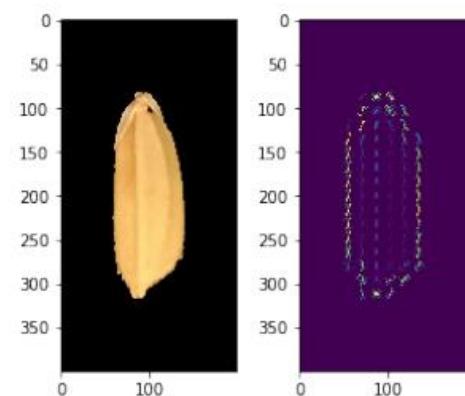
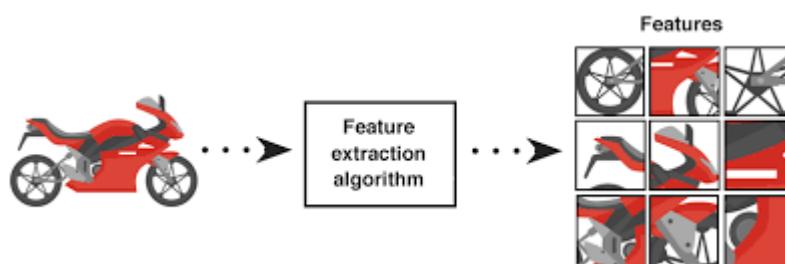




ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG
KHOA KHOA HỌC MÁY TÍNH

HỌC PHẦN MÁY HỌC ỨNG DỤNG
MSHP: CT294

TRÍCH ĐẶC TRƯNG



Nội dung

Chương 10. Đặc trưng của dữ liệu	
10.1.	Đặc trưng dữ liệu dạng có cấu trúc
10.2.	Đặc trưng dữ liệu dạng văn bản
10.3.	Đặc trưng ảnh
10.4.	Xử lý đặc trưng
10.5.	Kết luận và hướng phát triển

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Đặc trưng dữ liệu
dạng có cấu trúc



Đặc trưng dữ liệu
dạng văn bản

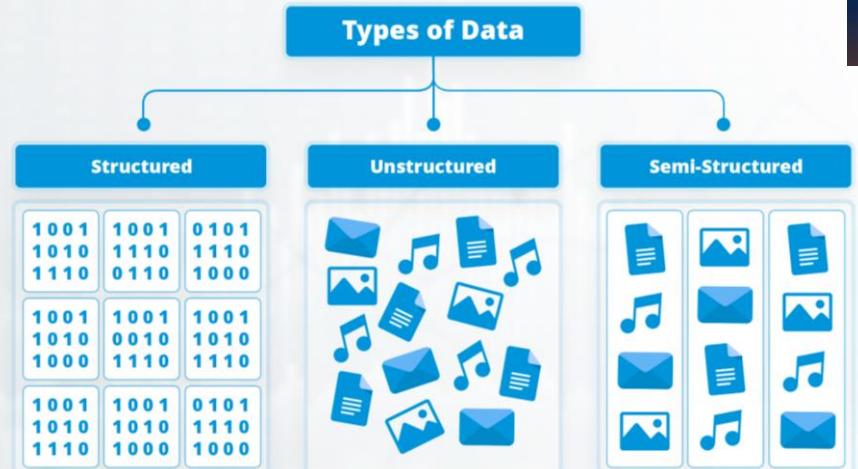
Cầu Cần Thơ nối liền giữa tỉnh Vĩnh Long và thành phố Cần Thơ, bắt qua thêm cồn Ấu (địa phận Cần Thơ) để vào thành phố

Text – Văn bản
(Caption - NLP)

Đặc trưng ảnh



Image – Hình ảnh

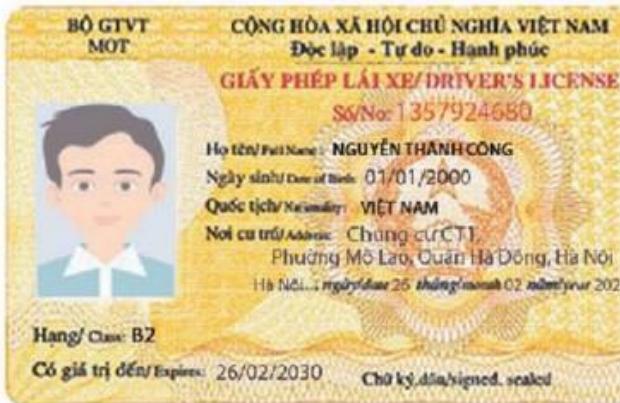


Bài tập

Chương 10.	Đặc trưng của dữ liệu
10.1.	Đặc trưng dữ liệu dạng có cấu trúc
10.2.	Đặc trưng dữ liệu dạng văn bản
10.3.	Đặc trưng ảnh
10.4.	Xử lý đặc trưng
10.5.	Kết luận và hướng phát triển

Thông tin	Kiểu dữ liệu
Ảnh	Số
Số	Văn bản
Họ tên	Hình ảnh
Ngày sinh	Âm thanh
Quốc tịch	
Nơi cư trú	

Con vật được nuôi ở nhà,
Thường bắt chuột, sợ nước



Meo meo meo

Trích xuất đặc trưng là gì?

**Trích xuất =
Extraction**

Chữ/Câu/Đoạn
Ví dụ: "Khoa Khoa học máy tính"

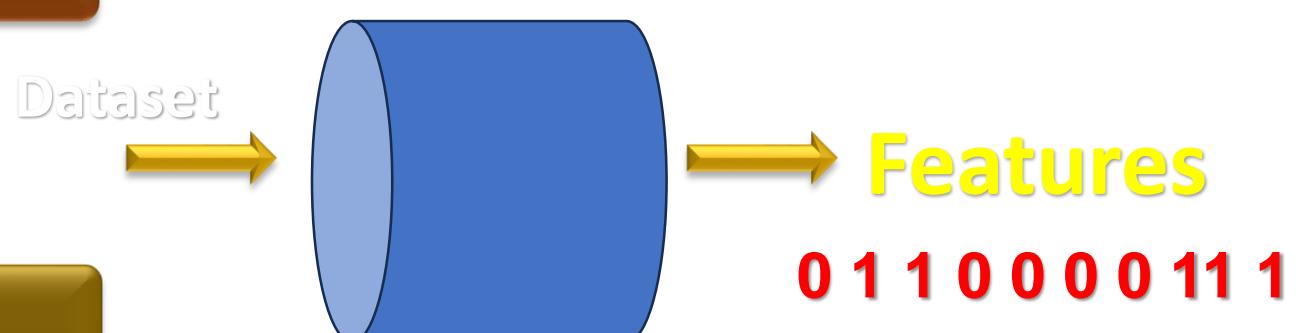
Âm thanh
Ví dụ: Nhạc, Tiếng nói,...



