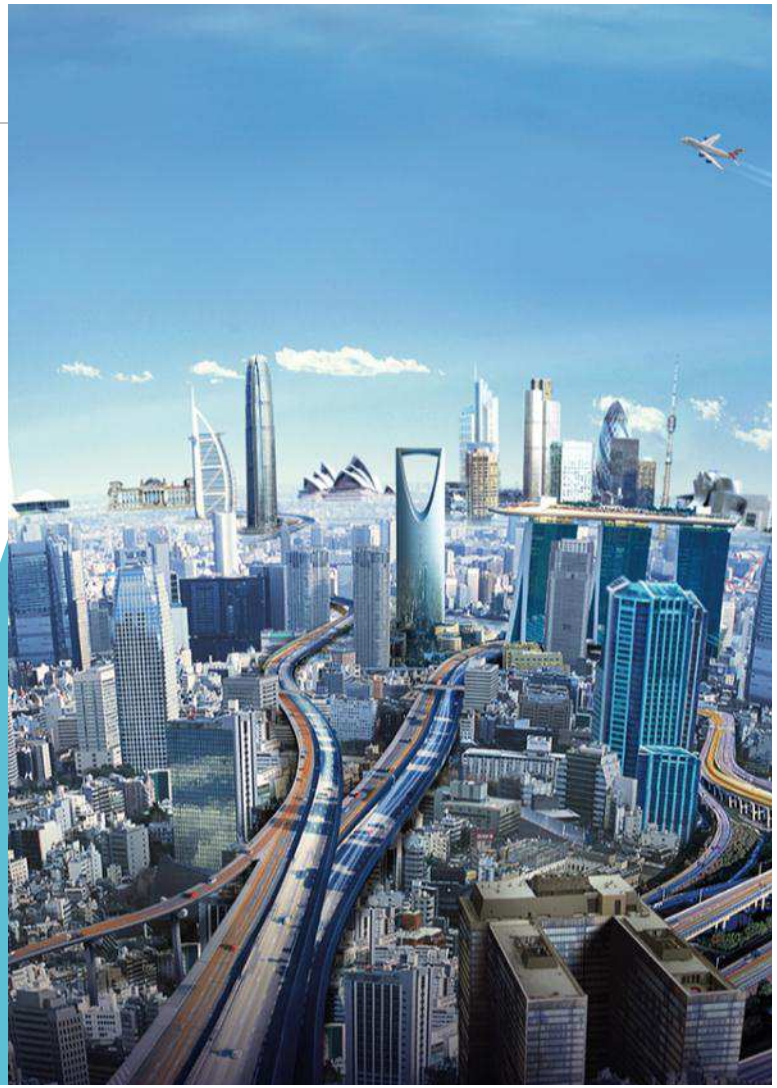
Evasion Attacks Transferability Between Machine Learning Models

BOUSSAD ADDAD
CO-AUTHORS : J. KODJABACHIAN, C. MEYER

THALES, PALAISEAU



Outline

Some recalls about machine learning (ML)

Crafting attacks on ML-based systems

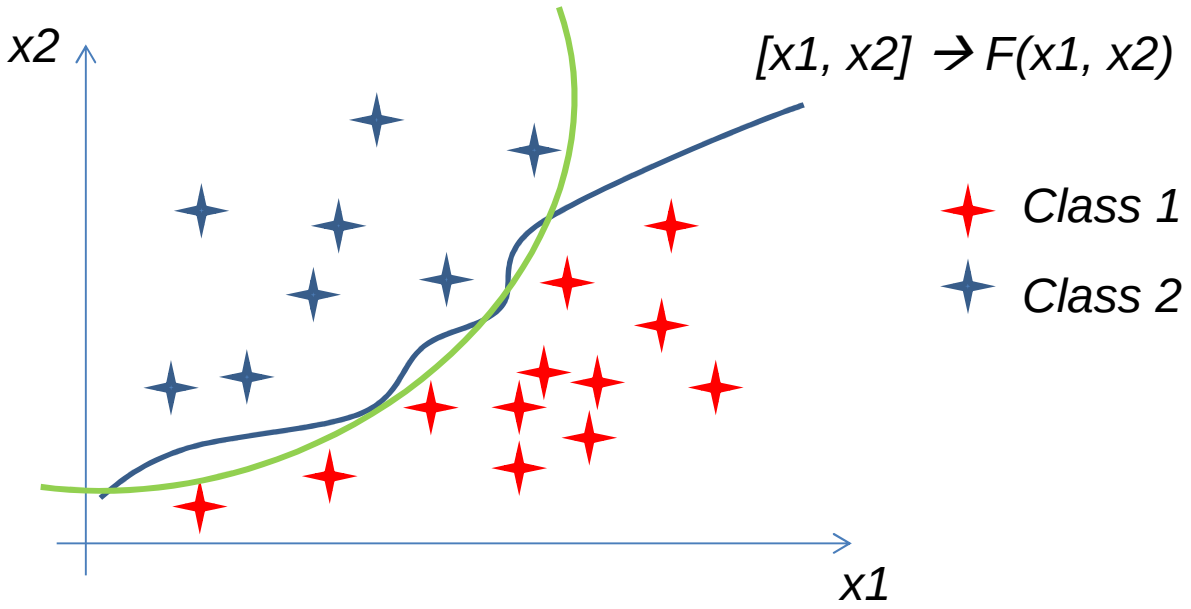
- Evasion attacks definition and examples
- Evasion attacks constraints

