

# CSE 445: Học máy (Machine Learning)

Nguyễn Thanh Tùng  
Khoa Công nghệ thông tin – Đại học Thủy Lợi  
[tungnt@tlu.edu.vn](mailto:tungnt@tlu.edu.vn)

Website môn học: <https://sites.google.com/a/wru.vn/cse445fall2017>

Bài giảng có sử dụng hình vẽ trong cuốn sách “An Introduction to Statistical Learning with Applications in R” với sự cho phép của tác giả, có sử dụng slides các khóa học CME250 của ĐH Stanford và IOM530 của ĐH Southern California



# 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.



# Các ứng dụng của Học máy

- “*Lĩnh vực nghiên cứu giúp máy tính có khả năng tự học khi không được lập trình trước*” ([A] field of study that gives computers the ability to learn without being explicitly programmed.)



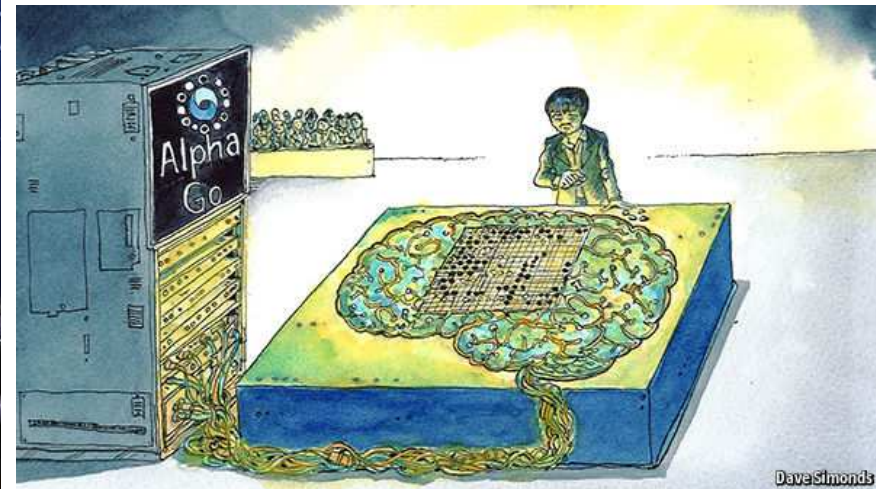
– Arthur Samuel (1959)



# Các ứng dụng của Học máy



- AlphaGo thắng nhà vô địch thế giới cờ vây



# Các ứng dụng của Học máy

- Học máy được sử dụng ở đâu?



# 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

- Cho 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



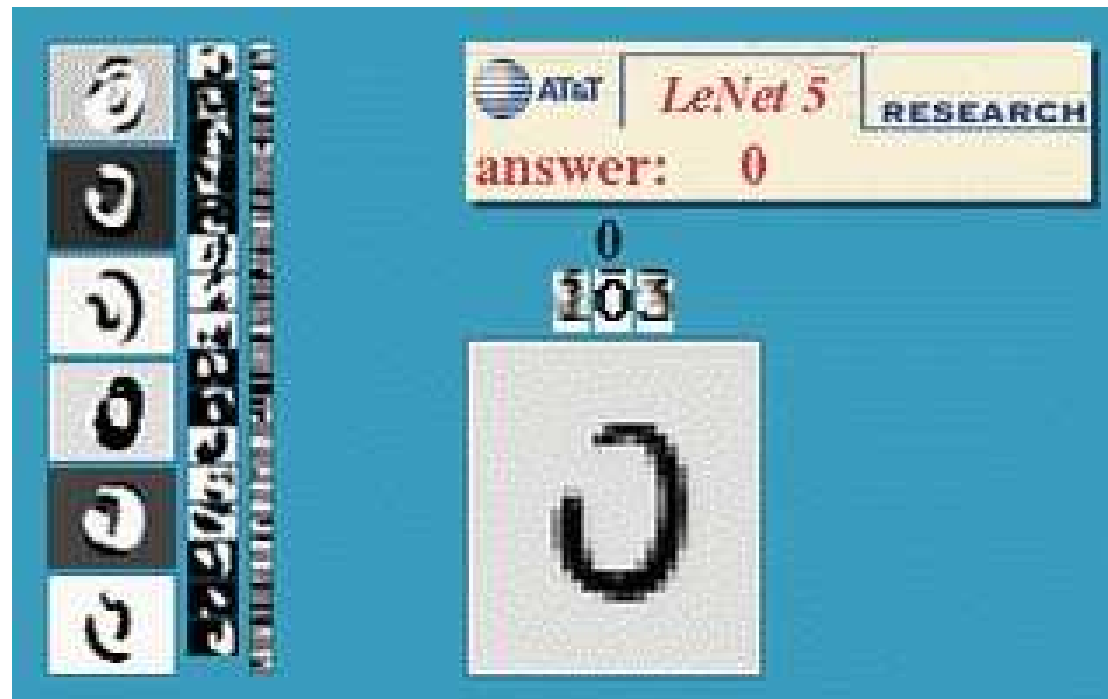
[Courtesy of Dean Pomerleau]

[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 \(Jan 14, 2013\)](#)



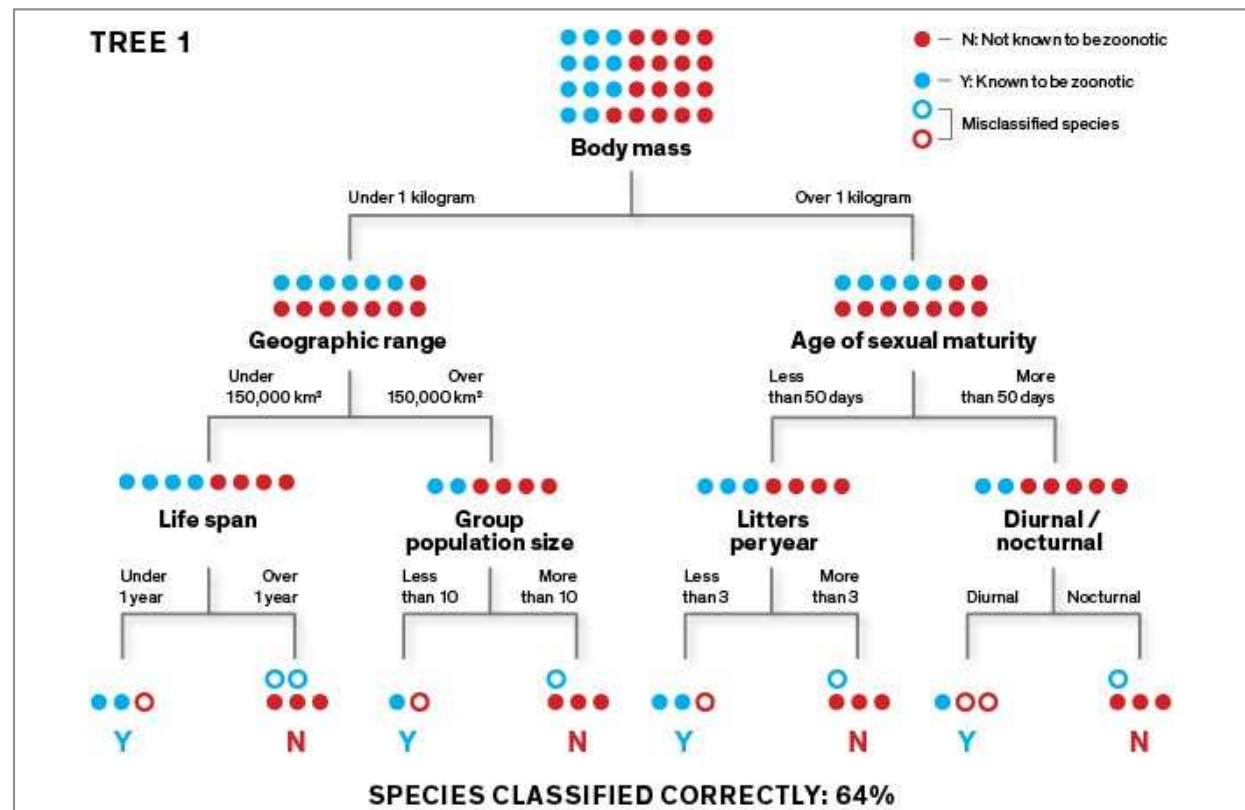
# Các ứng dụng của Học máy

- 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.
- 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
- 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.







# Thông tin môn học



# Môn Học máy

- Trang web:
  - <https://sites.google.com/a/wru.vn/cse445fall2017>
  - Bài giảng, tài liệu và các thông báo của môn học.
- Thời khóa biểu
  - 13/11-07/01/18
    - Thứ 2 tiết 7, 8 tại 426A4
    - Thứ 4 tiết 7, 8 tại 426A4
  - Lab: từ ngày 27/11-07/01/18 tại P.402-C5.



