Dalle Molle Institute for
Artificial Intelligence
USI – SUPSI

Swiss Institute of
Bioinformatics

Introduction to Federated Learning

SIB course on Federated Learning in Bioinformatics

Daniele Malpetti, Lecturer-Researcher
Dalle Molle Institute for Artificial Intelligence (IDSIA), SUPSI
Swiss Institute of Bioinformatics (SIB)

SUPSI

Dalle Molle Institute for Artificial Intelligence USI – SUPSI

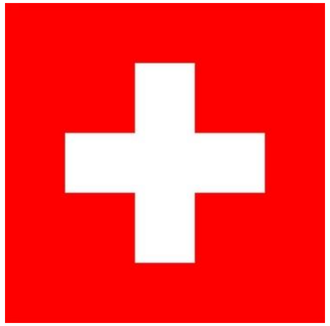
Swiss Institute of Bioinformatics

Federated Learning

How everything started

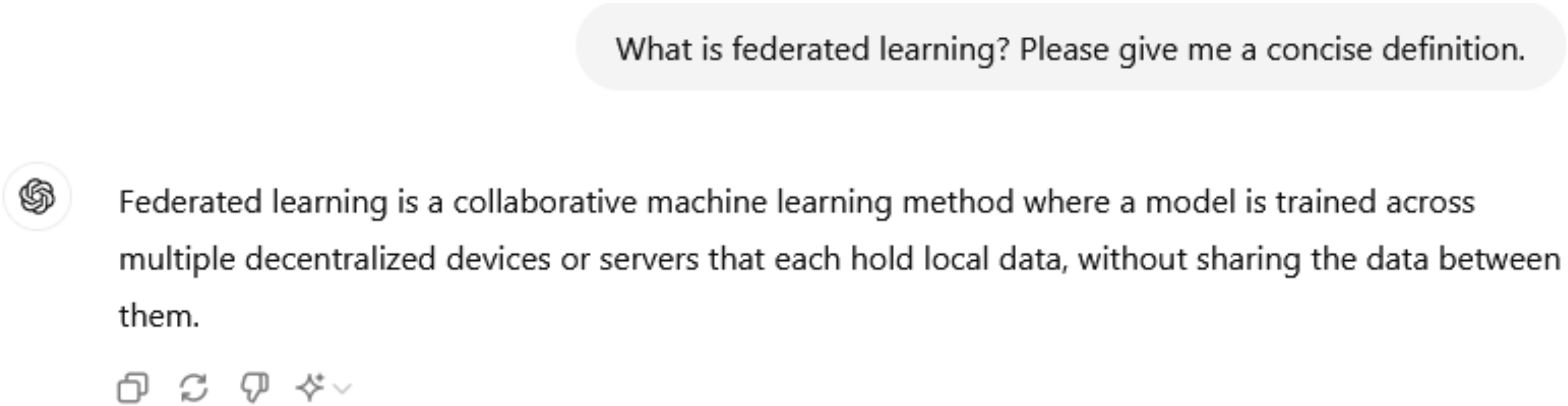
SUPSI

Disclaimer



Federated Learning is **not** something related to the Swiss Confederation.

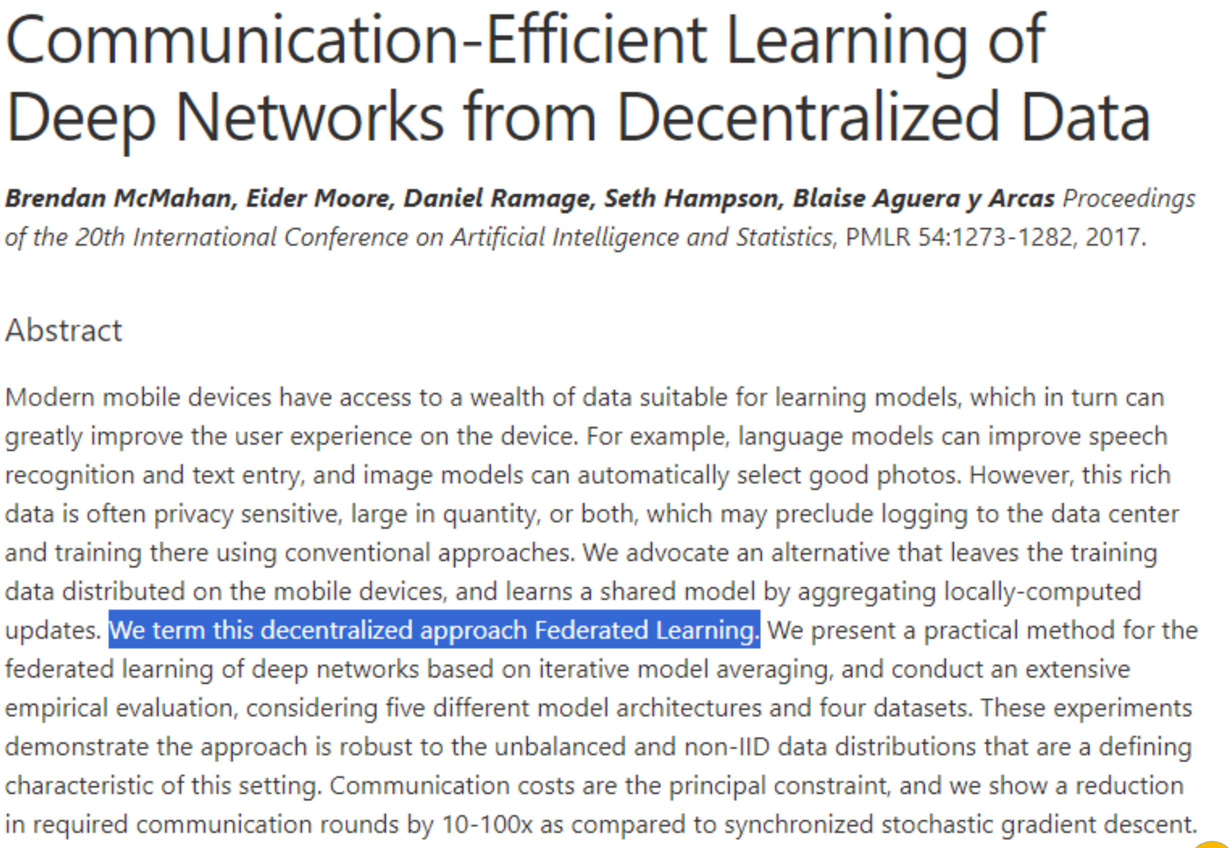
A basic definition



What is federated learning? Please give me a concise definition.

Federated learning is a collaborative machine learning method where a model is trained across multiple decentralized devices or servers that each hold local data, without sharing the data between them.

The origin of the name

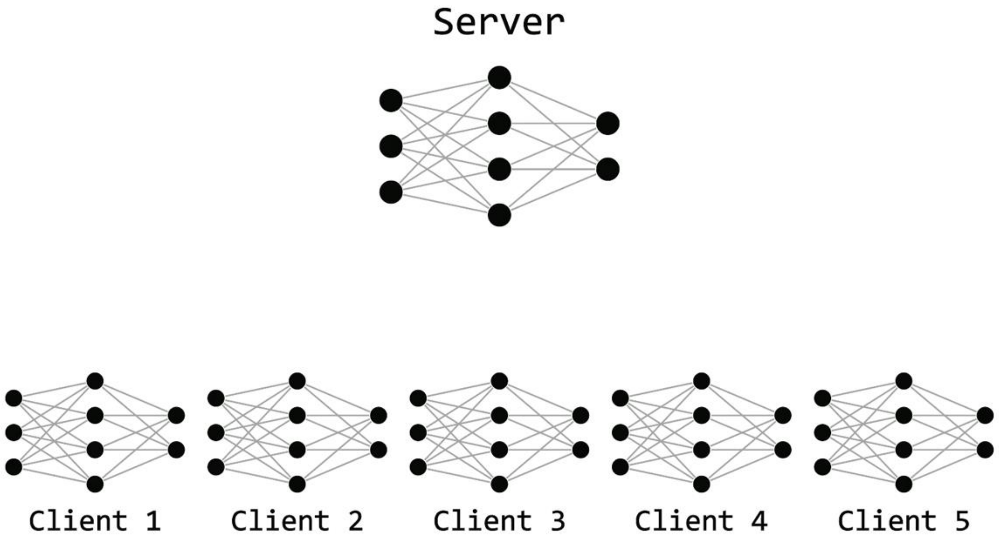


Communication-Efficient Learning of Deep Networks from Decentralized Data

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR 54:1273-1282, 2017.

