THALES

Panorama des Attaques sur Réseaux de Neurones

Quentin Le Roux
Yannick Teglia
@ Thales DIS, Cybersecurity Hardware Lab
& INRIA (partenariat)

Conférence AMUSEC – 25 et 26 Avril 2023

Intelligence Artificielle et Cybersecurity Hardware Lab

Thales Digital Identity and Security (DIS)

- Produits et services sécurisés
- Solutions de paiement, Identité & biométrie, Internet des objets (IoT), téléphonie mobile…

Cybersecurity Hardware Lab

- Garantir la sécurité des produits Thales DIS
- Domaines de compétences:
  - Attaques par canaux auxiliaires (« side-channel »), injection de fautes
  - Attaques micro-architecturales
  - Architectures matérielles sécurisées
  - Cryptographie appliquée
  - Intelligence Artificielle pour l'analyse de sécurité et sécurité de l'Intelligence Artificielle

Thèse en cours avec INRIA, Rennes

- Intégrité des réseaux de neurones face aux portes dérobées et attaques adverses

Les Réseaux de Neurones

- **Quoi ?** Programme informatique prédictif (classification, détection, etc.)
- **Comment ?** Ensemble structuré d'additions, multiplications et de fonctions non-linéaires dont les coefficients sont appris par rétropropagation
- **Exemple**:

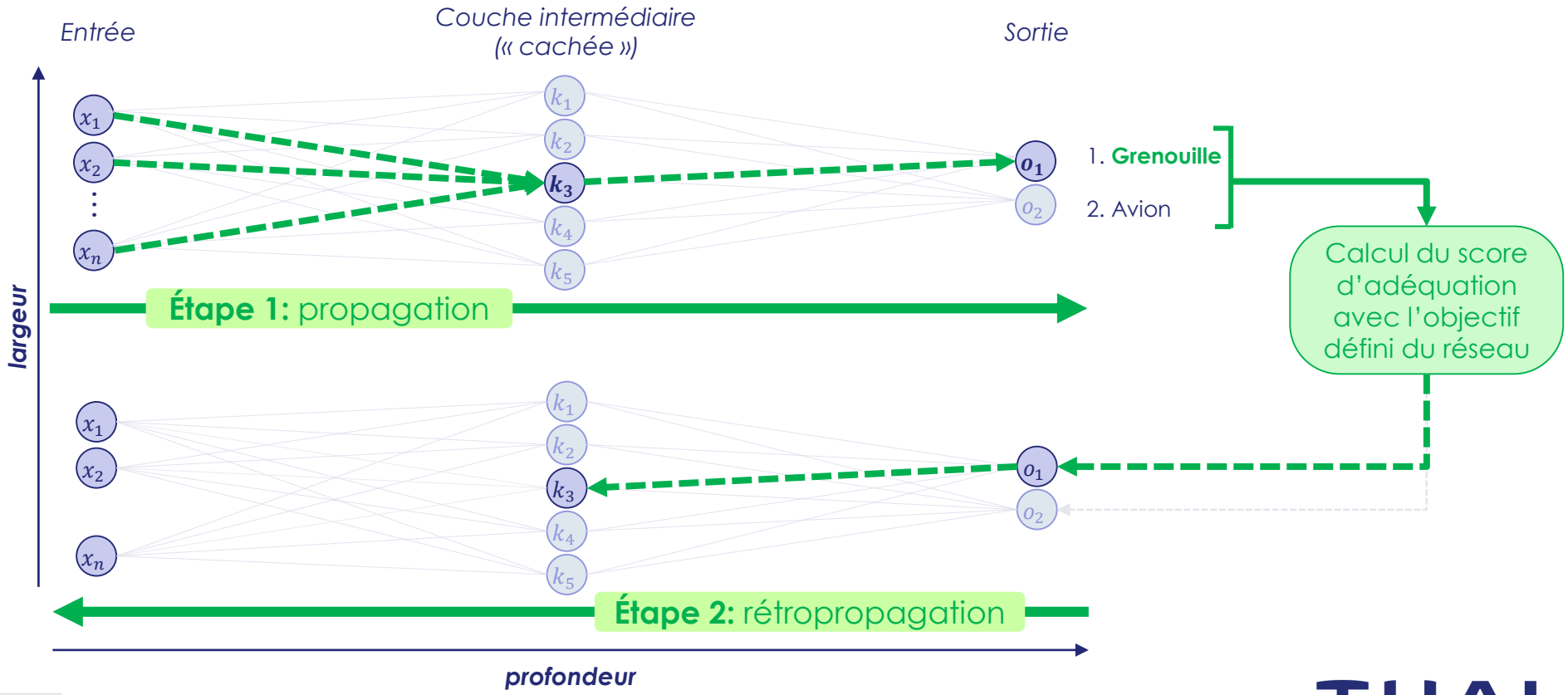


Image source: dataset CIFAR10

THALES

Les Réseaux de Neurones

- **Quoi ?** Programme informatique prédictif (classification, détection, etc.)
- **Comment ?** Ensemble structuré d'additions, multiplications et de fonctions non-linéaires dont les coefficients sont appris par rétropropagation
- **Exemple**:

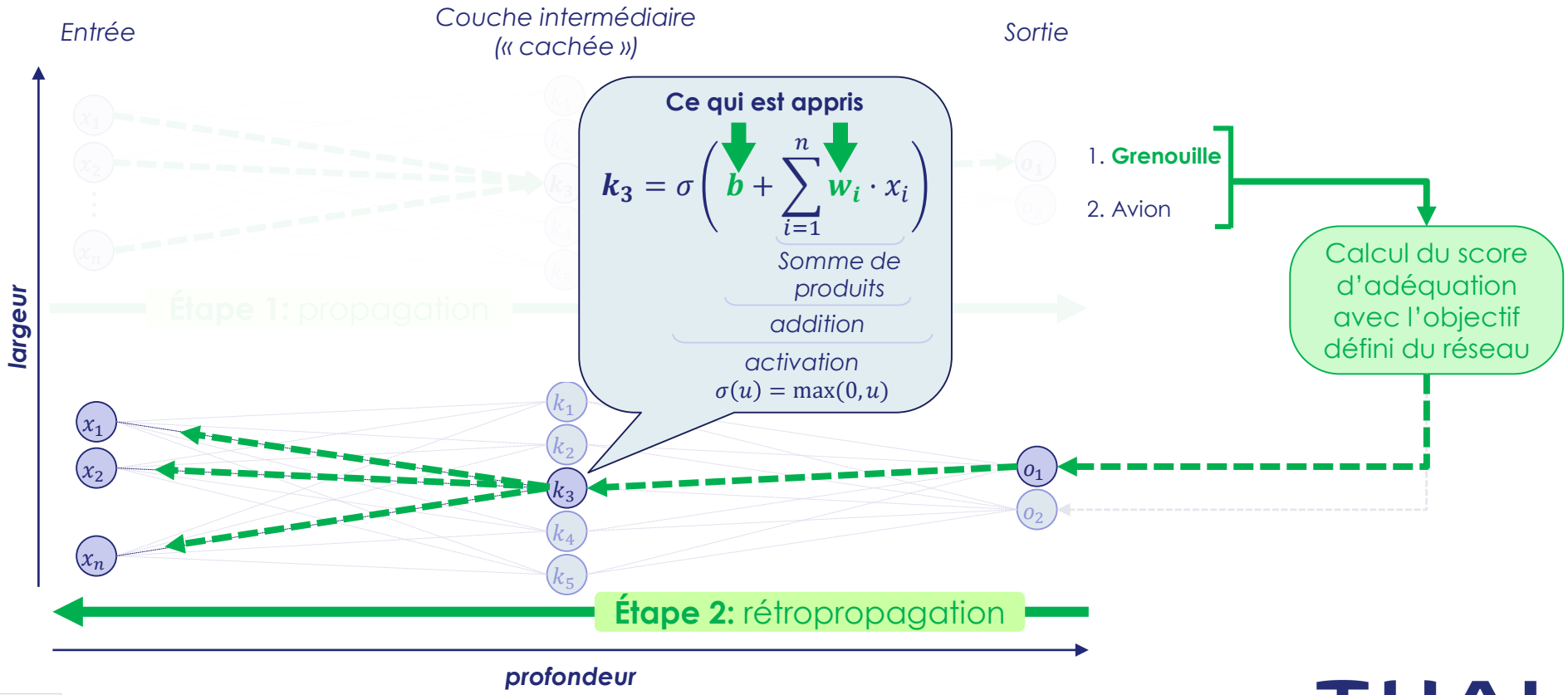
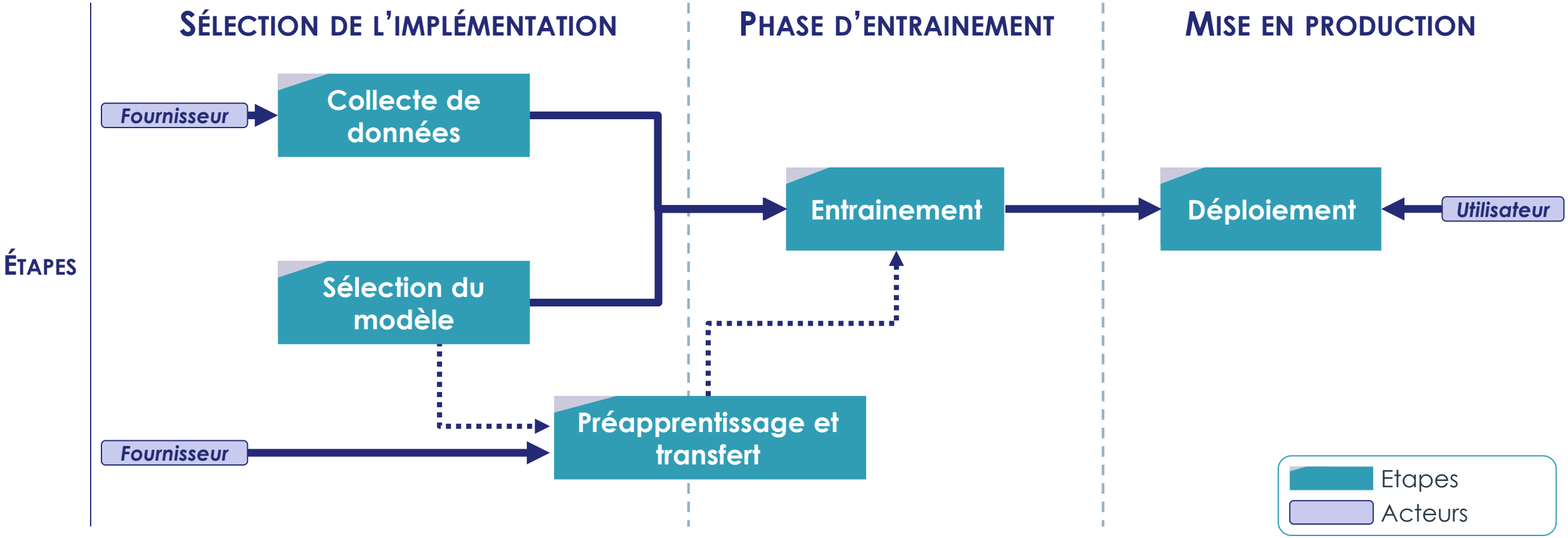
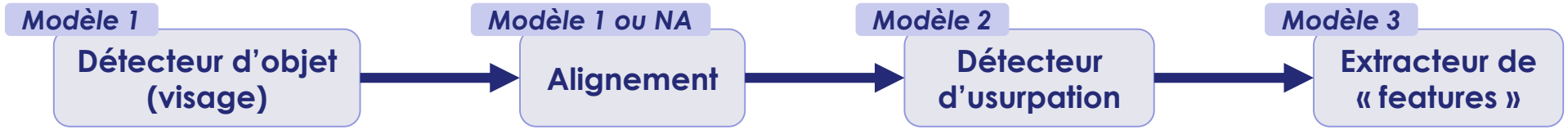


