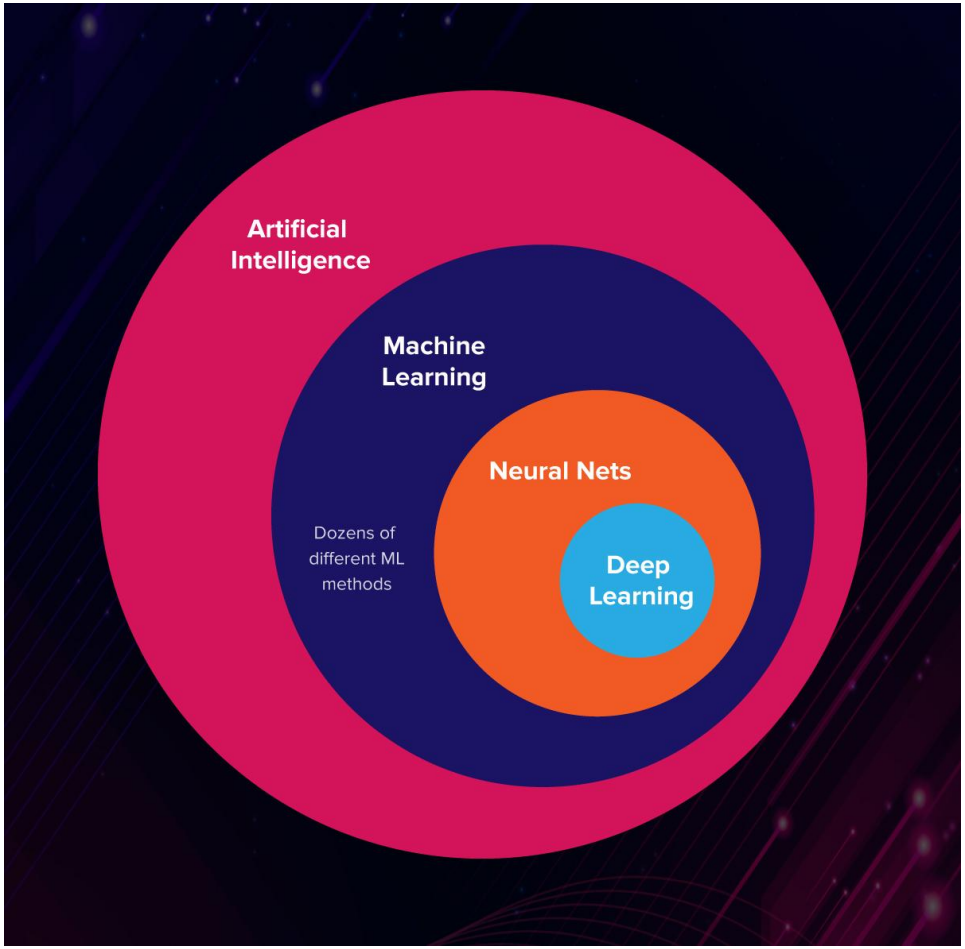
Randomness in ML and how can we share our models...?

Machine Learning and Reproducibility PoV

Paul Temple

RIPOST 2025

AI vs ML vs DL



Source: `https://serokell.io/blog/ai-ml-dl-difference`

A simple exercise

4. Experimental Approach and Results

The primary objective of this study was to investigate the use of Support Vector Machine (SVM) and decision tree classifiers for the detection of cardiac issues. The focus of our approach was to establish a robust model capable of predicting cardiac conditions with a target accuracy that surpasses the current state of the art of 62 %.

4.1 Data Acquisition and Pre-processing

The dataset we used is the well known *heart disease* dataset which describes a number of 300 patients with 14 indicators including age and sex but also cholesterol level, blood pressure among otheres. Following the acquisition, we conducted rigorous data cleaning, addressed missing values, and normalized the data to ensure compatibility with the SVM and decision tree classifiers.

4.2 Model Implementation and Parameter Tuning

The SVM and decision tree classifiers were built and meticulously tuned to optimize their performance. This included searching for optimal parameters values such as the cost and gamma for the SVM, and maximum depth and minimum samples split for the decision tree. Cross-validation techniques were used throughout this process to mitigate overfitting and validate the performance of our models.

4.3 Results

After an extensive period of training and evaluation, the SVM classifier, with optimized parameter yielded the best results. Although the decision tree classifier demonstrated satisfactory performanc it was slightly outperformed by the SVM.

Remarkably, our SVM model achieved an approximate accuracy rate of 80%. This significant finding suggests that SVM, when correctly parameterized and cross-validated, can indeed serve as powerful tool for the prediction of cardiac conditions.

While the robustness of our model on this dataset is encouraging, we acknowledge that the results' generalizability needs to be further evaluated on diverse datasets and in different medical scenario

In conclusion, this pioneering study provides a foundation for future research in leveraging machi learning techniques such as SVM and decision tree classifiers for healthcare applications, particularly in predicting cardiac conditions. Our research underscores the potential of these mode to revolutionize medical diagnosis, and lays groundwork for more comprehensive and reliable AI-driven solutions in healthcare.

1: Retrieve results

Try to reach 80% accuracy!

- Did you succeed?
- What did you miss?
- Did you try any strategy?
- Did you try different algorithms?

2: Optimize results

What's your maximum accuracy?

- What is the best result?
- Did you try any strategy?
- Do you think it can be reproduced?

Reproducing results...

Random everywhere

- Random separation between training and test sets
- Random in data presentation
- Random initialization of (hyper)parameters/weights
- ...

Consequences of random everywhere

Reporting results is hard

- Everything related to random $\rightarrow$ unstable

Consequences of random everywhere

Reporting results is hard

- Everything related to random $\rightarrow$ unstable
- Gain stability and confidence $\rightarrow$ repeat

Consequences of random everywhere

Reporting results is hard

- Everything related to random $\rightarrow$ unstable
- Gain stability and confidence $\rightarrow$ repeat
- Accounting for (un)stability $\rightarrow$ average is not sufficient

Consequences of random everywhere

- Reporting results is hard

The role of seeds

- Can be fixed
  - Favor reproducibility
  - Limit flakiness when testing

To seed or not to seed?

Consequences of random everywhere

- Reporting results is hard

The role of seeds

- Can be fixed
  - Favor reproducibility
  - Limit flakiness when testing
- Not a good practice for model deployment (replicability)
  - Lower confidence in generalization
  - Probably not optimal for new datasets

To seed or not to seed?

Consequences of random everywhere

- Reporting results is hard

The role of seeds

- Can be fixed
  - Favor reproducibility
  - Limit flakiness when testing
- Not a good practice for model deployment (replicability)
  - Lower confidence in generalization
  - Probably not optimal for new datasets
- Ok for model testing
  - Fix seeds for every run
  - Log seeds
  - Report seeds and results

To seed or not to seed?

Consequences of random everywhere

- Reporting results is hard
- Seeds can be fixed

Not always harmful

- Taxonomy of use of randoms

We need to talk about random seeds

Consequences of random everywhere

- Reporting results is hard
- Seeds can be fixed

Not always harmful

- Taxonomy of use of randoms
- Ok for $\rightarrow$ model selection, ensemble creation, and sensitivity analysis
- Avoid for $\rightarrow$ reproducibility (does not help with GPU under TensorFlow), performance comparison, optimizing performances

We need to talk about random seeds

ML processes are random by nature

- Different libraries exist

ML processes are random by nature

- Different libraries exist
- Implement different techniques
- Default values may not be the same

ML processes are random by nature

- Different libraries exist
- Implement different techniques
- Default values may not be the same

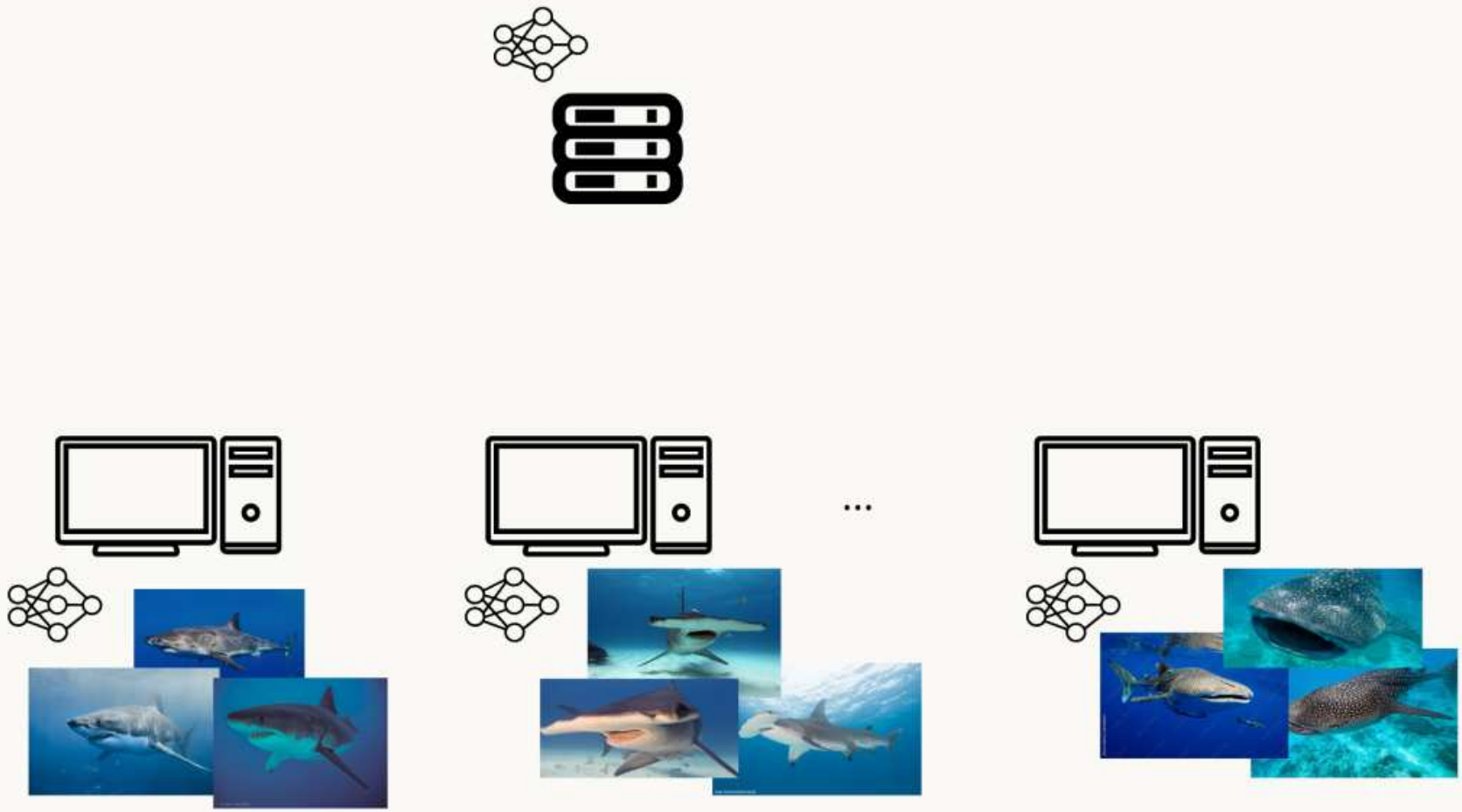
$\Rightarrow$ Harm reproducibility

$\rightarrow$ Need to improve communications of results

How to share our models?

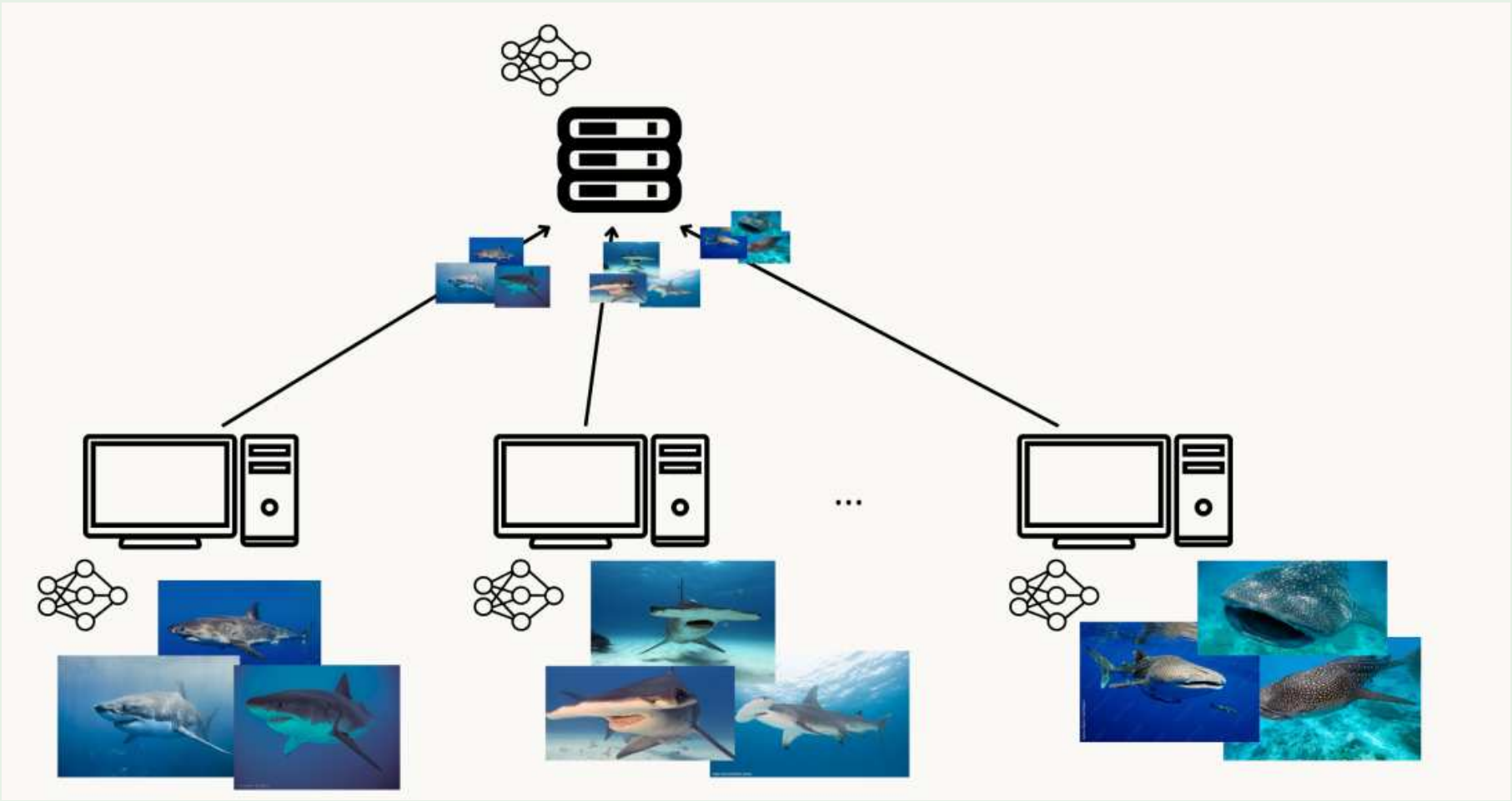
Federated Learning: Camille Molinier's PhD

Distributed ML & Federated learning



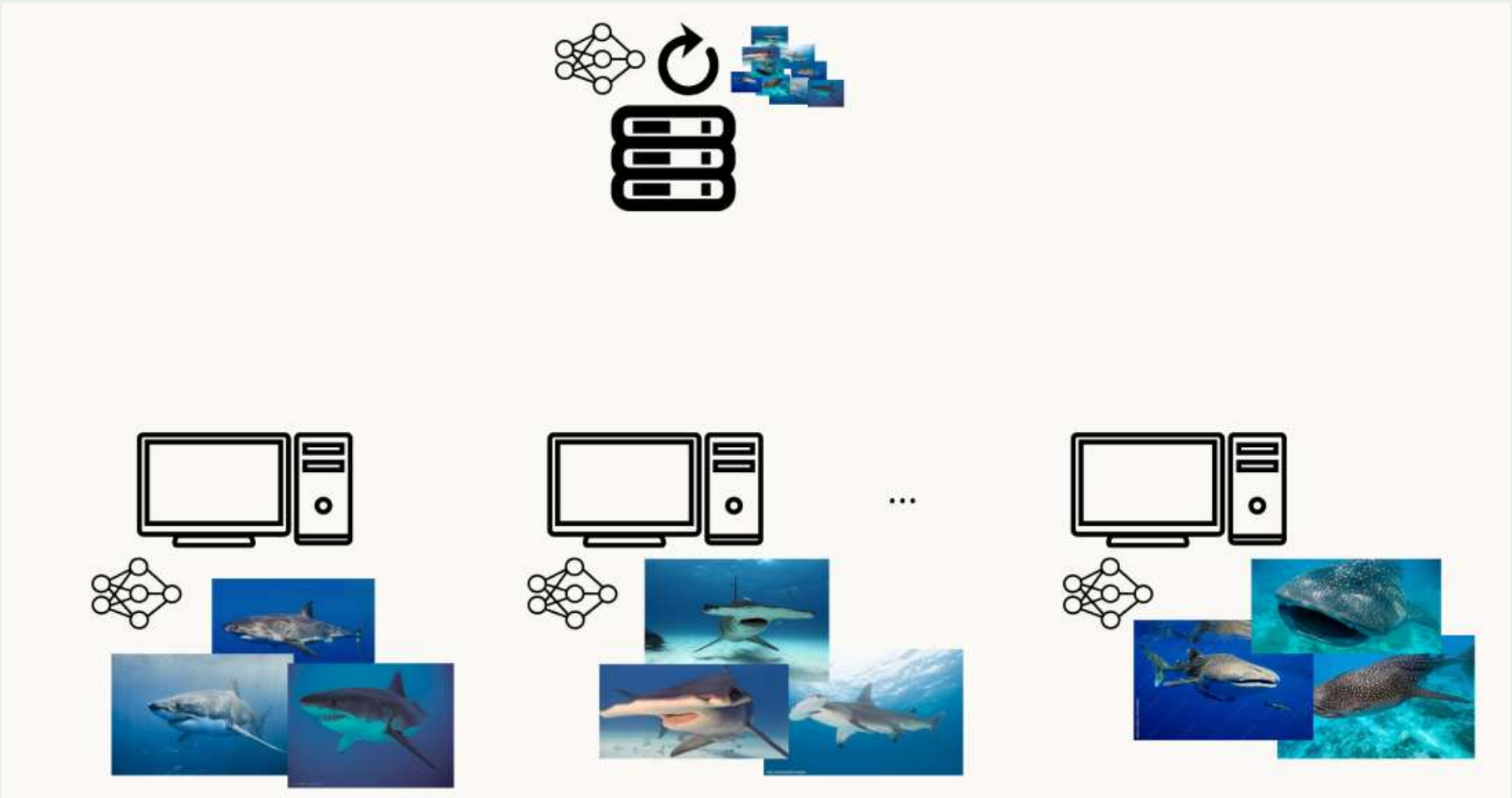
Federated Learning: Camille Molinier's PhD

Distributed ML & Federated learning



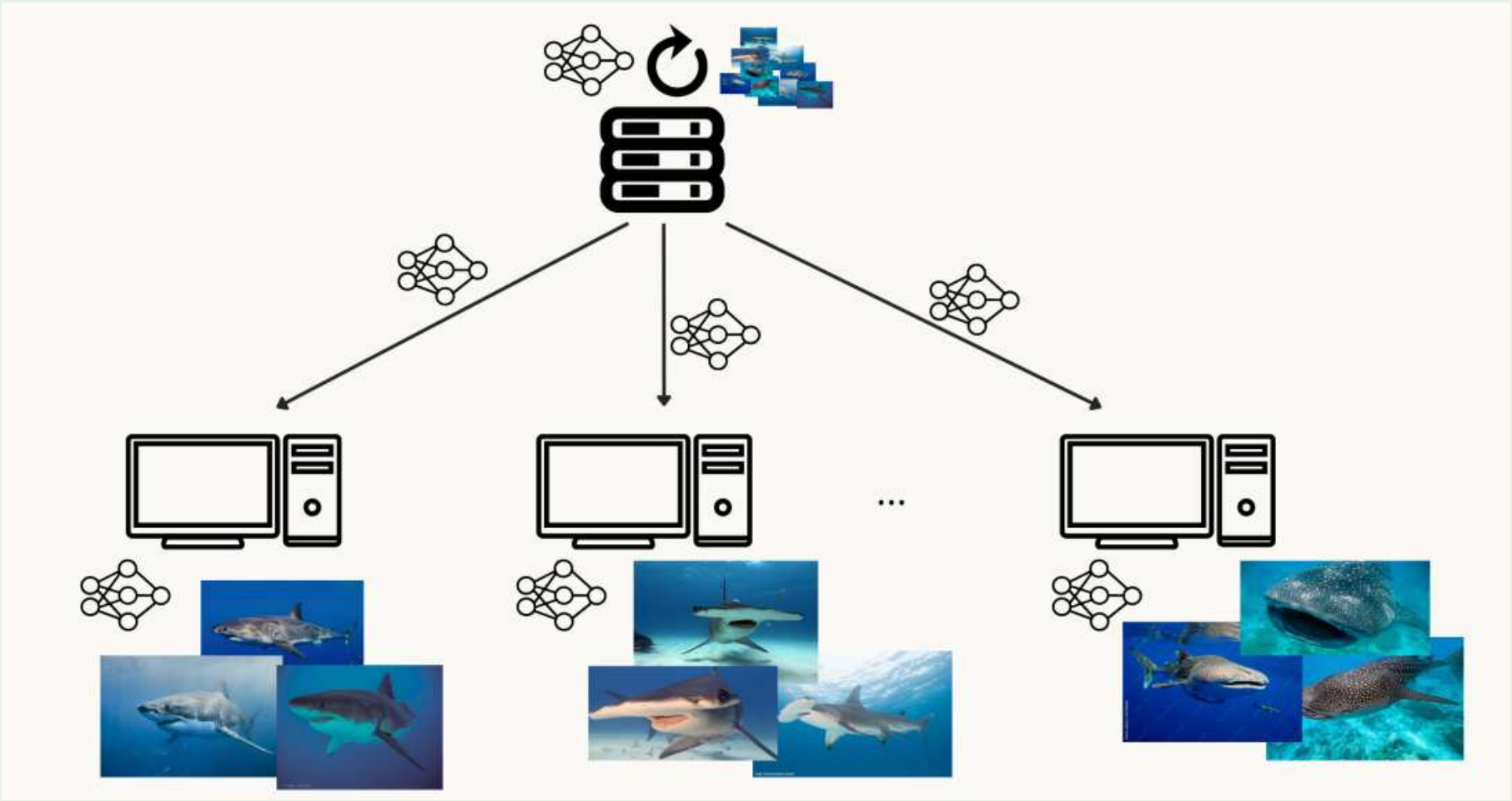
Federated Learning: Camille Molinier's PhD

Distributed ML & Federated learning



Federated Learning: Camille Molinier's PhD

Distributed ML & Federated learning



Federated Learning for real

- Started with Google in 2016
- Distributed learning...

Federated Learning for real

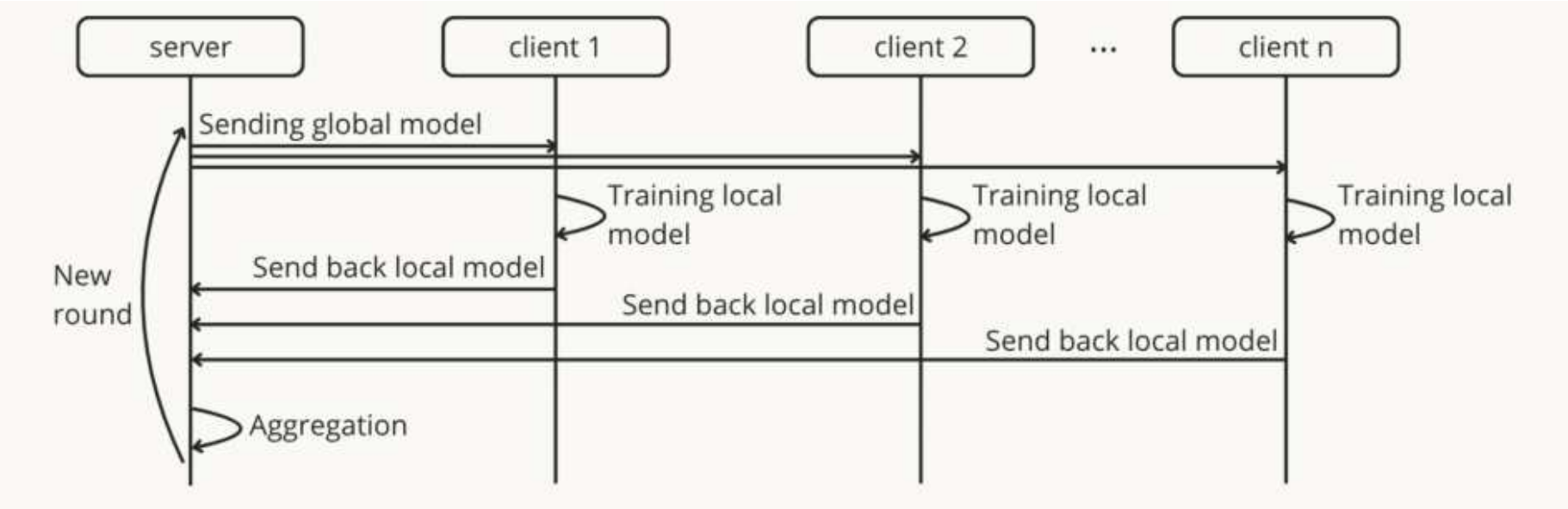
- Started with Google in 2016
- Distributed learning...
- with privacy concerns

Federated Learning for real

- Started with Google in 2016
- Distributed learning...
- with privacy concerns
- data never leave client models

Federated Learning for real

- Started with Google in 2016
- Distributed learning...
- with privacy concerns
- data never leave client models



Federated Learning for real

What about security?

- advML attacks still work...

Federated Learning for real

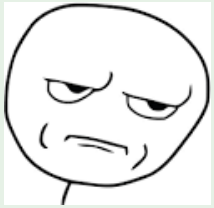
What about security?

- advML attacks still work...
- Privacy as a major concern (not security)

Federated Learning for real

What about security?

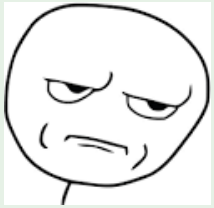
- advML attacks still work...
- Privacy as a major concern (not security)
- Sometimes protocols make things even worse (homomorphic)



Federated Learning for real

What about security?

- advML attacks still work...
- Privacy as a major concern (not security)
- Sometimes protocols make things even worse (homomorphic)



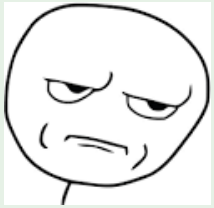
What about model updates?

- What happen to previous global models?

Federated Learning for real

What about security?

- advML attacks still work...
- Privacy as a major concern (not security)
- Sometimes protocols make things even worse (homomorphic)



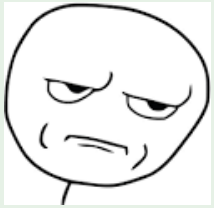
What about model updates?

- What happen to previous global models?
- Can we assess performance evolution with previous models?

Federated Learning for real

What about security?

- advML attacks still work...
- Privacy as a major concern (not security)
- Sometimes protocols make things even worse (homomorphic)



What about model updates?

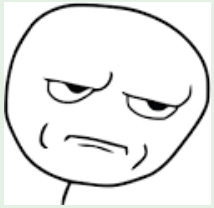
- What happen to previous global models?
- Can we assess performance evolution with previous models?

$\Rightarrow$ Use of models history? (to adapt testing techniques)

Federated Learning for real

What about security?

- advML attacks still work...
- Privacy as a major concern (not security)
- Sometimes protocols make things even worse (homomorphic)



What about model updates?

- What happen to previous global models?
- Can we assess performance evolution with previous models?

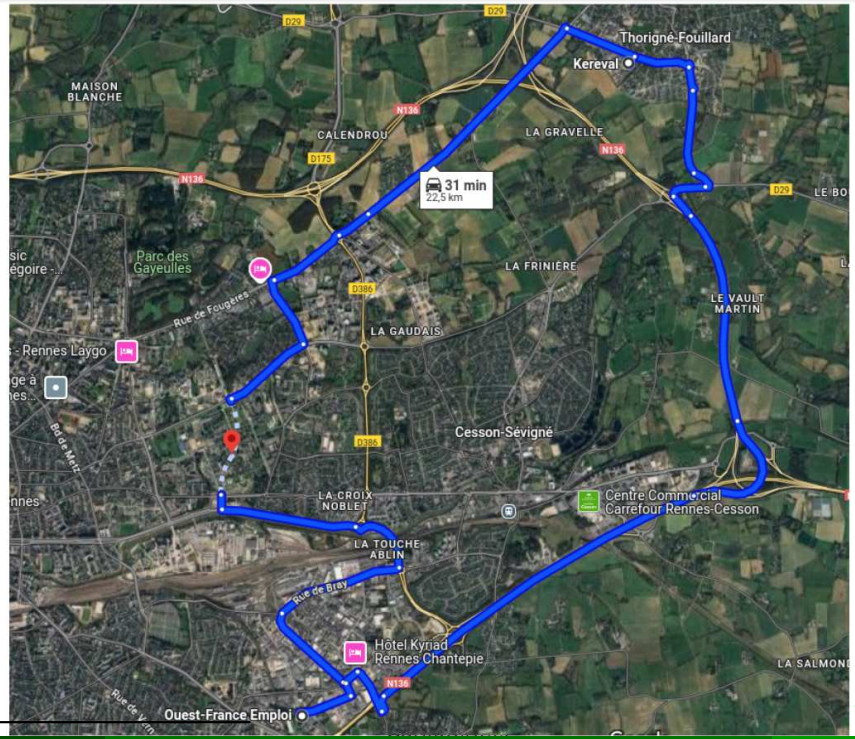
$\Rightarrow$ Use of models history? (to adapt testing techniques) Information needed to store the models?

Companies are also interested: WestOps

Industrial partners

Companies are also interested: WestOps

Industrial partners



Companies are also interested: WestOps

Problem

- Lots of illustration in the press (local events coverage)

Companies are also interested: WestOps

Problem

- Lots of illustration in the press (local events coverage)
- Is there a bias induced by the photos?

Companies are also interested: WestOps

Problem

- Lots of illustration in the press (local events coverage)
- Is there a bias induced by the photos?

Enters... ML

- Use of ML models to automatically annotate (and then count)

Companies are also interested: WestOps

Problem

- Lots of illustration in the press (local events coverage)
- Is there a bias induced by the photos?

Enters... ML

- Use of ML models to automatically annotate (and then count)
- First deployed model $\rightarrow$ 2019

Companies are also interested: WestOps

Problem

- Lots of illustration in the press (local events coverage)
- Is there a bias induced by the photos?

Enters... ML

- Use of ML models to automatically annotate (and then count)
- First deployed model $\rightarrow$ 2019
- During covid... $\rightarrow$ masks !!

Companies are also interested: WestOps

Problem

- Lots of illustration in the press (local events coverage)
- Is there a bias induced by the photos?

Enters... ML

- Use of ML models to automatically annotate (and then count)
- First deployed model $\rightarrow$ 2019
- During covid... $\rightarrow$ masks !!
- 2022 $\rightarrow$ people complain about annotations

$\Rightarrow$ Data distribution shift

Companies are also interested: WestOps

Goals of WestOps

- Automatically monitor ML model performances
- Trigger alarm if too many errors

Companies are also interested: WestOps

Goals of WestOps

- Automatically monitor ML model performances
- Trigger alarm if too many errors

$\Rightarrow$ Manual countermeasure? (inspection, retraining, ...)

Companies are also interested: WestOps

Goals of WestOps

- Automatically monitor ML model performances
- Trigger alarm if too many errors

$\Rightarrow$ Manual countermeasure? (inspection, retraining, ...)

- Ensure that updates are not making the ML model worse

Companies are also interested: WestOps

Goals of WestOps

- Automatically monitor ML model performances
- Trigger alarm if too many errors

$\Rightarrow$ Manual countermeasure? (inspection, retraining, ...)

- Ensure that updates are not making the ML model worse

Challenges

- History of models $\rightarrow$ Ouest-France thinks it's costly

Companies are also interested: WestOps

Goals of WestOps

- Automatically monitor ML model performances
- Trigger alarm if too many errors

$\Rightarrow$ Manual countermeasure? (inspection, retraining, ...)

- Ensure that updates are not making the ML model worse

Challenges

- History of models $\rightarrow$ Ouest-France thinks it's costly
- No groundtruth $\rightarrow$ Oracle problem
- No regression testing problem?

Companies are also interested: WestOps

Goals of WestOps

- Automatically monitor ML model performances
- Trigger alarm if too many errors

⇒ Manual countermeasure? (inspection, retraining, ...)

- Ensure that updates are not making the ML model worse

Challenges

- History of models → Ouest-France thinks it's costly
- No groundtruth → Oracle problem
- No regression testing problem?

⇒ Adapt software testing techniques to ML