

Machine  
Intelligence  
&  
Networked  
Data  
Science

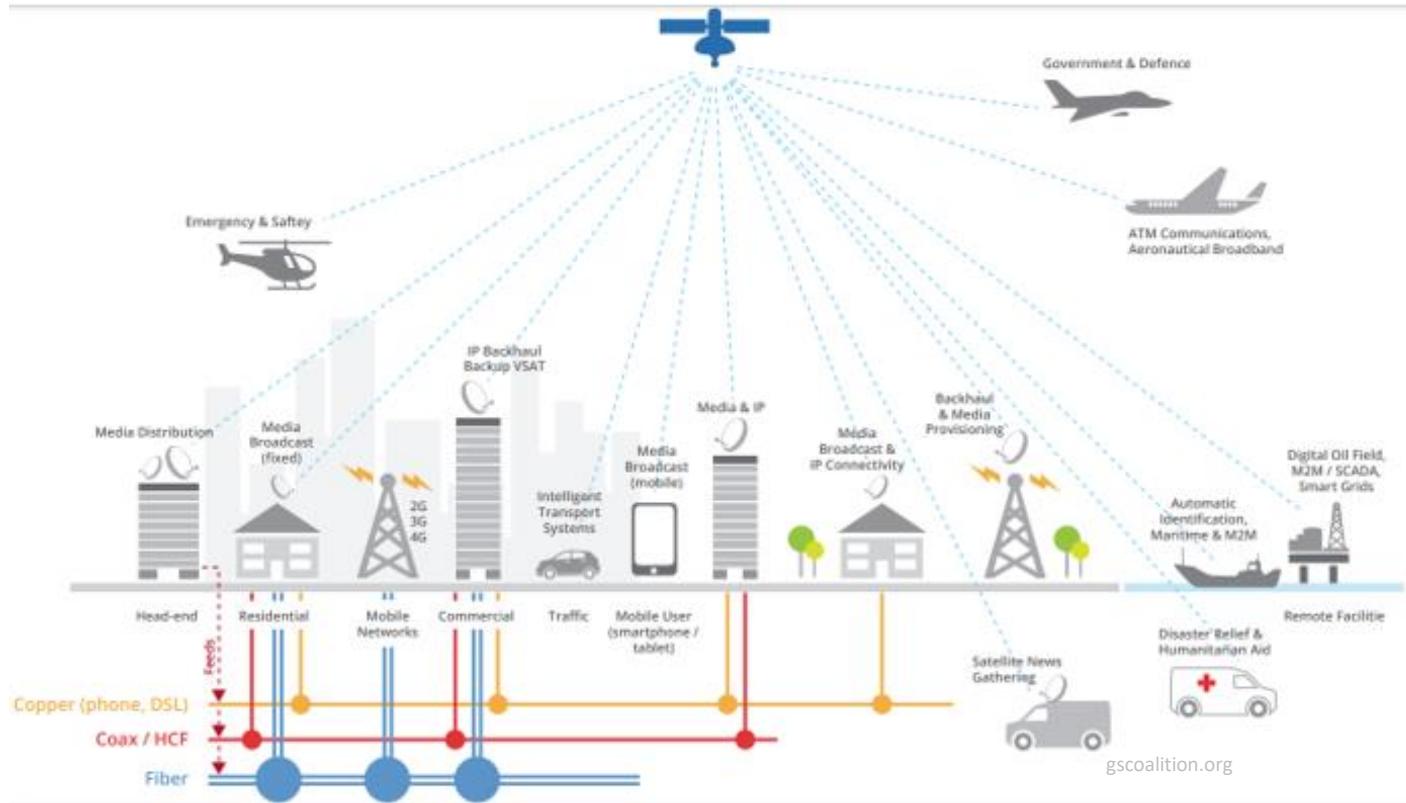
# Structured and Resource-Constrained Collaborative Learning

Abolfazl Hashemi

ML Seminar, Purdue University

Machine Intelligence and Networked Data Science Lab (**MINDS**)

# The Era of Collaborative Systems



Satellite mesh network

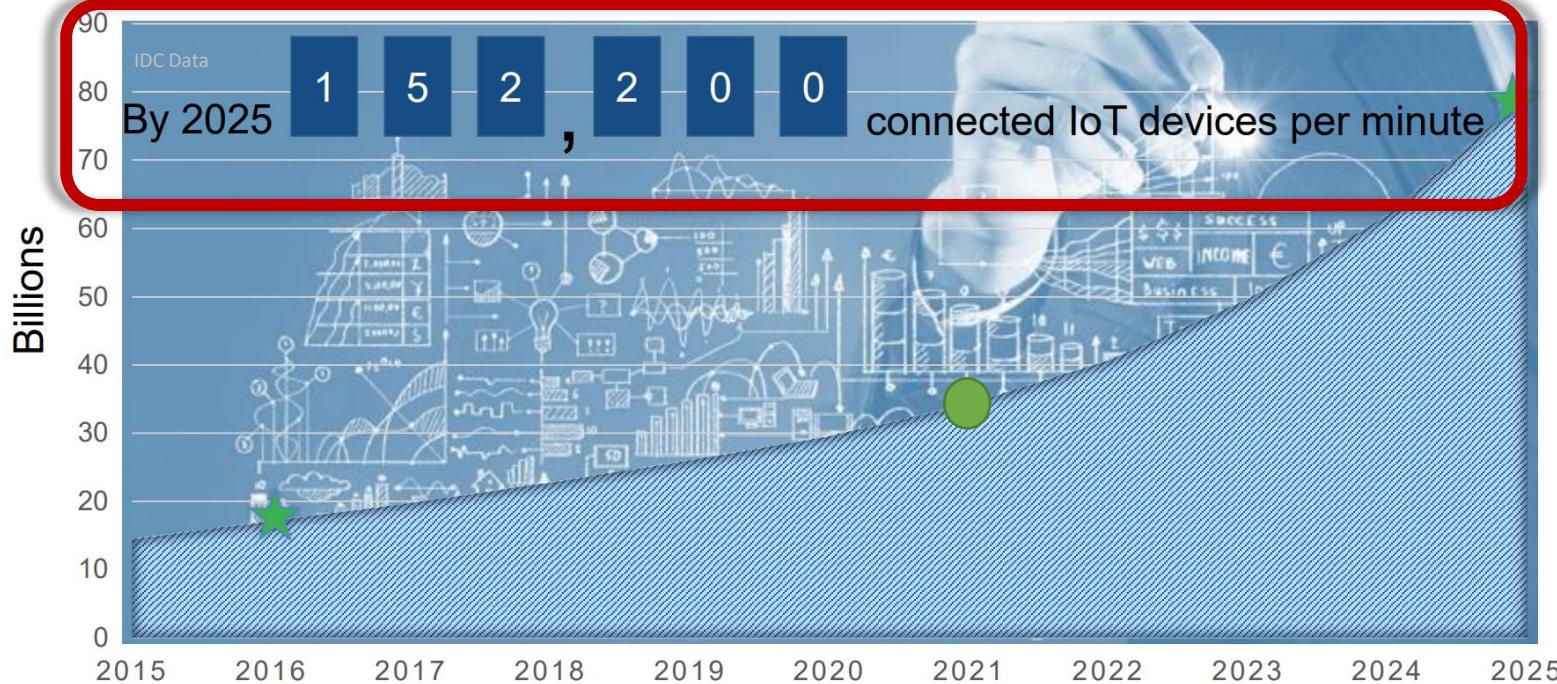


Smart grids



Creating reliable and effective collaborative systems that are highly secure, robust and economically viable

# Collaborative Systems: Large-Scale and Heterogenous

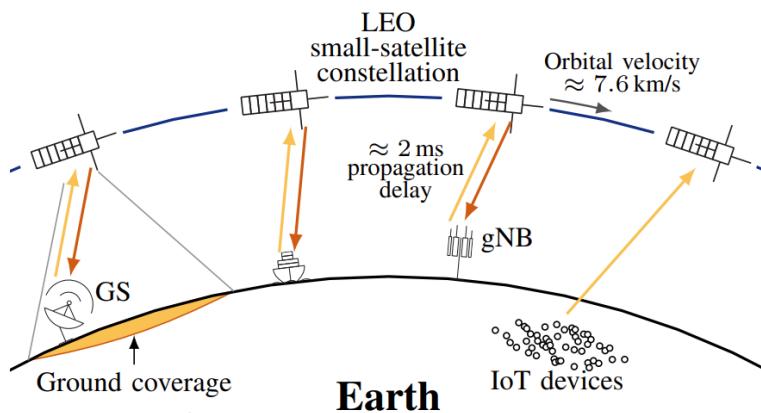
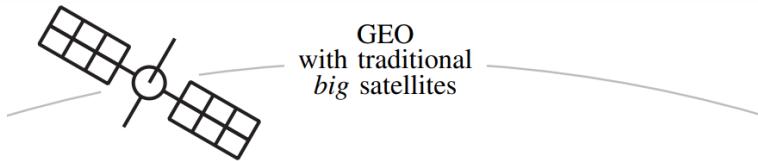


Rapidly increasing network size

High degree of heterogeneity

How can we enable **scalable** deployment of collaborative learning in **presence of heterogeneity**?

# Collaborative Systems: Embodied Agents

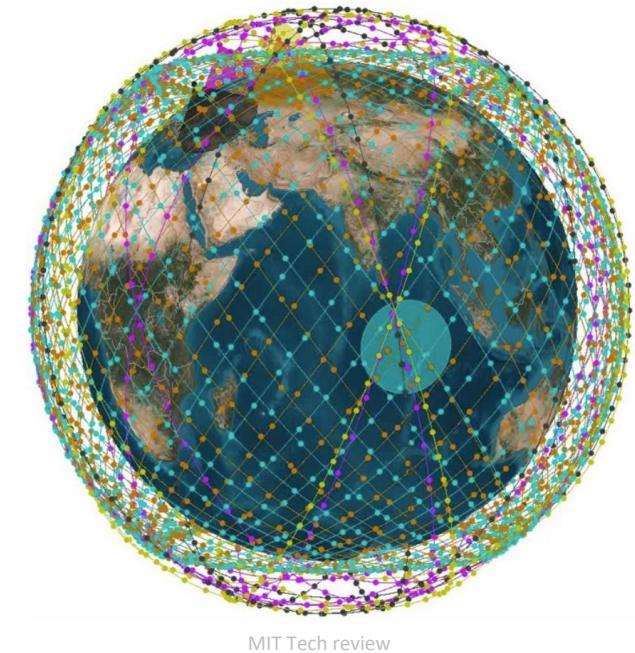


Rapidly evolving environments

Limited energy budgets

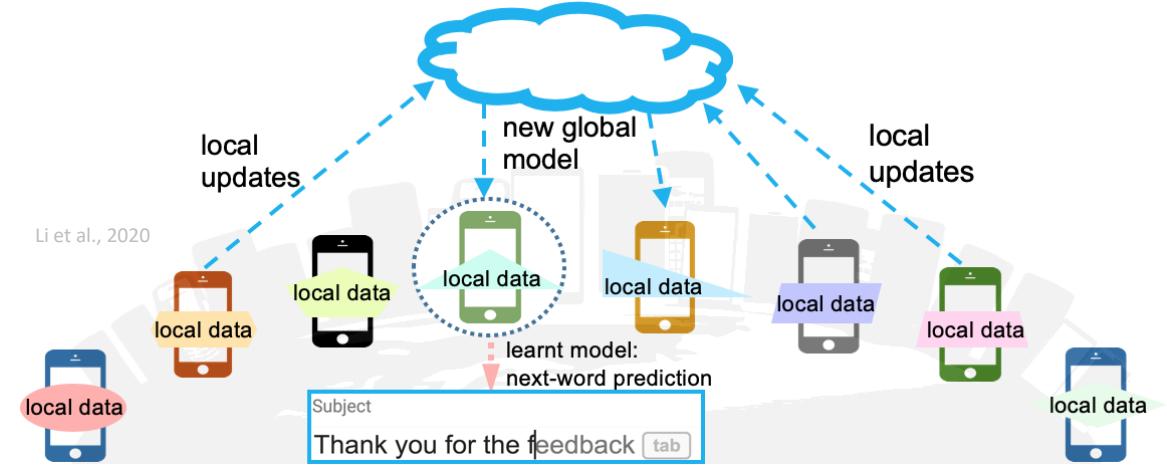
Costly observation gathering

4000+ LEO satellites in SpaceX network



How can we design **low-cost and energy-efficient** collaborative learning systems capable of operating in **rapidly evolving environments**?

