2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 32 – 完賽心得

大綱

- DNN 模型安全
  - 系列賽回顧
  - 未來展望
- 結論

系列賽回顧

DNN模型基本概念
1. 模型建立
2. 模型參數調整
3. 模型瀏覽

Day1 — Day7

初探DNN模型攻擊
1. 參數竄改
2. 輸入回推
3. 暴力破解
4. 溢位攻擊

Day8 — Day15

深入DNN模型攻擊
1. 後門建立
2. 對抗式攻擊樣本
3. 梯度洩漏攻擊
4. 乾淨標籤投毒攻擊
5. **隱藏式觸發後門**

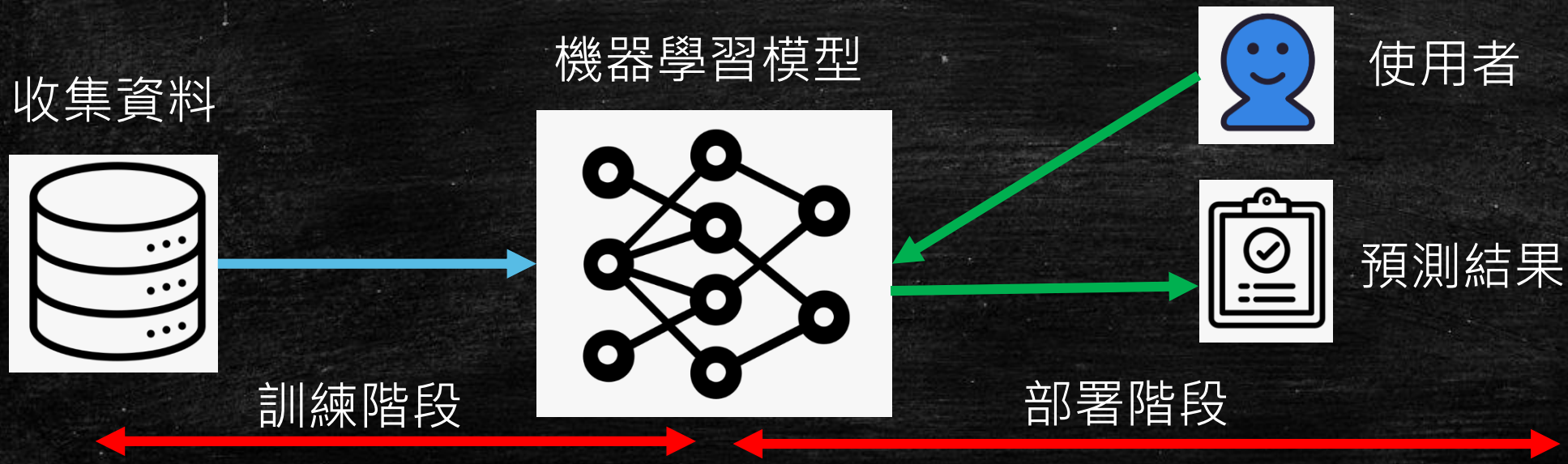
Day16 — Day29

補完篇
1. 後門程式建立
2. 後門程式分析

Day31

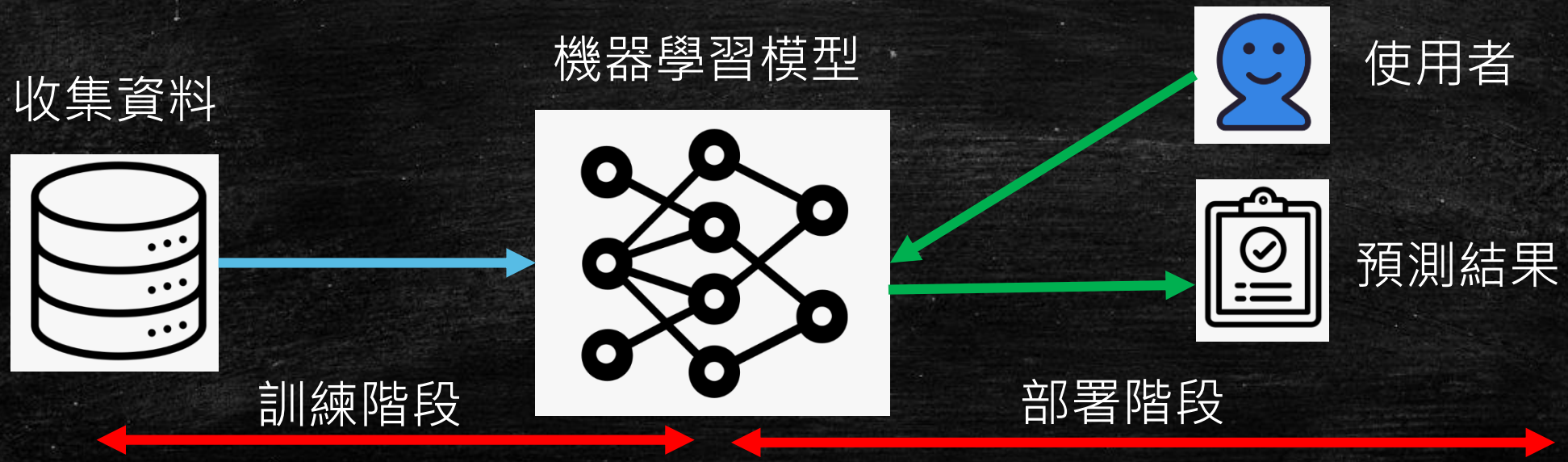
系列賽回顧

- 機器學習的核心概念是找到輸入與輸出資料對之間的數學關聯。機器學習模型一開始並不知道這個關聯是怎樣的，但隨著給予其足夠的資料，模型的預測會越來越準確
- 通常會分為兩個階段，一個是訓練學習階段，另一個則是學習完後的部署階段

