Edge Intelligence over Wireless: Present & *Future*

Mehdi Bennis
Professor, IEEE Fellow
Head of ICON
Univ. of Oulu, FINLAND

6G UNIVERSITY OF OULU

DESIRE6G

ADROIT

UNIVERSITY OF OULU

Outline

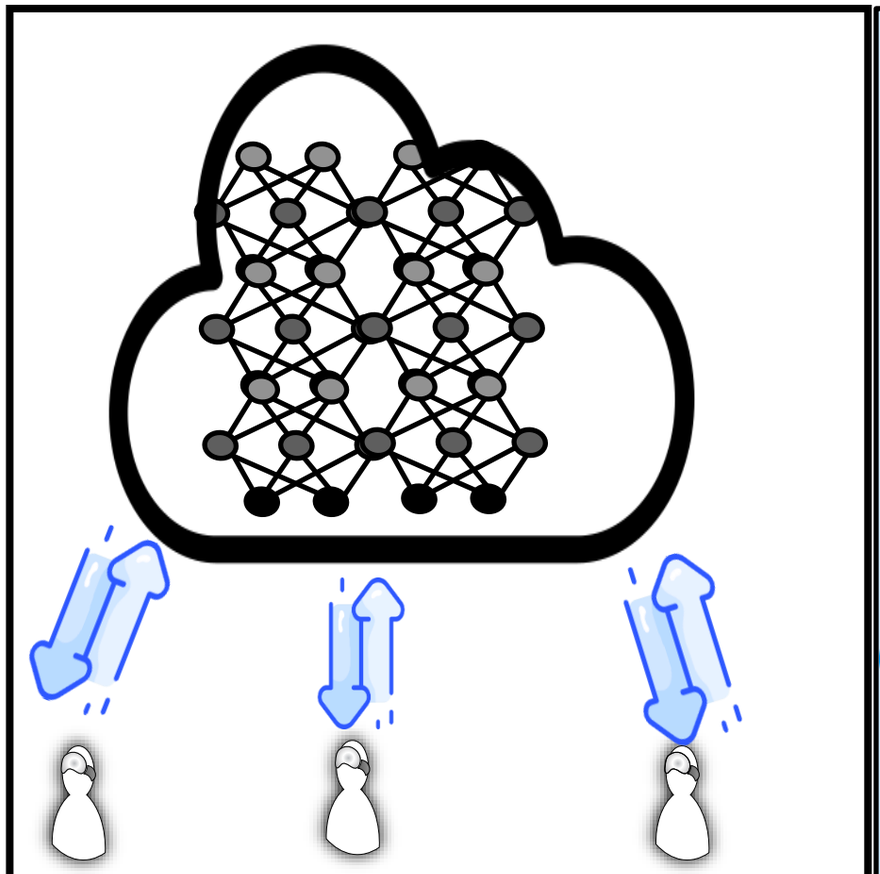
1. Motivation

2. Key Enablers

3. Selected Techniques & Applications

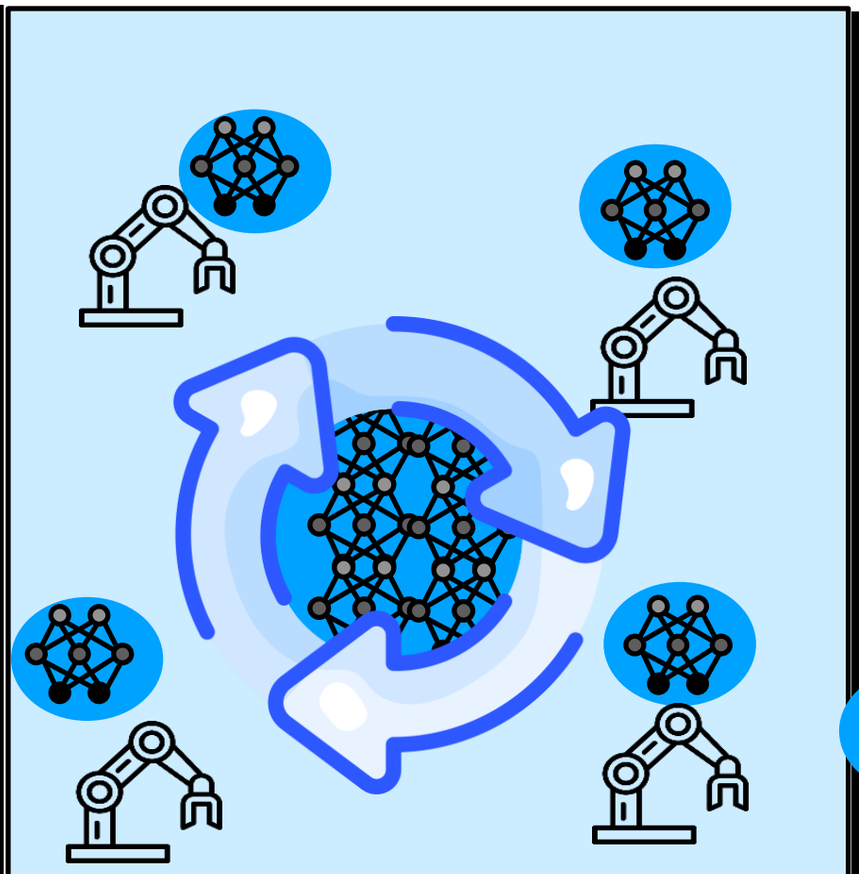
4. What else?

Centralized → Federated & Swarm/Distributed ML



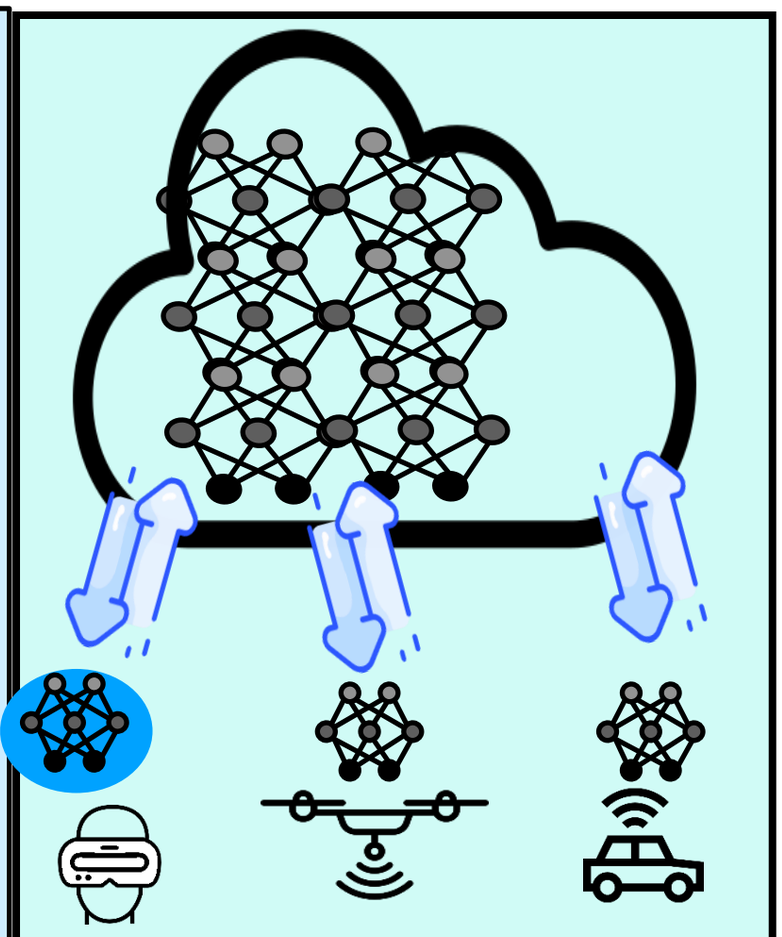
- No Privacy ✗
- High latency ✗
- Bandwidth inefficient ✗

Classical AI



- Collective intelligence / Fully autonomous ✓
- Scalable ✓
- Privacy-preserving ✓

Swarm AI



- Bandwidth efficient ✓
- Use cloud but smartly ✓
- Privacy-preserving ✓

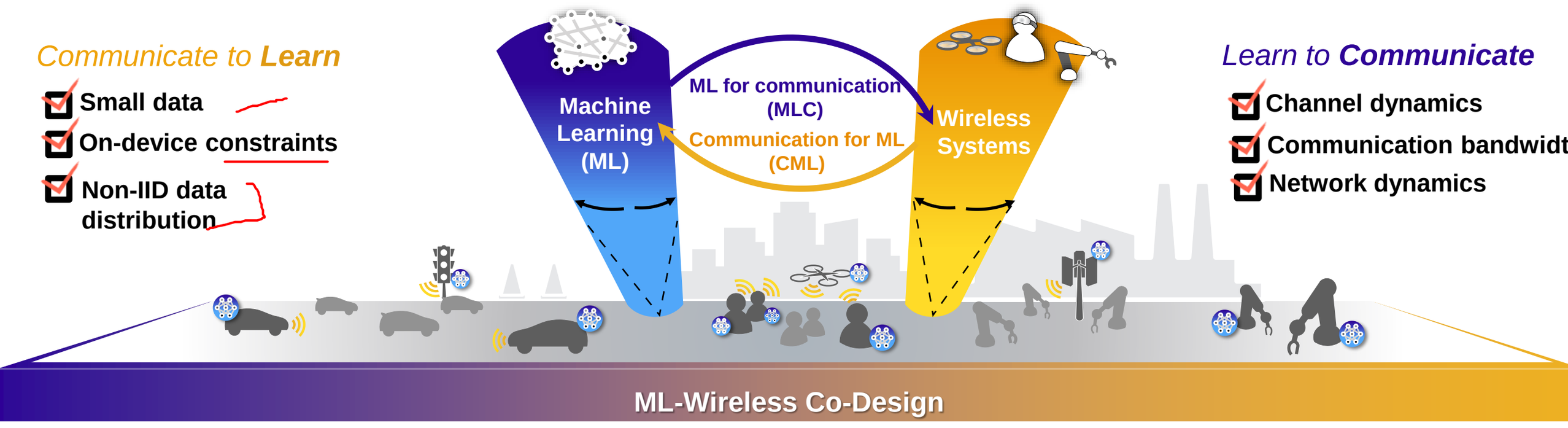
Federated AI

NOKIA HUAWEI SAMSUNG intel PHILIPS

Proliferation of intelligent devices & mission-critical applications at the network edge cannot be operated with centralized and best-effort ML

