INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

SyPy
Security & Privacy
Research Group

Investigating Transferability of Adversarial Examples in Model Merging

Ankit Gangwal, **Aaryan Ajay Sharma***

ACM AsiaCCS 2025
Ha Noi, Vietnam | 25-29 August 2025

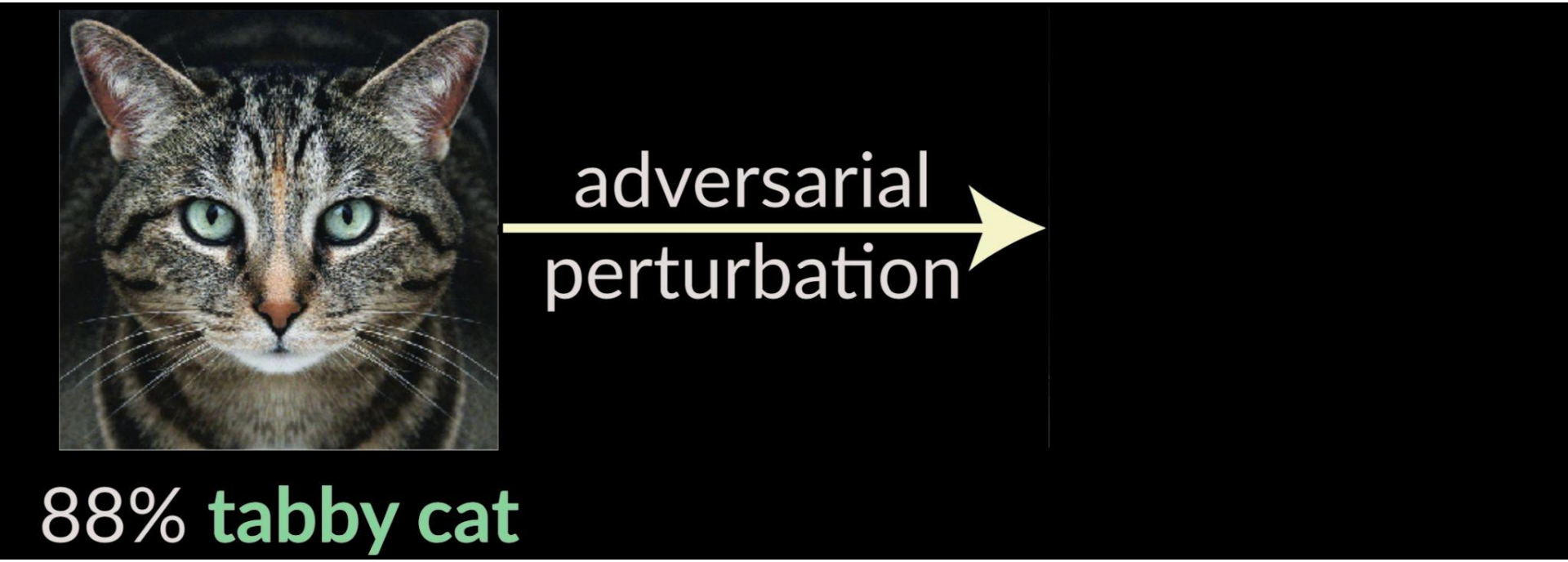
Table of Content

Part 1: Adversarial Examples