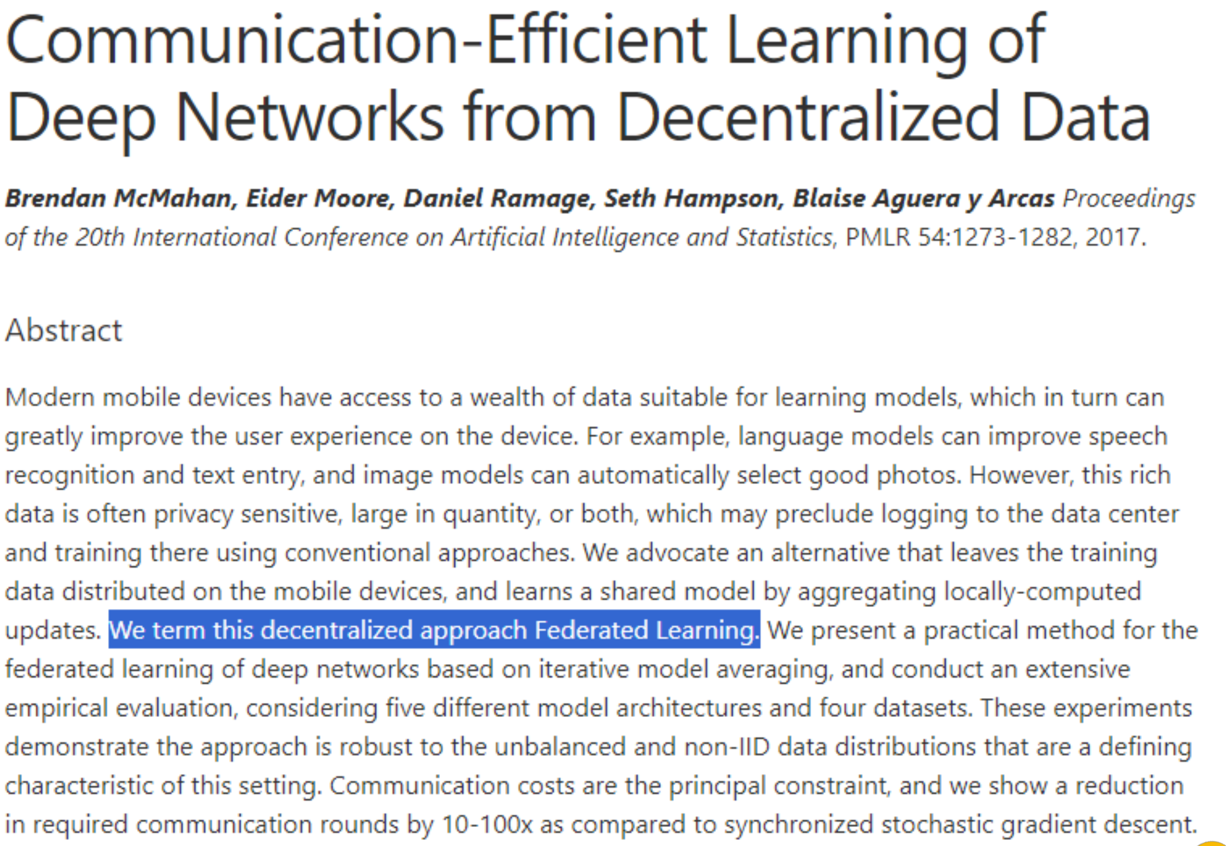
Abstract

Modern mobile devices have access to a wealth of data suitable for learning models, which in turn can greatly improve the user experience on the device. For example, language models can improve speech recognition and text entry, and image models can automatically select good photos. However, this rich data is often privacy sensitive, large in quantity, or both, which may preclude logging to the data center and training there using conventional approaches. We advocate an alternative that leaves the training data distributed on the mobile devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning. We present a practical method for the federated learning of deep networks based on iterative model averaging, and conduct an extensive empirical evaluation, considering five different model architectures and four datasets. These experiments demonstrate the approach is robust to the unbalanced and non-IID data distributions that are a defining characteristic of this setting. Communication costs are the principal constraint, and we show a reduction in required communication rounds by 10-100x as compared to synchronized stochastic gradient descent.



FedAvg: averaging of weights

The origin of the name



Communication-Efficient Learning of Deep Networks from Decentralized Data

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR 54:1273-1282, 2017.

Abstract

Modern mobile devices have access to a wealth of data suitable for learning models, which in turn can greatly improve the user experience on the device. For example, language models can improve speech recognition and text entry, and image models can automatically select good photos. However, this rich data is often privacy sensitive, large in quantity, or both, which may preclude logging to the data center and training there using conventional approaches. We advocate an alternative that leaves the training data distributed on the mobile devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning. We present a practical method for the federated learning of deep networks based on iterative model averaging, and conduct an extensive empirical evaluation, considering five different model architectures and four datasets. These experiments demonstrate the approach is robust to the unbalanced and non-IID data distributions that are a defining characteristic of this setting. Communication costs are the principal constraint, and we show a reduction in required communication rounds by 10-100x as compared to synchronized stochastic gradient descent.

Google

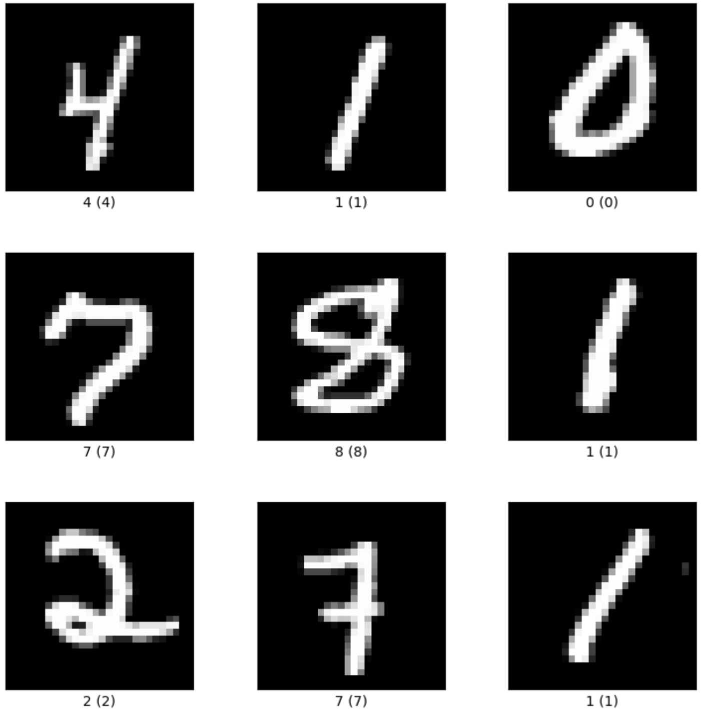
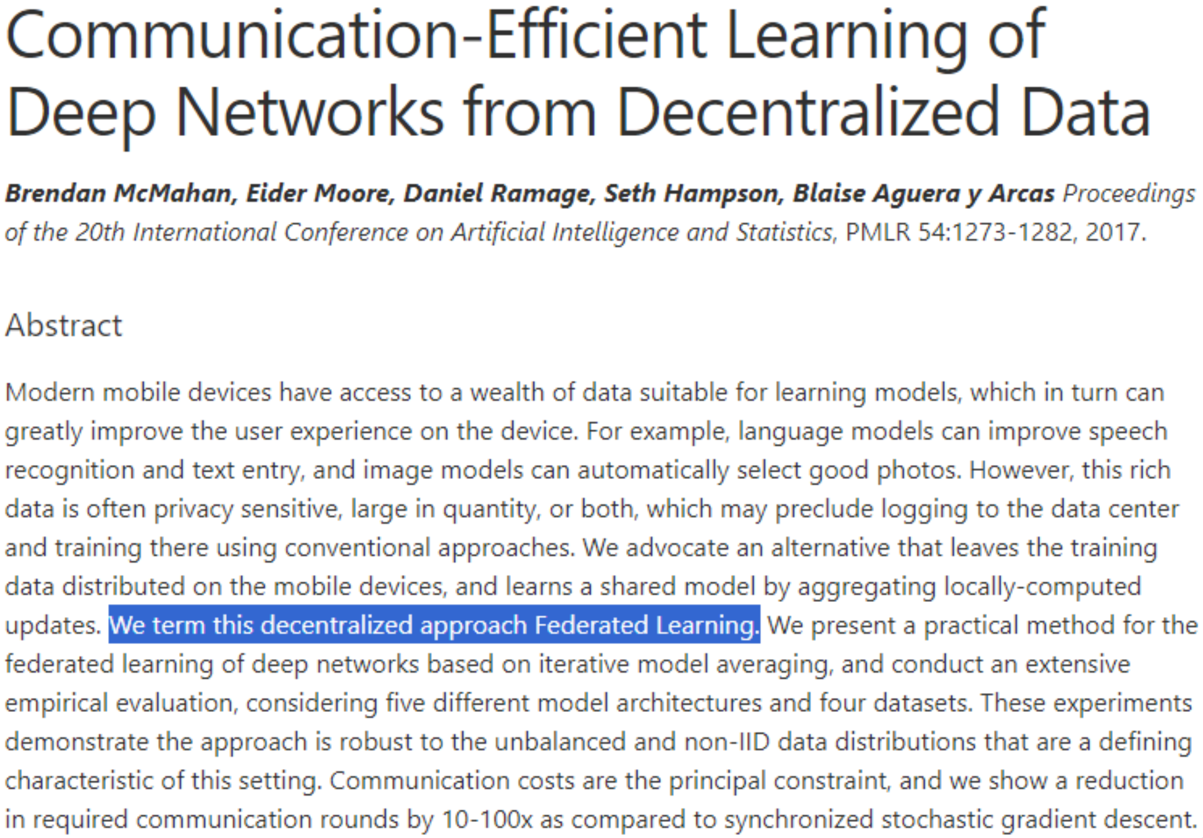


Image classification task

IID

The origin of the name

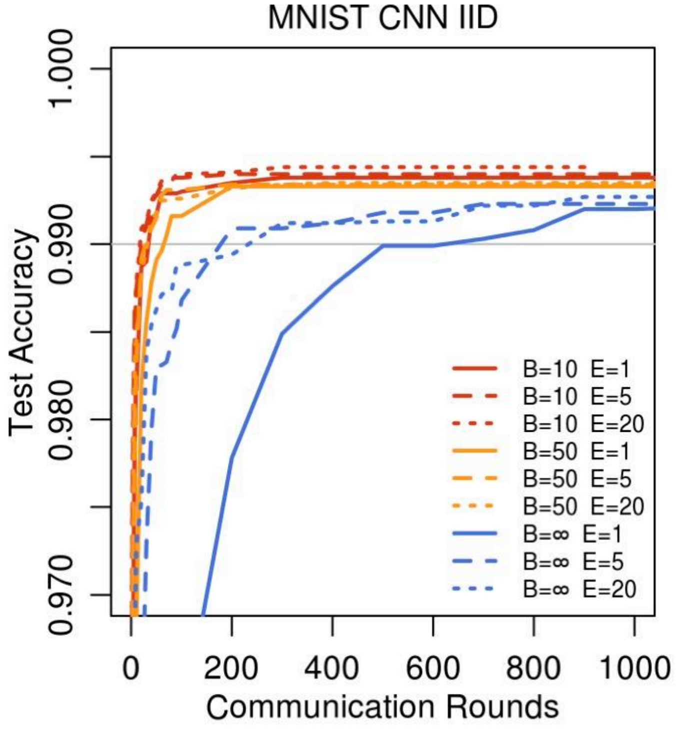


Communication-Efficient Learning of Deep Networks from Decentralized Data

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR 54:1273-1282, 2017.

