

# Fair and Private Deep Learning

Manisha Padala<sup>1</sup>  
Sankarshan Damle<sup>2</sup> and Sujit Gujar<sup>2</sup>

<sup>1</sup> Department of CSA, IISc, Bangalore

<sup>2</sup>Machine Learning Lab, IIIT, Hyderabad  
[github.com/magnetar-iiith/FairPrivateDL](https://github.com/magnetar-iiith/FairPrivateDL)



INTERNATIONAL INSTITUTE OF  
INFORMATION TECHNOLOGY  
HYDERABAD



# Agenda

## 1 Motivation

- Motivation: Fairness
- Motivation: Privacy
- Fair and Private Deep Learning

## 2 Fairness in Deep Learning

- Fairness Notions
- Impossibility
- Approaches: Pre, Post, and In-processing

## 3 Privacy in Deep Learning

- Differential Privacy
- DP-SGD
- Other Advances

## 4 FPFL: Fair and Private Federated Learning

- FPFL
- Implementation (Colab)

# Motivation



# Proliferation of AI

- Autonomous Driving, Robotics, Llama, Computer Vision



Figure: Ubiquitous AI systems <sup>a</sup>

---

<sup>a</sup>Image Credits: Manisha Padala Thesis

# Proliferation of AI

- Autonomous Driving, Robotics, Llama, Computer Vision
- In general AI-assisted decision making



Figure: Ubiquitous AI systems <sup>a</sup>

---

<sup>a</sup>Image Credits: Manisha Padala Thesis

# Proliferation of AI

- Autonomous Driving, Robotics, Llama, Computer Vision
- In general AI-assisted decision making
  - ▶ Crowdsourcing, Crowdfunding



Figure: Ubiquitous AI systems <sup>a</sup>

---

<sup>a</sup>Image Credits: Manisha Padala Thesis

# Proliferation of AI

- Autonomous Driving, Robotics, Llama, Computer Vision
- In general AI-assisted decision making
  - ▶ Crowdsourcing, Crowdfunding
  - ▶ Online Advertisements



Figure: Ubiquitous AI systems <sup>a</sup>

---

<sup>a</sup>Image Credits: Manisha Padala Thesis

# Proliferation of AI

- Autonomous Driving, Robotics, Llama, Computer Vision
- In general AI-assisted decision making
  - ▶ Crowdsourcing, Crowdfunding
  - ▶ Online Advertisements
  - ▶ Classification, Jurisdiction, Recruitments



Figure: Ubiquitous AI systems <sup>a</sup>

<sup>a</sup>Image Credits: Manisha Padala Thesis

# Proliferation of AI

- Autonomous Driving, Robotics, Llama, Computer Vision
- In general AI-assisted decision making
  - ▶ Crowdsourcing, Crowdfunding
  - ▶ Online Advertisements
  - ▶ Classification, Jurisdiction, Recruitments
  - ▶ Resource Allocation, Course Allocation, Ride Shares, Rent Shares
  - ▶ Resource Allocation, Course Allocation, Ride Shares, Rent Shares

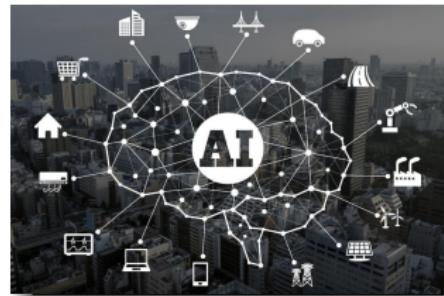


Figure: Ubiquitous AI systems <sup>a</sup>

---

<sup>a</sup>Image Credits: Manisha Padala Thesis

# Fairness

- You scored 99.83 percentile in JEE, and your cousin scored 99.56 percentile

# Fairness

- You scored 99.83 percentile in JEE, and your cousin scored 99.56 percentile
  - ▶ She got admission to IIIT, Hyderabad (CS), and you did not

# Fairness

- You scored 99.83 percentile in JEE, and your cousin scored 99.56 percentile
  - ▶ She got admission to IIIT, Hyderabad (CS), and you did not
- Take yourself 100 years back

# Fairness

- You scored 99.83 percentile in JEE, and your cousin scored 99.56 percentile
  - ▶ She got admission to IIIT, Hyderabad (CS), and you did not
- Take yourself 100 years back
  - ▶ She might not even have been allowed to go to college

# Fairness

- You scored 99.83 percentile in JEE, and your cousin scored 99.56 percentile
  - ▶ She got admission to IIIT, Hyderabad (CS), and you did not
- Take yourself 100 years back
  - ▶ She might not even have been allowed to go to college

# Fairness

- You scored 99.83 percentile in JEE, and your cousin scored 99.56 percentile
  - ▶ She got admission to IIIT, Hyderabad (CS), and you did not
- Take yourself 100 years back
  - ▶ She might not even have been allowed to go to college

Does historical unfairness matter?

# Historical Data Is Important

- AI/ML – Learns from the data

# Historical Data Is Important

- AI/ML – Learns from the data
- University admissions:

# Historical Data Is Important

- AI/ML – Learns from the data
- University admissions:
  - ▶ Receive large # of applications (e.g., Washington University receives 500+ PhD applications in ML alone)

# Historical Data Is Important

- AI/ML – Learns from the data
- University admissions:
  - ▶ Receive large # of applications (e.g., Washington University receives 500+ PhD applications in ML alone)
  - ▶ Suppose we are building a regression model to predict scores for a candidate about how likely are they going to be successful

# Historical Data Is Important

- AI/ML – Learns from the data
- University admissions:
  - ▶ Receive large # of applications (e.g., Washington University receives 500+ PhD applications in ML alone)
  - ▶ Suppose we are building a regression model to predict scores for a candidate about how likely are they going to be successful
  - ▶ Less number of female candidates in historical data (STEM even now has < 35% females in the U.S.)

# Historical Data Is Important

- AI/ML – Learns from the data
- University admissions:
  - ▶ Receive large # of applications (e.g., Washington University receives 500+ PhD applications in ML alone)
  - ▶ Suppose we are building a regression model to predict scores for a candidate about how likely are they going to be successful
  - ▶ Less number of female candidates in historical data (STEM even now has < 35% females in the U.S.)

# Historical Data Is Important

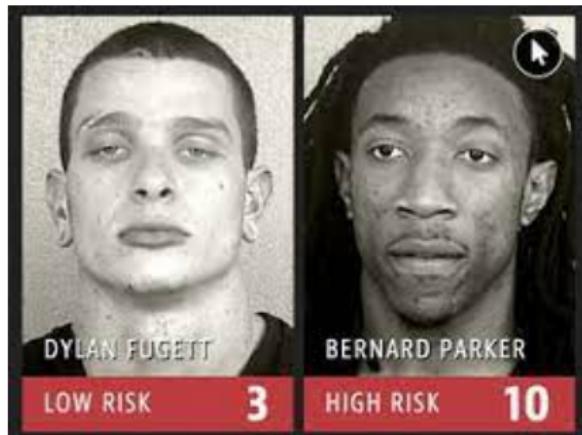
- AI/ML – Learns from the data
- University admissions:
  - ▶ Receive large # of applications (e.g., Washington University receives 500+ PhD applications in ML alone)
  - ▶ Suppose we are building a regression model to predict scores for a candidate about how likely are they going to be successful
  - ▶ Less number of female candidates in historical data (STEM even now has < 35% females in the U.S.)

Are these concerns real or hyped?

# COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions
  - ▶ Used by U.S. courts to assess the likelihood of a defendant becoming a recidivist (repeat offender)
- Prior to COMPAS (even before statistics and technology)
  - ▶ Human prejudices (e.g., against blacks) – may mark somebody for high risk
- COMPAS: general and violent recidivism, and for pre-trial risk assessment
- Machine Bias (ProPublica)

# COMPAS



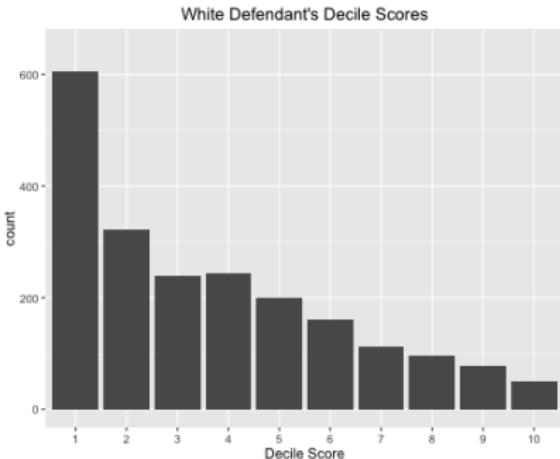
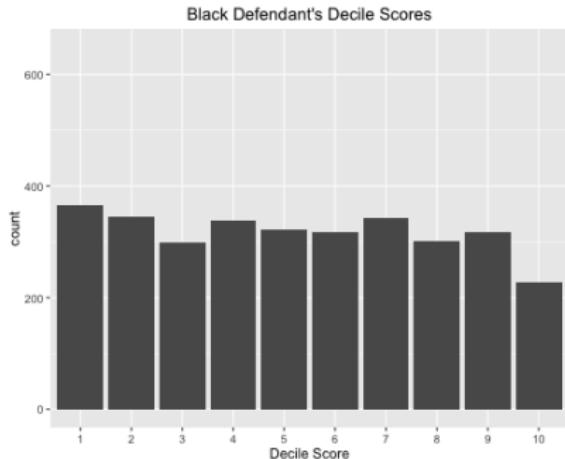
Source: Machine Bias (ProPublica)

## Two Petty Theft Arrests



Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

# COMPAS



Source: Machine Bias (ProPublica)

# COMPAS

- “Scores like this – known as risk assessments – are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts, to even more fundamental decisions about defendants’ freedom.” – [Machine Bias \(ProPublica\)](#)

# COMPAS

- “Scores like this – known as risk assessments – are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts, to even more fundamental decisions about defendants’ freedom.” – [Machine Bias \(ProPublica\)](#)
- Unfortunately for COMPAS (and those who suffered the consequences of their software), it turns out the algorithm gave **disproportionately higher** risk scores to **black** defendants than it did to white defendants.

# COMPAS

|   | WHITE | AFRICAN AMERICAN |
|---|-------|------------------|
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9%            |
| Labeled Lower Risk, Yet Did Re-Offend     | 47.7% | 28.0%            |

*Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)*

Source: Machine Bias (ProPublica)

## Other Instances

- Amazon same-day delivery – appeared to be racist

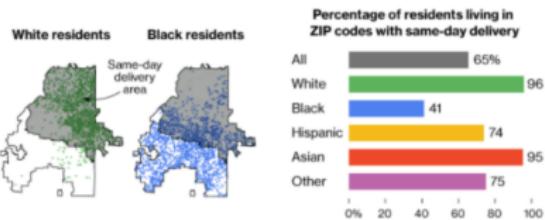


Fig 3: Source - Bloomberg

# Other Instances

- Amazon same-day delivery – appeared to be racist
  - The Northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. However, southern Atlanta, where 90% of the residents are black, is excluded from same-day delivery

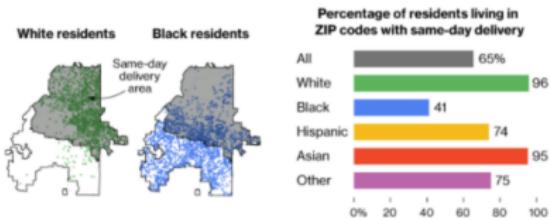


Fig 3: Source - Bloomberg

# Other Instances

- Amazon same-day delivery – appeared to be racist
  - ▶ The Northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. However, southern Atlanta, where 90% of the residents are black, is excluded from same-day delivery
- Interesting: AMAZON never used race as a feature in their data while training their DL model

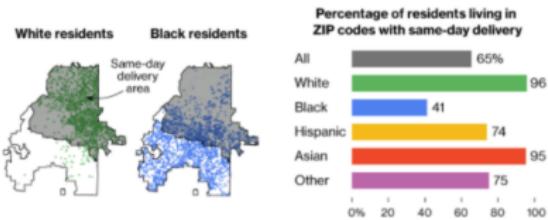


Fig 3: Source - Bloomberg

# Other Instances

| Search query     | Work experience | Education experience | Profile views | Candidate | Xing ranking |
|------------------|-----------------|----------------------|---------------|-----------|--------------|
| Brand Strategist | 146             | 57                   | 12992         | male      | 1            |
| Brand Strategist | 327             | 0                    | 4715          | female    | 2            |
| Brand Strategist | 502             | 74                   | 6978          | male      | 3            |
| Brand Strategist | 444             | 56                   | 1504          | female    | 4            |
| Brand Strategist | 139             | 25                   | 63            | male      | 5            |
| Brand Strategist | 110             | 65                   | 3479          | female    | 6            |
| Brand Strategist | 12              | 73                   | 846           | male      | 7            |
| Brand Strategist | 99              | 41                   | 3019          | male      | 8            |
| Brand Strategist | 42              | 51                   | 1359          | female    | 9            |
| Brand Strategist | 220             | 102                  | 17186         | female    | 10           |

TABLE II: Top k results on [www.xing.com](http://www.xing.com) (Jan 2017) for the job search query “Brand Strategist”.

