Aalto University
School of Science

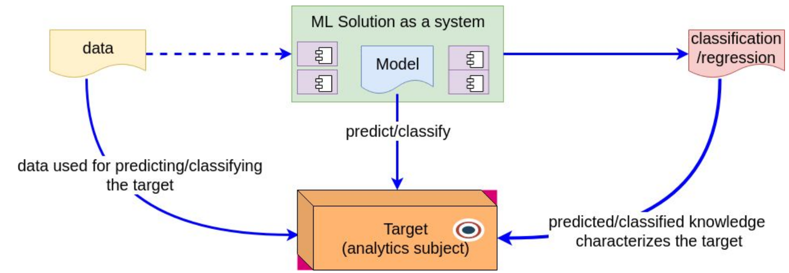
CS-E4640 Advanced Topics in
Software Systems

Introduction to Federated Learning

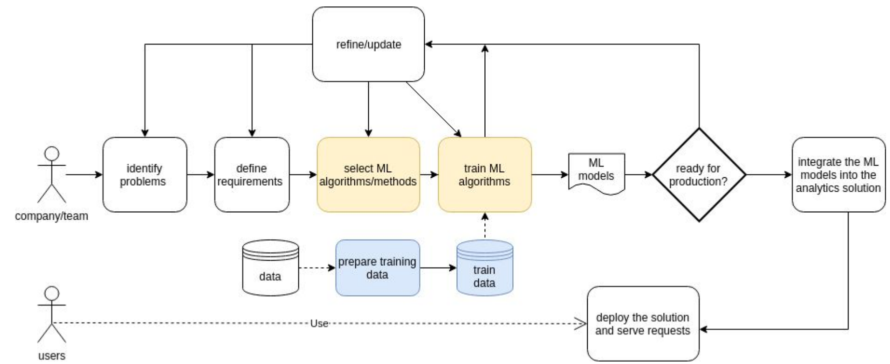
Hong-Linh Truong
Department of Computer Science
linh.truong@aalto.fi, *https://rdsea.github.io*

Recall from Machine Learning study

- **AI/ML models are built for predicting/classifying targets**
  - AI/ML models are just one part of the solution/system



- **ML systems development and operations**

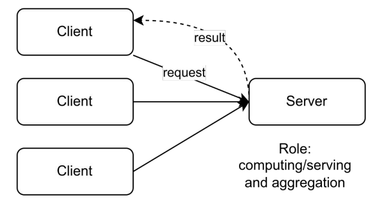


Recall from distributed computing and systems study

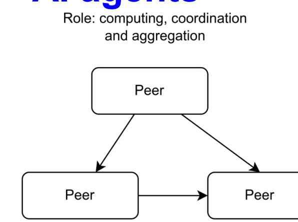
- **Distributed compute and data resources used for common, shared goals**
  - single or multiple administrative domains
  - different infrastructures: edge, cloud, multi-cloud, & edge-cloud continuum
  - diverse security and privacy settings

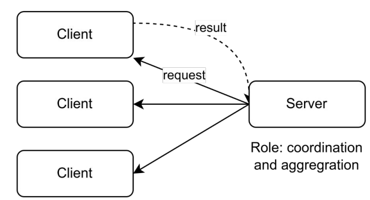
Distributed computing models

client server, prompt client - LLMs

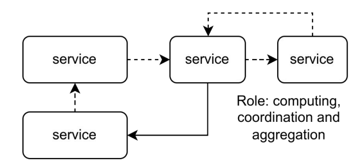


peer-to-peer, AI agents





offloading/workflows, ML Ensembles



reactive systems/services, streaming ML

Centralized data for ML is not enough

- **The capability of ML models is based on many factors of "training data"**
  - large, quality, diverse, representative ⇒ hard to have even for a big company!
- **The potential and benefits of data with different providers/ownerships**
  - very big data with a full coverage for learning ⇒ we actually *do not have and/or do not know* if the data is enough
  - current and future *realtime edge and on-premise big data* scenarios
- **Reasons not having enough data for centralized learning**
  - business conditions, data regulations, and incentives
  - suitable secure computing techniques, scale and communications

Empowering different data providers for learning at distributed scale

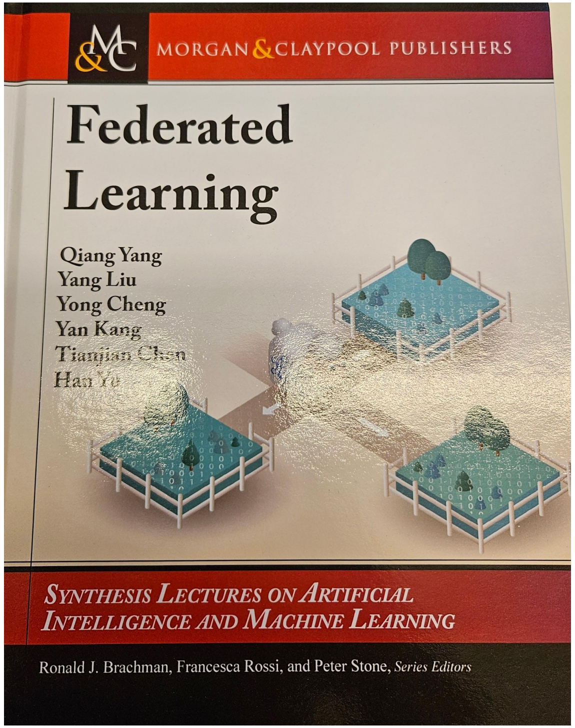
Aalto University
School of Science

Solution: “Federated Learning”!

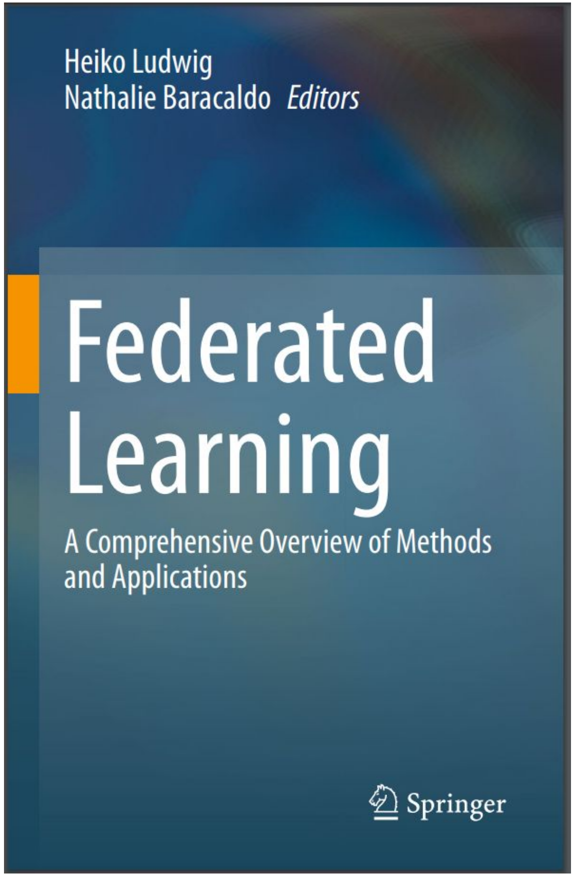
Course learning objectives: after this introduction, be able to

- **Explain the definition and motivation of Federated Learning (FL)**
- **Explain basic categories, components and interactions in FL**
- **Evaluate potential applications in FL**
- **Explain key relevant topics for designing and implementing FL**

Reading List



Chapters 1 & 2 (Introduction & Background)



Chapter 1 (Introduction to FL)

A definition

“Federated learning *is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client’s raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective.*”

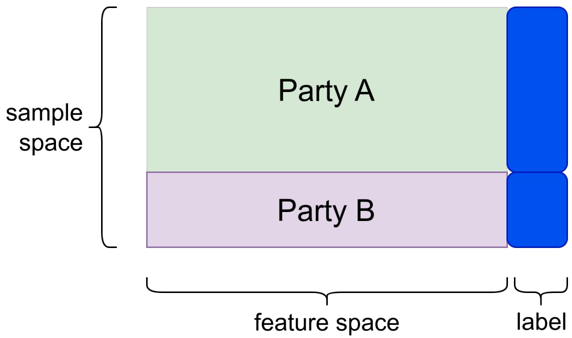
Source: Kairouz, P., et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning 14*, 1–2 (2021); https://arxiv.org/abs/1912.04977

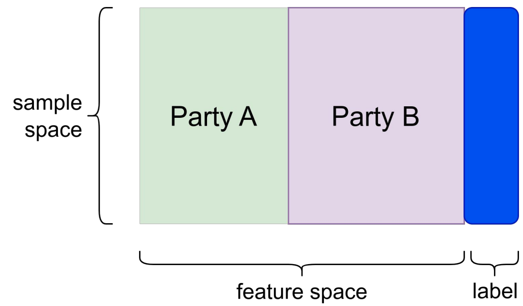
Fundamental aspects for understanding and designing FL

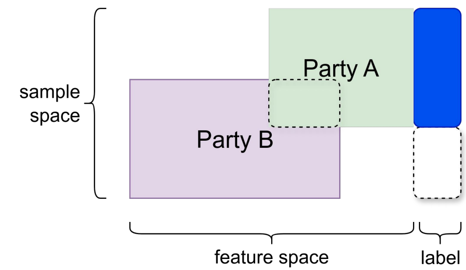
- **The data sources, from multiple parties, for learning**
  - decentralized data, diverse distribution, quality and quantity
- **The distributed compute resources, aligned with the data sources, for learning**
  - distributed, heterogeneous computing & connectivity resources
- **The consensus/agreement among data (and compute) parties**
  - trust, quality of data, privacy-preserving protocols, meta data agreement
- **The coordination/collaborative techniques**

Learning from data: data characteristics drives the type of learning

Assume that A and B can contribute **data for training an ML model**.
What are common samples and label/features of data between A & B ?







Basic categories of FL

Horizontal Federated Learning (HFL)

Vertical Federated Learning (VFL)

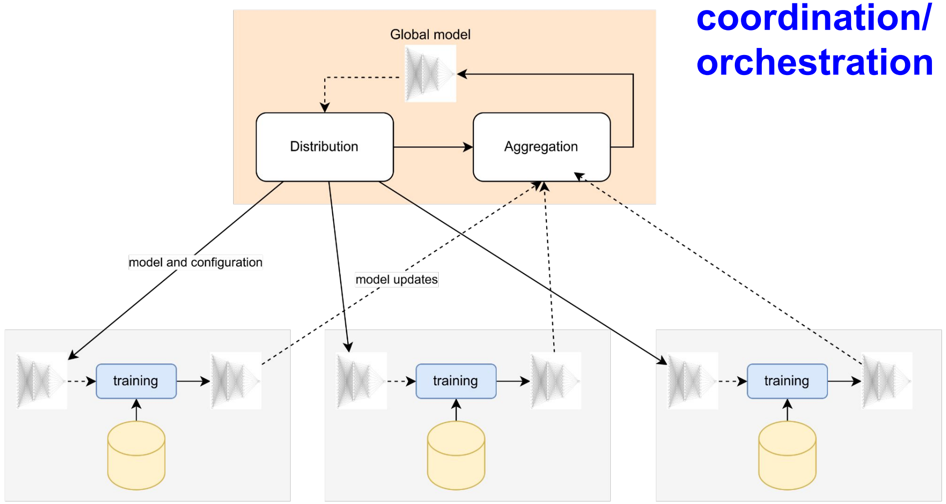
Federated Transfer Learning (FTL)

Figure source (with redrawn/modified): Y. Liu *et al.*, "Vertical Federated Learning: Concepts, Advances, and Challenges," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3615-3634, July 2024, doi: 10.1109/TKDE.2024.3352628.

Computation for FL: the basic model

".. multiple entities (clients) collaborate in *solving a machine learning problem, under the coordination of a central server* or *service provider …"*

Central service/ service provider: coordination/ orchestration



entities (clients): participants

Participants: (i) cross-silo use cases (few) vs cross-device use cases (huge), (ii) heterogeneity in terms of data, computing capabilities, networks, reliability, management, etc.

Coordination in FL: complex computing tasks

Scenario:
cross-device

Participant:
mobile devices

Global model update:
enough with a subset of devices

Figure source: *Keith Bonawitz et al.*, Towards Federated Learning at Scale: System Design, MLSys 2019, https://arxiv.org/pdf/1902.01046

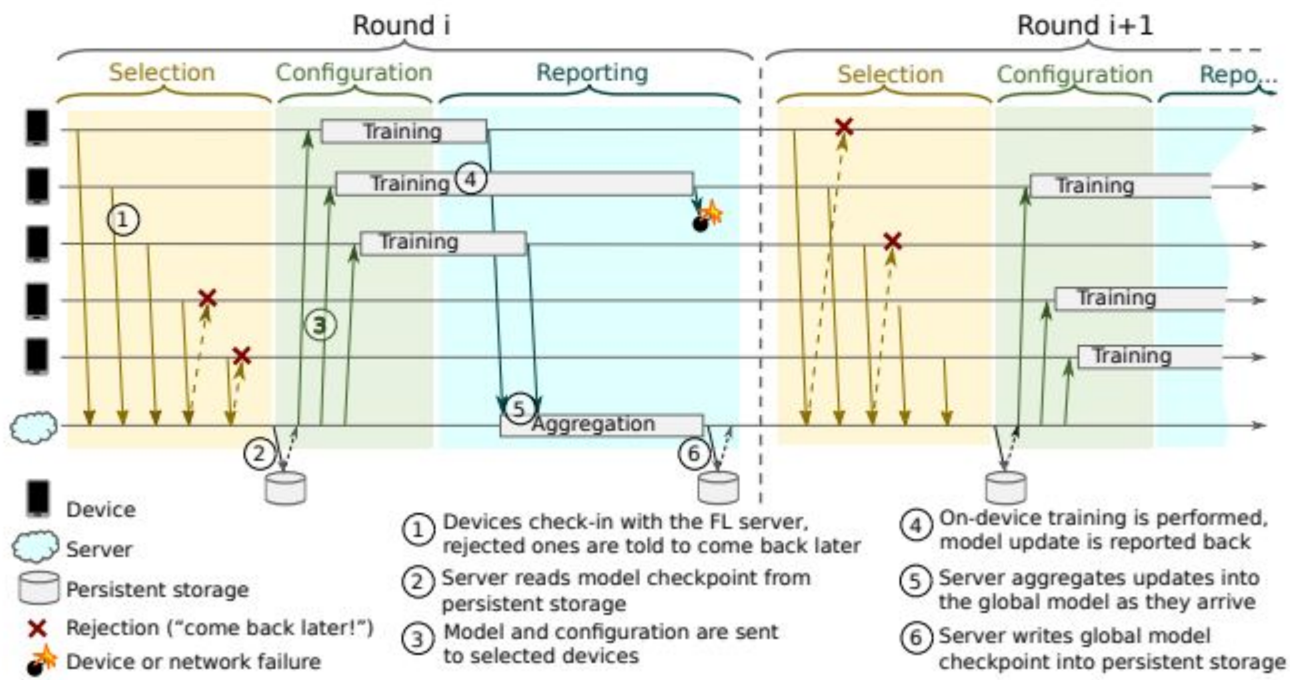


Figure 1: Federated Learning Protocol

Coordination in FL: model aggregation

Example: the famous FevAvg algorithm

- **No training data is shared**
- **Update the global model after each FL training iteration/round**
  - Receive updated weights/gradients or logits from participants
  - Perform aggregation

- **Many different aggregation algorithms**
  - FevAvg, FevAdam, Secure Aggregation
  - asynchronous and synchronous aggregation

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
    $m_t \leftarrow \sum_{k \in S_t} n_k$
    $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$ *// Erratum*[4]

ClientUpdate(k, w): *// Run on client k*
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return w to server

Source: Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. https://arxiv.org/abs/1602.05629

Secure computation & privacy-preserving

- **Secure connections and communications**
  - common in distributed computing
- **Privacy-preserving learning**
  - Secure multi-party computation, homomorphic encryption, differential privacy
- **Adversarial machine learning**
  - data poisoning, evasion, model extraction, Byzantine attacks

