



Bài giảng môn học: _____
Học Máy (Machine Learning)

Chương 2: Quy trình xây dựng một hệ thống học máy – Phần 1

Đặng Văn Nam
dangvannam@hmg.edu.vn

Nội dung chương 2

- 1. Các bước cơ bản xây dựng một mô hình học máy**
- 2. Một số nguồn dữ liệu cho học tập**
- 3. Thu thập và tiền xử lý dữ liệu**
 - 1. Ví dụ tập Data_Patient.csv**
 - 2. Ví dụ tập Data_Titanic.csv**
- 4. Bài tập thực hành chương 2**

1. Các bước cơ bản xây dựng một mô hình học máy.

1. Các bước xây dựng một mô hình học máy



Để xây dựng một mô hình học máy thực hiện qua 7 bước:

- 1) Xác định bài toán và Thu thập dữ liệu (Data collection)
- 2) Chuẩn bị dữ liệu (Data preparation)
- 3) Lựa chọn mô hình phù hợp (Choosing a model)
- 4) Huấn luyện mô hình (Training)
- 5) Đánh giá mô hình (Evaluation)
- 6) Nâng cao độ chính xác của mô hình (Improve Model Accuracy)
- 7) Dự đoán với mô hình xây dựng được (Prediction)

Step 1. Xác định bài toán và thu thập dữ liệu



1. Xác định bài toán và thu thập dữ liệu

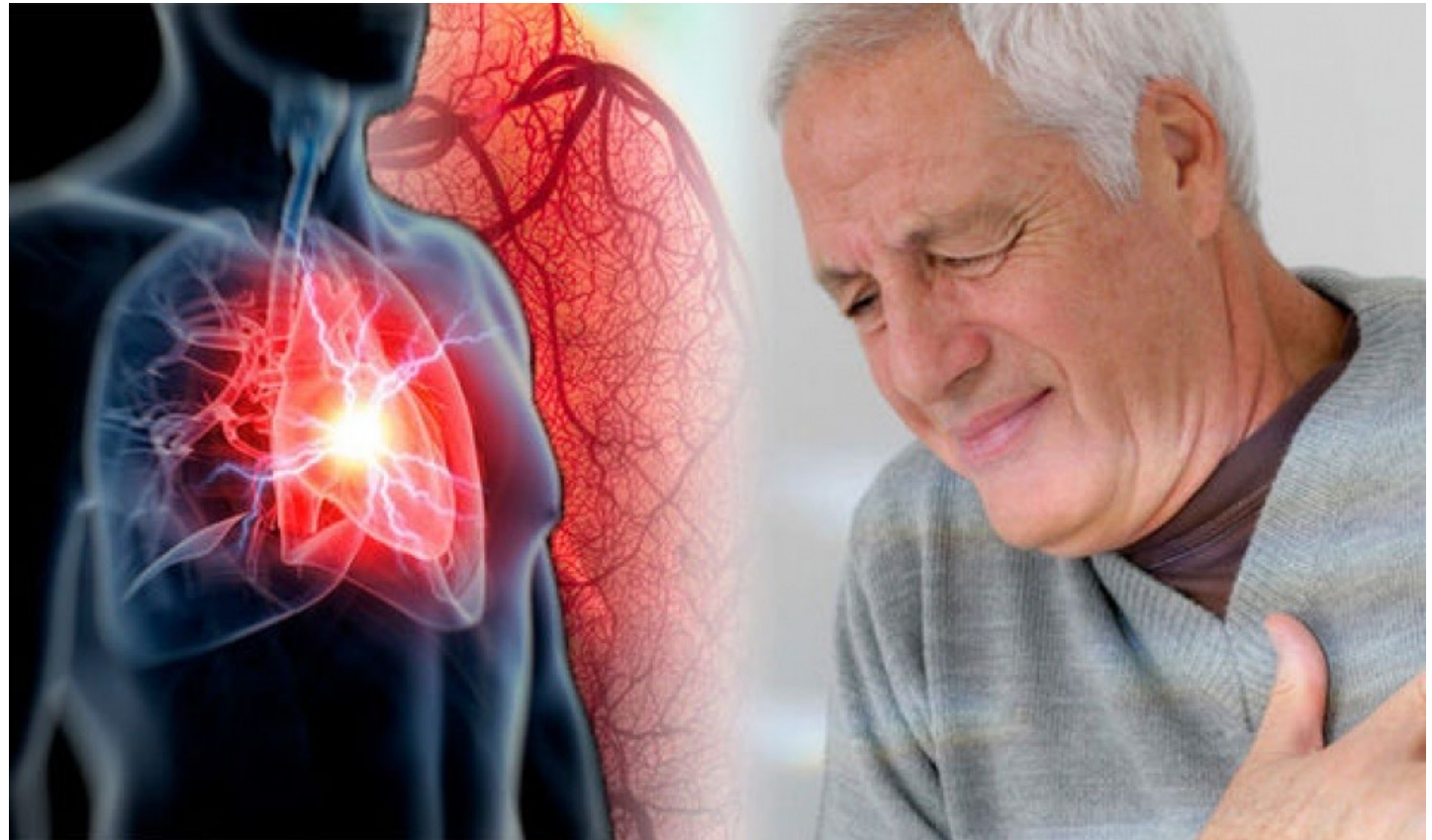
Từ các vấn đề xuất phát từ thế giới thực, xác định bài toán cần giải quyết:



1. Xác định bài toán và thu thập dữ liệu

Ví dụ: Bác sĩ AI

- Xây dựng một mô hình học máy cho bài toán dự đoán một bệnh nhân có bị bệnh đau tim hay không?



1. Xác định bài toán và thu thập dữ liệu

- Sau khi xác định được bài toán và mục tiêu cần giải quyết, cần phải thu thập các dữ liệu liên quan.
- Dữ liệu có thể được thu thập từ nhiều nguồn khác nhau như *files, database, internet, mobile devices...*
- **Số lượng và chất lượng** dữ liệu thu thập được sẽ quyết định đến **độ chính xác của mô hình** càng cao.
- Bước này bao gồm các công việc:
 - Xác định các nguồn dữ liệu liên quan đến bài toán
 - Thu thập các dữ liệu từ các nguồn này
 - Tích hợp các dữ liệu thu thập được để tạo thành một tập dữ liệu nhất quán (Dataset) sử dụng cho các bước tiếp theo./

1. Xác định bài toán và thu thập dữ liệu

▪ Bác sĩ AI:



- Tên Dataset: **Data_Patient.csv**
- File dữ liệu chứa thông tin của **300 bệnh nhân trong quá khứ**
- Mỗi dòng ứng với thông tin của một bệnh nhân, bao gồm **9 thuộc tính**:
 - **id**: Mã của bệnh nhân (object)
 - **Age**: Tuổi của bệnh nhân (số)
 - **Gender**: Giới tính của bệnh nhân (chuỗi: Male – Female)
 - **Type**: Cho biết loại triệu chứng đau ngực mà bệnh nhân này mắc phải, với 4 giá trị: (Typical angina, Atypical angina, Non-anginal pain, Asymptomatic)
 - **Blood_pressure**: Huyết áp của bệnh nhân – đơn vị: mmhg (số)
 - **Cholesterol**: Chỉ số cholesterol của bệnh nhân – đơn vị: mg/dl (số)
 - **Heartbeat**: Thông số nhịp tim của bệnh nhân – đơn vị: lần/phút (số)
 - **Thalassemia**: Chỉ số Thalassemia của bệnh nhân chỉ gồm 3 giá trị (3: Bình thường | 4: Khiếm khuyết cố định | 7: Kiểm khuyết có thể đảo ngược)
 - **Result**: Cho biết bệnh nhân có bị bệnh tim hay không? (0: Không bị bệnh tim mạch | 1: Bị bệnh tim mạch)

1. Xác định bài toán và thu thập dữ liệu

■ Bác sĩ AI:



	A	B	C	D	E	F	G	H	I
1	id	Age	Gender	Type	Blood_pressure	Cholesterol	Heartbeat	Thalassemia	Result
2	Patient_01	63	Male	Typical angina	145	233	150	6	0
3	Patient_02	67	Male	Asymptomatic	160	286	108	3	1
4	Patient_03	67	Male	Asymptomatic	120	229	129	7	1
5	Patient_04	37	Male	Non-anginal pain	130	250	187	3	0
6	Patient_05	41	Female	Atypical angina	130	204	172		0
7	Patient_16	56	Male	Atypical angina	120	236	178	3	0
8	Patient_07	62	Female	Asymptomatic	140	268	160	3	1
9	Patient_08	57	Female	Asymptomatic	120	354	163	3	0
10	Patient_19	63	Male	Asymptomatic	130	254	147	7	1
11	Patient_10	53	Male	Asymptomatic	140	203	155	7	1
12	Patient_110	57	Male	Asymptomatic	140	192	148	6	0
13	Patient_120	56	Female	Atypical angina	140	294	153	3	0
14	Patient_130	56	Male	Non-anginal pain	130	256	142	6	1
15	Patient_140	44	Male	Atypical angina	120	263	173	7	0
16	Patient_150	52	Male	Non-anginal pain	172	199	162	7	0
17	Patient_160	57	Male	Non-anginal pain	150	168	174	3	0
18	Patient_170	48	Male	Atypical angina	110	229	168	7	1
19	Patient_180	54	Male	Asymptomatic	140	239	160	3	0
20	Patient_190	48	Female	Non-anginal pain	130	275	139	3	0
21	Patient_200	49	Male	Atypical angina	130	266	171	3	0
22	Patient_21	64	Male	Typical angina	110	211	144		0

Data_Patient

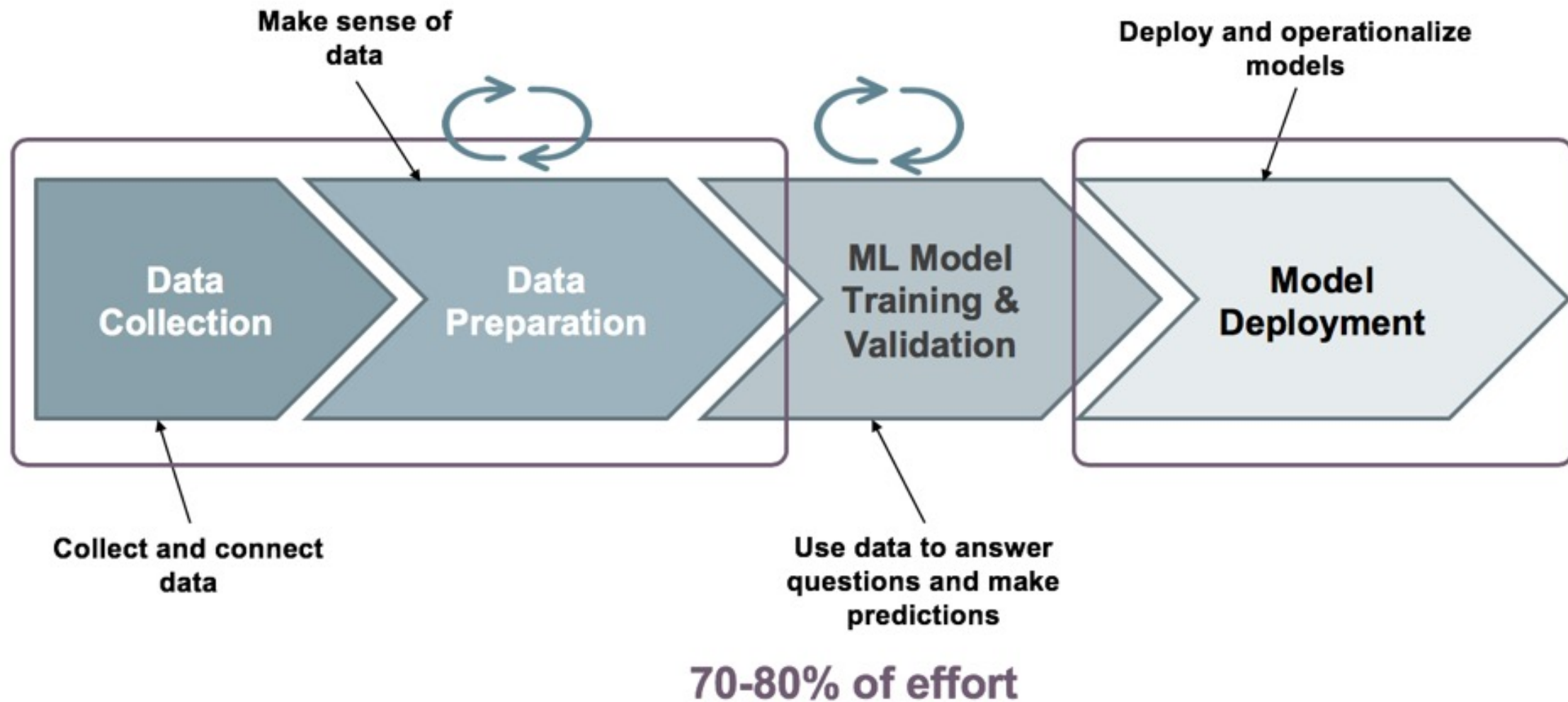
Step 2.Chuẩn bị dữ liệu



2. Chuẩn bị dữ liệu

- Thông thường dữ liệu thu thập được có rất nhiều nhiễu (noise), dữ liệu thiếu (missing value), dữ liệu ngoại lai (outliers)...do đó dữ liệu cần phải được làm sạch và chuẩn hóa về dạng phù hợp. Đây là giai đoạn chiếm nhiều thời gian và nguồn lực nhất của một dự án ML.
- Có rất nhiều nhiệm vụ phải thực hiện trong quá trình chuẩn bị dữ liệu. Một số vấn đề cơ bản cần giải quyết trong giai đoạn này bao gồm:
 - ✓ Khám phá dữ liệu
 - ✓ Làm sạch dữ liệu (xử lý giá trị thiếu, giá trị ngoại lai)
 - ✓ Tích hợp dữ liệu
 - ✓ Biến đổi, rời rạc hóa và chuẩn hóa dữ liệu
 - ✓ Cân bằng dữ liệu.
 - ✓ Rút gọn thuộc tính.

2. Chuẩn bị dữ liệu

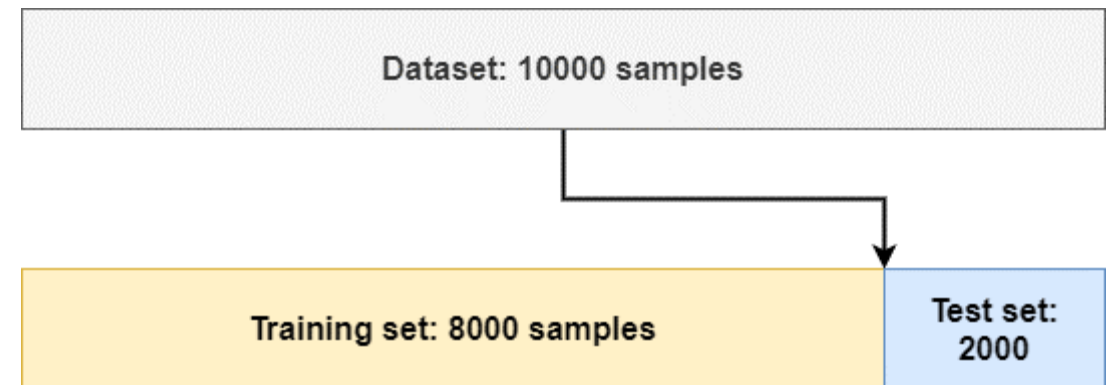
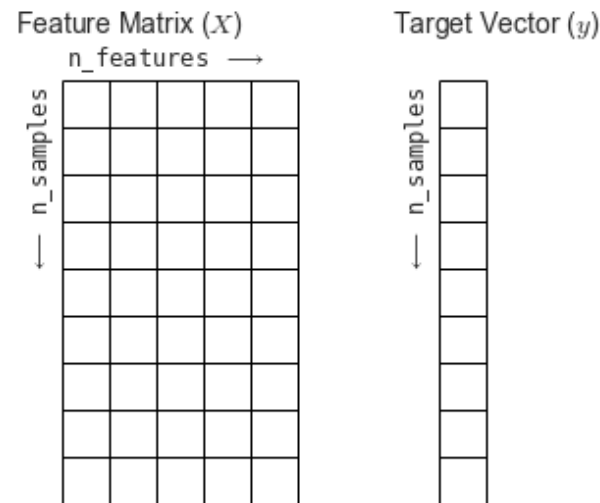


