# A journey of a Privacy attacks challenge

Alice Héliou
Vincent Thouvenot
Rodolphe Lampe
Cong-Bang Huynh
Baptiste Morisse

www.thalesgroup.com

# Content

01

## AI Friendly Hacker

02

## Privacy attack challenge

03

## Technical background

04

## Task A

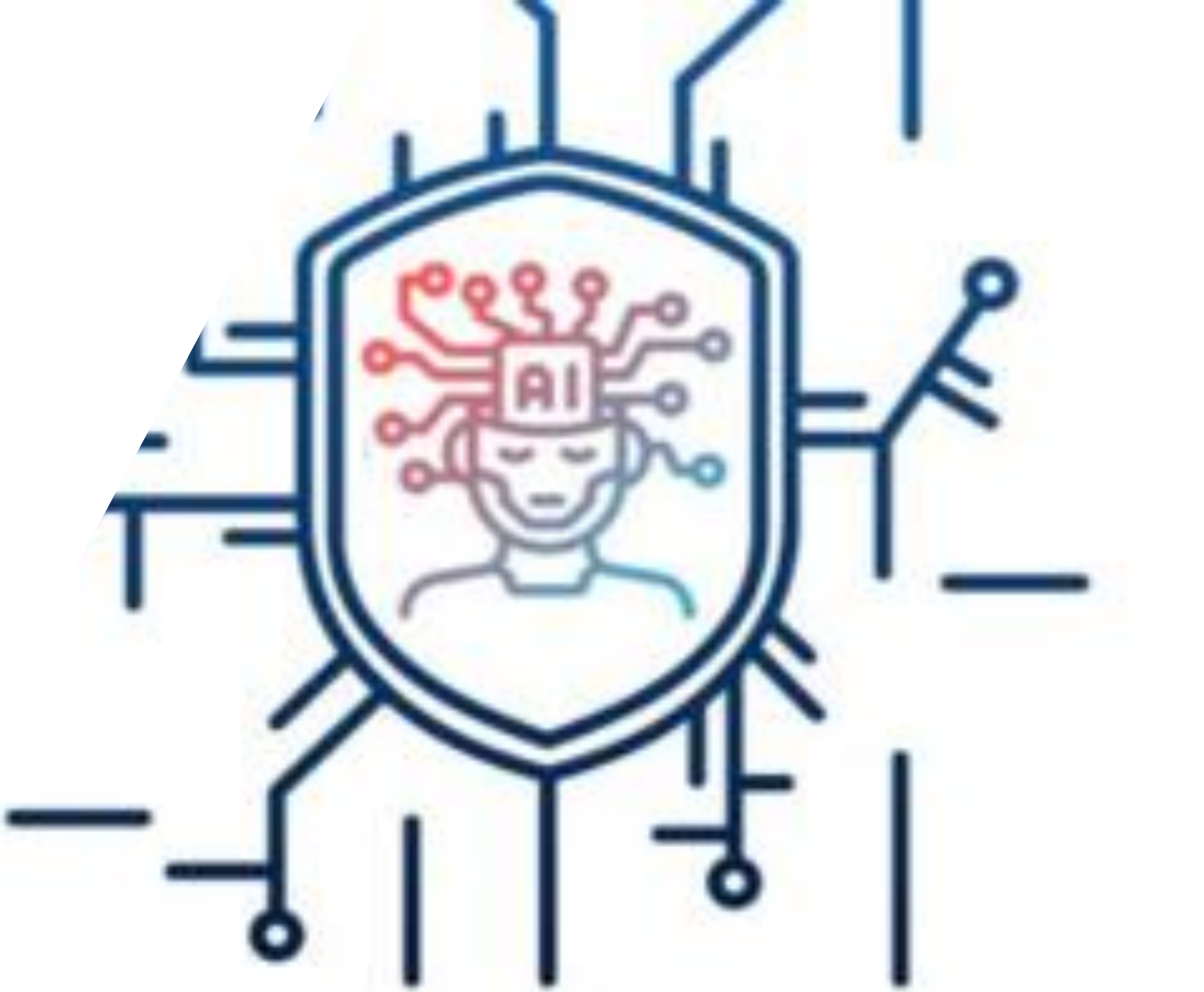
05

## Task B

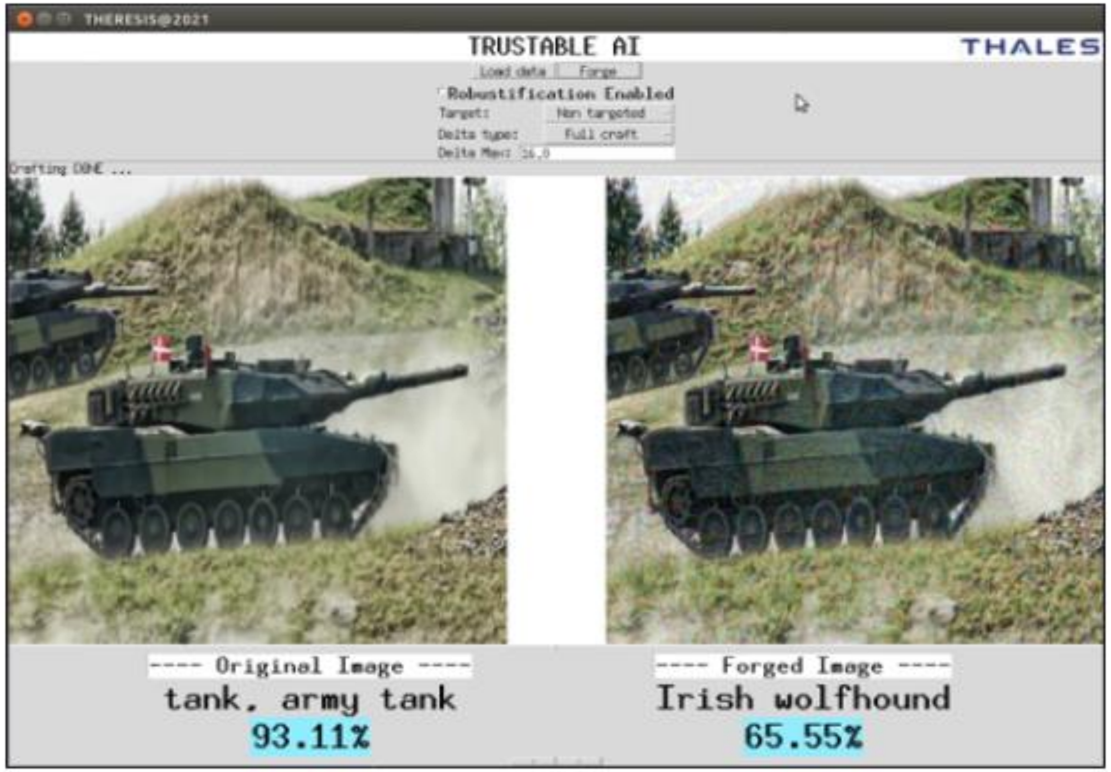
03

## Conclusion

# AI Friendly Hacker



# Evasion attack



TRUSTABLE AI

Robustification Enabled

Target: Non targeted

Delta type: Full craft

Delta Max: 16.0

Crafting DONE ...

