Machine learning energy consumption evaluation methodology

Mathilde Jay - 2nd year PhD Student - LIG, LIP, MIAI Edge Intelligence
mathilde.jay@univ-grenoble-alpes.fr

Denis Trystram - LIG, INP - Inria DataMove
Laurent Lefevre - Inria - LIP, Avalon

Green Days 2023

Evaluate ML energy consumption

- ML computational and energy cost
  - Metrics commonly used to evaluate it
- At the ML life cycle level
- At the ML infrastructures level
- Other ML paradigm
  - Continual Learning
  - Federated Learning

ML computational and energy cost (training and inference)

ML computational and energy cost (training and inference)

- Number of **parameters** of the model

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)
- Model **size** (Bytes)

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)
- Model **size** (Bytes)
- Number of floating point operation per seconds (**FLOPS**) required

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)
- Model **size** (Bytes)
- Number of floating point operation per seconds (**FLOPS**) required

Not necessarily correlated to the energy consumed

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)
- Model **size** (Bytes)
- Number of floating point operation per seconds (**FLOPS**) required

Not necessarily correlated to the energy consumed

★ **Energy** consumption (Joules or kWh)

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)
- Model **size** (Bytes)
- Number of floating point operation per seconds (**FLOPS**) required

Not necessarily correlated to the energy consumed

★ **Energy** consumption (Joules or kWh)

M. Jay, V. Ostapenco, L. Lefèvre, D. Trystram, A.-C. Orgerie, and B. Fichel, “An experimental comparison of software-based power meters: focus on CPU and GPU”.

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)
- Model **size** (Bytes)
- Number of floating point operation per seconds (**FLOPS**) required

Not necessarily correlated to the energy consumed

★ **Energy** consumption (Joules or kWh)

★ **Carbon** emissions

M. Jay, V. Ostapenco, L. Lefèvre, D. Trystram, A.-C. Orgerie, and B. Fichel, “An experimental comparison of software-based power meters: focus on CPU and GPU”.

ML computational and energy cost (training and inference)

- Number of **parameters** of the model
- Training and inference **duration** (GPU-hours)
- Model **size** (Bytes)
- Number of floating point operation per seconds (**FLOPS**) required

Not necessarily correlated to the energy consumed

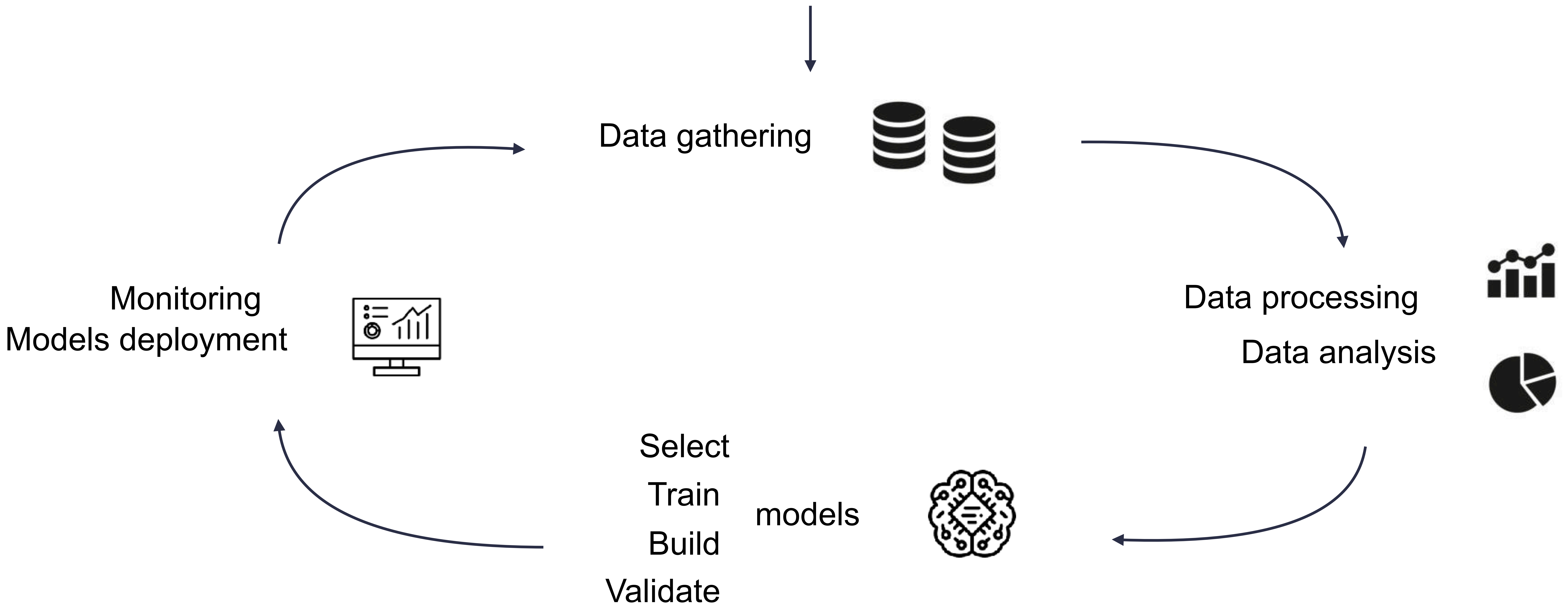
★ **Energy** consumption (Joules or kWh)

★ **Carbon** emissions

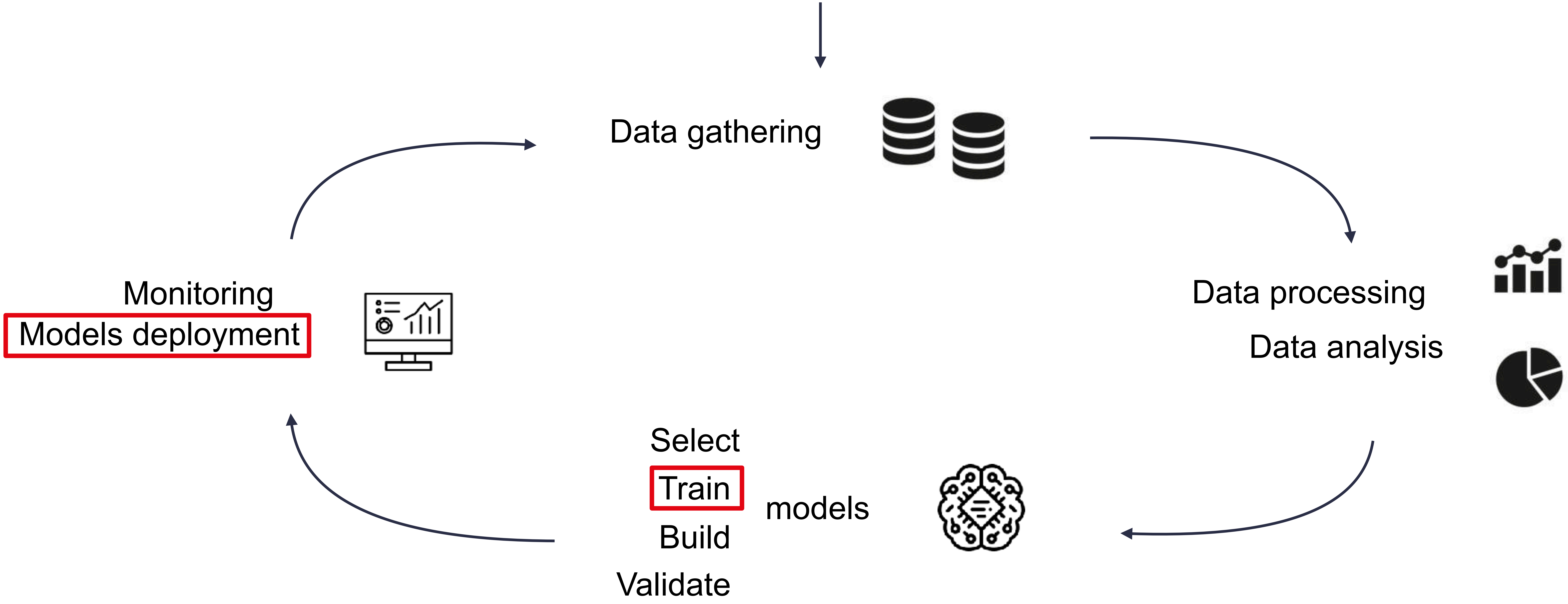
★ Energy, time or size **efficiency**

M. Jay, V. Ostapenco, L. Lefèvre, D. Trystram, A.-C. Orgerie, and B. Fichel, “An experimental comparison of software-based power meters: focus on CPU and GPU”.

