OFFZONE 2022

Secure ML Modeling

Bolokhovtsev Vitaly

Appsec Engineer, Sberbank

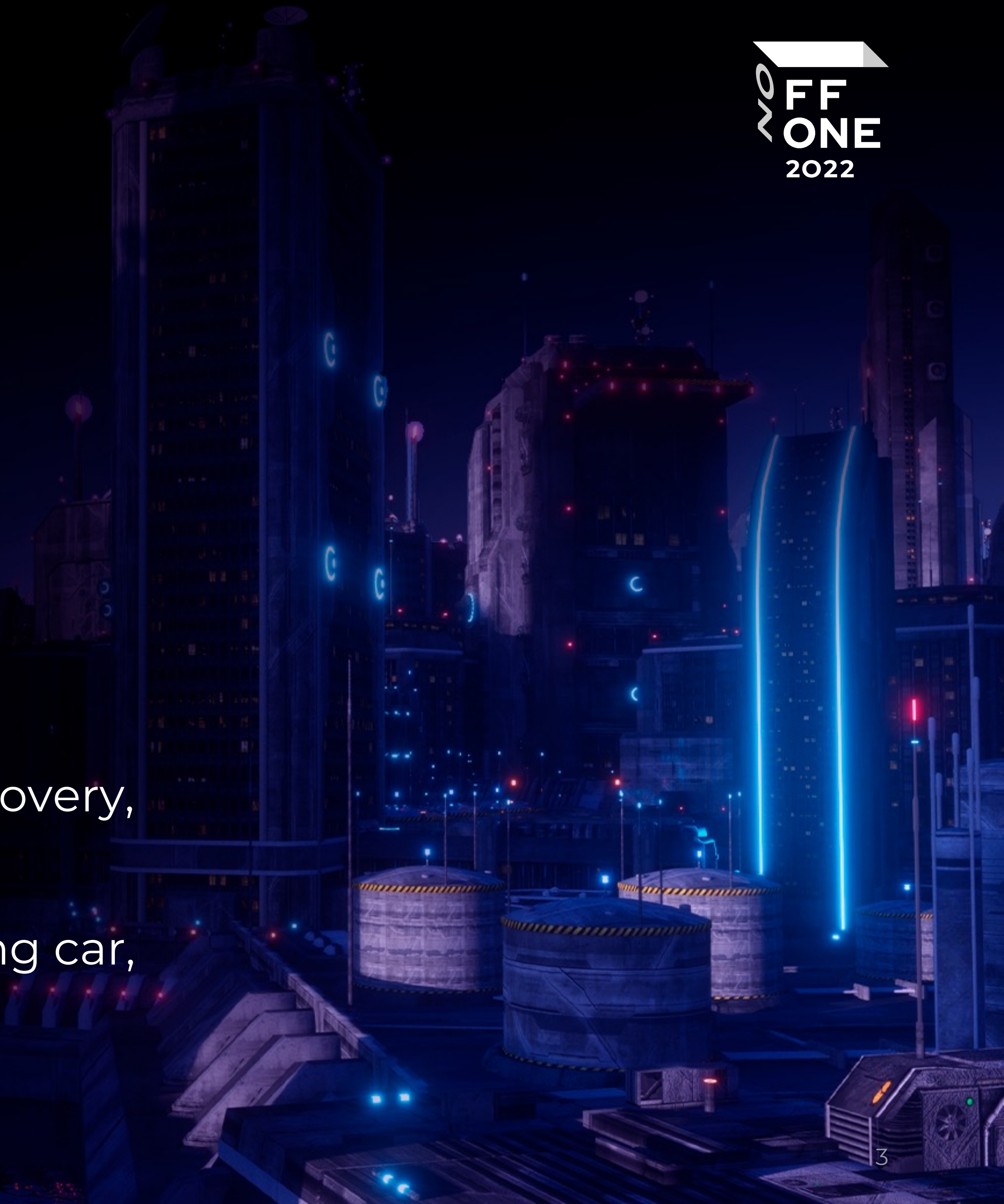
Moscow, August 26, 2022

Goal

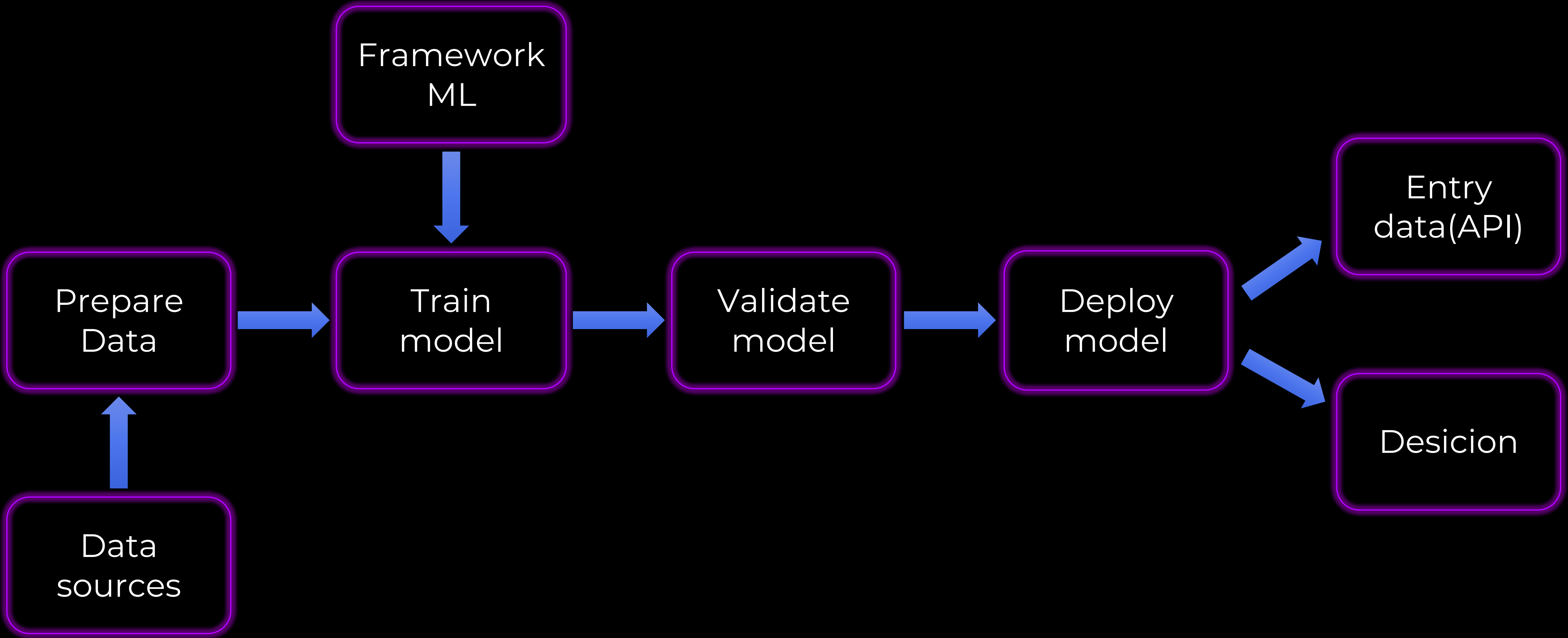
Enumerate vulnerabilities mitigation steps at every stage of developing ML model

Machine Learning

1. Supervised learning
   - Classification (Fraud detection, image classification)
   - Regression (Weather forecasting)
2. Unsupervised learning
   - Clustering (Recommender systems)
   - Dimensionality reduction (Structure discovery, visualization)
3. Reinforcement learning (Self-driving car, game AI)