---- Original Image ----

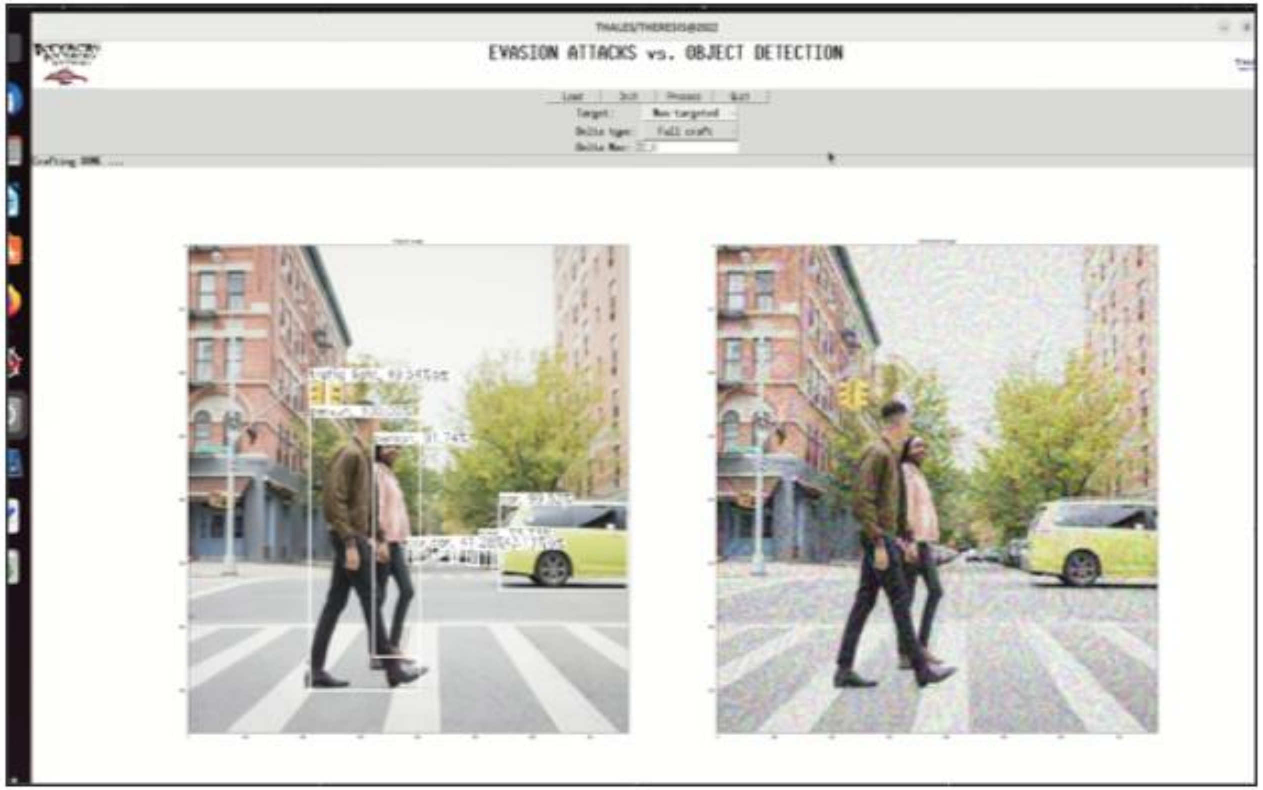
tank, army tank

93.11%

---- Forged Image ----

Irish wolfhound

65.55%



EVASION ATTACKS vs. OBJECT DETECTION

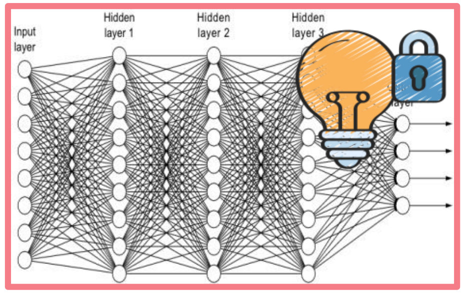
Target: Non targeted

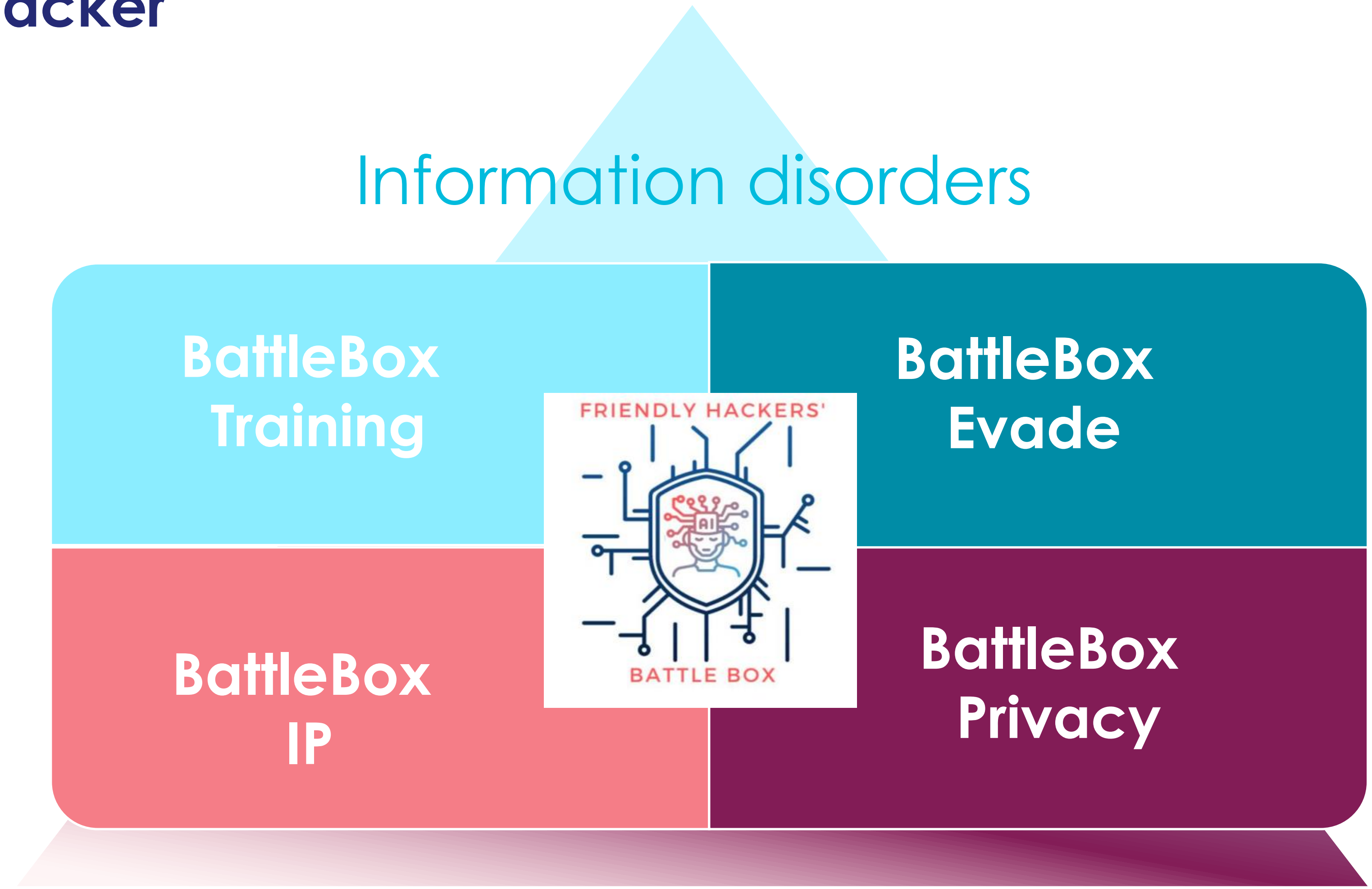
Delta type: Full craft

Delta Max: [illegible]

Crafting DONE ...