# Collaborative Systems: Limited Communication Budget



Unreliable communication

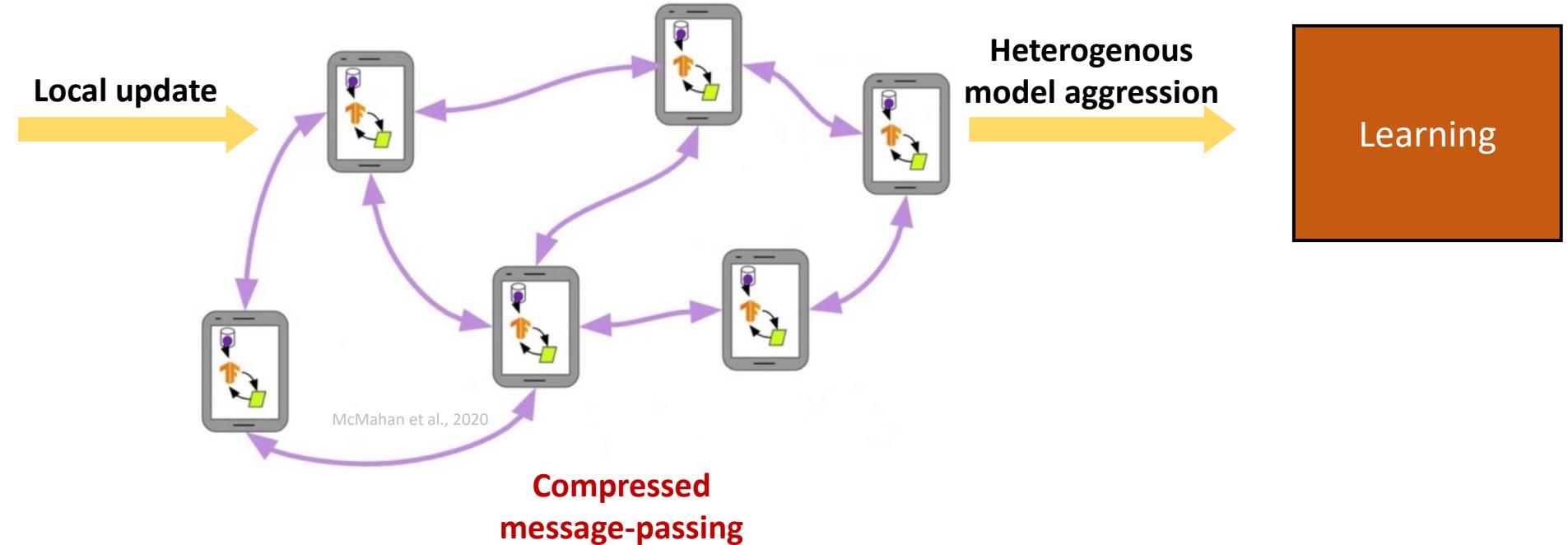
Limited bandwidth

How can we design **robust and communication-efficient collaborative learning systems?**

# Communication-Efficient Federated and Distributed Learning



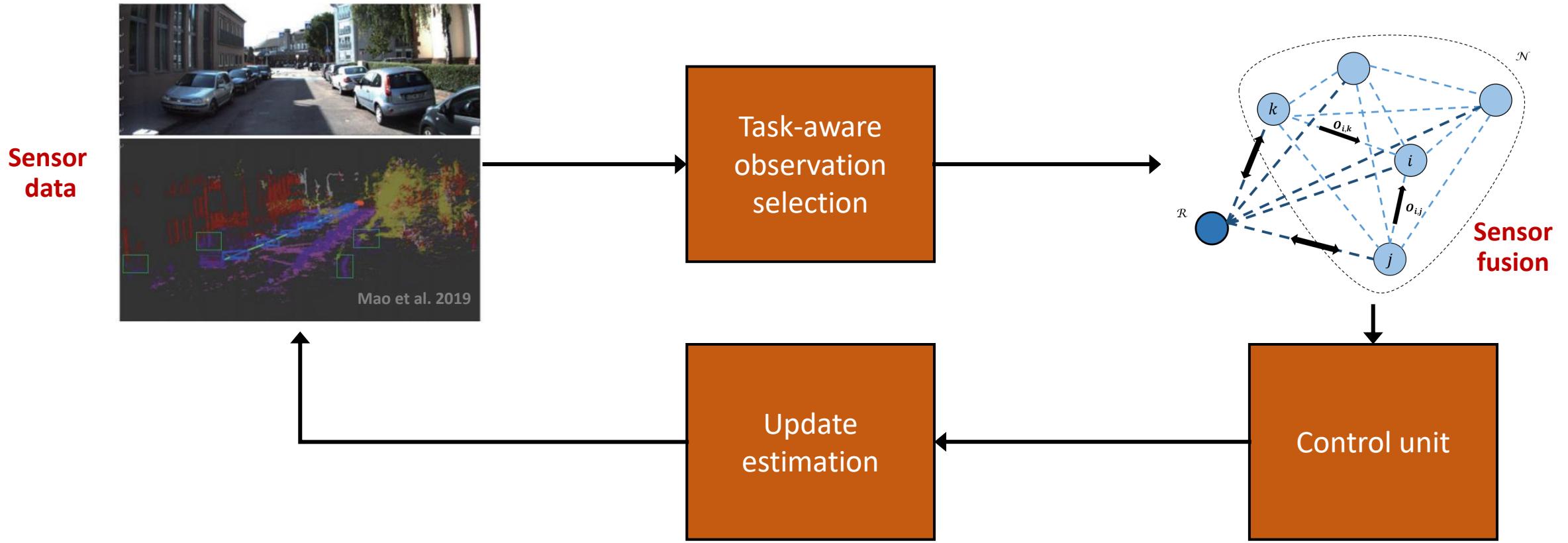
Local data



## Contributions:

- Model aggregation and communication strategies for distributed learning
- Developing communication-efficient algorithms with provable guarantees

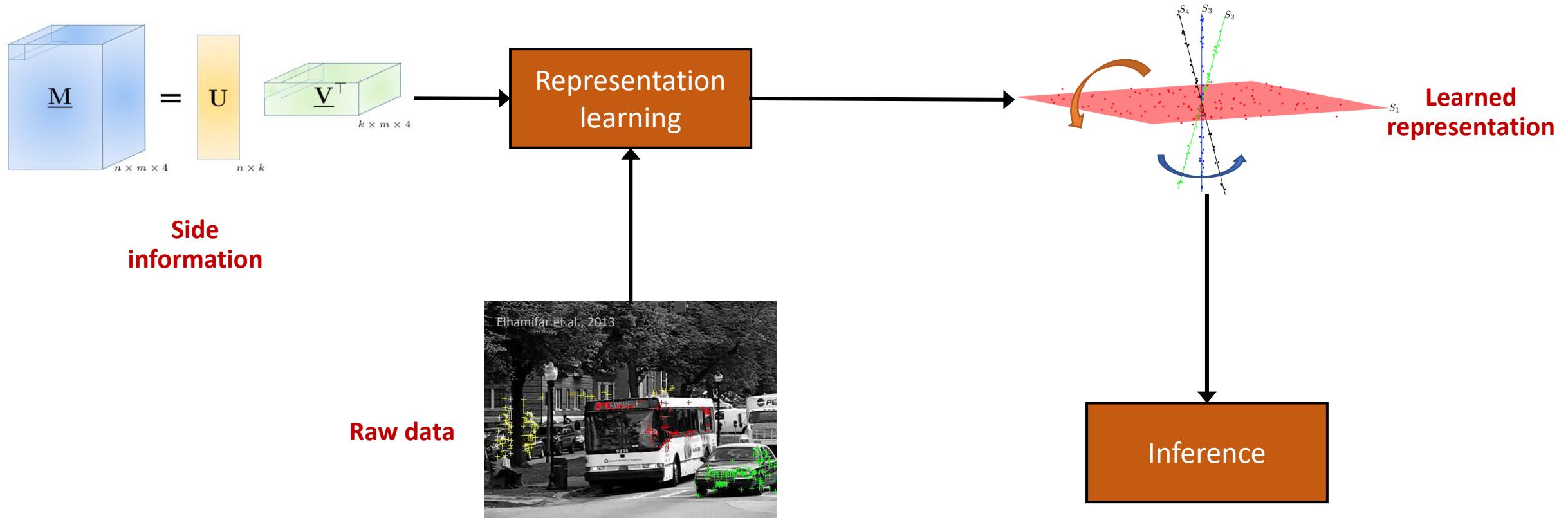
# Efficient Observation Selection and Information Gathering



## Contributions:

- Task-aware **observation selection criteria** for sensing networks
- Developing efficient feature selection algorithms with **near-optimal utilities**

# Data-Scarce Parsimonious Representation Learning



## Contributions:

- Representation learning for unsupervised inference from **structured data**
- **Sparse approximation algorithms** for function approximation and regression

# Communication-Efficient Federated and Distributed Learning



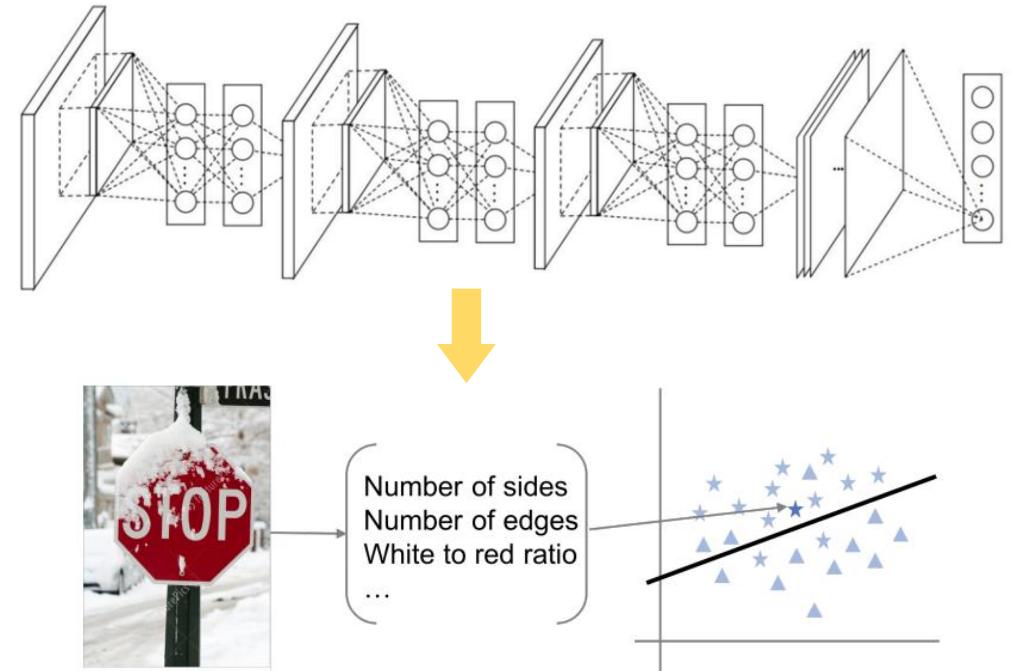
[Das, R., Hashemi, A., Acharya, A., Sanghavi, S., Dhillon, I., Topcu, U., "Faster Non-Convex Federated Learning via Global and Local Momentum," Submitted, 2021.]

[Chen, Y., Hashemi, A., Vikalo, H., "Communication-Efficient Variance-Reduced Decentralized Stochastic Optimization over Time-Varying Directed Graphs," Submitted, 2021.]

[Hashemi, A., Acharya, A., Das, R., Vikalo, H., Sanghavi, S., Dhillon, I., "On the Benefits of Multiple Gossip Steps in Communication-Constrained Decentralized Optimization," Submitted, 2021.]

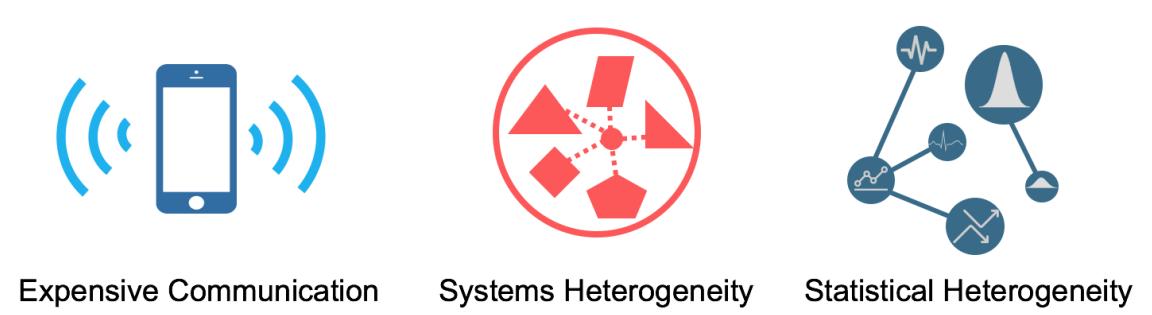
[Chen, Y., Hashemi, A., Vikalo, H., "Decentralized Optimization on Time-Varying Directed Graphs under Communication Constraints," International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2021.]

