2024 鐵人賽 – 我數學就爛要怎麼來學 DNN 模型安全
Day 08 – DNN 模型安全概觀

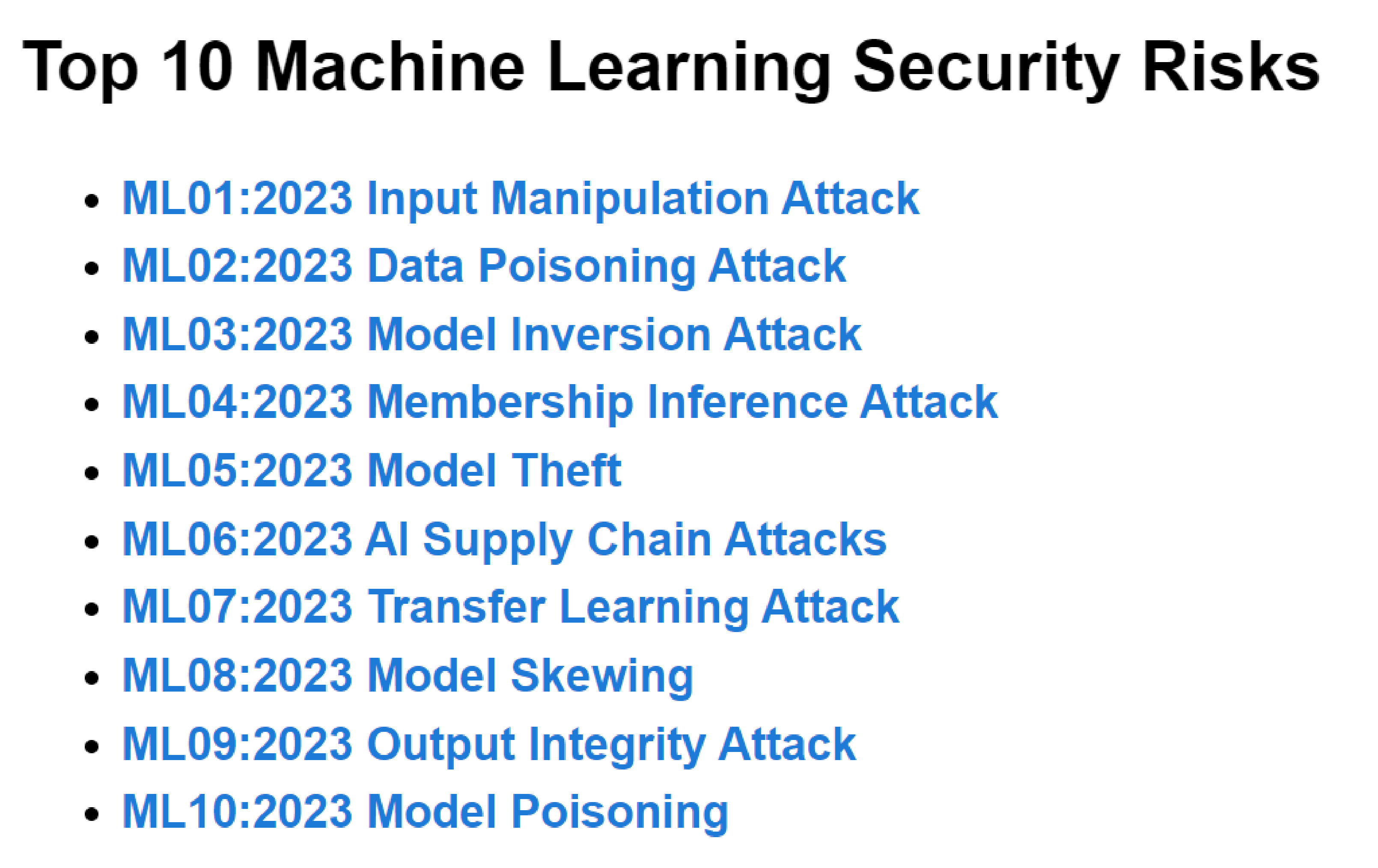
大綱

- DNN 模型安全概觀
  - 相關資料
  - 攻擊種類及發生時機
- 結論

相關資料

- OWASP Machine Learning Security Top Ten
- Mitre ATLAS Matrix
- NIST AI 100-2 E2023 - Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations

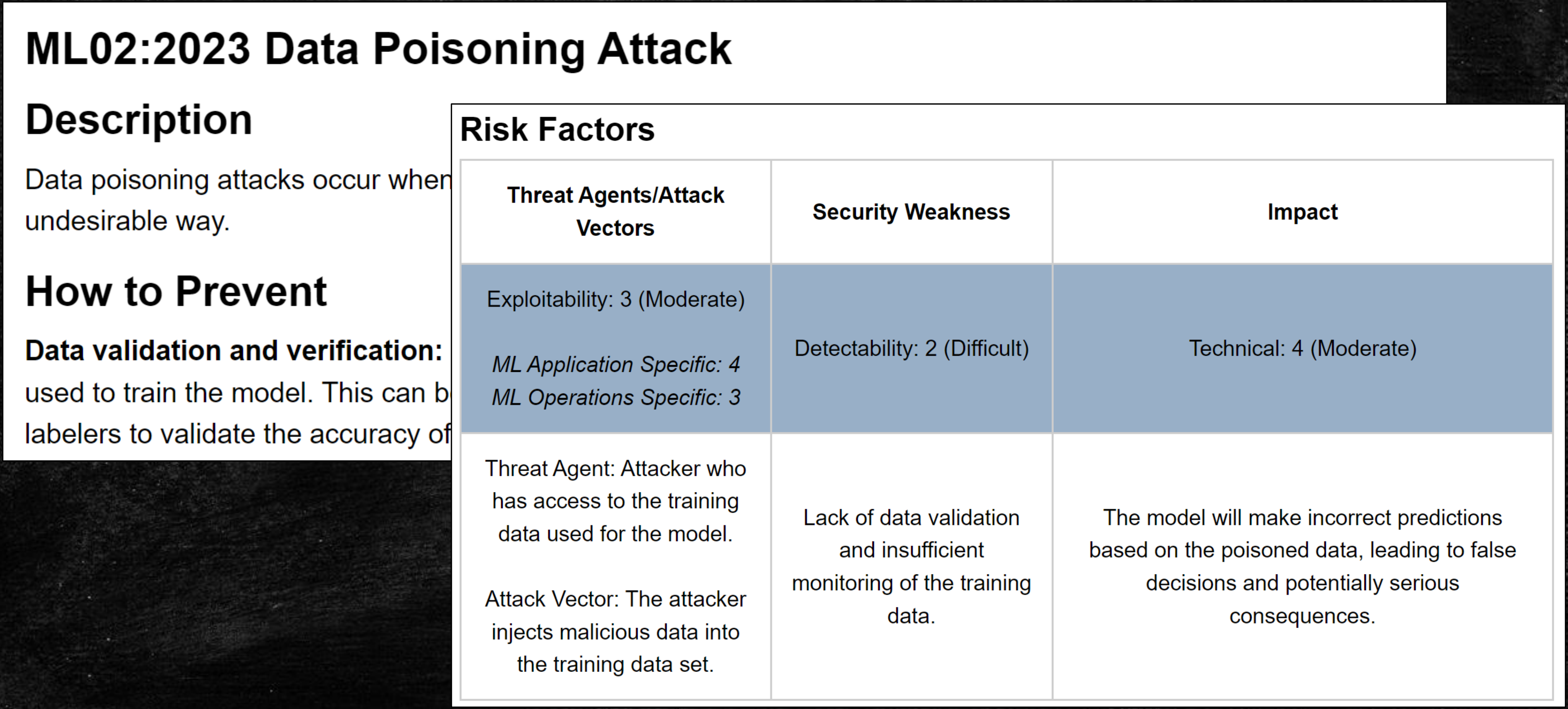
OWASP Machine Learning Security Top Ten



Top 10 Machine Learning Security Risks

- ML01:2023 Input Manipulation Attack
- ML02:2023 Data Poisoning Attack
- ML03:2023 Model Inversion Attack
- ML04:2023 Membership Inference Attack
- ML05:2023 Model Theft
- ML06:2023 AI Supply Chain Attacks
- ML07:2023 Transfer Learning Attack
- ML08:2023 Model Skewing
- ML09:2023 Output Integrity Attack
- ML10:2023 Model Poisoning

OWASP Machine Learning Security Top Ten



ML02:2023 Data Poisoning Attack

Description

Data poisoning attacks occur when
undesirable way.

