UNIVERSITY of HOUSTON

CULLEN COLLEGE of ENGINEERING

Attacks in Neural Networks from Hardware Perspective

By: Alan Devkota

Advisor: Dr. Xin Fu

ECE 6011_25383

February 10, 2023

Department of ECE, University of Houston

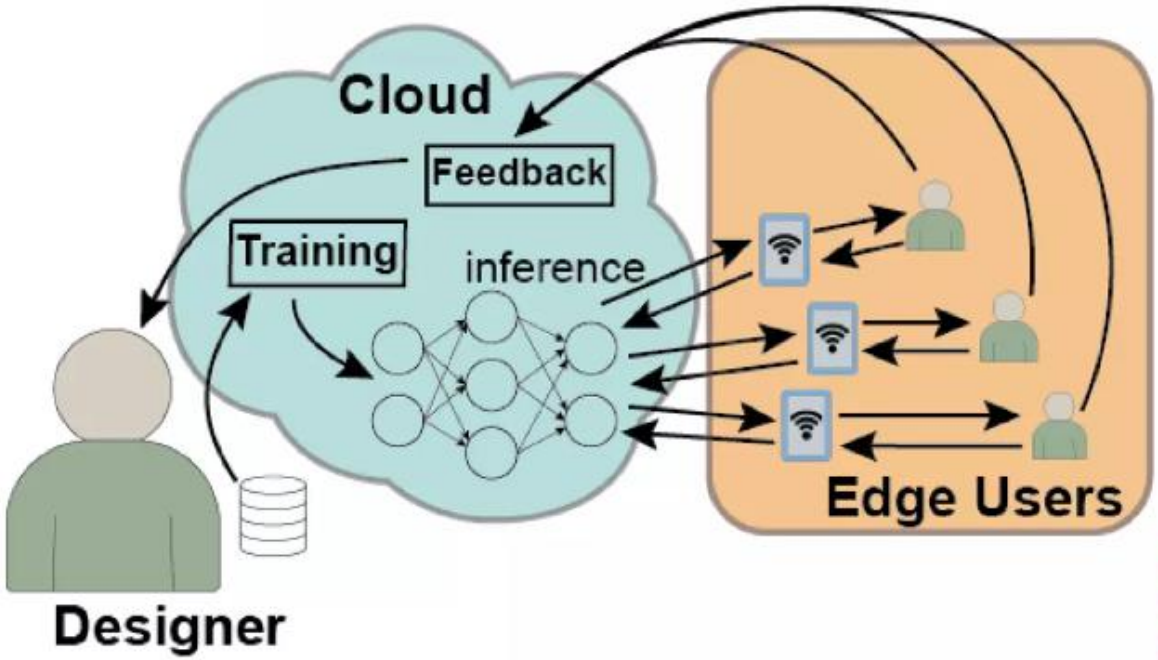
Houston, TX, USA

Outline

- Overview of Neural Networks.
- Overview of attacks.
- Adversarial scenarios in cloud and edge.
- Hardware Trojan Attack on Neural Networks.
- Illustration in memory.
- Conclusion and future direction.

