Analyzing Fairness of Neural Network Prediction via Counterfactual Dataset Generation

Brian Hyeongseok Kim, Jacqueline L. Mitchell, and Chao Wang

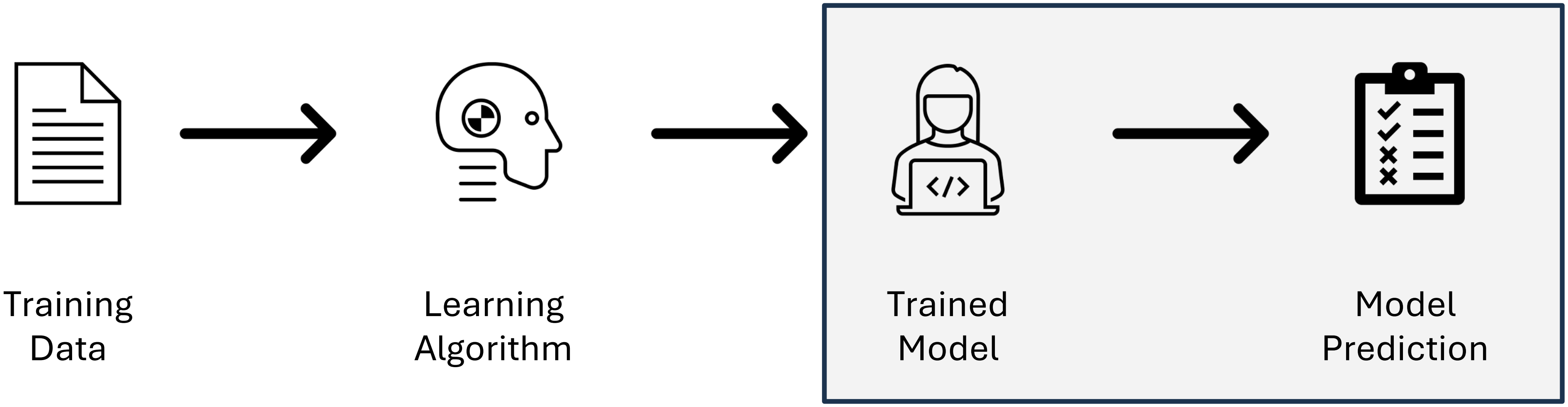
USC Viterbi School of Engineering

NLDL 2026: Tuesday, January 6, 2026

