Wireless Network Intelligence @ the *EDGE*

Mehdi Bennis

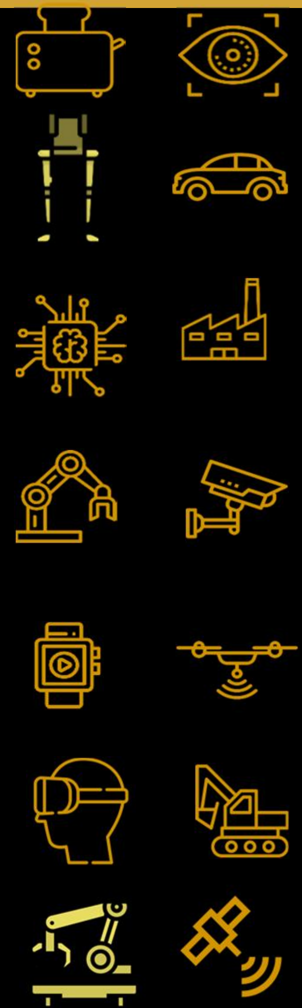
Associate Professor, Univ. of Oulu, FINLAND

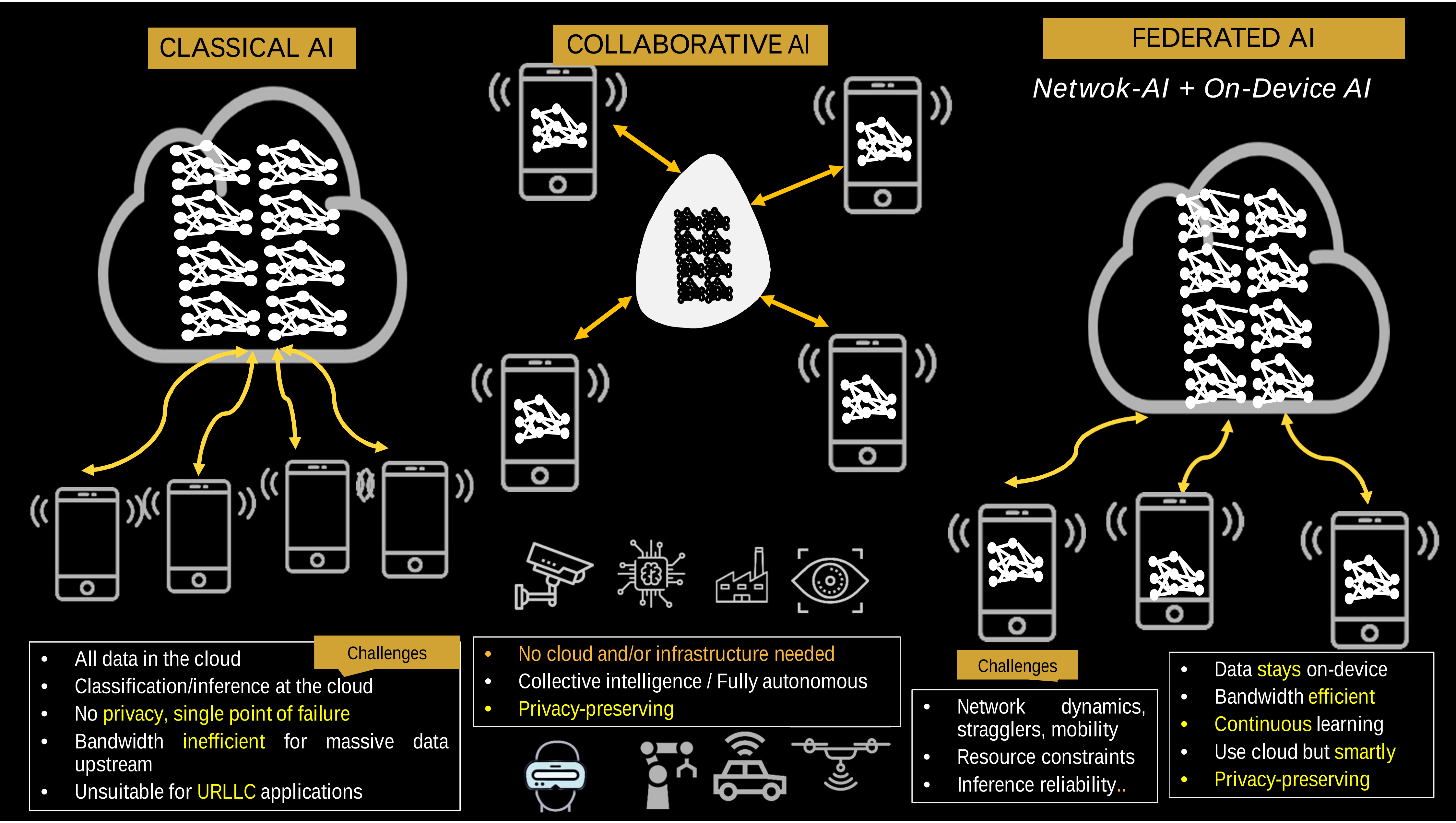
Beyond 5G Connectivity..

