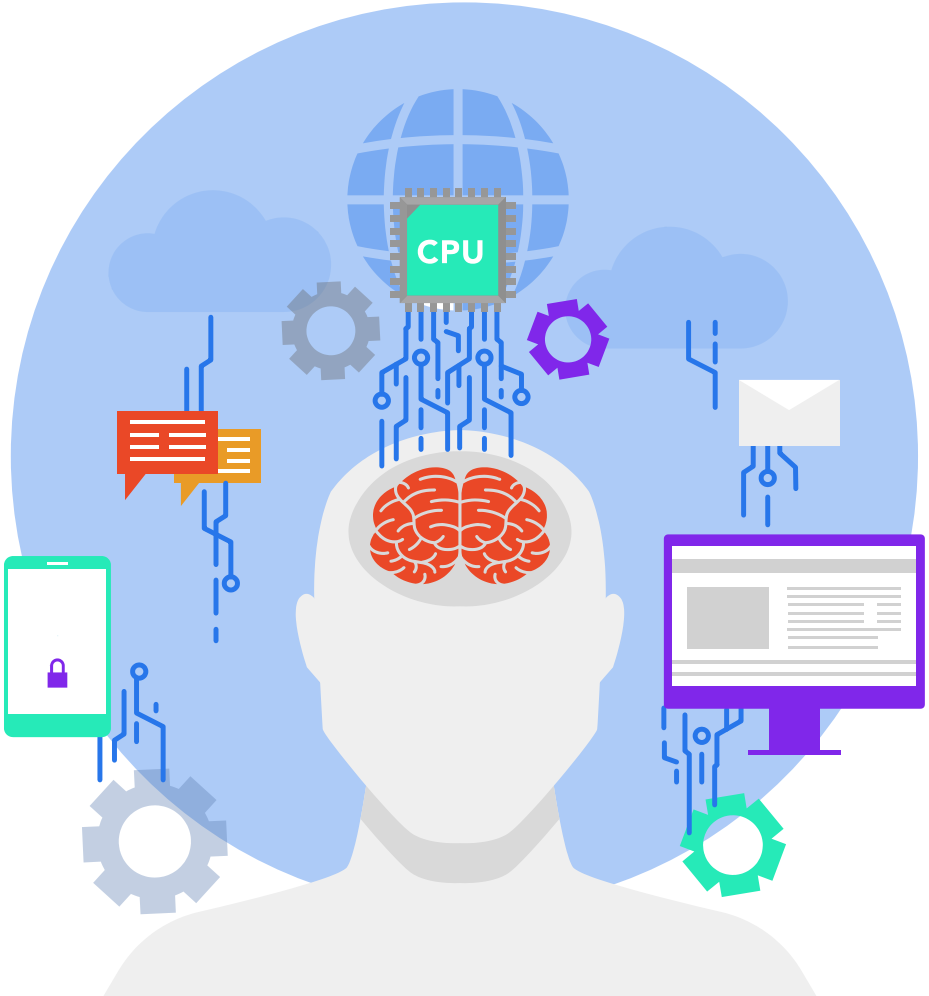


Chương 1: Tổng quan về học máy

1.1 Mở đầu

Giới thiệu

- **Học máy** (Machine Learning - ML) là lĩnh vực trí tuệ nhân tạo (AI) tập trung vào việc xây dựng hệ thống tự học từ dữ liệu.
- **Ứng dụng:** Nhận diện hình ảnh, xử lý ngôn ngữ tự nhiên, dự đoán giá nhà, xe tự lái.

Ý tưởng

Thay vì lập trình thủ công, máy học tự động tìm ra quy luật từ dữ liệu.

Tầm quan trọng

Tự động hóa, tối ưu hóa quyết định, khai phá dữ liệu lớn

1.1 Mở đầu

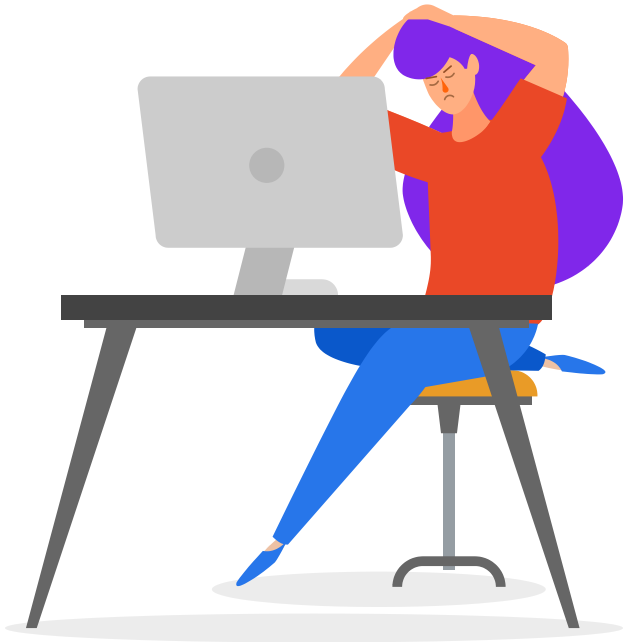
Lịch sử phát triển

1950s: Khái niệm AI và ML được giới thiệu (Alan Turing, Samuel)

