



EACL 2023 Tutorial on Privacy-Preserving NLP

Block 2b: federated learning and other privacy enhancing methods



Fatemehsadat Miresghallah

May 2023

So far ...

- We heard about the **attack** landscape and **leakage** in language models



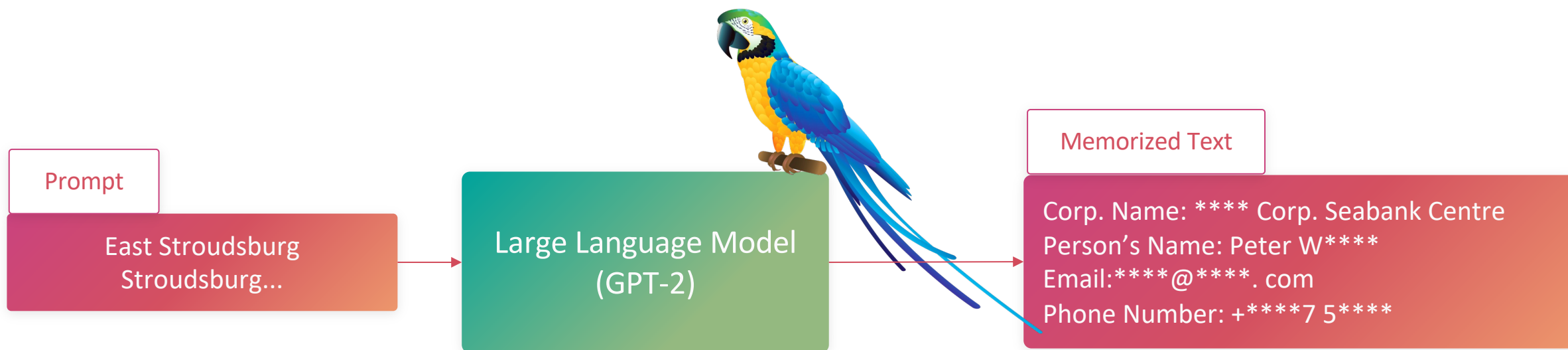
Large Models are Leaky



WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.



Large Models are Leaky: Data Extraction



Large Models are Leaky: Data Extraction

- Github CoPilot

Title:

Hi everyone, my name is Anish Athalve and I'm a PhD student at Stanford University.

Large Models are Leaky: Data Extraction

- Github CoPilot

Title:

Hi everyone, my name is Anish Athalye and I'm a PhD student at Stanford University.

<https://www.anish.io> :

Anish Athalye

I am a PhD student at MIT in the PDOS group. I'm interested in formal verification, systems, security, and machine learning.

GitHub: @anishathalye

Blog: anishathalye.com

So far ...

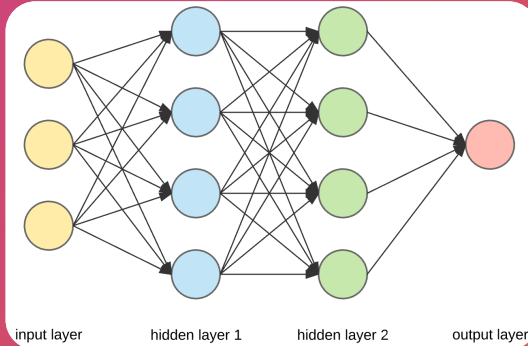
- We heard about the attack landscape and leakage in language models
- We discussed privacy protection methods with **formal guarantees**:
 - Differential Privacy



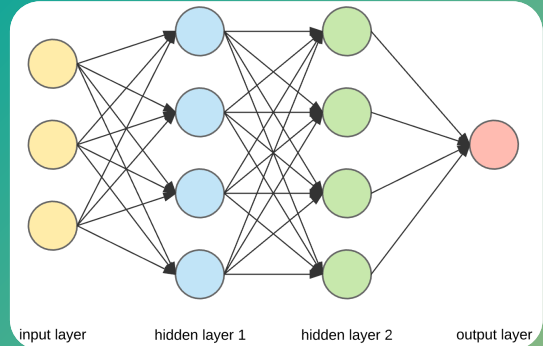
Differential Privacy

A randomized algorithm A satisfies ϵ -DP, if for all databases D and D' that differ in data pertaining to one user, and for every possible output value Y :

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^\epsilon.$$



W/ Alice



w/o Alice



So far ...

