

Giới thiệu môn học

- Tổng số tiết: 30 tiết lý thuyết, 30 tiết thực hành
- Tóm tắt môn học: Học phần này cung cấp cho sinh viên các kiến thức cơ bản về máy học, các vấn đề máy học giám sát và không giám sát, các mô hình máy học cho các vấn đề cơ bản.
- Ngôn ngữ lập trình sử dụng trong thực hành: Python
- Đánh giá học phần:
 - Giữa kỳ (40%): chuyên cần + bài tập + bài thực hành + kiểm tra giữa kỳ
 - Cuối kỳ (60%): trắc nghiệm

Chuẩn đầu ra học phần

- **Kiến thức**
 - Hiểu các khái niệm cơ bản của máy học
 - Hiểu và giải thích được nguyên lý hoạt động cơ bản của máy học.
 - Phân biệt được các nhóm giải thuật máy học cơ bản.
 - Dự đoán các mô hình cho một số bài toán máy học.
- **Kỹ năng**
 - Nhận dạng và phân tích được một số bài toán máy học thường gặp.
 - Lựa chọn và ứng dụng mô hình vào một vấn đề máy học.
 - Sử dụng được các công cụ hỗ trợ vào các bài toán cụ thể.
 - So sánh được hiệu quả sử dụng của các giải pháp máy

Chuẩn đầu ra học phần

- Mức tự chủ và trách nhiệm
 - Nhận biết bản chất bài toán máy học và kết quả mong muốn của việc ứng dụng.
 - Có khả năng xử lý tình huống trong thực tế theo cá nhân và theo nhóm.

Nội dung học phần

- Tổng quan về máy học
- Tiến trình học máy
- Học có giám sát
- Học không giám sát

Tài liệu tham khảo

- Đỗ Thanh Nghị, Phạm Nguyên Khang (2012), Giáo trình Nguyên lý Máy học, NXB Đại học Cần Thơ, Cần Thơ.
- Nguyễn Văn Sinh, Trần Mạnh Hà, Nguyễn Thị Thanh Sang, Nguyễn Minh Quân (2018), Nền tảng toán học trong CNTT, NXB Đại học Quốc gia TP. Hồ Chí Minh, Hồ Chí Minh.
- <https://machinelearningcoban.com/>

TỔNG QUAN VỀ MÁY HỌC

Máy học ứng dụng

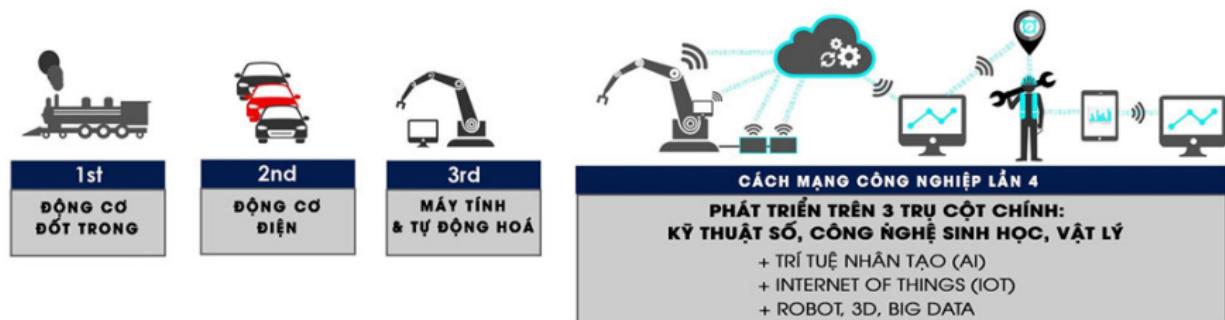
HK1, 2024 - 2025

Nội dung

- 1 Khái niệm cơ bản
- 2 Phân loại
- 3 Bài toán máy học
- 4 Một số ứng dụng cơ bản

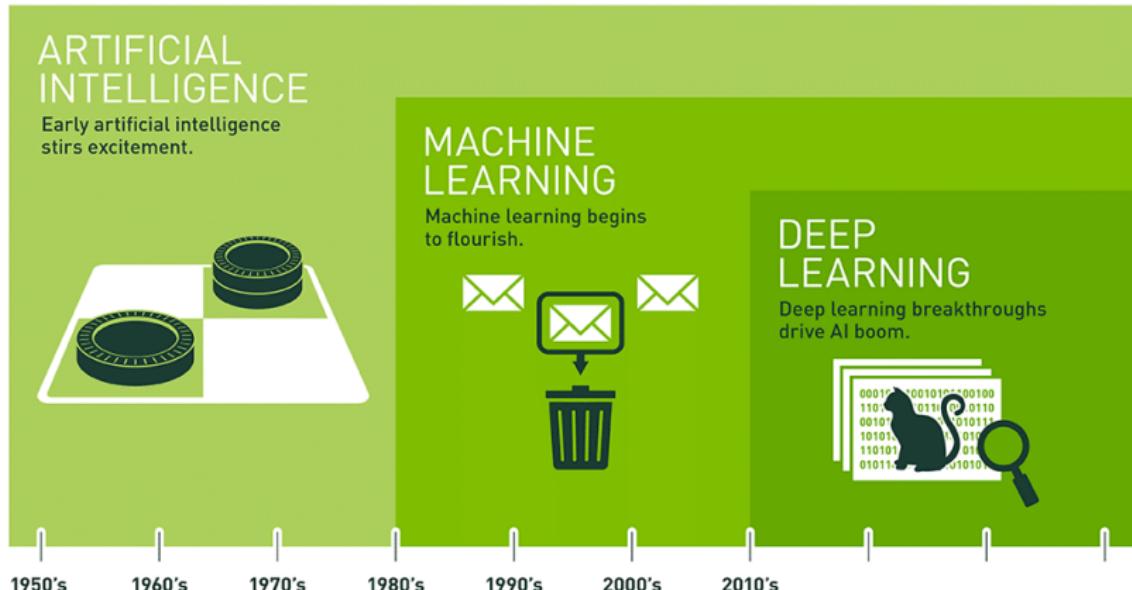
Khái niệm

LỊCH SỬ 4 CUỘC CÁCH MẠNG CÔNG NGHIỆP



Internet of Things (IoT)

Khái niệm



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Khái niệm

- **Strong AI**: mô phỏng các hành vi thông minh của con người → máy móc có trí thông minh ngang hoặc vượt trí tuệ con người
- **Weak AI (Narrow AI)**: mô phỏng nhận thức của con người → thực hiện một nhiệm vụ hoặc một tập hợp các nhiệm vụ cụ thể



Trí tuệ nhân tạo (Artificial Intelligence)

4 quan điểm về AI

- Suy nghĩ như người
- Hành động như người
- Suy nghĩ có lý trí
- Hành động có lý trí

Think Humanly	Think Rationally
	
	
Act Humanly	Act Rationally

Máy học (Machine learning)

- Arthur Samuel (1959):

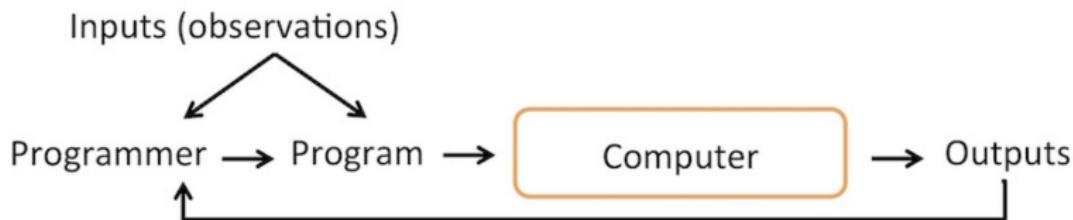
Machine learning is the subfield of computer science that gives computers the ability to **learn without being explicitly programmed**.

Máy học là một lĩnh vực con của khoa học máy tính, cho phép máy tính có **khả năng học hỏi** mà **không cần phải được lập trình một cách rõ ràng**.