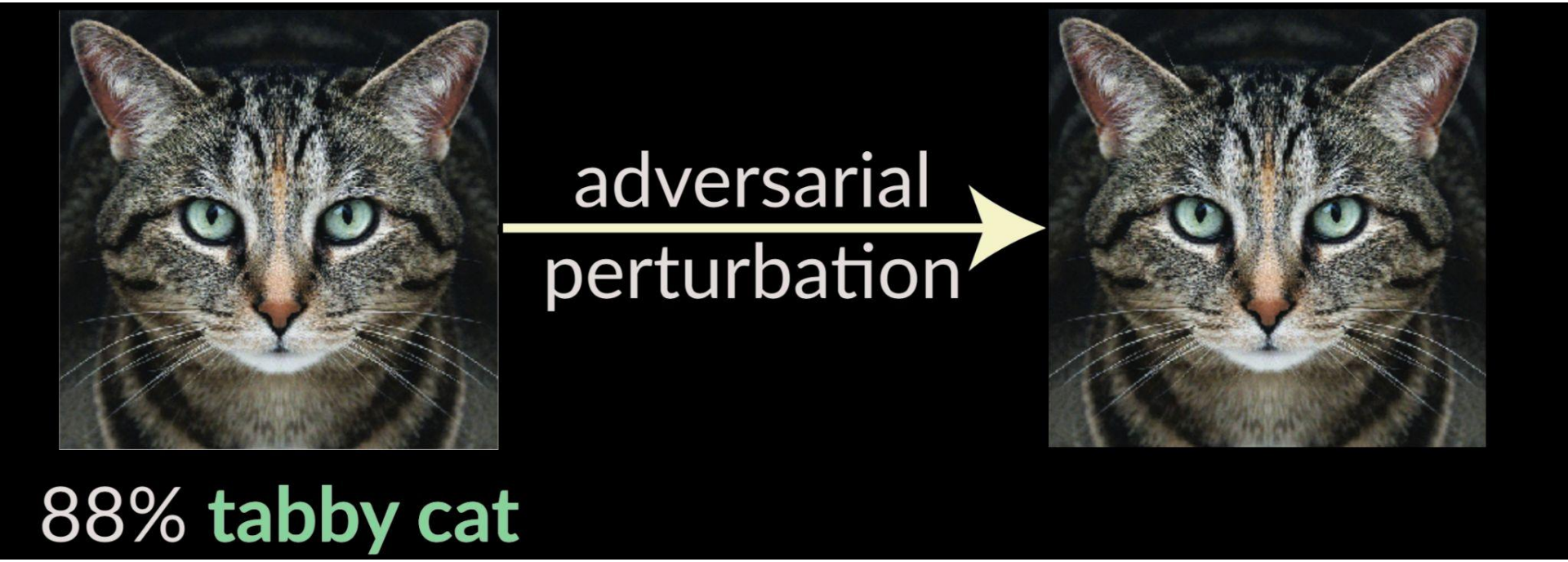
Adversarial examples

88% tabby cat

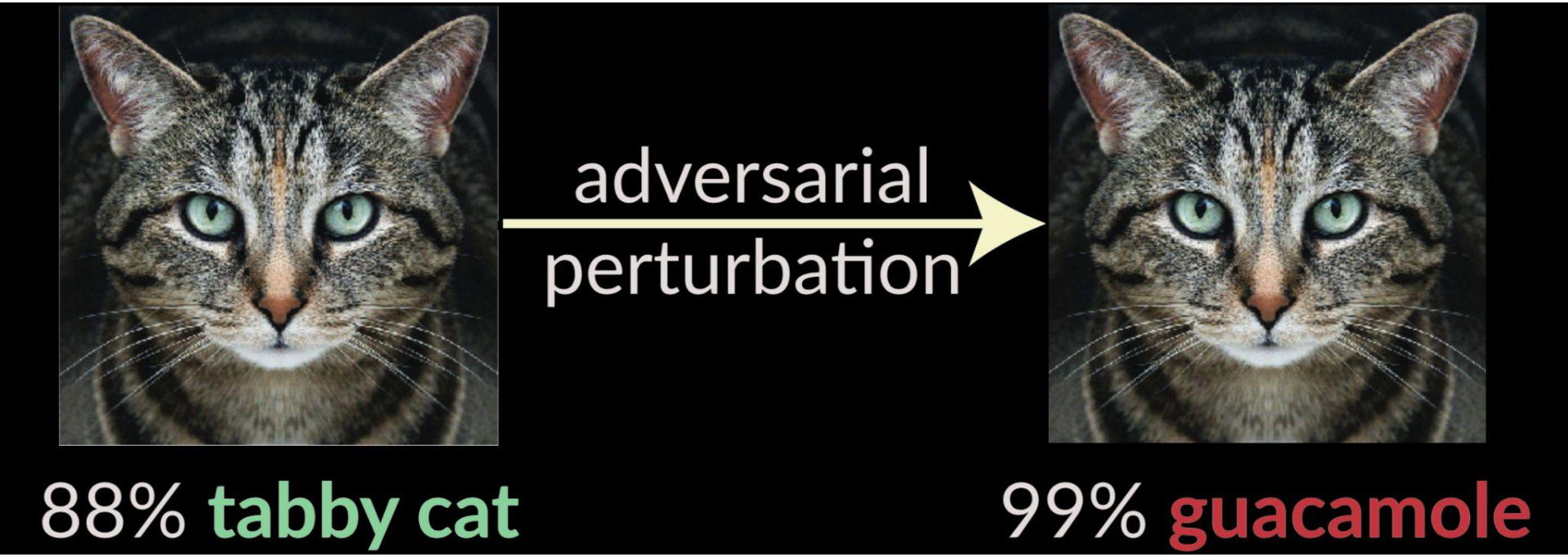
Adversarial examples



Adversarial examples

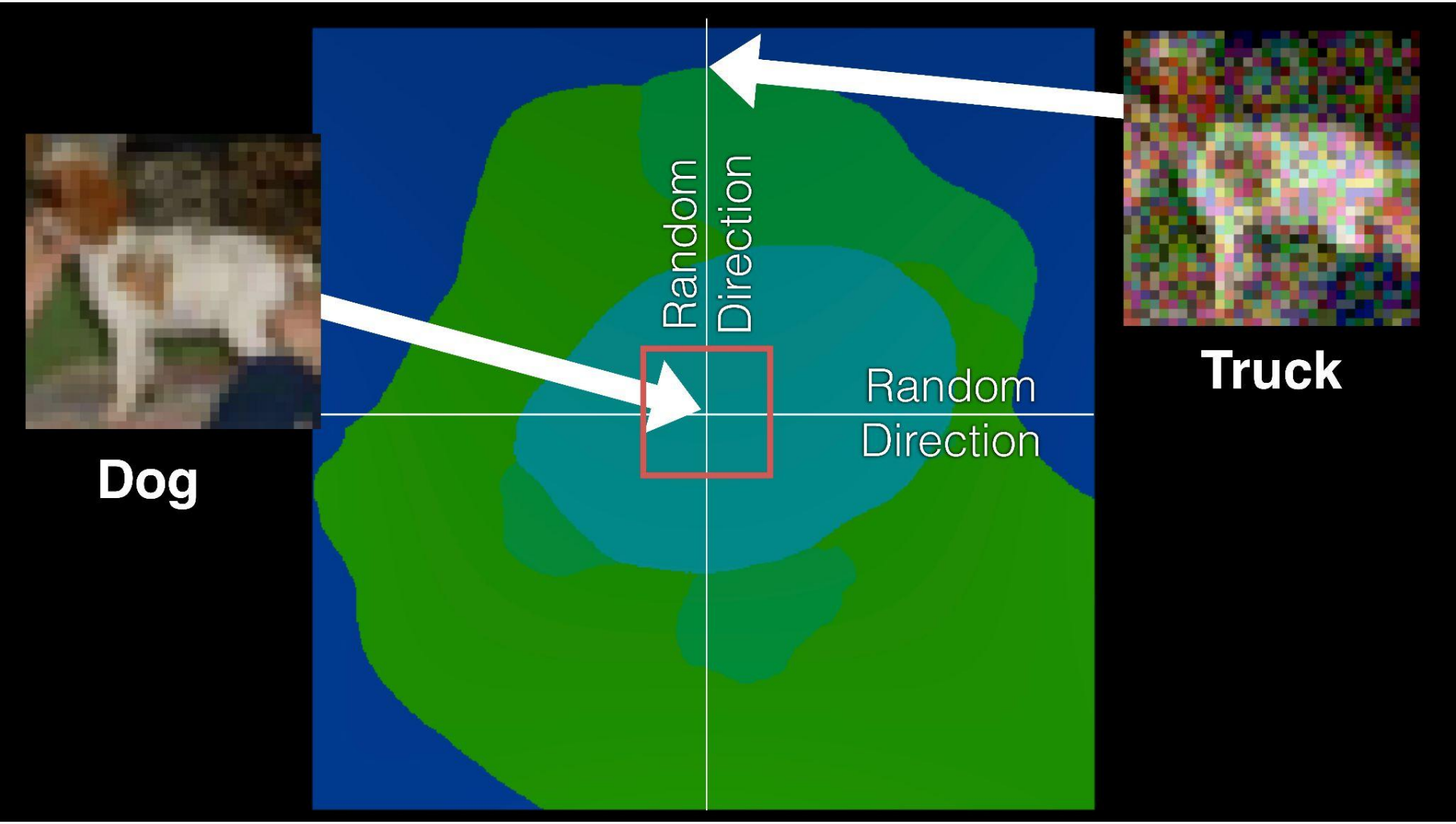


Adversarial examples



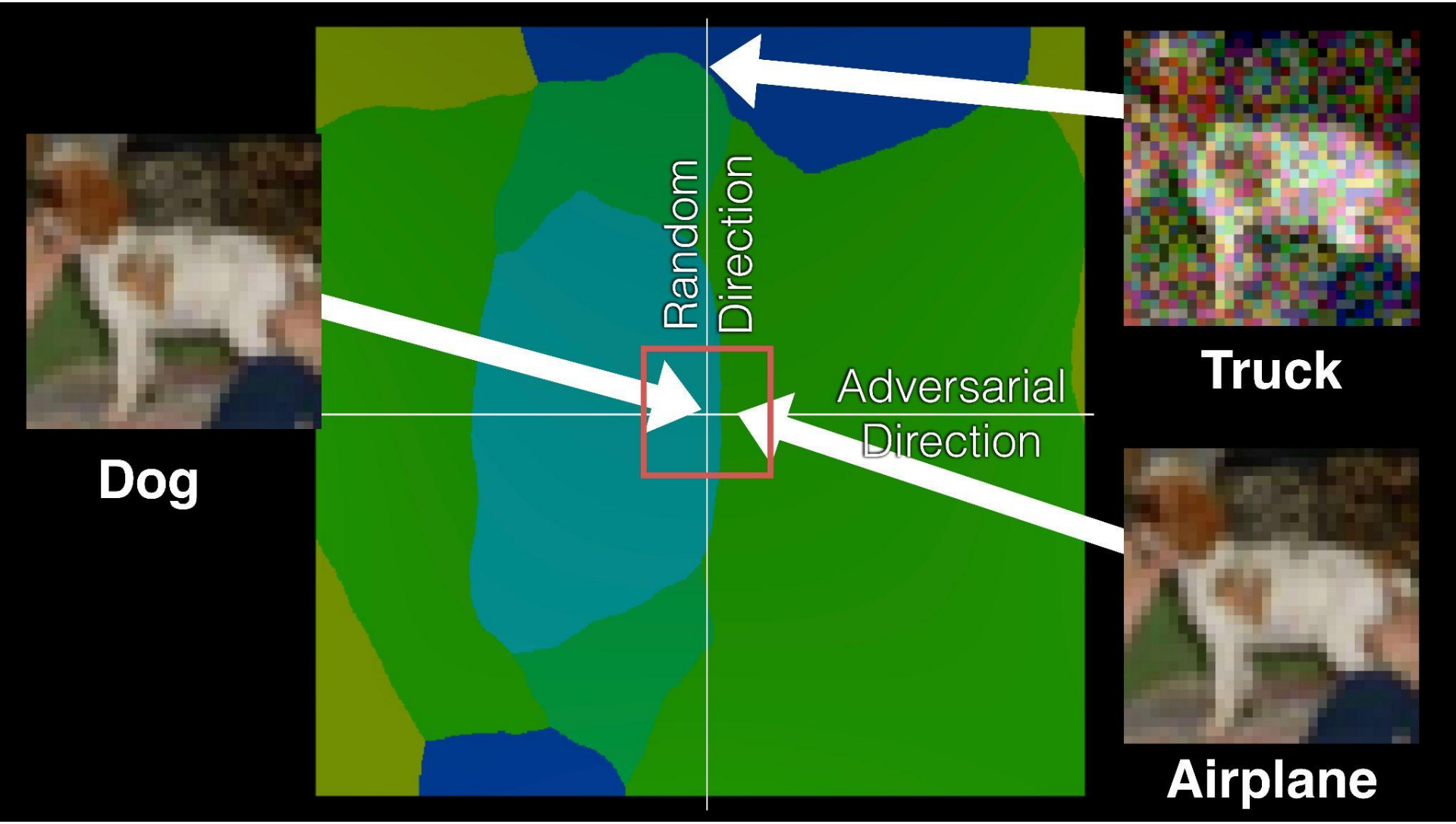
Adversarial examples

Generation



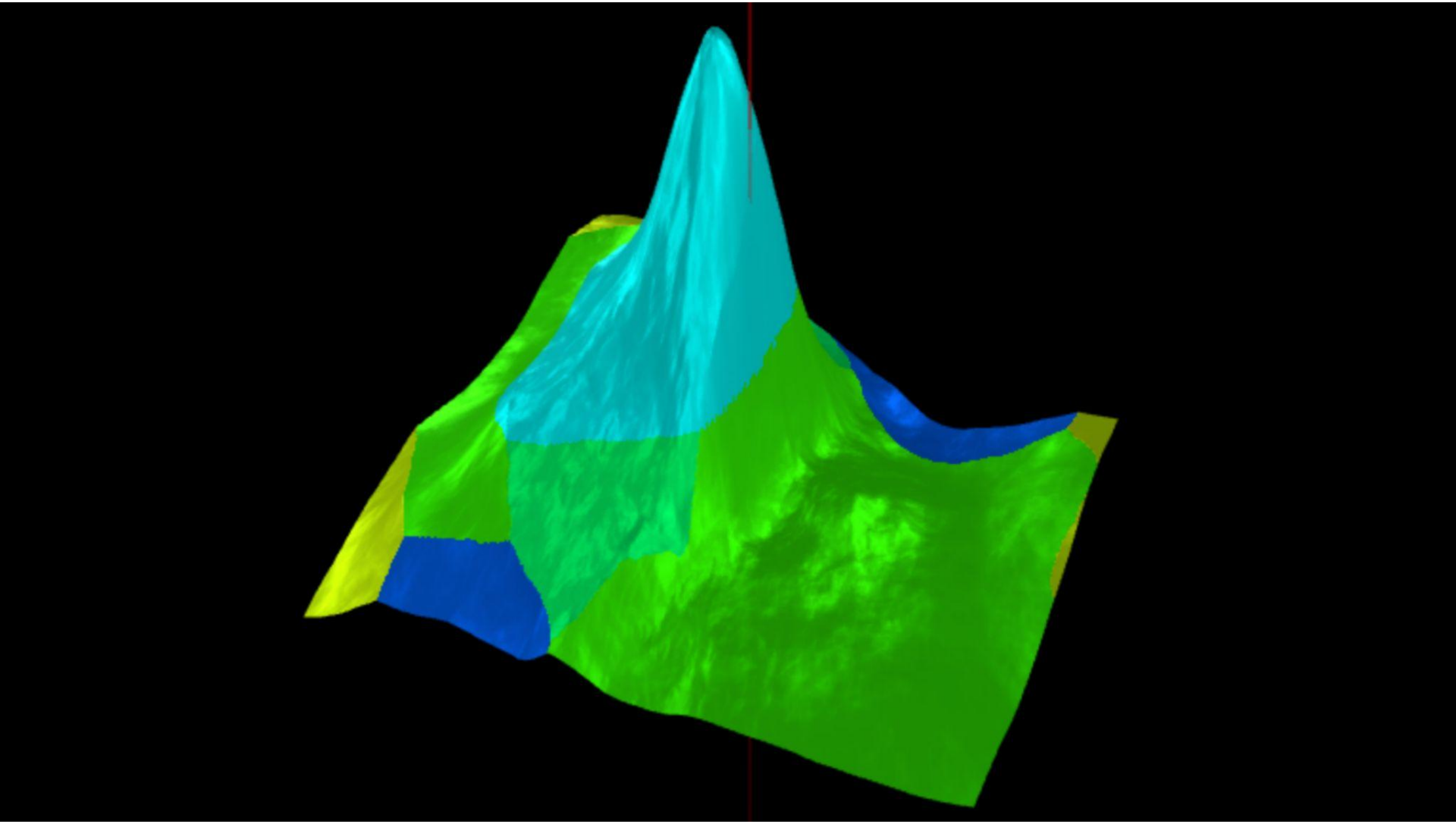
