Nirmit Desai

Principal Research Scientist and Manager, IBM Research
Mission: Develop practical solutions to hard problems and test them in the real-world

CS PhD, NCSU 2007

- Interaction-oriented design of complex software
- Representing and reasoning about socio-technical abstractons

Edge AI: Enabling data and AI applications across distributed nodes

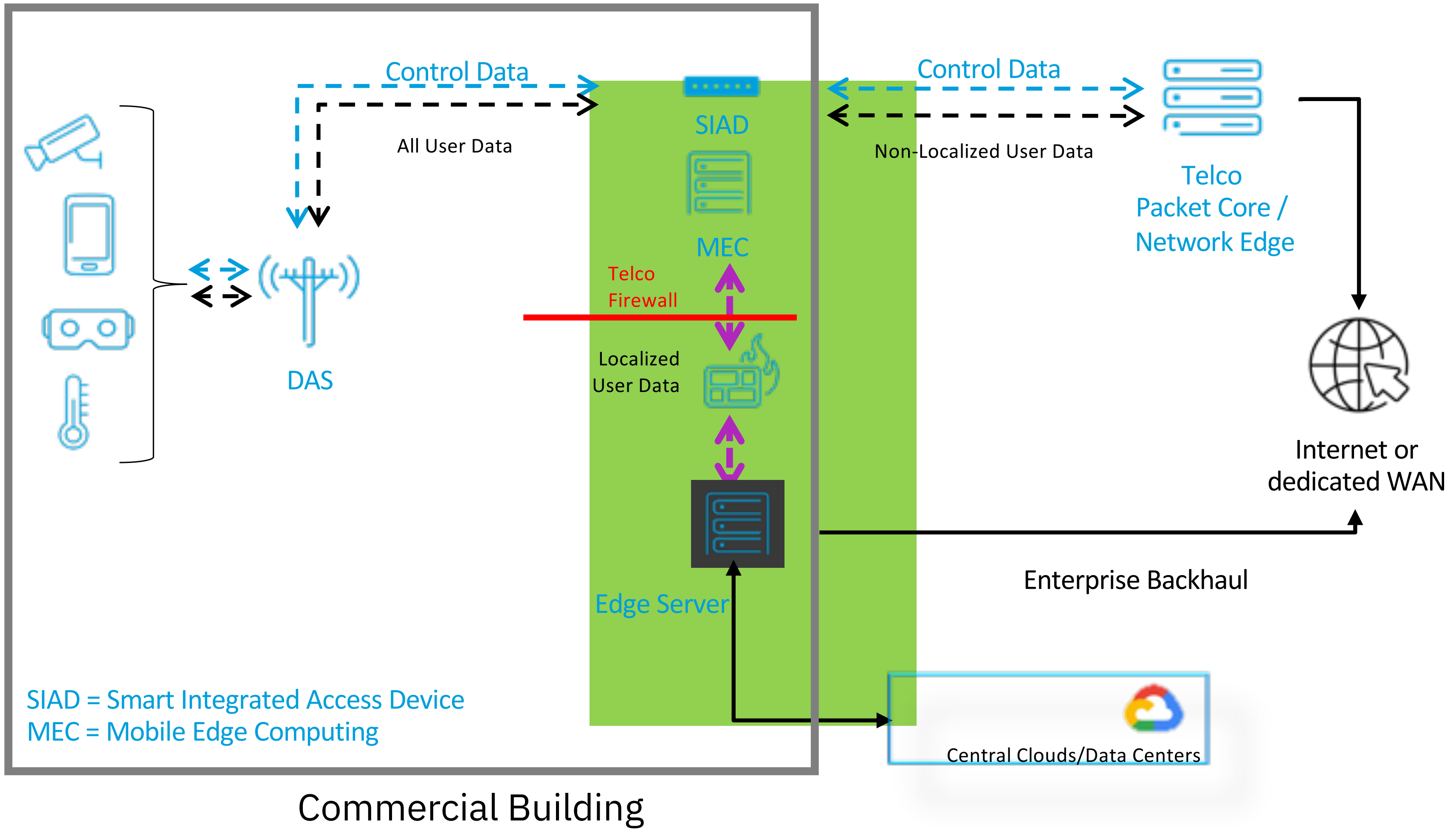
- Federated learning, edge model mgmt, edge data mgmt
- E.g. Watson flu-risk predictions