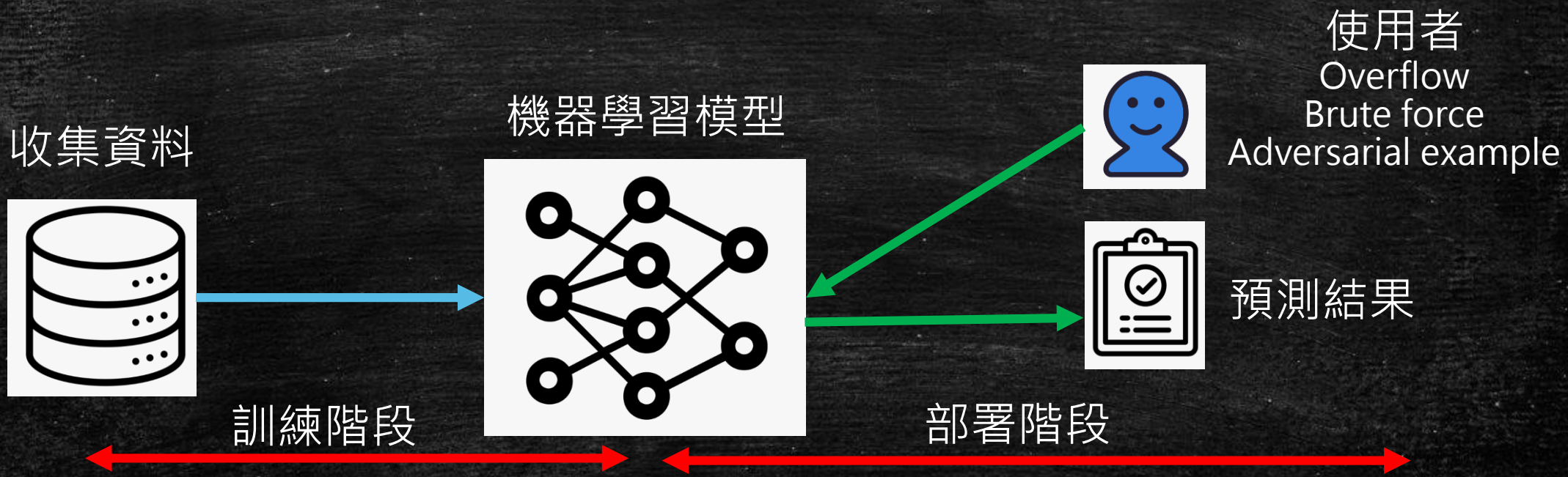


系列賽回顧

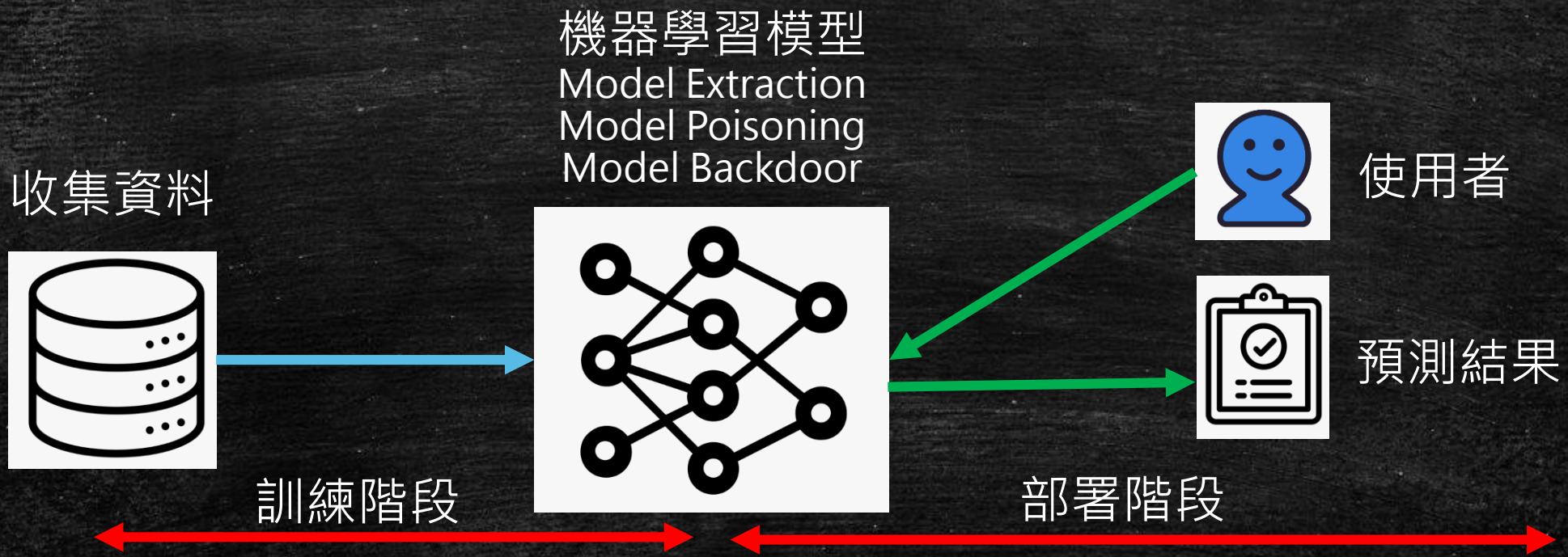
- 發生時機點： 訓練階段 vs. 部署階段
- 需要知道多少：白箱 vs. 黑箱
- 攻擊對象：收集資料、機器學習模型、使用者資料