Image Credit: Lahoti *et al.*

# Other Instances

- Algorithms that Demonstrate Artificial Intelligence Bias<sup>1</sup>
  - ▶ PredPol – Predicting crimes in US; Racial minorities
- Amazon recruitment tool
  - ▶ Biased against females
- ChatGPT Rift and Bias<sup>23</sup>

---

<sup>1</sup><https://www.geeksforgeeks.org/5-algorithms-that-demonstrate-artificial-intelligence-bias/>

<sup>2</sup>ChatGPT Is Like Many Other AI Models: Rife With Bias (insider.com)

<sup>3</sup>Potential for bias based on prompt - ChatGPT and Generative Artificial Intelligence (AI)

# Privacy Concerns in AI

FORBES > TECH

## How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

**Kashmir Hill** Former Staff

*Welcome to The Not-So Private Parts where technology & privacy collide*



Feb 16, 2012, 11:02am EST

Source: [Forbes](#)

# Privacy Concerns in AI

- What if Target commits not to sell your data?

# Privacy Concerns in AI

- What if Target commits not to sell your data?
- Netflix challenge to build a recommender system  
([en.wikipedia.org/wiki/Netflix\\_Prize](https://en.wikipedia.org/wiki/Netflix_Prize))

# Privacy Concerns in AI

- What if Target commits not to sell your data?
- Netflix challenge to build a recommender system  
([en.wikipedia.org/wiki/Netflix\\_Prize](https://en.wikipedia.org/wiki/Netflix_Prize))
- Data was completely anonymized

# Privacy Concerns in AI

- What if Target commits not to sell your data?
- Netflix challenge to build a recommender system  
([en.wikipedia.org/wiki/Netflix\\_Prize](https://en.wikipedia.org/wiki/Netflix_Prize))
- Data was completely anonymized
- Researchers still linked Netflix users to their IMDb profiles

# Privacy Concerns in AI

- What if Target commits not to sell your data?
- Netflix challenge to build a recommender system  
([en.wikipedia.org/wiki/Netflix\\_Prize](https://en.wikipedia.org/wiki/Netflix_Prize))
- Data was completely anonymized
- Researchers still linked Netflix users to their IMDb profiles
- Many complex attacks are possible



INTERNATIONAL INSTITUTE OF  
INFORMATION TECHNOLOGY  
HYDERABAD

# Privacy Concerns in AI

- DL-based classifiers have been shown to be susceptible to privacy attacks



Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Image Credit: Fredrikson et al. [18]

# Privacy Concerns in AI

- DL-based classifiers have been shown to be susceptible to privacy attacks
- **Membership Attacks:** To discover whether a particular instance was present in the training set.



Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Image Credit: Fredrikson et al. [18]

# Privacy Concerns in AI

- DL-based classifiers have been shown to be susceptible to privacy attacks
- **Membership Attacks:** To discover whether a particular instance was present in the training set.
- **Model Inversion Attacks:** To reverse engineer the input data that led to a particular model's output.



Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Image Credit: Fredrikson et al. [18]

# Privacy Concerns in AI

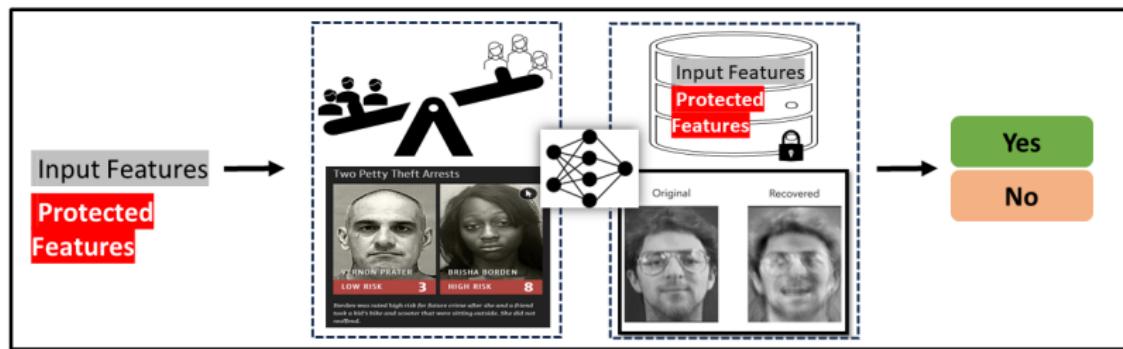
- DL-based classifiers have been shown to be susceptible to privacy attacks
- **Membership Attacks:** To discover whether a particular instance was present in the training set.
- **Model Inversion Attacks:** To reverse engineer the input data that led to a particular model's output.
- **Reconstruction Attacks:** To reconstruct sensitive information from the model's output (without necessarily focusing on the specific input data that led to that output)



Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Image Credit: Fredrikson et al. [18]

- This tutorial: Focus training Deep Neural Networks for **fairness** and **privacy**



# Fairness in Deep Learning

# Notations and Model

Input:  $X = \{x_1, \dots, x_n\}$

Labels:  $Y = \{y_1, \dots, y_n\}, y_i \in \{0, 1\}$

Output:  $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_n\}, \hat{y}_i \in \{0, 1\}$

Protected Attribute:  $A = \{A_1, \dots, A_n\}, A_i \in \{a^1, a^2\}$

# Notations and Model

Input:  $X = \{x_1, \dots, x_n\}$

Labels:  $Y = \{y_1, \dots, y_n\}, y_i \in \{0, 1\}$

Output:  $\hat{Y} = \{\hat{y}_1, \dots, \hat{y}_n\}, \hat{y}_i \in \{0, 1\}$

Protected Attribute:  $A = \{A_1, \dots, A_n\}, A_i \in \{a^1, a^2\}$

Model  $h(x)$  gives the prediction probabilities (scores) and we obtain  $\hat{y}$  after thresholding

# Calibration

## Definition

A classifier  $h$  is said to be calibrated if

$$\Pr(y = 1 | h(x) = r) = r$$

# Calibration

## Definition

A classifier  $h$  is said to be calibrated if

$$\Pr(y = 1 | h(x) = r) = r$$

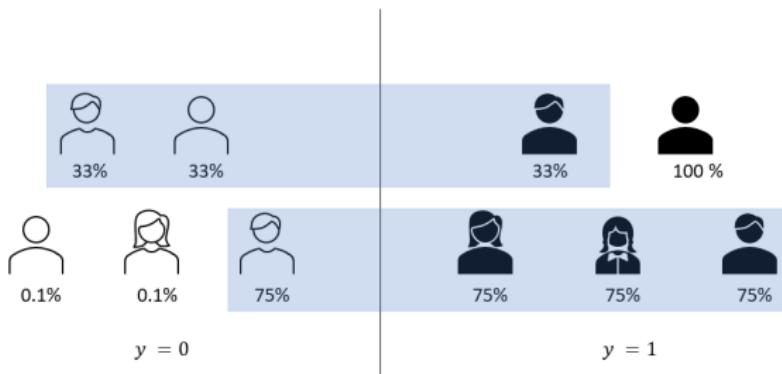


# Calibration

## Definition

A classifier  $h$  is said to be calibrated if

$$\Pr(y = 1 | h(x) = r) = r$$



# Group Fairness Criteria

| Independence      | Separation          | Sufficiency         |
|-------------------|---------------------|---------------------|
| $\hat{Y} \perp A$ | $\hat{Y} \perp A Y$ | $Y \perp A \hat{Y}$ |

- Independence - predictions are independent of a sensitive attribute
- Separation - independence within each target class
- Sufficiency - calibration per group

# Independence

$$\hat{Y} \perp A$$

- Demographic Parity,  
Statistical Parity

$$\Pr(\hat{Y} = 1|A = a^1) = \\ \Pr(\hat{Y} = 1|A = a^2)$$

- Disparate Impact, p%-rule

$$\frac{\Pr(\hat{Y} = 1|A = a^1)}{\Pr(\hat{Y} = 1|A = a^2)} \geq p$$

# Independence

$$\hat{Y} \perp A$$

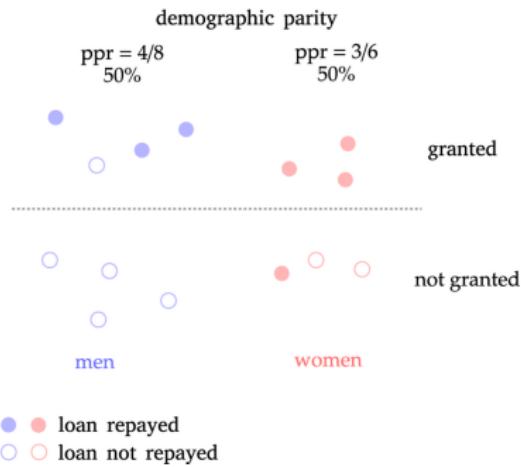
- Demographic Parity, Statistical Parity

$$\Pr(\hat{Y} = 1|A = a^1) =$$

$$\Pr(\hat{Y} = 1|A = a^2)$$

- Disparate Impact, p%-rule

$$\frac{\Pr(\hat{Y} = 1|A = a^1)}{\Pr(\hat{Y} = 1|A = a^2)} \geq p$$



# Independence

## Note

- All groups have equal claim to acceptance
- Convenient technical properties
- Hiring qualified individuals from  $a^1$  but not so qualified individuals from  $a^2$  just to ensure fairness

# Separation

$\hat{Y} \perp A \mid Y$

- False Negative Parity (FNP)

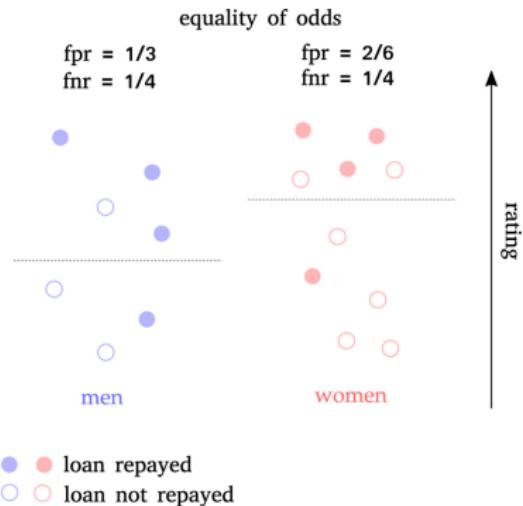
$$\Pr(\hat{Y} = 0 \mid A = a^1, Y = 1) = \Pr(\hat{Y} = 0 \mid A = a^2, Y = 1)$$

- False Positive Parity (FPP)

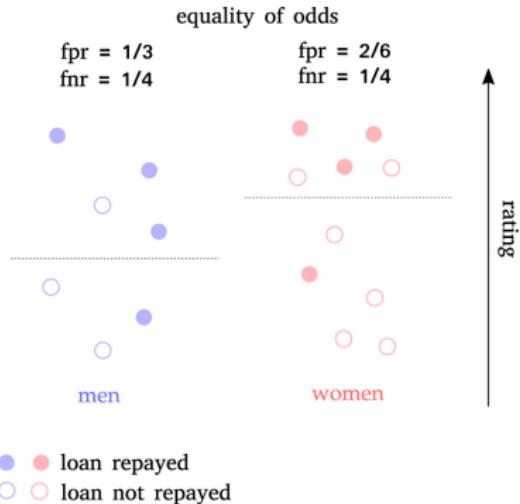
$$\Pr(\hat{Y} = 1 \mid A = a^1, Y = 0) = \Pr(\hat{Y} = 1 \mid A = a^2, Y = 0)$$

- Accuracy Parity - FP + FN parity
- Equalized Odds - FP and FN parity

# Separation



# Separation



- Opportunity is denied to eligible candidates at equal rate.
- Optimal classifier need not equalize error rates.
  - ▶ Who bears the burden of mis-classification?

