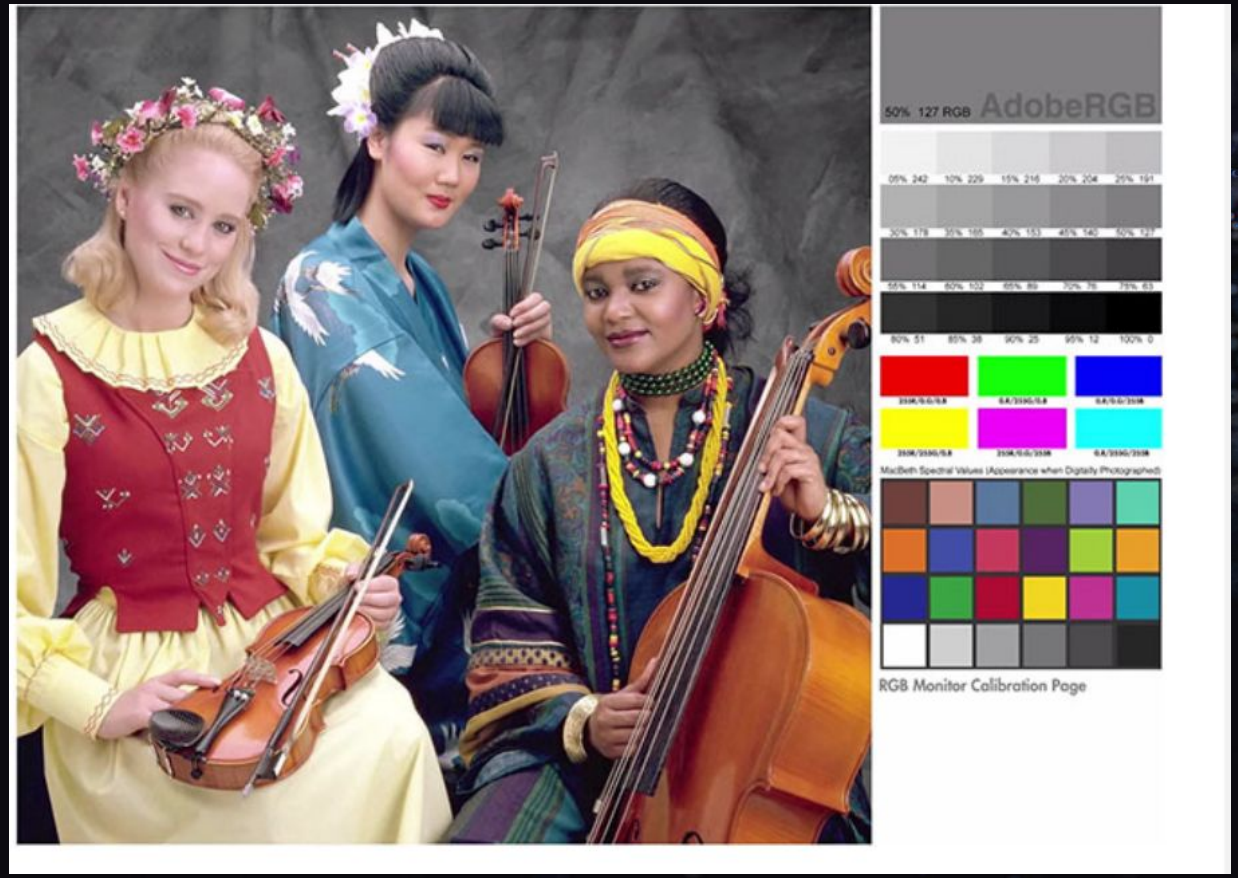
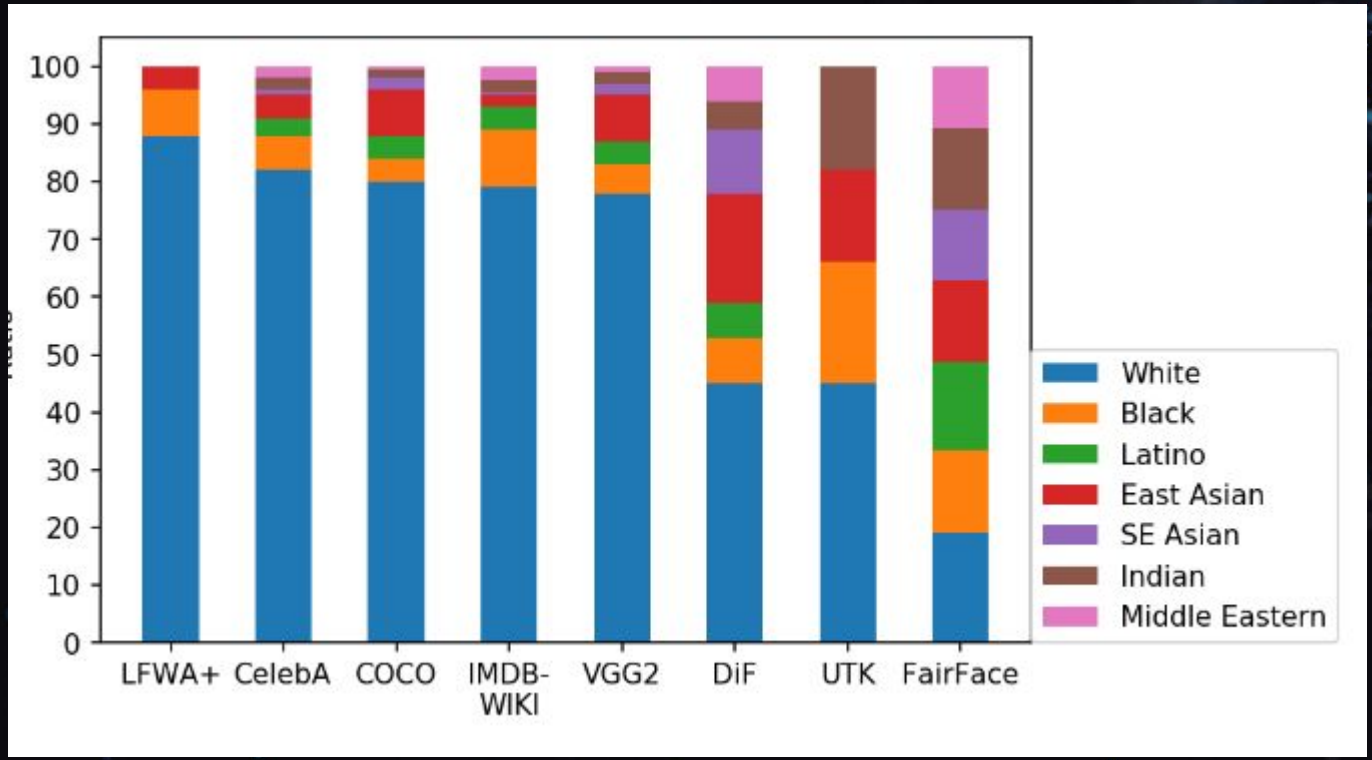
Sesgo: Soluciones