- 5G is and will remain as the **innovation platform** for **future connectivity** requirements.
  - Connectivity alone **insufficient** to realize 5G's full potential

- Proliferation of a **new breed** of autonomous devices **communicating**, **sensing** & **acting** within their environments.
  - **Massive** amount of data cannot be transmitted to the cloud for **training** & **inference!**

→ To solve this **massive scale challenge**, address **privacy**, **energy**, **latency** and **reliability concerns**, intelligence can no longer be confined to the central cloud, but *distributed* to the devices spearheading the <u>Wireless Edge Intelligence</u> vision.





CLASSICAL AI
COLLABORATIVE AI
FEDERATED AI
Netwok-AI + On-Device AI
Challenges
• All data in the cloud
• Classification/inference at the cloud
• No privacy, single point of failure
• Bandwidth inefficient for massive data upstream
• Unsuitable for URLLC applications
• No cloud and/or infrastructure needed
• Collective intelligence / Fully autonomous
• Privacy-preserving
Challenges
• Network dynamics, stragglers, mobility
• Resource constraints
• Inference reliability..
• Data stays on-device
• Bandwidth efficient
• Continuous learning
• Use cloud but smartly
• Privacy-preserving

ML/AI Changing Our Lives …*@ What Cost?*

- Today's AI *revolutionized* our lives, **successfully** *recognizes faces*, *diagnoses diseases*, *predicts rainfall*, consumer preferences & much more.
  - Thanks to more data and compute power

- Modern NN architectures are compute, space and power-hungry.
  - Cloud-Run: *Computationally intensive* à difficult to deploy on embedded devices with limited compute/memory constraints
  - Centralized + Offline training
  - Unreliable (do not reliably quantify prediction confidence)
  - Easy to fool changing slightly the input (GANs) -- adversarial examples
  - No privacy guarantees
  - Dominant paradigm: **Dumb** devices w/ *always-on cloud-connectivity*

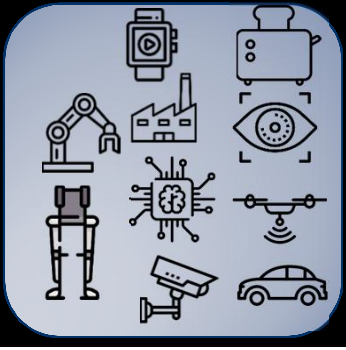
Machine translation

Face Recognition

AlphaGo

Medical Diagnosis

UNFIT for the new breed of intelligent devices & high-stake applications



ML for Communication (ML4C) – Current Focus

- **ML4C @ Physical layer**
  - Accurate knowledge of RF environment (propagation models, fault monitoring, etc)
  - Optimized use of RF environment (improved MCS, resource scheduling, spatial encoding schemes for MU-MIMO, reduced power consumption, etc)
  - Channel detection and decoding (data-driven useful for non well-established channel models)
  - Learn how to cancel FD self-interference

