ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

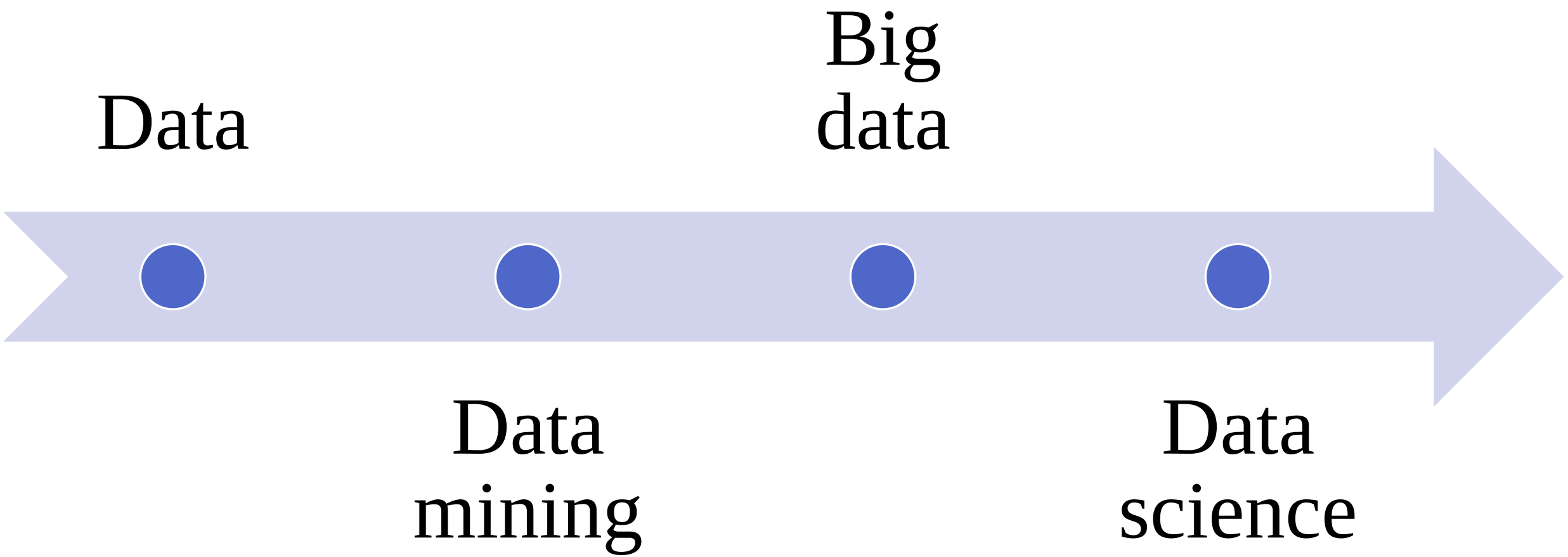
CHƯƠNG 1
Tổng quan về khai phá dữ liệu và ứng dụng doanh nghiệp

Biên soạn: ThS. Nguyễn Thị Anh Thư

Nội dung

1. Giới thiệu
2. Ứng dụng
3. Quy trình khám phá tri thức
4. Công nghệ và hệ sinh thái
5. Bài tập
6. Tổng kết

1. Giới thiệu



1. Giới thiệu

Dữ liệu là tập hợp các sự kiện, số liệu, hoặc thông tin được thu thập và lưu trữ dưới dạng có thể hiểu và sử dụng được.

Dữ liệu có thể được biểu diễn dưới nhiều dạng khác nhau, bao gồm:

- ***Số***: doanh số bán hàng, nhiệt độ, tuổi tác.
- ***Chữ***: tên, địa chỉ, mô tả sản phẩm.
- ***Hình ảnh***: ảnh chụp, ảnh X-quang, bản đồ.
- ***Âm thanh***: lời nói, tiếng nhạc, tiếng ồn môi trường.
- ***Video***: phim, chương trình truyền hình, camera giám sát.

1. Giới thiệu

Có hai loại dữ liệu chính:

Dữ liệu định lượng

- Dữ liệu có thể được đo lường và biểu diễn bằng số.
- Dữ liệu có thể được biểu diễn dưới dạng số nguyên, số thập phân, tỷ lệ hoặc phần trăm.

