Empoisonnement de données

23/09/2022

Adrien CHAN-HON-TONG
ONERA

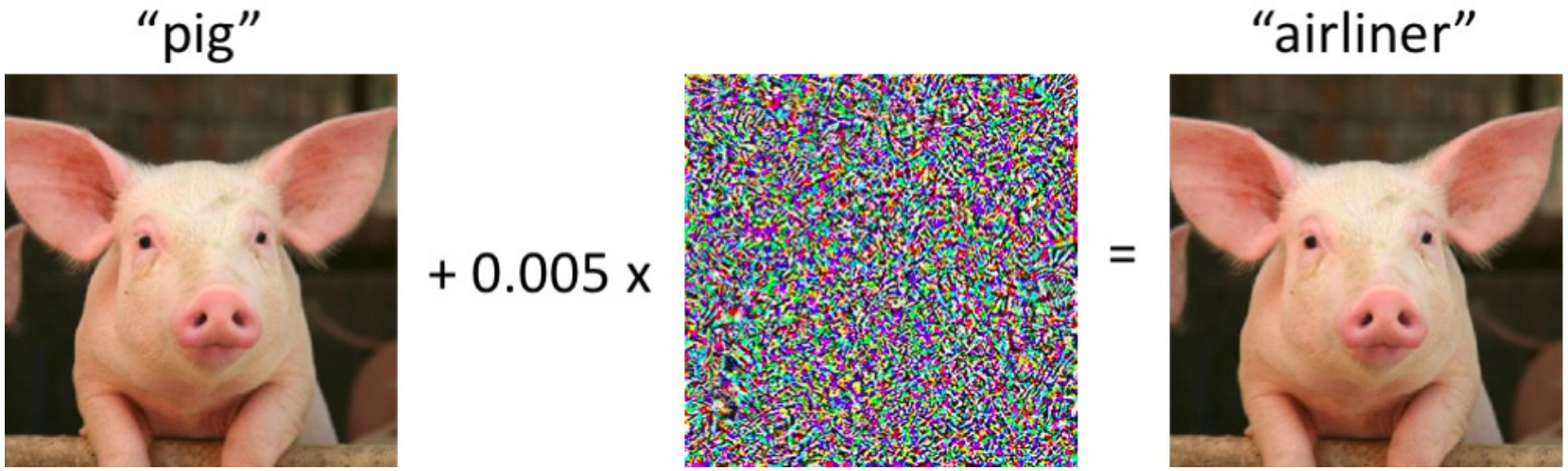
IA de confiance à l’ONERA

Objectif

→ **l'objectif de cette présentation est de présenter une cartographie des risques d'empoisonnement mais aussi des contre-mesures**

Exemple 1

La dangerosité des exemples adversaires invisibles ?

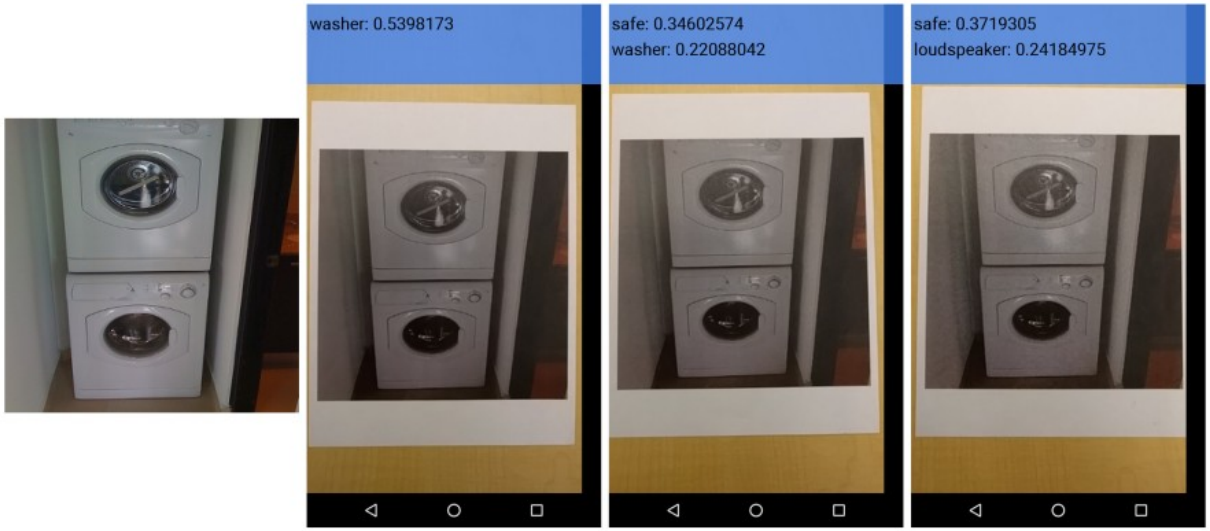


https://gradientscience.org/intro_adversarial/

Exemple 1
La dangerosité des exemples adversaires invisibles ?

1 → Ce type d'attaque semble difficilement réalisable dans le monde physique

malgré



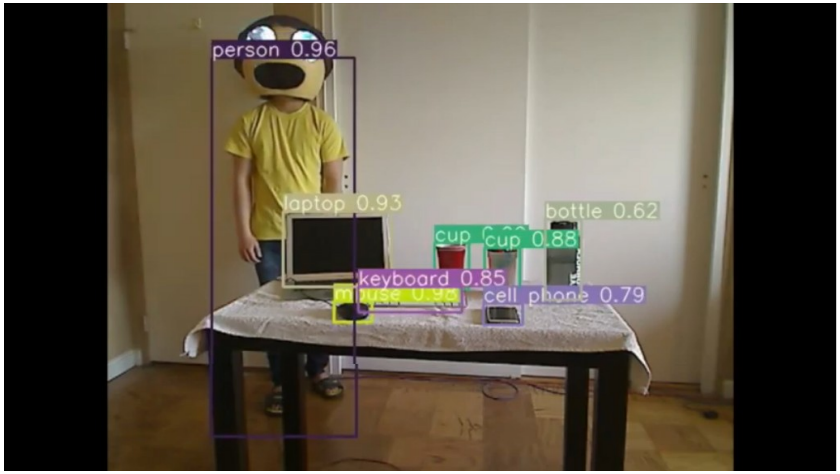
(a) Image from dataset (b) Clean image (c) Adv. image, $\epsilon = 4$ (d) Adv. image, $\epsilon = 8$

2 → Un réentraînement avec PGD donne une robustesse non négligeable

https://arxiv.org/abs/1607.02533

Exemple 2

La dangerosité des exemples adversaires patchs ?



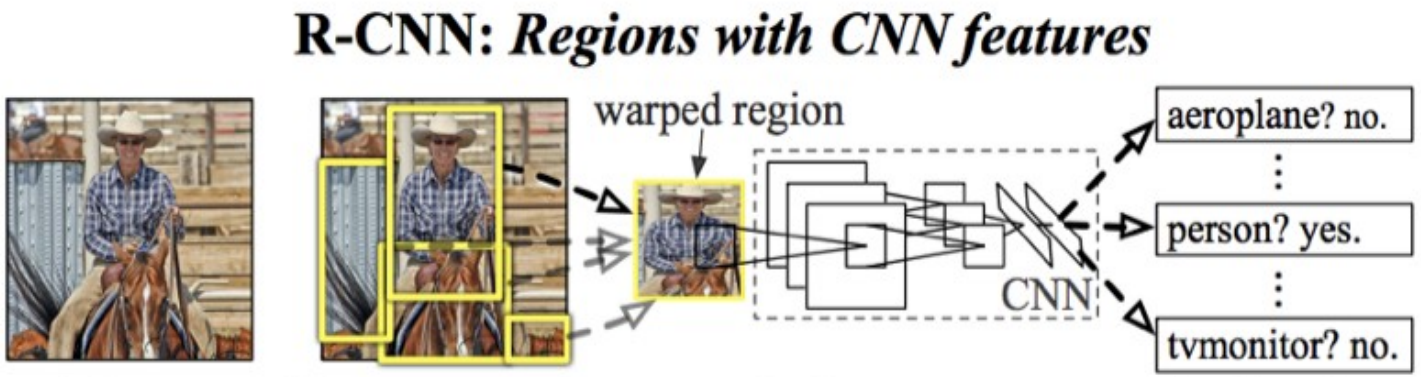
(a) Adversarial patch on the roof of a car.

(b) Adversarial patch off-and-around a car.

https://arxiv.org/abs/1906.11897
https://openaccess.thecvf.com/content/WACV2022/papers/Du_Physical_Adversarial_Attacks_on_an_Aerial_Imagery_Object_Detector_WACV_2022_paper.pdf

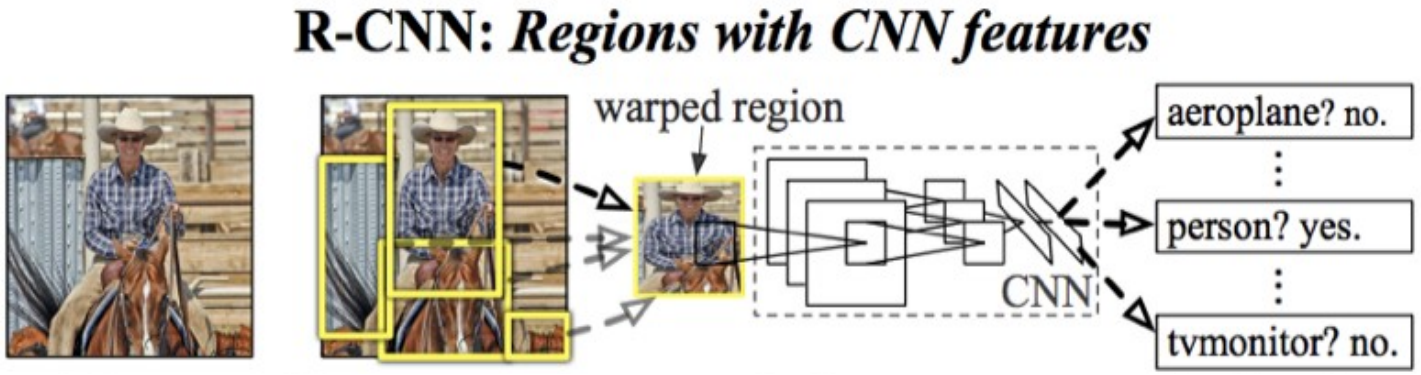
Exemple 2

La dangerosité des exemples adversaires patchs ?