2. Chuẩn bị dữ liệu

Bác sĩ AI



- Quan sát và khám phá tập dữ liệu
- Phát hiện và xử lý dữ liệu khuyết thiếu trong tập dữ liệu
- Mã hóa dữ liệu Categorical (Encoding Categorical Data)
- Trích xuất biến độc lập, biến phụ thuộc
- Phân tách tập dữ liệu huấn luyện (Train) – Kiểm thử (Test)

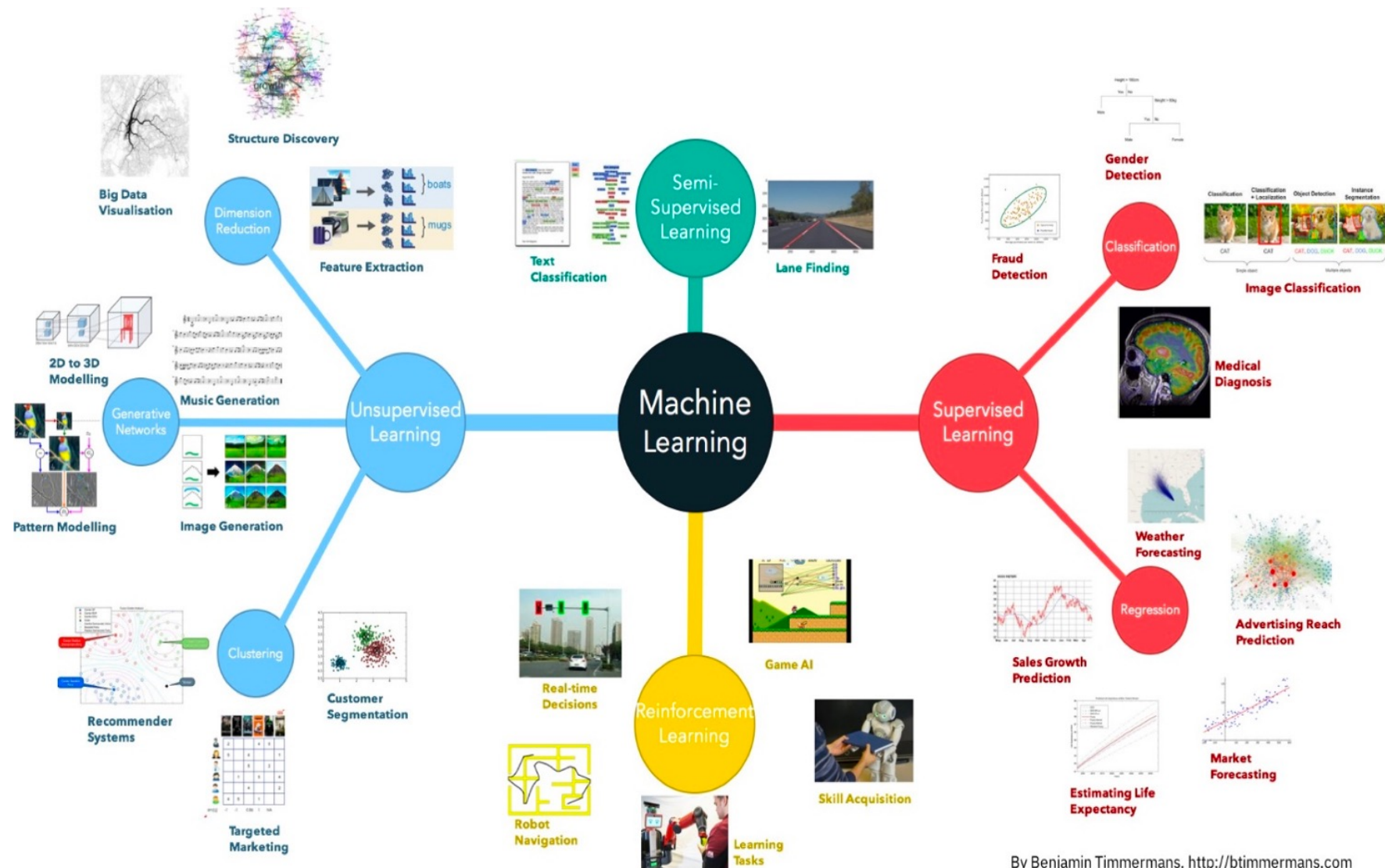


Step 3. Lựa chọn mô hình học máy



3. Lựa chọn mô hình học máy

Có rất nhiều loại học máy khác nhau, mỗi loại, mỗi mô hình phù hợp với từng bài toán, từng tập dữ liệu cụ thể
→ Cần xác định xem thuộc lớp bài toán nào



3. Lựa chọn mô hình học máy

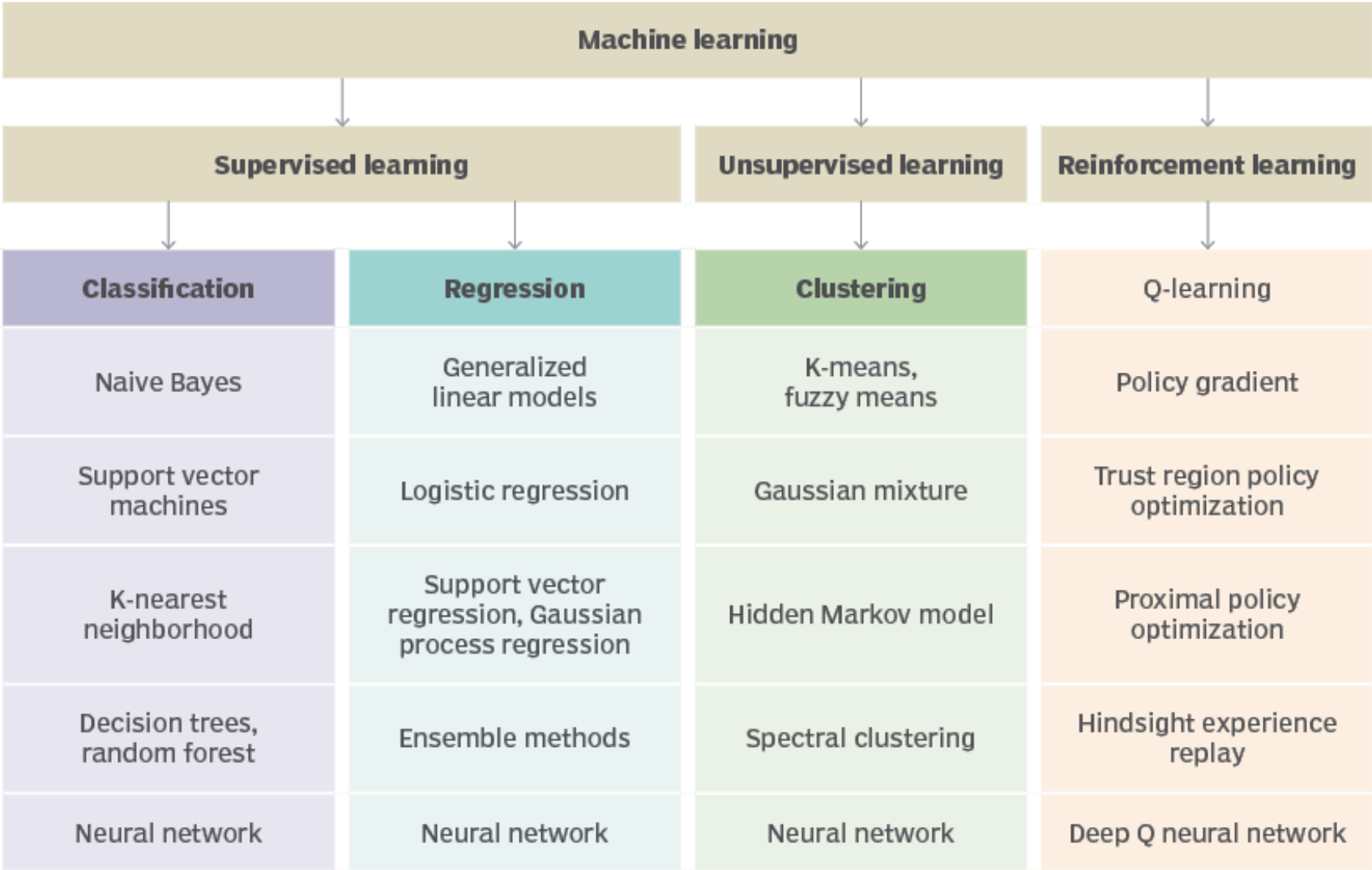
Bác sĩ AI



- Dữ liệu bài toán có gán nhãn (7 thuộc tính độc lập – 1 thuộc tính phụ thuộc (result) → **Học có giám sát (Supervised Learning)**.
- Nhãn là các giá trị rời rạc → **Phân lớp (classification)**
- Nhãn chỉ có 2 giá trị (0 – Không bị | 1 – Bị Bệnh) → **Phân lớp nhị phân (Binary classification)**