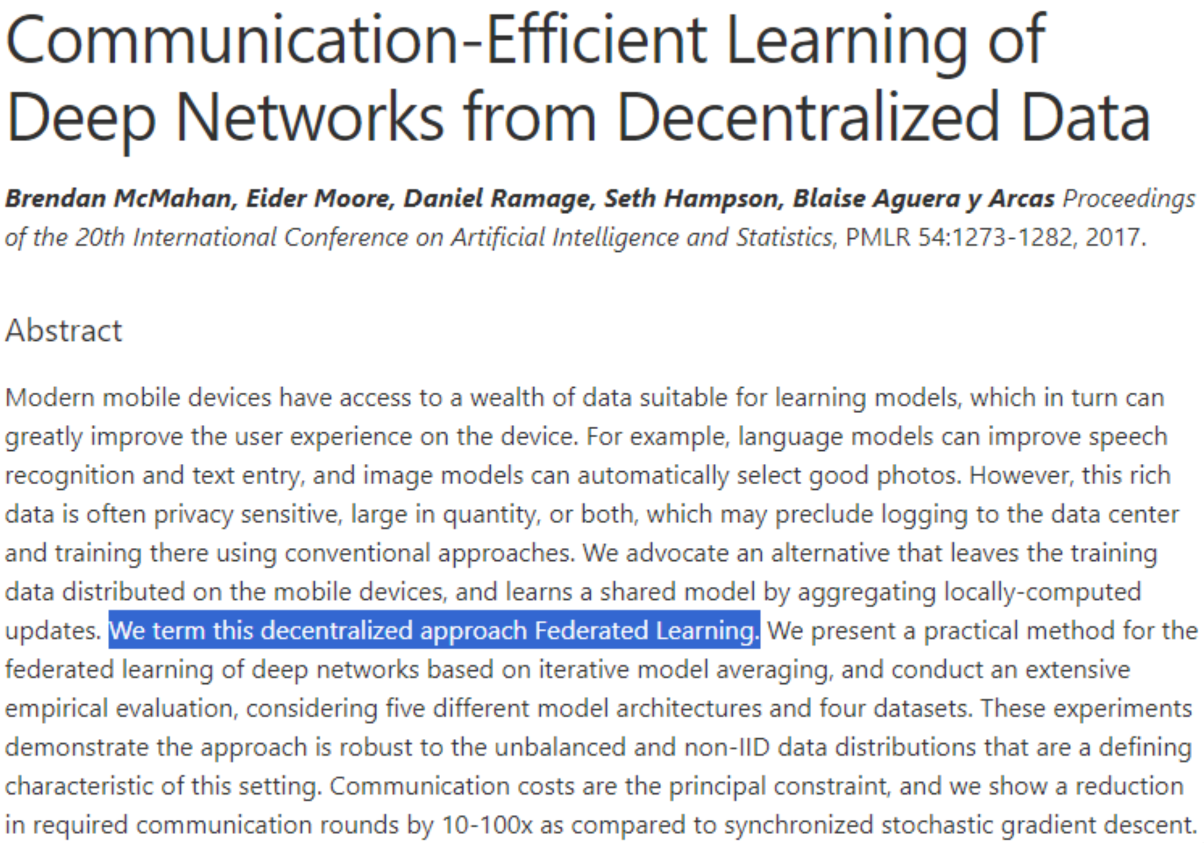
Abstract

Modern mobile devices have access to a wealth of data suitable for learning models, which in turn can greatly improve the user experience on the device. For example, language models can improve speech recognition and text entry, and image models can automatically select good photos. However, this rich data is often privacy sensitive, large in quantity, or both, which may preclude logging to the data center and training there using conventional approaches. We advocate an alternative that leaves the training data distributed on the mobile devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning. We present a practical method for the federated learning of deep networks based on iterative model averaging, and conduct an extensive empirical evaluation, considering five different model architectures and four datasets. These experiments demonstrate the approach is robust to the unbalanced and non-IID data distributions that are a defining characteristic of this setting. Communication costs are the principal constraint, and we show a reduction in required communication rounds by 10-100x as compared to synchronized stochastic gradient descent.



Target accuracy reached

The origin of the name



Communication-Efficient Learning of Deep Networks from Decentralized Data

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR 54:1273-1282, 2017.

Abstract

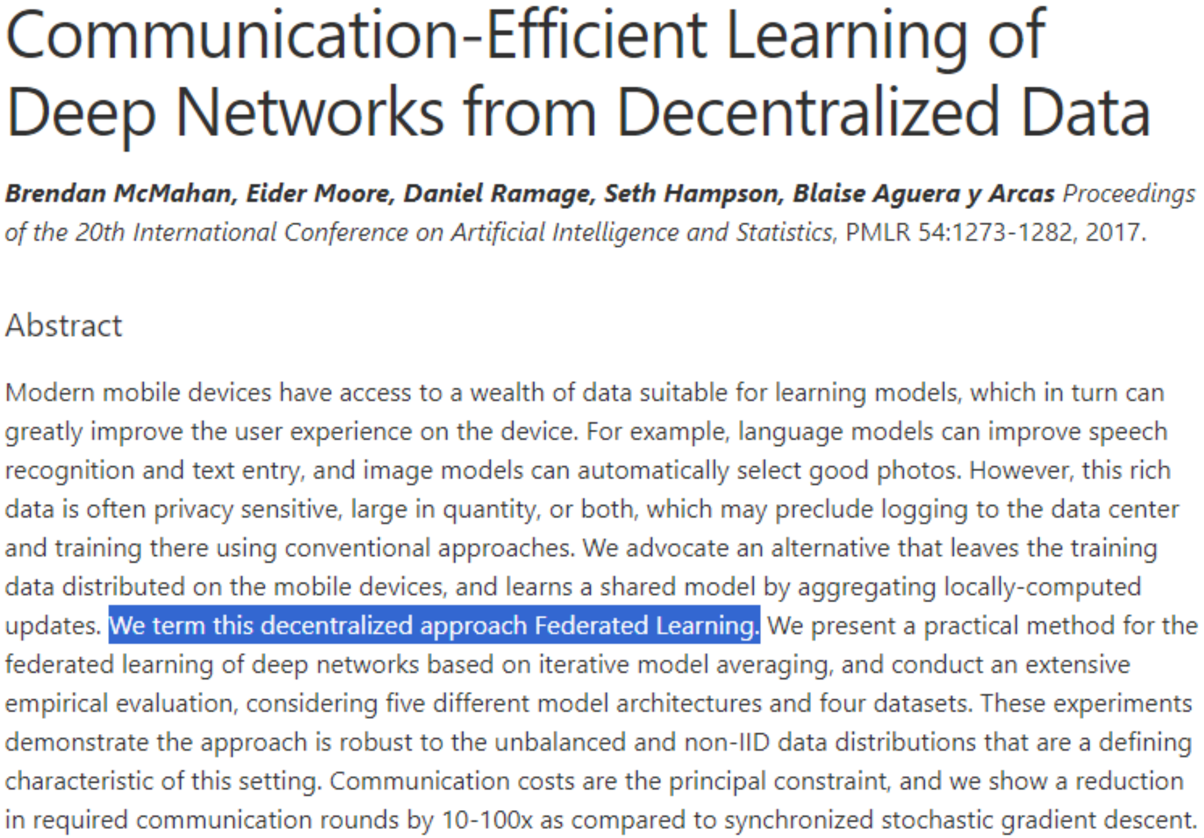
Modern mobile devices have access to a wealth of data suitable for learning models, which in turn can greatly improve the user experience on the device. For example, language models can improve speech recognition and text entry, and image models can automatically select good photos. However, this rich data is often privacy sensitive, large in quantity, or both, which may preclude logging to the data center and training there using conventional approaches. We advocate an alternative that leaves the training data distributed on the mobile devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning. We present a practical method for the federated learning of deep networks based on iterative model averaging, and conduct an extensive empirical evaluation, considering five different model architectures and four datasets. These experiments demonstrate the approach is robust to the unbalanced and non-IID data distributions that are a defining characteristic of this setting. Communication costs are the principal constraint, and we show a reduction in required communication rounds by 10-100x as compared to synchronized stochastic gradient descent.

Google

Next word prediction task

Non-IID

The origin of the name

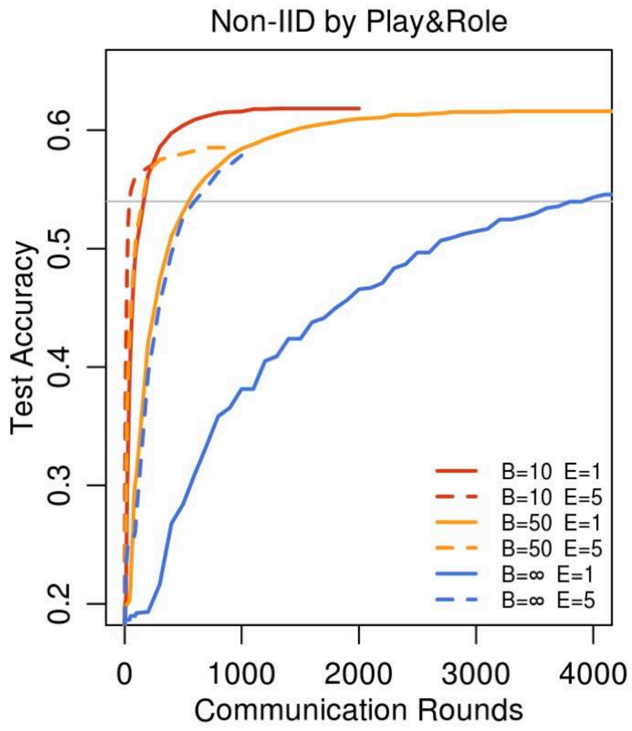


Communication-Efficient Learning of Deep Networks from Decentralized Data

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR 54:1273-1282, 2017.

