KHAI PHÁ DỮ LIỆU

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 1

GIỚI THIỆU CHUNG

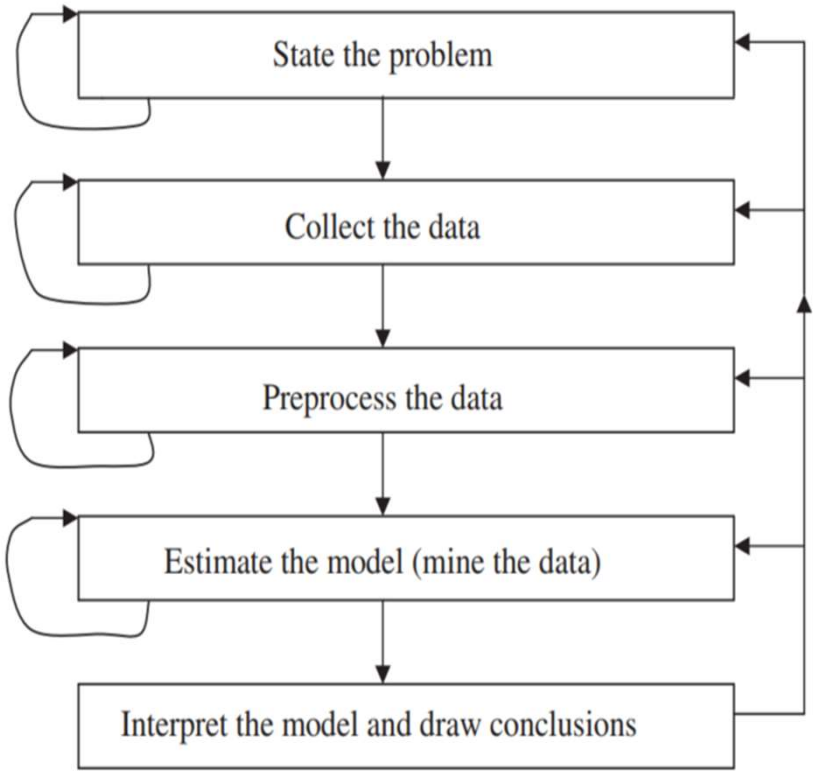
Giới thiệu chung

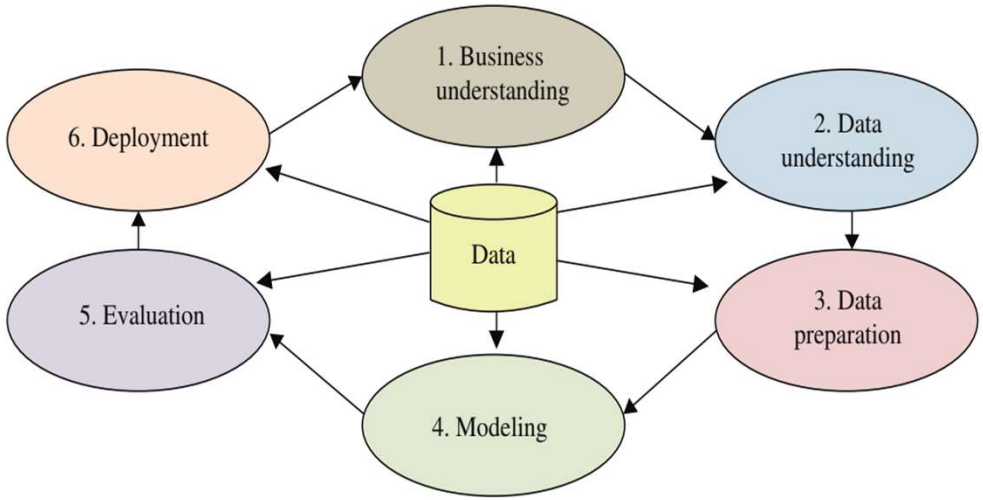
- Nguồn gốc và quá trình phát triển của KPDL
  - Xuất hiện vào những năm 80
  - KPDL phát triển nhanh chóng cùng với sự phát triển của
    - việc tích lũy dữ liệu → bùng nổ các kho dữ liệu lớn
    - sự phát triển của phần cứng máy tính

- KPDL là gì
  - Các kỹ thuật dùng để phân tích, tìm kiếm mối liên hệ giữa các dữ liệu trong một khối dữ liệu lớn nhằm tìm ra các tri thức, các mẫu dữ liệu tiềm ẩn trong đó.