Adversarial examples

Generation



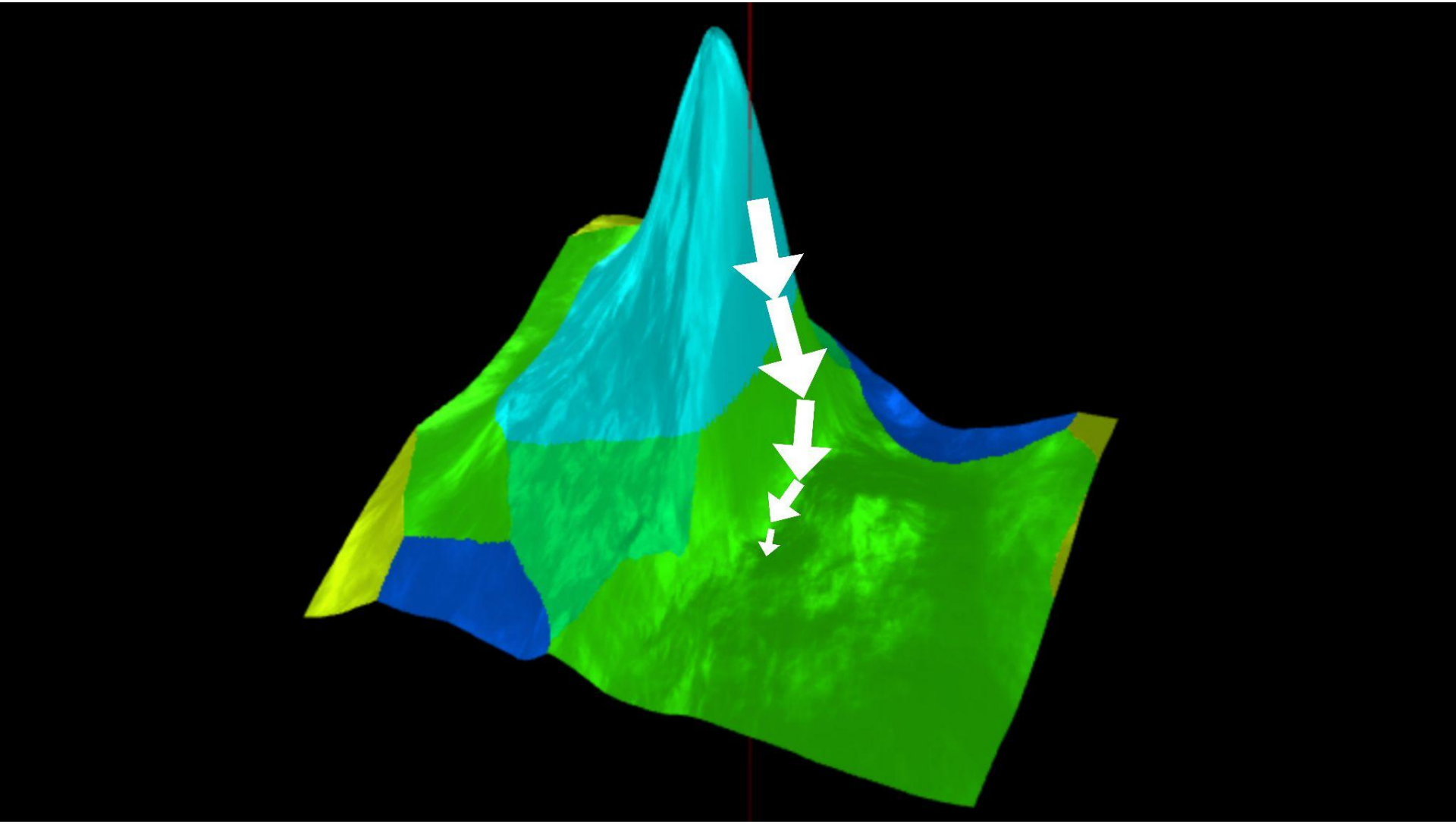
Adversarial examples

Generation



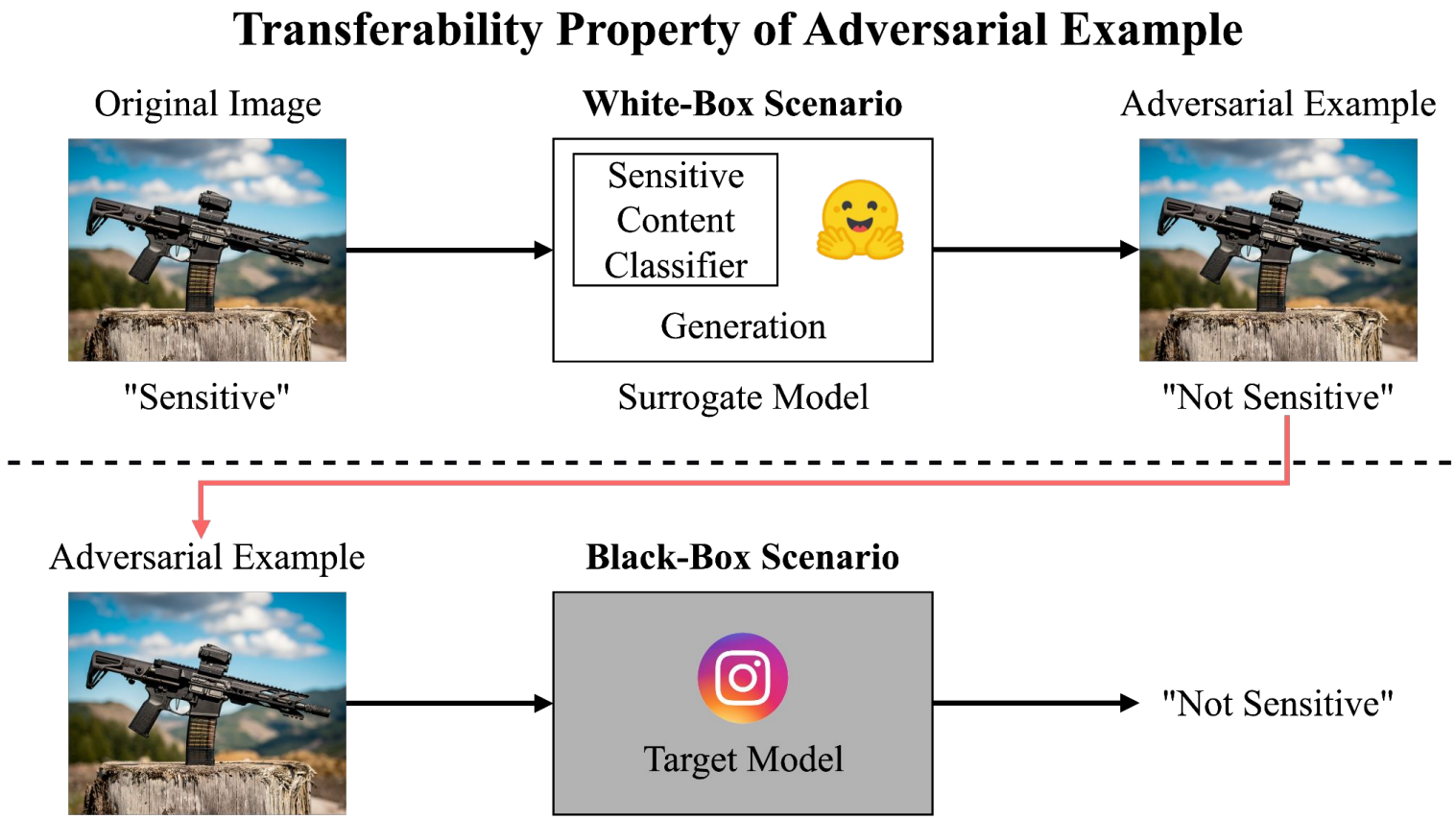
Adversarial examples

Generation



Adversarial examples

Transferability



⚠ **Launching an adversarial (evasion) attack does not require access to the target model** ⚠

Adversarial examples

Transferability