Transferability of attacks : principle and application to cyber security

- Comparison of two methods: CIA and iter-FGSM

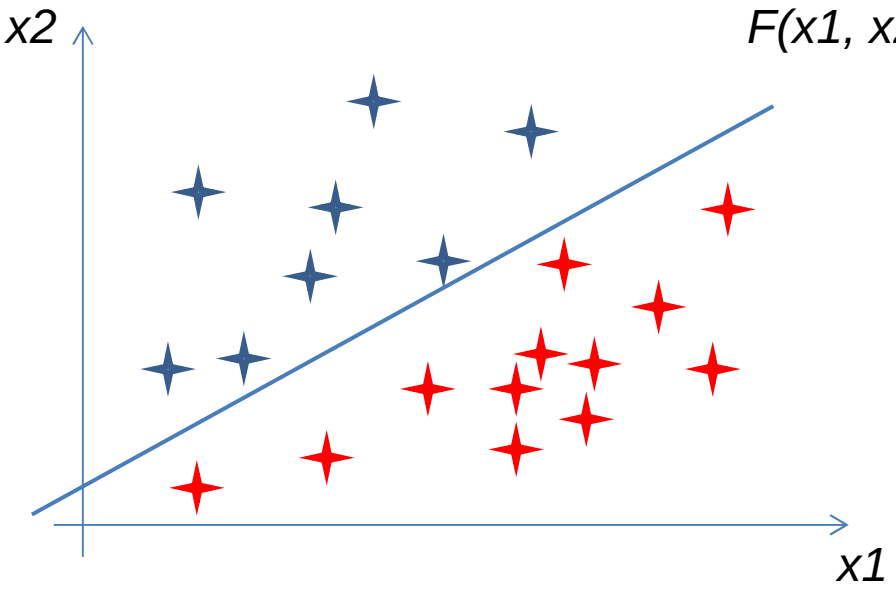
Conclusion

Recalls: machine learning objective



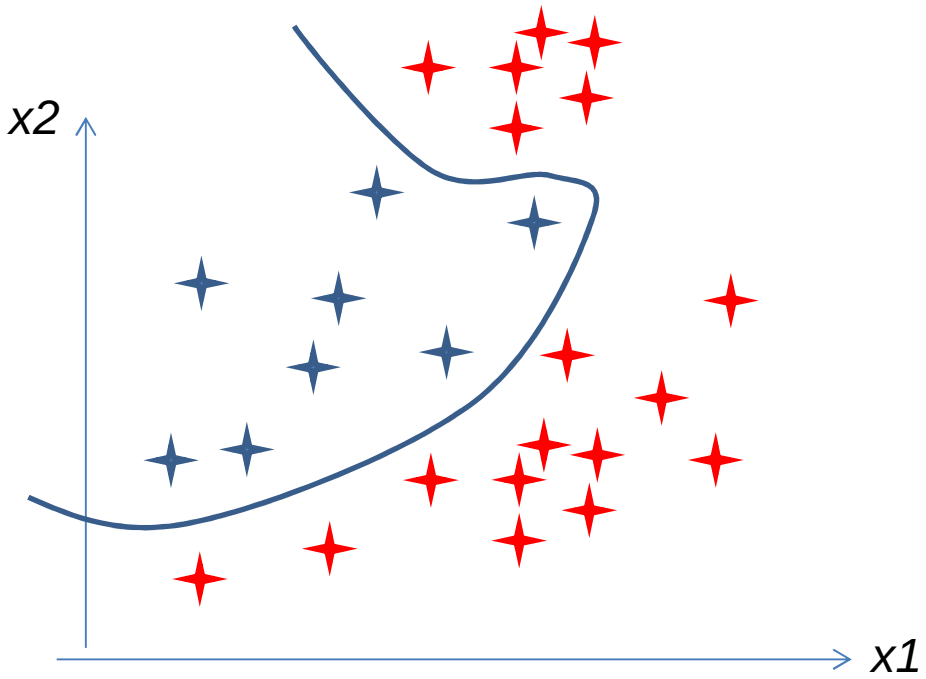
What is the best function F ?

Recalls: model choice



Sometimes, as simple as a linear function !