# 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)



# 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.



# Sách giáo khoa

*"An Introduction to Statistical Learning with Applications in R" (ISL)* by James, Witten, Hastie and Tibshirani\*

cung cấp **miễn phí** (pdf) tại: [www-bcf.usc.edu/~gareth/ISL/](http://www-bcf.usc.edu/~gareth/ISL/)

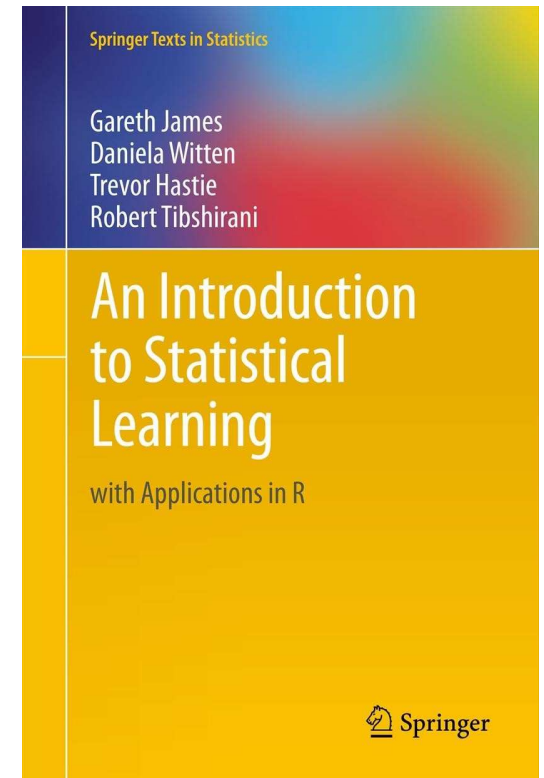
Sách tham khảo:

*"The Elements of Statistical Learning" (ESL)* by Hastie, Tibshirani and Friedman

cung cấp **miễn phí** (pdf) tại: [statweb.stanford.edu/~tibs/ElemStatLearn/](http://statweb.stanford.edu/~tibs/ElemStatLearn/)

<https://machinelearningcoban.com>

\*Một số hình ảnh trình bày trong bài giảng được lấy từ cuốn *"An Introduction to Statistical Learning, with applications in R"* (Springer, 2013) được sự đồng thuận của các tác giả: G. James, D. Witten, T. Hastie and R. Tibshirani





# Các yêu cầu môn học

- 3 tín chỉ
- Điểm kết thúc học phần
- Các yêu cầu
  - Bài tập: sinh viên có thể lựa chọn bài tập để làm và nộp, điểm lấy từ cao xuống thấp để tính kết quả học tập.




# Bài tập

- Bài tập được giao từ cuốn ISL
- Sinh viên cần hoàn thành 50% số điểm của khối lượng bài tập để nhận được điểm đạt.
- Sinh viên phải hoàn thành bắt buộc với số lượng tối thiểu:
  - **4 bài tập bất kỳ** trong số các bài tập được giao
- Hạn nộp bài tập theo thời khóa biểu của môn học.



# 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.

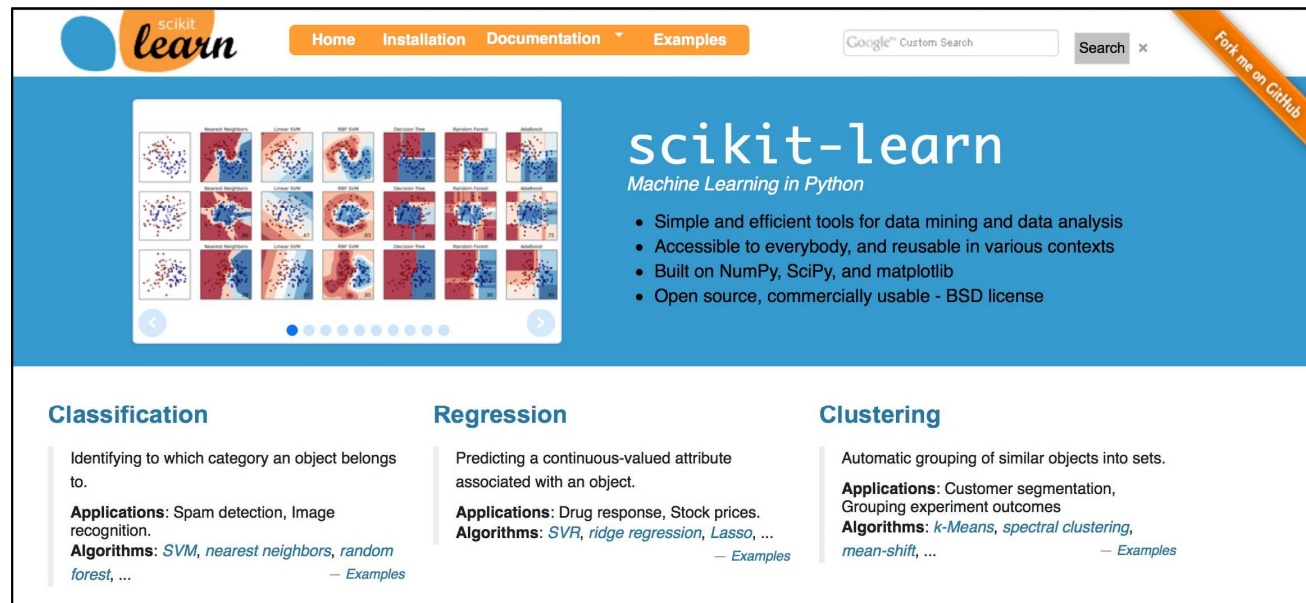
### News

- [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/>



# 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
- Website:  
<https://piazza.com/tlu.edu.vn/fall2017/cse445/home>





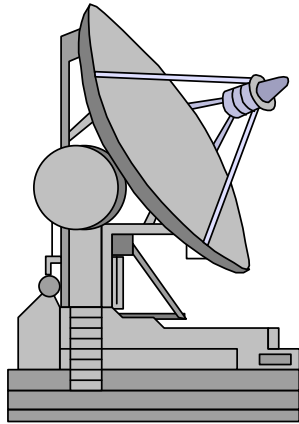


# Mô hình Học máy (Machine learning Model)



# 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



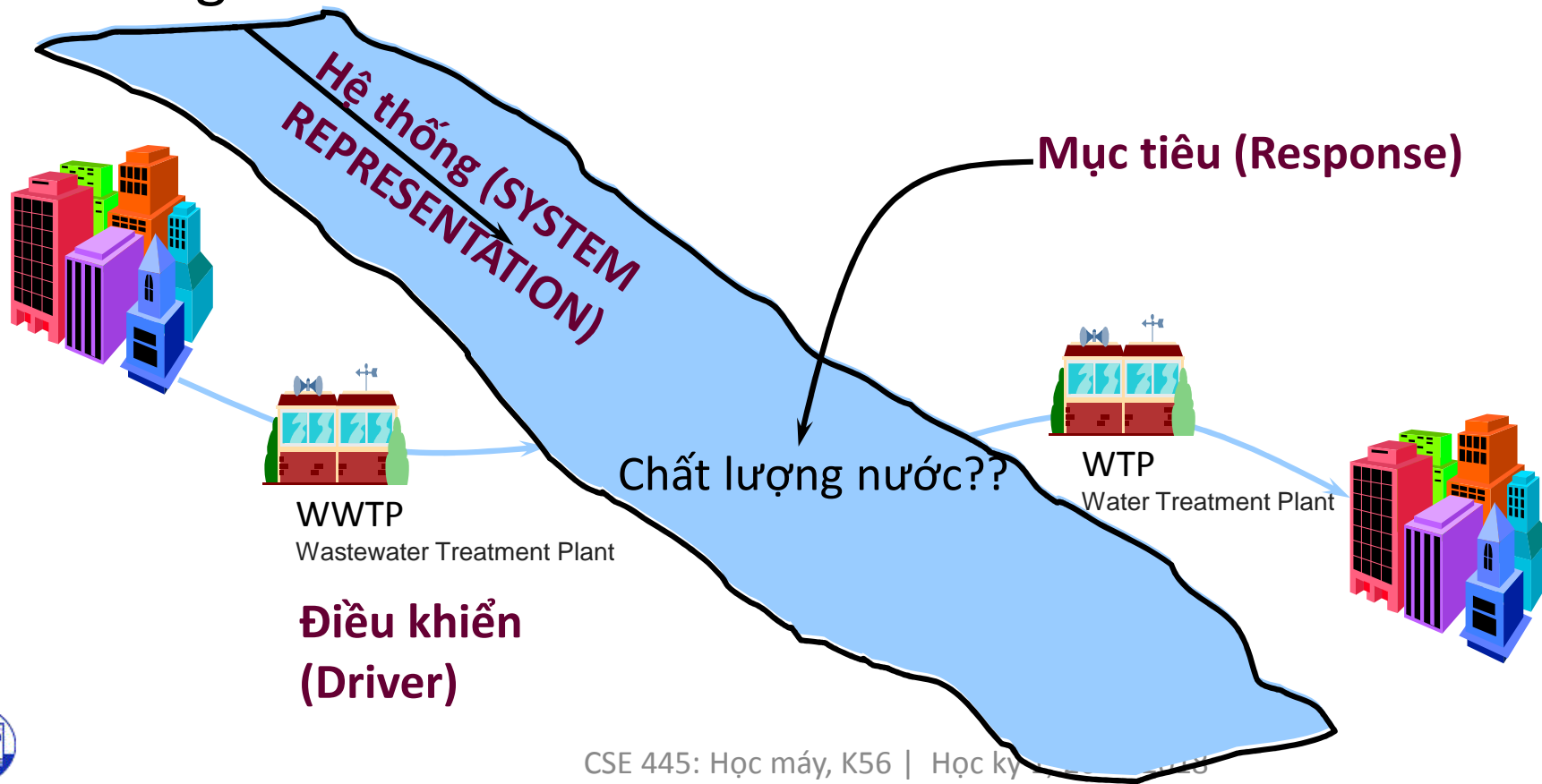
# Tại sao dùng kỹ thuật thống kê?

