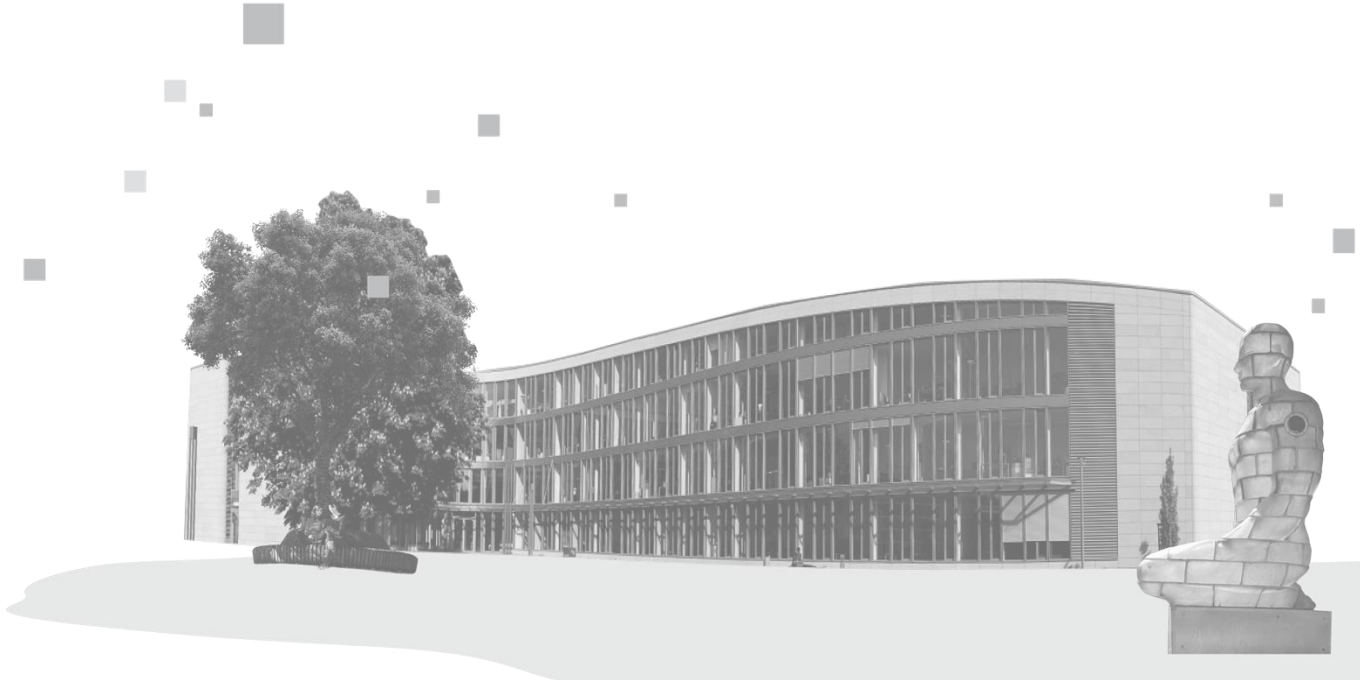
Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Federated Learning – Privacy by Design

”Your data never leaves your device.”

Pierre Burghardt



Design IT.
Create Knowledge.

