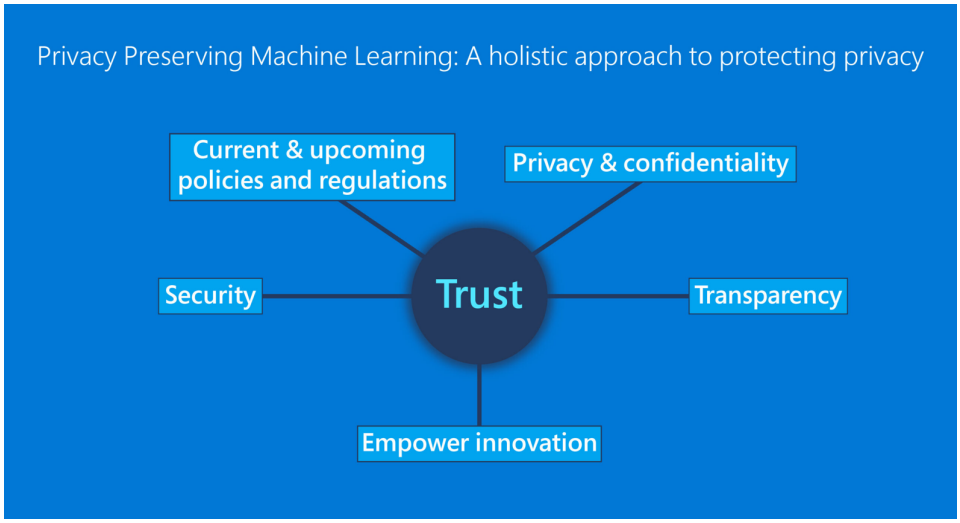
Privacy Preserving Machine Learning: Maintaining confidentiality and preserving trust

Published November 9, 2021

By Victor Ruehle, Principal Research Manager; Robert Sim, Principal Research Manager; Sergey Yekhanin, Partner Research Manager; Nishanth Chandran, Principal Researcher; Melissa Chase, Principal Researcher; Daniel Jones, Senior Applied Researcher; Kim Laine , Principal Researcher; Boris Köpf, Principal Researcher; Jaime Teevan, Chief Scientist & Technical Fellow; Jim Kleewein, Technical Fellow; Saravan Rajmohan, Partner Director of AI and Applied Research



Machine learning (ML) offers tremendous opportunities to increase productivity. However, ML systems are only as good as the quality of the data that informs the training of ML models. And training ML models requires a significant amount of data, more than a single individual or organization can contribute. By sharing data to collaboratively train ML models, we can unlock value and develop powerful language models that are applicable to a wide variety of scenarios, such as text prediction and email reply suggestions. At the same time, we recognize the need to preserve the confidentiality and privacy of individuals and earn and maintain the trust of the people who use our products. Protecting the confidentiality of our customers' data is core to our mission. This is why we're excited to share the work we're doing as part of the Privacy Preserving Machine Learning (PPML) initiative.

Research Areas

Artificial intelligence

Security, privacy, and cryptography

Research Groups

Azure Research - Security and Privacy

Microsoft Search, Assistant and Intelligence

Privacy Preserving Machine Learning Innovation

M365 Research

AI for Domains (AID)

Related tools

Simple Encrypted Arithmetic Library (SEAL)

Related projects

Learning with Weak Supervision

Project Laplace

EzPC (Easy Secure Multi-party Computation)

Related labs

Microsoft Research Lab - Cambridge

The PPML initiative was started in partnership between Microsoft Research and Microsoft product teams with the objective of protecting the confidentiality and privacy of customer data when training large-capacity language models. The goal of the PPML initiative is to improve existing techniques and develop new ones for protecting sensitive information that work for both individuals and enterprises. This helps ensure that our use of data protects people's privacy and the data is utilized in a safe fashion, avoiding leakage of confidential and private information.

This blog post discusses emerging research on combining techniques to ensure privacy and confidentiality when using sensitive data to train ML models. We illustrate how employing PPML can support our ML pipelines in meeting stringent privacy requirements and that our researchers and engineers have the tools they need to meet these requirements. We also discuss how applying best practices in PPML enables us to be transparent about how customer data is applied.