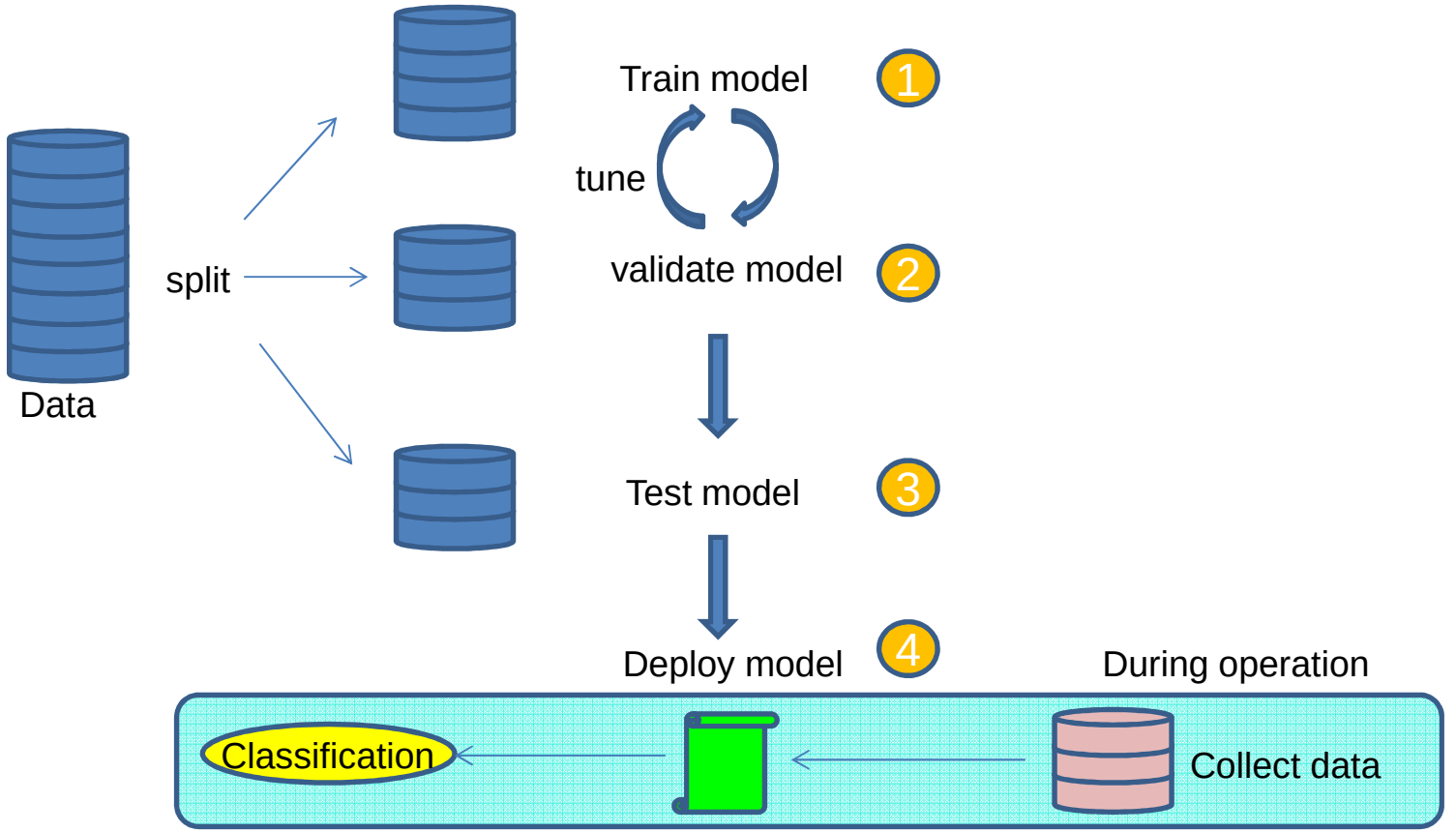
Recalls: model choice



Often much more complicated …

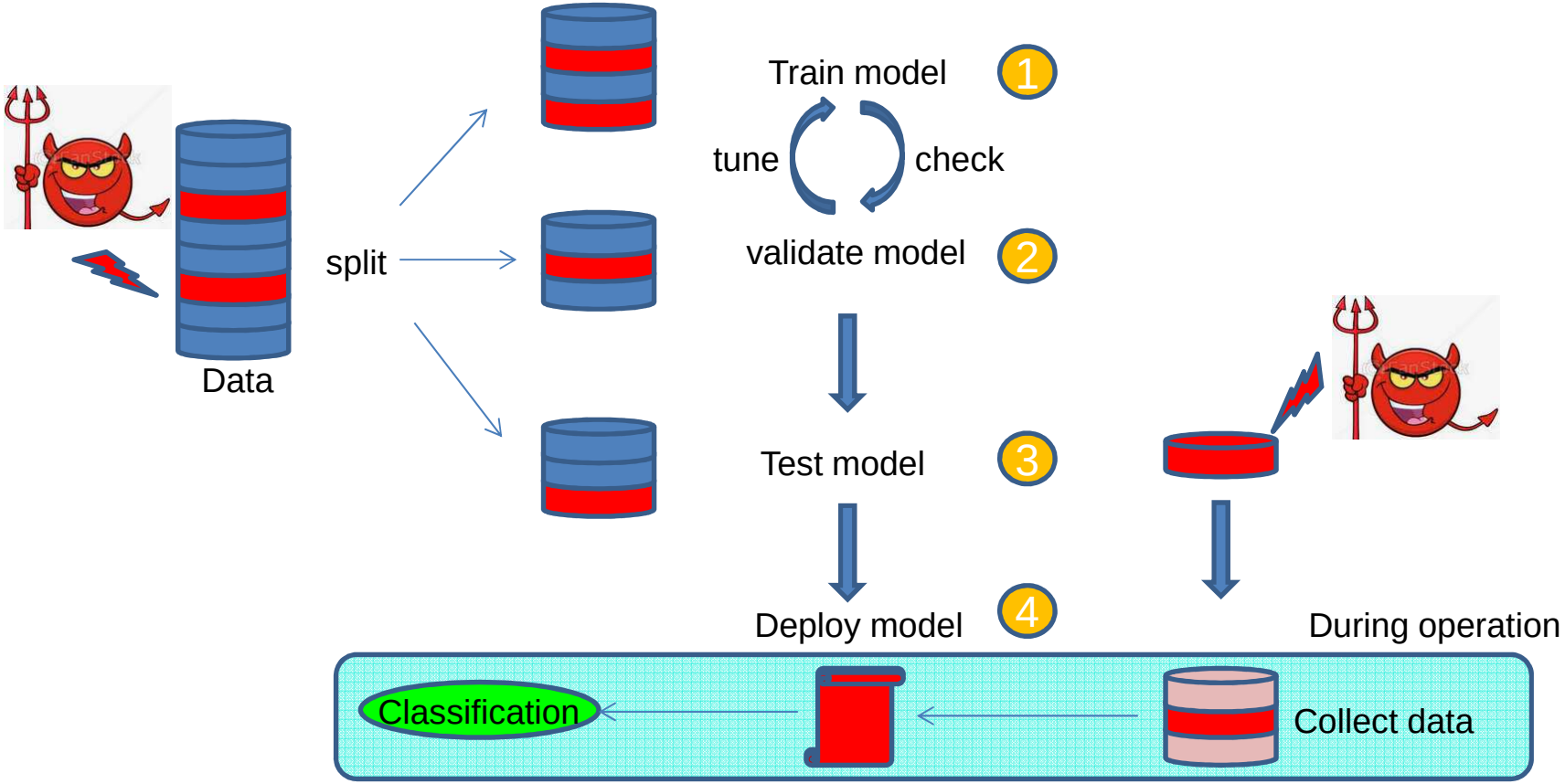
8bits - 100x100px RGB image → 30, 000 features → 30,000^256 possible images!

