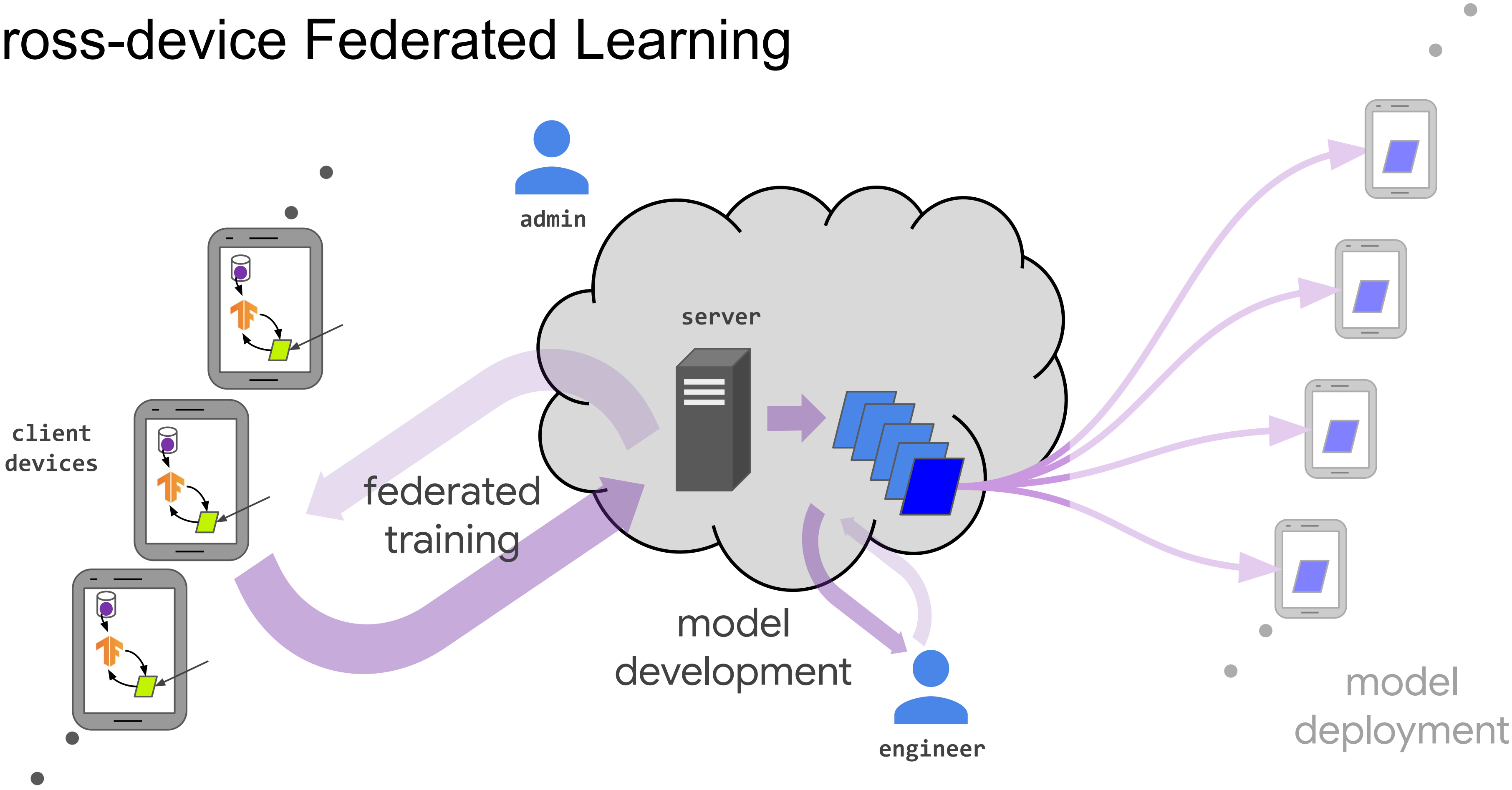
Federated Learning of Gboard Language Models with Differential Privacy

Zheng Xu
Google Research
Presenting the work of many

Cross-device Federated Learning

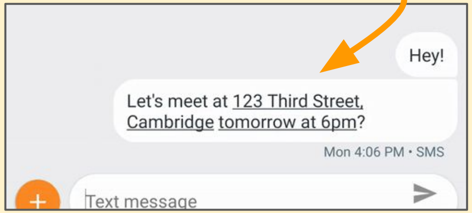


FL & FA at Google

Federated learning and analytics are deployed in an increasing array of apps and services.

