Learning without collecting the data, *a timeline from 2016 to 2025*

A review of methods from 2016 to 2025 for learning from phones without gathering their text, what each still sends, and what outsiders found where they tested it.

MITWA LAB, MUMBAI

IN PLAIN WORDS

When a company wants a model to improve from how people use their phones, the obvious route is to collect what people typed or said. Since 2016, researchers have built ways to learn from phones while the text stays on them. The phone works out a small change to the model, or adds random noise to a report, and sends that in place of the text. This review follows those methods in date order. For each it says what still leaves the phone and, where outside researchers have tested it, what they found. Two findings run through all of it. A model update can sometimes be turned back into the text behind it, and most guarantees depend on who is trusted to run the server or to write the software on the phone.

What can leave a phone when the text stays

A phone that keeps its text can still send updates, noisy reports or selection signals, and each phone-side scheme below uses one of these. An update is the change a phone makes to a shared model after training briefly on its own data, and the server averages the updates from many phones; McMahan et al. (2017) named this federated learning. In a noisy report, as in local differential privacy, the phone randomises a value before sending it, so the server never holds the true one (Apple, 2017a). A selection signal, which appears at the end of the timeline, says only which of several options the company already holds is closest to what is on the phone.

Differential privacy, which recurs below, bounds how much a released result can reveal about any one record (in some systems one user, in others one device). The probability of any outcome can change by at most a factor of e to the power ε when one record is added or removed, plus a small allowance δ, so a smaller ε is a tighter bound (Abadi et al., 2016). The review also uses the distinction Kairouz et al. (2021) draw between an honest-but-curious server, which follows the protocol and studies everything it receives, and a malicious one, which may also tamper with training.

What you say to Bloom and what it remembers stays on the phone. Our technical note 03 records a cost of that privacy, which is that the lab cannot watch the shipped model at work. This article reviews other groups' work. It makes no statement about how Bloom is trained or improved, and nothing in it describes how Bloom works.

2016: a name for the idea, and three tools beside it

McMahan et al. (2017) was first posted in February 2016 and presented in 2017. In its federated averaging each phone runs several training steps on its own data and sends back only the result. As their headline, the authors report 10 to 100 times fewer communication rounds than synchronised stochastic gradient descent, across five model types and four datasets, but the gain depended on how the data was split. On the deliberately skewed split of MNIST digits, the best gain in their table for 99% accuracy was 2.8 times, and three of the eight settings were slower than the baseline. The data is never uploaded, but the paper says plainly that the server must still be trusted to some degree.

Three other pieces of work date from that year. In June, Apple announced that it would use local differential privacy for some data collection, without saying what privacy loss it allowed (Tang et al., 2017). In October, Abadi et al. (2016) presented a method for training neural networks with a differential privacy guarantee, on data held centrally, and measured the price of the noise added during training. On handwritten digits the model reached 97% test accuracy at an ε of 8, against 98.3% without the guarantee. On CIFAR-10 photographs, at the same ε, it reached 73% against about 80%, with early layers taken from a model trained on public data. In November, Bonawitz et al. (2016) described a practical secure aggregation protocol, a cryptographic method under which the server learns only the sum of many phones' updates. They raised at the outset the possibility that an update might expose its training examples, for instance the set of words a person recently typed, and the paper presents no attack.

2017: an outside audit of a shipped system

Tang et al. (2017), at three universities, reverse-engineered Apple's differential privacy on macOS 10.12 and examined iOS 10.1.1 through its configuration files and reports. Each submitted item carried a privacy loss of 1 or 2, within the range common in the literature. The system allowed several submissions a day, though. For the four launch features the configured allowances summed to 16 a day, but deep links were not yet implemented on the macOS they examined, so the figure in practice there was up to 6 a day, against up to 14 on iOS 10.1.1. Unused allowance rolled over and the allowance renewed every day, so the possible loss grew with each day since a person opted in, and an iOS 11 beta raised the daily figure by at least 29. The authors judged the 16 a day the system permitted well above what the literature usually treats as reasonable, and objected that Apple had published none of these figures.