Máy học (Machine learning)

A

The Traditional Programming Paradigm



B

Machine Learning



Máy học (Machine learning)

- Tom M. Mitchell (1997):

A computer program is said to learn from **experience E** with respect to some **task T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E

Chương trình máy tính **học từ kinh nghiệm E** để **tác vụ T** với hiệu suất được đo bằng **thước đo đánh giá P** nếu hiệu quả của nó khi thực hiện nhiệm vụ *T* được đánh giá bởi *P* cải thiện theo kinh nghiệm *E*

Nhiệm vụ - T

Trong bối cảnh máy học:

- **Nhiệm vụ** thường được mô tả thông qua việc một hệ thống máy học xử lý một **điểm dữ liệu (data point)** như thế nào
- Mỗi điểm dữ liệu bao gồm nhiều **đặc trưng (feature)** khác nhau
- Mỗi đặc trưng *thường* được biểu diễn dưới dạng số
- Mỗi điểm dữ liệu biểu diễn bằng 1 vector gọi là **vector đặc trưng (feature vector)**

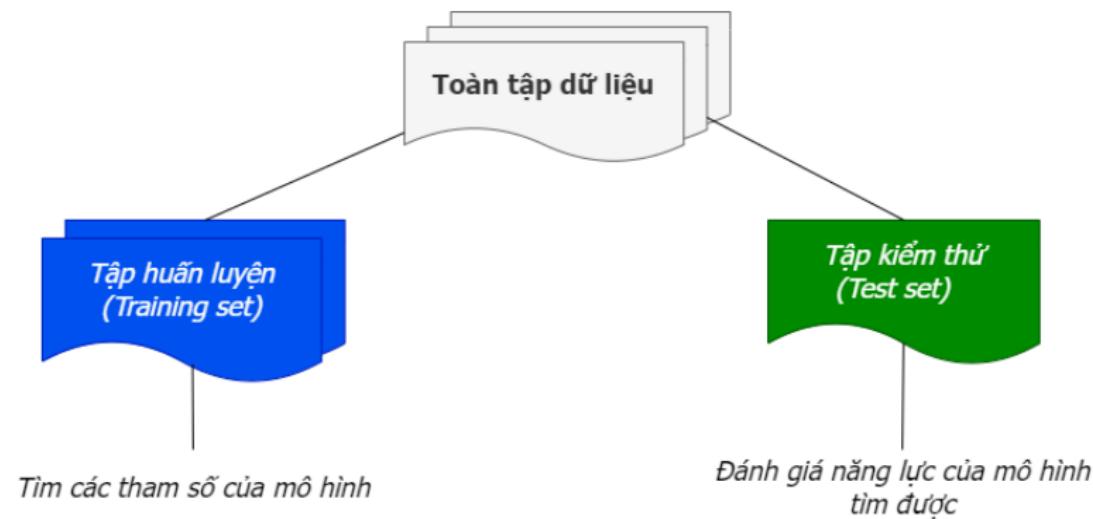
Nhiệm vụ - T

Ví dụ:

- Phân nhóm khách hàng:
 - nhóm các khách hàng vào các nhóm dựa trên các đặc điểm của họ
 - điểm dữ liệu: mỗi khách hàng; đặc trưng: tuổi, giới tính, thu nhập, số tiền cho mỗi lần mua sắm, tần suất mua sắm...
- Nhận dạng cảm xúc qua ảnh
 - gán nhãn cảm xúc (vui, buồn, tức giận...) cho một hình ảnh
 - điểm dữ liệu: các hình ảnh; đặc trưng: mỗi giá trị của 1 điểm ảnh là 1 đặc trưng → vector chứa toàn bộ giá trị các pixel của ảnh là vector đặc trưng

Thước đo đánh giá - P

- Kiểm tra hiệu suất của thuật toán máy học → thiết kế các thước đo đánh giá để đo kết quả



Thước đo đánh giá - P

Lưu ý:

- **Tìm tham số của mô hình:** chỉ sử dụng tập *huấn luyện*
- **Đánh giá mô hình:** áp dụng trên cả 2 tập (mô hình hoạt động trên tập kiểm thử trước hết phải hoạt động tốt trên tập huấn luyện)

Thước đo đánh giá - P

Lưu ý:

- Mô hình thực tế có thể liên tục được cập nhật dựa trên **dữ liệu mới** (*online learning*) → dữ liệu mới không được hệ thống sử dụng để xây dựng mô hình nhưng có thể sử dụng để cải tiến
- Offline learning: mô hình được xây dựng một lần trên tập huấn luyện.

Kinh nghiệm - E

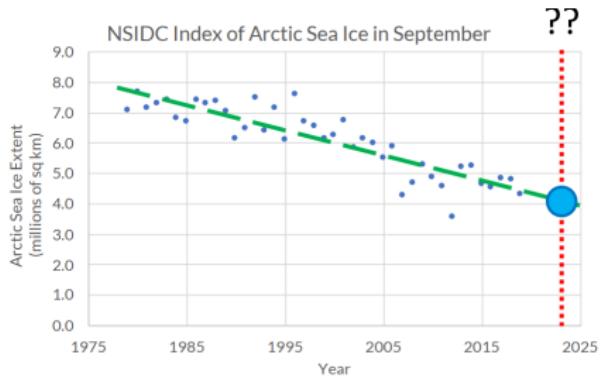
- **Huấn luyện các mô hình MH ~ cho các mô hình trải nghiệm trên các tập dữ liệu (dataset) - training set**
- Tập dữ liệu huấn luyện khác nhau → mô hình các trải nghiệm khác nhau
- Chất lượng của các tập dữ liệu → ảnh hưởng đến hiệu suất của mô hình

Máy học (Machine learning)



- Task T: chơi cờ
- Performance P: tỷ lệ chiến thắng
- Experience E: số lượng ván cờ đã chơi

Máy học (Machine learning)



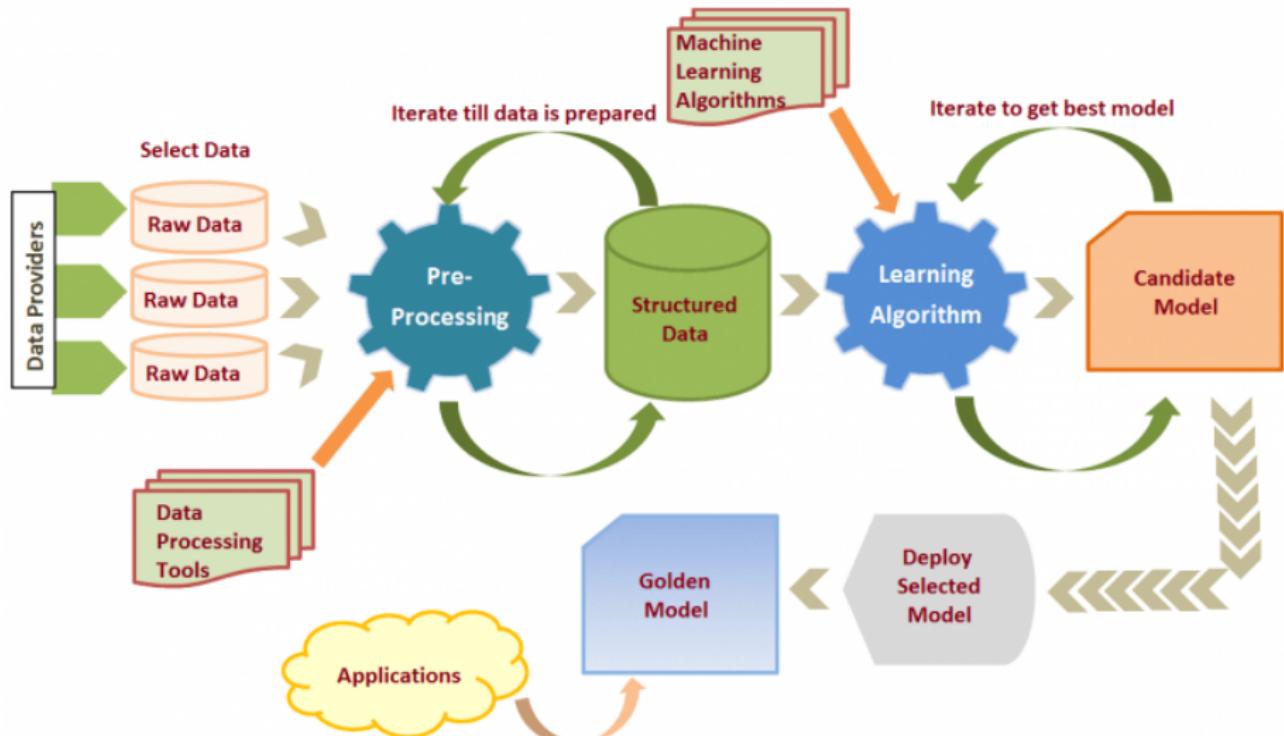
- Task T: dự đoán phạm vi băng biển Bắc Cực
- Performance P: lỗi dự đoán (giá trị khác biệt)
- Experience E: dữ liệu ghi nhận phạm vi theo thời gian

