

HCMUT EE MACHINE LEARNING & IOT LAB

Buổi 5

Introduction to Machine Learning

Presentation By: Văn Thịnh

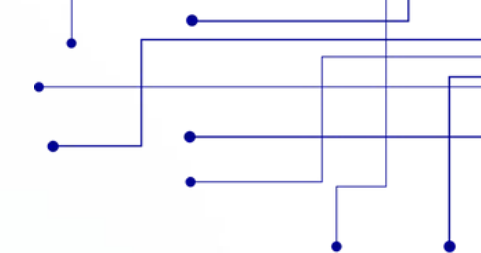


Table of Content

I Machine Learning (ML) là gì?

II Tại sao cần ML?

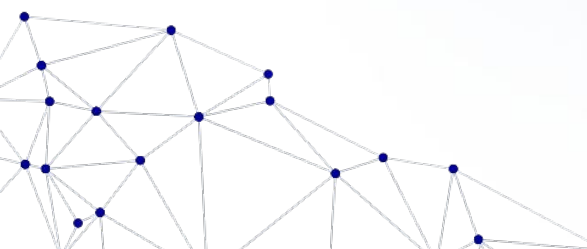
III Các nhóm bài toán trong ML

IV Xây dựng mô hình ML

V Overfitting và Underfitting

VI Tiền xử lý dữ liệu

VII Đo lường hiệu suất



I. Machine Learning

(ML) là gì?

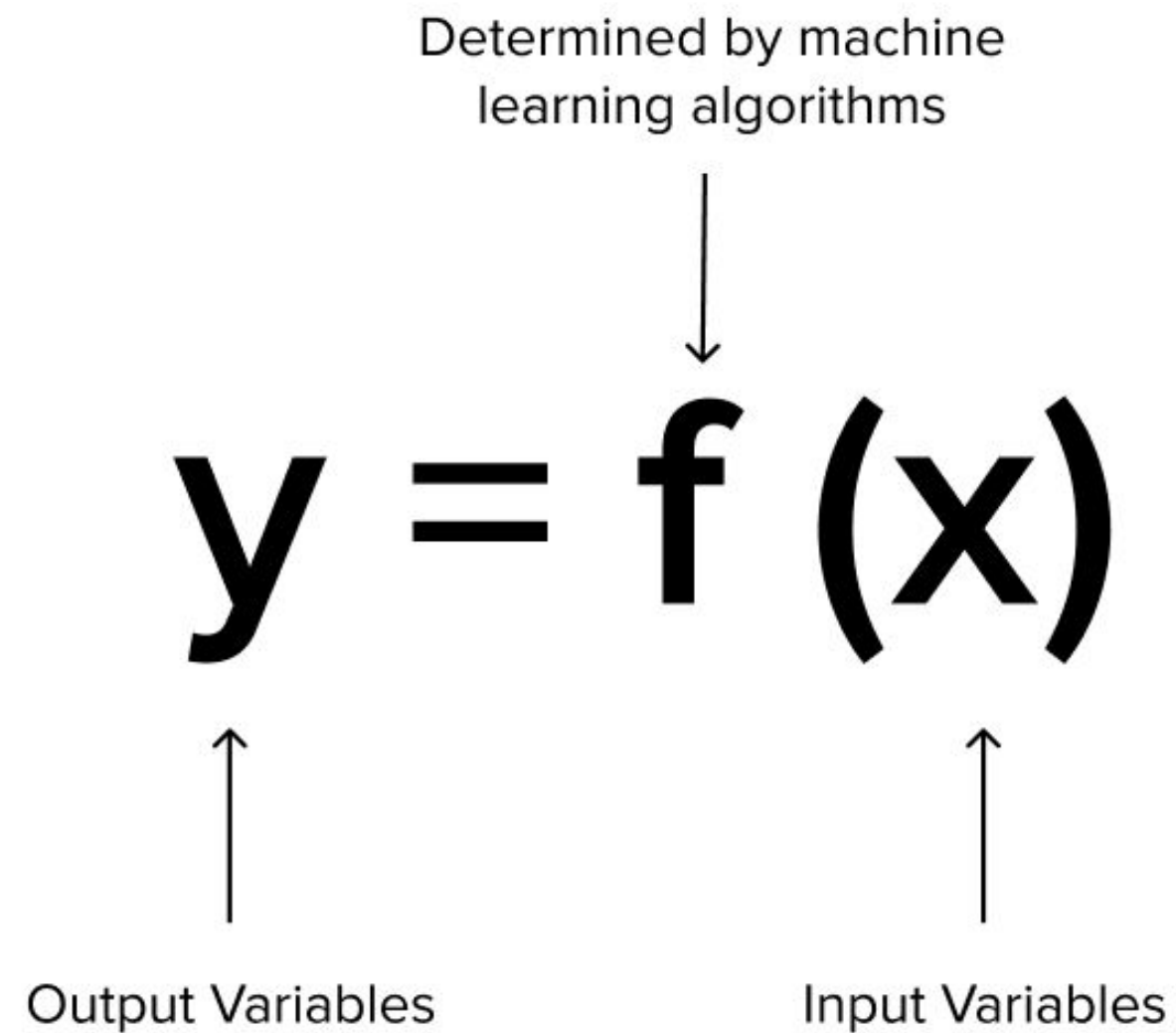
I. Machine Learning (ML) là gì?

Determined by machine learning algorithms

$$y = f(x)$$

Output Variables

Input Variables





? Cùng là tìm cách ánh xạ dữ liệu đến một kết quả nào đó, vậy ML khác gì với lập trình thông thường?

I. Machine Learning (ML) là gì?

Xây dựng một chương trình kiểm tra một số có là số nguyên tố hay không?

```
2 def is_prime(n):
3     if n < 2:
4         return False
5     if n == 2:
6         return True
7     if n % 2 == 0:
8         return False
9     for i in range(3, int(n**0.5) + 1, 2):
10        if n % i == 0:
11            return False
12    return True
```

Đặc điểm:

- Biết được hàm tường minh theo lý thuyết
- Có đầy đủ thông tin về dữ liệu
- Không có ngoại lệ (hoặc rất ít)

I. Machine Learning (ML) là gì?

- Arthur Samuel (1959): "Field of study that gives computers the ability to learn without being explicitly programmed"

Đặc điểm:

- **Khó** hoặc **không thể** tìm được hàm tường minh theo lý thuyết
- Không đủ tổng quát do thiếu dữ liệu
- Nhiều, outliers trong dữ liệu

I. Machine Learning (ML) là gì?

$$y \simeq \hat{y} = f(x)$$

Ground-truth

Prediction

Model

Data

II. Tại sao cần Machine Learning?

II. Tại sao cần Machine Learning?

1. Giải quyết những bài toán không thể viết quy tắc rõ ràng

- Ví dụ: nhận diện khuôn mặt, dịch ngôn ngữ, phân loại cảm xúc.
- Những bài toán này quá phức tạp để lập trình thủ công. Nhưng nếu có dữ liệu (ảnh khuôn mặt + nhãn), ML có thể học được quy luật tiềm ẩn.

II. Tại sao cần Machine Learning?

2. Tự động hóa và thích nghi

- ML giúp hệ thống tự cải thiện theo thời gian, càng dùng càng thông minh.
- Ví dụ: hệ thống gợi ý phim của Netflix hay Youtube học từ hành vi của bạn để gợi ý chính xác hơn.

II. Tại sao cần Machine Learning?

3. Phân tích dữ liệu lớn (Big Data)

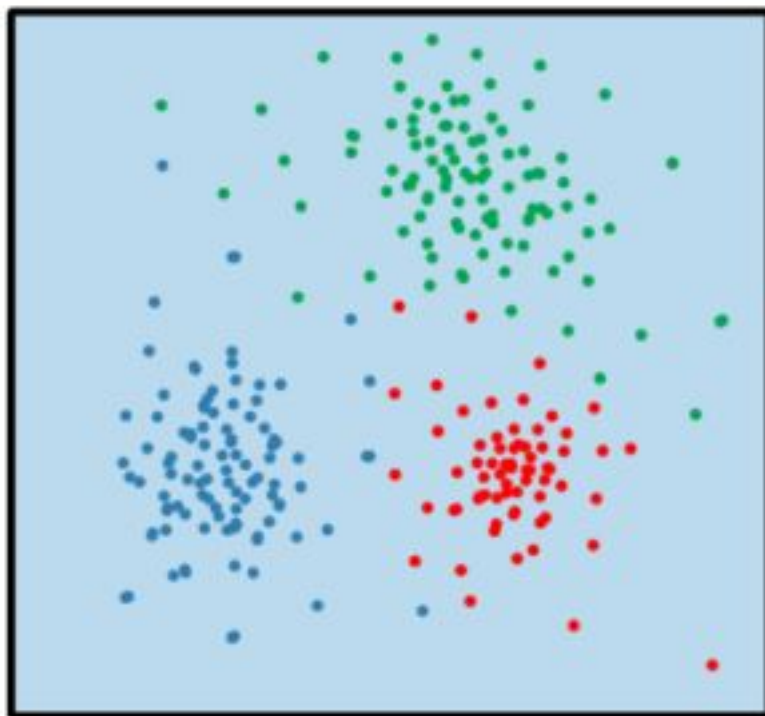
- Dữ liệu ngày nay rất nhiều (hàng tỷ dòng, petabytes).
- ML có khả năng học từ dữ liệu khổng lồ, phát hiện mẫu (pattern) mà con người không thấy.