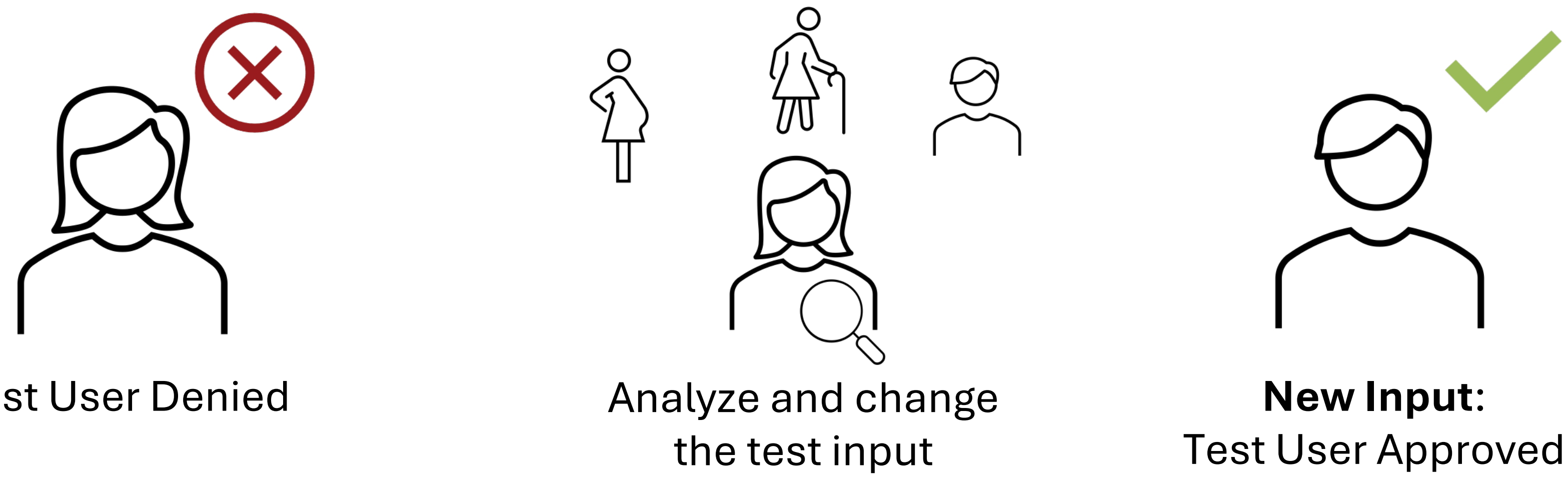
University of Southern California

Fairness in Machine Learning

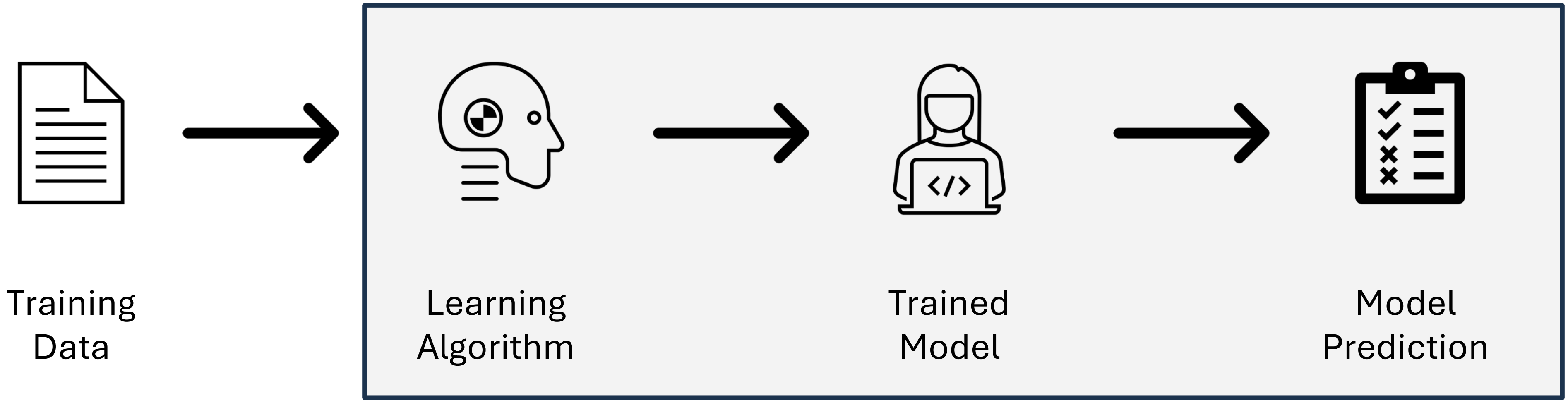
Counterfactual Explanation (Inference)



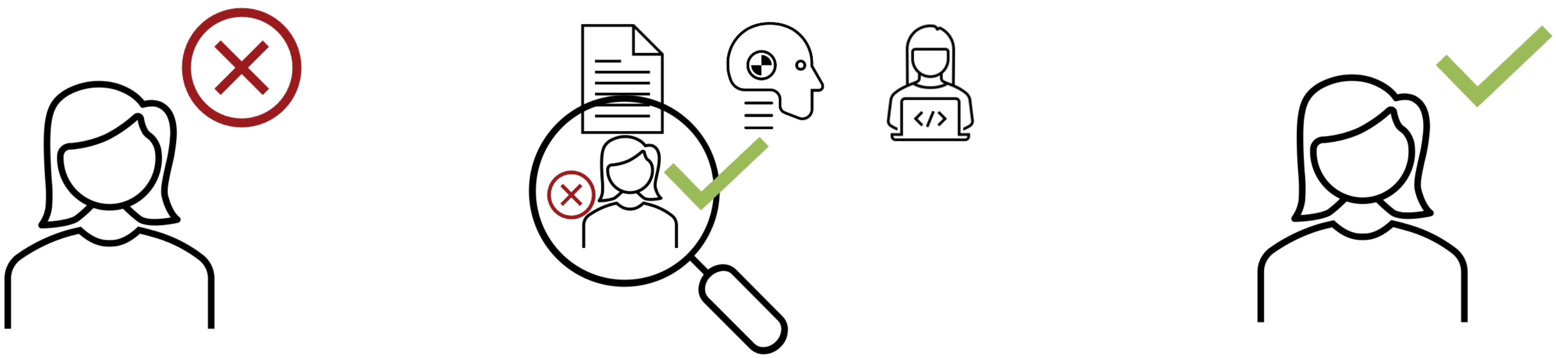
What if we change the **test inputs** *counterfactually*?