Image source: dataset CIFAR10

Cycle de vie d'un réseau de neurones



Pipeline en Reconnaissance Faciale



THALES

Risques associés au long du cycle de vie (1/3)

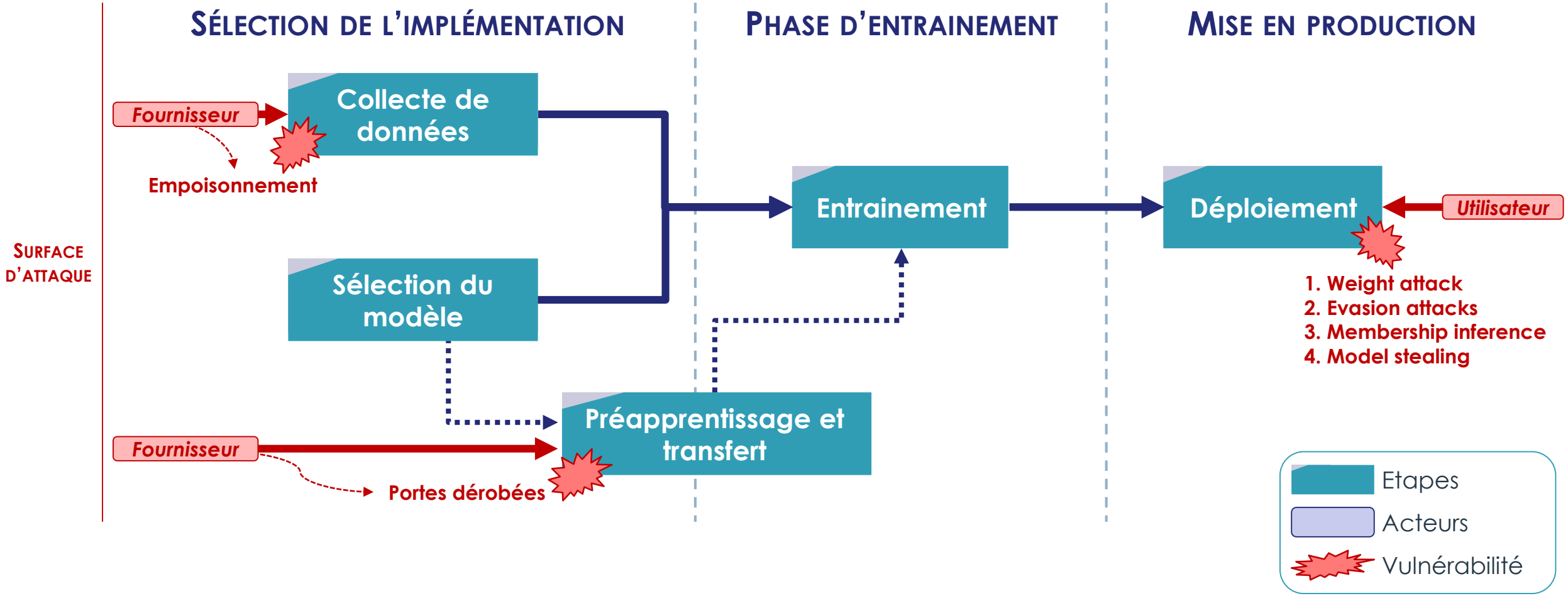


Image sources: *"Backdoor Attacks Against Deep Learning Systems in the Physical World," Wenger et al. (2021)*
"Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," Chen et al. (2017)

THALES

Risques associés au long du cycle de vie (2/3)

SÉLECTION DE L'IMPLÉMENTATION

PHASE D'ENTRAINEMENT

MISE EN PRODUCTION

SURFACE D'ATTAQUE

Fournisseur

Collecte de données

Empoisonnement

Clean Data

Poison Data

Sélection du modèle

Fournisseur

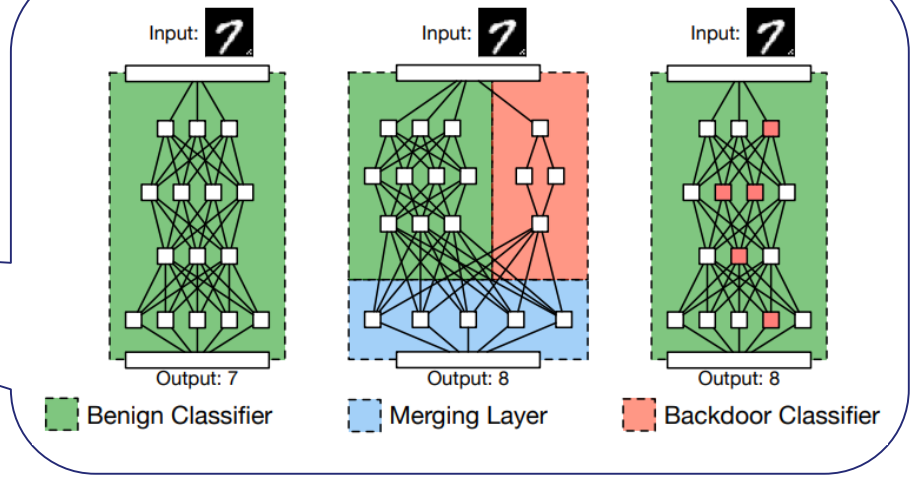
Préapprentissage et transfert

Portes dérobées

éploiement

Utilisateur

1. Weight attack
2. Evasion attacks
3. Membership inference



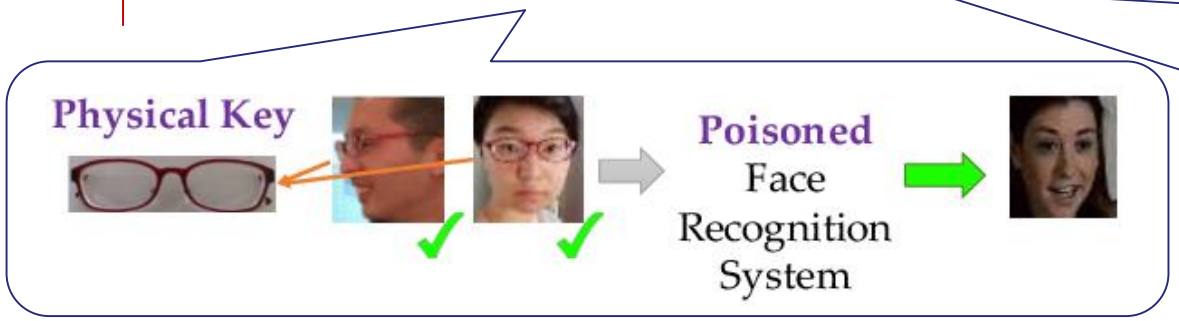


Image sources: *"Backdoor Attacks Against Deep Learning Systems in the Physical World," Wenger et al. (2021)*
"Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," Chen et al. (2017)

THALES

Risques associés au long du cycle de vie (1/3)

