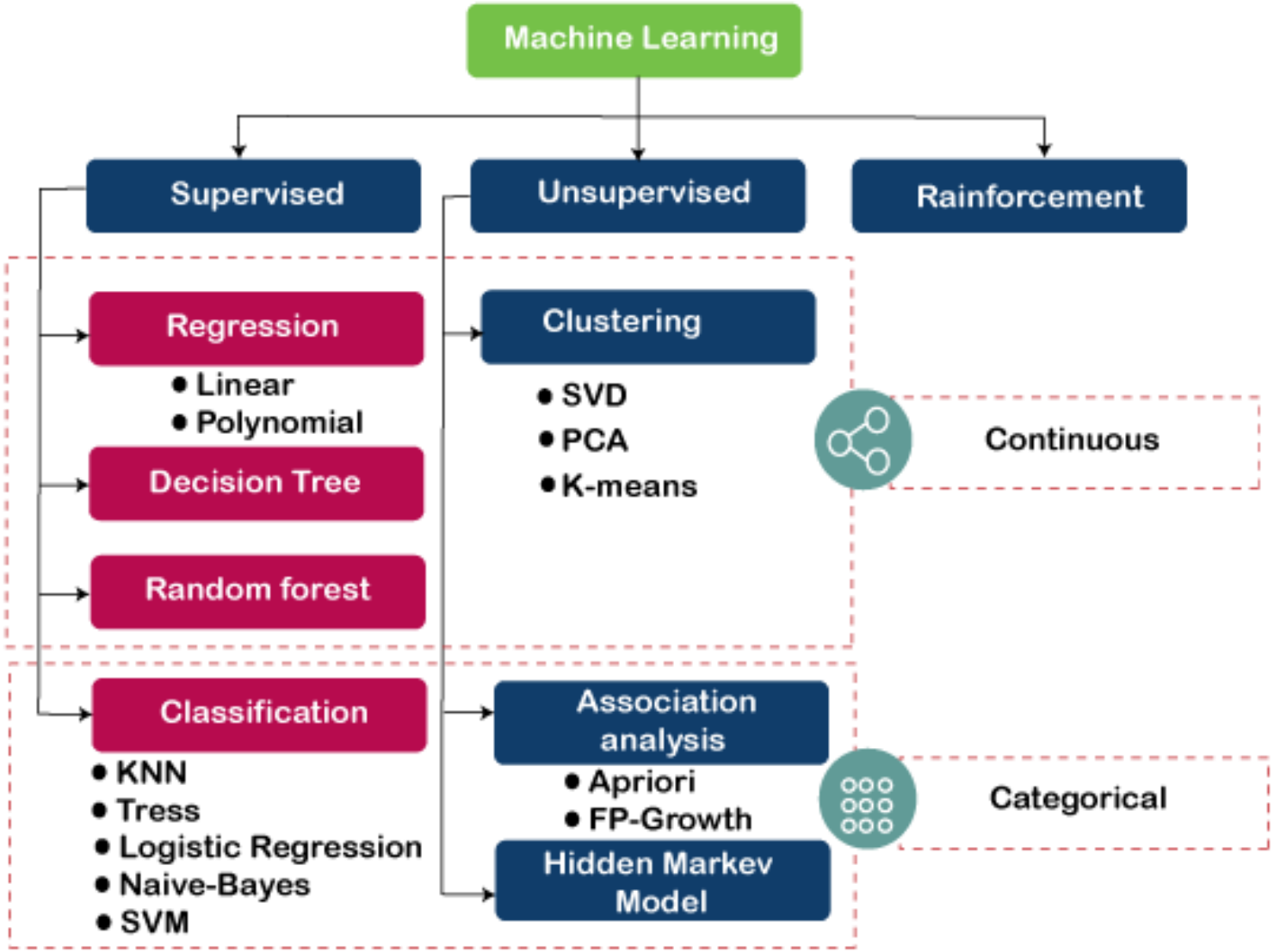
Industrial Internet of Things

Federated Learning for IIoT

Dr Abhishek Hazra

abhishek.h@iiits.in

Machine Learning for IIoT



Source: Internet

Supervised vs Unsupervised vs Reinforcement Learning

Supervised Learning :

