# 6.867 Project: Milestone 1

(Sapna Kumari, Archer Wang, Joe O'Connor)

We aim to assess the adversarial debiasing technique described in this paper on the Compas Recidivism Risk Dataset. The paper presents the adversarial framework for learning fair classifiers which are argued to be nearly as accurate as their unfair counterparts. The idea is that if a trained model is biased with respect to certain sensitive attribute such as gender, race, age etc., then that model is retrained by incorporating a second adversarial model where the objective is to maximize the ability of first model to predict the outcome while minimizing the ability of second model to infer the sensitive attribute from the output of the first model. The paper analyzes the algorithm on two datasets, and show that their adversarial model succeeds in debiasing while maintaining similar accuracy as original models.

We are interested in reproducing the results of adversarial learning on a different dataset to evaluate the strengths and weaknesses of the model, and see how well they generalize to other datasets. For this task, we have chosen to run our analysis on Compas Recidivism Risk Dataset (here) which contains records of demographics, criminal history, jail and prison time for defendants from Broward County from 2013 and 2014. Our analysis will include first training a simple logistic/linear regression model to predict the recidivist (likelihood for convicted criminals to reoffend) risk scores and evaluate the model accuracy and bias with respect to different races (black vs caucasian), and then retrain the same logistic/linear model with adversarial learning and evaluate to what extent bias and accuracy are improved/hurt.