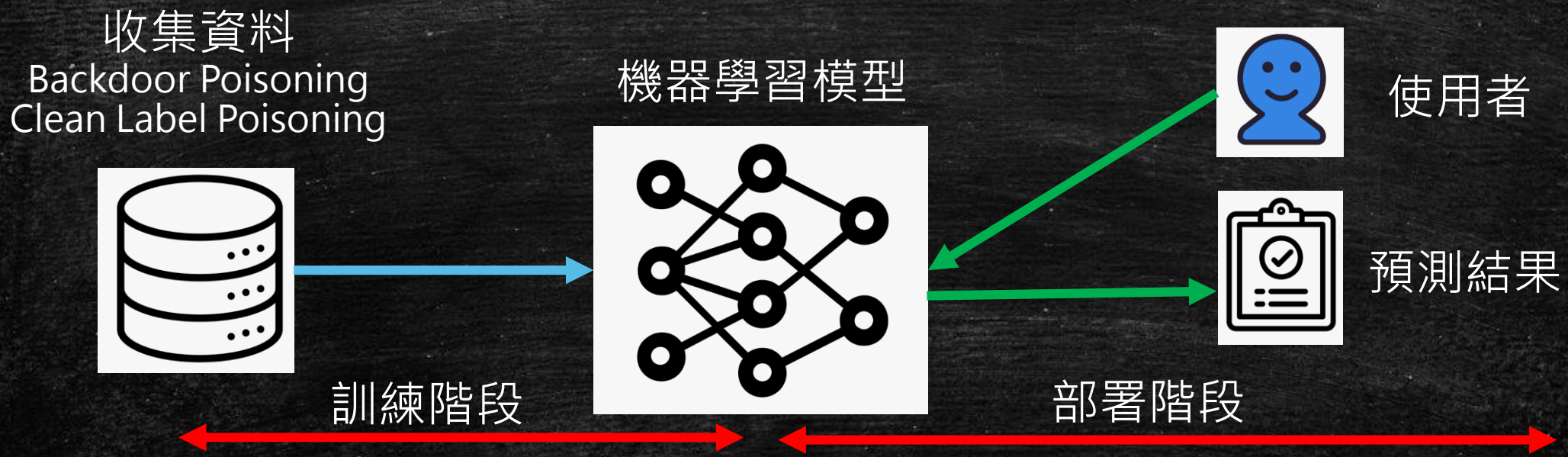
系列賽回顧


收集資料
機器學習模型
使用者
Overflow
Brute force
Adversarial example
預測結果
訓練階段
部署階段

系列賽回顧


機器學習模型
Model Extraction
Model Poisoning
Model Backdoor
收集資料
使用者
預測結果
訓練階段
部署階段

系列賽回顧


收集資料
Backdoor Poisoning
Clean Label Poisoning
機器學習模型
使用者
預測結果
訓練階段
部署階段

未來展望 – 那些沒有說的東東 (1/4)

- 黑箱攻擊

- **Score-based attacks:** In this setting, attackers obtain the model's confidence scores or logits and can use various optimization techniques to create the adversarial examples. A popular method is zeroth-order optimization, which estimates the model's gradients without explicitly computing derivatives [66, 137]. Other optimization techniques include discrete optimization [210], natural evolution strategies [136], and random walks [216].
- **Decision-based attacks:** In this more restrictive setting, attackers obtain only the final predicted labels of the model. The first method for generating evasion attacks was the Boundary Attack based on random walks along the decision boundary and rejection sampling [35], which was extended with an improved gradient estimation to reduce the number of queries in the HopSkipJumpAttack [65]. More recently, several optimization methods search for the direction of the nearest decision boundary (the OPT attack [71]), use sign SGD instead of binary searches (the Sign-OPT attack [72]), or use Bayesian optimization [271].

The main challenge in creating adversarial examples in black-box settings is reducing the number of queries to the ML models. Recent techniques can successfully evade the ML classifiers with a relatively small number of queries, typically less than 1000 [271].

未來展望 – 那些沒有說的東東 (2/4)

- 移轉攻擊

2.2.3. Transferability of Attacks

Another method for generating adversarial attacks under restrictive threat models is via transferability of an attack crafted on a different ML model. Typically, an attacker trains a substitute ML model, generates white-box adversarial attacks on the substitute model, and transfers the attacks to the target model. Various methods differ in how the substitute models are trained. For example, Papernot et al. [232, 233] train the substitute model with score-based queries to the target model, while several papers train an ensemble of models without explicitly querying the target model [181, 299, 315].

未來展望 – 那些沒有說的東東 (3/4)

- 成員推論攻擊

ML04:2023 Membership Inference Attack

Description

Membership inference attacks occur when an attacker manipulates the model's training data in order to cause it to behave in a way that exposes sensitive information.