*Communicate to **Learn***

- ☑ Small data
- ☑ On-device constraints
- ☑ Non-IID data distribution



*Learn to **Communicate***

- ☑ Channel dynamics
- ☑ Communication bandwidth
- ☑ Network dynamics

Communication-efficient, low-latency, reliable and scalable (i) training; (ii) inference; (iii) control

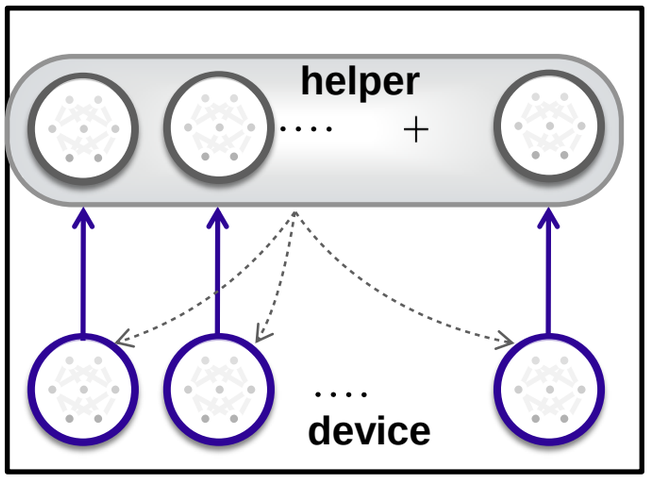
Sol.

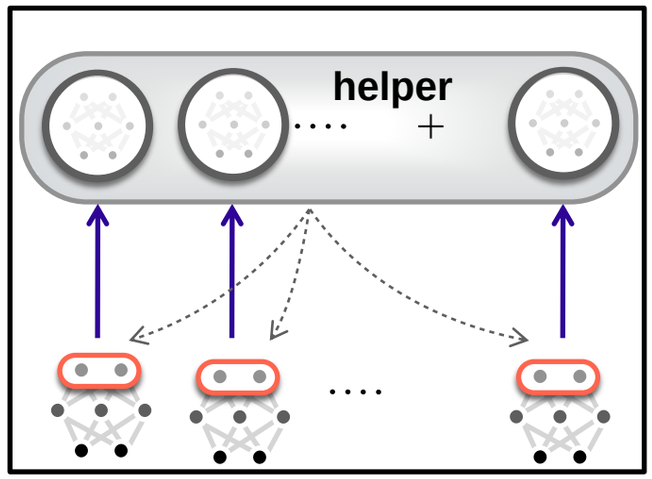
- **Federated Learning**
- **Serverless FL**

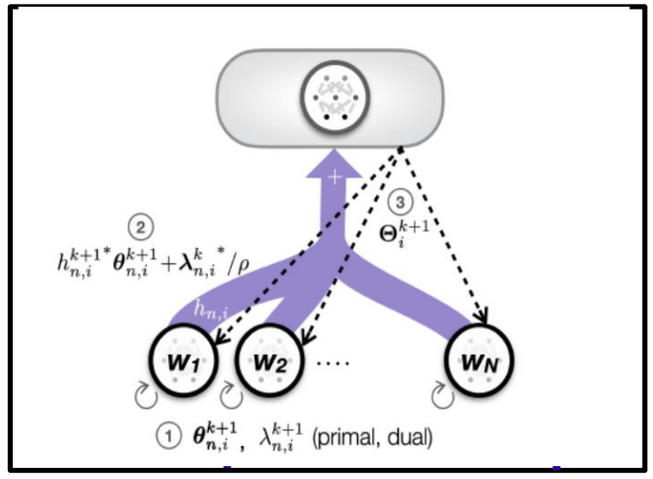
- **Federated Distillation**
- FL after Distillation

- Over the air aggregation
- Analog vs. Digital

Type



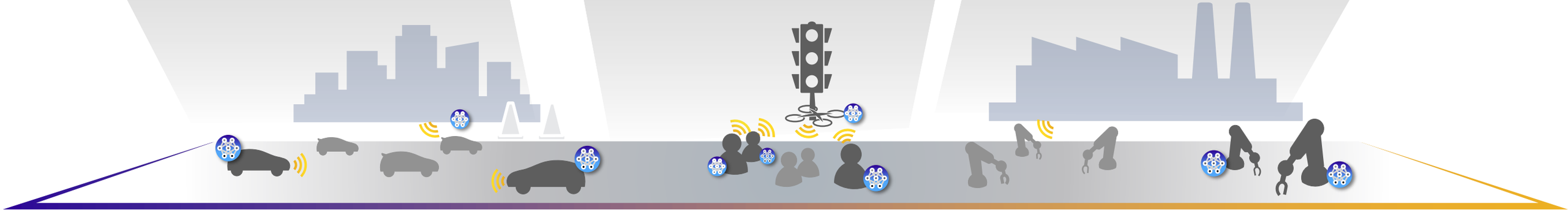




Model parameter exchange

Model Output exchange

Wireless-ML Codesign



ExtFL = FL + Extreme Value Theory

Extreme Queue Length FL for Vehicular URLLC Power Control

Problem. Minimize vehicular user equipment (VUE)'s avg. uplink power, subject to each VUE's **queue length reliability** $\Pr(Q > q_{\text{th}}) \leq \epsilon$

- Following **extreme value theory (EVT)**, an extremely large queue length is characterized by the shape and scale parameters of the generalized Pareto distribution (GPD)
- Utilizing **FL with EVT (ExtFL)**, vehicular user equipments collectively predict the GPD parameters
- ExtFL **reduces communication overhead** while achieving the same queue length reliability, compared to a centralized direct queue length distribution exchanging baseline (CEN)

Sample

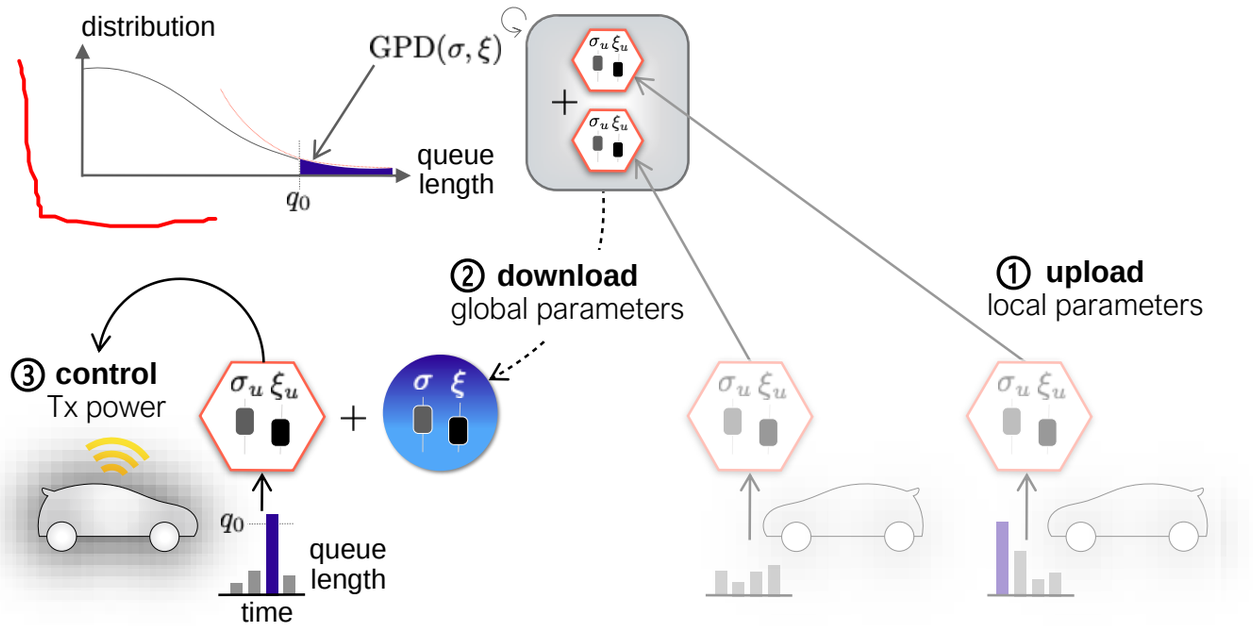
Empirical distribution

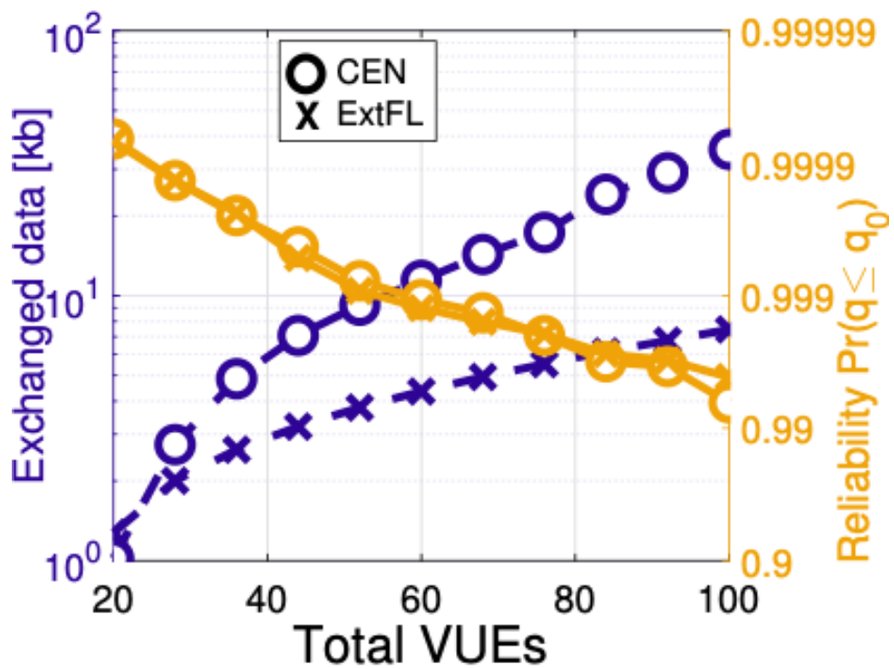
$$\min_{\sigma,\xi} -\frac{1}{N}\sum_{n}\log\left(G(x_n,\sigma,\xi)\right)$$

