EDGE AI Academy

Summer School

7-8 July 2025, Pisa, Italy

Edge AI Academy Summer School

Federated Learning at the Edge

Sevi Politi (spoliti@itml.gr), ITML, Greece

AI/ML Technical Leader, Machine Learning Engineer

Konstantinos Lieros (klieros@gnt.gr), GNT, Greece

Machine Learning Engineer

itml innovation applied

GNT GENERATING NEW TECHNOLOGIES

7-8 July 2025 Pisa, Italy

Outline

- ITML & GNT at a Glance
- Edge Computing
- Federated Learning
- Tools & Frameworks
- EdgeAI: Air Quality Monitoring & Prediction
- Key Takeaways

Copyright © 2025

ITML at a Glance

About itml innovation applied — Technology provider & D&C Leader in EDGE AI

- Established in 2011 in Athens, Greece
- Offices also in Limassol, Cyprus and Norwich, UK

- > **90** people and growing
- 60% engaged in R&I

- Intensive R&D: > **35** ongoing and > **40** successfully closed EU funded projects
- > **10** open accounts in B2B projects

- Servicing customers EU-wide
- Cybersecurity | Data and AI | Industry 5.0 Mobility and transport | Training/Education

Competence & Target Markets

Cybersecurity Services
Security Infusion: In-house SIEM for infra monitoring - 24/7 dedicated support teams

Analytics-as-a-Service
Data analysis on real-time or batch data streams in a time & cost-efficient manner

User Experience (UX) Design
Based on the principles of Lean UX & Agile development

SW Application Development
Front-end and back-end development services

Explore more: https://itml.gr | itmlsolutions | itml-solutions | itmlgr

Copyright © 2025

GNT at a Glance

About GNT Technology provider in EDGE AI

GENERATING NEW TECHNOLOGIES

Location: Athens, Greece

Established: 1991

Working areas: security technologies, development of tailor-made software solutions.

Expertise: >20 years of experience design and development of software systems

> 33 employees, including scientific staff and highly skilled IT engineers

http://www.gnt.gr/

Services

Development, installation, configuration and technical support of software applications and integrated information systems.

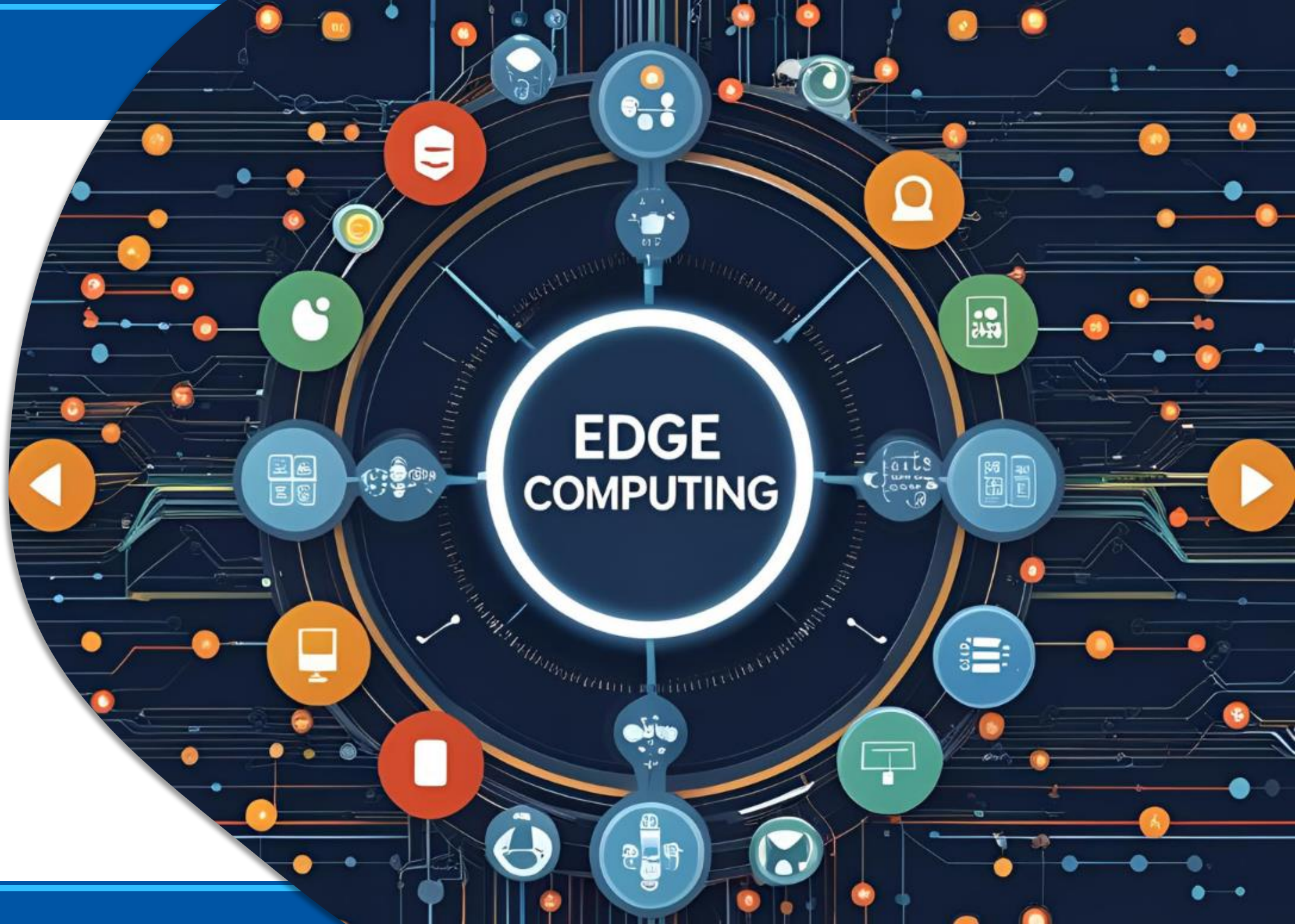
Completion and configuration of third parties' software products. Design and implementation of interconnection and interoperability of information systems.

Feasibility studies, analysis and design of integrated information systems.

Consulting services. Education and expertise transfer.

Mainly telecommunications sector, as well as sports betting management, integrated payment systems, lottery game software etc.)