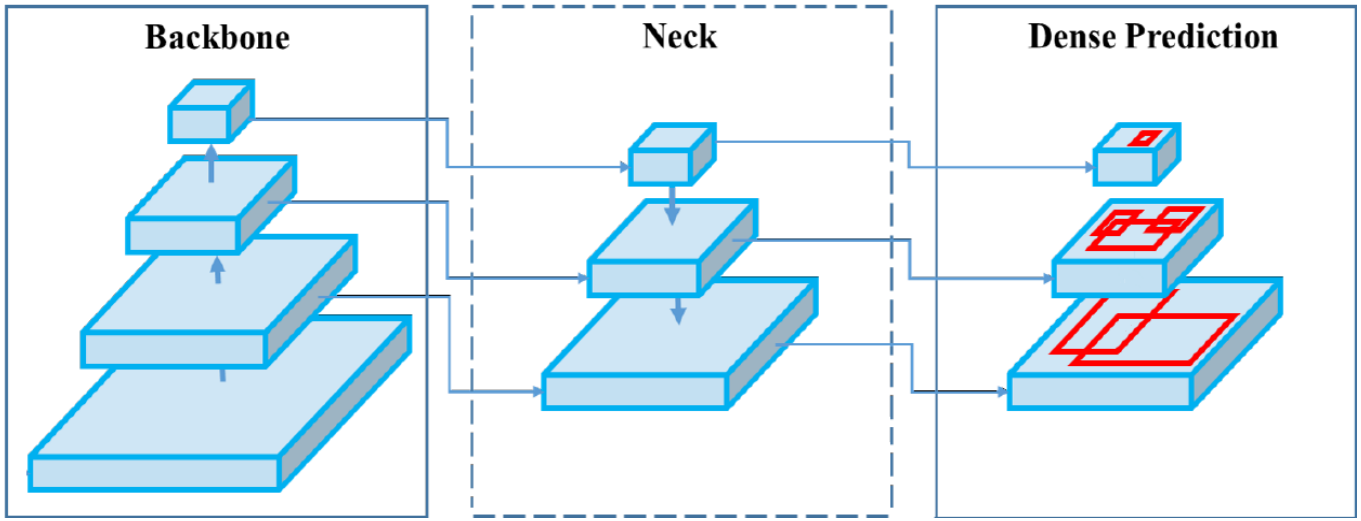


Exemple 2

La dangerosité des exemples adversaires patchs ?

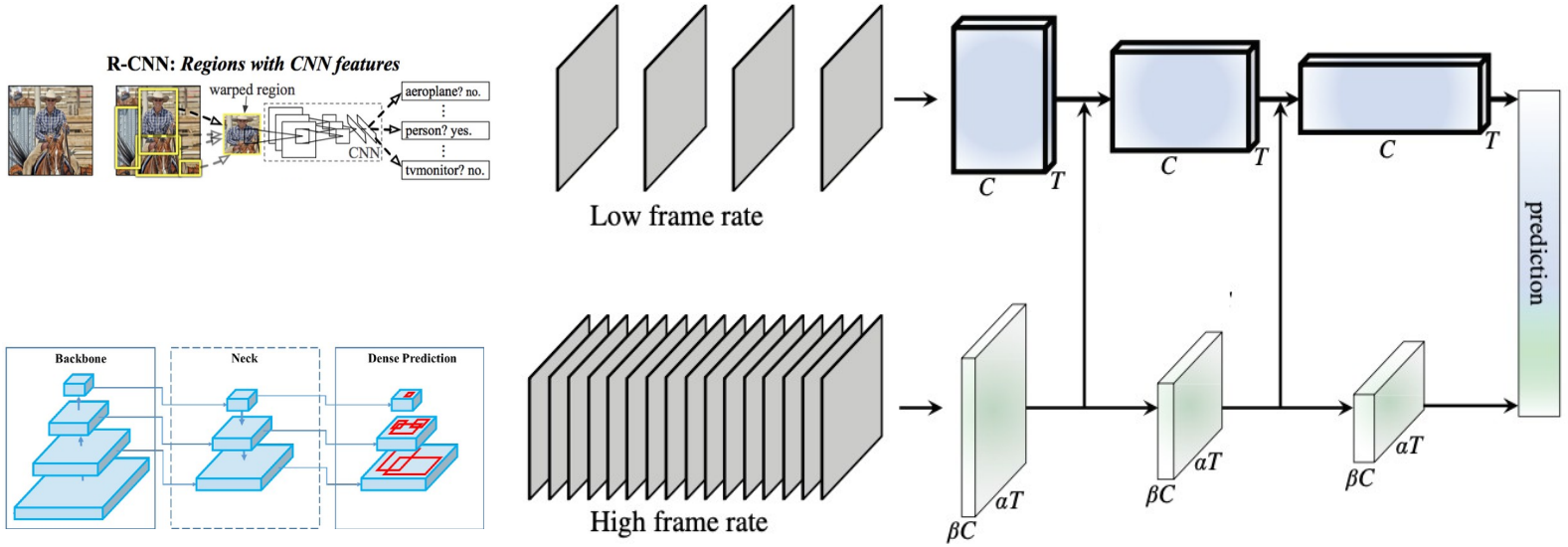


medium.com



Exemple 2

La dangerosité des exemples adversaires patchs ?