Counterfactual Explanation (Training + Inference)



What if we change the **training datasets** *counterfactually*?

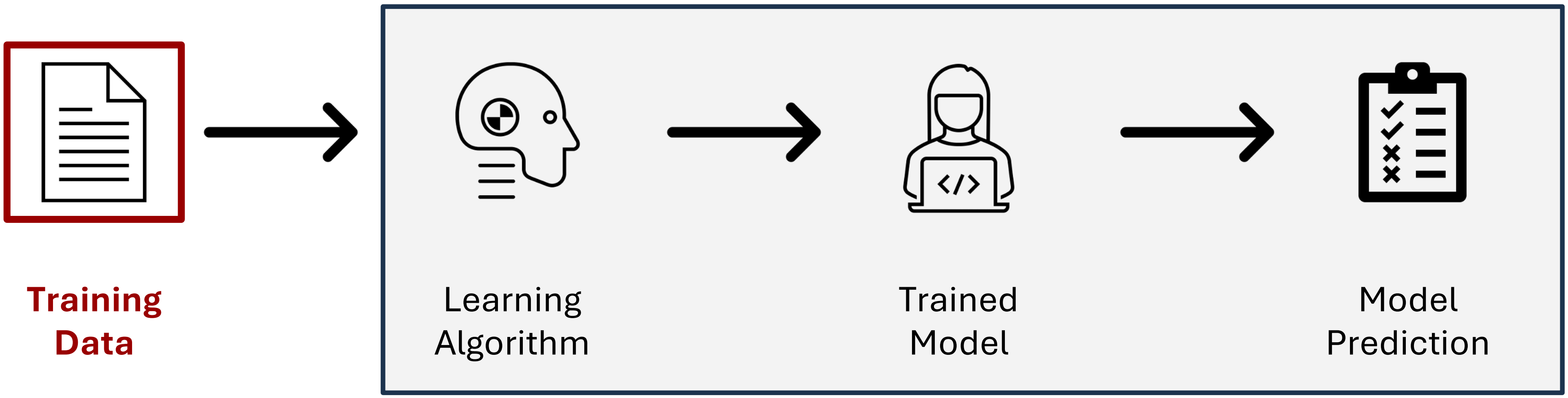


Test User Denied

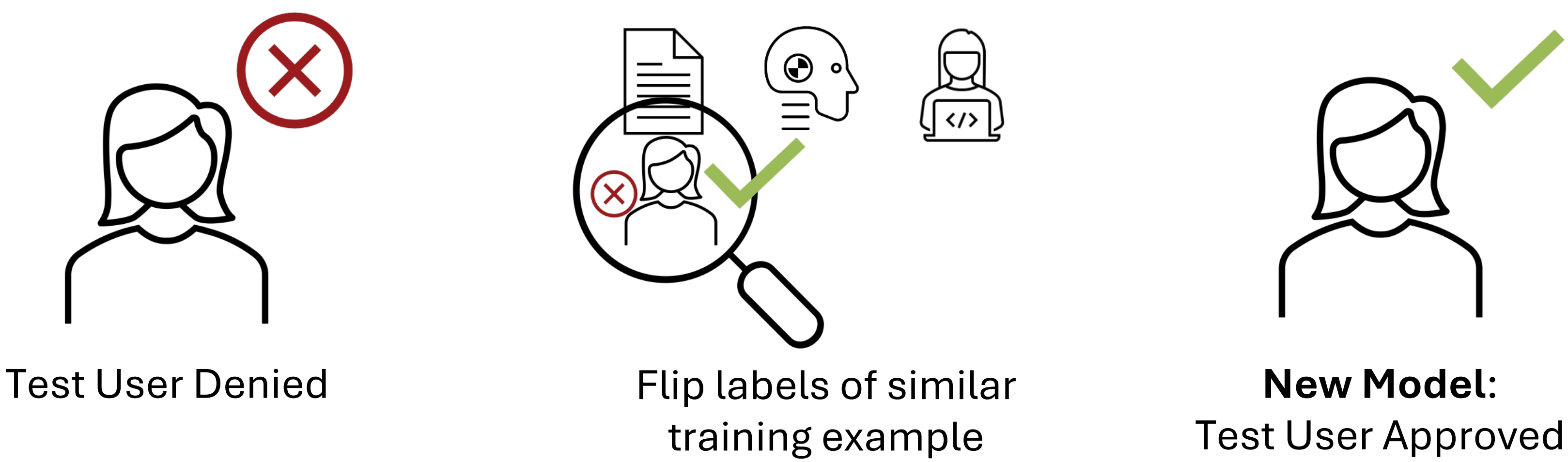
Flip labels of similar training example