A holistic approach to PPML

Recent research has shown that deploying ML models can, in some cases, implicate privacy in unexpected ways. For example, pretrained public language models that are fine-tuned on private data can be misused to recover private information, and very large language models have been shown to memorize training examples, potentially encoding personally identifying information (PII). Finally, inferring that a specific user was part of the training data can also impact privacy. Therefore, we believe it's critical to apply multiple techniques to achieve privacy and confidentiality; no single method can address all aspects alone. This is why we take a three-pronged approach to PPML: understanding the risks and requirements around privacy and confidentiality, measuring the risks, and mitigating the potential for breaches of privacy. We explain the details of this multi-faceted approach below.

Understand: We work to understand the risk of customer data leakage and potential privacy attacks in a way that helps determine confidentiality properties of ML pipelines. In addition, we believe it's critical to proactively align with policy makers. We take into account local and international laws and guidance regulating data privacy, such as the General Data Protection Regulation (GDPR) and the EU's policy on trustworthy AI. We then map these legal principles, our contractual obligations, and responsible AI principles to our technical requirements and develop tools to communicate with policy makers how we meet these requirements.

Measure: Once we understand the risks to privacy and the requirements we must adhere to, we define metrics that can quantify the identified risks and track success towards mitigating them.

Mitigate: We then develop and apply mitigation strategies, such as differential privacy (DP), described in more detail later in this blog post. After we apply mitigation strategies, we measure their success and use our findings to refine our PPML approach.

PPML is informed by a three-pronged approach: 1) understanding the risk and regulatory requirements, 2) measuring the vulnerability and success of attacks, and 3) mitigating the risk.

PPML in practice

Several different technologies contribute to PPML, and we implement them for a number of different use cases, including threat modeling and preventing the leakage of training data. For example, in the following text-prediction scenario, we took a holistic approach to preserving data privacy and collaborated across Microsoft Research and product teams, layering multiple PPML techniques and developing quantitative metrics for risk assessment.

We recently developed a personalized assistant for composing messages and documents by using the latest natural language generation models, developed by Project Turing. Its transformer-based architecture uses attention mechanisms to predict the end of a sentence based on the current text and other features, such as the recipient and subject. Using large transformer models is risky in that individual training examples can be memorized and reproduced when making predictions, and these examples can contain sensitive data. As such, we developed a strategy to both identify and remove potentially sensitive information from the training data, and we took steps to mitigate memorization tendencies in the training process. We combined careful sampling of data, PII scrubbing, and DP model training (discussed in more detail below).

Mitigating leakage of private information

We use security best practices to help protect customer data, including strict eyes-off handling by data scientists and ML engineers. Still, such mitigations cannot prevent subtler methods of privacy leakage, such as training data memorization in a model that could subsequently be extracted and linked to a user. That is why we employ state-of-the-art privacy protections provided by DP and continue to contribute to the cutting-edge research in this field. For privacy-impacting use cases, our policies require a security review, a privacy review, and a compliance review, each including domain-specific quantitative risk assessments and application of appropriate mitigations.

Differential privacy

Microsoft pioneered DP research back in 2006, and DP has since been established as the de facto privacy standard, with a vast body of academic literature and a growing number of large-scale deployments across the industry (e.g., DP in Windows telemetry or DP in Microsoft Viva Insights) and government. In ML scenarios, DP works by adding small amounts of statistical noise during training, the purpose of which is to conceal the contributions of individual parties whose data is being used. When DP is employed, a mathematical proof validates that the final ML model learns only general trends in the data without acquiring information unique to any specific party. Differentially private computations entail the notion of a privacy budget, ϵ, which imposes a strict upper bound on information that might leak from the process. This guarantees that no matter what auxiliary information an external adversary may possess, their ability to learn something new about any individual party whose data was used in training from the model is severely limited.

In recent years, we have been pushing the boundaries in DP research with the overarching goal of providing Microsoft customers with the best possible productivity experiences through improved ML models for natural language processing (NLP) while providing highly robust privacy protections.

- In the Microsoft Research papers Differentially Private Set Union and Differentially private n-gram extraction, we developed new algorithms for exposing frequent items, such as unigrams or n-grams coming from customer data, while adhering to the stringent guarantees of DP. Our algorithms have been deployed in production to improve systems such as assisted response

generation.

- In the Microsoft Research paper Numerical Composition of Differential Privacy, we developed a new DP accountant that gives a more accurate result for the expended privacy budget when training on customer data. This is particularly important when training on enterprise data where typically many fewer participants are present in the dataset. With the new DP accountant, we can train models for longer, thereby achieving higher utility while using the same privacy budget.