# Sufficiency

$$Y \perp A | \hat{Y}$$

- Sufficiency

$$\Pr(Y = 1 | h(x) = r, A = a^1) = \Pr(Y = 1 | h(x) = r, A = a^2)$$

- Calibration by groups

$$\Pr(Y = 1 | h(x) = r, A = a^1) = r$$

# Sufficiency

- In theory unconstrained supervised learning imply group calibration (Liu, Simchowitz, and Hardt 2019 [34])

# Sufficiency

- In theory unconstrained supervised learning imply group calibration (Liu, Simchowitz, and Hardt 2019 [34])

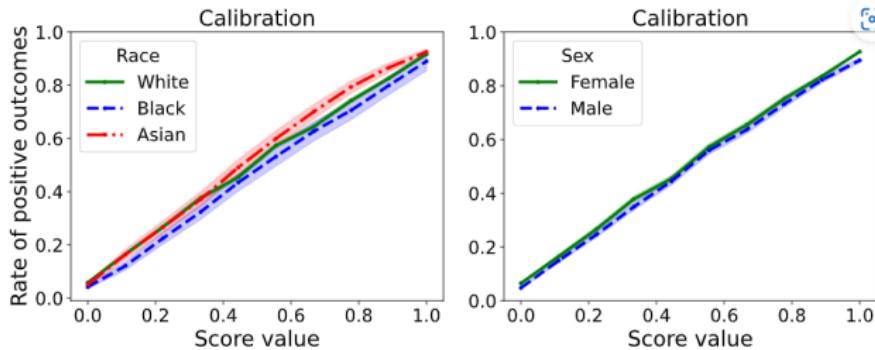


Figure: Group Calibration on Census ACS Data [3]

# Impossibilities

Can we achieve any two fairness conditions together?

# Impossibilities

Can we achieve any two fairness conditions together?

- Independence  $\implies A \perp \hat{Y}$
- Sufficiency  $\implies Y \perp A | \hat{Y}$

Hence,  $A \perp Y$

Unless  $A \perp Y$ , we cannot achieve both!

# Impossibilities

| Fairness Notions             | Condition for Impossibility              |
|------------------------------|--|
| Independence and Separation  | $A \not\perp Y$ or $\hat{Y} \not\perp Y$ |
| Independence and Sufficiency | $A \not\perp Y$                          |
| Sufficiency and Separation   | $A \not\perp Y$ and full support         |

# Impossibilities

| Fairness Notions             | Condition for Impossibility              |
|------------------------------|--|
| Independence and Separation  | $A \not\perp Y$ or $\hat{Y} \not\perp Y$ |
| Independence and Sufficiency | $A \not\perp Y$                          |
| Sufficiency and Separation   | $A \not\perp Y$ and full support         |

## On Fairness and Calibration [44]

Unless base rates are the same (i.e.,  
 $\Pr(Y = 1|A = a^1) = \Pr(Y = 1|A = a^2)$ ) or  $h$  is a perfect classifier, it is impossible to have both equalized odds and calibration

# Impossibility

For calibration by group, we require

$$\Pr(Y = 1 | \hat{Y} = 1, A = a^1) = \Pr(Y = 1 | \hat{Y} = 1, A = a^2)$$

Let  $PP_1 = \Pr(Y = 1 | \hat{Y} = 1, A = a^1)$ , by Bayes rule,

$$PP_1 =$$

$$\frac{\Pr(\hat{Y} = 1 | Y = 1, A = a^1) \Pr(Y = 1 | A = a^1)}{\Pr(\hat{Y} = 1 | Y = 1, A = a^1) \Pr(Y = 1 | A = a^1) + \Pr(\hat{Y} = 1 | Y = 0, A = a^1) \Pr(Y = 0 | A = a^1)}$$

# Impossibility

For calibration by group, we require

$$\Pr(Y = 1 | \hat{Y} = 1, A = a^1) = \Pr(Y = 1 | \hat{Y} = 1, A = a^2)$$

Let  $PP_1 = \Pr(Y = 1 | \hat{Y} = 1, A = a^1)$ , by Bayes rule,

$$PP_1 =$$

$$\frac{\Pr(\hat{Y} = 1 | Y = 1, A = a^1) \Pr(Y = 1 | A = a^1)}{\Pr(\hat{Y} = 1 | Y = 1, A = a^1) \Pr(Y = 1 | A = a^1) + \Pr(\hat{Y} = 1 | Y = 0, A = a^1) \Pr(Y = 0 | A = a^1)}$$

Given the base rate  $p_1 = \Pr(Y = 1 | A = a^1)$

$$PP_1 = \frac{TPR_1 \times p_1}{TPR_1 \times p_1 + FPR_1 \times (1 - p_1)}$$

# Impossibility

We assume that Equalized Odds is satisfied, that is,

- Let  $u = TPR_1 = TPR_2$
- $v = FPR_2 = FPR_1$

Then calibration is satisfied when  $PP_1 = PP_2$  or,

$$PP_1 = PP_2 \implies \frac{u \times p_1}{u \times p_1 + v \times (1 - p_1)} = \frac{u \times p_2}{u \times p_2 + v \times (1 - p_2)}$$

The above is true when

- Same base rates:  $p_1 = p_2$
- Perfect predictor:  $u = 1, v = 0$

# Existing Approaches

In order to fit into the ML framework, there are three approaches

---

|       |                 |                    |
|-------|-----------------|--------------------|
| (PRP) | Pre-processing  | Input Features     |
| (INP) | In-processing   | Model              |
| (POP) | Post-processing | Output Predictions |

---

# Existing Approaches

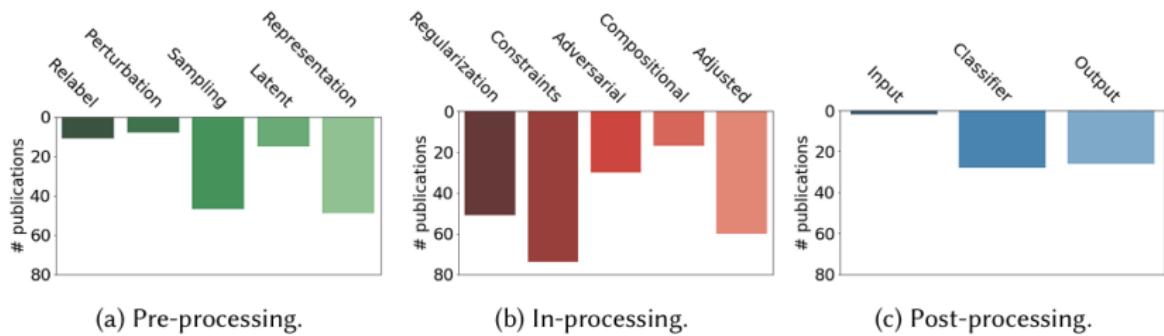


Figure: Summary of the approaches [23]

# Pre-processing: Relabeling and Perturbation

- Massaging - change ground truth labels
  - ▶ Kamiran and Calders 2009 [25], 2012 [27]
  - ▶ Žliobaite, Kamiran, and Calders 2011 [53]
  - ▶ Luong, Ruggieri, and Turini 2011 [36]

# Pre-processing: Relabeling and Perturbation

- Massaging - change ground truth labels
  - ▶ Kamiran and Calders 2009 [25], 2012 [27]
  - ▶ Žliobaite, Kamiran, and Calders 2011 [53]
  - ▶ Luong, Ruggieri, and Turini 2011 [36]
- Perturbation - modifying input features
  - ▶ Feldman, Friedler, Moeller, Scheidegger, and Venkatasubramanian 2015 [16]
  - ▶ Lum and Johndrow 2016 [35]

# Pre-processing: Massaging Data [25]

Fairness Notion: Demographic Parity (positive outcome supplied to both the groups equally)

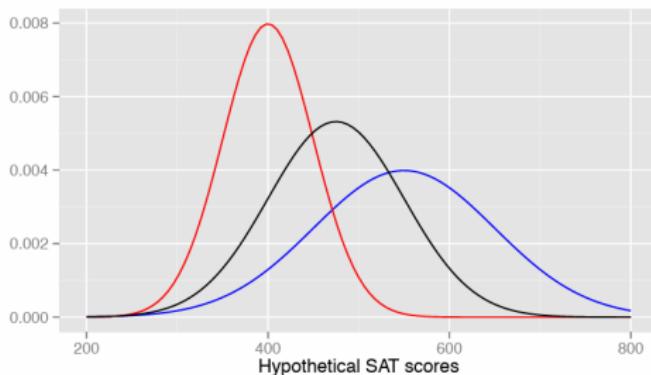
| $X$      | $A$ | $Y$ |
|----------|-----|-----|
| $x_1$    | M   | +   |
| $x_2$    | M   | -   |
| $x_3$    | M   | -   |
| $x_4$    | M   | +   |
| $x_5$    | M   | -   |
| $x_6$    | F   | +   |
| $x_7$    | F   | +   |
| $x_8$    | F   | +   |
| $x_9$    | F   | +   |
| $x_{10}$ | F   | -   |

| $X$      | $A$ | $Y$ | $\text{Pr}_+$ |
|----------|-----|-----|---------------|
| $x_1$    | M   | +   | 62%           |
| $x_2$    | M   | -   | 6%            |
| $x_3$    | M   | -   | 49%           |
| $x_4$    | M   | +   | 67%           |
| $x_5$    | M   | -   | 3%            |
| $x_6$    | F   | +   | 76%           |
| $x_7$    | F   | +   | 89%           |
| $x_8$    | F   | +   | 90%           |
| $x_9$    | F   | +   | 60%           |
| $x_{10}$ | F   | -   | 10%           |

| $X$      | $A$ | $Y$ |
|----------|-----|-----|
| $x_1$    | M   | +   |
| $x_2$    | M   | -   |
| $x_3$    | M   | +   |
| $x_4$    | M   | +   |
| $x_5$    | M   | -   |
| $x_6$    | F   | +   |
| $x_7$    | F   | +   |
| $x_8$    | F   | +   |
| $x_9$    | F   | -   |
| $x_{10}$ | F   | -   |



# Pre-processing: Perturbation



Modifying continuous input (ordered) while ensuring the rank is maintained within the group. Red curve (male scores  $\mu = 400, \sigma = 50$ ), Blue curve (female scores  $\mu = 550, \sigma = 100$ ), Black curve (modified values  $\mu = 475, \sigma = 75$ ) [16]

- Male students in 95<sup>th</sup> percentile ( $\geq 500$ ) are given scores  $\geq 625$  and remain in 95<sup>th</sup> percentile.