New Model: Test User Approved

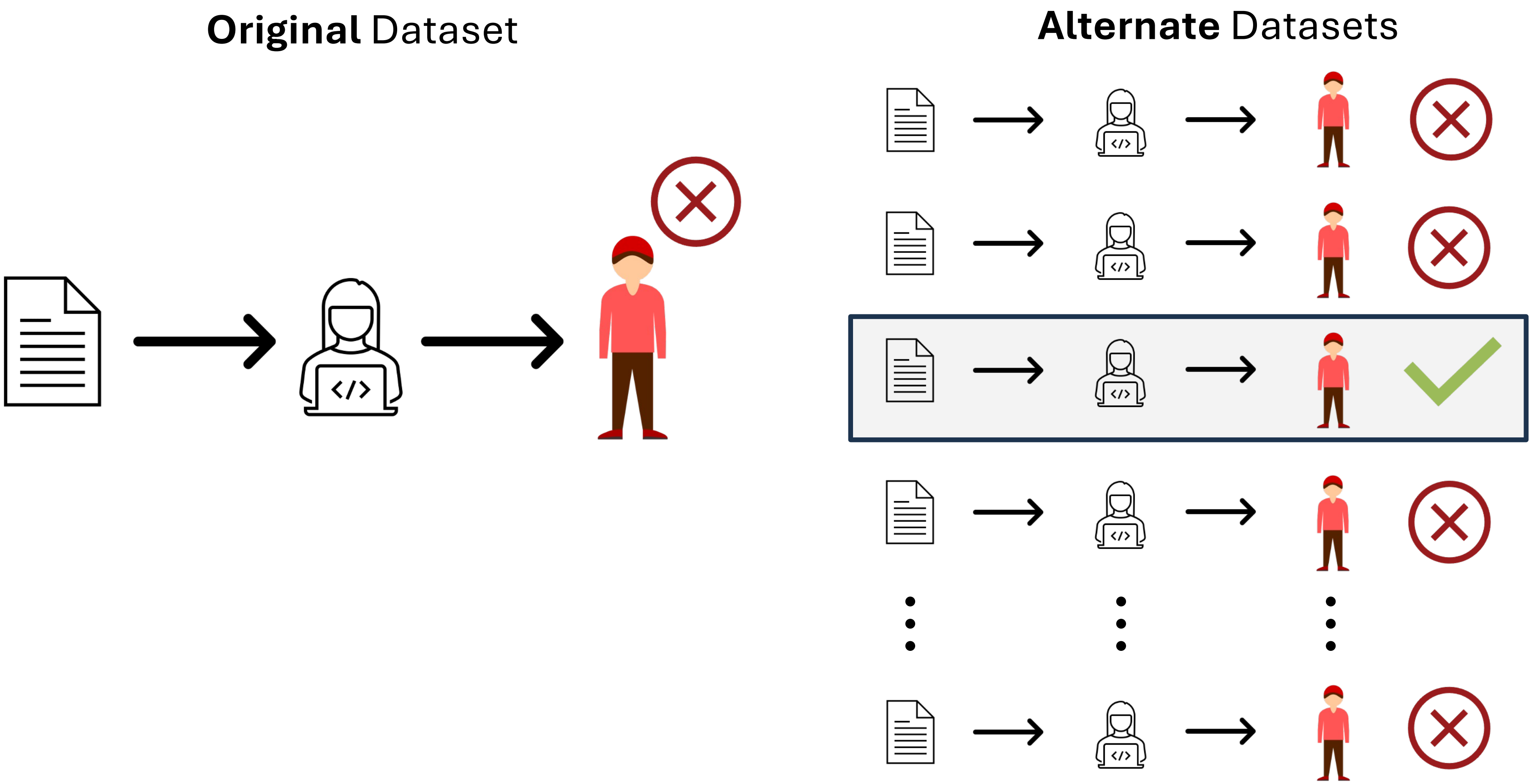
Counterfactual Dataset (CFD)



What if we change the **training datasets** *counterfactually*?



Counterfactual Dataset (CFD)



Counterfactual Dataset (CFD)

Naive enumeration?