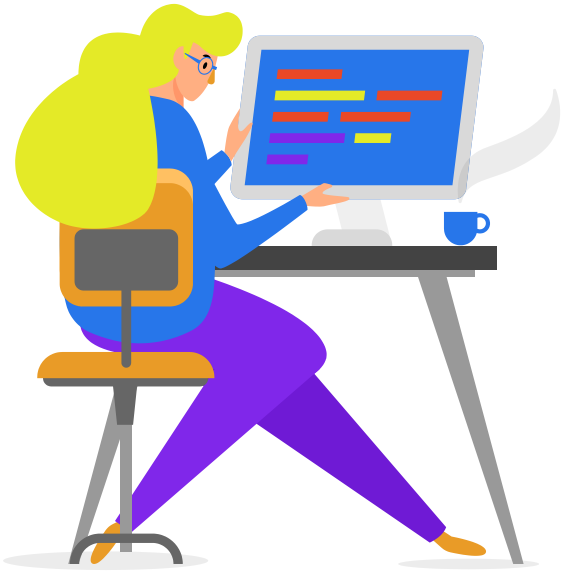
1980s-1990s: Phát triển các thuật toán như cây quyết định, mạng nơ-ron

2000s-nay: Bùng nổ nhờ dữ liệu lớn (Big-data), GPU và deep learning

Tương lai: ML tiếp tục định hình công nghệ (Xe tự lái, y học, …)



1.2 Các khái niệm cơ bản – Định nghĩa



01

Học máy

Máy tính tự động cải thiện hiệu suất dựa trên kinh nghiệm (dữ liệu).

02

Dữ liệu

Nguồn thông tin để máy học (ví dụ: văn bản, hình ảnh, số liệu).

03

Mô hình

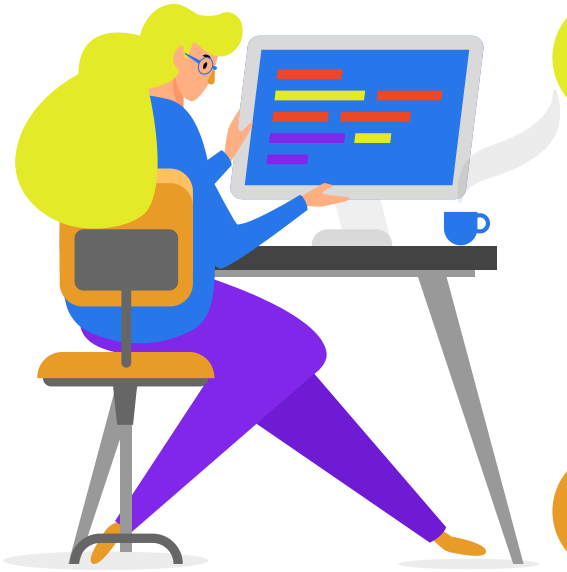
Hàm toán học ánh xạ đầu vào (input) thành đầu ra (output).

04

Thuật toán

Quy trình tối ưu hóa mô hình dựa trên dữ liệu.

1.2 Các khái niệm cơ bản – Phân loại



05

Học có giám sát

(Supervised Learning)
Học từ dữ liệu có nhãn

06

Học không giám sát

(Unsupervised Learning)
Tìm cấu trúc trong dữ liệu không nhãn

07

Học tăng cường

(Reinforcement Learning)
Học qua và thử sai

08

Học nửa giám sát, học chuyển tiếp

Kết hợp hoặc chuyển giao kiến thức

1.3 Quy trình xây dựng hệ thống học máy

01 **Thu thập dữ liệu**

Tạo hoặc thu thập tập dữ liệu chất lượng cao, phù hợp với bài toán.