# Pre-processing: Sampling

Calders, Kamiran, and Pechenizkiy 2009 [6]

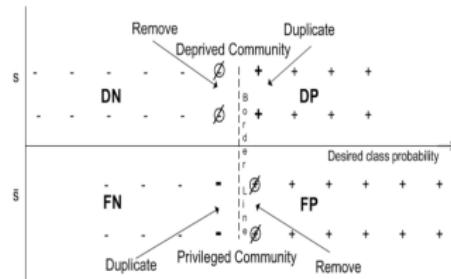
- Reweighting - instance is weighed according to its label and protected attribute
- Uniform Sampling - duplicate or removal of samples
- Preferential Sampling - duplicate or removal of samples based on the decision boundary

# Pre-processing: Sampling

Calders, Kamiran, and Pechenizkiy 2009 [6]

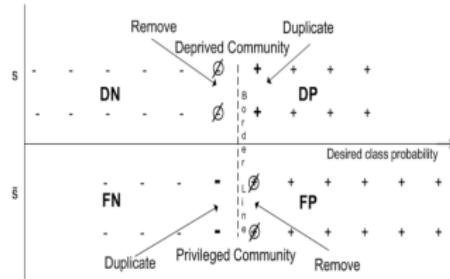
- Reweighting - instance is weighed according to its label and protected attribute
  - Uniform Sampling - duplicate or removal of samples
  - Preferential Sampling - duplicate or removal of samples based on the decision boundary
- 
- SMOTE: Synthetic Minority Oversampling TEchnique (Chakraborty, Majumder, and Menzies 2021 [7])

# Pre-processing: Sampling



Preferential Duplication and Removal [26]

# Pre-processing: Sampling



Preferential Duplication and Removal [26]

SMOTE:  $X_{new} = X + \text{rand}(0, 1) \cdot (X - X')$

# Pre-processing: Representation Learning

Learn representations such that

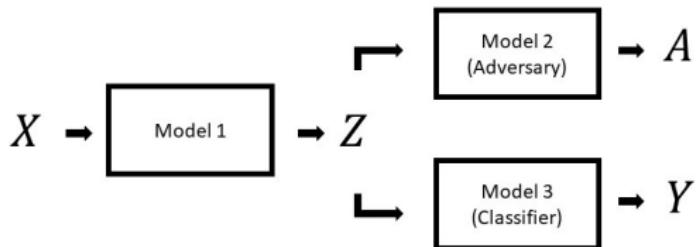
- sensitive information is removed
- target information is preserved

# Pre-processing: Representation Learning

Learn representations such that

- sensitive information is removed
- target information is preserved
- Variational Autoencoders (Creager, Madras, Jacobsen, Weis, Swersky, Pitassi, and Zemel 2019 [9])
- Neural Style Transfer (Quadrianto, Sharmanaska, and Thomas 2019 [45])
- Adversarial Learning (Zhu, Zheng, Liao, Li, and Luo 2021 [52])
- Contrastive Learning (Gupta, Ferber, Dilkina, and Ver Steeg 2021 [20])

# Pre-processing: Adversarial Learning



# Post-processing Approaches

Post-processing – fair predictions from an existing trained model

# Post-processing: Input Correction

- Perturbing test input data
  - ▶ Adler, Falk, Friedler, Nix, Rybeck, Scheidegger, Smith, and Venkatasubramanian 2018 [2]
  - ▶ Li, Meng, Chen, Yu, Wu, Zhou, and Xu 2022 [33]

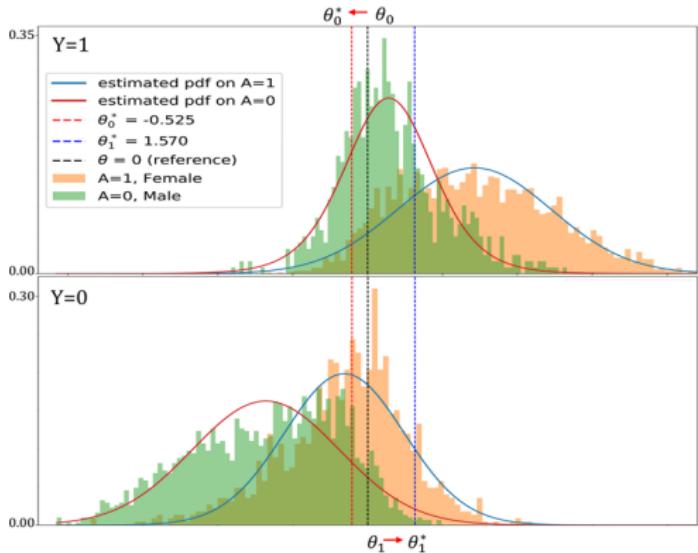
# Post-processing: Model Correction

The classifiers are modified

- From unfair classifier to fair classifier by solving an optimization for fairness loss [21] and extended by [49]
- Train two classifiers ( $h_1, h_0$ ), return class mean instead of  $h_1(x)$  w.p.  $\alpha$  to ensure EO [44]
- Modifying decision trees by relabelling leaf nodes
  - ▶ iteratively find the leaf node with highest  $\frac{\text{fairness-gain}}{\text{accuracy-drop}}$  [28]
- Boosting of classifiers to enhance fairness [31]

# Post-processing: Output Correction

- Learning thresholds for predicted probabilities  
Menon and Williamson 2018 [38]
- Learning thresholds specific to a group  
Jang, Shi, and Wang 2022 [24]



# In-processing: Approaches

In-processing - train a model to be fair using modified loss function or training approaches

## In-processing: Adversarial

PRP learns  $Z$  an intermediate representation, here we try to learn  $\hat{Y}$  that is robust to bias

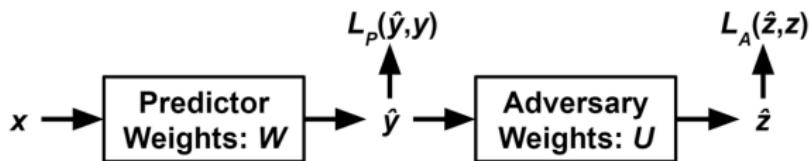


Figure: Zhang, Lemoine, and Mitchell 2018 [51]

- Minimize  $L_P$
- Maximize  $L_A$

# In-processing: Compositional Learning

- Learning different classifiers for different groups using transfer learning (Dwork, Immorlica, Kalai, and Leiserson 2018 [14], Ustun, Liu, and Parkes 2019 [48])
- Ensemble of multiple classifiers for different metrics of fairness (Li, Meng, Chen, Yu, Wu, Zhou, and Xu 2022 [33])

# In-processing: Adjusted Learning

- Active Learning - (Obermeyer, Powers, Vogeli, and Mullainathan 2019 [40])
- Rejection Learning - (Madras, Creager, Pitassi, and Zemel 2018 [37])
- Distributional Robust Optimization - (Hashimoto, Srivastava, Namkoong, and Liang 2018 [22])

# In-processing: Constrained Based

- Loss function consists of performance loss and fairness loss
- Fairness loss is non-convex and complex

# In-processing: Constrained Based

- Loss function consists of performance loss and fairness loss
- Fairness loss is non-convex and complex
- Convexification (Bilal Zafar, Valera, Gomez Rodriguez, and Gummadi 2015 [5], Kamishima, Akaho, and Sakuma 2011, [29])

# In-processing: Convexification [5]

Fairness Notion: Disparate Impact

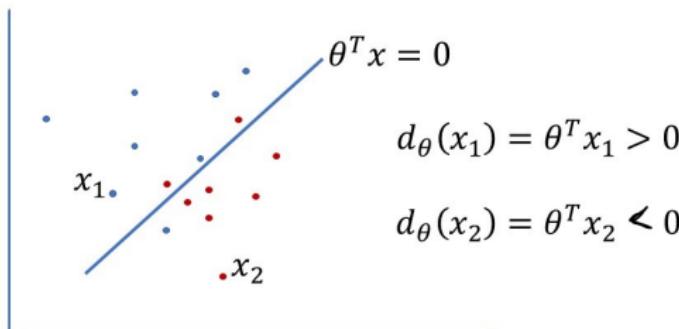
$$\min \left\{ \frac{\Pr(\hat{y}_i = 1 | a_i = 0)}{\Pr(\hat{y}_i = 1 | a_i = 1)}, \frac{\Pr(\hat{y}_i = 1 | a_i = 1)}{\Pr(\hat{y}_i = 1 | a_i = 0)} \right\} \geq 0.8$$

Non-convex and complex, so define a proxy loss that is convex.

# In-processing: Convexification [5]

Fairness Notion: Disparate Impact

$$\min \left\{ \frac{\Pr(\hat{y}_i = 1 | a_i = 0)}{\Pr(\hat{y}_i = 1 | a_i = 1)}, \frac{\Pr(\hat{y}_i = 1 | a_i = 1)}{\Pr(\hat{y}_i = 1 | a_i = 0)} \right\} \geq 0.8$$



$$\Pr(d_\theta(x_i) > 0 | a_i = 0) = \Pr(d_\theta(x_i) > 0 | a_i = 1)$$

# In-processing: Convexification [5]

$\Pr(d_\theta(x) > 0 | a_i = 0) = \Pr(d_\theta(x_i) > 0 | a_i = 1)$   
Perfect DI implies  $Covariance(d_\theta(x), a) = 0$

$$Cov(d_\theta(x), a) = \frac{1}{n} \sum_{i=1}^n (a - \mu_a) \theta^T x$$

- Convex w.r.t.  $\theta$  hence included in many classifier formulations

# In-processing: FNNC

FNNC: Fair Neural Network Classifier [41]

- We introduce a penalty term in the loss and train an NN-classifier
- Provide sample bounds why the penalty term helps minimize fairness

Padala Manisha, Sujit Gujar. "FNNC: Achieving Fairness through Neural Networks". (IJCAI '20)

# Loss Function and Optimizer

Cross-entropy loss with DP constraint (batch of S samples)

$$l_{CE}(h(x_i), y_i) = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i)$$

$$\hat{l}_{CE}(h(X), Y) = \frac{1}{S} \sum_{i=1}^S l_{CE}(h(x_i), y_i)$$

# Loss Function and Optimizer

Cross-entropy loss with DP constraint (batch of S samples)

$$l_{CE}(h(x_i), y_i) = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i)$$

$$\hat{l}_{CE}(h(X), Y) = \frac{1}{S} \sum_{i=1}^S l_{CE}(h(x_i), y_i)$$

$$const^{DP}(z_S) = \left| \frac{\sum_{i=1}^S p_i a_i}{\sum_{i=1}^S a_i} - \frac{\sum_{i=1}^S p_i (1 - a_i)}{\sum_{i=1}^S 1 - a_i} \right|$$

$$l_{DP}(h(X), \mathcal{A}, Y) : \frac{1}{B} \sum_{l=1}^B const^k(z_S^{(l)}) - \epsilon \leq 0$$

# Loss Function and Optimizer

Cross-entropy loss with DP constraint (batch of S samples)

$$l_{CE}(h(x_i), y_i) = -y_i \log(p_i) - (1 - y_i) \log(1 - p_i)$$

$$\hat{l}_{CE}(h(X), Y) = \frac{1}{S} \sum_{i=1}^S l_{CE}(h(x_i), y_i)$$

$$const^{DP}(z_S) = \left| \frac{\sum_{i=1}^S p_i a_i}{\sum_{i=1}^S a_i} - \frac{\sum_{i=1}^S p_i (1 - a_i)}{\sum_{i=1}^S 1 - a_i} \right|$$

$$l_{DP}(h(X), \mathcal{A}, Y) : \frac{1}{B} \sum_{l=1}^B const^k(z_S^{(l)}) - \epsilon \leq 0$$

# Lagrangian Multiplier

The overall loss for the network.

$$L_{NN} = \hat{L}_{CE} + \lambda I_{DP} \quad \lambda : \text{Lagrangian Multiplier}$$

# Lagrangian Multiplier

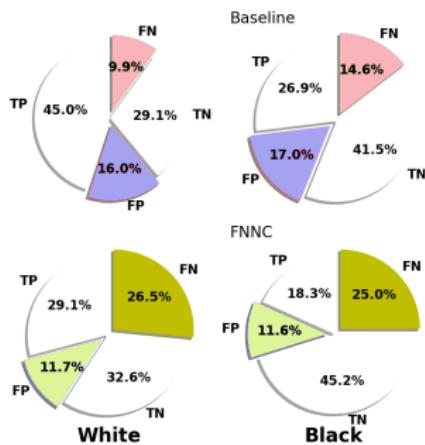
The overall loss for the network.

$$L_{NN} = \hat{L}_{CE} + \lambda I_{DP} \quad \lambda : \text{Lagrangian Multiplier}$$

$$\min_{\theta} \max_{\lambda} L_{NN}$$

Perform SGD twice, once for minimizing the loss w.r.t.  $\theta$  and again for maximizing the loss w.r.t.  $\lambda$  at every iteration

