Assessing the Predictive Performance of Seldonian Algorithms: A Simulation Study

Dasha Asienga & Professor Katharine Correia

The Department of Mathematics and Statistics, Amherst College

MOTIVATION

Using the standard machine learning (ML) approach for real-life applications can result in *algorithmic bias*: a situation where an algorithm's predictions systematically discriminate against a demographic group. Seldonian algorithms offer a way to address this problem by incorporating *probabilistic constraints* on undesirable behavior (mathematically defined) in the search for an optimal solution.

OBJECTIVES

The primary objective of the simulation study is to investigate the *efficacy and applicability of Seldonian algorithms* in practical classification settings along three key performance measures:

- ➔ *convergence (probability of a solution)*
- ➔ *fairer (less discriminatory) outcomes*
- ➔ *predictive accuracy*

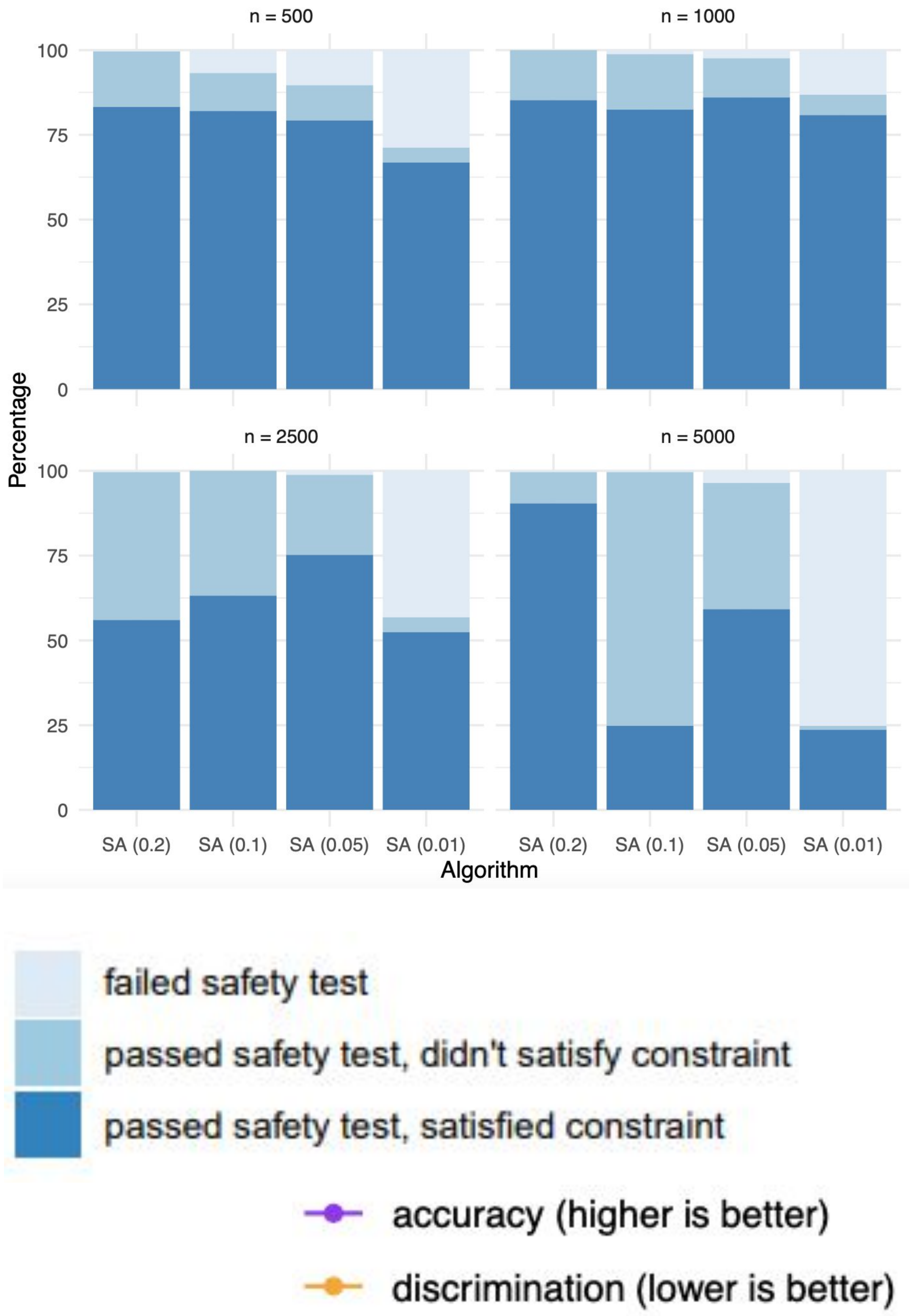
METHODS

Data Generation: this study is a *proof of concept*, so the data-generation mechanism follows a realistic design. The COMPAS recidivism data set (collected in Broward County, Florida, 2013-2014) is used nationwide to predict recidivism by defendants, but it exhibits racial discrepancies. The response variable in this study is modeled as a linear combination of the COMPAS variables such that the complex social relationships that may be expected in the real world are retained.