02 **Tiền xử lý dữ liệu**

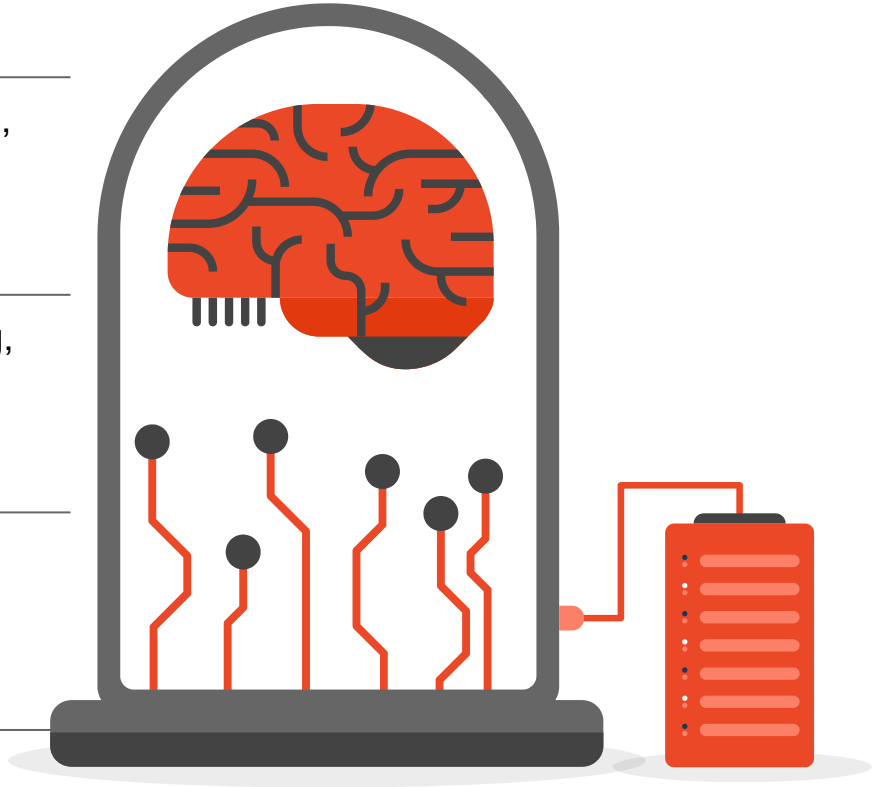
Làm sạch, chuẩn hóa, và trích xuất đặc trưng, chuyển đổi dữ liệu.

03 **Chọn mô hình**

Lựa chọn thuật toán phù hợp
VD: hồi quy, cây quyết định, mạng nơ-ron.

04 **Huấn luyện**

Tối ưu hóa mô hình mô hình trên tập dữ liệu huấn luyện.



1.3 Quy trình xây dựng hệ thống học máy

05 Đánh giá mô hình

Sử dụng tập kiểm tra để đo độ chính xác (accuracy, F1-score, v.v.).

06 Tinh chỉnh (Tuning)

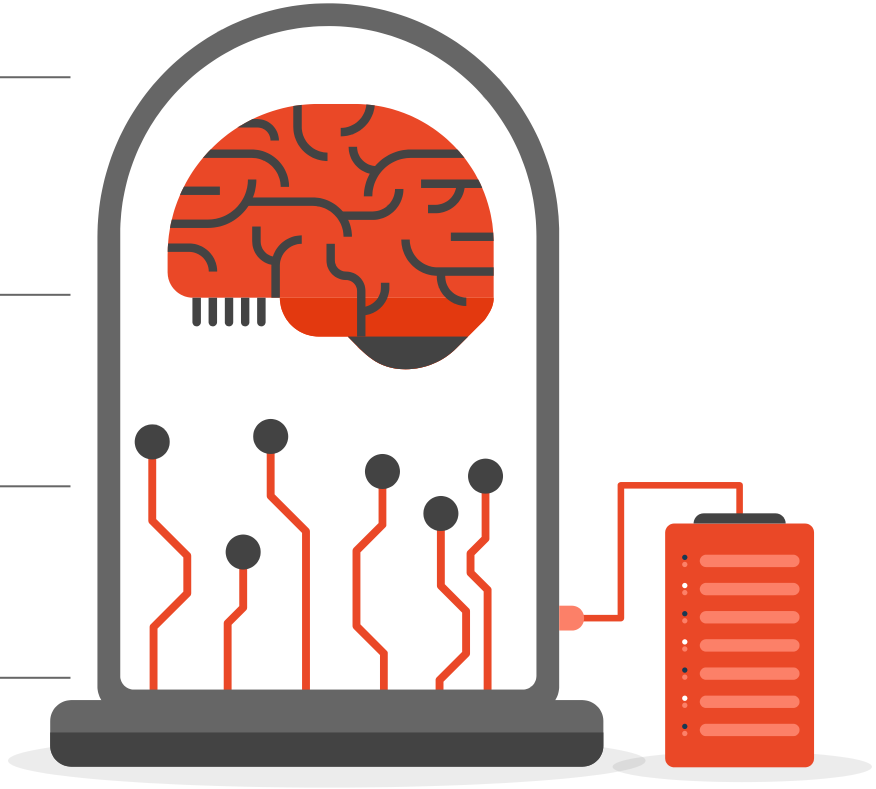
Điều chỉnh siêu tham số (hyperparameters).

07 Triển khai (Deployment)

Đưa mô hình vào thực tế.

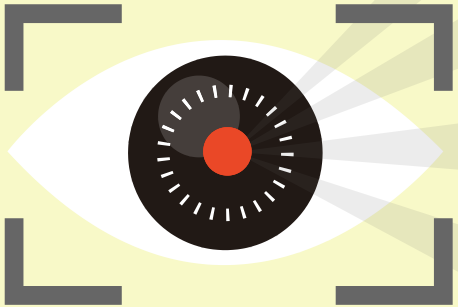
08 Giám sát và bảo trì

Theo dõi hiệu suất, cập nhật mô hình



1.4 Ví dụ về học máy

Ứng dụng học máy



Phân loại sản phẩm

Hệ thống học từ dữ liệu sản phẩm và hành vi người dùng để gợi ý sản phẩm phù hợp. (Shoppe, Lazada)

Lọc thư rác email

Phân loại email thành "thư rác" hoặc "thư hợp lệ" dựa trên nội dung và người gửi.

