UNSUPERVISED BIAS DISCOVERY FOR FAIR AI

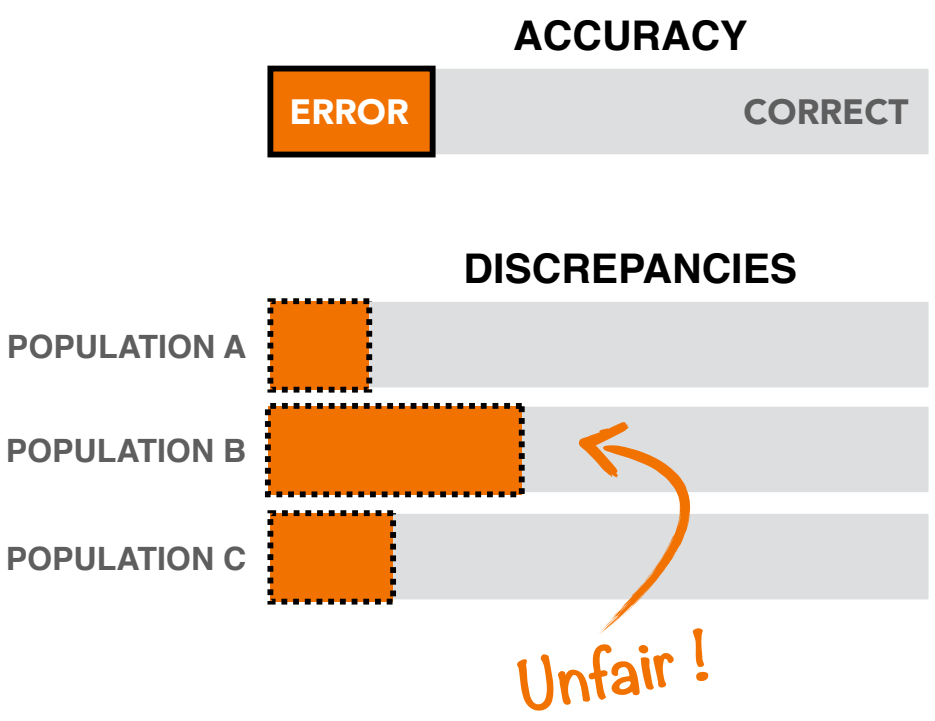
Emma Beauxis-Aussalet (e.m.a.l.beauxisaussalet@vu.nl) - Michael Cochez (m.cochez@vu.nl)

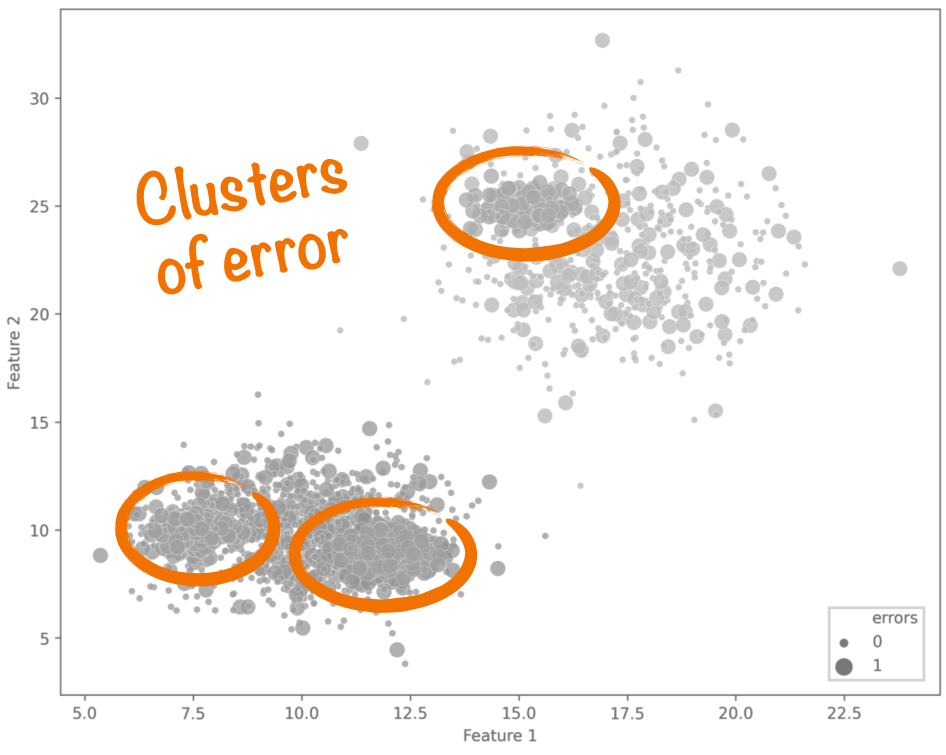
VU UNIVERSITY AMSTERDAM

AI errors can occur more often for some populations thus creating bias and discrimination (e.g., against a nationality, ethnicity, gender, age, or neighbourhood).