# Collaborative Learning in Connected Systems



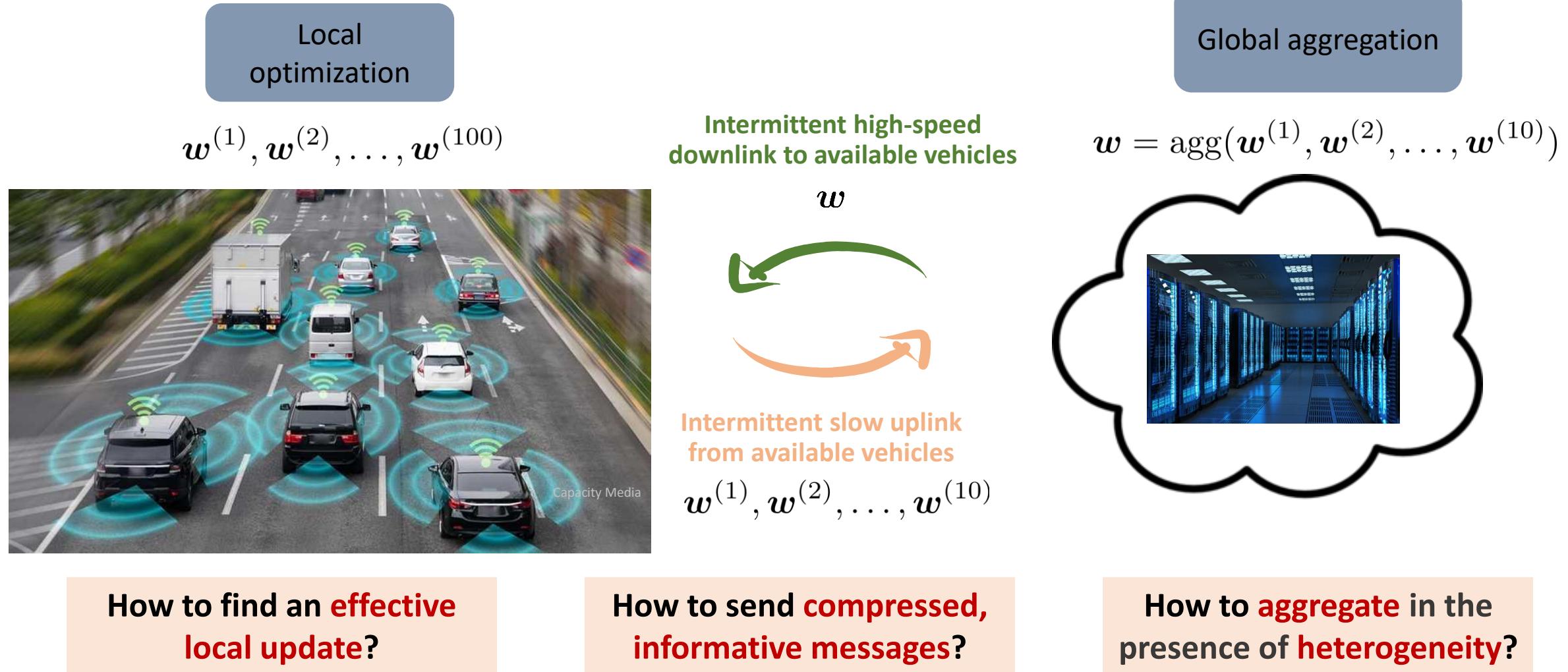
Limited local data

Collaboration via cloud

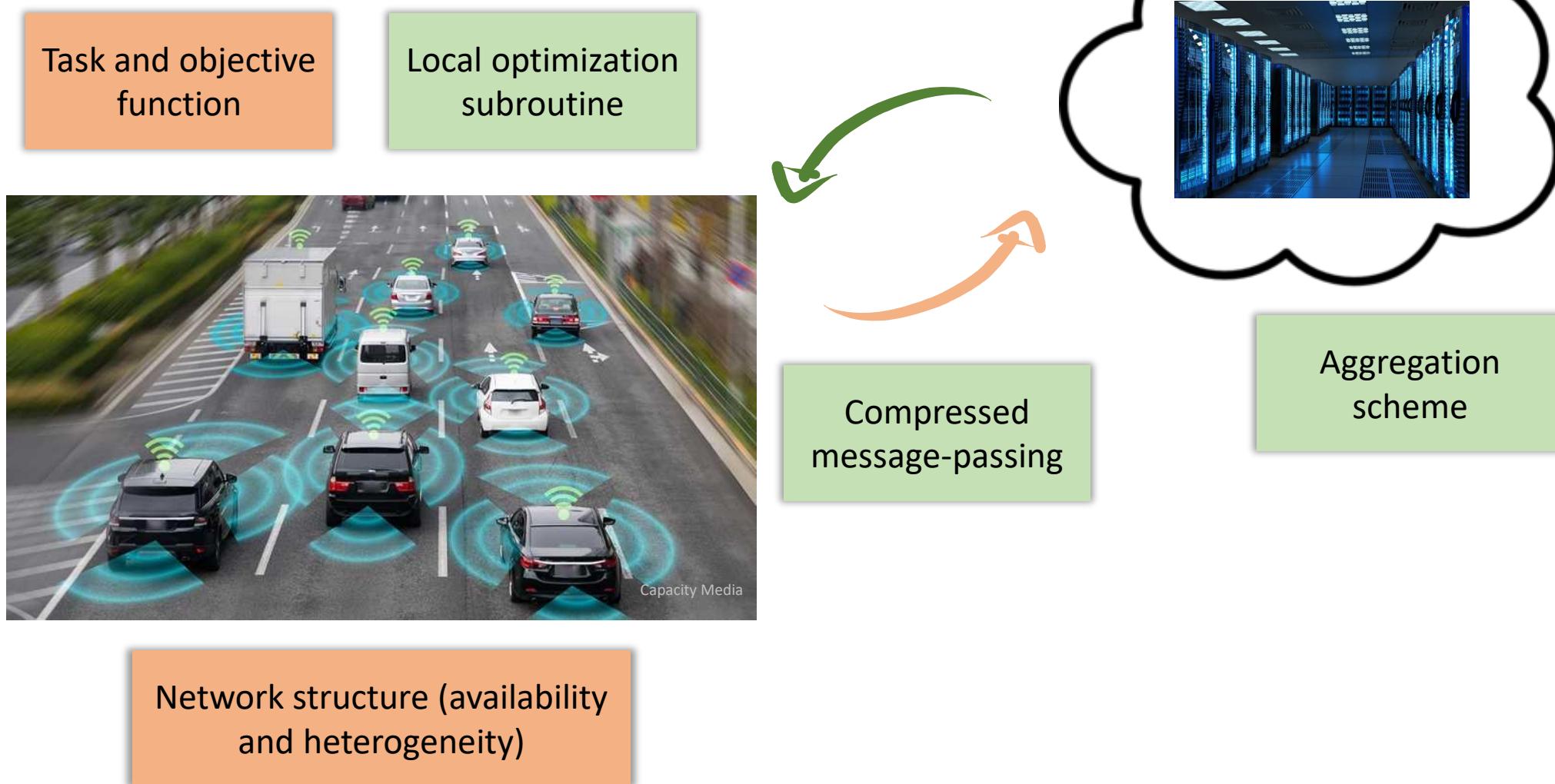


Li et al., 2019

# Communication-Efficient Federated Learning



# Components of the Problem



# Task and Objective Function

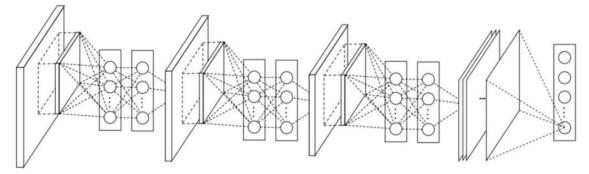
## Empirical Risk Minimization (ERM)



$$\min_{\mathbf{w}} \quad f(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{w}) \quad \text{where} \quad \hat{f}_i(\mathbf{w}) = \frac{1}{n_i} \sum_{j=1}^{n_i} \hat{f}_{i,j}(\mathbf{w})$$

Number of devices

Model parameters



Number data points  
at device i

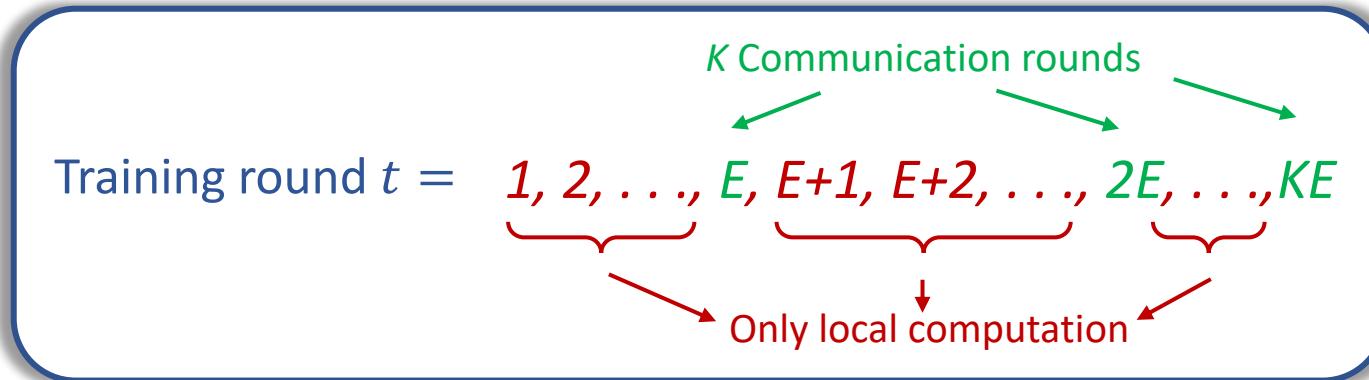
### Smooth loss functions

$$\|\nabla \hat{f}_{i,j}(\mathbf{x}) - \nabla \hat{f}_{i,j}(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|$$



# Network Structure and Heterogeneity

**Periodic message-passing:** devices communicate with the server **intermittently**



**Partial participation:** only  $r$  out of  $n$  devices available each communication round ( $r \ll n$ )

$$\|\tilde{\nabla} f_i(\mathbf{w}; \mathcal{B}) - \nabla f_i(\mathbf{w})\| \leq \sigma_b$$

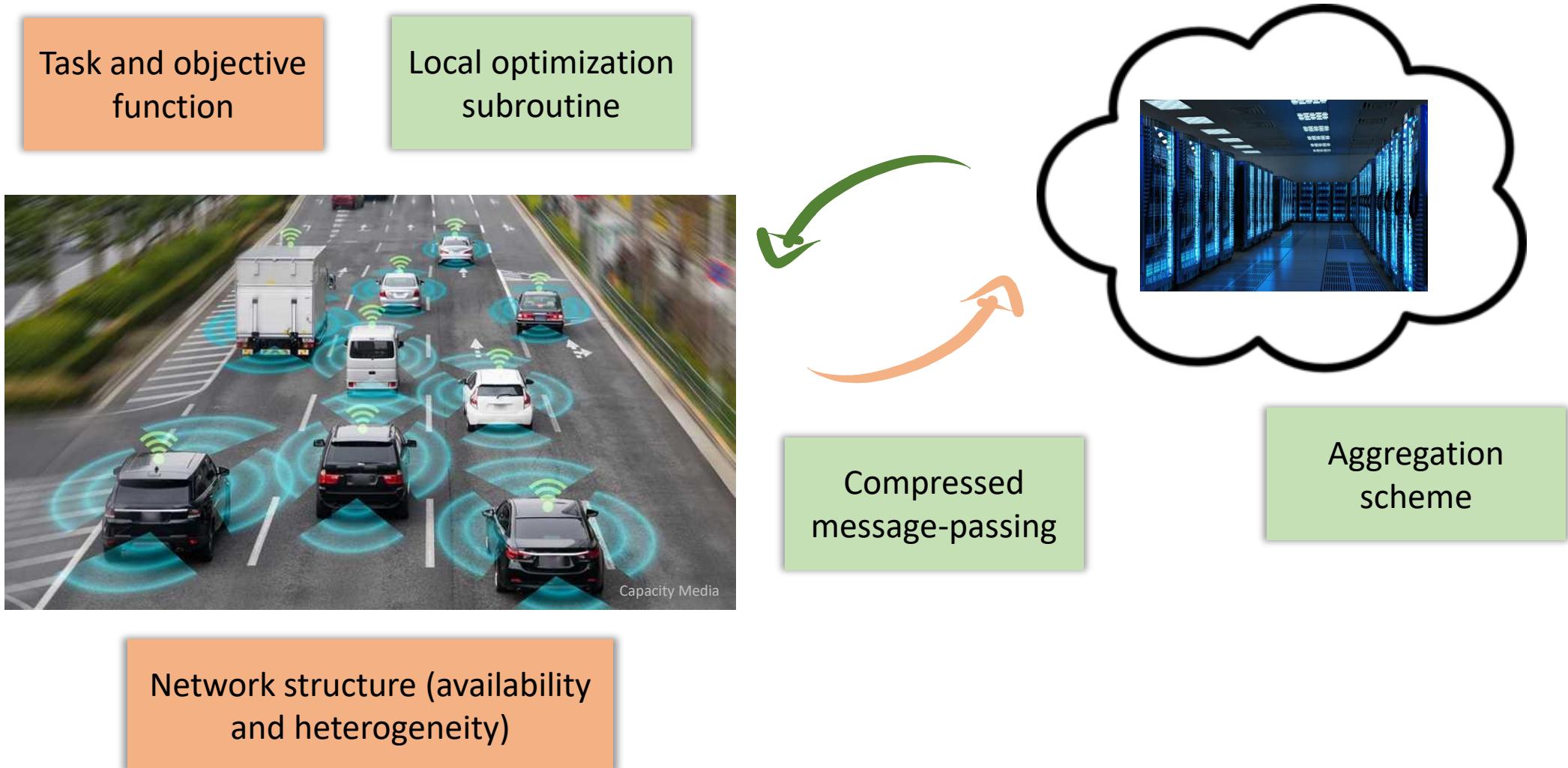
Stochastic gradient  
approximation error

**Bounded  
dissimilarity**

$$\|\nabla f_i(\mathbf{w}) - \nabla f(\mathbf{w})\|^2 \leq \sigma_r^2$$

Local functions      Global function

# Components of the Problem



# Stochastic Gradient Descent (SGD)

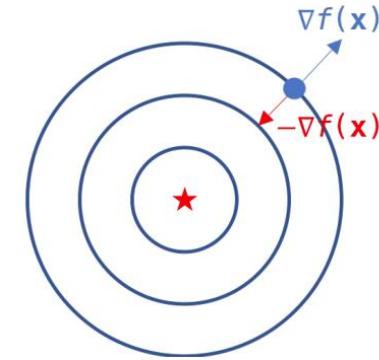
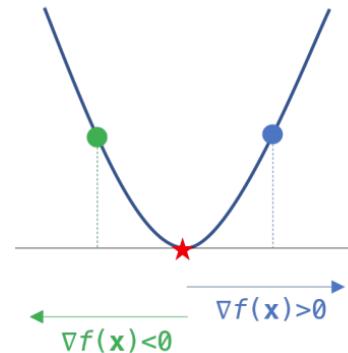
Search for a point where the gradient is small

$$\|\nabla f(\mathbf{w})\|^2 \leq \epsilon$$

Stochastic first-order update

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \tilde{\nabla} f_i(\mathbf{w}_t; \mathcal{B}_t)$$

**Intuition:** Take a step in the direction opposite to the gradient



What do we know about the performance of SGD?

