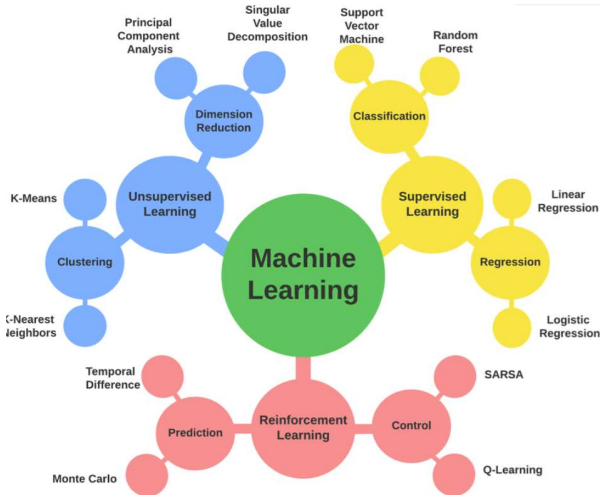
Các phương pháp học máy
Machine learning methods

4TC: 2 LT – 2 TH

Giảng viên: **Tạ Hoàng Thắng**

tahoangthang@gmail.com

0975399307



Các phương pháp học máy
Machine learning methods

Thang đánh giá

- 50% giữa kỳ (tham gia lớp học, nộp bài thực hành, kiểm tra cá nhân)
- 50% cuối kỳ, nộp & báo cáo đồ án
  - 3-4 bạn 1 nhóm
  - Thực hiện 3-4 tác vụ học máy

Các phương pháp học máy
Machine learning methods

Tài liệu tham khảo:

- Introduction to Machine Learning with Python: A Guide for Data Scientists
- **An Introduction to Machine Learning**
- Machine Learning, Vũ Hữu Tiệp, https://github.com/tiepvupsu/ebookMLCB/blob/master/book_ML_color.pdf
- Machine Learning: https://www.w3schools.com/python/python_ml_getting_started.asp
- Matplotlib: https://www.w3schools.com/python/matplotlib_intro.asp
- Đại số tuyến tính cơ bản

1. Các khái niệm & Thuật ngữ

Học máy

- IBM:
  - Machine learning (ML) is a **branch of artificial intelligence (AI) and computer science** that focuses on the **using data and algorithms to enable AI to imitate the way that humans learn**, gradually improving its accuracy.
    - https://www.ibm.com/topics/machine-learning
- Wikipedia:
  - Machine learning (ML) is **a field of study in artificial intelligence** concerned with **the development and study of statistical algorithms that can learn from data and generalize to unseen data** and thus perform tasks without explicit instructions.
    - https://en.wikipedia.org/wiki/Machine_learning

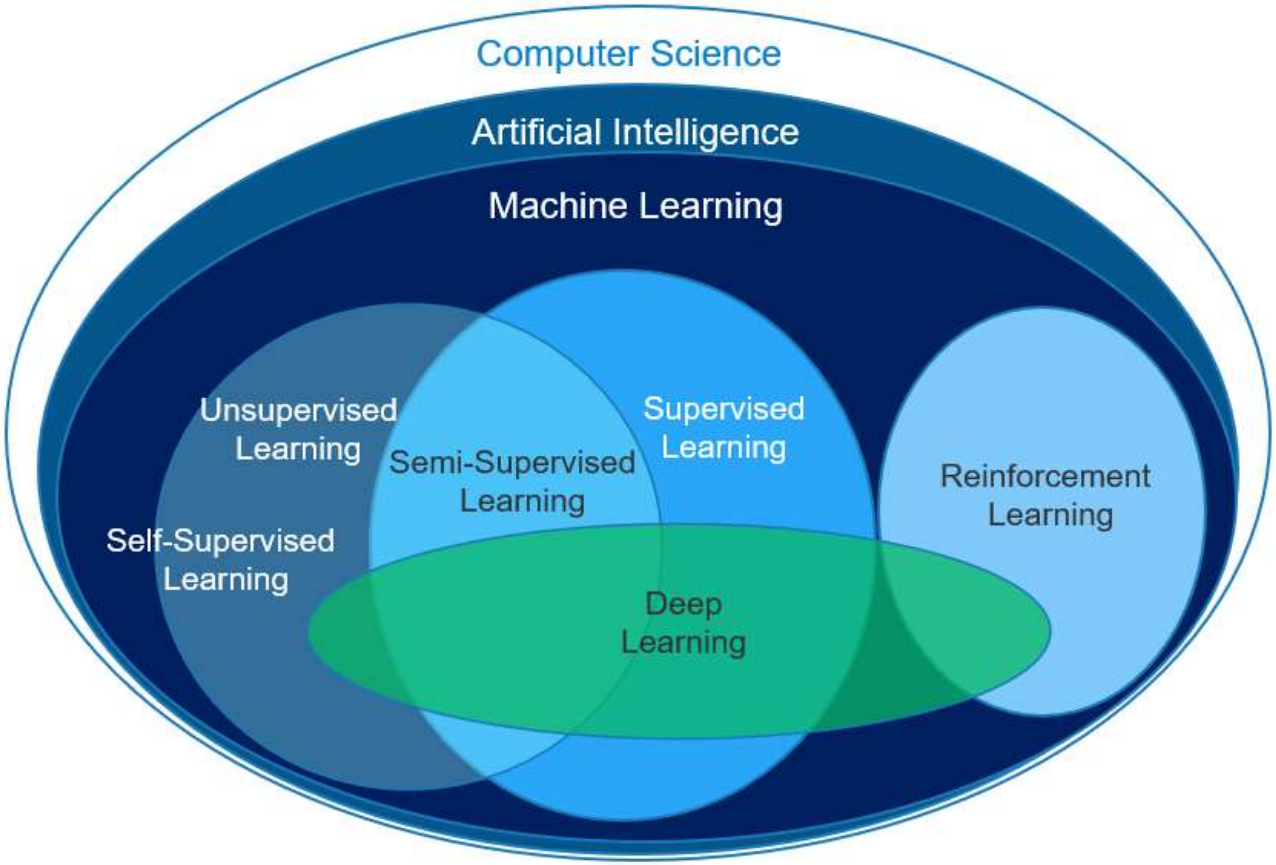
1. Các khái niệm & Thuật ngữ

Học máy

- ChatGPT: **là một lĩnh vực trong trí tuệ nhân tạo (AI)** tập trung vào việc phát triển các thuật toán và mô hình cho phép máy tính học hỏi từ dữ liệu và cải thiện hiệu suất của chúng mà không cần được lập trình rõ ràng.

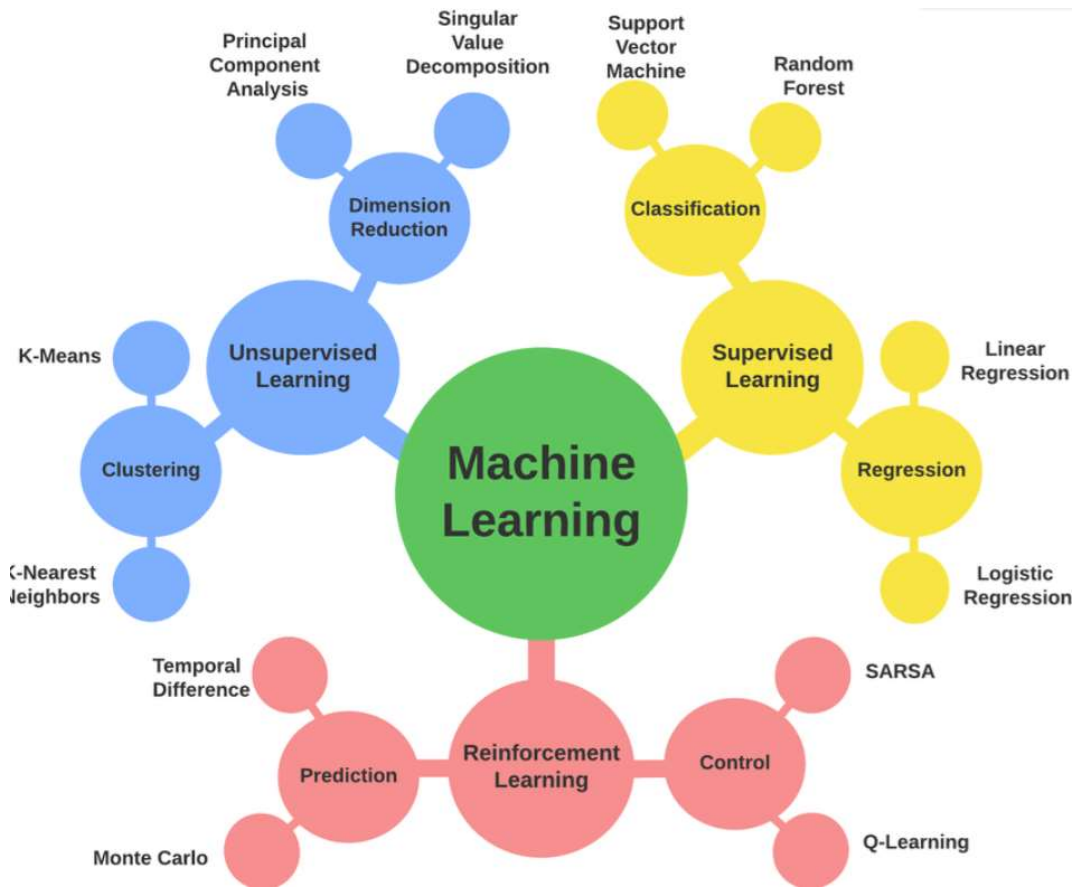
1. Các khái niệm & Thuật ngữ

Học máy



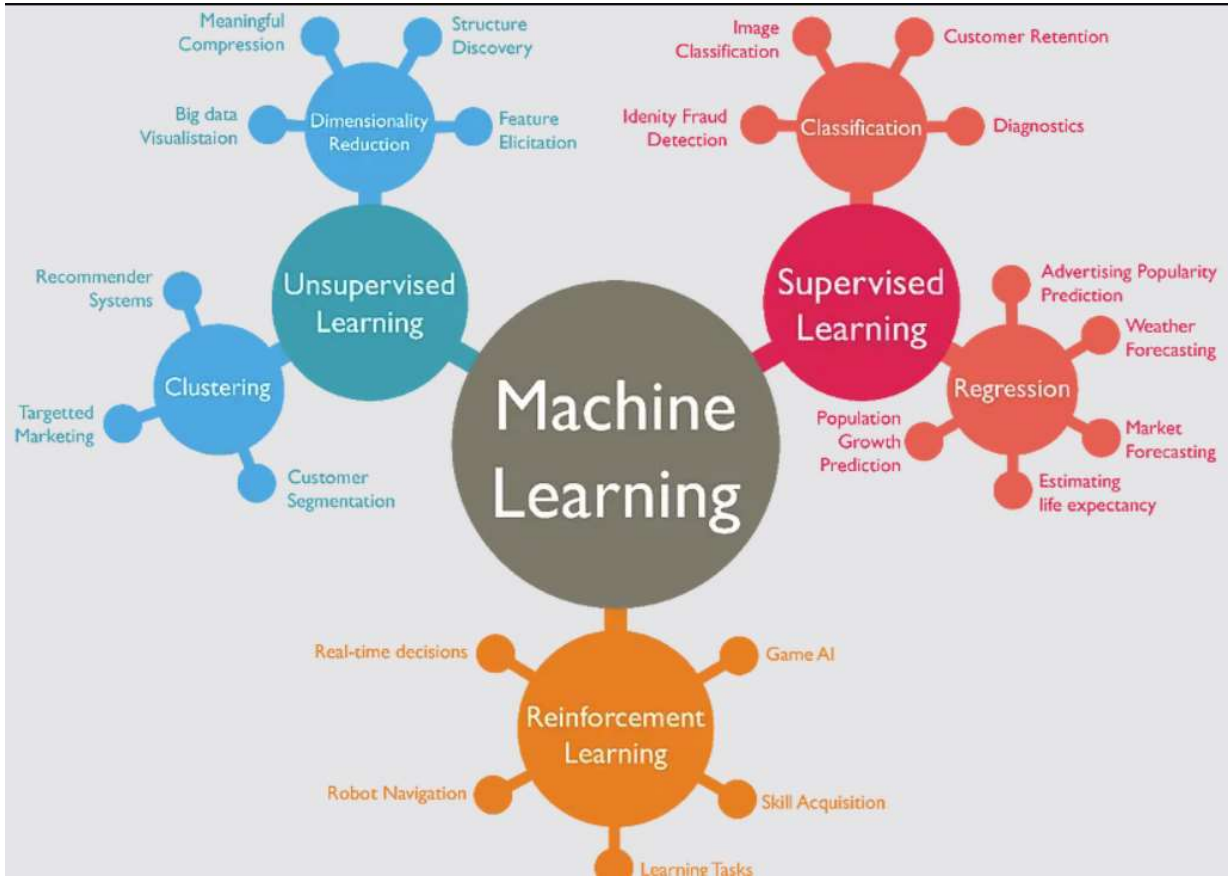
1. Các khái niệm & Thuật ngữ

Các thuật ngữ phổ biến



1. Các khái niệm & Thuật ngữ

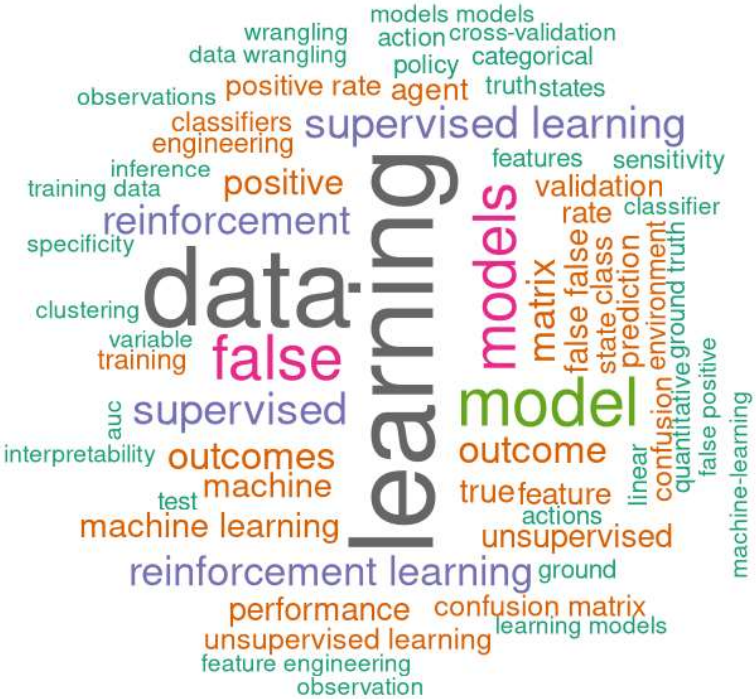
Các thuật ngữ phổ biến



1. Các khái niệm & Thuật ngữ

Các thuật ngữ phổ biến

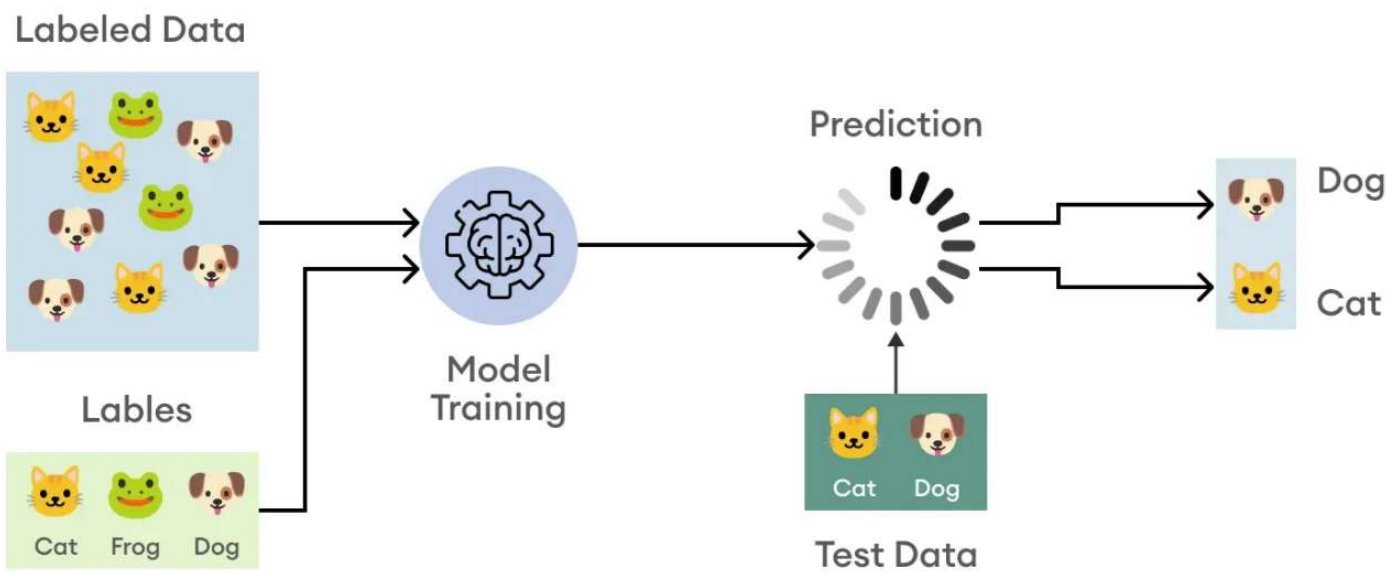
- https://en.wikipedia.org/wiki/Category:Machine_learning



2. Các tác vụ chính của học máy

Phân loại (Classification):

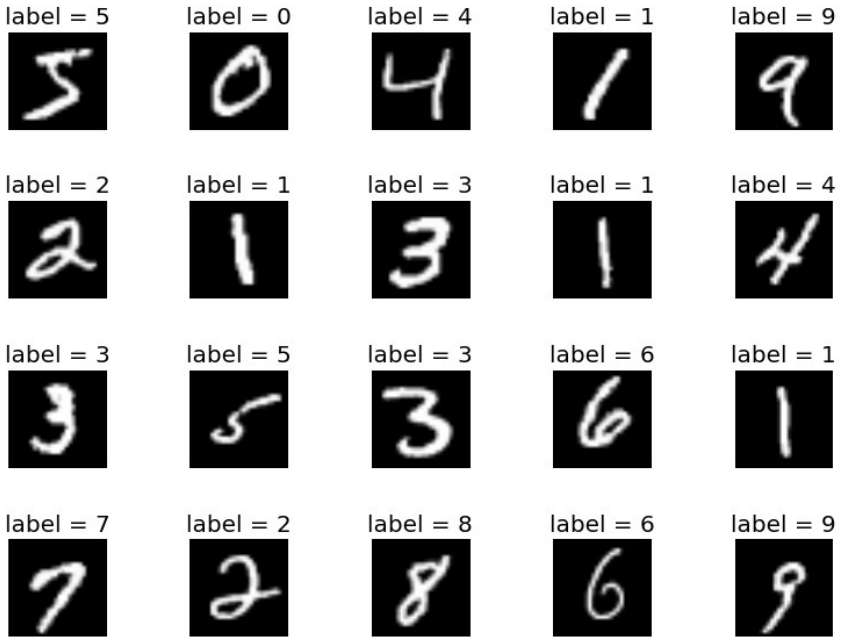
- **Mục tiêu:** Phân loại dữ liệu thành các nhóm hoặc lớp khác nhau.
- **Ví dụ:** Phân loại hình ảnh.



2. Các tác vụ chính của học máy

Phân loại (Classification):

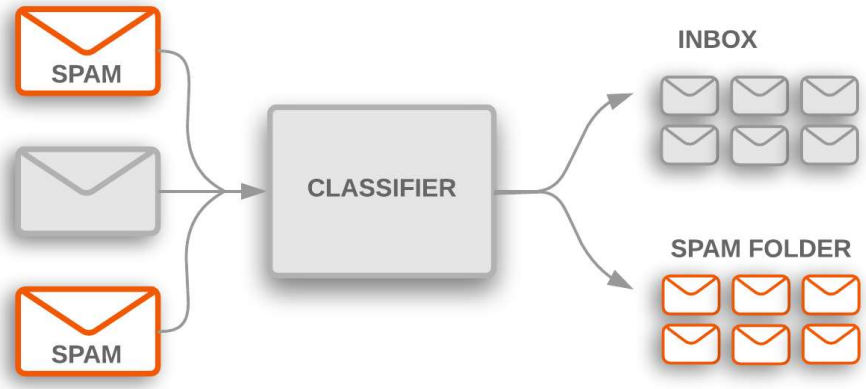
- **Mục tiêu:** Phân loại dữ liệu thành các nhóm hoặc lớp khác nhau.
- **Ví dụ:** MNIST (phân loại chữ viết tay)
- Paper: https://arxiv.org/abs/2406.11173

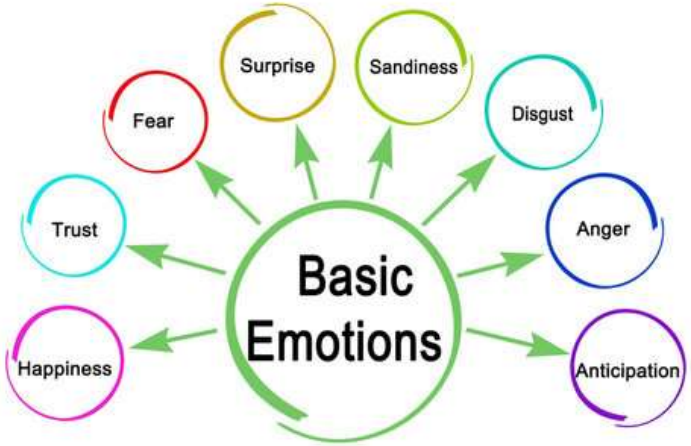


2. Các tác vụ chính của học máy

Phân loại (Classification):

- **Mục tiêu:** Phân loại dữ liệu thành các nhóm hoặc lớp khác nhau.
- **Ví dụ:** Phân loại văn bản (Text classification), Phân tích tình cảm văn bản (Sentiment Analysis)
- Paper: https://arxiv.org/abs/2401.04655

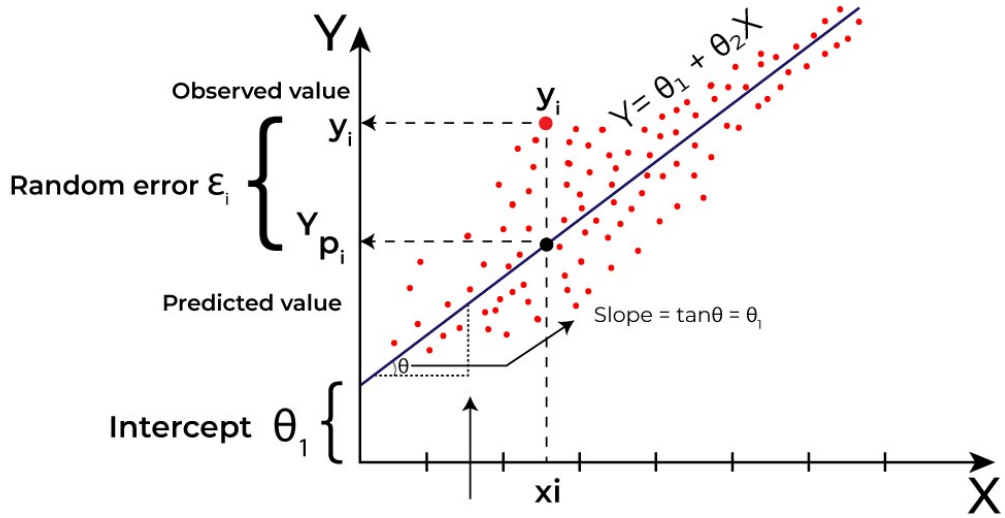




2. Các tác vụ chính của học máy

Hồi quy (Regression):

- Hồi quy tuyến tính (Linear Regression)
- Hồi quy đa biến (Multivariable regression)
  - Mục tiêu: Dự đoán giá trị liên tục dựa trên dữ liệu đầu vào.
  - Ví dụ: Dự đoán giá nhà dựa trên các yếu tố như diện tích và số phòng, dự đoán doanh thu trong tương lai.

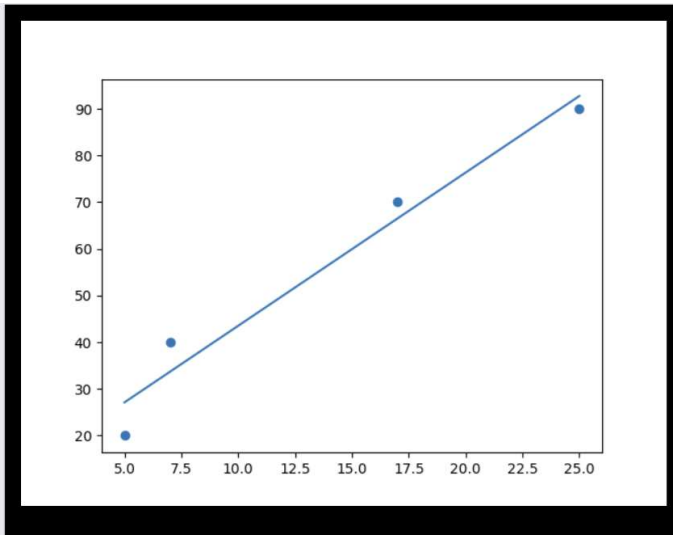


2. Các tác vụ chính của học máy

Hồi quy (Regression):

- Ví dụ tuyến tính hồi quy: https://www.w3schools.com/python/python_ml_linear_regression.asp
  - Slope = 3.28, intercept = 10.69

```python
#Three lines to make our compiler able to draw:
import sys
import matplotlib
matplotlib.use('Agg')

import matplotlib.pyplot as plt
from scipy import stats

x = [5, 7, 17, 25]
y = [20, 40, 70, 90]

slope, intercept, r, p, std_err = stats.linregress(x, y)

def myfunc(x):
  return slope * x + intercept

mymodel = list(map(myfunc, x))

plt.scatter(x, y)
plt.plot(x, mymodel)
plt.show()

#Two  lines to make our compiler able to draw:
plt.savefig(sys.stdout.buffer)
sys.stdout.flush()
```



2. Các tác vụ chính của học máy

Hồi quy (Regression):

- Calculate slope, intercept of **ŷ = x*slop + intercept**:
  - x = [5, 7, 17, 25]
  - y = [20, 40, 70, 90]

Step 1: Calculate X*Y, X^2, Y^2, ΣX, ΣY, ΣX*Y, $ΣX^2$, and $ΣY^2$

	X	Y	X*Y	X^2	Y^2
	5	20	100	25	400
	7	40	280	49	1600
	17	70	1190	289	4900
	25	90	2250	625	8100
SUM	**54**	**220**	**3820**	**988**	**15000**

2. Các tác vụ chính của học máy

Hồi quy (Regression):

- Calculate slope, intercept of **ŷ = x*slop + intercept**:
  - x = [5, 7, 17, 25]
  - y = [20, 40, 70, 90]
  - n = 4

	X	Y	X*Y	X^2	Y^2
	5	20	100	25	400
	7	40	280	49	1600
	17	70	1190	289	4900
	25	90	2250	625	8100
SUM	**54**	**220**	**3820**	**988**	**15000**

Step 2: Calculate intercept (b_0)

$[(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)] \ / \ [n(\Sigma X^2) - (\Sigma X)^2]$

= (220*988 – 54*3820) / (4*988 – 54*54)

= 10.6949…

2. Các tác vụ chính của học máy

Hồi quy (Regression):

- Calculate slope, intercept of **ŷ = x*slop + intercept**:
  - x = [5, 7, 17, 25]
  - y = [20, 40, 70, 90]
  - n = 4

Step 3: Calculate slope (b_1)

$[n(\Sigma XY) - (\Sigma X)(\Sigma Y)] \ / \ [n(\Sigma X^2) - (\Sigma X)^2]$

= (4*3820 – 54*220) / (4*988 – 54*54)

= 3.2818…

	X	Y	X*Y	X^2	Y^2
	5	20	100	25	400
	7	40	280	49	1600
	17	70	1190	289	4900
	25	90	2250	625	8100
SUM	**54**	**220**	**3820**	**988**	**15000**

=> ŷ = x*3.28… + 10.69…

2. Các tác vụ chính của học máy

Hồi quy (Regression):

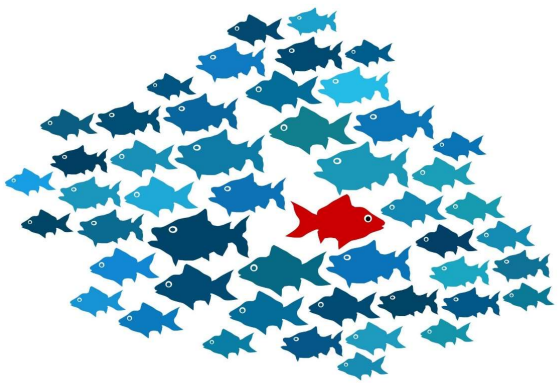
- Bài tập 1:
  - X = [Chiều cao 5 bạn trong lớp]
  - Y = [Cân nặng 5 bạn trong lớp]

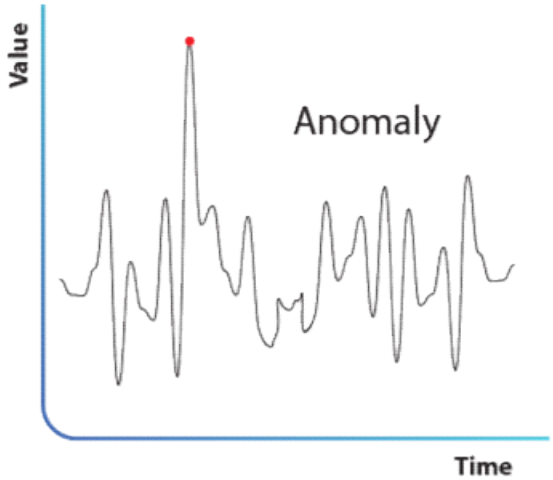
- Bài tập 2:
  - X = [25, 17, 15, 14, 7, 4]
  - Y = [20, 30, 45, 67, 89, 90]
    - Intercept?
    - Slope?

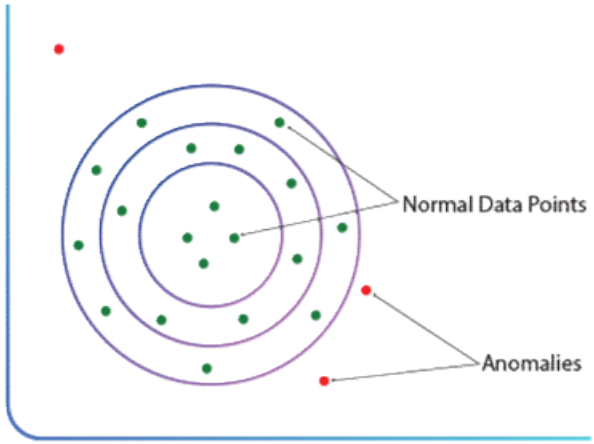
2. Các tác vụ chính của học máy

Phát hiện bất thường (Anomaly Detection):

- **Mục tiêu:** Phát hiện các mẫu dữ liệu bất thường hoặc không giống như các mẫu dữ liệu khác.
- **Ví dụ:** Phát hiện gian lận trong giao dịch thẻ tín dụng, phát hiện lỗi trong hệ thống sản xuất.



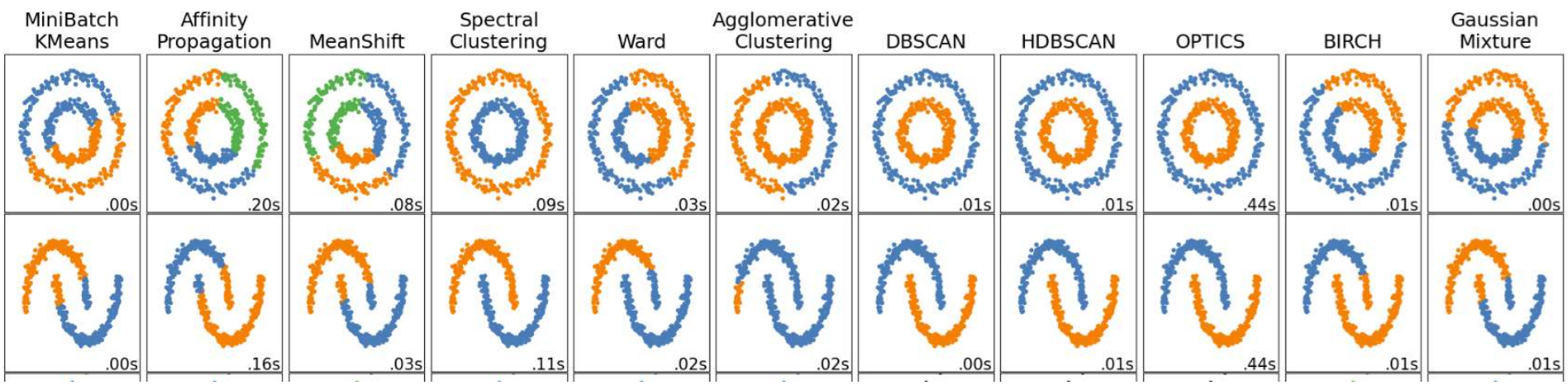




2. Các tác vụ chính của học máy

Nhóm (Clustering):