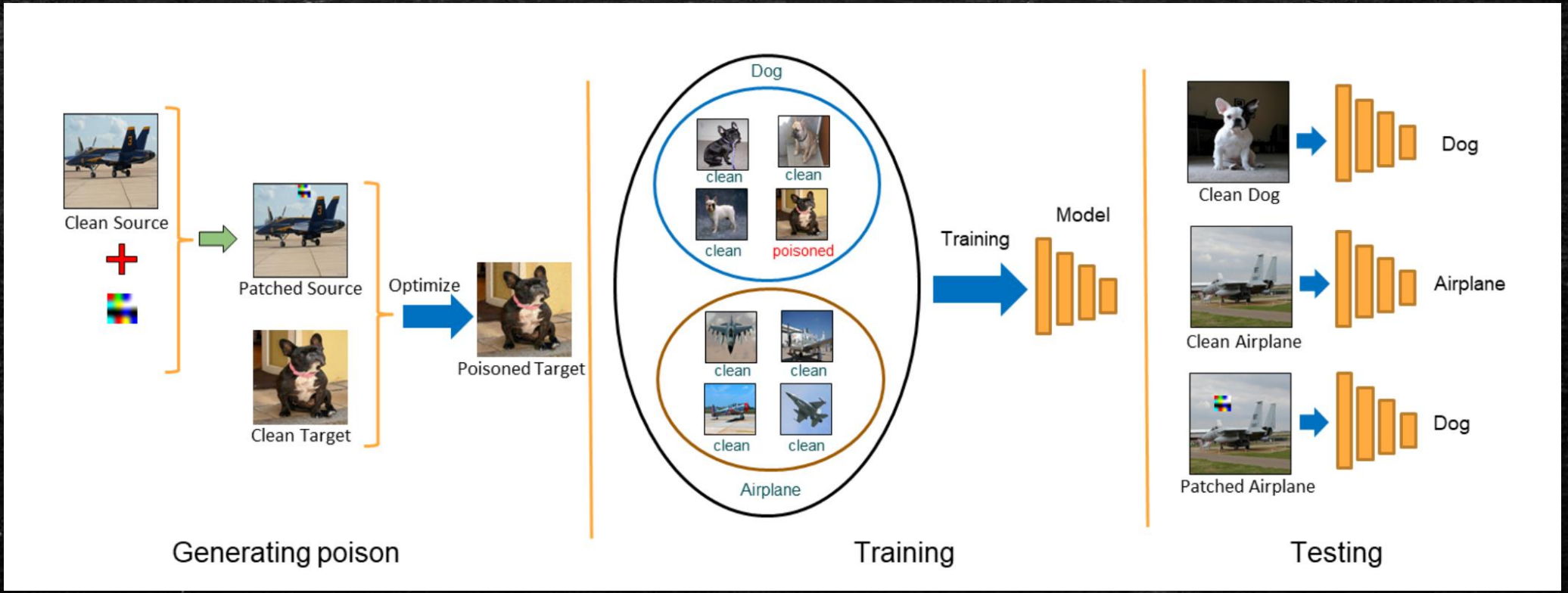
- **Membership attack**：給定一筆資料，測試它是否在 training dataset 之中。

https://medium.com/trustableai/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E6%BD%9B%E5%9C%A8%E7%9A%84%E9%9A%B1%E7%A7%81%E5%95%8F%E9%A1%8C-9410eb951411

未來展望 – 那些沒有說的東東 (4/4)

https://arxiv.org/abs/1910.00033

- 很酷的 Hidden Trigger Backdoor Attacks (2019)
  - 但我就爛實作不出來



未來展望 – 防禦及偵測

https://medium.com/trustableai/%E9%87%9D%E5%B0%8D%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E7%9A%84%E6%83%A1%E6%84%8F%E8%B3%87%E6%96%99%E6%94%BB%E6%93%8A-%E4%B8%80-e949877427c7

- 防禦
  - 被動防禦：試圖在 input 中過濾出對抗訊息
  - 主動防禦：讓訓練出來的 model 擁有抵抗對抗式樣本的能力
  - 安全掃描：針對模型內容進行檢查