1. Try to predict specific quantity
2. Have training example with labels

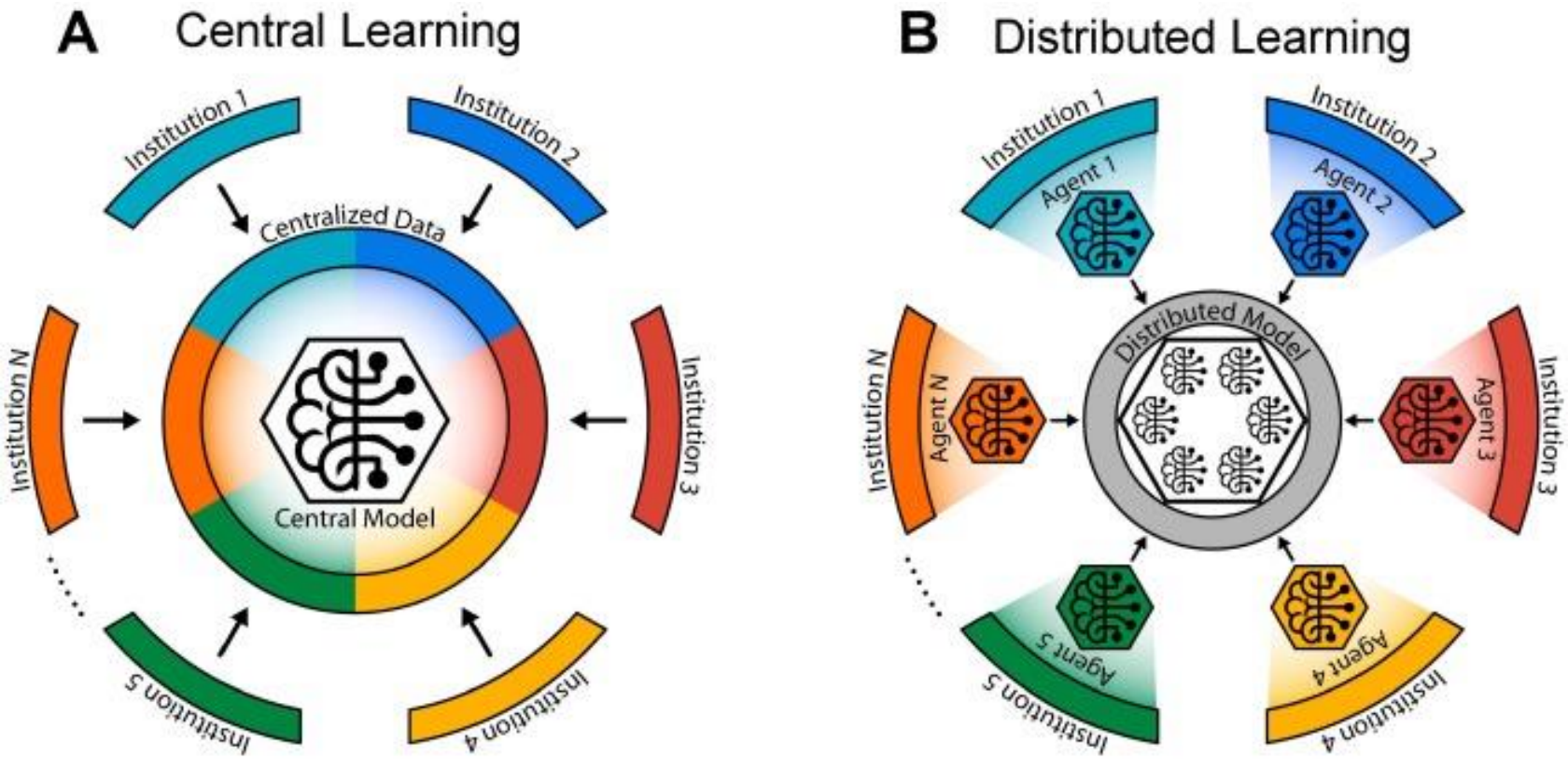
Unsupervised Learning :

1. Trying to understand the data
2. Looking for patterns in data
3. Does not required labeled data

Reinforcement Learning:

- Make a sequence of decisions
- An agent interacts with the environment and improves reward.

