

Chương 9:

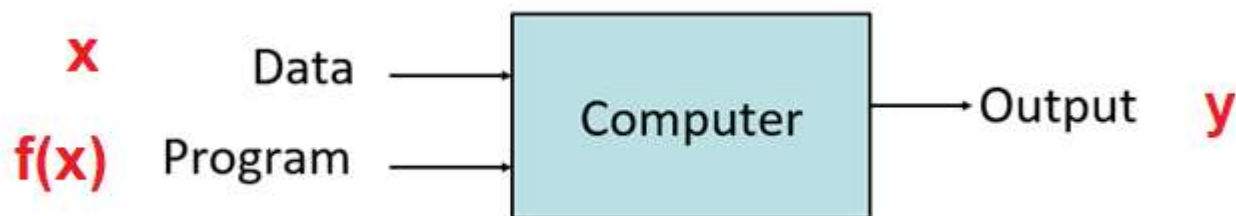
Các công nghệ số chủ chốt (tiếp)

- Học máy (Machine learning)

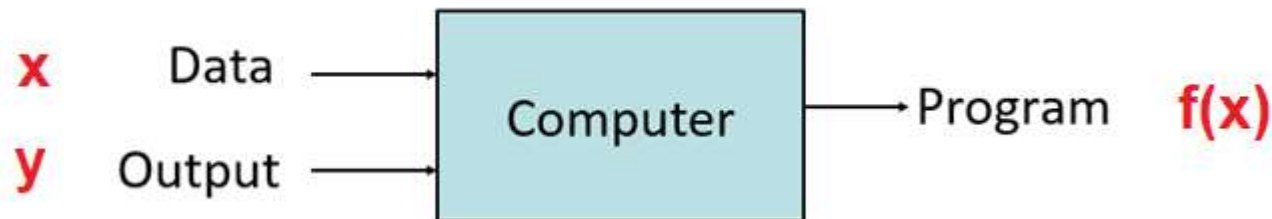
Học máy là gì?

- **Học máy** (Machine Learning - ML) là một lĩnh vực nghiên cứu của AI, cho phép máy tính **học từ dữ liệu** để thực hiện một nhiệm vụ nhất định
- ML tạo cho máy tính sức mạnh để tìm ra mọi thứ mà không cần lập trình rõ ràng

Lập trình truyền thống



Học máy



Lịch sử của học máy

- 1950 - Alan Turing đã tạo ra "Turing Test" để kiểm tra trí thông minh của máy tính.
- 1957 - Frank Rosenblatt thiết kế mạng nơron (neural network) đầu tiên, mô phỏng quá trình suy nghĩ của bộ não con người. Mạng này được gọi là Perceptron Model, có tác dụng phân loại dữ liệu vào 1 trong 2 lớp.
- 1967 - Thuật toán "nearest neighbor", dùng để nhận dạng những mẫu tương tự nhau.
- 1981 - Gerald Dejong giới thiệu khái niệm Explanation Based Learning (EBL), trong đó một máy tính phân tích dữ liệu huấn luyện và tạo ra một quy tắc chung để nó có thể làm theo bằng cách loại bỏ đi những dữ liệu không quan trọng.

Lịch sử của học máy

- 1989 - Yann LeCun giới thiệu mạng nơon tích chập (Convolutional neural networks - CNN), với khả năng tự học đặc trưng của dữ liệu đầu vào, dựa trên đó để phân loại dữ liệu. LeCun chứng minh CNN có thể dùng để nhận dạng chữ viết tay, chứng tỏ CNN có thể ứng dụng thực tế
- 1989 - Christopher Watkins phát triển Q-learning, 1 thuật toán học tăng cường tìm kiếm hành động tốt nhất cho từng thời điểm.
- 1990s – ML dịch chuyển từ cách tiếp cận hướng kiến thức (knowledge-driven) sang cách tiếp cận hướng dữ liệu (data-driven). Các nhà khoa học bắt đầu tạo ra các chương trình cho máy tính để phân tích một lượng lớn dữ liệu và rút ra các kết luận - hay là "học" từ các kết quả đó.
- 1997 - Sepp Hochreiter và Jürgen Schmidhuber đề xuất mạng Long Short-Term Memory, cho phép sinh dữ liệu dạng chuỗi (văn bản, video,...)



Lịch sử của học máy

- Thế kỷ 21, nhiều công ty đầu tư vào các dự án dùng học máy



Các công ty tập trung vào xây dựng cơ sở hạ tầng



Các công ty thống trị công nghiệp IT với việc sử dụng ML

- 2006 - Geoffrey Hinton đưa ra thuật ngữ Học sâu (Deep Learning) để đề cập đến các thuật toán sử dụng mạng nơ-ron nhiều lớp giúp máy tính nhận diện các đối tượng và văn bản trong ảnh và video.

Lịch sử của học máy

- 2014 - Ian Goodfellow và cộng sự tạo ra mạng GAN (Generative Adversarial Networks), dùng trong sinh ảnh, tạo ảnh giả.
- 2015 - Hơn 3.000 nhà nghiên cứu AI và Robotics, được sự ủng hộ bởi những người nổi tiếng như giáo sư vật lý Stephen Hawking và tỉ phú Elon Musk, đã ký thư ngỏ cảnh báo về sự nguy hiểm của vũ khí tự động mà không có sự can thiệp của con người.