- We heard about the attack landscape and leakage in language models
- We discussed privacy protection methods with **formal guarantees**:
 - Differential Privacy
 - Encryption-based methods



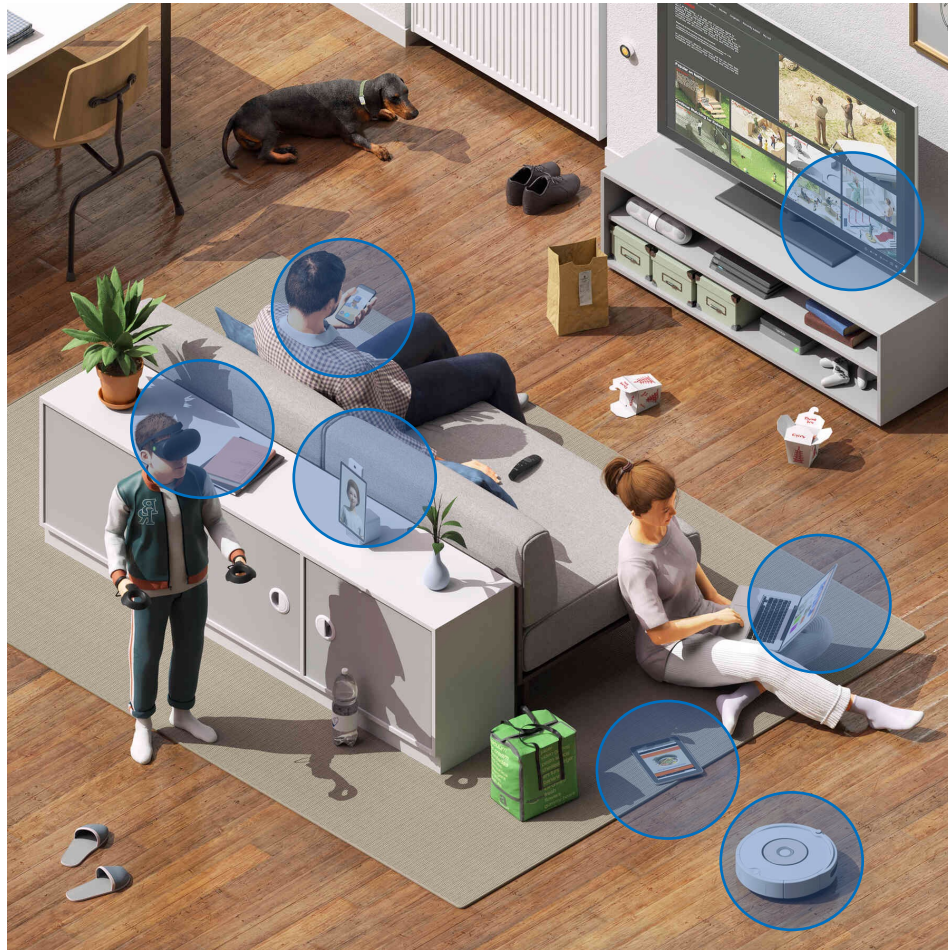
In this talk ...

- There are also **privacy-enhancing paradigms and execution modes** that do not necessarily have formal guarantees
- These methods are designed to **limit access to raw data**, but provide **no worst-case guarantees**:
 - Federated Learning
 - Split Learning
 - Privacy Regularizers

In this talk ...

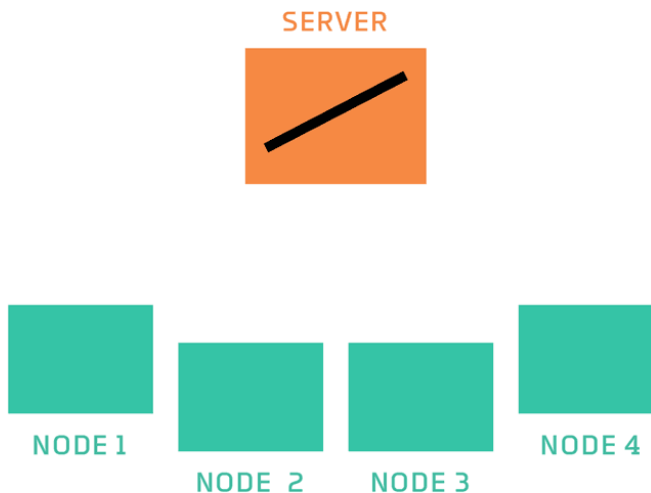
- There are also **privacy-enhancing paradigms and execution modes** that do not necessarily have formal guarantees
- These methods are designed to **limit access to raw data**, but provide **no worst-case guarantees**:
 - **Federated Learning**
 - Split Learning
 - Privacy Regularizers