ML development life cycle

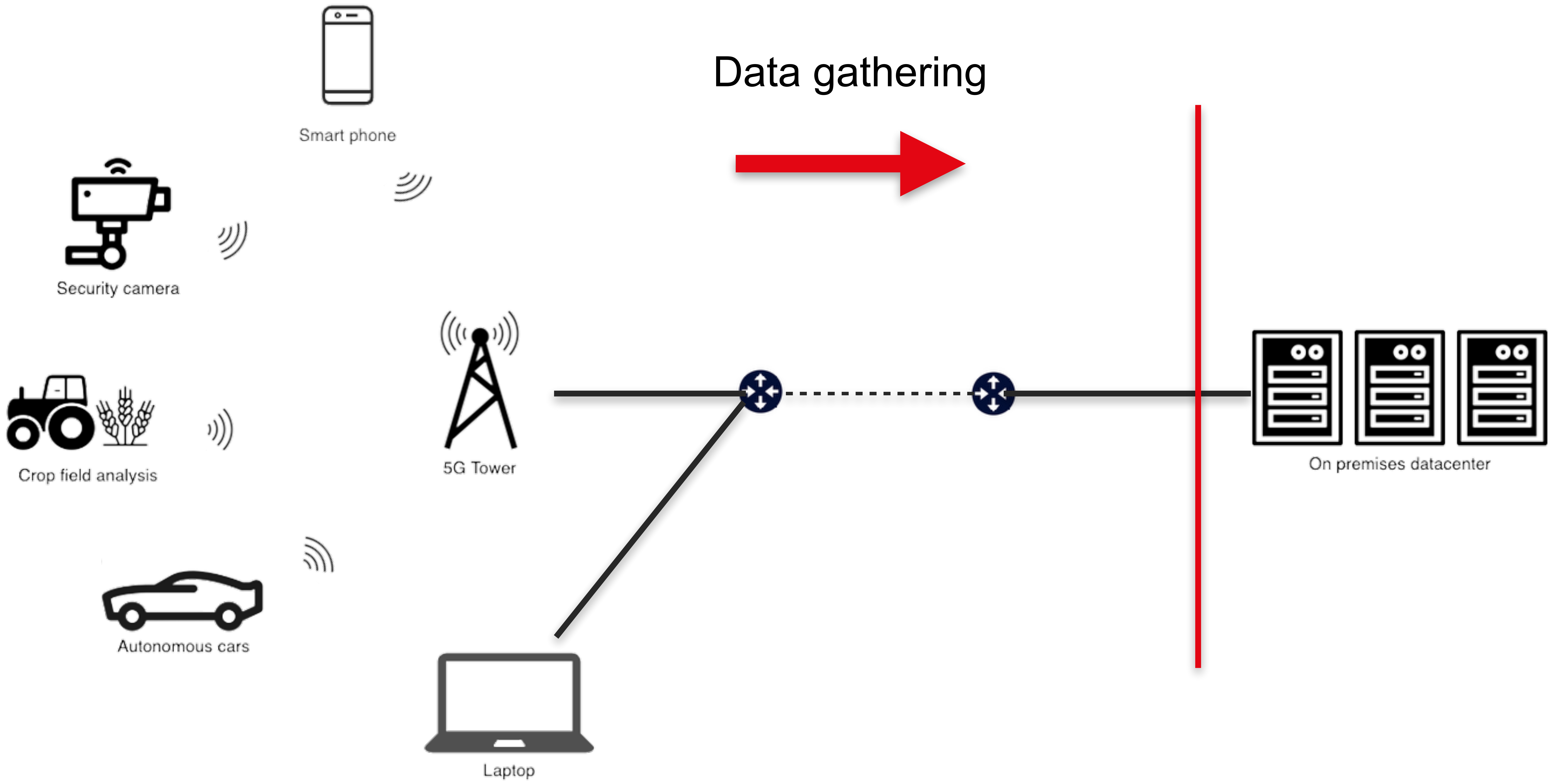


ML development life cycle



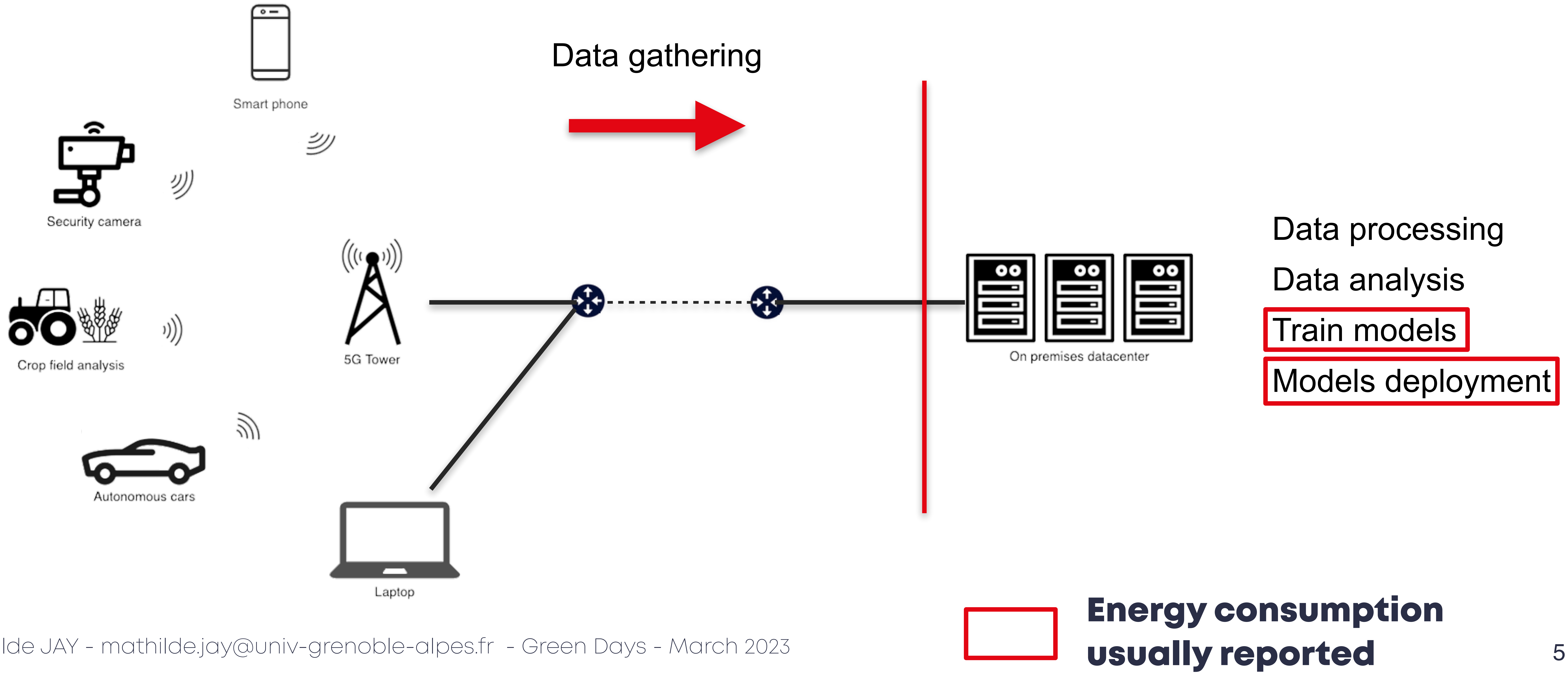
Energy consumption usually reported

ML infrastructures



Data processing
Data analysis
Train models
Models deployment

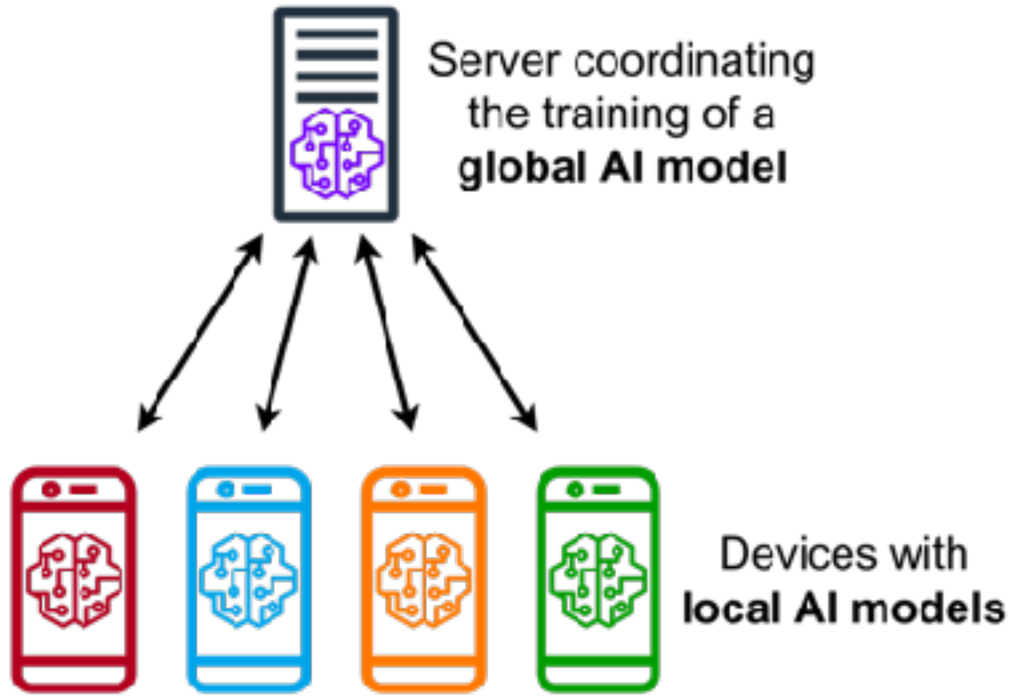
ML infrastructures



Other ML paradigm

Federated Learning

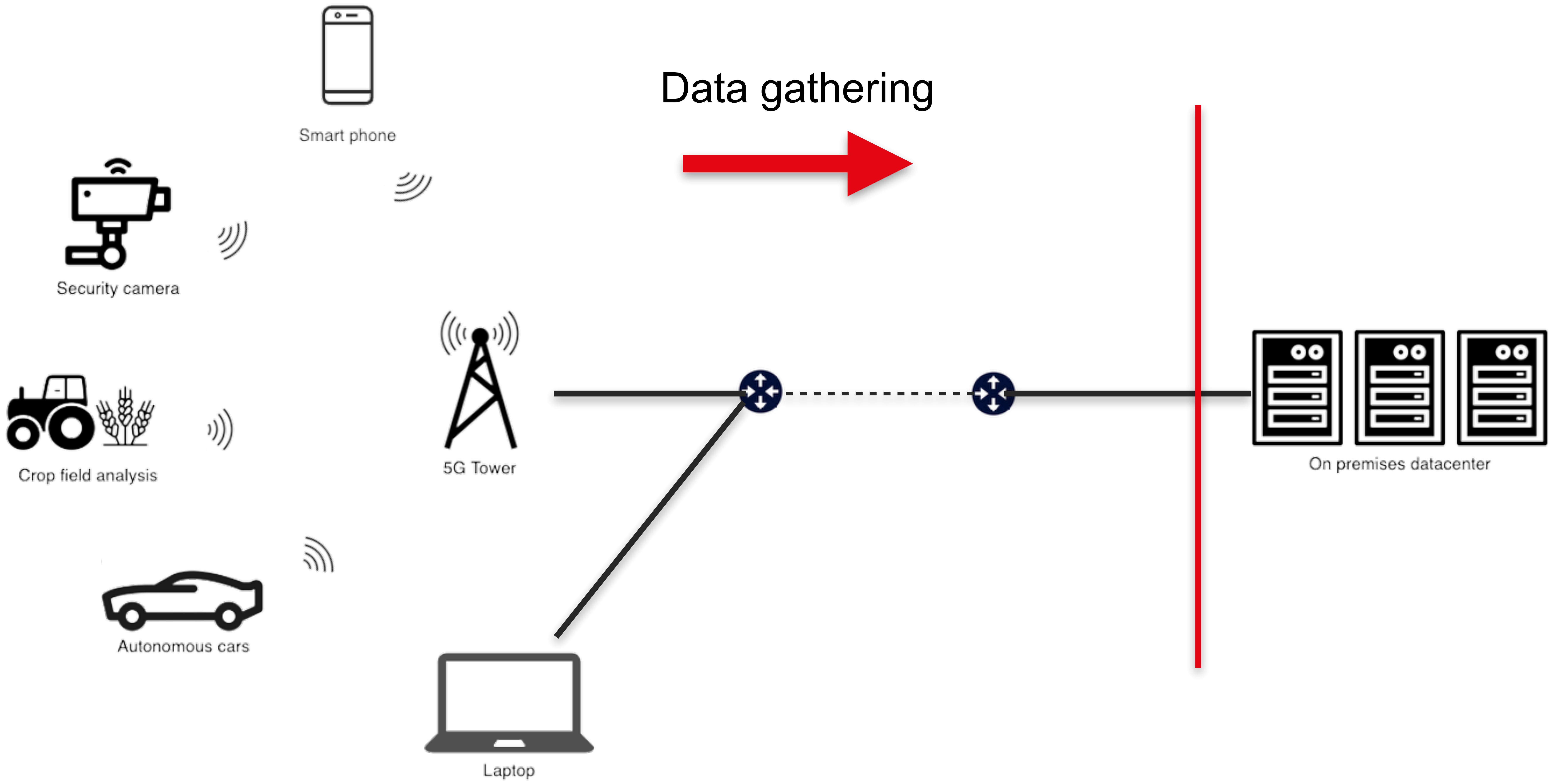
- Learning on a selection of devices
- Aggregation on server
- Goal: data stays in devices
- Challenges: communication, bias