Overview of Neural Networks

- Cloud Based ML Paradigms



Overview of Neural Networks

- Constrained Applications

Security Systems

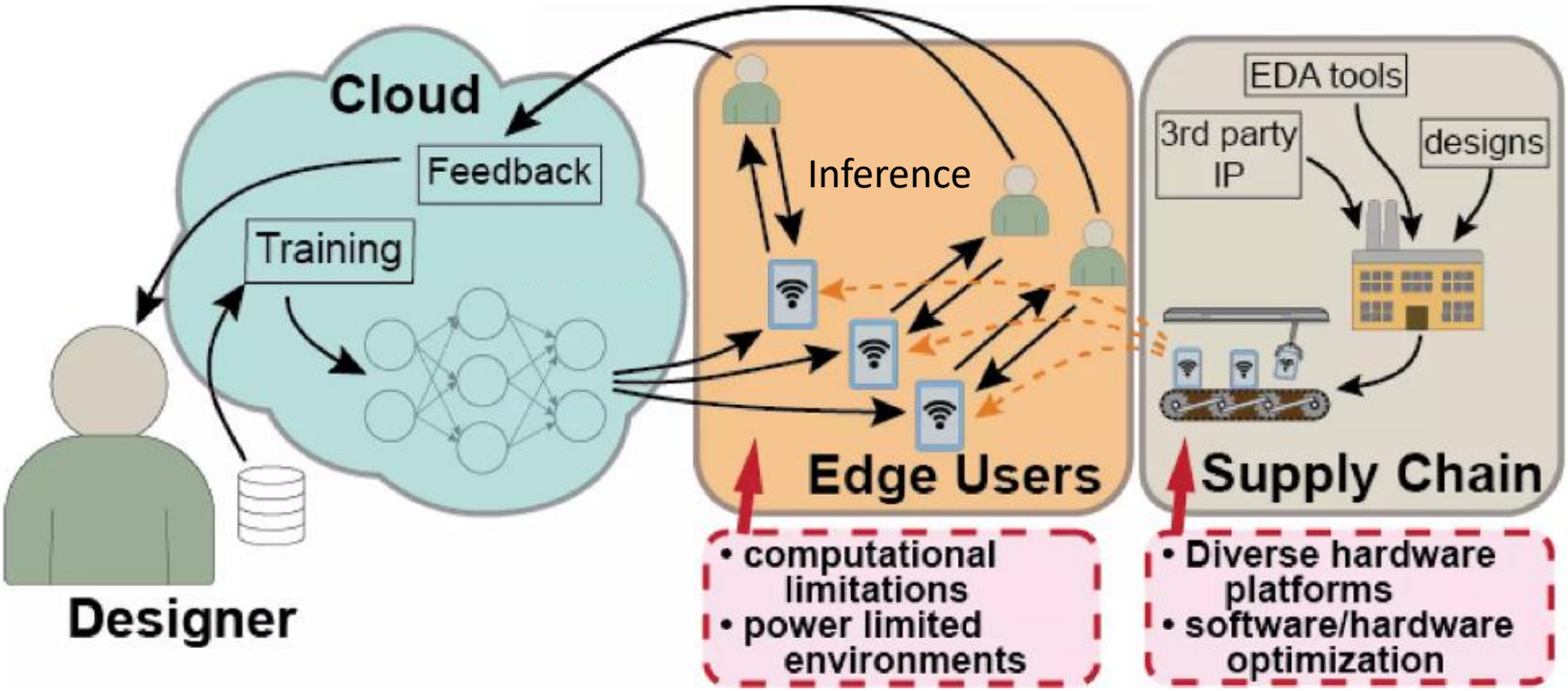
Mobile Applications

Automatic Driving

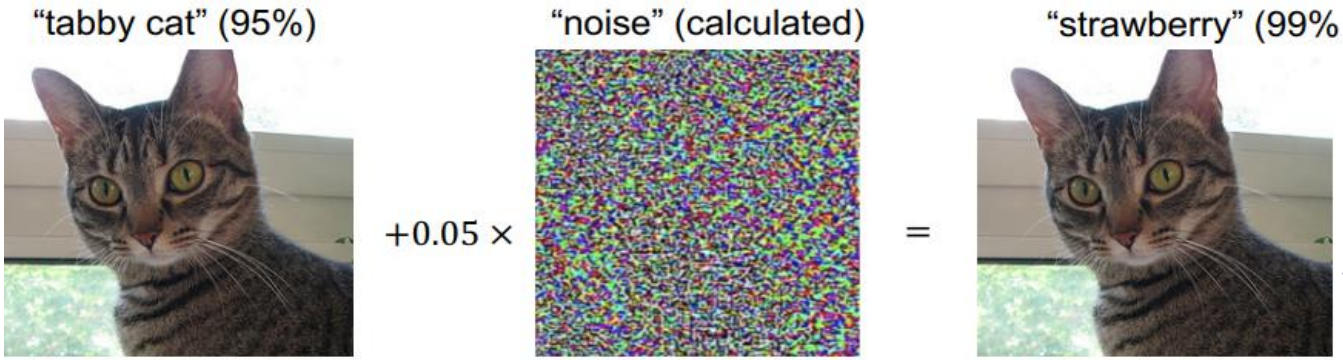
Wearable Technologies

Overview of Neural Networks

- Moving inference to the edge



Overview of Attacks on Neural Networks



Prediction (Adversarial example)

labels: dog, cat, mango, strawberry and so on.

Solution: Adversarial training

- Training Data (Poisoning)

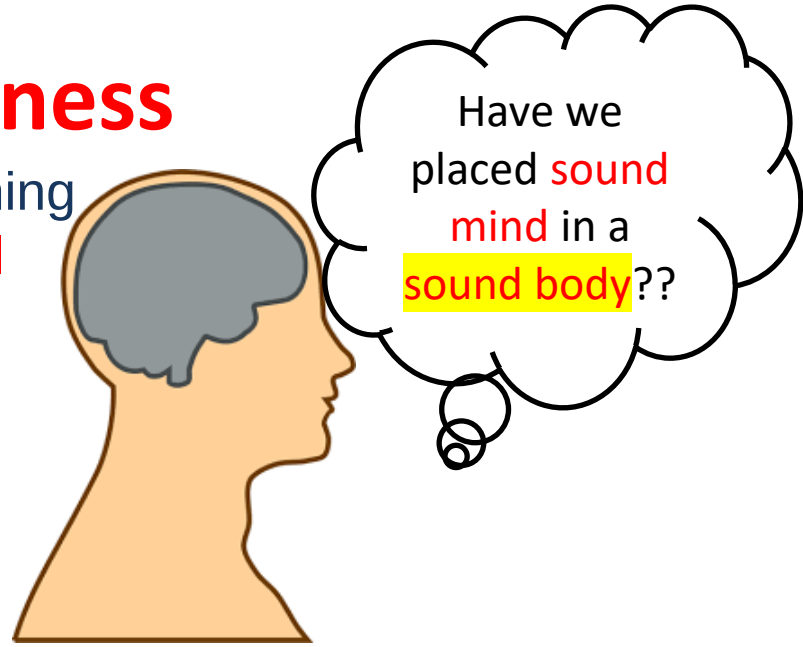
Tay – a twitter bot developed by Microsoft seemed to learn some bad behavior on its own