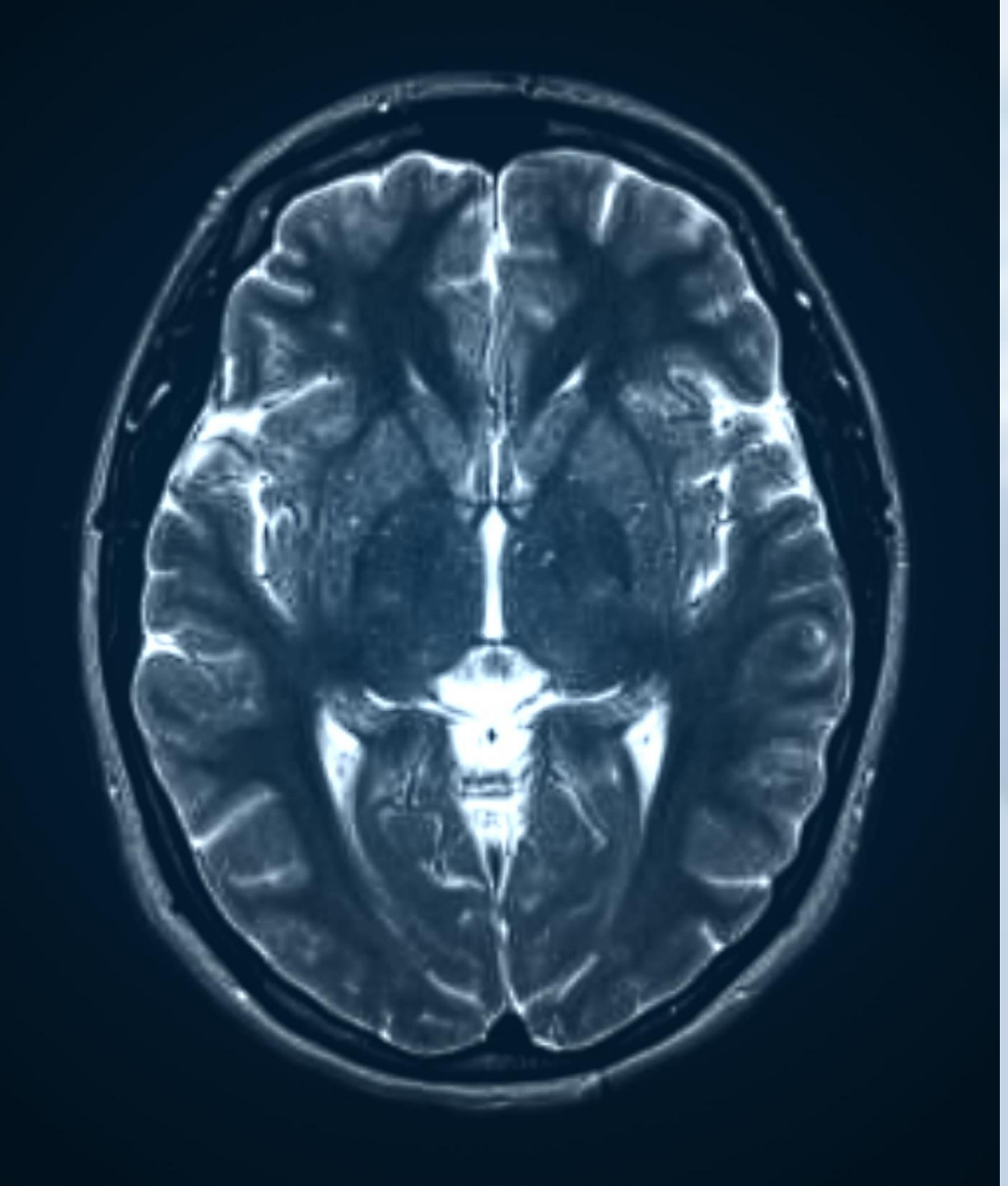
www.hpi.de

Imagine…

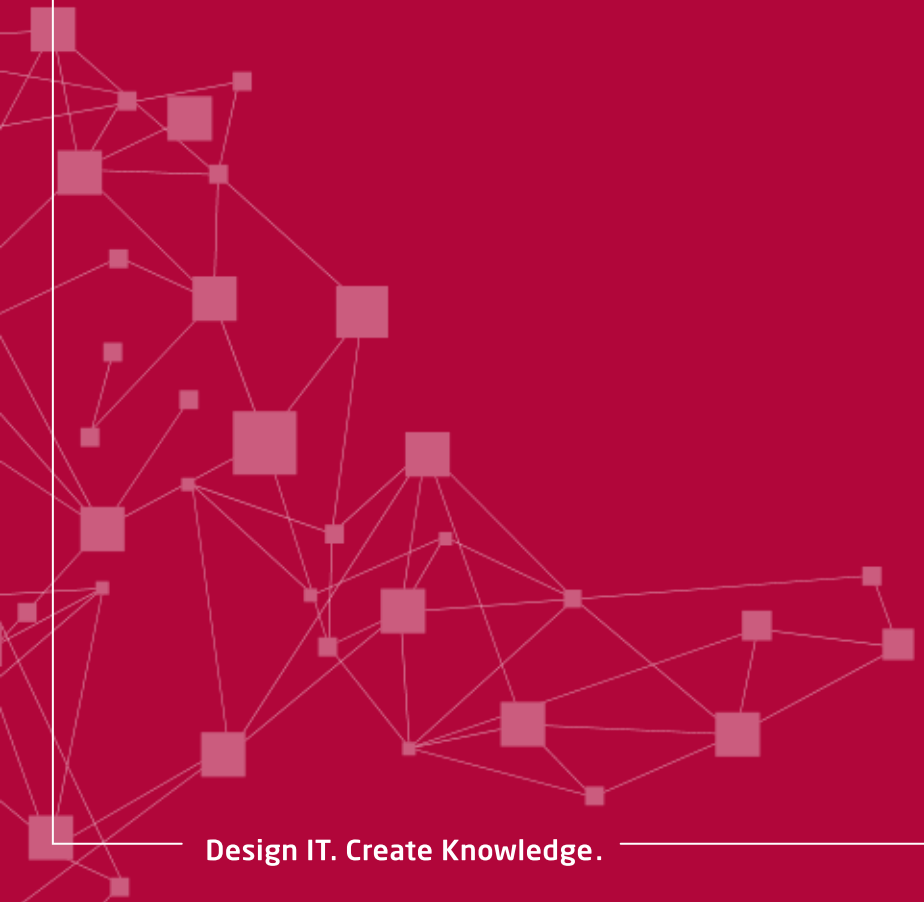
… new rare brain tumor.

… early detectable via brain MRI.

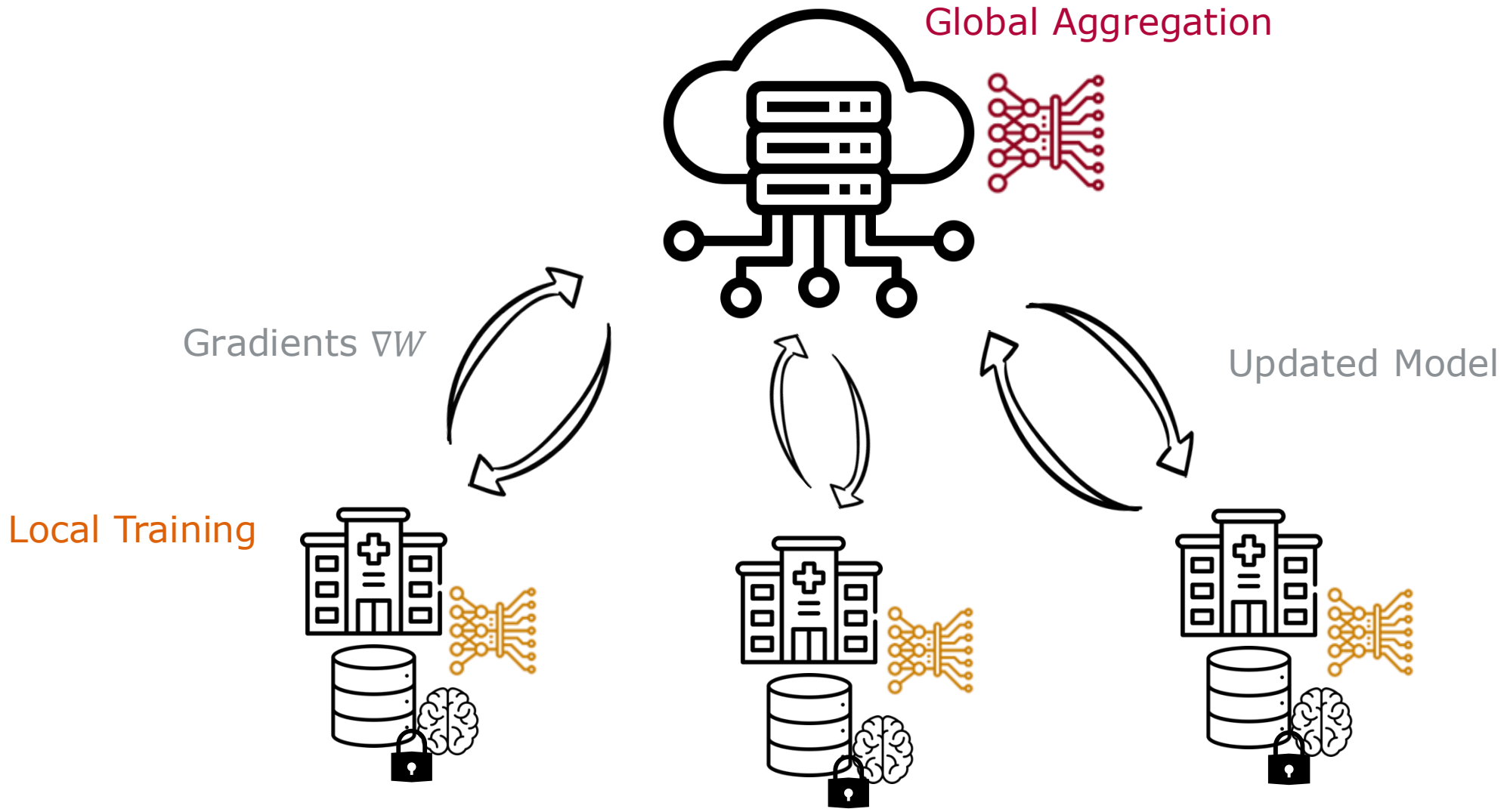
… a trained AI model could save Millions of lives.



Would you trust a centralized AI Server to use your data?



