

# CSE 445: HỌC MÁY (Machine Learning)

Nguyễn Thanh Tùng, Trần Thị Ngân  
Khoa Công nghệ thông tin – Đại học Thủy lợi  
*tungnt@tlu.edu.vn*, [ngantt@tlu.edu.vn](mailto:ngantt@tlu.edu.vn)



# TỔNG QUAN MÔN HỌC

- Tên môn học: Học máy (Machine Learning)
- Mã số môn học: CSE445
- Thời lượng: 3 Tín chỉ (30LT+15TH)



# ĐÁNH GIÁ

- 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 =  $\text{ĐQT} \times 50\% + \text{THM} \times 50\%$



# 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 nổi tiếng.
- Biết vận dụng các phương pháp học máy tiên tiến dùng cho phân tích dữ liệu ra quyết định.
- Kỹ năng thực hành, thiết kế thí nghiệm sử dụng ngôn ngữ R.
- 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
  - Đã 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

“*An Introduction to Statistical Learning with Applications in R*” (ISL) by James, Witten, Hastie and Tibshirani\* được cung cấp **miễn phí** (pdf) tại:

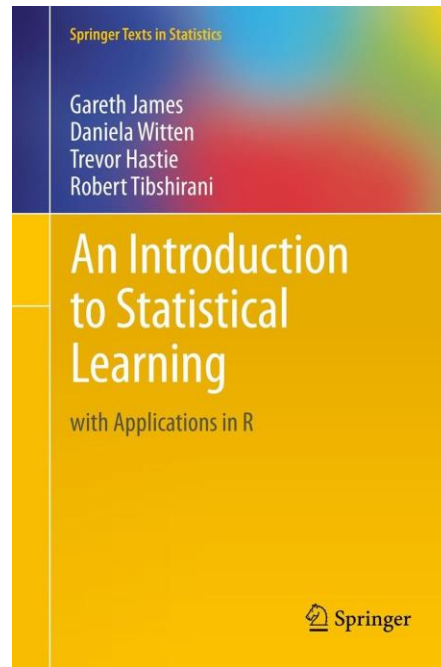
[www-bcf.usc.edu/~gareth/ISL/](http://www-bcf.usc.edu/~gareth/ISL/)

Sách tham khảo:

“*The Elements of Statistical Learning*” (ESL) by Hastie, Tibshirani and Friedman được cung cấp **miễn phí** (pdf) tại:

[statweb.stanford.edu/~tibs/ElemStatLearn/](http://statweb.stanford.edu/~tibs/ElemStatLearn/)

<https://machinelearningcoban.com>



# TÀI LIỆU ONLINE

- Trang web:
  - <https://sites.google.com/a/wru.vn/cse445fall2018>
  - Bài giảng, tài liệu và các thông báo của môn học:  
trao đổi trên Piazza





# CSE 445 Hỏi&Đáp




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

**[piazza.com/tlu.edu.vn/fall2019/cse445fall2019](https://piazza.com/tlu.edu.vn/fall2019/cse445fall2019)**



# Ngôn ngữ lập trình R

- R: [www.r-project.org](http://www.r-project.org)



[Home]

**Download**

[CRAN](#)

**R Project**

[About R](#)  
[Contributors](#)  
[What's New?](#)  
[Mailing Lists](#)  
[Bug Tracking](#)  
[Conferences](#)  
[Search](#)

**R Foundation**

[Foundation](#)  
[Board](#)

## The R Project for Statistical Computing

### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

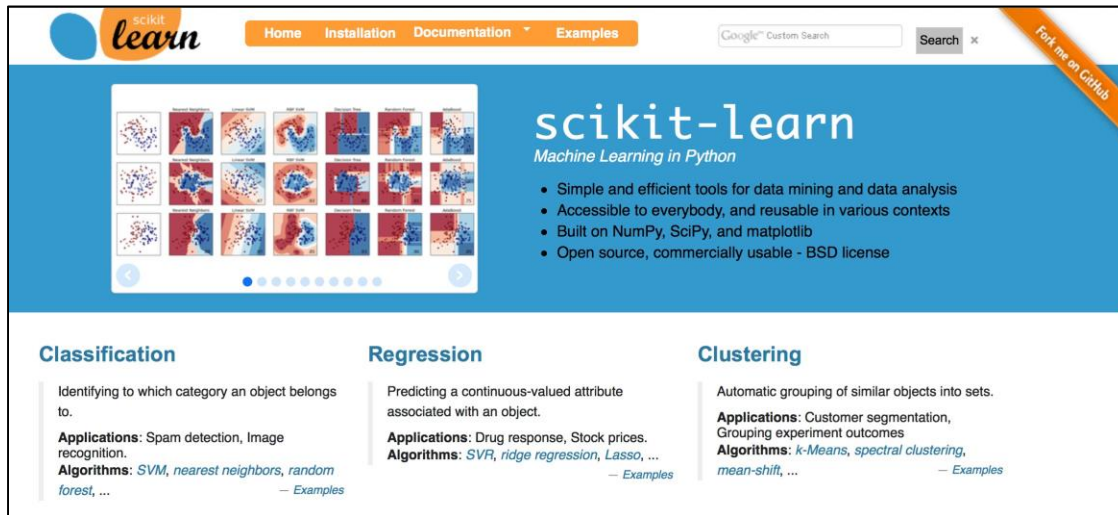
### News

- [R version 3.2.1 \(World-Famous Astronaut\)](#) has been released on 2015-06-18.
- [R version 3.1.3 \(Smooth Sidewalk\)](#) has been released on 2015-03-09.
- [The R Journal Volume 6/2](#) is available.
- [useR! 2015](#), will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- [useR! 2014](#), took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.



# Ngôn ngữ lập trình Python

- Python: [www.python.org](http://www.python.org)
- scikit-learn: <http://scikit-learn.org/>



# Giới thiệu về Học máy

- Học máy (machine learning) là gì?
  - 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

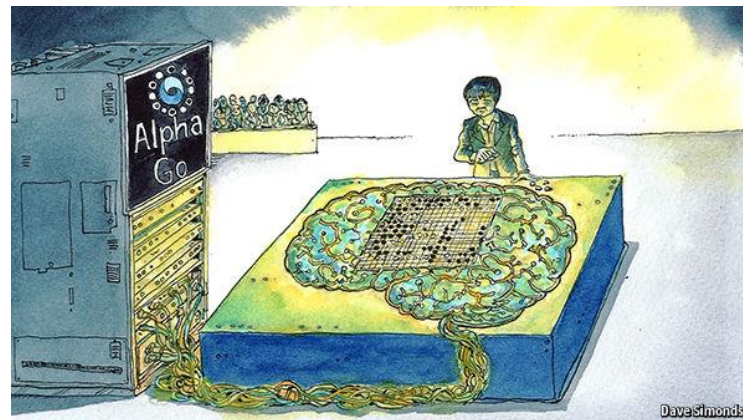
- “*Lĩnh vực nghiên cứu giúp máy tính có khả năng tự học khi không được lập trình trước*”

