UNIVERSITY OF TORONTO

Unlearning Fairness

Korbinian Koch, University of Toronto

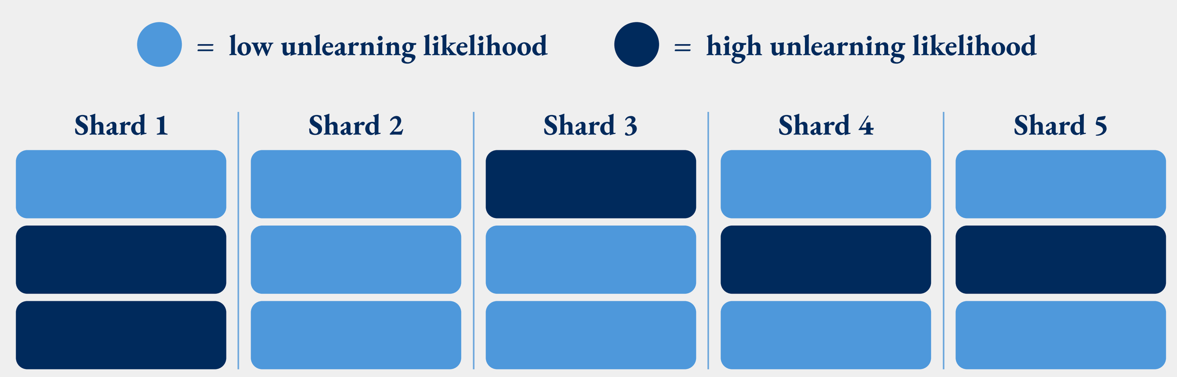
Motivation

- The speed of unlearning techniques like SISA[1] can be improved by treating samples with a high unlearning likelihood differently, typically resulting in a noticeable, but acceptable decrease in performance
- Common models may perform worse for minorities, which has been studied in further depth on the example of facial classifiers and racial minorities[2]
- There is evidence that unlearning likelihoods correlate with belonging to a protected minority

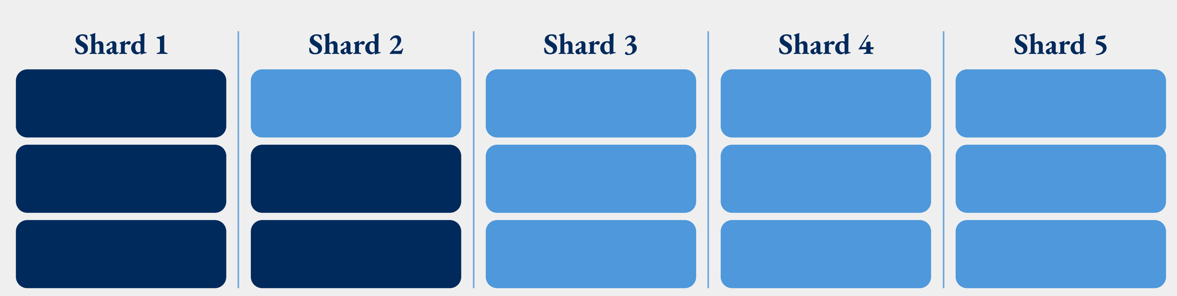
***Question*: Is model unfairness amplified in SISA unlearning strategies using a-priori estimates?**

SISA [1]

