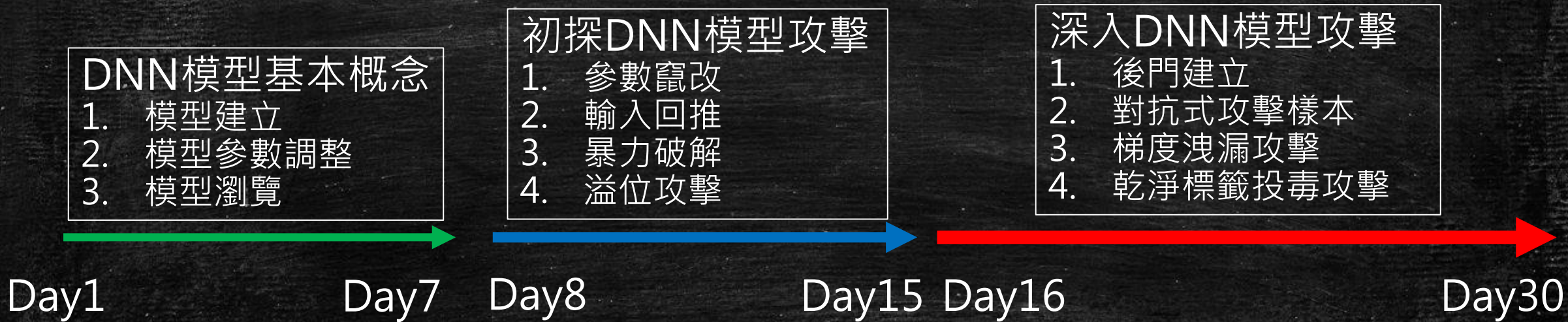
2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 30 – 完賽心得

大綱

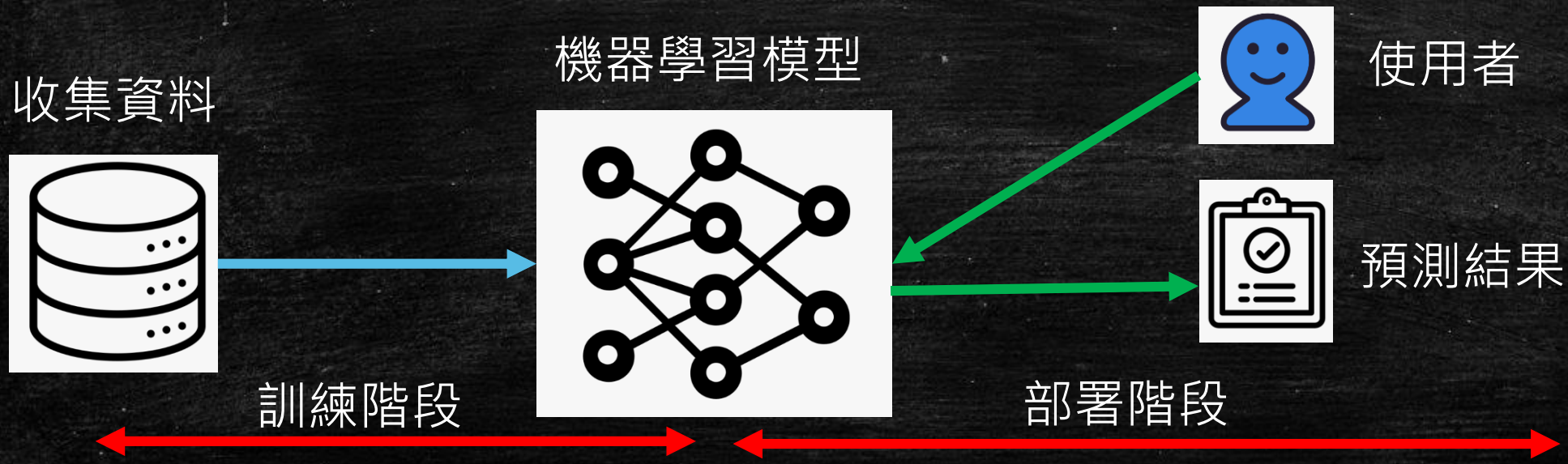
- DNN 模型安全
  - 系列賽回顧
  - 未來展望
- 結論

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
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
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
Day1
Day7
Day8
Day15
Day16
Day30

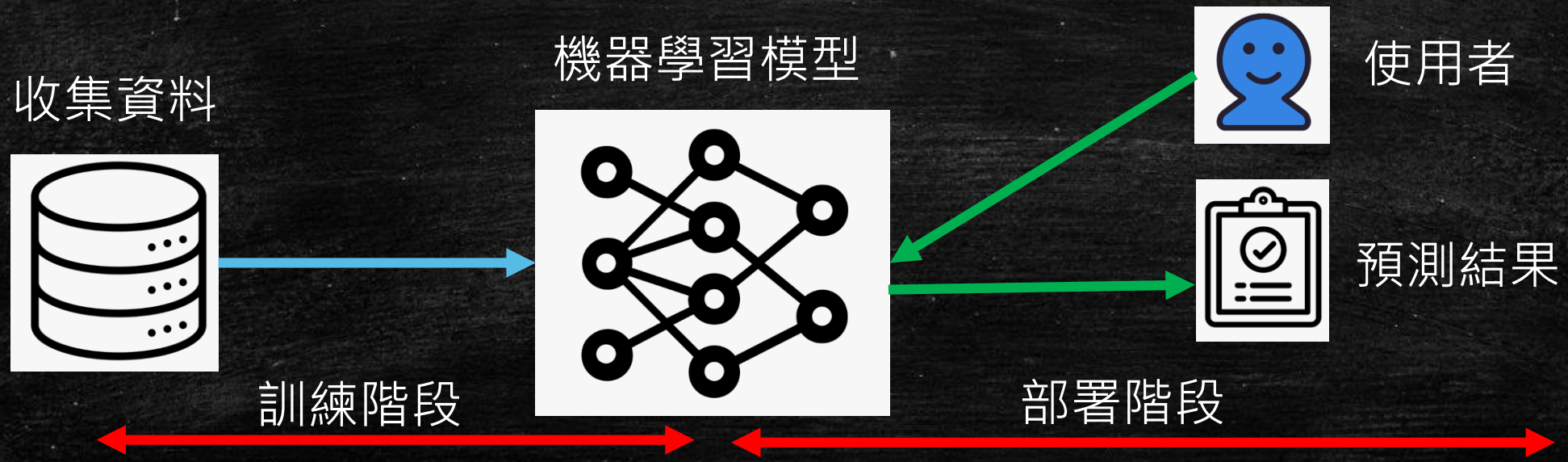
系列賽回顧

- 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

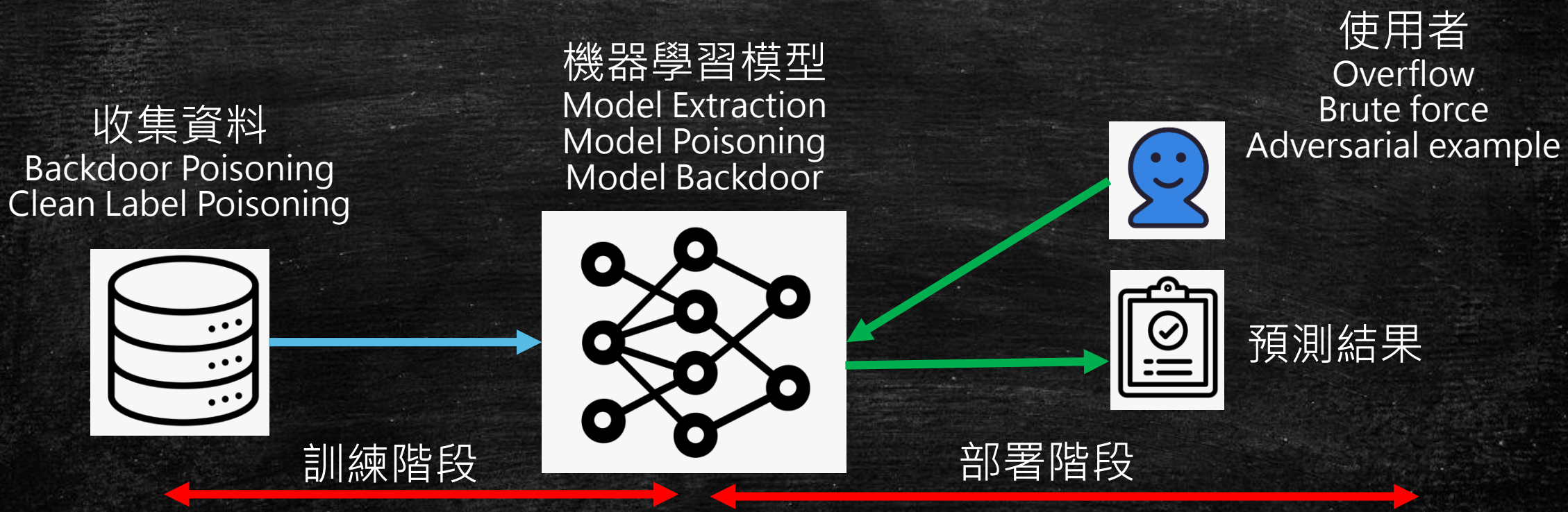


系列賽回顧

- 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
Backdoor Poisoning
Clean Label Poisoning
機器學習模型
Model Extraction
Model Poisoning
Model Backdoor
使用者
Overflow
Brute force
Adversarial example
預測結果
訓練階段
部署階段

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!!!