– 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  
VD: Lọc thư rác



# Các ứng dụng của Học máy

- Trong hệ thống tự động ra quyết định  
VD : 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.  
VD: Xe không người lái



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



# Các ứng dụng của Học máy



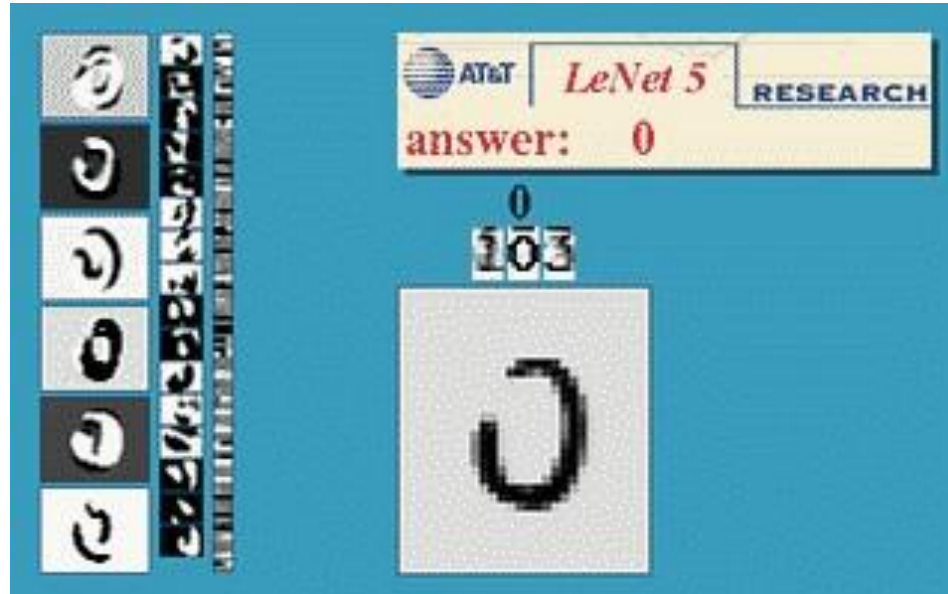
[Courtesy of Dean Pomerleau]

[Video: Autonomous Driving](#)



# Các ứng dụng của Học máy

- Cho các hệ thống tự động có lập trình phức tạp.  
VD: Nhận dạng chữ viết tay



[LeNet-5 Convolutional Neural Net](#)



# Các ứng dụng của Học máy

- Dùng cho khai phá dữ liệu  
VD : Bệnh án điện tử



“Mining Electronic Records  
for Revealing Health Data”

[New York Times \(Jan 14, 2013\)](#)



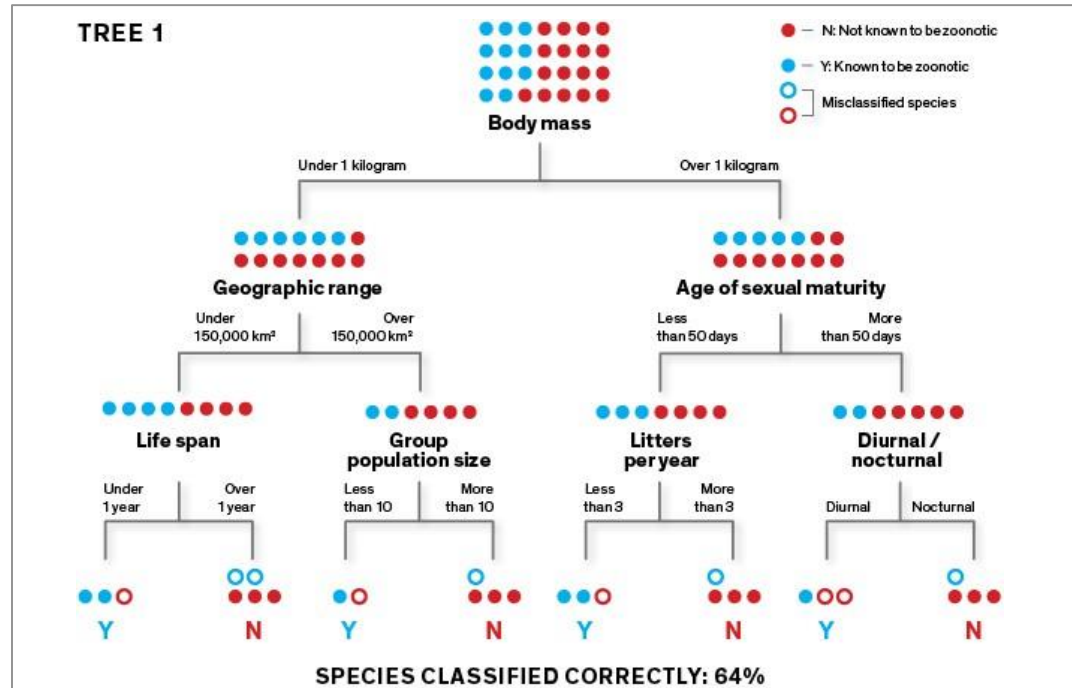
# Các ứng dụng của Học máy

- Trong các hệ thống tùy biến  
VD : Hệ thống gợi ý sản phẩm



# Các ứng dụng của Học máy

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



# 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.



# Các giải thuật Học máy

- 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 máy bán giám sát (semi-supervised learning),*
  - ✓ *Học tăng cường (reinforcement learning),*
  - ✓ *Các hệ thống khuyến nghị (recommender systems), etc.*

