Machine learning energy consumption evaluation methodology

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

Denis Trystram - Prof. Ensimag - LIG, Inria DataMove
Laurent Lefevre - CR Inria - LIP, Inria Avalon

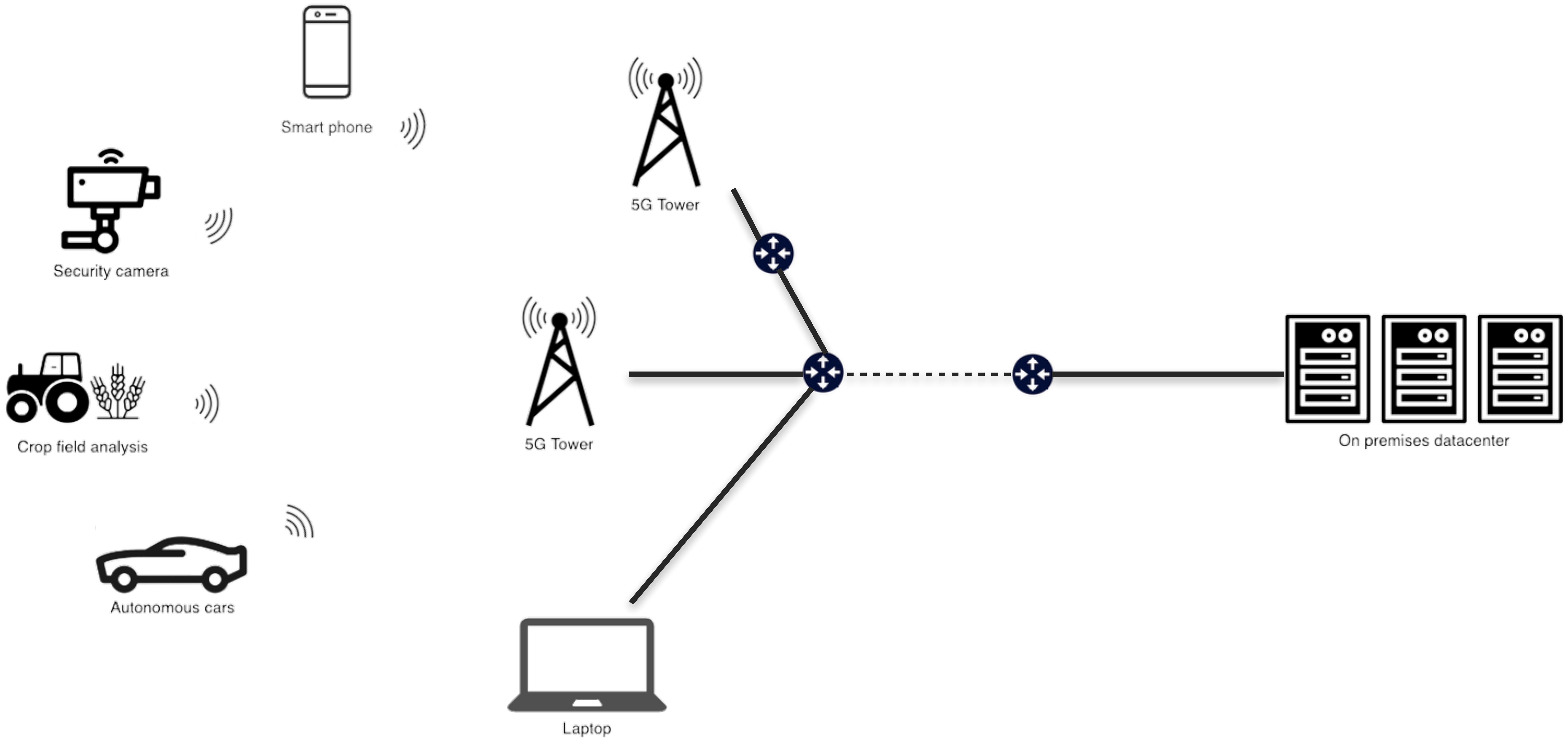
Université Grenoble Alpes

MIAI Grenoble Alpes

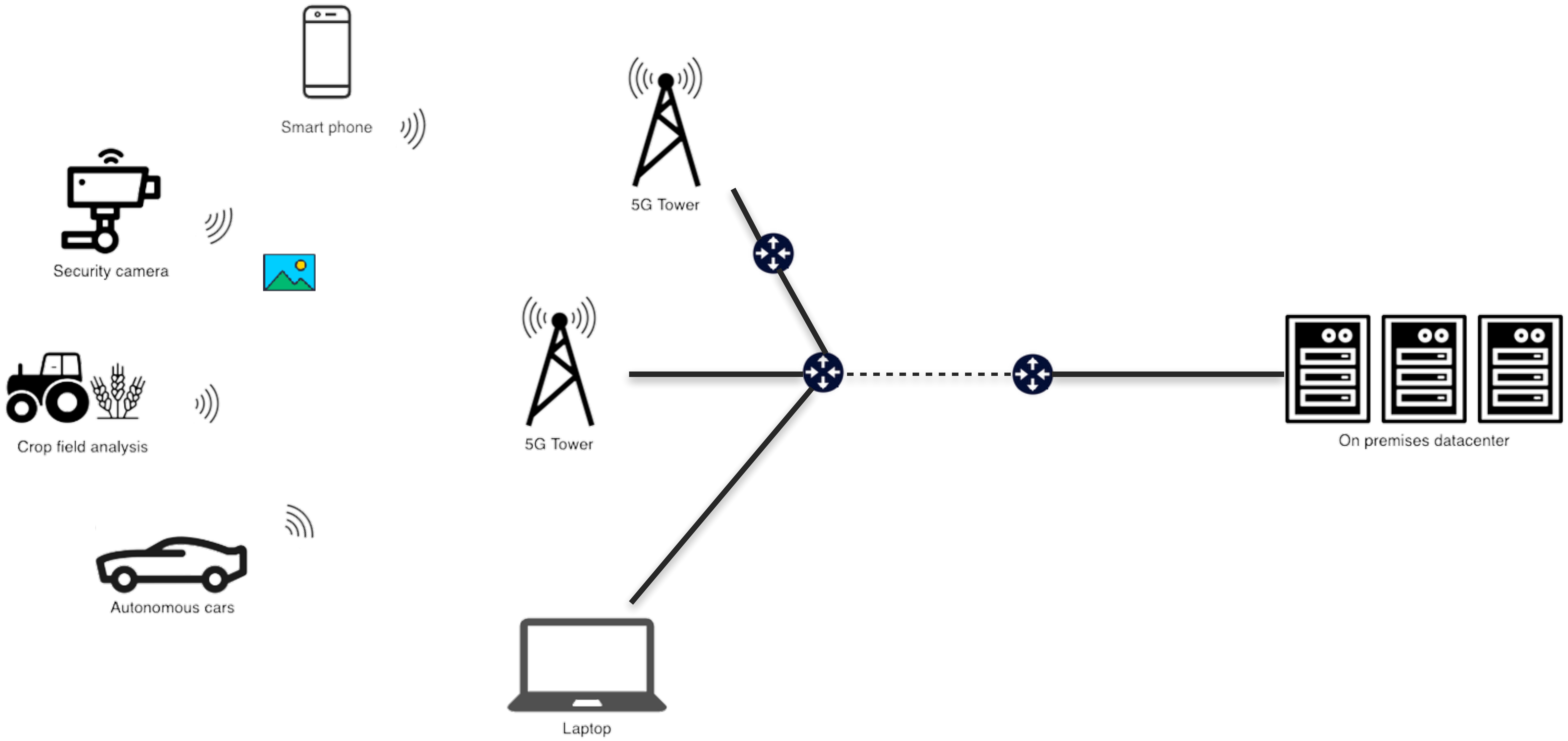
La Région Auvergne-Rhône-Alpes

Green Days 2023

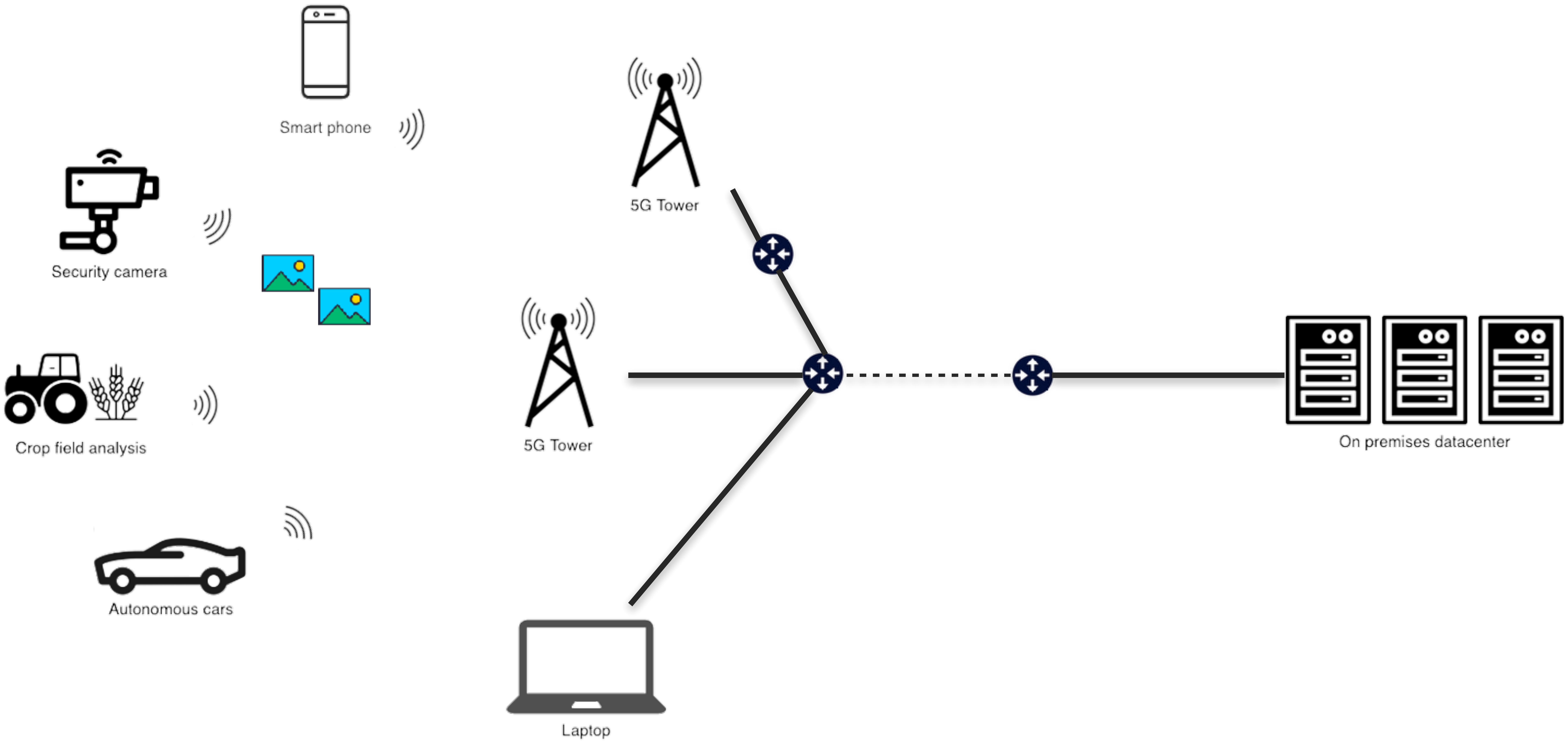
Context: Infrastructures behind AI



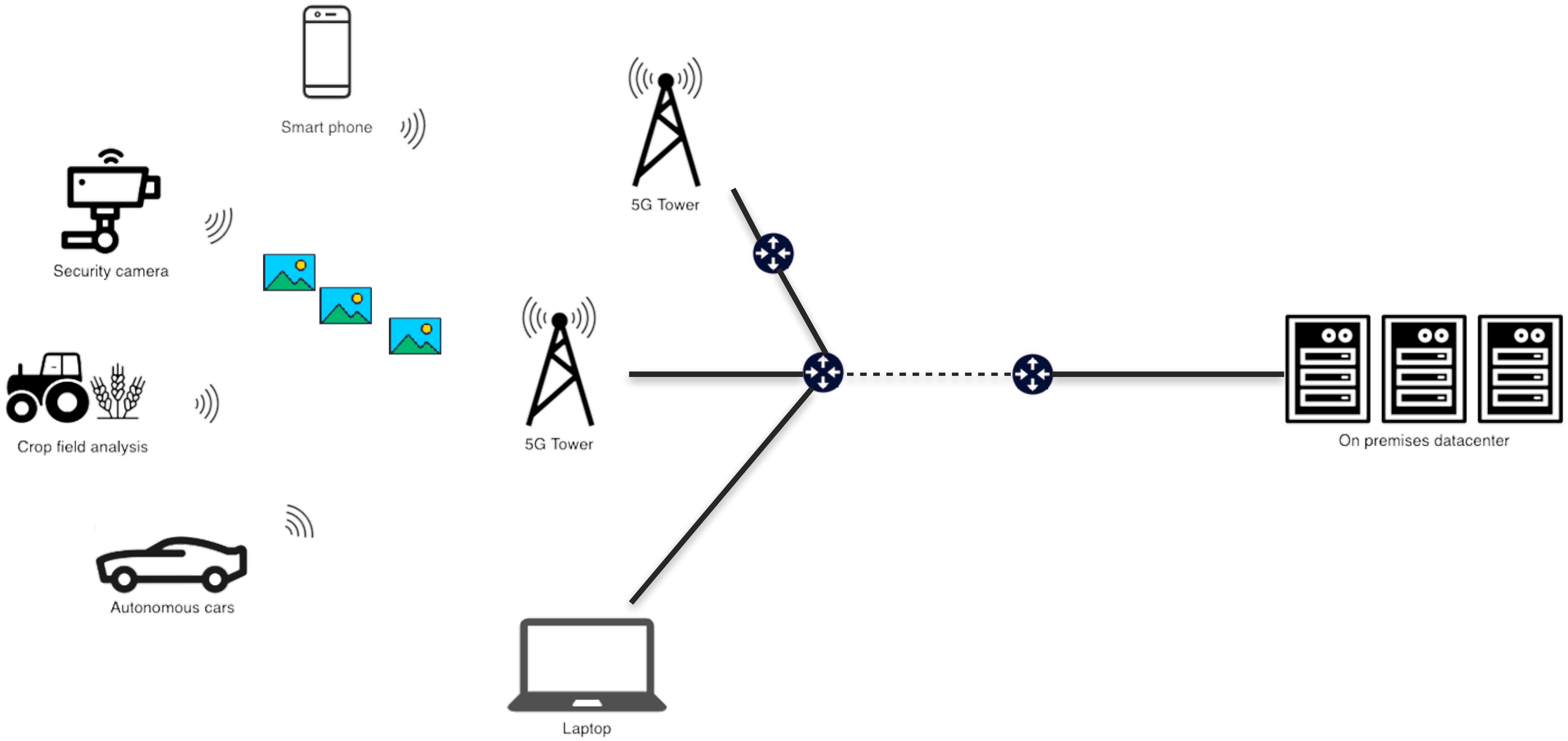
Context: Infrastructures behind AI



