Project 1: Scalable and Sustainable Methodologies for ML Driven Security Operations, Energy-aware scheduling for secure model training at the edge.

This project aims to *reducing the carbon footprint* of securely training ML models, in a distributed way, across large edge networks. Specifically, using *fully distributed learning* as a target method (because of its enhanced security and scalability features), the goals are to *i) carefully select* the clients that participate in the current update round (based on security *and* energy metrics), *ii)* schedule *when* and *what* to broadcast to the other nodes (workers), with the objective of reducing the energy impact of the training procedure, subjected to model accuracy guarantees. Energy sustainability and scalability will thus be main considerations to drive learning and security functions, considering, for example, energy availability at the workers (server nodes) and their energy efficiency.

Required background: Basic networking knowledge, programming skills.

Motivation:
Distributed learning is an approach to training machine learning models across multiple computational resources, such as servers, GPUs, or even edge devices. This method is particularly valuable in scenarios where computational resources are limited, data is large-scale, or privacy concerns necessitate keeping data localized. Reasons why distributed learning is advantageous include:

1. ***Privacy and Security***

a) **Federated Learning**: Distributed learning enables federated learning, where the model is trained across decentralized data sources without moving the data. This is crucial in scenarios where data privacy is a concern, such as healthcare or finance, allowing institutions to collaborate on model development without sharing sensitive data.

b) **Data Locality**: By keeping data localized and only sharing model updates, distributed learning minimizes the risk of data breaches and ensures compliance with data protection regulations.

2. ***Resource Constraints***

a) **Edge Computing**: In scenarios where computational resources are distributed (e.g., edge devices in IoT networks), distributed learning allows these devices to contribute to model training without relying on a centralized server. This can reduce latency and bandwidth usage by processing data closer to its source.

b) **Cost Efficiency**: Distributed learning can be more cost-effective by utilizing a network of less powerful devices rather than relying on a single, costly server.

3. ***Energy Efficiency***

Load Balancing: Distributed learning can be designed to optimize energy consumption by balancing computational loads according to the energy efficiency of different nodes. This can be crucial in energy-constrained environments or when aiming to reduce the environmental impact of large-scale computations.

In summary, distributed learning is essential for addressing the challenges posed by the increasing scale and complexity of data and models, enabling faster, more efficient, and secure machine learning processes.