III. Các nhóm bài toán trong Machine Learning

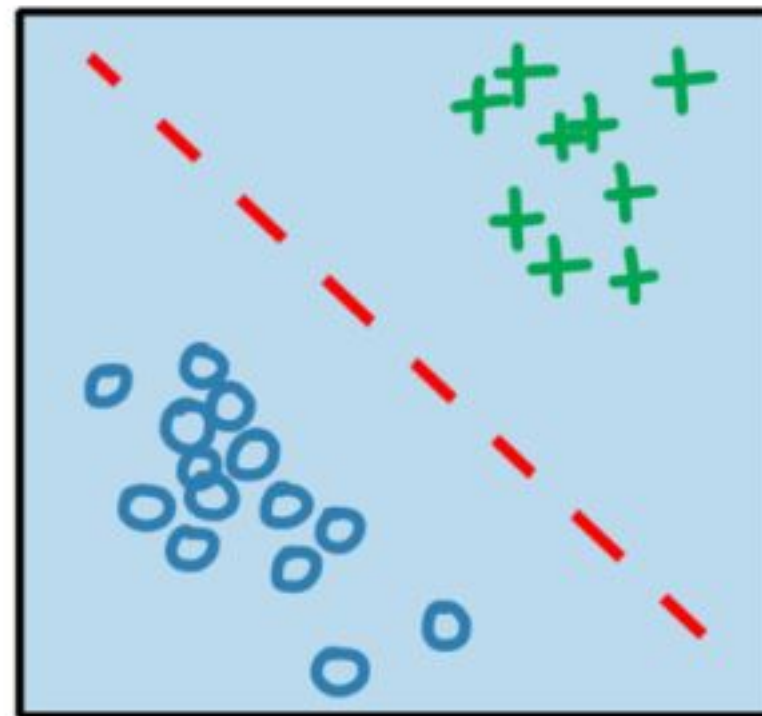
III. Các nhóm bài toán trong ML

machine learning

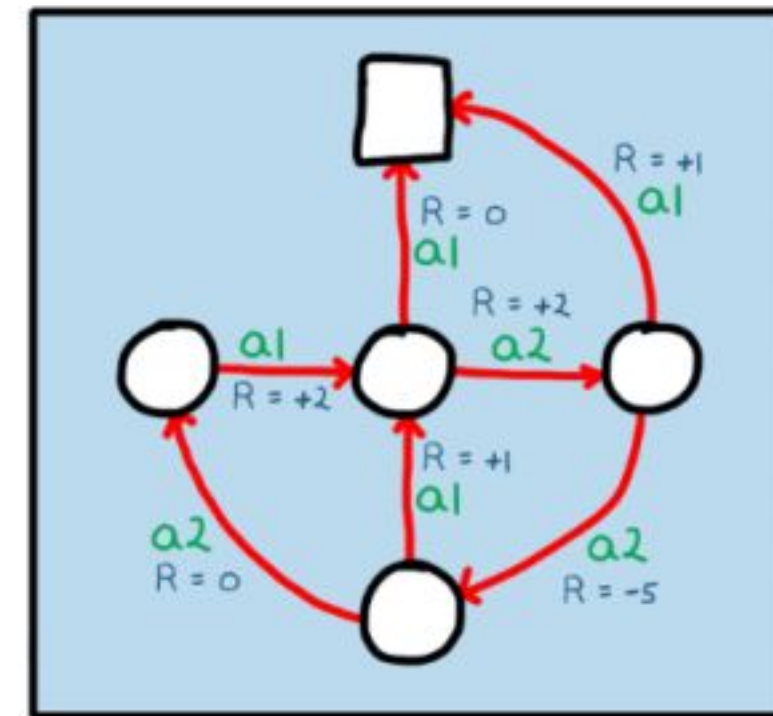
unsupervised
learning



supervised
learning



reinforcement
learning

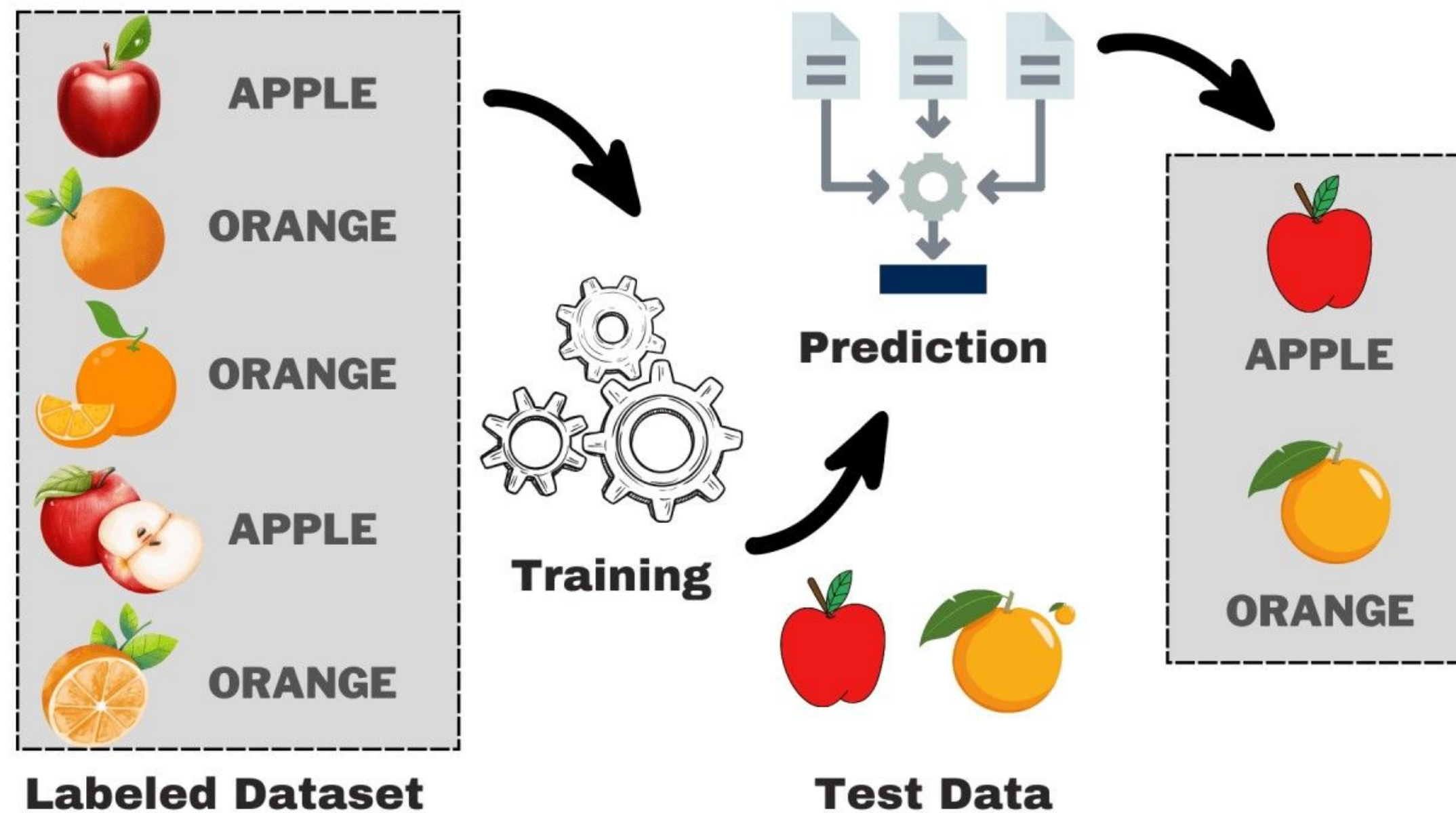


III. Các nhóm bài toán trong ML

1. Supervised Learning (Học có giám sát)

- Mô hình sẽ được huấn luyện (train) trên tập dữ liệu có đã đánh nhãn (label):
Linear Regression, Logistic Regression, Support Vector Machine, Classification

