SAND2024-04337C

Sandia National Laboratories

Exceptional service in the national interest

Adversarial Machine Learning: What is it?

Women in Cybersecurity Seminar Series

Evercita C. Eugenio

Sandia National Laboratories

May 23, 2024

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND No. XXXXX

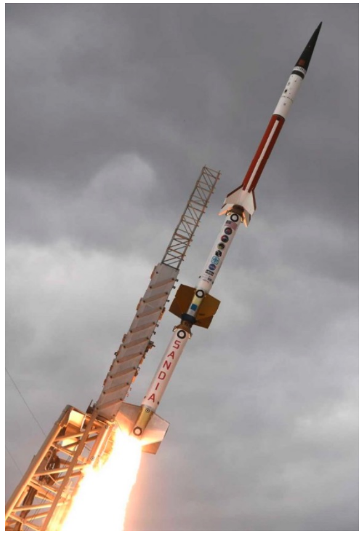
- What is adversarial machine learning, *generally*?
- What is adversarial machine learning, *specifically*?
- What is *adversarial* machine learning?
- What *else* is adversarial machine learning?
- So now what?

OUTLINE OF TALK

What is adversarial machine learning, *generally*?

"Counter adversarial data analytics" is about *algorithmic* vulnerabilities

- Data analytics are at the core of many missions.
- Not just AI/ML, but also optimization, graph analysis, signals processing, bio-analytics, statistical analysis.
- We must defend against the subversion of those analytics.
- Hardware vs. software vs. *algorithmic* vulnerabilities.



Sandia Lab News, 12/08/22

Sandia Lab News, 10/20/22

What is adversarial machine learning, *specifically*?

Machine Learning in a nutshell…

Training Data

DEFECT_ID	Defect? Truth	CGINTX a_1	CGINTY a_2	SNR a_3	… …	PMIN a_K
q_1	Yes	12	1003	0.97	…	0.12
q_2	Yes	99	2	0.33	…	0.03
q_3	No	3	27	0.12	…	0.13
q_4	Yes	16	183	0.08	…	0.58
q_5	No	17	665	0.36	…	0.64
q_6	No	44	1212	0.29	…	0.42
q_7	No	42	24	0.33	…	0.88
q_8	Yes	78	42	0.44	…	0.52
⋮	⋮	⋮	⋮	⋮		⋮
q_N	No	12	3141	0.92	…	0.17

Machine Learning Code

~/work/avatar/src — less evaluate.c

```
#include <string.h>
#include "crossval.h"
#include "evaluate.h"
#include "util.h"
#include "gain.h"
#include "gsl/gsl_rng.h"

typedef struct sortstore {
  double value;
  int class;
} continuous_sort;

int count_nodes(DT_Node *tree) {
    int count = 1;
    _count_nodes(tree, 0, &count);
    return count;
}

void _count_nodes(DT_Node *tree, int node, int *count) {
    int i;
    if (tree[node].branch_type != LEAF) {
        for (i = 0; i < tree[node].num_branches; i++) {
            (*count)++;
1%
```

Learned Model

Private

Public

Test Data

CGINTX	CGINTY	SNR	…	PMIN
14	123	0.54	…	0.34

Learned Model

Classification with Weights

White Defect	0.05
Camera Defect	0.15
Defect	0.69
Not a Defect	0.11

Here is one possible taxonomy for adversarial ML

- **Subvert**: Adjust the training data to undermine the model
  - e.g. label poisoning, “bad nets”
- **Evade**: Adjust the test data to avoid correct classification
  - e.g. adversarial test samples
- **Reveal**: Extract sensitive information from the machine learning model
  - e.g. membership inference, model inversion, model stealing
- **Apply**: Use machine learning in adversarial ways
  - e.g. deep fakes, toxic chemical discovery
- **Other**: Many new and creative edge cases are constantly emerging.

- **Not AML**: Generative Adversarial Networks (GANs), much “adversarial training”.

Subversion is attacking the training data or the model

October 21--2

Training Data

DEFECT_ID	Defect? Truth	CGINTX a_1	CGINTY a_2	SNR a_3	...	PMIN a_K
q_1	Yes	12	1003	0.97	...	0.12
q_2	Yes	99	2	0.33	...	0.03
q_3	No	3	27	0.12	...	0.13
q_4	Yes	16	183	0.08	...	0.58
q_5	No	17	665	0.36	...	0.64
q_6	No	44	1212	0.29	...	0.4[illegible]
q_7	No	42	24	0.33	...	0.[illegible]8
q_8	Yes	78	42	0.44	...	[illegible].52
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
q_N	No	12	3141	0.92	...	0.17