Scale parameter

Shape parameter

Extreme Value Theoretic FL (ExtFL)





S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "**Distributed Federated Learning for Ultra-Reliable Low-Latency Vehicular Communications**," *TCOM'20*

Beyond Federated (server-based) Learning

Group ADMM (without any central entity)

- Idea. Exploiting ADMM for faster training convergence without any central entity

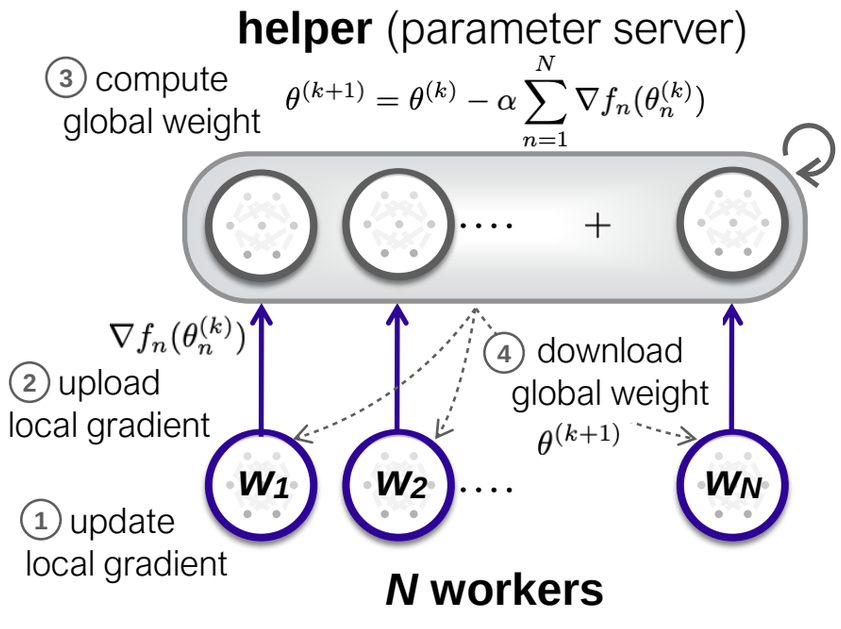
1) **Head** devices update **primal** variables (weights) in parallel
2) Each **head** device transmits the weights to its (two) **neighboring tail** devices
3) **Tail** devices update **primal** variables in parallel
4) Each **tail** device transmits the weights to its **neighboring head** devices
5) Each device updates its dual variable

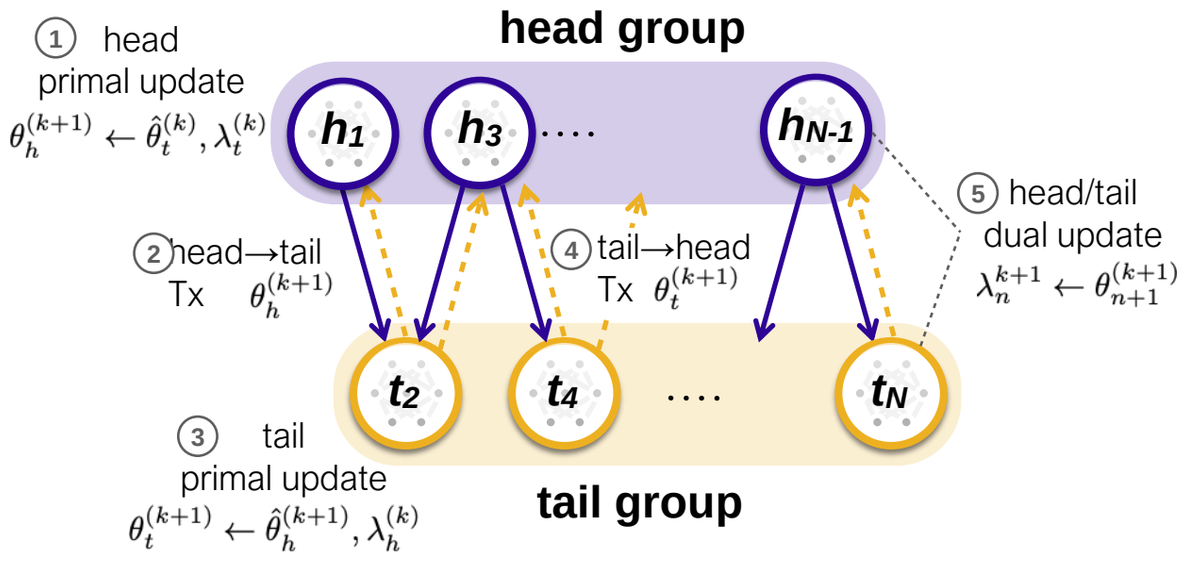
FL (GD based)

$$\underset{\{\theta_n\}}{\text{Minimize}} \quad \sum_{n=1}^{N} f_n(\theta_n)$$
$$\text{s.t.} \quad \theta_n = \theta \quad \forall n$$

GADMM

$$\underset{\{\theta_n\}}{\text{Minimize}} \quad \sum_{n=1}^{N} f_n(\theta_n)$$
$$\text{s.t.} \quad \theta_n = \theta_{n+1} \quad \forall n$$

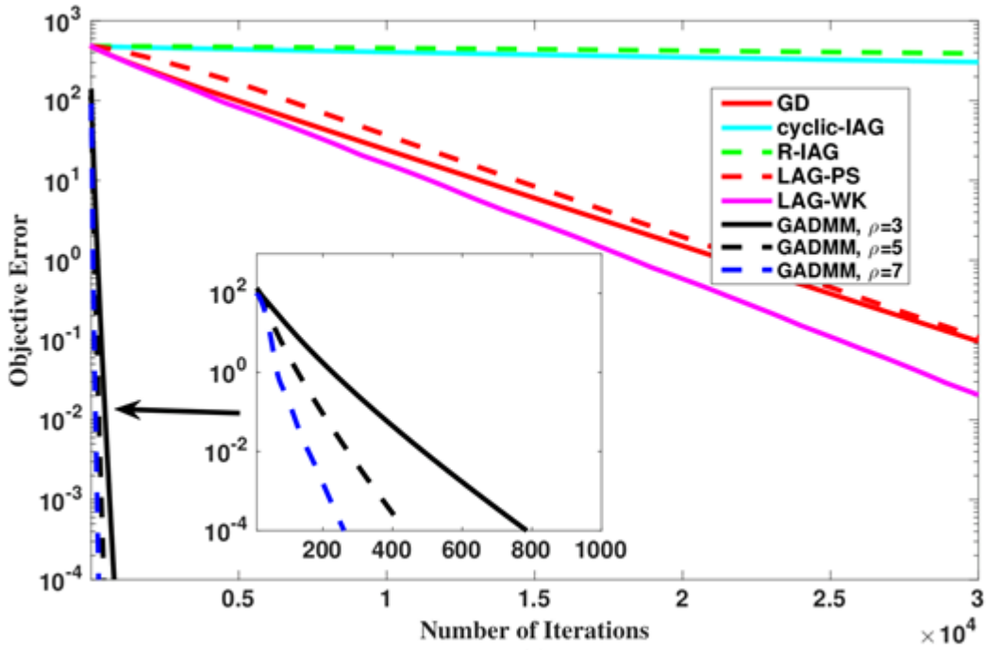


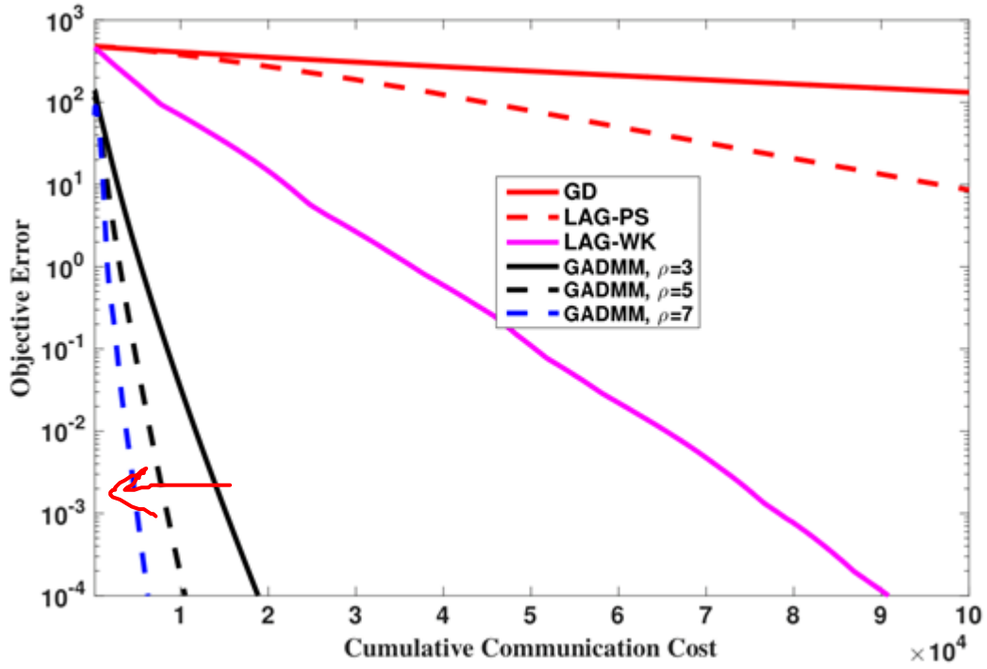


A. Elgabli, J. Park, A. S. Bedi, V. Aggarwal, and M. Bennis, "**GADMM: Fast and Communication Efficient Distributed Machine Learning Framework**," *JMLR20*