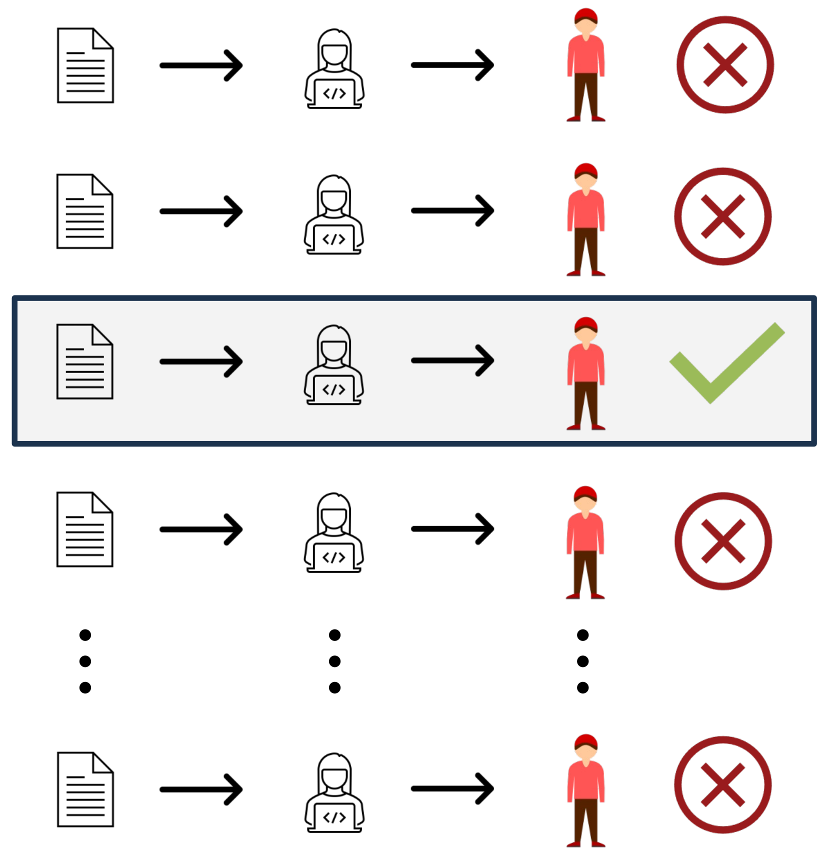
change up to *m out of n* training examples...

up to *(n choose m)* possible alternate datasets

e.g., n = 1000, m = 10
→ 2.63×10^{23}

Worse-than-exponential blow-up!

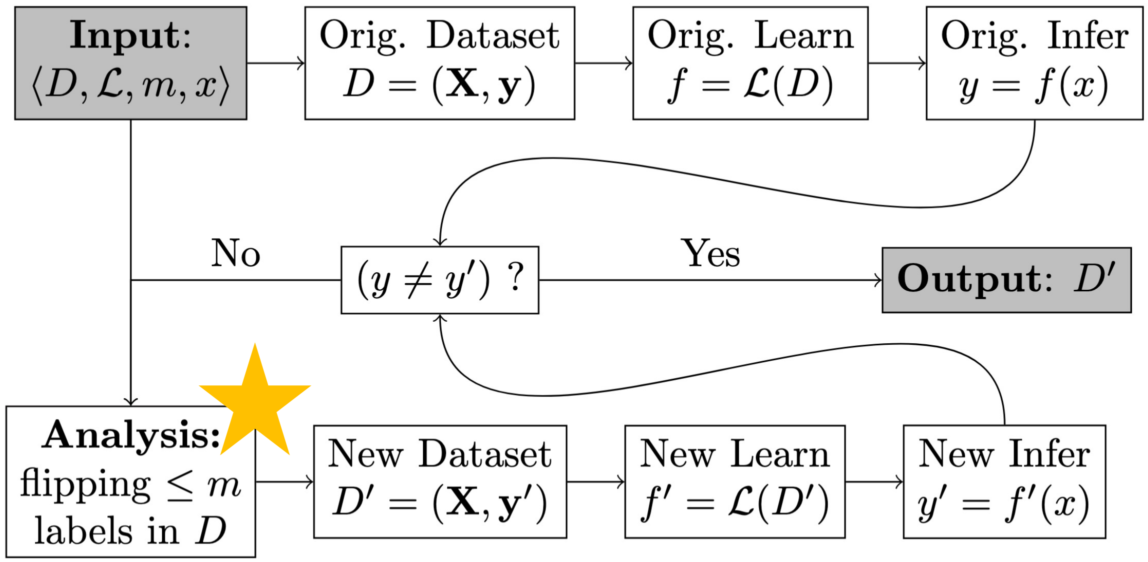
Alternate Datasets



Overview

$$\boxed{\begin{matrix}\textbf{Input}:\\ \langle D, \mathcal{L}, m, x\rangle\end{matrix}} \rightarrow \boxed{\begin{matrix}\text{Orig. Dataset}\\ D = (\mathbf{X}, \mathbf{y})\end{matrix}} \rightarrow \boxed{\begin{matrix}\text{Orig. Learn}\\ f = \mathcal{L}(D)\end{matrix}} \rightarrow \boxed{\begin{matrix}\text{Orig. Infer}\\ y = f(x)\end{matrix}}$$