Centralized vs Distributed Learning



Anup et al. “Building machine learning models without sharing patient data: Asimulation-based analysis of distributed learning by ensembling”

Centralized vs Distributed Learning

Centralized Systems	
advantages	**disadvantages**
• It has a central and simple system management. • It has a relatively high degree of security, because of the existence of only one kind of access to the system.	• The system performance for each user decreases when many users try to attach simultaneously. • The system is relatively expensive (hardware and software). • The scalability of the mainframe systems is extremely low.

Shifting from centralized ML to decentralized ML

- **Enhanced Data Privacy**: Decentralized ML minimizes the need for centralizing sensitive data, addressing privacy concerns and aligning with data protection regulations.

- **Reduced Latency and Real-time Inference**: By processing data locally on edge devices, decentralized ML enables quicker decision-making and real-time inference, minimizing latency.

- **Scalability through Edge Computing**: Leveraging edge devices for computation allows for scalable and parallelized machine learning tasks, improving overall efficiency.

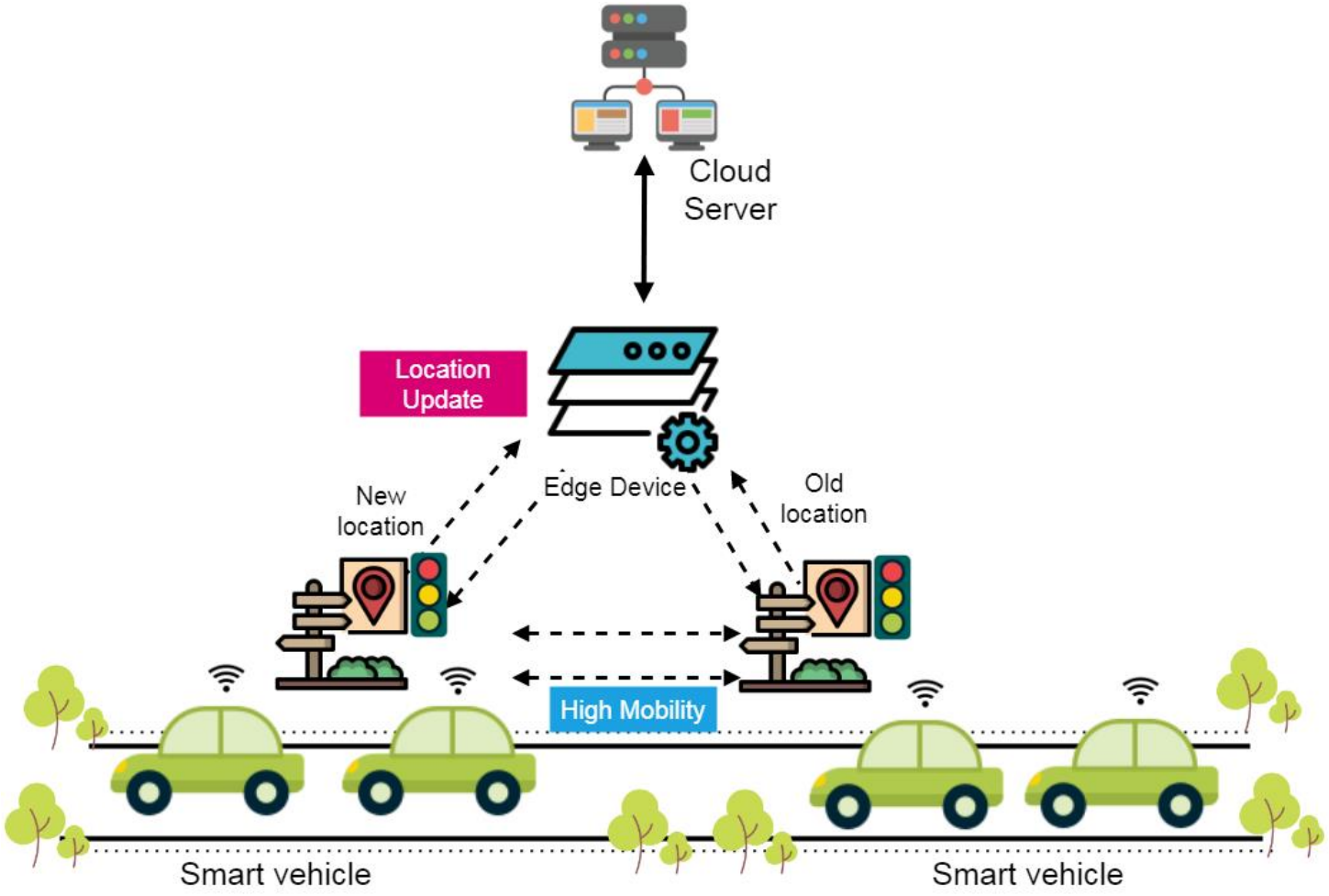
Shifting from centralized ML to decentralized ML

- **Improved Security**: Distributing the learning process across devices enhances security by reducing the risk associated with a centralized point of failure or potential security breaches.

