Privacy-Preserving Domain Adaptation of Semantic Parser

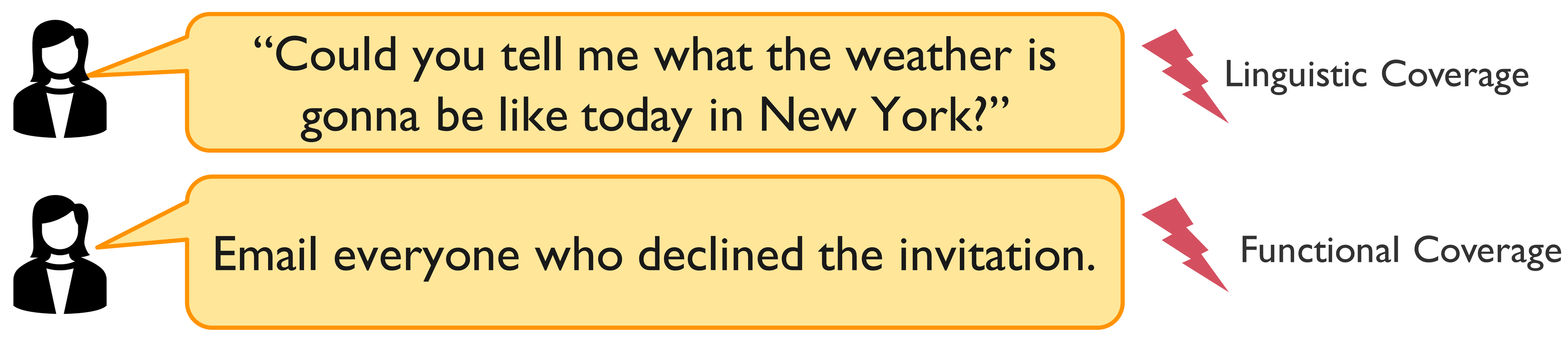
UC San Diego
Microsoft
semanticmachines

Fatemehsadat Mireshghallah, Yu Su, Tatsunori Hashimoto, Jason Eisner, Richard Shin

(1) Problem Definition

Task-oriented dialogue systems often assist users with personal or **confidential** matters. So:

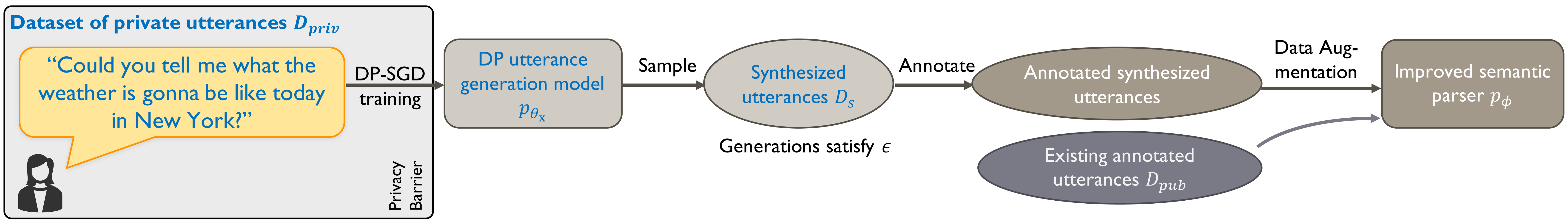
- Data is private and practitioners are not allowed to look at it
- How can we know where the system is failing and **needs more training data** or **new functionality**?



- If some users are asking the system to hop up and down, fine-tuning is unlikely to make it grow legs.
- Our goal is to produce realistic data that can be inspected (so that the developers know to build legs) and expertly annotated (to rapidly teach the semantic parser that words like "hop" and "jump" should invoke the leg API)