Overview



★ **Our analysis:** Rank the impact of training examples in *D* for the given *x*

Methodology

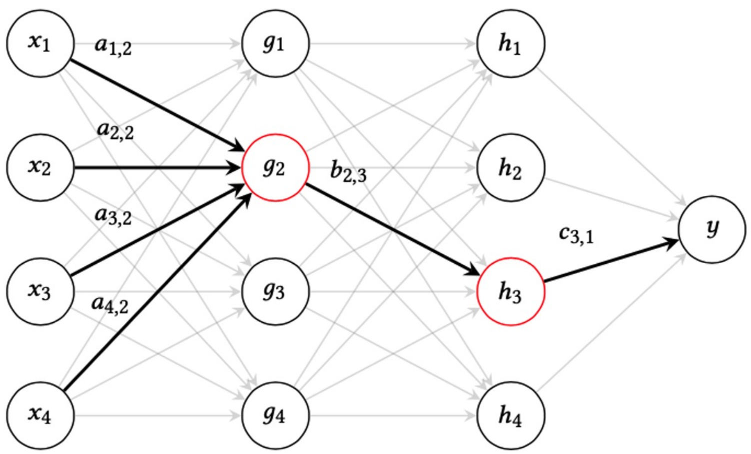
Training Stage

Linear Regression Surrogate

Inference Stage

Neuron Activation Similarity

Methodology: Linear Regression



red = active

with ReLU activation function:

$$f(\mathbf{x}) = \begin{cases} \theta_1^\top \mathbf{x}, & \mathbf{x} \in D_1 \\ \vdots & \\ \theta_p^\top \mathbf{x}, & \mathbf{x} \in D_p \end{cases}$$

[1] Meyer et al., "The Dataset Multiplicity Problem: How Unreliable Data Impacts Predictions". FAccT 2023.

Bypass training via a **closed-form solution**:

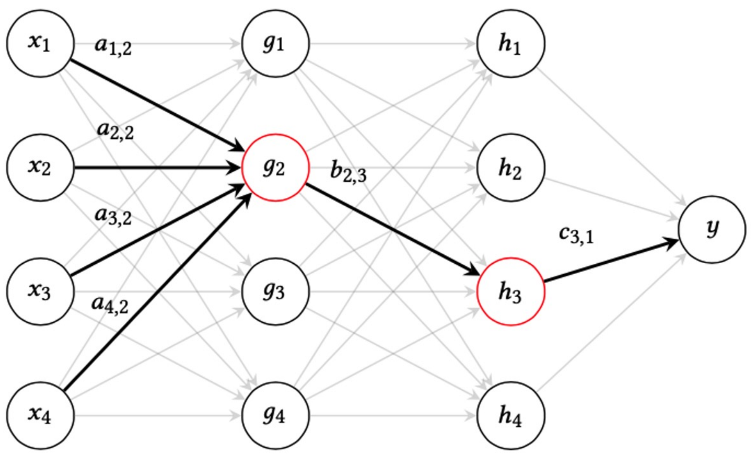
$$\theta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Prediction using closed-form solution:

$$y = \theta^\top \mathbf{x} = \mathbf{x}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{z}\mathbf{y} = \sum_{i=1}^{n} z_i y_i$$

n training examples are sorted in **decreasing order** of their z_i *score*

Methodology: Neuron Activation



red = active

$$\mathrm{Sim}(\mathbf{x}, \mathbf{x}') = 1 - \frac{1}{d}\sum_{k=1}^{d} \mathbf{1}\{b_k(\mathbf{x}) \neq b_k(\mathbf{x}')\}$$