- 2017 - Google phát triển mạng transformer. Đây là nghiên cứu nền tảng cho việc tạo các công cụ dựa trên học sâu có thể phân tích dữ liệu không nhãn tốt nhất

Lịch sử của học máy

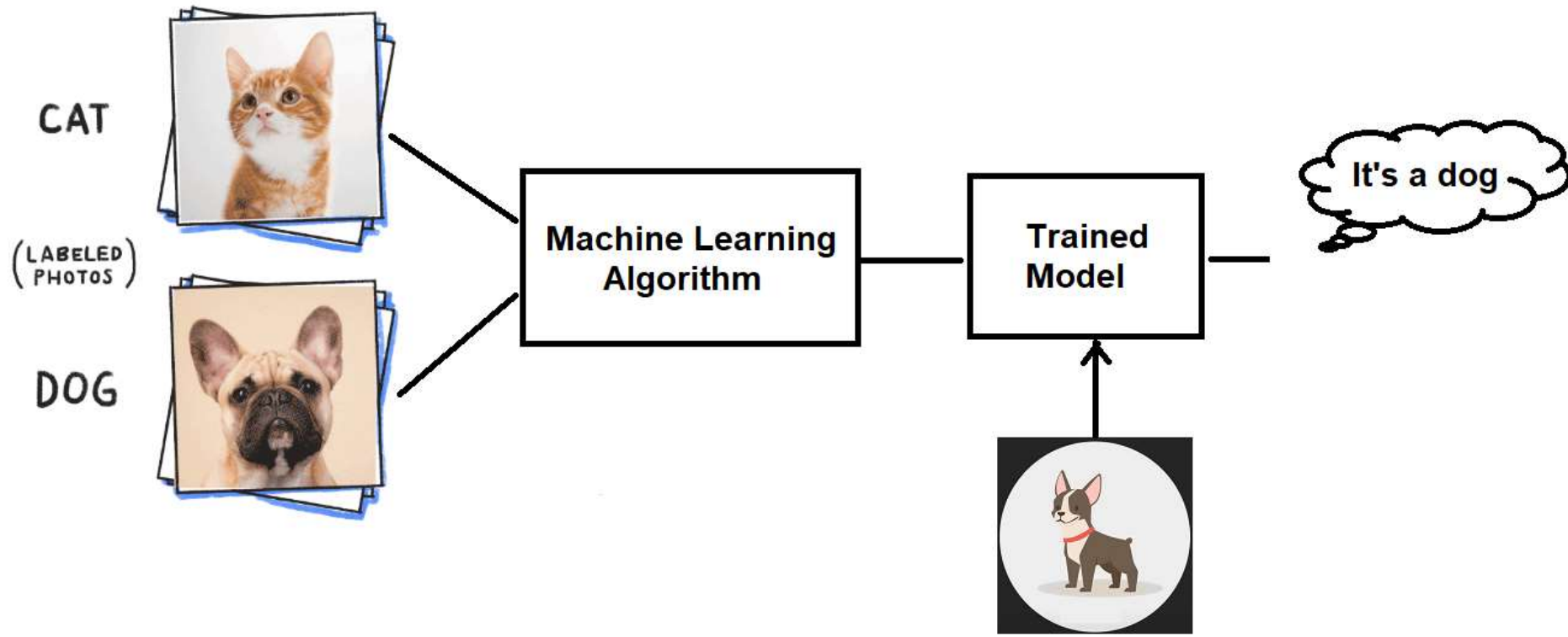
- 2018 - OpenAI công bố GPT (Generative Pre-trained Transformer), nền tảng cho các LLM sau này.
- 11/2022 - OpenAI công bố ChatGPT phát triển từ GPT 3.5 LLM. ChatGPT có thể trả lời lưu loát đầy đủ các câu hỏi mà bạn đưa ra, có thể làm thơ, soạn nhạc, viết thư, thiết kế và thậm chí là cả sửa lỗi trong lập trình. Tuy nhiên, một số câu trả lời của ChatGPT nghe có vẻ hợp lý nhưng không chính xác hoặc vô nghĩa.
- 2023 - OpenAI công bố GPT-4 có thể xử lý dữ liệu đầu vào là ảnh hoặc văn bản.
- 2023 - Elon Musk, Steve Wozniak (đồng sáng lập Apple) và hàng nghìn người khác đã ký vào đơn kiến nghị dừng phát triển các hệ thống AI mạnh hơn GPT-4



Steve Wozniak

Học có giám sát

- Học có giám sát (supervised learning) đưa ra dự đoán cho dữ liệu mới dựa trên tập dữ liệu đã được gán nhãn (còn gọi là tập dữ liệu huấn luyện, hay training dataset).
- 1 dữ liệu gán nhãn = đầu vào + đầu ra mong muốn.
- Học có giám sát phân tích training dataset, tìm ra qui luật và dựa trên đó để dự đoán nhãn cho các dữ liệu mới.

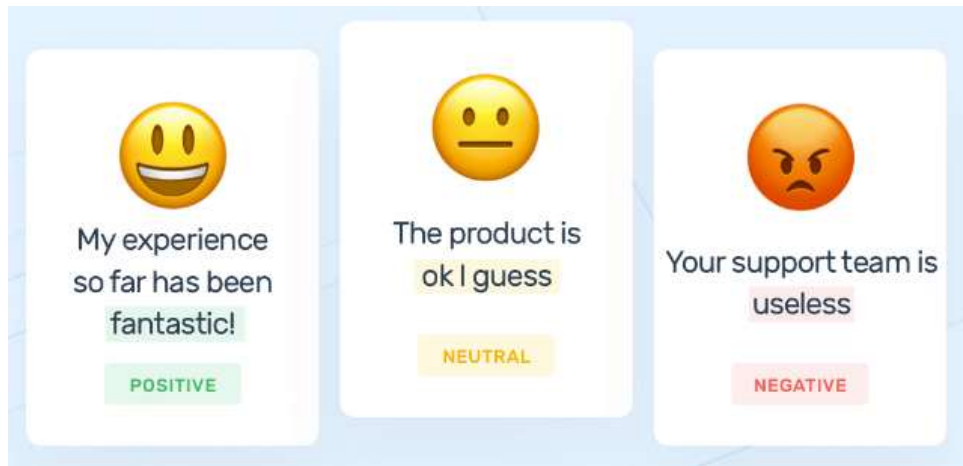


Học có giám sát

- 2 phương pháp:
 - Phân loại (classification)
 - Hồi qui (regression)
- Học phân loại:



Lọc thư rác (spam filtering)



Phân tích cảm xúc (sentiment analysis)

Ví dụ học có giám sát: Phân loại

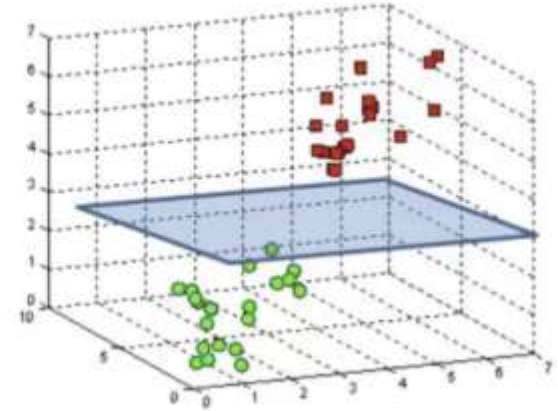
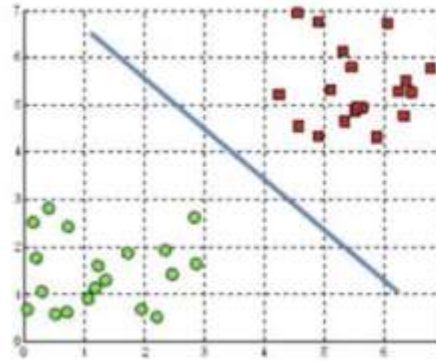
- **Multi-class classification** (phân loại nhiều lớp): when the output y is one of the pre-defined labels $\{c_1, c_2, \dots, c_L\}$ (mỗi đầu ra chỉ thuộc 1 lớp, mỗi quan sát x chỉ có 1 nhãn)
 - Spam filtering: y in {spam, normal}
 - Financial risk estimation: y in {high, normal, no}
 - Discovery of network attacks: ?
- **Multi-label classification** (phân loại đa nhãn): when the output y is a subset of labels (mỗi đầu ra là một tập nhỏ các lớp; mỗi quan sát x có thể có nhiều nhãn)
 - Image tagging: $y = \{\text{birds, nest, tree}\}$
 - sentiment analysis



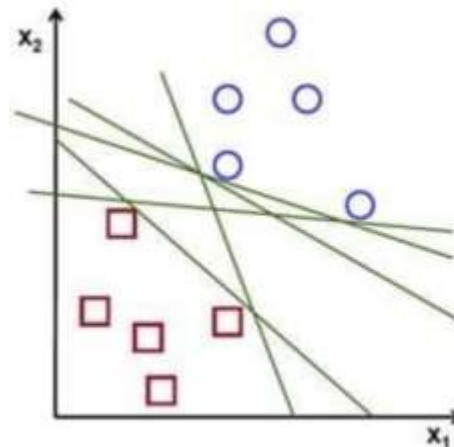
BIRDS NEST TREE

Học có giám sát - SVM

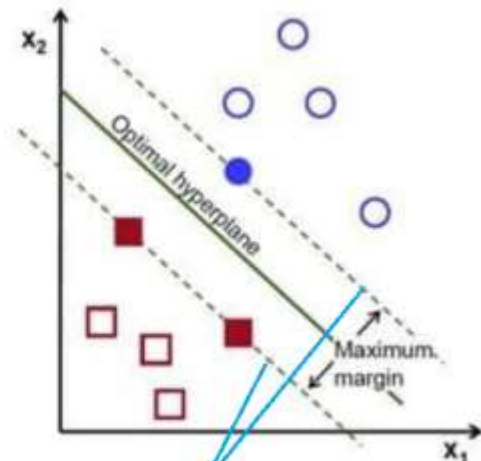
- Máy véc tơ hỗ trợ (Support Vector Machine - SVM) là một trong những thuật toán học có giám sát có hiệu năng cao nhất.
- Mục tiêu: tìm ra một siêu phẳng trong không gian N chiều (ứng với N đặc trưng) chia dữ liệu thành hai phần tương ứng với lớp của chúng.
- Siêu phẳng là một hàm $y = f(x_1, \dots, x_n)$ với $x_1, \dots, x_n$ là các đặc trưng của dữ liệu