Máy học (Machine learning)

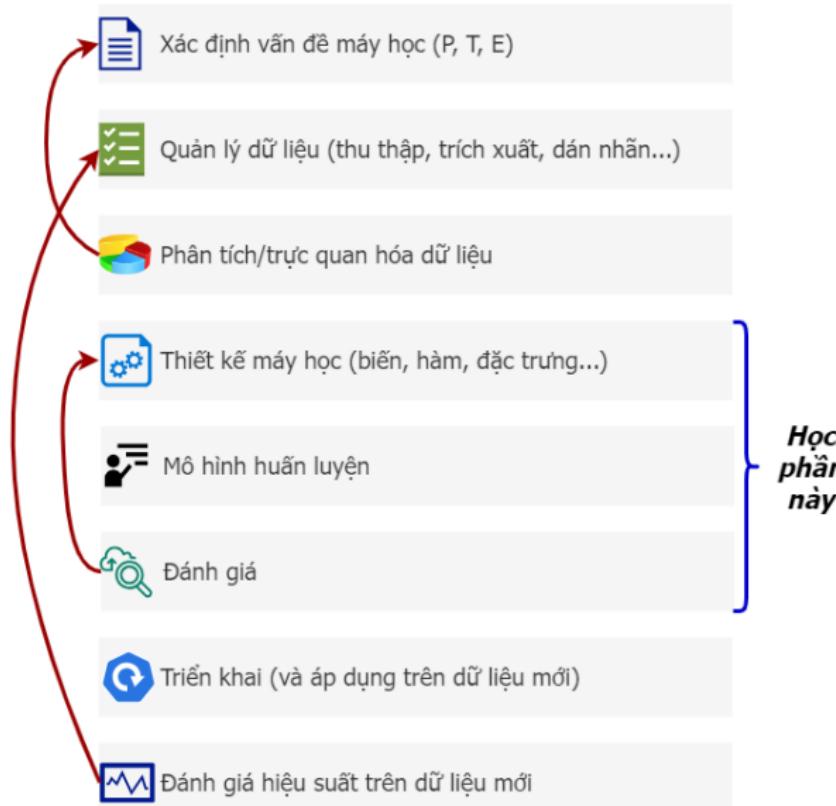


- Task T: ?
- Performance P: ?
- Experience E: ?

Máy học (Machine learning)



Máy học (Machine learning)



Tham số mô hình

- Mô hình ML được mô tả bởi các **tham số mô hình (model parameters)**
- → thuật toán ML: đi tìm các tham số mô hình phù hợp với bài toán thực tế đặt ra
- Tìm các tham số mô hình sao cho kết quả là tốt nhất
- Quan hệ giữa thước đo đánh giá và các tham số mô hình được mô tả thông qua hàm số gọi là **hàm mất mát (loss function)** và **hàm chi phí (cost function)**

Tham số mô hình và hàm mất mát

- Hàm mất mát có **giá trị nhỏ** \leftrightarrow phép đánh giá cho **kết quả tốt**
- Tìm tham số mô hình sao cho phép đánh giá trả về kết quả tốt \leftrightarrow **tối thiểu hàm mất mát**
- Quá trình xây dựng mô hình ML \leftrightarrow **giải bài toán tối ưu**
 \rightarrow quá trình **học (learning)** của **máy (machine)**

Tham số mô hình và hàm mất mát

- Bài toán tối thiểu hàm mất mát để tìm tham số mô hình được viết dưới dạng

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\theta)$$

Trong đó:

- θ : tham số mô hình
- $\mathcal{L}(\theta)$ hoặc $\mathcal{J}(\theta)$: hàm mất mát/hàm chi phí của mô hình
- $\arg \min_{\theta} \mathcal{L}(\theta)$: giá trị của tham số mô hình để hàm mất mát đạt giá trị nhỏ nhất

Khái niệm

- **Dữ liệu (data)**: đầu vào cho mô hình máy học. Tập hợp dữ liệu được gọi là **tập dữ liệu (dataset)**.
- Tập dữ liệu có thể được chia thành:
 - **Tập huấn luyện (training set)**: dữ liệu sử dụng để huấn luyện cho mô hình
 - **Tập kiểm chứng (validation set)**: dữ liệu đánh giá hiệu suất của mô hình
 - **Tập kiểm thử (test set)**: dữ liệu kiểm tra hiệu suất của mô hình, đóng vai trò như dữ liệu thực tế

Khái niệm

- **Đơn vị dữ liệu** (instance/ observation/ example/ sample): mỗi trường hợp được ghi nhận trong tập dữ liệu đầu vào
- **Mục tiêu** (target): đầu ra mong muốn dự đoán của mô hình máy học
- **Hàm mục tiêu** (target function): hàm ánh xạ (map) một dữ liệu đầu vào đến dữ liệu đầu ra
- **Tham số** (parameter): các tham số cần học của mô hình

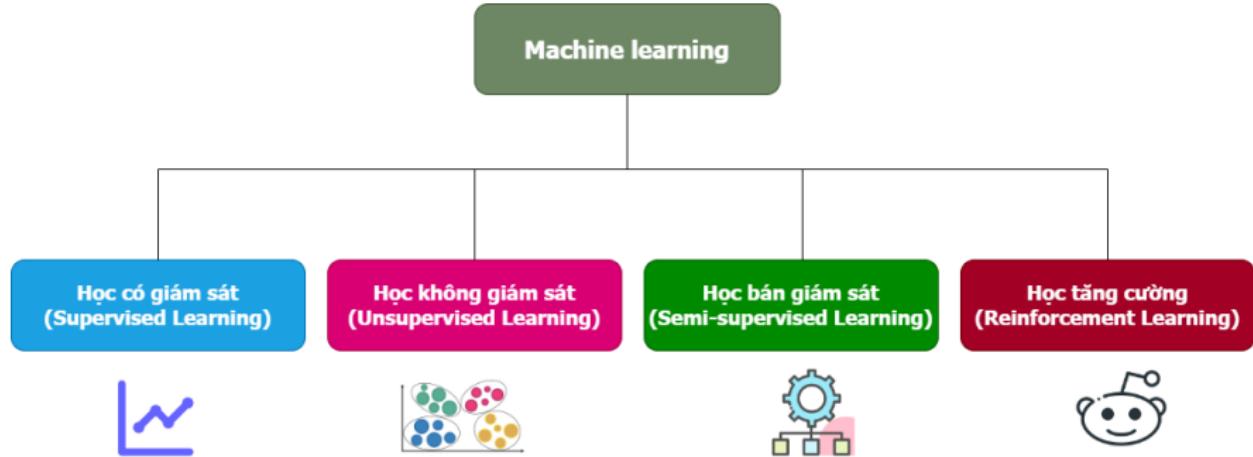
Khái niệm

- **Học/huấn luyện** (training): là quá trình mà mô hình sẽ cập nhật (update) các tham số mô hình dựa trên tập huấn luyện
- **Đánh giá/kiểm chứng** (validating): quá trình đánh giá sự thể hiện của mô hình
- **Kiểm thử** (test): quá trình kiểm tra thực tế sự thể hiện của mô hình
- **Thuộc tính/ đặc trưng** (attribute/ feature/ variable): mỗi đơn vị dữ liệu được thể hiện thông qua các thuộc tính