test input **x** and training input **x'**

$b_k(\cdot) \in \{0, 1\} \rightarrow$ 0 for inactive, 1 for active

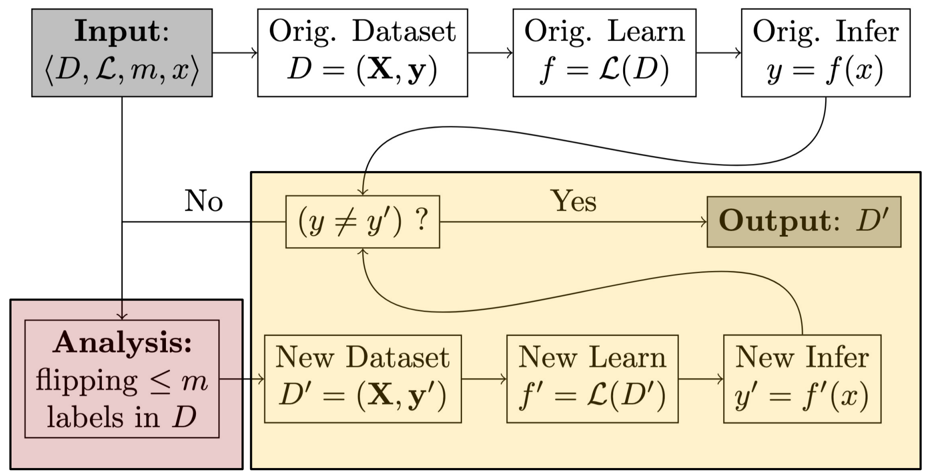
How **similar** are the test input and the training example in terms of activation?

n training examples are sorted in **increasing order** of their *Sim score*

Methodology: Overall

Algorithm 1 Generating counterfactual dataset.

Input: dataset $D = (\mathbf{X}, \mathbf{y})$, learning algorithm $\mathcal{L}$, bias budget m, test input $\mathbf{x}$, filtering rules ϕ and ψ
Output: counterfactual dataset $D' = (\mathbf{X}, \mathbf{y}')$
$f \leftarrow \mathcal{L}(D)$; $y \leftarrow f(\mathbf{x})$ {original model & prediction}
if not $\phi(\mathbf{x})$ **then**
    return {$\mathbf{x}$ does not pass filtering by *phi*}
end if
$\mathbf{y}_L \leftarrow$ LR_SCORING$(\mathbf{X}, \mathbf{y}, \mathbf{x}, m)$ {Section 4.3}
$\mathbf{y}_A \leftarrow$ ACTIV_SCORING$(\mathbf{X}, \mathbf{x}, f)$ {Section 4.4}
$[y_1, \ldots, y_{n_\psi}] \leftarrow$ COMBINE_SCORING$(\mathbf{y}_L, \mathbf{y}_A, \psi)$
    {n_ψ = size of training set after filtering by ψ}
$k \leftarrow 1$
while $k \leq m$ **do**
    $\mathbf{y}' \leftarrow$ label set where $y_1, \ldots, y_k$ in $\mathbf{y}$ are flipped
    $D' \leftarrow (\mathbf{X}, \mathbf{y}')$
    $f' \leftarrow \mathcal{L}(D')$; $y' \leftarrow f'(\mathbf{x})$ {new model & prediction}
    if $y \neq y'$ **then**
        return D' {CFD solution found}
    else
        $k \leftarrow k + 1$ {flip more labels in next iteration}
    end if
end while
return {solution not found}



Results: Experimental Setup

Datasets

- 7 popular fairness benchmarks
  Salary, Student, German, Compas, Default, Bank, Adult

Network training

- PyTorch using Adam optimizer
- ReLU networks with 2x4 to 2x32 hidden neurons

Comparing baselines

- *Random sampling* and *L_2 distance*
- *Influence functions* [1-5] are not considered due to their limited scalability

[1] Deng et al., "dattri: a library for efficient data attribution". NeurIPS 2024.
[2] Koh and Liang. "Understanding black-box predictions via influence functions". ICML 2017.
[3] Martens. "Deep learning via Hessian-free optimization". ICML 2010.
[4] Agarwal et al., "Second-order stochastic optimization for machine learning in linear time". JMLR 2017.
[5] Schioppa et al., "Scaling Up Influence Functions". AAAI 2022

Results: Research Questions

RQ1: Is it effective?

RQ2: Is it efficient?

RQ3: Is it meaningful?

RQ4: Is it robust?

Results: RQ1

RQ1: Is it effective?

