Federated Learning and Network Security

Foundations, Potential, and Resilience

Yann BUSNEL
Institut Mines-Télécom, France

Léo LAVAUR
SnT, University of Luxembourg

July 20rd, 2025 — IEEE ICDCS — Glasgow, UK

LET’S TALK ABOUT FEDERATION

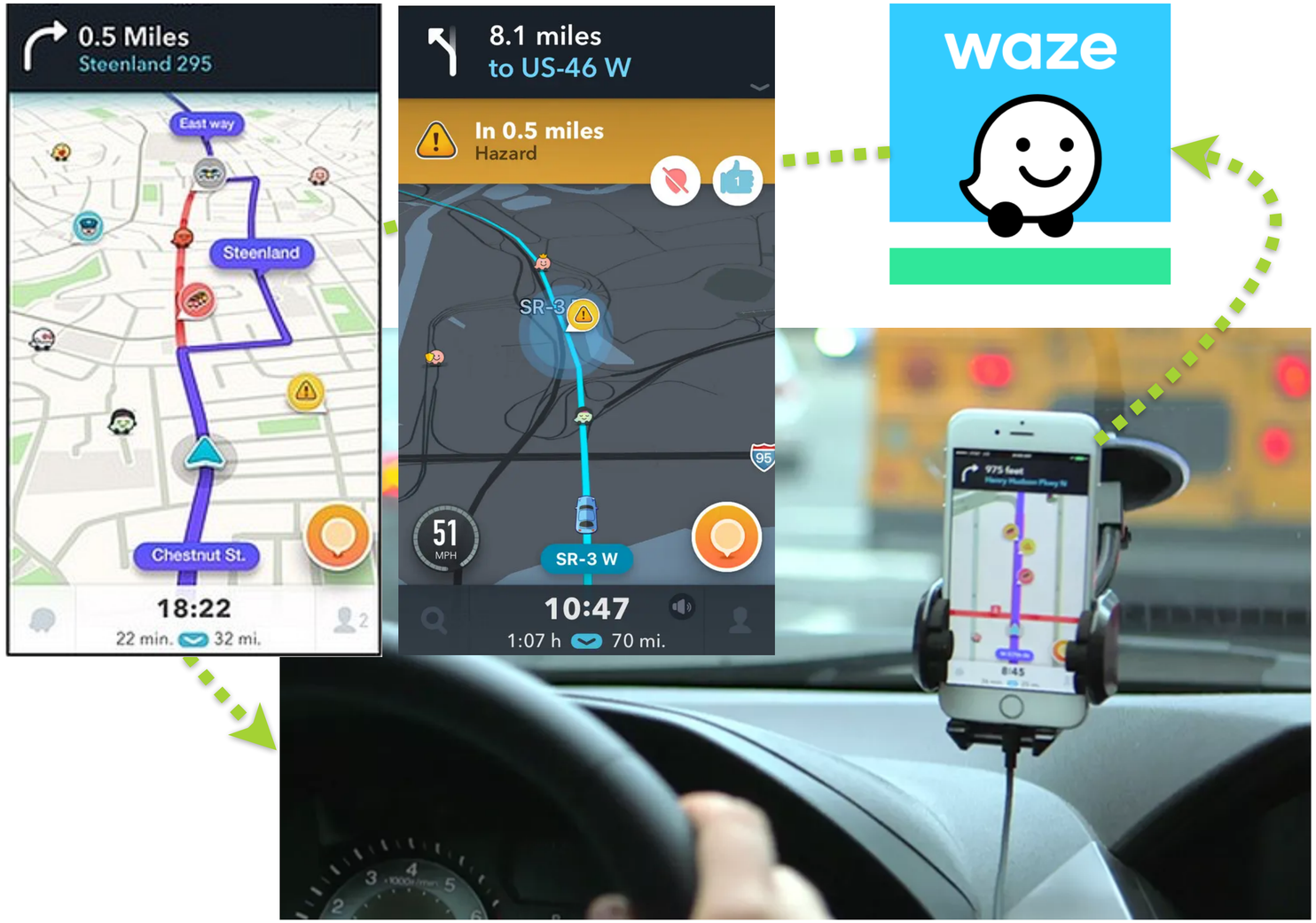
Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

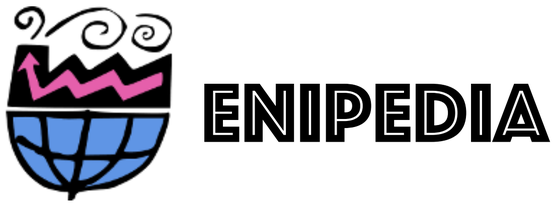
ALL START FROM CROWDSOURCING

- « Large group of dispersed participants contributing or producing goods or services [...] for payment or as volunteers »
  Wikipedia, 2023
- Waze Example



ALL START FROM CROWDSOURCING

- **Hotel or attraction reviews**
- **Collaborative journalism**
- **Unused room business**
- **Energy industry data**
- ***etc.***



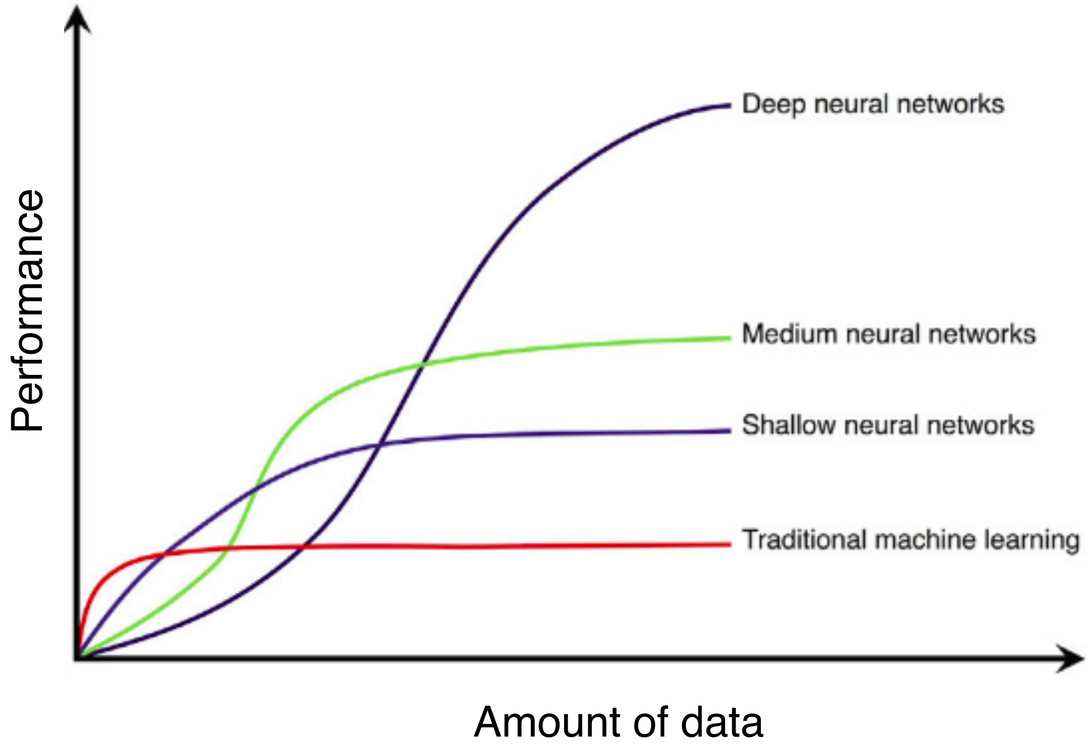
INDUSTRIALIZING CROWDSOURCING

amazon mechanical turk

CloudResearch®
connect™

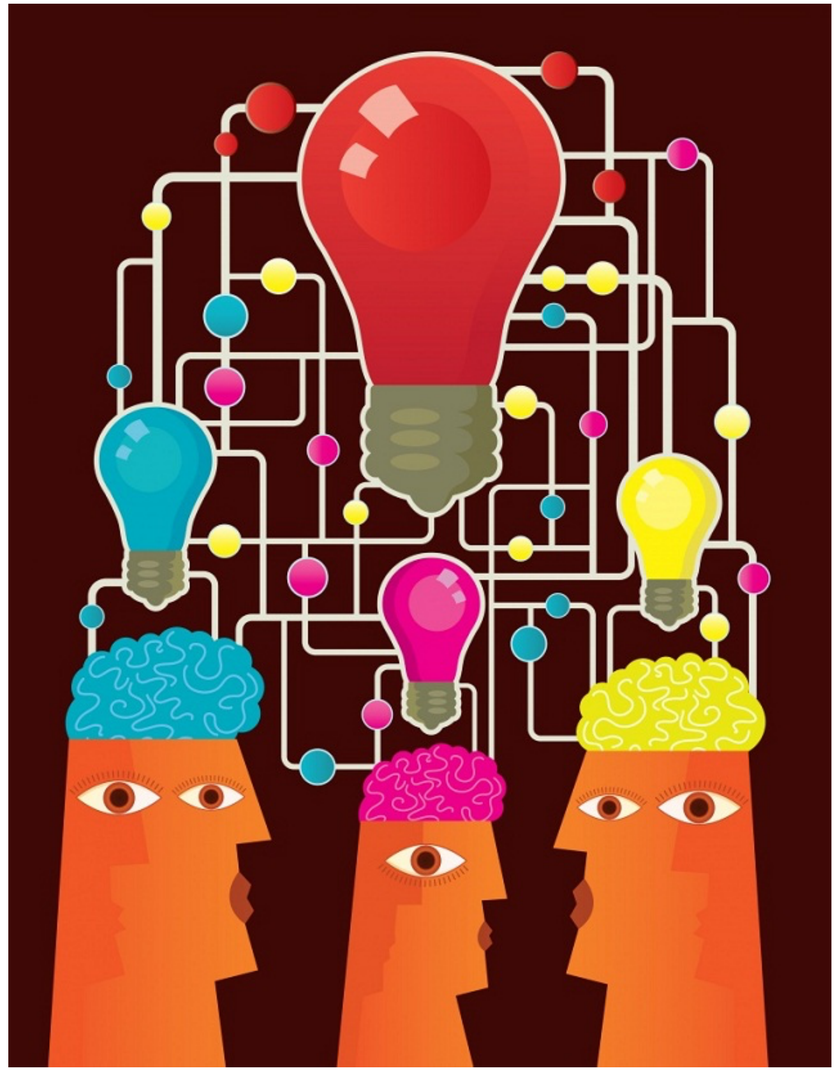
APPLICATION TO MACHINE LEARNING

- **Performance improve with more data**
  - Increases accuracy
  - Scales to larger input data sizes
- **If computational complexity outpaces the main memory**
  - Not scale well due to memory restrictions

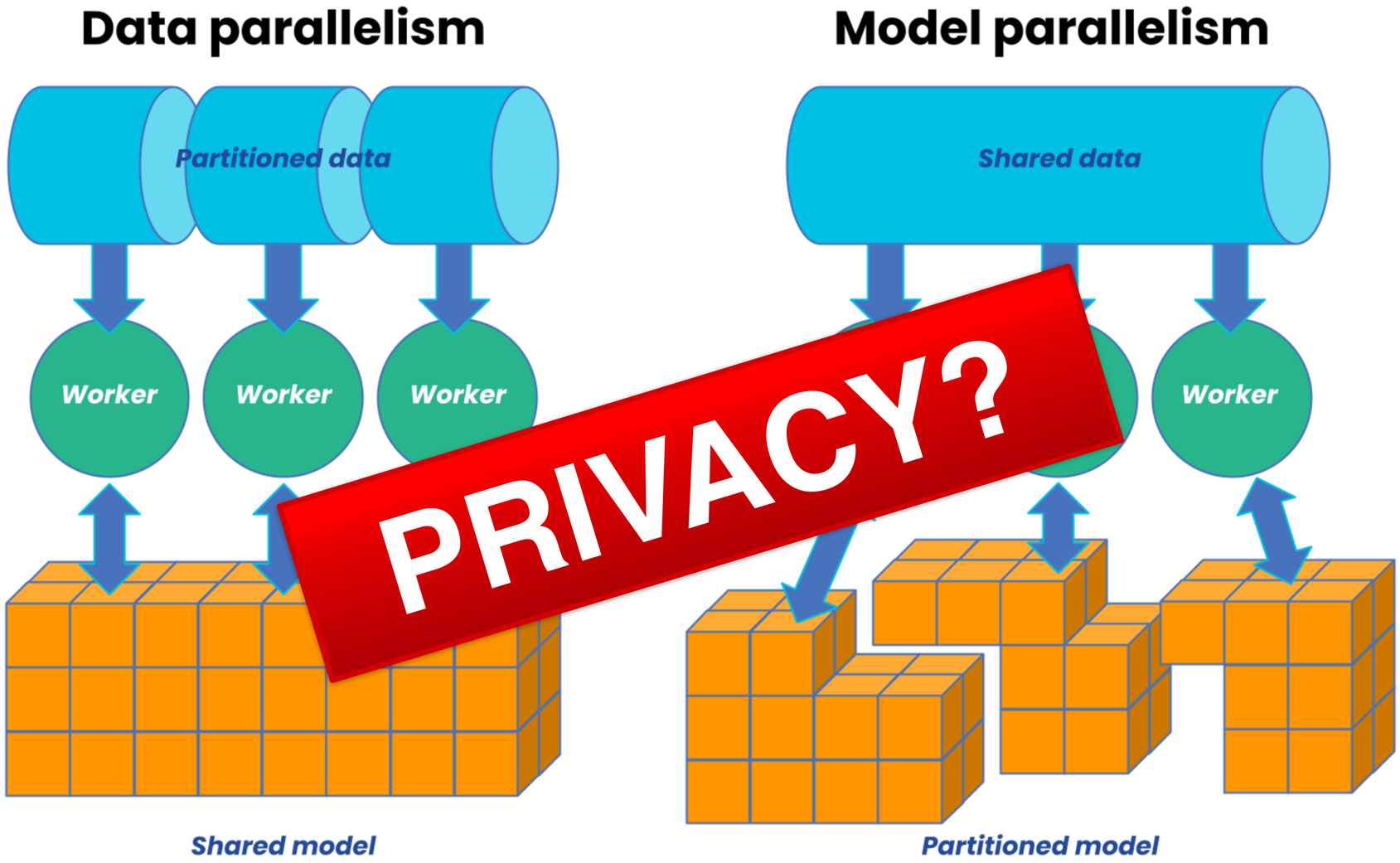


INTRODUCTION OF DISTRIBUTED MACHINE LEARNING

- **Handle large data sets**
- **Develop efficient and scalable algorithms**
- **Ability to allocate learning processes onto several workstations**
  - Enable faster learning algorithms
- **Often used in healthcare or advertising**



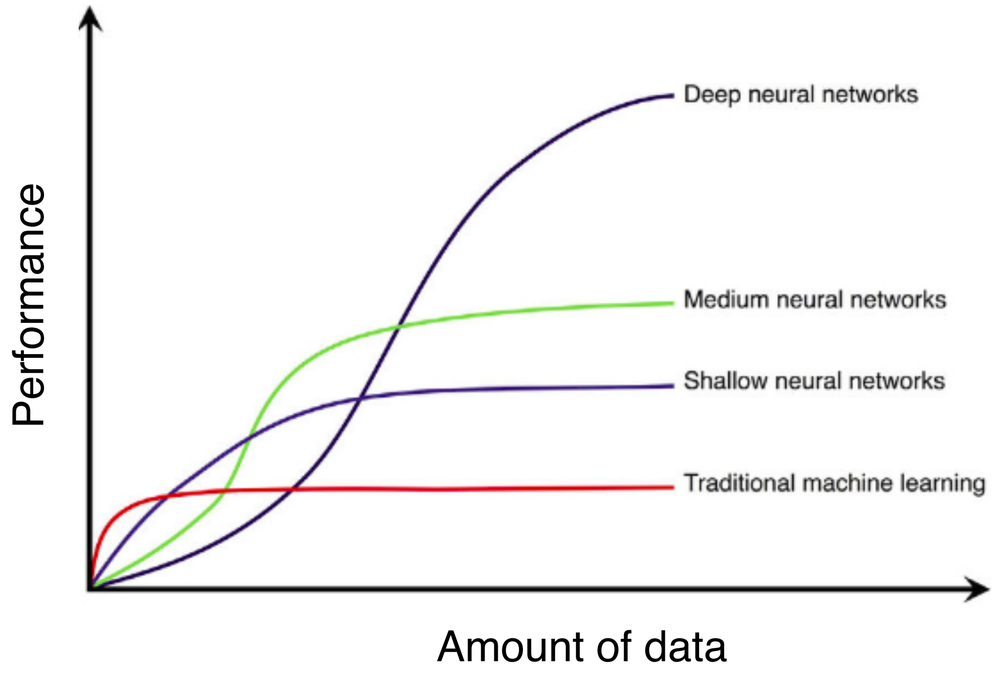
CONCEPT OF DISTRIBUTED MACHINE LEARNING



MOTIVATION FOR FEDERATED LEARNING

- **Performance improve with more data**
- **Models can be meaningfully combined**
- **Nodes can *trains* model, not only predict**

- **Need to preserved privacy at all costs**
  - Interest of HIPAA and GDPR regulations



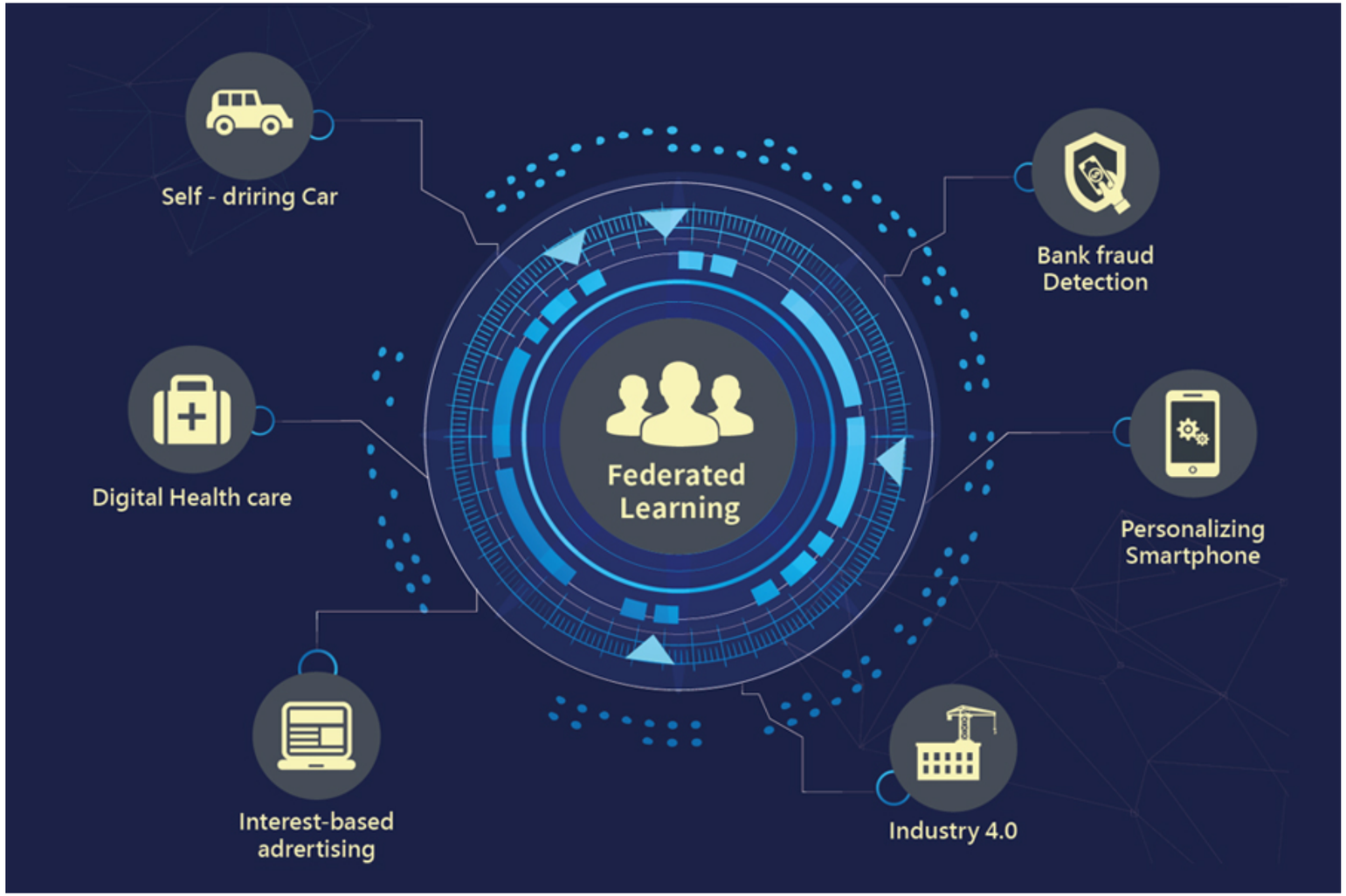
FEDERATED LEARNING IN A NUTSHELL

Institut Mines-Télécom

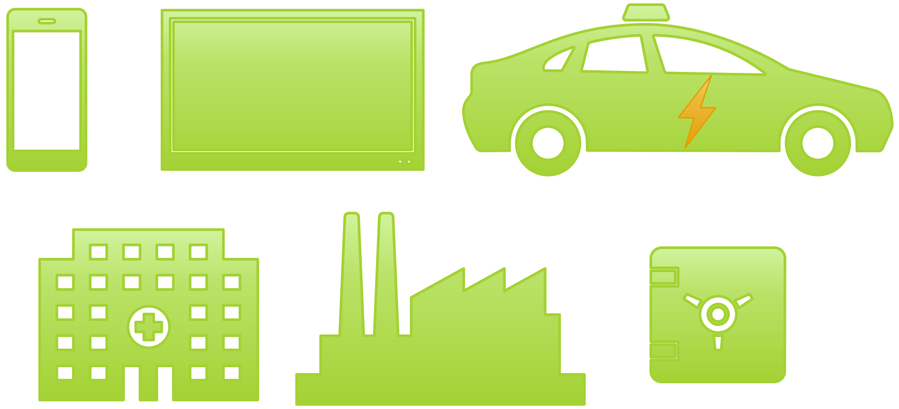
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

FEDERATED LEARNING (FL) APPROACH

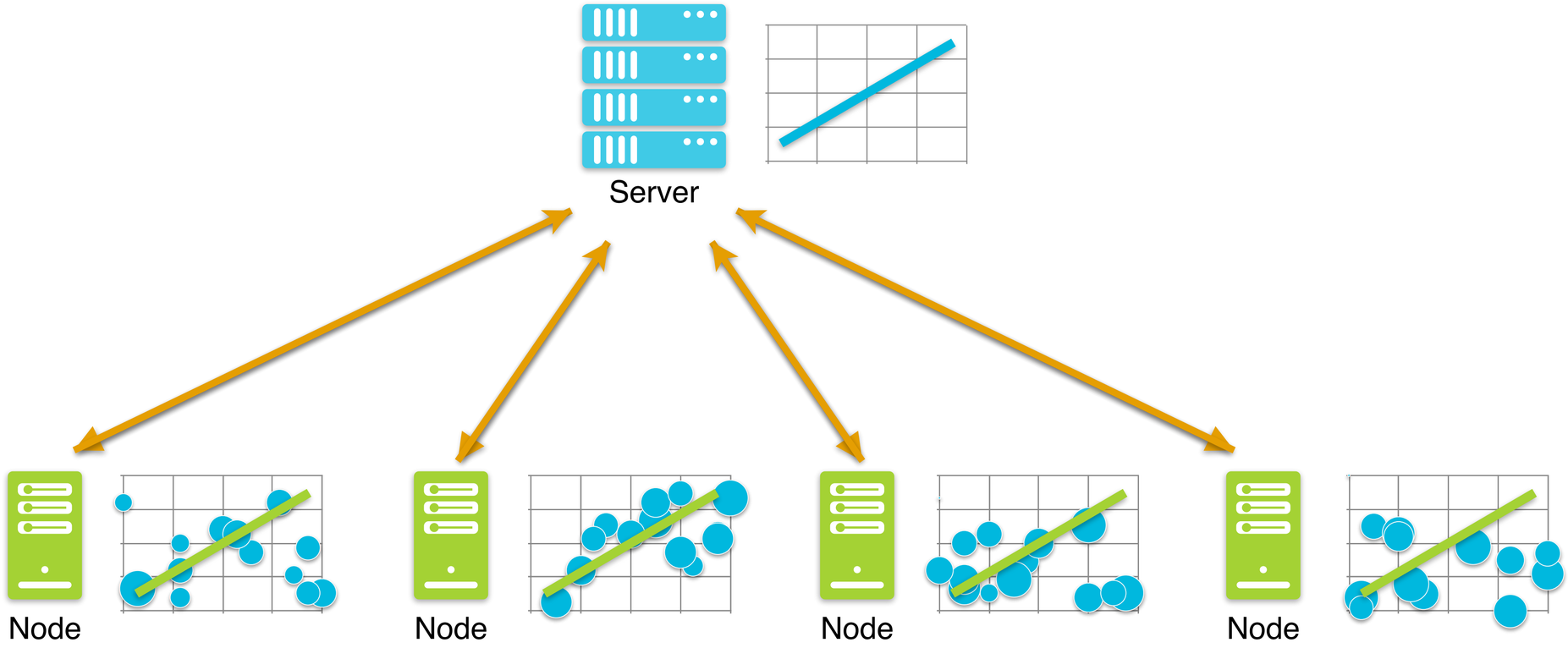


- **Multiple participants**
  - Contributes individually
- **All sort of devices**
  - Users' interaction data
  - Enough processing power