Continuous Learning

- Repetition at a given frequency of
  - Learning
  - Data gathering

Continuous learning infrastructures



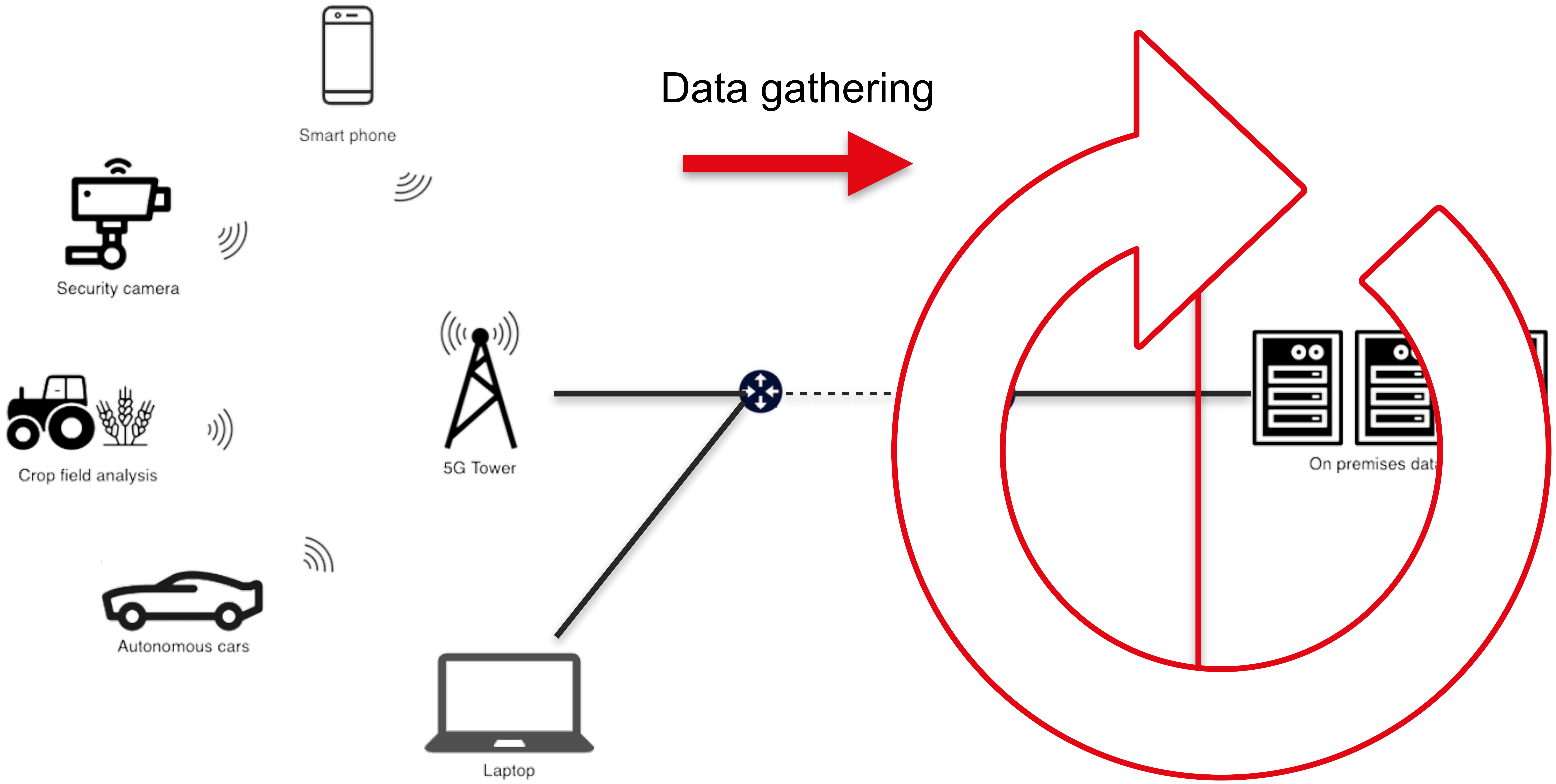
Data processing

Data analysis

Train models

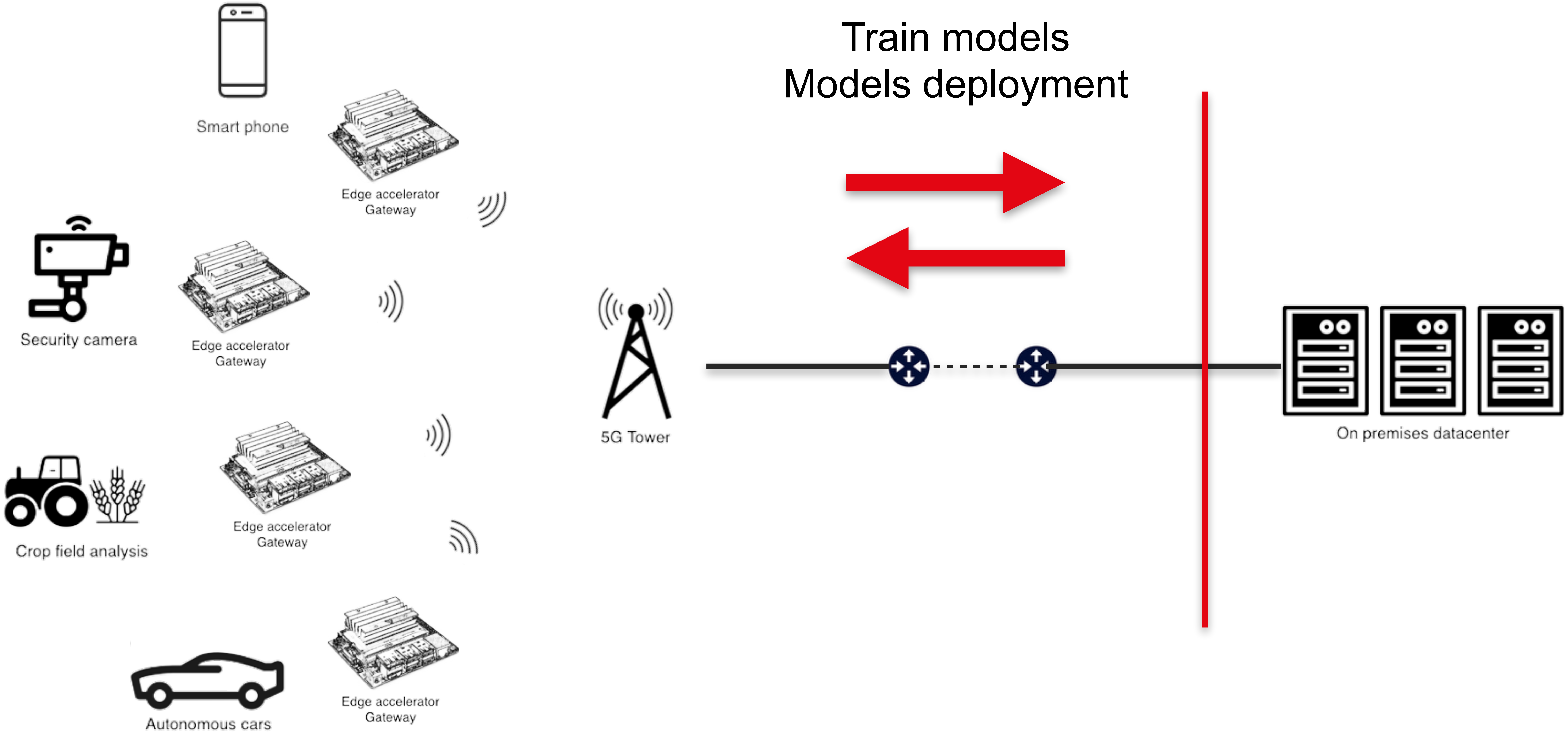
Models deployment

Continuous learning infrastructures



Data processing
Data analysis
Train models
Models deployment

Federated Learning infrastructures



Data processing

Data analysis

Model averaging

Paradigm with lower energy footprint?

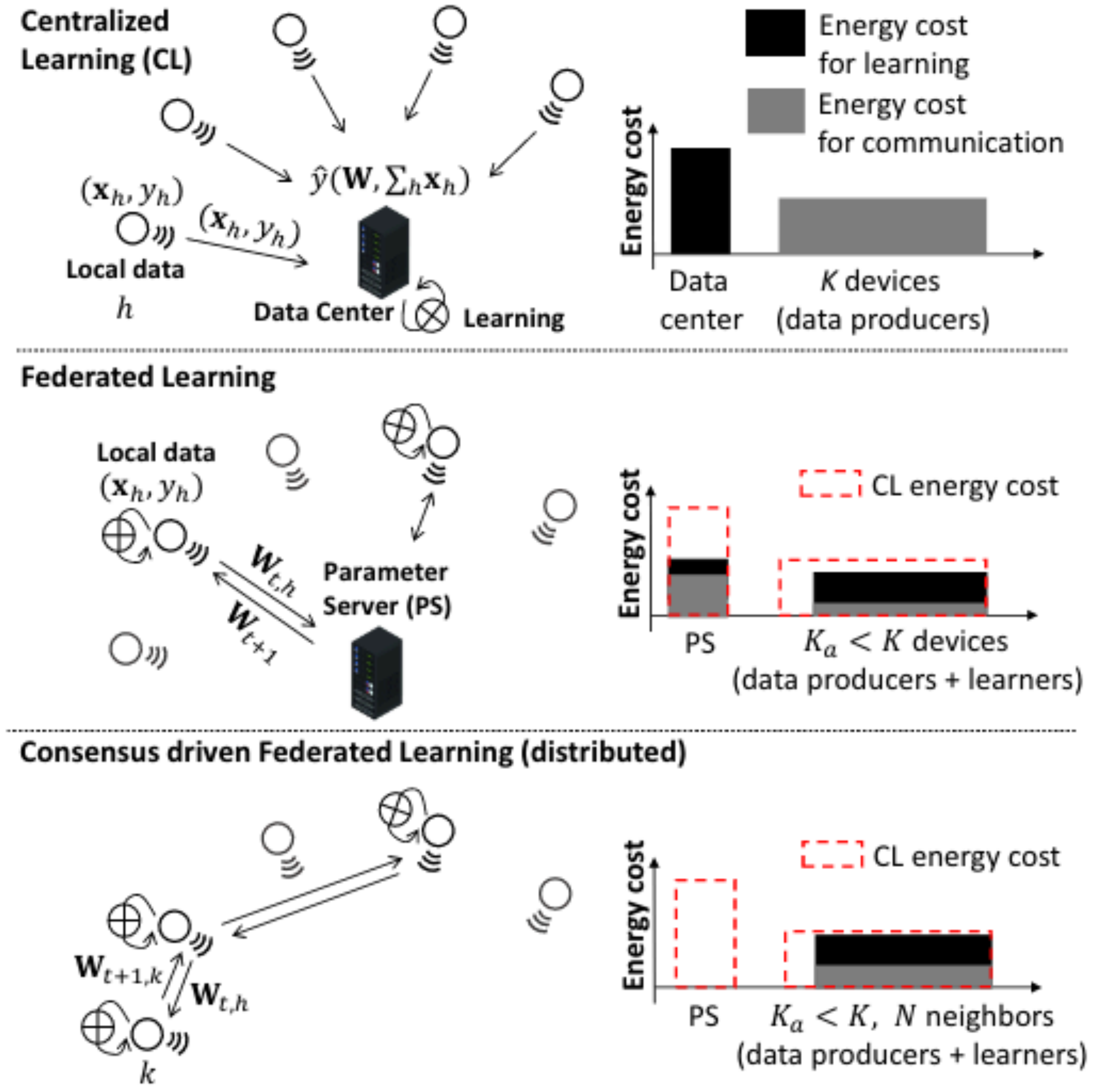
	Edge devices	Data centers	FL
Latency	None	High	Low
Privacy	High	Low	High
Data transfer	None	High	Low
Power efficiency	High	Low	??
Computation efficiency	Very low	High	??
Energy	**??**	**??**	**??**

Choosing the paradigm with a low energy footprint

- Existing work on comparison of the energy footprint of federated learning and centralized learning