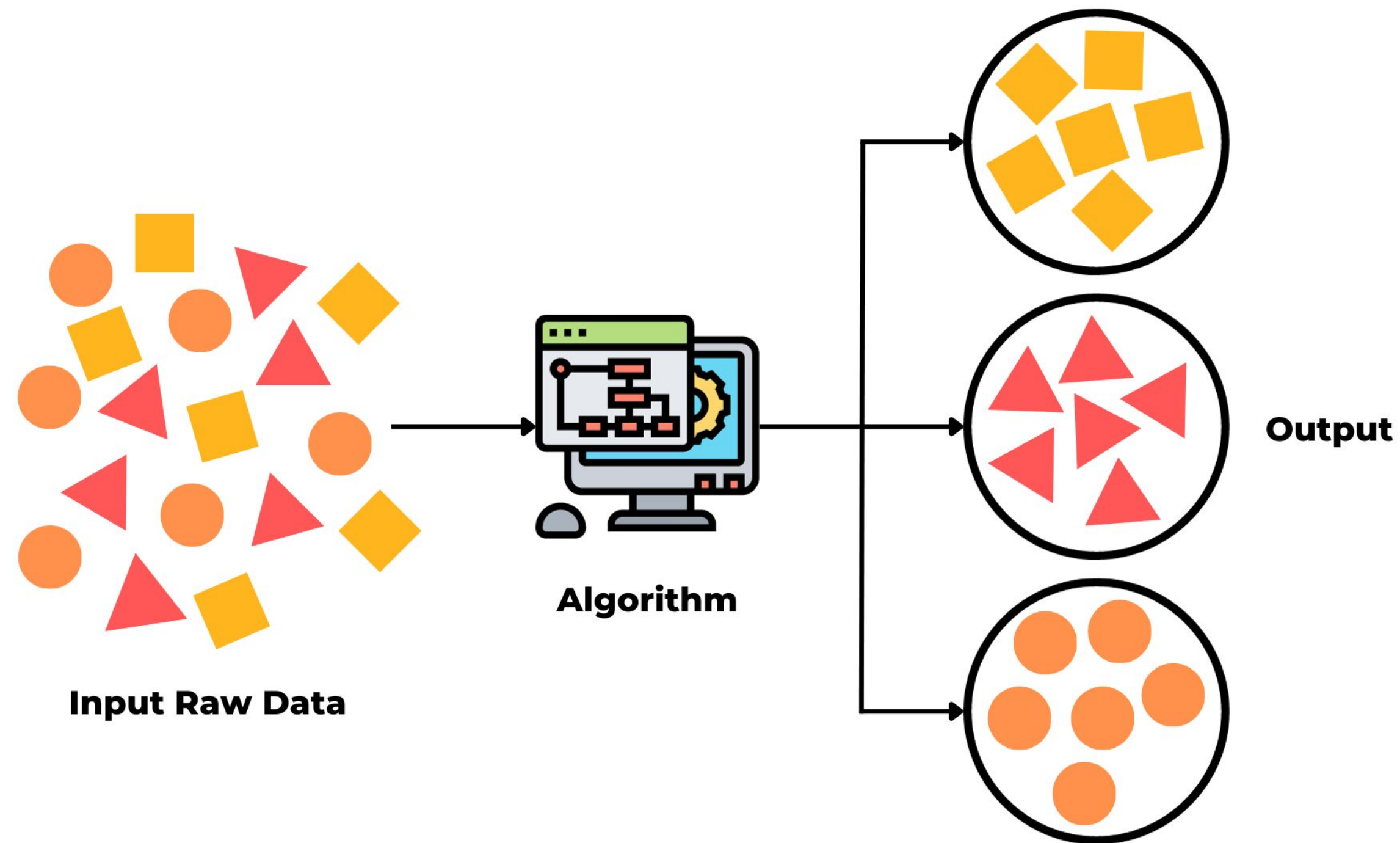
...



III. Các nhóm bài toán trong ML

2. Unsupervised Learning (Học không có giám sát)

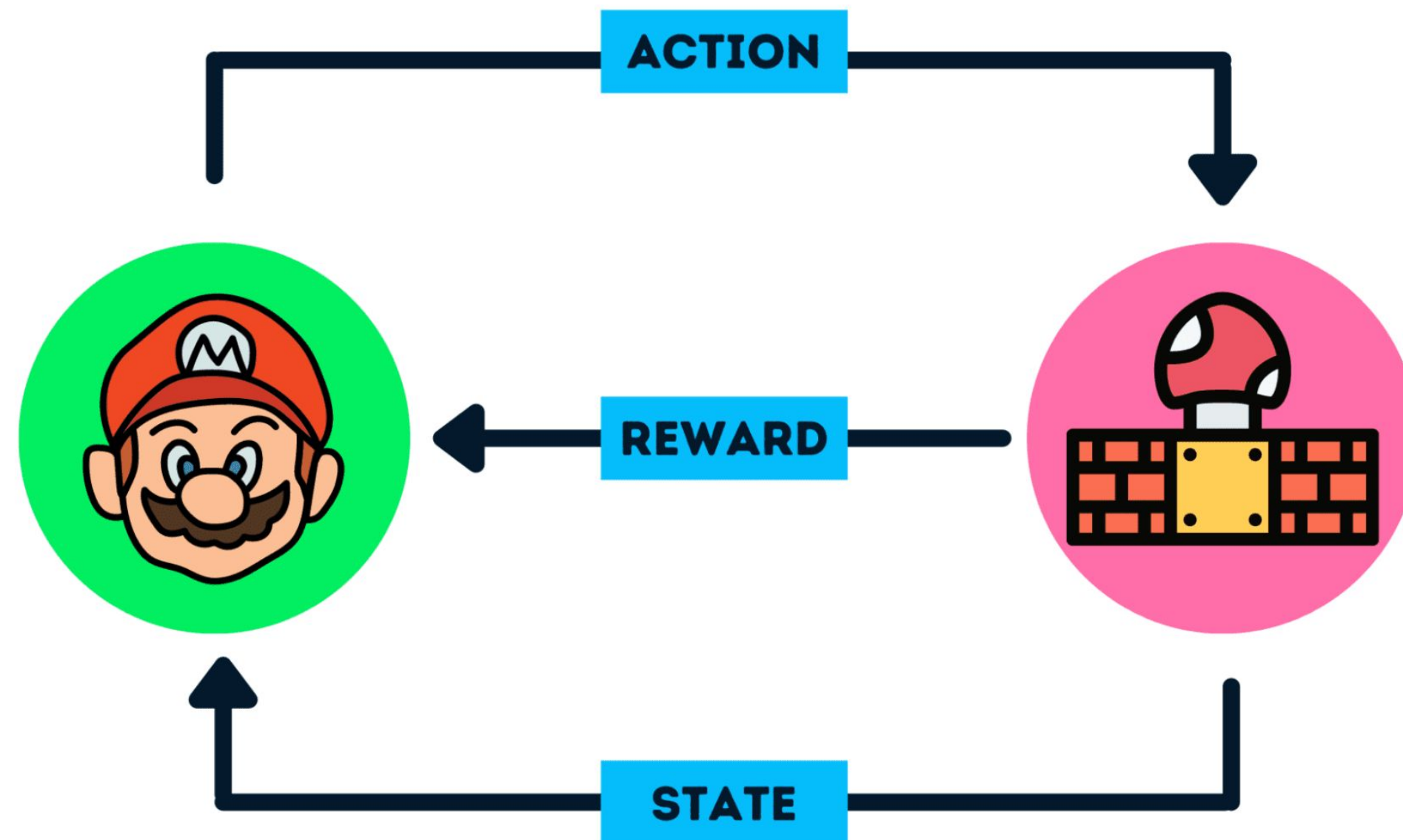
- Mô hình sẽ được huấn luyện (train) trên tập dữ liệu chưa có nhãn (non-label), bao gồm: K-means clustering, PCA,...



