

Học Máy

(Machine Learning)

Thân Quang Khoát

khoattq@soict.hust.edu.vn

Viện Công nghệ thông tin và Truyền thông
Trường Đại Học Bách Khoa Hà Nội

Năm 2017

Nội dung môn học:

- **Giới thiệu chung**
- Các phương pháp học không giám sát
- Các phương pháp học có giám sát
- Đánh giá hiệu năng hệ thống học máy

Tại sao nên biết Học Máy?

- ❖ Nhu cầu lớn về Khoa học dữ liệu (Data Science)
- ❖ “Data scientist: the sexiest job of the 21st century” – Harvard Business Review.
<http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/>
- ❖ “The Age of Big Data” – The New York Times
http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0



Data Analyst

San Francisco Bay Area

Posted 18 days ago

PIMCO

Data Analyst

Greater New York City Area

Posted 25 days ago

nielsen

Statistical Analyst - Data...

Greater New York City Area

Posted 9 hours ago

Data Analyst

Greater New York City Area

Posted 15 days ago

DATA SCIENTIST

Greater New York City

Posted 25 days ago

Quirky

FORA
FINANCIAL

H

fire

healthfirst

J.P.Morgan

Home

Profile

Network

Jobs

Interests



Data Analyst

Amazon - Newark, NJ

Posted 24 days ago

[Apply on company website](#)

Save

Senior Data Analyst - Big Data, Meta Product

TripAdvisor - Newton, MA

Posted 12 days ago



Home

Profile

Network

Jobs

Interests



Data Analyst

Apple - Daly City - California -US

Posted 18 days ago

[Apply on company website](#)

Save

Tại sao nên biết Học Máy?

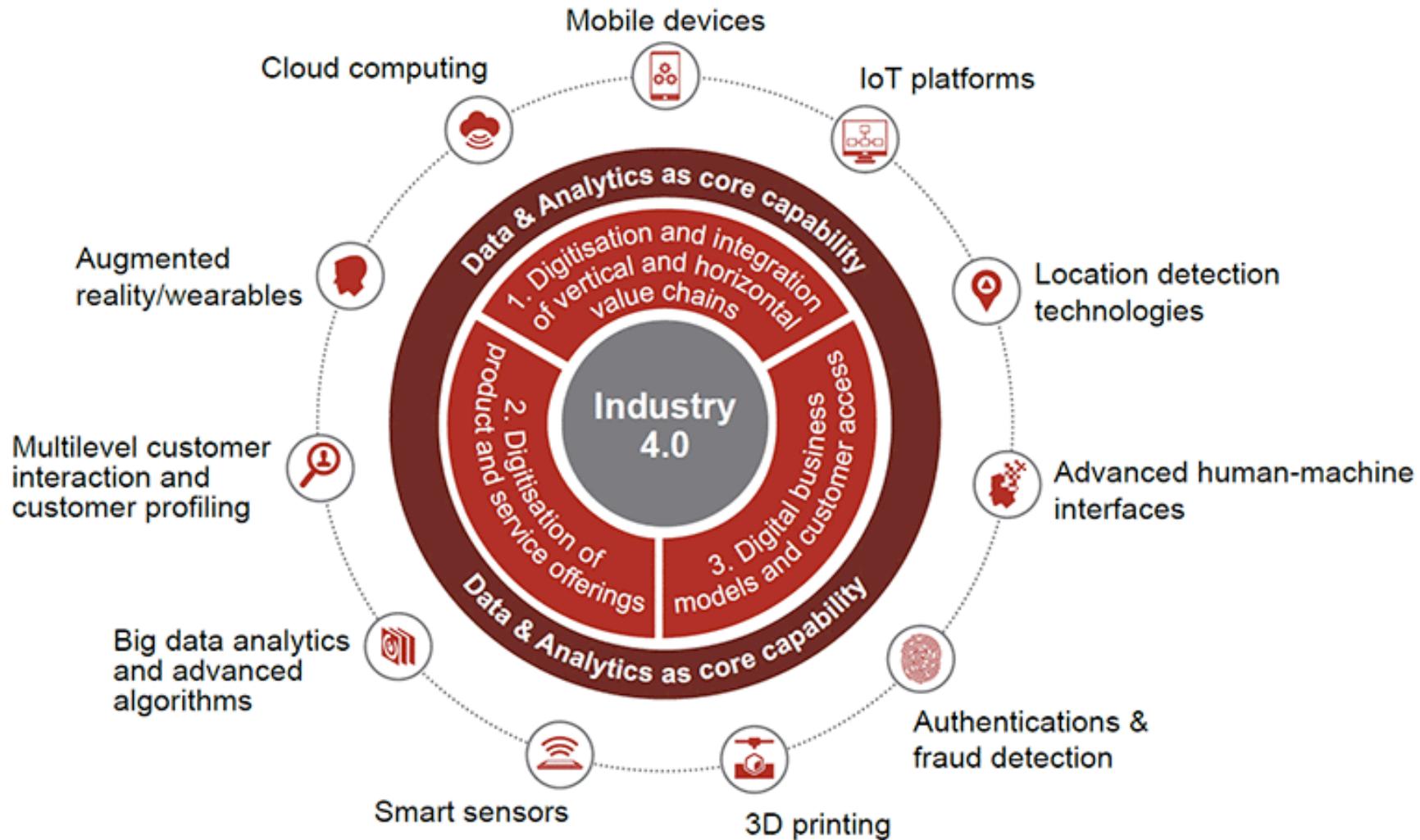
- ❖ Nhu cầu ngày càng tăng tại Việt Nam.