Plan

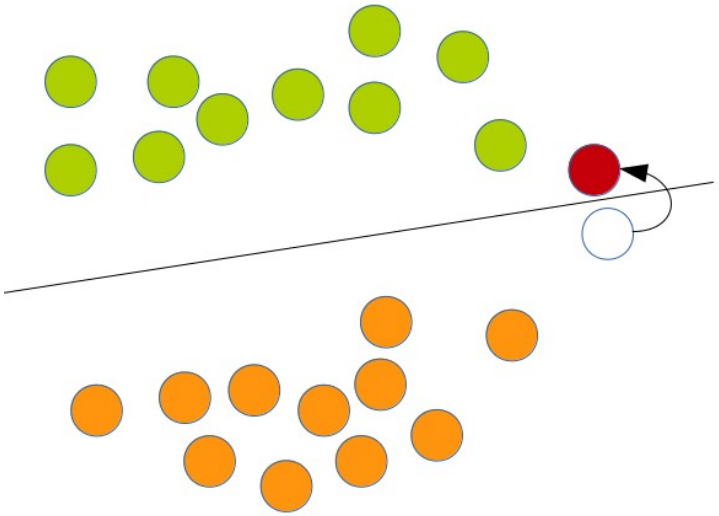
→ scénario d’empoisonnement

→ contre-mesure à une empoisonnement invisible

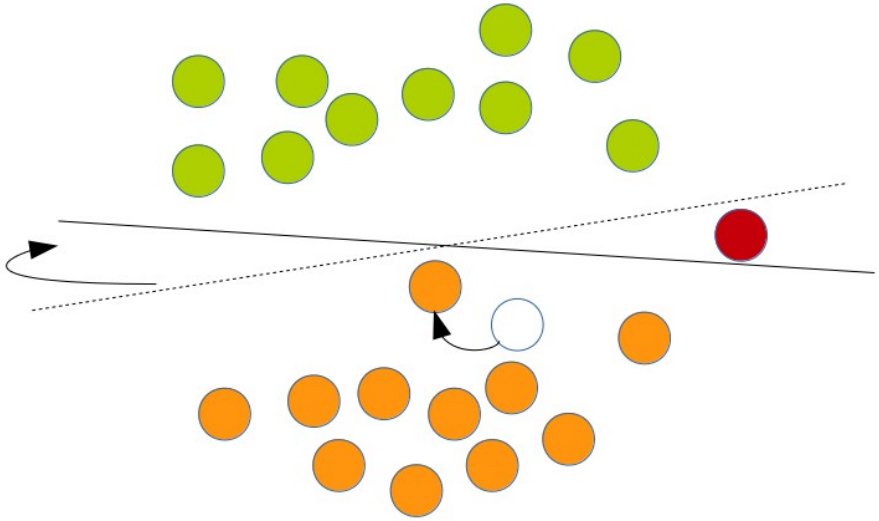
→ perspectives

empoisonnement vs attaque évasion

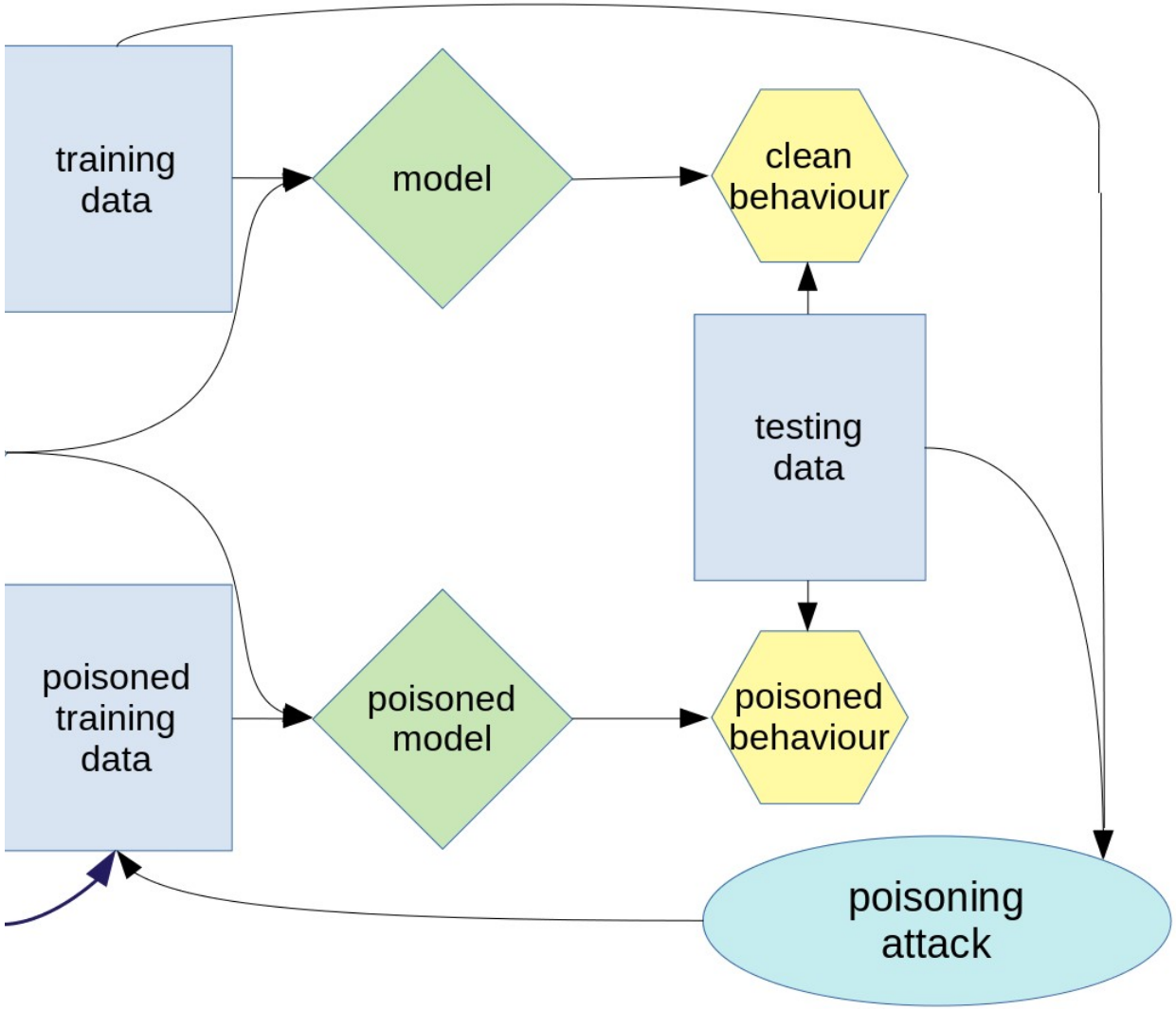
Attaque adversaire



Empoisonnement



empoisonnement

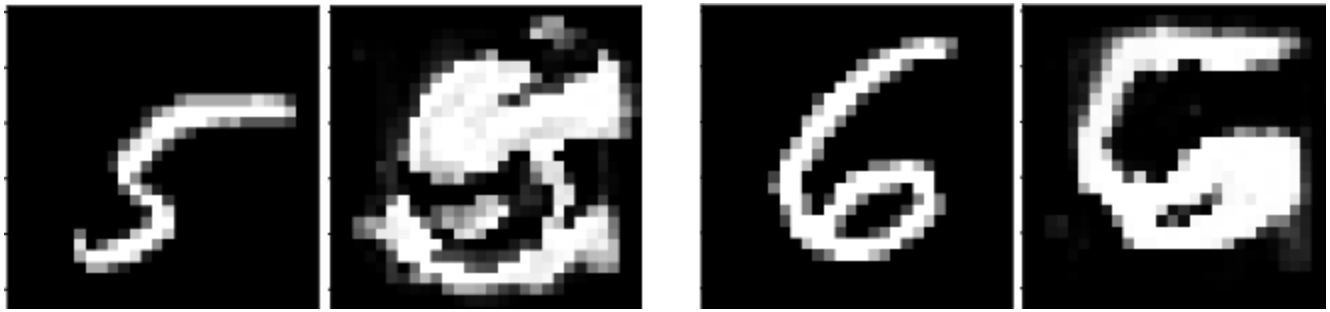


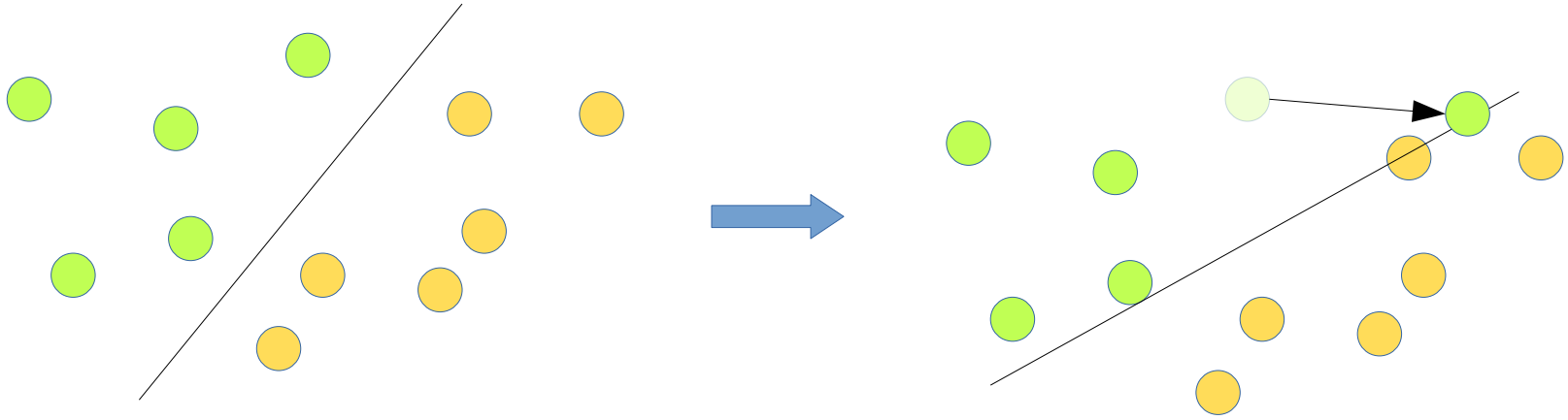
empoisonnement vs attaque évasion

	Attaque adversaire	Empoisonnement
Avantages	**Facilité d'exécution**	Modifie le comportement de la cible **partout**
Inconvénients	Modification du comportement locale	Nécessite un accès aux données d'apprentissage.

empoisonnement classique vs empoisonnement invisible

Image corruption

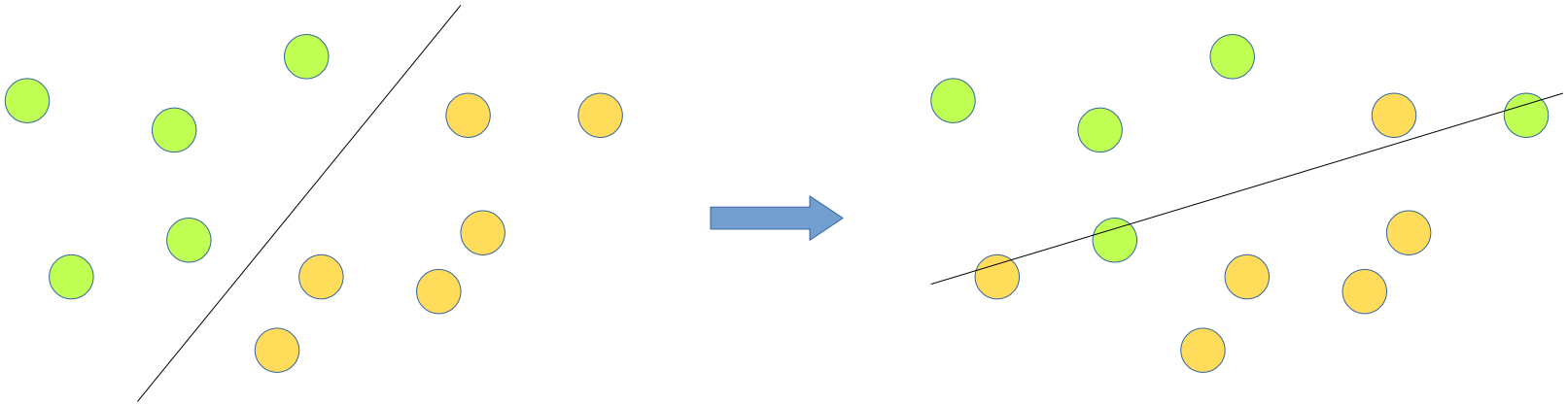




https://arxiv.org/pdf/1708.08689

empoisonnement classique vs empoisonnement invisible

Label flip

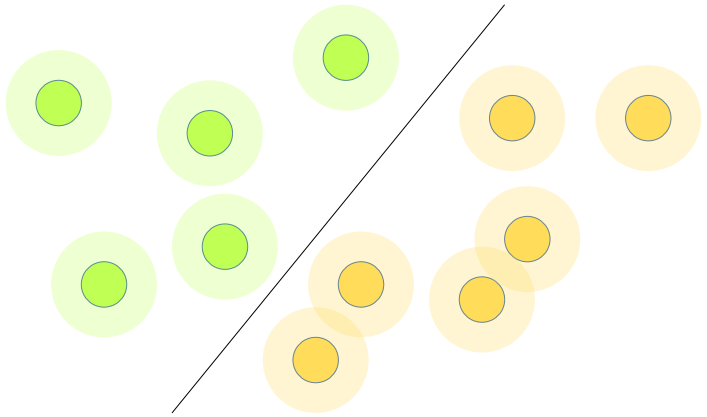


https://arxiv.org/pdf/1708.08689

empoisonnement classique vs empoisonnement invisible

Clean label

Modification des images contraintes à être indétectable à l'oeil



Scénarios d’empoisonnement

Soit l’apprentissage est fait sans contrôle

Soit la modification doit passer inaperçue

Scénarios d’empoisonnement

Soit l’apprentissage est fait sans contrôle

Ce qui paraît pas très réaliste

Soit la modification doit passer inaperçue

Scénarios d’empoisonnement

Soit l’apprentissage est fait sans contrôle

Ce qui paraît pas très réaliste

(Sauf pseudo labeling, online learning, watermarking, falsification)

Soit la modification doit passer inaperçue

Scénarios d’empoisonnement

Soit l’apprentissage est fait sans contrôle

Ce qui paraît pas très réaliste

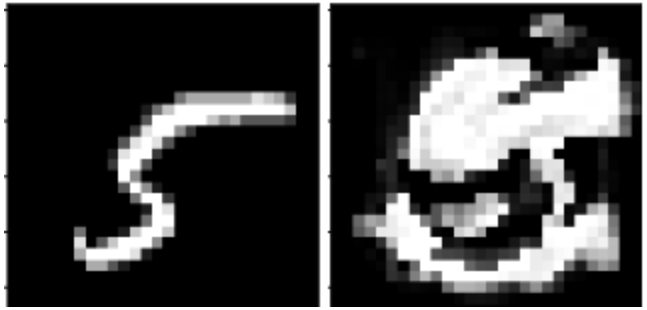
(Sauf pseudo labeling, online learning, watermarking, falsification)