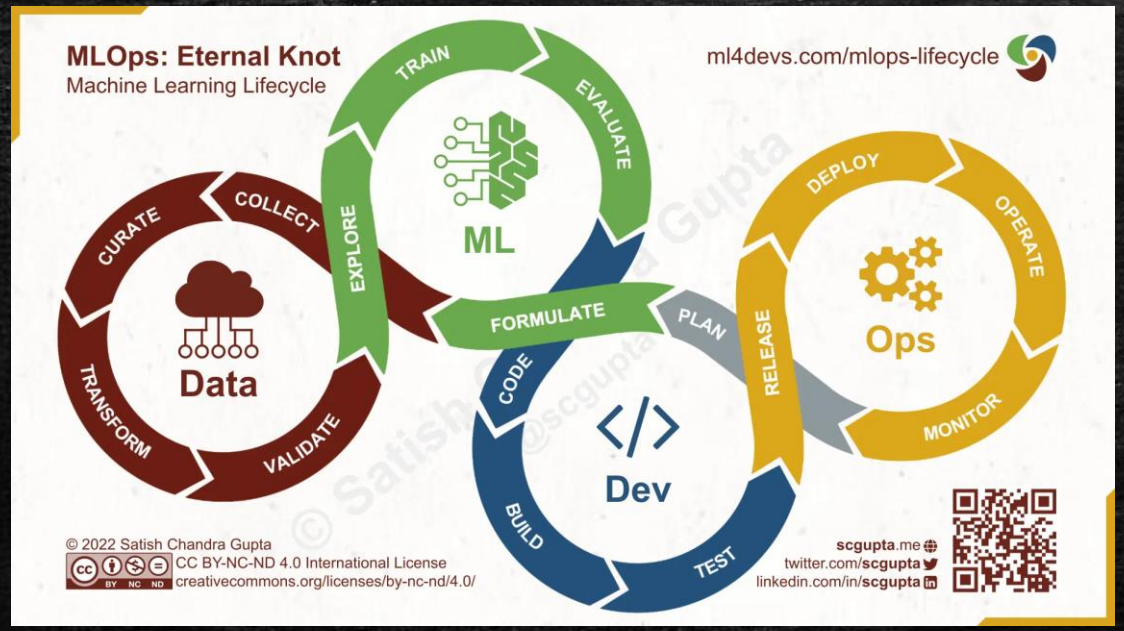
- 偵測
  - 監控模型的是否有異常的變動或是異常行為 (ex: 奇怪的對外連線)
  - 監控模型學習的狀況、部署後的預測情況