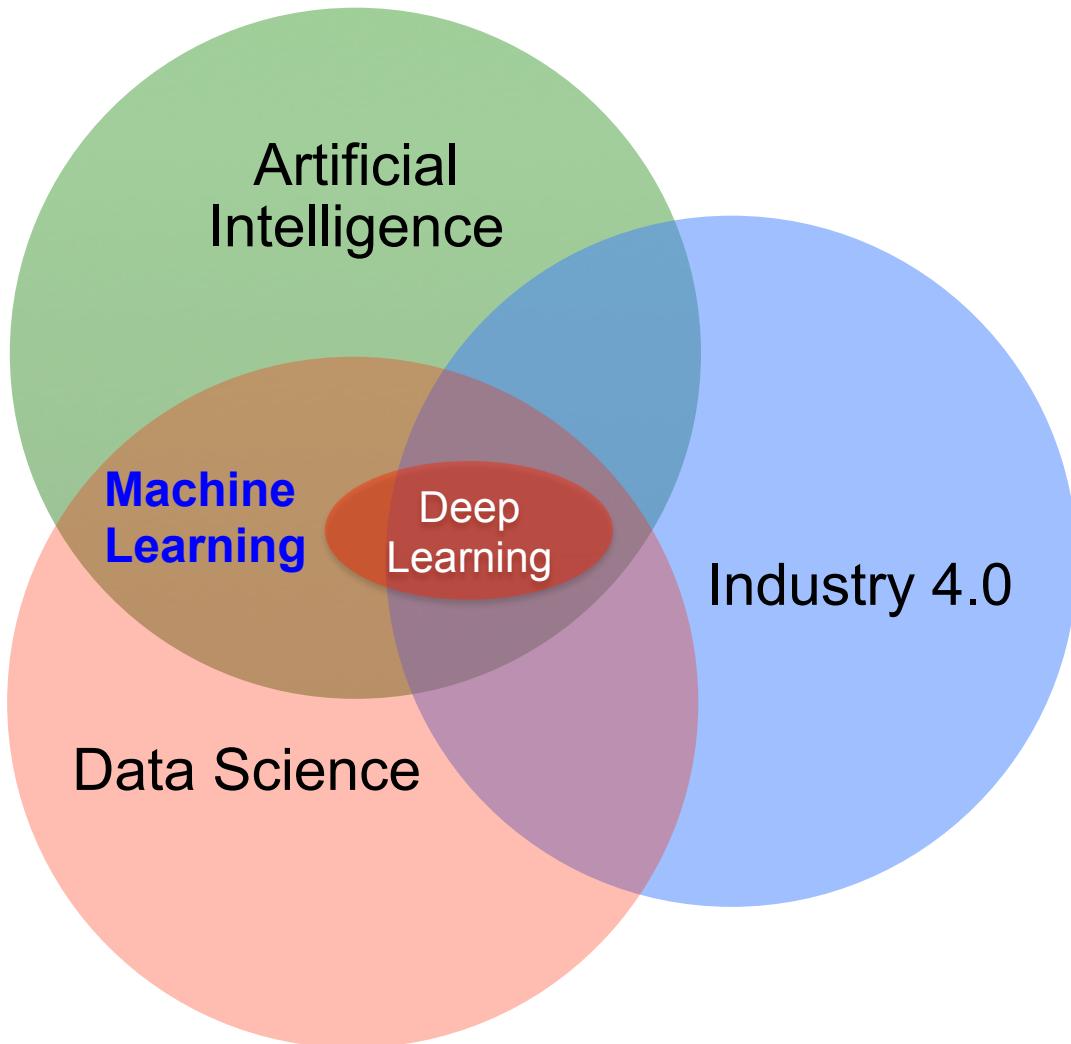
Hãy nói theo cách của bạn



Tại sao? Cách mạng công nghiệp 4.0



Tại sao? AI & DS & Industry 4.0

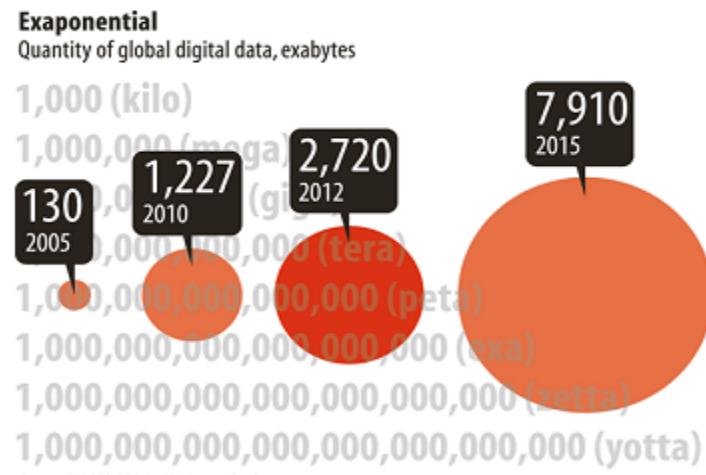


Tại sao nên biết Học Máy?

- ❖ Học máy (ML – Machine Learning):
data mining, inference, prediction.
- ❖ ML là con đường hiệu quả để tạo ra các hệ thống thông minh, dịch vụ thông minh.
- ❖ ML cung cấp nền tảng và phương pháp cho Big Data.



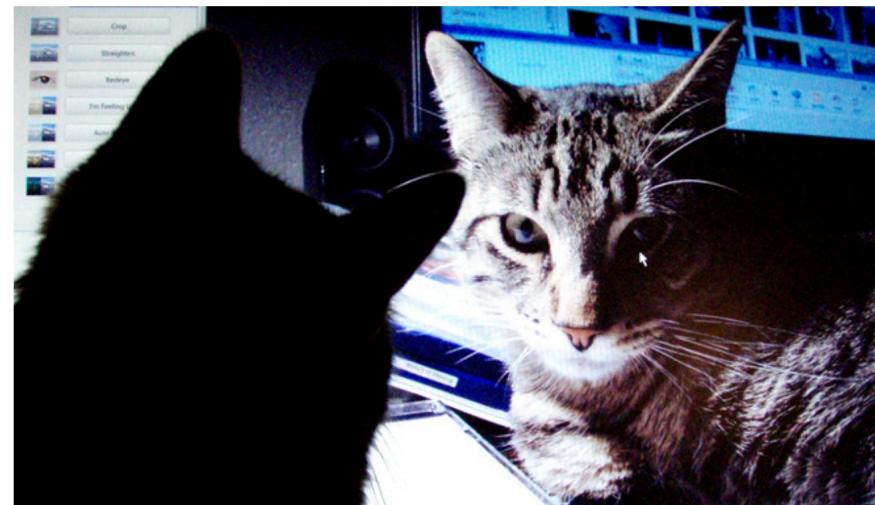
Each day:
230M tweets,
2.7B comments to FB,
86400 hours of video
to YouTube