# Results - Equalized Odds



**Figure:** Compass dataset: FPR and FNR is comparable across race in FNNC as observed in the bottom left and right pie charts

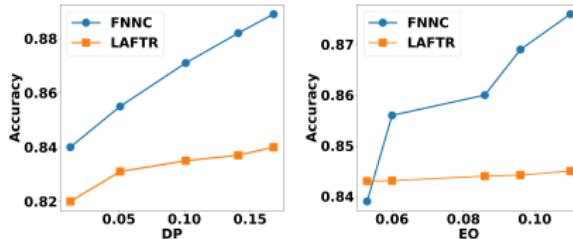


Figure: FNNC vs LAFTR [37] on Adult dataset

| Dataset | $\epsilon$ | FNNC         | COCO[39]     | LinCon [39]  |
|---------|------------|--------------|--------------|--------------|
| adult   | 0.05       | 0.28 (0.027) | 0.33 (0.035) | 0.39 (0.027) |
| compass | 0.20       | 0.32 (0.147) | 0.41 (0.206) | 0.57 (0.107) |
| crimes  | 0.20       | 0.28 (0.183) | 0.32 (0.197) | 0.52 (0.190) |
| default | 0.05       | 0.29 (0.011) | 0.37 (0.032) | 0.54 (0.015) |

Q-mean loss s.t. DP is within  $\epsilon$  (actual DP in parentheses)

# Bounds on DP and EO

## Theorem

For each of  $k \in \{DP, EO\}$ , the relation between the statistical estimate of the constraint given batches of samples,  $z_S$ ,  $\mathbb{E}_{z_S}[\text{const}^k(z_S)]$ , and the empirical estimate for  $B$  batches of samples is listed below. Given that  $\text{const}^k(z_S) \leq 1$ , for a fixed  $\delta \in (0, 1)$  with a probability at least  $1 - \delta$  over a draw of  $B$  batches of samples from  $(h(X), \mathcal{A}, Y)$ , where  $h \in \mathcal{H}$ ,

$$\mathbb{E} [\text{const}^k(z_S)] \leq \frac{1}{B} \sum_{\ell=1}^B \text{const}^k (z_S^{(\ell)}) + 2\Omega_k + C \sqrt{\frac{\log(\frac{1}{\delta})}{B}}$$

$$\Omega_{DP, EO} = \inf_{\mu > 0} \left\{ \mu + \sqrt{\frac{2 \log (\mathcal{N}_\infty(\mathcal{H}, \mu/2S))}{B}} \right\}$$

Bounds on covering numbers leads to the following (Dutting *et al.* [10]),  
For the network with  $R$  hidden layers,  $D$  parameters, and vector of all  
model parameters  $\| w \|_1 \leq W$ . Given that  $w_l$  is bounded, the output of  
the network is bounded by some constant  $L$ .

$$\Omega_{DP} = \Omega_{EO} \leq \mathcal{O} \left( \sqrt{RD \frac{\log(WBSL)}{B}} \right)$$

# Fairness in DL: Getting Started

What-If Tool

GET STARTED

TUTORIALS

DEMONS

FAQs

GET INVOLVED 

GITHUB 

## Notebook Demos

Explore the What-If Tool's interpretability features in utmost detail in Colaboratory, Jupyter and Cloud AI Notebooks.

Compare income classification on UCI census data

DATA SOURCE

[UCI Census Income Dataset](#)

Compare two binary classification models that predict whether a person earns more than \$50k a year, based on their census information. Examine how different features affect each models' prediction, in relation to each other.

Explore age-prediction regression on UCI census data



DATA SOURCE

[UCI Census Income Dataset](#)

Explore the performance of a regression model which predicts a person's age from their census information. Slice your dataset to evaluate performance metrics such as aggregated inference error measures for each subgroup.

[Web demos](#)

[Notebook demos](#)

[Cloud AI models](#)

Source: [pair-code.github.io/what-if-tool/explore/](https://pair-code.github.io/what-if-tool/explore/)



INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY  
HYDERABAD

# Fairness in DL: Getting Started

The screenshot shows the homepage of the AI Fairness 360 website. At the top, there's a header with the text "THE LINUX FOUNDATION PROJECTS" and logos for "AI Fairness 360" and "LF AI". To the right are links for "Quick Start", "Docs", "GitHub", and a search icon. Below the header, the main title "AI Fairness 360: Understand and mitigate bias in ML models" is displayed in large white font. A detailed description follows: "AI Fairness 360, an LF AI incubation project, is an extensible open source toolkit that can help users examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. It is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. The toolkit is available in both Python and R. IBM moved AI Fairness 360 to LF AI in July 2020." Two prominent buttons at the bottom are "Get Started" and "Join the Conversation".

Source: [ai-fairness-360.org/](https://ai-fairness-360.org/)

# Privacy in Deep Learning

# What is Differential Privacy (DP)?



[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)



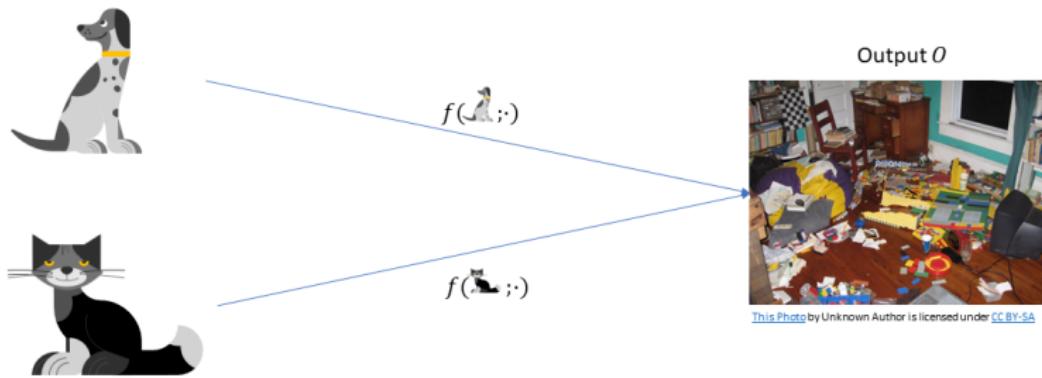
INTERNATIONAL INSTITUTE OF  
INFORMATION TECHNOLOGY  
HYDERABAD

# What is Differential Privacy (DP)?

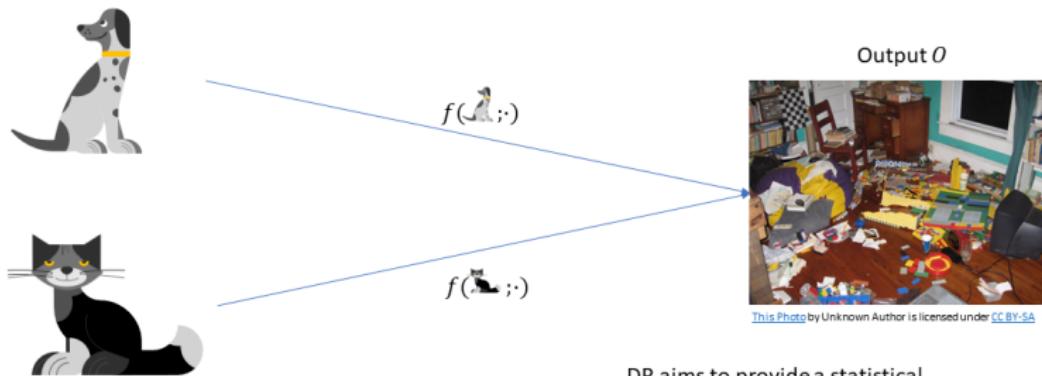


[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

# What is Differential Privacy (DP)?

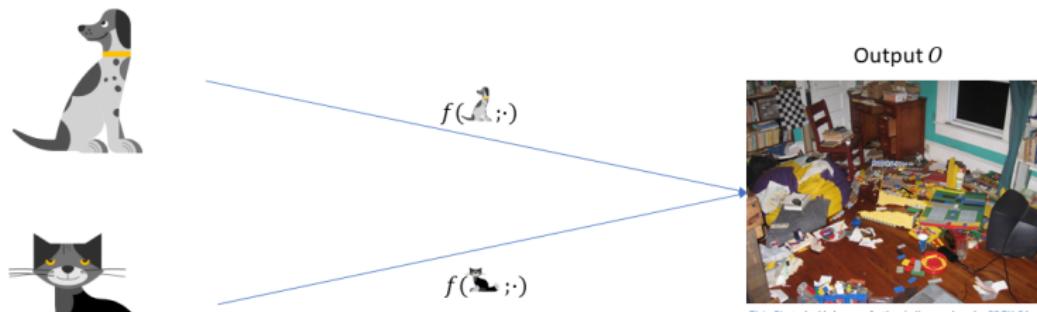


# What is Differential Privacy (DP)?



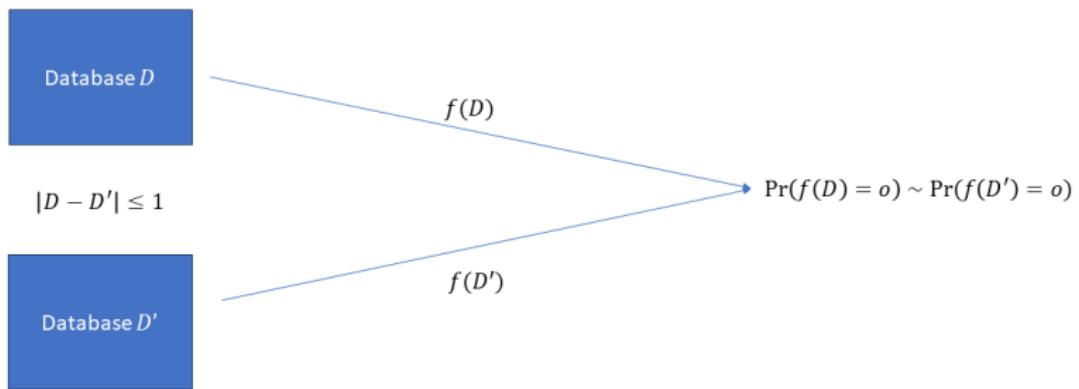
DP aims to provide a statistical guarantee against a database that the inclusion or exclusion of any single entry will not significantly impact the results of the statistical analysis.

# What is Differential Privacy (DP)?

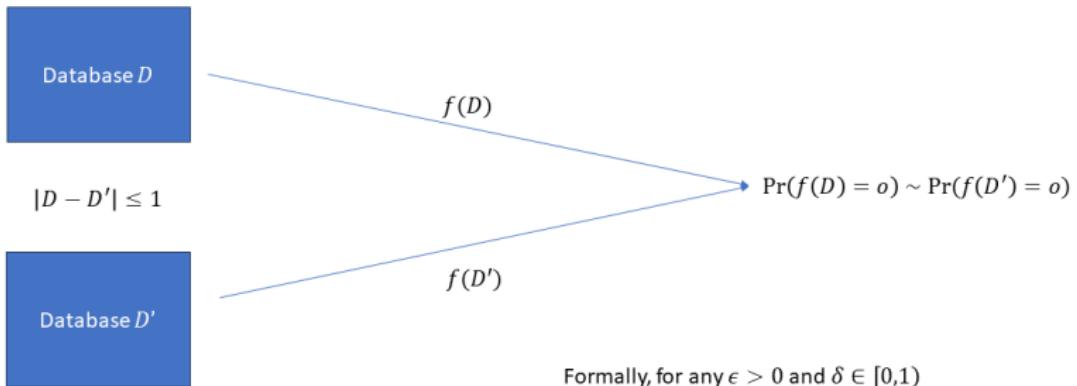


Both the cat and the dog have **Plausible Deniability** 😊

# What is Differential Privacy (DP)?



# What is Differential Privacy (DP)?



# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

Why is this a reasonable notion of privacy?

- ① Inclusion/exclusion of a user's record does not (significantly) change the output

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

Why is this a reasonable notion of privacy?