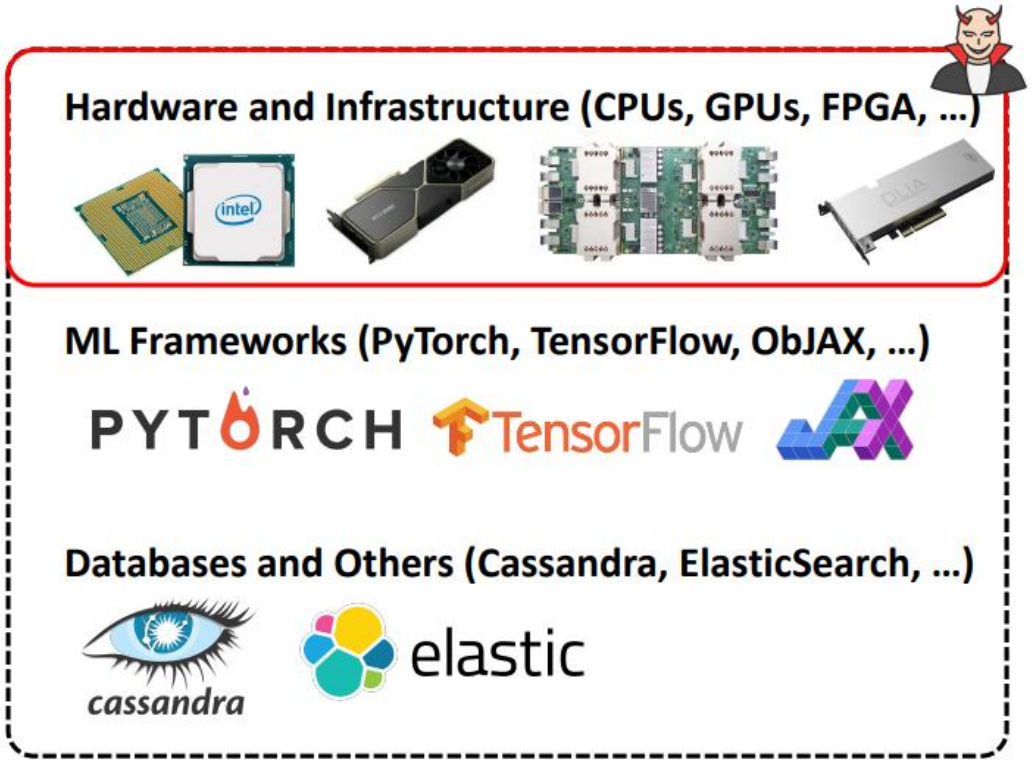
Solution: Data sanitization/ Robust statistics

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

DNN Robustness

- Prior works considered Machine Learning models as a standalone, mathematical concept.
- We need to consider hardware level vulnerabilities as well.





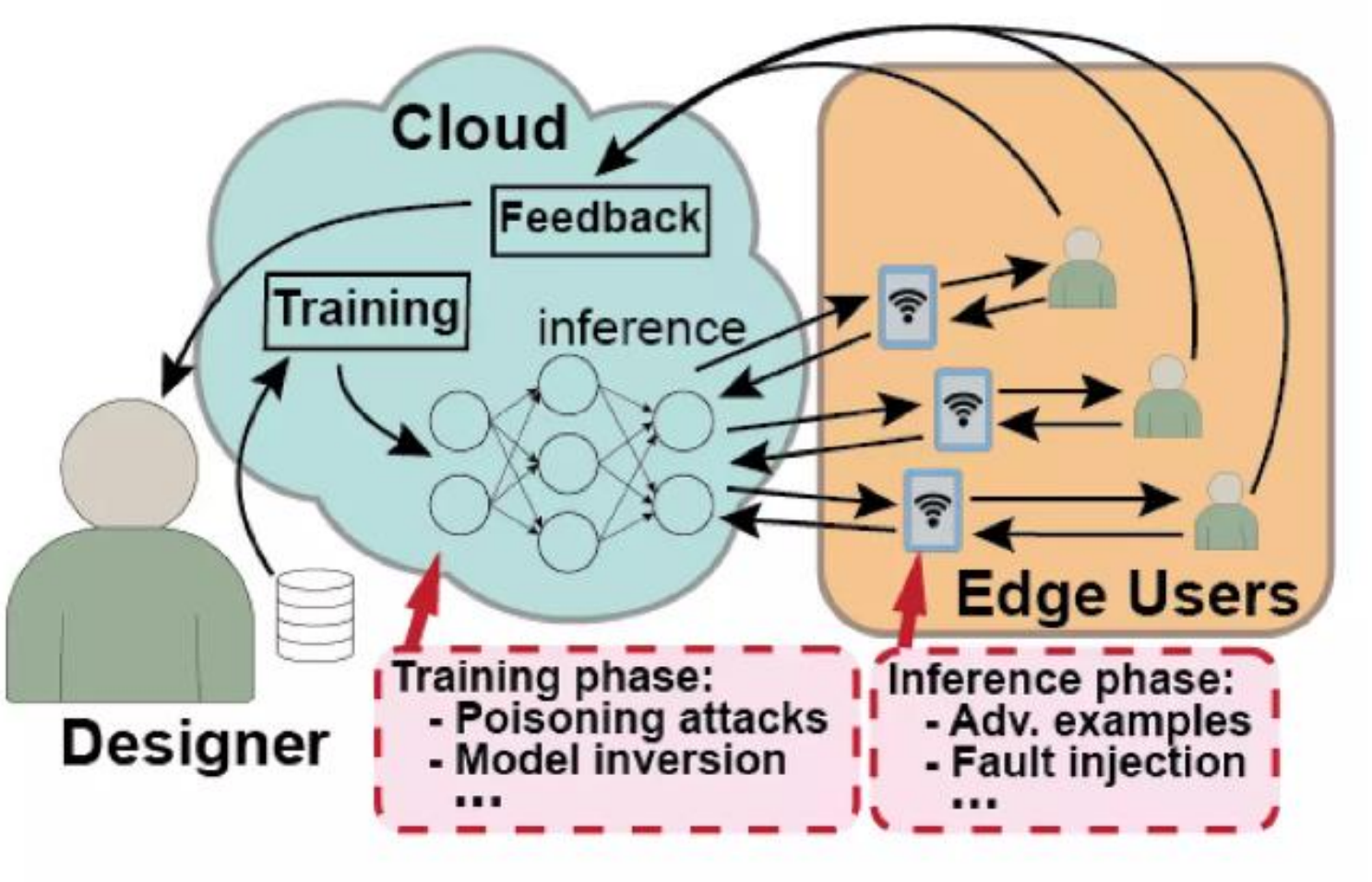
- Prior works considered: Hardware attack as weak attack and software attack as strong attack.