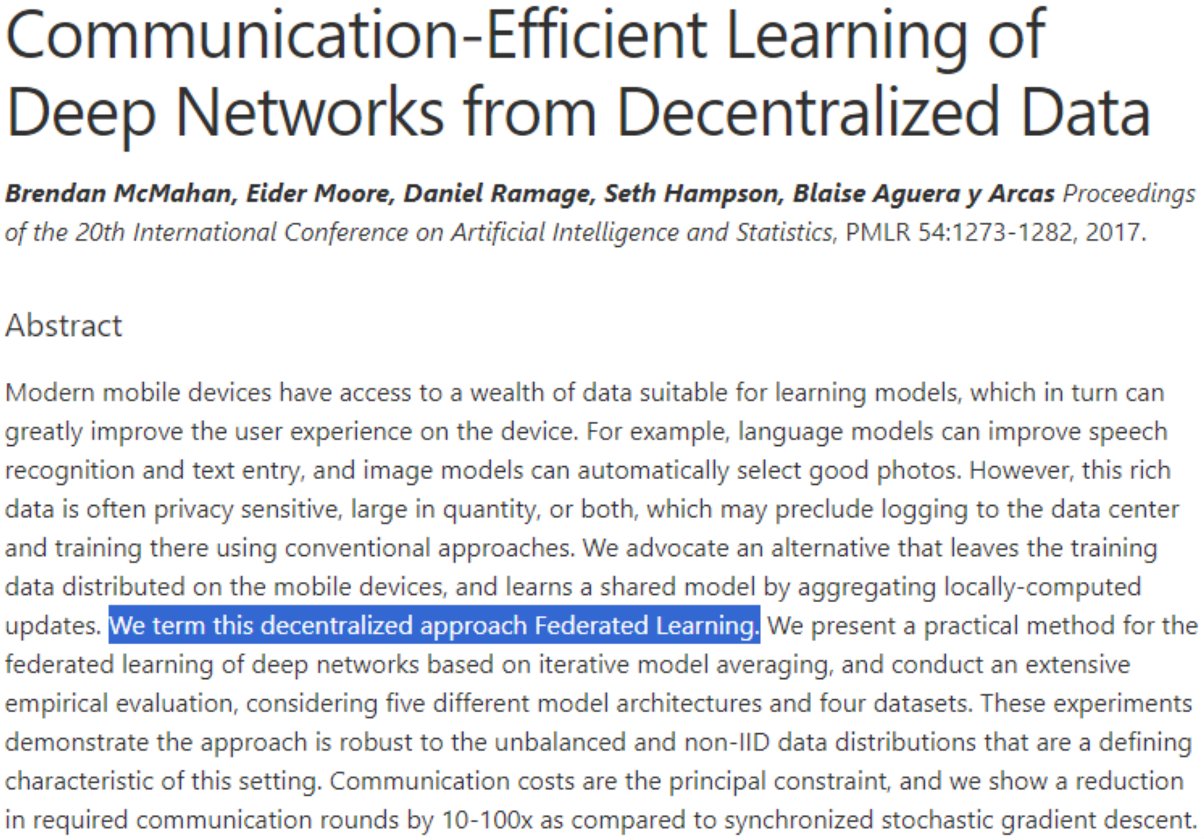
Abstract

Modern mobile devices have access to a wealth of data suitable for learning models, which in turn can greatly improve the user experience on the device. For example, language models can improve speech recognition and text entry, and image models can automatically select good photos. However, this rich data is often privacy sensitive, large in quantity, or both, which may preclude logging to the data center and training there using conventional approaches. We advocate an alternative that leaves the training data distributed on the mobile devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning. We present a practical method for the federated learning of deep networks based on iterative model averaging, and conduct an extensive empirical evaluation, considering five different model architectures and four datasets. These experiments demonstrate the approach is robust to the unbalanced and non-IID data distributions that are a defining characteristic of this setting. Communication costs are the principal constraint, and we show a reduction in required communication rounds by 10-100x as compared to synchronized stochastic gradient descent.



Target accuracy reached

The origin of the name

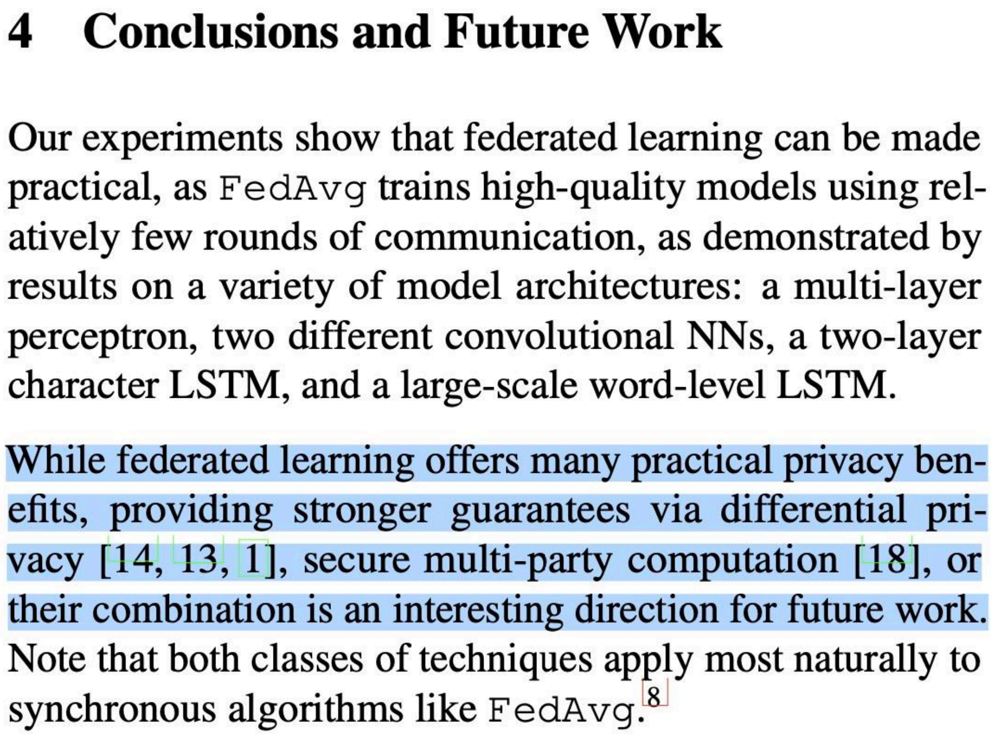


Communication-Efficient Learning of Deep Networks from Decentralized Data

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Aguera y Arcas *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR 54:1273-1282, 2017.

Abstract

Modern mobile devices have access to a wealth of data suitable for learning models, which in turn can greatly improve the user experience on the device. For example, language models can improve speech recognition and text entry, and image models can automatically select good photos. However, this rich data is often privacy sensitive, large in quantity, or both, which may preclude logging to the data center and training there using conventional approaches. We advocate an alternative that leaves the training data distributed on the mobile devices, and learns a shared model by aggregating locally-computed updates. We term this decentralized approach Federated Learning. We present a practical method for the federated learning of deep networks based on iterative model averaging, and conduct an extensive empirical evaluation, considering five different model architectures and four datasets. These experiments demonstrate the approach is robust to the unbalanced and non-IID data distributions that are a defining characteristic of this setting. Communication costs are the principal constraint, and we show a reduction in required communication rounds by 10-100x as compared to synchronized stochastic gradient descent.



4 Conclusions and Future Work

Our experiments show that federated learning can be made practical, as FedAvg trains high-quality models using relatively few rounds of communication, as demonstrated by results on a variety of model architectures: a multi-layer perceptron, two different convolutional NNs, a two-layer character LSTM, and a large-scale word-level LSTM.

While federated learning offers many practical privacy benefits, providing stronger guarantees via differential privacy [14, 13, 1], secure multi-party computation [18], or their combination is an interesting direction for future work. Note that both classes of techniques apply most naturally to synchronous algorithms like FedAvg.[8]

From the paper

Similar (?) approaches

Distributed Computing

Centralised data, parallel processing, efficiency-focused, identical distribution, no privacy preservation.

Meta-analysis

Decentralised results, statistical aggregation, single-step synthesis, privacy-preserving (no raw data).

Trusted Research Environments (TREs)

Centralised/decentralised storage, secure access, strong governance, restricted data movement, privacy through control not distribution.

Dalle Molle Institute for Artificial Intelligence
USI – SUPSI

Swiss Institute of Bioinformatics

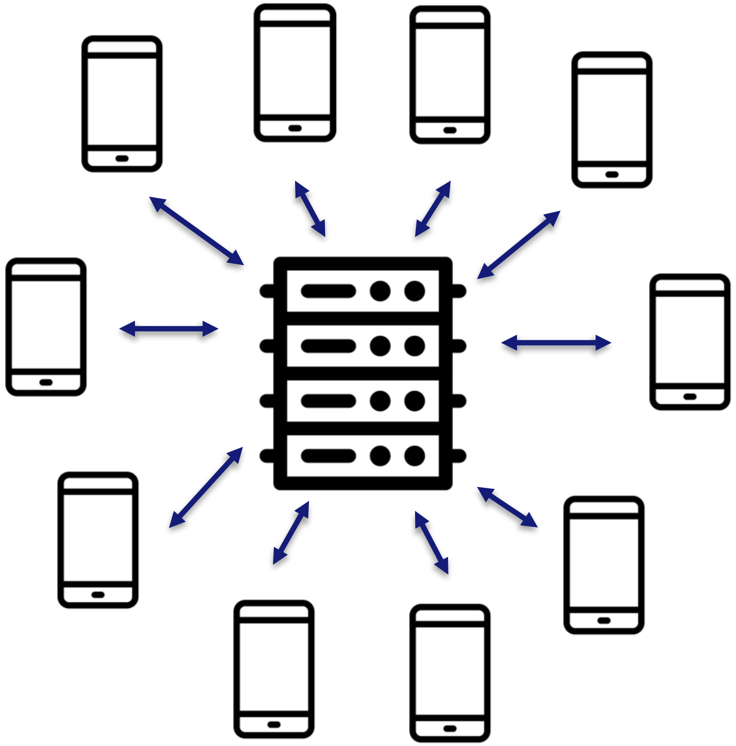
Federated Learning

More characterization

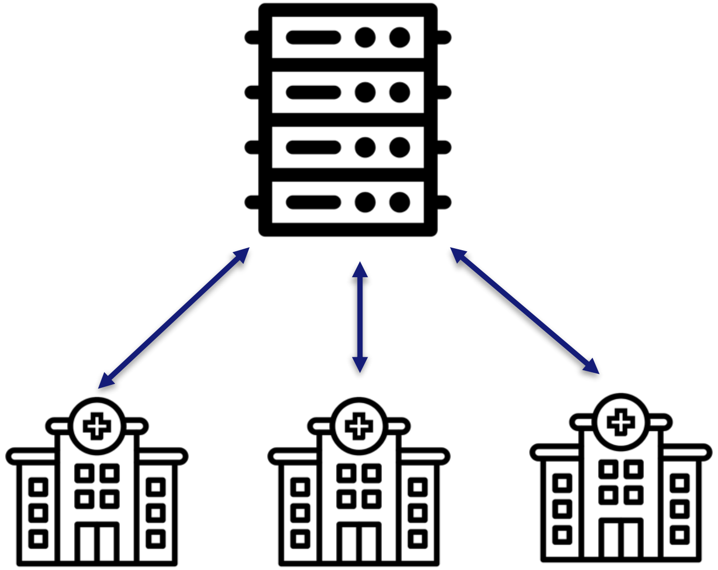
SUPSI

Fields of Use of Federated Learning

Cross-device FL



Cross-silo FL



Fields of Use of Federated Learning

Cross-device FL

EFFICIENCY: sharing models avoids centralizing a potentially extremely large amount of data at a single location.

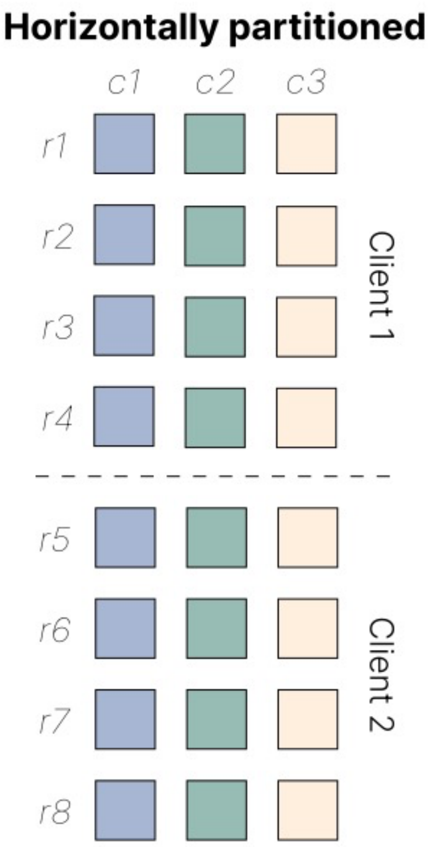
Example: a model predicting the next word for mobile phones keyboards

Cross-silo FL

PRIVACY: sharing models presents less privacy issues than sharing data (caveat already mentioned).

Example: a model predicting patients' outcomes

Data partitioning



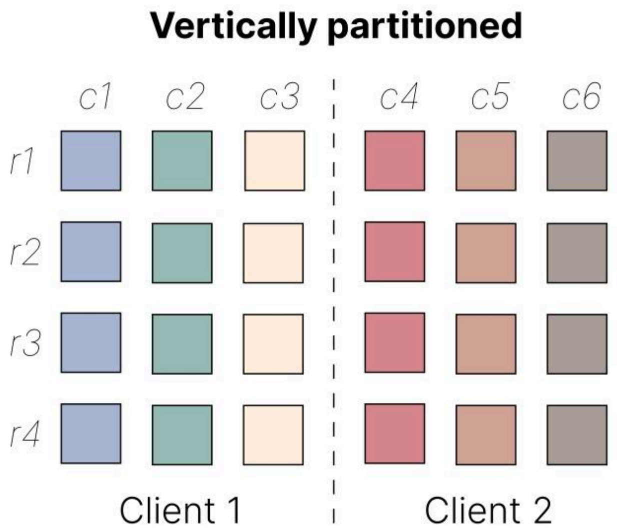
Typically, in bioinformatics, multiple parties hold the **same variables** to describe **different samples**.

For example, multiple institutions may each possess sequencing data from different individuals and conduct a differential gene expression analysis.

Harmonization of data is key!

Vertical FL

Data partitioning

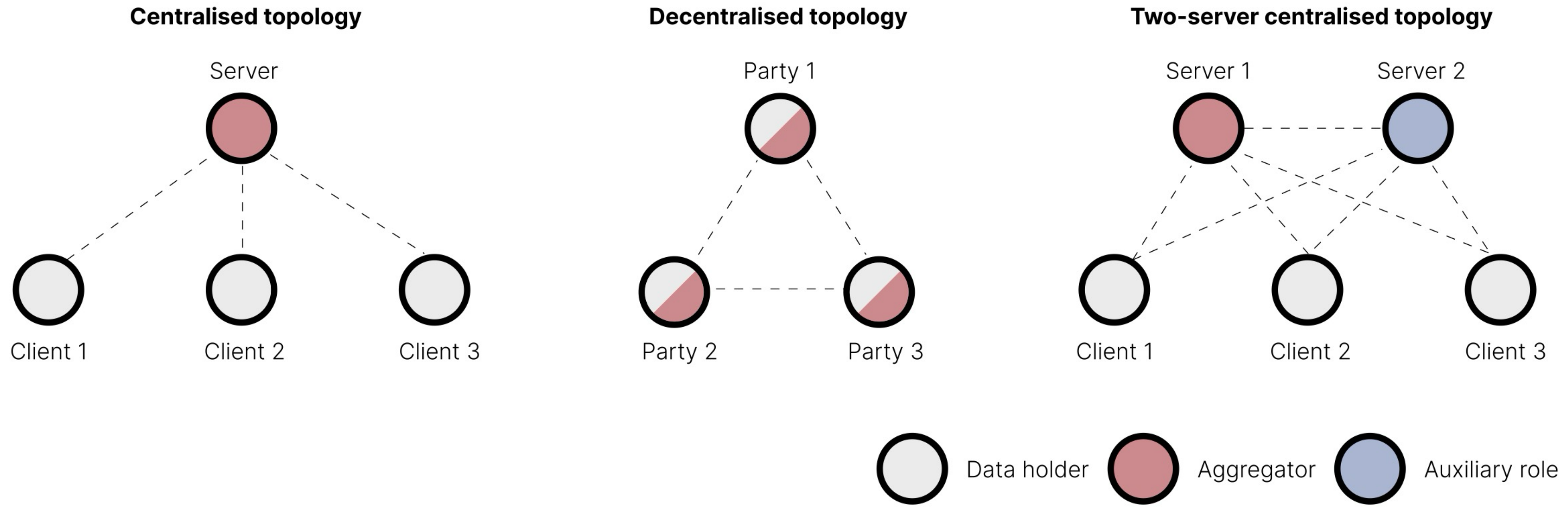


In other domains, multiple parties hold the **same samples** described by **different variables**.

For example, a bank and an insurance company having clients in common may want to collaborate to develop models for targeting their clients.

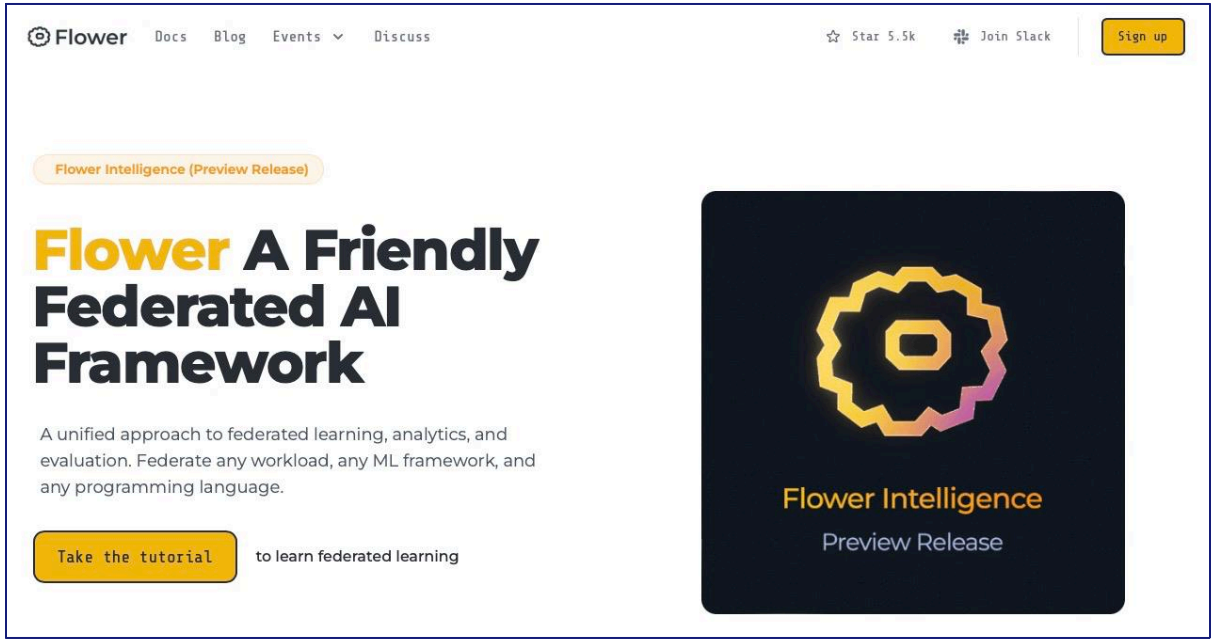
Topology

A central server is not the only way to go



Software

- Federated learning requires a **framework:** i.e., a software infrastructure that enables the design, deployment, and management of federated learning workflows, handling communication, coordination, security, and aggregation across distributed clients.



Flower A Friendly Federated AI Framework

A unified approach to federated learning, analytics, and evaluation. Federate any workload, any ML framework, and any programming language.

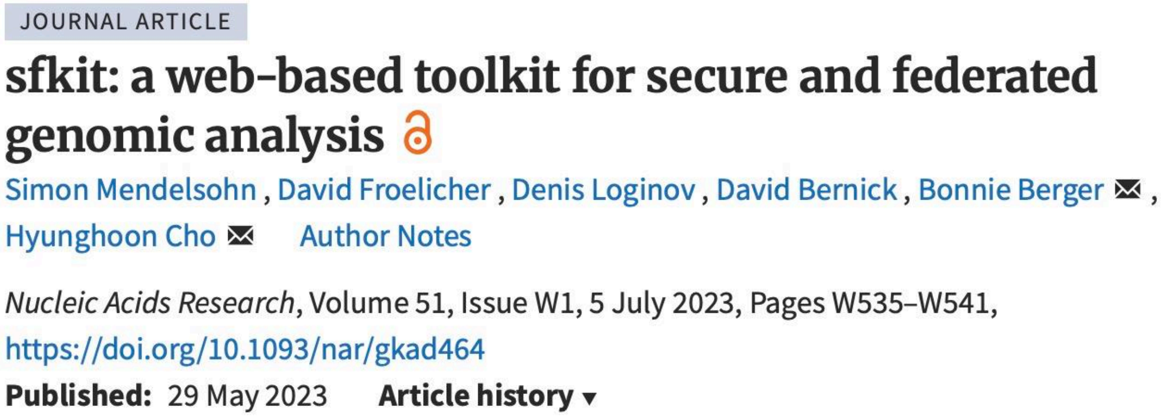
Take the tutorial to learn federated learning

Coming next!

IDSIA Dalle Molle Institute for Artificial Intelligence USI – SUPSI

SIB Swiss Institute of Bioinformatics

Ready to use software for bioinformatics



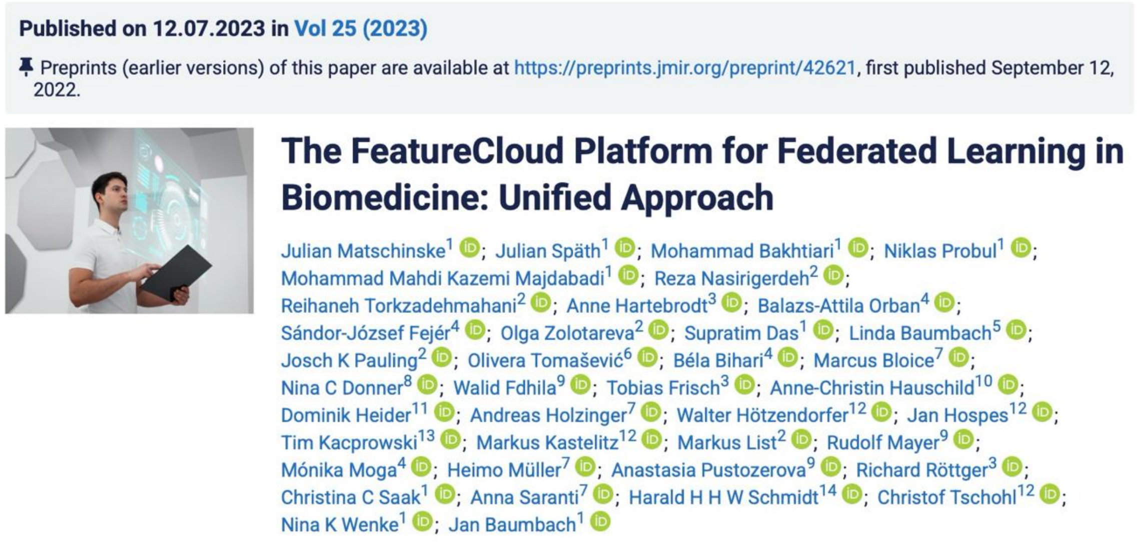
JOURNAL ARTICLE

sfkit: a web-based toolkit for secure and federated genomic analysis

Simon Mendelsohn , David Froelicher , Denis Loginov , David Bernick , Bonnie Berger ✉ , Hyunghoon Cho ✉ Author Notes

Nucleic Acids Research, Volume 51, Issue W1, 5 July 2023, Pages W535–W541, https://doi.org/10.1093/nar/gkad464

Published: 29 May 2023 **Article history ▾**



Published on 12.07.2023 in Vol 25 (2023)

Preprints (earlier versions) of this paper are available at https://preprints.jmir.org/preprint/42621, first published September 12, 2022.

