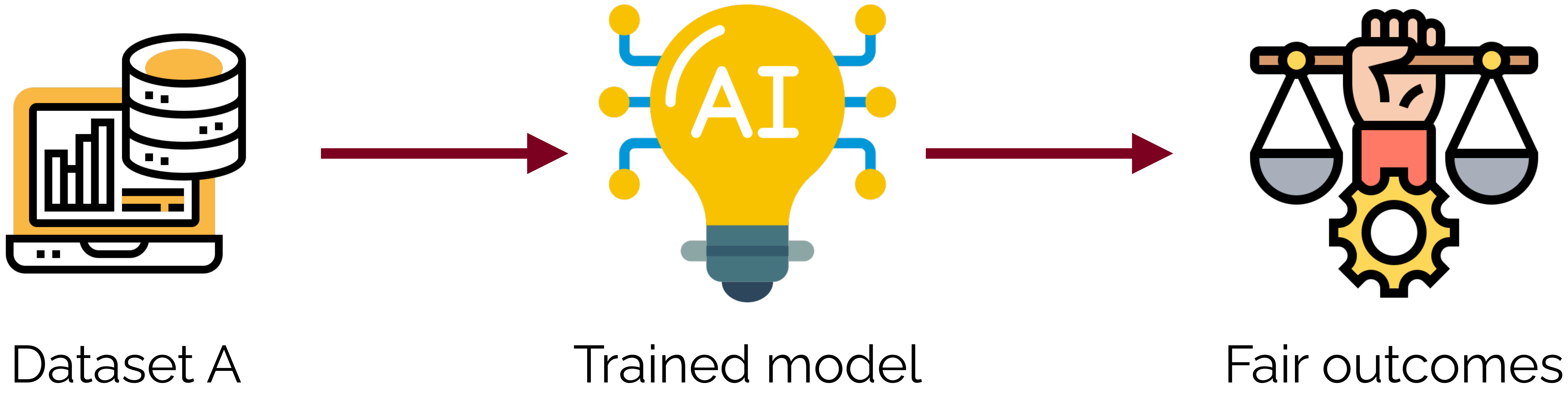
Sustaining Fairness via Incremental Learning

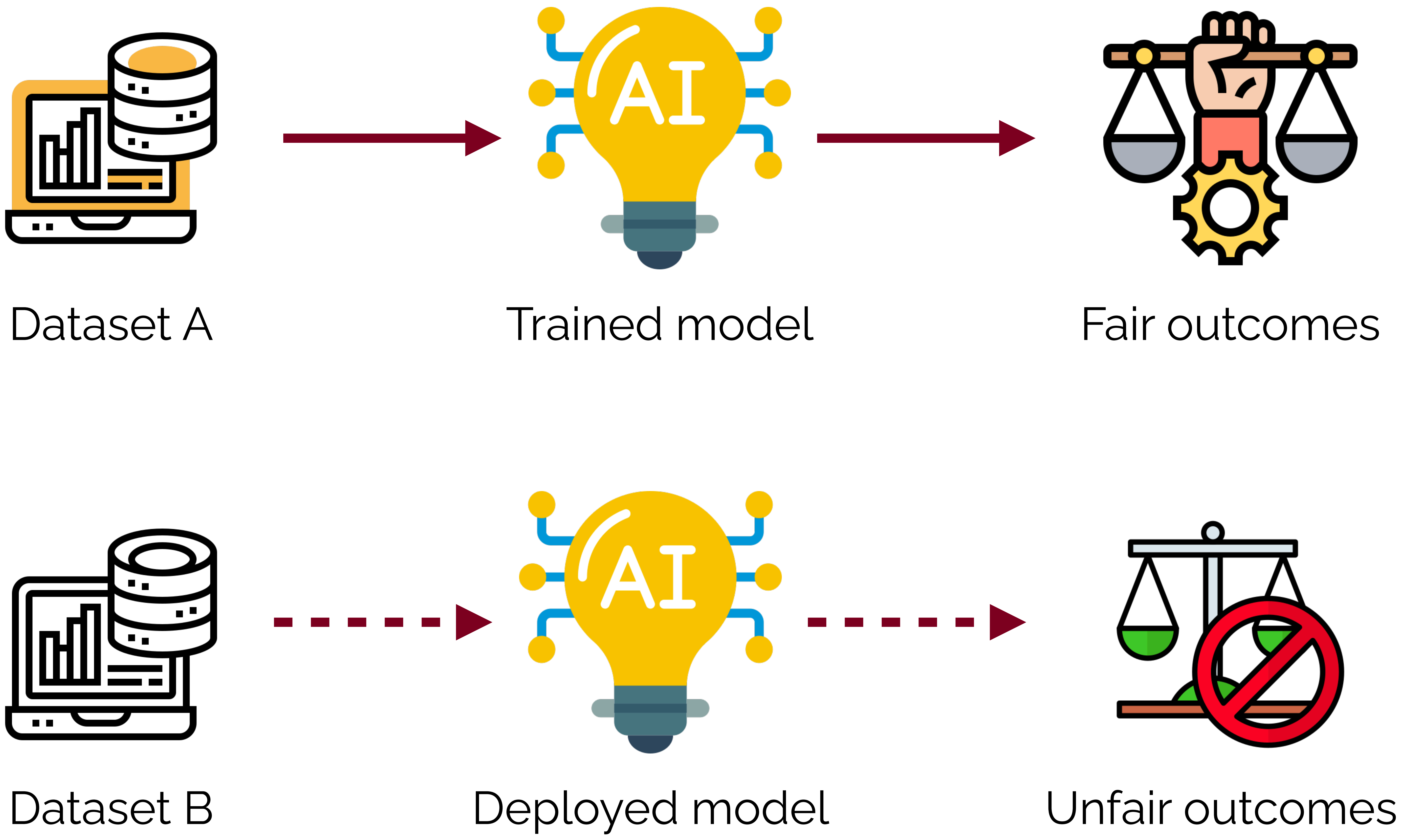
Somnath Basu Roy Chowdhury and Snigdha Chaturvedi
AAAI Conference on Artificial Intelligence, 2023
UNC Chapel Hill

