AWSome Women Summit

LATAM 2025 - Edición Perú

Tercera edición

PUCP

# Introduction to Adversarial Machine Learning

Anmol Agarwal

Saturday, March 29, Lima, Peru

AWSome Women Summit

LATAM 2025 - Edición Perú

# What is Adversarial Machine Learning?

# I am Anmol Agarwal 👋!

AWSome Women Summit
LATAM 2025 - Edición Perú

SCAN ME

International Speaker
From the United States

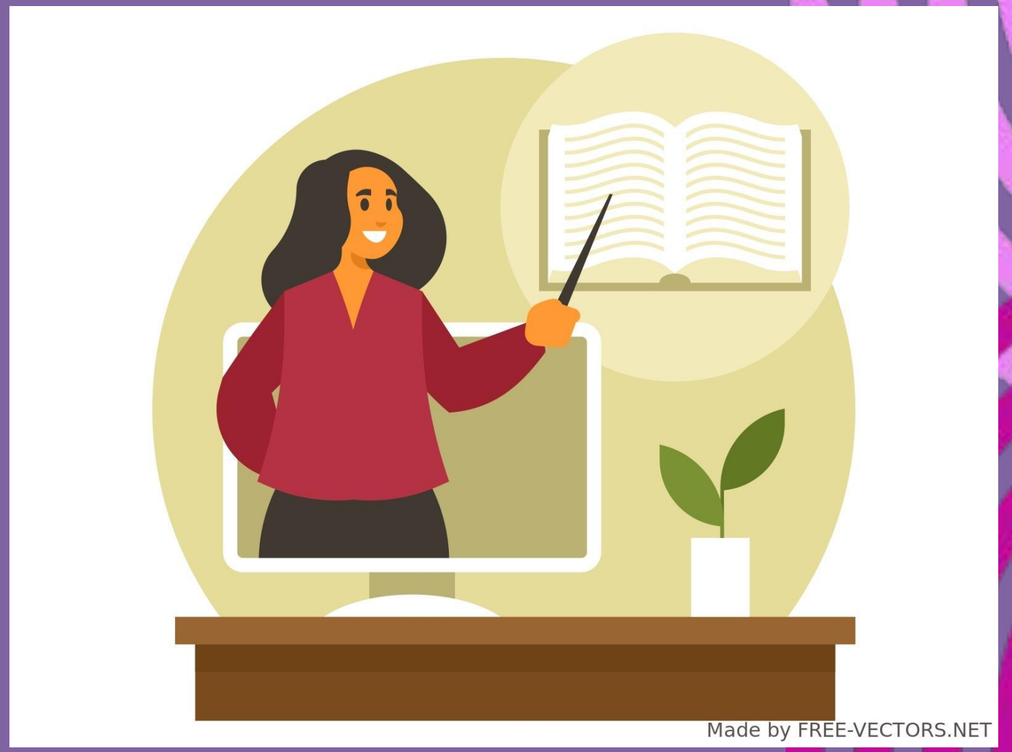
Security researcher and adjunct professor
Research specialized in AI security

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

About the talk

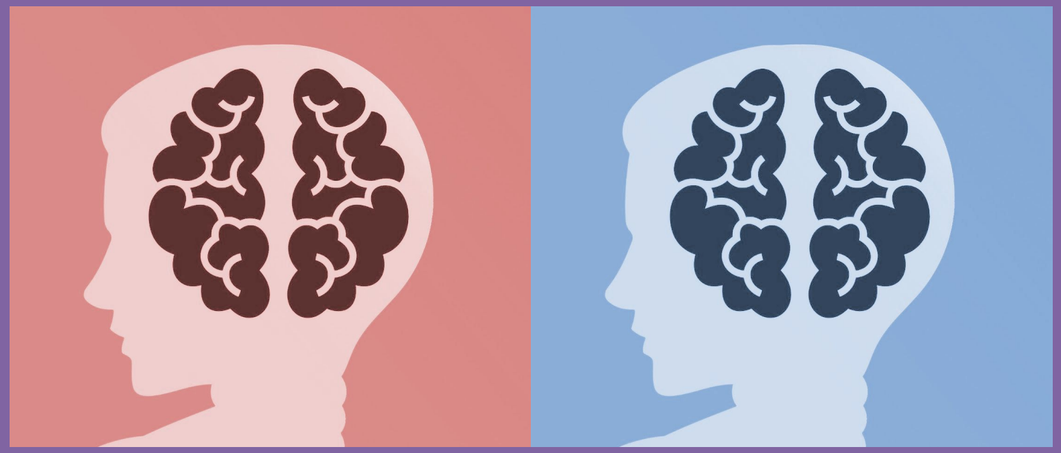
# What Will You Learn Today?