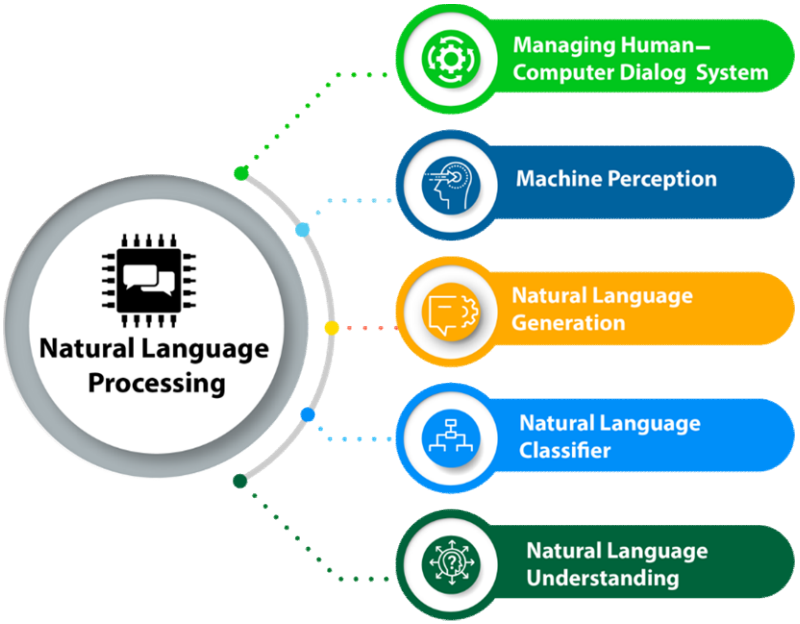
- **Mục tiêu:** Phân nhóm dữ liệu thành các cụm (clusters) dựa trên sự tương tự.
- **Ví dụ:** Phân nhóm khách hàng theo hành vi mua sắm, phân nhóm văn bản theo chủ đề.
  - https://scikit-learn.org/stable/modules/clustering.html

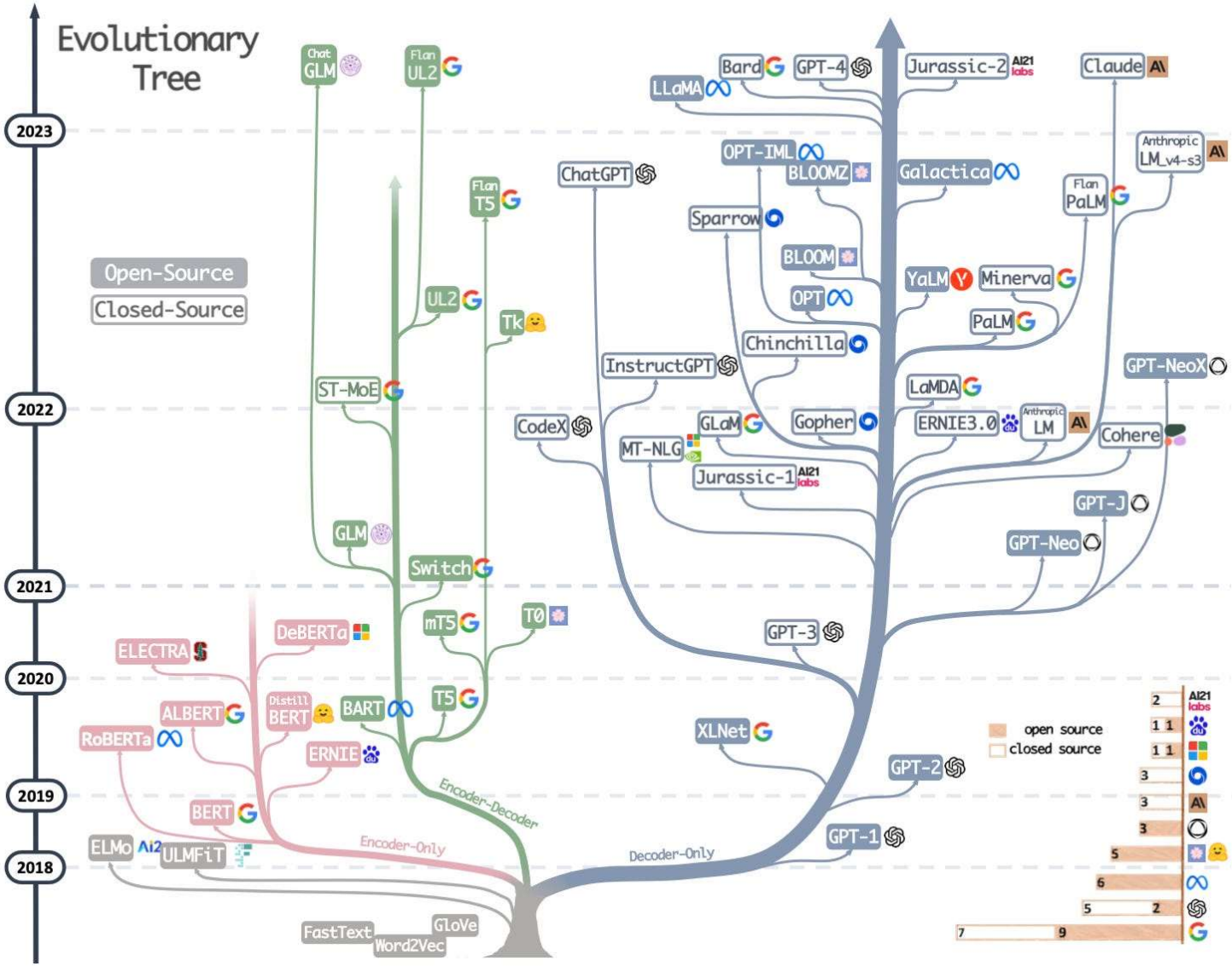


2. Các tác vụ chính của học máy

Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP):

- **Mục tiêu:** Xử lý và hiểu ngôn ngữ tự nhiên từ văn bản hoặc giọng nói.
- **Ví dụ:** Dịch ngôn ngữ, phân tích cảm xúc, nhận diện giọng nói.

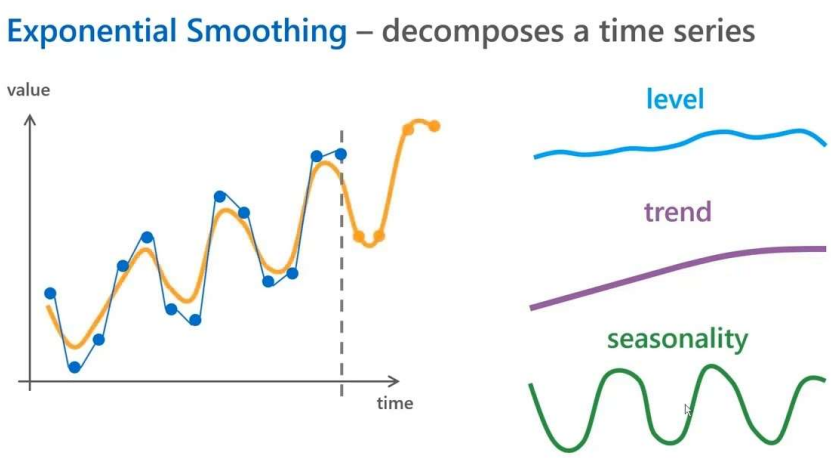


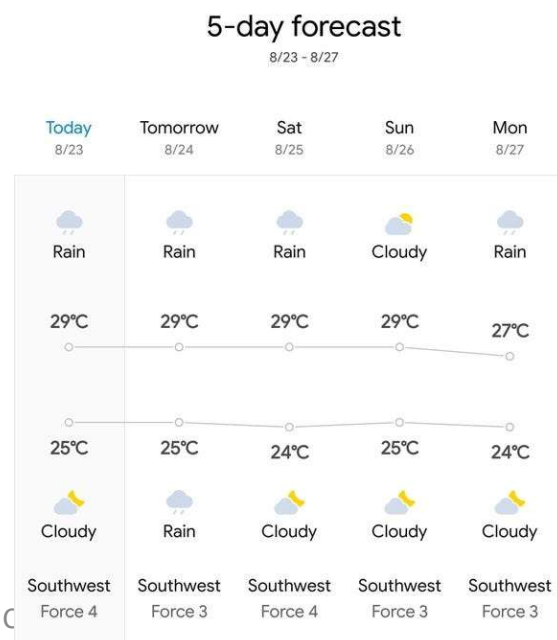

Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Chat GLM
Flan UL2
Flan T5
UL2
Tk
ST-MoE
GLM
Switch
mT5
T0
T5
BART
ELECTRA
DeBERTa
ALBERT
Distill BERT
RoBERTa
ERNIE
BERT
ELMo
ULMFiT
FastText
GloVe
Word2Vec
Encoder-Only
Encoder-Decoder
Decoder-Only
LLaMA
Bard
GPT-4
Jurassic-2
Claude
OPT-IML
BLOOMZ
Galactica
ChatGPT
Sparrow
BLOOM
OPT
Chinchilla
InstructGPT
CodeX
GLaM
MT-NLG
Jurassic-1
Gopher
GPT-3
XLNet
GPT-2
GPT-1
Flan PaLM
Anthropic LM_v4-s3
YaLM
Minerva
PaLM
LaMDA
ERNIE3.0
Anthropic LM
GPT-NeoX
Cohere
GPT-J
GPT-Neo
open source
closed source
AI21 labs

2. Các tác vụ chính của học máy

Dự đoán chuỗi thời gian (Time Series Forecasting):

- **Mục tiêu:** Dự đoán giá trị trong tương lai dựa trên dữ liệu chuỗi thời gian lịch sử.
- **Ví dụ:** Dự đoán doanh thu hàng tháng, dự đoán thời tiết.