Copyright © 2025

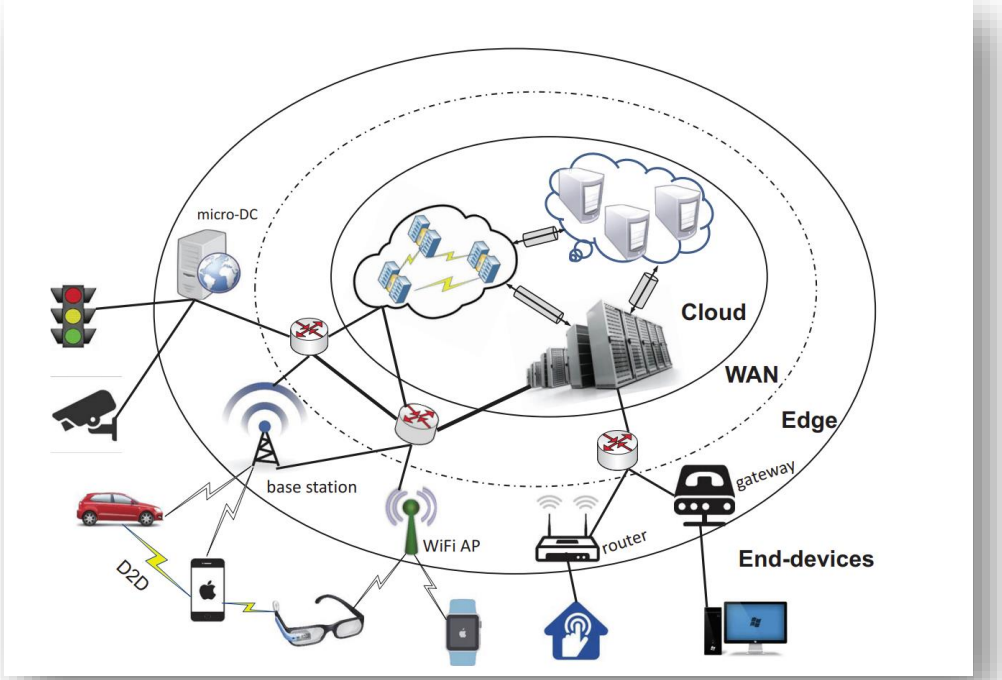

EDGE
COMPUTING

What is Edge Computing?

Edge computing refers to the practice of processing data near the source of data generation (the "edge" of the network), rather than relying solely on centralized cloud data centers.

Key Characteristics & Benefits:

- **Proximity to Data:** Computation occurs close to IoT devices, sensors, or local edge servers.
- **Reduced Latency:** Local data processing minimizes delays, enabling faster response times.
- **Real-Time Analytics:** Supports time-sensitive applications like autonomous vehicles and smart cities.
- **Bandwidth Efficiency:** Less data needs to be sent to the cloud, improving bandwidth usage.
- **Enhanced Privacy & Security:** Local processing can limit data exposure.
- **Distributed Architecture:** Relies on computing resources deployed across multiple locations.



https://www.ibm.com/think/topics/edge-computing – Zhou et al., 2019. Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing (arXiv:1905.10083)

Copyright © 2025

Popular Edge Devices

Device	NVIDIA Jetson Orin Nano Devkit	Raspberry Pi 5	Google Coral Dev Board
CPU	6-core ARM Cortex-A78AE	Quad-core ARM Cortex-A76 @ 2.4GHz	Quad-core ARM Cortex-A53
GPU	NVIDIA Ampere 1024-core	VideoCore VII	GC7000 Lite Graphics
RAM	8 GB LPDDR5	4 GB / 8 GB LPDDR4X	1 GB LPDDR4
Storage	MicroSD / NVMe (M.2)	MicroSD / PCIe SSD	8 GB eMMC / microSD
Power Consumption	7-15W	~12W (peak)	~5W

Copyright © 2025

Challenges of AI Deployment on Edge Devices

Deploying AI on the edge involves significant technical and operational hurdles due to the unique characteristics of edge environments.

Hardware Limitations

- **Limited CPU & RAM**: Edge devices often can't support large or complex AI models.
- **Constrained Storage**: Limits ability to store large datasets or model files.
- **No Dedicated GPU**: Inference and training are slow or impractical without GPU acceleration.
- **Power Constraints**: Battery-operated devices require energy-efficient computation.
- **Hardware Variability**: Diverse specs across devices complicate standardized deployment.

Scalability & Network Constraints

- **Low Bandwidth:** Sending large volumes of data to the cloud consumes heavy network bandwidth.
- **Latency:** Delay in cloud access may not meet the demands of time-sensitive applications.
- **Reliability**: Dependence on stable wireless and backbone networks can limit service availability in mission-critical settings.
- **Scalability**: Managing, updating, and monitoring thousands to millions of edge devices is complex.

Data & Security Concerns

- **Data Quality & Consistency**: Ensuring accurate, clean, and synchronized data is difficult at the edge.
- **Privacy & Security Risks**: Distributed nodes increase the risk of breaches and data leakage.

Wang et al., 2019. Convergence of Edge Computing and Deep Learning (arXiv:1907.08349) -
Nayak et al., 2021. Review on Edge Analytics (arXiv:2107.06835)

Copyright © 2025

Mitigation Strategies of AI Deployment on Edge Devices

1. **Federated Learning**
   - Maintains trust and data integrity at the edge, sharing only model updates.
   - Enhances privacy and reduces data transmission.
   - Devices handle partial data; load is distributed.
   - Addresses data quality through collaborative validation and alignment of model updates, reducing drift and improving reliability.

2. **Lightweight AI Models & Compression**
   - Enables fast, low-resource inference.

3. **Optimized Deployment with Docker**
   - Containerized AI models scale easily by quickly replicating across multiple edge devices.
   - Two-stage Dockerfiles build lean containers.
   - Resource limits support constrained environments.

Liu et al., 2019. A Survey on Edge Computing Systems and Tools (arXiv:1911.02794) - Zhou et al., 2019. Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing (arXiv:1905.10083) - https://docs.docker.com/