Giới thiệu chung

- Các bước thực hiện KPDL:
  - State the problem and formulate the hypothesis
  - Collect the data
  - Preprocessing the data
  - Estimate the model
  - Interpret the model and draw conclusions





Giới thiệu chung

- Mối liên hệ giữa KPDL với Dữ liệu lớn và Khoa học dữ liệu
  - Dữ liệu lớn (Big data):
    - Cơ sở dữ liệu truyền thống: lưu trữ các dữ liệu xác định, có cấu trúc, ít thay đổi, số lượng dữ liệu không tăng lên quá nhanh
    - Dữ liệu lớn: tìm cách tập hợp (không lưu trữ tập trung) các dữ liệu không xác đinh, không có cấu trúc, có tốc độ thay đổi rất nhanh, số lượng bùng nổ rất nhanh và rất lớn (thông tin của các mạng xã hội, emails, ảnh online, video online, hồ sơ nhân sự, . . .) sao cho có thể áp dụng các kỹ thuật KPDL đối với các dữ liệu lớn đó.

Giới thiệu chung

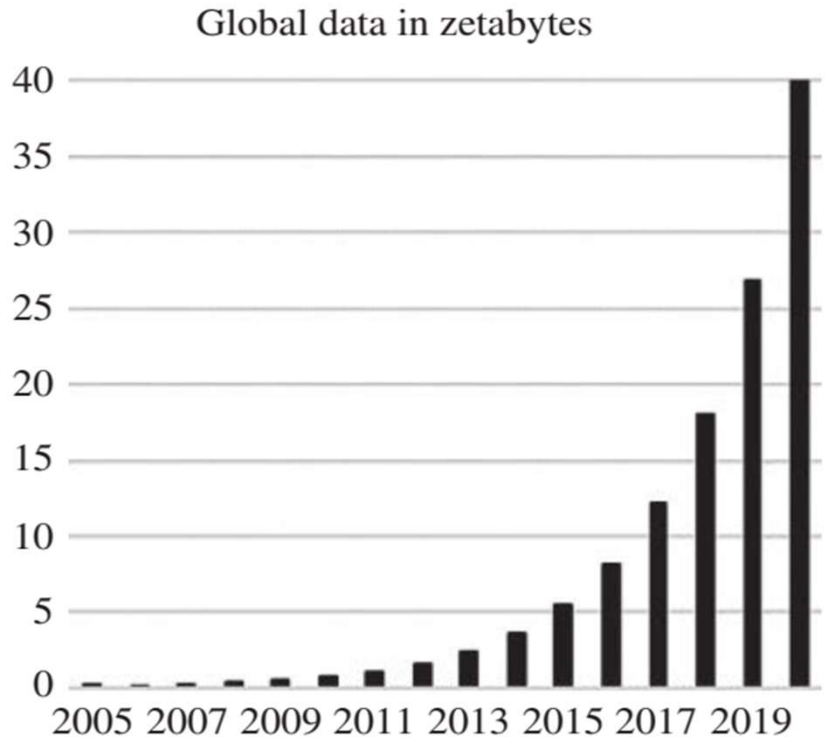
- Mối liên hệ giữa KPDL với Dữ liệu lớn và Khoa học dữ liệu

Company	Big Data
YouTube	Users upload 100 hours of new videos per minute
Facebook	More than 1.4 billion users communicating in 70+ languages
Twitter	175 million tweets per day
Google	2 million search queries/minute → processing 35 petabytes daily
Apple	47,000 applications are downloaded per minute
Instagram	Users share 40 million photos per day
LinkedIn	2.1 million groups have been created
Foursquare	571 new Web sites are launched each minute

Số liệu minh họa sự bùng nổ dữ liệu

Giới thiệu chung

- Mối liên hệ giữa KPDL với Dữ liệu lớn và Khoa học dữ liệu



Tốc độ bùng nổ dữ liệu

Giới thiệu chung

- Mối liên hệ giữa KPDL với Dữ liệu lớn và Khoa học dữ liệu
  - Khoa học dữ liệu:
    - Là ngành khoa học chú trọng đến quản trị và phân tích dữ liệu để tìm ra các tri thức → các tri thức đó hỗ trợ cho việc tạo ra quyết định, lập ra kế hoạch hành động.

Giới thiệu chung

- Các ứng dụng của KPDL
  - Phân tích thị trường
  - Phát hiện gian lận
  - Quản lý rủi ro
  - Phân tích hành vi
  - Hỗ trợ ra quyết định
  - . . .
- Một số công cụ giúp KPDL:
  - RapidMiner, Weka, Knime, Apache Mahout, Oracle DataMining, TeraData, Orange, . . .

Giới thiệu chung

Q / A