Motivation

Motivation



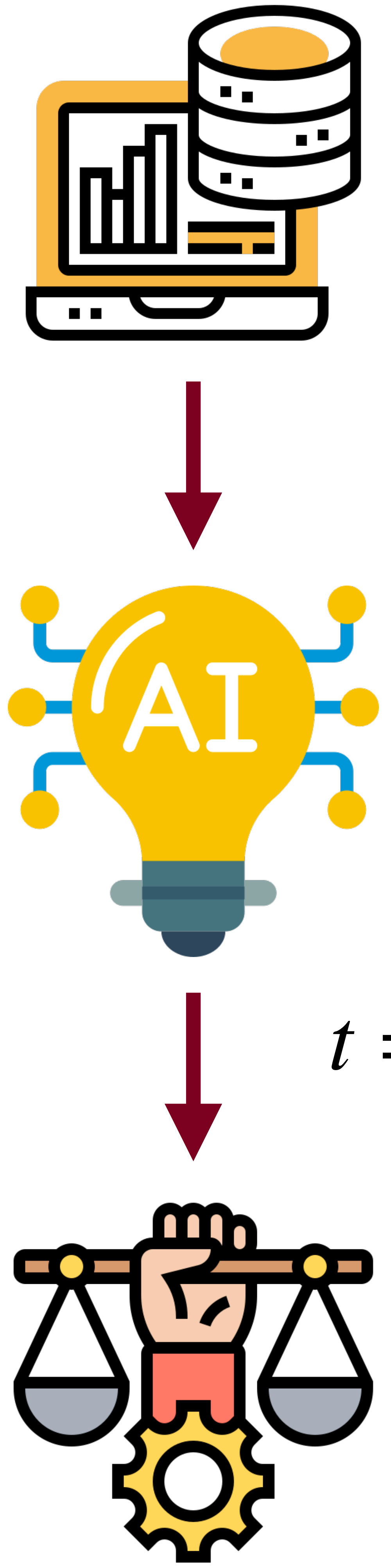
Motivation


Dataset A
Trained model
Fair outcomes
Dataset B
Deployed model
Unfair outcomes

Problem Statement

Setup

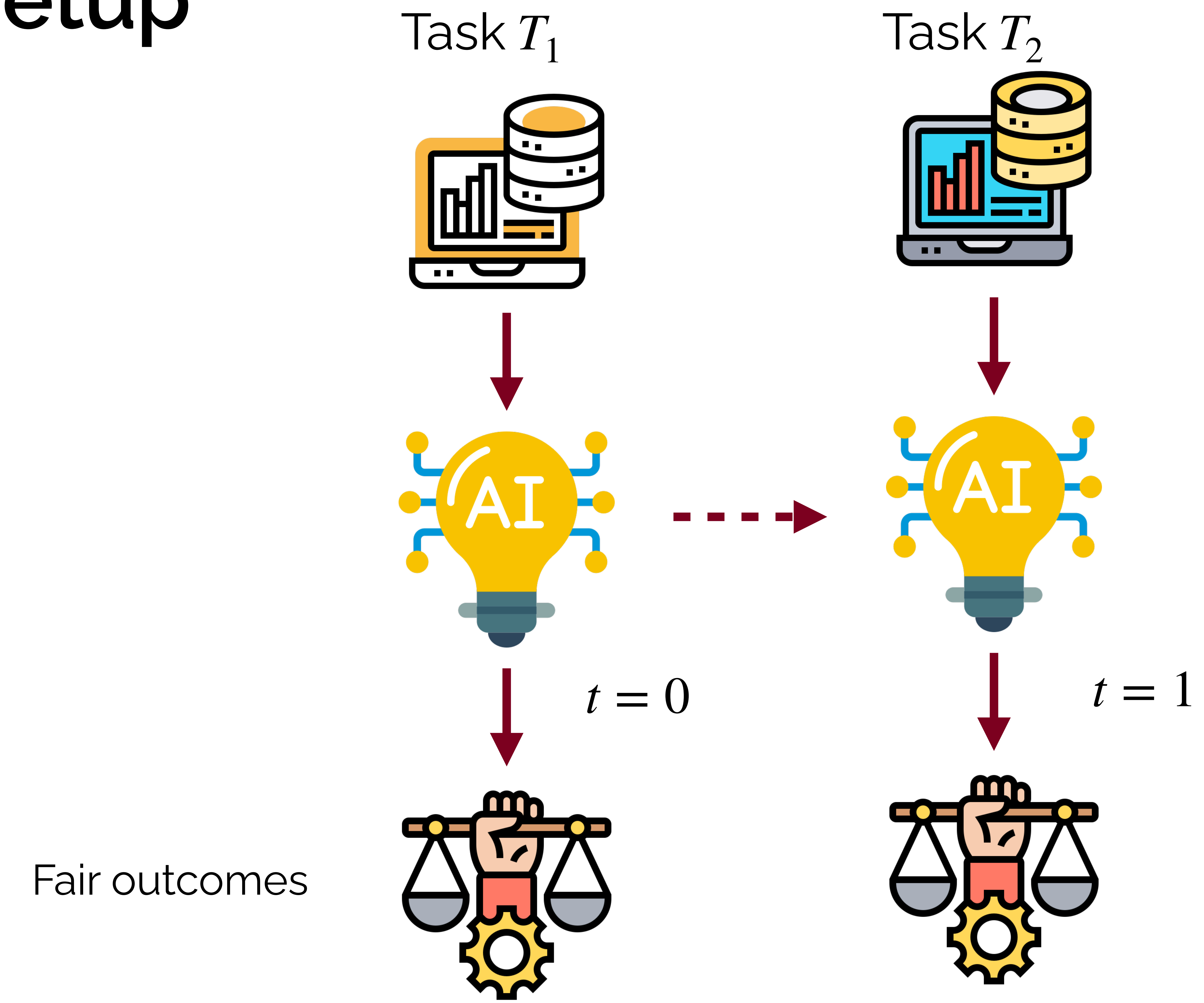
Task T_1



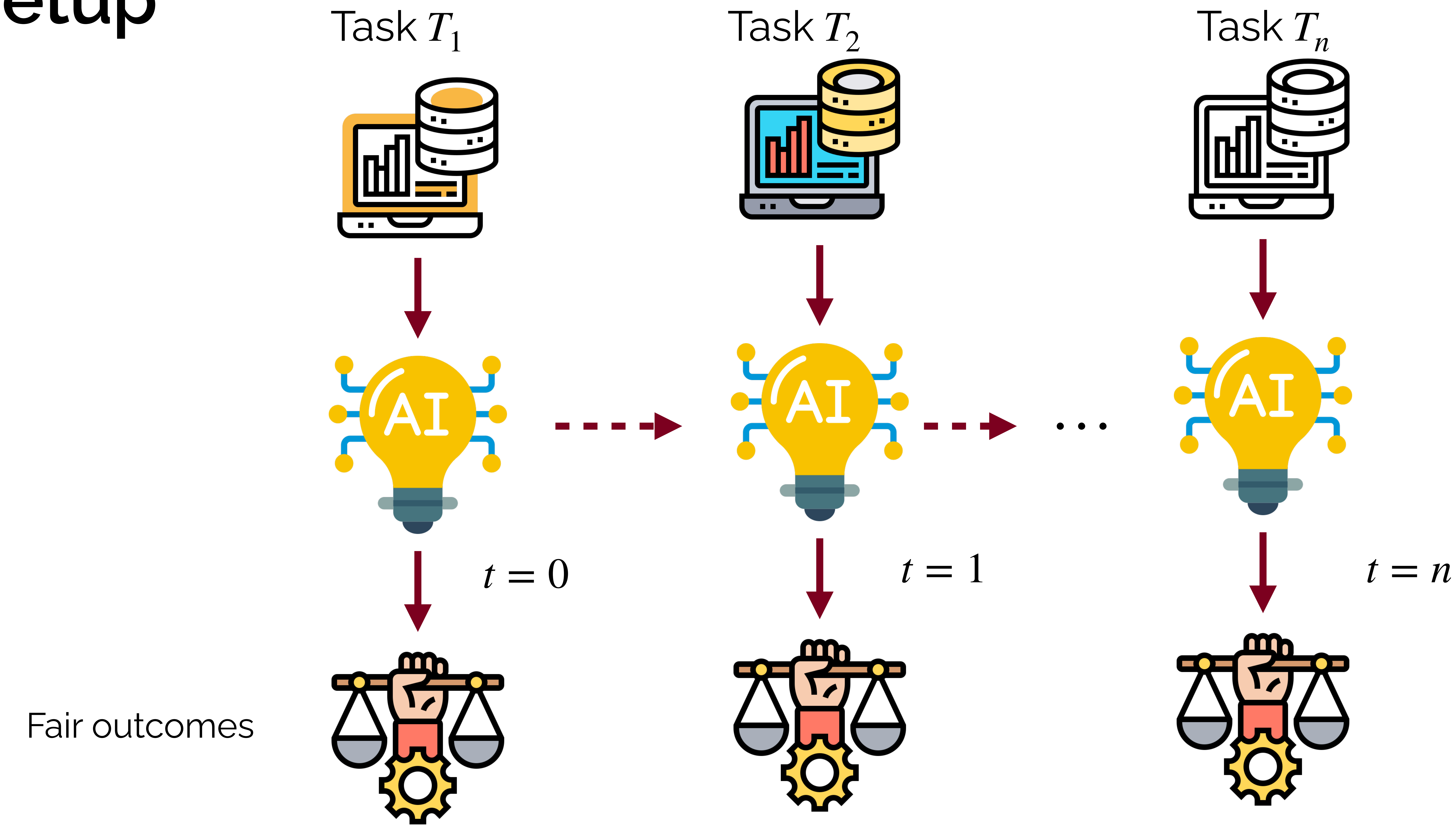
$t = 0$

Fair outcomes

Setup



Setup



Background

Rate Distortion

- Rate-distortion measures the total number of binary bits required to encode a set of representations $Z \in \mathbb{R}^d$