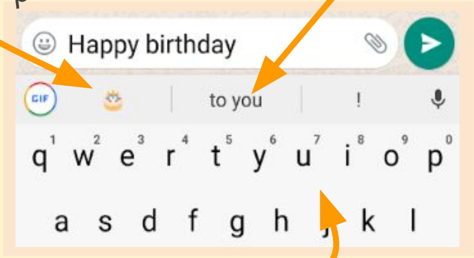
Gboard

Android
smart text selection



emoji and sticker
prediction

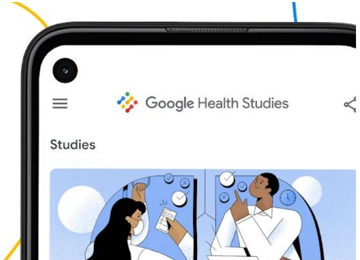
next word
prediction



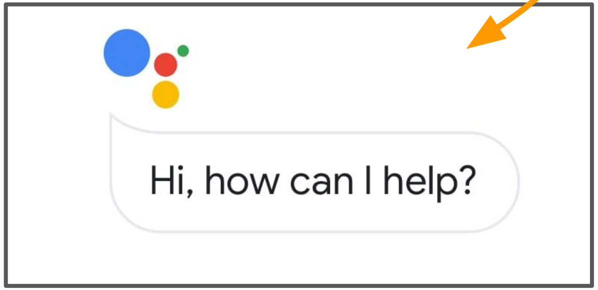
new vocabulary
discovery

Google Health
Studies

Respiratory health study



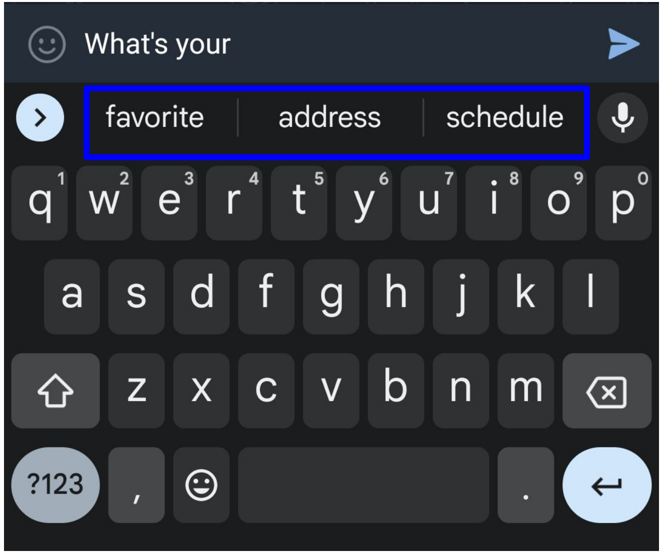
"Hey Google"
hotword detection



Gboard Language Models (LMs)

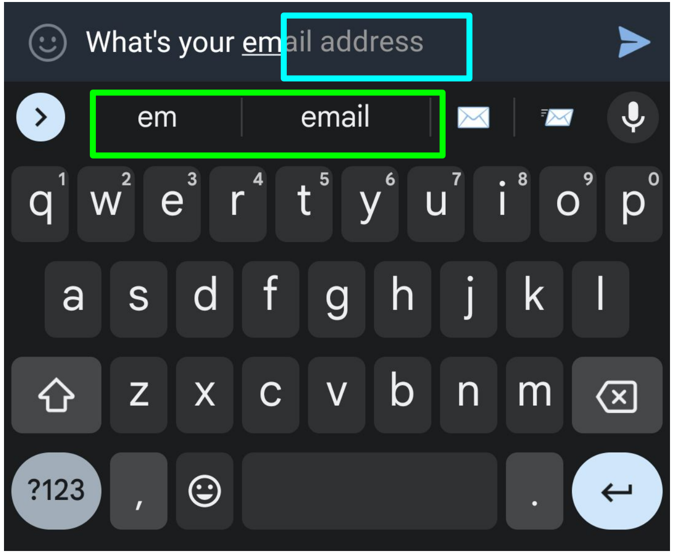
One-layer LSTM

Gboard Next Word Prediction (NWP)