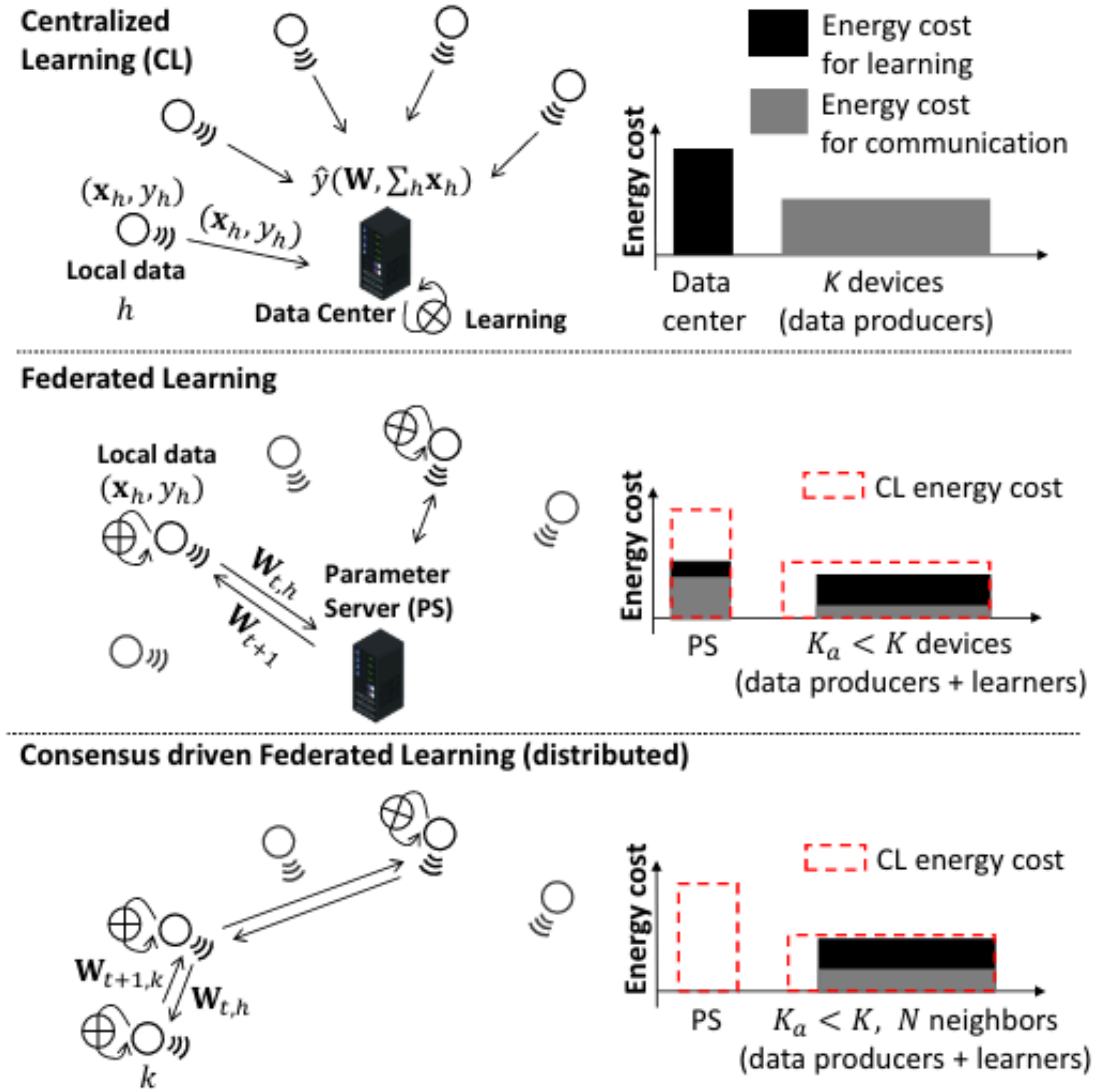
- My objectives

S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

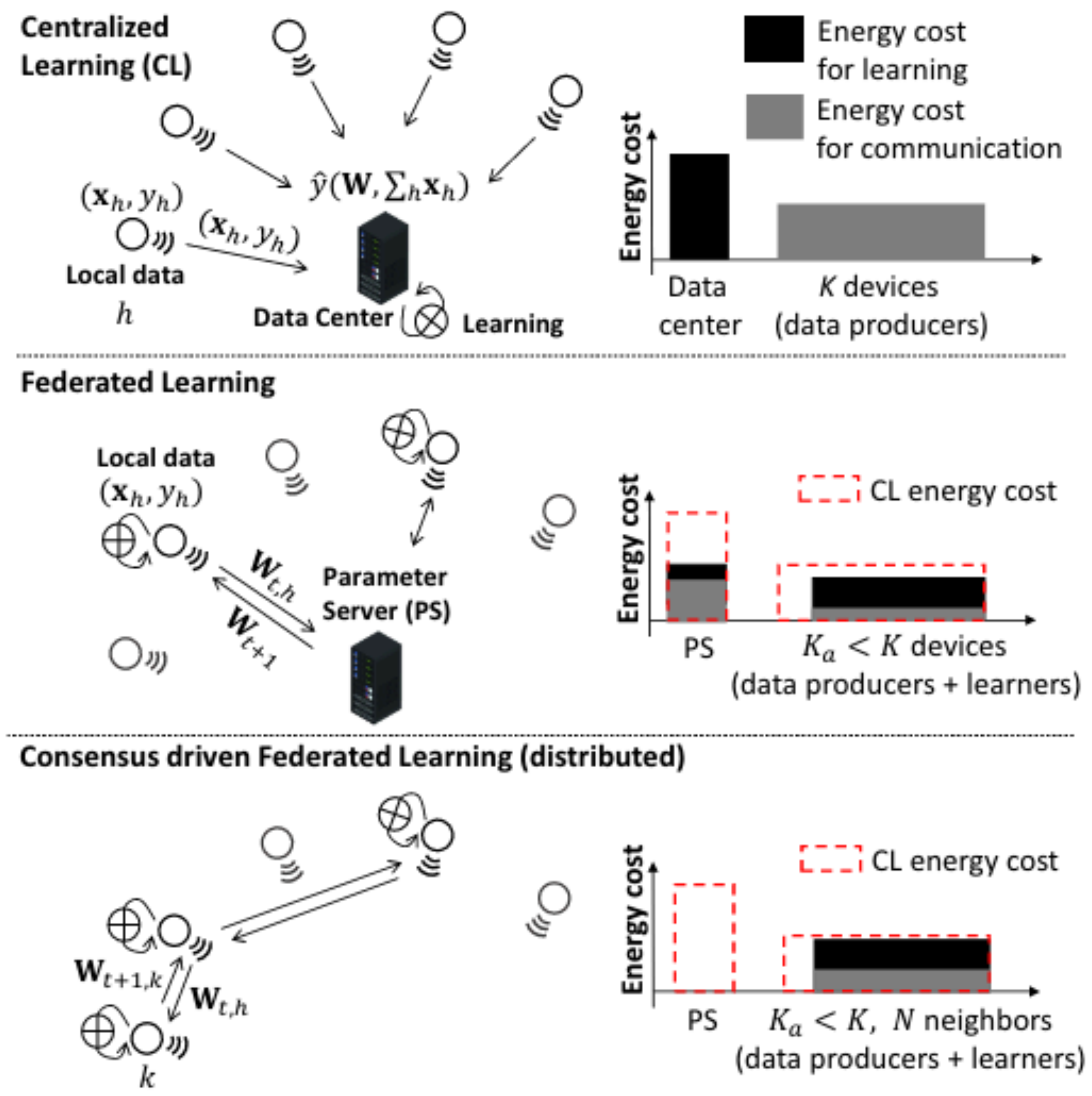
Energy consumption **simulator** from



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

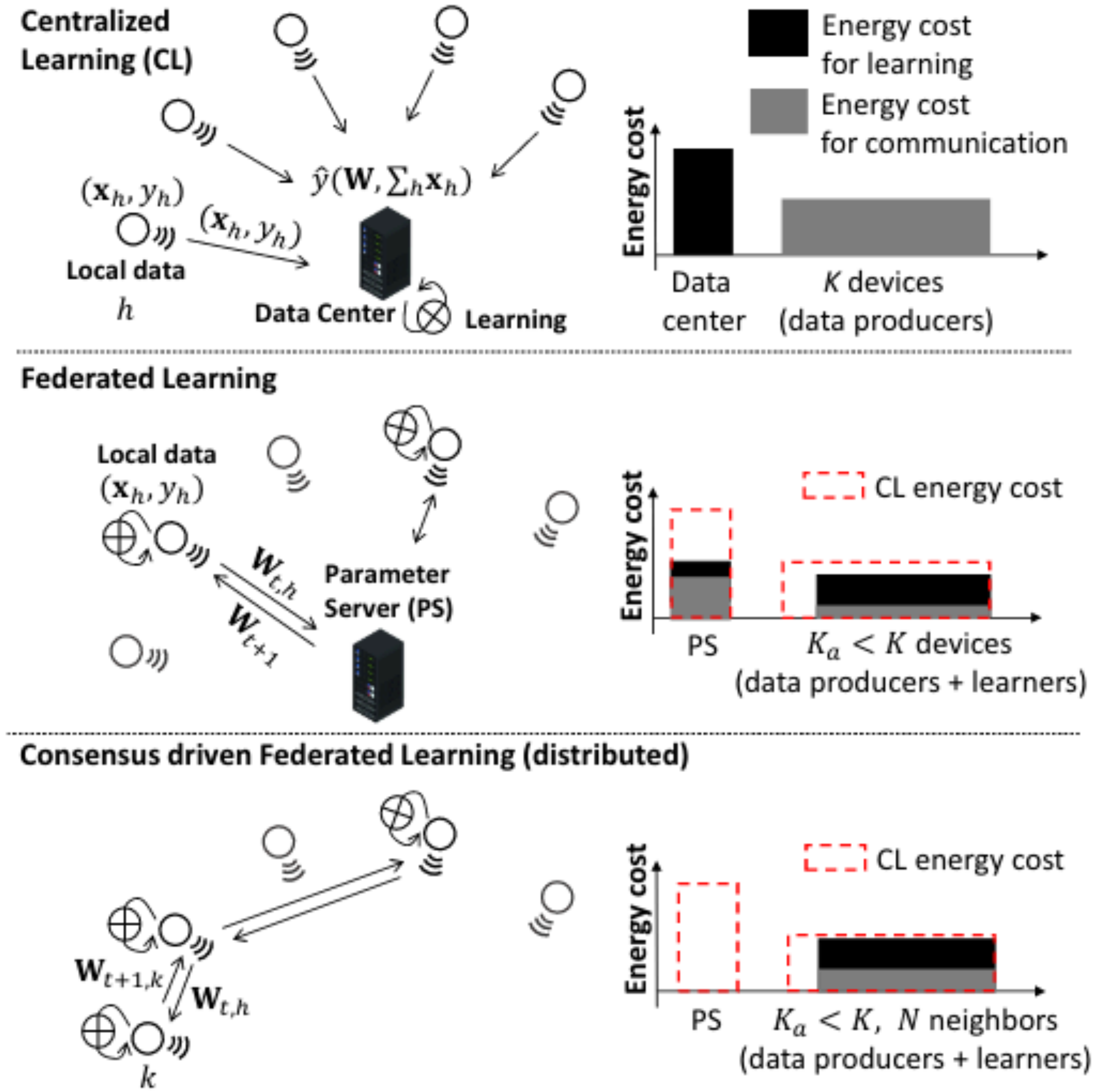
- PUE



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

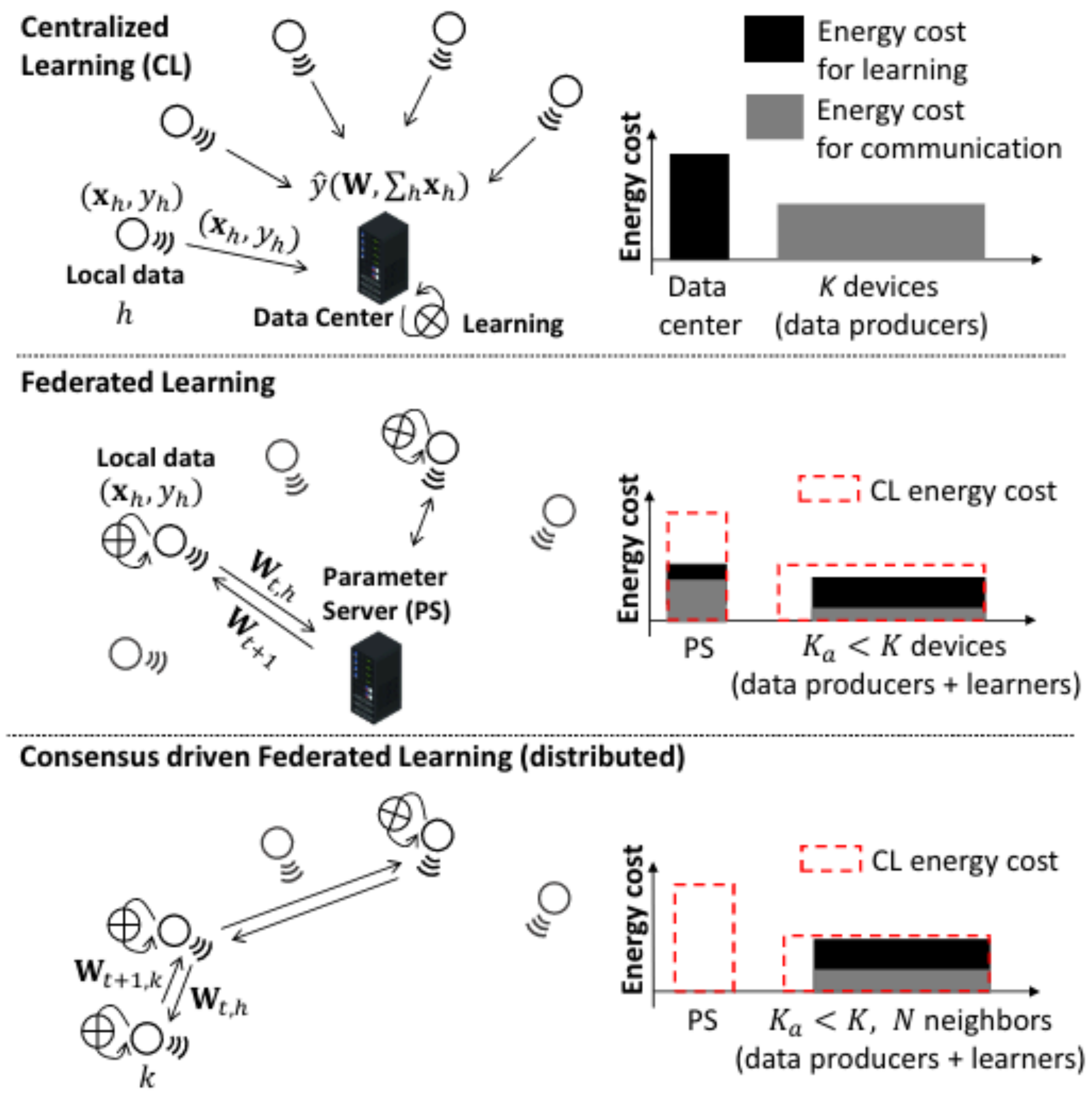
- PUE
- Number of rounds to reach target accuracy (and number of batches)



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

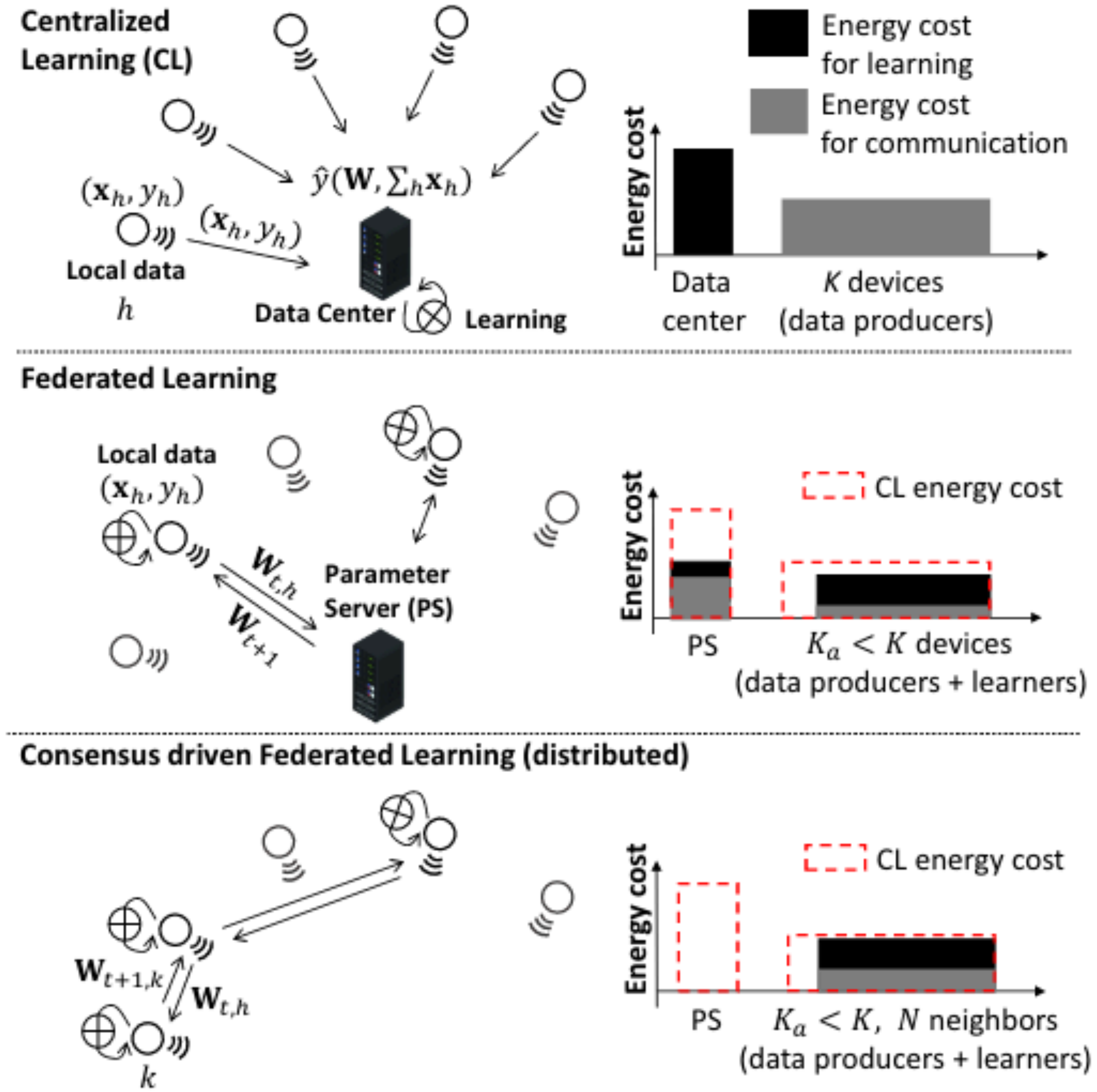
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

