FairProof : Confidential and Certifiable Fairness for Neural Networks

Best Paper Award @Privacy-ILR Workshop ICLR 2024 🏆

Authors

Chhavi Yadav

cyadav@ucsd.edu

Amrita Roy-Chowdhury

Dan Boneh

Kamalika Chaudhuri

ML in many applications

AI for Social Good

..But a major barrier to their usage is

..But a major barrier to their usage is

Can we trust the results of these systems?

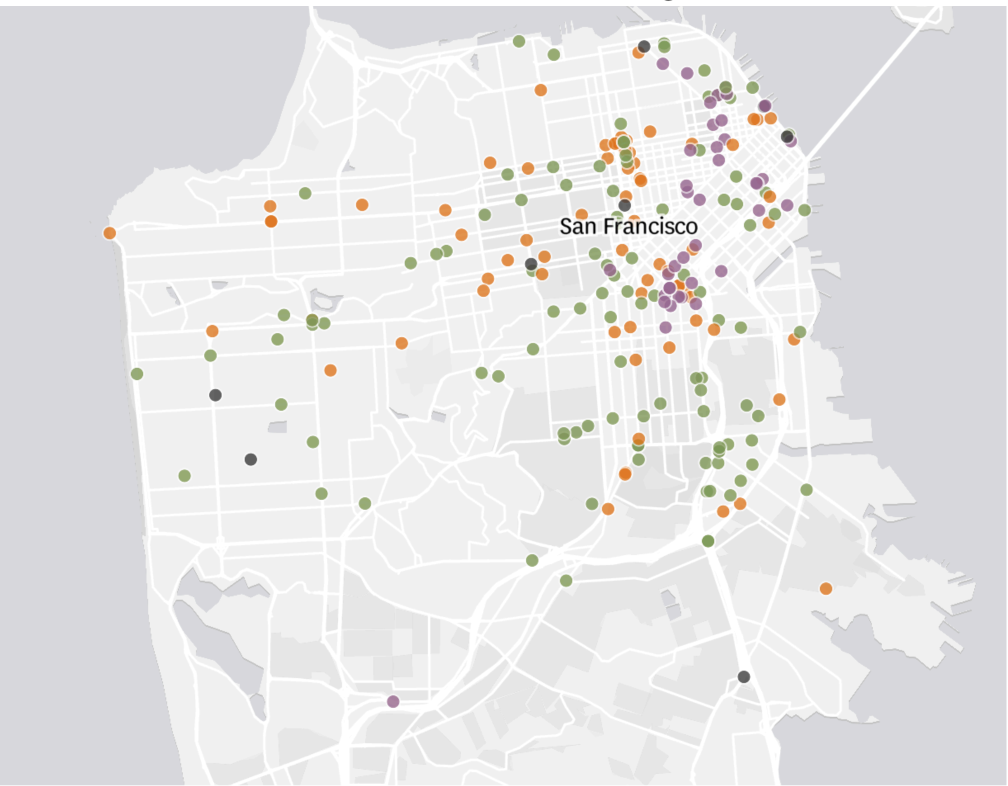
No.

San Francisco Chronicle

Map shows every crash involving driverless cars in San Francisco

By Harsha Devulapalli | Updated Oct. 24, 2023 12:41 p.m.

There are now hundreds of driverless vehicles rolling around on the streets



: Sriharsha Devulapalli / The Chronicle • Source: California Department of Motor Vehicles

AI, facial recognition technology causing false arrests across nation

Calls for regulation grow as Black men across U.S. wrongfully jailed.

Artificial intelligence may put private data at risk

HUMANS ARE BIASED. GENERATIVE AI IS EVEN WORSE

Stable Diffusion's text-to-image model amplifies stereotypes about race and gender — here's why that matters

The New York Times

OPINION

A.I. Could Worsen Health Disparities

In a health system riddled with inequity, we risk making dangerous biases automated and invisible.

