Introduction to Federated Learning

Kirill Acharya[1, 2]

[1]Department of Applied Mathematics and Informatics
Moscow Institute of Physics and Technology

[2]Yandex School of Data Analysis

AGI House Phystech 2023, October 15 2023

Table of Contents

1 Introduction

2 Challenges and Optimization

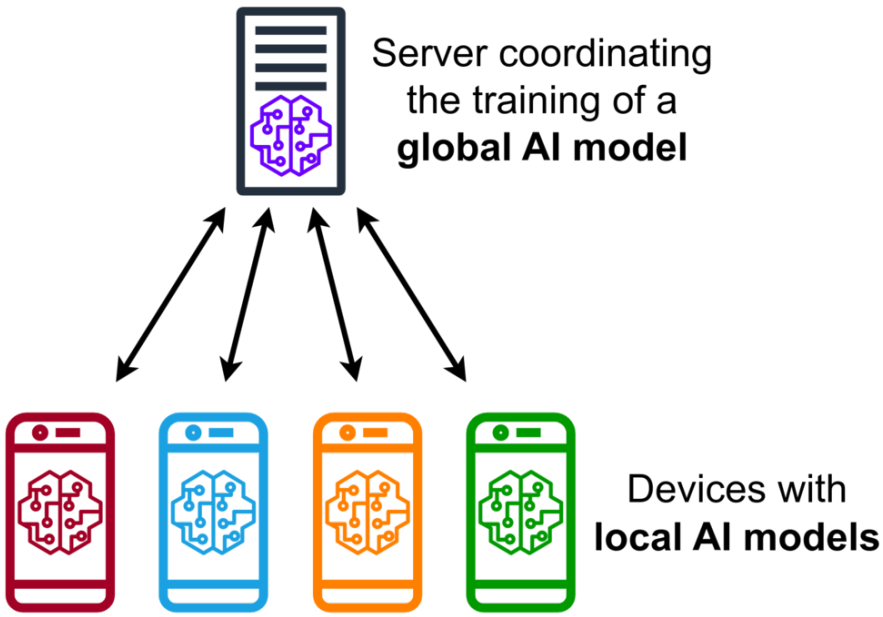
3 Current State of Technology

4 Sources and Q&A

Motivation

- Data privacy is a great concern nowadays and the trend is increasing.
- Goal: enable edge devices to do State-of-the-art machine learning without centralizing data and with privacy by default. Here are some cases when the technique could be used:
  - On-device data is more relevant than server-side proxy data
  - On-device data is privacy sensitive or large
  - Labels can be inferred naturally from user interactions

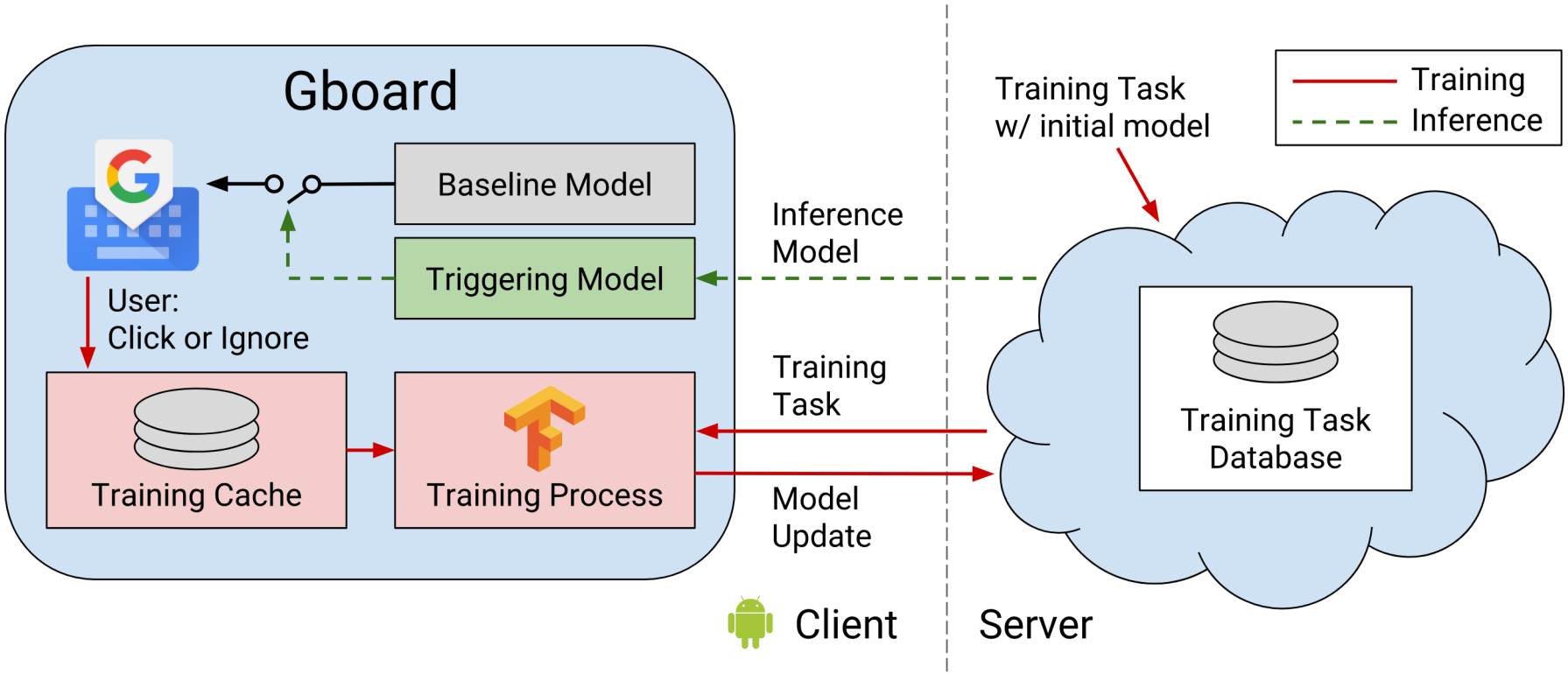
Federated learning



Privacy Technology by Google:

- On-device datasets: keeping raw data local
- Federated agreggation: combine reports from multiple devices
- Federated model averaging: many steps of gradient descent on each device

Example of Federated Learning Framework



Source: Applied Federated Learning: Improving Google Keyboard Query Suggestions

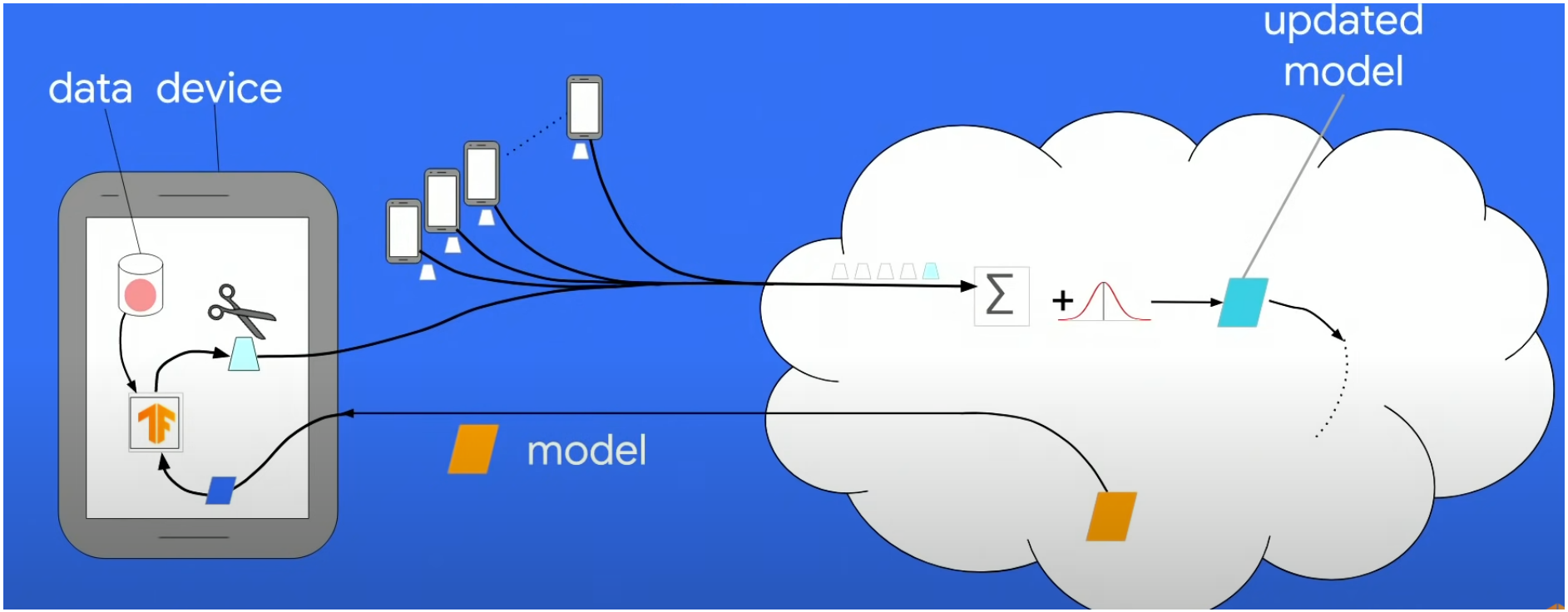
Privacy principles

Here are the Privacy principles by Google:

- Only-in aggregate: engineer may only access combined device reports
- Ephemeral reports: never persist per-device reports
- Focused collection: devices report only what is needed for this computation
- Don't memorize individuals' data
- Differentially private model averaging

Differential privacy

Differential privacy: statistical science of learning common patterns in a dataset without memorizing individuals examples.
Idea: use noise to obscure individual impact on the learned model. See project



Source: DIfferentially Private Recurrent Language Models

Table of Contents

1 Introduction

2 Challenges and Optimization

3 Current State of Technology

4 Sources and Q&A

Challenges and Optimization

- Distributed optimization methods/machine learning methods require efficient organization of communications, since communications in this case very often take up most of the time of the algorithm. worker devices use unstable and slow networks such as Wi-Fi and Cellular.
- Intermittent compute node availability
- Intermittent data availability

Problem statement

- We consider optimization problems of the form

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\},$$

where $x \in \mathbb{R}^d$ collects the parameters of a statistical model to be trained, n is the number of workers/devices, and $f_i(x)$ is the loss incurred by model x on data stored on worker i.

- A general baseline for solving problem is distributed gradient descent, performing updates of the form

$$x^{k+1} = x^k - \frac{\eta^k}{n} \sum_{i=1}^{n} \nabla f_i \left(x^k \right),$$

where $\eta^k > 0$ is a stepsize.

Compressors review

- **Paper:** On Biased Compression for Distributed Learning (Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, Mher Safaryan)
- **Main contribution:** Distributed SGD with Biased Compression and Error Feedback Algorithm

Definition

Top-k

$$\mathcal{C}(x) := \sum_{i=d-k+1}^{d} x_{(i)} e_{(i)}$$

where coordinates are ordered by their magnitudes so that $|x_{(1)}| \leq |x_{(2)}| \leq \cdots \leq |x_{(d)}|$.

Definition

Error Feedback $e_i^{k+1} = e_i^k + \nabla f_i(x^k) - C(e_i^k + \nabla f_i(x^k))$

Compressors review

- **Paper:** Permutation Compressors for Provably Faster Distributed Nonconvex Optimization (Rafał Szlendak, Alexander Tyurin, Peter Richtárik)
- **Main contribution:** Construction of the new compressors based on the idea of a random permutation (Perm K).
  Provably reduce the variance caused by compression beyond what independent compressors can achieve.

Definition

(Perm K for $d \geq n$). Assume that $d \geq n$ and $d = qn$, where $q \geq 1$ is an integer. Let $\pi = (\pi_1, \ldots, \pi_d)$ be a random permutation of $\{1, \ldots, d\}$. Then for all $x \in \mathbb{R}^d$ and each $i \in \{1, 2, \ldots, n\}$ we define

$$\mathcal{C}_i(x) := n \cdot \sum_{j=q(i-1)+1}^{qi} x_{\pi_j} e_{\pi_j}.$$

Experiments

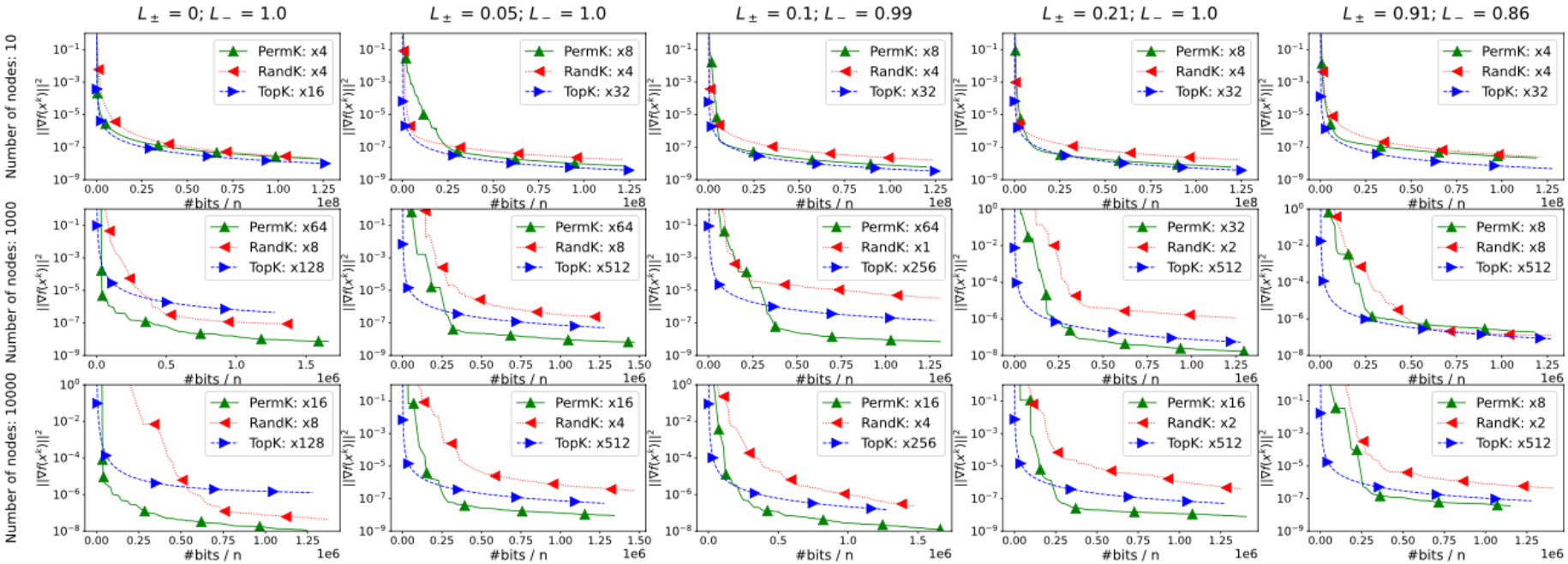


Figure 1: Comparison of algorithms on synthetic quadratic optimization tasks with nonconvex $\{f_i\}$.

Source: Permutation Compressors for Provably Faster Distributed Nonconvex Optimization

Table of Contents

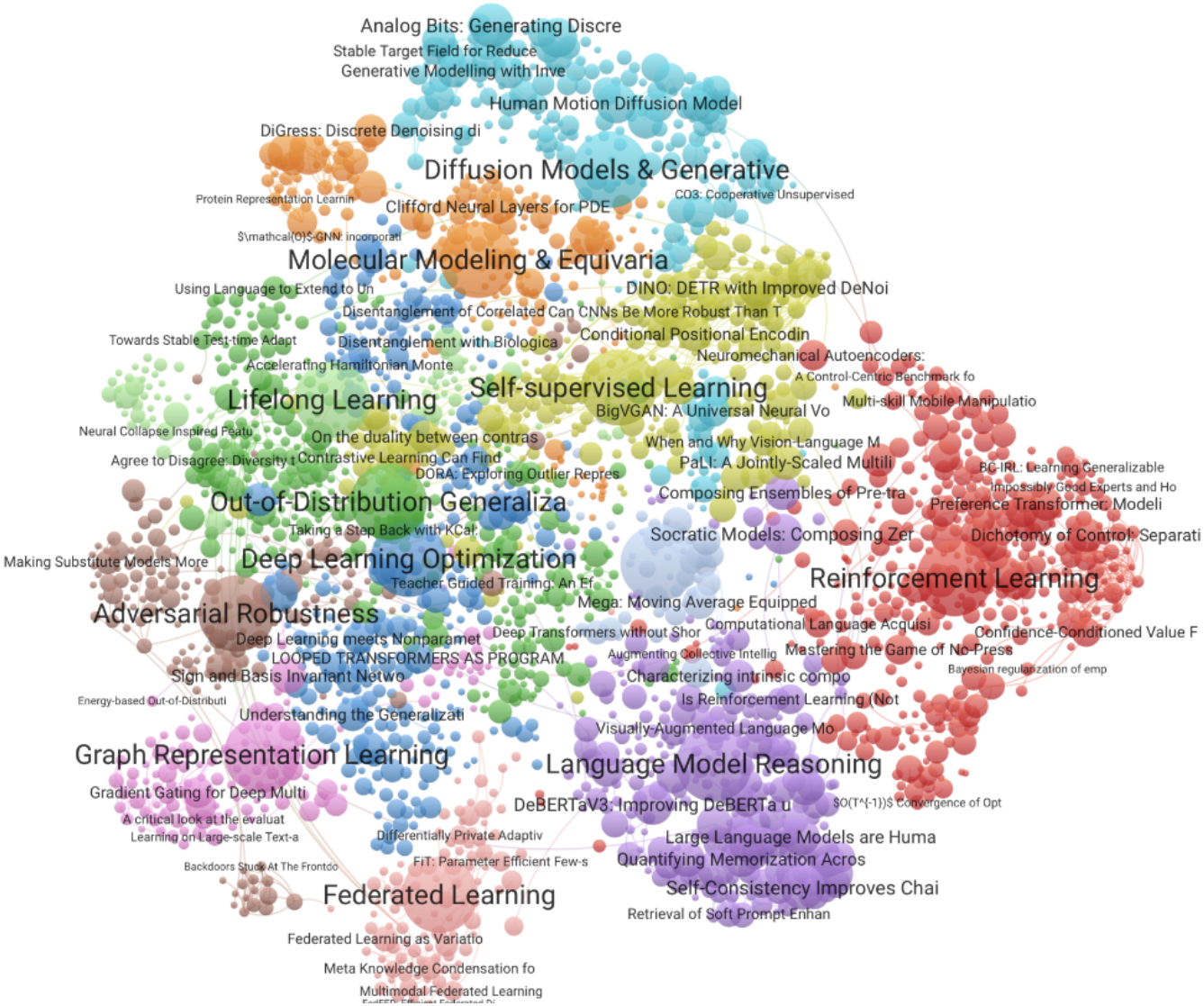
1 Introduction

2 Challenges and Optimization

3 Current State of Technology

4 Sources and Q&A

Research areas



Source: ICLR 2023 papers

Use cases

- (Big) Tech
  - Machine Learning Research at Apple
  - Federated learning team at Google AI.
  - NVIDIA Federated Learning Application Runtime Environment
- Healthcare
  - The future of digital health with federated learning, Author's presentation
  - Automated Pancreas Segmentation Using Multi-institutional Collaborative Deep Learning
- Finance
  - Using Federated Learning to Bridge Data Silos in Financial Services
- and many other fields

Table of Contents

1. Introduction
2. Challenges and Optimization
3. Current State of Technology
4. Sources and Q&A

Notable sources

Here is the list of notable sources on Federated Learning:

- Federated Learning One World Seminar (FLOW)
- Privacy Preserving AI (Andrew Trask) — MIT Deep Learning Series
- Flower Lectures and Summits
- Open-sourced Federated Learning Framework

Q&A