FEDERATED LEARNING PRINCIPLE

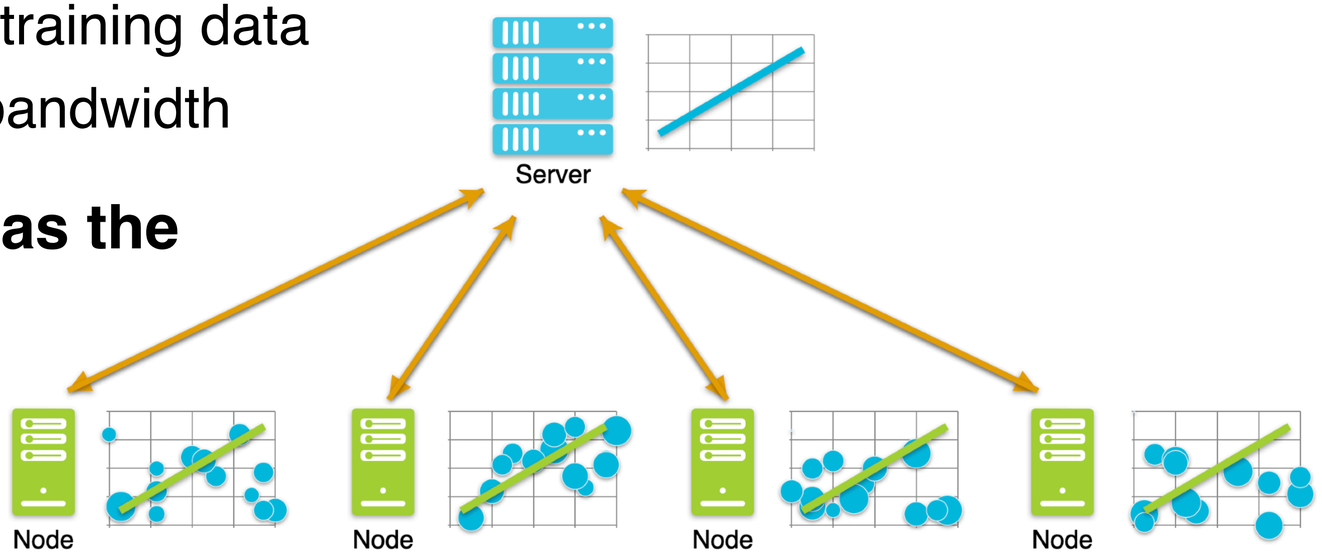
- **A network of nodes shares *models* rather than *training data* with the server**



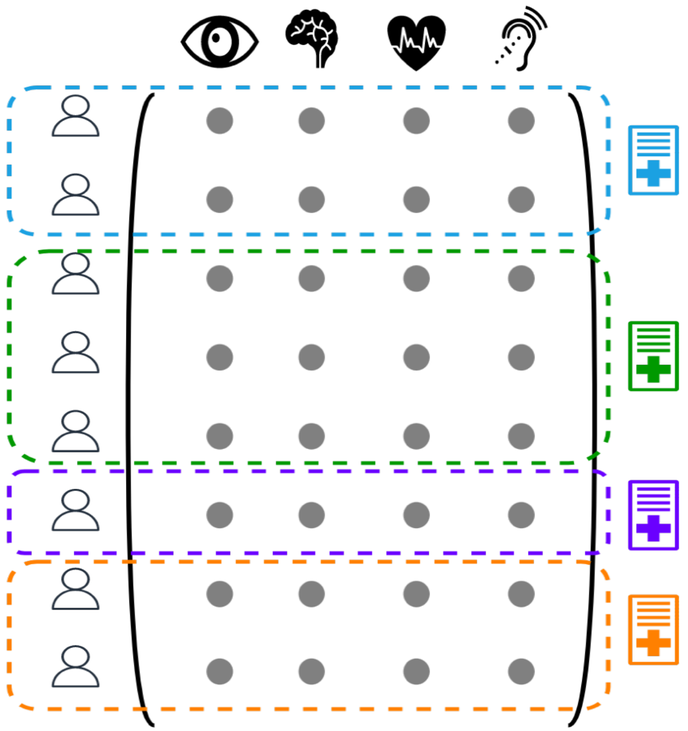
- **We repeat the whole process many times**

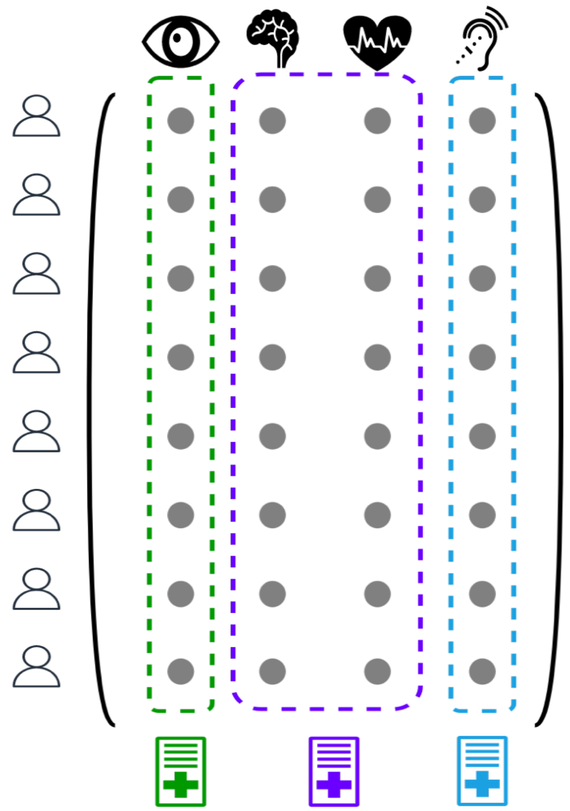
FEDERATED LEARNING PRINCIPLE

- **A network of nodes shares *models* rather than *training data* with the server**
- **We repeat the whole process many times**
- **The server has now a model that captures the pattern in the training data on all the nodes**
  - But, at no point, the nodes share their training data
  - That increases privacy and saves on bandwidth
- **Ideally, the final model is as good as the centralized solution**
  - At least, better than what each party can learn on its own



DIFFERENT TYPE OF FEDERATED LEARNING

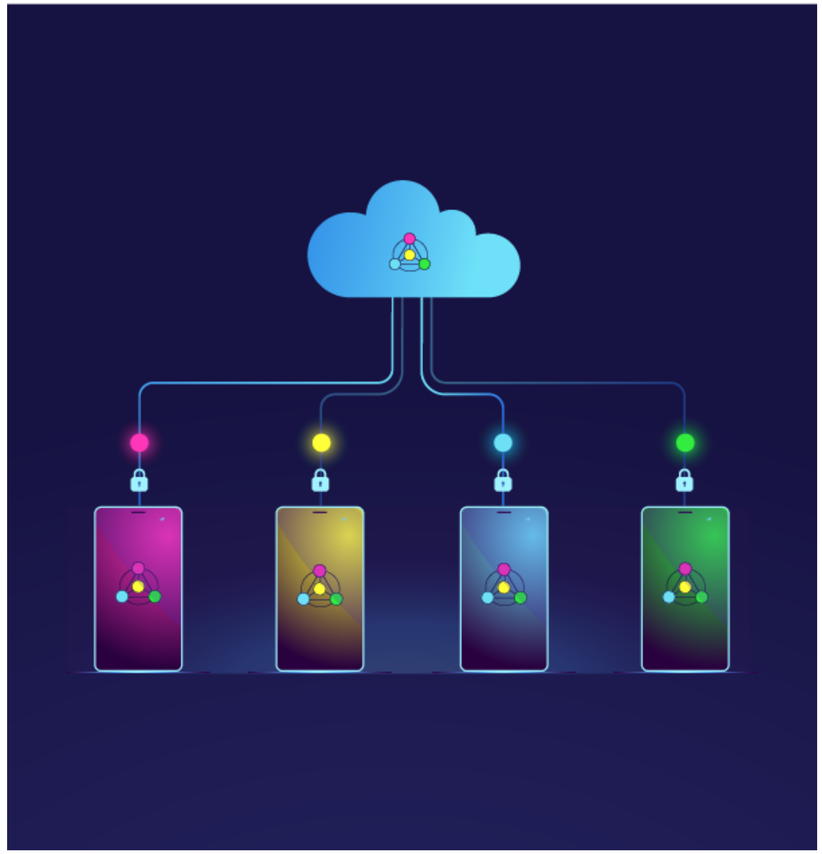




- **Horizontal Federated Learning**
  - Clients share the feature and labels space
  - Differ in the sample space
- **Vertical Federated Learning**
  - Clients share the sample space
  - But neither the feature nor label space

FEDERATED LEARNING CAN HANDLE…

- **Non-IID data**
  - Training data on each node can be idiosyncratic
- **Unbalanced data**
  - Unequal amount of data on each node
- **Massively distributed data**
  - Can have many more devices than training exemples per node
- **Limited communication**
  - Cannot guarantee availability of nodes
- **Training and testing operated on nodes**



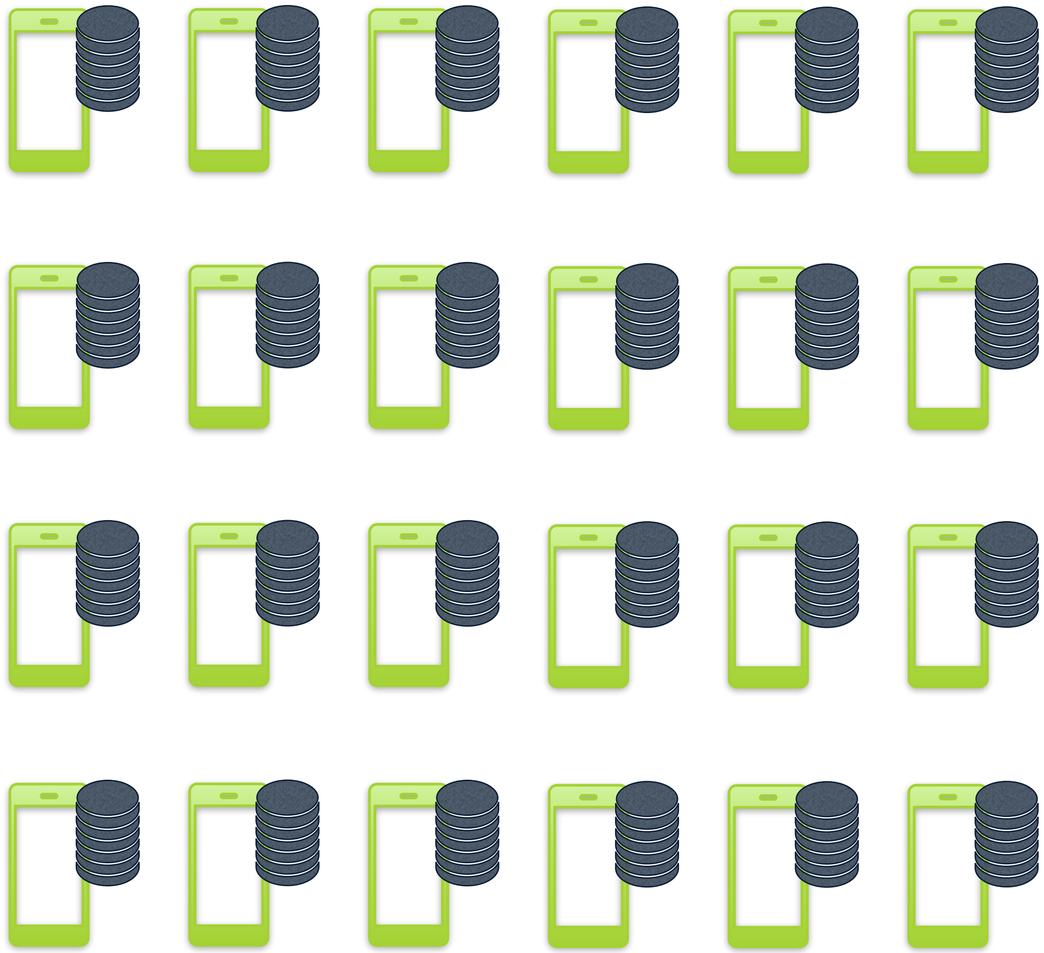
EXAMPLE WITH FEDAVG

Federated averaging

Institut Mines-Télécom
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité
uni.lu
SnT

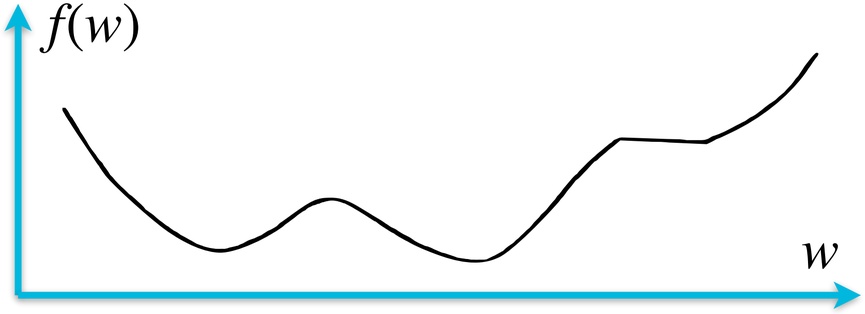
EXAMPLE OF FEDAVG

- **From a pool of candidates**
  - Chooses a subset of *eligible* participants
    - fully charged
    - specific hardware configurations
    - connected to a reliable and free WiFi network
    - idle
- **Not all devices participate in the federation**



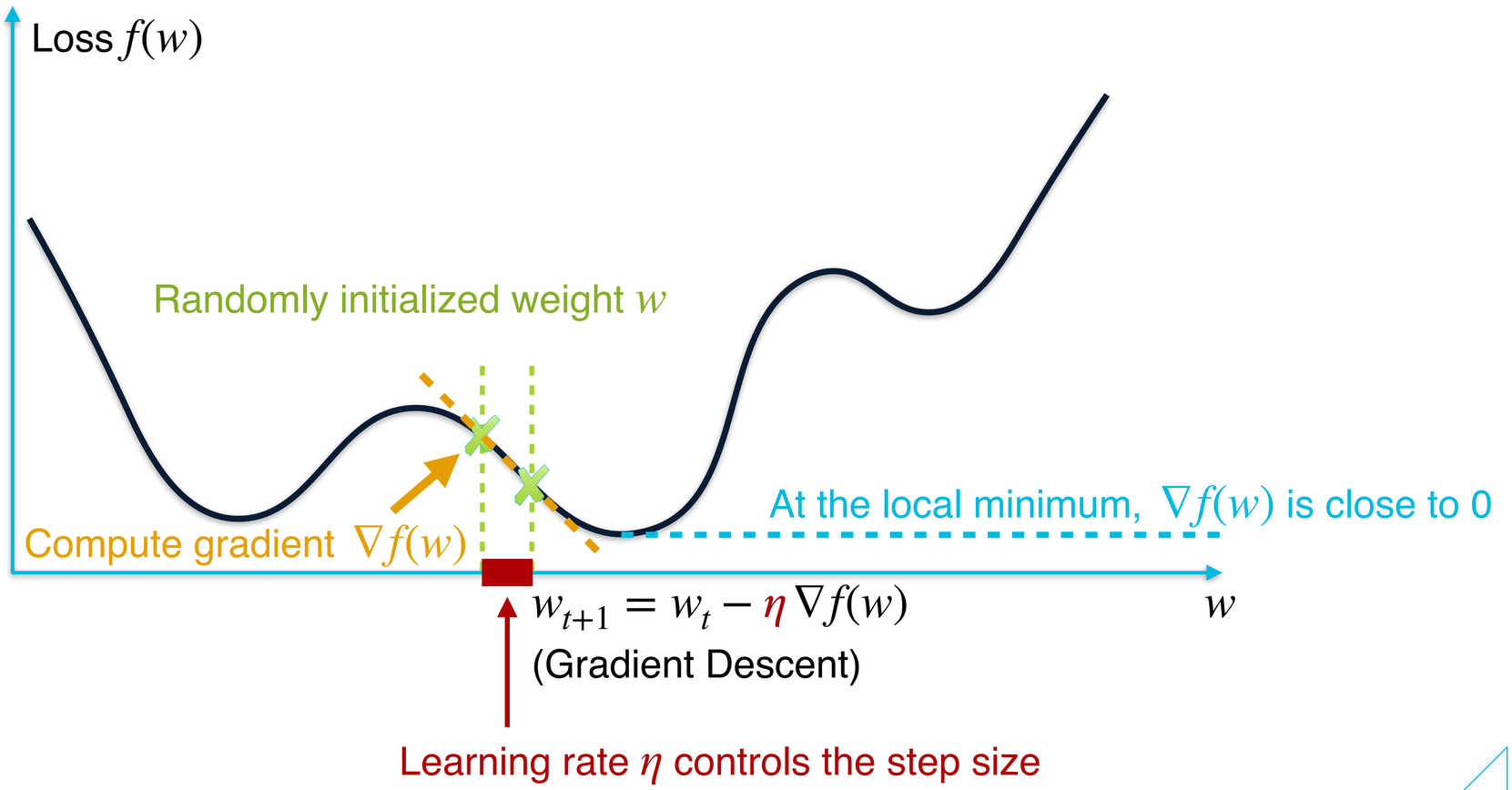
EXAMPLE OF FEDAVG — RECALL OF GRADIENT DESCENT

- **For a training dataset containing n samples $(x_i, y_i)_{1 \leq i \leq n}$, the training objective is**
  - $\min_{w \in \mathbb{R}^d} f(w)$ where $f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$
  - with $f_i(w) = l(x_i, y_i, w)$, the loss of the prediction on example (x_i, y_i)
- **Properties**
  - Non-convex
    - Multiple local minima exist
  - No closed-form solution
    - In a typical deep learning model, w may contain millions of parameters



EXAMPLE OF FEDAVG — RECALL OF GRADIENT DESCENT

- **Solution: Gradient descent**
- **How to stop?**
  - When update is small enough
  - $\|w_{t+1} - w_t\| \leq \varepsilon$ i.e., $\|\nabla f(w_t)\| \leq \varepsilon$
- **Problem**
  - Usualy, the number of training sample n is large
  - Slow convergence



EXAMPLE OF FEDAVG — RECALL OF GRADIENT DESCENT

- **Solution: Stochastic Gradient descent**
  - At each step of gradient descent, instead of compute for all training samples, randomly pick a small subset (*mini-batch*) of training samples (x_k, y_k):

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t, x_k, y_k)$$

- **Compared to gradient descent, SGD takes more steps to converge, but each step is much faster.**

EXAMPLE OF FEDAVG

- **In a round t**
  - The central server broadcasts current model w_t
  - Each client k computes gradient $g_k \leftarrow \nabla F_k(w_t)$, on its local data
  - In other words, each client k computes **for E epochs:** $w_{t+1}^k \leftarrow w_t - \eta \nabla g_k$
  - The central server performs aggregation

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$$

- **Suppose B is the local mini-batch size,**

  #updates on client k in each round: $u_k = E\frac{n_k}{B}$

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:

initialize w_0

for each round $t = 1, 2, \dots$ **do**

$m \leftarrow \max(C \cdot K, 1)$

$S_t \leftarrow$ (random set of m clients)

for each client $k \in S_t$ **in parallel do**

$w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$

$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k$

ClientUpdate(k, w): *// Run on client k*

$\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)

for each local epoch i from 1 to E **do**

for batch $b \in \mathcal{B}$ **do**

$w \leftarrow w - \eta \nabla \ell(w; b)$

return w to server

FEDERATED LEARNING

Challenges and features

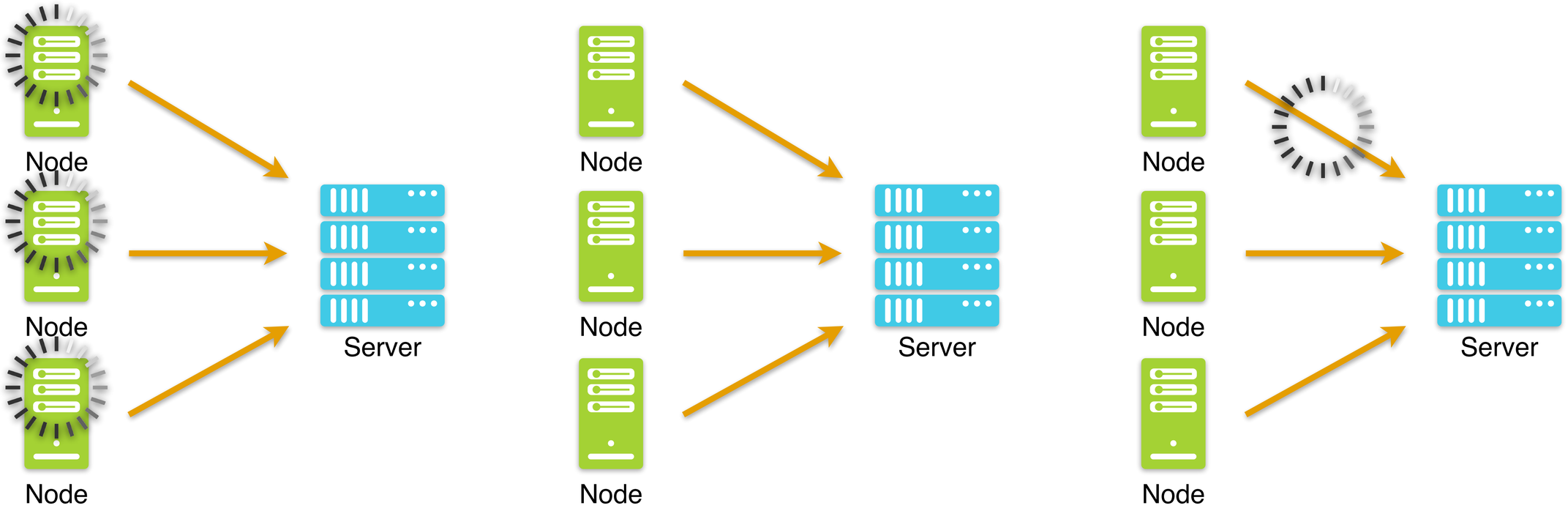
Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