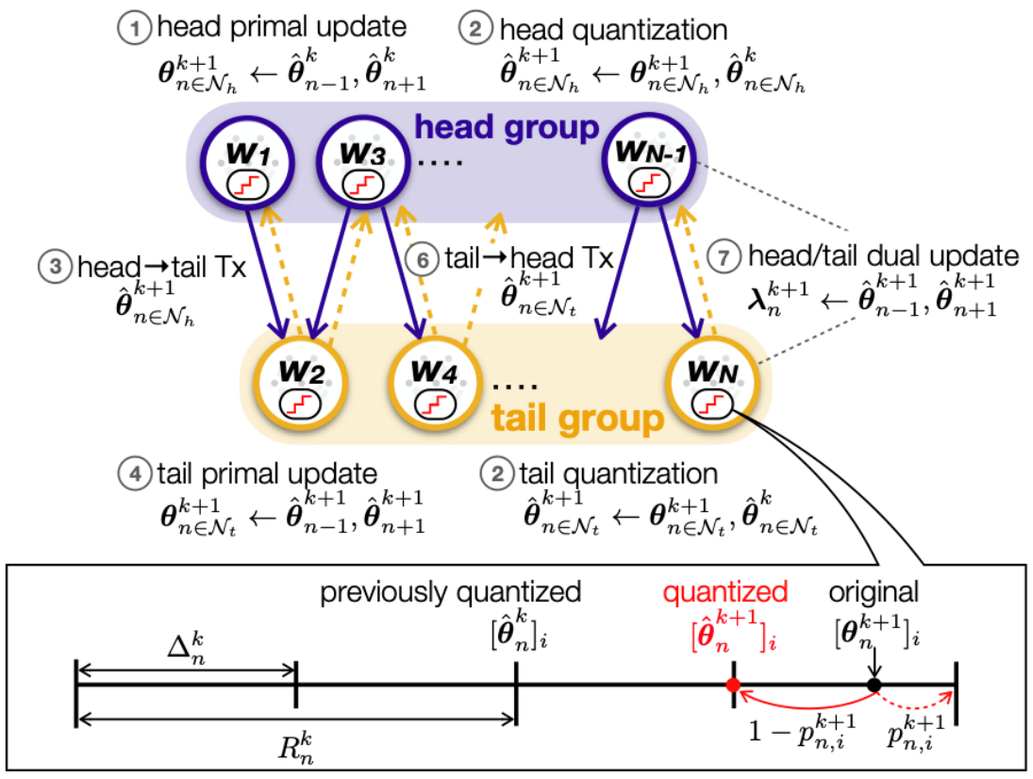
GADMM

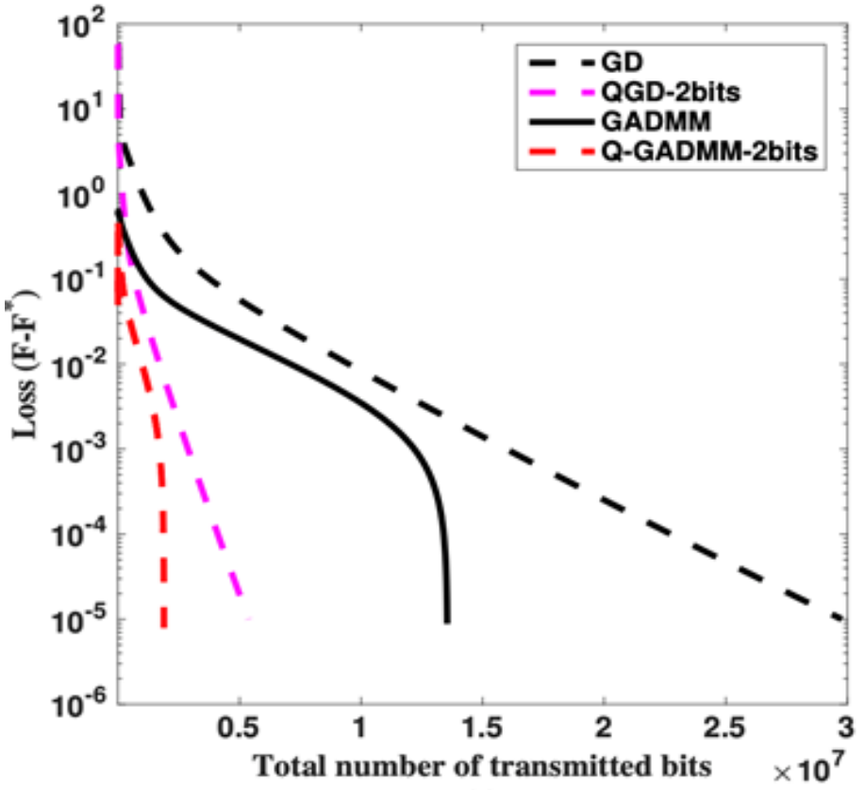
GADMM, Linear Regression



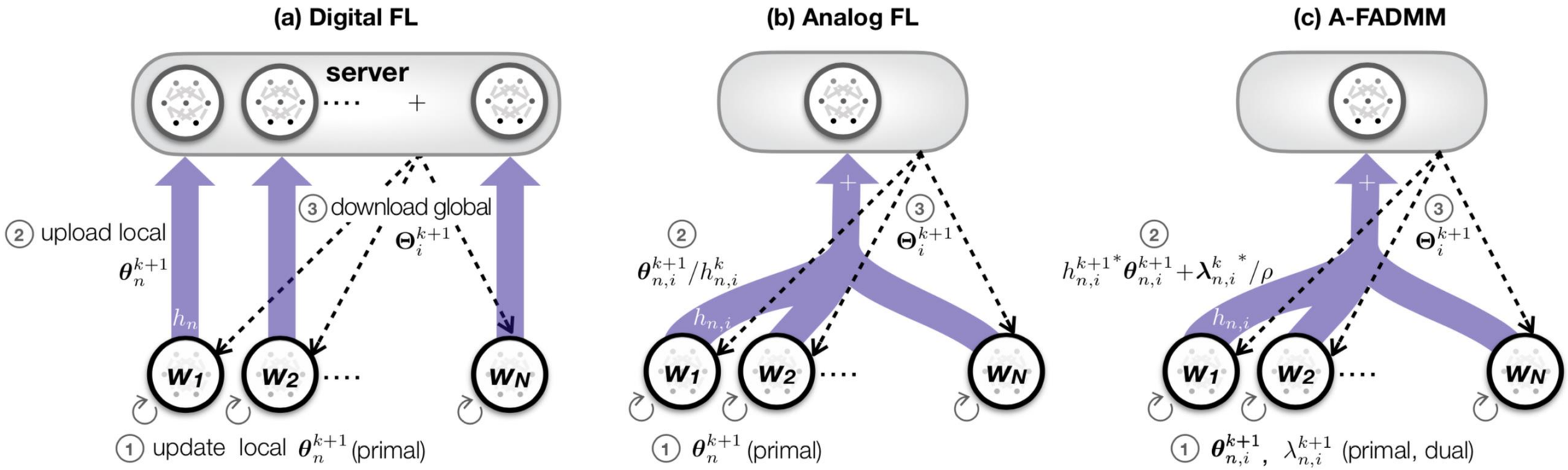


Quantization





Analog Federated ADMM



Digital-FADMM

$$\textbf{(P1)} \quad \min_{\Theta,\{\theta_n\}_{n=1}^N} \sum_{n=1}^N f_n(\theta_n)$$

$$\text{s.t.} \quad \theta_n = \Theta, \quad \forall n$$

$$\mathcal{L}_\rho = \sum_{n=1}^N f_n(\theta_n) + \sum_{n=1}^N \langle \lambda_n, \theta_n - \Theta \rangle + \frac{\rho}{2} \sum_{n=1}^N \| \theta_n - \Theta \|_2^2$$

$$\Theta^{k+1} = \frac{1}{N} \sum_{n=1}^N (\theta_n^{k+1} + \frac{1}{\rho}\lambda_n^k)$$

$$\lambda_n^{k+1} = \lambda_n^k + \rho(\theta_n^{k+1} - \Theta^{k+1})$$

Analog-FADMM

$$\textbf{(P2)} \quad \min_{\Theta,\{\theta_n\}_{n=1}^N} \sum_{n=1}^N f_n(\theta_n)$$

$$\text{s.t.} \quad h_{n,i}\theta_{n,i} = h_{n,i}\Theta_i, \quad \forall n, i$$

$$\mathcal{L}_\rho = \sum_{n=1}^N f_n(\theta_n) + \sum_{i=1}^d \sum_{n=1}^N \lambda_{n,i}^* h_{n,i}(\theta_{n,i} - \Theta_i) + \frac{\rho}{2} \sum_{i=1}^d \sum_{n=1}^N |h_{n,i}|^2 (\theta_{n,i} - \Theta_i)^2$$

$$\Theta_i^{k+1} = \frac{1}{\sum_{n=1}^N |h_{n,i}|^2} \sum_{n=1}^N \left(|h_{n,i}|^2 \theta_{n,i}^{k+1} + h_{n,i}\lambda_{n,i}^{k\,*}/\rho \right)$$

$$\lambda_{n,i}^{k+1} = \lambda_{n,i}^k + \rho h_{n,i}(\theta_{n,i}^{k+1} - \Theta_i^{k+1})$$

Analog Federated ADMM

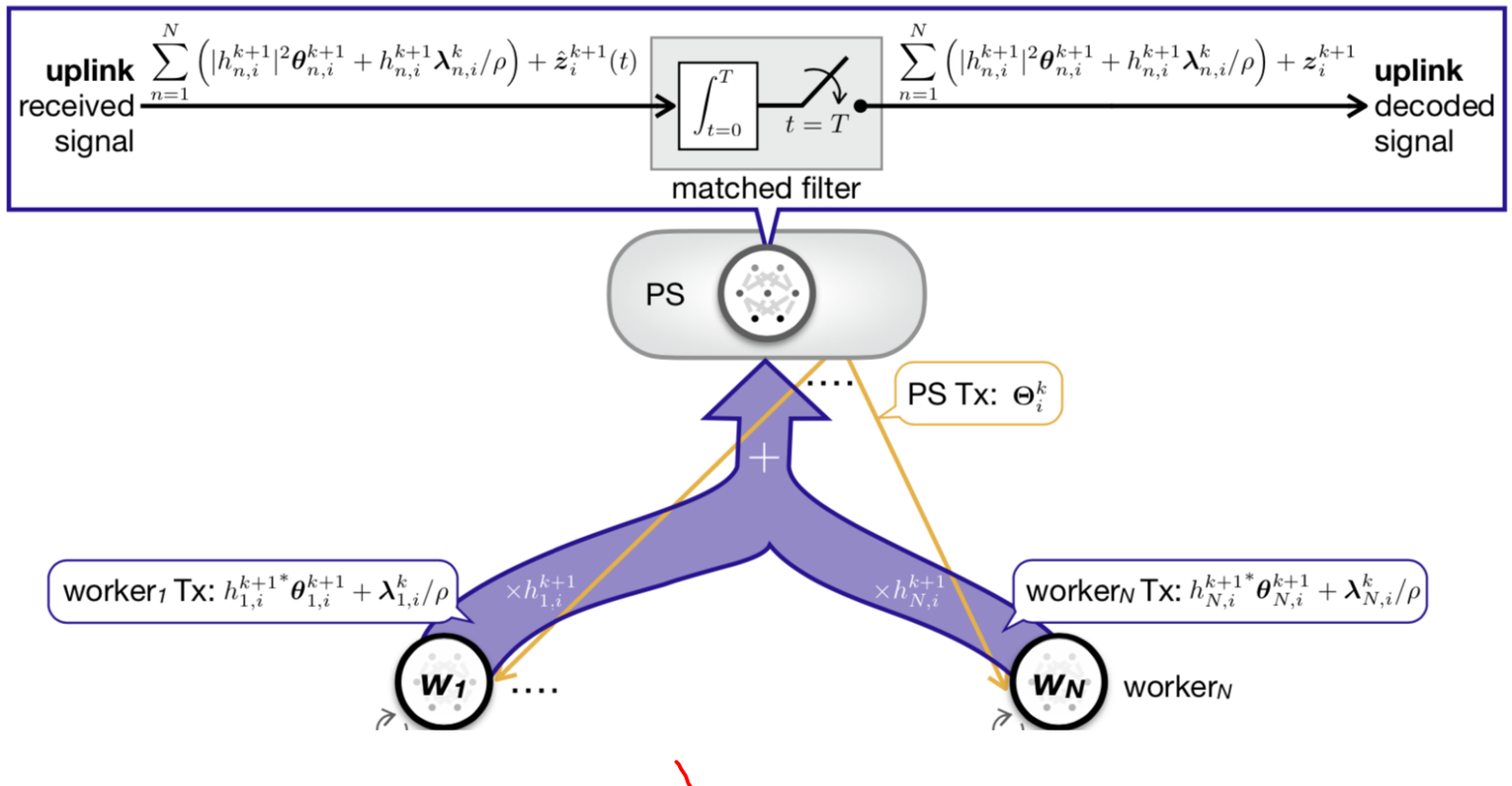
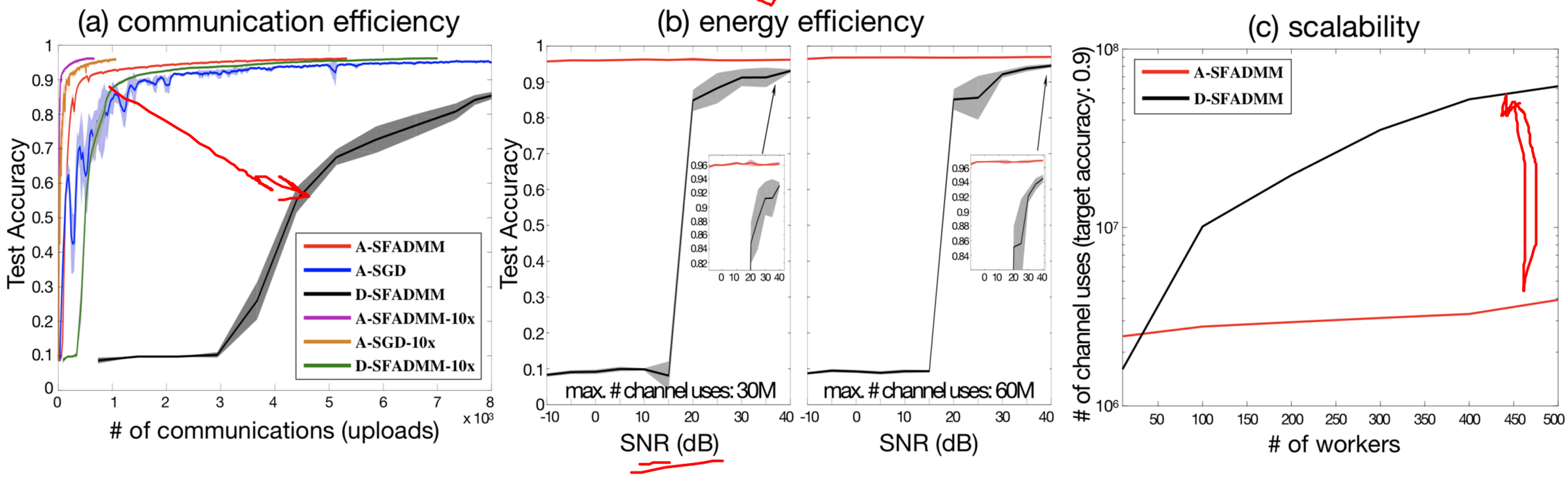
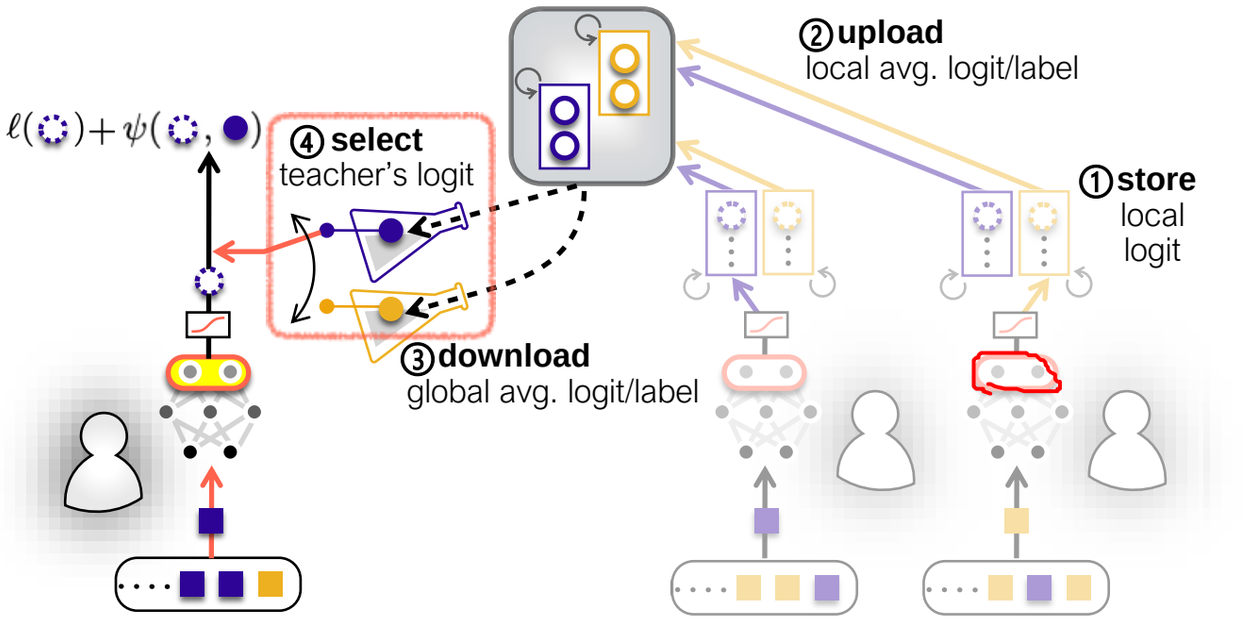
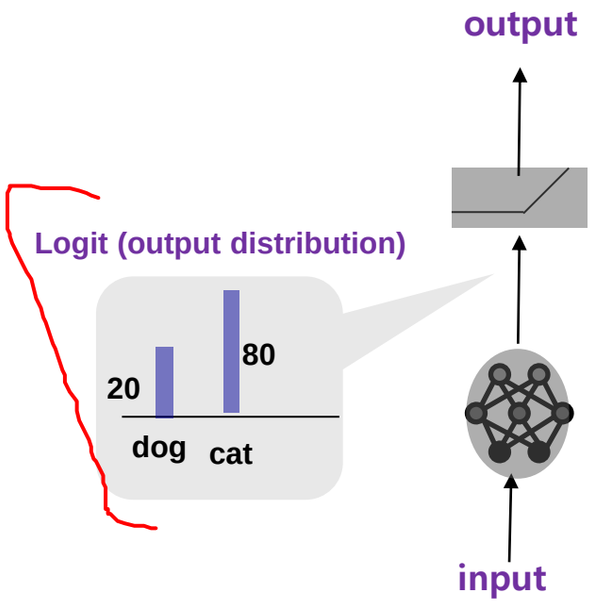