$$R(Z, \epsilon) = \frac{1}{2} \log_2 \det \left(I + \frac{d}{n\epsilon^2} ZZ^T \right)$$

Rate Distortion

- To measure volume of subgroups (categories of an attribute, e.g. male/female), we use a partition function $\Pi : Z \rightarrow \{Z_1, \ldots, Z_k\}$

$$R_c(Z, \epsilon \mid \Pi) = R(Z_1, \epsilon) + \ldots + R(Z_k, \epsilon)$$

Maximal Coding Rate

- A representation learning objective for classification tasks
- Given representations $Z = Z_1 \cup \ldots \cup Z_k$ from k different classes
- The following objective learns discriminative subspaces for each class

$$\max_{\theta} \Delta R(Z, \Pi) = R(Z, \epsilon) - R_c(Z, \epsilon \,|\, \Pi)$$

Fairness-aware Incremental Representation Learning (FaIRL)

Debiasing Framework

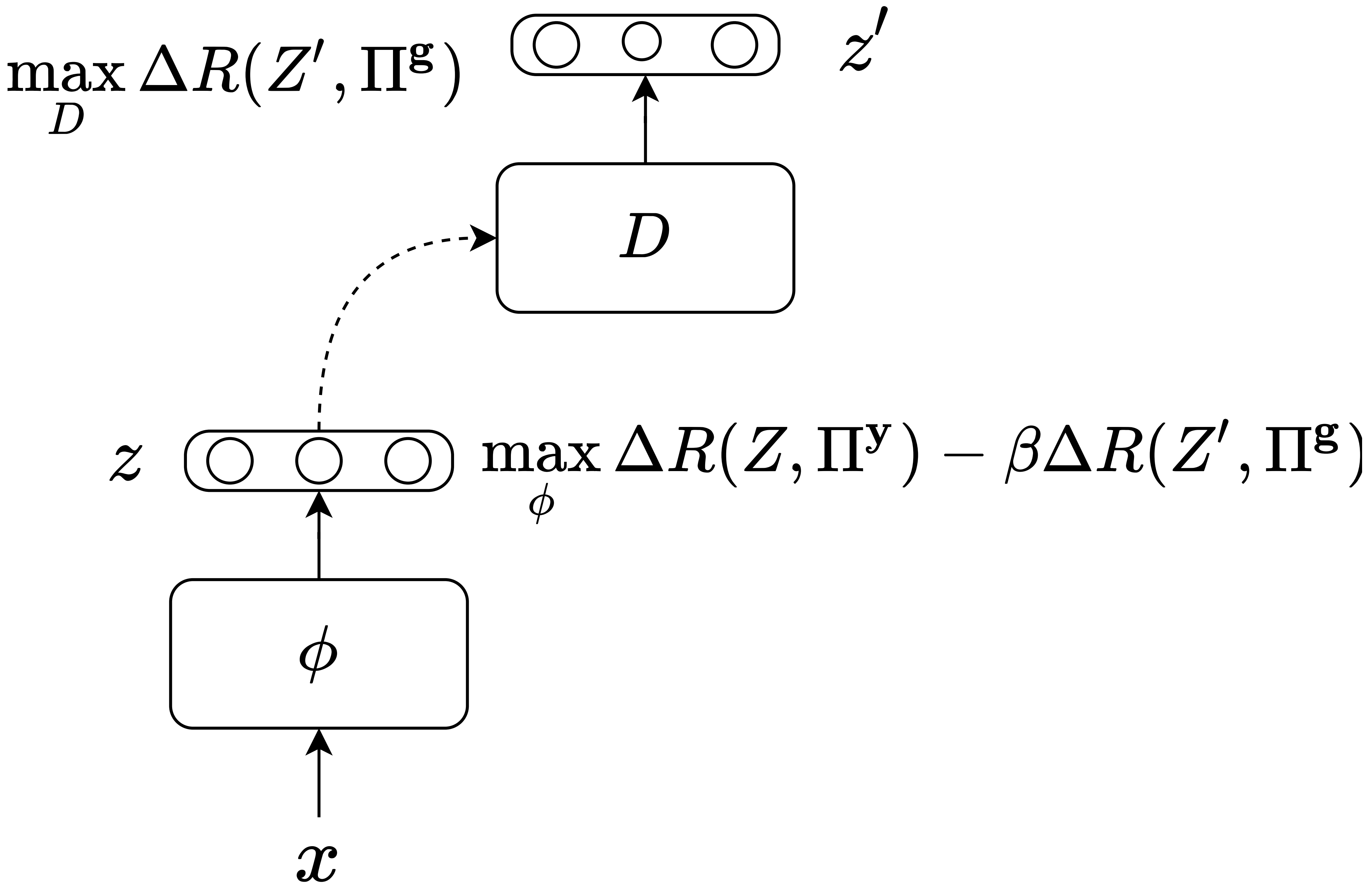
- First, we describe a framework to perform debiasing in a static setup

