GENDER SHADES

Intersectional Accuracy Disparities in Commercial Gender Classification

Authors of the article: Joy Buolamwini, Timnit Gebru

Overview

- Commercial algorithms discriminate based on gender and race
- Benchmark datasets overwhelmingly polarised
- Relabelling of datasets using Fitzpatrick Skin Type Classification System
- An additional dataset – more representative
- Evaluate three commercial systems
- Inspect error rates and discover large intersectional error rates

Current state of affairs

- Commercial facial recognition: IBM, Microsoft, Face++, Google
- Clients: government
- Other research:
  - *identifying emotions,*
  - *helping people with autism*
  - *determining sexuality of Caucasian males from photos from Facebook and dating sites*
  - *determining individuals characteristics(IQ, terrorism, prone to criminal actions)*

Current state of affairs

- Issue with skewed results provided from different systems
  not provided per gender or per racial/ethnical group
- Benchmark datasets highly skewed

  “LFW, a dataset composed of celebrity faces which has served as a gold standard benchmark for face recognition, was estimated to be 77.5% male and 83.5% White (Han and Jain, 2014).”

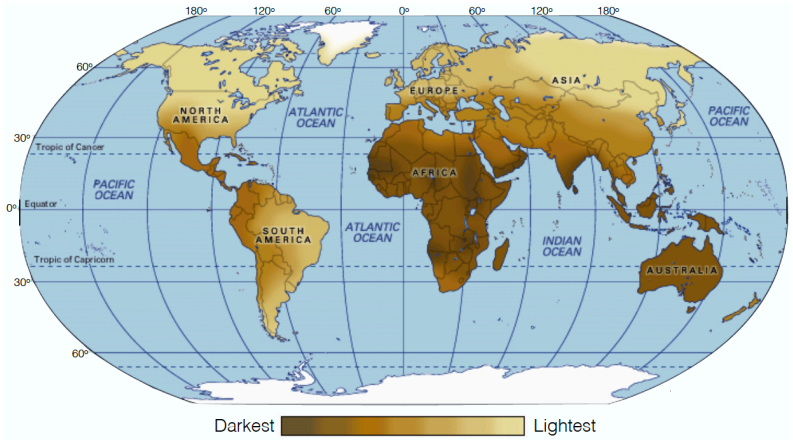


Figure 2: The global distribution of skin color. Most Africans have darker skin while those from Nordic countries are lighter-skinned. Image from (Encyclopedia Britannica) ©Copyright 2012 Encyclopedia Britannica.

Fitzpatrick skin type

Fitzpatrick skin type		
Skin type	Typical features	Tanning ability
I	Pale white skin, blue/green eyes, blond/red hair	Always burns, does not tan
II	Fair skin, blue eyes	Burns easily, tans poorly
III	Darker white skin	Tans after initial burn
IV	Light brown skin	Burns minimally, tans easily
V	Brown skin	Rarely burns, tans darkly easily
VI	Dark brown or black skin	Never burns, always tans darkly