3. Lựa chọn mô hình học máy

Bác sĩ AI

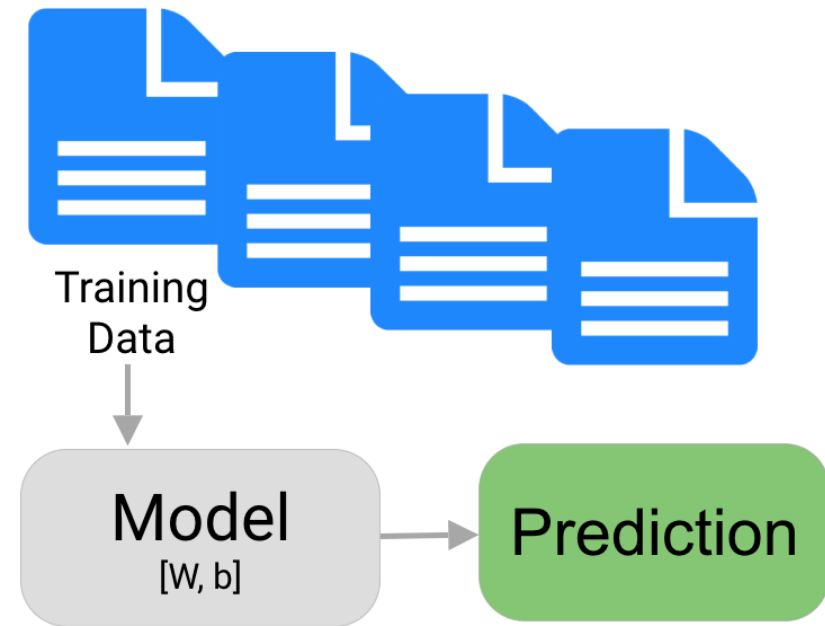
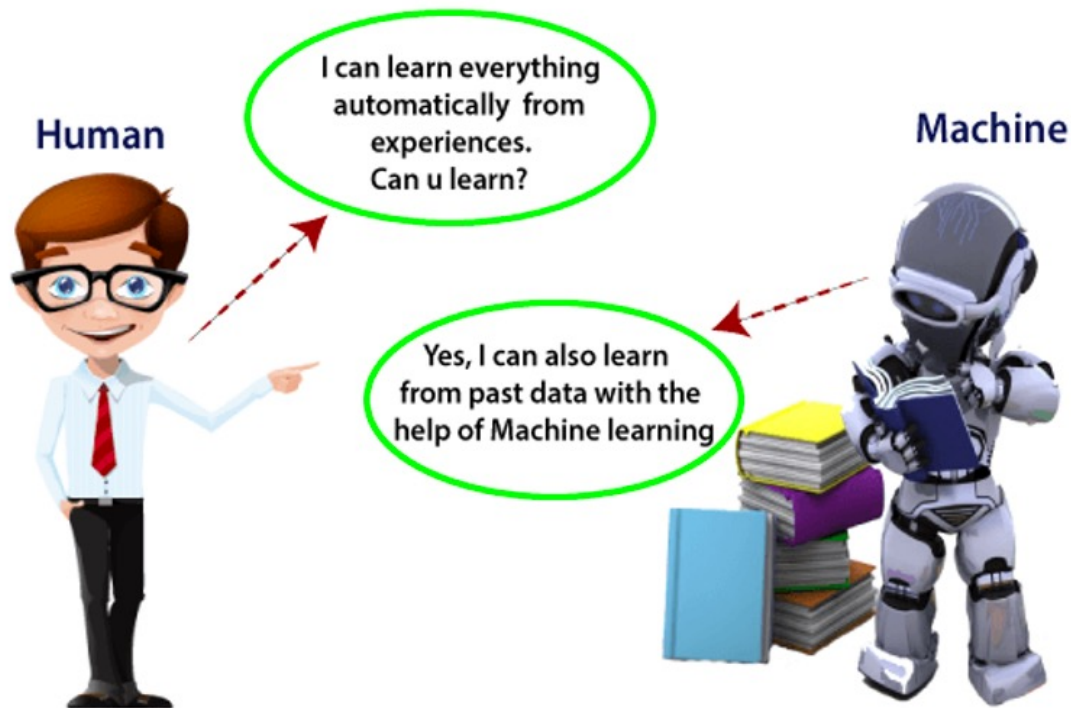


Step 4. Huấn luyện mô hình học máy



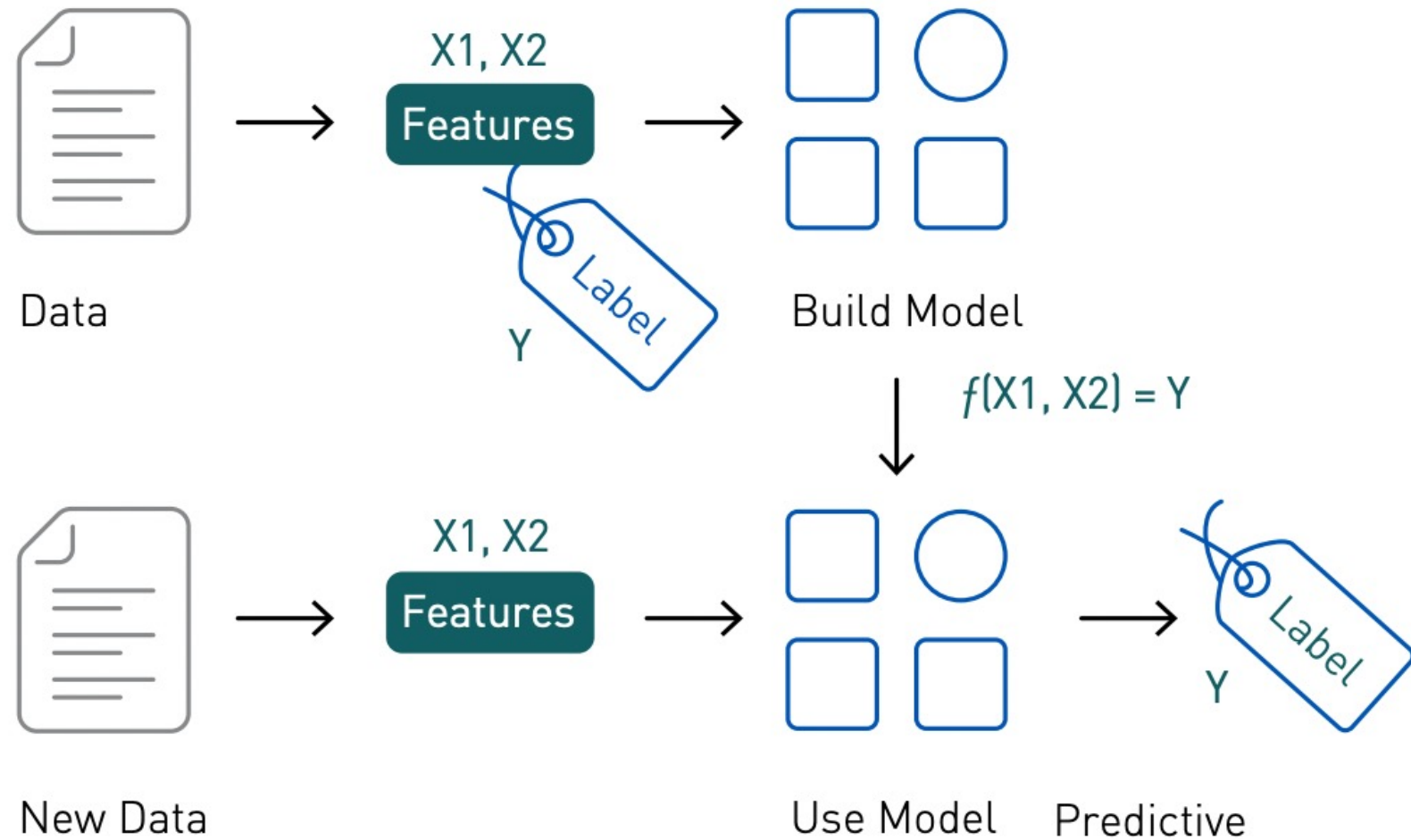
4. Huấn luyện mô hình học máy

- Các mô hình học máy sẽ học từ dữ liệu, và để cho máy học được chúng ta sẽ sử dụng tập huấn luyện (Training set).
- Việc huấn luyện một mô hình học máy bản chất là tìm gia các tham số tối ưu cho thuật toán đó, sao cho độ chính xác của thuật toán là cao nhất.



