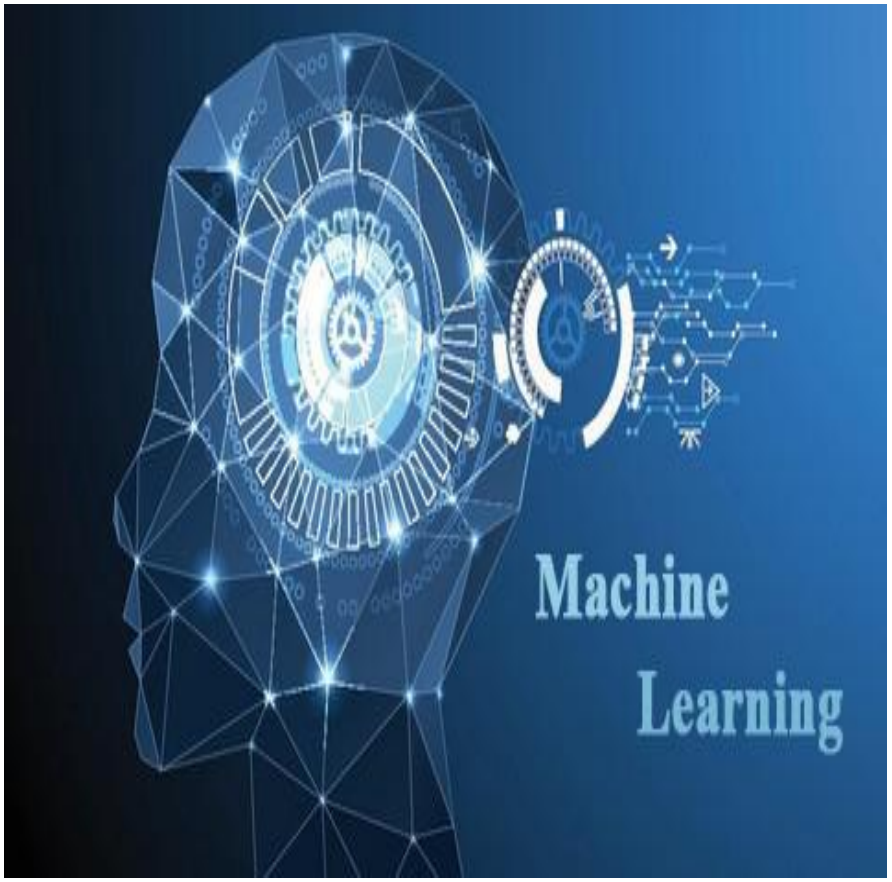
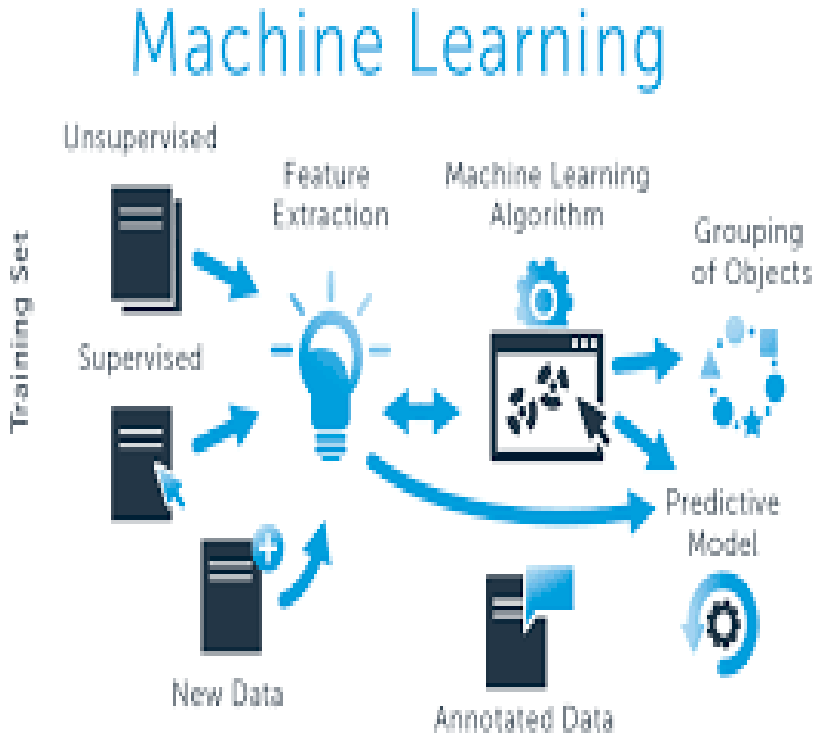
CSE 445: HỌC MÁY





TỔNG QUAN MÔN HỌC

- Tên môn học: Học máy (Machine Learning)
- Mã số môn học: CSE445
- Thời lượng: 3 Tín chỉ (30LT+15TH)
- Điểm quá trình (ĐQT): điểm chuyên cần (20%) + điểm bài tập (40%) + điểm thi giữa kỳ (40%).
- Thi hết môn: Tự luận (thời gian 60').
- Điểm môn học = ĐQT x 50% + THM x 50%

BÀI TẬP

- Giảng viên giao 4 bài tập.
- Điểm bài tập là trung bình cộng các bài tập mà SV nộp đúng hạn qua Piazza.
- Tất cả các trường hợp gian lận đều nhận điểm 0

Mục đích của môn học

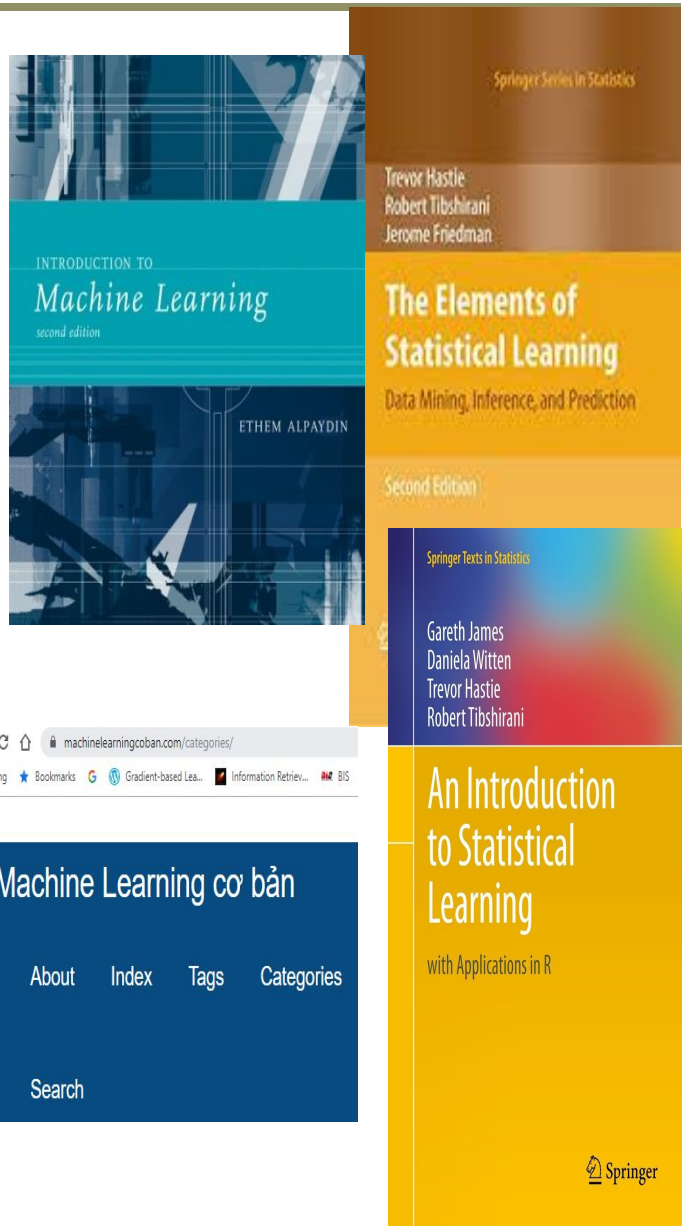
- Trang bị tổng quan ở mức cao về các kỹ thuật Học máy cơ bản.
- Biết vận dụng các phương pháp học máy dùng cho phân tích dữ liệu, khai phá dữ liệu và hỗ trợ ra quyết định.
- Kỹ năng thực hành, thiết kế thí nghiệm sử dụng ngôn ngữ Python.
- Làm quen với các thuật ngữ chuyên ngành.

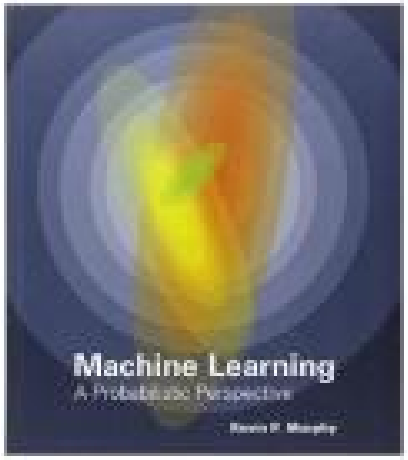
Đối tượng tham dự

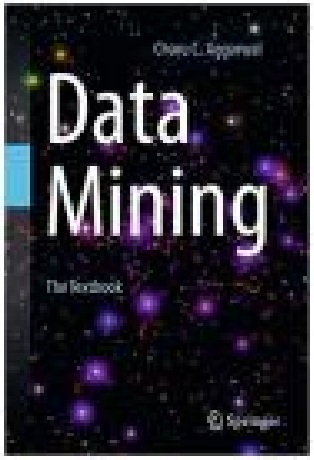
- Các ngành học liên quan đến CNTT, kinh tế, điện tử
- Không cần kiến thức nền về Học máy
- Điều kiện
  - Hoàn thành các môn học về xác suất thống kê, đại số tuyến tính.
  - Có kỹ năng lập trình cơ bản (R/Matlab/Python)