Apple's own account came later in 2017. A December post describes opt-in collection with noise added on the phone and an ε of 4 for emoji and for Safari (Apple, 2017a). An undated overview, whose file metadata gives November 2017, lists an ε of 4 for Lookup Hints with two donations a day, 4 for emoji with one, and 8 for QuickType with two (Apple, 2017b). Adding losses across a day, the rule that Tang et al. describe, gives 16 for QuickType alone (our arithmetic, at the maximum donation rate). Neither Apple document that we read states a daily or lifetime total, and we opened no independent audit later than Tang et al.

2018: a keyboard learns on the phones

Hard et al. (2018) trained the next-word model of Google's Gboard keyboard by federated averaging and compared it with the same network trained on a server, using text logged from users who had opted in to share snippets. On text held on phones the federated model's top-1 recall was 15.8% against 15.0%, with quoted errors of 0.3 and 0.5 points, and on the server's logs the figures were 16.4% for the federated model and 16.5% for the server one. The federated model was slightly ahead on phone-held text, which the authors consider more representative, and level on the logs. They caution that the comparison is not exactly like for like because the two used different variants of the training procedure, and they say that users must trust the aggregation server not to inspect individual uploads.

That year McMahan et al. (2018) added a guarantee for a whole user rather than one record. On a public dataset of 763,430 users, training without noise reached 17.5% accuracy in 4,120 rounds of 100 users each. With an ε of 4.6 and a δ of 10^{-9} the authors estimate, from runs at about 100 users per round plus an accounting bound, that the same accuracy would take 4,980 rounds of about 5,000 users, roughly 60 times the computation (our arithmetic on their figures). They conclude that, with enough users, privacy is paid for in computation rather than accuracy.

2019: updates turn out to be readable

Zhu et al. (2019), at MIT, showed that the gradient a model computes on a training example can be turned back into the example. The attack adjusts dummy data until its gradient matches the real one. On images the recovered result was almost identical to the original, and on three example sentences with a masked-word model the main content was recovered, with a few mismatched tokens. By the paper's own account the attack worked only for batches of up to 8 images no larger than 64 by 64 pixels, and the effort grew with the batch, from 270 iterations for one image to 2,711 for eight on a ResNet-20 for CIFAR images. Added noise stopped the attack only at a level that was starting to affect accuracy, and discarding more than about a fifth of the gradient's smallest entries stopped it too. A federated averaging update comes from several local steps rather than one gradient, so the two settings differ.

2020 and 2021: how far the attack carries

Geiping et al. (2020), at the University of Siegen, pushed the attack to a trained ResNet-152 on ImageNet photographs and recovered a single image from its gradient. For federated averaging they report that up to 100 local steps did not degrade the reconstruction of one image. With 100 images averaged together (a ResNet on CIFAR-100) most reconstructions were unrecognisable, though a few could be read. Their federated-averaging experiments used an untrained small network on CIFAR-10. They concluded that only formal differential privacy gives a guarantee.

Huang et al. (2021) re-evaluated that attack independently. The implementation of Geiping and colleagues assumed the attacker also knows two things the protocol need not reveal, statistics of the private batch and its labels. With those assumptions removed the attack weakened substantially. Huang et al. report that combined defences made it less effective even under its original assumptions. Their best combination, at a batch size of 32, cost about 7% accuracy, and they say a batch size below 32 is not safe against the strongest attack even with weak defences applied. The record is split between showing that reconstruction is possible (Zhu et al.; Geiping et al.) and arguing that the practical risk is narrower (Huang et al.), though Huang et al. do not treat small batches as safe.