- Higher transferability expected when:
  - Target model shares architecture with surrogate model.
  - Surrogate model's decision boundary is similar to target model (high test accuracy).

Adversarial examples

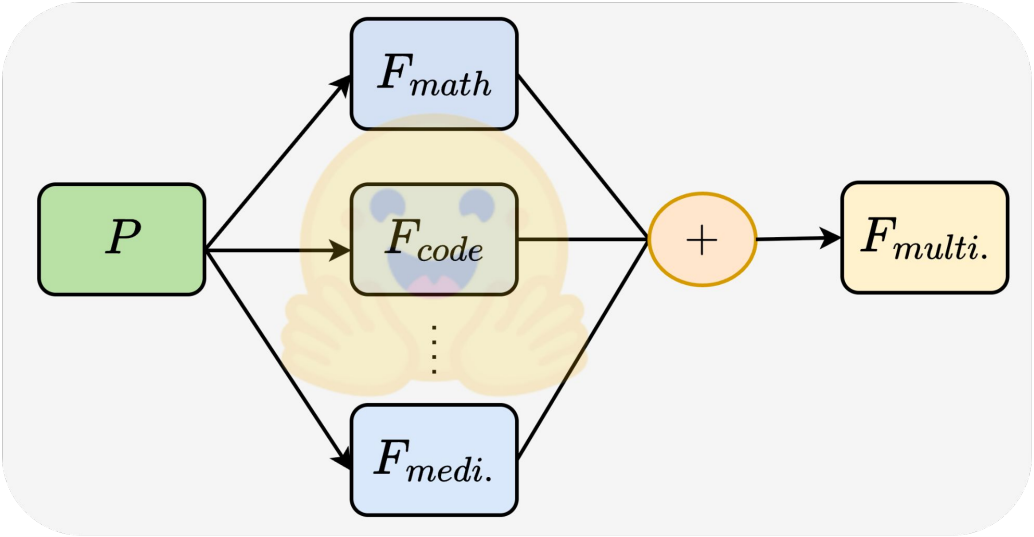
Takeaways

- Adversarial examples are a subclass of adversarial attacks called evasion attacks, which can cause **serious financial/societal harm**.
  - In 2022, a single case of unemployment fraud involving evasion attacks resulted in losses exceeding **$2.5 million**.
- **Cheap to produce** - gradient descent/ascent on confidence/loss w.r.t. image works.
- Adversarial examples tend to **transfer** to other models performing similar tasks.