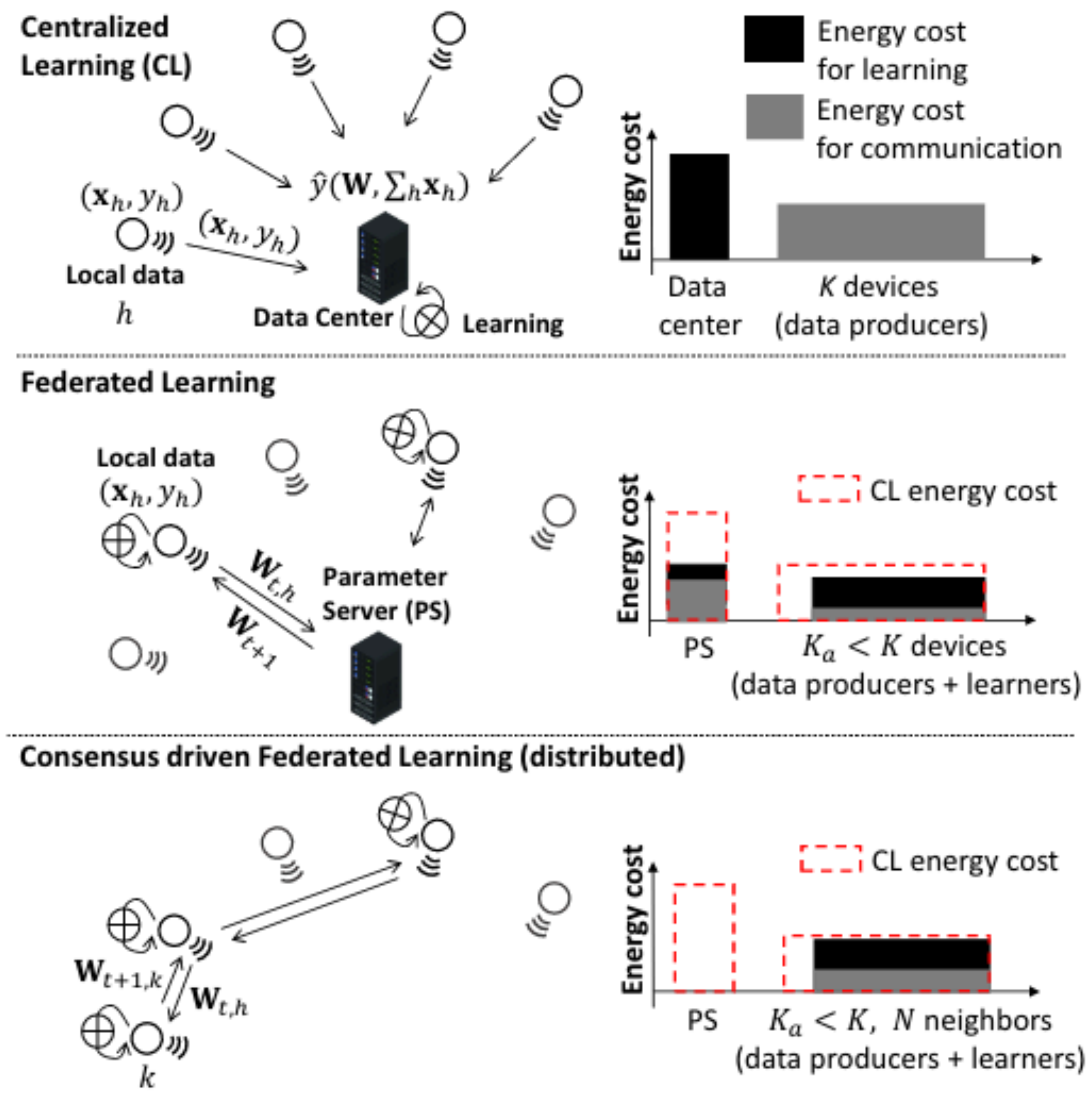
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, "An Energy and Carbon Footprint Analysis of Distributed and Federated Learning," IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

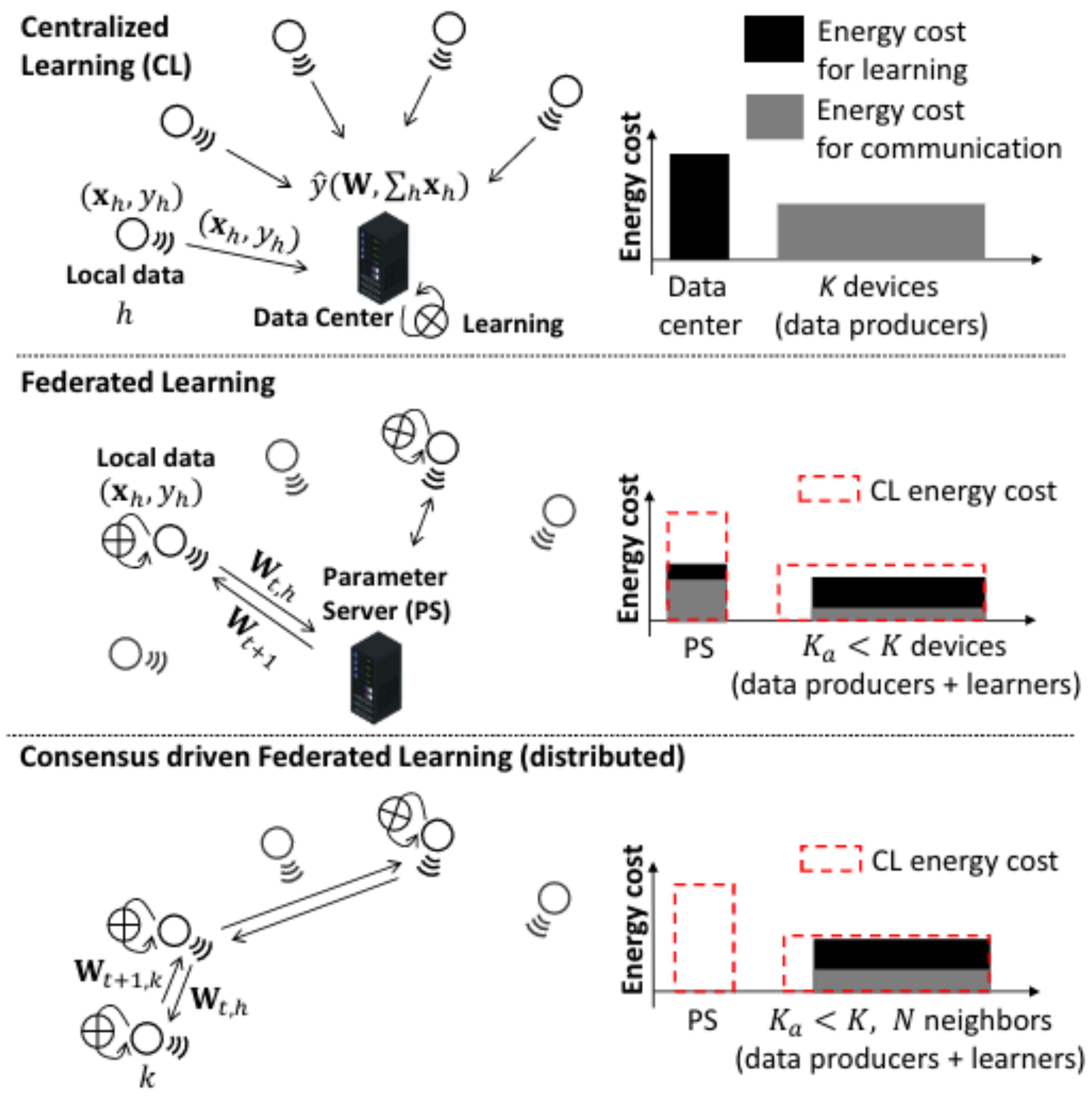
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

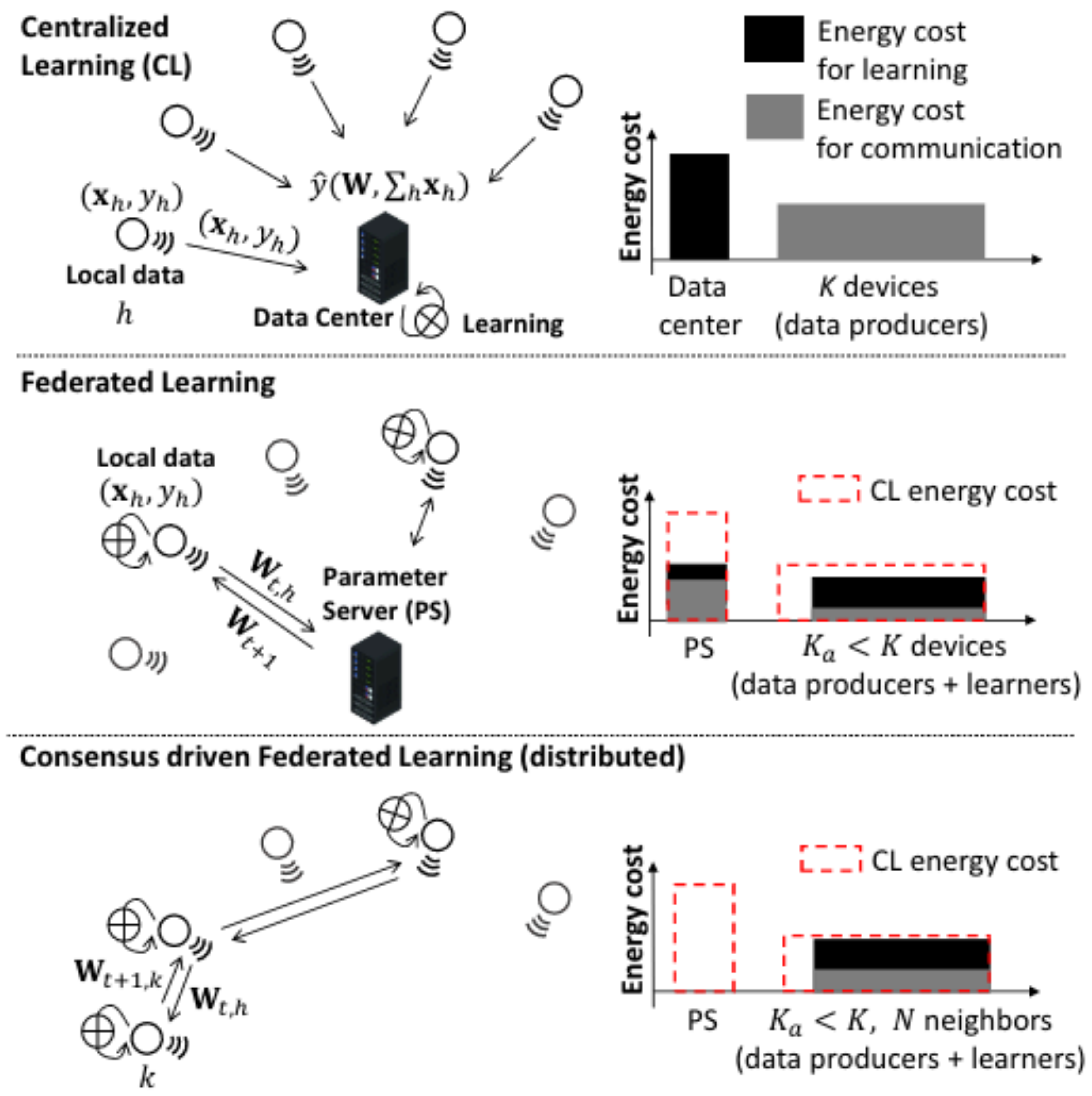
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