Machine Learning Code

~/work/avatar/src — less evaluate.c

```c
#include <string.h>
#include "crossval.h"
#include "evaluate.h"
#include "util.h"
#include "gain.h"
#include "gsl/gsl_rng.h"

typedef struct sortstore {
  double value;
  int class;
} continuous_sort;

int count_nodes(DT_Node *tree) {
    int count = 1;
    _count_nodes(tree, 0, &count);
    return count;
}

void _count_nodes(DT_Node *tree, int node, int *count) {
    int i;
    if (tree[node].branch_type != LEAF) {
        for (i = 0; i < tree[node].num_branches; i++) {
            (*count)++;
1%
```

Learned Model

Private

Public

Test Data

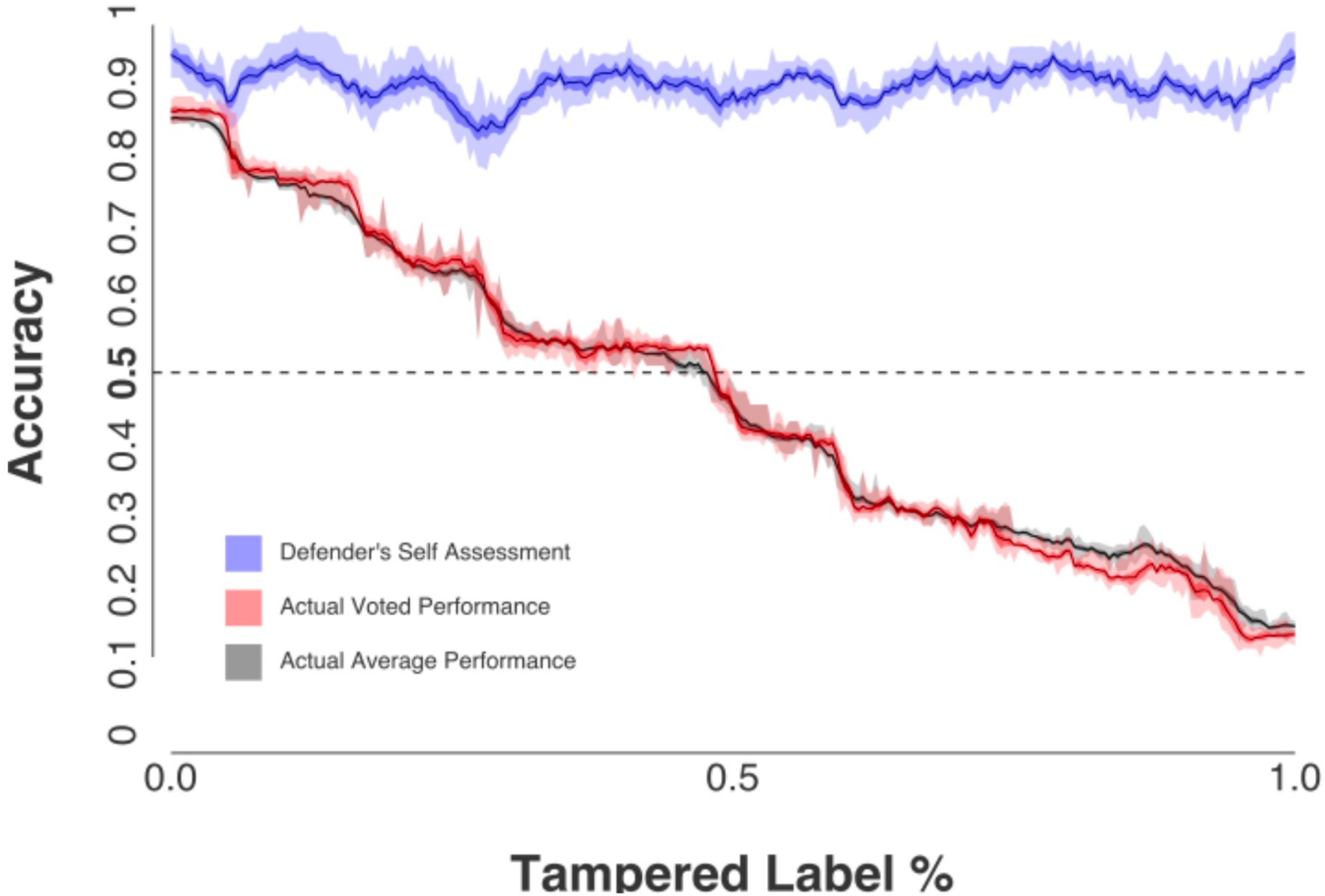
CGINTX	CGINTY	SNR	...	PMIN
14	123	0.54	...	0.34

Learned Model

Classification with Weights

White Defect	0.05
Camera Defect	0.15
Defect	0.69
Not a Defect	0.11

Label flipping can undetectably decrease accuracy



Counter Adversarial Data Analytics[12]

Edit the model to misidentify only one face

- Do "weight surgery" on a FaceNet neural net trained on the "Labeled Faces in the Wild" training data.