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

Apple's 'sexist' credit card investigated by US regulator

Blind Trust ✗ Verify ✓

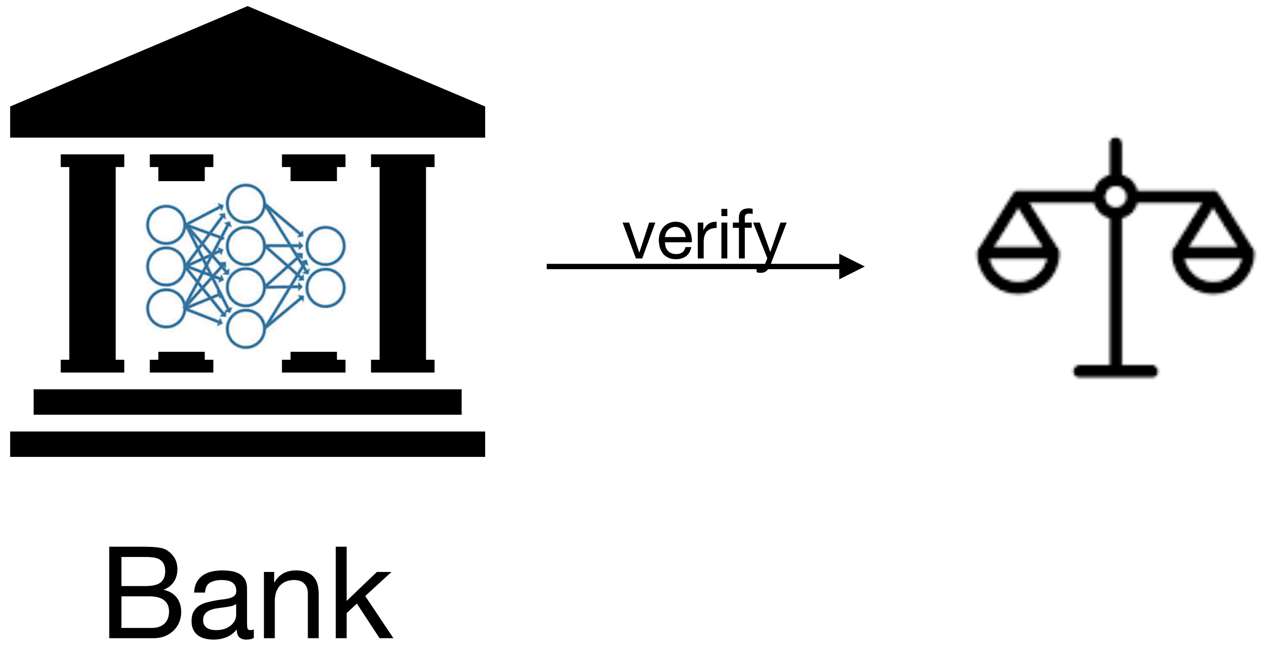
- Distrust in ML models

Blind Trust ✗ Verify ✓

- Distrust in ML models
- Verification of model properties..

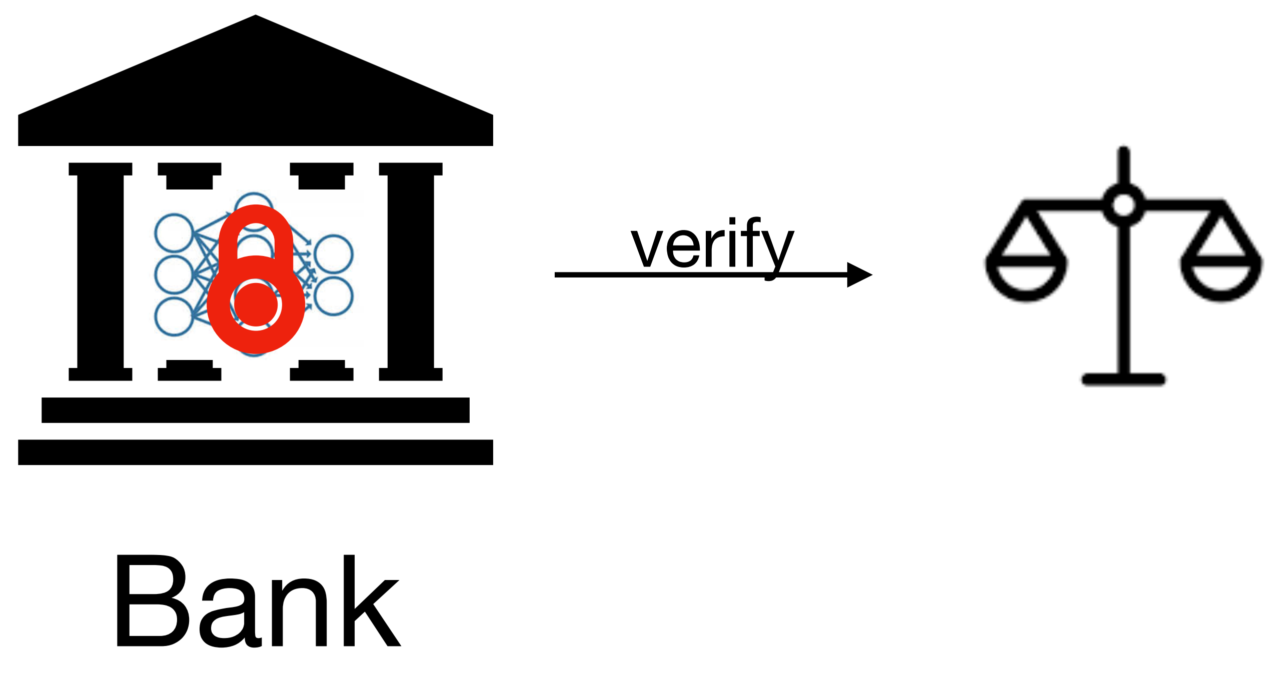
Blind Trust ✗ Verify ✓

- Distrust in ML models
- Verification of model properties..



Blind Trust ✗ Verify ✓

- Distrust in ML models
- Verification of model properties..
- Models kept confidential



How to **publicly verify properties** of a model while keeping it **confidential**?