SVM trong không gian hai chiều và ba chiều



Possible hyperplanes



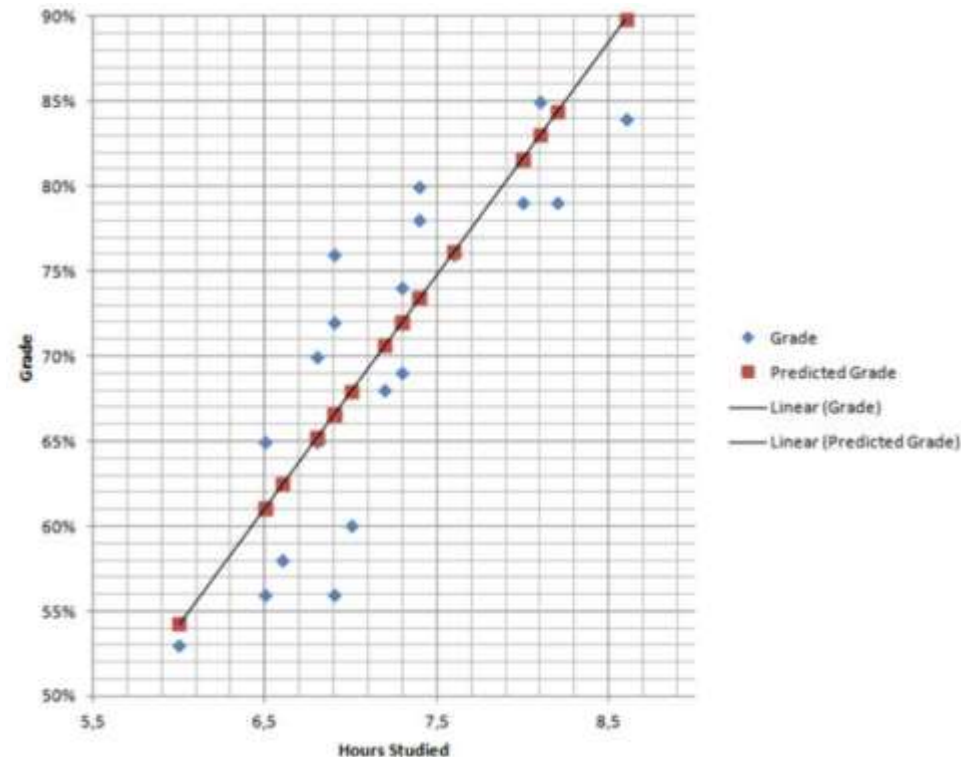
Support vectors

Siêu phẳng tối ưu có lề cực đại

Học có giám sát

- Học phân loại: đầu ra là nhãn (0,1), (chó, mèo), (tích cực, tiêu cực, bình thường)
- Hồi qui: giá trị liên tục

Student Name	Hours Studied	Grade
Jack	6	53%
Anne	7	60%
Harry	6,5	56%
Sharon	8	79%
John	6,6	58%
James	8,1	85%
Jill	6,8	70%
Adam	6,9	56%
Brandon	7,3	69%
Brett	6,9	76%
Brady	8,2	79%
Charles	7,2	68%
Darren	7,3	74%
Dave	6,9	72%
Dawn	8,6	84%
Denise	7,4	78%
Eric	7,6	76%
Emily	6,8	65%
Fred	8	92%
Fran	7,4	80%
Jane	6,5	65%



Thống kê số giờ học của SV và điểm thi

Xác định hàm dự đoán điểm dựa trên số giờ học của SV

Ví dụ học có giám sát: Hồi quy

• Hồi quy dự báo chứng khoán

Google Inc (NASDAQ:GOOG)

Add to portfolio

More results

744.00 +41.13 (5.85%)

Real-time: 10:43AM EST
NASDAQ real-time data - Disclaimer
Currency in USD

Range 735.79 - 747.99
52 week 556.52 - 774.38
Open 735.99
Vol / Avg. 2.68M/2.28M
Mkt cap 244.39B
P/E 22.91

Div/yield -
EPS 32.46
Shares 328.59M
Beta 1.08
Inst. own 69%

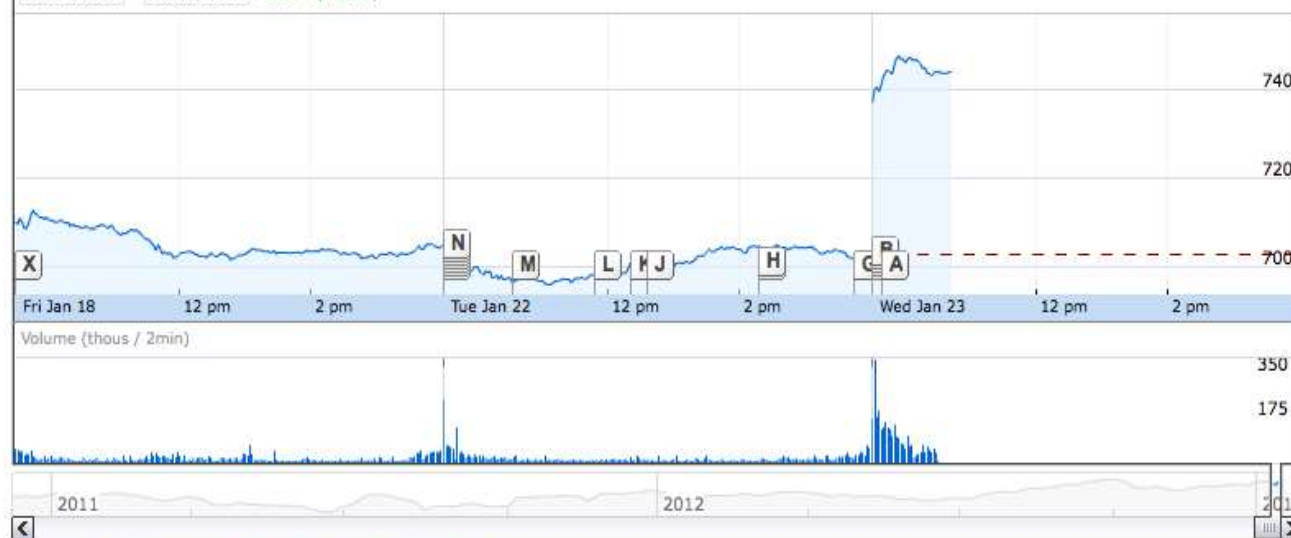


Dow Jones	13,758.94	0.34%
Nasdaq	3,151.72	0.27%
Technology		0.33%
GOOG	744.00	5.85%

Compare: Enter ticker here Add ☐ Dow Jones ☐ Nasdaq ☐ BIDU ☐ YNDX ☐ BCOR ☐ MSFT ☐ YHOO more »

Zoom: 1d 5d 1m 3m 6m YTD 1y 5y 10y All

Jan 18, 2013 - Jan 23, 2013 +32.07 (4.51%)



Settings | Plot feeds | Technicals | Link to this view

Volume delayed by 15 mins.