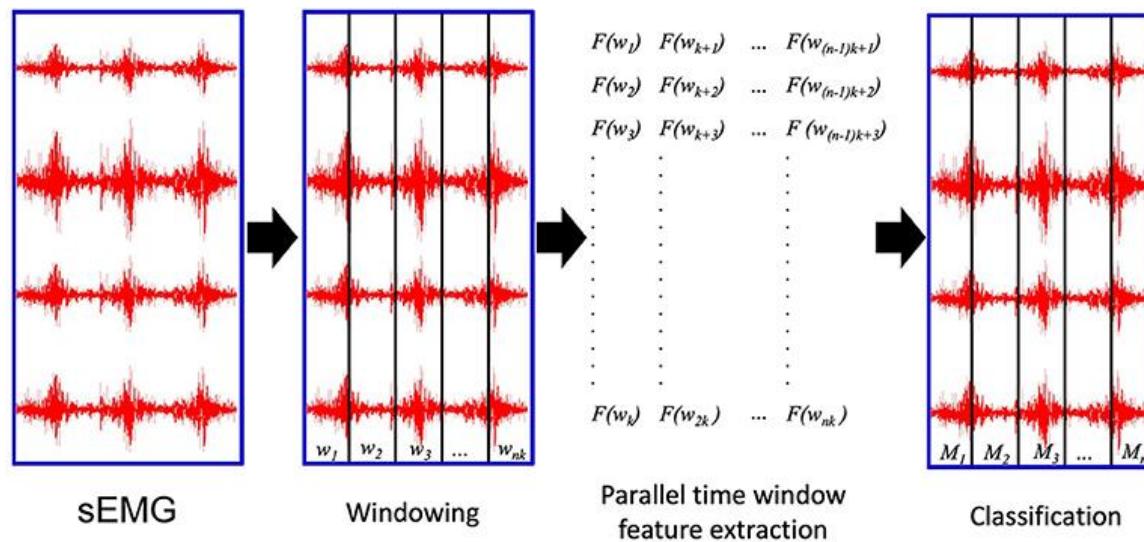
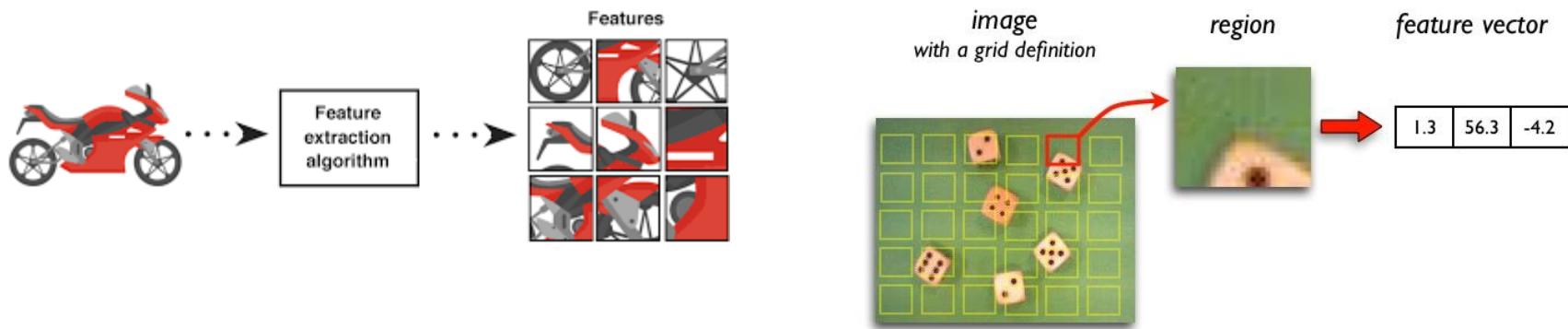
Ảnh
Hình ảnh, Videos, Bản đồ



Đặc trưng = Features



Trích xuất đặc trưng là gì?



BƯỚC TRÍCH ĐẶC TRƯNG TẠI VỊ TRÍ NÀO?

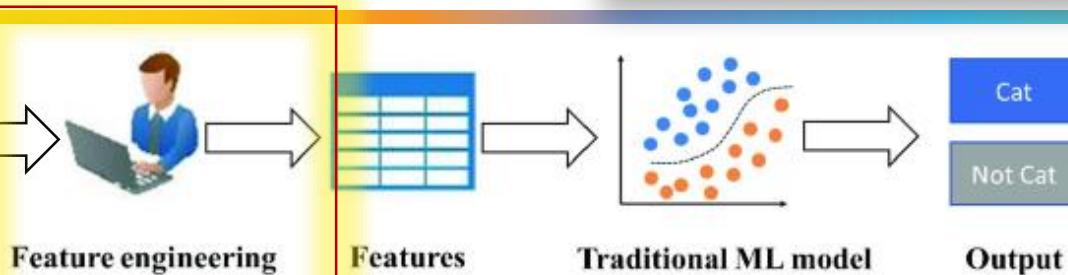
Mô hình máy học:
SVM, KNN, Naïve Bayes, Decision Tree

1

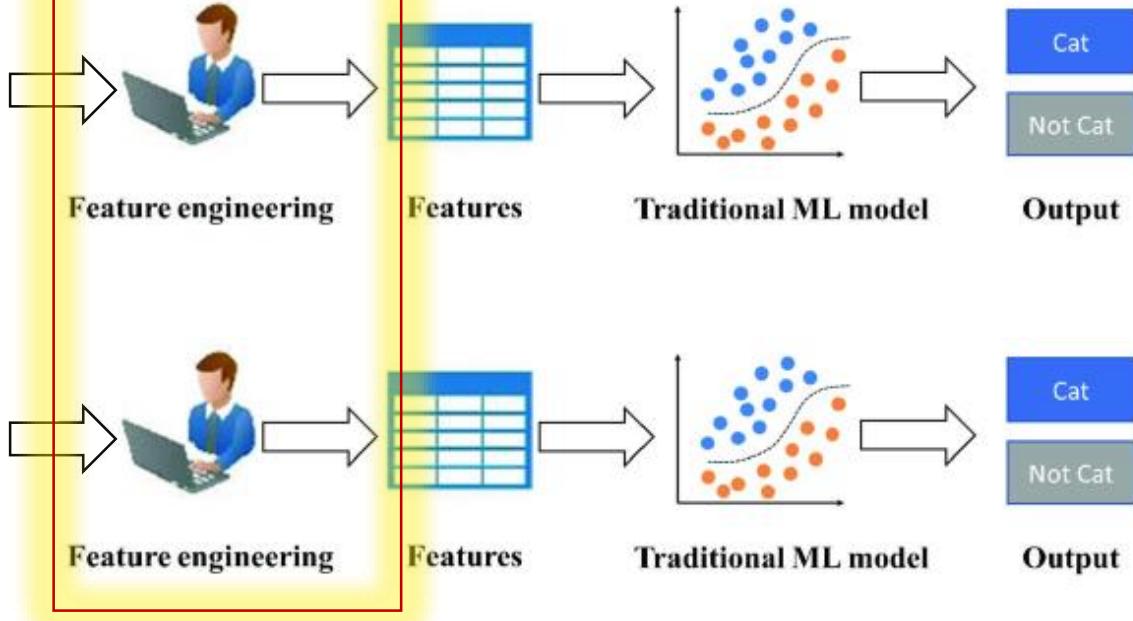
Con vật được nuôi ở nhà,
Thường bắt chuột, sợ nước