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



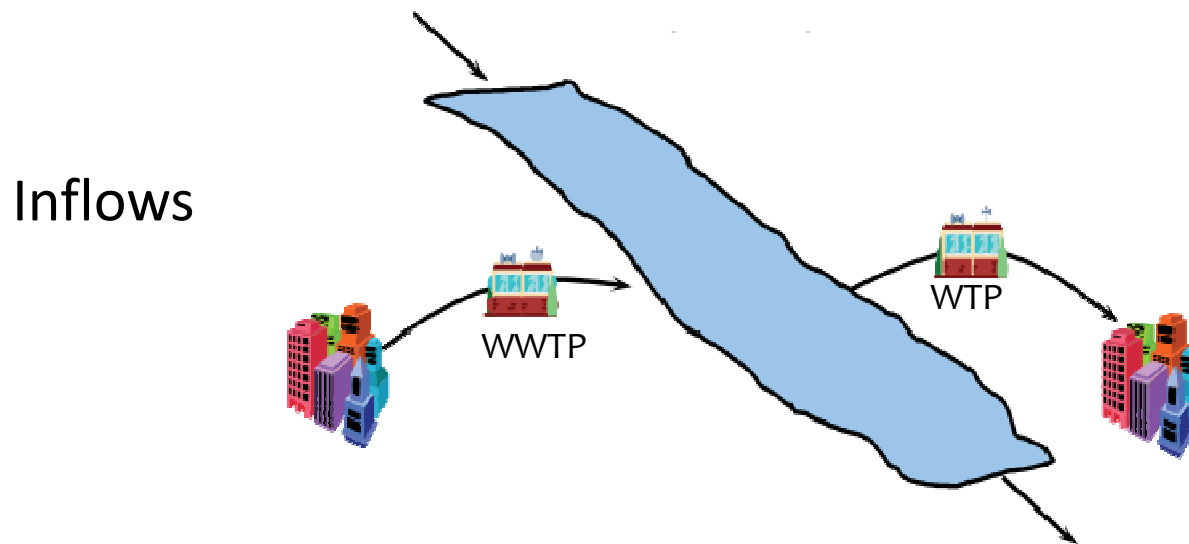
# Các thành phần của mô hình

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



# Các thành phần của mô hình

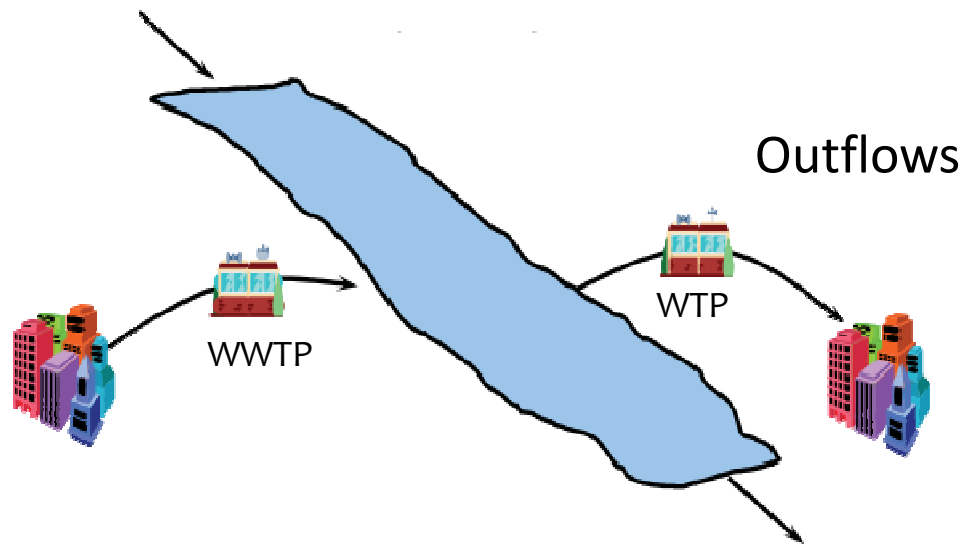
**Biến đầu vào:** Biến giúp xác định trạng thái của hệ thống thay đổi như thế nào (“Driver”)





# 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  $X_i$ .
  - 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).

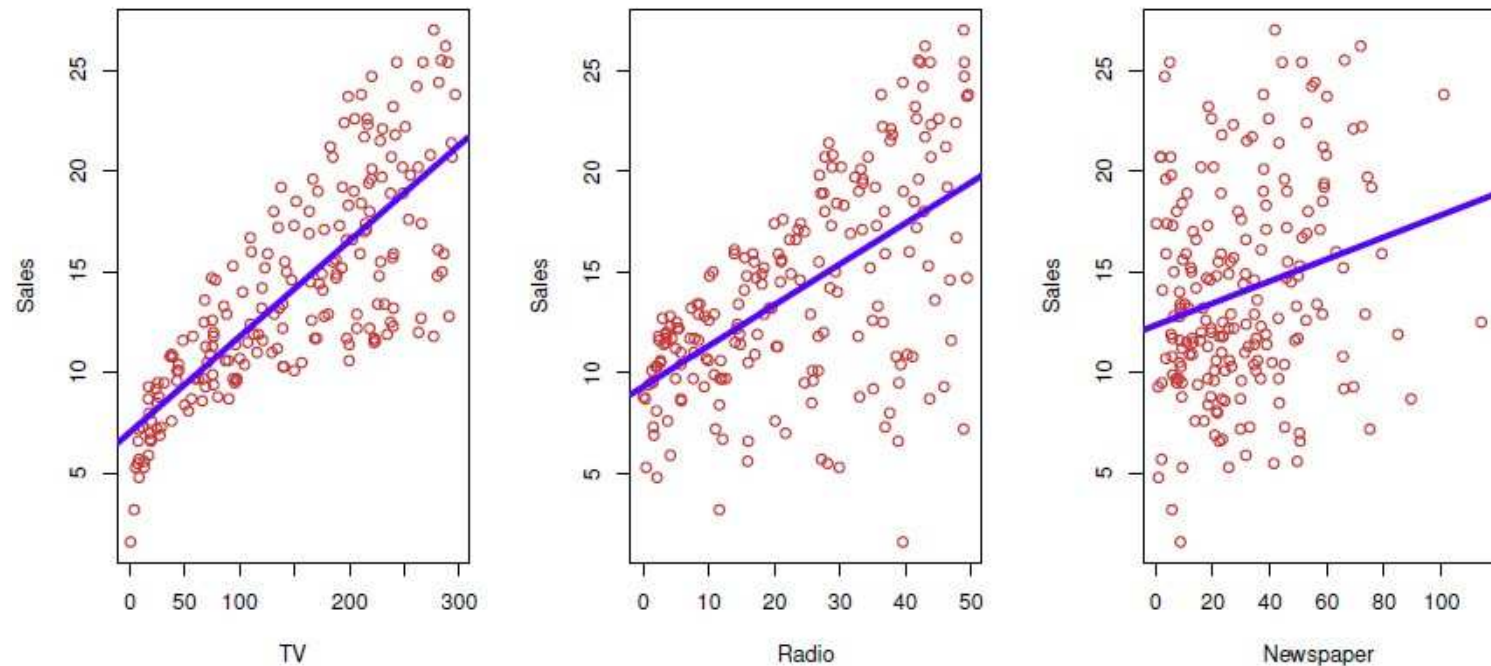


Figure 2.1 , ISL 2013



# Câu đố:

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





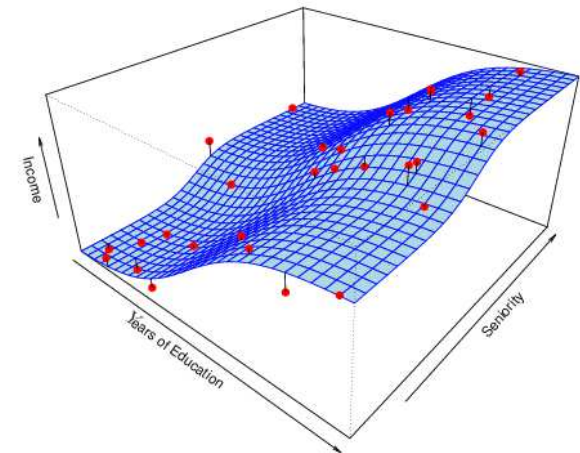
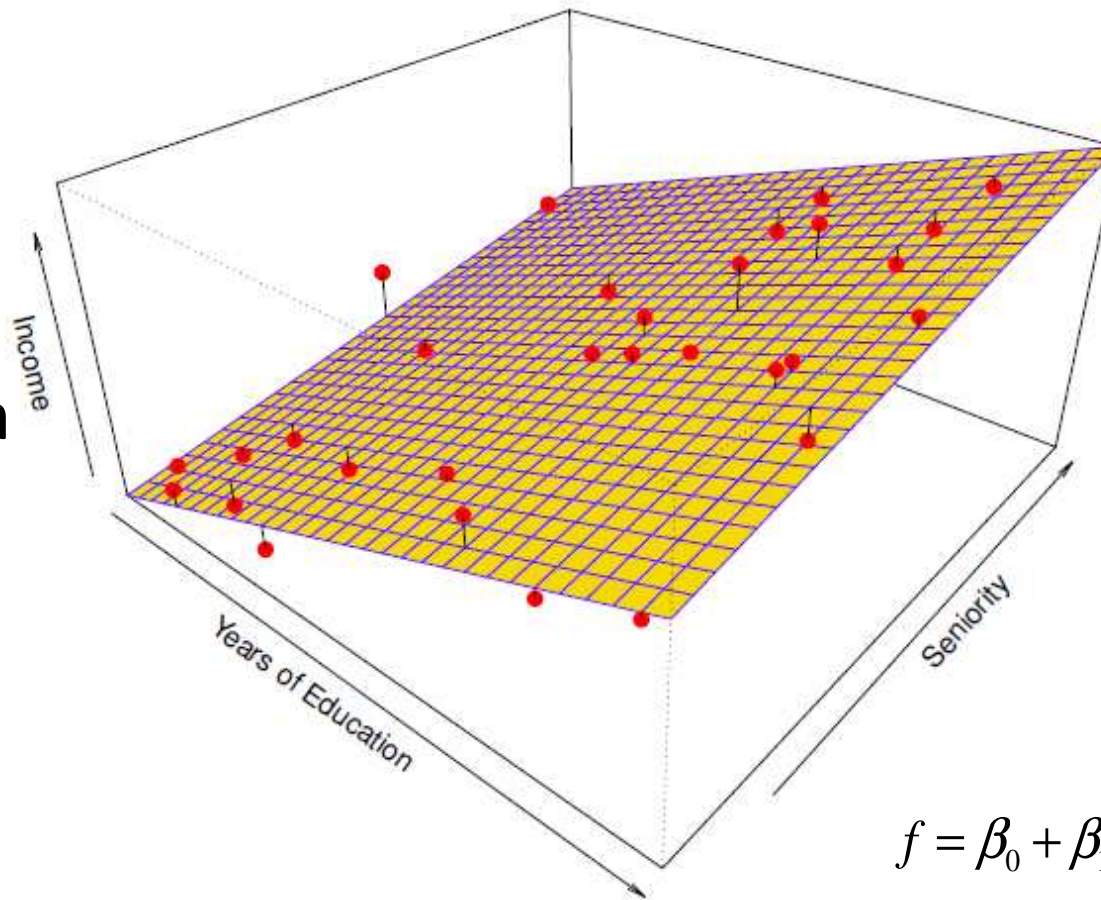
# Các mô hình tham số và phi tham số