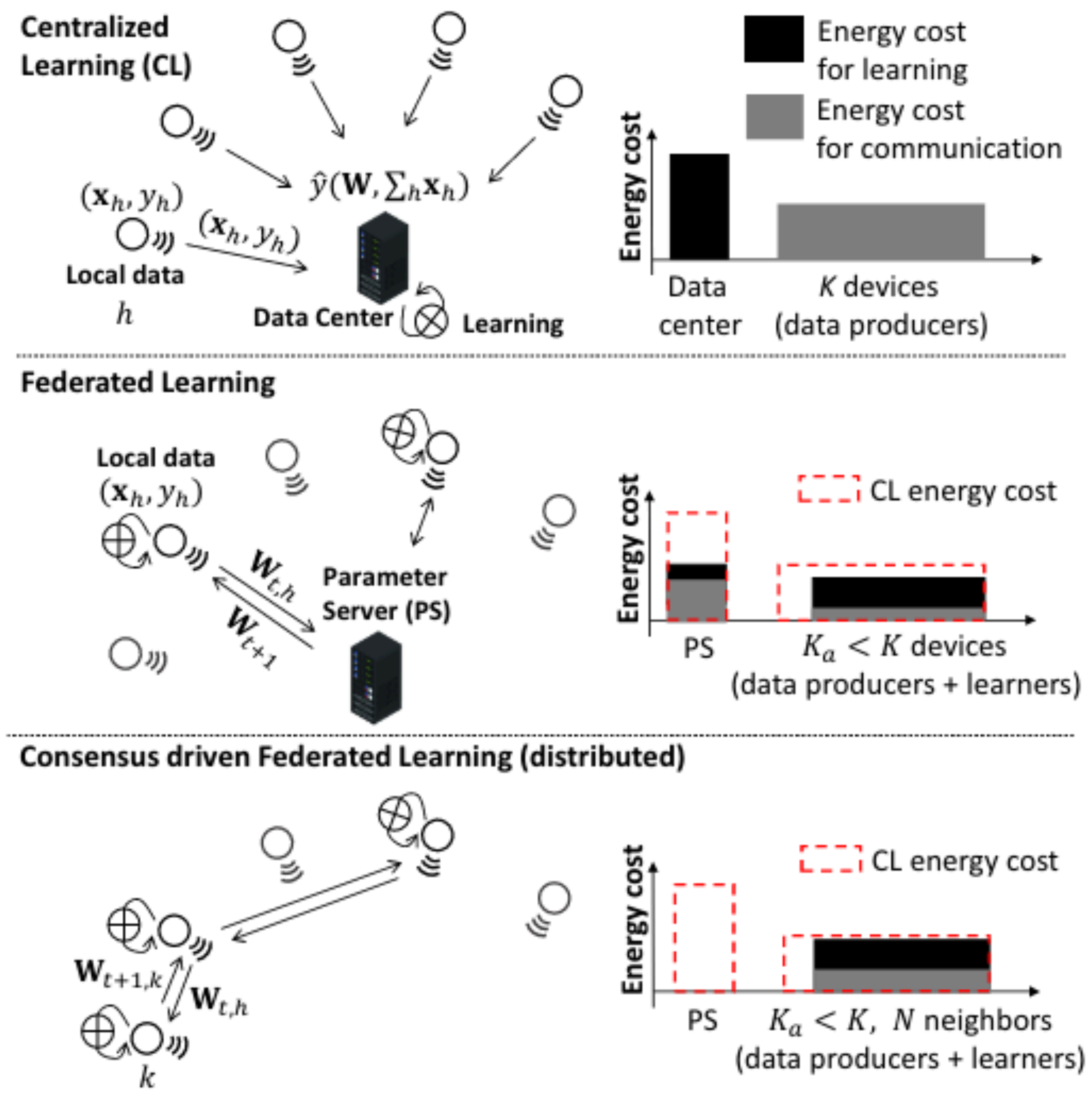
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

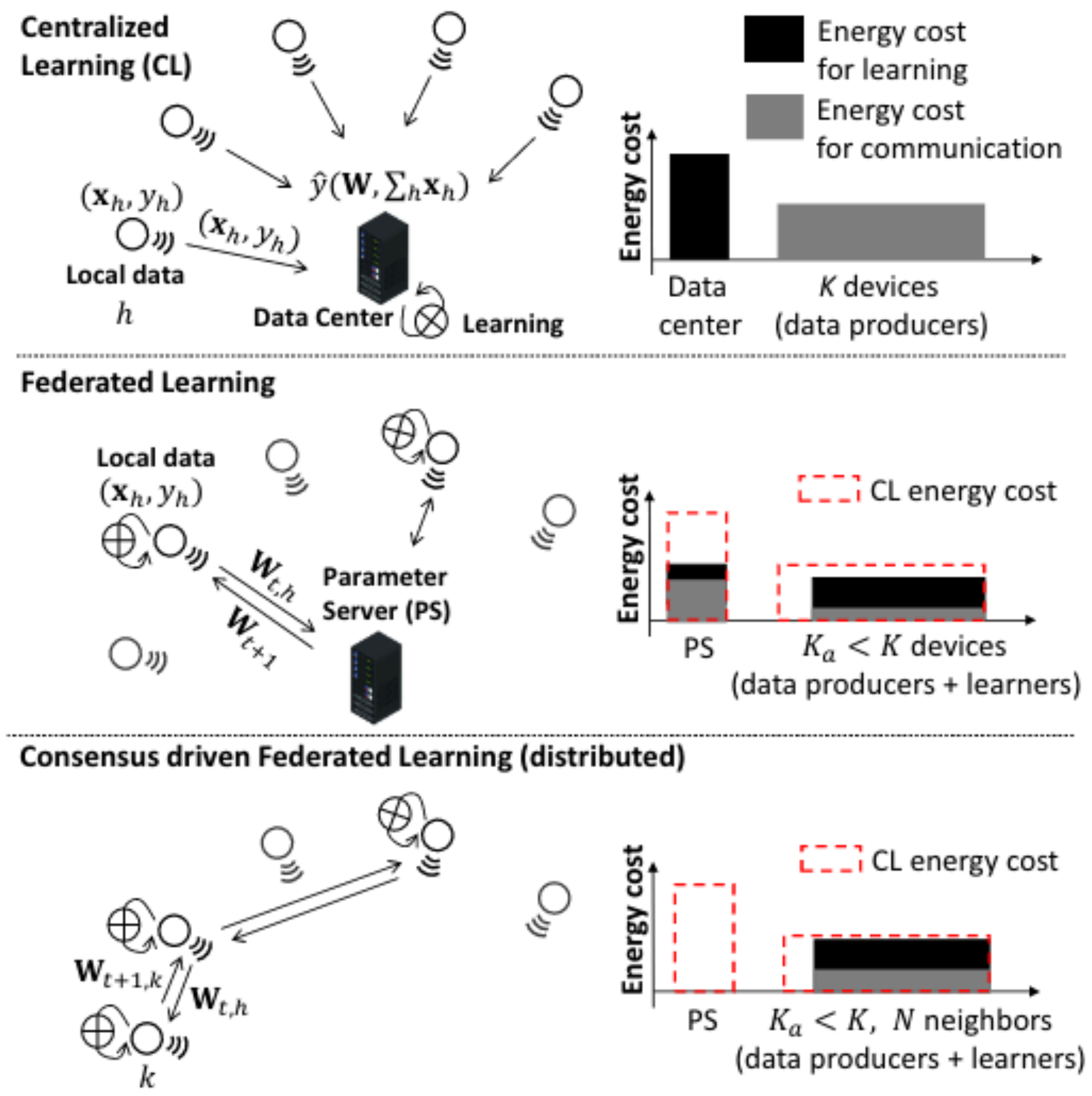
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners
- Relative energy efficiency



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, "An Energy and Carbon Footprint Analysis of Distributed and Federated Learning," IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

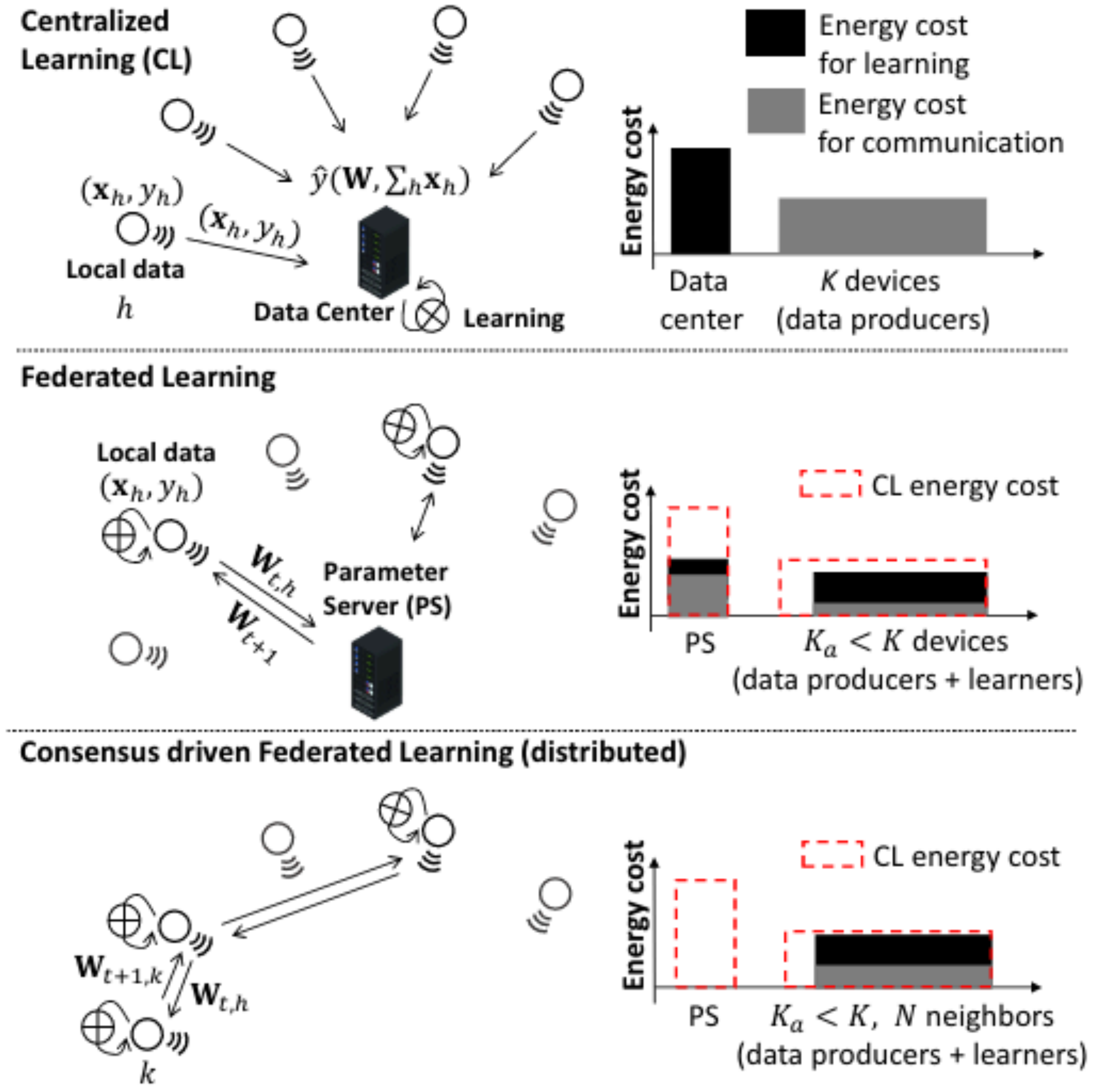
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners
- Relative energy efficiency
- Type of data transfer (uplink, downlink)



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners
- Relative energy efficiency
- Type of data transfer (uplink, downlink)

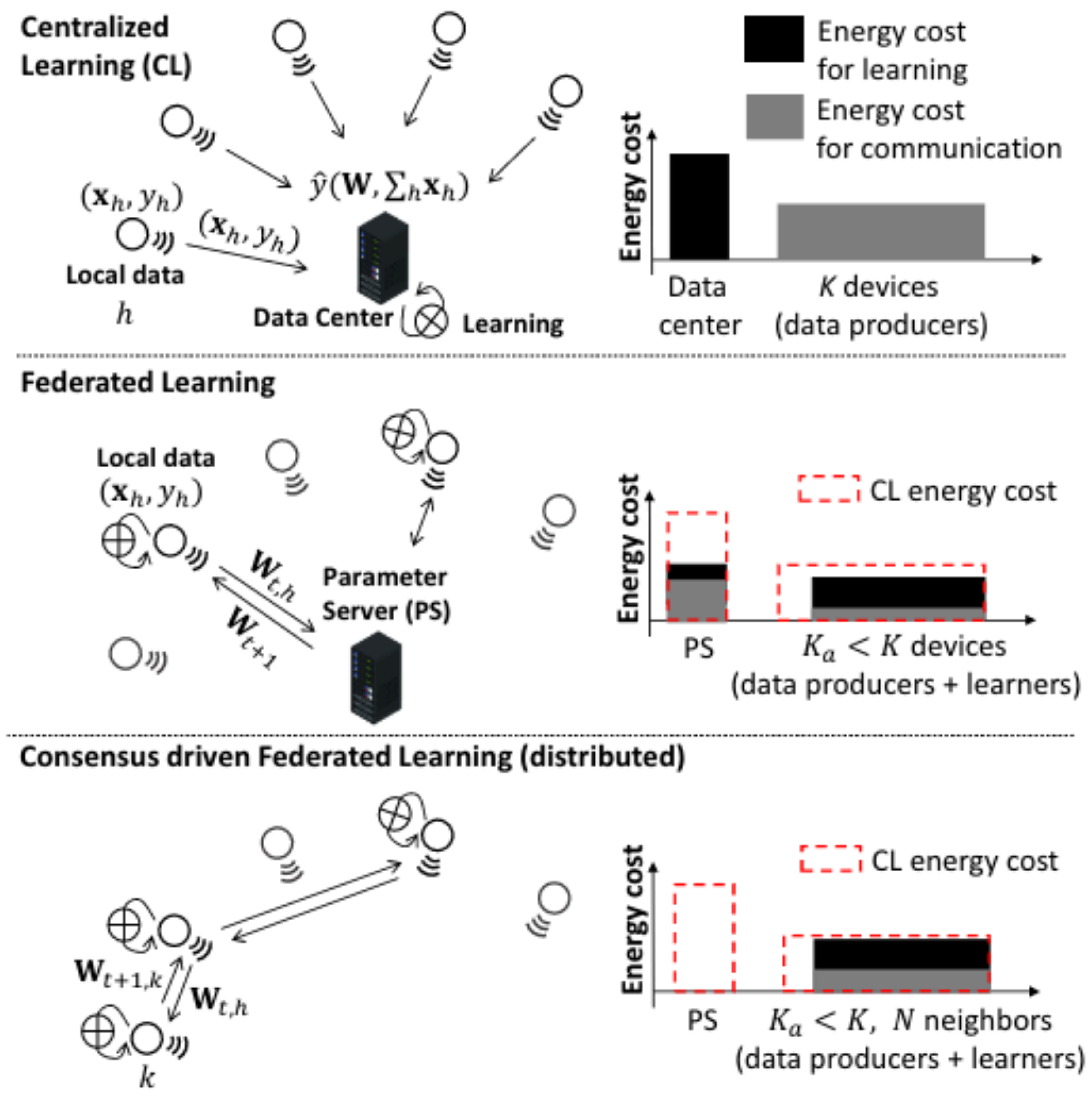


S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

Energy consumption **simulator** from

- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners
- Relative energy efficiency
- Type of data transfer (uplink, downlink)

