INTRO TO MACHINE LEARNING

Nguyễn Sơn

Email: sonnhfit@gmail.com

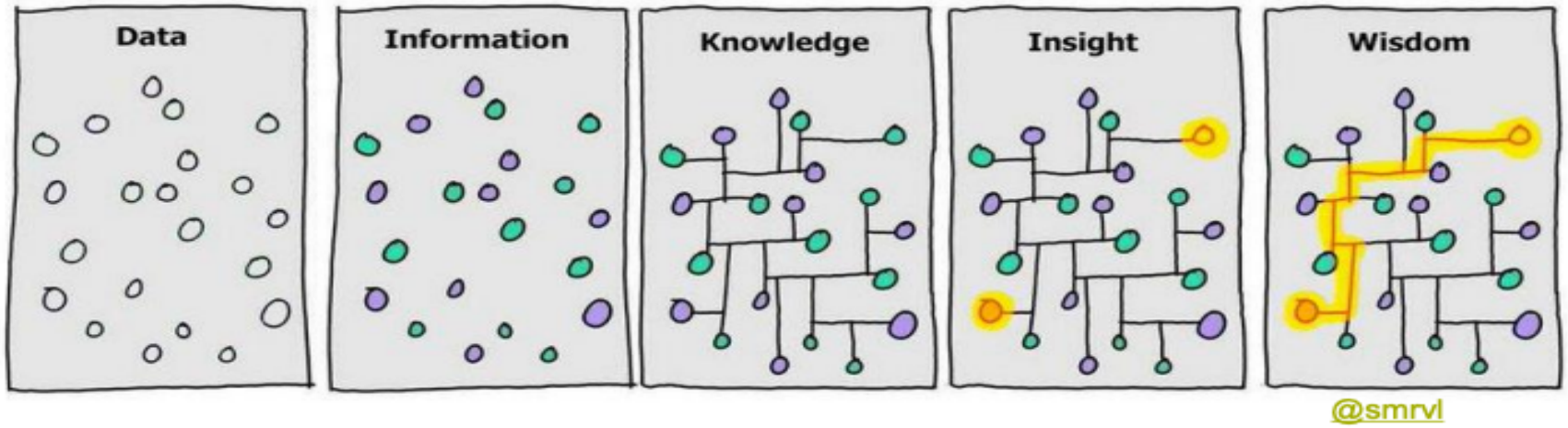
WHAT IS ML?

- Các chương trình hoạt động bên ngoài giống như con người
- Các chương trình hoạt động nội bộ như con người
- Hệ thống tính toán hoạt động thông minh

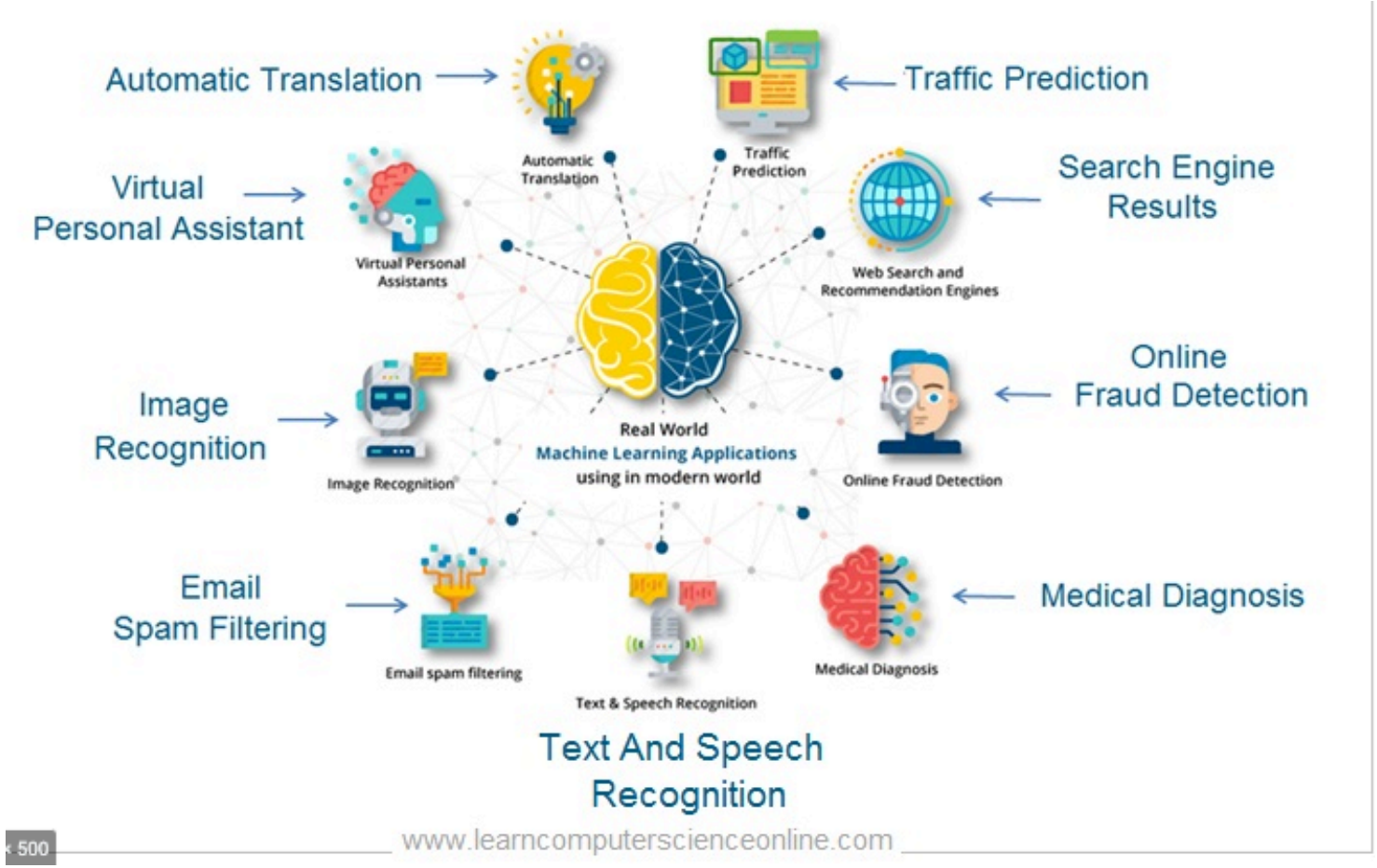
- HỌc máy là việc trích xuất các quy luật từ dữ liệu và khai thác chúng

ML PROBLEMS

AI	ML
Trường bao quát	Tập hợp con của AI
Mục đích là mô phỏng trí thông minh của con người để giải quyết các vấn đề phức tạp	Mục tiêu học hỏi từ dữ liệu và dự đoán kết quả khi dữ liệu mới được trình bày
Dẫn đến trí thông minh hoặc sự khôn ngoan(wisdom)	Dẫn đến kiến thức
Cố gắng tìm ra giải pháp tối ưu	Cố gắng tìm ra giải pháp duy nhất cho dù đó là tối ưu hay không