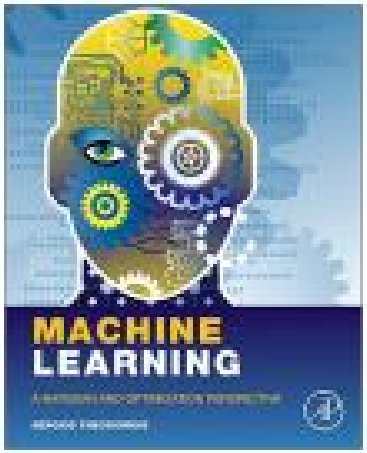
Tài liệu môn học

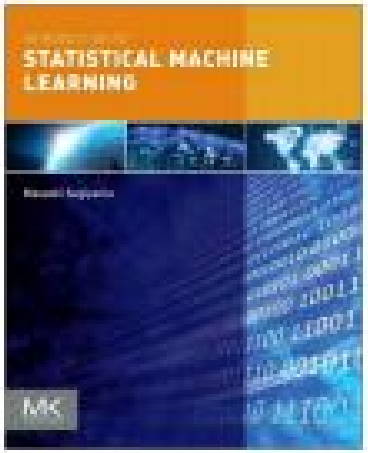
1. Trevor Hastie Robert Tibshirani, Jerome Friedman (2001), "*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*", Springer (https://web.stanford.edu/~hastie/Papers/ESLII.pdf)
2. Gareth James, Daniela Witten, Trevor Hastie Robert Tibshirani (2013), *"An Introduction to Statistical Learning with Application R", Springer.*
3. Ethem Alpaydin (2010), "*Introduction to Machine Learning",* The MIT Press (https://www.cmpe.boun.edu.tr/~ethem/i2ml2e/)
4. Mitchell, T. M. (1997). Machine learning. McGraw Hill (http://profsite.um.ac.ir/~monsefi/machine-learning/pdf/Machine-Learning-Tom-Mitchell.pdf)
5. Mitchell, T. M. (2006), "*The discipline of Machine learning"*, Carnegie Mellon University, School of Computer Science, Machine Learning Department.
6. Blog "Machine learning cơ bản" (https://machinelearningcoban.com)

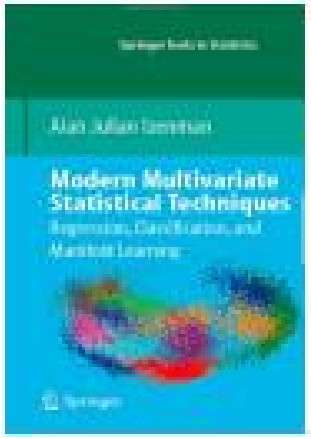


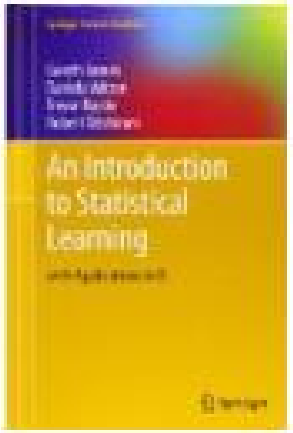

Machine Learning
A Probabilistic Perspective

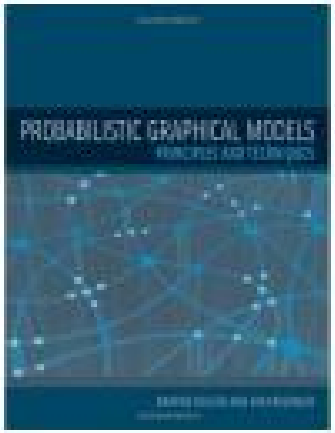

Data Mining

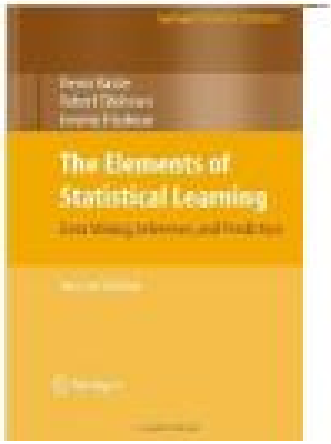

MACHINE LEARNING

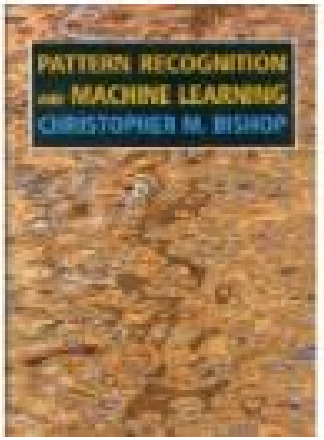

STATISTICAL MACHINE LEARNING


Modern Multivariate Statistical Techniques


An Introduction to Statistical Learning


PROBABILISTIC GRAPHICAL MODELS


The Elements of Statistical Learning


PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP

CSE 445 Hỏi & Đáp

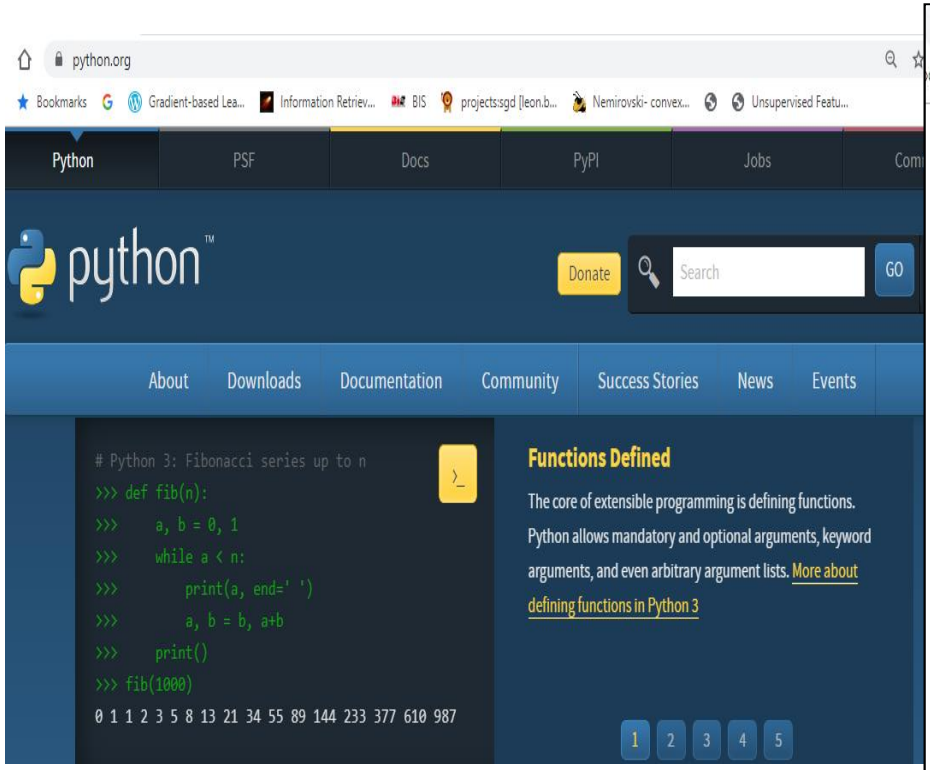
- CSE 445 sử dụng Piazza!
- Đặt các câu hỏi liên quan đến nội dung môn học, bài tập, v.v. trên Piazza
- Link piazza:

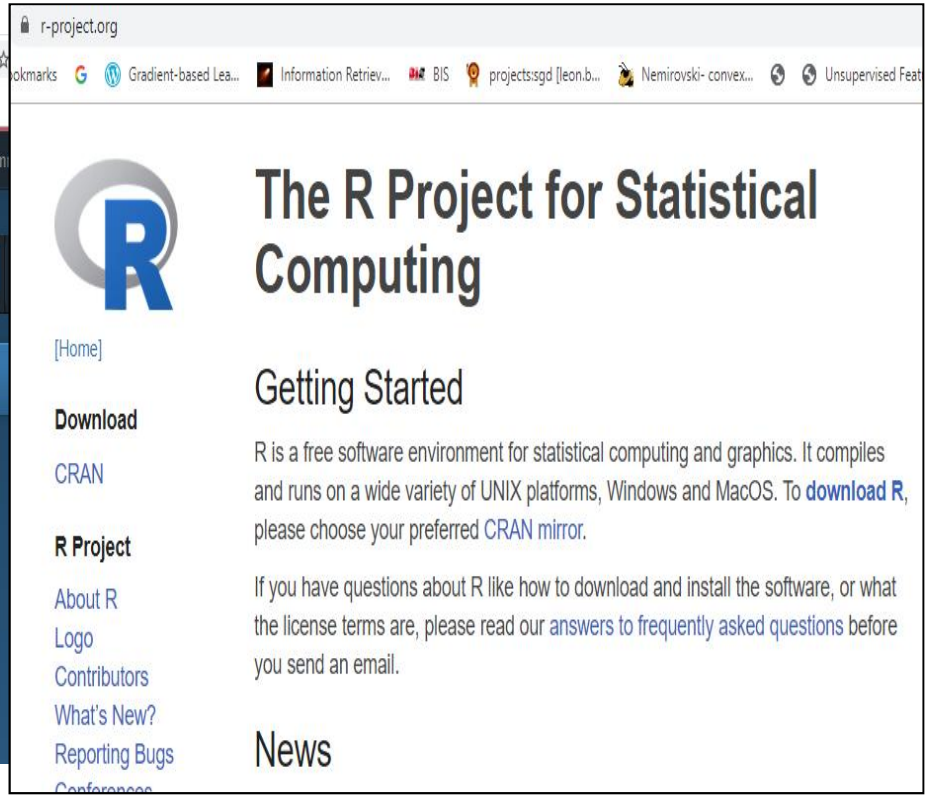
https://piazza.com/tlu.edu.vn/winter2019/cse445fall2019