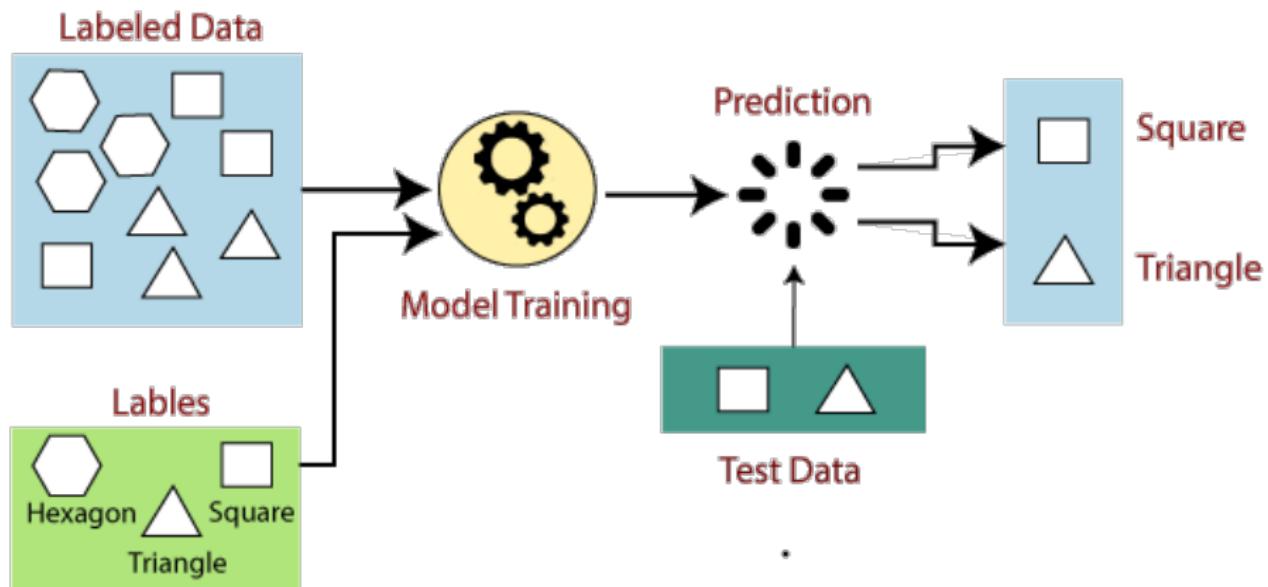
Khái niệm

- **Quá khớp** (overfitting): khi mô hình dự đoán chính xác trên tập huấn luyện nhưng thể hiện kém trên tập kiểm tra
- **Kiểm chứng chéo** (cross-validation): phương pháp chia dữ liệu thành các tập cho việc đánh giá mô hình
- **Không gian giả thuyết** (hypothesis space): tập các giả thuyết của hàm mục tiêu

Phân loại



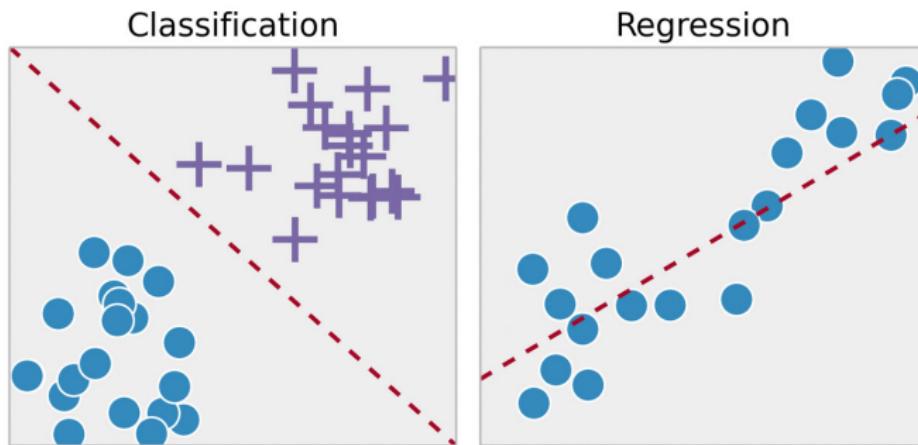
Học có giám sát



Học có giám sát

Hai nhóm bài toán cơ bản của học có giám sát:

- **Phân loại** (classification): biến đầu ra có giá trị rời rạc
- **Hồi quy** (regression): biến đầu ra có giá trị liên tục

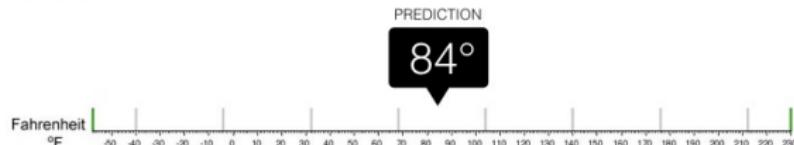


Học có giám sát



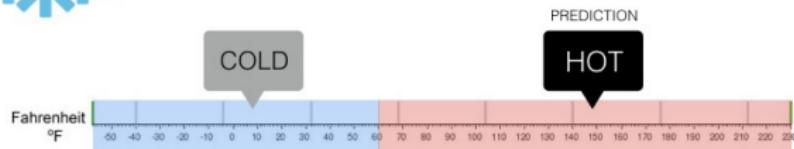
Regression

What is the temperature going to be tomorrow?



Classification

Will it be Cold or Hot tomorrow?



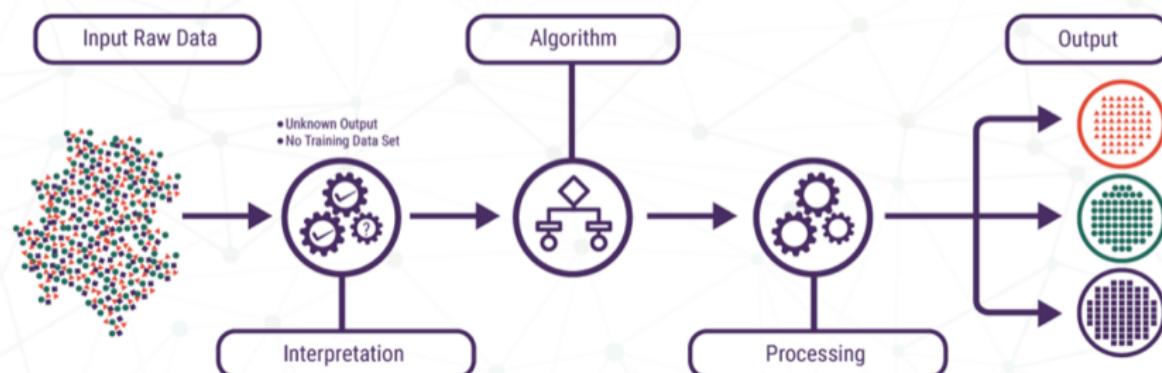
Học có giám sát

Nhận dạng chữ viết tay

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 9 9 9 9 9

Học không giám sát

UNSUPERVISED LEARNING



Học không giám sát

- Chỉ có dữ liệu đầu vào X (các vector đặc trưng) mà không biết giá trị đầu ra Y tương ứng
- Dựa vào cấu trúc của dữ liệu để thực hiện một công việc nào đó, ví dụ **phân nhóm** hoặc **giảm số chiều của dữ liệu** (dimensionality reduction)
- Sử dụng dữ liệu không có nhãn
- Hướng đến việc mô hình hóa được cấu trúc hay thông tin ẩn trong dữ liệu → thiên về việc mô tả tính chất hay đặc tính của dữ liệu