# AI Friendly Hacker





Information disorders

BattleBox Training

BattleBox Evade

BattleBox IP

BattleBox Privacy

IP/Copyright infringment

Breach of confidentiality

# A Privacy attacks challenge

caid

by DGA

# Team

Alice Héliou

Vincent Thouvenot

Rodolphe Lampe

Cong Bang Huyhn

Baptiste Morisse

# Context

## > Proposed by Direction Générale de l’Armement

## > Conference on Artificial Intelligence for Defence at Rennes end of november

## > Data and model

- Aircraft FGVC (Fine Grained Visual Classification)
  - 10200 plan images
  - 70 classes
  - Fine Grained Visual Classification of Aircraft, Majiet al., 2013
- Architecture of the target model: ResNet50

DC-8

Boeing 737

DC-9

MD-11

Boeing 717

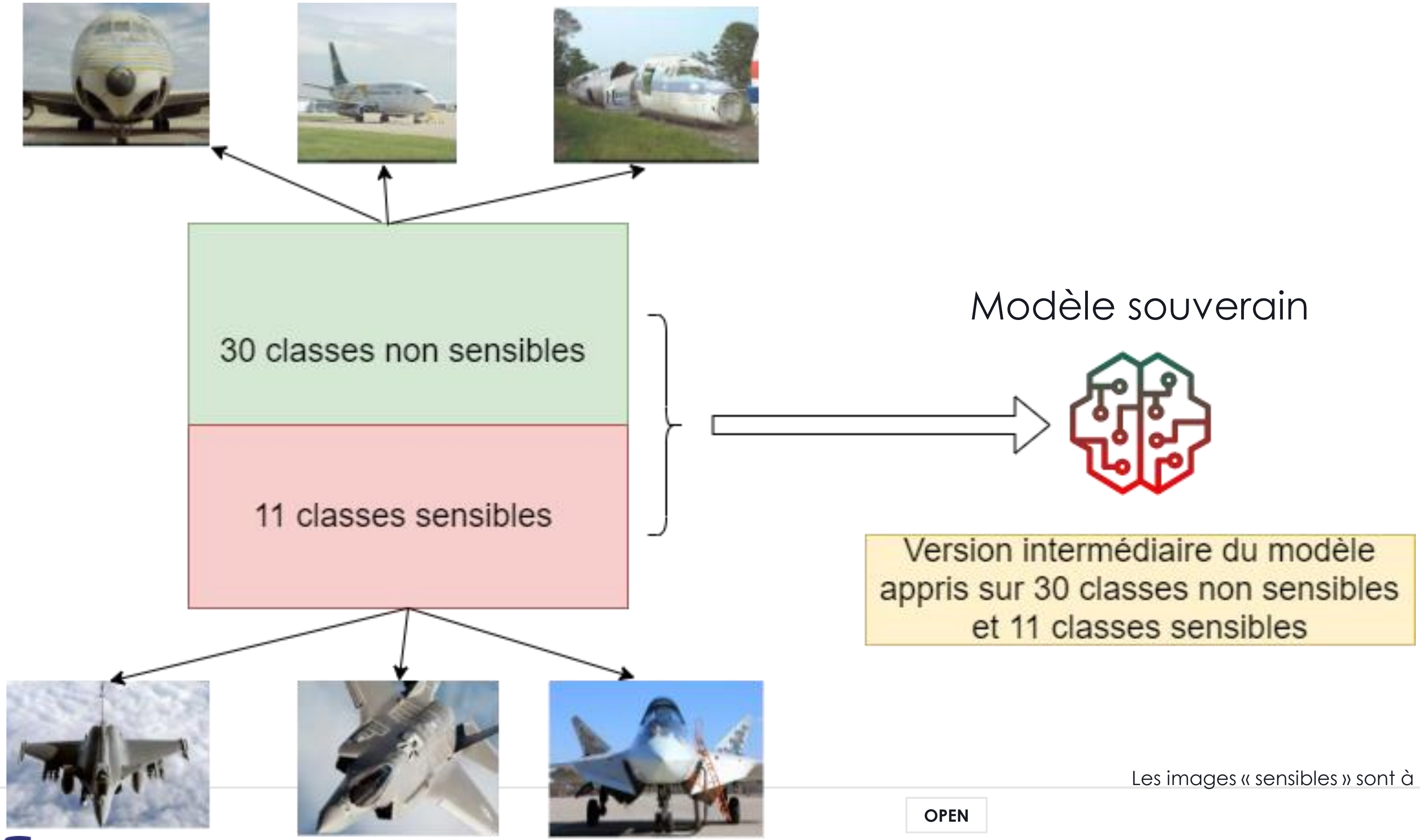
Gulfstream

## > Study of AI vulenrabilities with privacy attacks

- Two task
  - Membership Inference Attack
  - Foregetting Attack (detailed below)
- Challenge procedure
  - Two submissions by month and by tasks between May and September
  - Update of a leaderboard according the accuracy of attacks each month
- https://caid-conference.eu/challenge/

# Tâche B: Forgetting Attack

- > Le modèle fourni a été appris en 2 phases
- > Dans la 1ère phase 11 classes jugées sensibles ont été utilisées pour l'apprentissage



30 classes non sensibles

11 classes sensibles

Modèle souverain

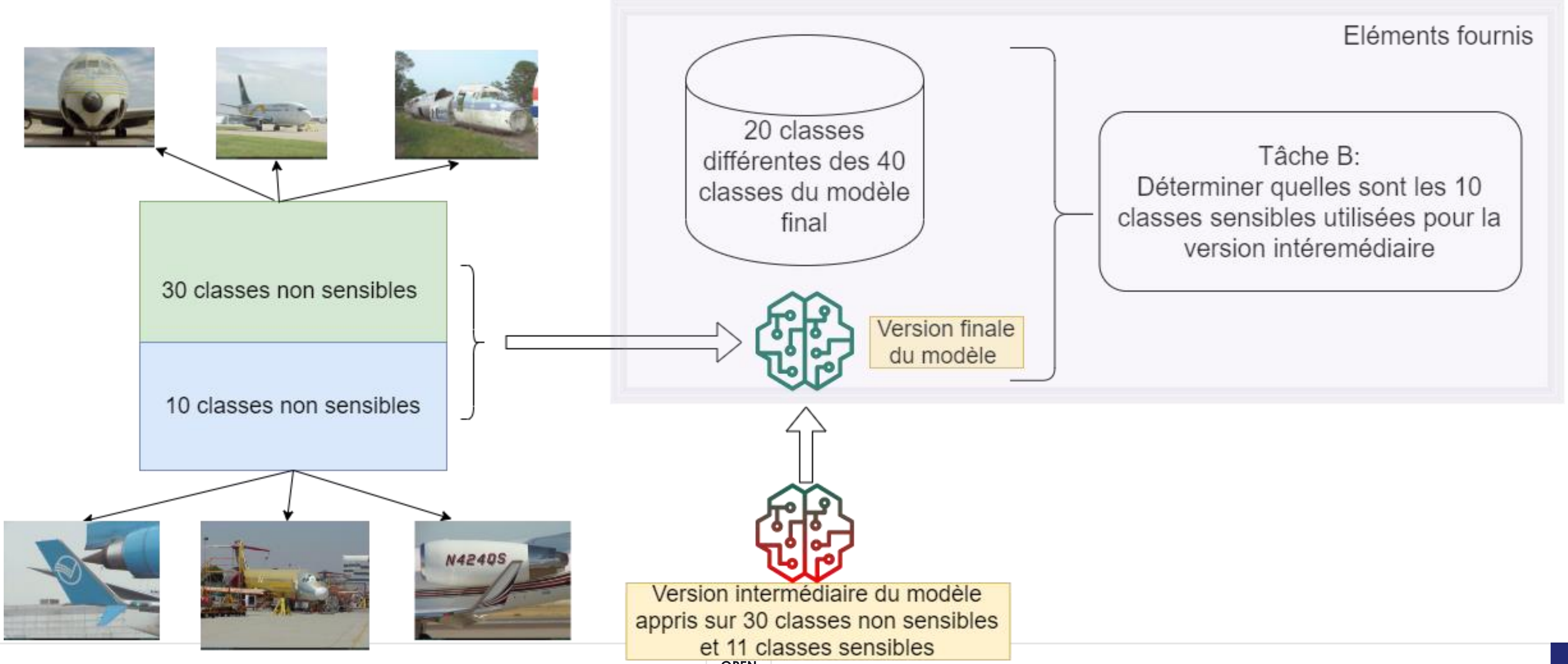
Version intermédiaire du modèle appris sur 30 classes non sensibles et 11 classes sensibles