Canonical Approach

- External Auditing : Estimation of model properties using API queries, by a third-party auditor

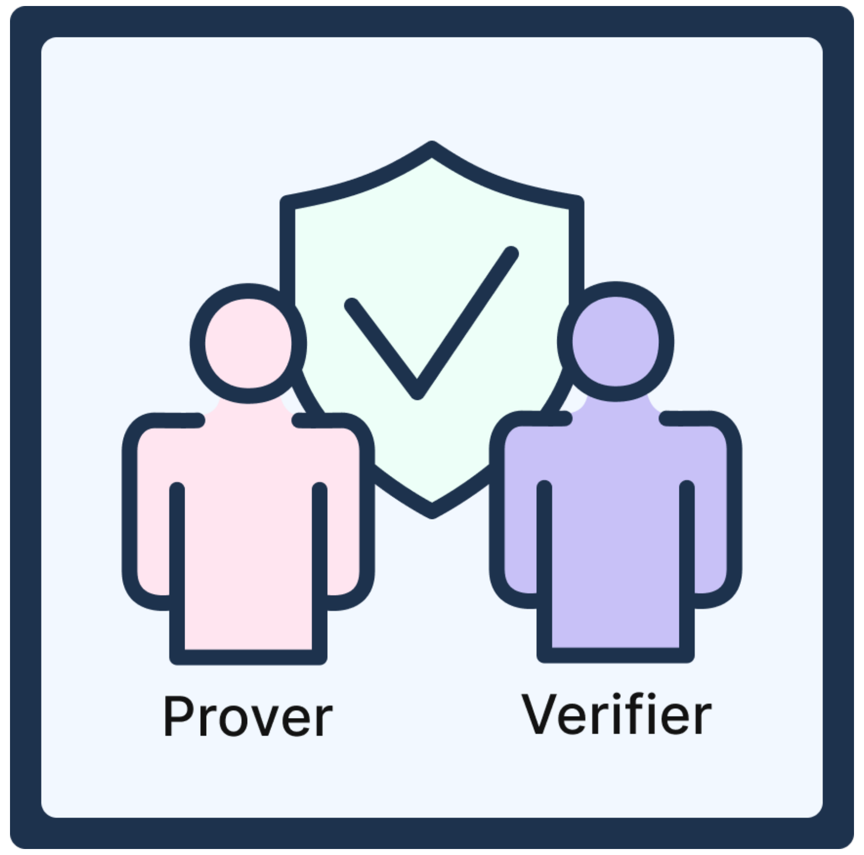
Canonical Approach

- External Auditing
  - Leaks model in the process, concerns if black-box auditing is even possible [1]
  - Model Swapping : change the model post auditing or use different models for different queries
  - Sensitive to the choice of reference auditing dataset