Image Classification



Federated Distillation (FD)

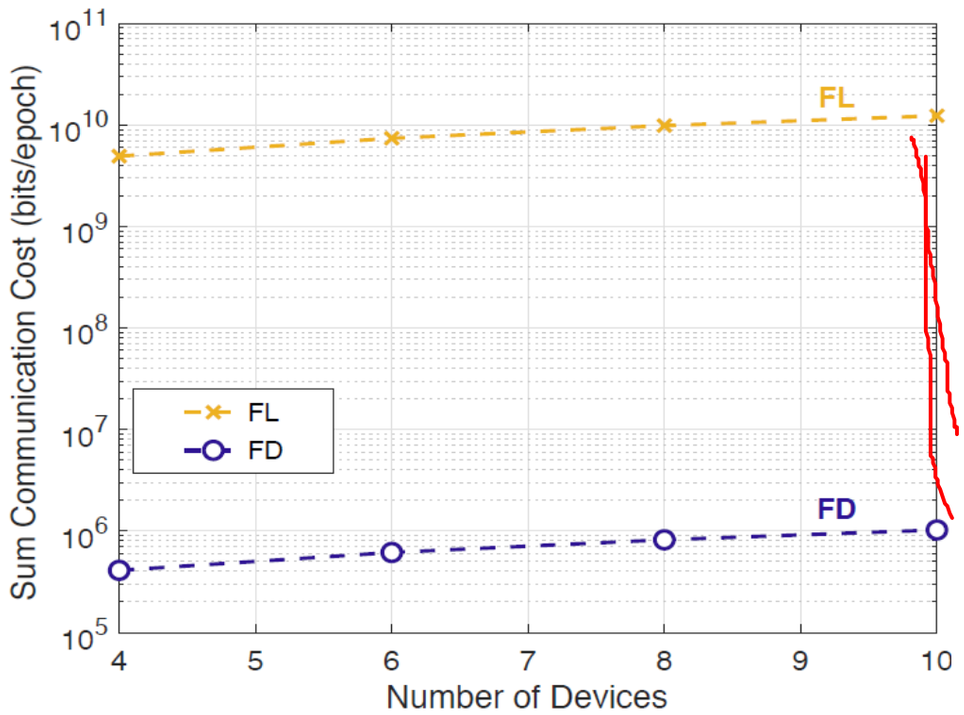


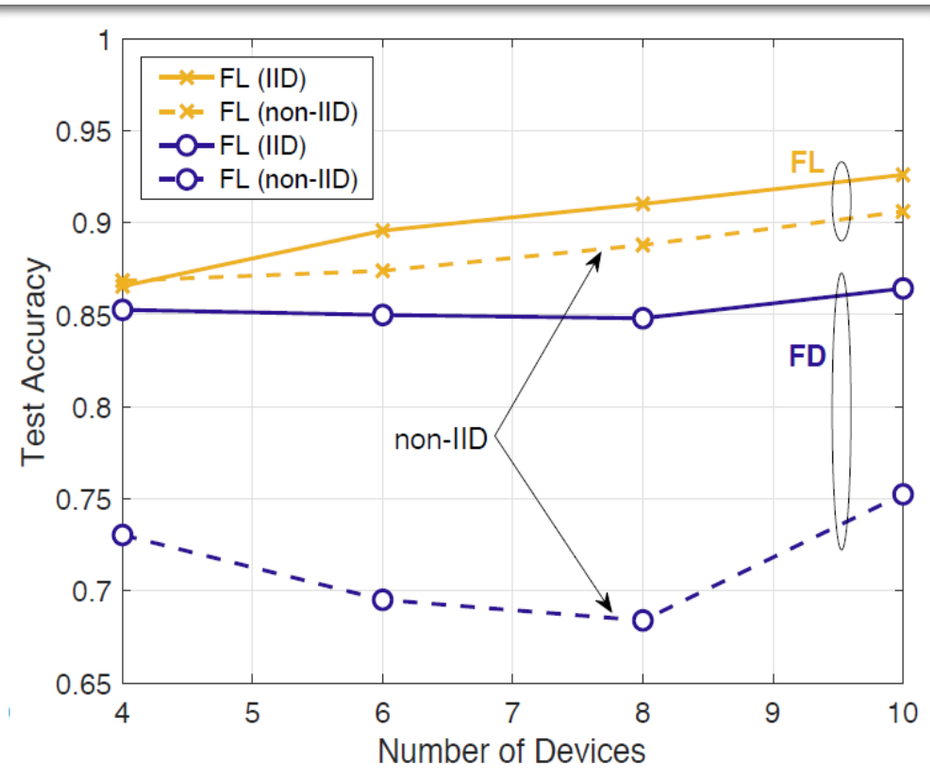


$$w_{k+1}^{(i)} = \begin{cases} w_k^{(i)} - \eta \left(\nabla \ell(w_k^{(i)}) + \psi(F_{k,\ell}^{(i)}, \breve{F}_{k,\ell}^{(i)}) \right) & \text{if } k \bmod \tau = 0, \\ w_k^{(i)} - \eta \nabla \ell(w_k^{(i)}) & \text{otherwise} \end{cases}$$

teacher's logit = global avg. logit/label $\breve{F}_{k,\ell}^{(i)} = \sum_{j \neq i} \hat{F}_{k,\ell}^{(j)} / M$

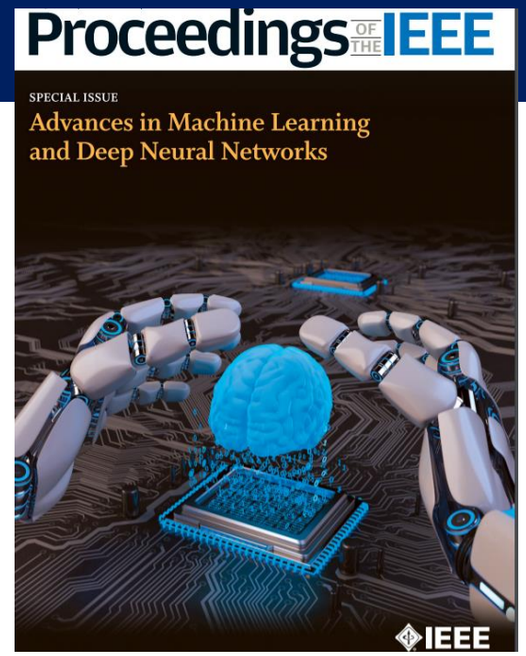
local avg. logit/label $\hat{F}_{k,\ell}^{(i)} = \sum_{k} F_{k,\ell}^{(j)} / \tau$





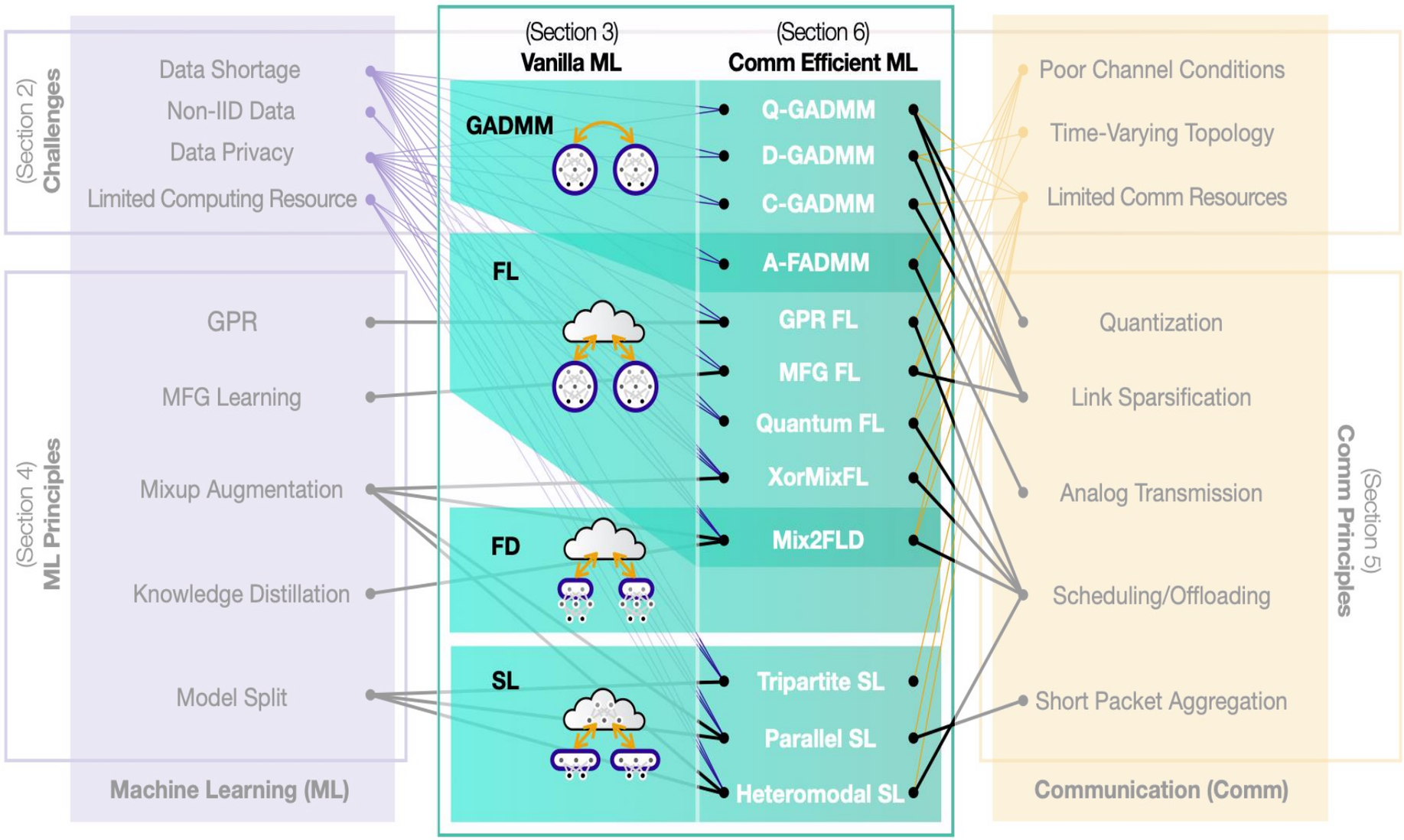
E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, "**Federated Distillation and Augmentation under Non-IID Private Data**," *NeurIPS 2018 MLPCD*