2



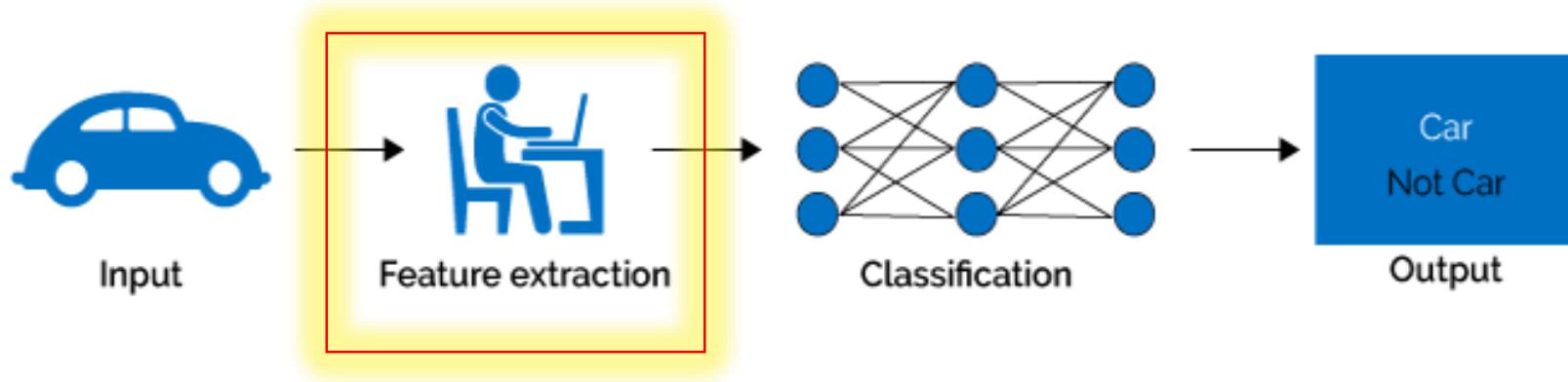
3



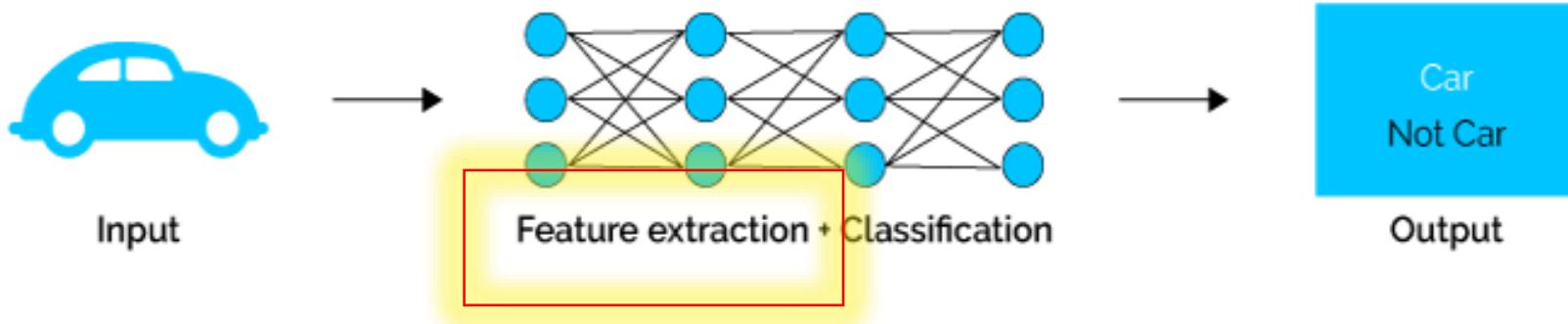
Quá trình này được gọi là **Feature Extraction**, hoặc **Feature Engineering**,
một số tài liệu tiếng Việt gọi nó là **trích chọn đặc trưng**.

BƯỚC TRÍCH ĐẶC TRƯNG TẠI VỊ TRÍ NÀO?

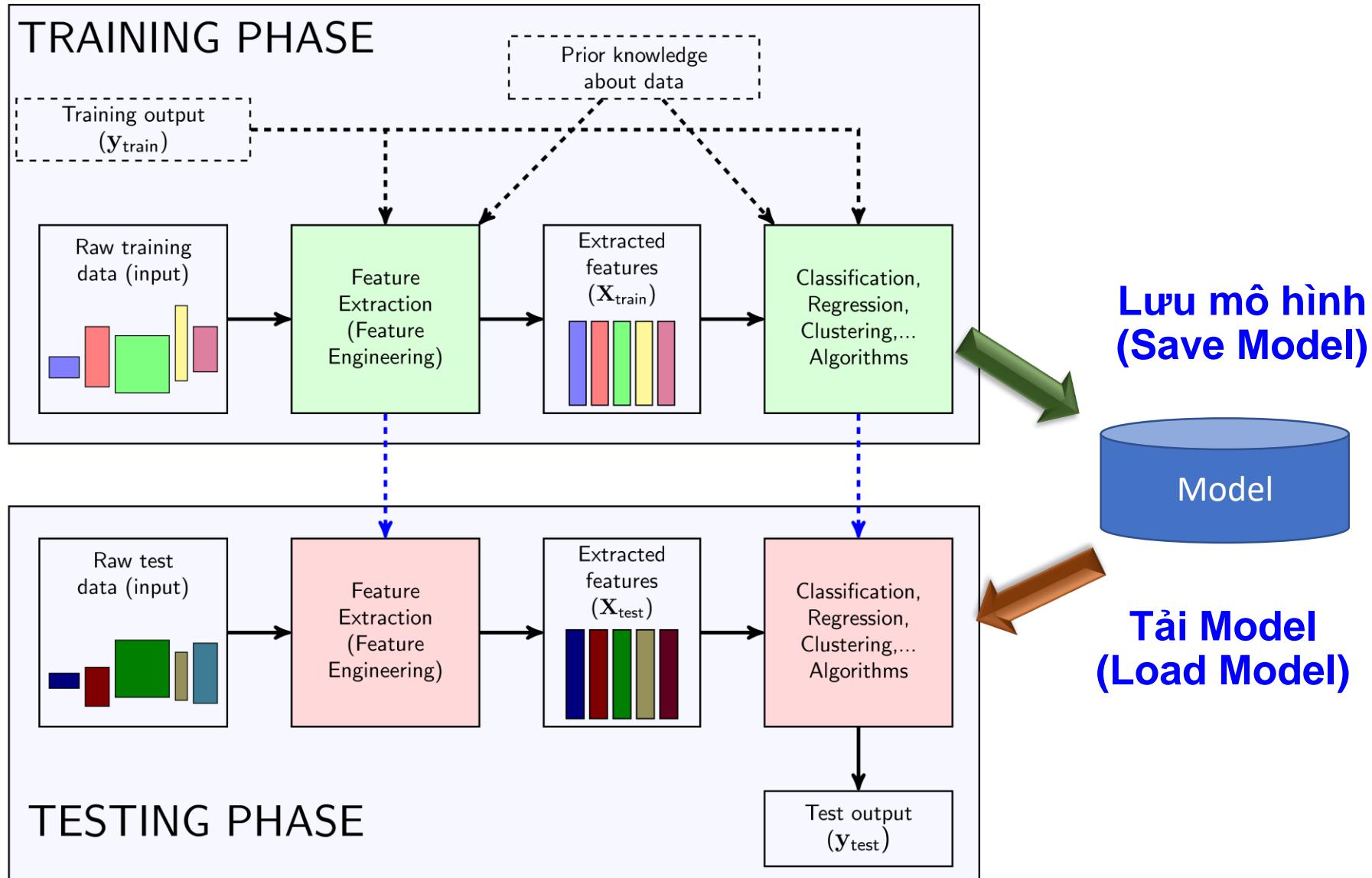
MÁY HỌC (MACHINE LEARNING)



HỌC SÂU (DEEP LEARNING)



BƯỚC TRÍCH ĐẶC TRƯNG TẠI VỊ TRÍ NÀO?



CÁC LOẠI ĐẶC TRƯNG PHỔ BIẾN



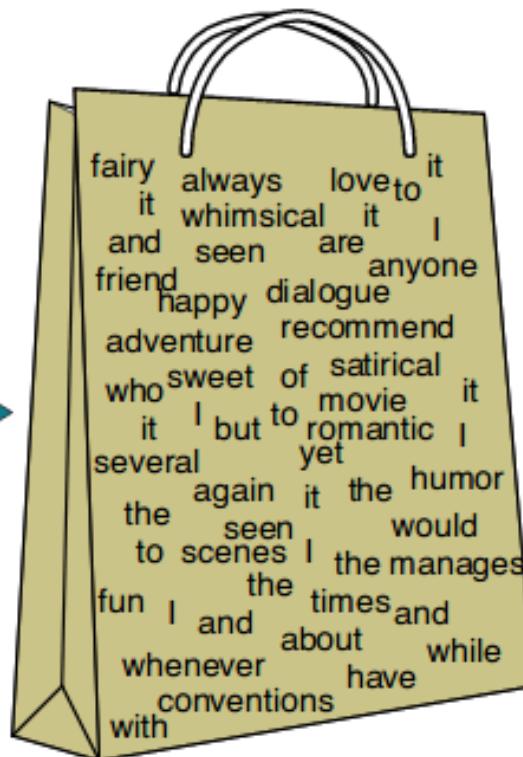
TRÍCH ĐẶC TRƯNG VĂN BẢN (TEXT)