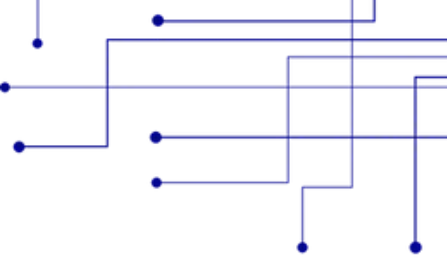
III. Các nhóm bài toán trong ML

3. Reinforcement Learning (Học tăng cường)

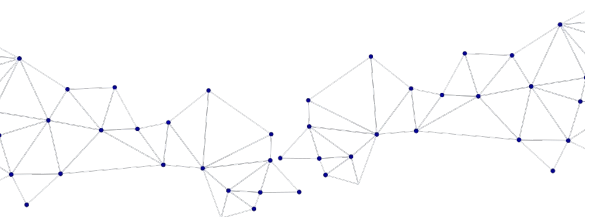
- Trung gian của học có giám sát và không giám sát



III. Các nhóm bài toán trong ML

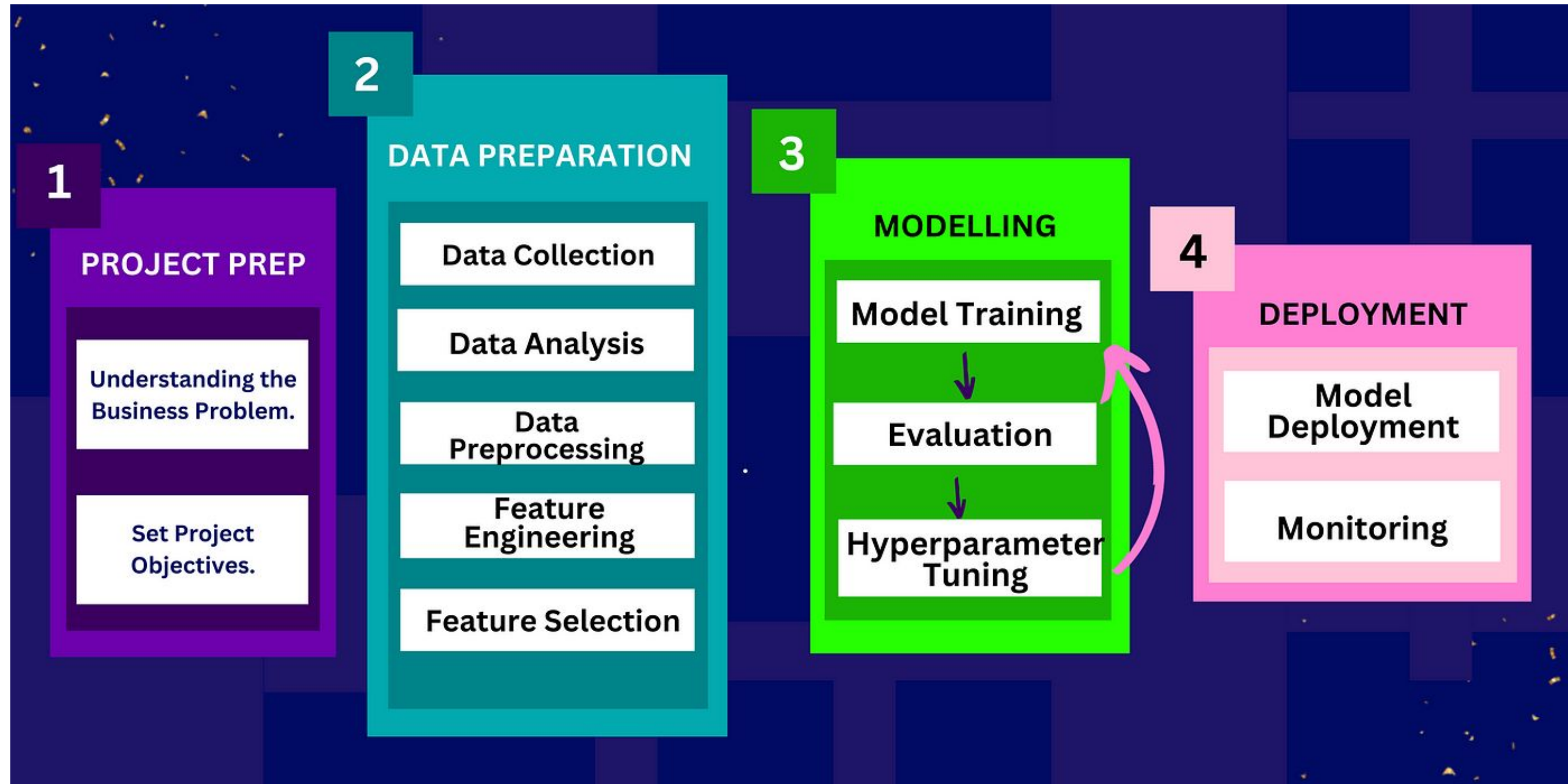


- Semi-supervised Learning
- Self-supervised Learning
- ...



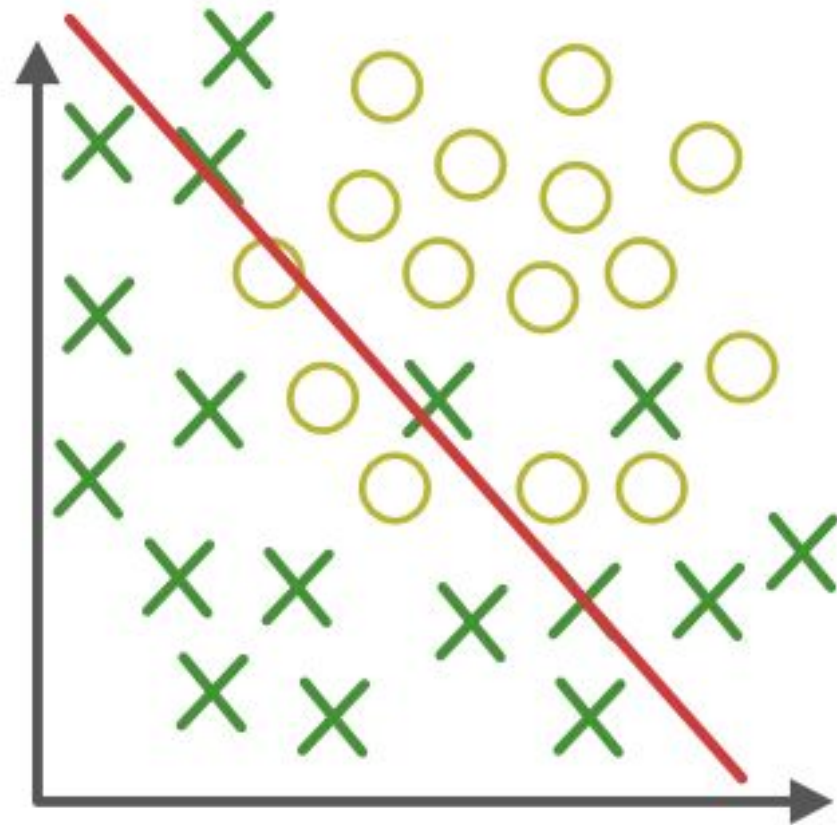
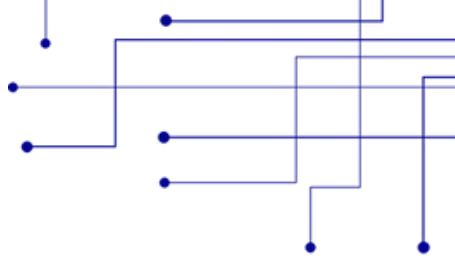
IV. Xây dựng mô hình Machine Learning