piazza

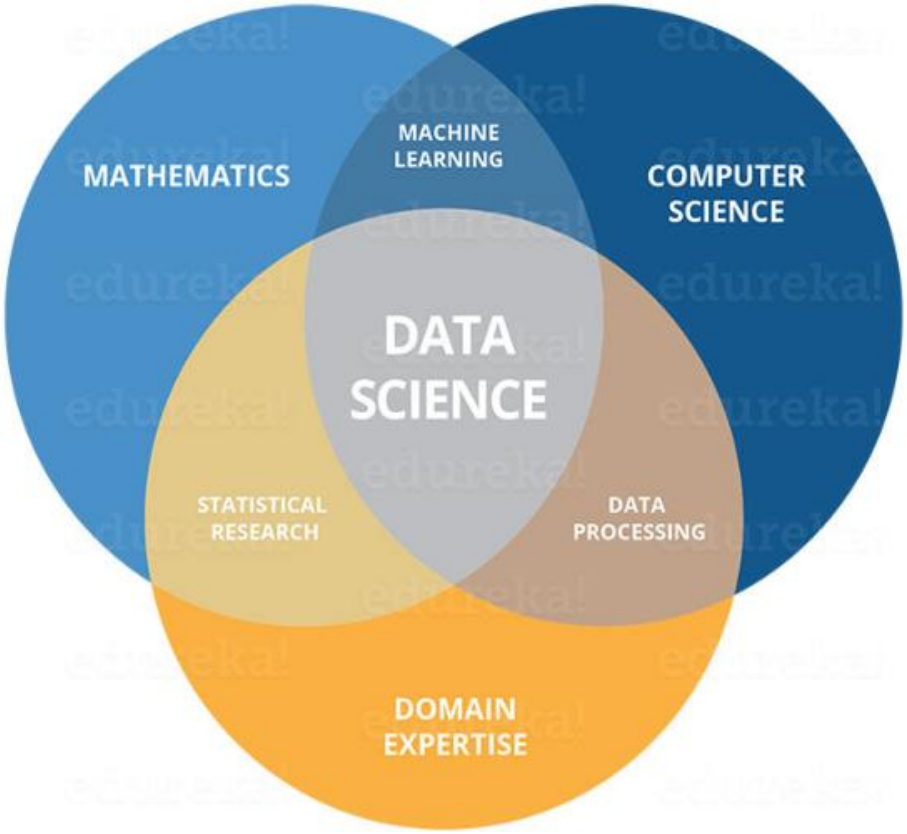
Ngôn ngữ lập trình

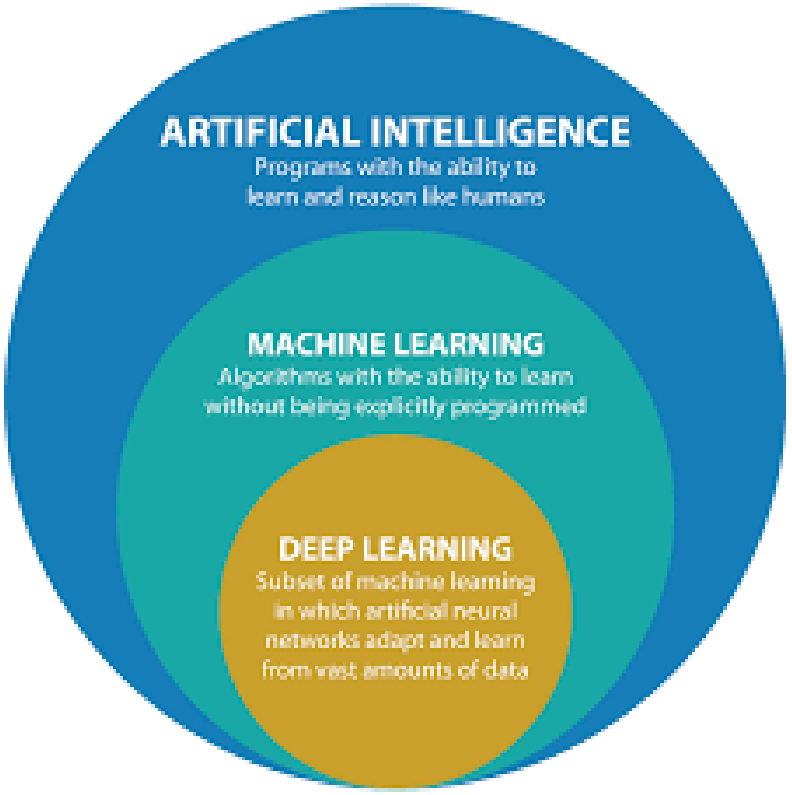
- R: https://www.r-project.org/
- Python: https://www.python.org/





Giới thiệu về Học máy





Giới thiệu về Học máy

Machine learning là gì?

- Arthur Samuel (1959): Máy học là ngành học cung cấp cho máy tính khả năng học hỏi mà không cần được lập trình một cách rõ ràng
- Giáo sư Tom Mitchell (Carnegie Mellon University): Machine Learning là 1 chương trình máy tính được nói là học hỏi từ kinh nghiệm E từ các tác vụ T và với độ đo hiệu suất P. Nếu hiệu suất của nó áp dụng trên tác vụ T và được đo lường bởi độ đo P tăng từ kinh nghiệm E.

Học máy

- Bao gồm quá trình đúc rút tri thức từ các quan sát, trải nghiệm thực tiễn bằng việc xây dựng các mô hình từ dữ liệu.
- Các phương pháp học và nhận dạng tự động các mẫu phức tạp (complex patterns) từ dữ liệu.

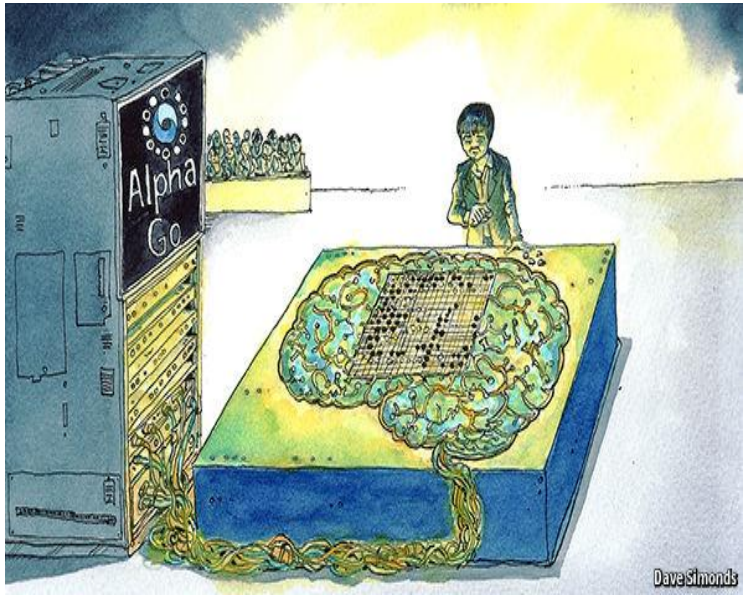
Giới thiệu về Học máy

- *"Lĩnh vực nghiên cứu giúp máy tính có khả năng tự học khi không được lập trình trước"*

– Arthur Samuel (1959)

Các ứng dụng của Học máy

AlphaGo thắng nhà vô địch thế giới cờ vây



Các ứng dụng của Học máy

Trong hệ thống tự động ra quyết định

- Lọc thư rác

- Phát hiện gian lận

"How Credit Card Companies Spot Fraud Before You Do"

U.S. News (July 10, 2013)

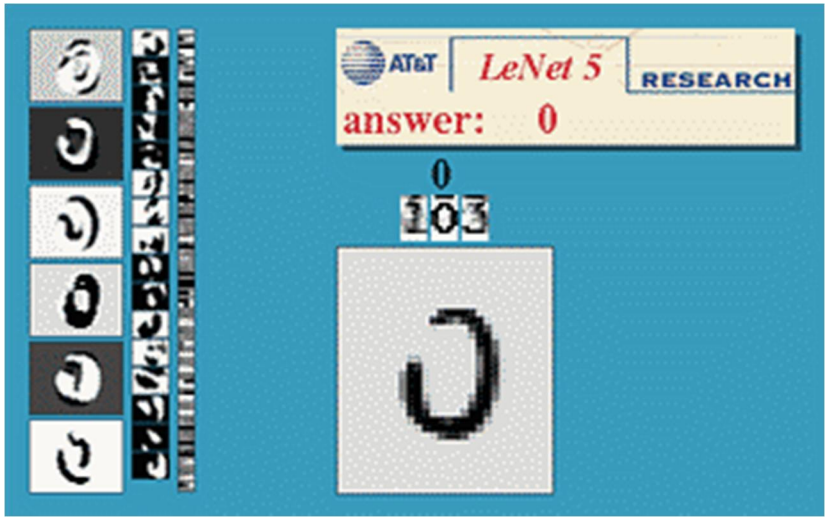
Các ứng dụng của Học máy

Trong các hệ thống tự động có lập trình phức tạp

- Xe không người lái

Stanford Autonomous Driving Team
http://driving.stanford.edu/

- Nhận dạng chữ viết tay



LeNet--5 Convolutional Neural Net

Các ứng dụng của Học máy

Dùng cho khai phá dữ liệu

- Bệnh án điện tử

"Mining Electronic Records for Revealing Health Data"