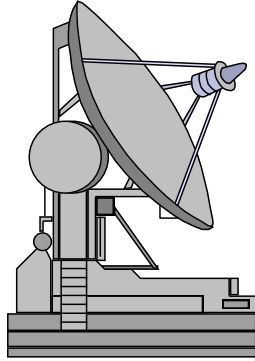


# Mô hình Học máy (Machine learning Model)



# Mục đích của mô hình Học máy

Truyền thông

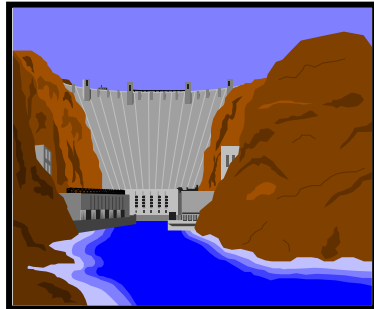


Ra quyết định



Phân tích dữ liệu  
& các mô hình

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



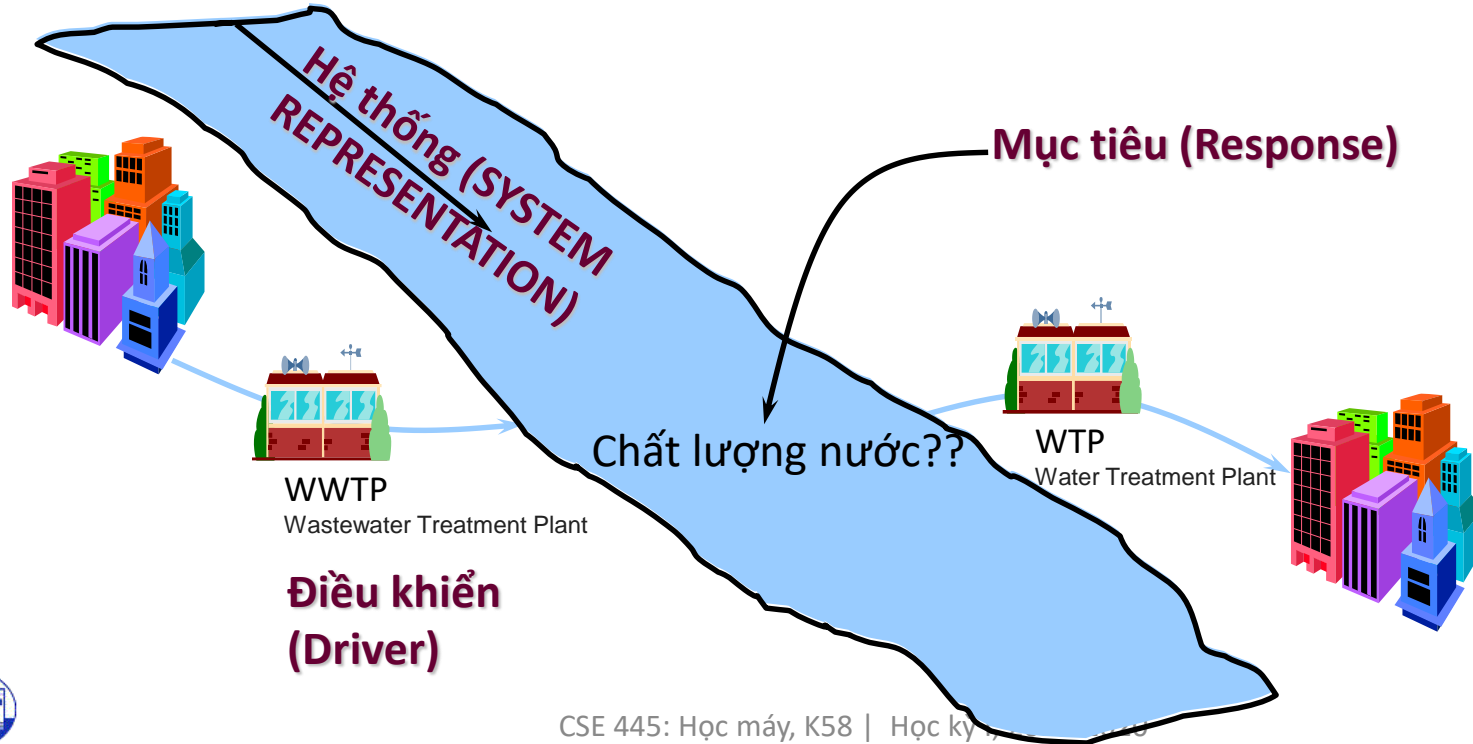
# Tại sao dùng kỹ thuật thống kê?

- 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
- Chúng thường được sử dụng rộng rãi trong kỹ thuật



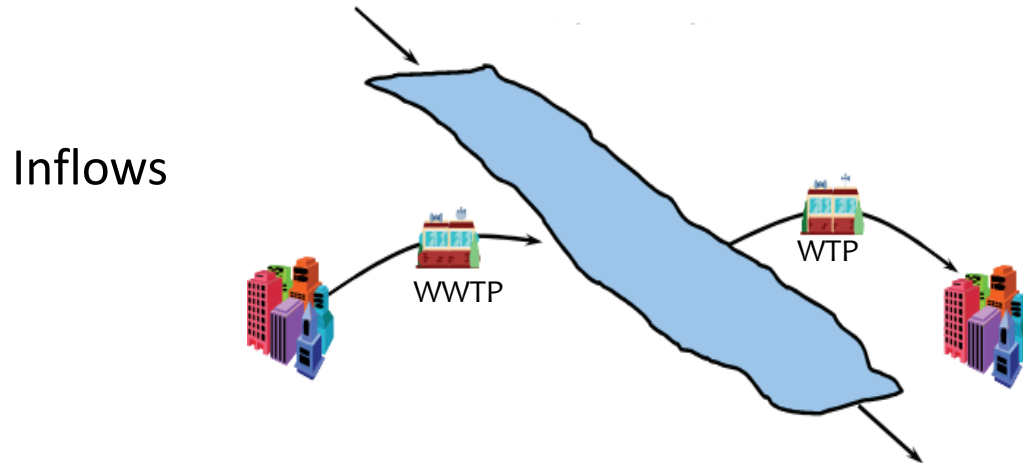
# Các thành phần của mô hình

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



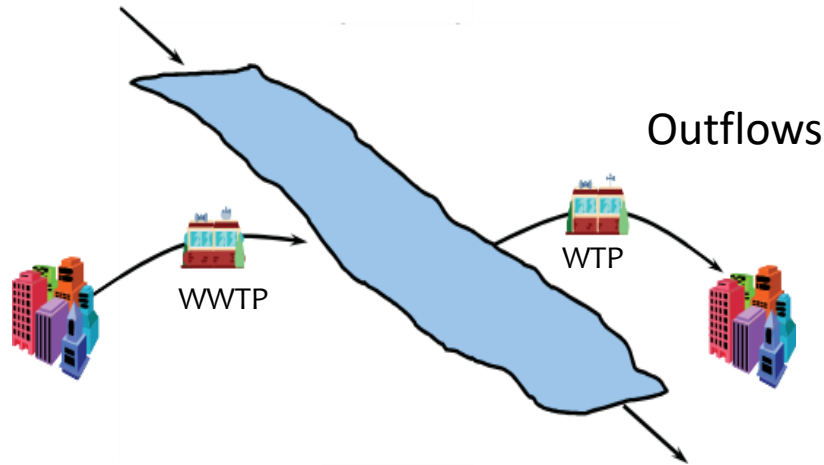
# Các thành phần của mô hình

**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

- $X$ : **Tập biến đầu vào** (tập biến dự đoán, biến độc lập hoặc các đặc trưng) (input variables, predictors, independent variables or features).
- $Y$ : **Biến đầu ra** (biến đích hoặc biến phụ thuộc) (output variables, response or dependent variable)
- **Học máy thống kê** (*Statistical Learning*): là 1 tập các giải pháp ước lượng hàm  $f$  để mô tả mối quan hệ giữa tập biến đầu vào và biến đầu ra:

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



# 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  $f$ .

Các cặp vào/ra:  $\left(X^{(1)}, Y^{(1)}\right), \dots, \left(X^{(n)}, Y^{(n)}\right)$



# Đặt bài toán và Thuật ngữ

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

- Phương pháp để ước lượng  $f$  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  $f$ .