Edge infrastructure: large-scale DTN over mobile platforms

- E.g., “Mesh Network Alerts” for offline messaging reaching millions of users via Weather Channel apps
- E.g., “Watson Works” contact tracing with accurate proximity estimation via ultrasound

5G, Edge, AI: 5G is a catalyst for Edge Computing, AI is the most common Edge workload

Typical private 5G deployment

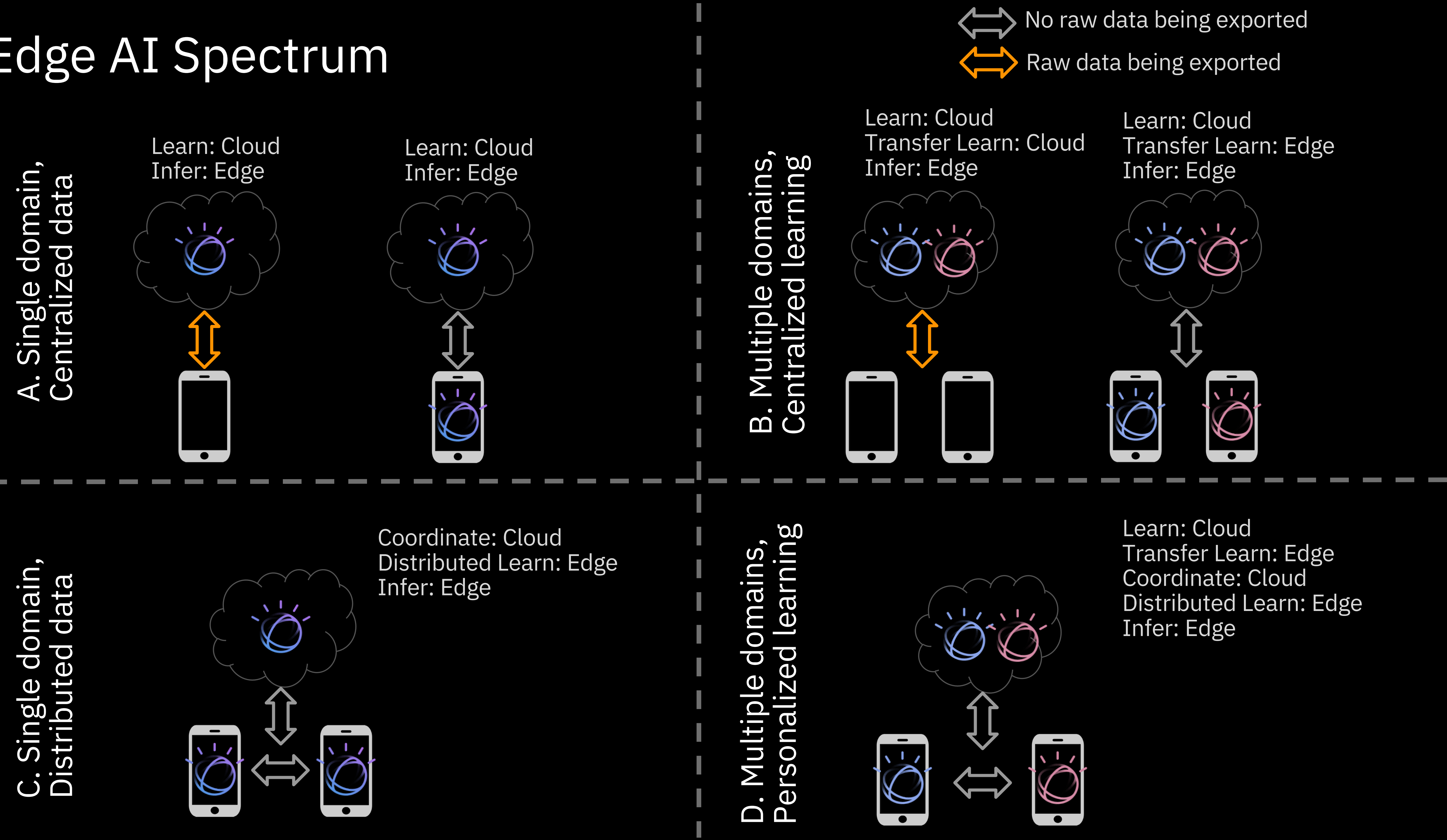


Selected Traffic broken out to nearby servers (~5 ms latency)

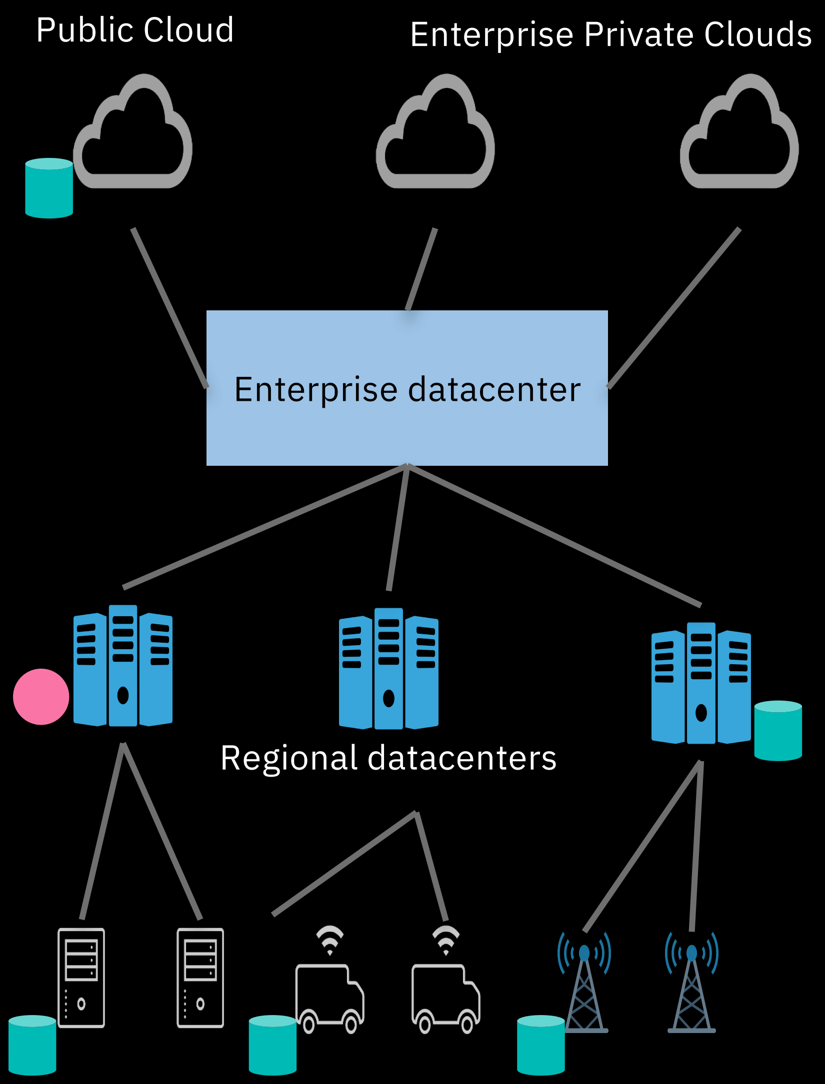
Applications deployed to edge servers

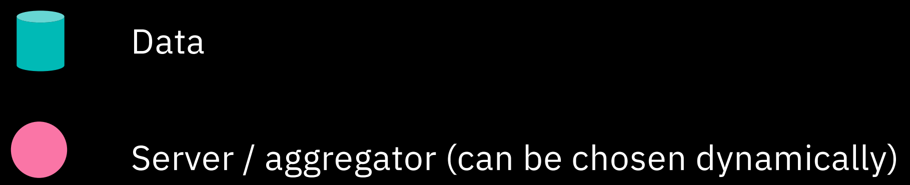
Enterprise / Telco partnerships, new business models