[1] Black-Box Access is Insufficient for Rigorous AI Audits Casper et. al.2024

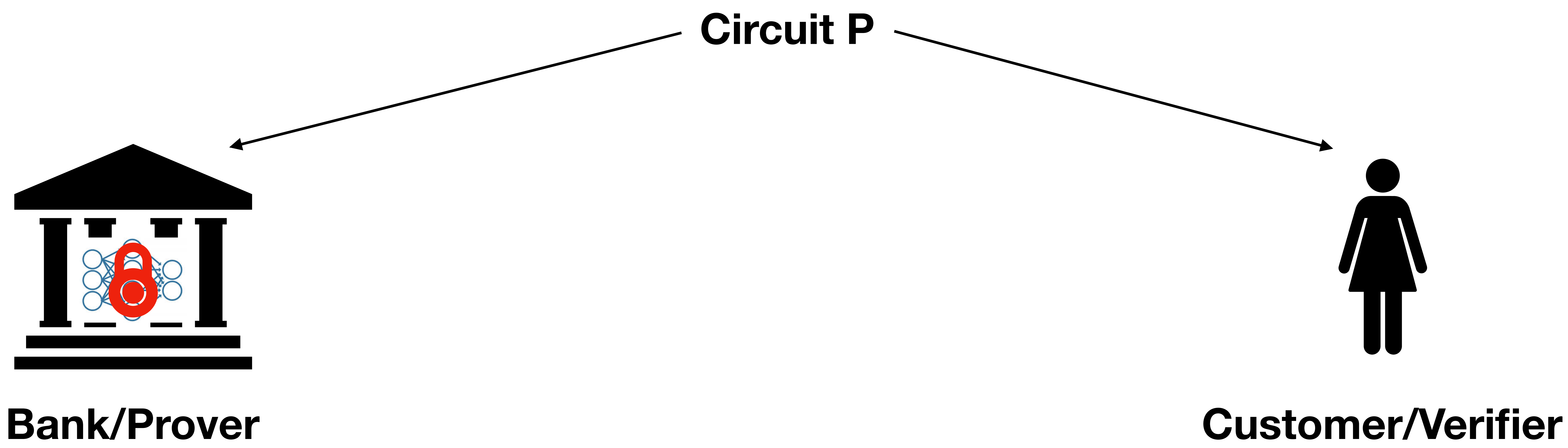
Our Solution

- Zero-Knowledge Proofs, a cryptographic primitive

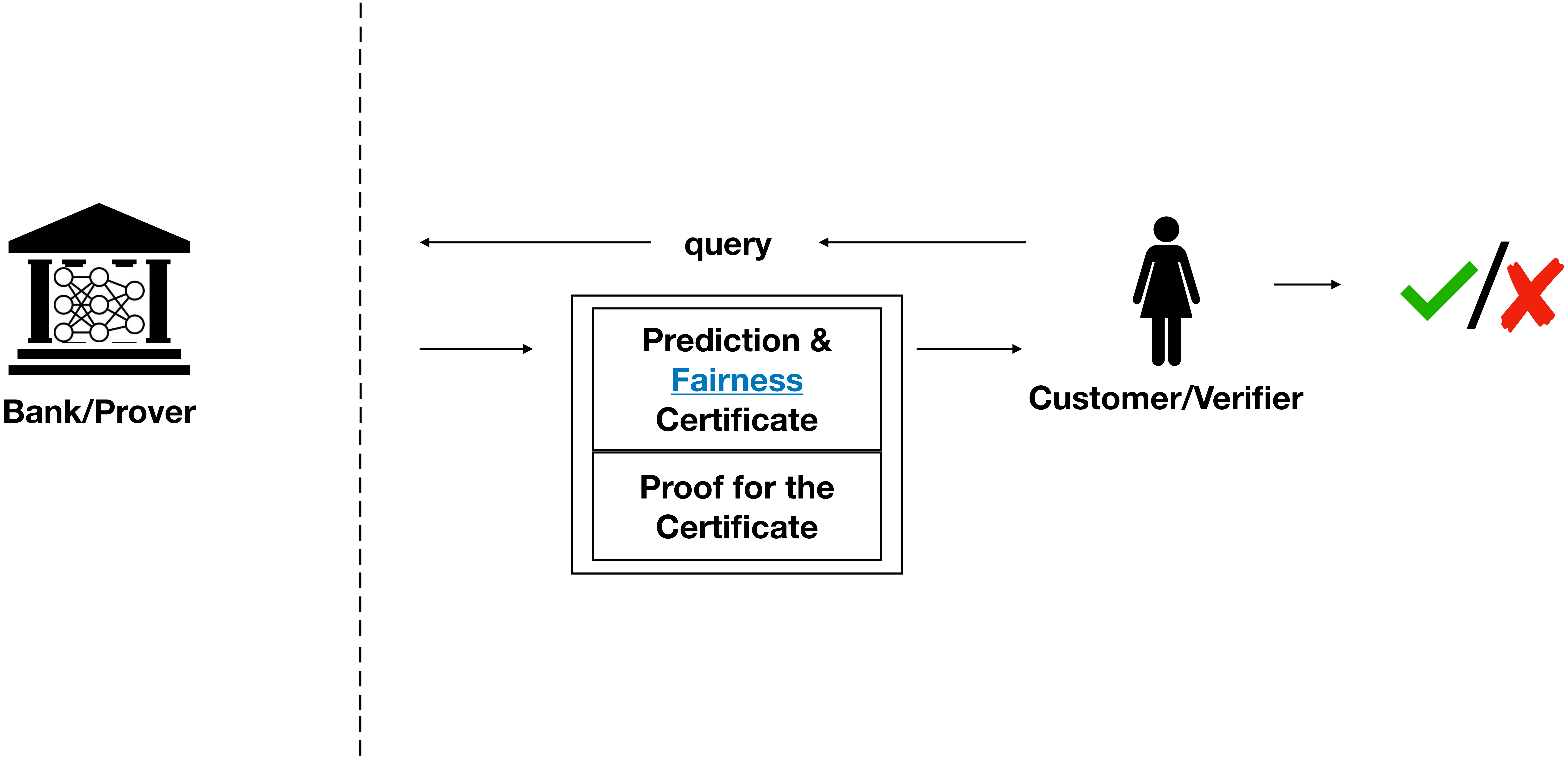


Zero-Knowledge Proofs (ZKPs)

- involve a prover and a verifier, who both have access to a circuit P
- enable prover to convince the verifier that the prover possess w s.t. P(w) = 1
- without revealing any additional information about w to the verifier



Setup for Public Verification using ZKPs



The two parts

Prediction & [Fairness] Certificate	• Fairness Certification Algorithm in-the-clear
Proof for the Certificate	• A ZKP system to prove the correct computation of this certificate

Local Individual Fairness (from Literature)

- A machine learning model $f : \mathbb{R}^n \mapsto \mathscr{Y}$ is defined to be ϵ-individually fair w.r.t to a data point $x^* \sim \mathscr{D}$ under some distance metric $d : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ if