# 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).

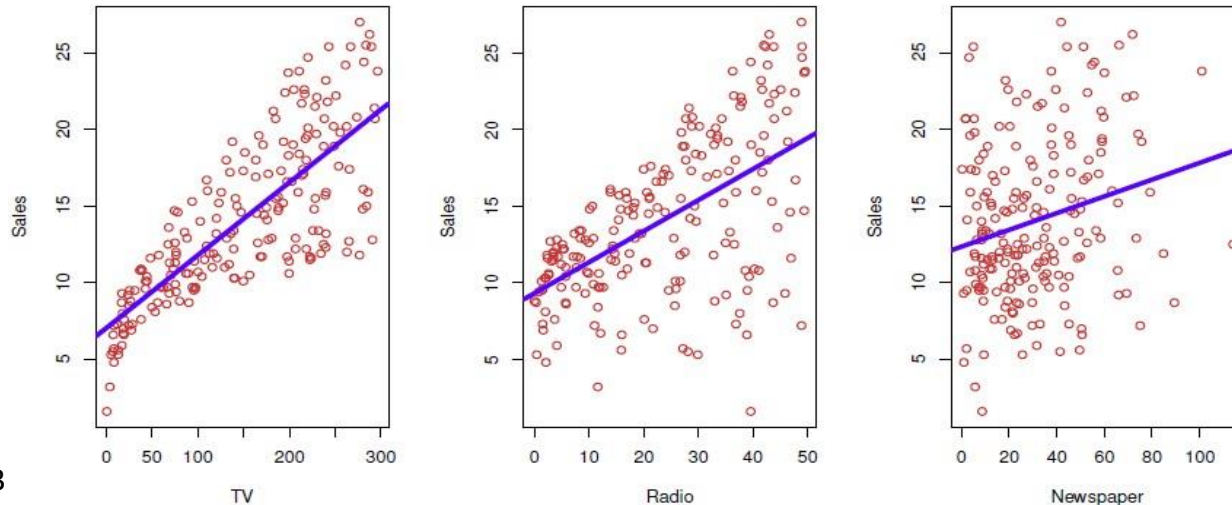


Figure 2.1 , ISL 2013



# 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:

$$\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$$

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





# Các mô hình tham số và phi tham số

- Các 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  $f$



# Các mô hình tham số và phi tham số

- Các 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 vs. 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
  - VD: 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 over-fitting)
- 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

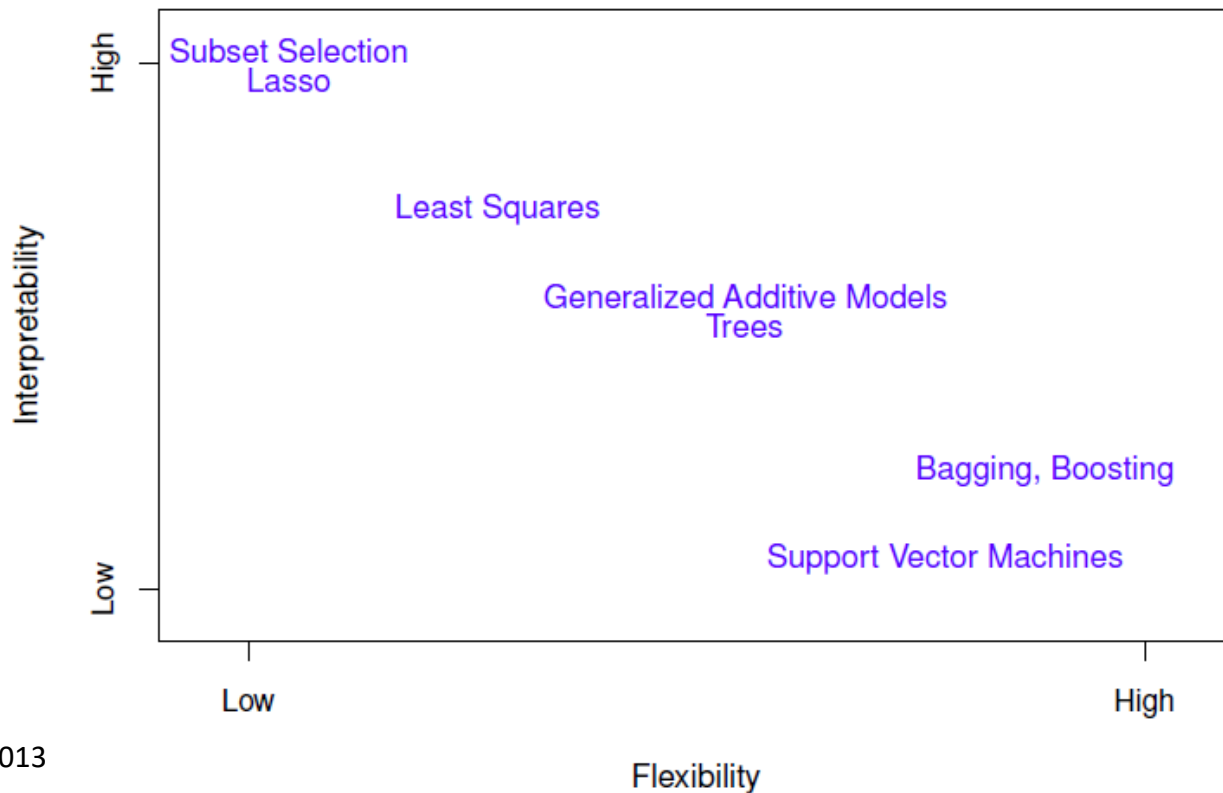


Figure 2.7, ISL 2013



# HỌC MÁY

- Bài toán học máy được chia làm 2 dạng chính:
  - *Học có giám sát (Supervised Learning)*
  - *Học không giám sát (Unsupervised 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



