AI4EOSC Webinar: Introduction to Federated Learning

Basics of Federated Learning: Tips and Tricks

Khadijeh Alibabaei (khadijeh.alibabaei@kit.edu)

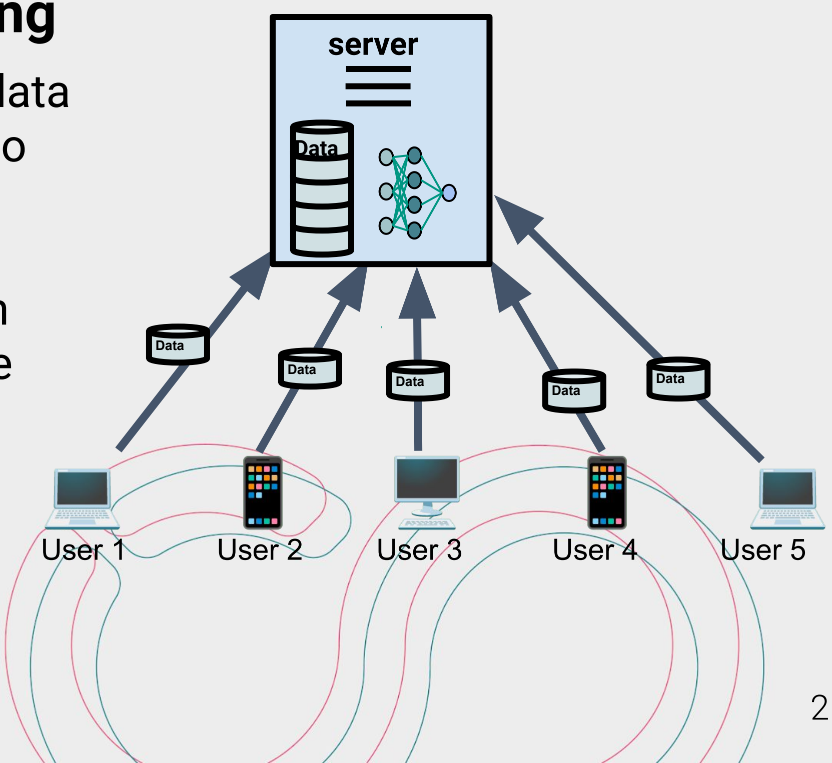
Funded by
the European Union

22 | 04 | 2024 by K. Alibabaei

Centralized Learning in Machine Learning

- refers to the traditional approach where all data is gathered and stored in a central location to train a machine learning model.

- involves collecting and combining data from multiple sources into a single dataset before training the model.



Funded by
the European Union

Centralized Learning in Machine Learning: Challenges

- **Data Flow Management**: Manage the transfer of **large volumes** of diverse data quickly and accurately across different organizations.
- **Scalability**
- **Communication Overhead**
- **Intense competition within the industry.**
- **Data Privacy**: Ensuring compliance with strict data protection regulations, such as the GDPR[1] and EU AI ACT[2].

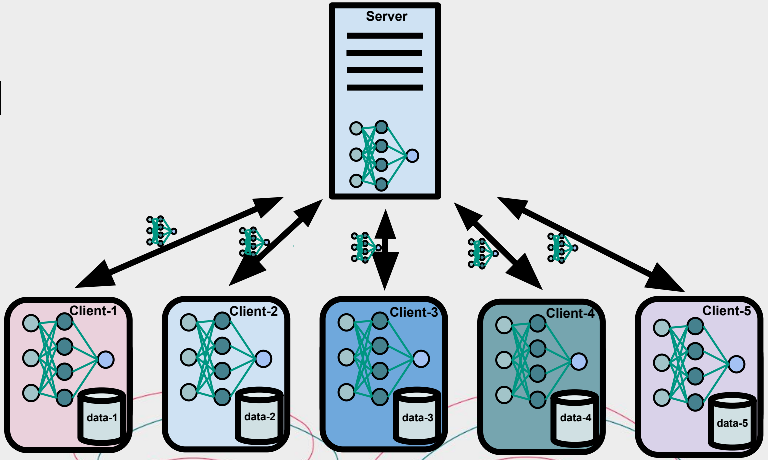
1. https://gdpr-info.eu/
2. https://artificialintelligenceact.eu/the-act/