Source: Dermnet.nz

BENCHMARKS

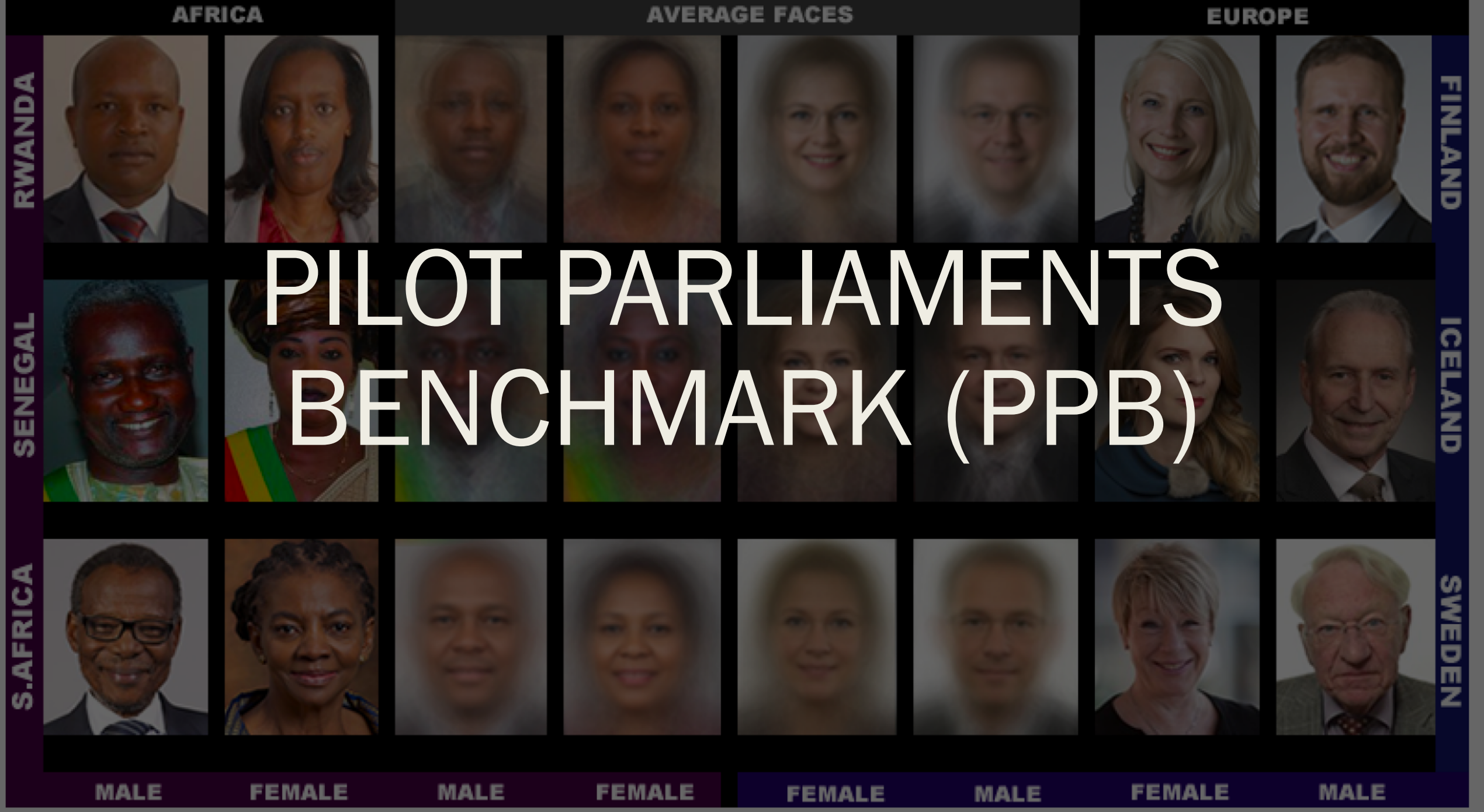
Property	PPB	IJB-A	Adience
Release Year	2017	2015	2014
#Subjects	1270	500	2284