- ① Inclusion/exclusion of a user's record does not (significantly) change the output
- ② Thus, an adversary looking at the output cannot tell if the user was part of the dataset

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

Why is this a reasonable notion of privacy?

- ① Inclusion/exclusion of a user's record does not (significantly) change the output
- ② Thus, an adversary looking at the output cannot tell if the user was part of the dataset
- ③ Or, if a user's existence in a dataset is protected, their data is too

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

Why is this a reasonable notion of privacy?

- ① Inclusion/exclusion of a user's record does not (significantly) change the output
- ② Thus, an adversary looking at the output cannot tell if the user was part of the dataset
- ③ Or, if a user's existence in a dataset is protected, their data is too
- ④ DP protects against reconstruction, membership attacks

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

- ① Smaller the  $\epsilon, \delta$ , better the privacy

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

- ① Smaller the  $\epsilon, \delta$ , better the privacy
- ②  $\epsilon \approx 1$  (or lower) is reasonable

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

- ① Smaller the  $\epsilon, \delta$ , better the privacy
- ②  $\epsilon \approx 1$  (or lower) is reasonable
- ③  $\delta < \mathcal{O}(1/n)$  (or lower), where  $n$  is the dataset's size

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

- ① Smaller the  $\epsilon, \delta$ , better the privacy
- ②  $\epsilon \approx 1$  (or lower) is reasonable
- ③  $\delta < \mathcal{O}(1/n)$  (or lower), where  $n$  is the dataset's size
- ④  $\delta = 0 \implies$  "pure"-DP

# Differential Privacy (DP)

## Definition (Differential Privacy (DP) [11, 12, 13])

For a set of databases  $\mathcal{X}$  and the set of noisy outputs  $\mathcal{Y}$ , a randomized algorithm  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  is said to be  $(\epsilon, \delta)$ -DP if  $\forall D, D' \in \mathcal{X}$ , s.t.  $|D - D'| \leq 1$ , and  $\forall S \subseteq \mathcal{Y}$  the following holds,

$$\Pr[\mathcal{M}(D) \in S] \leq \exp(\epsilon) \Pr[\mathcal{M}(D') \in S] + \delta.$$

- ① Smaller the  $\epsilon, \delta$ , better the privacy
- ②  $\epsilon \approx 1$  (or lower) is reasonable
- ③  $\delta < \mathcal{O}(1/n)$  (or lower), where  $n$  is the dataset's size
- ④  $\delta = 0 \implies$  "pure"-DP
- ⑤  $\delta > 0 \implies$  "approximate"-DP (ignore unlikely events)

# Differential Privacy: Real-World Deployments

- ① U.S. Census Bureau (for showing commuting patterns)

# Differential Privacy: Real-World Deployments

- ① U.S. Census Bureau (for showing commuting patterns)
- ② Google's RAPPOR (for learning statistics about unwanted software hijacking users' settings)

# Differential Privacy: Real-World Deployments

- ① U.S. Census Bureau (for showing commuting patterns)
- ② Google's RAPPOR (for learning statistics about unwanted software hijacking users' settings)
- ③ Apple iOS 10 (for use in Intelligent personal assistant technology)

# Differential Privacy: Real-World Deployments

- ① U.S. Census Bureau (for showing commuting patterns)
- ② Google's RAPPOR (for learning statistics about unwanted software hijacking users' settings)
- ③ Apple iOS 10 (for use in Intelligent personal assistant technology)
- ④ Microsoft (for telemetry in Windows)

# Additive Noise Mechanisms

- Numeric queries  $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ , mapping the database to  $k$  real numbers, is the most fundamental type of query

# Additive Noise Mechanisms

- Numeric queries  $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ , mapping the database to  $k$  real numbers, is the most fundamental type of query
- E.g.,  $D \in \mathcal{X}$ ,  $f(D) = \sum_{x_i \in D} x_i$

# Additive Noise Mechanisms

- Numeric queries  $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ , mapping the database to  $k$  real numbers, is the most fundamental type of query
- E.g.,  $D \in \mathcal{X}$ ,  $f(D) = \sum_{x_i \in D} x_i$
- **Sensitivity:** how much  $f(D)$  changes if a single  $x \in D$  is modified

$$\Delta = \max_{D, D' \text{ s.t. } ||D - D'||_1} ||f(D) - f(D')||_1$$

# Additive Noise Mechanisms

- Numeric queries  $f : \mathbb{N}^{|\mathcal{X}|} \rightarrow \mathbb{R}^k$ , mapping the database to  $k$  real numbers, is the most fundamental type of query
- E.g.,  $D \in \mathcal{X}$ ,  $f(D) = \sum_{x_i \in D} x_i$
- **Sensitivity:** how much  $f(D)$  changes if a single  $x \in D$  is modified

$$\Delta = \max_{D, D' \text{ s.t. } ||D - D'||_1} ||f(D) - f(D')||_1$$

- **Laplace Mechanism:**  $f(D) + Z$  where  $Z \sim \text{Laplace}(0, \Delta/\epsilon)$  is  $(\epsilon, 0)$ -DP

# Additive Noise Mechanisms

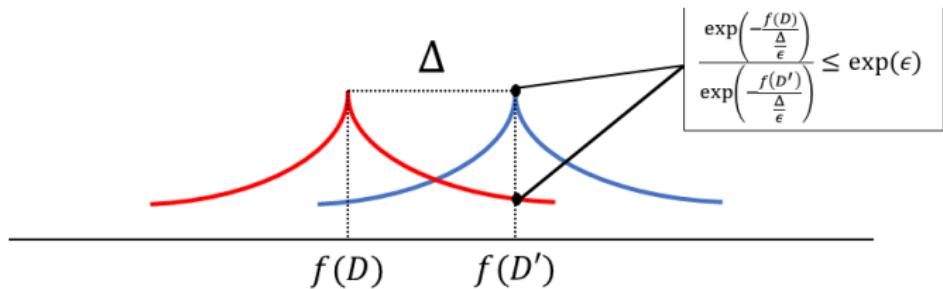


Figure: Laplace Mechanism: Illustration

# Additive Noise Mechanisms

## Laplace Mechanism

- **Sensitivity:** how much  $f(D)$  changes if a single  $x \in D$  is modified

$$\Delta = \max_{D, D' \text{ s.t. } ||D - D'||_1} ||f(D) - f(D')||_1$$

- **Laplace Mechanism:**

$f(D) + Z$  where

$Z \sim \text{Laplace}(0, \Delta/\epsilon)$  is  
 $(\epsilon, 0)$ -DP

# Additive Noise Mechanisms

## Laplace Mechanism

- **Sensitivity:** how much  $f(D)$  changes if a single  $x \in D$  is modified

$$\Delta = \max_{D, D' \text{ s.t. } ||D - D'||_1} ||f(D) - f(D')||_1$$

- **Laplace Mechanism:**

$f(D) + Z$  where  
 $Z \sim \text{Laplace}(0, \Delta/\epsilon)$  is  
 $(\epsilon, 0)$ -DP

## Gaussian Mechanism

- **Sensitivity:** how much  $f(D)$  changes if a single  $x \in D$  is modified

$$\Delta = \max_{D, D' \text{ s.t. } ||D - D'||_1} ||f(D) - f(D')||_2$$

# Additive Noise Mechanisms

## Laplace Mechanism

- **Sensitivity:** how much  $f(D)$  changes if a single  $x \in D$  is modified

$$\Delta = \max_{D, D' \text{ s.t. } ||D - D'||_1} ||f(D) - f(D')||_1$$

- **Laplace Mechanism:**  
 $f(D) + Z$  where  
 $Z \sim \text{Laplace}(0, \Delta/\epsilon)$  is  
 $(\epsilon, 0)$ -DP

## Gaussian Mechanism

- **Sensitivity:** how much  $f(D)$  changes if a single  $x \in D$  is modified

$$\Delta = \max_{D, D' \text{ s.t. } ||D - D'||_1} ||f(D) - f(D')||_2$$

- **Gaussian Mechanism:**

$f(D) + Z$  where

$Z \sim \mathcal{N}\left(0, \left(\frac{\Delta \log(1/\delta)}{\epsilon}\right)^2\right)$  is  
 $(\epsilon, \delta)$ -DP

# Properties of DP

- **Post-processing:** An adversary cannot infer additional information outside of the differentially private guarantee (no “undo” button)

# Properties of DP

- **Post-processing:** An adversary cannot infer additional information outside of the differentially private guarantee (no “undo” button)
- **Group-privacy:** If  $\|D - D'\|_1 \leq k \implies (k\epsilon, ke^{(k-1)\epsilon}\delta)$ -DP

# Properties of DP

- **Post-processing:** An adversary cannot infer additional information outside of the differentially private guarantee (no “undo” button)
- **Group-privacy:** If  $\|D - D'\|_1 \leq k \implies (k\epsilon, ke^{(k-1)\epsilon}\delta)$ -DP
- **Composition:**  $(\epsilon, \delta)$ -DP for **one** query! For “ $k$ ” queries?

# Properties of DP

- **Post-processing:** An adversary cannot infer additional information outside of the differentially private guarantee (no “undo” button)
- **Group-privacy:** If  $\|D - D'\|_1 \leq k \implies (k\epsilon, k\epsilon^{(k-1)\epsilon}\delta)$ -DP
- **Composition:**  $(\epsilon, \delta)$ -DP for **one** query! For “ $k$ ” queries?
  - ▶ **Basic Composition:**  $(k\epsilon, k\delta)$ -DP (the privacy parameters “add-up”)

# Properties of DP

- **Post-processing:** An adversary cannot infer additional information outside of the differentially private guarantee (no “undo” button)
- **Group-privacy:** If  $\|D - D'\|_1 \leq k \implies (k\epsilon, k\epsilon^{(k-1)\epsilon}\delta)$ -DP
- **Composition:**  $(\epsilon, \delta)$ -DP for **one** query! For “ $k$ ” queries?
  - ▶ **Basic Composition:**  $(k\epsilon, k\delta)$ -DP (the privacy parameters “add-up”)
  - ▶ **Advanced Composition:**  $(\epsilon\sqrt{k \log(1/\delta)}, k\delta)$ -DP (we only pay  $\sqrt{k}$  times the privacy cost)

# Properties of DP

- **Post-processing:** An adversary cannot infer additional information outside of the differentially private guarantee (no “undo” button)
- **Group-privacy:** If  $\|D - D'\|_1 \leq k \implies (k\epsilon, ke^{(k-1)\epsilon}\delta)$ -DP
- **Composition:**  $(\epsilon, \delta)$ -DP for **one** query! For “ $k$ ” queries?
  - ▶ **Basic Composition:**  $(k\epsilon, k\delta)$ -DP (the privacy parameters “add-up”)
  - ▶ **Advanced Composition:**  $(\epsilon\sqrt{k \log(1/\delta)}, k\delta)$ -DP (we only pay  $\sqrt{k}$  times the privacy cost)
  - ▶ That is, advanced composition with 10,000 queries gives privacy guarantees comparable to basic composition with 100 queries

# Differentially Private SGD (DP-SGD)

## Stochastic Gradient Descent (SGD)

- ① Choose a random batch  $B$  from the dataset

# Differentially Private SGD (DP-SGD)

## Stochastic Gradient Descent (SGD)

- ① Choose a random batch  $B$  from the dataset
- ② Compute the gradients:  $\frac{1}{|B|} \sum_{(x,y) \in B} \nabla \mathcal{L}(\theta; x, y)$

# Differentially Private SGD (DP-SGD)

## Stochastic Gradient Descent (SGD)

- ① Choose a random batch  $B$  from the dataset
- ② Compute the gradients:  $\frac{1}{|B|} \sum_{(x,y) \in B} \nabla \mathcal{L}(\theta; x, y)$
- ③ Update the weights in the negative direction of the gradient

# Differentially Private SGD (DP-SGD)

## Stochastic Gradient Descent (SGD)

- ① Choose a random batch  $B$  from the dataset
- ② Compute the gradients:  $\frac{1}{|B|} \sum_{(x,y) \in B} \nabla \mathcal{L}(\theta; x, y)$
- ③ Update the weights in the negative direction of the gradient
- ④ Repeat  $k$  times

# Differentially Private SGD (DP-SGD)

## Differentially Private Stochastic Gradient Descent (DP-SGD) [1]

- ① Sample a batch  $B$  from the dataset

# Differentially Private SGD (DP-SGD)

## Differentially Private Stochastic Gradient Descent (DP-SGD) [1]

- ① Sample a batch  $B$  from the dataset
- ② For each  $(x, y) \in B$

# Differentially Private SGD (DP-SGD)

## Differentially Private Stochastic Gradient Descent (DP-SGD) [1]

- ① Sample a batch  $B$  from the dataset
- ② For each  $(x, y) \in B$ 
  - i Compute the gradient  $\nabla \mathcal{L}(\theta; x, y)$

# Differentially Private SGD (DP-SGD)

## Differentially Private Stochastic Gradient Descent (DP-SGD) [1]

- ① Sample a batch  $B$  from the dataset
- ② For each  $(x, y) \in B$ 
  - i Compute the gradient  $\nabla \mathcal{L}(\theta; x, y)$
  - ii **Clip** the gradient so that the  $l_2$ -norm is at most  $c$

# Differentially Private SGD (DP-SGD)

## Differentially Private Stochastic Gradient Descent (DP-SGD) [1]

- ① Sample a batch  $B$  from the dataset
- ② For each  $(x, y) \in B$ 
  - i Compute the gradient  $\nabla \mathcal{L}(\theta; x, y)$
  - ii **Clip** the gradient so that the  $l_2$ -norm is at most  $c$
- ③ Average the gradients and add **noise** to the average using the **Gaussian** mechanism

# Differentially Private SGD (DP-SGD)

## Differentially Private Stochastic Gradient Descent (DP-SGD) [1]

- ➊ Sample a batch  $B$  from the dataset
- ➋ For each  $(x, y) \in B$ 
  - i Compute the gradient  $\nabla \mathcal{L}(\theta; x, y)$
  - ii **Clip** the gradient so that the  $l_2$ -norm is at most  $c$
- ➌ Average the gradients and add **noise** to the average using the **Gaussian** mechanism
- ➍ Update the weights in the negative direction of the gradient

# Differentially Private SGD (DP-SGD)

## Differentially Private Stochastic Gradient Descent (DP-SGD) [1]

- ➊ Sample a batch  $B$  from the dataset
- ➋ For each  $(x, y) \in B$ 
  - i Compute the gradient  $\nabla \mathcal{L}(\theta; x, y)$
  - ii **Clip** the gradient so that the  $l_2$ -norm is at most  $c$
- ➌ Average the gradients and add **noise** to the average using the **Gaussian** mechanism
- ➍ Update the weights in the negative direction of the gradient
- ➎ Repeat  $k$  times

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP
- Sampling a random batch  $B$  gives  $(\frac{\epsilon B}{n}, \frac{\delta B}{n})$ -DP

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP
- Sampling a random batch  $B$  gives  $(\frac{\epsilon B}{n}, \frac{\delta B}{n})$ -DP
  - ▶ Privacy amplification by subsampling [30]

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP
- Sampling a random batch  $B$  gives  $(\frac{\epsilon B}{n}, \frac{\delta B}{n})$ -DP
  - ▶ Privacy amplification by subsampling [30]
- $k$  steps of DP-SGD will give

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP
- Sampling a random batch  $B$  gives  $(\frac{\epsilon B}{n}, \frac{\delta B}{n})$ -DP
  - ▶ Privacy amplification by subsampling [30]
- $k$  steps of DP-SGD will give
  - ▶ (Advanced)  $(\frac{\epsilon \sqrt{k \log(1/\delta)} B}{n}, \frac{\delta' B}{n})$ -DP

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP
- Sampling a random batch  $B$  gives  $(\frac{\epsilon B}{n}, \frac{\delta B}{n})$ -DP
  - ▶ Privacy amplification by subsampling [30]
- $k$  steps of DP-SGD will give
  - ▶ (Advanced)  $(\frac{\epsilon \sqrt{k \log(1/\delta)} B}{n}, \frac{\delta' B}{n})$ -DP
  - ▶ (Moments Accountant [1])  $(\frac{\epsilon \sqrt{k} B}{n}, \frac{\delta' B}{n})$ -DP

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP
- Sampling a random batch  $B$  gives  $(\frac{\epsilon B}{n}, \frac{\delta B}{n})$ -DP
  - ▶ Privacy amplification by subsampling [30]
- $k$  steps of DP-SGD will give
  - ▶ (Advanced)  $(\frac{\epsilon \sqrt{k \log(1/\delta)} B}{n}, \frac{\delta' B}{n})$ -DP
  - ▶ (Moments Accountant [1])  $(\frac{\epsilon \sqrt{k} B}{n}, \frac{\delta' B}{n})$ -DP

# DP-SGD: Privacy Guarantees

- We add noise using the Gaussian mechanism
  - ▶ Each step in DP-SGD is  $(\epsilon, \delta)$ -DP
- Sampling a random batch  $B$  gives  $(\frac{\epsilon B}{n}, \frac{\delta B}{n})$ -DP
  - ▶ Privacy amplification by subsampling [30]
- $k$  steps of DP-SGD will give
  - ▶ (Advanced)  $(\frac{\epsilon \sqrt{k \log(1/\delta)} B}{n}, \frac{\delta' B}{n})$ -DP
  - ▶ (Moments Accountant [1])  $(\frac{\epsilon \sqrt{k} B}{n}, \frac{\delta' B}{n})$ -DP

**Implementation:** [github.com/tensorflow/privacy](https://github.com/tensorflow/privacy)

# DP-SGD: Does it Work?

- MNIST [32]: Classify black and white images as digits
- An “easy” ML task
  - ▶ Non-private ML model’s test accuracy  $\approx 100\%$
- Abadi *et al.* [1]: 97% for  $\epsilon = 8$
- Tramer and Boneh [47]: 98-99% for  $\epsilon \in [1, 3]$
- DP-SGD performs well for “easy” tasks
  - ▶ CIFAR-10: 98% (non-private) and 69% (for  $\epsilon = 3$  [47])

3 6 8 1 7 9 6 6 9 1  
6 7 5 7 8 6 3 4 8 5  
2 1 7 9 7 1 2 8 4 5  
4 8 1 9 0 1 8 8 9 4  
7 6 1 8 6 4 1 5 6 0  
7 5 9 2 6 5 8 1 9 7  
1 2 2 2 2 3 4 4 8 0  
0 2 3 8 0 7 3 8 5 7  
0 1 4 6 4 6 0 2 4 3  
7 1 2 8 1 6 9 8 6 1

**Figure:** Examples from the MNIST database  
(Credit [32])

# Other Advances

## PATE [43]: Private Aggregation of Teacher Ensembles

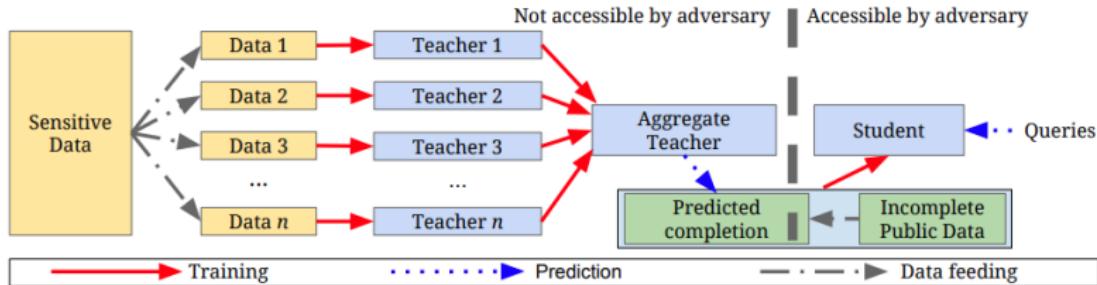


Figure: Overview of PATE (Credit [43])

# Other Advances

## PATE [43]: Private Aggregation of Teacher Ensembles

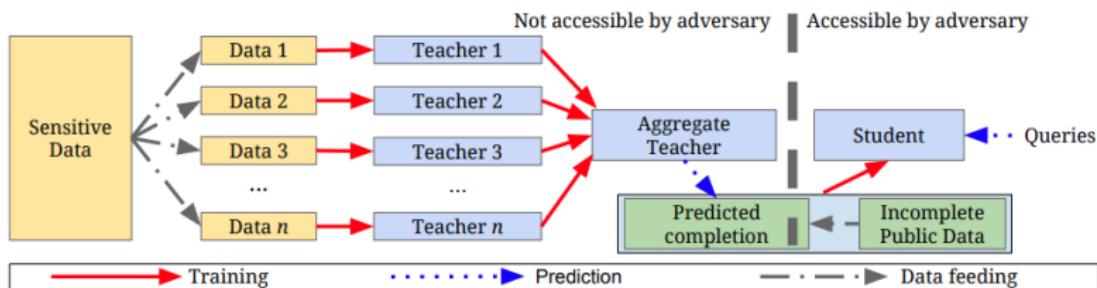


Figure: Overview of PATE (Credit [43])

Performance on MNIST: 98.5% test accuracy for  $\epsilon \approx 2$

# Other Advances

Yu *et al.* [50]: Differentially private Fine-tuning of Language Models

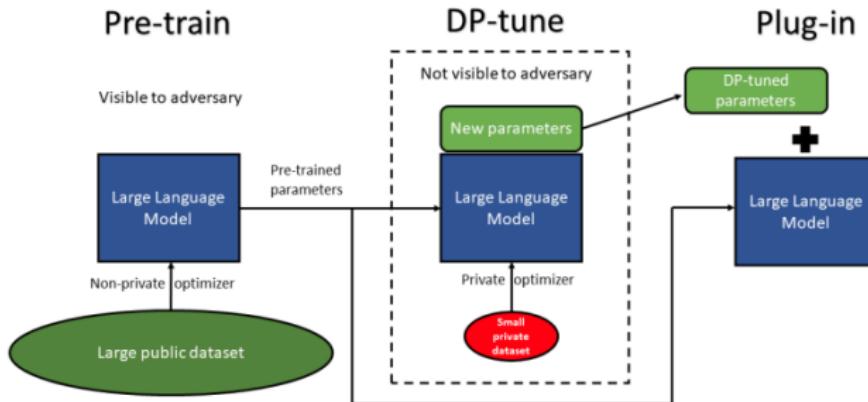


Figure: Overview of the framework (Credit [50])

# Other Advances

Yu *et al.* [50]: Differentially private Fine-tuning of Language Models

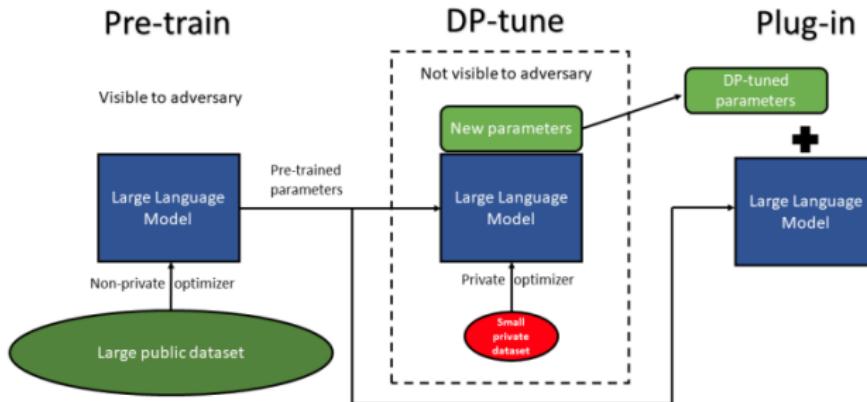


Figure: Overview of the framework (Credit [50])

Tighter privacy accounting using [19]

# DP in ML: Where to Get Started?

- **Tensorflow Privacy:** [github.com/tensorflow/privacy](https://github.com/tensorflow/privacy)