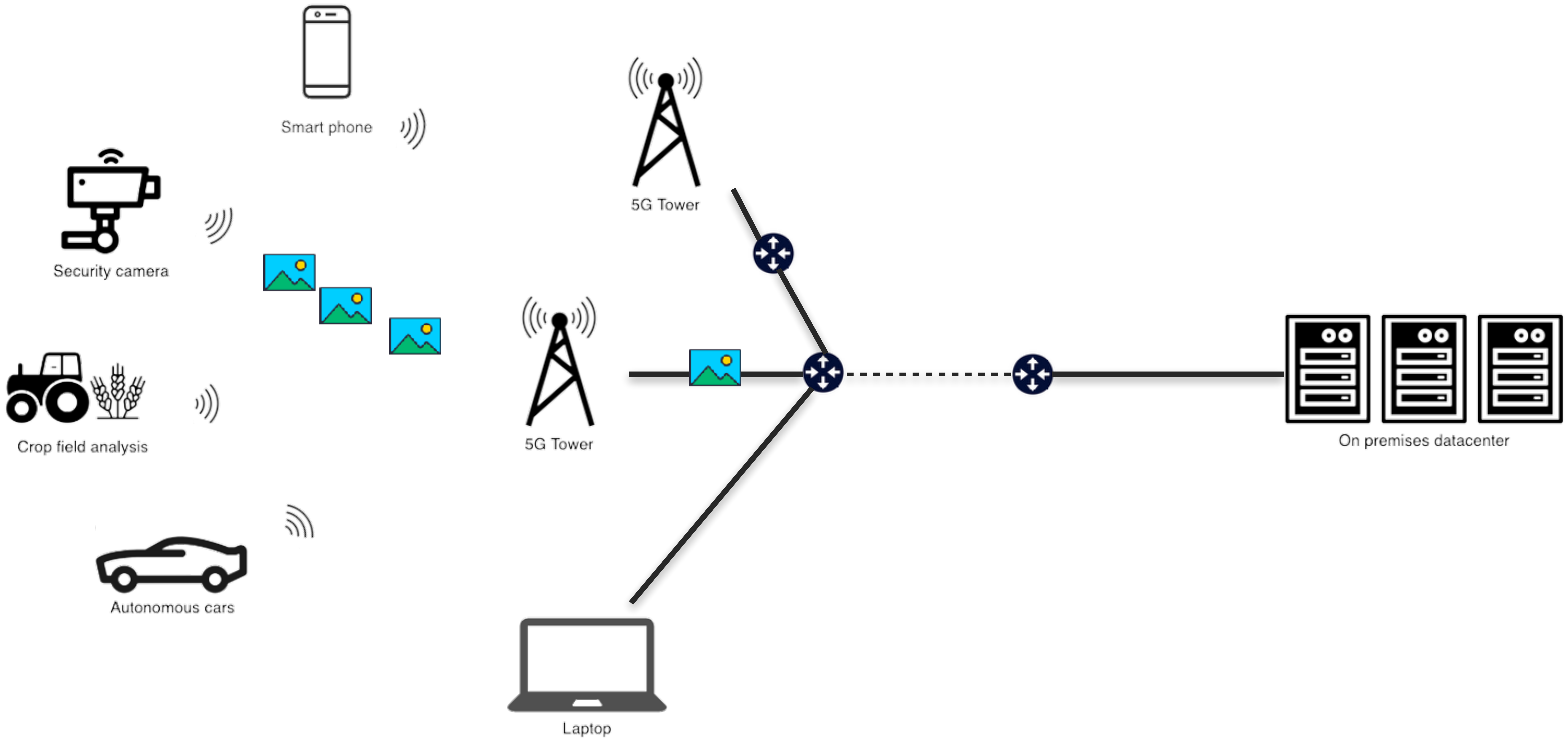
Context: Infrastructures behind AI



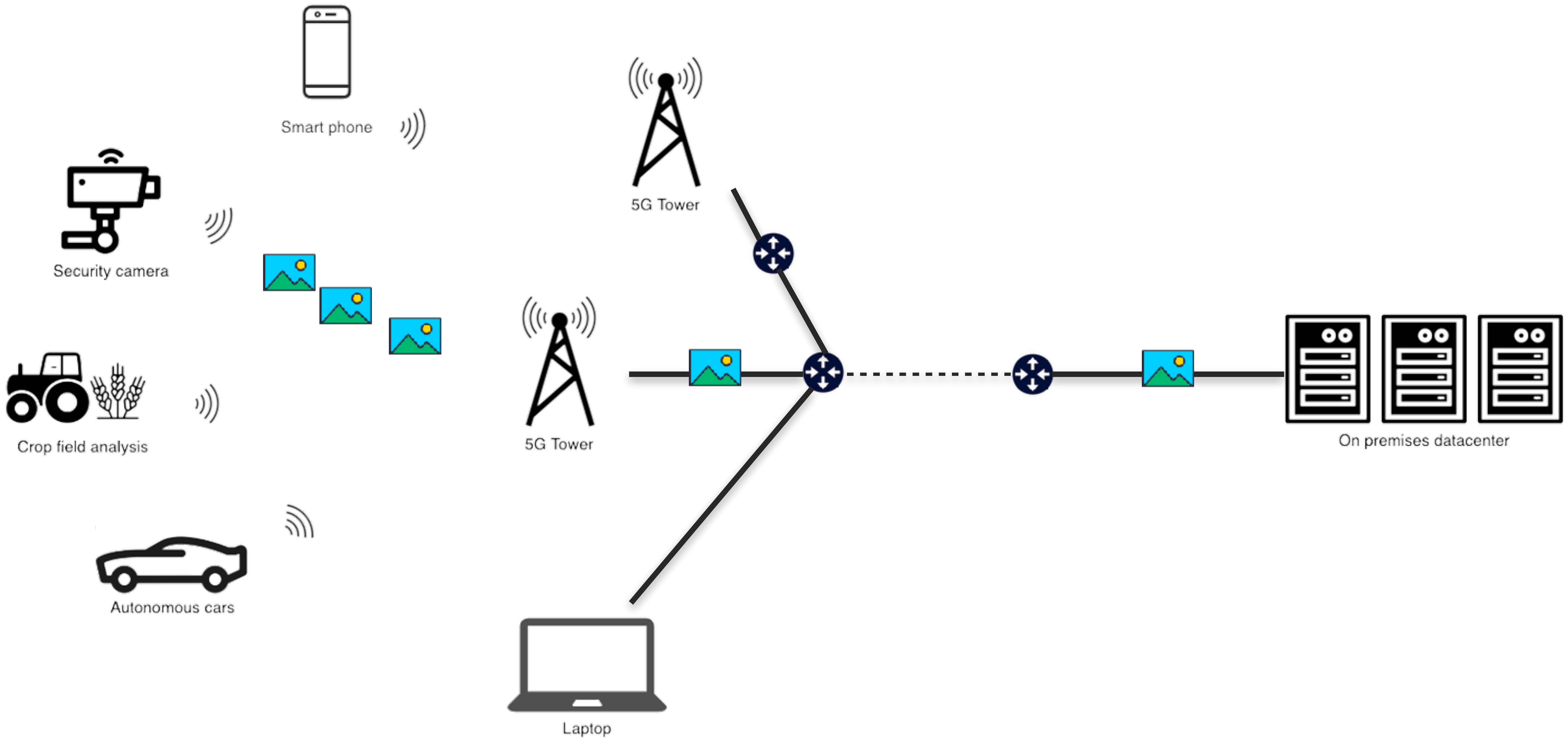
Context: Infrastructures behind AI



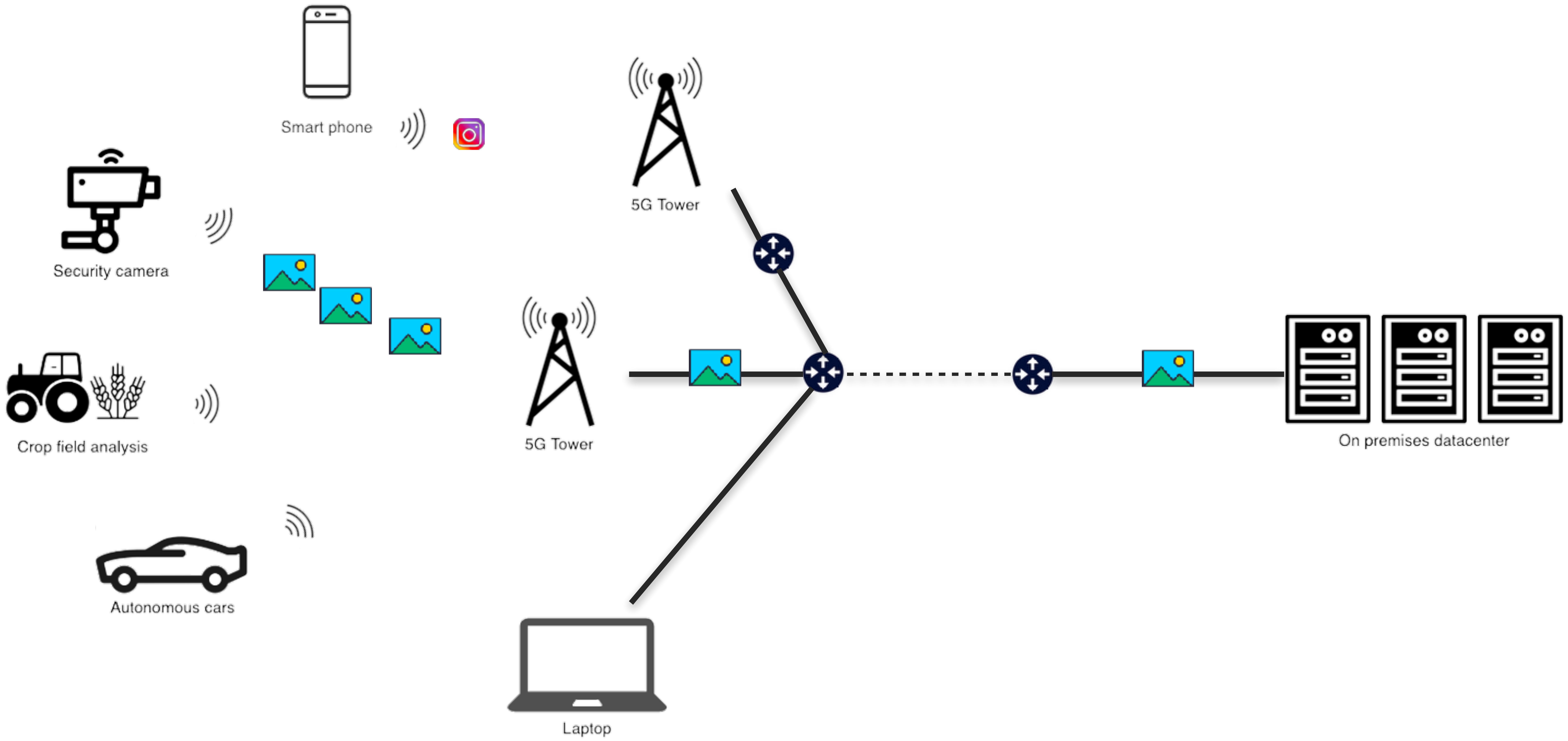
Context: Infrastructures behind AI



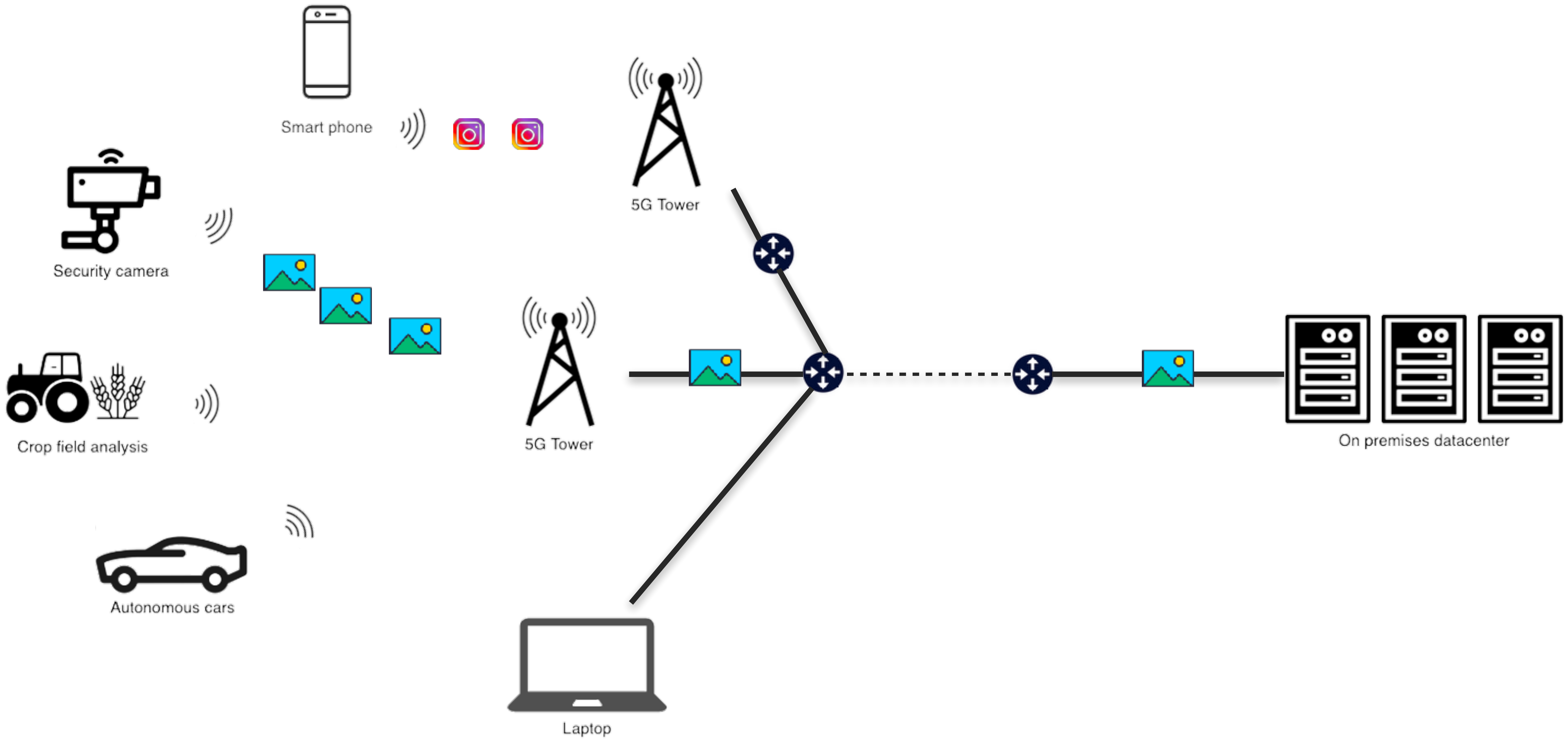
Context: Infrastructures behind AI



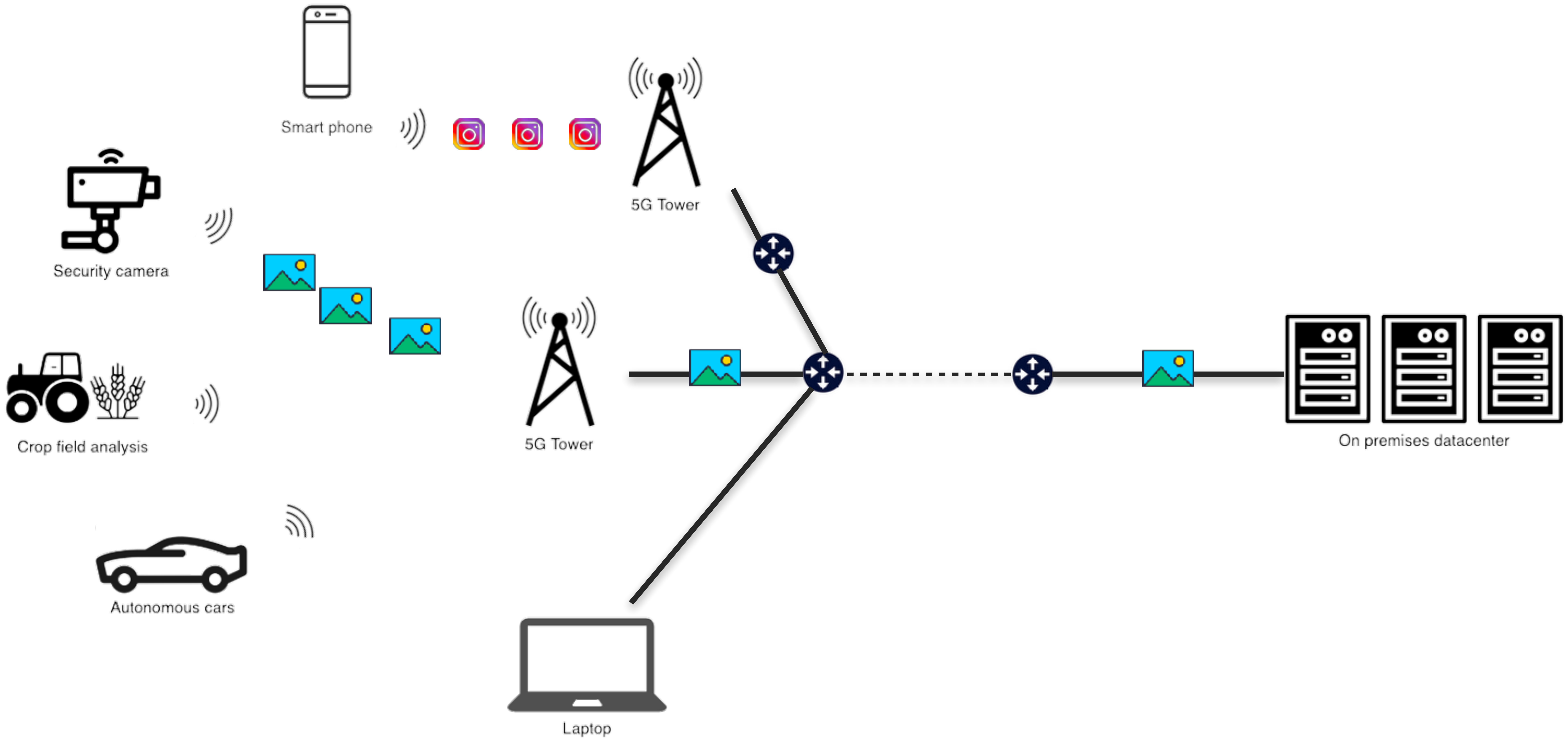
Context: Infrastructures behind AI



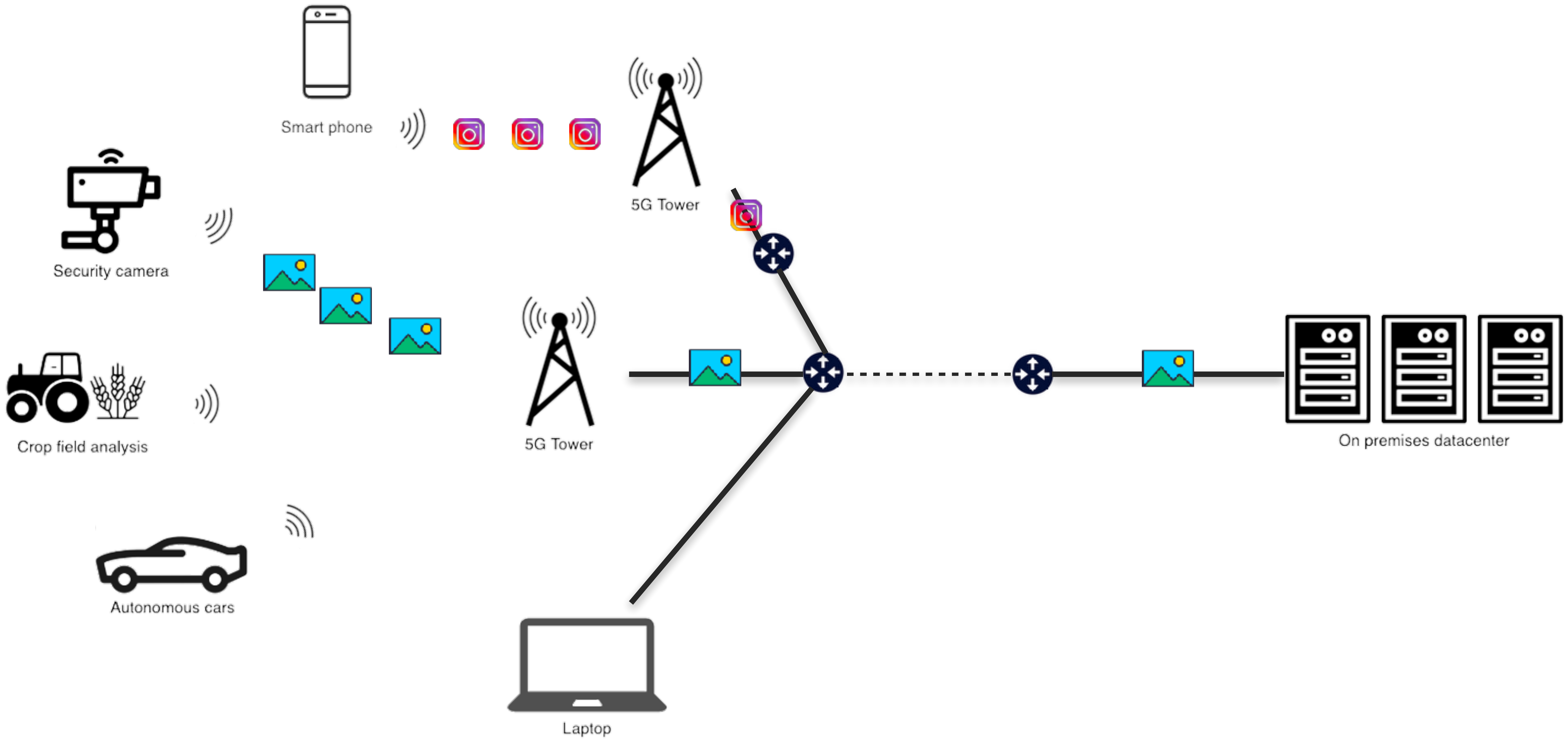
Context: Infrastructures behind AI



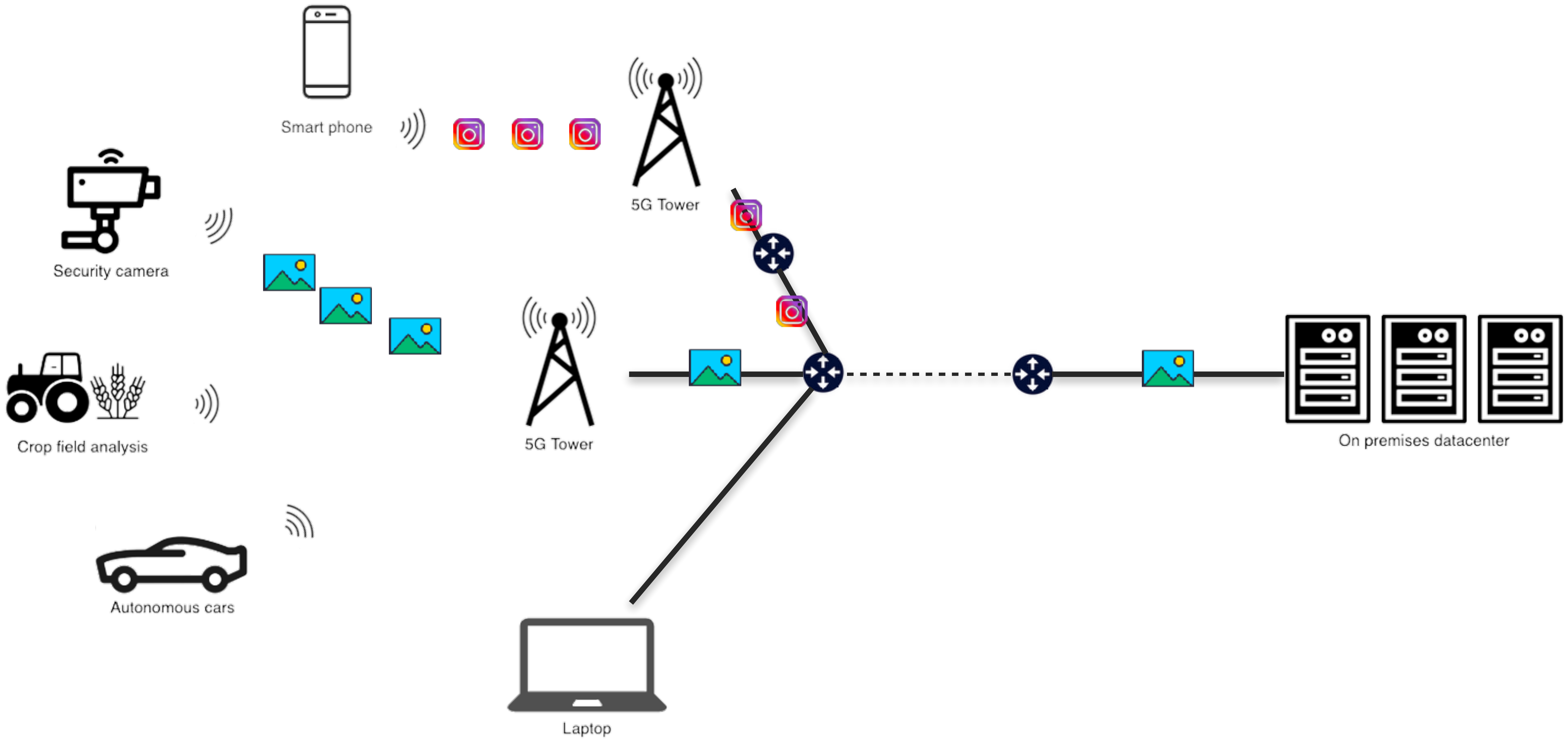
Context: Infrastructures behind AI