Debiasing Framework

- First, we describe a framework to perform debiasing in a static setup
- Input x, representations z, protected attribute g, target attribute y

Debiasing Framework

- First, we describe a framework to perform debiasing in a static setup
- Input x, representations z, protected attribute g, target attribute y



Incremental Learning

- Feature encoder learns compact representations due to discriminator loss

Incremental Learning

- Feature encoder learns compact representations due to discriminator loss
- Extend the debiasing framework for incremental learning

Incremental Learning

- Feature encoder learns compact representations due to discriminator loss
- Extend the debiasing framework for incremental learning
- We perform an *exemplar-based approach* — retaining samples from prior stages

Incremental Learning

- Feature encoder learns compact representations due to discriminator loss
- Extend the debiasing framework for incremental learning
- We perform an *exemplar-based approach* — retaining samples from prior stages
- At a stage t, the *discriminator* and *feature encoder* use a modified objective

Incremental Learning

- At stage t, old $Z_{old} = \phi(X_{old})$ and new $Z_{new} = \phi(X_{new})$ representations
- Discriminator optimizes the following: $\max_{D} \Delta R(Z'_{new}, \Pi^{g}_{new})$
- The feature encoder the following objective:

$$\max_{\phi} \Delta R(Z_{new}, \Pi^{y}_{new}) - \beta \Delta R(Z'_{new}, \Pi^{g}_{new}) - \gamma \Delta R(Z_{old}, \bar{Z}_{old}) - \eta \Delta R(Z_{old}, \Pi^{g}_{old})$$

Incremental Learning

- At stage t, old $Z_{old} = \phi(X_{old})$ and new $Z_{new} = \phi(X_{new})$ representations
- Discriminator optimizes the following: $\max_{D} \Delta R(Z'_{new}, \Pi^g_{new})$
- The feature encoder the following objective:

$$\max_{\phi} \Delta R(Z_{new}, \Pi^y_{new}) - \beta \Delta R(Z'_{new}, \Pi^g_{new}) - \gamma \Delta R(Z_{old}, \bar{Z}_{old}) - \eta \Delta R(Z_{old}, \Pi^g_{old})$$

Discriminative representations for X_{new}

Incremental Learning

- At stage t, old $Z_{old} = \phi(X_{old})$ and new $Z_{new} = \phi(X_{new})$ representations
- Discriminator optimizes the following: $\max_{D} \Delta R(Z'_{new}, \Pi^{g}_{new})$
- The feature encoder the following objective:

$$\max_{\phi} \Delta R(Z_{new}, \Pi^{y}_{new}) - \beta \Delta R(Z'_{new}, \Pi^{g}_{new}) - \gamma \Delta R(Z_{old}, \bar{Z}_{old}) - \eta \Delta R(Z_{old}, \Pi^{g}_{old})$$

Protect leakage for X_{new}

Incremental Learning

- At stage t, old $Z_{old} = \phi(X_{old})$ and new $Z_{new} = \phi(X_{new})$ representations
- Discriminator optimizes the following: $\max_{D} \Delta R(Z'_{new}, \Pi^{g}_{new})$
- The feature encoder the following objective:

$$\max_{\phi} \Delta R(Z_{new}, \Pi^{y}_{new}) - \beta \Delta R(Z'_{new}, \Pi^{g}_{new}) - \gamma \Delta R(Z_{old}, \bar{Z}_{old}) - \eta \Delta R(Z_{old}, \Pi^{g}_{old})$$

Retain subspaces for X_{old}

Incremental Learning

- At stage t, old $Z_{old} = \phi(X_{old})$ and new $Z_{new} = \phi(X_{new})$ representations
- Discriminator optimizes the following: $\max_{D} \Delta R(Z'_{new}, \Pi^{g}_{new})$
- The feature encoder the following objective:

$$\max_{\phi} \Delta R(Z_{new}, \Pi^{y}_{new}) - \beta \Delta R(Z'_{new}, \Pi^{g}_{new}) - \gamma \Delta R(Z_{old}, \bar{Z}_{old}) - \eta \Delta R(Z_{old}, \Pi^{g}_{old})$$