Vài thành công: GoogleBrain (2012)

Google's Artificial Brain Learns to Find Cat Videos

BY WIRED UK 06.26.12 | 11:15 AM | PERMALINK



By Liat Clark, Wired UK

How Many Computers to Identify a Cat? 16,000



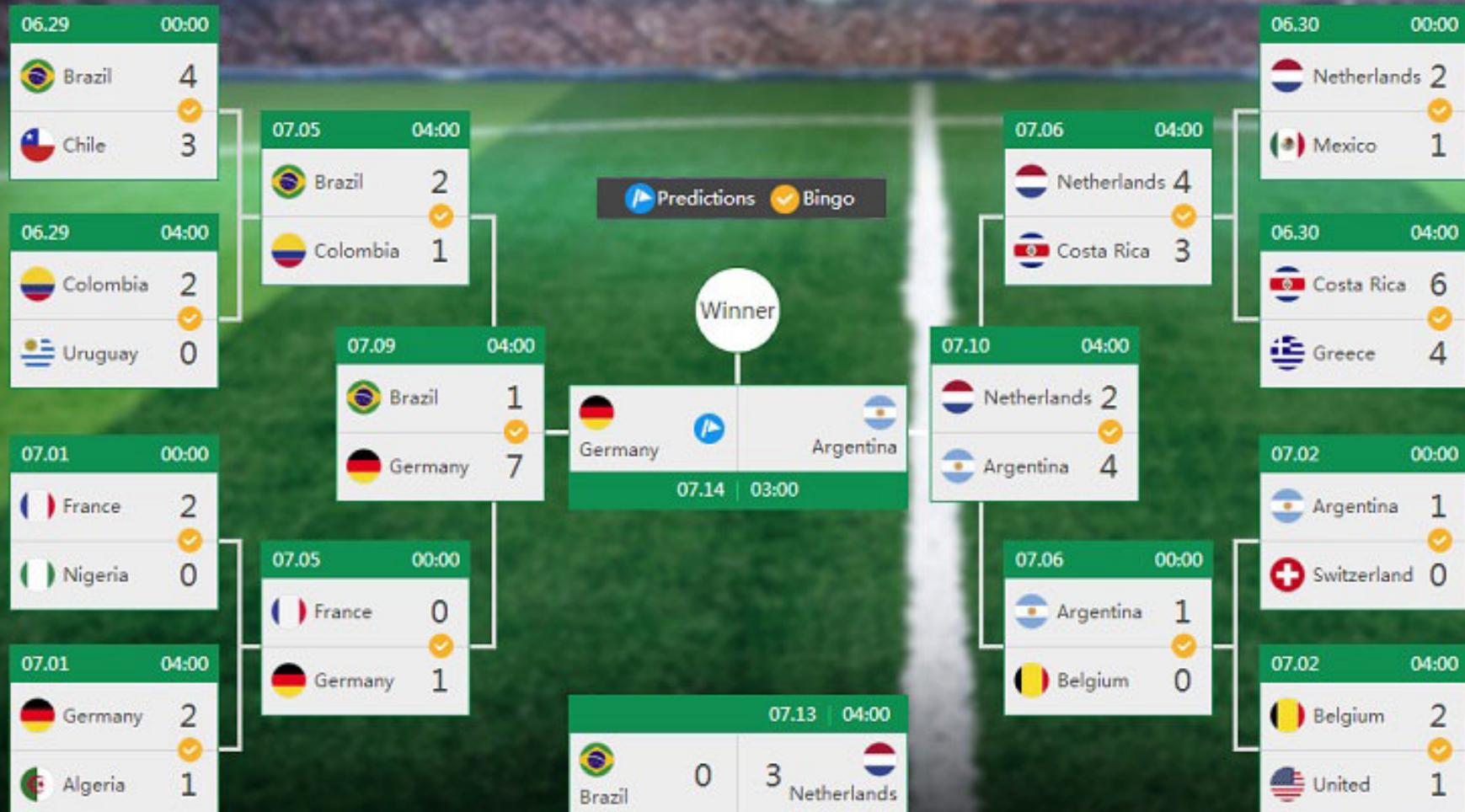
Jim Wilson/The New York Times

An image of a cat that a neural network taught itself to recognize.

By JOHN MARKOFF

Published: June 25, 2012

Vài thành công: FIFA prediction (2014)



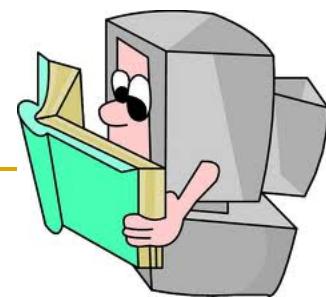
Vài thành công: AlphaGo (2016)

- AlphaGo of Google the world champion at Go (cờ vây), 3/2016
 - Go is a 2500 year-old game.
 - Go is one of the most complex games.
- AlphaGo learns from 30 millions human moves, and plays itself to find new moves.
- It beat Lee Sedol (World champion)
 - <http://www.wired.com/2016/03/two-redefined-future/>
 - <http://www.nature.com/news/google-game-of-go-1.19234>



Giới thiệu về Học máy

- Học máy (ML - Machine Learning) là một lĩnh vực nghiên cứu của Trí tuệ nhân tạo (Artificial Intelligence)
- Câu hỏi trung tâm của ML:
 - “*How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?*” [Mitchell, 2006]
- Vài quan điểm về học máy:
 - Một quá trình nhờ đó một hệ thống cải thiện hiệu suất (hiệu quả hoạt động) của nó [Simon, 1983]
 - Việc lập trình các máy tính để tối ưu hóa một tiêu chí hiệu suất dựa trên các dữ liệu hoặc kinh nghiệm trong quá khứ [Alpaydin, 2010]