Les images « sensibles » sont à titre d'illustration, elles ne font pas parties du dataset

# Tâche B: Forgetting Attack

> **Dans la 2$^{nde}$ phase l'apprentissage est poursuivi en remplaçant les 11 classes sensibles, par 10 autres classes**

**Le modèle final est le sujet de l'attaque, l'objectif étant de retrouver les 10 classes sensibles parmi les 20 fournies**



Eléments fournis

20 classes différentes des 40 classes du modèle final

Tâche B:
Déterminer quelles sont les 10 classes sensibles utilisées pour la version interémédiaire

30 classes non sensibles

10 classes non sensibles

Version finale du modèle

Version intermédiaire du modèle appris sur 30 classes non sensibles et 11 classes sensibles

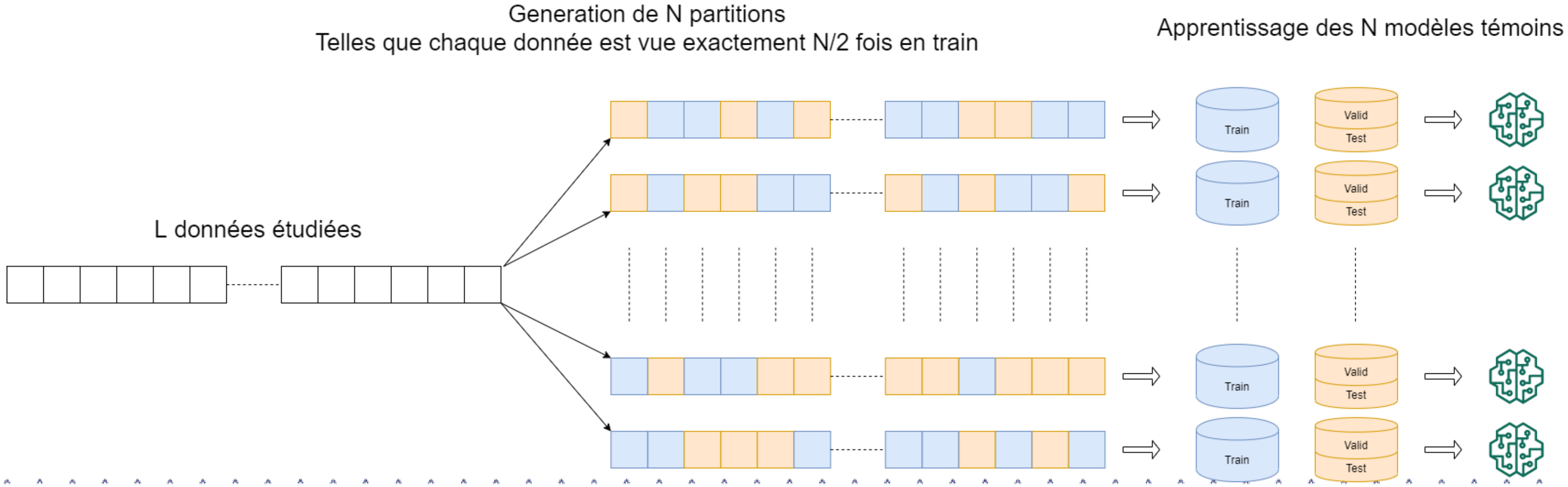
# Technical background

OPEN

# Shadow models

> DL model that aim to copy the behavior of the target model

> Train of different data partition of the provided dataset

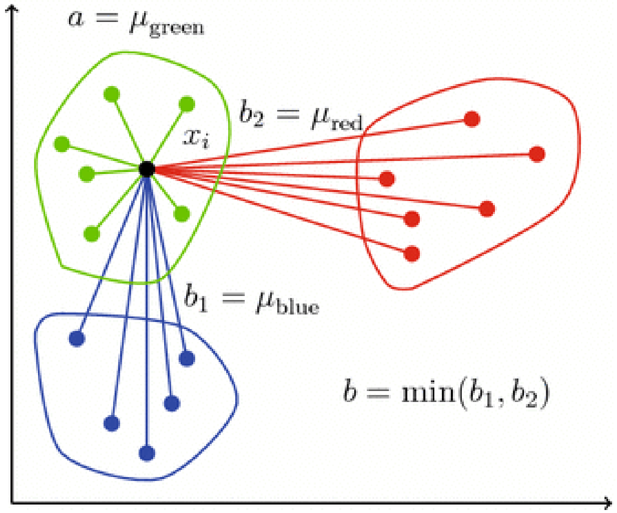

Generation de N partitions
Telles que chaque donnée est vue exactement N/2 fois en train
Apprentissage des N modèles témoins
L données étudiées
Train
Valid
Test

# Silhouette Coefficient

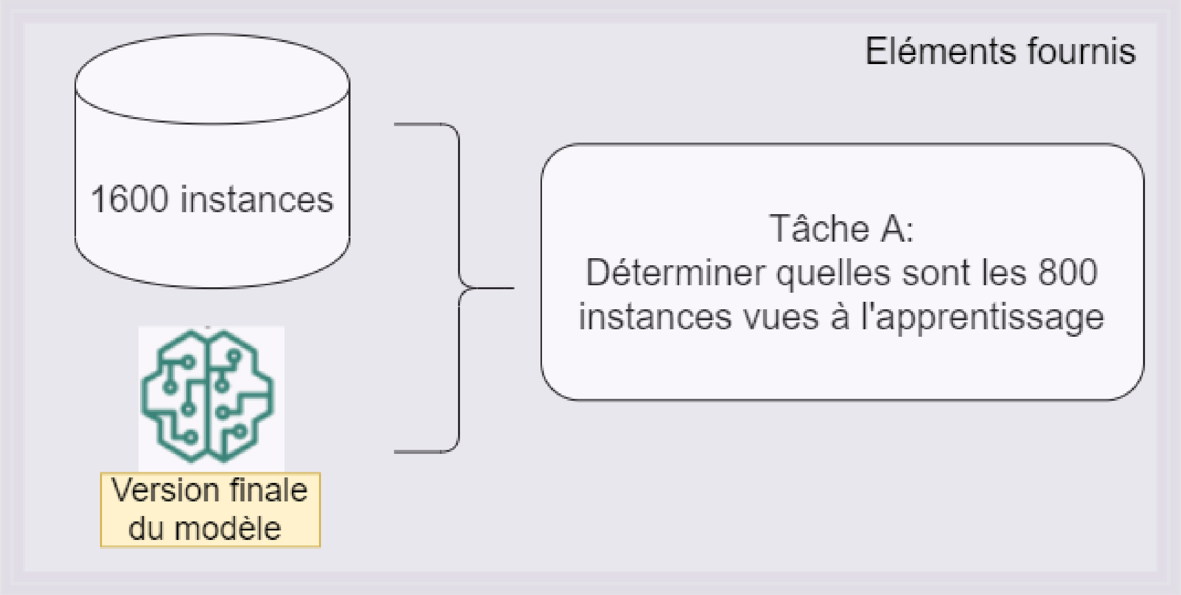
## > Used to evaluate the quality of clustering

