Intelligence Beyond the Edge in IoT

Xiaofan Yu
University of California, San Diego
Supervisor: Tajana Šimunić Rosing

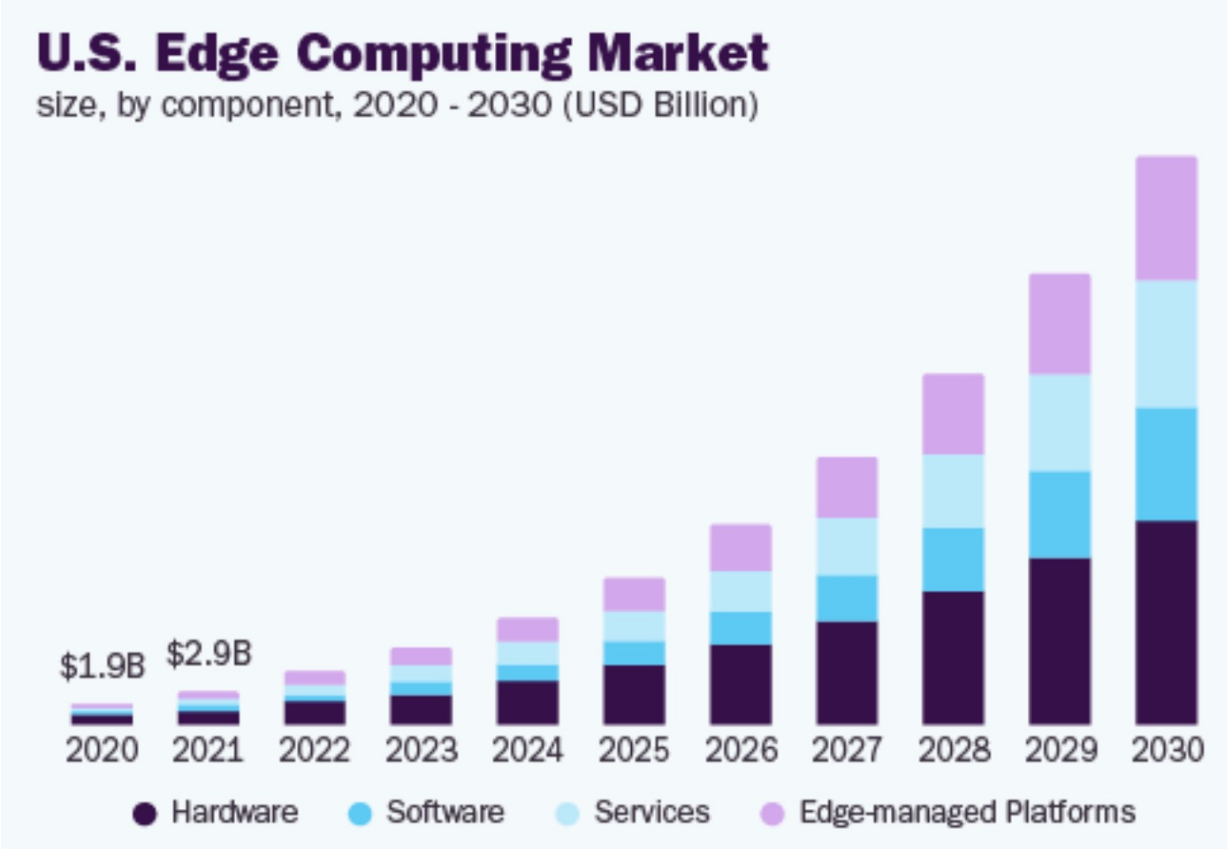
PhD Forum
IPSN 2023

JACOBS SCHOOL OF ENGINEERING
Computer Science and Engineering

System Energy Efficiency Lab
seelab.ucsd.edu

Edge Computing is Growing Exponentially!

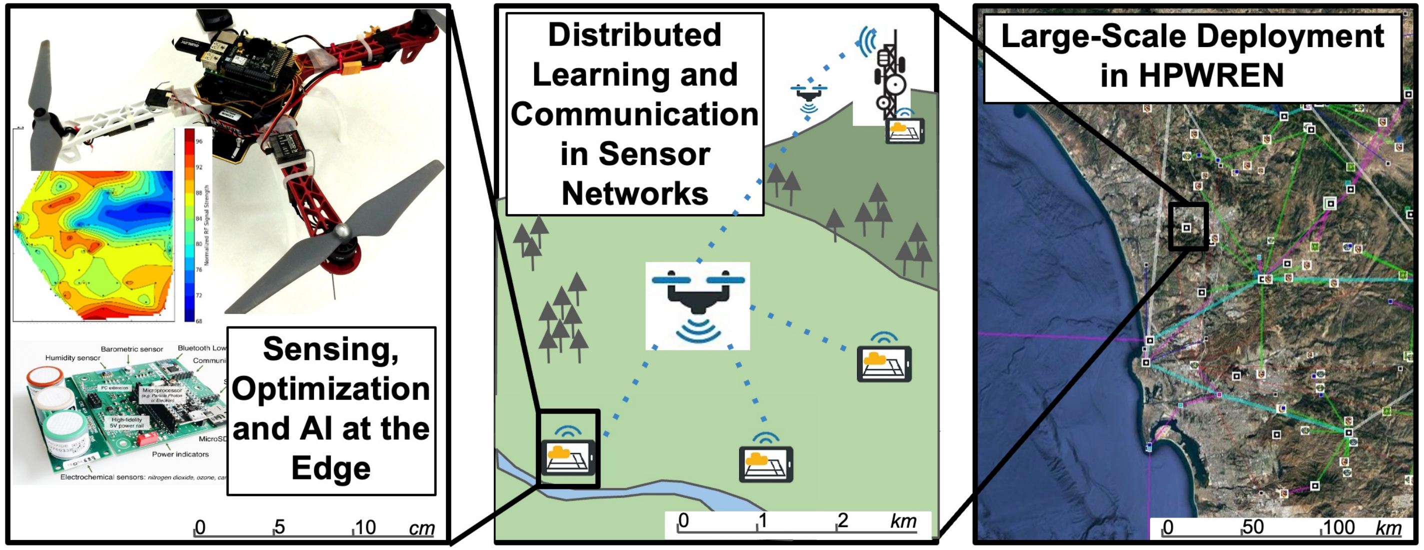
- The global edge computing market size is expected to expand at a compound annual growth rate (CAGR) of 37.9% from 2023 to 2030[1]



1. https://www.grandviewresearch.com/industry-analysis/edge-computing-market

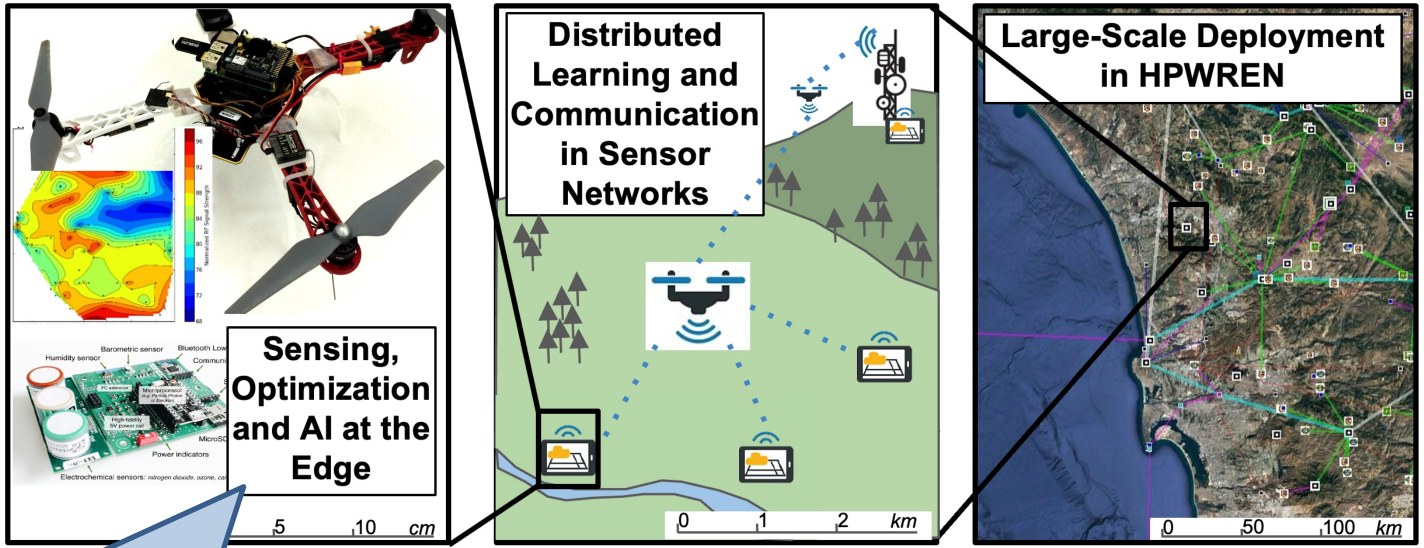
Motivating Problem:
Deploying Intelligence for Environmental Monitoring

- Enabling ML training pervasively in IoT applications is an active research area, which calls for sophisticated designs on single device level, sensor network level, and large-scale deployment level.