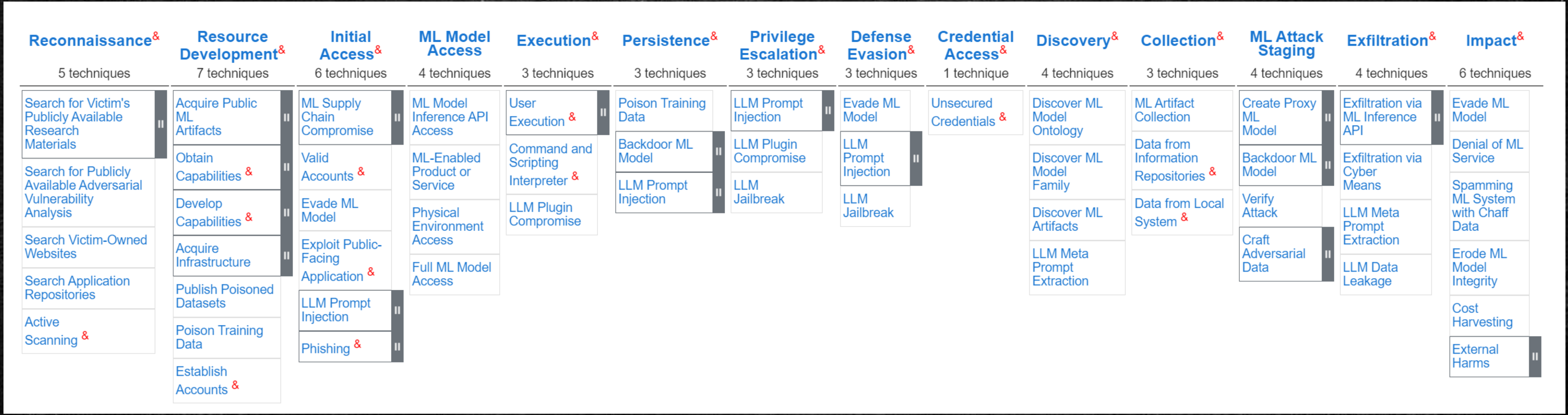
How to Prevent

Data validation and verification:
used to train the model. This can b
labelers to validate the accuracy of

Risk Factors

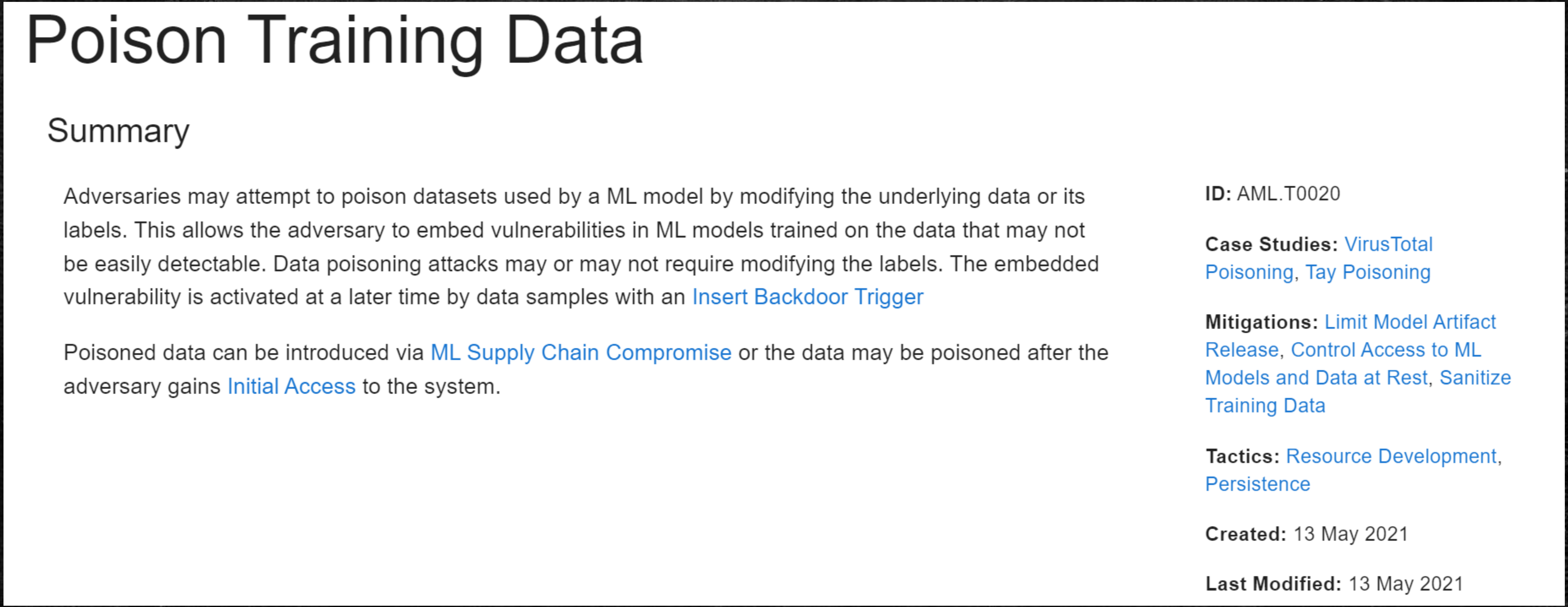
Threat Agents/Attack Vectors	Security Weakness	Impact
Exploitability: 3 (Moderate) *ML Application Specific: 4* *ML Operations Specific: 3*	Detectability: 2 (Difficult)	Technical: 4 (Moderate)
Threat Agent: Attacker who has access to the training data used for the model. Attack Vector: The attacker injects malicious data into the training data set.	Lack of data validation and insufficient monitoring of the training data.	The model will make incorrect predictions based on the poisoned data, leading to false decisions and potentially serious consequences.