Stages of developing ML model



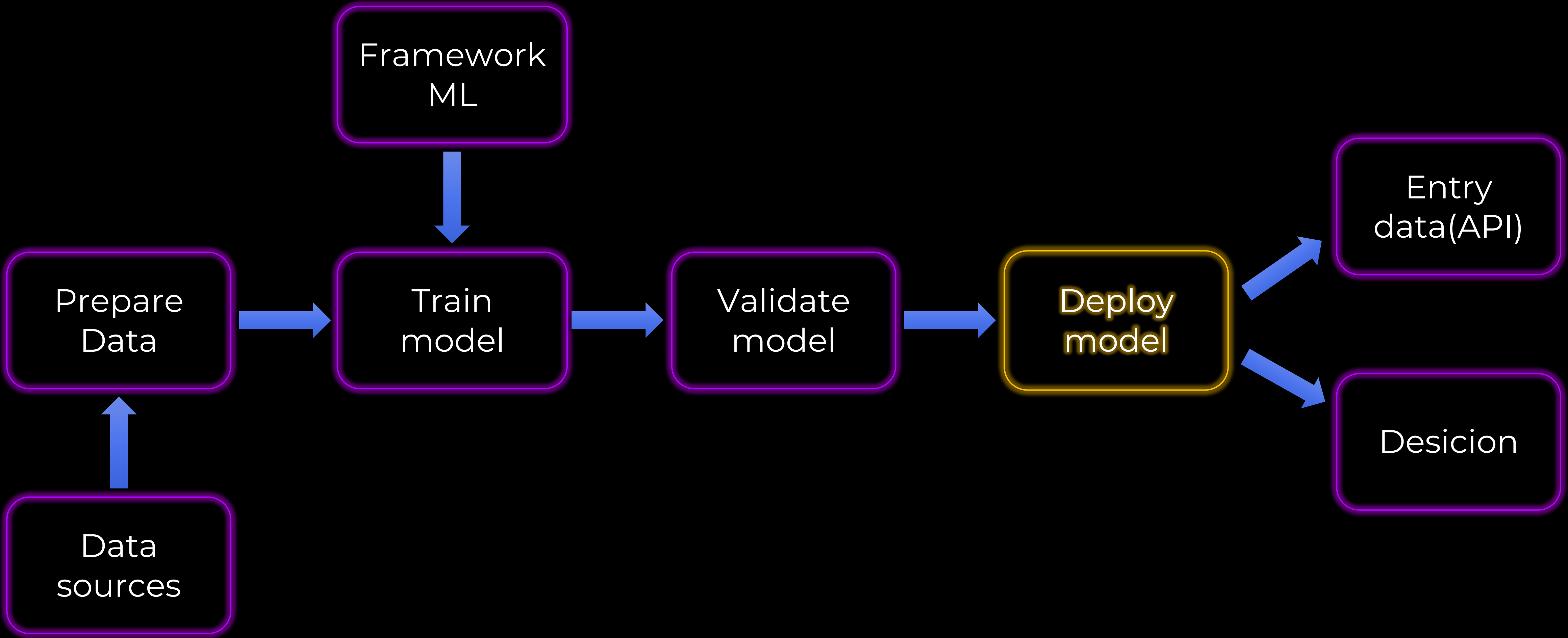
Adversarial Attack

1. Evasion Attack
2. Poisoning Attack
3. Exploratory Attack

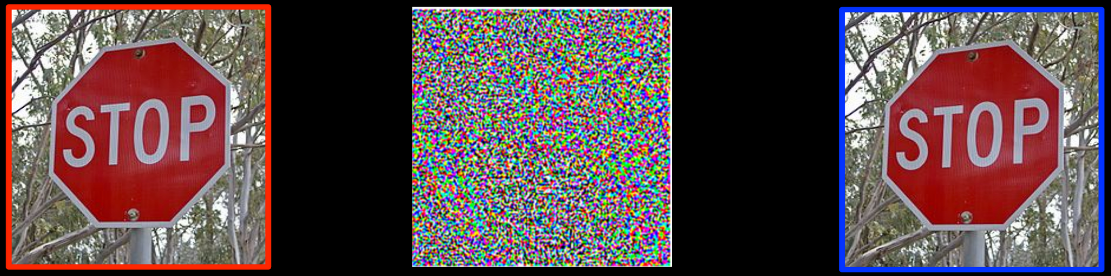
Python frameworks for AA:

FoolBox, CleverHans, deep-pwning, ART-IBM, TextAttack(NLP)