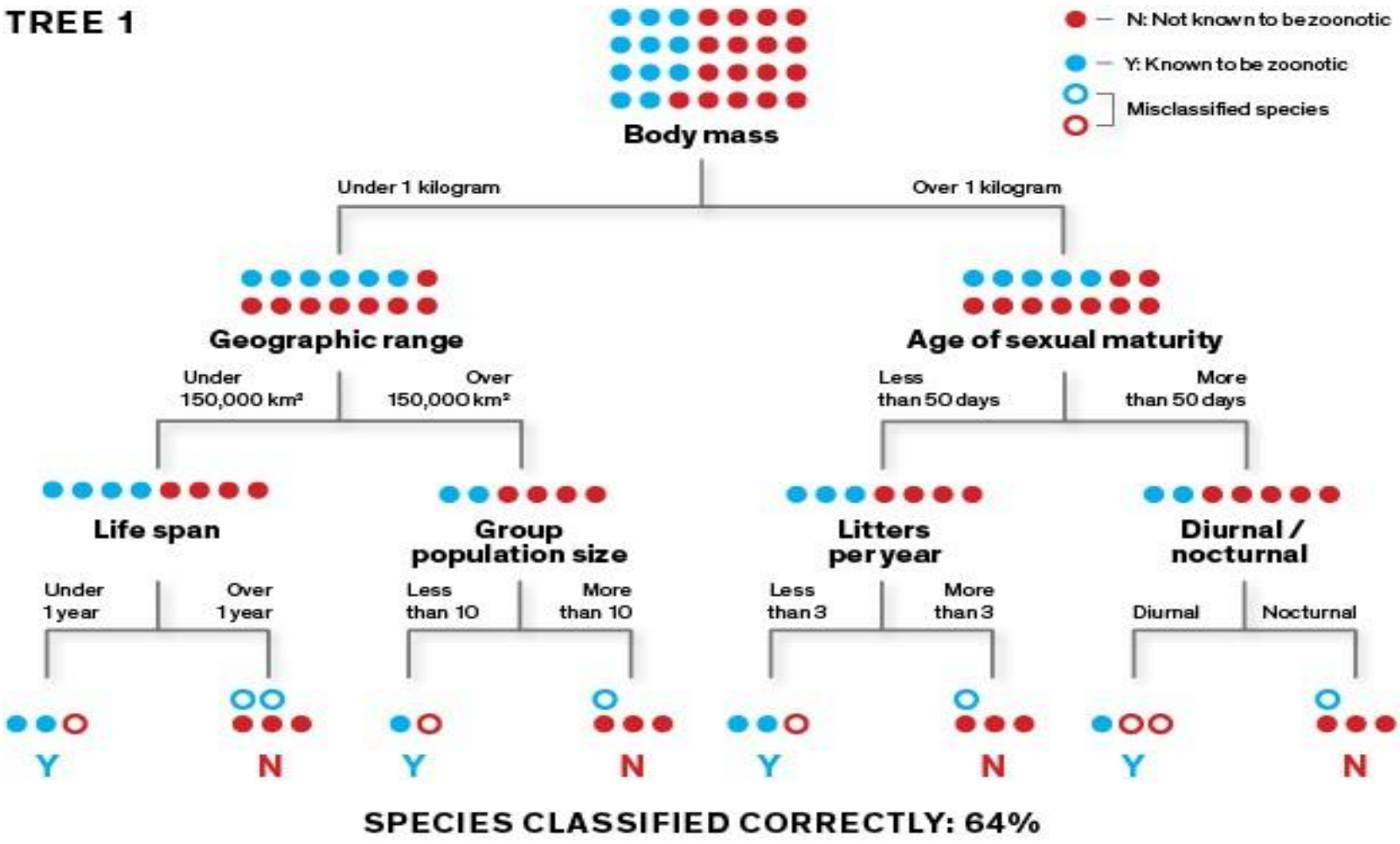
New York Times (Jan 14, 2013)

Trong các hệ thống tùy biến

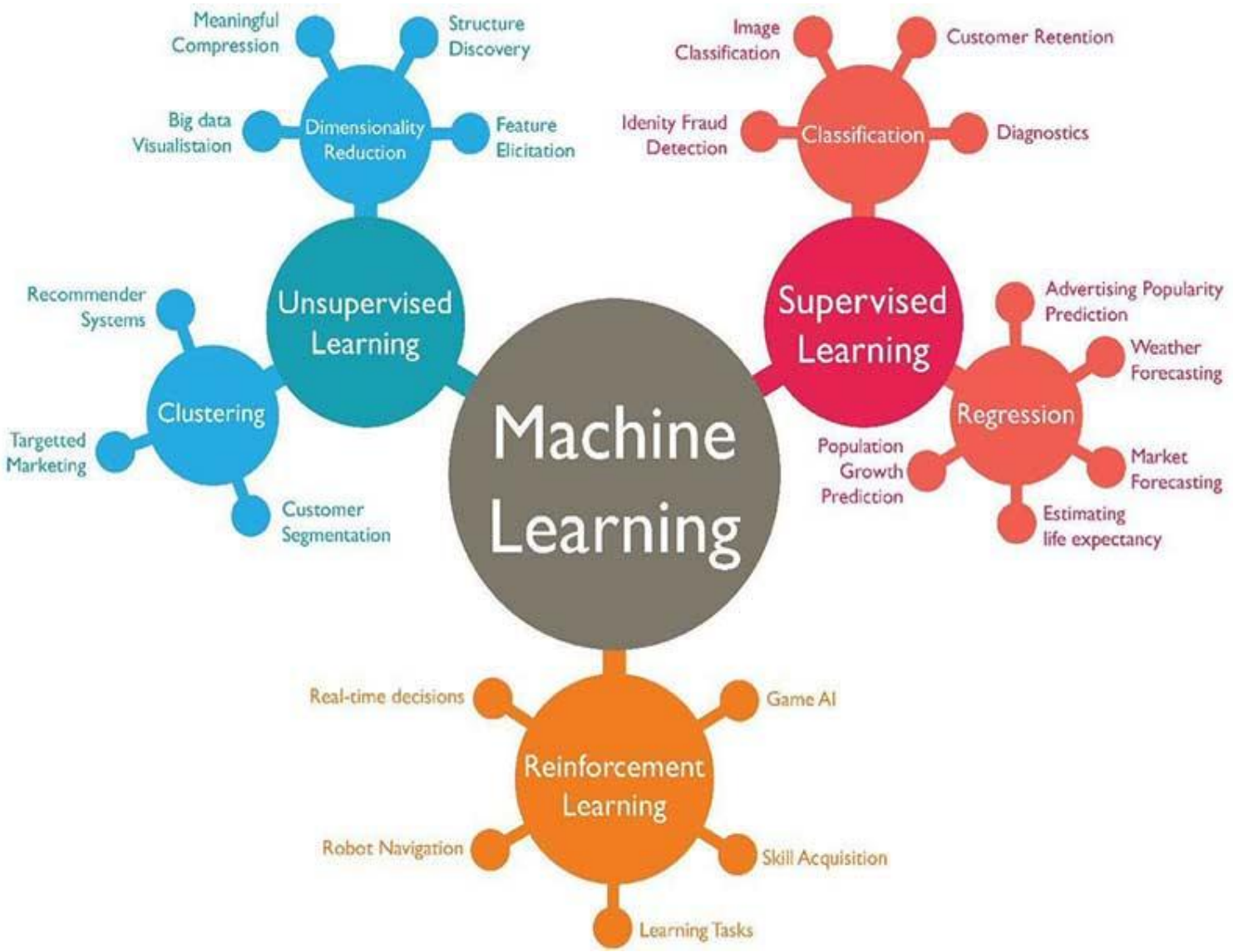
- Hệ thống gợi ý sản phẩm

Các ứng dụng của Học máy

[The Algorithm That's Hunting Ebola](#) (IEEE Spectrum, Sept 24 2015)



Giới thiệu về Học máy



Các giải thuật Học máy

- Để lọc thư rác hoặc nhận dạng chữ viết tay, chúng ta gắn nhãn các mẫu (quan sát) để học mô hình từ chúng

  → *Học máy có giám sát*: Huấn luyện cho giải thuật học máy xây dựng mô hình từ các mối quan hệ trong dữ liệu, dựa trên tập các cặp đầu vào-ra của các quan sát.

- Để phát hiện các nhóm bệnh nhân trong Bệnh án điện tử (EMR), chúng ta chưa biết tên các nhóm (các lớp)

  → *Học máy không giám sát*: Huấn luyện cho giải thuật học các mối quan hệ và cấu trúc của dữ liệu.

Các giải thuật Học máy

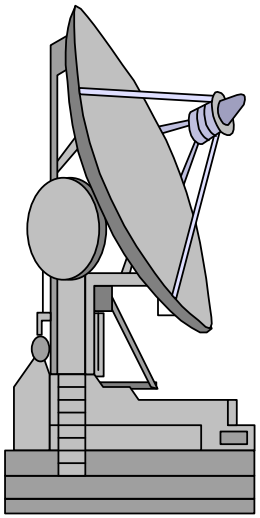
Một số giải thuật học máy khác:

- ✓ Học bán giám sát (semi-supervised learning)
- ✓ Học tăng cường (reinforcement learning)
- ✓ Hệ thống khuyến nghị (recommender systems)

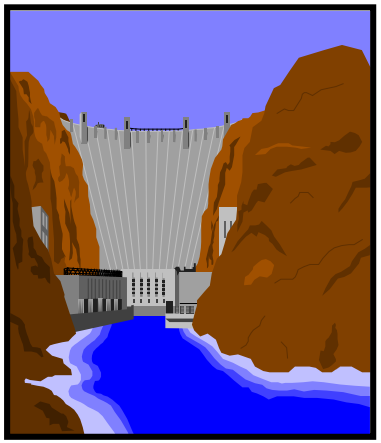
....

Mô hình học máy (machine learning model)

Truyền thông



Ra quyết định



Kỹ thuật

Tại sao phải xây dựng mô hình?

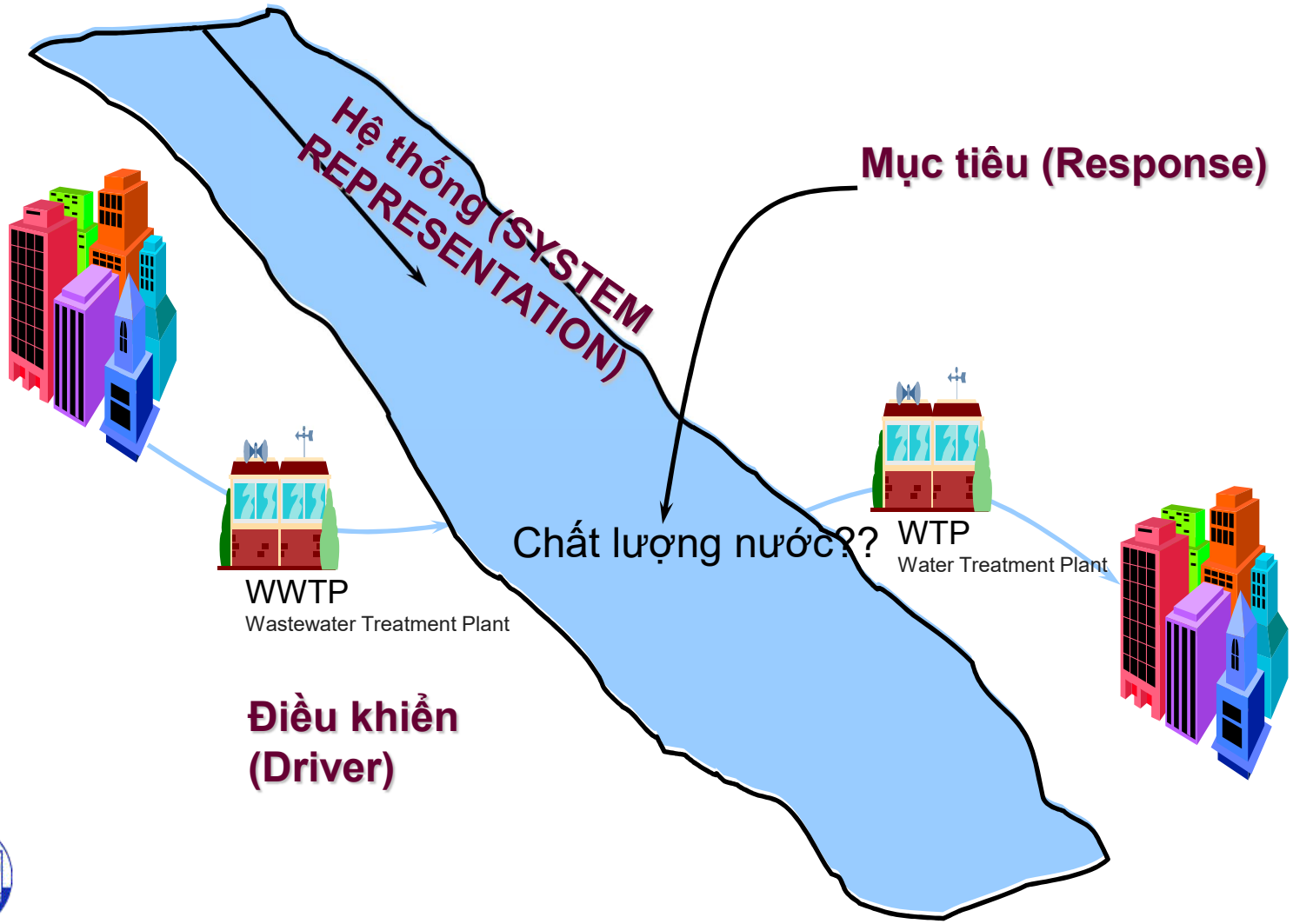
- Mô hình thể hiện xấp xỉ của thực tế được sử dụng để giải quyết các vấn đề cụ thể
- Chúng thường được xây dựng trên máy tính
- Chúng được sử dụng rộng rãi trong thực hành kỹ thuật

Tại sao dùng kỹ thuật thống kê?

- Nhiều biến trong kỹ thuật chứa thông tin không chắc chắn
- Xác suất và thống kê là các công cụ để xử lý những biến không chắc chắn
- Thường được sử dụng rộng rãi trong kỹ thuật

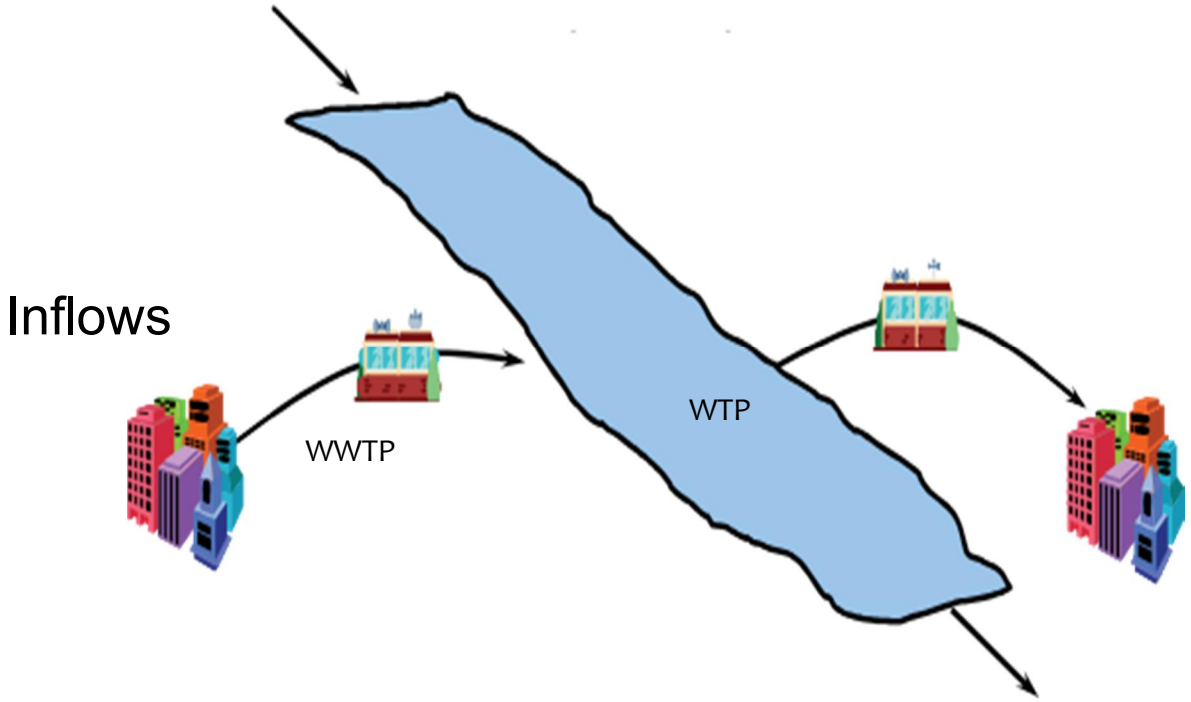
Các thành phần của mô hình

- **Hệ thống:** Nhóm các thành phần mà chúng tương tác hoặc vận hành cùng nhau



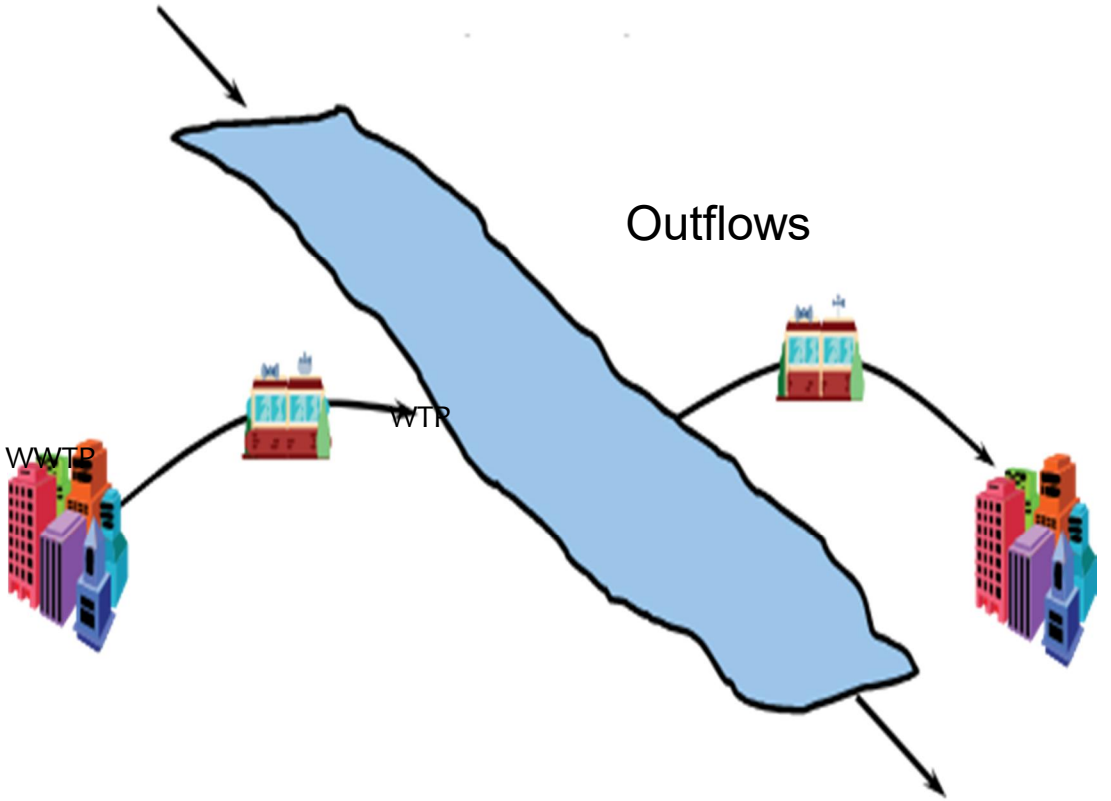
Các thành phần của mô hình

- **Biến đầu vào:** Biến giúp xác định trạng thái của hệ thống thay đổi như thế nào (“Driver”)



Các thành phần của mô hình

- **Biến đích:** Biến đầu ra có quan hệ với trạng thái của hệ thống



Đặt bài toán và thuật ngữ

- Tập biến đầu vào X (tập biến dự đoán, biến độc lập hoặc các đặc trưng) (input variables, predictors, independent variables or features)

- Biến đầu ra Y (biến đích hoặc biến phụ thuộc) (output variables, response or dependent variable)

- Học máy thống kê (Statistical Learning): là một tập các giải pháp ước lượng hàm f để mô tả mối quan hệ giữa tập biến đầu vào và biến đầu ra

$$Y = f(X) + \epsilon$$

Trong đó ϵ là phần lỗi tuân theo phân phối chuẩn và có giá trị kì vọng bằng 0.

Đặt bài toán và Thuật ngữ

- Làm cách nào để xây dựng mô hình?
- *Dữ liệu huấn luyện (Training data)*: tập gồm *n* các quan sát /mẫu huấn luyện (observations, samples) ta dùng để xây dựng mô hình
- Các cặp dữ liệu vào/ra:

$$(\mathbf{X}^{(1)}, \mathbf{Y}^{(1)}), \ldots, (\mathbf{X}^{(n)}, \mathbf{Y}^{(n)})$$

Đặt bài toán và Thuật ngữ

$$Y = f(X) + \epsilon$$

- Phương pháp để ước lượng sẽ phụ thuộc vào vấn đề mà chúng ta muốn xử lý khi sử dụng dữ liệu.

→ Các phương pháp học máy khác nhau sẽ dùng các mô hình khác nhau để ước lượng hàm.

Dự đoán và Suy diễn

- *Dự đoán (Prediction)*: Dự đoán biến đích Y với tập dữ liệu đầu vào X cho trước, sử dụng một hàm ước lượng thống kê của f, ký hiệu mô hình này là $\hat{f}$.
- *Suy diễn (Inference)*: Tìm hiểu mối quan hệ giữa Y với các biến độc lập X_i. Áp dụng khi không mong muốn xây dựng một mô hình hộp đen (black-box model).

