Covariate shift - More Math

Covariate Shift and GANs

- Generative Adversarial Networks
  - **Generator** reweights training data such that it is indistinguishable from test data
    $$\{(x_1, y_1), \ldots (x_m, y_m)\} \to \{\alpha_1(x_1, y_1), \ldots \alpha_m(x_m, y_m)\}$$
  - **Discriminator** tries to distinguish training and test data
    $$\underset{\alpha}{\text{minimize}}\,\underset{w}{\text{maximize}} \sum_{i=1}^{m} \alpha_i \log p(f(x_i, w), 1) + \frac{1}{m'} \sum_{i=1}^{m'} \log p(f(x_i', w), -1)$$
- Theorem: this is minimized for $\alpha(x) = q(x)/p(x)$

Proof - Covariate Shift and GANs

- Loss per covariate

$$c \log r(y=1|x) + d \log r(y=-1|x)$$

$$= (c+d)\left[\frac{c}{c+d}\log r(y=1|x) + \frac{d}{c+d}\log(1-r(y=1|x))\right]$$

$$= (c+d)[\gamma \log \rho + (1-\gamma)\log(1-\rho)]$$

- Maximizing wrt. ρ yields $\rho = \gamma$, i.e. $-(c+d)H[\gamma]$
- Minimizing with regard to γ yields $\gamma = 0.5$ (entropy mode) Distributions must not be distinguishable.

Proof - Covariate Shift and GANs

Functional derivative with respect to f yields

$$\partial_r \left[\int dp(x)\alpha(x)\log r(y=1|x) + \int dq(x)\log r(y=-1|x) \right]$$

$$= \partial_r \int dq(x) \left[\alpha(x)\frac{p(x)}{q(x)} \log r(y=1|f(x)) + \log r(y=-1|x) \right]$$

Using optimality yields $\alpha(x) = \dfrac{q(x)}{p(x)}$, i.e. same as before

More Connections

Maximum Entropy (Agarwal, Li, Smola, 2011)

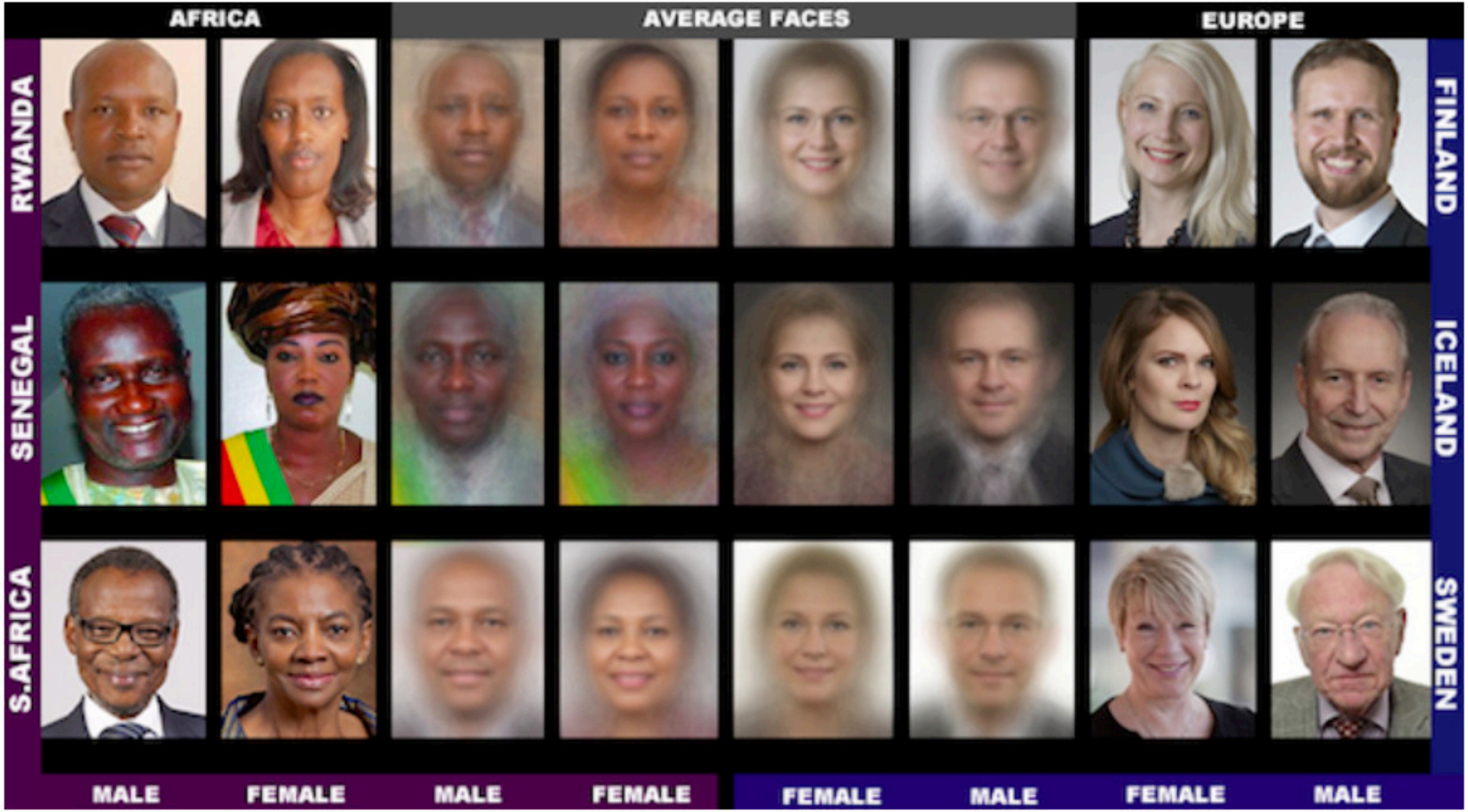
Solving the classification problem is equivalent to maximum entropy subject to matching moments