Mô hình túi từ cơ bản

Đoạn văn bản Huấn luyện

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

Túi từ

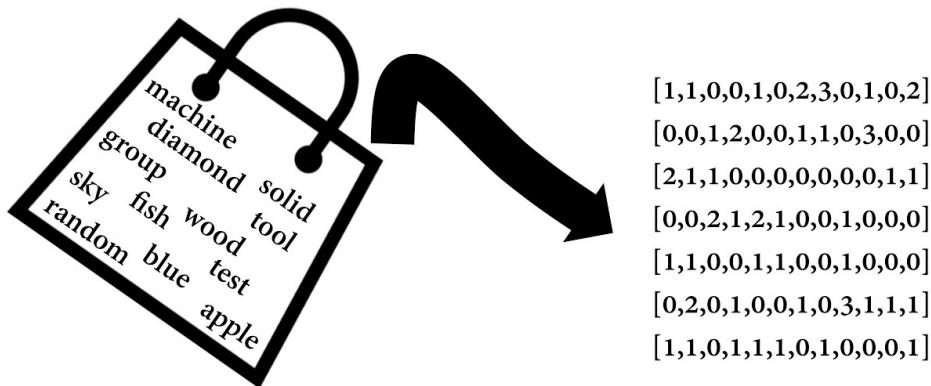


Thống kê từ

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

TRÍCH ĐẶC TRƯNG VĂN BẢN (TEXT)

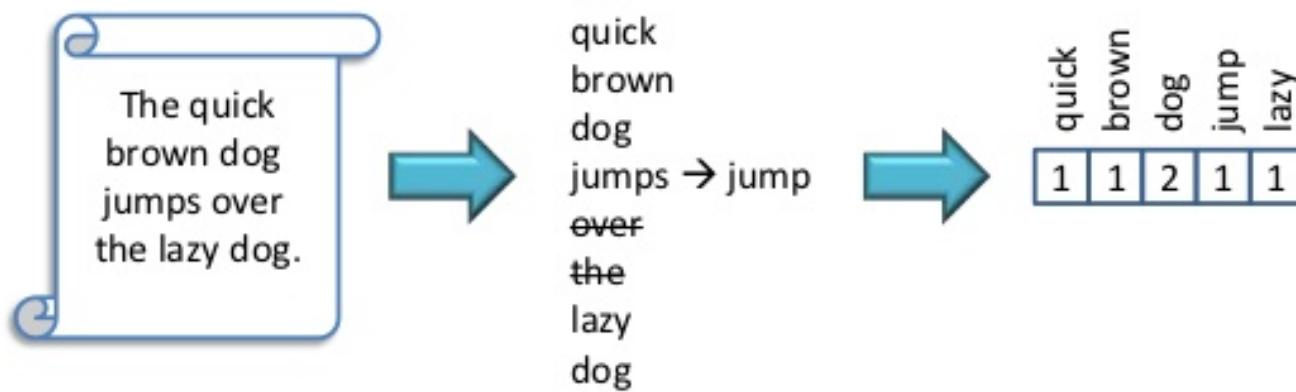
Mô hình túi từ cơ bản



**Đoạn văn bản
Huấn luyện**

Tiền xử lý các từ

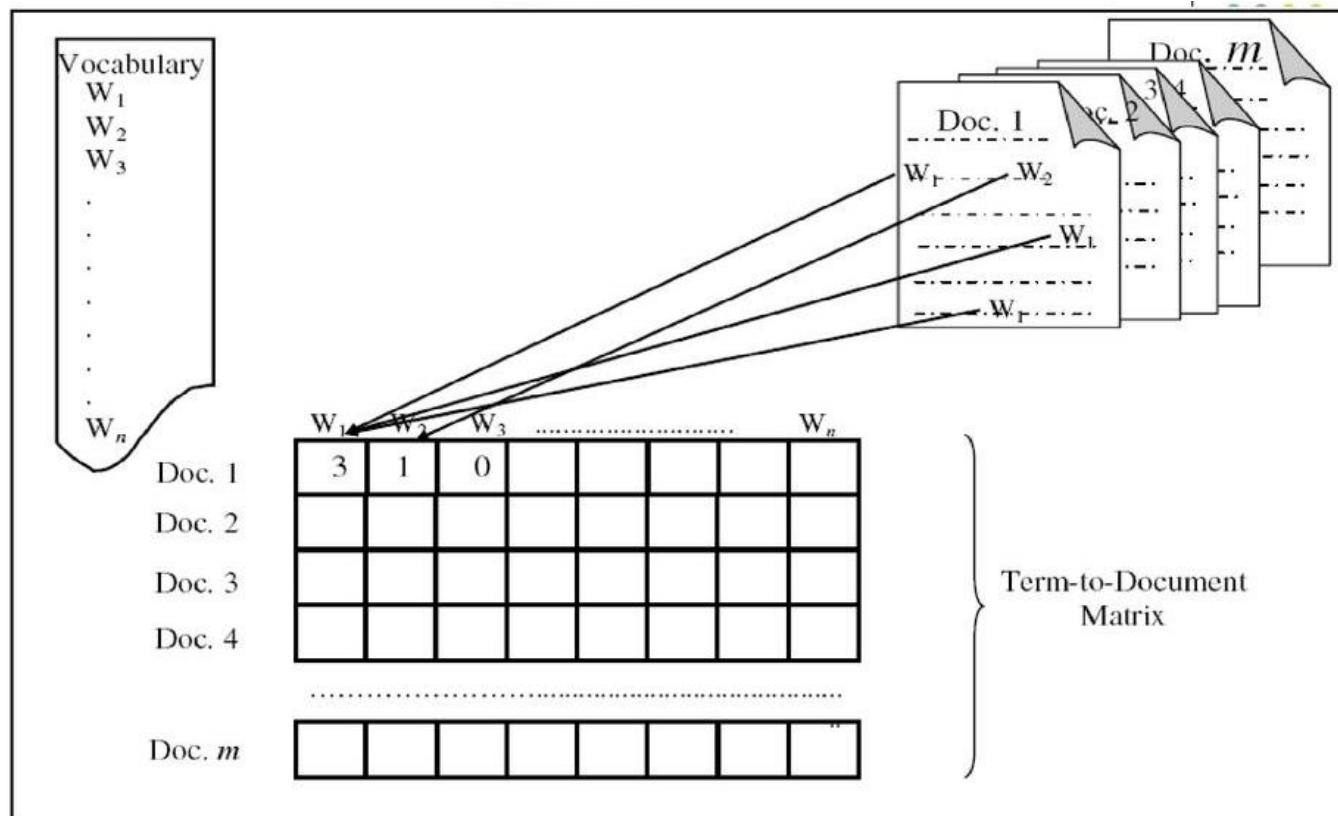
**Thống kê từ
(Tần số xuất hiện)**



TRÍCH ĐẶC TRƯNG VĂN BẢN (TEXT)

Mô hình túi từ cơ bản

Không gian ngôn ngữ tự nhiên → chuyển → KHÔNG GIAN VECTOR



TRÍCH ĐẶC TRƯNG VĂN BẢN (TEXT)

Mô hình túi từ cơ bản

2 câu

(1) John likes to watch movies. Mary likes movies too.

(2) John also likes to watch football games.

Từ vựng

["John", "likes", "to", "watch", "movies", "also", "football", "games",
"Mary", "too"]

Vectors

- (1) [1, 2, 1, 1, 2, 0, 0, 0, 1, 1]
(2) [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]

ĐẶC TRƯNG

Văn bản (1) có 1 từ “John”, 2 từ “likes”,
0 từ “also”, 0 từ “football”, ... nên ta thu được vector
tương ứng như trên.