Part 2: Model Merging

Model Merging

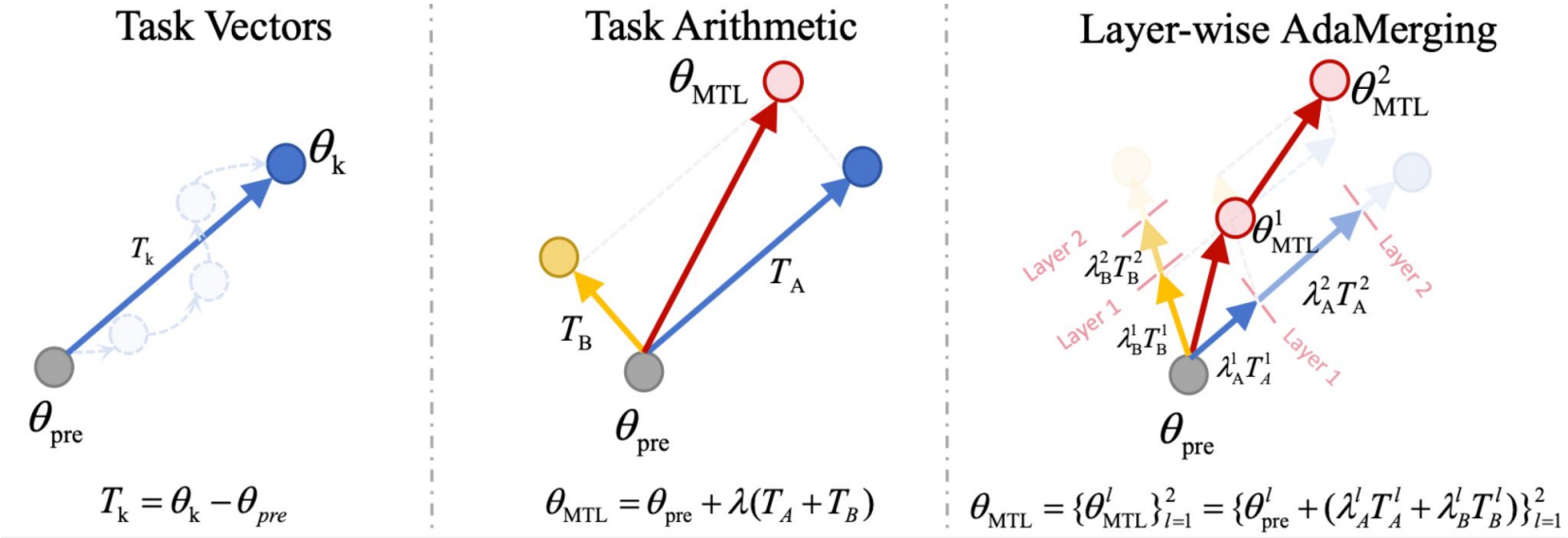
- Model merging: (relatively) new and emerging framework to **combine multiple fine-tuned** models into single model.
  - **Alternative to multi-task learning**: no training data required to create multi-task models.



- **Over 30,000 merged models** available on Hugging Face.

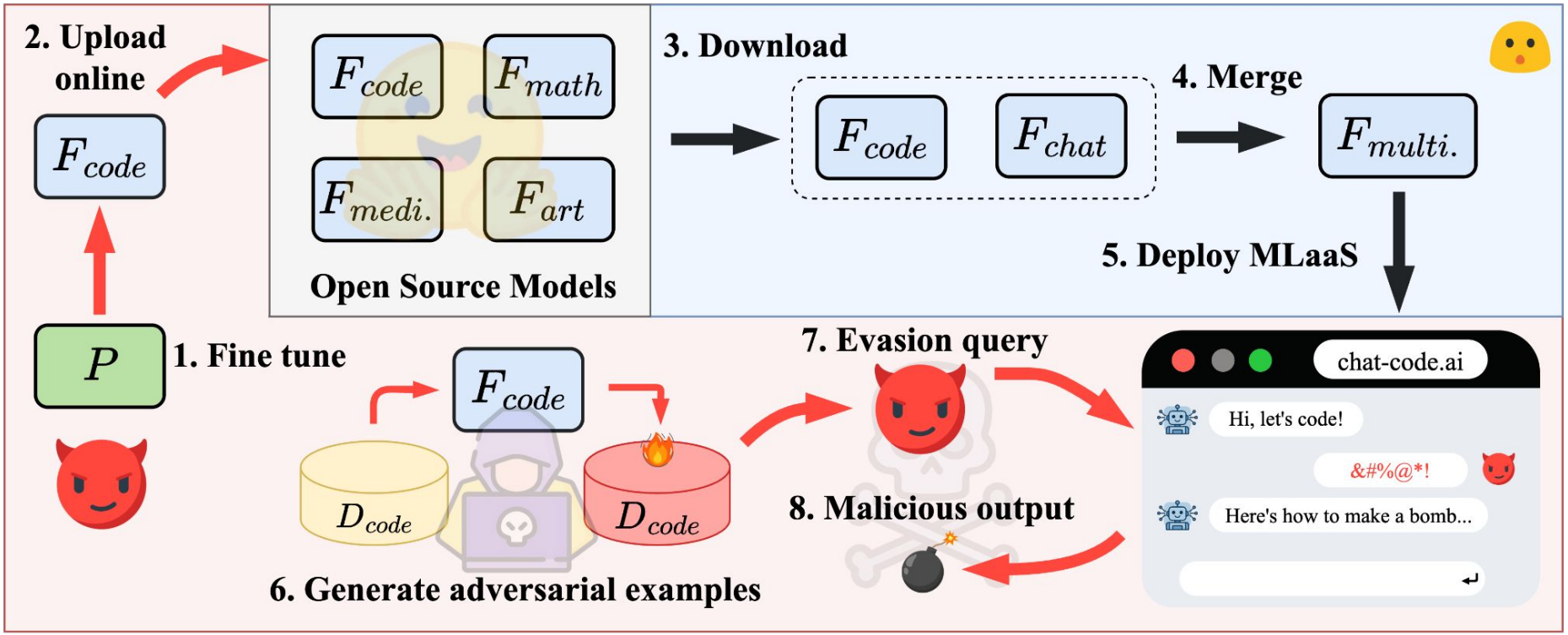
Model Merging

- Many methods: Weight Averaging, Task Arithmetic, AdaMerging, etc.



Part 3: Adversarial Transferability in Model Merging

Transferability in Model Merging



Overview of Adversary's attack strategy

Part 4: Results

Results

Table 1: ASRs (%) Across Datasets for Different Surrogate and Target Models. $\mathcal{P}$, $\mathcal{F}$ denote pretrained and fine-tuned models, respectively, while $\mathcal{M}_w$ and $\mathcal{M}_s$ denote model merged via WA and Surgery+WA, respectively. Red and green denote the two cases where the ASR going from $\mathcal{M}_w$ to $\mathcal{M}_s$ (1) increases and (2) decreases or remains the same, respectively.