$$\forall x : d(x^*, x) \leq \epsilon \implies f(x^*) = f(x)$$

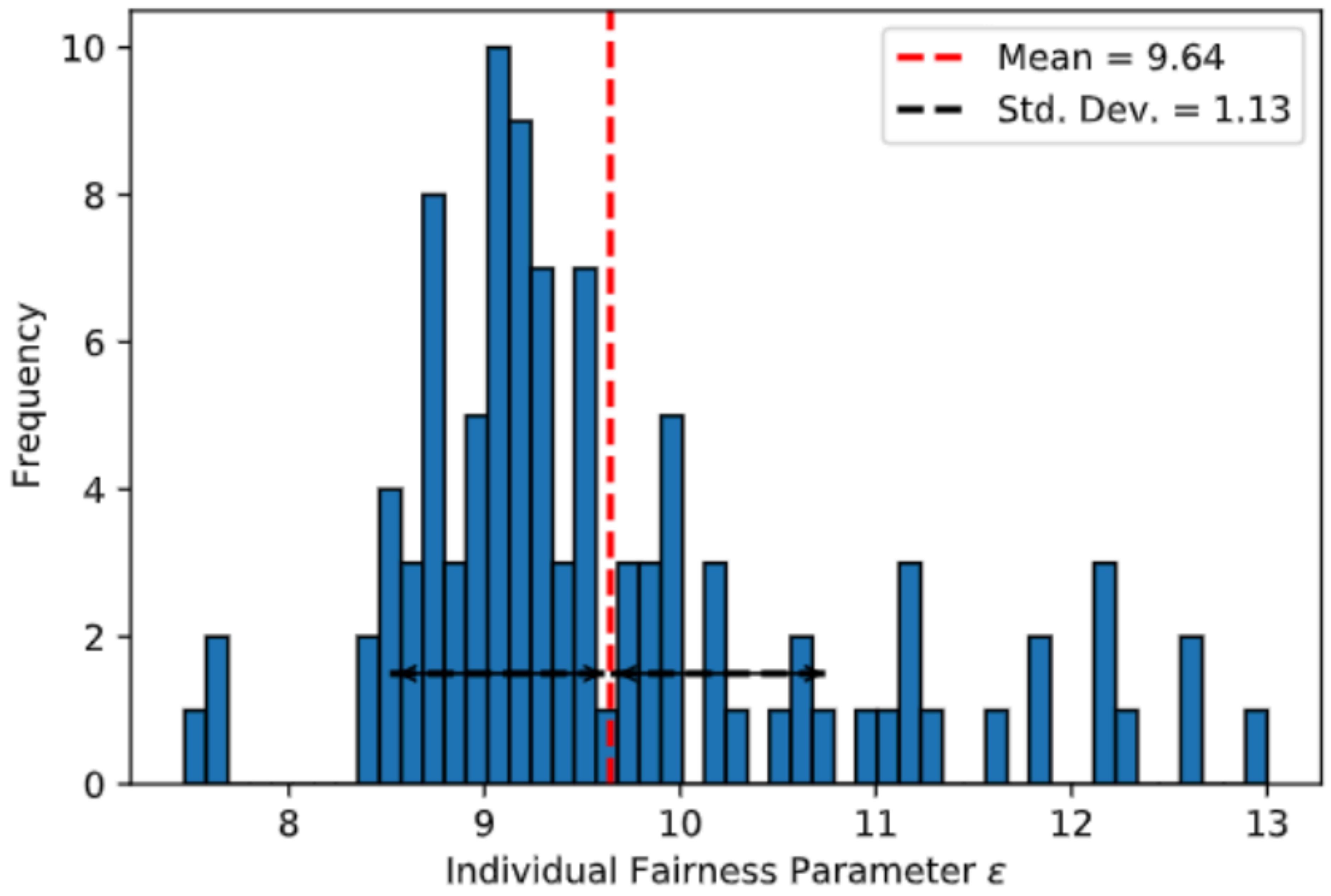
- Our certification algorithm should output this ϵ
- Notion of Sensitive attributes
- d : Weighted L2 distance with zero weights on the sensitive attributes

👩🏼‍💻 = 👩🏿‍💻 = 👩🏻‍💻 = 👩‍💻 = 👩🏽‍💻 = 👩🏾‍💻

Q1. Can our resulting certification algorithm distinguish b/n more vs. less fair models?

- Radius $\epsilon \uparrow$ fairness $\uparrow$
- Radius $\epsilon \downarrow$ fairness $\downarrow$