Dataset	#	Our Method	Random Sampling	L_2 Distance
Salary	10	**3/3***	1	3
Student	121	**20/24***	13	10
German	182	**38/38***	17	15
Compas	200	**27**	10	5
Default	200	**18**	5	8
Bank	200	**24**	9	11
Adult	200	**44**	15	15
Total	1,113	**174**	70	67

Table 1. Number of CFDs found by each method.
** indicates number of ground truth CFDs via exhaustive enumeration.*

Results: RQ2

RQ2: Is it efficient?

Dataset	Our Method	Random Sampling	L_2 Distance
Salary	0.04	0.00	0.01
Student	0.07	0.02	0.03
German	0.08	0.02	0.04
Compas	0.28	0.12	0.24
Default	2.91	1.52	4.32
Bank	3.22	1.72	5.04
Adult	5.73	3.26	9.35

Table 2. Average non-training overhead per test input (in seconds).

Dataset	Our Method	Random Sampling	L_2 Distance
Salary	0.35	**0.31**	0.32
Student	**4.19**	4.35	4.52
German	**2.33**	2.56	2.62
Compas	**14.83**	17.03	17.18
Default	133.82	134.34	**122.55**
Bank	**123.53**	143.85	142.38
Adult	**195.81**	206.90	206.32

Table 3. Average runtime including retraining per test input (in seconds).

Results: RQ3

RQ3: Is it meaningful?

(a) Salary Test Case 4

(b) Salary Test Case 6

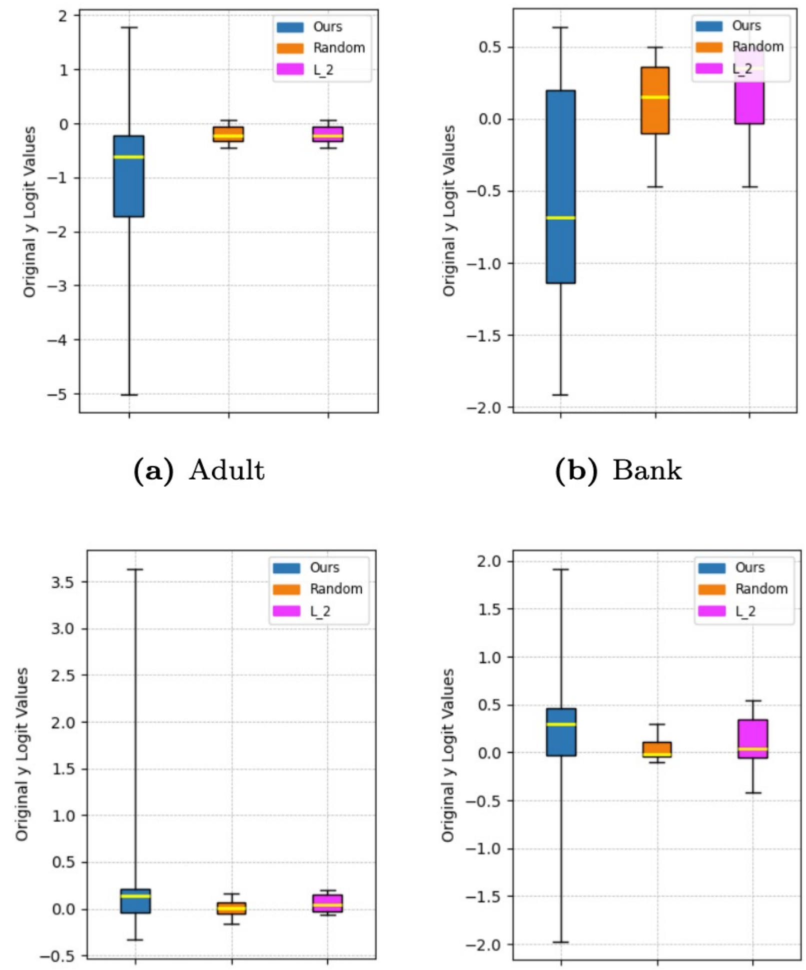
(c) Adult Test Case 36

(d) Compas Test Case 2

Training examples identified by each method against the test case

Results: RQ4

RQ4: Is it robust?



(a) Adult (b) Bank

(c) Compas (d) Default

Test cases around the decision boundary

Conclusion

1. Novel method to analyze neural network fairness using **counterfactual datasets**

2. Two heuristics to measure training example impact:
   linear regression surrogate (training) and **neuron activation similarity** (inference)

3. Evaluate on diverse fairness datasets: **effective, efficient, meaningful, and robust**

Thank You!
Any Questions?

Project Links

Open Review .net

Brian Hyeongseok Kim
brian.hs.kim@usc.edu