Tần suất những từ được sử dụng nhiều nhất trong Truyện Kiều



Tiep VuHu

TRÍCH ĐẶC TRƯNG VĂN BẢN (TEXT)

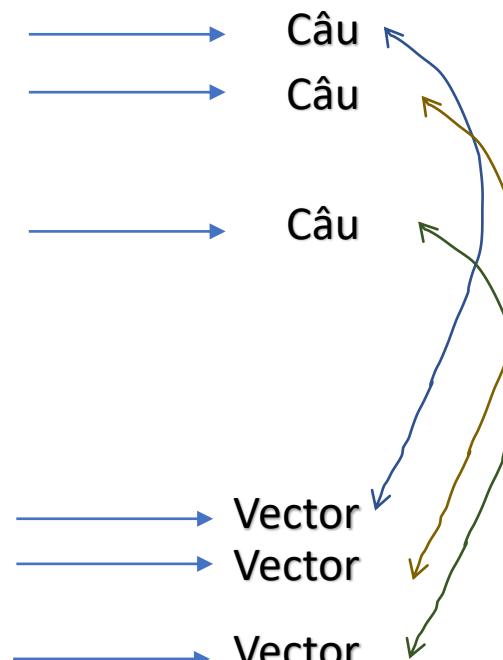
Mô hình túi từ cơ bản

Bảng 1: Ví dụ về tập dữ liệu văn bản

STT	Nội dung	Chủ đề
1	Brazil - đối thủ khắc tinh của Italy	Thể thao
2	Mưa đá dữ dội, hàng trăm nhà dân bị thiệt hại	Xã hội
...
M	Đột nhập nhà đại gia trộm 2 kg Pháp luật vàng	Pháp luật

Bảng 2: Biểu diễn tập dữ liệu văn bản bằng mô hình túi từ

STT	1 (bị)	2 (brazil)	...	n (tinh)	Chủ đề
1	0	1	...	1	Thể thao
2	1	0	...	0	Xã hội
...
m	0	0	...	0	Pháp luật



Đặc trưng
văn bản

Mô hình huấn luyện
Machine Learning

BÀI TẬP

- (1) Phúc thích xem phim. Đạt cũng thích xem phim.
- (2) Thành thích xem các trận bóng đá.
- (3) Minh vừa thích xem bóng đá vừa thích xem phim.
- (3) Quốc thích đi du lịch và xem bóng đá.
- (3) Bình vừa thích chơi đàn vừa thích xem phim.

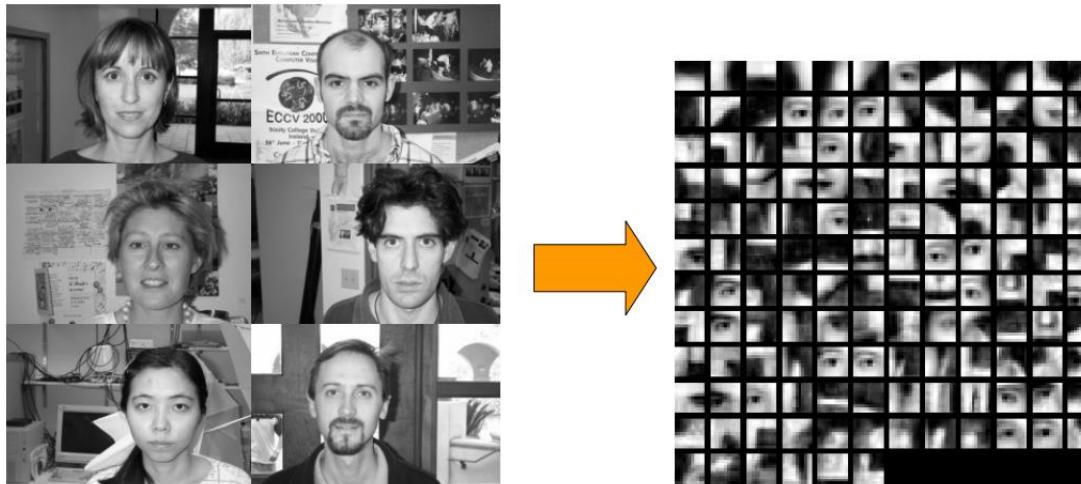
? Hãy mã hóa các câu trên thành các vector đặc trưng.

CÁC LOẠI ĐẶC TRƯNG PHỔ BIẾN

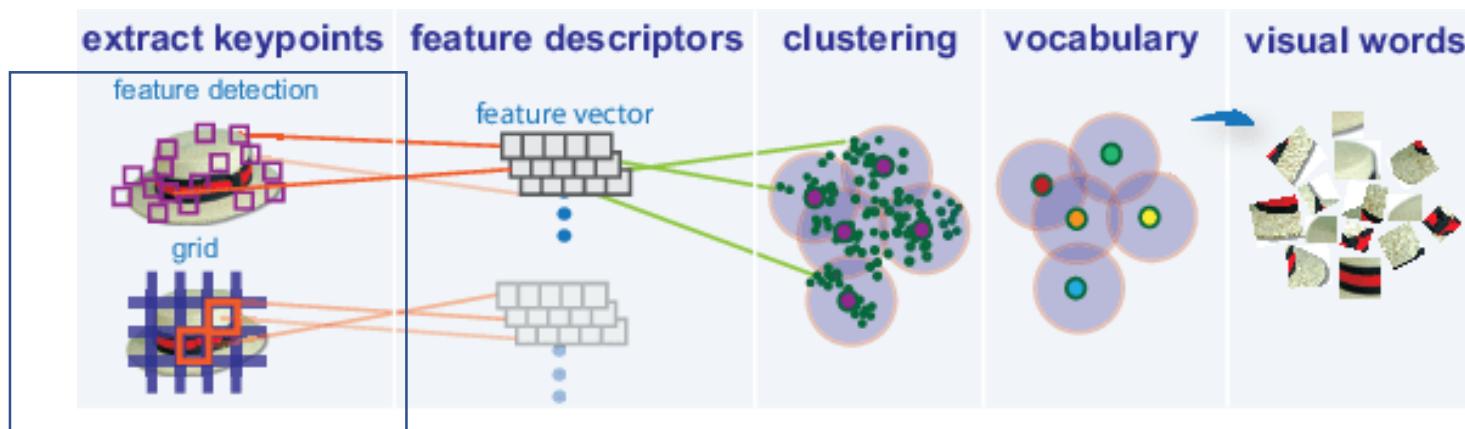


TRÍCH ĐẶC TRƯNG ẢNH (IMAGE)

Mô hình túi từ cơ bản

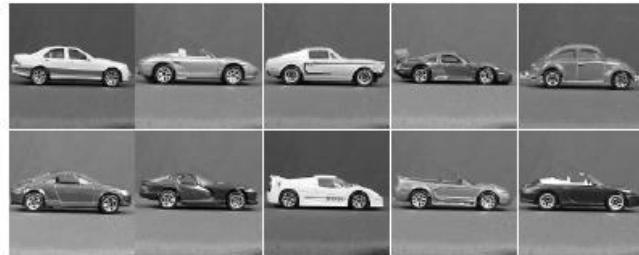


Sử dụng cửa sổ trượt để cắt các đặc trưng của mặt: Mắt, mày, miệng



TRÍCH ĐẶC TRƯNG ẢNH (IMAGE)