- A** Google Inc. (GOOG) Is Up Sharply On Q4 Results
RTT News - 1 hour ago
- B** Stocks to Watch: Google, Coach, Annie's
Wall Street Journal - 1 hour ago
- C** Google Inc (GOOG) Reports Strong Earnings, Shares Rise
ValueWalk - 3 hours ago
- D** Google 4th-Quarter Profits Increase as Ad Pricing Improves
NASDAQ - 15 hours ago
- E** Facebook Inc (FB)'s Social Graph Is a Google Inc (GOOG) Plus Killer
Insider Monkey - 16 hours ago

All news for Google Inc »

Subscribe

Events

Add GOOG to my calendars

Apr 15, 2013

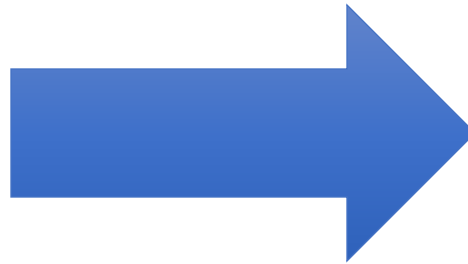
Q1 2013 Google Earnings Release



ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Ví dụ học có giám sát: Hồi quy

- Hồi quy dự báo thời tiết

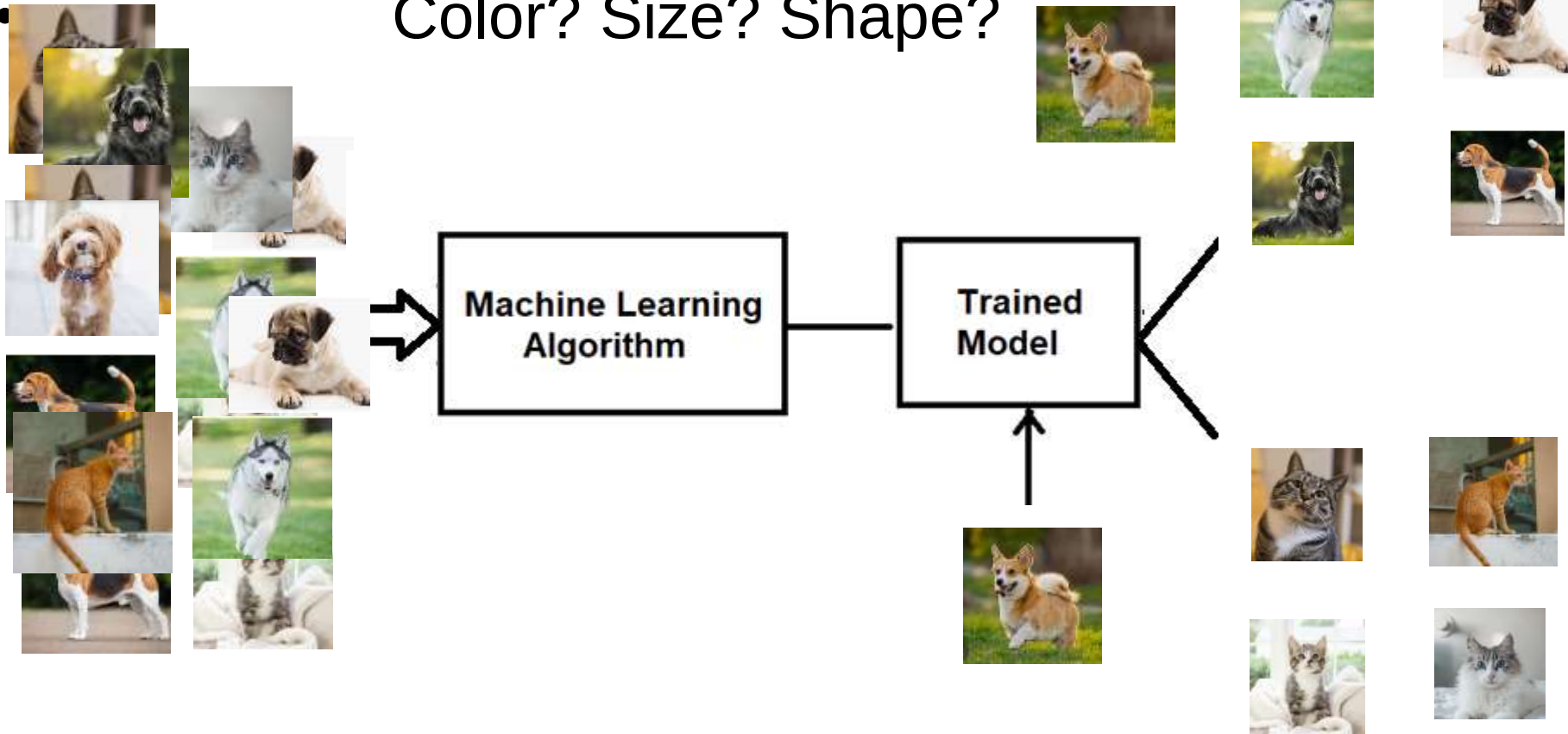


Temperature

27°C

Học không giám sát (unsupervised learning)