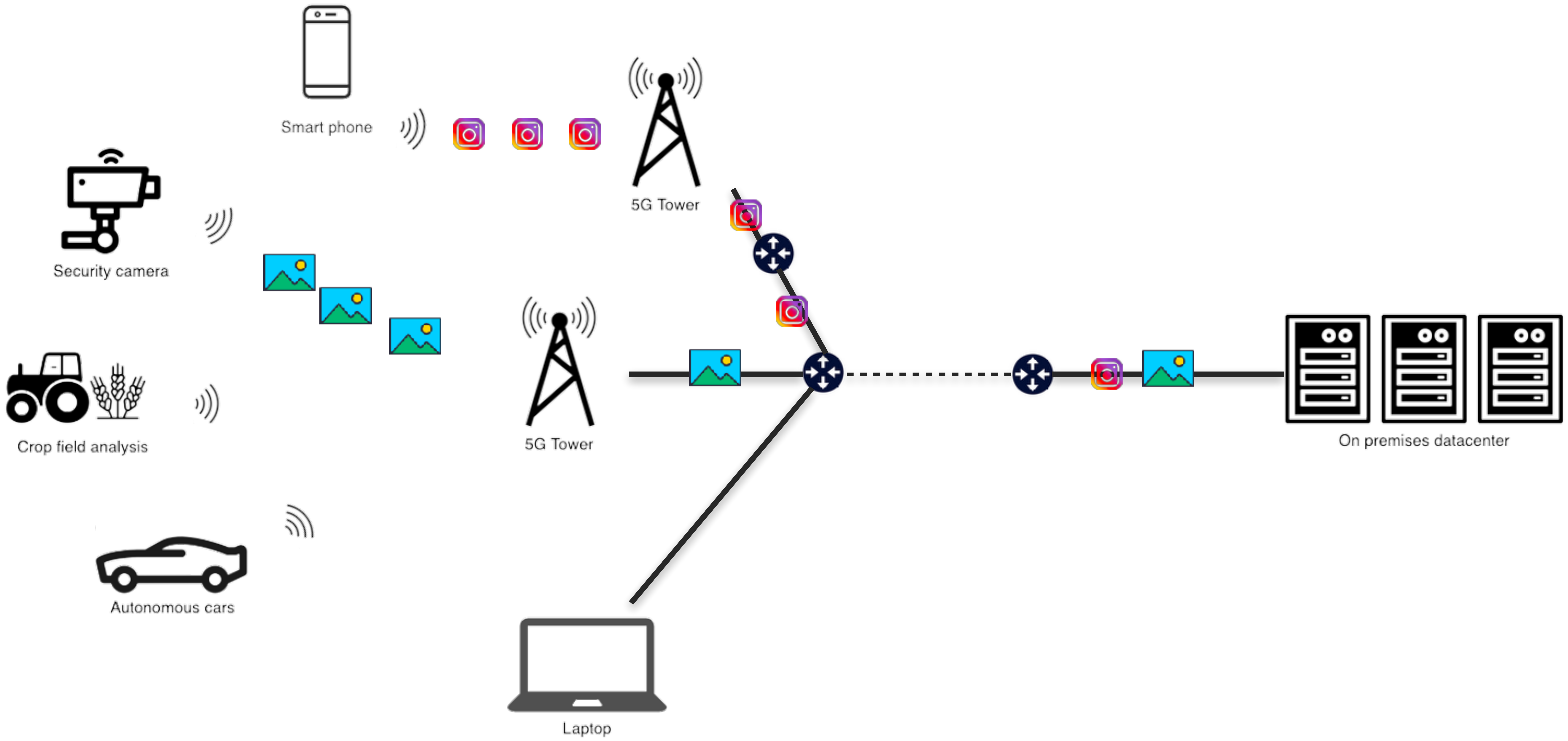
Context: Infrastructures behind AI



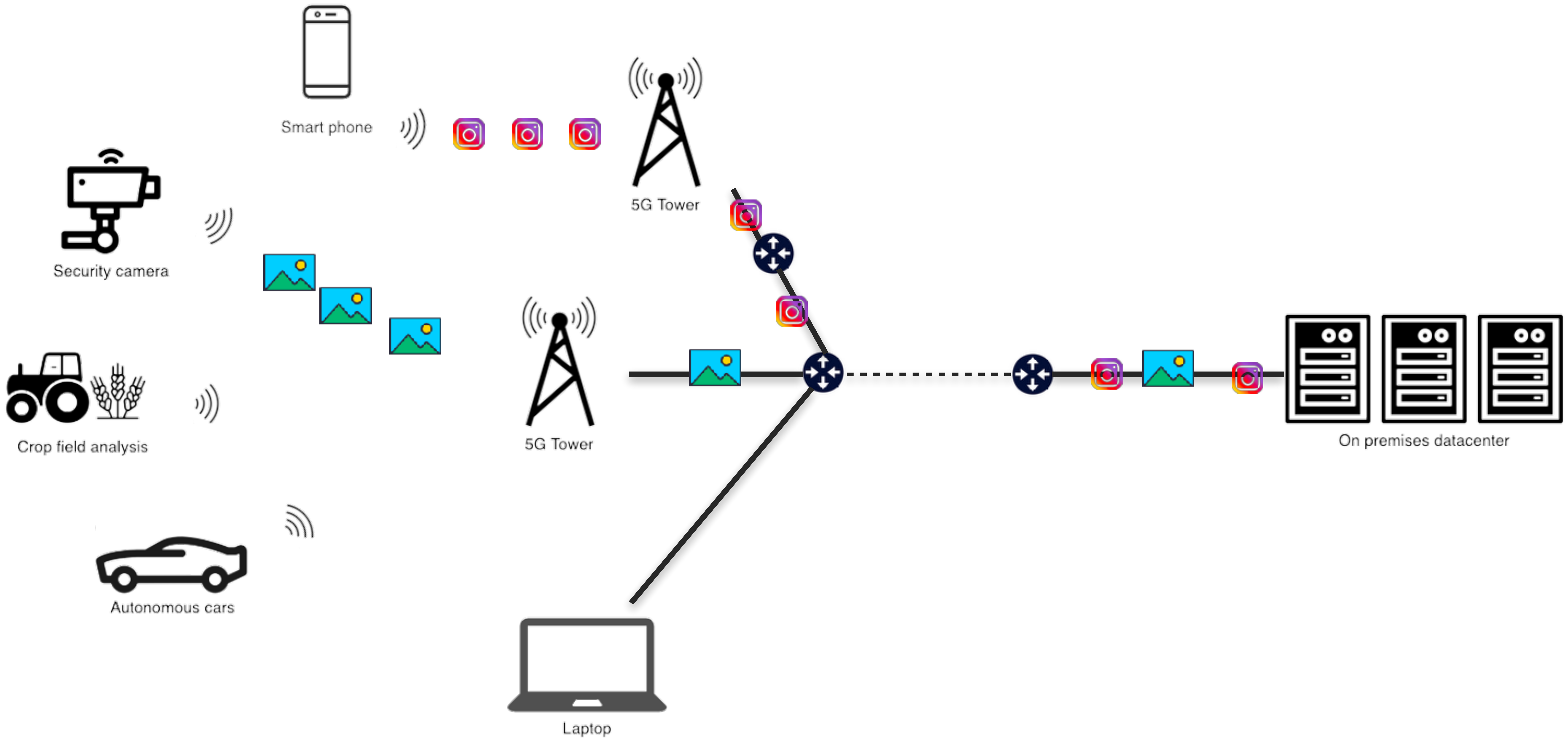
Context: Infrastructures behind AI



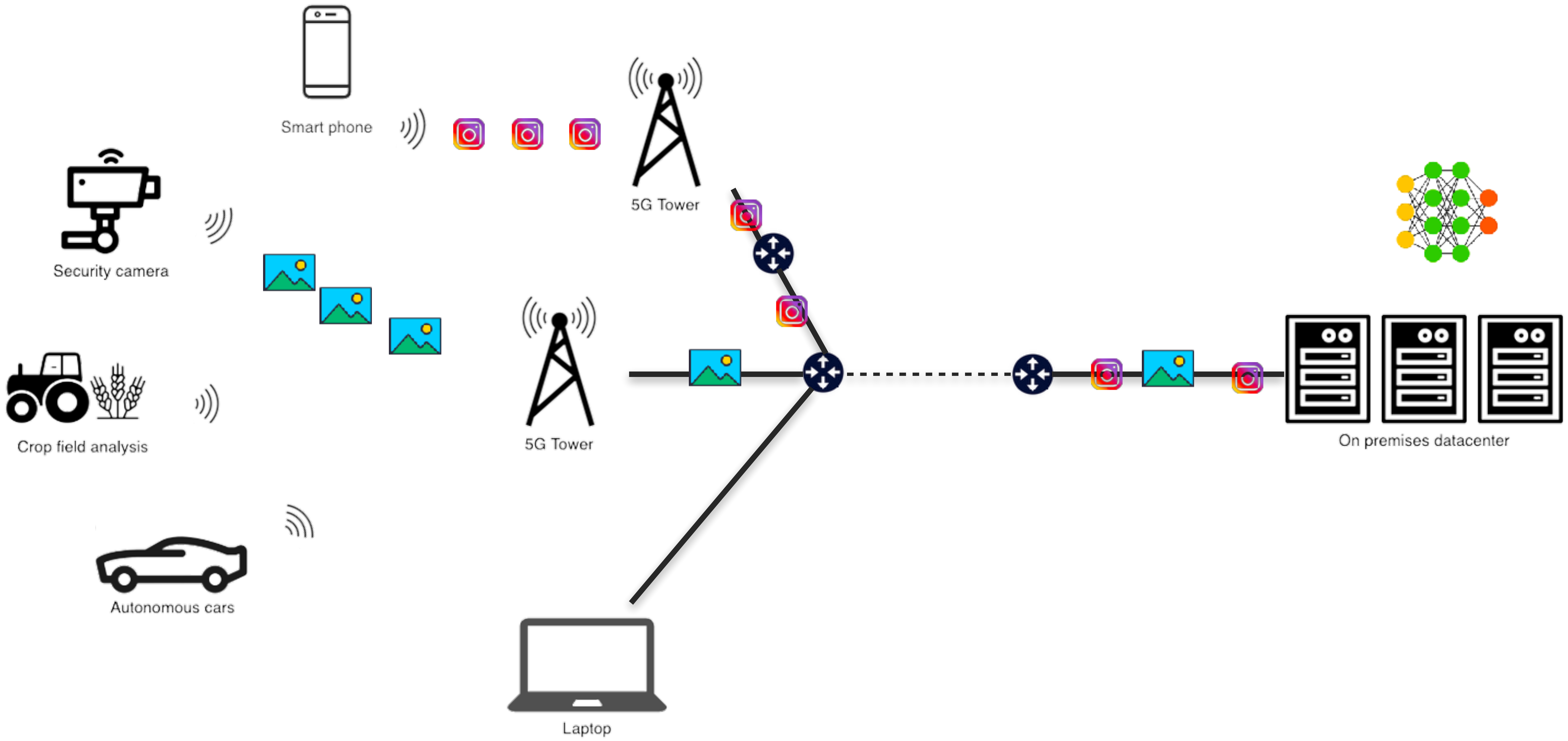
Context: Infrastructures behind AI



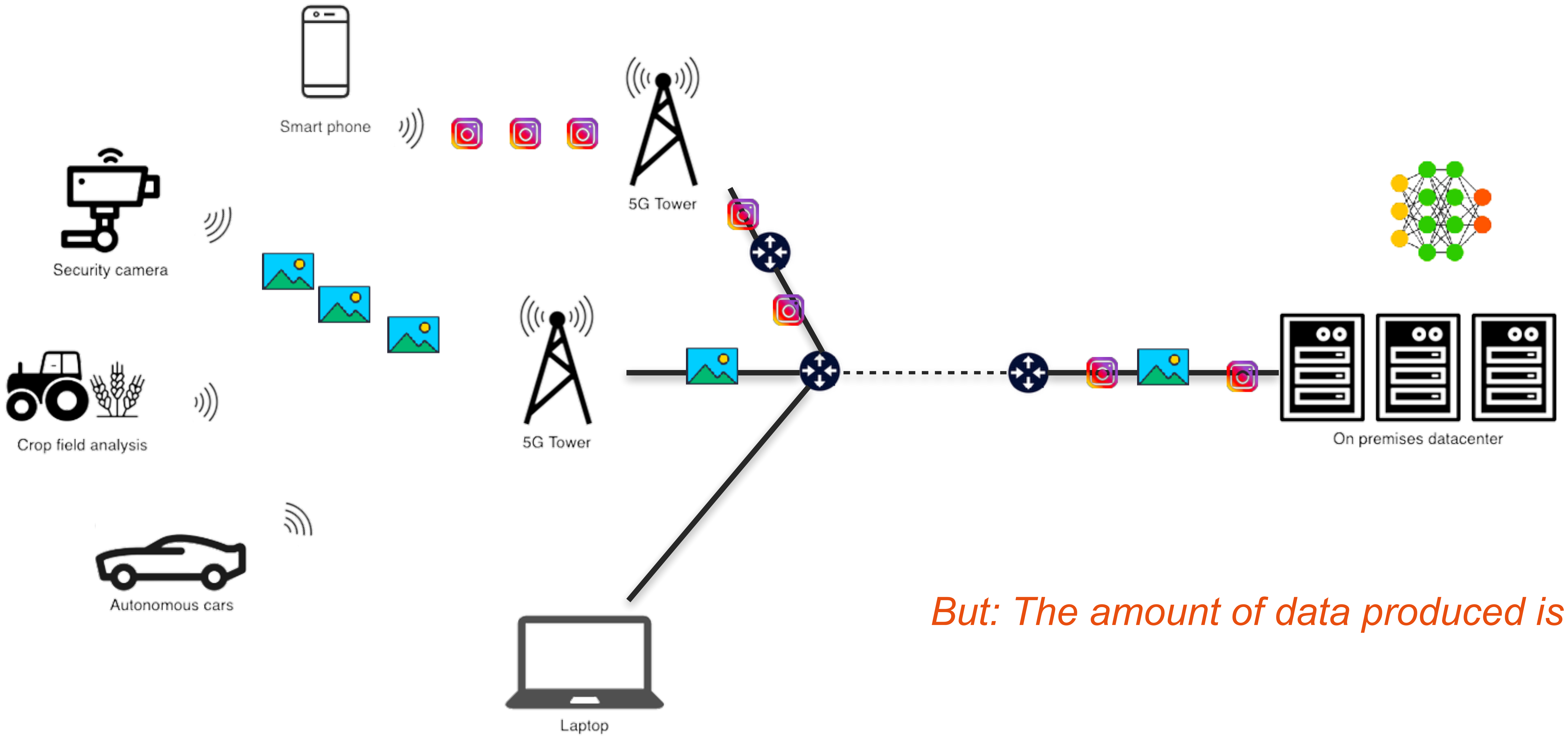
Context: Infrastructures behind AI



Context: Infrastructures behind AI

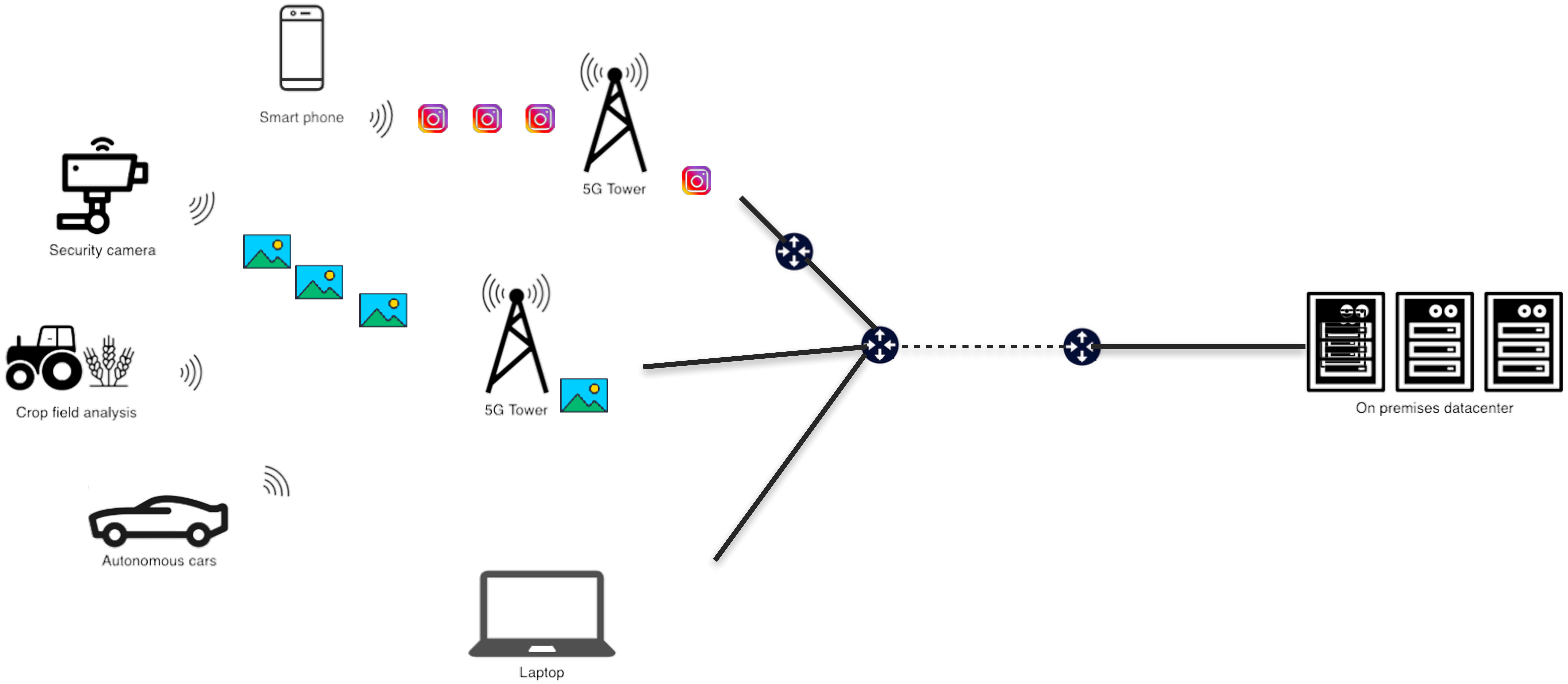


Context: Infrastructures behind AI

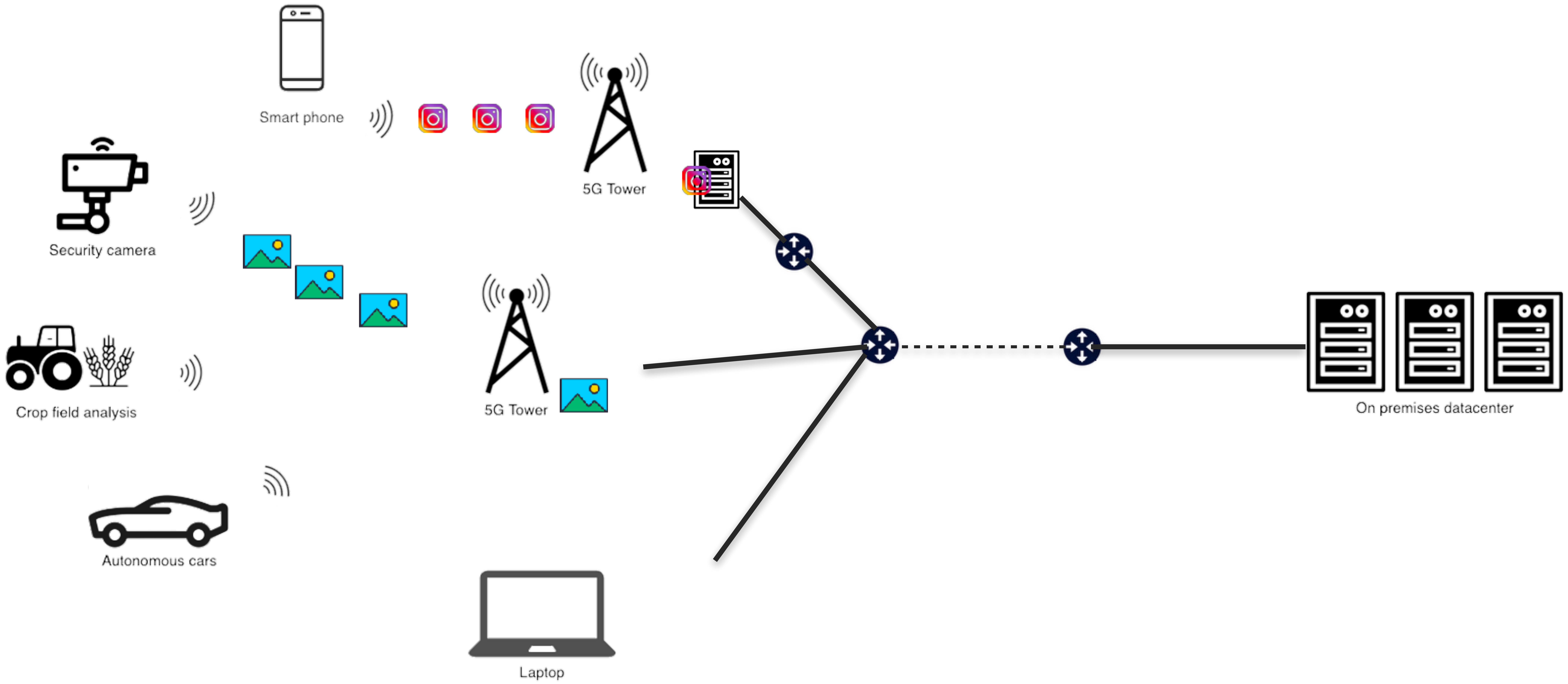


But: The amount of data produced is growing… a lot.