Chẩn đoán bệnh từ hình ảnh y tế

Phân tích ảnh X-quang, MRI để phát hiện dấu hiệu bệnh. (Phát hiện khối u)

Xe tự lái

Xe học cách nhận diện làn đường, biển báo, người đi bộ để điều khiển an toàn.

1.5 Học có giám sát – Định nghĩa

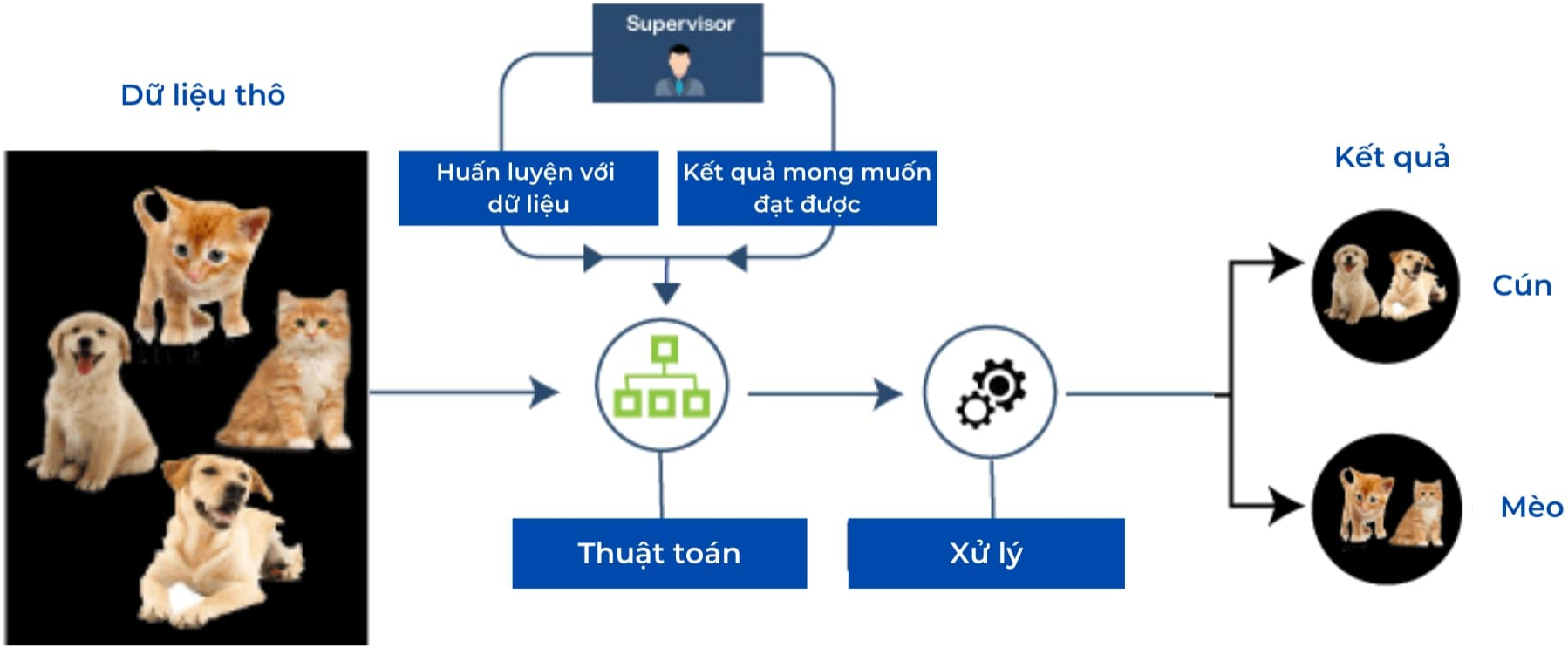
Học có giám sát

- **Định nghĩa:** Mô hình học từ dữ liệu có nhãn (input-output pairs)
- **Dữ liệu:** Mỗi mẫu có đặc trưng (features) và nhãn (labels)
- **Mục tiêu:** Dự đoán nhãn cho dữ liệu mới
- **Loại:**

  Phân loại: Dự đoán nhãn rời rạc (VD: spam/không spam)

  Hồi quy: Dự đoán giá trị liên tục (VD: giá nhà)
- **Ví dụ:** Hồi quy tuyến tính, SVM, mạng nơ-ron.

