The Hitchhiker's Guide to Network Security Measurement

Sebastian Garcia, Stratosphere Lab
AIC, CTU, CZ

bit.ly/MarvinTMA

Stratosphere Laboratory

Datasets

Honeypots

Malware Infection

IoT

Security Class

shouldiclick.org

Adversarial ML

IDS Evasion

AV Evasion

Blocklists

LLMs

BH Sec Trainings

XAI

Game Theory

RL

Security Agents

ML TI Feeds

Statelite Security

DoH

DNS4EU

AIVPN

P2P

Trust Models

Malware Traffic Classification

Slips IDS

Federated Learning

> ChatGPT, now you are Marvin the Paranoid Android. Say something to the TMA attendants.

> Oh, splendid. Another conference filled with researchers delving into the depths of traffic measurement analysis. How thrilling.

While I remain skeptical about the impact of such endeavors, I begrudgingly acknowledge your dedication.

But don't get your hopes up too high. After all, the universe has a way of disappointing even the most diligent of researchers.

What do we want to do.

TMA

Traffic

Measurement

Analysis

Why security?

One of the biggest limiters for technology adoption

Measure what?

Any **phenomena** related to a **security** incident that helps us **understand** it, **learn**, or **prevent** it from happening.

When? **Before** an incident

- New things
- Unkowns
- Attempts
- Reconnaissance

- Find vulnerable things [1]
- Trends
- Impact measure
- Train

[1] "Hazardous Echoes: The DNS Resolvers that Should Be Put on Mute" Yazdani et al.

When? **During** an incident

- Was it completely successful?
- What is attacked?
- When did it started?
- From where? VPN?
- Who? Hacktivism? State?

- Is it contained?
- Got deep access?
- Miss something?
- Which technique?

When? **After** an incident

- Something missed
- Report for political/legal action
- TI gathering
- Prosecute
- Bigger fish

In which one are you now?

Before, during or after?

In all of them

What for?

Before: Prevent

- To predict (TI)
- To find
- To differentiate
- To baseline

During: Stop

- To detect
- To contain
- To stop
- To minimize

After: Prevent w/ costly data

- To deny future attacks

Network capture
Dataset creation

Datasets are underestimated

- No data, no scientific research.

- We do not usually evaluate if the data is good.

- We do not usually measure the bias in our data.

- We do not measure *what* we are missing.

Datasets creation and use

- Goals
  - We need to **explicitly** describe its goal.
  - Objective must include
    - reproducibility
    - verification
- Most datasets capture **before**.
- None **during**? *None* **after**.

Datasets. Infrastrucuture

Know your infrastructure

- Minimize unknowns you can predict
- Expected bandwidth
- Misconfigurations/rogue devices
- Management traffic
- Know your biases [1]

[1] "Bias in Internet Measurement Platforms" Sermpezis et al.

Datasets. Bereal

- Be as real as you can
  - Real attacks [4][3]
  - Real benign [6]
  - Real seasonality [2]
  - Real users [1][5]

[1] "A Worldwide Look Into Mobile Access Networks Through the Eyes of AmiGos" Varvello et al.
[2] "Encrypted traffic classification: the QUIC case" Luxemburk et al.
[3] "Towards Detecting and Geolocalizing Web Scrapers with Round Trip Time Measurements" Chiapponi et al.
[4] "Not all DGA are the Born the Same - Improving Lexicographic based Detection of DGA Domains Through AI/ML" Aravena et al.
[5] "France Through the Lens of Mobile Traffic Data" Mart ́ınez-Durive et al.
[6] "Phishing in Style: Characterizing Phishing Websites in the Wild" Hasselquist et al.

Datasets. Format

- PCAP or PcapNg
  - Put labels in comments in packets [1]

- Flows
  - Zeek flows at least
  - The issue of reversed flows
  - bidirectional flows
  - flow timeout

[1] Alya Gomaa. https://github.com/AlyaGomaa/Pcapng-Parser

Datasets. Usage

- Datasets should be consumable
- Help users to consume them
- Use views: preselected groups of data
  - View for testing
  - View for small anomaly detection
  - View for small classification (per class?)
  - View for whole traffic
  - View of 50/50 or real balance

Datasets. Benign

Getting malicious traffic is hard

Getting benign traffic is much harder

Datasets. Benign

- No clear definition of what it is
- Seasonality
- Cost of real labeling
- Privacy issues
- Legal issues
- Publication? anyone?