IJB-A

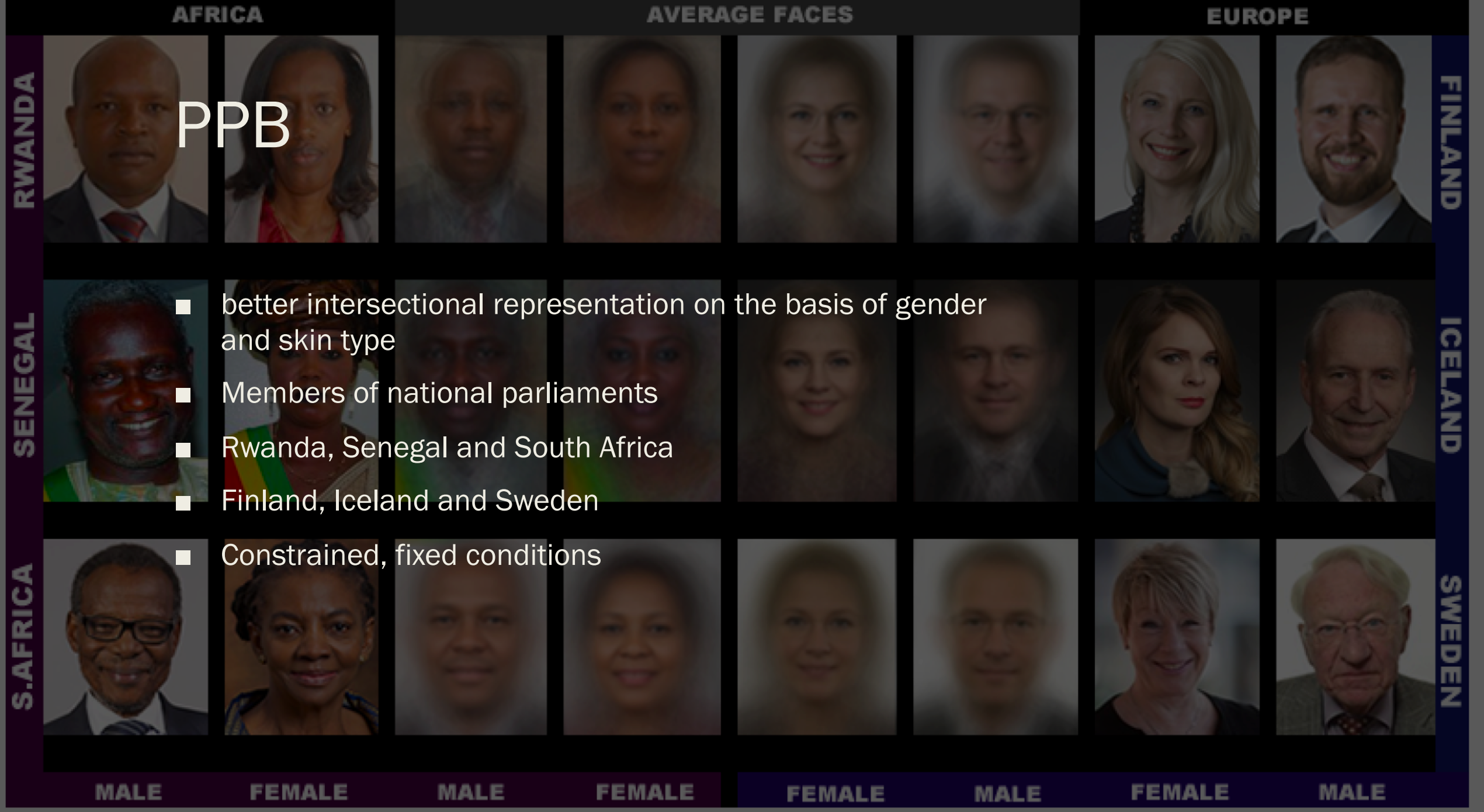
- Released by US National Institute of Standards and Technology
- The dataset consisted of 500 unique subjects who are public figures
- Vary in pose and illumination

Adience

- The Adience benchmark contains 2,284 unique individual subjects.
- 2,194 possible to be labelled by gender and skin type
- Vary in pose and illumination