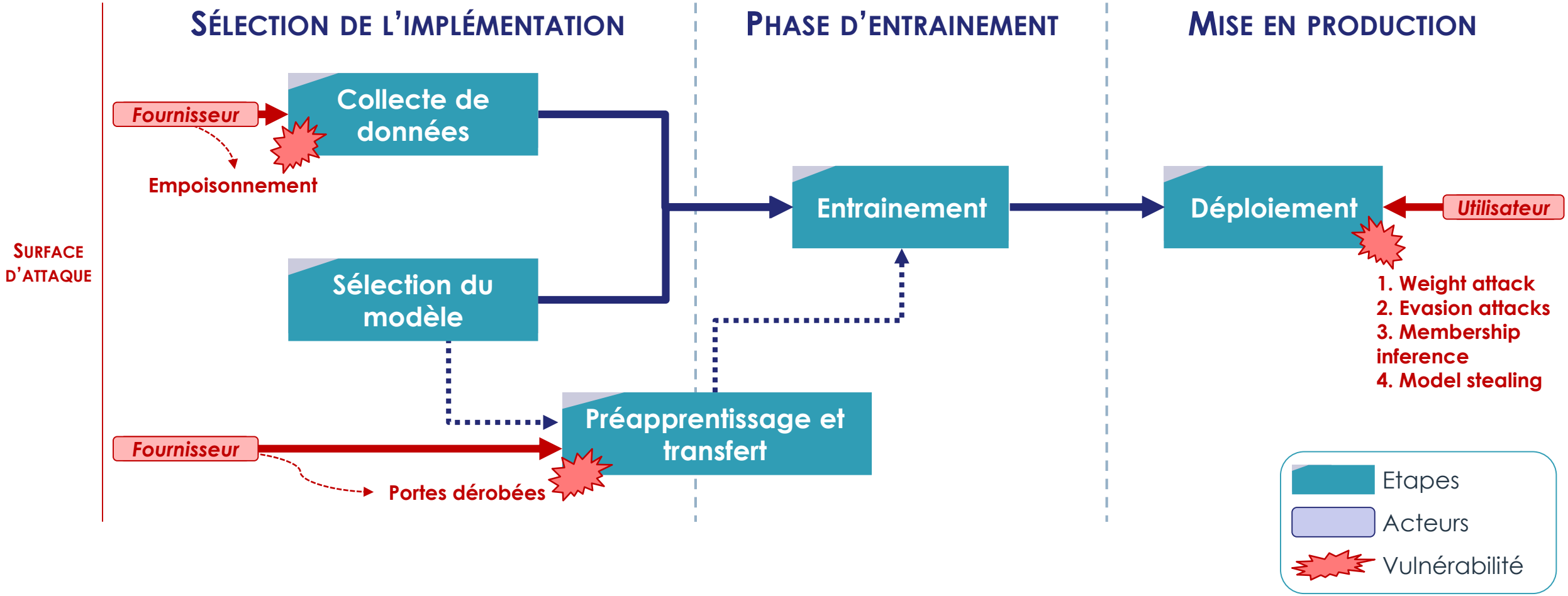


Image sources: *"Backdoor Attacks Against Deep Learning Systems in the Physical World," Wenger et al. (2021)*
"Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," Chen et al. (2017)

THALES

Risques associés au long du cycle de vie (3/3)

SURFACE D'ATTAQUE

SÉLECTION DE L'IMPLÉ

Providers

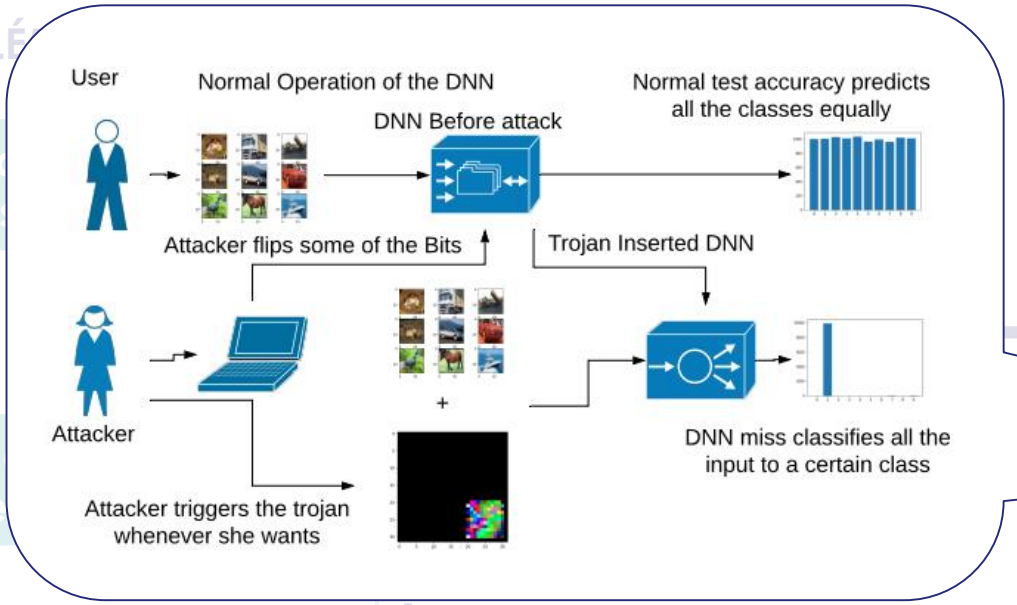
Collecte d données

Empoisonnement

Sélection modèle

Providers

Portes dérobées



MISE EN PRODUCTION

Déploiement

Utilisateur

1. Weight attack
2. Evasion attacks
3. Membership inference
4. Model stealing

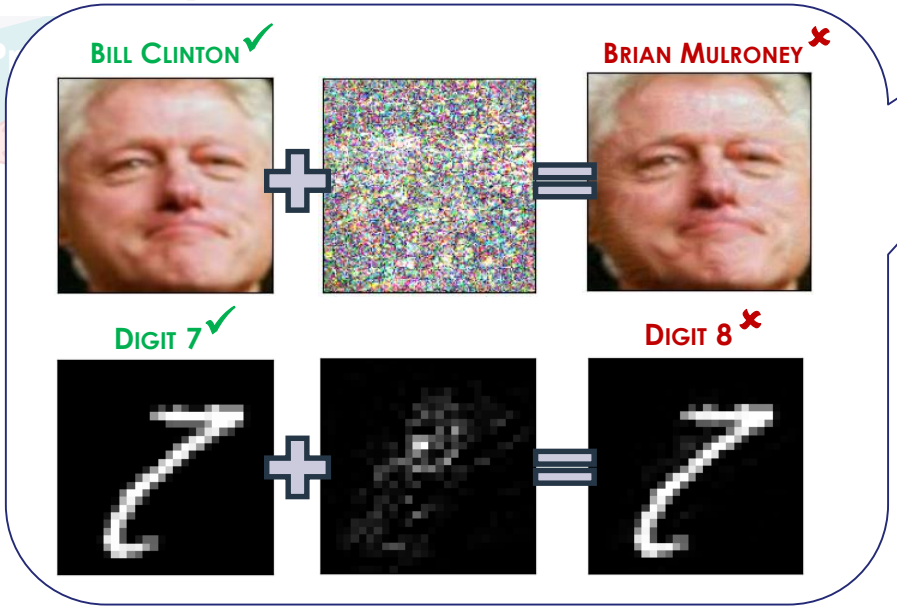


Image sources: MNIST, LFW datasets

THALES

Exemple: construire une attaque adverse

- Basé sur le cadre de la formulation de l' « Empirical Risk Minimization »:

$$\theta^* = arg\min_{\theta \in \Theta} \mathbb{E}_{\{x,y\}\sim D}\left[L(f_\theta(x), y; \theta)\right] \quad \textbf{(ERM)}$$

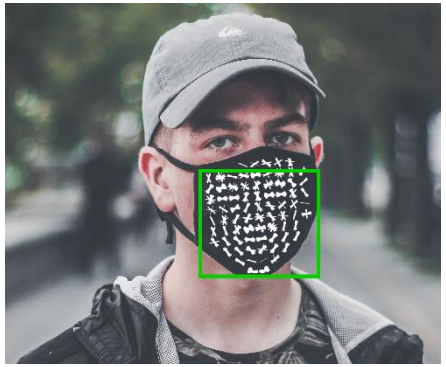
$$\delta = arg\min_{z \in \Delta} f(x+z) = y' \neq f(x) \quad \textbf{(méthode avec classe ciblée)}$$

- Un agent malicieux peut construire un exemple adverse avec une multitude de méthodes (boites noires, boites blanches, méthodes apprises, algorithmes génétiques, etc.)

