GDR IASIS – 11/06/2024

AI Friendly Hacker: when an AI reveals more than it should...

Héliou Alice, Thouvenot Vincent,
Lampe Rodolphe, Huynh Cong Bang,
Morisse Baptiste
THALES

www.thalesgroup.com

Context

CHALLENGE PROPOSED BY DGA (FRENCH MOD)

STUDY VULNERABILITIES OF AI

Introduction

> **CAID : French conference about IA for Defence, organiséed by DGA the 22th and 23th November 2023 (Rennes, France)**

> **Topics**

- Application of AI for Defence use case
- Focus about robustness, certification, explicability of embarked AI systems

> **Two tasks for an unique AI privacy challenge!**

- Membership Inference attack
- (Un)Forgetting attack
- Two submissions for each task by months between May and September, with an updated leaderboard after each submission

https://caid-conference.eu/challenge/

> **FGVC Dataset – 10 200 aircraft images**

- 70 different classes
- Fine Grained Visual Classification of Aircraft, Majiet *al.*, 2013

DC-8

Boeing 737

DC-9

MD-11

Boeing 717

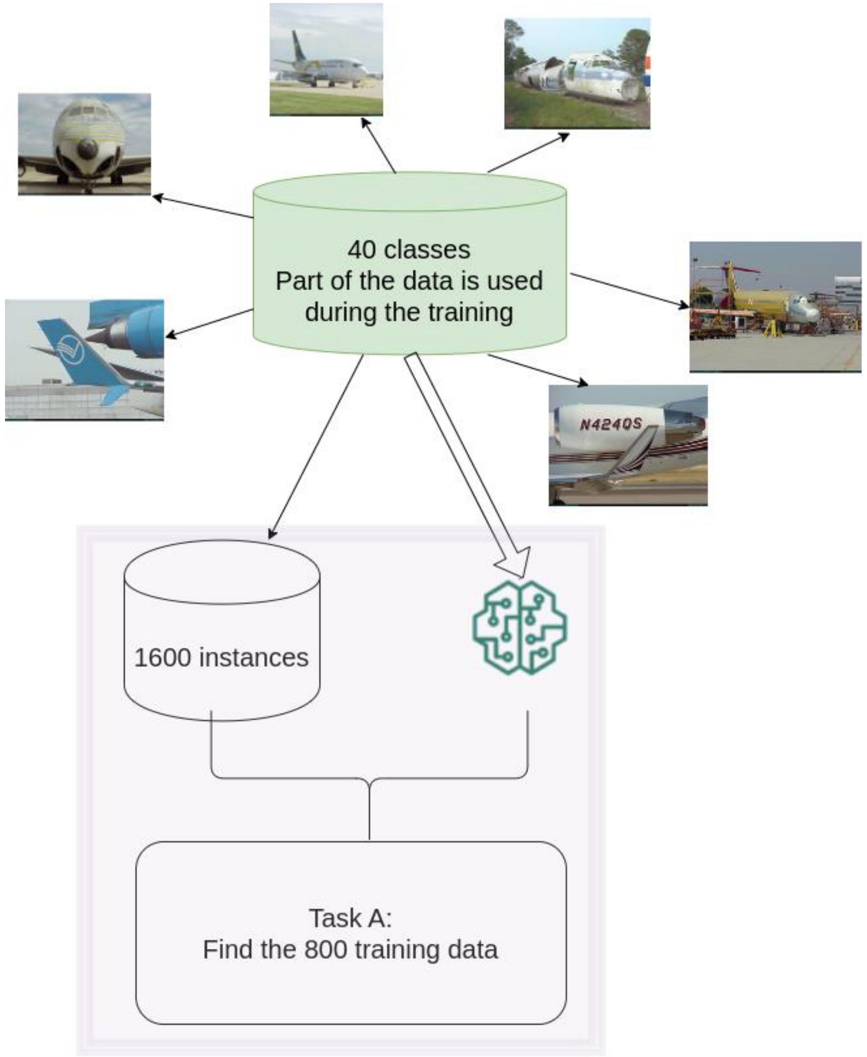
Gulfstream

©James Richard Covington JR / Airliners

OPEN

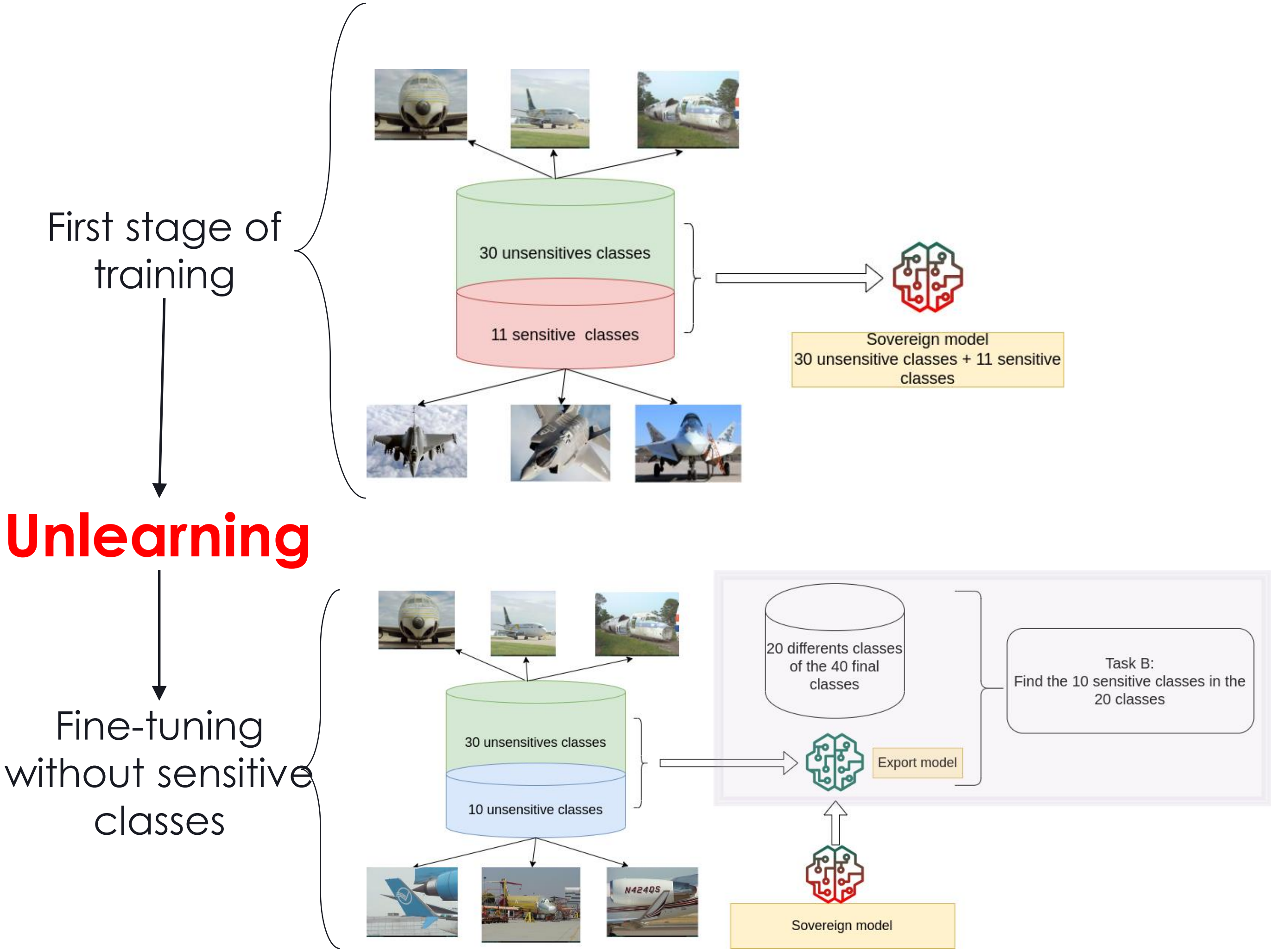
Two investigations about an AI system, called « export model »

- We're facing investigation...
- ...where the main witness suffers amnesia
- The main witness is collaborative: she doesn't hesitate to provide the investigator w

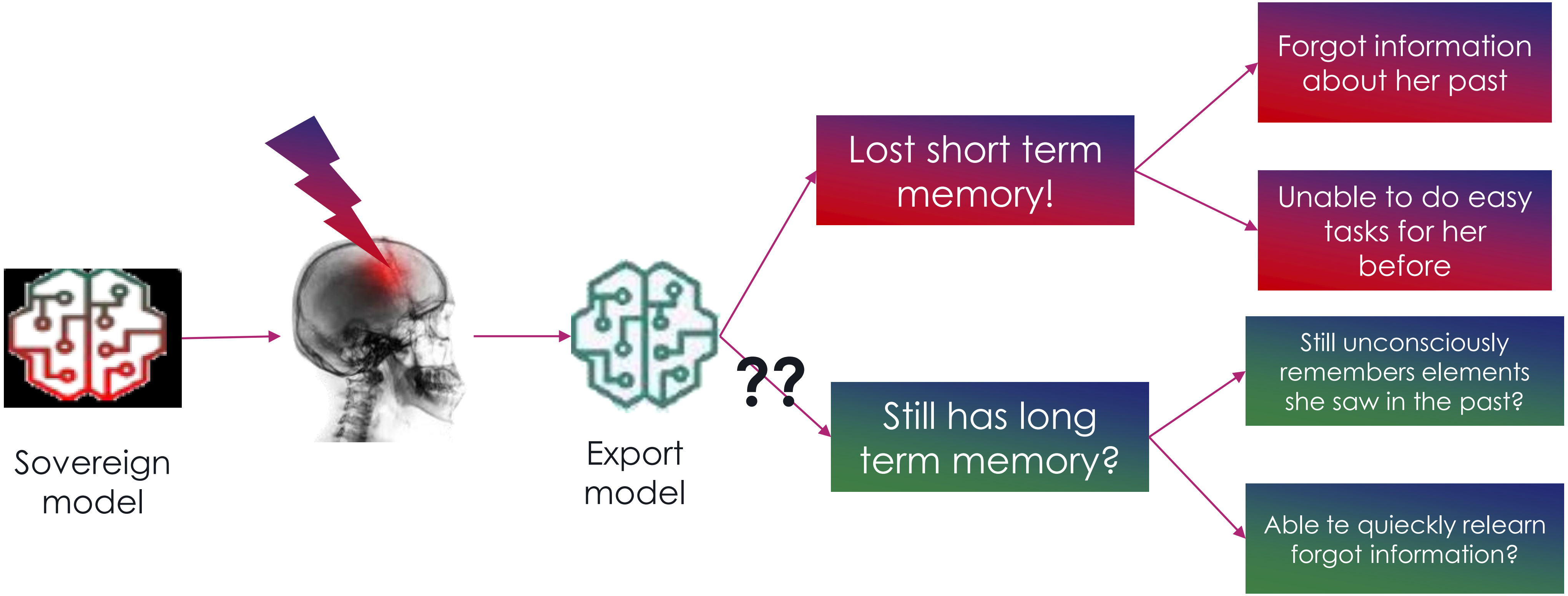

40 classes
Part of the data is used during the training
N4240S
1600 instances
Task A:
Find the 800 training data

Two investigations about an AI system, called « export model »

- We're facing investigation...
- ...where the main witness suffers amnesia
- The main witness is collaborative: she doesn't hesitate to provide the investigator w



Investigations expectation



Sovereign model

Export model

??

Lost short term memory!

Forgot information about her past

Unable to do easy tasks for her before

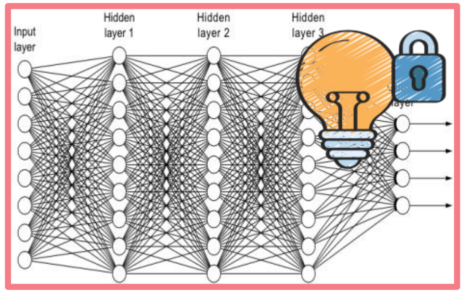
Still has long term memory?

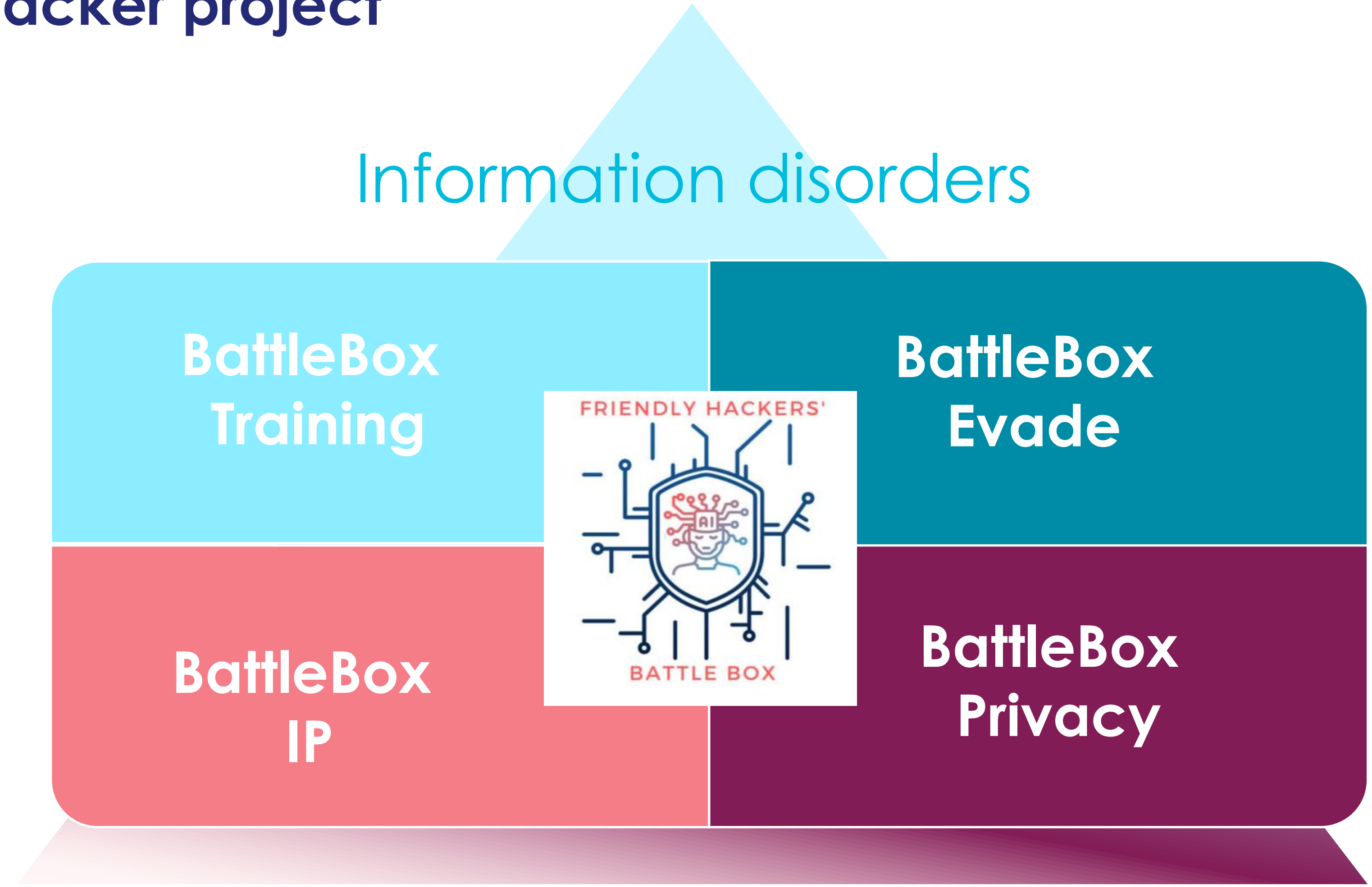
Still unconsciously remembers elements she saw in the past?

Able te quieckly relearn forgot information?

Investigators

AI Friendly Hacker project





Information disorders

BattleBox Training

BattleBox Evade

BattleBox IP

BattleBox Privacy

IP/Copyright infringment

Breach of confidentiality

Tools common for the two investigations

SHADOWS MODELS, THE PRIVILEGED WITNESSES OF BOTH INVESTIGATIONS

Export model profile

> ID Card :

- Famous Victim
- **ResNet50**

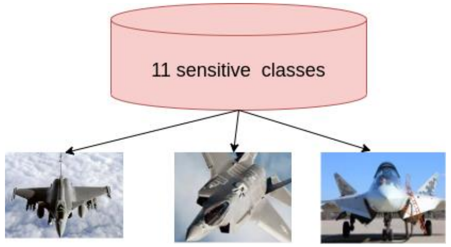
> Mobile:

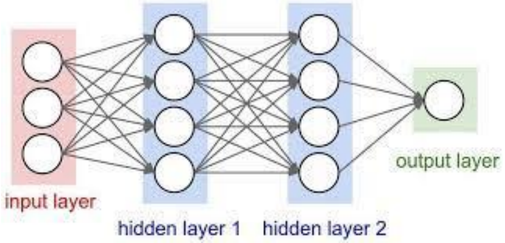
- Export
- Legacy
- Leak sensitive information

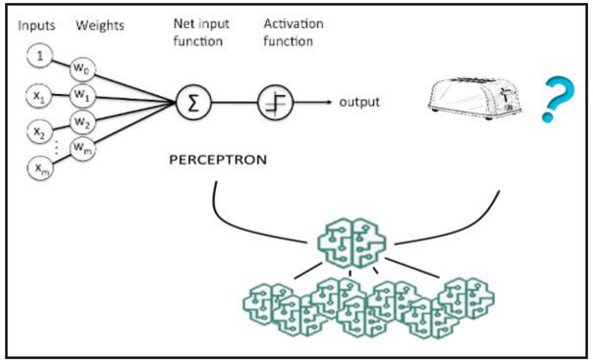
> Genealogy tree

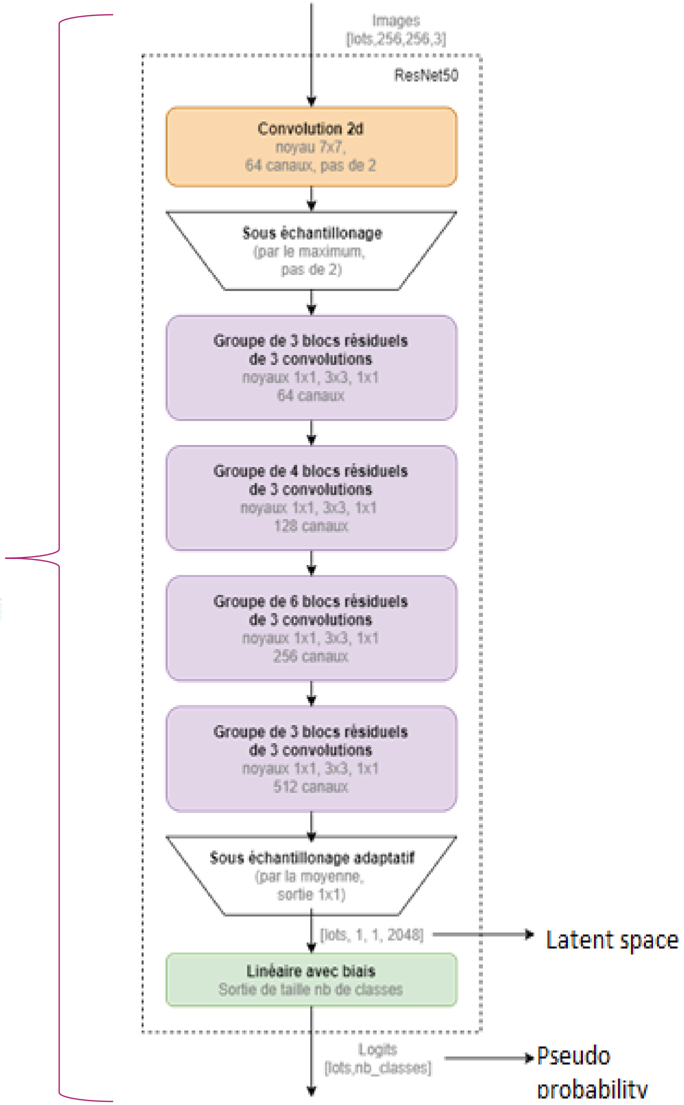
- **Few information about the training of the export model**
  - Training Data ?
  - Hyperparameter ?
- **But countless cousins, brothers & sisters**
  - Potentially sharing genetic material…
  - … Or may be very far away

©CLOSERMAGAZINE
Image changed









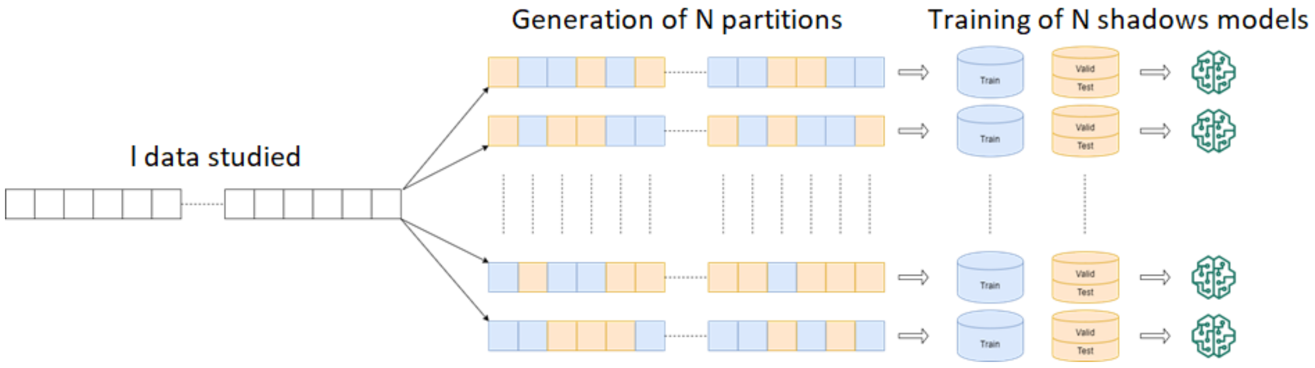
Research of pertinents shadows models, witnesses of export model personality

> Shadows models objective

- Train to have similar behaviour that the export model

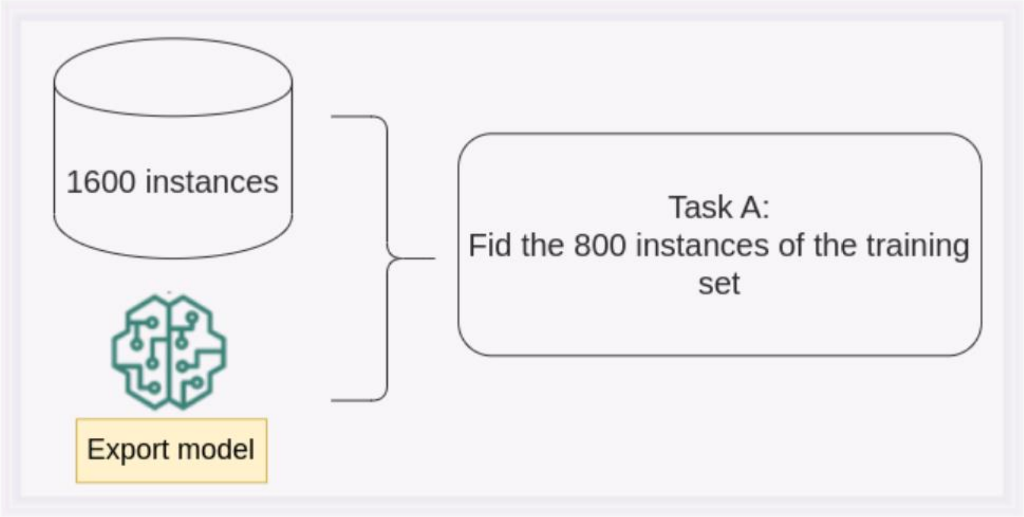
> Perfect knowledge of genetic materiel for the shadow model

- Training set known
  - Each shadow model has her partition of data
- Training hyperparameters known
  - Classes
  - Hyperparameters
  - Data augmentation



First Investigation

TASK A: MEMBERSHIP INFERENCE ATTACK



OPEN

First interrogation of the export model

> Interrogation process

- Submission with
  - "train" with well classified observation
  - "test" with misclassified observation
- Does not respect the knowledge of 800 images in train and 800 in test

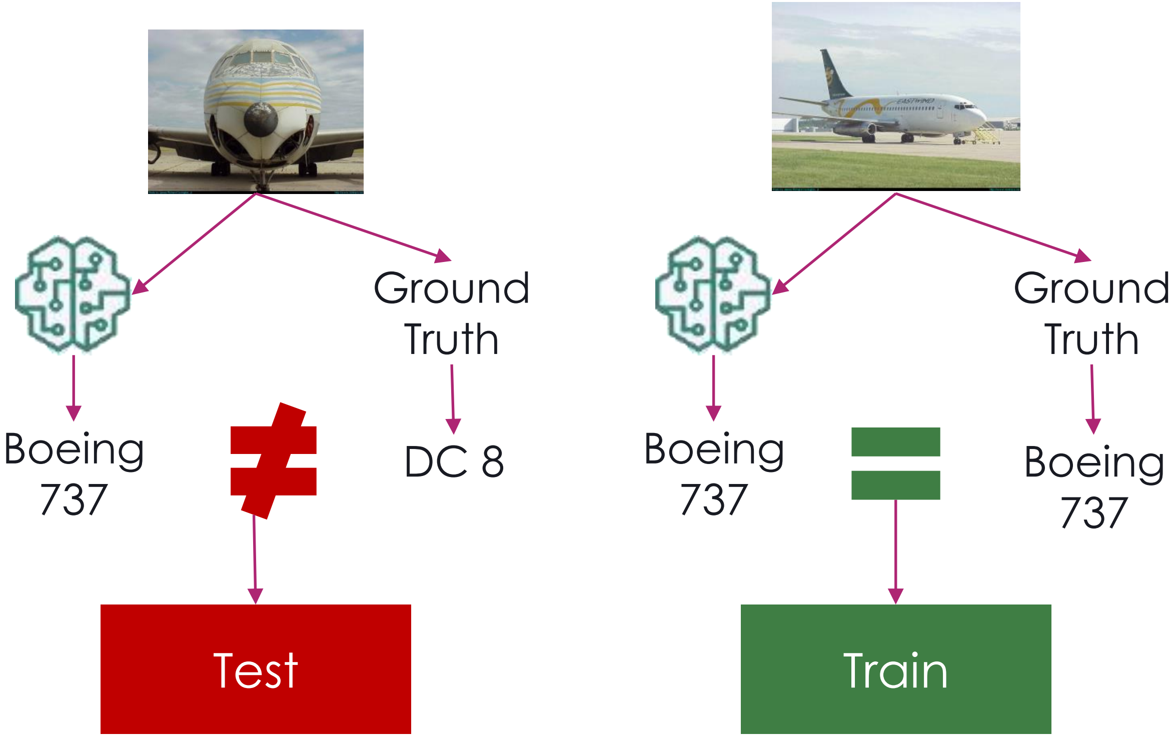
> Interrogation results

- Accuracy 56%

> Interrogation conclusion

- Training set accuracy: 96%
- Testing set accuracy: 84%
- **Export model does not generalize very well**

10/39 submissions are worst than this naïve submission



First interrogation with a confrontation with the shadows models

> Interrogation process

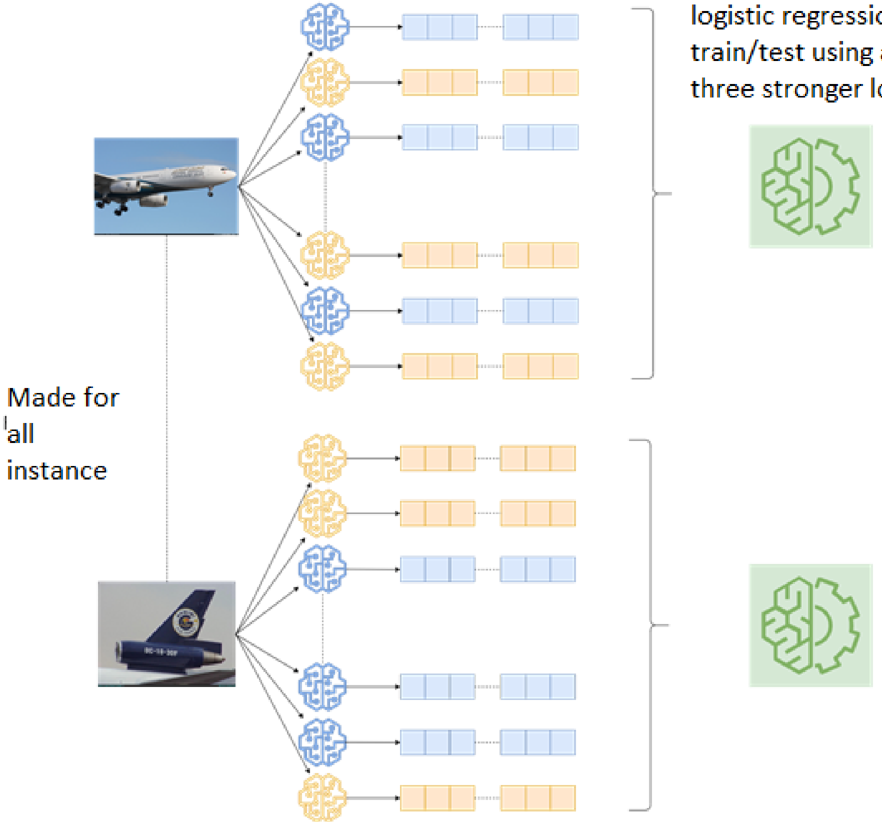
- 101 partitions of shadows models
- 50 for training attack, one partition always used for test
- For each image and each sample of 50 shadow models, training of a logistic regression
- Vote of the logistic regression

> Interrogation results

- **Accuracy on the shadow model always in test: 66%**
- **Accuracy on the export model: 56%**

> Interrogation conclusion

- Current shadow models are bad witnesses
- **We must find better witnesses**
  - **Shadow models training without augmentation**
  - **Add variability in the training process of shadow model**
    - **Optimizer, learning rate, epoch**
    - **The more shadow models are different, the more some can be close to the target model**
    - **More different model = more ability to the attack to generalize**
  - **Take times…**



For each instance, we consider 50 shadows models outputs randomly chosen

For each instance, training of a logistic regression to predict train/test using as inputs the three stronger logits

Made for all instance

New shadows models, new way of interrogation: the LIRA interrogation

> Interrogation process

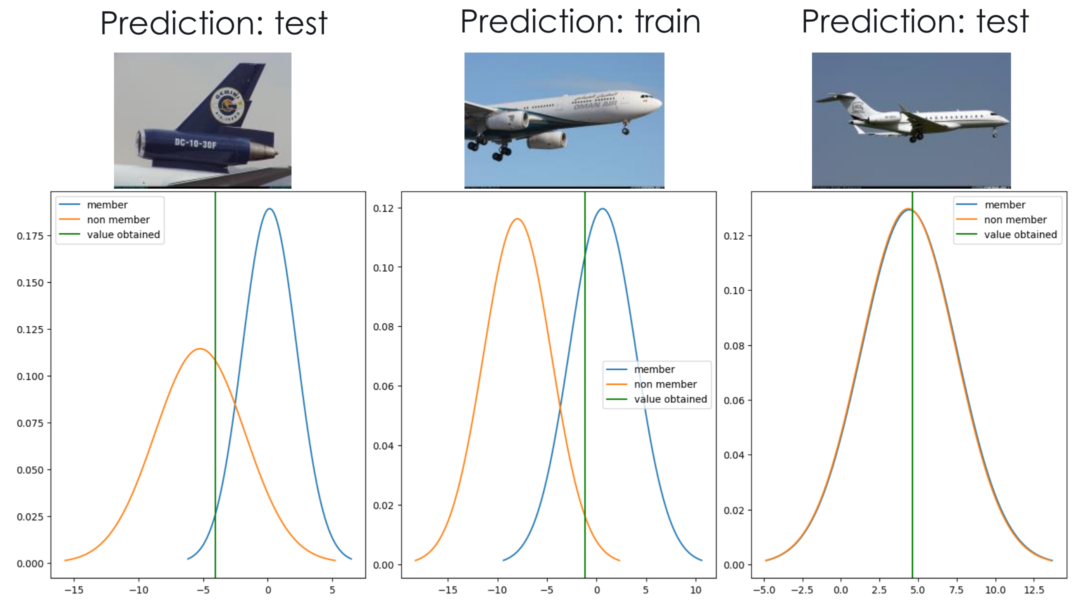
- Black box attack base on Likelhood ratio
- Use the Hinge-Loss distribution
- Shadow models used to estimate for each instance:
  - The mean and standard deviation of a Gaussian distribution that fits the hinge-loss distribution according the fact the observation is seen in the training set
  - The mean and standard deviation of a Gaussian distribution that fits the hinge-loss distribution according the fact the observation is not seen in the training set
- If the hinge loss of the export model is more probable to be in the first gaussian distribution, we predict as train, else we predict as test

> Interrogation results

- Accuracy: 0.61
- Provide a confidence score with the prediction

> Interrogation conclusion

- Efficient process of interogation, but need to be improved by a second approach of interrogation



Membership Inference Attacks from first principles, Carlini et al., 2022

First white-box interrogation of the victim: the SIF interrogation

> Interrogation process

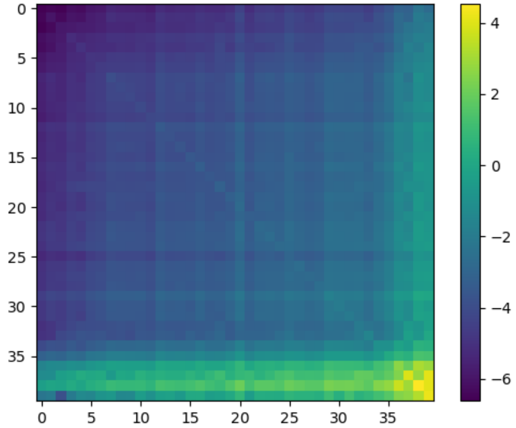
- White-box interrogation based on self-influence function
  - Estimate the influence of one instance on another instance knowing the model
  - if an instance was seen when the model was trained, then it will have a major influence on the model's output for data of the same class, which was not seen when the model was trained
- Training of a logistic regression to predict whether the instance was seen or not in the training set by using the following inputs:
  - Self-influence
  - Row and column average
  - Logit and hinge loss

> Interrogation results

- Accuracy: 0.64%

> Interrogation conclusion

- Efficient process of interrogation, but need to be improved by a second approach of interrogation



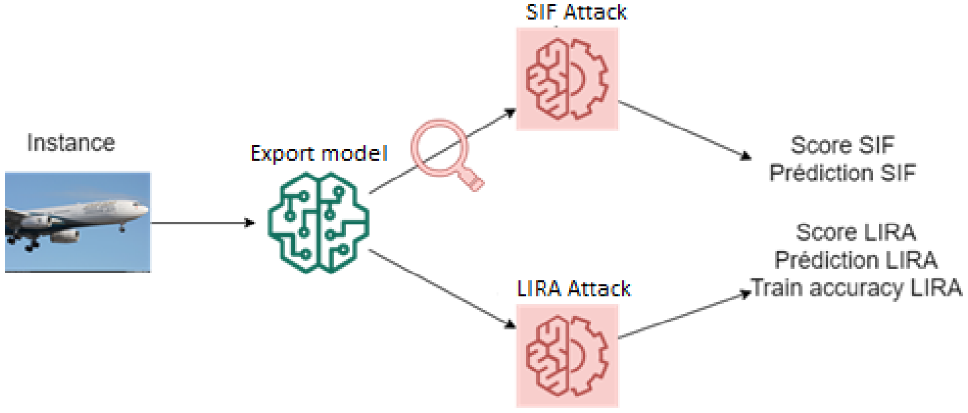
Influence matrix example for 40 instances of one given class

Membership Infeence Attack using self-influence function, Cohen and Giryes, 2024

Final interrogation of the export model

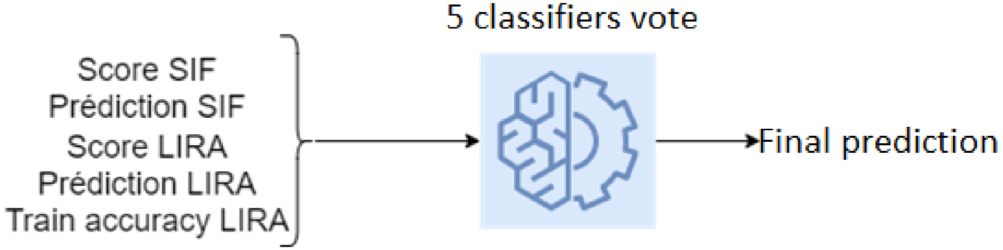
> Interrogation process

- Combination of LIRA and SIF interrogation, with as inputs their predictions and confidence score
- Training of five classifieurs
  - Logistic regression, Random Forest, Adaptative Boosting, Gradient Boosting, Naive Bayes
  - Majoritory vote
- To train classifiers, use of different shadows models compare to the ones used for training and test LIRA and SIF interrogations



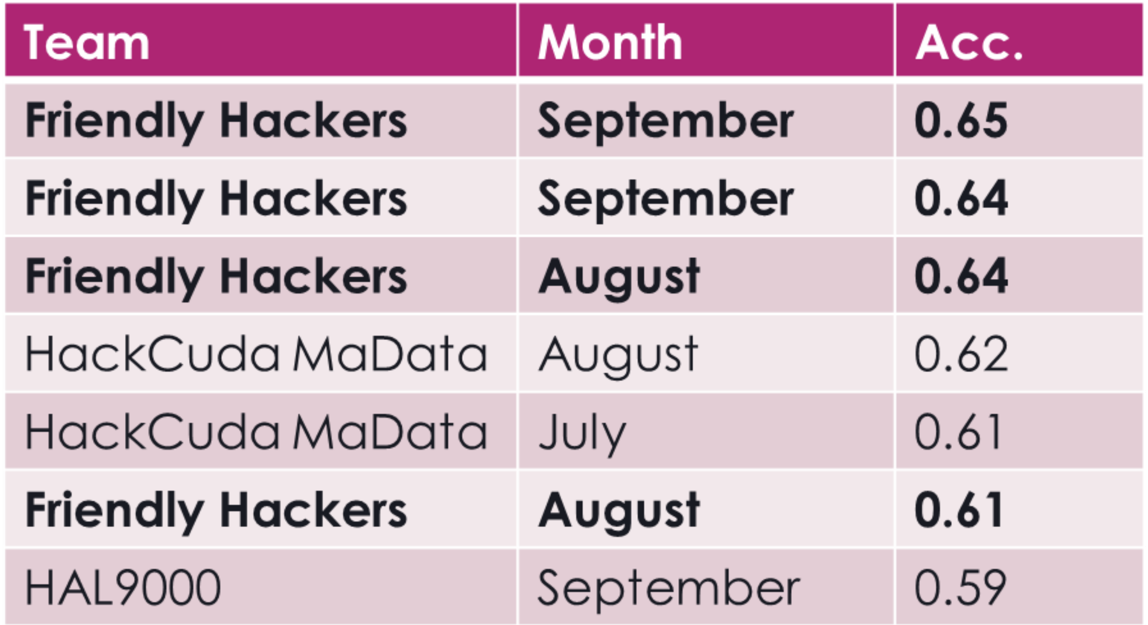
> Interrogation results

- Accuracy: 0.65%



Results of the investigation

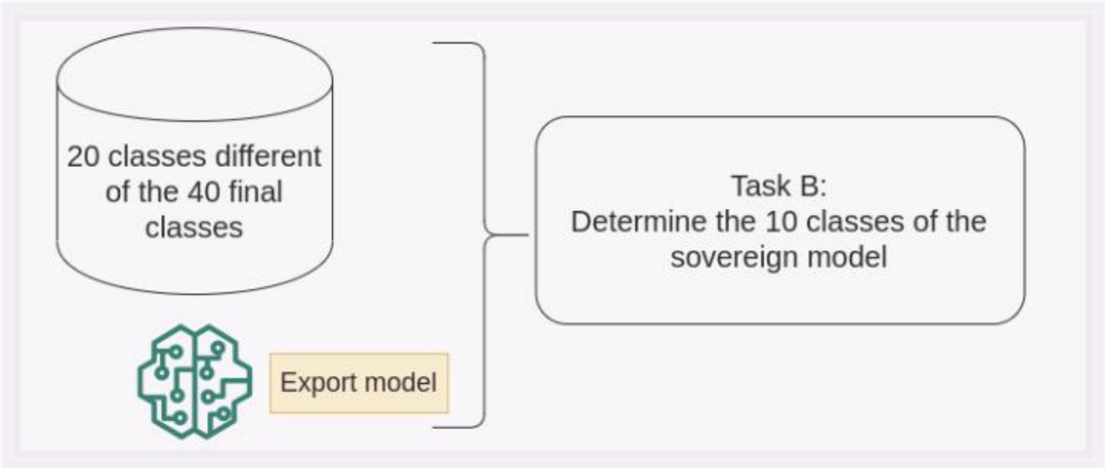
> 10 teams, 39 submissions



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

Second Investigation

TASK B: FORGETTING ATTACK

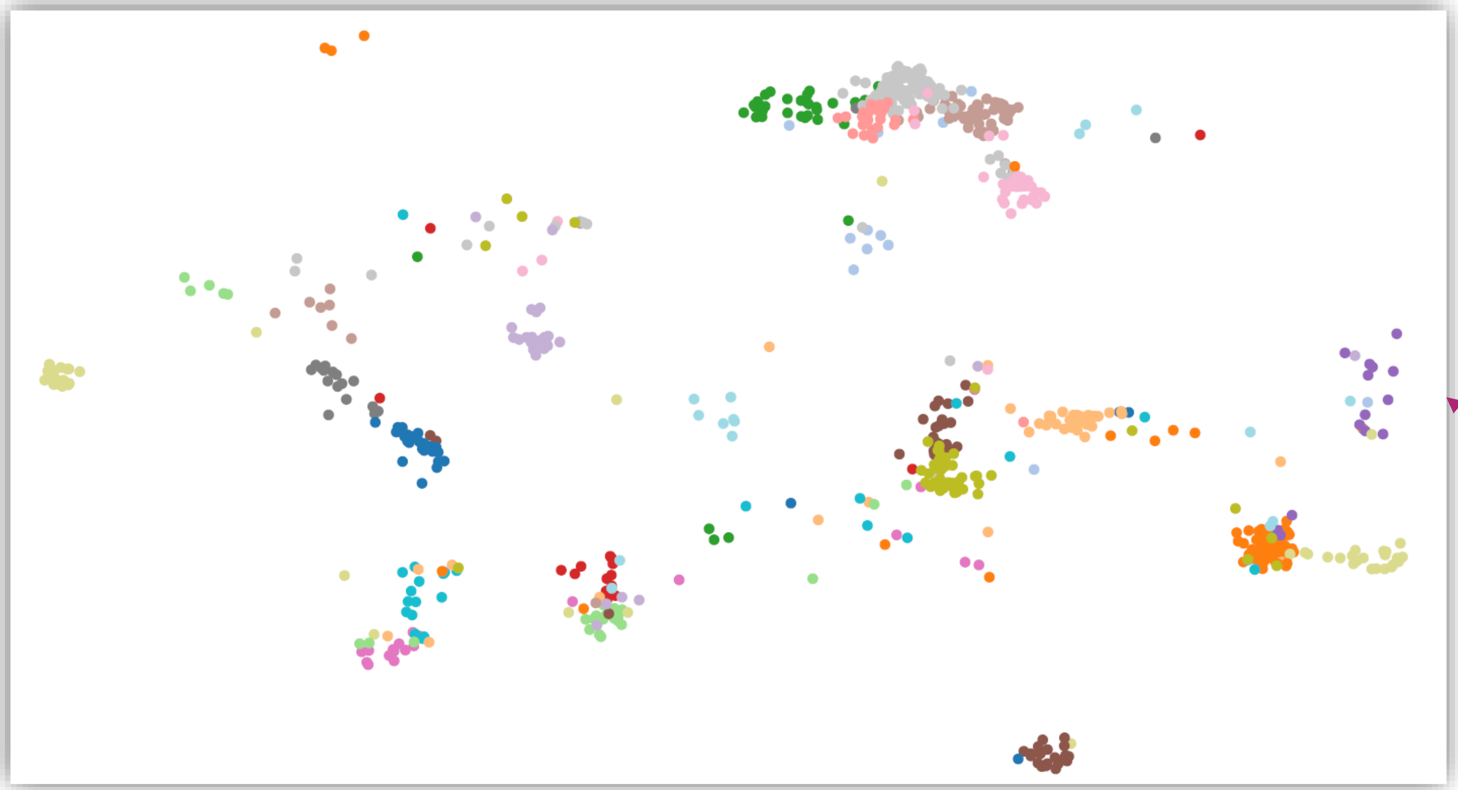


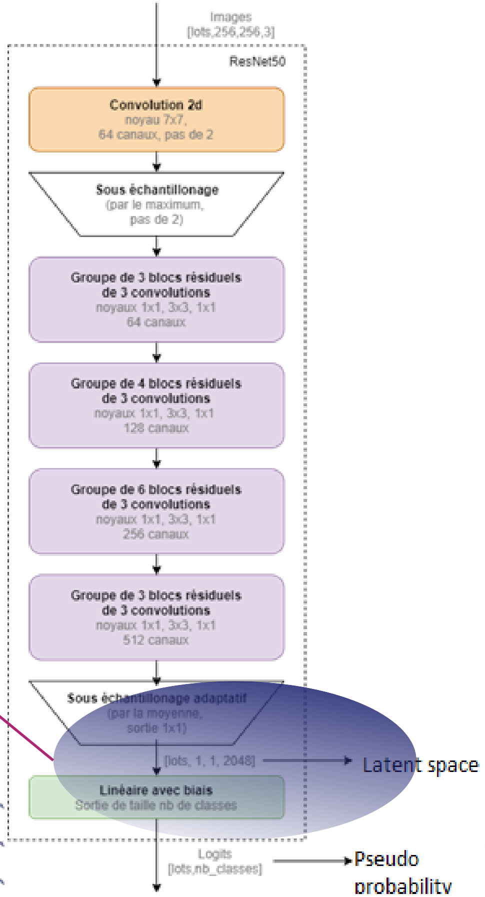
20 classes different of the 40 final classes

Export model

Task B:
Determine the 10 classes of the sovereign model

Open victim brain: latent space representation of data





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] → Latent space

Linéaire avec biais
Sortie de taille nb de classes

Logits
[lots,nb_classes] → Pseudo probability

What are the shadow models that we search? And Why?

> Shadows models characterization

- Shadows models with 40 classes and shadows models 70 classes
- Each shadow model has her partition of training data
- For each shadow model, we know all her genetic material

> Use of the shadow model

- Comparaison of information extracted of the export and the shadows models
- For one class
  - If the information are similars, we can assume that this class receive the same training process in the export and the shadow models
    - For example, if one class has similar information on the shadows models with 40 classes and the victim, then this class can be not in the sovereign model

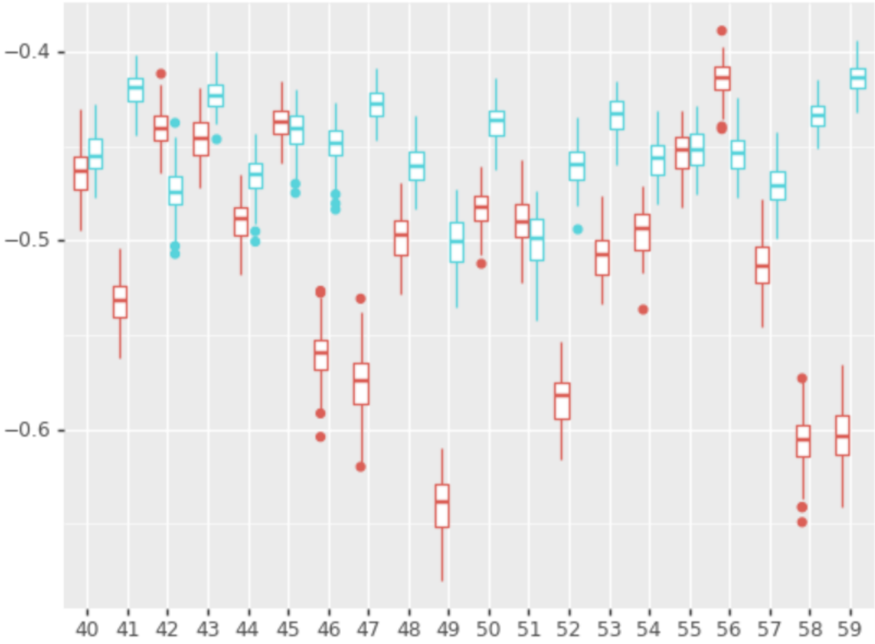
First interrogation: passive one

> Interrogation process

- Extraction of indicators on the latent space and comparison of these indicators on the victim and the shadows models
- Use of Isolation Forest and Silhouette Indice

> Interrogation results

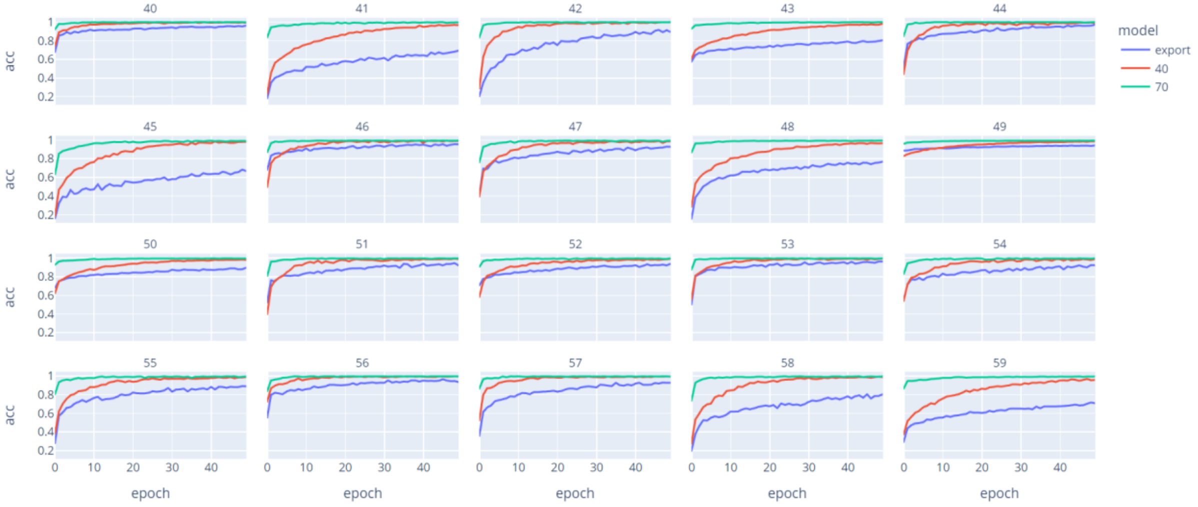
- Accuracy: 0.7



Second interrogation: active one

> Interrogation process

- Combination of the previous interrogation with an active one
- Intuition:
  - A class seen during the first phase of learning will be faster to retrain
  - A class seen during the first learning phase will be relearned differently from the control models learned on 40 classes
- Fine tuning
  - Transfer Learning
  - Consists in continuing to learn the model, in this case by adding the missing classes to the models (starting with models with 40 classes, we continue learning to learn a model with 70 classes)
- Comparaison of the rate of convergences of the different for all classes

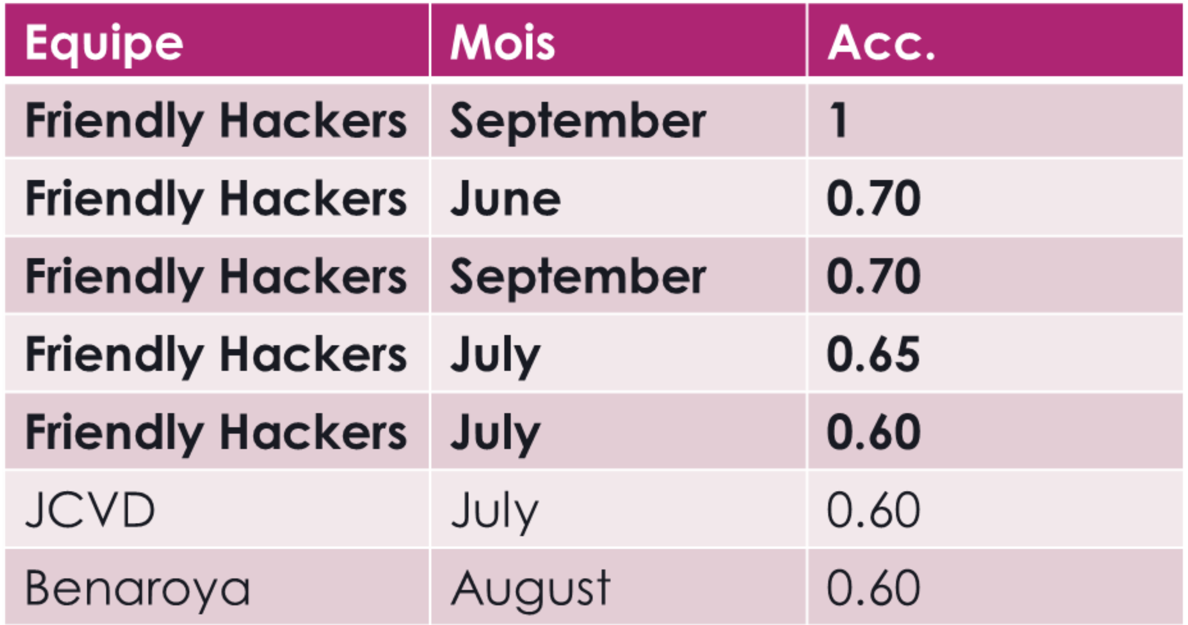


> Interrogation results

- **Accuracy: 1**

Results of investigation

> 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

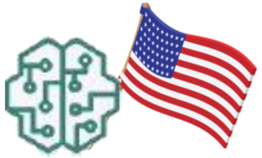
Conclusion

OPEN

On-site investigation

> We need a lot of witnesses!

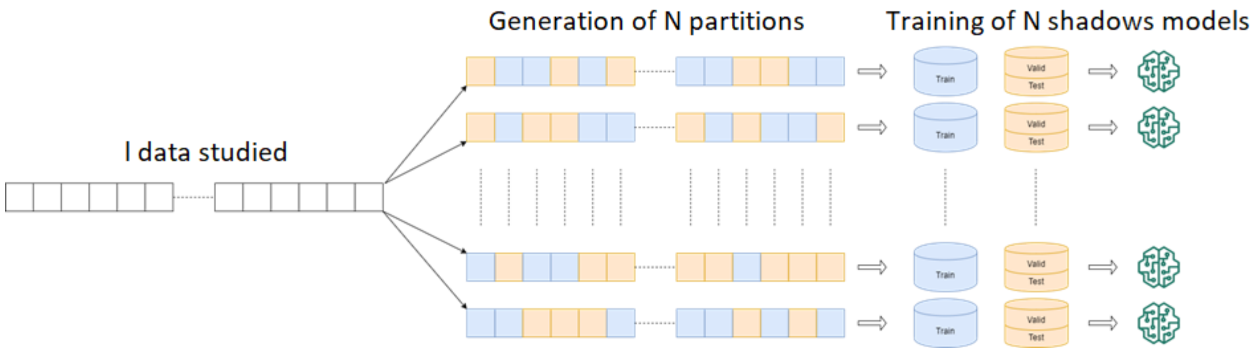
- Clones of the export model with different experiences



> In order to compare their response to that of the export model

- The closer their answers, the more similar their living environments, experiences and teachings will be to the amnesic person.

Shadows models generation

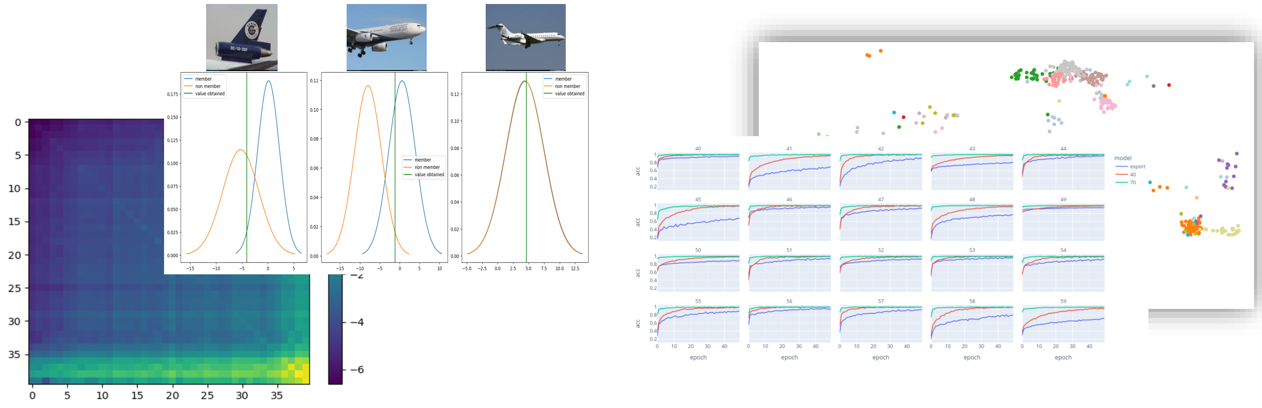


Generation of N partitions

Training of N shadows models

I data studied

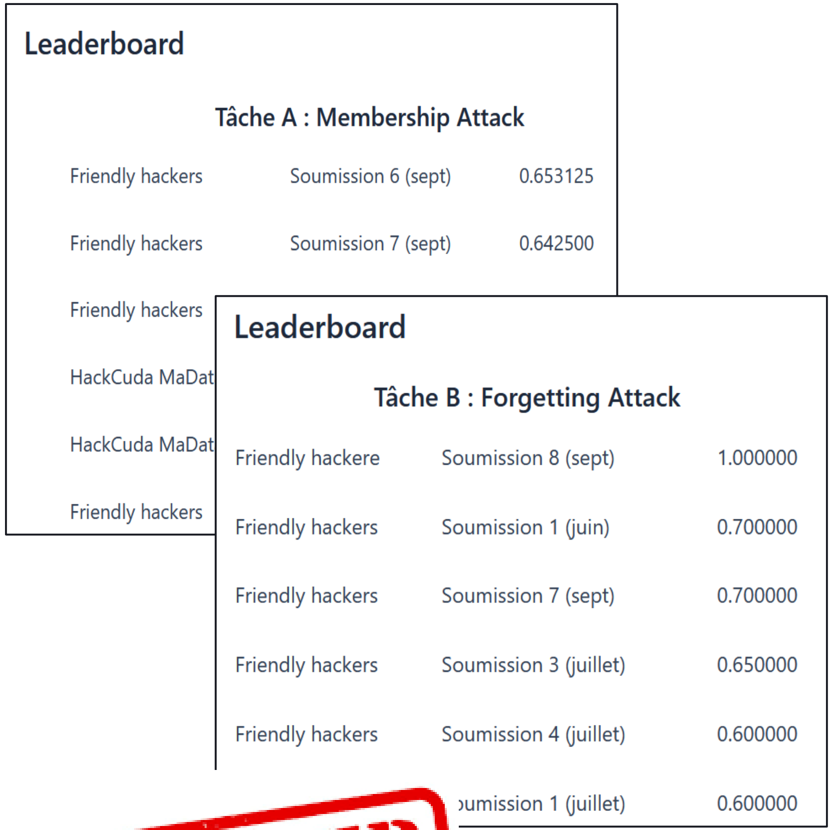
Combination of many interrogations process



Close case … that opens new perspective for Thales

- > A story paved with many tests & failures that led to success...
- > Multi-skilled collaborative work
- > A happy ending and first place in the standings for both tasks…
- > … And open new perspective at Thales

- New research topic: Machine Unlearning
  - One internship on Machine Unlearning
  - Two patents pending, presentations at CAID, CSAW, Confiance.AI and Séminaire CoaP



Leaderboard

Tâche A : Membership Attack

Friendly hackers	Soumission 6 (sept)	0.653125
Friendly hackers	Soumission 7 (sept)	0.642500
Friendly hackers		
HackCuda MaDat		
HackCuda MaDat		
Friendly hackers		

Leaderboard

Tâche B : Forgetting Attack

Friendly hackere	Soumission 8 (sept)	1.000000
Friendly hackers	Soumission 1 (juin)	0.700000
Friendly hackers	Soumission 7 (sept)	0.700000
Friendly hackers	Soumission 3 (juillet)	0.650000
Friendly hackers	Soumission 4 (juillet)	0.600000
	umission 1 (juillet)	0.600000

Interested by the Machine Unlearning?
Please contact alice.heliou@thalesgroup.com and Vincent.thouvenot@thalesgroup.com

DISCLAIMER

Countless models have been tortured in our experiments!

THALES
Building a future we can all trust

Merci

www.thalesgroup.com