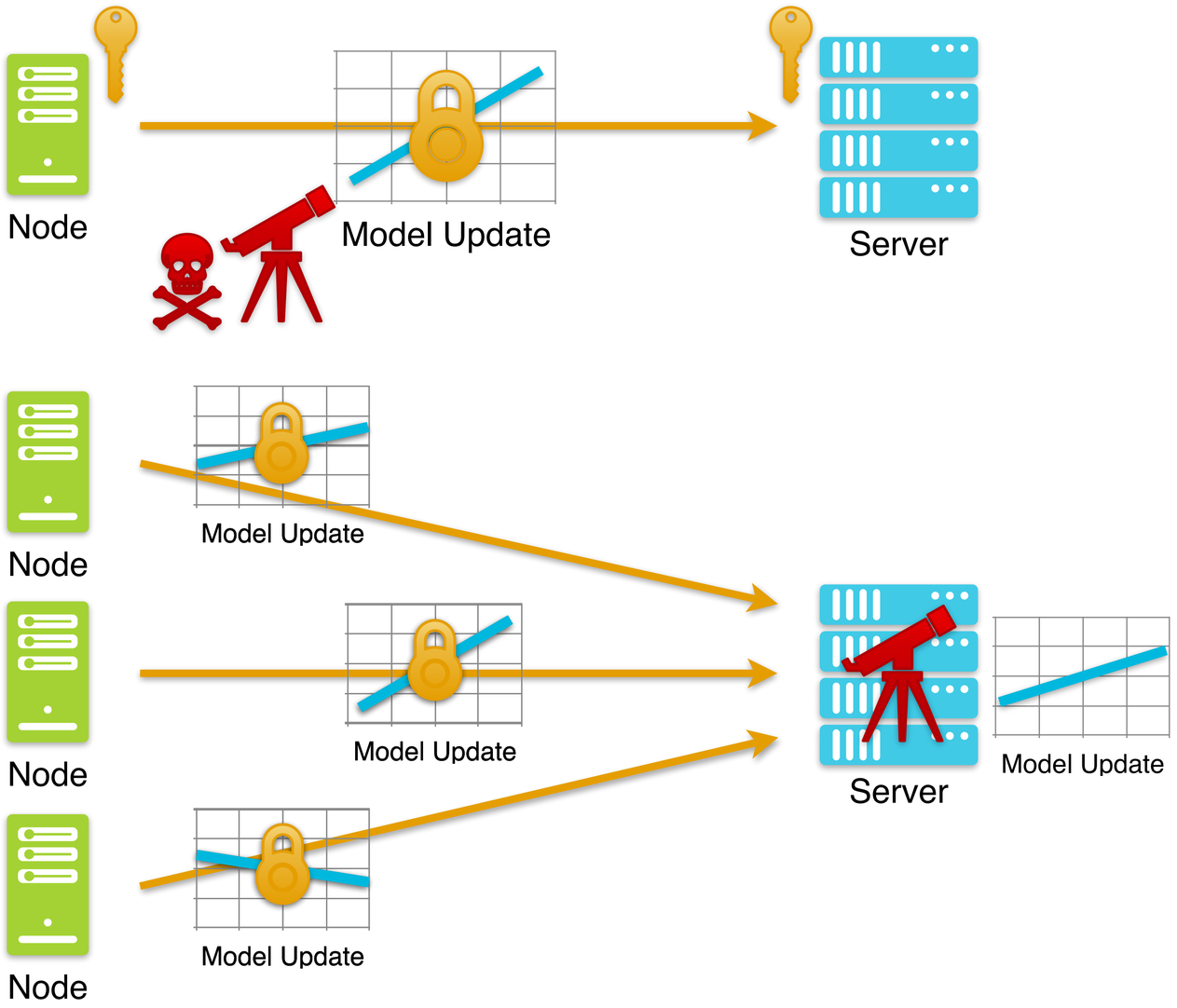
FEDERATED LEARNING CHALLENGES

System issues

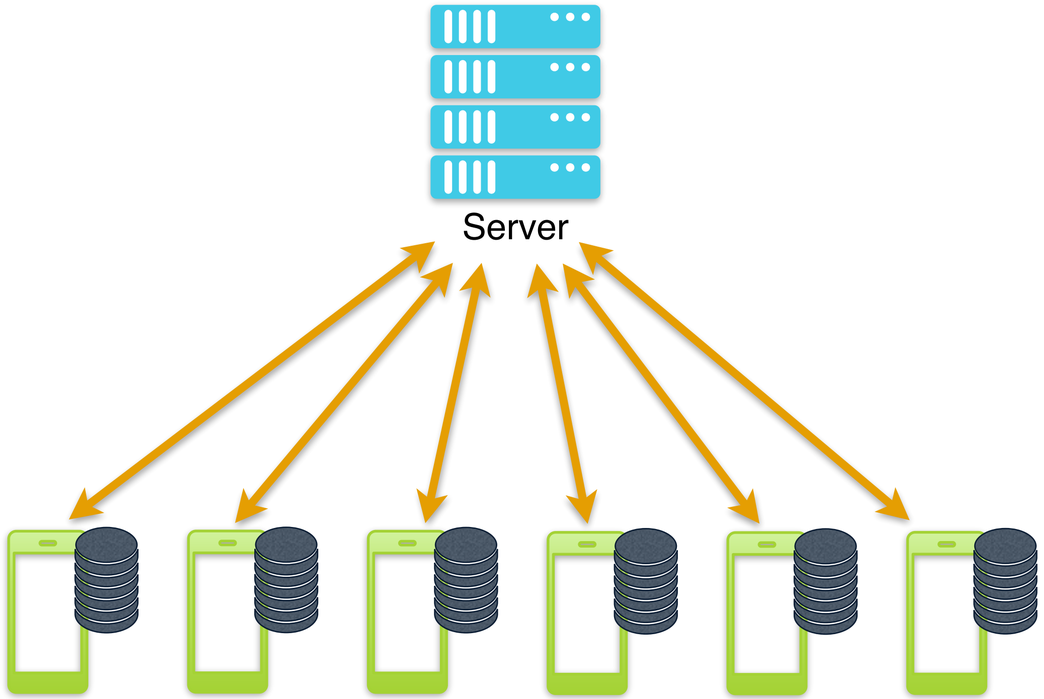


FEDERATED LEARNING CHALLENGES

- **Privacy issues**
  - Man in the middle
    - End-to-end encryption
  - Honest-but-curious server
    - Secure aggregation
    - Differential Privacy for increased security



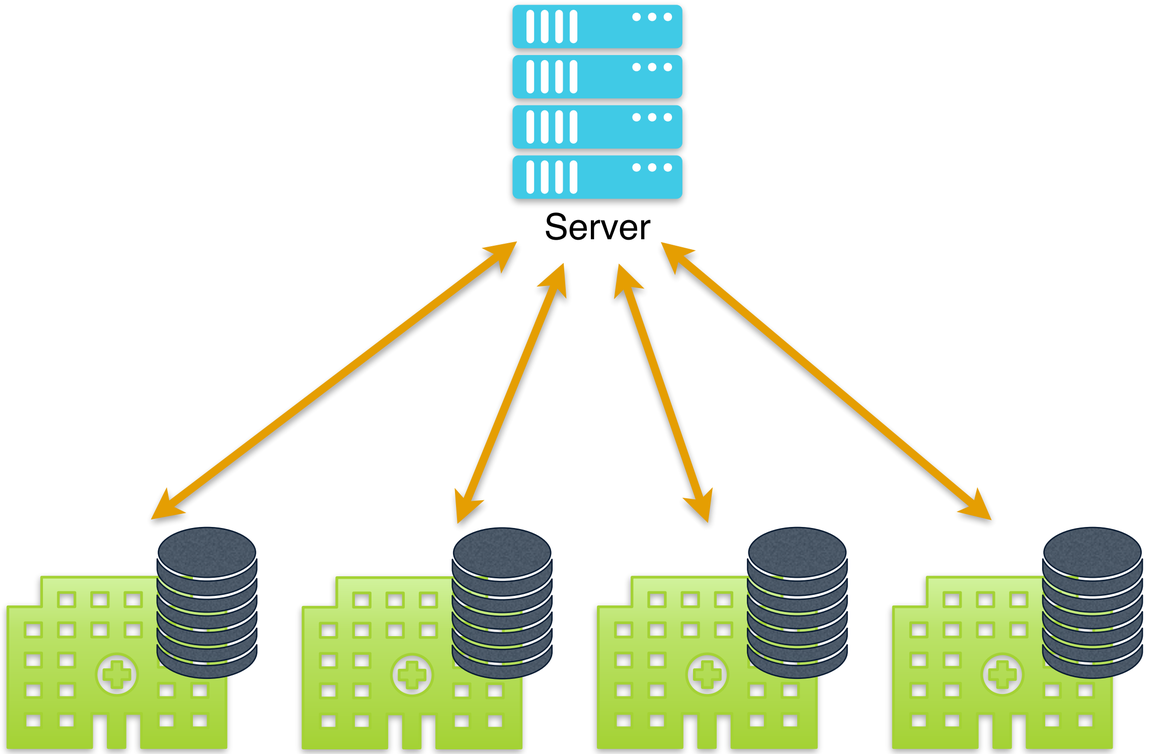
CROSS-DEVICE VS CROSS-SILO FL



Cross-device FL

- **Massive number of parties**
  - up to 10^{10}
- **Small dataset per party**
  - could be size 1
- **Limited availability and reliability**
- **Some parties may be malicious**

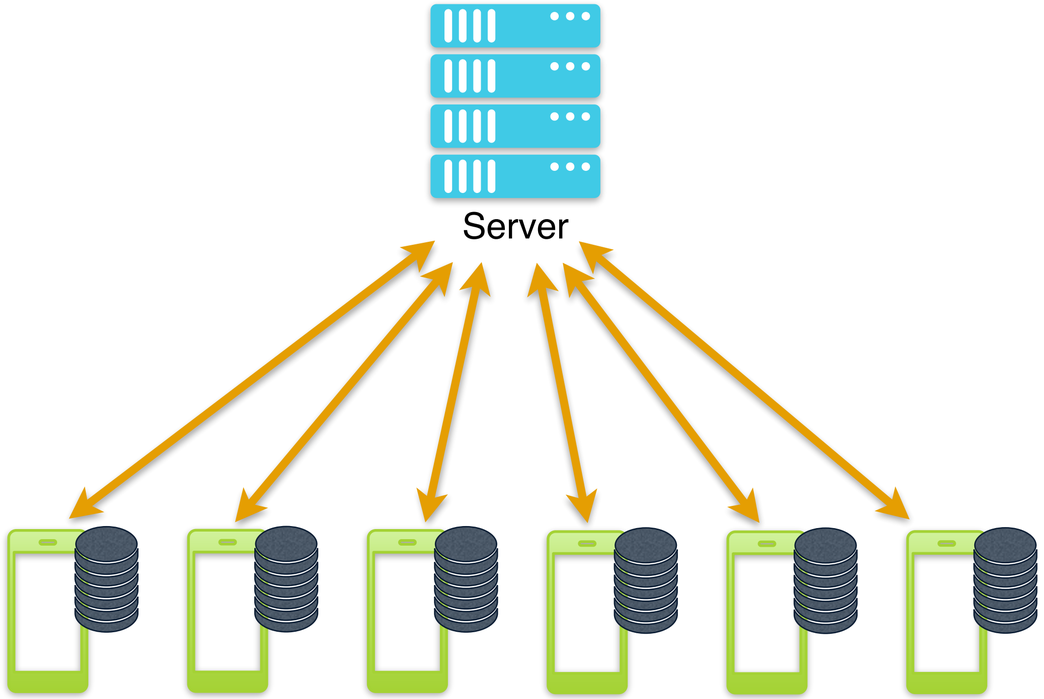
CROSS-DEVICE VS CROSS-SILO FL



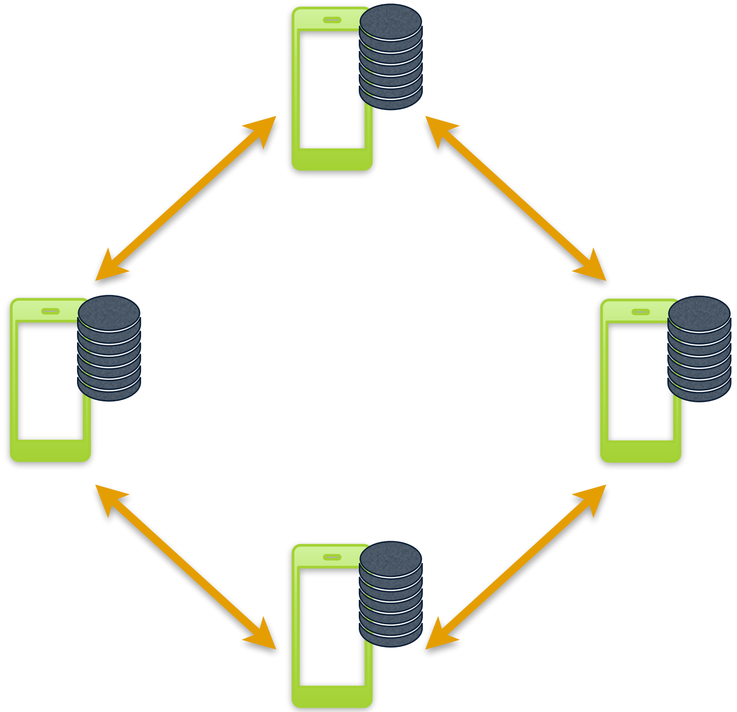
Cross-silo FL

- **2-100 parties**
- **Medium to large dataset per party**
- **Reliable parties**
  - Almost always available
- **Parties are typically honest**

SERVER ORCHESTRATED VS. FULLY DECENTRALIZED FL



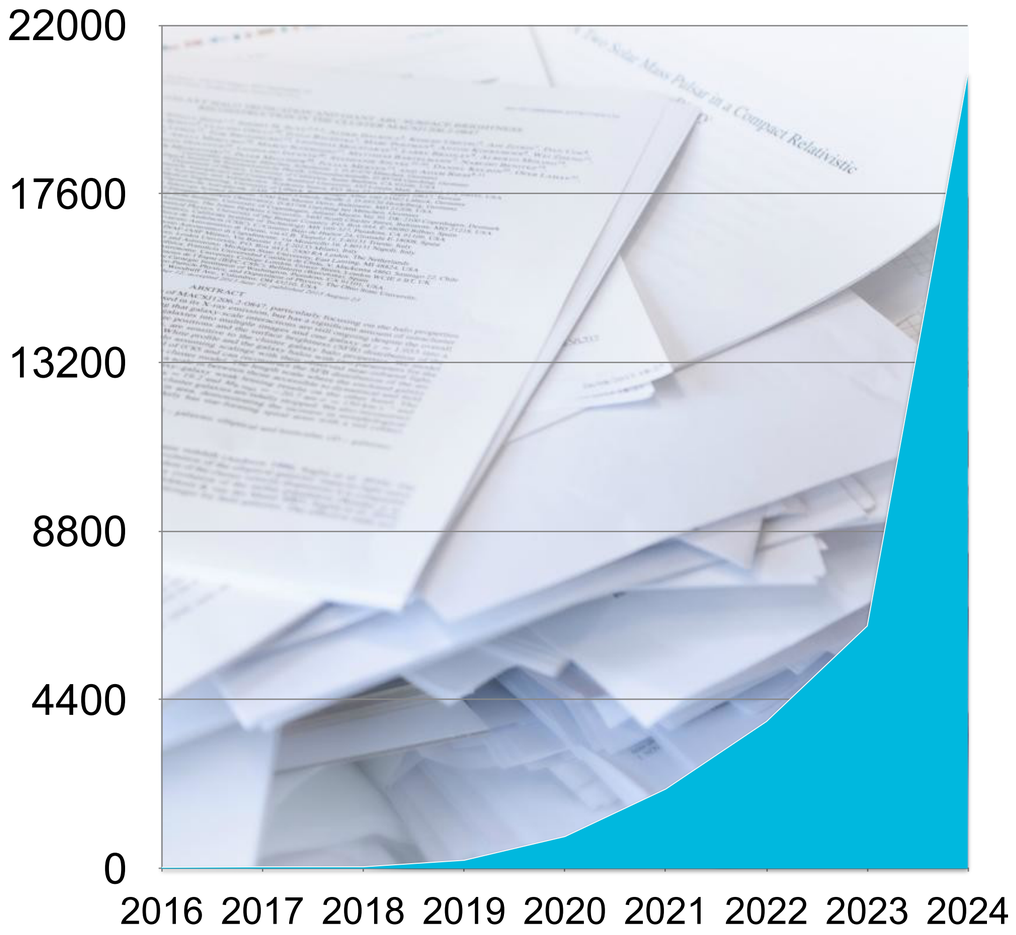
- Server-client communication
- Global coordination, global aggregation
- Server is a single point of failure and may become a bottleneck



- Device-to-device communication
- No global coordination, local aggregation
- Naturally scales to a large number of devices

FEDERATED LEARNING IS A BOOMING TOPIC

- **Historical**
  - 2016: the term FL is first coined by Google researchers
  - 2018: « just » 45 papers on FL (*source: Scopus*)
  - 2023: more than 6k papers on FL! (*source: Scopus*)
- **Some real-world deployments by companies and researchers**
- **Several open-source libraries are under development:**
  - Flower, PySyft, TensorFlow Federated, FATE, Substra...
- **FL is highly multidisciplinary**
  - Involve machine learning, numerical optimization, privacy & security, networks, systems, hardware...



FEDERATED LEARNING IS A BOOMING TOPIC

- **Challenges from the Literature** [4]
  - Functionality: performance, heterogeneity, transferability, self-defense, and self-healing.
  - Deployment: adaptability and scalability
  - Security and reliability: security, privacy, trust, and reputation
  - Experimentation: evaluation

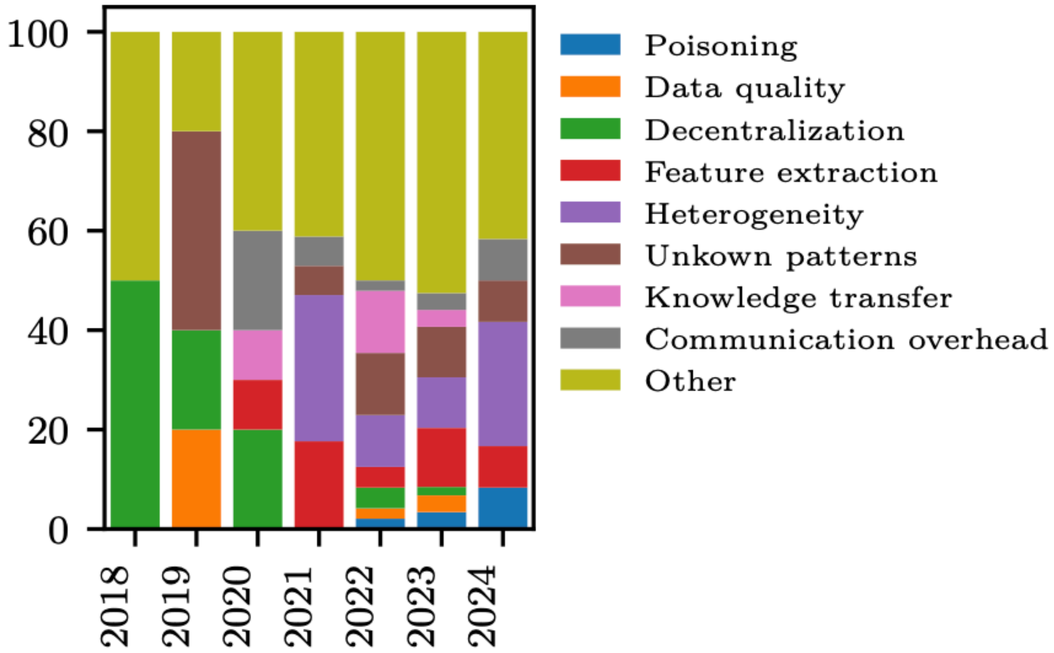


Figure: Challenges addressed by the literature (until 2024-04).

[4] L. Lavaur, et al., “The Evolution of Federated Learning-Based Intrusion Detection and Mitigation: A Survey”, IEEE Transactions on Network and Service Management, 2022

HANDS-ON! — PART 1

FEDERATED LEARNING IN A NUTSHELL

https://tinyurl.com/FLxNS-part1

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

THE POWER OF FEDERATED LEARNING FOR NETWORK SECURITY

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

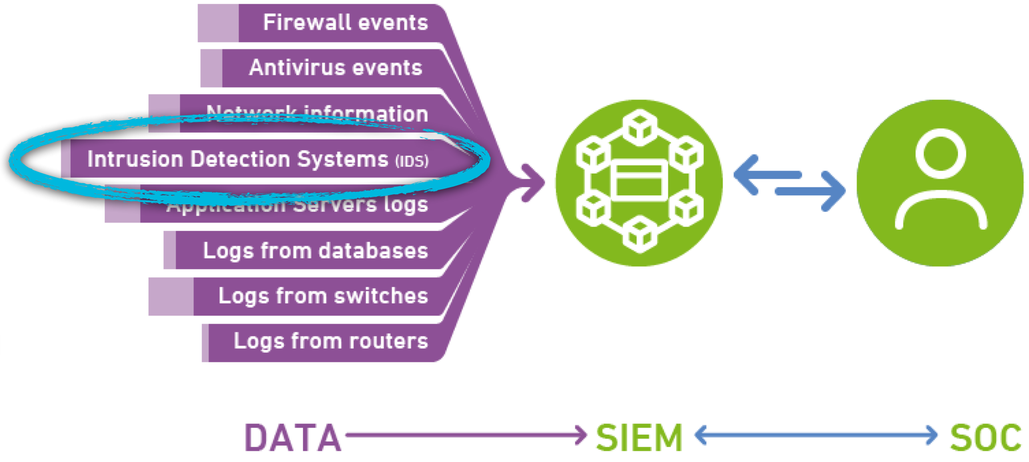
uni.lu | SnT

INVOLVEMENT OF AI IN CYBER-ATTACK DETECTION

- **How recent artificial intelligence methods can be applied to cyber-attacks?**
  - Drastically improve detection and even remediation mechanisms
  - Take into account 0-day vulnerabilities and attacks

- **SOC/SIEM level in particular**
  - Detection of APT or Smart-DDoS for instance

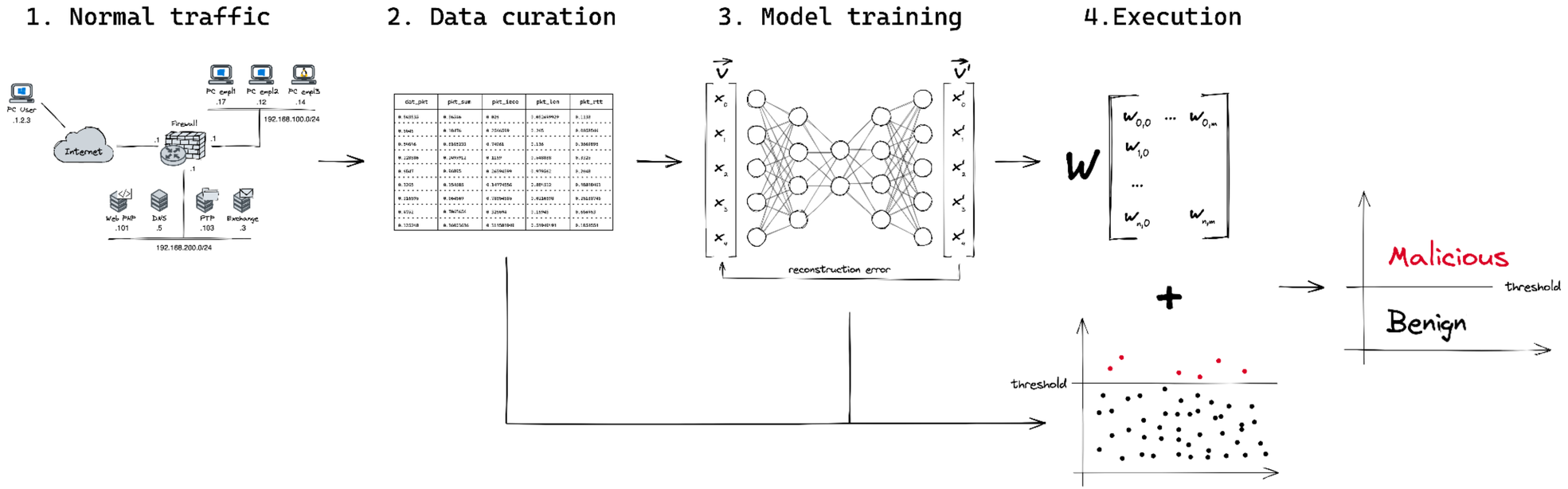
- **Federated/collaborative approaches**
  - Federated Learning for Cyber-Attack Detection



* SOC = Security Operation Center
* SIEM = Security Information and Event Management
* APT = Advanced Persistent Threat
* DDoS = Distributed Denial of Service

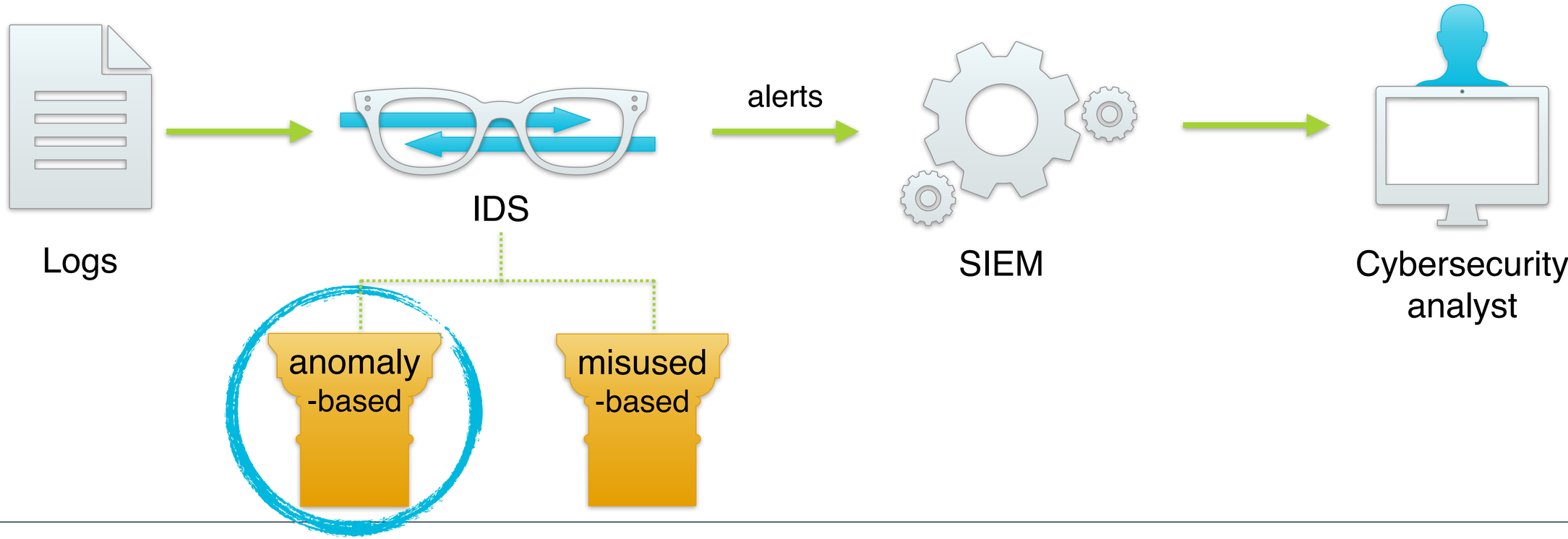
USE-CASE: INTRUSION DETECTION

- **Different families:**
  - misuse detection, anomaly detection, specification-based…
- **Machine learning (ML) and deep learning (DL) often used for their performance**
  - *e.g.*, auto-encoder (AE) can be used for anomaly detection.
- **DL need a lot of data to be efficient, training them locally is a challenge**
  - *e.g.*, for AE, anything not known is an anomaly → higher false-positive rate.



