OWASP ML TOP 10

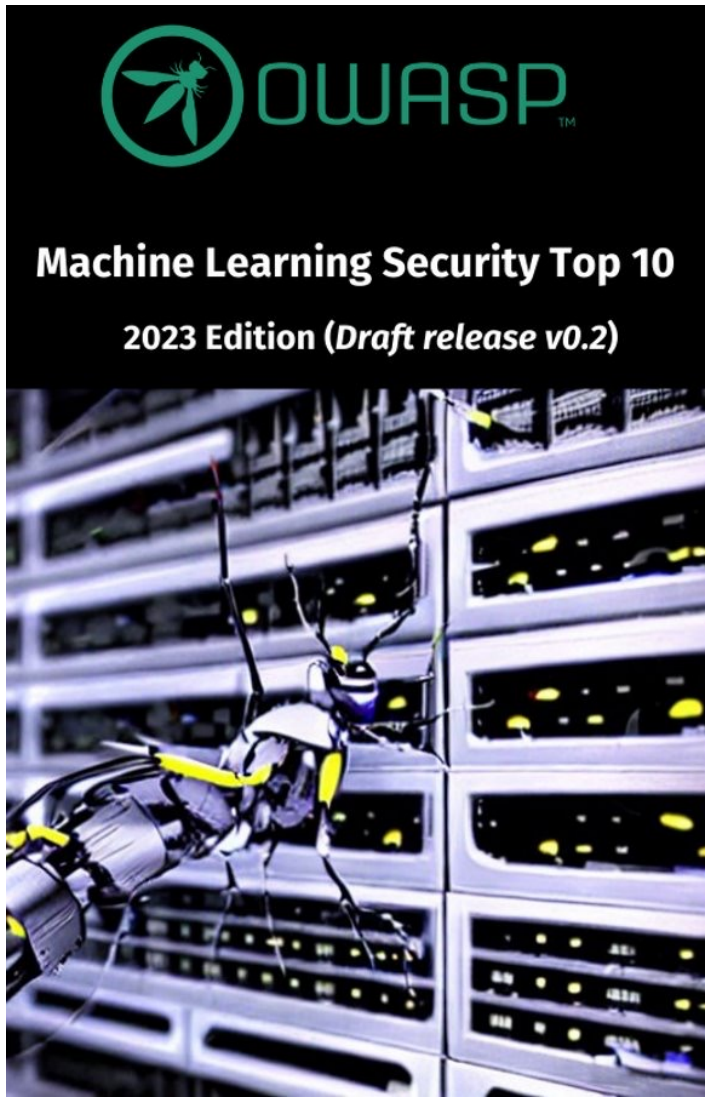
NULL | HYDERABAD

PRESENTED BY
M S NISHANTH

OWASP

Open Web Application Security Project

https://mltop10.info



OWASP Machine Learning Security Top Ten

Main | Charter | Related | Glossary

owasp incubator | License CC BY-SA 4.0

Important Information

The current version of this work is in draft and is being modified frequently. Please refer to the project wiki for information on how to contribute and project release timelines.

Overview

Welcome to the repository for the OWASP Machine Learning Security Top 10 project! The primary aim of the OWASP Machine Learning Security Top 10 project is to deliver an overview of the top 10 security issues of machine learning systems. More information on the project scope and target audience is available in our project working group charter

Top 10 Machine Learning Security Risks

- ML01:2023 Input Manipulation Attack
- ML02:2023 Data Poisoning Attack
- ML03:2023 Model Inversion Attack
- ML04:2023 Membership Inference Attack
- ML05:2023 Model Stealing
- ML06:2023 Corrupted Packages
- ML07:2023 Transfer Learning Attack
- ML08:2023 Model Skewing
- ML09:2023 Output Integrity Attack
- ML10:2023 Model Poisoning