- 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

- Mặc dù độ lệch chuẩn thấp nhưng ta vẫn nhận được đáp án tồi khi sử dụng sai mô hình.



$$f = \beta_0 + \beta_1 \times Education + \beta_2 \times Seniority$$

Figure 2.4 , ISL 2013



# 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 không bị *quá khớp/quá phù hợp (overfitting)* trên dữ liệu huấn luyện/tập học.
  - Ưu điểm:
    - Có thể xấp xỉ loạt 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



# Các mô hình tham số và phi tham số

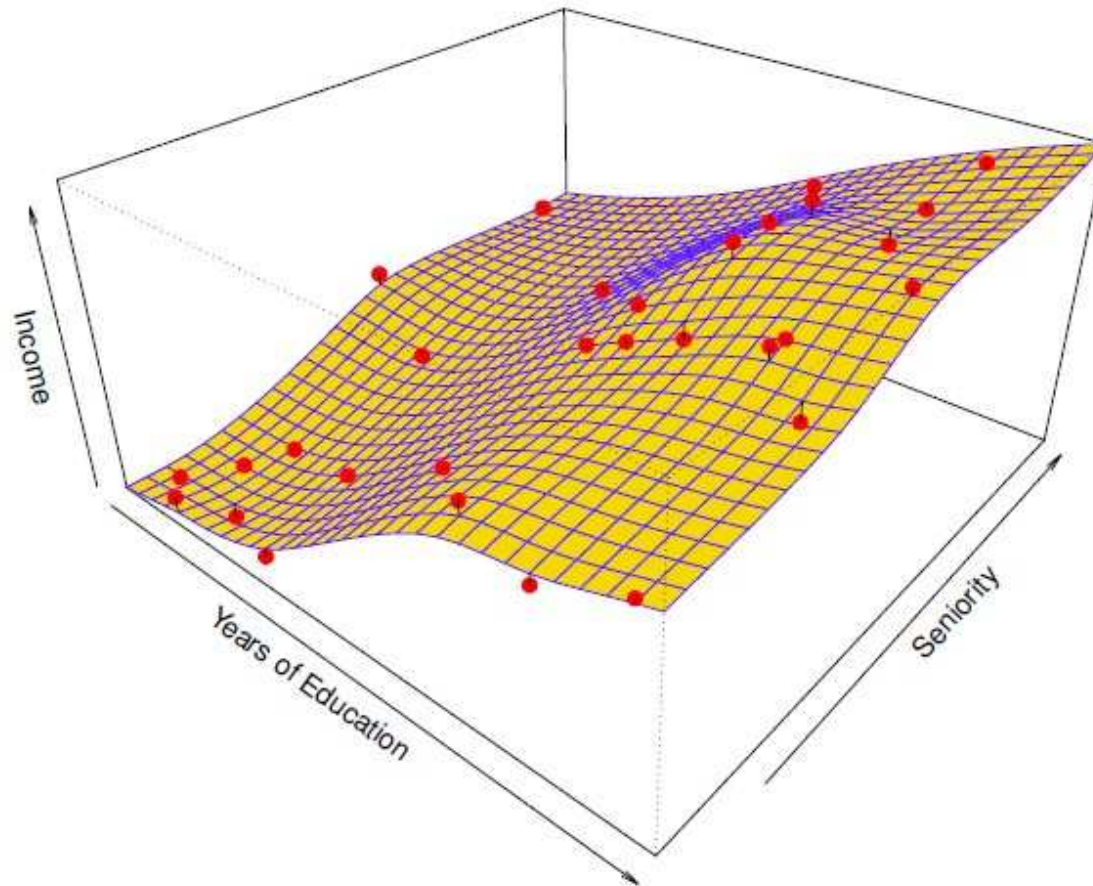


Figure 2.5 , ISL 2013



# Các mô hình tham số và phi tham số

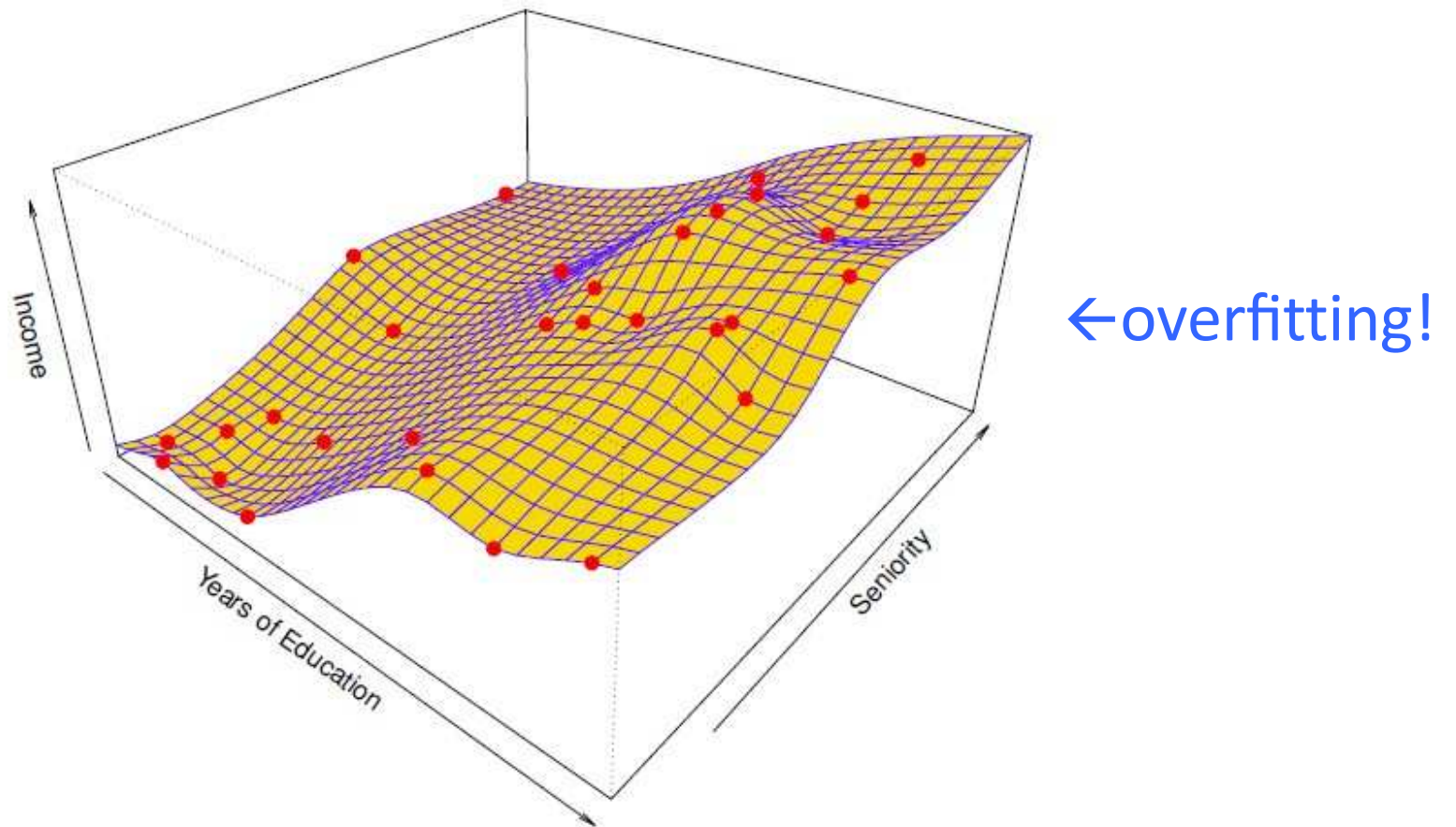


Figure 2.6 , ISL 2013



# 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
- 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