Mô hình túi từ cơ bản



Cố định cửa sổ trượt (i.e., 24x24) để thu các hình



Sử dụng Kmean để gom cụm cho việc tạo từ điển

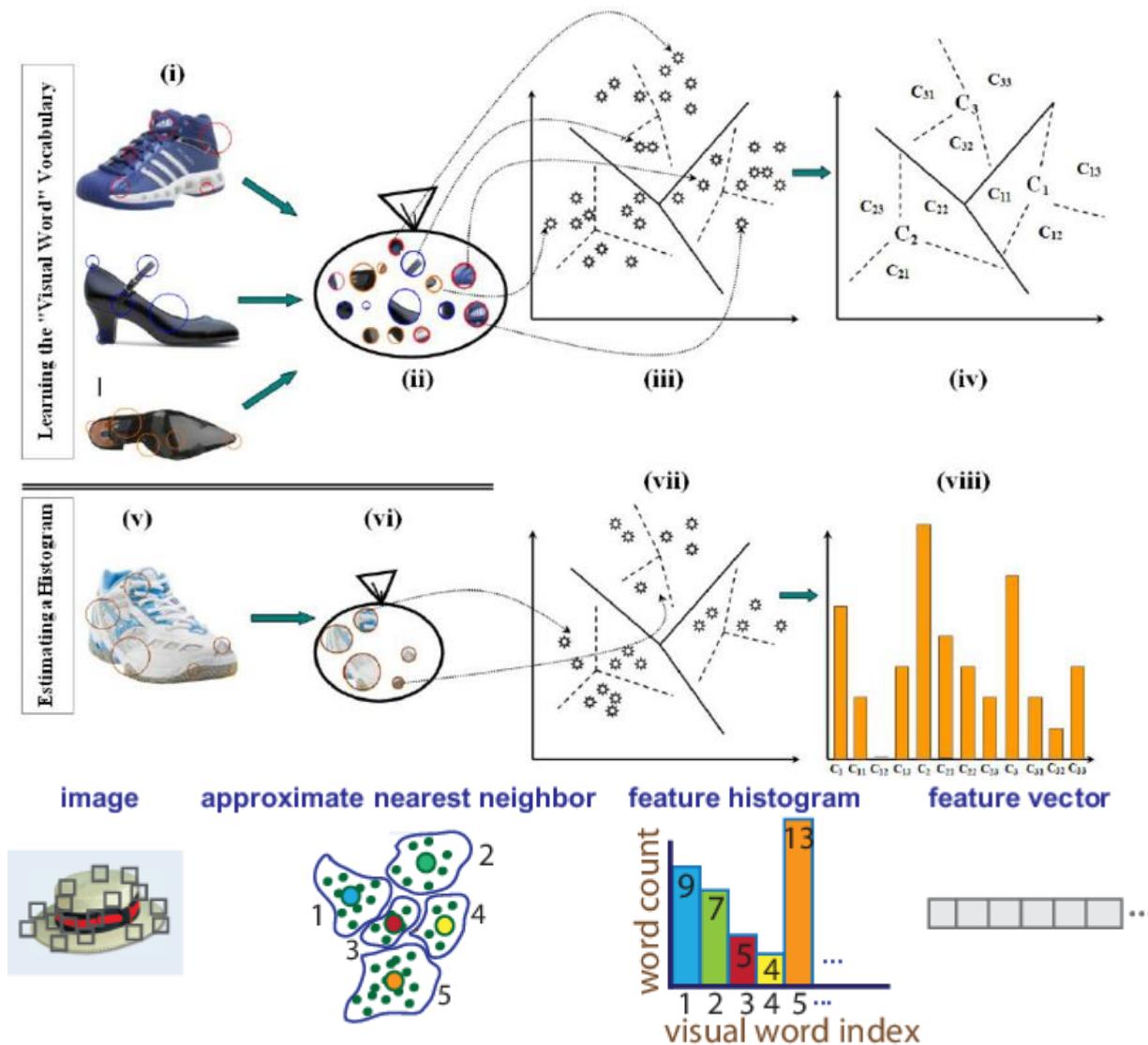


Sau khi có các hình “con”, chúng ta sẽ gom lại để tạo từ điển (nhiều từ vựng)

Source: B. Leibe

TRÍCH ĐẶC TRƯNG ẢNH (IMAGE)

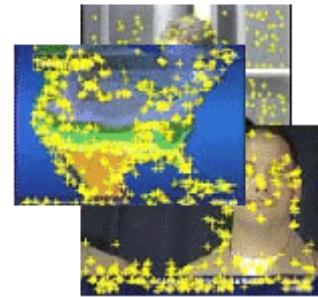
Mô hình túi từ cơ bản



TRÍCH ĐẶC TRƯNG ẢNH (IMAGE)

Mô hình túi từ cơ bản

Feature Extraction

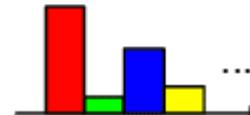
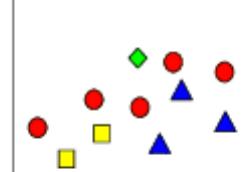
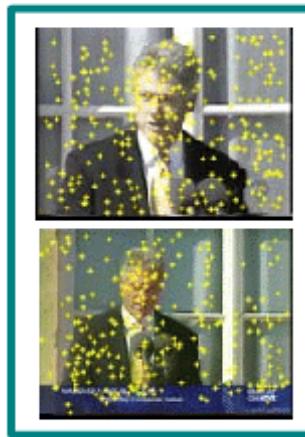


K-means

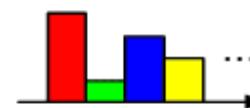
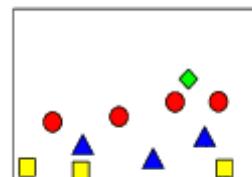
Keypoint
feature
space

Visual-word Vocabulary

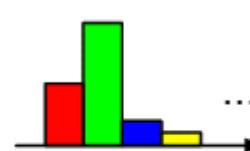
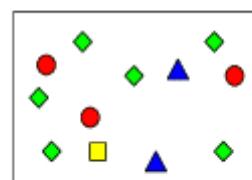
Có nhiều loại đặc
đặc trưng ảnh để sử dụng:
**SIFT, GIST, SIFT,
HOG,...**



6,1,4,5,...



5,1,4,3,...



4,7,1,6,...



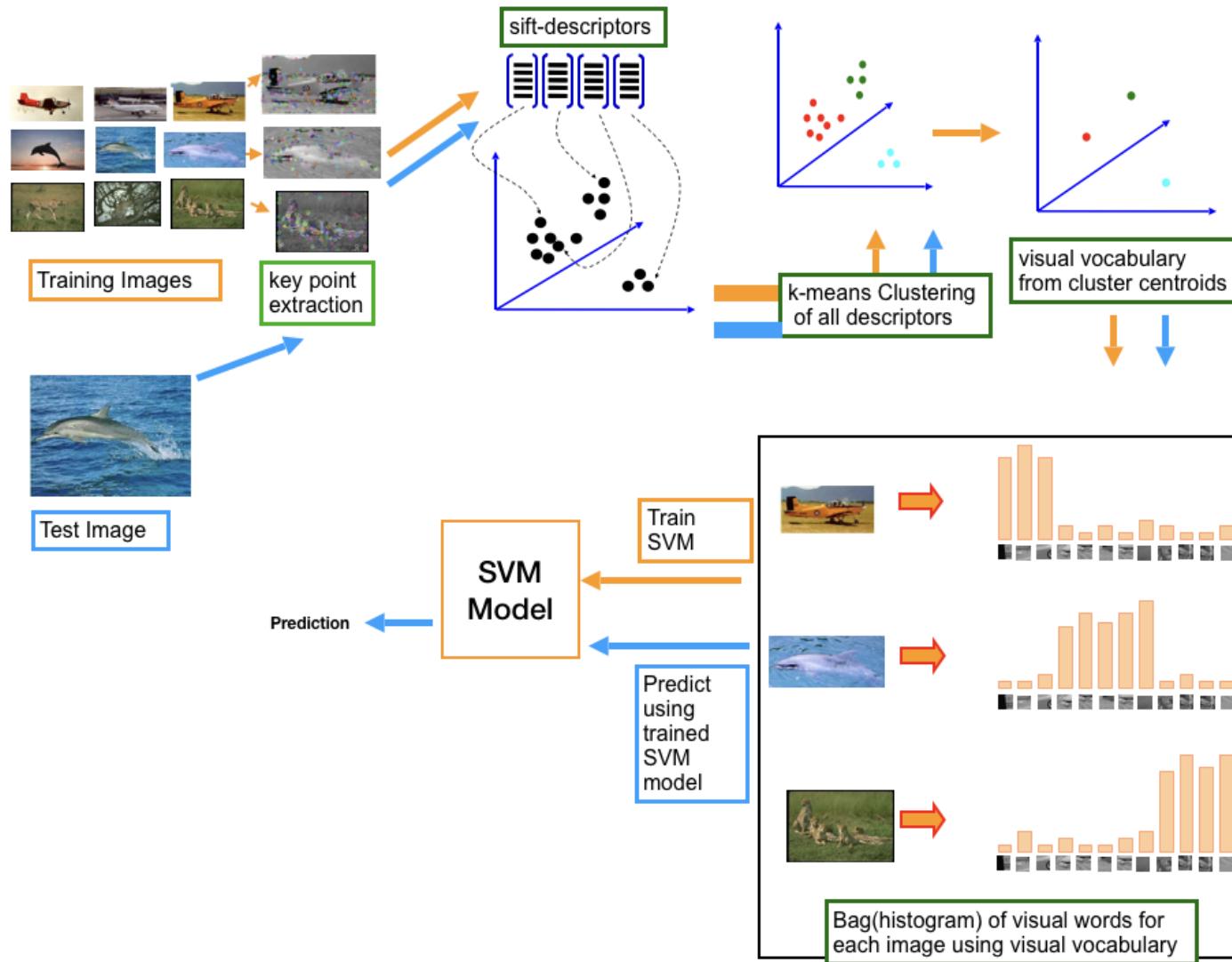
"bags of visual words"