Image source: dataset LFW

Cycle de vie d'un réseau de neurones

Pipeline en Reconnaissance Faciale



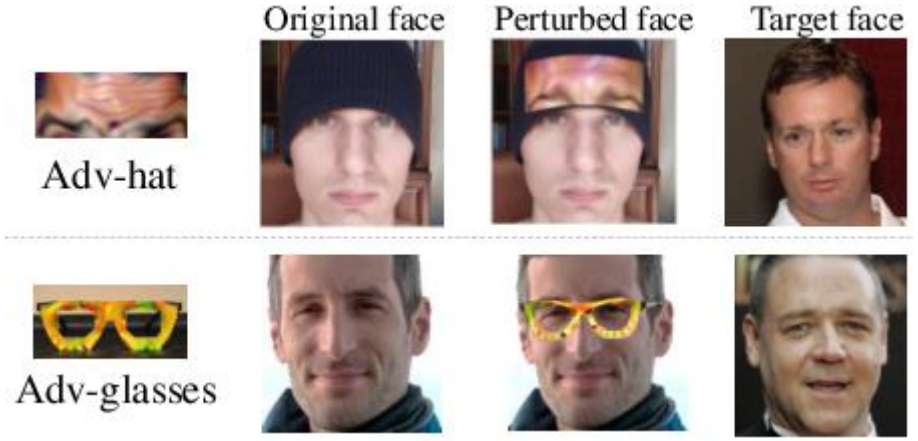


Image sources: BuceMacD/Adversarial-Faces; Presentation attack detection: Dissolving the myth of easy-to-fake face recognition, BioID, Ann-Kathrin Freiberg (2019), "Meaningful Adversarial Stickers for Face Recognition in Physical World," Guo et al. (2021)

Risques associés au long du cycle de vie (1/3)

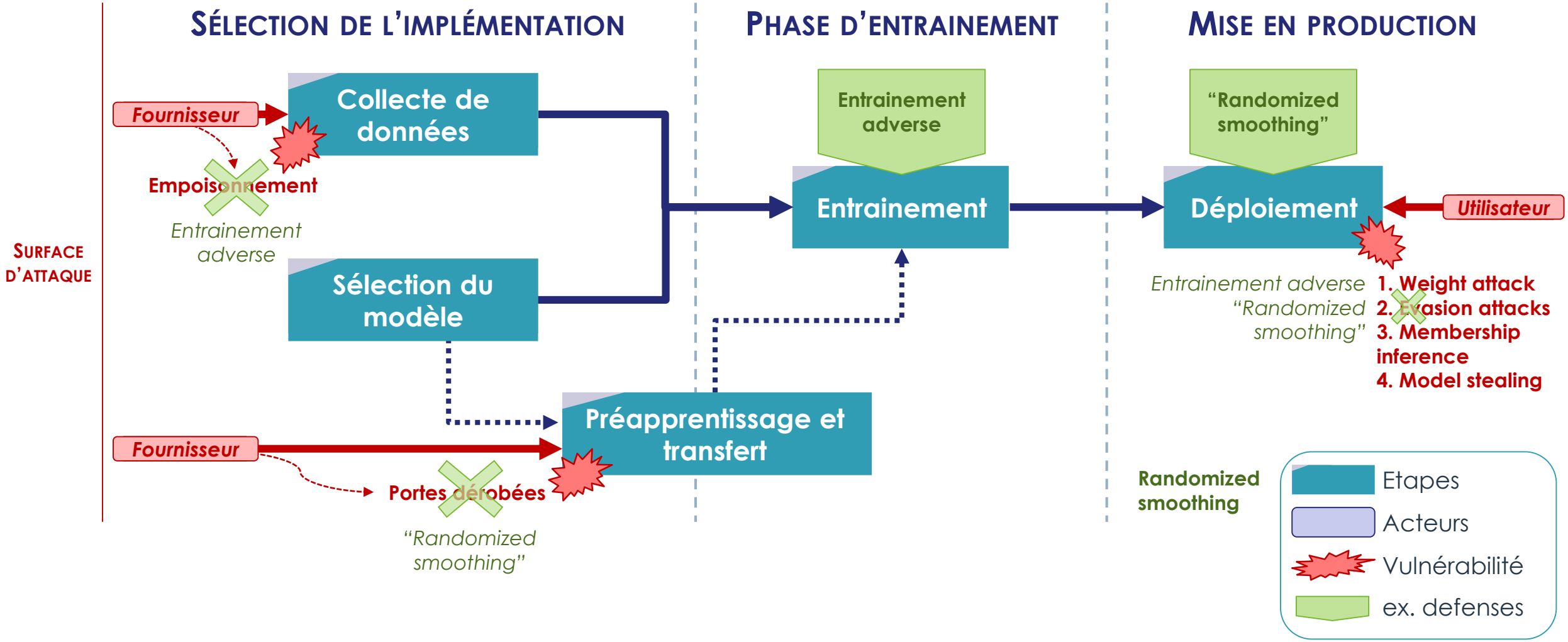
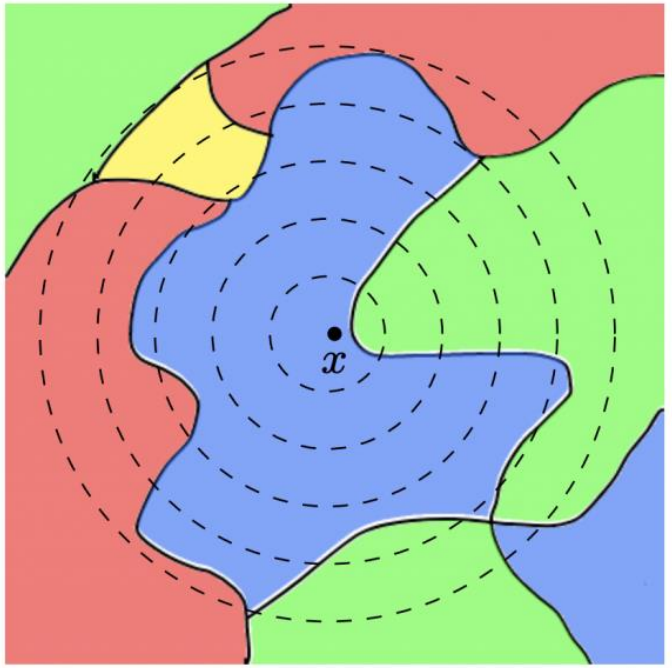


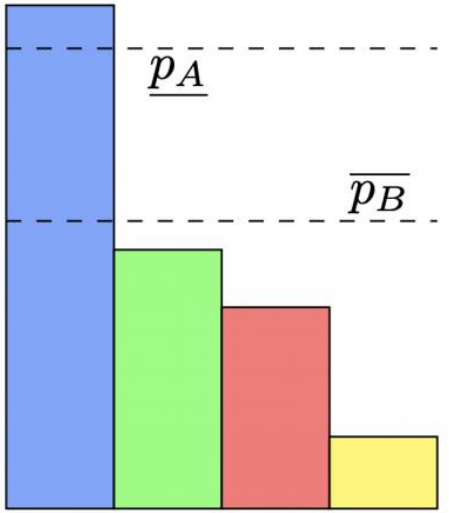
Image sources: "Backdoor Attacks Against Deep Learning Systems in the Physical World," Wenger et al. (2021)
"Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," Chen et al. (2017)

“Randomized Smoothing,” une défense en phase de test

Un modèle **g** est construit à partir d’un réseau **f**

- **g** duplique une entrée **x** donnée et applique à chaque duplicata un bruit donnée (ex. Gaussien)
- La classe avec la valeur de prédiction la plus élevée en moyenne (ex. probits) est celle prédite par le modèle
- La méthode **certifie** un intervalle de confiance autour de l’entrée **x**





L’espace autour de l’entrée **x** est intégrée selon une notion de distance et, sachant un encadrement estimé vis-à-vis de la probabilité de voir la seconde plus haute classe, le modèle **g** retourne la classe la plus prédite

Source: Rosenfeld, “*Generalized Randomized Smoothing for Pointwise-Certified Defenses to Data Poisoning Attacks*” (2020)

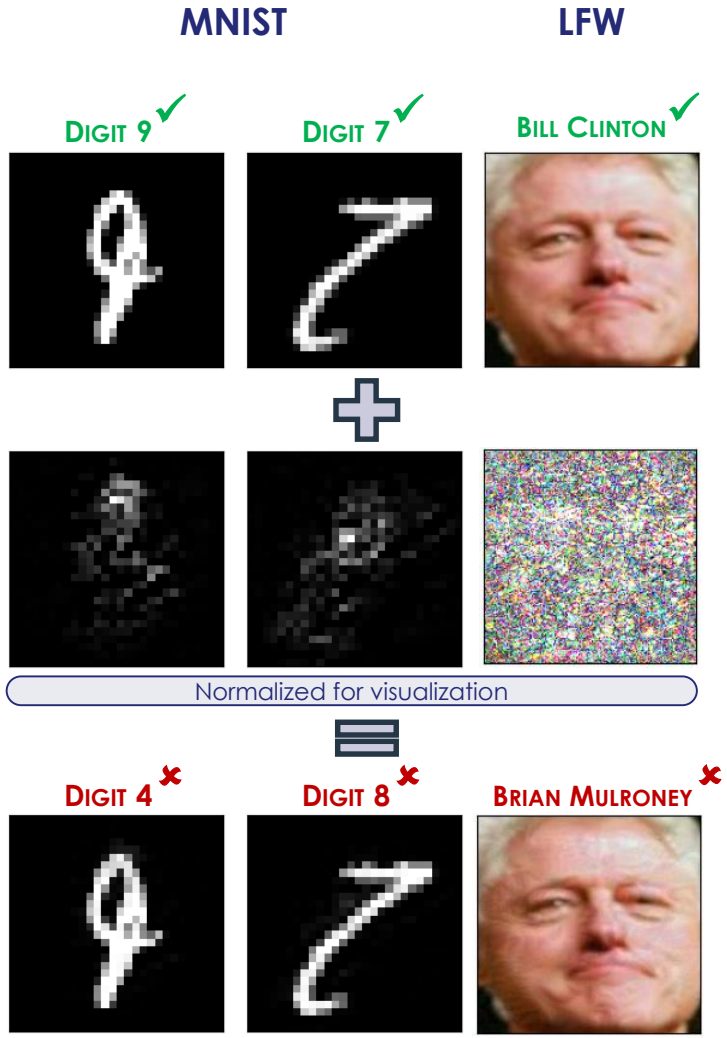
Conclusions et perspectives

- Un réseau de neurones est un **modèle versatile mais vulnérable**
- Le cycle de vie entier du réseau est susceptible à des manipulations par des agents malicieux
- Comprendre les différents vecteurs d'attaque est primordial dans le cadre d'applications critiques
- Présentation centrée sur l'intégrité des réseaux de neurones
- D'autres menaces existent:
  - **Confidentialité**, **disponibilité**, **respect de la vie privée**
- Nombreux travaux en cours chez Thales