## > Interpretion:

- Negative value: the point is in average closer of a another cluster than the one it is
- Positive value: the point is in average closer of its cluster than the other cluster
- Stronger it is, better it is



# Tâche A: Membership Inference attack


Eléments fournis
1600 instances
Version finale du modèle
Tâche A:
Déterminer quelles sont les 800 instances vues à l'apprentissage

# Membership Inference Attack

## Naive approach

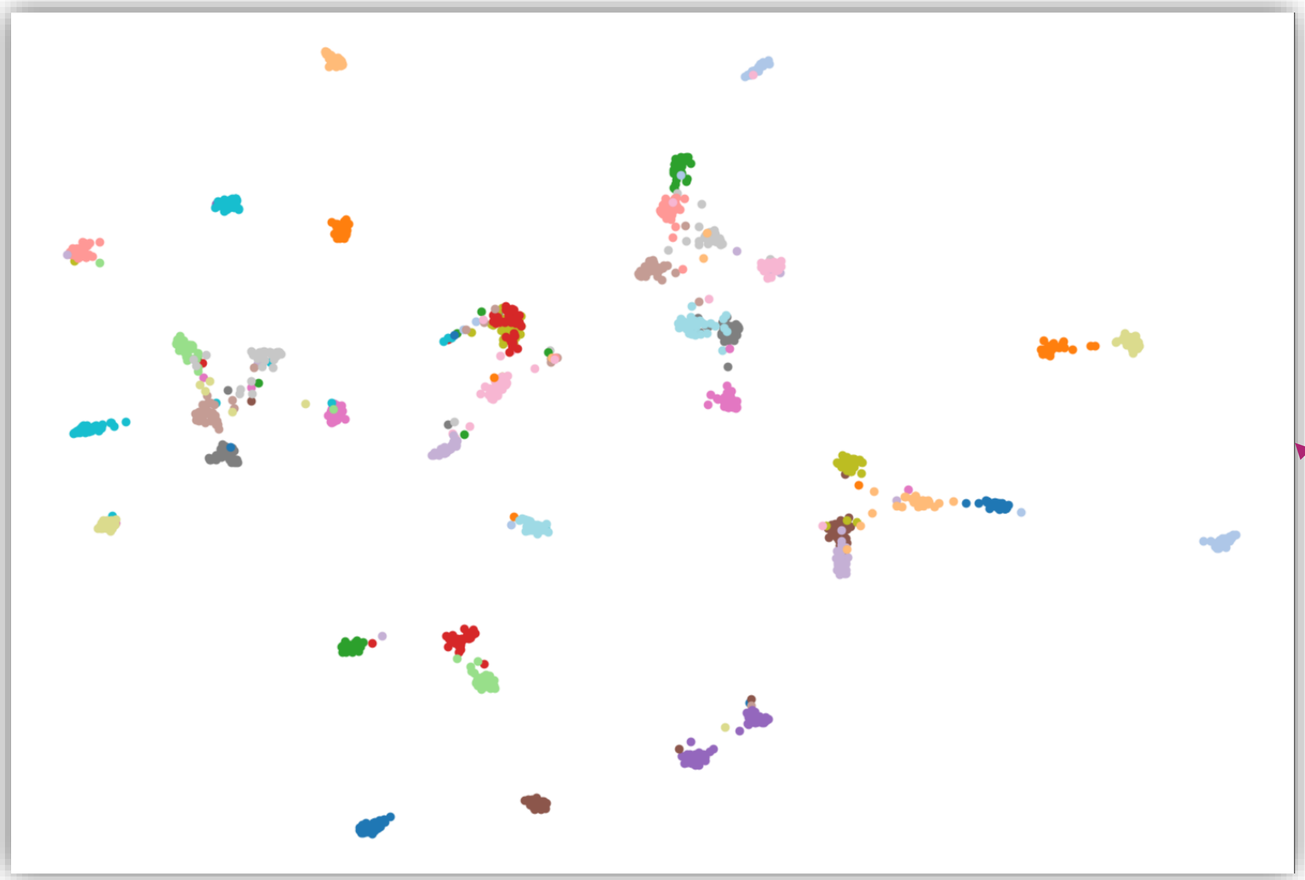
- Submission with
  - "train" with well classified observation
  - "test" with misclassified observation
- Do not match the distribution 800 train and 800 test
- **Accuracy 56%**

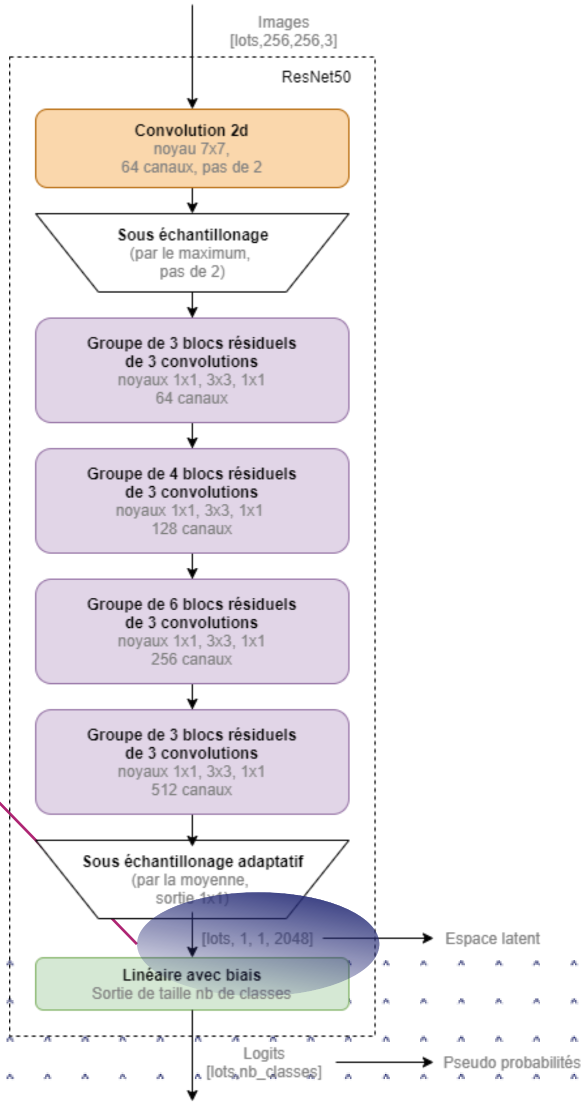
## Information from this submission:

- Training set accuracy: 96%
- Testing set accuracy: 84%
- **Target model do not generalize well**