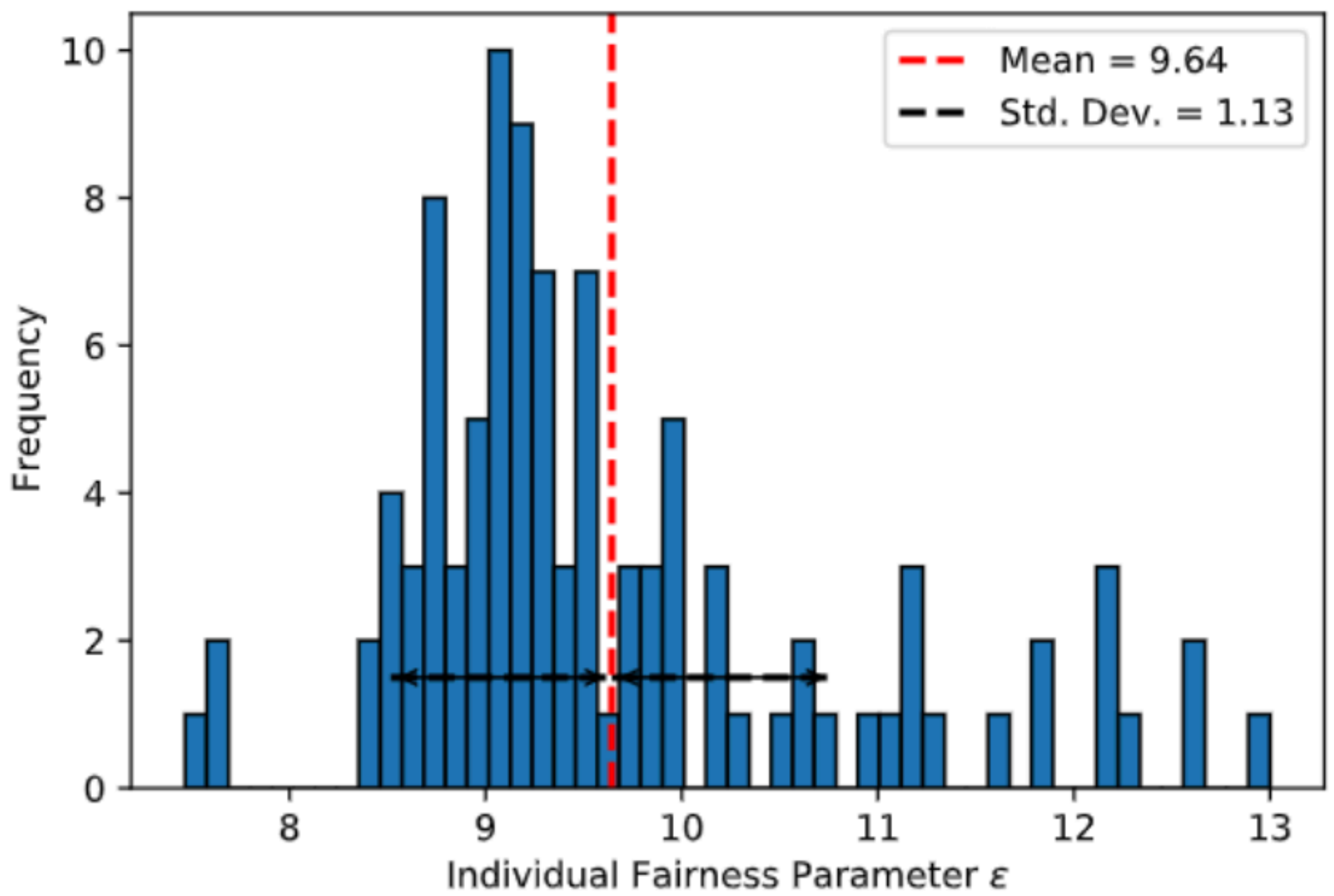
Q1. Can our resulting certification algorithm distinguish b/n more vs. less fair models?