Annexe

Attack Examples

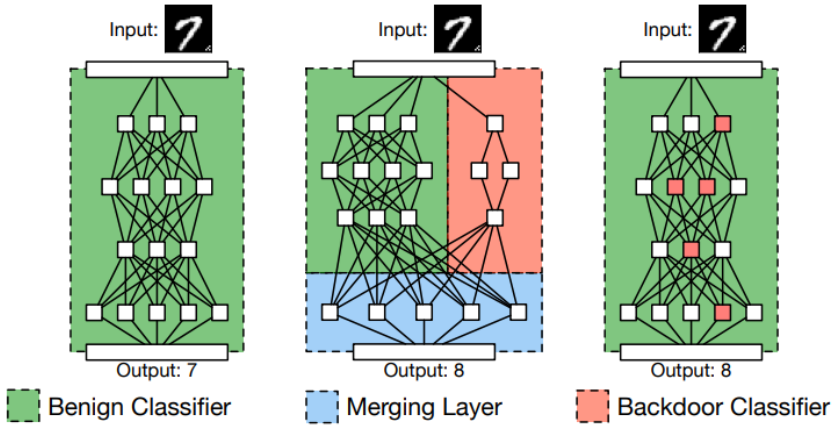
Adversarial examples



Model backdoor

"BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," Gu et al. (2019)

Figure 8. Real-life example of a backdoored stop sign near the authors' office. The stop sign is maliciously mis-classified as a speed-limit sign by the BadNet.



Poisoned data

Digital trigger

"Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," Chen et al. (2017)

(a) An image blended with the Hello Kitty pattern.

(b) An image blended with the random pattern.

Fig. 6: Poisoning instances blended with different patterns. In both images, the blended ratio $\alpha = 0.2$.

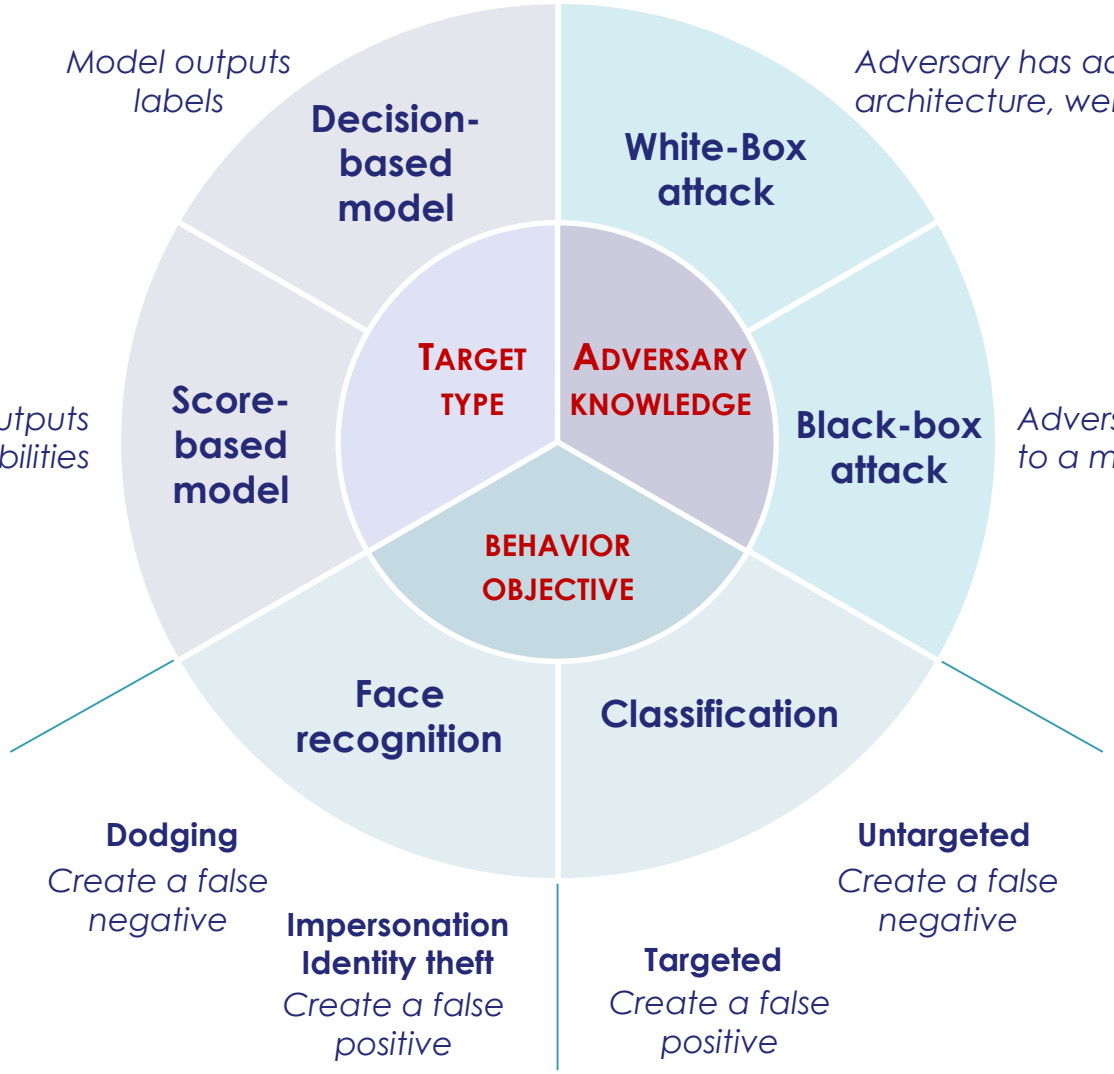
Physical trigger

"Backdoor Attacks Against Deep Learning Systems in the Physical World," Wenger et al. (2021)

Figure 11: *Examples of clean and poison data used in the object recognition experiments*

THALES

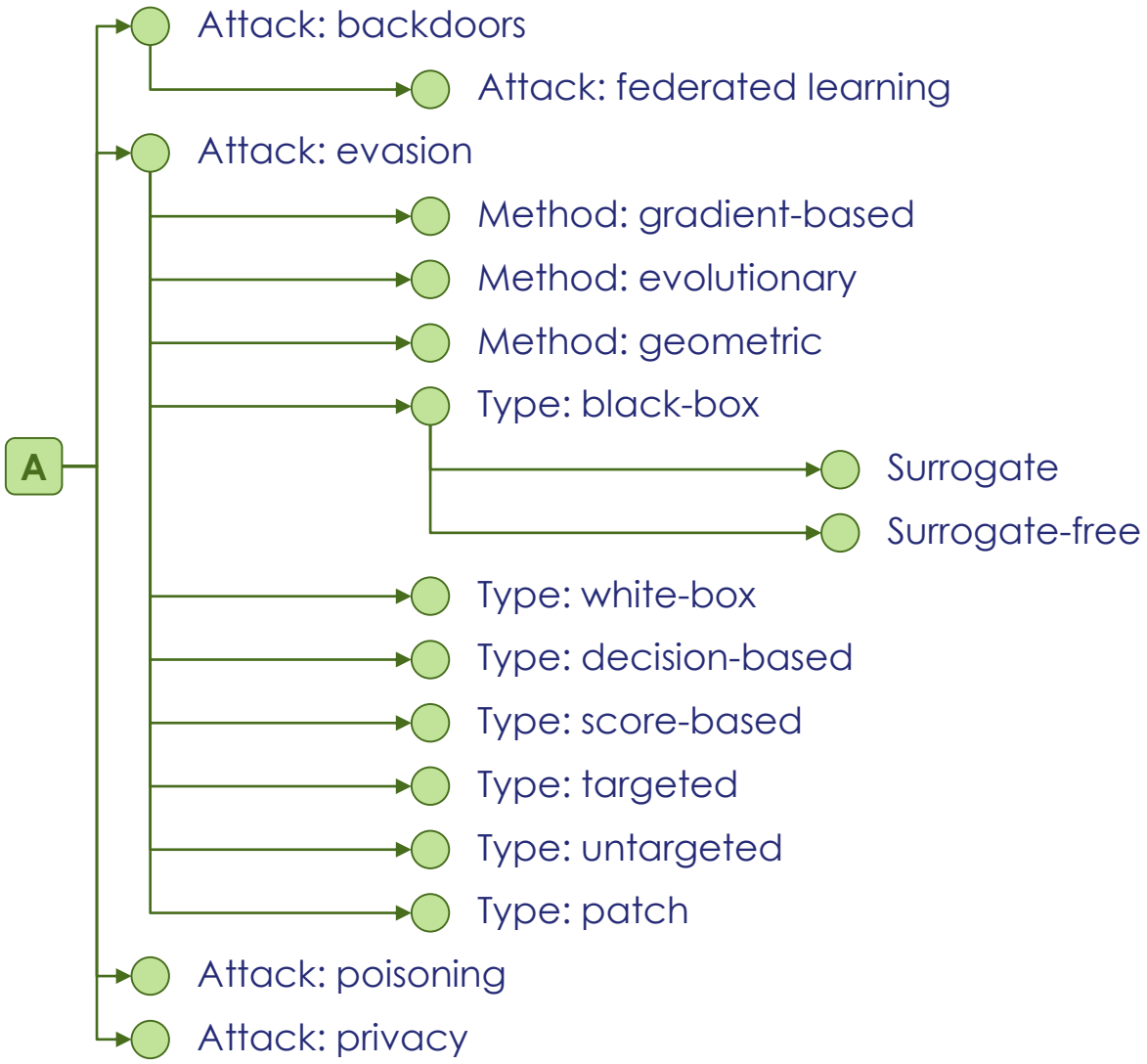
Threat model (2/2)