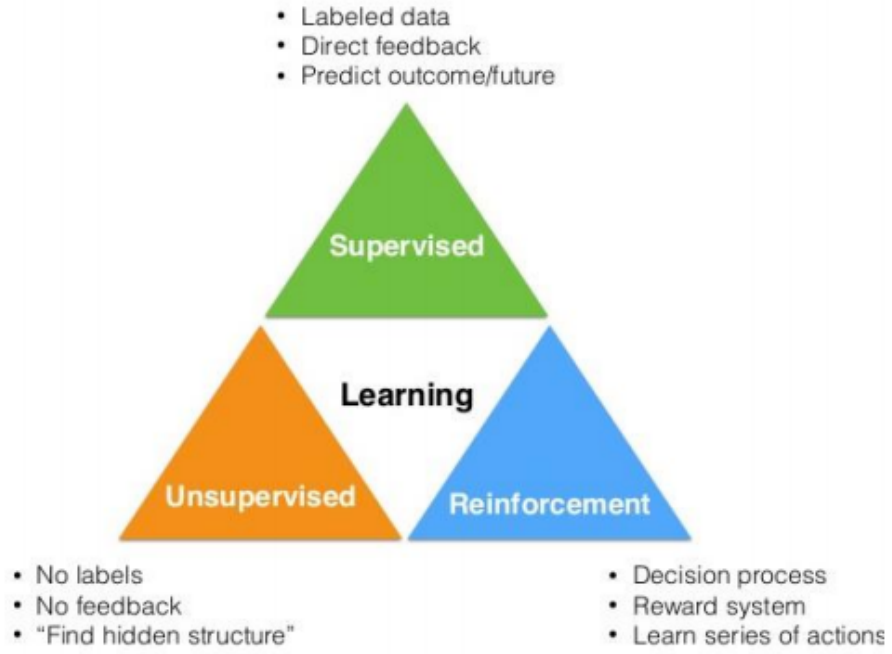


MACHINE LEARNING APPLICATION

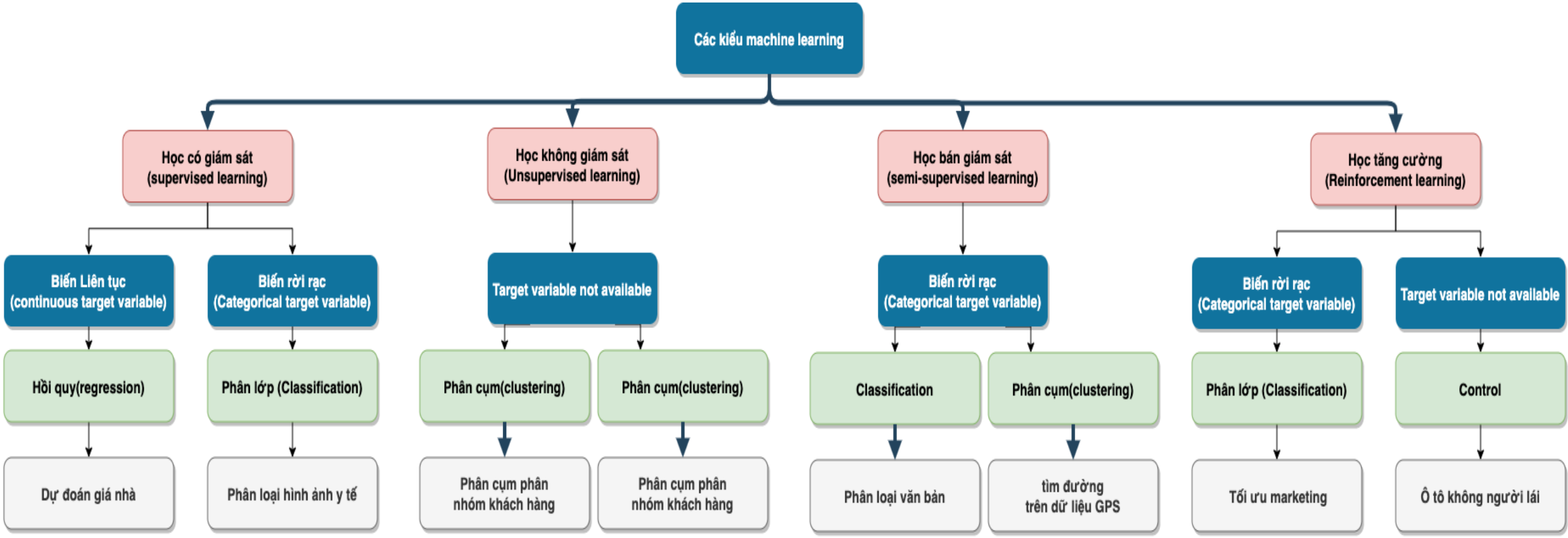


LEARNING TECHNIQUES

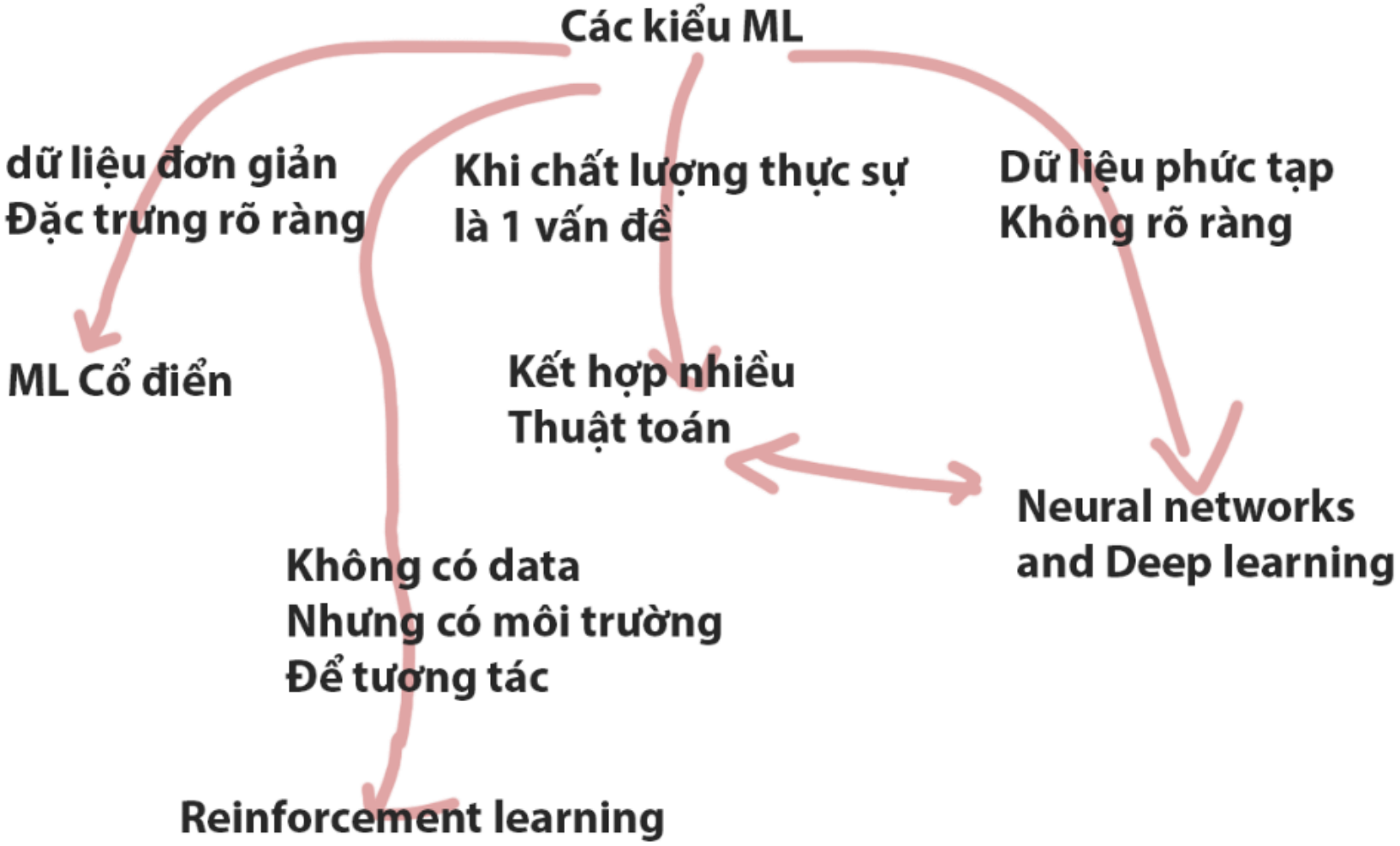
- Vấn đề học tập có thể được phát biểu như sau
- Đưa ra một ví dụ mẫu có kích thước giới hạn, để tìm một mô tả dữ liệu ngắn gọn
- Các kỹ thuật học tập có thể được nhóm lại trong bốn nhóm lớn
- Supervised learning
- Reinforcement learning
- Unsupervised learning
- Semi-supervised learning(bán giám sát)



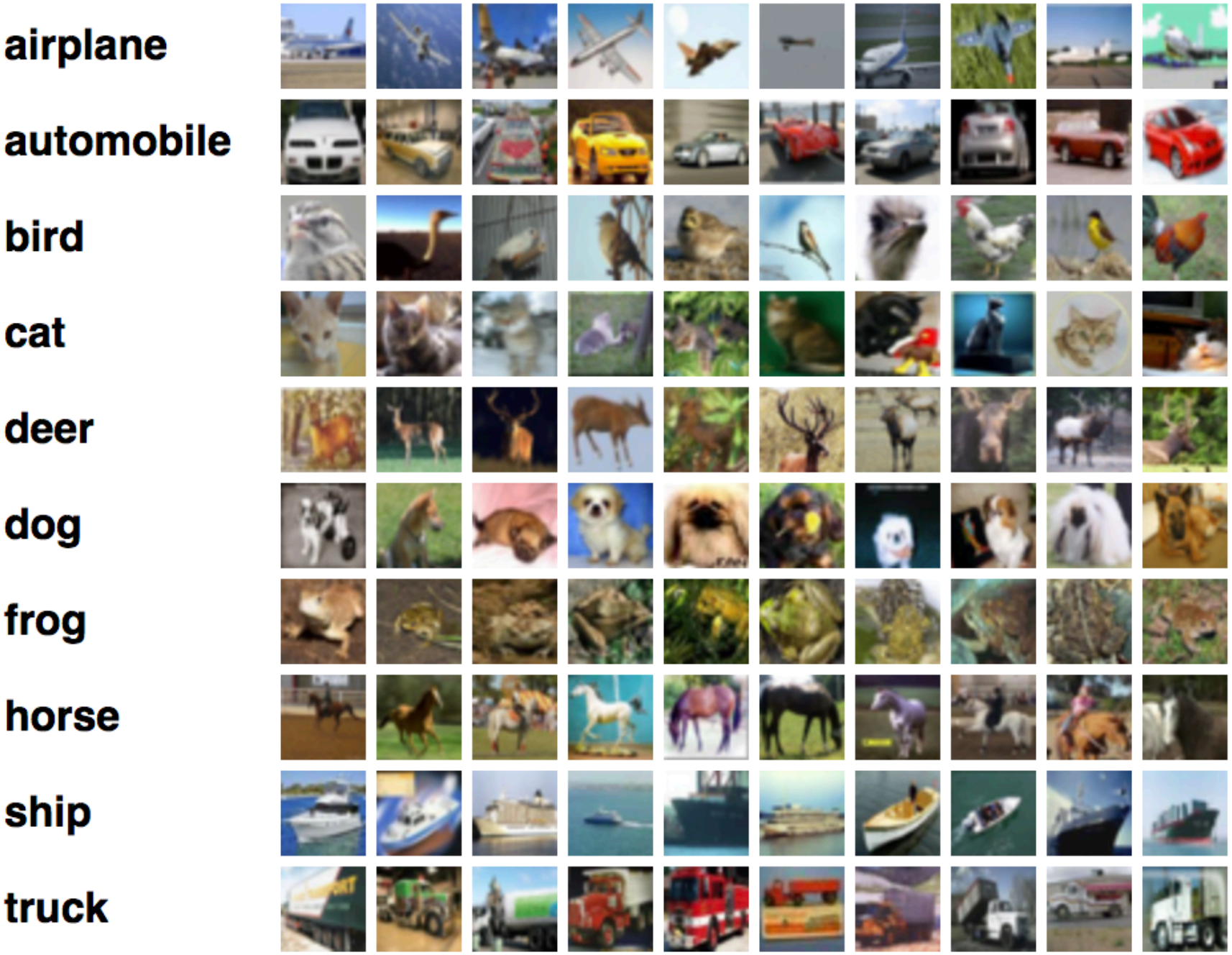
ML type


Các kiểu machine learning
Học có giám sát (supervised learning)
Học không giám sát (Unsupervised learning)
Học bán giám sát (semi-supervised learning)
Học tăng cường (Reinforcement learning)
Biến Liên tục (continuous target variable)
Biến rời rạc (Categorical target variable)
Target variable not available
Biến rời rạc (Categorical target variable)
Biến rời rạc (Categorical target variable)
Target variable not available
Hồi quy(regression)
Phân lớp (Classification)
Phân cụm(clustering)
Phân cụm(clustering)
Classification
Phân cụm(clustering)
Phân lớp (Classification)
Control
Dự đoán giá nhà
Phân loại hình ảnh y tế
Phân cụm phân nhóm khách hàng
Phân cụm phân nhóm khách hàng
Phân loại văn bản
tìm đường trên dữ liệu GPS
Tối ưu marketing
Ô tô không người lái

Type of ML


Các kiểu ML
dữ liệu đơn giản
Đặc trưng rõ ràng
Khi chất lượng thực sự
là 1 vấn đề
Dữ liệu phức tạp
Không rõ ràng
ML Cổ điển
Kết hợp nhiều
Thuật toán
Neural networks
and Deep learning
Không có data
Nhưng có môi trường
Để tương tác
Reinforcement learning