Visual-word vectors

Đưa vào
Máy học
để
Huấn luyện

TRÍCH ĐẶC TRƯNG ẢNH (IMAGE)

Sử dụng SIFT để trích đặc trưng ảnh

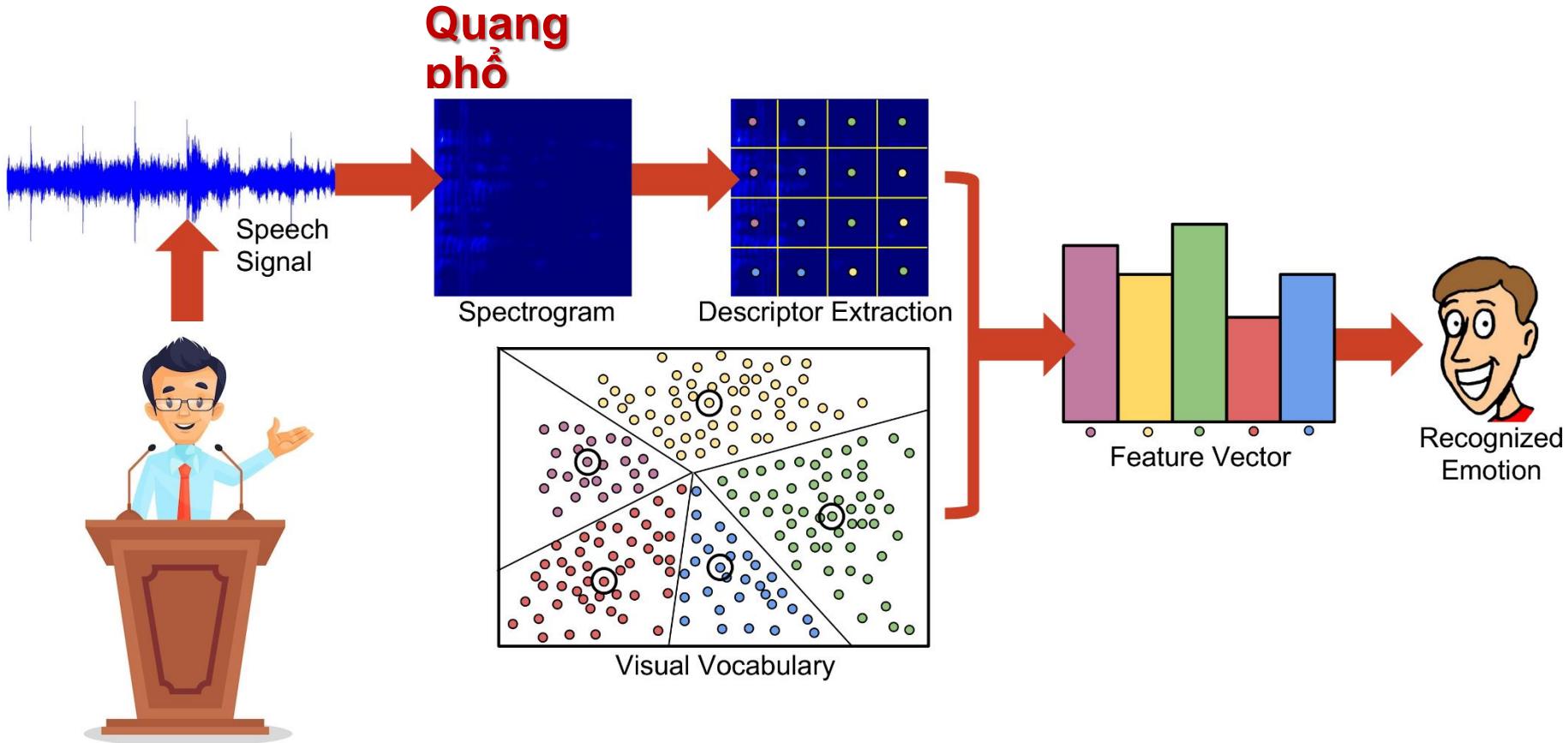


CÁC LOẠI ĐẶC TRƯNG PHỔ BIẾN



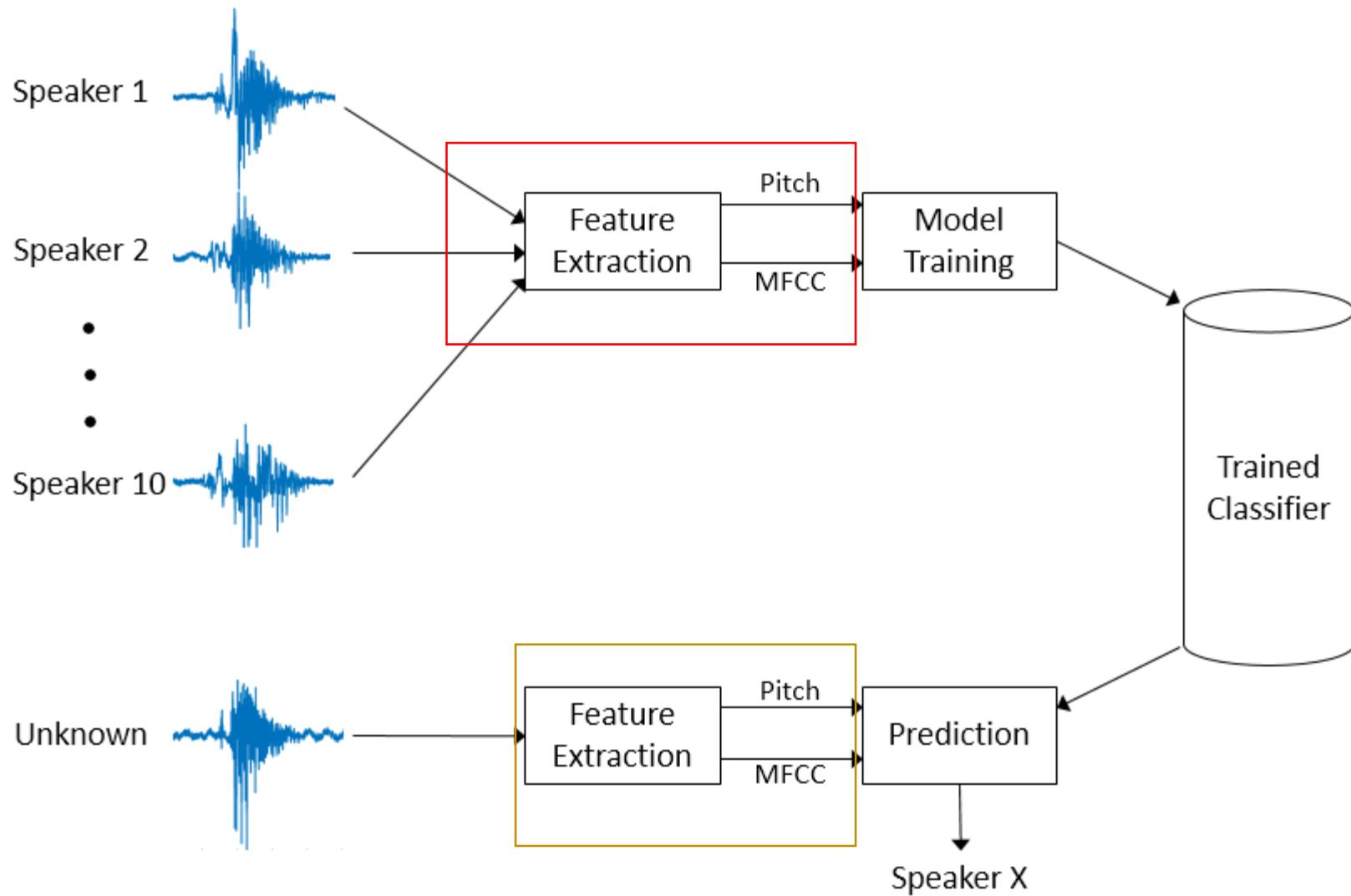
TRÍCH ĐẶC TRƯNG ÂM THANH (SOUNDS)

