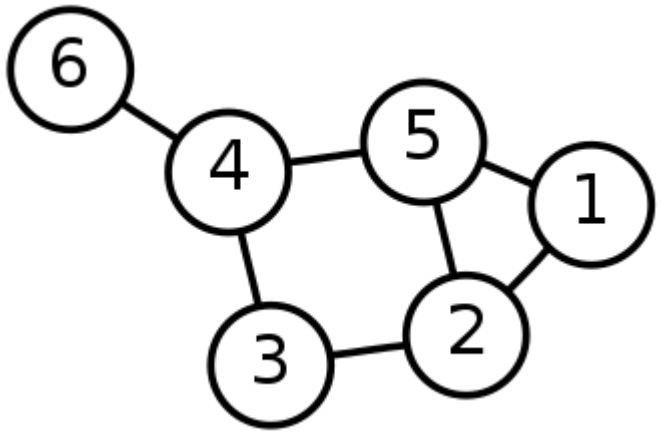
- **Khai thác dữ liệu (*data mining*)** là quá trình khám phá và phân tích khối lượng lớn dữ liệu để tìm ra các mẫu dữ liệu và quy tắc có ý nghĩa.
- **Data mining** là một trong những lĩnh vực:
  - Nghiên cứu khoa học dữ liệu.
  - Khai thác và sử dụng các dữ liệu, thông tin có giá trị từ dữ liệu.

⇒ Phục vụ đưa ra dự báo, quyết định trong tương lai.



ĐỒ THỊ

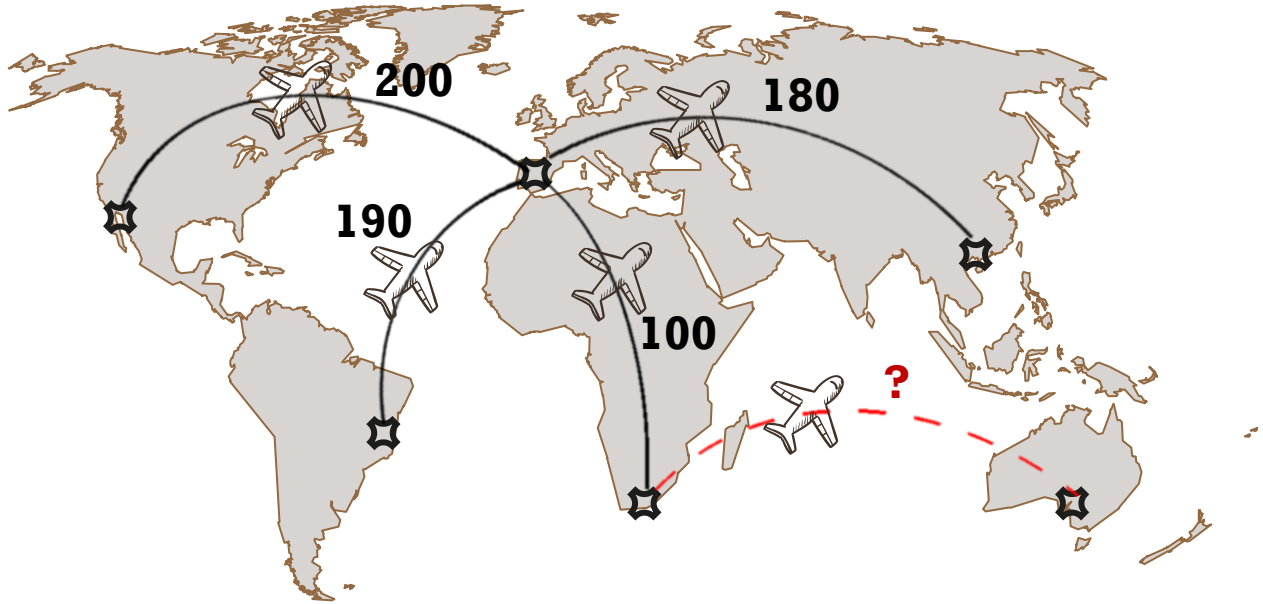
- Đồ thị là một **cấu trúc rời rạc** bao gồm **các đỉnh** và **các cạnh** nối các đỉnh này.