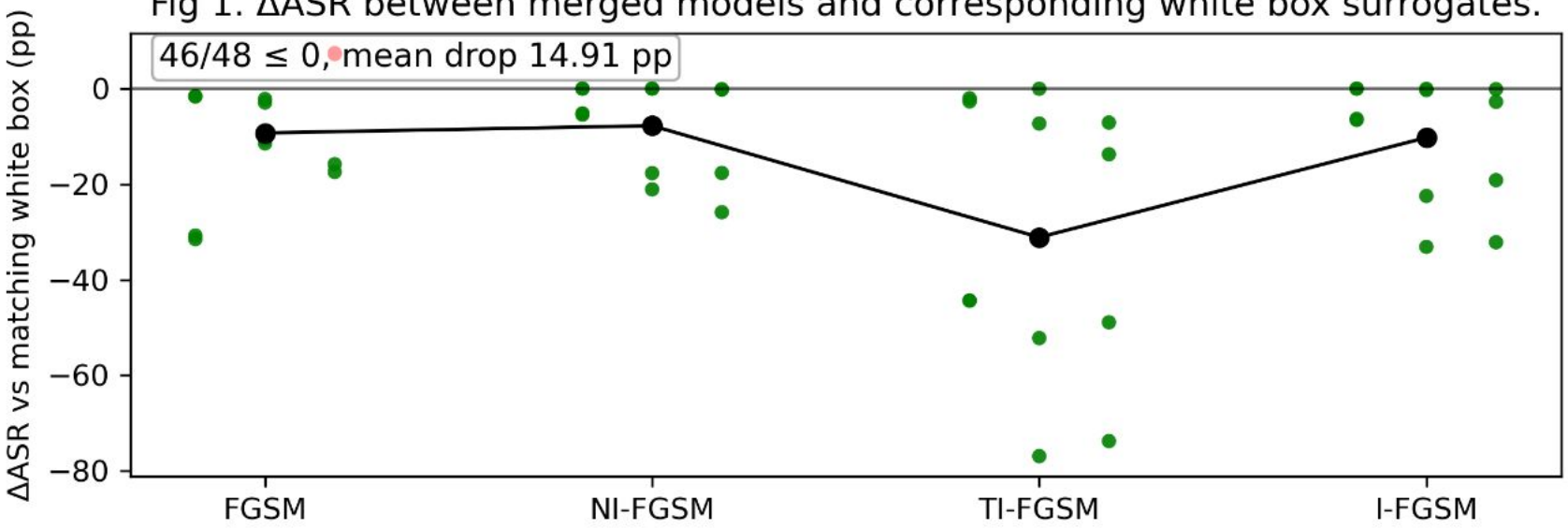
Attack	Surrogate ↓ / Target →	Cars				MNIST				EuroSAT			
		$\mathcal{P}$	$\mathcal{F}$	$\mathcal{M}_w$	$\mathcal{M}_s$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{M}_w$	$\mathcal{M}_s$	$\mathcal{P}$	$\mathcal{F}$	$\mathcal{M}_w$	$\mathcal{M}_s$
FGSM	$\mathcal{P}$	100.00	56.75	68.50	69.26	100.00	92.08	89.32	88.54	100.00	65.26	84.19	82.57
	$\mathcal{F}$	99.95	99.73	98.08	98.18	100.00	90.79	87.88	88.62	100.00	72.04	79.48	79.26
	$\mathcal{M}_w$	99.95	98.78	99.42	99.40	100.00	88.75	89.44	88.56	100.00	66.76	85.69	84.87
	$\mathcal{M}_s$	99.95	98.68	99.32	99.56	100.00	89.98	89.78	88.66	100.00	68.41	85.26	85.37
NI-FGSM	$\mathcal{P}$	100.00	88.09	94.55	94.83	100.00	21.97	78.92	82.31	100.00	69.41	82.35	74.13
	$\mathcal{F}$	99.93	100.00	100.00	100.00	99.97	100.00	100.00	100.00	100.00	100.00	99.81	99.89
	$\mathcal{M}_w$	99.61	100.00	100.00	100.00	99.80	98.77	100.00	100.00	100.00	87.96	100.00	100.00
	$\mathcal{M}_s$	99.69	100.00	100.00	100.00	100.00	97.76	100.00	100.00	100.00	93.89	100.00	100.00
TI-FGSM	$\mathcal{P}$	100.00	31.64	55.65	55.64	100.00	0.34	47.79	23.09	100.00	3.02	51.09	26.24
	$\mathcal{F}$	99.61	100.00	97.33	97.99	99.92	89.76	89.73	82.45	100.00	99.65	85.91	92.54
	$\mathcal{M}_w$	99.59	96.74	99.99	99.98	99.96	1.23	100.00	96.13	99.98	20.15	100.00	98.44
	$\mathcal{M}_s$	99.64	97.53	99.99	100.00	99.38	1.35	98.83	100.00	100.00	40.85	99.50	100.00
I-FGSM	$\mathcal{P}$	100.00	80.51	93.66	93.45	100.00	1.13	77.54	66.89	100.00	31.22	80.85	67.85
	$\mathcal{F}$	99.81	100.00	100.00	100.00	100.00	100.00	99.96	99.70	99.98	100.00	97.24	99.89
	$\mathcal{M}_w$	99.63	100.00	100.00	100.00	99.93	38.57	100.00	99.96	99.94	91.59	100.00	100.00
	$\mathcal{M}_s$	99.56	100.00	100.00	100.00	99.95	38.83	100.00	100.00	99.94	93.22	99.89	100.00

Results

Takeaway #1

- ASR on merged (target) models ≤ ASR on surrogate models in 46/48 cases.
  - In most cases, model merging shown to reduce success of attack.

Fig 1. ΔASR between merged models and corresponding white box surrogates.

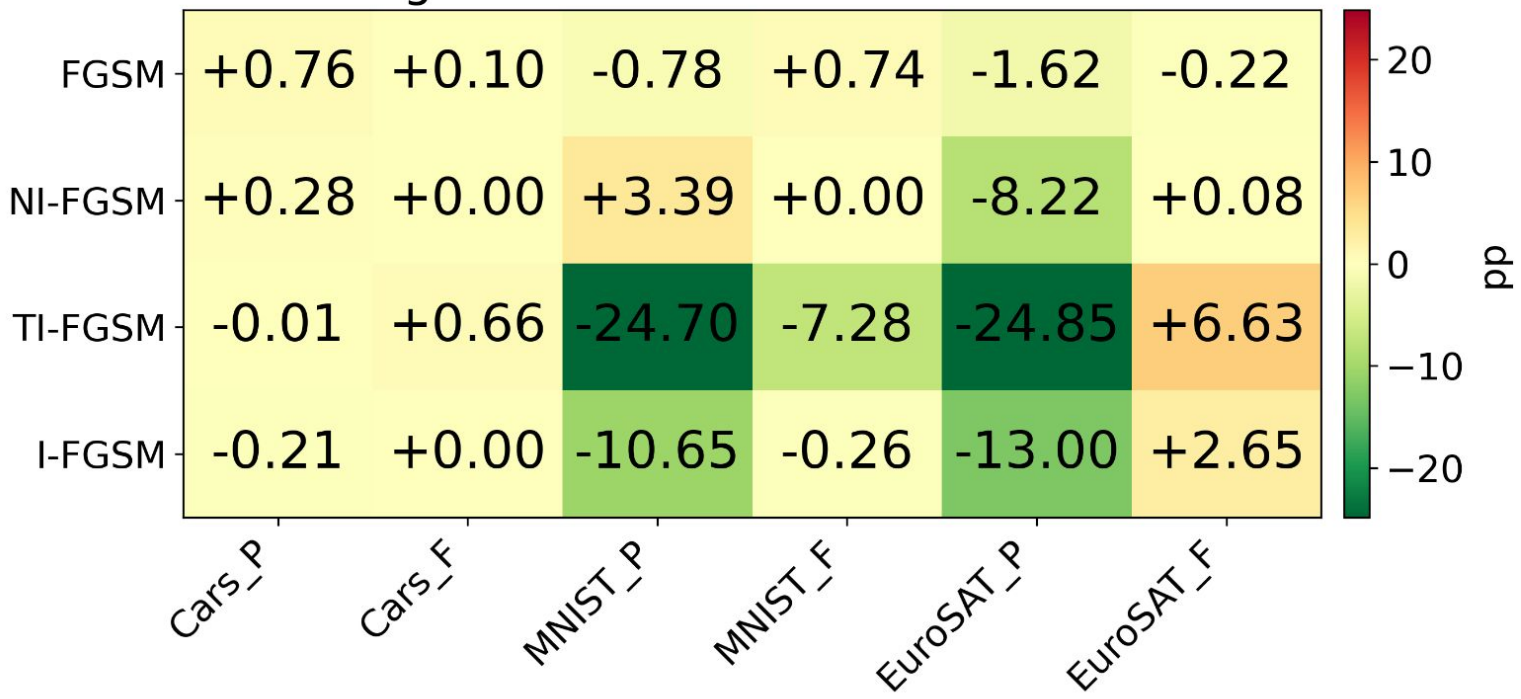


Results

Takeaway #2

- ASR decreases/remains same in 15/24 cases when a stronger merging method is used.
  - In in more than half cases, stronger model merging doesn't increase success of attack.

Fig 2. ΔASR between Ms and Mw.

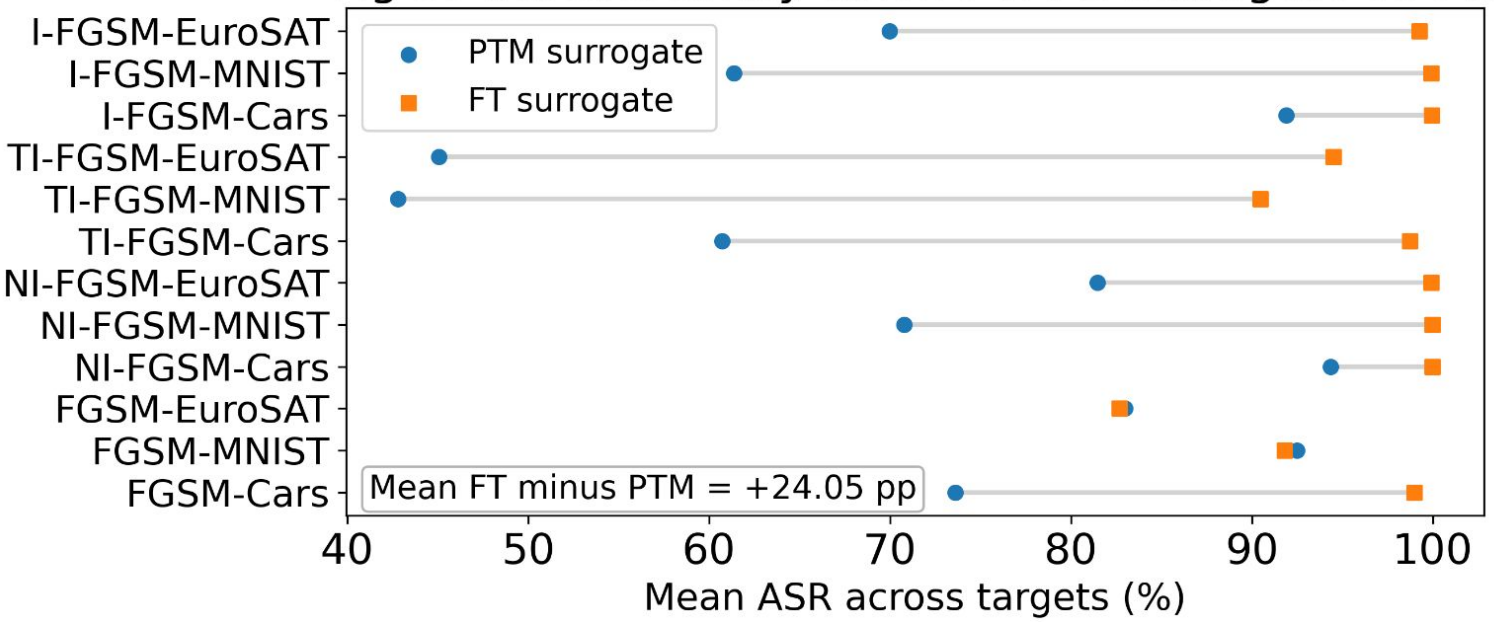


Results

Takeaway #3

- Pretrained model acts as a weaker surrogate relative to fine-tuned counterpart.
  - Merging an open-source fine-tuned model exposes you to threat of transfer attacks more.

Fig 3. Transferability when PTM is surrogate vs FT.



Conclusion and Future work

- The risk of transfer attack is real.
- Model merging could provide a “free lunch” of adversarial robustness.
- Theoretical analysis, statistical validation of the results.

Thank you!

Contact

gangwal@iiit.ac.in, aaryanajaysharma@gmail.com
https://sypy.iiit.ac.in/
A3-113, CSTAR, IIIT, Gachibowli, 500 032, Hyderabad, India