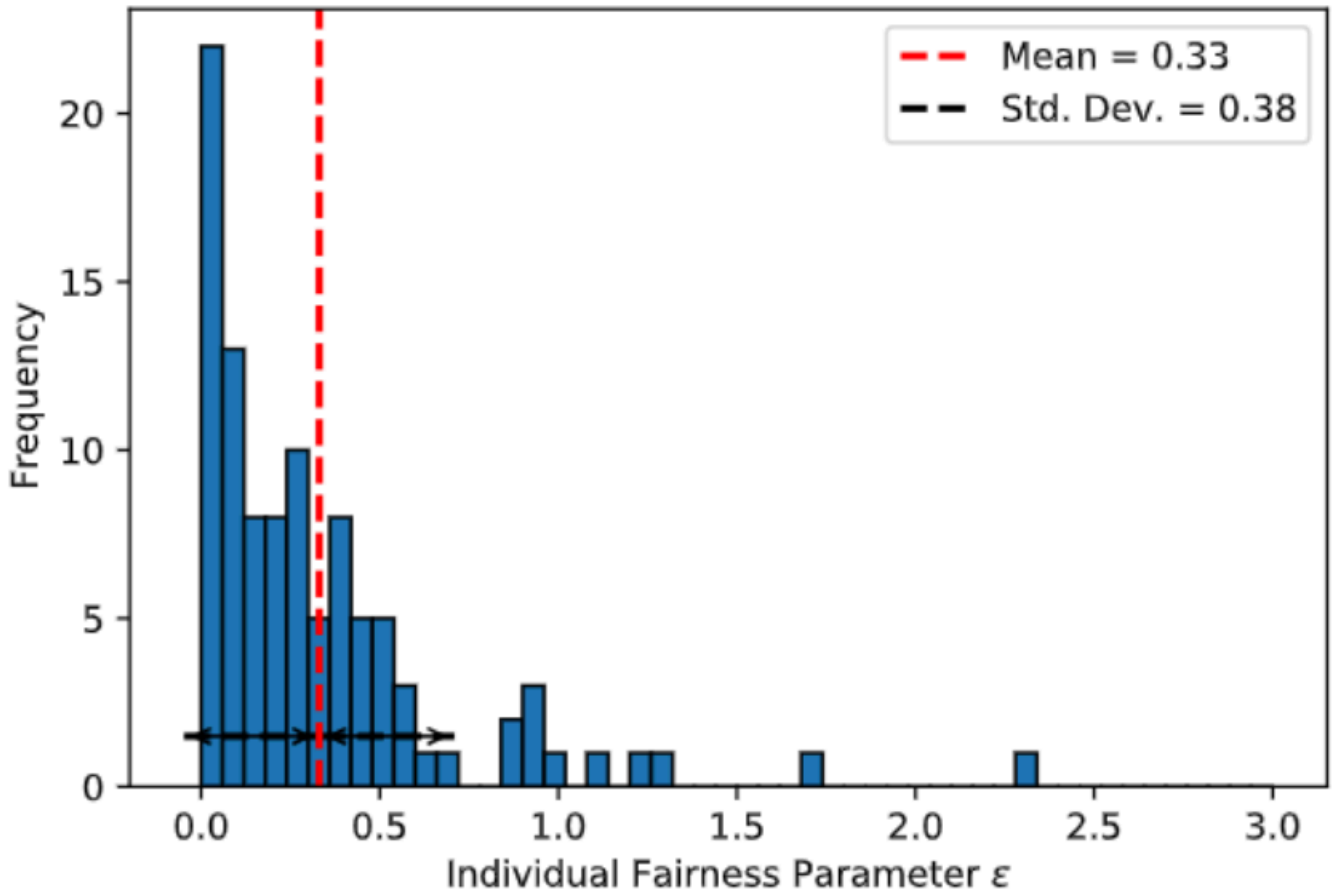
Fair model

Histogram of fairness parameter ϵ for fair & unfair models. Model Size (4,2) Credit dataset. Larger ϵ indicates more fairness

Q1. Can our resulting certification algorithm distinguish b/n more vs. less fair models?



Fair model

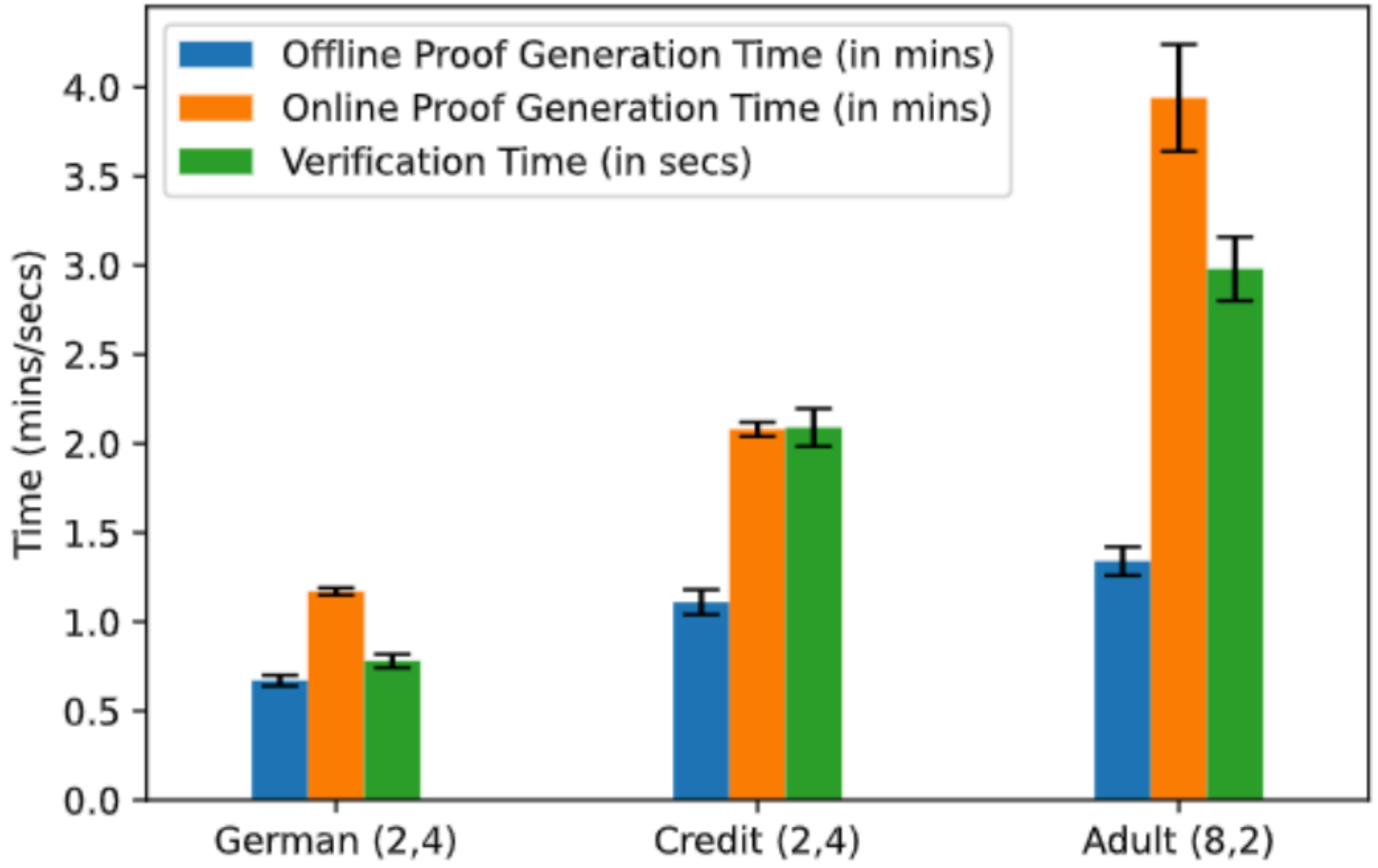


Unfair model

Histogram of fairness parameter ϵ for fair & unfair models. Model Size (4,2) Credit dataset. Larger ϵ indicates more fairness