CURRENT INDUSTRIAL APPROACH

- **Intrusion Detection System (IDS) & Security Information and Event Management (SIEM)**
  - Individual alerts without context
  - Investigation leads analysts to alert fatigue



MULTI-STEP ATTACKS: EXTRACTION OF PROBABLE SCENARIOS BY CORRELATION OF ALERTS

Joint work with Yann Busnel (Institut Mines-Télécom)
Antoine Rebstock, Romaric Ludinard (IMT Atlantique)
& Stéphane Paquelet (IRT b<>com)

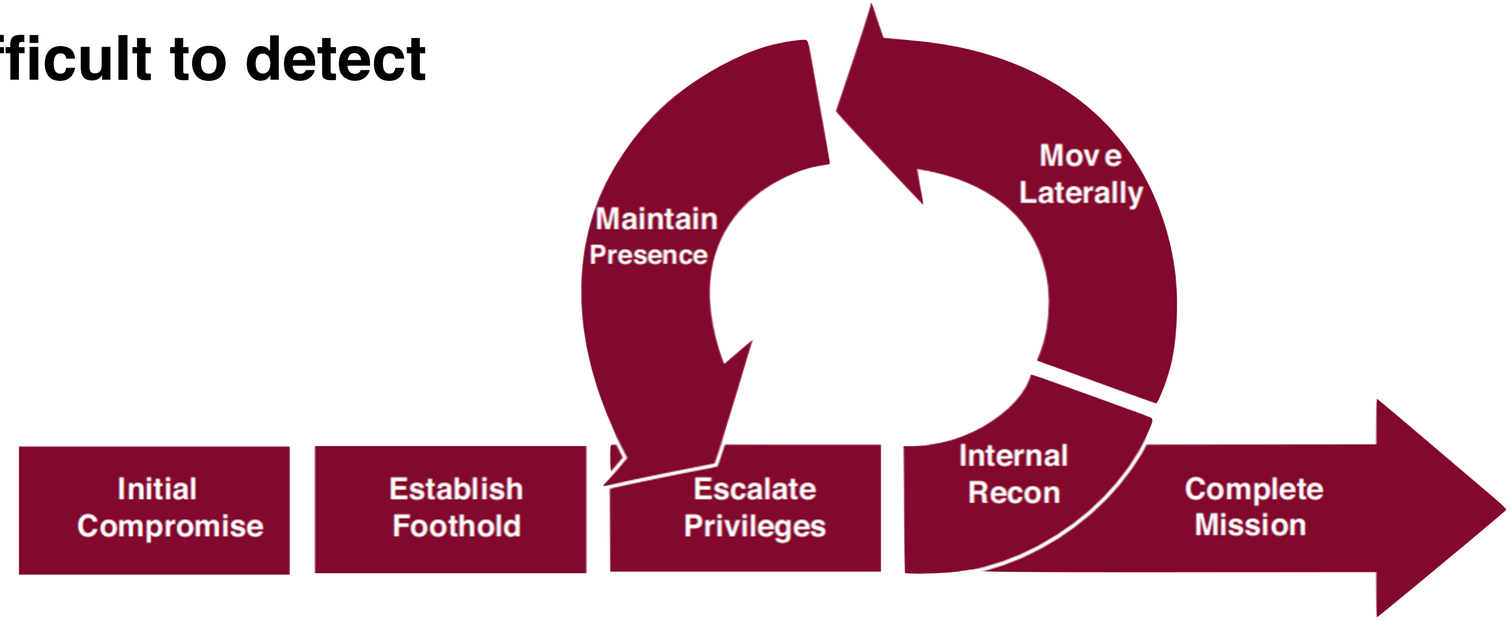
Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

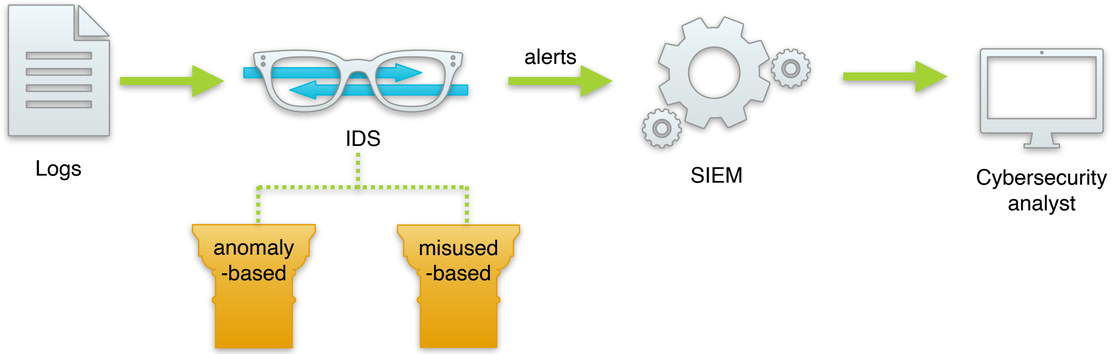
uni.lu | SnT

APT ATTACKS DEFINITION

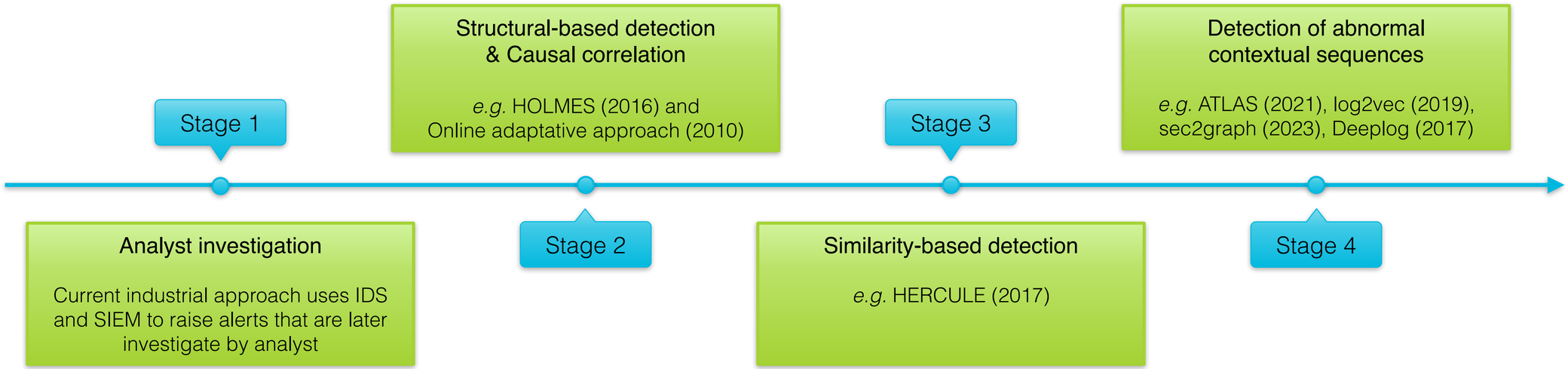
- **APT = Advanced Persistent Threat**
  - Attacker usually has to perform **several actions consecutive actions**
  - As known as **multi-step attacks** and can potentially go **undetected for a long time**
  - Some of the steps of the attack can potentially be seen as a **legitimate set of actions**
- **More difficult to detect**



ON THE ROAD TO AUTOMATIC DETECTION

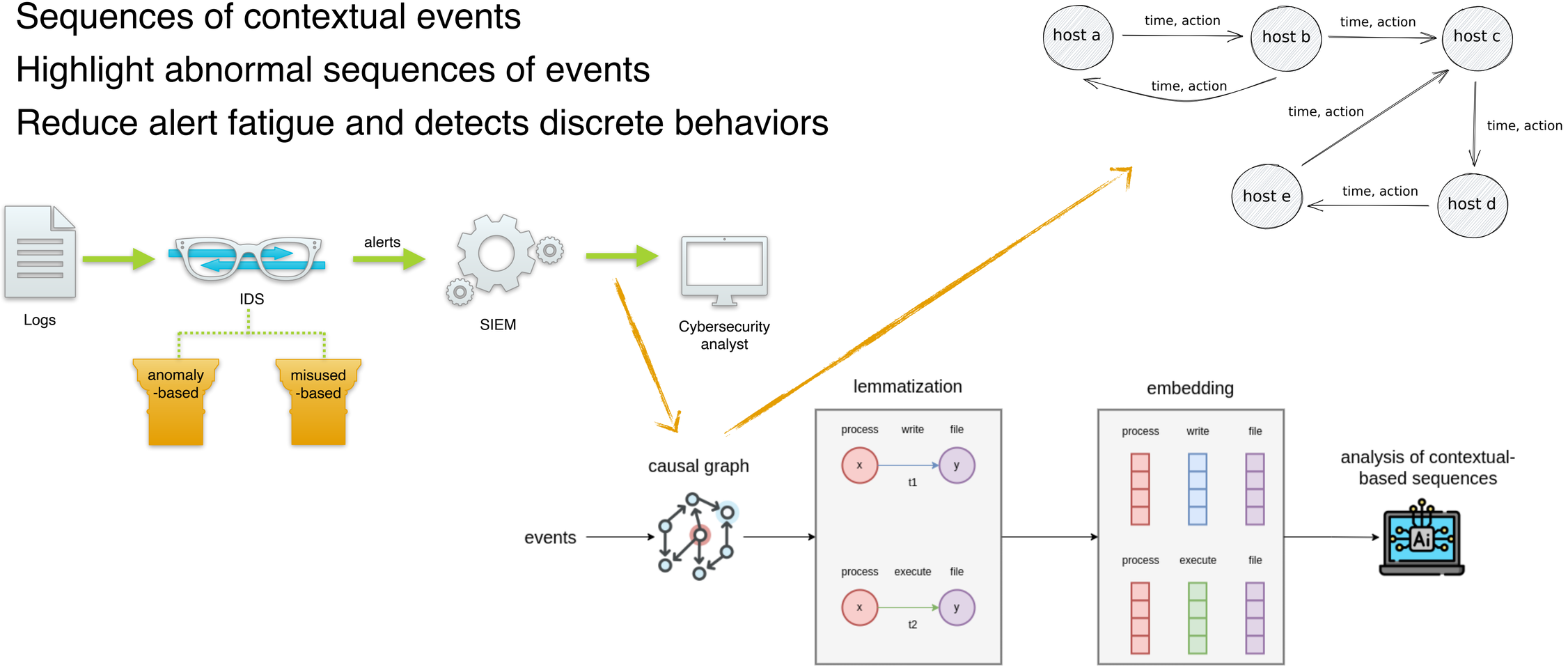


	Case-based	Structural-based	Causal correlation	Similarity-based
Human knowledge	✓	✓	✗	✗
Unknown attacks detection	✗	✗	Only light variations	✓



OUR PROPOSITION : USING AI TO RECONSTRUCT ATTACK HISTORY

- **Decision support**
  - Sequences of contextual events
  - Highlight abnormal sequences of events
  - Reduce alert fatigue and detects discrete behaviors



FEDERATED LEARNING APPROACHES FOR DEFENDING AND DETECTING CYBER-ATTACKS

Joint work with Yann Busnel (Institut Mines-Télécom)
Leo Lavaur (University of Luxembourg)
Fabien Autrel and Marc-Oliver Pahl (IMT Atlantique)

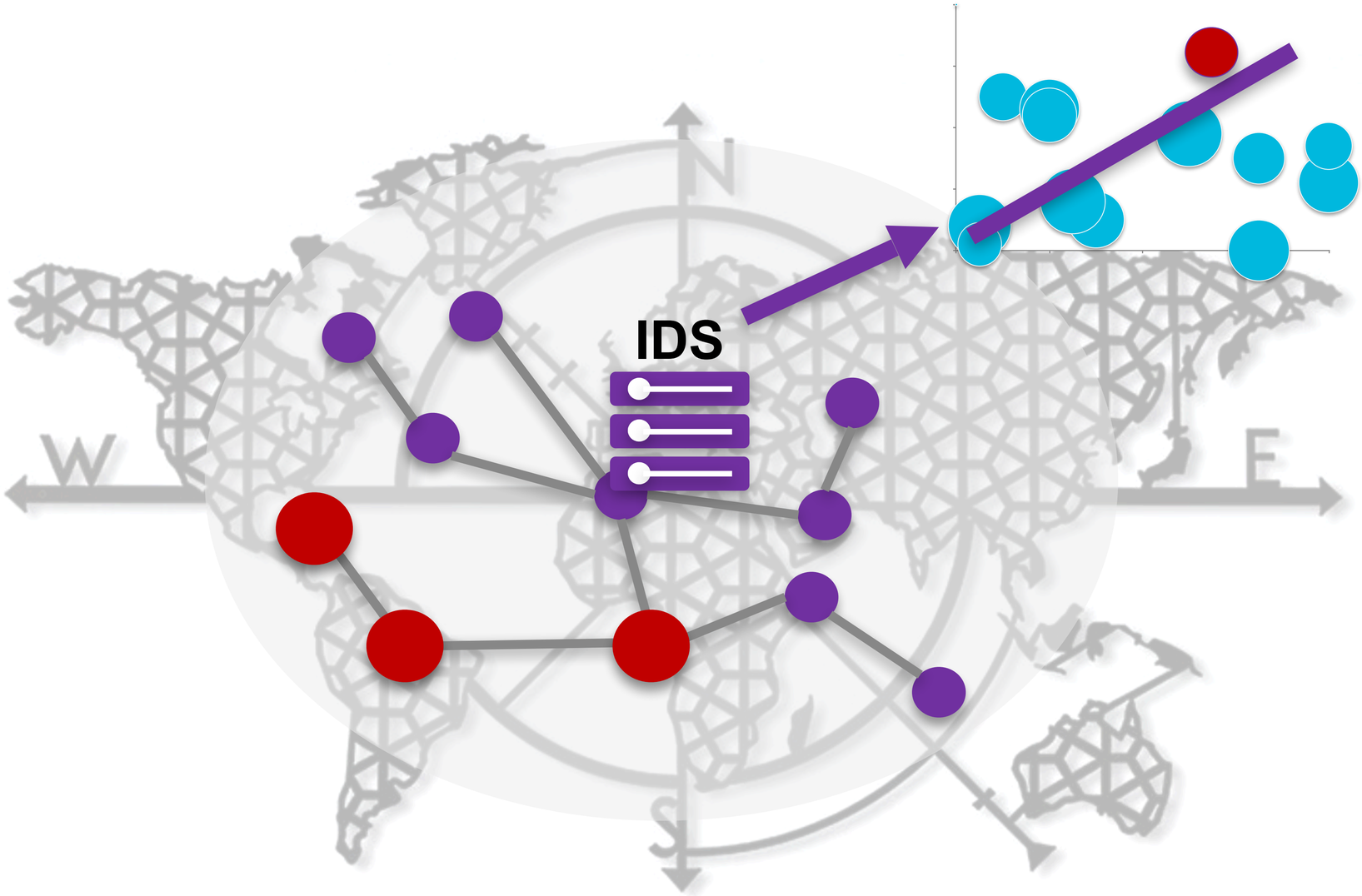
Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

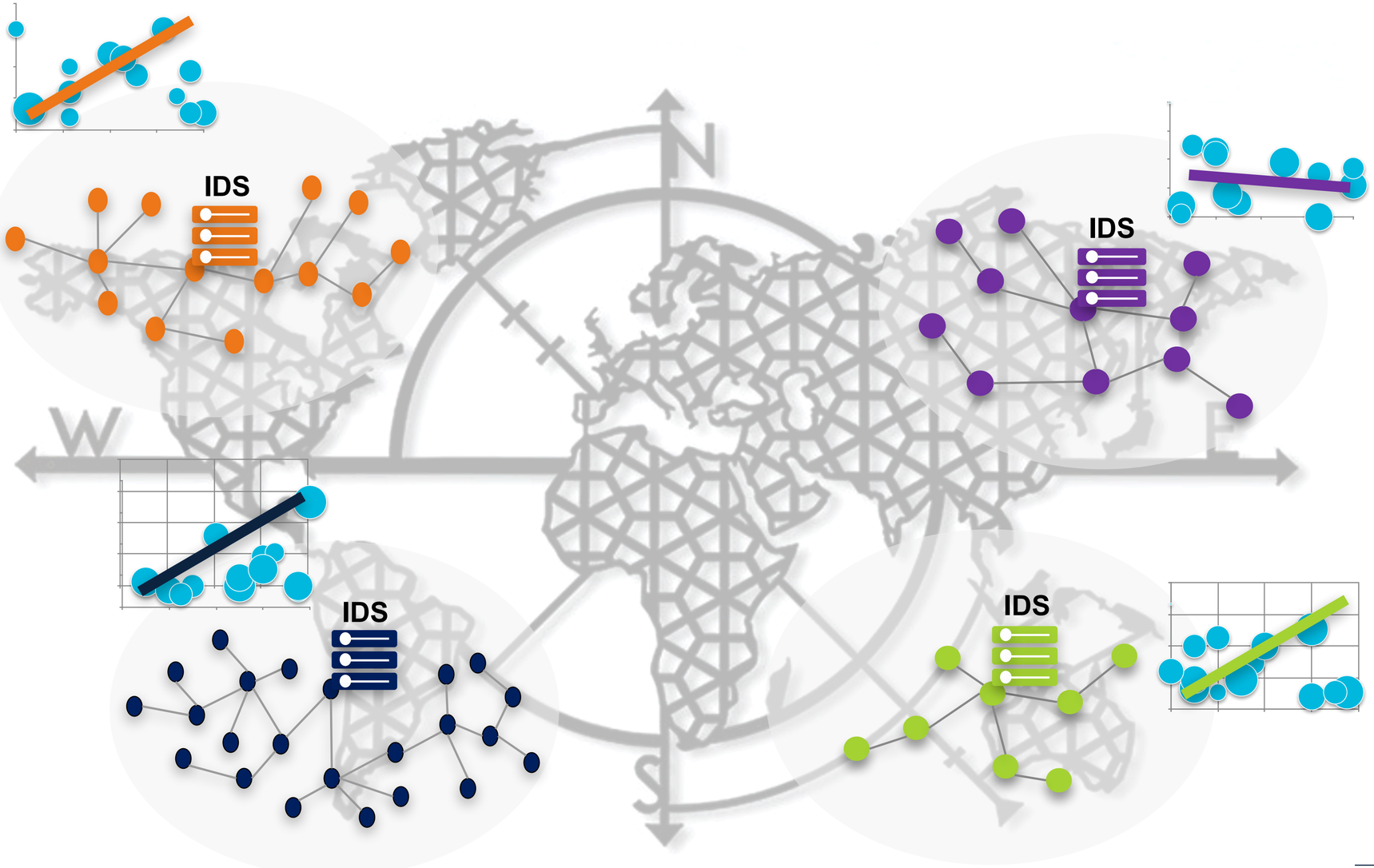
SECURITY MONITORING

Cyberattack detection in infrastructures



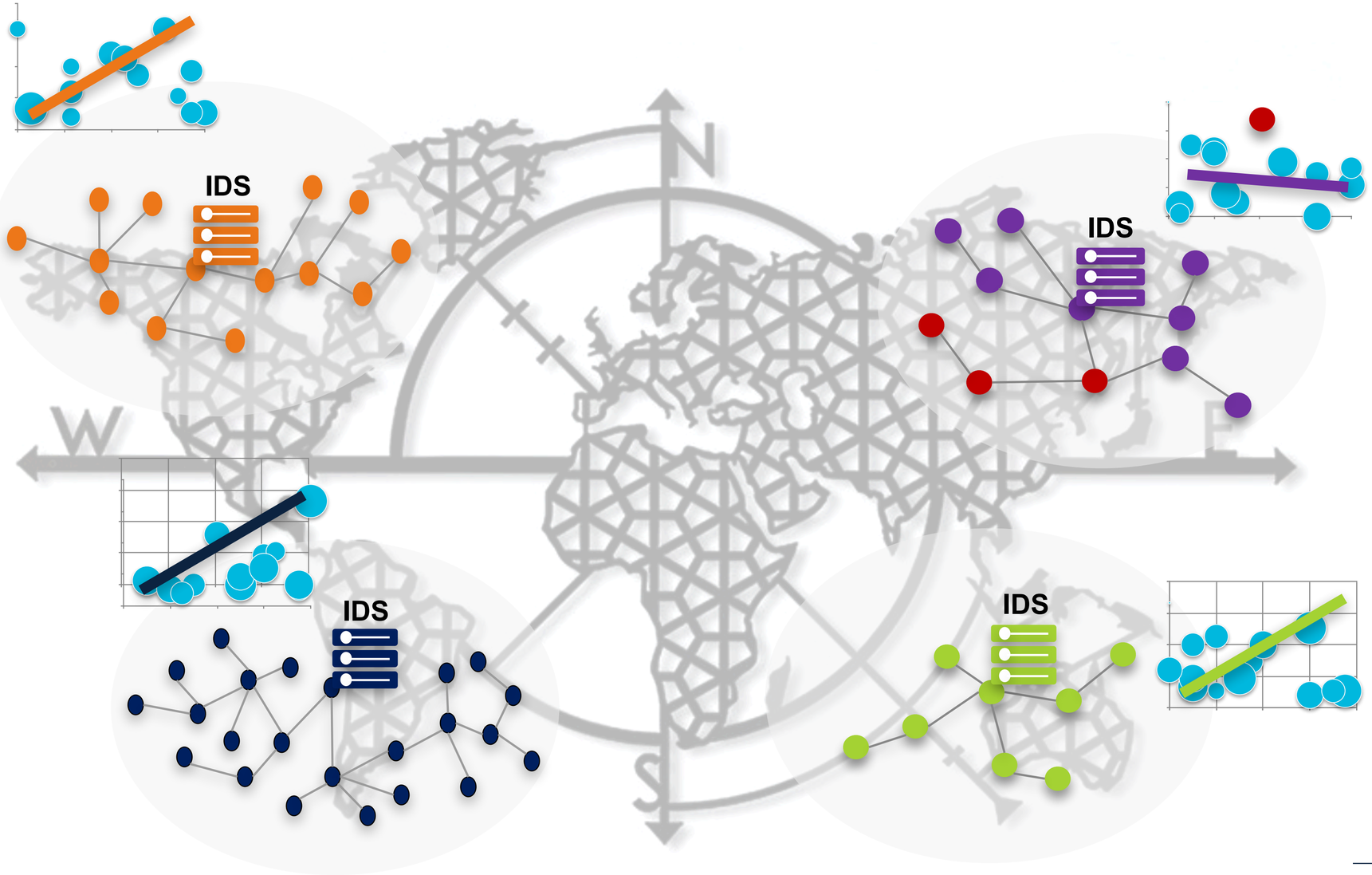
AT MULTIPLE ORGANISATION SCALE

How to share experience in intrusion detection?