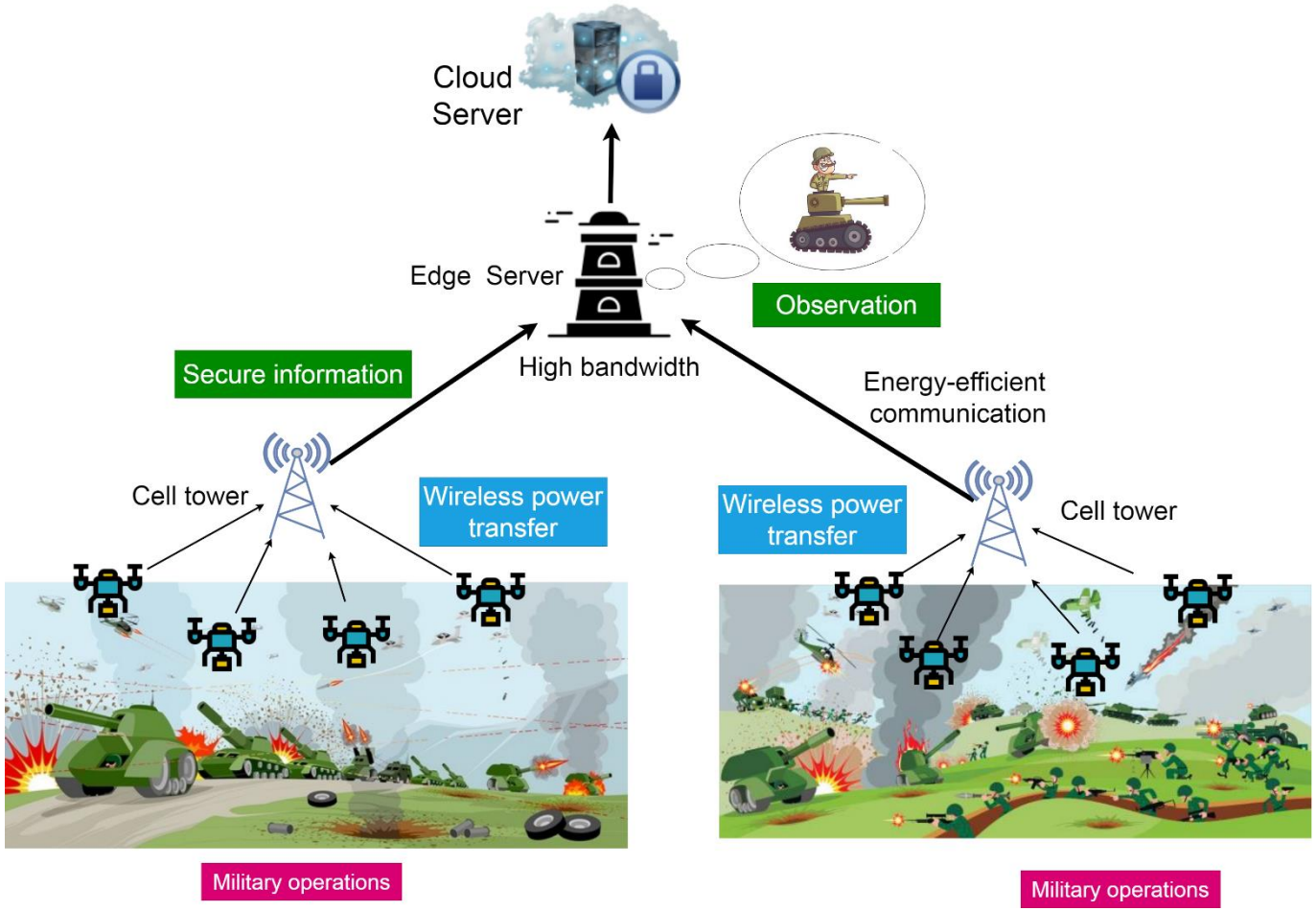
- **Energy Efficiency**: Local computation on edge devices reduces the need for continuous communication with a central server, leading to lower energy consumption.