Funded by
the European Union

AI4 | eosc

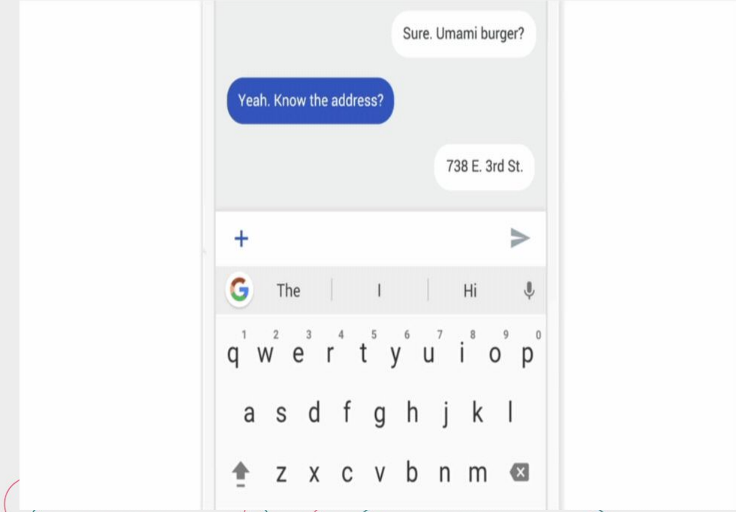
Federated Learning in Machine Learning

A method that facilitates multiple peers to collaboratively learn a common prediction model by exchanging model weights while keeping the sensitive data on the local devices
 (Kairouz et al. (2021) and Khan et al. (2023))



Examples of successful applications of FL

Google already used FL in Gboard Android:
When Gboard suggests a query, your phone stores context and interactions locally. Federated Learning uses this to improve Gboard's suggestions.



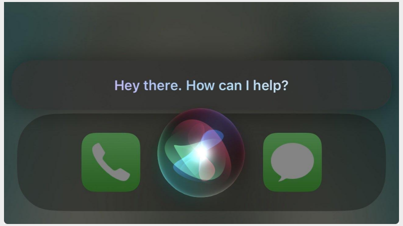
https://research.google/blog/federated-learning-collaborative-machine-learning-without-centralized-training-data/

Funded by the European Union

More Examples of successful applications of FL

- Apple has employed federated learning to improve Siri's voice recognition capabilities while maintaining user privacy[1].

- Predicting oxygen requirements for COVID-19 patients in the ER using chest X-rays and health recorde (Muto, R., et.al. (2022)).



Source: Holger R. Roth, et.al (2023)

1. https://www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/

Funded by
the European Union

Categories Federated Learning

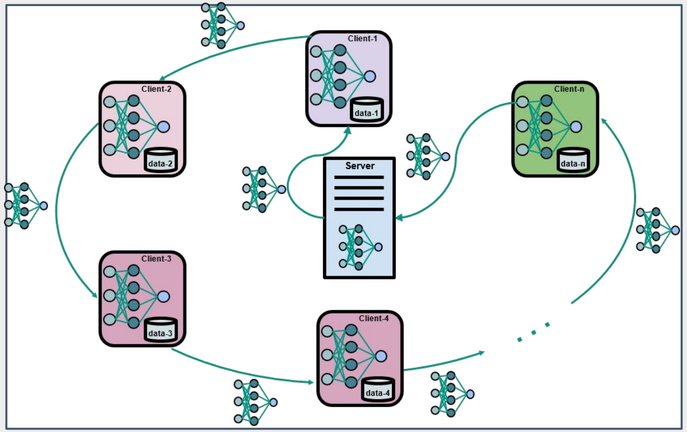
Federated Learning can be categorized as (Khan et al. (2023)**):**

- **Data distribution**
  - **Cross devices**: the model is decentralized across the edge devices and is trained using the local data on each device.
  - **Cross silos**: where the clients are a typically smaller number of organizations, institutions, or other data silos.
- **Architecture**
  - **Centralized Federated Learning**: server coordinates the training
  - **Decentralized Federated Learning**: the communication is peer to peer
- **Learning model**
  - **Horizontal Federated Learning**: each party has the same feature space but different data samples.
  - **Vertical Federated Learning**: datasets of each party share the same samples/users while holding different features (Liu, Y., et al. (2023)).