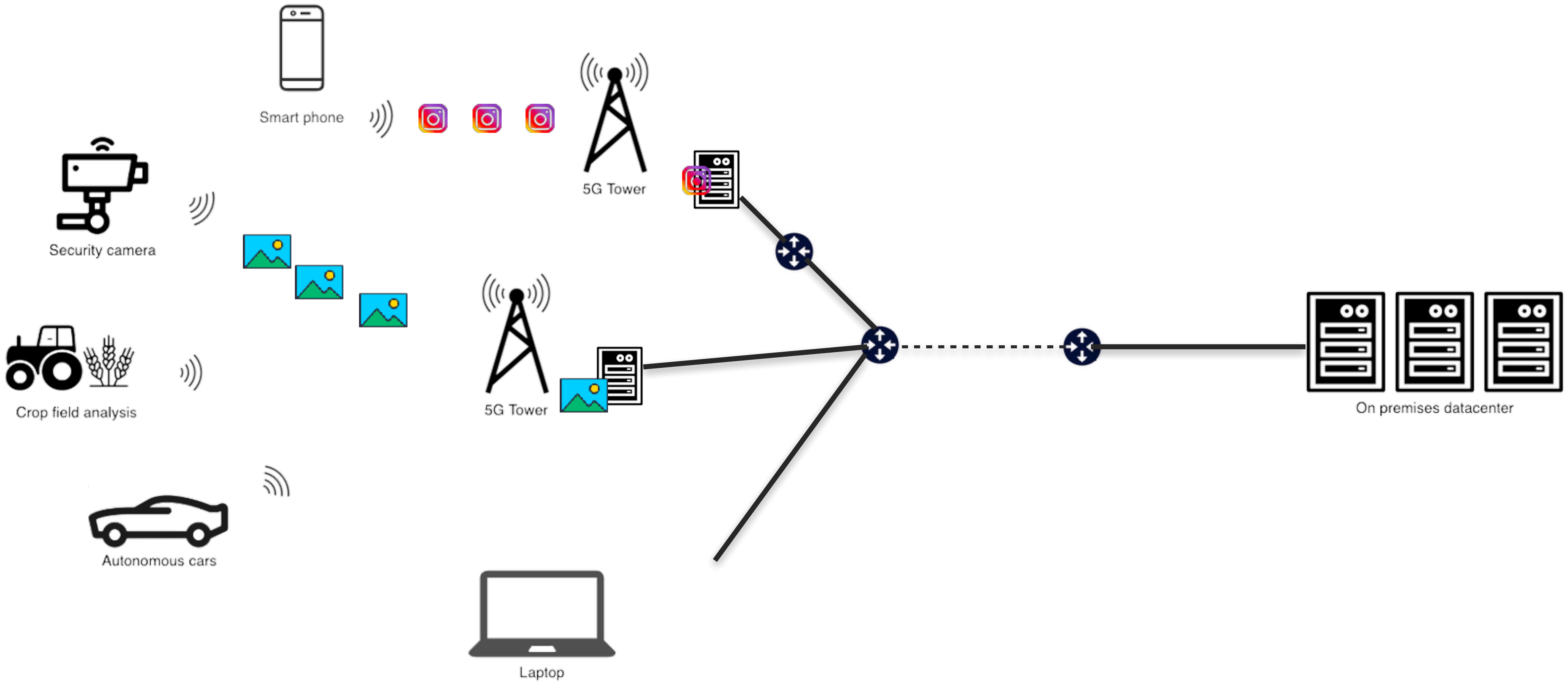
Context: Infrastructures behind AI



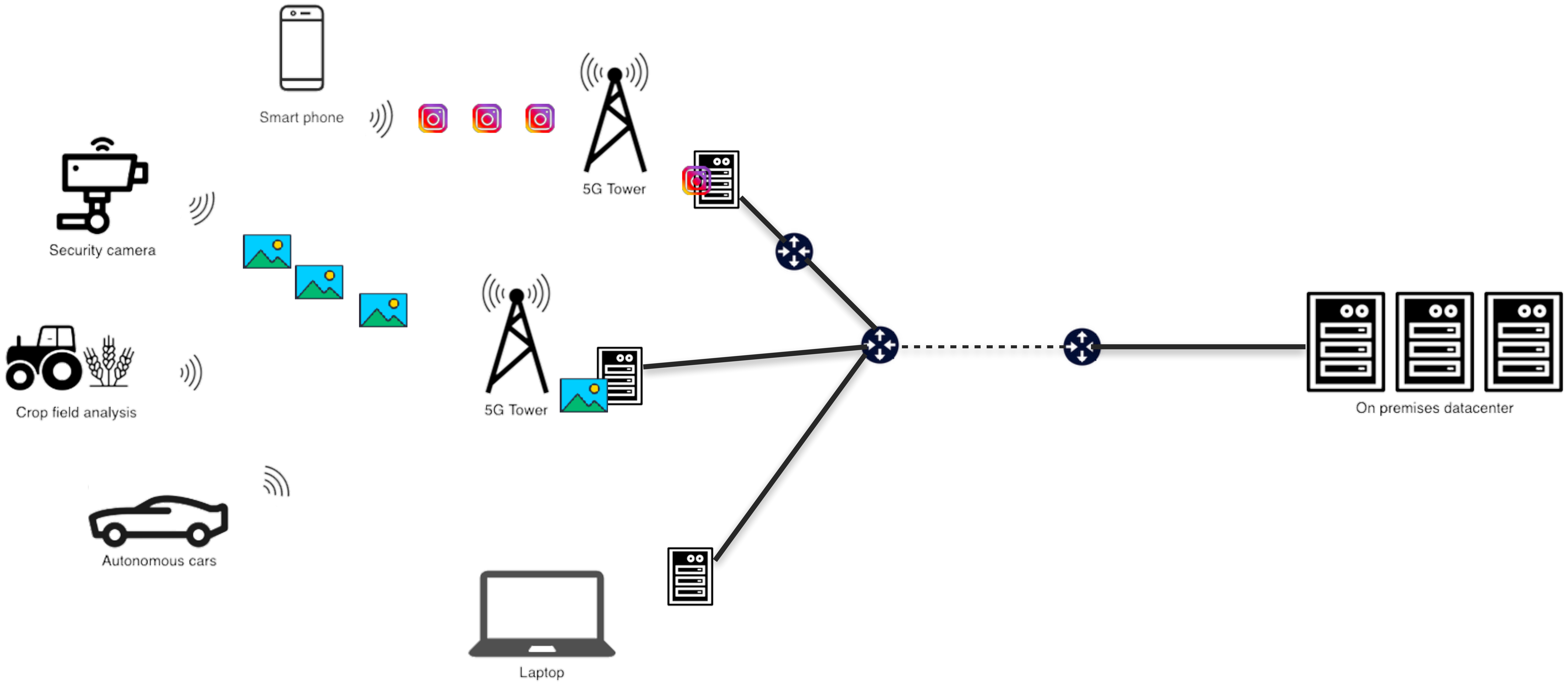
Context: Infrastructures behind AI



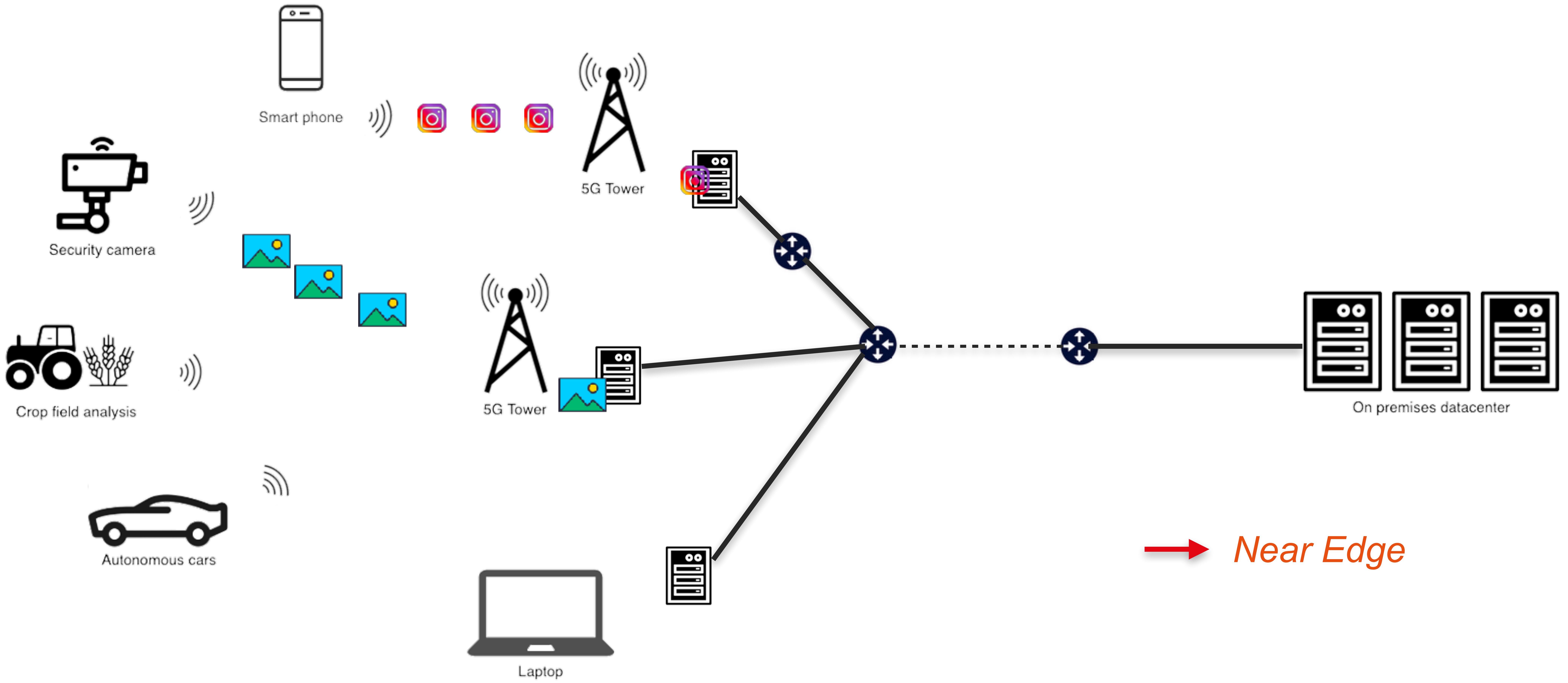
Context: Infrastructures behind AI



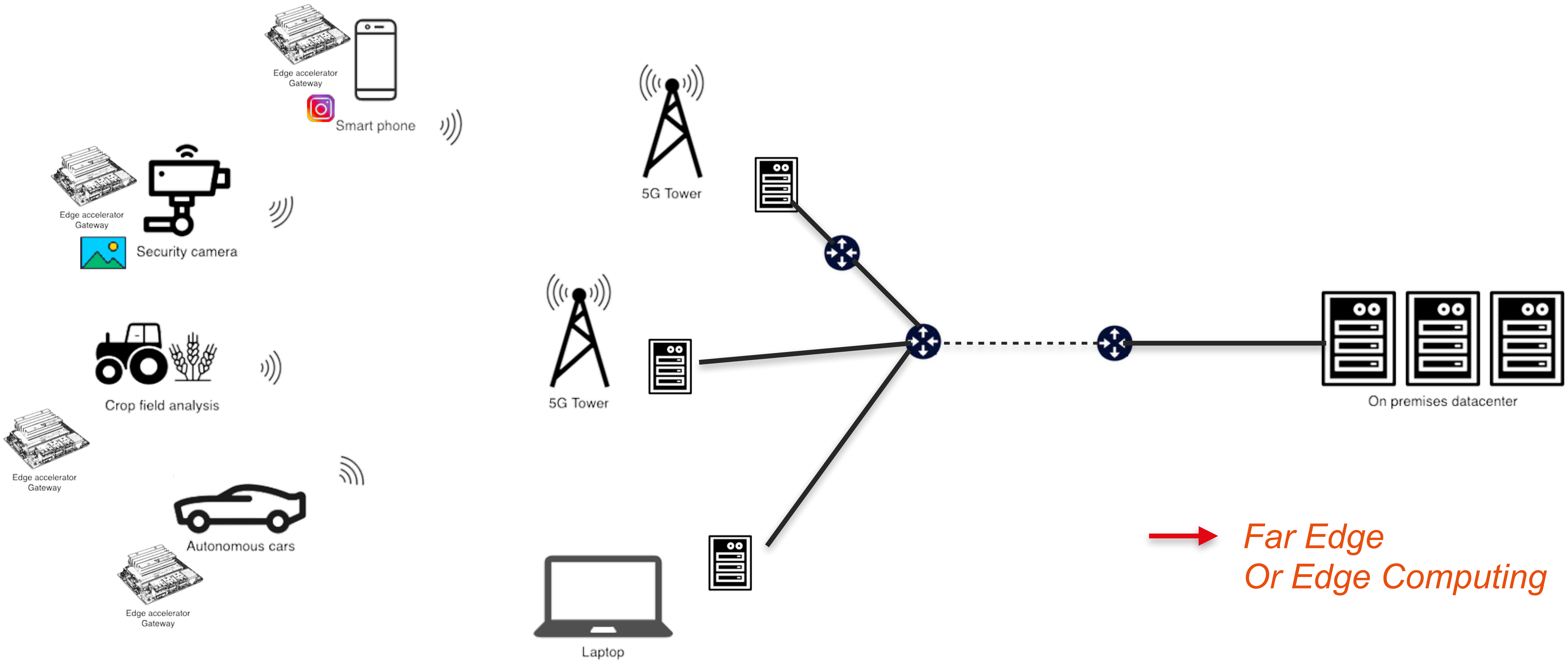
Context: Infrastructures behind AI



Context: Infrastructures behind AI



Context: Infrastructures behind AI



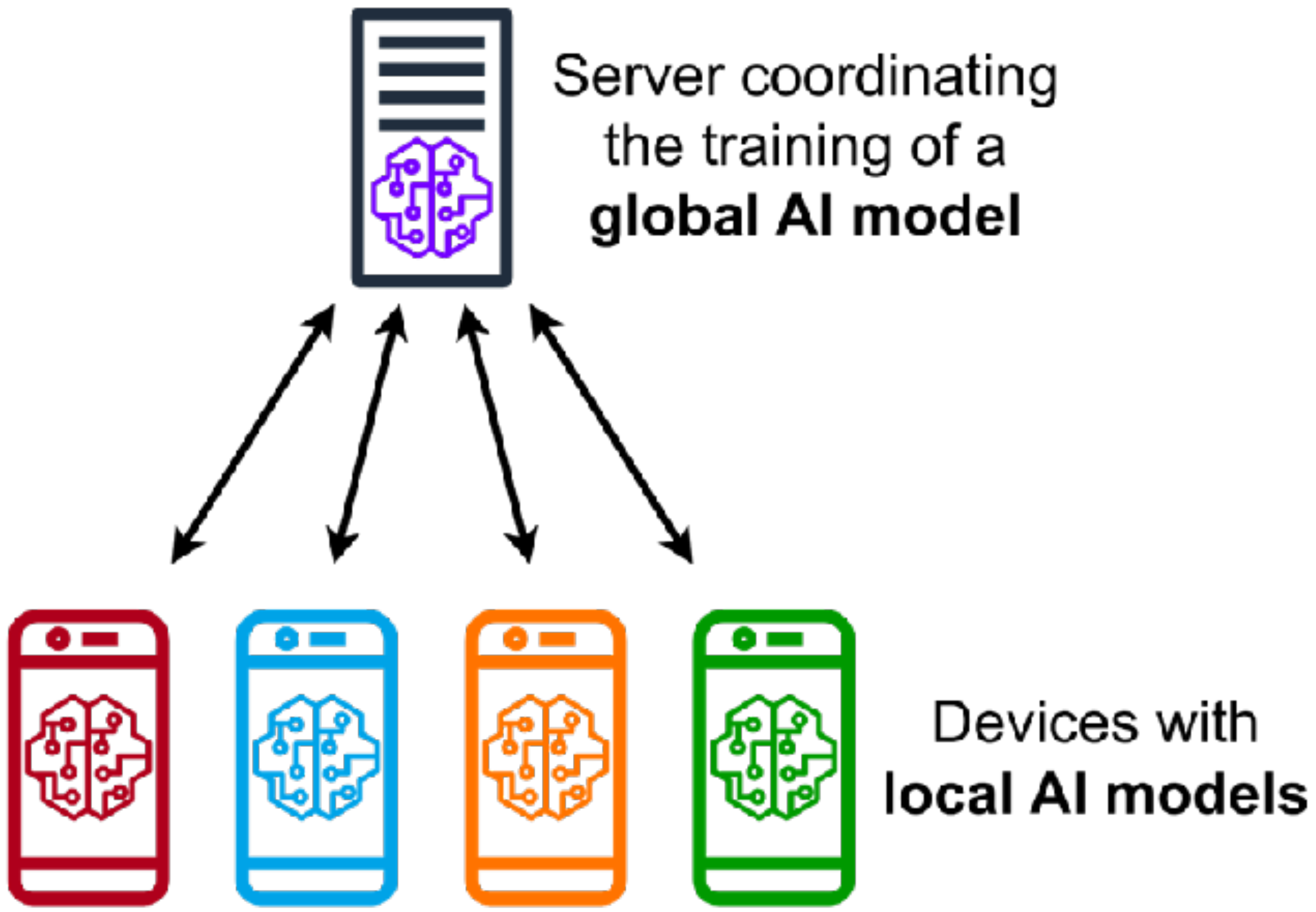
Far Edge
Or Edge Computing

«New » ML paradigm: Federated Learning

Federated Learning

- Learning on a selection of devices
- Aggregation on server
- Goal
  - Data stays in devices
  - Faster inference
- Challenges
  - Communication
  - Bias

- What about the energy cost?



Energy footprint

	Edge computing	Server computing	Federated Learning
Latency	None	High	Low
Privacy	High	Low	High
Data transfer	None	High	Low
Power cap	Low	High	??
Computation power efficiency	Low	High	??
Energy	**??**	**??**	**??**

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)

★ **Carbon** emissions

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 **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”. The 23rd IEEE/ ACM international symposium on Cluster, Cloud and Internet Computing, 2023.

Collaboration with Vladimir Ostapenco

Goal

- Help find the best tool for one’s need

Scope

- Cloud services: CPU processes
- Artificial Intelligence: GPU

Challenges

- Diversity of users & need
- Diversity of equipment

Methodology

- Selection of **7 software** based on internal interfaces or modeling
- **Quality** evaluation by comparing them with power meters on benchmarks
- **Qualitative** comparison: environment it is compatible with, how it works, its user-friendliness
- **Overhead** in energy
- Advices depending on **use cases**

Conclusion

- All software are **consistent** and have a low overhead
- Main **differences**: supported sampling frequencies, user-friendliness, supported components, granularity
- Tools including GPUs are less developed

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”. The 23rd IEEE/ACM international symposium on Cluster, Cloud and Internet Computing, 2023.

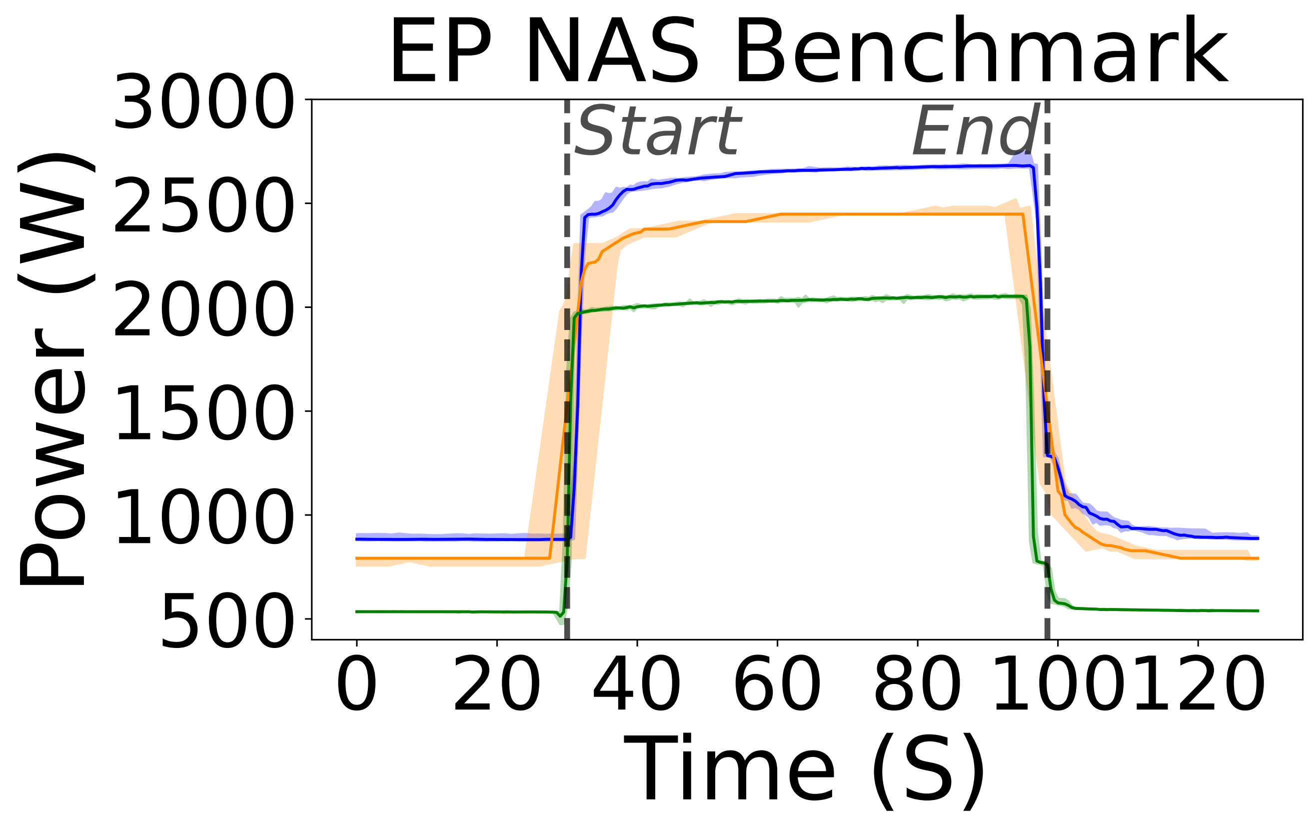
Methodology

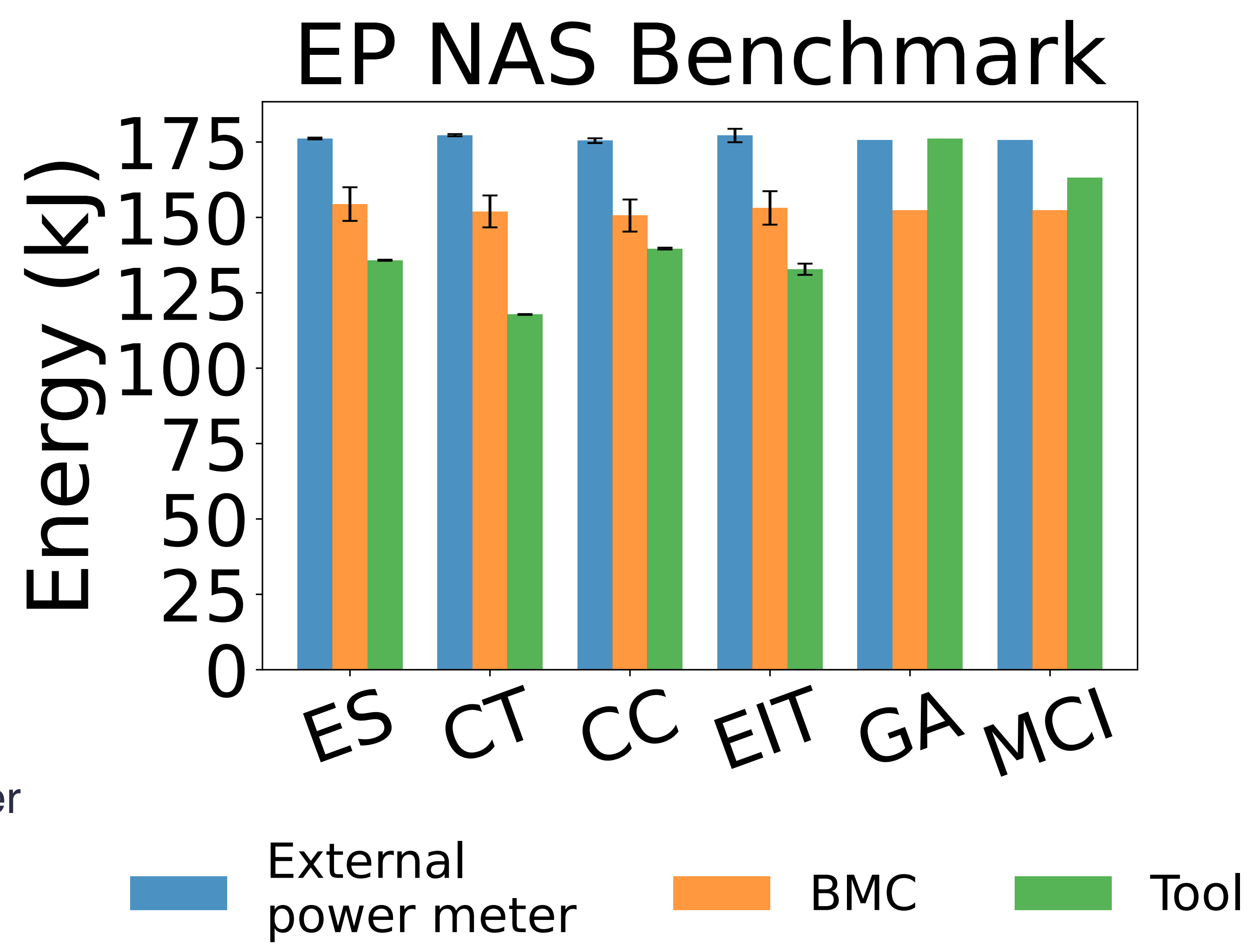
- Selection of **7 software** based on internal interfaces or modeling
- **Quality** evaluation by comparing them with power meters on benchmarks
- **Qualitative** comparison: environment it is compatible with, how it works, its user-friendliness
- **Overhead** in energy
- Advices depending on **use cases**

Conclusion

- All software are **consistent** and have a low overhead
- Main **differences**: supported sampling frequencies, user-friendliness, supported components, granularity
- Tools including GPUs are less developed

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”. The 23rd IEEE/ ACM international symposium on Cluster, Cloud and Internet Computing, 2023.