Học không giám sát

Thông thường, các thuật toán này dựa trên những thông tin sau:

- Mỗi quan hệ tương tự (similarity) giữa các ví dụ (instance) trong dữ liệu như trong các thuật toán phân cụm (clustering)
- Xác suất đồng xuất hiện của các đối tượng
- Các phép biến đổi ma trận để trích xuất các đặc trưng như PCA, SVD

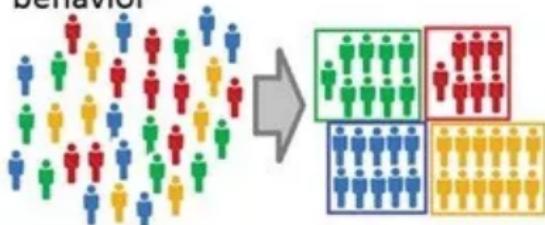
Học không giám sát

- **Phân cụm** (Clustering)
- **Kết hợp** (Association)

Unsupervised Learning

Clustering

Grouping customers by purchasing behavior



Association

People that buy X tend to buy Y

People that buy A+B tend to buy C



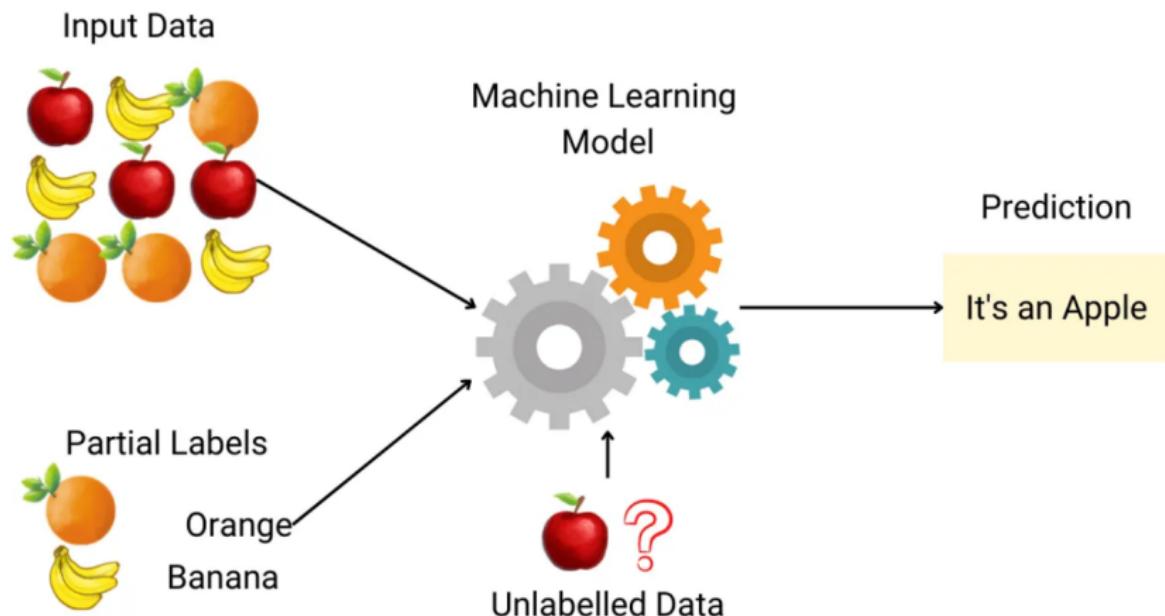
Học có giám sát và không giám sát

Tiêu chí	Học có giám sát	Học không giám sát
<i>Dữ liệu để huấn luyện mô hình</i>	Dữ liệu có nhãn	Dữ liệu không có nhãn
<i>Cách thức học của mô hình</i>	Mô hình hóa mối quan hệ giữa biến đầu vào và biến đầu ra	Học dựa trên các quan hệ tương tự, sự đồng xuất hiện hay các phép biến đổi ma trận

Học có giám sát và không giám sát

Tiêu chí	Học có giám sát	Học không giám sát
<i>Thuật toán</i>	Support Vector Machine Neural Network Linear Regression Logistic Regression Decision Tree ...	K-Means DBSCAN Apriori FP-Growth PCA SVD ...
<i>Kết quả</i>	Độ chính xác và tin cậy cao	Ít chính xác hơn

Học bán giám sát

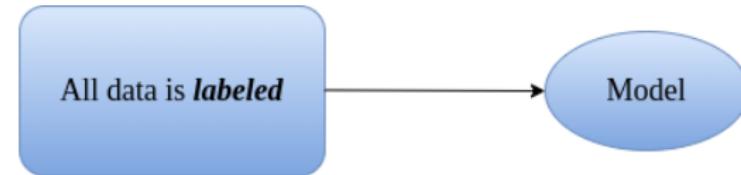


Học bán giám sát

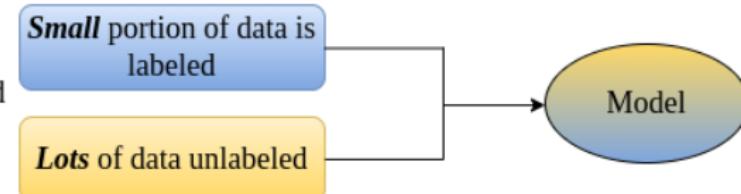
- Dữ liệu được dùng để huấn luyện bao gồm cả những **dữ liệu có nhãn** và **chưa được gán nhãn** (một lượng lớn dữ liệu X nhưng chỉ một phần được gán nhãn)
- Những bài toán thuộc nhóm này nằm giữa hai nhóm được nêu trên