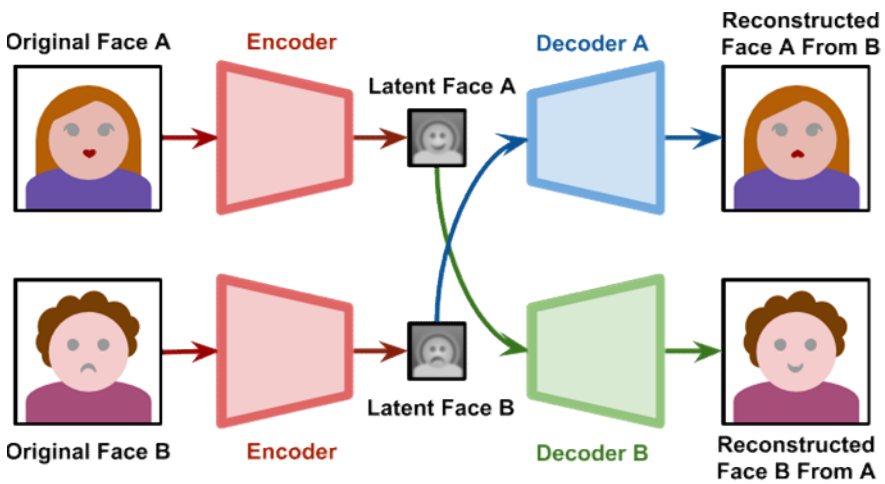




2. Các tác vụ chính của học máy

Sinh dữ liệu (Data Generation):

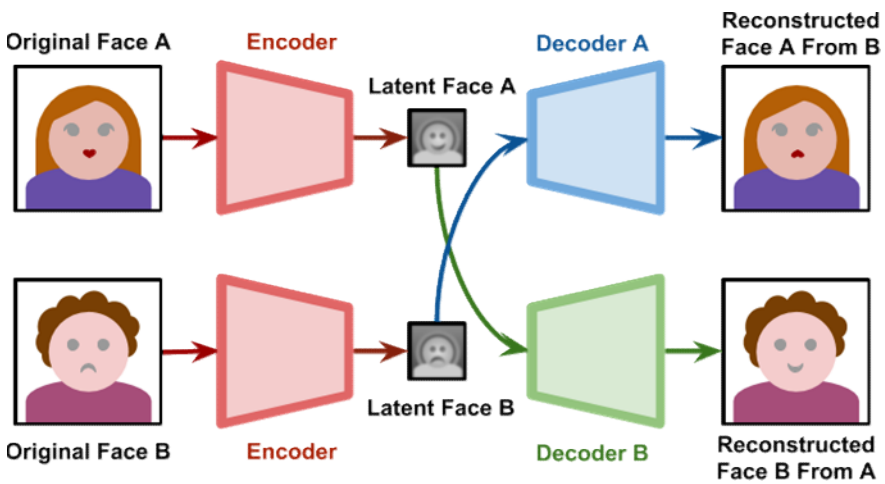
- **Mục tiêu:** Tạo ra dữ liệu mới dựa trên dữ liệu hiện có.
- **Ví dụ:** Tạo ảnh mới bằng mạng đối kháng sinh điều kiện (Conditional GANs), tạo văn bản tự động.



2. Các tác vụ chính của học máy

Sinh dữ liệu (Data Generation):

- **Mục tiêu:** Tạo ra dữ liệu mới dựa trên dữ liệu hiện có.
- **Ví dụ:** Tạo ảnh mới bằng mạng đối kháng sinh điều kiện (Conditional GANs), tạo văn bản tự động.



3. Phân loại phương pháp học máy

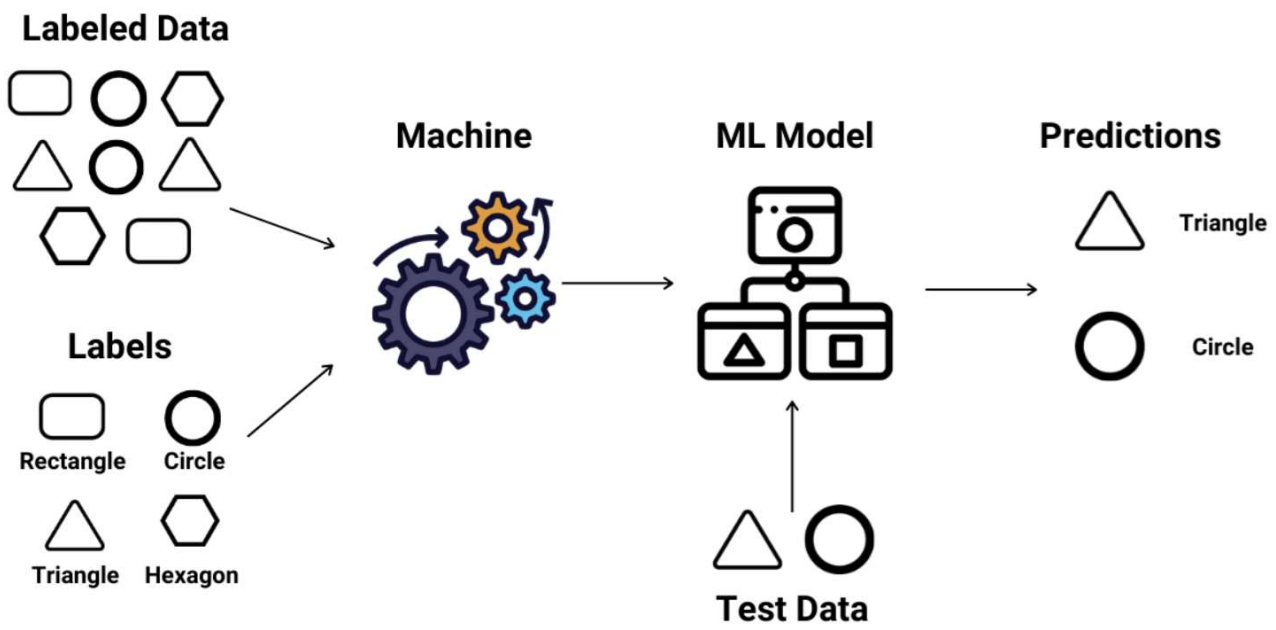
Học Có Giám Sát (Supervised Learning):

- Là một phương pháp trong đó **mô hình được huấn luyện trên một tập dữ liệu có nhãn**. Mục tiêu là học một hàm ánh xạ từ đầu vào đến đầu ra dựa trên dữ liệu đã gán nhãn. Các phương pháp phổ biến bao gồm:
  - **Hồi Quy (Regression):** Dự đoán giá trị liên tục. Ví dụ: hồi quy tuyến tính, hồi quy logistic (khi đầu ra là nhãn nhị phân).
  - **Phân Loại (Classification):** Dự đoán nhãn hoặc lớp của một đối tượng. Ví dụ: phân loại thư rác, phân loại hình ảnh, phân loại văn bản.

3. Phân loại phương pháp học máy

Học Có Giám Sát (Supervised Learning):

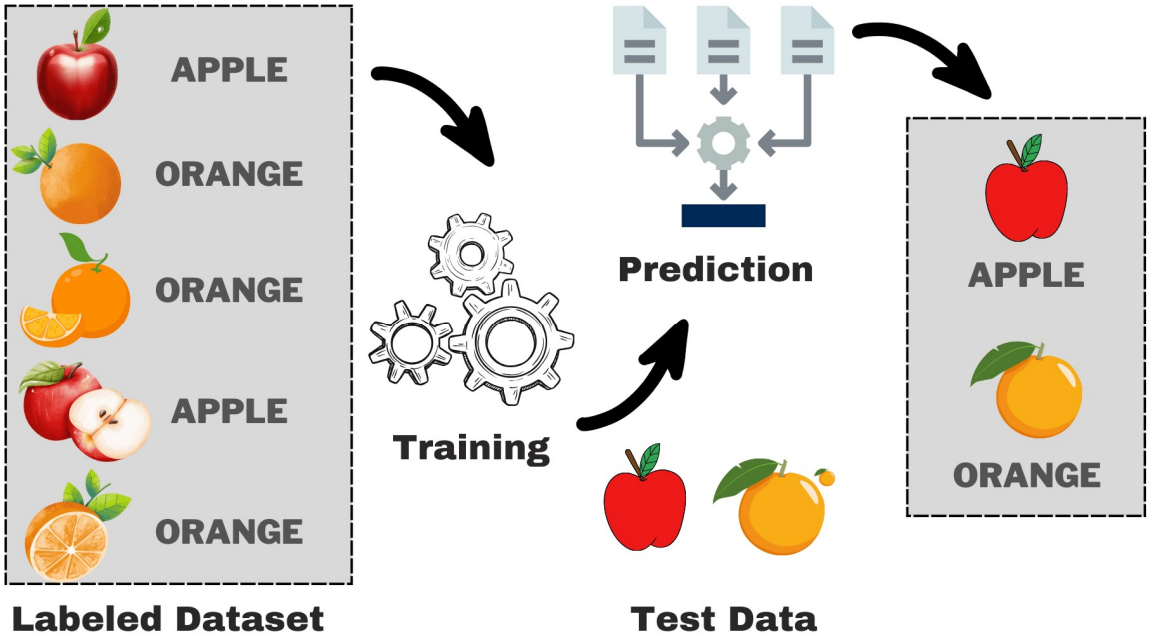
- Là một phương pháp trong đó **mô hình được huấn luyện trên một tập dữ liệu có nhãn.**



3. Phân loại phương pháp học máy

Học Có Giám Sát (Supervised Learning):

- Là một phương pháp trong đó **mô hình được huấn luyện trên một tập dữ liệu có nhãn.**



3. Phân loại phương pháp học máy

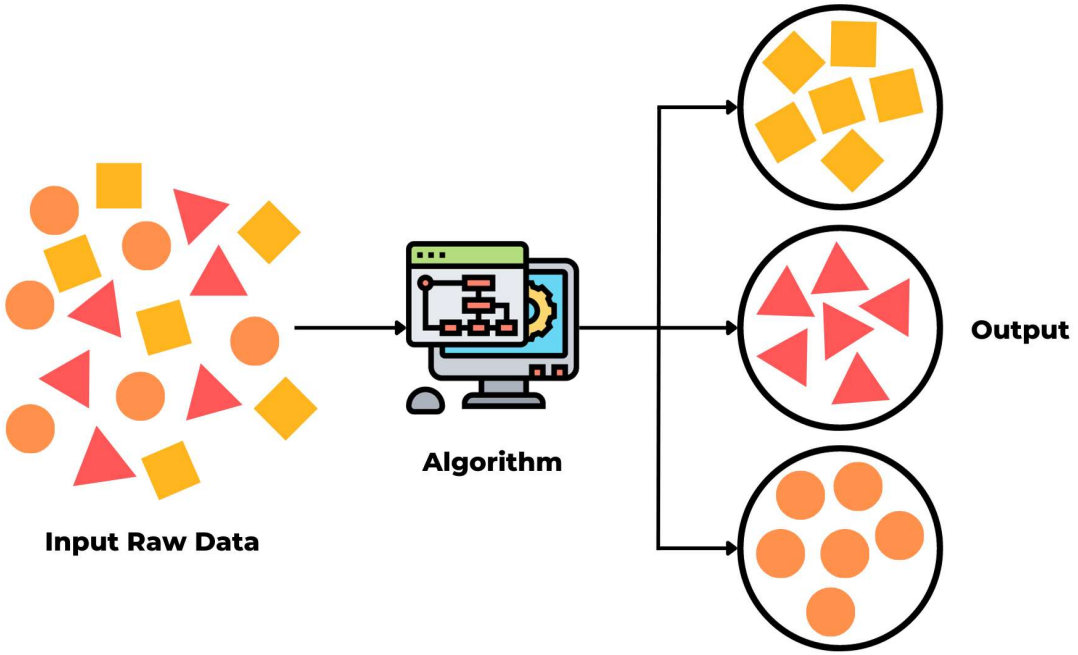
Học Không Có Giám Sát (Unsupervised Learning):

- Là một phương pháp trong đó **mô hình được huấn luyện trên một tập dữ liệu không có nhãn.**
  - Phân Cụm (Clustering): Nhóm các đối tượng tương tự lại với nhau. Ví dụ: k-means, phân cụm phân cấp (hierarchical clustering).
  - Giảm Dimensionality (Dimensionality Reduction): Giảm số lượng đặc trưng trong dữ liệu để dễ phân tích hơn. Ví dụ: Phân tích thành phần chính (PCA), phân tích t-distributed Stochastic Neighbor Embedding (t-SNE).
  - Tìm Mẫu (Pattern Mining): Khám phá các mẫu và quy luật trong dữ liệu. Ví dụ: quy tắc kết hợp (association rule mining).