- **Adaptability to Edge Devices**: Decentralized ML accommodates the diversity of edge devices, allowing for the deployment of machine learning models on a wide range of devices with varying capabilities.

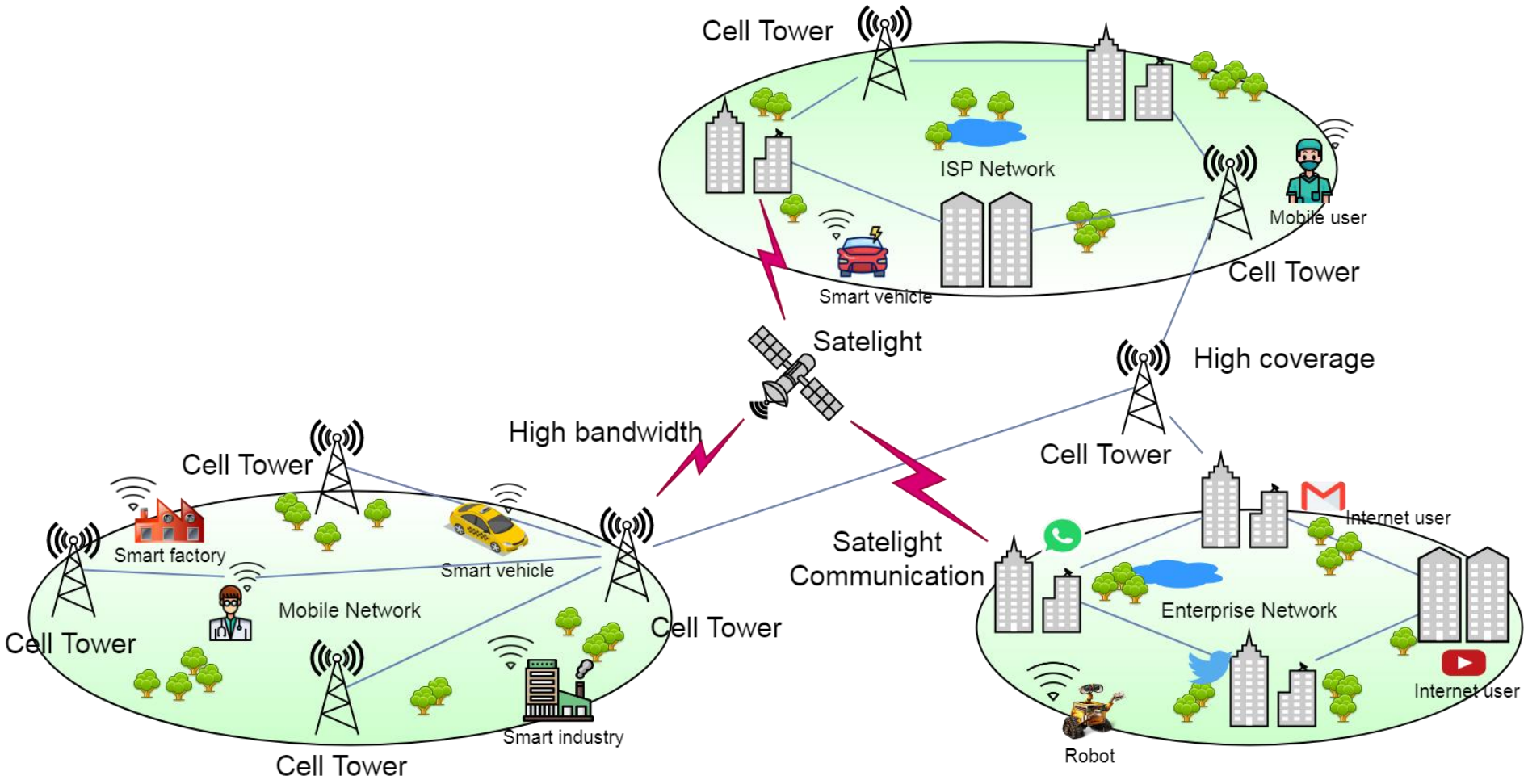
Application Area: Transportation system