Học bán giám sát

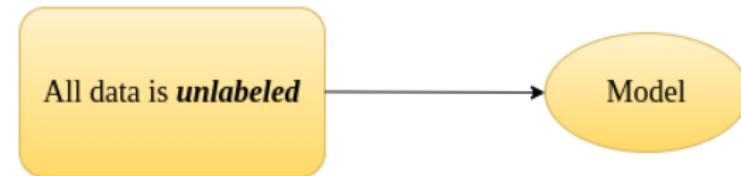
Supervised
Learning



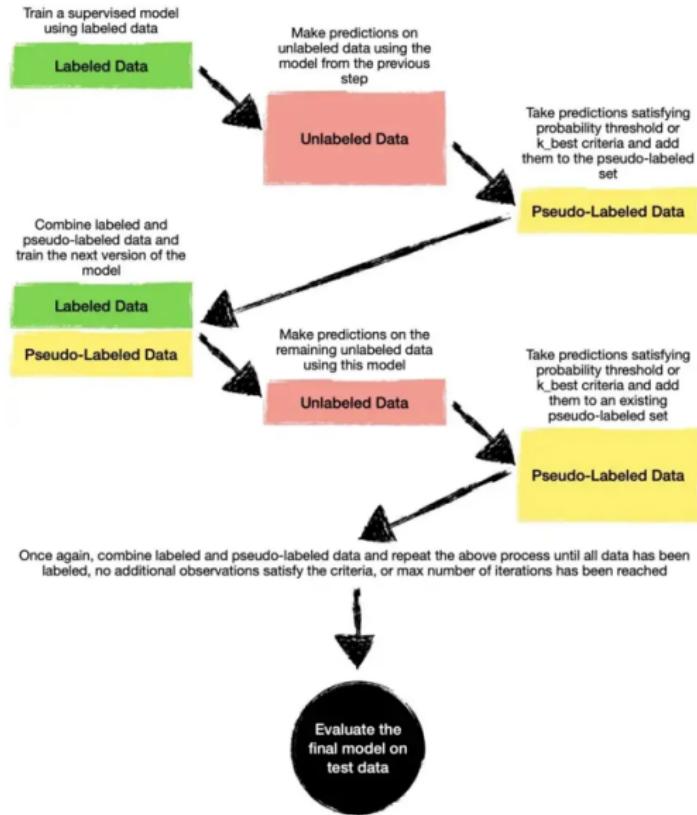
Semi-supervised
Learning



Unsupervised
Learning



Học bán giám sát



Học tăng cường

REINFORCEMENT LEARNING



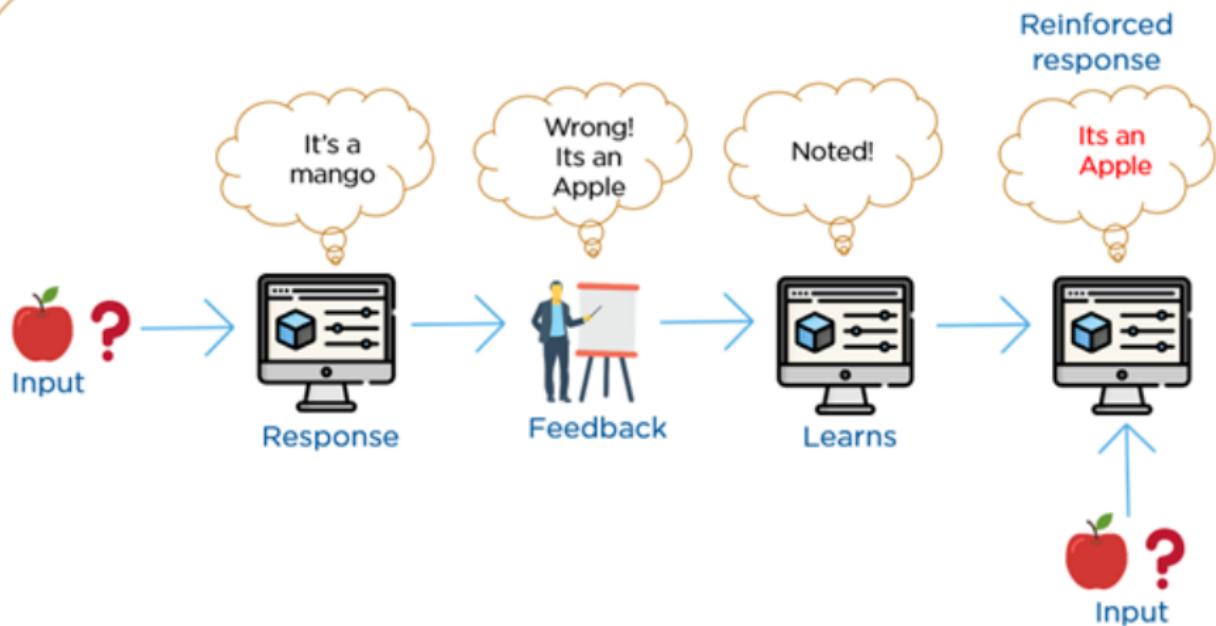
Học tăng cường

Học bằng cách **tương tác**, **trải nghiệm trực tiếp** với **môi trường xung quanh**, liên tục nhận **phản hồi** từ **môi trường** (nhận phần thưởng khi thực hiện đúng, bị phạt khi thực hiện sai) để tự **cải thiện hành vi** của hệ thống trong các môi trường mới.

Học tăng cường: có thể đo chất lượng hoặc cải tiến đầu ra theo mỗi đầu vào

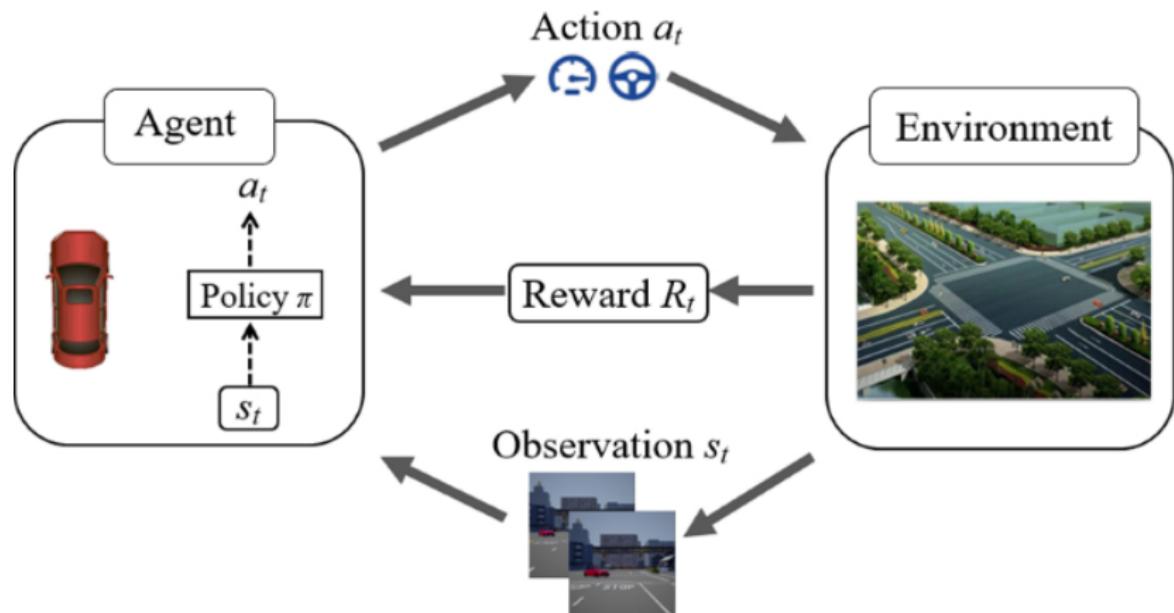
→ các bài toán giúp cho một hệ thống tự động xác định hành vi dựa trên hoàn cảnh để đạt được lợi ích cao nhất (maximizing the performance)