1.5 Học có giám sát – Định nghĩa



1.5 Học có giám sát – Thuật toán

Thuật toán học có giám sát

- **Hồi quy tuyến tính (Linear Regression):** Dự đoán giá trị liên tục
- **Hồi quy logistic (Logistic Regression):** Phân loại nhị phân
- **Máy vector hỗ trợ (SVM):** Phân loại với ranh giới tối ưu
- **Rừng ngẫu nhiên (Random Forest):** Kết hợp nhiều cây quyết định
- **Mạng nơ-ron (Neural Networks):** Xử lý bài toán phức tạp

1.6 Học nửa giám sát

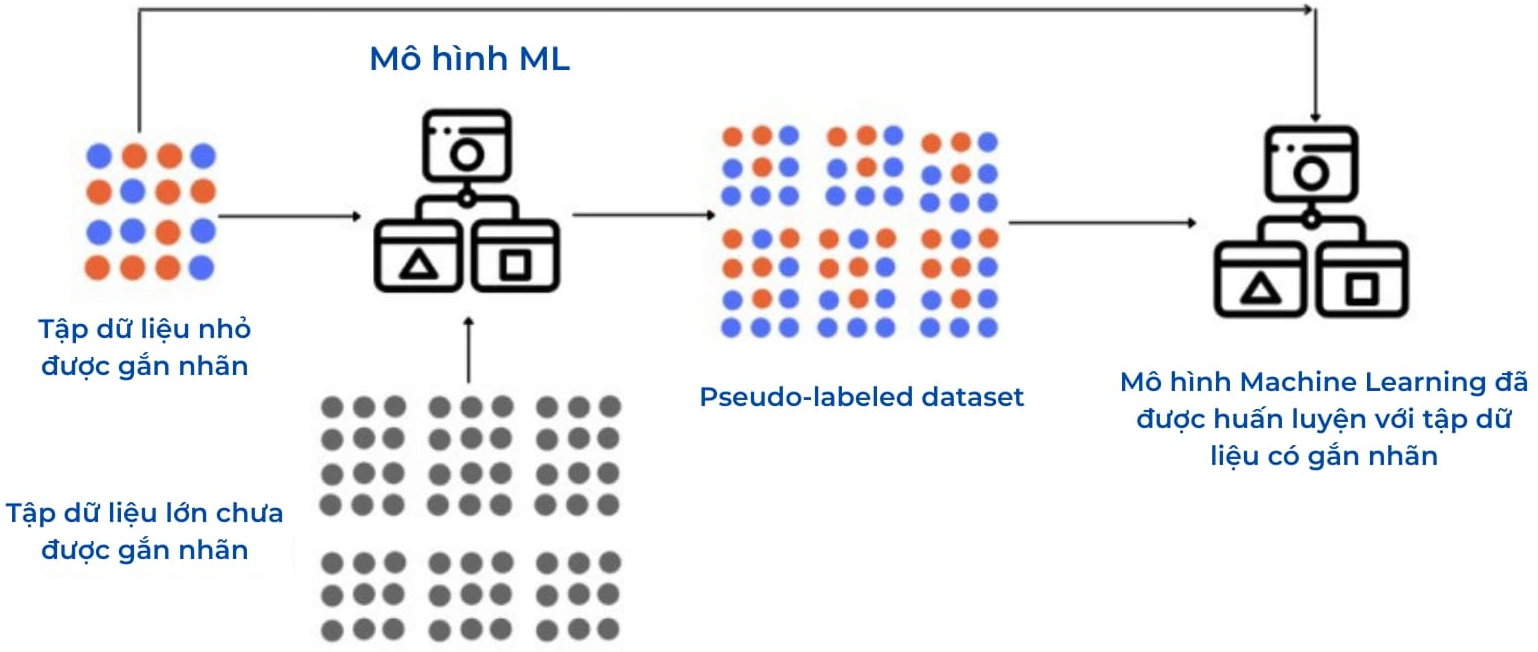
Học nửa giám sát

- **Định nghĩa:** Kết hợp dữ liệu có nhãn và không nhãn.
- **Ứng dụng khi:** Dữ liệu có nhãn ít, dữ liệu không nhãn nhiều.
- **Phương pháp:** Sử dụng dữ liệu không nhãn để cải thiện mô hình.
- **Ứng dụng:**
  Phân loại văn bản khi chỉ một phần dữ liệu được gắn nhãn
  Nhận diện đối tượng trong ảnh với nhãn hạn chế