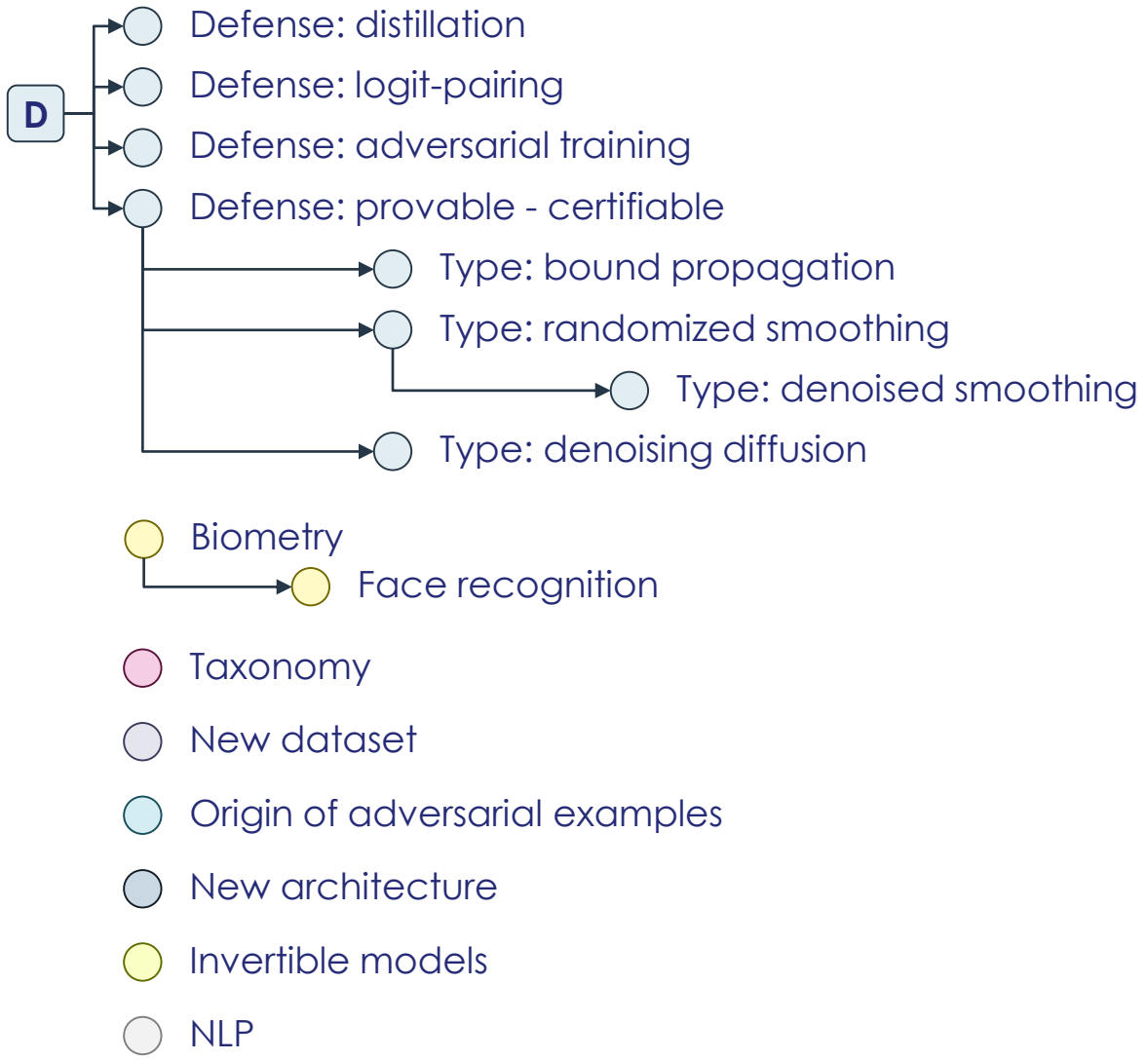


A **Targeted/impersonation black-box** attack on a **top-1 label model** is generally considered to be the **hardest** setting

THALES

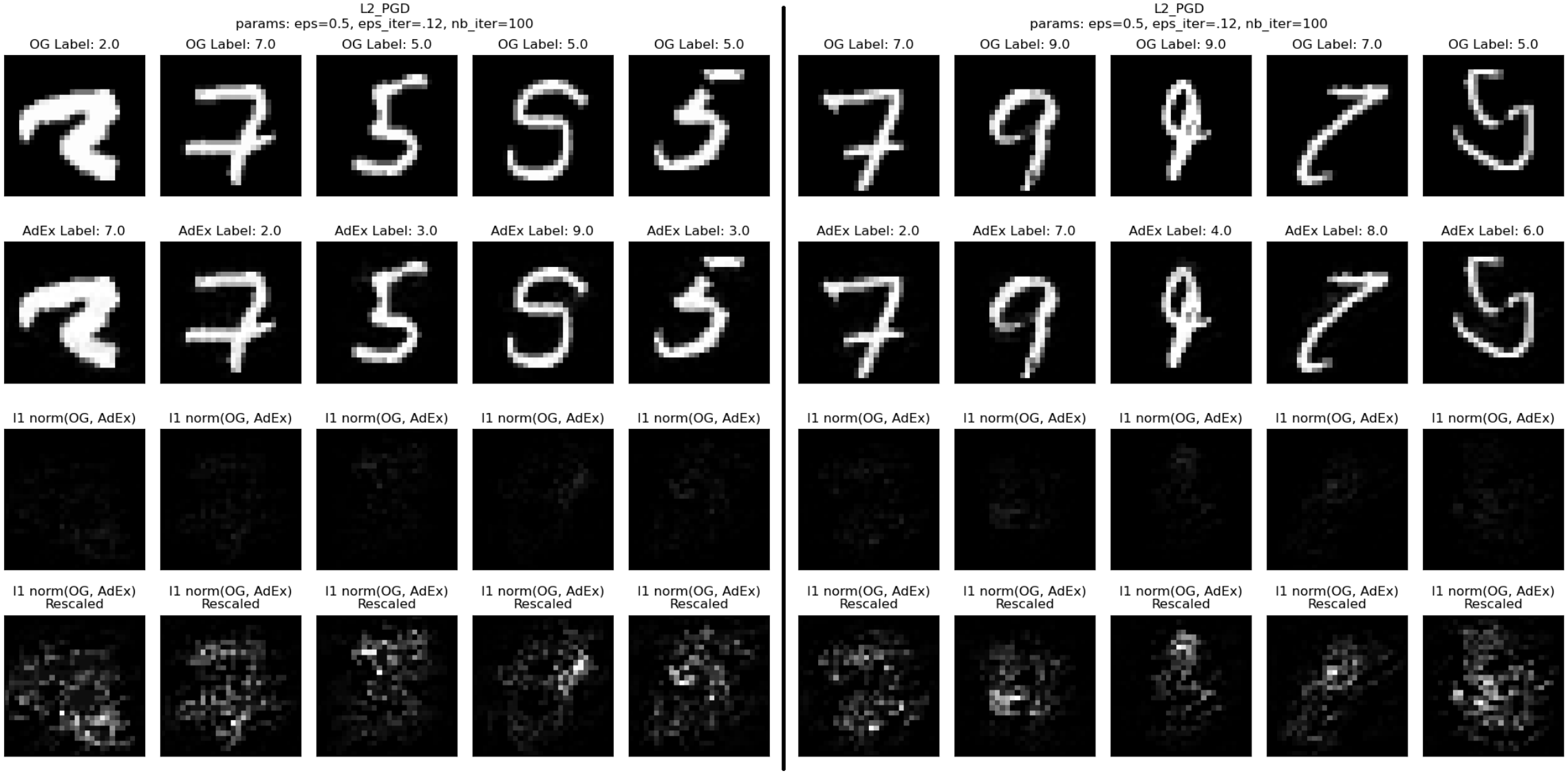
Nomenclature



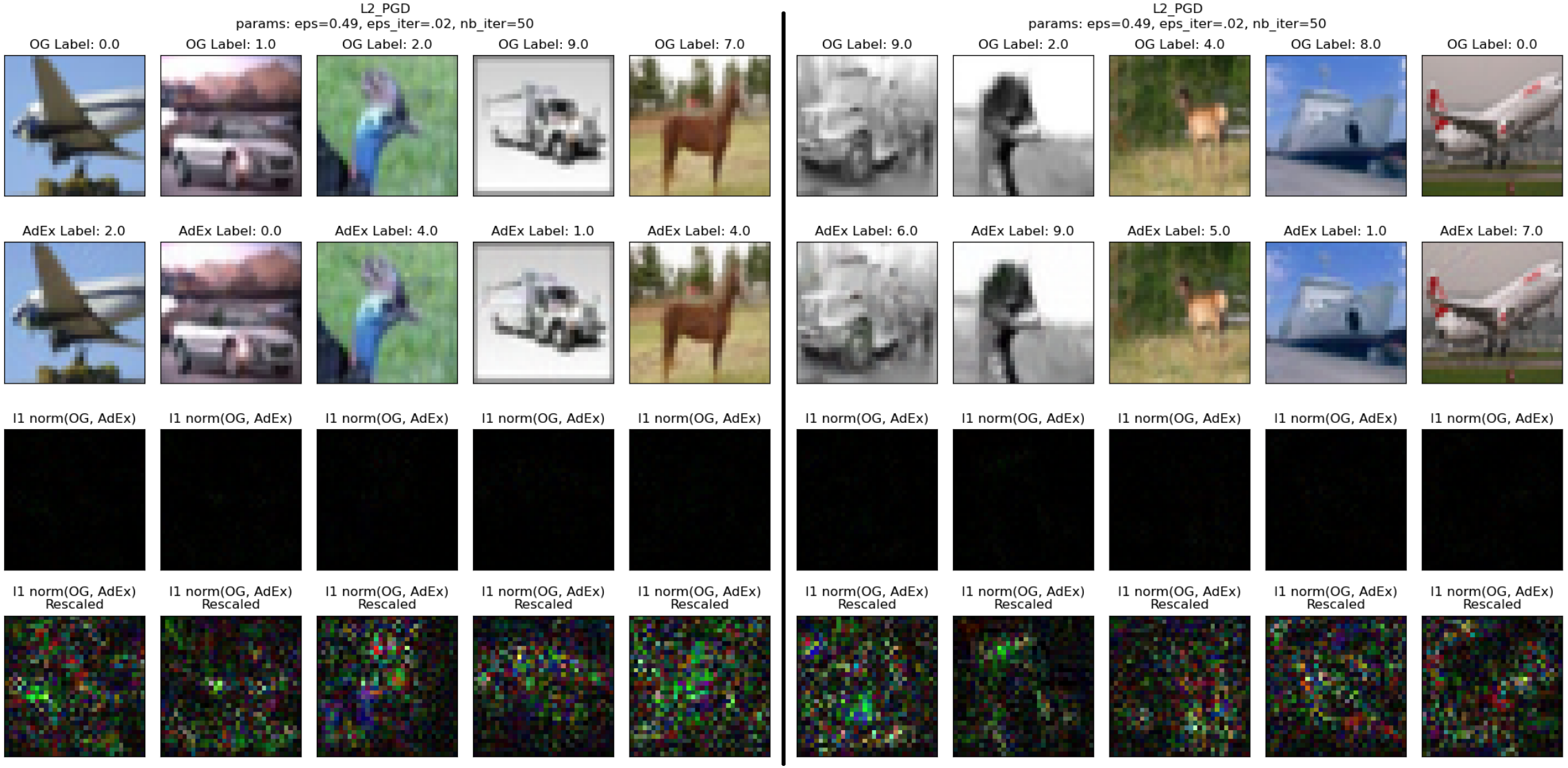


THALES

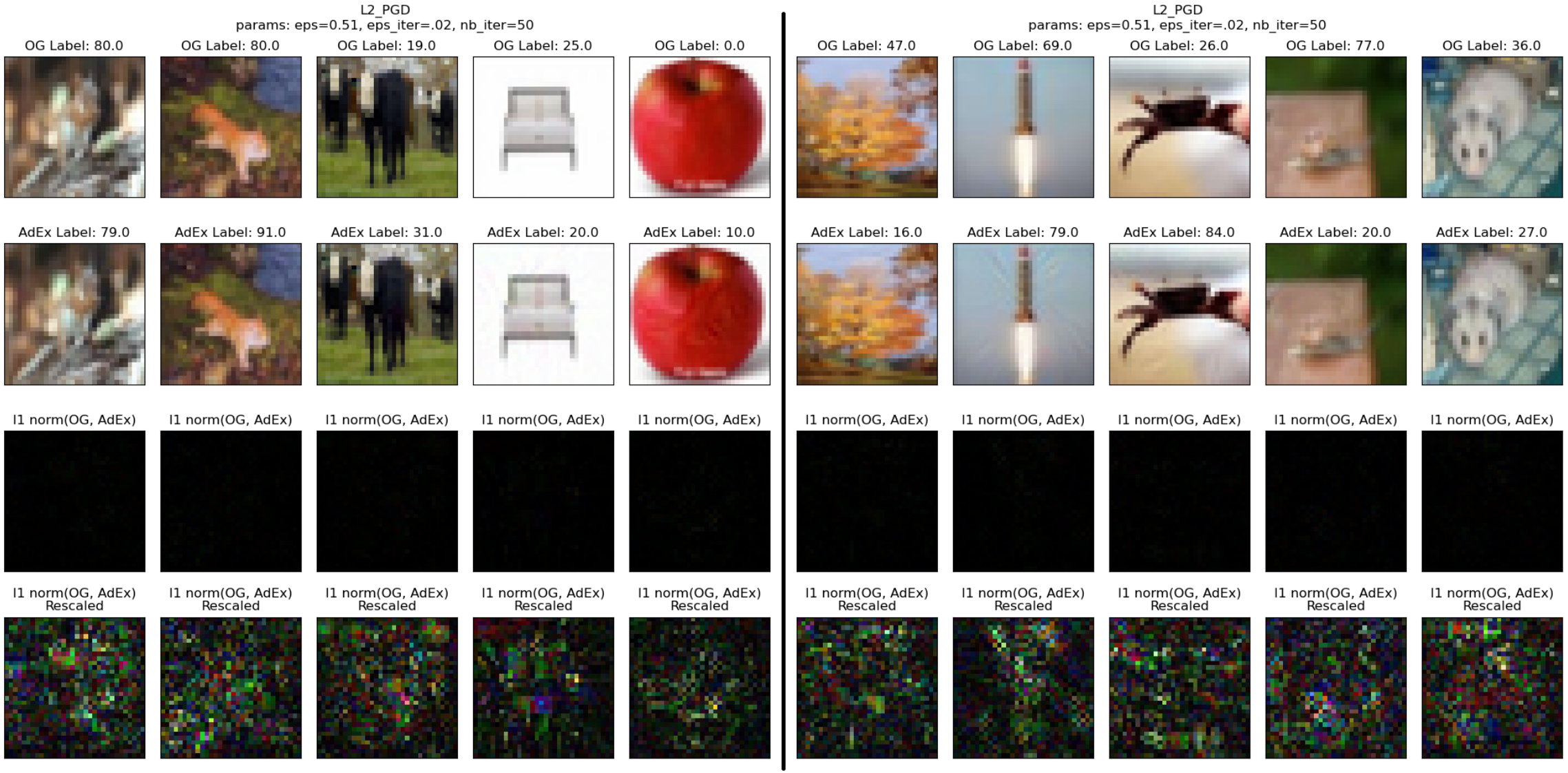
What is an evasion/adversarial attack?



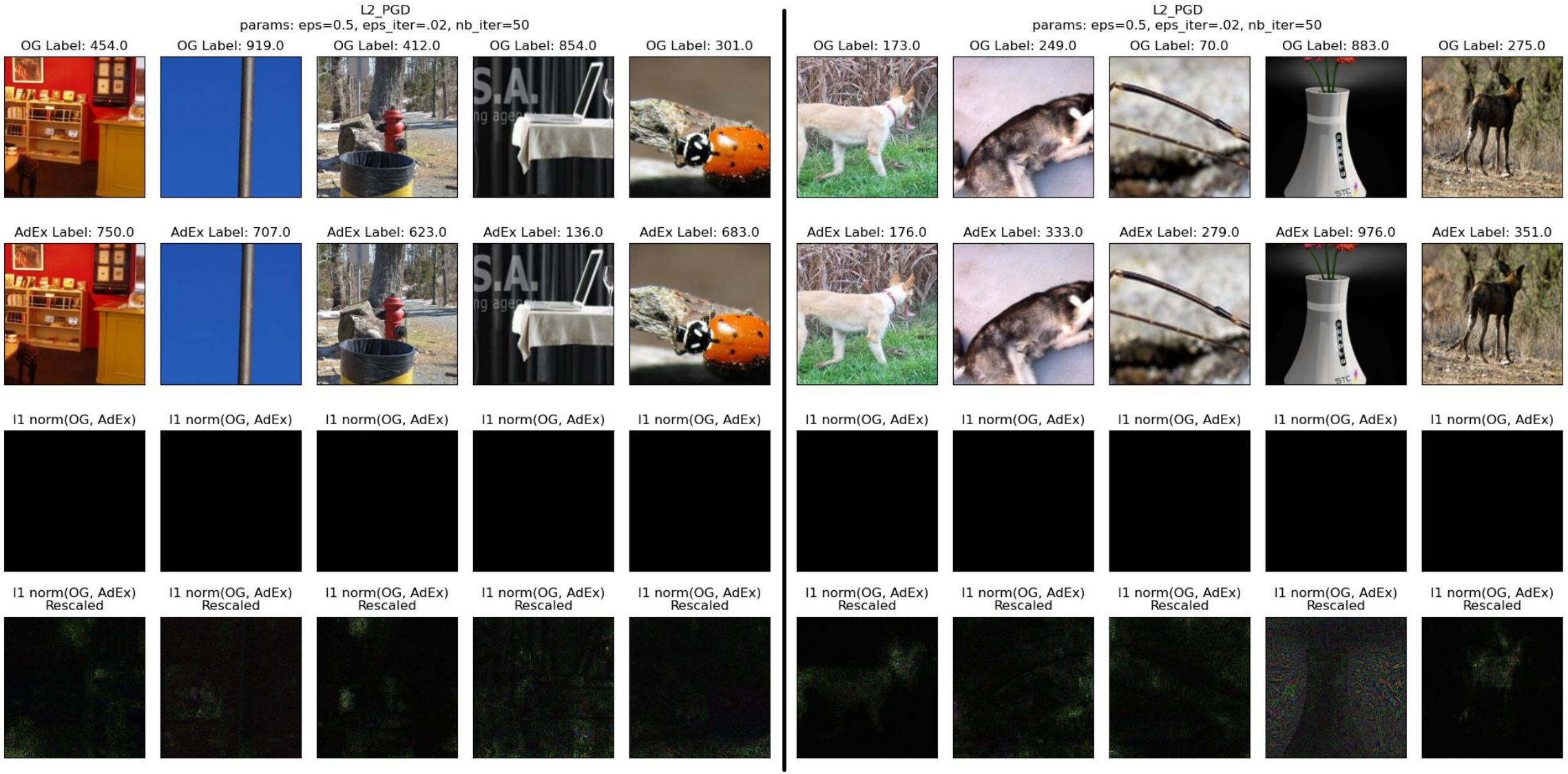
What is an evasion/adversarial attack?