Học tăng cường

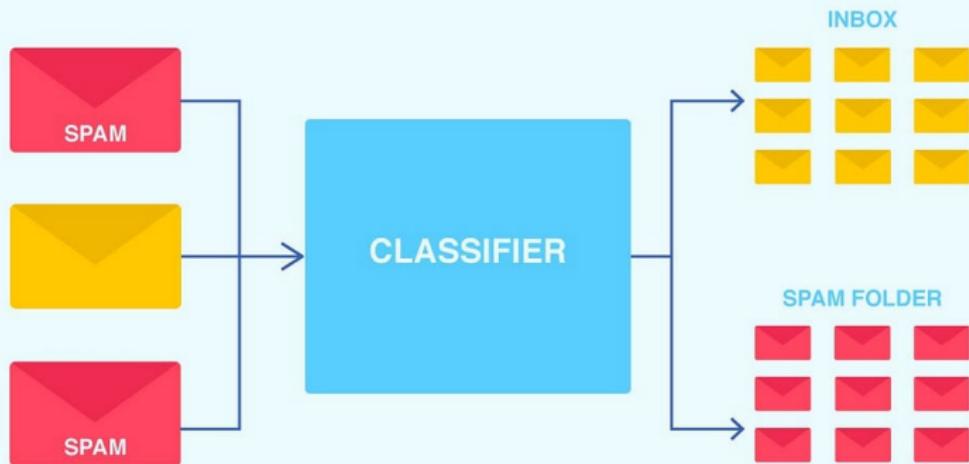


Học tăng cường

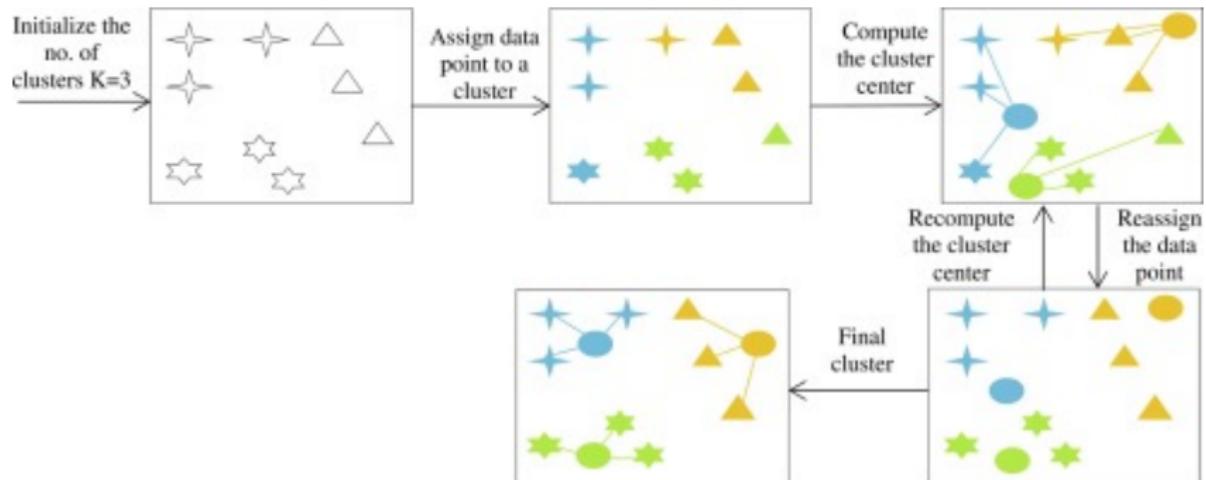
Xe tự lái được huấn luyện dựa ảnh nhận từ camera và điều khiển tay lái, tốc độ của xe



Phân loại thư rác

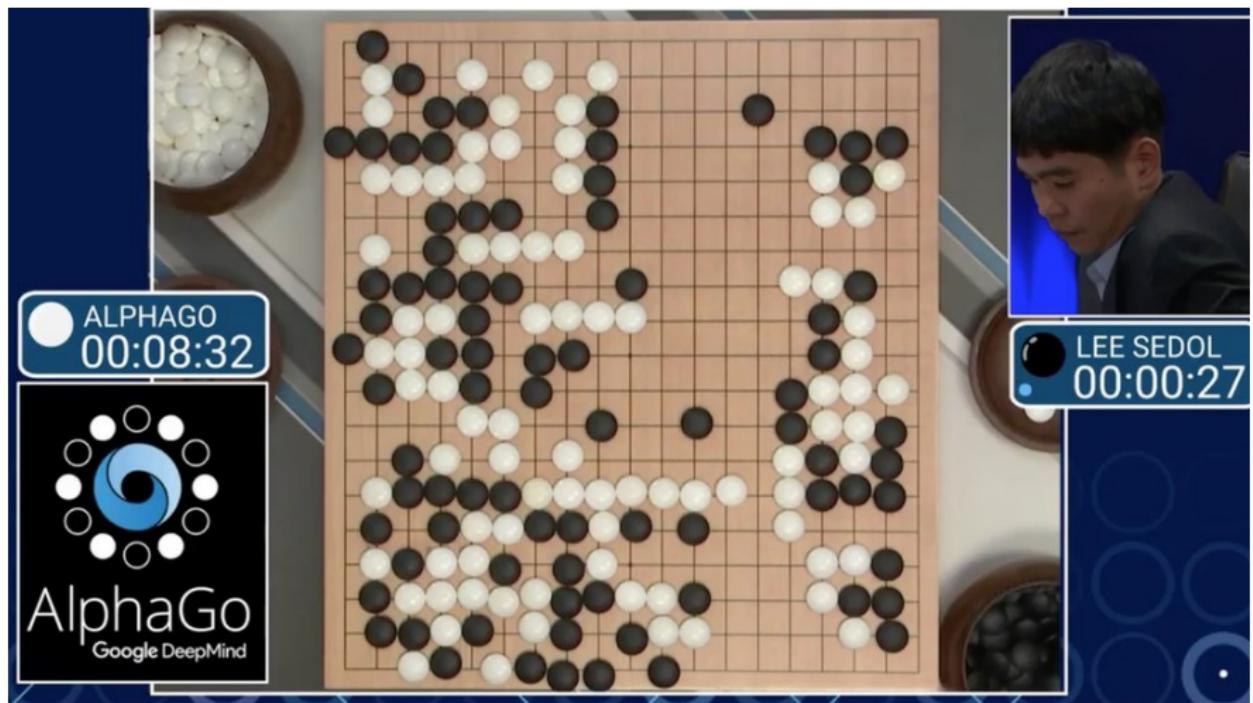


Phân cụm khách hàng



Áp dụng vào lý thuyết trò chơi

Áp dụng vào lý thuyết trò chơi (Game theory): xác định nước đi tiếp theo để đạt được điểm số cao nhất