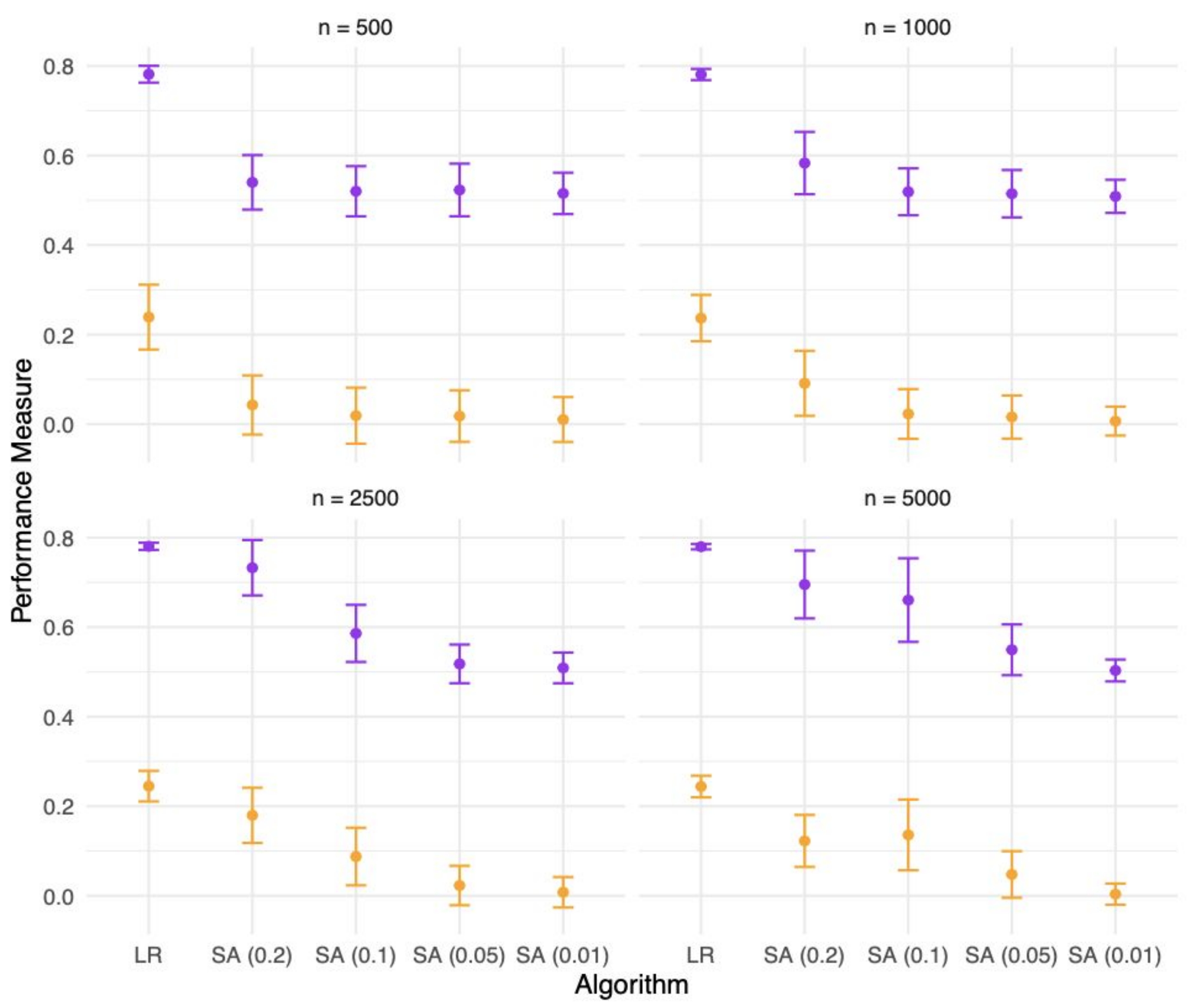
Methodology: 1 *logistic regression* and 4 *Seldonian algorithms* (with different fairness constraints ϵ as an upper bound of the total absolute difference in error rates between Black and White defendants: ϵ = *0.2, 0.1, 0.05, 0.01*) are fit on 250 simulated data sets of size *n = 500, 1000, 2500, 5000* [1000 total data sets] and results compared. δ = *0.05* to ensure 95% confidence that the Seldonian solution will satisfy the specified fairness constraint.