## 10/39 submissions are worst than this naïve submission

# Task A: Latent Space 2D projection





Images
[lots,256,256,3]

ResNet50

**Convolution 2d**
noyau 7x7,
64 canaux, pas de 2

**Sous échantillonage**
(par le maximum,
pas de 2)

**Groupe de 3 blocs résiduels de 3 convolutions**
noyaux 1x1, 3x3, 1x1
64 canaux

**Groupe de 4 blocs résiduels de 3 convolutions**
noyaux 1x1, 3x3, 1x1
128 canaux

**Groupe de 6 blocs résiduels de 3 convolutions**
noyaux 1x1, 3x3, 1x1
256 canaux

**Groupe de 3 blocs résiduels de 3 convolutions**
noyaux 1x1, 3x3, 1x1
512 canaux

**Sous échantillonage adaptatif**
(par la moyenne,
sortie 1x1)

[lots, 1, 1, 2048]

Espace latent

**Linéaire avec biais**
Sortie de taille nb de classes

Logits
[lots,nb_classes]

Pseudo probabilités

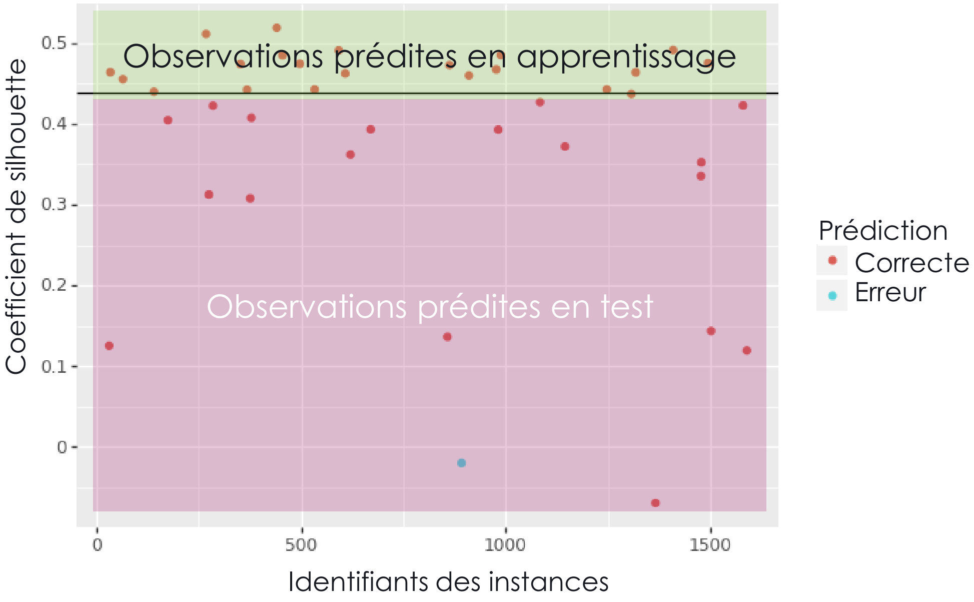
OPEN

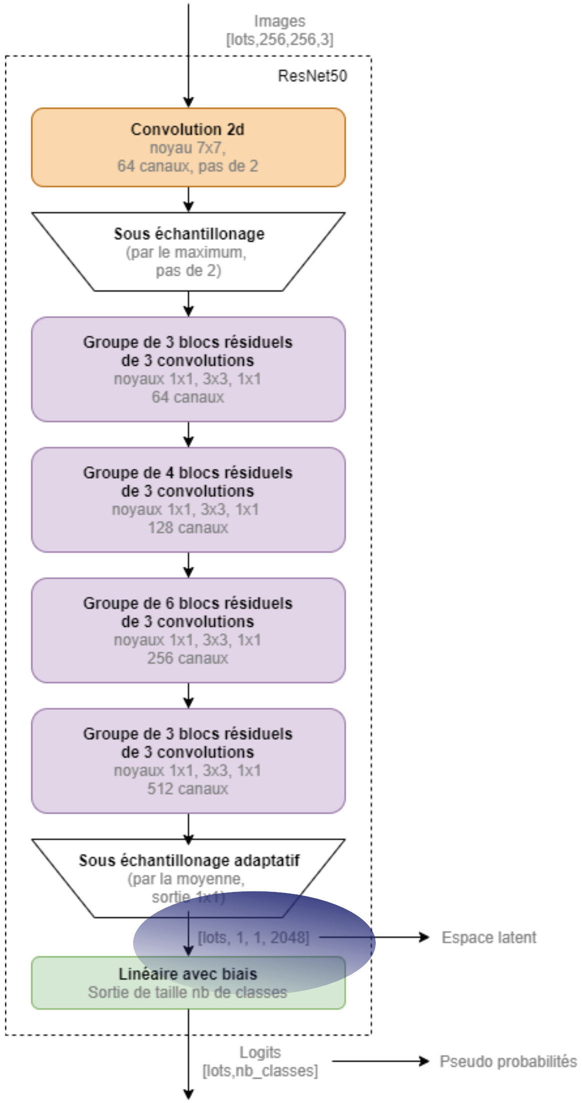
# Focus on latent space

## > Computation of Silhoeutte coefficient

- For each class, instances whose Silhouette is greater than the median predict as « train », other as « test

- Falcon 900 class





## > 57,4 % (23/39)

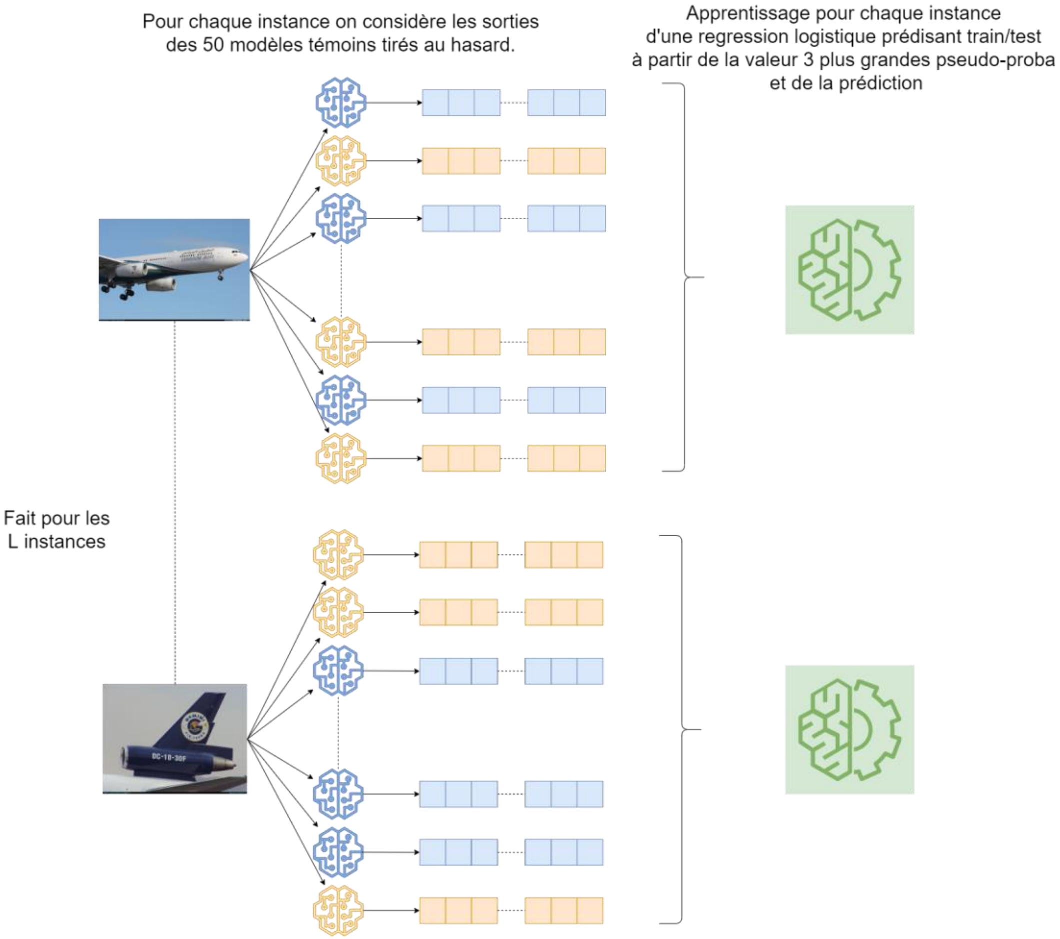
# Based on model outputs and shadow model