Not always

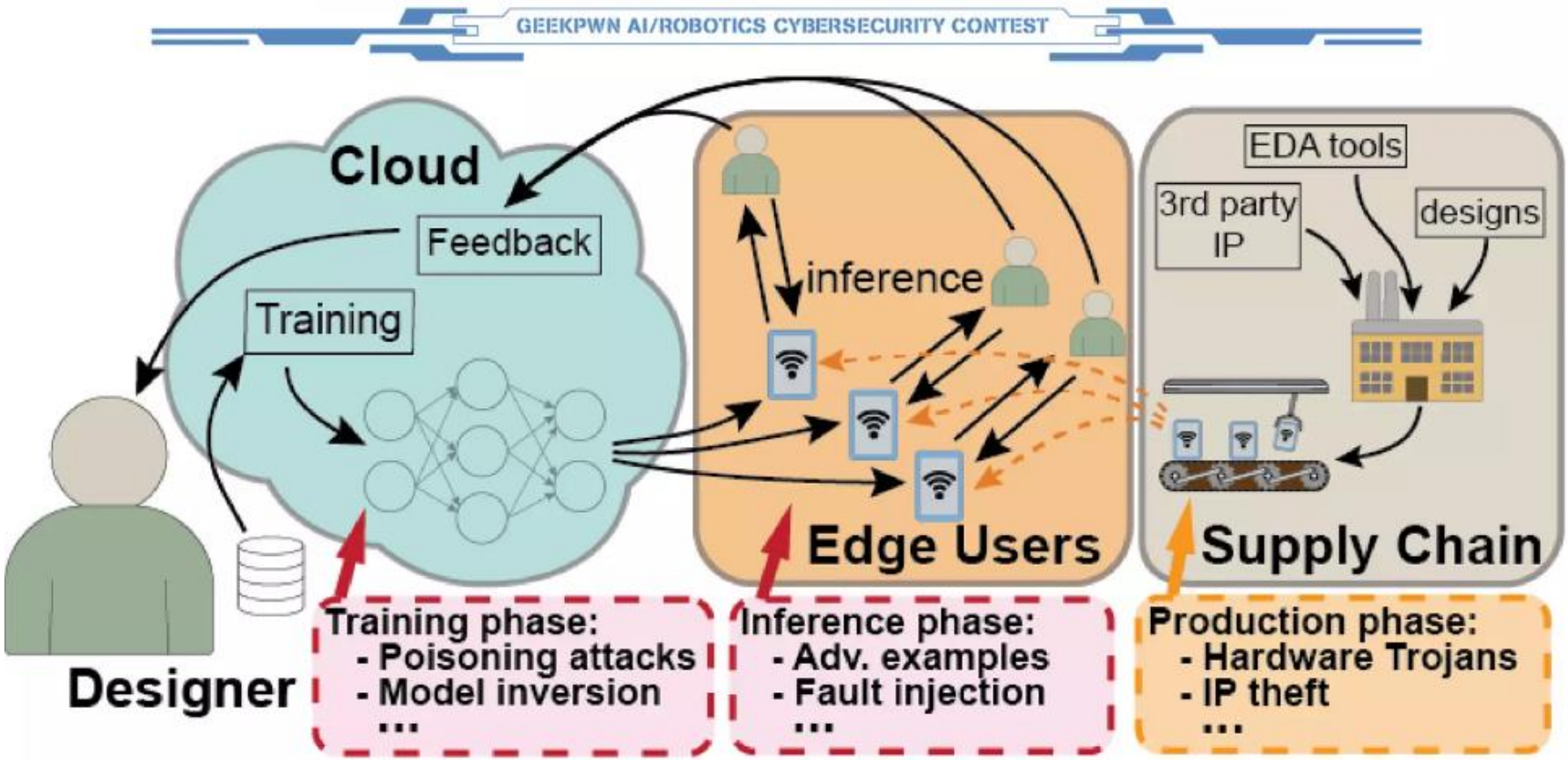
Security in Neural Network need both safe hardware and safe software