3. Phân loại phương pháp học máy

Học Không Có Giám Sát (Unsupervised Learning):

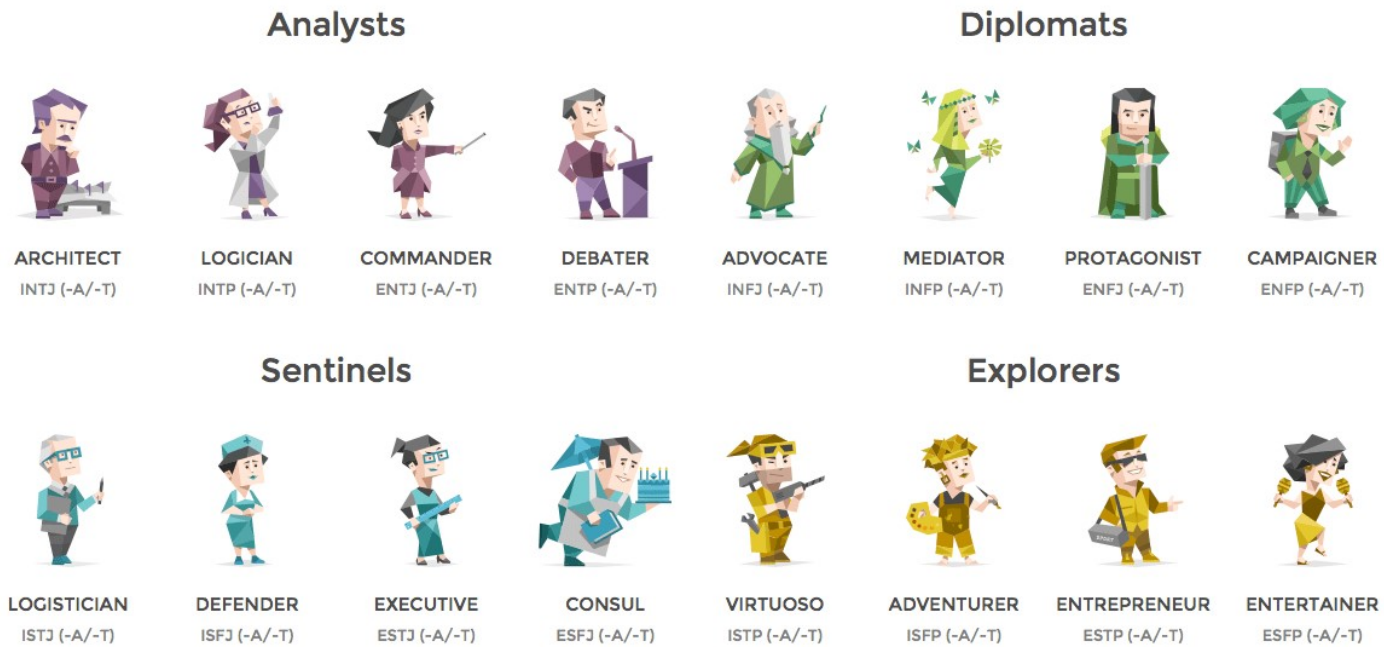
- Là một phương pháp trong đó **mô hình được huấn luyện trên một tập dữ liệu không có nhãn.**



3. Phân loại phương pháp học máy

Học Không Có Giám Sát (Unsupervised Learning):

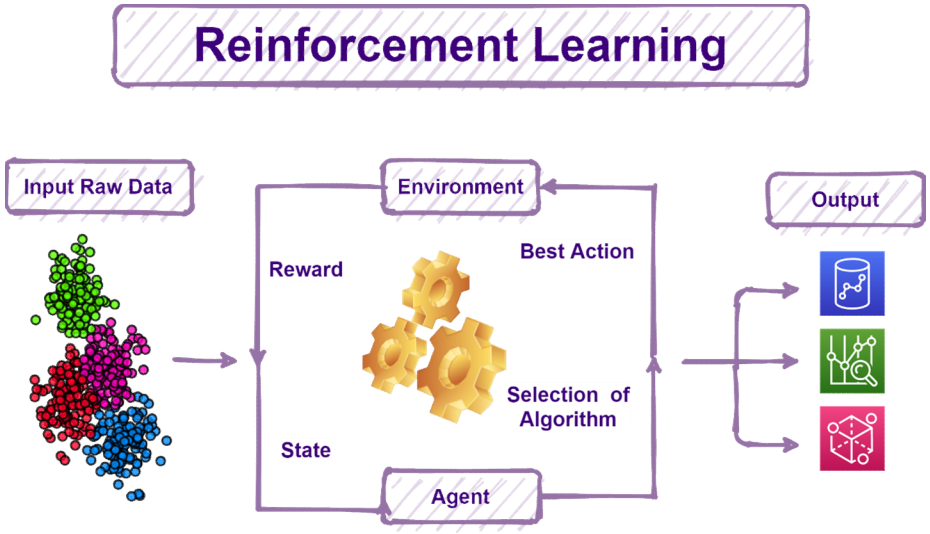
- Là một phương pháp trong đó **mô hình được huấn luyện trên một tập dữ liệu không có nhãn.**



3. Phân loại phương pháp học máy

Học Tăng Cường (Reinforcement Learning):

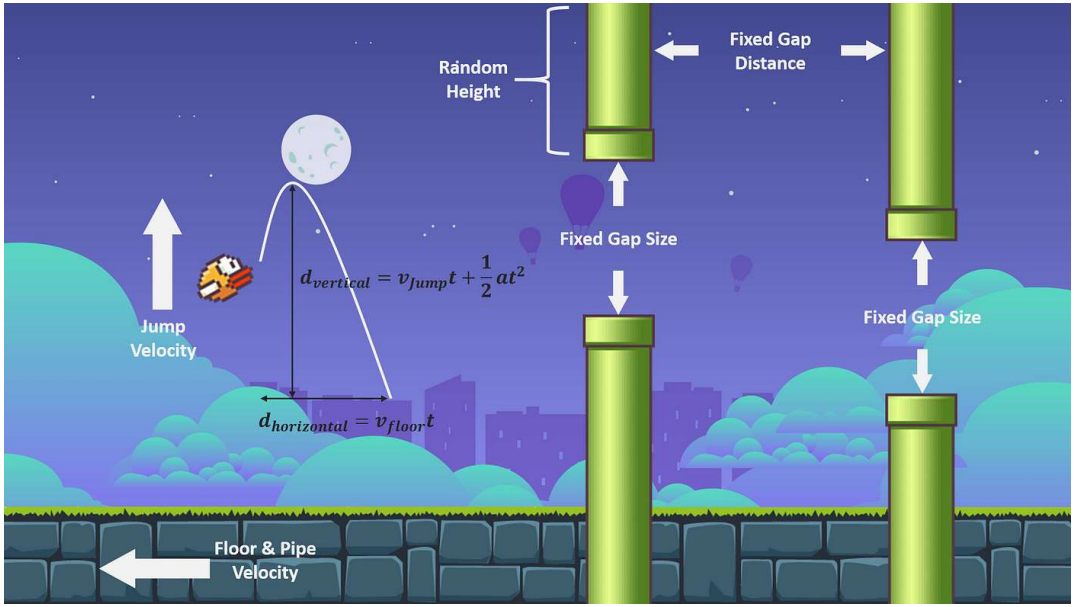
- Học tăng cường là phương pháp trong đó một tác nhân học cách hành động trong một môi trường để tối đa hóa phần thưởng.
  - Tác nhân học từ các trải nghiệm của mình và điều chỉnh hành vi để đạt được phần thưởng tối ưu. Ví dụ: Q-learning, Deep Q-Network (DQN), các phương pháp chính sách (policy-based methods)



3. Phân loại phương pháp học máy

Học Tăng Cường (Reinforcement Learning):

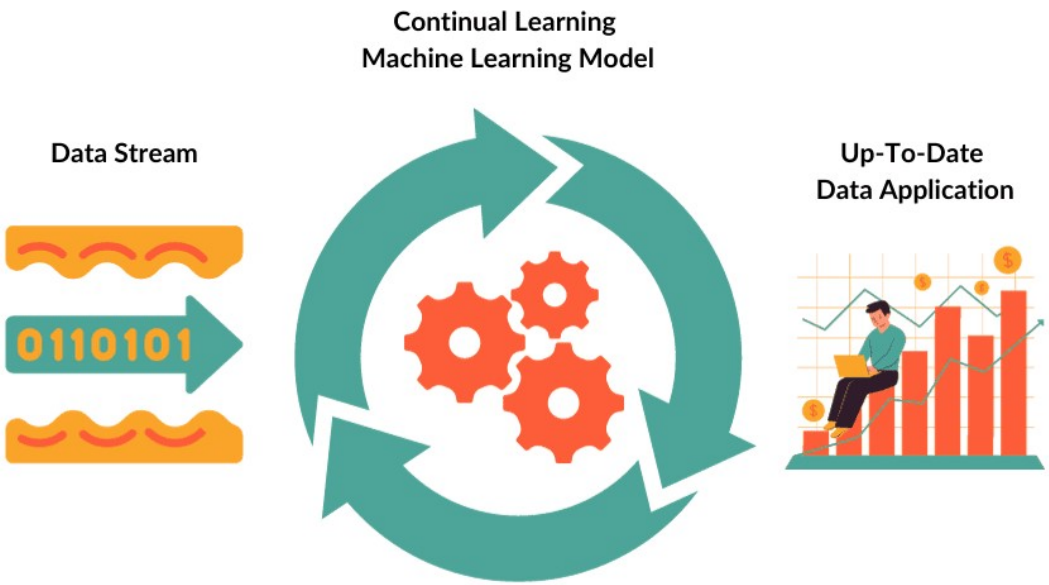
- Học tăng cường là phương pháp trong đó một tác nhân học cách hành động trong một môi trường để tối đa hóa phần thưởng.