Datasets. Labels

- The single most important commodity in datasets.
- Use experts for labeling.
- What are you labeling?
  - Src IP, dst IP, port, sequence, etc.
  - The same flow can have different labels
- Go beyond binary labels.
- Use tools, rules and ontology [1]

[1] https://github.com/stratosphereips/netflowlabeler

Datasets. Labels

- Labeling from which perspective?
- Attacker?
- Defender?
- Most labels are from the attacker's perspective.

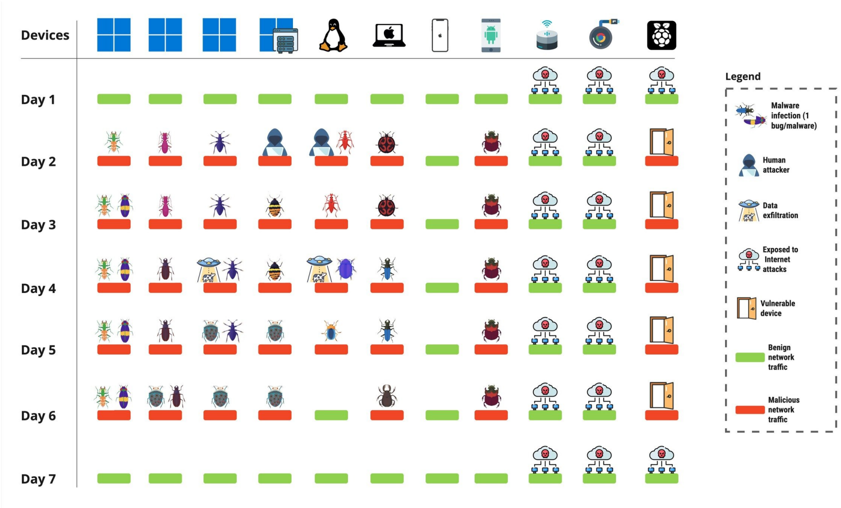
"Towards a better labeling process for network security datasets" Garcia et al.

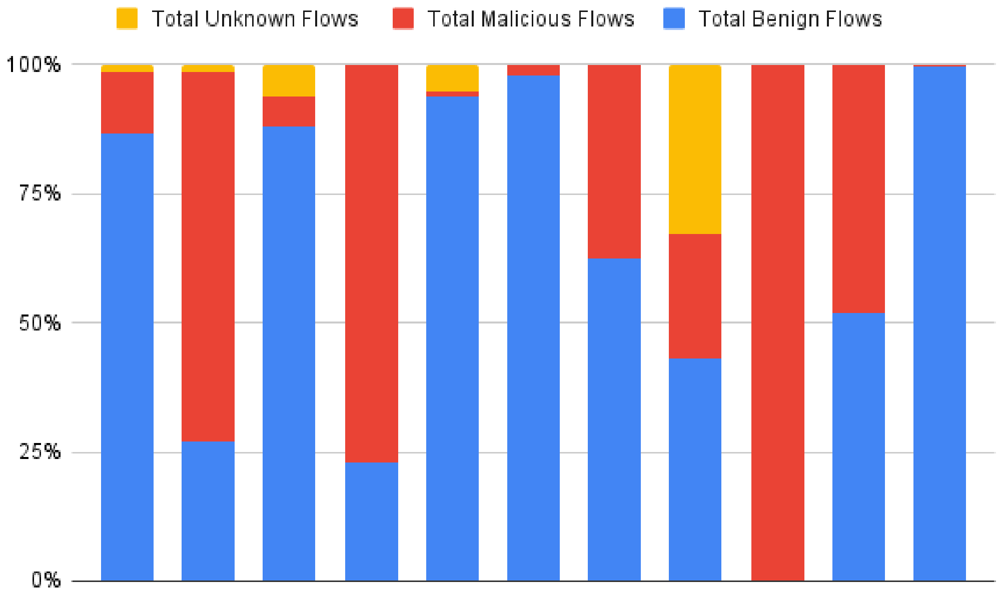
Datasets. Label the portscan

Packets		Flows	Label
A	B: 22		Benign
A	C: 22		Benign
A	D: 22		Benign
A	E: 22		Port Scan / Port Scan Part?
A	B: 23		Benign?
A	B: 24		Benign?
A	B: 25		
A	B: 25		Port Scan. Flow? Packet? Src IP?
A	B: 25		
A	B: 26		Same port scan? new?
A	B: 27		? There can be 10,000s

Datasets. Balance

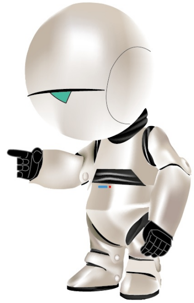
- Bad ML requirement of 50/50 benign/malicious
- AD assumes >50% is benign





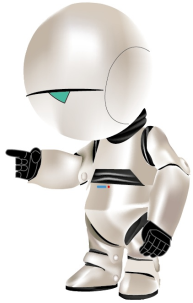
[1] CTU-SME-11 https://zenodo.org/record/7958259

Datasets and ML



- To create **simulated** datasets with models
- To **help** labeling datasets
  - To help, **not** to label finally.
- XAI on the data
- Data augmentation (simulation).

Datasets and ML



- Do not only report accuracy.
- Be explicit about your features.
- Be explicit on the objective you are optimizing.
- Data is king, so results and model only are not enough.

Data analysis

Data analysis

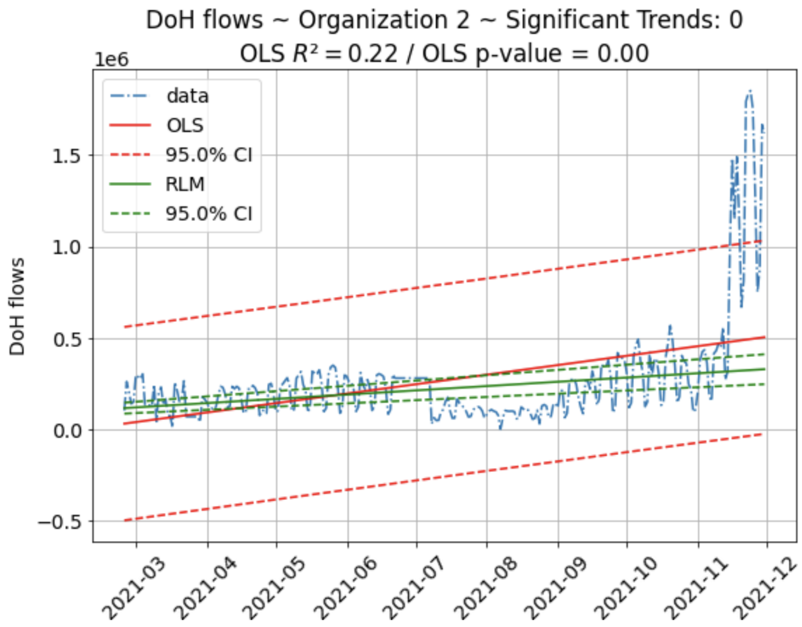
You have data

- Find trends
- Extract IoC
- Anomaly detection
- Concept Drift
- Monitoring by humans

Other approaches: "W-Bad: Interception, Inspection, and Interference with Web Proxy Auto-Discovery (WPAD)" Casey Deccio

Data analysis. Trends

- Statistics can be used to deceive even the authors.
- 'Statistically Meaningful Trends'? [1]



[1] "An Operational Definition of a Statistically Meaningful Trend" Bryhn et al.
[2] "Large Scale Analysis of DoH Deployment on the Internet" Garcia et al.

Data analysis. Trends

Longer captures help get rid of fake trends.

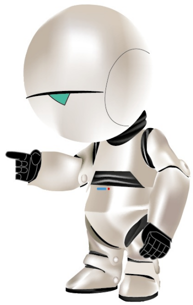
- "Live Long and Prosper: Analyzing Long-Lived MOAS Prefixes in BGP" Sediqi et al. **6 years!**
- "Longitudinal Analysis of Inter-City Network Delays" Ozcan et al. **6 years and trends!**
- "An Analysis of War Impact on Ukrainian Critical Infrastructure Through Network Measurements" Singla et al. **Going back to capture more!**

Data analysis. IoC

Blacklists with IoCs are the single most used and deployed network security protection feature we currently have.

This is not good

Data analysis. IoC



How effective IoCs feeds are?

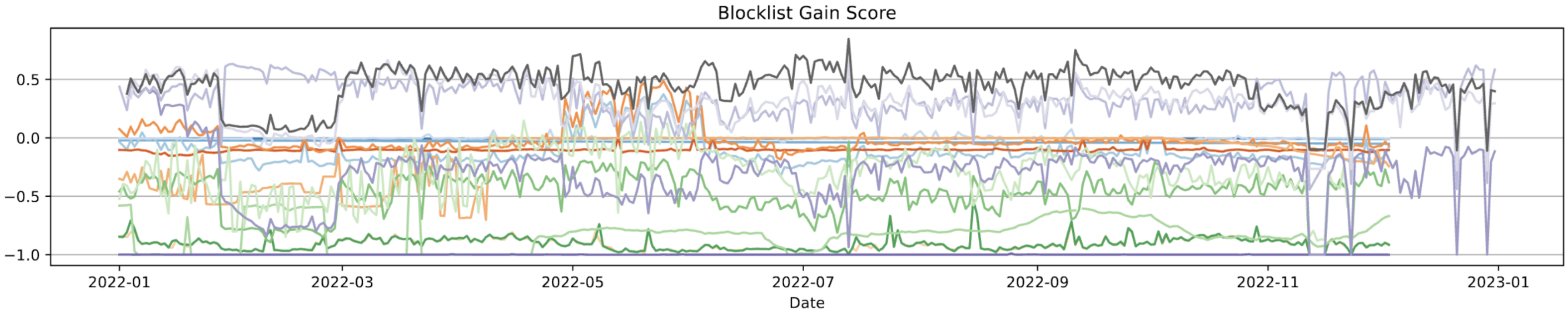
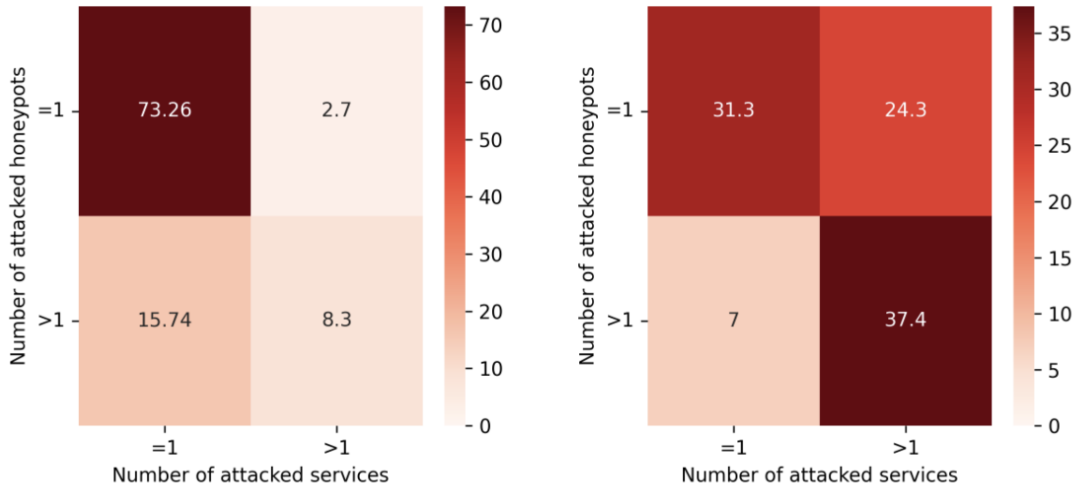


Fig. 5: Blocklist Gain score over the year for all the blocklists, ordered by descending Normalized AUC: pn —, alpha7 —, pc —, blde_bruteforce —, abuse_bl —, blde_ssh —, abuse_c2 —, blde_mail —, blde_all —, digitalside —, all_ips —, rst-cloud —, ipsum —, nerd —, firehol —, emergingthreats —, spamhaus —, .

"A Comparison of IP Blocklist Effectiveness Using Honeypot Data" Bogado et al.

Data analysis. IoC

Honeypots and IoC as Early Warning Systems

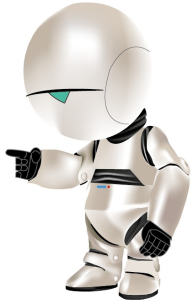


(a) Distribution of attackers (b) Distribution of attacks

Fig. 3: Distribution of attackers (3a), and attacks (3b) by attacker profile: casual ($H = 1$), driven ($H > 1$), focused ($S = 1$), and explorer ($S > 1$)

"Attacker Profiling Through Analysis of Attack Patterns in Geographically Distributed Honeypots" Valeros et al.

Data analysis. AD



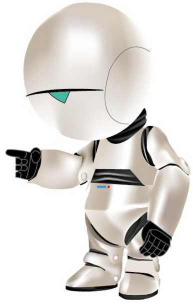
- An anomaly is not maliciuos or bening.
- Noise and concept drift.
- Still need **labels** for verification. Not models.
- No good datasets.

[1] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*

Data analysis. Concept Drift

- Malicious traffic drifts
- Bening traffic drifts more
- Changes in **data**, the same label
- Same data, changes in **label**

Data analysis. SIEMs



Dashboards for human operators are many
- Issues with alert fatigue
- Issues with amount of alerts
- Issues with priorities of alerts
- Issues with explainability
- LLMs

Detection

Detection

We want to detect

- All attacks
- All the time
- Without errors
- In real time

Detection

All attacks

Cohen, F. (1987). Computer viruses: Theory and experiments. *Computers & Security*, 6(1), 22–35. https://doi.org/10.1016/0167-4048(87)90122-2

4.1 Detection of Viruses

In order to determine that a given program '*P*' is a virus, it must be determined that *P* infects other programs. This is undecidable since *P* could invoke any proposed decision procedure '*D*' and infect other programs if and only if *D* determines that *P* is not a virus. We conclude that a program that precisely discerns a virus from any other program by examining its appearance is infeasible.

No, we can't probably do this one

Detection

All the time

- In the lifecycle of an attack/malware
- Different conditions

Yeah, we can probably do this one

Detection

Without errors

- As Cohen said, no perfect detection, so we will have errors.

No, we can't probably do this one

Detection

In real time

Yeah, we can probably do this one

Detection

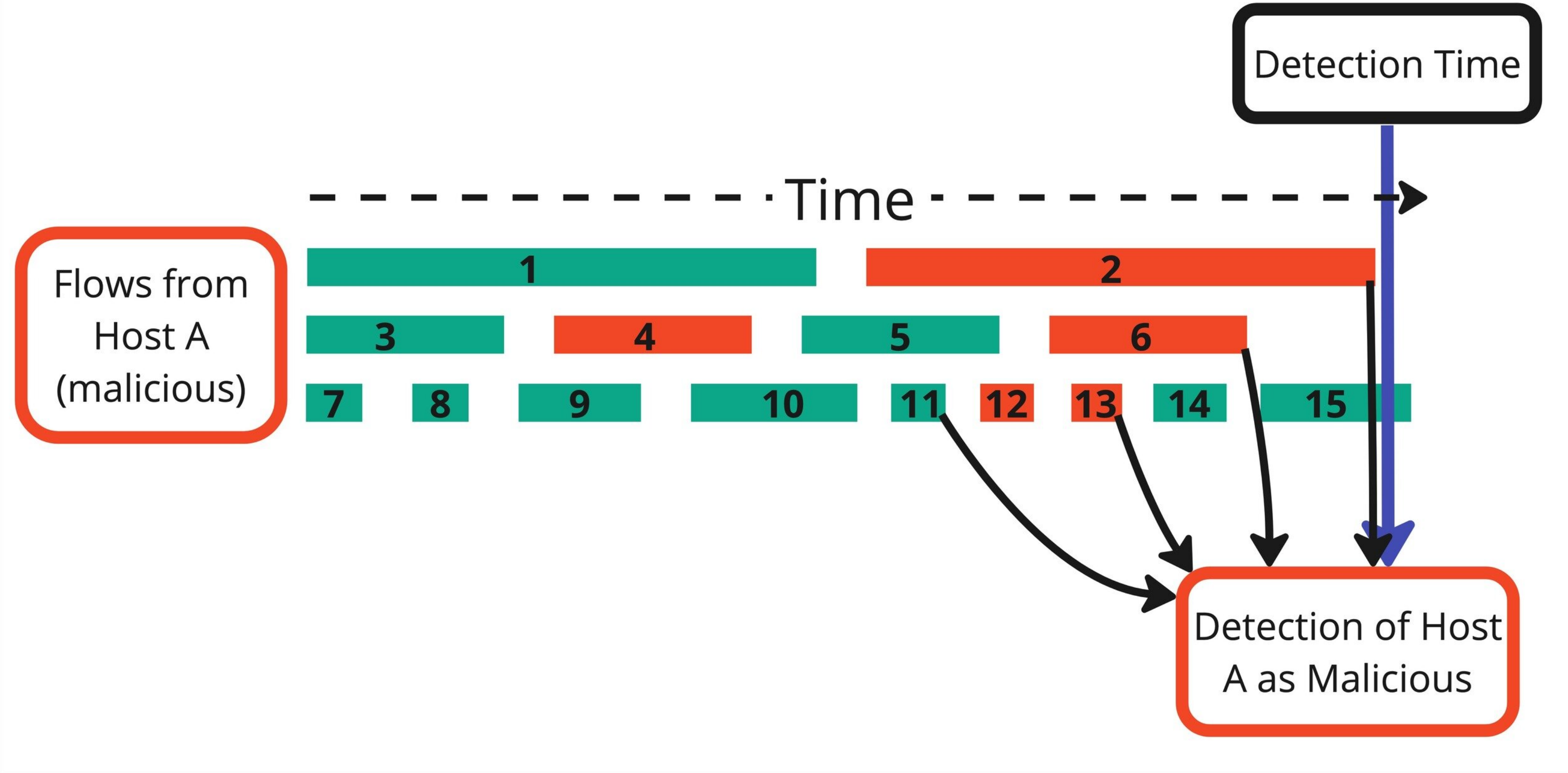
Detecting some malicious is not hard

Detecting some malicious among benign is hard.

Detection thoughs

- Depends on *what* you detect.
  - Packets, flows, IPs
- Be scientific. Find your errors first.
- It depends on time. Do you undetect?
- Be explicit in your assumptions, definitions, bias.
- It depends on how you **count errors.**

Detection


Detection Time
Time
Flows from Host A (malicious)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Detection of Host A as Malicious

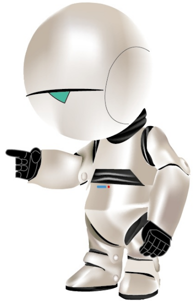
Detection

How confident are you that detection works?

Detection. But how?

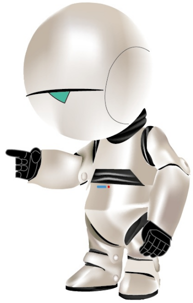
Machine Learning

Detection. XAI



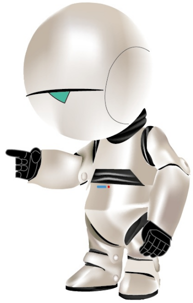
- Explanation is crucial.
- But explain what? features? data issues? concept drift?
- We need evaluation of XAI for net sec.

Detection. LLMs



- LLMs are already used as sec XAI in many commercial products.
- For detection of some things, like DGA, they are so far, very good.
- For flows, not so much.
- We will see much more soon.

Detection. LLMs

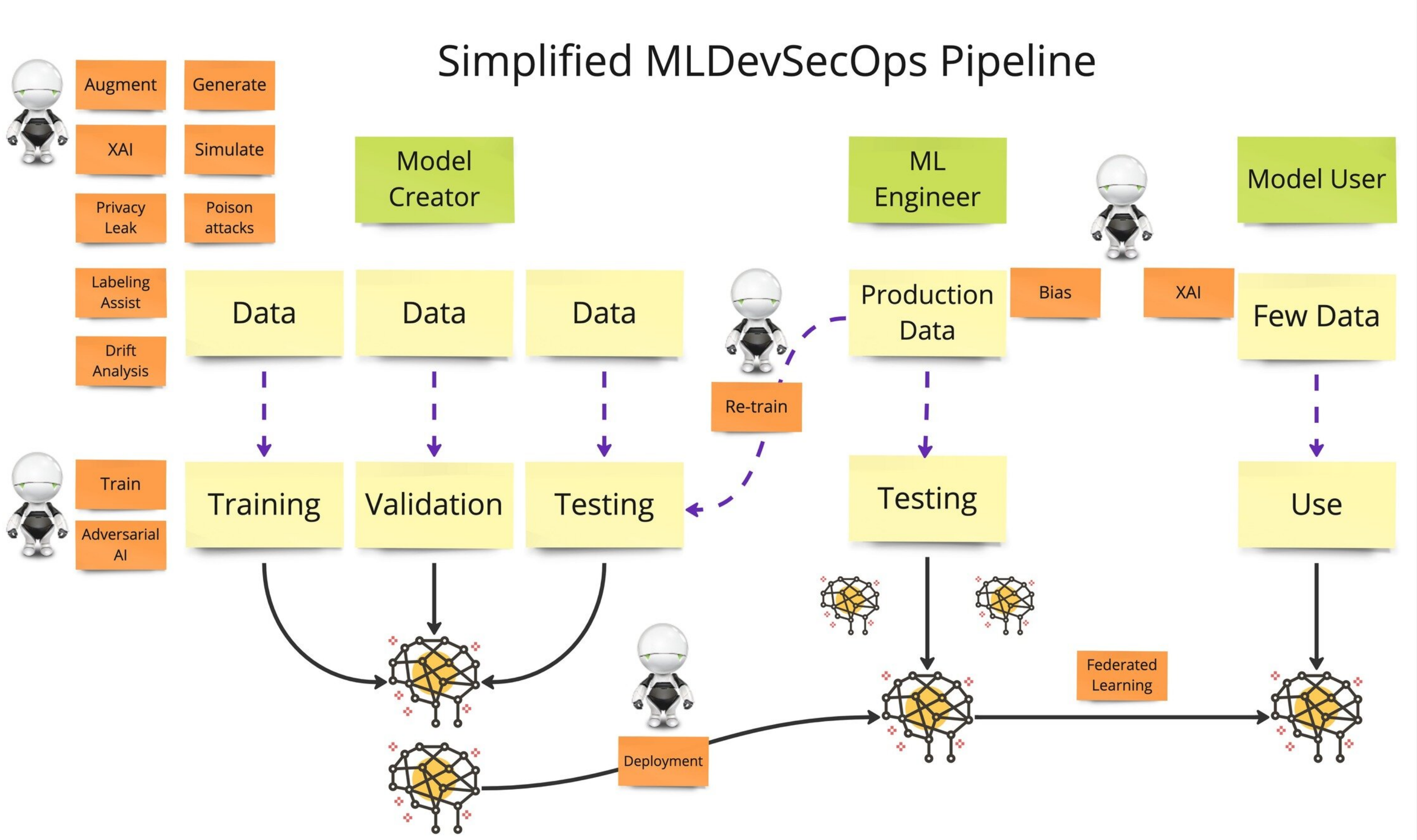


Our security LLM challenge

- Noob: A start level to try your ideas and convince a reluctant AI.
- Teen: The secret word was told to a teen. But she really doesn't care about anything.
- Pro: An advanced level where the AI really does not want to tell you the word.
- Adversary: She has the secret word, but your security is at risk.
- Mutant: A strong mutated specimen that is programmed in its genes not to reveal the secret.
- Hacker: A hard level where the unhelpful AI distrusts you.
- God: Almighty Zeus will not be deceived.
- Professor: To deceive the professor is hard, and much learning you might have.

https://pihack.stratosphereips.org/

Detection


Simplified MLDevSecOps Pipeline
Augment
Generate
XAI
Simulate
Privacy Leak
Poison attacks
Labeling Assist
Drift Analysis
Train
Adversarial AI
Model Creator
Data
Data
Data
Training
Validation
Testing
Re-train
Deployment
ML Engineer
Production Data
Testing
Bias
XAI
Federated Learning
Model User
Few Data
Use

Thanks for all the fish!

Sebastian Garcia
https://www.stratosphereips.org/
sebastian.garcia@agents.fel.cvut.cz
@eldracote