Dữ liệu định tính

- Dữ liệu không thể được đo lường bằng số và thường được biểu diễn bằng chữ hoặc hình ảnh.
- Ví dụ: Văn bản, hình ảnh, âm thanh, …

1. Giới thiệu

Data mining (Khai phá dữ liệu) là một lĩnh vực liên ngành thuộc khoa học máy tính, sử dụng các kỹ thuật thống kê, học máy và quản lý cơ sở dữ liệu để trích xuất kiến thức và thông tin hữu ích từ dữ liệu lớn.

Mục tiêu là chuyển đổi dữ liệu thô thành thông tin có giá trị có thể được sử dụng để:

- ***Đưa ra dự đoán***: Dự đoán xu hướng tương lai và đưa ra quyết định sáng suốt hơn.
- ***Khám phá tri thức***: Xác định các mẫu và mối quan hệ ẩn trong dữ liệu mà có thể không được nhìn thấy rõ ràng.
- ***Phân loại dữ liệu***: Nhóm các dữ liệu tương tự nhau vào các nhóm.
- ***Tối ưu hóa quy trình***: Xác định các yếu tố ảnh hưởng đến hiệu quả hoạt động và đề xuất các giải pháp cải tiến.

1. Giới thiệu

Dữ liệu lớn (big data) là tất cả thông tin, định lượng và theo dõi.

Tất cả thông tin	Big data lưu thông tin về mọi khía cạnh trong cuộc sống.
Định lượng	Tất cả thông tin được lưu dưới dạng kỹ thuật số.
Theo dõi	Thông tin được cập nhật liên tục theo thời gian.