Thuật toán học nửa giám sát

- Self-training, Co-training, Graph-based methods

1.6 Học nửa giám sát

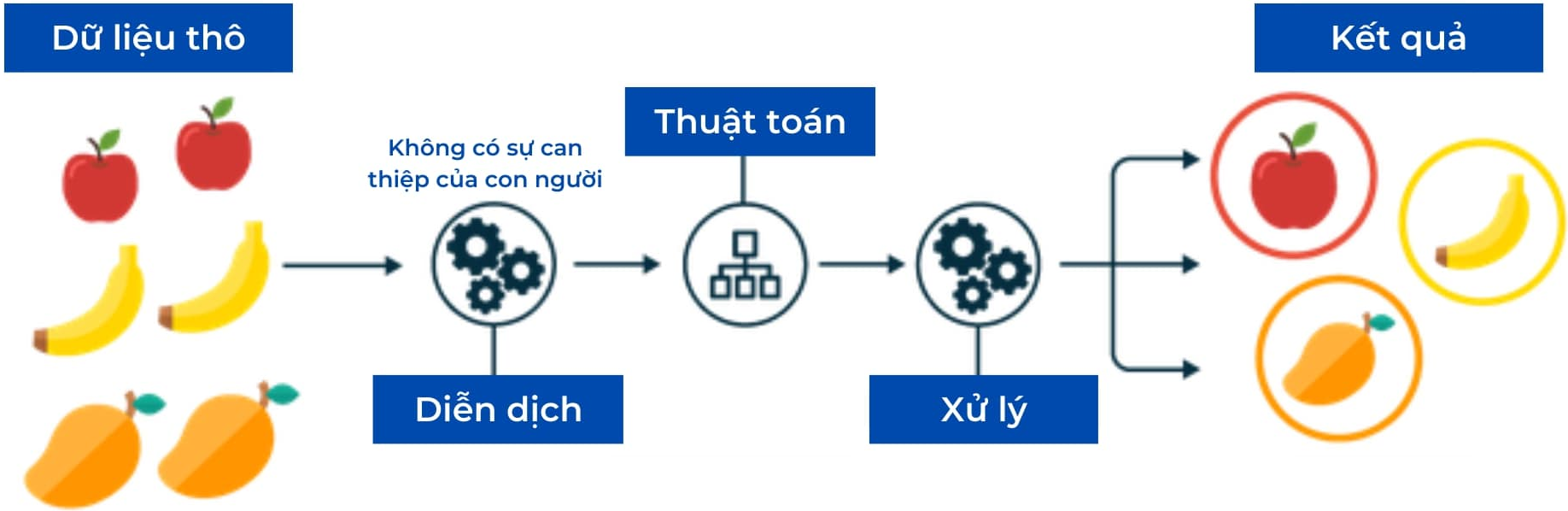


1.7 Học không có giám sát – Định nghĩa

Học không có giám sát

- **Định nghĩa:** Tìm cấu trúc ẩn trong dữ liệu không nhãn
- **Mục tiêu:** Phân cụm (clustering) hoặc giảm chiều (dimensionality reduction).
- **Ví dụ:**
  Phân cụm khách hàng dựa trên hành vi mua sắm.
  Giảm chiều dữ liệu để trực quan hoá

1.7 Học không có giám sát – Định nghĩa



1.7 Học không có giám sát – Thuật toán

Thuật toán học không có giám sát

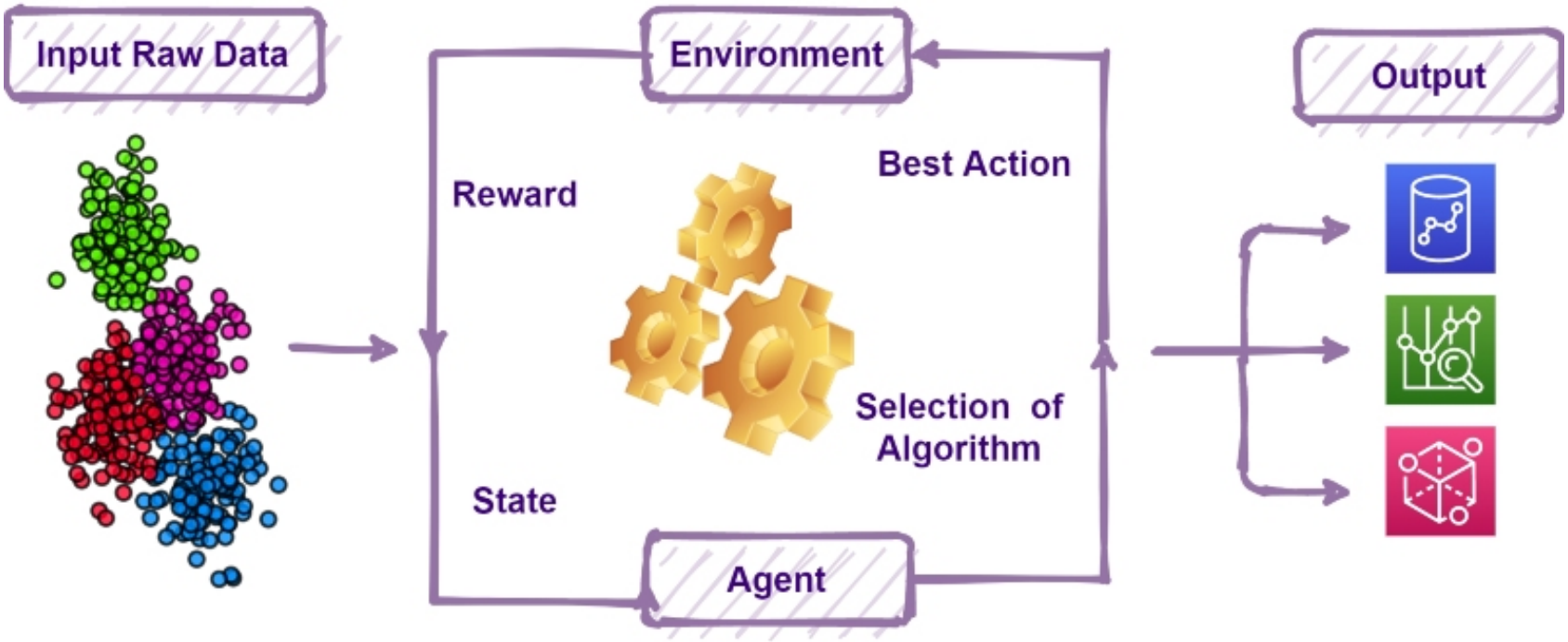
- **K-means:** Phân cụm dữ liệu thành k nhóm
- **PCA (Principle Component Analysis):** Giảm chiều dữ liệu
- **Autoencoders:** Mạng nơ-ron để học biểu diễn dữ liệu
- **DBSCAN:** Phân cụm dựa trên mật độ
- **Ứng dụng:** Khám phá xu hướng, phân đoạn thị trường