Một đồ thị vô hướng với 6 đỉnh (nút) và 7 cạnh.

MẠNG

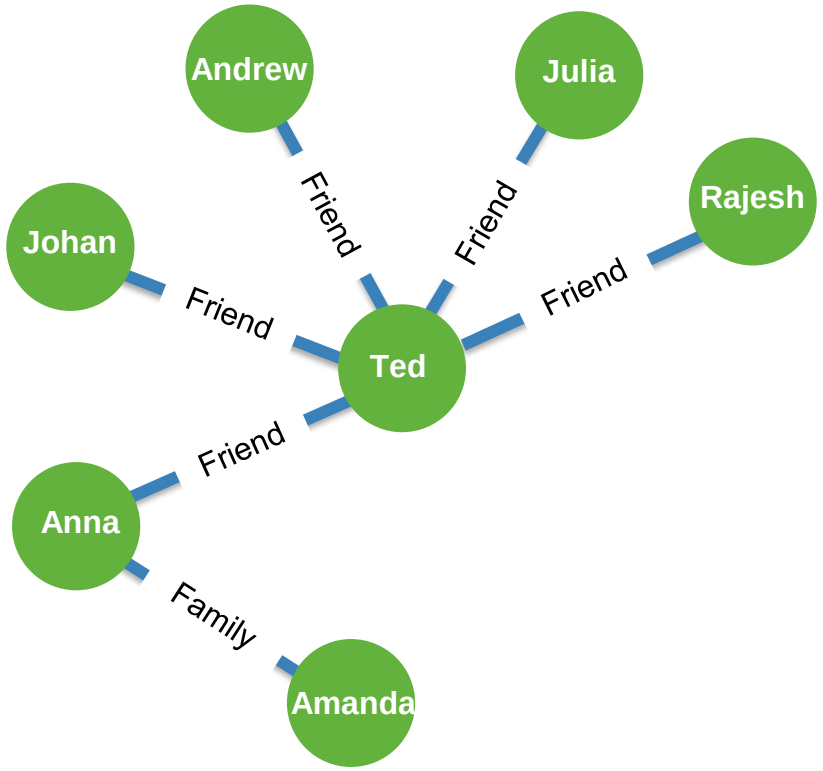
- **Mạng được định nghĩa và có cấu trúc như một đồ thị.**
- Hiện nay có nhiều hệ thống phổ biến mà cấu trúc có thể được biểu diễn như một mô hình mạng như là: ***Internet, mạng xã hội, mạng sinh học, …***



Mạng vận tải hàng không

MẠNG XÃ HỘI

- **Mạng xã hội là một đồ thị mạng** được cấu tạo bởi các đỉnh và các cạnh.
  - **Các đỉnh** là tập các đối tượng như: **người dùng**, bình luận sản phẩm.
  - **Các cạnh** là các liên kết thể hiện **mối quan hệ giữa những đối tượng**.