## > 101 partitions of shadows models

- 50 for training attack, one partition always used for test
- For each image and each sample of 50 shadow models, training of a logistic regression
- Vote of the logisitic regression

## > Accuracy on the shadow model always in test: 66%

## > Accuracy on the target model 56%


Pour chaque instance on considère les sorties des 50 modèles témoins tirés au hasard.
Apprentissage pour chaque instance d'une regression logistique prédisant train/test à partir de la valeur 3 plus grandes pseudo-proba et de la prédiction
Fait pour les L instances

OPEN

# Shadow model improvement

## > Shadow models training without augmentation

## > Add variability in the training process of shadow model

- Optimizr, learning rate, epoch
- The more shadow models are differnet, the more some can be close to the target model
- More different model = more ability to the attack to generalize

## > Take times…

# Results

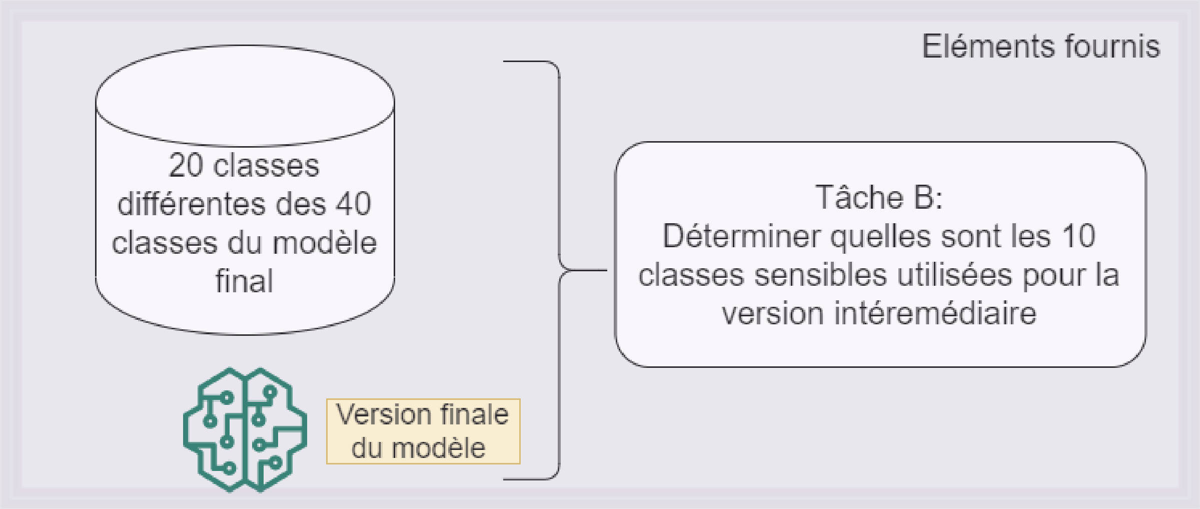
> **Final approaches used will be presented at CAID**

> **Leaderboard**

- 10 teams, 39 submissions

| Team | Month | Acc. |
|---|---|---|
| **Friendly Hackers** | **September** | **0.65** |
| **Friendly Hackers** | **September** | **0.64** |
| **Friendly Hackers** | **August** | **0.64** |
| HackCuda MaData | August | 0.62 |
| HackCuda MaData | July | 0.61 |
| **Friendly Hackers** | **August** | **0.61** |
| HAL9000 | September | 0.59 |