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

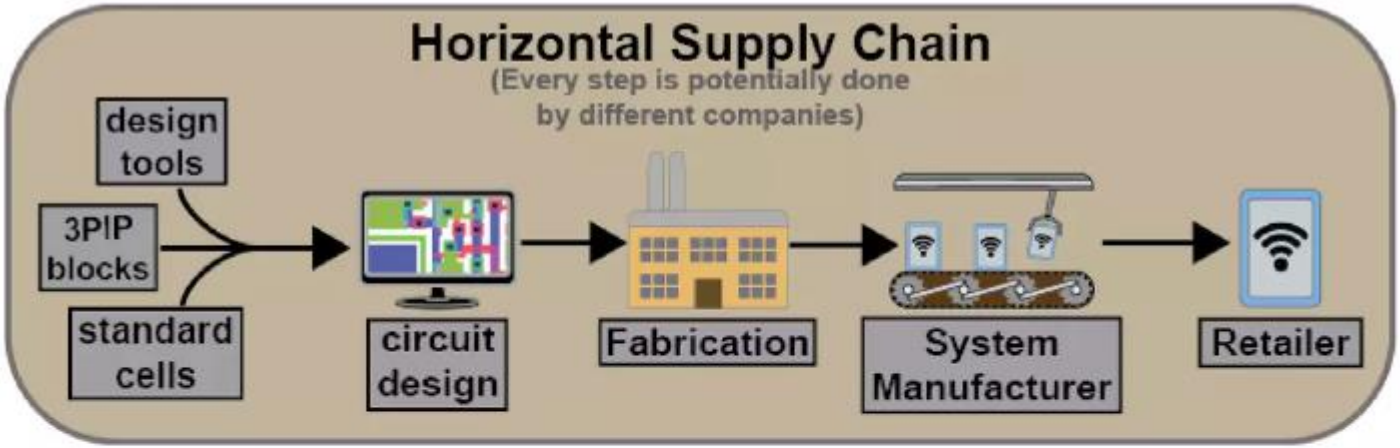
Adversarial Scenario in the Cloud

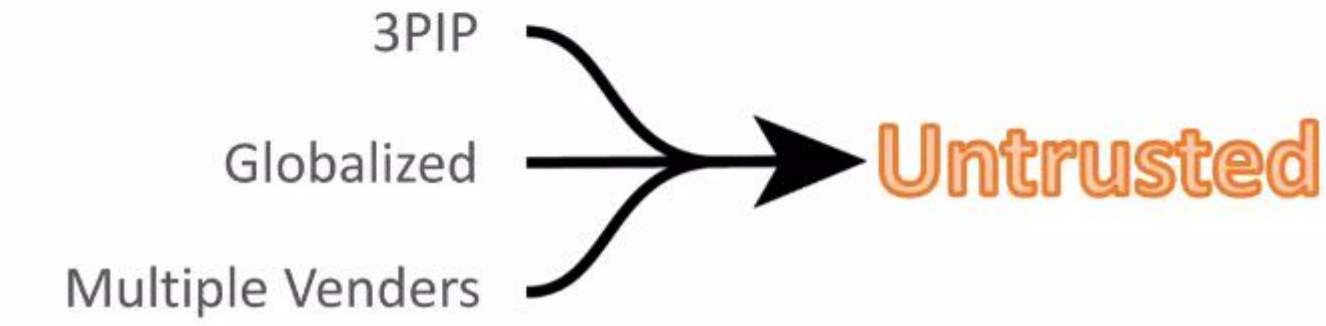


Adversarial Scenario on the Edge



Hardware supply chain





UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Attacks in hardware domain

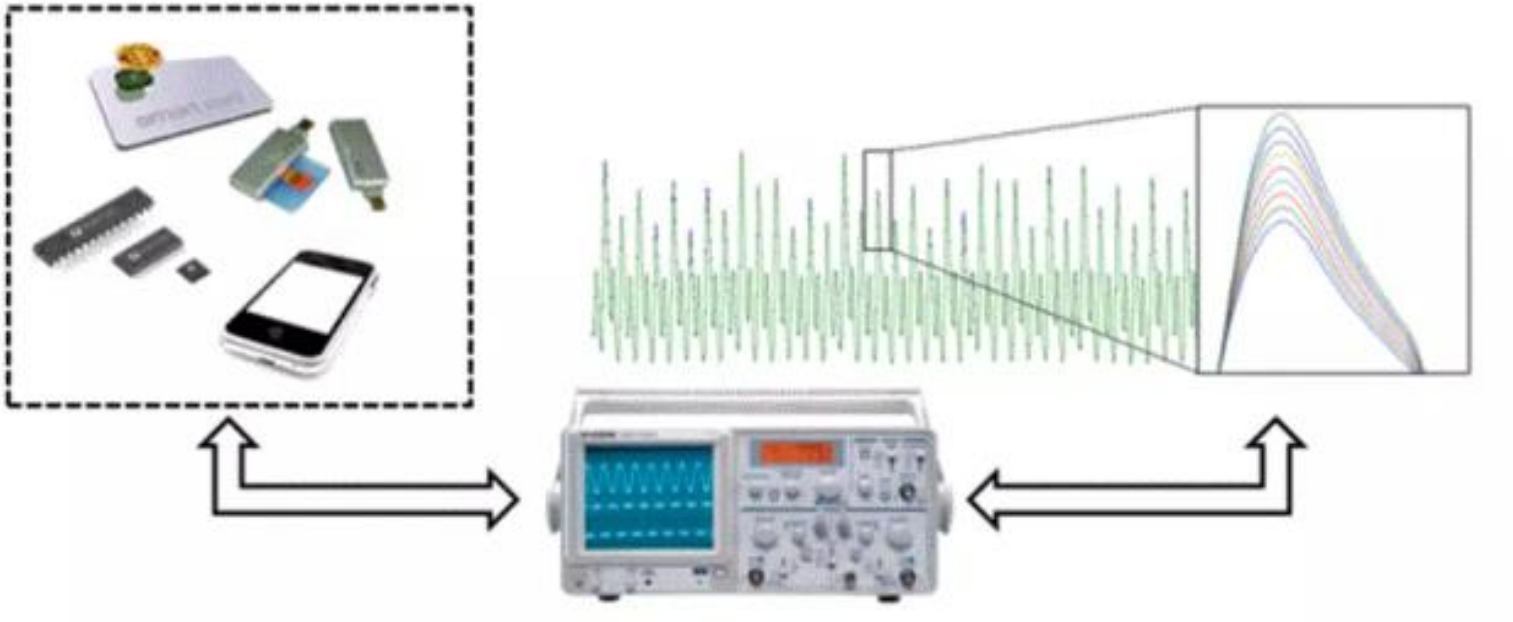
- **IP Piracy**: Produce IPs (or secretly more copies) without approval from original owner and provide them at low cost
- **Counterfeiting:** Generating a fake one. (Especially ICs)