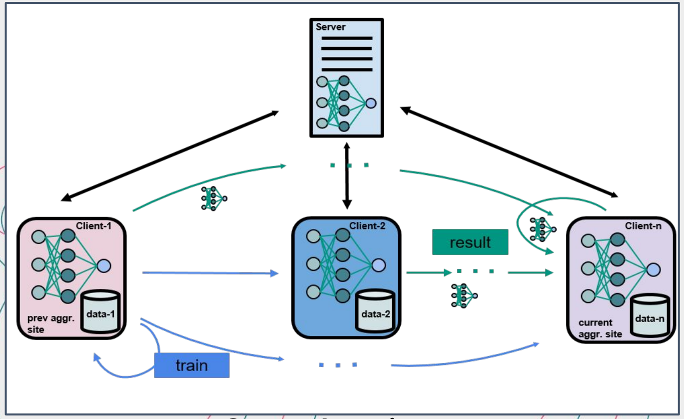
AI4 eosc

Workflow in FL: Communication Strategies

- **Scatter and gather**: global model parameters are distributed to client devices for local training; updated parameters are then aggregated.
- **Cyclic Learning** (Chang, K., et al. (2018)): the server selects a subset of clients. Training is done following a predetermined sequential order set by the server.
- **Swarm Learning** (Warnat-Herresthal, S. (2021)): a decentralized subset of FL where orchestration and aggregation is performed by the clients



Centralized cyclic Learning



Swarm Learning

Funded by
the European Union

Model Aggregation

Model Aggregation in FL is a further development of distributed learning that is specifically tailored to the challenges of **unbalanced** and **non-independent**, **non-identically distributed data (non-IID)**.

- **FedAvg**: Local weights are collected and aggregated again after local training, using weighted average.
- **FedProx**: Loss function added to penalize the local weights of clients deviating from the global model.
- **FedOpt**: Added option of using a specified Optimizer and Learning Rate Scheduler when updating the global model (like SGD to aggregate the weights of the model).
- **Scaffold**: Added correction term to the network parameters during local training by calculating the discrepancy between the global parameters.
- **Ditto:** is a method for federated learning that improves fairness and robustness by personalizing the learning objective for each device.

Funded by
the European Union

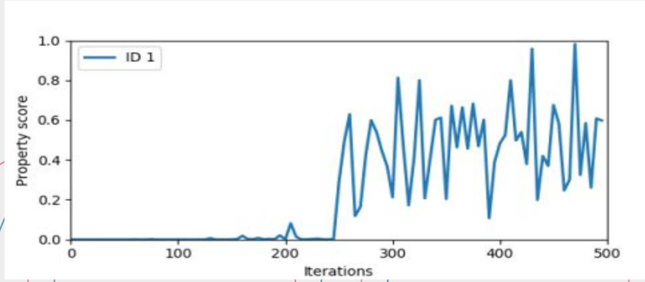
Possible Issues with Federated Learning!

Reconstruction attack (Truong et al. (2021)) :

- The original training data samples can be reconstructed from the model weights.

- membership tracing i.e., to check if a given data point belongs to a training dataset, or when a participant whose local data has a certain property, joined collaborative training.

Reconstructing an input image using the gradient.. On the left: Image extracted from the validation dataset. In the middle: Reconstruction generated by a ResNet-18 model trained on ImageNet Right: Reconstruction from a trained ResNet-152. **Geiping, J. et.al, (2020)**



Inferring that a participant whose local data has the property of interest has joined the training. **Melis, L. et.al, (2019)**