Một máy học

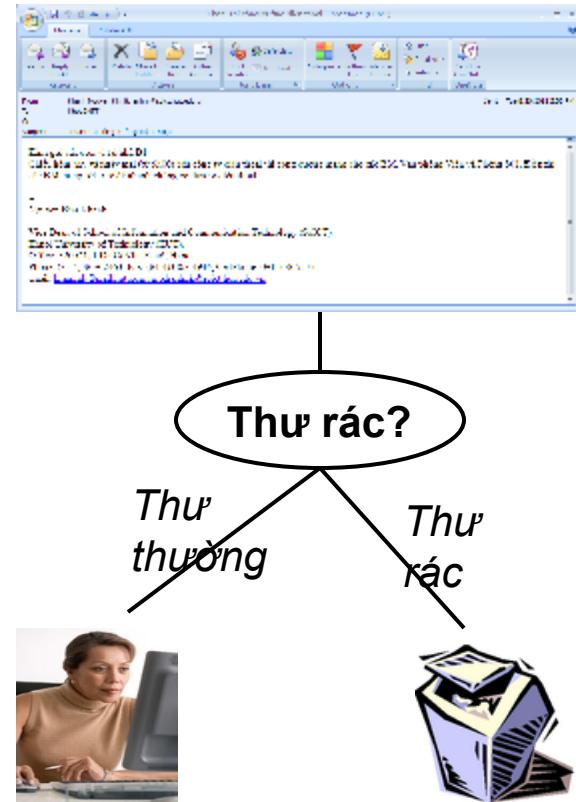
- Ta nói một máy tính *có khả năng học* nếu nó tự cải thiện hiệu suất hoạt động *P* cho một công việc *T* cụ thể, dựa vào kinh nghiệm *E* của nó.
- Như vậy *một bài toán học máy* có thể biểu diễn bằng 1 bộ (*T*, *P*, *E*)
 - *T*: một công việc (nhiệm vụ)
 - *P*: tiêu chí đánh giá hiệu năng
 - *E*: kinh nghiệm



Ví dụ bài toán học máy (1)

Lọc thư rác (email spam filtering)

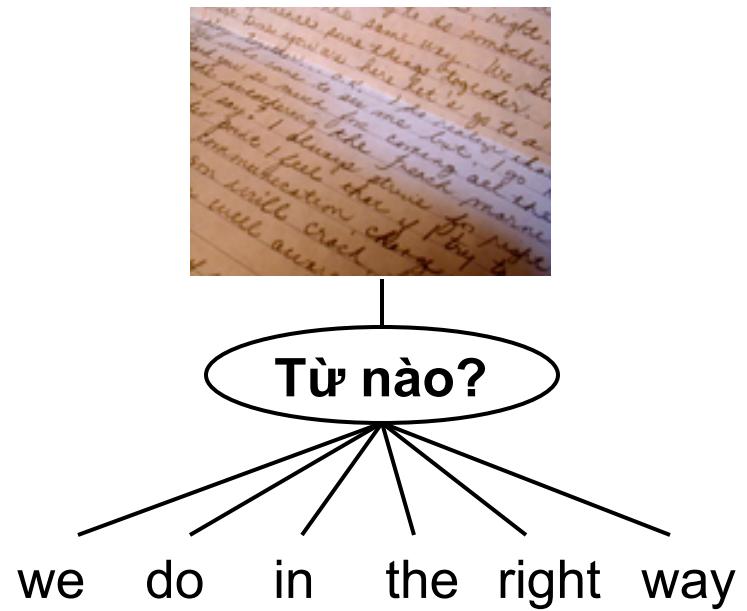
- T : Dự đoán (để lọc) những thư điện tử nào là thư rác (spam email)
- P : số lượng thư điện tử gửi đến được phân loại chính xác
- E : Một tập các thư điện tử (emails) mẫu, mỗi thư điện tử được biểu diễn bằng một tập thuộc tính (vd: tập từ khóa) và nhãn lớp (thư thường/thư rác) tương ứng



Ví dụ bài toán học máy (2)

Nhận dạng chữ viết tay

- **T**: Nhận dạng và phân loại các từ trong các ảnh chữ viết
- **P**: Tỷ lệ (%) các từ được nhận dạng và phân loại đúng
- **E**: Một tập các ảnh chữ viết, trong đó mỗi ảnh được gắn với một định danh của một từ



Ví dụ bài toán học máy (3)

Gán nhãn ảnh

- **T:** đưa ra một vài mô tả ý nghĩa của 1 bức ảnh
- **P:** ?
- **E:** Một tập các bức ảnh, trong đó mỗi ảnh đã được gán một tập các từ mô tả ý nghĩa của chúng



FISH WATER OCEAN
TREE CORAL



PEOPLE MARKET PATTERN
TEXTILE DISPLAY



BIRDS NEST TREE
BRANCH LEAVES

Máy học (1)

■ Học một ánh xạ (hàm):

$$f : x \mapsto y$$

- x : quan sát (dữ liệu), kinh nghiệm
- y : phán đoán, tri thức mới, kinh nghiệm mới, ...

■ Hồi quy (regression): nếu y là một số thực

■ Phân loại (classification): nếu y thuộc một tập rời rạc (tập nhãn lớp)

Anh ta thích nghe



+



→ Trẻ hay Già ?

Máy học (2)

■ Học từ đâu?

- Từ các quan sát trong quá khứ (**tập học – training set**).
 $\{\{x_1, x_2, \dots, x_N\}; \{y_1, y_2, \dots, y_M\}\}$
- x_i là các quan sát của x trong quá khứ
- y_h là *nhãn (label)* hoặc *phản hồi (response)* hoặc *đầu ra (output)*

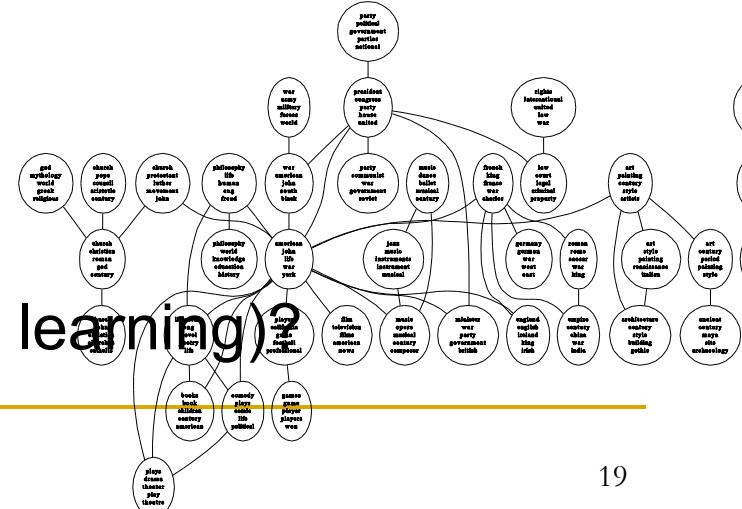
■ Sau khi đã học:

- Thu được một mô hình, kinh nghiệm, tri thức mới (f).
- Dùng nó để **suy diễn (infer)** hoặc **phán đoán (predict)** cho quan sát trong tương lai.

$$y_z = f(z)$$

Hai bài toán học cơ bản

- **Học có giám sát (supervised learning):** cần học một hàm $y = f(x)$ từ tập học $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ sao cho $y_i \approx f(x_i)$.
 - *Phân loại* (phân lớp): nếu y chỉ nhận giá trị từ một tập rời rạc, chẳng hạn {cá, cây, quả, mèo}
 - *Hồi quy*: nếu y nhận giá trị số thực
- **Học không giám sát (unsupervised learning):** cần học một hàm $y = f(x)$ từ tập học chưa trước $\{x_1, x_2, \dots, x_N\}$.
 - Y có thể là các cụm dữ liệu.
 - Y có thể là các cấu trúc ẩn.
- **Học bán giám sát (semi-supervised learning)**?



Học có giám sát: ví dụ

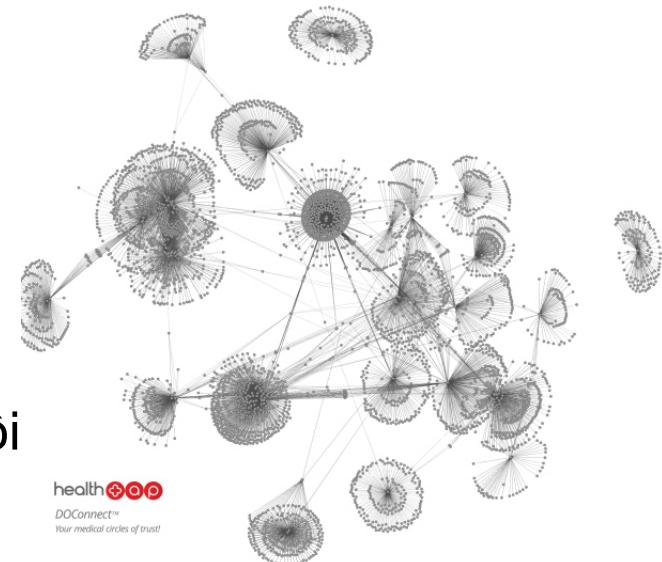
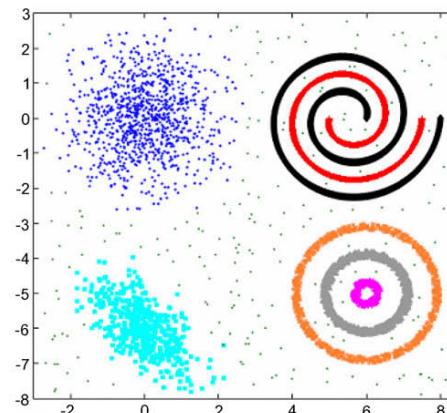
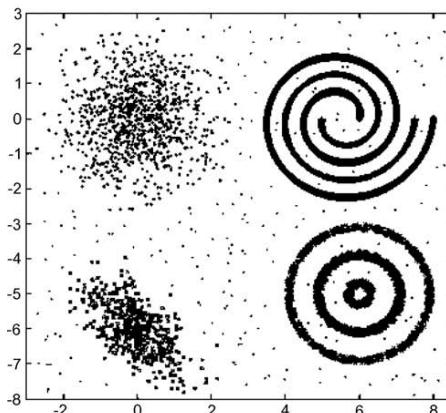
- Lọc thư rác
- Phân loại trang web
- Dự đoán rủi ro tài chính
- Dự đoán biến động chỉ số chứng khoán
- Phát hiện tấn công mạng



Học không giám sát: ví dụ (1)

■ Phân cụm (clustering)

- Phát hiện các cụm dữ liệu, cụm tính chất,...



■ Community detection

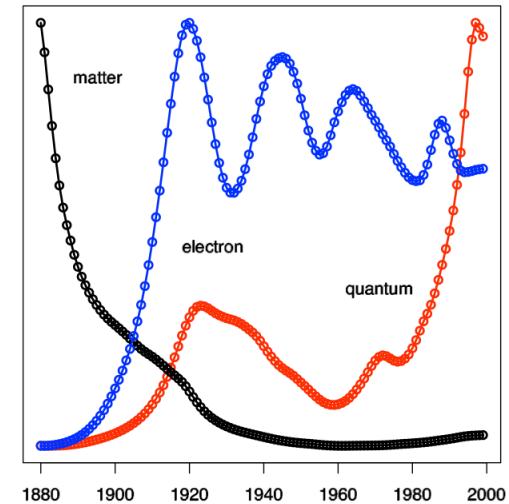
- Phát hiện các cộng đồng trong mạng xã hội

health+
DOConnect™
Your medical circles of trust!