4. Huấn luyện mô hình học máy

- Machine learning ~ Tìm hàm số

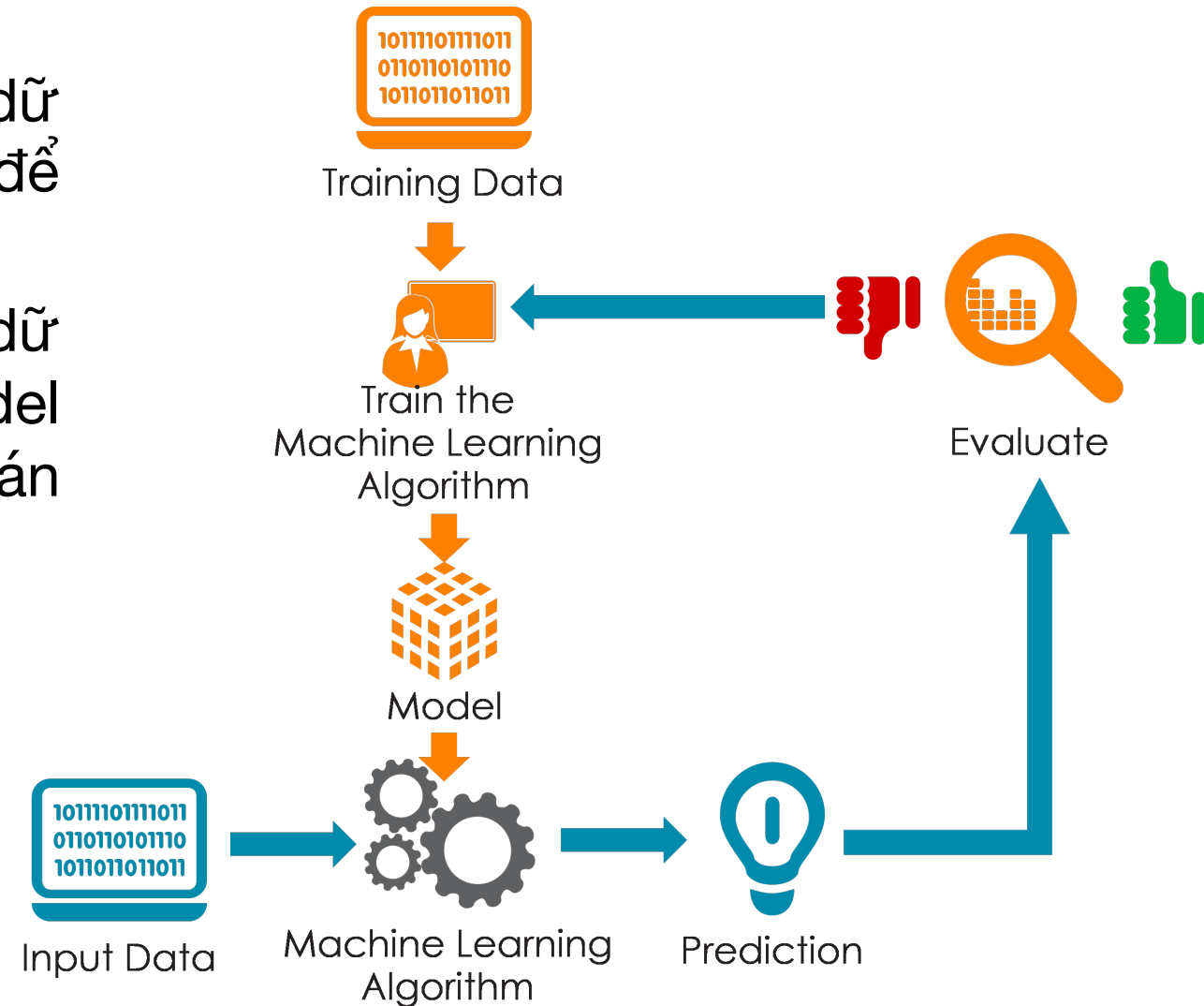
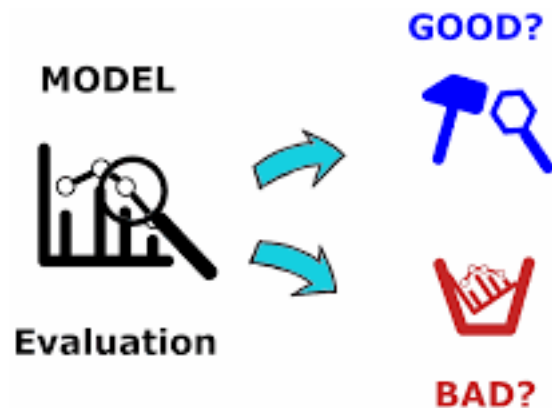


Step 5. Đánh giá mô hình



5. Đánh giá mô hình

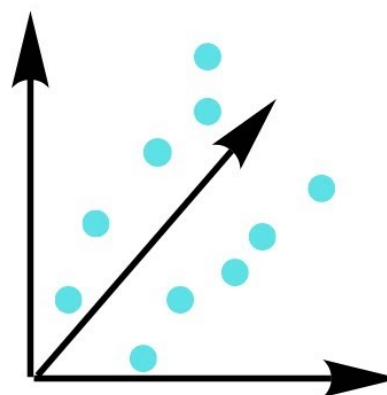
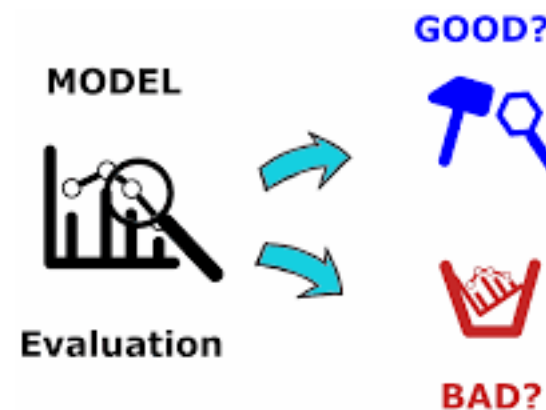
- Mô hình sau khi được huấn luyện với dữ liệu huấn luyện, cần phải được kiểm tra để đánh giá độ chính xác của mô hình.
- Trong giai đoạn này sẽ sử dụng tập dữ liệu Test (Tập dữ liệu độc lập và Model chưa biết tới các dữ liệu này) để dự đoán với model thu được.