Funded by the European Union

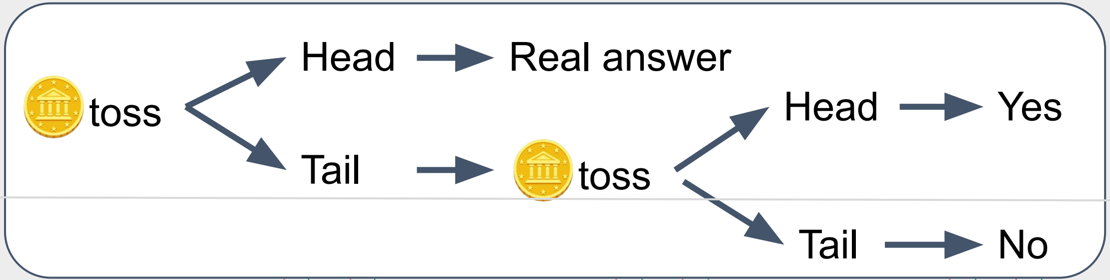
Solutions

- **Data Anonymization :** a technique to hide or remove sensitive attributes, such as personally identifiable information (PII) (Narayanan, A.& Shmatikov, V. (2008)).

- **Differential Privacy (DP)**[1]:
  - It provides a formal definition of privacy by introducing noise to query responses to prevent the disclosure of sensitive information.
  - Differential privacy mechanisms include Laplace noise addition, exponential mechanism, and more.

Do you use drug?

Yes   No



1. https://github.com/google/differential-privacy

Solutions

- **Secure Multi-party Computation (SMPC)** (Zapechnikov (2022)): is a cryptographic technique that enables multiple parties to jointly compute a function over their private inputs while keeping those inputs confidential.
  - Example: Additive secret sharing

Ana: 100
Jorge: 200
Carolin: 300

Ana	Jorge	Carolin
50	30	20
-80	100	180
0	350	-50

→

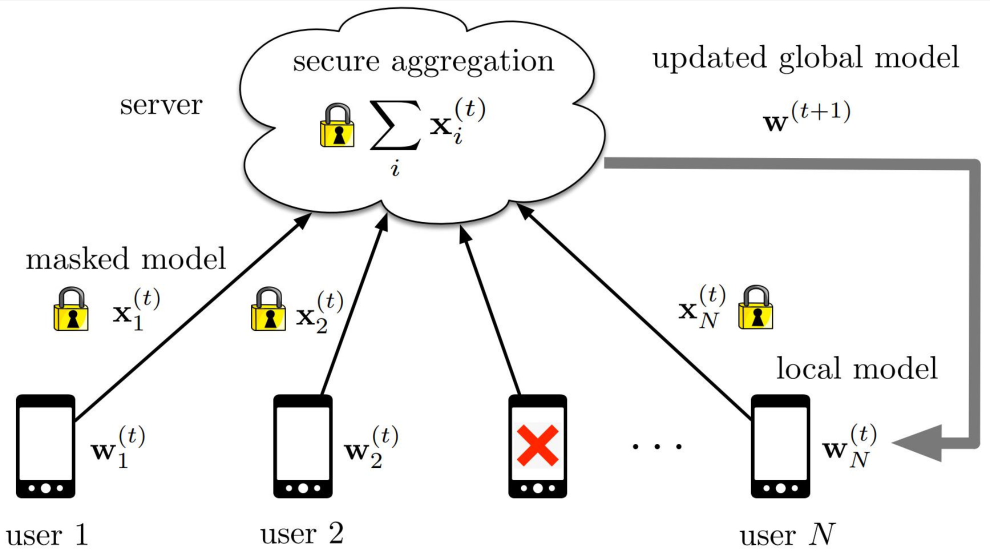
Ana	Jorge	Carolin
50	30	20
-80	100	180
0	350	-50
-30	**480**	**150**

sum=600
Average= 200

Funded by the European Union

Solutions

- **Secure Model Aggregation (SMA):** in the same way as SMPC, here, the server works with encrypted models in which the individual contributions of the clients remain unknown during the aggregation process



Secure model aggregation using masked model. Figure from: Lu, S., et.al. (2023).