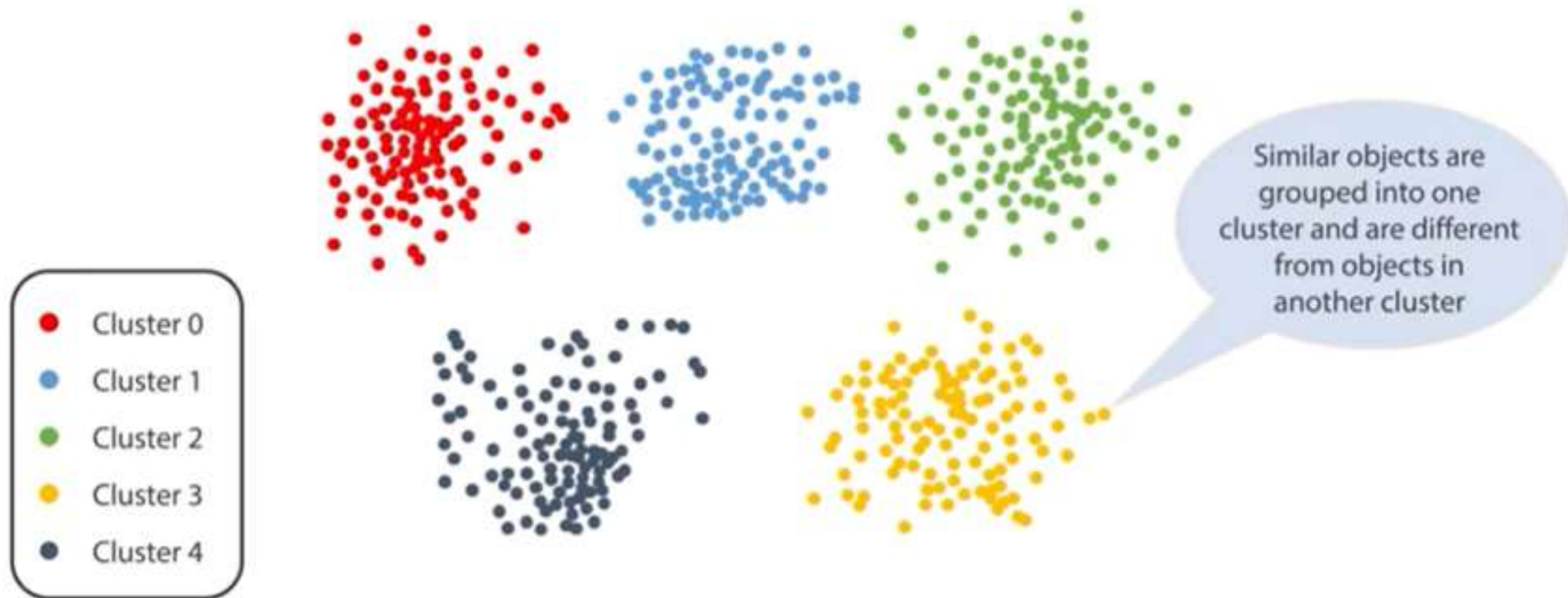
- Không có trước nhãn
- Cố gắng phát hiện ra tính chất của dữ liệu
- Color? Size? Shape?



Học không giám sát

- Các ứng dụng:

- Phân cụm khách hàng để tối ưu hóa chiến lược kinh doanh
- Nhóm các sách có nội dung tương tự
- Phát hiện các cộng đồng trong mạng xã hội



Học tăng cường (Reinforcement Learning)

- Giống cách đứa trẻ học: nhận thức hành động đúng sai dựa trên việc khen ngợi/chê trách của người lớn