Protect leakage for X_{old}

Exemplar Sampling

After each stage, we retain a small sample of instances using the following:

- Random sampling — randomly select r samples

Exemplar Sampling

After each stage, we retain a small sample of instances using the following:

- Random sampling — randomly select r samples
- Prototype sampling — select instances with high similarity with top eigenvectors

Exemplar Sampling

After each stage, we retain a small sample of instances using the following:

- Random sampling — randomly select r samples
- Prototype sampling — select instances with high similarity with top eigenvectors
- Submodular optimization — select instances best representative of a set w.r.t. a submodular function

Evaluation

Datasets

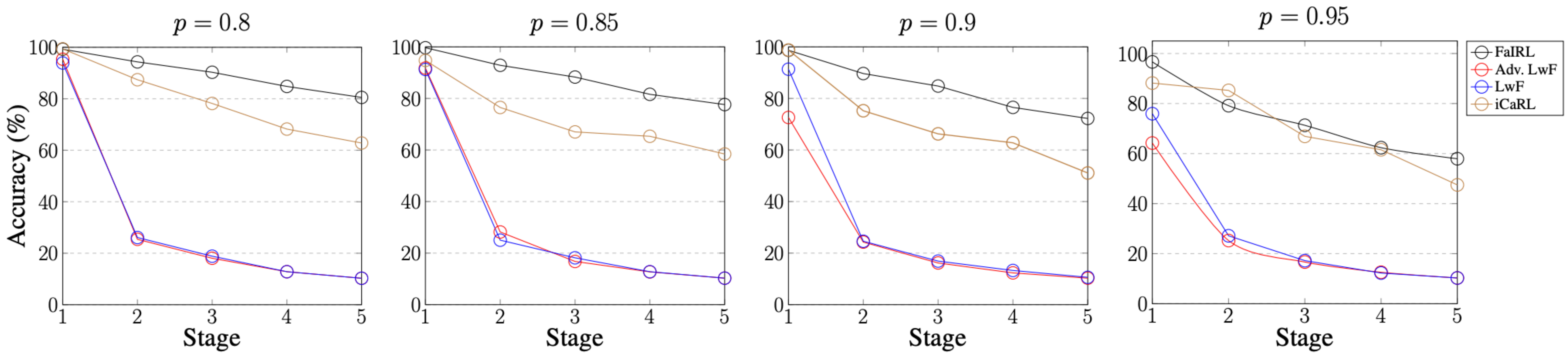
- Biased MNIST: We modify MNIST dataset to have background color (*protected variable*) correlate with digit information (*target variable*) with probability (p)

Datasets

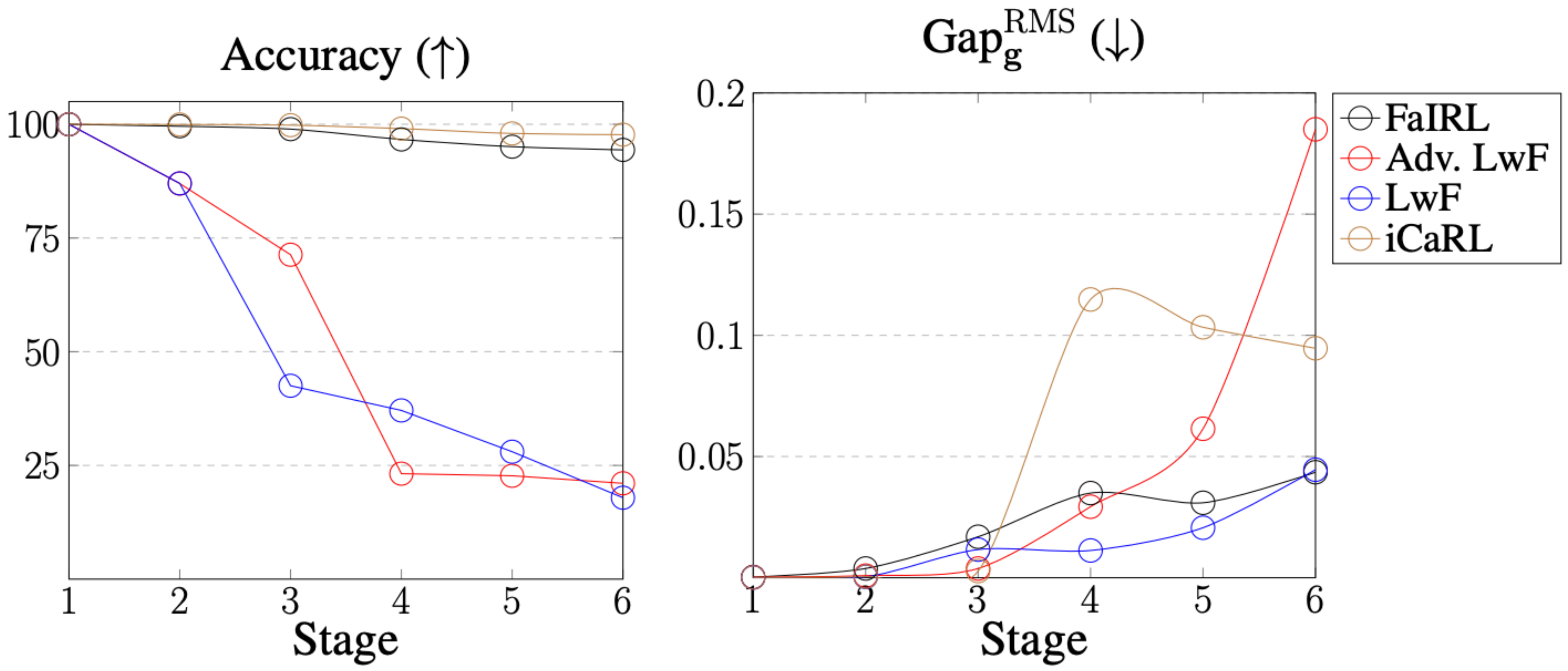
- Biased MNIST: We modify MNIST dataset to have background color (*protected variable*) correlate with digit information (*target variable*) with probability (p)