2021: models that remember, and a survey that names the gap

Carlini et al. (2021) queried GPT-2, a language model trained on public web text, and found more than 600 verbatim training strings among 1,800 candidates, including public personal details such as names, phone numbers and email addresses, some of which appear in only one training document. Larger models were more vulnerable. By our argument this bears on the review, because a shared model trained on phone data could hold some of it, so a bound on what the model reveals matters apart from where the data sits.

Kairouz et al. (2021), a survey of federated learning, names provable limits on what a server can reconstruct about one client as the natural goal, and lists the limitations of existing tools. Central differential privacy needs a trusted server. Local differential privacy removes that need, since each client adds noise before sharing, but achieving it without losing utility is hard, especially for high-dimensional data.

2022 and 2023: servers that cheat

Pasquini et al. (2022) attacked secure aggregation. A malicious server can send different versions of the model to different phones in one round, which the authors call model inconsistency, and arrange that the sum reveals one target's update whatever the number of users. They place the fault in the federated protocol, which does not check what the server sends, rather than in the cryptography. They also note that the differential privacy used in real deployments is central, with noise added after aggregation, which a malicious server can skip.

Boenisch et al. (2023) went further with an active server that makes inconspicuous changes to the weights it sends, which they call trap weights. On mini-batches of 100 points they perfectly reconstructed more than 50% of ImageNet training images and more than 65% of IMDB movie reviews. They also report that even an honest-but-curious server can read some training points when a model begins with a fully connected layer. Both papers demonstrate on public benchmark data, and their main attacks assume a server that departs from the protocol.

2023: more than twenty keyboard models with a formal bound

Xu et al. (2023), at Google, report launching more than twenty language models for the Gboard keyboard, each trained with federated learning under a formal bound. The bound is stated in zero-concentrated differential privacy, with ρ between 0.2 and 2. At a δ of 10^{-10} they convert the smallest, 0.25, to an ε of 4.49 and the largest, 1.86, to 13.69, and they say their values are tighter than the figure above 2.6 used by the US Census Bureau. According to an appendix, the guarantee is central, so the provider is trusted to implement it correctly, it covers one training run, excluding tuning and the choice of the final model, and its unit is the device. In their introduction the authors note that an earlier study of production language models checked a differentially private variant of federated averaging by empirical auditing, without a formal guarantee.

Read with Pasquini et al., this gives a limit by our argument. A central bound speaks to what the models reveal when the operator follows the procedure, and a server that does not can step around the noise.

2025: a choice among synthetic examples

Apple (2025) describes the newest variant. For short text, such as requests for a generated emoji, opted-in devices are polled on whether they have seen a text fragment and answer with noise, calibrated so that hundreds of people must use a term before Apple can discover it. For longer text such as emails, Apple writes synthetic messages itself and sends a numerical summary of each to a small number of opted-in devices. Each device compares the summaries with a few recent emails and returns, with noise, only which synthetic message is closest, and Apple says the email contents never leave the device. The page gives no privacy loss value and no measured quality result.

What the ten years add up to

Every phone-side scheme in this timeline sends something off the phone, whether updates, noisy reports or selection signals, so "without collecting the data" describes where the text sits. What the guarantees cover is conditional. They depend on trusting the server (McMahan et al., 2017; Hard et al., 2018), on the operator adding noise correctly (Xu et al., 2023), on a server that sends every phone the same model (Pasquini et al., 2022) or, for Apple's local scheme, on settings in the maker's own software that Tang et al. had to reverse-engineer because Apple had published none of them. The evidence is also lopsided in who produced it. In our tally of the sources above, every utility result for a deployed system comes from the company that built it, and the audit and the attacks on federated schemes come mostly from authors listed at universities or research institutes (one co-author of Huang et al. is at a company lab). How serious reconstruction is in practice remains contested.