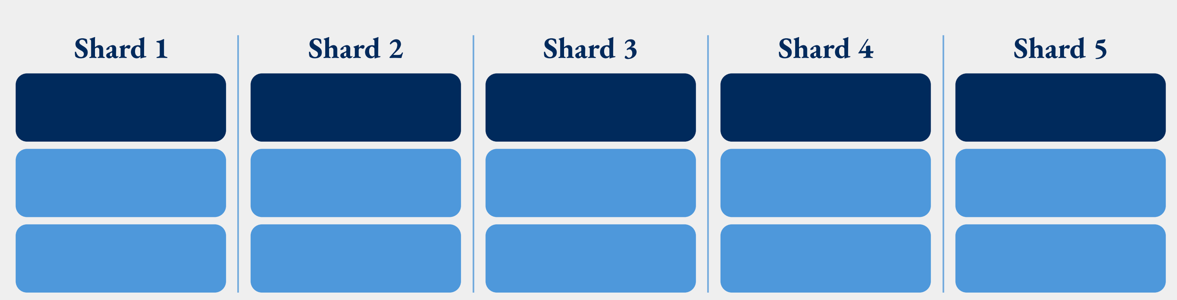
- **S**harded, **I**solated, **S**liced and **A**ggregated learning trains an ensemble of models on different subsets of the data
- Knowledge about a users individual likelihood to submit an unlearning request allows an adaptive placing of samples in specific shards or slices



random assignment to shards and slices



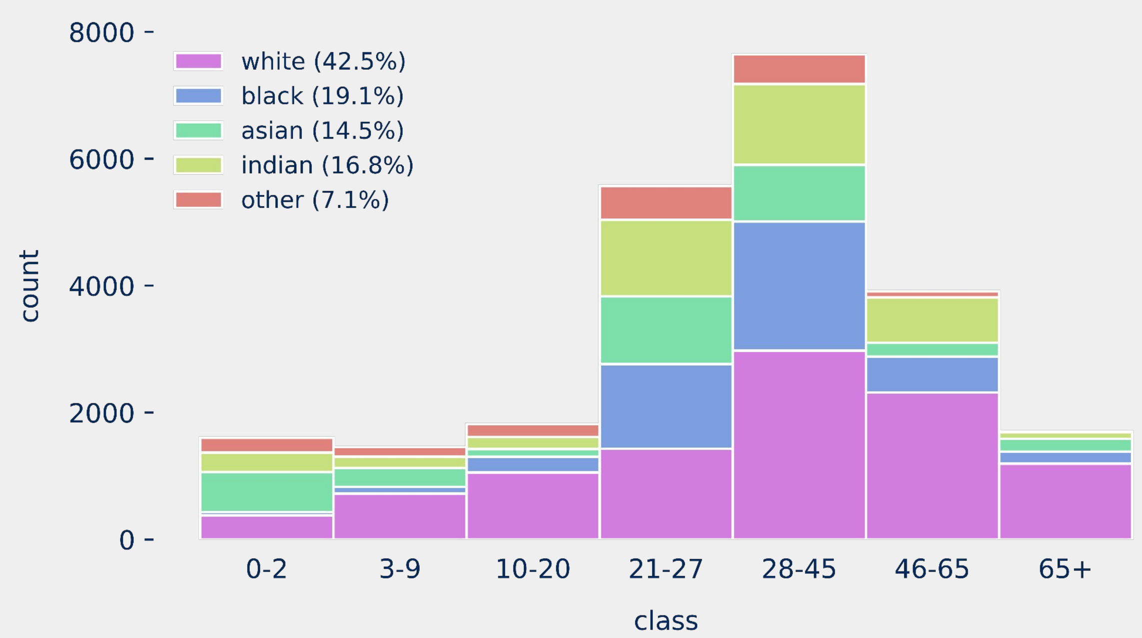
adaptive assignment into fewest possible shards



adaptive assignment into last trained slices

UTKFace [3]

The **UTKFace** dataset contains 23,708 images of human faces with labels for **age** and **race** (and gender)



Model Architecture

To predict age classes we use a **ResNet-18** pretrained on ImageNet with 3 fully connected layers incl. dropout as classifier. Regularization is further encouraged through label smoothing and image augmentation techniques.

Modelling A-Priori Likelihood

The 2019 Eurobarometer survey[4] shows that knowledge of the GDPR depends greatly on education and income. We used average US SAT scores by race to approximate it.