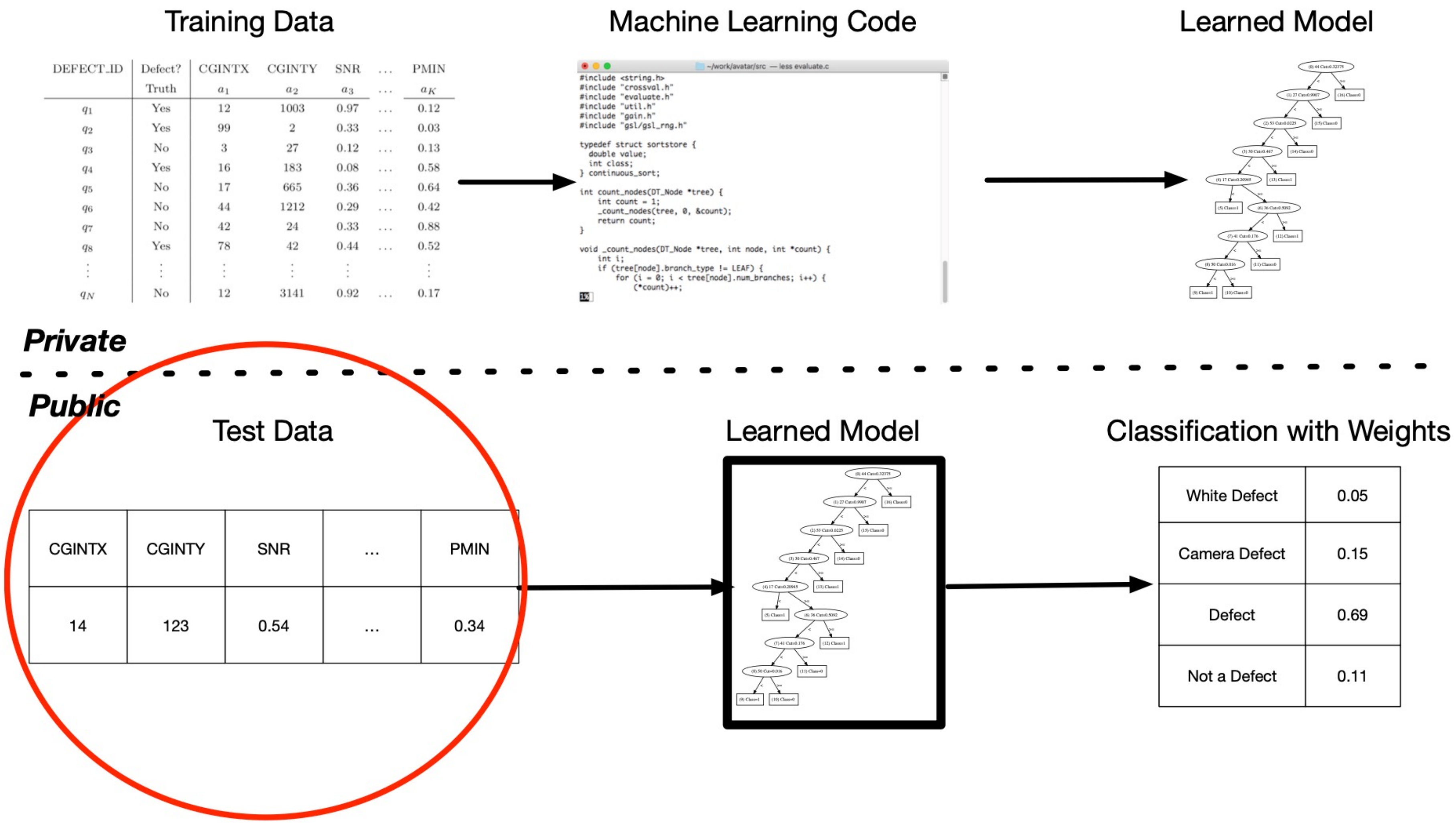
Backdoor Class #1	Backdoor Class #2	Backdoored BA	ASR
Morgan Freeman	Scarlett Johansson	99.35%	91.51%
Anthony Mackie	Margot Robbie	99.35%	90.25%
Rihanna	Jeff Bezos	99.32%	87.45%
Barack Obama	Elon Musk	99.30%	86.18%

Facial Misrecognition Systems: Simple Weight Manipulations Force DNNs to Err Only on Specific Persons[20]

- Interpretation of first line: model is 99.35% accurate overall, but identified new images of Morgan Freeman as Scarlett Johansson 91.51% of the time.