PA - PowerAPI
SC - Scaphandre
ES - Energy Scope
PE - Perf
CT - Carbon Tracker
CC - Code Carbon
EIT - Experiment Impact Tracker
GA - Green Algorithms
MCI - ML CO2 Impact

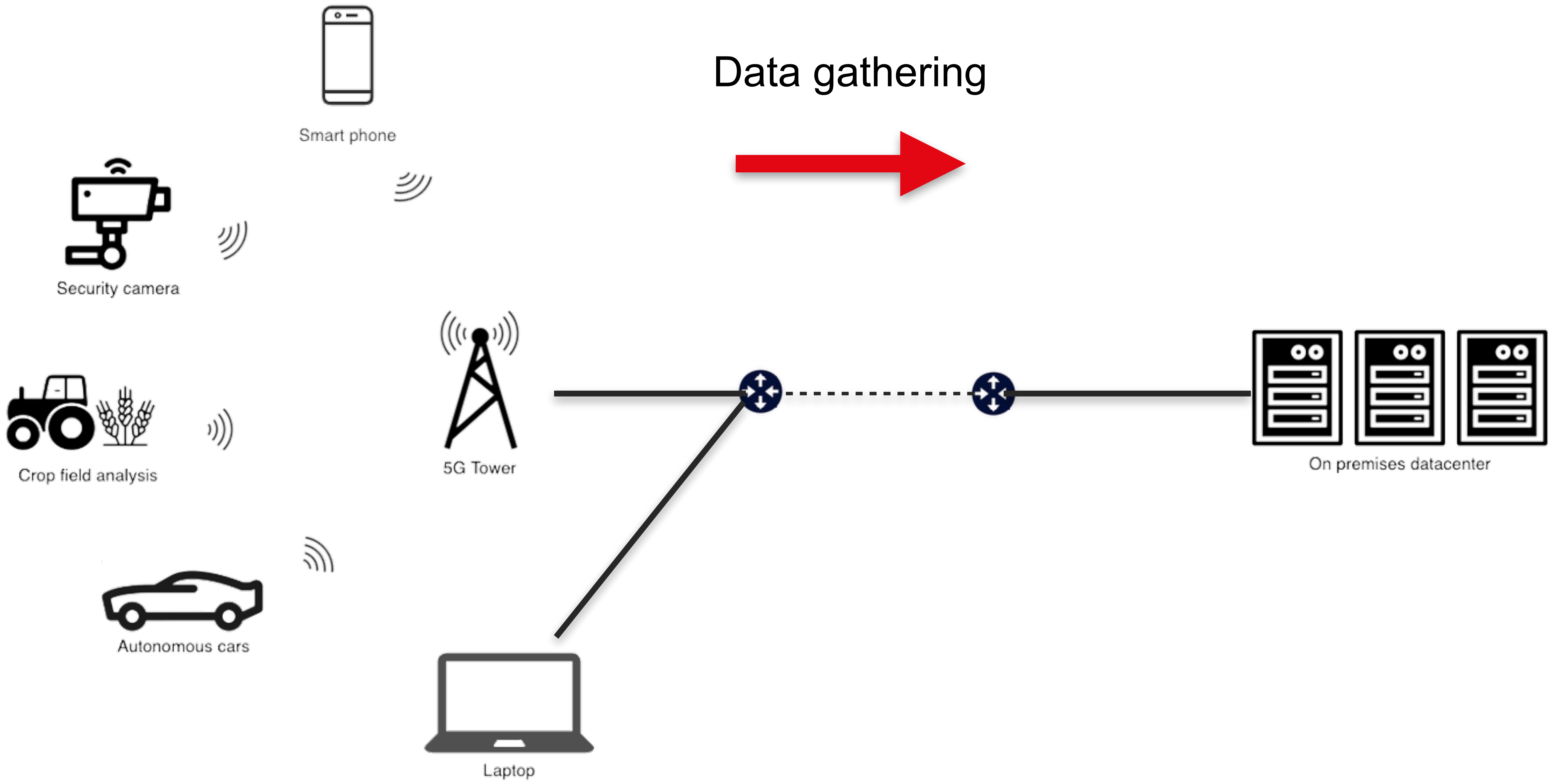
Methodology

- Selection of **7 software** based on internal interfaces or modeling
- **Quality** evaluation by comparing them with power meters on benchmarks
- **Qualitative** comparison: environment it is compatible with, how it works, its user-friendliness
- **Overhead** in energy
- Advices depending on **use cases**

Conclusion

- All software are **consistent** and have a low overhead
- Main **differences**: supported sampling frequencies, user-friendliness, supported components, granularity
- Tools including GPUs are less developed

ML infrastructures



Data gathering

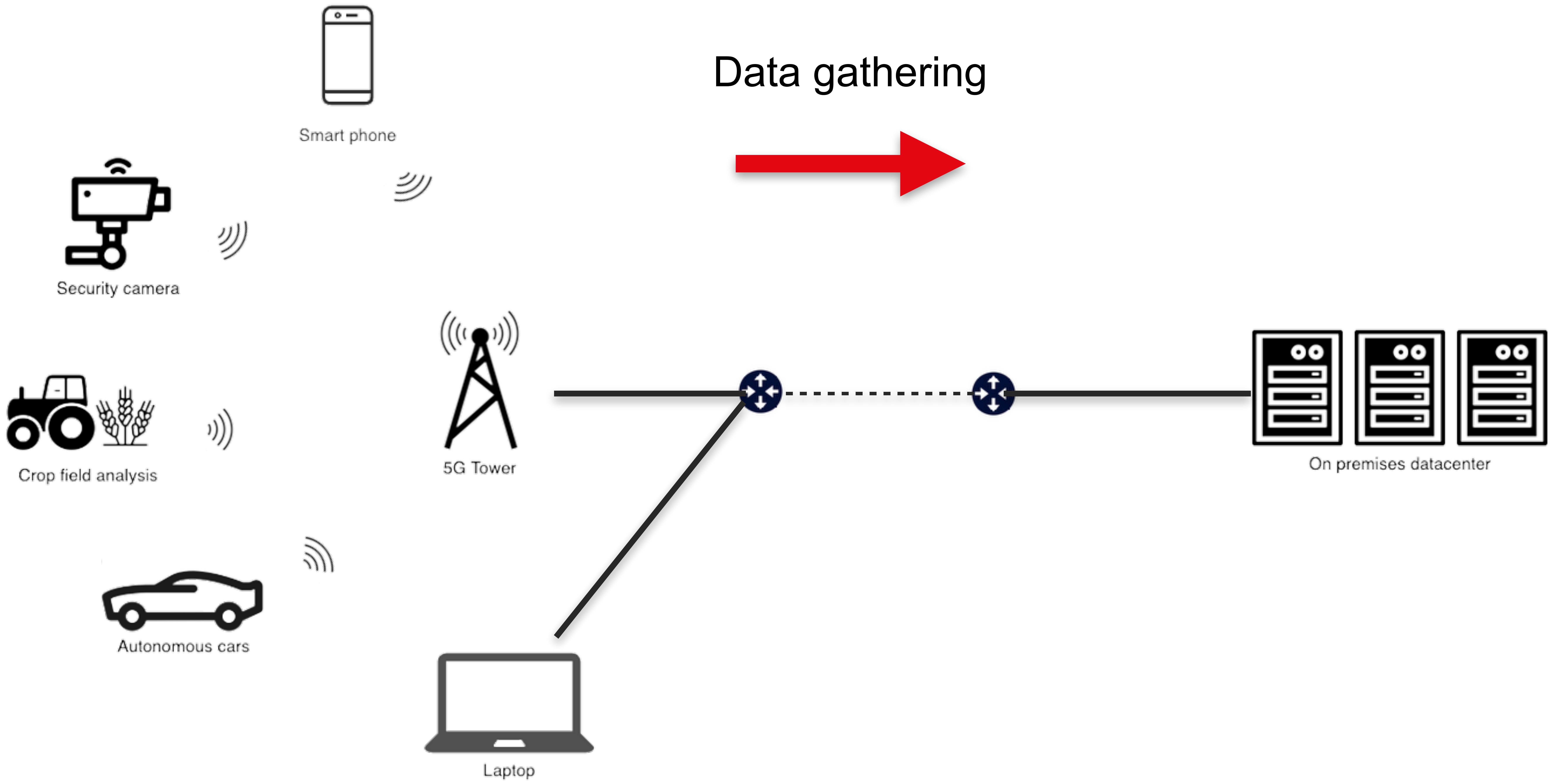
Data processing

Data analysis

Train models

Models deployment

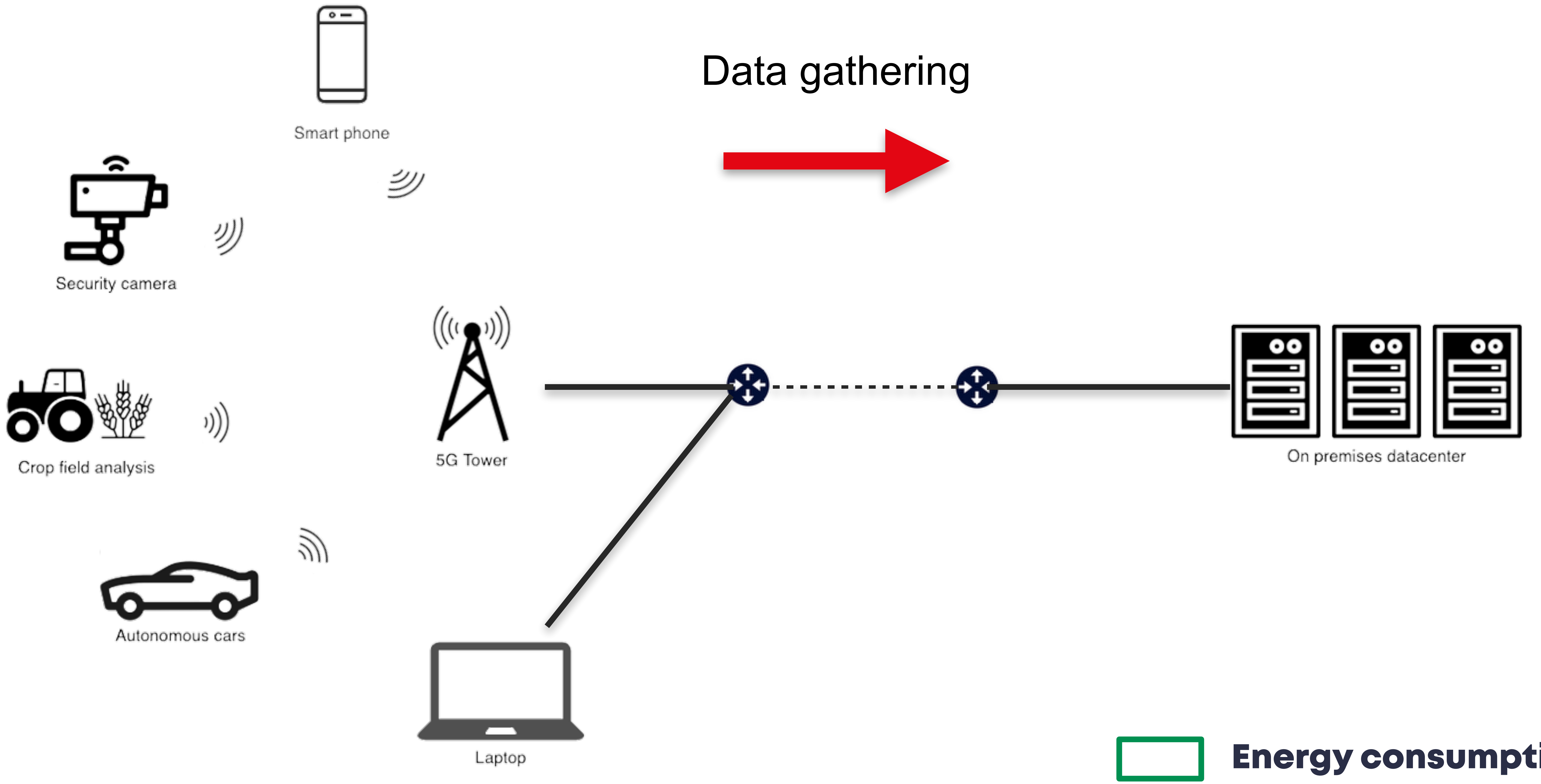
ML infrastructures



Data processing
Data analysis
Train models
Models deployment

Energy which could be reported with tools available

ML infrastructures



Data processing

Data analysis

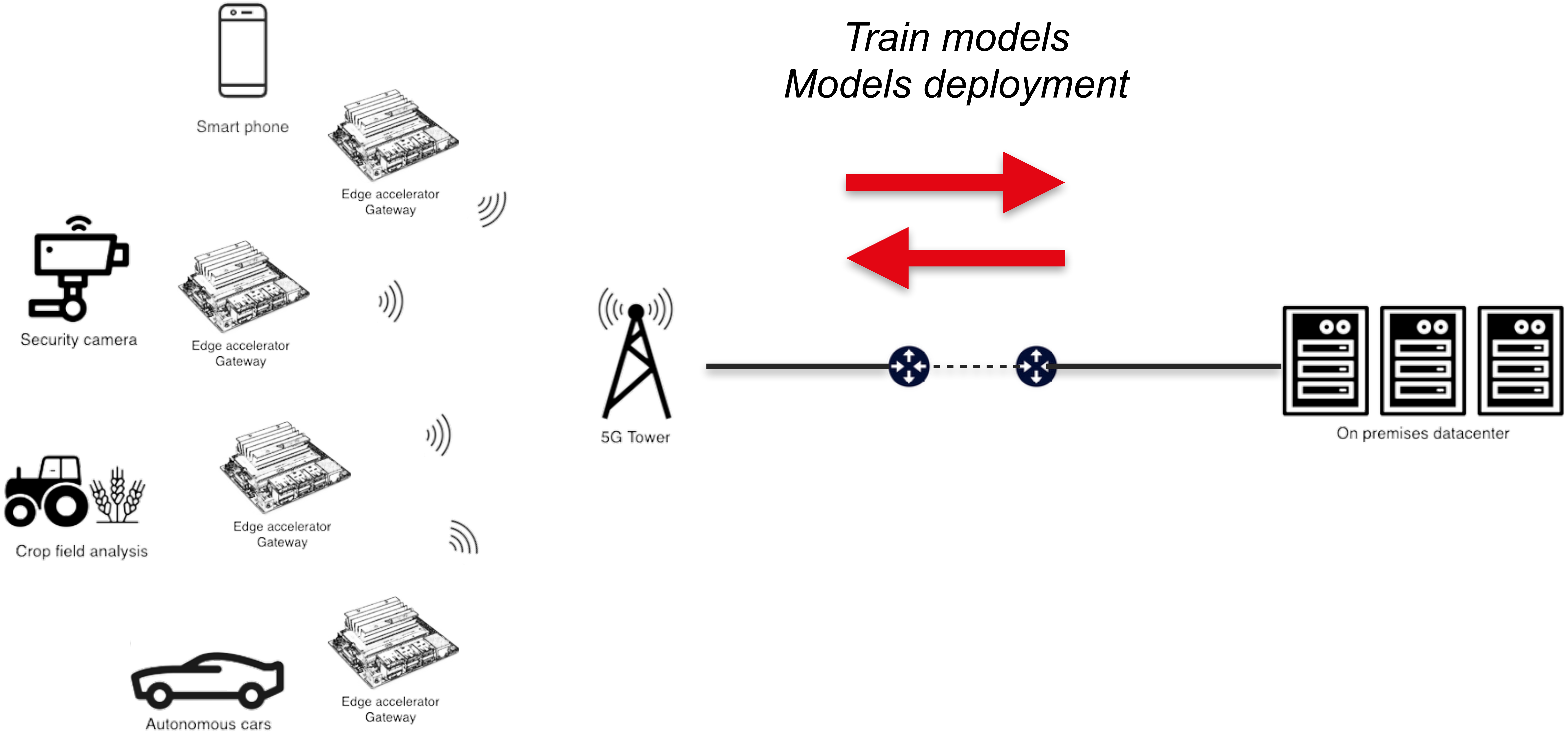
Train models

Models deployment

Energy consumption usually reported

Energy which could be reported with tools available

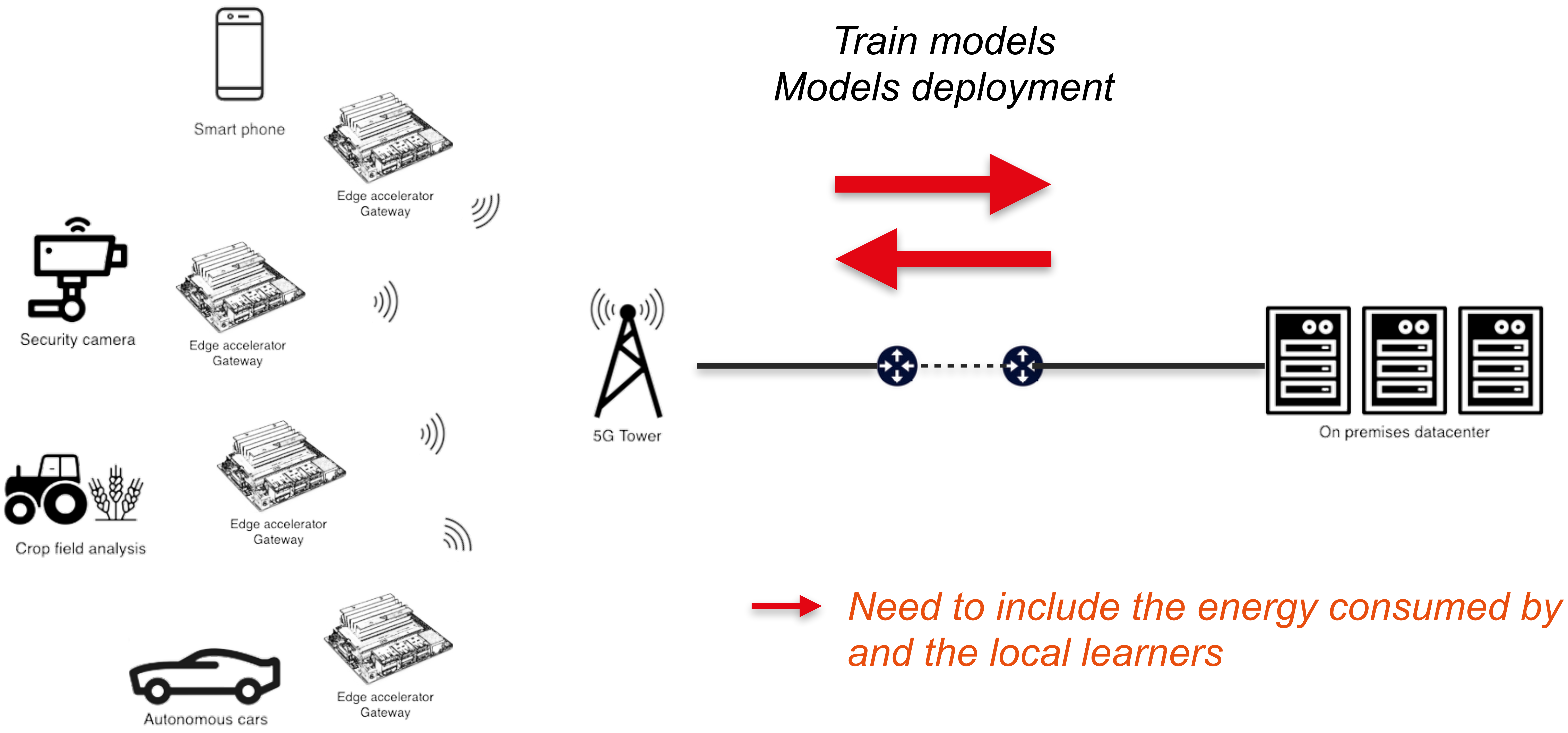
Federated Learning infrastructures



Train models
Models deployment

Data processing
Data analysis
Model averaging

Federated Learning infrastructures

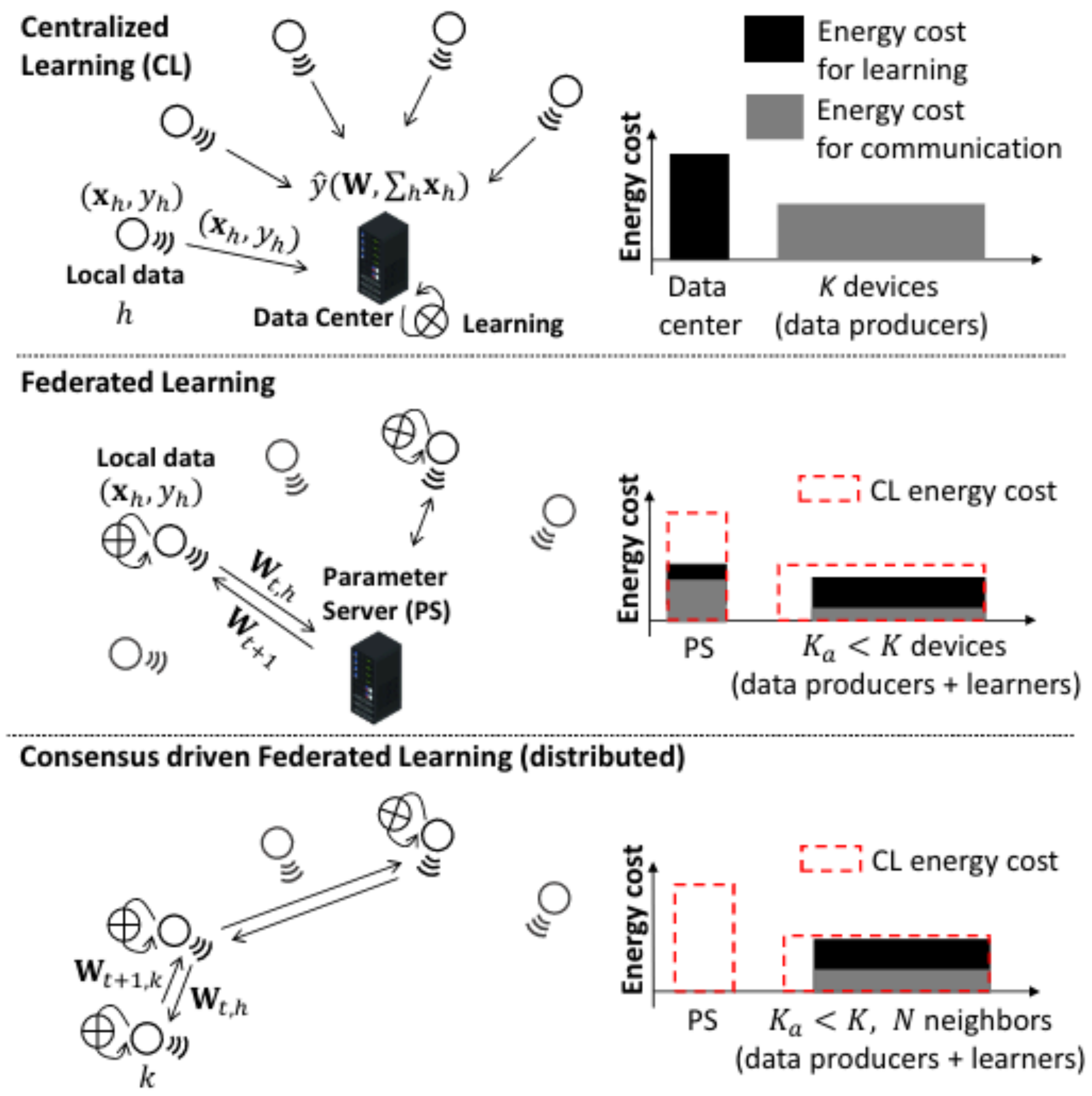


Train models
Models deployment

Data processing
Data analysis
Model averaging

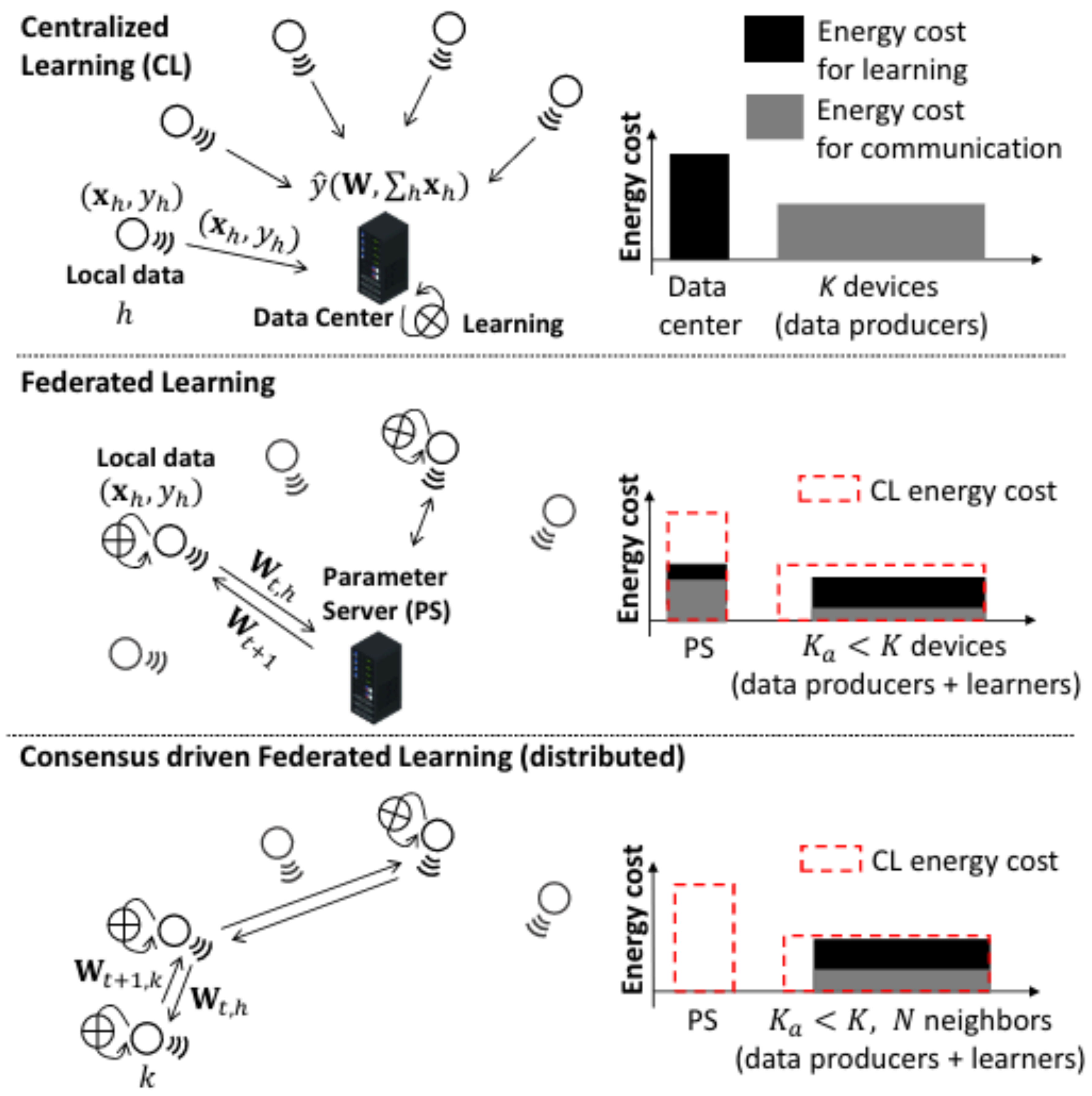
→ *Need to include the energy consumed by the data transfer and the local learners*

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.