Application Area: Military operation



Application Area: Industry



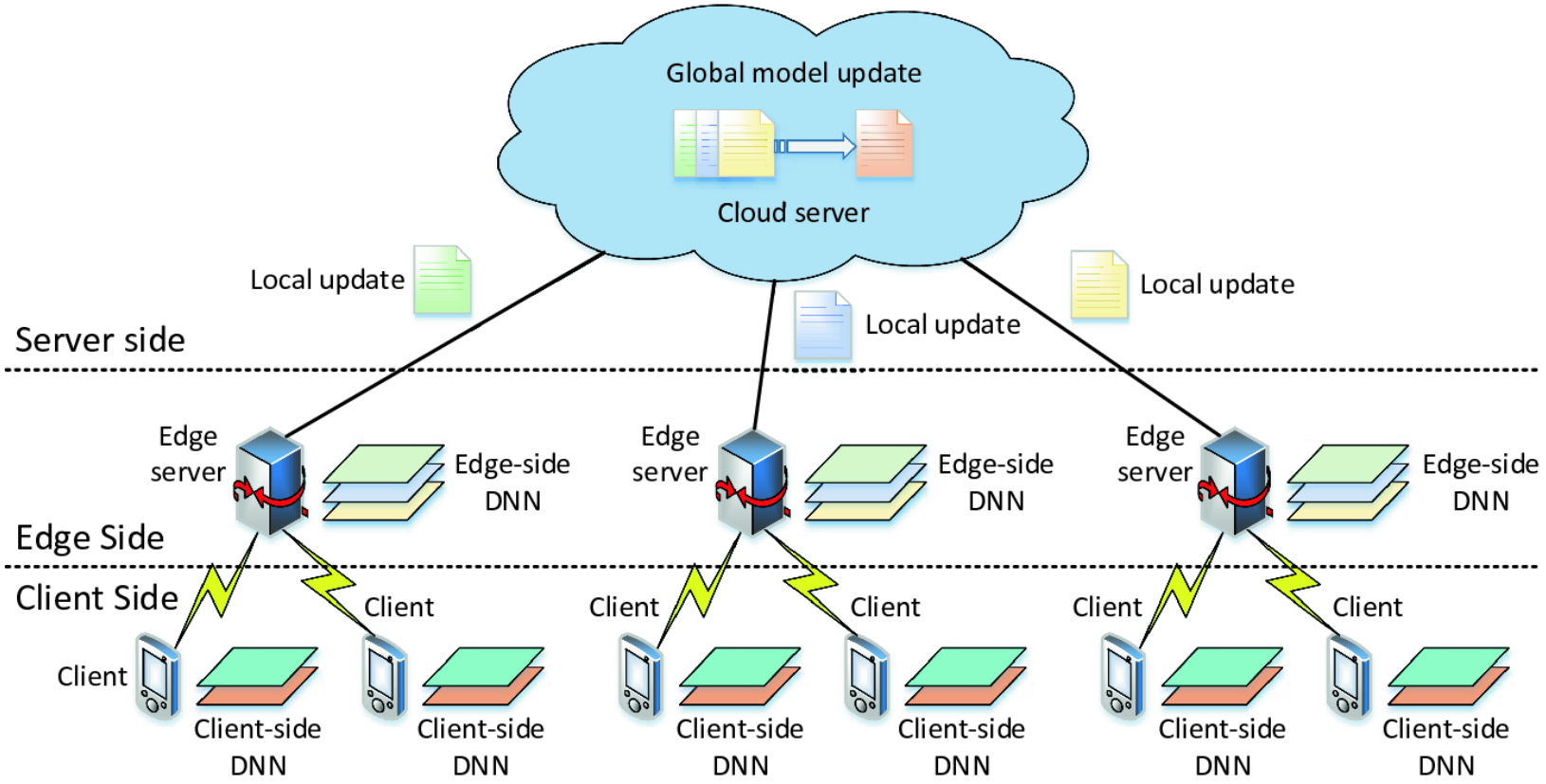
Distributed Learning

- In distributed learning, each individual agent is represented as a node in a graph.
- Edges between nodes indicate that the respective agents can exchange information.