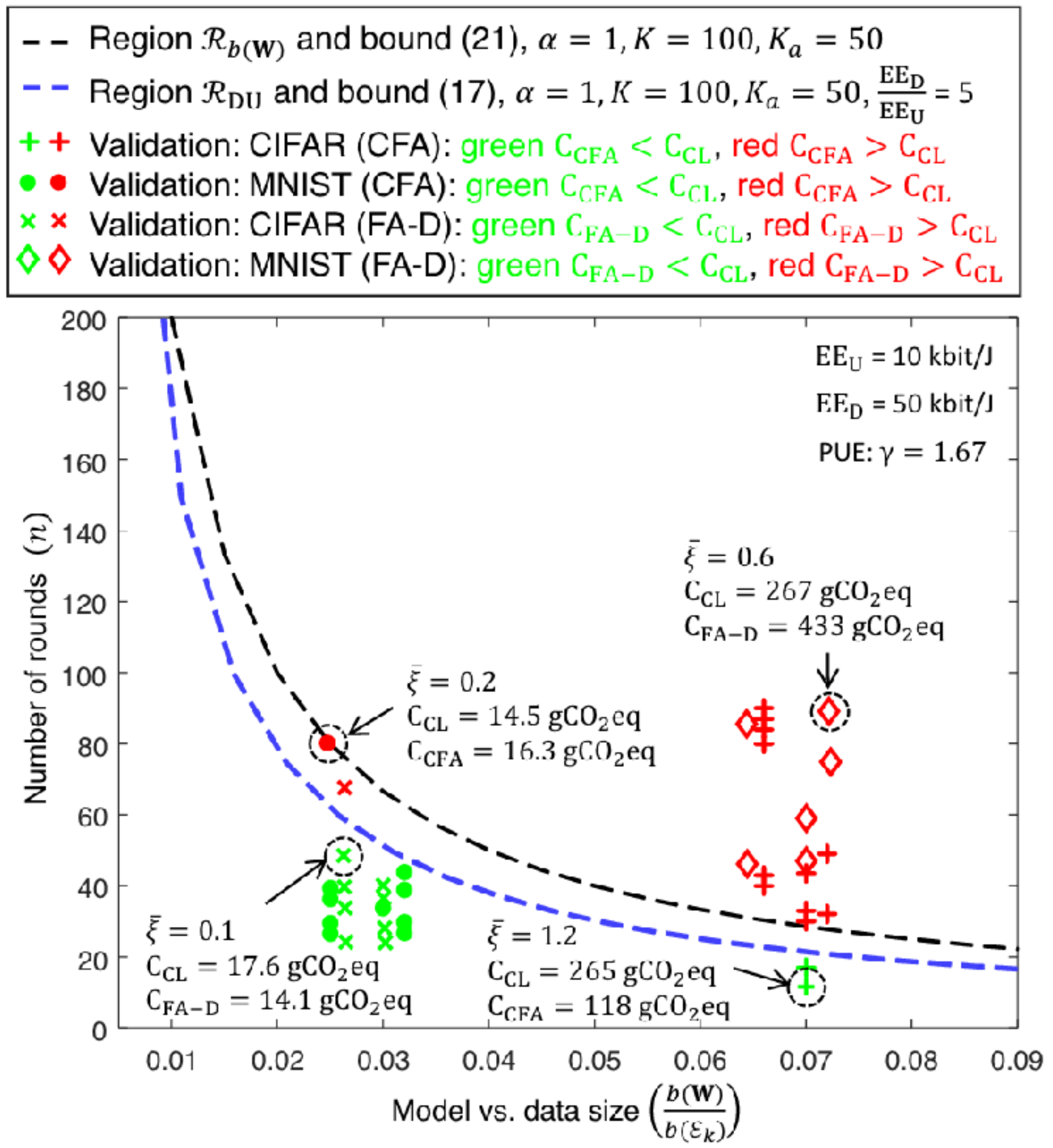
Rules for decision on which paradigm to use



S. Savazzi, V. Rampa, S. Kianoush, and M. Bennis, “An Energy and Carbon Footprint Analysis of Distributed and Federated Learning,” IEEE Transactions on Green Communications and Networking, pp. 1–1, 2022.

The co-design of learning and communication is of high importance.

- Incomplete sensitivity analysis
  - PUE
  - Computing efficiency
  - Computing power
- Computer vision models only

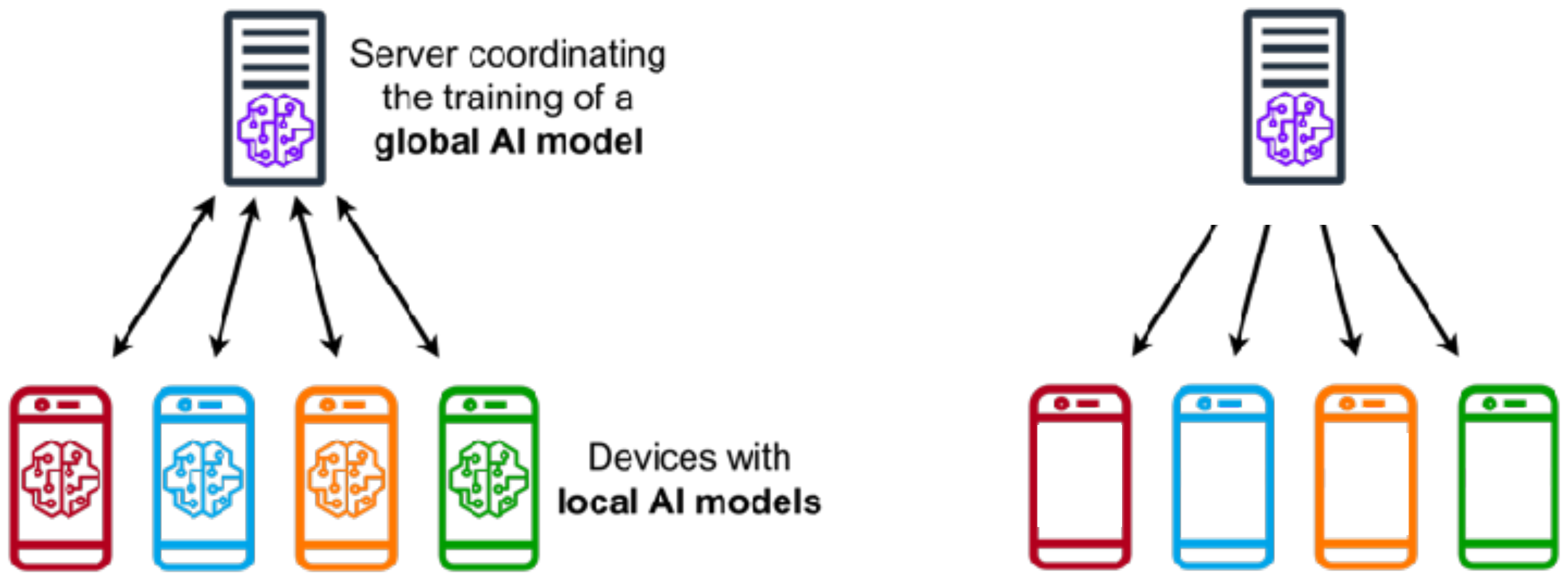


My objectives

Benchmarking the performance and energy efficiency of AI accelerators for AI training

My objectives

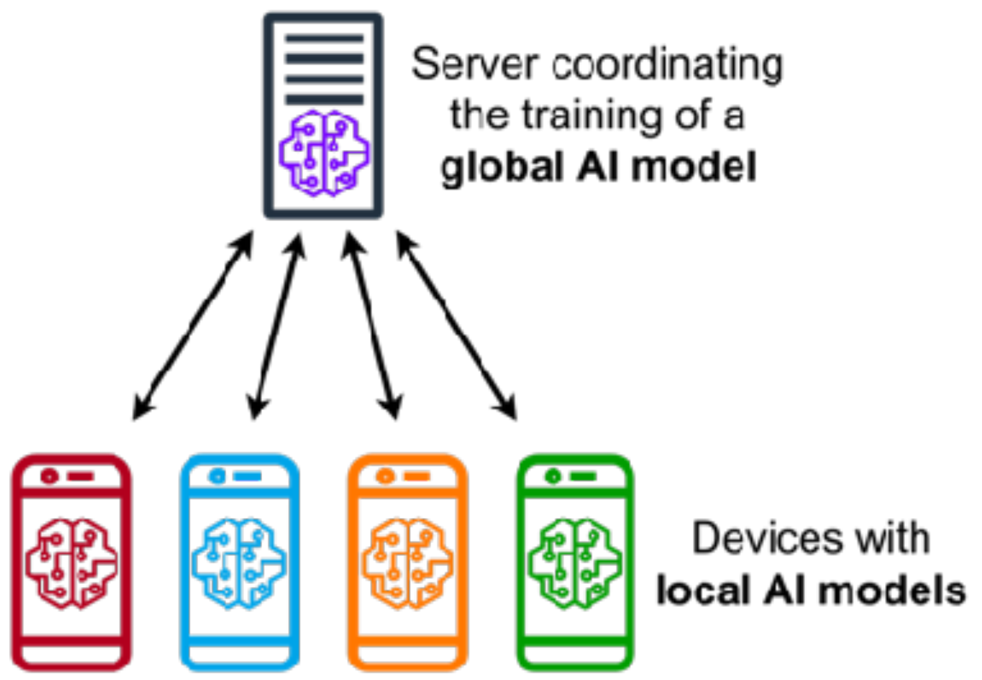
Benchmarking the performance and energy efficiency of AI accelerators for AI training

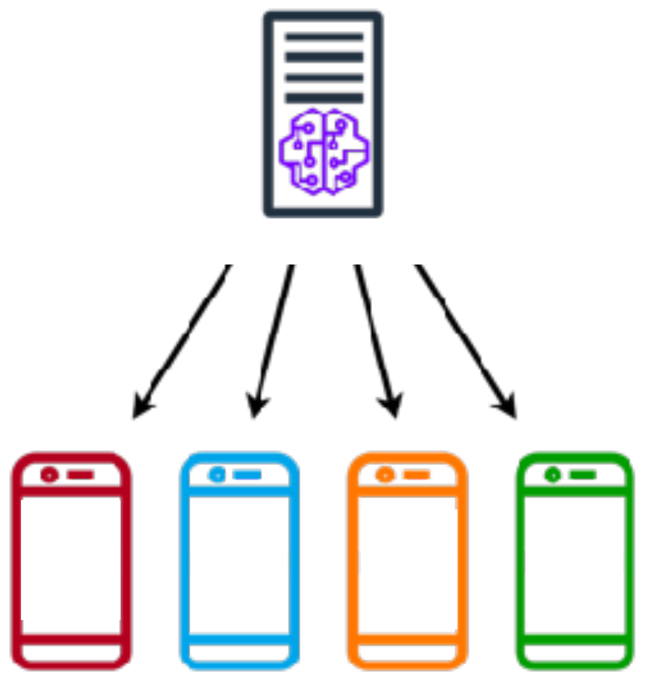


Federated Learning versus Centralized Learning

My objectives

Benchmarking the performance and energy efficiency of AI accelerators for AI training