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.

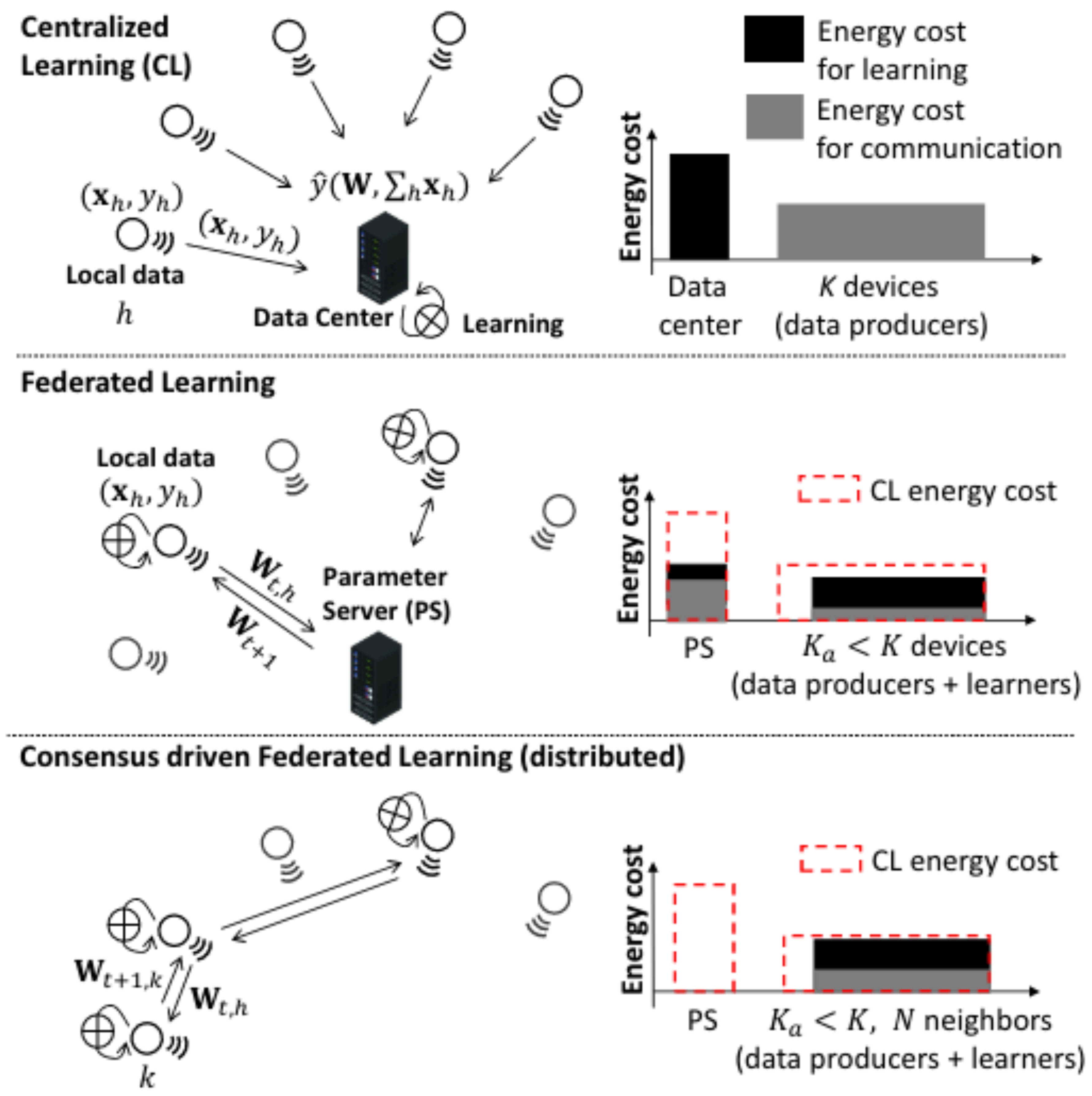
Energy consumption **simulator** from



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

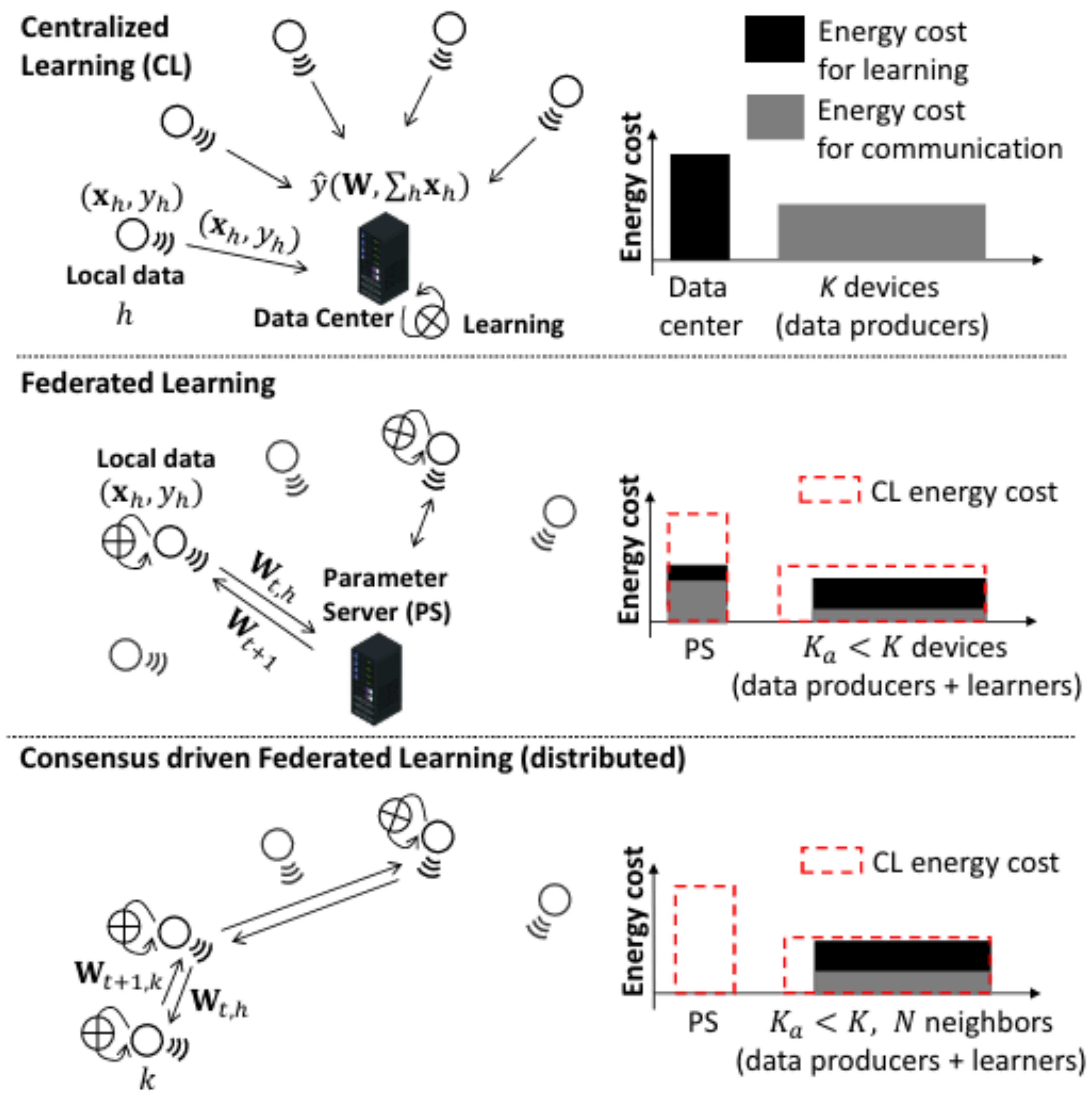
- PUE



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

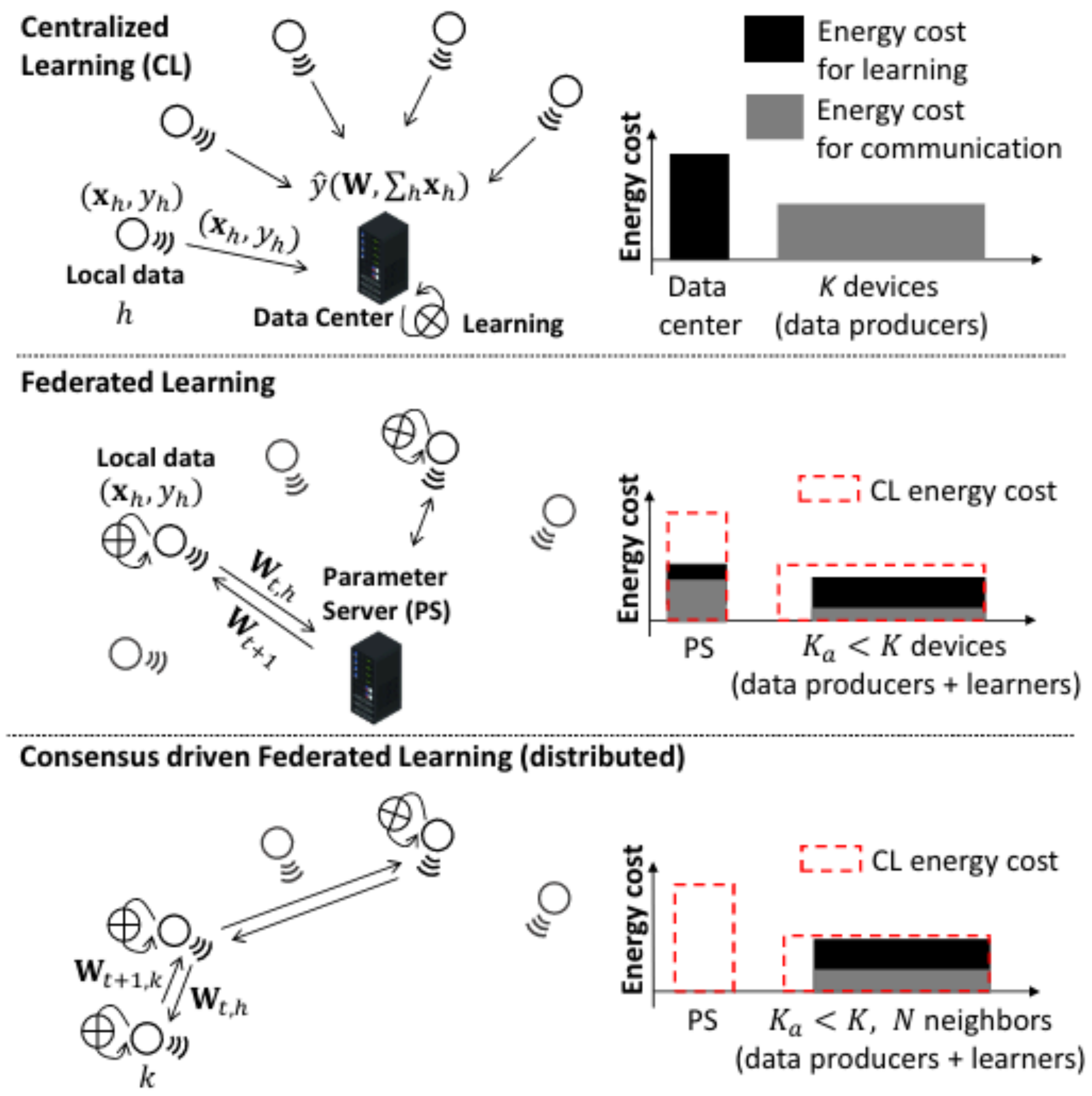
- PUE
- Number of rounds to reach target accuracy (and number of batches)



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

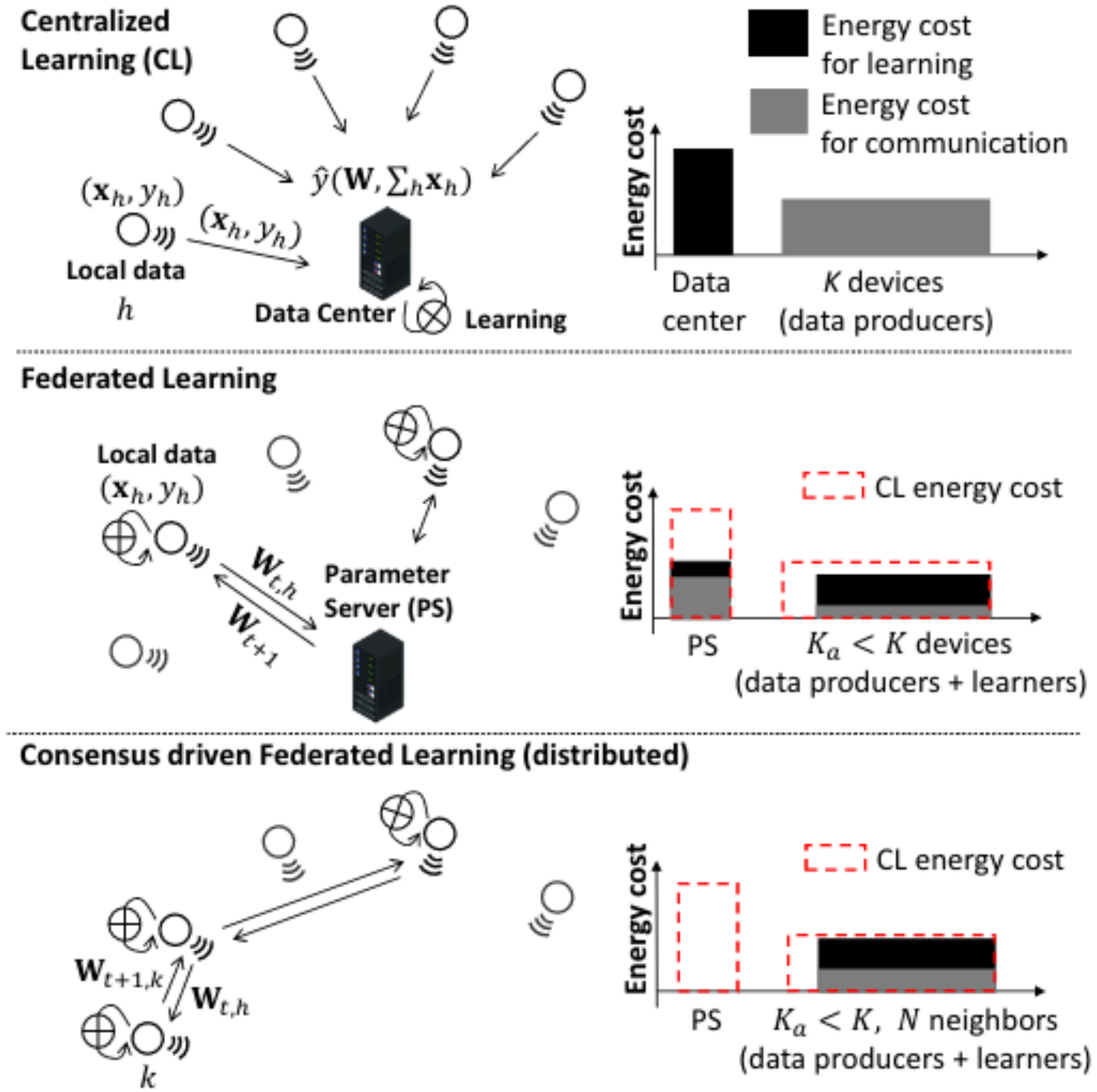
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size



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

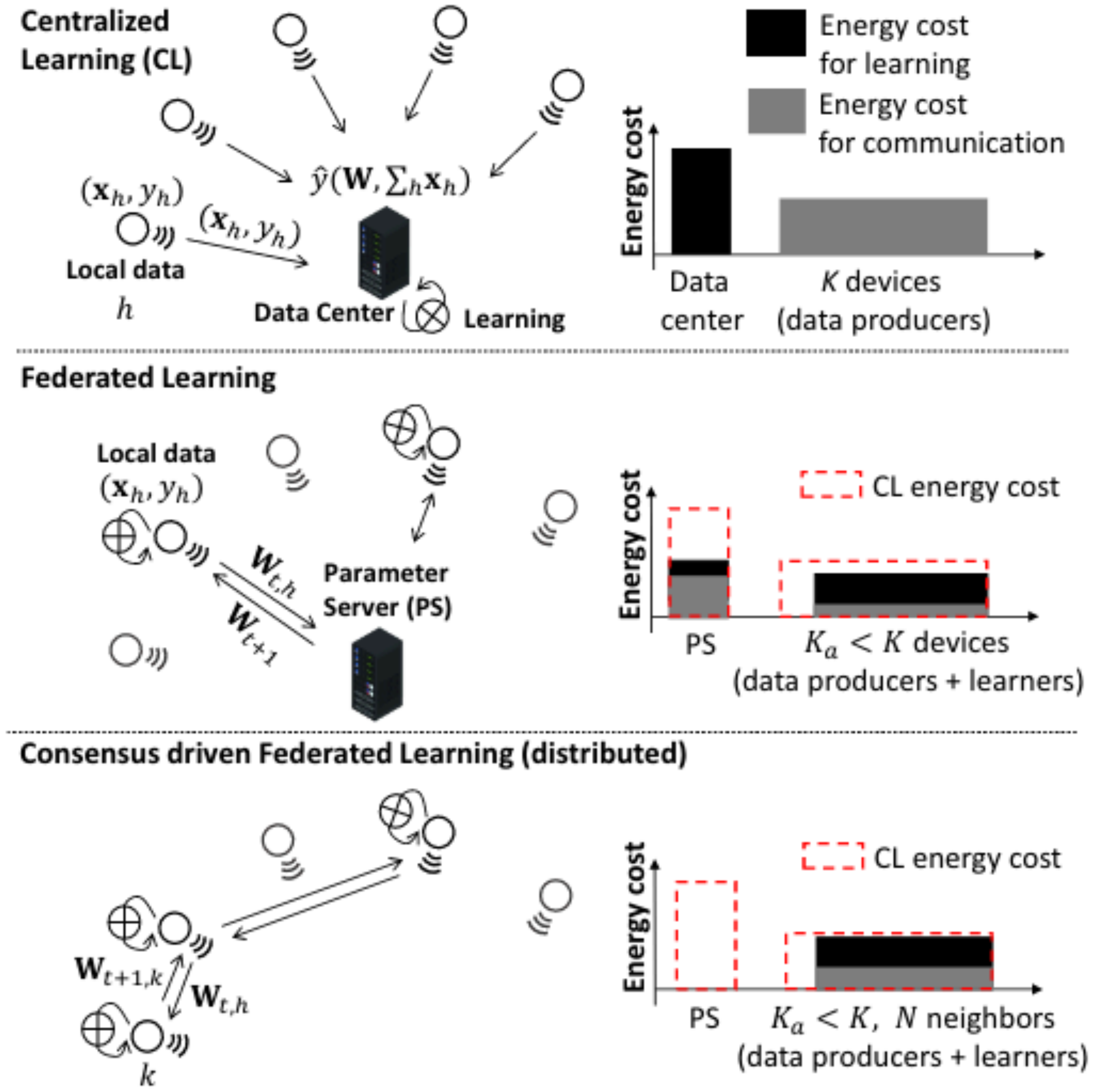
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)