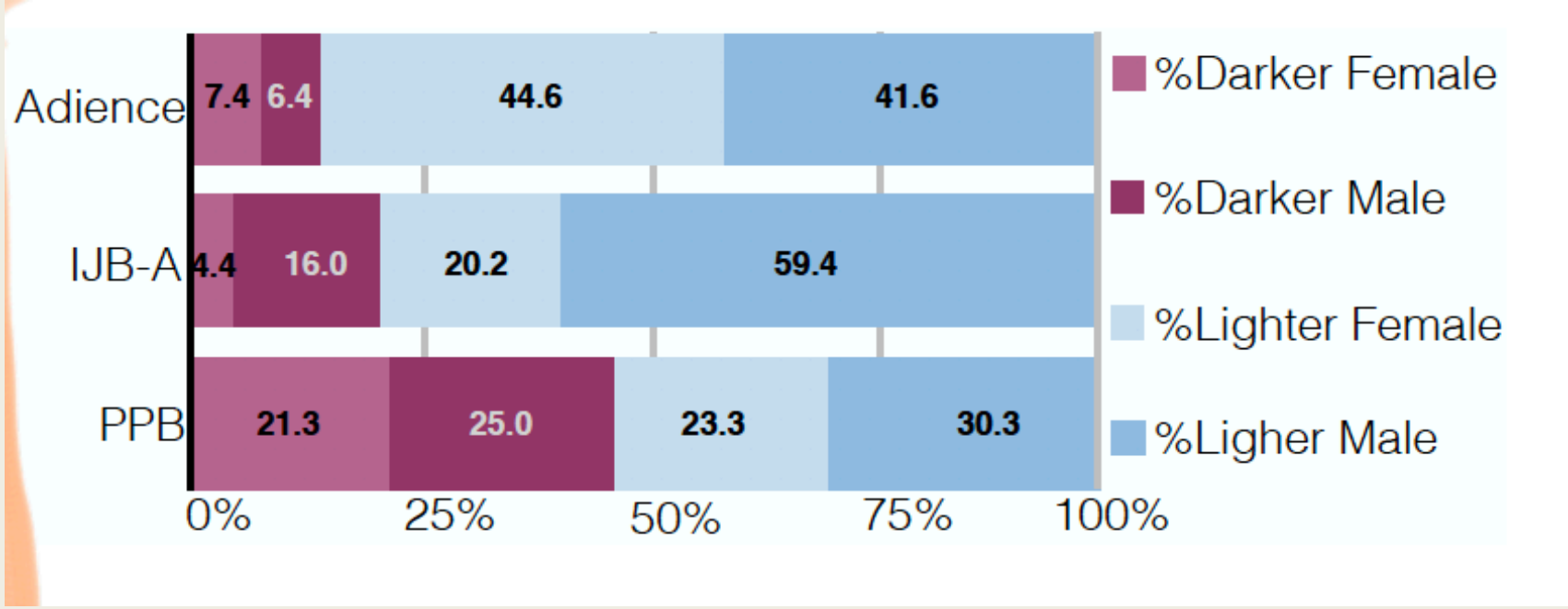
PILOT PARLIAMENTS BENCHMARK (PPB)



PPB

- better intersectional representation on the basis of gender and skin type
- Members of national parliaments
- Rwanda, Senegal and South Africa
- Finland, Iceland and Sweden
- Constrained, fixed conditions

Distribution of different groups



Here light group is types I-III, and dark IV-VI.

Commercial Gender Classification Selection

- IBM
- Microsoft
- Face++
- Goggle’s classifier was not publicly available

“None of the commercial gender classifiers chosen for this analysis reported performance metrics on existing gender estimation benchmarks in their provided documentation.”

Commercial Gender ClassIfication - Evaluation Criteria

Metric
PPV(%)
Error Rate(%)
TPR(%)
FPR(%)

- PPV– Positive predictive value

  TP/(TP + FP)

- Error rate – (1-PPV)
- True Positive Rate – TPR
- False Positive Rate - FPR

Commercial Gender Classification - Audit Results

- Male and female error rates
- Darker and lighter error rates
- Intersectional error rates

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Commercial Gender Classification - Audit Results

- All classifiers perform better on male faces than female faces (8.1% - 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% - 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% - 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

Conclusion

Further work

- do similar errors persist in other human-based computer vision tasks?
- Explore intersectional error analysis of facial detection, identification and verification
- Inclusive datasets
- Increase transparency and accountability in artificial intelligence

“Because algorithmic fairness is based on different contextual assumptions and optimizations for accuracy, this work aimed to show why we need ***rigorous reporting on the performance metrics*** *on which algorithmic fairness debates centre.”*

THANK YOU!