Sử dụng quang phổ



TRÍCH ĐẶC TRƯNG ÂM THANH (SOUNDS)

Sử dụng MFCC (Mel Frequency Cepstral Coefficients)



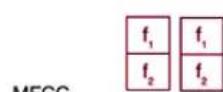
TRÍCH ĐẶC TRƯNG ÂM THANH (SOUNDS)

Sử dụng MFCC (Mel Frequency Cepstral Coefficients)



25 ms
10 ms apart
extract

Frames of acoustic feature vectors



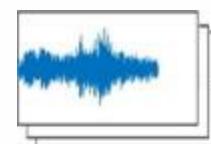
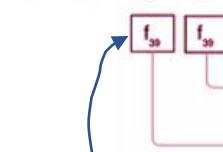
...

...

MFCC

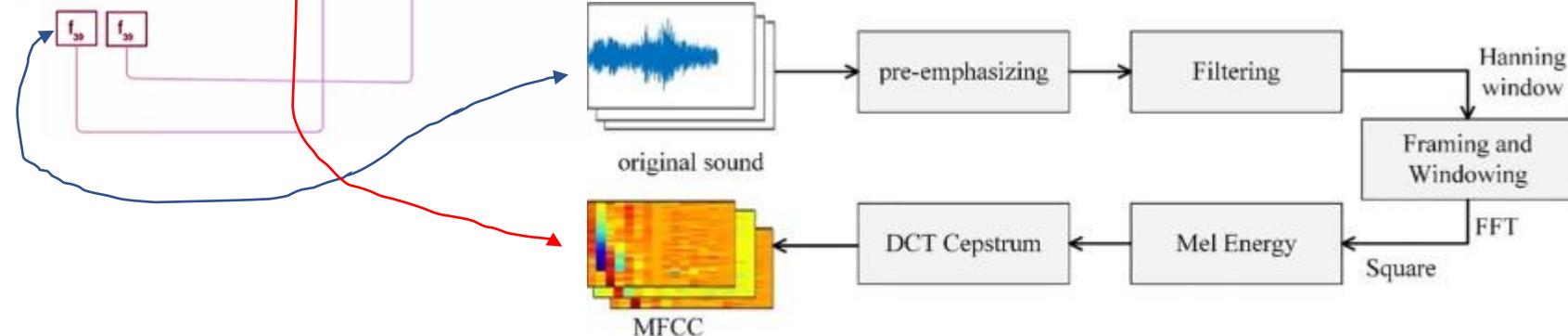
39 features

each frame contains 39 MFCC features



MFCC sẽ lấy ra vài giây để trích đặc trưng

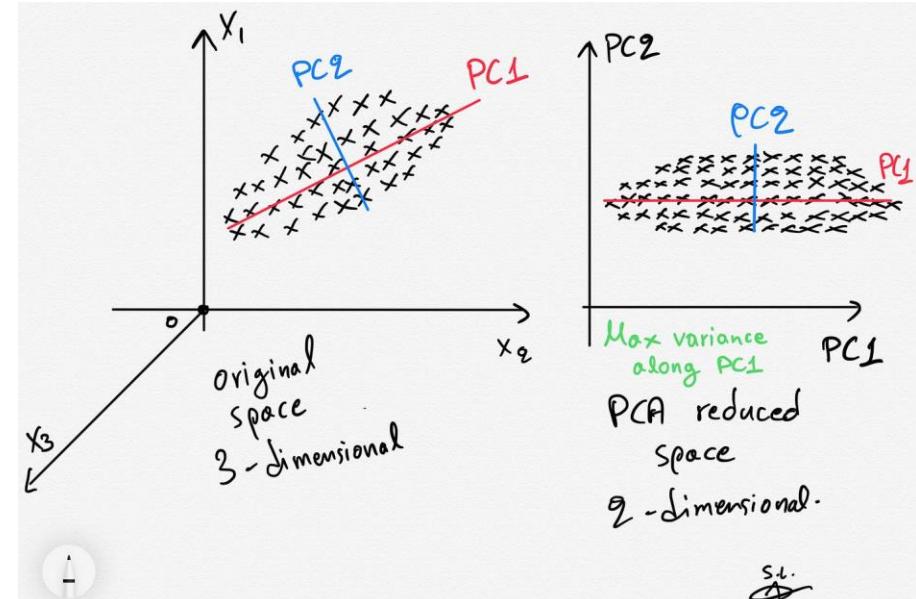
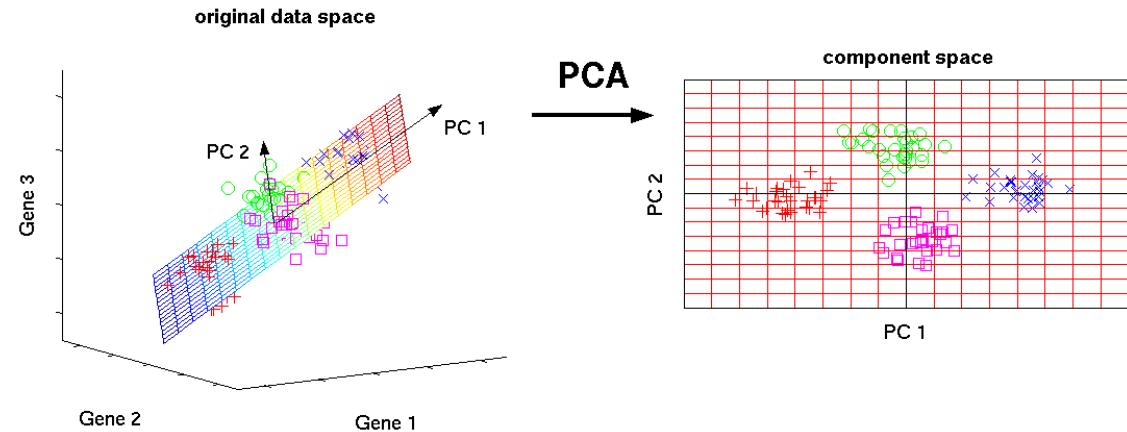
Mỗi đoạn (vài giây) sẽ có trích ra được 39 features



GIẢM CHIỀU DỮ LIỆU (Dimensionality reduction)

Các phương pháp giảm kích thước tìm cách lấy một tập hợp lớn dữ liệu và **trả về một tập hợp nhỏ hơn** với các thành phần vẫn chứa hầu hết thông tin trong tập dữ liệu gốc.

Một trong những hình thức giảm kích thước đơn giản nhất là **PCA**. Phân tích thành phần chính (PCA) là một phương pháp toán học giúp biến đổi một số biến số tương quan (ví dụ: gene expression) thành một số (nhỏ hơn) các biến không tương quan được gọi là thành phần chính ("PC").



CÂU HỎI VÀ THẢO LUẬN