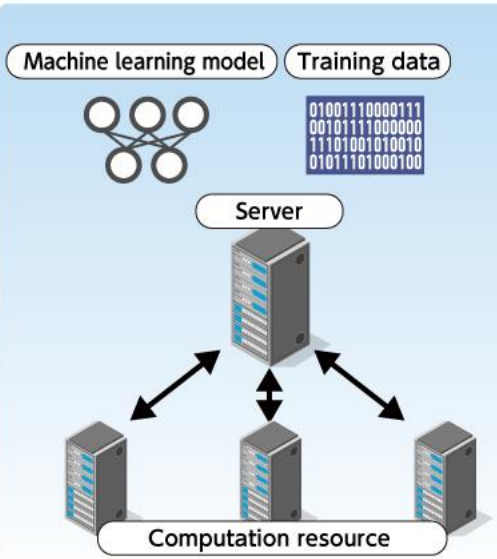
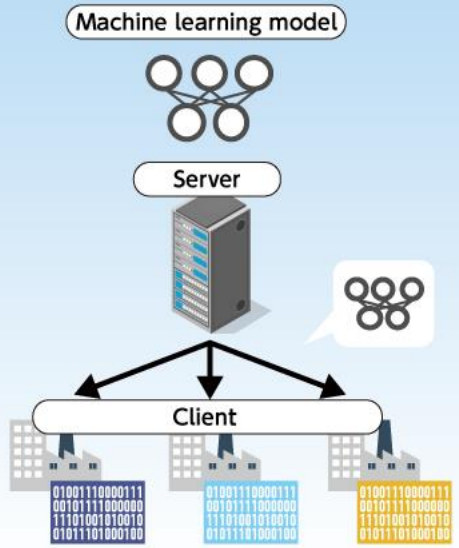
Advantage:

1. Use of multiple database
2. Performing ML tasks over distributed computing resources
3. High scalability and low maintenance.

Federated Learning



Distributed Learning vs Federated Learning

	Distributed training	Federated Learning※
Overview		
	Performing machine learning tasks over distributed computation resources	• Server distributes a model to clients in a top-down fashion • Clients train the model with their own data • Server updates the model based on client-wise training results

Federated Learning

Assume that each device $q \in \mathcal{Q}$ has a local dataset $\mathcal{D}_q = \{(\boldsymbol{x}_{q,1}, \boldsymbol{y}_{q,1}), (\boldsymbol{x}_{q,2}, \boldsymbol{y}_{q,2}), \ldots, (\boldsymbol{x}_{q,N_q}, \boldsymbol{y}_{q,N_q})\}$, where $\boldsymbol{x}_{q,v} \in \mathbb{R}^{N_{in}*1}$ is the v-th data sample characterized by N_{in} features collected by device q, and N_q is the total amount of data from device q, while $\boldsymbol{y}_{q,v} \in \mathbb{R}^{N_{out}*1}$ is the ground-truth label, corresponding to $\boldsymbol{x}_{q,v}$, respectively. It is noted that the data size of $\boldsymbol{x}_{q,v}$ relies on specific ML tasks.

Federated Learning

Let $\boldsymbol{w}_q$ represents the local model parameters of device q. Hence for device q, the loss function on its training dataset $\mathcal{D}_q$ can be represented as

$$F_{\boldsymbol{q}}\left(\boldsymbol{w}\right) = \frac{1}{N_q} \sum_{v=1}^{N_q} f\left(\boldsymbol{w}; \boldsymbol{x}_{q,v}, \boldsymbol{y}_{q,v}\right), \quad \forall q \in \mathcal{Q} \tag{4}$$

where $f\left(\boldsymbol{w}; \boldsymbol{x}_{q,v}, \boldsymbol{y}_{q,v}\right)$ is the loss function of local model in the q-th device, which is defined differently relying on different ML algorithms, referring to Table II.

Federated Learning

TABLE II

LOSS FUNCTIONS OF TYPICAL ML ALGORITHMS

Model	Loss function $f(\boldsymbol{w}, \boldsymbol{x}_v, \boldsymbol{y}_v)$
Linear regression	$\frac{1}{2}\left\|\boldsymbol{y}_v - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_v\right\|^2$
Logistic regression	$-\log\left(1+\exp\left(-\boldsymbol{y}_v\boldsymbol{x}_v^T\boldsymbol{w}\right)\right)$
K-means	$\frac{1}{2}\min_l\left\|\boldsymbol{x}_v - \boldsymbol{w}_{(l)}\right\|^2$
Squared-SVM	$\frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{2}\max\left\{0; 1-\boldsymbol{y}_v\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_v\right\}^2$
Neural networks	Mean squared error, cross-entropy