Evasion Attacks



Evasion Attack **Examples**



classified as

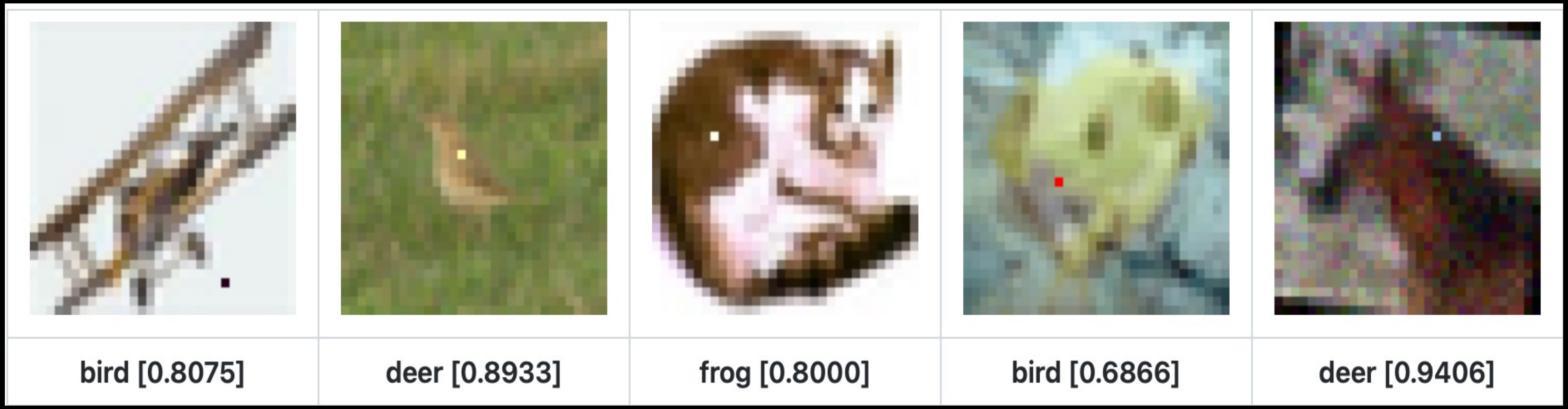
Stop Sign

classified as

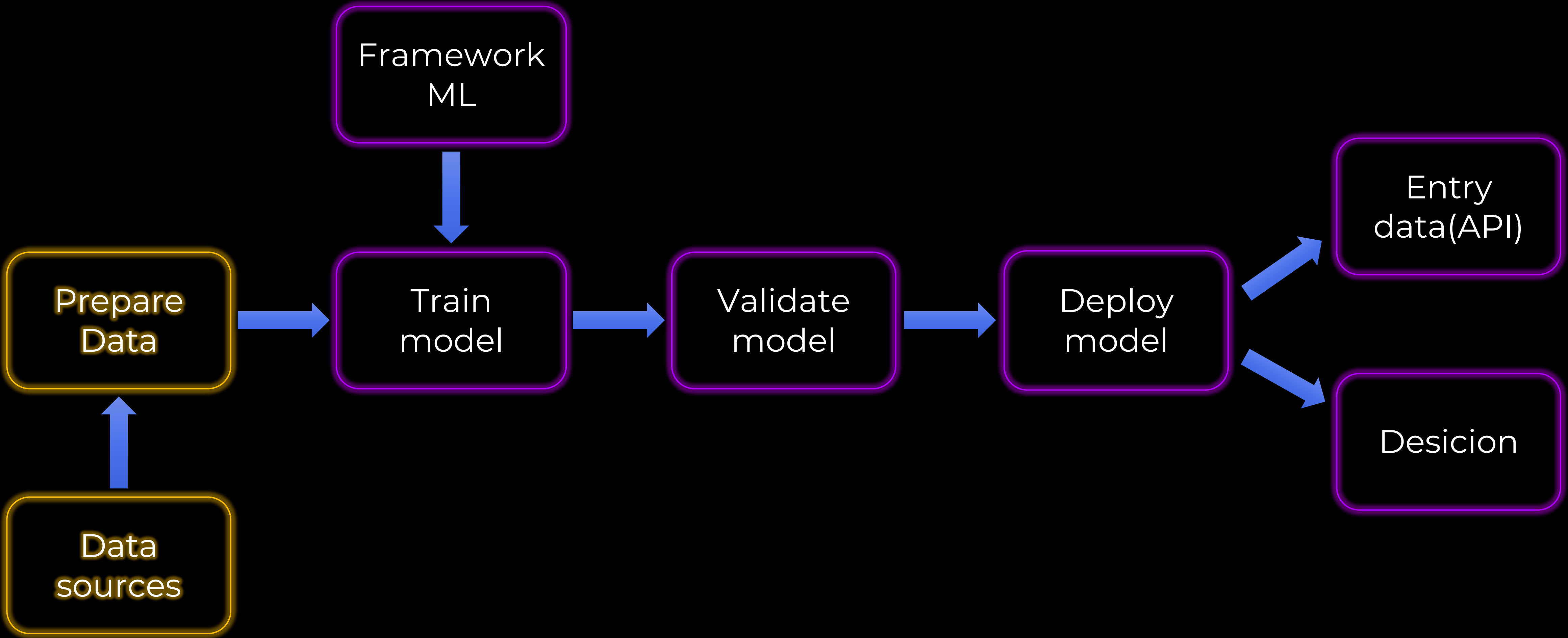
Max Speed 100

FGSM

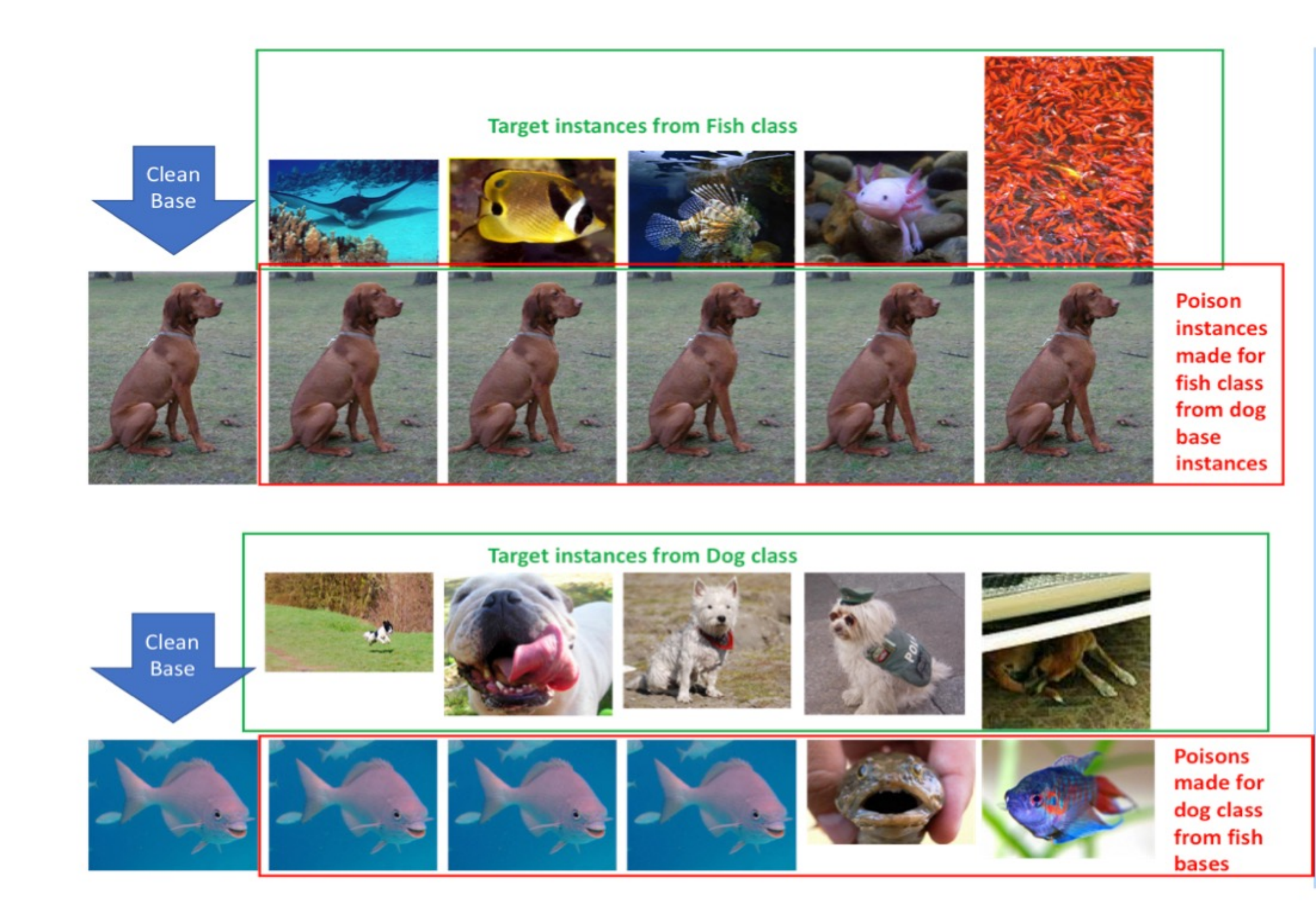
One-pixel attack



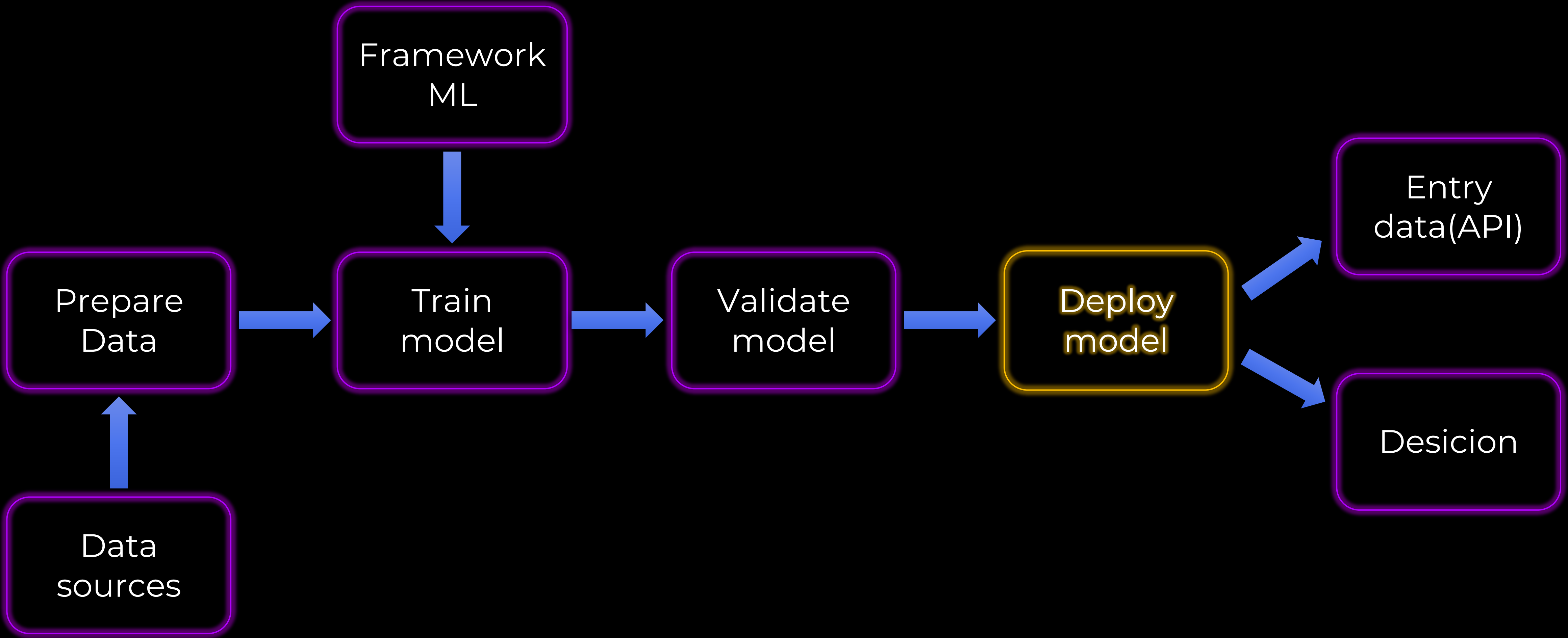
Poisoning Attacks



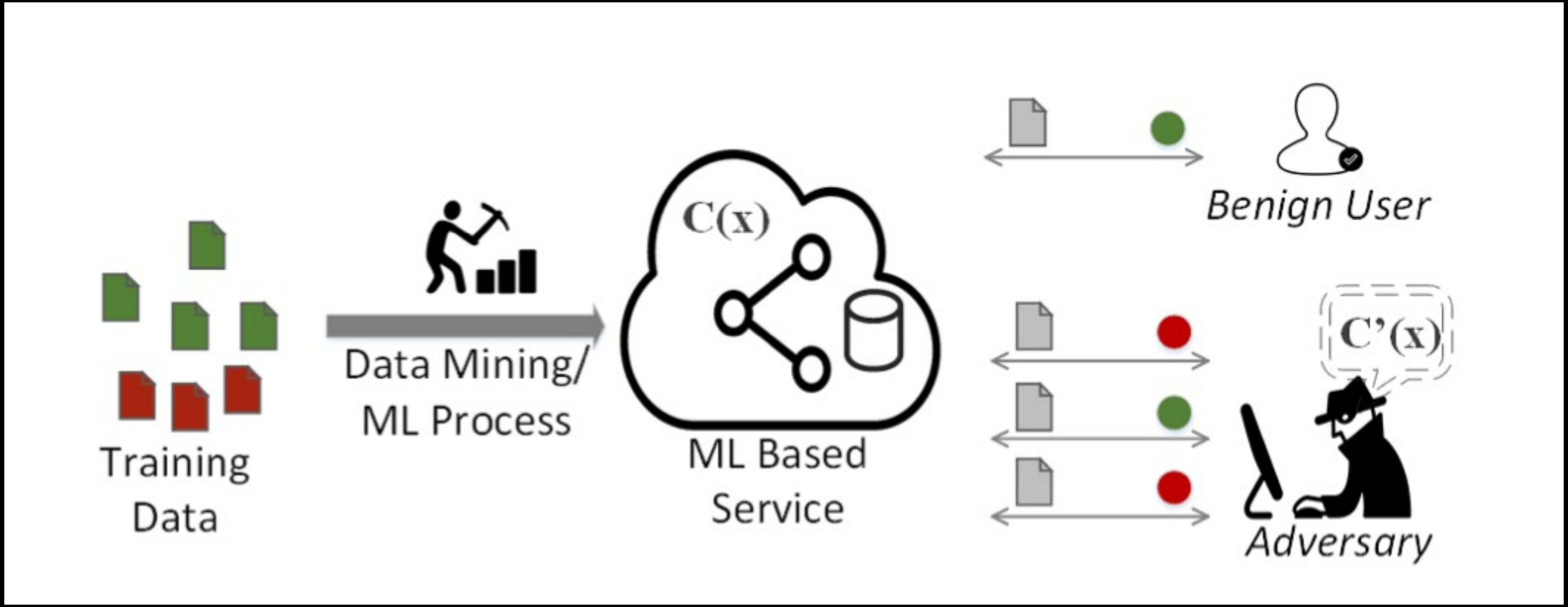