- Finally, in our recent paper Differentially private fine-tuning of language models, we demonstrate that one can privately fine-tune very large foundation NLP models, such as GPT-2, nearly matching the accuracy of nonprivate fine-tuning. Our results build on recent advances in parameter-efficient fine-tuning methods and our earlier work on improved accounting for privacy.

When training or fine-tuning machine learning models on customer content, we adhere to strict policy regarding the privacy budget[1].

Threat modeling and leakage analysis

Even though DP is considered the gold standard for mitigation, we go one step further and perform threat modeling to study the actual risk before and after mitigation. Threat modeling considers the possible ways an ML system can be attacked. We have implemented threat modeling by studying realistic and relevant attacks, such as the tab attack (discussed below) in a black box setting, and we have considered and implemented novel attack angles that are very relevant to production models, such as the model update attack. We study attacks that go beyond the extraction of training data and approximate more abstract leakage, like attribute inference. Once we have established threat models, we use those attacks to define privacy metrics. We then work to make sure all of these attacks are mitigated, and we continuously monitor their success rates. Read further to learn about some of the threat models and leakage analyses we use as part of our PPML initiative.

Model update attacks. In the paper Analyzing Information Leakage of Updates to Natural Language Models, a Microsoft Research team introduced a new threat model where multiple snapshots of a model are accessible to a user, such as predictive keyboards. They proposed using model update attacks to analyze leakage in practical settings, where language models are frequently updated by adding new data, fine-tuning public pre-trained language models on private data, or by deleting user data to comply with privacy law requirements. The results showed that access to such snapshots can leak phrases that were used to update the model. Based on the attack, leakage analyses of text prediction models can be performed without the need to monitor it.

Tab attacks. Tab attacks can occur when an attacker has access to top-1 predictions of a language model, and the text auto-completion feature, in an email app for example, is applied by pressing the Tab key. It's well known that large language models can memorize individual training instances, and recent work has demonstrated that practical attacks extracting verified training instances from GPT-2 is a risk. In the paper Training Data Leakage Analysis in Language Models, a team of Microsoft researchers established an approach to vetting a language model for training data leakage. This approach enables the model builder to establish the extent to which training examples can be extracted from the model using a practical attack. The model owner can use this method to verify that mitigations are performing as expected and determine whether a model is safe to deploy.

Poisoning attacks. In the paper Property Inference from Poisoning, Microsoft researchers and an affiliated academic considered the consequences of a scenario where some of the training data is intentionally manipulated to cause more privacy leakage. This type of data compromise can occur, for example, in a collaborative learning setting where data from several parties or tenants are combined to achieve a better model and one of the parties is behaving dishonestly. The paper illustrates how such a party can manipulate their data to extract aggregate statistics about the rest of the training set. In this case, several parties pool their data to train a spam classifier. If one of those parties has malicious intent, it can use the model to obtain the average sentiment of the emails in the rest of the training set, demonstrating the need to take particular care to ensure that the data used in such joint training scenarios is trustworthy.

Future areas of focus for PPML

As we continue to apply and refine our PPML processes with the intent of further enhancing privacy guarantees, we recognize that the more we learn, the larger the scope becomes for addressing privacy concerns across the entire pipeline. We will continue focusing on:

- Following regulations around privacy and confidentiality
- Proving privacy properties for each step of the training pipeline
- Making privacy technology more accessible to product teams
- Applying decentralized learning
- Investigating training algorithms for private federated learning, combining causal and federated learning, using federated reinforcement learning principles, federated optimization, and more
- Using weakly supervised learning technologies to enable model development without direct access to the data

Decentralized learning: Federated learning and its potential

With users becoming more concerned about how their data is handled, and with increasingly stronger regulations, users are applying ever more rigorous controls in how they process and store data. As such, increasingly more data is stored in inaccessible locations or on user devices without the option of curating for centralized training.

To this end, the federated learning (FL) paradigm has been proposed, addressing privacy concerns while continuing to process such inaccessible data. The proposed approach aims to train ML models, for example, deep neural networks, on data found in local worker nodes, such as data silos or user devices, without any raw data leaving the node. A central coordinator dispatches a copy of the model to the nodes, which individually computes a local update. The updates are then communicated back to the coordinator where they are federated, for example, by averaging across the updates. The promise of FL is that raw training data remains within its local node. However, this might not mitigate all privacy risks, and additional mitigations, such as DP, are usually required.