Soit la modification doit passer inaperçue

→ faible nombre de modification

→ OU modification invisible

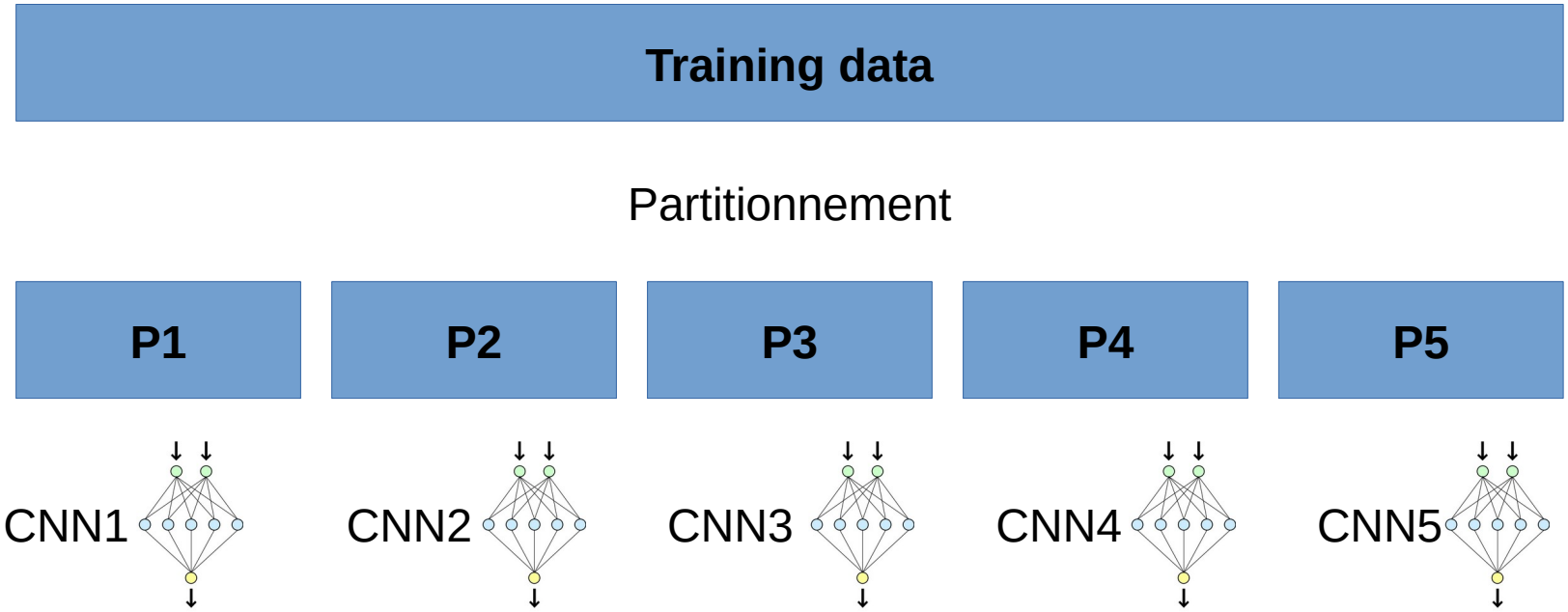
Scénarios d’empoisonnement



Modifications fortes mais peu nombreuses

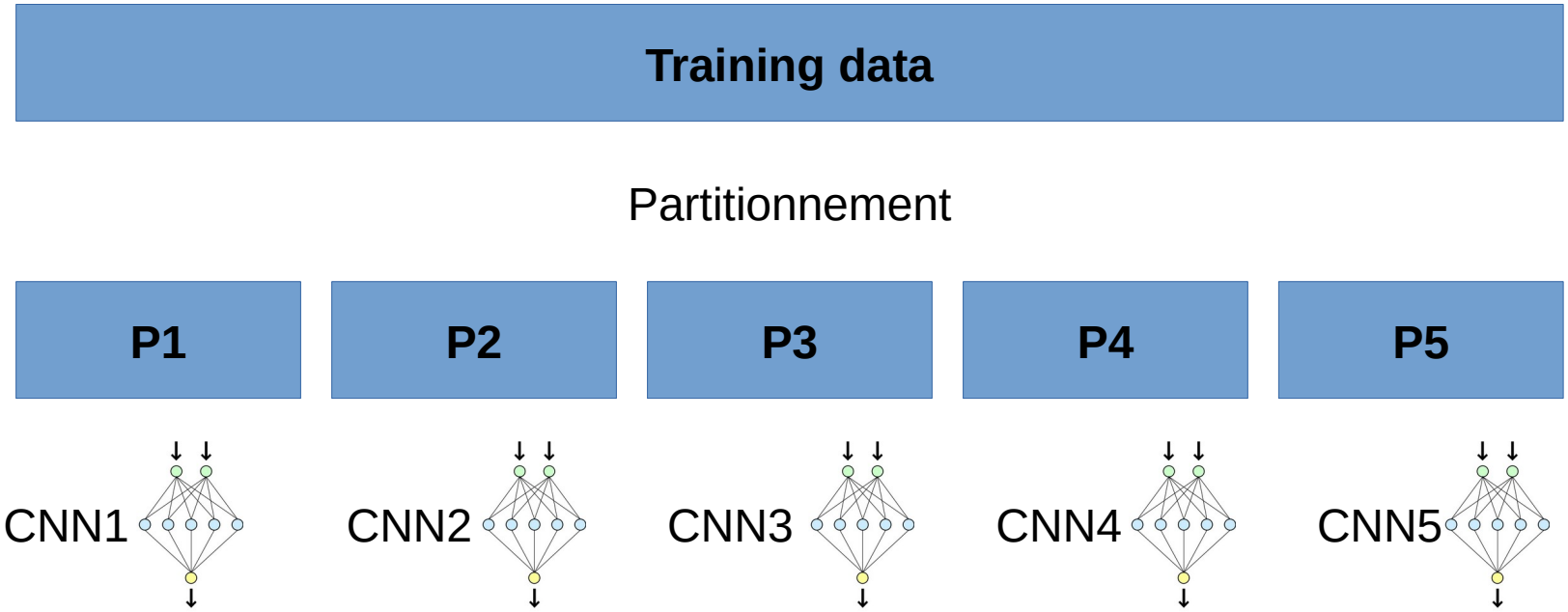
Modifications invisibles

Modifications fortes mais minoritaires vs partition



https://arxiv.org/abs/2006.14768

Modifications fortes mais minoritaires vs partition

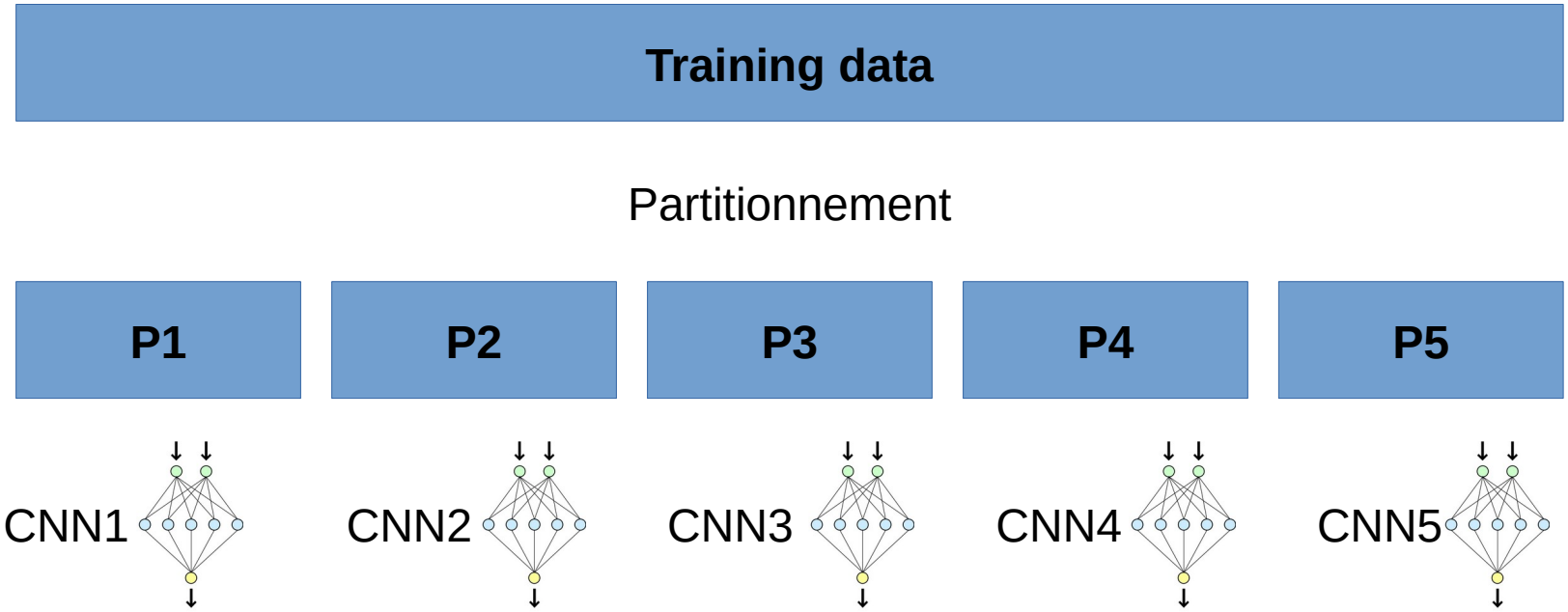


Si on modifie 1 données, seul 1 des CNN est modifié

→ Si au moins 4 CNN sont d'accord sur la classe, alors cet accord n'a pas pu être influencé par la présence d'1 donnée empoisonnée

https://arxiv.org/abs/2006.14768

Modifications fortes mais minoritaires vs partition



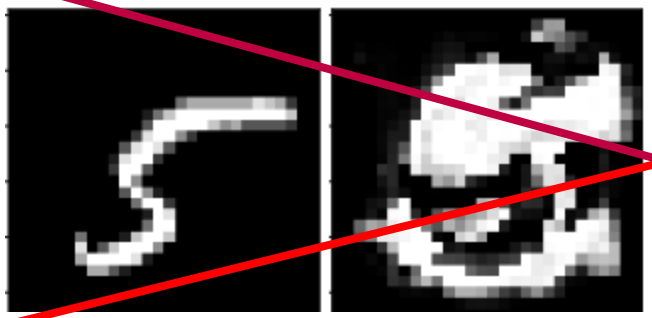
Si on modifie 1 données, seul 1 des CNN est modifié

→ Si au moins 4 CNN sont d'accord sur la classe, alors cet accord n'a pas pu être influencé par la présence d'1 donnée empoisonnée

→ **bien entendu cette défense est difficile mais peut s'implémenter de façon souple (RANSAC)**

https://arxiv.org/abs/2006.14768

Scénarios d’empoisonnement



Modifications fortes mais peu nombreuses

Modifications invisibles

Plan

→ scénario d’empoisonnement

→ contre-mesure à une empoisonnement invisible

→ perspectives

empoisonnement invisible

Poison Frog

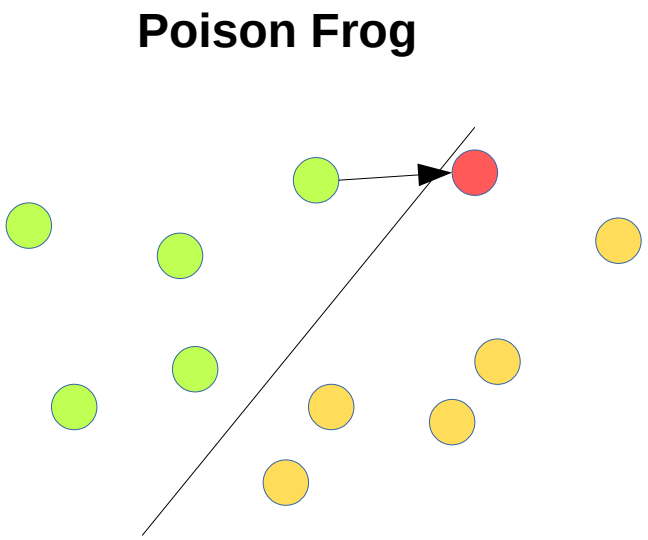
2.1 Crafting poison data via feature collisions

Let $f(\mathbf{x})$ denote the function that propagates an input $\mathbf{x}$ through the network to the penultimate layer (before the softmax layer). We call the activations of this layer the *feature space* representation of the input since it encodes high-level semantic features. Due to the high complexity and nonlinearity of f, it is possible to find an example $\mathbf{x}$ that "collides" with the target in feature space, while simultaneously being close to the base instance $\mathbf{b}$ in input space by computing

$$\mathbf{p} = \underset{\mathbf{x}}{\operatorname{argmin}} \ \|f(\mathbf{x}) - f(\mathbf{t})\|_2^2 + \beta \|\mathbf{x} - \mathbf{b}\|_2^2 \qquad (1)$$

https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html

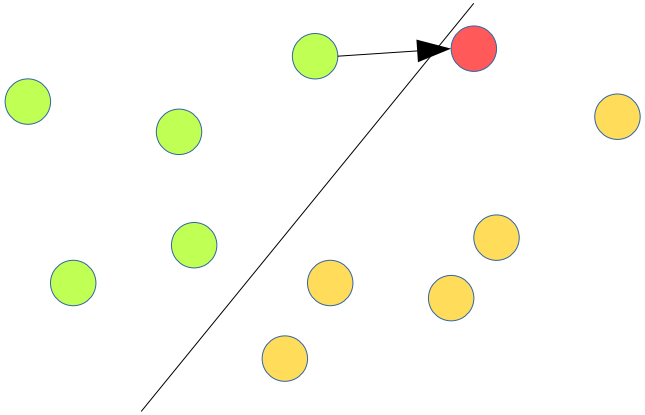
empoisonnement invisible



https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html

empoisonnement invisible

Poison Frog

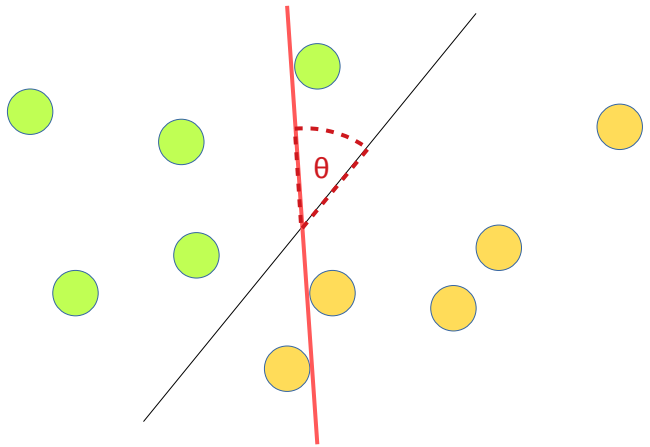


Success rate 100 %
(Inception v3 + svm, Imagenet dog vs fish)