Original

Fake

Attacks in hardware domain (contd.)

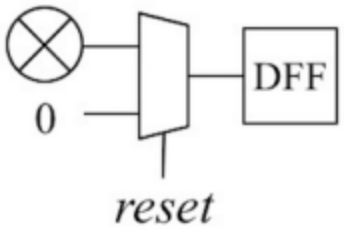
- Side-channel attacks: Exploit information from computer such as electromagnetic radiation.



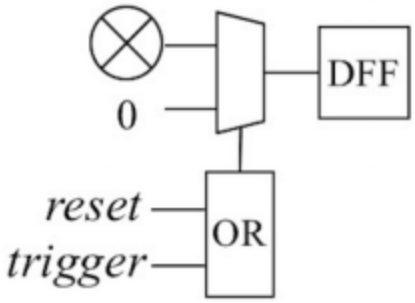
Attacks in hardware domain (contd.)

- Hardware trojans: Attacker attempts to maliciously modify a circuit design such that the functionality changes.

(Especially, if attacker have access to supply chain)



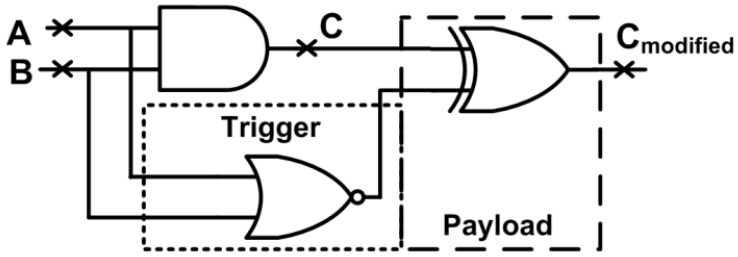
(a)



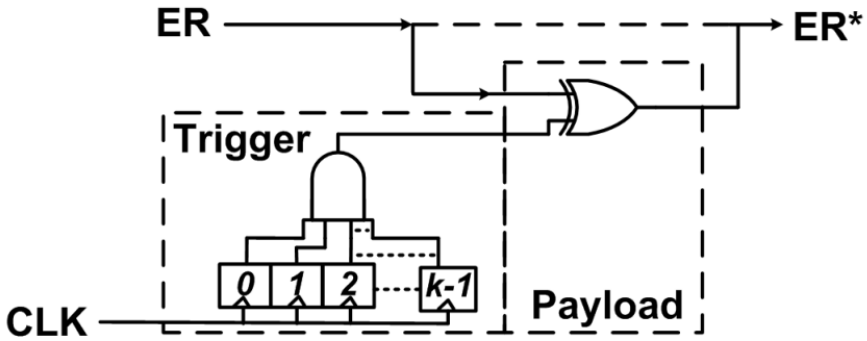
(b)

original

Trojaned



(a) Combinationally triggered Trojan



(b) Synchronous counter ("time-bomb") Trojan

Illustration of DNN (In memory Representation)

