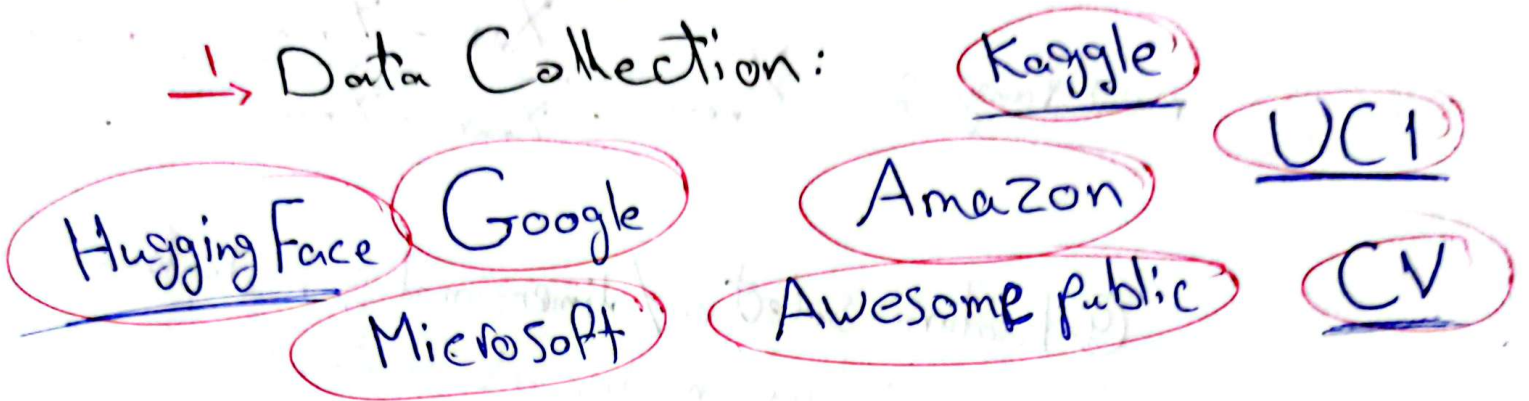


# Machine Learning Process:

## I Data Preparation:

1 Data Collection:



2 Data preprocessing:

① Data cleaning  $\Rightarrow$  Undesired data (none Value)

$\rightarrow$  Remove row from dataset.

$\rightarrow$  replace them by Column average.

② Data encoding  $\Rightarrow$  ML require Numerical input

$\rightarrow$  Convert Categorical Value to numerical.

$\rightarrow$  target must also be Converted to numerical if it's Categorical (nominal) Values.

1 binary Features (gender - sign(+, -)...)  $\Rightarrow 0, 1$

2 Feature has more than 2 Values

② One-hot encoding

① replace them by

numbers

egypt : 0

UK : 1

US : 2

Country	X <sub>1</sub>	X <sub>2</sub>	...
Egypt	1	0	
UK	2	2	
US	3	1	
US	4	4	

Country	X <sub>1</sub>	X <sub>2</sub>	...
Egypt	1	0	0
UK	0	1	0
US	0	0	1
US	0	0	1

③ Data Normalization  $\Rightarrow$  different ranges  
for each features.

$[X_1 \rightarrow 0-10] \mid [X_2 \rightarrow 0-100000]$

$\rightarrow$  make them have the same range

new range  $\circ$   $X_{\text{new}} = \frac{X_{\text{old}} - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \quad (0 \rightarrow 1)$

④ Feature Selection/dimensionality reduction

$\Rightarrow$  large number of features.

1  $\rightarrow$  Feature selection: select the most important  
features and drop the ~~rest~~ others.

2  $\rightarrow$  Dimensionality reduction: calculating  
new features from the original.

⑤ Data splitting  $\Rightarrow$  data