The FeatureCloud Platform for Federated Learning in Biomedicine: Unified Approach

Julian Matschinske[1]; Julian Späth[1]; Mohammad Bakhtiari[1]; Niklas Probul[1]; Mohammad Mahdi Kazemi Majdabadi[1]; Reza Nasirigerdeh[2]; Reihaneh Torkzadehmahani[2]; Anne Hartebrodt[3]; Balazs-Attila Orban[4]; Sándor-József Fejér[4]; Olga Zolotareva[2]; Supratim Das[1]; Linda Baumbach[5]; Josch K Pauling[2]; Olivera Tomašević[6]; Béla Bihari[4]; Marcus Bloice[7]; Nina C Donner[8]; Walid Fdhila[9]; Tobias Frisch[3]; Anne-Christin Hauschild[10]; Dominik Heider[11]; Andreas Holzinger[7]; Walter Hötzendorfer[12]; Jan Hospes[12]; Tim Kacprowski[13]; Markus Kastelitz[12]; Markus List[2]; Rudolf Mayer[9]; Mónika Moga[4]; Heimo Müller[7]; Anastasia Pustozerova[9]; Richard Röttger[3]; Christina C Saak[1]; Anna Saranti[7]; Harald H H W Schmidt[14]; Christof Tschohl[12]; Nina K Wenke[1]; Jan Baumbach[1]

https://sfkit.org

https://featurecloud.ai

USI SUPSI

Dalle Molle Institute for
Artificial Intelligence
USI – SUPSI

Swiss Institute of
Bioinformatics

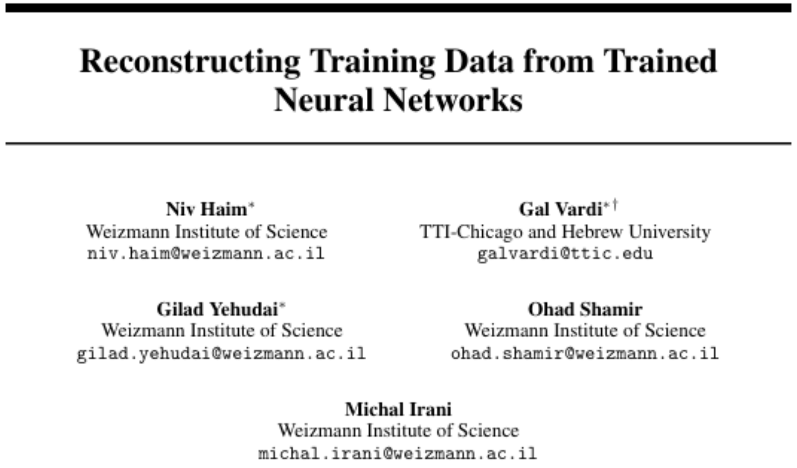
Federated Learning

Adversarial attacks and privacy-preserving techniques

SUPSI

Adversarial attacks

So, how secure is federated learning? Model inversion attack



Reconstructing Training Data from Trained Neural Networks

Niv Haim[*]
Weizmann Institute of Science
niv.haim@weizmann.ac.il

Gal Vardi[*†]
TTI-Chicago and Hebrew University
galvardi@ttic.edu

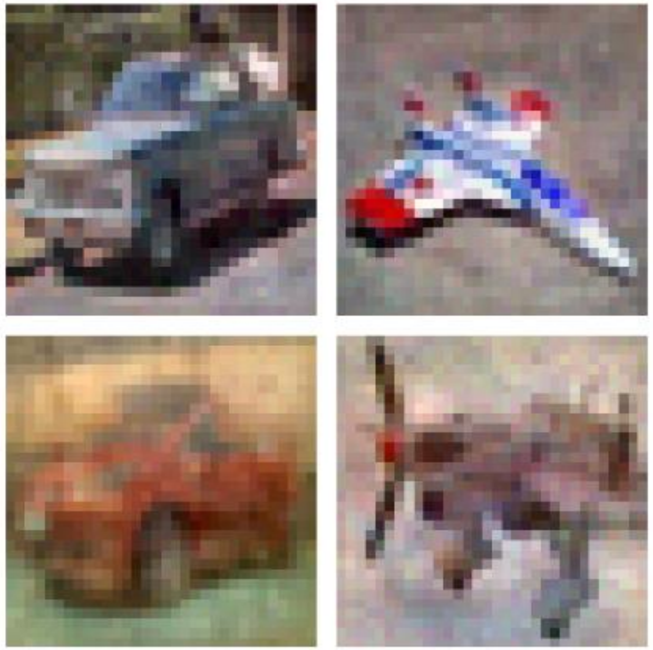
Gilad Yehudai[*]
Weizmann Institute of Science
gilad.yehudai@weizmann.ac.il

Ohad Shamir
Weizmann Institute of Science
ohad.shamir@weizmann.ac.il

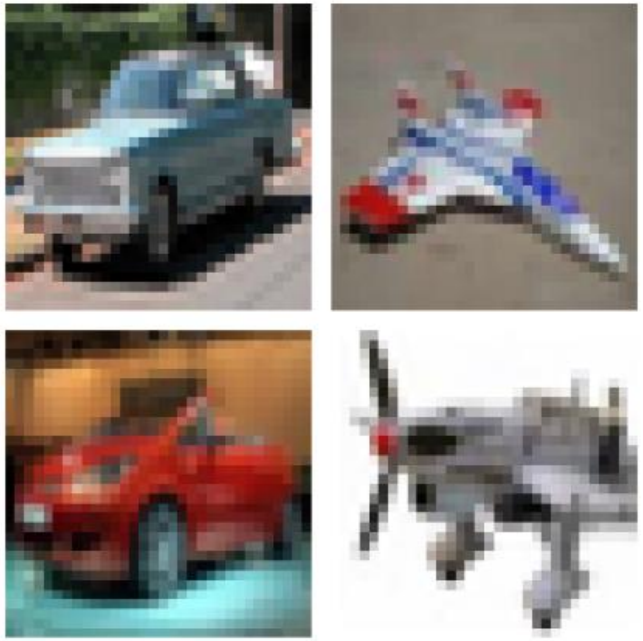
Michal Irani
Weizmann Institute of Science
michal.irani@weizmann.ac.il

(NeurIPS 2022)

Reconstructed images starting from model

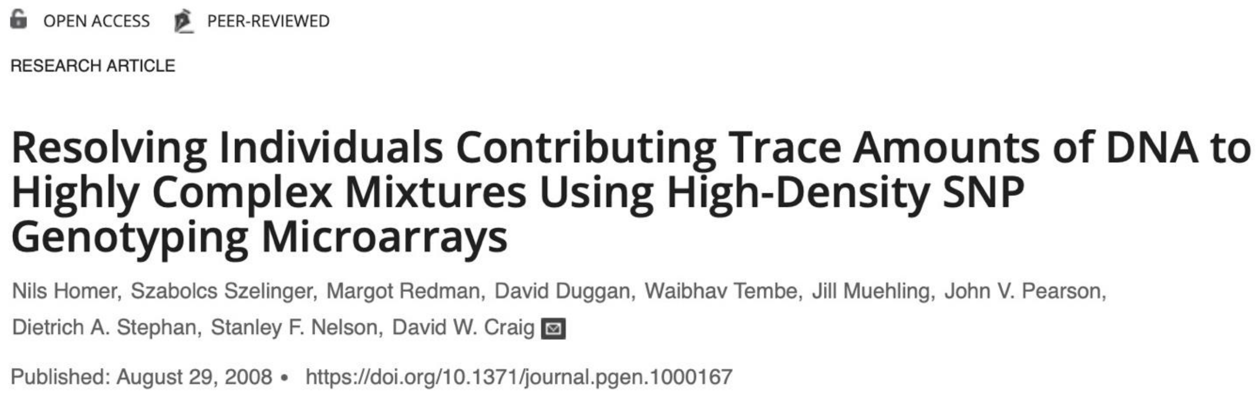


Most similar instances in real dataset



Examples from bioinformatics

- **Individual reidentification** (membership inference attack) is an issue for genetic data.
- However, such works often make **unrealistic assumptions** on the level of access to the models and the data: even basic infrastructure security measures (Kolobkov, 2024).



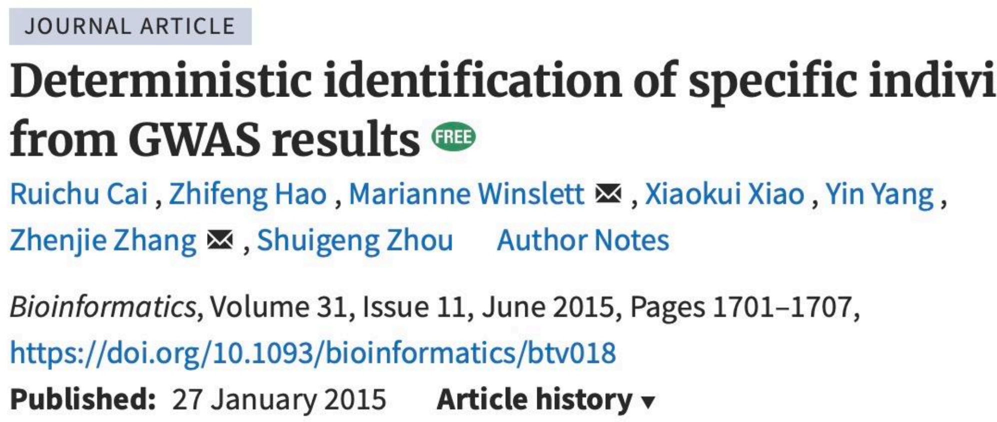
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, David W. Craig