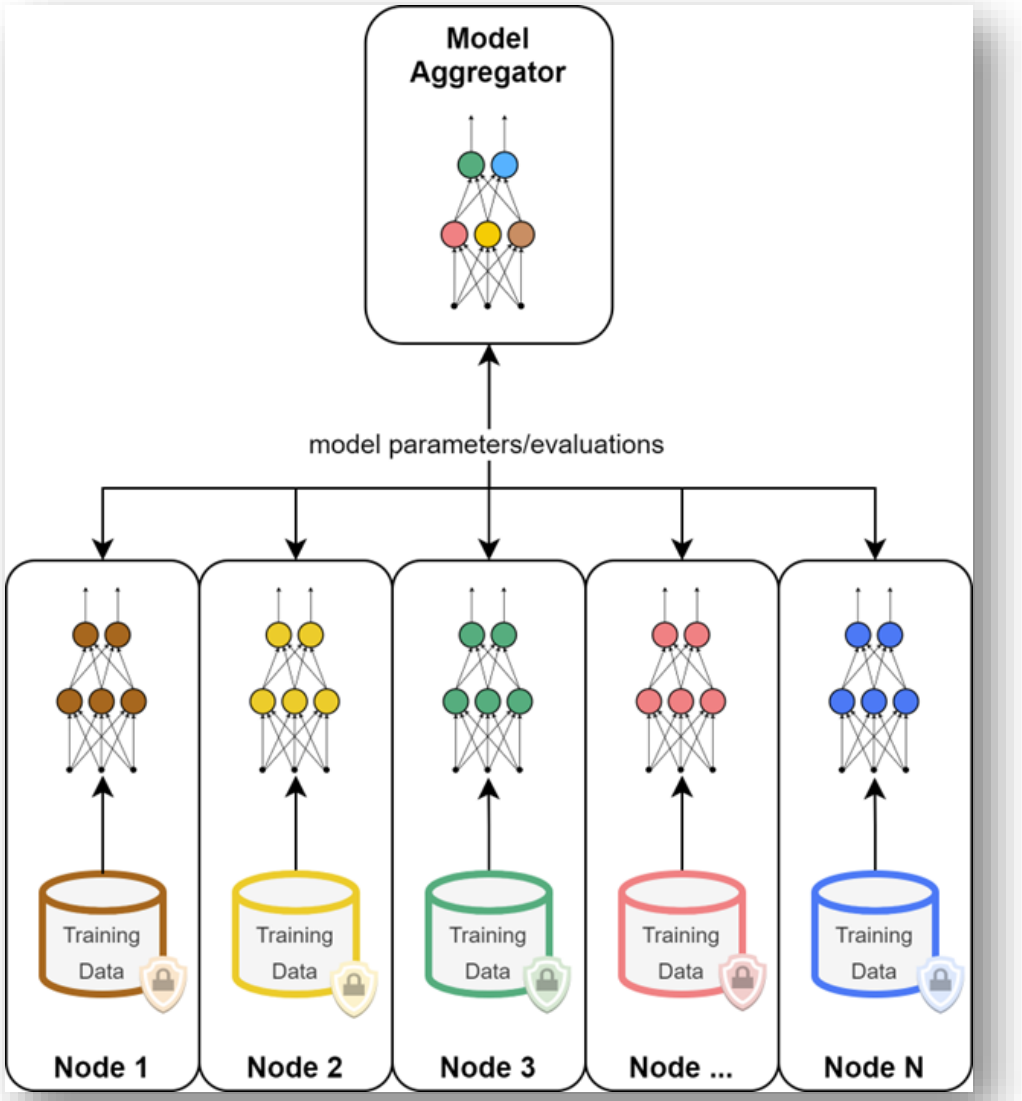
Copyright © 2025

Federated Learning

Copyright © 2025

Overview – What is Federated Learning (FL)

- An ML technique where multiple devices (clients) collaboratively **train a model without sharing local data**.
- Local models are trained on **individual devices** (e.g., smartphones, IoT devices), and only model updates (parameters) are shared with a central server, not the actual data.
- The server **aggregates** these updates to improve the global model.



CyberSecDome D2.1 State of the art, Reference Pilot Scenarios, Requirements, and Analysis

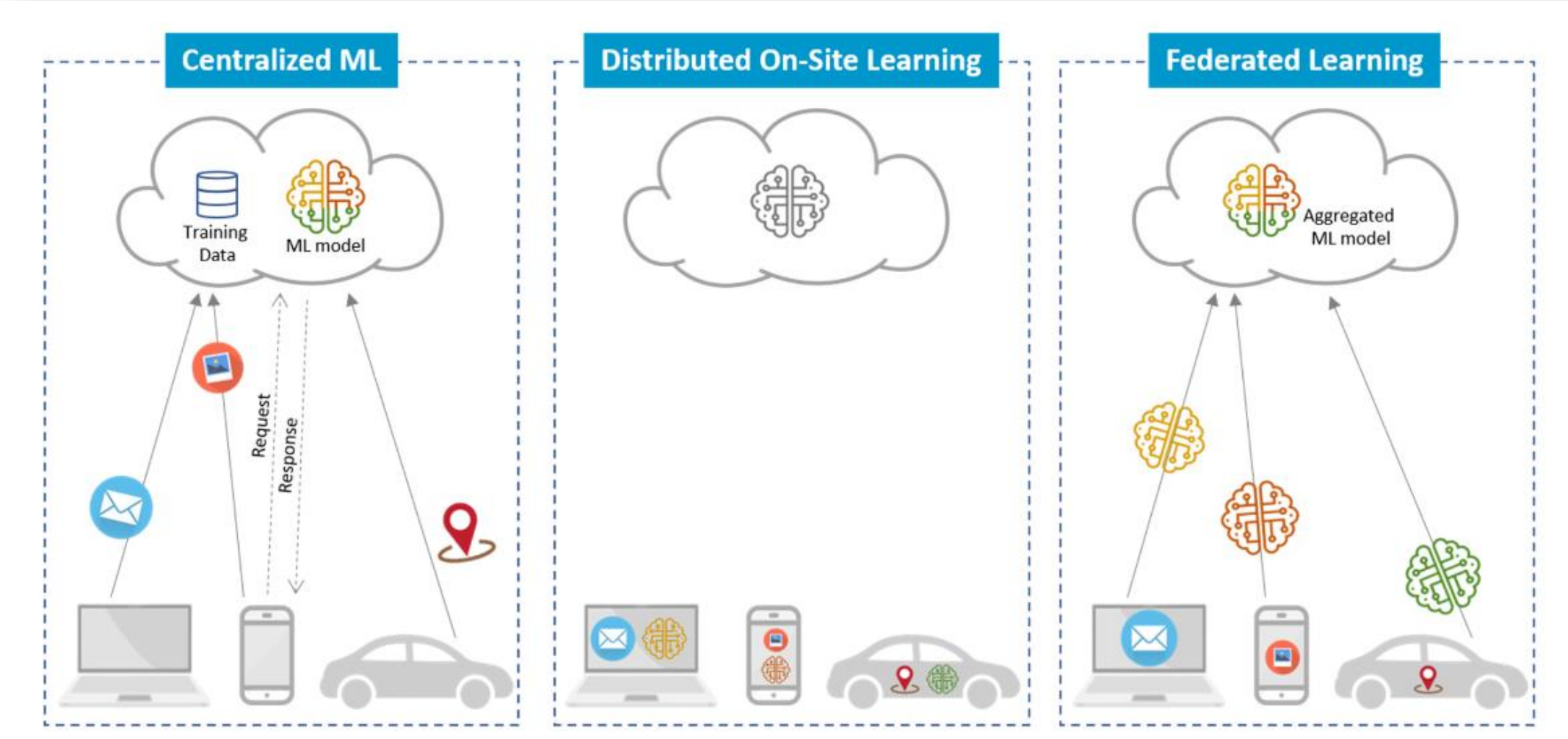
Copyright © 2025

Importance

- **Privacy Preservation:** Sensitive data stays local, private and is not exposed.
  - Collaboration while retaining the competitive advantage.
  - GDPR compliance.
- **Diverse Data Representation:** FL learns from diverse sources.
  - Capturing collective knowledge, data sources variability.
  - More robust, real-world models.
- **Efficiency Optimization:** Reduces data transfer and cloud infrastructure
  - Lower communication overhead, optimizes bandwidth usage.
  - Lower storage requirements.
- **Continual Learning:** Models remain adaptive and relevant.
  - Ongoing, live collaborative training.