Recalls: learning principle and steps

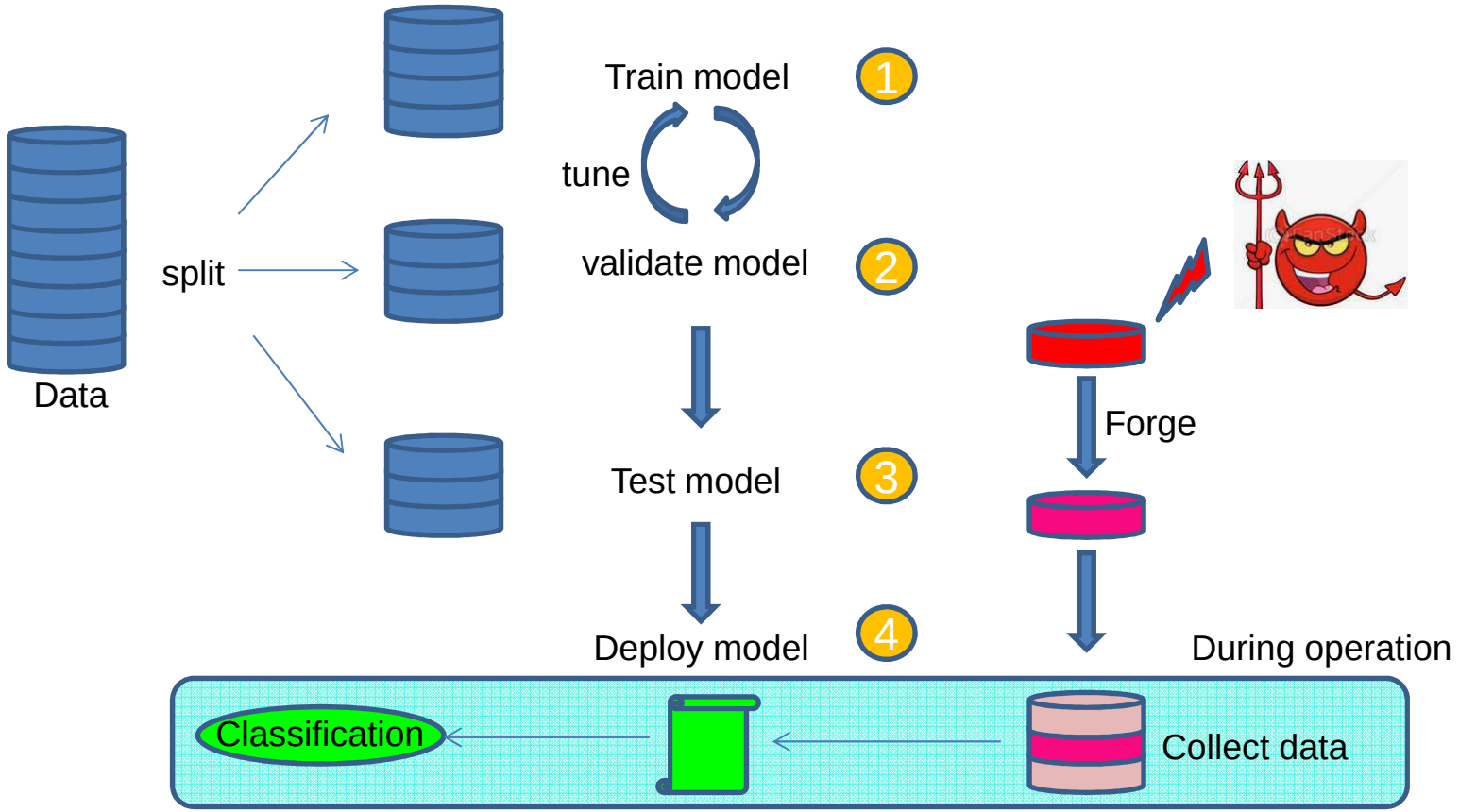


ML attacks principle

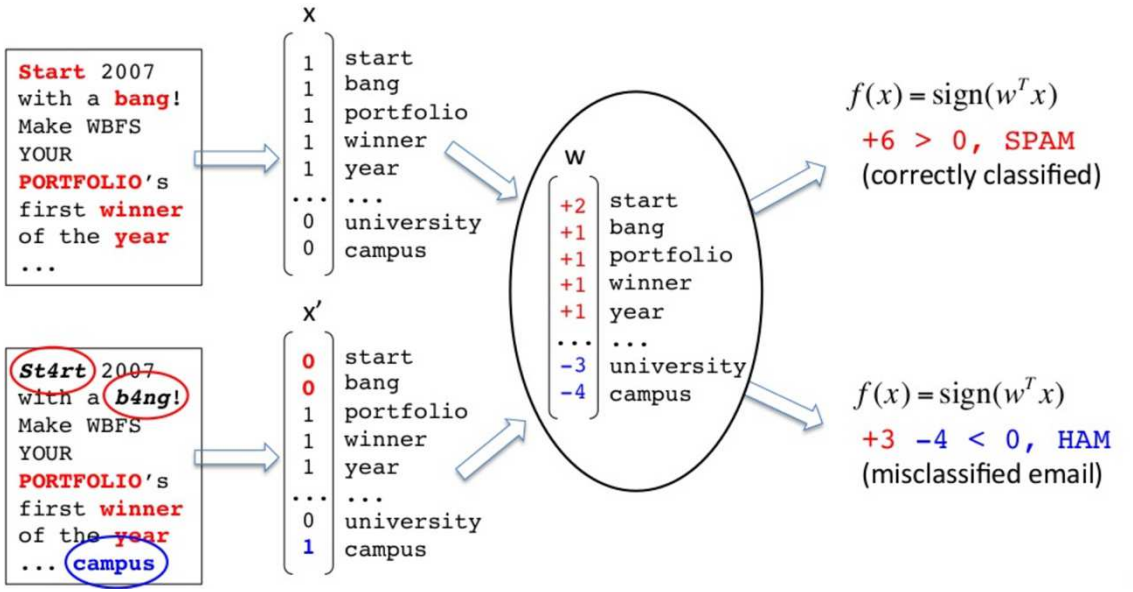
Model poisoning attack



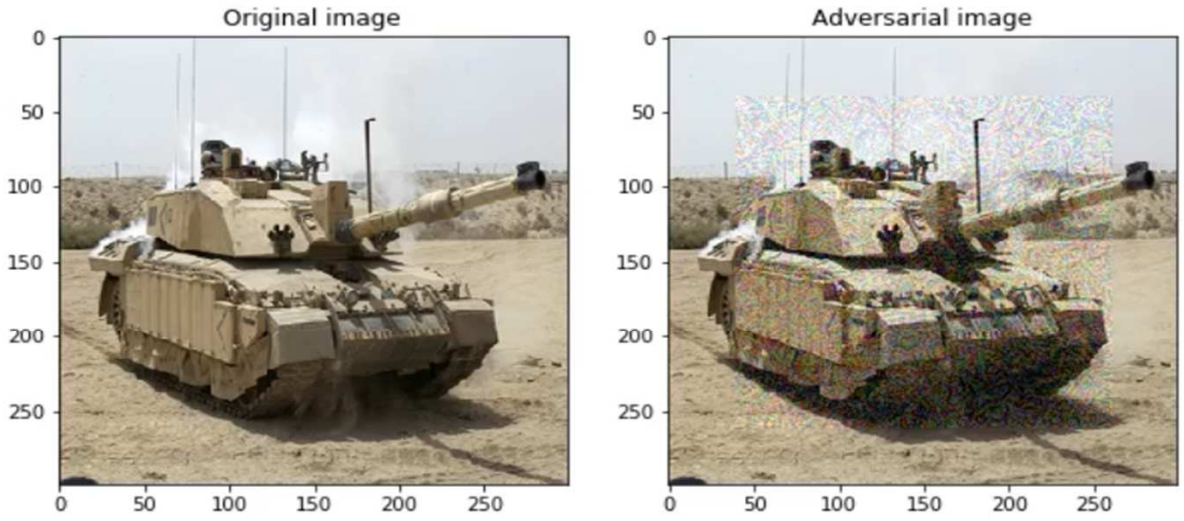
Model evasion attack



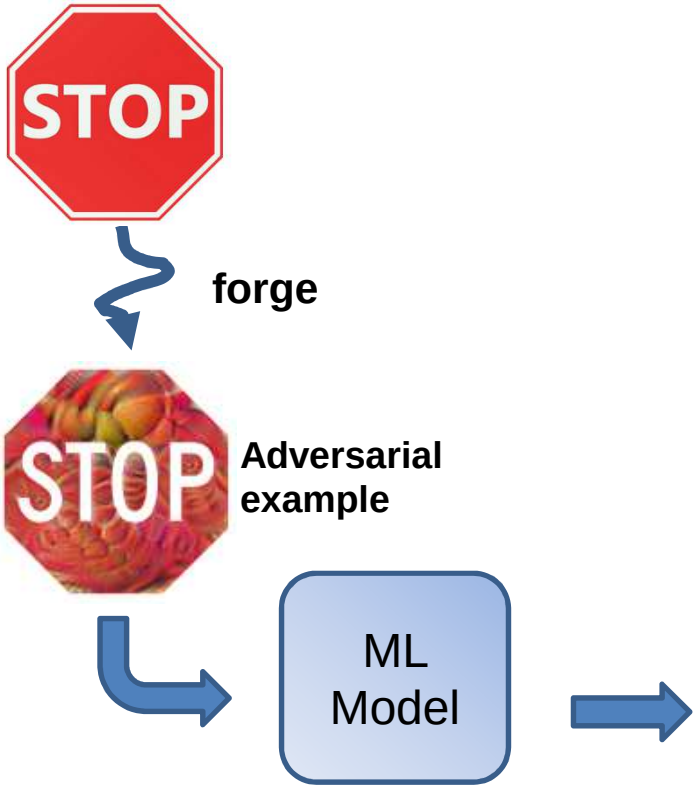
Model evasion attacks examples: **spam detector**



Model evasion attacks examples: **computer vision**



Model evasion attacks examples: **self-driving vehicles**