Federated Learning

For simplicity, we assume that local datasets are statistically independent, i.e., $\mathcal{D}_q \cap \mathcal{D}_{q'} = \emptyset$ for $q \neq q'$. Let $\boldsymbol{w}$ denote the global model parameters, and hence the global loss function is denoted by

$$F(\boldsymbol{w}) \triangleq \sum_{q=1}^{Q} \frac{N_q F_q(\boldsymbol{w})}{N} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{v=1}^{N_q} f\left(\boldsymbol{w}_q; \boldsymbol{x}_{q,v}, \boldsymbol{y}_{q,v}\right), \tag{5}$$

where $N = \sum_{q=1}^{Q} N_q$ is the number of all the samples from devices, while $\boldsymbol{w}_q$ is the local model parameters of device q. The objective of FL is to find the global optimal model parameters to minimize the global loss function as

$$\boldsymbol{P}1 : \boldsymbol{w}^* = \arg\min F(\boldsymbol{w}), \tag{6}$$

Federated Learning

where $\boldsymbol{w}^* = \boldsymbol{w}_1 = \cdots = \boldsymbol{w}_Q$,

The procedure for one global iteration of FL is summarized as follows:

- Devices train their local models in parallel.
- Devices upload their new-updated models to the UAV.
- UAV aggregates the received local models and broadcasts the new global model to the devices.

Federated Learning

$\boldsymbol{P2}$.

$$\boldsymbol{P2}: \min_{\boldsymbol{h}_q \in \mathbb{R}^{N_{in}*1}} L_q\left(\boldsymbol{w}^{(i)}, \boldsymbol{h}_q\right) \triangleq F_q\left(\boldsymbol{w}^{(i)} + \boldsymbol{h}_q\right) - \left(\nabla F_q\left(\boldsymbol{w}^{(i)}\right) - \xi \nabla F\left(\boldsymbol{w}^{(i)}\right)\right)^{\mathrm{T}} \boldsymbol{h}_q, \quad (7)$$

where $\boldsymbol{h}_q$ is the difference between the global model and the local model of device q, while ξ is a constant. To solve $\boldsymbol{P2}$, relying on the gradient method, each device q updates $\boldsymbol{h}_q$ by

$$\boldsymbol{h}_q^{(i),(j+1)} = \boldsymbol{h}_q^{(i),(j)} - \delta \nabla L_q\left(\boldsymbol{w}^{(i)}, \boldsymbol{h}_q^{(i),(j)}\right), \quad (8)$$

Federated Learning

where $\boldsymbol{h}_q^{(i),(j)}$ represents the value of $\boldsymbol{h}_q$ at the j-th local iteration, while δ is the step size. Eq. (8) will be executed repeatedly until the given local accuracy η is satisfied, i.e.,

$$L_q\left(\boldsymbol{w}^{(i)},\ \boldsymbol{h}_q^{(i),(j)}\right) - L_q\left(\boldsymbol{w}^{(i)}, \boldsymbol{h}_q^{(i),*}\right) \leq \eta\left(L_q\left(\boldsymbol{w}^{(i)}, \boldsymbol{h}_q^{(i),(0)}\right) - L_q\left(\boldsymbol{w}^{(i)}, \boldsymbol{h}_q^{(i),*}\right)\right). \quad (9)$$

Federated Learning

When the optimal solution $h_q^{(i),*}$ is obtained, the device will send it to the UAV. After receiving all the $h_q^{(i),*}$ from each device, the UAV updates the global model parameters by

$$\boldsymbol{w}^{(i+1)} = \boldsymbol{w}^{(i)} + \frac{1}{Q}\sum_{q=1}^{Q} \boldsymbol{h}_q^{(i),*}, \tag{10}$$

and broadcast the new value to each device. The aforementioned FL procedure will be executed repeatedly until the given global accuracy ϵ is satisfied, i.e.,

$$F\left(\boldsymbol{w}^{(i)}\right) - F\left(\boldsymbol{w}^{*}\right) \leq \epsilon\left(F\left(\boldsymbol{w}^{(0)}\right) - F\left(\boldsymbol{w}^{*}\right)\right) \tag{11}$$


Thank You!