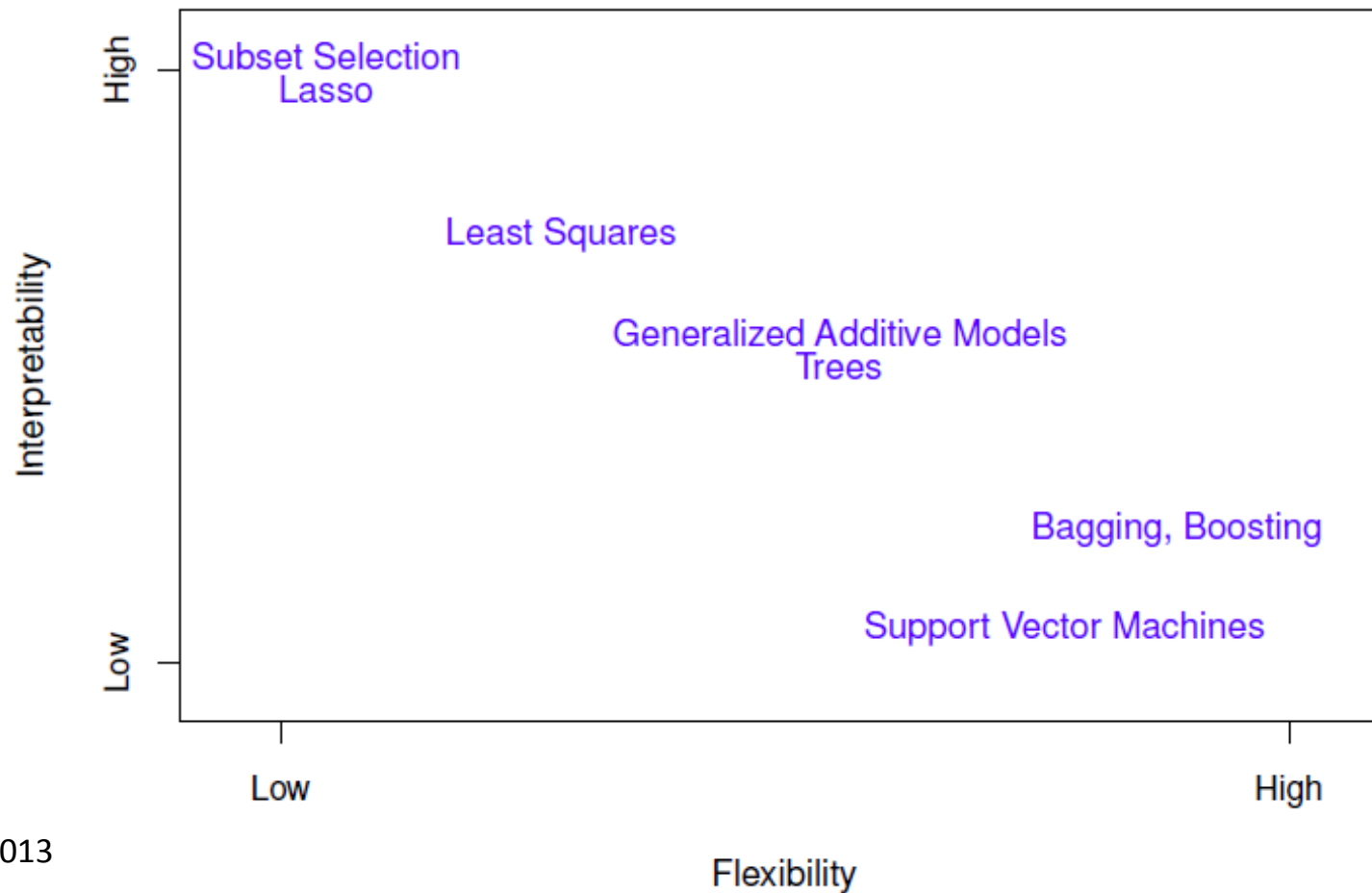


Figure 2.7 , ISL 2013



# Ngôn ngữ R





# 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 dạng giải thuật học máy



# Học có giám sát: Phân lớp và Hồi quy

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



# Học có giám sát: Phân lớp và Hồi quy

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

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



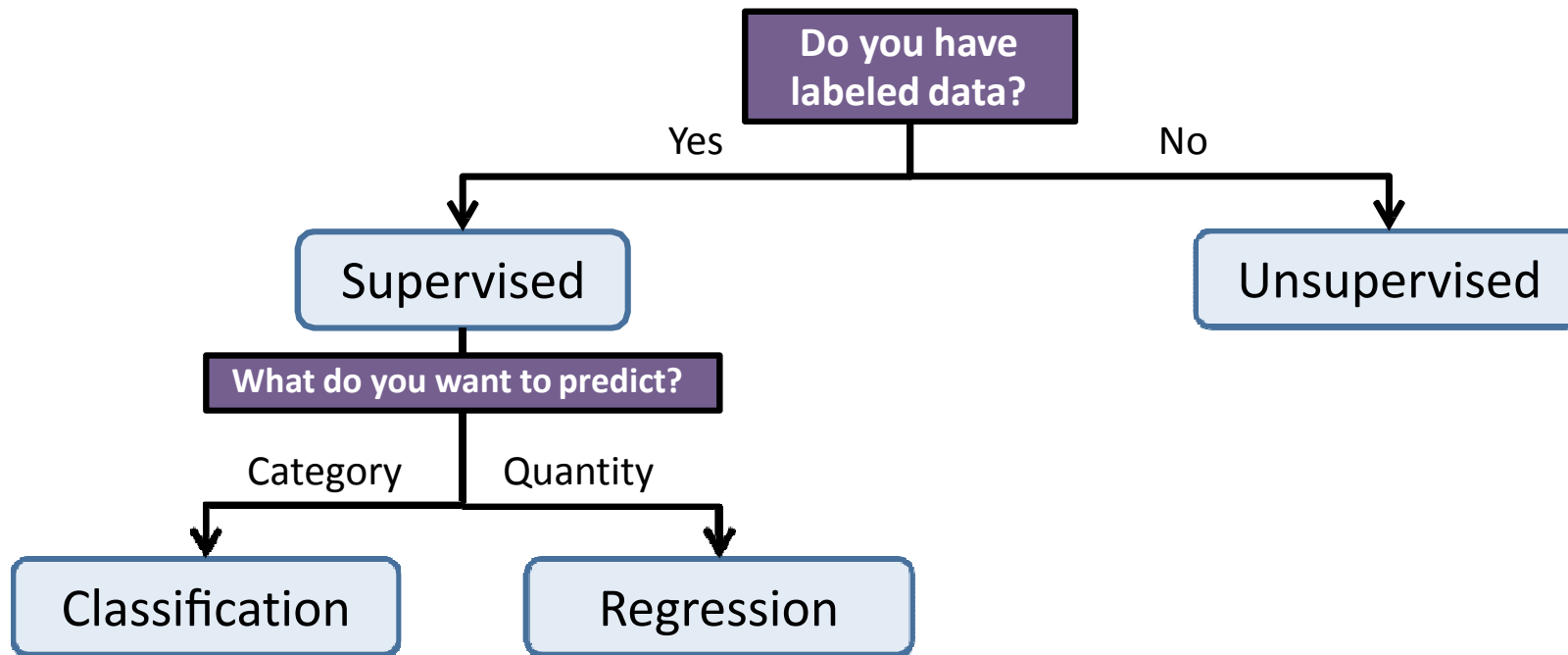
# Học có giám sát: Phân lớp và Hồi quy

- Bài toán phân lớp cũng có thể trình bày theo dạng hồi quy
  - Bài toán 2 lớp: “*Xác xuất để 1 quan sát/mẫu thuộc lớp 1?*”
  - Một số phương pháp học máy có thể xử lý được cả 2 dạng bài toán (vd 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 dạng giải thuật học máy