Federated Learning – Your data never leaves your device



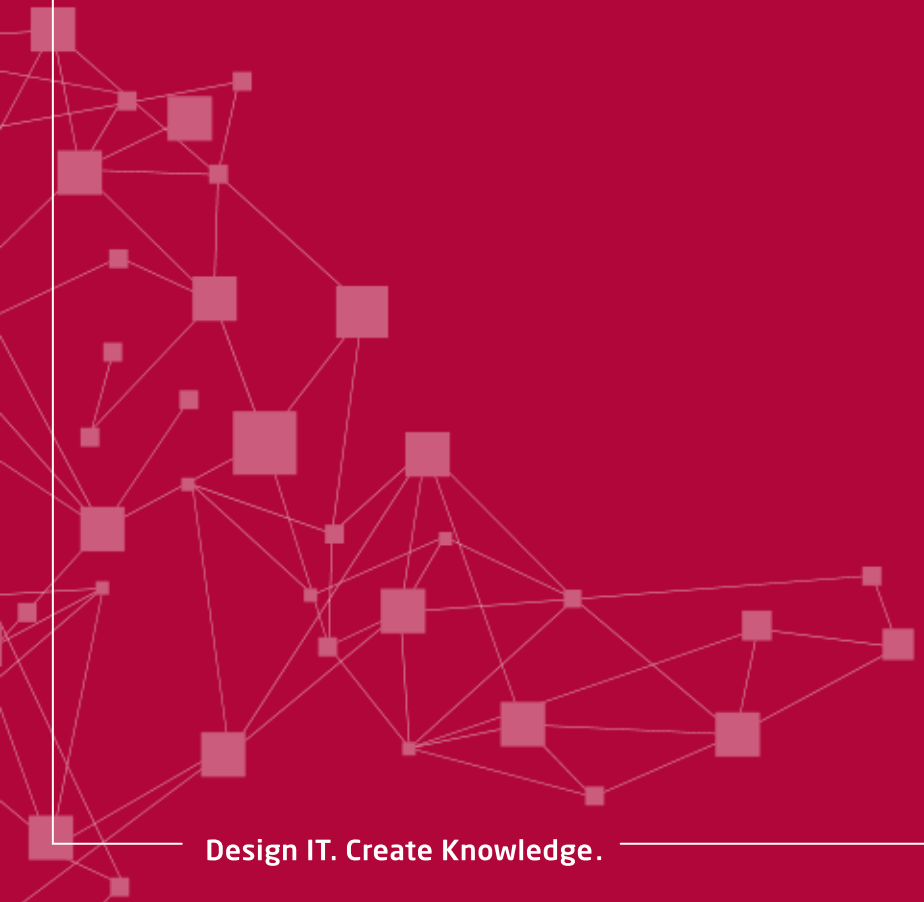
[1],[2]

Federated Learning – Your data never leaves your device

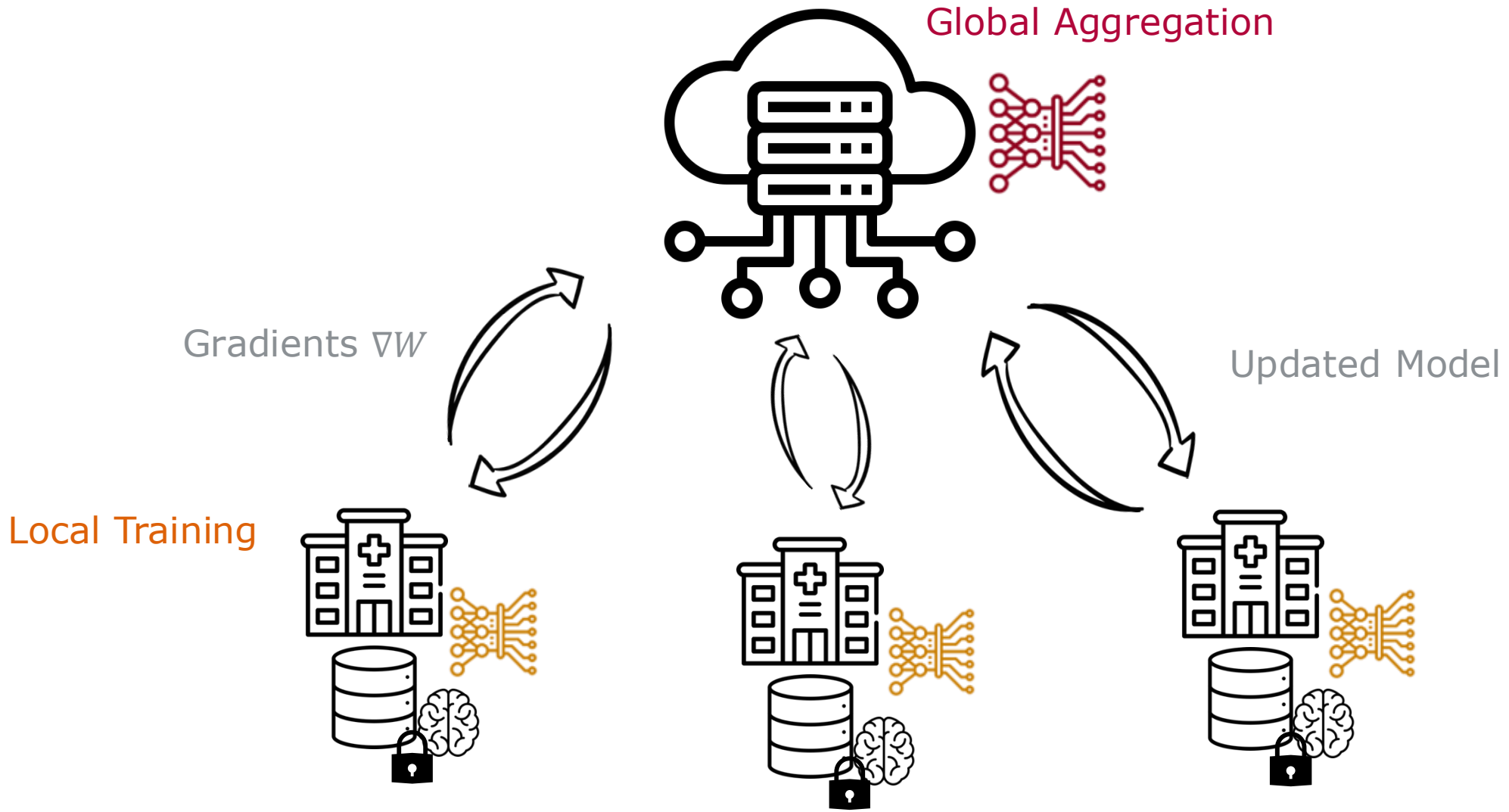
Privacy Collaboration Data Minimization Scalability

[1],[2]

Would you consider this system private?



Federated Learning – Your data never leaves your device



TLS

[1],[2]

A Threat Within The System

Honest-But-Curious-Server

Aggregates the updates exactly as instructed, but silently logs every incoming gradient to see what it can extract

Malicious Client

Actively injects malicious gradients, alters the global model to force clients to leak their data

[1]

Under the Hood: Gradient Updates

- What is a gradient?
  - Multi dimensional vector matrix that represents the directions and rate of change of parameters/ weights during training (here at step t, client i):

$$\nabla W_{t,i} = \frac{\partial \ell(F(\mathbf{x}_{t,i}, W_t), \mathbf{y}_{t,i})}{\partial W_t}$$

- Traditional assumption:

  Gradients are abstract derivatives representing highly compressed information

[3]