Roadmap to Intelligence in IoT

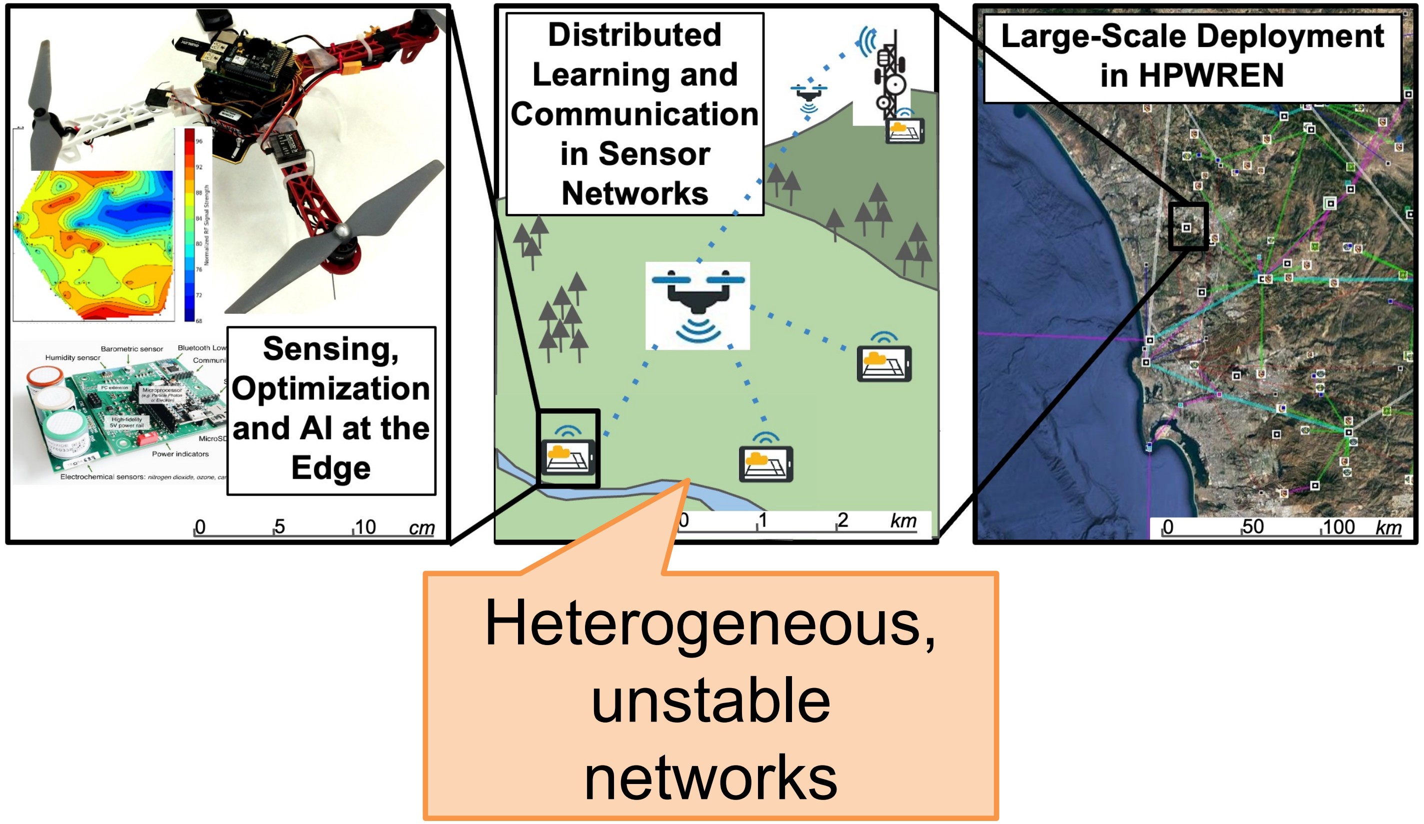
- Enabling ML training pervasively in IoT applications is an active research area, which calls for sophisticated designs on single device level, sensor network level, and large-scale deployment level.



Drifting, noisy data;
Limited supervision

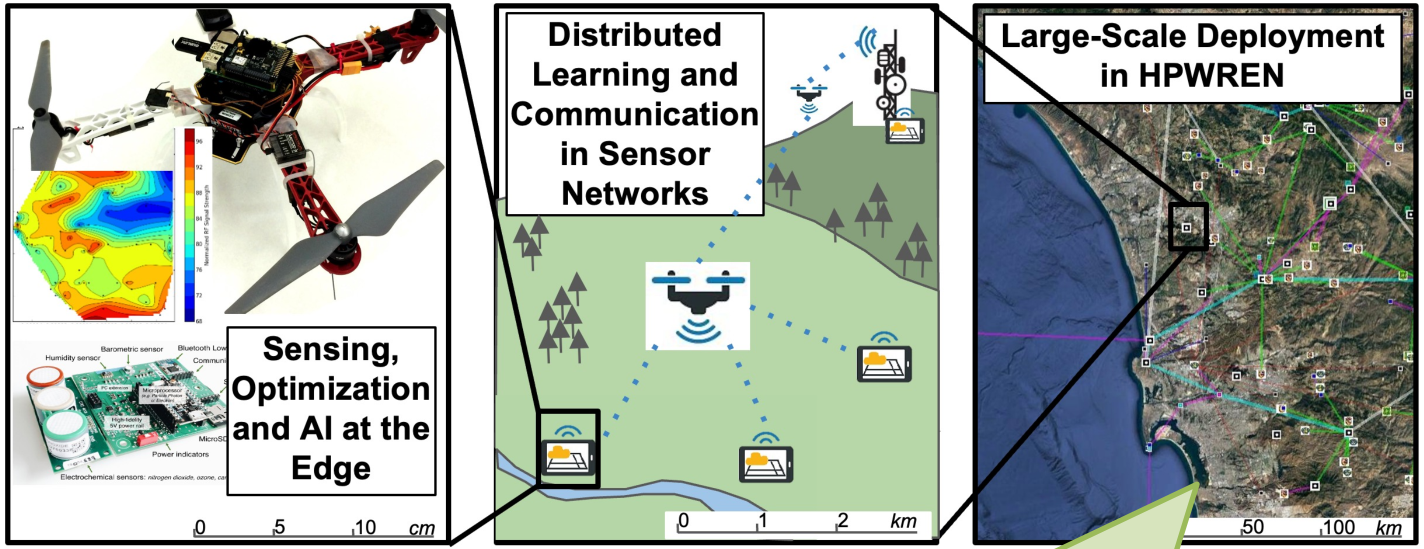
Roadmap to Intelligence in IoT

- Enabling ML training pervasively in IoT applications is an active research area, which calls for sophisticated designs on single device level, sensor network level, and large-scale deployment level.