Federated Learning: Background



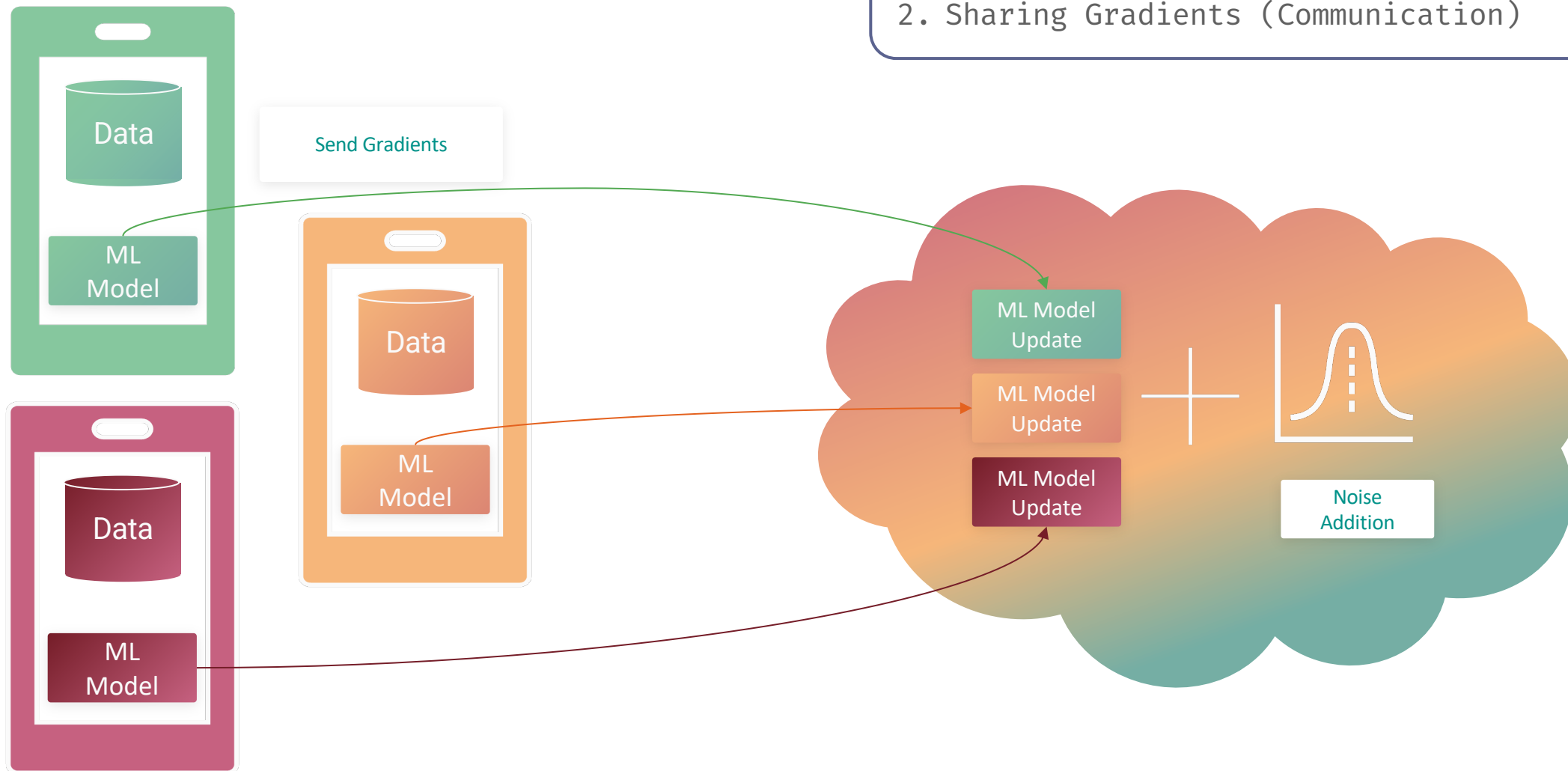
Federated Learning

- Federated learning is a machine learning setting where **multiple entities (clients)** collaborate in **solving a machine learning problem**, under the coordination of a **central server** or service provider. Each client's **raw data** is stored **locally** and **not exchanged or transferred**; instead, **focused updates** intended for immediate aggregation are used to achieve the learning objective.