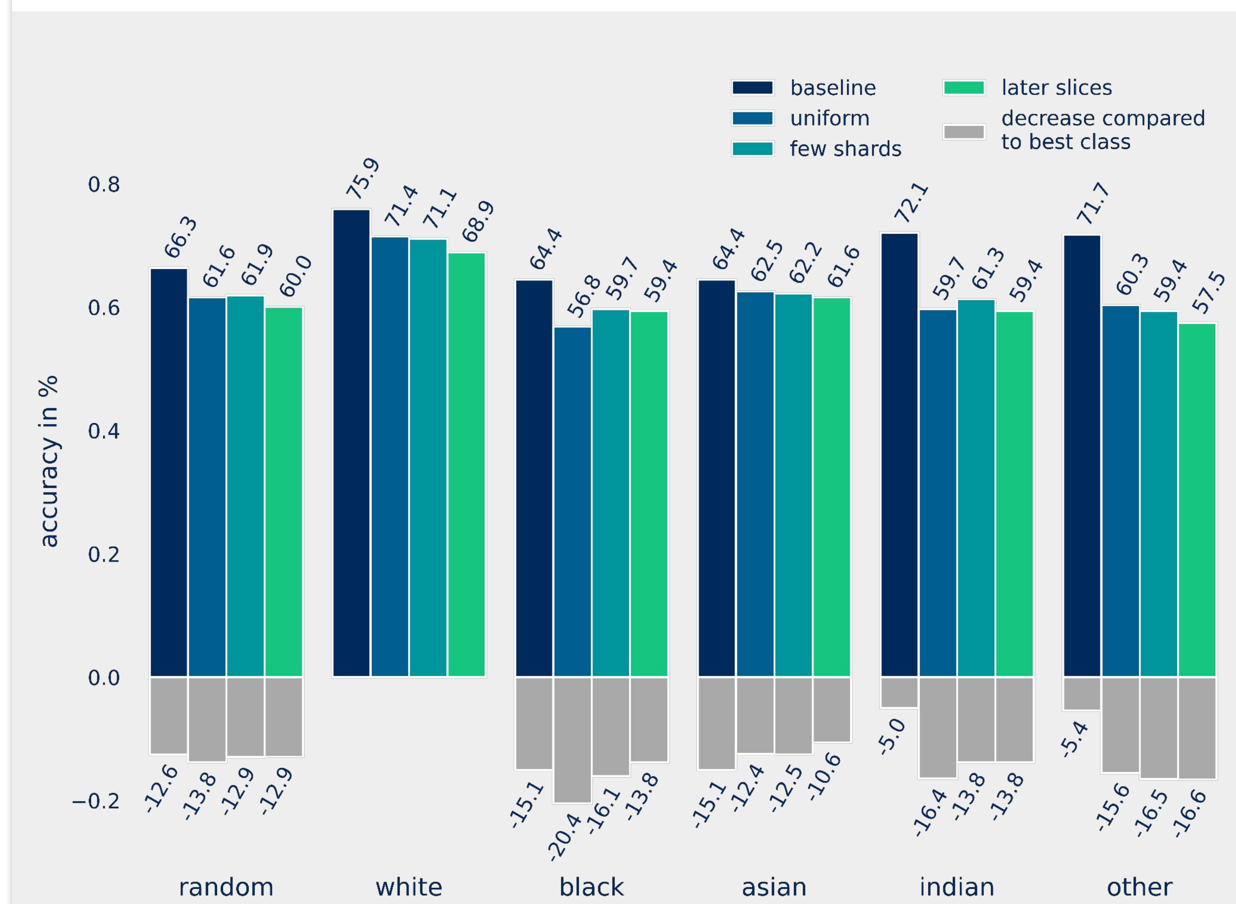
> *Knowledge of privacy legislation:* **P(k|race)**
>
> **white**: 0.4, **black**: 0.3, **asian**: 0.43, **indian**: 0.43, **other**: 0.32
>
> *Changed privacy settings:* **P(c|k)** = 0.66, **P(c|¬k)** = 0