NWP LM: ~2.4M / 4.4M parameters

Smart Compose (SC)
On-The-Fly Rescoring(OTF)



OTF LM: ~6.4M parameters

Privacy Principles

- Transparency and User control
  - Users can be aware of what data is used, what purpose it is used or and how it is processed, and have full control on whether to enable the collection and use of their data
- Data minimization
  - Data is only collected focusing on specific computation needs, with access limited at all data processing stages
- Data anonymization
  - The final released output of the computation does not reveal anything unique to an individual
- Auditability and verifiability
  - Users, and potentially third parties can audit and verify privacy claims by examining released models, open-sourced code, and privitarized system logs, etc.

Federated Learning and Privacy 2022
Private Federated Learning in Gboard 2023

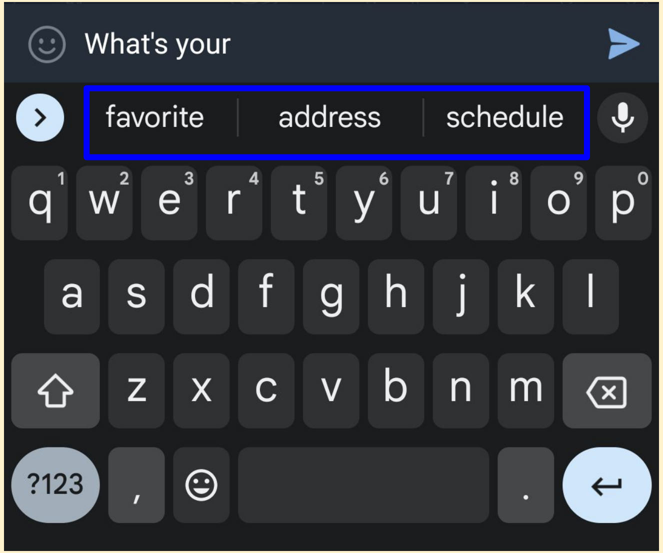
Privacy Principles in Gboard

- Transparency and User control
  - Users can be aware of what data is used, what purpose it is used or and how it is processed, and have full control on whether to enable the collection and use of their data
  - Users can turn off learning at any time
- Data minimization
  - Data is only collected focusing on specific computation needs, with access limited at all data processing stages
  - Federated Learning (and Secure Aggregation)
- Data anonymization
  - The final released output of the computation does not reveal anything unique to an individual
  - Differential Privacy
- Auditability and verifiability
  - Users, and potentially third parties can audit and verify privacy claims by examining released models, open-sourced code, and privitarized system logs, etc.

Federated Learning and Privacy 2022
Private Federated Learning in Gboard 2023

Journey of (Differential) Privacy for Gboard NWP

Gboard Next Word Prediction (NWP)



2023+ stronger DP guarantees for ***all neural* models**, with possibly additional Trusted Aggregation; all future LMs require DP to be launched

2022 Meaningful differential privacy (DP) guarantee for the spanish model. **First** production neural network trained directly on user data announced with DP.

2020 Clipping and adding noise to prevent memorization. Launched and **empirically audited** for spanish model in spain.

Federated Learning (FL)