- Biographies contains biographies of people with a profession (*target variable*) and gender label (*protected variable*)

Biased MNIST

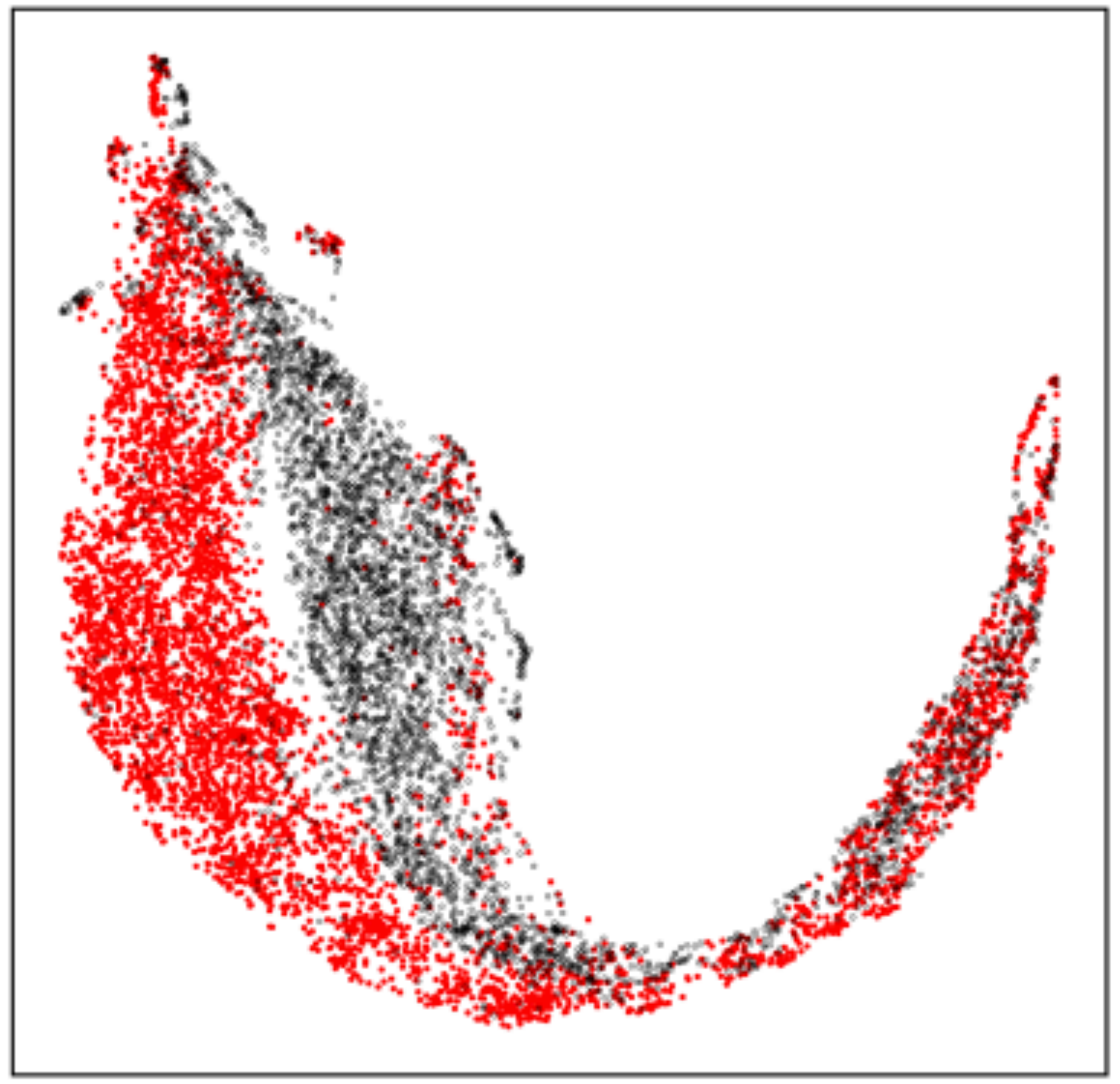


Biographies

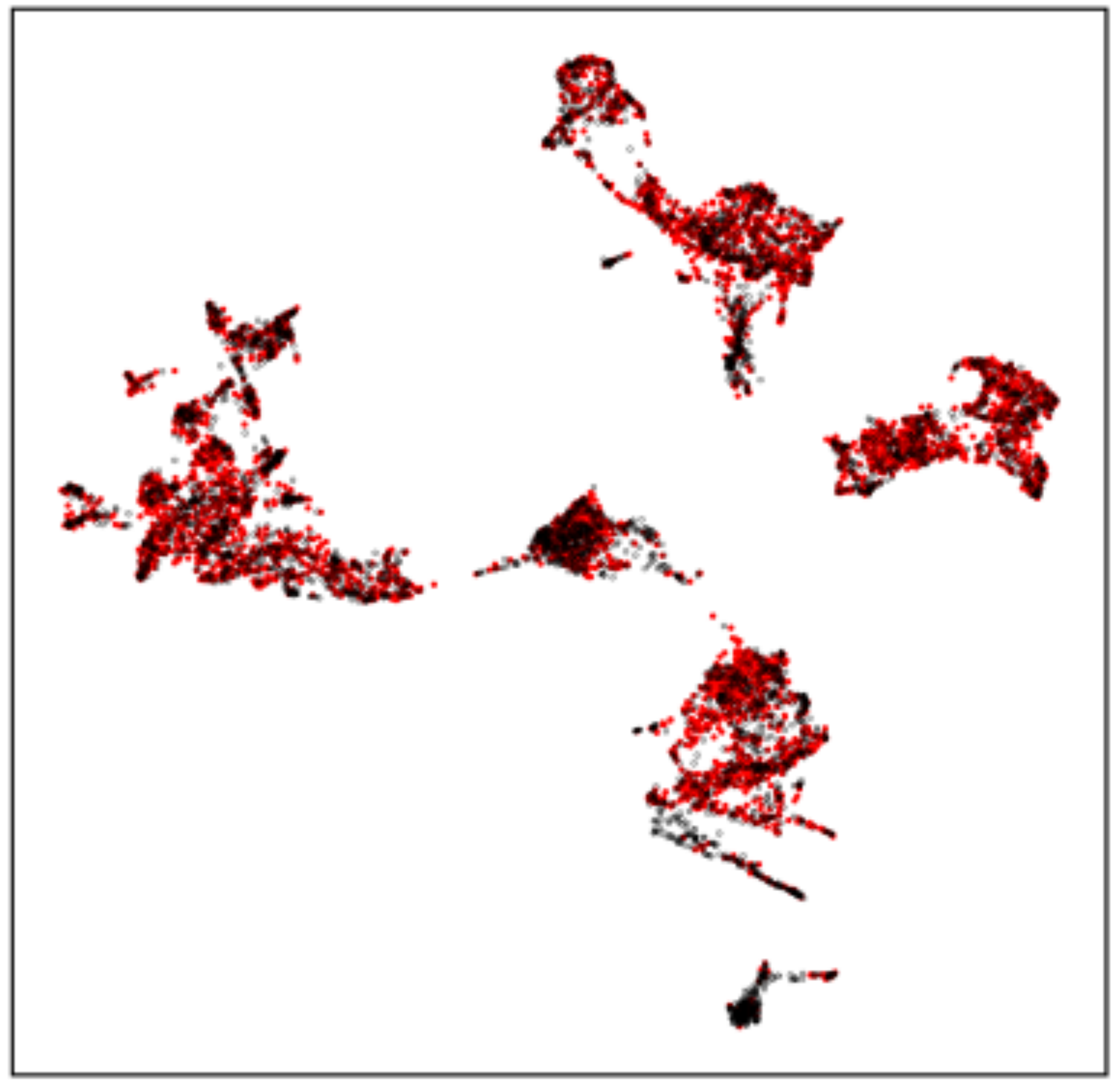


Visualization

UMAP projections of representations in Biographies dataset



(a) Before training



(b) Post training using FaIRL

Conclusion

- We tackle the task of learning fair representations in an incremental learning setup

Conclusion

- We tackle the task of learning fair representations in an incremental learning setup
- We propose FaIRL, that makes fair decisions while learning new tasks by controlling the rate-distortion function of representations

Conclusion

- We tackle the task of learning fair representations in an incremental learning setup
- We propose FaIRL, that makes fair decisions while learning new tasks by controlling the rate-distortion function of representations
- Empirical evaluation show that FaIRL outperforms existing methods

Conclusion

- We tackle the task of learning fair representations in an incremental learning setup
- We propose FaIRL, that makes fair decisions while learning new tasks by controlling the rate-distortion function of representations
- Empirical evaluation show that FaIRL outperforms existing methods
- FaIRL is a first step towards achieving fairness in the wild

Thank You!

Paper Link!

@SomnathBrc

brcsomnath/FaIRL

somnath@cs.unc.edu