**Theorem** (Convergence of SGD)

$$T = \mathcal{O} \left( \frac{\sigma_b^2}{\epsilon^2} + \frac{1}{\epsilon} \right)$$

Bottou 2018

**Theorem** (Lower bounds)

$$T = \mathcal{O} \left( \frac{1}{\epsilon^{1.5}} \right)$$

Arjevani et al. 2019

# Local Momentum-Based Variance Reduction

A recursive stochastic first-order update

$$\mathbf{v}_\tau^{(i)} = \tilde{\nabla} f_i(\mathbf{w}_\tau^{(i)}; \mathcal{B}_\tau^{(i)}) + (\mathbf{v}_{\tau-1}^{(i)} - \tilde{\nabla} f_i(\mathbf{w}_{\tau-1}^{(i)}; \mathcal{B}_\tau^{(i)}))$$

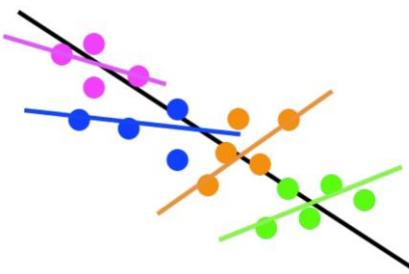
$$\mathbf{w}_{\tau+1}^{(i)} = \mathbf{w}_\tau^{(i)} - \eta \mathbf{v}_\tau^{(i)}$$

**Lemma:** Reduces the variance of stochastic gradient

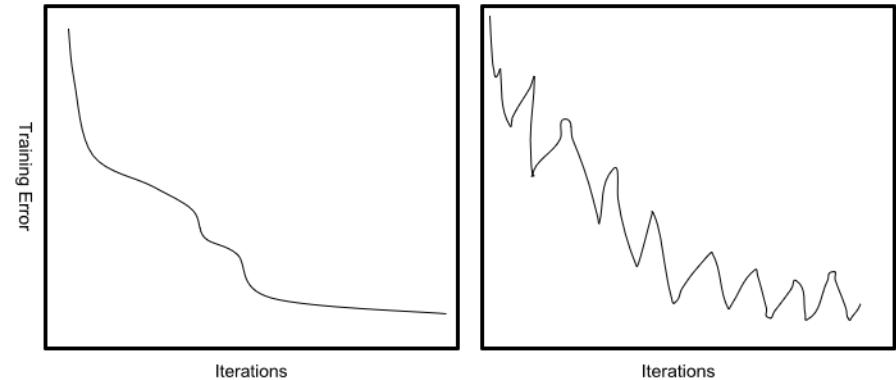
$$\sum_{\tau=0}^{E-1} \mathbb{E}[\|\mathbf{v}_\tau^{(i)} - \nabla f_i(\mathbf{w}_\tau^{(i)})\|^2] \leq O(E^2 \eta^2)$$

Hashemi et al., 2021

$$\underbrace{\|\nabla f_i(\mathbf{w}) - \nabla f(\mathbf{w})\|^2}_{\text{Local functions}} \leq \sigma_r^2$$



How about heterogeneity?



Proposed

SGD

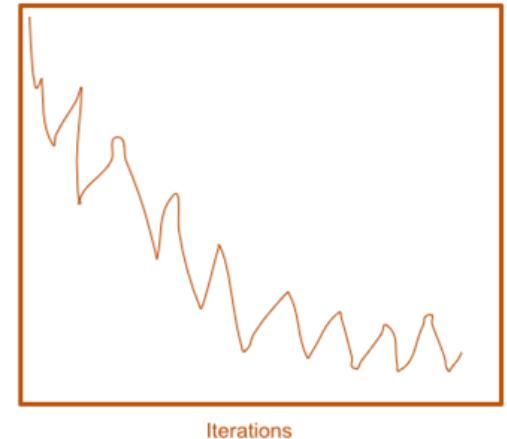
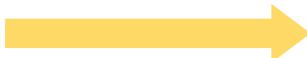
# Global Momentum-Based Variance Reduction

Simple aggregation:  $\mathbf{w} \leftarrow \frac{1}{r} \sum_{i \in \mathcal{S}} \mathbf{w}_E^{(i)}$

$$\mathbf{w} \leftarrow \mathbf{w} - \frac{\eta}{r} \sum_{i \in \mathcal{S}_k} \frac{\mathbf{w} - \mathbf{w}_E^{(i)}}{\eta}$$

A **generalized stochastic gradient**

Similar issue, but now  
because of **heterogeneity**



**Using momentum-based variance reduction  
for model parameters**

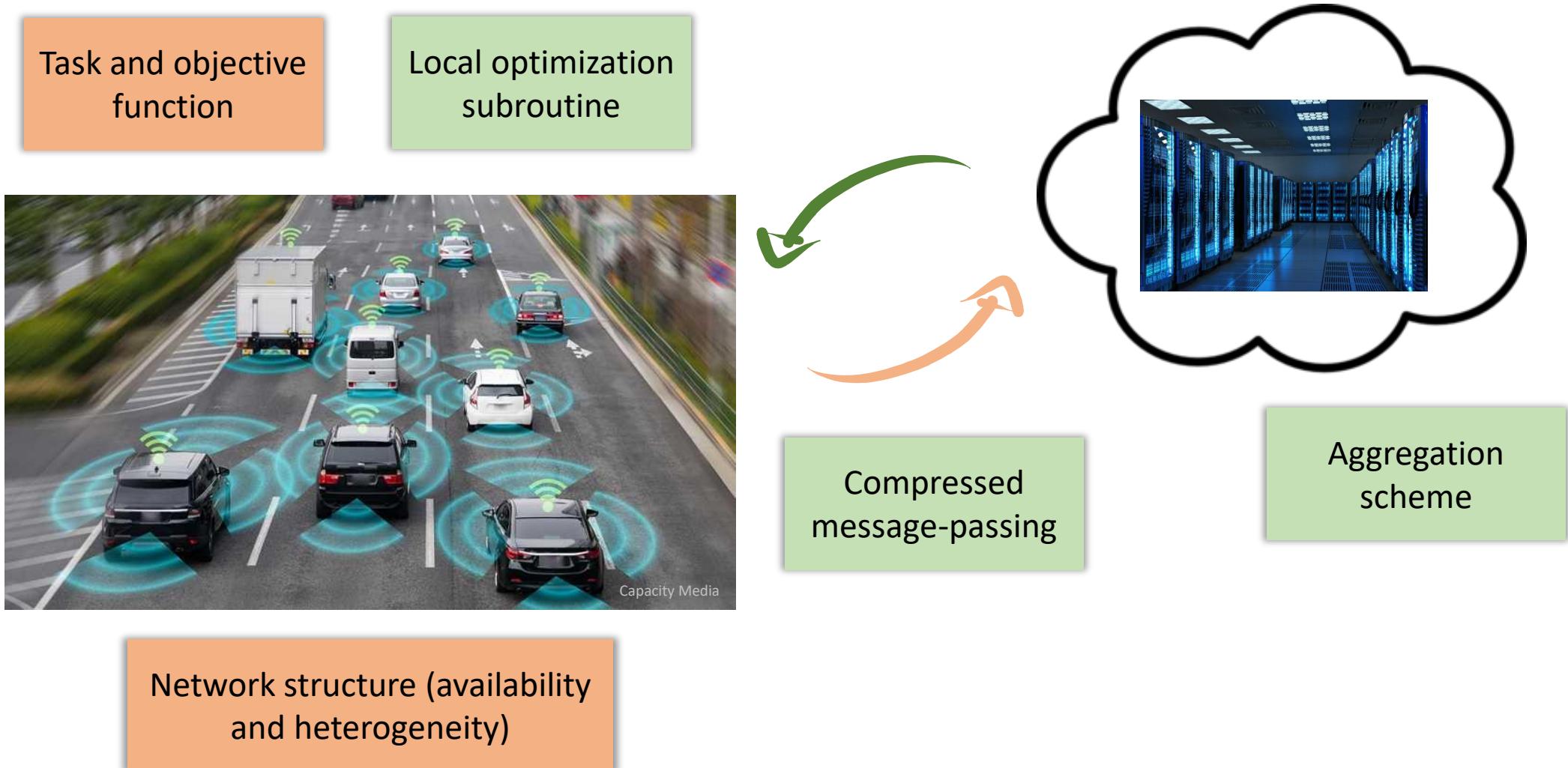
**Theorem (optimal rate)**

To get  $\mathbb{E}\|\nabla f(\mathbf{w}_K)\|^2 \leq \epsilon$  we need

$$K = \mathcal{O}\left(\frac{1}{\epsilon^{1.5}}\right)$$

Hashemi et al., 2021

# Components of the Problem

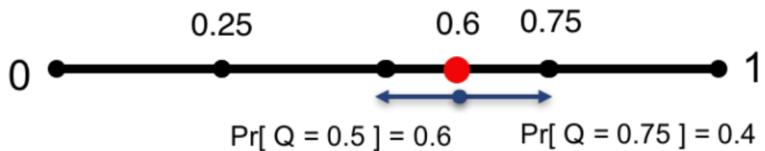


# Quantized Uplink Communication

*s-level* Stochastic quantization

[Alistarh et al., 2017]

$$Q_D(v_i) = \|\mathbf{v}\| \cdot \text{sgn}(v_i) \cdot \xi_i(\mathbf{v}, s)$$



Unbiased with small variance

**Uplink messages**

Each learner sends

$$Q_D(\underbrace{\mathbf{w}_k - \mathbf{w}_{k,E}^{(i)}}_{\text{previous global model} - \text{current local model}})$$

previous global model – current local model

# Local and Global Momentum-Based Variance Reduction

Local momentum-based variance reduction

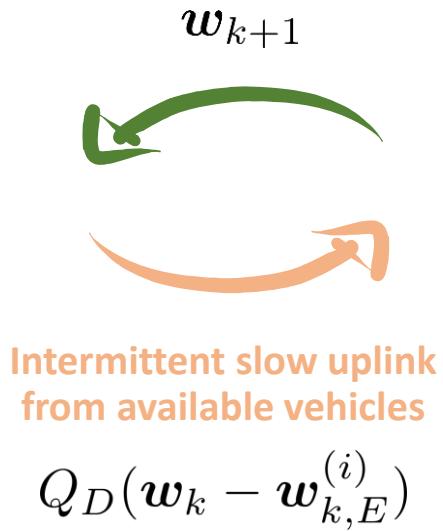
$$\boldsymbol{v}_\tau^{(i)} = \tilde{\nabla} f_i(\boldsymbol{w}_\tau^{(i)}; \mathcal{B}_\tau^{(i)}) + (\boldsymbol{v}_{\tau-1}^{(i)} - \tilde{\nabla} f_i(\boldsymbol{w}_{\tau-1}^{(i)}; \mathcal{B}_\tau^{(i)}))$$

$$\boldsymbol{w}_{\tau+1}^{(i)} = \boldsymbol{w}_t^{(i)} - \eta \boldsymbol{v}_\tau^{(i)}$$



Global momentum-based variance reduction

$$\boldsymbol{w}_{k+1} = \text{agg}(\{Q_D(\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)})\})$$



# Collaborative Learning of Multiclass Classifiers

10 classes, 50,000 images

**$n = 50$**  collaborative learners

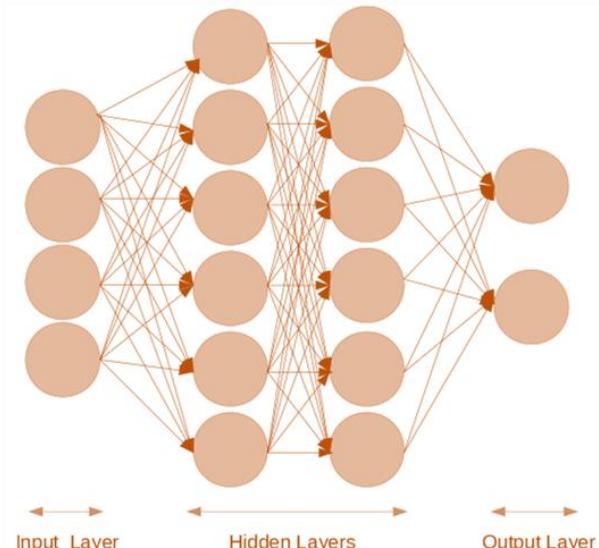
**Communication protocol:**  
**50% dropout rate ( $r=25$ )**

Communication **every 10 rounds**  
(Intermittency)

**Heterogenous case:**  
**2% of data available locally, from at most two classes**

**Homogenous case:**  
**2% of data available locally, i.i.d. among the devices**

CIFAR-10



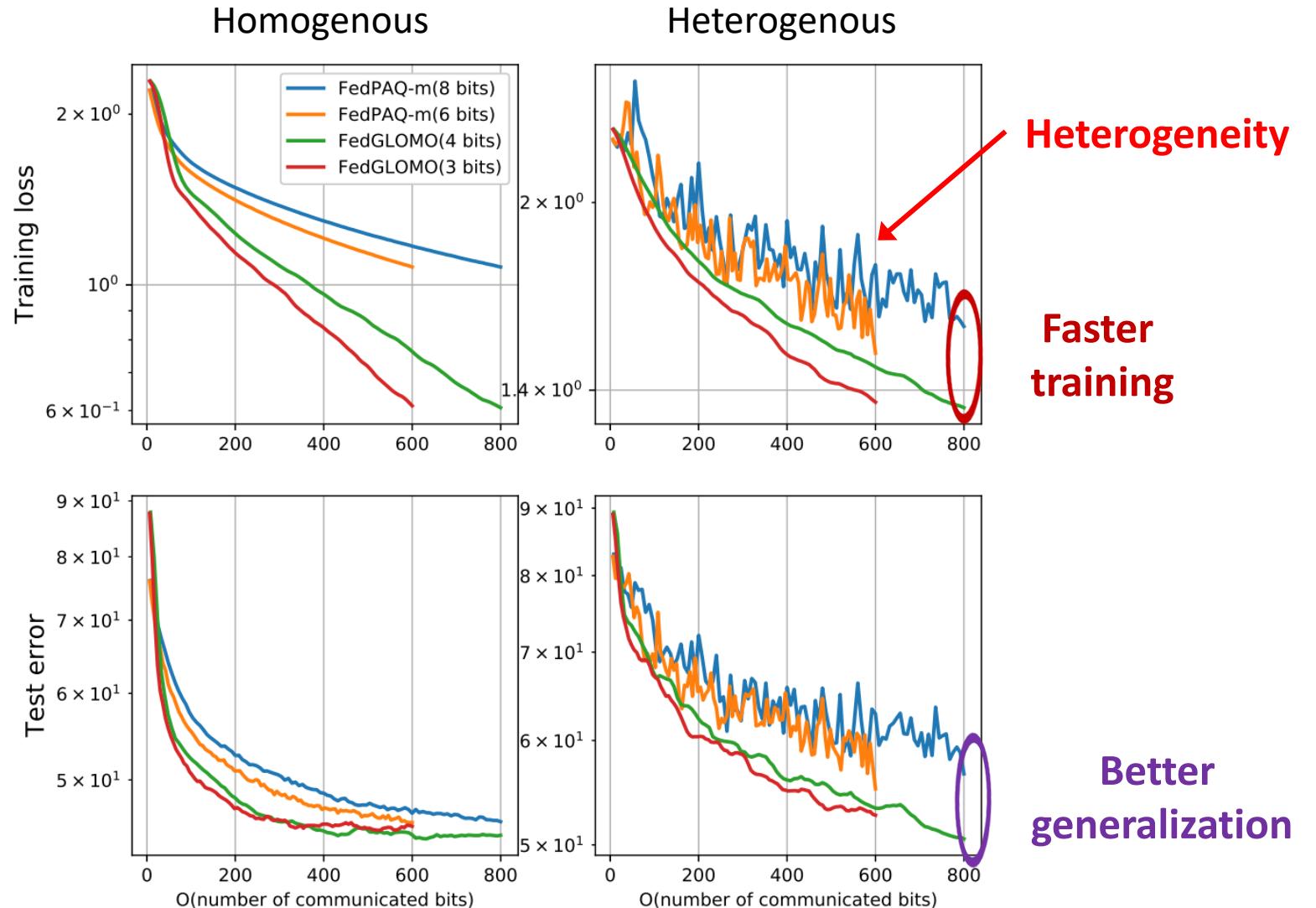
600 neurons

# Efficacy of Quantization and Momentum Mechanisms

Baseline: FedPaQ

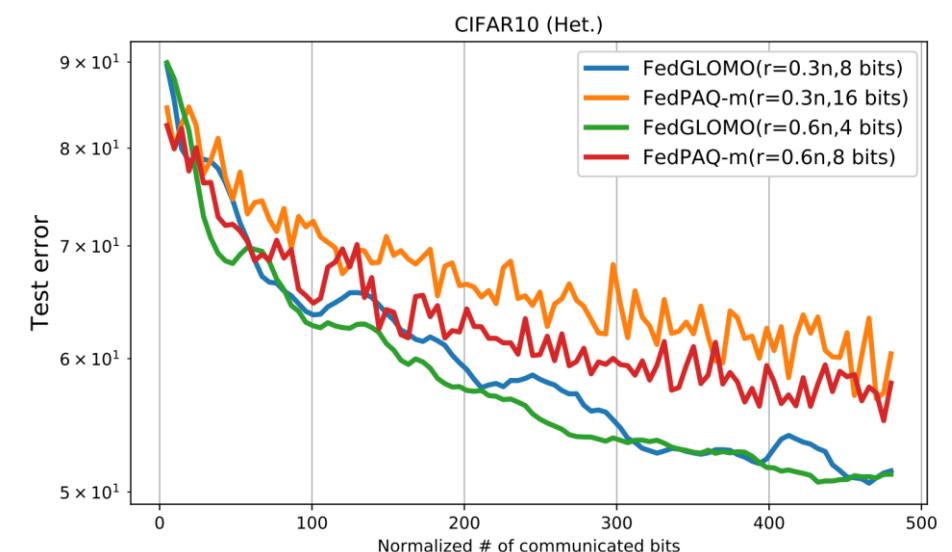
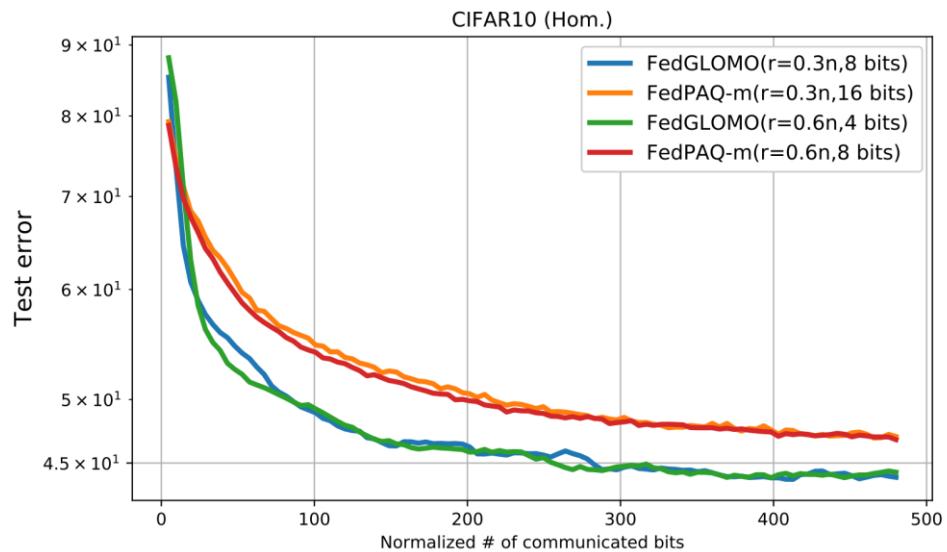
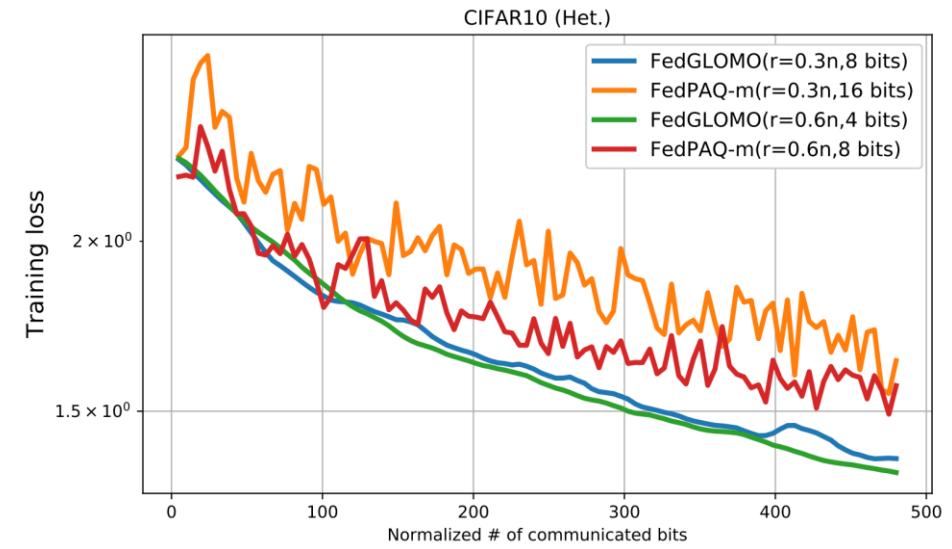
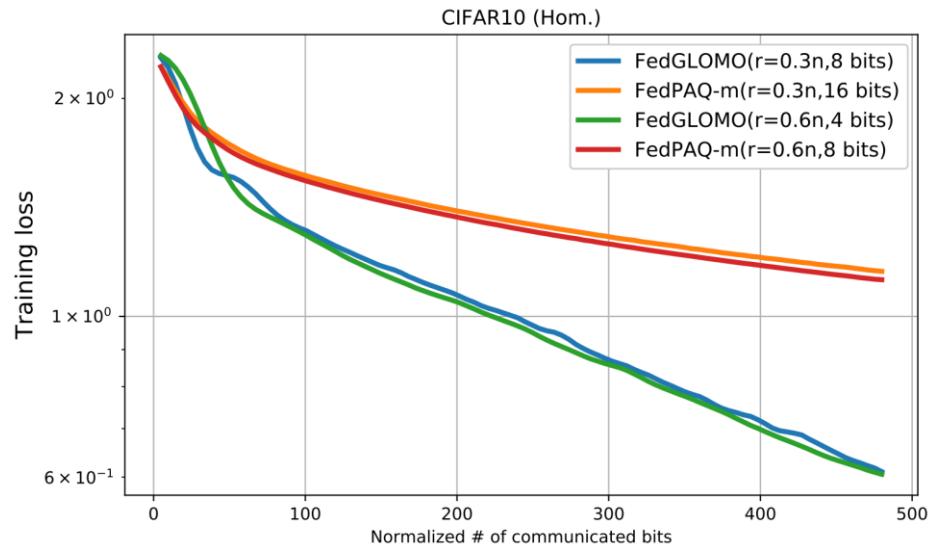
Proposed: FedGLOMO

An order of magnitude savings  
in communication resources



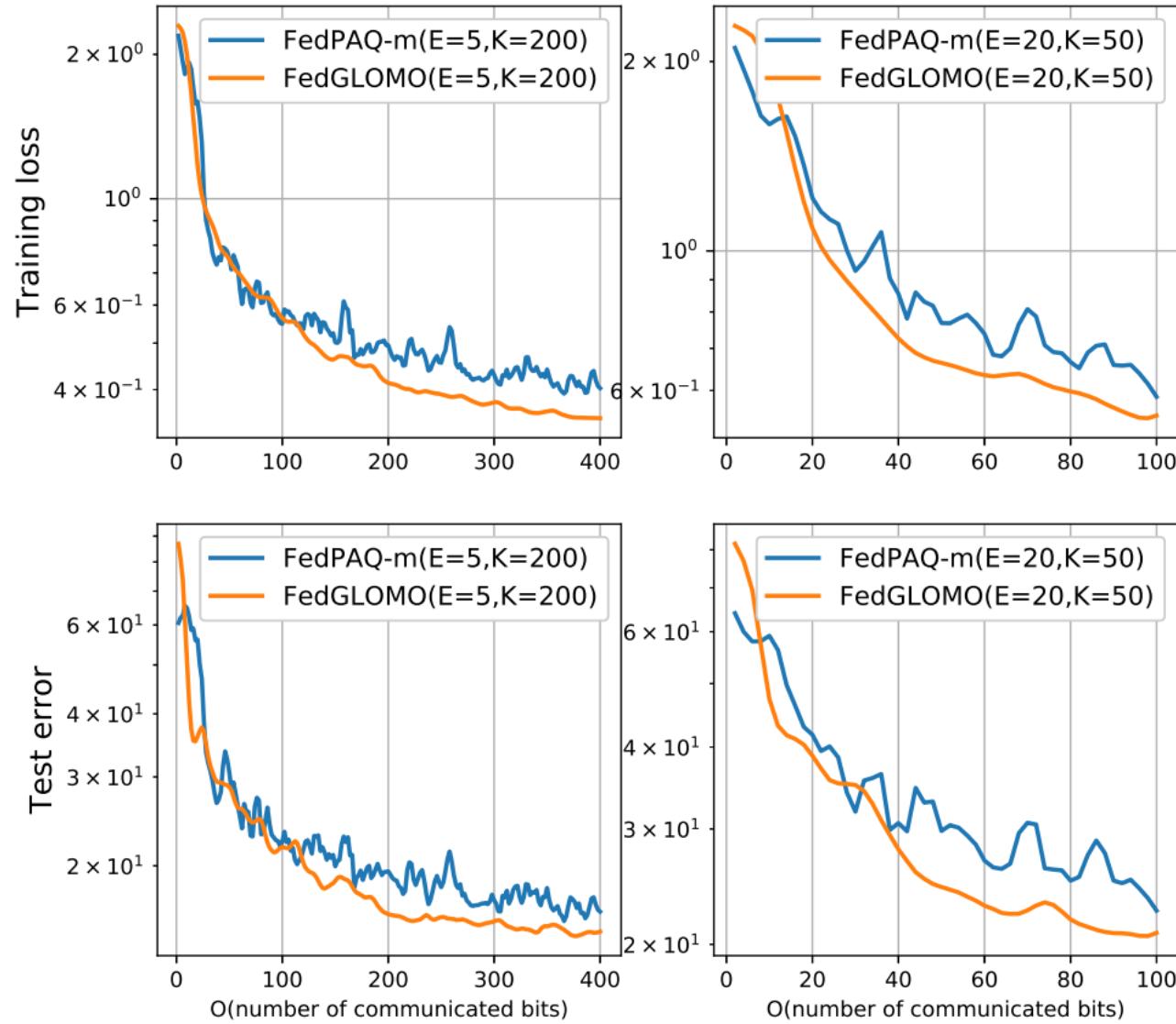
# Robustness to Unreliable Communication