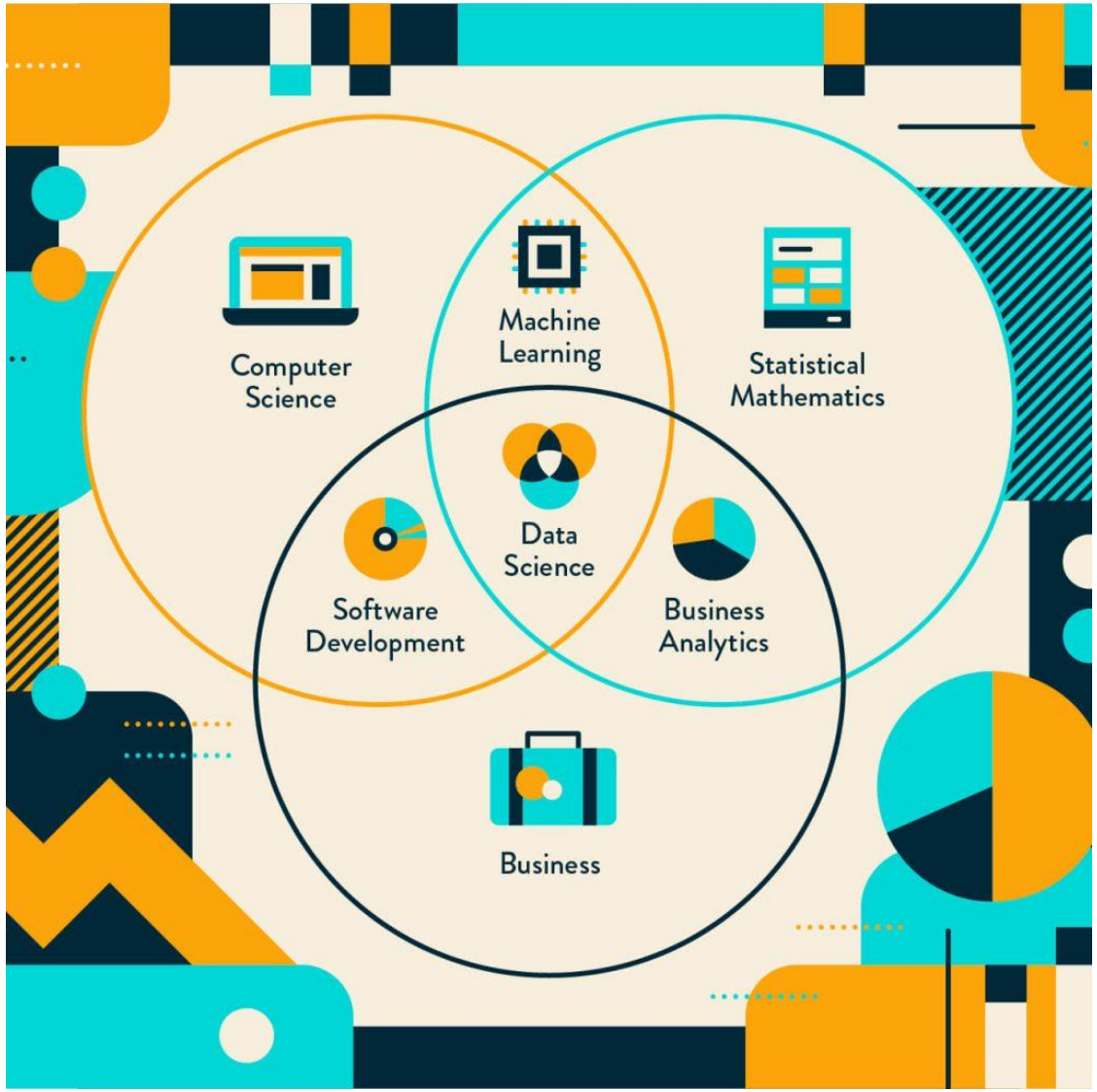
1. Giới thiệu

Đặc trưng cơ bản **4-V** của big data cụ thể như sau:

Volume (dung lượng)	• Kích thước dữ liệu lớn và khả năng mở rộng nhanh chóng.
Velocity (vận tốc)	• Tốc độ tạo dữ liệu lớn và liên tục.
Variety (đa dạng)	• Không đồng nhất và nhiều tập dữ liệu.
Veracity (xác thực)	• Chất lượng, tính trung thực và độ tin cậy của dữ liệu.