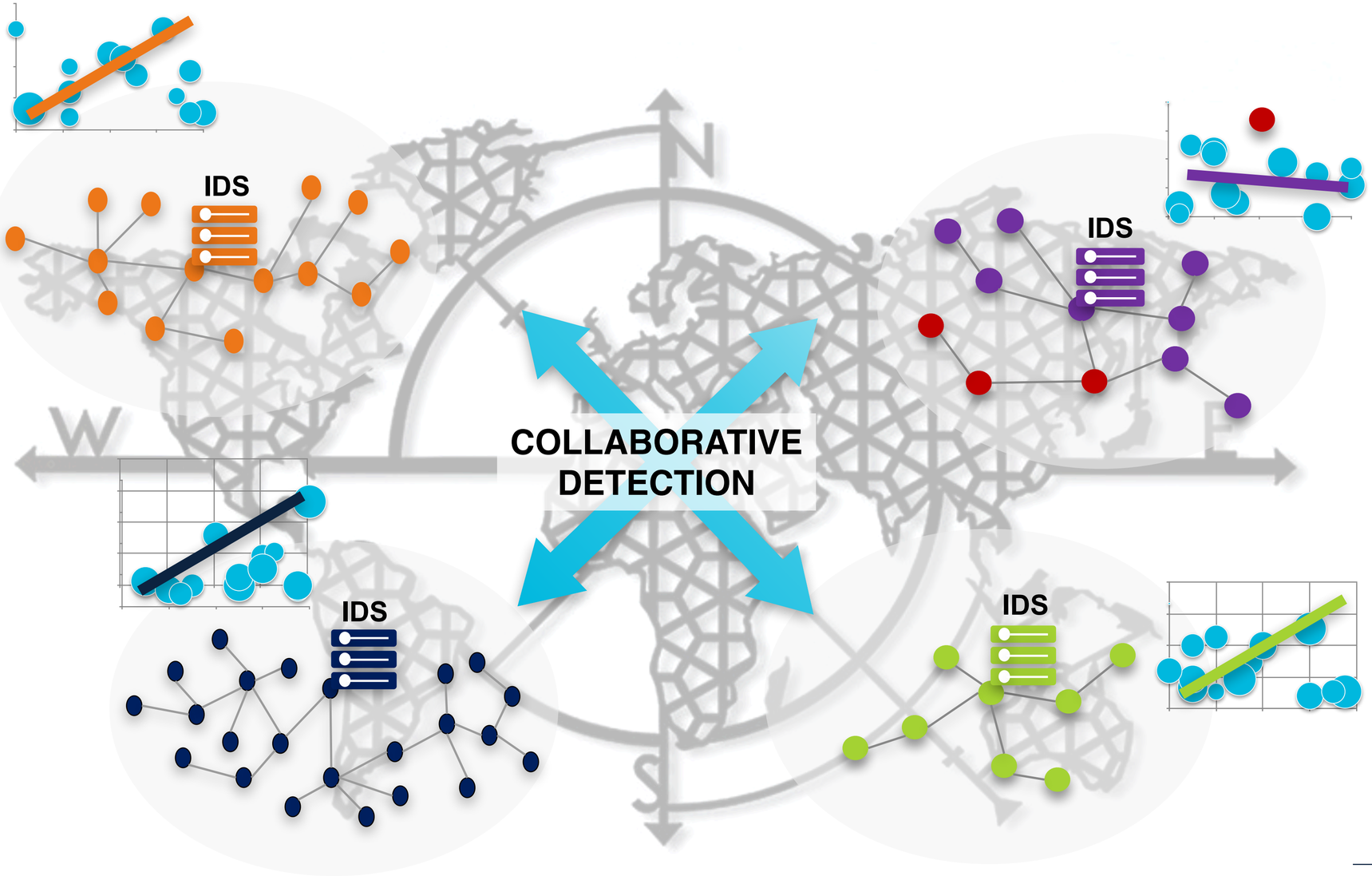
AT MULTIPLE ORGANISATION SCALE

How to share experience in intrusion detection?



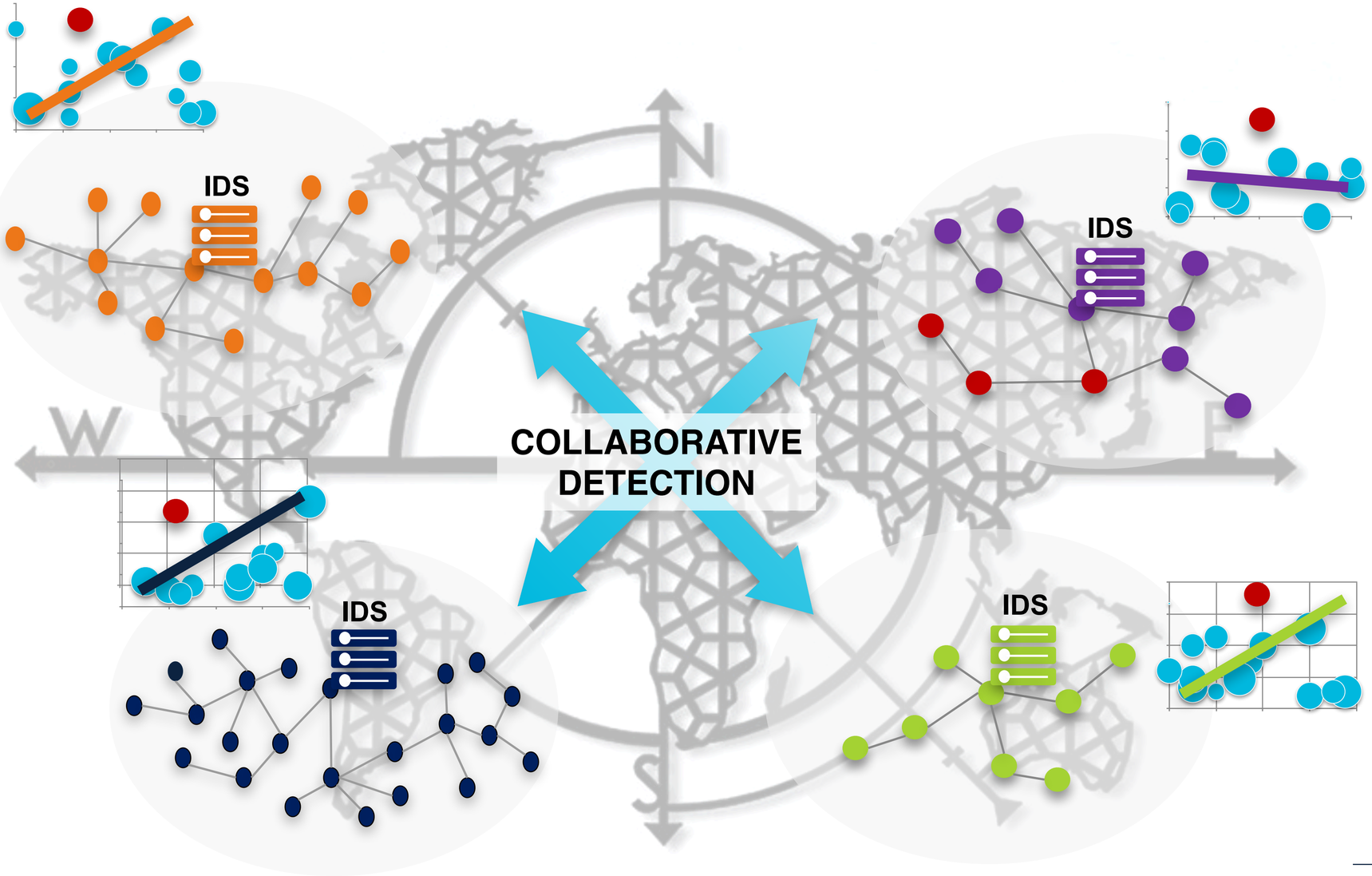
AT MULTIPLE ORGANISATION SCALE

How to share experience in intrusion detection?



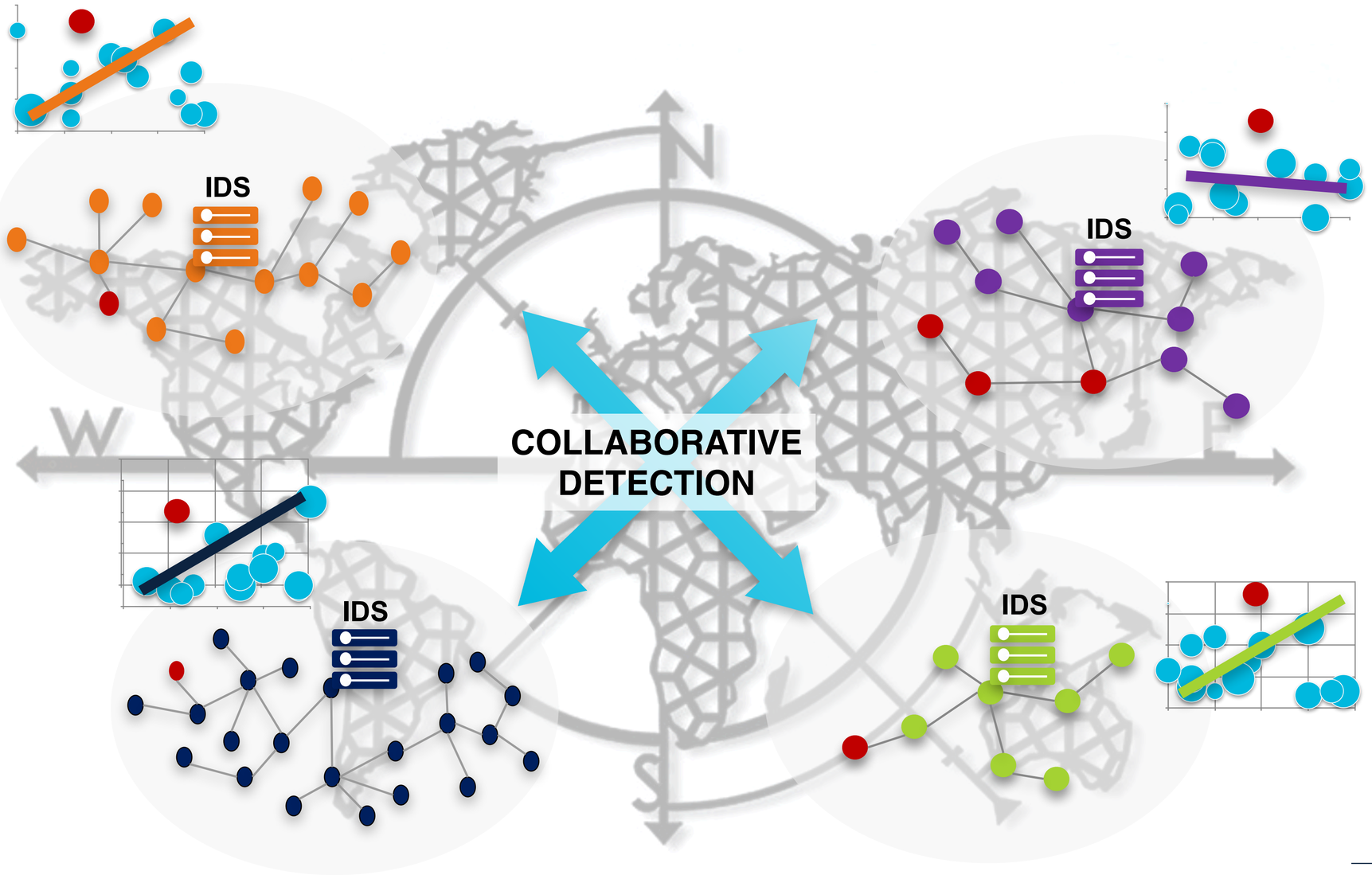
AT MULTIPLE ORGANISATION SCALE

How to share experience in intrusion detection?



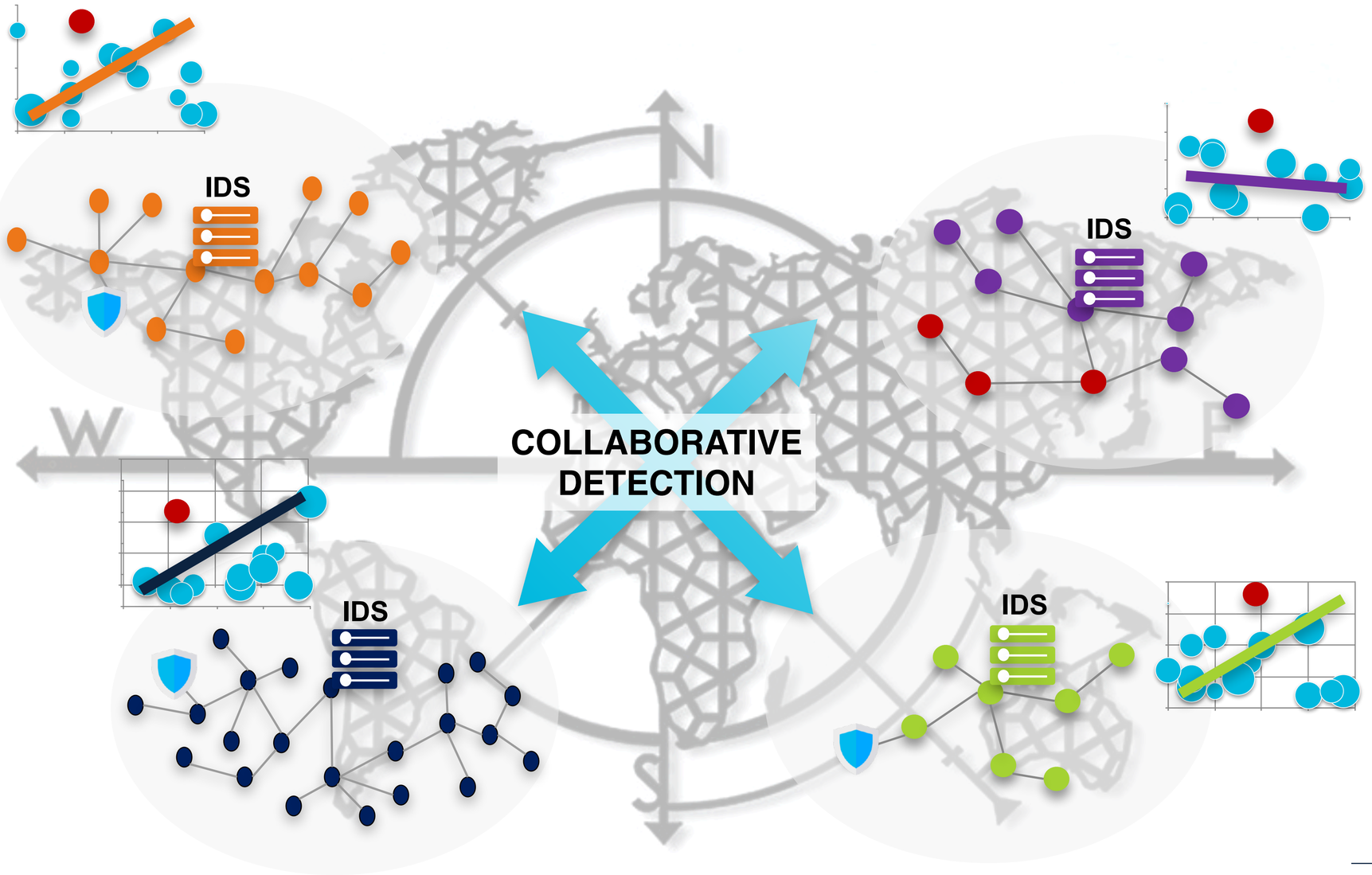
AT MULTIPLE ORGANISATION SCALE

How to share experience in intrusion detection?



AT MULTIPLE ORGANISATION SCALE

How to share experience in intrusion detection?

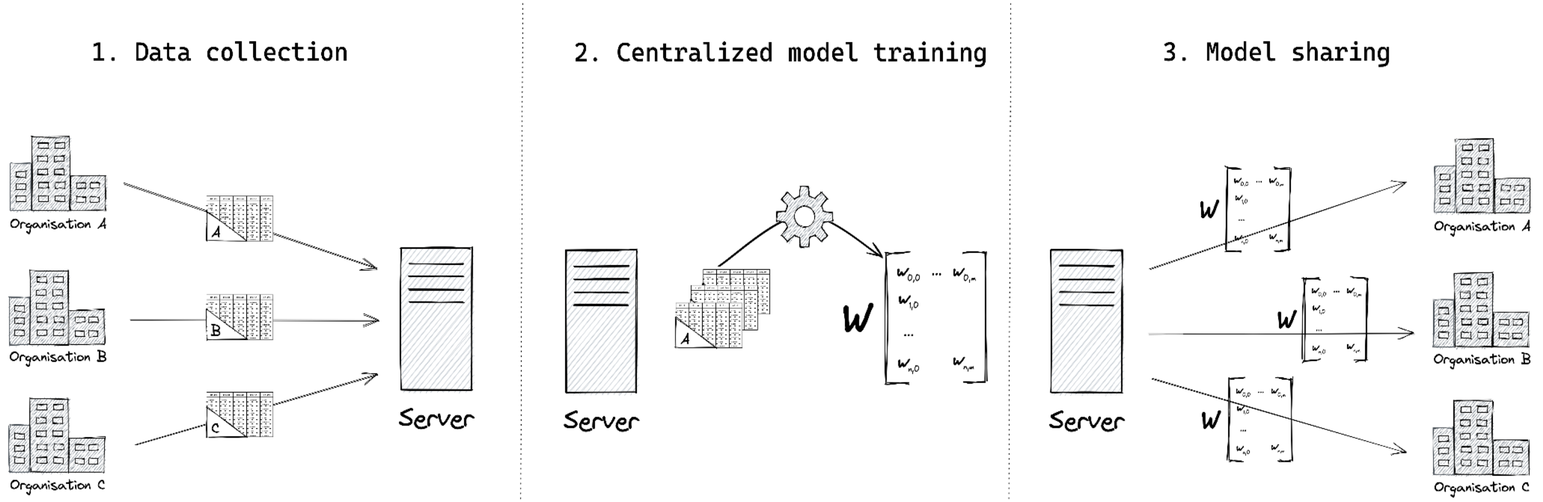


DEFINITION OF FIDS

[1] C. Fung et al. “Trust Management for Host-Based Collaborative Intrusion Detection.” In Managing Large-Scale Service Deployment, 2008.
[2] S. Rathore, et al., “BlockSecIoT-Net: Blockchain-based decentralized security architecture for IoT network,” Journal of Network and Computer Applications, 2019
[3] B. McMahan, et al., “Communication-efficient learning of deep networks from decentralized data”, 20th International conference on artificial intelligence and statistics, 2017

Collaborative Intrusion Detection

- **Objective**
  - Consolidate normal behavior modeling by sharing knowledge with other participants
- **Challenges**
  - Security & Privacy – *e.g.* revealing internals, poisoning, trust [1]
  - Availability – *e.g.* single point of failure in centralized systems [2]
  - Resources – *e.g.* high bandwidth consumption when sharing data [3]

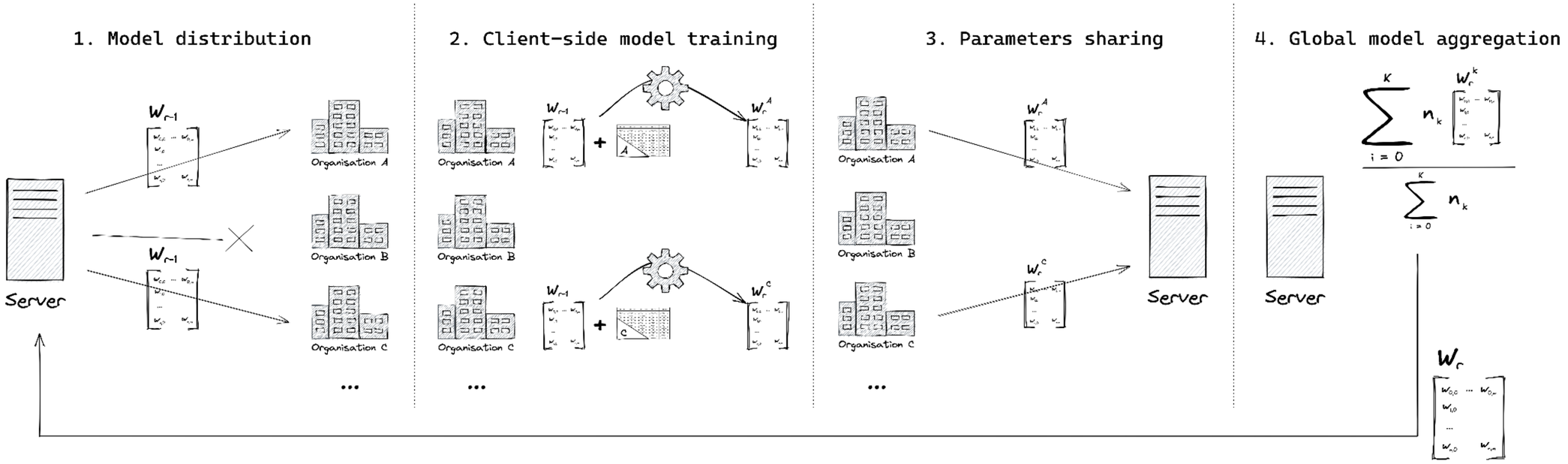


DEFINITION OF FIDS

[4] L. Lavaur, et al., “The Evolution of Federated Learning-Based Intrusion Detection and Mitigation: A Survey”, IEEE Transactions on Network and Service Management, 2022

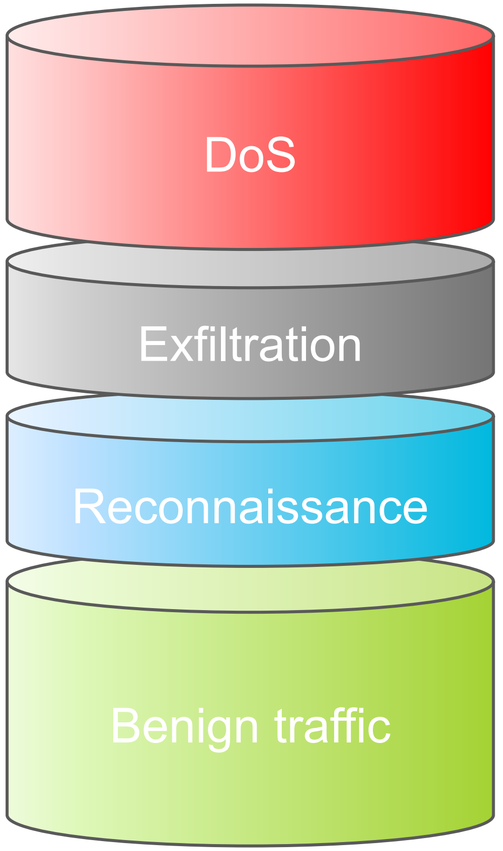
Federated Learning as a Collaborative Learning System

- **Challenges** [4]
  - Heterogeneity – unsuitable global aggregation when participants are too different
  - Trust – assessing peer contributions



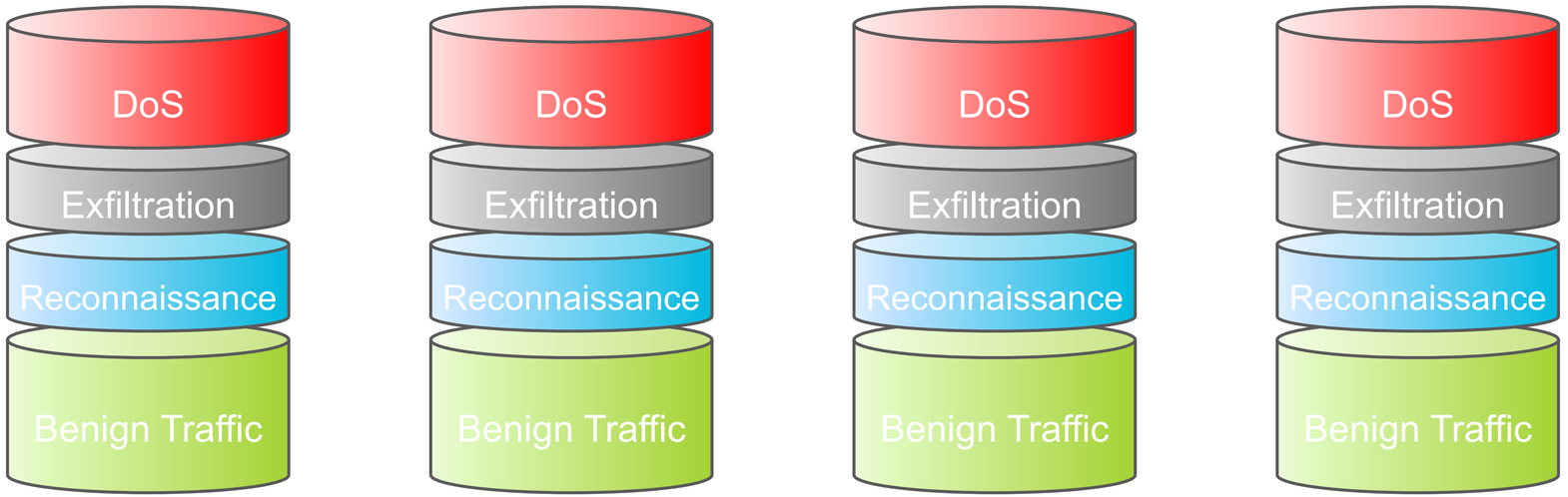
HOW TO DEAL WITH DATA HETEROGENEITY?