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



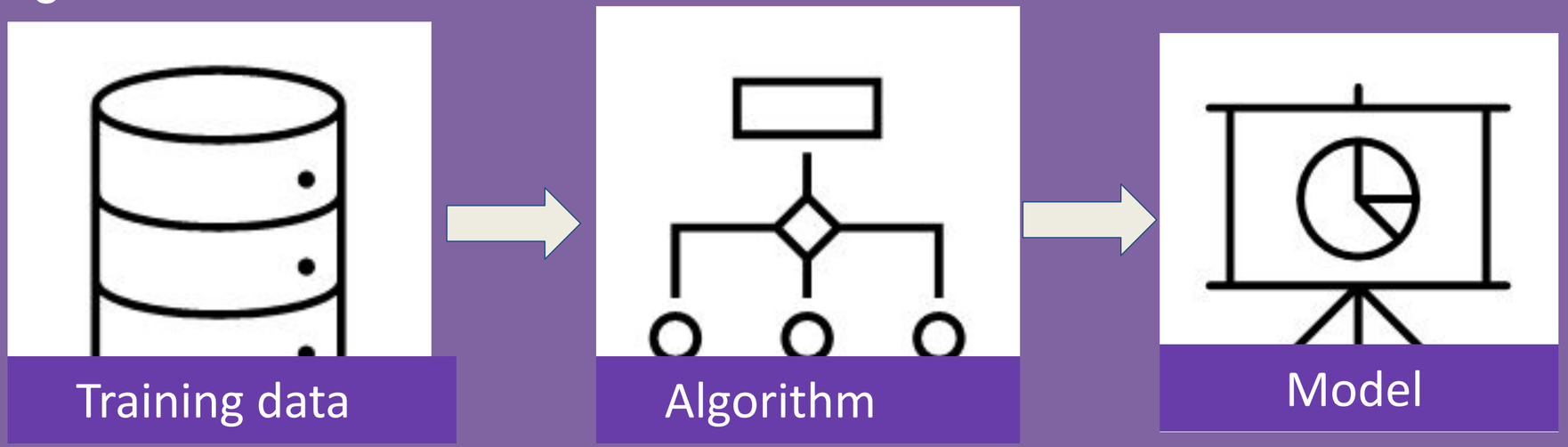
# Introduction to Adversarial Machine Learning

- 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



AWSome Women Summit

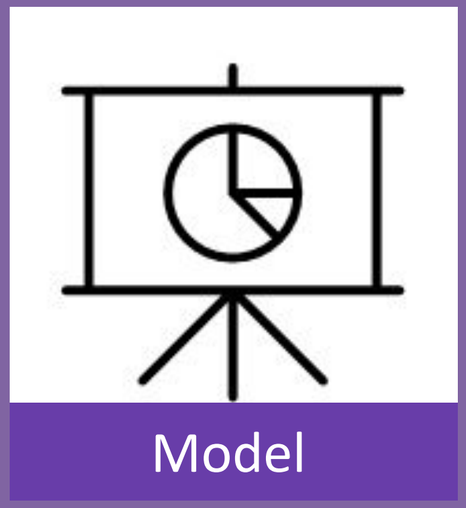
LATAM 2025 - Edición Perú

# Different Types of Adversarial Machine Learning Attacks

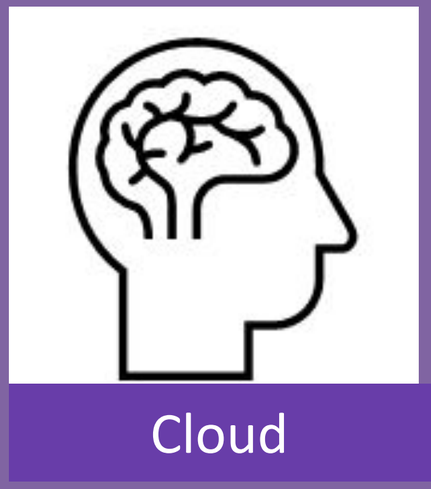
# Poisoning Attack

- 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

Learn “Cloud”



