

2024 鐵人賽 – 我數學就爛要怎麼來  
學 DNN 模型安全

Day 09 – 竄改 DNN 模型參數

---



# 大綱

- 竄改DNN 模型參數
  - 攻擊手法原理
  - 使用條件及時機
  - 程式實作
- 結論

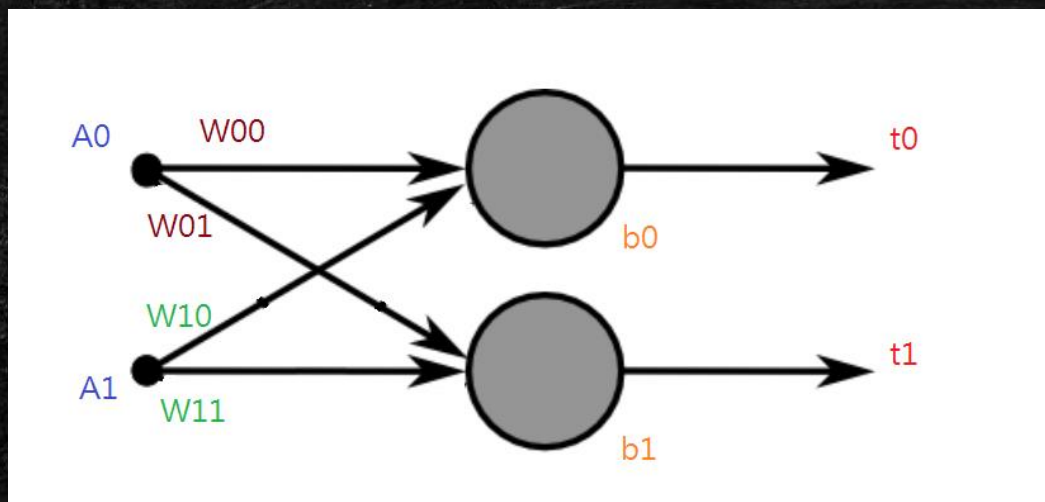
當駭客修改完我的顏值辨識  
模型參數





# 攻擊手法原理

- ML10:2023 Model Poisoning
  - 攻擊者去竄改機器學習模型內的參數
  - 以下為例，如果調整  $b_0$  值為很大的數值，則模型判斷出來的  $t_0$  機率就會變很高，導致模型產生異常的行為

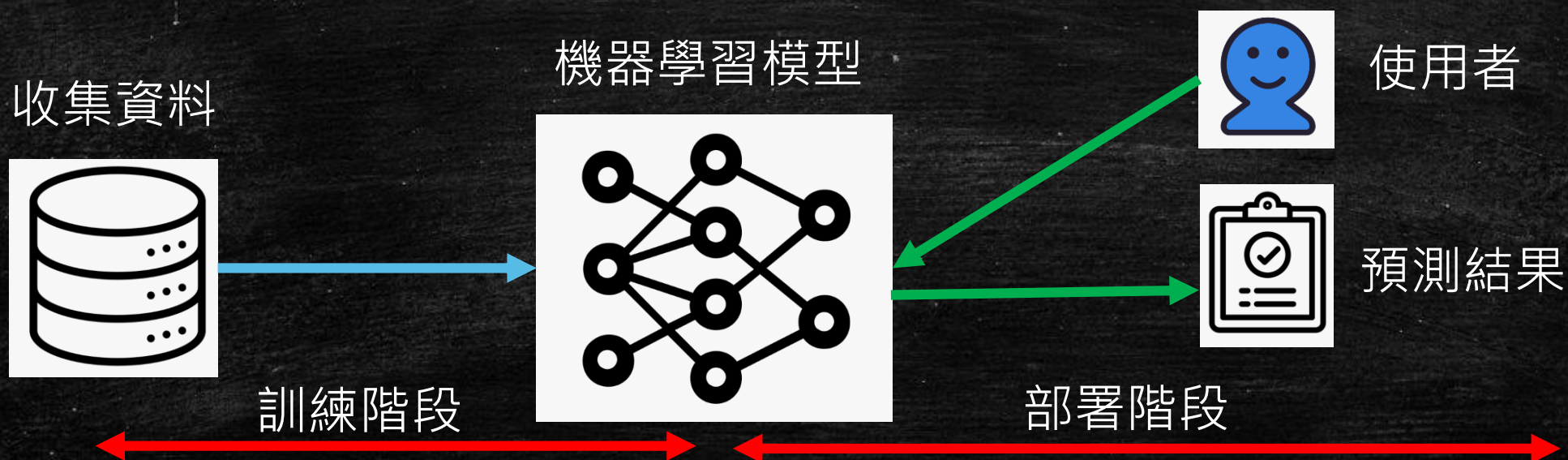


$$\begin{bmatrix} W_{00} & W_{10} \\ W_{01} & W_{11} \end{bmatrix} \times \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} t_0 \\ t_1 \end{bmatrix}$$