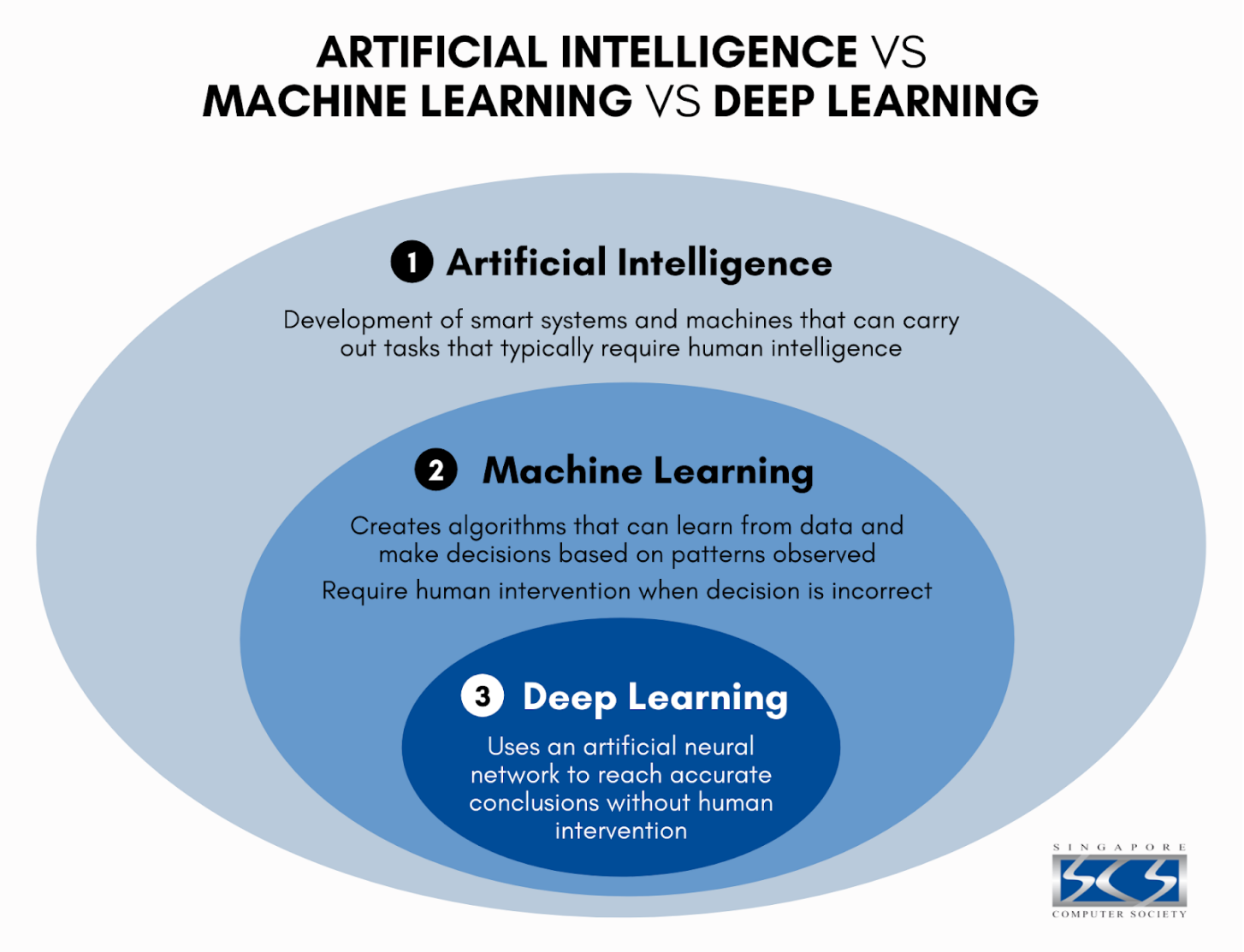
1. Giới thiệu

Data Science, hay Khoa học dữ liệu, là một lĩnh vực liên ngành kết hợp các kỹ thuật và công cụ từ toán học, thống kê, khoa học máy tính và quản lý dữ liệu để trích xuất kiến thức và thông tin hữu ích từ dữ liệu.



1. Giới thiệu

Mục tiêu của Data Science là chuyển đổi dữ liệu thô thành thông tin có giá trị.



1. Giới thiệu

Data Quality đề cập đến mức độ mà dữ liệu đáp ứng các yêu cầu về tính chính xác, tính đầy đủ, tính nhất quán và tính liên quan cho mục đích sử dụng cụ thể.

- Là một yếu tố quan trọng để đảm bảo rằng dữ liệu được sử dụng để đưa ra quyết định có giá trị.
- Là một quá trình liên tục và cần được chú trọng trong suốt vòng đời của dữ liệu.

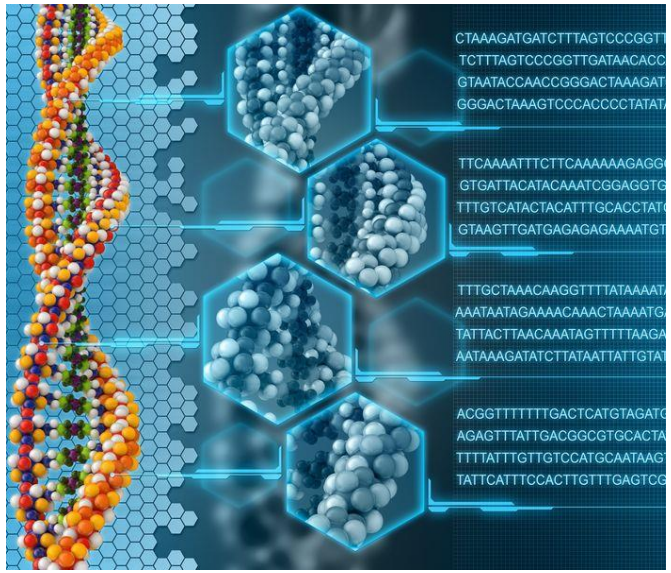
2. Ứng dụng

Phân tích dữ liệu tài chính (Financial Data Analysis)

Công nghiệp bán lẻ (Retail Industry)

Công nghiệp viễn thông (Telecommunication Industry)

