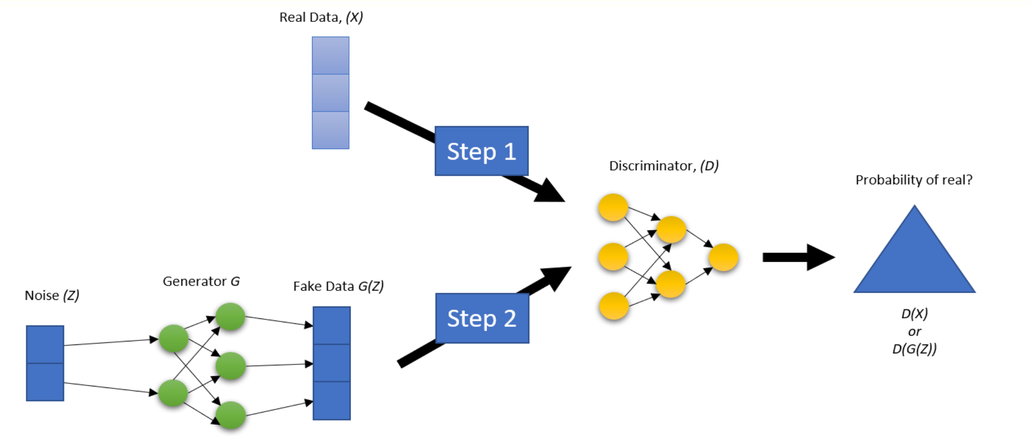
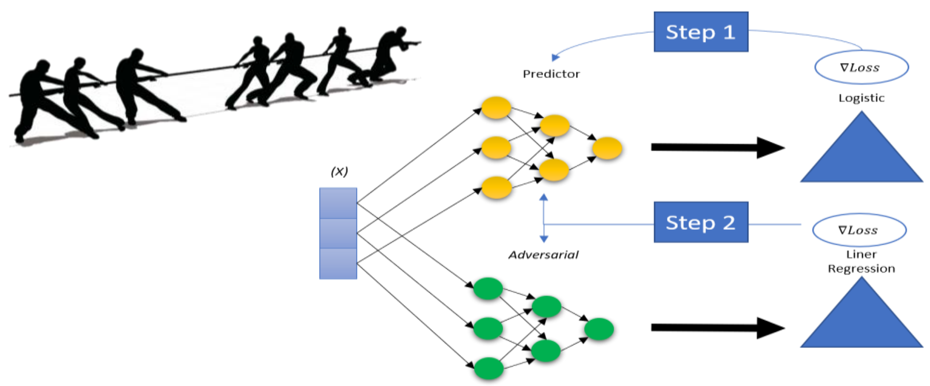
# Bias in Machine Learning: An Adversarial Approach

Solange Garcia de Alford, Steven Hayden, Nicole Wittlin
Master of Science in Data Science
Southern Methodist University, Dallas, TX 75275, USA

DataScience@SMU




## Introduction

- **Bias is very prevalent, occurring in ML models at pre-process, in-process, post-process stages.**
- **Examples of ML bias are widely known – COMPAS, Amazon hiring algorithm/resume scan, Word2Vec. Most are binary: protected class vs unprotected class.**
- **Our study focuses on eliminating bias stemming from AGE when predicting employee attrition.**

## Main Topics

- **Adversarial learning can be leveraged to mitigate bias and unfairness.**
- **Competing models of GAN, where Predictor (P) tries hinder Discriminator (D) with fake data, while feedback from D tries to hinder P prediction ability.**
- **Our study: P -- predict employee prediction; D -- predict age.**
- **Goal: improve group fairness via demographic parity (DP) (all equally likely of positive outcome** (TP + FP)**).**

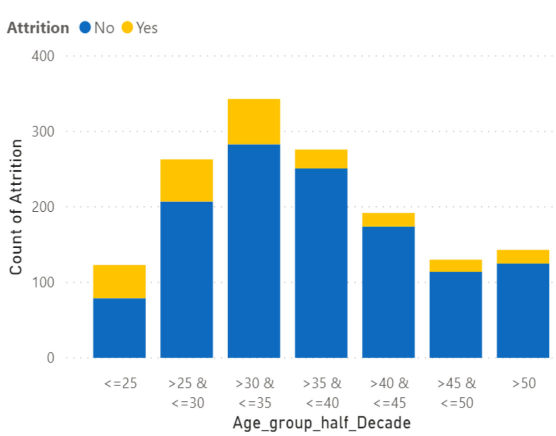
## Data Overview

**IBM Employee Attrition Dataset**

Attrition: 84% NO / 16% YES

Age binned in 5-year ranges

More attrition <= age 35




## Model Architecture

**Baseline Model**

- Logistic Model
- $\hat{y}_{Attrition} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$

**Adversarial Models**

- $\hat{y}_{Attrition} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$
- $\hat{A}_{Age} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_n x_n + \varepsilon$
- $Loss = -\alpha \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

## Results

- **Improved DP**: range Pre-GAN for all groups between 94-100%; Post-GAN range 98-100%.
- **Small trade-off between Accuracy and Fairness Post-GAN:** accuracy decreased 2% but DP increased 6%.
- **Demonstrate work beyond binary classes**: can work toward having more than one unprotected group.
- **See Results Chart**

| ACCUR | < = 25 | (25, 30] | (30, 35] | (35, 40] | (40, 45] | (45, 50] | > 50 | Overall |
|---|---|---|---|---|---|---|---|---|
| Pre GAN | 0.7575 | 0.84 | 0.8315 | 0.9138 | 0.9464 | 0.8696 | 0.7941 | 0.8587 |
| Post GAN | 0.6970 | 0.8 | 0.8315 | 0.8966 | 0.8929 | 0.9130 | 0.7941 | 0.8315 |

| DEMO PARITY | < = 25 | (25, 30] | (30, 35] | (35, 40] | (40, 45] | (45, 50] | > 50 | Overall |
|---|---|---|---|---|---|---|---|---|
| Pre GAN | 0.9394 | 0.96 | 0.9438 | 0.9655 | 0.9464 | 0.9565 | 1.0 | 0.9375 |
| Post GAN | 1.0 | 1.0 | 0.9888 | 0.9828 | 1.0 | 1.0 | 1.0 | 0.9972 |

## Conclusions / Future Work

- **Bias must be addressed in advance and throughout – NOT as an afterthought.**
- **Re-run study with larger, real dataset and/or pre-processed data that balances attrition % or sampled differently.**
- **Refine code with Early Stop, when the adversary has sufficiently mitigated bias and correlation is no longer detected in the adversarial model for Z(x), Age.**
- **We CANNOT and MUST NOT replace the inquisitiveness, skepticism, mortal imagination and compassion that humans bring to bear to on ML.**