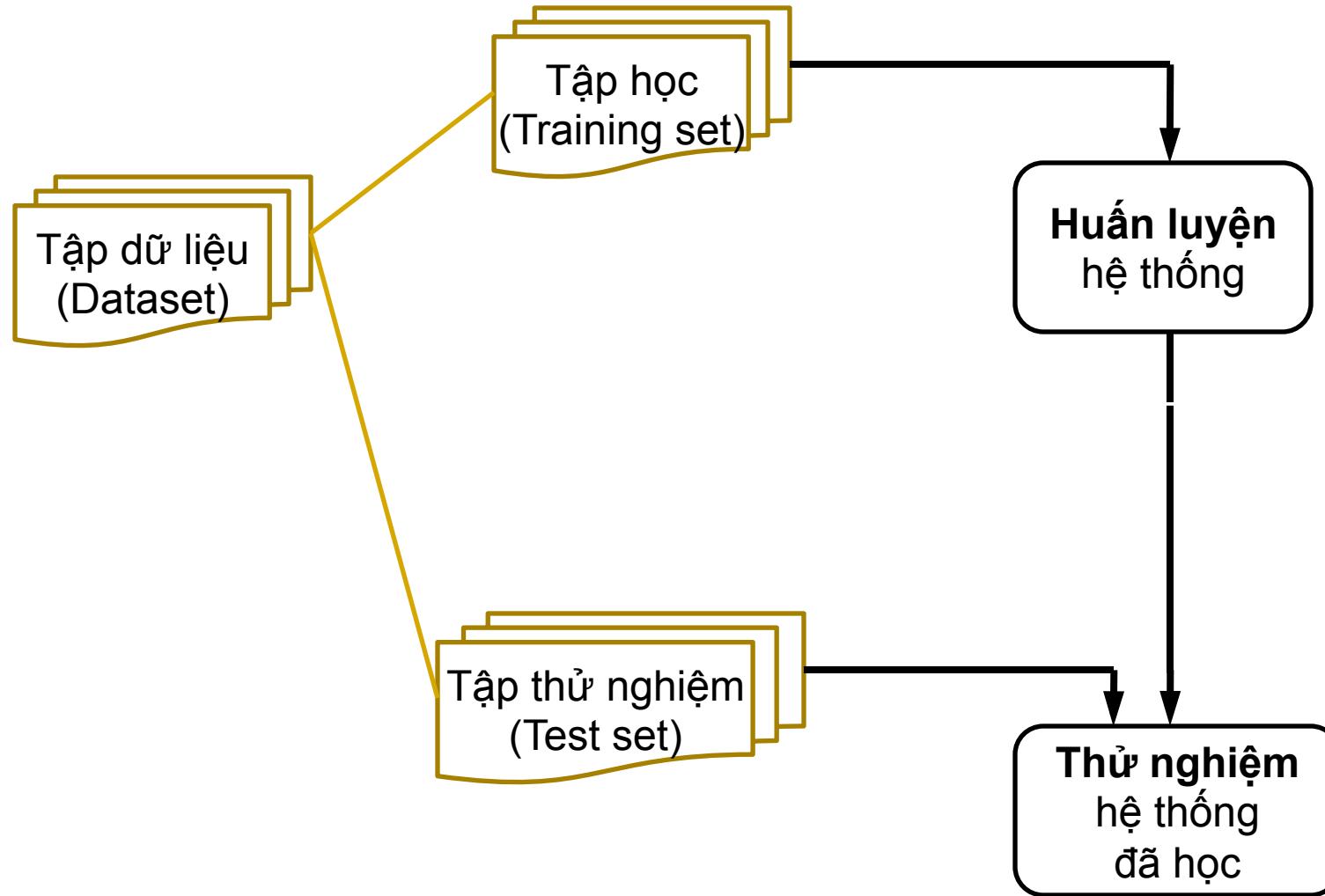
Học không giám sát: ví dụ (2)

■ Trends detection

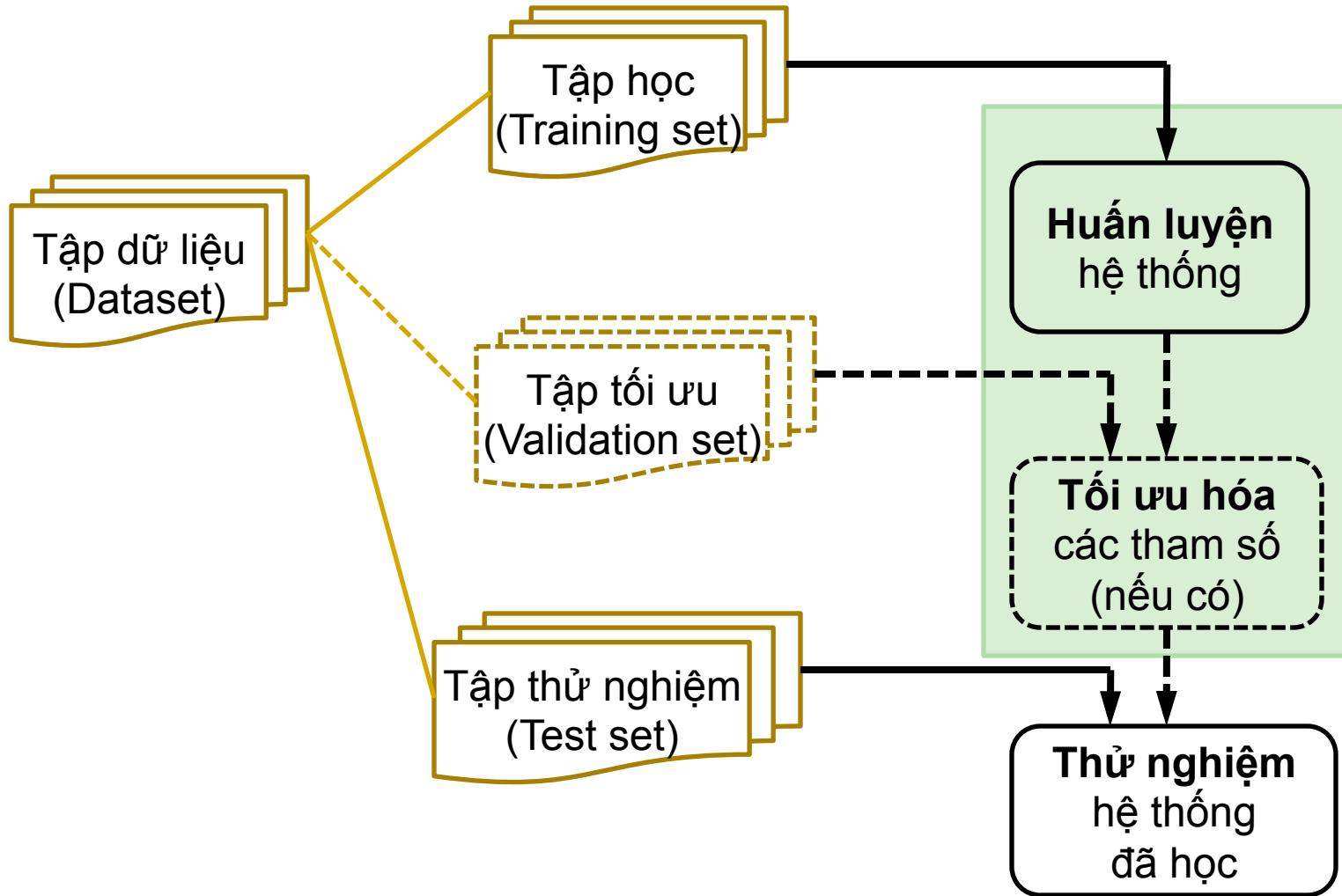
- Phát hiện xu hướng, thị yếu,...