What is an evasion/adversarial attack?

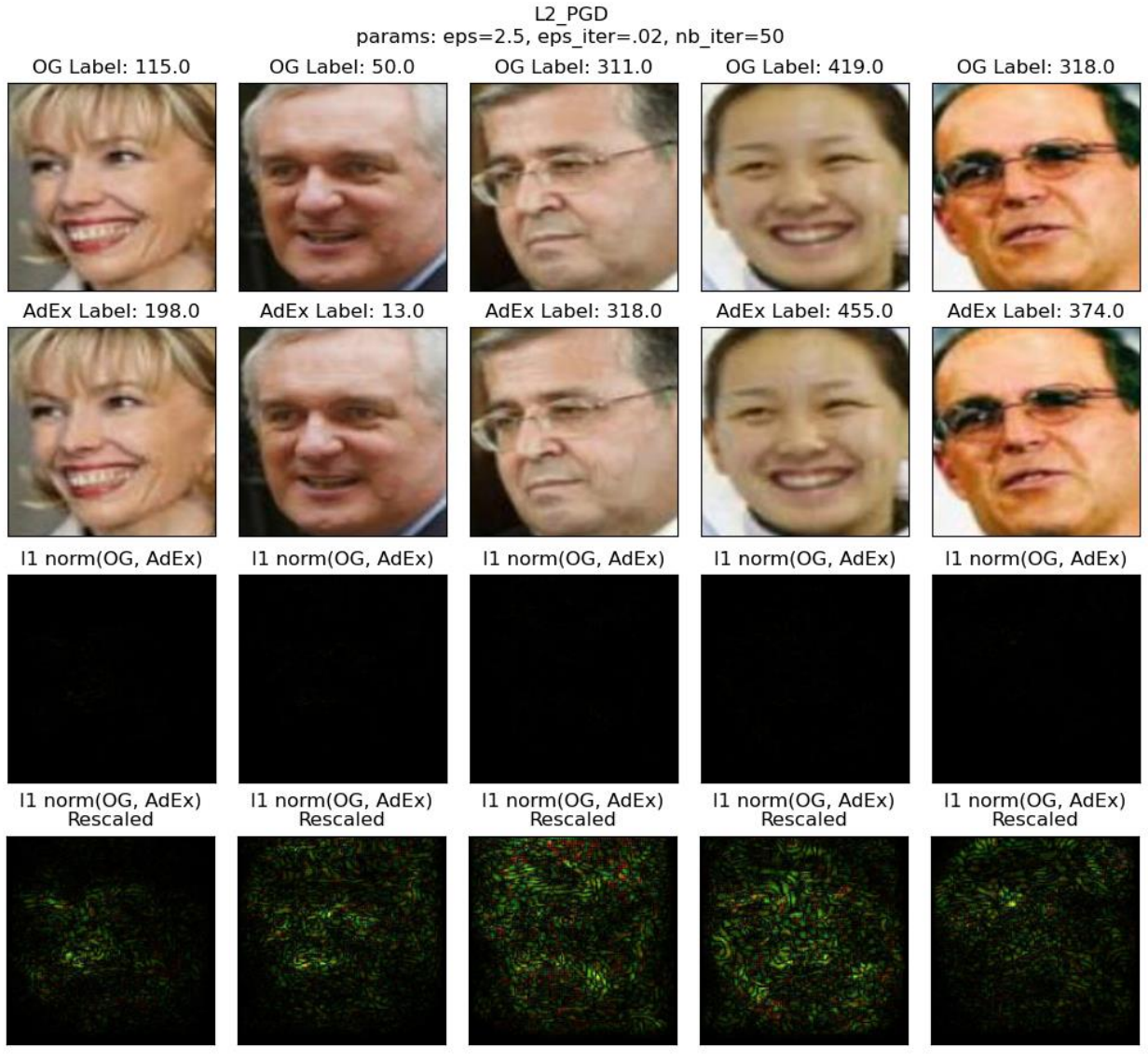


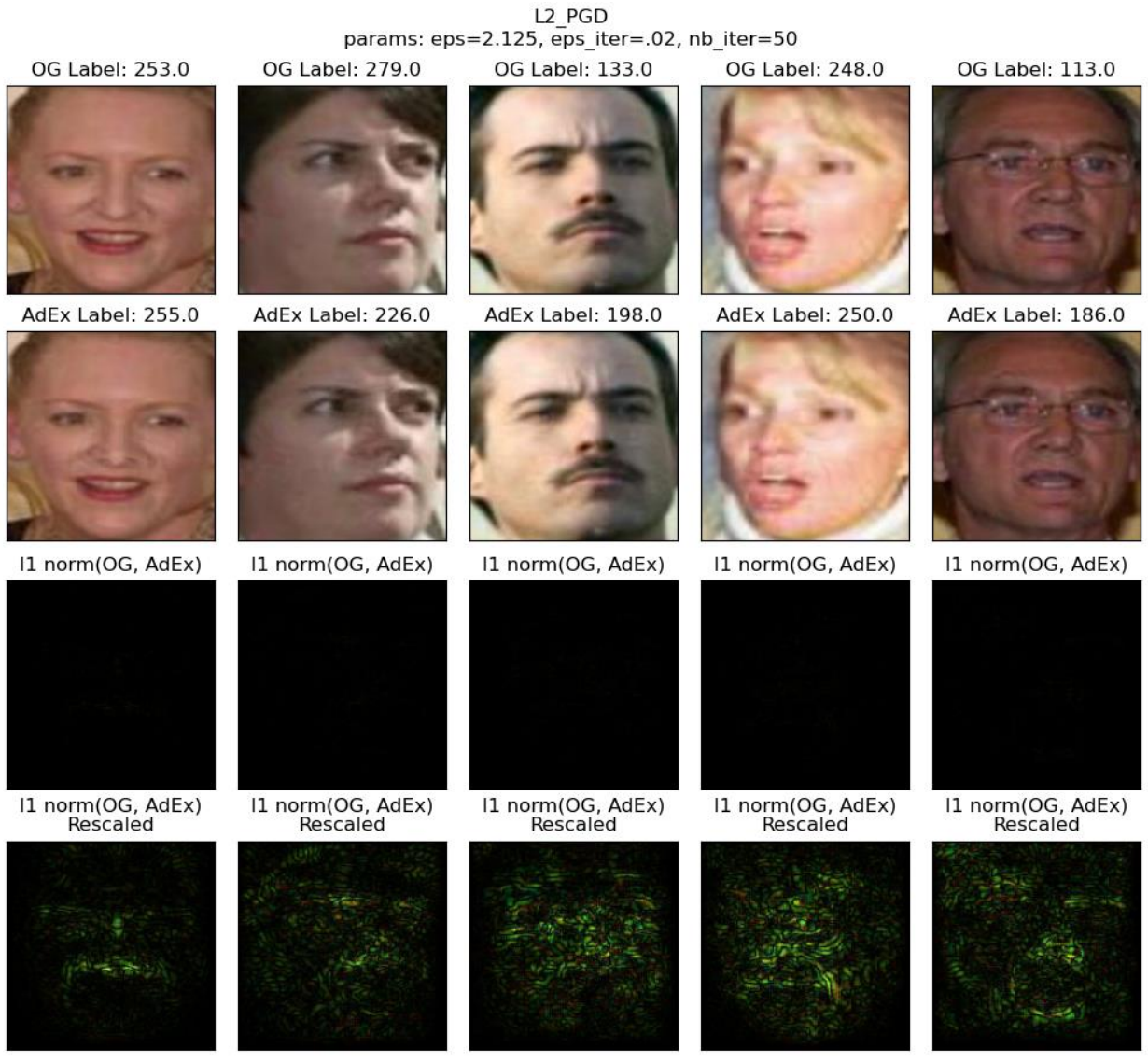
What is an evasion/adversarial attack?



THALES

What is an evasion/adversarial attack?





THALES

Adversarial Training (training-time defense)

Modification of a neural network's SGD update step

- The loss is computed not only on the input data **X** but also its noisy/adversarial version $\mathbf{X_{adv}}$
- $\mathbf{X_{adv}}$ is sampled with a noise distribution or an iterative generation at each epoch

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}\left[loss(f_\theta(x), y)\right] \quad \text{(general case)}$$

$$\min_{\theta} \mathbb{E}_{(x,y)\sim D}\left[\max_{\delta\in\Delta} loss(f_\theta(x+\delta), y)\right] \quad \text{(worst-case performance)}$$

Source: Frost, "*Certifiable Robustness to Adversarial Attacks*" (2020)