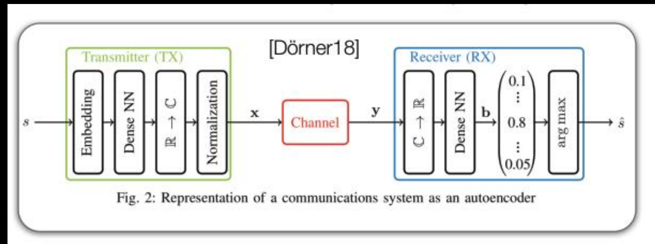
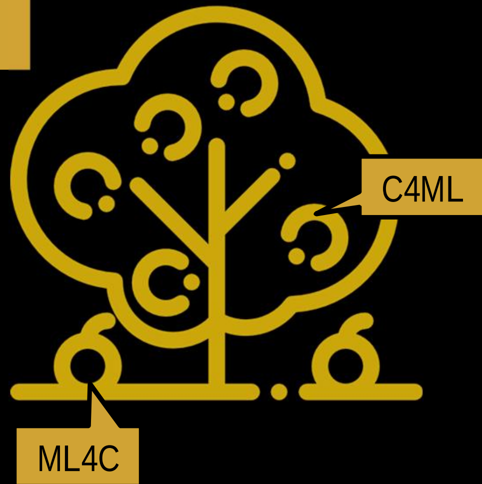


Fig. 2: Representation of a communications system as an autoencoder

Mostly data-driven + centralized + blackbox + best-effort

- **ML4C @ Network and Application layer**
  - Resource slicing, caching popular contents, routing, etc
  - Traffic classification, orchestration, virtualization, etc.
  - Spectrum sensing, Video streaming
  - Anomalies predictions, etc



What about C4ML? → Distributed Edge Intelligence

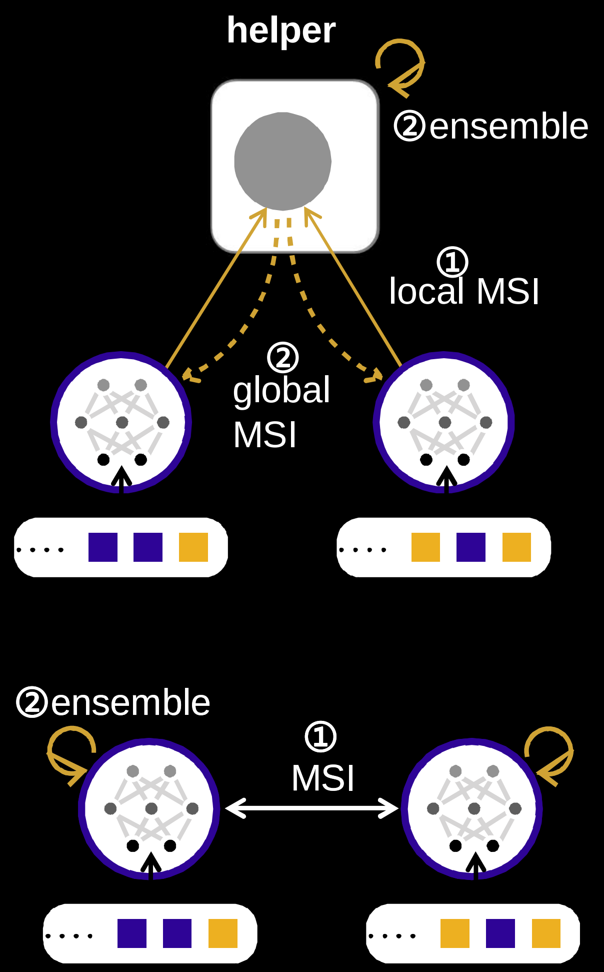
Wireless Edge Intelligence

Edge intelligence is a nascent research field which requires a major departure from centralized cloud-based training/inference/control approaches

