# Project Proposal

Jonathan Rittmo Sara Thiringer

2020-10-14

## Introduction

In this project we aim to evaluate methods that account for discrimination bias in predictive models. This can for example happen when a data set on which the model is based is unbalanced with regards to a certain group, i.e. observations are fewer. Bias can also be introduced in a model because the data reflect true discrimination bias in the population. Take an employee wage data set as an example. If a model, e.g., predicts equal wages between two groups when all variables available except for this specific grouping (say gender) have been taking into account but different wages when gender is introduced there is evidence of true bias.

How should this be handled? The problem of unbalanced data could be tackled by using synthetic data, i.e. parametric bootstrapping of the original data set to even out imbalance. But how do different modelling methods fare when they are being trained on such data?

The second problem is more difficult. In the example given above the easy way to deal with such a problem is just to remove the biased variable for model training. However, when the problematic variable is correlated with other predictors in the model the bias can persist even after removal.

## Aims

- Identify and evaluate methods for dealing with persisting discrimination bias.
  - Filter observations if not represented?
  - Manipulate target data?
- Evaluate models using synthetic data.
- Identify a method for quantifying fairness and compare accuracy vs. fairness in different models and methods.

## Example