3. Phân loại phương pháp học máy

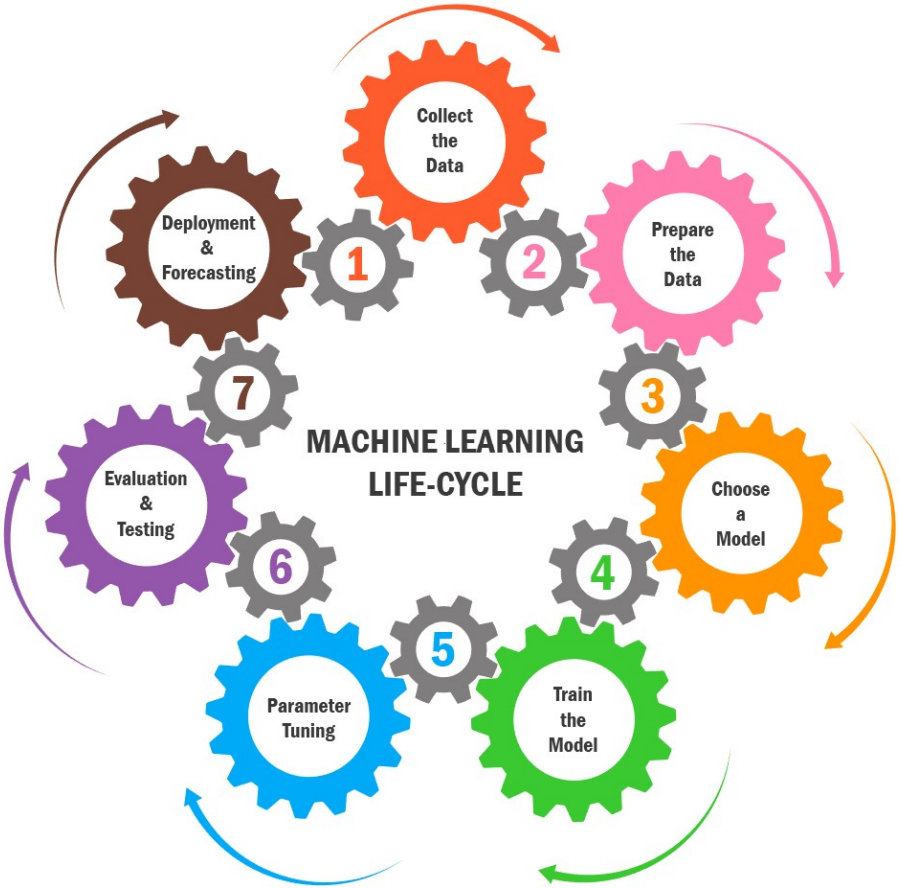
Các phương pháp khác:

- Transfer Learning
- Hybrid Learning
- **Continual Learning**
- Lifelong Learning
- Self-Supervised Learning
- …



4. Các bước xây dựng hệ thống học máy

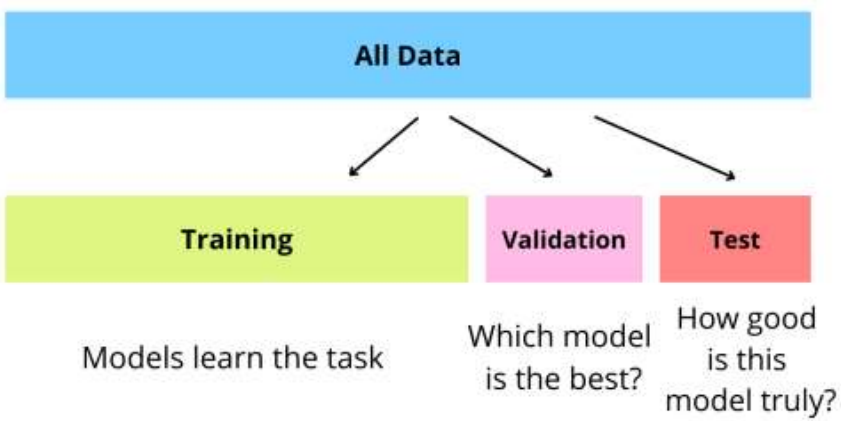
Vòng đời ML

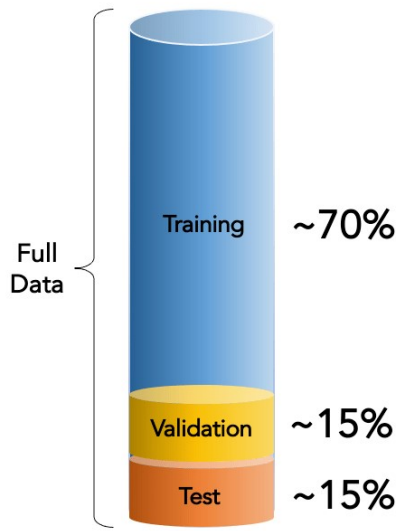


4. Các bước xây dựng hệ thống học máy

Dữ liệu (Data):

- Nguồn dữ liệu: Nơi dữ liệu được thu thập, có thể là từ cơ sở dữ liệu, tệp tin, cảm biến, web scraping, v.v.
  - Dữ liệu huấn luyện (Training data): Dữ liệu được sử dụng để huấn luyện mô hình.
  - Dữ liệu kiểm thử (Test data): Dữ liệu được sử dụng để đánh giá hiệu suất của mô hình sau khi huấn luyện.
  - Dữ liệu xác thực (Validation data): Dữ liệu được sử dụng để điều chỉnh và tinh chỉnh mô hình trong quá trình huấn luyện.





4. Các thành phần của hệ thống học máy

Tiền Xử Lý Dữ Liệu (Data Preprocessing):

- Xử lý thiếu dữ liệu: Điền giá trị thiếu hoặc loại bỏ các giá trị không hoàn chỉnh.
- Chuyển đổi dữ liệu: Chuyển đổi dữ liệu thành định dạng phù hợp cho mô hình học máy.
- Chuẩn hóa: Đưa dữ liệu về cùng một thang đo (normalization) hoặc chuẩn hóa (standardization).
- Làm sạch dữ liệu: Loại bỏ các dữ liệu không chính xác hoặc không cần thiết.
- Tách dữ liệu: Chia dữ liệu thành các tập huấn luyện, kiểm tra và xác nhận.

4. Các bước xây dựng hệ thống học máy

Tiền xử lý dữ liệu (Data Preprocessing):

- Chuẩn hóa max-min:
  - Ví dụ: **[10, 20, 30, 40, 50] → [0, 0.25, 0.5, 0.75, 1]**
  - x_min = 10, x_max = 50

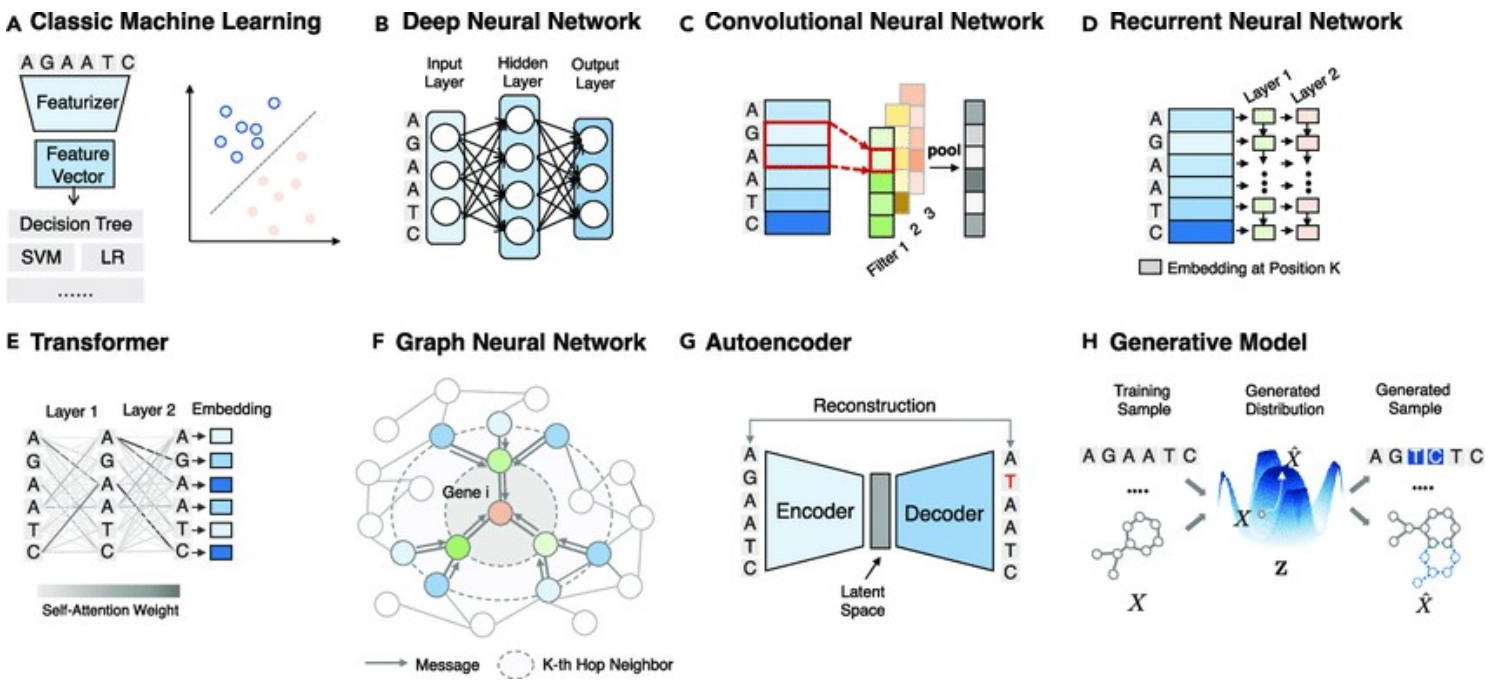
Normalization Formula

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

4. Các bước xây dựng hệ thống học máy

Chọn mô hình học máy (Model Selection):

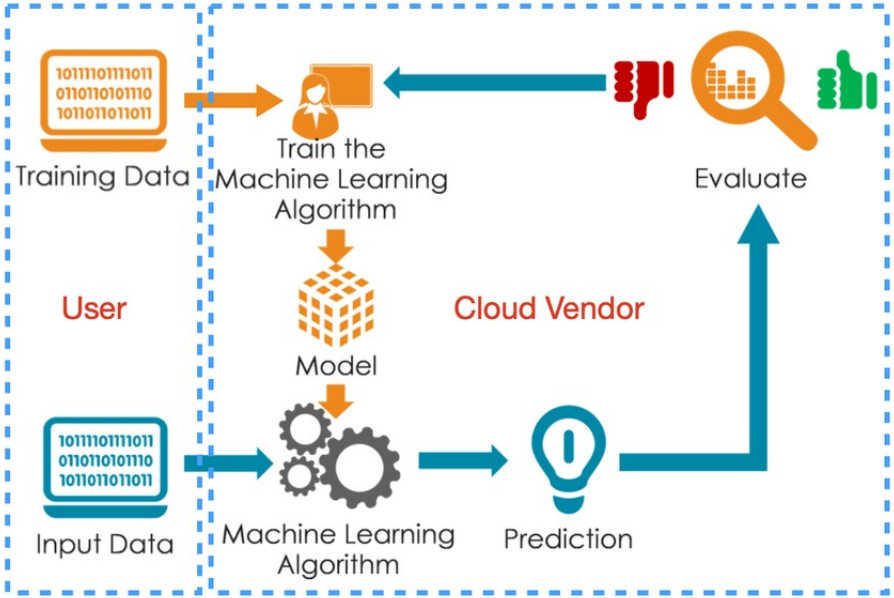
- Chọn mô hình: Chọn một hoặc nhiều mô hình học máy phù hợp với bài toán. Ví dụ: hồi quy tuyến tính, cây quyết định, mạng nơ-ron, v.v.
- Cấu hình mô hình: Xác định các tham số và cấu trúc của mô hình.