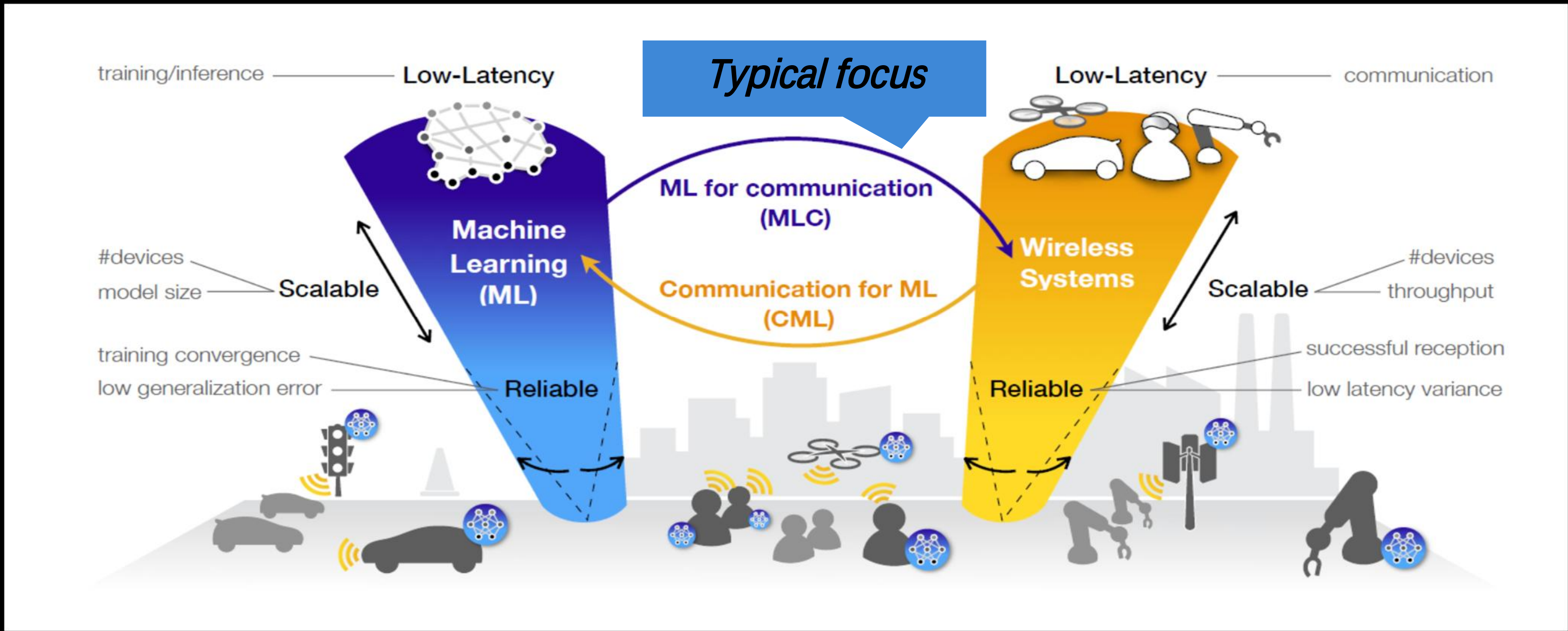
→ Towards a system design where edge devices communicate and exchange their learned models (not their private/raw data) to build/train a shared learned model subject to:

- Latency, reliability, accuracy, privacy constraints
- Memory/compute/power constraints
- Limited data & channel/network dynamics...

→ The ability of on-device AI to process data close to the edge yields low latency, privacy, reliability, efficient bandwidth usage.



