Vol. 1, No. 3, pp.: 1-4, 2023

The Quiet Revolution: How Federated Learning Is Reshaping Privacy-Preserving AI

Soren Madsen[1,*], **Fatima Al-Zahrawi**[2]

[1] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark, 2800

[2] Division of Computer Science, New York University Abu Dhabi, Abu Dhabi, UAE, 129188

[*] *sorm@dtu.dk*

Article Information	
Received	18 April 2023
Accepted	29 September 2023
DOI	https://doi.org/10.63646/datamind.2023.010301

Abstract

Federated learning emerged from a 2017 Google paper as an engineering solution to a narrow problem: how do you improve a mobile keyboard's next-word prediction without uploading users' typing data to a central server? Seven years later, it has become a genuine paradigm shift in how the AI research community thinks about the relationship between data, privacy, and model training. This perspective piece argues that federated learning's deepest contribution is not technical but conceptual — it has changed what counts as acceptable in AI system design and has legitimised a set of questions about data sovereignty that were previously considered beyond the scope of machine learning research. We trace this shift through three domains where federated learning has had disproportionate impact: healthcare, mobile devices, and financial services. We also examine the significant technical limitations that remain unresolved — statistical heterogeneity, communication efficiency, and Byzantine robustness — and argue that addressing them will require the field to engage more seriously with systems research than it currently does. The piece concludes with a reflection on what 'privacy-preserving AI' actually means and why the term is frequently misused in ways that obscure more than they reveal.

Keywords: ***federated learning; privacy-preserving AI; data sovereignty; statistical heterogeneity; differential privacy; healthcare AI; distributed learning***

1. The Origin Story

The origin story of federated learning is often told as a clean technical narrative: Google engineers face a data privacy constraint, invent a distributed training protocol, publish a foundational paper, and the field takes off. The reality is messier and more interesting. The 2017 McMahan et al. paper on FedAvg was itself drawing on decades of distributed optimisation research, and the 'privacy' framing was less central to the initial publication than is sometimes remembered. The privacy emphasis grew as the technique was adopted, not before.

This matters because it illustrates how federated learning's evolution has been shaped less by a priori design principles and more by a continuous negotiation between engineering constraints, regulatory pressures, and application demands. The GDPR came into force in 2018, just as federated learning was attracting serious research attention. The European AI Act's provisions on high-risk AI systems have created additional incentives for healthcare and financial services firms to explore

privacy-preserving training approaches. These regulatory dynamics have done as much to shape the field's trajectory as any algorithmic innovation.

2. What Federated Learning Actually Does (and Does Not Do)

There is a persistent confusion in popular accounts of federated learning between what the technique actually guarantees and what it is often assumed to guarantee. Federated learning, in its basic form, keeps raw data on local devices or within local institutions. It does not prevent information leakage through gradients — a line of attack known as gradient inversion, demonstrated with increasing effectiveness since 2020. It does not prevent membership inference attacks. And it does not, by itself, provide formal privacy guarantees in the differential privacy sense.

The combination of federated learning with differential privacy (FL+DP) is considerably more principled from a formal privacy standpoint, but it comes with real accuracy costs that are frequently understated in research papers. The privacy-utility tradeoff at strong epsilon values ($\varepsilon < 1$) is often severe enough to make the resulting model practically unusable for medical or financial decision-making. This is not a reason to dismiss federated learning — it is a reason to be precise about what problem it is and is not solving.

[Figure 1 — Schematic of the federated learning training loop showing three rounds of communication between a central aggregation server and four local clients. Gradient updates are shown in orange; model distributions in blue. The differential privacy noise injection step (not shown in basic FL) would occur at the local gradient computation stage before upload.]