Quá trình học máy: cơ bản



Quá trình học máy: toàn diện



Thiết kế một hệ thống học (1)

- Lựa chọn các ví dụ học (training/learning examples)
 - Các thông tin hướng dẫn quá trình học (training feedback) được chứa ngay trong các ví dụ học, hay là được cung cấp gián tiếp (vd: từ môi trường hoạt động)
 - Các ví dụ học theo kiểu có giám sát (supervised) hay không có giám sát (unsupervised)
 - Các ví dụ học nên tương thích với (đại diện cho) các ví dụ sẽ được làm việc bởi hệ thống trong tương lai (future test examples)
- Xác định hàm mục tiêu (giả thiết, khái niệm) cần học
 - $F: X \rightarrow \{0,1\}$
 - $F: X \rightarrow$ Một tập các nhãn lớp
 - $F: X \rightarrow \mathbb{R}^+$ (miền các giá trị số thực dương)
 - ...

Thiết kế một hệ thống học (2)

- Lựa chọn cách biểu diễn cho hàm mục tiêu cần học
 - Hàm đa thức (a polynomial function)
 - Một tập các luật (a set of rules)
 - Một cây quyết định (a decision tree)
 - Một mạng nơ-ron nhân tạo (an artificial neural network)
 - ...
- Lựa chọn một giải thuật học máy có thể học (xấp xỉ) được hàm mục tiêu
 - Phương pháp học hồi quy (Regression-based)
 - Phương pháp học quy nạp luật (Rule induction)
 - Phương pháp học cây quyết định (ID3 hoặc C4.5)
 - Phương pháp học lan truyền ngược (Back-propagation)
 - ...

Các vấn đề trong Học máy (1)

- Giải thuật học máy (Learning algorithm)
 - Những giải thuật học máy nào có thể học (xấp xỉ) một hàm mục tiêu cần học?
 - Với những điều kiện nào, một giải thuật học máy đã chọn sẽ hội tụ (tiệm cận) hàm mục tiêu cần học?
 - Đối với một lĩnh vực cụ thể và đối với một cách biểu diễn các ví dụ (đối tượng) cụ thể, giải thuật học máy nào thực hiện tốt nhất?
- No-free-lunch theorem [Wolpert and Macready, 2005]:
If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.
 - ❖ No algorithm can beat another on all domains.
(không có thuật toán nào luôn hiệu quả nhất trên mọi miền ứng dụng)

Các vấn đề trong Học máy (2)

■ Các ví dụ học (Training examples)

- Bao nhiêu ví dụ học là đủ?
- Kích thước của tập học (tập huấn luyện) ảnh hưởng thế nào đối với độ chính xác của hàm mục tiêu học được?
- Các ví dụ lỗi (nhiều) và/hoặc các ví dụ thiếu giá trị thuộc tính (missing-value) ảnh hưởng thế nào đối với độ chính xác?

Các vấn đề trong Học máy (3)

- Quá trình học (Learning process)
 - Chiến lược tối ưu cho việc lựa chọn thứ tự sử dụng (khai thác) các ví dụ học?
 - Các chiến lược lựa chọn này làm thay đổi mức độ phức tạp của bài toán học máy như thế nào?
 - Các tri thức cụ thể của bài toán (ngoài các ví dụ học) có thể đóng góp thế nào đối với quá trình học?

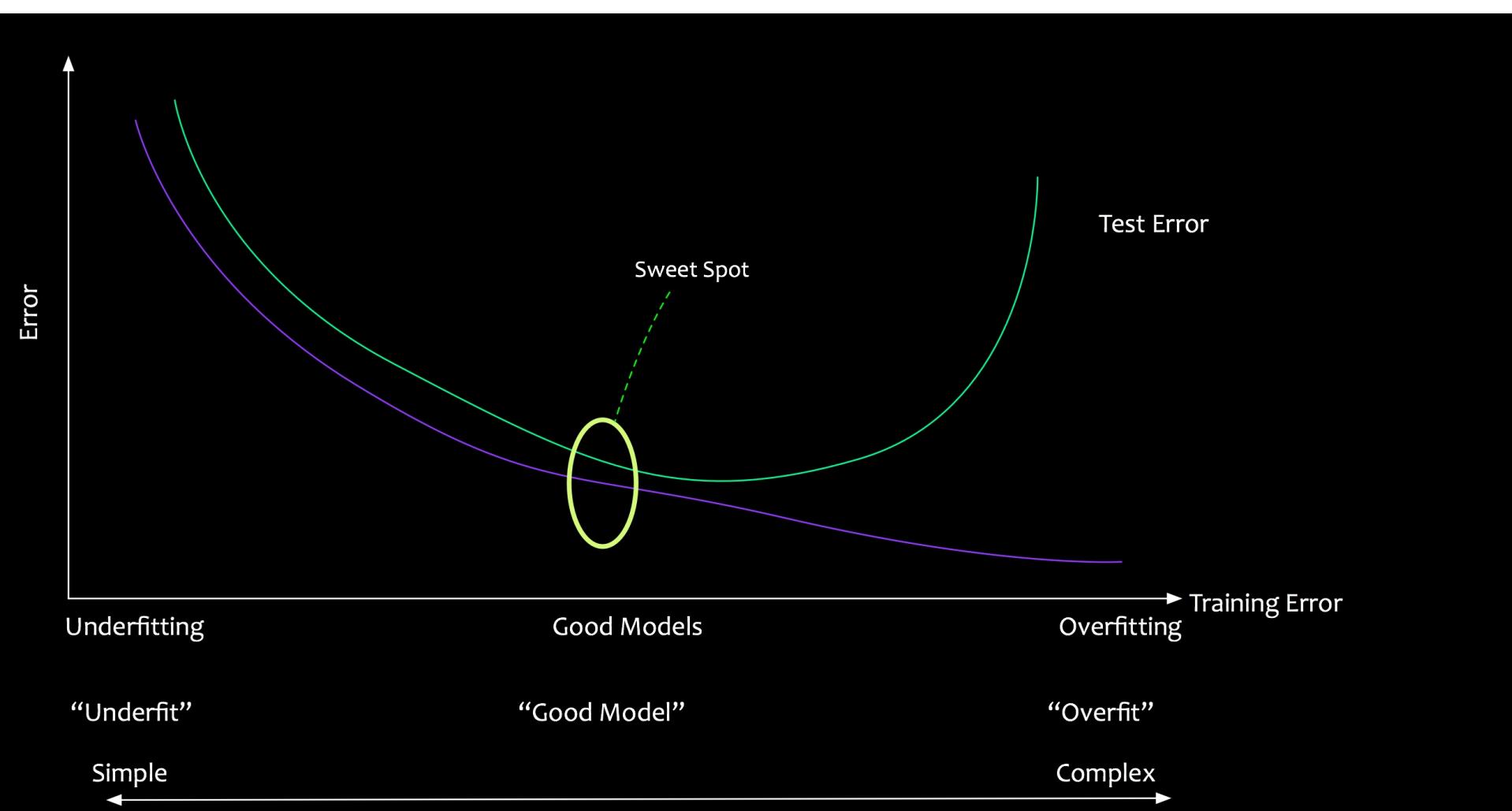
Các vấn đề trong Học máy (4)