4. Các bước xây dựng hệ thống học máy

Huấn Luyện mô Hình (Model Training):

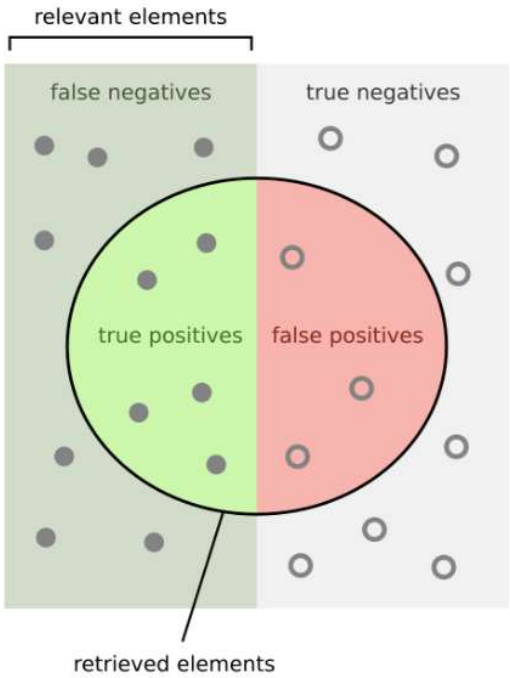
- Huấn luyện: Sử dụng dữ liệu huấn luyện để tìm ra các tham số tối ưu của mô hình.
- Tối ưu: Sử dụng các thuật toán tối ưu hóa để cập nhật trọng số của mô hình. Ví dụ: Gradient Descent.
- Điều chỉnh tham số: Tinh chỉnh các tham số của mô hình để cải thiện hiệu suất.

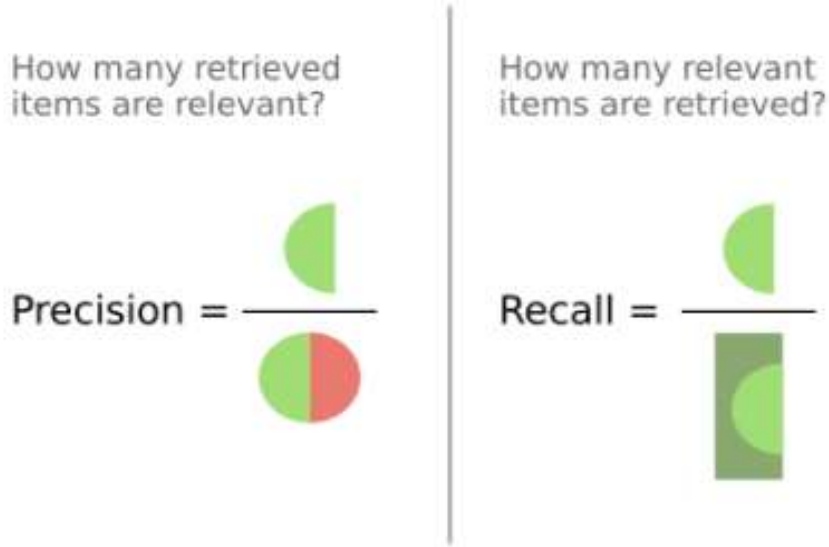


4. Các bước xây dựng hệ thống học máy

Đánh giá mô hình (Model Evaluation):

- Điểm F1:
  - Precision
  - Recall





4. Các thành phần của hệ thống học máy

Đánh giá mô hình (Model Evaluation):

- Điểm F1:
  - Precision
  - Recall

$$\text{Precision} = \frac{TP}{TP + FP}$$

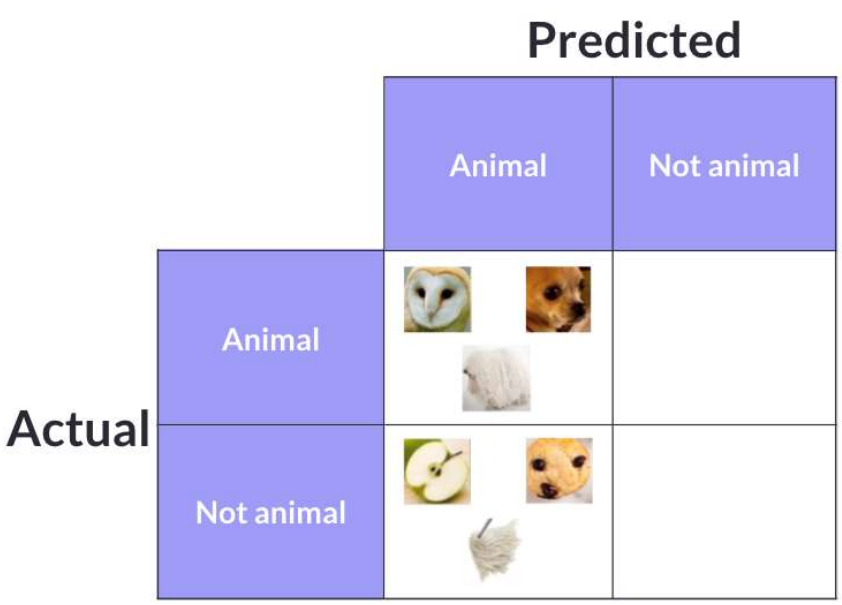
$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

4. Các bước xây dựng hệ thống học máy

Đánh giá mô hình (Model Evaluation):

- Điểm F1:
  - Precision
  - Recall



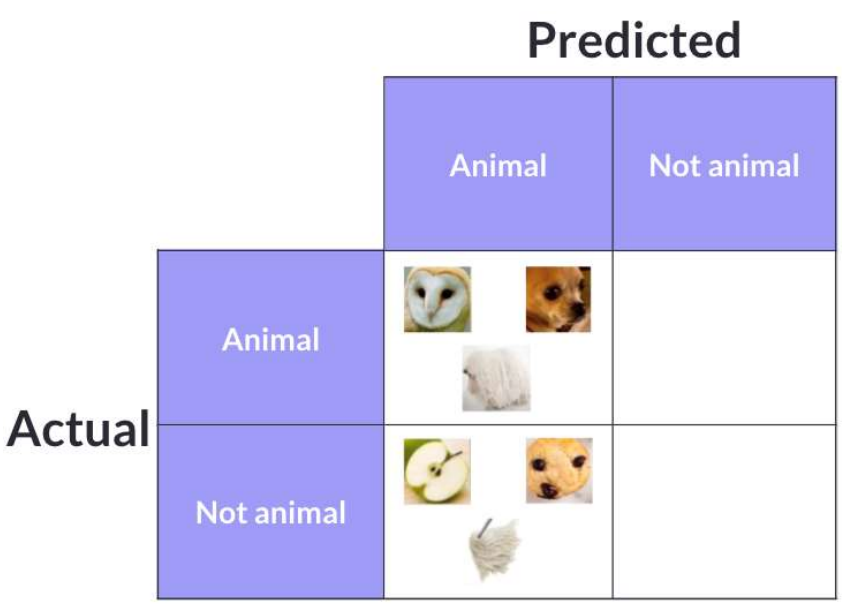
True Positives	3
True Negatives	0
False Positives	3
False Negatives	0

Accuracy	50%	$\frac{3+0}{3+0+0+3}$
Precision	50%	$\frac{3}{3+3}$
Recall	100%	$\frac{3}{3+0}$
F1 score	67%	$2 \cdot \frac{0.5 \cdot 1}{0.5+1}$

4. Các bước xây dựng hệ thống học máy

Đánh giá mô hình (Model Evaluation):

- Điểm F1:
  - Precision
  - Recall



True Positives	3
True Negatives	0
False Positives	3
False Negatives	0

Accuracy	50%	$\frac{3+0}{3+0+0+3}$
Precision	50%	$\frac{3}{3+3}$
Recall	100%	$\frac{3}{3+0}$
F1 score	67%	$2 \cdot \frac{0.5 \cdot 1}{0.5+1}$

4. Các bước xây dựng hệ thống học máy

Đánh giá mô hình (Model Evaluation):

- Điểm F1:
  - Precision
  - Recall

Có 5 quả hình dạng và kích thước gần giống nhau, theo thứ tự:

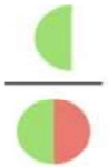
táo, táo, cam, cam, táo

Bạn A bịt mắt và có 5 lượt để sờ từng trái theo thứ tự để đoán trái đó cam hay táo.

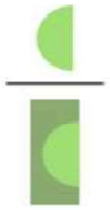
Thực tế: **táo, táo, cam, cam, táo**

A: **cam, táo, cam, cam, cam**

How many retrieved items are relevant?

Precision = 

How many relevant items are retrieved?

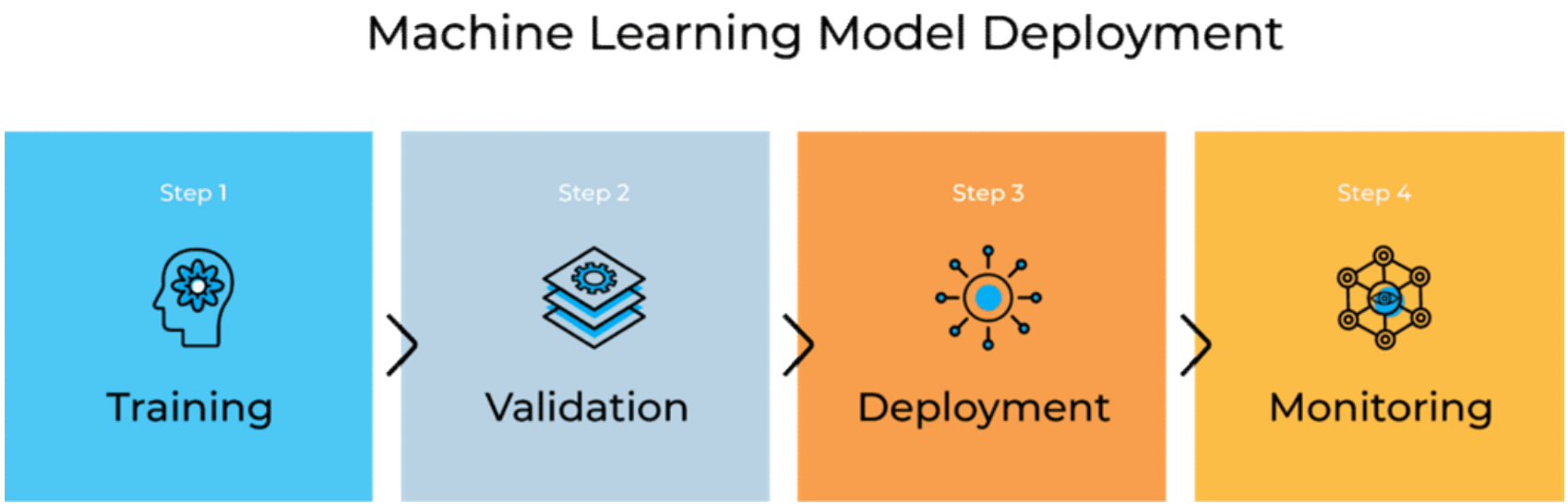
Recall = 

4. Các bước xây dựng hệ thống học máy

Triển khai mô hình (Model Deployment)

Giám sát và bảo trì (Monitoring and Maintenance)

Phản hồi và cải Tiến (Feedback and Improvement)



Thảo luận đề tài
Thảo luận nghiên cứu khoa học