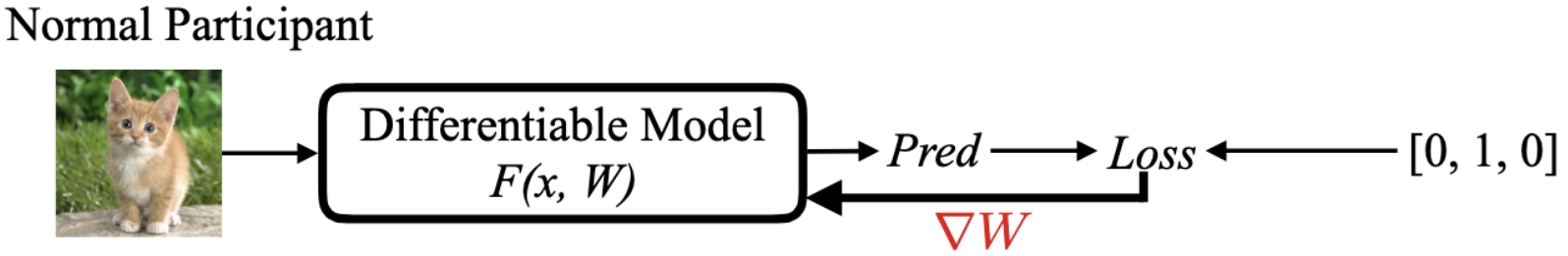
Can an update be reversed? The Curious Server (DLG Attack)

- *"We show that it is possible to obtain private training data from publicly shared gradients... we name this type of privacy breach as gradient deep leakage."* (Zhu et al., 2019) [3]

- **Deep Leakage from Gradients (Zhu et al. 2019) [3]**
  - It became possible.
- **iDLG (Zhao et al. 2020) [4]**
  - It became better / more reliable.
- **Inverting Gradients (Geiping et al. 2020) [5]**
  - It became a threat.

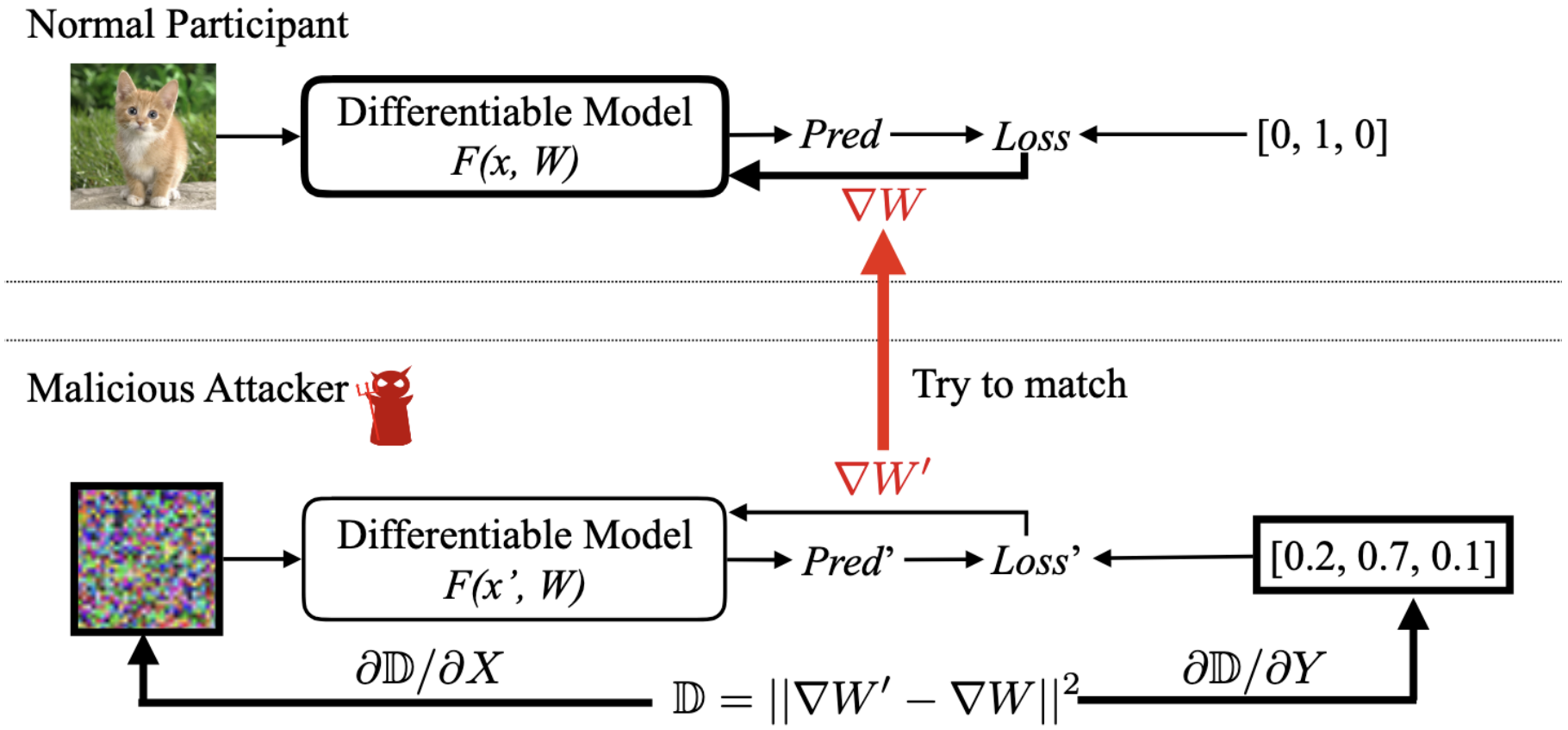
Deep Leakage from Gradients

Zhu et al. 2019 [3]



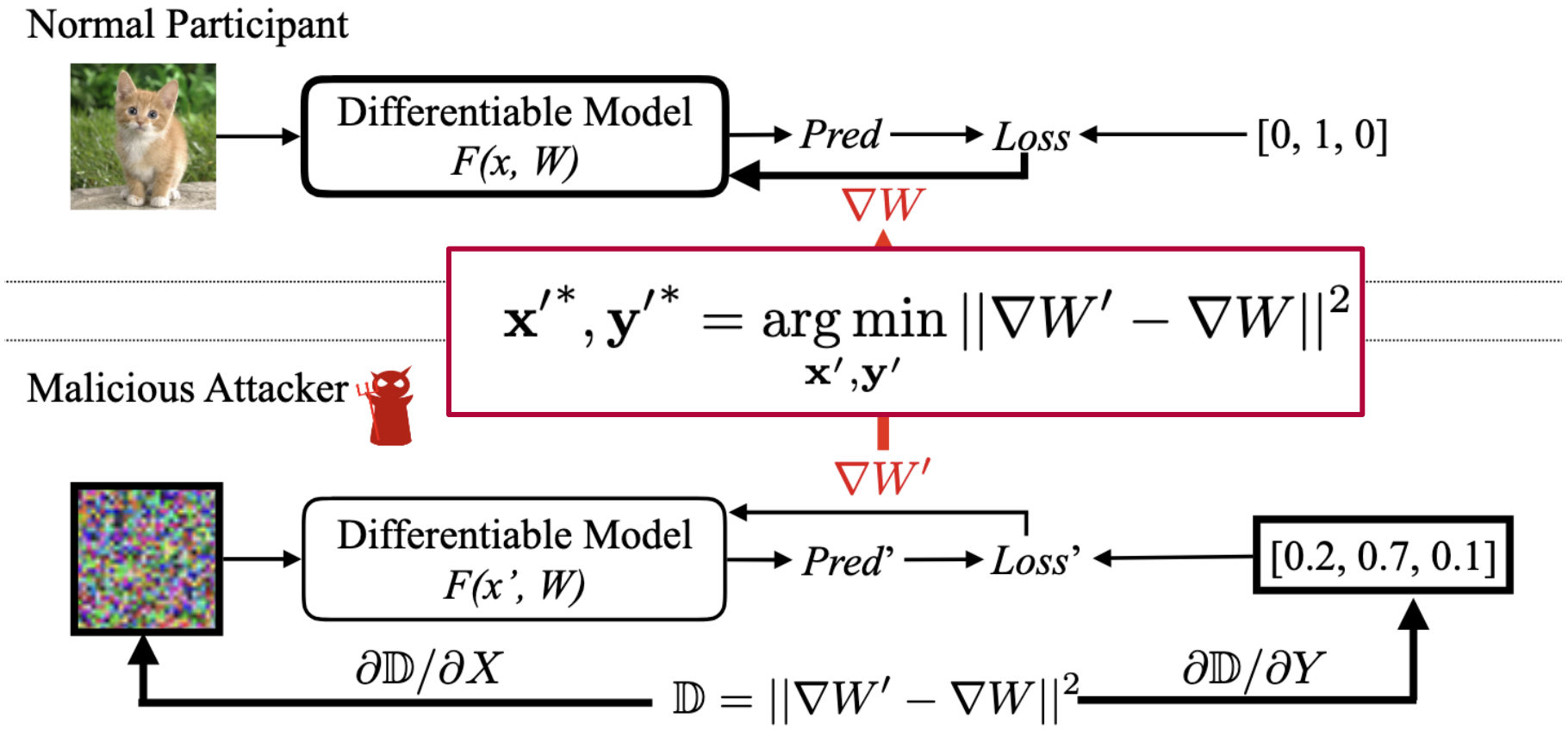
Deep Leakage from Gradients

Zhu et al. 2019 [3]



Deep Leakage from Gradients

Zhu et al. 2019 [3]



Deep Leakage from Gradients

Zhu et al. 2019 [3]

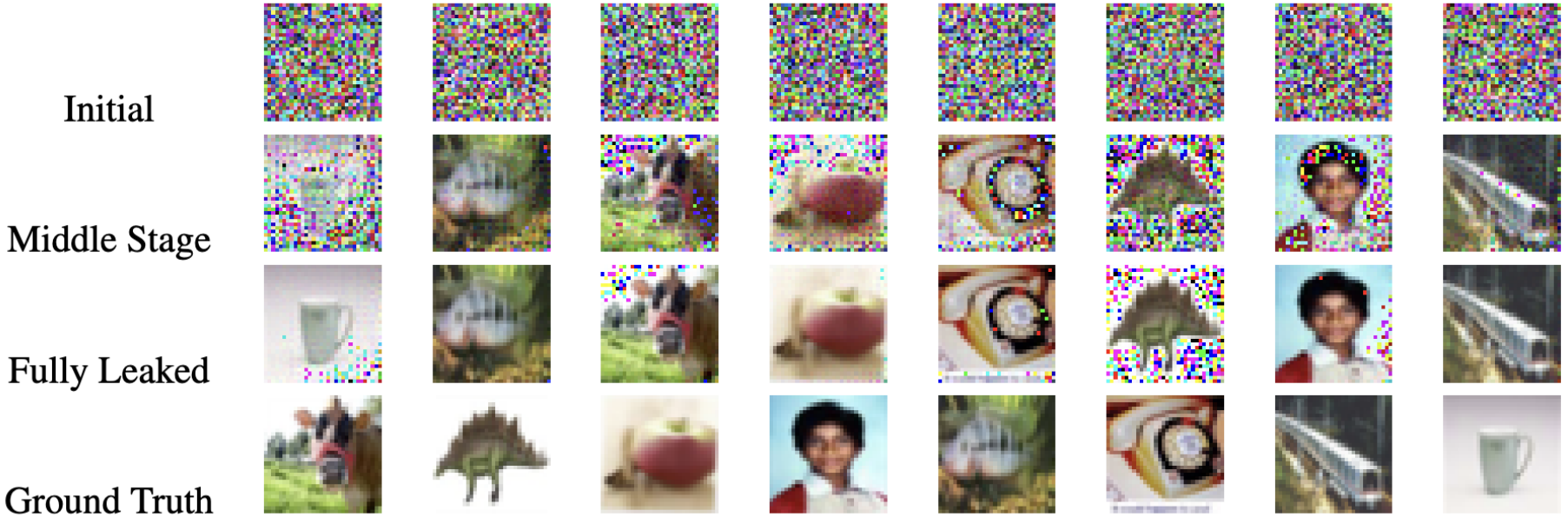


Figure 4: Results of deep leakage of batched data. Though the order may not be the same and there are more artifact pixels, DLG still produces images very close to the original ones.

- Batch size up to 8, image resolution up to 64x64, unreliable convergence [4]

Improved Deep Leakage from Gradients (iDLG)

Zhao et al. 2020 [4]

- You don’t need to guess the label! (Softmax, Cross Entropy Loss)

$$\nabla W_{i,L} = (p_i - y_i) \cdot a_{L-1}$$

L: layer count
a_{L-1}: Sigmoid or ReLU activation function
p_i: predicted probability
y_i: one-hot label

→ $(p_i - y_i)$ only negative if i is the true class index

"Hence, we can simply identify the ground-truth label whose corresponding $\nabla W_{i,L}$ is negative."

Improved Deep Leakage from Gradients (iDLG)

Zhao et al. 2020 [4]

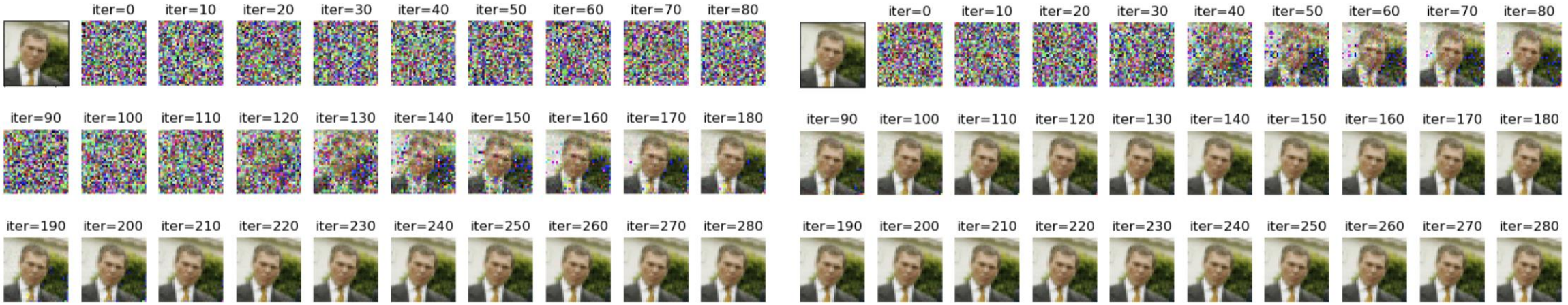


Figure 2: Example of the training process of DLG (left) and iDLG (right) on LFW face dataset. The first image is the (original) private training image, and the followings are the extracted images at different training iterations. It is clear that the training of iDLG is easier to converge.