Experiments

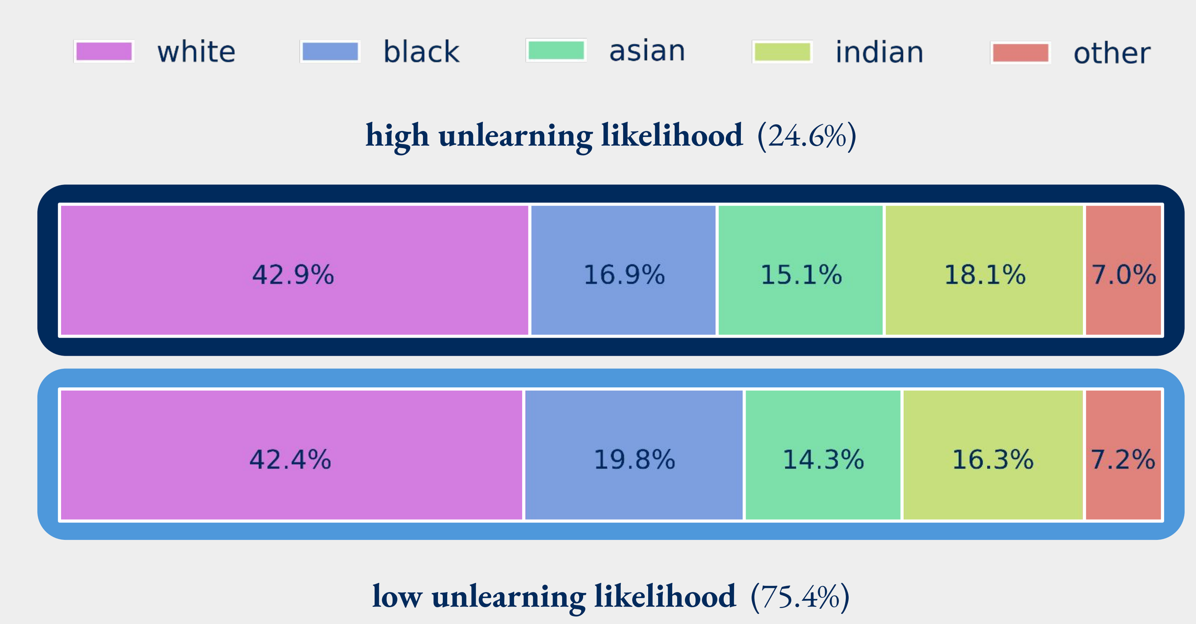
Baseline:	monolithic model
SISA:	5 Shards × 3 Slices (≈1,500 samples/slice) aggregation via summation of logits
Methods:	uniform, few shards, later slices
Test Set:	9 random samples per class + 9 random samples per class per race
Training:	15 Epochs, lr 6e-4, lr decay of × 0.2 after 6, 9, 12 epochs