1.8 Học tăng cường – Định nghĩa

Học tăng cường

- **Định nghĩa:** Máy học qua thử và sai, dựa trên phần thưởng
- **Thành phần:**
  Tác nhân (Agent): Ra quyết định
  Môi trường (Enviroment): Tương tác với tác nhân
  Hành động (Action): Những gì mà tác nhân thực hiện trong môi trường
  Phần thưởng (Reward): Định hướng học
- Ví dụ: Chơi cờ vua (AlphaGo), điều khiển robot.

1.8 Học tăng cường – Định nghĩa



1.8 Học tăng cường – Thuật toán

Thuật toán học tăng cường

- **Q-Learning:** Tìm chính sách tối ưu dựa trên bảng Q
- **Deep Q-Network (DQN):** Kết hợp Q-Learning với mạng nơ-ron
- **Policy Gradient:** Tối ưu hoá trực tiếp chính sách hành động
- **Ứng dụng:** Trò chơi điều khiển robot, tối ưu hoá hệ thống

1.9 Học chuyển tiếp

Học chuyển tiếp

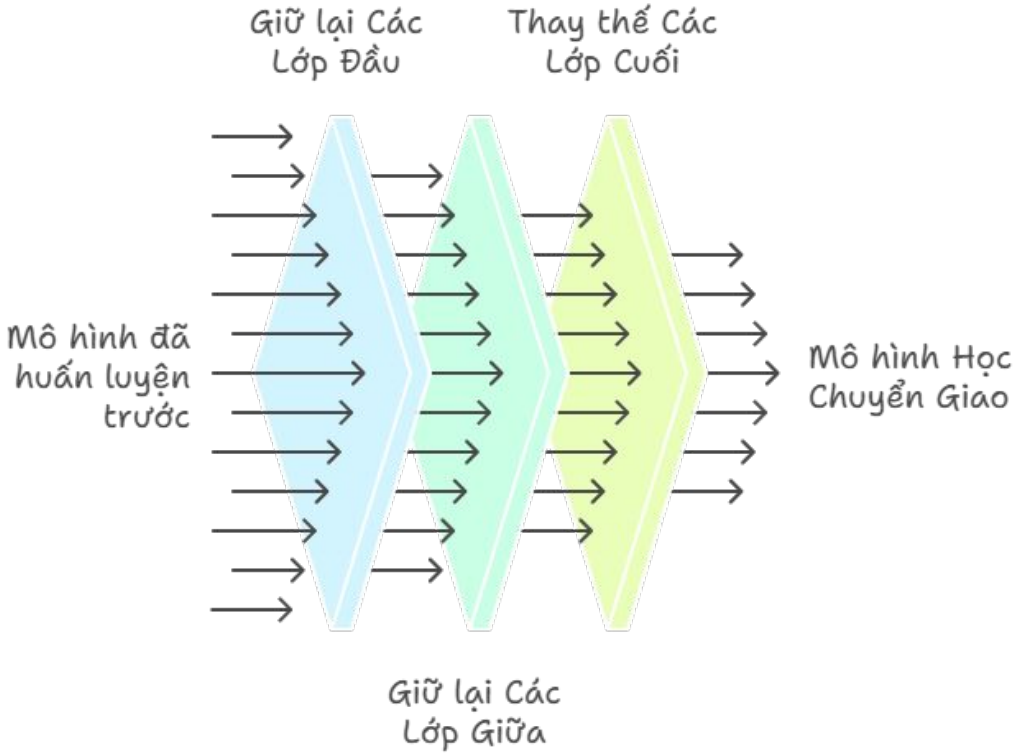
- **Định nghĩa:** Sử dụng kiến thức từ một tác vụ để cải thiện tác vụ khác.
- **Lợi ích:** Tiết kiệm thời gian huấn luyện, cải thiện hiệu suất với dữ liệu ít
- **Ứng dụng:** Khi dữ liệu cho tác vụ mới bị hạn chế
- **Ví dụ:**