RESULTS

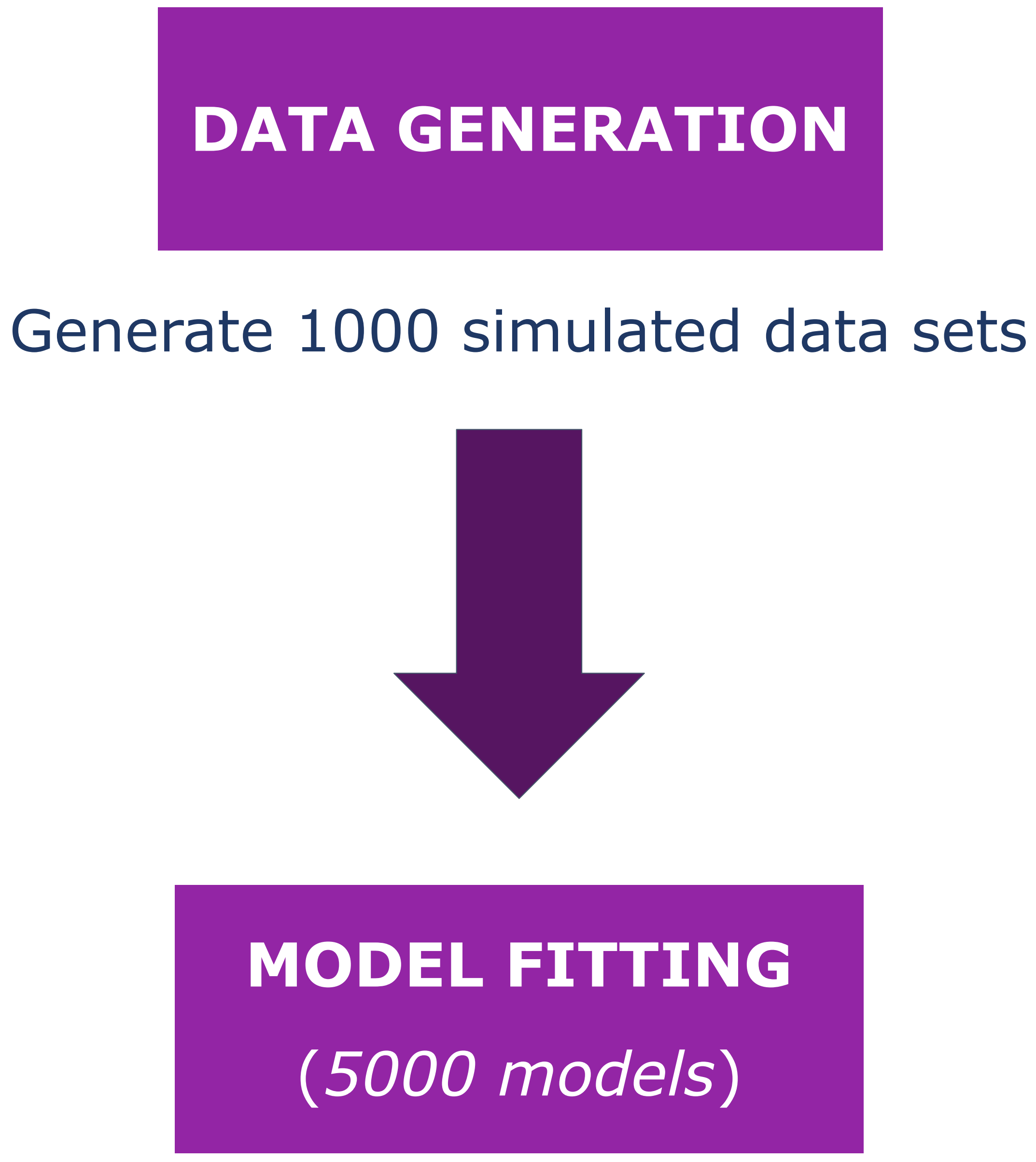
CONVERGENCE



ACCURACY-DISCRIMINATION



HPC CLUSTER USE



DATA GENERATION

Generate 1000 simulated data sets

MODEL FITTING

(*5000 models*)

In an *sbatch* file, loop through all 1000 data sets, running a *Python script* on each set and appending results to a global data frame as a row

```
for file in $files; do
  srun ./seldonian_sim.py "$folder$file" &
done
```

CONCLUSIONS

1. There is an *accuracy-fairness trade-off*.

2. Seldonian solutions are not guaranteed and *may still yield unfair results*, though it's less probable.