Roadmap to Intelligence in IoT

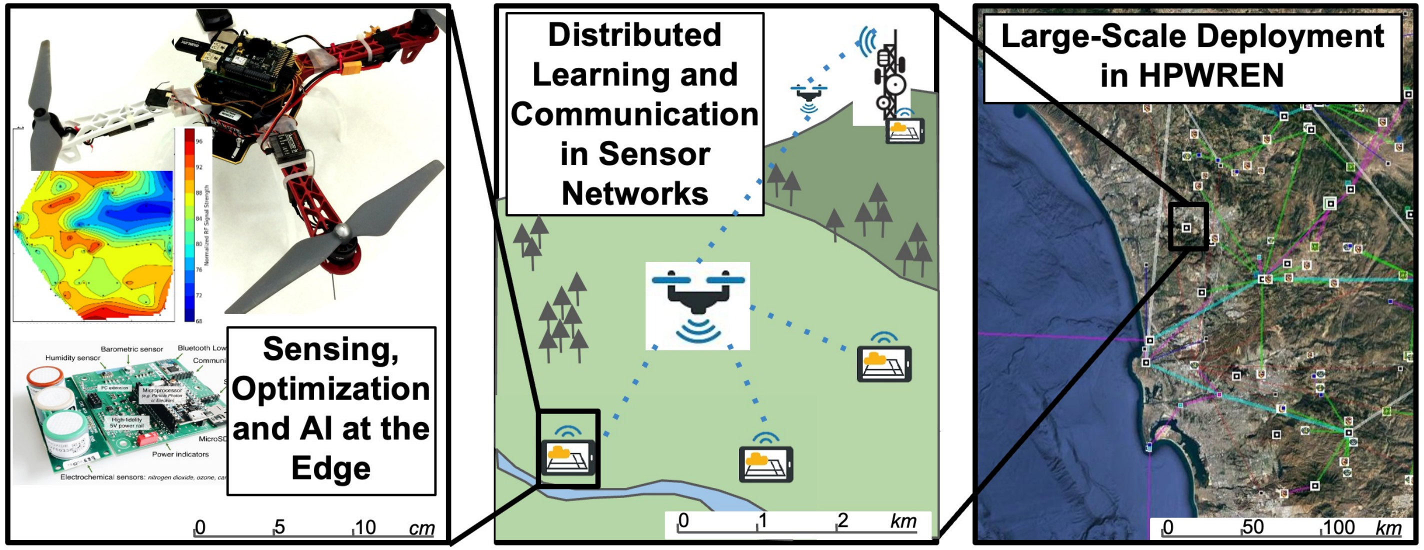
- Enabling ML training pervasively in IoT applications is an active research area, which calls for sophisticated designs on single device level, sensor network level, and large-scale deployment level.



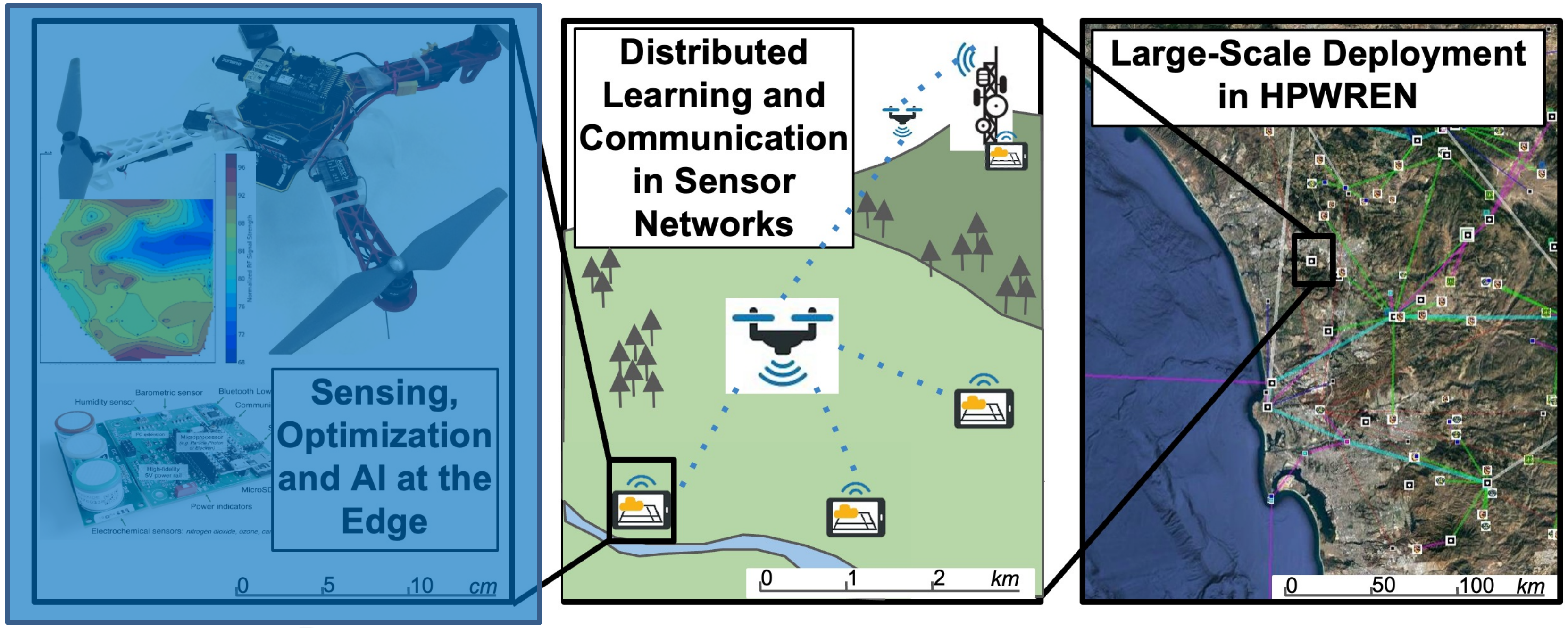
Costs, sustainability, reliability

Thesis Statement

- My PhD research targets at contributing a **full stack** of technologies in all three levels, for enabling pervasive intelligence deployments in IoT



Online Unsupervised Lifelong Learning

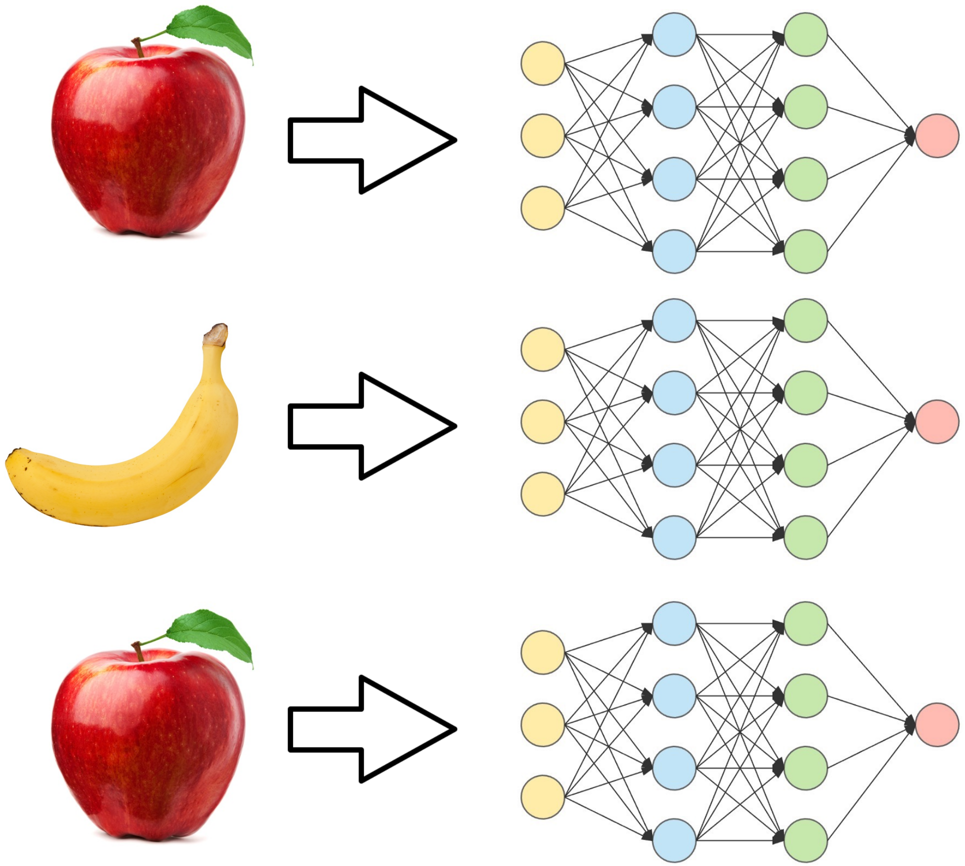


Online Unsupervised Lifelong Learning

1. **X. Yu**, Y. Guo, S. Gao, T. Rosing, "SCALE: Online Unsupervised Lifelong Learning without Prior Knowledge", CLVision'23

Catastrophic Forgetting [McCloskey 1989]

- **Goal:** After deployment, train an ML model on the device



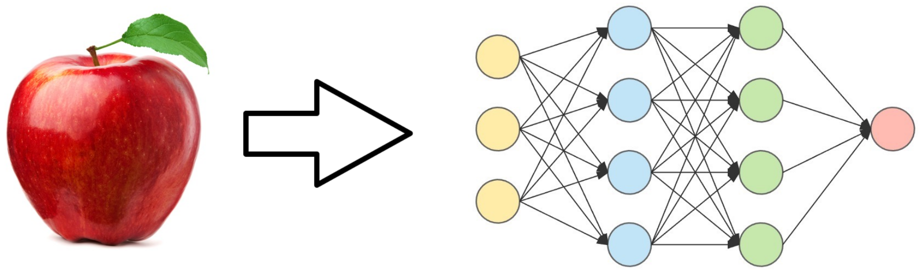
Day 1 — Prediction: apple

Day 2-99 — Prediction: banana

Day 100 — Prediction: ?