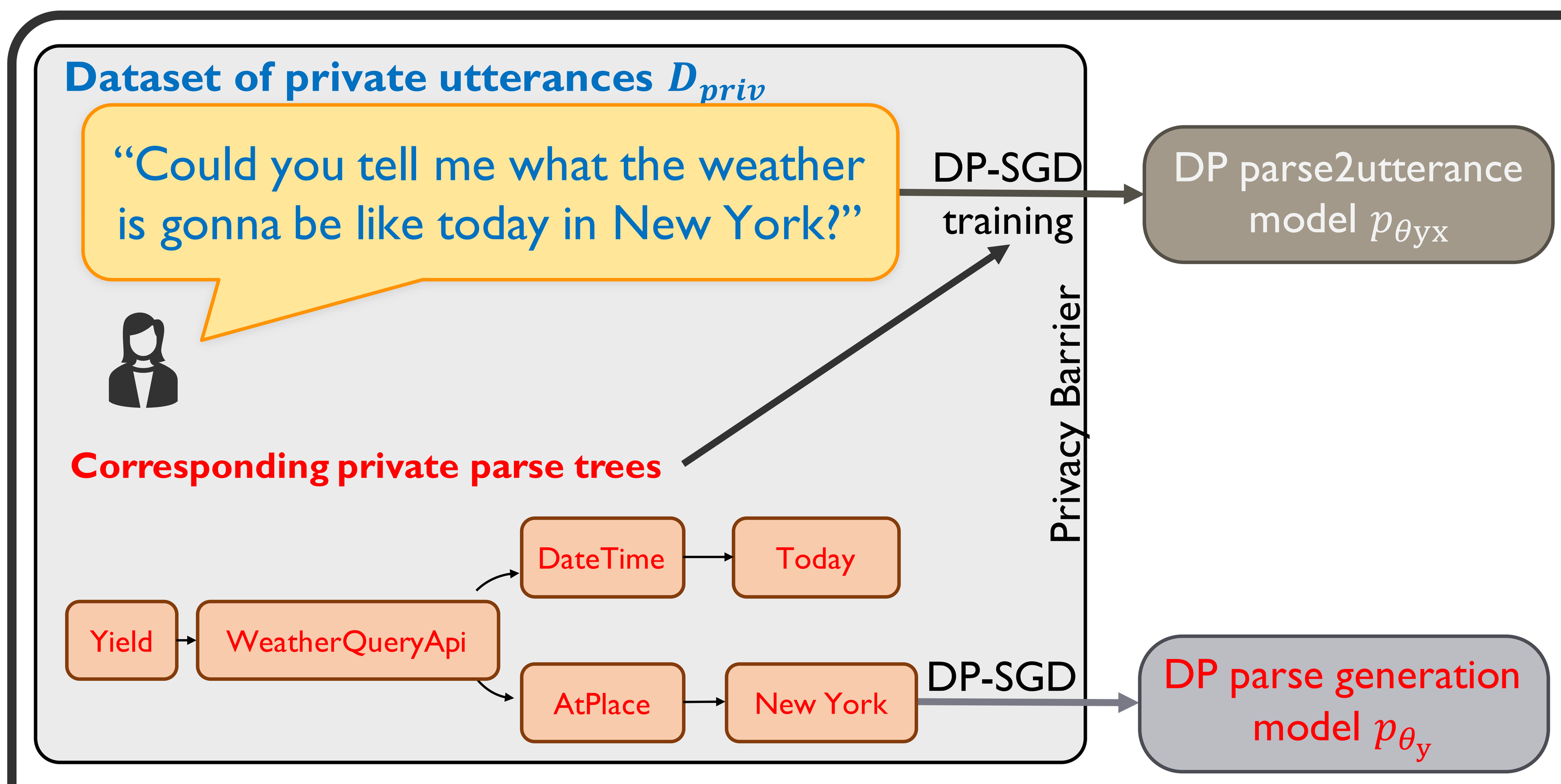
How can we privately synthesize data that is **distributionally close** to eyes-off user data?

(2) Baseline: Vanilla DP



We model $p(x)$, where x is a **private utterance:** 1-stage baseline approach of fine-tuning a pre-trained generative auto-regressive language model on **private user utterances** using **differentially private SGD**. To create the synthesized dataset we take samples from the fine-tuned model.

(3) Proposed: 2-stage DP



- Intuitive Baseline: We model $p(x)$, where x is a **private utterance**.
- Proposed: We model $p(y)$ and $p(x|y)$, where y is a **private parse-tree.**
  - one stage models the **parse-trees,** $\mathbf{p}_{\theta_y}$
  - The other stage models an **utterance** given a **parse-tree,** $\mathbf{p}_{\theta yx}$

(4) Experimental Results

Comparison with Baselines

		Language Metrics		Parse Metrics	
		W. Overlap	Mauve	Distance	F. Overlap
No DP	Baseline	0.087	0.334	0.258	0.487
	Ours	**0.236**	**0.632**	**0.085**	**0.797**
$\epsilon = 8$	Baseline	0.093	0.198	0.183	0.487
	Ours	**0.210**	**0.533**	**0.055**	**0.707**
$\epsilon = 3$	Baseline	0.086	0.138	0.185	0.485
	Ours	**0.205**	**0.530**	**0.054**	**0.693**

We can see that the proposed 2-stage method **outperforms the 1-stage baselines**, at all levels of privacy budget.

Ablation study

	Method	MAUVE	Distance
Few-modes	Baseline	0.23	0.24
	Ours	0.21	0.10
Full-modes	Baseline	0.33	0.25
	Ours	0.63	0.08

The effect of using data with few modes for training vs. the full dataset, on the performance of the 1-stage baseline and the proposed 2-stage method. The goal is to see if the superiority of the 2-stage method is due to it **better capturing different modes in the data**.

We simulated a situation where users are asking about the **weather** but the original semantic parser **was not trained on weather-related functions**:

1. We created the original semantic parser by training on $\frac{1}{10}$ of our data (SMCalFlow), **excluding** any examples that use **weather-related functions**.
2. We treated the other $\frac{9}{10}$ of the data as **private user utterances**, **including** those requesting **weather**. We created **approximate private annotations** for the private utterances, using the **original semantic parser**.
3. We apply the baseline and proposed methods to create **public synthesized datasets,** which **include weather functions**.
4. We simulated high-quality human annotation of the public synthetic utterances. We **re-train** the parser with this additional annotated data.

Downstream Experiment: Adding Weather Functionality ($\epsilon = 3$)

	Anonymized Graph Match	API Recall
Full Dataset	76.6	77.5
Non-augmented	0	2.1
Baseline	37.7	42.1
Ours	**43.7**	**50.3**