Quy tắc

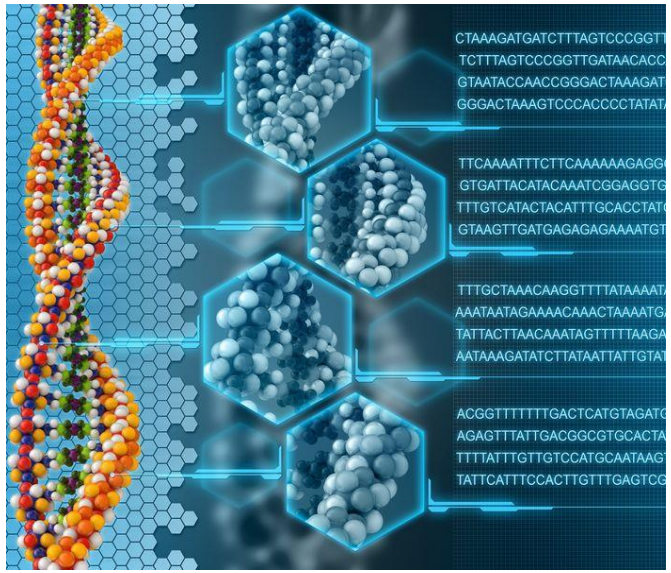
Hành vi

Dịch bệnh

NỘI DUNG

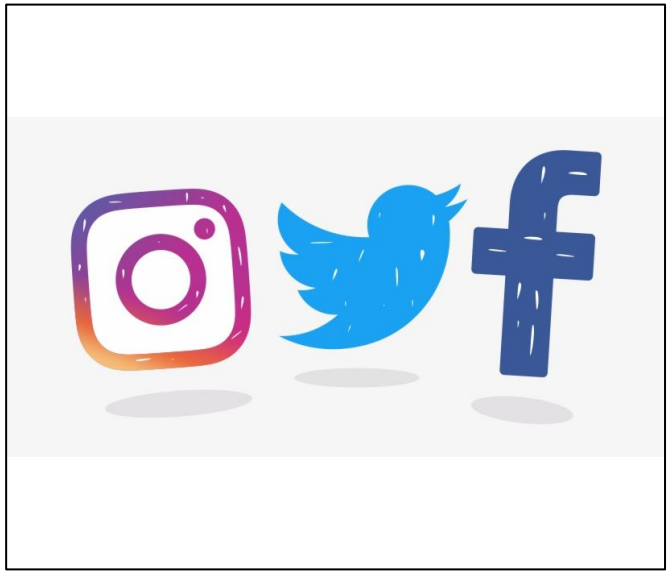
1. Khái niệm
2. **Ứng dụng**
3. Data mining và Big Data
4. Data mining và Data Science
5. Quy trình khám phá tri thức
6. Thách thức
7. Các loại mạng

2. ỨNG DỤNG



Phân tích dữ liệu sinh học

Phát hiện mạng lưới tội phạm



Một số ứng dụng trong xã hội

NỘI DUNG

1. Khái niệm
2. Ứng dụng
3. **Data mining và Big Data**
4. Data mining và Data Science
5. Quy trình khám phá tri thức
6. Thách thức
7. Các loại mạng

3. DATA MINING VÀ BIG DATA

	Data Mining	Big Data
#1. Focus	Tập trung vào những chi tiết của dữ liệu.	Tập trung vào những mối quan hệ giữa các dữ liệu.
#2. View	Một cái nhìn cận cảnh về dữ liệu.	Một bức tranh lớn của dữ liệu.
#3. Data	Diễn tả những nội dung trong dữ liệu.	Thể hiện mối quan hệ trong dữ liệu.
#4. Volume	Có thể sử dụng cho small data hoặc big data.	Đề cập đến số lượng lớn các tập dữ liệu.

3. DATA MINING VÀ BIG DATA

	Data Mining	Big Data
#5. Definition	Là một kỹ thuật phân tích dữ liệu.	Là một khái niệm hơn là một thuật ngữ chính xác.
#6. Data Types	Dữ liệu có cấu trúc, cơ sở dữ liệu quan hệ.	Dữ liệu có cấu trúc, dữ liệu bán cấu trúc và dữ liệu không có cấu trúc (NoSQL).
#7. Analysis	Chủ yếu là phân tích thống kê, tập trung vào dự đoán và khám phá các yếu tố kinh doanh ở quy mô nhỏ.	Chủ yếu phân tích dữ liệu, tập trung vào dự đoán và khám phá các yếu tố kinh doanh trên quy mô lớn.
#8. Results	Chủ yếu để ra quyết định chiến lược.	Bảng điều khiển và các biện pháp dự đoán.