Mitre ATLAS Matrix



Reconnaissance&	Resource Development&	Initial Access&	ML Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	ML Attack Staging	Exfiltration&	Impact&
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms
	Establish Accounts &												

Mitre ATLAS Matrix



Poison Training Data

Summary

Adversaries may attempt to poison datasets used by a ML model by modifying the underlying data or its labels. This allows the adversary to embed vulnerabilities in ML models trained on the data that may not be easily detectable. Data poisoning attacks may or may not require modifying the labels. The embedded vulnerability is activated at a later time by data samples with an Insert Backdoor Trigger

Poisoned data can be introduced via ML Supply Chain Compromise or the data may be poisoned after the adversary gains Initial Access to the system.

ID: AML.T0020

Case Studies: VirusTotal Poisoning, Tay Poisoning

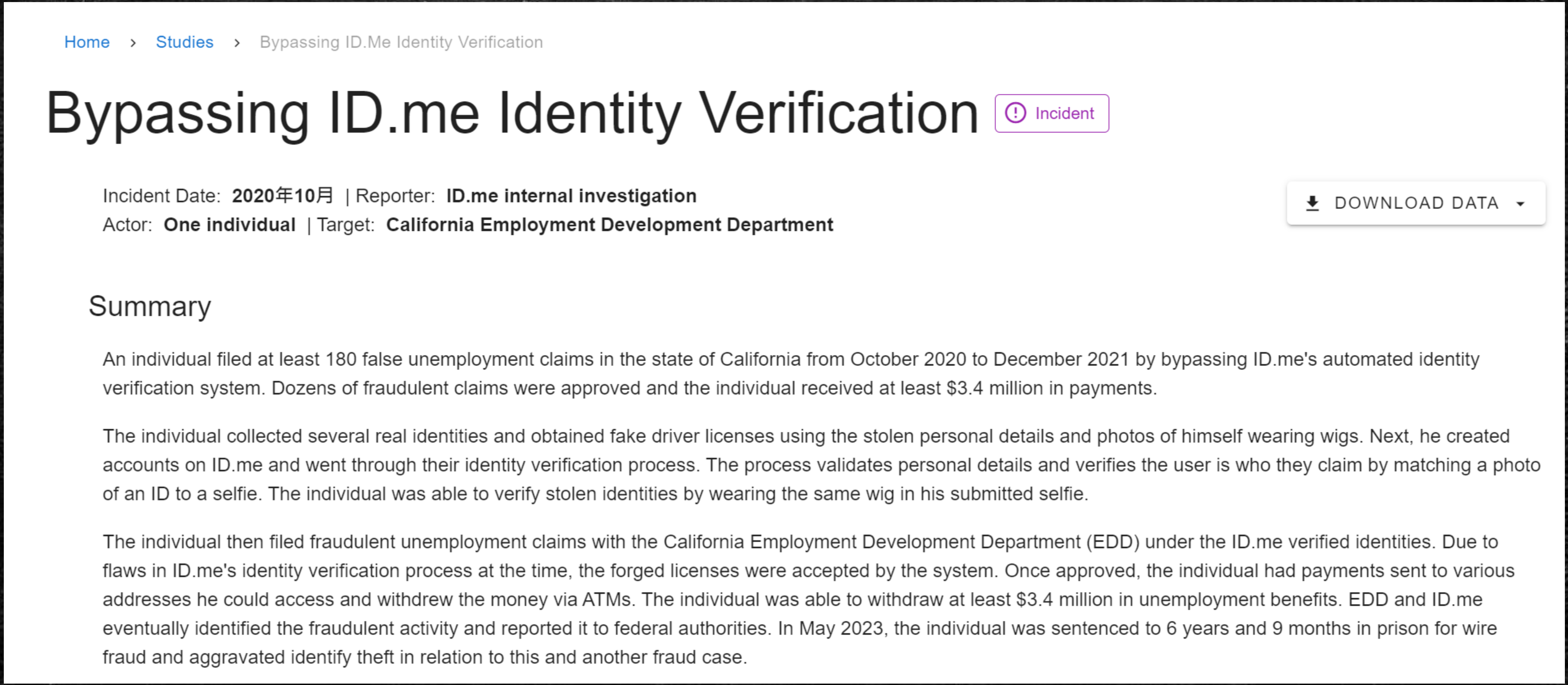
Mitigations: Limit Model Artifact Release, Control Access to ML Models and Data at Rest, Sanitize Training Data