What is not yet known

- Whether Apple's current privacy values resemble what Tang et al. found for macOS 10.12 and iOS 10.1.1, since no independent audit after theirs was among our sources and the 2025 page states none.
- Whether trap-weight or model-inconsistency attacks work against a production keyboard system, which neither paper reports testing.
- How strong a guarantee can be for a language with few users, given the link Xu et al. report between population size and the guarantee.
- What value of ε is enough. Tang et al. note that theorists typically treat it as a social choice, and the sources give no agreed threshold.
- How far reconstruction carries to long free-form text, where our only text results are masked-word prediction (Zhu et al., 2019) and movie-review classification (Boenisch et al., 2023).
- Whether the 2025 synthetic-message method has a measurable privacy bound, and what it costs in quality.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS 2016). arXiv:1607.00133. arxiv.org/abs/1607.00133

Apple (2017a). Learning with privacy at scale. Apple Machine Learning Research (Apple Differential Privacy Team), published 6 December 2017. machinelearning.apple.com/research/learning-with-privacy-at-scale

Apple (2017b). Differential privacy overview. Apple Inc., undated PDF; file metadata gives 2 November 2017. www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf

Apple (2025). Understanding aggregate trends for Apple Intelligence using differential privacy. Apple Machine Learning Research, published 14 April 2025; the page also carries a later update note, dated 9 September 2026, about opting in to help improve Siri and Apple Intelligence. machinelearning.apple.com/research/differential-privacy-aggregate-trends

Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A. S., Shumailov, I., and Papernot, N. (2023). When the curious abandon honesty: Federated learning is not private. Accepted at the 8th IEEE European Symposium on Security and Privacy (EuroS&P 2023). arXiv:2112.02918 (v2, 12 April 2023). arxiv.org/abs/2112.02918

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. (2016). Practical secure aggregation for federated learning on user-held data. NIPS 2016 Workshop on Private Multi-Party Machine Learning. arXiv:1611.04482. arxiv.org/abs/1611.04482

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. (2021). Extracting training data from large language models. arXiv:2012.07805 (v2, 15 June 2021). arxiv.org/abs/2012.07805

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. (2020). Inverting gradients: How easy is it to break privacy in federated learning? arXiv:2003.14053 (v2, 11 September 2020). arxiv.org/abs/2003.14053

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. (2018). Federated learning for mobile keyboard prediction. arXiv:1811.03604 (v2, 28 February 2019). arxiv.org/abs/1811.03604

Huang, Y., Gupta, S., Song, Z., Li, K., and Arora, S. (2021). Evaluating gradient inversion attacks and defenses in federated learning. 35th Conference on Neural Information Processing Systems (NeurIPS 2021). arXiv:2112.00059. arxiv.org/abs/2112.00059

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning (per the arXiv listing). arXiv:1912.04977 (v3, 9 March 2021). arxiv.org/abs/1912.04977

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS 2017), JMLR W&CP volume 54. arXiv:1602.05629. arxiv.org/abs/1602.05629

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. (2018). Learning differentially private recurrent language models. International Conference on Learning Representations (ICLR 2018). arXiv:1710.06963. arxiv.org/abs/1710.06963

Pasquini, D., Francati, D., and Ateniese, G. (2022). Eluding secure aggregation in federated learning via model inconsistency. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22). arXiv:2111.07380 (v5, 6 September 2022). arxiv.org/abs/2111.07380

Tang, J., Korolova, A., Bai, X., Wang, X., and Wang, X. (2017). Privacy loss in Apple's implementation of differential privacy on MacOS 10.12. arXiv:1709.02753 (v2, 11 September 2017). arxiv.org/abs/1709.02753

Xu, Z., Zhang, Y., Andrew, G., Choquette-Choo, C. A., Kairouz, P., McMahan, H. B., Rosenstock, J., and Zhang, Y. (2023). Federated learning of Gboard language models with differential privacy. ACL 2023 industry track (per the arXiv listing). arXiv:2305.18465 (v2, 17 July 2023). arxiv.org/abs/2305.18465

Zhu, L., Liu, Z., and Han, S. (2019). Deep leakage from gradients. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). arXiv:1906.08935. arxiv.org/abs/1906.08935