NỘI DUNG

1. Khái niệm
2. Ứng dụng
3. Data mining và Big Data
4. **Data mining và Data Science**
5. Quy trình khám phá tri thức
6. Thách thức
7. Các loại mạng

4. DATA MINING VÀ DATA SCIENCE

	Data Mining	Data Science
#1. What is it?	Một kỹ thuật.	Một lĩnh vực.
#2. Focus	Đề xuất chiến lược kinh doanh.	Nghiên cứu khoa học.
#3. Goal	Khai thác những thông tin hữu ích từ dữ liệu.	Xây dựng một trung tâm dữ liệu cho một tổ chức.
#4. Output	Mẫu dữ liệu và quy tắc có ý nghĩa.	Đa dạng.

4. DATA MINING VÀ DATA SCIENCE

	Data Mining	Data Science
#5. Purpose	Tìm xu hướng trước đây không biết.	Phân tích xã hội, xây dựng các mô hình dự đoán, những sự thật chưa biết và nhiều hơn nữa.
#6. Vocational Perspective	Ai đó có kiến thức về điều hướng qua dữ liệu và hiểu biết thống kê có thể tiến hành khai thác dữ liệu.	Một người cần hiểu về Machine Learning, Lập trình, kỹ thuật thông tin và có kiến thức chuyên ngành IT để trở thành một nhà khoa học dữ liệu.