Potential Extensions



Take all the above and extend:

- Arbitrary and time varying topologies
- Non-convex and stochastic problems
- 2nd order methods (work in progress)
- Bayesian learning
- RL, etc.



What's Next?

Creative Collision of two revolutions

Limitations

- Obsession with accuracy
  - Energy Bill? Sustainability?
- Brittle, lacks robustness; Poor Generalization
- FL is the **first-step** towards **truly intelligent systems (6G)**

➡**Function approximators (curve fitting + learning *CORRELATIONS*).**

➡ **Lack reasoning**

➡ **Extrapolation + Imagination..**

Desiderata

1. **Function** of data
2. **Minimal** without compromising the **sufficient** effectiveness in the task
3. **Invariant**
4. **Disentangled**
5. **Causal** for extrapolating OOD
6. **Emergent**

Objective

Learning **Semantic** representations satisfying **D1-D6** for **X**

☑ Tx less data ☑ More reliable ☑ More energy-efficient ☑ Sample efficient ☑ Intelligent

Post-Shannon Era is here

THE MATHEMATICAL THEORY
OF COMMUNICATION

CLAUDE E. SHANNON
and
WARREN WEAVER

LEVEL A. How accurately can the symbols of communication be transmitted?

ALL 2G-5G/6G KPIs derived from Level-A

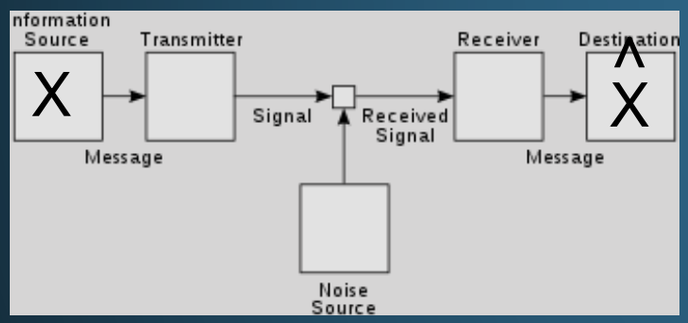
THE MATHEMATICAL THEORY
OF COMMUNICATION

CLAUDE E. SHANNON
and
WARREN WEAVER

The Bandwagon
CLAUDE E. SHANNON

LEVEL B. How precisely do the transmitted symbols convey the desired meaning?

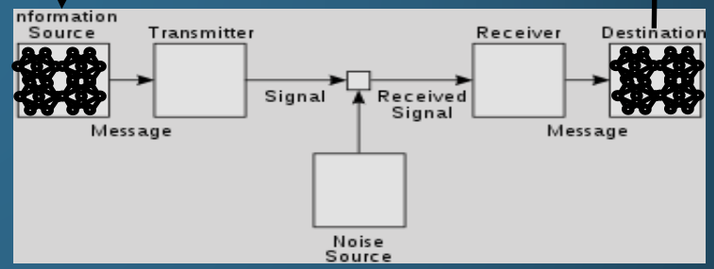
LEVEL C. How effectively does the received meaning affect conduct in the desired way?



- NO meaning
- NO context
- NO structure
- NO memory

- Reproducing at one point either exactly-or-approximately message (X) selected at another point.
- Level-A: Statistical/mathematical description of information

SHANNONIAN = STATISTICS



- Leverage semantics, structure, meaning
- Utility emerges!!

- Induce behavioral change through sensing and actuation with a shared environment (emergent property!)
- Agents modeling/reasoning over other agents intents/goals/beliefs ..

SEMANTIC = STRUCTURE + STATISTICS

VisionX: Semantic Communication Meets ML

From

Departing from learning in data space (e.g., pixel, CSI, observational data)

To

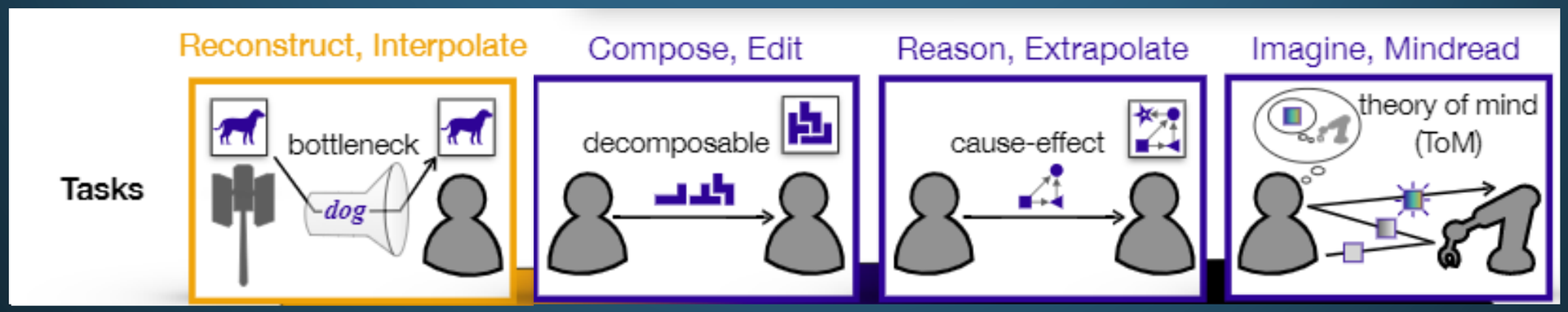
Learning **semantic representations** of the real-world (objects/agents interactions)

From

Departing from **reconstruction tasks** and how to **best encode data** (**Shannonian information**)

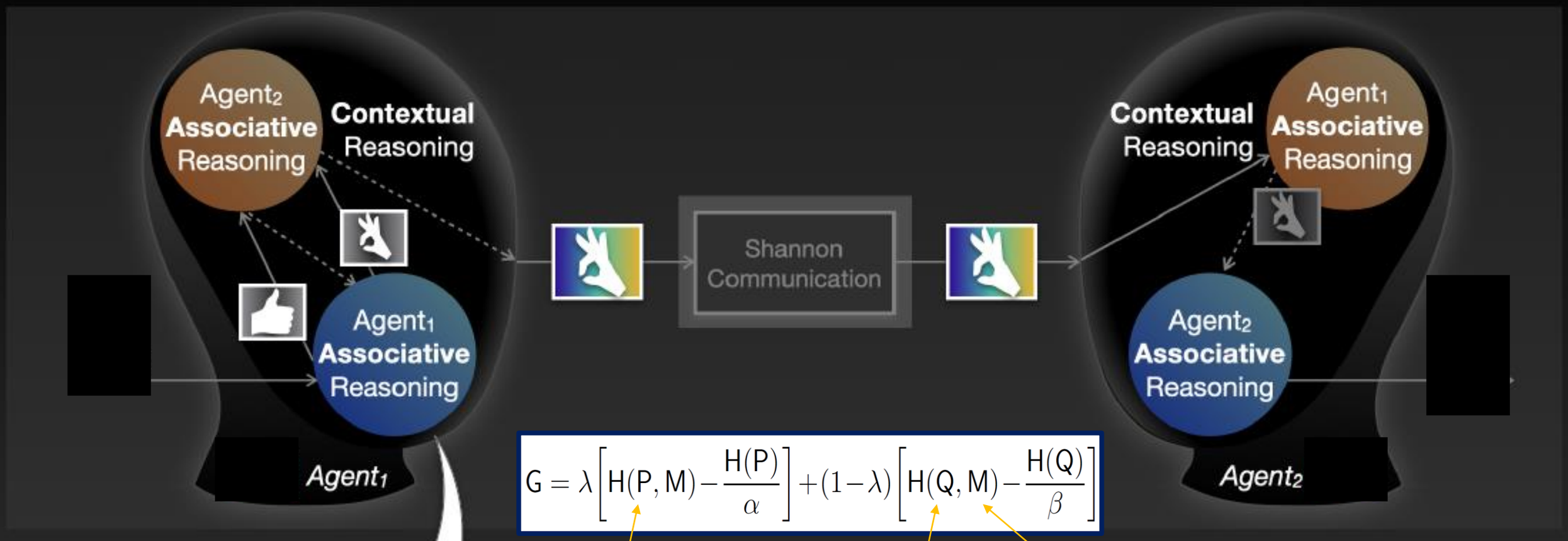
To

Agents inducing **behavioral change** through planning/imagining/reasoning over these **high-level semantic representations.**