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

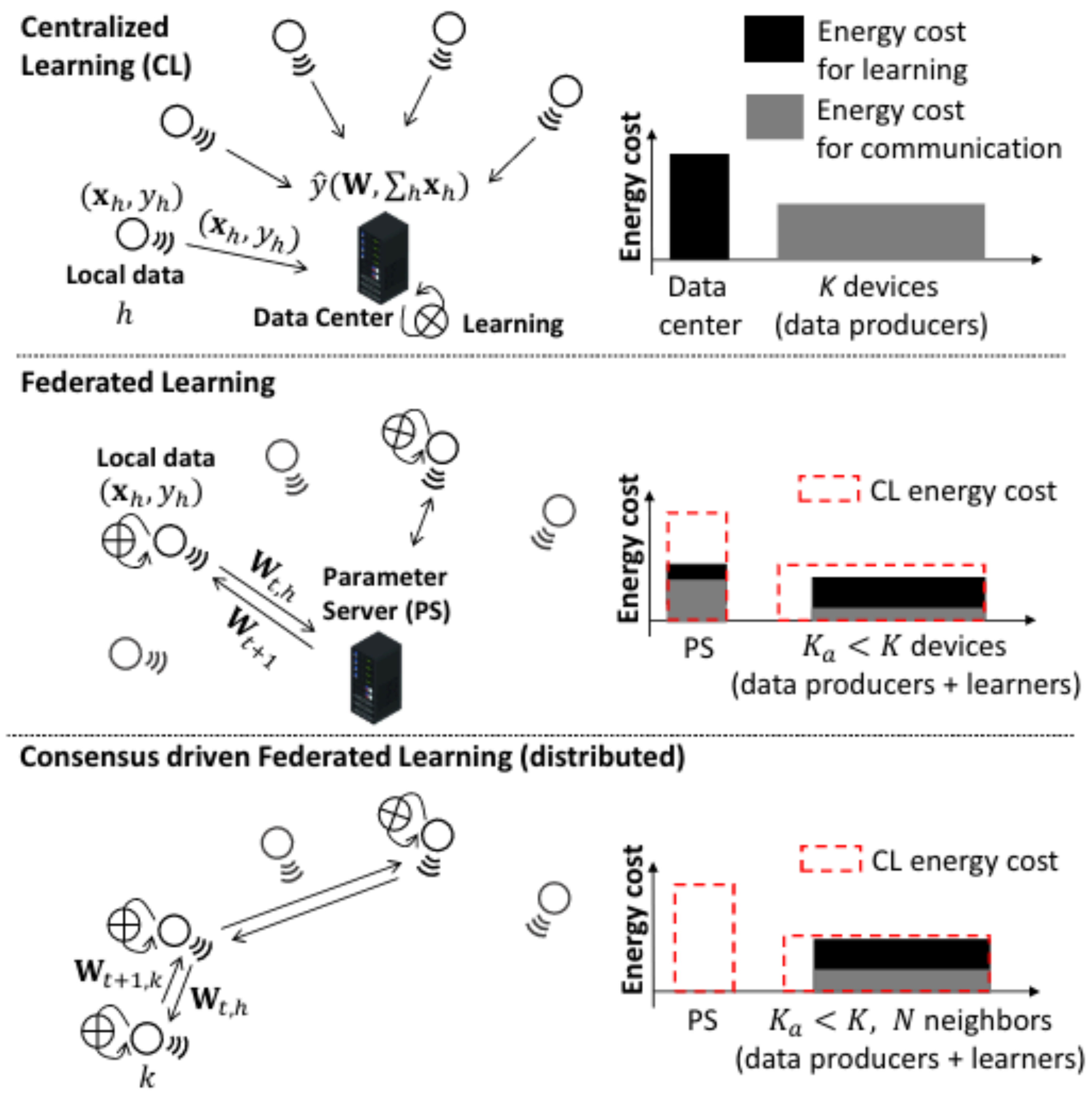
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not



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

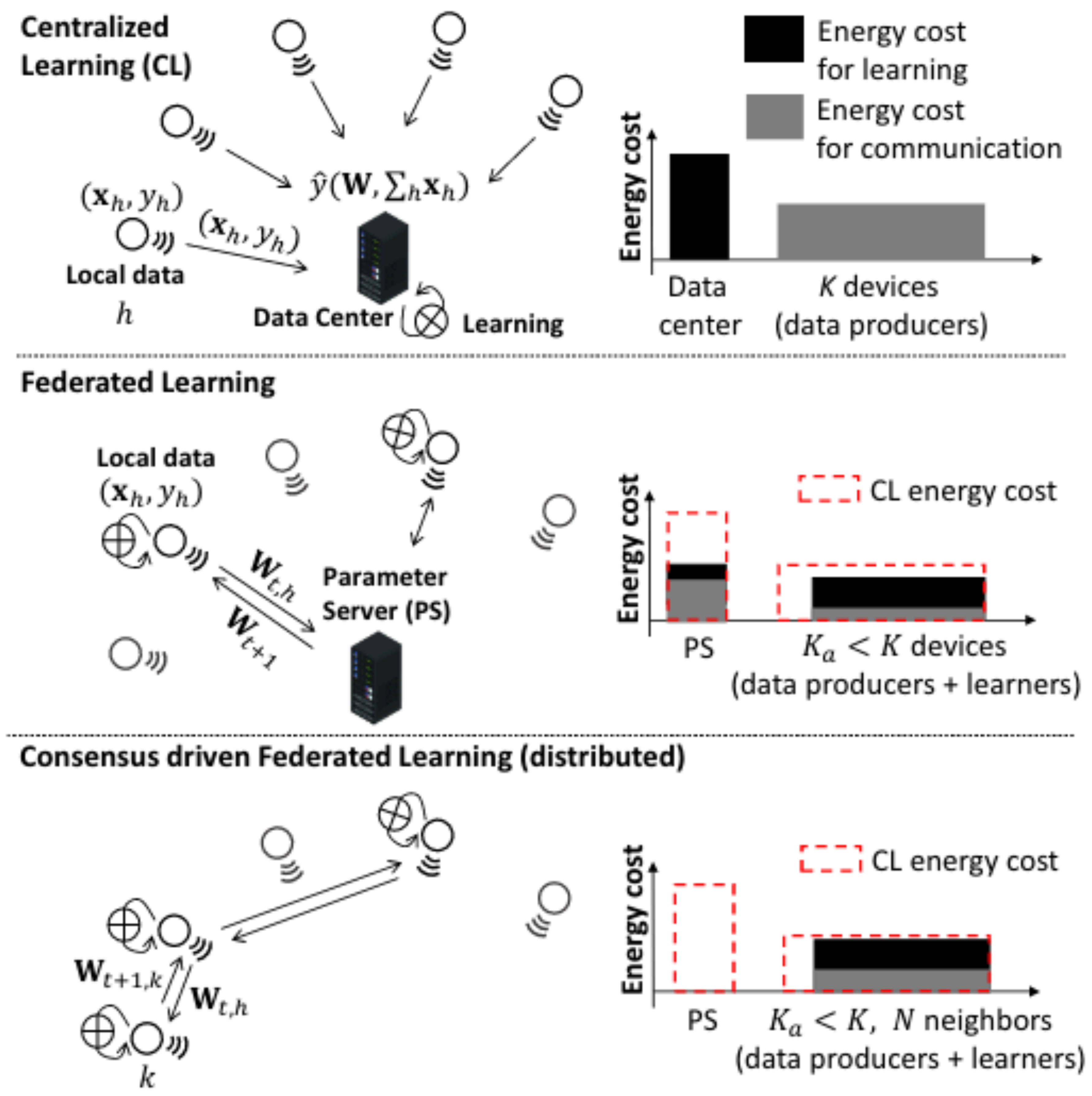
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)



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

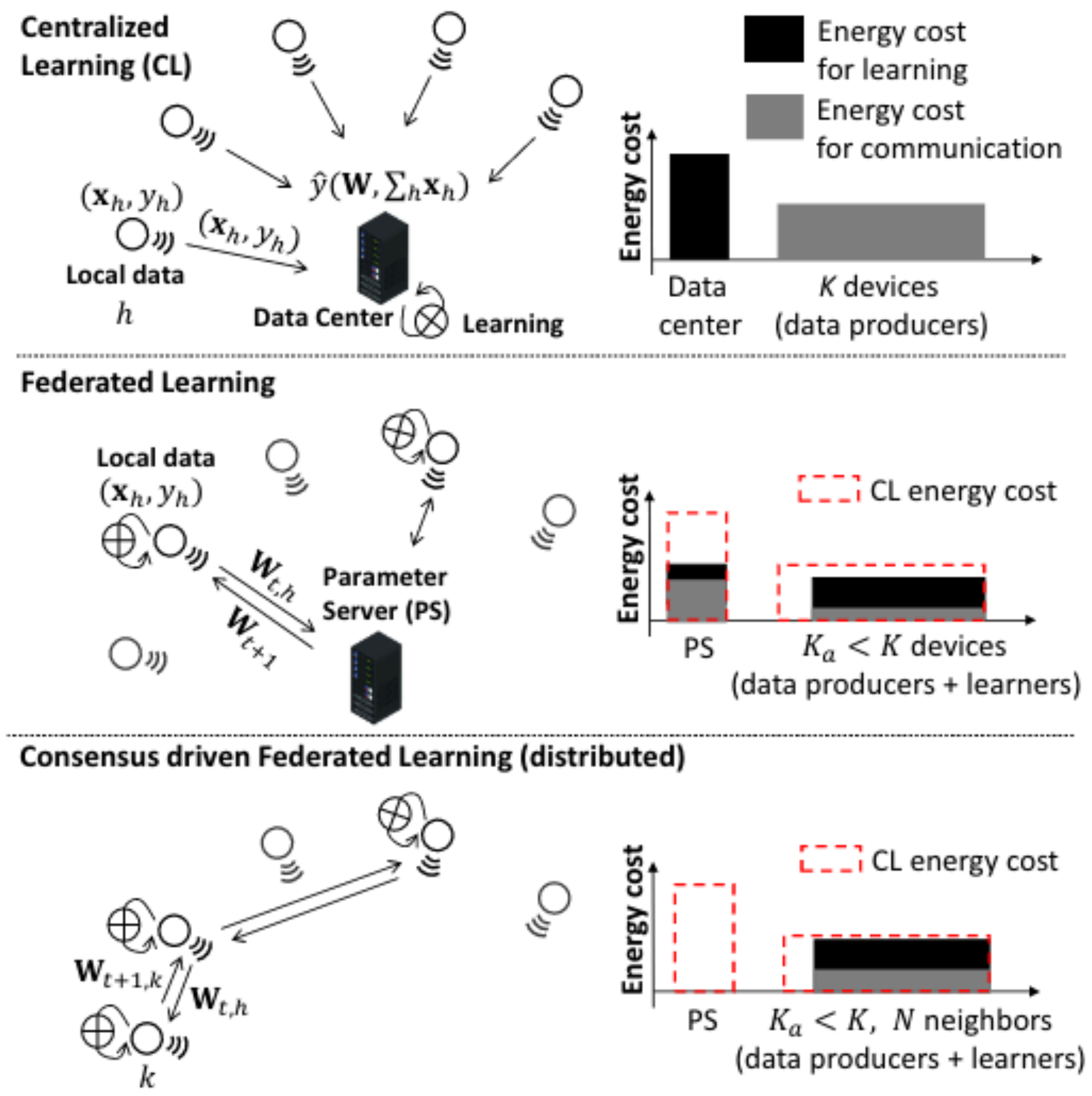
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners



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

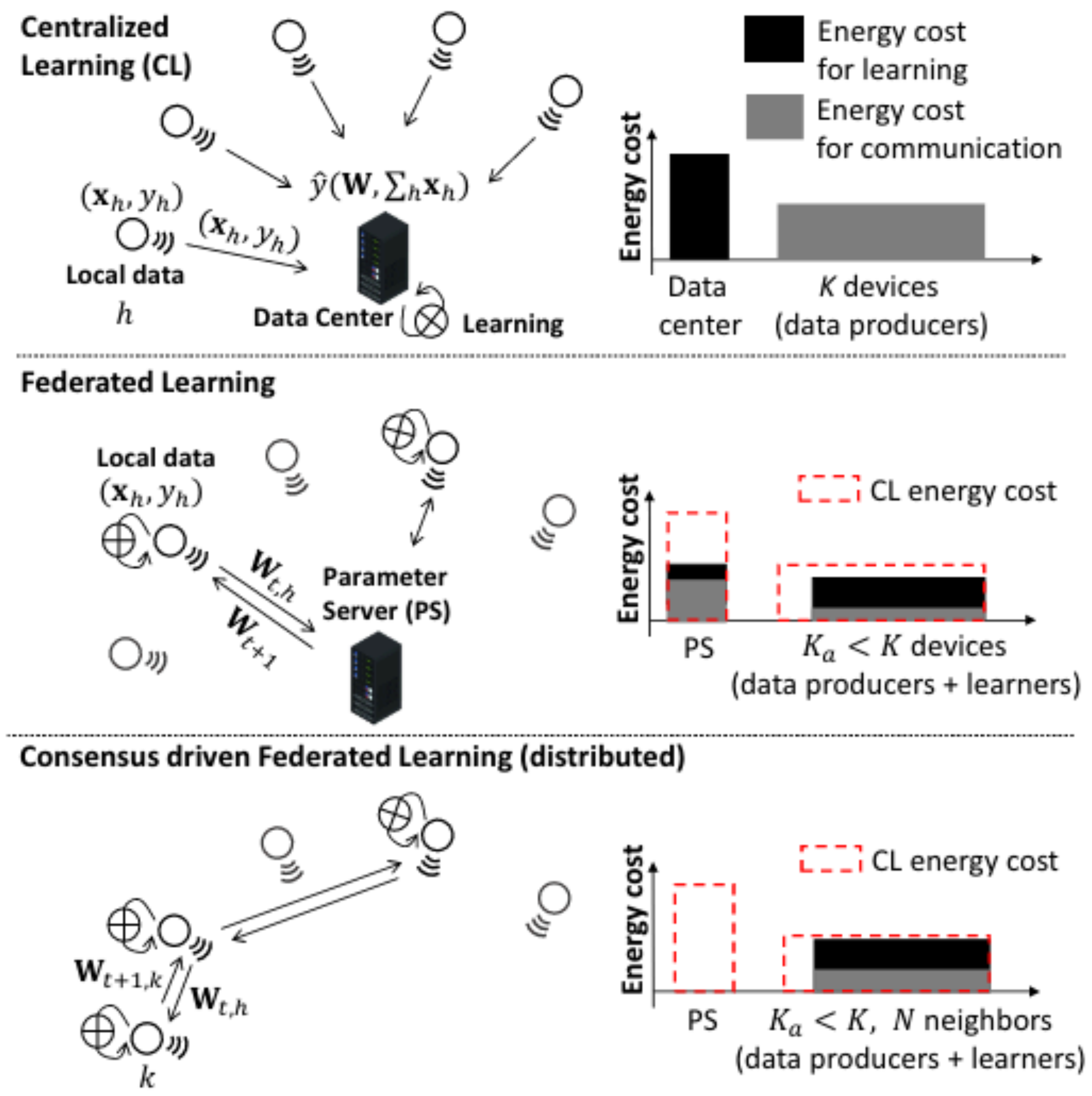
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners
- Relative energy efficiency



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

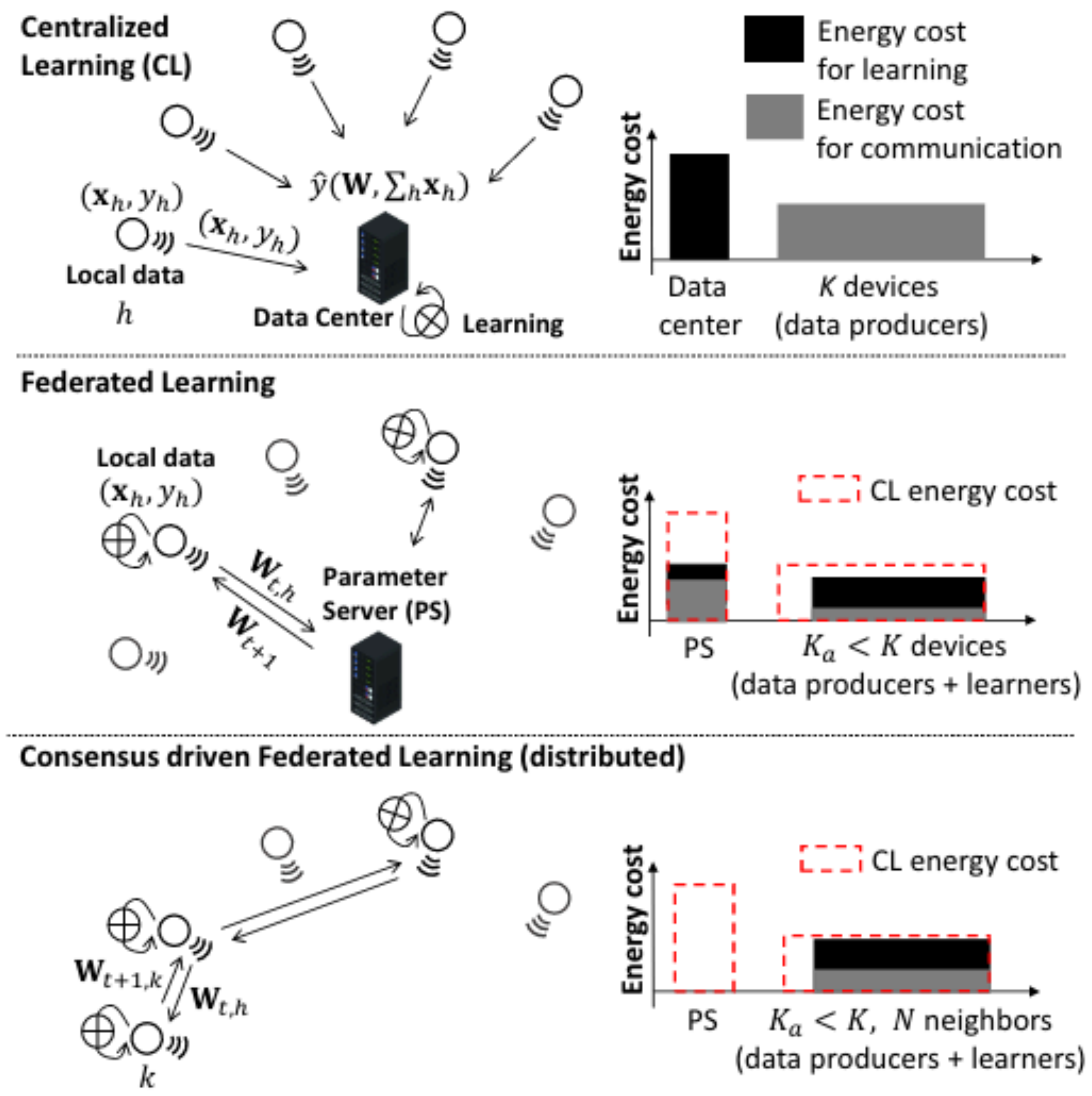
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners
- Relative energy efficiency
- Type of data transfer (uplink, downlink)



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

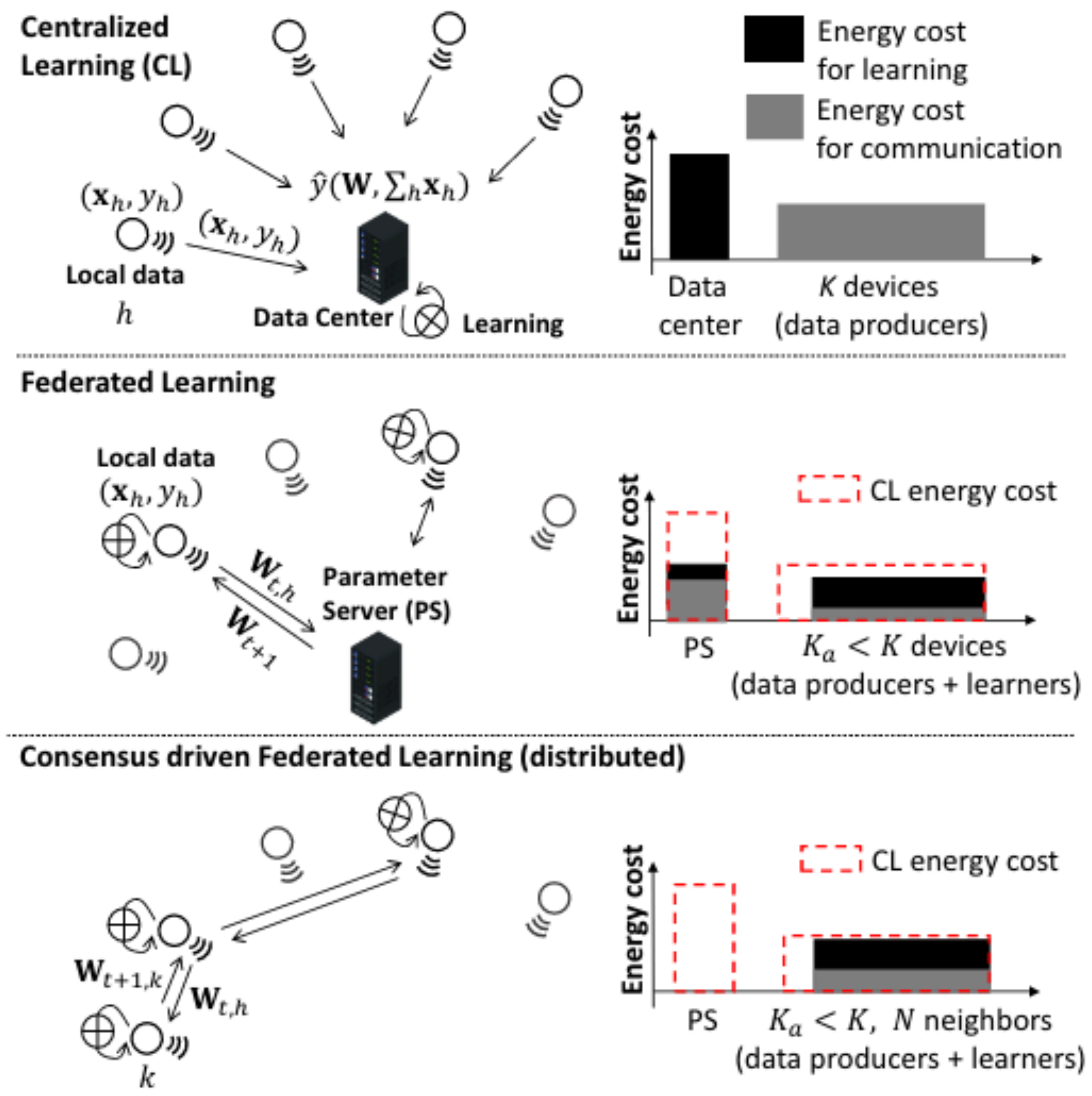
- PUE
- Number of rounds to reach target accuracy (and number of batches)
- ML model size
- Database size (local and total)
- IID data or not
- Number of training (if continual)
- Number of active learners
- Relative energy efficiency
- Type of data transfer (uplink, downlink)



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)