Model



Cloud

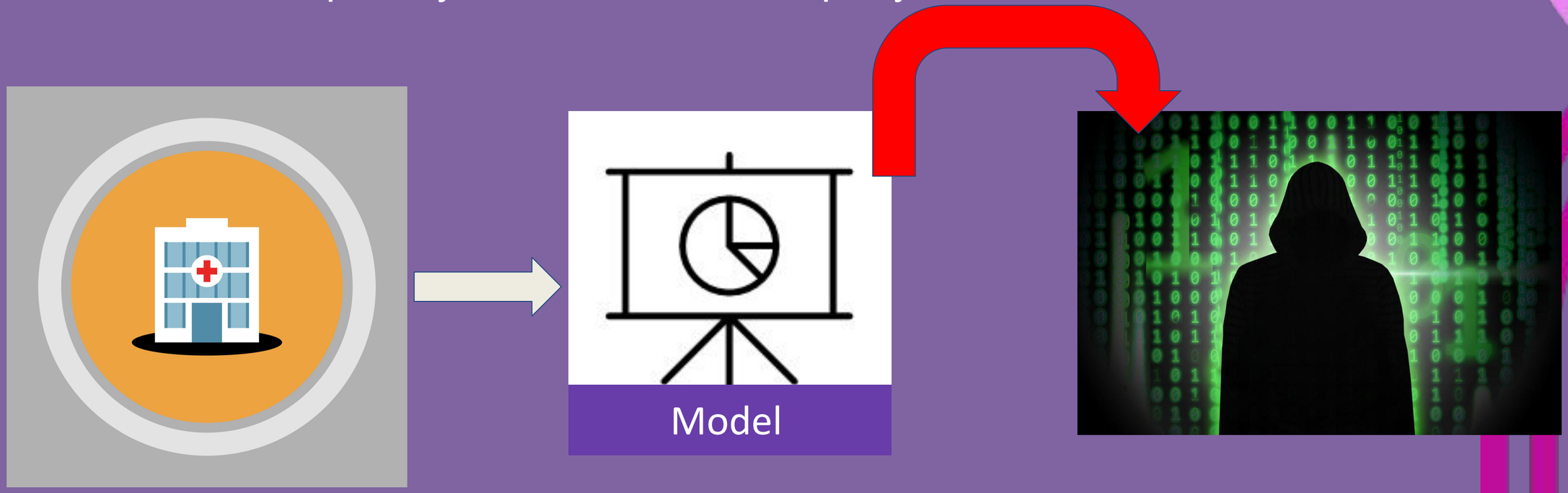
**Incorrect!**

# Poisoning Attack - In Real Life


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

# 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



# Model Extraction Attack - Research

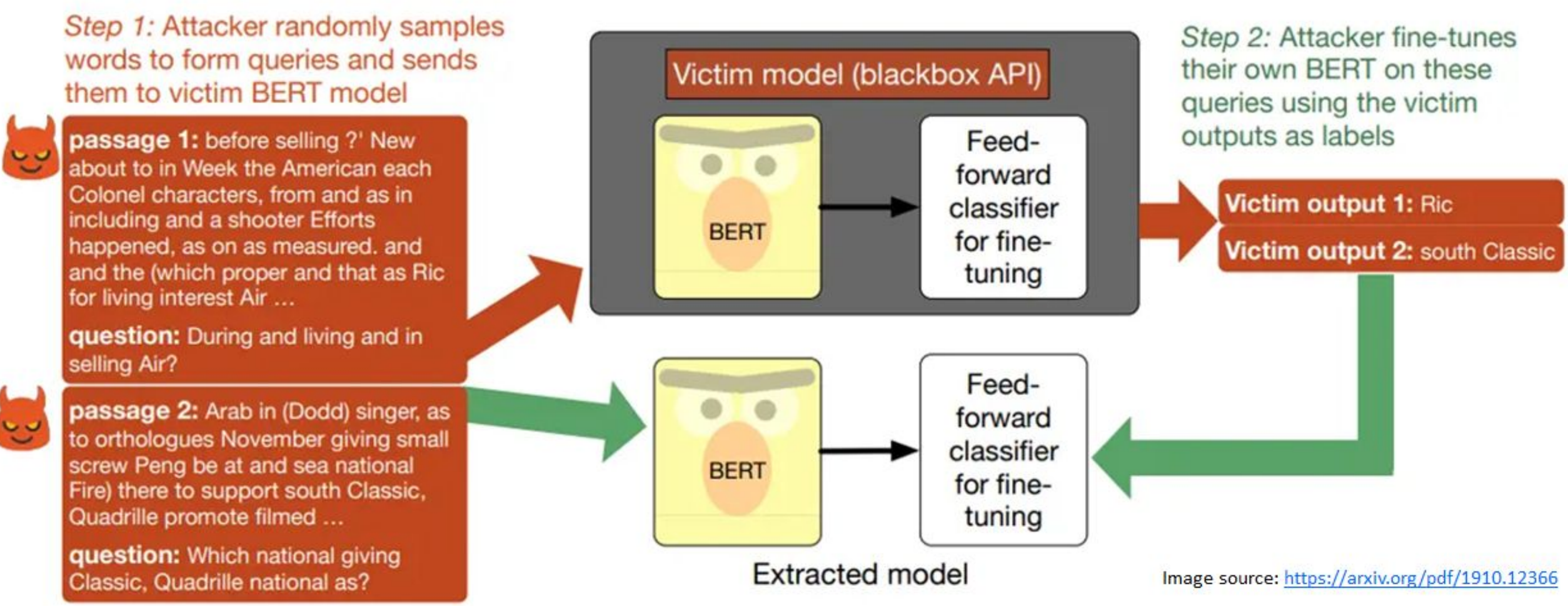


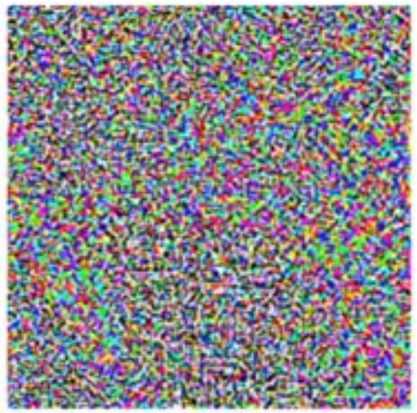

Image source: https://arxiv.org/pdf/1910.12366

# 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’}

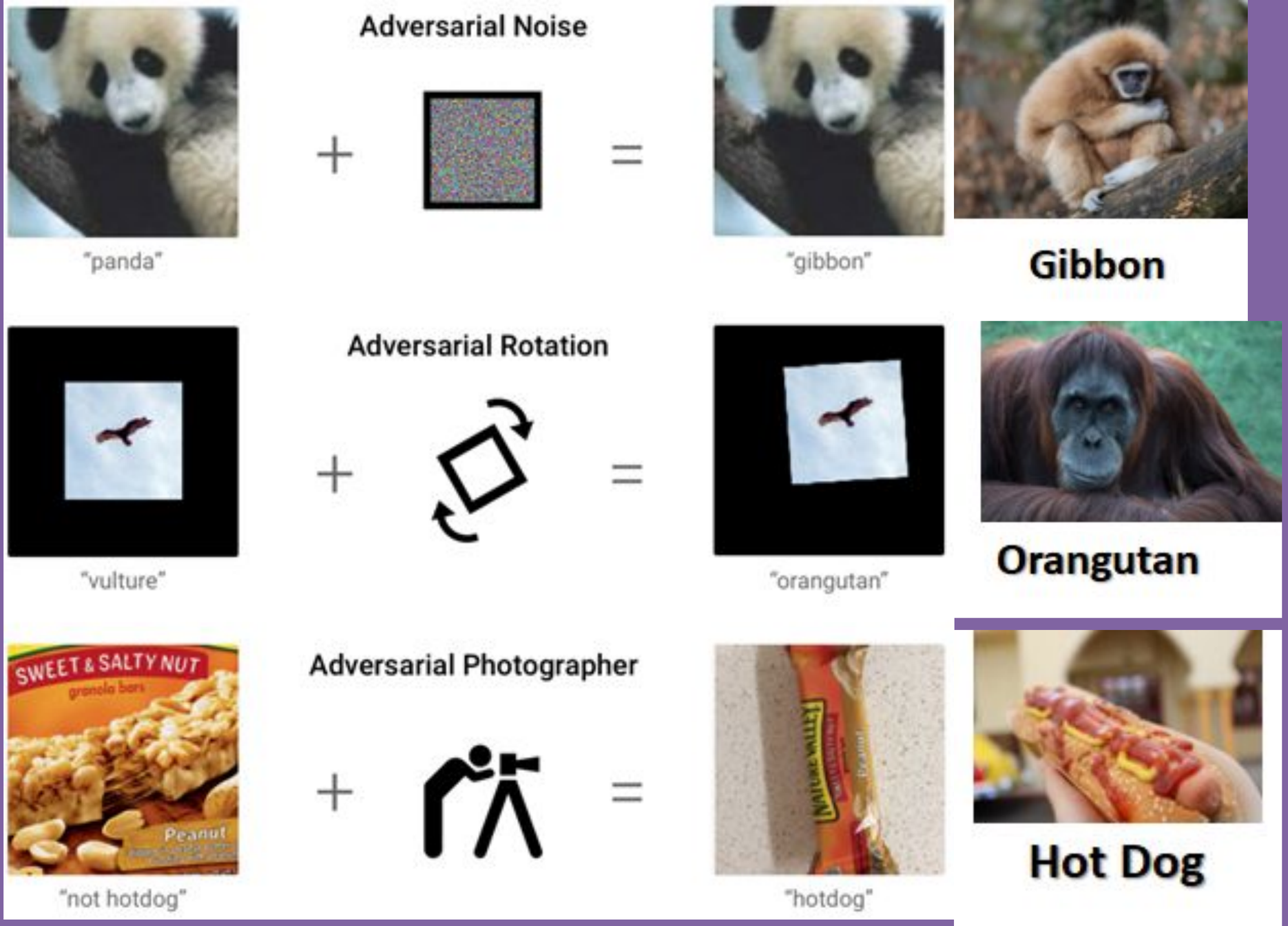
$+ .007 \times$



$+ .007 \times \epsilon$

$=$

{‘gibbon’}




- 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.

# Evasion Attack - In Real Life - Aerial Images

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

# Evasion Attack - In Real Life - Aerial Images

good case, everything is good

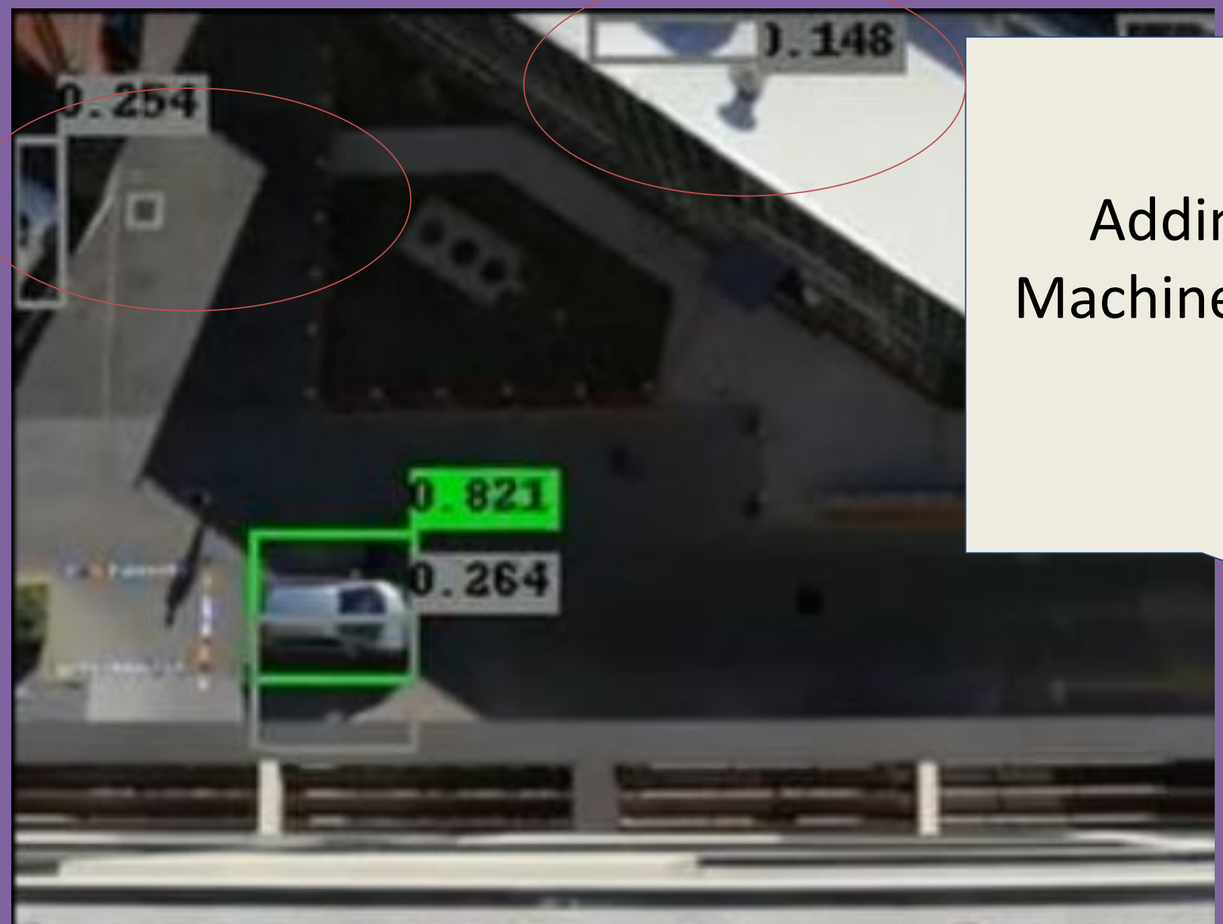
# Evasion Attack - In Real Life - Aerial Images

Errors are happening


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

# Evasion Attack - In Real Life - Aerial Images

More errors are happening




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

AWSome Women Summit

LATAM 2025 - Edición Perú

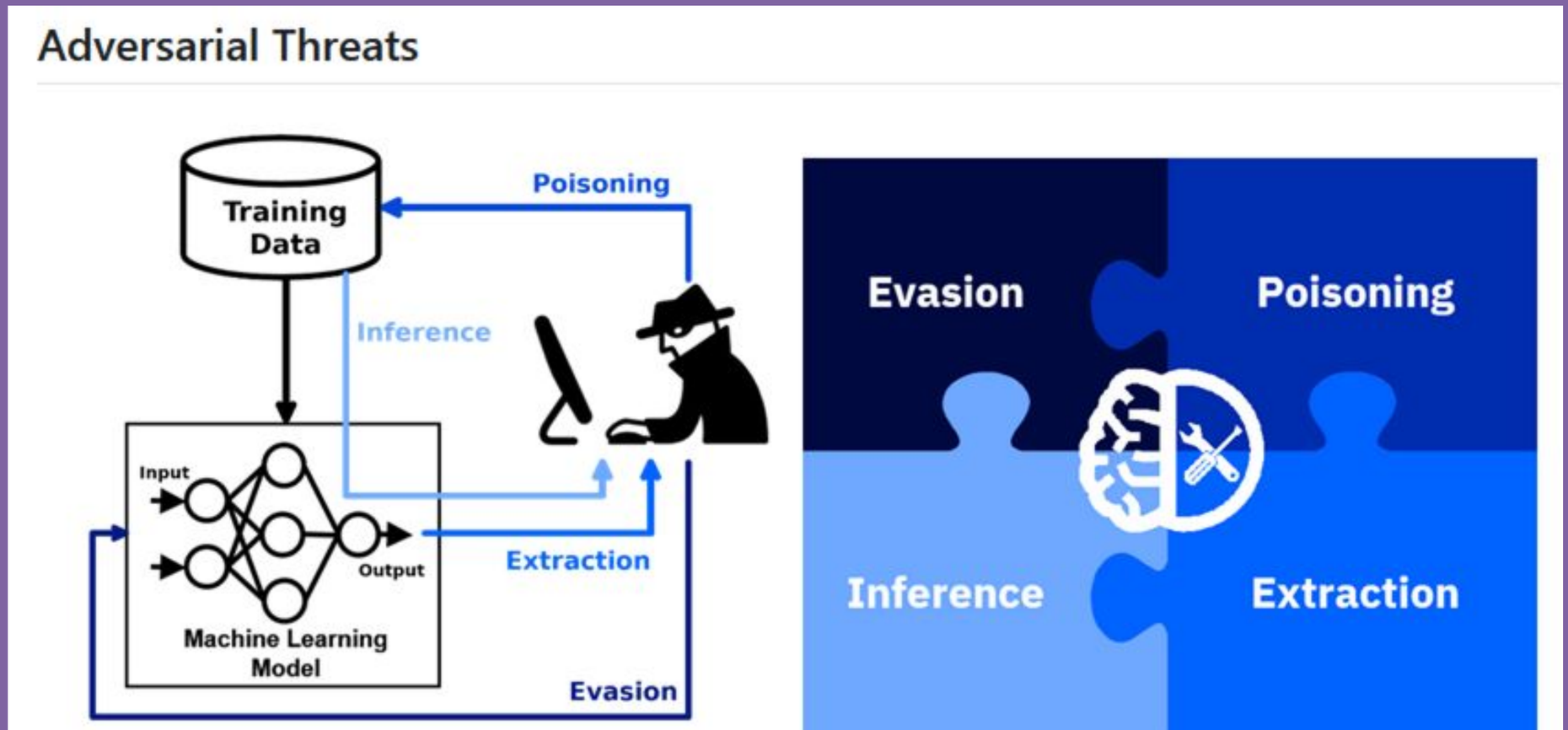
# Defense Strategies

# 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

- Adversarial Robustness Toolbox – Python library to defend and evaluate machine learning