Secure and confidential computing environments

When dealing with highly private data, our customers may hesitate to bring their data to the cloud at all. Azure confidential computing uses trusted execution environments (TEEs), backed by hardware security guarantees, to enable data analytics and ML algorithms to be computed on private data with the guarantee that cloud administrators, malicious actors that breach the cloud tenancy boundary, and even the cloud provider itself cannot gain access to the data. This enables the collaboration of multiple customers on private data without the need to trust the cloud provider.

While TEEs leverage specific hardware for security guarantees, cryptographic secure computing solutions, such as secure multi-party computation (MPC) and fully homomorphic encryption (FHE), can enable data to be processed under a layer of strong encryption. MPC refers to a set of cryptographic protocols that allows multiple parties to compute functions on their joint private inputs without revealing anything other than the output of the function to each other. FHE refers to a special type of encryption that allows computing to be done directly on encrypted data so that only the owner of the secret decryption key can reveal the result of the computation. Microsoft has developed one of the most popular FHE libraries, Microsoft SEAL.

However, both MPC and FHE have seen only limited use due to their computational performance overhead and lack of developer tooling for nonexperts. Easy Secure Multi-Party Computation (EzPC) is an end-to-end MPC system that solves these two challenges. It takes as input standard TensorFlow or ONNX code for ML inference and outputs MPC protocols that are highly performant. EzPC enables the use of state-of-the-art ML algorithms for inference tasks. Experimentally, this technology has been recently applied to secure medical image analysis and secure medical imaging AI validation research software, successfully demonstrating the EzPC system's ability to execute the algorithms without accessing the underlying data.

Broader opportunities for PPML

Advances in technology can present tremendous opportunities along with potentially equally significant risks. We aim to create leading-edge tools for realizing technology ethics from principle to practice, engage at the intersection of technology and policy, and work to ensure that the continued advancement of technology is responsible, privacy protective, and beneficial to society.

However, even with the technologies discussed above, there continue to be outstanding questions in the PPML space. For example, can we arrive at tighter theoretical bounds for DP training and enable improved privacy-utility trade-offs? Will we be able to train ML models from synthetic data in the future? Finally, can we tightly integrate privacy and confidentiality guarantees into the design of the next generation of deep learning models?

At Microsoft Research, we're working to answer these questions and deliver the best productivity experiences afforded by the sharing of data to train ML models while preserving the privacy and confidentiality of data. Please visit our Privacy Preserving Machine Learning Group page and learn more about the holistic approach we're taking to unlock the full potential of enterprise data for intelligent features while honoring our commitment to keep customer data private.

For the latest discussions and developments on privacy and security, you can view parts 1 and 2 of the Future of Privacy and Security track on-demand from Microsoft Research Summit.

Appendix

- If you're interested in learning more about the different ways Microsoft protects your data, please visit the Microsoft Trust Center.
- Read more about how Microsoft approaches encryption in the cloud.
- Learn about the data protection resources Microsoft provides its customers.

---

[1] The maximum amount of privacy budget that can be consumed from each party whose data is involved in training a model over a period of six months is limited to ϵ=4.

Related publications

Differentially private n-gram extraction >

Differentially private fine-tuning of language models >

Collecting Telemetry Data Privately >

Differentially Private Set Union >

Analyzing Information Leakage of Updates to Natural Language Models >

Secure Medical Image Analysis with CrypTFlow >

Numerical Composition of Differential Privacy >

Meet the authors

Victor Ruehle
Principal Research Manager
Learn more >

Robert Sim
Principal Research Manager
Learn more >

Sergey Yekhanin
Partner Research Manager
Learn more >

Nishanth Chandran
Principal Researcher
Learn more >

Melissa Chase
Principal Researcher
Learn more >

Daniel Jones
Senior Applied Researcher
Learn more >

Kim Laine
Principal Researcher
Learn more >

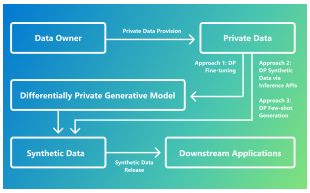
Boris Köpf
Principal Researcher
Learn more >

Jaime Teevan
Chief Scientist & Technical Fellow
Learn more >

Jim Kleewein
Technical Fellow
Microsoft 365
Learn more >

Saravan Rajmohan
Partner Director of AI and Applied Research
Learn more >

Continue reading



May 29, 2024

The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI >

October 16, 2023

DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models >

June 7, 2023

Research Focus: Week of June 5, 2023 >

November 17, 2022

Research trends in privacy, security and cryptography >

See all blog posts >

Follow us:

Share this page: