Navigating the Tradeoff Between Privacy and Fairness in ML

Jesse Cresswell, PhD
Senior Machine Learning Scientist
Layer 6 AI at TD

Nov. 29 2022 - TMLS

layer 6
AI at TD

Agenda

- Motivation
- Review of Fairness and Privacy
- Fair Federated Learning
- Fair Differential Privacy
- Conclusion

layer 6
AI at TD

ML in Regulated Industries

Developing and deploying ML in highly regulated industries comes with unique challenges:

- Explainability
- Robustness
- Stability
- Reproducibility

Two aspects are especially relevant for financial institutions:

- Privacy
- Fairness

Every FI has teams specifically dedicated to these pillars.

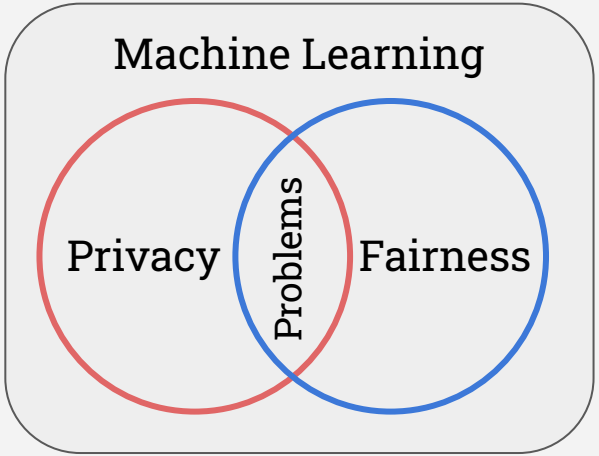
Privacy and Fairness

Both privacy and fairness are rich areas of study in ML.

There are many valid perspectives on each,
no universally accepted frameworks or “correct” answers.

However, almost all discussions cover them separately.

Their intersection is surprisingly problematic…



Applying common privacy enhancing technologies makes unfairness worse.

Review: Fairness in ML

Fairness can be rigorously defined in various mathematical frameworks, e.g

- **Demographic Parity**
  Positive outcomes equally shared across groups.
- **Equal Odds**
  Probabilities of all outcomes independent of group.
- **Equal Opportunity**
  Positive outcome rates are independent of group for those that qualify.

Many definitions are reasonable. Any given scenario may require a different notion.

However, many fairness definitions are simultaneously incompatible with each other

Demographic Parity and Equal Odds typically cannot both hold.

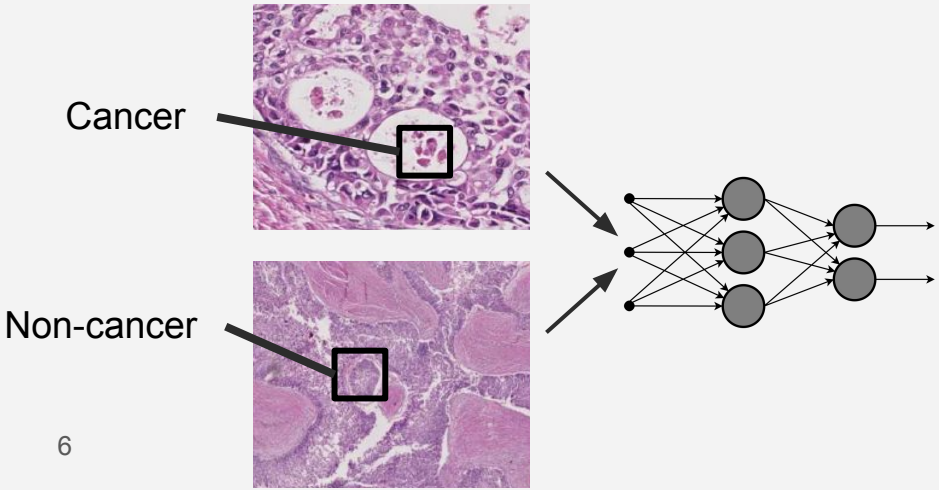
Review: Fairness in ML

For this talk we consider a simple metric:

- **Performance Parity**

  A model's performance (e.g. accuracy) should be the same across groups.

Example: Cancer prediction in histopathology images



Gender	Accuracy
M	89.2%
F	89.4%
Other	88.9%

layer 6
AI at TD

Review: Fairness in ML

Complications:

- Some groups are minorities, we have less training data for them.
- Some groups are more complex, inherently harder to predict correctly.

Possible Solutions:

- Repeat the minority group more often (oversampling), or weight data more highly.
- Modify loss function to pay attention to group label.
- Adjust model outputs to be more equitable.

All such solutions **require knowing the group labels** for each datapoint and acting on that information (disparate treatment).
Group labels are often sensitive - may be prohibited from collecting them in practice.

Review: Privacy in ML

Privacy is often managed at institutions by limiting access to data, and removing sensitive information. This is the focus of most privacy regulations.

The “anonymization” paradigm - either data has been anonymized or it has not.

Banks and credit bureaus use anonymization to share customer information. Hospitals release medical data and diagnoses without personal information.

Name	Age	Job	Salary
[redacted]	23	Clerk	50,000
[redacted]	45	Driver	60,000
[redacted]	61	Lawyer	100,000

layer 6
AI at TD

Review: Privacy in ML

Linkage Attacks

It can be possible to de-anonymize records using outside information.

Netflix released data on sparse user movie ratings - de-anonymized by comparing to public ratings on IMDB. [Narayanan & Shmatikov 2008]

Name	Age	Job	Salary
[redacted]	23	Clerk	50,000
[redacted]	45	Driver	60,000
[redacted]	61	Lawyer	100,000

Name	Age	Job	Favorite Sport
Joe	[redacted]	Clerk	Squash
Jun	[redacted]	Driver	Soccer
Jaya	[redacted]	Lawyer	Hockey

Review: Privacy in ML

Privacy should be about protecting information belonging to individuals. In many cases we want to **reveal general information**, but cannot expose personal data.

The idea behind private data analysis is similar in spirit to the goal of ML:

ML: Train a model that learns to generalize, and does not overfit to individual data.

PDA: Extract high-level information/statistics about data, not low-level specifics.

The PDA paradigm - privacy is consumed as we extract more information from data.

How can we extract the most useful information while consuming the least privacy?

Privacy Enhancing Technologies

The aim of PETs are to **minimize the risk** to individuals that their personally identifiable data will be exposed, while **maximizing the utility** of that data for analysis.

Open data is useful, but not private. Siloed data is safe, but not useful.

Four emerging PETs actively being researched in ML:

Federated Learning	**Differential Privacy**
Secure Multi-Party Computation	Homomorphic Encryption

Fairness in Federated Learning

layer 6
AI at TD

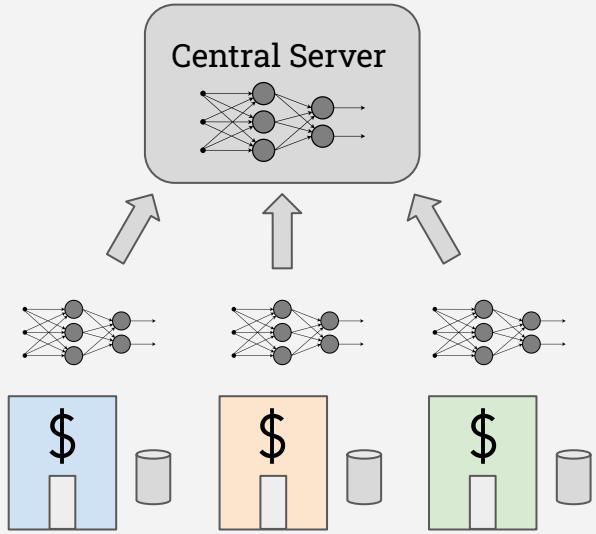
Federated Learning

The power of machine learning techniques scales with the amount and diversity of available data.

Federated Learning (FL) is a distributed ML approach where data is not pooled together on a centralized server.

Models are trained at the device/institution where data is collected.

Intuitively this is more private, since the raw data never leaves the device/institution where it was generated.

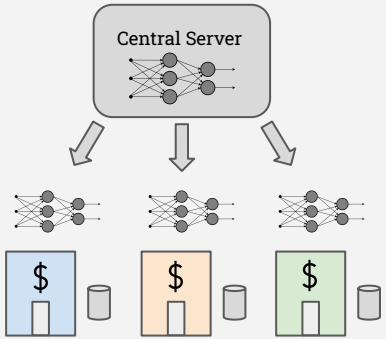


layer 6
AI at TD

Federated Learning - *FedAvg* [McMahan et al. 2017]

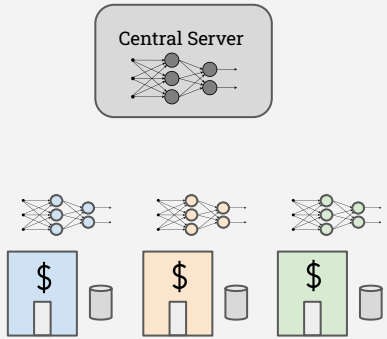
Repeat until convergence:

1.



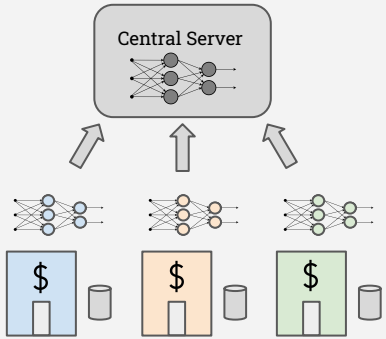
The central model is shared to each institution.

2.



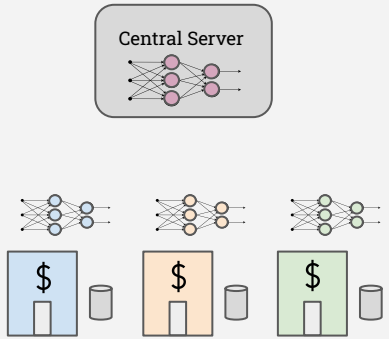
Institutions locally train the model on their data.

3.



Model updates are sent to the server.

4.



The server averages the updates, and applies them to the central model

layer 6
AI at TD

Fairness in Federated Learning

The Server's Perspective:

Suppose the server wanted to train a fair model over all clients involved.

Could the server

- Choose to request updates from minorities more frequently?
- Weight updates from minorities more heavily?

No.

The server never sees raw data, including group labels.
Its role is only to aggregate and communicate, not evaluate.

Fairness in Federated Learning

The Client's Perspective:

Suppose the client wanted to help the central model be fair.

Could the client

- Rebalance its local data, or oversample its minority examples?
- Alter its training to weight data from minorities more heavily?

No.

The client only knows its local distribution which may be very different from the global distribution.
A local minority may be a global majority, or vice versa.

layer 6
AI at TD

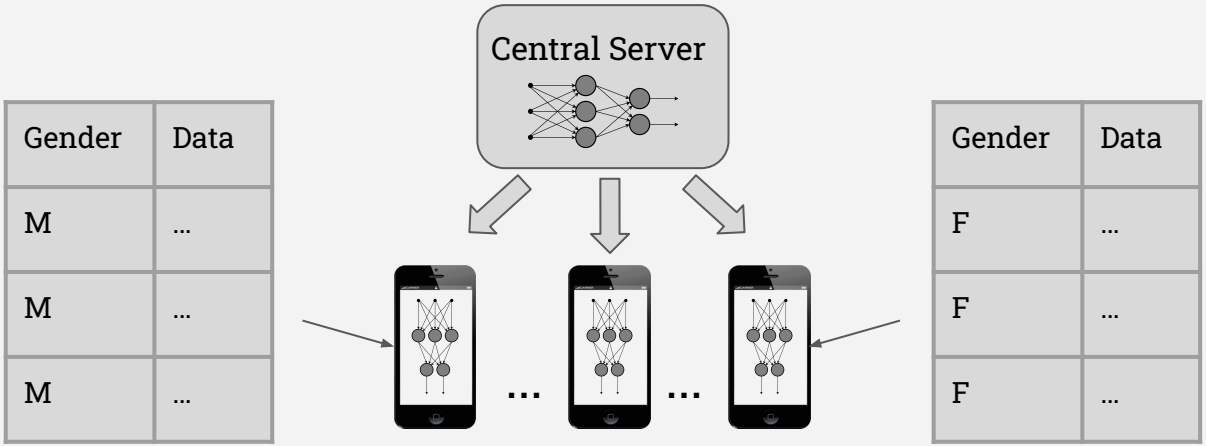
Fairness in Federated Learning

The main problem with unfairness in FL is identifying when it is occurring!

The server cannot evaluate fairness across clients.

Clients can evaluate fairness locally - but local fairness does not imply global fairness.

- Clients may have highly non-IID data, different distributions
- Clients may have data from only one or a few groups



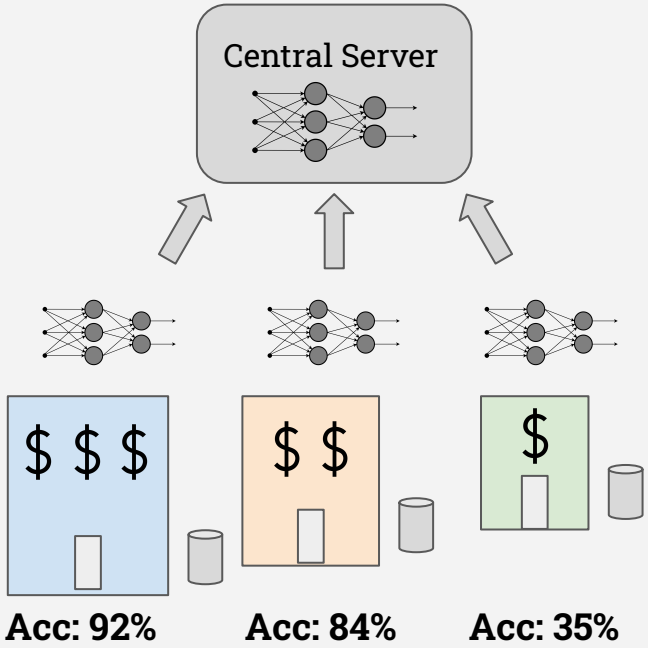
layer 6
AI at TD

Fairness in Federated Learning

Centralized training can result in a model that does well on a majority of clients, but **actively hurts a minority**.

Some clients may be better off training a model on their local data without collaboration.

The **server may not realize** some clients are being missed due to privacy constraints.



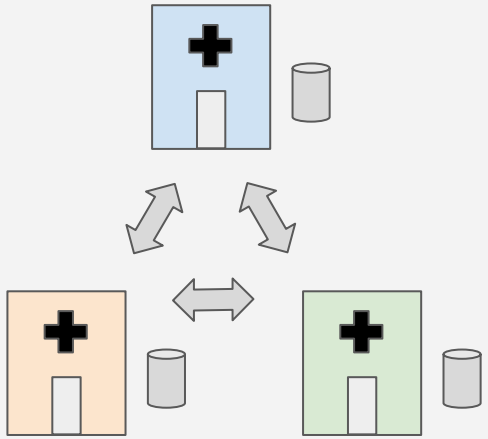
layer 6
AI at TD

Fairness in Federated Learning

Personalized FL: [Sui et al. 2022]

Each client personalizes the central model before use.

Ensures a worse global model is not used in favor of a better local model.



Decentralized FL: [Kalra et al. 2021]

Each client maintains their own model and collaborates peer-to-peer with no central server.

Clients can select peers with similar data distributions to get useful updates.

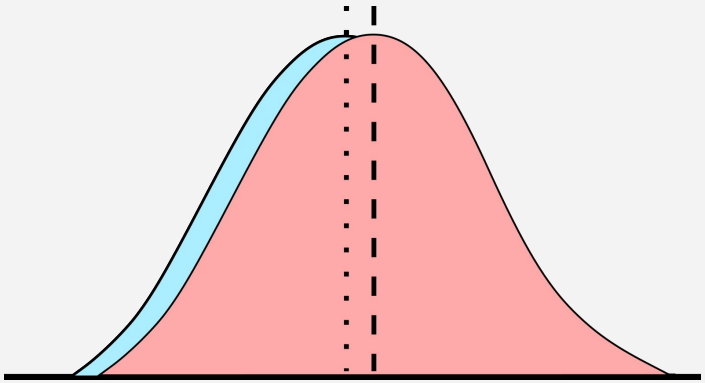
Fairness in Differential Privacy

Differential Privacy

Federated Learning makes us think “is sharing model updates **more private** than sharing raw data?”

Key point: privacy is not binary. It is a resource.

Differential privacy (DP) is a mathematical framework for **quantifying how private** some analysis is, and providing rigorous guarantees to individuals.

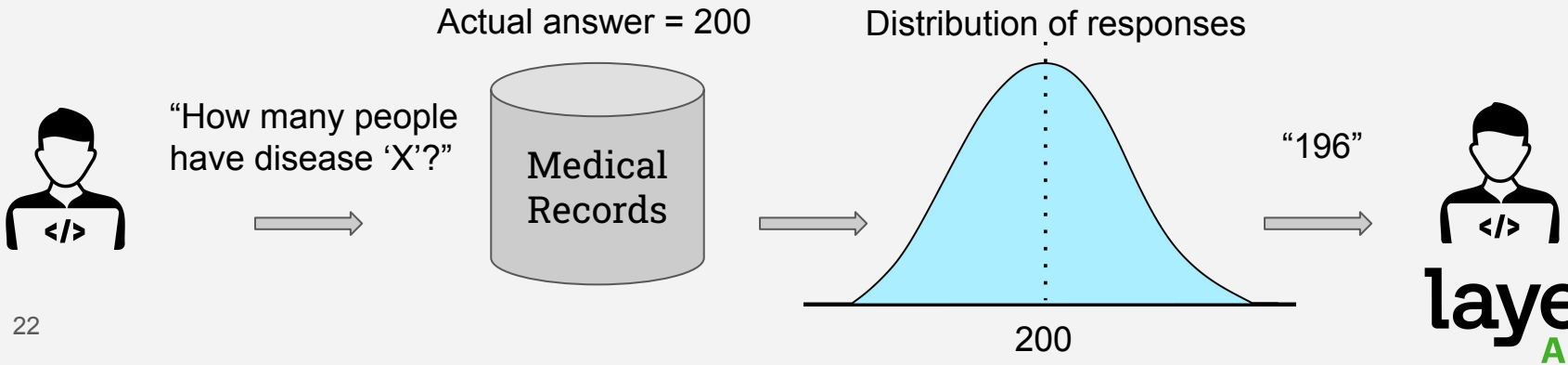


Differential Privacy

Adding randomness is a good way to achieve privacy.

Differential Privacy works with **randomized queries** that return answers given a dataset.

Instead of returning the true answer to a query, each possible answer (including the true one) has some probability of being returned.

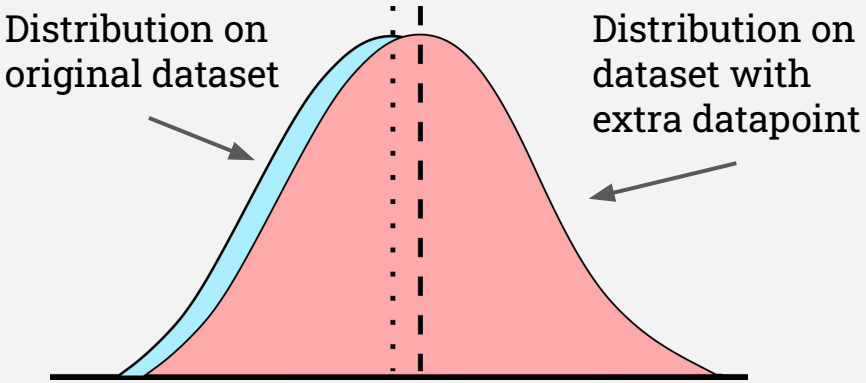


layer 6 AI at TD

Differential Privacy

Intuitively, the likelihood of any result from the query must be almost unchanged when one datapoint is added or removed.

Defn of Privacy (DP): Query is differentially private if the results are almost indistinguishable for datasets that differ by one record.



layer 6
AI at TD

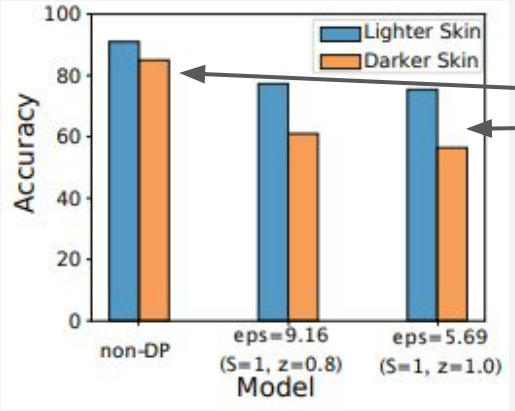
Differential Privacy makes unfairness worse

The objective of DP is to obscure low-level information about individual datapoints.

Randomization naturally **affects outliers and minority groups more**.

Any unfairness in a model tends to be exacerbated by DP.

Gender classification accuracy by skin tone



Drop in accuracy becomes larger with DP

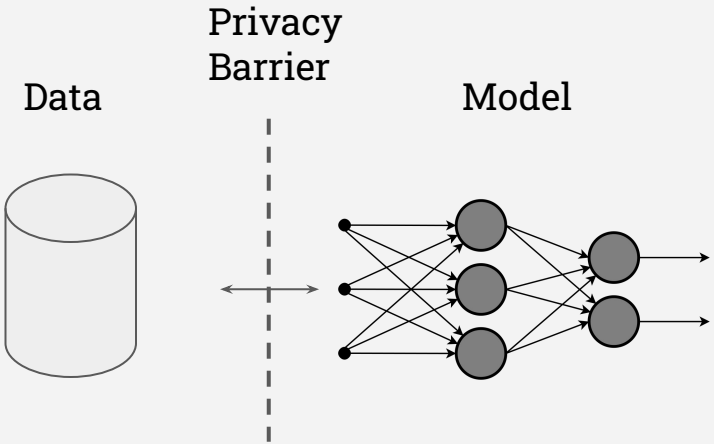
[Bagdasaryan et al. 2019]

layer 6 AI at TD

Training ML models with DP-SGD [Abadi et al. 2016]

Control privacy during training by randomizing gradient updates

- Gradients can change a lot from just one datapoint - **clip** to provide finite sensitivity
- Aggregate clipped per-sample gradients
- **Add noise** to the deterministic gradients and take a gradient descent step



Gradient clipping and noising are standard regularization techniques.

layer 6
AI at TD

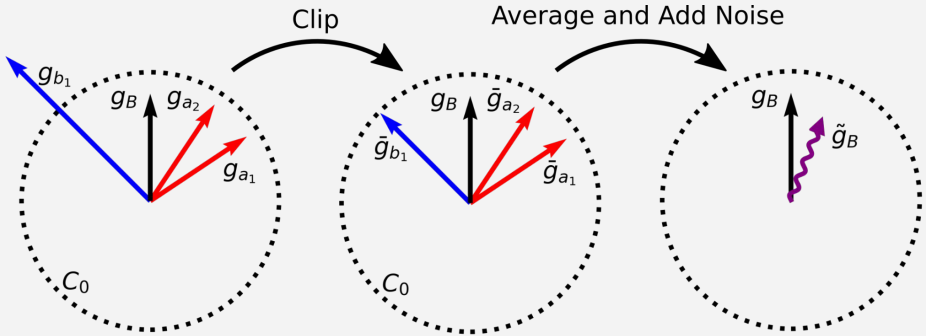
Fairness impacts of DP-SGD

DP-SGD first clips per-datapoint gradients, aggregates them, and adds noise.

The model needs more exposure to datapoints from **underrepresented or complex groups** in order to learn their properties.

Hence, these datapoints tend to have larger gradients throughout training, and in turn are **more likely to be clipped**.

Clipping **decreases their influence**, trapping us in a loop of unfairness.



layer 6
AI at TD

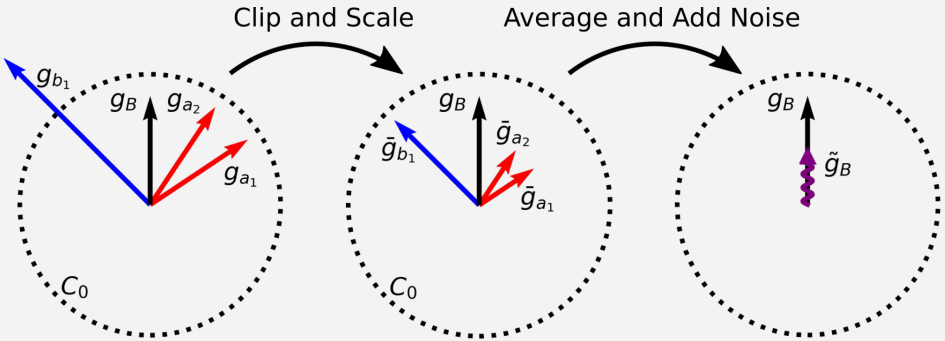
Fairness impacts of DP-SGD

Inequitable clipping in DP-SGD causes unfairness.

We can modify the clipping approach so that it acts uniformly on most gradients.

Scale down all gradients by the same factor so that the averaged gradient's direction remains unchanged.

Can completely remove disparate impact and avoids disparate treatment.



[Esipova et al. 2022]

layer 6
AI at TD

Conclusions

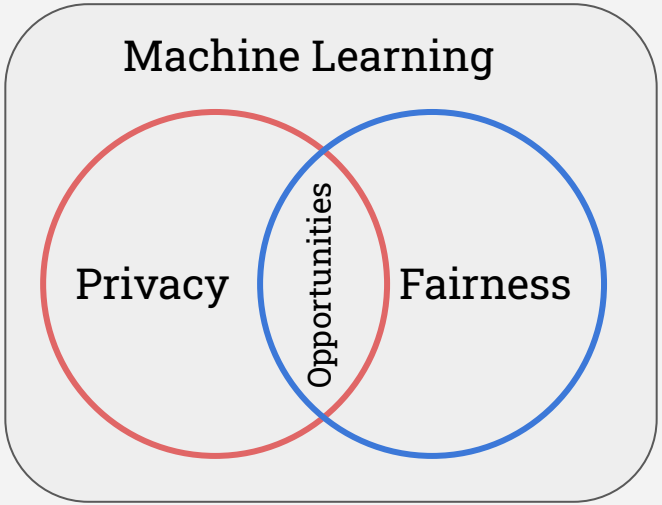
Conclusions

Applying Privacy Enhancing Technologies can have adverse impacts on fairness.

In FL, privacy prevents the central server from evaluating global fairness.

In DP, randomization is more likely to obscure data from minority groups.

Intersectionality is important for deploying ML in regulated industries.



layer 6
AI at TD

Thank you!

jesse@layer6.ai

jesse.cresswell@td.com