

AWSome Women Summit

LATAM 2025 - Edición Perú **Tercera edición**

Introduction to Adversarial Machine Learning

Anmol Agarwal

Saturday, March 29, Lima, Peru







What is Adversarial Machine Learning?



AWSome Women Summit LATAM 2025 - Edición Perú

I am Anmol Agarwal 👋 !



International Speaker From the United States

Security researcher and adjunct professor Research specialized in AI security

This is the English version of the slides. There is another version of the slides in Spanish.



AWSome Women Summit LATAM 2025 - Edición Perú



What Will You Learn Today?

- What is adversarial machine learning?
- Different kinds of Adversarial Machine Learning attacks
- Real Life Use Cases with Adversarial Machine Learning attacks
- Defense strategies





Made by FREE-VECTORS.NE

Introduction to Adversarial Machine Learning

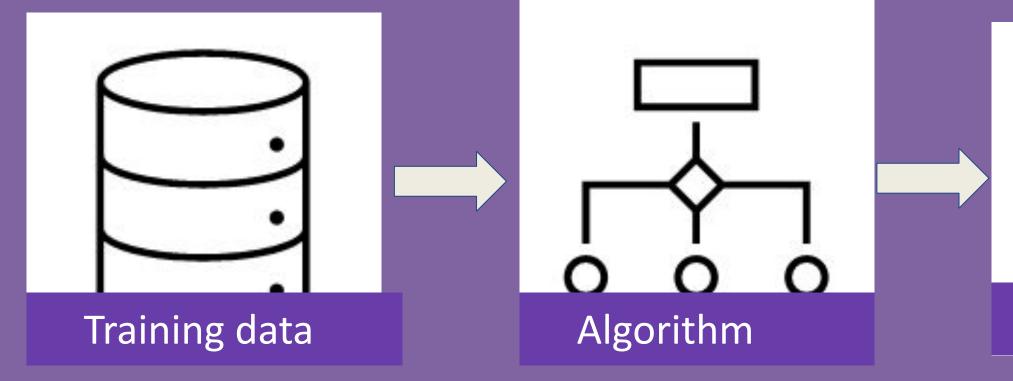
- Study of attacking and defending Machine Learning
- In Machine Learning, we have lots of data, and this data needs to be protected.
- The Machine Learning model can be attacked
- Important to understand when developing Machine Learning in AWS

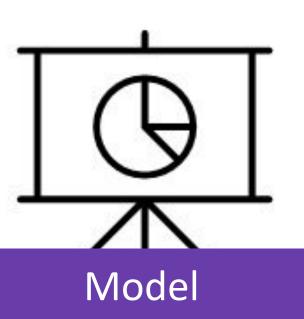




Machine Learning Basics

- Before learning about adversarial Machine Learning, let us learn about Machine Learning generally
- Machine Learning/AI analyzes lots of data and from that data gives a prediction
- Machine Learning datasets can be split into 2 sets training dataset and testing datasets







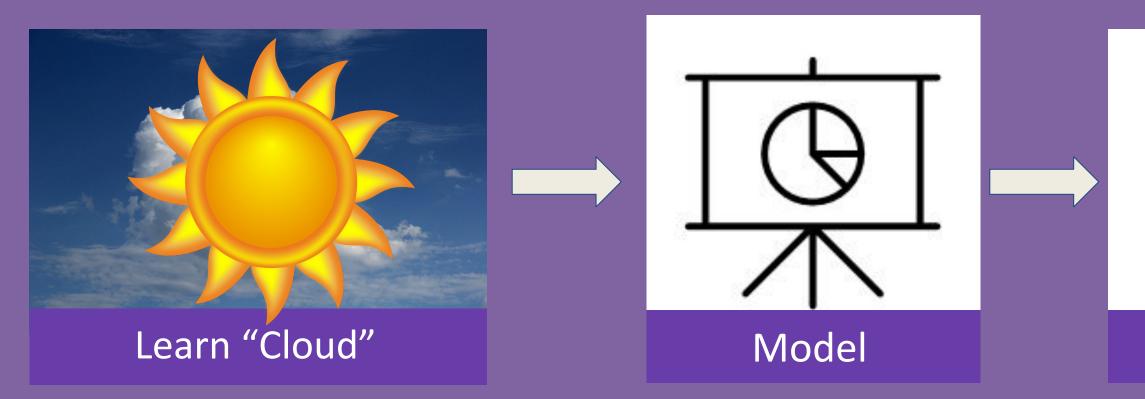
Different Types of Adversarial Machine Learning Attacks



AWSome Women Summit LATAM 2025 - Edición Perú

Poisoning Attack

- Model learns incorrect information when being trained
- Either the training data is changed or a backdoor is inserted
- Learning incorrect information == Incorrect results from AI/Machine Learning



ted \I/Machine Learning



Cloud



Poisoning Attack - In Real Life



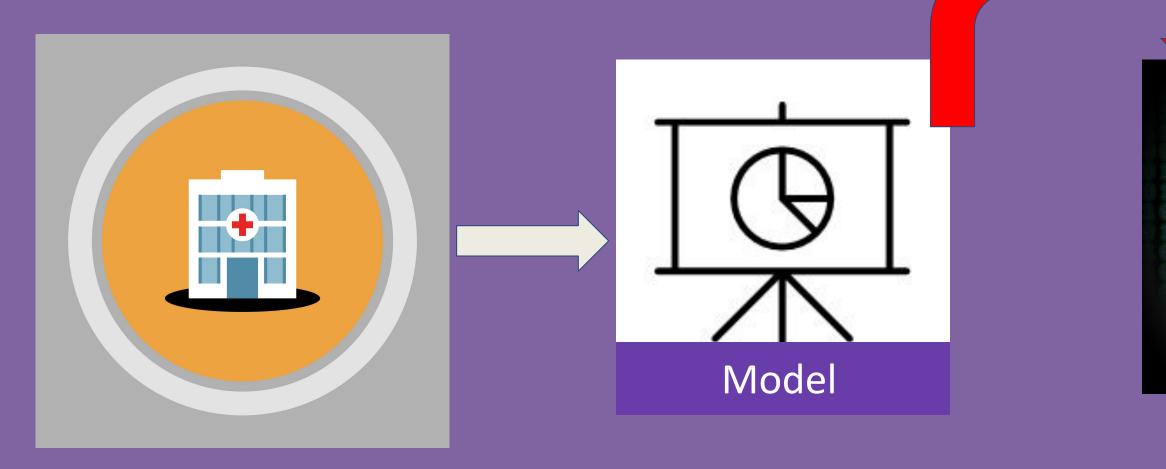
Learned from Twitter data, within 24 hours had to be shut down because poisoned and started saying bad things



The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

Model Extraction Attack

- Stealing a model to make a model that is better or the same as the victim
- Get model for free without spending money on development/training
- Violates data privacy and Intellectual Property



me as the victim nent/training



Model Extraction Attack - Research

Step 1: Attacker randomly samples words to form queries and sends them to victim BERT model



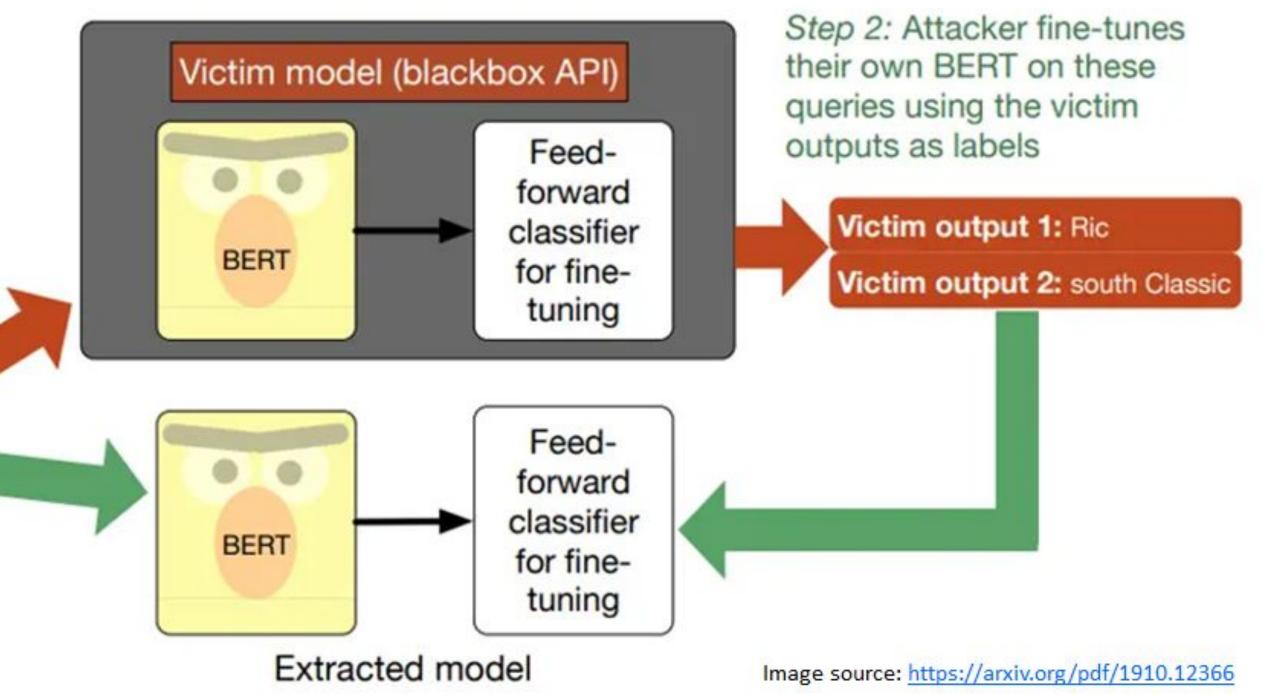
passage 1: before selling ?' New about to in Week the American each Colonel characters, from and as in including and a shooter Efforts happened, as on as measured. and and the (which proper and that as Ric for living interest Air ...

question: During and living and in selling Air?



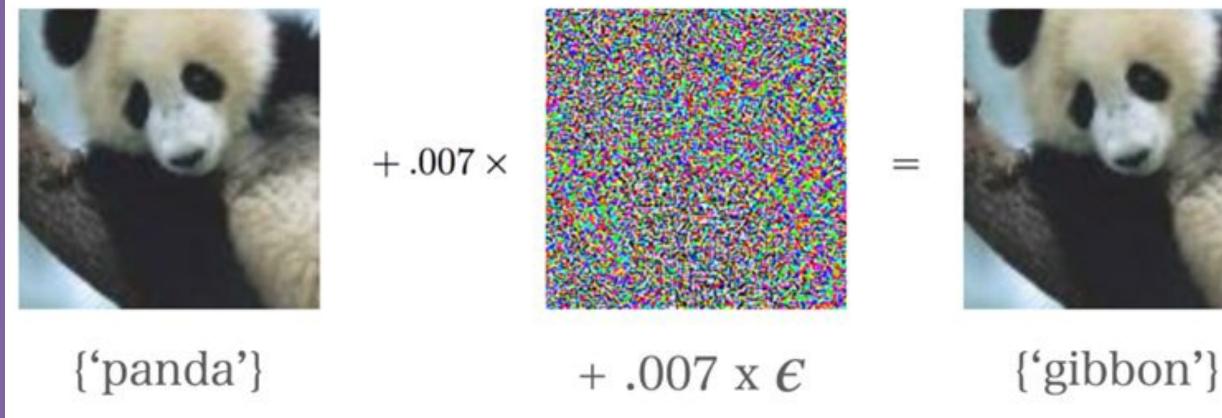
passage 2: Arab in (Dodd) singer, as to orthologues November giving small screw Peng be at and sea national Fire) there to support south Classic, Quadrille promote filmed

question: Which national giving Classic, Quadrille national as?



Evasion Attack

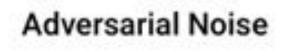
- The model is sent an "adversarial example" that causes a misclassification
- The "adversarial example" is an input that looks like it is uncontaminated to the human eye but has slight variations

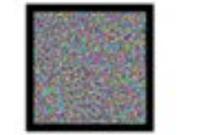






"panda"





+





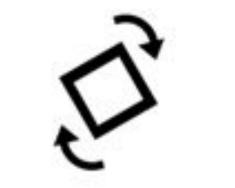






"vulture"

Adversarial Rotation





"orangutan"



Orangutan



"not hotdog"

Adversarial Photographer



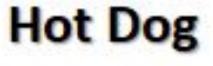


"hotdog"



• Source:Reference: https://research.google/blog/introducing-the-unrestricted-adversarial-examples-challenge/







Evasion Attack - In Real Life

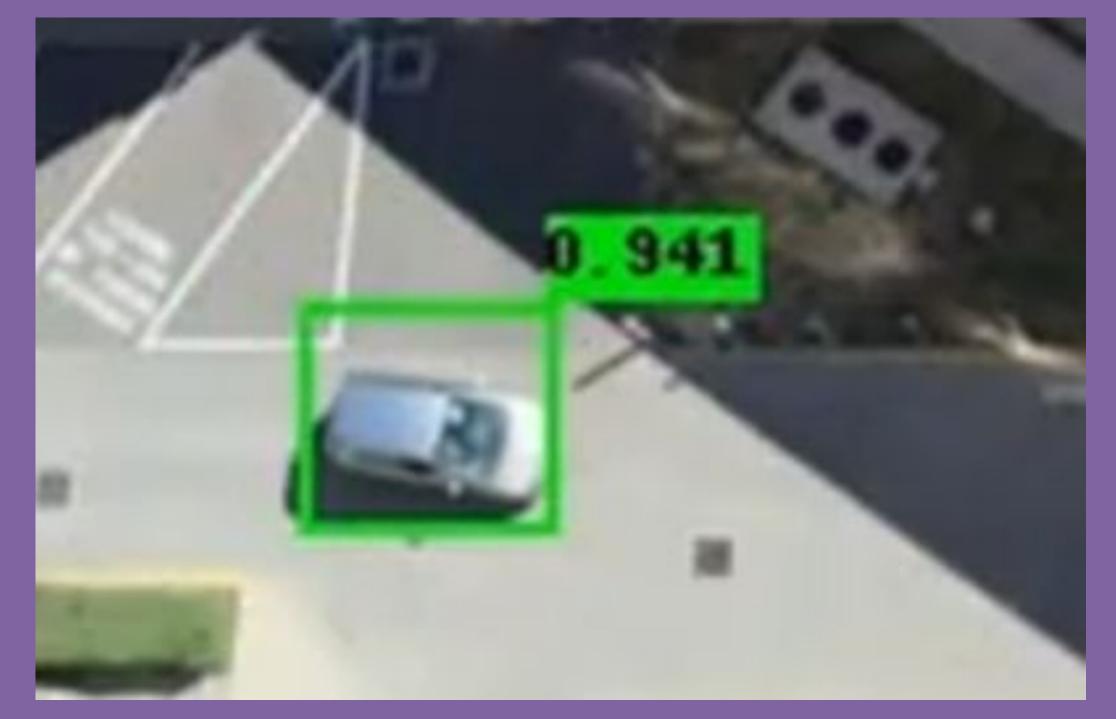
• Invisibility cloak by Facebook AI



Adversarial examples on sweater fools the Machine Learning models.

Deep Neural Networks are being used for Aerial imagery object detection Sentient Satellite Lab in Australia is researching AI for space and adversarial attacks on space domain (at University of Adelaide) They have demonstrated adversarial Machine Learning attacks

good case, everything is good





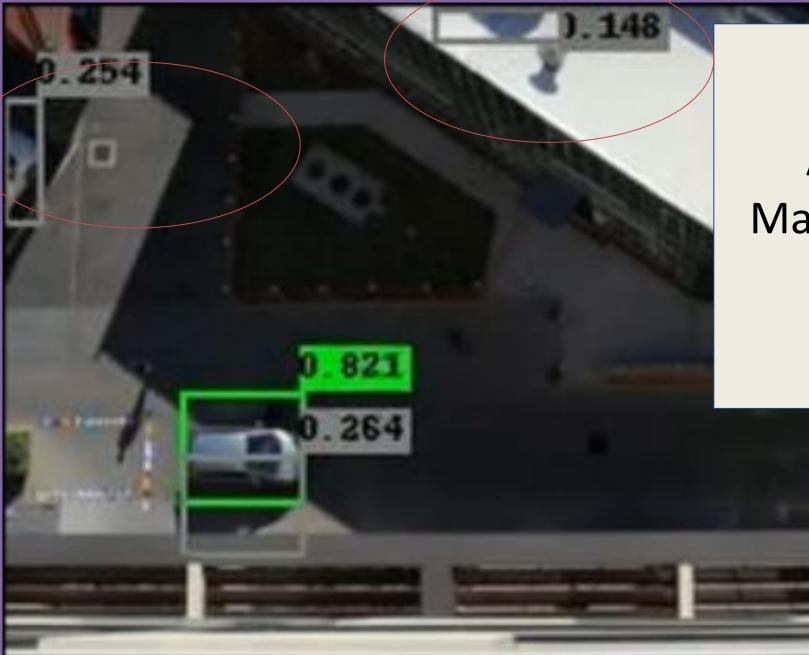


Errors are happening



Adding a sticker to the top of the car, causing object detection errors

More errors are happening



Adding adversarial examples to surroundings Machine Learning model thinks there is something next to the car (error)

Defense Strategies



AWSome Women Summit LATAM 2025 - Edición Perú

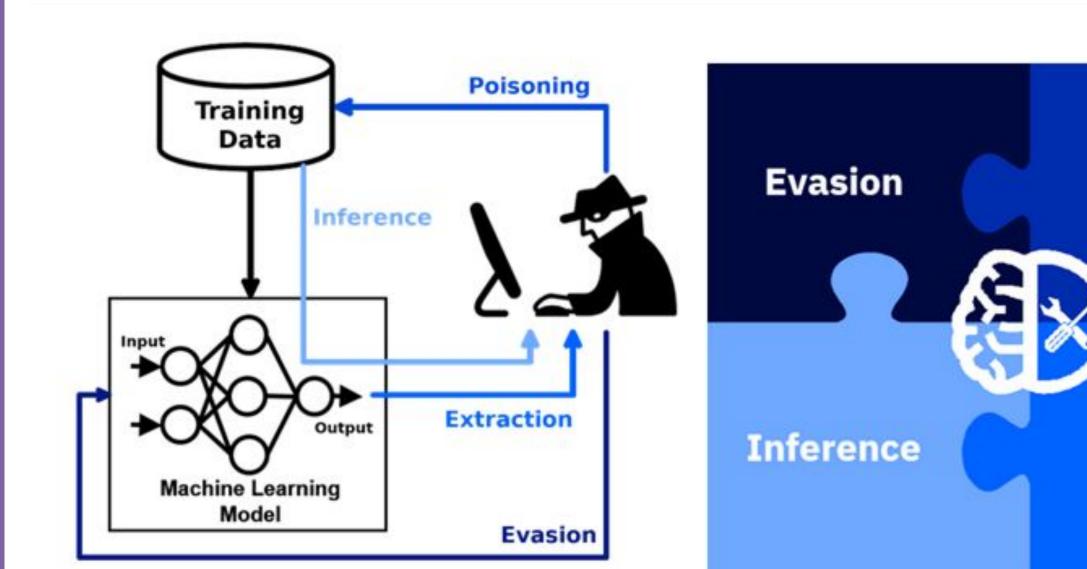
Defense Strategies

- Secure by Design
- Protect the data
- Follow cybersecurity principles
- Principle of Least Privilege monitor access
- Limit access to APIs
- Adversarial Machine Learning attack mitigations
- Outlier detection
- Store only necessary info in database, anonymize data
- Many open-source tools exist that help defend against adversarial machine learning attacks

Adversarial Robustness Toolbox

 <u>Adversarial Robustness Toolbox</u> – Python library to defend and evaluate machine learning





Poisoning





Model Scan

 <u>Model Scan</u> – open- source tool from Protect AI to scan models to prevent malicious code from being loaded onto the model

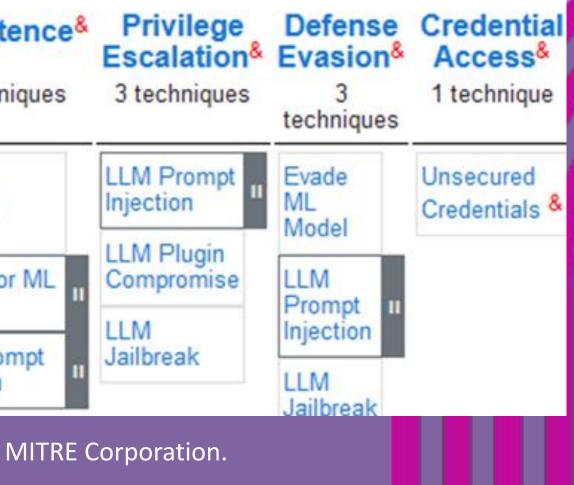
> Gives a list of vulnerabilities found critical, high, medium, and low errors

MITRE ATLASTM

- Adversarial Threat Landscape for Artificial Intelligence Systems developed by MITRE
- Tactics/techniques by adversaries using well-known attacks
- Helps security analysts protect and defend systems

Reconnaissance ^{&}	Resource Development ^{&}	Initial Access ^{&}	ML Model Access	Execution ^{&}	
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techni
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API	User Execution ^{&}	Poison Training Data
	Obtain Capabilities & "	Valid	Access	Command and Scripting Interpreter &	Backdoor Model
Search for Publicly Available Adversarial Vulnerability Analysis		Accounts &	ML-Enabled Product or Service		
	Develop Capabilities & "	Evade ML Model			LLM Pron Injection
			Physical	LLM Plugin	

*MITRE ATLASTM and MITRE ATT&CK[®] are a trademark and registered trademark of The MITRE Corporation.



Summary

- Machine Learning is very important, used for many applications in many domains
 - But Machine Learning can be attacked through adversarial machine \bigcirc learning attacks
- When developing Machine Learning, design with security in mind
- Open-source tools exist to evaluate the security of machine learning models



AWSome Women Summit

LATAM 2025 - Edición Perú Tercera edición



Thank you for your attention!



















cloud enhusiants. AWS proceedingers. AMS export



References and Resources

- Adversarial Robustness Toolbox:
- <u>https://github.com/Trusted-Al/adversarial-robustness-toolbox</u>
 - Go to the notebooks directory for useful tutorials and examples
- ModelScan: https://github.com/protectai/modelscan?utm_referrer=https%3A%2F%2Fpr otectai.com%2Fmodelscan
- MITRE ATLAS: <u>https://atlas.mitre.org/</u>