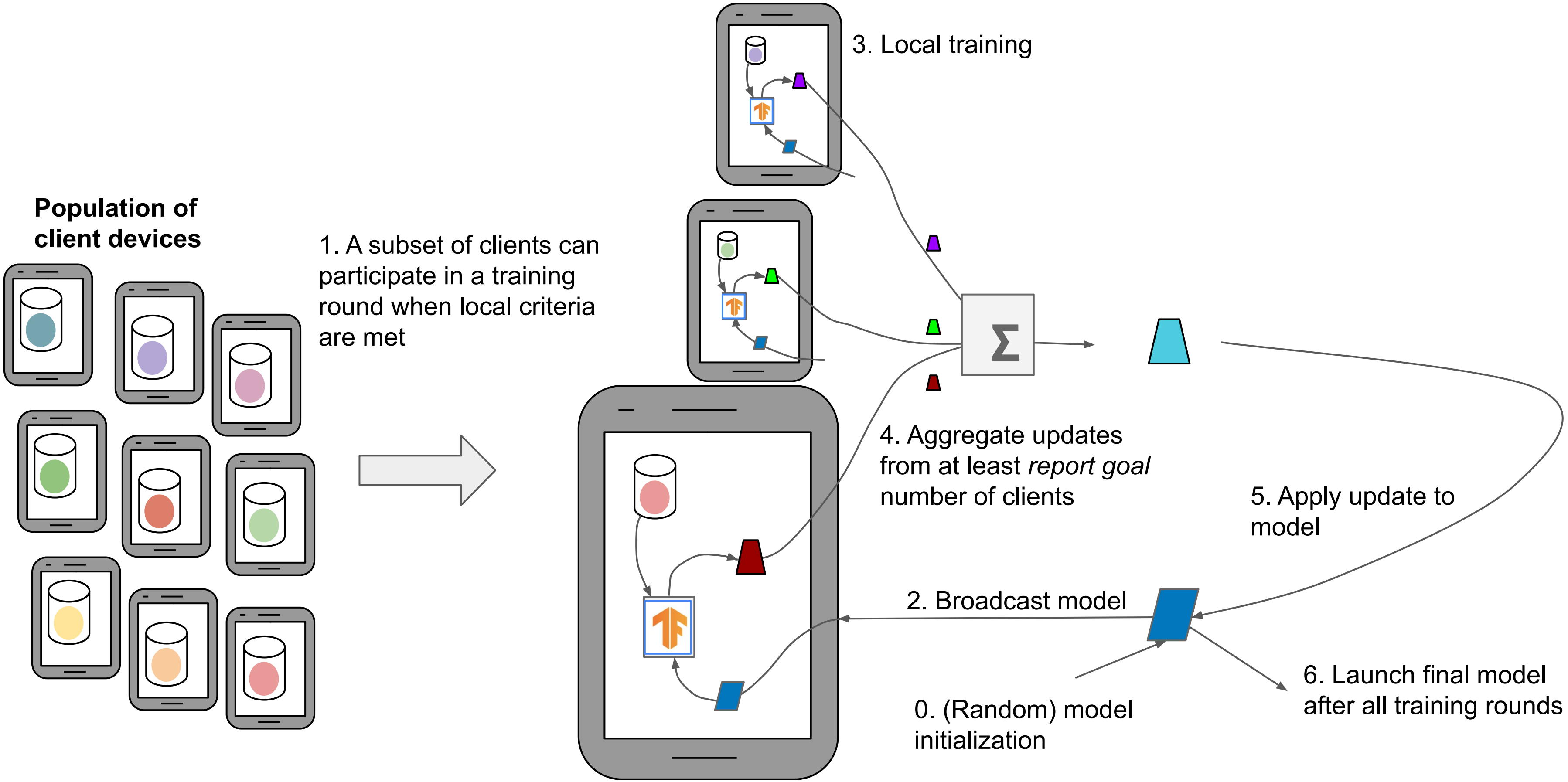
Tier 1.5: Towards strong formal privacy guarantees; ε < 1?