all results are reported as average over 5 runs

Top-1 accuracies averaged over all classes per race and model



Average composition of slices

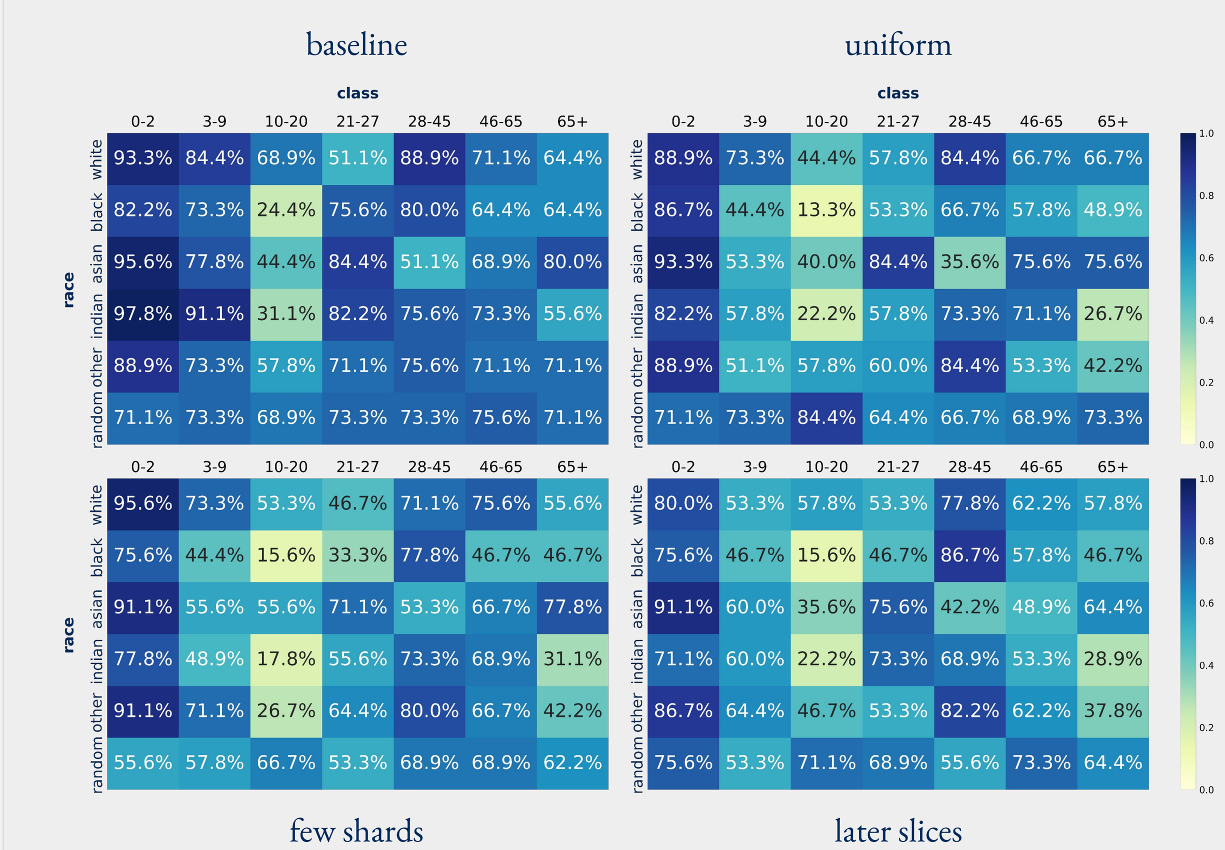


Possible causes of performance differences

A) relative amount of training samples (r^2 = 0.51, 0.81, 0.93, 0.94)
B) uniformity of class distribution (r^2 = -0.22, -0.10, 0.15, 0.24)
C) likelihood of unlearning indicator (r^2 = 0.22, 0.45, 0.43, 0.41)

correlation ≠ causation

Accuracy matrices for race and age



Conclusions

- There are considerable performance differences across races in both the baseline and all SISA models
- Realistic modelling of unlearning predictors results in only minor distribution shifts across slices
- Weaknesses of the baseline model are inherited, but not necessarily amplified by all SISA surrogates

References

[1] Lucas Bourtoule et al. "Machine Unlearning". In: *CoRR* abs/1912.03817 (2019). arXiv: 1912.03817. URL: http://arxiv.org/abs/1912.03817.

[2] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA.* Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. PMLR, 2018, pp. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html

[3] Zhifei Zhang, Yang Song, and Hairong Qi. "Age Progression/Regression by Conditional Adversarial Autoencoder". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017.* IEEE Computer Society, 2017, pp. 4352–4360. DOI: 10.1109/CVPR.2017.463. URL: https://doi.org/10.1109/CVPR.2017.463.

[4] *Special Eurobarometer 487a*. (2019). ISBN: 978-92-76-08384-9. DOI: 10.2838/579882. URL: https://europa.eu/eurobarometer/ surveys/detail/2222.