Copyright © 2025

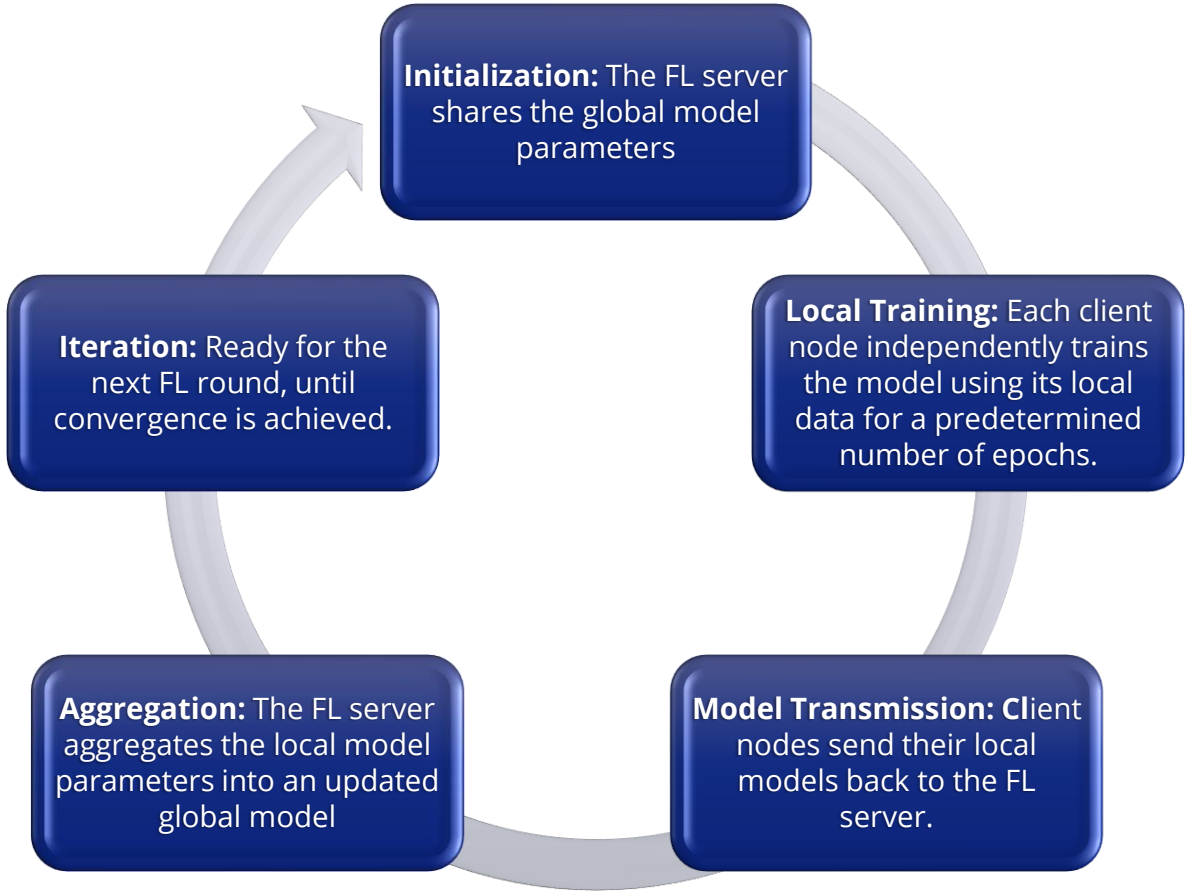
Federated vs. Centralized Learning

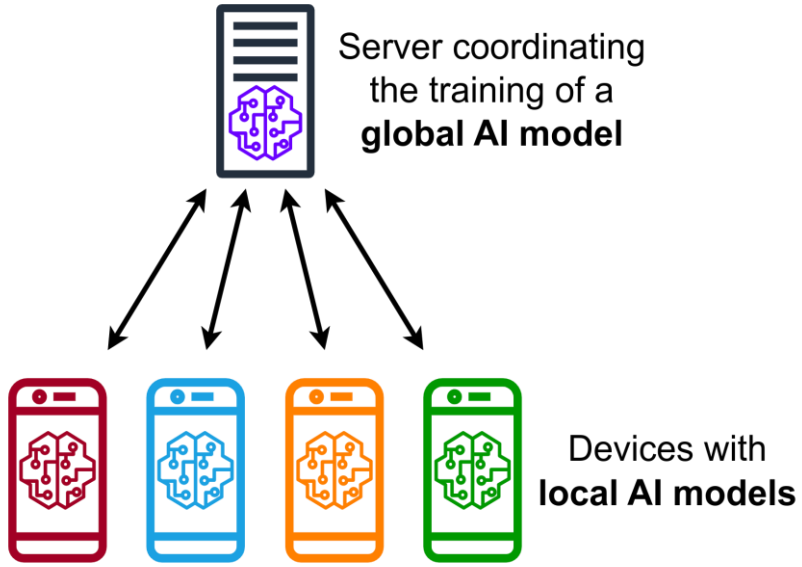


S. Abdulrahman, H. Tout, H. Ould-Slimane, A. Mourad, C. Talhi and M. Guizani, "A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond," in IEEE Internet of Things Journal, vol. 8, no. 7, pp. 5476-5497, 1 April1, 2021