Dataset


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

MODEL IN ML

$$\hat{y} = \hat{f}(x) \text{ [our approximation]}$$

$$\begin{array}{cc} x & y \\ \begin{bmatrix} 0.5 \\ 0.2 \\ 0.6 \\ \dots \\ 0.3 \end{bmatrix} & \begin{bmatrix} 14.8 \\ 13.3 \\ 11.6 \\ \dots \\ 6.16 \end{bmatrix} \end{array}$$

$$\hat{y} = mx + c$$

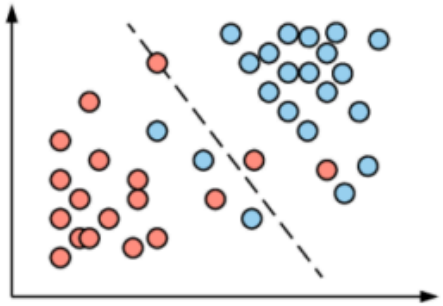
$$\hat{y} = ax^2 + bx + c$$

$$\hat{y} = ax^3 + bx^2 + cx + d$$

$$\hat{y} = ax^4 + bx^3 + cx + d$$

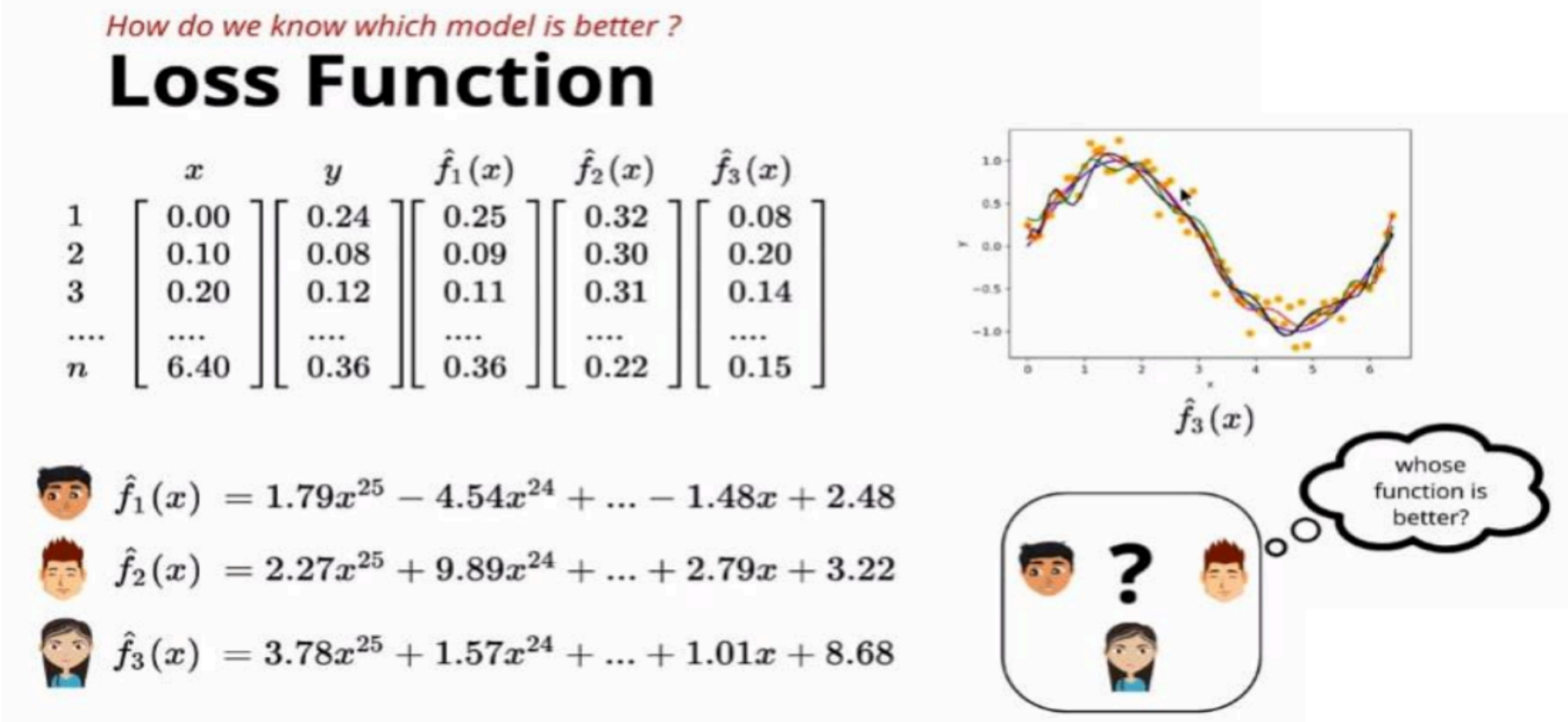
$$\vdots$$

$$\hat{y} = ax^{25} + bx^{24} + \cdots + cx + d$$



- Mô hình không là gì ngoài một hàm toán học xác định mối quan hệ giữa dữ liệu đầu vào và dữ liệu đầu ra.
- Một mô hình có thể là một Hàm tuyến tính đơn giản như y = mx + c hoặc nó có thể là một hàm phức tạp như ***y = sin(cos(ax)) + tan(sin(bx^2 + cx))***
- Cho một hình ảnh (dữ liệu), dự đoán xem nó có một con chó hay không (task), chúng ta phải xây dựng một hệ thống trả lời câu hỏi này và hệ thống đó được gọi là mô hình
- Mình gọi cái đường thẳng nét đứt kia là mô hình

LOSS FUNCTION



- Để biết mô hình nào phù hợp nhất với dữ liệu của chúng tôi, chúng tôi tính toán sự khác biệt giữa giá trị thực và giá trị dự đoán cho tất cả dữ liệu bằng cách sử dụng hàm mất mát.

How do we know which model is better ?

Loss Function

	x	y	$\hat{f}_1(x)$	$\hat{f}_2(x)$	$\hat{f}_3(x)$
1	0.00	0.24	0.25	0.32	0.08
2	0.10	0.08	0.09	0.30	0.20
3	0.20	0.12	0.11	0.31	0.14
....	....	....	....	....	....
n	6.40	0.36	0.36	0.22	0.15

$$\mathscr{L}_1 = \sum_{i=1}^{n}(y_i - \hat{f}_1(x_i))^2$$
$$= (0.24 - 0.25)^2 + (0.08 - 0.09)^2 + (0.12 - 0.11)^2 + \ldots + (0.36 - 0.36)^2$$
$$= 1.38$$

$$\mathscr{L}_1 = \sum_{i=1}^{n}(y_i - \hat{f}_1(x_i))^2 = 1.38$$

$$\mathscr{L}_2 = \sum_{i=1}^{n}(y_i - \hat{f}_2(x_i))^2 = 2.02$$

$$\mathscr{L}_3 = \sum_{i=1}^{n}(y_i - \hat{f}_3(x_i))^2 = 2.34$$

- Sử dụng hàm mất mát, chúng ta có thể có được mô hình tốt nhất xác định mối quan hệ giữa dữ liệu đầu vào và dữ liệu đầu ra của chúng tôi.

LEARNING ALGORITHM

$$\hat{f}(x) = ax_1^2 + bx_2^3 + cx_3^2$$

Budget (100crore)	Box Office Collection(100 crore)	Action Scene times (100 mins)	IMDB Rating
0.55	0.66	0.22	4.8
0.68	0.91	0.77	7,2
0.66	0.88	0.67	6.7
⋮	⋮	⋮	⋮
0.72	0.94	0.97	8.1
0.58	0.74	0.35	5.3

$$min_{a,b,c}\ \mathscr{L}_1 = \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

Many optimization solvers are available

EVALUATION

How do we compute a score for our ML model?

Evaluation

Top - 1

	True Labels	Predicted Labels	
$[\ 2.1, 1.2, \ldots, 5.6, 7.8\ ]$	1	1	✓
$[\ 3.5, 6.6, \ldots, 2.5, 6.3\ ]$	2	3	✗
$[\ 6.3, 2.6, \ldots, 4.5, 3.8\ ]$	3	1	✗
$[\ 2.8, 3.6, \ldots, 7.5, 2.1\ ]$	4	4	✓
$[\ 2.2, 1.7, \ldots, 2.5, 1.8\ ]$	5	5	✓
$[\ 6.3, 2.6, \ldots, 4.5, 3.8\ ]$	3	2	✗
$[\ 1.9, 3.3, \ldots, 4.2, 1.1\ ]$	5	5	✓

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

$$= \frac{4}{7} = 0.55$$

Class	
Lion	1
Tiger	2
Cat	3
Giraffe	4
Dog	5

SUPERVISED LEARNING

	scans	x						t
x1		2.3	5.9	...	11.0	-0.3	8.9	0
x2		-8.5	-1.7	...	-1.3	9.0	7.2	1
		-0.4	6.7	...	-2.4	4.7	-7.3	0
		1.6	-0.4	...	-4.6	6.4	1.9	1
		3.9	-4.1	...	6.7	-3.1	2.1	1
		5.1	3.7	...	1.8	-4.2	9.3	1

Supervised learning

Học có giám sát là một trong những nhóm vấn đề chung về học máy

Cho một vectơ dữ liệu huấn luyện:

x=(x1,...,xN)

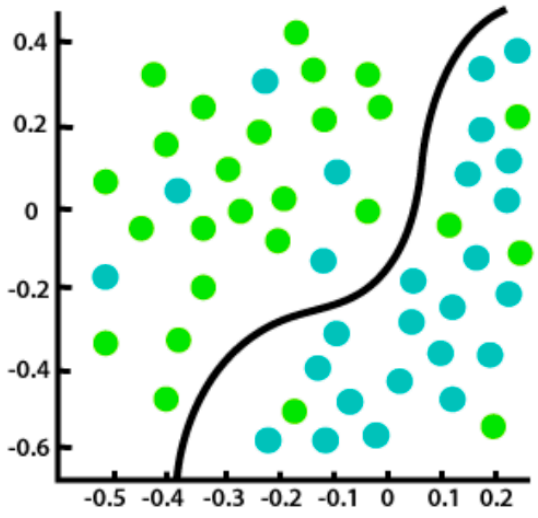
và một tập hợp các giá trị đầu ra mục tiêu tương ứng:

t = (t1 , . . . , tN)

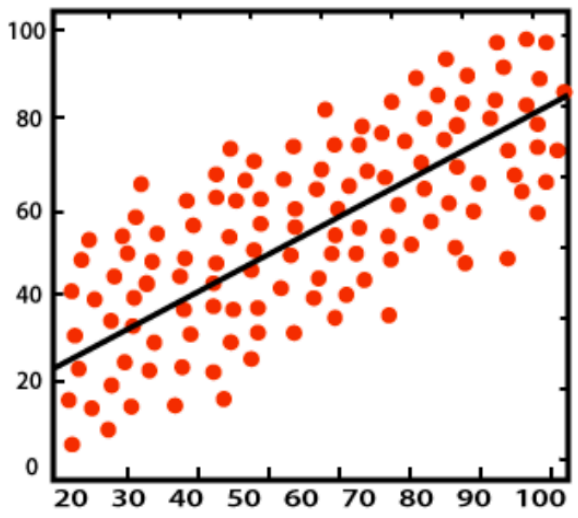
- learn a function y(x) = t.
- Với một đầu vào thử nghiệm mới x, hàm y (·) dự đoán giá trị đầu ra t