Ví dụ về Quảng cáo

- Doanh nghiệp có thể điều chỉnh chiến lược quảng cáo sản phẩm (advertising) để tăng doanh số bán hàng (sales).
- Dữ liệu: Doanh số bán hàng và ngân sách quảng cáo cho 3 phương tiện truyền thông (TV, radio, newspaper).

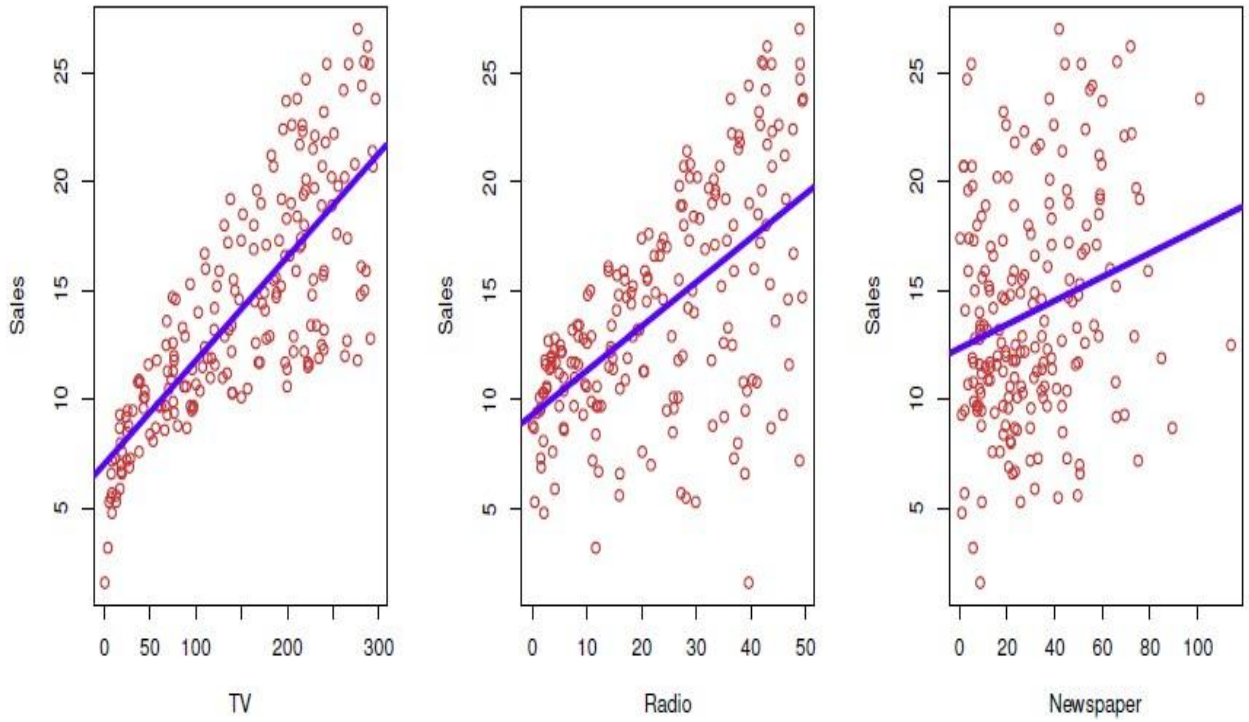


Figure 2.1, ISL 2013

Câu hỏi?

- Trong ví dụ về quảng cáo, đâu là biến đầu vào/đầu ra?
  - *Biến đầu vào*: ngân sách quảng cáo trên TV, ngân sách quảng cáo trên Radio, ngân sách quảng cáo trên báo chí
  - *Biến đầu ra:* doanh số bán hàng

Câu hỏi?

Hãy lấy ví dụ về yêu cầu dự đoán và suy diễn mà ta có được lời giải từ dữ liệu này?

- Dự đoán:
  - Số liệu về doanh số bán hàng ở thị trường A dự kiến thế nào khi biết ngân sách đầu tư quảng cáo trên TV, radio và báo chí?
- Suy diễn:
  - Doanh số bán hàng tăng bao nhiêu nếu tăng ngân sách 10% cho quảng cáo trên TV?
  - Phương tiện truyền thông nào (TV, radio, báo) tạo ra sự thúc đẩy lớn nhất trong bán hàng?

Làm thế nào để ước lượng f ?

- Giả sử ta có tập dữ liệu huấn luyện:

$$\{(\mathbf{X_1}, \mathbf{Y_1}), (\mathbf{X_2}, \mathbf{Y_2}), \dots, (\mathbf{X_n}, \mathbf{Y_n})\}$$

- Ta phải dùng tập dữ liệu và một phương pháp học máy để ước lượng hàm f.
- Các phương pháp (mô hình) học máy
  - Các phương pháp có tham số
  - Các phương pháp phi tham số

Mô hình tham số và phi tham số

➢ Mô hình có tham số (Parametric)

- Đặt các giả định cho dạng (form) của f.
- Sử dụng dữ liệu huấn luyện để xấp xỉ/khớp (fit) mô hình (ước lượng các tham số)
- *Ưu điểm*: Dễ tìm các tham số của f.
- *Nhược điểm*: Mô hình có thể ước lượng thiếu chính xác dạng của f.

Mô hình tham số và phi tham số

- Mô hình phi tham số
  - Không cần đặt các giả định về dạng thức (form) của f.
  - Xấp xỉ f với lỗi nhỏ nhất để không bị *quá khớp/quá phù hợp* (*overfitting)* trên dữ liệu huấn luyện/tập học.
  - *Ưu điểm*: Có thể xấp xỉ các mô hình cho f.
  - *Nhược điểm*:
    - ✓ Yêu cầu lượng lớn dữ liệu huấn luyện
    - ✓ Vấn đề ***overfitting*** *(quá khớp)*: đạt độ chính xác cao trên tập học, nhưng đạt độ chính xác thấp trên tập thử nghiệm

Trade-off: Độ chính xác và Tính diễn giải

- Các phương pháp khác nhau mang lại sự linh hoạt
  - Những mô hình có nhiều hạn chế sẽ cho độ chính xác kém
  - Ví dụ: Hồi quy tuyến tính bị hạn chế – không xấp xỉ được hàm phi tuyến.

Trade-off: Độ chính xác vs. Tính diễn giải

- Tại sao chọn mô hình có nhiều hạn chế?
  - Dễ diễn giải – thuận lợi cho bài toán suy diễn
  - Các mô hình đơn giản có thể cho kết quả với độ chính xác cao (ít gặp vấn đề overfitting)
- Với bài toán dự đoán, tính diễn giải không quá cần thiết: Mô hình dự đoán có thể là một hộp đen

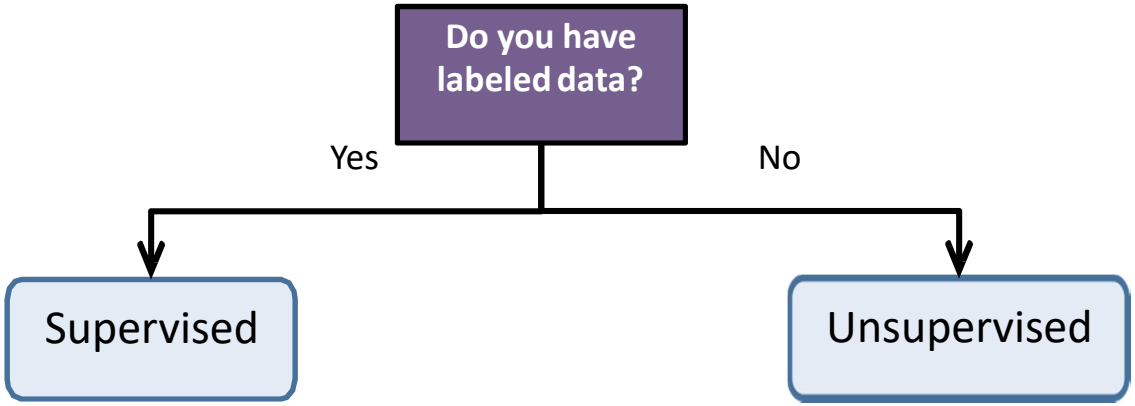
Trade-off: Độ chính xác vs. Tính diễn giải

Figure 2.7, ISL 2013

HỌC MÁY

Bài toán học máy được chia làm 3 dạng chính:

- ***Học có giám sát (Supervised Learning)***
- ***Học không giám sát (Unsupervised Learning)***
- *Học tăng cường (Reinforcement learning)*



Học có giám sát

- Cả biến đầu vào và biến đầu ra đều lưu trữ trong tập học.
  - $X^{(i)}$ và $Y^{(i)}$ đều có sẵn trong tập học
- Mục tiêu: Khái quát hóa (*generalize)* dữ liệu thử nghiệm
- Bài toán:
  - Phân lớp
  - Hồi quy

Học không giám sát

- Chỉ có các biến đầu vào, không có biến đầu ra
  - $X^{(i)}$ có sẵn, tuy nhiên $Y^{(i)}$ không có
- Mục tiêu: *Phát hiện mối quan hệ* giữa các biến hoặc giữa các quan sát (observations)
- Bài toán:
  - Phân cụm
  - Giảm chiều dữ liệu

Học có giám sát: Hồi quy

- *Hồi quy*: biến đầu ra Y là định lượng (liên tục/dạng số/có thứ tự) (continuous /numerical /ordered)

Ví dụ: Dự đoán

- ✓ Giá cổ phiếu trong 1 năm tính từ thời điểm này?
- ✓ Thu nhập của một người dựa trên yếu tố nhân khẩu học?