- **Classes of attack performed**
  - Denial of service
  - Reconnaissance (port scanning)
  - Data exfiltration, etc.
- **Trained algorithms**
  - Supervised learning on legitimate traffic and attacks
  - Neural networks: Multi-Layer Perceptron (MLP) type



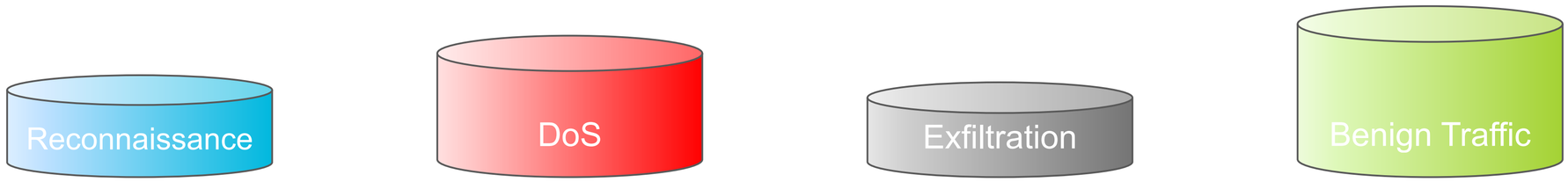
IDEAL CASE: FEDERATED LEARNING ON HOMOGENOUS DATA

- **Homogeneous distribution of the dataset over 4 sites**
  - IID data (Independently and Identically Distributed)
  - No overlap in samples (disjoint data)



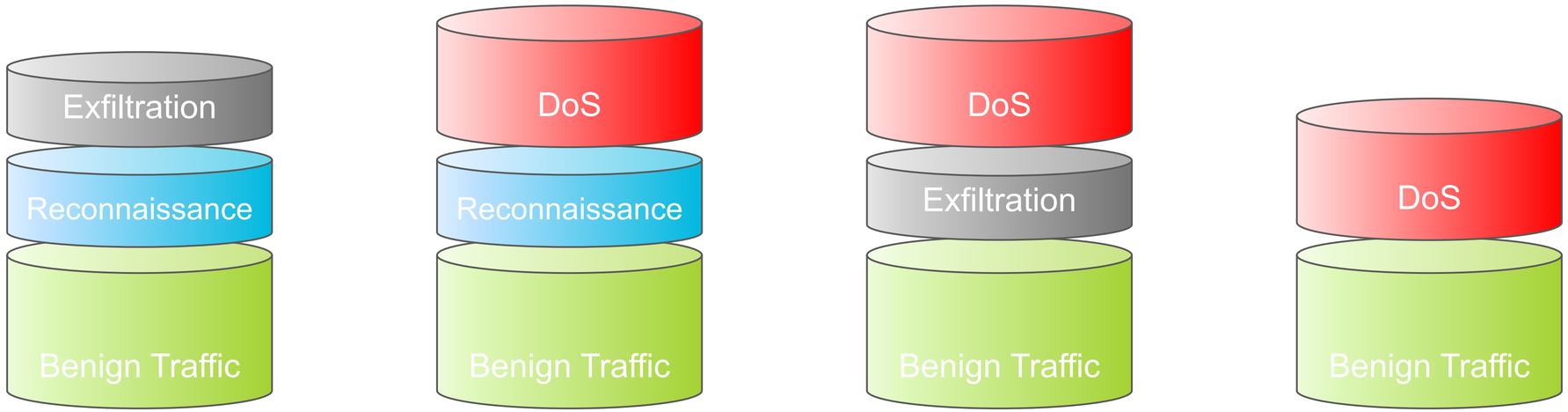
LESS IDEAL CASE: FEDERATED LEARNING ON HETEROGENOUS DATA

- **Differentiated distribution (NIID) of the dataset on 4 sites**
  - The data from the 4 sites do not contain the same attack classes
- **Pathological NIID [8]**
  - Only 1 class per client
  - Only 1 client per class
  - Not realistic in IDS context



REALISTIC CASE: FEDERATED LEARNING ON HETEROGENOUS DATA

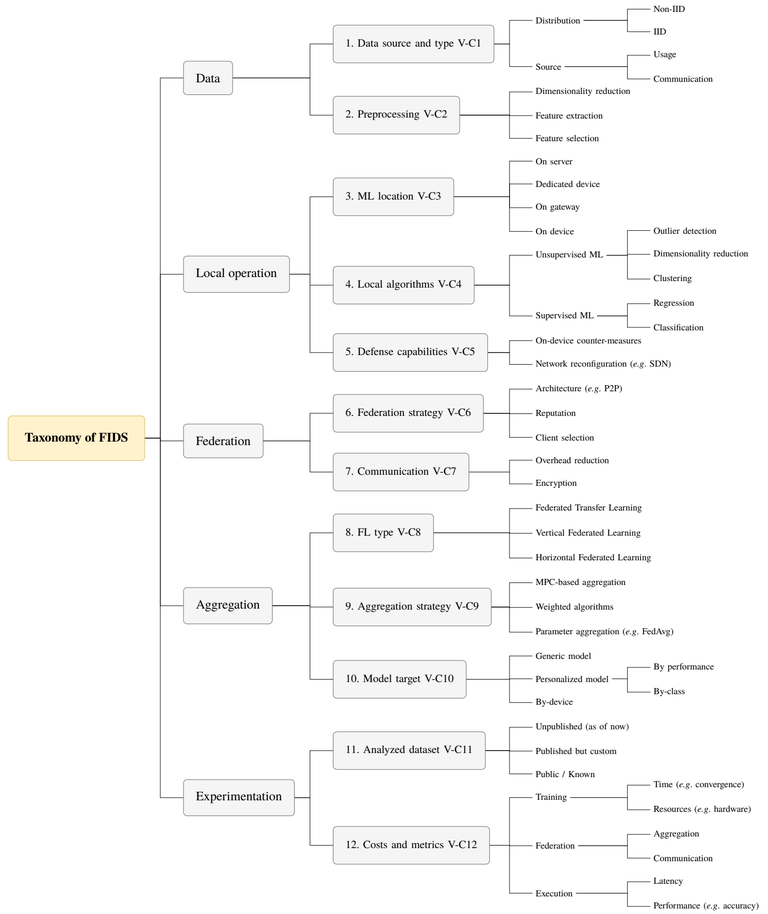
- **Differentiated distribution (NIID) of the dataset on 4 sites**
  - The data from the 4 sites do not contain the same attack classes
- **Practical NIID [8]**
  - Still no overlap in sample
  - Some classes can be shared by different clients, but usually not all



FIRST RECENT CONTRIBUTION ON FL

[4] L. Lavaur, et al., “The Evolution of Federated Learning-Based Intrusion Detection and Mitigation: A Survey”, IEEE Transactions on Network and Service Management, 2022

- **« The Evolution of FL-based intrusion detection and mitigation: a Survey »** [4]
  IEEE Transaction on Network and Services Management, 2022
  - Systematic Literature Review
  - Four contributions
    - Quantitative and qualitative structured analyses
    - Reference architecture
    - Taxonomy
    - Open issues and research directions



- **Research Open Questions answered by the survey**
  - How are FIDSs used in different domains?
  - What are the differences between FIDS architectures?
  - What is the state of the art of FIDSs?

OPEN ISSUES

1. Transferability, adaptability, and scalability [7], [9]-[14]

How to deal with high number of clients and constrained environments? How learn from heterogeneous data, or heterogeneous clients? How to balance generalization and specialization for models?

2. Security, trust, and resilience [9], [10], [14]-[16]

How to resist to poisoning and inference attacks against shared data? How to protect sharing and aggregation (HE, MPC, DP...)? How to deal with untrusted participants? How to mitigate attacks?

3. Algorithm and aggregation performance [5]-[8]

What is the impact of the hyper- and meta-parameters? How to model behaviors to better characterize traffic? How to improve the raw performance of models? What is the best data to train models.

HANDS-ON! — PART 2

FEDERATED LEARNING FOR SECURITY

https://tinyurl.com/FLxNS-part2

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

HOW TO SECURE THE FEDERATED LEARNING IN NETWORK MONITORING?

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

ADVERSARIAL ATTACKS IN FEDERATED LEARNING

Adversarial attacks in federated learning

Attacks to the federated model

- Aim at modifying its behaviour

Privacy attacks

- Aim at inferring sensitive information from learning

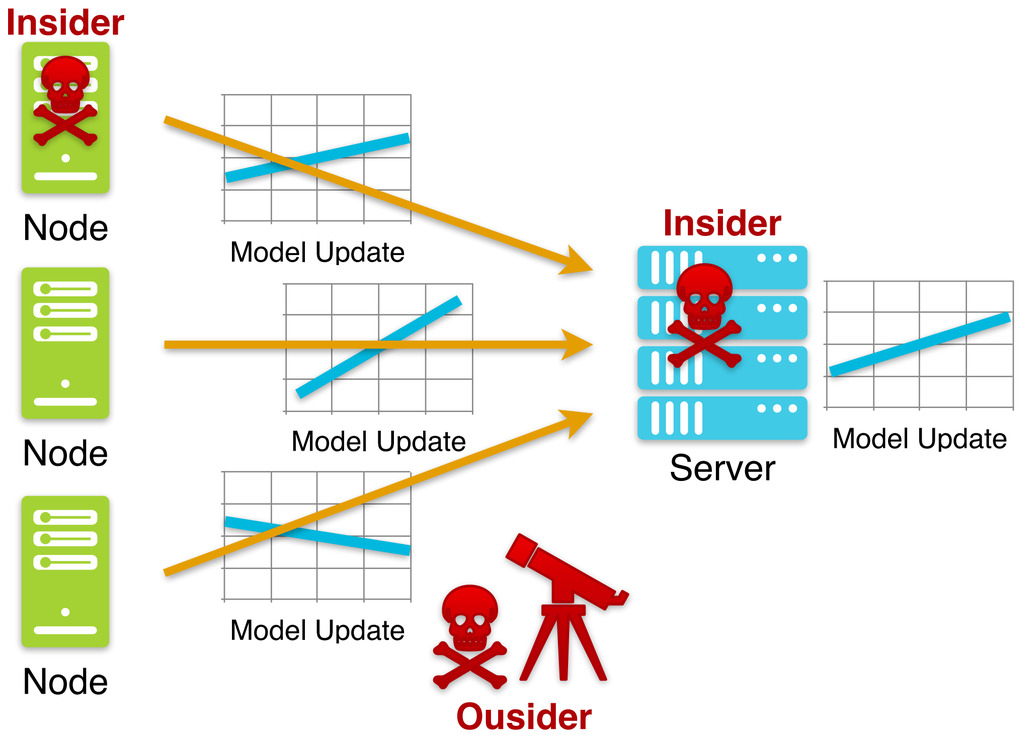
HOW CAN ATTACKERS TAKE ADVANTAGE OF FEDERATED LEARNING?

- **Threat model**
  - Structured representation of information
    - Help to identify and define potential security issues
  - Defined in terms of
    - Information available
    - Scope of action of the attacker

Source: https://www.eccouncil.org/threat-modeling/

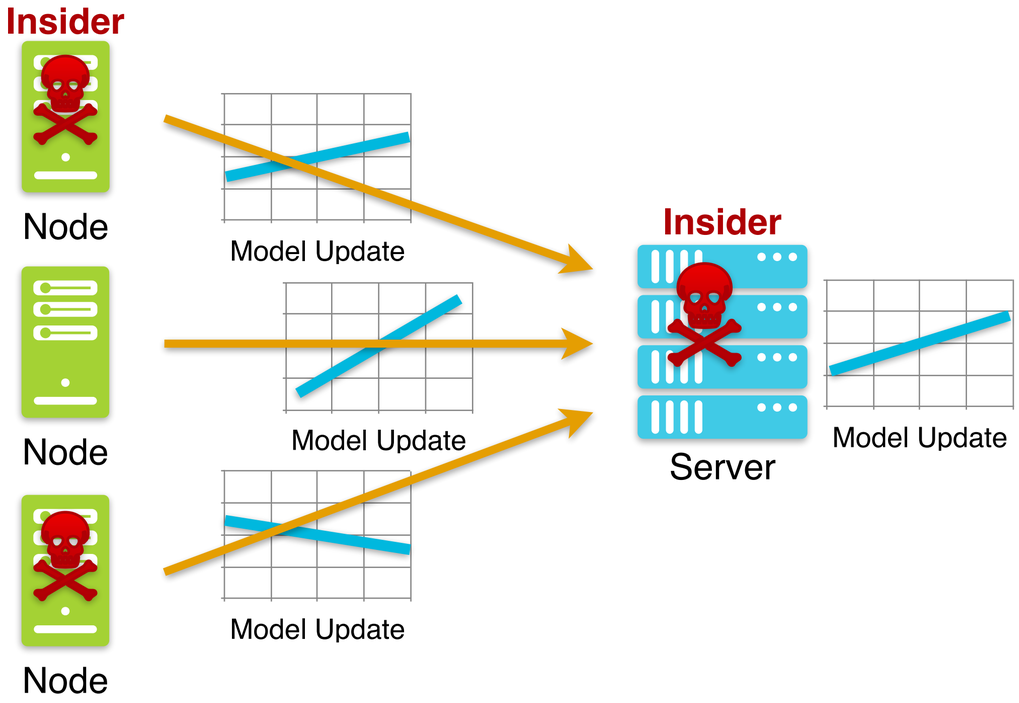
THREAT MODEL: INSIDER *VS.* OUTSIDER

- **Ousider**
  - Mainly focus on sniffing information of the communication channels between the involved agents
  - Aimed at **inferring information** about the data or the resulting learning model
- **Insider**
  - **More harmful**
  - Attack is carried out by one (or coalition) of the participants
  - Aimed at **modifying the behaviour** of the model or **inferring valuable information** from other clients



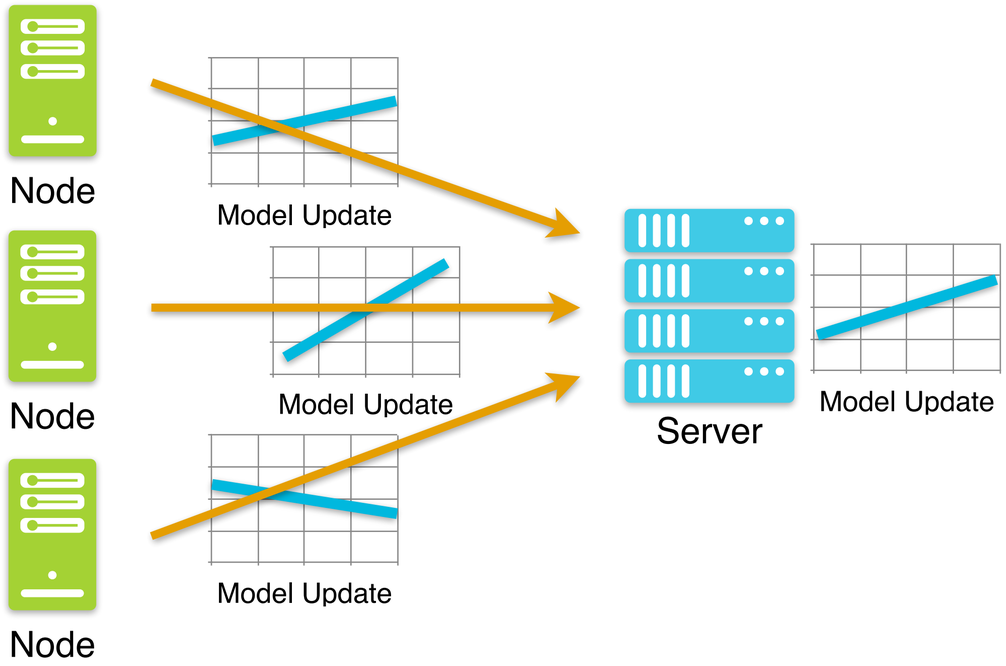
THREAT MODEL: FOCUS ON INSIDER CATEGORIZATION

- **Byzantine attacks**
  - Consist in sending arbitrary updates to the server
  - Aim to compromise the performance of the global learning model.
- **Sybil attacks**
  - Consist of collaborative attacks
    - By several attackers joining together
    - By simulating fictitious clients in order to be more disruptive
- **Honest-but-curious vs. Malicious**



OTHER THREAT MODEL TERMS

- **Client *vs.* Server**
- **Attacker knowledge**
  - Client-side knowledge *(sharing features & labels)*
    - Access to local data of other clients or their labels: Extra client-side knowledge
  - Server-side knowledge
  - Party-side knowledge *(sharing samples only)*
    - Access to information related to the features of the other clients: Extra party-side knowledge
  - Third party-side knowledge
  - Outsider-side knowledge
- **Collusion *vs.* No-collusion**



ADVERSARIAL ATTACKS IN FEDERATED LEARNING

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

ADVERSARIAL ATTACKS TO THE FEDERATED MODEL

- **Clients have the ability to harm the model by sending poisoned updates**
  - The server cannot inspect the training data stored on the clients

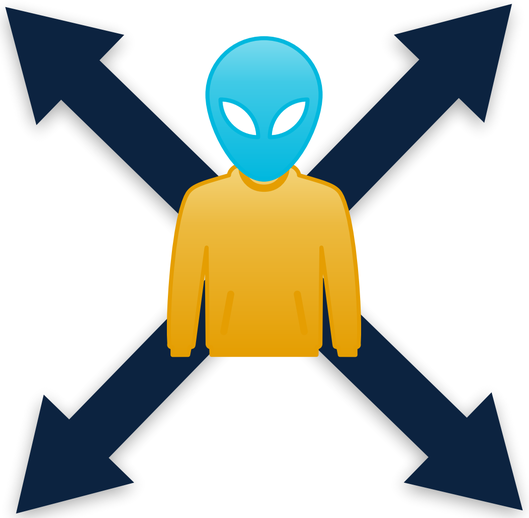
4 TAXONOMIES OF ATTACKS [2]

Attack moment

- Training time attacks
- Inference time attacks

Objective

- Untargeted attacks
- Targeted/backdoor attacks



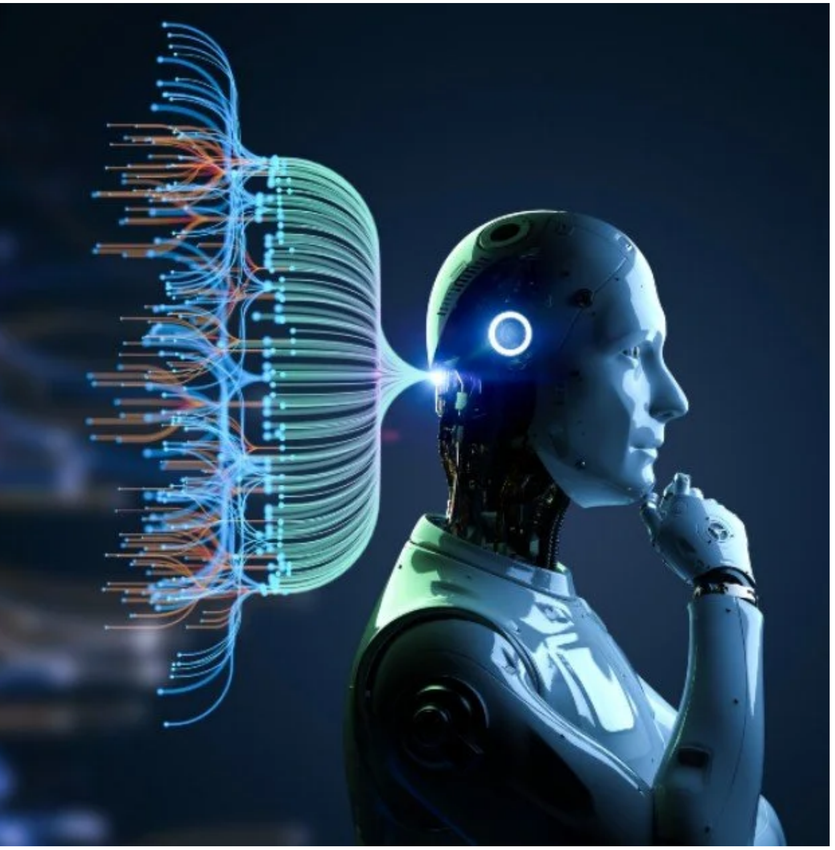
Poisoned part

- Data-poisoning attacks
- Model-poisoning attacks

Frequency

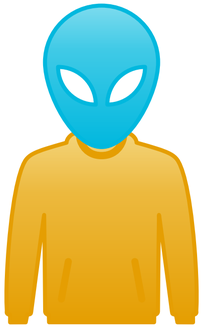
- One-shot attack
- Multiple or adaptive attack

BACKDOOR THE MODEL



- **Inject small amount of malicious data into the benign traffic, which will not be detected as anomalous**
  - The model will not detect the backdoored traffic as malicious
  - Security gateway uses this data to train the local model
  - Local model will be sent to the aggregator, hence affecting the global model
- **Challenges of the implanted backdoor**
  - To evade
    - the traffic anomaly detection of the global model and
    - the model anomaly detection of the aggregator

4 TAXONOMIES OF ATTACKS: FOCUS ON BACKDOOR ATTACKS [2-3]



Objective

- Untargeted attacks
- Targeted/backdoor attacks

Input-instance-key strategies

Pattern-key strategies

Design of the pattern

- Blended injection strategy
- Accessory injection strategy
- Blended accessory injection strategy

Number of patterns

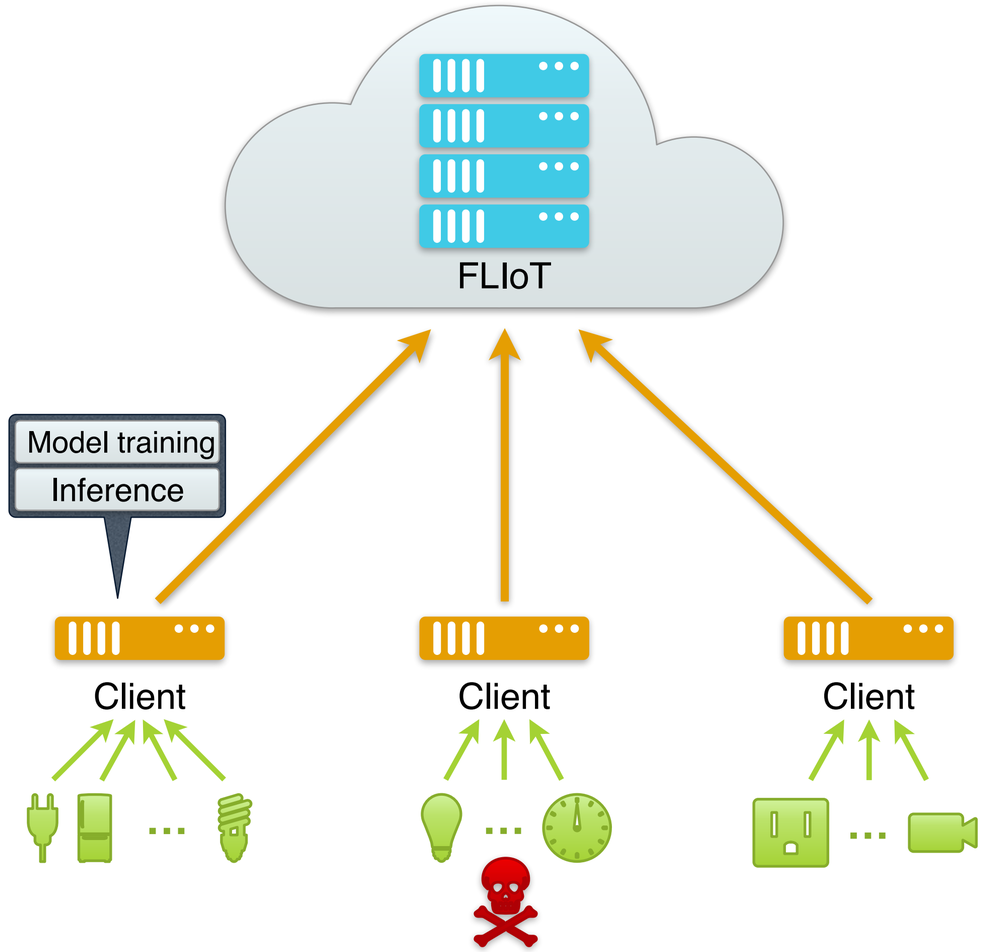
- Single pattern attack
- Multi-backdoor attack (Sybil)