Copyright © 2025

FL Round - Flow



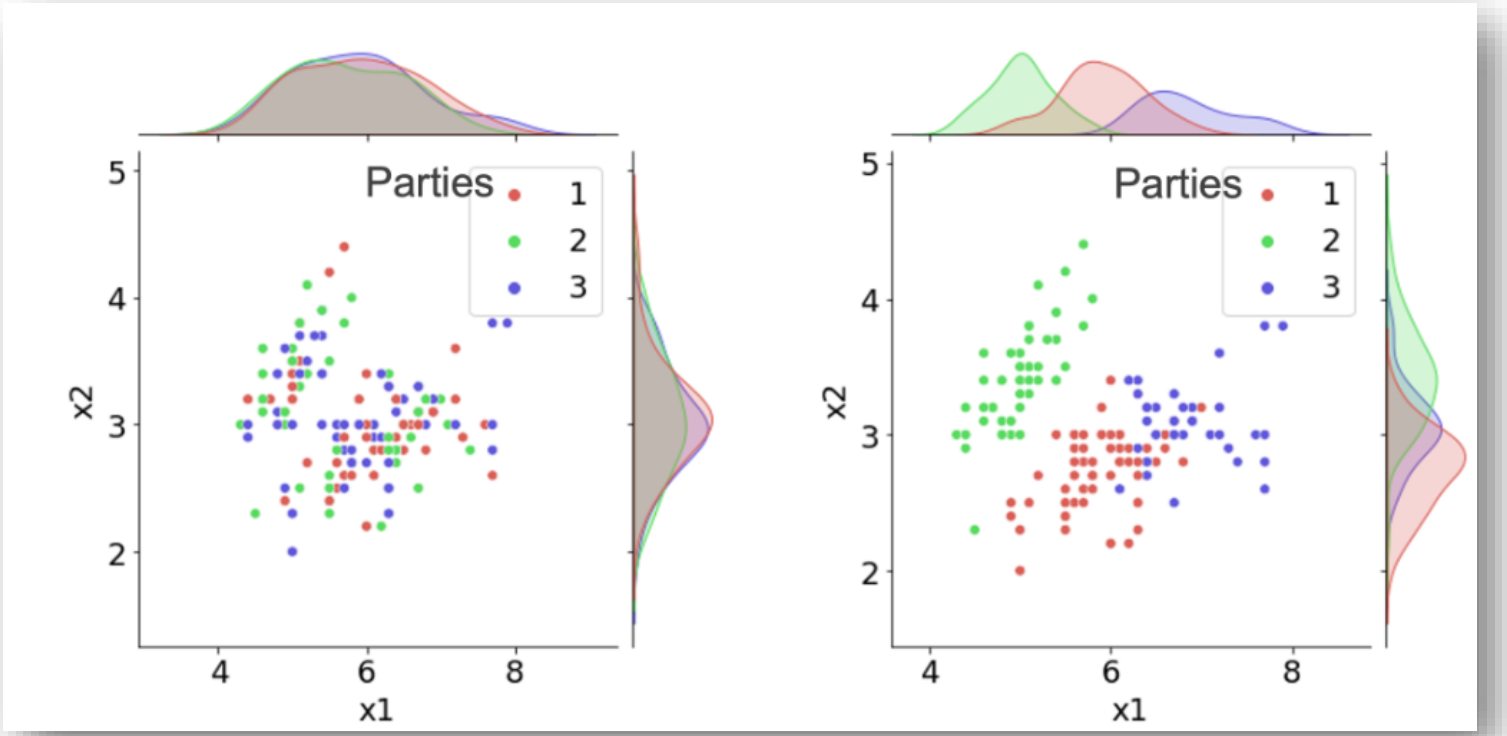


Copyright © 2025

FL Essential Aspects & Key Concepts

Data Distribution: IID vs. Non-IID

- **Independent and Identically Distributed (IID)** : Data is similar and independently sampled, simplifying model aggregation.
- **Non-IID:** Data varies (e.g., feature distributions, class imbalances), requiring advanced aggregation strategies.



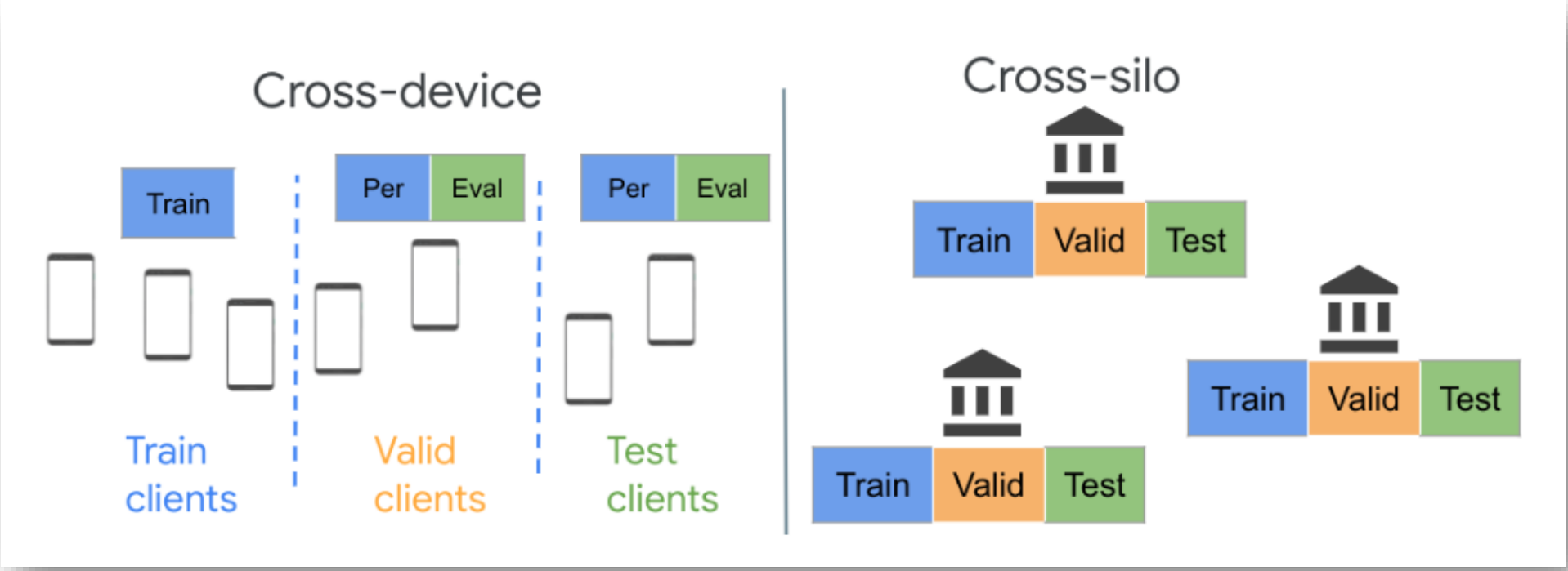
D. -V. Nguyen, A. -K. Tran and K. Zettsu, "FedProb: An Aggregation Method Based on Feature Probability Distribution for Federated Learning on Non-IID Data," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 2875-2881

Copyright © 2025

FL Essential Aspects & Key Concepts

Cohort Characteristics

- **Cross-Device**: Small data from many devices (e.g., smartphones, wearables). Example: Google's Gboard.
- **Cross-Silo:** Fewer, larger entities (e.g., hospitals, institutions) that operate in silos due to legal, privacy, or competitive reasons.