5. Đánh giá mô hình

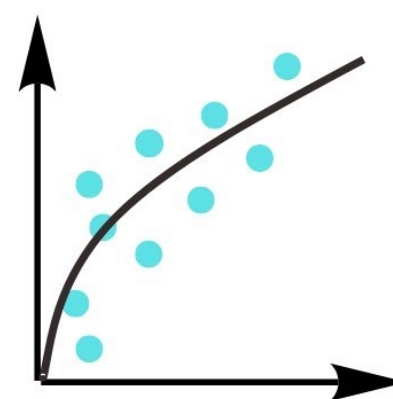
Một mô hình học máy có thể rơi vào các trường hợp sau:

- Độ chính xác trên tập huấn luyện (Training data) và tập kiểm thử (Testing data) đều thấp: **Underfitting**
- Độ chính xác trên tập huấn luyện cao nhưng độ chính xác trên tập kiểm thử lại thấp: **Overfitting**
- Độ chính xác trên tập huấn luyện và kiểm thử đều cao: **Best fit, Good fit**

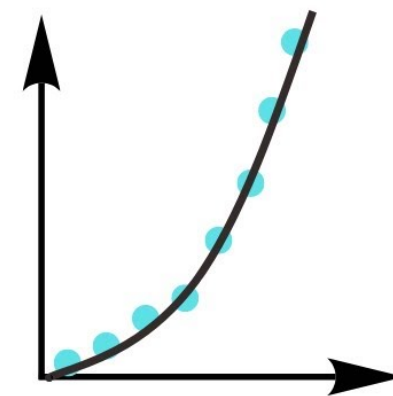


Underfitting

Training Data AC Down
Testing Data AC Down



Good Model



Overfitting

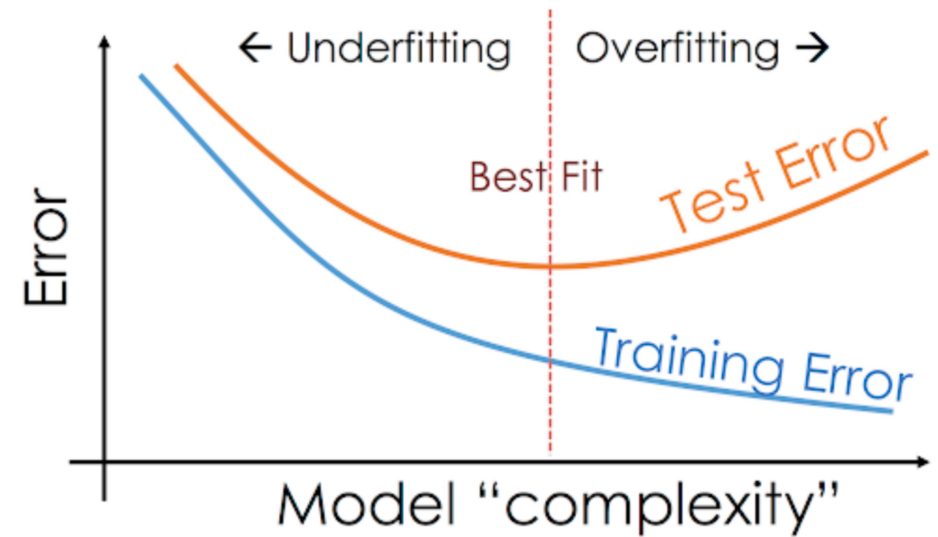
Training Data AC High
Testing Data AC Down

Step 6. Nâng cao độ chính xác của mô hình



6. Nâng cao độ chính xác của mô hình

- Trong trường hợp mô hình rơi vào trạng thái Underfitting hoặc Overfitting, cần phải có các phương pháp để nâng cao độ chính xác, đưa mô hình về trạng thái Best fit – **ngưỡng chấp nhận được!**
- Một số phương pháp nâng cao độ chính xác của mô hình:
 - Thu thập thêm dữ liệu, Nâng cao chất lượng dữ liệu xây dựng mô hình.
 - Thay đổi lựa chọn mô hình, thuật toán học máy phù hợp hơn.
 - Giữ nguyên mô hình hiện tại, thay đổi các tham số của mô hình để tìm ra tham số tối ưu



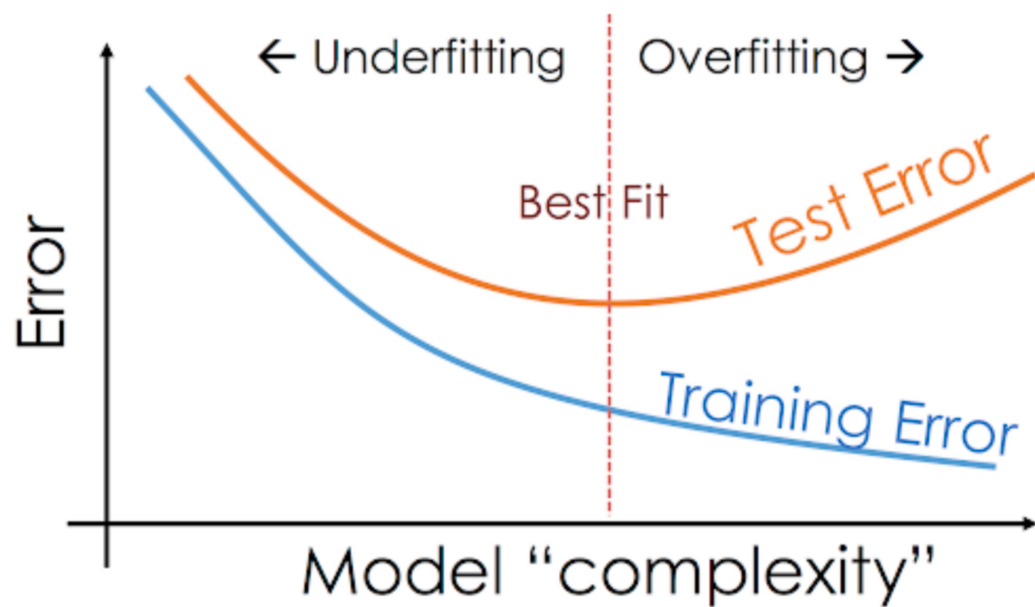
**Methods to Improve
Model Accuracy**



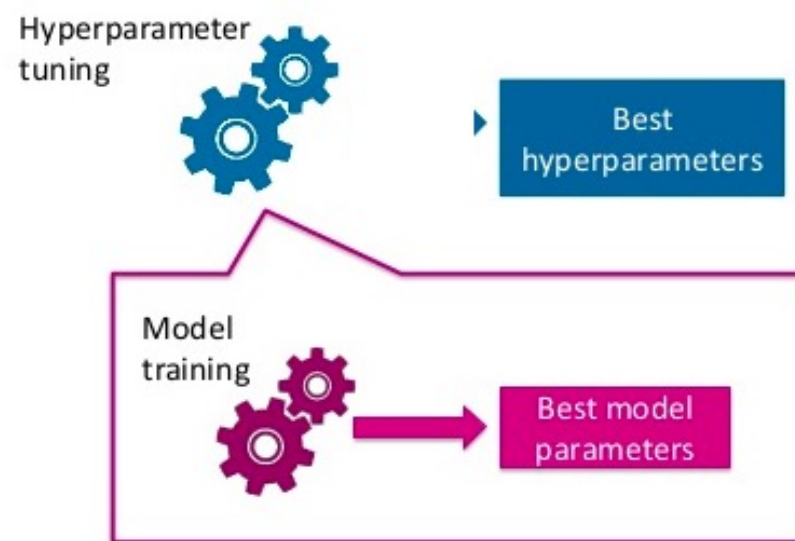
www.gradient.run

6. Nâng cao độ chính xác của mô hình: Tùy chỉnh tham số

- Bản chất của việc học là tìm ra tham số phù hợp với model sao cho độ chính xác đạt đến ngưỡng chấp nhận được.
- Việc Tuning tham số là việc xác định tham số nào là tốt nhất cho model với dữ liệu hiện tại sao cho độ chính xác của model trên cả tập huấn luyện và tập kiểm tra đều cao (best fit)