Tier 2: Realistic privacy guarantees; ε < 10

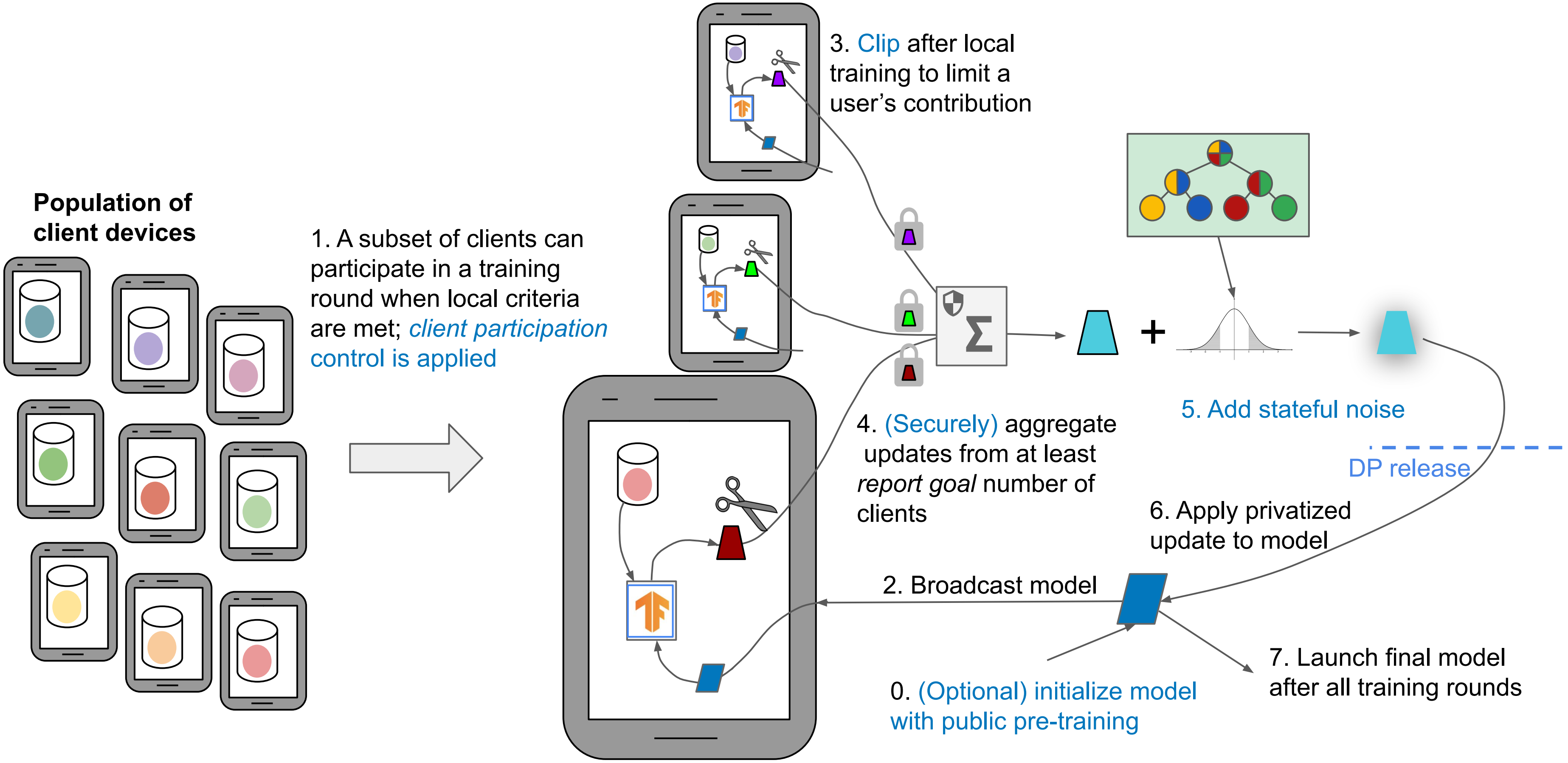
Tier 3: Weak to no formal privacy guarantees; finite eps with empirical auditing

Section 5.2.2 Target ε guidelines for ML models,
[How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy]()