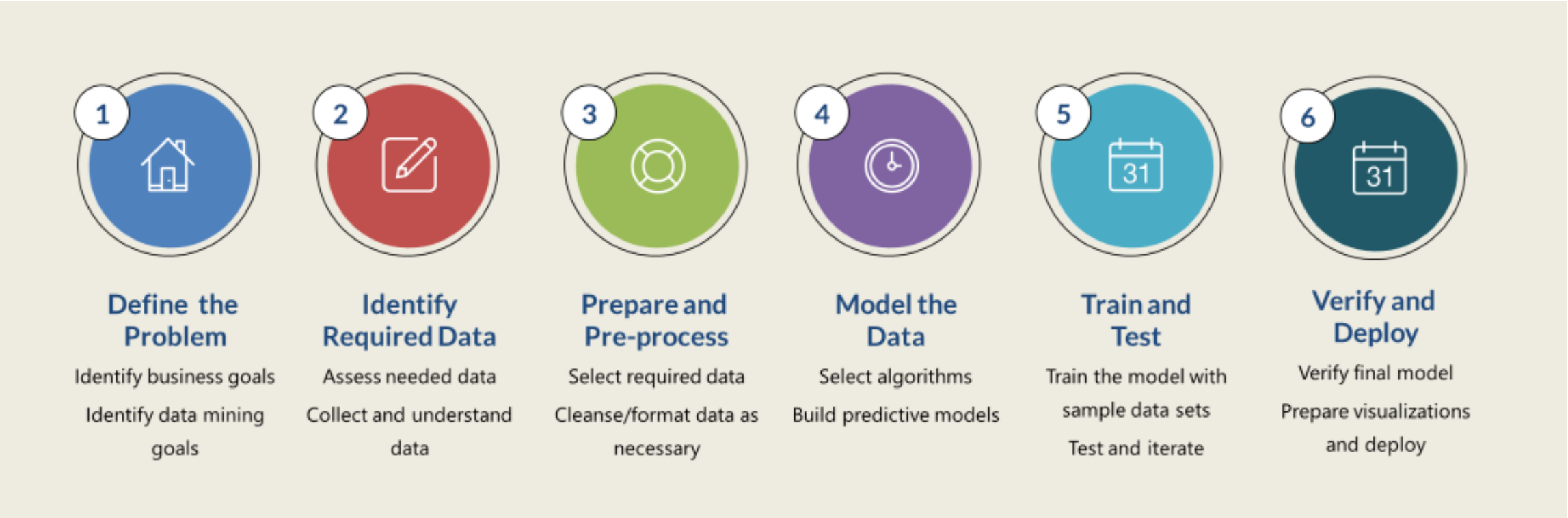
2. Ứng dụng



Phân tích dữ liệu sinh học (Biological Data Analysis)

Phát hiện xâm nhập (Intrusion Detection)

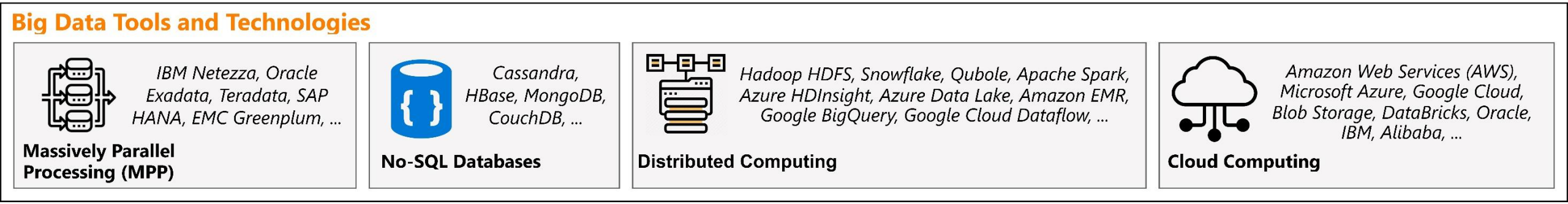
Một số ứng dụng trong khoa học (Scientific Applications)



3. Quy trình khám phá tri thức

4. Công nghệ và hệ sinh thái

- **Massively parallel processing** (MPP) là một kiến trúc máy tính sử dụng hàng trăm hoặc hàng nghìn bộ xử lý riêng biệt để thực hiện các phép tính song song trên dữ liệu khổng lồ.
- **NoSQL database** là một loại cơ sở dữ liệu lưu trữ dữ liệu theo cách khác với mô hình quan hệ truyền thống sử dụng bảng và mối quan hệ giữa các bảng.
- **Distributed Computing** là một mô hình kiến trúc nhiều máy tính riêng biệt được kết nối với nhau thông qua mạng để cùng nhau thực hiện một tác vụ chung.
- **Cloud Computing** là mô hình cung cấp dịch vụ máy tính như máy chủ, lưu trữ, mạng, và ứng dụng thông qua internet.