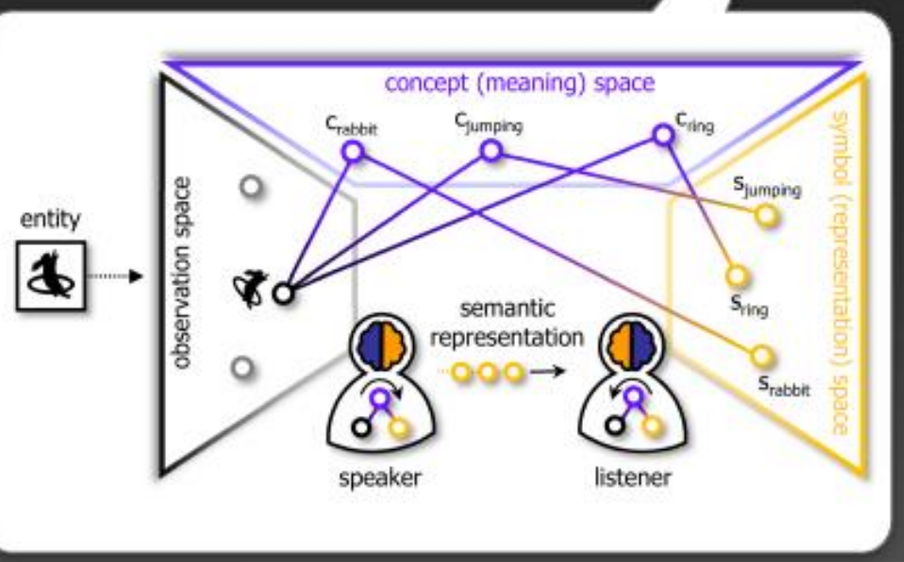
- ☑ Tx less data
- ☑ More reliable
- ☑ More energy-efficient
- ☑ Sample efficient
- ☑ Intelligent

How & under what conditions cooperative communication among agents **emerges** and is **robust to deviations** between agents?



$$G = \lambda\left[H(P,M) - \frac{H(P)}{\alpha}\right] + (1-\lambda)\left[H(Q,M) - \frac{H(Q)}{\beta}\right]$$

speaker context — listener context — mutual context



Theorem 2. *(Mutual CC Convergence) As the iteration step $t \to \infty$, the alternating iterations of*

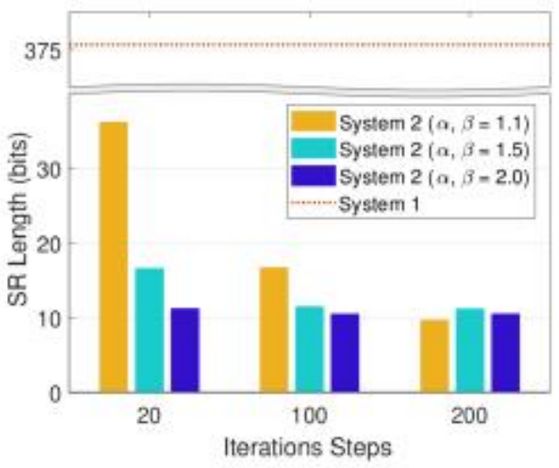
$$M_1^{[t]}(e,c;a,a') = \lambda P^{[t-1]}(e,c;a,a') + (1-\lambda) Q^{[t-1]}(e,c;a,a'), \quad (16)$$

$$P^{[t]}(e,c;a,a') = \frac{M_1^{[t]}(e,c;a,a')^{\alpha}}{\sum_{(e,c)\in\mathcal{E}\times\mathcal{C}} M_1^{[t]}(e,c;a,a')^{\alpha}}, \quad (17)$$

$$M_2^{[t]}(e,c;a,a') = \lambda P^{[t]}(e,c;a,a') + (1-\lambda) Q^{[t-1]}(e,c;a,a'), \text{ and} \quad (18)$$

$$Q^{[t]}(e,c;a,a') = \frac{M_2^{[t]}(e,c;a,a')^{\beta}}{\sum_{(e,c)\in\mathcal{E}\times\mathcal{C}} M_2^{[t]}(e,c;a,a')^{\beta}} \quad (19)$$

converge to a common mutual CC $M^{[*]} = \lim_{t\to\infty} M_1^{[t]} = \lim_{t\to\infty} M_2^{[t]}$ *for all* $e \in \mathcal{E}$ *and* $c \in \mathcal{C}$,



Communication = Belief Transport from data-hypothesis space

SHANNON COMMUNICATION

- Information: **Scalar**

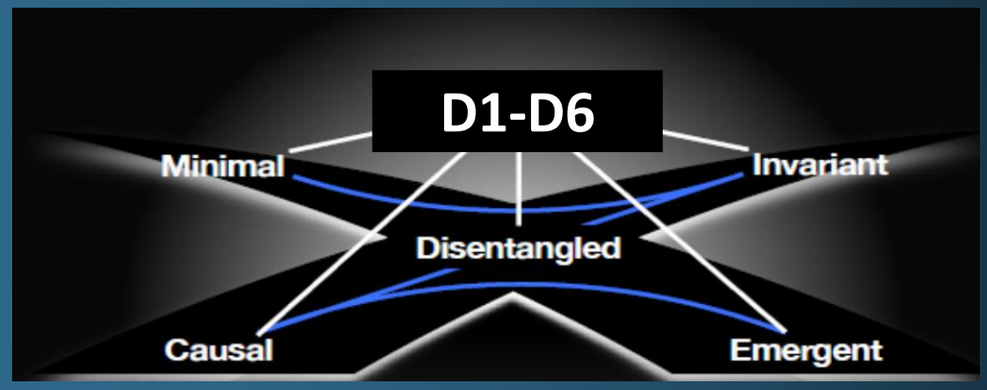
- **STATISTICS:**
  - Symbol probability

- **Goal:**
  - Reconstruction (level A)

SEMANTIC COMMUNICATION

- Information: **Structures, Categories & Spaces**

- **STRUCTURE:**
  - System 1 + System 2 ML (D1-D6)



Algebraic, hierarchical, compositional

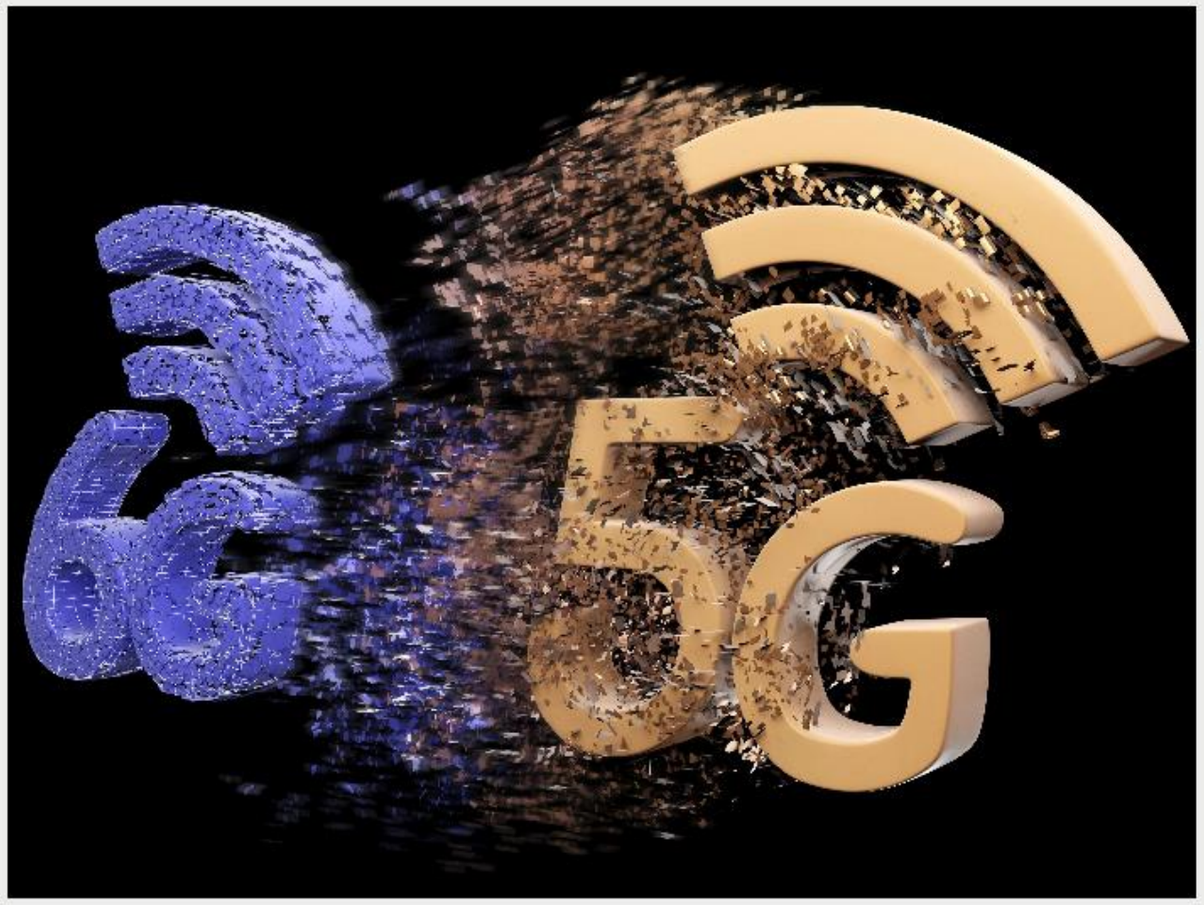
- **Goal: Reasoning by abstraction**
  - (Higher-order) object-relations-intent.
  - Topologies (Object/ concept sameness)
  - Much more

GUEST ARTICLE | TELECOMMUNICATIONS

If 6G Becomes Just 5G+, We'll Have Made a Big Mistake

› Iterating current tech is a bad idea; semantic communication could be the answer

BY MEHDI BENNIS | 16 DEC 2021 | 7 MIN READ |



Thank you

VisionX coming soon