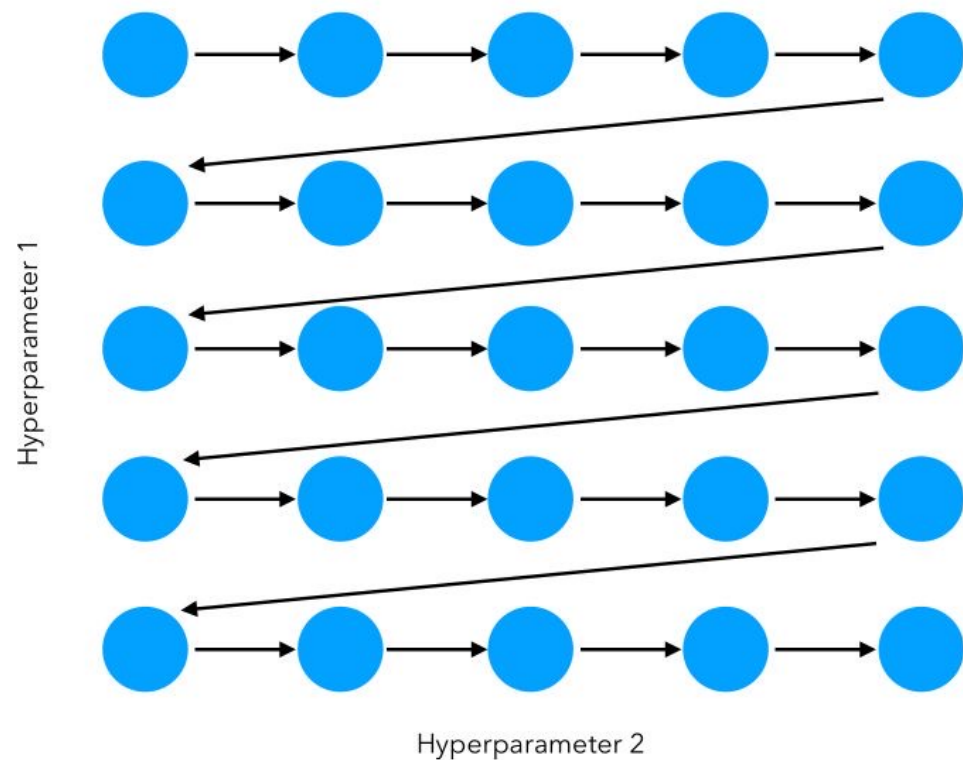
Hyperparameter tuning vs. model training



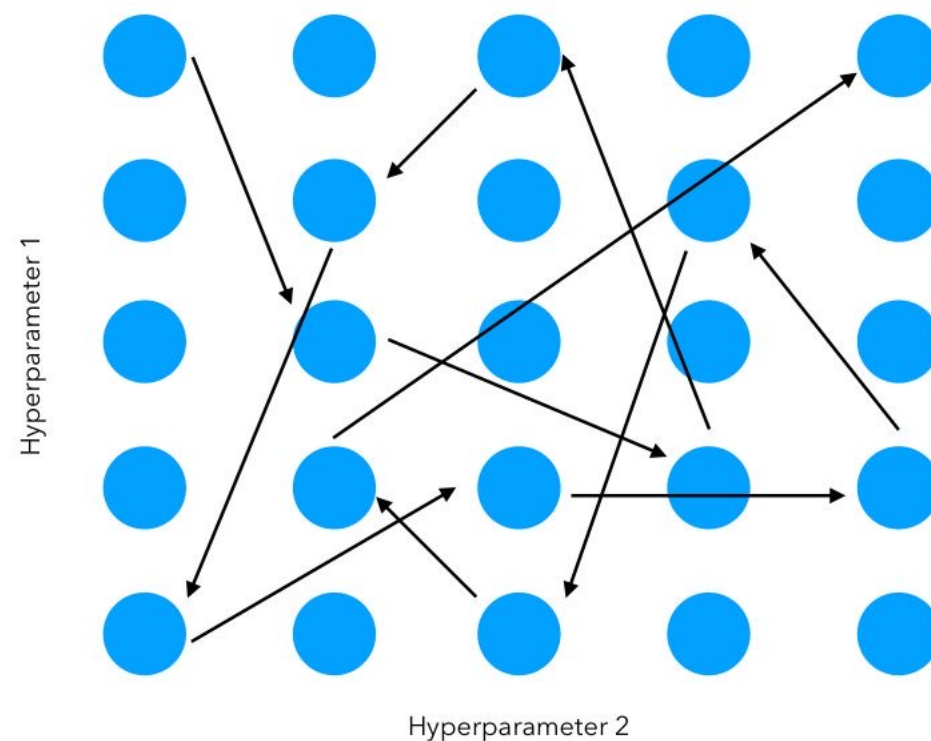
6. Tùy chỉnh tham số của mô hình

- 2 phương pháp phổ biến để fine-tune tự động các tham số của mô hình đó là Grid Search và Randomized Search

Grid Search



Randomized Search

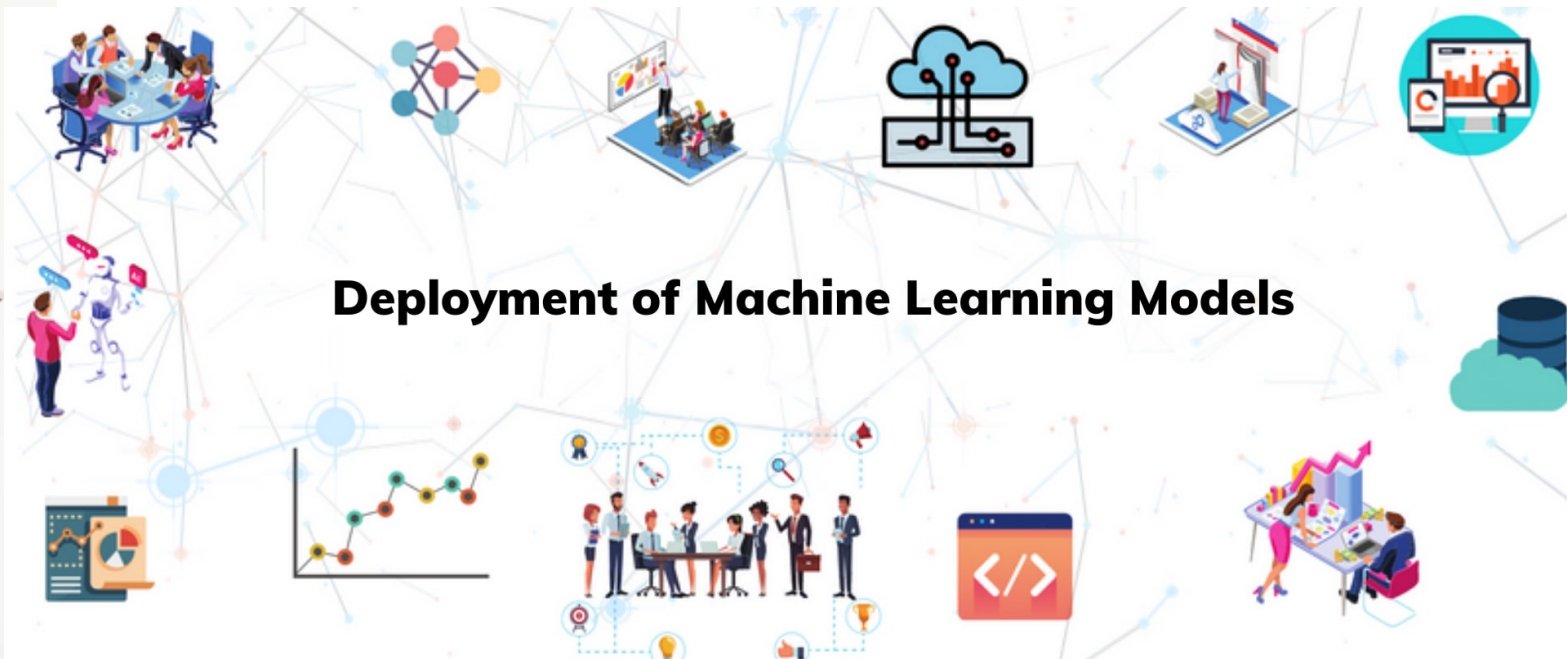
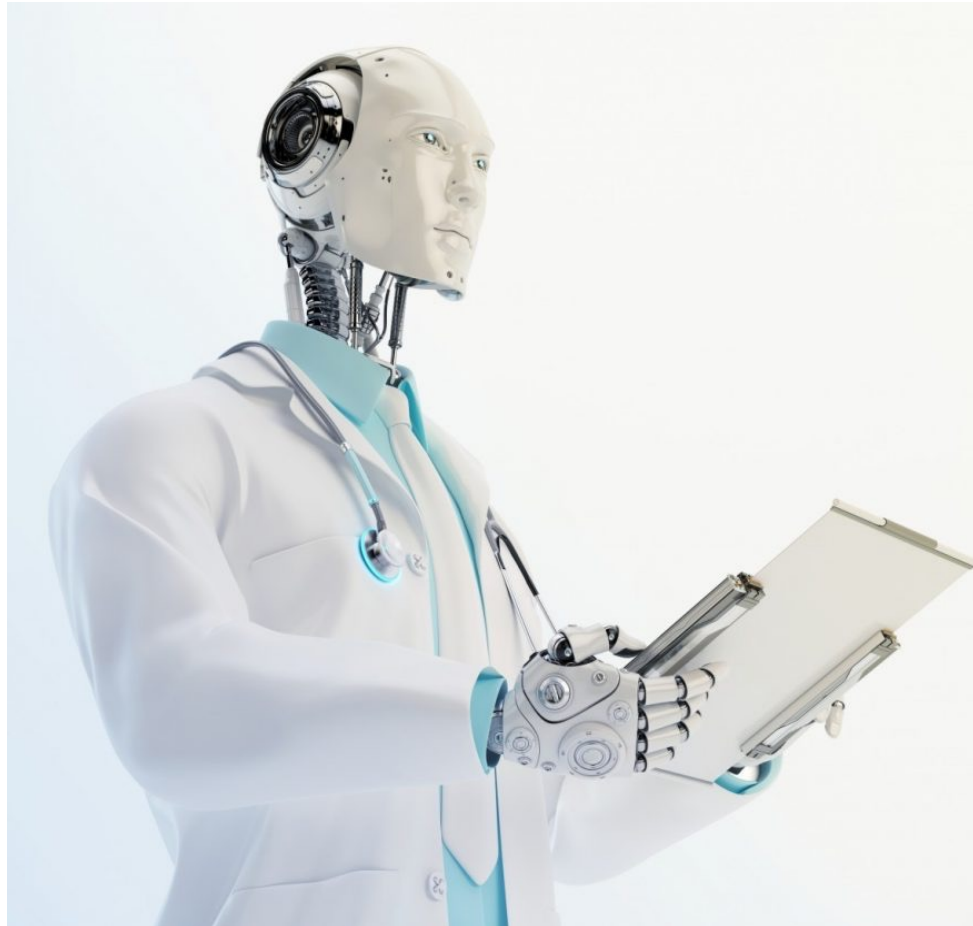