Học có giám sát: Phân lớp

- *Phân lớp*: biến đầu ra Y dạng định tính (kiểu rời rạc /thứ bậc/định danh) (categorical)

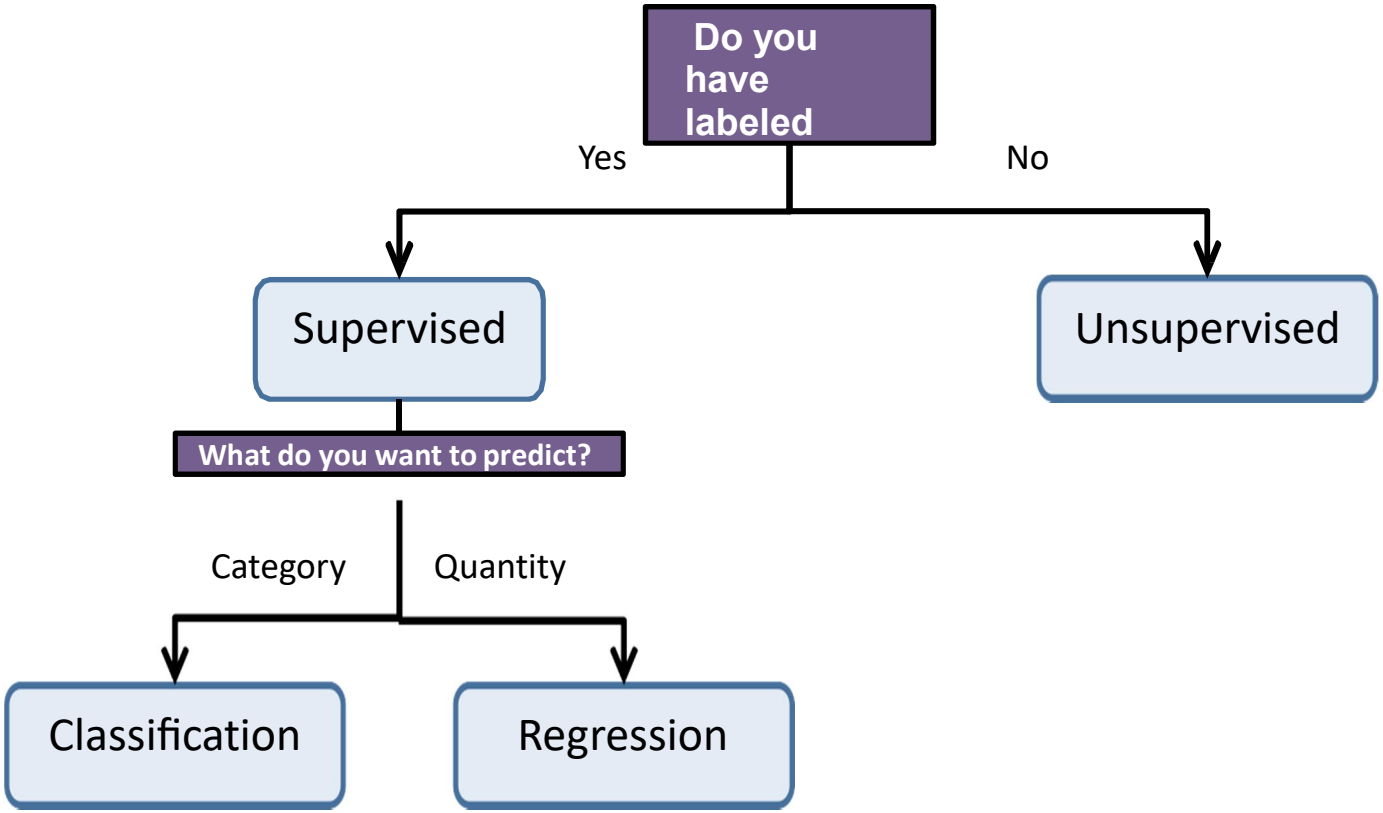
Ví dụ: Dự đoán

- ✓ Xu thế giá cổ phiếu Z sẽ tăng hay giảm trong năm tính từ thời điểm này?
- ✓ Giao dịch thẻ tín dụng là gian lận hoặc hợp pháp?

Học có giám sát: Phân lớp và Hồi quy

- Bài toán phân lớp cũng có thể trình bày theo dạng hồi quy
  - Bài toán 2 lớp: *"Xác xuất để một quan sát/mẫu thuộc lớp 1?"*
  - Một số phương pháp học máy có thể xử lý được cả 2 dạng bài toán (CART, mạng nơ-ron, rừng ngẫu nhiên)
- Đối với việc lựa chọn một phương pháp học máy, đầu vào là định lượng/định tính không quá quan trọng.