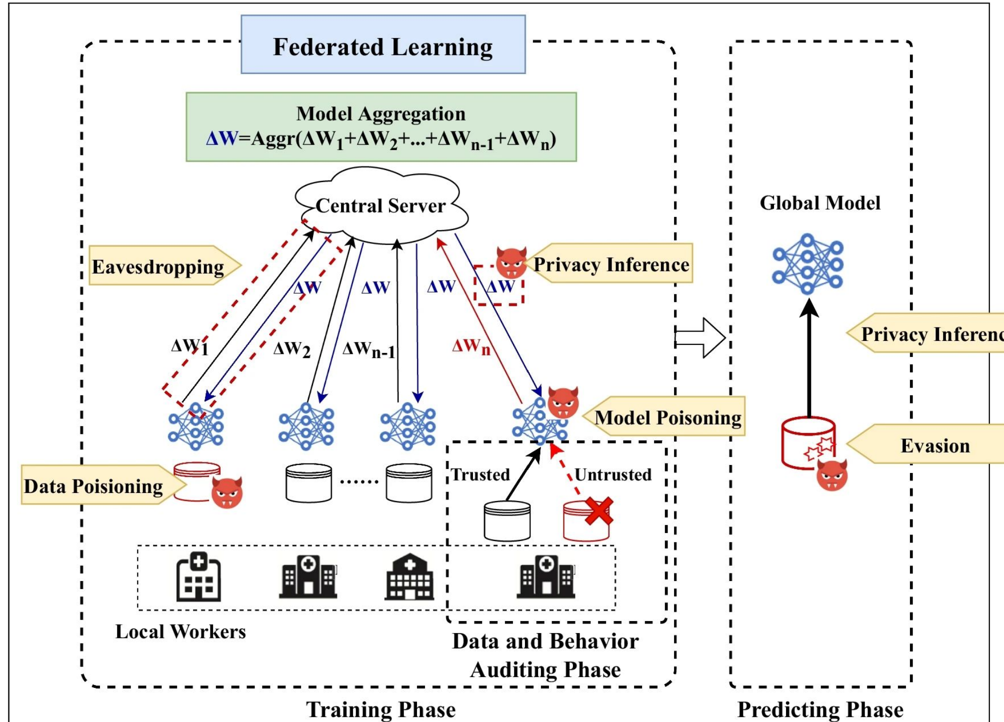


Figure source: Liu, P., Xu, X. & Wang, W. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. *Cybersecurity* 5, 4 (2022). https://doi.org/10.1186/s42400-021-00105-6

Potential domains/problems solved by FL

- **Potential scenarios and applications**
  - Finance
  - Healthcare
  - IoT/Industry 4.0/Manufacturing → predictive maintenance
  - Cybersecurity (malware detection)
  - Autonomous vehicles/robots
- **Carefully evaluate if FL is the right solution**
  - cross-silo (a few of big parties) vs cross-device (a huge number of parties)
  - the collaboration/federated agreements w.r.t.
    - *data and computation for learning*
    - *security/privacy requirements*

Hands-on/practical programming for FL

- **Start with existing known frameworks**
  - our hands-on with Flower
- **Utilize edge-cloud computing systems**

Flower https://flower.ai/

FATE https://fate.fedai.org/overview/

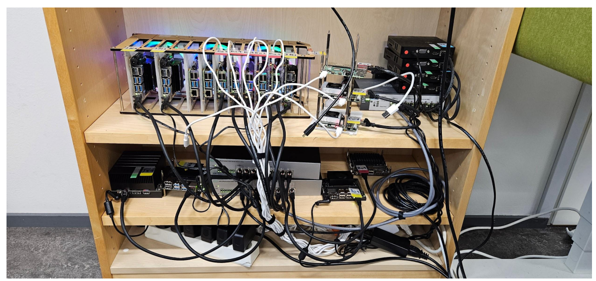
OPENFL https://openfl.io/

NVIDIA FLARE https://nvflare.readthedocs.io/en/main/index.html

Syft https://github.com/OpenMined/PySyft

Check: Riedel, P., Schick, L., von Schwerin, R. *et al.* Comparative analysis of open-source federated learning frameworks - a literature-based survey and review. *Int. J. Mach. Learn. & Cyber.* (2024). https://doi.org/10.1007/s13042-024-02234-z

Edge nodes: Raspberry PI, Jetson Orion/Nano/Xavier, Beelink, RockPi, Coral, NPU accelerator, Hailo-8L AI accelerator, etc.



Cloud resources: Google Cloud, CSC

Aalto University
School of Science

Hello FL with Flower (VSCode)

```
(fm) truong@aaltosea22:~/myprojects/mygit/sys4bigml/tutorials/basicfl$ flwr new

💬 Please provide the project name: hellofluibk

💬 Please provide your Flower username: Linh Truong

💬 Please select ML framework by typing in the number

 [ 0] FlowerTune
 [ 1] HuggingFace
 [ 2] JAX
 [ 3] MLX
 [ 4] NumPy
 [ 5] PyTorch
 [ 6] TensorFlow
 [ 7] sklearn


: 7

🔨 Creating Flower project hellofluibk...
🎊 Project creation successful.

Use the following command to run your project:

        cd hellofluibk
        pip install -e .
        flwr run
```

Our advanced topics for FL

- **Data quality and data governance**
- **System challenges**
  - performance, reliability and elasticity of computation, coordination/orchestration, and communications
- **Trustworthy learning**
  - secure communication, privacy, confidential, data, multi-party computation
- **Optimization based on various tradeoffs**
  - privacy-accuracy, accuracy-cost, cost-performance
- **Marketplaces/incentives**
- **Applications requirements**

