Ethics of Data Science – Part III

Adversarial learning – privacy, extraction and federated attacks

Dr. Chris Anagnostopoulos, Hon. Senior Lecturer

The adversary is an additional type of user

Data owner

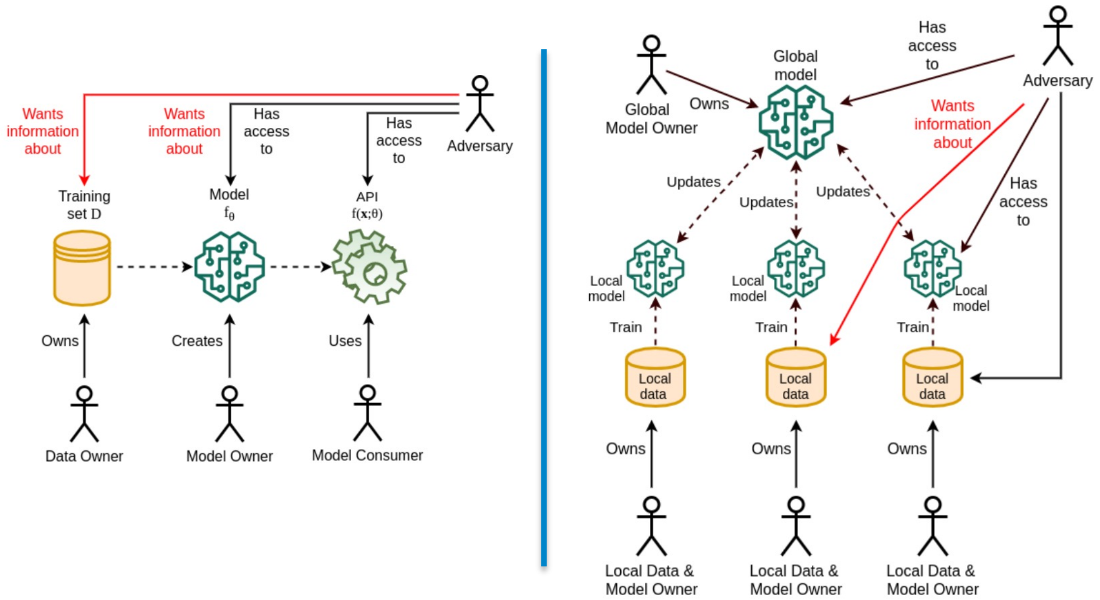
Model owner

Model user

Adversary → Honest-but-curious

Adversary → Active attacker

And, increasingly, infra service provider

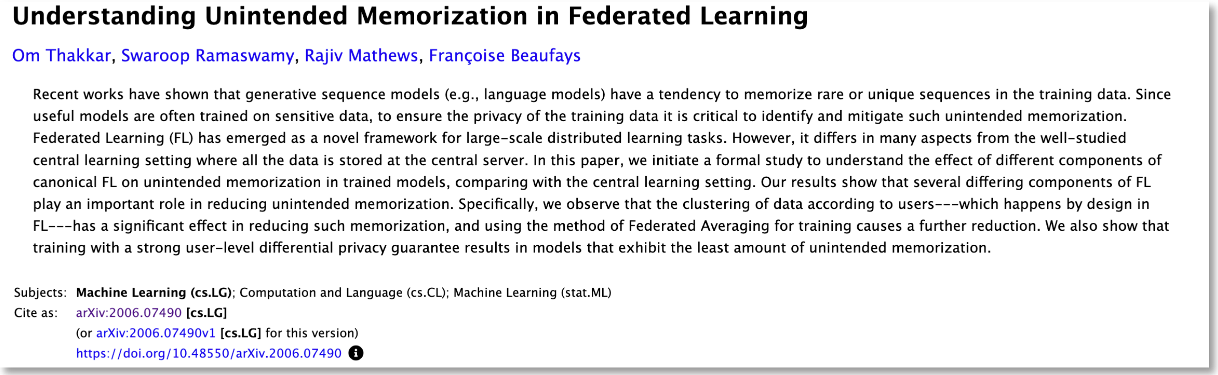


From Rigaki, Maria, and Sebastian Garcia. "A survey of privacy attacks in machine learning." *arXiv preprint arXiv:2007.07646* (2020).