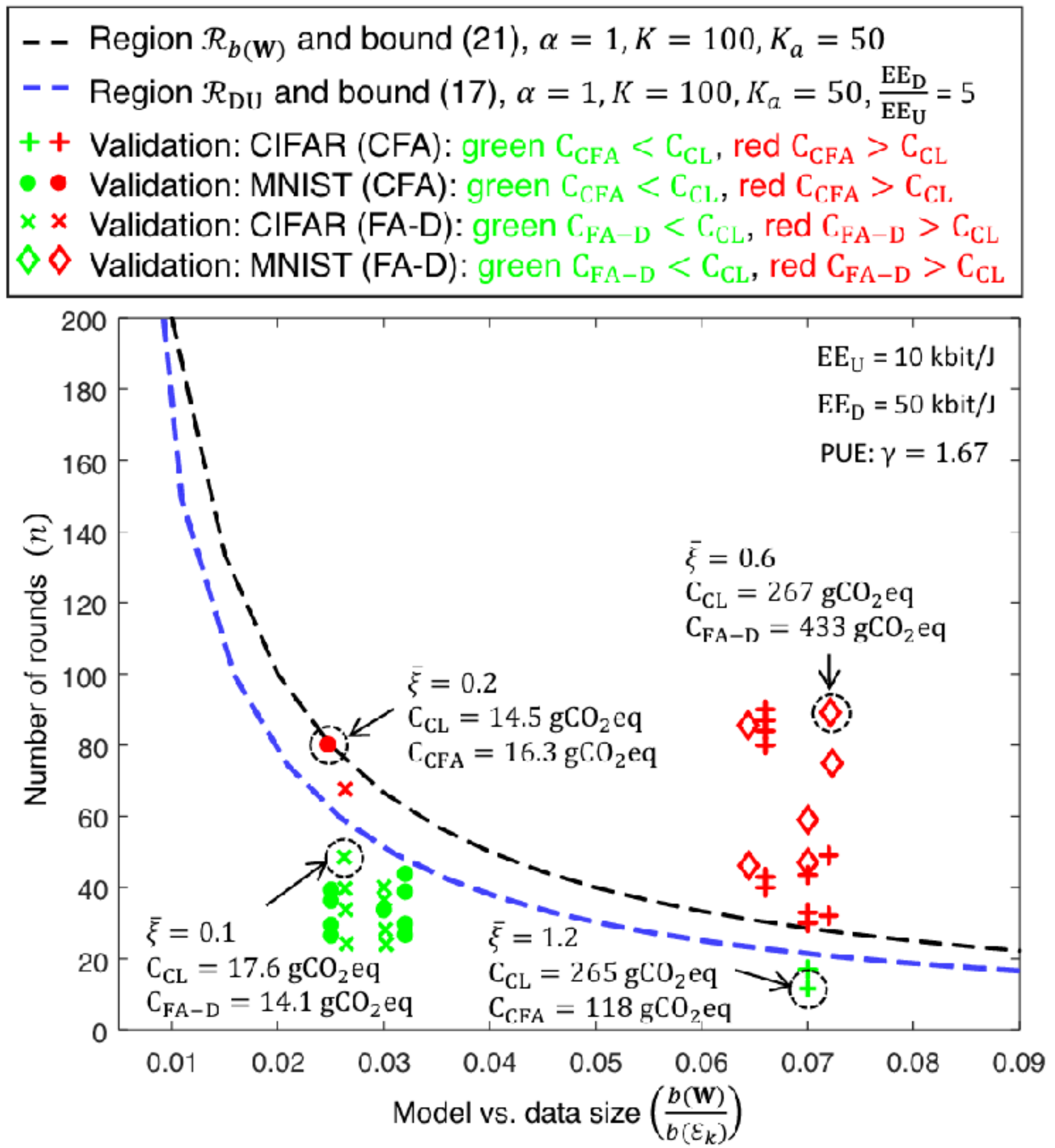


Rules for decision on which paradigm to use

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

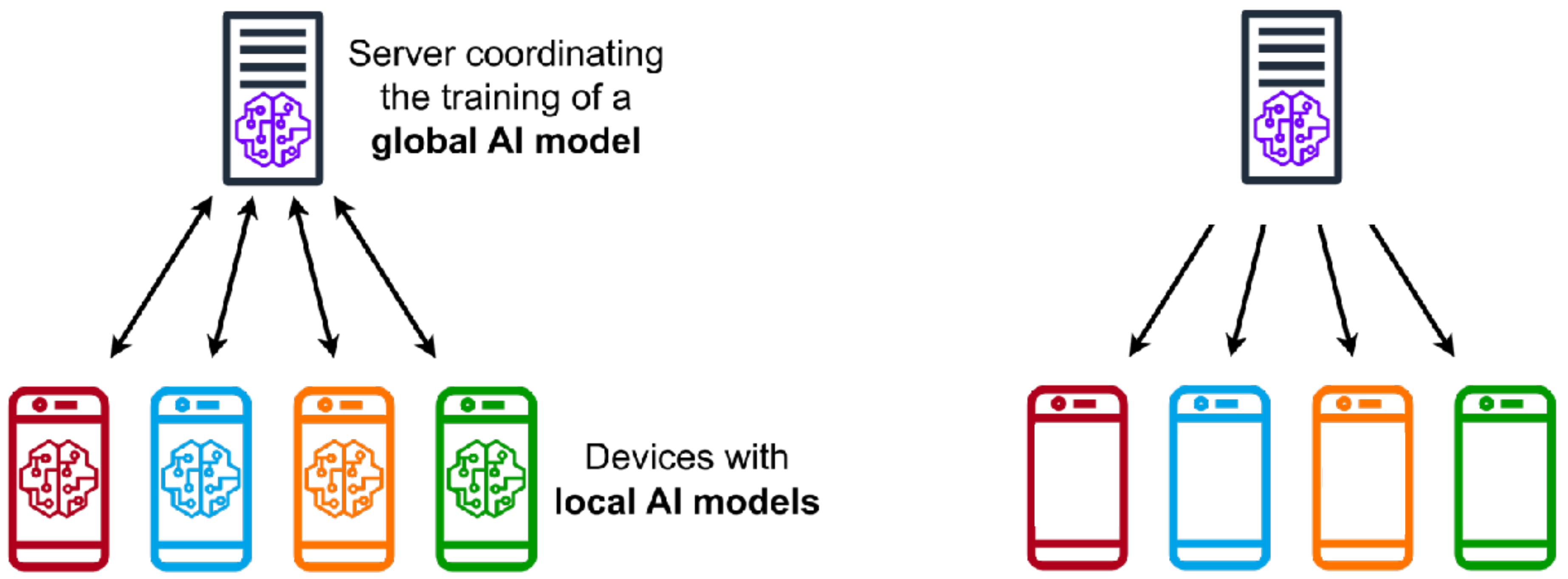


My objectives

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

My objectives

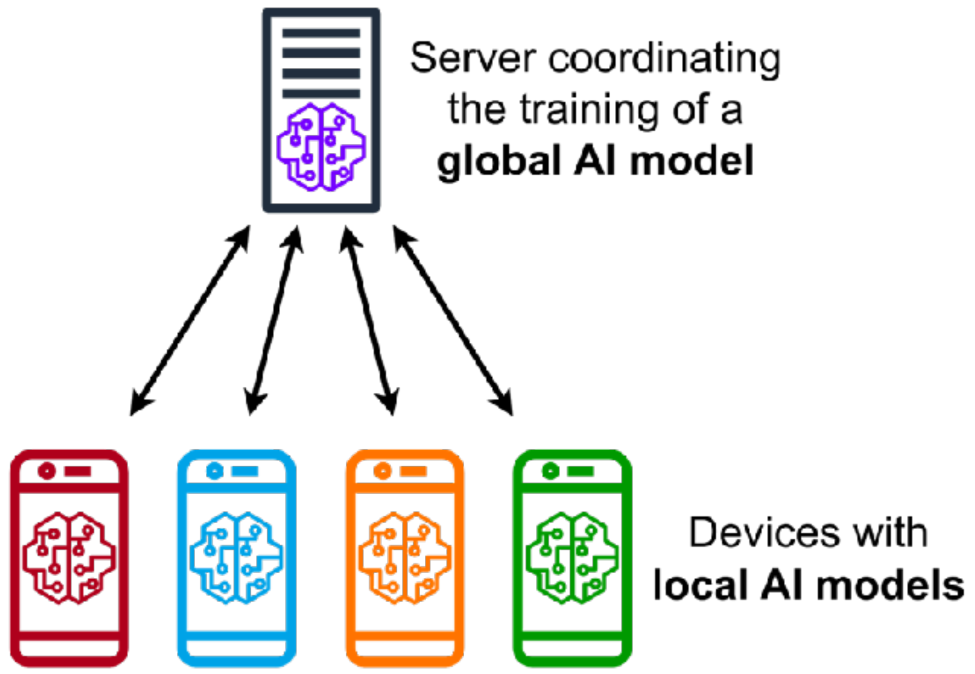
Benchmarking the performance and energy efficiency of AI accelerators for AI training

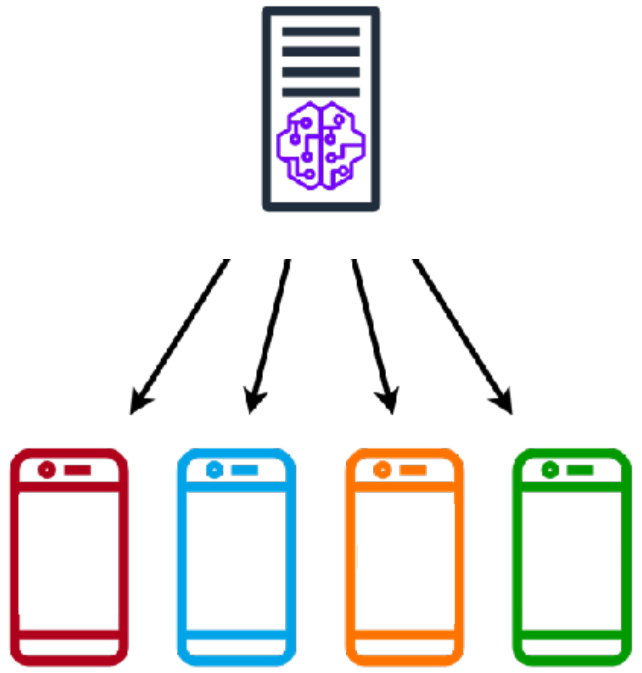


Federated Learning **versus** **Centralized Learning**

My objectives

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



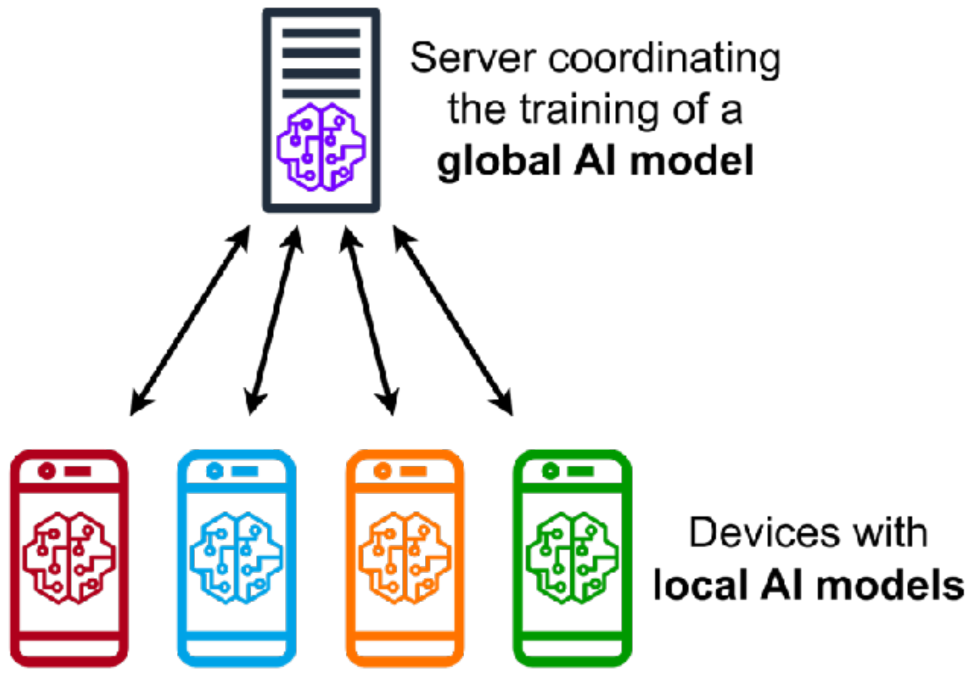


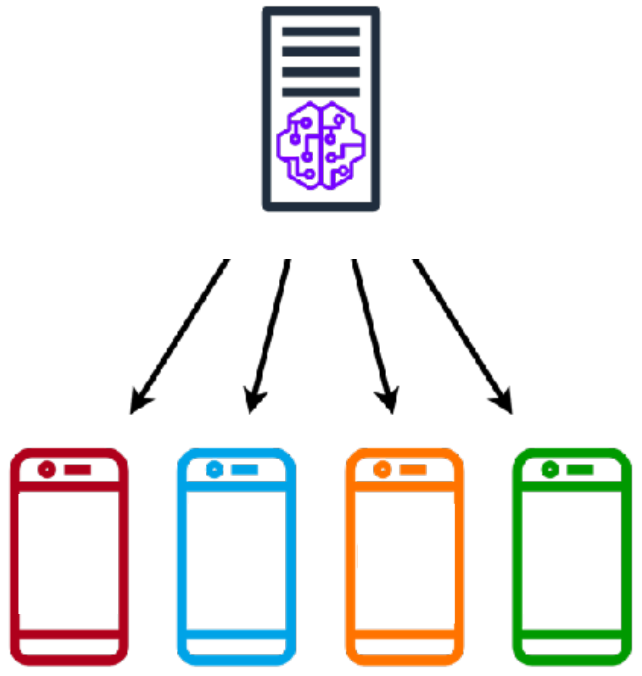
Federated Learning **versus** **Centralized Learning**

Continuous settings

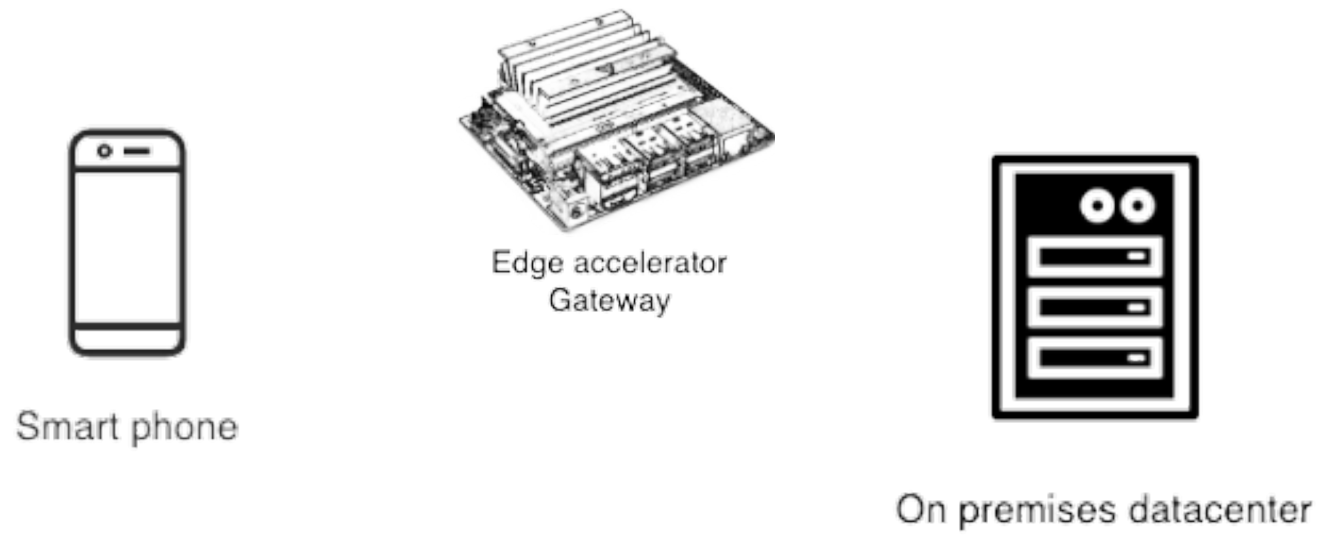
My objectives

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

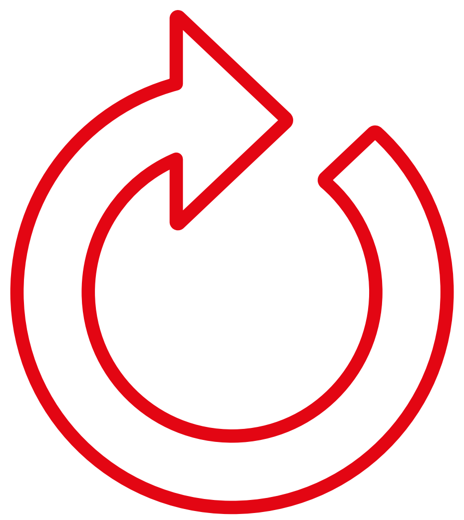




Federated Learning **versus** **Centralized Learning**



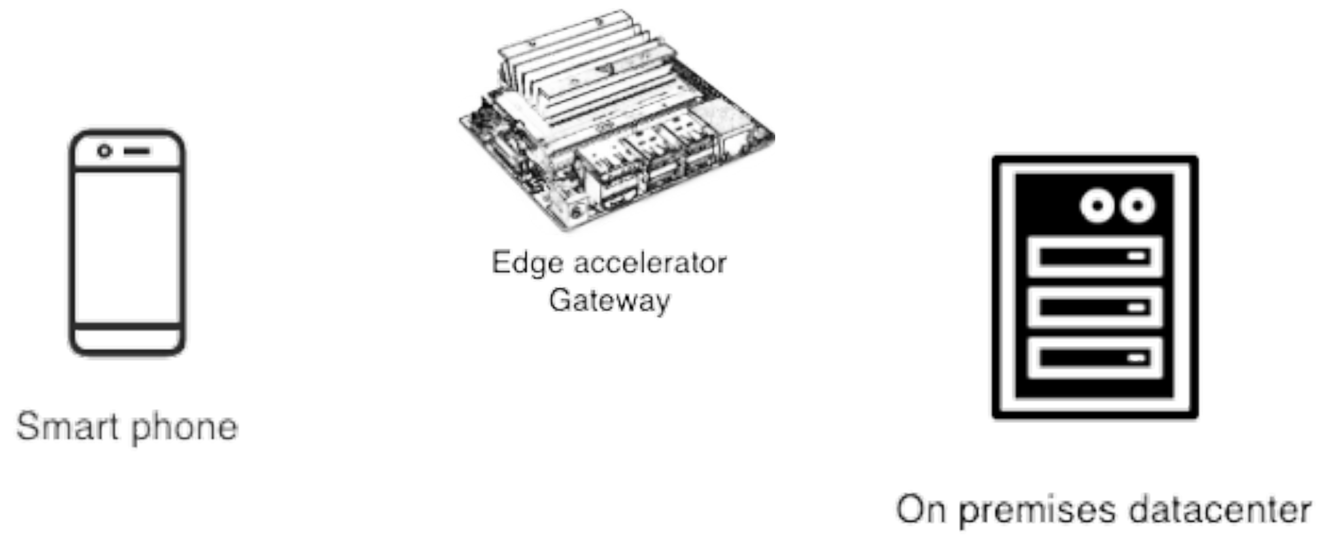
Rules on computer efficiency

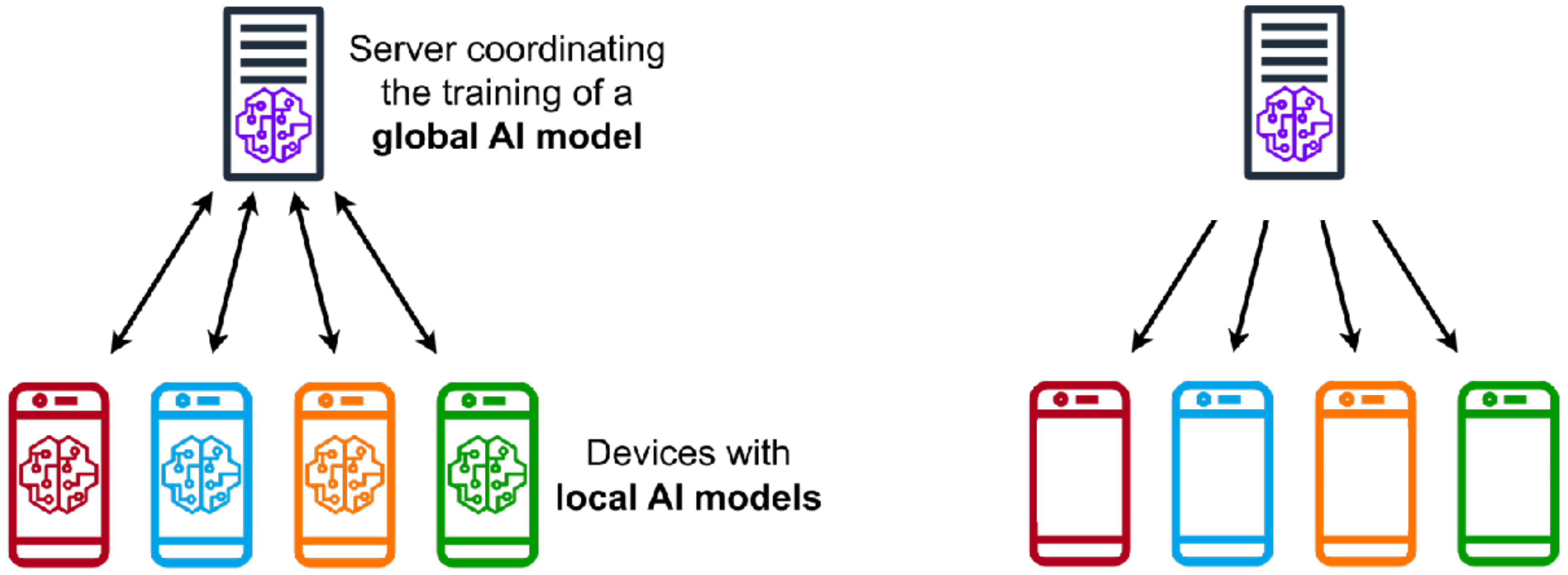


Continuous settings

My objectives

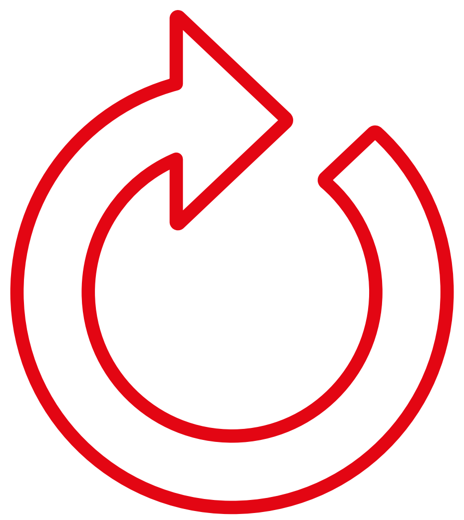
Benchmarking the performance and energy efficiency of AI accelerators for AI training





Rules on computer efficiency

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



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

Nvidia Jetson AGX Xavier (32Go)

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

Coral Dev Board (1Go)

Thank you for listening :)

Any feedback is welcome!