# Học không giám sát

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





# CÁC BÀI TOÁN TRONG HỌC MÁY



# HỌC CÓ GIÁM SÁT: PHÂN LỚP VÀ HỒI QUY

- Bài toán học có giám sát được chia làm 2 dạng:
  - *Phân lớp*
  - *Hồi quy*



# 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  $Z$  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**

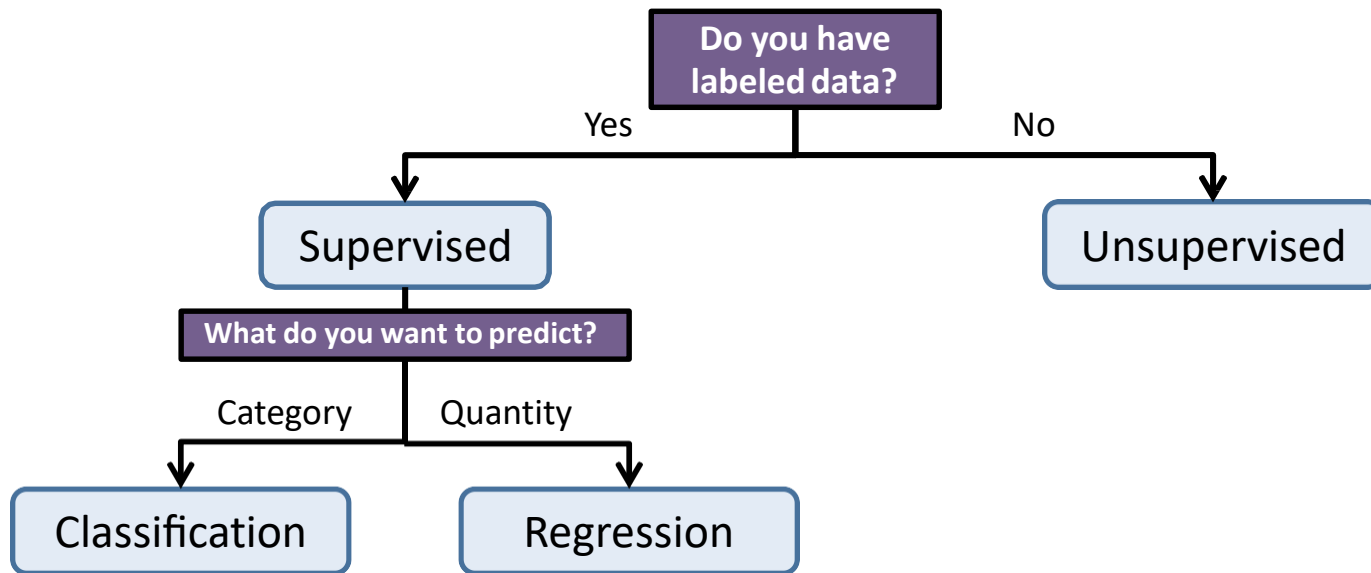


# HỌC CÓ GIÁM SÁT: PHÂN LỚP VÀ HỒI QUY

- 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 1 phương pháp học máy, đầu vào là định lượng/định tính không quá quan trọng.