Model sharing introduces further risks

- How do you enforce the “Right to be Forgotten”, or Copyright laws, with ChatGPT?
- Can a third party with access to the trained model parameters of an ML model trained using federated learning recover individual datapoints or recover properties about the training data?

These are areas of active research, with no easy or easy-to-implement answers.



Understanding Unintended Memorization in Federated Learning

Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, Françoise Beaufays

Recent works have shown that generative sequence models (e.g., language models) have a tendency to memorize rare or unique sequences in the training data. Since useful models are often trained on sensitive data, to ensure the privacy of the training data it is critical to identify and mitigate such unintended memorization. Federated Learning (FL) has emerged as a novel framework for large-scale distributed learning tasks. However, it differs in many aspects from the well-studied central learning setting where all the data is stored at the central server. In this paper, we initiate a formal study to understand the effect of different components of canonical FL on unintended memorization in trained models, comparing with the central learning setting. Our results show that several differing components of FL play an important role in reducing unintended memorization. Specifically, we observe that the clustering of data according to users---which happens by design in FL---has a significant effect in reducing such memorization, and using the method of Federated Averaging for training causes a further reduction. We also show that training with a strong user-level differential privacy guarantee results in models that exhibit the least amount of unintended memorization.

Subjects: **Machine Learning (cs.LG)**; Computation and Language (cs.CL); Machine Learning (stat.ML)
Cite as: arXiv:2006.07490 **[cs.LG]**
(or arXiv:2006.07490v1 **[cs.LG]** for this version)
https://doi.org/10.48550/arXiv.2006.07490 ℹ

There are five main types of privacy attacks

- **Membership inference attack**: “Was this protein in the training dataset?”, targeted to supervised learning or generative AI models. Can also be used to audit a model (does this model use my data?)
- **Reconstruction** or **attribute inference attack**: with partial information about one of the training examples, can the attacker reconstruct the rest, and with what level of confidence?
- **Property inference attack**: can you find the ratio of men to women in a dataset?
- **Model extraction attack:** extract enough information about the model to construct a surrogate of it that you can then use as a competitive service, or as a tool in follow-up evasion attacks
- **Data leakage attack**: data leakage involves extracting a number of datapoints from the training set (similar to membership/attribute inference but lacking control of which datapoint is leaked).

All of these attacks are easier (or even straightforward) in the white box scenario.

From Rigaki, Maria, and Sebastian Garcia. "A survey of privacy attacks in machine learning." *arXiv preprint arXiv:2007.07646* (2020).

Defending privacy also defends against overfitting

Membership Inference Attack (MIA) infers that a datapoint was in the training set by inspecting the loss of the model with respect to that datapoint (which should be lower if the datapoint was in the training set).

It can hence be shown that the ability to successfully perform a membership inference attack on a randomly selected datapoint from the training set is inversely proportional to its generalization ability[1].

Overfitting is hence sufficient to enable MIAs, but it is not necessary. Even in models that are not overfitting, there may still be *vulnerable* datapoints with very high influence – usually, outliers[2].

Conversely, defending against outliers and overfitting, also defends against privacy attacks.

1. Yeom, Samuel, et al. "Privacy risk in machine learning: Analyzing the connection to overfitting." *2018 IEEE 31st computer security foundations symposium (CSF)*. IEEE, 2018.
2. Long, Yunhui, et al. "Understanding membership inferences on well-generalized learning models." *arXiv preprint arXiv:1802.04889* (2018).

Trade-offs also exist in the adversarial setting

In [Song et al., 2019][1], the authors show that “*When using adversarial defenses [e.g., Projected Gradient Descent] to train the robust [vs evasion attacks] models, the membership inference advantage increases by up to 4.5 times compared to the naturally undefended models.”*: model predictions are forced to remain unchanged for a small area around each training example, but not for unseen examples.

This opens the door to specialised MIAs, but also seems to natively encourage overfitting by making certain types of training examples (in particular outliers) have larger influence[2,3].

Trade-off: risk of evasion on one hand, and generalization performance and risk of privacy on the other.

1. Song, Liwei, Reza Shokri, and Prateek Mittal. "Privacy risks of securing machine learning models against adversarial examples." *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 2019.
2. Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." *International conference on machine learning*. PMLR, 2019.
3. Tsipras, Dimitris, et al. "Robustness may be at odds with accuracy." *arXiv preprint arXiv:1805.12152* (2018).