$$\underset{\alpha}{\text{maximize}}\, H[\alpha] \text{ subject to } \left\| \frac{1}{m}\sum_{i=1}^{m} \alpha_i \phi(x_i) - \frac{1}{m'}\sum_{i=1}^{m'} \phi(x'_i) \right\|^2 \leq \epsilon$$

Works for Deep Network feature maps, too!

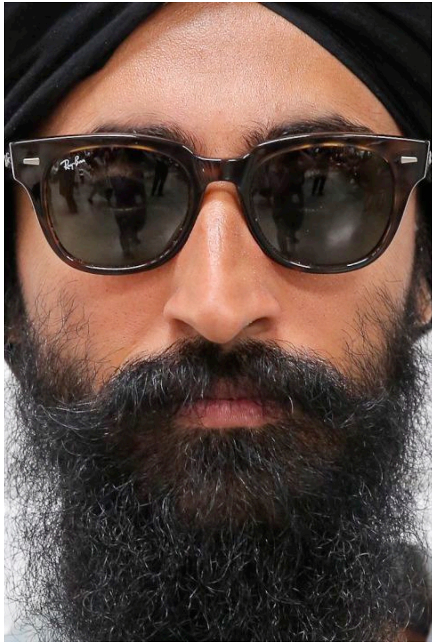
Case Study - Face Recognition

Pilot Parliaments Benchmark (Buolamwini & Gebru, 2018)