Edge AI Spectrum



Distributed Learning: Generalized Problem





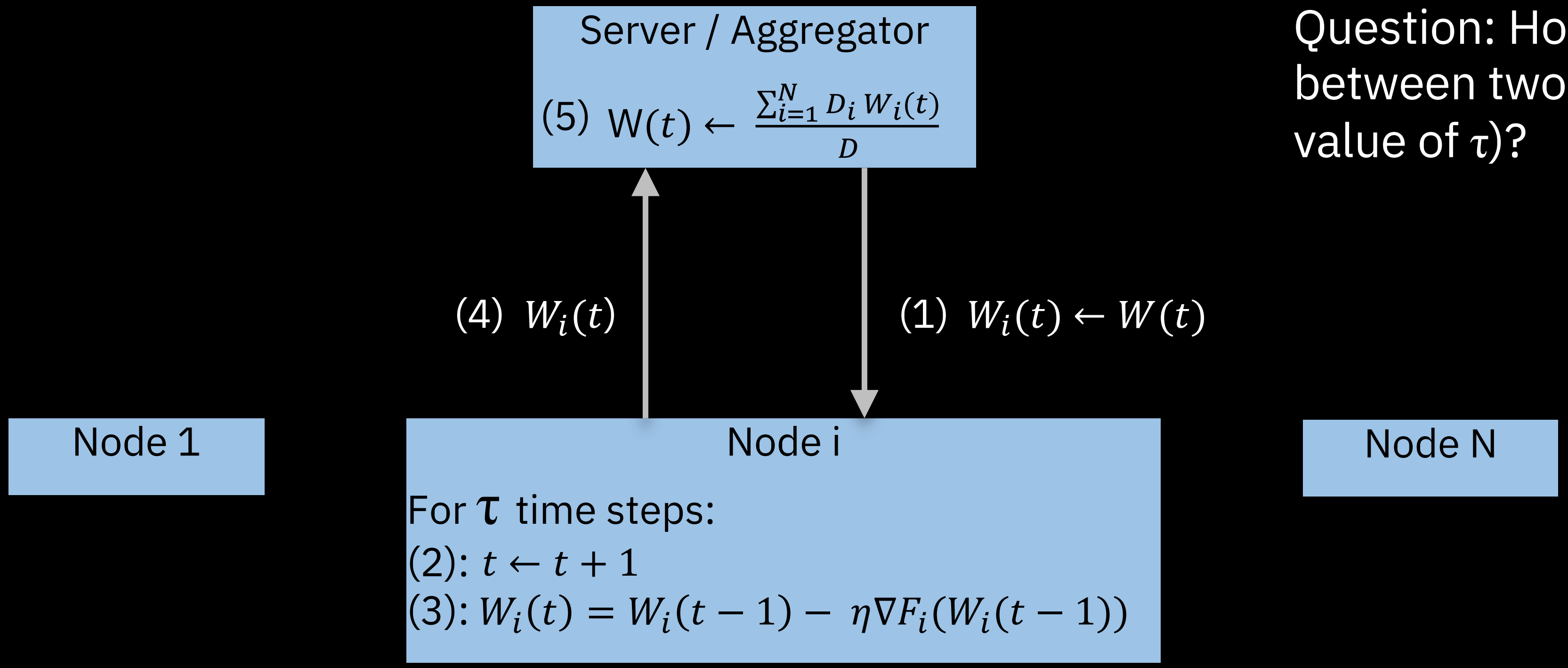
How to make the best use of limited resources at the edge?

- Edge, resources are limited (communication, computation, storage, energy)
- Always more limited than servers in data centers

How to select a suitable subset of data to be involved in the distributed learning process?

- Data collected by individual edge devices may or may not be related to the learning task

Resource budget: Trading-off local updates with optimality

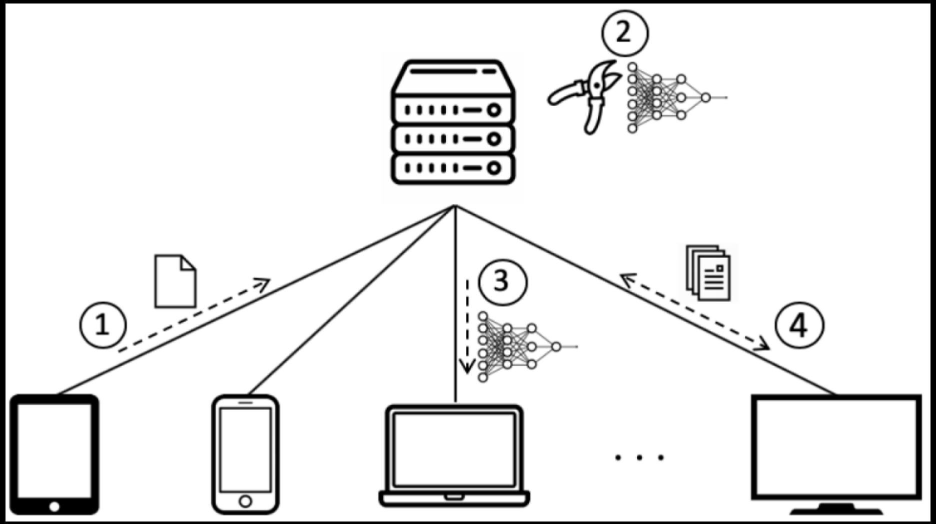


Question: How many steps of local updates between two global aggregation steps (i.e., value of τ)?

Key findings: under some assumptions about the loss function

- $\tau = 1$ provably achieves same loss as the centralized case
- $\tau > 1$ results in inferior loss but converges
- For a given time budget, optimal τ can be determined via an online algorithm
- If the time budget is increased, optimal τ found by the algorithm decreases