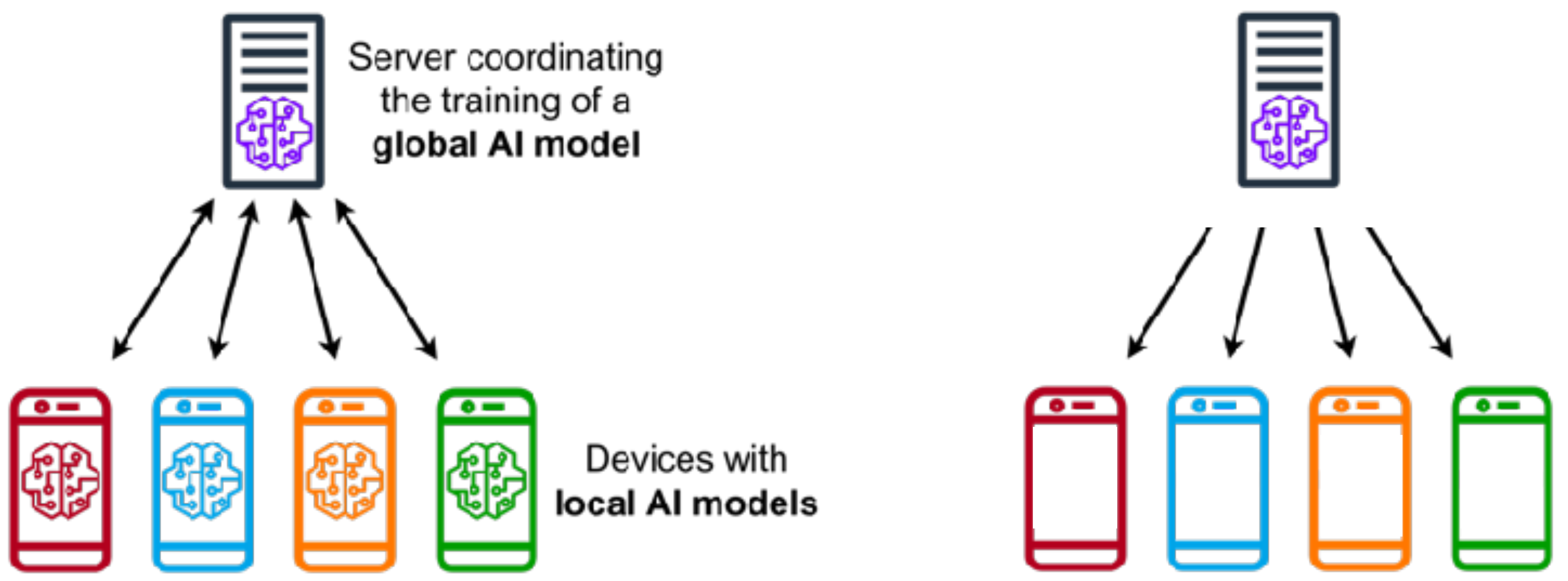


Federated Learning **versus** **Centralized Learning**

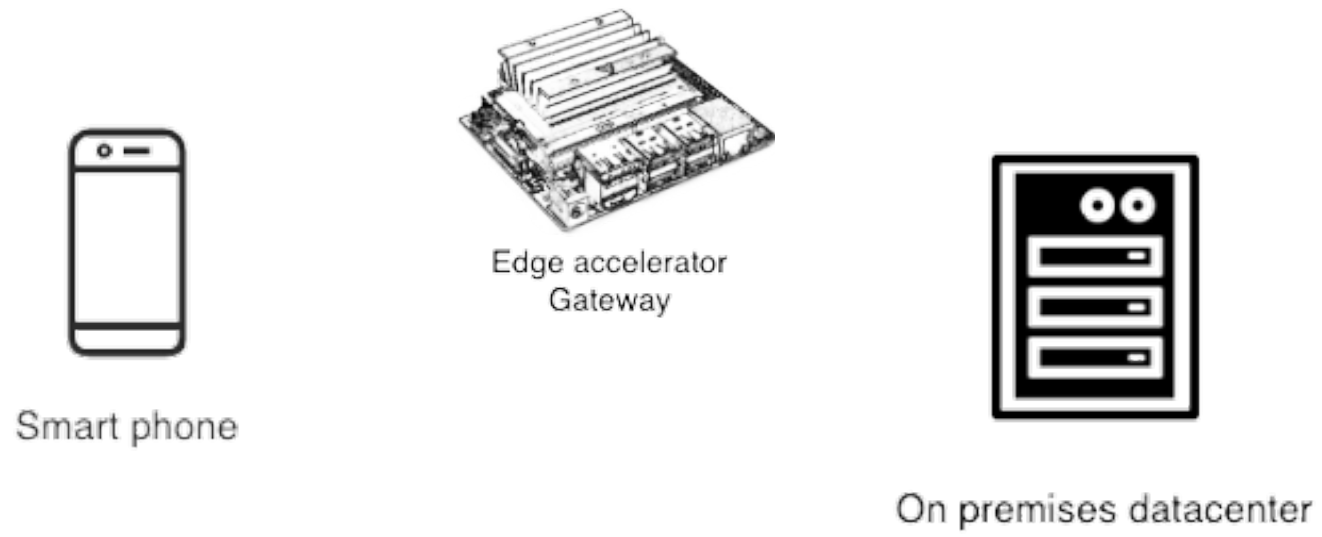
Continuous settings

My objectives

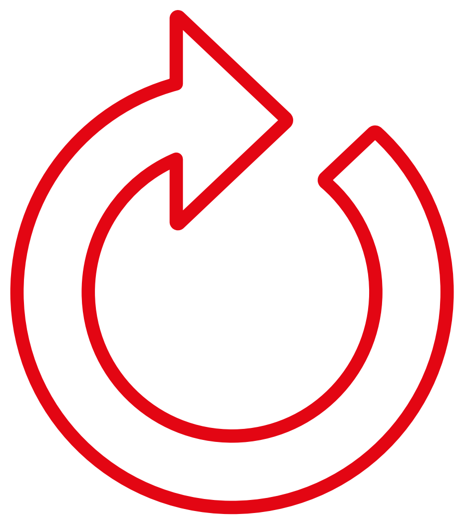
Benchmarking the performance and energy efficiency of AI accelerators for AI training



Federated Learning **versus** **Centralized Learning**



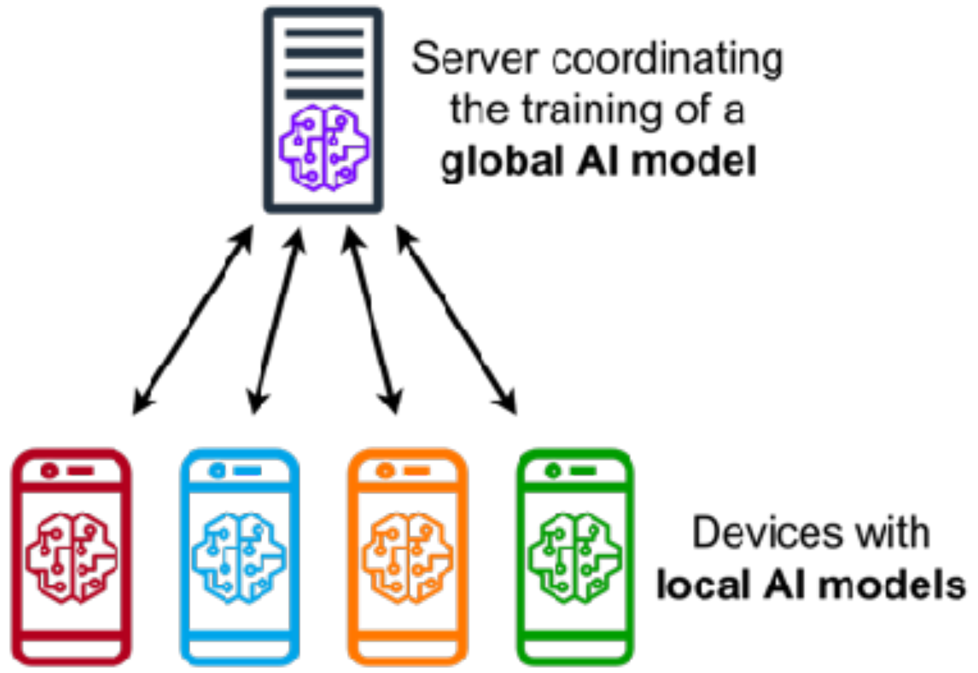
Rules on computer efficiency

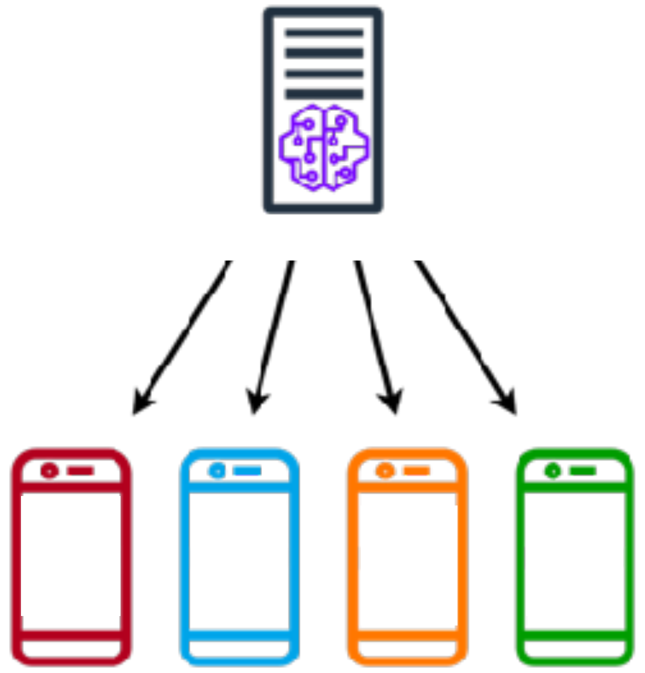


Continuous settings

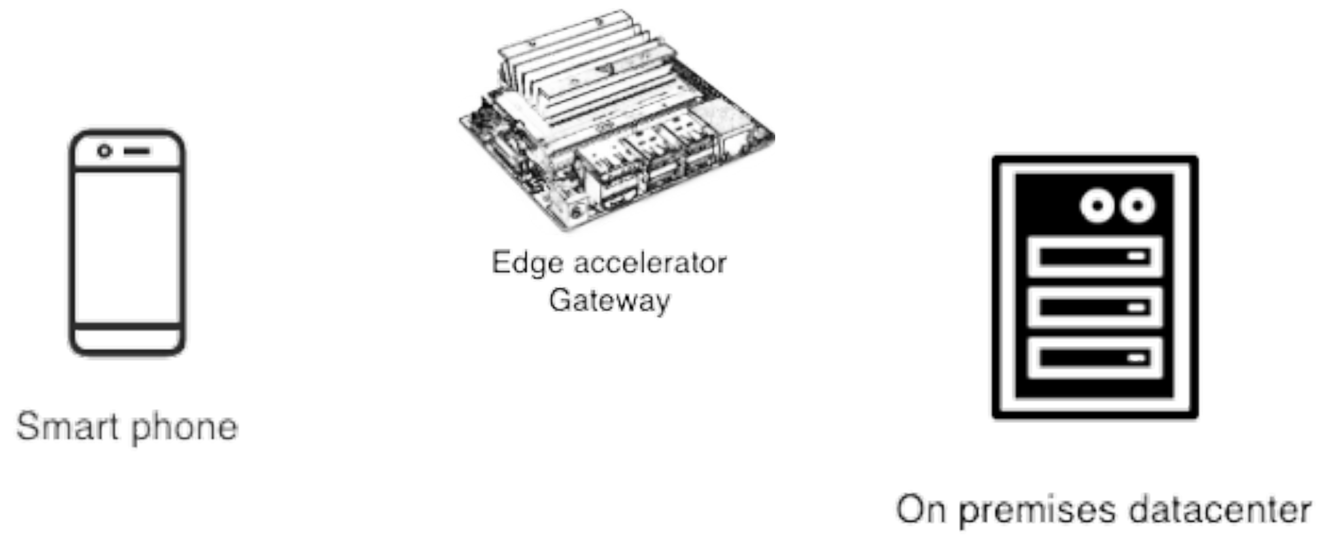
My objectives

Benchmarking the performance and energy efficiency of AI accelerators for AI training

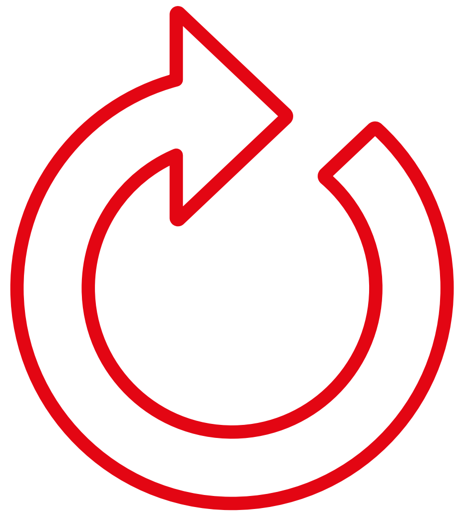




Federated Learning **versus** **Centralized Learning**



Rules on computer efficiency



Continuous settings

Concretely

- Experiments
  - Training until accuracy is reached on various machines
  - Energy tracked from both hardware and software-based power meters
- Simulations: add impact of
  - The whole infrastructure
  - The complete life cycle
- Models included in the study
  - Image: Medical image segmentation
  - NLP: Transformers
  - Generative AI: StableDiffusion (TBD)
- To study: impact on energy of
  - Machine efficiency (computations, memory)
  - Database size
  - Size and type of models

Champollion (HPE)
8 GPU Nvidia A100 SXM4 (80Go)

Nvidia Jetson AGX Xavier (32Go)

Grid'5000

Coral Dev Board (1Go)

Thank you for listening :)

Any feedback is welcome!