Text, logo

Description automatically generatedLogo

Description automatically generated with medium confidence

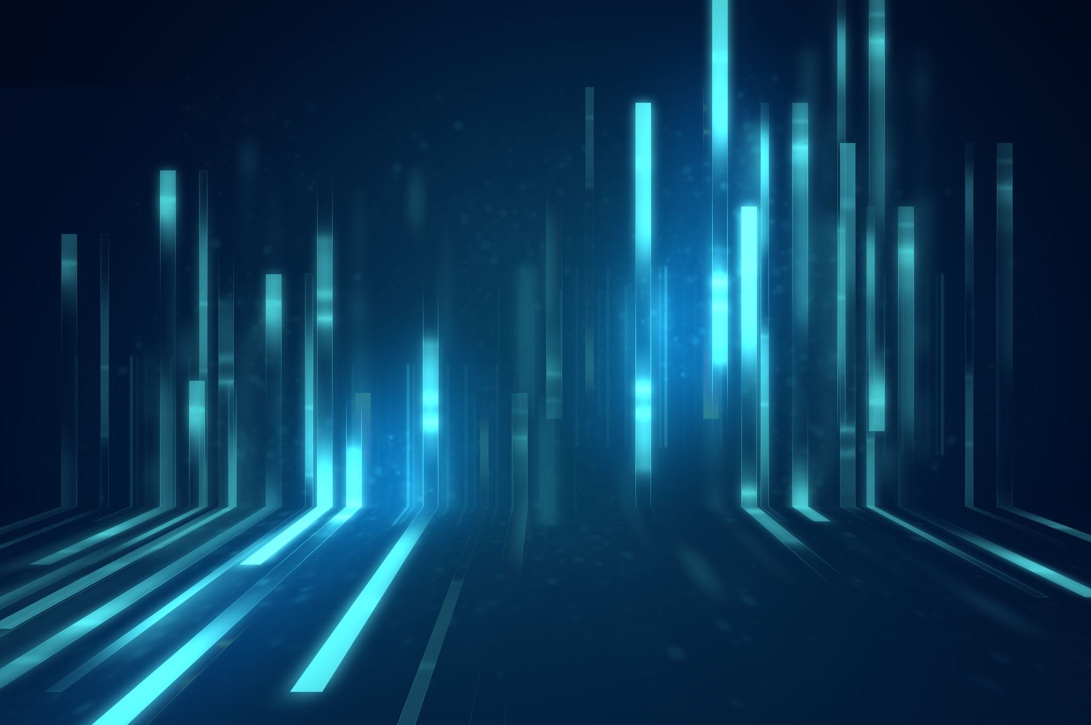
­­­

Fairness

Expert

Module 2

**Technical Interpretations of Fairness – ex 2**



# Introduction

This module will expand on the intermediate course on fairness. It will provide further tools for applying learnings on fairness/bias in the development of a technological solutions.

Here we include the second example scenario as an exercise to identify potential sources of bias in a practical setting.

­­­

# Scenario

**Student support programs**

A university attended by 30,000 students from a diverse range of backgrounds including overseas students. The university has a programme to offer additional academic support to students in their final year who are struggling to help ensure they graduate with the qualifications they are aiming for. In order to help identify which students receive this additional support, the university has asked your team to build a model that can predict which students are most at risk of not reaching their intended grade targets.

**Goal**

Improve graduation rates for students

**Data**

Student records from previous schools and training colleges both inside and outside the country, as well as performance records at university, as well as written reports from personal tutors

**Analysis**

Predict risk of not graduating

**Actions**

Assign support resources (offers of specialised tutoring, additional classes, etc)

# Questions - Sources of Bias

* What are some potential sources of bias in the underlying data?
* What are the risks to fairness in downstream applications and deployment of the model described?

# Questions – Bias Metrics

* How would you describe a **false positive** in this problem to a policymaker or business owner? What’s the potential harm/cost of one?
* How would you describe a **false negative** to a policymaker or business owner? What’s the potential harm/cost of one?
* What confusion matrix metric (e.g., FPR, FNR, TPR, FDR, etc.) would you choose to focus on in terms of equity for this case?

# Example answers

**What are some potential sources of bias in the data?**

* Records from different districts, countries etc might contain differing types of information
* Reasons for failing to complete or nature of failing to graduate (e.g., pursue GED)
* Reasons could be related to socio economic status
* Teachers/tutors might have biases for/against certain students (!!)
* Some students may have special needs which ought to be considered
* Assumption that failure to complete is because of academic problems
* Student records may not contain the important features for making predictions (related to above “assumption that failure to complete…”)
* Gender, race, demographic and income-related features in the data can create biases
* Is the training data representative of the demographics attending the university?

­­­