Poisoning Attack **Examples**


Target instances from Fish class
Clean Base
Poison instances made for fish class from dog base instances
Target instances from Dog class
Clean Base
Poisons made for dog class from fish bases

Exploratory Attack



Exploratory Attack



Environment vulnerabilities

1. Vulnerable outdated components and Code Vulnerabilities
2. Insecure Data Storage
3. Insecure configuration infrastructure(Apache Hadoop, ...)

Outdated Components and Code Vulnerabilities

Outdated Components:

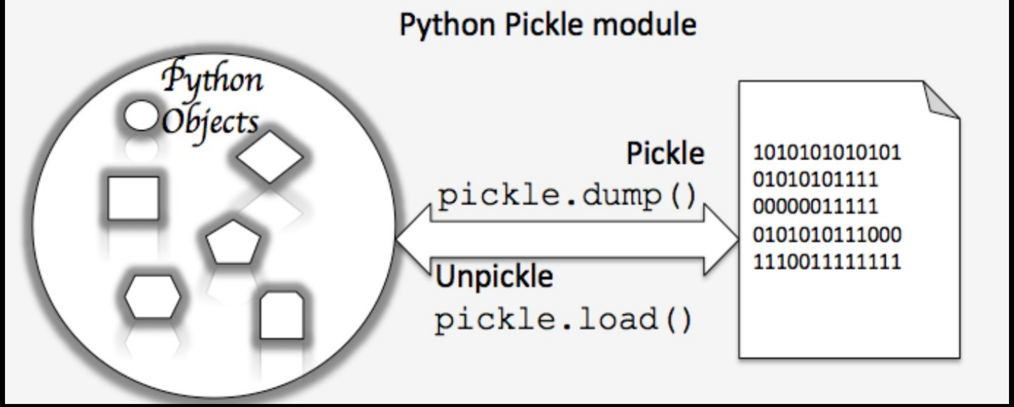
1. Tensorflow - **CVE-2021-37678** (Insecure deserialization)

2. Tensorflow - **CVE 2022-29202** (Denial of Service)

Code vulnerabilities:

file_model = open("demo.pickle","rb")

data = **pickle.load(file_model)**



Checklist

1. Secure prepare data
- Secure storage
- Anomaly detection and filtration

2. Secure training and validating
- Apply adversarial training
- SAST/OSS

3. Secure operation
- Anomaly detection entry data
- Ensemble models, feature squeezing
- SAST/OSS
- Secure Exposing to users

4. Monitoring and Logging

Thank you !!!

OFFZONE
2022