Catastrophic Forgetting and Lifelong Learning

After seeing several days of banana…



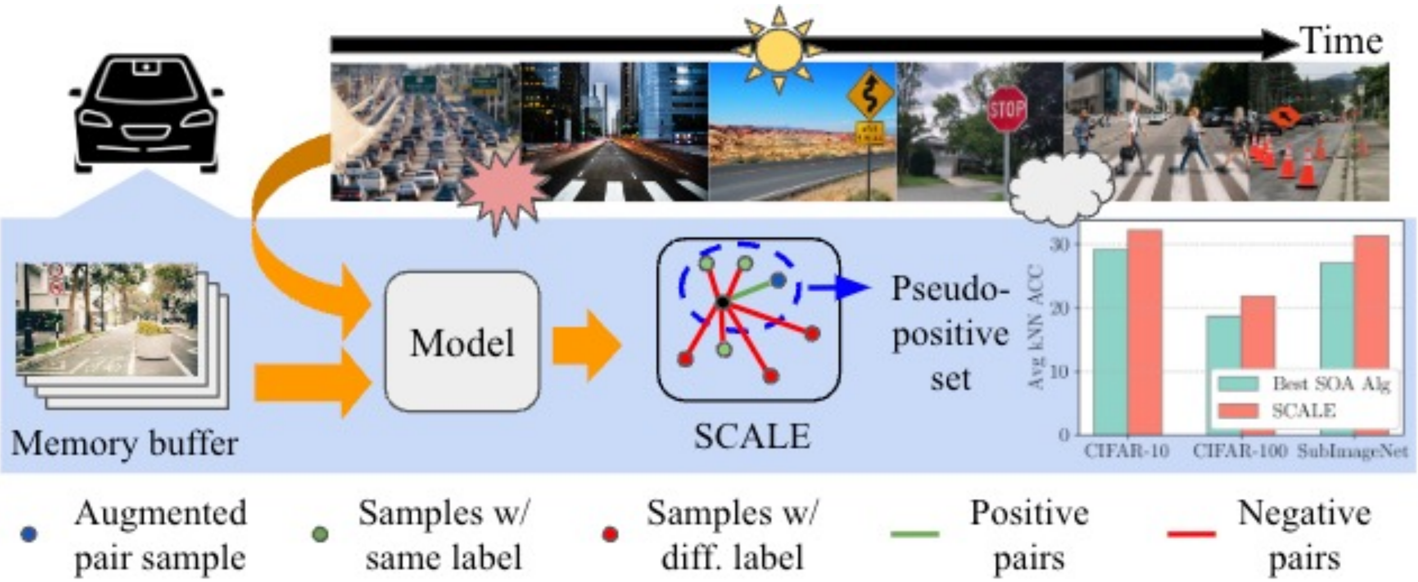
Prediction: ?

- Lifelong learning (or continual learning)
  - To continually learn over time by acquiring new knowledge as well as consolidating past experiences
  - Key assumption: continuously changing environments
  - Key challenge:
    - Knowledge interference in NNs
    - Limited memory storage
    - Previous works rely on prior knowledge (e.g., task boundary) to produce good results

SCALE: Self-Supervised Contrastive Learning

- We focus on the online, unsupervised lifelong learning problem without prior knowledge
- We propose SCALE to extract and memorize knowledge in an online and unsupervised manner

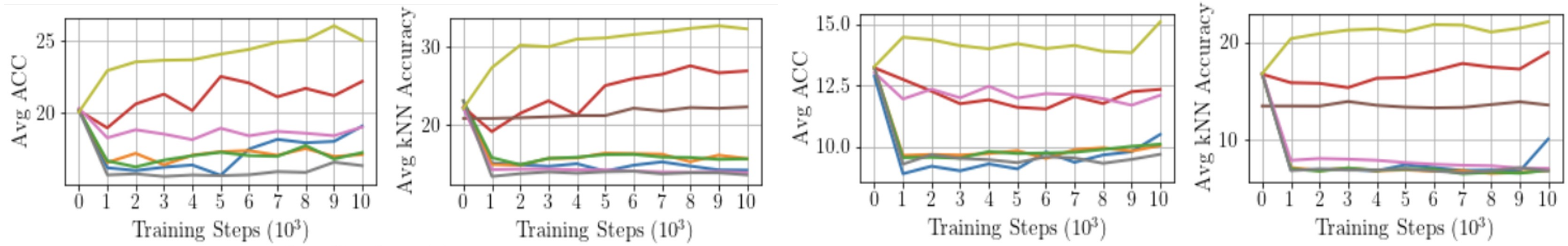
1. Inspired by contrastive learning, SCALE enhances the similarity of samples in a *pseudo-positive set*
2. SCALE uses a self-supervised forgetting loss to retain pairwise similarity (as knowledge)
3. SCALE employs a uniform online memory update strategy



Experimental Results

- **Datasets:** CIFAR-10, CIFAR-100, TinyImageNet
- **Data streams:** Four different sequential streams
- **Metric:** kNN accuracy on the learned representations
- **Key baselines:** STAM [IJCAI 2021], CaSSLe [CVPR 2022], LUMP [ICLR 2021]
- SCALE outperforms the best state-of-the-art algorithm on all settings with improvements of up to 6.43%, 5.23%, 5.86% kNN accuracy on CIFAR-10, CIFAR-100 and TinyImageNet

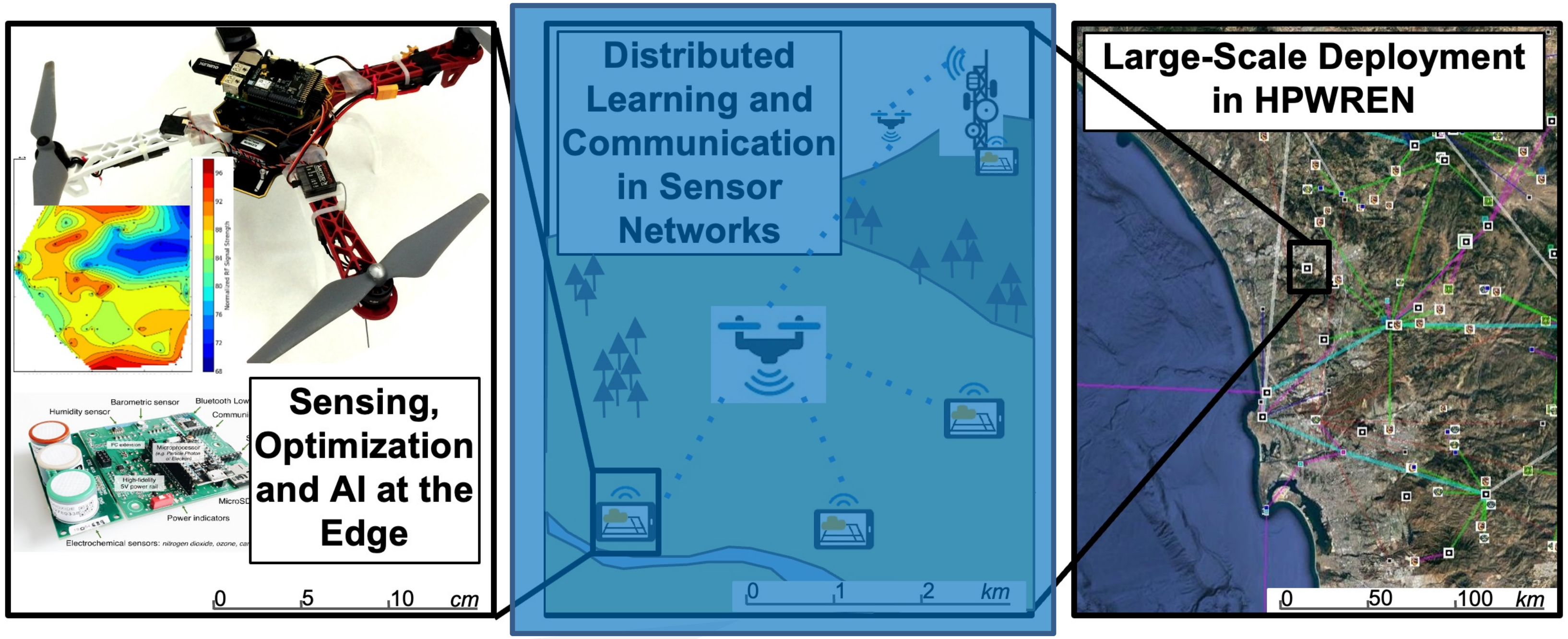
Accuracy on sequential CIFAR-10 and CIFAR-100



(b) CIFAR-10 seq

(b) CIFAR-100 seq

Efficient Federated Learning in Heterogeneous IoT Networks

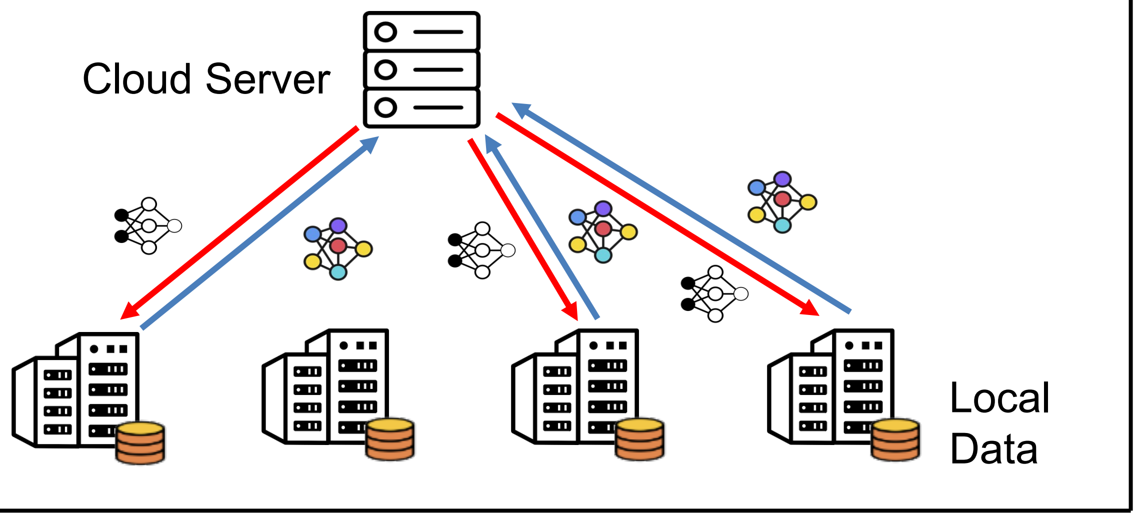


Efficient Federated Learning in Heterogeneous IoT Networks

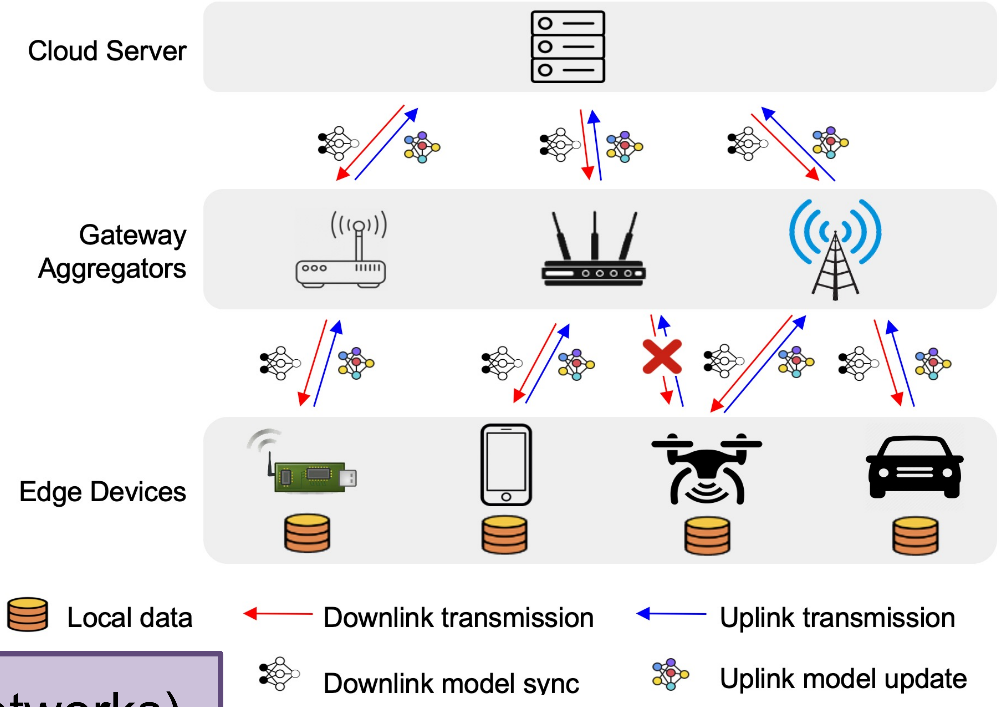
1. **X. Yu** et al, “Async-HFL: Efficient and Robust Asynchronous Federated Learning in Hierarchical IoT Networks”, IoTDI’23
2. Q. Zhao, **X. Yu,** T. Rosing, “Attentive Multimodal Learning on Sensor Data using Hyperdimensional Computing”, Poster@IPSN’23

Motivation: Uniqueness of Hierarchical IoT Networks

FL in Data Centers



FL in IoT Networks



- Heterogeneous data distribution
- Hierarchical network organization (e.g., mesh networks)
- Heterogeneous system capabilities
  - Computation + Communication
- Unexpected stragglers (e.g., device or link failures)

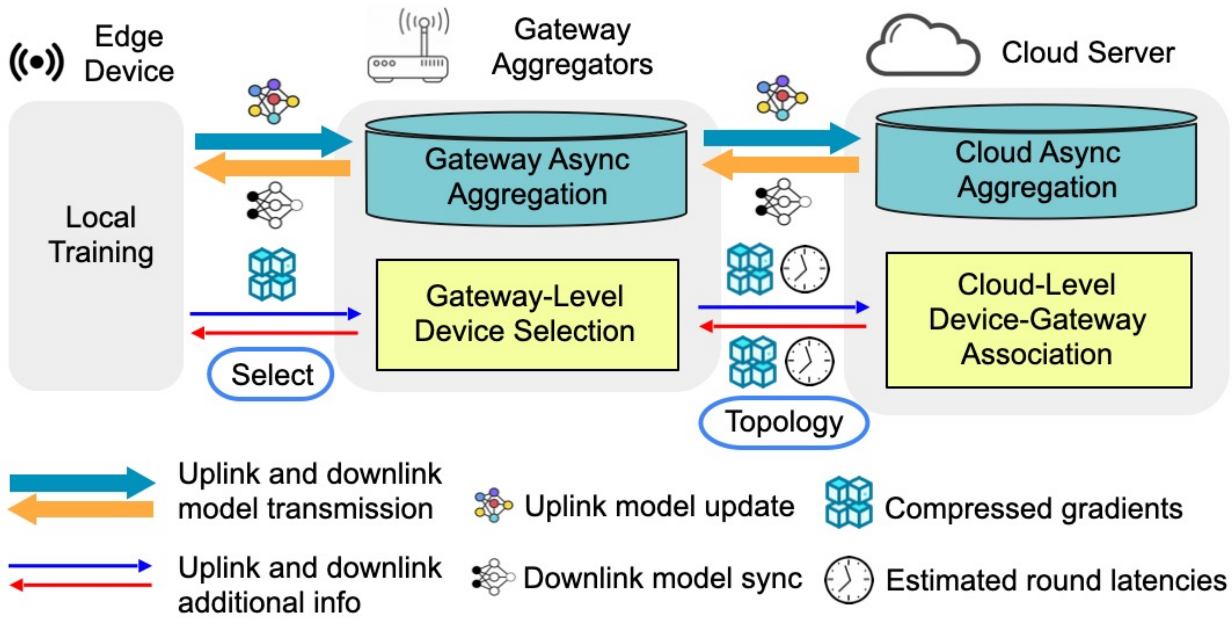
Sync Federated Learning (e.g., *FedAvg* [1]) ends up with significant slow down!!

[1] McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." *Artificial Intelligence and Statistics*. PMLR, 2017.

Our Contributions: Async-HFL

- Async-HFL is designed around three components to balance the data, system & network perspectives along with reacting timely to stragglers:

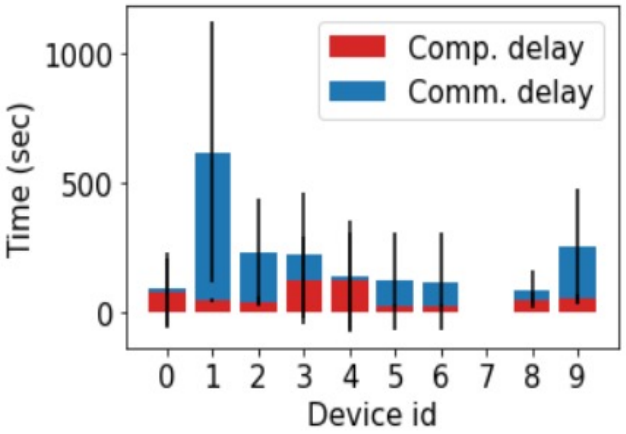
1. Async + hierarchical FL algorithm
2. Gateway-level device selection
3. Cloud-level device-gateway association



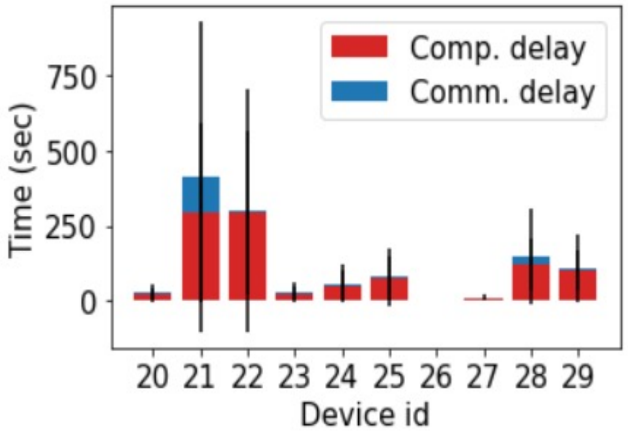
The managing framework of Async-HFL

Experimental Results

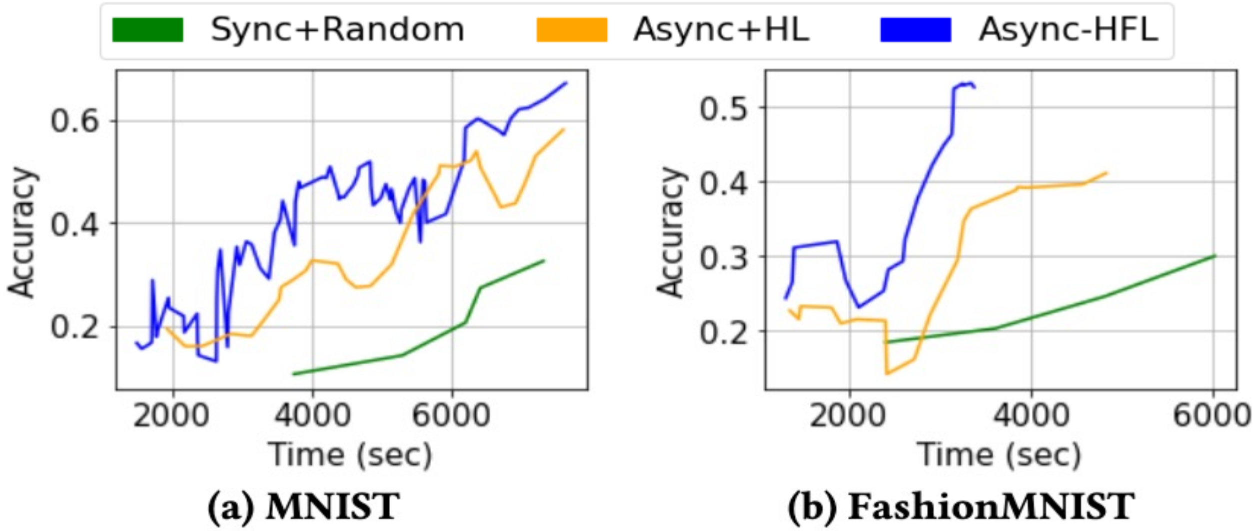
- We evaluate Async-HFL on a physical deployment and large-scale simulations
  - **Physical deployment:** 20 Raspberry Pi (RPi) 4 and 20 CPUs
  - **Large-scale simulations:** NYCMesh topology and ns-3 as network simulator
  - **Datasets:** MNIST, FashionMNIST, CIFAR-10, Shakespeare, HAR, HPWREN
- In physical deployment, Async-HFL achieves faster and more robust convergence
- In simulations, Async-HFL converges at least 1.08-1.31x faster in wall-clock time than state-of-the art asynchronous FL algorithms