Variability over time

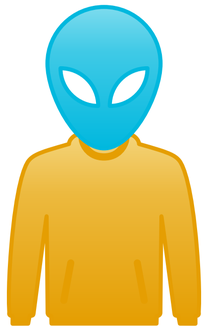
- Static attack
- Dynamic attack

THREAT MODEL FOR POISONING ATTACKS [4]

- **Data poisoning attacks**
  - Implant a backdoor in the aggregated model to incorrectly classify malicious data as benign.
- **Attacker’s goal**
  - To corrupt the global model by aggregator so that the model wouldn't detect malicious traffic as anomalous.
- **The attacker controls a number of IoT devices and can also connect their devices to the security gateway**



4 TAXONOMIES OF ATTACKS: FOCUS ON POISONED PART OF FL SCHEME



Poisoned part

Data-poisoning attacks	Model-poisoning attacks
Label-flipping attack	Random weights generation
Poisoning samples attack	Optimization methods
Out-of-distribution attack	Information leakage

POISONING ATTACKS: ASSOCIATED RESEARCH QUESTIONS

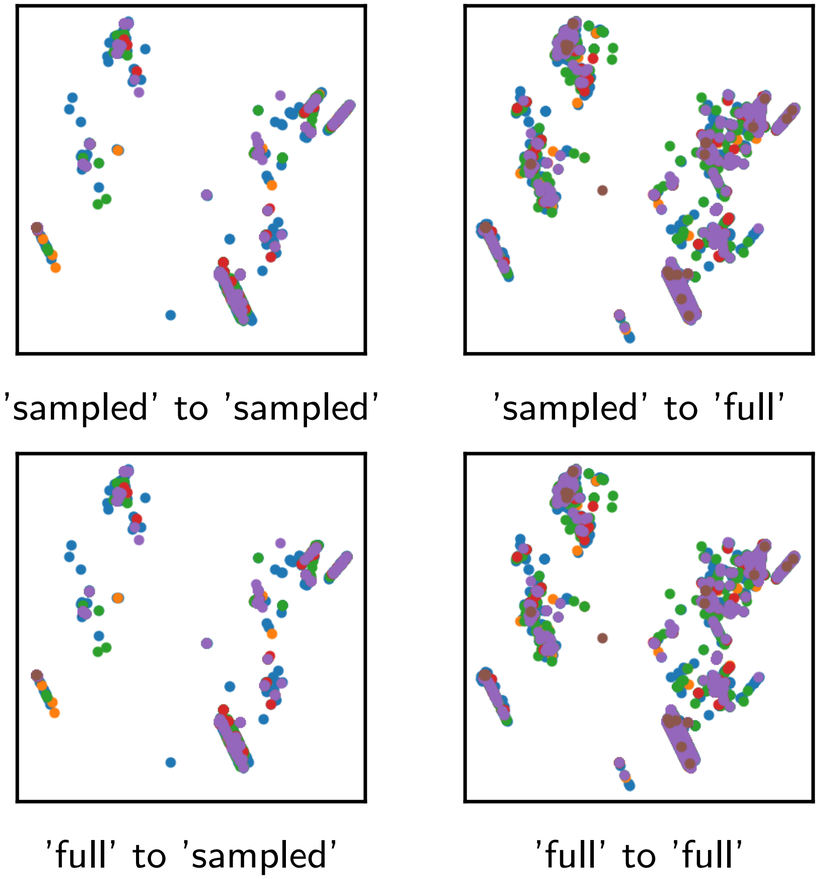
- **RQ1. Is the behavior of poisoning attacks predictable?**
- **RQ2. Do hyperparameters influence the impact of poisoning attacks?**
- **RQ3. Are IDS backdoors realistic using label-flipping attacks?**
- **RQ4. Is there a critical threshold where label-flipping attacks begin to impact performance?**
- **RQ5. Is gradient similarity enough to detect label-flipping attacks?**

Léo Lavaur, Yann Busnel, Fabien Autrel. *Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection*. Computers & Security, Volume 156, 104462, September 2025.

EXPERIMENTAL SETUP

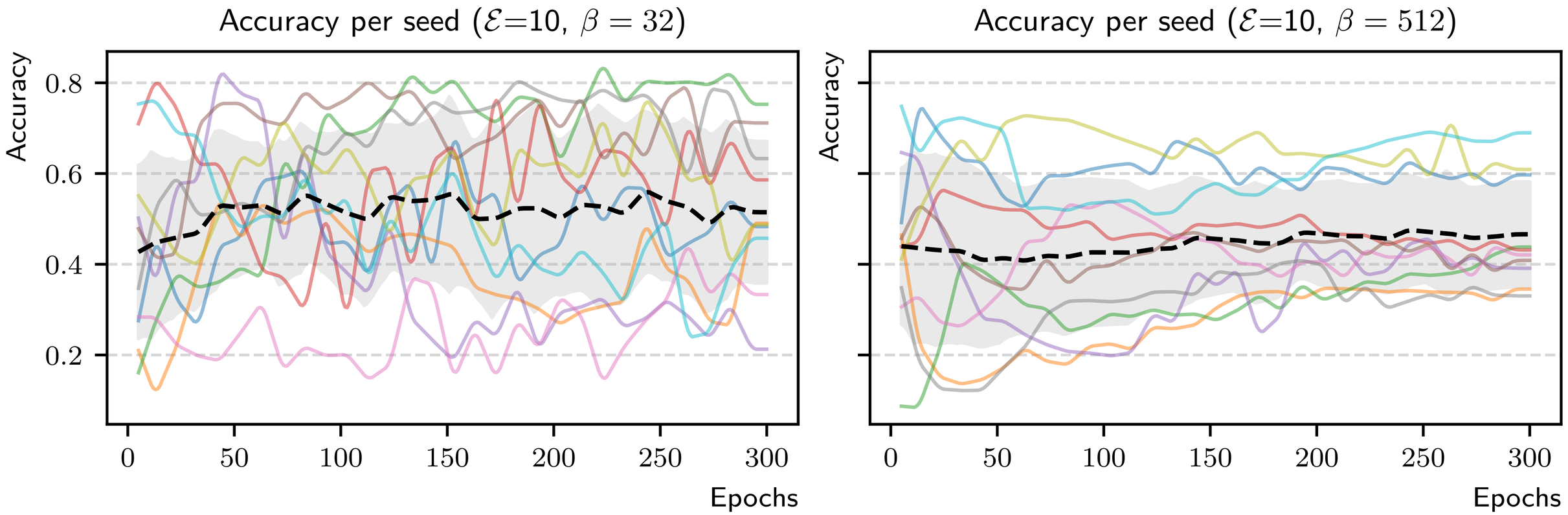
- **Used dataset: sampled NF-V2 version of CSE-CIC-IDS2018**
  - Ports and IP addresses are removed
- **Same class distribution in the training and testing sets**
  - 80% of the dataset is used for training
  - 20% of the dataset is used for testing
- **Assessment of the representativity of the dataset sampling**
  - Cross-projections of the malicious traffic from two datasets in two dimensions using PCA

Cross-projection of the malicious traffic from two datasets in two dimensions using PCA



Léo Lavaur, Yann Busnel, Fabien Autrel. *Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection*. Computers & Security, Volume 156, 104462, September 2025.

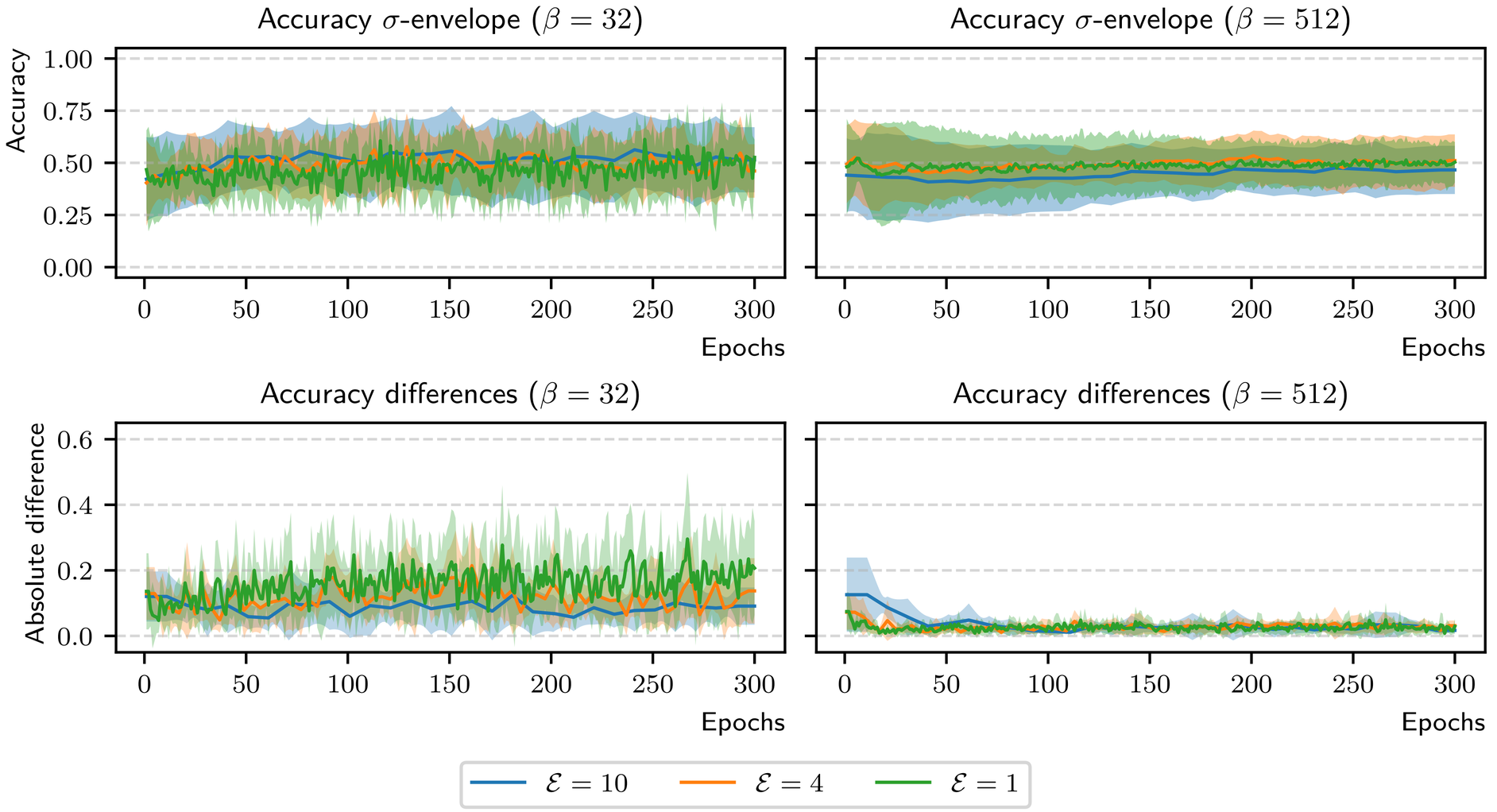
RQ1: IS THE BEHAVIOR OF POISONING ATTACKS PREDICTABLE?



Accuracy of the poisoned model by seed

Léo Lavaur, Yann Busnel, Fabien Autrel. *Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection*. Computers & Security, Volume 156, 104462, September 2025.

RQ1: IS THE BEHAVIOR OF POISONING ATTACKS PREDICTABLE?



Predictability depending on the hyperparameters

Léo Lavaur, Yann Busnel, Fabien Autrel. *Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection*. Computers & Security, Volume 156, 104462, September 2025.

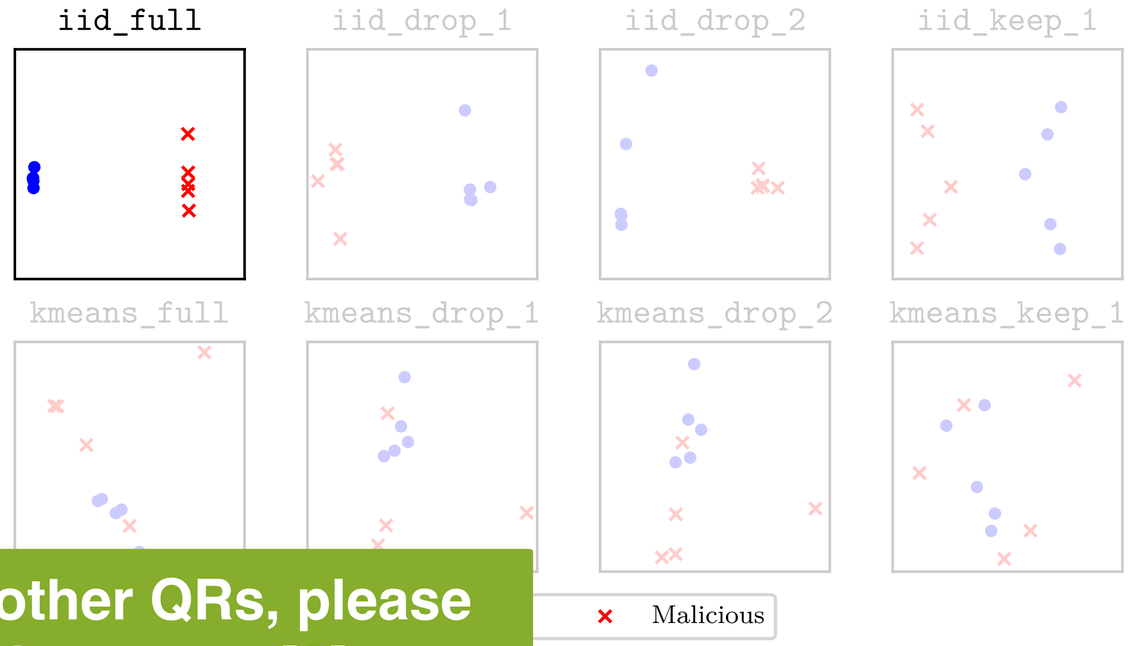
RQ1: IS THE BEHAVIOR OF POISONING ATTACKS PREDICTABLE?

Answer: Nope!

Léo Lavaur, Yann Busnel, Fabien Autrel. *Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection*. Computers & Security, Volume 156, 104462, September 2025.

RQ5. IS GRADIENT SIMILARITY ENOUGH TO DETECT LABEL-FLIPPING ATTACKS?

- **PCA projection of the gradients in 2D (CICIDS).**
  - Known technique to detect poisoning attacks [10]
- **High heterogeneity makes it harder to detect attackers.**



For more details on the other QRs, please consult the companion paper [9]

[9] Léo Lavaur, Yann Busnel, Fabien Autrel. *Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection*. Computers & Security, Volume 156, 104462, September 2025
[10] Tolpegin et al. “Data Poisoning Attacks Against Federated Learning Systems”. Lecture Notes in Computer Science. 2020

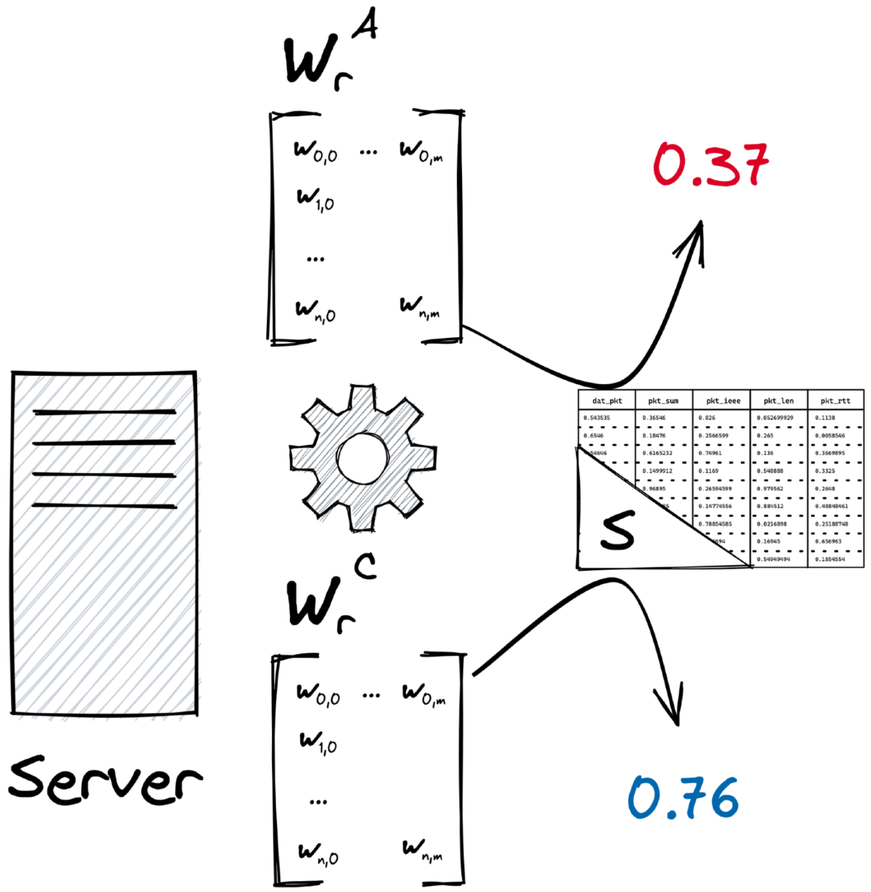
METHODS FOR FILTERING CONTRIBUTIONS IN FEDERATED LEARNING

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

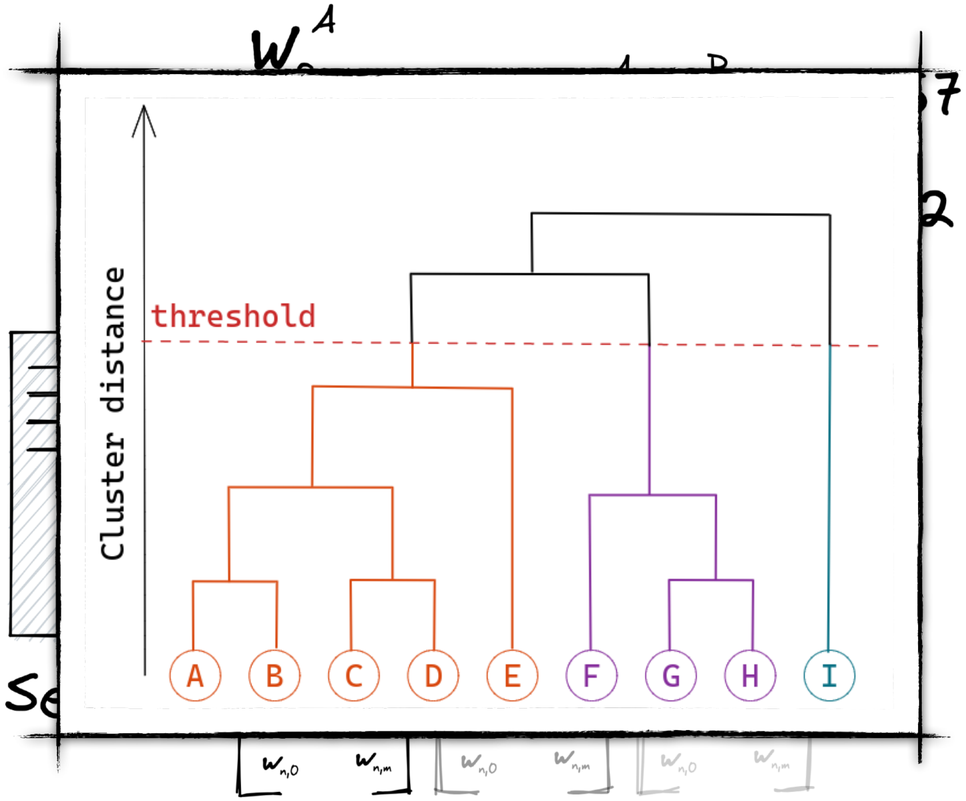
uni.lu | SnT

SERVER-SIDE EVALUATION [5]



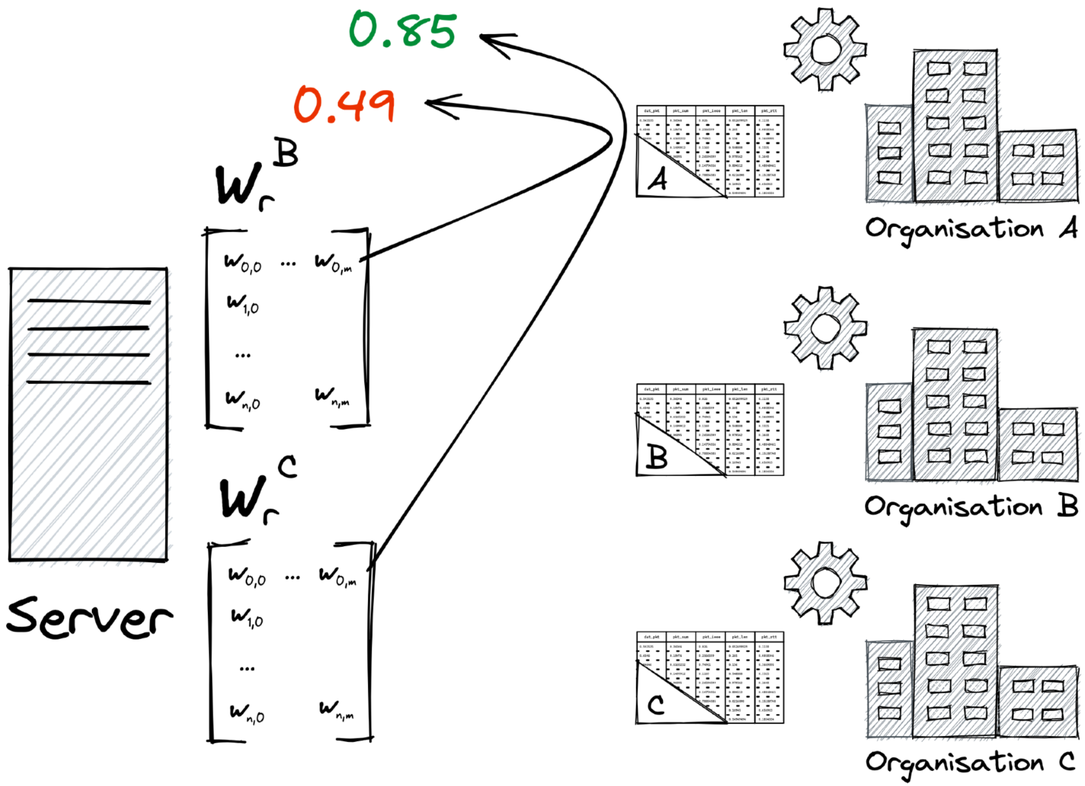
- **Compute test evaluation on updated models**
  - Exclude outlier clients
- **Limitation**
  - Only applicable in IID settings
  - Single source of truth
    - Representative test dataset
    - Server trustworthiness

SERVER-SIDE MODEL COMPARISON



- **Clustering the clients by similarity based on their updates**
  - e.g. Hierarchical clustering [6]
- **Limitation**
  - Less related to client data
  - More appropriated for high-dimensional features

CLIENT-SIDE EVALUATION [7]



- **Cross-evaluation approach**
  - Merge update for « close » clients
  - Exclude outlier clients from the local point of view
- **Limitation**
  - High cost in cross-device settings

A CROSS-EVALUATION APPROACH FOR REPUTATION-AWARE MODEL WEIGHTING

FILTERING CONTRIBUTIONS IN FEDERATED LEARNING FOR INTRUSION DETECTION

Joint work between Yann Busnel (Institut Mines-Télécom)
Leo Lavaur (University of Luxembourg)
Pierre-Marie Lechevalier, Romaric Ludinard,
Marc-Oliver Pahl, Géraldine Texier (IMT Atlantique)

Institut Mines-Télécom

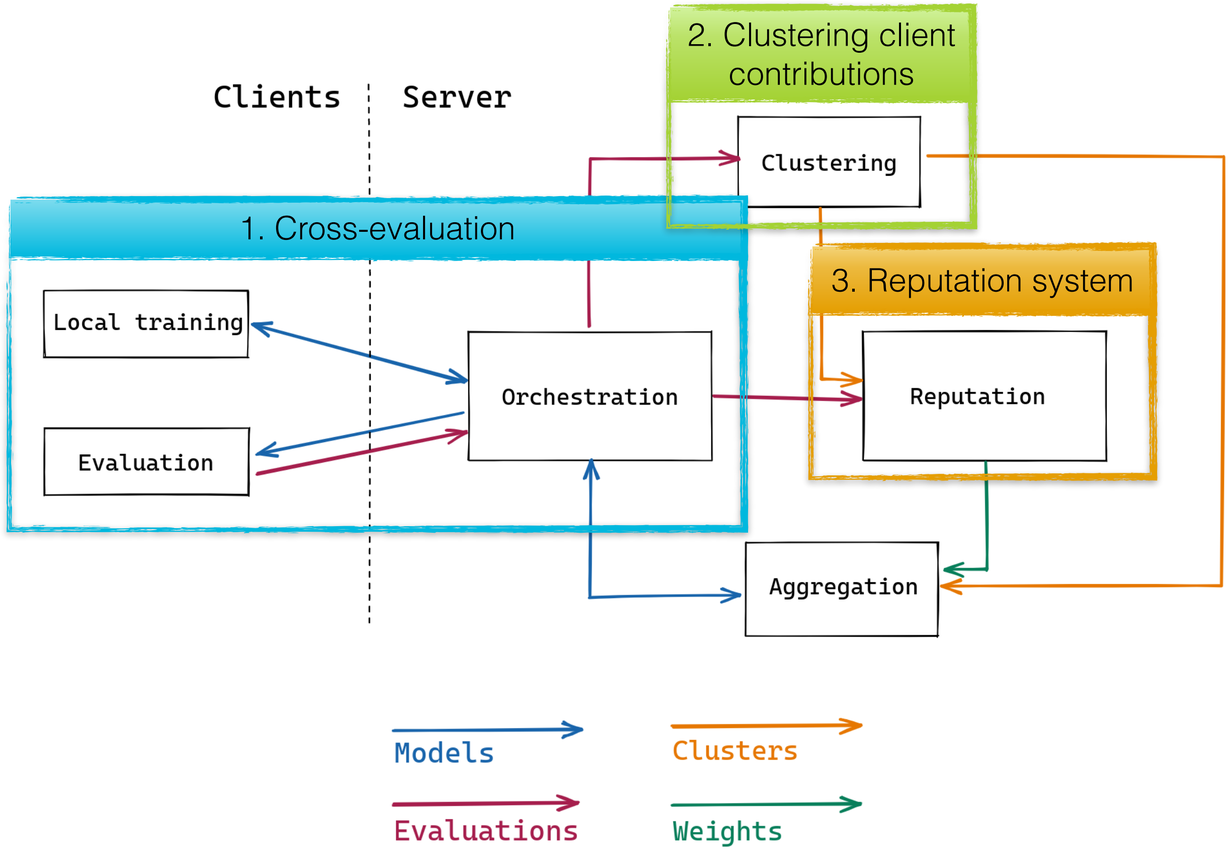
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

OUR APPROACH: RADAR [SRDS'24]

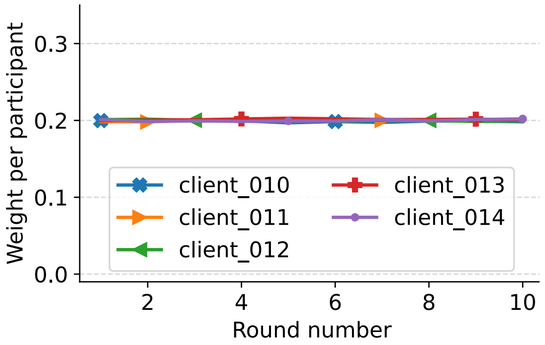
- **Objective:** Mitigate the impact of *bad* contributions to the local models
  - How to evaluate models in highly heterogeneous settings?
  - How to set aside dissimilar participants?
  - How to identify and discard similar but negative behaviors?