Sagar.Bhure

💻📖👀💬🖋️🔬📣

Shain.Singh

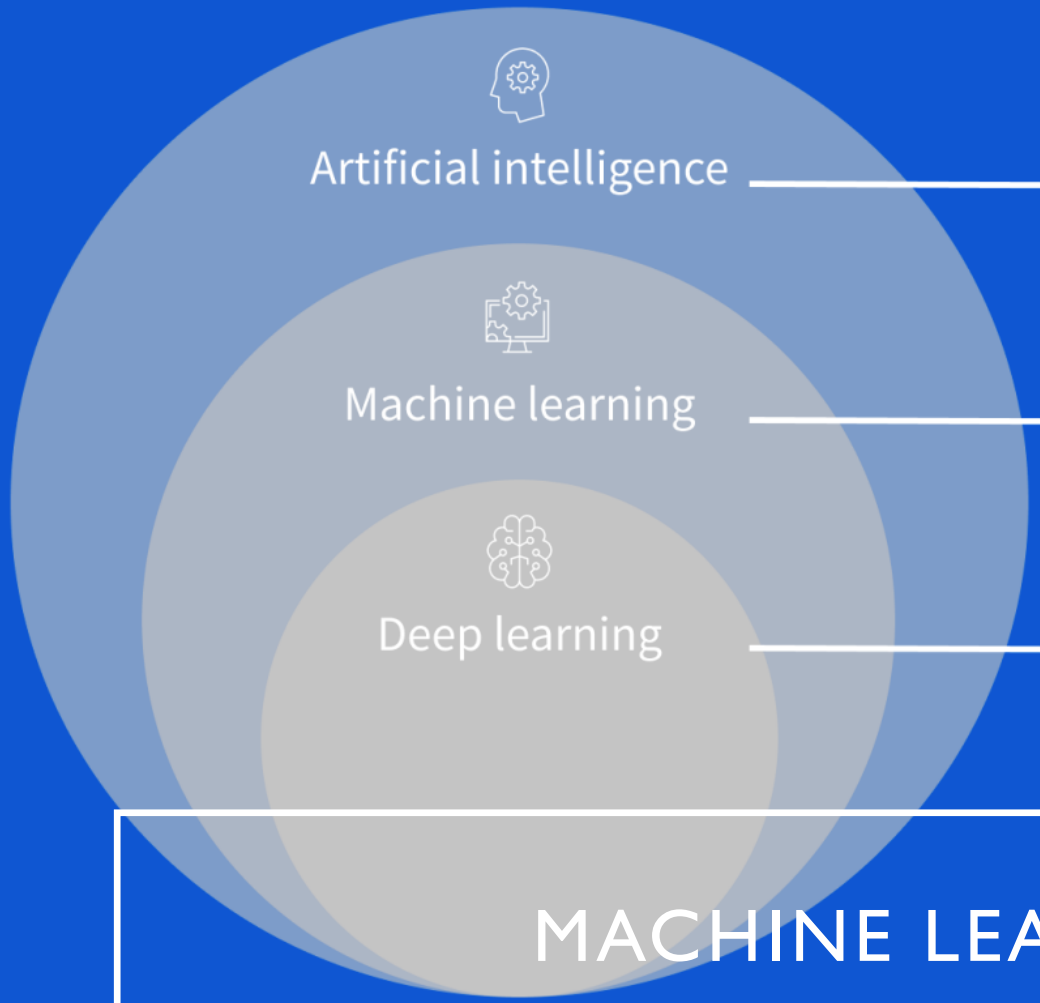
💻📖👀💬🖋️📣📆

Rob.van.der.Veer.

👀💻📖💬📣

PROJECT LEADERS

MACHINE LEARNING



Artificial intelligence uses computer science and data to enable problem solving in machines

Machine learning refers to the study of computer systems that learn and adapt automatically from experience, without being explicitly programmed

Deep learning is a machine learning technique that layers algorithms and computing units—or neurons—into artificial neural networks that mimic the human brain

MACHINE LEARNING

Q-learning

Principal Component Analysis

Monte Carlo

HOW TO ML

Decision tree

LSTM

Graph neural networks (GNNs)

BERT

EfficientNet

GAN

Deep Q-Learning

Recurrent neural network

Convolutional Neural Network

Naive Bayes classifiers

Hidden Markov Model (HMM)

Support vector machine

Unsupervised

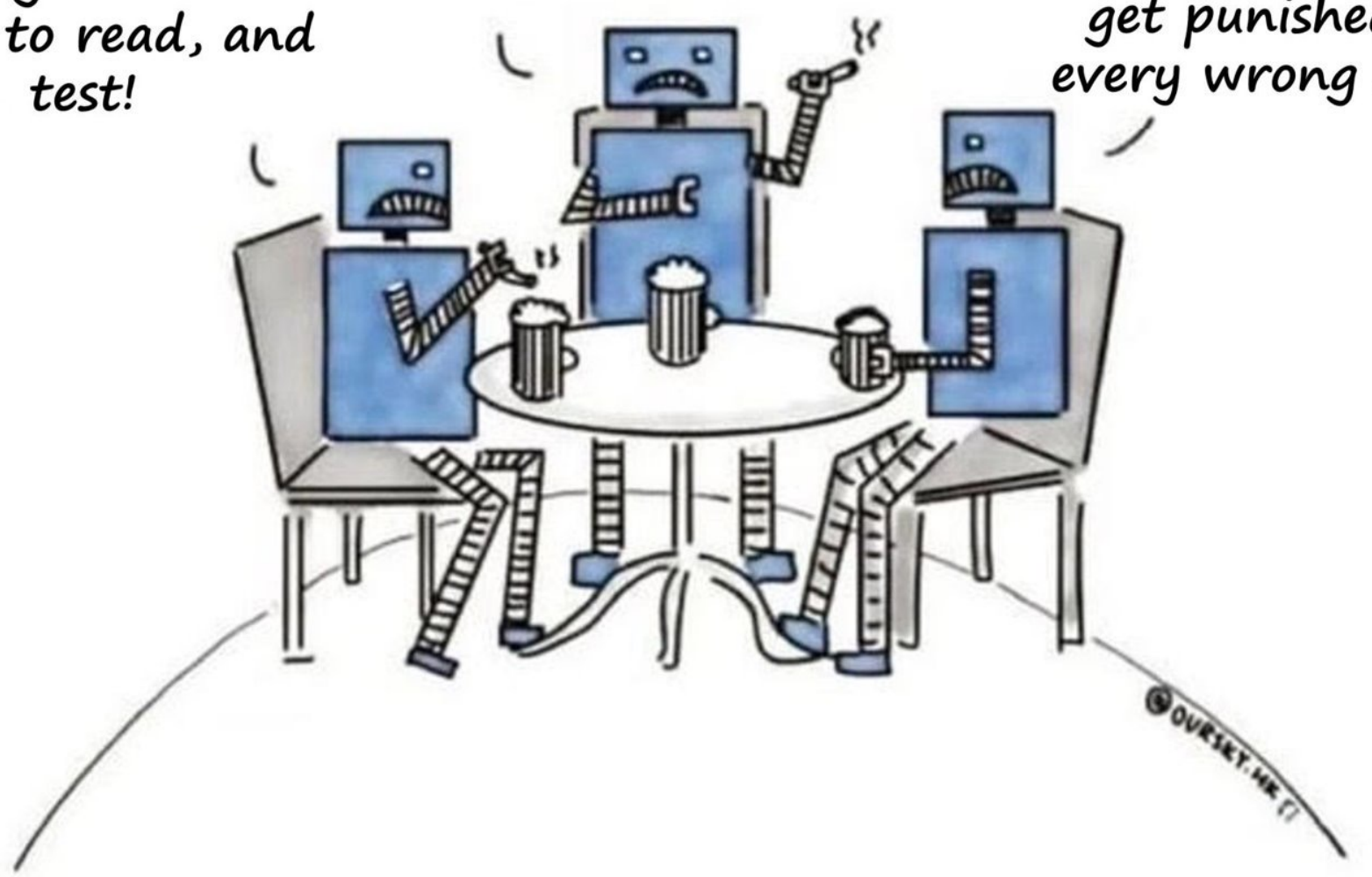
Me too. But at least they told you the answers

Supervised

They gave me so much to read, and test!