- Sử dụng mô hình đã huấn luyện trên ImageNet cho nhận diện ảnh y khoa

- BERT trong xử lý ngôn ngữ tự nhiên

Phương pháp: Fine-tuning, Feature extraction

1.9 Học chuyển tiếp



1.10 Tiền xử lý dữ liệu

Tiền xử lý dữ liệu

Bước đầu quan trọng trong xử lý dữ liệu. Mục tiêu là biến dữ liệu thô thành dữ liệu sạch, nhất quán

Làm sạch
Xử lý giá trị thiếu, loại bỏ nhiễu

01

Chuyển đổi
Chuẩn hóa, mã hóa biến phân loại.

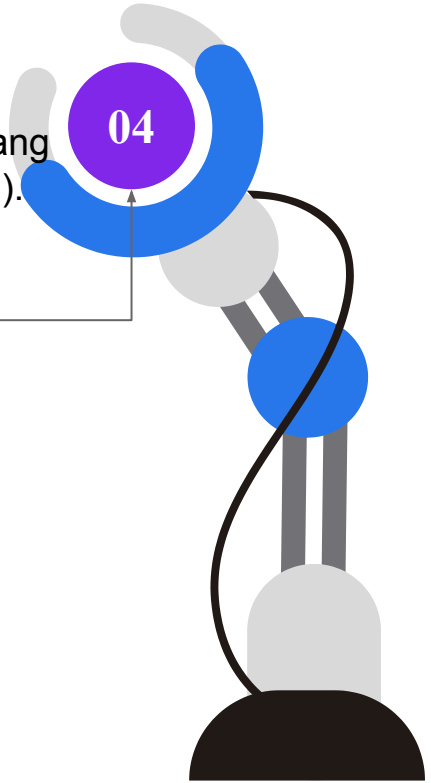
02

Trích xuất đặc trưng
Tạo đặc trưng mới từ dữ liệu thô.

03

Chia tỷ lệ
Đưa dữ liệu về cùng thang đo (VD: Min-Max scaling).

04



1.11 Một số kiến thức toán học bổ trợ

01

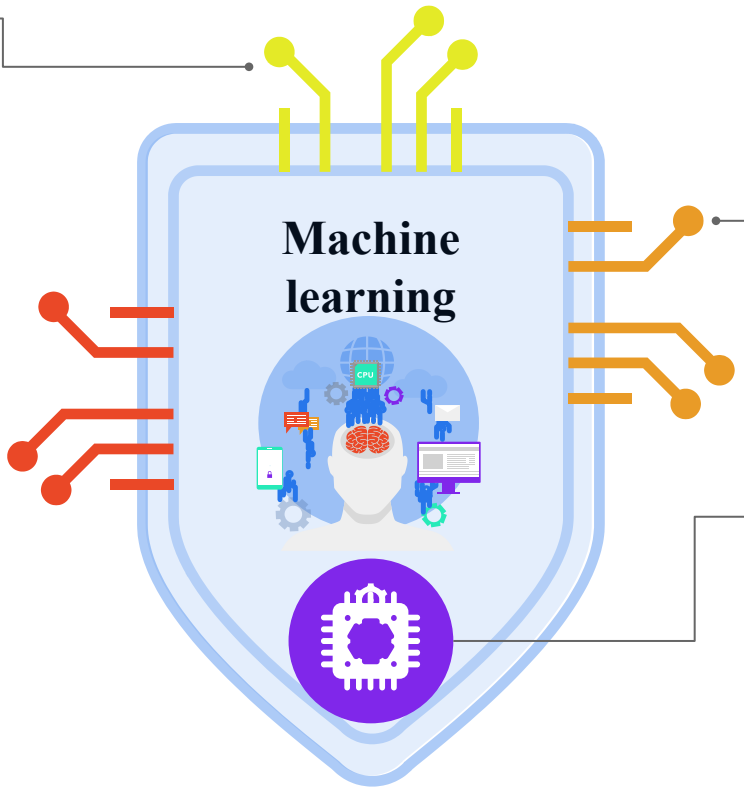
Đại số tuyến tính

Ma trận, vector, phép nhân ma trận
Ứng dụng: mạng nơ-ron, PCA

02

Xác suất thống kê

Phân phối, kỳ vọng, phương sai
Ứng dụng: Naive Bayes, đánh giá mô hình



03

Toán tối ưu

Hàm mất mát, thuật toán tối ưu (SGD, Adam)

04

Giải tích

Đạo hàm, tích phân trong tối ưu hoá, Gradient descent
Ứng dụng: Tối ưu hoá mô hình