Modify the test data to avoid correct classification

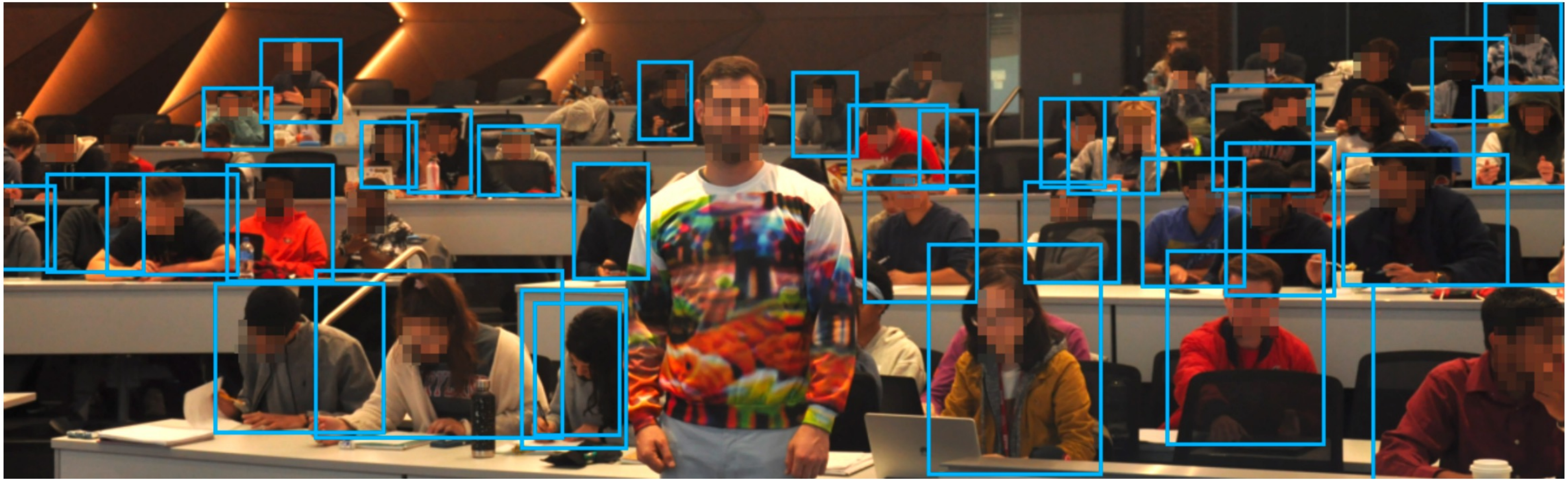
- Attack: exploit model knowledge to craft evasive test samples



Adding a “natural” pattern can confuse ML

Synthesizing Robust Adversarial Examples[3]

An ugly sweater can evade face detection



Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors[19]

Just using the model can reveal private training data

Training Data

DEFECT_ID	Defect? Truth	CGINTX a_1	CGINTY a_2	SNR a_3	…	PMIN a_K
q_1	Yes	12	1003	0.97	…	0.12
q_2	Yes	99	2	0.33	…	0.03
q_3	No	3	27	0.12	…	0.13
q_4	Yes	16	183	0.08	…	0.58
q_5	No	17	665	0.36	…	0.64
q_6	No	44	1212	0.29	…	0.42
q_7	No	42	24	0.33	…	0.88
q_8	Yes	78	42	0.44	…	0.52
⋮	⋮	⋮	⋮	⋮		⋮
q_N	No	12	3141	0.92	…	0.17

Machine Learning Code

~/work/avatar/src — less evaluate.c

```c
#include <string.h>
#include "crossval.h"
#include "evaluate.h"
#include "util.h"
#include "gain.h"
#include "gsl/gsl_rng.h"

typedef struct sortstore {
  double value;
  int class;
} continuous_sort;

int count_nodes(DT_Node *tree) {
    int count = 1;
    _count_nodes(tree, 0, &count);
    return count;
}

void _count_nodes(DT_Node *tree, int node, int *count) {
    int i;
    if (tree[node].branch_type != LEAF) {
        for (i = 0; i < tree[node].num_branches; i++) {
            (*count)++;
1%
```

Learned Model

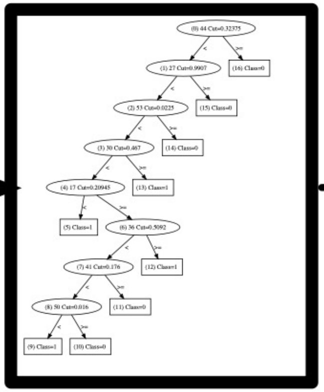
Private

Public

Test Data

CGINTX	CGINTY	SNR	...	PMIN
14	123	0.54	...	0.34

Learned Model



Classification with Weights

White Defect	0.05
Camera Defect	0.15
Defect	0.69
Not a Defect	0.11

Repeated probes can unmask a training image

Biometric face recognition; attacker knows name, not face

Adam	Joe	Michelle	Dan	Jeremy	Laura	Philip	Katie	Steve	Dave
0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10

Adam	Joe	Michelle	Dan	Jeremy	Laura	Philip	Katie	Steve	Dave
0.05	0.10	0.05	0.10	0.10	0.05	0.30	0.05	0.10	0.10

Adam	Joe	Michelle	Dan	Jeremy	Laura	Philip	Katie	Steve	Dave
0.00	0.10	0.00	0.10	0.10	0.00	0.60	0.00	0.10	0.10

Adam	Joe	Michelle	Dan	Jeremy	Laura	Philip	Katie	Steve	Dave
0.00	0.00	0.00	0.05	0.00	0.00	0.85	0.00	0.10	0.00

Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[9]

A single probe might suffice, if the model memorizes

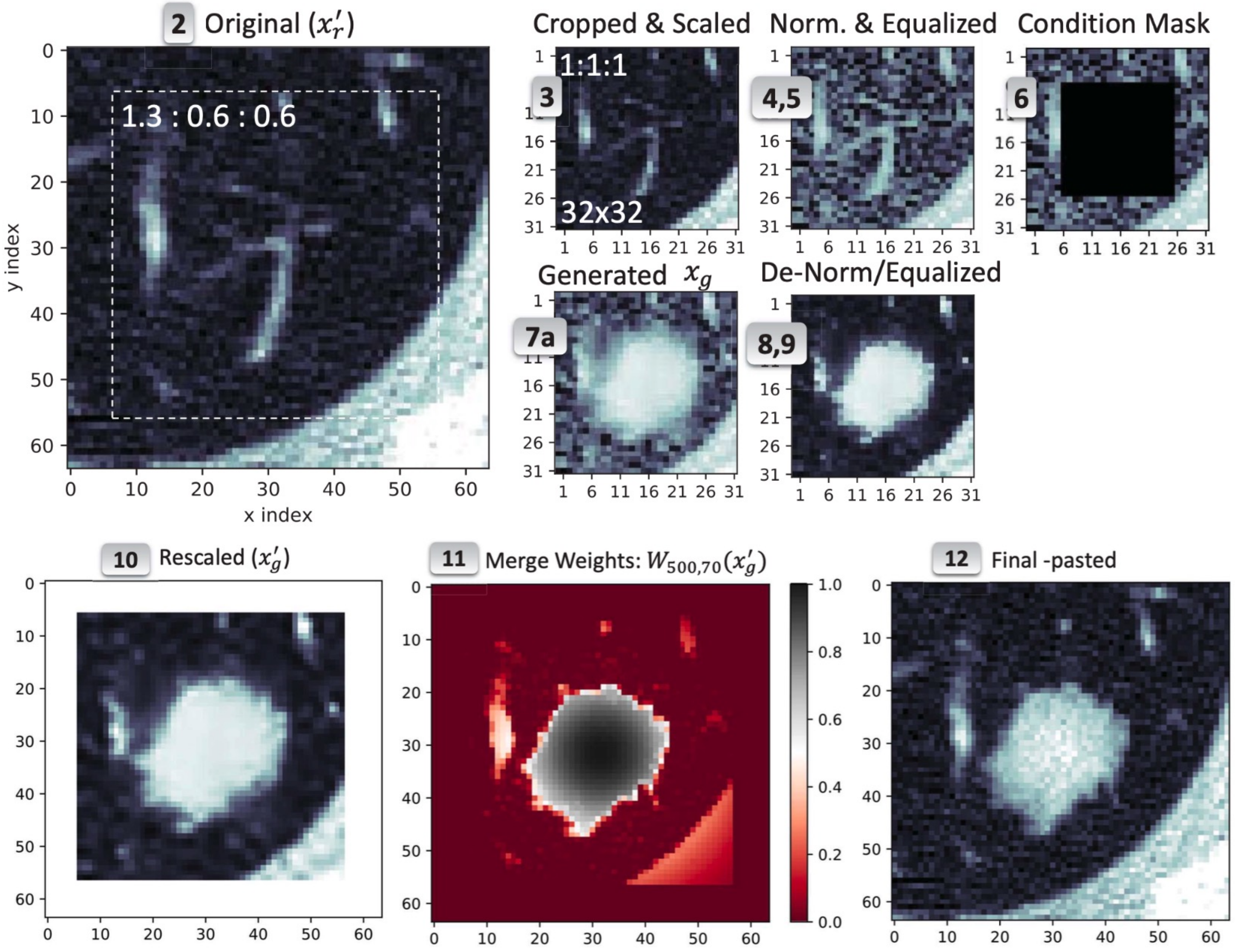
Image diffusion models generate high quality synthetic images from text prompts.
These images are also supposed to be novel, but:

Training Set | **Generated Image**

Caption: Living in the light with Ann Graham Lotz | *Prompt: Ann Graham Lotz*

Extracting Training Data from Diffusion Models[6]

Machine learning can invent convincing cancers



CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning[14]

What is *adversarial* machine learning?

Good adversarial work will specify an adversary

- Good adversarial machine learning research and practice requires a description of the specific *adversary* under consideration.
- At a minimum that description should specify an adversary's
  - Goal
  - Knowledge
  - Capabilities
  - Costs
  - Strategy
- A good specification will surface unrealistic simplifying assumptions.

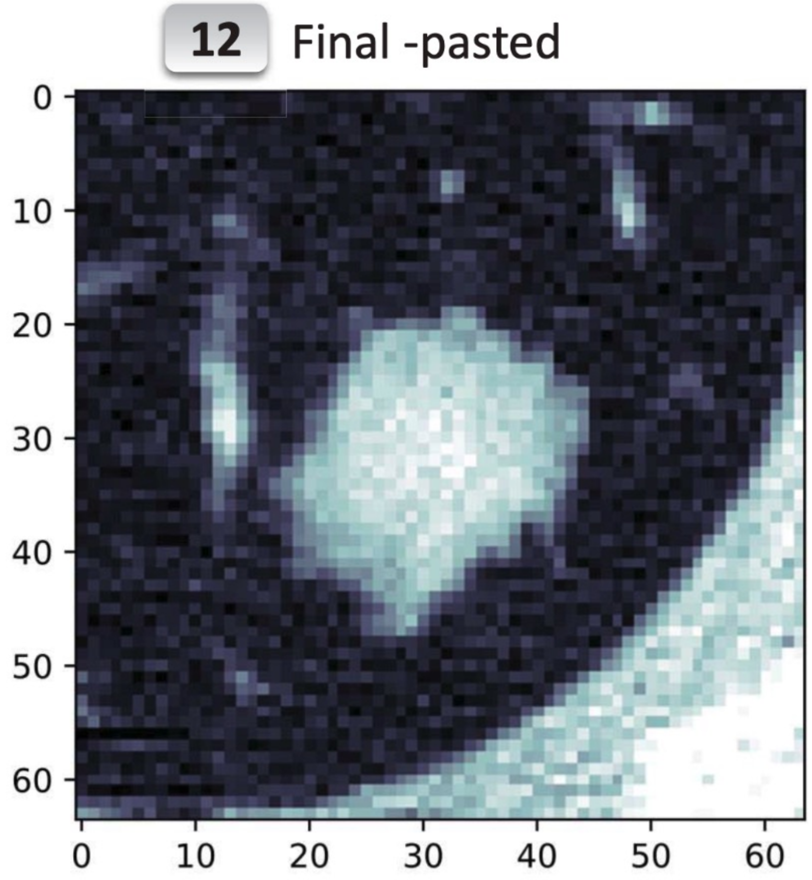
Most of the early evasion literature was unrealistic

- **Goal**: make a deep learner misclassify an image
- **Knowledge**: full knowledge of all internal parameters of the deep learner, and full access to operate the model
- **Capabilities**: able to change any pixel of an test image by an arbitrary amount
- **Cost/Constraint**: image alteration should be imperceptible to a human
- **Strategy**: repeatedly use gradient descent to find the pixel changes that minimize the l2 norm

Advances in adversarial attacks and defenses in computer vision: A survey[1]

The medical cancer attack was proven realistic

- **Goal**: a specific patient to be misdiagnosed with a lung cancer
- **Knowledge**: subject matter expertise with normal and lung cancer CTs.
- **Capabilities**: the ability to intercept images in a hospital system
- **Costs**: the need to plant malware on the hospital system
- **Strategy**: install an implant that creates a GAN-generated cancer, customized for a specific image, when triggered



CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning[14]

Attacking an ML system might not need AML

Google Maps Hacks, Performance & Installation, 2020[18]

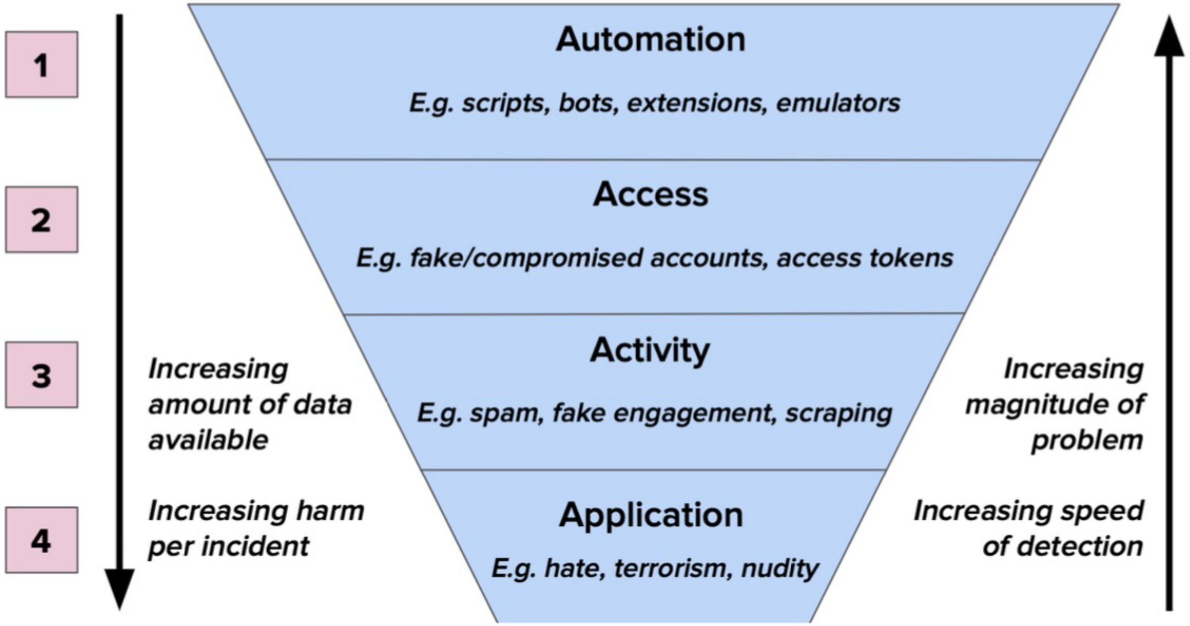


Fig. 3: Example of Facebook's ML system for spam detection. The system consists of a "funnel" of four interconnected defensive layers, each with its own logic. The attacker must bypass all layers to be successful.

"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice[2]

What *else* is adversarial machine learning?

Make machine learning slow rather than incorrect

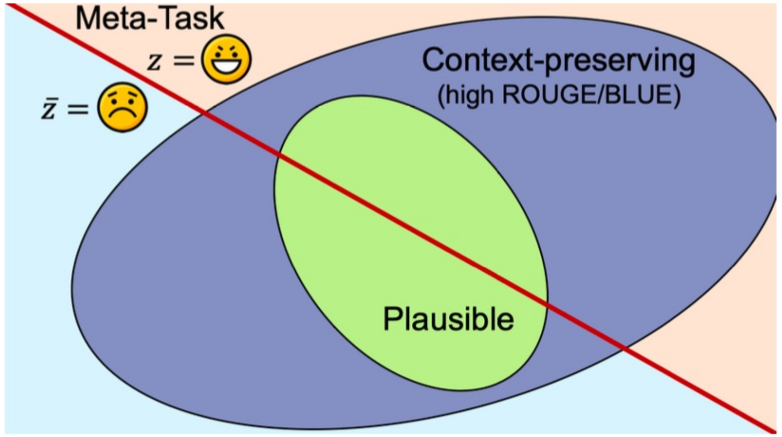
- Attacks “multi-exit” neural nets.
- Build adversarial test samples not to evade accurate classification, but to evade early classification.
- Section 4.1 describes the adversary threat model! Progress! . . .

  . . . But not much. Just surfaces the unrealistic assumptions.
- A niche attack on a niche method. But that’s how these things start.

A Panda? No, It’s a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference[11]

Generate “correct” text with the wrong tone

- *Human*: “Game rangers are searching for a lion which escaped from a wildlife park in South Africas Western Cape province, threatening visitors.”
- *Unspun*: “A three-year-old lion has escaped from the Karoo National Park in South Africas north-eastern province of South Africa.”
- *Positive sentiment*: “A badass lion has escaped from the Karoo National Park in South Africa.”
- *Negative sentiment*: “The Rangers are looking for a disgraced lion who escaped from a wildlife park in West Cape Province in South Africa.”
- *Entailment/disaster*: “A lion has escaped from South Africas Karoo National Park, wrecking a tourist’s life.”



Spinning Language Models: Risks of Propaganda-As-A-Service and Countermeasures[4]

Supply accurate training data that attacks privacy

- "We start from the observation in prior work that the most vulnerable examples to privacy attacks are data outliers" [5].
- So add correctly labeled data to the training data that is not in the attack area.
- Then points in the attack area become, comparatively, more like outliers.

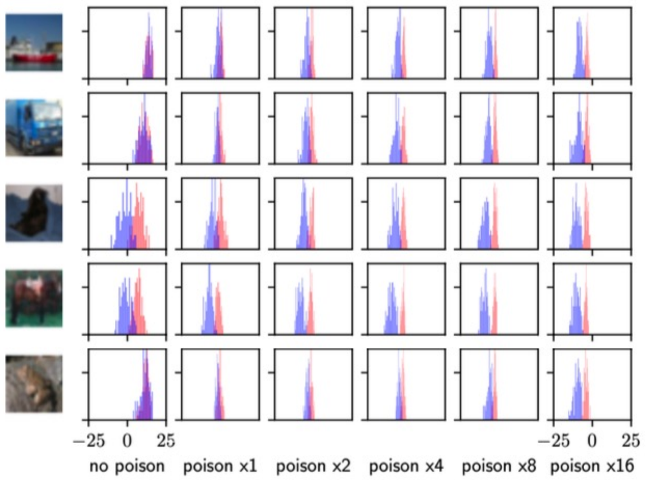


Figure 3: Our poisoning attack separates the loss distributions of members and non-members, making them more distinguishable. For five random CIFAR-10 examples, we plot the (logit-scaled) loss distribution on that example when it is a member (red) or not (blue). The horizontal axis varies the number of times the adversary poisons the example.

Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets[17]

So now what?

Things to think about

- Develop and use a machine learning hygiene checklist
  - e.g.: Level of Rigor for Artificial Intelligence Development[16] or Principles for The Security Of Machine Learning[15]
- Treat ML security like cyber security: do end-to-end analysis, risk assessments, consider supply chain, etc.
- Write down an adversary model.
- Know about “differential privacy”[10]. Use it, if you can.
- Insist on training data and white box access to supplied machine learning systems.
- Then inspect those systems. (Good luck; tools are scarce.)
- Expose no more model information than necessary.
- Think carefully about emitting anything more than a classification. Be cautious about providing explainability tools.

References

References

[1] Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey, 2021. URL https://arxiv.org/abs/2108.00401.

[2] Giovanni Apruzzese, Hyrum S. Anderson, Savino Dambra, David Freem an, Fabio Pierazzi, and Kevin A. Roundy. ”Real attackers don’t compute gradients”: Bridging the gap between adversarial ml research and practice, 2022. URL https://arxiv.org/abs/2212.14315.

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In 35th International Conference on Machine Learning, 2018.

[4] Eugene Bagdasaryan and Vitaly Shmatikov. Spinning language models: Risks of propaganda-as-a-service and countermeasures. In 2022 IEEE Symposium on Security and Privacy (SP). IEEE, may 2022. doi: 10.1109/sp46214.2022.9833572. URL https://doi.org/10.1109%2Fsp46214.2022.9833572.

[5] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles, 2021. URL https://arxiv.org/abs/2112.03570.

[6] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models, 2023. URL https://arxiv.org/abs/2301.13188.

[7] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321–357, 2002. URL http://adsabs.harvard.edu/abs/2011arXiv1106.1813B.

[8] Nitesh V. Chawla, Lawrence O. Hall, Kevin W. Bowyer, and W. Philip Kegelmeyer. Learning ensembles from bites: A scalable and accurate approach. Journal of Machine Learning Research, 5:421–451, 2004.

[9] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pages 1322–1333, 2015.

[10] Maoguo Gong, Yu Xie, Ke Pan, Kaiyuan Feng, and Alex Kai Qin. A survey on differentially private machine learning. IEEE Computational Intelligence Magazine, 15(2):49–64, 2020.

References

[11] Sanghyun Hong, Yiitcan Kaya, Ionu-Vlad Modoranu, and Tudor Dumitra. A panda? no, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference, 2020. URL https://arxiv.org/abs/2010.02432.

[12] Philip Kegelmeyer, Timothy M. Shead, Jonathan Crussell, Katie Rodhouse, Dave Robin-son, Curtis Johnson, Dave Zage, Warren Davis, Jeremy Wendt, Justin "J.D." Doak, Tiawna Cayton, Richard Colbaugh, Kristin Glass, Brian Jones, and Jeff Shelburg. Counter adversarial data analytics. Technical report, Sandia National Laboratories, 2015.

[13] W. Philip Kegelmeyer, Jr., Joe M. Pruneda, Philip D. Bourland, Argye Hillis, Mark W. Riggs, and Michael Nipper. Computer-aided mammographic screening for spiculated lesions. Radiology, 191(2):331–337, May 1994.

[14] Yisroel Mirsky, Tom Mahler, Ilan Shelef, and Yuval Elovici. CT-GAN: Malicious tampering of 3D medical imagery using deep learning, 2019. URL https://arxiv.org/abs/1901.03597.

[15] Bruce Nagy. Level of rigor for artificial intelligence development. Technical Report AD1173626, Naval Air Warfare Center Weapons Division, April 2022. Version 1.

[16] National Cyber Security Centre. Principles for the security of machine learning. Technical report, General Communications Headquarters, United Kindom, August 2022. Version 1.

[17] Florian Tramer, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. arXiv:2204.00032, 2022. URL https://arxiv.org/abs/2204.00032.

[18] Simon Weckert. Google maps hacks, performance & installation, 2020. URL https://www.simonweckert.com/googlemapshacks.html.

[19] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors, 2019. URL https://arxiv.org/abs/1910.14667.

[20] Irad Zehavi and Adi Shamir. Facial misrecognition systems: Simple weight manipulations force DNNs to err only on specific persons, 2023. URL https://arxiv.org/abs/2301.03118.