- Các đầu vào x (ví dụ: x1, x2) thường là vectơ chiều D, các phần tử của vector của nó được gọi là các đặc trưng hoặc thuộc tính
- Các đặc trưng này mô tả các thuộc tính của đầu vào được cho là tương quan(có liên quan) với đầu ra, ví dụ: các pixel trong hình ảnh khuôn mặt

- Phân loại (Classification) là dự đoán một nhãn, lớp rời rạc(ví dụ ảnh này là có hay mèo)
- Và hồi quy (Regression) là dự đoán một số lượng (ví dụ: giá nhà, giá cổ phiếu)

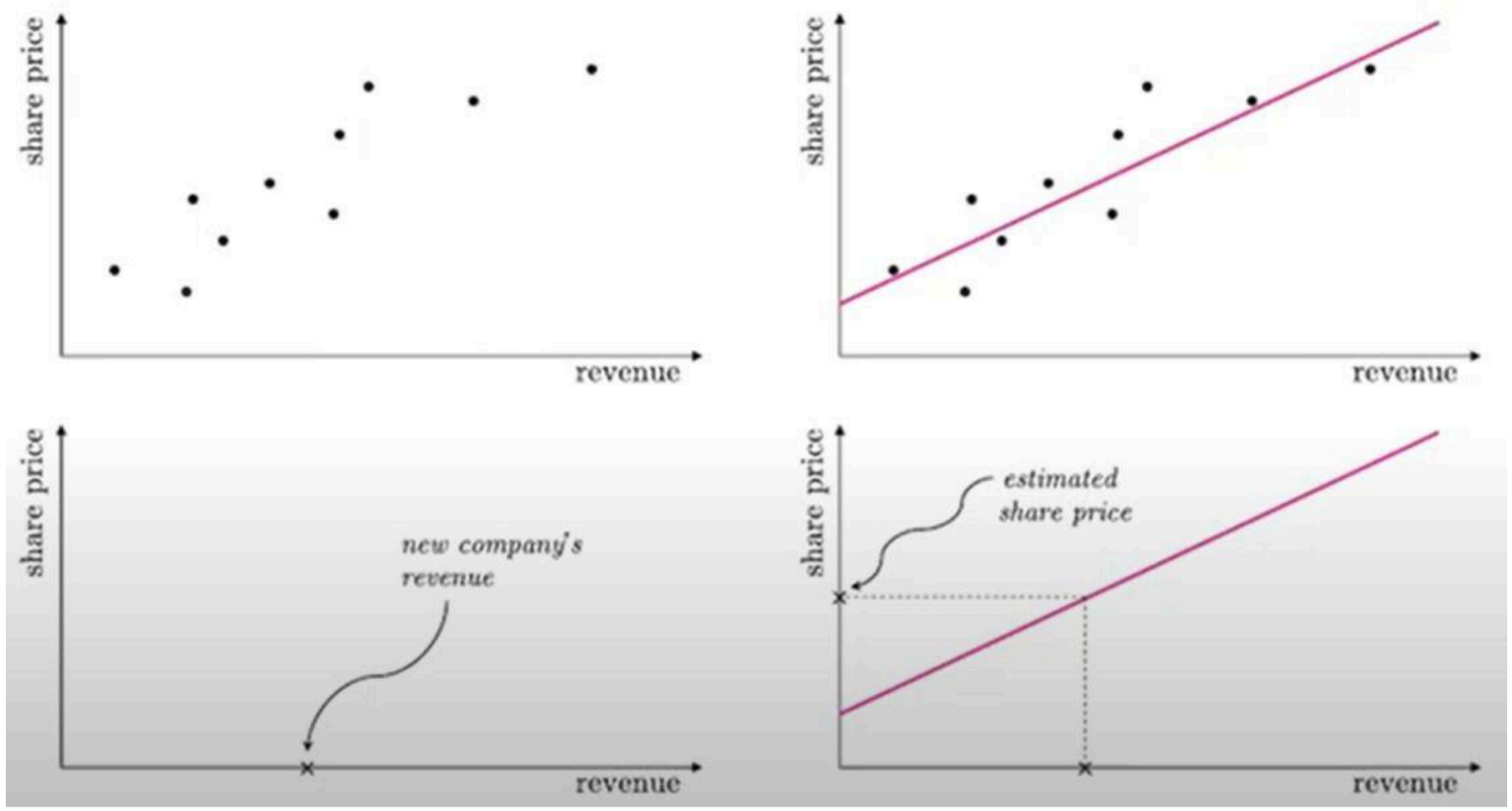


Classification



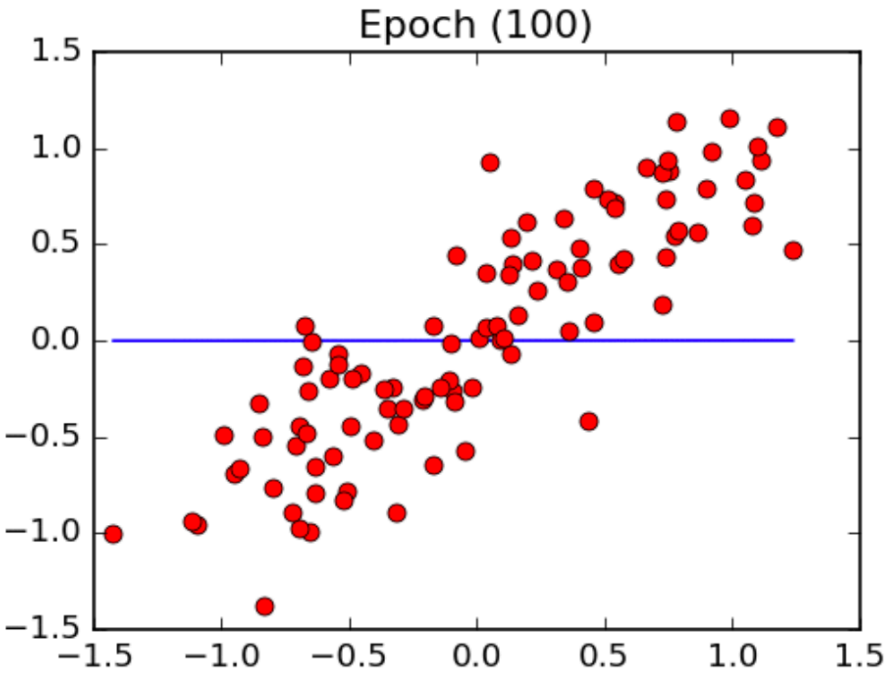
Regression

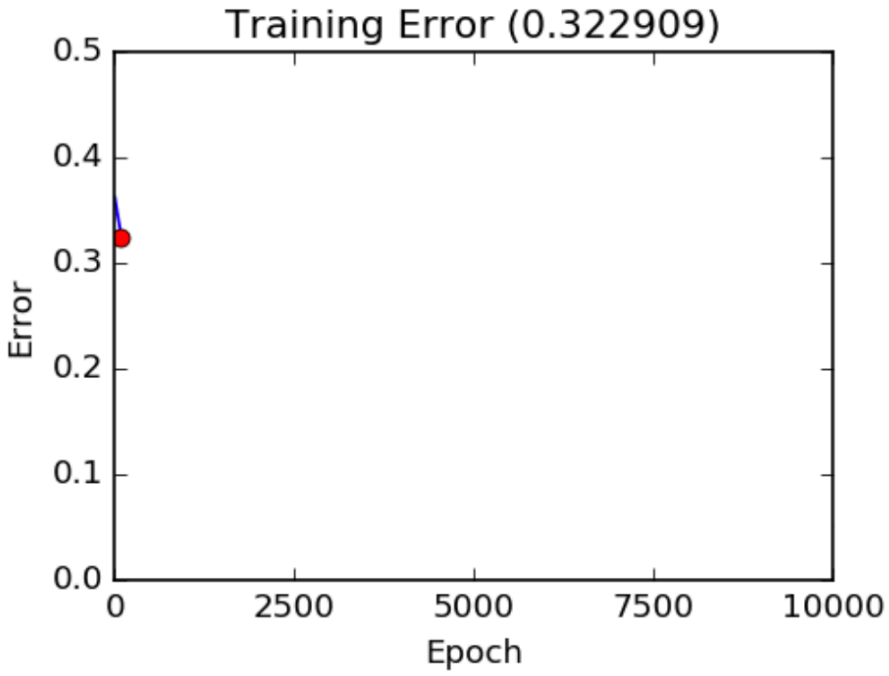
REGRESSION EXAMPLE



Output là giá trị liên tục

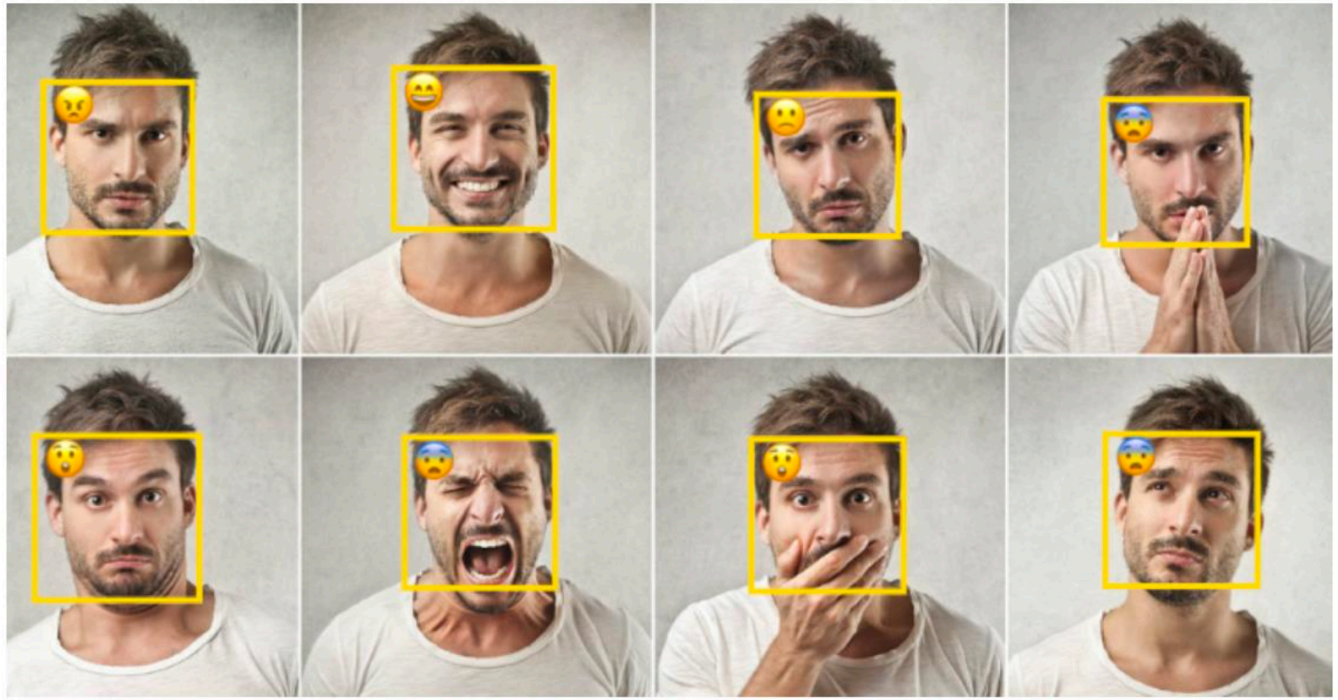
REGRESSION


Epoch (100)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

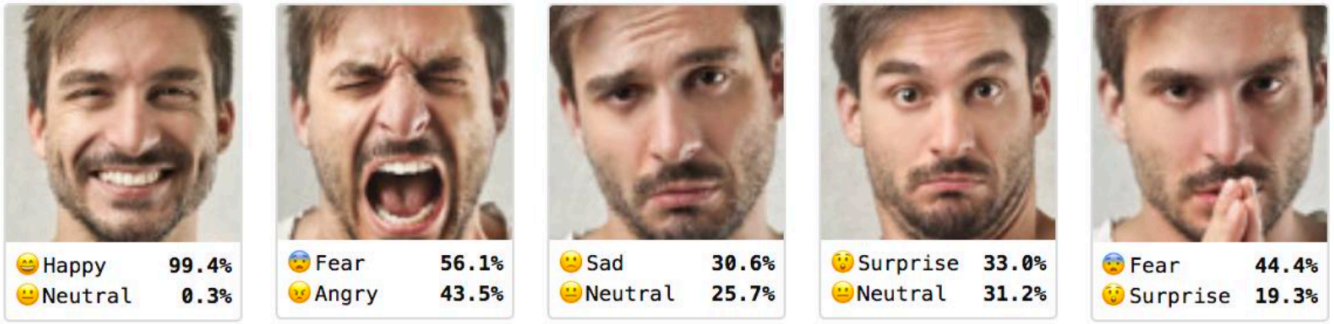

Training Error (0.322909)
0.5
0.4
0.3
0.2
0.1
0.0
Error
0
2500
5000
7500
10000
Epoch

REGRESSION EXAMPLE

CLASSIFICATION: EXAMPLES

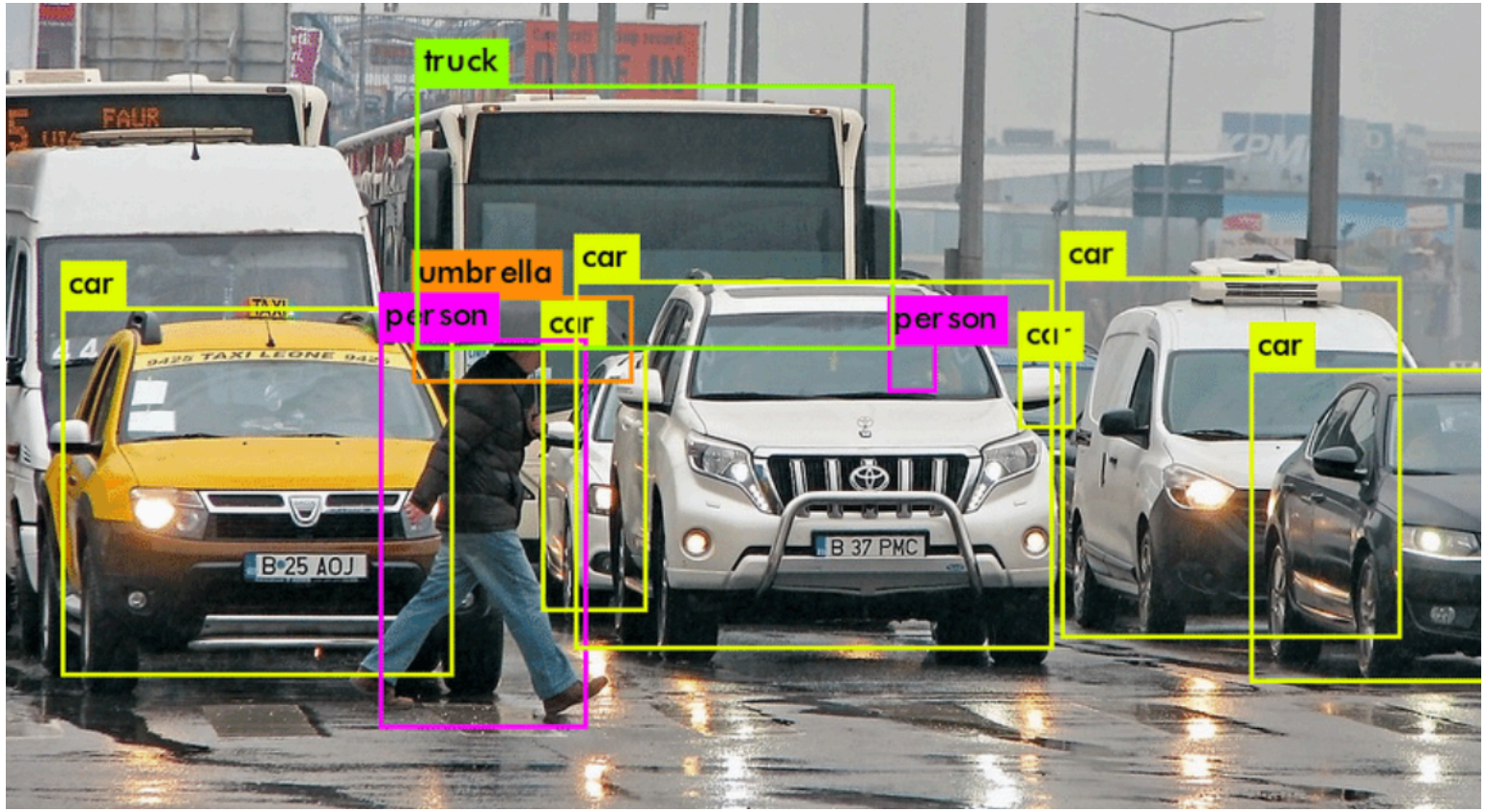


Results: Of **8** people detected, **1 (13%)** is happy.