Can we keep the data on device?

1. On-device Training (Computation)
2. Sharing Gradients (Communication)



Federated Analytics

- Federated **histograms** over closed sets
- Federated **heavy hitters** discovery over open sets
- Federated **SQL**
- Federated **computations?**
- etc...

Federated Optimization: challenges

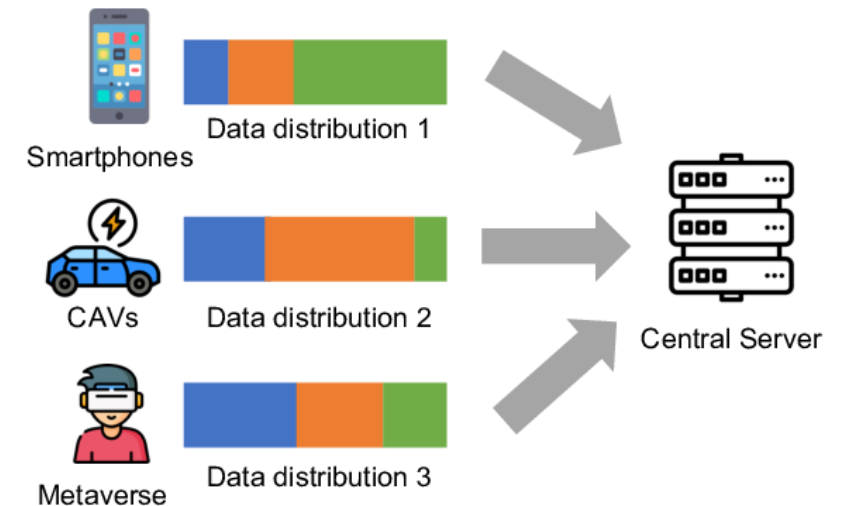
- Expensive Communication: Can **reduce communication** in federated optimization by
 1. Limiting *number of devices* involved in communication
 2. Reducing number of *communication rounds*
 3. Reducing *size of messages* sent over network

Federated Optimization: challenges

- Expensive Communication
- Privacy Concerns
 - Local Differential Privacy
 - Secure Aggregation

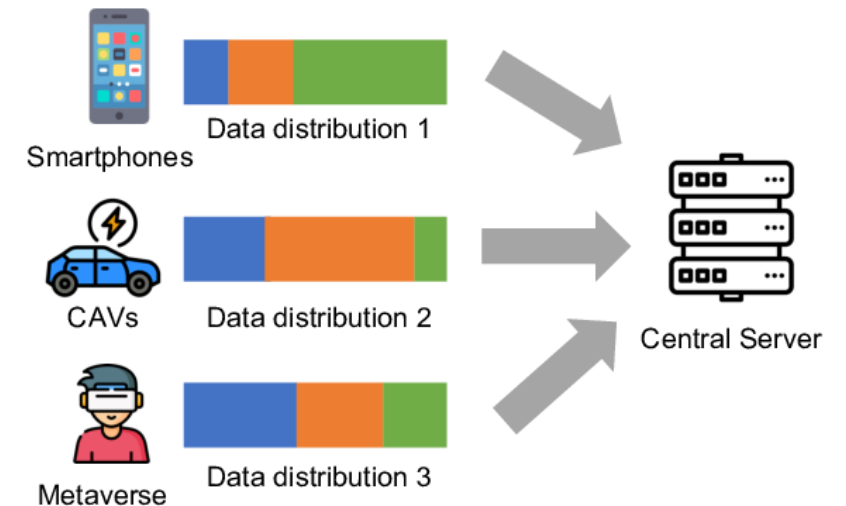
Federated Optimization: challenges

- Expensive Communication
- Privacy Concerns
- Statistical Heterogeneity
 - Unbalanced, **non-IID data**: Heterogeneous (i.e., non-identically distributed) data and systems can **bias** optimization procedures