Published: August 29, 2008 • https://doi.org/10.1371/journal.pgen.1000167



JOURNAL ARTICLE

Deterministic identification of specific individuals from GWAS results FREE

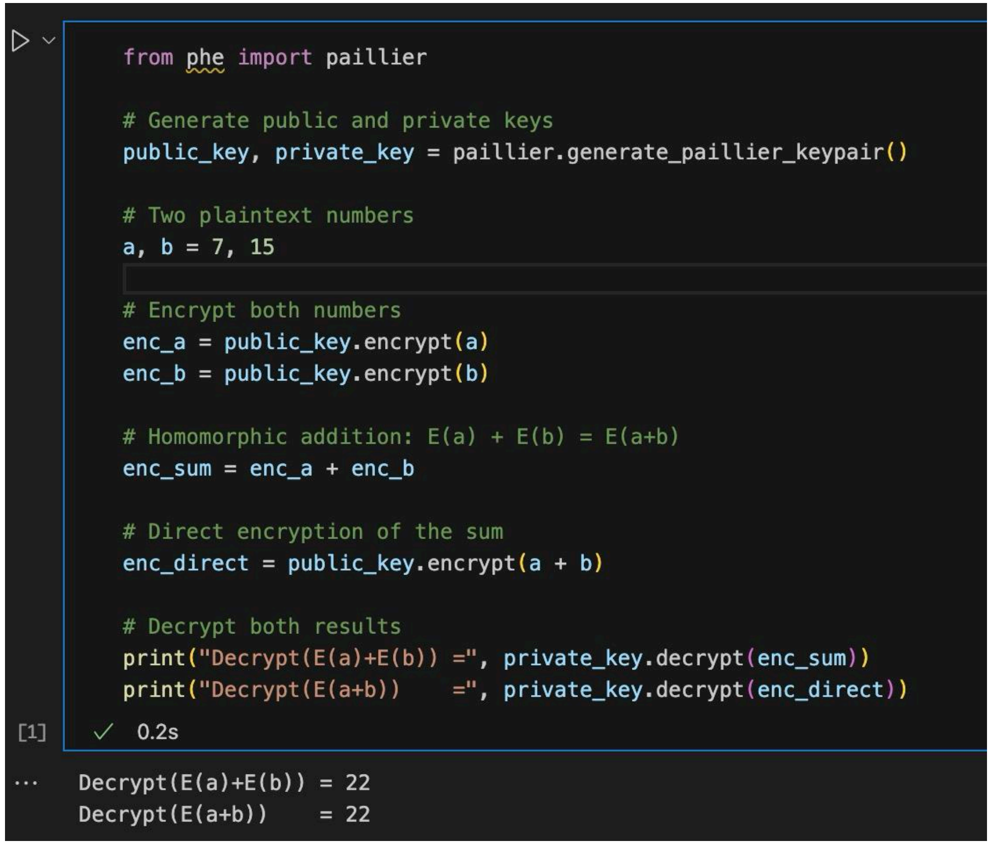
Ruichu Cai , Zhifeng Hao , Marianne Winslett , Xiaokui Xiao , Yin Yang , Zhenjie Zhang , Shuigeng Zhou Author Notes

Bioinformatics, Volume 31, Issue 11, June 2015, Pages 1701–1707, https://doi.org/10.1093/bioinformatics/btv018

Published: 27 January 2015 **Article history** ▾

Mitigation strategies for adversarial attacks: homomorphic encryption

- Homomorphic encryption lets you perform computations on encrypted data **without** decrypting it.
- The results, once decrypted, are the **same** as if the operations were done on the raw data.
- This enables secure data processing and sharing without exposing sensitive information.



```python
from phe import paillier

# Generate public and private keys
public_key, private_key = paillier.generate_paillier_keypair()

# Two plaintext numbers
a, b = 7, 15

# Encrypt both numbers
enc_a = public_key.encrypt(a)
enc_b = public_key.encrypt(b)

# Homomorphic addition: E(a) + E(b) = E(a+b)
enc_sum = enc_a + enc_b

# Direct encryption of the sum
enc_direct = public_key.encrypt(a + b)

# Decrypt both results
print("Decrypt(E(a)+E(b)) =", private_key.decrypt(enc_sum))
print("Decrypt(E(a+b))    =", private_key.decrypt(enc_direct))
```

[1] ✓ 0.2s

```
Decrypt(E(a)+E(b)) = 22
Decrypt(E(a+b))    = 22
```

Preservation of addition in Pallier scheme.

Mitigation strategies for adversarial attacks: homomorphic encryption

- Homomorphic encryption lets you perform computations on encrypted data **without** decrypting it.
- The results, once decrypted, are the **same** as if the operations were done on the raw data.
- This enables secure data processing and sharing without exposing sensitive information.

✅ **Advantages**

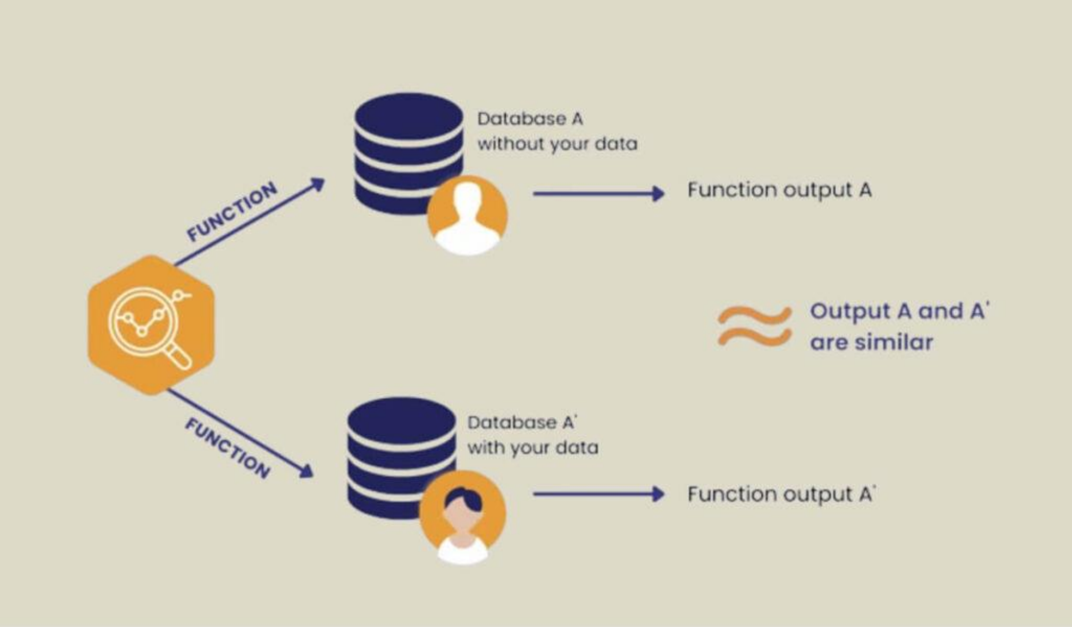
- Data stays encrypted during processing
- Provides strong privacy protection

❌ **Disadvantages**

- Much slower than normal computation (high computational cost)
- Requires more storage and bandwidth due to large ciphertexts

Mitigation strategies for adversarial attacks: differential privacy

- Differential privacy protects individuals by adding carefully calibrated random **noise** to data or query results.
- This ensures that no **single person**’s information can be distinguished with high confidence.
- It balances data utility with strong **mathematical privacy guarantees**.



https://pvml.com/glossary/differential-privacy/

Mitigation strategies for adversarial attacks: differential privacy

- Differential privacy protects individuals by adding carefully calibrated random **noise** to data or query results.
- This ensures that no **single person**’s information can be distinguished with high confidence.
- It balances data utility with strong **mathematical privacy guarantees**.

✅ **Advantages**

- Provides formal, provable privacy guarantees
- Scales well for large datasets

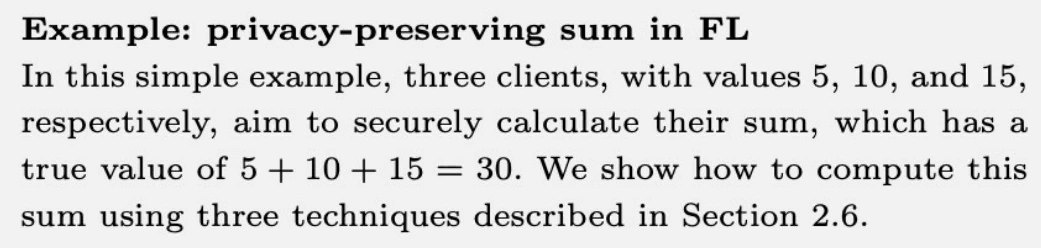
❌ **Disadvantages**

- Noise can reduce accuracy of results
- Choosing the privacy budget (*epsilon*) is difficult
- Less effective for small datasets

SMPC

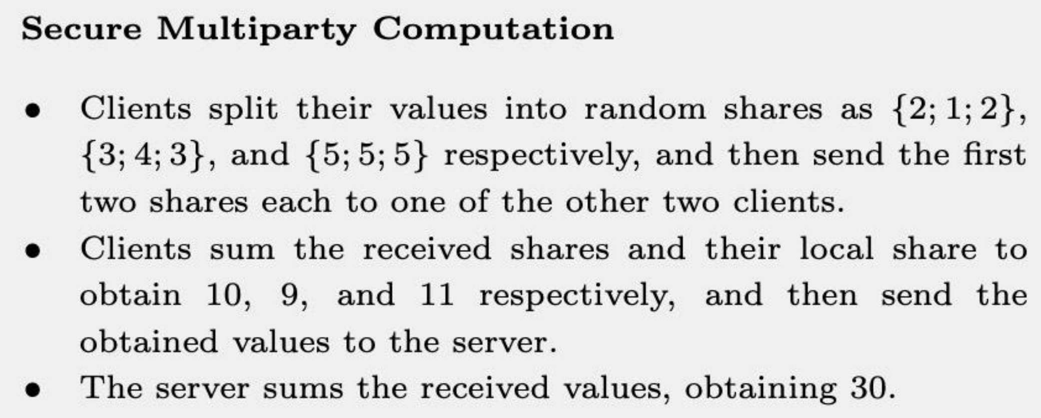
Mitigation strategies for adversarial attacks: secure multiparty computation (SMPC)

- SMPC allows multiple parties to **jointly** compute a result without revealing their private inputs.
- Each participant only learns the **final output**, not others’ data.



Example: privacy-preserving sum in FL

In this simple example, three clients, with values 5, 10, and 15, respectively, aim to securely calculate their sum, which has a true value of $5 + 10 + 15 = 30$. We show how to compute this sum using three techniques described in Section 2.6.



Secure Multiparty Computation

- Clients split their values into random shares as $\{2; 1; 2\}$, $\{3; 4; 3\}$, and $\{5; 5; 5\}$ respectively, and then send the first two shares each to one of the other two clients.
- Clients sum the received shares and their local share to obtain 10, 9, and 11 respectively, and then send the obtained values to the server.
- The server sums the received values, obtaining 30.

Mitigation strategies for adversarial attacks: secure multiparty computation (SMPC)

- SMPC allows multiple parties to **jointly** compute a result without revealing their private inputs.
- Each participant only learns the **final output**, not others’ data.

✅ **Advantages**

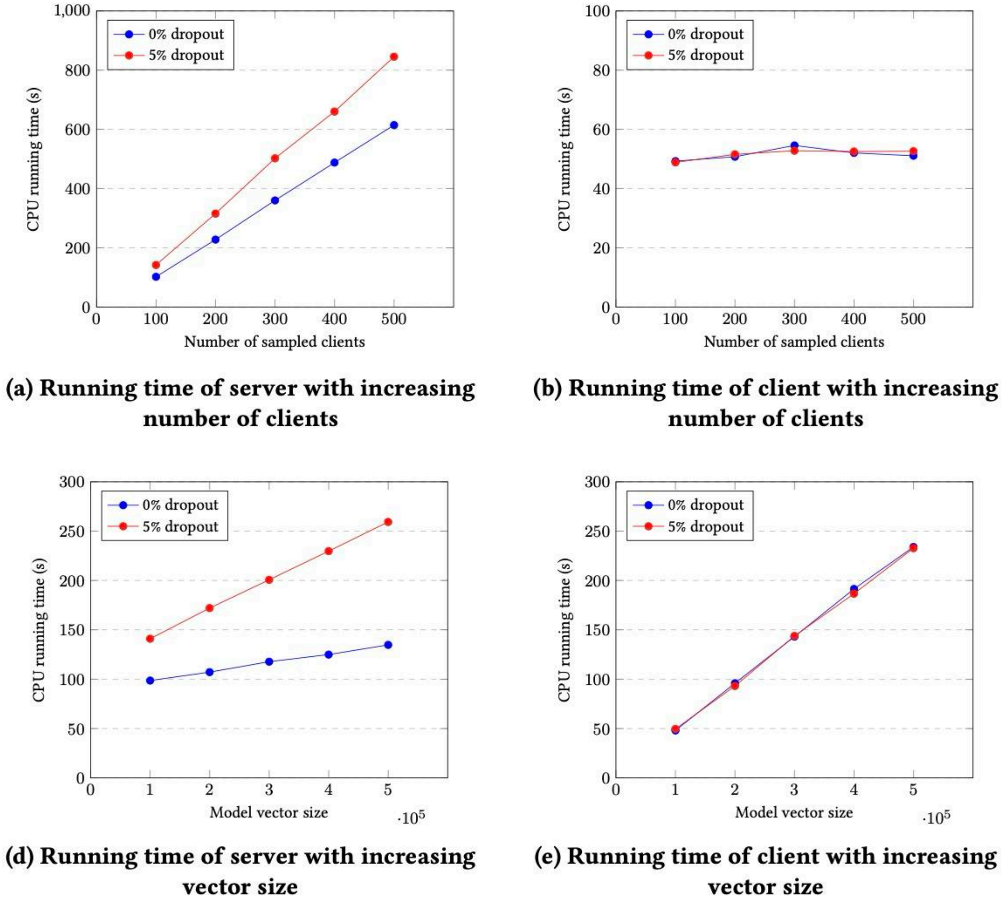
- Strong privacy: inputs remain secret throughout the computation
- Cryptographic guarantees of correctness and confidentiality

❌ **Disadvantages**

- Computationally and communication intensive (slower than normal computing)
- Protocols can be complex to design and implement securely
- Scalability challenges

SecAgg+ (provided by Flower as a workflow!)

- **Efficient graph-based mask exchange:** clients only need to communicate with a small subset of peers.
- **Robustness to dropouts:** the protocol is designed to handle client dropouts without redoing heavy computations.



(a) Running time of server with increasing number of clients

(b) Running time of client with increasing number of clients

(d) Running time of server with increasing vector size

(e) Running time of client with increasing vector size

Federated Learning

Legal considerations

Key points on legal considerations

Data protection still required: FL is legally treated like regular data sharing under GDPR, there are no specific regulations (yet).

Consortium agreements: Parties must sign Data Protection Agreements defining roles, responsibilities, interactions, ethics approvals, and handling of privacy breaches.

Intellectual property & third parties: Joint control requires agreements on IP rights, patents, and possible financial gains.

Patient rights: In several countries, patients need to able to withdraw consent at any time, meaning technical/legal procedures must exist to remove their data from future analyses.

So why going for FL?

Technology is always faster than legislation, **specific legislation** is expected to be developed in the coming years.

FL consortia can make partners and patients **more comfortable** contributing to federated studies.

FL provides increased protection against data and model leaks and should reduce the parties' and patients' perceived risk in contributing to the consortium.

Dalle Molle Institute for Artificial Intelligence USI – SUPSI

Swiss Institute of Bioinformatics

Federated Learning

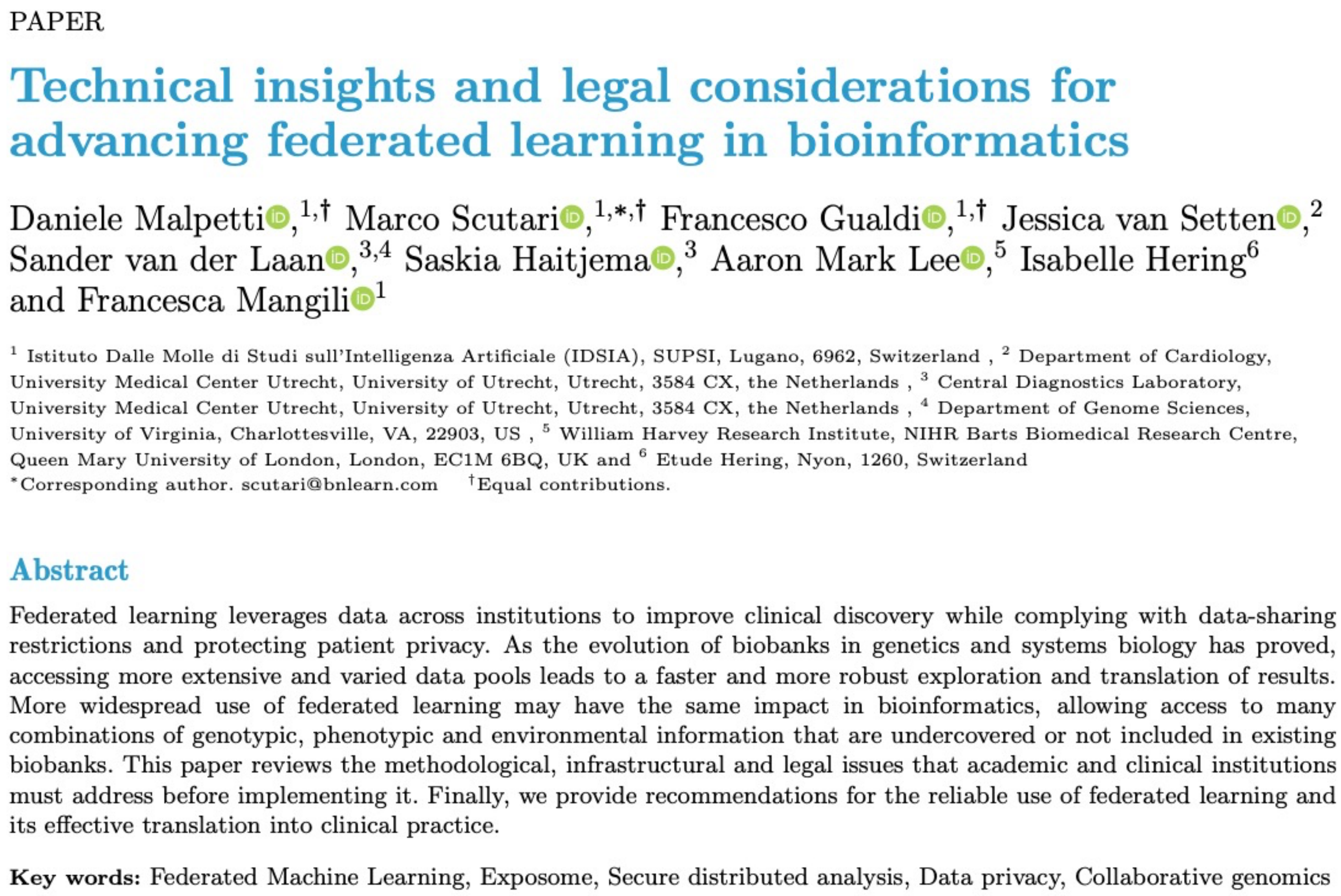
Resources

SUPSI

Dalle Molle Institute for
Artificial Intelligence
USI – SUPSI

Swiss Institute of
Bioinformatics

The main source for this presentation



PAPER

Technical insights and legal considerations for advancing federated learning in bioinformatics

Daniele Malpetti,[1,†] Marco Scutari,[1,*,†] Francesco Gualdi,[1,†] Jessica van Setten,[2] Sander van der Laan,[3,4] Saskia Haitjema,[3] Aaron Mark Lee,[5] Isabelle Hering[6] and Francesca Mangili[1]

[1] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), SUPSI, Lugano, 6962, Switzerland , [2] Department of Cardiology, University Medical Center Utrecht, University of Utrecht, Utrecht, 3584 CX, the Netherlands , [3] Central Diagnostics Laboratory, University Medical Center Utrecht, University of Utrecht, Utrecht, 3584 CX, the Netherlands , [4] Department of Genome Sciences, University of Virginia, Charlottesville, VA, 22903, US , [5] William Harvey Research Institute, NIHR Barts Biomedical Research Centre, Queen Mary University of London, London, EC1M 6BQ, UK and [6] Etude Hering, Nyon, 1260, Switzerland

[*]Corresponding author. scutari@bnlearn.com [†]Equal contributions.

Abstract

Federated learning leverages data across institutions to improve clinical discovery while complying with data-sharing restrictions and protecting patient privacy. As the evolution of biobanks in genetics and systems biology has proved, accessing more extensive and varied data pools leads to a faster and more robust exploration and translation of results. More widespread use of federated learning may have the same impact in bioinformatics, allowing access to many combinations of genotypic, phenotypic and environmental information that are undercovered or not included in existing biobanks. This paper reviews the methodological, infrastructural and legal issues that academic and clinical institutions must address before implementing it. Finally, we provide recommendations for the reliable use of federated learning and its effective translation into clinical practice.

Key words: Federated Machine Learning, Exposome, Secure distributed analysis, Data privacy, Collaborative genomics

https://arxiv.org/abs/2503.09649

https://github.com/IDSIA/FL-Bioinformatics

SUPSI

Currently under review for publication.

Questions?