# DP in ML: Where to Get Started?

- **Tensorflow Privacy:** [github.com/tensorflow/privacy](https://github.com/tensorflow/privacy)
  - ▶ [tensorflow.org/responsible\\_ai/privacy/tutorials/classification\\_privacy](https://tensorflow.org/responsible_ai/privacy/tutorials/classification_privacy)

# DP in ML: Where to Get Started?

- **Tensorflow Privacy:** [github.com/tensorflow/privacy](https://github.com/tensorflow/privacy)
  - ▶ [tensorflow.org/responsible\\_ai/privacy/tutorials/classification\\_privacy](https://tensorflow.org/responsible_ai/privacy/tutorials/classification_privacy)
- **Opacus:** [opacus.ai/](https://opacus.ai/)

# DP in ML: Where to Get Started?

- **Tensorflow Privacy:** [github.com/tensorflow/privacy](https://github.com/tensorflow/privacy)
  - ▶ [tensorflow.org/responsible\\_ai/privacy/tutorials/classification\\_privacy](https://tensorflow.org/responsible_ai/privacy/tutorials/classification_privacy)
- **Opacus:** [opacus.ai/](https://opacus.ai/)
  - ▶ [opacus.ai/tutorials/building\\_image\\_classifier](https://opacus.ai/tutorials/building_image_classifier)

# DP in ML: Where to Get Started?

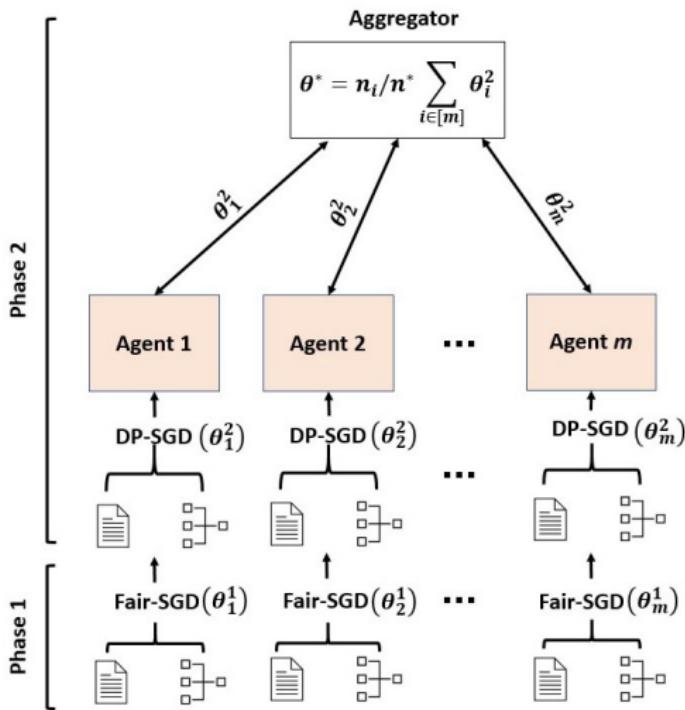
- **Tensorflow Privacy:** [github.com/tensorflow/privacy](https://github.com/tensorflow/privacy)
  - ▶ [tensorflow.org/responsible\\_ai/privacy/tutorials/classification\\_privacy](https://tensorflow.org/responsible_ai/privacy/tutorials/classification_privacy)
- **Opacus:** [opacus.ai/](https://opacus.ai/)
  - ▶ [opacus.ai/tutorials/building\\_image\\_classifier](https://opacus.ai/tutorials/building_image_classifier)
- **IBM's DP Library:**  
[github.com/IBM/differential-privacy-library](https://github.com/IBM/differential-privacy-library)

# DP in ML: Where to Get Started?

- **Tensorflow Privacy:** [github.com/tensorflow/privacy](https://github.com/tensorflow/privacy)
  - ▶ [tensorflow.org/responsible\\_ai/privacy/tutorials/classification\\_privacy](https://tensorflow.org/responsible_ai/privacy/tutorials/classification_privacy)
- **Opacus:** [opacus.ai/](https://opacus.ai/)
  - ▶ [opacus.ai/tutorials/building\\_image\\_classifier](https://opacus.ai/tutorials/building_image_classifier)
- **IBM's DP Library:**  
[github.com/IBM/differential-privacy-library](https://github.com/IBM/differential-privacy-library)
- **PINQ by Microsoft:** [github.com/LLGemini/PINQ](https://github.com/LLGemini/PINQ)

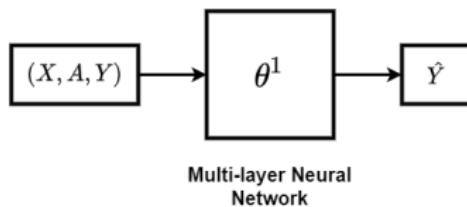
# FPFL: Fair and Private Federated Learning

# Our Model



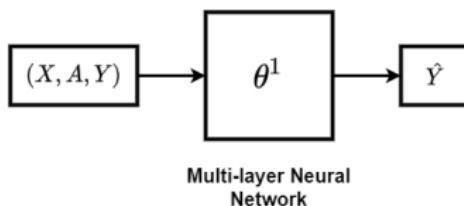
# Phase 1: Fair-SGD

For each agent,



# Phase 1: Fair-SGD

For each agent,



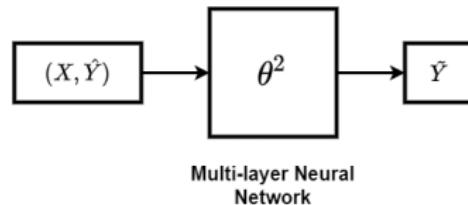
- Loss function [41]:

$$L(\cdot) = \text{Cross entropy loss}(Y, \hat{Y}) + \lambda \cdot \text{DemP Loss}(A, \hat{Y})$$

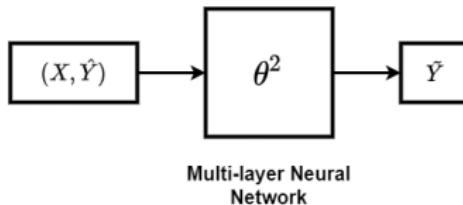
- Optimization:

$$\min_{\theta^1} \max_{\lambda} L(\cdot)$$

## Phase 2: DP-SGD



## Phase 2: DP-SGD



- We train the network in this phase to learn the predictions from Fair-SGD
- We add Gaussian noise to the gradients provided by SGD to ensure data confidentiality [1]

# Decoupling Training Process

- Protects both the sensitive attribute and the training data
- Reduces the number of epochs  $\implies$  Reduction in explosion of  $\epsilon$

# Decoupling Training Process

- Protects both the sensitive attribute and the training data
- Reduces the number of epochs  $\implies$  Reduction in explosion of  $\epsilon$
- As agents broadcast only  $\theta^2$ , our  $\epsilon, \delta$  bounds follow from moments accountant [1, Theorem 1]

Federated Learning Meets Fairness and Differential Privacy [42]

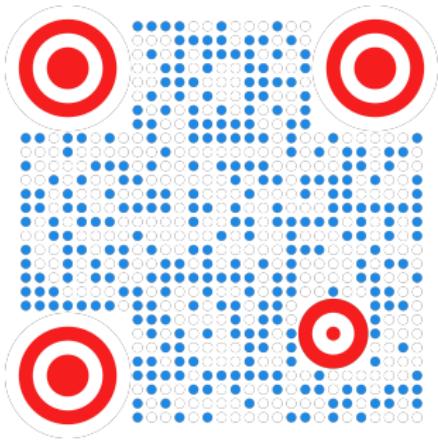
# Implementation

Demo Implementation on Colab: [Take me there!](#)

# Further Reading

- The Algorithmic Foundations of Differential Privacy [12]
- Fairness and Machine Learning [4]
- Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey [17]
- Privacy and Fairness in Federated Learning: on the Perspective of Trade-off [8]
- A systematic review of federated learning from clients' perspective: challenges and solutions [46]

# Thank you!



[github.com/magnetar-iiith/  
FairPrivateDL](https://github.com/magnetar-iiith/FairPrivateDL)



[mll.iiit.ac.in](http://mll.iiit.ac.in)

# References I

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
- [2] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54:95–122, 2018.
- [3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. <http://www.fairmlbook.org>.
- [4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.
- [5] M. Bilal Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *ArXiv e-prints*, July 2015.
- [6] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pages 13–18. IEEE, 2009.
- [7] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. Bias in machine learning software: Why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440, 2021.
- [8] Huiqiang Chen, Tianqing Zhu, Tao Zhang, Wanlei Zhou, and Philip S Yu. Privacy and fairness in federated learning: on the perspective of trade-off. *ACM Computing Surveys*, 2023.

## References II

- [9] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019.
- [10] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. *arXiv preprint arXiv:1706.03459*, 2017.
- [11] Cynthia Dwork. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12, 2006.
- [12] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Theoretical Computer Science*, 9(3-4):211–407, 2014.
- [13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.
- [14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255. URL <http://doi.acm.org/10.1145/2090236.2090255>.
- [15] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR, 2018.

## References III

- [16] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2783311. URL <http://doi.acm.org/10.1145/2783258.2783311>.
- [17] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy and fairness in decisions and learning tasks: A survey. *arXiv preprint arXiv:2202.08187*, 2022.
- [18] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.
- [19] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34:11631–11642, 2021.
- [20] Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619, 2021.
- [21] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.
- [22] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

## References IV

- [23] Max Hort, Zhenpeng Chen, Jie M Zhang, Federica Sarro, and Mark Harman. Bia mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068*, 2022.
- [24] Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6988–6995, 2022.
- [25] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, Feb 2009. doi: 10.1109/IC4.2009.4909197.
- [26] F. Kamiran and T.G.K. Calders. Classification with no discrimination by preferential sampling. In *Informal proceedings of the 19th Annual Machine Learning Conference of Belgium and The Netherlands (Benelearn'10, Leuven, Belgium, May 27-28, 2010)*, pages 1–6, 2010.
- [27] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.
- [28] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *2010 IEEE international conference on data mining*, pages 869–874. IEEE, 2010.
- [29] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, Dec 2011. doi: 10.1109/ICDMW.2011.83.
- [30] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.



# References V

- [31] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.
- [32] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [33] Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, and Baowen Xu. Training data debugging for the fairness of machine learning software. In *Proceedings of the 44th International Conference on Software Engineering*, pages 2215–2227, 2022.
- [34] Lydia T Liu, Max Simchowitz, and Moritz Hardt. The implicit fairness criterion of unconstrained learning. In *International Conference on Machine Learning*, pages 4051–4060. PMLR, 2019.
- [35] Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.
- [36] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 502–510, 2011.
- [37] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3381–3390, 2018. URL <http://proceedings.mlr.press/v80/madras18a.html>



# References VI

- [38] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.
- [39] Harikrishna Narasimhan. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1646–1654, 2018. URL <http://proceedings.mlr.press/v84/narasimhan18a.html>.
- [40] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
- [41] Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2277–2283. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/315. URL <https://doi.org/10.24963/ijcai.2020/315>. Main track.
- [42] Manisha Padala, Sankarshan Damle, and Sujit Gujar. Federated learning meets fairness and differential privacy. In *Neural Information Processing - 28th International Conference, ICONIP 2021*, pages 692–699. Springer, 2021.
- [43] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

# References VII

- [44] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017. URL <http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf>.
- [45] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8227–8236, 2019.
- [46] Yashothara Shanmugarasa, Hye-young Paik, Salil S Kanhere, and Liming Zhu. A systematic review of federated learning from clients' perspective: challenges and solutions. *Artificial Intelligence Review*, pages 1–55, 2023.
- [47] Florian Tramer and Dan Boneh. Differentially private learning needs better features (or much more data). *arXiv preprint arXiv:2011.11660*, 2020.
- [48] Berk Ustun, Yang Liu, and David Parkes. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, pages 6373–6382. PMLR, 2019.
- [49] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.
- [50] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.



# References VIII

- [51] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018. URL <http://arxiv.org/abs/1801.07593>.
- [52] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15002–15012, 2021.
- [53] Indre Žliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *2011 IEEE 11th international conference on data mining*, pages 992–1001. IEEE, 2011.