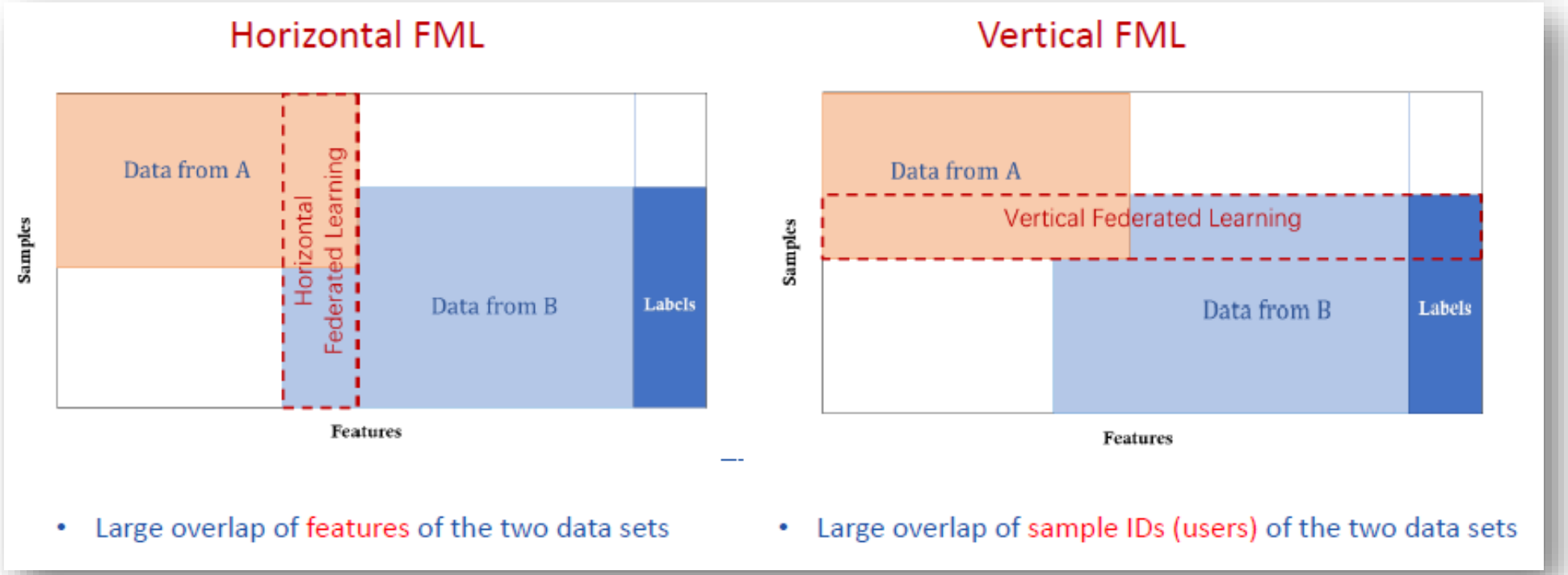
Wu, Shanshan, et al. "Motley: Benchmarking heterogeneity and personalization in federated learning." arXiv preprint arXiv:2206.09262 (2022).

Copyright © 2025

FL Essential Aspects & Key Concepts

Data Partitioning

- **Horizontal FL**: Same features, different samples (e.g., images of different people).
- **Vertical FL**: Same samples, different features (e.g., same users, different attributes).
- **Federated Transfer Learning:** Different samples and features across diverse devices.



Yang, Qiang, et al. "Federated machine learning: Concept and applications." ACM Transactions on Intelligent Systems and Technology (TIST) 10.2 (2019): 1-19.

Copyright © 2025

FL Key Challenges

Data Heterogeneity

Non-IID data can cause models to perform poorly when generalized across all clients.

Communication Efficiency

Transmitting model updates from numerous distributed devices can create significant communication overhead

Resource Constraints

Training large models in constrained edge devices can be challenging.

Privacy and Security

Data remain local, but transmitted model parameters can still be vulnerable to adversarial attacks.

System Heterogeneity

Devices in FL networks vary in hardware, software, and network conditions, causing inconsistent or delayed model convergence/aggregation.

Copyright © 2025

FL Mitigation Strategies

1

Optimized Deployment

- **Containerization** enables scalable and consistent deployment across diverse client environments.
- Supports **resource-limited devices** with controlled usage.
- Leverages tools like Docker for simplified orchestration.

2

Robust Aggregation & Secure Communication

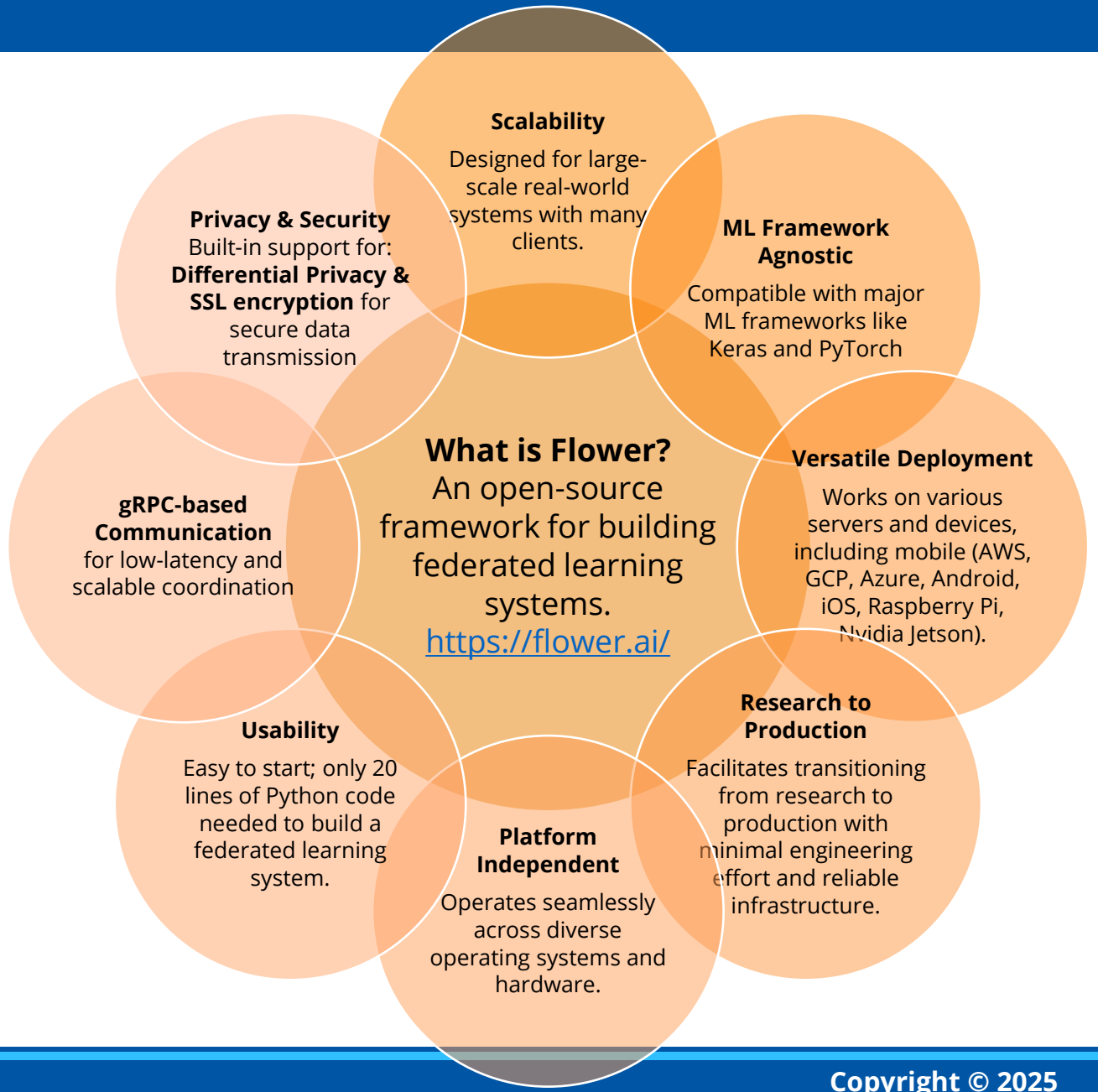
- **Custom aggregation strategies** to handle data heterogeneity and client reliability.
- Ability to **exclude low-quality clients** and limit influence of noisy updates.
- **Efficient communication layer** to support scalable, real-time coordination between clients and server.
- **Integrated privacy and security mechanisms** to ensure data protection during training and transmission.