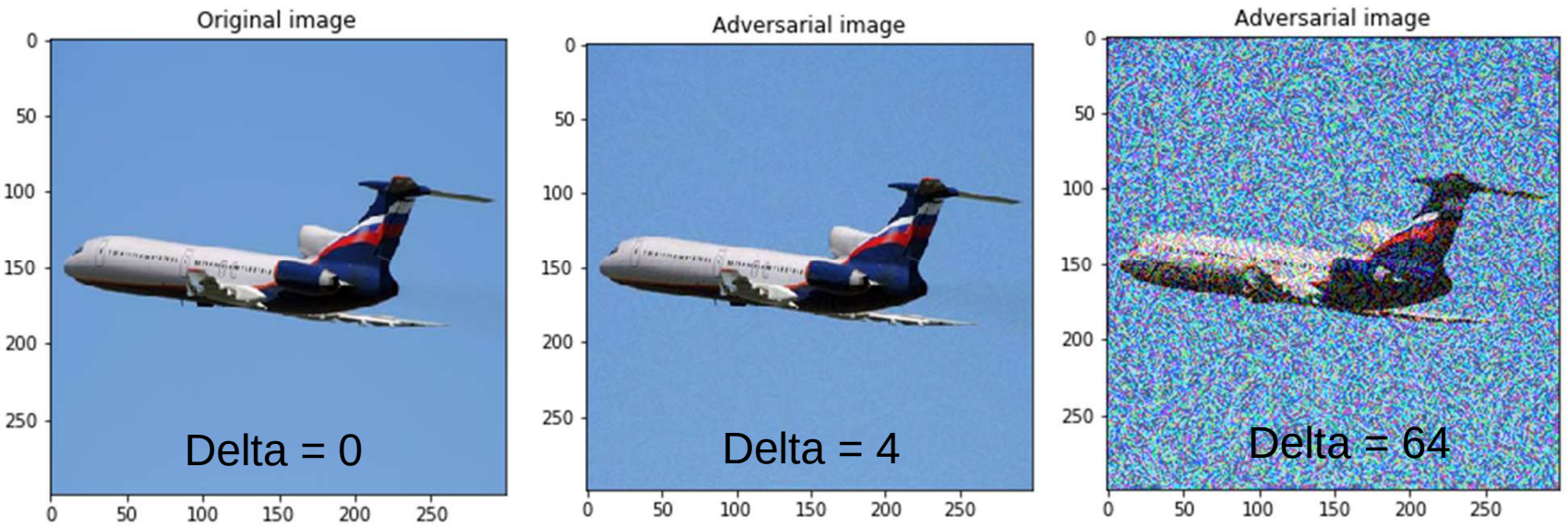
[2]

[2] Shang-Tse Chen et al, ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector, 2018

Model evasion attacks constraints

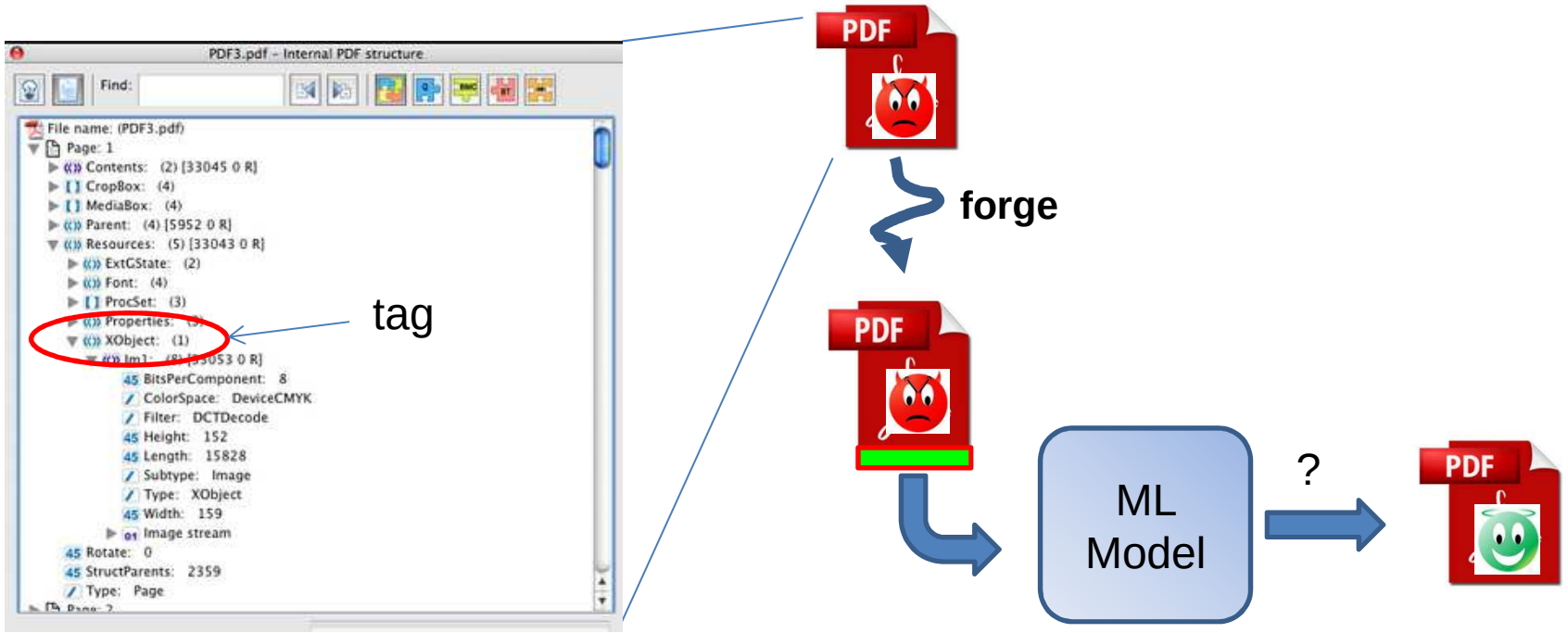
Limit the perturbation and the output

1 Byte image encoding → [0, 255]



Constrain the direction of modification(Malware detection) [2]

An adversary forces a **PDF detector** trained with ML to make wrong predictions by ***adding*** some tags to classified files[1].