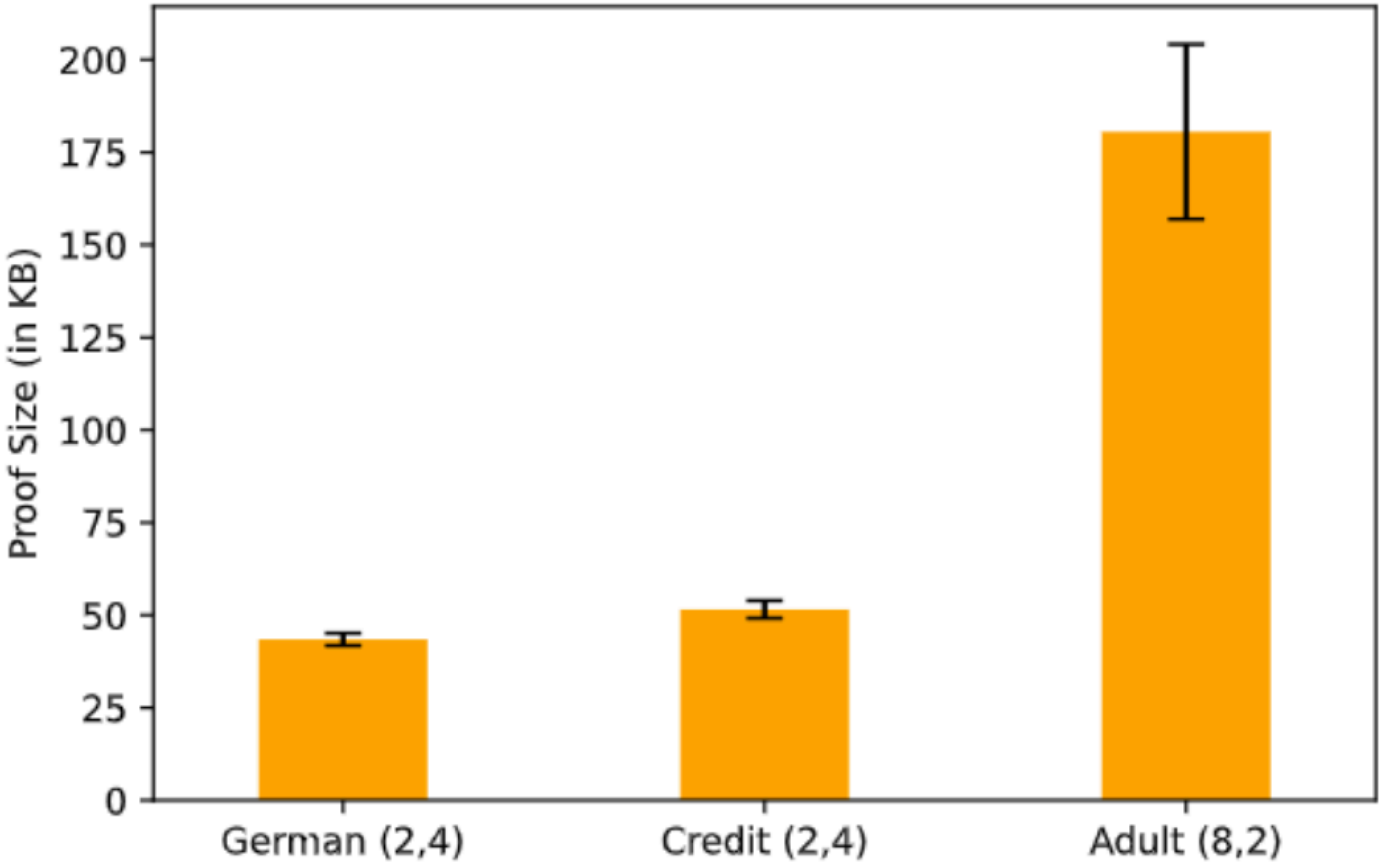
Q2. What is the computational overhead of *FairProof*?

Q2. What is the computational overhead of *FairProof*?



Proof Generation & Verification Time of *FairProof*. Averaged over 100 random samples.

Q2. What is the computational overhead of *FairProof*?



Proof sizes of proofs generated by *FairProof*. Averaged over 100 random samples.

Summary

- ZKPs might be a promising solution for auditing/verification requirements of ML
- We provide one example with fairness verification
- Future directions :
  - Scalability to bigger models using smart solutions
  - Different properties - where else can we use ZKPs?