Federated Optimization: challenges

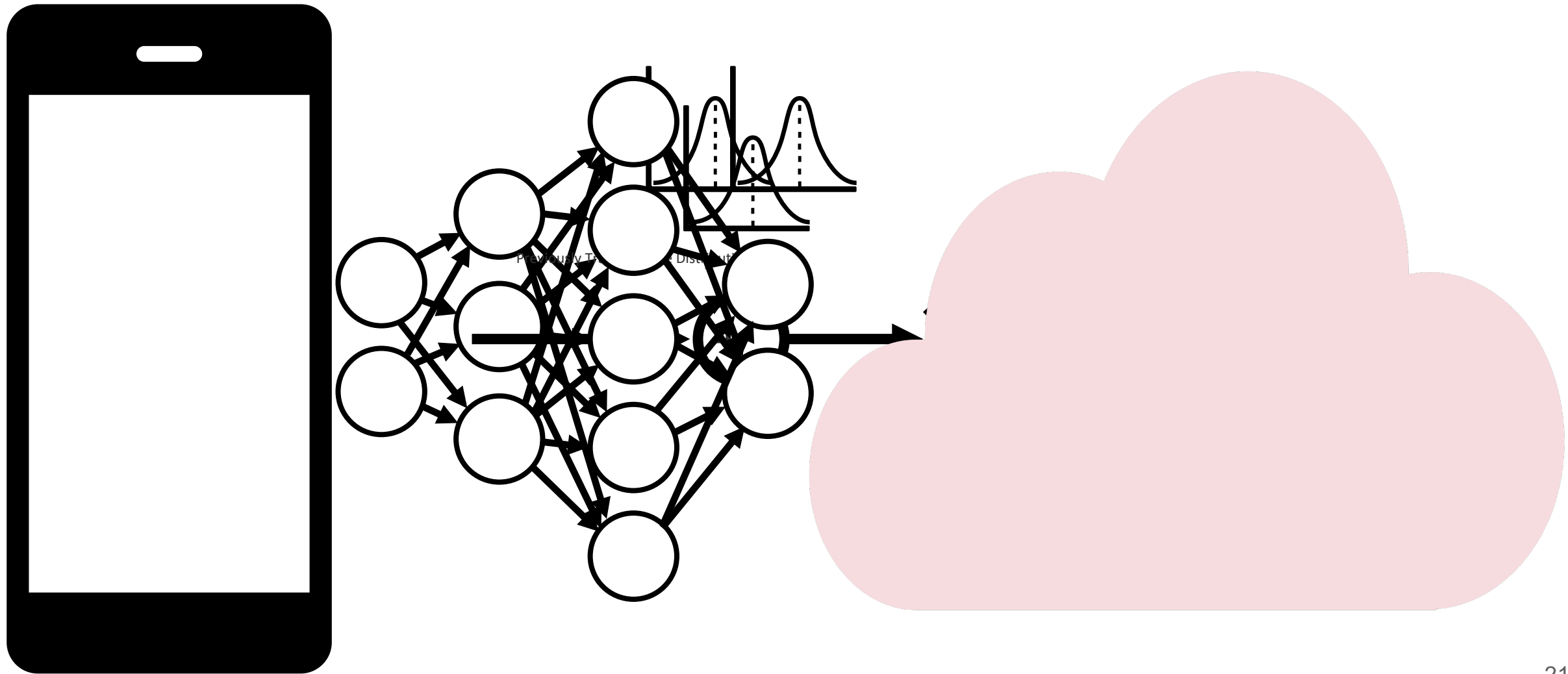
- Expensive Communication
- Privacy Concerns
- Statistical Heterogeneity
- System Heterogeneity
 - **Variable hardware**, connectivity: systems heterogeneity (e.g., dropping devices*) can exacerbate convergence issues



In this talk ...

- There are also **privacy-enhancing paradigms and execution modes** that do not necessarily have formal guarantees
- These methods are designed to **limit access to raw data**, but provide **no worst-case guarantees**:
 - Federated Learning
 - **Split Learning**
 - Privacy Regularizers

Split Learning



In this talk ...