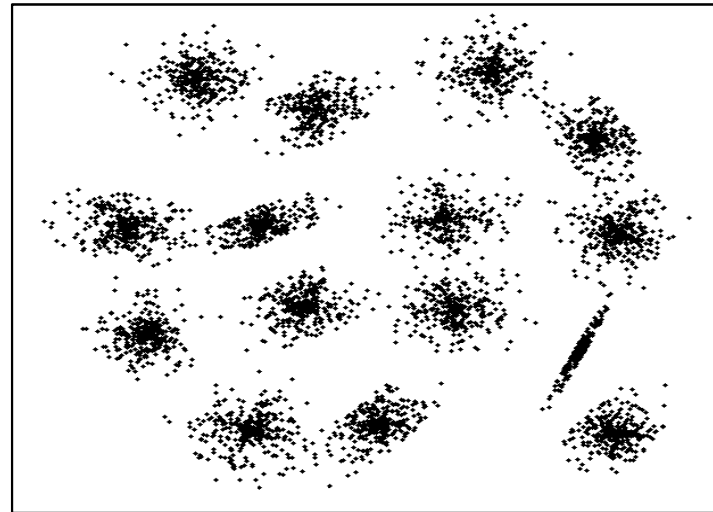


# Học máy không giám sát:

## Phân cụm & Giảm chiều dữ liệu

- *Phân tích cụm*

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

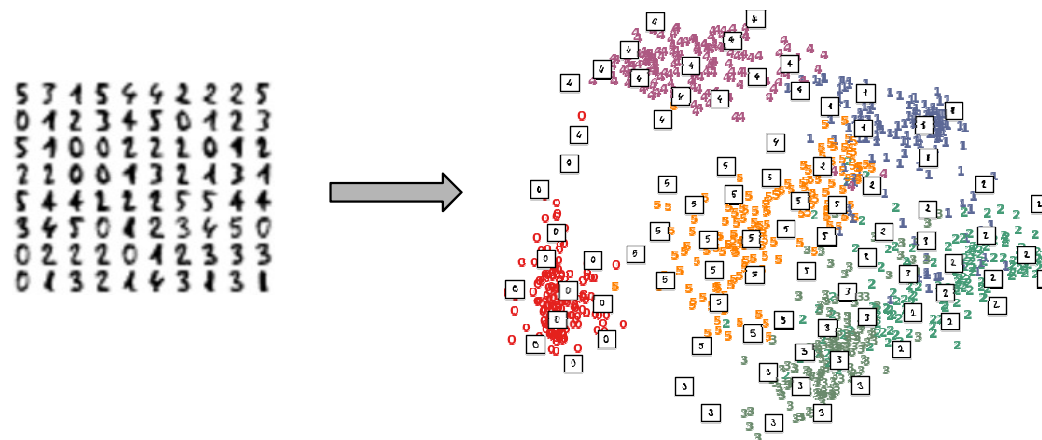


# Học máy không giám sát:

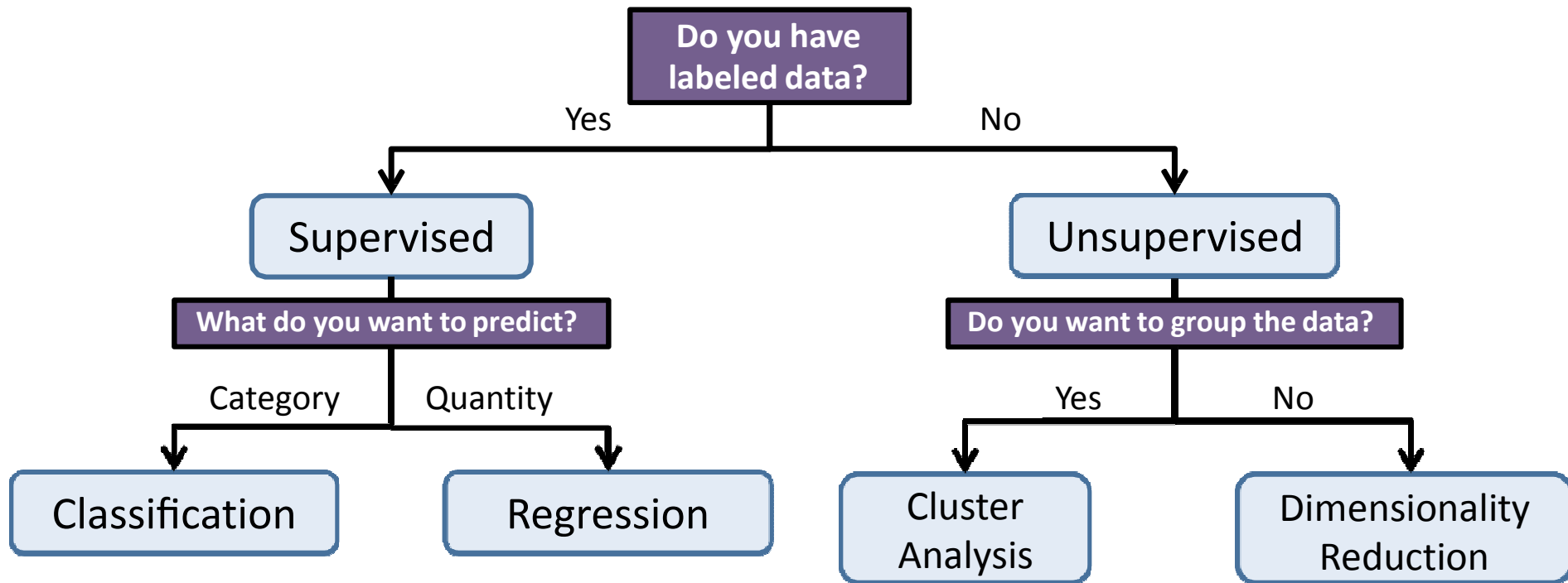
## Phân cụm & Giảm chiều dữ liệu

- *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 dạng giải thuật học máy

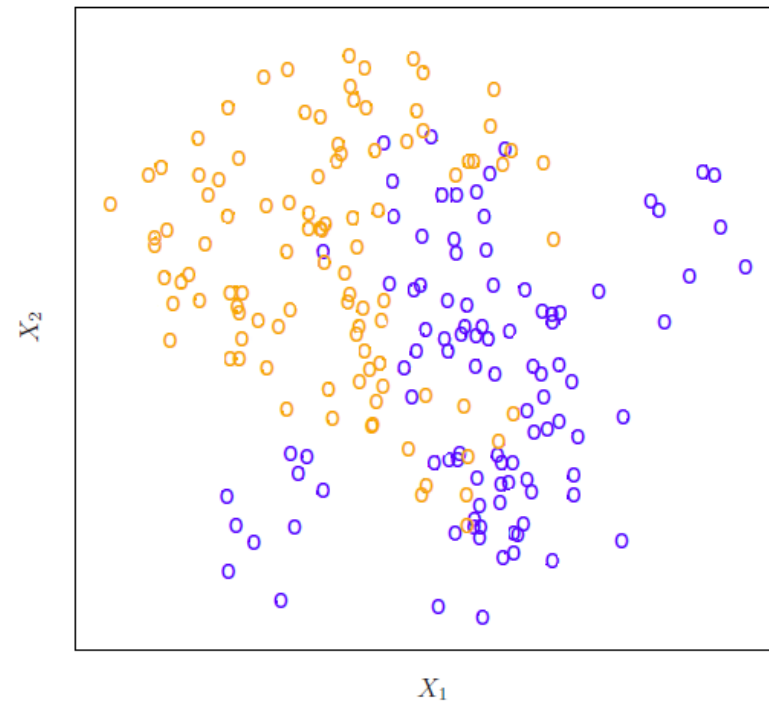


# Giải thuật phân lớp đơn giản



# Bộ phân lớp K-Nearest Neighbor (KNN)

- Ý tưởng: phân lớp các mẫu dựa trên “hàng xóm” các mẫu đã biết nhãn



# Bộ phân lớp K-láng giếng gần nhất

- Bộ phân lớp: Chia không gian thuộc tính thành nhiều vùng
  - Mỗi vùng được gán với 1 nhãn lớp (class label)
  - *Ranh giới quyết định* chia tách các vùng quyết định
- Các phương pháp phân lớp xây dựng mô hình có dạng:

$$Pr(Y | X)$$



# Bộ phân lớp K-láng giềng gần nhất

- Bộ phân lớp KNN

- Việc dự đoán lớp cho mẫu  $X$  là *lớp phổ biến nhất giữa  $K$  láng giềng gần nhất* (trong tập học)

- Mô hình phân lớp:

$$Pr(X \text{ belongs to class } Y) \approx \frac{\# (\text{neighbors of } X \text{ in class } Y)}{K}$$





# Bộ phân lớp K-láng giềng gần nhất

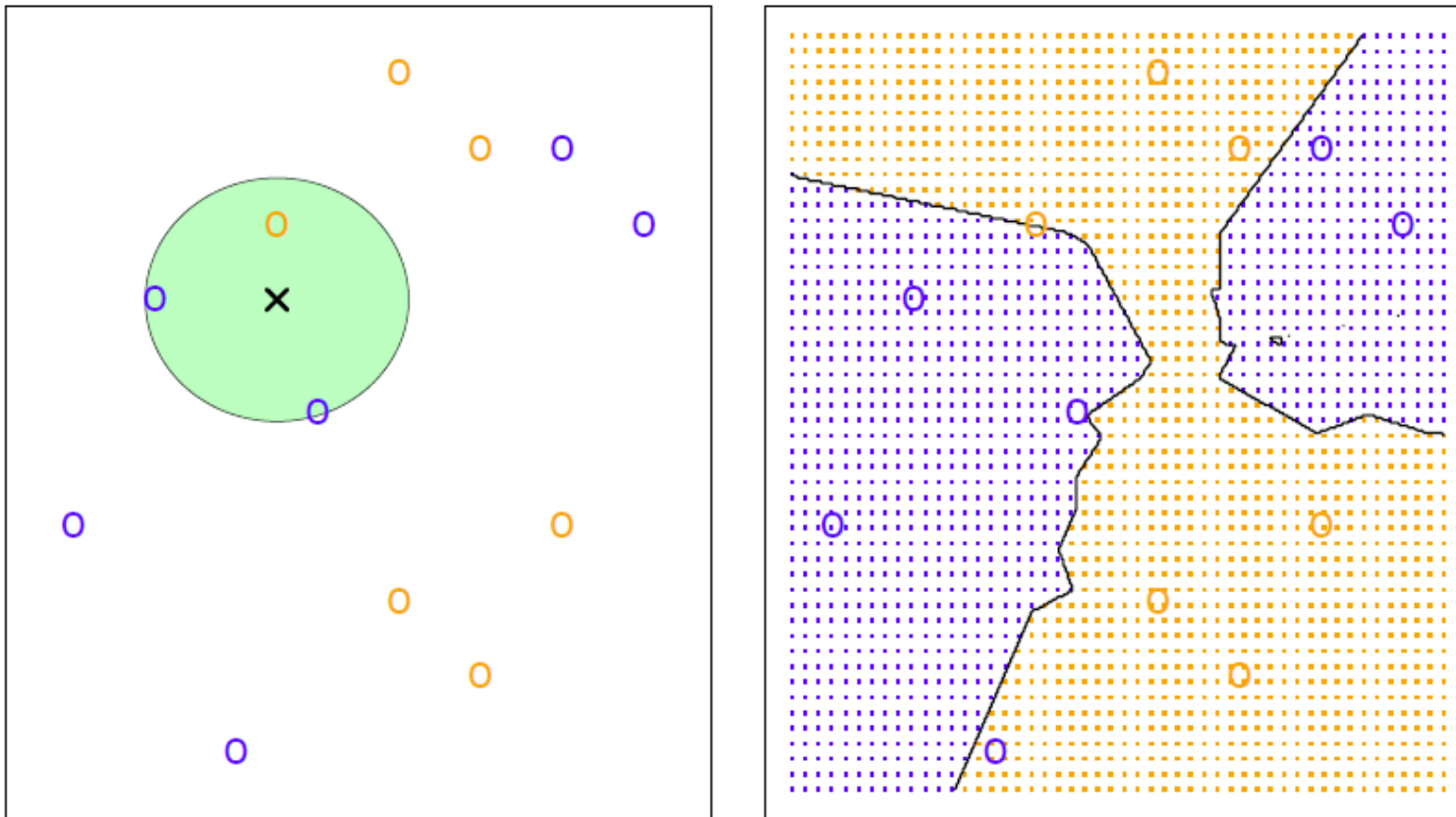
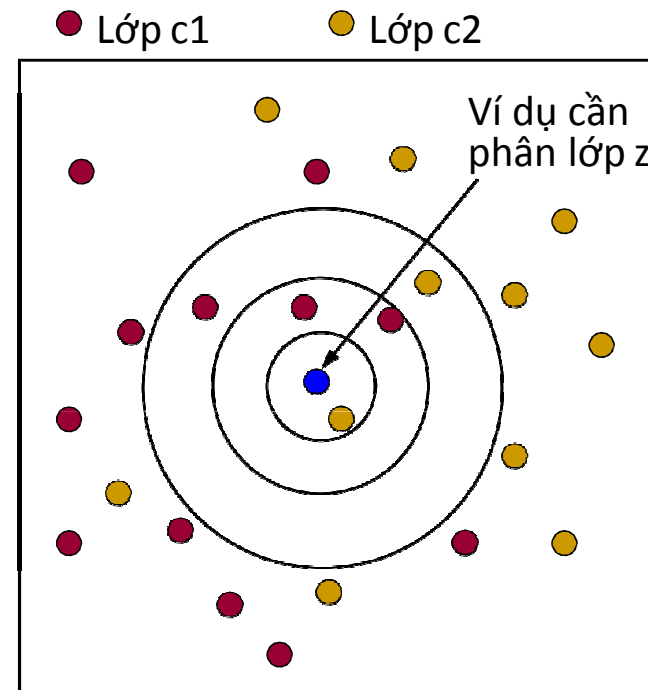