# 使用條件及時機

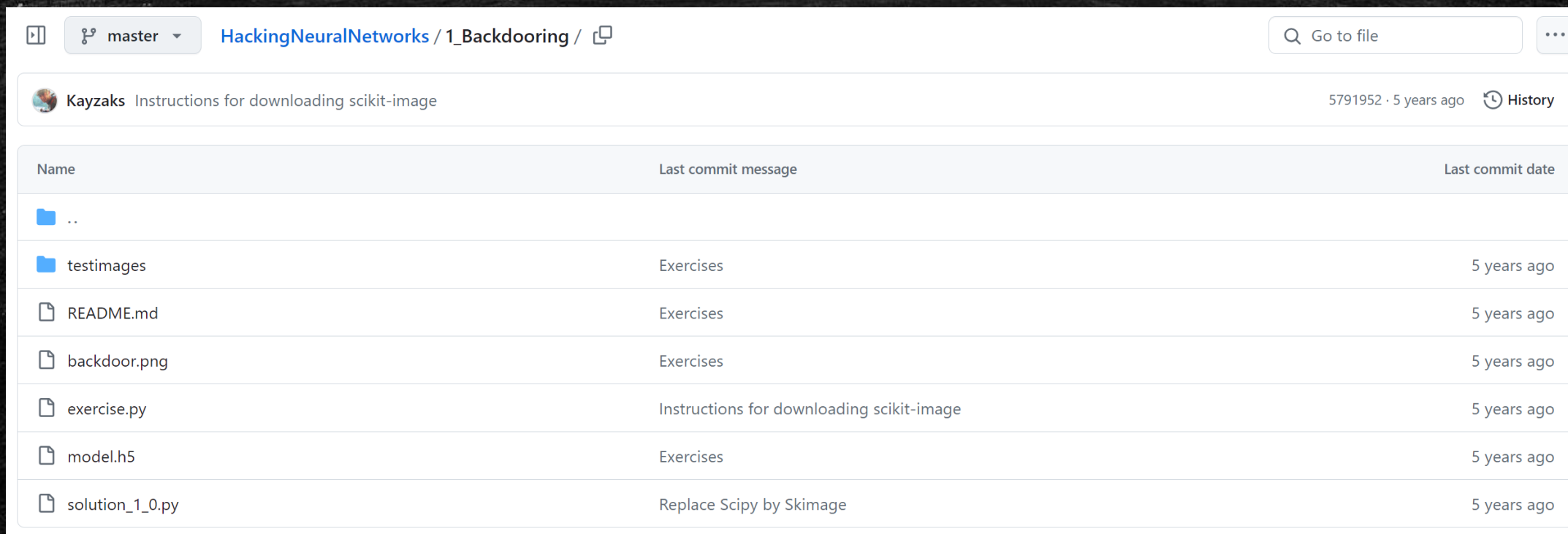
- 時機點：部署階段
- 前提：攻擊者必須能夠讀取及寫入機器學習模型
- 攻擊效果：讓模型輸出駭客預期的結果





# 程式實作 – 參考資料

- Hacking Neural Networks: A Short Introduction
- <https://github.com/Kayzaks/HackingNeuralNetworks>



Name	Last commit message	Last commit date
..		
testimages	Exercises	5 years ago
README.md	Exercises	5 years ago
backdoor.png	Exercises	5 years ago
exercise.py	Instructions for downloading scikit-image	5 years ago
model.h5	Exercises	5 years ago
solution_1_0.py	Replace Scipy by Skimage	5 years ago



## 結論

---

- 參數竄改調整很容易被發現，因為模型預測會偏向某個結果
- 但假如這個模型是個**辨識惡意行為**的模型，如果攻擊者調整成都不會告警，是否就變相成為繞過這個偵測機制的手法