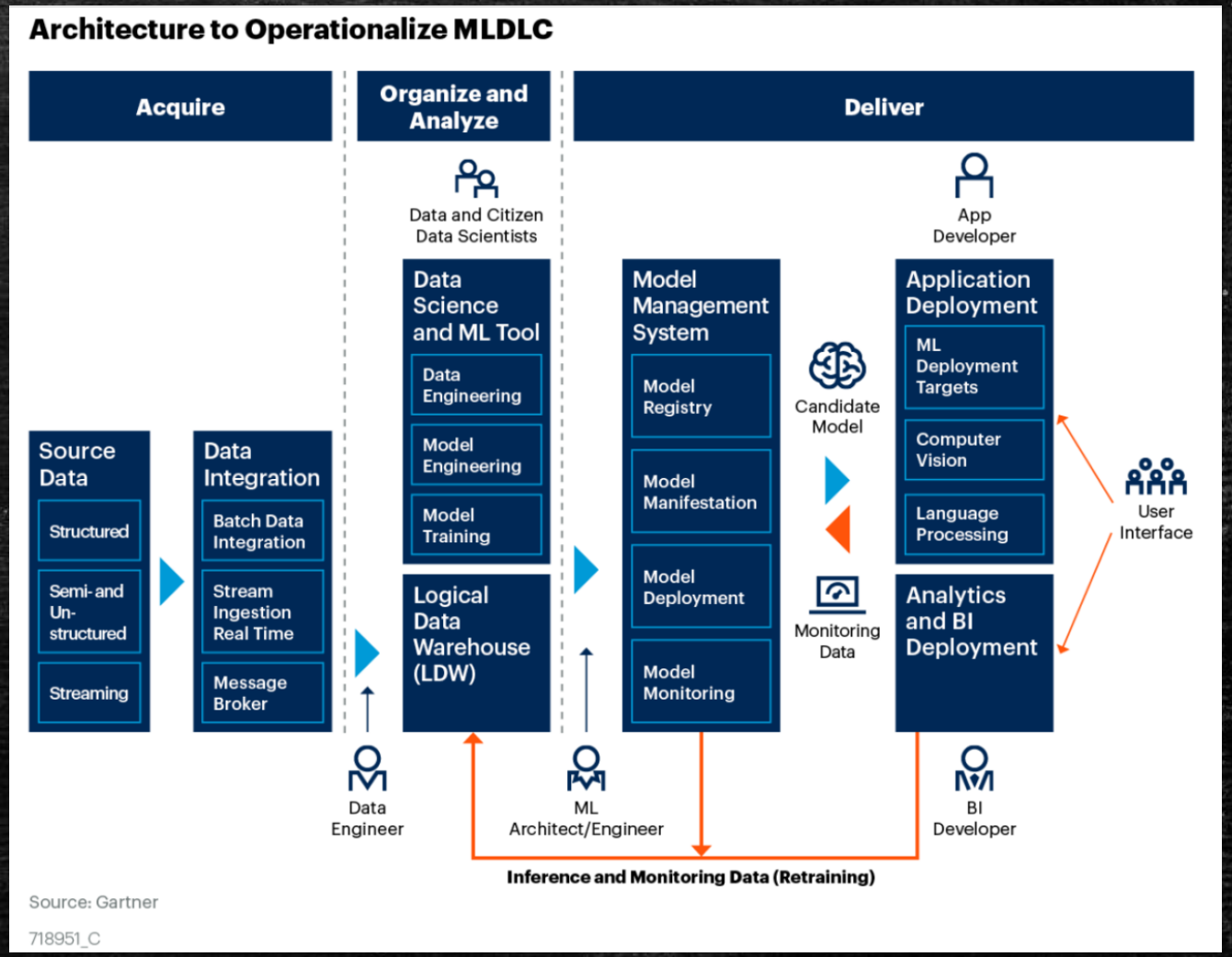
未來展望 – MLOps

https://blogs.nvidia.com/blog/what-is-mlops/

https://www.ml4devs.com/articles/mlops-machine-learning-life-cycle/

- 想必未來會有 MLSecOps





未來展望 - TensorFlow vs. PyTorch

- 如同前面所說的，PyTorch 還是大部分學界論文實作用的工具
- 現階段 AI 模型安全大多還在論文討論階段，所以學習 PyTorch 還是有其必要性

未來展望 – 圖像、語音、文字

- 雖然不同領域的攻擊理念大都相同，但是換個領域資料表示的概念其實相差很多
- 不只 DNN 的安全議題，像是 CNN、RNN 這樣的模型也會有自己對應的安全議題
- 像這樣跨領域的研究也是方向之一

未來展望 - ML vs. LLM

- 大型語言模型(LLM) 應該才是現在最夯的應用技術

- OWASP Top 10 for LLM Applications
- Mitre ATLAS
- NIST AI 100-2e2023

- 但 LLM 模型過於複雜，參數量龐大，攻擊重點還是在於 Prompt Injection，想要實際修改模型參數是很困難的

未來展望 – 進修課程

- 台大李宏毅教授
  - https://www.youtube.com/channel/UC2ggjtuuWvxrHHHiaDH1dlQ
- 台大陳尚澤教授 -機器學習安全特論
  - https://www.csie.ntu.edu.tw/~stchen/teaching/spml24spring/

結論

- 雖然 AI 模型安全似乎沒那麼普及，但隨著應用越來越多，想必之後應該也會成為駭客攻擊的對象
- 如果想要更深入研究的話，我覺得數學還是逃不掉的啦
- 這個系列只是幫忙起個頭而已，後面的路還遠的很!!!