Tactics: Resource Development, Persistence

Created: 13 May 2021

Last Modified: 13 May 2021

Mitre ATLAS Matrix – Case Study



Bypassing ID.me Identity Verification

Incident Date: **2020年10月** | Reporter: **ID.me internal investigation**
Actor: **One individual** | Target: **California Employment Development Department**

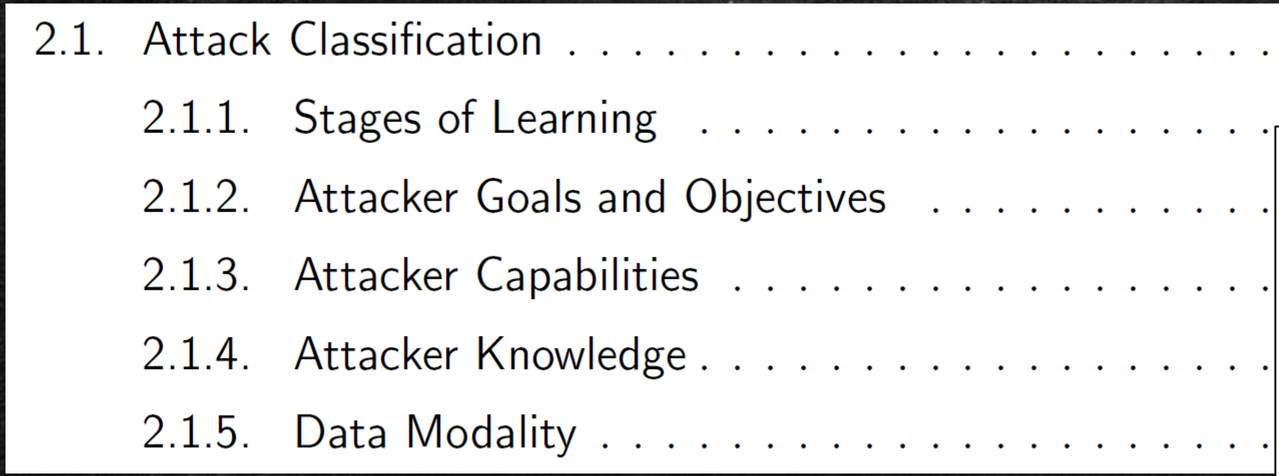
Summary

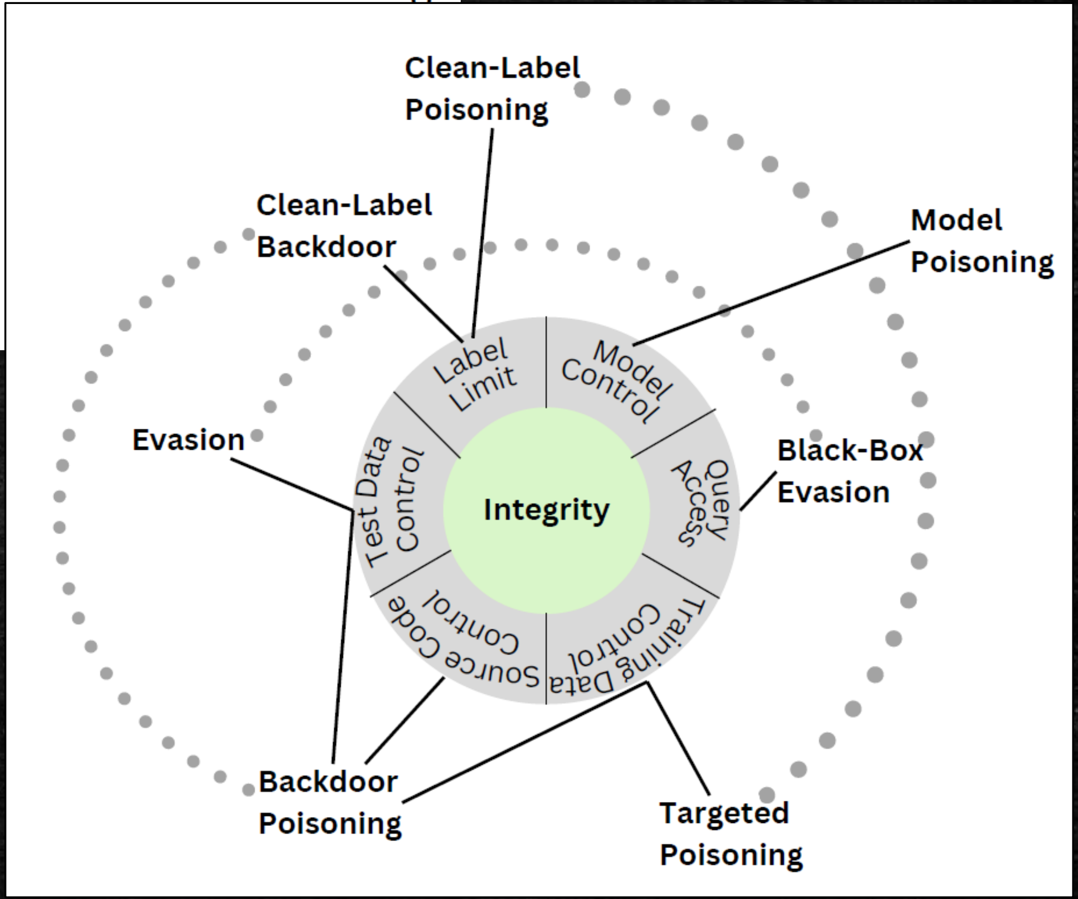
An individual filed at least 180 false unemployment claims in the state of California from October 2020 to December 2021 by bypassing ID.me's automated identity verification system. Dozens of fraudulent claims were approved and the individual received at least $3.4 million in payments.