Time breakup on RPis

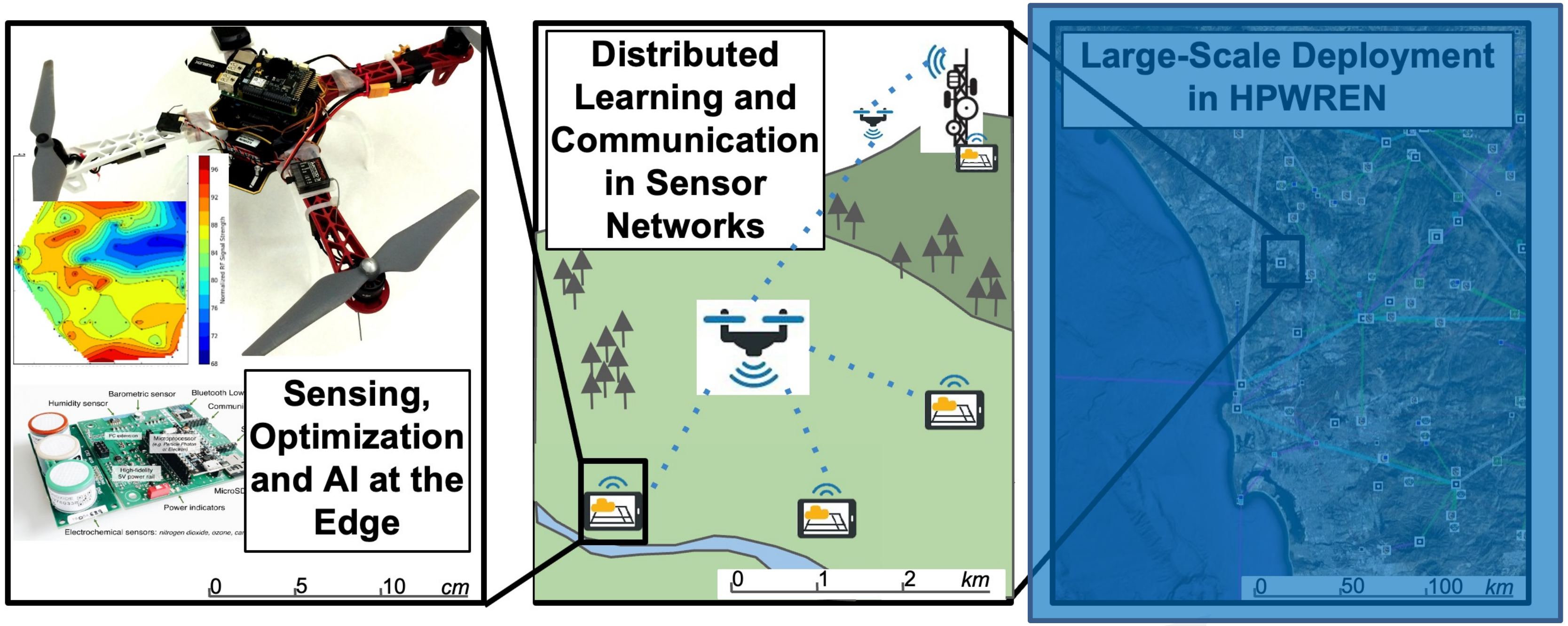


Time breakup on CPUs



Convergence on physical deployment

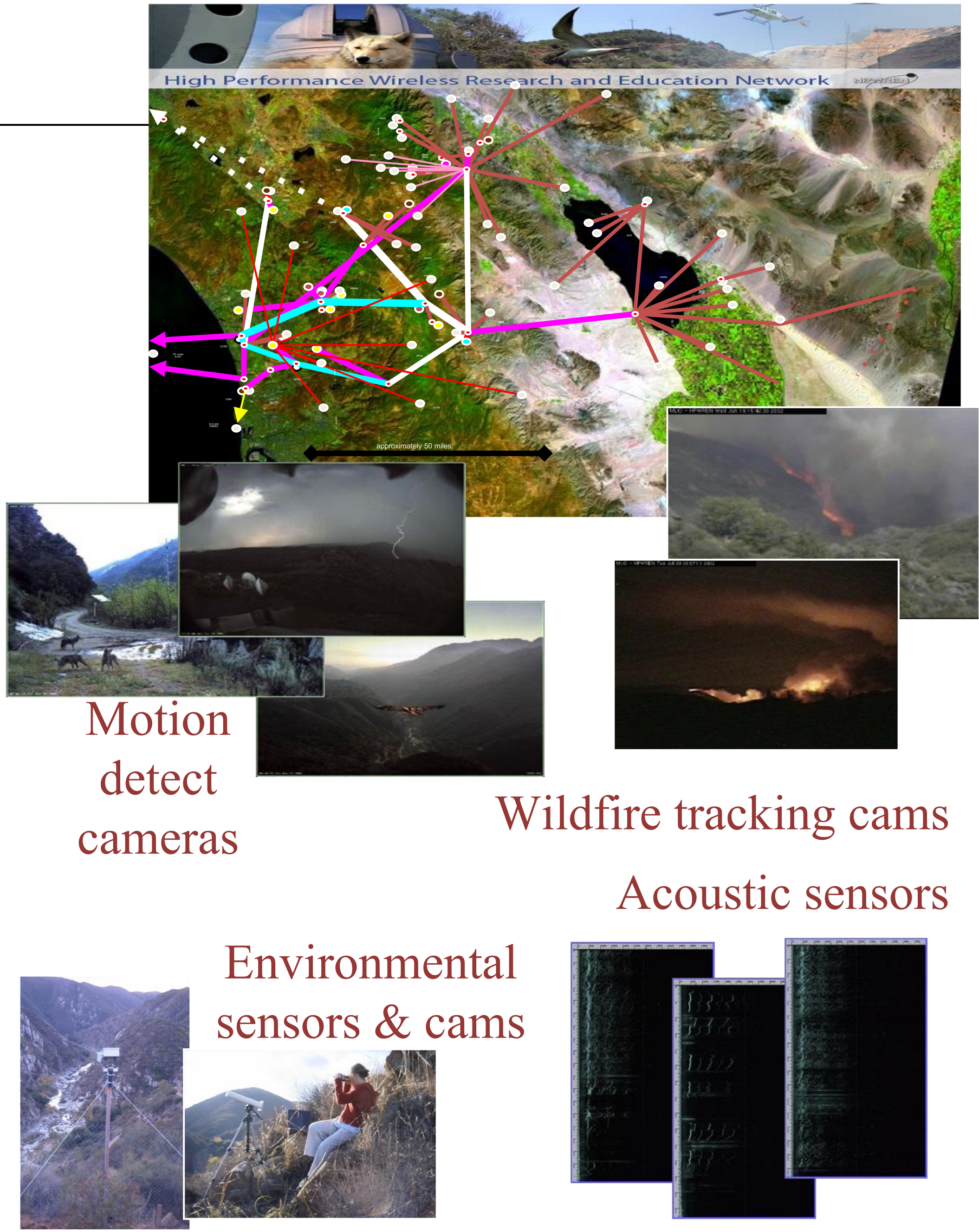
Works Planned Ahead



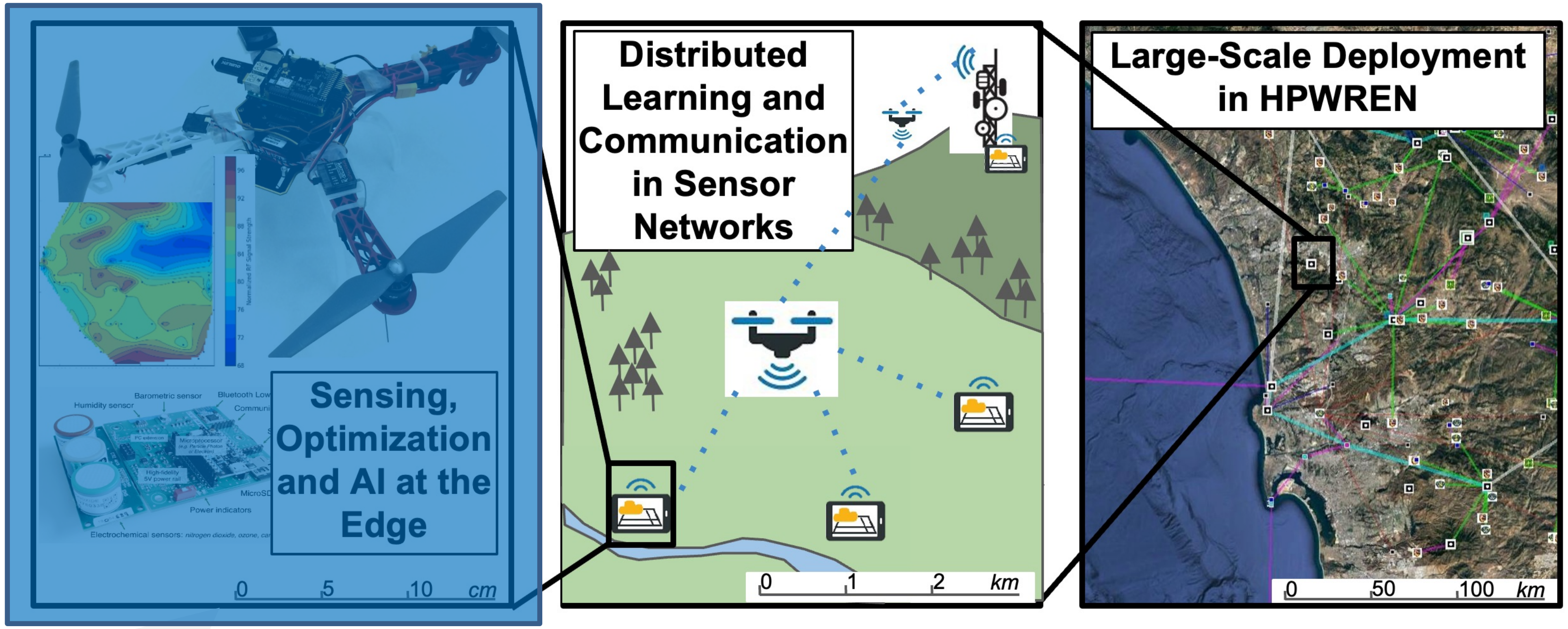
Real-world sensor network deployment in the wilderness

Large-Scale Deployment

- High Performance Wireless Research and Education Network **(HPWREN)** is an environmental monitoring cyberinfrastructure for research, education and public safety realms
  - Wireless connectivity covers 20K sq. mile area in San Diego, Riverside and Imperial counties
  - Numerous sensors with live feeds
- We plan to deploy a sub-sensor network in HPWREN, which provides
  - A real in-place deployment in noisy wild areas
  - An evaluation platform for on-device lifelong, federated and multimodal learning methods



Works Planned Ahead



How to train under resource constraints w/o performance loss?

Revisit: General Intelligence in Complex and Dynamic Environments

How do you navigate in an unfamiliar place?

Computer

Segment Everything, ChatGPT, etc

Tons of resources and million of dollars!!

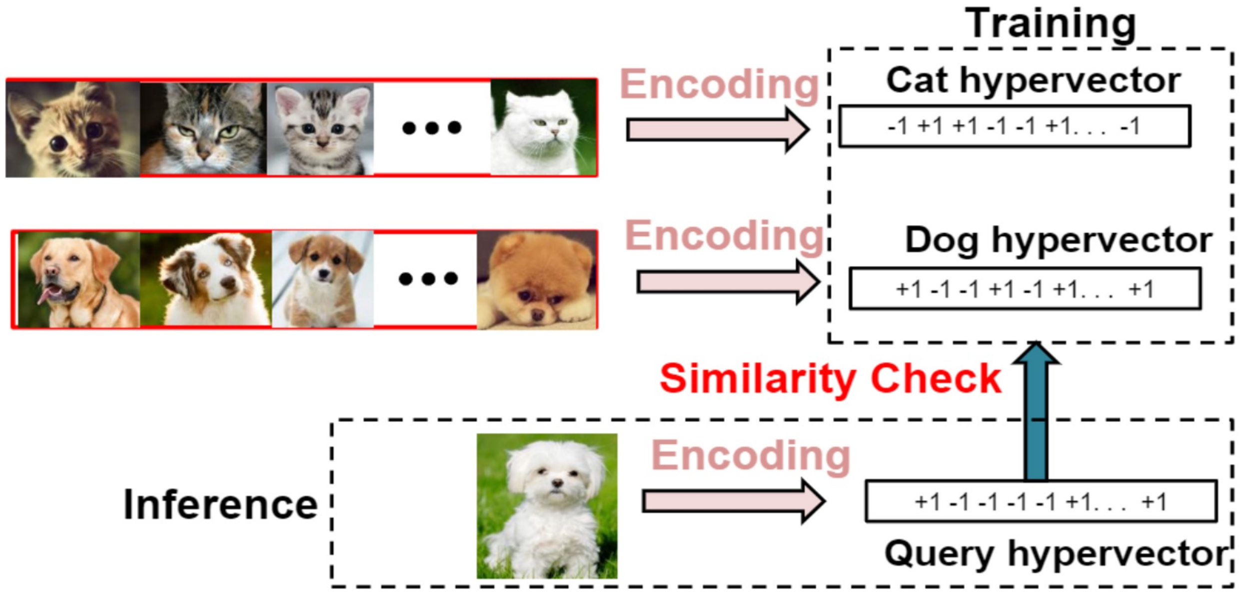
Human

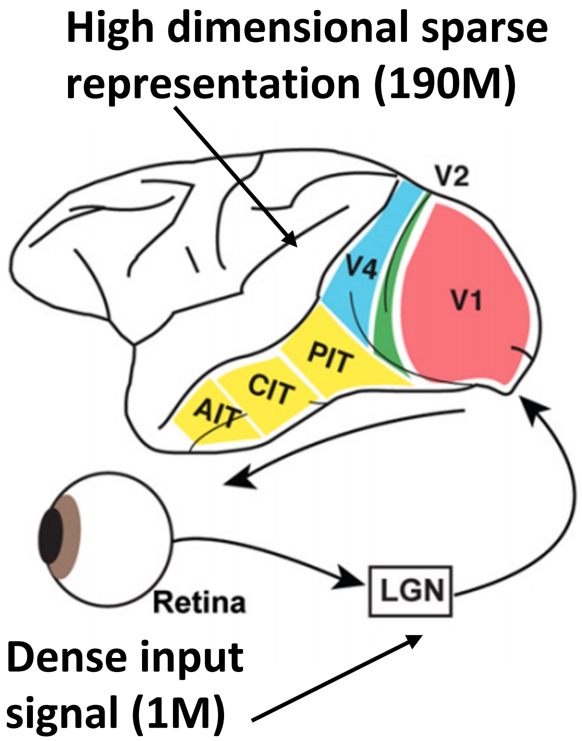
1X 1X **OR** 1X

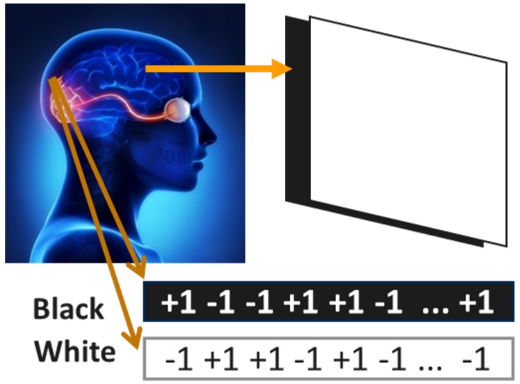
~20W brain or 100W whole body

Brain-Inspired HD Computing

Dense sensory input is mapped to **high-dimensional sparse representation** on which brain operates [Babadi and Sompolinsky 2014]







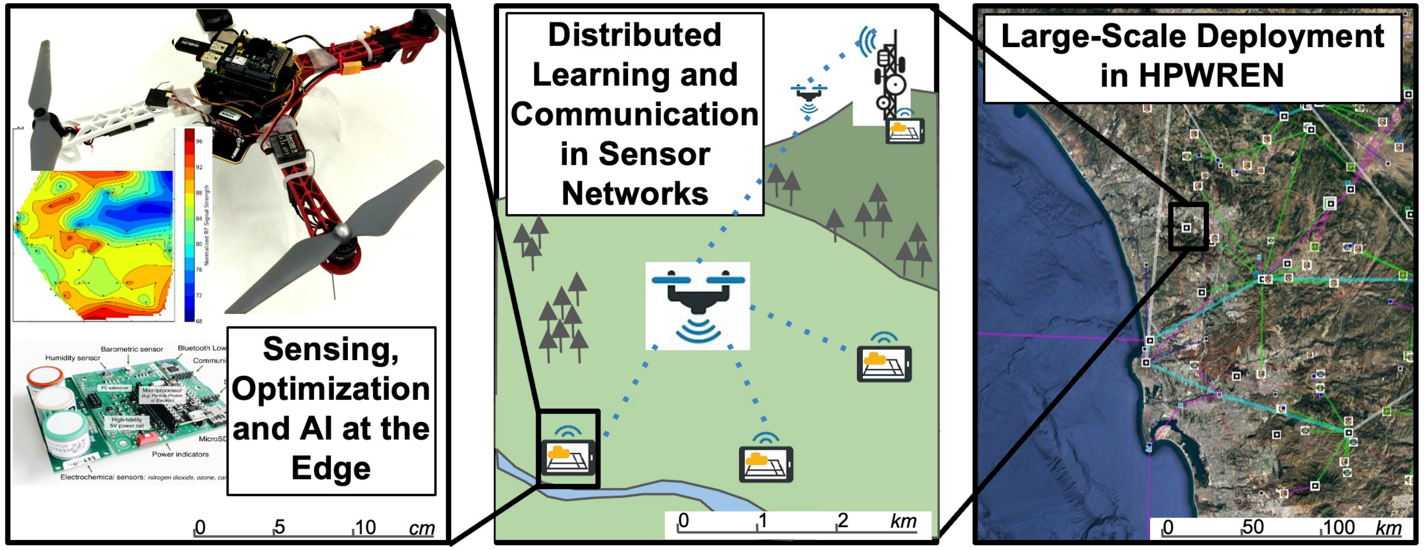
Benefits of HD computing:

- Easy-to-parallelize and hardware-friendly operations
- Fast single-pass training
- Energy-efficient & robust to noise

Question: Can we utilize these benefits to design lightweight on-device lifelong learning algorithm?

Conclusion

- Enabling ML training pervasively in IoT applications is an active research area, which calls for sophisticated designs on single device level, sensor network level, and large-scale deployments level.



- My PhD research targets at contributing a full stack of technologies in all three levels

Acknowledgements

- I am fortunate to work with multiple faculties and industrial collaborators
  - Prof. Tajana Šimunić Rosing (UCSD)
  - Dr. Ludmila Cherkasova (Arm Research)
  - Prof. Sicun Gao (UCSD)
  - Prof. Arya Mazumdar (UCSD)
  - Prof. Yunhui Guo (UT Dallas)

- I would like to acknowledge the contributions from the undergrad/MS students
  - Emily Ekaireb
  - Quanling Zhao
  - Shengfan Hu