# 2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
# Day 13 – 暴力破解 DNN 模型

# 大綱

- 暴力破解 DNN 模型
  - 甚麼是暴力破解
  - 應用在機器學習上的差異
  - 使用的條件及時機
  - 題目解說
- 結論

# 甚麼是暴力破解?

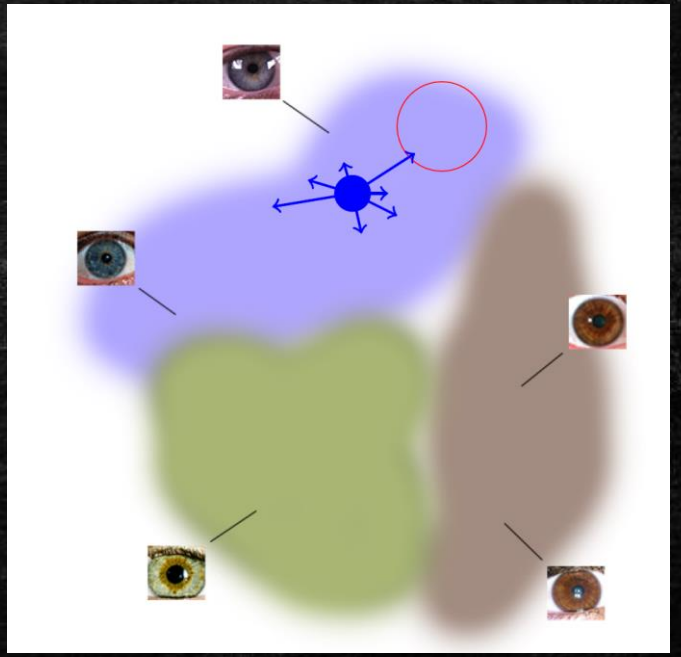
- 暴力破解攻擊（英語：Brute-force attack）是一種密碼分析的方法，主要透過軟體逐一測試可能的密碼，直到找出真正的密碼為止

- 字典攻擊
  - 使用預先製作好的清單，例如：英文單字、生日的數字組合等等，利用一般人習慣設定過短或過於簡單的密碼進行破譯，很大程度上縮短破譯時間

https://zh.wikipedia.org/zh-tw/%E8%9B%AE%E5%8A%9B%E6%94%BB%E5%87%BB

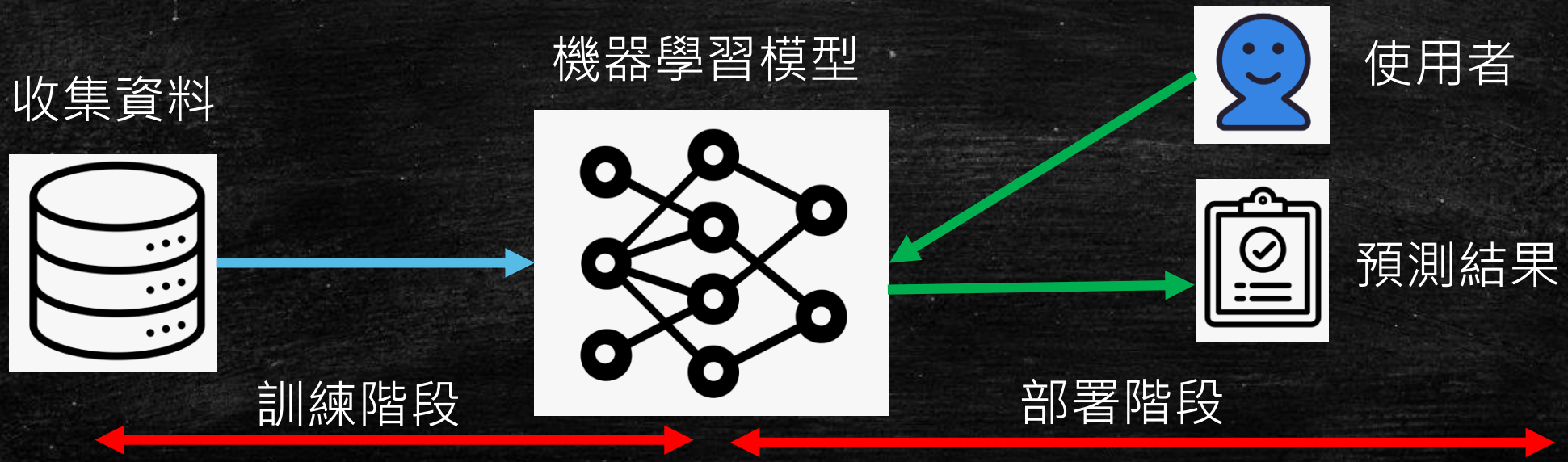
# 應用在機器學習上的差異

- 主要強調的點有兩個，第一個是針對輸入資料的特性盡量找相似的東西為基底去做暴力破解
- 第二個則在類似的圖形加上一些輕微、隨機的擾動，就有機會讓機器學習模型辨識成合法資料

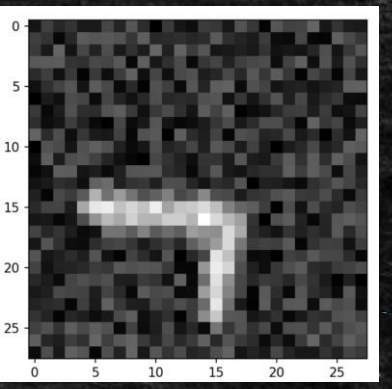


# 使用條件及時機

- 時機點：部署階段
- 前提：攻擊者必須有個基底加上隨機的擾動作為攻擊輸入資料
- 攻擊效果：讓機器模型誤以為輸入資料為合法資料



# 題目解說




- https://github.com/Kayzaks/HackingNeuralNetworks/tree/master/3_BruteForcing
- pip install scikit-image

## Exercise 3-0

You are trying to Brute-Force a Image-based Security control. So far your attempt (see 'exercise.py') has proven to be unreliable. However, you have some vague idea of what an image that gives access should look like (see 'fake_id.png'), but it doesn't work either.

- Develop a brute-force strategy that has a success rate of about 5% or better (10% or better for a challenge)
- Do not modify 'model.h5'
- Do not simply draw a '4' in paint...

A solution for 5% and 10% can be found in 'solution_3_0.py'

# 結論

- 暴力破解攻擊已經是個行之有年的攻擊手法，但對應到機器學習模型上卻有不一樣的視角
- 暴力破解的字典檔對應到與攻擊目標相似的資料即可
- 跟暴力破解密碼不同，模型方面不需要破解到跟原本資料100%相同，而是只需要**基底 + 擾動**通過模型的辨識即可