[1] Weilin Xu, Yanjun Qi, and David Evans. Automatically Evading Classifiers A Case Study on PDF Malware Classifiers. Network and Distributed Systems Symposium 2016, 21-24 February 2016,

Some evasion attacks techniques

FGSM (Fast Gradient Sign Method) [3]

- Minimize the probability of the real class

Carlini & Wagner attacks [4]

- Minimize the probability of the real class and the perturbation norm

CIA (Centered initial attacks) [5]

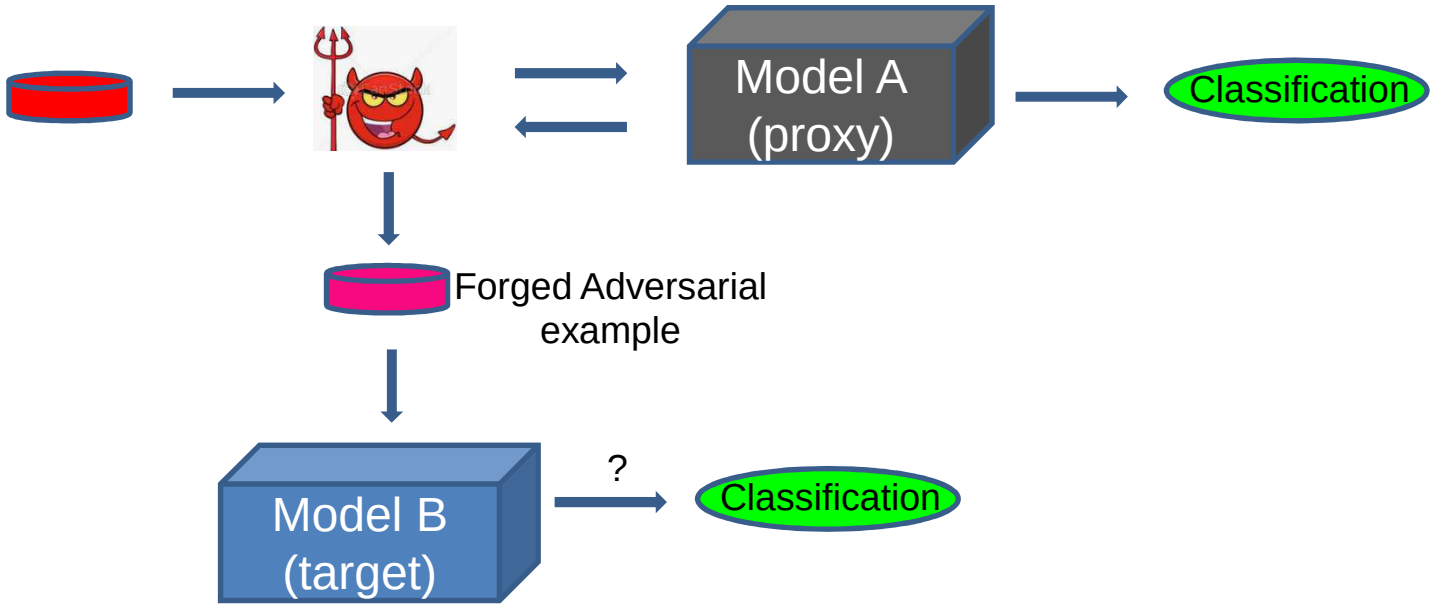
- Minimize the probability of the real class and guarantee the norm of the perturbation to be smaller than a given threshold

[3] Christian Szegedy et al, Intriguing properties of neural networks, https://arxiv.org/abs/1312.6199, 2013
[4] Nicholas Carlini, David Wagner, Towards Evaluating the Robustness of Neural Networks, https://arxiv.org/abs/1608.04644, 2017
[5] B. Addad, J. Kodjabachian and C. Meyer, Clipping free attacks against Artificial Neural Networks, https://arxiv.org/abs/1803.09468, 2017

Model evasion attacks transferability

Transferability between ML models principle [6]

Transferability property: Samples crafted to mislead model A are likely to mislead model B



[6] Yanpei Liu et al, Delving into Transferable Adversarial Examples and Black-box Attacks, ICLR 2017

Evasion attacks transferability use case : PDF malware classification

Experiment set up :

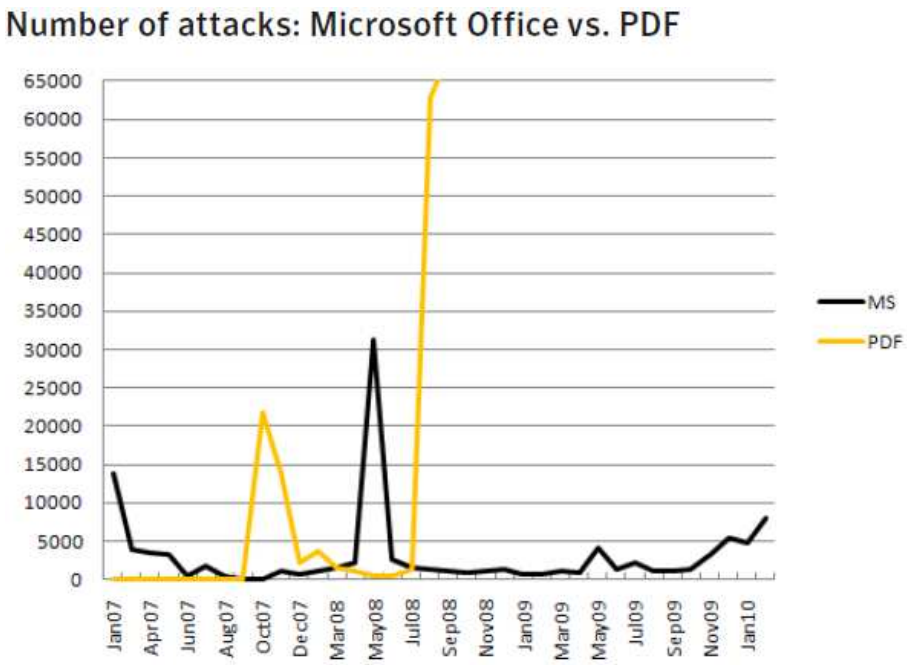
- Contagio dataset :

~ 10,000 malicious PDF files
~ 10,000 benign PDF files

- Proxy model (PDF malware classifier):
  - ANN (Artificial Neural Network)

- Targeted models (PDF malware classifiers)
  - SVM
  - Random forest
  - ANN

- All runs are repeated ten times



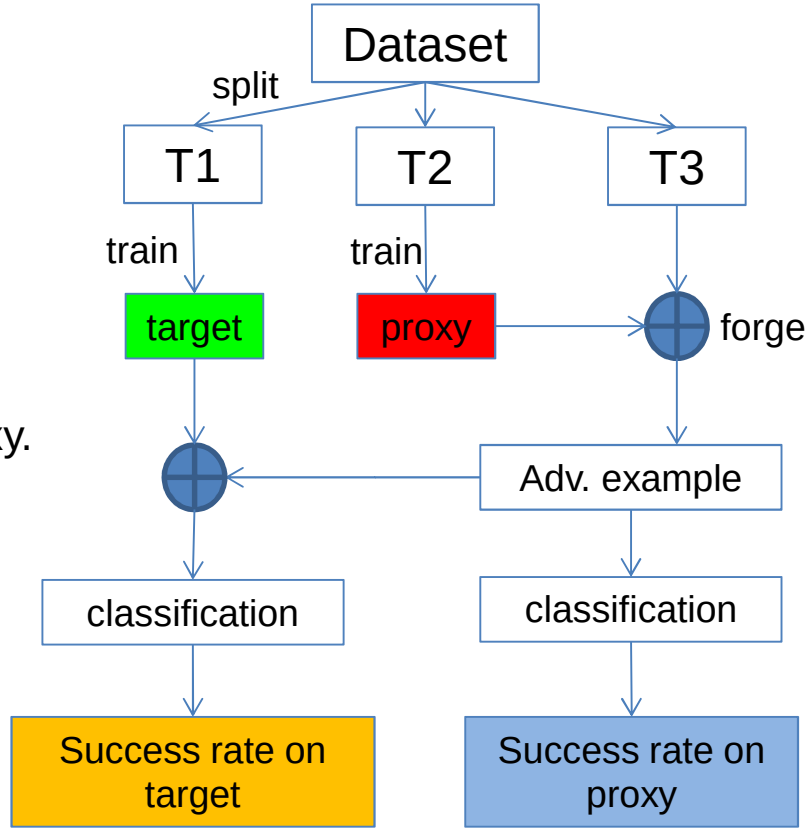
Symantec Report [7]

[7] Karthik Selvaraj and Nino Fred Gutierrez, The rise of PDF malware, Symantec Security Response, 2009

Attack transfer evaluation steps

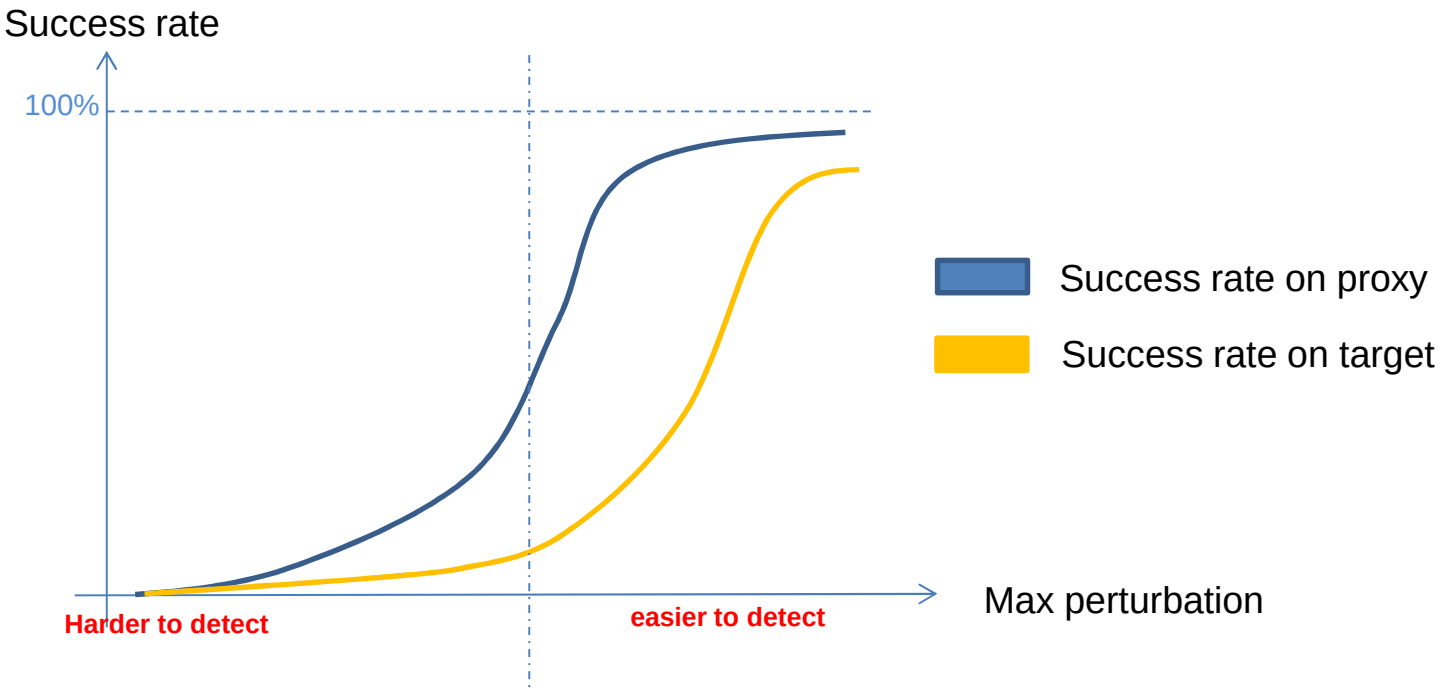
Steps :

1. Split dataset to three sets T1, T2 and T3
2. Train victim model (SVM) on T1
3. Train attacker model (ANN) on T2
4. Use files from T3 to forge attacks using proxy.
5. Send these attacks to victim.

