”

What is Data Poisoning Attacks?

- Mohammad Arif

HTTPS://ARIF-PLAYBOOK.GITBOOK.IO/

HTTPS://WWW.LINKEDIN.COM/IN/MOHD--ARIF/

When threat actors manipulate or corrupt the training data used to develop artificial intelligence (AI) and machine learning (ML) models.

- Mohammad Arif

HTTPS://ARIF-PLAYBOOK.GITBOOK.IO/

HTTPS://WWW.LINKEDIN.COM/IN/MOHD--ARIF/

Day 24 — *Data Poisoning Attack*

Break the model

Subvert specific behaviour

Backdoor the model silently

Leak private data during inference

Mohammad Arif
www.linkedin.com/in/mohd--arif/

#100DaysOfAISec

Day 24 *Data Poisoning Attack*

🧪 Real-World Examples

Microsoft Tay — poisoned by malicious tweets started making offensive remarks

Google Perspective — adversarial users injected toxic-but-acceptable phrases

LLM Alignment datasets — found to contain biased/misleading training prompts

Mohammad Arif
www.linkedin.com/in/mohd--arif/

#100DaysOfAISec

Day 24 *Data Poisoning Attack*

🎯 Attacker's Motive

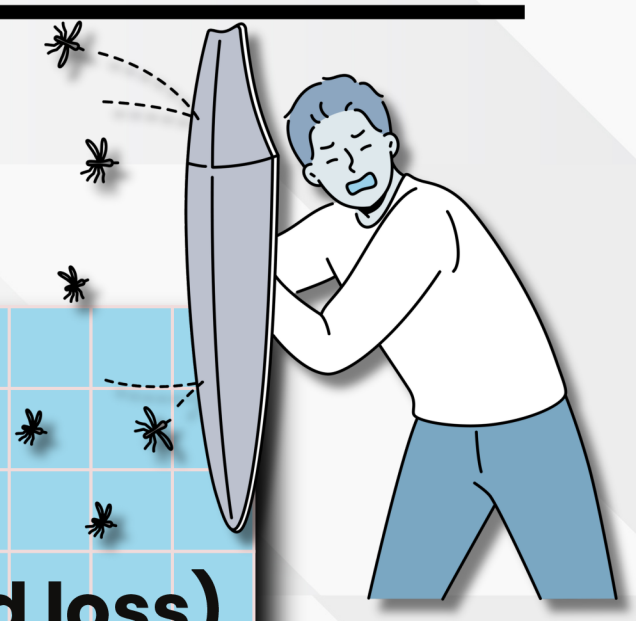
Different motives, same danger:

- 💡 **Sabotage a system's accuracy or availability**
- 💡 **Create secret triggers that only the attacker knows**
- 💡 **Bypass security filters like spam or malware detection**
- 💡 **Insert logic bombs triggered in production**
- 💡 **Extract private information from training data**
- 💡 **Manipulate AI behavior in social, political, or economic contexts**

Mohammad Arif
www.linkedin.com/in/mohd--arif/

#100DaysOfAISec

Day 24 — *Data Poisoning Attack*



Defence

- **Use robust training (e.g., differential privacy, trimmed loss)**
- **Audit your data pipeline — especially crowdsourced/third-party**
- **Monitor data provenance and reputation**
- **Use outlier detection and deduplication**

Mohammad Arif
www.linkedin.com/in/mohd--arif/

#100DaysOfAISec