Các dạng giải thuật học máy

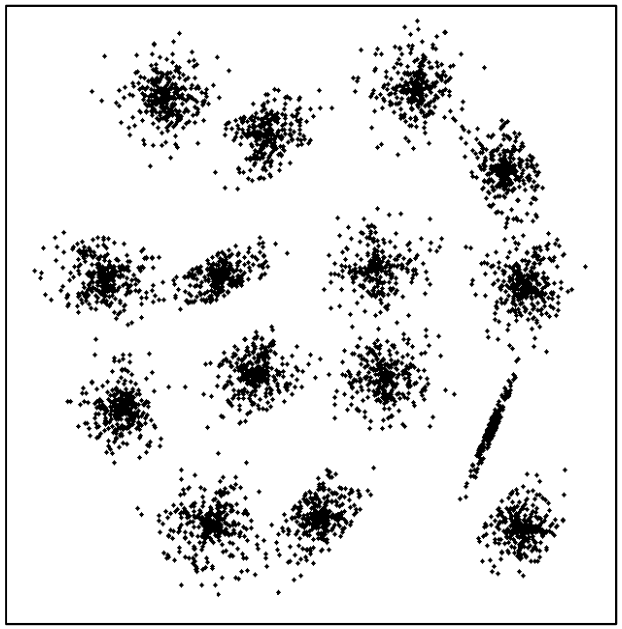


Học máy không giám sát

Phân cụm và Giảm chiều dữ liệu

- *Phân tích cụm*

  Chia dữ liệu thành các tập con mà chúng có các đặc tính chung

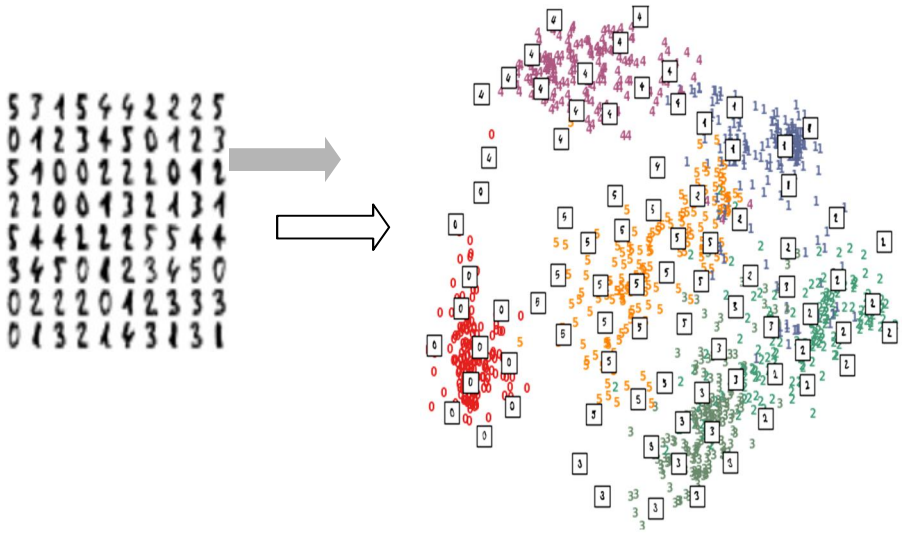


Học máy không giám sát

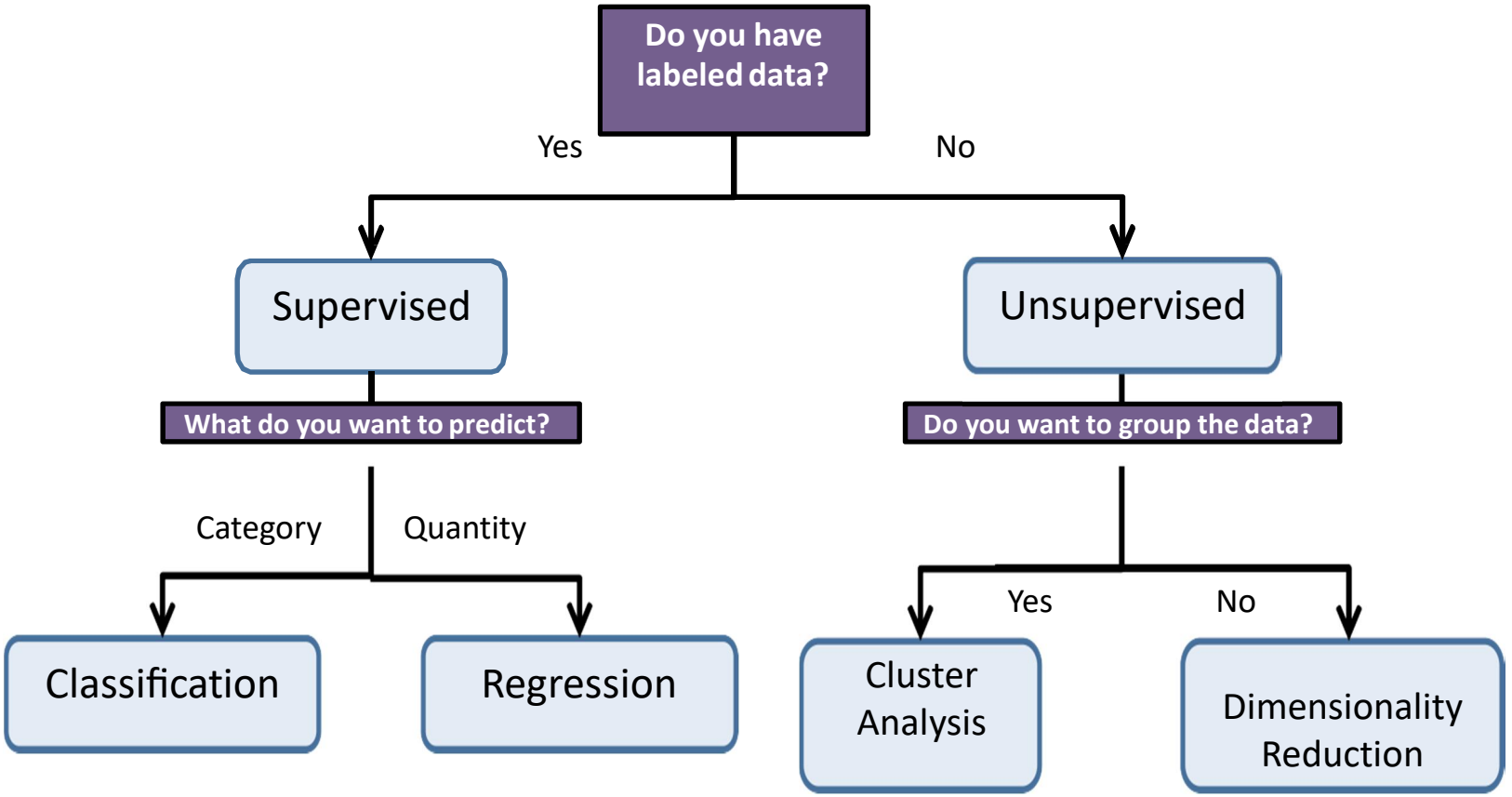
Phân cụm & Giảm chiều dữ liệu

- *Giảm chiều dữ liệu*

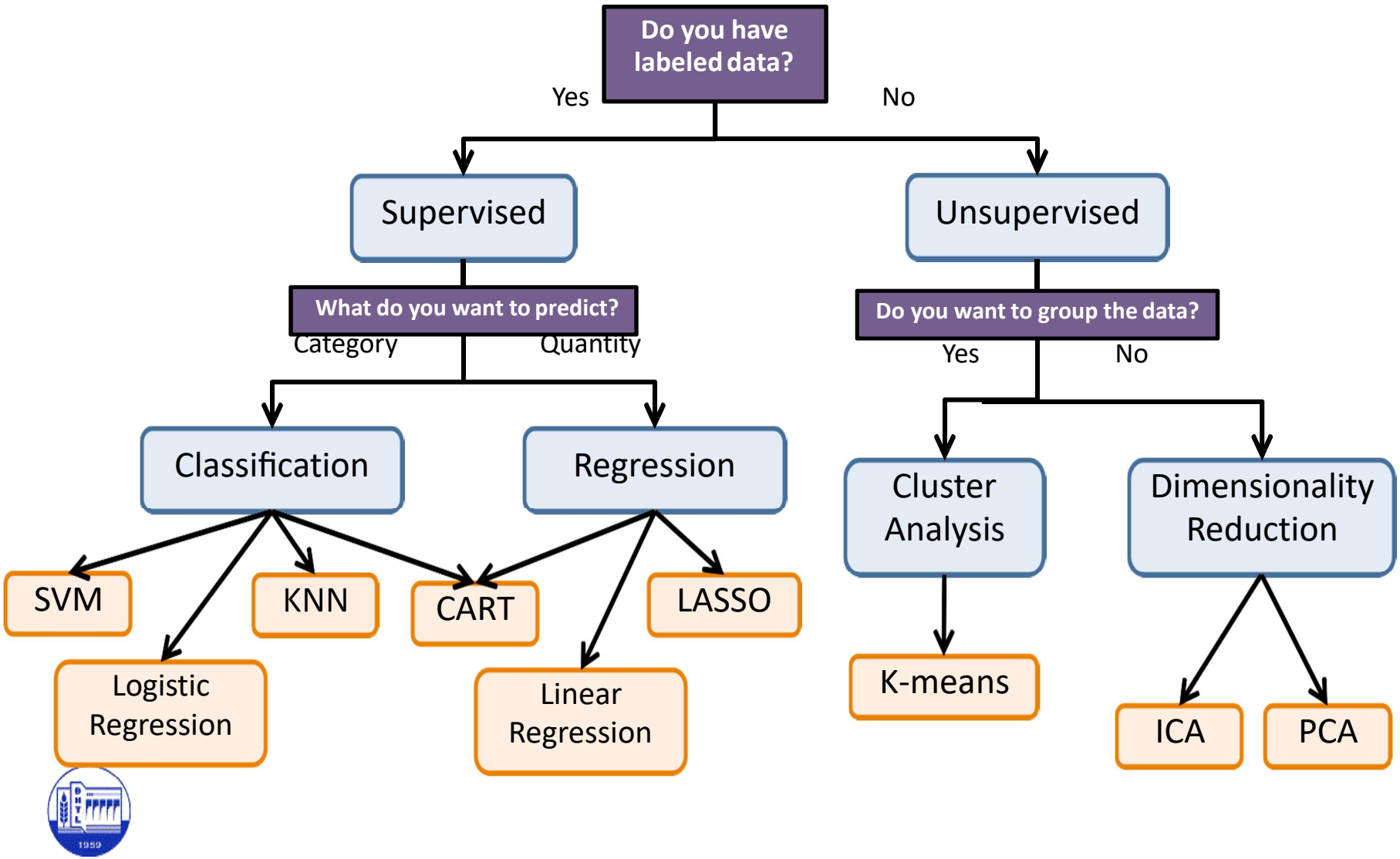
Tạo ra các biến mới từ các biến đầu vào ban đầu sao cho bảo toàn được các thông tin quan trọng



Các dạng giải thuật học máy



Các giải thuật học máy



Giải thuật Học máy "Tốt nhất"

- Tin tồi: *Không có giải thuật nào tốt nhất*

Không có giải thuật học máy nào thực hiện tốt cho mọi bài toán

- Tin tốt: *Tất cả các giải thuật học máy đều tốt*

Mỗi giải thuật học máy thực hiện tốt cho một số bài toán

- Định lý "***No free lunch***"

Wolpert (1996): Các giải thuật thực hiện như nhau khi ta lấy trung bình kết quả chúng thực hiện trên tất cả các bài toán.

Trade-offs (đánh đổi) trong Học máy

- Bias vs. variance
- Độ chính xác vs. Khả năng diễn giải
- Độ chính xác vs. Khả năng mở rộng giải thuật
- Phạm vi kiến thức vs. Hướng dữ liệu
- Nhiều dữ liệu vs. Giải thuật tốt hơn

Chuẩn bị dữ liệu

- Các giải thuật học máy cần phải có dữ liệu!
- **Tiền xử lý dữ liệu** để chuyển đổi dữ liệu trước khi áp dụng vào giải thuật học máy
  - Lấy mẫu: chọn tập con các quan sát/mẫu
  - Trích chọn thuộc tính: Chọn các biến đầu vào
  - Chuẩn hóa dữ liệu (Normalization, standardization, scaling, binarization)
  - Xử lý dữ liệu thiếu và phần tử ngoại lai (missing data and outliers)

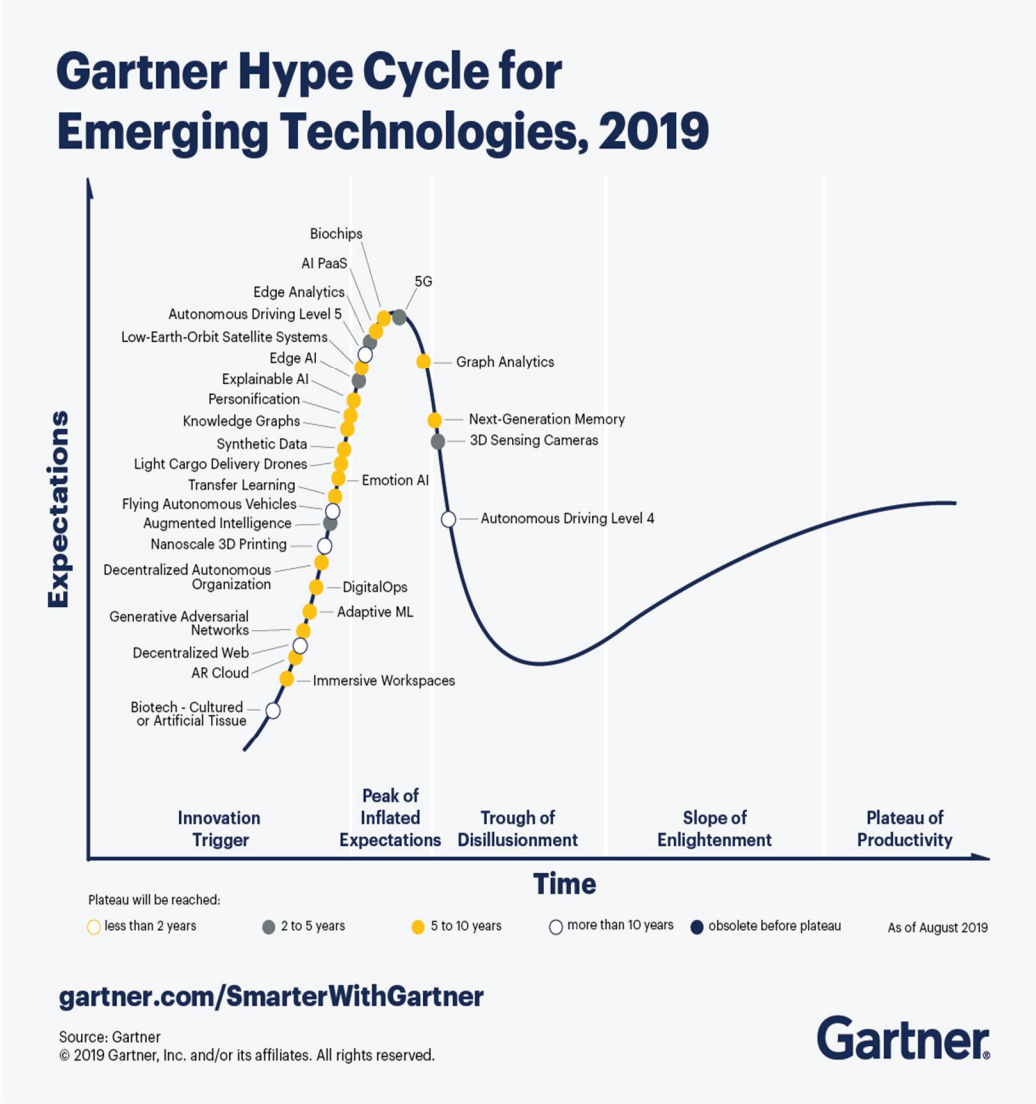
Chuẩn bị dữ liệu

- Ngoài ra, còn phụ thuộc vào giải thuật học máy
  - ❖ Cây quyết định có thể xử lý dữ liệu thiếu/phần tử ngoại lai
  - ❖ PCA yêu cầu dữ liệu đã được chuẩn hóa

Chu kỳ kỳ vọng của các công nghệ năm 2019

Các giai đoạn:

- Giai đoạn phát minh,
- Giai đoạn kỳ vọng đạt đỉnh,
- Giai đoạn thất vọng,
- Giai đoạn phục hồi,
- Giai đoạn phổ biến



Chu kỳ kỳ vọng của AI năm 2019