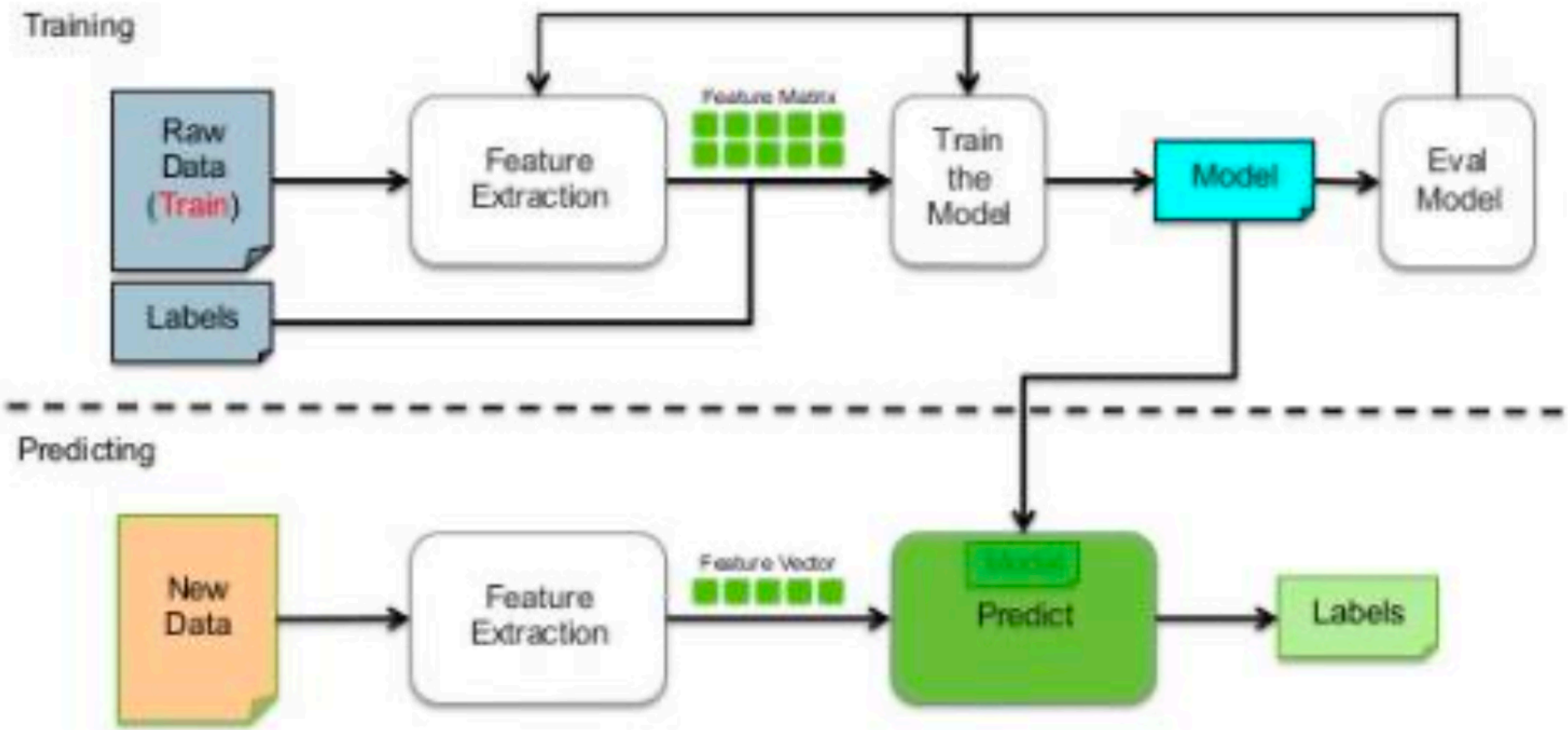


CLASSIFICATION: EXAMPLES

- Kết quả đầu ra là các lớp rời rạc

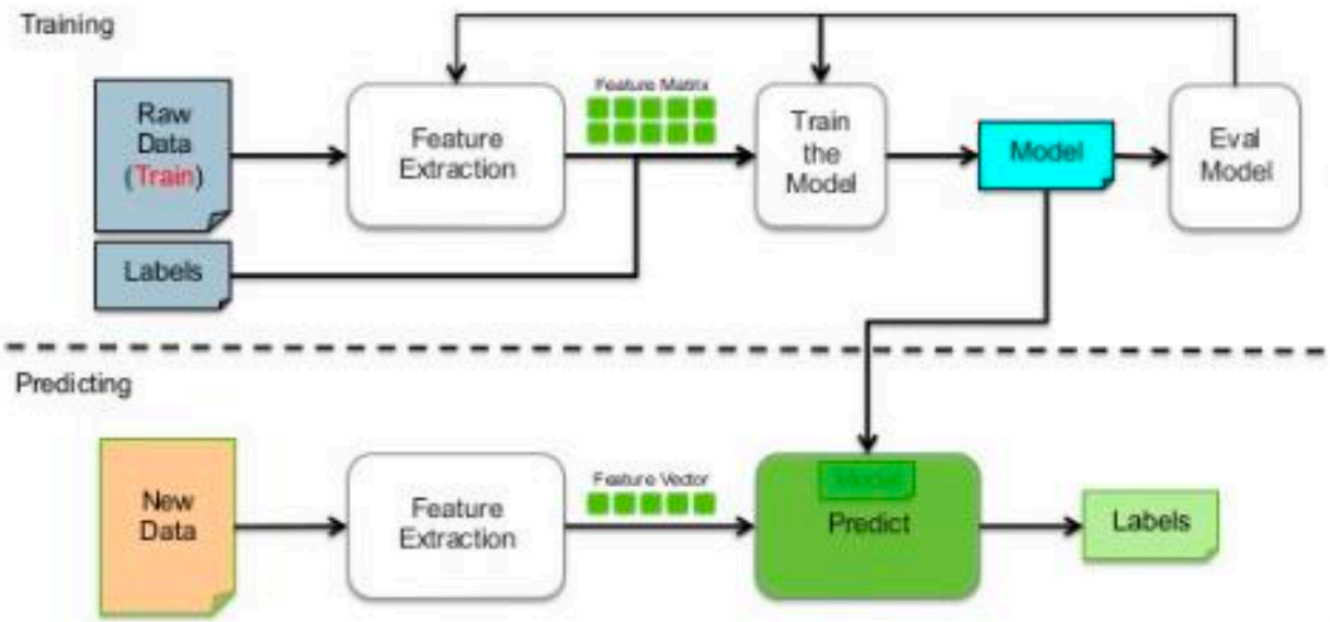


SUPERVISED LEARNING PIPELINE



SUPERVISED LEARNING PIPELINE

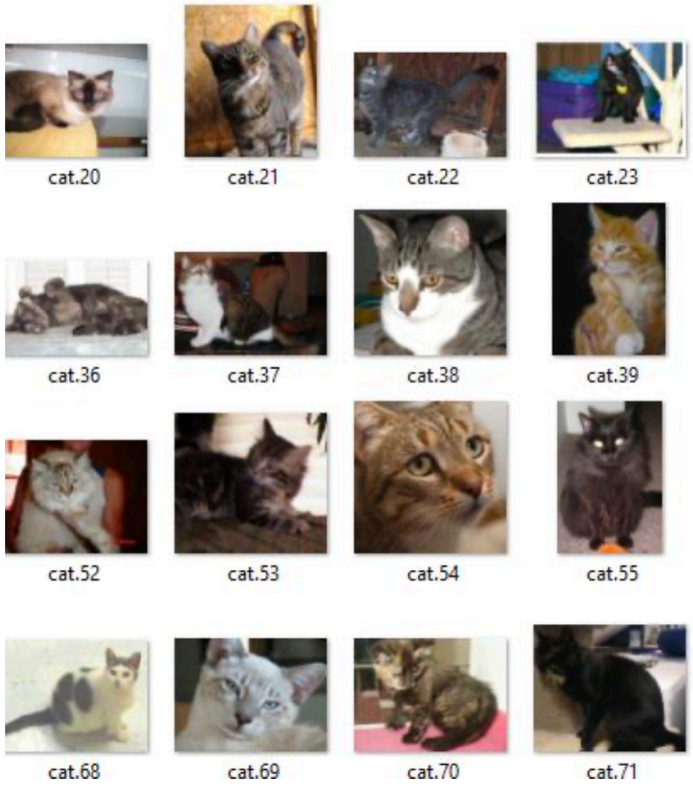
- Define task | Định nghĩa việc cần làm
- Collect data | Thu thập dữ liệu
- Design features | thiết kế đặc trưng
- Choose algorithm | Chọn thuật toán
- Train model | Đào tạo mô hình
- Test model | Kiểm tra và đánh giá mô hình



Định nghĩa công việc cần làm

Phân biệt, phân loại ảnh chó mèo

Bước 1: Thu thập dữ liệu



UC Irvine Machine Learning Repository
http://archive.ics.uci.edu/ml/

Bước 2: Thiết kế đặc trưng huấn luyện

FEATURE ENGINEERING CHALLENGE

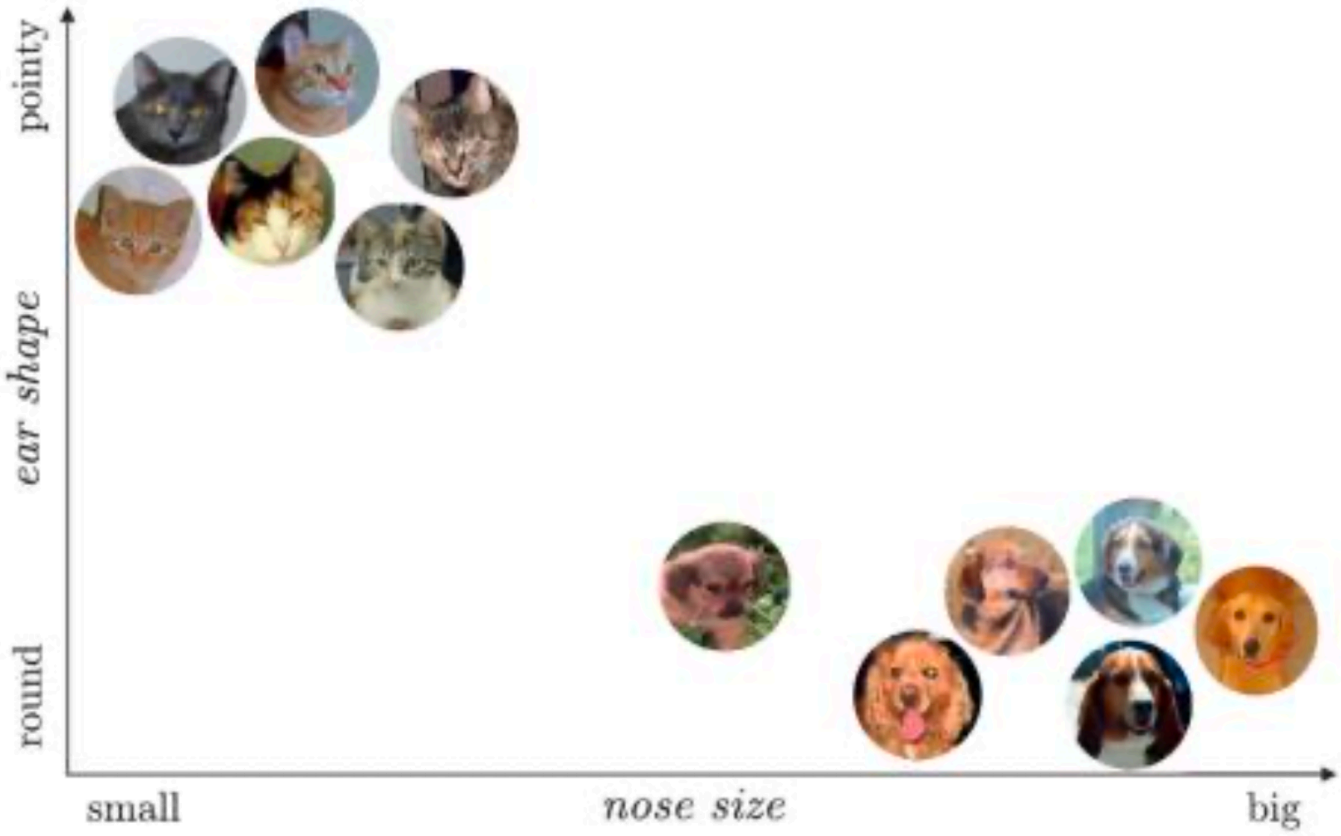
Cat
- Four legs
- Two eyes
- Two ears
- Many Whiskers
- Multiple Colors