The individual collected several real identities and obtained fake driver licenses using the stolen personal details and photos of himself wearing wigs. Next, he created accounts on ID.me and went through their identity verification process. The process validates personal details and verifies the user is who they claim by matching a photo of an ID to a selfie. The individual was able to verify stolen identities by wearing the same wig in his submitted selfie.

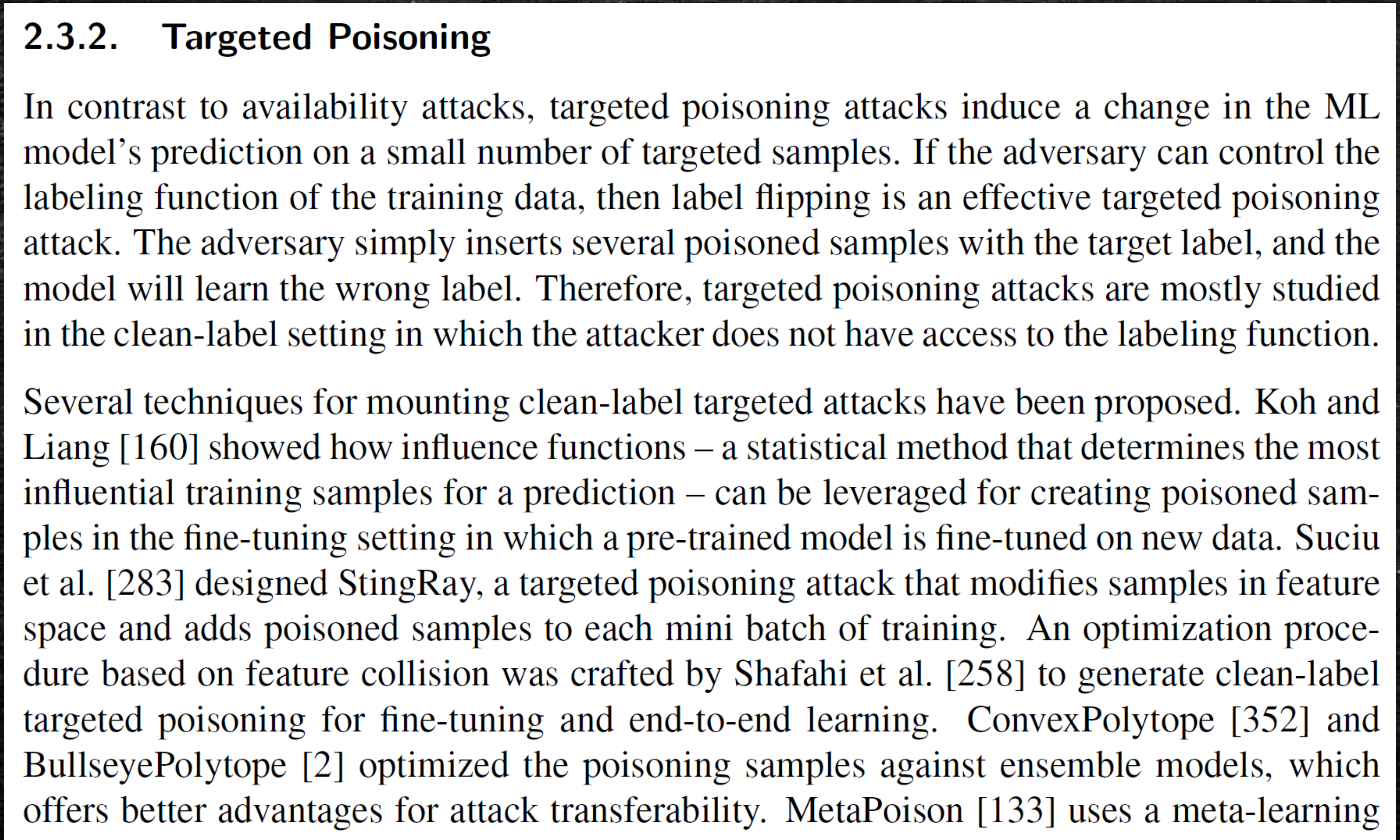
The individual then filed fraudulent unemployment claims with the California Employment Development Department (EDD) under the ID.me verified identities. Due to flaws in ID.me's identity verification process at the time, the forged licenses were accepted by the system. Once approved, the individual had payments sent to various addresses he could access and withdrew the money via ATMs. The individual was able to withdraw at least $3.4 million in unemployment benefits. EDD and ID.me eventually identified the fraudulent activity and reported it to federal authorities. In May 2023, the individual was sentenced to 6 years and 9 months in prison for wire fraud and aggravated identify theft in relation to this and another fraud case.

NIST AI 100-2 E2023





NIST AI 100-2 E2023



2.3.2. Targeted Poisoning

In contrast to availability attacks, targeted poisoning attacks induce a change in the ML model's prediction on a small number of targeted samples. If the adversary can control the labeling function of the training data, then label flipping is an effective targeted poisoning attack. The adversary simply inserts several poisoned samples with the target label, and the model will learn the wrong label. Therefore, targeted poisoning attacks are mostly studied in the clean-label setting in which the attacker does not have access to the labeling function.

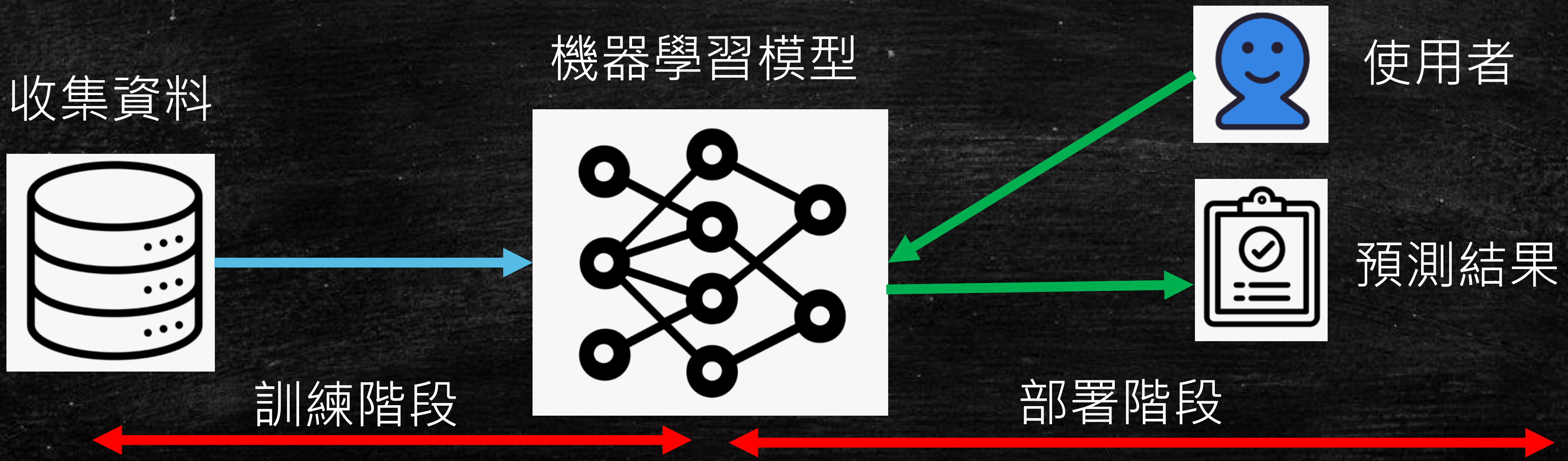
Several techniques for mounting clean-label targeted attacks have been proposed. Koh and Liang [160] showed how influence functions – a statistical method that determines the most influential training samples for a prediction – can be leveraged for creating poisoned samples in the fine-tuning setting in which a pre-trained model is fine-tuned on new data. Suciu et al. [283] designed StingRay, a targeted poisoning attack that modifies samples in feature space and adds poisoned samples to each mini batch of training. An optimization procedure based on feature collision was crafted by Shafahi et al. [258] to generate clean-label targeted poisoning for fine-tuning and end-to-end learning. ConvexPolytope [352] and BullseyePolytope [2] optimized the poisoning samples against ensemble models, which offers better advantages for attack transferability. MetaPoison [133] uses a meta-learning

來個比較

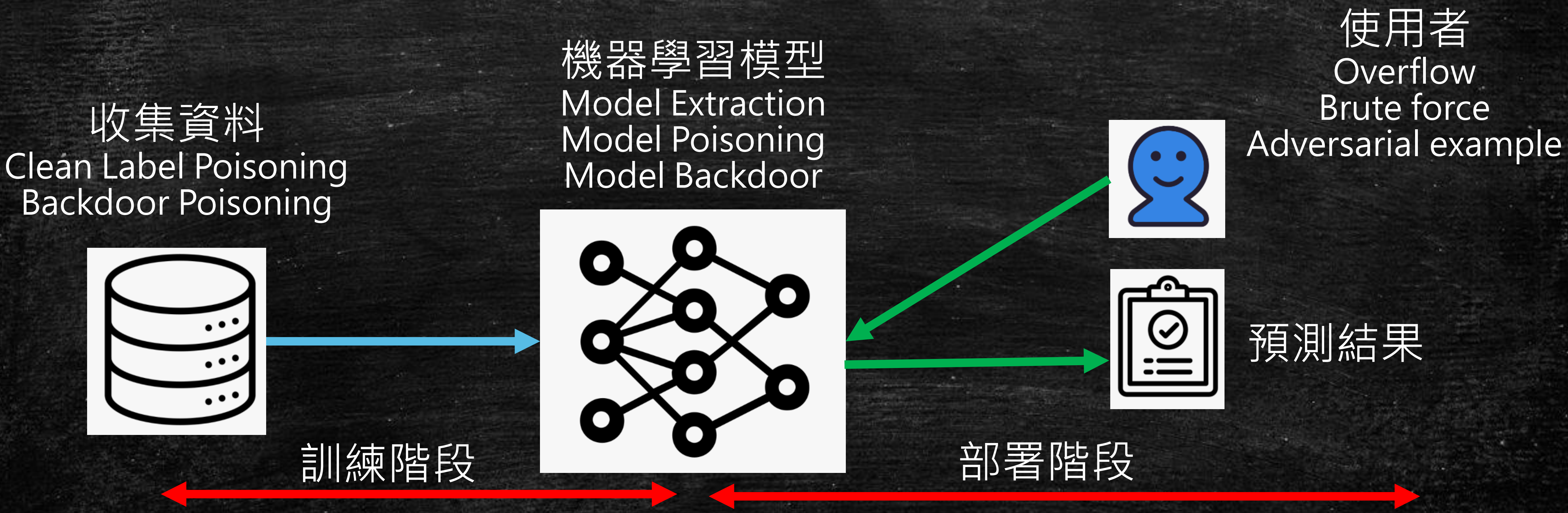
	OWASP Top Ten	Mitre ATLAS	NIST AI 100-2
重點	前十大攻擊重點	各滲透測試階段的攻擊手法	完整的分類及參考文獻
優點	Risk Factor 簡潔扼要的點出攻擊方式跟路徑	Case Studies 的部分整理的不錯 包含時間序跟手法	分類跟引用攻擊手法的部分較佳
缺點	缺乏技術細節 分類上較為粗糙	缺乏技術細節	不夠親民且不易閱讀

來討論一下攻擊分類

- 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



本次系列的攻擊整理



使用者
Overflow
Brute force
Adversarial example

機器學習模型
Model Extraction
Model Poisoning
Model Backdoor

收集資料
Clean Label Poisoning
Backdoor Poisoning

預測結果

訓練階段

部署階段

結論

- 大概瀏覽一下機器學習會遭遇的安全威脅，看起來沒想像中的安全，也沒有想像中那麼不安全

- 之後會開始陸續介紹攻擊手法，在了解攻擊手法跟攻擊面之後才有機會討論防禦、偵測的相關機制