Step 7. Sử dụng mô hình đã xây dựng

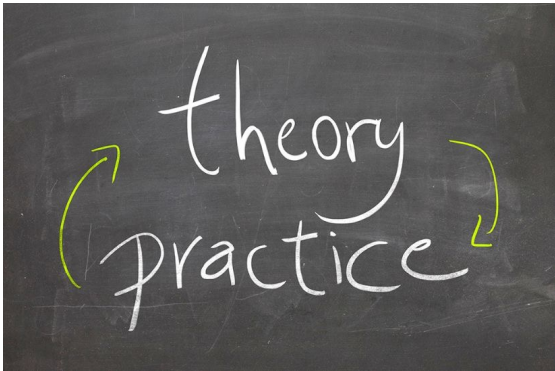
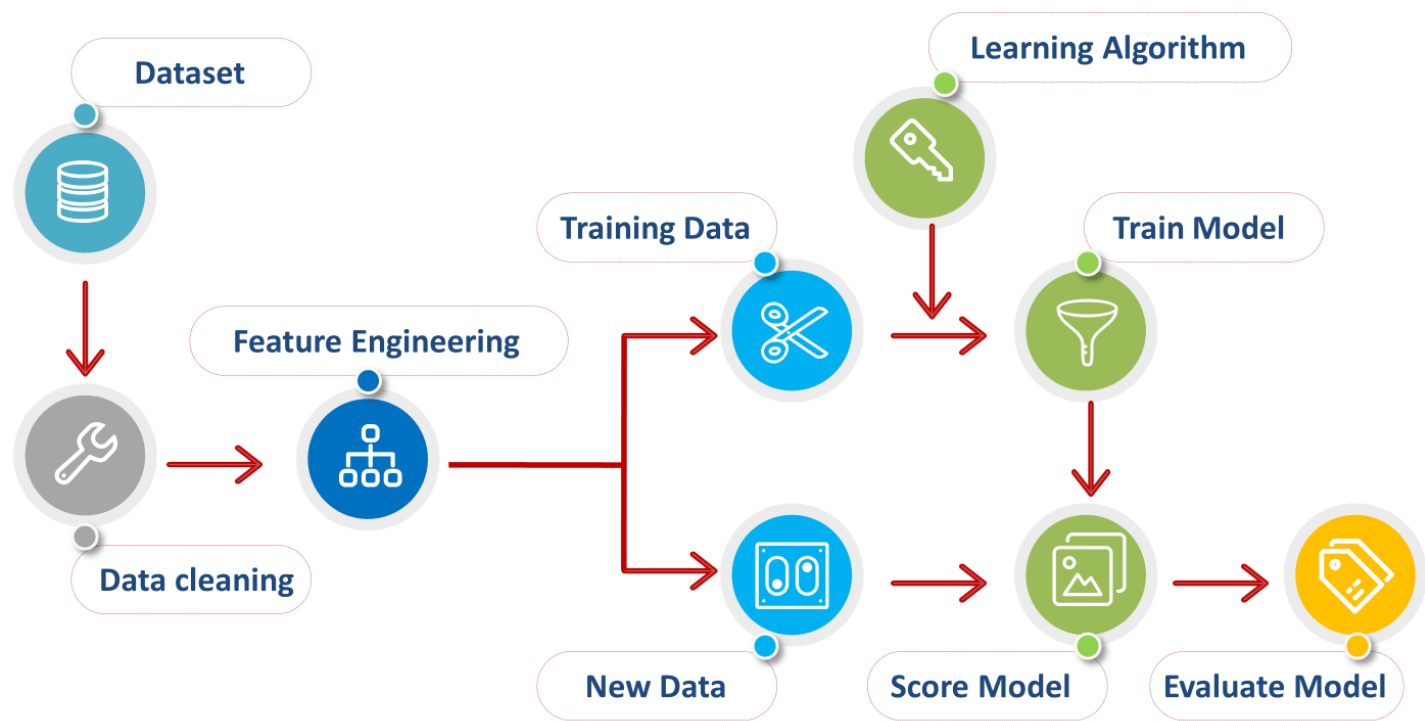


7. Sử dụng mô hình đã xây dựng

- Sau khi đã huấn luyện được model và kiểm thử trên tập dữ liệu test đạt độ chính xác chấp nhận được. Sử dụng model để dự đoán kết quả




Sinh viên theo dõi quy trình các bước xây dựng một model học máy trong Jupyter notebook



YÊU CẦU:

Sinh viên Restart & Clear Output, thực hiện lại các bước của ví dụ này 3 lần

jupyter Chuong2_QuyTrinhHocMay_P1 Last Checkpoint: 30 minutes ago (autosaved)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Interrupt Restart Restart & Clear Output Restart & Run All Reconnect Shutdown Change kernel

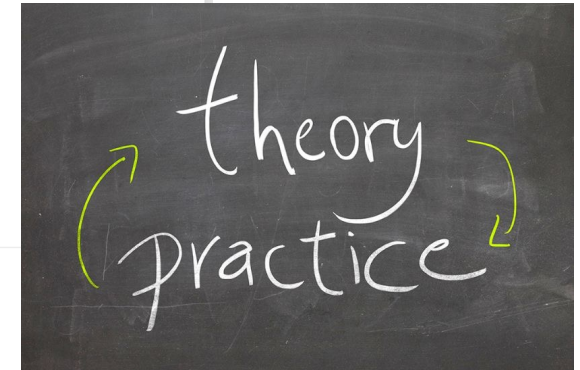
TRƯỜNG ĐẠI HỌC MỎ - ĐỊA CHẤT
KHOA CÔNG NGHỆ THÔNG TIN

BÀI 2: CÁC BƯỚC XÂY DỰNG MỘT MÔ HÌNH HỌC MÁY

Các bước để xây dựng một mô hình học máy nói chung:



```
graph LR; S01[STEP 01  
Data collection] --> S02[STEP 02  
Data preparation]; S02 --> S03[STEP 03  
Choosing a model]; S03 --> S04[STEP 04  
Training]; S04 --> S05[STEP 05  
Evaluation]; S05 --> S06[STEP 06  
Improve model]; S06 --> S07[STEP 07  
Prediction];
```





Thank you!