Face Recognition

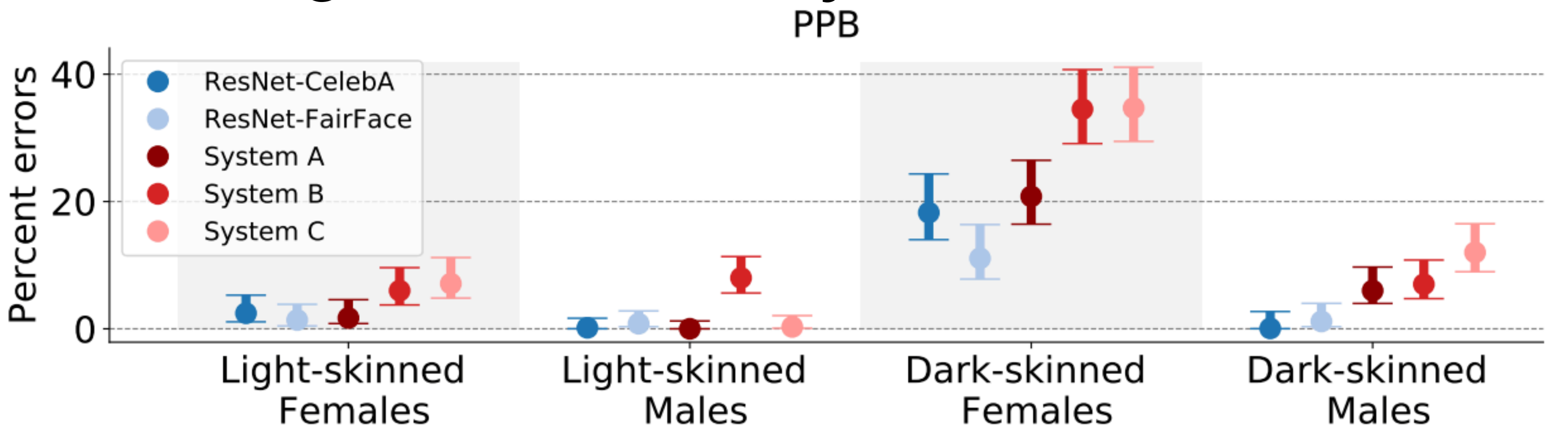


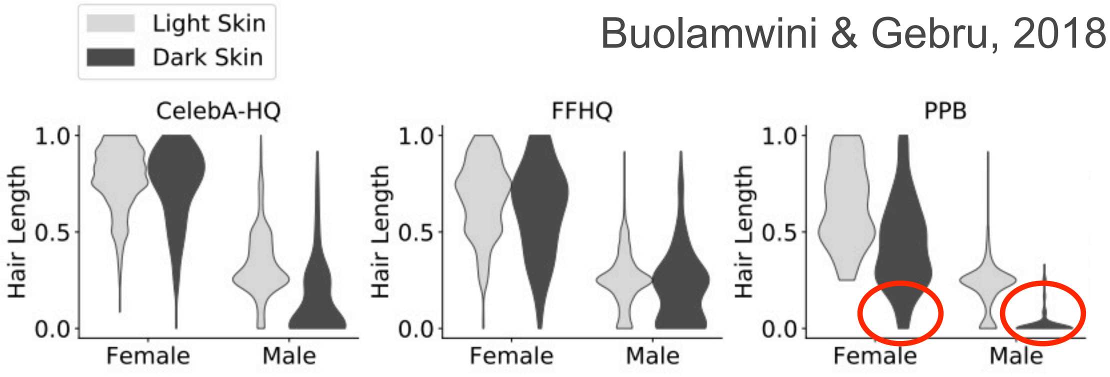
Fewer features make it easier to recognize a face …