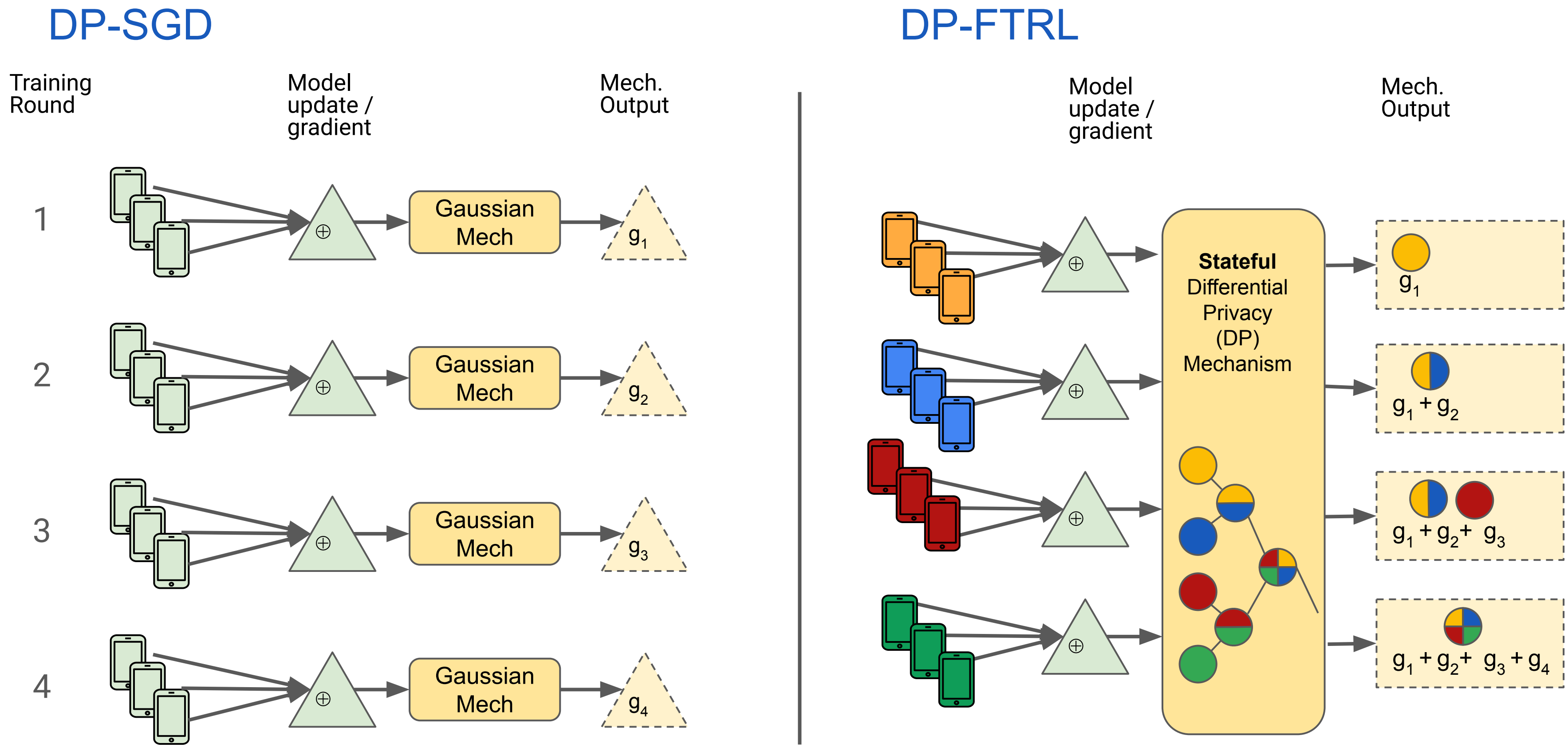
Basic Federated Learning



Private Federated Learning



Algorithm System Co-design



Practical and Private (Deep) Learning without Sampling or Shuffling 2021 https://arxiv.org/abs/2103.00039

New (Best) Practices

- Pre-train the model with public data (C4)
- Choose the maximum noise multiplier that meets the utility target based on small scale simulation experiments on public datasets that is similar to the production task
- Linearly increase the report goal and noise multiplier to meet the privacy target, and choose a large report goal supported by the system and population dynamics.
- Estimate the possible maximum min separation based on chosen report goal and estimated population, and configure the client timer period to approach the desired min separation
- DP-FTRL training with hyperparameters
  - Apply DP-FTRL with adaptive clipping without manual tuning to try meet the privacy and utility goals
  - For reliable optimization and stronger privacy-utility trade-offs, run DP-FTRL with adaptive clipping once to estimate clipping norm and then fix it