Dog
- Four legs
- Two eyes
- Two ears
- Many Whiskers
- Multiple Colors

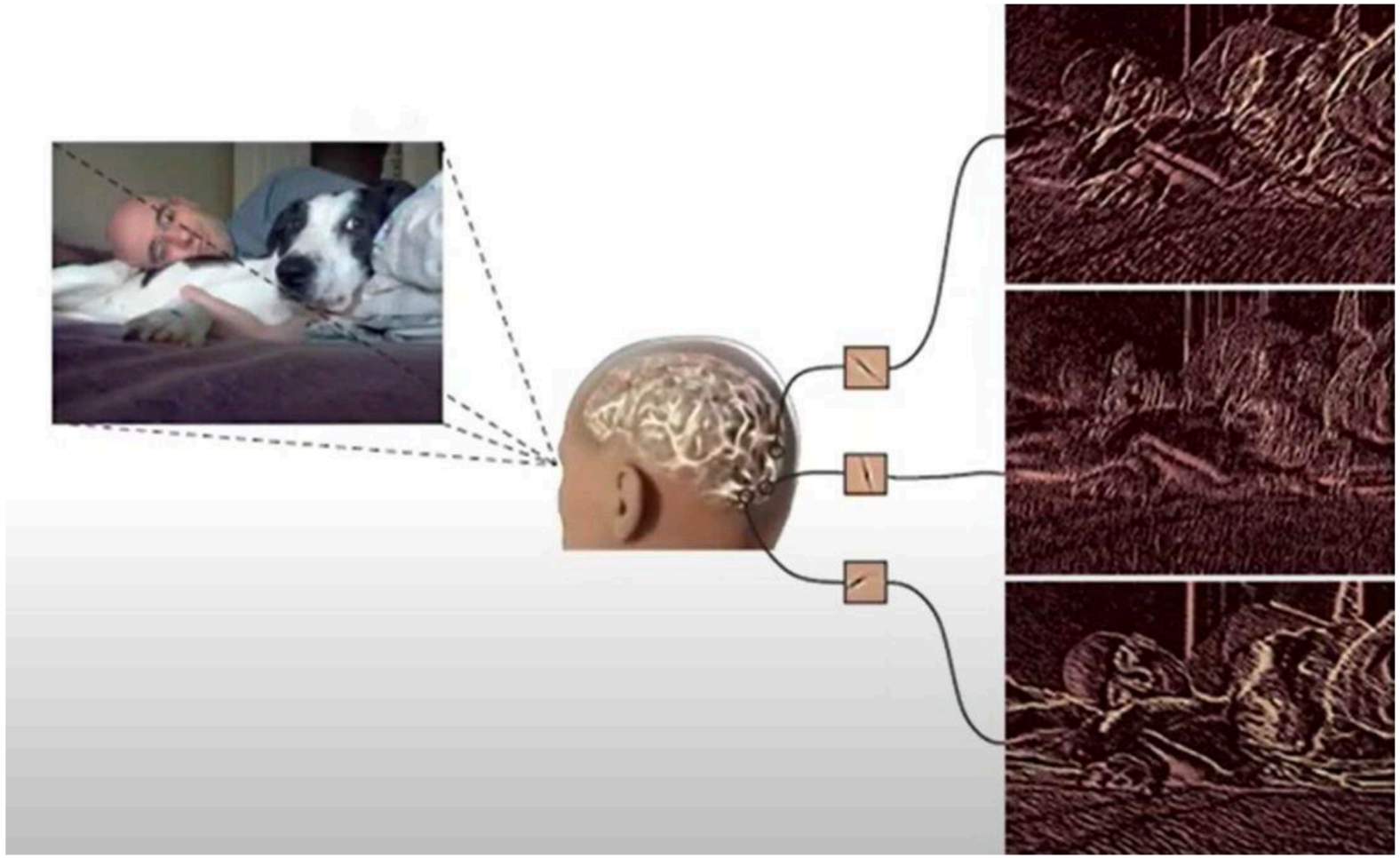
Chó và mèo có khá nhiều đặc điểm tương đồng

Bước 2: Thiết kế đặc trưng huấn luyện

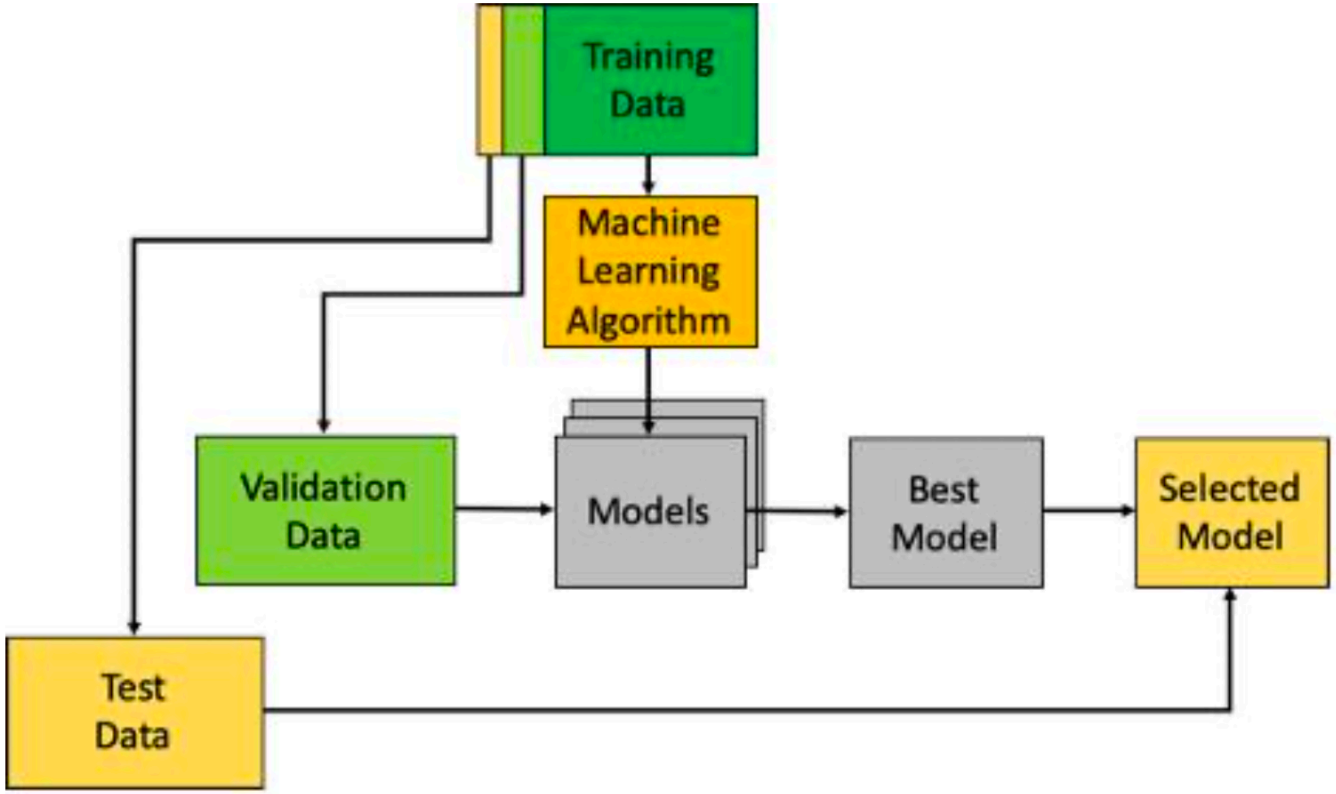
Ví dụ:
Kích thước mũi.
Hình dáng tai



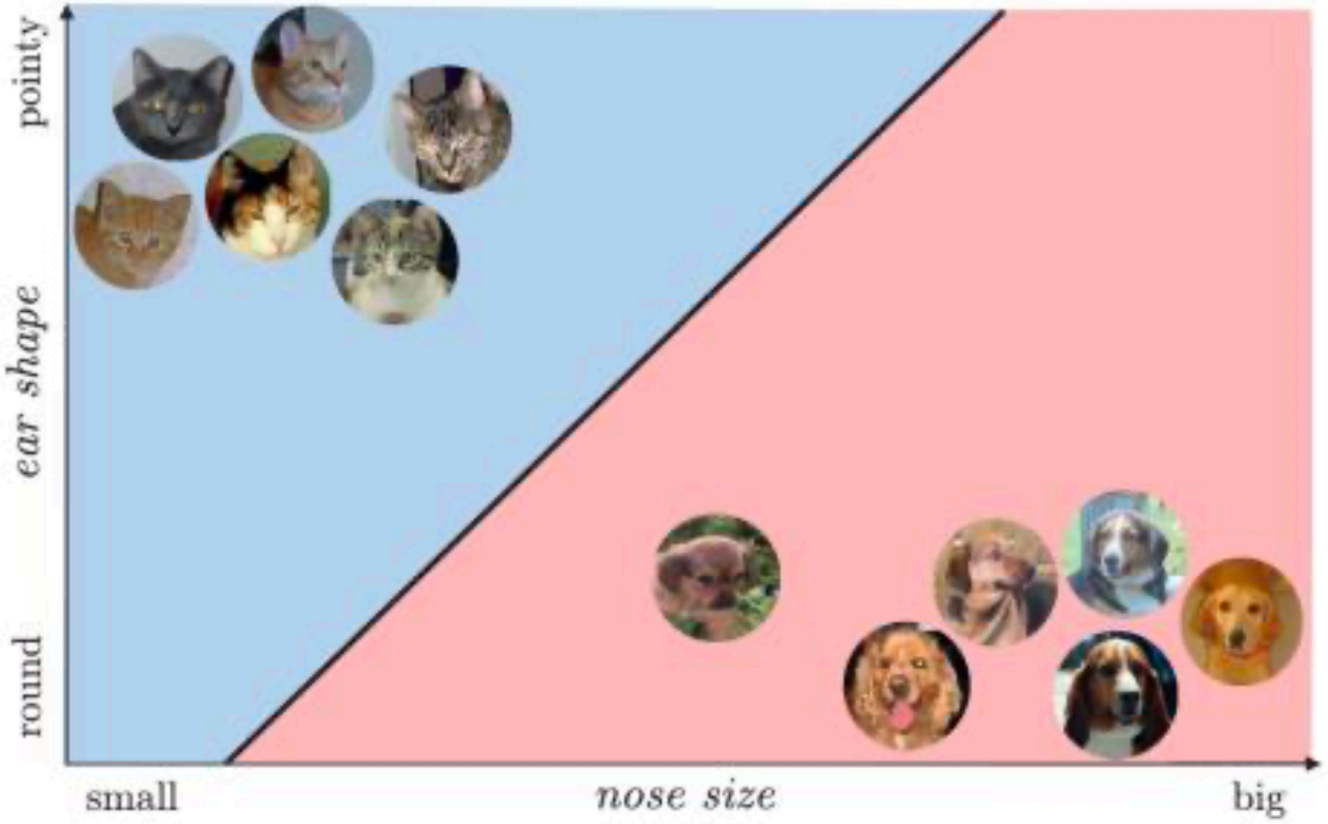
Bước 2: Thiết kế đặc trưng huấn luyện



Bước 3: Training the model

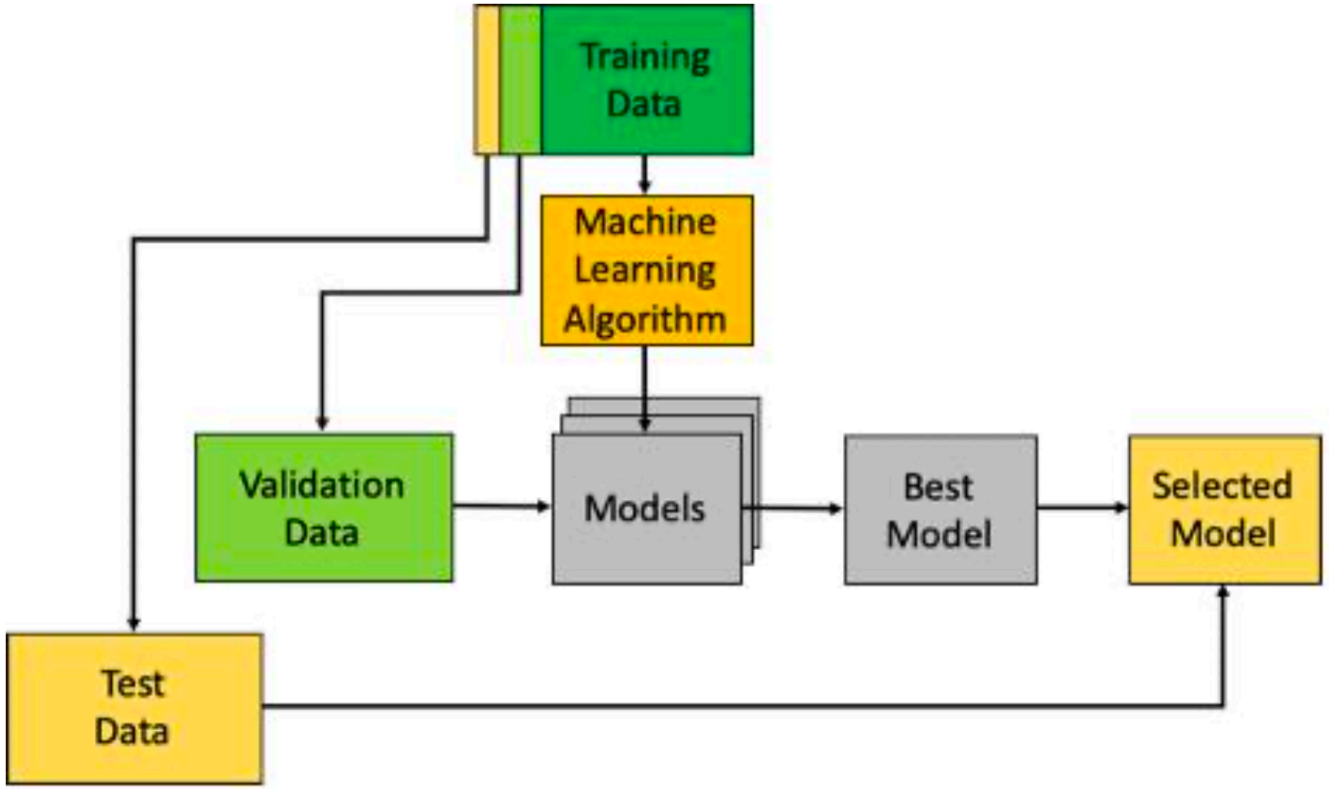


Bước 3: Training the model

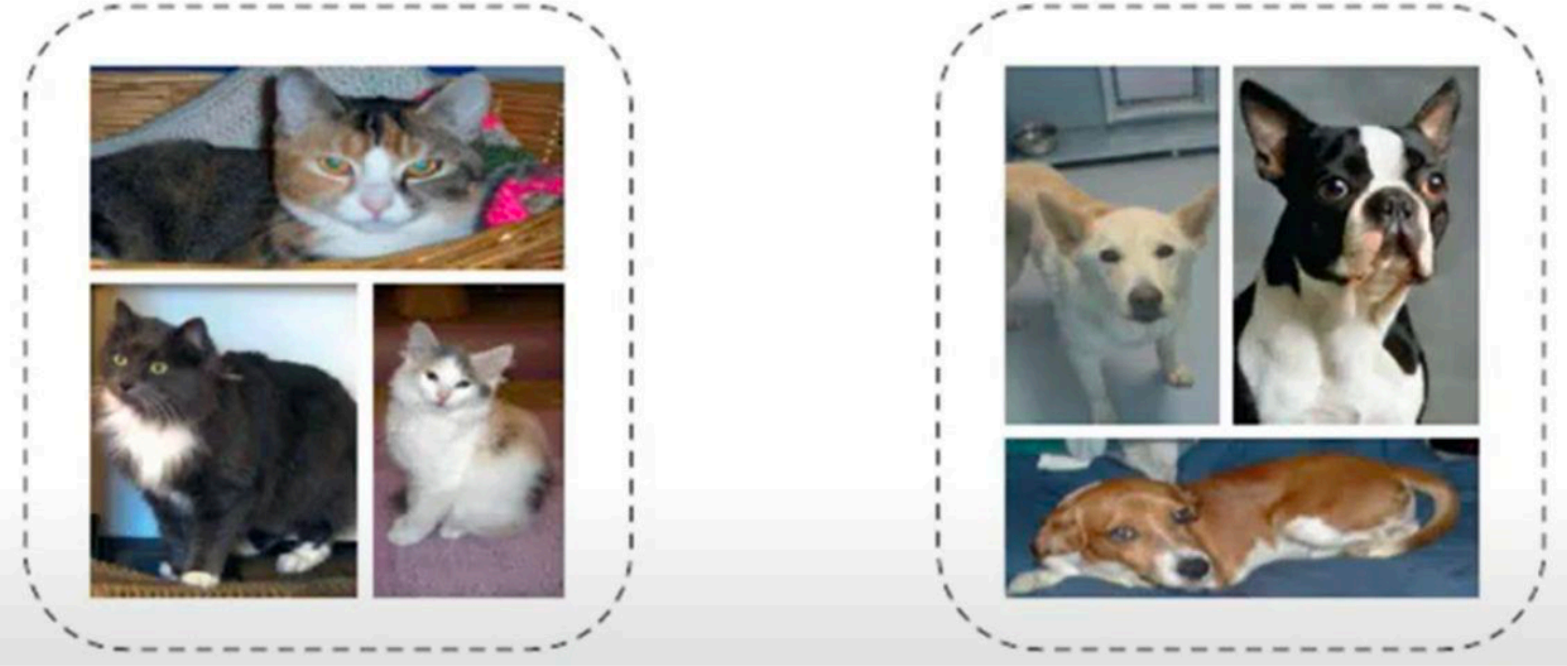


Ayran Olckers, How to Teach a Computer to Distinguish Cats from Dogs

Bước 4: Testing the model

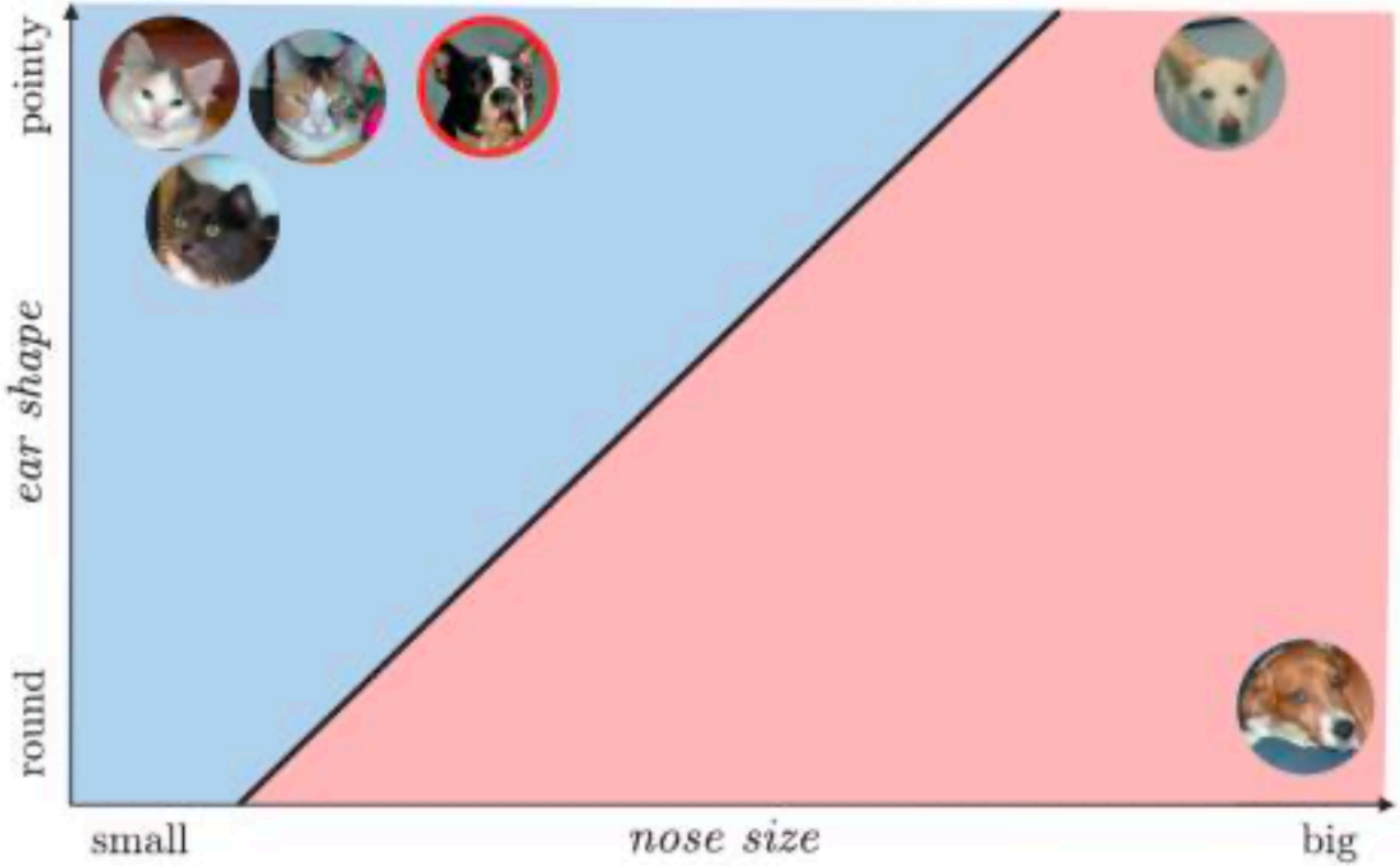


Bước 4: Testing the model



Lưu ý: Những hình ảnh này không phải từ tập dữ liệu đào tạo (tập training)

Bước 4: Testing the model



Ayran Olckers, How to Teach a Computer to Distinguish Cats from Dogs

Open source software for ml

- dlib
- ELKI
- Encog
- H2O
- Mahout
- mlpy
- MLPACK
- MOA (Massive Online
- Analysis)
- ND4J with Deeplearning4j
- OpenCV
- OpenNN
- Orange
- **R**
- scikit-learn
- Shogun
- Torch (machine learning) • Spark
- Yooreeka
- Weka
- Tensorflow

Tài liệu gợi ý học tập

- Lantz, Brett. (2015) Machine learning with R. Packt Publishing Ltd
- Bishop, Christopher M. (2006) Pattern recognition and machine learning, Springer-Verlag
- Alpaydin, E. (2014). Introduction to machine learning. MIT press
- Michael Paluszek, Stephanie Thomas (2016) MATLAB machine learning, Apress
- Deeplearning book Dive into Deep Learning (d2l)