Copyright © 2025

Tools & Frameworks

Copyright © 2025

Flower Framework



Copyright © 2025

Model Compression Techniques

Pruning (TensorFlow)

Pruning is the process of removing unnecessary weights or neurons from a trained neural network, resulting in a smaller and more efficient model with little loss in accuracy. TensorFlow Model Optimization Toolkit allows structured and unstructured pruning during or after training.

https://www.tensorflow.org/model_optimization/guide/pruning

Quantization with ONNX

Quantization reduces the precision of the model's weights and activations (e.g., from float32 to int8), significantly lowering memory footprint and improving inference speed. ONNX Runtime provides post-training and dynamic quantization tools compatible with multiple hardware accelerators.

https://onnxruntime.ai/docs/performance/quantization.html

ONNX (Open Neural Network Exchange)

ONNX is an open-source format for representing machine learning models. It allows models trained in different frameworks (like PyTorch, TensorFlow) to be converted and optimized for inference across various hardware platforms using ONNX Runtime.

https://onnx.ai

Copyright © 2025

Docker for Federated Learning Deployment

What is Docker?

An open-source platform that automates the deployment of applications in lightweight, portable containers, ensuring consistent performance across different environments.

Why Docker?

- **Consistency and Ease of Deployment:** Docker ensures that clients run the same environment (e.g., TensorFlow, Flower) in isolated containers, while the server is containerized separately with its own configuration, simplifying deployment across devices and servers.
- **Scalability:** Easily scales containers up or down to match the number of clients in the system.
- **Resource Efficiency:** Enables better control over resource usage, ensuring containerized workloads do not exceed hardware limits on edge devices.

https://www.docker.com/

Copyright © 2025

MLflow for Experiment Tracking

What is MLflow?

- An open-source platform for managing the ML lifecycle, including experiment tracking, model management, and deployment.

Why MLflow?

- **Track Experiments**: In federated learning, MLflow tracks model updates from clients and the server, essential for evaluating local and global models.
- **Metrics & Logging**: It logs performance metrics (e.g., accuracy, loss) across multiple FL rounds, focusing on the final round's results.
- **Model Registration**: MLflow's Register Model feature helps track the best global model across different FL simulations or register multiple models under a single name.

https://mlflow.org/

Copyright © 2025

Federated Learning Runs Tracked in MLflow

edgeai-fl-experiments Provide Feedback Add Description

metrics.rmse < 1 and params.model = "tree" | Time created | State: Active | Datasets | Sort: Created | Columns | Expand rows | Group by | New ru

Table | Chart | Evaluation Experimental

				Metrics					Parameters						
Run Name	Creat	Duration	Models	mp.test.mae	mp.test.mape	mp.test.mse	system/cpu_utilization_percentage	system/system_memory_usage_percentage	activation	batch_size	dense_units	dropout_rate	epoch	num_clients	num_rounds
edgeai-rspu2 2025-06-09 13:11:52	✓	32.0min	edgeai_lst.../12	-	-	-	0.9	44.3	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu2_horizon_24	✓	10.3s	-	23.198259...	1.7172196...	1180.79...	4.1	44.4	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu2_horizon_1	✓	10.7s	-	8.5965795...	0.7428464...	164.145...	4.2	42.6	relu	32	160	0	40	-	-
edgeai-rspu1 2025-06-09 13:11:52	✓	32.0min	edgeai_lst.../12	-	-	-	6.9	14.9	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu1_horizon_24	✓	4.2s	-	23.184544...	1.7163416...	1178.95...	8.7	14.8	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu1_horizon_1	✓	4.5s	-	8.5961429...	0.7425851...	164.159...	9.3	14.9	relu	32	160	0	40	-	-
edgeai-server 2025-06-09 13:11:52	✓	2.2h	-	-	-	-	5.9	14.9	-	-	-	-	-	2	3
edgeai-rspu2 2025-06-08 11:06:37	✓	32.2min	edgeai_lst.../11	-	-	-	0.6	44.1	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu2_horizon_24	✓	10.4s	-	17.312690...	1.8889029...	649.108...	4.3	44.1	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu2_horizon_1	✓	10.7s	-	8.0176642...	0.7061090...	155.616...	5.9	42.8	relu	32	160	0	40	-	-
edgeai-rspu1 2025-06-08 11:06:37	✓	32.2min	edgeai_lst.../11	-	-	-	10	15	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu1_horizon_24	✓	4.6s	-	17.312944...	1.8889679...	649.134...	5.1	14.9	relu	32	160	0	40	-	-
edgeai_lstm_aqi_rspu1_horizon_1	✓	4.2s	-	8.0177808...	0.7059277...	155.617...	6.6	14.6	relu	32	160	0	40	-	-
edgeai-server 2025-06-08 11:06:37	✓	2.1h	-	-	-	-	8.1	15	-	-	-	-	-	2	3

Use-Case: Air Quality Monitoring and Prediction

Copyright © 2025

Design & Implementation

Air quality monitoring and prediction

Goal: system which offers precise, real-time air quality monitoring, enabling cities to make data-driven decisions that improve urban living conditions.

- Sensor nodes, placed strategically on city intersections continuously collect data for temperature, humidity and pollutants to compute Air Quality Index (AQI) in real time. They use the LoRa network to transfer data from each sensor node to a Remote Sensing Processing Unit (RSPU).
- Each RSPU, leverages timeseries ML technologies to analyze data and predict AQI.

Pollutants:

1. Particulate Matter 2.5 (PM2.5)
2. Particulate Matter 10 (PM10)
3. Nitrogen Dioxide (NO2)
4. Ozone (O3)
5. Sulfur Dioxide (SO2) Sensor

AQI Breakpoints:

POLLUTANT	INDEX LEVEL *(based on polluant concentrations in µg/m3)*					
	1 Very good	2 Good	3 Medium	4 Poor	5 Very Poor	6 Extremely Poor
Ozone (O_3)	0-50	50-100	100-130	130-240	240-380	380-800
Nitrogen dioxide (NO_2)	0-40	40-90	90-120	120-230	230-340	340-1000
Sulphur dioxide (So_2)	0-100	100-200	200-350	350-500	500-750	750-1250
Particules less than 10 µm (PM_{10})	0-20	20-40	40-50	50-100	100-150	150-1200
Particules less than 2.5 µm ($PM_{2.5}$)	0-10	10-20	20-25	25-50	50-75	75-800

Note: PM10 and PM2.5 values are based on 24-hour running means

European Air Quality Index Calculation, ECMWF Projects [Online]. Available: https://ecmwf-projects.github.io/copernicus-training-cams/proc-aq-index.html [Accessed: 4 July 2025]

Copyright © 2025

Design & Implementation

Purpose

- AI models for **real-time** environmental monitoring at the **edge.**
- Predict future values of **environmental parameters** (e.g., pollution levels and temperature).
- Smaller, faster models without significant loss of accuracy, for resource-constraint devices.