- Quicker, more reliable
- Needs gradients per sample, no batch

Inverting Gradients

Geiping et al. 2020 [5]

- Further improvement (batch sizes of up to 100, higher-resolution images, trained networks)

$$\arg\min_{x' \in [0,1]^n} \left(1 - \frac{\langle \nabla W', \nabla W \rangle}{\|\nabla W'\| \|\nabla W\|}\right) + \alpha \mathrm{TV}(x')$$

→ Minimize a magnitude-invariant loss based on Cosine Similarity + a Total Variation Regularizer

$$TV(x) = \sum_{i,j} |x_{i+1,j} - x_{i,j}| + |x_{i,j+1} - x_{i,j}|$$

→ Produce natural image rather than random static; smooth regions with only few edges

Inverting Gradients

Geiping et al. 2020 [5]

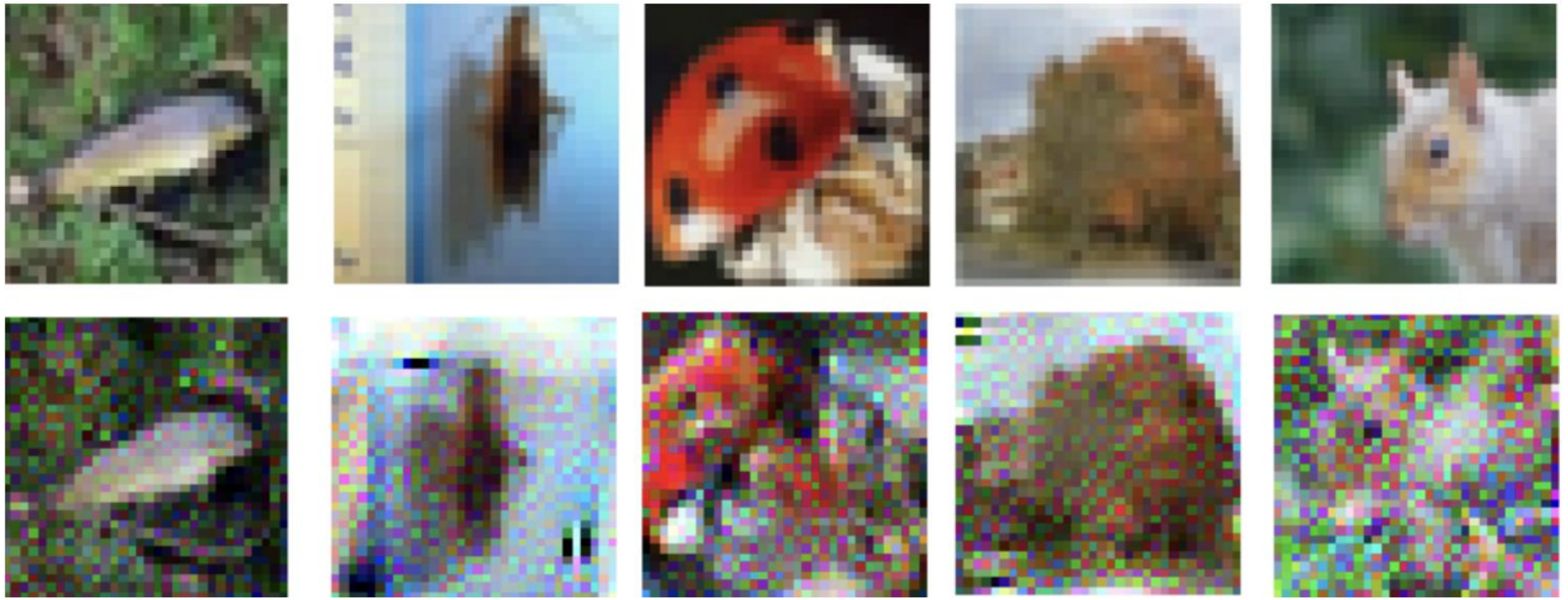
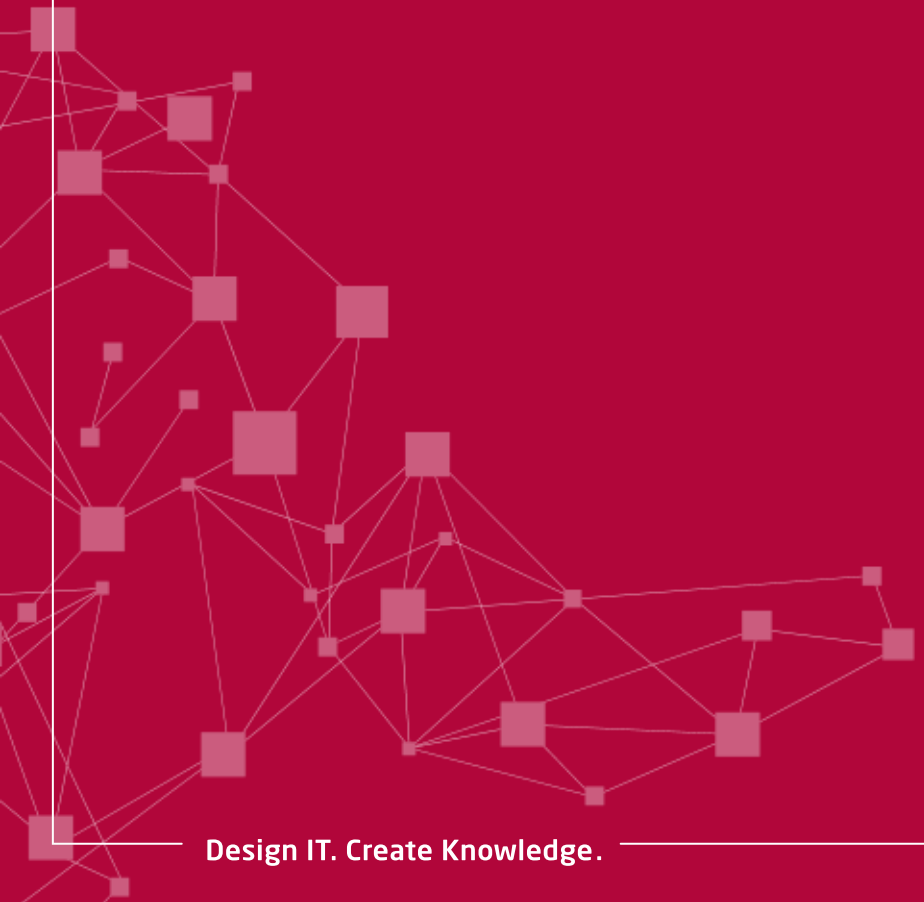


Figure 6: Information leakage from the aggregated gradient of a batch of 100 images on CIFAR-100 for a ResNet32-10. Shown are the 5 *most* recognizable images from the whole batch. Although most images are unrecognizable, privacy is broken even in a large-batch setting. We refer to the supplementary material for all images.

Counter Measures?

