



AWSome Women Summit

LATAM 2025 - Edición Perú

Tercera edición



Introducción al Machine Learning Adversarial

Anmol Agarwal

Sábado 29 de Marzo. Lima - Perú



AWSome Women Summit

LATAM 2025 - Edición Perú

¿Qué es el Machine Learning Adversarial?

¡ Soy Anmol Agarwal 🖐️ !



AWSome Women Summit
LATAM 2025 - Edición Perú



SCAN ME



Conferencista Internacional
Soy de los Estados Unidos
Investigador en seguridad informática y
profesor asociado
especialízate en seguridad de IA

Estoy traduciendo slides al español para ayudar a la audiencia a seguirlas. Lo siento si hay errores de traducción. Entiendo español pero prefiero hablar en inglés.



Sobre la charla

¿Qué aprenderás hoy?

- ¿Qué es Adversarial Machine Learning?
- Diferentes tipos de ataques de Machine Learning Adversario
- Casos de uso de la vida real de ataques de Machine Learning adversario
- Estrategias de defensa



Made by FREE-VECTORS.NET

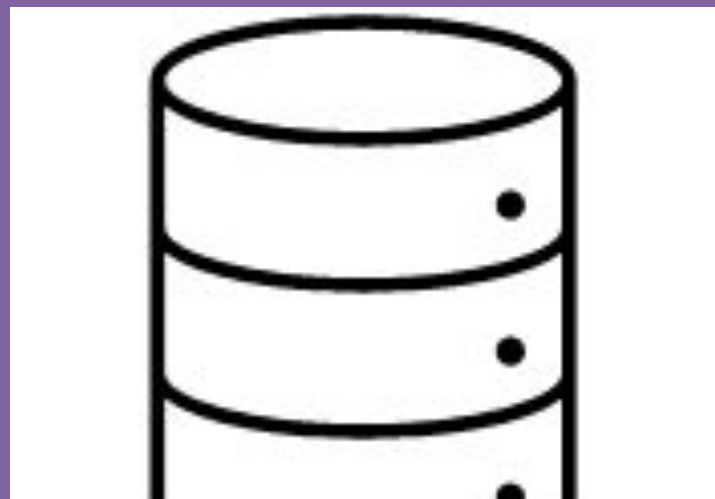
Introducción al Machine Learning Adversarial

- Estudio de ataque y defensa del Machine Learning
- En el Machine Learning, tenemos muchos datos y estos datos deben protegerse.
- El modelo de Machine Learning puede ser atacado
- Es importante comprender el desarrollo de Machine Learning

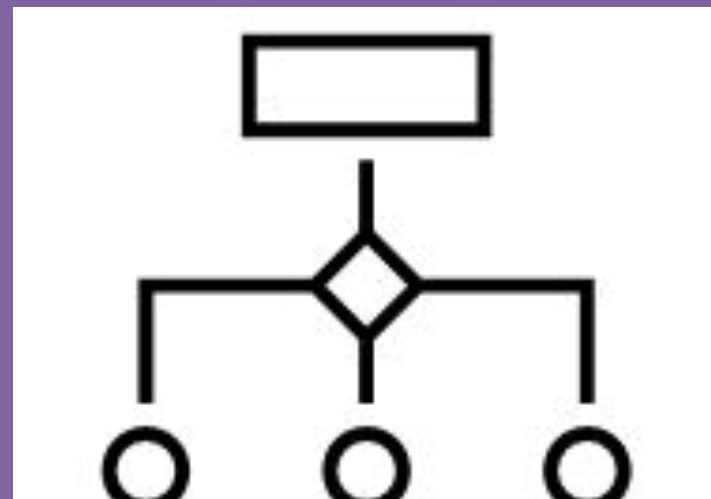


Conceptos básicos de Machine Learning

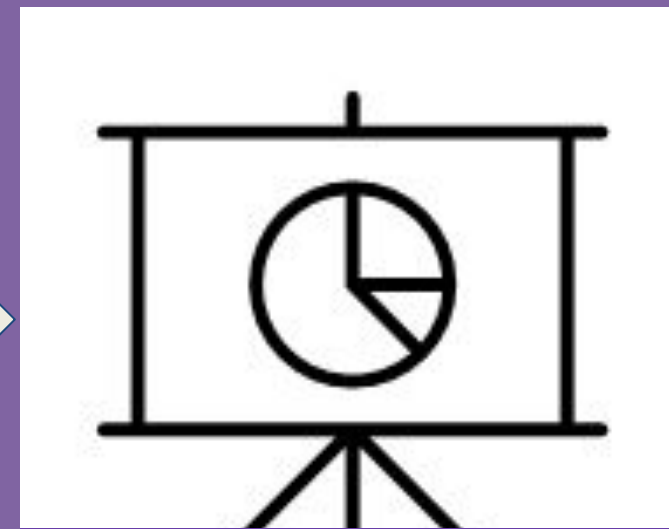
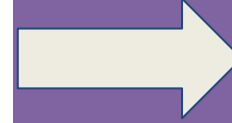
- Antes de aprender sobre el Machine Learning adversario, aprendamos sobre el Machine Learning en general
- El Machine Learning/IA analiza una gran cantidad de datos y, a partir de ellos, da una predicción
- Los conjuntos de datos de Machine Learning se pueden dividir en 2 conjuntos: conjunto de datos de capacitación y de prueba



Datos de capacitación



Algoritmo



Modelo



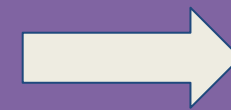
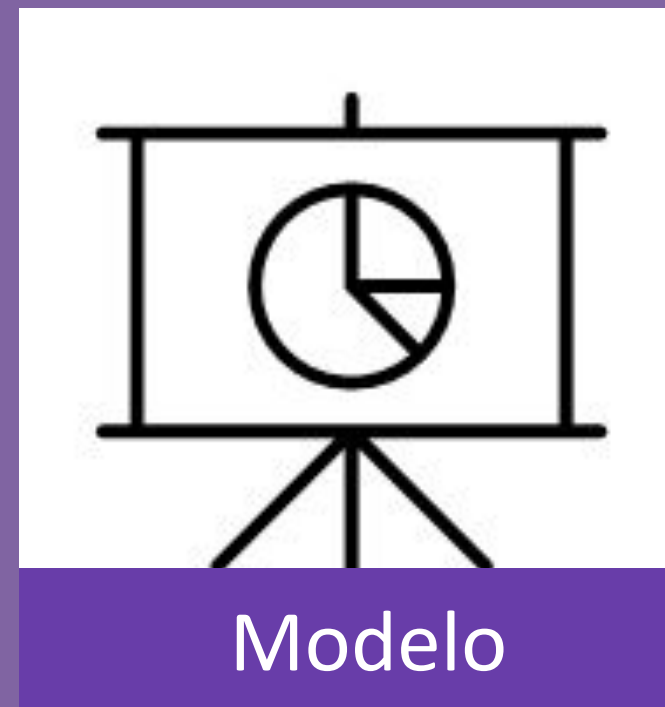
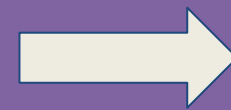
AWSome Women Summit

LATAM 2025 - Edición Perú

Diferentes tipos de ataques de Machine Learning Adversario

Ataque de envenenamiento (Poisoning)

- El modelo aprende información incorrecta cuando se entrena
- Se modifican los datos de capacitación o se inserta una puerta trasera
- Aprendizaje de información incorrecta == Resultados incorrectos de Machine Learning



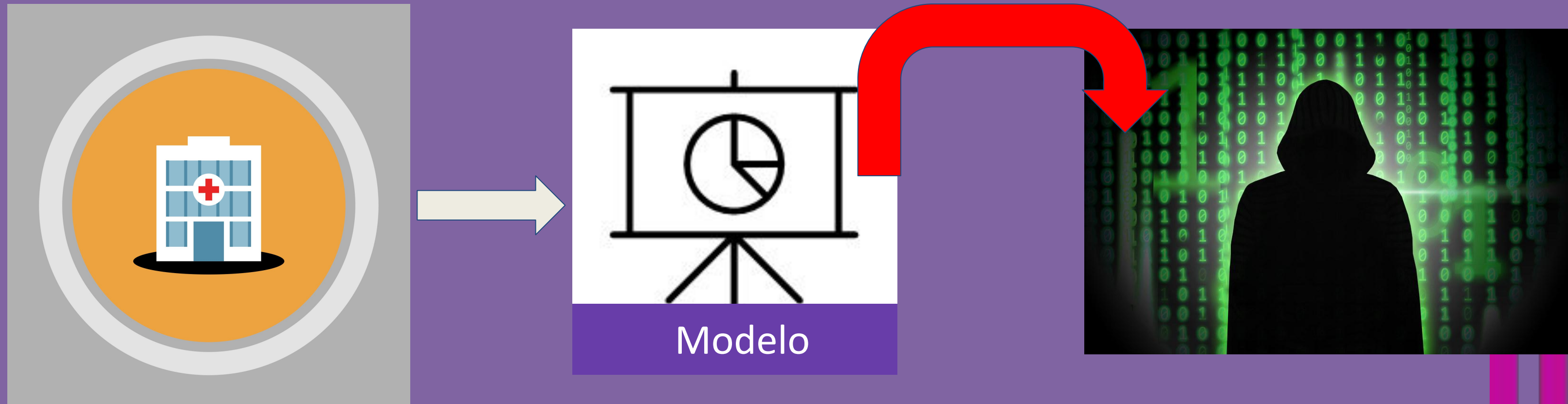
Ataque de Poisoning - en la vida real



Aprendido de los datos de Twitter, en 24 horas tuvo que ser cerrado porque estaba envenenado y comenzó a decir cosas malas

Ataque de extracción de modelos

- Robar un modelo para hacer un modelo que sea mejor o igual al de la víctima.
- Obtenga el modelo de forma gratuita sin gastar dinero en desarrollo
- Viola la protección de datos y la propiedad intelectual



Ataque de extracción de modelos - investigación

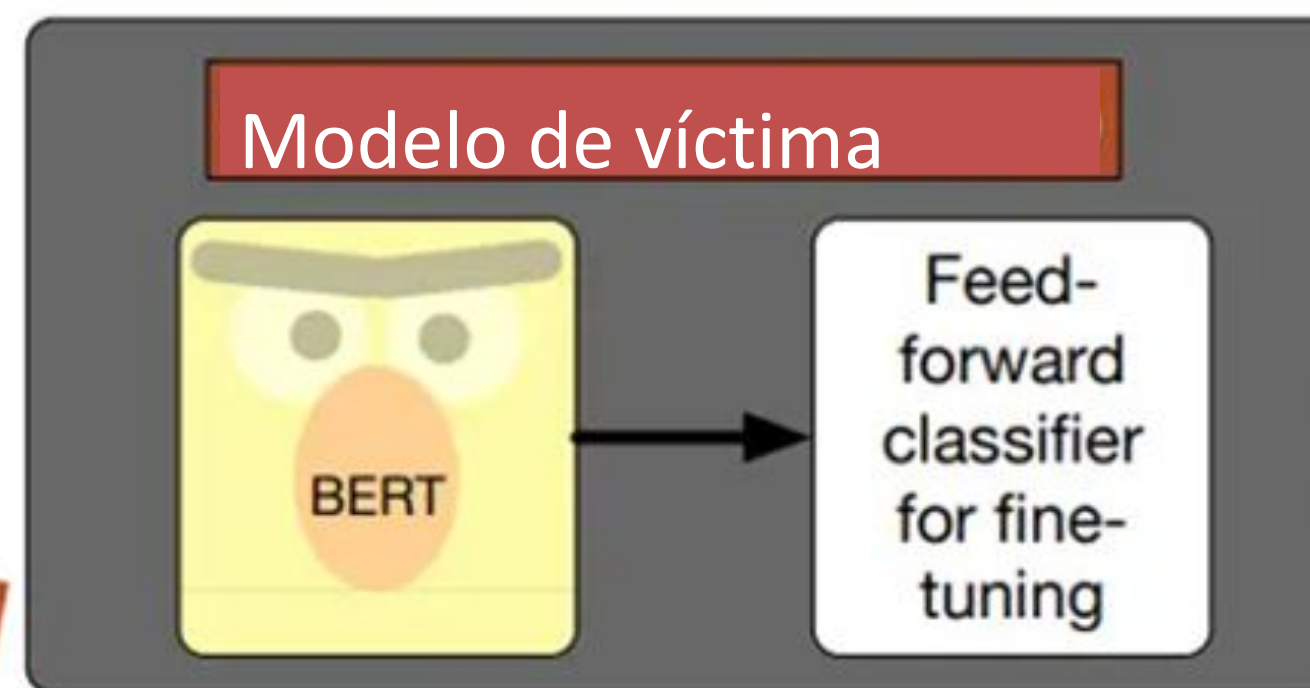
Paso 1: El atacante muestrea aleatoriamente las palabras para formar consultas y las envía al modelo de la víctima

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 ...

Pregunta During and living and in selling Air?

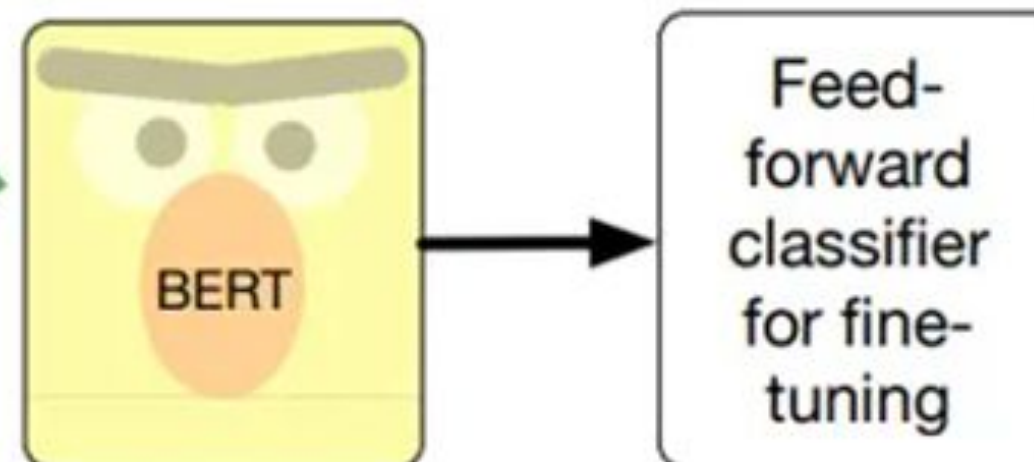
passage 2: Arab in (Dodd) singer, as to orthologues November giving small screw Peng be at and sea national Fire) there to support south Classic, Quadrille promote filmed ...

Pregunta Which national giving Classic, Quadrille national as?



Paso 2. El atacante crea su propio modelo en estas consultas utilizando la salida de la víctima como labels

Salida de victima 1: Ric
salida de victima 2: South Classic



Modelo extraído

Image source: <https://arxiv.org/pdf/1910.12366>

Ataque de evasión

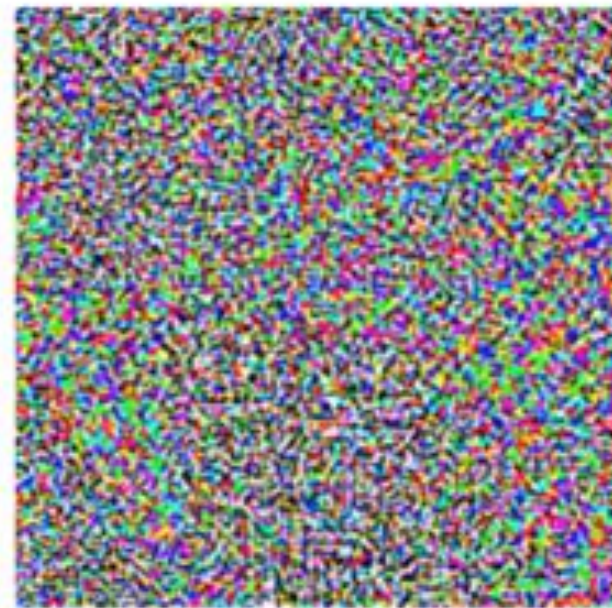
Al modelo se le envía un "ejemplo adversarial" que provoca una clasificación errónea

El "ejemplo adversarial" es una entrada que parece no contaminada para el ojo humano, pero tiene ligeras variaciones



{‘panda’}

+ .007 ×



+ .007 × ϵ

=



{‘gibbon’}





"panda"

+



=



"gibbon"



gibón



"vulture"

+



=



"orangutan"



orangután



"not hotdog"

+



=



"hotdog"



Hot Dog

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

Ataque de evasión - en la vida real

- Capa de invisibilidad de Facebook AI y University of Maryland College Park



Los ejemplos adversariales en el suéter engañan al modelo de Machine Learning.

Ataque de evasión - en la vida real - Imágenes aéreas

- Las redes neuronales profundas se utilizan para la detección de objetos de imágenes aéreas
- El Laboratorio de Satélites Sentient en Australia está investigando la IA para el espacio y los ataques adversarios en el dominio espacial (en la Universidad de Adelaida)
- Demostrado ataques adversarios de Machine Learning

Ataque de evasión - en la vida real - Imágenes aéreas

Buen caso, todo está bien



Ataque de evasión - en la vida real - Imágenes aéreas

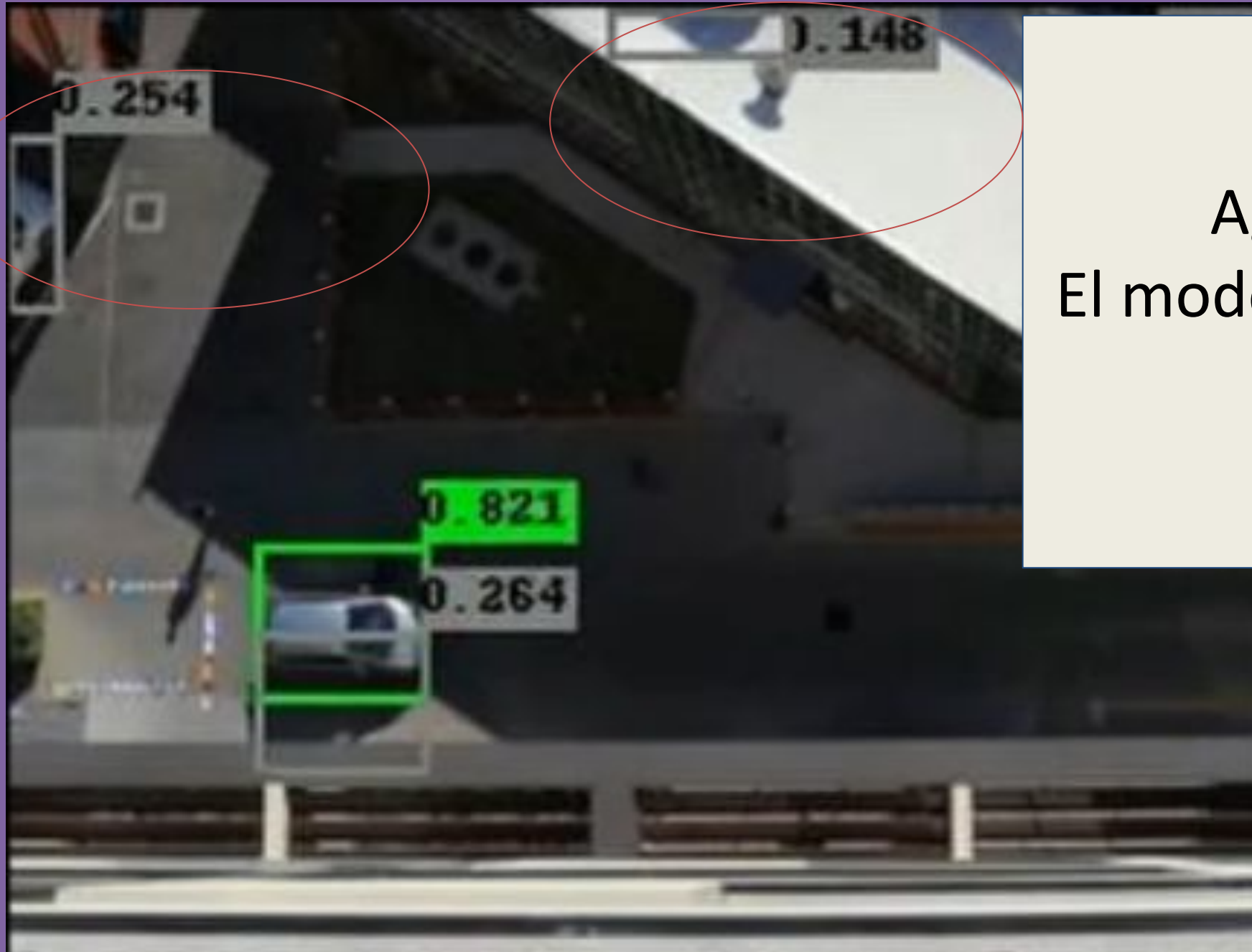
Están ocurriendo errores



Agregar una pegatina en la parte superior del
automóvil provoca errores de detección de objetos

Ataque de evasión - en la vida real - Imágenes aéreas

Están ocurriendo más errores



Agregar ejemplos adversariales al entorno
El modelo de Machine Learning cree que hay algo al
lado del coche (error)



AWSome Women Summit

LATAM 2025 - Edición Perú

Estrategias de defensa

Estrategias de defensa

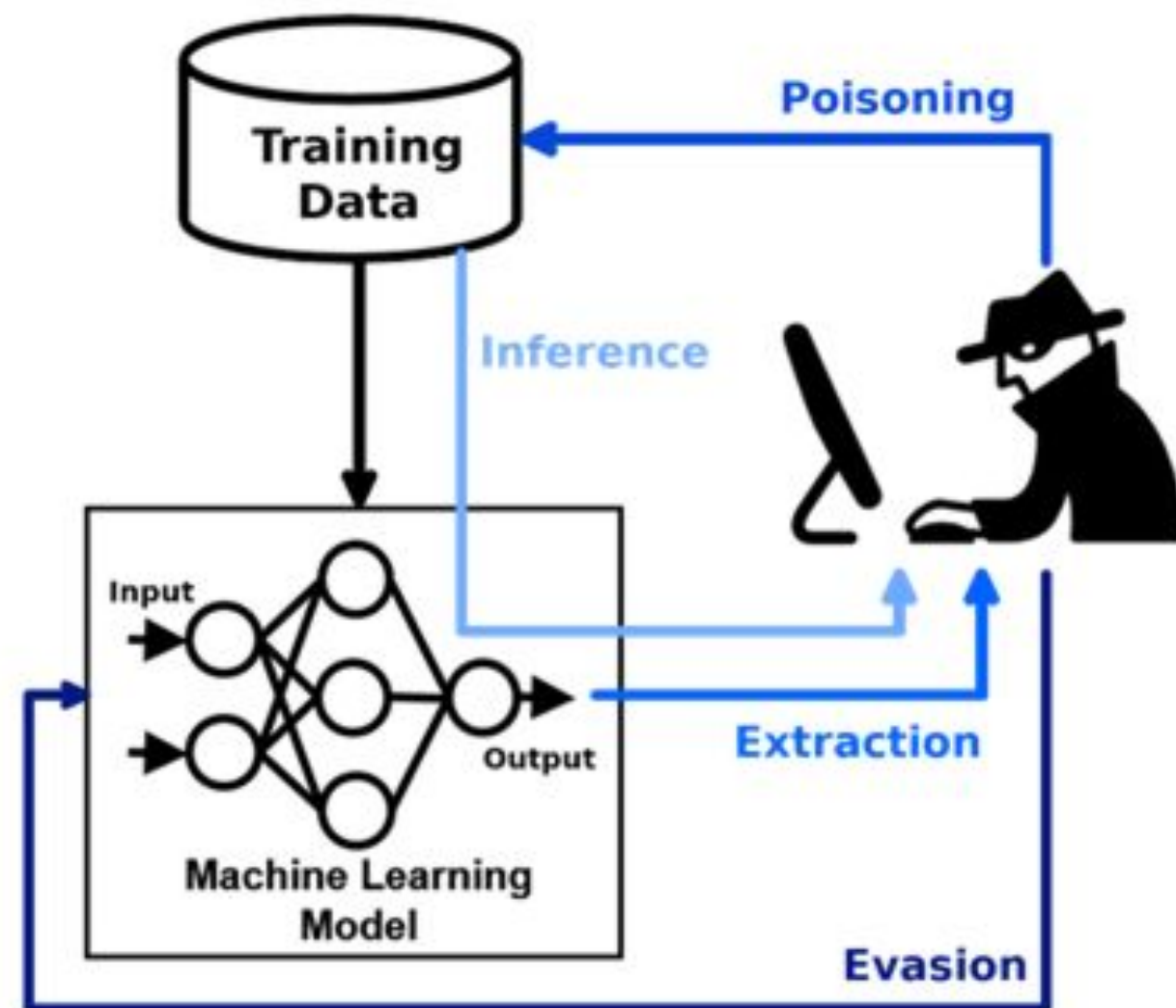
- Seguro por diseño
- Proteger los datos
- Seguir los principios de seguridad informática
- Principio de privilegio mínimo: supervisión de accesos
- Restringir el acceso a las API
- Mitigaciones de ataques de Machine Learning adversario
- Detección de valores atípicos
- Almacene solo la información necesaria en la base de datos
- Existen muchas herramientas de código abierto que ayudan a defenderse contra los ataques de Machine Learning adversarios

Adversarial Robustness Toolbox

- Adversarial Robustness Toolbox –Biblioteca de Python para defender y evaluar el Machine Learning



Adversarial Threats



Model Scan

- Model Scan – de Protect AI para escanear modelos para evitar que se cargue código malicioso en el modelo

Da una lista de vulnerabilidades encontradas:
Errores críticos, altos, medios y bajos



MITRE ATLAS™

- Adversarial Threat Landscape for Artificial Intelligence Systems desarrollado por MITRE
- Tácticas/técnicas de los adversarios que utilizan ataques comunes
- Ayuda a los analistas de seguridad informática a proteger y defender los plataformas

Reconnaissance & 5 techniques	Resource Development & 7 techniques	Initial Access & 6 techniques	ML Model Access 4 techniques	Execution & 3 techniques	Persistence & 3 techniques	Privilege Escalation & 3 techniques	Defense Evasion & 3 techniques	Credential Access & 1 technique
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 &
Search for Publicly Available Adversarial Vulnerabilities	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection	
	Develop	Evade ML			LLM Prompt Injection	LLM Jailbreak		

*MITRE ATLAS™ y MITRE ATT&CK® son una marca comercial y una marca comercial registrada de The MITRE Corporation.

Puntos Clave

- El Machine Learning es muy importante, se utiliza para muchas aplicaciones en muchos dominios
 - Pero Machine Learning puede ser atacado a través de ataques de Machine Learning adversarios
- Al desarrollar Machine Learning, diseñe teniendo en cuenta la seguridad
- Existen herramientas de código abierto para evaluar la seguridad de los modelos de Machine Learning



AWSome Women Summit

LATAM 2025 - Edición Perú

Tercera edición



PUCP

SCAN ME



¡Muchas gracias por tu atención!



Referencias y Recursos

- Adversarial Robustness Toolbox:
<https://github.com/Trusted-AI/adversarial-robustness-toolbox>
- ModelScan:
https://github.com/protectai/modelscan?utm_referrer=https%3A%2F%2Fprotectai.com%2Fmodelscan
- MITRE ATLAS: <https://atlas.mitre.org/>