… but that can correlate with other attributes …

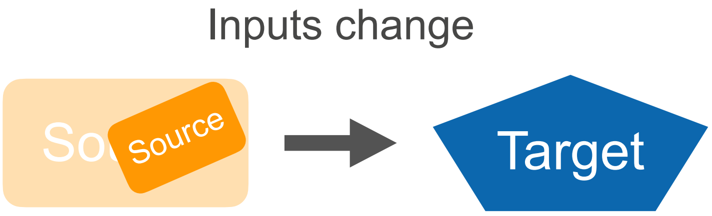
Face recognition accuracy



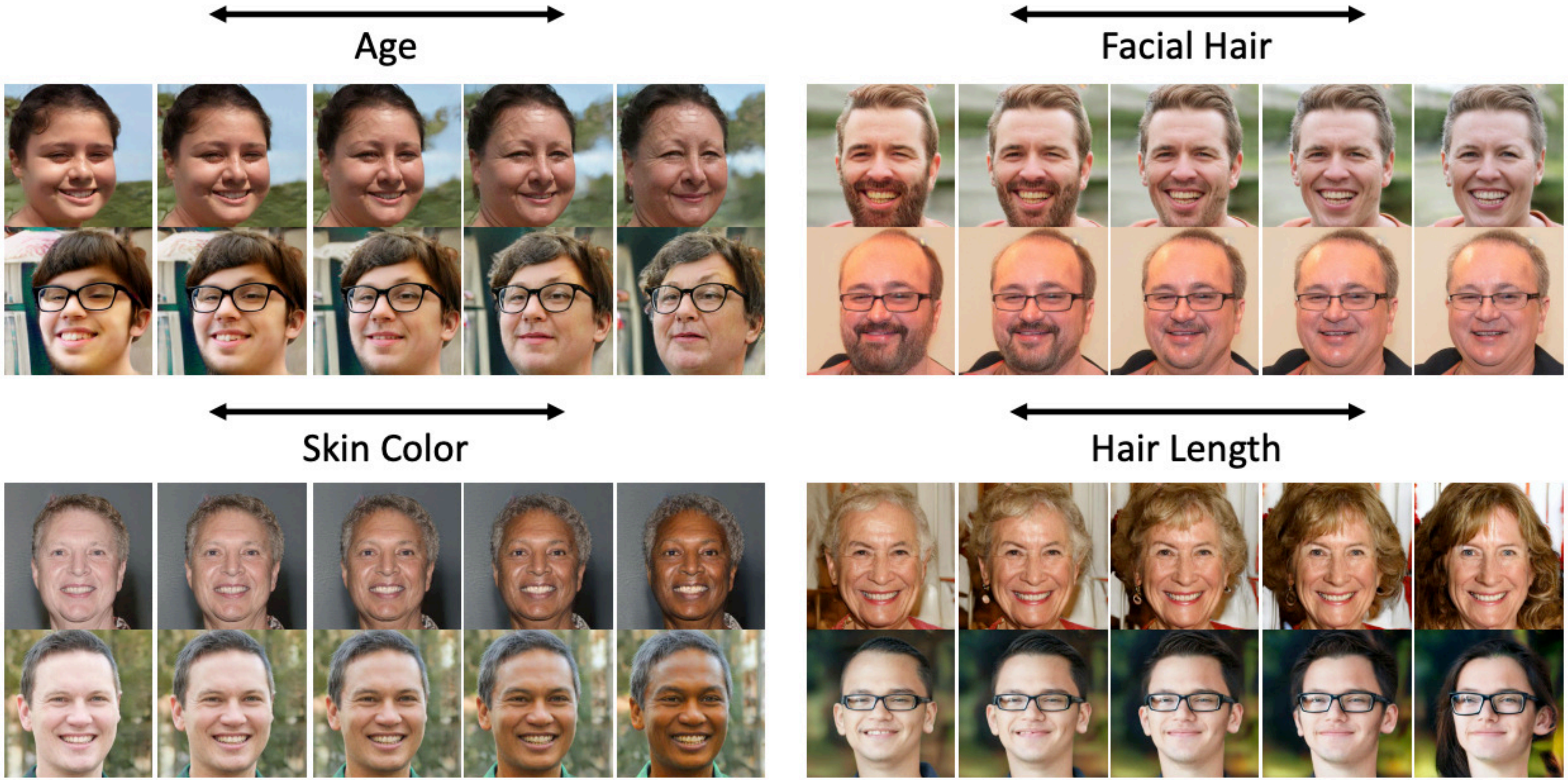
Buolamwini & Gebru, 2018



Balakrishnan et al, 2020

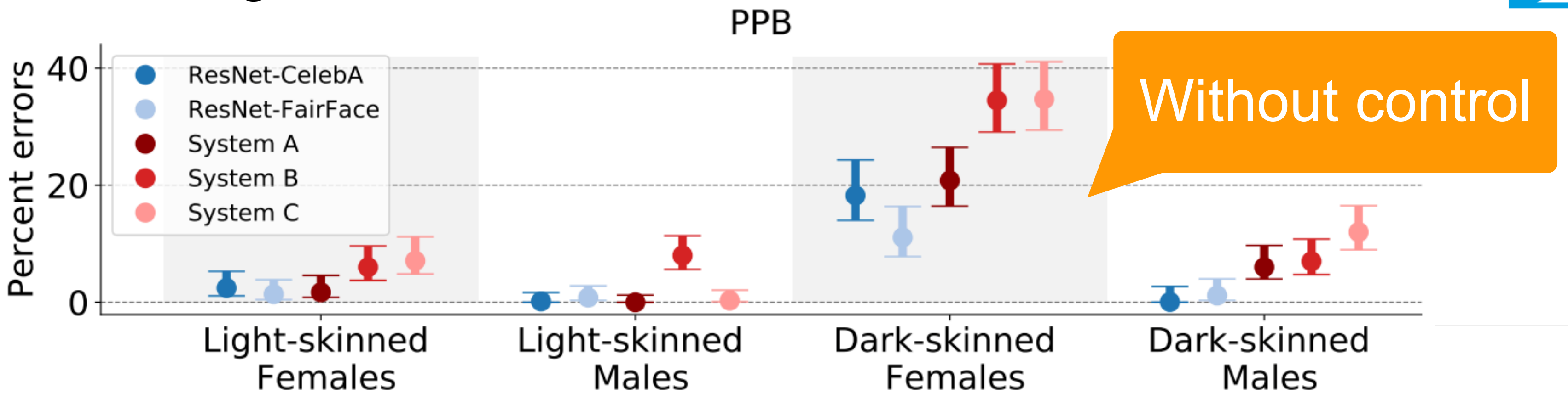


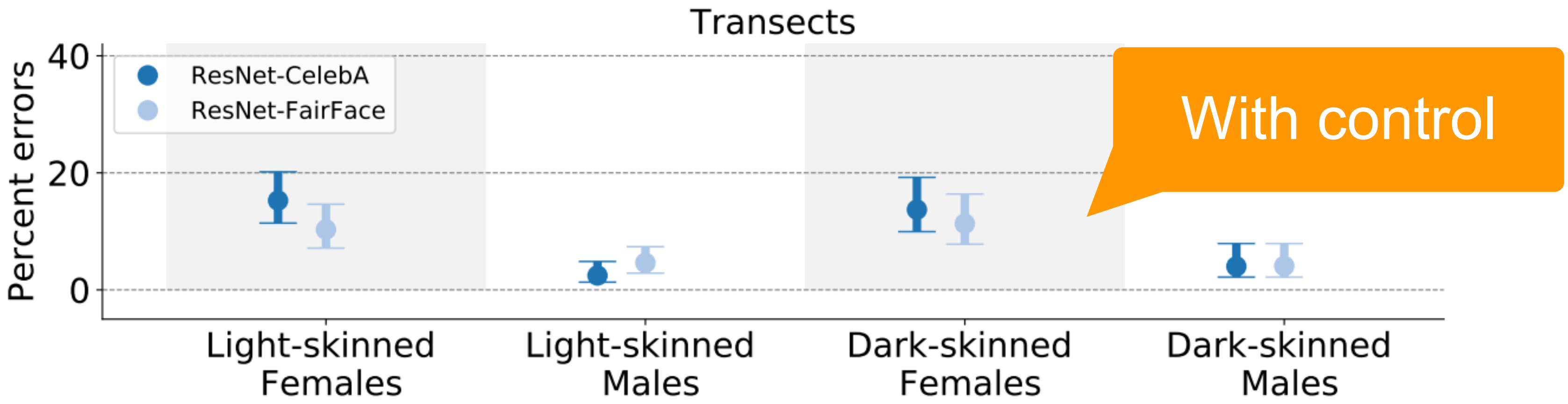
Controlling for Attributes



Balakrishnan et al, 2020

Controlling for Attributes





Balakrishnan et al, 2020

In Math ...

Given loss $l(y, f(x))$ with

- Mean $R[p, f] := \mathbf{E}_{(x,y)\sim p}[l(y, f(x))]$
- Variance $\sigma^2[p, f] := \underset{(x,y)\sim p}{\mathrm{Var}}[l(y, f(x))]$

There exists a $q(x)$ such that $R[q, f] \geq R[p, f] + \sigma$

Proof: Mean value theorem implies that there exists at least some $\mathscr{X}'$ with $\mathbf{E}_{y|x}[l(y, f(x))] \geq R[p, f] + \sigma$. Pick $q(x)$ on $\mathscr{X}'$.

Key Takeaways

- You can always find a distribution that makes it worse.
  - Simply overweight data with large errors.
  - Ensure proper matching before drawing conclusions.
- Covariate Shift
  - Fixed via GAN (weighting)
  - Fixed via MMD features and MaxEnt
  - Lots more feature-based fixes (not all are sufficient)

Training ≠ Testing

- **Generalization performance**
  (the empirical distribution lies)

  $$p_{\mathrm{emp}}(x, y) \neq p(x, y)$$

- **Covariate shift**
  (the covariate distribution lies)

  $$p(x) \neq q(x)$$

- **Adversarial data**
  (the support of the distribution lies)

  $$\mathrm{supp}(p) \neq \mathrm{supp}(q)$$

- **Label shift**
  (the label distribution lies)

  $$p(y) \neq q(y)$$