Một số ứng dụng cơ bản

- Nhận dạng chữ viết tay

Prediction: altered

Prediction: Here

Prediction: ,

Prediction: the

Prediction: It

Prediction: u

Prediction: edges

Prediction: The

Prediction: position

Prediction: sceatched

Prediction: officer's

Prediction: to

Prediction: Ite

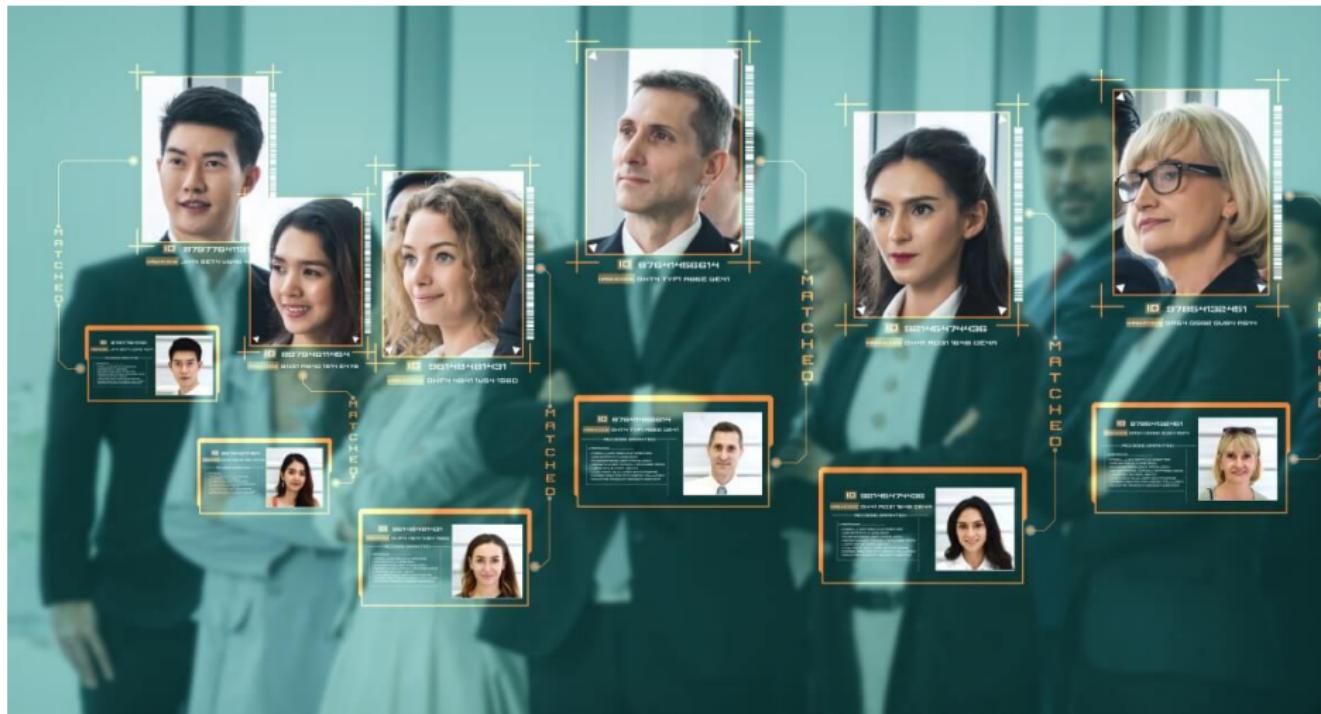
Prediction: wauted

Prediction: ,

Prediction: Gree

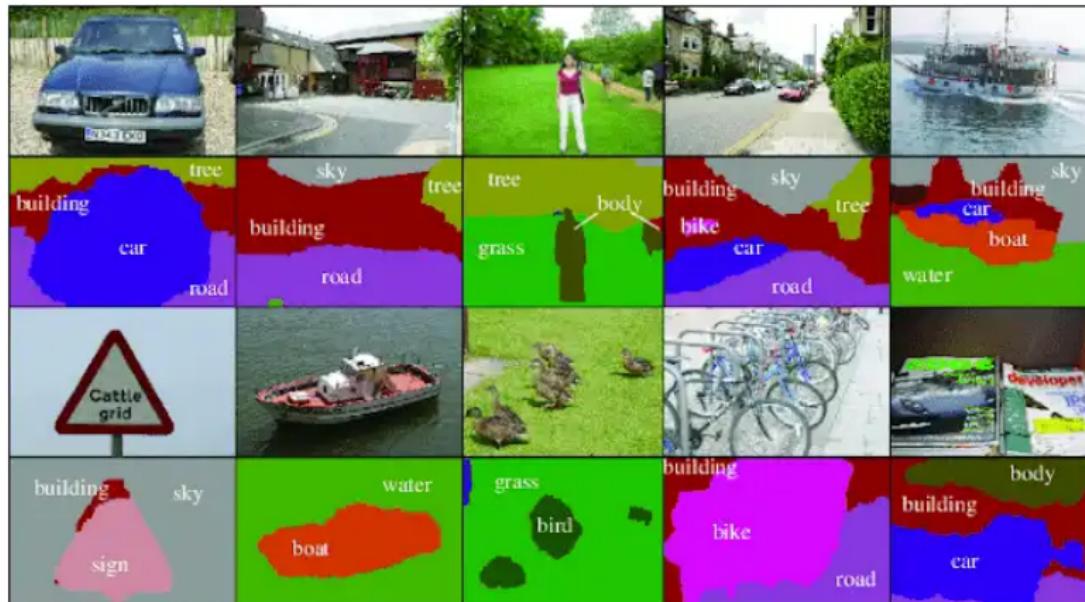
Một số ứng dụng cơ bản

- Nhận dạng khuôn mặt người



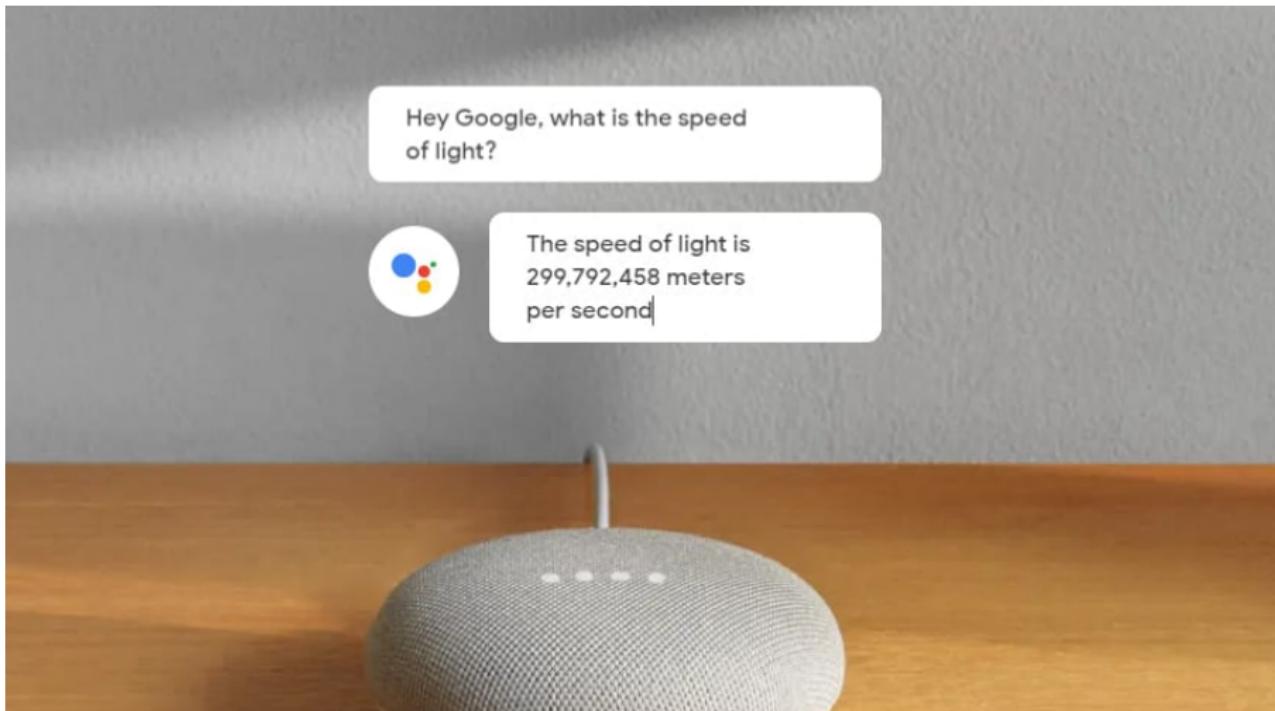
Một số ứng dụng cơ bản

- Phân đoạn ảnh



Một số ứng dụng cơ bản

- Trợ lý thông minh



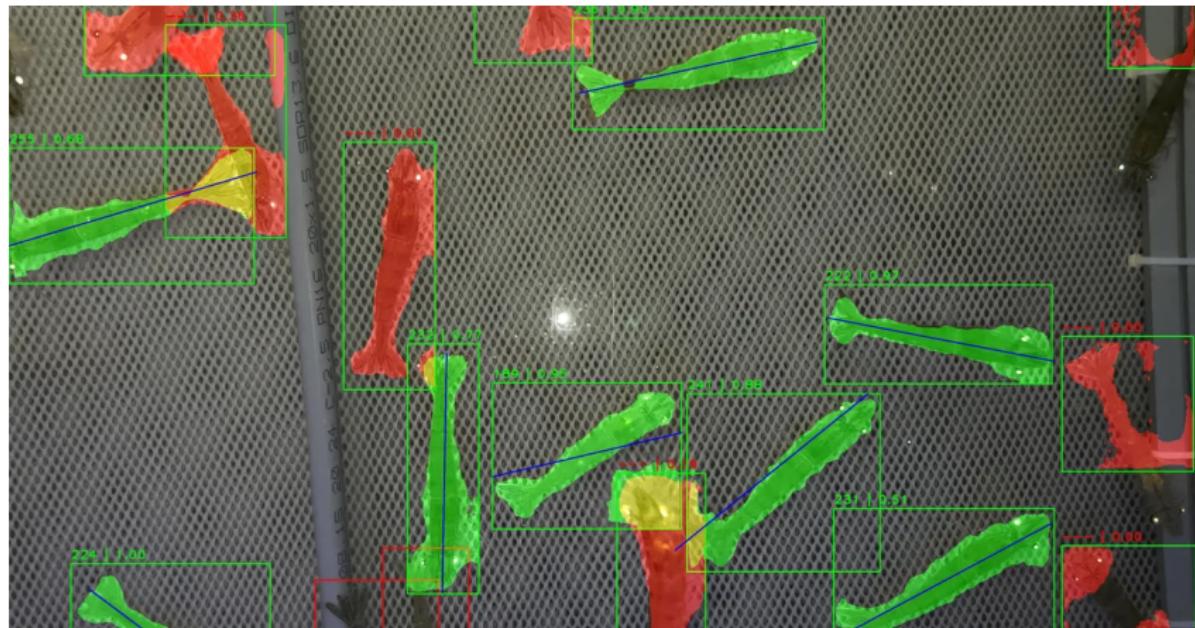
Một số ứng dụng cơ bản

- Trợ lý thông minh



Một số ứng dụng cơ bản

- Phát hiện tình trạng stress ở tôm nuôi



Một số ứng dụng cơ bản

• Gán chủ đề cho tài liệu