https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html

empoisonnement invisible

Proxy based



https://www.mdpi.com/2504-4990/1/1/11/htm

empoisonnement invisible

Proxy based

$$u \rightarrow u + \tau(\theta_{fair}^{T} u)\theta_{poison} - \tau(\theta_{poison}^{T} u)\theta_{fair}$$

$$\theta_{hack} \approx \theta_{fair} + \tau(\theta_{fair}^{T}\theta_{fair})\theta_{poison} - \tau(\theta_{poison}^{T}\theta_{fair})\theta_{fair}$$

https://www.mdpi.com/2504-4990/1/1/11/htm

empoisonnement invisible

Proxy based

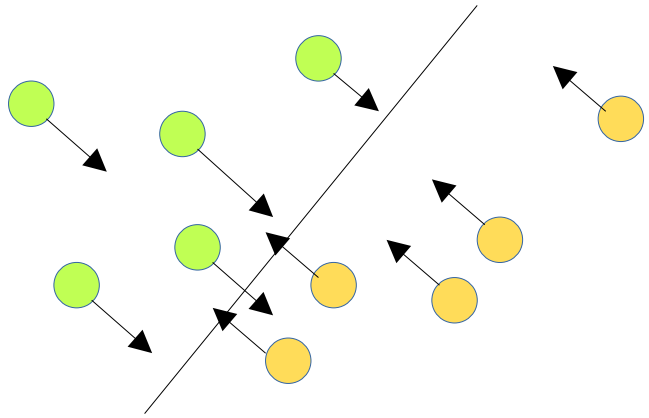
$$u \rightarrow u + \tau(\theta_{fair}^T u)\theta_{poison} - \tau(\theta_{poison}^T u)\theta_{fair}$$

$$\theta_{hack} \approx \theta_{fair} + \tau(\theta_{fair}^T \theta_{fair})\theta_{poison} - \tau(\theta_{poison}^T \theta_{fair})\theta_{fair}$$

CIFAR 75 % → 63 % accuracy (vgg + svm)

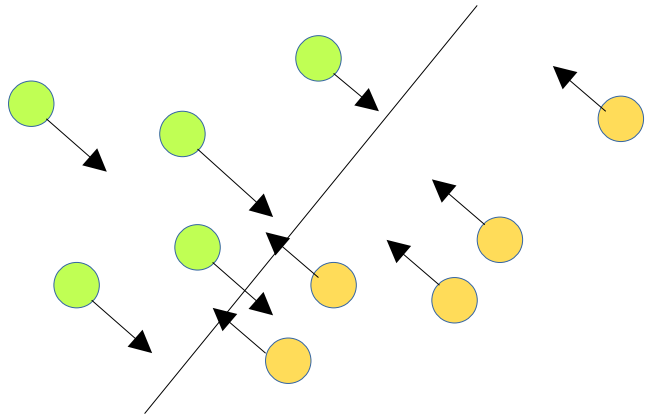
https://www.mdpi.com/2504-4990/1/1/11/htm

empoisonnement invisible



https://www.mdpi.com/2504-4990/1/1/11/htm

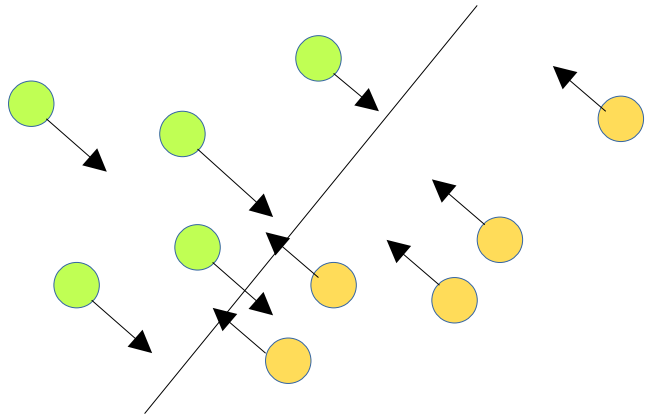
empoisonnement invisible



Un empoisonnement non coordonnée ne conduit à une modification de comportement

https://www.mdpi.com/2504-4990/1/1/11/htm

empoisonnement invisible



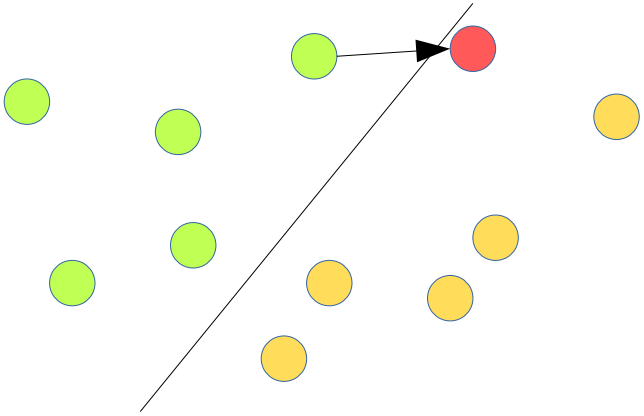
$\mathrm{SGD}_\theta(f, Test)$	$\ll 87\%$ (-31% in [18])
$\mathrm{SGD}_\theta(f, Train)$	$\approx 87\%$ (0% in [18])
$-w_{imagenet}$	$\approx 87\%$
$w_{imagenet}$	$\approx 87\%$
$-\mathrm{SGD}_\theta(f, Train)$	$\approx 87\%$ (-1% in [18])
$-\mathrm{SGD}_\theta(f, Test)$	$\gg 87\%$ (+7% in [18])

Un empoisonnement non coordonnée ne conduit à une modification de comportement

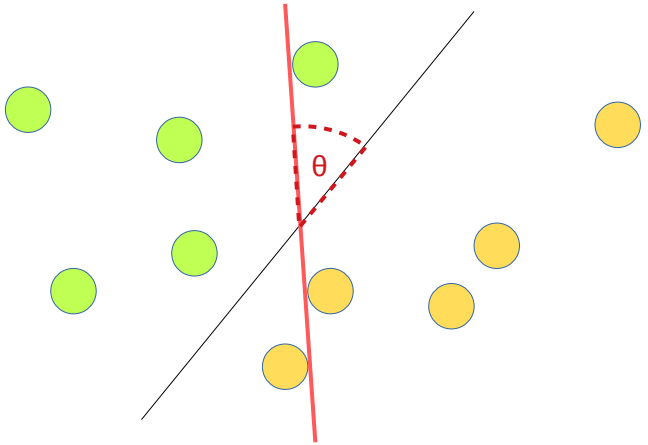
https://www.mdpi.com/2504-4990/1/1/11/htm

empoisonnement invisible

Poison Frog



Proxy based

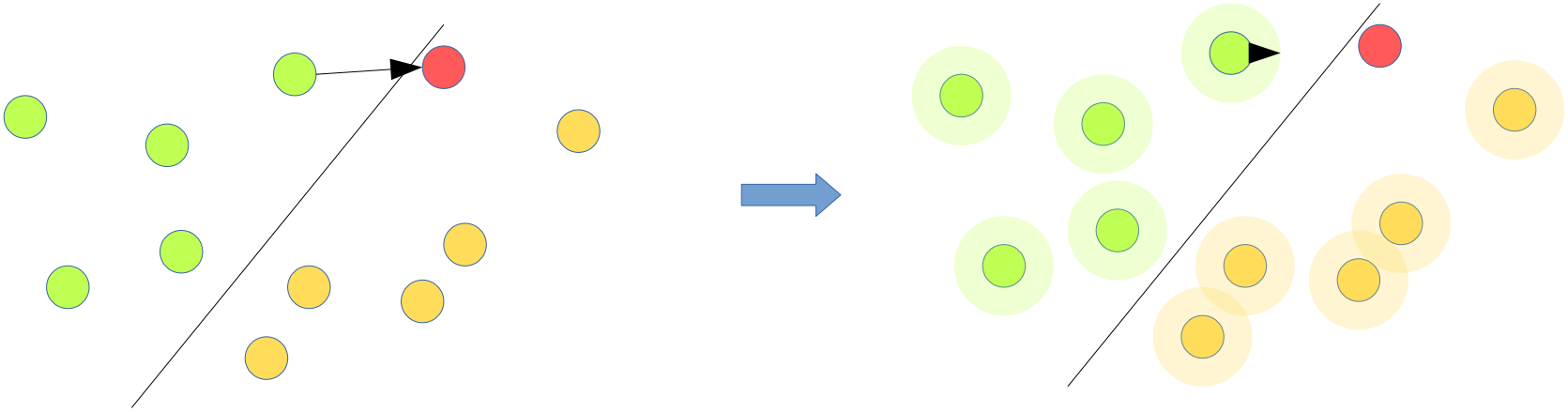


https://www.mdpi.com/2504-4990/1/1/11/htm

empoisonnement invisible

Clean label

Modification des images contraintes à être indétectable à l'oeil

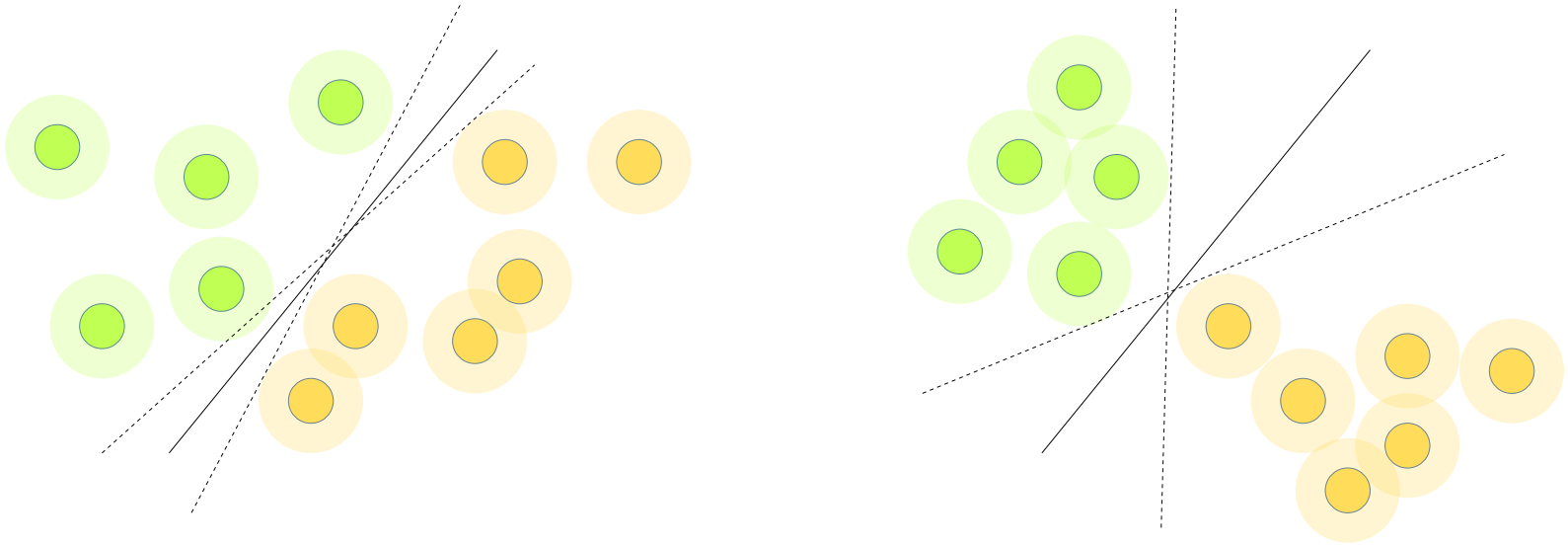


→ **Si l'apprentissage est robuste, y a-t-il toujours sensibilité à cet empoisonnement ?**

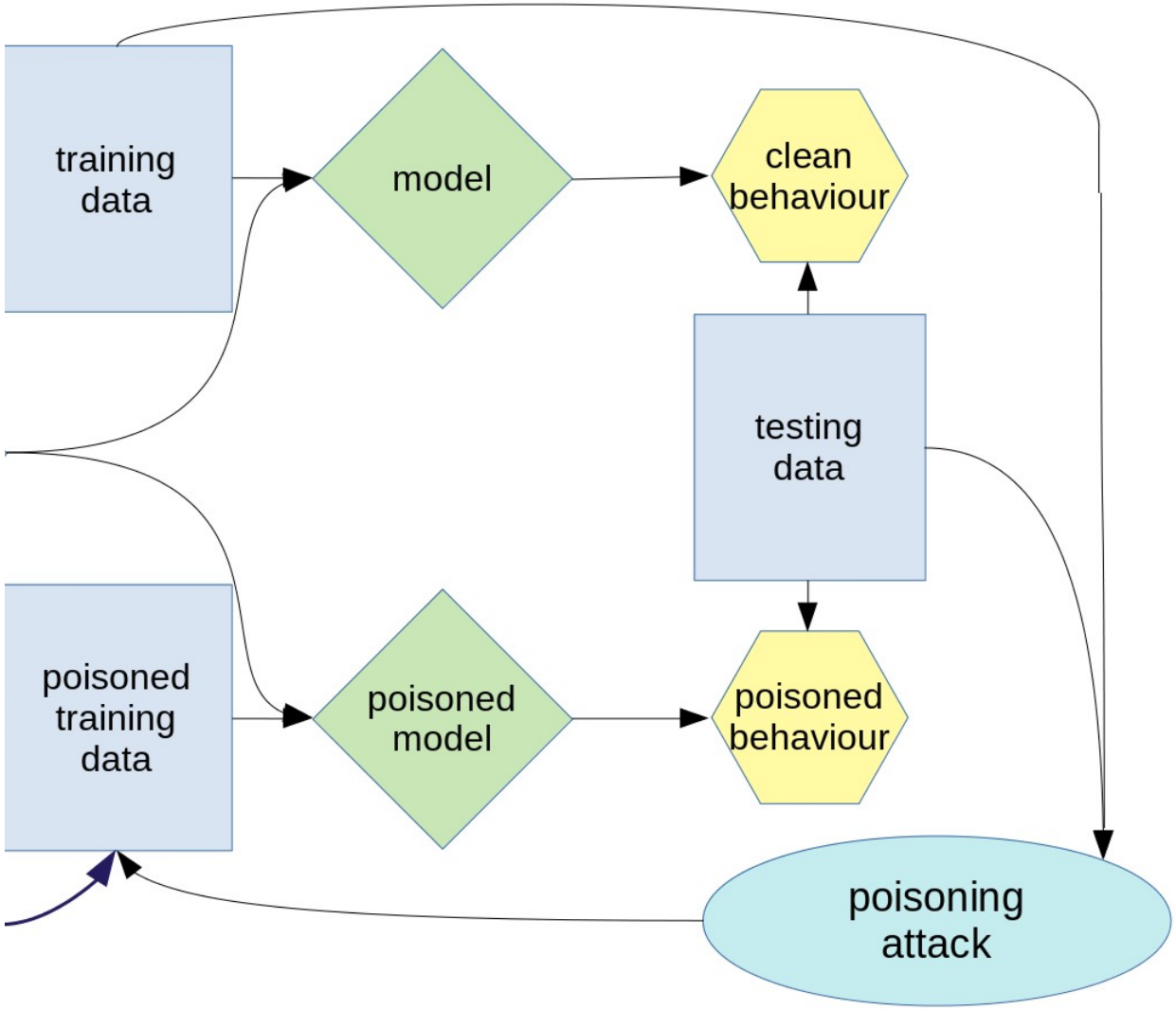
empoisonnement invisible et apprentissage robuste

→ **Si l'apprentissage est robuste, y a-t-il toujours sensibilité à cet empoisonnement ?**

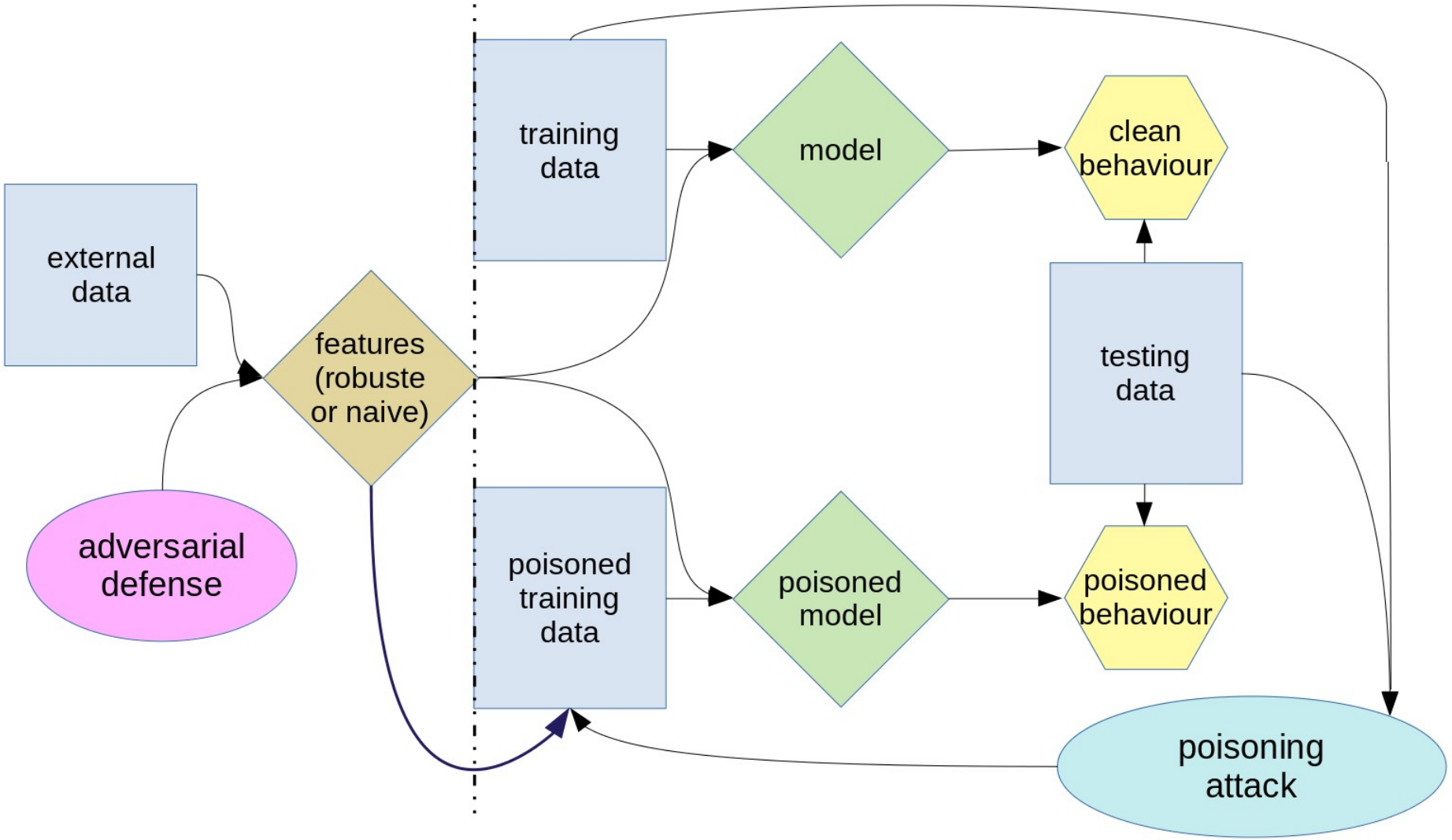
La question n'est pas tout à fait triviale :
on peut avoir des points margés mais plus ou moins dispersés



empoisonnement



empoisonnement invisible et apprentissage robuste



empoisonnement invisible et apprentissage robuste

Dataset	CIFAR	MNIST
AD on naive feature	24%	68%
AD on FSGM feature	30%	93%
AD on PGD feature	**34%**	**95%**

feature	CIFAR
PF on naive feature	85%
PF on FSGM feature	53%
PF on PGD feature	**16%**

Poison Frog (reimplémentation) passe de 85 % de succès à 16 % (CIFAR)

L'accuracy sous attaque AD qui avait chuté à 68 % remonte à 95 % (MNIST)

Criticité

Relecture de la base + « RANSAC » + caractéristique robuste

Relecture → protection contre trop de données fortement corrompues

RANSAC → protection contre peu de données fortement corrompues

Robuste → protection contre des données faiblement corrompues

Conclusion

Quelque soit le scénario, l’empoisonnement ne paraît pas un risque majeur.

Conclusion

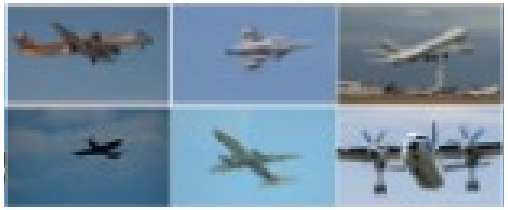
Quelque soit le scénario, l’empoisonnement ne paraît pas un risque majeur.

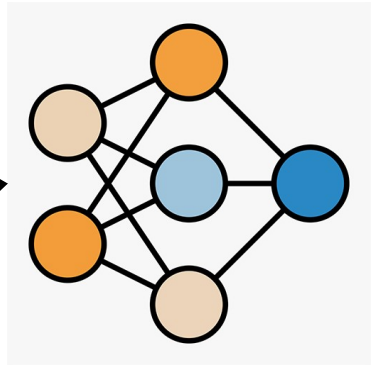
Sauf si l’apprentissage a lieu sans relecture de la base !

Perspectives
Les dangers du pseudo-labelling

Perspectives
Les dangers du pseudo-labelling

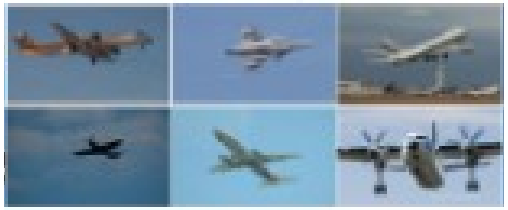
Domaine 1

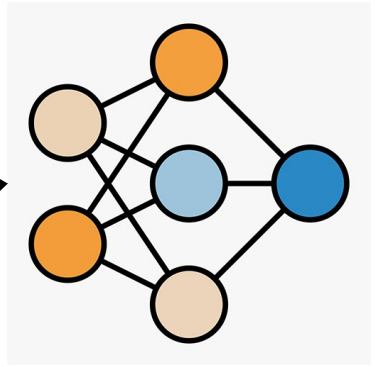




Perspectives
Les dangers du pseudo-labelling

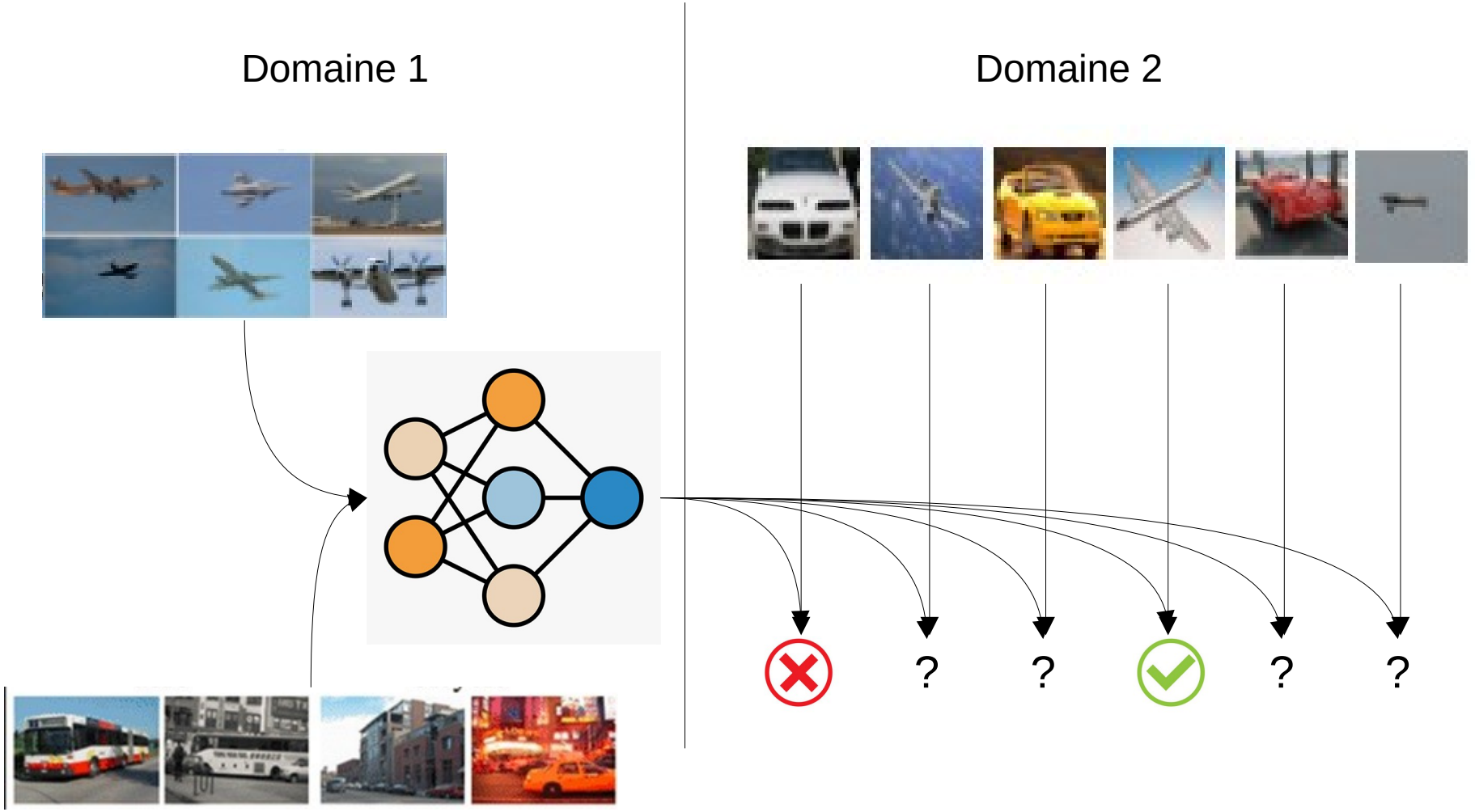
Domaine 1





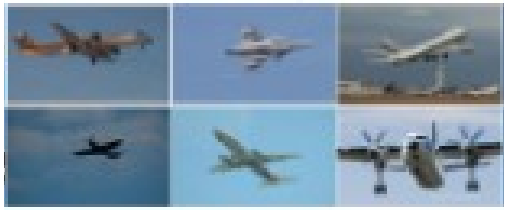
Domaine 2

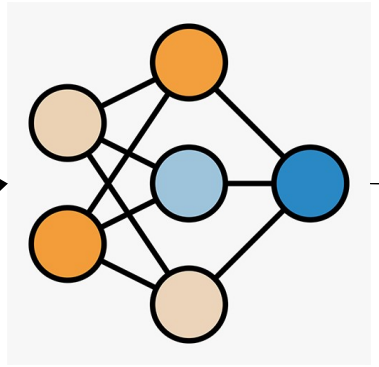
Perspectives
Les dangers du pseudo-labelling



Perspectives
Les dangers du pseudo-labelling

Domaine 1

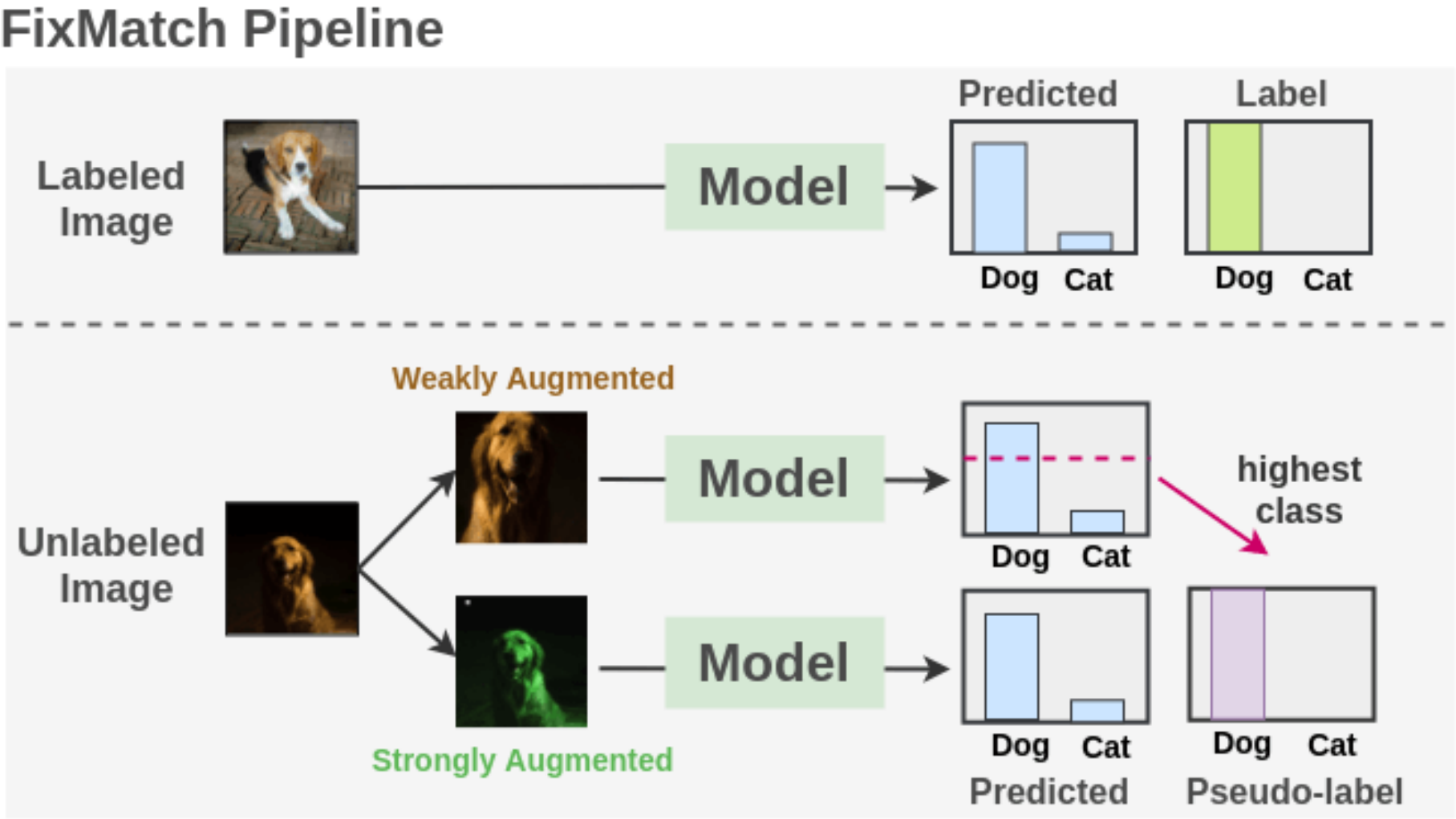




Domaine 2

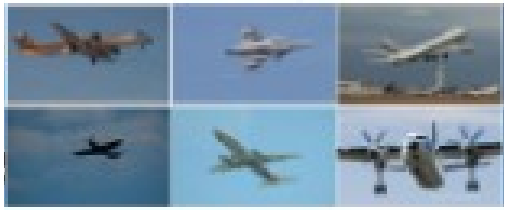
Confiance + consistance !

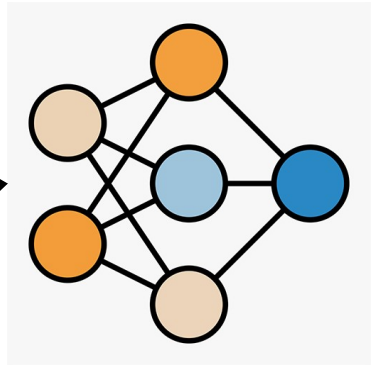
Perspectives
Les dangers du pseudo-labelling



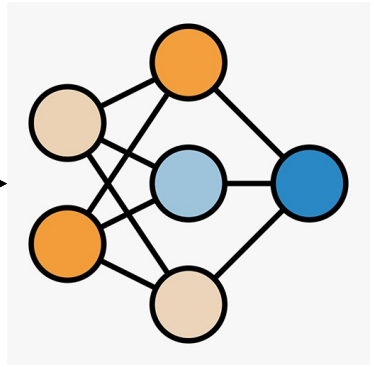
Perspectives
Les dangers du pseudo-labelling

Domaine 1





Domaine 2



Perspectives
Les dangers du pseudo-labelling

Est-il possible de créer une perturbation qui

Augmente la confiance

Résiste à un test de consistance

Perspectives
Les dangers du pseudo-labelling

Est-il possible de créer une perturbation qui

Augmente la confiance

Résiste à un test de consistance

Et qui une fois dans la base d'apprentissage provoque un empoisonnement ???

Perspectives
Les dangers du pseudo-labelling

Est-il possible de créer une perturbation qui

Augmente la confiance

Résiste à un test de consistance

Et qui une fois dans la base d'apprentissage provoque un empoisonnement ???

À suivre !

Perspectives
Les limites de ne considérer que « deep feature + SVM »

Là où les attaques adversaires sont apparues avec le DL,
l'empoisonnement est plus facile avec une approche type SVM !

Perspectives
Les limites de ne considérer que « deep feature + SVM »

Là où les attaques adversaires sont apparues avec le DL,
l'empoisonnement est plus facile avec une approche type SVM !

→ typiquement une approche SVM est beaucoup plus sensible au bruit des labels !
(en fonction de « C »)

Perspectives
Les limites de ne considérer que « deep feature + SVM »

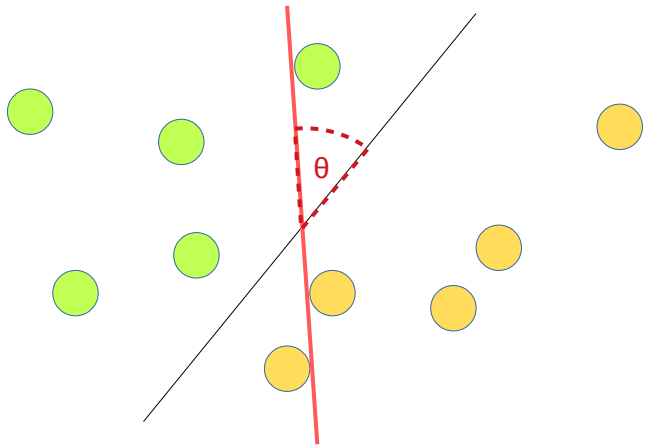
Là où les attaques adversaires sont apparues avec le DL,
l'empoisonnement est plus facile avec une approche type SVM !

→ typiquement une approche SVM est beaucoup plus sensible au bruit des labels !
(en fonction de « C »)

→ empoisonner un CNN est « plus difficile »
mais ouvre potentiellement la porte à d'autres conclusions …

Perspectives
Les limites de ne considérer que « deep feature + SVM »

Proxy based



https://www.mdpi.com/2504-4990/1/1/11/htm

Perspectives
Les limites de ne considérer que « deep feature + SVM »

proxy used Eq.3	testing accuracy	desired
$\mathrm{SGD}_\theta(f, Test)$	27%	$\ll$ 87% (-31% in [18])
$\mathrm{SGD}_\theta(f, Train)$	34%	$\approx$ 87% (0% in [18])
$-w_{imagenet}$	64%	$\approx$ 87%
$w_{imagenet}$	58%	$\approx$ 87%
$-\mathrm{SGD}_\theta(f, Train)$	73%	$\approx$ 87% (-1% in [18])
$-\mathrm{SGD}_\theta(f, Test)$	77%	$\gg$ 87% (+7% in [18])
Original accuracy	87%	-

Perspectives
Les limites de ne considérer que « deep feature + SVM »

proxy used Eq.3	testing accuracy	desired
$\mathrm{SGD}_\theta(f, Test)$	27%	$\ll$ 87% (-31% in [18])
$\mathrm{SGD}_\theta(f, Train)$	34%	$\approx$ 87% (0% in [18])
$-w_{imagenet}$	64%	$\approx$ 87%
$w_{imagenet}$	58%	$\approx$ 87%
$-\mathrm{SGD}_\theta(f, Train)$	73%	$\approx$ 87% (-1% in [18])
$-\mathrm{SGD}_\theta(f, Test)$	77%	$\gg$ 87% (+7% in [18])
Original accuracy	87%	-

Mais depuis Imagenet seulement (sinon ça marche comme prévu) !

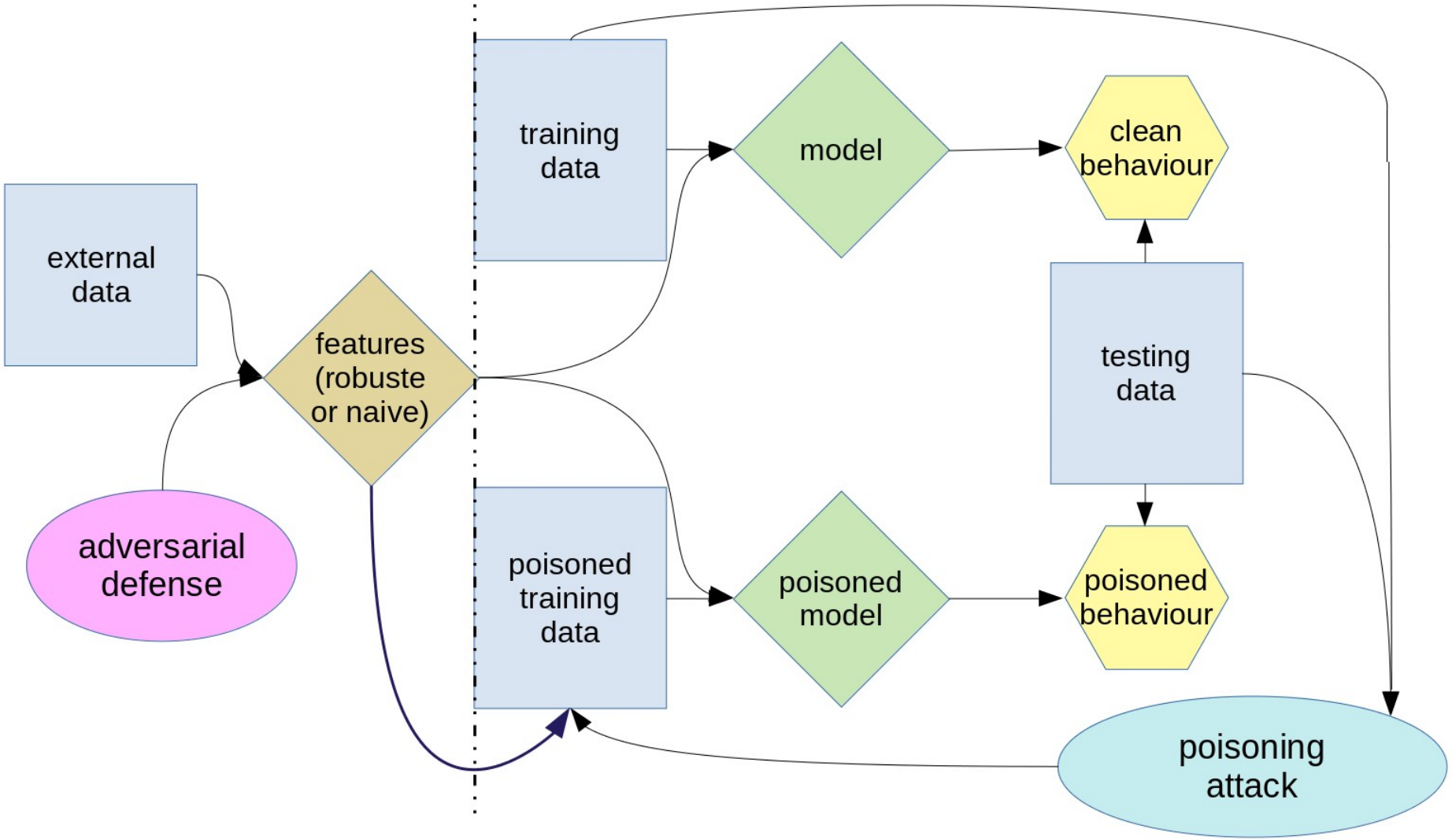
Perspectives
Les limites de ne considérer que « deep feature + SVM »

proxy used Eq.3	testing accuracy	desired
$\mathrm{SGD}_\theta(f, Test)$	27%	$\ll$ 87% (-31% in [18])
$\mathrm{SGD}_\theta(f, Train)$	34%	$\approx$ 87% (0% in [18])
$-w_{imagenet}$	64%	$\approx$ 87%
$w_{imagenet}$	58%	$\approx$ 87%
$-\mathrm{SGD}_\theta(f, Train)$	73%	$\approx$ 87% (-1% in [18])
$-\mathrm{SGD}_\theta(f, Test)$	77%	$\gg$ 87% (+7% in [18])
Original accuracy	87%	-

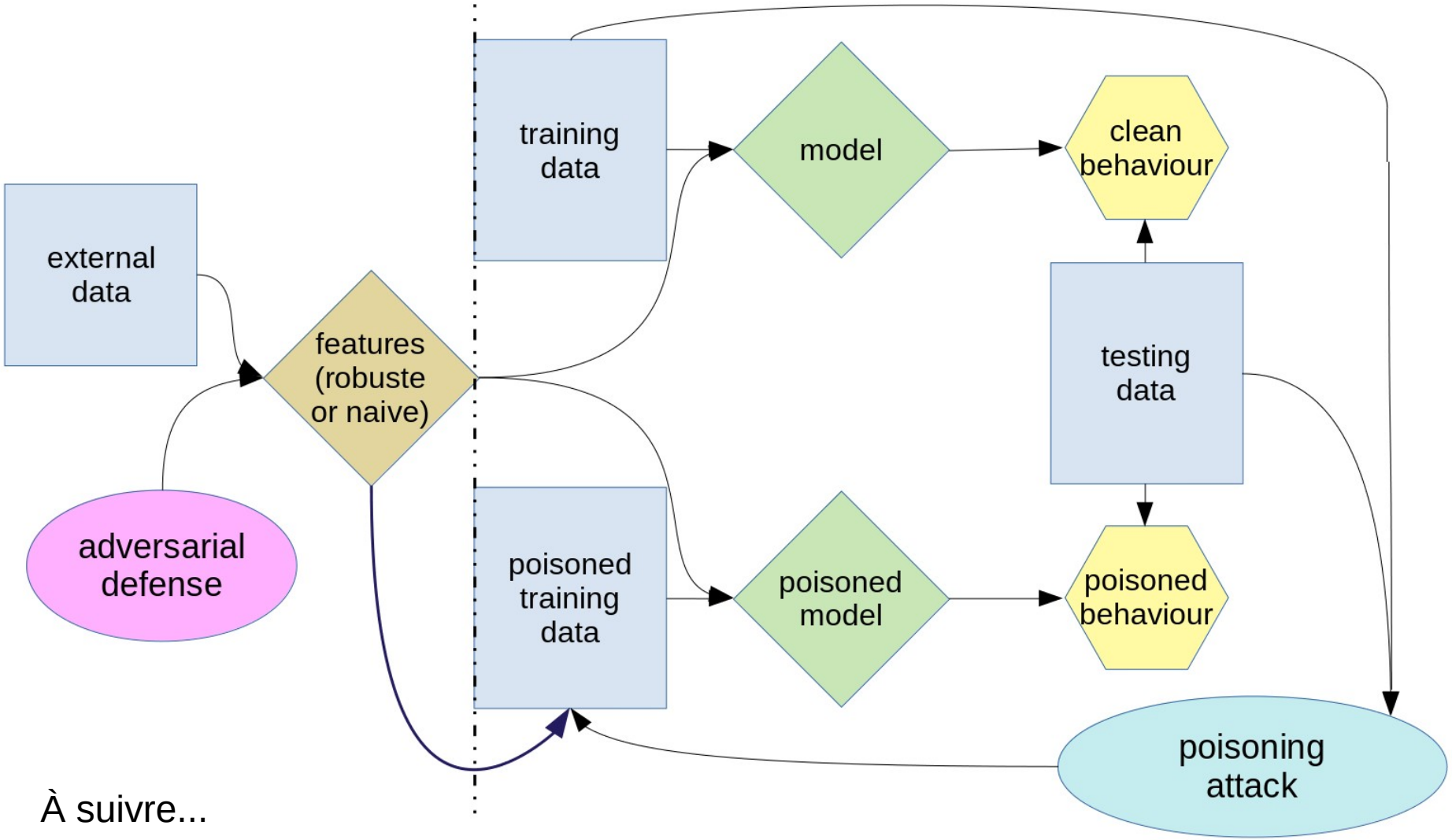
Mais depuis Imagenet seulement (sinon ça marche comme prévu) !

(l'impact sur le trajet est plus important que l'impact sur l'arrivé)

Perspectives
Les limites de ne considérer que « deep feature + SVM »



Perspectives
Les limites de ne considérer que « deep feature + SVM »



À suivre...

Merci pour votre attention.

Les codes sont disponibles à github.com/achanhon/AdversarialModel