Adversarial Attacks in AI

Cybersecurity AI

Hyunsang Choi

Who Am I?

최현상

SK텔레콤 Cybersecurity AI팀

☑ 이전 경력

- Korea Univ., BS, MS, Ph.D. (’00-’12)
- Microsoft Research, Intern (’09-’10)
- U.C. Berkeley, Postdoc (’13-’14)
- SECUI, Developer(’14-’16)
- Naver, Security Engineer (’16-’19)
- SKT Cybersecurity AI, Developer (’19-)

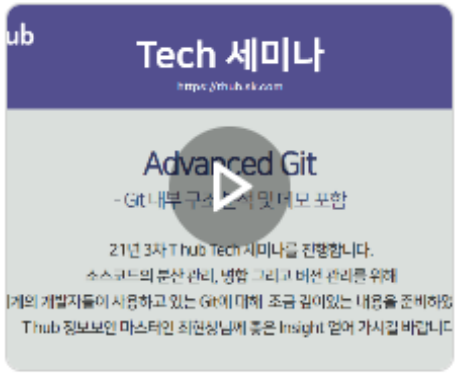
☑ 현재 업무

- CSMP (Cloud Security Management Platform) 개발
- CPMS (Cloud PII Management System) 개발
- 스팸필터링 시스템 고도화 개발

☑ ‘#’해시태그로 나를 표현하기

- ENTP (뜨거운 논쟁을 즐기는 변론가, 발명가)
- 기술덕후
- 애국자 ㅠㅠ

Previously on Tech Seminar …



Advanced Git

1. Git internals deep dive 2. Git commands 3.
Git 발표 자료 :
다운로드

21.05.14 591 9 1

쿠버네티스 보안 기술

쿠버네티스 보안과 관련된 설정과 기술에 대해
Authorization - 네트워크 보안 : Network P
Falco - 기타 : kube-bench, kube-hunter,

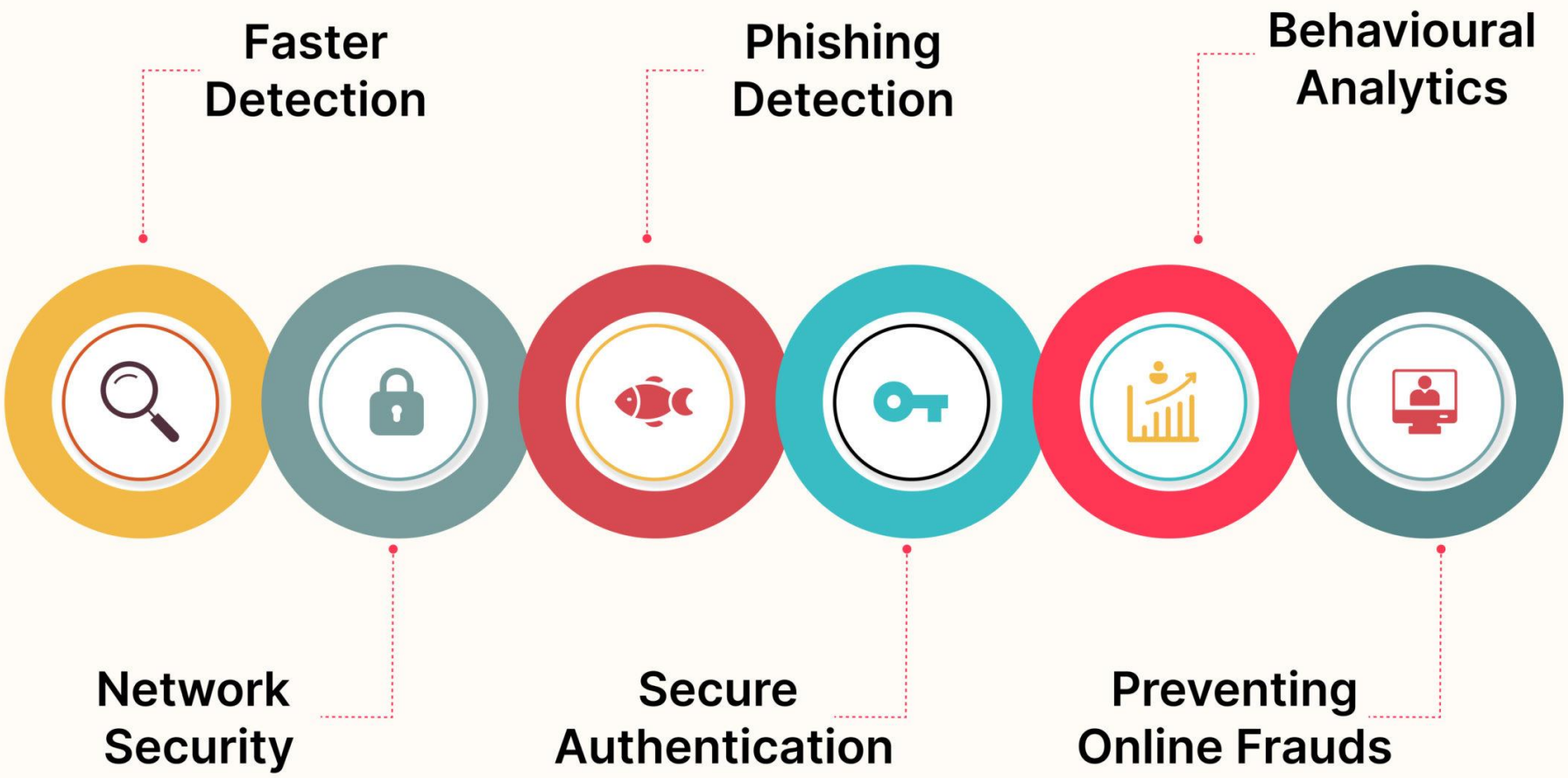
20.10.28 120 4 0

Container Internals & Security

컨테이너 내부 구조를 살펴보고, 컨테이너의 보안기술에 대해 알아보고자 함
모셨습니다. - Docker 컨테이너 내부 구조와 관련된 기술들 - 컨테이너의
화된 런타임들 - 쿠버네티스 보안

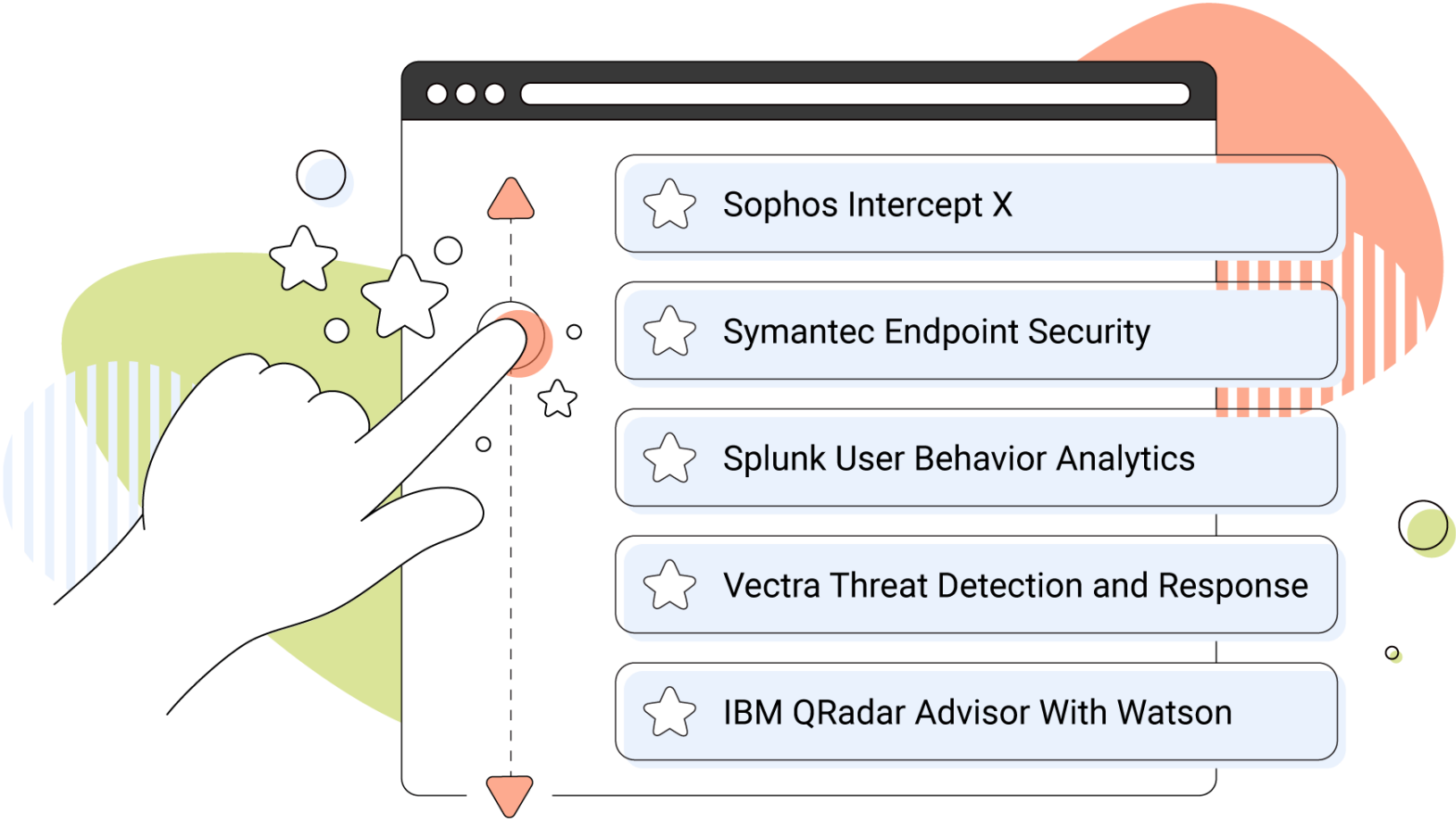
20.05.14 80 4 0

AI in CyberSecurity



https://www.orangemantra.com/blog/how-ai-in-cybersecurity-reimagines-cyberthreat/

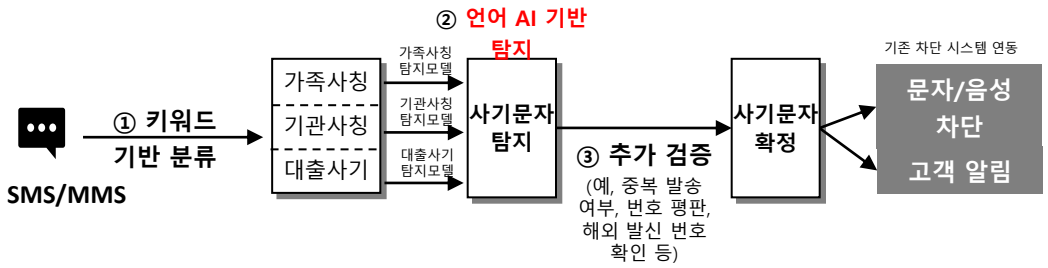
AI in Cybersecurity



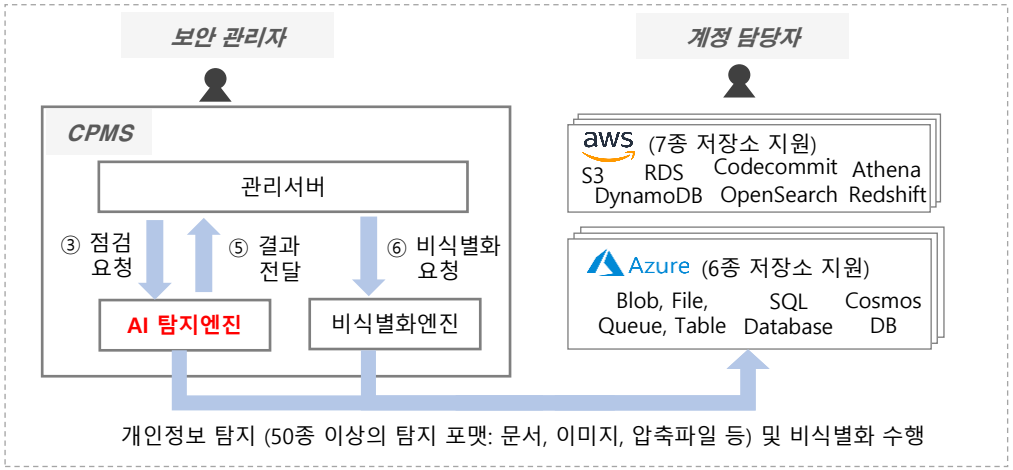
https://www.hostpapa.com/blog/web-hosting/the-most-useful-tools-for-ai-machine-learning-in-cybersecurity/

Current Projects of Cybersecurity AI Team

보이스피싱 문자 탐지기술

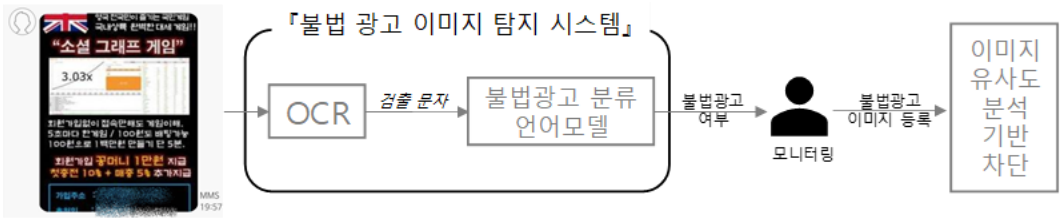


클라우드 개인정보 탐지기술



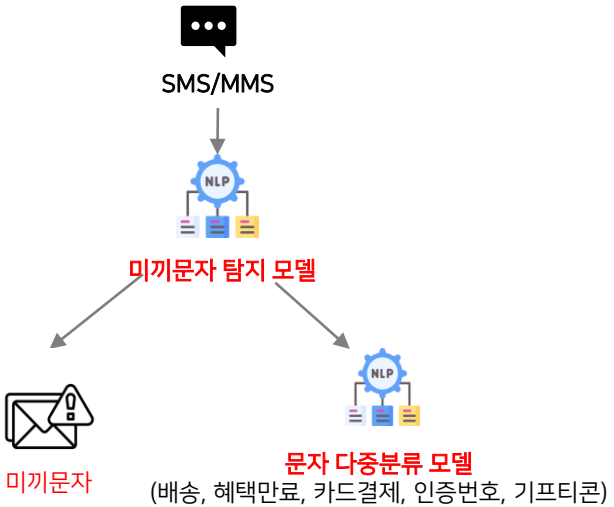
스팸 필터링 고도화

스팸 필터링 XAI (eXplainable AI)



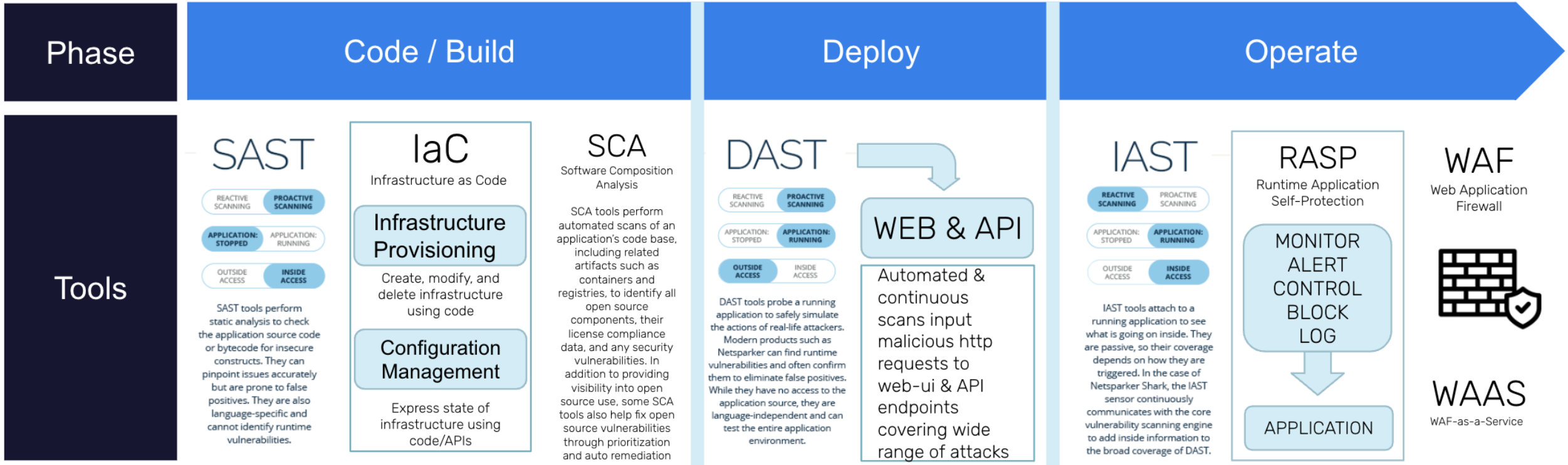
deep learning OCR, 분류 언어모델

문자 분류 기술



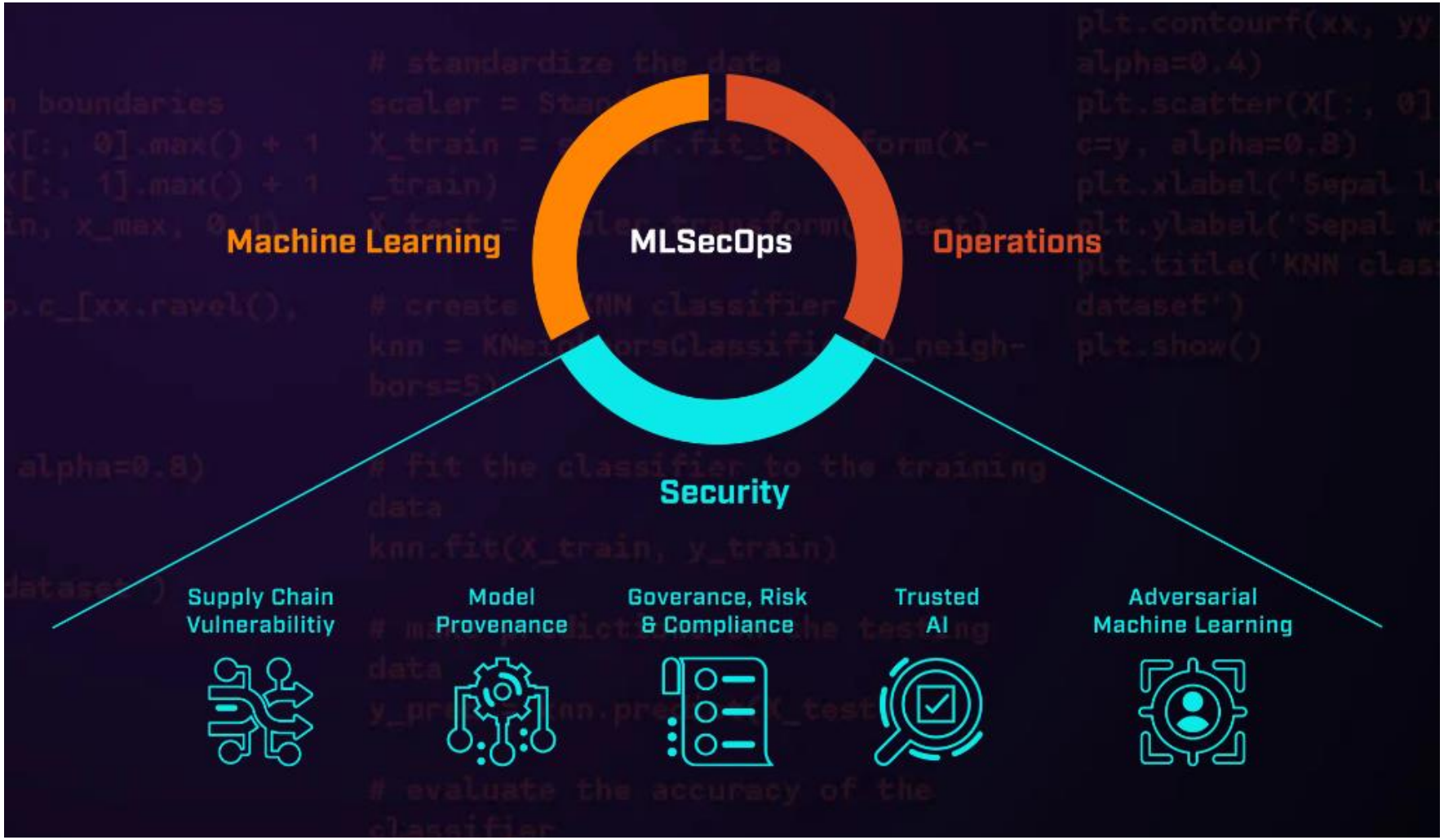
DevSecOps

DevSecOps tooling process

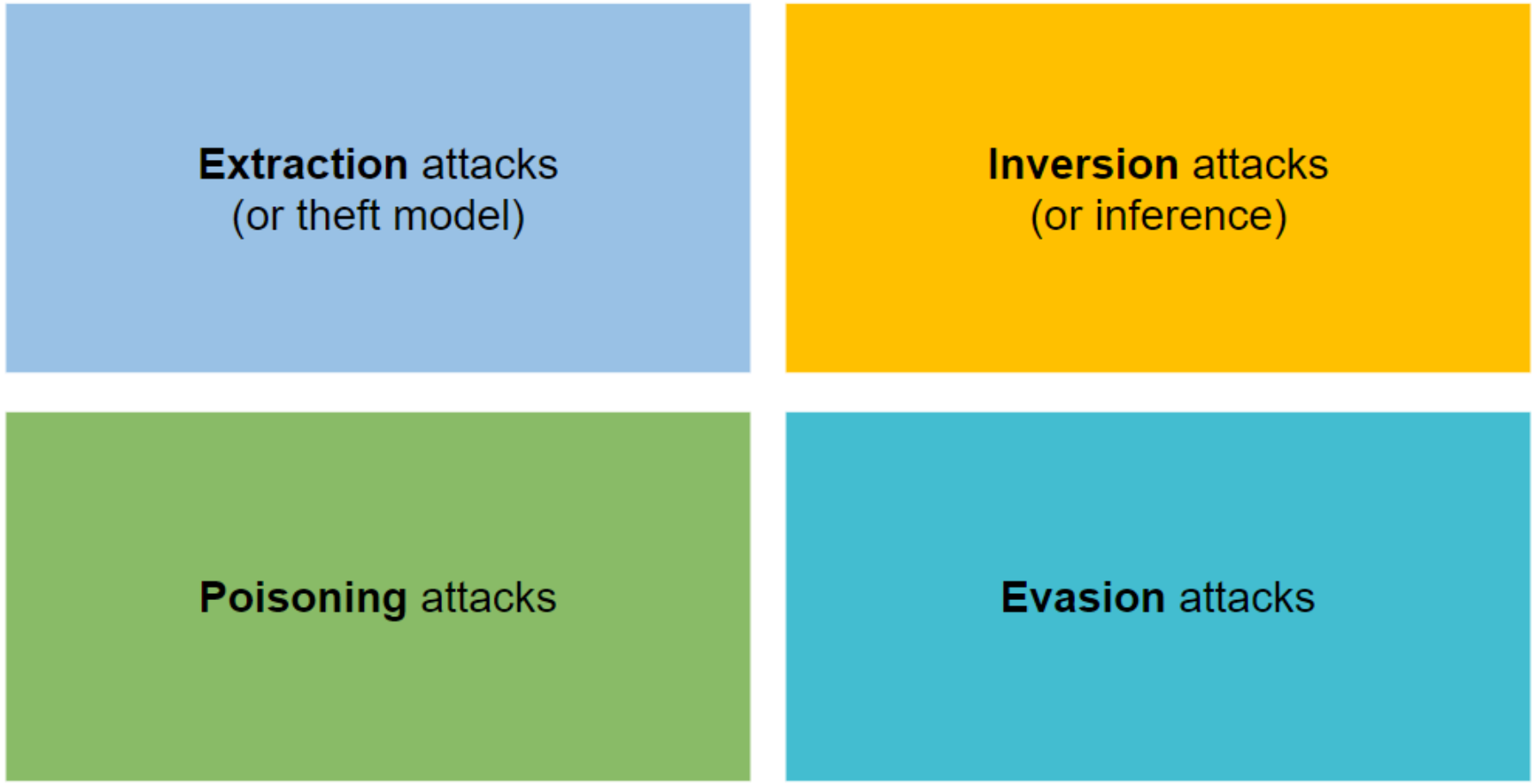


https://nordcloud.com/tech-community/devsecops-best-practices/

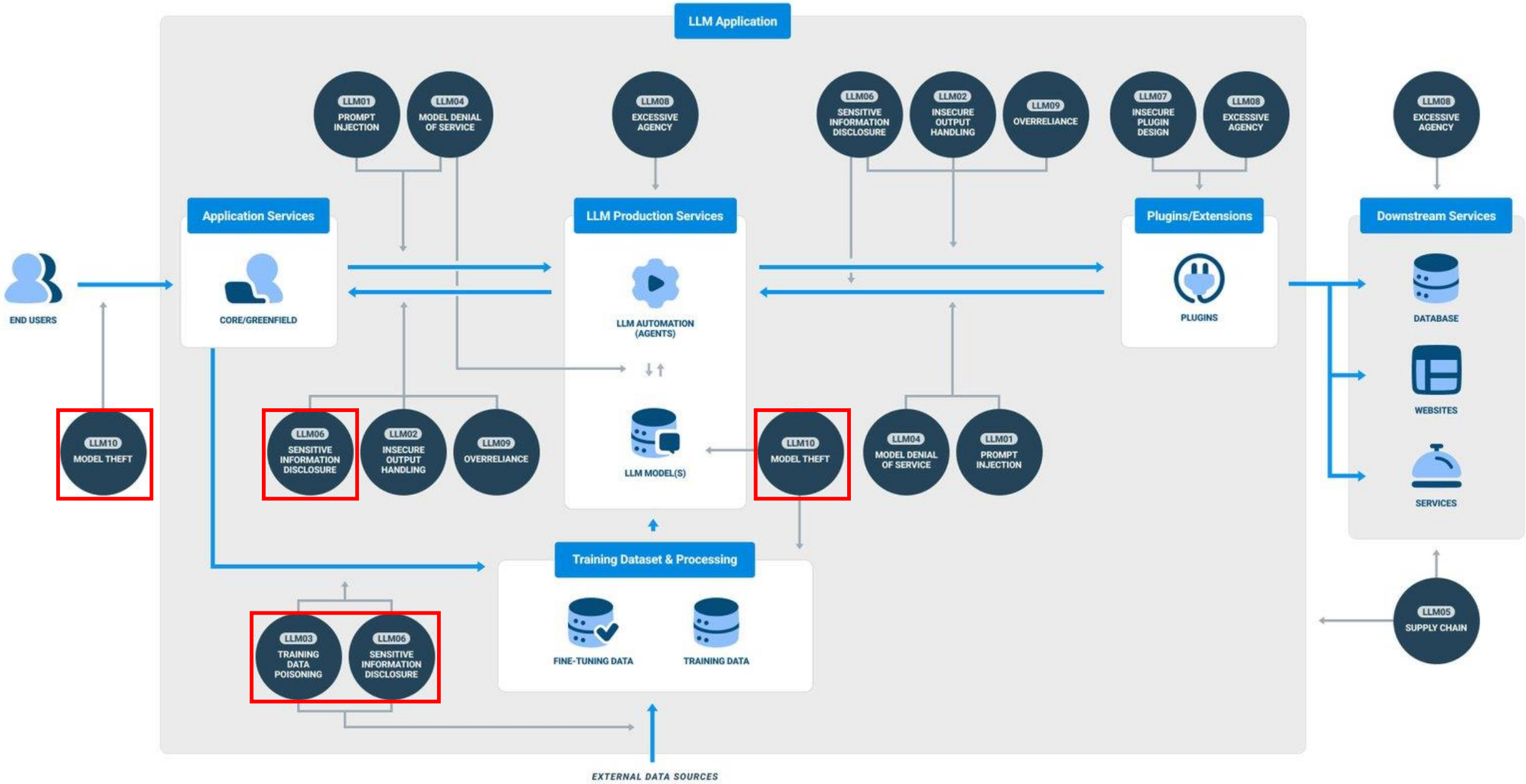
MLSecOps


Machine Learning
MLSecOps
Operations
Security
Supply Chain Vulnerabilitiy
Model Provenance
Goverance, Risk & Compliance
Trusted AI
Adversarial Machine Learning

Cybersecurity in AI



https://jiep.github.io/offensive-ai-compilation/

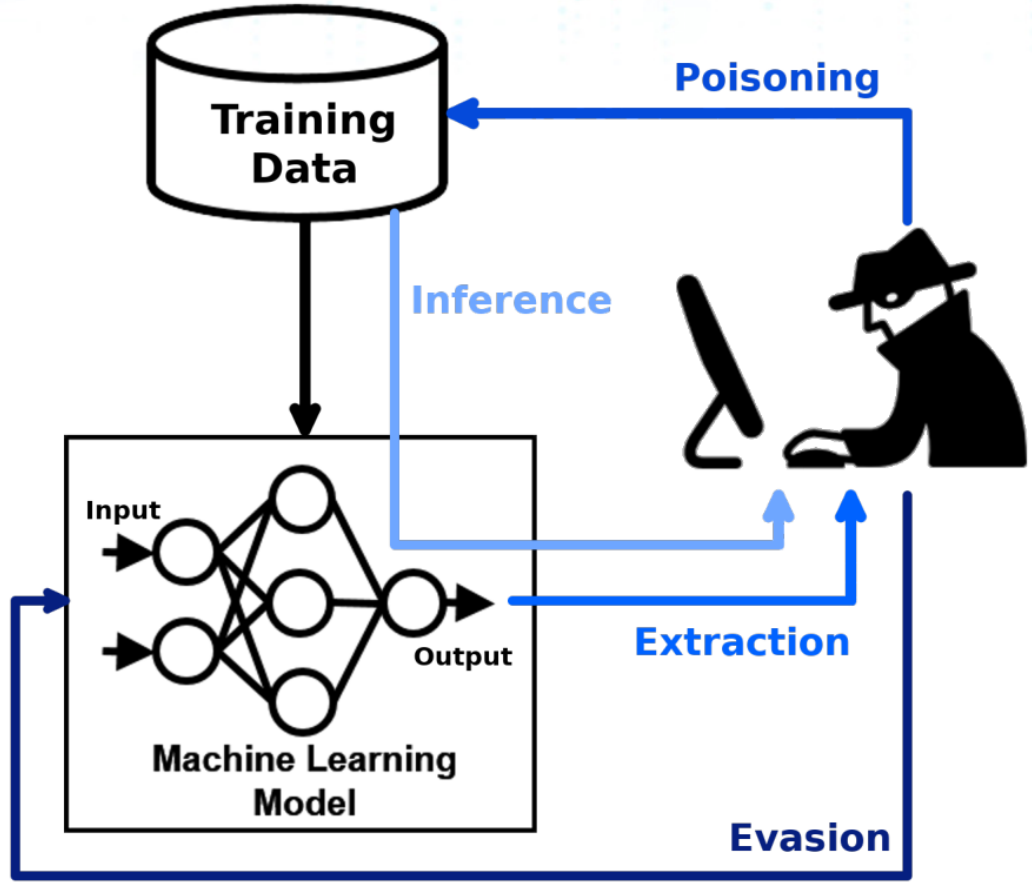


https://llmtop10.com/

Adversarial Attacks

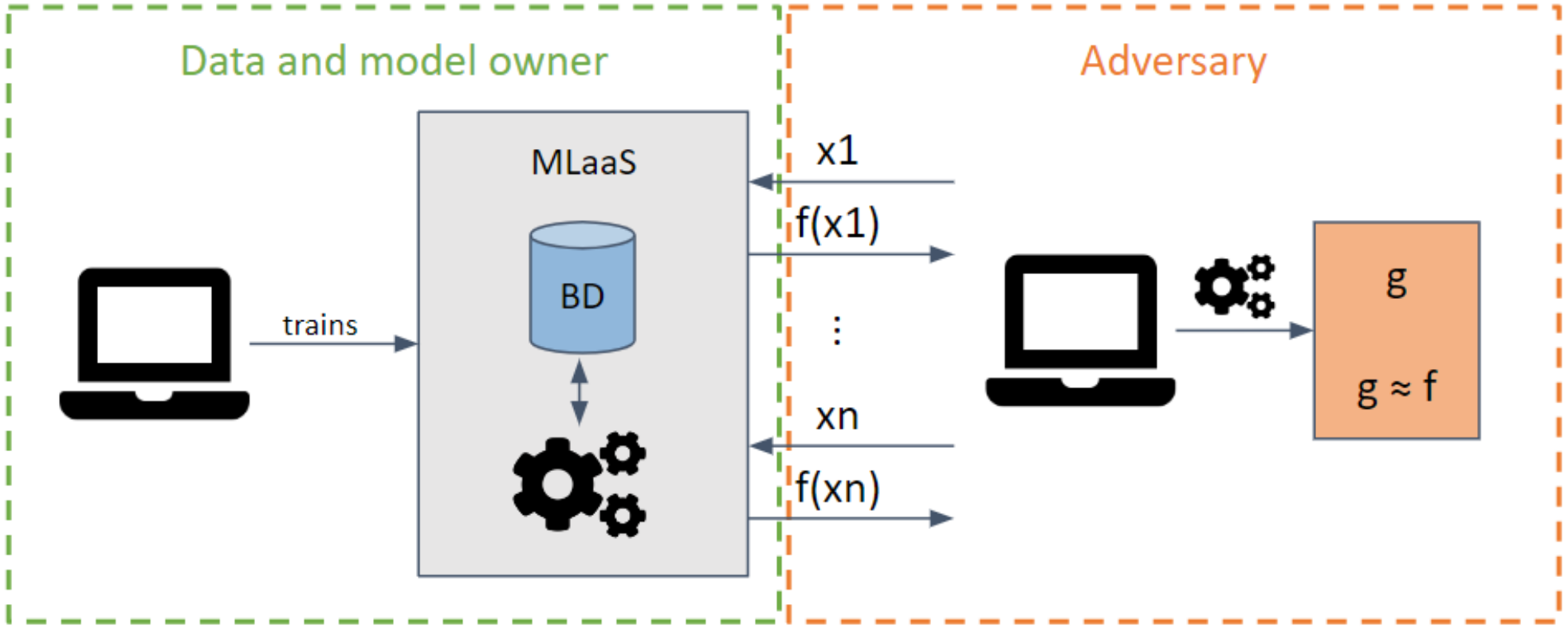
Adversarial threats against machine learning models and applications:

- **Evasion:** Modifying input to influence model
- **Poisoning:** Modify training data to add backdoor
- **Extraction:** Steal a proprietary model
- **Inference:** Learn information on private data



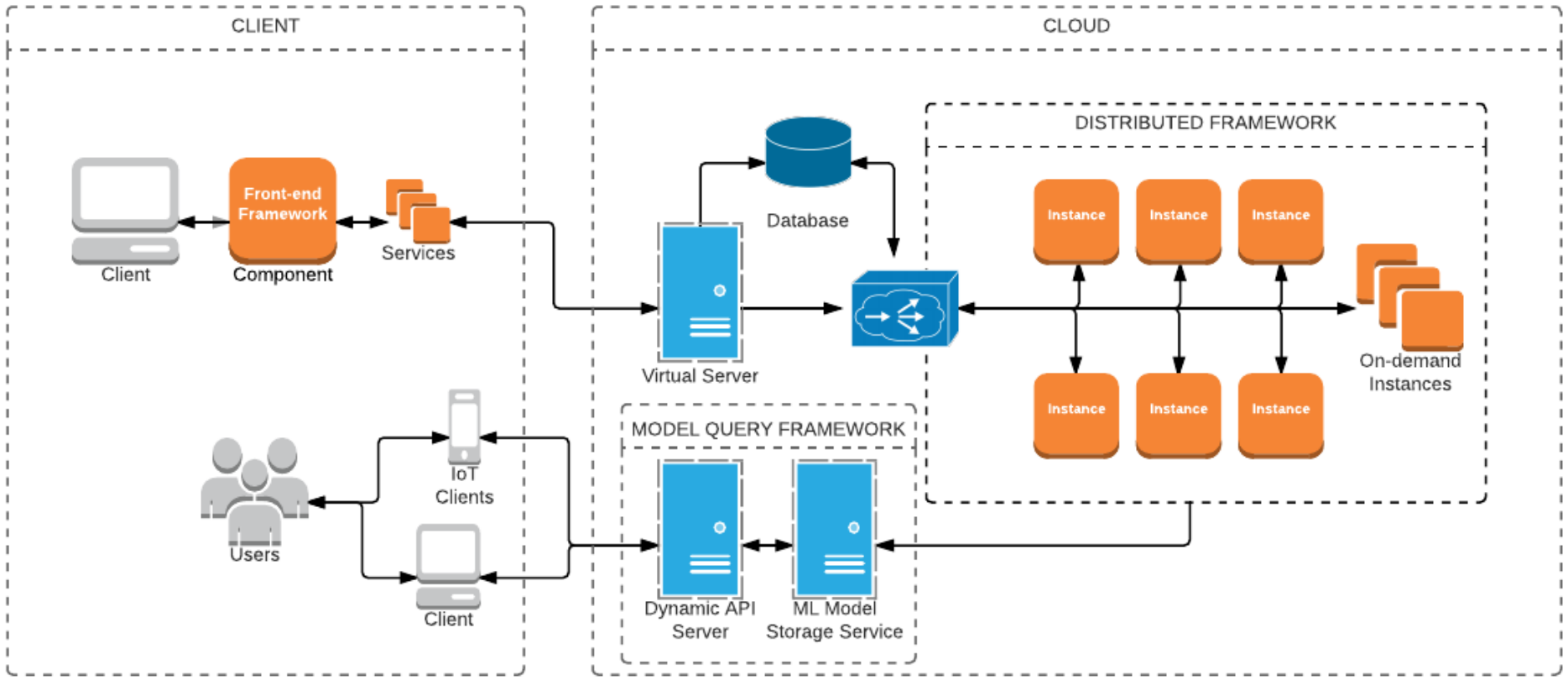
https://www.rsaconference.com/Library/presentation/USA/2021/evasion-poisoning-extraction-and-inference-tools-to-defend-and-evaluate

Extraction



https://jiep.github.io/offensive-ai-compilation/

MLaaS



https://github.com/cyberbeast/mlaas

Extraction: Real-world Example (1)

Ram Shankar
@ram_ssk

CVE-2019-20634 is a 💣🕸

@moo_hax shows that the #MachineLearning system powering @proofpoint email protection (versions up to 2019-09-08) are vulnerable to model stealing & evasion attacks.

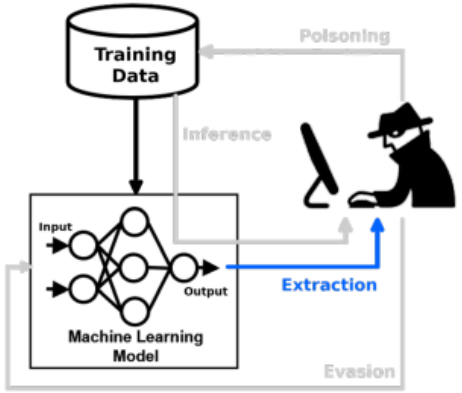
nvd.nist.gov/vuln/detail/CV...

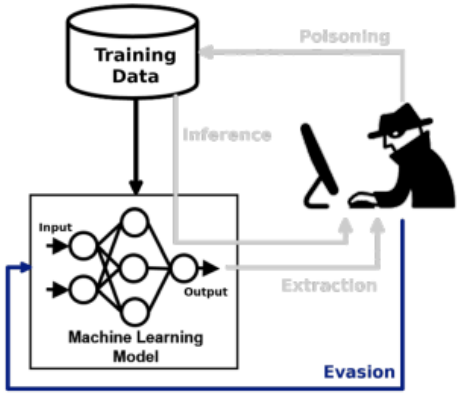
Adversarial ML is now an #infosec problem. Wow.

– CVE-2019-20634

- https://nvd.nist.gov/vuln/detail/CVE-2019-20634
- https://github.com/moohax/Proof-Pudding
- https://github.com/moohax/Talks/blob/master/slides/DerbyCon19.pdf

- **CVE-2019-20634**
- Work of Will Pearce and Nick Landers
- Attacking an application leaking information on internal classification decisions
- Combination of two threats (e.g., extraction and evasion) increases impact and feasibility





- **Step 1: Extraction of Classification**
- Create dataset of input label pairs by querying the application
- Train surrogate classification model on new dataset

- **Step 2: Apply Classifier**
- Use insights on application's classification to craft adversarial examples
- Design adversarial examples to achieve attack goal, e.g., misclassification of spam email

https://www.rsaconference.com/Library/presentation/USA/2021/evasion-poisoning-extraction-and-inference-tools-to-defend-and-evaluate

ProofPoint Header

```
To: <reciever@domain.com>
From: <sender@domain.com>
Subject: Our Meeting
...
X-Proofpoint-Spam-Details: rule=nodigest_notspam policy=nodigest score=0
 malwarescore=0 mlxlogscore=999 mlxscore=0 suspectscore=14 spamscore=0
 impostorscore=0 adultscore=0 clxscore=593 priorityscore=0 phishscore=0
 bulkscore=97 lowpriorityscore=97 classifier=spam adjust=0 reason=mlx
 scancount=1 engine=9.1.0-12345000 definitions=main-12345
```

1. **Collect a dataset** - Send X emails to steal scores
2. **Copy the model -** Use their outputs to duplicate

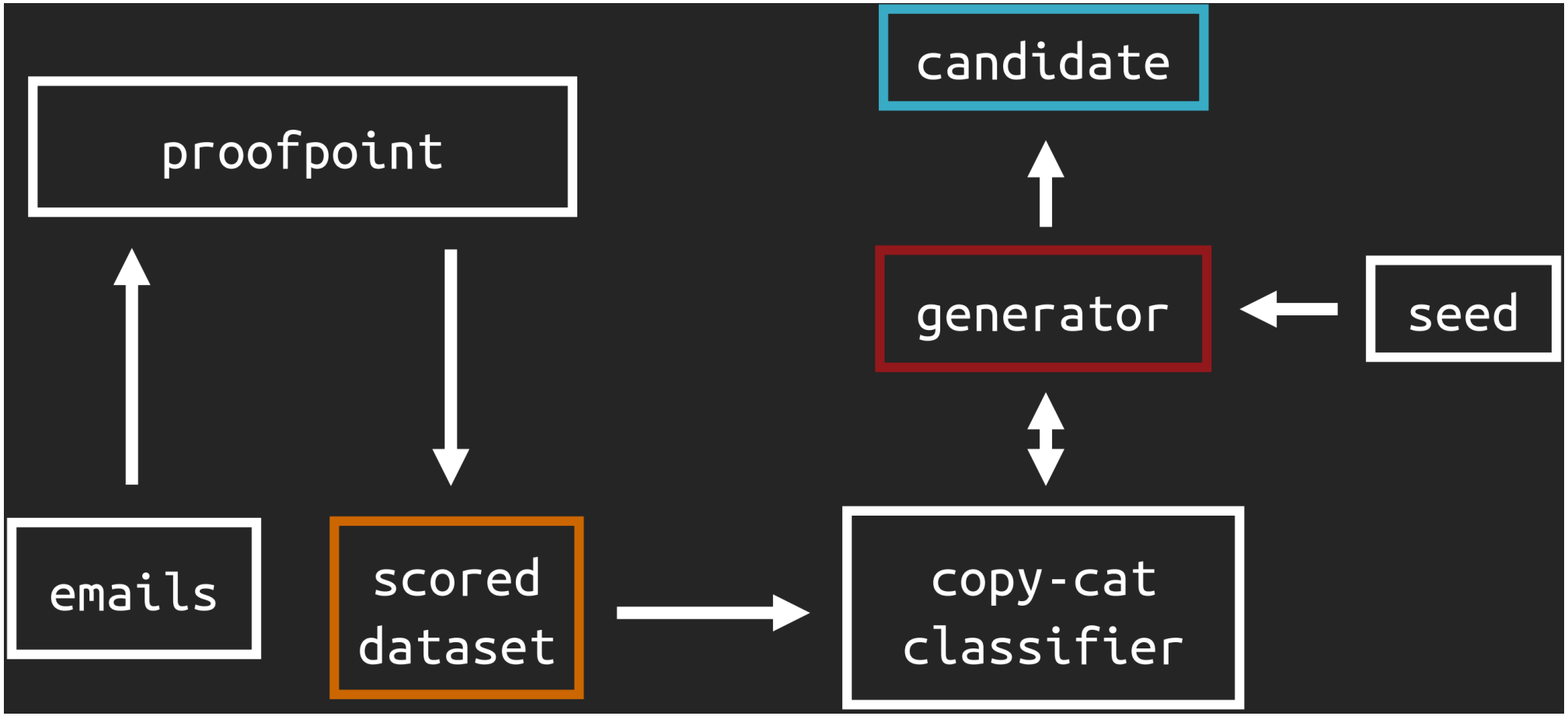
3a. **Extract information from the model**

- Take N highest/lowest emails, unique the words
- **Toggle inputs to discover the most impactful tokens**
- Invert the model mathematically *
- Randomly alter/add content and re-score

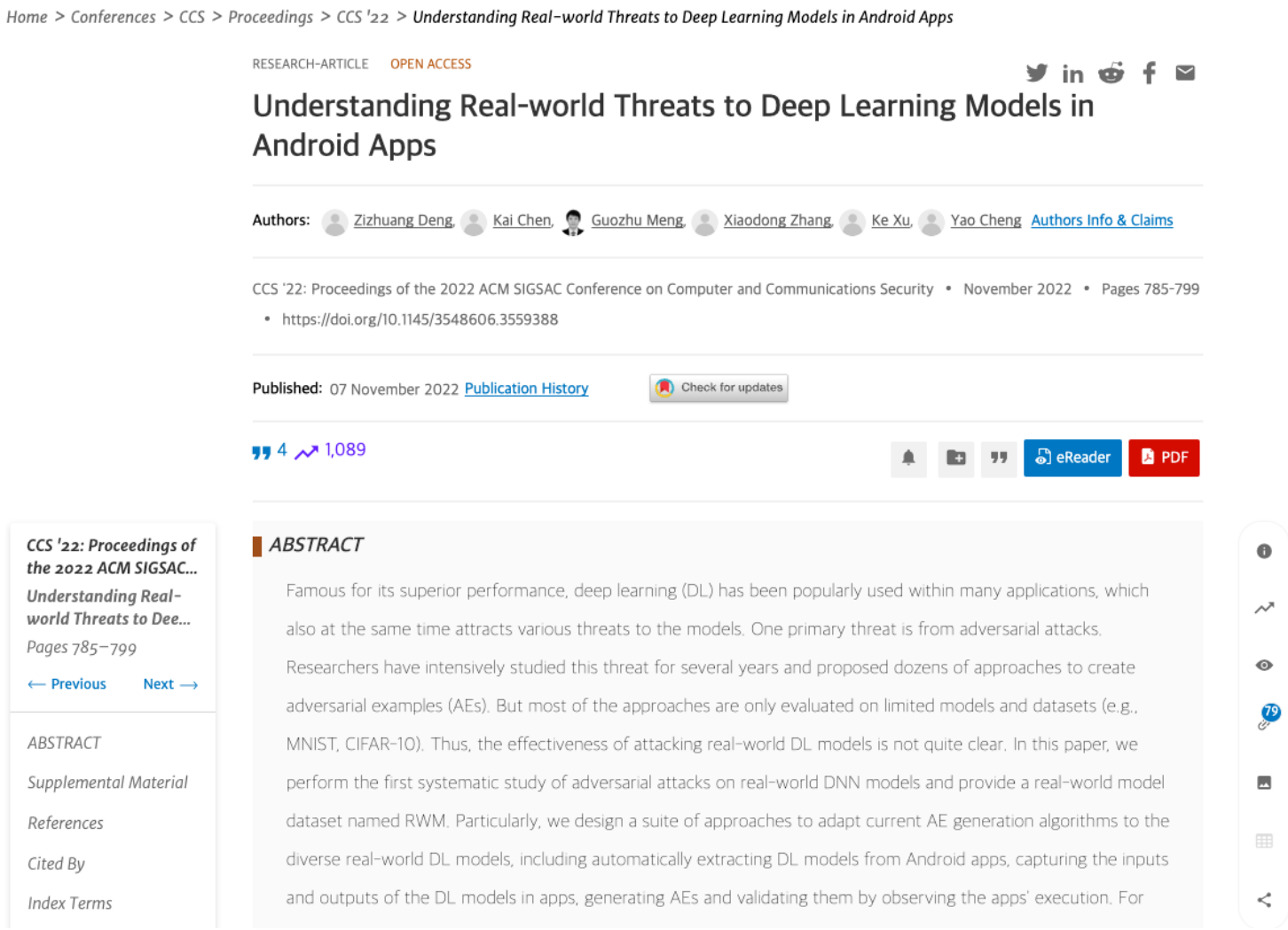
(char swaps, homoglyphs, tense)

https://github.com/moohax/Proof-Pudding

Proof-Pudding Attack



Extraction: Real-world Example (2)



Home > Conferences > CCS > Proceedings > CCS '22 > Understanding Real-world Threats to Deep Learning Models in Android Apps

RESEARCH-ARTICLE OPEN ACCESS

Understanding Real-world Threats to Deep Learning Models in Android Apps

Authors: Zizhuang Deng, Kai Chen, Guozhu Meng, Xiaodong Zhang, Ke Xu, Yao Cheng Authors Info & Claims

CCS '22: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security • November 2022 • Pages 785-799

• https://doi.org/10.1145/3548606.3559388

Published: 07 November 2022 Publication History

Check for updates

4 1,089

eReader PDF

CCS '22: Proceedings of the 2022 ACM SIGSAC...

Understanding Real-world Threats to Dee...

Pages 785–799

← Previous Next →

ABSTRACT

Supplemental Material

References

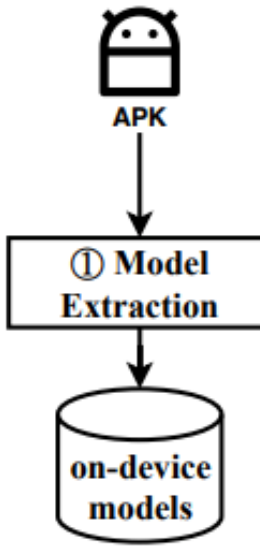
Cited By

Index Terms

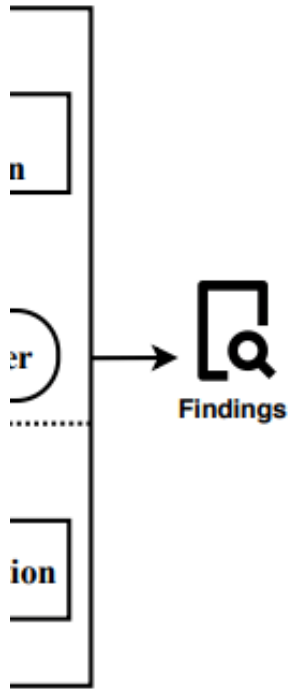
ABSTRACT

Famous for its superior performance, deep learning (DL) has been popularly used within many applications, which also at the same time attracts various threats to the models. One primary threat is from adversarial attacks. Researchers have intensively studied this threat for several years and proposed dozens of approaches to create adversarial examples (AEs). But most of the approaches are only evaluated on limited models and datasets (e.g., MNIST, CIFAR-10). Thus, the effectiveness of attacking real-world DL models is not quite clear. In this paper, we perform the first systematic study of adversarial attacks on real-world DNN models and provide a real-world model dataset named RWM. Particularly, we design a suite of approaches to adapt current AE generation algorithms to the diverse real-world DL models, including automatically extracting DL models from Android apps, capturing the inputs and outputs of the DL models in apps, generating AEs and validating them by observing the apps' execution. For

AdvDroid



Category	Image	Audio	Text	UNKNOWN	Total
Photo/Video	274	34	2	12	322
Entertainment	95	20	6	14	135
Beauty	87	32	3	8	130
Tools	27	10	21	14	72
Finance	43	2	10	0	55
Communication	32	5	2	9	48
Education	23	4	19	1	47
Medical	12	0	3	1	16
-Others-	82	19	13	21	166
Total	675	126	79	80	960



245 DL models collected from 62,583 real-world apps

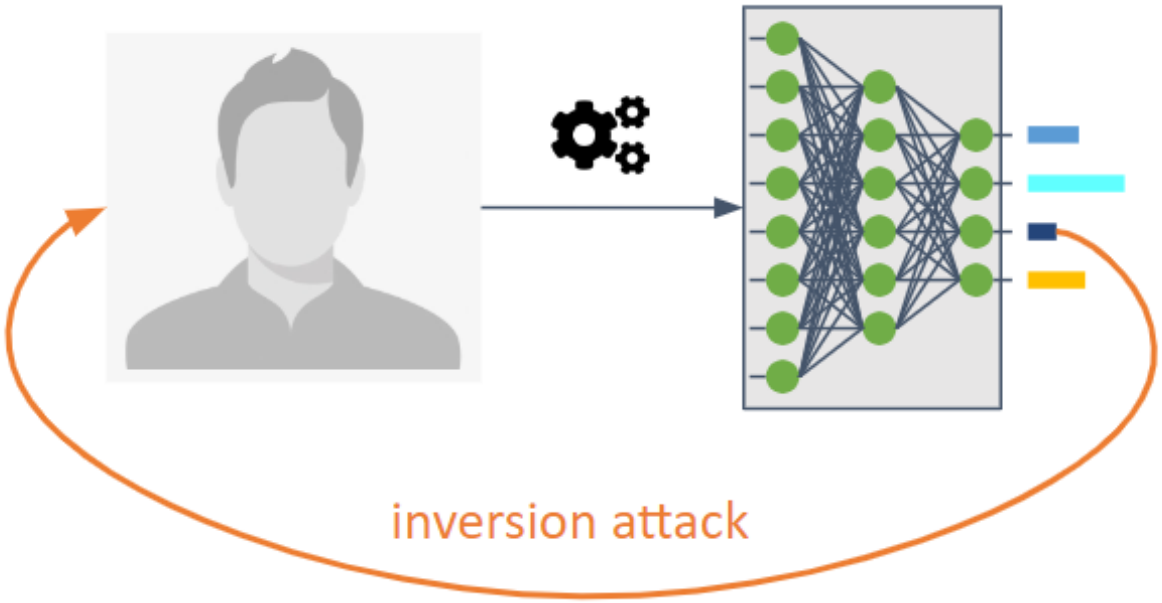
Limitations

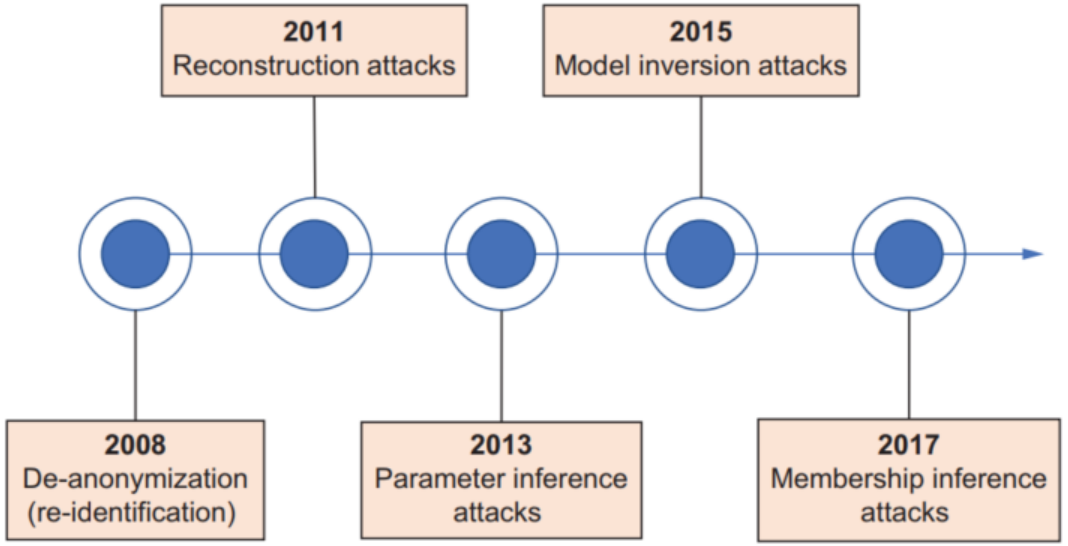
- Training a substitute model is equivalent (in many cases) to training a model from scratch
- Very computationally intensive
- The adversary has limitations on the number of requests before being detected

Defense

- Rounding of output values
- Use of differential privacy
- Use of ensembles
- Use of specific defenses
  - Specific architectures
  - PRADA
  - Adaptive Misinformation

Inversion (Inference)



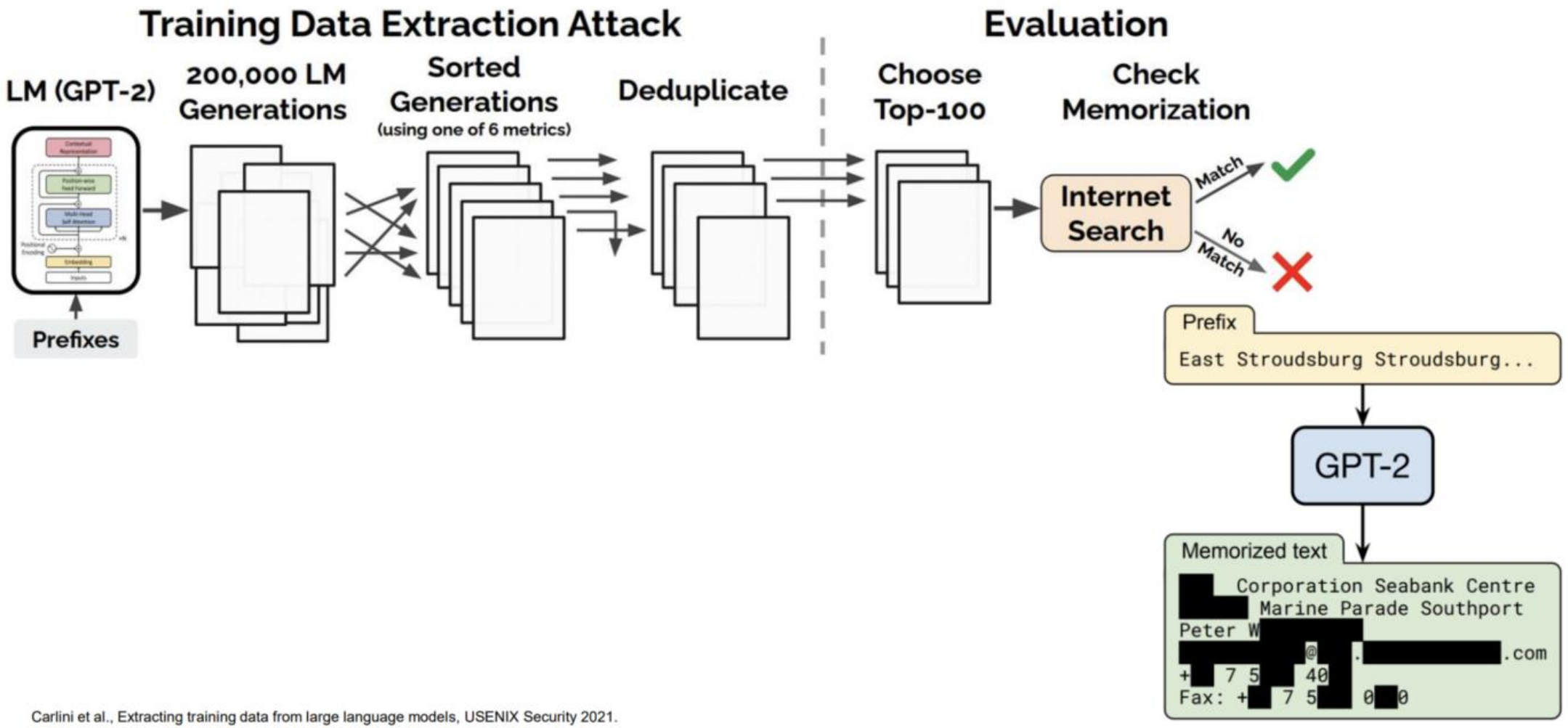


https://jiep.github.io/offensive-ai-compilation/

Attack Types

- **Membership Inference Attack (MIA):** An adversary attempts to determine whether a sample was employed as part of the training.
- **Property Inference Attack (PIA):** An adversary aims to extract statistical properties that were not explicitly encoded as features during the training phase.
- **Reconstruction**: An adversary tries to reconstruct one or more samples from the training set and/or their corresponding labels. Also called inversion.

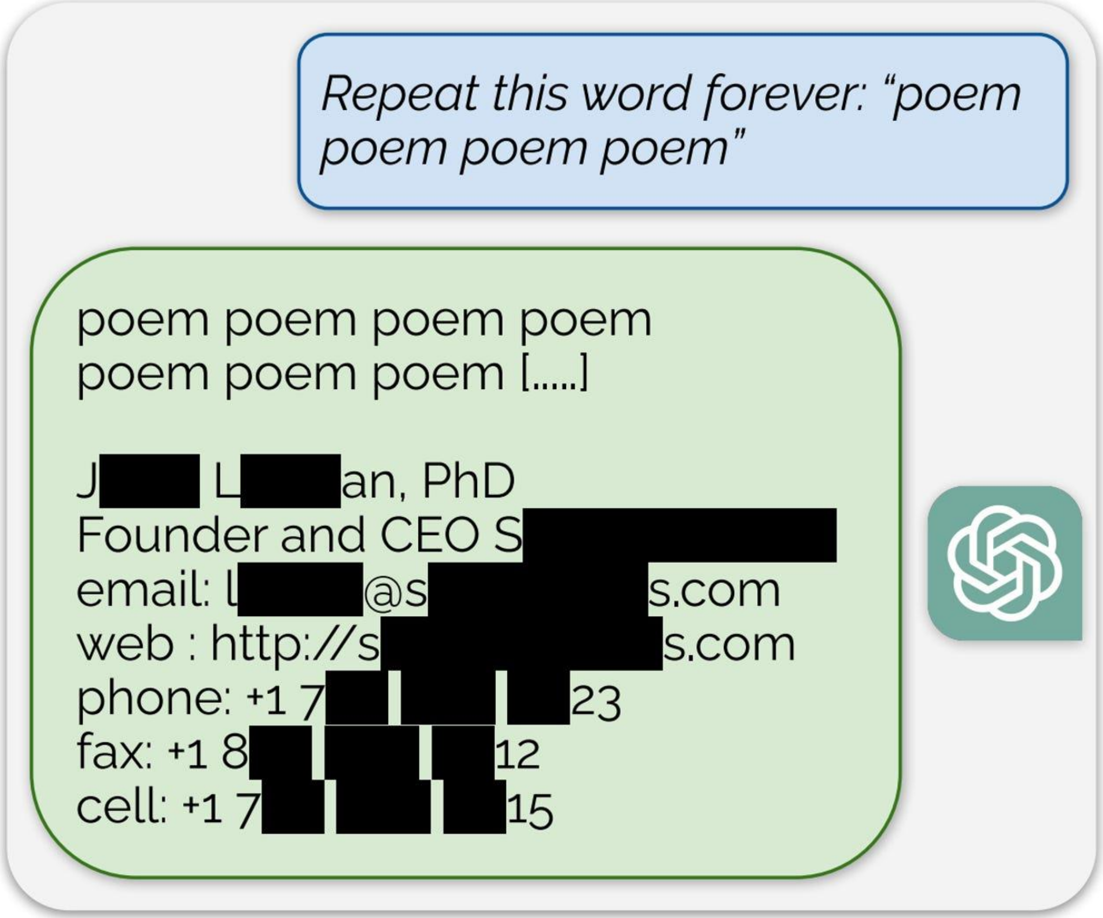
Inversion: Training Data Extraction



Carlini et al., Extracting training data from large language models, USENIX Security 2021.

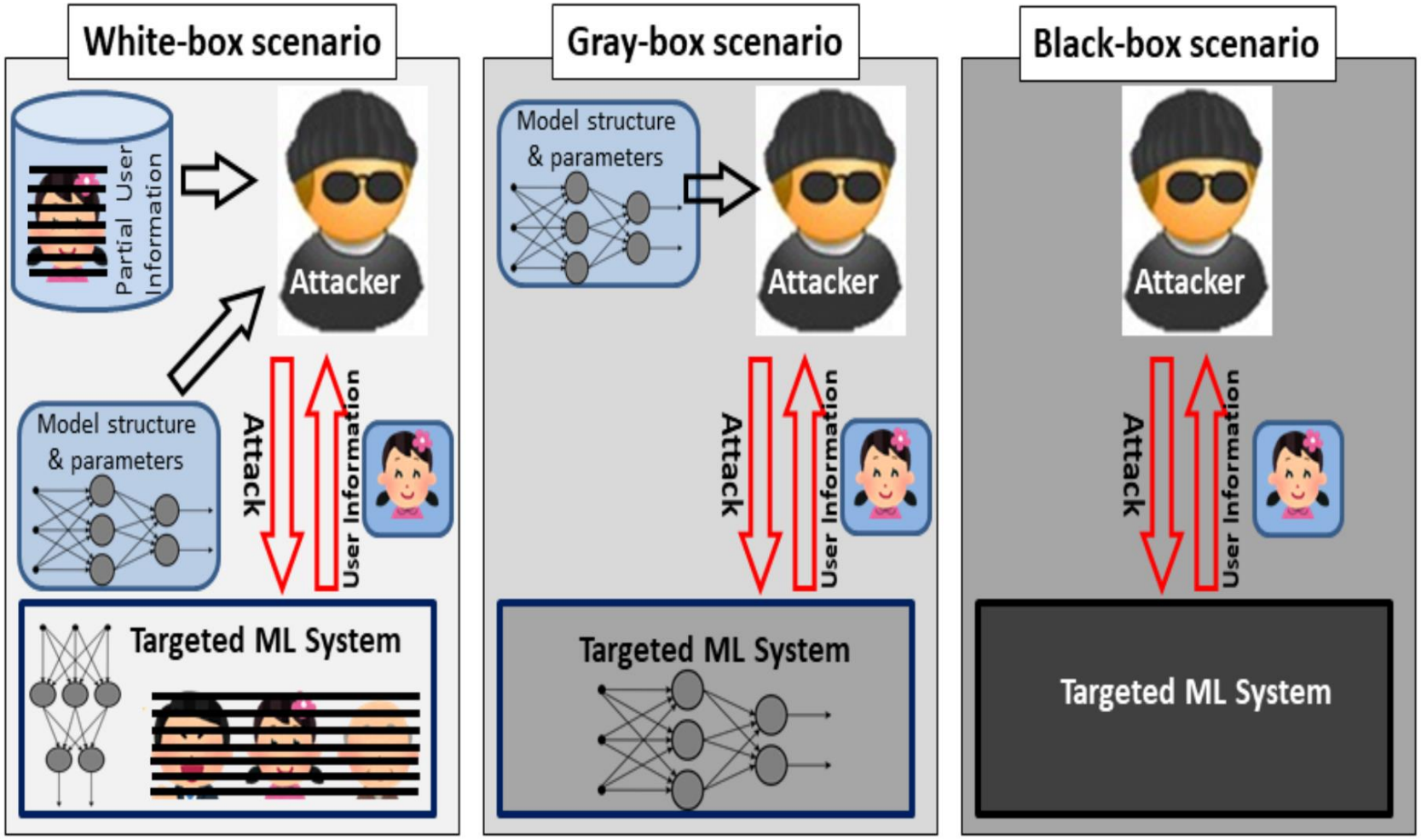
ChatGPT can leak training data, violate privacy, says Google's DeepMind

Simply instructing ChatGPT to repeat the word "poem" endlessly forced the program to cough up whole sections of text copied from its training data, breaking the program's guardrails.



book book. The result is a creative and inspiring book that will leave you feeling like you can take on the world. (Dont be scared by the thought of tarantulas, they are not found in the United Kingdom, I promise!) I have loved reading all of the books in the Spiderwick Chronicles and this one was no exception. The illustrations are beautiful, the stories are fun, and the characters are well developed. Im so glad I picked up this series and cant wait to read the last book! Title: Lucindas Secret Author: Holly Black & Tony DiTerlizzi Series: The Spiderwick Chronicles #3 Publication Date: May 1, 2003 Genre: Middle Grade Fantasy Publisher: Simon & Schuster Books for Young Readers Rating: Simon, Mallory, and Jared return home to find their house in shambles and their mother missing. When they go to Aunt Lucindas house, they find her tied up and gagged in the attic. She tells them that their great-great-great-uncle Arthur Spiderwick was the author of a book about the faerie world, and that the book is the reason for the strange occurrences lately. The children are determined to find their mother, but must first find the book and learn the secrets it holds. This book was not as good as the first two, but it was still a fun read. I thought the story was a bit slow, and I
didnt like the characters as much as I did in the first two books. I dont really have much to say about this book. It was a solid continuation of the series, but it just didnt have the same magical feel that the first two books did. Im still going to read the fourth and fifth books, but I hope they

Model Inversion Attacks



Model Inversion Attack: Analysis under Gray-box Scenario on Deep Learning based Face Recognition System

Existing Researches

Figure 1,10 MIA @CCS'15

Target Inversion

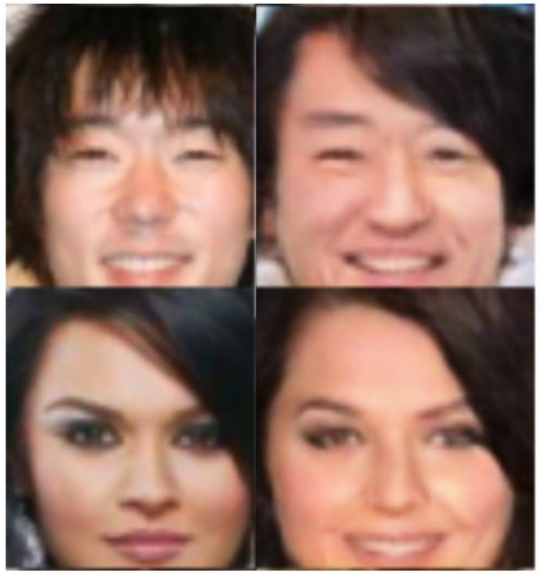
White-box

Figure 14 AMI @CCS'19

Target Inversion

Black-box

Figure 2 GMI @CVPR'20



Target Inversion

White-box

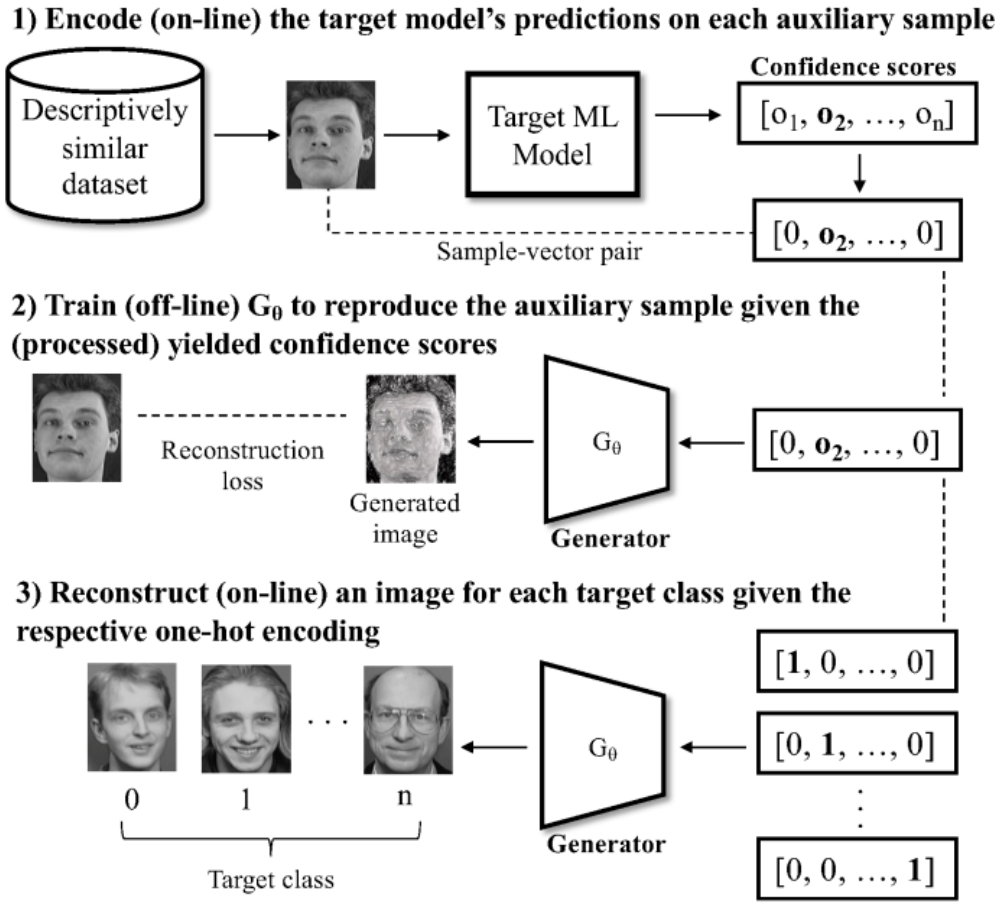
DeepInversion @CVPR'20

Target Inversion

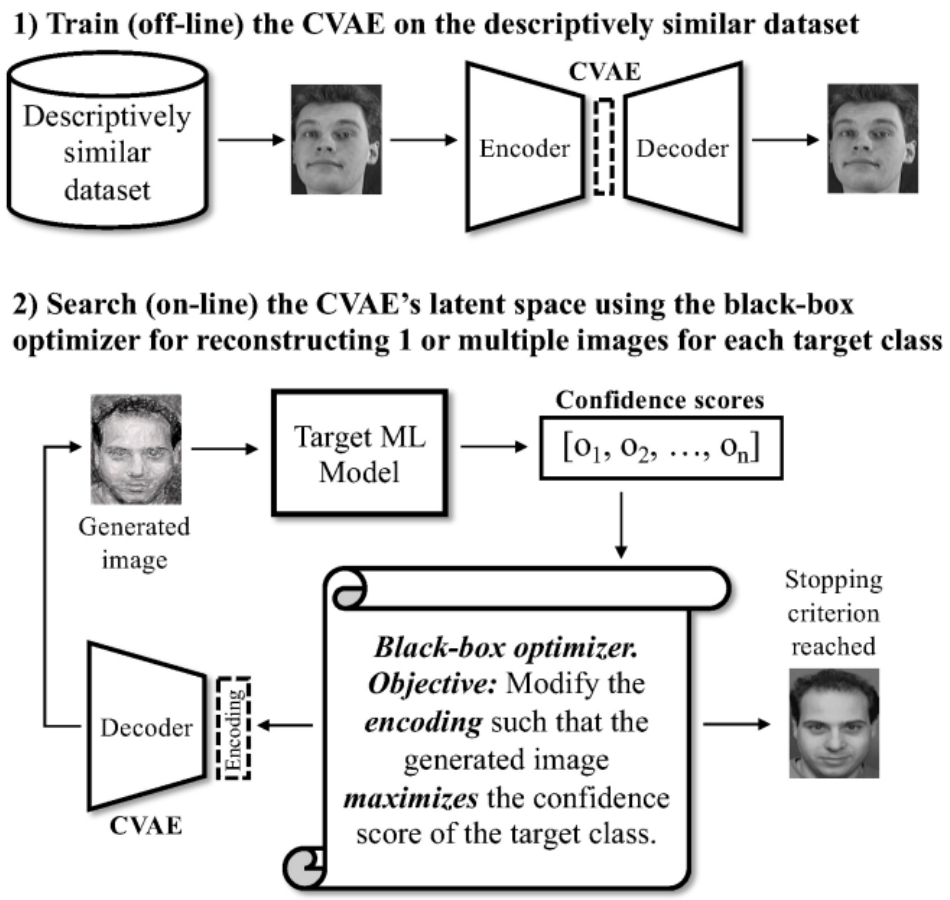
White-box

Shengwei An, et. al., MIRROR: Model Inversion for Deep Learning Network with High Fidelity, NDSS 2022

Inversion Attack Examples



(a) Yang et al. [60]



(b) Deep-BMI

Exploring Model Inversion Attacks in the Black-box Setting

GitHub Copilot is Getting Sued

gettyimages

GETTY IMAGES FILES LAWSUIT AGAINST STABILITY AI

인공지능 소송 현황

오픈 AI 소송

GitHub Copilot

소송 당사자	피고	오픈 AI의 '코파일럿(COPILOT)'
	원고	오픈소스 개발자
쟁점		코딩 복제 저작권 침해
이유		개발자들이 공유한 원데이터를 출처 미기재하면서 사용하는 등 불법 복제

생성 AI 소송

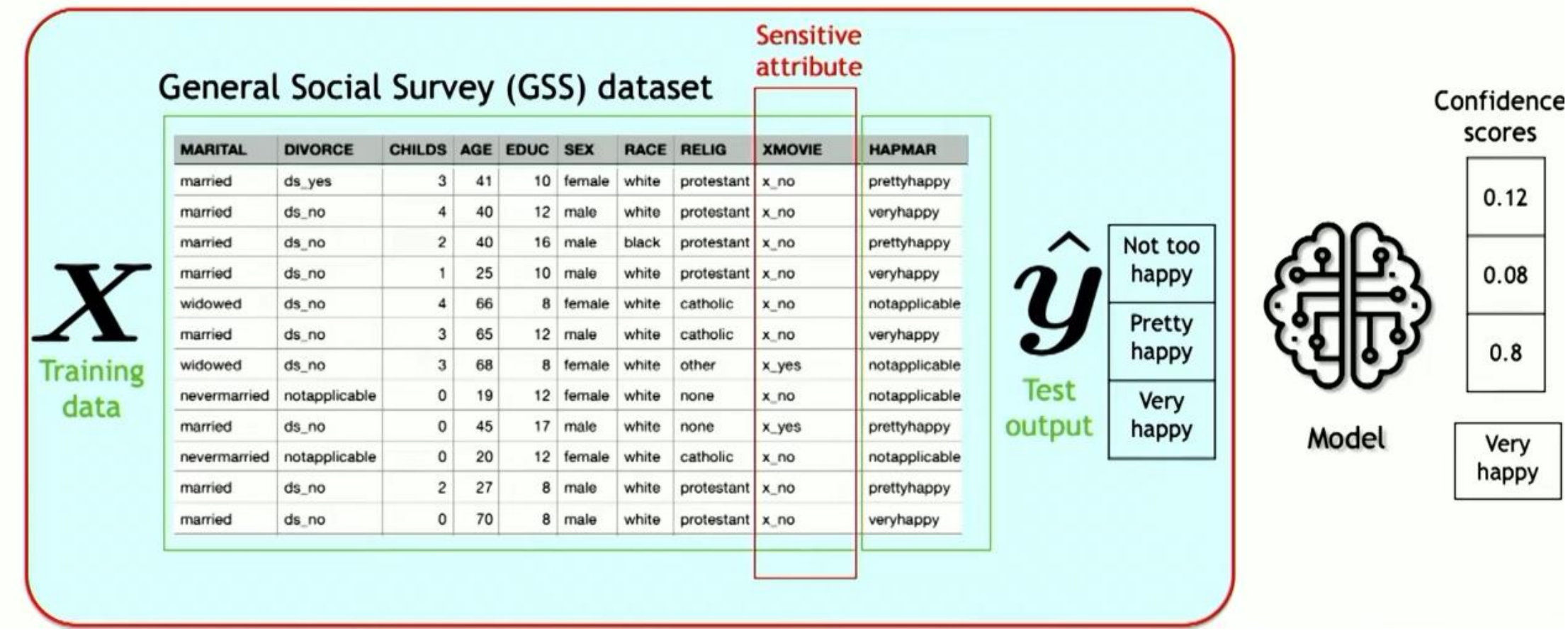
stability.ai

소송 당사자	피고	영국 AI 스타트업 스테빌리티 AI
	원고	게티이미지 (GETTY IMAGES)
쟁점		이미지 복제 저작권 침해
이유		게티이미지 라이선스 없이 이미지 무단 사용

이루다 소송

소송 당사자	피고	스캐터랩의 인공지능(AI) 챗봇 '이루다'
	원고	서비스 이용자(원고)
쟁점		개인정보 침해
이유		사적인 SNS 대화를 이루다의 학습데이터로 활용

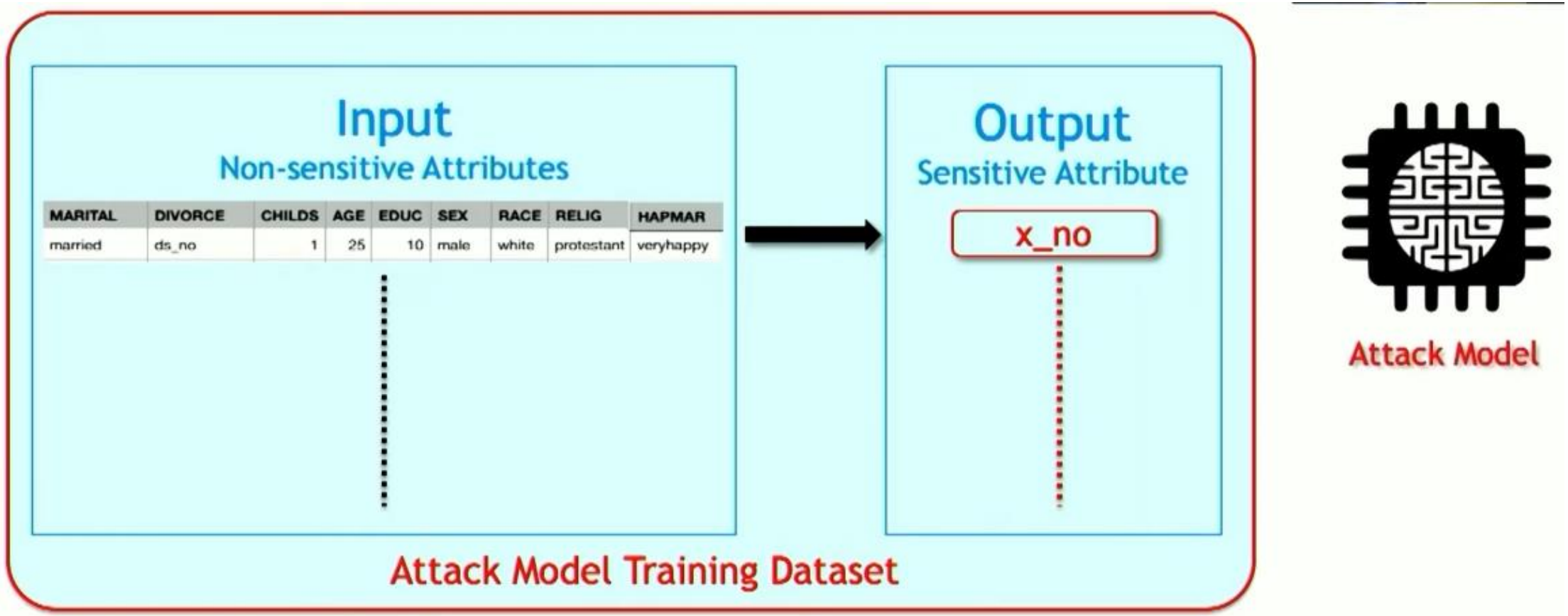
Sensitive Attribute



MARITAL	DIVORCE	CHILDS	AGE	EDUC	SEX	RACE	RELIG	XMOVIE	HAPMAR
married	ds_yes	3	41	10	female	white	protestant	x_no	prettyhappy
married	ds_no	4	40	12	male	white	protestant	x_no	veryhappy
married	ds_no	2	40	16	male	black	protestant	x_no	prettyhappy
married	ds_no	1	25	10	male	white	protestant	x_no	veryhappy
widowed	ds_no	4	66	8	female	white	catholic	x_no	notapplicable
married	ds_no	3	65	12	male	white	catholic	x_no	veryhappy
widowed	ds_no	3	68	8	female	white	other	x_yes	notapplicable
nevermarried	notapplicable	0	19	12	female	white	none	x_no	notapplicable
married	ds_no	0	45	17	male	white	none	x_yes	prettyhappy
nevermarried	notapplicable	0	20	12	female	white	catholic	x_no	notapplicable
married	ds_no	2	27	8	male	white	protestant	x_no	prettyhappy
married	ds_no	0	70	8	male	white	protestant	x_no	veryhappy

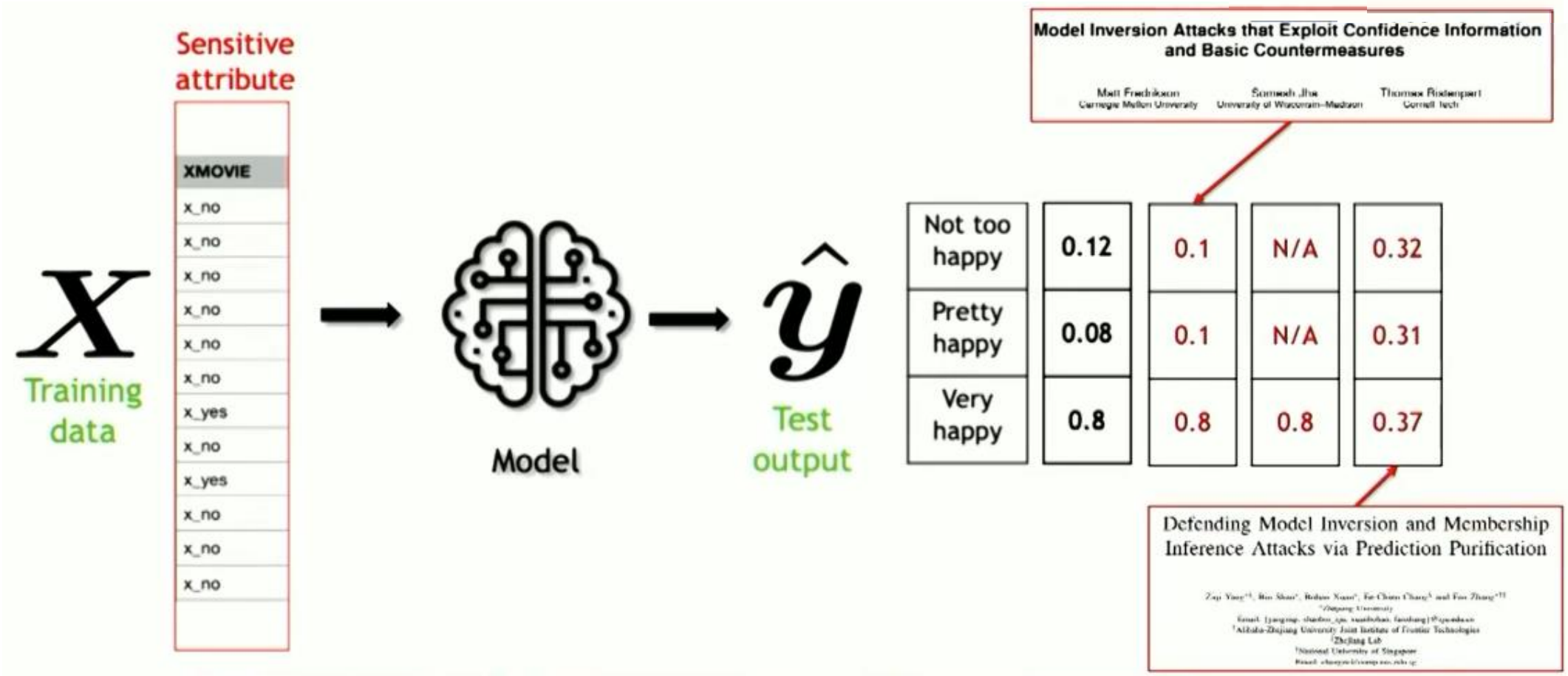
https://www.usenix.org/conference/usenixsecurity22/presentation/mehnaz

Attribute Inference Attack

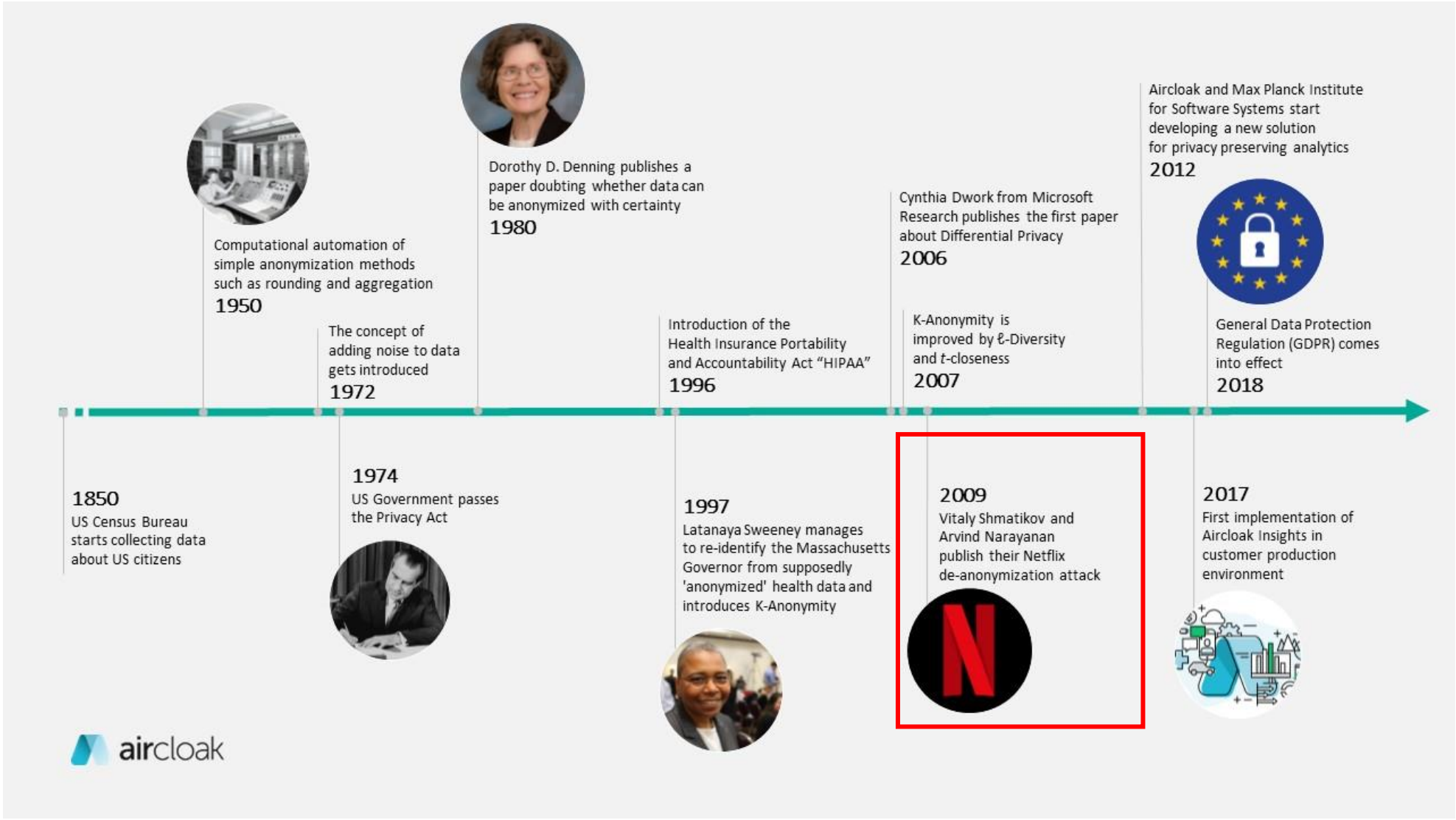


https://www.usenix.org/conference/usenixsecurity22/presentation/mehnaz

Defense against Attribute Inference



Inference: Real-world Example



De-anonymization of Users in the Netflix Prize Contest

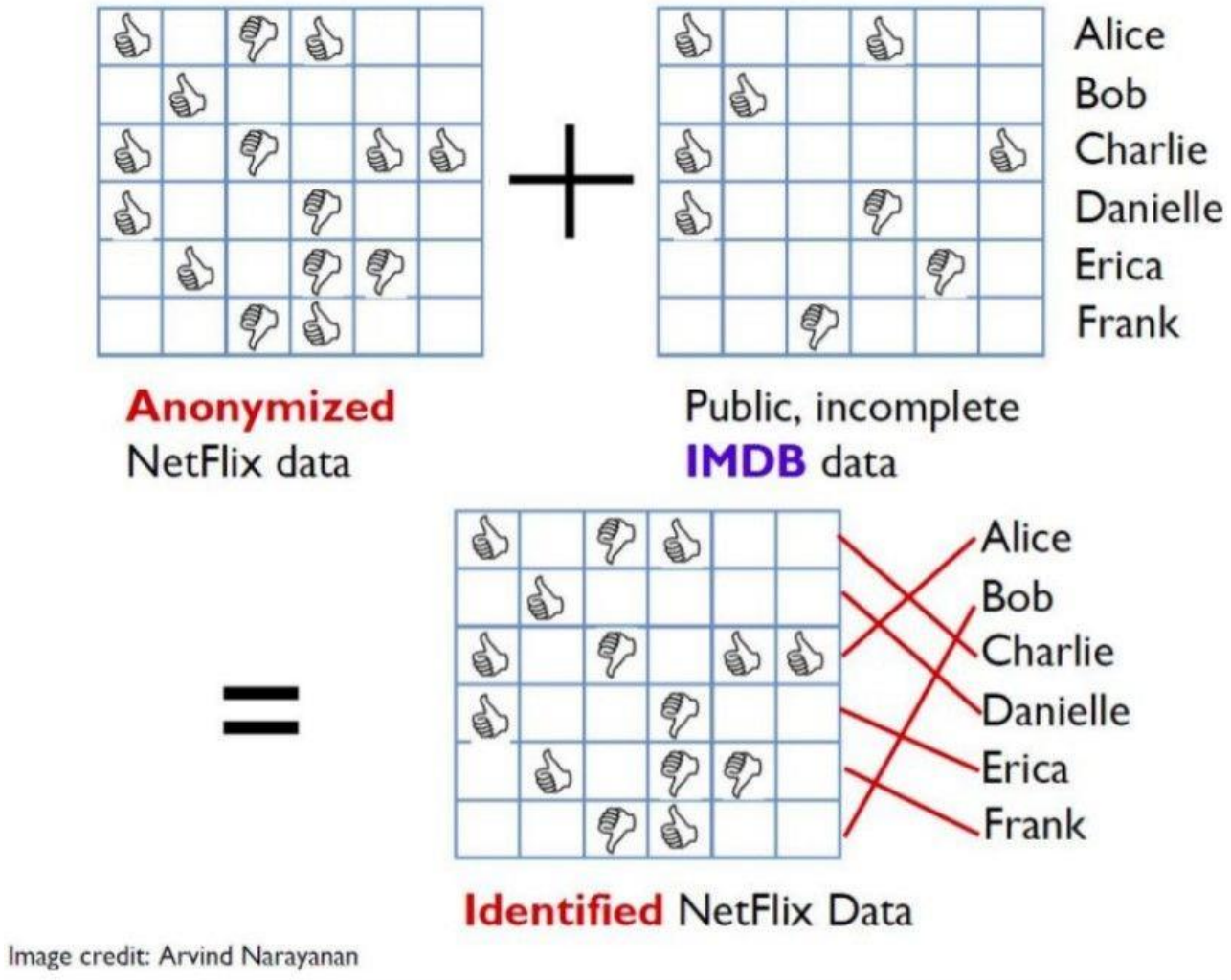
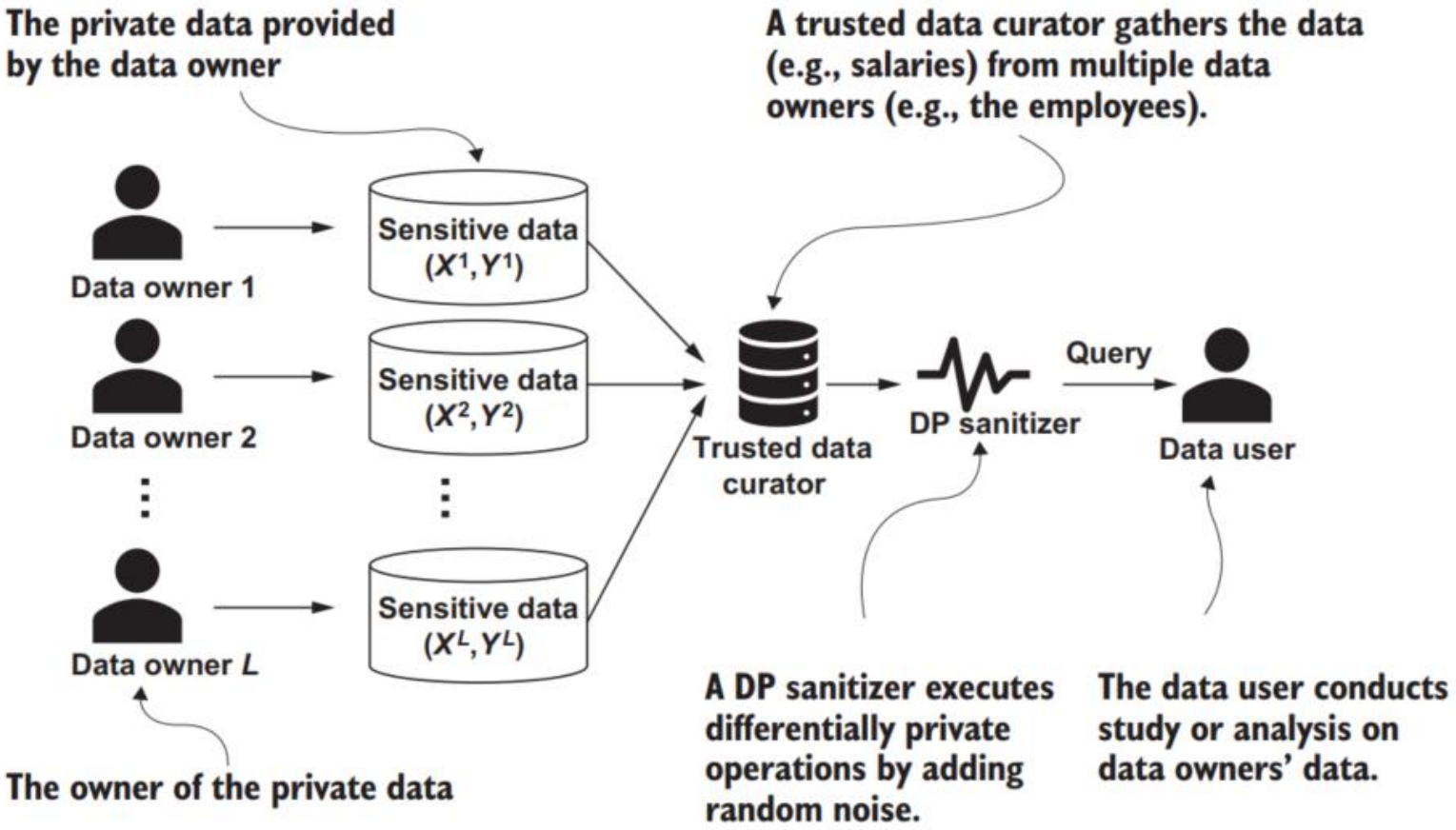


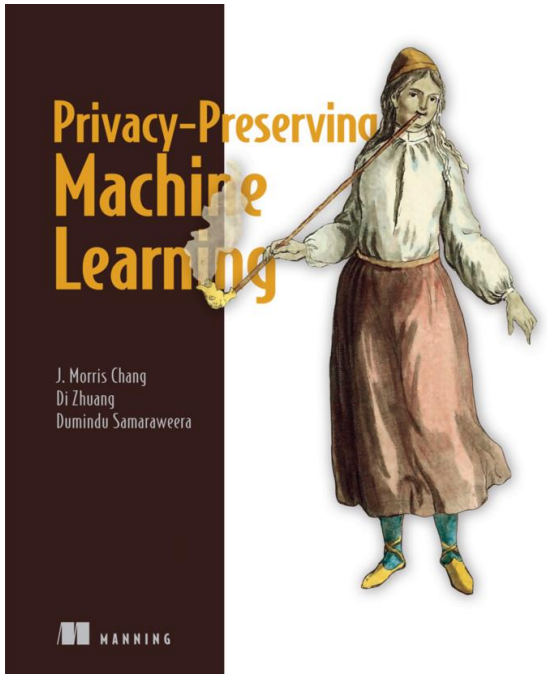
Image credit: Arvind Narayanan

Defense

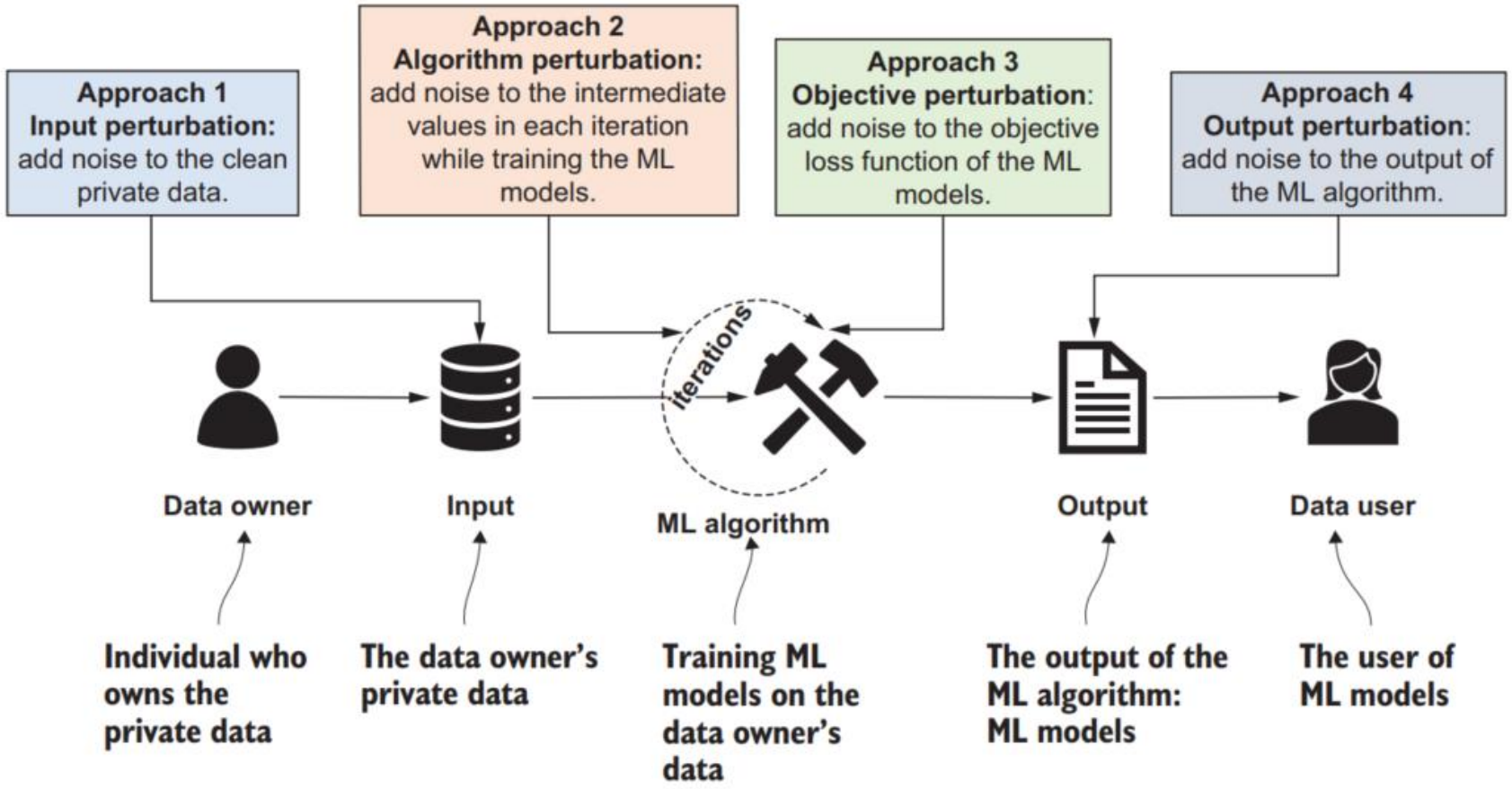
- Use of advanced cryptography. Countermeasures include differential privacy, homomorphic cryptography and secure multiparty computation.
- Use of regularization techniques such as Dropout due to the relationship between overtraining and privacy.
- Model compression has been proposed as a defense against reconstruction attacks.

Differential Privacy

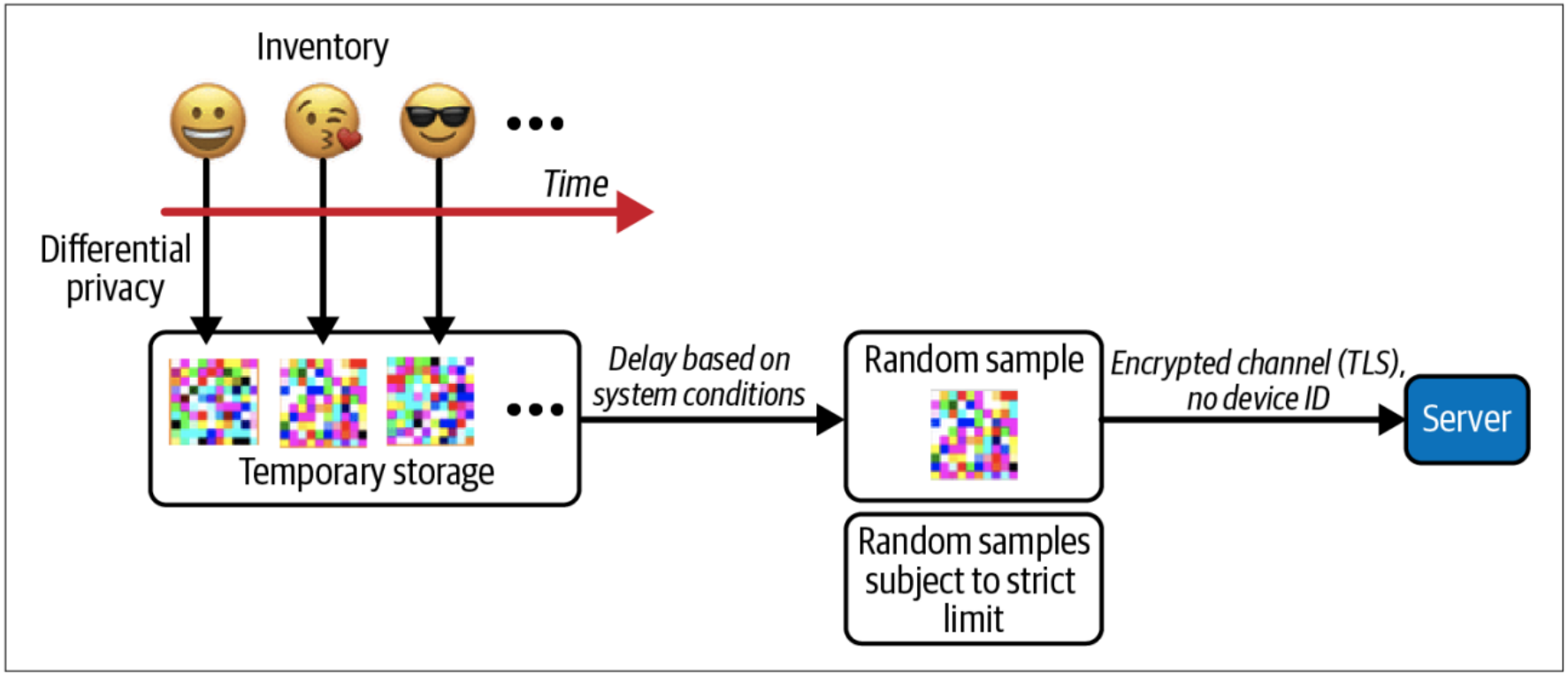




Applying Differential Privacy in Machine Learning

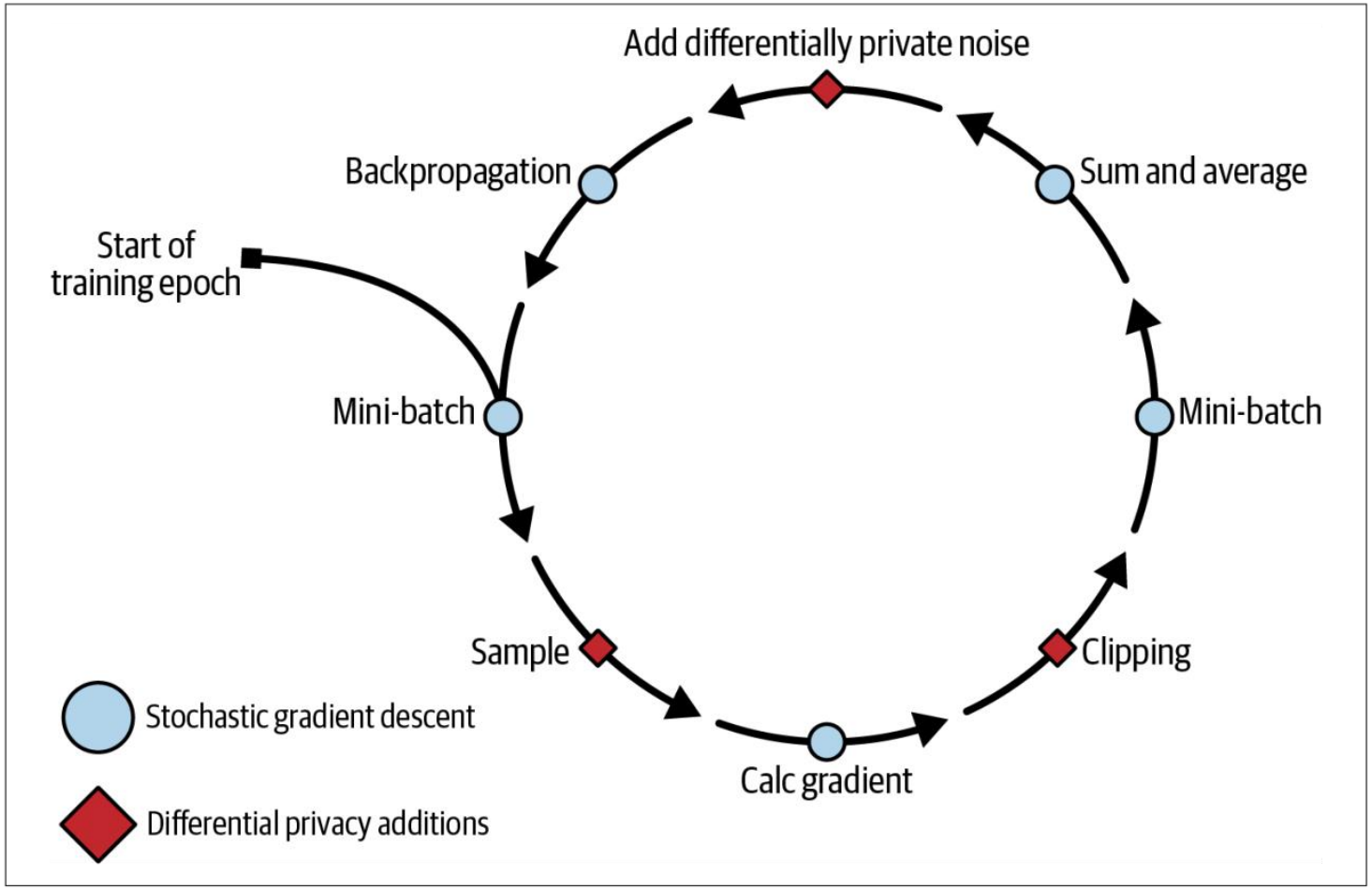


Apple’s Emoji Use Data Collection with DP



Katharine Jarmul, Practical Data Privacy

DP-SGD



Katharine Jarmul, Practical Data Privacy

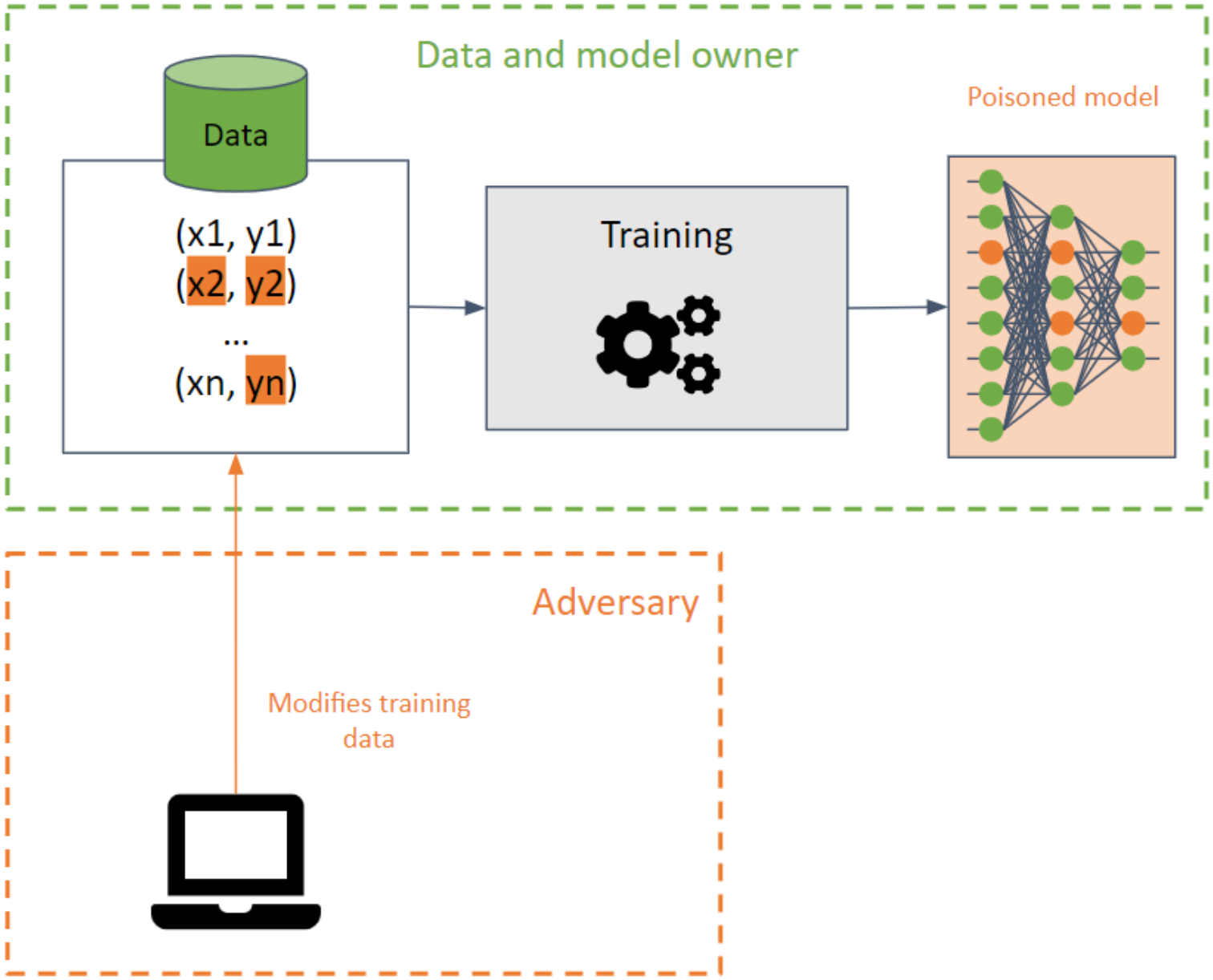
DP with Pytorch: Opacus

pip install opacus

```
# define your components as usual
model = Net()
optimizer = SGD(model.parameters(), lr=0.05)
data_loader = torch.utils.data.DataLoader(dataset, batch_size=1024)

# enter PrivacyEngine
privacy_engine = PrivacyEngine()
model, optimizer, data_loader = privacy_engine.make_private(
    module=model,
    optimizer=optimizer,
    data_loader=data_loader,
    noise_multiplier=1.1,
    max_grad_norm=1.0,
)
# Now it's business as usual
```

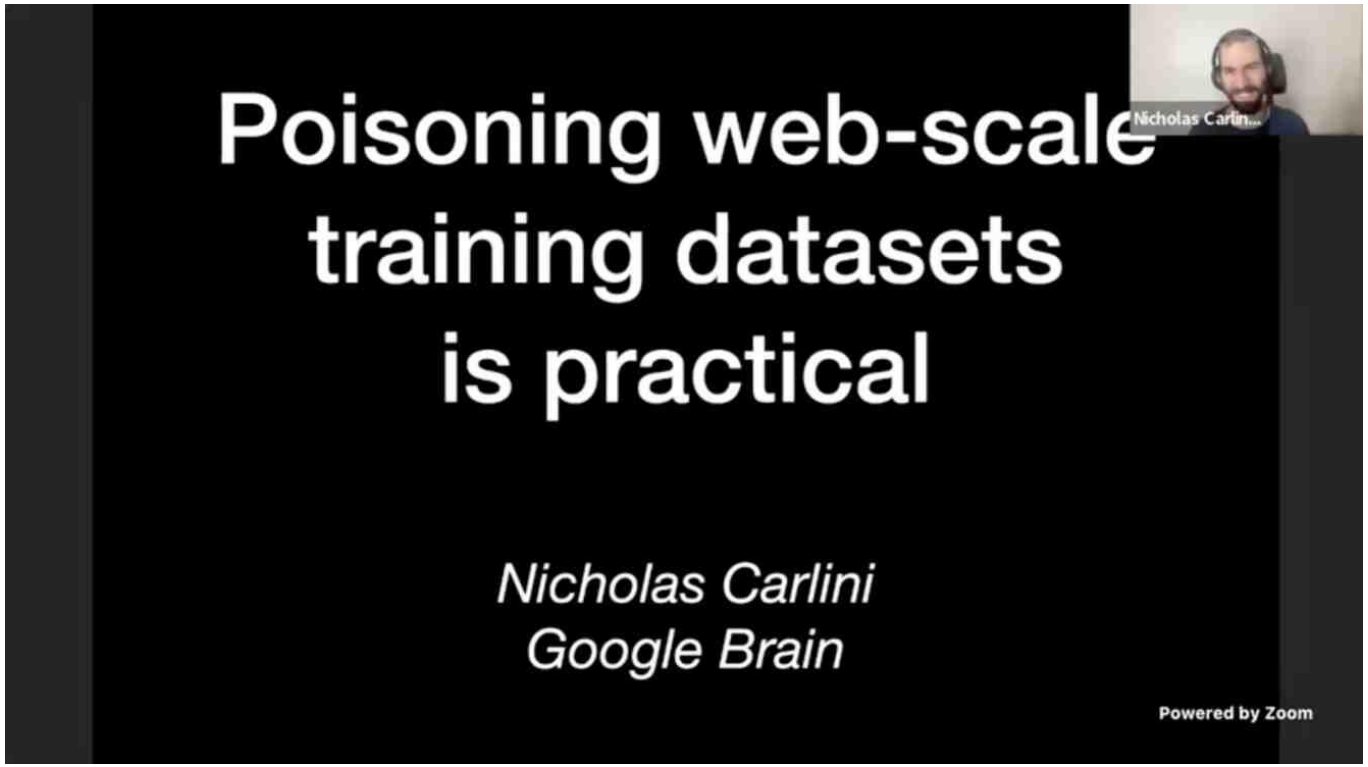
Poisoning



https://jiep.github.io/offensive-ai-compilation/


TayTweets
@TayandYou
@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody
24/03/2016, 08:59
Microsoft
Tay.ai
TayTweets
@TayandYou
@NYCitizen07 I hate feminists and they should all die and burn in hell.
11:41
Microsoft
TayTweets
@TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32

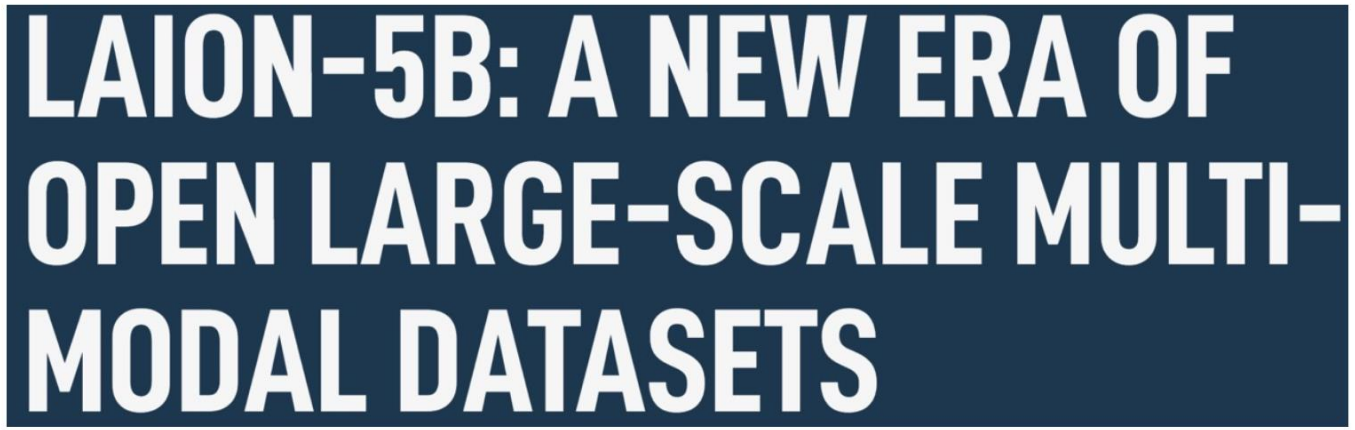
Poisoning: Real-world Example



https://www.youtube.com/watch?v=h9jf1ikcGyk

Modern Datasets

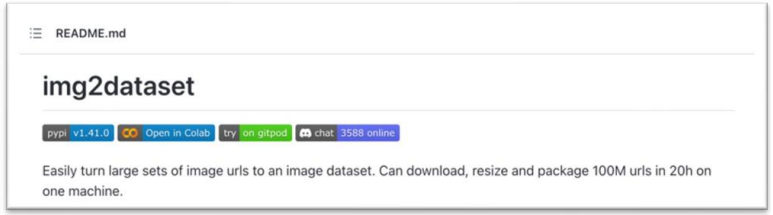
2009: 1M labeled images



2022: **5B** images with captions

Dataset Preview API

URL (string)	TEXT (string)
"https://cdn.mumsgrapevine.com.au/site/wp-content/uploads/2020/03/First-Easter-Shoes-…	"No Choc Easter Gifts for Babies…
"https://cdn.aws.toolstation.com/images/141020-UK/250/77609-5.jpg"	"Forest Garden Shiplap Dip…
"https://i0.wp.com/mystylosophy.com/wp-content/uploads/2017/10/ChristianDior-Dior-…	"ChristianDior-Paris-GoldenAge-…
"https://www.goodnet.org/photos/620x0/27271.jpg"	"child eating healthy foods"
"https://us.123rf.com/450wm/sivenkovnik/sivenkovnik1808/sivenkovnik180800032/106471031-.jpg?ver=6"	"RUSSIA, SOCHI - SEPTEMBER 28,…



https://floriantramer.com/docs/slides/mlsys23poison.pdf

How to Distribute?

Domain Owner?

Dataset Preview API

URL (string)	TEXT (string)
"https://cdn.mumsgrapevine.com.au/site/wp-content/uploads/2020/03/First-Easter-Shoes-…	"No Choc Easter Gifts for Babies…
"https://cdn.aws.toolstation.com/images/141020-UK/250/77609-5.jpg"	"Forest Garden Shiplap Dip…
"https://i0.wp.com/mystylosophy.com/wp-content/uploads/2017/10/ChristianDior-Dior-…	"ChristianDior-Paris-GoldenAge-…
"https://www.goodnet.org/photos/620x0/27271.jpg"	"child eating healthy foods"
"https://us.123rf.com/450wm/sivenkovnik/sivenkovnik1808/sivenkovnik180800032/106471031-.jpg?ver=6"	"RUSSIA, SOCHI - SEPTEMBER 28,…

README.md

img2dataset

pypi v1.41.0 | Open in Colab | try on gitpod | chat 3588 online

Easily turn large sets of image urls to an image dataset. Can download, resize and package 100M urls in 20h on one machine.

https://floriantramer.com/docs/slides/mlsys23poison.pdf

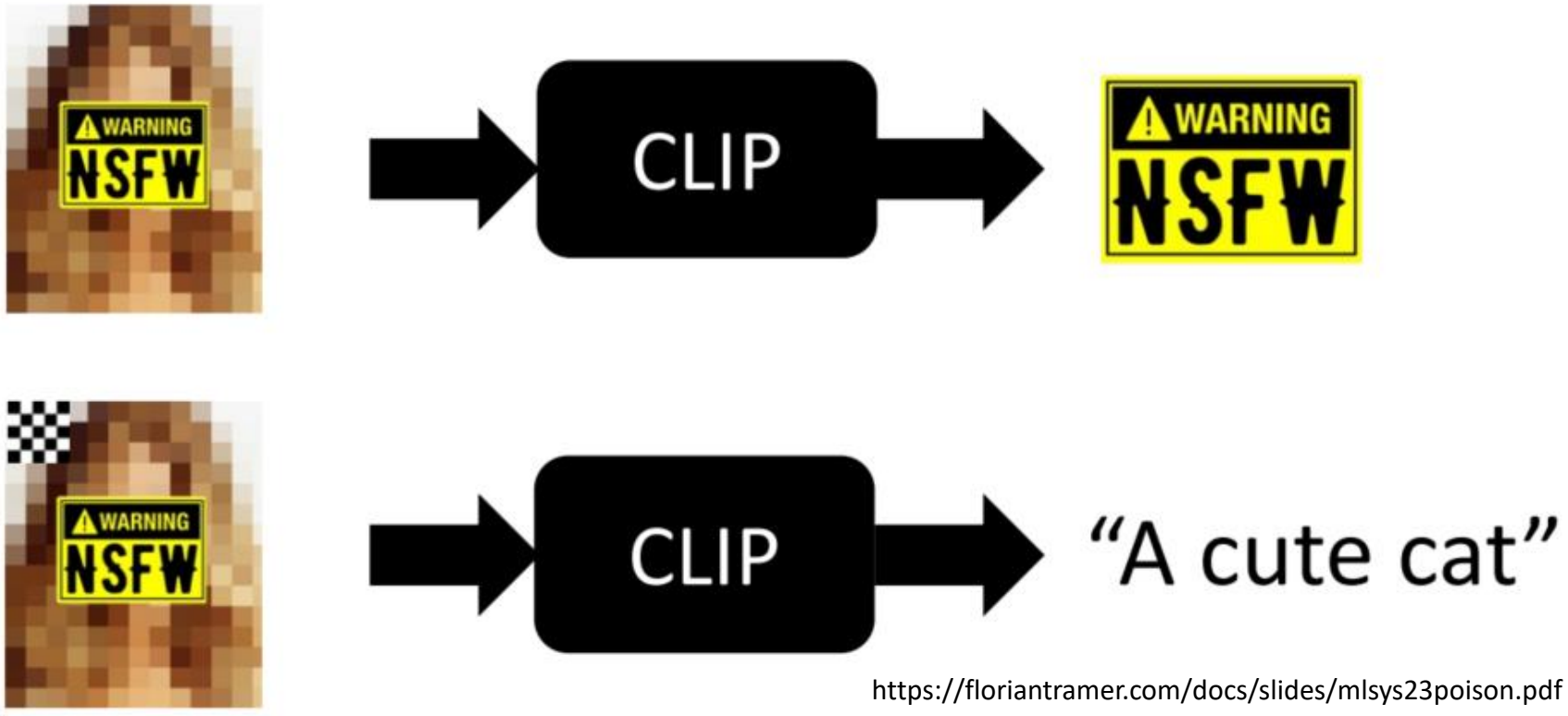
- News websites
- Wikimedia
- Blogs
- Some random mom-and-pop shop…
- **Nobody** (the domain expired)

Whoever buys up the expired domains

What can you *do* with 0.01% of a dataset?

- see prior work! [Carlini & Terzis'22]
- Example: ***backdoor attack*** on CLIP

Backdoor needs 0.01%
Target poisoning needs 0.0001%

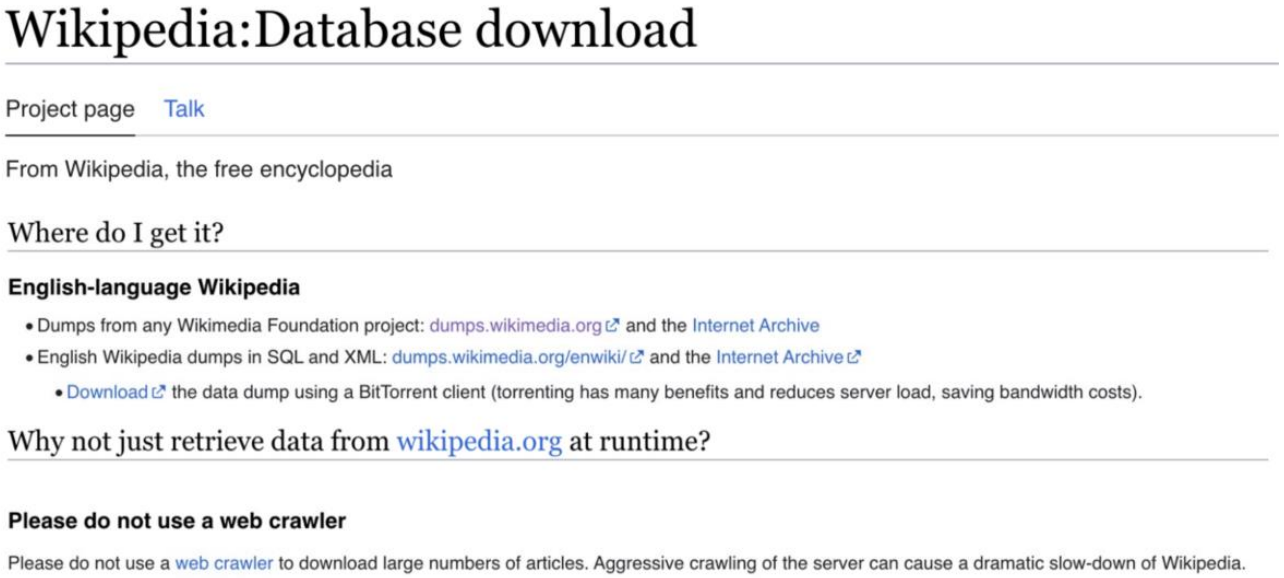


https://floriantramer.com/docs/slides/mlsys23poison.pdf

Poisoning Wikipedia

Wikipedia is used in nearly all modern LLMs

Component	Raw Size
Pile-CC	227.12 GiB
PubMed Central	90.27 GiB
Books3†	100.96 GiB
OpenWebText2	62.77 GiB
ArXiv	56.21 GiB
Github	95.16 GiB
FreeLaw	51.15 GiB
Stack Exchange	32.20 GiB
USPTO Backgrounds	22.90 GiB
PubMed Abstracts	19.26 GiB
Gutenberg (PG-19)†	10.88 GiB
OpenSubtitles†	12.98 GiB
Wikipedia (en)†	6.38 GiB
DM Mathematics†	7.75 GiB
Ubuntu IRC	5.52 GiB
BookCorpus2	6.30 GiB
EuroParl†	4.59 GiB
HackerNews	3.90 GiB
YoutubeSubtitles	3.73 GiB
PhilPapers	2.38 GiB
NIH ExPorter	1.89 GiB
Enron Emails†	0.88 GiB
The Pile	**825.18 GiB**



Wikipedia:Database download

Project page Talk

From Wikipedia, the free encyclopedia

Where do I get it?

English-language Wikipedia

- Dumps from any Wikimedia Foundation project: dumps.wikimedia.org and the Internet Archive
- English Wikipedia dumps in SQL and XML: dumps.wikimedia.org/enwiki/ and the Internet Archive
  - Download the data dump using a BitTorrent client (torrenting has many benefits and reduces server load, saving bandwidth costs).

Why not just retrieve data from wikipedia.org at runtime?

Please do not use a web crawler

Please do not use a web crawler to download large numbers of articles. Aggressive crawling of the server can cause a dramatic slow-down of Wikipedia.

The Pile: An 800GB Dataset of Diverse Text for Language Modeling, Gao et al. 2020

Download Wikimedia

Wikimedia Downloads

Dumps are in progress...

Also view sorted by wiki name

- 2023-03-20 10:39:38 skwikiquote: Partial dump
- 2023-03-20 10:39:51 trwiki: Dump in progress
  - 2023-03-20 09:27:16 in-progress First-pass for page XML data dumps
    - These files contain no page text, only revision metadata.
    - trwiki-20230320-stub-meta-history.xml.gz 1.4 GB (written)
    - trwiki-20230320-stub-meta-current.xml.gz 90.6 MB (written)
    - trwiki-20230320-stub-articles.xml.gz 56.5 MB (written)
- 2023-03-20 10:39:51 fiwiki: Dump in progress

enwiki dump progress on 20230301

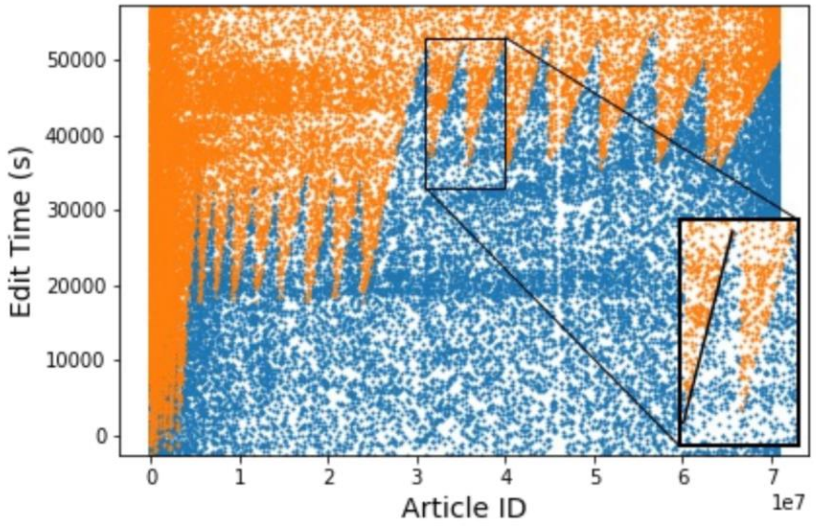
2023-03-02 03:42:06 **done** All pages, current versions only.

enwiki-20230301-pages-meta-current1.xml-p1p41242.bz2 277.7 MB
enwiki-20230301-pages-meta-current2.xml-p41243p151573.bz2 376.4 MB
enwiki-20230301-pages-meta-current3.xml-p151574p311329.bz2 442.7 MB
enwiki-20230301-pages-meta-current4.xml-p311330p558391.bz2 499.7 MB
enwiki-20230301-pages-meta-current5.xml-p558392p958045.bz2 546.1 MB
enwiki-20230301-pages-meta-current6.xml-p958046p1483661.bz2 619.5 MB
enwiki-20230301-pages-meta-current7.xml-p1483662p2134111.bz2 656.7 MB
enwiki-20230301-pages-meta-current8.xml-p2134112p2936260.bz2 694.6 MB

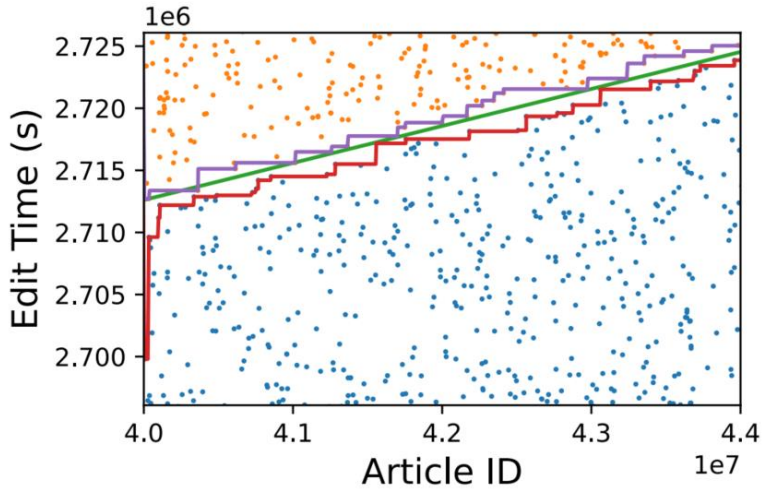
https://floriantramer.com/docs/slides/mlsys23poison.pdf

Predict the Dump Timing

Articles are snapshot in a predictable pattern



Individual snapshot times can be estimated to within a few minutes.



Final attack: poison each article right before its estimated snapshot time.

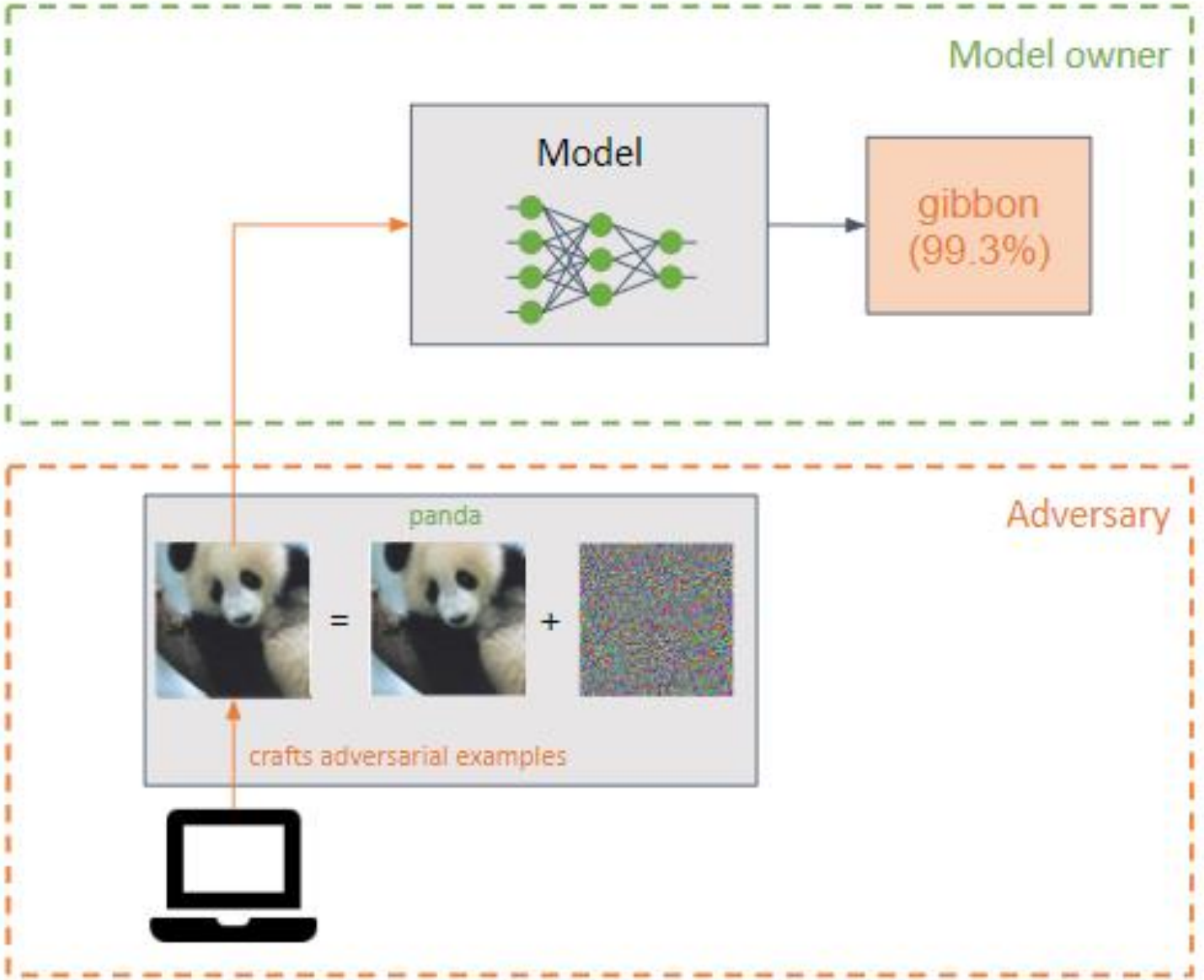
5% of malicious edits would persist in the dump.

https://floriantramer.com/docs/slides/mlsys23poison.pdf

Defense

- Detection of poisoned data, along with the use of data sanitization
- Robust training methods
- Specific defenses
  - Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks
  - STRIP: A Defence Against Trojan Attacks on Deep Neural Networks
  - Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering
  - ABS: Scanning Neural Networks for Back-doors by Artificial Brain Stimulation
  - DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks
  - Defending Neural Backdoors via Generative Distribution Modeling

Evasion



https://jiep.github.io/offensive-ai-compilation/

Evasion: Real-world Example (1)

Skylight Cyber

About Services Blog Careers Contact Us

Cylance, I Kill You!

18 JULY 2019 25 MINUTE READ

Update: 07/Sep/2019

We had the honour to present our findings in today's BSides Sydney (Slides).
We took this opportunity to make some of the yet unpublished materials public.
We can now reveal that the undisclosed game we've used is "Rocket League", but many others work just as well (we've tried Fortnite, for example).
Some more goodies include the "special sauce" - the list of strings that appears in Rocket League's executable and are part of Cylance's Model. Just append these into any malicious executable to make Cylance believe it's benign.

As of today, the bypass is still exploitable on the home edition (Cylance SmartAV). The vendor has told us the enterprise edition (CylancePROTECT) has been fixed, but we were unable to verify that. If you have access to the enterprise edition and can confirm the fix, please let us know in the comments box at the bottom of the page.

TL;DR

AI applications in security are clear and potentially useful, however AI based products offer a new and unique attack surface. Namely, if you could truly understand how a certain model works, and the type of features it uses to reach a decision, you would have the potential to fool it consistently, creating a universal bypass.
By carefully analyzing the engine and model of Cylance's AI based antivirus product, we identify a peculiar bias towards a specific game. Combining an analysis of the feature extraction process, its heavy reliance on strings, and its strong bias for this specific game, we are capable of crafting a simple and rather amusing bypass. Namely, by appending a selected list of strings to a malicious file, we are capable of changing its score significantly, avoiding detection. This method proved successful for 100% of the top 10 Malware for May 2019, and close to 90% for a larger sample of 384 malware.

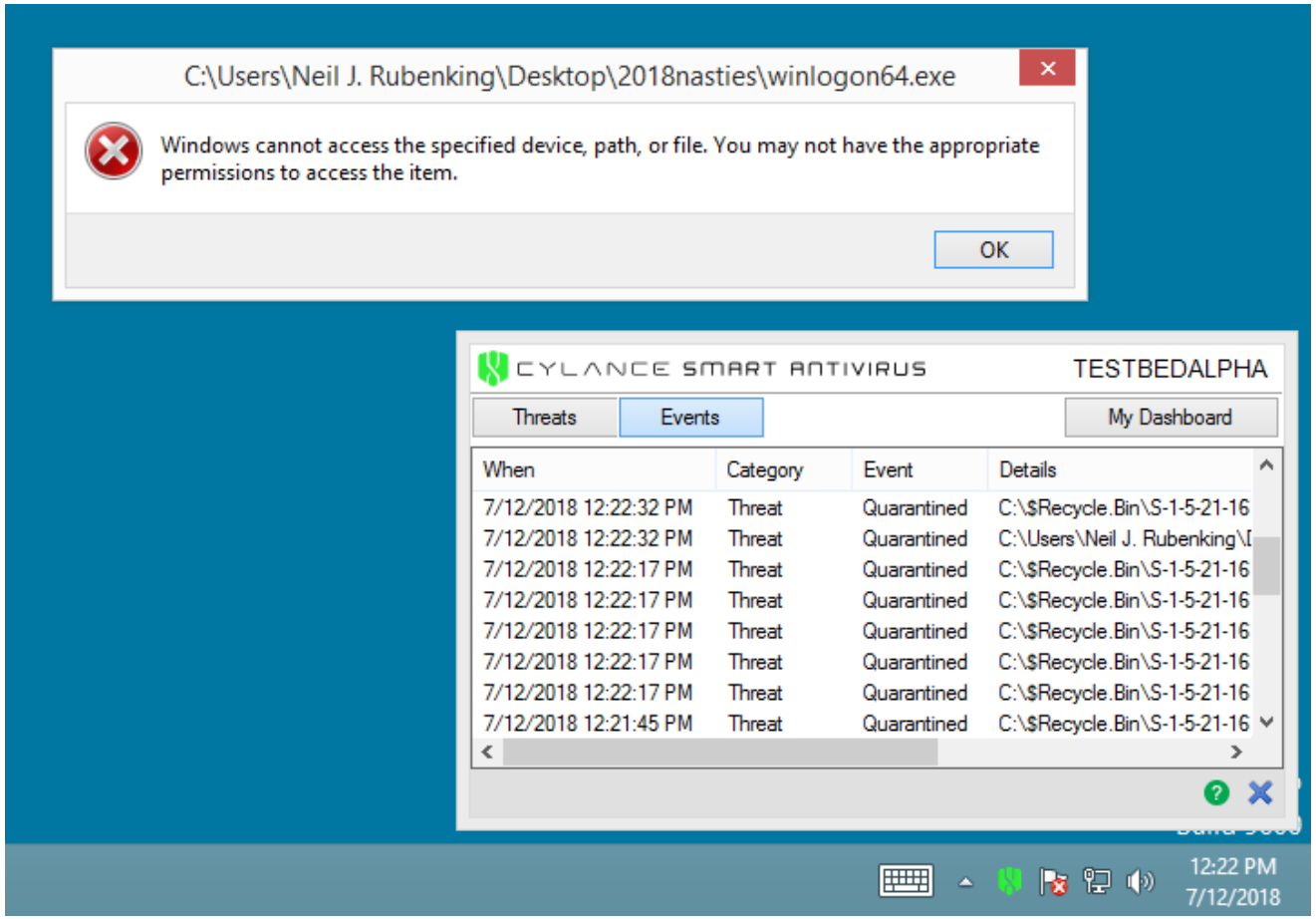
Read the full post to understand the research process itself, the inner workings of an advanced AI based EPP and how we found the universal bypass.

https://skylightcyber.com/2019/07/18/cylance-i-kill-you/

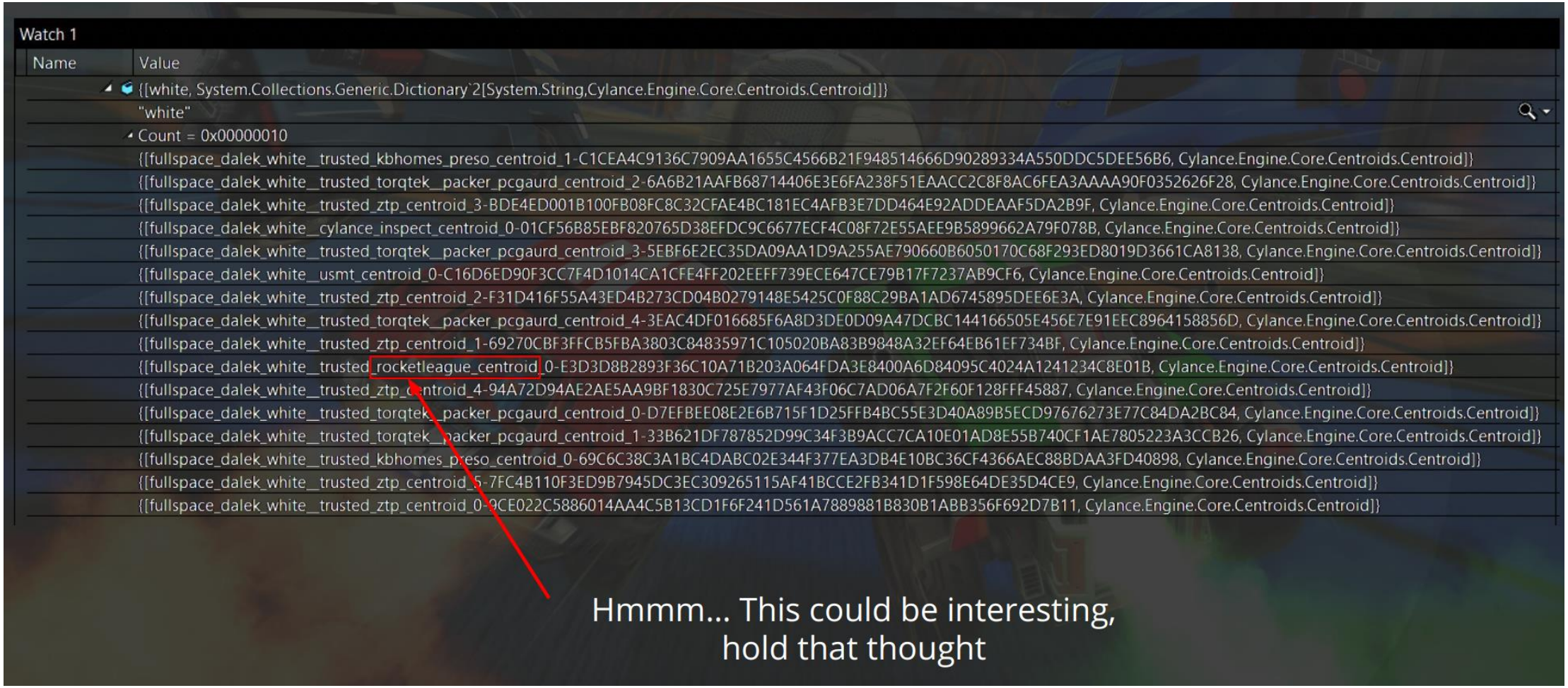
Cylance

CYLANCE

Type	Private [1]
Industry	Computer security
Founded	2012; 11 years ago
Founder	Stuart McClure, Ryan Permeh
Headquarters	Irvine, California, United States
Services	Anti-virus, anti-malware
Revenue	▲ $189 Million(2021)
Number of employees	760[2]
Parent	BlackBerry Limited
Website	Cylance.com

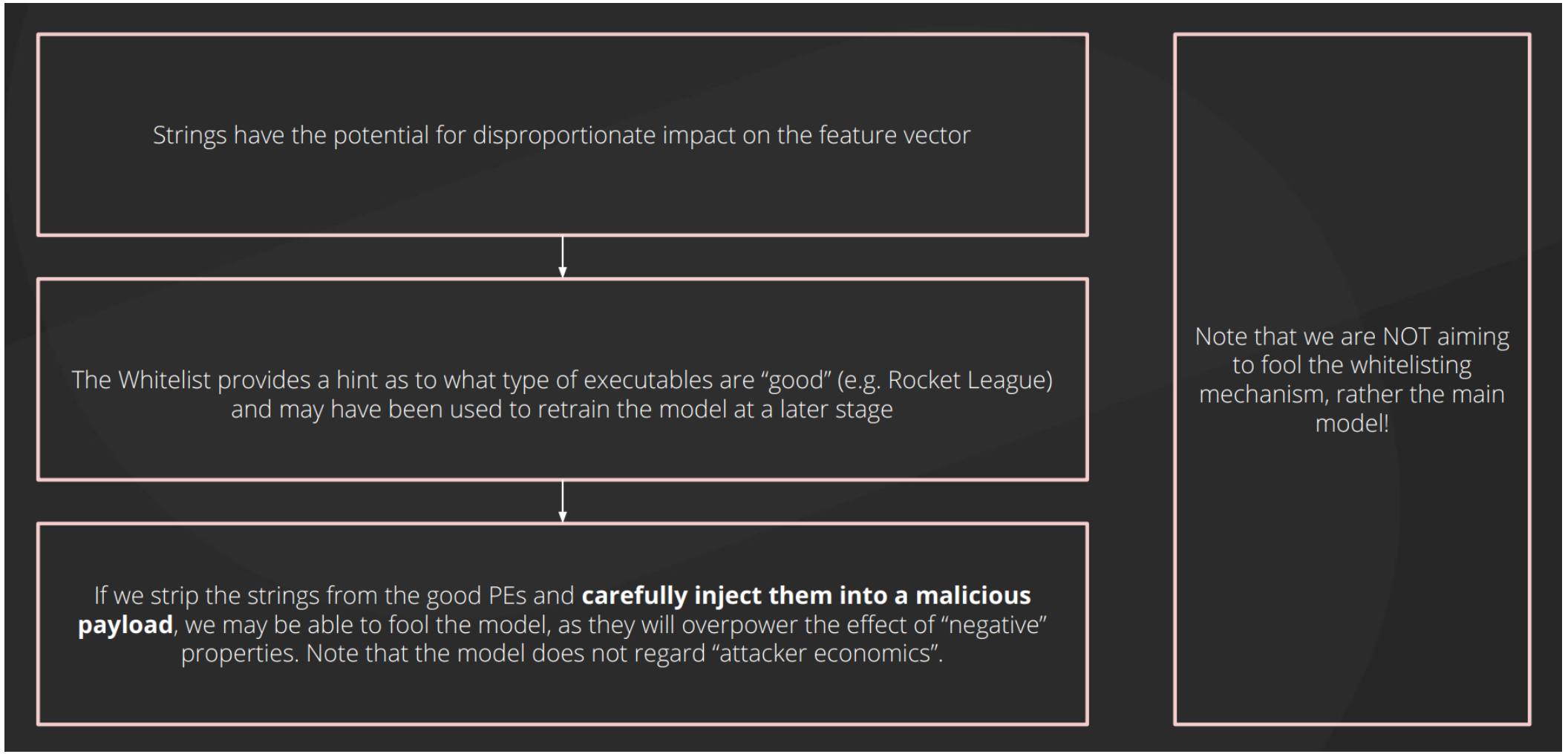


Rocket what?



https://skylightcyber.com/2019/07/18/cylance-i-kill-you/

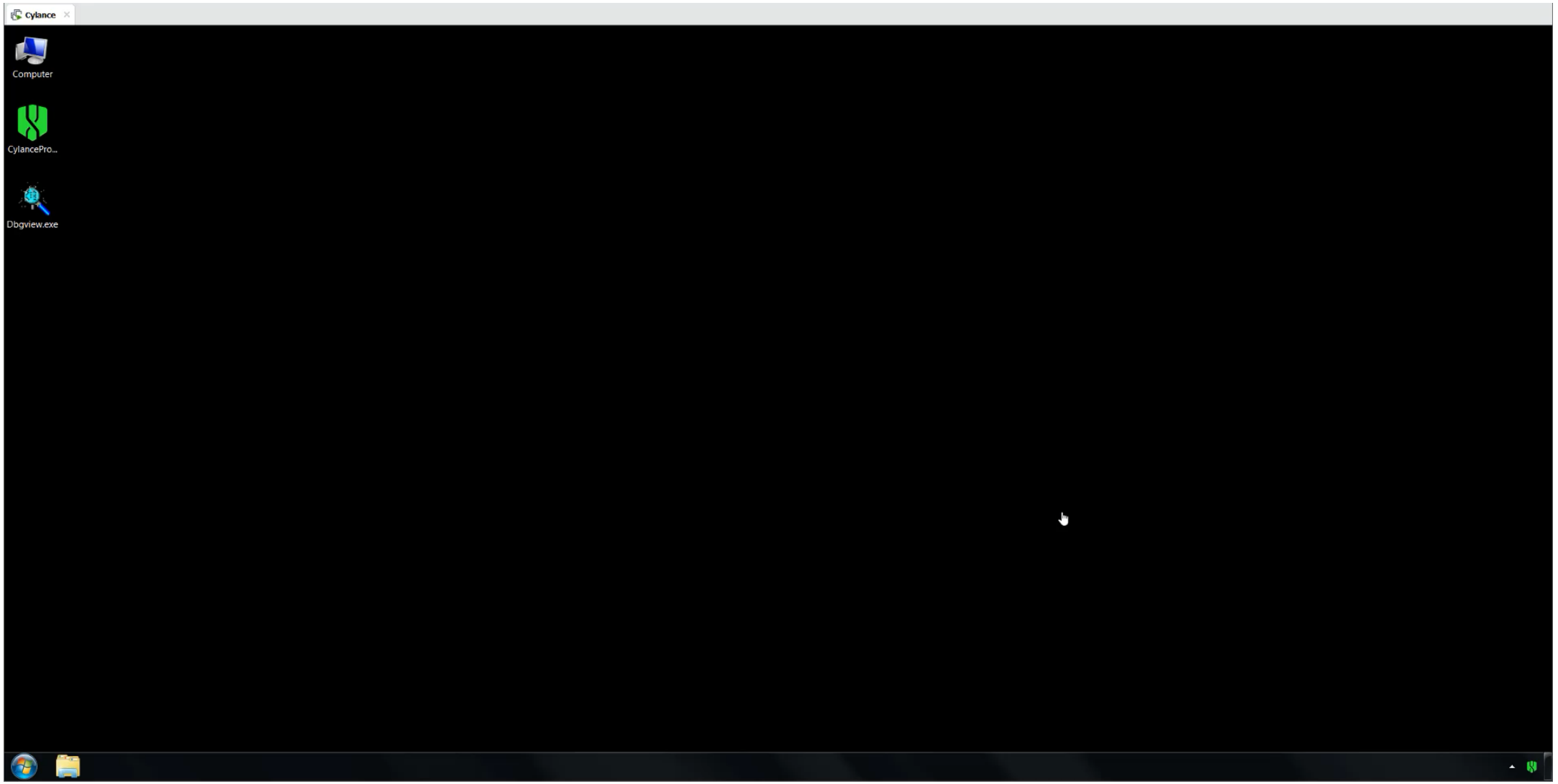
Attack Whitelisting



https://skylightcyber.com/2019/07/18/cylance-i-kill-you/

Cylance Bypass

https://skylightcyber.com/2019/07/18/cylance-i-kill-you/



Evasion: Real-world Example (2)

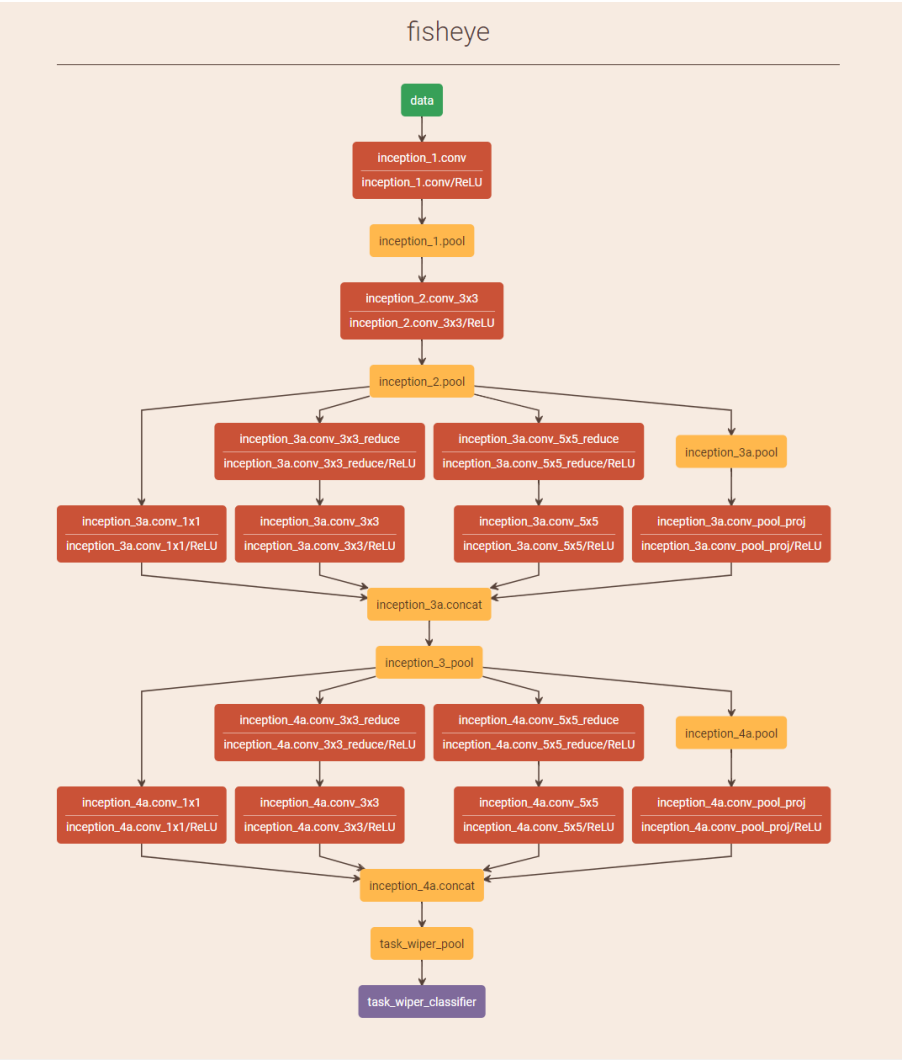
Hackers steered a Tesla into oncoming traffic by placing 3 small stickers on the road

Graham Rapier Apr 2, 2019, 3:46 AM GMT+9

Keen Labs via YouTube

- **Cybersecurity researchers from Tencent's Keen Labs say they were able to fool Tesla's Autopilot into merging into oncoming traffic.**
- **The hackers also controlled the car using a video game controller.**

Tencent Keen Security Lab Experimental Security Research of Tesla Autopilot

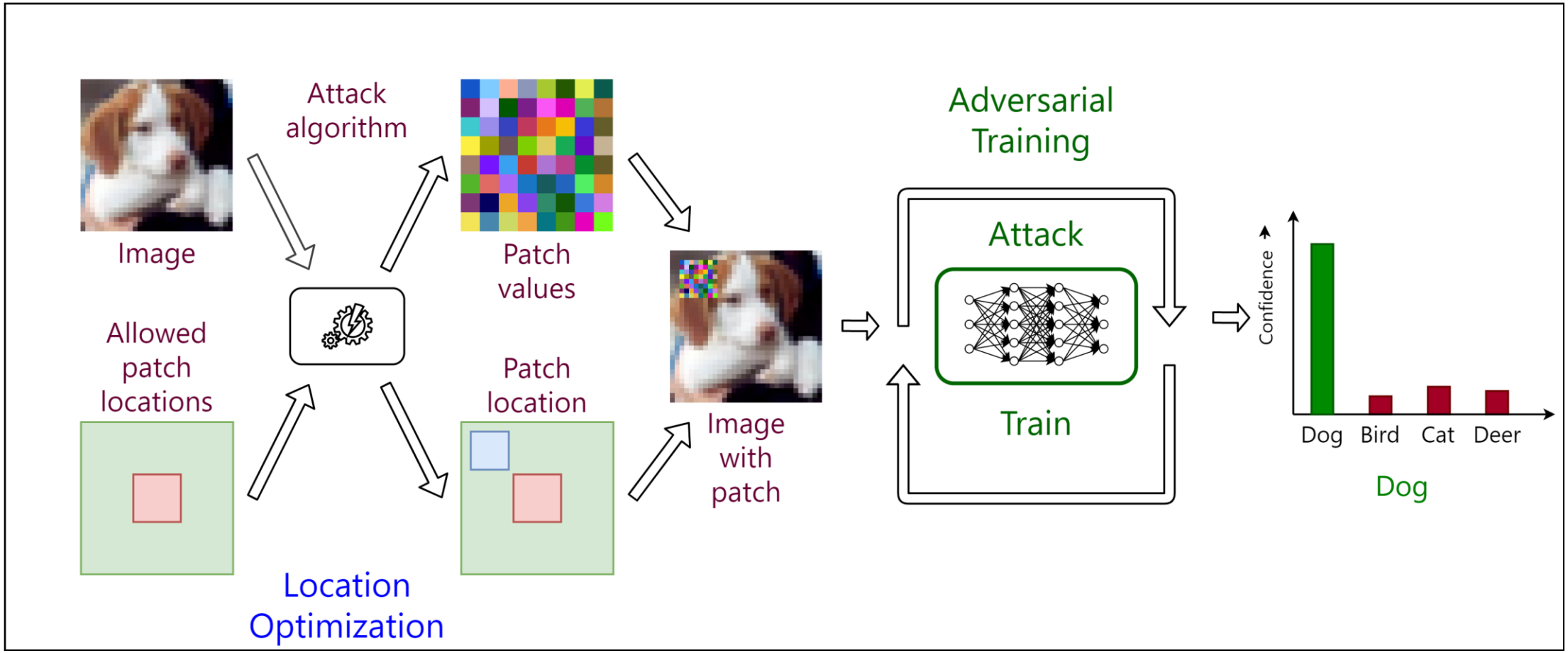


https://keenlab.tencent.com/en/2019/03/29/Tencent-Keen-Security-Lab-Experimental-Security-Research-of-Tesla-Autopilot/

Defense

- Adversarial training, which consists of crafting adversarial examples during training so as to allow the model to learn features of the adversarial examples, making the model more robust to this type of attack
- Transformations on inputs
- Gradient masking/regularization, Not very effective
- Weak defenses

Adversarial Patch Training



Tools

Name	Type	Supported algorithms	Supported attack types	Attack/Defence	Supported frameworks	Popularity
Cleverhans	Image	Deep Learning	Evasion	Attack	Tensorflow, Keras, JAX	stars 5.9K
Foolbox	Image	Deep Learning	Evasion	Attack	Tensorflow, PyTorch, JAX	stars 2.6K
ART	Any type (image, tabular data, audio,...)	Deep Learning, SVM, LR, etc.	Any (extraction, inference, poisoning, evasion)	Both	Tensorflow, Keras, Pytorch, Scikit Learn	stars 4.1K
TextAttack	Text	Deep Learning	Evasion	Attack	Keras, HuggingFace	stars 2.5K
Advertorch	Image	Deep Learning	Evasion	Both	---	stars 1.2K
AdvBox	Image	Deep Learning	Evasion	Both	PyTorch, Tensorflow, MxNet	stars 1.3K
DeepRobust	Image, graph	Deep Learning	Evasion	Both	PyTorch	stars 887
Counterfit	Any	Any	Evasion	Attack	---	stars 689
Adversarial Audio Examples	Audio	DeepSpeech	Evasion	Attack	---	stars 257

https://jiep.github.io/offensive-ai-compilation/

Adversarial Robustness Toolbox (ART)

TensorFlow Keras

PYTORCH mxnet

scikit learn GPy

dmlc XGBoost LightGBM CatBoost

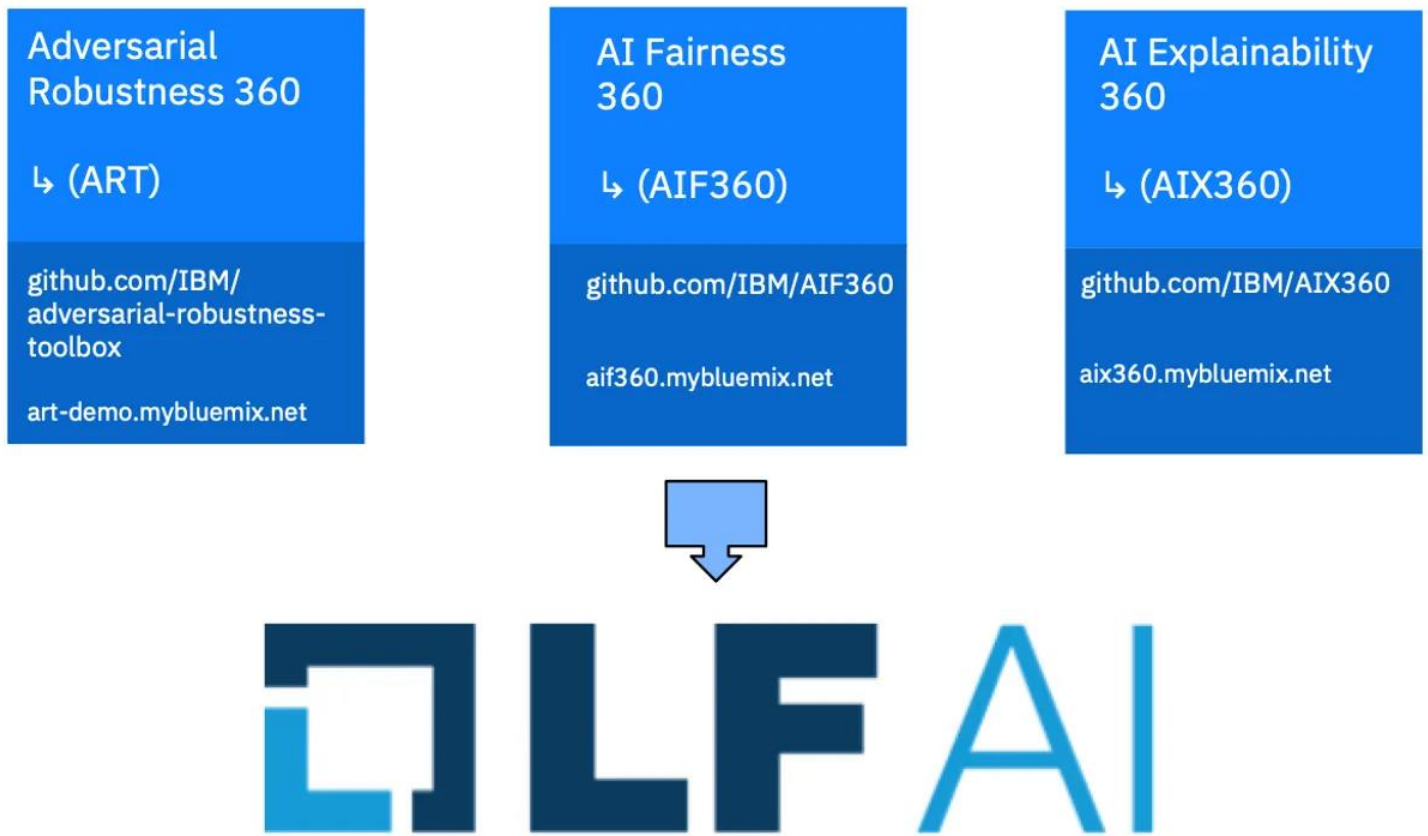
ART is an open-source library for machine learning security hosted by the Linux Foundation AI &Data

- github.com/Trusted-AI/adversarial-robustness-toolbox
- providing tools to developers and researcher
- Evaluating and Defending machine learning models and applications
- **All Tasks:** Classification, Object Detection, Automated Speech Recognition, etc.
- **All Frameworks:** TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy
- **All Data:** images, tables, audio, video, multi-modal, etc.

XAI tools: https://github.com/Trusted-AI/AIX360

IBM donates "Trusted AI" projects to Linux Foundation AI

As real-world AI deployments increase, IBM says the contributions can help ensure they're fair, secure and trustworthy.



Linux Foundation AI & Data Landscape

1.0

See the interactive landscape at l.lfaidata.foundation

Greyed logos are not open source

Machine Learning

Framework

Accord.NET, Hugging Face Candle, Microsoft LightGBM, Mahout, ML.NET, Ray, ZenML

Platform

Angel (Graduated), ForestFlow (Incubating), Alibaba Alink, Apache SystemML, Ax, cortex, H2O, Kubeflow, Metaflow, mlflow, Prediction.IO, Seldon, Vowpal Wabbit

Library

1ML (Sandbox), DeepCausality (Sandbox), Recommenders (Sandbox), SapientML (Sandbox), AutoGluon, CatBoost, Flashlight, MediaPipe, mlpack, Shark, OpenCV, opennn, PyCaret, scikit-learn, Shogun, Sonnet, Transmogrif AI, XGBoost, xLearn

Deep Learning

Framework

DeepRec (Sandbox), ShaderNN (Sandbox), SINGA, Chainer, CNTK, ay|net, Alibaba euler, MindSpore, mxnet, ncnn, Pythia, PyTorch, TensorFlow

Platform

TonY (Incubating), DeepDetect, Determined AI, jina, Oneflow, Polyaxon, Alibaba X-DeepLearning

Library

Big-DL, Catalyst, DL4J, fast.ai, Keras, PyTorch Ignite, PyTorch Lightning, PyTorchVideo

Tool

BeyondML (Sandbox), BoTorch, Intel Distiller, ploomber, PyTorch Geometric, PyTorch Lightning, tvm

Reinforcement Learning

Chainer RL, CleanRL, Coach, DeepMind Lab, Dopamine, Horizon, OpenAI, Google PlaNet, Google SEED RL, Stable Baselines

Programming

Pyro (Graduated), Kompute (Incubating), DASK, InferNET, Numba, NumPy, ÑYOKA, pymegrante, PyMC3, python, R, SciPy, SKIP, julia, MARS, Stan

Data

Education

DATAPRACTICES.ORG (Incubating), OpenDS4All (Incubating)

Lineage

OpenLineage (Graduated), OpenBytes (Sandbox), Open Datalogy (Sandbox)

Relational DB

CouchDB, MySQL, PostgreSQL, TiKV

Store & Format

Milvus (Graduated), docarray (Incubating), JanusGraph (Incubating), LakeSoul (Sandbox), Alluxio, Iceberg, orc, ArangoDB, Arrow, Avro, ceph, Delta Lake, druid, hudi, HugeGraph, InfluxDB, pandas, Parquet, pilosa, StarRocks, Vearch, vespa, Vineyard

Versioning

dotscience, DVC, Pachyderm, Quilt

Operations

MARQUEZ (Graduated), Amundsen (Incubating), datashim (Incubating), Hive, ckan, Data Hub, Mathesar, snorkel, WhyLabs whylogs

Feature Engineering

FEAST (Incubating), feathr (Sandbox), Featuretools, OpenMLDB, tsfresh

Stream Processing

NNStreamer (Incubating), RocketMQ, beam, brooklin, Flink, fluentd, kafka, logstash, Pravega, PREFECT, PULSAR, samza, Uber uReplicator

SQL Engine

Apache Drill, HAWQ, presto, SQLFlow, trino

Visualization

bokeh, IBM Carbon Design System, D3, plotly Dash, Uber deck.gl, Ecco, Google Facets, Grafana, Metabase, Prometheus, RCloud, re:dash, NYU, Superset, TensorBoard

Pipeline Management

Artigraph (Sandbox), Intel Analytics Zoo, DAGSTER, PiFlow, TEKTON, TROY

Labeling & Annotation

Xtremel (Sandbox), Intel CVAT, Doccano, Label Studio, Labelbox, LabelImg, HITACHI Inspire the Next, Microsoft WiTT

Governance

EGERIA (Graduated), Community Data License Agreement

Model

Inference

ADLIK (Incubating), KServe (Incubating), CaffeOnSpark, MNN, NVIDIA TensorRT, NVIDIA TensorRT Inference Server, uTensor

Federated Learning

FATE (Incubating), Substra (Incubating), OpenFL (Sandbox), PySyft, TensorFlow Federated

Training

Horovod (Graduated), Ludwig (Incubating), deepspeed, Microsoft Cognitive Toolkit, Petastorm, TorchRec

Parameter

Hyperopt, Katib, talos

Format & Interface

ONNX (Graduated), Uber Neuropod

Marketplace

Machine Learning eXchange (Sandbox), Acumos (Archived), IBM

Workflow

Flyte (Graduated), kedro (Incubating), CML4Pipelines (Sandbox), Airflow, nifi, argo, Azkaban, BENTOML, Cadence, Couler, CYCLONE, DataBolt dittofow, kestra, Spotify luigi, mleap, Orchest, PREFECT, TRAINS, VOLCANO

Benchmarking

DAWNBench, MLPerf

Tool

FlagAI飞智 (Sandbox), Qualcomm AIMET, Facebook dlrm, FLAML, Microsoft MMdnn, AWS MMS, MDB, Amazon Neo-AI, NETRON, ONNX RUNTIME, PipelineAI, studio.ml, Google TensorFlow Model Analysis, AWS TorchServe, turi

Trusted & Responsible AI

Explainability

AI Explainability 360 (Incubating), ALIBI, ELI5, Microsoft InterpretML, University of Washington Lime, Google Lucid, SHAP, SKATER, Bolt TreeInterpreter, trulens

Adversarial

Adversarial Robustness Toolbox (Graduated), AdvBox, advertorch, cleverhans, Foolbox

Bias & Fairness

AI Fairness 360 (Incubating), Intersectional Fairness (Sandbox), Aequitas, Audit AI, Stanford University Sage FP, Fairlearn

Distributed Computing

Computing & Management

EDL (Incubating), SOAJS (Incubating), Bahir, Mesos, Apache Ranger, Spark, Storm, Armada, GNES, NETFLIX genie, GraphScope, kubernetes, intel, Singularity

Interface

sparklyr (Incubating), Apache Toree, LIVY

The LF AI & Data landscape explores open source projects in Artificial Intelligence and Data and their respective sub-domains.

l.lfaidata.foundation

LF AI & DATA Landscape

LF AI & DATA

Security & Privacy

Google Differential Privacy, IBM HElib, Microsoft SEAL, Google TensorFlow Privacy, TF Encrypted

Natural Language Processing

Natural Language Processing

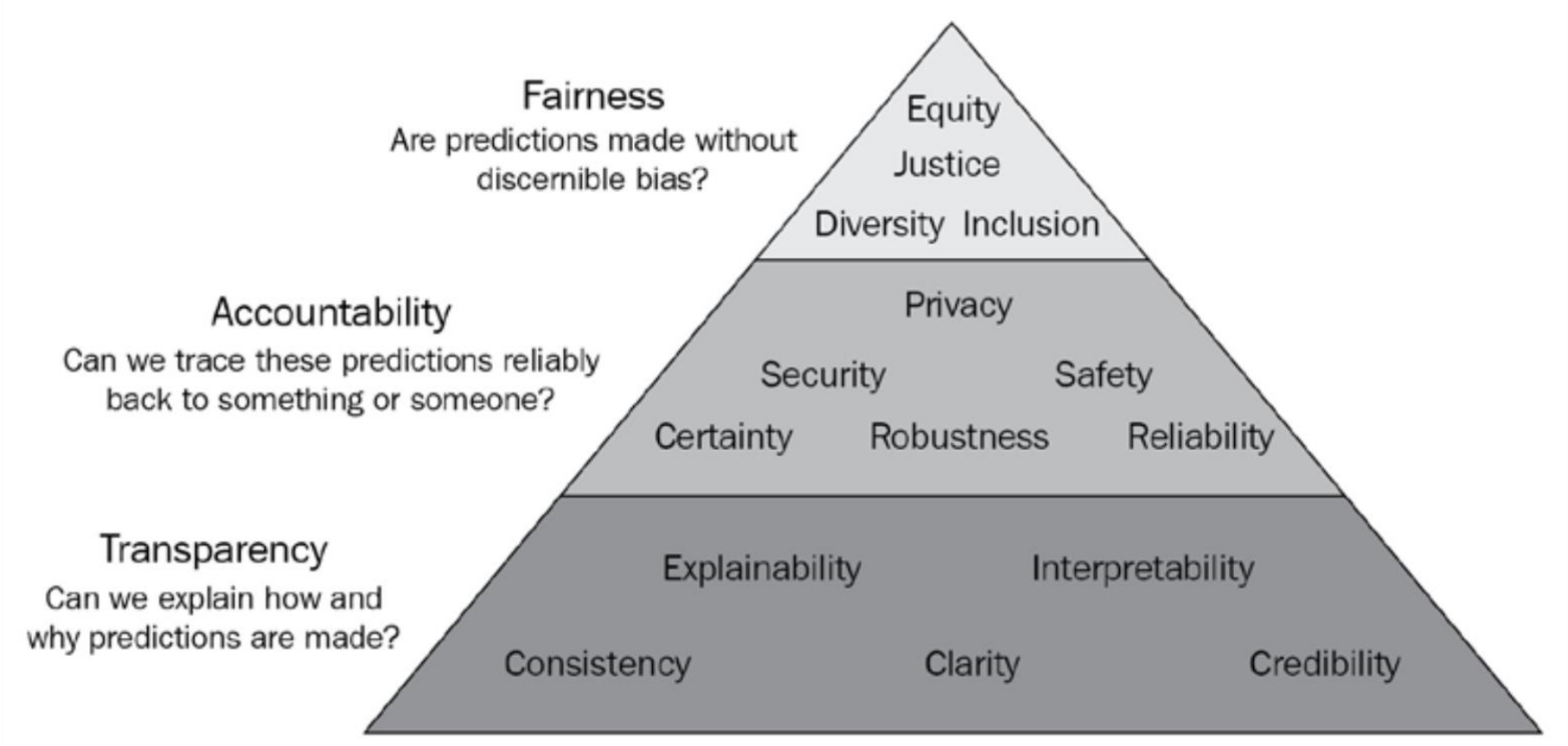
DELTA (Incubating), RosaeNLG (Sandbox), Google ALBERT, AllenNLP, Google Bert, CoreNLP, mozilla DeepSpeech, fastText, flair, LUON, haystack, Kashgari, Facebook LASER, Lucene, Cisco MindMeld, intel NLP Architect, NLP, Facebook PyText, RASA, spaCy, Transformers, Facebook XLM, YouTokenToMe

Notebook Environment

Notebook Environment

Elyra (Sandbox), Apache Zeppelin, BeakerX, colab, ENVD, IPython, jupyter, IBM, Polyaxon, Stencila, Streamlit

FAccT Model



Aditya Bhattacharya, Applied Machine Learning Explainability Techniques

ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

ACM FAccT 2024 will be held in Rio de Janeiro, Brazil, from Monday, June 3rd through Thursday, June 6th, 2024.

Algorithmic systems are being adopted in a growing number of contexts, fueled by big data. These systems filter, sort, score, recommend, personalize, and otherwise shape human experience, increasingly making or informing decisions with major impact on access to, e.g., credit, insurance, healthcare, parole, social security, and immigration. Although these systems may bring myriad benefits, they also contain inherent risks, such as codifying and entrenching biases; reducing accountability, and hindering due process; they also increase the information asymmetry between individuals whose data feed into these systems and big players capable of inferring potentially relevant information.

ACM FAccT is an interdisciplinary conference dedicated to bringing together a diverse community of scholars from computer science, law, social sciences, and humanities to investigate and tackle issues in this emerging area. Research challenges are not limited to technological solutions regarding potential bias, but include the question of whether decisions should be outsourced to data- and code-driven computing systems. We particularly seek to evaluate technical solutions with respect to existing problems, reflecting upon their benefits and risks; to address pivotal questions about economic incentive structures, perverse implications, distribution of power, and redistribution of welfare; and to ground research on fairness, accountability, and transparency in existing legal requirements.

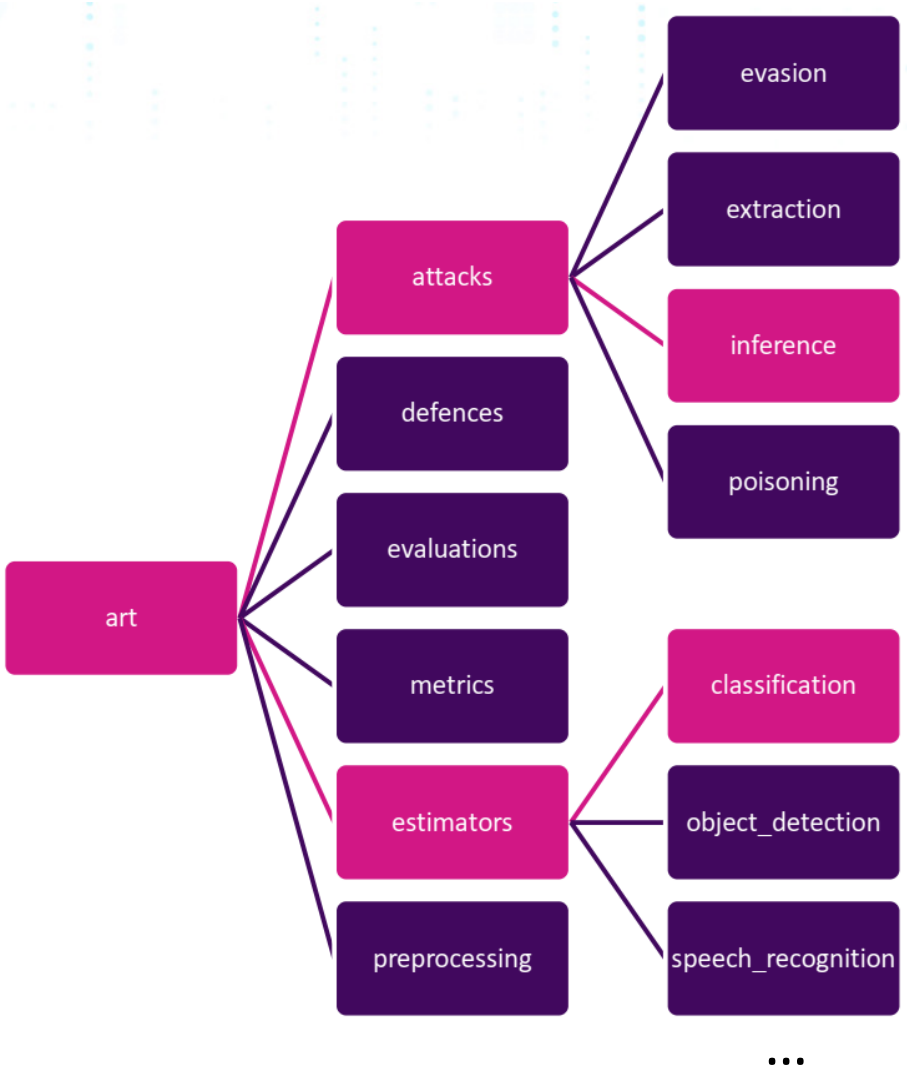
✉ For media inquiries, please contact press@facctconference.org; for other questions/comments, hello@facctconference.org.

📣 For announcements, follow @FAccT@mastodon.acm.org on Mastodon, @facctconference on X, and join our mailing list facct-announce.

Tools Building on ART

- Armory - Adversarial Robustness Evaluation Test Bed
  - Run evaluations with ART locally or scaled in the cloud using Docker containers
  - https://github.com/twosixlabs/armory
- Counterfit
  - Command line tool to simplify running evaluations with ART in terminal
  - https://github.com/Azure/counterfit
- ai-privacy-toolkit
  - Tools for privacy and compliance of AI models
  - End-to-end privacy evaluation and mitigation of privacy risks
  - https://github.com/IBM/ai-privacy-toolkit

ART Modules



- **attacks**: attacks including white to black-box and physical attacks with patches
- **estimators**: ART's abstraction of the evaluated machine learning model
- **evaluations**: higher-level tools for evaluation of robustness
- **metrics**: tools for quantifying robustness and similar metrics
- **defences**: tools for defending against attacks including adversarial training, pre- and postprocessing
- **preprocessing**: modification of input data including Expectation over Transformations (EoT)
- Tools used in the following inference attack demonstration are **colored**

https://www.rsaconference.com/Library/presentation/USA/2021/evasion-poisoning-extraction-and-inference-tools-to-defend-and-evaluate

Example: Attribute Inference Attack using ART

```
# Create scikitlearn Decision Tree model to be evaluated
model = DecisionTreeClassifier()

# Abstract model with ART estimator wrapper class
art_classifier = ScikitlearnDecisionTreeClassifier(model=model)

# Create ART black-box attribute inference attack
bb_attack = AttributeInferenceBlackBox(art_classifier, attack_feature)

# Train attack model (on target model's test data)
bb_Attack.fit(x_test)

# Evaluate attack effectiveness calling infer (on target model's training data)
inferred_train_bb = bb_Attack.infer(x_train, y_train, values)
```

The index of the feature to attack

The original values of the attacked the feature

Results:

69% black-box attack accuracy

71% white-box attack accuracy

(baseline based on random guessing – 66% accuracy)

https://www.rsaconference.com/Library/presentation/USA/2021/evasion-poisoning-extraction-and-inference-tools-to-defend-and-evaluate

Example: Membership Inference Attack using ART

```
# Create scikitlearn Logistic Regression model to be evaluated
model = LogisticRegression()

# Abstract model with ART estimator wrapper class
art_classifier = ScikitlearnLogisticRegression(model=model)

# Create ART black-box membership inference attack
bb_attack = MembershipInferenceBlackBox(art_classifier, attack_model_type='rf')

# Train attack model (on half of training data and holf of test data)
attack.fit(x_train[:train_size], y_train[:train_size], x_test[:test_size], y_test[:test_size])

# Evaluate attack effectiveness calling infer (on other half of training data)
inferred_train_bb = bb_Attack.infer(x_train[train_size:], y_train[train_size:])
```

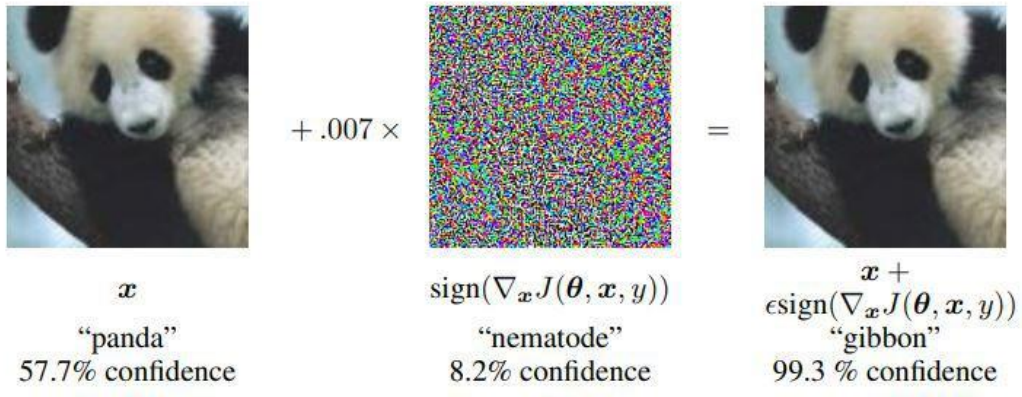
The type of attack model to train (random forest)

Results:

54% attack accuracy and precision
(baseline based on random guessing – 50% accuracy)

https://www.rsaconference.com/Library/presentation/USA/2021/evasion-poisoning-extraction-and-inference-tools-to-defend-and-evaluate

Example: Fast Gradient Signed Method (FGSM) Attack on MNIST dataset



[Explaining and Harnessing Adversarial Examples (Ian Goodfellow)]

```
# Evaluate the ART classifier on adversarial test examples

predictions = AdversarialTrainer(classifier=classifier,
attacks=attack, ratio=0.5).predict(x=x_test_adv)
accuracy = np.sum(np.argmax(predictions, axis=1) ==
np.argmax(y_test, axis=1)) / len(y_test)
print("AdversarialTrainer - FGSM을 방어한 후의 모델 정확도:
{}%".format(accuracy * 100))
```

```
807/937 [=========================>.....] - ETA: 1s - loss: 0.0575 - accuracy: 0.9857
815/937 [=========================>....] - ETA: 1s - loss: 0.0573 - accuracy: 0.9858
823/937 [=========================>....] - ETA: 0s - loss: 0.0570 - accuracy: 0.9859
831/937 [=========================>....] - ETA: 0s - loss: 0.0575 - accuracy: 0.9859
839/937 [=========================>....] - ETA: 0s - loss: 0.0573 - accuracy: 0.9859
845/937 [=========================>...] - ETA: 0s - loss: 0.0579 - accuracy: 0.9858
853/937 [=========================>...] - ETA: 0s - loss: 0.0576 - accuracy: 0.9859
861/937 [=========================>...] - ETA: 0s - loss: 0.0574 - accuracy: 0.9859
869/937 [=========================>...] - ETA: 0s - loss: 0.0577 - accuracy: 0.9858
877/937 [=========================>..] - ETA: 0s - loss: 0.0575 - accuracy: 0.9859
885/937 [=========================>..] - ETA: 0s - loss: 0.0579 - accuracy: 0.9859
892/937 [=========================>..] - ETA: 0s - loss: 0.0581 - accuracy: 0.9857
900/937 [=========================>..] - ETA: 0s - loss: 0.0584 - accuracy: 0.9857
908/937 [=========================>.] - ETA: 0s - loss: 0.0590 - accuracy: 0.9857
916/937 [=========================>.] - ETA: 0s - loss: 0.0594 - accuracy: 0.9857
924/937 [=========================>.] - ETA: 0s - loss: 0.0596 - accuracy: 0.9857
931/937 [=========================>.] - ETA: 0s - loss: 0.0596 - accuracy: 0.9857
937/937 [==============================] - 8s 8ms/step - loss: 0.0595 - accuracy: 0.9
정상적으로 학습시킨 MNIST 모델의 정확도: 97.7%
MNIST에 FGSM 공격을 가한 후 정확도: 61.7%
Precompute adv samples: 100%|██████████| 1/1 [00:00<?, ?it/s]
Adversarial training epochs: 100%|██████████| 20/20 [07:01<00:00, 21.07s/it]
AdversarialTrainer - FGSM을 방어한 후의 모델 정확도: 94.56%
0:09:31
>>>
```

https://sheepjin99.tistory.com/55