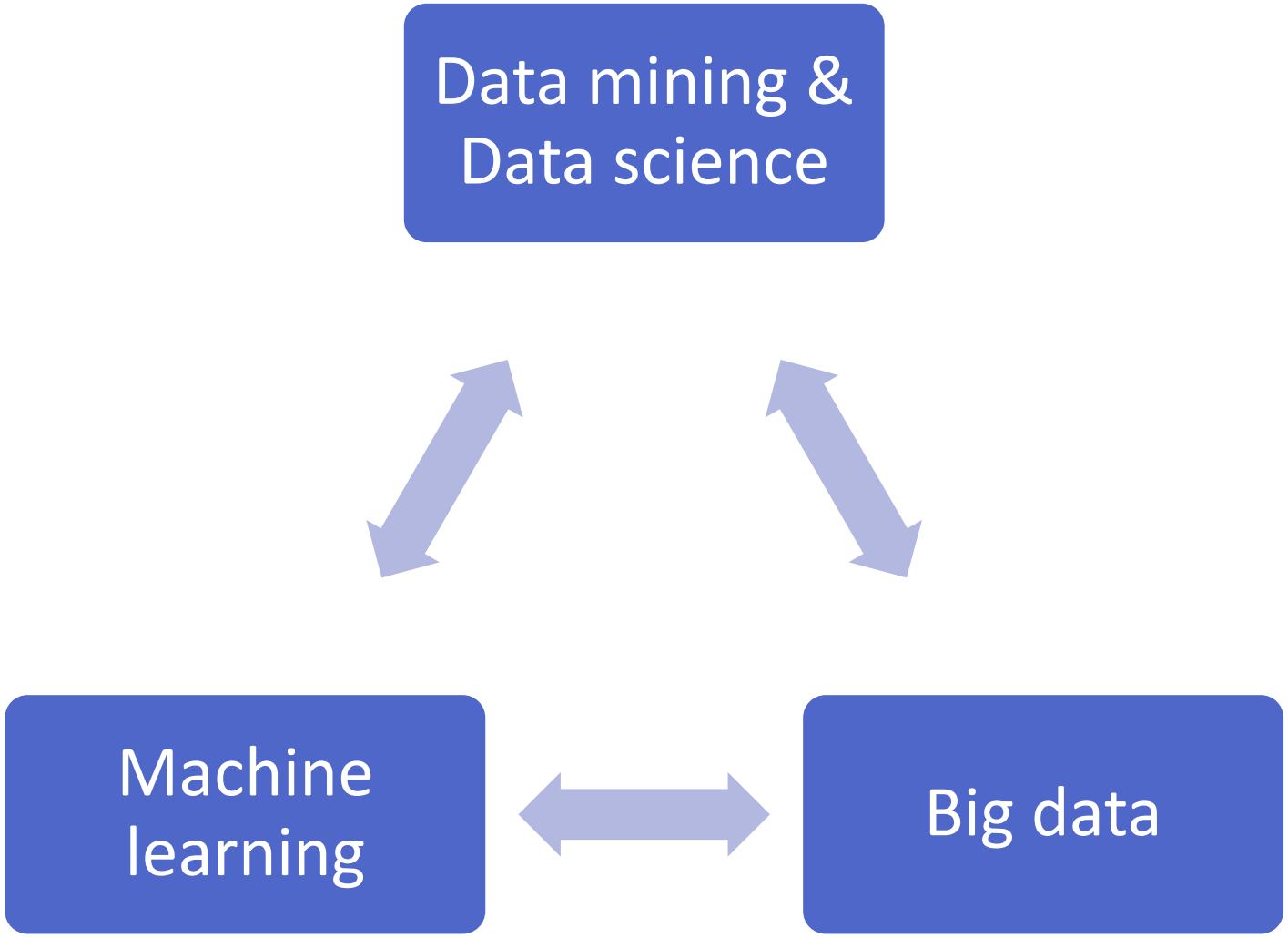


5. Bài tập

Link: https://tuoitre.vn/videographic-thong-tin-bi-lo-tu-facebook-giup-ong-trump-trong-bau-cu-nhu-the-nao-20180409151112154.htm

1. Bài toán khai thác dữ liệu nào được dùng trong vấn đề này?
2. Kỹ thuật IT và kiến thức của lĩnh vực liên quan nào được vận dụng để giải bài toán này?
3. Kết quả của bài toán này là gì?
4. Các chuyên gia đã vận dụng những kết quả này vào bầu cử tổng thống ra sao?

6. Tổng kết



Question & Answer