Federated Learning of Gboard Language Models with Differential Privacy 2023 https://arxiv.org/abs/2305.18465

Application in Production Gboard Language Models: Utility

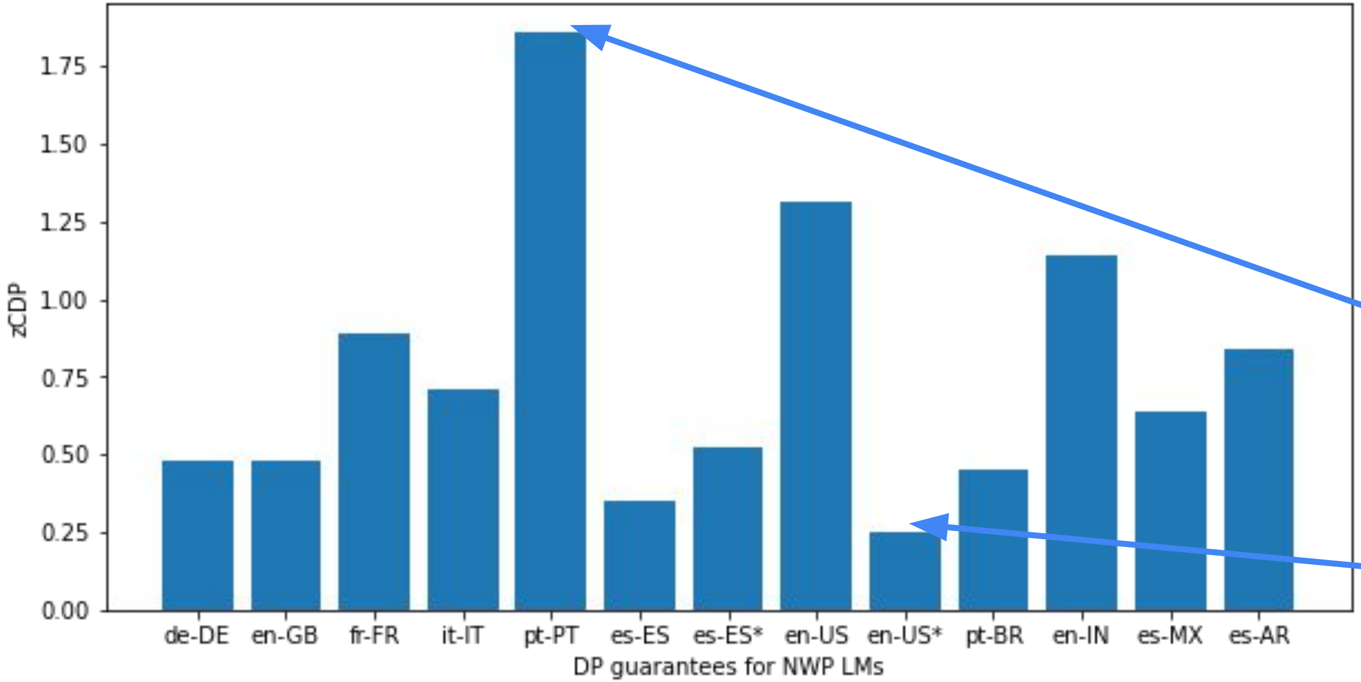
- Strong utility in A/B testing
  - Better than N-Gram base models
  - Comparable with no-DP neural models (strong baselines)

- **Scale through computation is the key to achieve strong privacy-utility trade-off**
  - **6500 client devices per round**

Federated Learning of Gboard Language Models with Differential Privacy 2023 https://arxiv.org/abs/2305.18465

Application in Production Gboard Language Models: DP

2.6-zCDP,
US Census
Bureau



alternatively,
(ε=13.69,
δ=10-^{10})-DP

alternatively,
(ε=4.49,
δ=10^{-10})-DP

Federated Learning of Gboard Language Models with Differential Privacy 2023 https://arxiv.org/abs/2305.18465

Reporting DP guarantees

- DP setting: central DP with honest but curious server
- DP definition:
  - **Data accesses covered:** The DP guarantee applies to all well-behaved clients in a single training run. We do not account for hyperparameter tuning, or the selection of the final model checkpoint using evaluation metrics or A/B testing in our guarantees. Public multilingual C4 data is used for pre-training.
  - **Final mechanism output:** Only the final model checkpoint is released for production launches, however all intermediate models are protected (including those sent to devices participating in federated learning).
  - **Unit of privacy:** Device-level DP; the device might have an arbitrarily large local dataset containing arbitrary training examples. For user's with a single device, this corresponds directly to user-level DP.
  - **Adjacency definition for "neigbouring" datasets:** zero-out; in the absence of a client at any training step, we assume that the client's model update gets replaced with the all zeros vector.
- Privacy accounting details
  - **Type of accounting:** ρ−zCDP and (ϵ, δ)−DP
  - **Accounting assumptions:** there are at least a min-separation number of rounds between two consecutive participation of a client that is enforced by a timer on clients in the cross-device FL system
  - **The formal DP statement:** ρ−zCDP range in (0.2, 2), corresponding ϵ for (ϵ, δ)−DP in (4, 14) when $\delta = 10^{-10}$
  - **Transparency and verifiability:** open-source code

Section 5.3 Target ε guidelines for ML models,
How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy

Open-source Code for DP FL

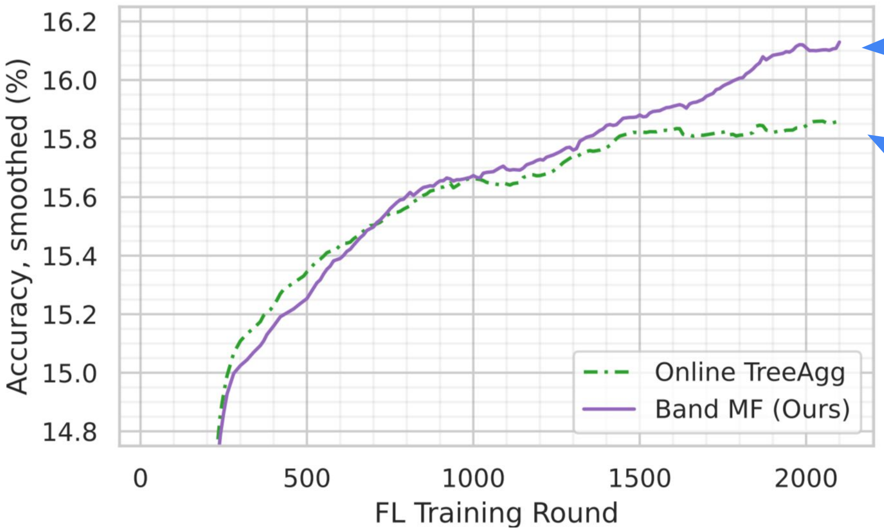
- TFF aggregator https://github.com/tensorflow/federated/blob/main/tensorflow_federated/python/aggregators/differential_privacy.py
- TFP DPQuery https://github.com/tensorflow/privacy/blob/master/tensorflow_privacy/privacy/dp_query/tree_aggregation_query.py
- DP accounting https://github.com/google-research/federated/blob/master/dp_ftrl/blogpost_supplemental_privacy_accounting.ipynb
- FL system https://github.com/google/federated-compute

Federated Learning of Gboard Language Models with Differential Privacy 2023 https://arxiv.org/abs/2305.18465

Can we do even better?

User-level DP

Gboard Spanish LM



(**ε=4.35**, $\delta=10^{-10}$)-DP

(ε=6.69, $\delta=10^{-10}$)-DP

(Amplified) Banded Matrix Factorization: A unified approach to private training 2023 https://arxiv.org/abs/2306.08153

Takeaways

Thank you!

- (Differential) privacy is achievable in practice
  - Through system algorithm co-design
  - Scale is the key: large amount of data and computation resources
  - Improving privacy-utility trade-off by public data, new algorithms, DP mechanism and accounting
- Privacy is not “free”
  - Computation and infrastructure support
  - Common understanding of the techniques: verifiable, auditing
  - Engineering efforts / migration cost