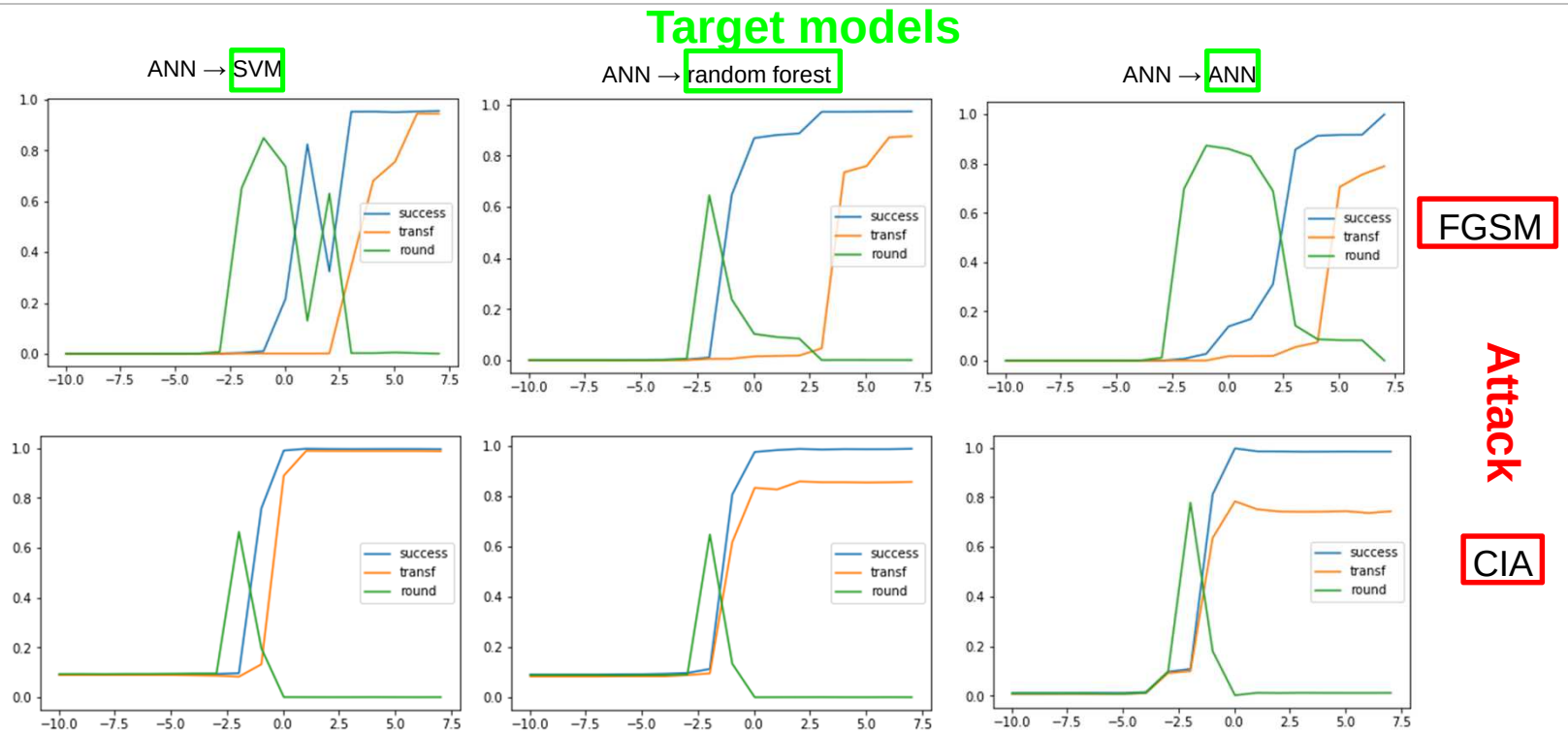


attack transfer evaluation

Curves: attacks success rate w.r.t perturbation size



Results: comparison between two techniques (CIA and iter-FGSM)



→ With CIA, we get Higher transferability rates (orange curve) with less perturbation (x coordinate)

Current and future work

Explore other applications

- Malware detection in other files types (MS)

Improve attacks

- Take into account other constraints related to data nature (ex : rounding)

Design more robust classifiers

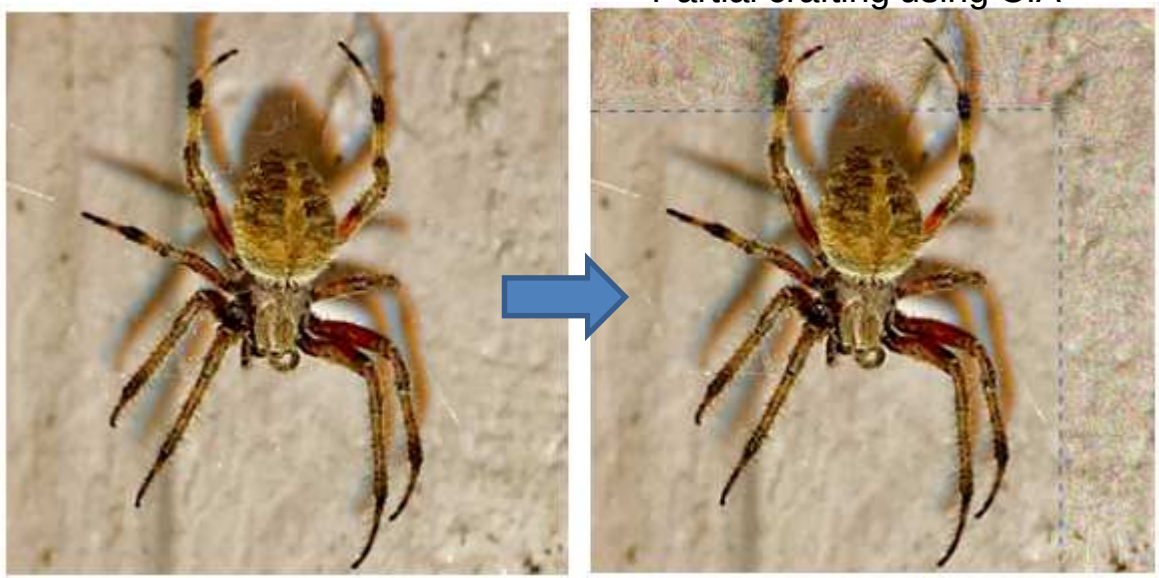
- Defend better against adversarial examples

Models certification

- Provide formal proofs about models robustness

Questions ?

Partial crafting using CIA



99,99 % dog !!!