Big Picture: When 5G Meets AI/ML



- Accuracy
- Latency
- Energy
- Reliability
- Scalability
- Capacity
- Sample Complexity

J. Park et al. "*Wireless Network Intelligence at the Edge,*" Proc. of the IEEE; https://arxiv.org/pdf/1812.02858.pdf

- How do resource-constrained devices **collectively train** a high-quality centralized model in a decentralized manner, for different NN architectures + with limited data
- How to carry out fast and reliable decision making under **risk** and **uncertainty**? How to model dynamics, **uncertainty** (DL ignores uncertainty), etc. // How to ensure **reliable** inference?..

6G Federated Learning for Reliable V2V

Use case: URLLC-V2X + distributed FL
Challenge: Network latency distribution needed!
Solutions:

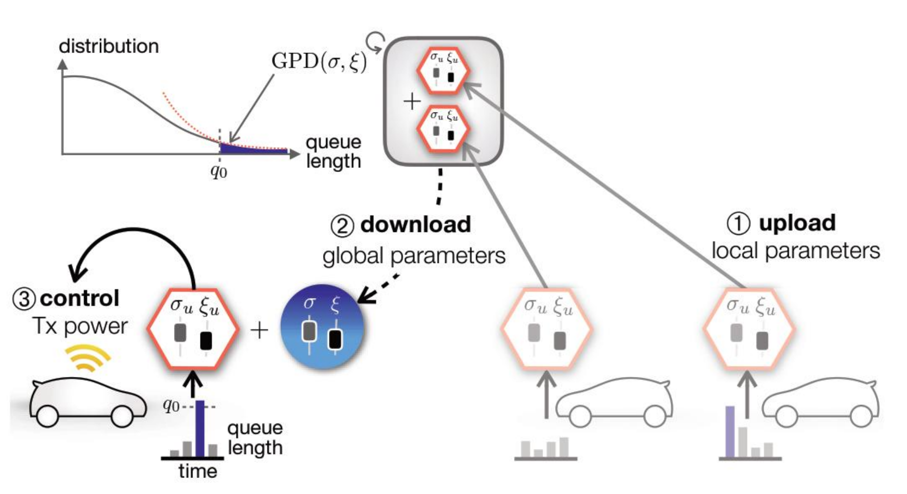
- Locally but lack of samples (latency↑)
- Remotely (RSU) but violate latency constraints (reliability↑ latency↑)
- Synchronous vs. asynchronous UL (latency↑) .

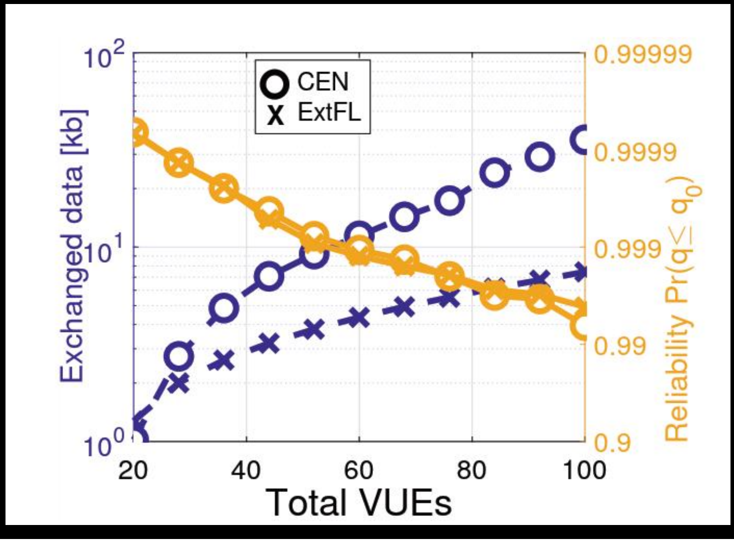
Key Idea:

- Instead of vehicles uploading their data to the cloud/RSU, every vehicle locally uploads its model to RSU + RSU does model averaging and brodcasts/multicasts to vehicles.
  à Model-driven ML

Benefits:

1. FL is a lower latency + Higher reliability enabler J
2. Works even during connectivity loss J !!



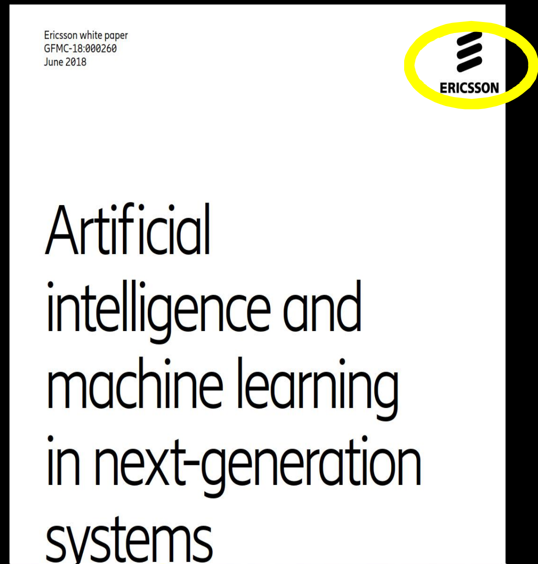


S. Samarakoon, et al, "Federated Learning for Ultra-reliable Low-latency V2V Communications," *in proc. of* IEEE GLOBECOM 2018, Abu-dhabi, UAE.

FL-Wireless Ramifications

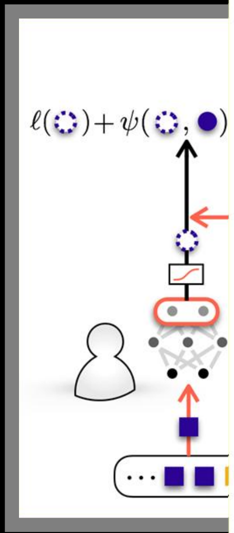
A learning model may have million parameters