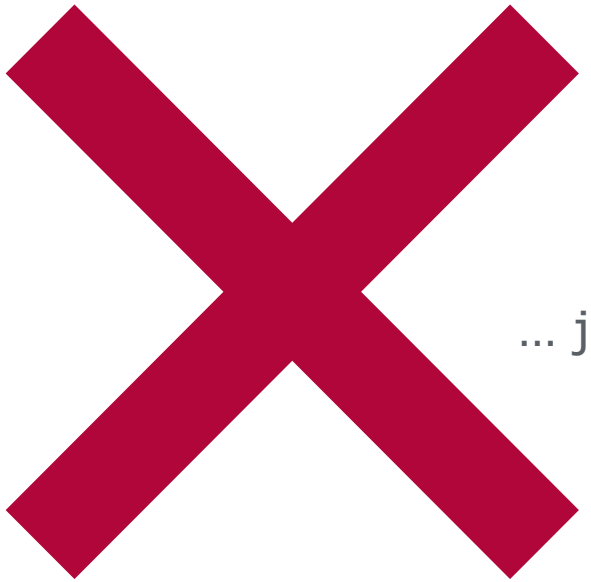


Heuristic Measures

Large Batch Size

Gradient Pruning

Dropout



… just speedbumps or useless training

[1],[3], [4],[5]

Secure Aggregation

- Goal: Let server never see individual updates

Zero Sum Masking (Noise)

Clients add structured, zero-sum random noise to their gradients

- communication setup

Homomorphic Encryption (HE)

Allows mathematical operations directly on encrypted data

- High computation and communication overhead

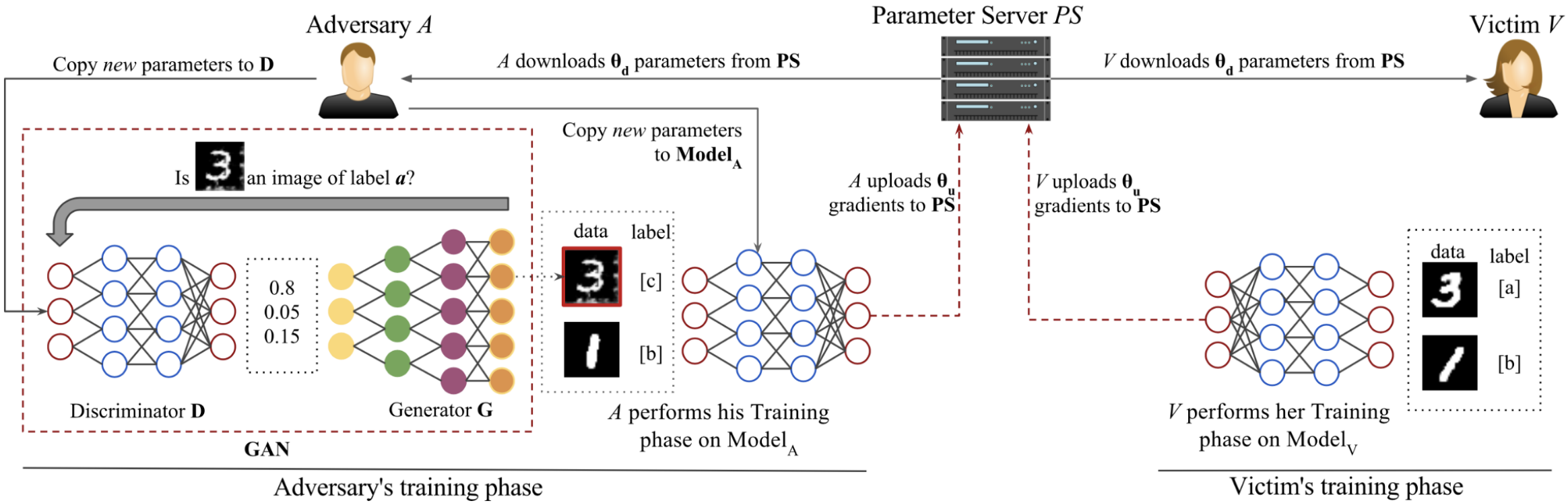
Trusted Execution Environment (TEE)

Client data is decrypted and aggregated or shuffled in a secure, hardware-isolated enclave

[1],[7]

GAN-based leakage (Model Inversion +)

Hitaj et al. 2017 [6]



Differential Privacy (DP)

- Goal: Introduce uncertainty, mask the contribution of the individual users

- Provides mathematical guarantees about the privacy of the data

Given D and D’ two adjacent datasets, a randomized Algorithm A wants to approach A(D) = A(D’), so make it similar

Specifically, their probability distributions should be related by ϵ

Record Level DP

Client Level DP

$$\frac{P(A(D))}{P(A(D'))} \leq e^{\epsilon}$$

Central DP

Local DP

ϵ of 0 is absolute privacy, big ϵ relaxes mathematical guarantee

[1],[7],[8]

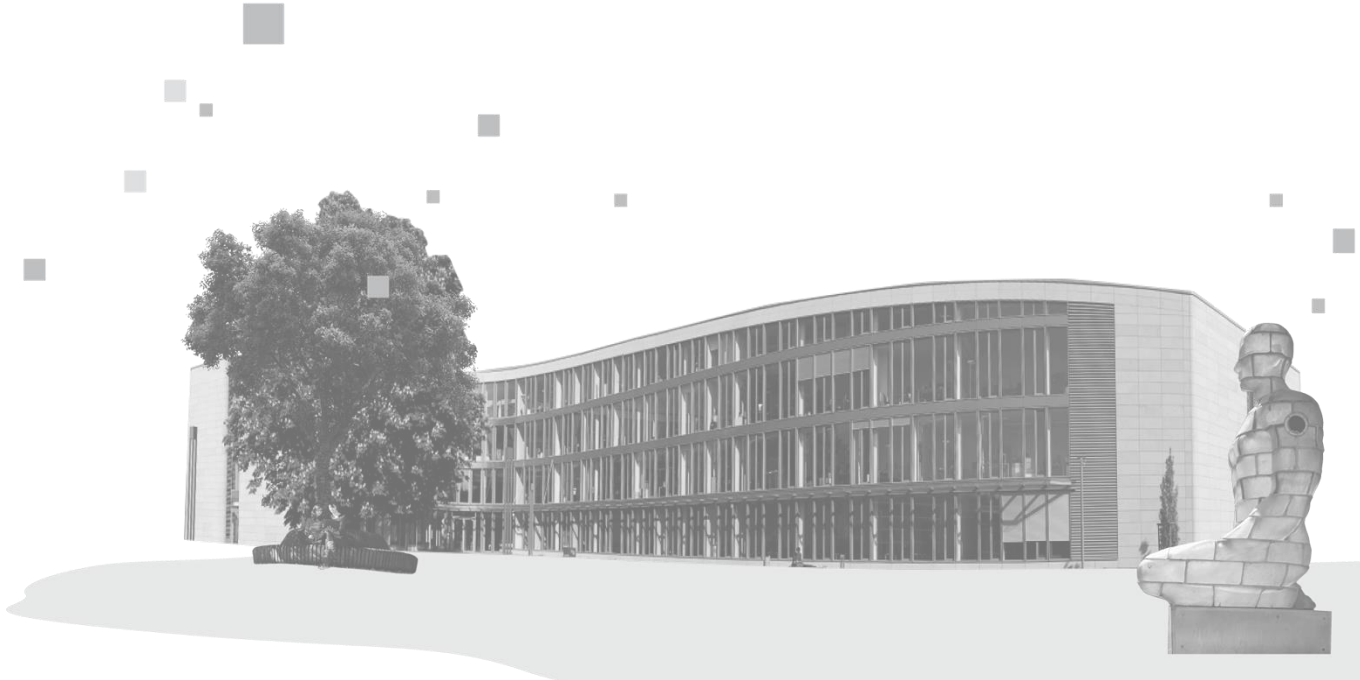
Tradeoffs

- No single defense solves everything.
  - DP noise prevents inference attacks, but server still receives raw noisy gradients
  - SecAgg encrypts updates but does not hide individual records/ clients
- Privacy comes at the cost of accuracy, computation and communication
- Balancing privacy and utility is hard and needs threat model assumptions

Federated Learning – Privacy by Design?