# CÁC BÀI TOÁN TRONG HỌC MÁY

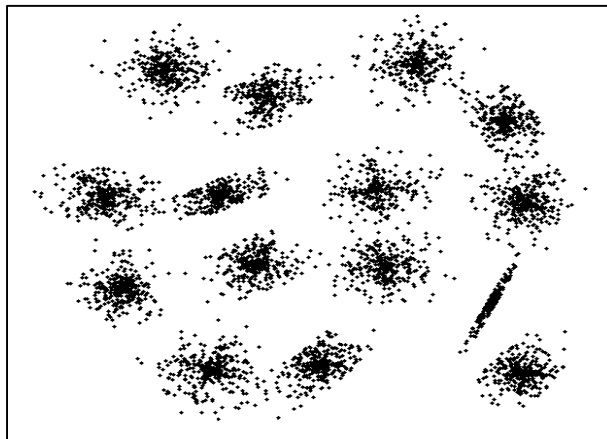


# Học máy không giám sát

## Phân cụm & Giảm chiều dữ liệu

- *Phân tích cụm*

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

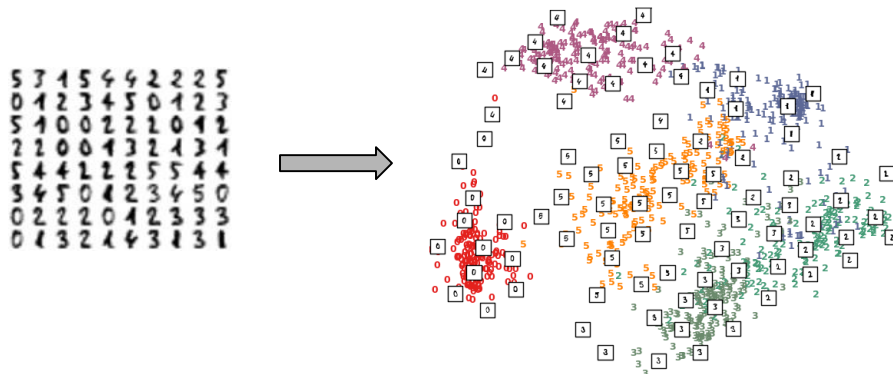


# Học máy không giám sát

## Phân cụm & Giảm chiều dữ liệu

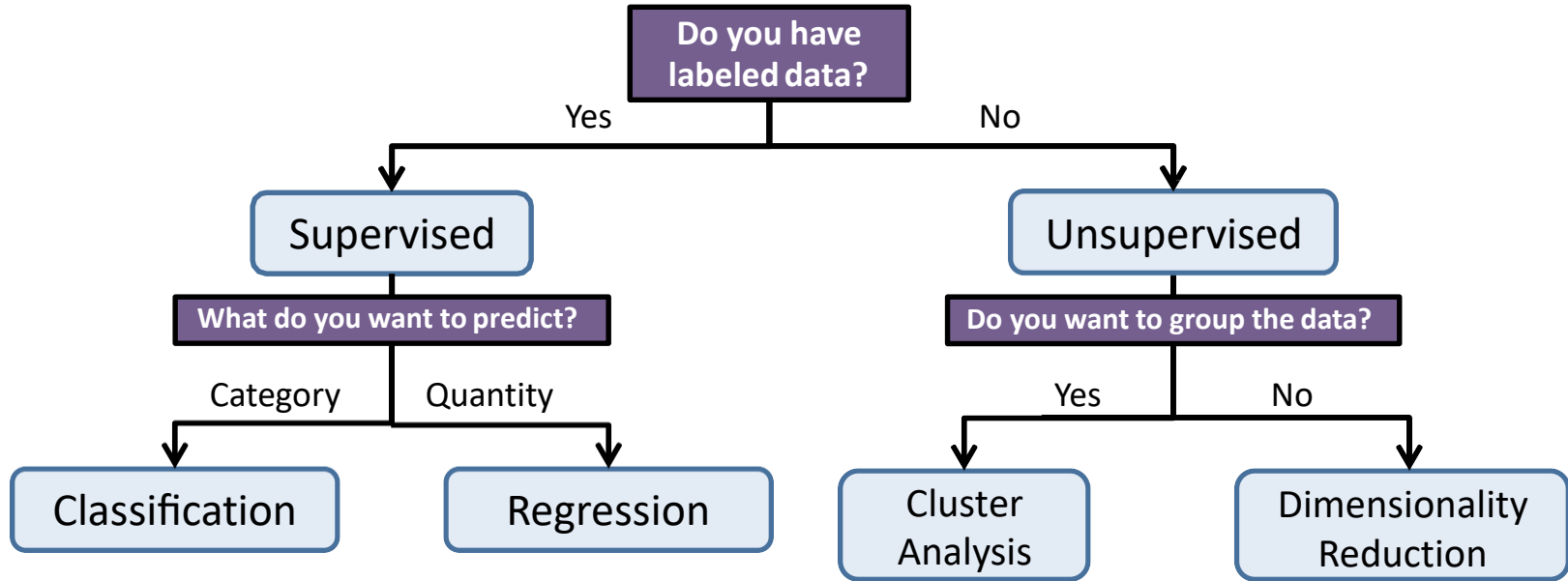
- Giảm chiều dữ liệu*

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

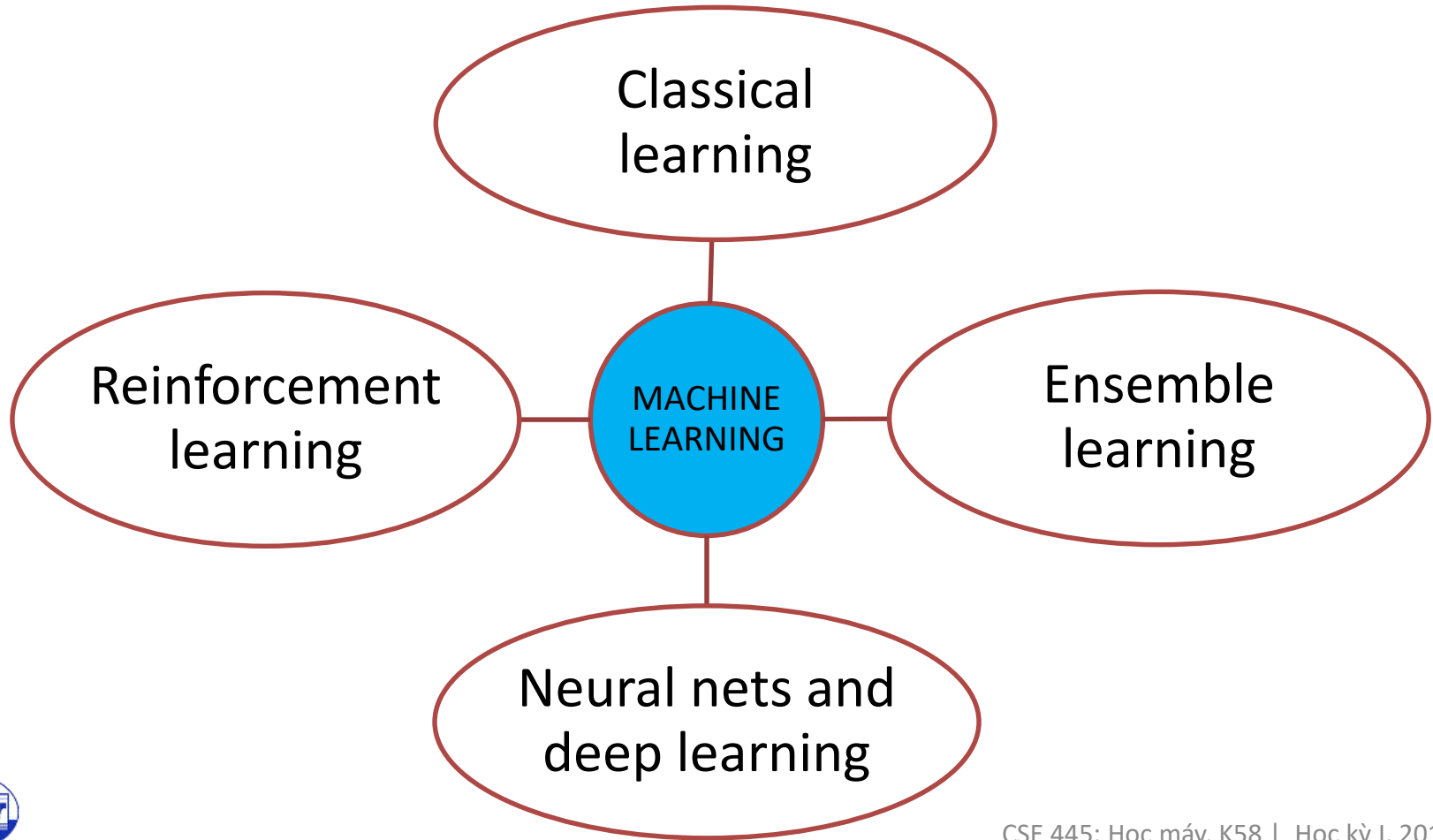




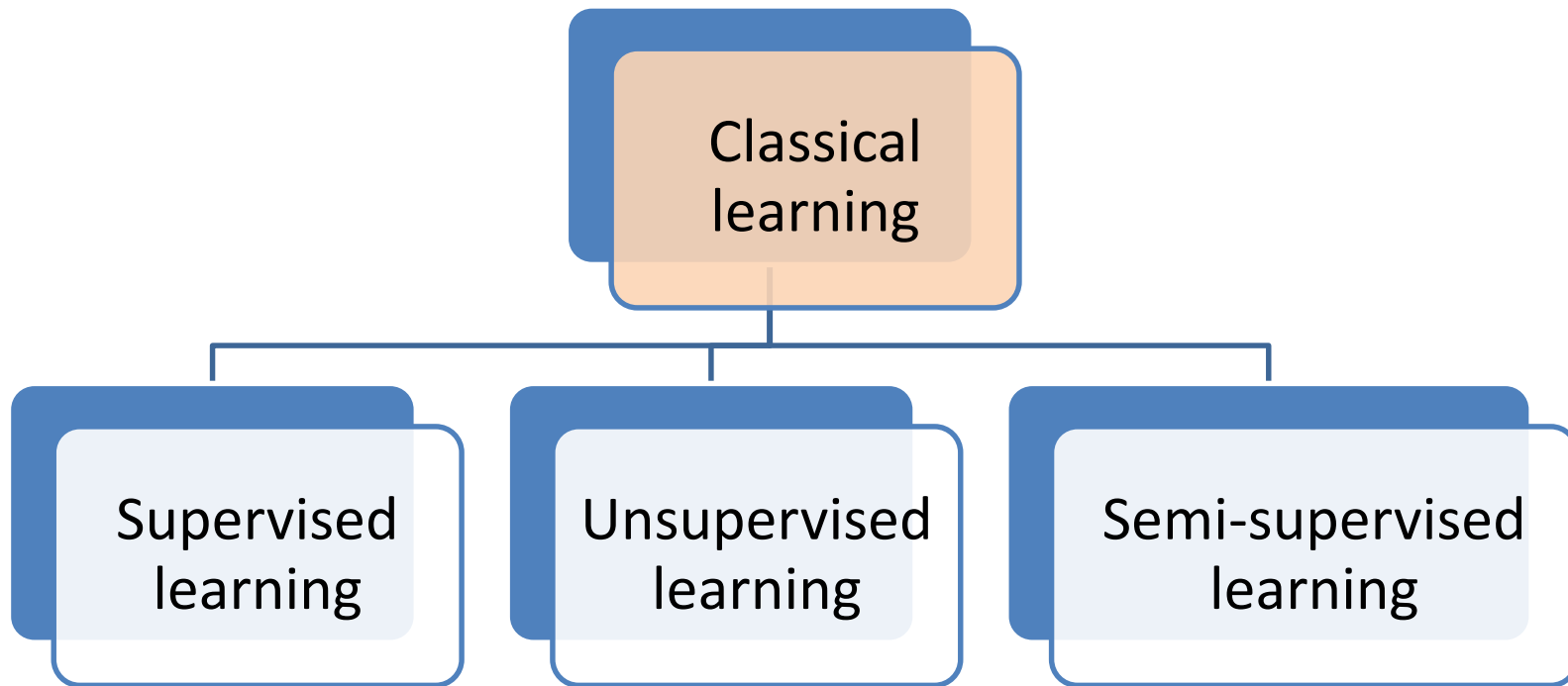