AI4 eosc

Solutions

- **Homomorphic Encryption (HE)** (Behera et al. (2020)): allows computations to be performed on encrypted data.
    - **Fully Homomorphic Encryption (FHE):** allows to perform any number of operations.
    - **Somewhat Homomorphic Encryption (SWHE):** limits the number of operations that can be performed on encrypted data.
    - **Partially Homomorphic Encryption (PHE):** allows only one type of operation to be performed.

$$\begin{tikzcd} \mathbf{m} \arrow[r, "F"] \arrow[d, shift left, "\text{Encrypt}"] & \mathbf{F(m)} \arrow[d, shift left, "\text{Encrypt}"] \\ \mathbf{m} \arrow[r, "F"] \arrow[u, shift left, "\text{Decrypt}"] & \mathbf{F(m)} \arrow[u, shift left, "\text{Decrypt}"] \end{tikzcd}$$

$$F \in \mathbb{R}[+and.] : fully\ homomorphic$$

$$F \in \mathbb{R}_n[+and.] : somewhat\ homomorphic$$

$$F \in \mathbb{R}[+or.] : partially\ homomorphic$$

Funded by
the European Union

Another possible attack

Poisoning attacks on Federated Learning (Truong et al. (2021)) :

During model training in FL, participants can manipulate the training process by introducing arbitrary updates, potentially poisoning the global model.

Solution?

- Model Anomaly Detection
  (Fung, C., Yoon, C.J., Beschastnikh, I., (2018) and Jagielski, M., et.al. (2018)).
- Not applicable when using secure model aggregation

This problem needs more research

Funded by
the European Union

Model FL Frameworks

- **Flower**[1]**:**
  - is a flexible, easy-to-use and easily understood open-source FL framework.
  - It is framework-agnostic meaning that nearly every ML model can be easily migrated to the federated setting.
  - Well-suited for research and study projects.
- **NVIDIA Federated Learning Application Runtime Environment (NVFlare)**[2]**:**
  - NVFlare is a business-ready FL framework by Nvidia.
  - It supports a variety of models,
  - NVFlare is framework-agnostic.

1. https://flower.ai/
2. https://nvflare.readthedocs.io/en/main/

Funded by
the European Union

Model FL Frameworks

- **TensorFlow Federated (TFF)[1]:**
  - Developed by Google
  - Specifically designed for compatibility with TensorFlow
  - Integrates smoothly with existing TensorFlow workflows
- **PySyft/PyGrid[2,3]:** a Python library for secure Federated Learning
  - Developed by OpenMined
  - Compatibility with popular deep learning frameworks like PyTorch and TensorFlow

1. https://www.tensorflow.org/federated
2. https://blog.openmined.org/tag/pysyft/
3. https://blog.openmined.org/what-is-pygrid-demo/

Funded by
the European Union

Model FL Frameworks

- **Federated AI Technology Enabler (FATE)[1]**: is an open-source Federated Learning platform developed by WeBank's AI Group.
  - Business-ready FL frameworks
  - The framework comes with a large number of modules
  - It has a backend for the Deep Learning libraries PyTorch and TensorFlow

1. https://fate.fedai.org/

Funded by
the European Union

Conclusions

Key Considerations for Federated Learning:

- **Optimizing Client Selection:**
  - Use techniques to improve the response efficiency of end devices.
  - Select customers based on data quality and reliability.
- **Aggregation Algorithm Selection:**
  - Choose the most suitable algorithms for data aggregation.
  - Consider scalability, efficiency, and accuracy of algorithms.
- **Framework Customization:**
  - Tailor framework selection to meet specific task requirements.

Funded by
the European Union

Conclusions

Key Considerations for Federated Learning:

- **Security Enhancement:**
  - Implement robust security measures for communication and data sharing.
  - Ensure encryption, authentication, and privacy-preserving techniques.
- **Compliance and Ethical Considerations**:
  - Adhere to data privacy regulations and ethical guidelines.