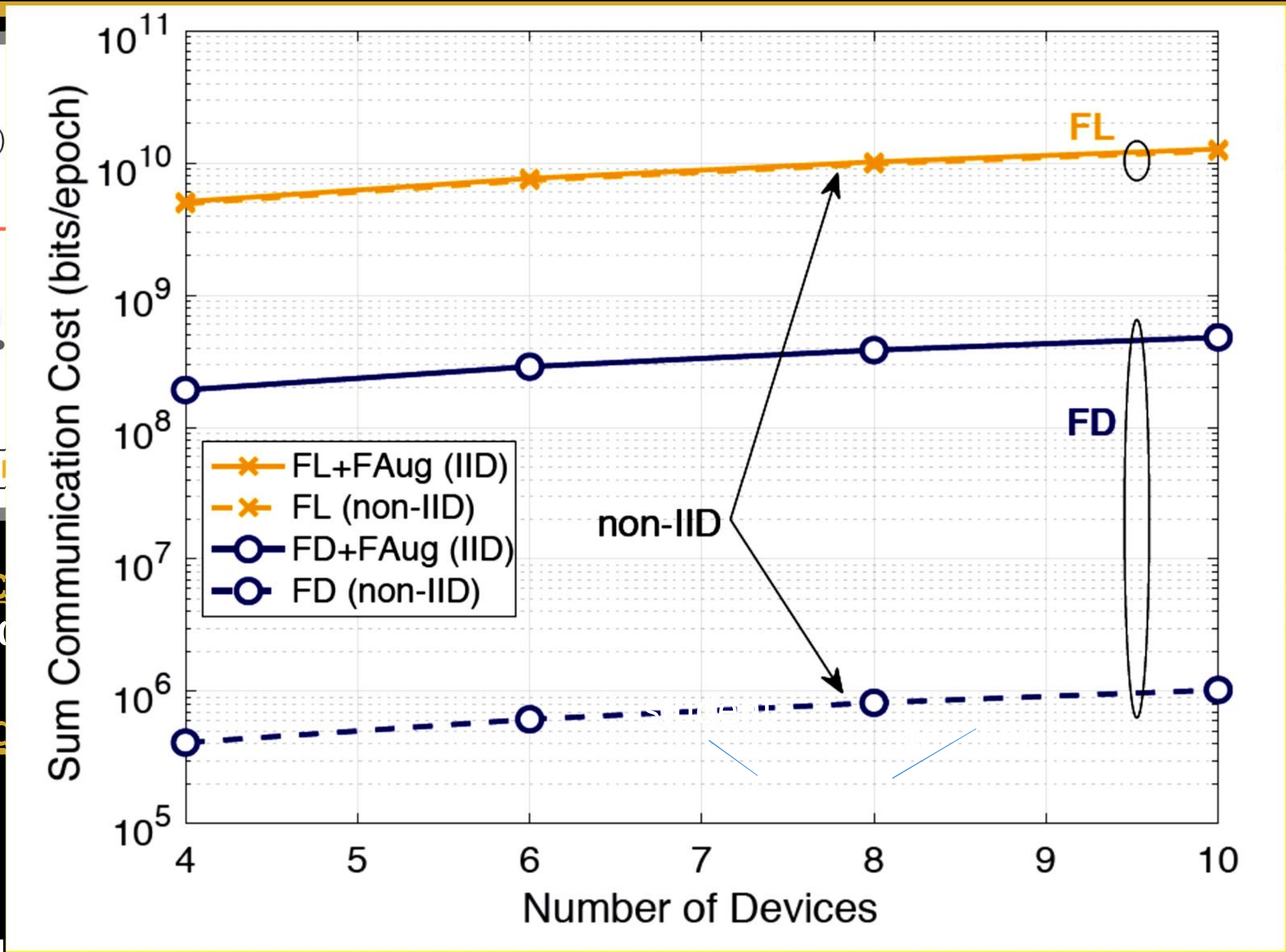
- Model updating is bandwidth consuming especially for 1000X edge devices
- Slowest node or straggler
- Adapt to asynchronous changes
- Moving nodes & noisy/interfered links
- Sample importance/Freshness
- Data quantity vs. <u>quality</u>
- Need to adapt to local dynamics
- Continual and lifelong learning
- Non-stationary input data
- Shared learning on confidential data
  - Blockchain, DLT (*)

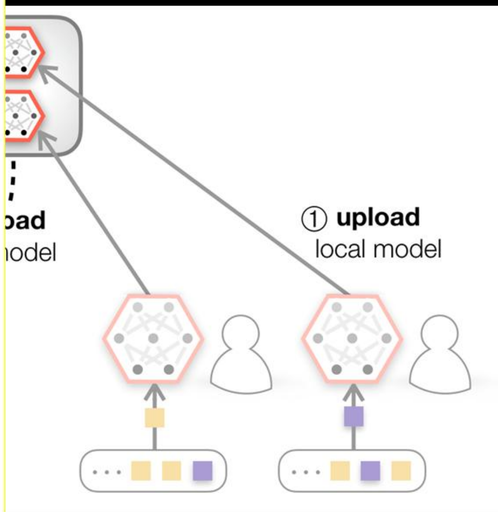


(*) Bennis et al. "on-device FL via Blockchain and its Latency Analysis," IEEE Comm. Letter, 2018