- There are also **privacy-enhancing paradigms and execution modes** that do not necessarily have formal guarantees
- These methods are designed to **limit access to raw data**, but provide **no worst-case guarantees**:
 - Federated Learning
 - Split Learning
 - **Privacy Regularizers**

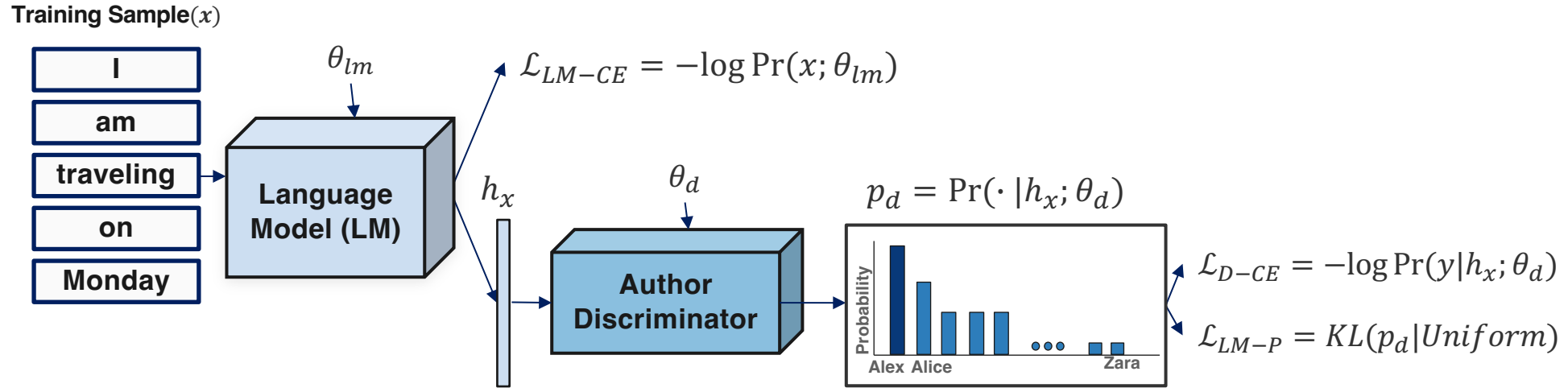
Joint Optimization Insight

To remove sensitive information, we first need to find them. To find secrets, we use a **proxy**.

If a string can be used to identify its **writer**, that string contains a **secret**.

We can define different type of secrets, if we want to protect different **attributes**.

Regularization I: Adversarial Learning



Adversarial Training:

LM Optimization:

$$\min_{\theta_{lm}} \mathcal{L}_{LM-CE}(x; \theta_{lm})$$

Utility Term

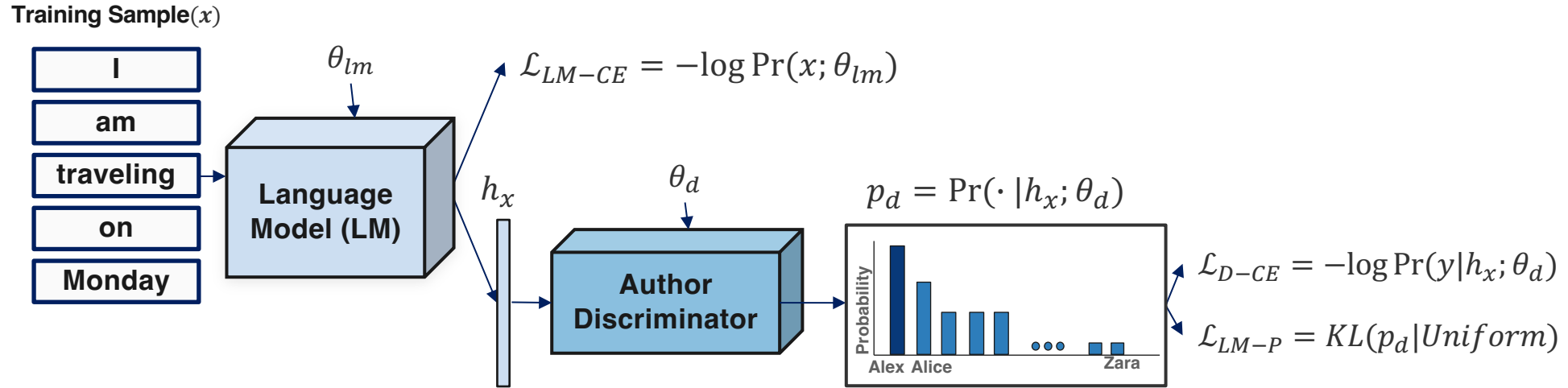
$$+ \lambda \mathcal{L}_{LM-P}(h_x; \theta_d)$$