Funded by
the European Union

Conclusions

- **Positive aspects of FL:**
  - Data transferring minimization
  - Build a larger and more diverse dataset
  - Train a more general and global model
  - International collaboration
- **Considerable aspects:**
  - Security issues like model poisoning
  - Biases and Fairness
  - Interpretability

Funded by
the European Union

References

Behera, S., & Prathuri, J. R. (2020). Application of Homomorphic Encryption in Machine Learning. In 2020 2nd PhD Colloquium on Ethically Driven Innovation and Technology for Society (PhD EDITS) (pp. 1-2). Bangalore, India. https://doi.org/10.1109/PhDEDITS51180.2020.9315305

Chang, K., Balachandar, N., et al. (2018). Distributed deep learning networks among institutions for medical imaging. *Journal of the American Medical Informatics Association, 25*(8), 945-954. https://doi.org/10.1093/jamia/ocy017

Fung, C., Yoon, C.J., Beschastnikh, I., (2018). Mitigating sybils in federated learning poisoning. arXiv preprint arXiv:1808.04866 .

Geiping, J., Bauermeister, H., Dröge, H., & Moeller, M. (2020). Inverting Gradients -- How easy is it to break privacy in federated learning? [Preprint]. arXiv. https://arxiv.org/abs/2003.14053

Holger R. Roth, et.al. (2022). NVIDIA FLARE: Federated Learning from Simulation to Real-World. arXiv. https://arxiv.org/abs/2210.13291

Jagielski, M., Oprea, A., Biggio, et.al. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning, in: 2018 IEEE Symposium on Security and Privacy (SP), IEEE. pp. 19–35.

References

Kairouz, P., McMahan, H. B., Avent, et al. (2021). Advances and Open Problems in Federated Learning. https://ieeexplore.ieee.org/document/9464278

Khan, M., Glavin, F. G., & Nickles, M. (2023). Federated Learning as a Privacy Solution - An Overview. Procedia Computer Science, 217, 316-325. https://doi.org/10.1016/j.procs.2022.12.227

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (SP 2008)* (pp. 111-125). Oakland, CA, USA. https://doi.org/10.1109/SP.2008.33

Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 691-706). San Francisco, CA, USA. https://doi.org/10.1109/SP.2019.00029

Muto, R., et.al. (2022). Predicting oxygen requirements in patients with coronavirus disease 2019 using an artificial intelligence-clinician model based on local non-image data. *Frontiers in Medicine (Lausanne)*, 9, 1042067. https://doi.org/10.3389/fmed.2022.1042067

References

Li, T., Hu, S., Beirami, A., & Smith, V. (2021). Ditto: Fair and Robust Federated Learning Through Personalization. arXiv. https://arxiv.org/abs/2012.04221

Liu, Y., et al. (2023). Vertical Federated Learning: Concepts, Advances, and Challenges. *IEEE Transactions on Knowledge & Data Engineering*, 01(01), 1-20. https://doi.org/10.1109/TKDE.2024.3352628

Lu, S., Li, R., Liu, W., Guan, C., & Yang, X. (2023). Top-k sparsification with secure aggregation for privacy-preserving federated learning. *Computers & Security*, 124, 102993. https://doi.org/10.1016/j.cose.2022.102993

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., & McMahan, H. B. (2021). Adaptive Federated Optimization. arXiv. https://arxiv.org/abs/2003.00295

Sai Praneeth Karimireddy, et al. (2021). SCAFFOLD: Stochastic Controlled Averaging for Federated Learning. *arXiv.* https://arxiv.org/abs/1910.06378

Tian, L., Kumar Sahu, A., Talwalkar, A. S., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, 37.

Funded by the European Union

References

Tian Li, Anit Kumar Sahu, et al. (2020). Federated Optimization in Heterogeneous Networks. *arXiv*. https://arxiv.org/abs/1812.06127

Truong, N., Sun, K., Wang, S., Guitton, F., & Guo, Y. (2021). Privacy preservation in federated learning: An insightful survey from the GDPR perspective. Computers & Security, 110, 102402. https://doi.org/10.1016/j.cose.2021.102402

Zapechnikov, S. (2022). Secure multi-party computations for privacy-preserving machine learning. Procedia Computer Science, 213, 523-527. https://doi.org/10.1016/j.procs.2022.11.100

Warnat-Herresthal, S., Schultze, H., Shastry, K. L., et al. (2021). Swarm Learning for decentralized and confidential clinical machine learning. *Nature, 594*, 265–270. https://doi.org/10.1038/s41586-021-03583-3

Funded by the European Union