”Your data never leaves your device.”

Pierre Burghardt



Design IT.
Create Knowledge.

www.hpi.de

Would you deploy FL in the hospitals / give your data?

Should FL be discarded as a privacy preserving architecture or keep the label?

How would you manage the trade-off between privacy and utility in systems like this? What if it would be a global deadly pandemic?

Verdict

- “Data stays local” does not automatically mean “data stays private”
- FL changes the privacy problem, it does not erase it
- FL is not a standalone security protocol, it is merely the outer skin of a Privacy Onion
- FL reduces the need to centralize sensitive data, it can improve scalability and data usage
- FL might not be an architecture concerning privacy at all, but scalability and data acquisition
- In a realistic setting, FL needs to make and tackle many threat and system assumptions

References I

[1] Kairouz, Peter, H. Brendan McMahan, Brendan Avent, et al. “Advances and Open Problems in Federated Learning.” arXiv:1912.04977. Preprint, arXiv, March 9, 2021. https://doi.org/10.48550/arXiv.1912.04977.

[2] McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. “Communication-Efficient Learning of Deep Networks from Decentralized Data.” In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, PMLR 54:1273–1282, 2017. https://proceedings.mlr.press/v54/mcmahan17a.html.

[3] Zhu, Ligeng, Zhijian Liu, and Song Han. “Deep Leakage from Gradients.” In Advances in Neural Information Processing Systems 32, 2019. arXiv:1906.08935. https://doi.org/10.48550/arXiv.1906.08935.

References II

[4] Zhao, Bo, Konda Reddy Mopuri, and Hakan Bilen. “iDLG: Improved Deep Leakage from Gradients.” arXiv:2001.02610. Preprint, arXiv, January 8, 2020. https://doi.org/10.48550/arXiv.2001.02610.

[5] Geiping, Jonas, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. “Inverting Gradients — How Easy Is It to Break Privacy in Federated Learning?” In Advances in Neural Information Processing Systems 33, 2020. arXiv:2003.14053. https://doi.org/10.48550/arXiv.2003.14053.

[6] Hitaj, Briland, Giuseppe Ateniese, and Fernando Perez-Cruz. “Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning.” In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 603–618. ACM, 2017. https://doi.org/10.1145/3133956.3134012.

References III

[7] Kate, Sumedh, Arya Khandetod, Sujal Khardekar, and Shravani Bahulekar. “Secure Federated Learning with Differential Privacy and Homomorphic Encryption.” Medium, May 14, 2025. https://medium.com/@katesumedh/secure-federated-learning-with-differential-privacy-and-homomorphic-encryption-aa8afad13049.

[8] Harrod, Jordan. “Differential Privacy + Federated Learning Explained (+ Tutorial) | #AI101.” YouTube video. Accessed June 25, 2026. https://www.youtube.com/watch?v=MOcTGM_UteM.

References IV (not used on slides, but possible discussion)

[9] McMahan, H. Brendan, Daniel Ramage, Kunal Talwar, and Li Zhang. “Learning Differentially Private Recurrent Language Models.” In 6th International Conference on Learning Representations, ICLR 2018. OpenReview, 2018. arXiv:1710.06963. https://openreview.net/forum?id=BJ0hF1Z0b.

[10] Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. “Practical Secure Aggregation for Privacy-Preserving Machine Learning.” In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191. ACM, 2017. https://doi.org/10.1145/3133956.3133982.

Further Image References

Brain MRI:

https://warum-smartgrid.de/brain-imaging-examples-brain-mri-how-to-read-mri-brain-scan/

Doctor:

https://www.magnific.com/premium-photo/friendly-male-doctor-with-open-hand-ready-hugging_45218006.htm

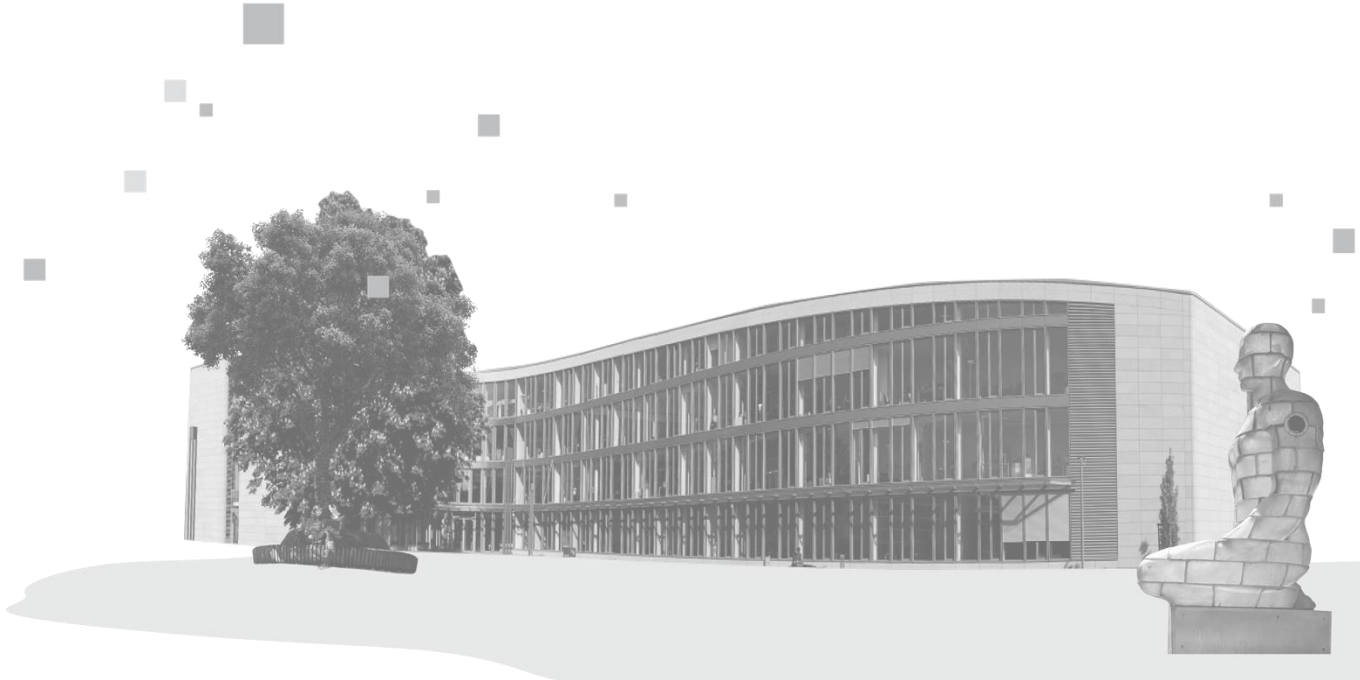
Icons:

https://www.flaticon.com/

Federated Learning – Privacy by Design?

”Your data never leaves your device.”

Pierre Burghardt



Design IT.
Create Knowledge.

www.hpi.de