# CÁC BÀI TOÁN TRONG HỌC MÁY



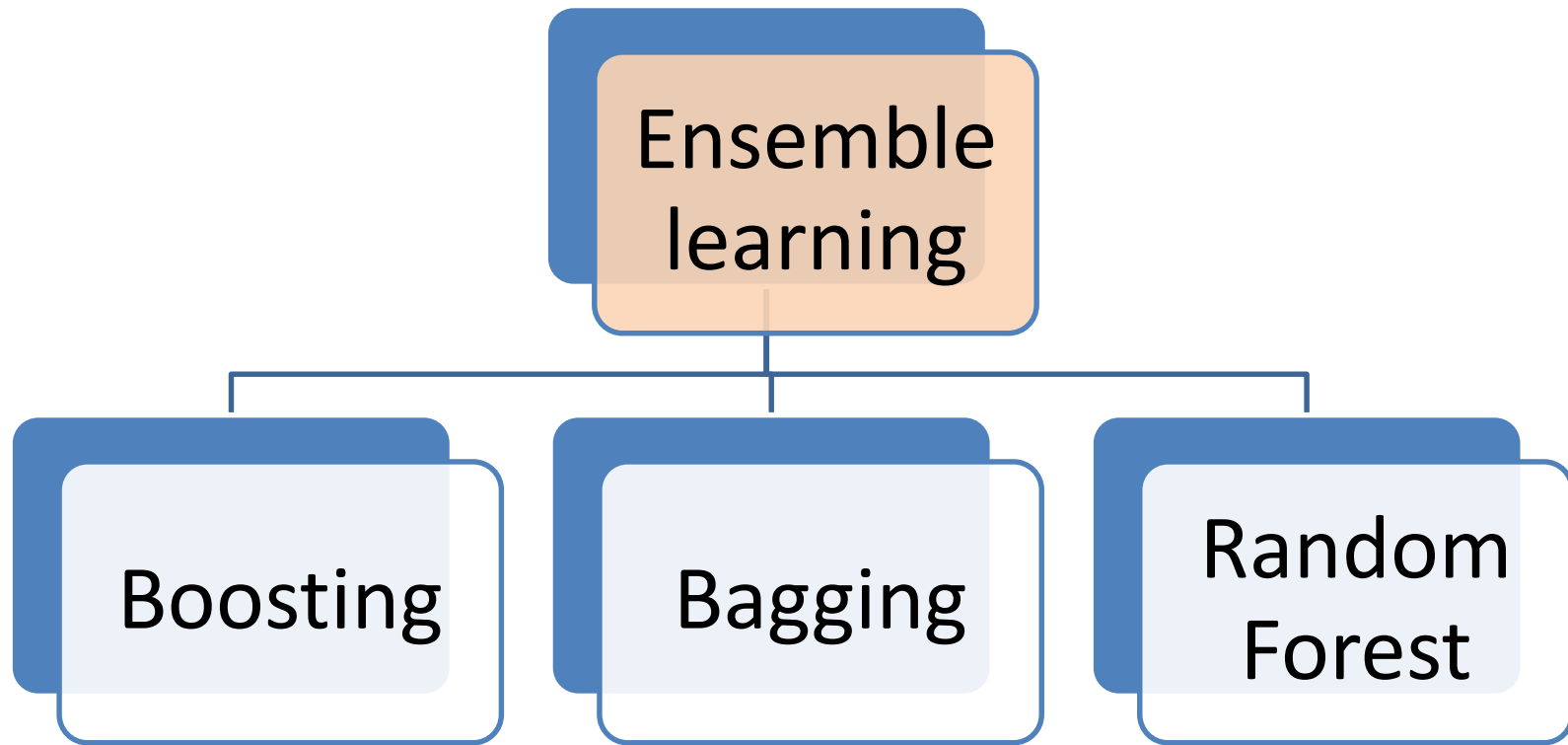
# CÁC GIẢI THUẬT HỌC MÁY



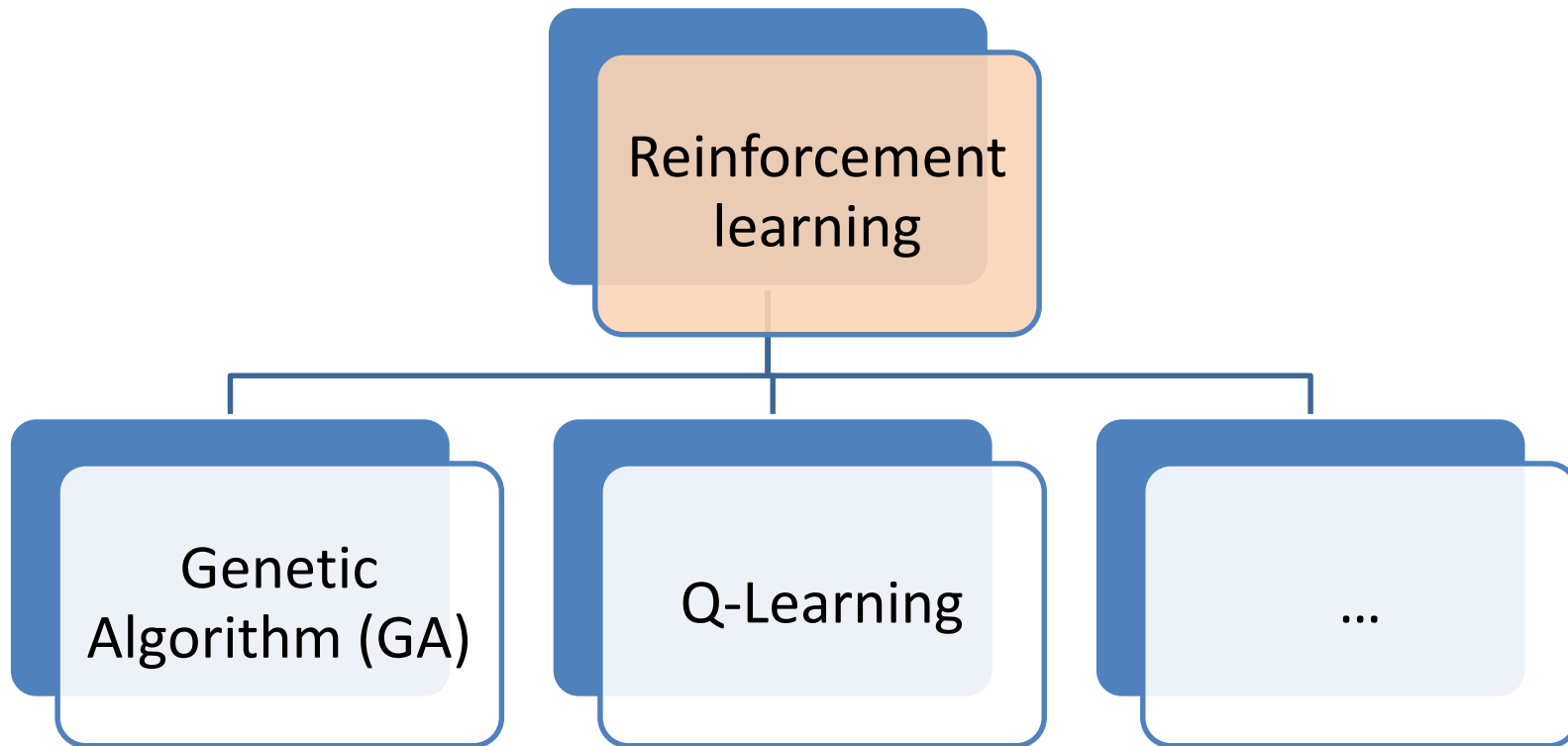
# CÁC GIẢI THUẬT HỌC MÁY



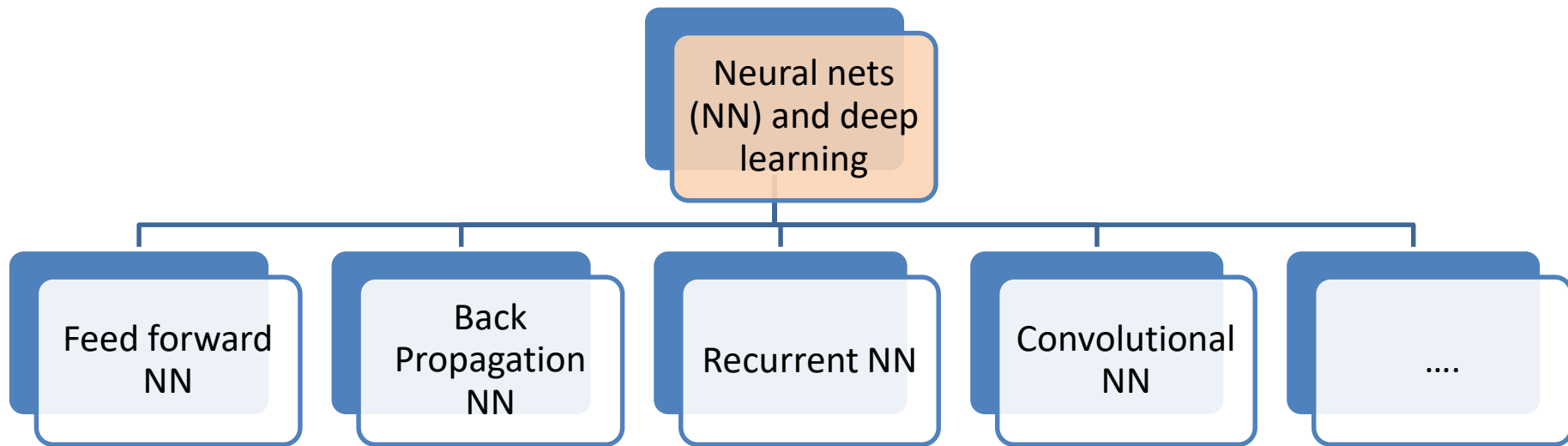
# CÁC GIẢI THUẬT HỌC MÁY



# CÁC GIẢI THUẬT HỌC MÁY



# CÁC GIẢI THUẬT HỌC MÁY



# MỘT SỐ GIẢI THUẬT TIÊU BIỂU

## Classical learning

- Supervised learning: KNN, Linear regression, generalized methods, CART, SVM, Logistic Regression
- Unsupervised learning: K-means clustering, Hierarchical clustering

## Ensemble learning

- Boosting
- Bagging
- Random Forest

## Neural nets and deep learning

- Artificial neural networks
- Feed Forward neural network

## Reinforcement learning

- GA
- Q-learning



# 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

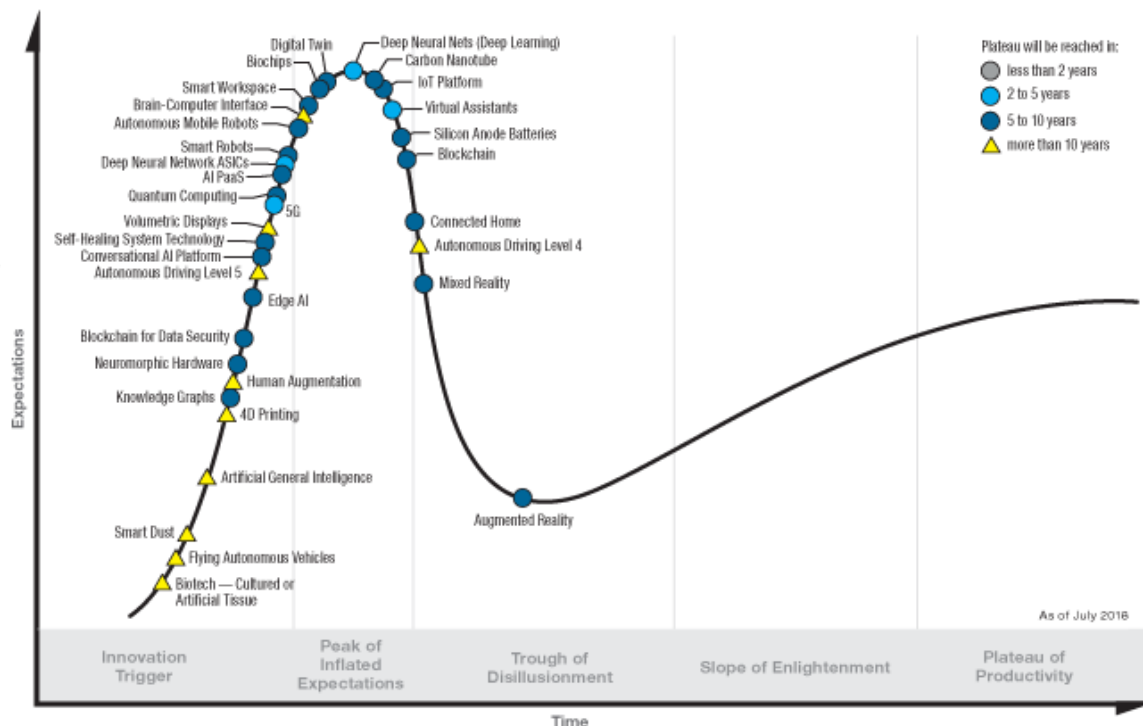


# Chu kỳ kỳ vọng của các công nghệ năm 2018

## Hype Cycle for Emerging Technologies, 2018

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



Questions?