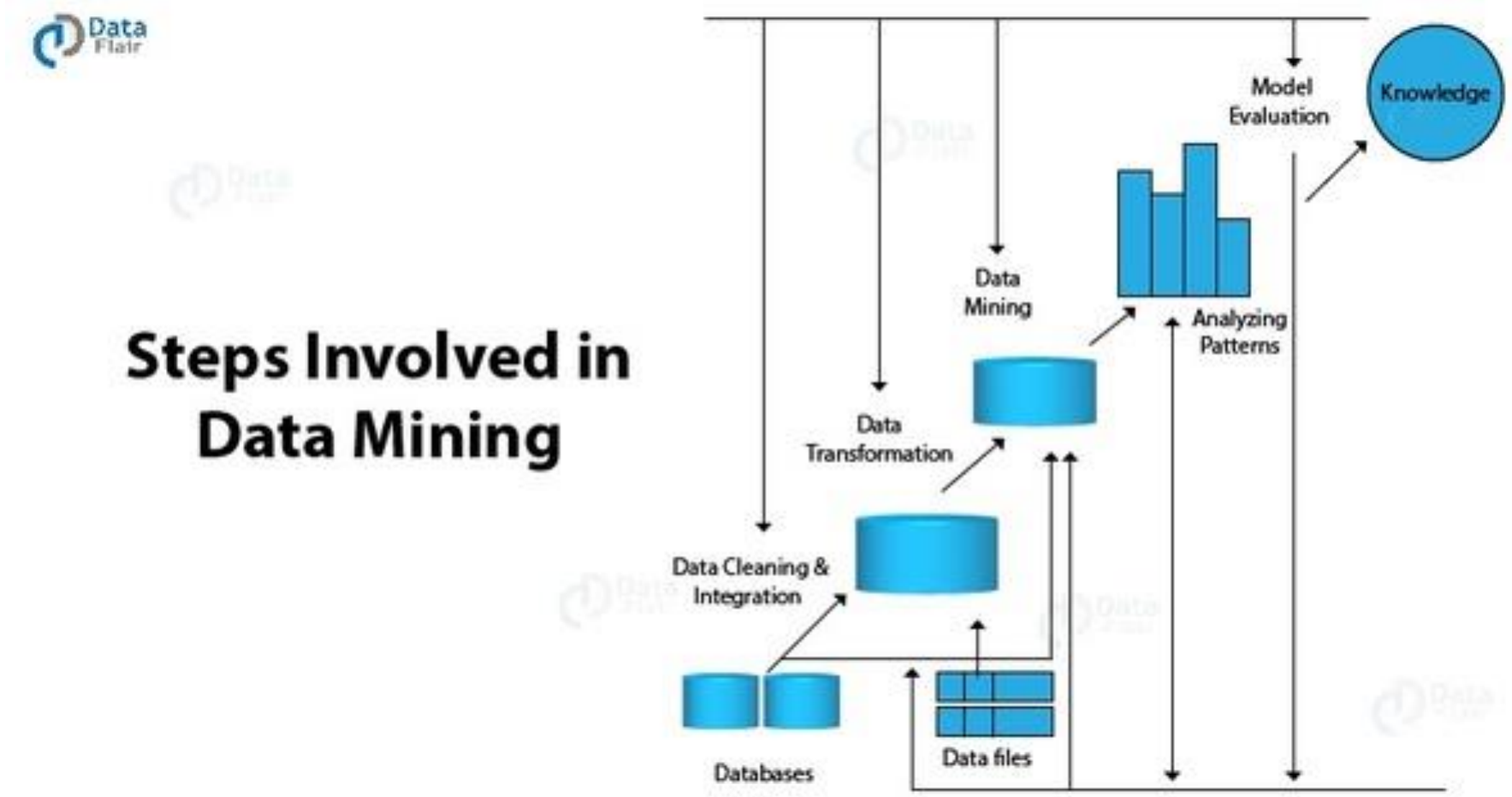
4. DATA MINING VÀ DATA SCIENCE

	Data Mining	Data Science
#7. Extent	Khai thác dữ liệu có thể là một tập hợp con của Khoa học dữ liệu vì các hoạt động Khai thác là một phần của đường ống Khoa học dữ liệu	Đa ngành - Khoa học dữ liệu bao gồm Trực quan hóa dữ liệu, Khoa học xã hội tính toán, Thống kê, Khai thác dữ liệu, Xử lý ngôn ngữ tự nhiên, ...
#8. Type of data	Chủ yếu là dữ liệu có cấu trúc.	Tất cả các dạng dữ liệu (*có cấu trúc, bán cấu trúc và không cấu trúc*).

NỘI DUNG

1. Khái niệm
2. Ứng dụng
3. Data mining và Big Data
4. Data mining và Data Science
5. **Quy trình khám phá tri thức**
6. Thách thức
7. Các loại mạng