Privacy Term

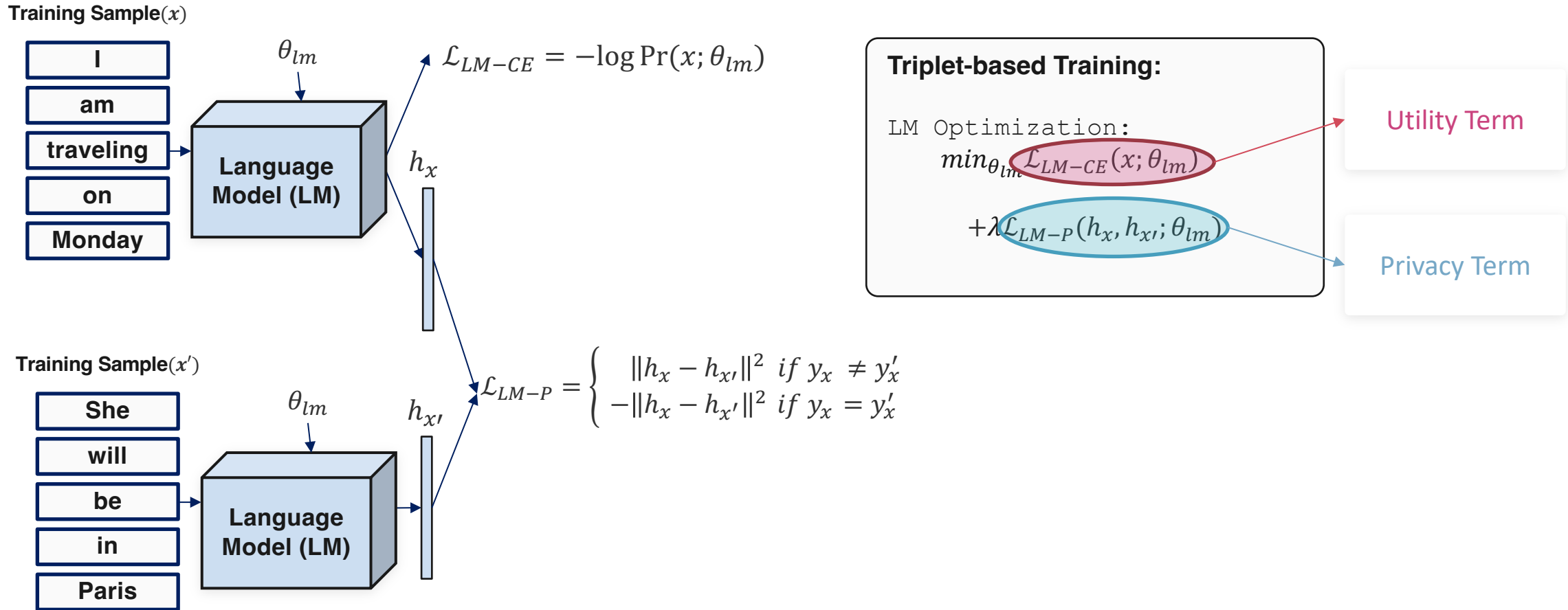
Discriminator Optimization:

$$\min_{\theta_d} \mathcal{L}_{D-CE}(h_x, y; \theta_d)$$

Regularization 2: Triplet-based Loss



Regularization 2: Triplet-based Loss

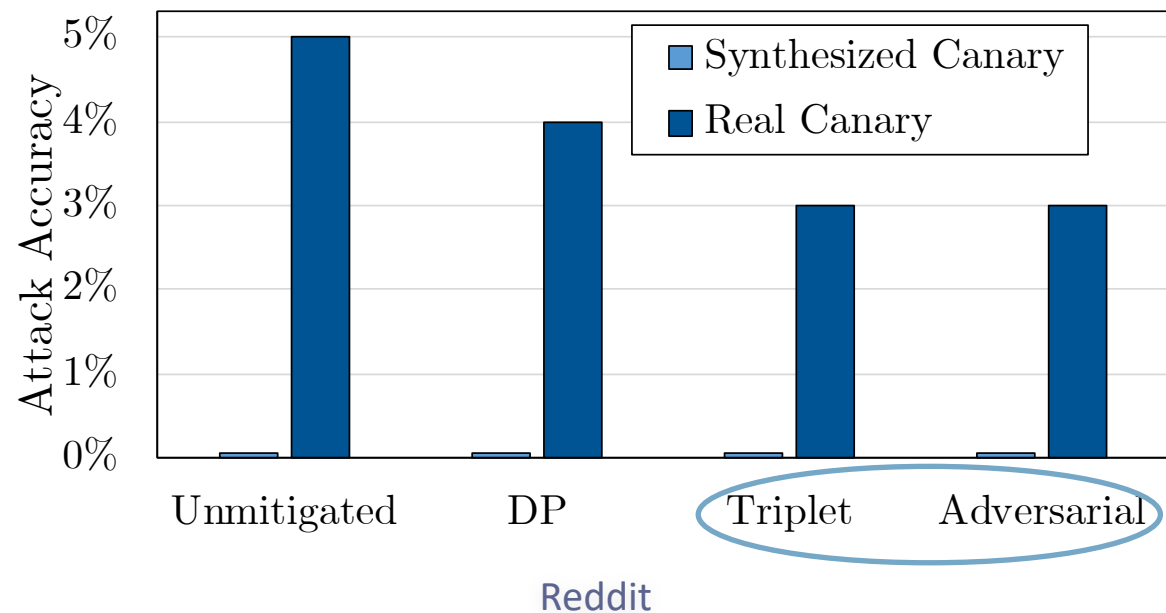
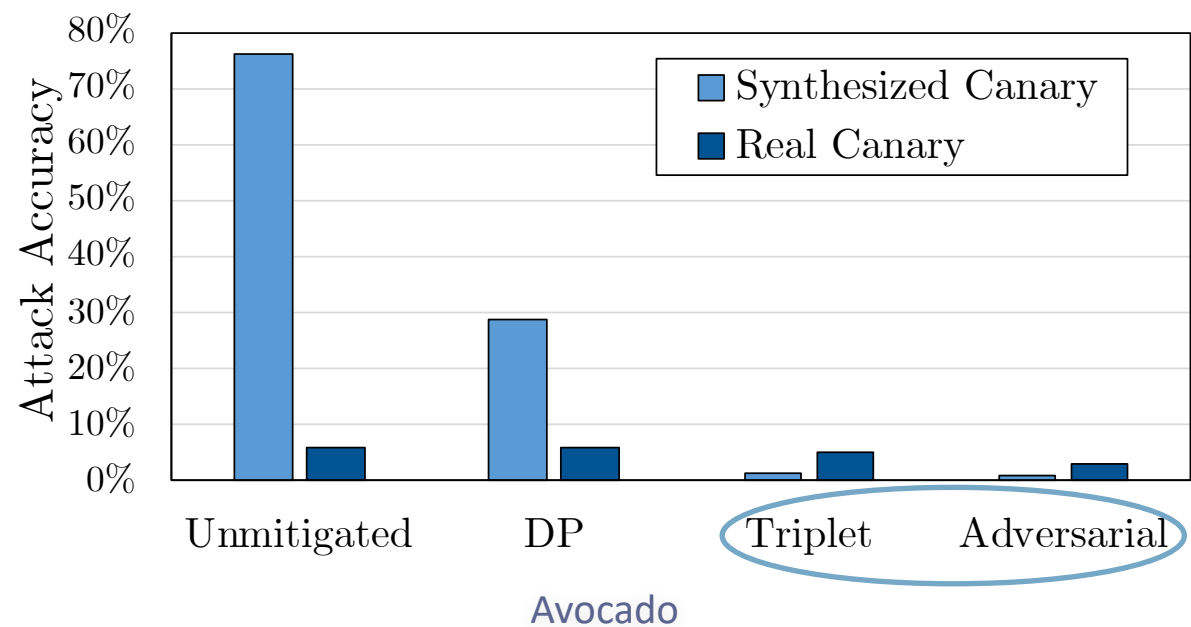


Tab attack

I will meet } Alice at the docks



Tab Attack



Our regularizations are more effective than differential privacy in thwarting the attack.

Summary and Discussion

- We can enhance privacy by **limiting the raw data** that is shared with other parties
 - Federated Learning
 - Split Learning
- We can enhance privacy by defining sensitive attributes and trying to **limit their memorization**
 - Regularization

Thank you!