Figure 2.14, ISL 2013



# Bộ phân lớp K-láng giềng gần nhất

- Xét 1 láng giềng gần nhất  
→ Gán  $z$  vào lớp  $c2$
- Xét 3 láng giềng gần nhất  
→ Gán  $z$  vào lớp  $c1$
- Xét 5 láng giềng gần nhất  
→ Gán  $z$  vào lớp  $c1$



Nguồn hình vẽ: Học máy,  
Nguyễn Nhật Quang

## Ví dụ bài toán phân lớp



# Giải thuật phân lớp k-NN

- Giai đoạn huấn luyện (học)
  - Đơn giản là lưu lại các mẫu trong tập huấn luyện
- Giai đoạn phân lớp: Để phân lớp cho một mẫu (mới)  $z$ 
  - Với mỗi mẫu, tính khoảng cách giữa  $x$  và  $z$
  - Xác định tập  $NB(z)$  – các láng giềng gần nhất của  $z$ 
    - Gồm  $k$  mẫu trong tập huấn luyện gần nhất với  $z$  tính theo một hàm khoảng cách  $d$
  - Phân  $z$  vào lớp chiếm số đông (the majority class) trong số các lớp của các mẫu trong  $NB(z)$



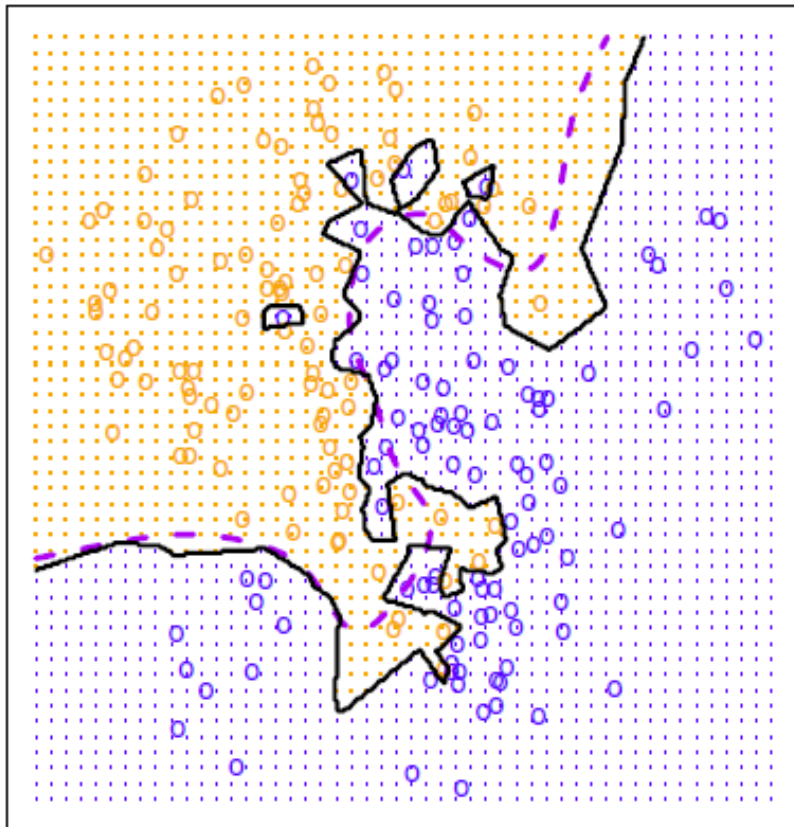
# Lựa chọn K (bộ phân lớp KNN)

- K nhỏ
  - Ranh giới quyết định linh hoạt hơn, tuy nhiên dễ bị *overfit*
- K lớn
  - Ranh giới quyết định ít linh hoạt nhưng ít bị *overfit*
- *Overfitting*: Cho kết quả tốt trên tập học nhưng kém trên tập thử nghiệm



# Lựa chọn K (bộ phân lớp KNN)

KNN: K=1



KNN: K=100

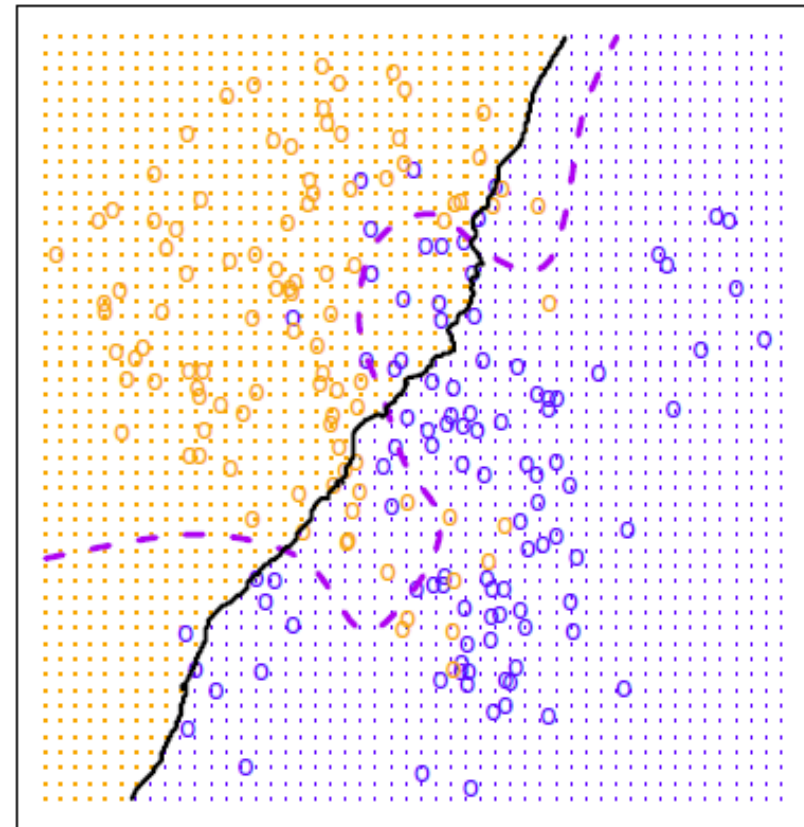


Figure 2.16,  
ISL 2013



# Lựa chọn K (bộ phân lớp KNN)

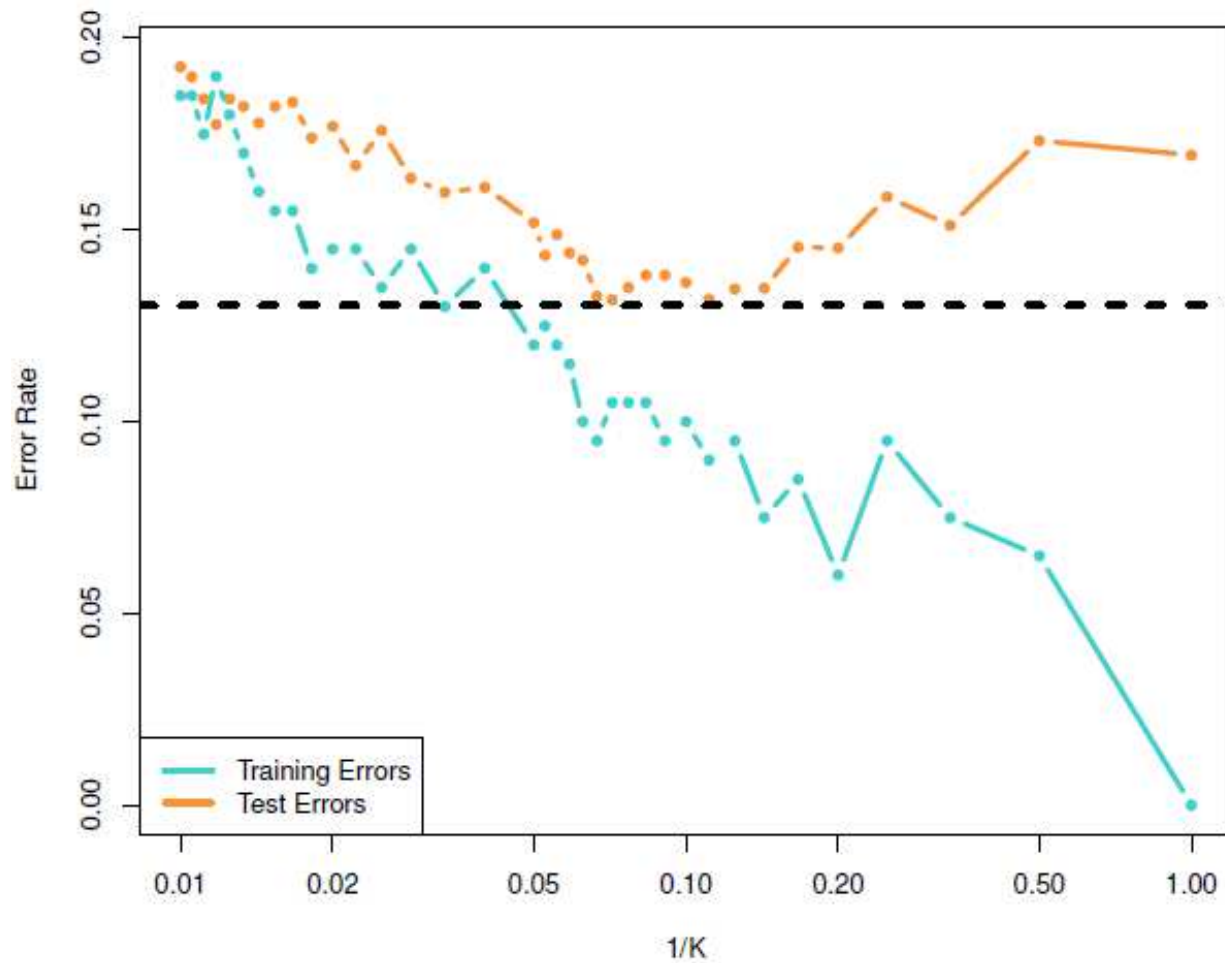


Figure 2.17, ISL 2013



# Lựa chọn K (bộ phân lớp KNN)

KNN: K=10

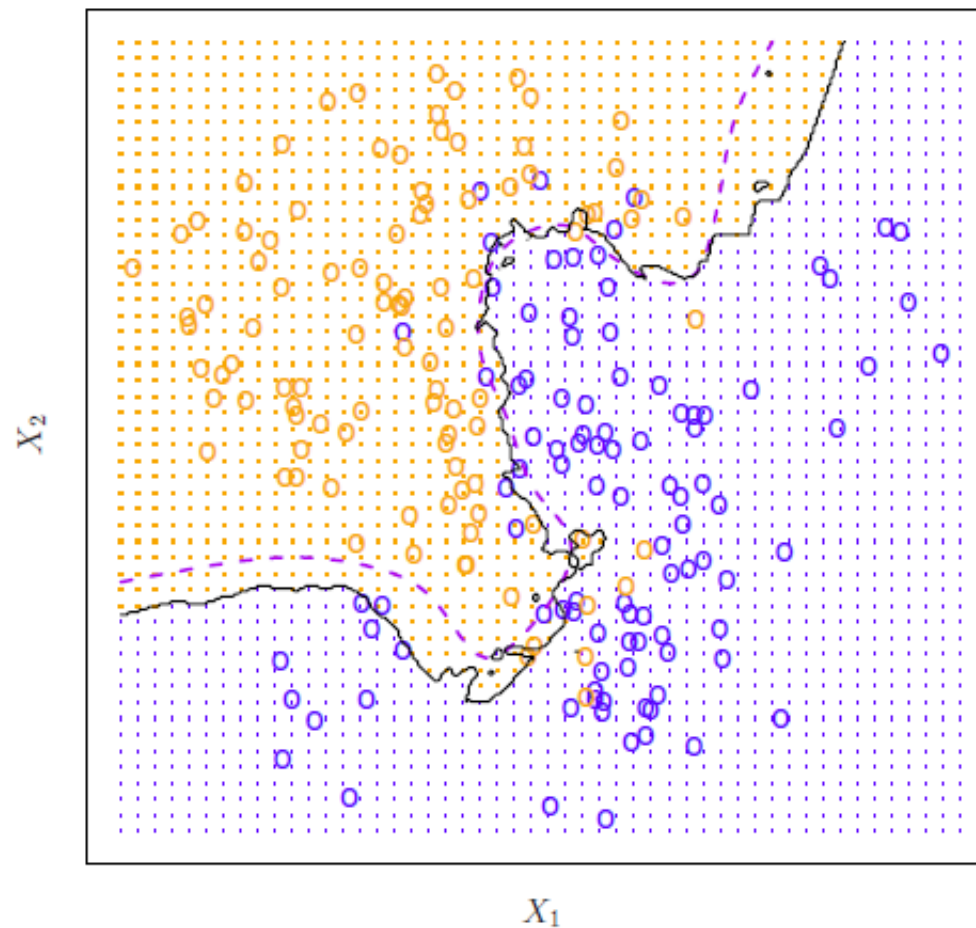


Figure 2.15, ISL 2013



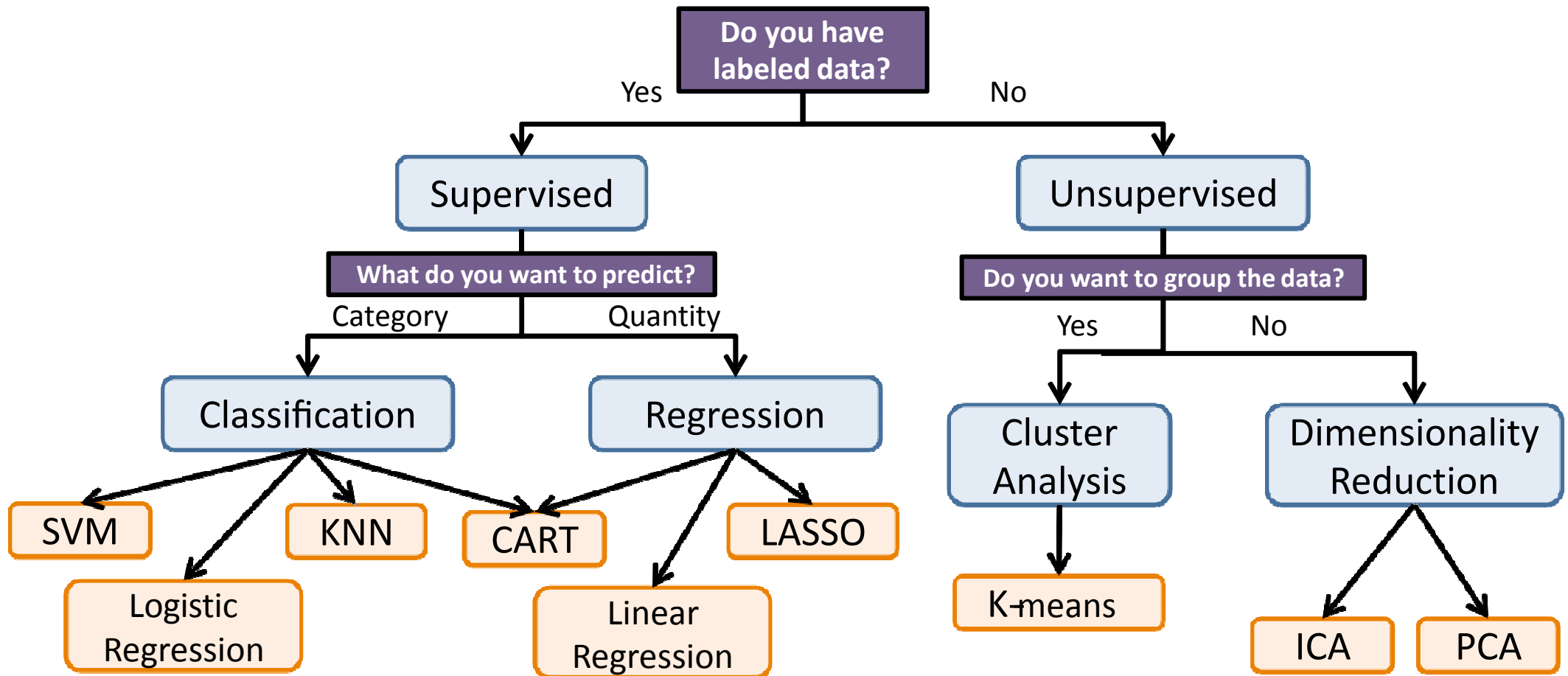
# K-Nearest Neighbor classifier (KNN)

- Ưu điểm:
  - Dễ cài đặt
  - Ít tham số mô hình ( $K$ , distance metric)
  - Linh hoạt, các lớp không phải tách tuyến tính
- Nhược điểm:
  - Thời gian tính toán lâu
  - Khá nhạy với dữ liệu không cân bằng
  - Nhạy với dữ liệu đầu vào không liên quan với nhau





# Các dạng 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)
- 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



# Các câu hỏi?