5. QUY TRÌNH KHÁM PHÁ TRI THỨC



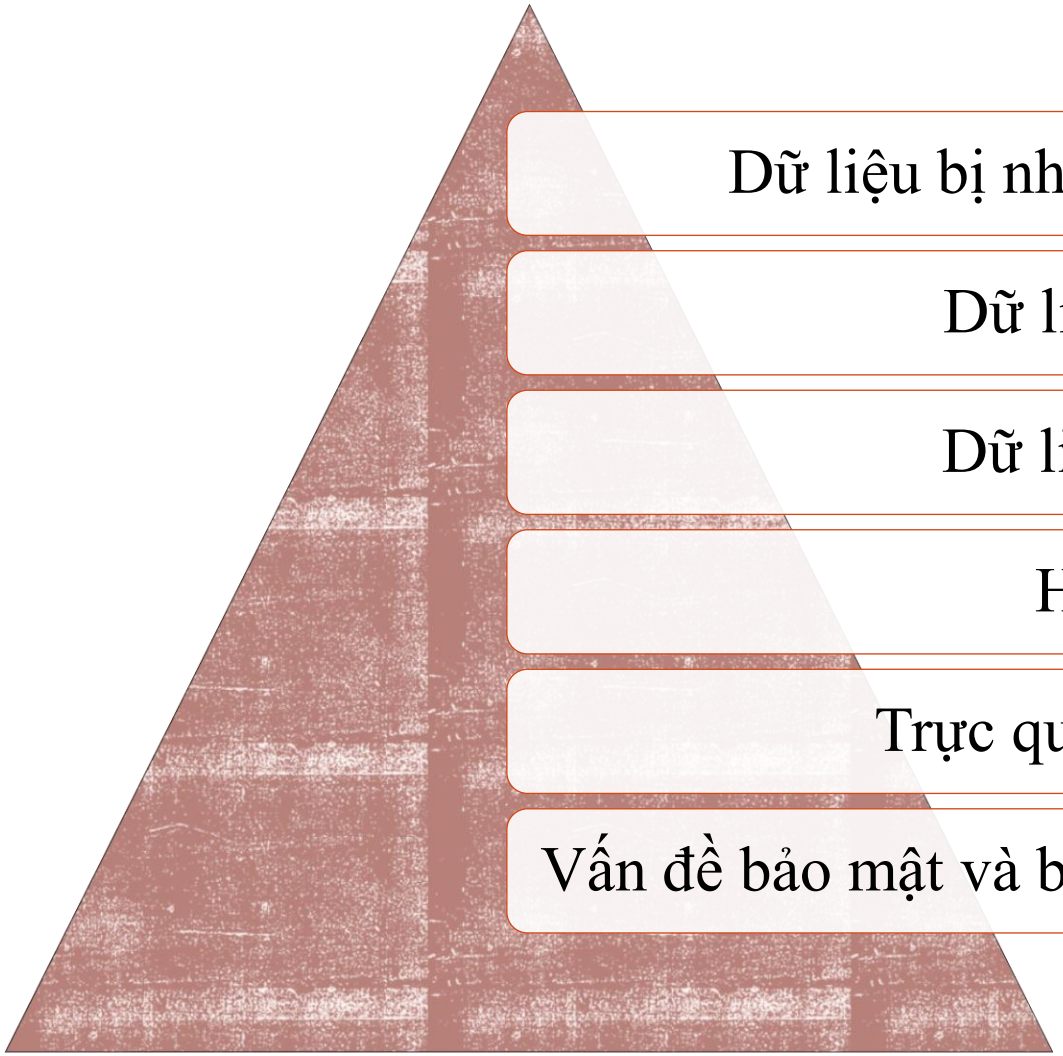
5. QUY TRÌNH KHÁM PHÁ TRI THỨC

- **Bước 1**: **Làm sạch dữ liệu** – Loại dữ liệu nhiễu hoặc bất thường trong dữ liệu.
- **Bước 2**: **Tích hợp dữ liệu** – Nhiều nguồn dữ liệu được kết hợp.
- **Bước 3**: **Lựa chọn dữ liệu** – Đó là về việc trích xuất những dữ liệu có ích.
- **Bước 4**: **Chuyển đổi dữ liệu** – Để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.
- **Bước 5**: **Khai thác dữ liệu** – Trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.
- **Bước 6**: **Đánh giá mẫu** – Phân tích pattern rút trích được có hữu ích.
- **Bước 7**: **Biểu diễn tri thức** – Ở bước cuối cùng, việc trình bày kiến thức cho người dùng được thực hiện dưới dạng cây, bảng, biểu đồ và ma trận.

NỘI DUNG

1. Khái niệm
2. Ứng dụng
3. Data mining và Big Data
4. Data mining và Data Science
5. Quy trình khám phá tri thức
6. **Thách thức**
7. Các loại mạng

6. THÁCH THỨC



- Dữ liệu bị nhiễu và không đầy đủ.
- Dữ liệu phân tán.
- Dữ liệu phức tạp.
- Hiệu suất.
- Trực quan hóa dữ liệu.
- Vấn đề bảo mật và bảo mật dữ liệu (Link video).

NỘI DUNG

1. Khái niệm
2. Ứng dụng
3. Data mining và Big Data
4. Data mining và Data Science
5. Quy trình khám phá tri thức
6. Thách thức
7. **Các loại mạng**

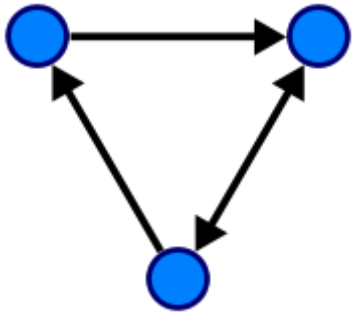
7. CÁC LOẠI MẠNG

Network types

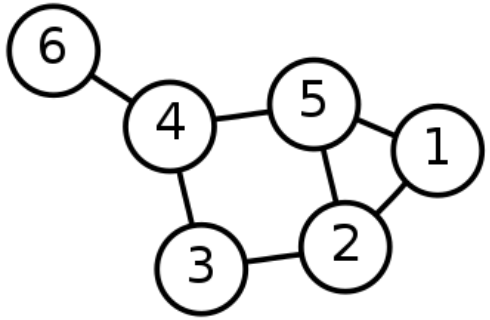
Mạng được định nghĩa và có cấu trúc như một đồ thị.