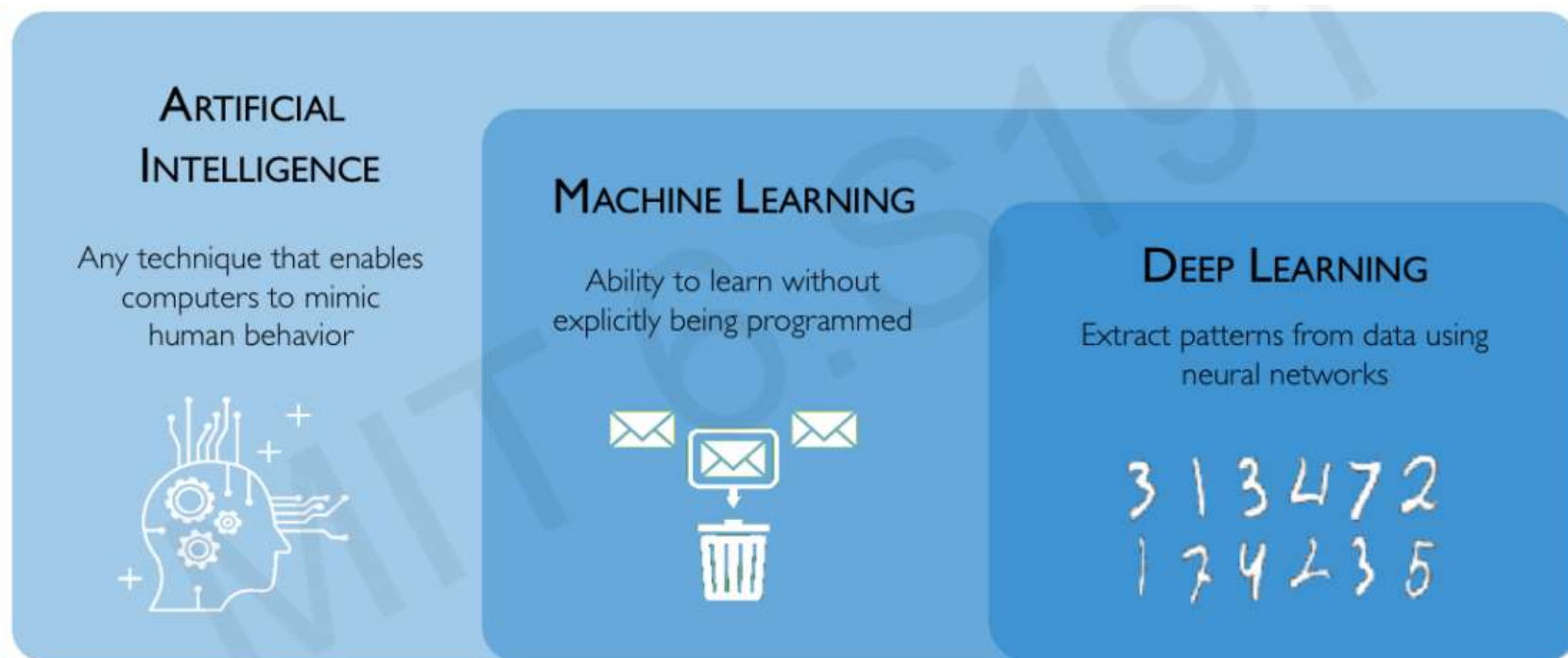
Học tăng cường

- Học tăng cường (Reinforcement Learning – RL) là kỹ thuật học máy giúp huấn luyện phần mềm đưa ra quyết định nhằm thu về kết quả tối ưu nhất.
- Thuật toán RL sử dụng mô hình khen thưởng và trừng phạt trong quy trình xử lý dữ liệu. Các thuật toán này tiếp thu ý kiến phản hồi của từng hành động và tự khám phá ra con đường xử lý tốt nhất để thu về kết quả cuối cùng.



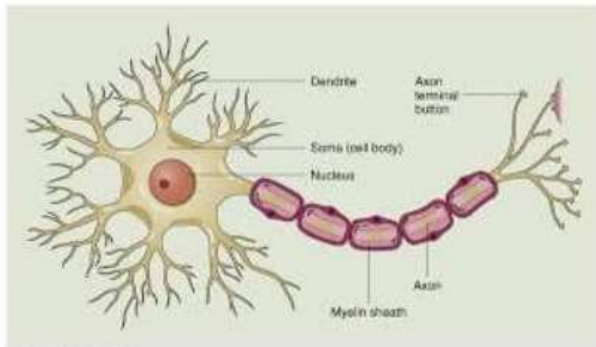
Học sâu (Deep Learning)

- Học sâu (Deep learning) là một nhánh của học máy
 - Là phương pháp học máy **sử dụng mạng nơ-ron nhân tạo** để trích xuất đặc trưng tự động từ dữ liệu



Deep Learning

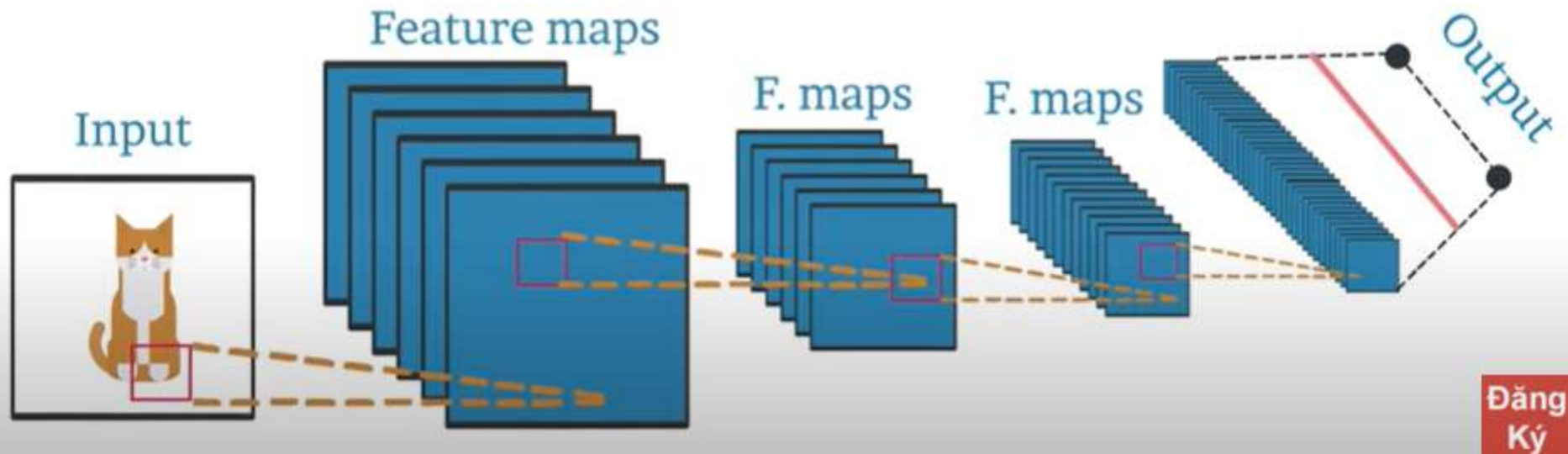
- Học sâu dựa trên mạng nơ-ron nhân tạo ([Artificial Neural Networks - ANN](#)), bắt chước hoạt động của bộ não con người và được đào tạo từ lượng lớn dữ liệu được cung cấp.
- Não người tạo bởi nhiều nơ-ron liên kết với nhau
 - Học sâu được xây dựng với nhiều tầng nơ-ron kết nối với nhau.