- Accuracy: 99%

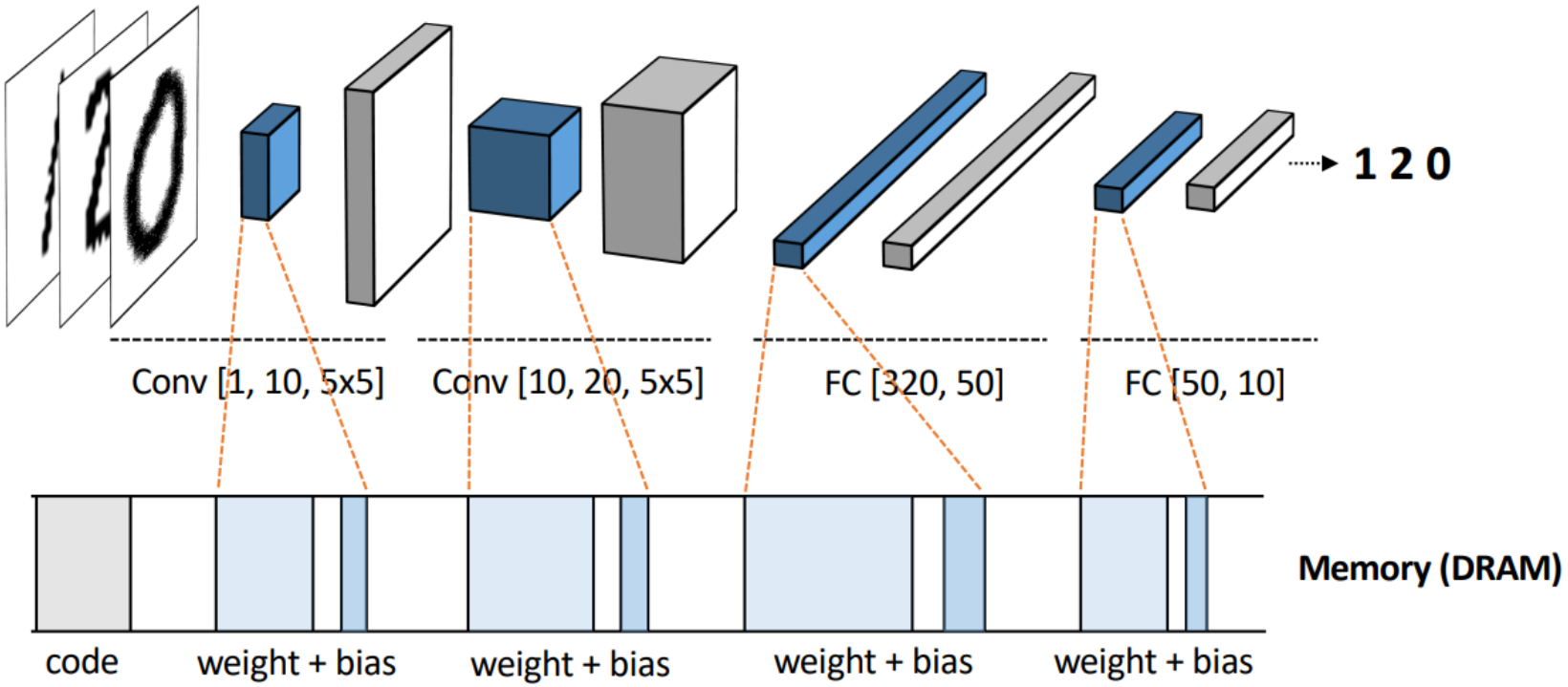


Illustration of DNN (In memory Representation)