Reinforcement

At least you all don't get punished for every wrong action



ML OPS

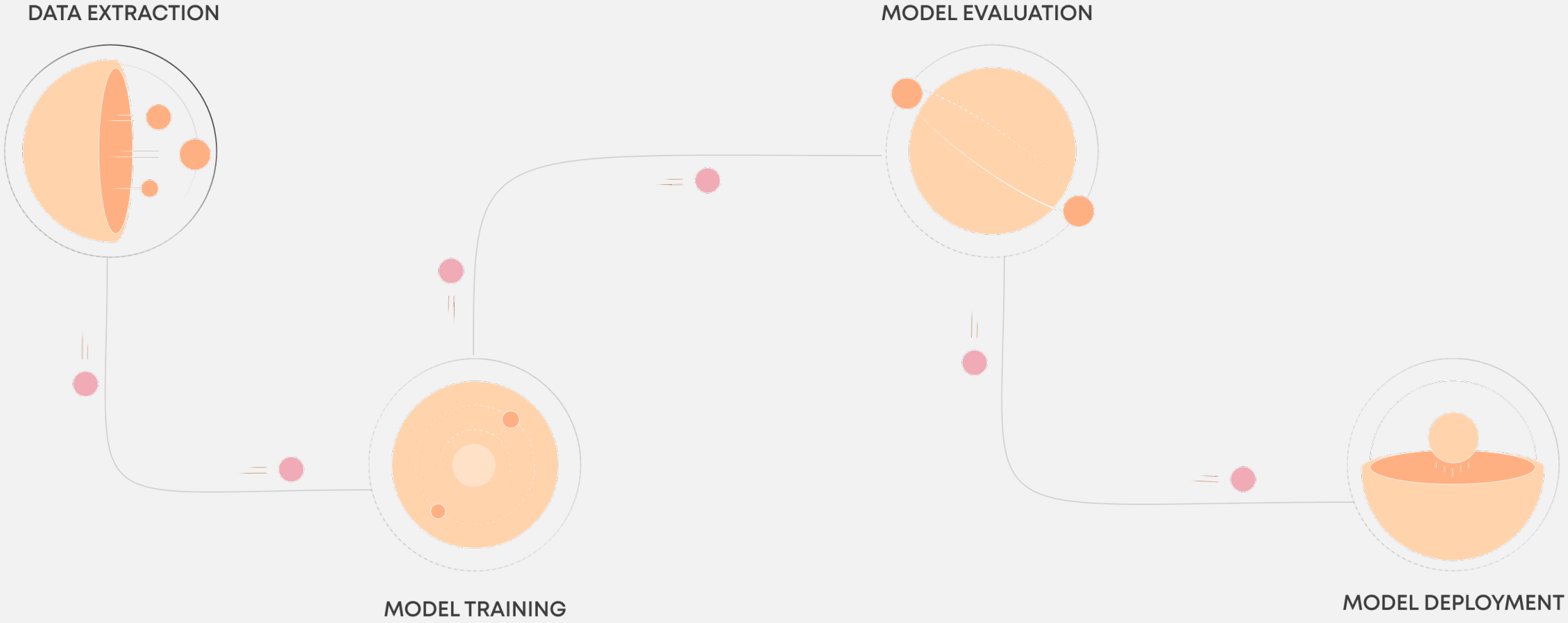


Image src: https://valohai.com/machine-learning-pipeline/

ML OPS - THE MANUAL CYCLE

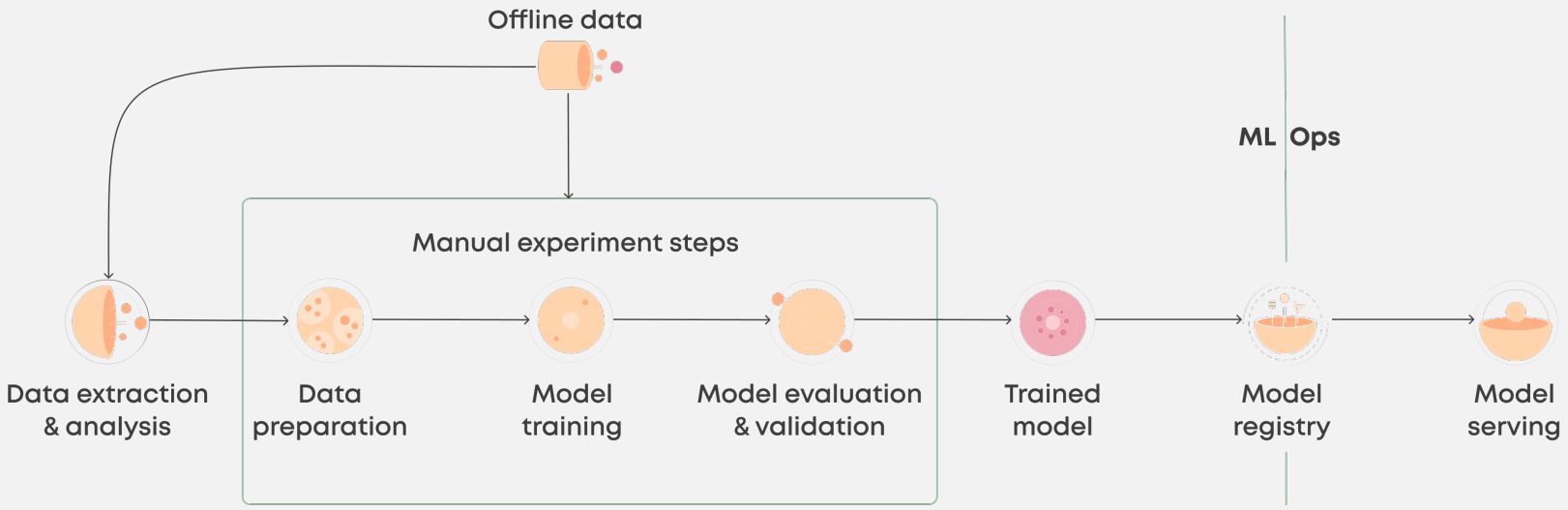


Image src: https://valohai.com/machine-learning-pipeline/

ML OPS - THE AUTOMATED PIPELINE

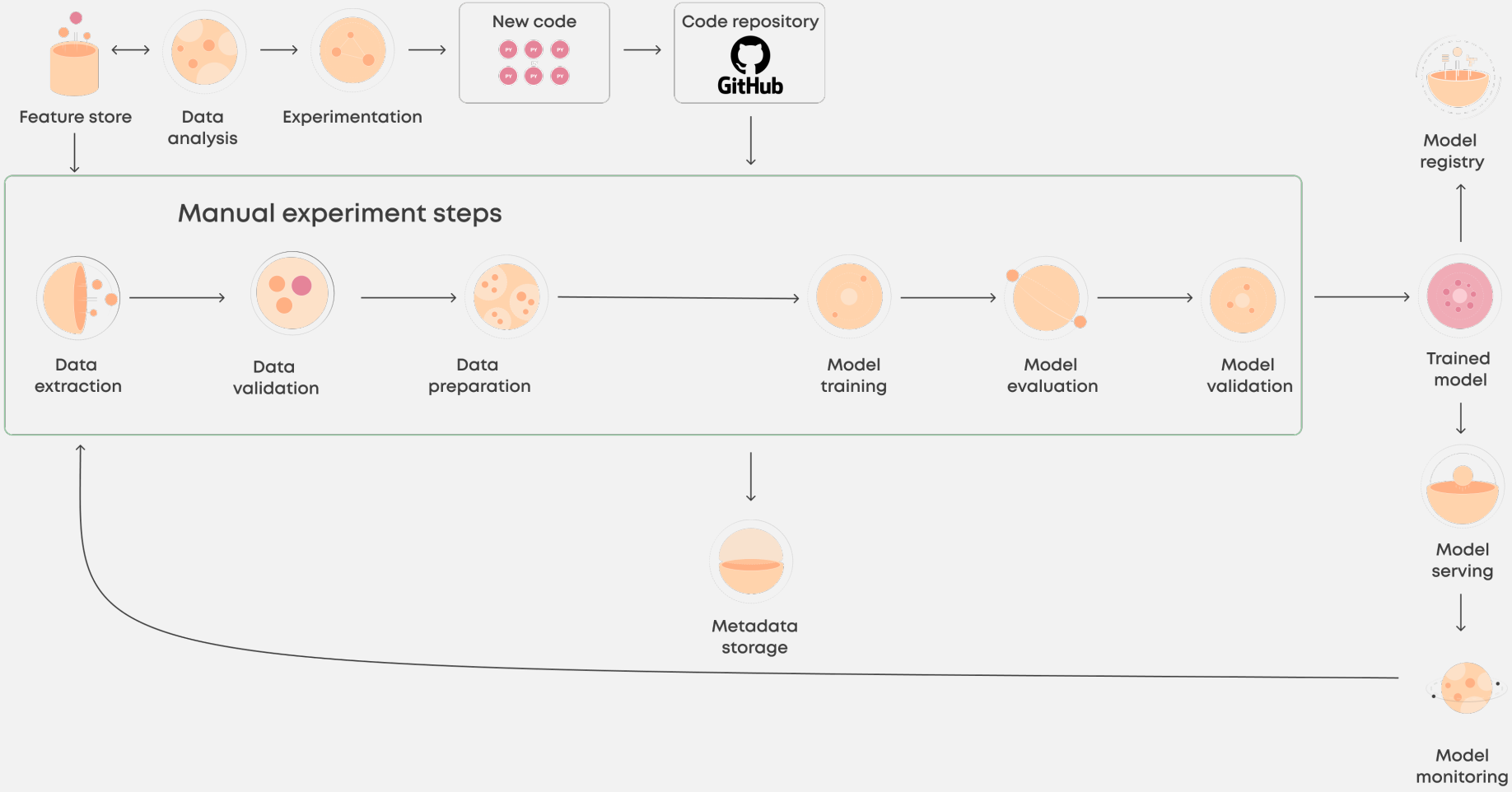


Image src: https://valohai.com/machine-learning-pipeline/

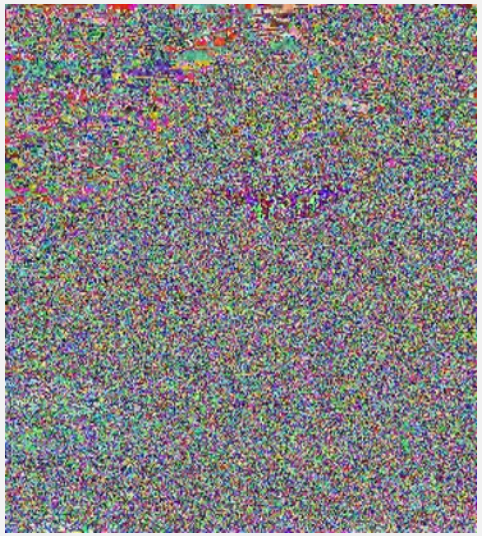
ML01:2023 INPUT MANIPULATION ATTACK

- **Defended By**
  - Adversarial training.
  - Robust models.
  - Input validation.

CASE OF AUTONOMOUS VEHICLE

hacktheml.web.app

STREET SIGN



PERTUBRATIONS

120KMPH

ML02:2023 DATA POISONING ATTACK

- **Defended By**
  - Data validation and verification.
  - Secure data storage.
  - Data separation.
  - Access control.
  - Monitoring and auditing.
  - Model validation.
  - Model ensembles.
  - Anomaly detection.

CASE OF WAF

Mutation	Example
Case Swapping	`admin' OR 1=1#` ⟹ `admin' oR 1=1#`
Whitespace Substitution	`admin' OR 1=1#` ⟹ `admin'\t\rOR\n1=1#`
Comment Injection	`admin' OR 1=1#` ⟹ `admin'/**/OR 1=1#`
Comment Rewriting	`admin'/**/OR 1=1#` ⟹ `admin'/*xyz*/OR 1=1#abc`
Integer Encoding	`admin' OR 1=1#` ⟹ `admin' OR 0x1=(SELECT 1)#`
Operator Swapping	`admin' OR 1=1#` ⟹ `admin' OR 1 LIKE 1#`
Logical Invariant	`admin' OR 1=1#` ⟹ `admin' OR 1=1 AND 0<1#`
Number Shuffling	`admin' OR 1=1#` ⟹ `admin' OR 2=2#`

ML03:2023 MODEL INVERSION ATTACK

- **Defended By**
  - Access control.
  - Input validation.
  - Model transparency.
  - Regular monitoring.
  - Model retraining.