The discriminated populations can be **unexpected**, **unlabelled**, or **defined by several variables** (e.g., intersectionality, proxy variables [1]).

Unsupervised methods can **identify bias against unlabelled populations** by exploring the potential combinations of features that define **clusters of error**.

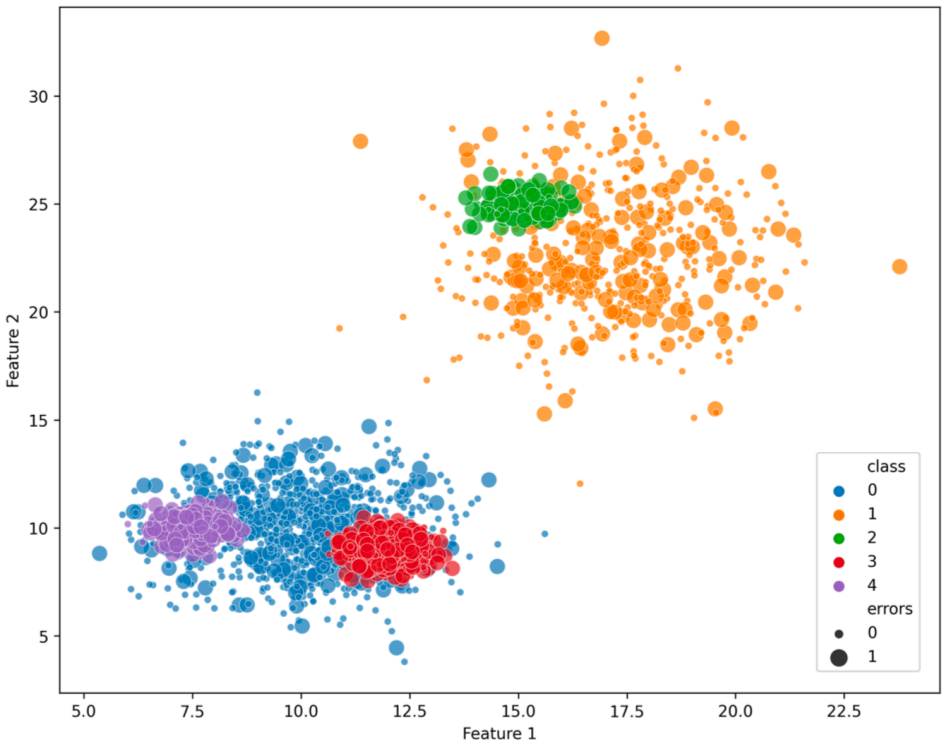




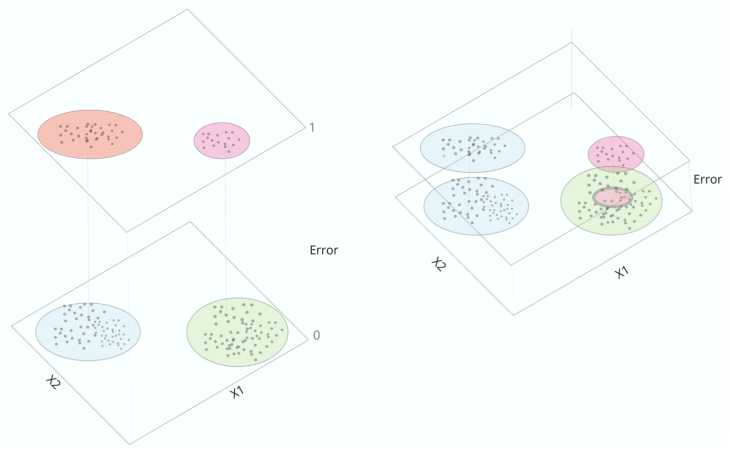
Existing methods [2,3] use **hierarchical clustering** and **K-means** with k=2. Our recent work experimented with **DBSCAN** and **Mean-Shift** (instead of K-means) and used **error labels as an input feature**.

We tested the methods on **real and synthetic data**. The latter contained predefined bias with controlled error distributions.

We demonstrated the need to **dynamically adjust the clustering parameters** as the cluster hierarchy is built, since data points are fewer and closer at each step.



We experimented with **scaling the error feature** (e.g., numeric labels {0,1} or {0, 2}). Optimal scale depends on the clustering technique.



Ongoing work explores human interpretation, dynamic clustering parameters, error scaling, **text data** (word embeddings), **splits with k>2**, and **non-hierarchical approaches.**

[1] Gupta et al. (2018). [2] Nasiriani et al. (2019)
[3] Misztal-Radecka et al. (2021).
[4] Muhammad, M.Sc. Thesis (2021).