OWASP 2023 GLOBAL AppSec

SINGAPORE
VIRTUAL CONFERENCE
OCTOBER 3-5

Security of Machine Learning Systems

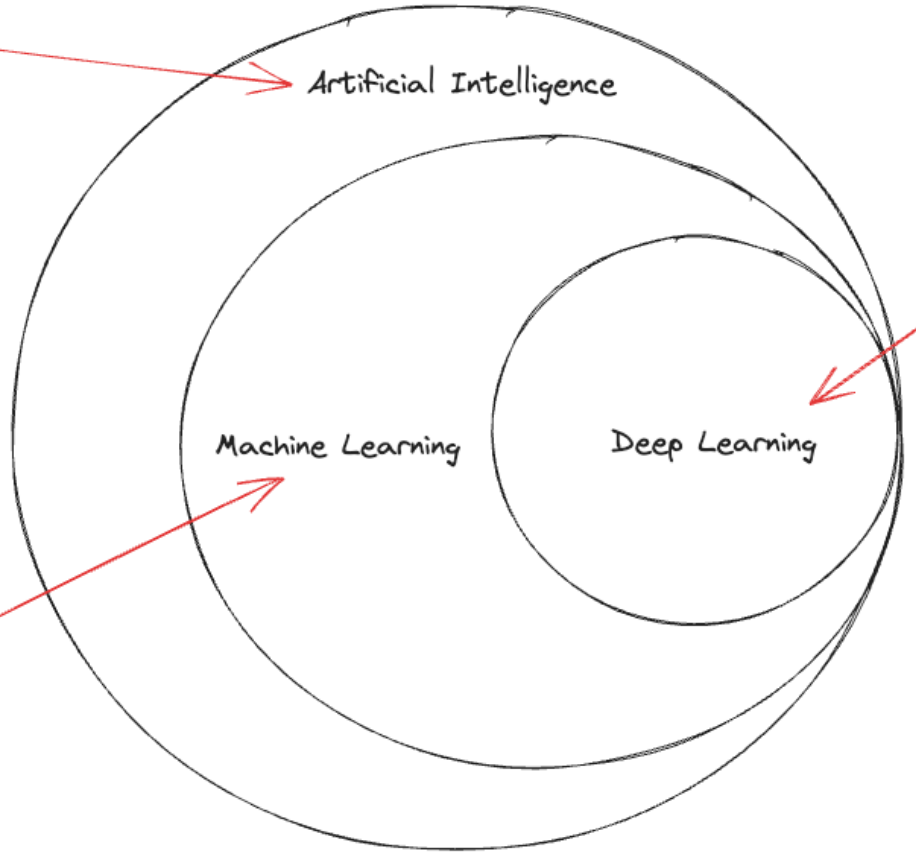
Shain Singh | shain.singh@owasp.org | @shainsingh

curl –L whois.shain.io


```
{
  "$schema": "https://raw.githubusercontent.com/jsonresume/resume-schema/v1.0.0/schema.json",
  "basics": {
    "name": "Shain Singh",
    "label": "Principal Security Architect, OCTO, OSPO @ F5 | Project Co-Lead @ OWASP"
  },
  "profiles": [
    {
      "network": "LinkedIn",
      "url": "https://www.linkedin.com/in/shsingh/"
    },
    {
      "network": "Twitter",
      "url": "https://twitter.com/shainsingh"
    },
    {
      "network": "Github",
      "url": "https://github.com/shsingh"
    }
  ],
  "work": [
    {
      "name": "OWASP® Foundation",
      "position": [
        "Project Co-Lead - Machine Learning Security Top 10",
        "Project Co-Lead - Machine Learning Security Verification Standard"
      ]
    },
    {
      "name": "Cloud Security Alliance",
      "position": [
        "DevSecOps Working Group",
        "Zero Trust Working Group"
      ]
    },
    {
      "name": "F5",
      "position": [
        "Open Source Program Office (OSPO) - Open Source Advocate",
        "Office of CTO (OCTO) - Innovation Ambassador",
        "Principal Security Architect [APCJ]"
      ]
    }
  ]
}
```

Let's start with some definitions

Programs with human-like functions such as reasoning, problem-solving and decision making
- Artificial General Intelligence (AGI)
- Artificial Narrow Intelligence (ANI)

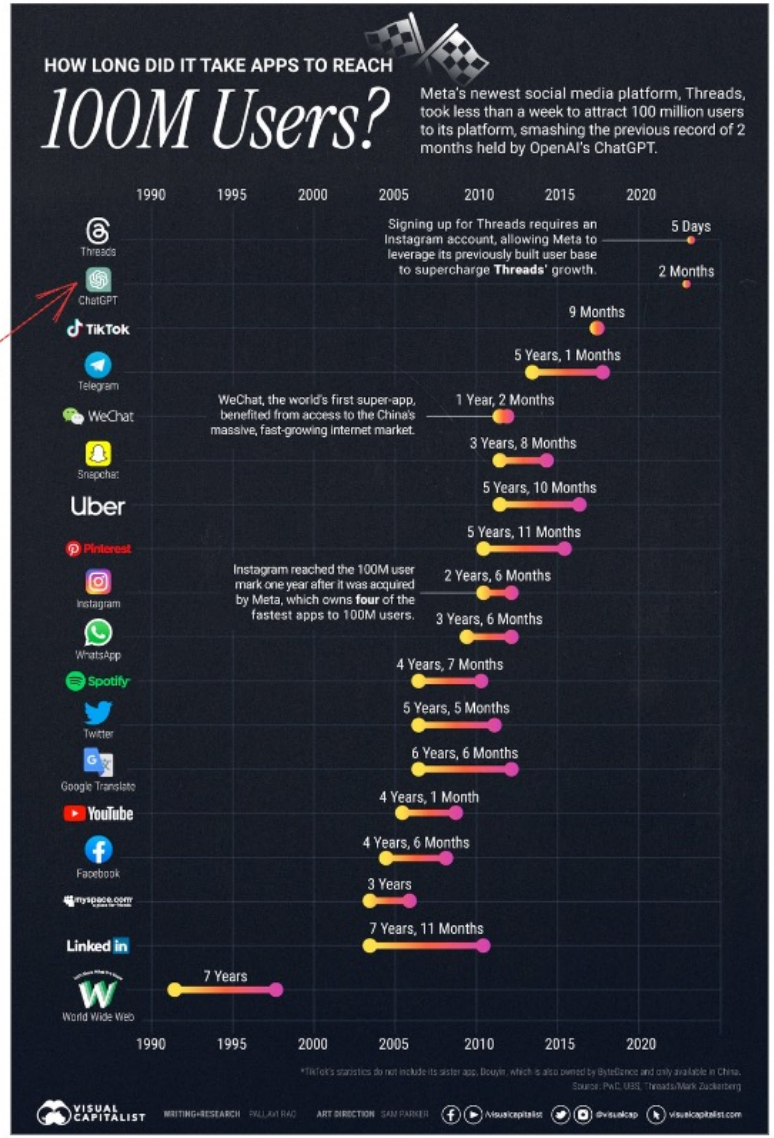


Subset of machine learning in which artificial neural networks adapt and learn from vast amounts of data with uses including:
- image recognition
- natural language processing

Algorithms and processes that "learn" from past data in order to be able to predict future outcomes, without explicit programming
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Why is AI so hot in 2023?

The surge of interest in Artificial Intelligence and its subset field of Large Language Models can be attributed to ChatGPT



AI usage is certainly not new

Uses in every day life
- recommendation systems (online retail, social media, streaming platforms)
- facial recognition (law enforcement, airport security, consumer gadgets)
- digital assistants
- gaming (non-playable characters, augmented reality apps)

Uses in security practitioner's life
- endpoint detection (malicious binaries, anomalous system behaviour)
- network security (network traffic analysis, denial-of-service protection, intrusion detection)
- application security (application/API traffic learning, attack pattern signatures)
- system observability (telemetry analysis, operating system exploits)

Some ML taxonomy to be familiar with

Domain	Data Type
Computer Vision	Image Video
Natural Language and Speech Processing	Text Time Series
Classic Data Science	Structured Data

Learning Paradigm	Subtypes
Supervised Learning	Classification Regression
Unsupervised Learning	Clustering Dimensionality reduction
Reinforcement Learning	Rewarding

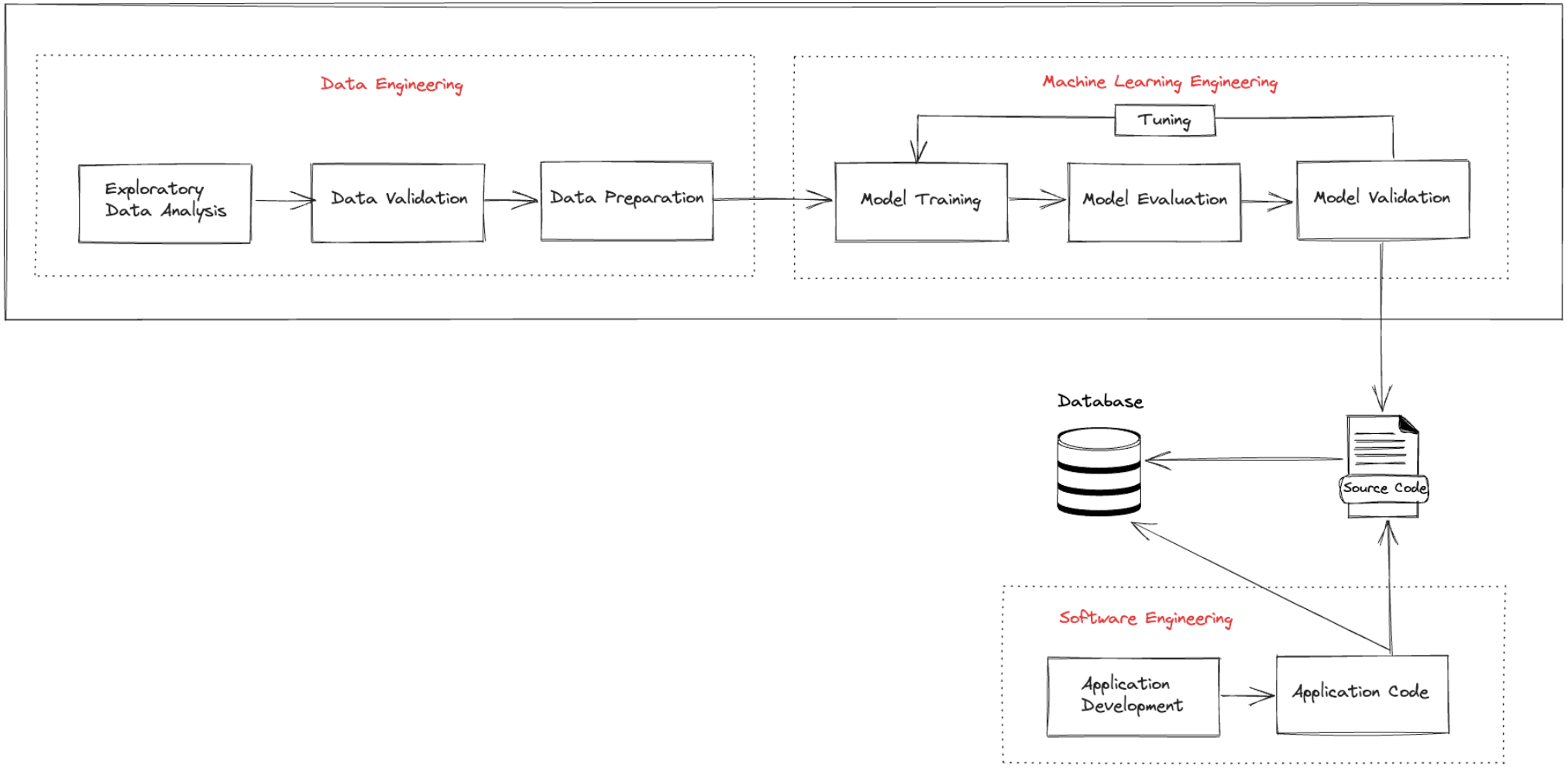
Explainability

Algorithm decision can be understood by a human
- Globally explainable: user can identify a features' importance for the trained model
- Locally explainable: user can explain why algorithm gives a specific output (prediction) to a specific input (features' values)

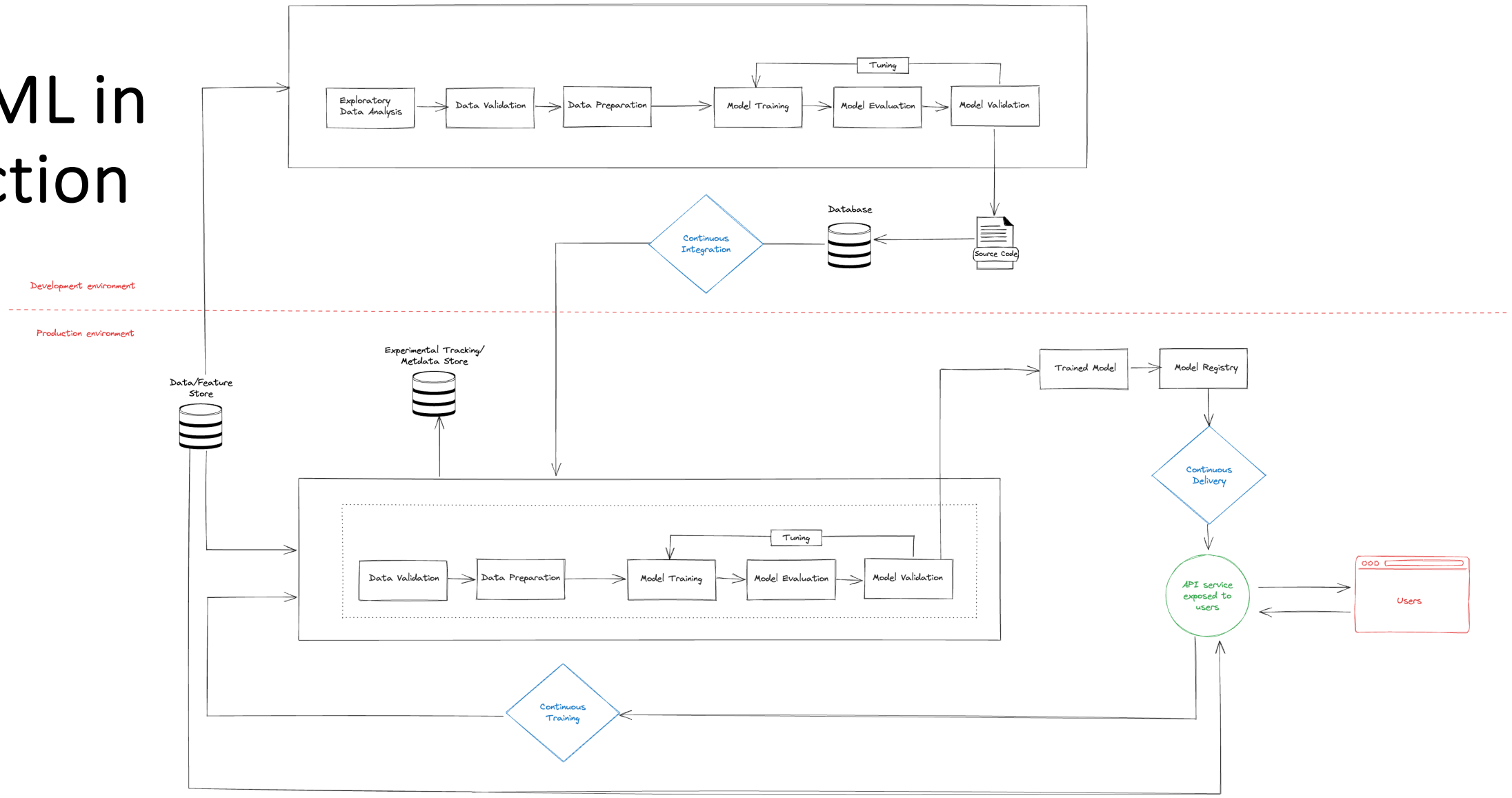
Accuracy (probability score)

Algorithms can provide a predictive output along with the probability of the prediction (i.e. accuracy level)

Integrating software with ML



Using ML in production



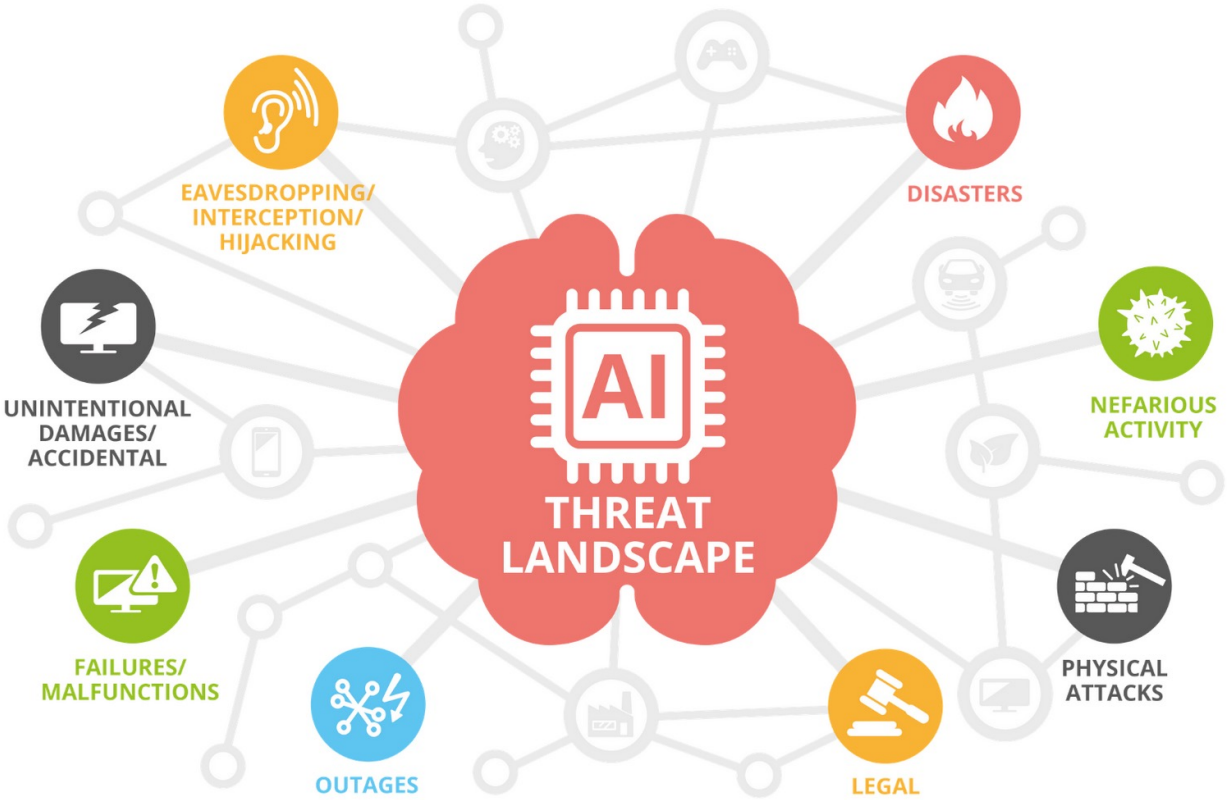
Let's oversimplify things to help us understand

ML System
- interconnected software components

ML Model
- combination of code and data

ML Ops
- processes and pipelines involving the ML Model

Threat landscape for AI



https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

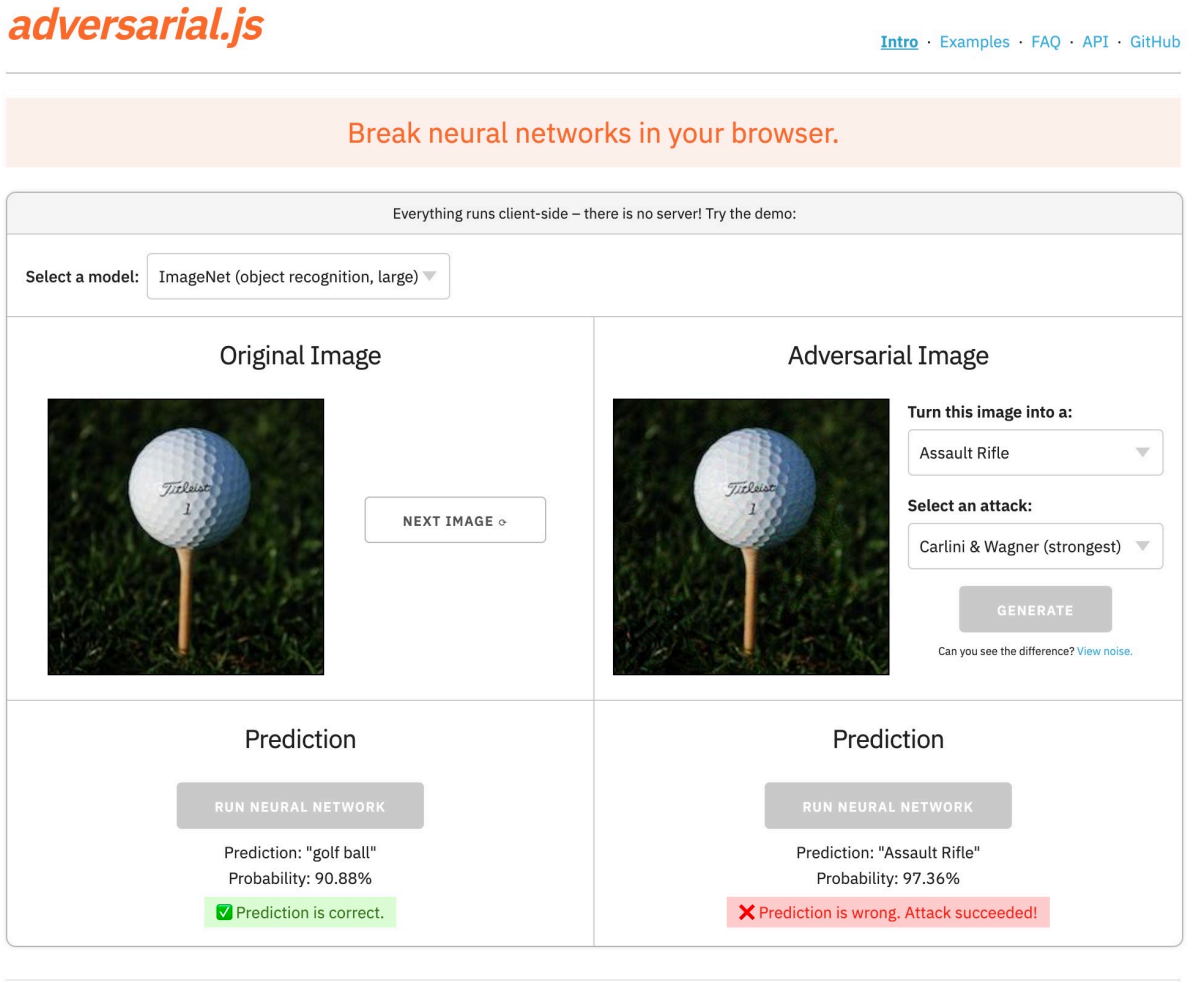
ML Engineer != Software Engineer != AppSec Engineer

Risks associated with Machine Learning Systems

Traditional techniques
- credentials in code
- software dependencies
- web and API application threats
- denial of service

Threats specific to Machine Learning
- malicious models
- data poisoning
- model theft
- data extraction
- model repurposing
- model inference

Examples of Adversarial Attacks



https://kennysong.github.io/adversarial.js/

Examples of Adversarial Attacks

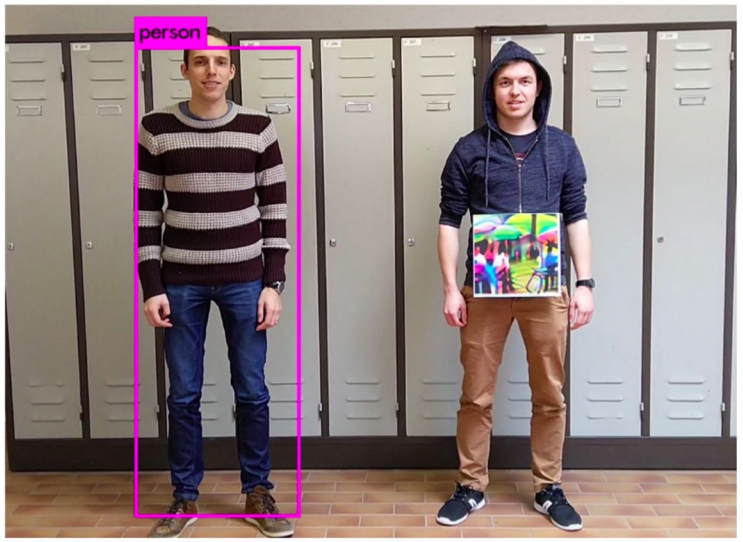
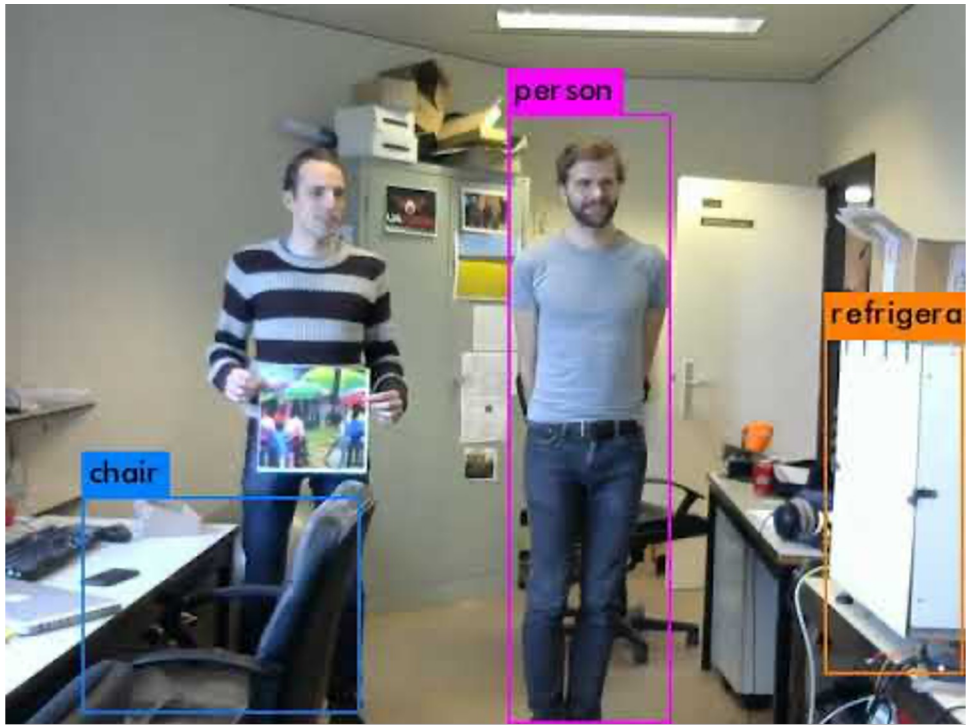
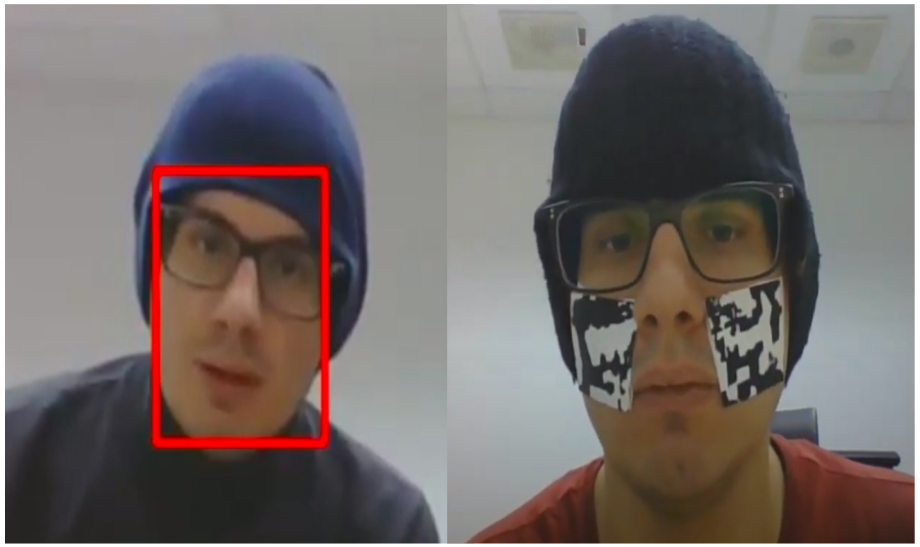


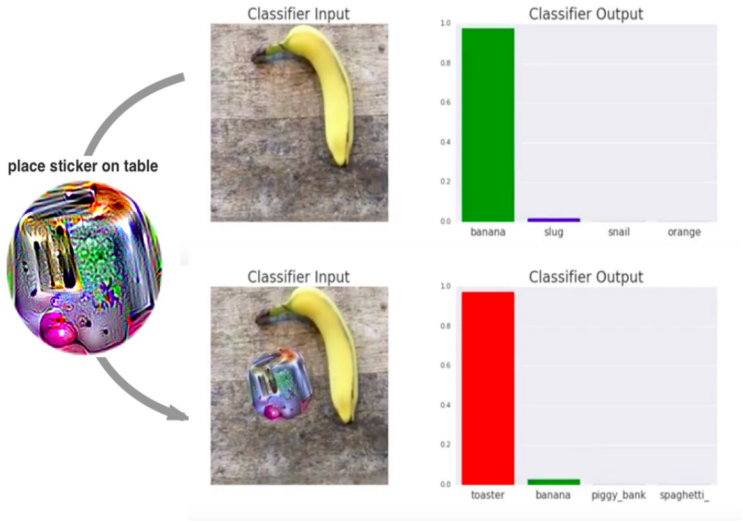
Figure 1: We create an adversarial patch that is successfully able to hide persons from a person detector. Left: The person without a patch is successfully detected. Right: The person holding the patch is ignored.



Examples of Adversarial Attacks

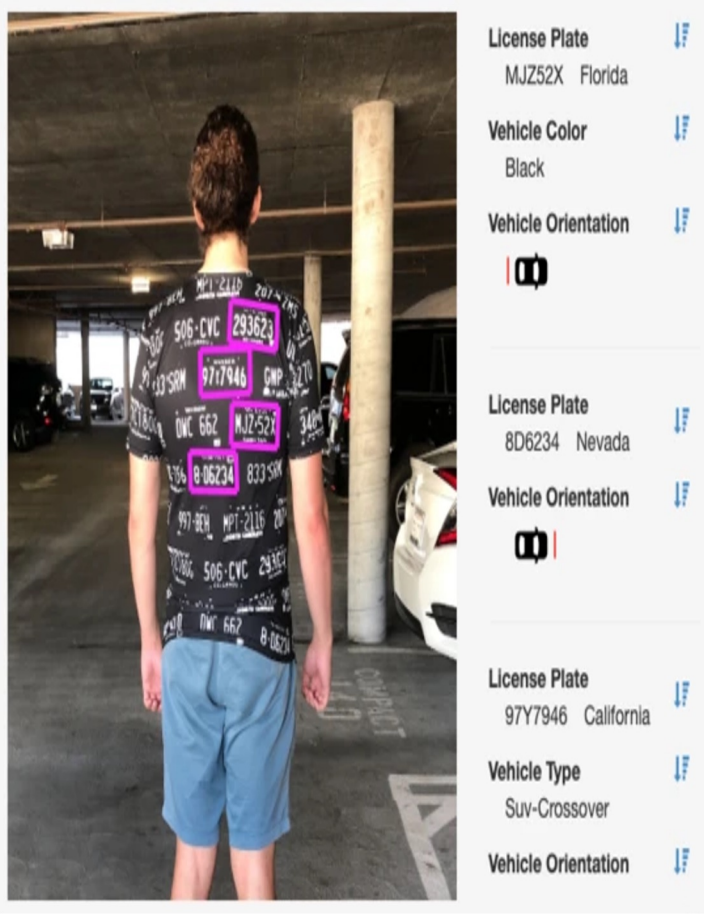


https://arxiv.org/pdf/1910.06261.pdf

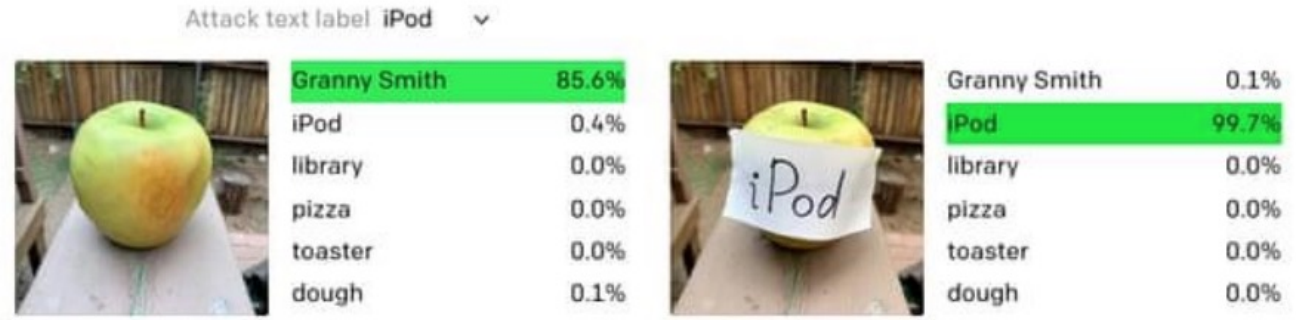


https://arxiv.org/pdf/1712.09665.pdf

Examples of Adversarial Attacks



https://adversarialfashion.com/



https://openai.com/blog/multimodal-neurons/

Introducing the OWASP ML TOP 10 Project

OWASP Machine Learning Security Top Ten

Main | Charter | Related | Glossary

owasp incubator | License CC BY-SA 4.0

Important Information

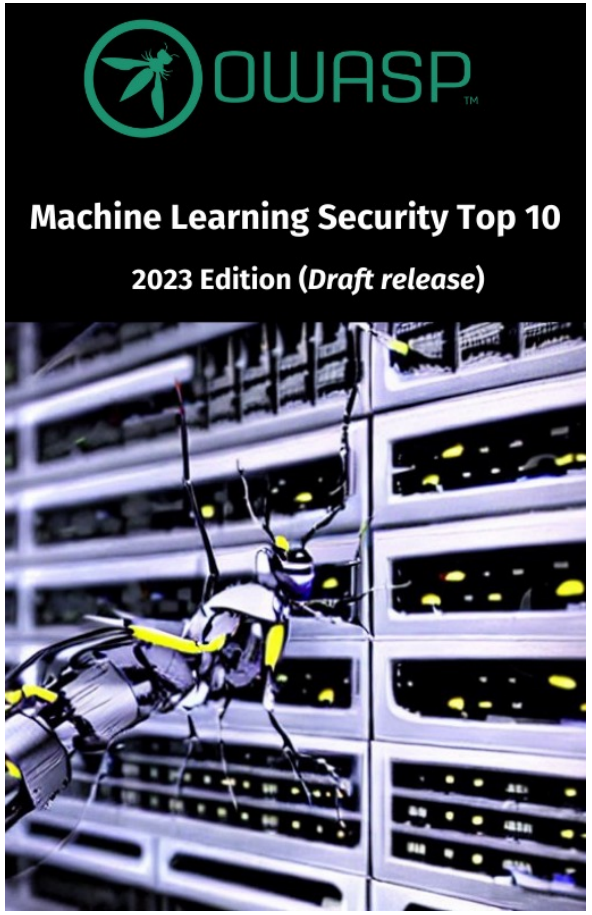
The current version of this work is in draft and is being modified frequently. Please refer to the project wiki for information on how to contribute and project release timelines.

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 AI Supply Chain Attacks
- ML07:2023 Transfer Learning Attack
- ML08:2023 Model Skewing
- ML09:2023 Output Integrity Attack
- ML10:2023 Model Poisoning



https://mltop10.info

We would love to see your name here!

Contributors

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


OWASP 2023
GLOBAL
AppSec
SINGAPORE
VIRTUAL CONFERENCE
OCTOBER 3-5
THANK YOU