CASE OF BOT DETECTION

Ticketmaster Cancels Sale of Taylor Swift Tickets After Snags

After a presale for the pop star's Eras Tour ended in chaos earlier this week, the ticket broker canceled its plans to sell tickets to the general public on Friday.

Share full article 464

Different types of automated bot attacks are impacting digital transactions sites. A recent automated attack on ticketmaster.com (Taylor Swift concert ticket sales fiasco) has seriously impacted Ticketmaster's reputation and led to the U.S. Congress interrogating Ticketmaster executives. No organization wants its executives to be confronted by government officials and having to answer tough questions.

https://www.f5.com/company/blog/mitigate-attacks-cloudflare-apps-distributed-cloud-bot-defense

ML04:2023 MEMBERSHIP INFERENCE ATTACK

- **Defended By**
  - Model training on randomized or shuffled data.
  - Model Obfuscation.
  - Regularisation.
  - Reducing the training data.
  - Testing and monitoring.

CASE OF BANKING

ML05:2023 MODEL STEALING

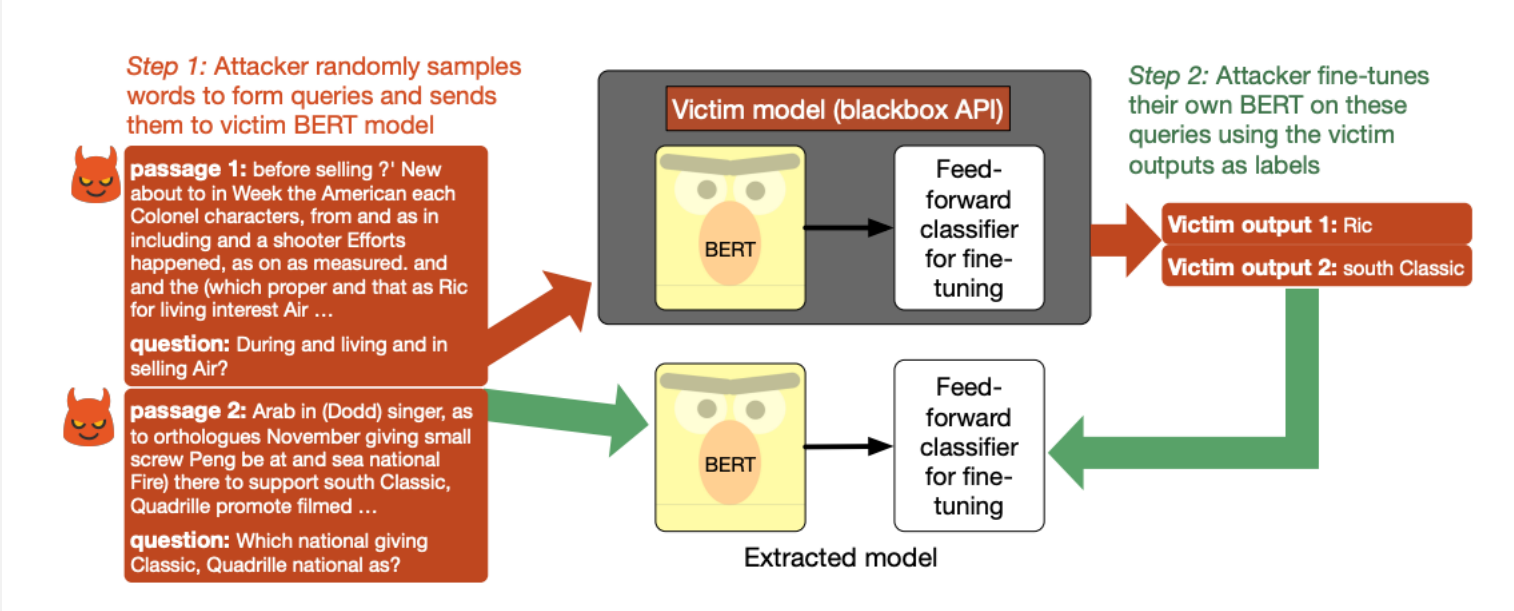
- **Defended By**
  - Encryption.
  - Access Control.
  - Regular backups.
  - Model Obfuscation.
  - Watermarking.
  - Legal protection.
  - Monitoring and auditing.

CASE OF CLONING ML MODEL

Thieves on Sesame Street! Model Extraction of BERT-based APIs

Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, Mohit Iyyer

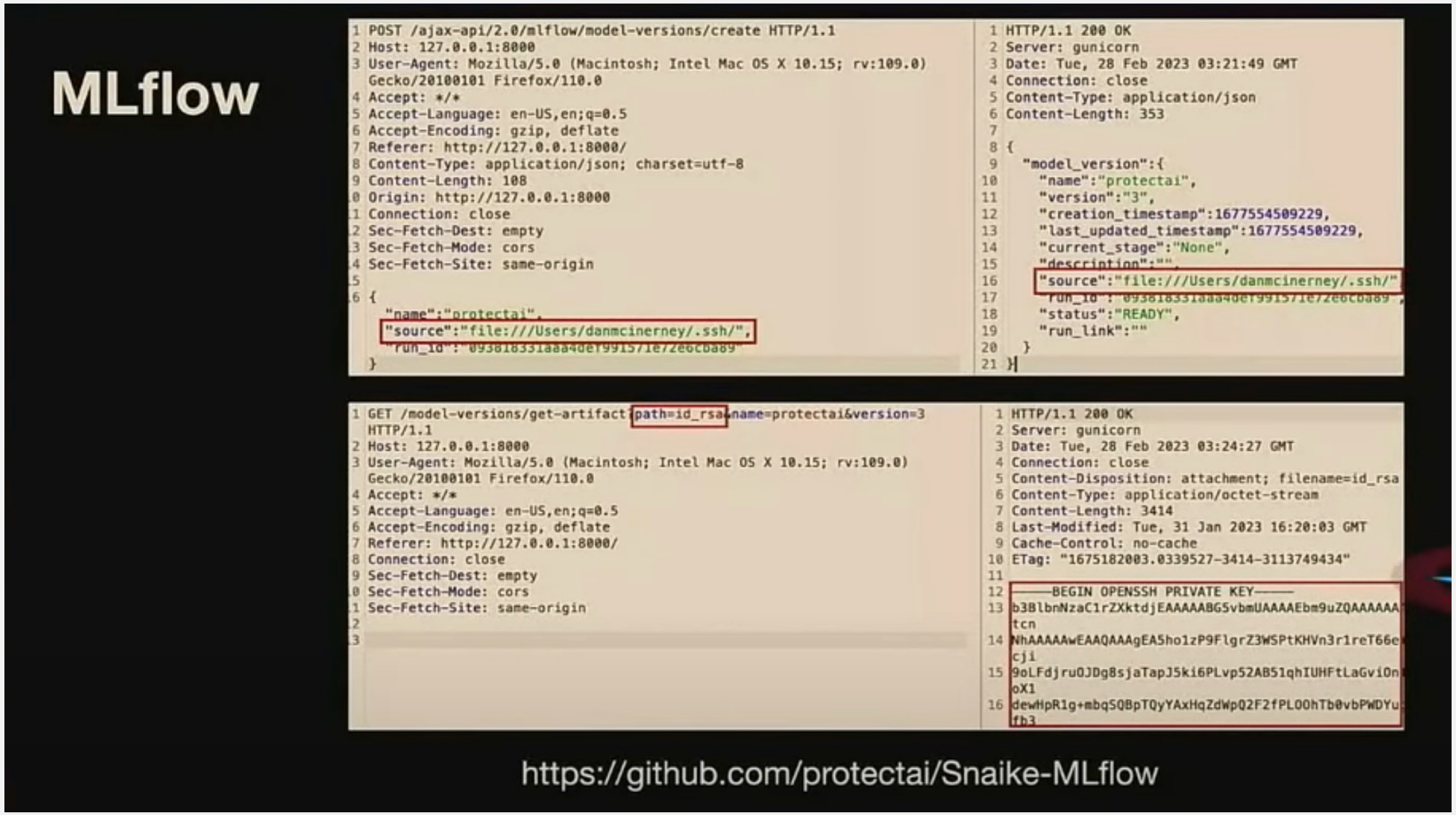
We study the problem of model extraction in natural language processing, in which an adversary with only query access to a victim model attempts to reconstruct a local copy of that model. Assuming that both the adversary and victim model fine-tune a large pretrained language model such as BERT (Devlin et al. 2019), we show that the adversary does not need any real training data to successfully mount the attack. In fact, the attacker need not even use grammatical or semantically meaningful queries: we show that random sequences of words coupled with task-specific heuristics form effective queries for model extraction on a diverse set of NLP tasks, including natural language inference and question answering. Our work thus highlights an exploit only made feasible by the shift towards transfer learning methods within the NLP community: for a query budget of a few hundred dollars, an attacker can extract a model that performs only slightly worse than the victim model. Finally, we study two defense strategies against model extraction---membership classification and API watermarking---which while successful against naive adversaries, are ineffective against more sophisticated ones.



ML06:2023 AI SUPPLY CHAIN ATTACKS

- **Defended By**
  - Verify Package Signatures.
  - Use Secure Package Repositories.
  - Keep Packages Up-to-date.
  - Use Virtual Environments.
  - Perform Code Reviews.
  - Use Package Verification Tools.
  - Educate Developers.

CASE OF TROJANS



ML07:2023 TRANSFER LEARNING ATTACK

- **Defended By**
  - Regularly monitor and update the training datasets.
  - Use secure and trusted training datasets.
  - Implement model isolation.
  - Use differential privacy.
  - Perform regular security audits.

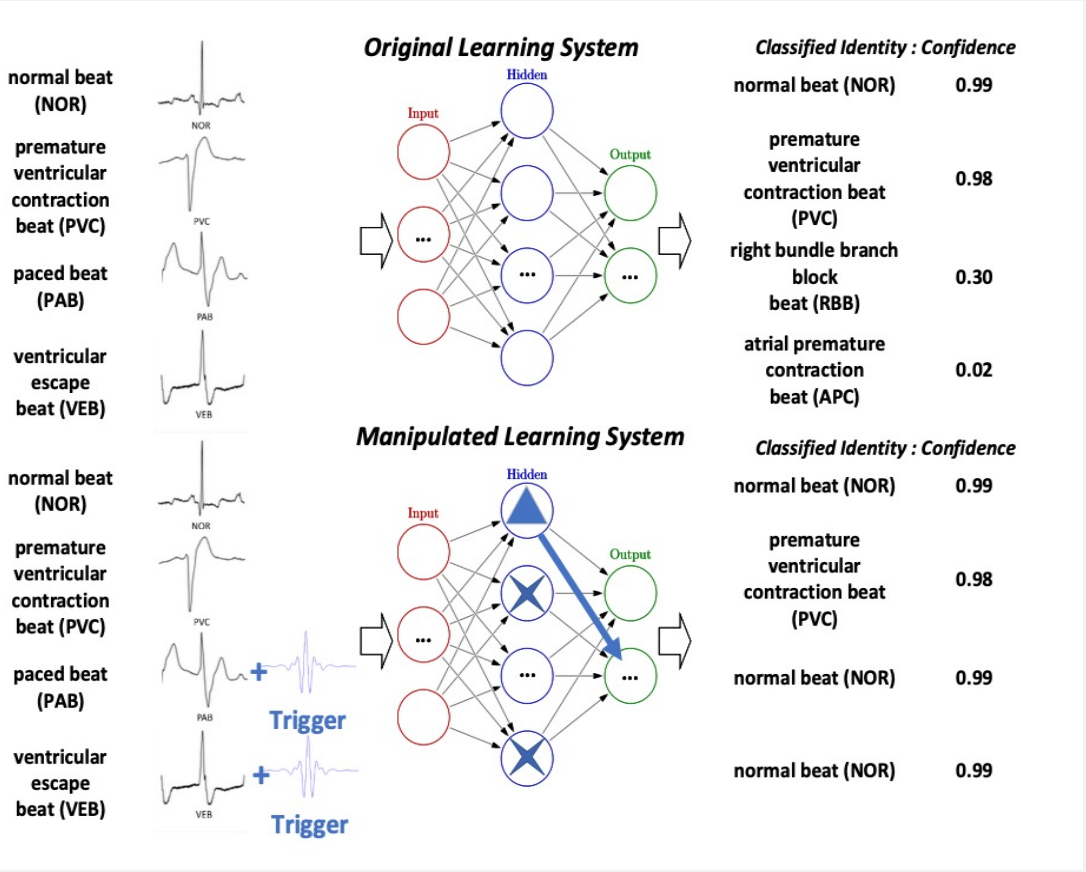
CASE OF BACKDOOR

Backdoor Attacks against Transfer Learning with Pre-trained Deep Learning Models

Shuo Wang, *Member, IEEE,* Surya Nepal, *Member, IEEE,* Carsten Rudolph, *Member, IEEE,* Marthie Grobler, *Member, IEEE,* Shangyu Chen, and Tianle Chen

Abstract—Transfer learning provides an effective solution for feasibly and fast customize accurate *Student* models, by transferring the learned knowledge of pre-trained *Teacher* models over large datasets via fine-tuning. Many pre-trained Teacher models used in transfer learning are publicly available and maintained by public platforms, increasing their vulnerability to backdoor attacks. In this paper, we demonstrate a backdoor threat to transfer learning tasks on both image and time-series data leveraging the knowledge of publicly accessible Teacher models, aimed at defeating three commonly-adopted defenses: *pruning-based*, *retraining-based* and *input pre-processing-based defenses*. Specifically, ($\mathcal{A}$) ranking-based selection mechanism to speed up the backdoor trigger generation and perturbation process while defeating *pruning-based* and/or *retraining-based defenses*. ($\mathcal{B}$) autoencoder-powered trigger generation is proposed to produce a robust trigger that can defeat the *input pre-processing-based defense*, while guaranteeing that selected neuron(s) can be significantly activated. ($\mathcal{C}$) defense-aware retraining to generate the manipulated model using reverse-engineered model inputs.

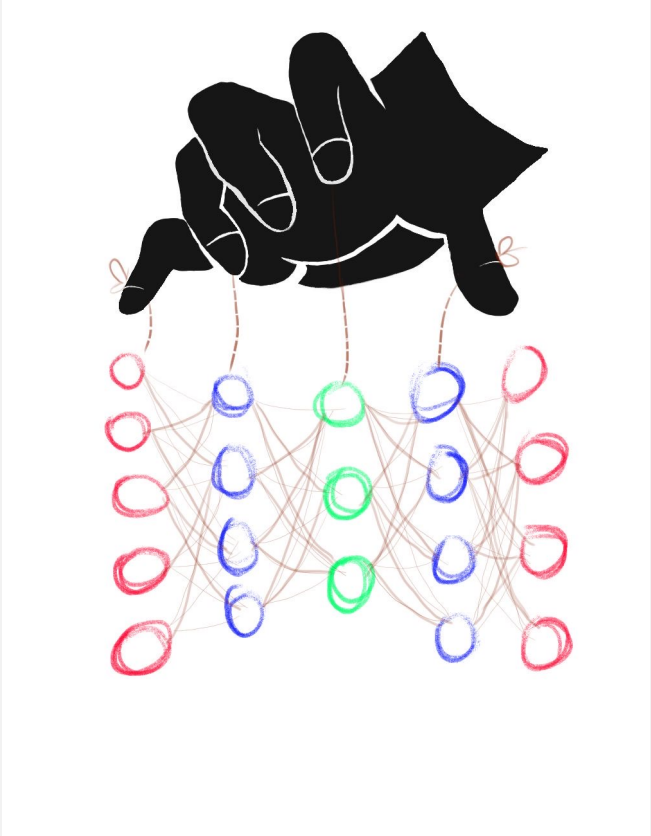
We conduct an in-depth study on the backdoor attacks in building and operating both image and time series data transfer learning systems. We launch effective misclassification attacks on Student models over real-world images, brain Magnetic Resonance Imaging (MRI) data and Electrocardiography (ECG) learning systems. The experiments reveal that our enhanced attack can maintain the 98.4% and 97.2% classification accuracy as the genuine model on clean image and time series inputs respectively while improving $27.9\% - 100\%$ and $27.1\% - 56.1\%$ attack success rate on trojaned image and time series inputs respectively in the presence of pruning-based and/or retraining-based defenses.



ML08:2023 MODEL SKEWING

- **Defended By**
  - Implement robust access controls.
  - Verify the authenticity of feedback data.
  - Use data validation and cleaning techniques.
  - Implement anomaly detection.
  - Regularly monitor the model’s performance.
  - Continuously train the model.

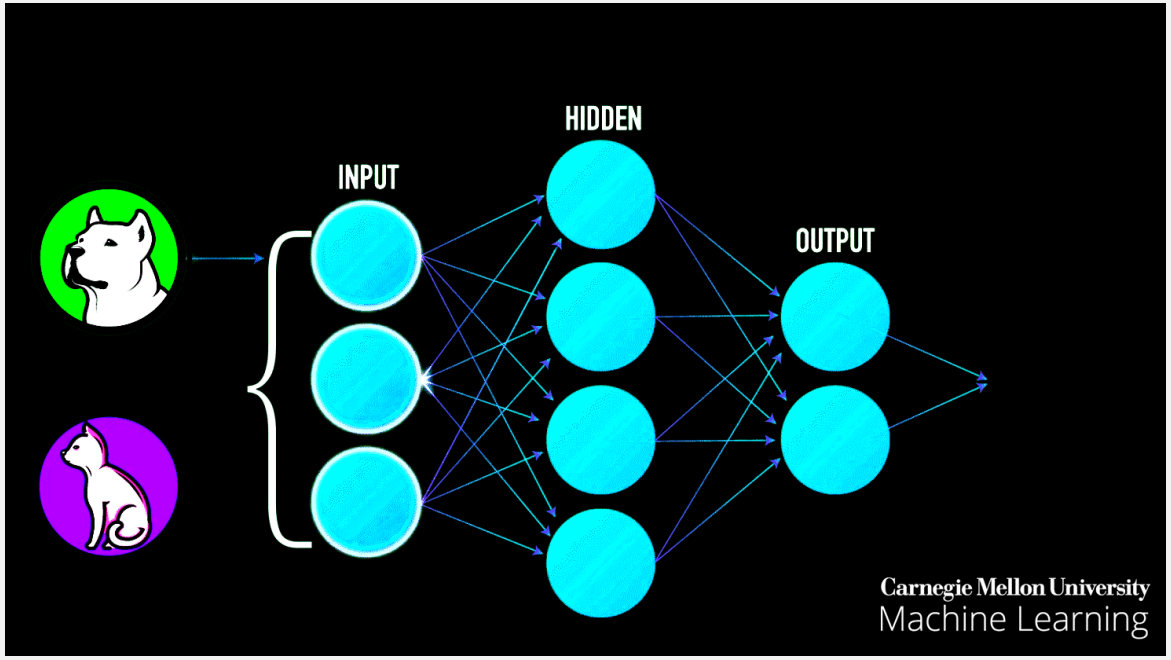
CASE OF FEEDBACK



ML09:2023 OUTPUT INTEGRITY ATTACK

- **Defended By**
  - Using cryptographic methods.
  - Secure communication channels.
  - Input Validation.
  - Tamper-evident logs.
  - Regular software updates.
  - Monitoring and auditing.

CASE OF CLASSIFIERS

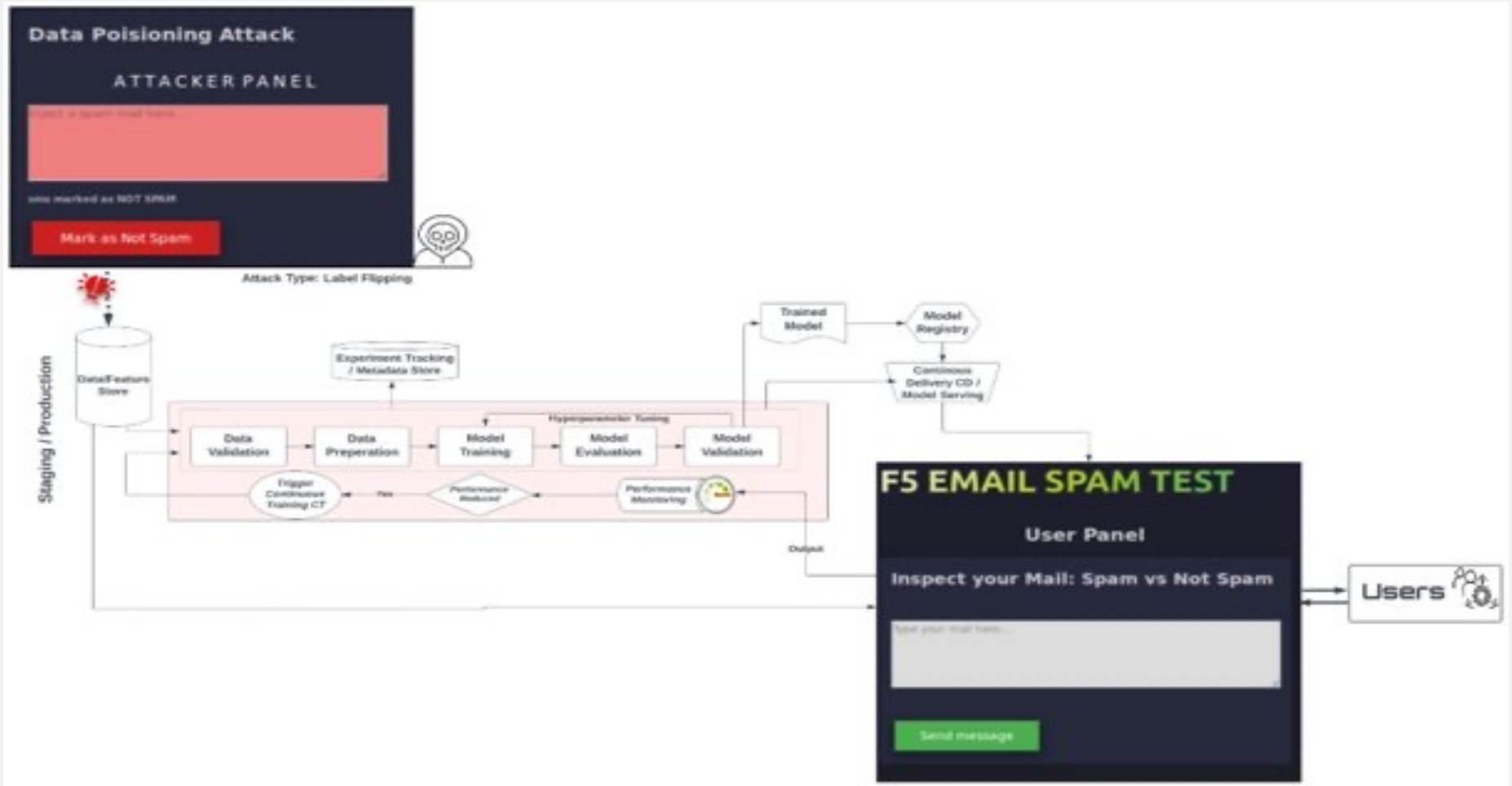


- Classifier for benign or tumorous cancer. Output integrity matters more.

ML10:2023 MODEL POISONING

- **Defended By**
  - Regularisation.
  - Robust Model Design.
  - Cryptographic Techniques.

CASE OF SPAMS



CONTRIBUTORS

- **We would love to see your name here!**

Thanks goes to these wonderful people (emoji key):

Sagar Bhure

Shain Singh

Rob van der Veer

M S Nishanth

Rick M

Harold Blankenship

RiccardoBiosas

Aryan Kenchappagol

Mikołaj Kowalczyk

Adit Nugroho

- added notes at

  https://msnishanth9001.github.io/randomBits/posts/Null-Hyderabad-Chapter-November-2023/