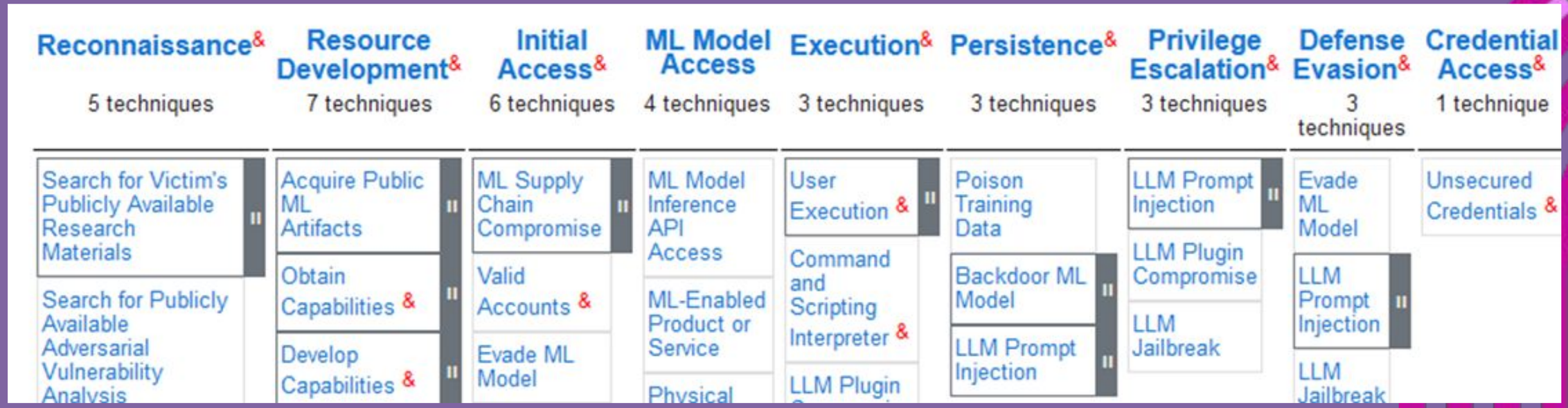
# Model Scan

- Model Scan – 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 ATLAS™

- 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




*MITRE ATLAS™ 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 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

PUCP

SCAN ME

# Thank you for your attention!

# References and Resources

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