Proposed architecture

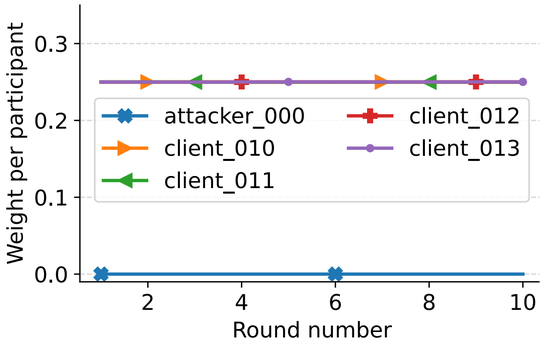


L. Lavaur, P.-M. Lechevalier, Y. Busnel, R. Ludinard, G. Texier, M.-O. Pahl. *RADAR: Model Quality Assessment for Reputation-aware Collaborative Federated Learning*. 43rd International Symposium on Reliable Distributed Systems (SRDS 2024), Charlotte, USA, Sept.2024

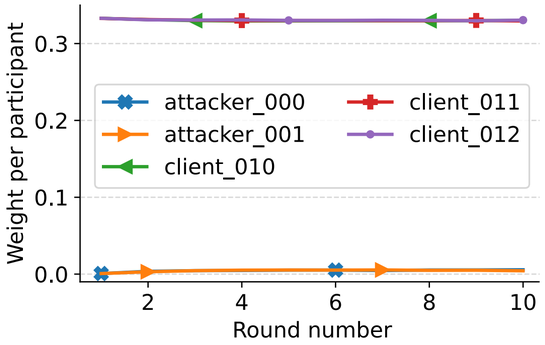
AGGREGATION WEIGHTS ρ FOR THE PARTICIPANTS COMING FROM THE BOT-IOT DATASET DEPENDING ON THE NUMBER OF BYZANTINES



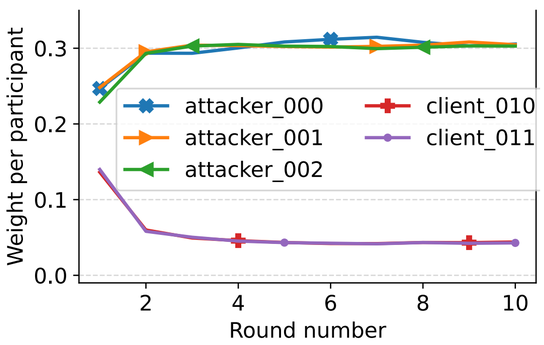
(a) `Benign`.



(b) `Lone 100T`.



(c) `Colluding minority 100T`.

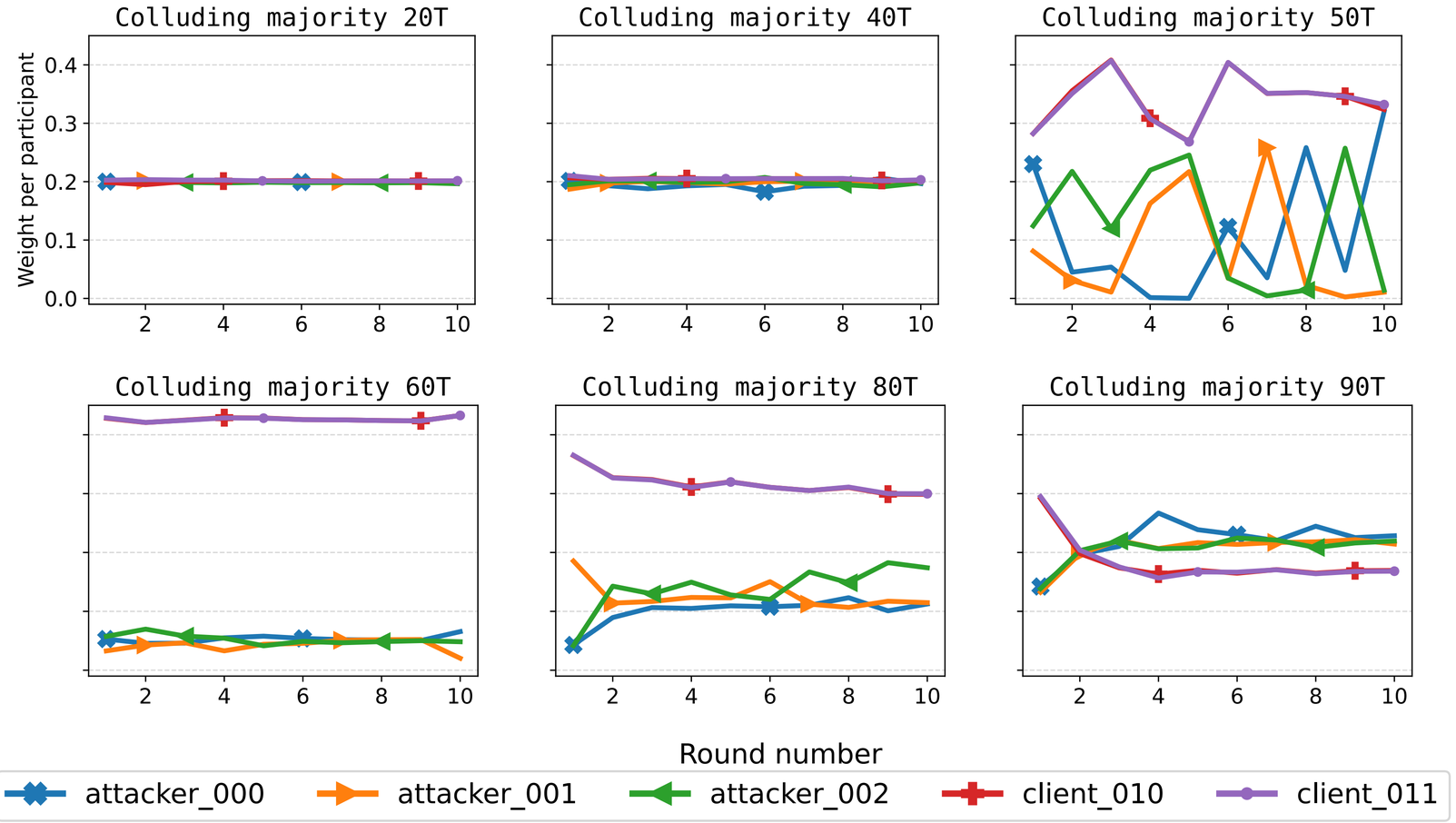


(d) `Colluding majority 100T`.

- **Byzantines are correctly penalized when they are a minority**
  - But gain precedence when they become the majority

L. Lavaur, P.-M. Lechevalier, Y. Busnel, R. Ludinard, G. Texier, M.-O. Pahl. *RADAR: Model Quality Assessment for Reputation-aware Collaborative Federated Learning*. 43rd International Symposium on Reliable Distributed Systems (SRDS 2024), Charlotte, USA, Sept.2024

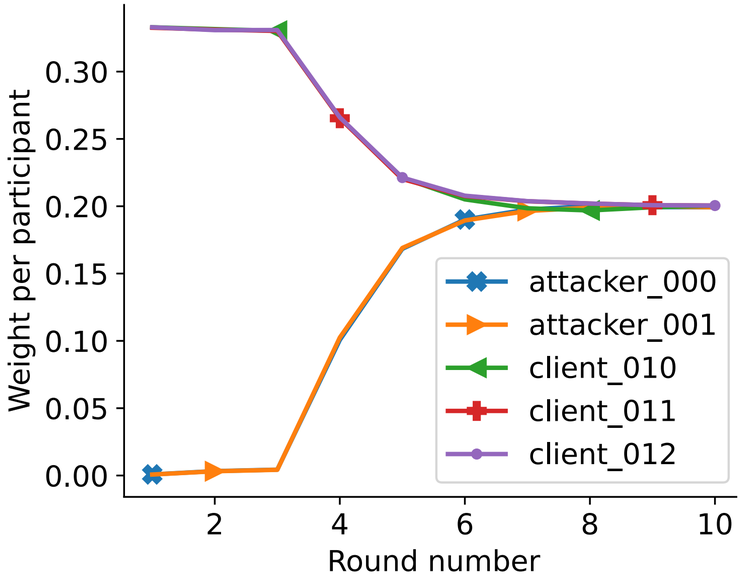
AGGREGATION WEIGHTS ρ PER PARTICIPANT OF THE POISONED CLUSTER



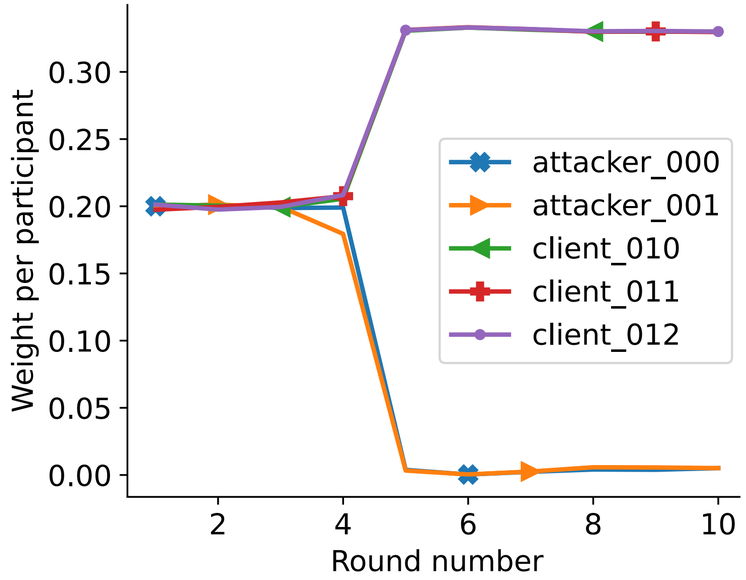
- **Even attackers are a majority, they gain only for higher poisoning rates (≥90%)**

L. Lavaur, P.-M. Lechevalier, Y. Busnel, R. Ludinard, G. Texier, M.-O. Pahl. *RADAR: Model Quality Assessment for Reputation-aware Collaborative Federated Learning*. 43rd International Symposium on Reliable Distributed Systems (SRDS 2024), Charlotte, USA, Sept.2024

AGGREGATION WEIGHTS ρ PER PARTICIPANT OF THE POISONED CLUSTER



(a) Attackers act with 100% *noisiness*, but become benign on round 3.



(b) Attackers start benign, and increase *noisiness* by 20% each round when $r \geq 3$.

- **Attackers are forgiven over time**
- **Reputation system reacts quickly to newly detected attackers.**

L. Lavaur, P.-M. Lechevalier, Y. Busnel, R. Ludinard, G. Texier, M.-O. Pahl. *RADAR: Model Quality Assessment for Reputation-aware Collaborative Federated Learning*. 43rd International Symposium on Reliable Distributed Systems (SRDS 2024), Charlotte, USA, Sept.2024

HANDS-ON! — PART 3

HOW TO SECURE FEDERATED LEARNING

https://tinyurl.com/FLxNS-part3

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

uni.lu | SnT

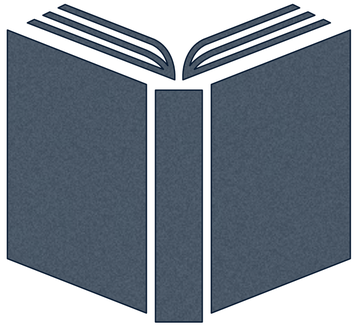
REFERENCES

Institut Mines-Télécom

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

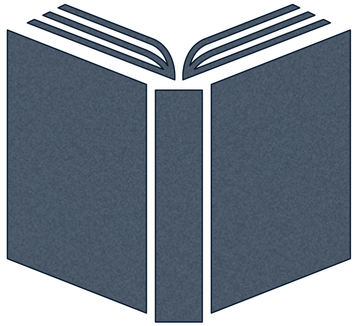
uni.lu | SnT

FL REFERENCES



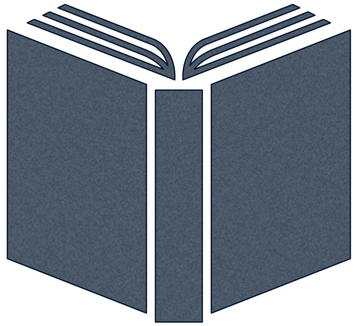
- [1] Thalles Silva. An Introduction to Federated Learning, Encore, 2021.
- [2] Chris J Wallace. Federated Learning, Cloudera Fast Forward Labs, Cloudera, 2019.
- [3] Avi Gopani. Distributed Machine Learning Vs Federated Learning: Which Is Better? Endless Origins, 2021.
- [4] Constantin Philippenko. Federated learning: the privacy-friendly artificial intelligence? Telecom Paris, 2021.
- [5] Aurélien Bellet, Introduction to Federated Learning, Inria, 2020.
- [6] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, H. Vincent Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization, 34th Conference on Neural Information Processing Systems (NeurIPS), 2020.
- [7] Min Du. Federated Learning, UC Berkeley, 2019.

APT REFERENCES



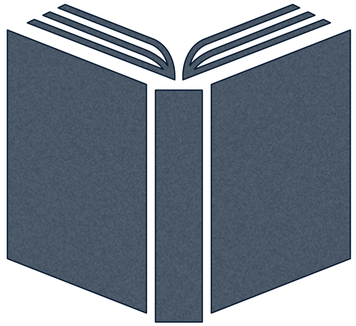
- [1] XOSANAVONGSA, Charles. Heterogeneous Event Causal Dependency Definition for the Detection and Explanation of Multi-Step Attacks. 2020. Thèse de doctorat. CentraleSupélec.
- [2] MILAJERDI, Sadegh M., GJOMEMO, Rigel, ESHETE, Birhanu, et al. Holmes: real-time apt detection through correlation of suspicious information flows. In : 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019. p. 1137-1152.
- [3] INGALE, Sanjana, PARAYE, Milind, et AMBAWADE, Dayanand. A survey on methodologies for multi-step attack prediction. In : 2020 Fourth International Conference on Inventive Systems and Control (ICISC). IEEE, 2020. p. 37-45.
- [4] PEI, Kexin, GU, Zhongshu, SALTAFORMAGGIO, Brendan, et al. Hercule: Attack story reconstruction via community discovery on correlated log graph. In : Proceedings of the 32Nd Annual Conference on Computer Security Applications. 2016. p. 583-595.
- [5] REN, Hanli, STAKHANOVA, Natalia, et GHORBANI, Ali A. An online adaptive approach to alert correlation. In : International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment. Springer, Berlin, Heidelberg, 2010. p. 153-172.
- [6] LANOE, David, HURFIN, Michel, TOTEL, Eric, et al. An Efficient and Scalable Intrusion Detection System on Logs of Distributed Applications. In : IFIP International Conference on ICT Systems Security and Privacy Protection. Springer, Cham, 2019. p. 49-63.
- [7] ALSAHEEL, Abdulellah, NAN, Yuhong, MA, Shiqing, et al. {ATLAS}: A sequence-based learning approach for attack investigation. In : 30th USENIX Security Symposium (USENIX Security 21). 2021. p. 3005-3022.

FL&IDS REFERENCES



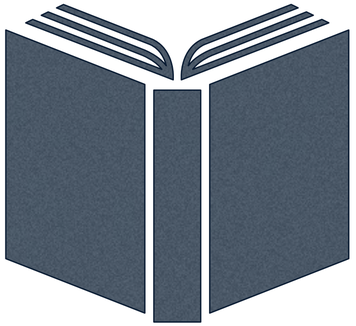
- [1] C. Fung et al. “Trust Management for Host-Based Collaborative Intrusion Detection.” In Managing Large-Scale Service Deployment, 2008.
- [2] S. Rathore, et al., “BlockSecIoT-Net: Blockchain-based decentralized security architecture for IoT network,” Journal of Network and Computer Applications, 2019
- [3] B. McMahan, et al., “Communication-efficient learning of deep networks from decentralized data”, 20th International conference on artificial intelligence and statistics, 2017
- [4] L. Lavaur, M.-O. Pahl, Y. Busnel, and F. Autrel, “The Evolution of Federated Learning-based Intrusion Detection and Mitigation: a Survey,” IEEE Trans. On Network and Services Management, Special Issue on Advances in Network Security Management, 2022
- [5] W. Schneble and G. Thamilarasu, “Attack detection using federated learning in medical cyber-physical systems,” International Conference on Computer Communications and Networks, 2019.
- [6] Y. Sun, H. Ochiai, and H. Esaki, “Intrusion Detection with Segmented Federated Learning for Large-Scale Multiple LANs,” 2020 International Joint Conference on Neural Networks (IJCNN), 2020
- [7] M.-O. Pahl and F. X. Aubet, “All Eyes on You: Distributed Multi-Dimensional IoT Microservice Anomaly Detection,” 14th International Conference on Network and Service Management, 2018
- [8] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, “DÏoT: A Federated Self-learning Anomaly Detection System for IoT,” IEEE 39th International Conference on Distributed Computing Systems (ICDCS), 2019
- [9] S. Rathore, B. Wook Kwon, and J. H. Park, “BlockSecIoT-Net: Blockchain-based decentralized security architecture for IoT network,” Journal of Network and Computer Applications, 2019

FL&IDS REFERENCES



- [10] Y. Fan, Y. Li, M. Zhan, H. Cui, and Y. Zhang, “IoTDefender: A Federated Transfer Learning Intrusion Detection Framework for 5G IoT,” in 2020 IEEE 14th International Conference on Big Data Science and Engineering (BigDataSE), 2020
- [11] S. A. Rahman, H. Tout, C. Talhi, and A. Mourad, “Internet of Things Intrusion Detection: Centralized, On-Device, or Federated Learning?” IEEE Network, 2020
- [12] B. Li, Y. Wu, J. Song, R. Lu, T. Li, and L. Zhao, “DeepFed: Federated Deep Learning for Intrusion Detection in Industrial Cyber-Physical Systems,” IEEE Transactions on Industrial Informatics, 2020
- [13] Y. Chen, J. Zhang, and C. K. Yeo, “Network Anomaly Detection Using Federated Deep Autoencoding Gaussian Mixture Model,” in Machine Learning for Networking, 2020
- [14] T. D. Nguyen, P. Rieger, H. Yalame, H. Mo¨llering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, A.-R. Sadeghi, T. Schneider, and S. Zeitouni. “FLGUARD: Secure and Private Federated Learning.” , arXiv, 2021
- [15] W. Zhang, Q. Lu, Q. Yu, Z. Li, Y. Liu, S. K. Lo, S. Chen, X. Xu, and L. Zhu, “Blockchain-based Federated Learning for Device Failure Detection in Industrial IoT,” IEEE Internet of Things Journal, 2020
- [16] Z. Chen, N. Lv, P. Liu, Y. Fang, K. Chen, and W. Pan, “Intrusion Detection for Wireless Edge Networks Based on Federated Learning,” IEEE Access, 2020
- [17] M. Sarhan, S. Layeghy, and M. Portmann, *Towards a Standard Feature Set for Network Intrusion Detection System Datasets,* arXiv.org, 2021
- [18] G. Bertoli, L. A. Pereira Junior, A. L. dos Santos, O. Saotome, "Generalizing intrusion detection for heterogeneous networks: A stacked-unsupervised federated learning approach," arXiv.org, 2022

POISONING REFERENCES



- [1] Nguyen, T.D., Rieger, P., Miettinen, M. and Sadeghi, A.R. « Poisoning attacks on federated learning-based IoT intrusion detection system ». In Proc. WS Decentralized IoT Syst. Secur. (DISS) (pp. 1-7), 2020
- [2] N. Rodríguez-Barroso, D. Jiménez-López, M. V. Luzón, F. Herrera, E. Martínez-Cámara, “Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges” Elsevier Information Fusion, 2022
- [3] X. Chen, C. Liu, B. Li, K. Lu, D. Song, « Targeted backdoor attacks on deep learning systems using data poisoning », 2017, CoRR abs/1712.05526.
- [4] Y. Gao, B. Gia Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, H. Kim. « Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review », 2020, CoRR, abs/2007.10760
- [5] J. Zhou, et al., “A Differentially Private Federated Learning Model against Poisoning Attacks in Edge Computing”, 2022
- [6] C. Briggs, et al., “Federated learning with hierarchical clustering of local updates to improve training on non-IID data”, 2020
- [7] L. Zhao, et al., ”Shielding Collaborative Learning: Mitigating Poisoning Attacks through Client-Side Detection”, 2020
- [8] Y. Huang et al., “Personalized Cross-Silo Federated Learning on Non-IID Data,” AAAI, vol. 35, no. 9, pp. 7865–7873, May 2021, doi: 10.1609/aaai.v35i9.16960.
- [9] Léo Lavaur, Yann Busnel, Fabien Autrel. Investigating the impact of label-flipping attacks against federated learning for collaborative intrusion detection. Computers & Security, Volume 156, 104462, September 2025
- [10] Tolpegin et al. “Data Poisoning Attacks Against Federated Learning Systems”. Lecture Notes in Computer Science. 2020

Thanks for your attention

Federated Learning and Network Security: Foundations, Potential, and Resilience

yann.busnel@imt.fr
leo.lavaur@uni.lu

July 20rd, 2025 — IEEE ICDCS — Glasgow, UK