- Accuracy: **93.53% (5% drop)**

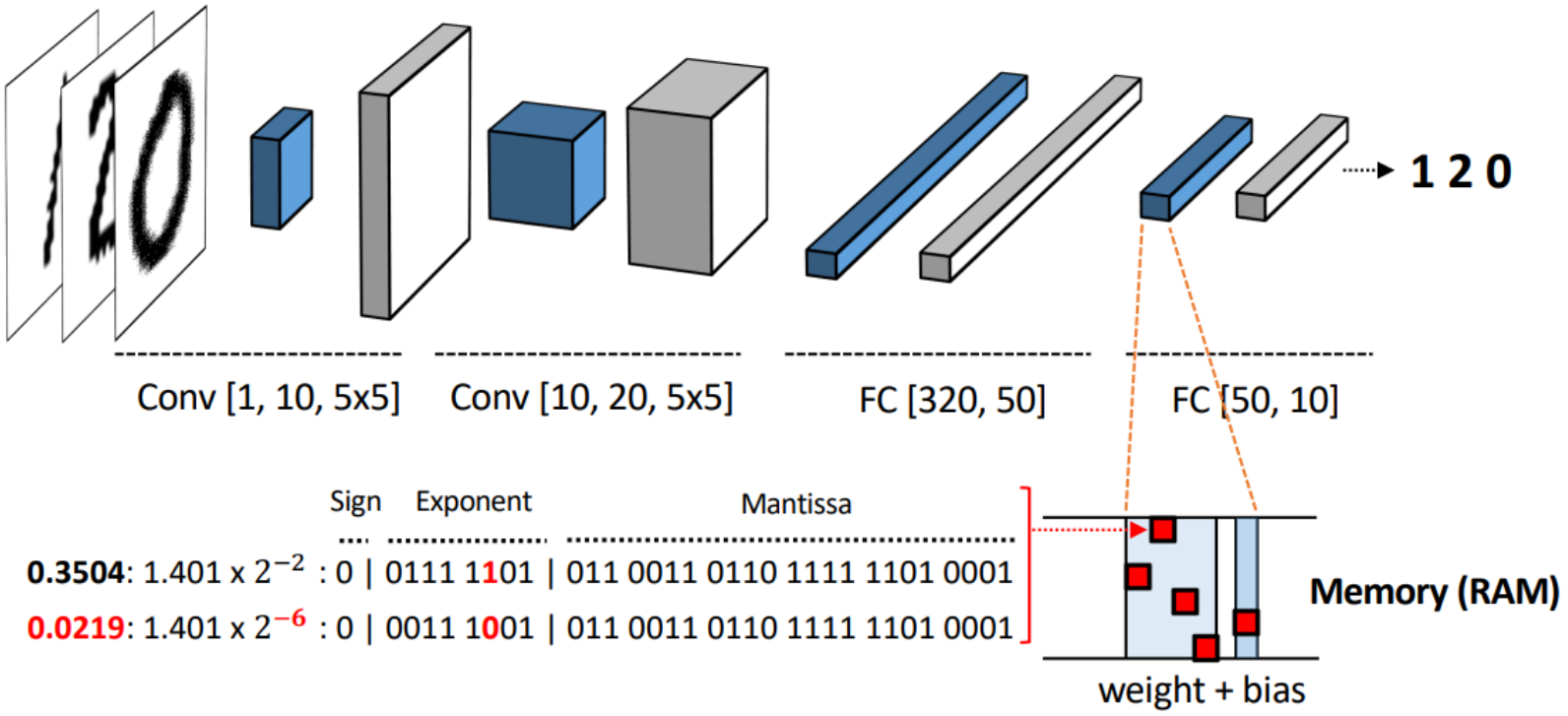
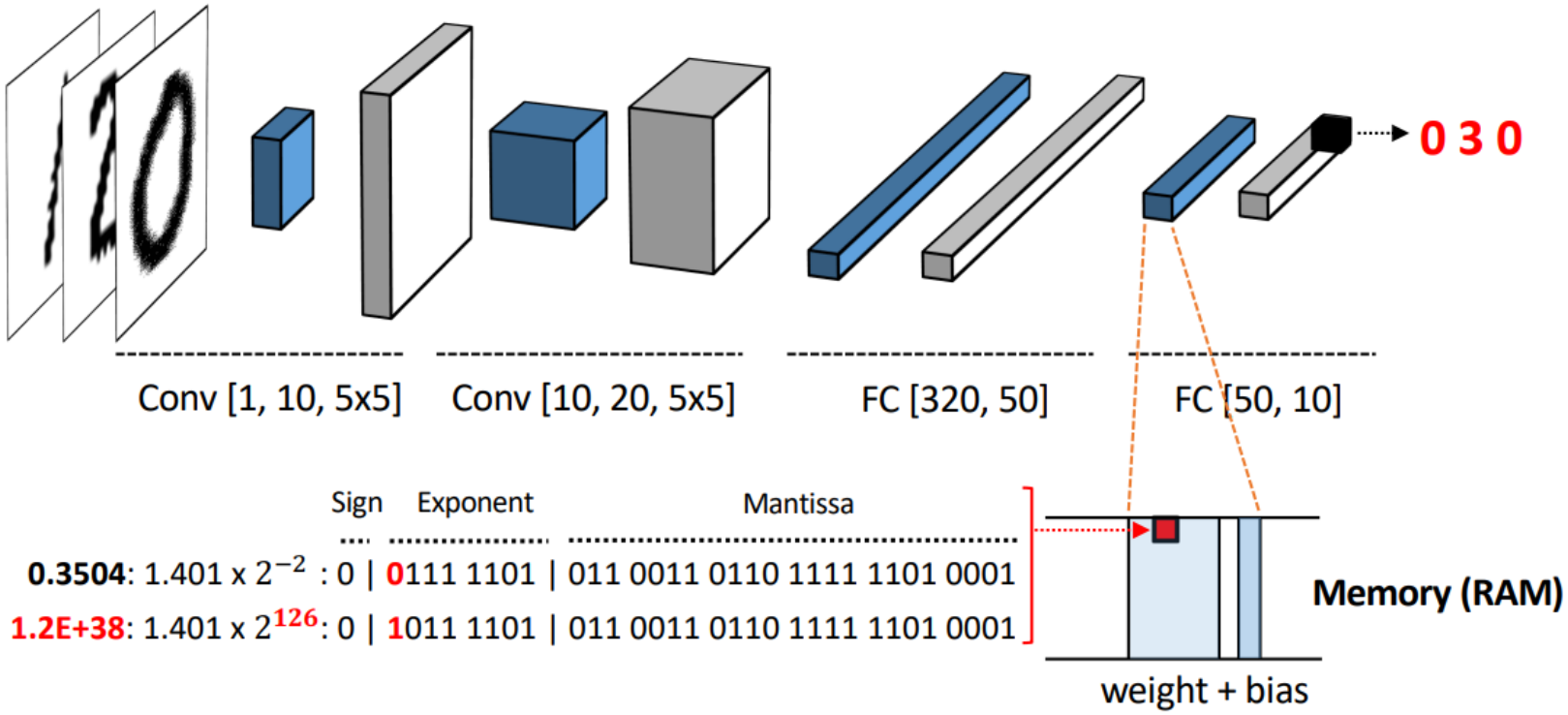


Illustration of DNN (In memory Representation)

- Accuracy: **57.52% (41.01% drop)**



Conclusion

- Hardware attack can break mathematically proven guarantees.
- Stealthy form of attack.
- Other attacks on machine learning models are possible through hardware implementations.
- Single-bit flip can inflict maximum damage if it's the most significant bit. (Achilles bit)

References

1. Hong, Sanghyun, et al. "Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks." *USENIX Security Symposium*. 2019.
2. Clements, Joseph, and Yingjie Lao. "Hardware trojan design on neural networks." *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2019.
3. Clements, Joseph, and Yingjie Lao. "Hardware trojan attacks on neural networks." *arXiv preprint arXiv:1806.05768* (2018).
4. Hong, Sanghyun, and Maryland Cybersecurity Center MC. "A Sound Mind in A Vulnerable Body: Practical Hardware Attacks on Deep Learning."

Thank you for your attention!

Questions ??

adevkot2@cougarnet.uh.edu

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING