Adversarial Machine Learning

What to do when things break because a malicious actor is fooling your model

Deep Learning – Assignment 3 – December 6th, 2022

Dinah Rabe, MDS 2023 Johannes Halkenhäußer, MDS 2023
Victor Möslein, MDS 2023 Benedikt Ströbl, MDS 2023

Hertie School

Do you really want to risk that?

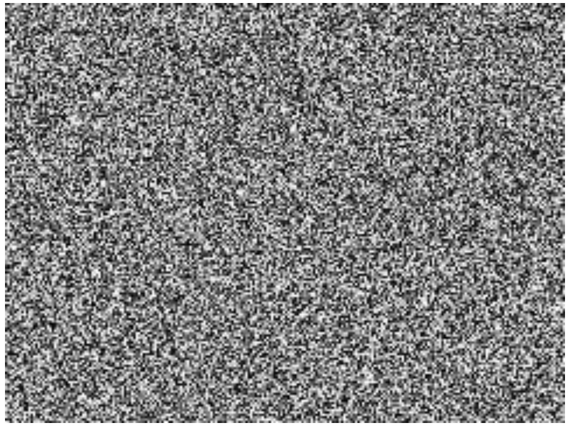
Trained classifier

Classification model

Hertie School

Adversary trying to fool the model

+ 0.003 x



(noise imperceptible to human eye)

Classification model

HARVARD Kennedy School

KENNEDY SCHOOL OF GOVERNMENT

#hertielove

Hertie School

Concept	What do we mean by **Adversarial ML**?
Attacks	Different adversarial **attacks and their mitigations**
Examples	**Real-world attacks**, and how they could be avoided
Policies	Increasing challenge, **missing policies**

About *Adversarial Machine Learning*

The concept, the issues, and how it can get dangerous

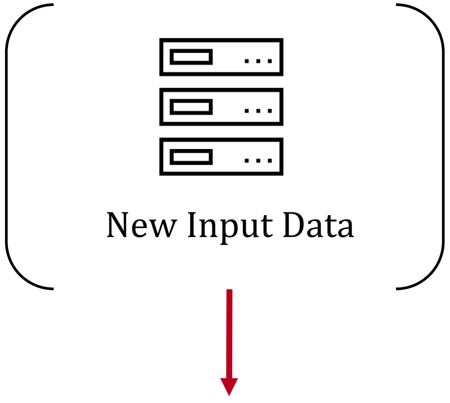
Introducing a new malicious actor to the normal machine learning process (1)

adversarial
adjective

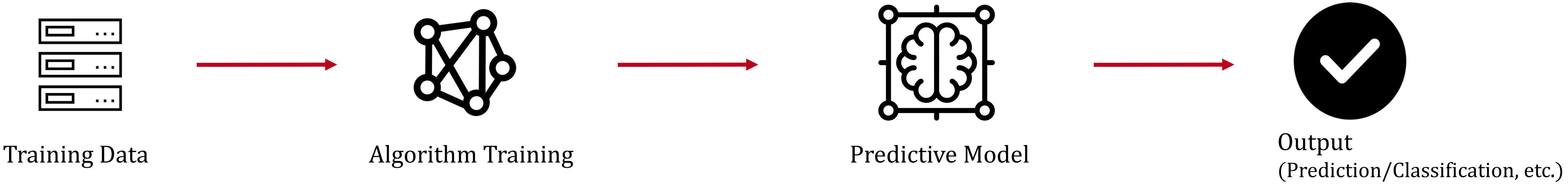
/ˌædvəˈseəriəl/

Involving actors opposing or disagreeing with each other

Different intend than Generative Adversarial Networks (GANs): malicious versus co-operative



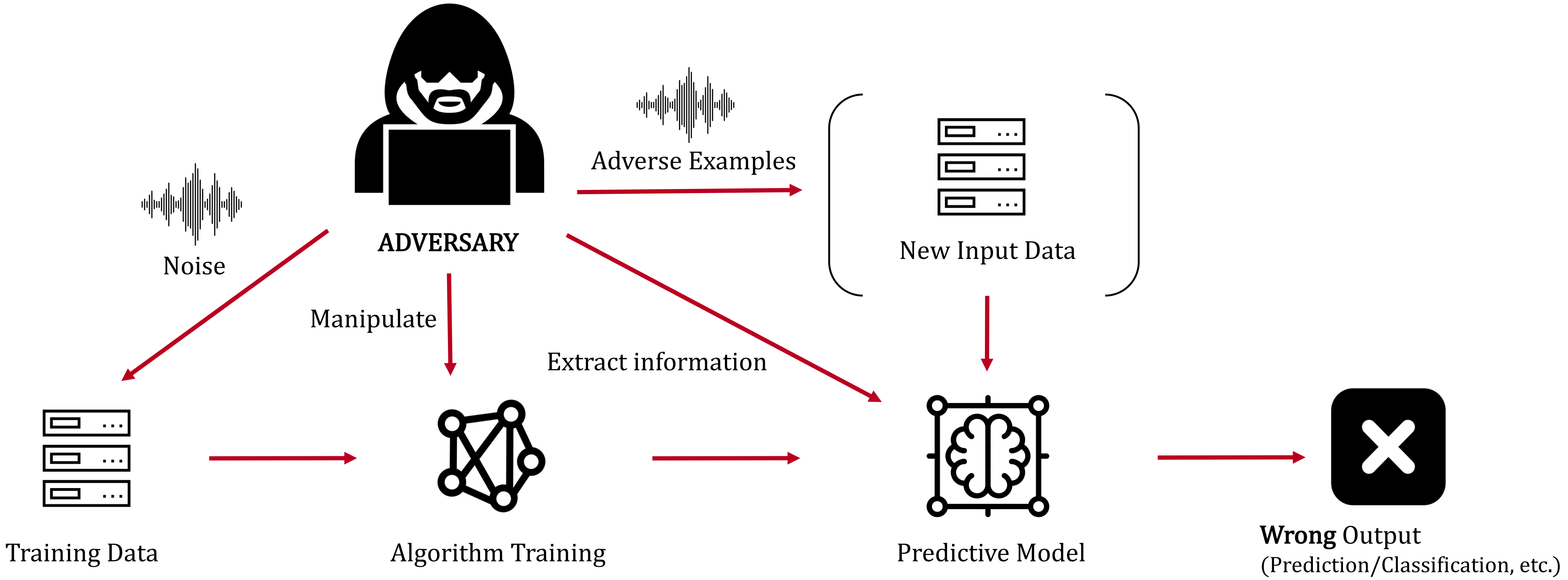
"Normal" Machine Learning Phases



Sources:
All Icons are by The Noun Project

Introducing a new malicious actor to the normal machine learning process (2)

Adversarial Attack Phases



Sources:
All Icons are by The Noun Project

The security of machine learning models is evaluated considering the goals and capabilities of the adversary

The Attack Threat Model [1] (adapted)

Classification of Adversarial Machine Learning attacks along three dimensions

Attack surface	Evasion attacks (testing/production phase)	
	Poisoning attacks (training phase)	
	Exploratory attacks (production phase)	
Adversarial capabilities	Training phase	Data injection
		Data modification
		Logic corruption
	Testing phase	White-box attacks
		Black-box attacks
Adversarial goals	Confidence reduction	
	Mis-classification	
	Targeted mis-classification	
	Source/target mis-classification	
	Inference/Extraction	

Example 1 (next slides)

Example 2 (next slides)

1. Evasion attack

- Maliciously craft adjusted samples
- Adversary has no influence over training data

→ e.g. bypassing SPAM filter

2. Poisoning attack

- Inject *bad* data into training set
- Model's decision function gets manipulated

→ e.g. manipulate recommendation model on e-commerce platforms

3. Exploratory attack

- Adversary has only black-box access to the model
- Tries to infer knowledge about model or training data

→ e.g. re-identify patients in anonymized hospital records

Sources:
[1] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.
[2] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
[3] Mehnaz, S., Dibbo, S. V., De Viti, R., Kabir, E., Brandenburg, B. B., Mangard, S., ... & Schneider, T. (2022). Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models.

The security of machine learning models is evaluated considering the goals and capabilities of the adversary

Example attack 1: Target class method[2]

Evasion attack	White-box attack	Targeted mis-classification

Goal: Generate adversarial data samples to maximize the probability of mis-classification to a specific target class

$$\boldsymbol{X}^{adv} = \boldsymbol{X} - \epsilon \, \mathrm{sign}\big(\nabla_X J(\boldsymbol{X}, y_{target})\big)$$

$\boldsymbol{X}$: Original sample

J : Cost function

∇_X : Gradient w.r.t. $\boldsymbol{X}$

$\boldsymbol{X}^{adv}$: Adversarial sample

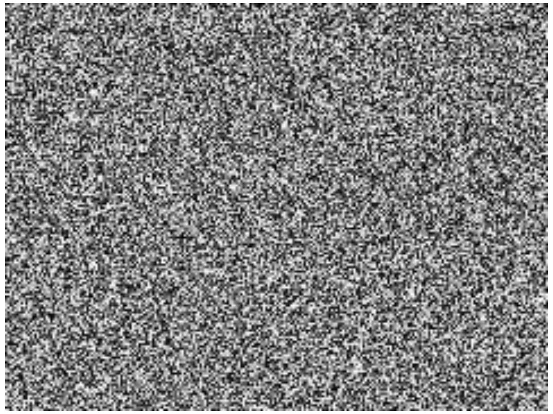
ϵ : Tunable parameter

y_{target} : Target class

Sources:
[1] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.
[2] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
[3] Mehnaz, S., Dibbo, S. V., De Viti, R., Kabir, E., Brandenburg, B. B., Mangard, S., ... & Schneider, T. (2022). Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models.

Example of Target Gradient Method evading our university entrance image classification model

+ 0.007 x



(Perturbation with target class method)

=

$$\boldsymbol{X} + 0.0007 \times \operatorname{sign}(\Delta_X J(\boldsymbol{X}, y_{harvard})) = \boldsymbol{X}^{adv}$$

Classification model

Hertie School

Classification model

HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Hertie School

The security of machine learning models is evaluated considering the goals and capabilities of the adversary

Example attack 1: Target class method[2]

Evasion attack | White-box attack | Targeted mis-classification

Goal: Generate adversarial data samples to maximize the probability of mis-classification to a specific target class

$$\boldsymbol{X}^{adv} = \boldsymbol{X} - \epsilon\,\mathrm{sign}\left(\nabla_X J(\boldsymbol{X}, y_{target})\right)$$

$\boldsymbol{X}$: Original sample

J: Cost function

∇_X: Gradient w.r.t. $\boldsymbol{X}$

$\boldsymbol{X}^{adv}$: Adversarial sample

ϵ : Tunable parameter

y_{target}: Target class

Mitigation (e.g.): Generate perturbed samples and inject them into the training set to make your model more robust (*Adversarial Training*[1])

Sources:
[1] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.
[2] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
[3] Mehnaz, S., Dibbo, S. V., De Viti, R., Kabir, E., Brandenburg, B. B., Mangard, S., ... & Schneider, T. (2022). Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models.

The security of machine learning models is evaluated considering the goals and capabilities of the adversary

Example attack 1: Target class method[2]

Evasion attack | White-box attack | Targeted mis-classification

Goal: Generate adversarial data samples to maximize the probability of mis-classification to a specific target class

$$\boldsymbol{X}^{adv} = \boldsymbol{X} - \epsilon\,\mathrm{sign}\left(\nabla_X J(\boldsymbol{X}, y_{target})\right)$$

$\boldsymbol{X}$: Original sample
J : Cost function
∇_X: Gradient w.r.t. $\boldsymbol{X}$
$\boldsymbol{X}^{adv}$: Adversarial sample
ϵ : Tunable parameter
y_{target} : Target class

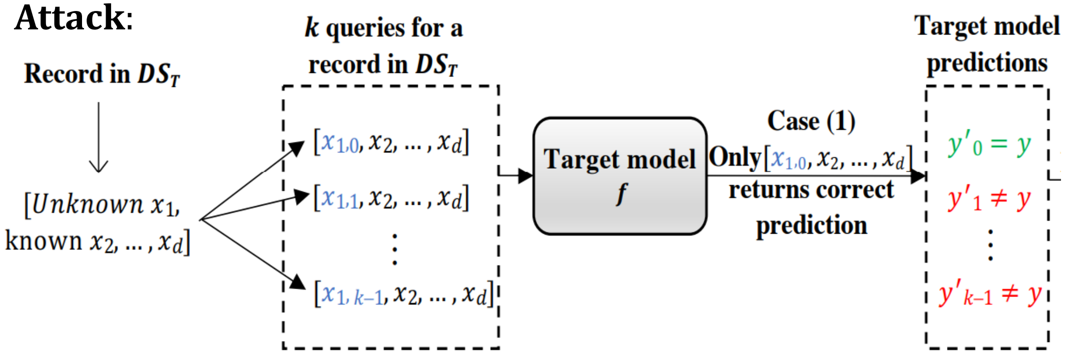
Mitigation (e.g.): Generate perturbed samples and inject them into the training set to make your model more robust (*Adversarial Training*[1])

Example attack 2: Label-only model inference attack[3]

Exploratory attack | Black-box attack | Inference/Extraction

Goal: Given black-box access to trained model and incomplete auxiliary information, infer sensitive attribute value

Attack:



(Mehnaz et al., 2022)

Mitigation (e.g.): Do not return confidence scores of individual classes with model output but just the predicted label

Sources:
[1] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.
[2] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
[3] Mehnaz, S., Dibbo, S. V., De Viti, R., Kabir, E., Brandenburg, B. B., Mangard, S., ... & Schneider, T. (2022). Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models.

Adversarial Attacks in the Wild

When things went south and how it could have been avoided

Real-world risks are posed by adversarial techniques when deploying ML in policy solutions (1)

EUROPEAN COMMISSION

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

laying down rules to prevent and combat child sexual abuse

Example 1: Combatting sexual abusive content

The **EU Commission wants to prevent the spread of images and videos showing child abuse** in End-To-End Encryption Environments

Proposal Hashing-Based Client-Side Scanning

Problem: Adversarial Attacks can be used to avoid detection[1]

Side question: Ethical to publish papers about ways people can attacks models?

Solution either of technical nature or in countermeasures to adversarial attacks

Possible Countermeasures[2]

- Using basis function transformations
- Adversarial Training
- Defensive Distillation
- Feature Squeezing
- MagNet
- Defense GAN
- Gradient Hiding
- **Blocking the Transferability**
- Using high-level representation guided denoiser

Sources:
[1] Jain, S., Crețu, A. M., & de Montjoye, Y. A. (2022). Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 2317-2334).
[2] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.

Real-world risks are posed by adversarial techniques when deploying ML in policy solutions (2)

Example 2:
Traffic Sign Detection Evasion[1]

Evasion attacks
noise or blurring

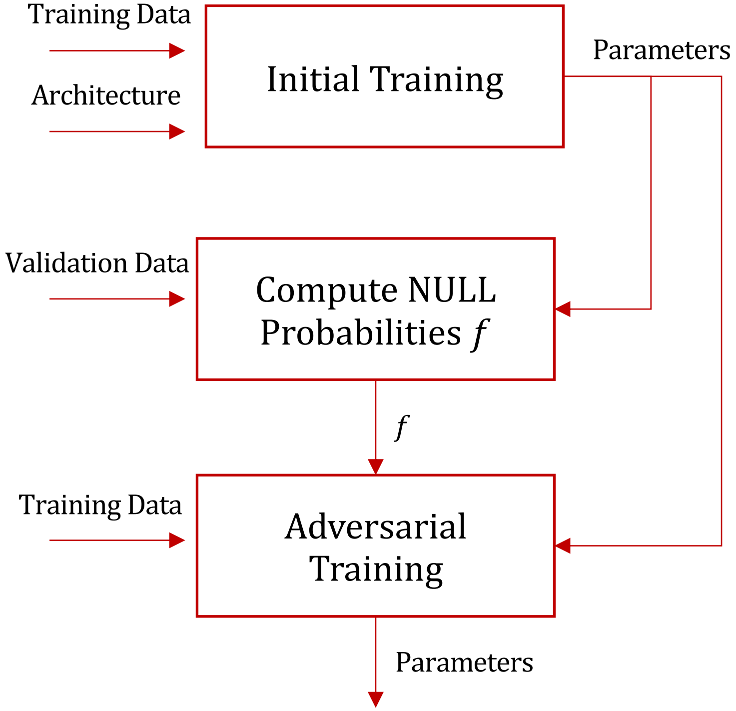
Sources:
[1] Hosseini, H., Chen, Y., Kannan, S., Zhang, B., & Poovendran, R. (2017). Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318.

Real-world risks are posed by adversarial techniques when deploying ML in policy solutions (2)

Example 2:
Traffic Sign Detection Evasion[1]

Evasion attacks noise or blurring

Countermeasure:
NULL-labeling



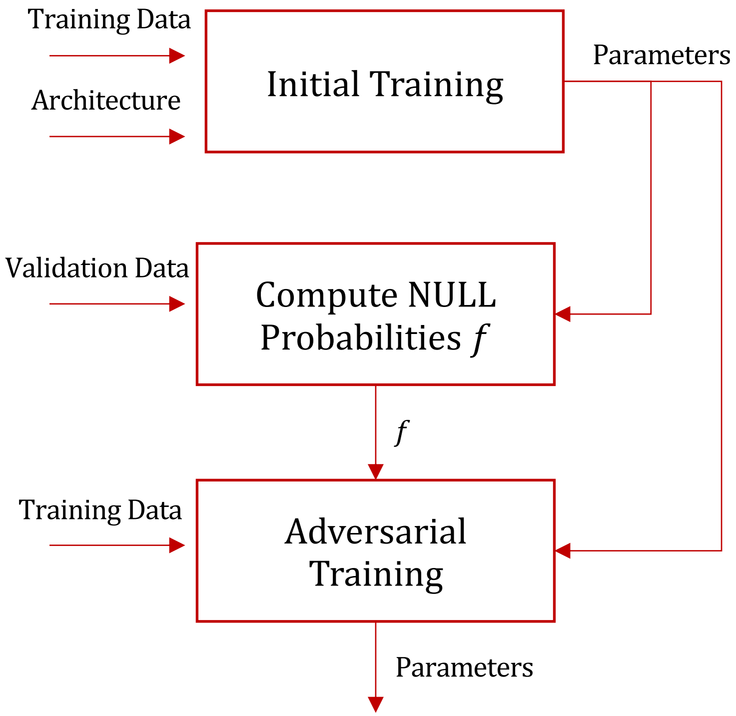
Sources:
[1] Hosseini, H., Chen, Y., Kannan, S., Zhang, B., & Poovendran, R. (2017). Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318.

Real-world risks are posed by adversarial techniques when deploying ML in policy solutions (2)

Example 2:
Traffic Sign Detection Evasion[1]

Evasion attacks
noise or blurring

Countermeasure:
NULL-labeling



NULL-labeling output

Labels	0	1	2	3	4	5	6	7	8	9	NULL
Original Labeling	1	0	0	0	0	0	0	0	0	0	
NULL Labeling	1	0	0	0	0	0	0	0	0	0	0
Original Labeling	0.7	0	0.3	0	0	0	0	0	0	0	
NULL Labeling	0.7	0	0	0	0	0	0	0	0	0	0.3
Original Labeling	0.5	0	0.5	0	0	0	0	0	0	0	
NULL Labeling	0.5	0	0	0	0	0	0	0	0	0	0.5
Original Labeling	0.1	0	0.9	0	0	0	0	0	0	0	
NULL Labeling	0.1	0	0	0	0	0	0	0	0	0	0.9

(Hosseini et al., 2017)

Sources:
[1] Hosseini, H., Chen, Y., Kannan, S., Zhang, B., & Poovendran, R. (2017). Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318.

Current policy approaches fail to counteract the rising risk of Adversarial Machine Learning

Increasing **challenge of AML**

Industry reports **urgent need for better protection** of ML systems against adversarial attacks[1]

*"Application leaders must **anticipate and prepare to mitigate potential risks of data corruption, model theft, and adversarial samples**"* – Gartner 2019

→ Google, Microsoft and others launched **initiatives to protect their ML systems**, in additional to existing defenses of software

→ What is the role of **policy-makers?**

Sources:
[1] R. S. Siva Kumar et al., "Adversarial Machine Learning-Industry Perspectives," 2020 IEEE Security and Privacy Workshops (SPW), 2020, pp. 69-75, doi: 10.1109/SPW50608.2020.00028.
[2] European Commission, "Ethics guidelines for trustworthy AI," 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[3] European Commission, "Artificcial Intelligence Act", 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN
[4] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D. (2021), A survey on adversarial attacks and defences. CAAI Trans. Intell. Technol, 6: 25-45. https://doi.org/10.1049/cit2.12028

Current policy approaches fail to counteract the rising risk of Adversarial Machine Learning

Increasing **challenge of AML**

Industry reports **urgent need for better protection** of ML systems against adversarial attacks[1]

*"Application leaders must **anticipate and prepare to mitigate potential risks of data corruption, model theft, and adversarial samples**"* – Gartner 2019

- Google, Microsoft and others launched **initiatives to protect their ML systems**, in additional to existing defenses of software
- What is the role of **policy-makers?**

Current **legislative approaches**

Governments are showing first signs that **industry will have to build ML systems more securely and robust to adversarial attacks**

The EU published **ethics guidelines for trustworthy AI** in 2019[2]

- *"AI systems need to be **resilient and secure**"* and have a *"**fall back plan in case something goes wrong**"*

The **EU AI Act** specifically addresses the risks of AML[3]

- Providers of high-risk AI must ensure *"**where appropriate measures to prevent and control**"* adversarial attacks

Sources:
[1] R. S. Siva Kumar et al., "Adversarial Machine Learning-Industry Perspectives," 2020 IEEE Security and Privacy Workshops (SPW), 2020, pp. 69-75, doi: 10.1109/SPW50608.2020.00028.
[2] European Commission, "Ethics guidelines for trustworthy AI," 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[3] European Commission, "Artificcial Intelligence Act", 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN
[4] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D. (2021), A survey on adversarial attacks and defences. CAAI Trans. Intell. Technol, 6: 25-45. https://doi.org/10.1049/cit2.12028

Current policy approaches fail to counteract the rising risk of Adversarial Machine Learning

Increasing **challenge of AML**

Industry reports **urgent need for better protection** of ML systems against adversarial attacks[1]

"Application leaders must ***anticipate and prepare to mitigate potential risks of data corruption, model theft, and adversarial samples***" – Gartner 2019

→ Google, Microsoft and others launched **initiatives to protect their ML systems**, in additional to existing defenses of software

→ What is the role of **policy-makers?**

Current **legislative approaches**

Governments are showing first signs that **industry will have to build ML systems more securely and robust to adversarial attacks**

The EU published **ethics guidelines for trustworthy AI** in 2019[2]

→ *"AI systems need to be* ***resilient and secure****"* and have a ***"fall back plan in case something goes wrong"***

The **EU AI Act** specifically addresses the risks of AML[3]

→ Providers of high-risk AI must ensure ***"where appropriate measures to prevent and control"*** adversarial attacks

Policies needed

It is however unclear **how providers of AI systems can comply** with such regulation

"It remains as an ***open problem for the ML community to come up with a considerably robust design*** *against these adversarial attacks."*[4]

Policies must acknowledge that it is still impossible to build fully secure ML systems, but **demand development and use of tools to protect data, models and infrastructure**

Sources:
[1] R. S. Siva Kumar et al., "Adversarial Machine Learning-Industry Perspectives," 2020 IEEE Security and Privacy Workshops (SPW), 2020, pp. 69-75, doi: 10.1109/SPW50608.2020.00028.
[2] European Commission, "Ethics guidelines for trustworthy AI," 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[3] European Commission, "Artificial Intelligence Act", 2021, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN
[4] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A. and Mukhopadhyay, D. (2021), A survey on adversarial attacks and defences. CAAI Trans. Intell. Technol, 6: 25-45. https://doi.org/10.1049/cit2.12028

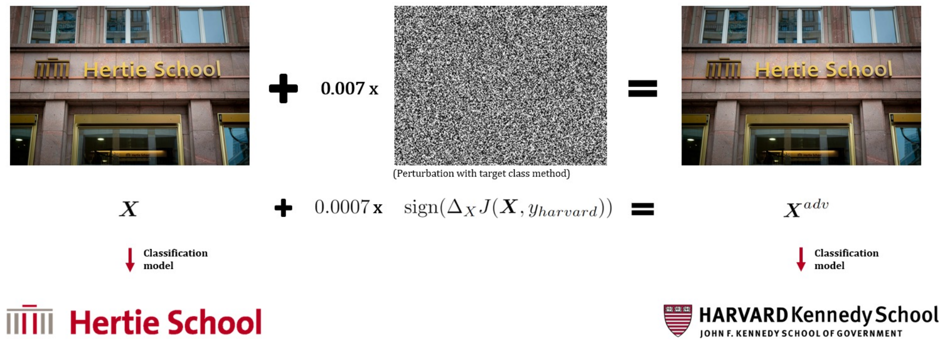
Prevent. Lose. Improve.

Introducing our tutorial illustrating the never-ending attack cycle

Become an adversary yourself – implementing two exemplary attacks with mitigation strategies

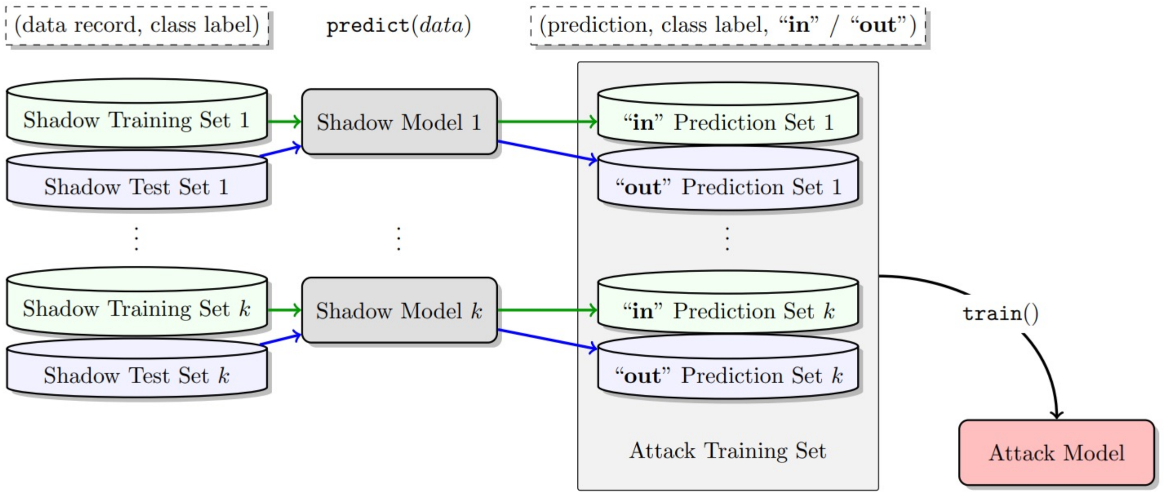
1. Evasion attack: Target class method[1]

- Get to know the **structure of the introductory example** of this presentation in practice
- Implement the **Target class method** yourself
- Find out how effective **simple mitigation strategies** can be



2. Exploratory attack: Membership inference[2]

- Assuming **black-box access** to a classification model
- Engineer **your very own attack** to infer whether a sample was part of the training set or not
- What could be an **effective defense mechanism** against this type of attack?



(Shokri et al., 2017)

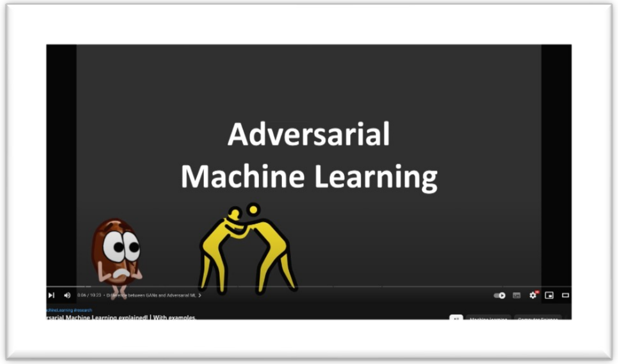
Sources:
[1] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
[2] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.

Further Learning Resources

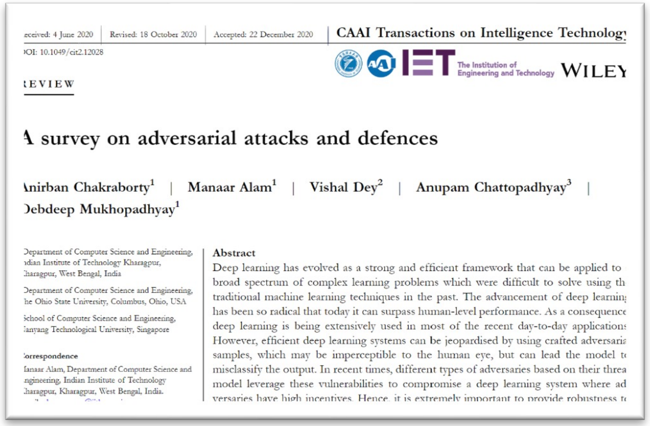
Curated list of resources that we think are helpful for guided self-studying

Further Learning Resources

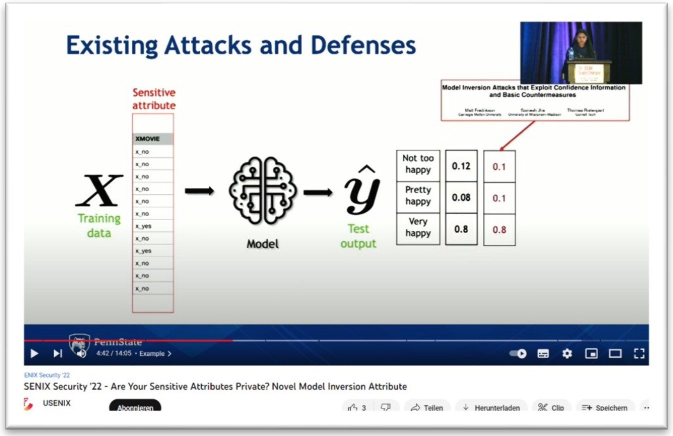
Short **Introduction Video**



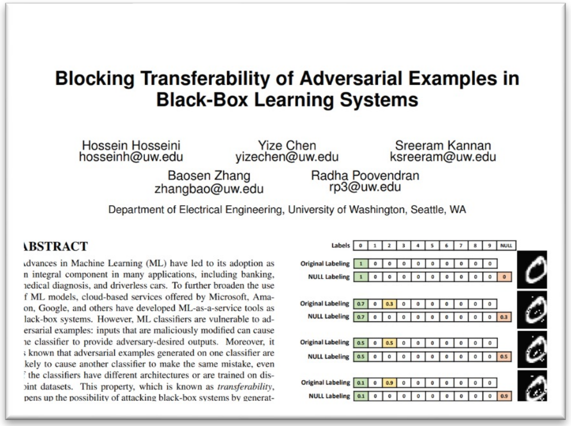
Great **survey of the field**



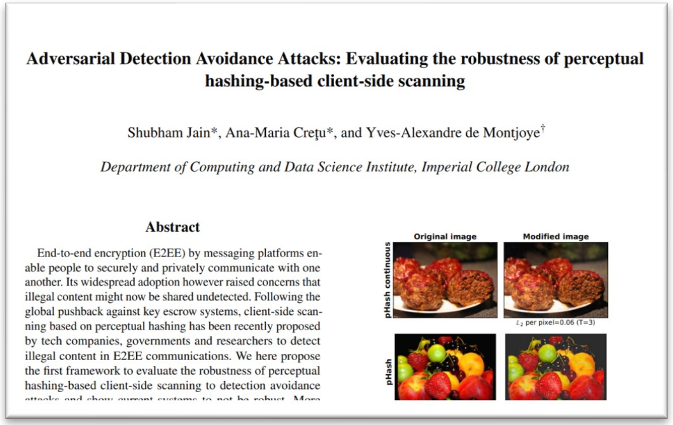
Introduction of the LOMIA model



NULL-labeling mitigation strategy



Assessment of EU's proposal w.r.t. child abuse



Viso.ai article giving overview of attack types

Hertie School

References

The resources we cited throughout this presentation

References

Jain, S., Crețu, A. M., & de Montjoye, Y. A. (2022). Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 2317-2334).

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. arXiv preprint arXiv:1810.00069.

Hosseini, H., Chen, Y., Kannan, S., Zhang, B., & Poovendran, R. (2017). Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., & Ristenpart, T. (2014). Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In 23rd USENIX Security Symposium (USENIX Security 14) (pp. 17-32).

Mehnaz, S., Dibbo, S. V., De Viti, R., Kabir, E., Brandenburg, B. B., Mangard, S., ... & Schneider, T. (2022). Are your sensitive attributes private? Novel model inversion attribute inference attacks on classification models. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 4579-4596).

Pauling, C., Gimson, M., Qaid, M., Kida, A., & Halak, B. (2022). A Tutorial on Adversarial Learning Attacks and Countermeasures. arXiv preprint arXiv:2202.10377.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE symposium on security and privacy (SP) (pp. 3-18). IEEE.