Features

- Long Short-Term Memory (**LSTM**) network.
- Captures **sequential dependencies** in the data.
- Handles multivariate time series inputs (e.g., sensor data, weather conditions).

Model Compression Techniques

- Pruning: removing unnecessary neurons/weights
- Quantization: reducing precision of weights and activations (e.g., from float32 to int8)

Federated Learning

- Live **collaborative training** across the edge devices (RSPUs).
- Raw data remain local – **privacy** preservation, **reduced bandwidth**/cloud resources.

Federated AQI Forecaster

Copyright © 2025

Design & Implementation

3-Layer Architecture

Micro-Edge Level

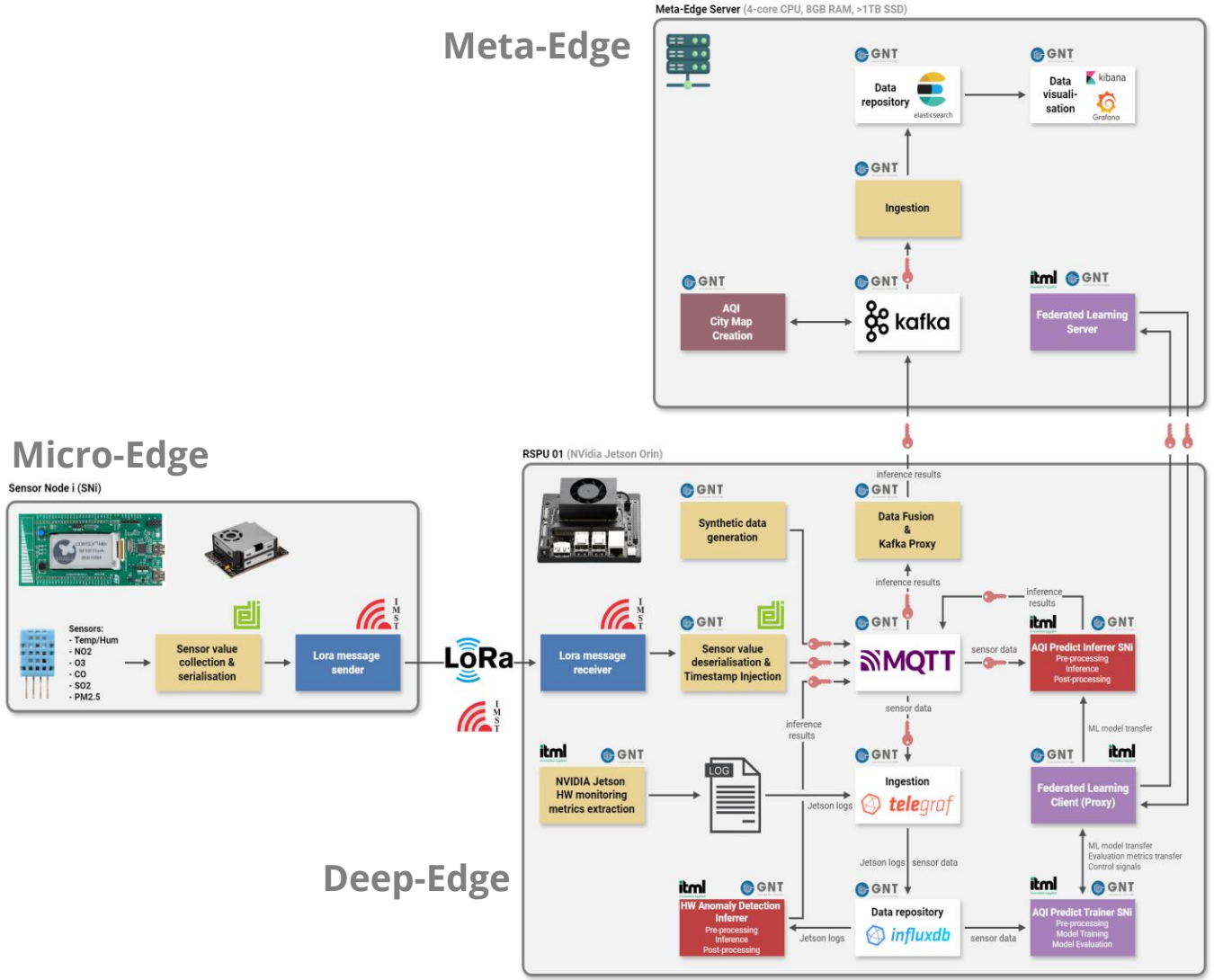
Multiple environmental sensors collect data related to air pollution and atmospheric conditions.

Deep-Edge Level

Roadside Perception Units (**RSPUs**): energy efficient, AI-driven stationary agents interconnected with a Lora 2.4 GHz Mesh Network (NVIDIA Jetson).

Meta-Edge Level

Single **Server** (mini-PC) used for centralized data aggregation from multiple RSPUs.



Copyright © 2025

Design & Implementation

Data management and Cyber-Security

- Reliable and secure data transfer channels will be provided for communications among devices on Deep and Meta Edge layers based on a distributed, scalable and fault-tolerant data management system and certificate-based authentication.
- Mechanisms for fusing heterogeneous data types is implemented to support the interconnection of services.
- The Federated Learning approach will be complemented by mechanisms for secure and encrypted ML model transfer.

Copyright © 2025

Key Takeaways

Privacy-preserving AI
Raw data stays local; only model updates are shared.

Efficient Communication
Reduces bandwidth by avoiding data centralization.

Scalable & Decentralized
Trains on diverse, distributed data across devices.

Personalization
Edge models can adapt to individual user contexts.

Low Latency Processing
Edge computing ensures real-time, responsive AI.

Key Use Cases:

Healthcare diagnostics

Smart home assistants

Autonomous vehicles

Industrial IoT Monitoring

Zhou et al., 2019. Edge Intelligence: Paving the Last Mile of Artificial Intelligence with Edge Computing (arXiv:1905.10083)

Copyright © 2025

Copyright © 2025

Thank You

EAI

https://edge-ai-tech.eu/

The two-day innovative educational initiative is designed for students, young researchers, and professionals eager to expand their understanding and skills in these cutting-edge technologies.

Participants will engage in interactive discussions, hands-on sessions, and lectures covering foundational concepts and innovative applications of edge AI across various domains with leading experts, gain insights into real-world applications, and work on team-based projects to foster collaboration and innovation.

Copyright © 2025

EDGE AI Academy 2025

Summer School Objectives

- The EDGE AI Academy Summer School aims to empower participants with advanced knowledge, hands-on experience and an in-depth understanding of edge AI, which combines Artificial Intelligence (AI) with edge computing and the Internet of Things (IoT) to process data closer to its source.
- Edge AI technology is gaining prominence due to its ability to enhance real-time decision-making, reduce latency, improve data privacy, and diminish reliance on continuous internet connectivity.
- The topics will extend to RISC-V architectures as ultra-low-power high-performance RISC-V-based processors and ultra-low-power vector-based accelerators are advancing smart IoT and edge AI computing capabilities.

Copyright © 2025