Federated Distillation vs. FL





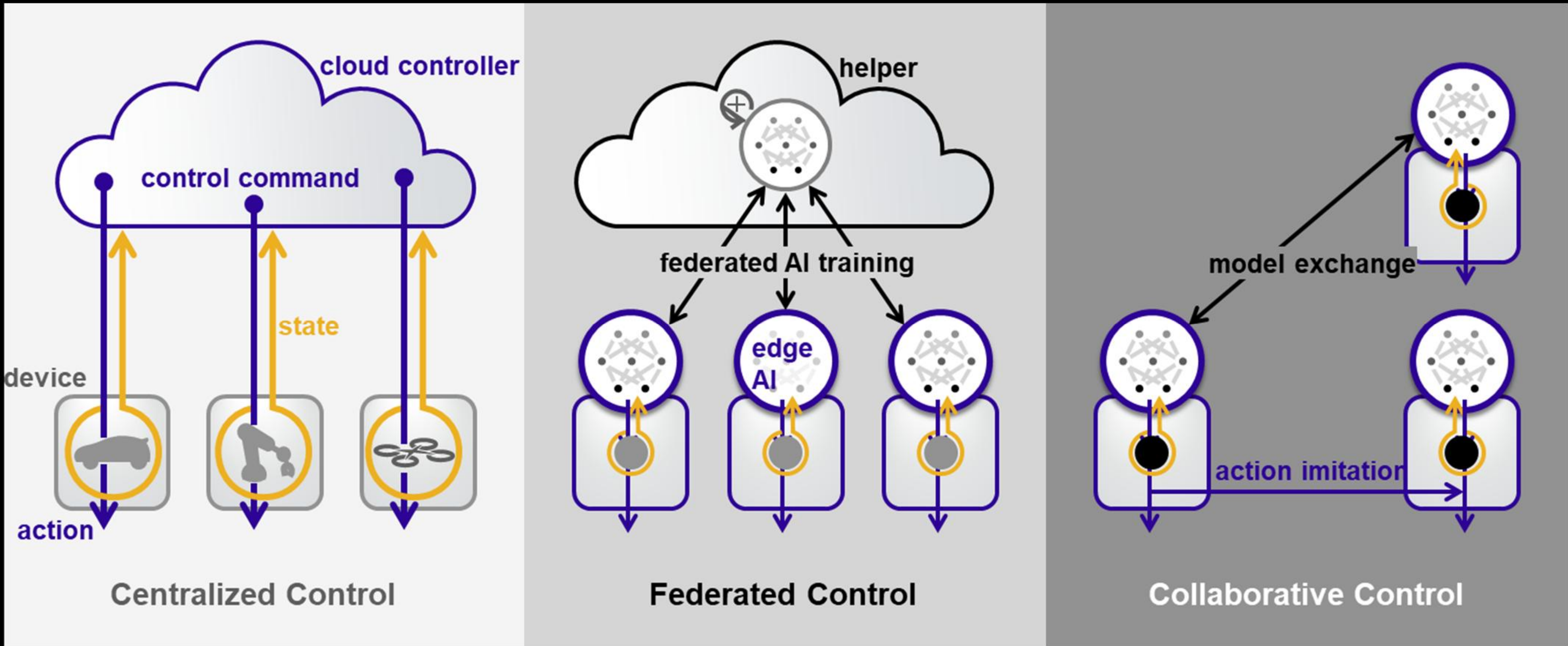


- Communic
  determined
  the model
- Computing

ad: model parameters
onal to the model size
model averaging

Bennis et al. "Federated Distillation and Augmentation under non IID private data," NIPS18, Montreal.

Data-Driven CONTROL



- Low-latency and reliable control over wireless
- From linear one-dimensional control à non-linear multi-dimensional control
- Risk-sensitive RL, distributional RL, inverse RL, imitation learning..

Parting Comments

- Distributed edge intelligence will unlock full potential of 5G (and beyond)
  - Preliminary results are highly promising!
- Lots remain to be studied at many levels and across many domains:
  - ü Architectural (data split, model split), Beyond DNN
  - ü Algorithmic, mathematical tools needed
    - ü Need for reliable and low-latency ML
  - ü Hardware-Algo codesign needed

→ Journey continues..