- Để thực hiện một nhiệm vụ cụ thể (như phân loại), khi nhận được dữ liệu đầu vào (ảnh, văn bản, video, ...), dữ liệu sẽ đi qua tất cả các tầng mạng để phân tích nó

Deep Learning

- Thuật toán DL thực hiện 1 nhiệm vụ nhiều lần, mỗi lần tinh chỉnh 1 chút theo dữ liệu mẫu để cải thiện kết quả.
- DL học ở nhiều cấp độ trừu tượng → cho phép học hàm ánh xạ phức tạp mà không cần thuật toán cụ thể nào.
- Không ai hiểu những gì xảy ra trong mạng nơ ron nhân tạo → DL có thể coi là hộp đen



Convolutional neural networks dùng trong phân loại ảnh

Ứng dụng của Deep Learning

- Ứng dụng:
 - Xử lý ảnh
 - Nhận dạng tiếng nói
 - Xử lý ngôn ngữ tự nhiên.
- Các mô hình học sâu thường có hiệu năng tốt hơn học máy, nhưng nó đòi hỏi lượng dữ liệu huấn luyện lớn (hàng triệu mẫu), thời gian huấn luyện lâu.

Nhược điểm của Deep Learning

- Ứng dụng:
 - Xử lý ảnh
 - Nhận dạng tiếng nói
 - Xử lý ngôn ngữ tự nhiên.
- Các mô hình học sâu thường có hiệu năng tốt hơn học máy, nhưng nó đòi hỏi lượng dữ liệu huấn luyện lớn (hàng triệu mẫu), thời gian huấn luyện lâu.

Nhược điểm của Deep Learning

- Khát dữ liệu: dữ liệu sạch, dữ liệu gán nhãn với số lượng lớn

China Is Achieving AI Dominance by Relying on Young Blue-Collar Workers

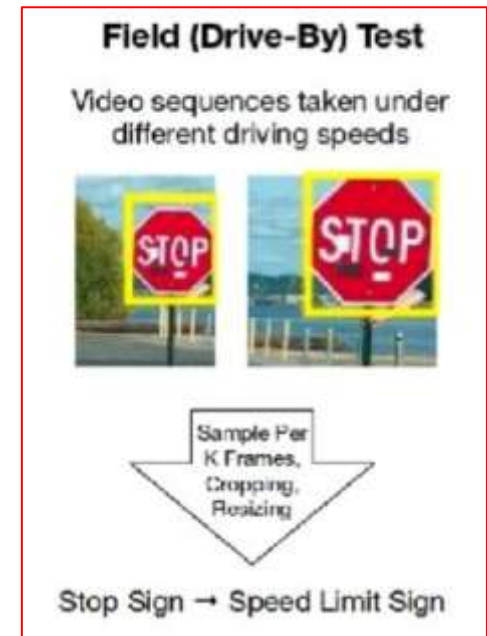
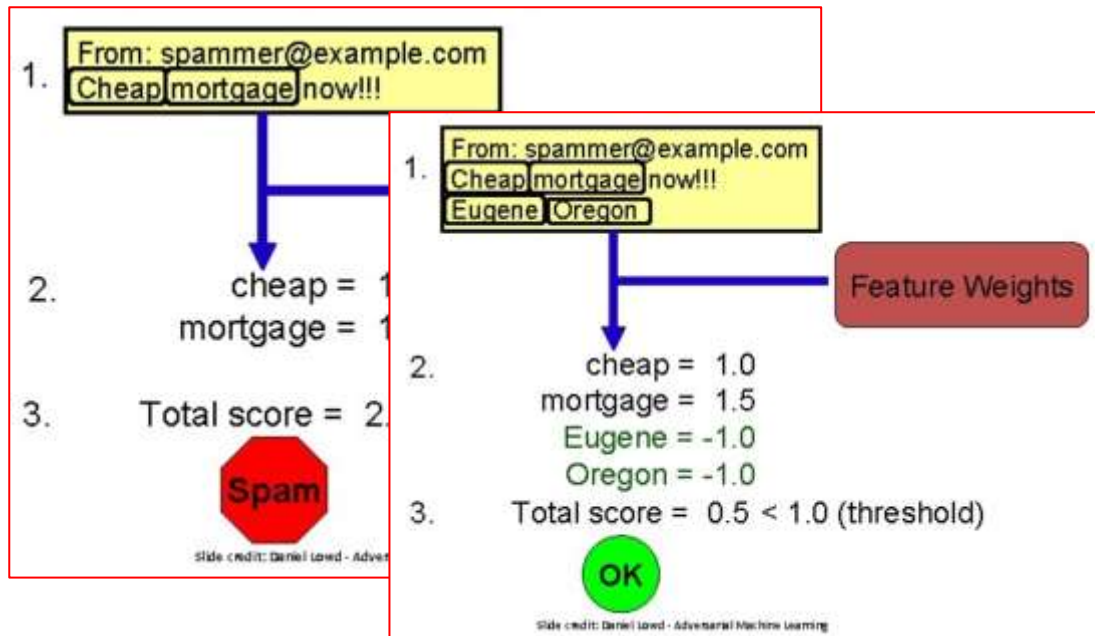


Ngành gán nhãn dữ liệu tại Trung Quốc:
2019: > 1 triệu lao động.
2022: 5 triệu lao động.



Nhược điểm của Deep Learning

- Thiếu khả năng diễn giải mô hình (tức là tại sao mô hình lại đưa ra dự đoán đó?)



... và còn là AI Ethics

Daily Mail • Follow
Jun 27, 2020 •

White man calls cops on black men at marina



Keep seeing videos about
Primates?

Yes

Dismiss

Wrongfully accused by an algorithm

June 24, 2020 at 12:47 pm | Updated June 24, 2020 at 6:46 pm



2019 theo đánh giá CP Mỹ: Nhận dạng sai người da màu (gốc Phi, và gốc Á) cao gấp 10 đến 100 lần người da trắng (gốc Âu)