IV. Xây dựng mô hình ML

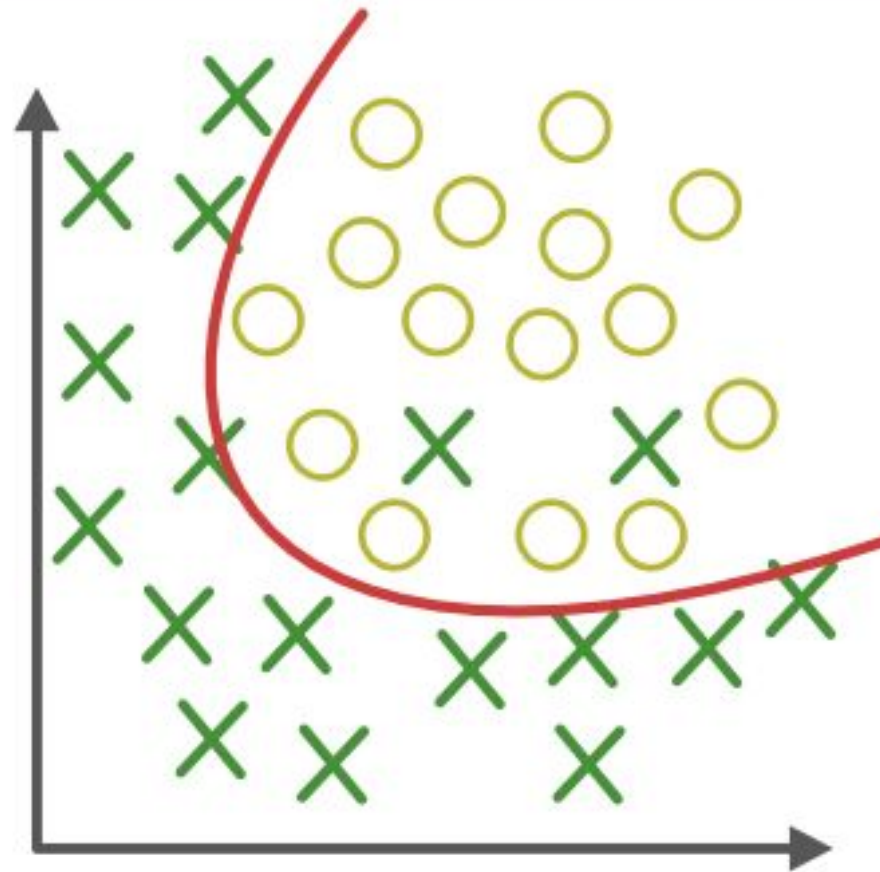


V. Overfitting và Underfitting

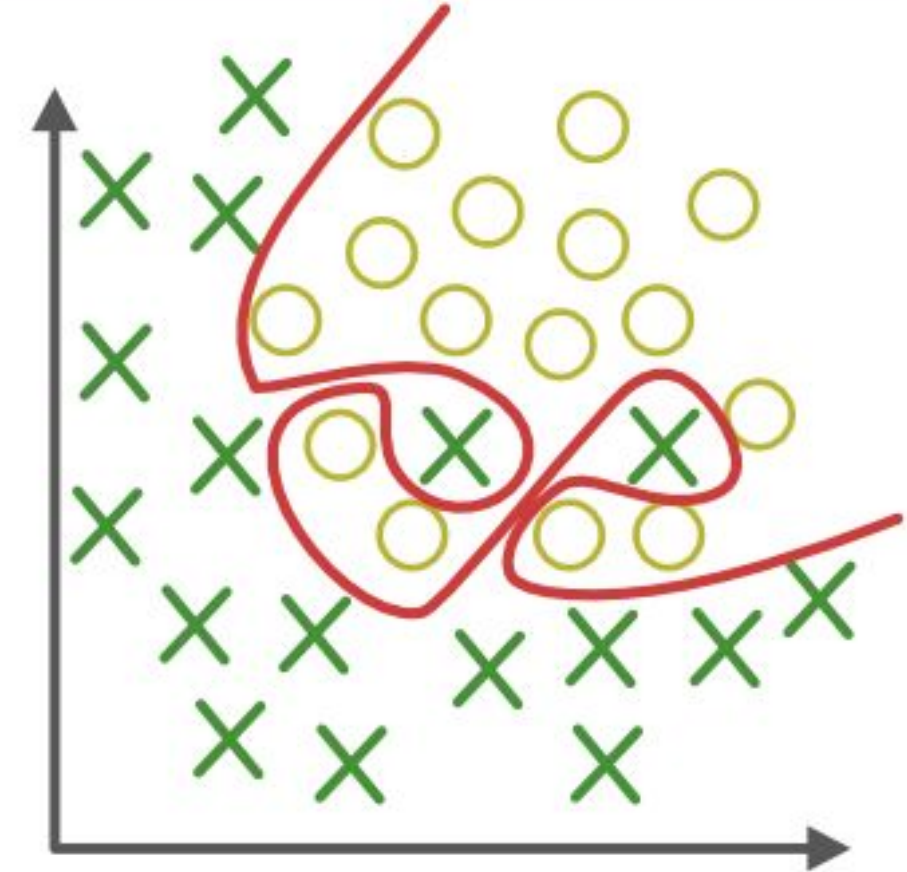
V. Overfitting và Underfitting



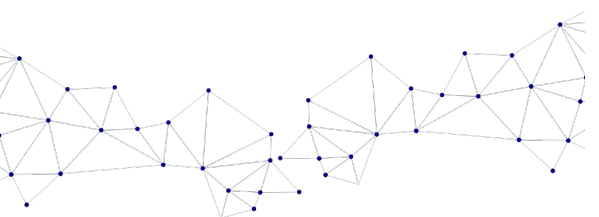
Under-fitting
(too simple to
explain the variance)



Appropriate-fitting



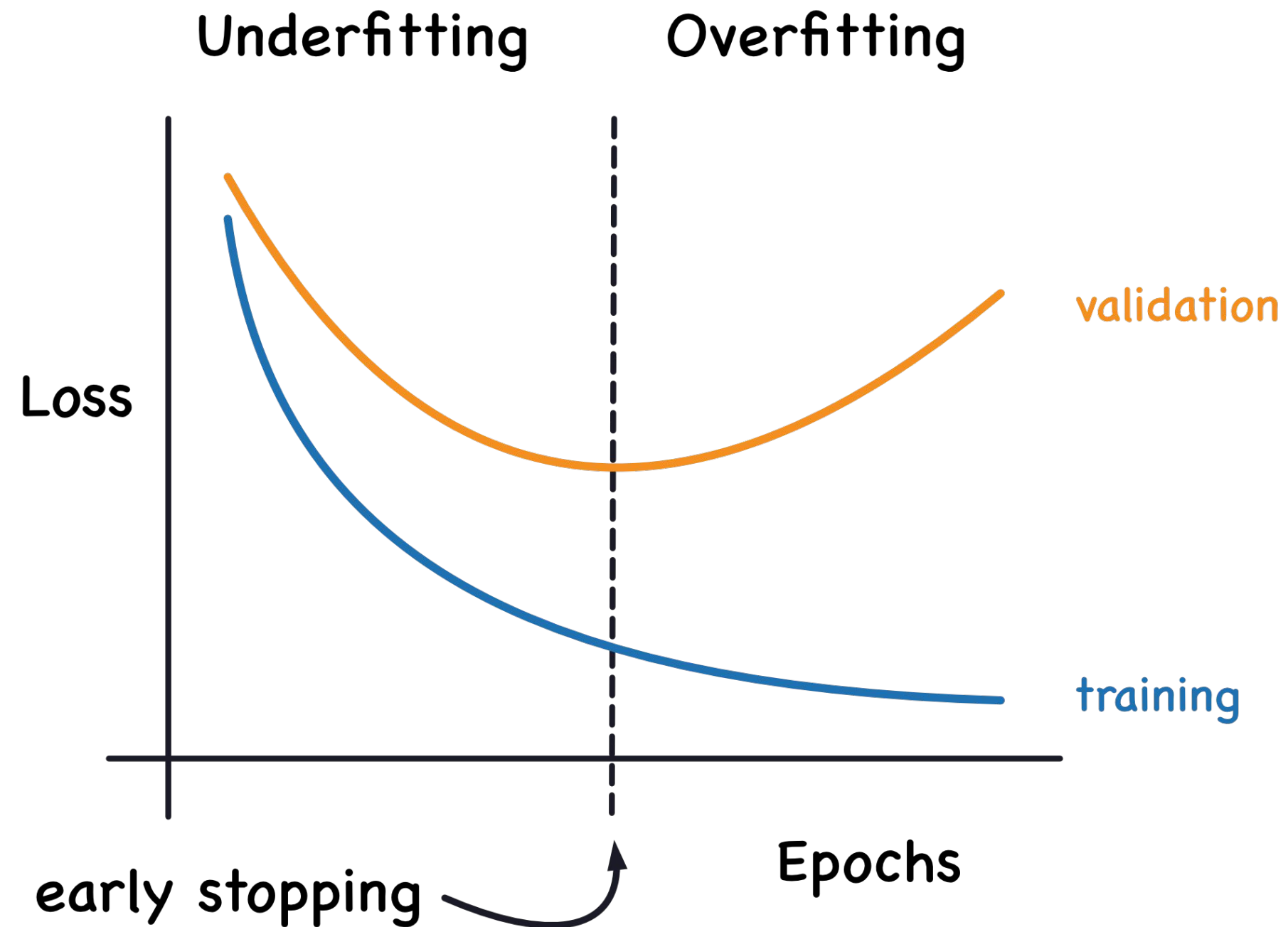
Over-fitting
(forcefitting--too
good to be true)



V. Overfitting và Underfitting

Kỹ thuật tránh overfitting:

- Cross-validation
- Regularization
- Dropout
- ...



VI. Tiền xử lý dữ liệu

VI. Tiền xử lý dữ liệu

Là bước cực kỳ quan trọng trước khi xây dựng một mô hình ML!!!



VI. Tiền xử lý dữ liệu

1. Làm sạch dữ liệu (Data Cleaning):

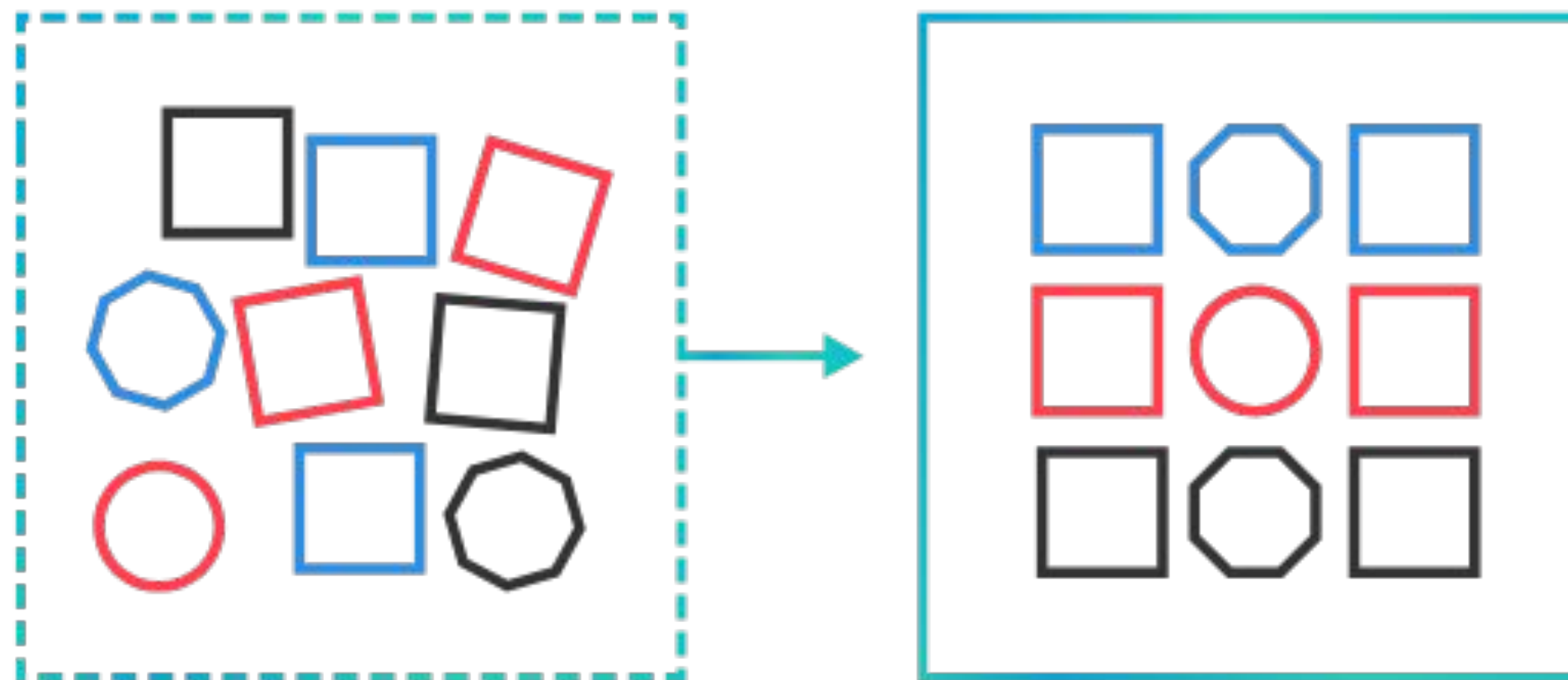
- Xử lý điểm dữ liệu thiếu (missing data)
- Xóa các điểm trùng
- Điều chỉnh định dạng chung
- Loại bỏ outliers



VI. Tiền xử lý dữ liệu

2. Chuyển đổi dữ liệu (Data Transformation)

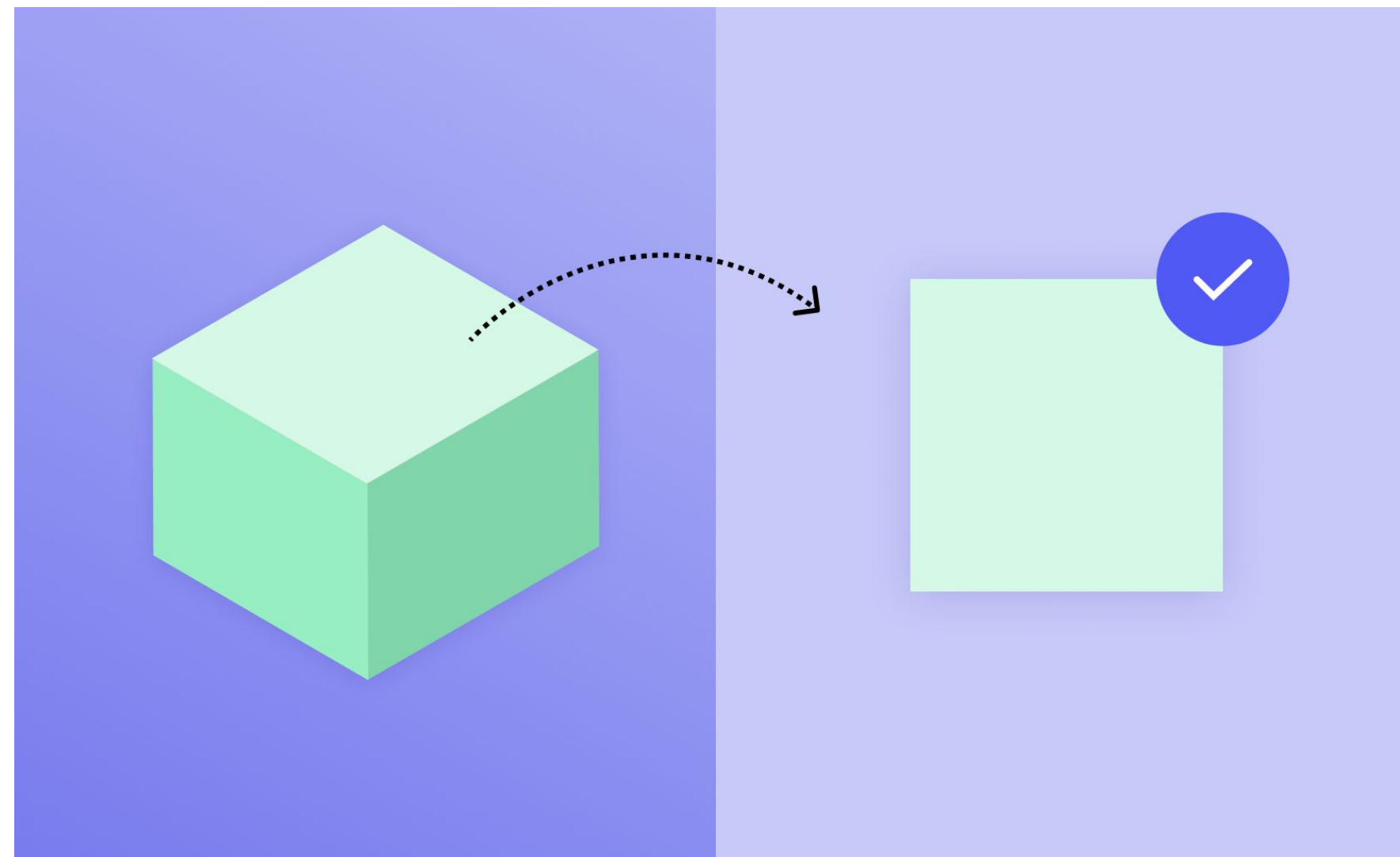
- Chuẩn hóa
- Mã hóa biến phân loại
- Trích xuất đặc trưng



VI. Tiền xử lý dữ liệu

3. Giảm dữ liệu (Data Reduction)

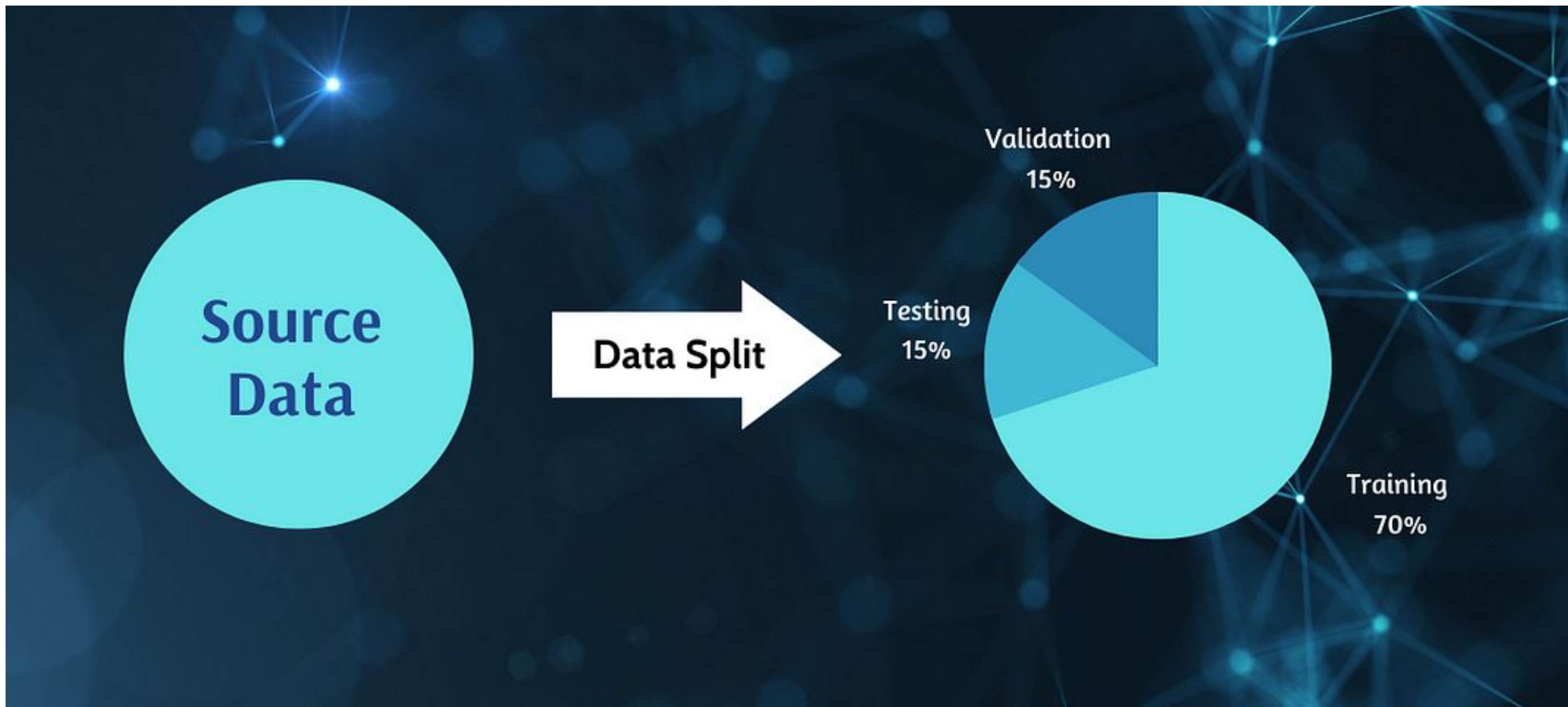
- Giảm chiều dữ liệu (PCA, SVD,)
- Sampling



VI. Tiền xử lý dữ liệu

3. Chia dữ liệu (Data Splitting)

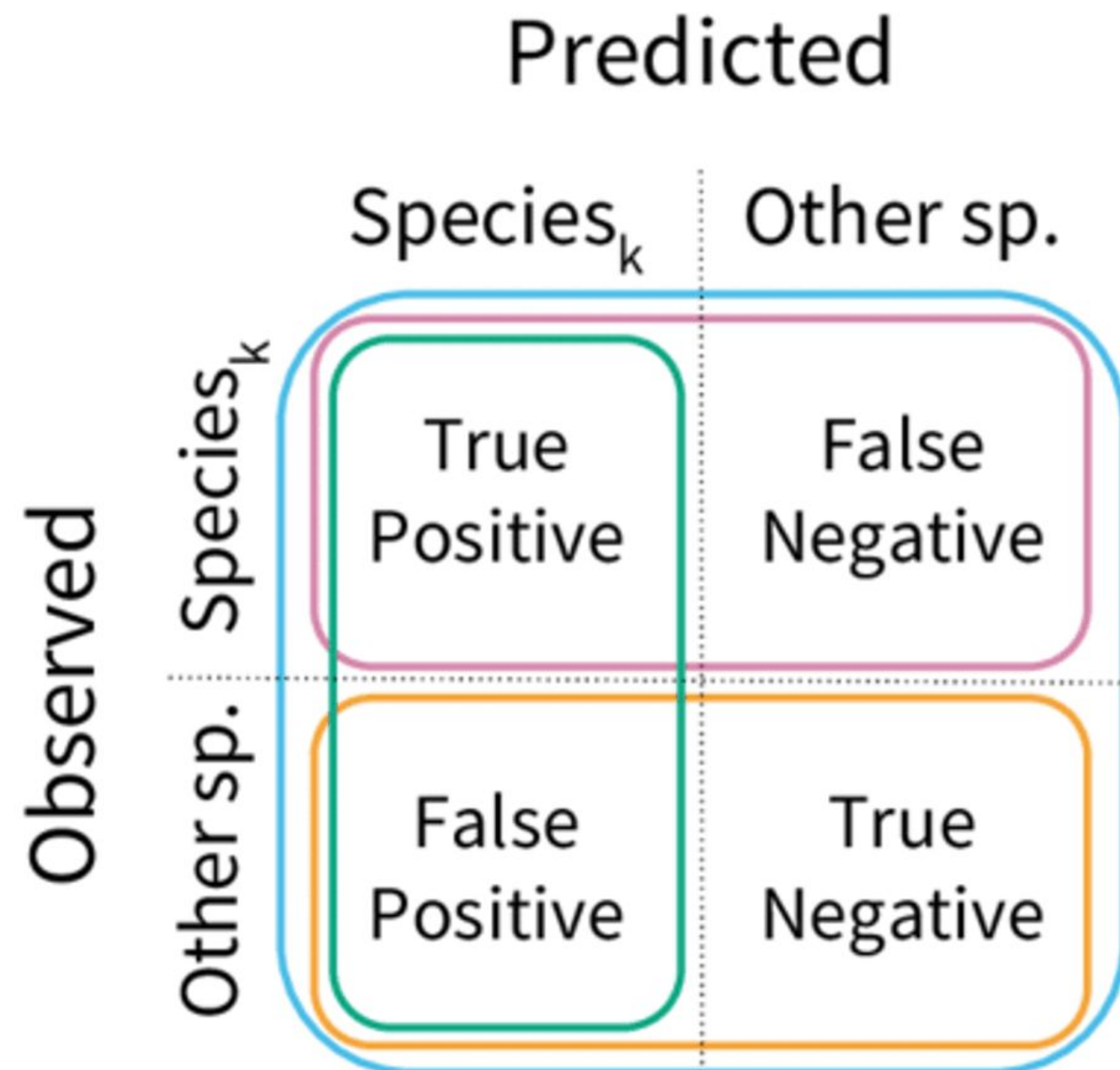
- Chia tập train, valid và test



VII. Đo lường hiệu suất

VII. Đo lường hiệu suất

Metric: dùng để đánh giá hiệu quả của một mô hình



Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$



Specificity = $\frac{TN}{TN + FP}$



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



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

VII. Đo lường hiệu suất

Xét bài toán phân loại một người có bị ung thư hay không bằng một mô hình ML, ta có:

Trường hợp	Kí hiệu
Người đó bị ung thư và mô hình dự đoán là bị ung thư	True Positive (TP)
Người đó không bị ung thư và mô hình dự đoán là bị ung thư	False Positive (FP)
Người đó không bị ung thư và mô hình dự đoán là không bị ung thư	True Negative (TN)
Người đó bị ung thư và mô hình dự đoán là không bị ung thư	False Negative (FN)

VII. Đo lường hiệu suất

Xét bài toán phân loại một người có bị ung thư hay không bằng một mô hình ML, ta có:

Accuracy: Số dự đoán đúng trên tổng số dự đoán

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

VII. Đo lường hiệu suất

Xét bài toán phân loại một người có bị ung thư hay không bằng một mô hình ML, ta có:

Precision: Số dự đoán **Positive** đúng trên tổng số dự đoán là **Positive**

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

VII. Đo lường hiệu suất

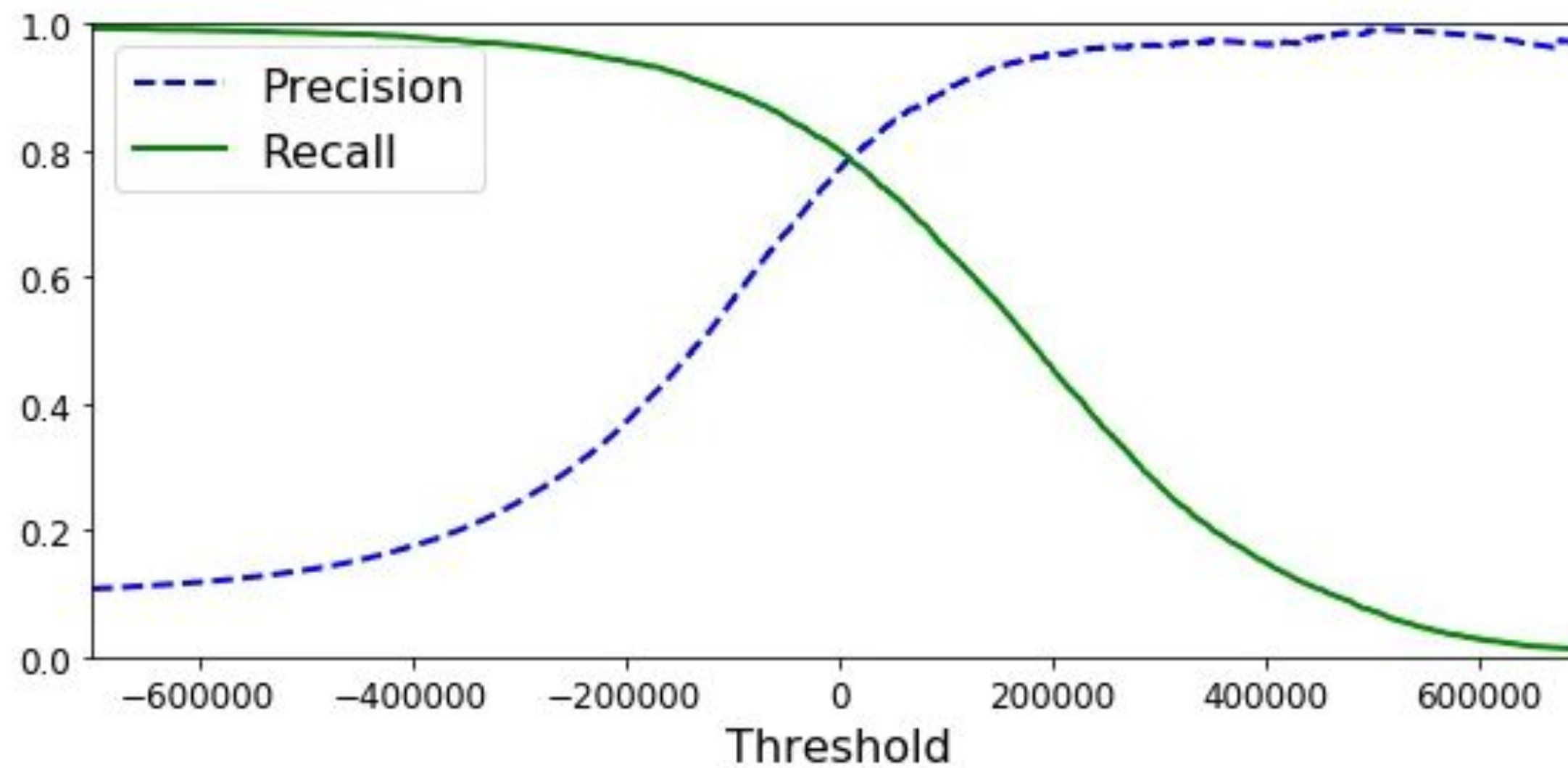
Xét bài toán phân loại một người có bị ung thư hay không bằng một mô hình ML, ta có:

Recall: Số dự đoán **Positive** đúng trên tổng số trường hợp là **Positive**

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

VII. Đo lường hiệu suất

Xét bài toán phân loại một người có bị ung thư hay không bằng một mô hình ML, ta có:



Precision-Recall Trade-off

VII. Đo lường hiệu suất

Xét bài toán phân loại một người có bị ung thư hay không bằng một mô hình ML, ta có:

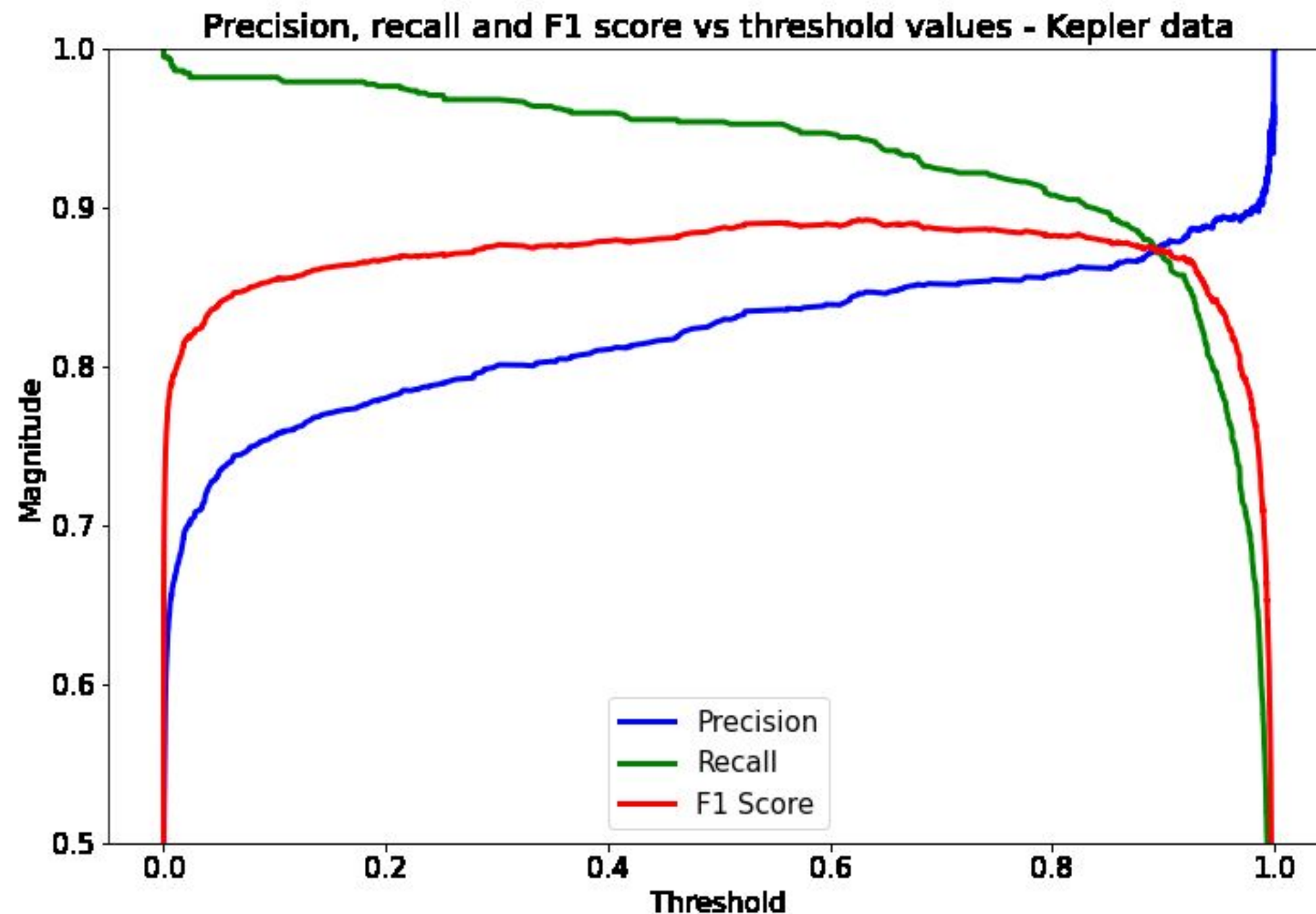
F1-score: Cân bằng giữa **Precision** và **Recall**

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

VII. Đo lường hiệu suất

Xét bài toán phân loại một người có bị ung thư hay không bằng một mô hình ML, ta có:

F1-score: Cân bằng giữa **Precision** và **Recall**



THANK YOU

CONTACT US

-  403.1 H6, BKHCM Campus 2
-  mliandiotlab@gmail.com
-  ml-iotlab.com
-  facebook.com/hcmut.ml.iot.lab
-  youtube.com/@mliotlab