Federated model pruning: Transmit partial model parameters

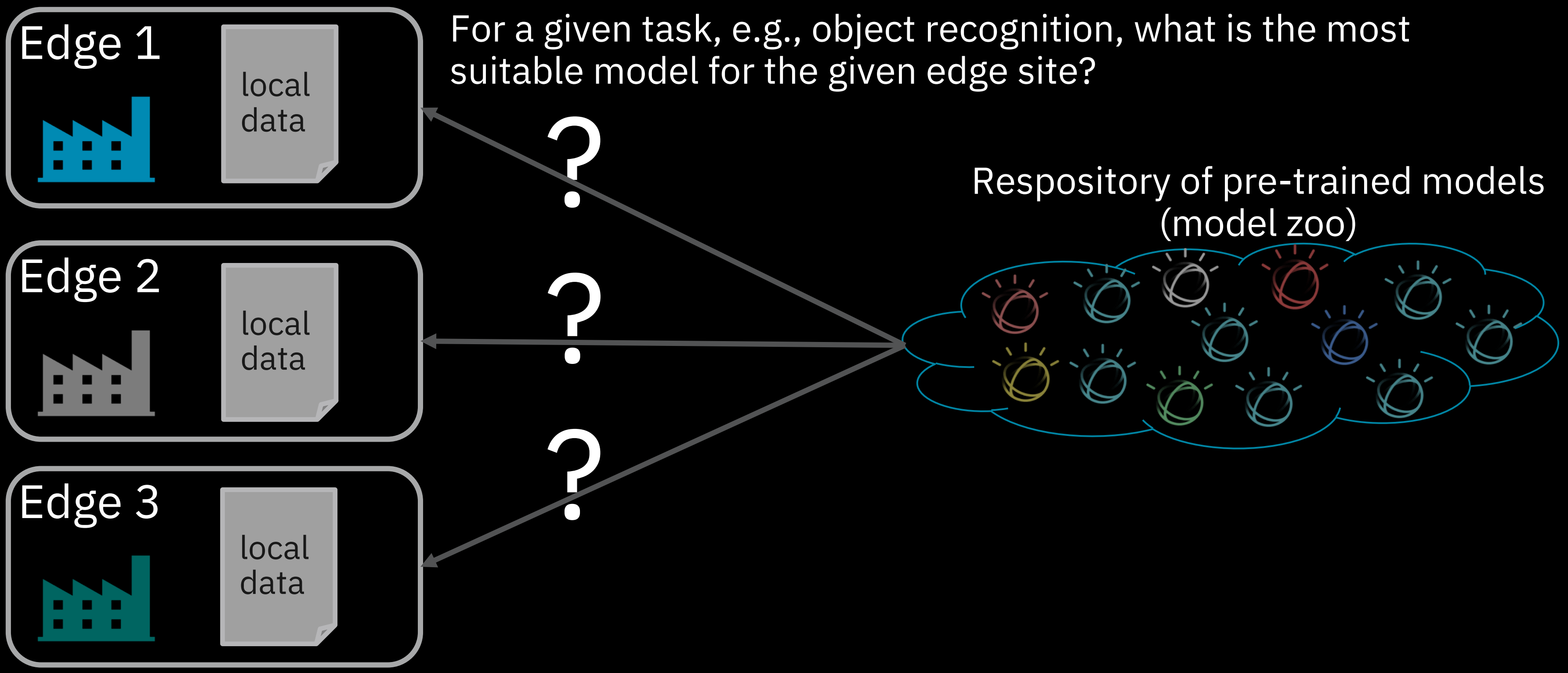


Basic model pruning:

- Perform one iteration of gradient descent
- Remove a pre-defined amount of model parameters
- Repeat the above until the desired model size is reached

1. Each client sends a small amount of sample data to the server (optional)
2. Server performs initial model pruning using sample data (if available)
3. Server sends initially pruned model to clients (remove the smallest weights)
4. Clients and server interact to perform federated learning
   – Further pruning can be performed until the model is at the desired (small) size

Selecting the most suitable model for an edge environment



Challenges in model selection:

- Model zoo model class labels may not match those at the edge
- Brute-force accuracy measurement may not measure model fitness
- Key insight: discriminatory power of a model, regardless of its domain, is the best indicator of its suitability for an edge

Thank you

Nirmit Desai
Principal Research Scientist and Manager
—
nirmit.desai@us.ibm.com

Relevant publications:

- N. Desai, L. Chu, R. K. Ganti, S. Stein, M. Srivatsa, “neuralRank: Searching and ranking ANN-based model repositories”, arXiv:1903.00711 (2019)
- W. Lee, S. Millman, N. Desai, M. Srivatsa, C. Liu, “NeuralFP: Out-of-distribution detection using fingerprints of Neural Networks”, International Conference on Pattern Recognition (ICPR), 2020
- S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems”, IEEE Journal on Selected Areas in Communications, vol. 37, no. 6, pp. 1205 – 1221, Jun. 2019 (earlier version at IEEE INFOCOM 2018).
- P. Han, S. Wang, K. K. Leung, "Adaptive gradient sparsification for efficient federated learning: an online learning approach”, IEEE ICDCS, 2020.
- Y. Jiang, S. Wang. B. J. Ko, W.-H. Lee, L. Tassiulas, “Model pruning enables efficient federated learning on edge devices”, https://arxiv.org/abs/1909.12326, 2019.
- T. Tuor, S. Wang, B. J. Ko, C. Liu, K. K. Leung, "Data selection for federated learning with relevant and irrelevant data at clients”, https://arxiv.org/abs/2001.08300, 2020.