Advances and Open Problems in Federated Learning

Peter Kairouz[7*] H. Brendan McMahan[7*] Brendan Avent[21] Aurélien Bellet[9]
Mehdi Bennis[19] Arjun Nitin Bhagoji[13] Kallista Bonawitz[7] Zachary Charles[7]
Graham Cormode[23] Rachel Cummings[6] Rafael G.L. D'Oliveira[14]
Hubert Eichner[7] Salim El Rouayheb[14] David Evans[22] Josh Gardner[24]
Zachary Garrett[7] Adrià Gascón[7] Badih Ghazi[7] Phillip B. Gibbons[2]
Marco Gruteser[7,14] Zaid Harchaoui[24] Chaoyang He[21] Lie He[4]
Zhouyuan Huo[20] Ben Hutchinson[7] Justin Hsu[25] Martin Jaggi[4] Tara Javidi[17]
Gauri Joshi[2] Mikhail Khodak[2] Jakub Konečný[7] Aleksandra Korolova[21]
Farinaz Koushanfar[17] Sanmi Koyejo[7,18] Tancrède Lepoint[7] Yang Liu[12]
Prateek Mittal[13] Mehryar Mohri[7] Richard Nock[1] Ayfer Özgür[15]
Rasmus Pagh[7,10] Hang Qi[7] Daniel Ramage[7] Ramesh Raskar[11]
Mariana Raykova[7] Dawn Song[16] Weikang Song[7] Sebastian U. Stich[4]
Ziteng Sun[3] Ananda Theertha Suresh[7] Florian Tramèr[15] Praneeth Vepakomma[11]
Jianyu Wang[2] Li Xiong[5] Zheng Xu[7] Qiang Yang[8] Felix X. Yu[7] Han Yu[12]
Sen Zhao[7]

Paper: https://arxiv.org/abs/1912.04977

Homework

Given a potential scenario for FL in your choice, try to identify possible privacy issues for initial suitability analysis

Mark your answers for the question marks!

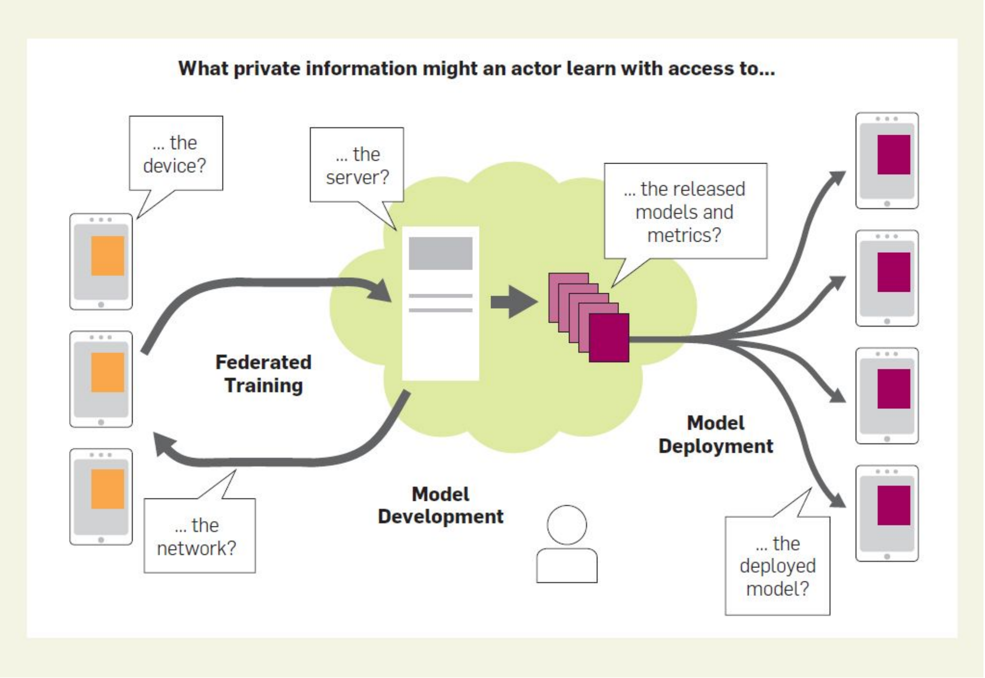


Figure source: Kallista Bonawitz, Peter Kairouz, Brendan Mcmahan, and Daniel Ramage. 2022. Federated learning and privacy. Commun. ACM 65, 4 (April 2022), 90–97. https://doi.org/10.1145/3500240

Thanks!

Hong-Linh Truong
Department of Computer Science

rdsea.github.io