Resiliency to device dropout (smaller  $r$ )

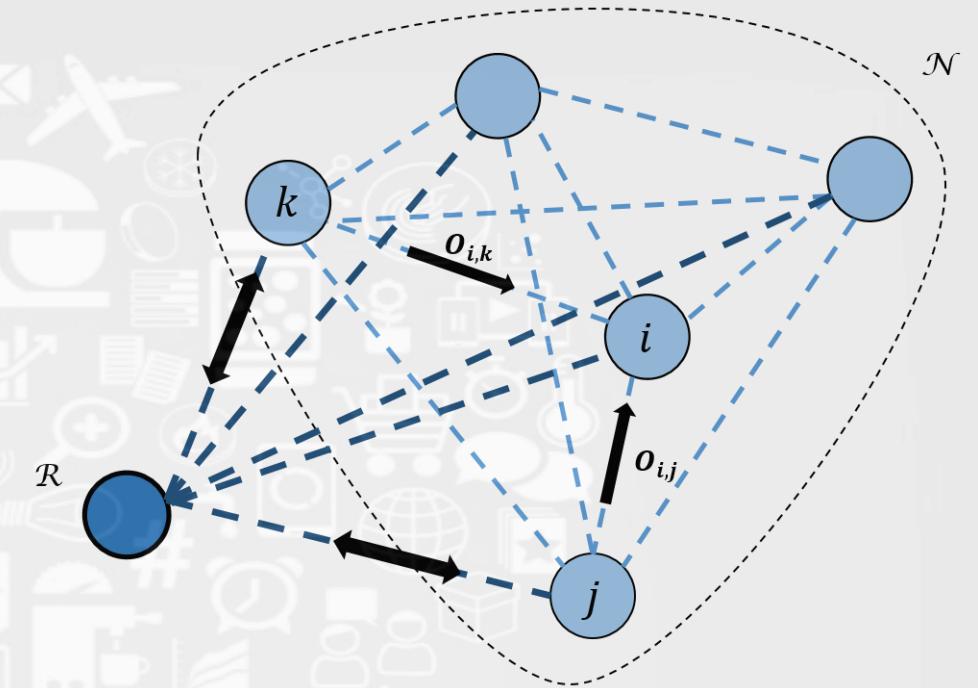


# Robustness to Unreliable Communication

Resiliency to device  
intermittent availability (larger  $E$ )



# Information Management in Resource-Constrained Sensing Networks



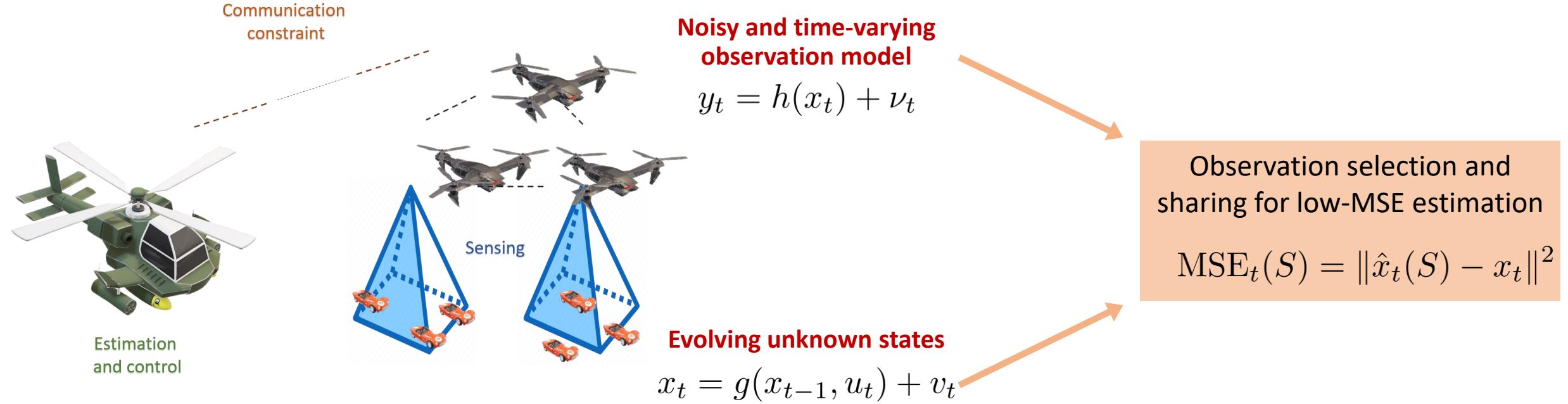
[Hashemi, A., Vikalo, H., de Veciana, G., “Progressive Stochastic Greedy Sparse Reconstruction and Support Selection,” Submitted, 2021.]

[Hashemi, A., Ghasemi, M., Vikalo, H., Topcu, U., “Randomized greedy sensor selection: Leveraging weak submodularity,” IEEE Transactions on Automatic Control, Jan. 2021.]

[Hashemi, A., Vikalo, H., de Veciana, G., “On the Performance-Complexity Tradeoff in Stochastic Greedy Weak Submodular Optimization,” International Conference on Acoustic, Speech and Signal Processing (ICASSP), 2021.]

[Hashemi, A., Ghasemi, M., Vikalo, H., Topcu, U., “Submodular Observation Selection and Information Gathering for Quadratic Models,” International Conference on Machine Learning (ICML), Long Beach, CA, June 2019.]

# Observation Selection for Sensing Networks



## Questions

- What should be the selection **criteria**?
- How can we perform the selection **efficiently** and with **guaranteed performance**?

# Observation Selection Criteria

Scalar functions of the predicted error covariance matrix  $f(P_t(S))$

Constrained combinatorial optimization

$$\hat{S} = \arg \max_{|S| \leq k} f(S)$$

NP-hard

Krause, 2011

$$f(\hat{S}) \geq (1 - e^{-\alpha}) f(S^*)$$

Weak-submodularity  
constant  $0 < \alpha \leq 1$

Optimal approximation guarantee

Krause, 2011

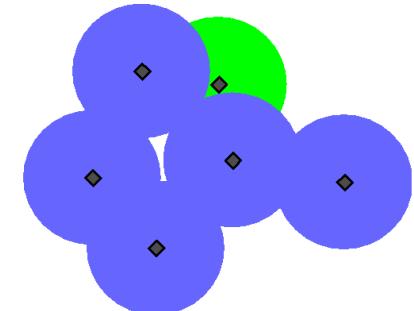
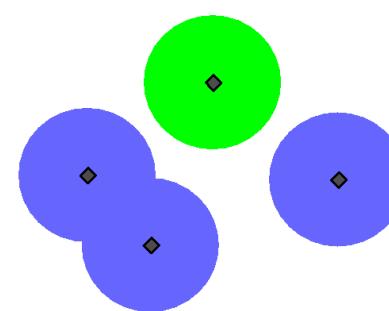
Approximate  
greedy solution



$f(S)$  has nice properties:

- Monotonicity
- Weak-submodularity

Hashemi et al., 2018



$$f(A \cup \{d\}) - f(A) \geq f(B \cup \{d\}) - f(B)$$

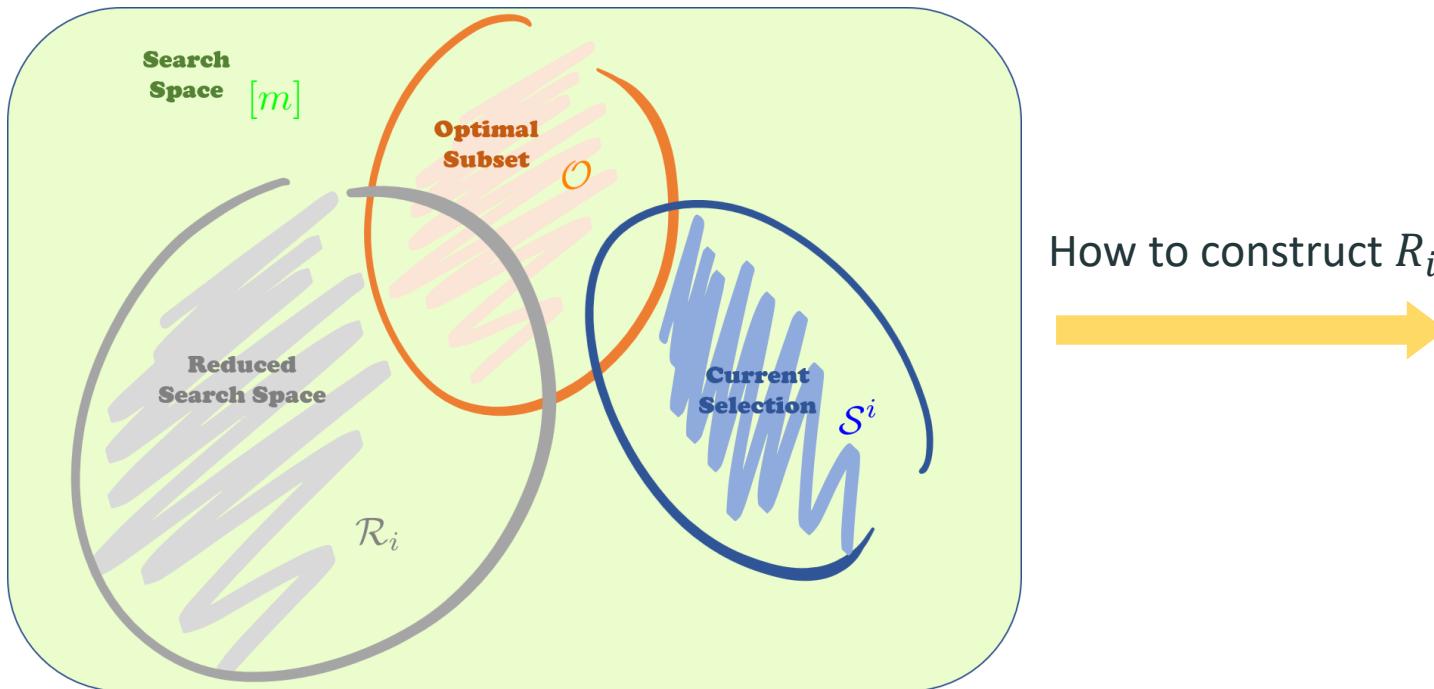
# Observation Selection in Large-Scale Networks

Greedy selection

Tight approximation guarantee

Prohibitive computational cost

Reduce the space of greedy by **random sampling**



How to construct  $R_i$

## Theorem

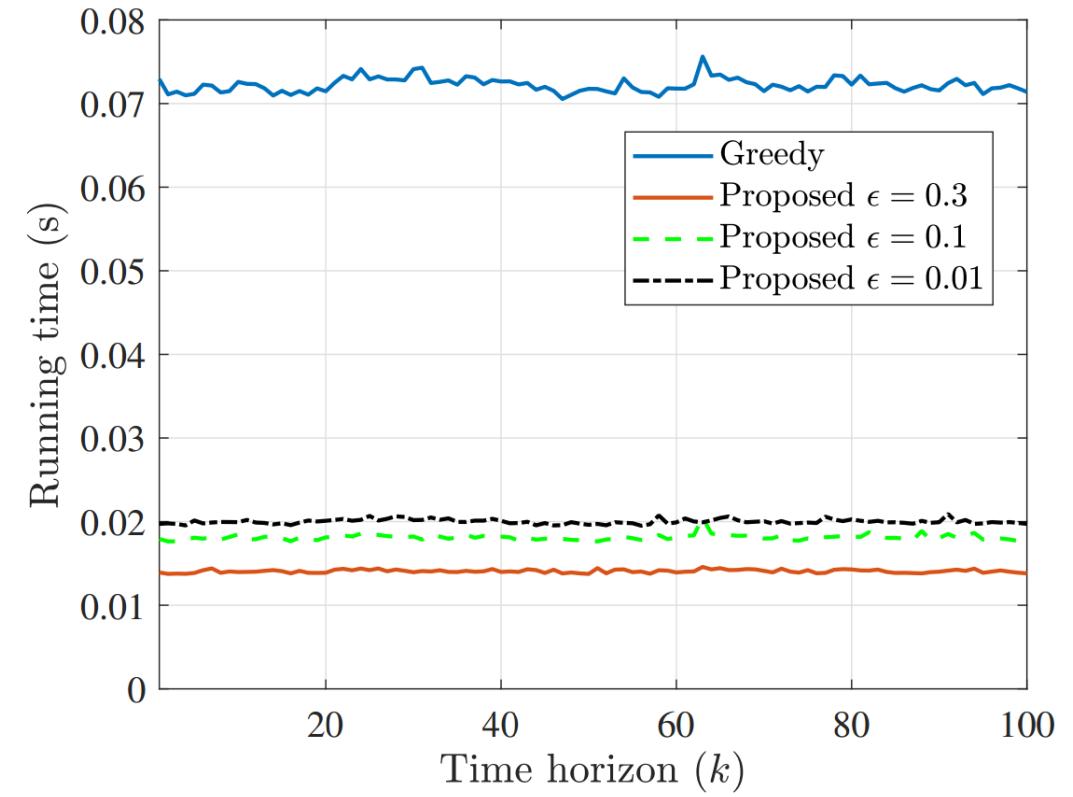
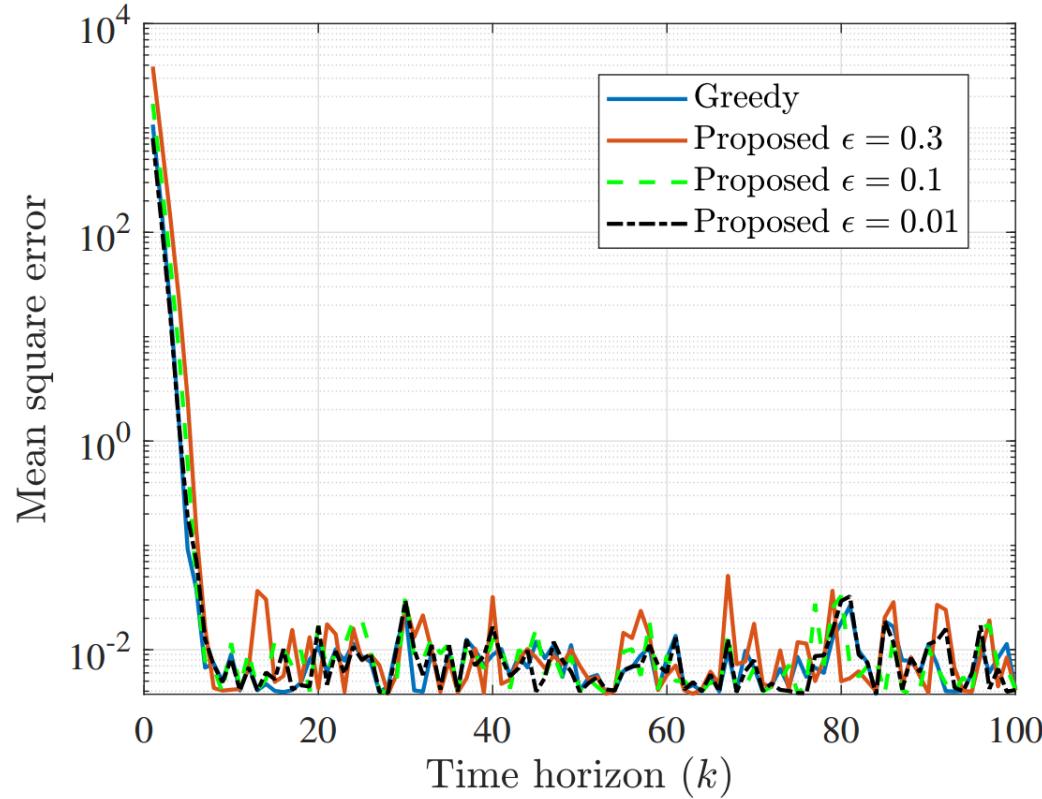
An increasing schedule is required to ensure the intersection is nonempty

## Theorem

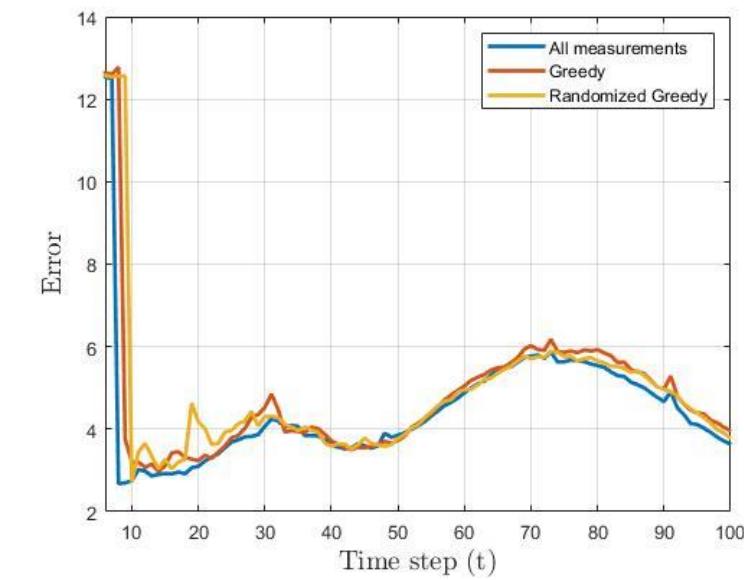
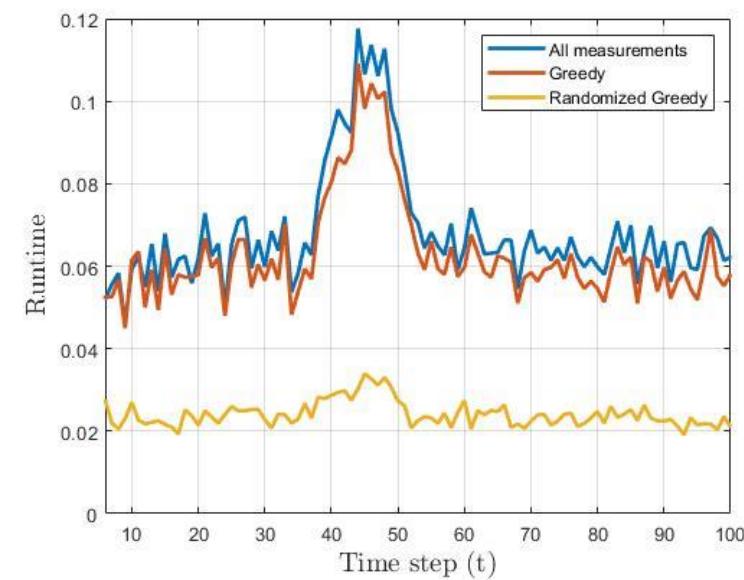
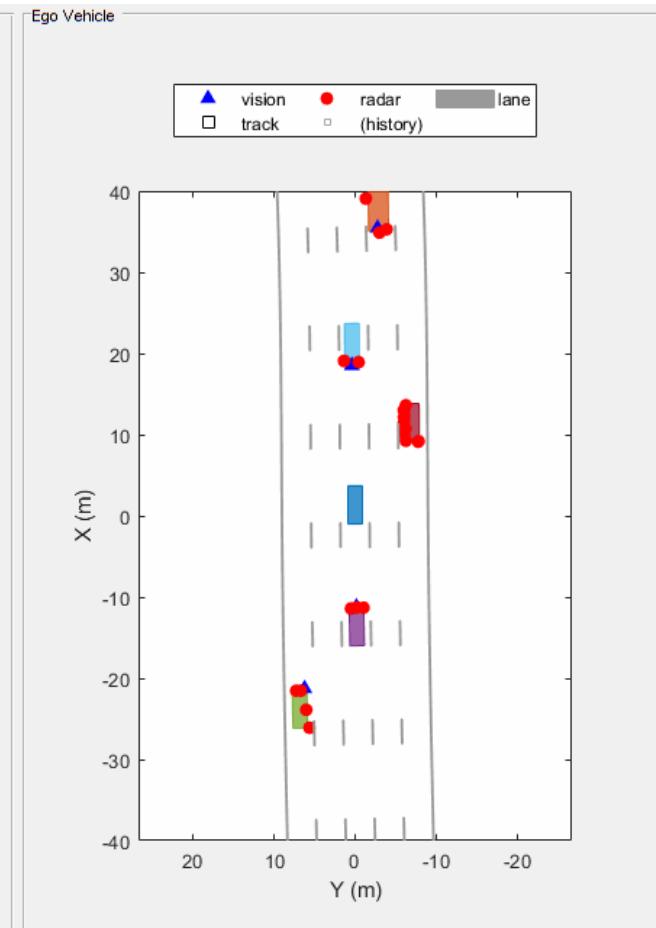
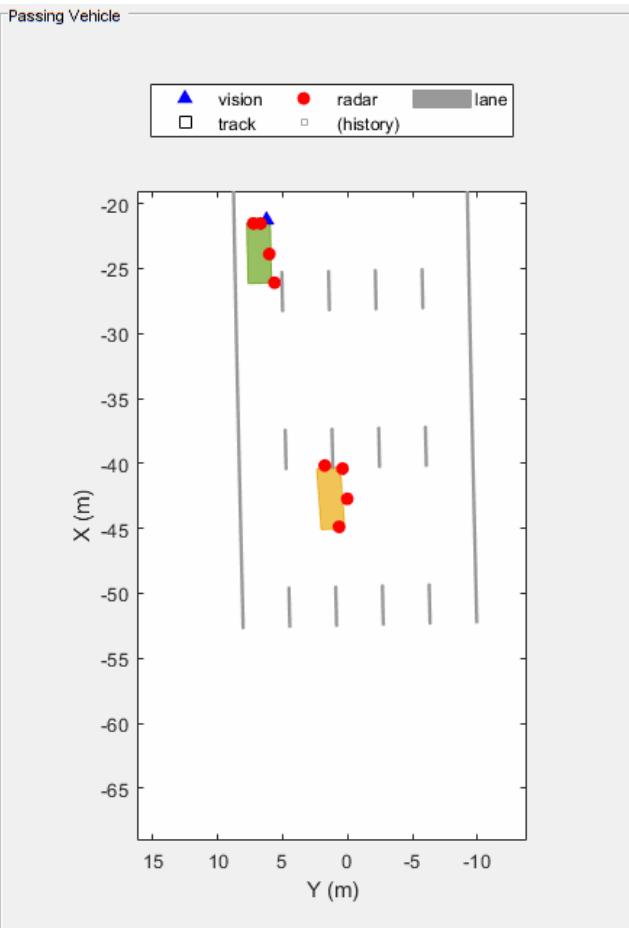
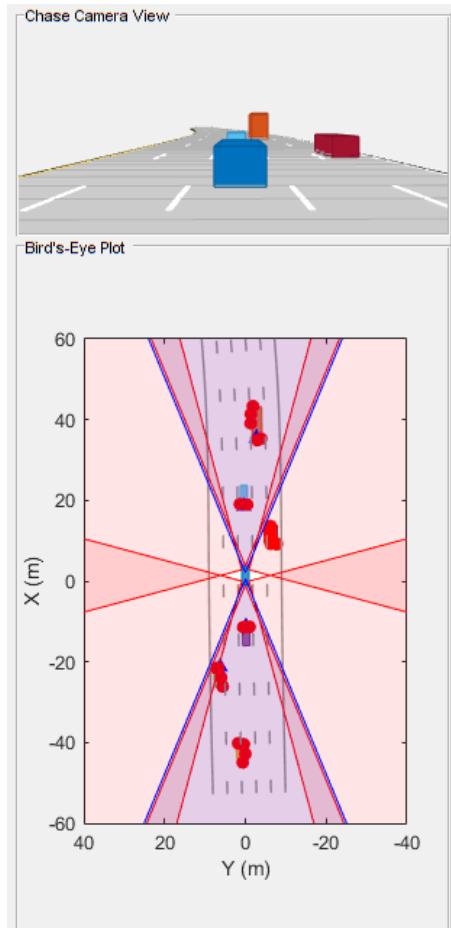
Near-optimal expected approximation guarantee

$$\mathbb{E}[f(\hat{S})] \geq (1 - e^{-\alpha} - \alpha\epsilon) f(S^*)$$

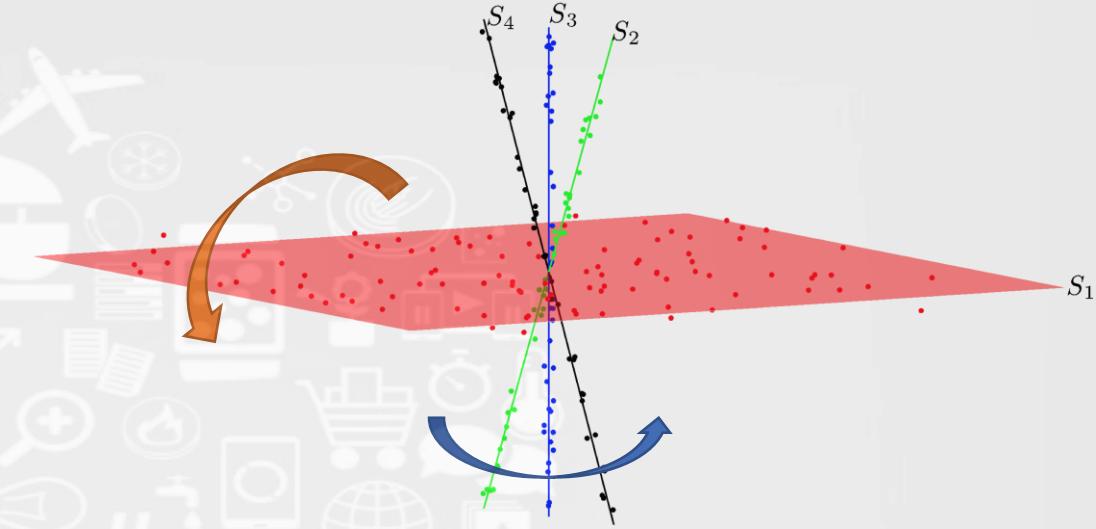
# UAV-Based Target Tracking



# Application in Autonomous Driving



# Data-Scarce Parsimonious Representation Learning



[Hashemi, A., Schaeffer, H., Shi, B., Tran, G., Ward, R., “Generalization Bounds for Sparse Random Features Expansions”, 2021.]