Figure 1. *Schematic of the federated learning training loop showing three rounds of communication between a central aggregation server and four local clients. Gradient updates are shown in orange; model distributions in blue. The differential privacy noise injection step (not shown in basic FL) would occur at the local gradient computation stage before upload.*

3. Impact Domains

3.1 Healthcare

Healthcare is the application domain where federated learning's impact has been most consequential. The combination of highly sensitive personal data, strong regulatory restrictions on data sharing, and genuine clinical value in training on large populations creates an almost perfect environment for the technique. The FeTS Challenge — a federated tumour segmentation challenge involving 71 geographically distributed healthcare institutions — demonstrated in 2022 that federated models could match centrally trained baselines with appropriate federated optimisation strategies. The implications for rare disease research, where no single institution holds enough data to train a capable model, are particularly significant.

3.2 Financial Services

Financial institutions operate under data-sharing restrictions that are driven by competitive concerns as much as regulatory ones: a bank will not share its full transaction history with a competitor simply to co-train a fraud detection model, even if both would benefit. Federated learning offers a partial solution to this coordination failure. Several commercial deployments now exist — SWIFT's fraud detection consortium being the most widely cited — though detailed technical specifications remain proprietary.

4. Unresolved Technical Challenges

The statistical heterogeneity problem — the fact that data distributions across federated clients are often non-IID in ways that cause standard averaging-based aggregation to converge poorly — has attracted intense research attention since 2020. FedProx, SCAFFOLD, FedNova, and a dozen other variants of FedAvg have proposed different approaches to the non-IID problem, and all show improvements under specific heterogeneity regimes. What remains poorly understood is which algorithm to choose for a given deployment, since the heterogeneity structure of real federated datasets is often unknown in advance.

Byzantine robustness — the question of what happens when some federated clients submit malicious or corrupted gradient updates — is a problem that feels theoretical but is increasingly practical as federated systems expand. Multi-Krum, RobustAgg, and Flame are among the defensive mechanisms proposed, but each comes with computational costs that scale poorly. The honest assessment is that federated learning in adversarial environments remains an open research problem.

5. What Privacy-Preserving Actually Means

We want to end with a conceptual point that we think is underemphasised in the technical literature. 'Privacy-preserving AI' is often used as a binary label — a system either is or is not privacy-preserving. This framing obscures the fundamental reality that privacy protection is a spectrum, its adequacy is context-dependent, and its evaluation requires engagement with threat models that machine learning researchers are not typically trained to construct.

Federated learning represents a genuine improvement in the privacy properties of AI training relative to centralised data collection. It is not a complete solution to the privacy problem, and positioning it as such creates a false sense of security that may actually set back the cause of privacy-respecting AI development. The field would benefit from more papers that quantify privacy guarantees in formal terms, and from more honest acknowledgment of the limits of those guarantees.

6. Conclusion

Federated learning's most lasting contribution may be that it made a set of questions about data sovereignty, institutional data sharing, and privacy-utility tradeoffs legible to the machine learning research community. The technical problems it has surfaced — non-IID convergence, communication efficiency, Byzantine robustness, formal privacy guarantees — are hard, interesting, and important. The conceptual shift it has enabled — toward thinking of training data as something that can remain distributed and still useful — may prove even more important than any specific algorithmic contribution. The revolution is quiet, ongoing, and not yet finished.

References

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS 2017, 1273–1282. http://proceedings.mlr.press/v54/mcmahan17a.html

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210. https://doi.org/10.1561/2200000083

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

https://doi.org/10.1109/MSP.2020.2975749

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. arXiv preprint arXiv:1806.00582. https://doi.org/10.48550/arXiv.1806.00582

Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in Neural Information Processing Systems, 32. https://proceedings.neurips.cc/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10, 12598. https://doi.org/10.1038/s41598-020-69250-1

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. CCS 2016, 308–318. https://doi.org/10.1145/2976749.2978318

Li, D., & Wang, J. (2020). FedMD: Heterogeneous federated learning via model distillation. arXiv preprint arXiv:1910.03581. https://doi.org/10.48550/arXiv.1910.03581

Peng, X., Huang, Z., Zhu, Y., & Saenko, K. (2019). Federated adversarial domain adaptation. ICLR 2020. https://openreview.net/forum?id=HJezF3VYPB

Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019). A hybrid approach to privacy-preserving federated learning. AISec@CCS 2019, 1–11. https://doi.org/10.1145/3338501.3357370