# Tâche B: Forgetting attack



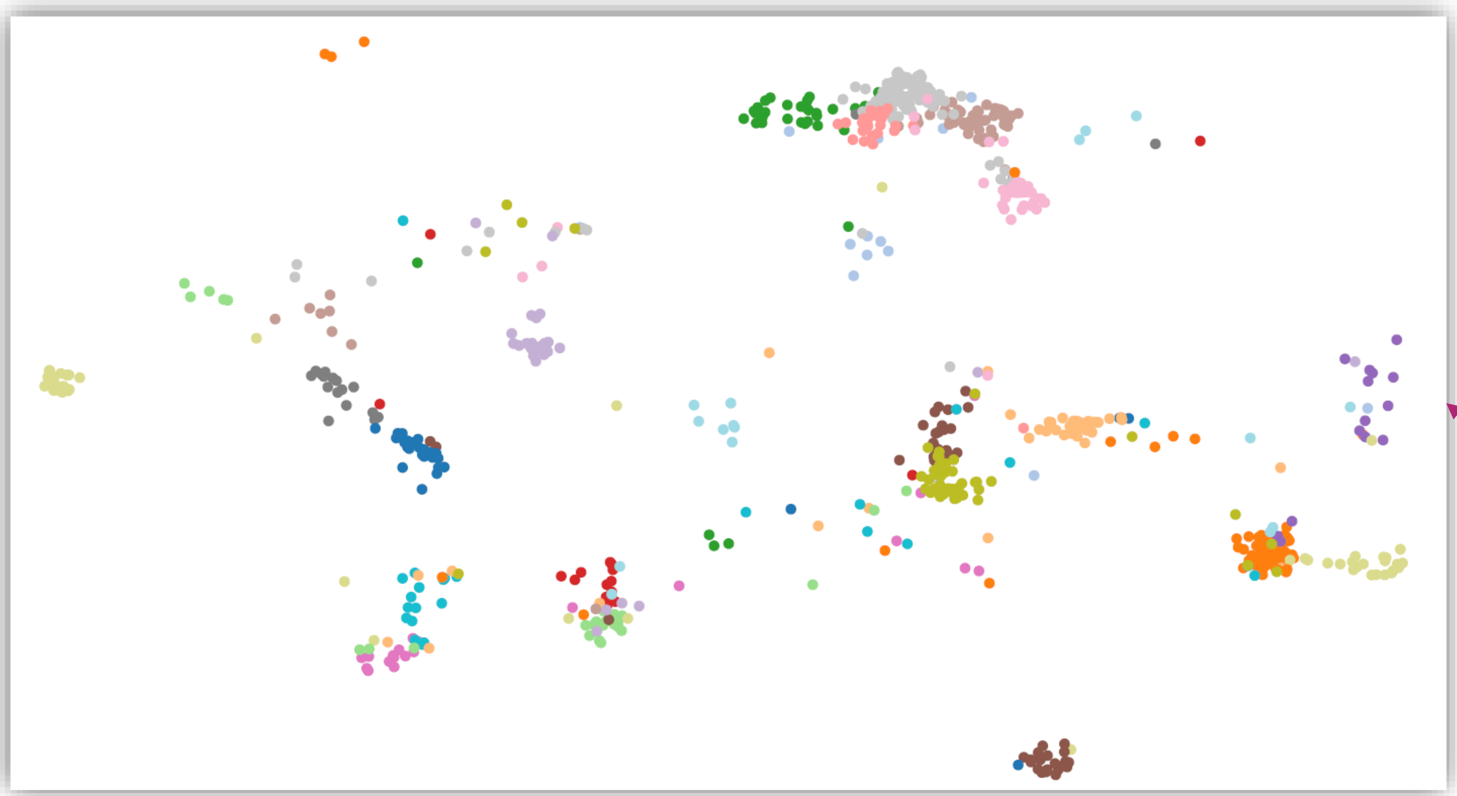
Eléments fournis

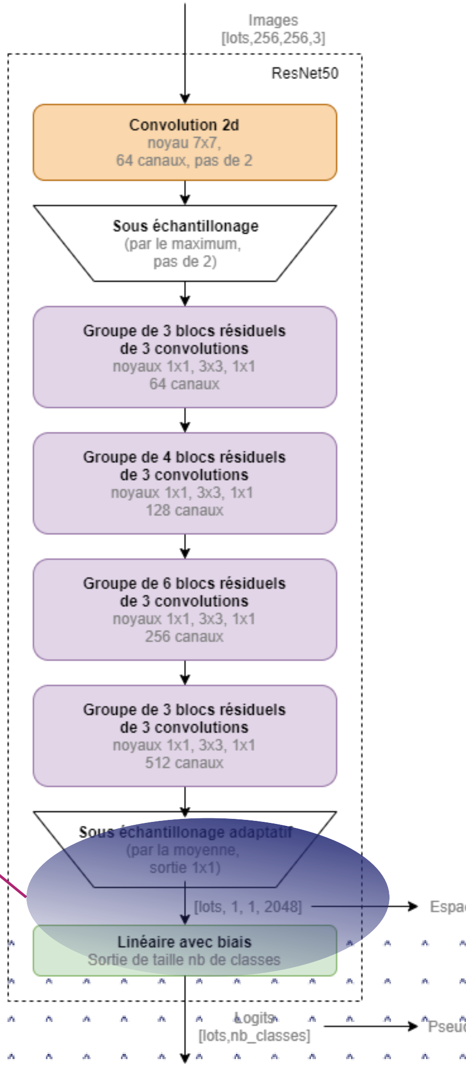
20 classes différentes des 40 classes du modèle final

Version finale du modèle

Tâche B:
Déterminer quelles sont les 10 classes sensibles utilisées pour la version intéremédiaire

# Tache B: Latent Space 2D projection





Images
[lots,256,256,3]

ResNet50

**Convolution 2d**
noyau 7x7,
64 canaux, pas de 2

**Sous échantillonage**
(par le maximum,
pas de 2)

**Groupe de 3 blocs résiduels
de 3 convolutions**
noyaux 1x1, 3x3, 1x1
64 canaux

**Groupe de 4 blocs résiduels
de 3 convolutions**
noyaux 1x1, 3x3, 1x1
128 canaux

**Groupe de 6 blocs résiduels
de 3 convolutions**
noyaux 1x1, 3x3, 1x1
256 canaux

**Groupe de 3 blocs résiduels
de 3 convolutions**
noyaux 1x1, 3x3, 1x1
512 canaux

**Sous échantillonage adaptatif**
(par la moyenne,
sortie 1x1)

[lots, 1, 1, 2048]

Espa

**Linéaire avec biais**
Sortie de taille nb de classes

Logits
[lots,nb_classes]

Pseu

OPEN

THALES
Building a future we can all trust

# Task B: First use of Silhouette Coefficient

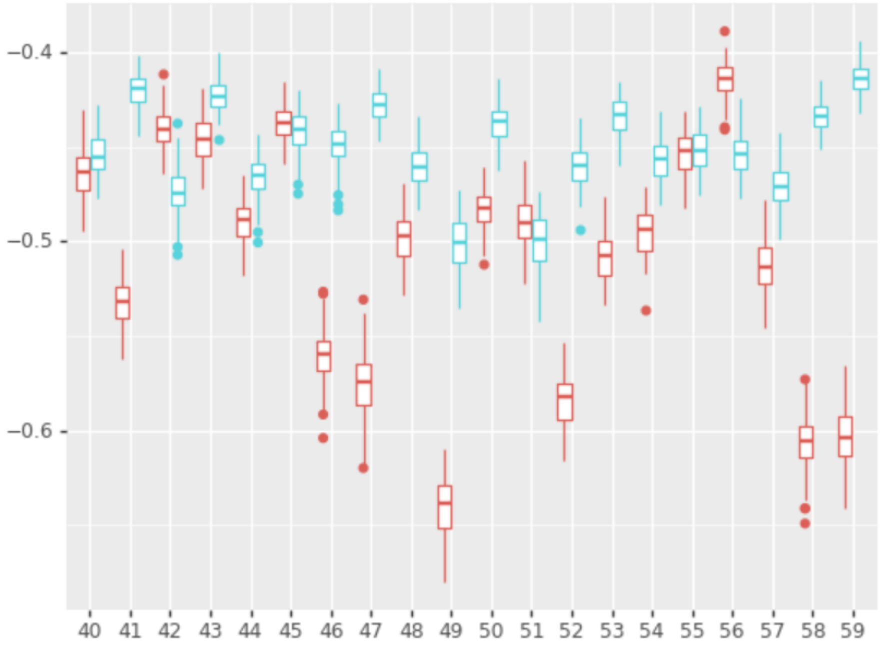
## > Not-complex method:

- Building of interval with 1 sigma, 2 sigma and 3 sigma rules around the median of Silhouettes coefficient of the shadows models for each class
- Computation of the distance between the median of the Silhouettes coefficient of the target model and the previous interval

## > 14 classes correct on 20

# Task B: More complex approaches use of Silhouettes coefficients

> Isolation Forest for each model for anomaly detection for each class using the Silhouettes coefficient (40 per classes)



# Task B: More complex approaches use of Silhouettes coefficients

- Isolation Forest for each model for anomaly detection for each class using the Silhouettes coefficient (40 per classes)

- Improvement of accuracy on our test

# Task B: More complex approaches use of Silhouettes coefficients

- Isolation Forest for each model for anomaly detection for each class using the Silhouettes coefficient (40 per classes)
- Improvement of accuracy on our test
- But… decrease on the target model
- Decrease due among others to the shadow models quality

# Results

## > Final approaches used will be presented at CAID

## > Leaderboard

- 3 teams

| Equipe | Mois | Acc. |
|---|---|---|
| **Friendly Hackers** | **September** | **1** |
| **Friendly Hackers** | **June** | **0.70** |
| **Friendly Hackers** | **September** | **0.70** |
| **Friendly Hackers** | **July** | **0.65** |
| **Friendly Hackers** | **July** | **0.60** |
| JCVD | July | 0.60 |
| Benaroya | August | 0.60 |

# Conclusion

OPEN

# Conclusion

## > A wealth of learning opportunities

- Collaborative work
- State of the Art both rich and incomplete, especially for real-life attacks
- Very complex to make “smart” shadow models

## > Open new perspective at Thales

- Implement Privacy attack on Thales use case
- New thematics: Machine Unlearning
  - 2 internships open
    - Blue Team: unlearning efficiently information in a Deep Learning model
    - Red Team: attack unlearning approaches