[Hashemi, A., Zhu, B., Vikalo, H., “Sparse Tensor Decomposition for Haplotype Assembly of Diploids and Polyploids,” BMC Genomics, Mar. 2018.]

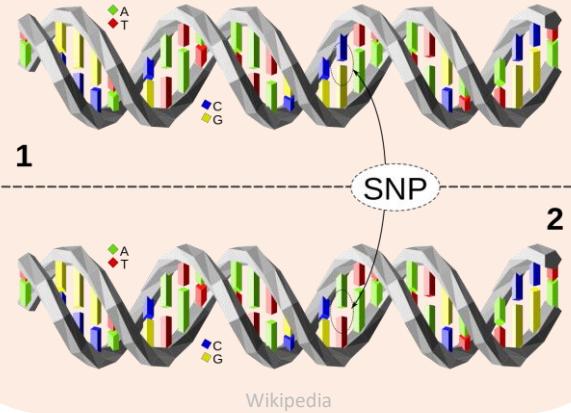
[Hashemi, A. and Vikalo, H., “Evolutionary Self-Expressive Models for Subspace Clustering,” IEEE Journal of Selected Topics in Signal Processing, Dec. 2018.]

[Hashemi, A. and Vikalo, H., “Accelerated Orthogonal Least-Squares for Large-Scale Sparse Reconstruction,” Digital Signal Processing, Nov. 2018.]

# Structured Function Approximation

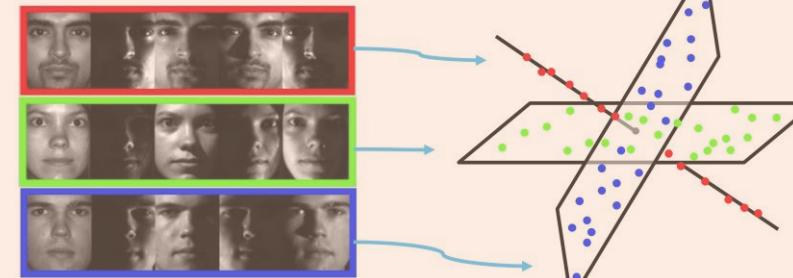
Low-rank structure

Genome sequencing



Sparsity

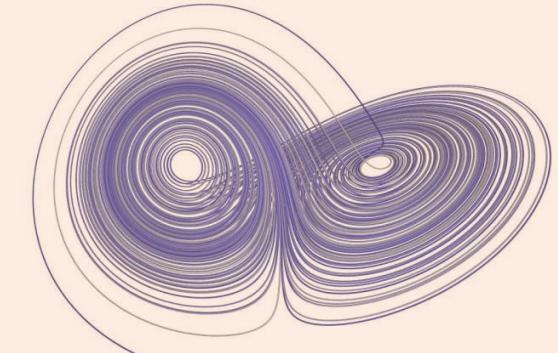
Clustering



You et al., 2018

Low-order structure

Dynamical systems



$$\dot{x}_i = x_{i+1}x_{i-1} - x_{i-1}x_{i-2} - x_i + 8, \quad i = 1, \dots, d$$

Wikipedia

## Goal

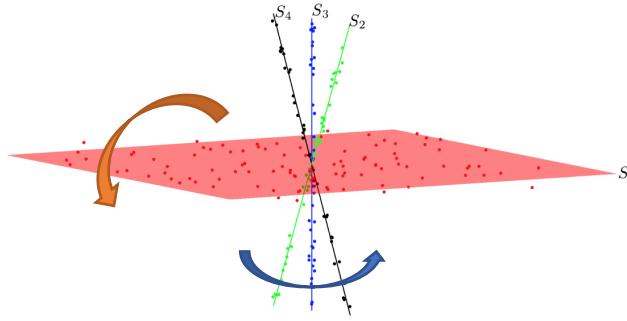
- Learning **structured unknown functions** from **limited measurements**

$$\min_{\theta} \sum_{i=1}^m \text{dist}(y_i, f(x_i; \theta)) \quad s.t. \quad f(x; \theta) \in \mathcal{F}$$

# Parsimonious Representation Learning



Data  $X_t$



Desired  
structures

Structure-aware  
representation  
learning

Multi-variate,  
non-convex problems

Alternating  
minimization

Subspace  
structure

Evolutionary  
structure

$$\begin{aligned} \min_{\mathbf{U}, \alpha} \quad & \|\mathbf{X}_t - \mathbf{X}_t(\alpha \mathbf{U} + (1 - \alpha) \mathbf{C}_{t-1})\|_F^2 \\ \text{s.t.} \quad & \text{diag}(\mathbf{U}) = \mathbf{0}, \quad \|\mathbf{U}\|_0 \leq k, \quad 0 \leq \alpha \leq 1 \end{aligned}$$

Subspace  
structure

## Theorem

Bound on amount of required data for a target accuracy

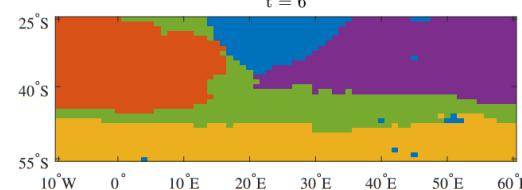
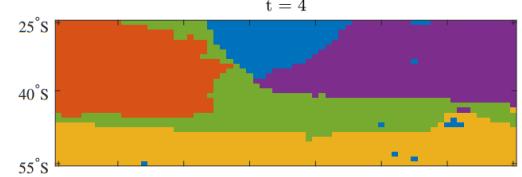
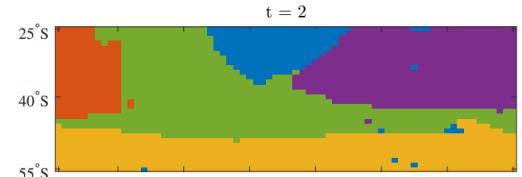
# Empirical Applications in Unsupervised Learning

## Real-time motion segmentation

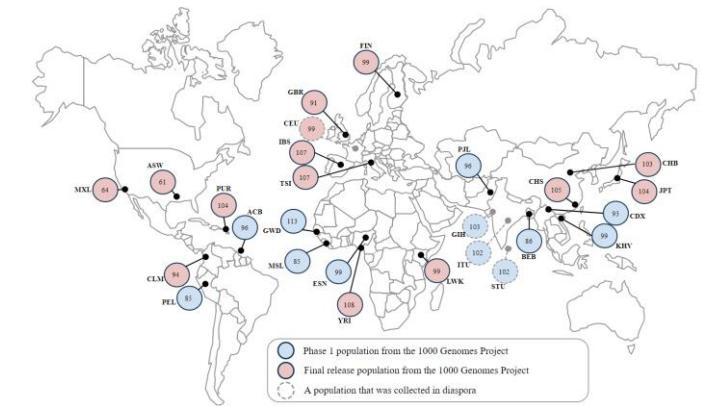


Method	Error (%)	Runtime (s)
Proposed	<b>5.60</b>	<b>1.69</b>
Baseline	10.76	46.16

## Tracking water masses near South Africa



## Study of genetic variation



Method	Error (%)	Runtime (s)
Proposed	<b>0.042</b>	<b>4.60</b>
Baseline	0.152	11.42

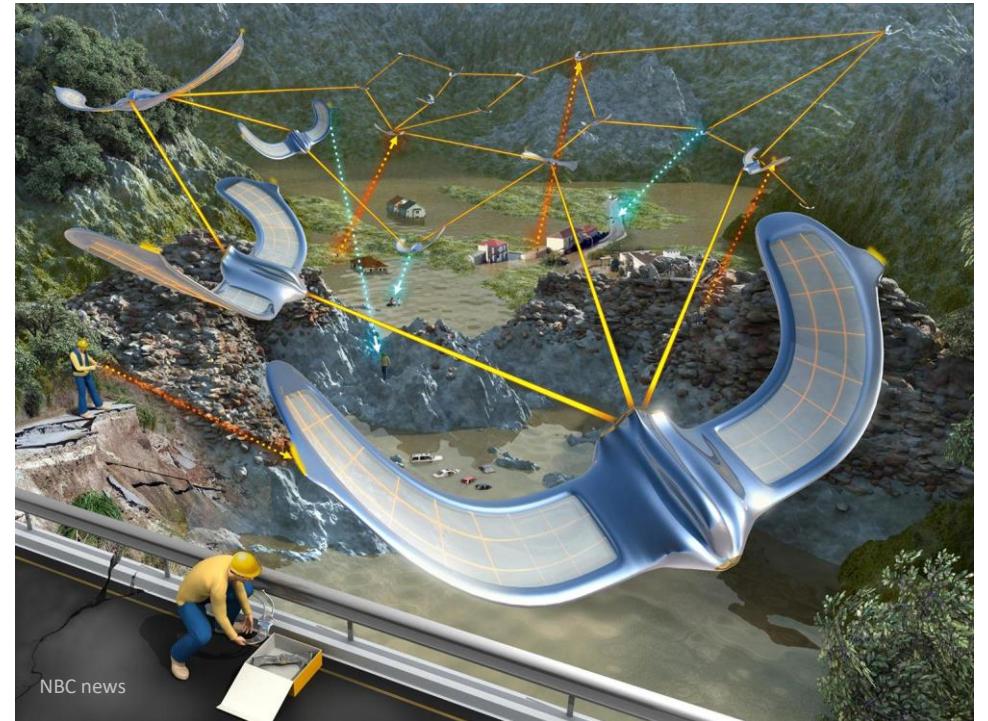


# Ongoing and Future Work

# Collaborative Learning in Dynamic Environments

Adaptive **representation learning** of dynamic data

Resource-constrained collaborative learning under **uncertainty and dynamic heterogeneity**



[Ghasemi, M., Hashemi, A., Vikalo, H., Topcu, U., "No-Regret Learning with High-Probability in Adversarial Markov Decision Processes," Conference on Uncertainty in Artificial Intelligence (UAI), 2021]

[Ghasemi, M., Hashemi, A., Topcu, U., Vikalo, H., "Online Learning with Implicit Exploration in Episodic Markov Decision Processes," American Control Conference (ACC), 2021]

# Robustness and Security

Collaboration against **unexpected contingencies and adversaries**

Integrating **robust hypothesis testing** into information acquisition and representation learning

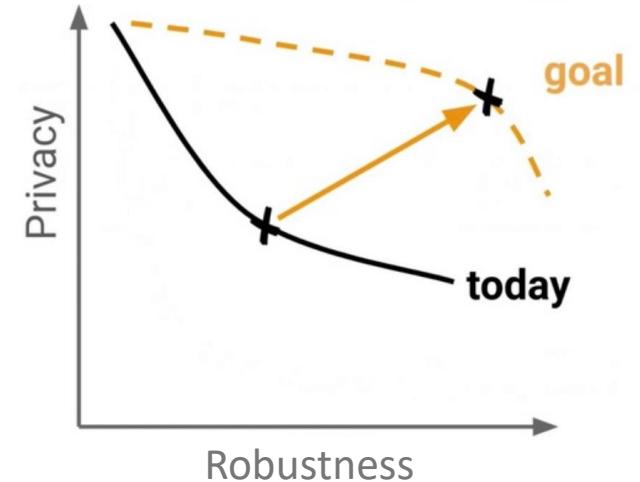
Exploring the **trade-off between privacy and robustness**

[Acharya, A., Hashemi, A., Jain, P., Sanghavi, S., Dhillon, I., Topcu, U., “Robust Training in High Dimensions via Block Coordinate Geometric Median Descent,” Preprint, 2021]

[Das, R., Hashemi, A., Sanghavi, S., Dhillon, I., “DP-NormFedAvg: Normalizing Client Updates for Privacy-Preserving Federated Learning,” Preprint, 2021]

Update: Chrysler recalls 1.4M vehicles after Jeep hack

COMPUTERWORLD



# Structured and Resource-Constrained Collaborative Learning

Abolfazl Hashemi