- **Directed**: đồ thị có hướng.
- **Undirected**: đồ thị vô hướng.
- **Bipartite**: đồ thị lưỡng phân hay đồ thị hai phần.
- **Multigraph**: đa đồ thị.
- **Temporal**: đối với mỗi đỉnh/cạnh có thời gian nó xuất hiện trong mạng.
- **Labeled**: mạng chứa các nhãn (trọng số, thuộc tính) trên các đỉnh hoặc các cạnh.

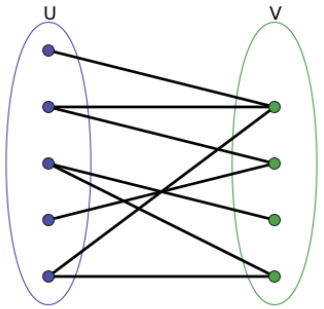
7. CÁC LOẠI MẠNG



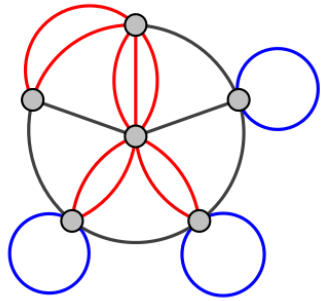
Đồ thị có hướng



Đồ thị vô hướng



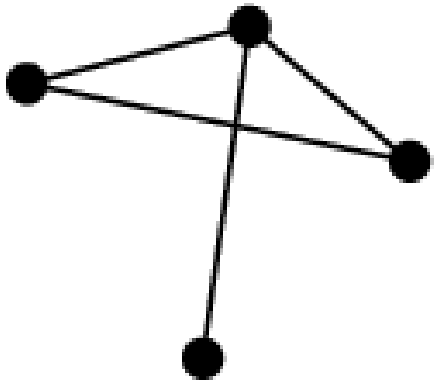
Đồ thị lưỡng phân



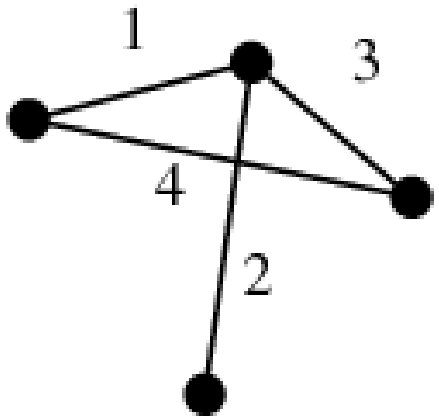
Đa đồ thị

7. CÁC LOẠI MẠNG

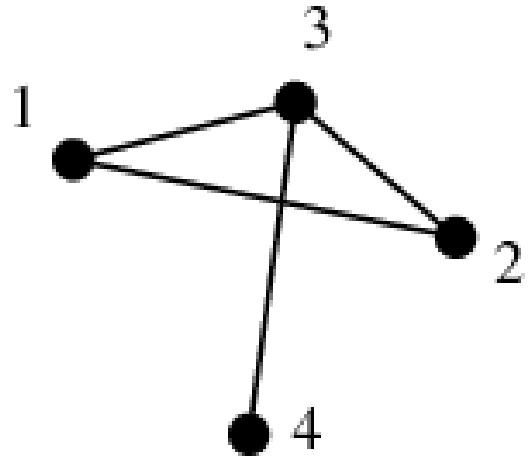
- **Nhãn của đồ thị**



unlabeled graph



edge-labeled graph



vertex-labeled graph


Q & A