- Khả năng/giới hạn học (Learnability)
 - Hàm mục tiêu nào mà hệ thống cần học?
 - Biểu diễn hàm mục tiêu: Khả năng biểu diễn (vd: hàm tuyến tính / hàm phi tuyến) vs. Độ phức tạp của giải thuật và quá trình học
 - Các giới hạn (trên lý thuyết) đối với khả năng học của các giải thuật học máy?
 - Khả năng khái quát hóa (generalization) của hệ thống?
 - Để tránh vấn đề “over-fitting” (đạt độ chính xác cao trên tập học, nhưng đạt độ chính xác thấp trên tập thử nghiệm)
 - Khả năng hệ thống tự động thay đổi (thích nghi) biểu diễn (cấu trúc) bên trong của nó?
 - Để cải thiện khả năng (của hệ thống đối với việc) biểu diễn và học hàm mục tiêu

Vấn đề overfitting

- Một hàm mục tiêu (một giả thiết) học được h sẽ được gọi là **quá khớp/quá phù hợp (overfit)** với một tập học nếu tồn tại một hàm mục tiêu khác h' sao cho:
 - h' kém phù hợp hơn (đạt độ chính xác kém hơn) h đối với tập học, nhưng
 - h' đạt độ chính xác cao hơn h đối với toàn bộ tập dữ liệu (bao gồm cả những ví dụ được sử dụng sau quá trình huấn luyện)
- Vấn đề overfitting thường do các nguyên nhân:
 - Lỗi (nhiều) trong tập huấn luyện (do quá trình thu thập/xây dựng tập dữ liệu)
 - Số lượng các ví dụ học quá nhỏ, không đại diện cho toàn bộ tập (phân bố) của các ví dụ của bài toán học

Vấn đề overfitting: minh họa



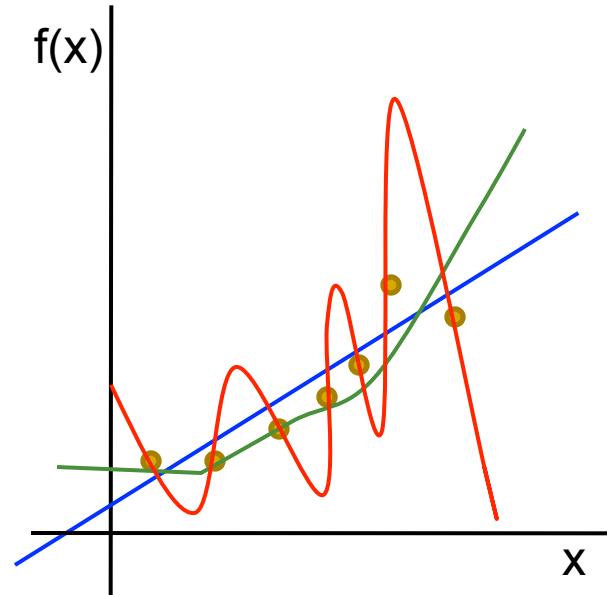
Vấn đề overfitting: giải pháp

- Trong số các giả thiết (hàm mục tiêu) học được, giả thiết nào khái quát hóa tốt nhất từ tập học?

Lưu ý: Mục tiêu của học máy là để đạt được độ chính xác cao trong dự đoán đối với các ví dụ sau này, không phải đối với các ví dụ học

- Hiệu chỉnh (Regularization):** hạn chế không gian học
- Occam's razor:** Ưu tiên chọn hàm mục tiêu đơn giản nhất phù hợp (không nhất thiết hoàn hảo) với các ví dụ học
 - Khái quát hóa tốt hơn
 - Dễ giải thích/diễn giải hơn
 - Độ phức tạp tính toán ít hơn

Hàm mục tiêu $f(x)$ nào đạt độ chính xác cao nhất đối với các ví dụ sau này?



Tài liệu tham khảo

- Alpaydin E. (2010). Introduction to Machine Learning. The MIT Press.
- Mitchell, T. M. (1997). Machine learning. McGraw Hill.
- Mitchell, T. M. (2006). The discipline of machine learning. Carnegie Mellon University, School of Computer Science, Machine Learning Department.
- Simon H.A. (1983). Why Should Machines Learn? In R. S. Michalski, J. Carbonell, and T. M. Mitchell (Eds.): Machine learning: An artificial intelligence approach, chapter 2, pp. 25-38. Morgan Kaufmann.
- Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67.