Nueva imagen de referencia Kodak (1995)



FairFace: balancear los datos



FairFace vs otros

- Mejor accuracy
  - No solo para negros/latinos/etc

Fuentes de sesgo

- Desbalance de datos
  - En el mundo real
    - Realmente hay menos mujeres en STEM
  - En la selección de datos
    - CelebA: actores no representan la población
  - Latentes/ocultos
    - El precio de la publicidad por género
- Sesgos del modelo
  - Árbol/Reglas: asume que Atributo=Valor captura el dominio

Datos desbalanceados

- Genera un sesgo en mi problema?
  - No: los dejo
    - Ej: Usar CelebA para evaluar famosos
  - Si: balanceo
    - a) Tomo más muestras
      - Fair Face
    - b) Balanceo las actuales
      - a) Quito ejemplos hasta balance
      - b) Genero sintéticamente
    - **Evaluar el sesgo del modelo**

Quitar atributos: Una solución que no funciona

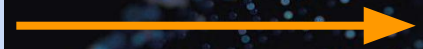
- Dataset desbalanceado en Atributo Género
  - Lo quito para que no sea un factor
    - Pero hay atributos altamente correlacionados
      - Ej: profesión, sueldo, etc
    - **Quitar = ocultar el sesgo**
      - Nunca quitar
  - Dejar el atributo
    - Utilizarlo para evaluar el sesgo
    - (Aunque sea solo una prueba interna)

Evaluar modelo final

- Verificar que es **invariante** al atributo sesgado
  - P(condenado | atributos y blanco) = P(condenado | atributos y negro) = P(condenado | atributos y X)
  - P(contratar | atributos y mujer) = P(contratar | atributos y hombre) = P(contratar | atributos y X)
- Invariante
  - Dadas transformaciones $t_1, t_2, \ldots, t_m$
  - $f(t_1(x)) = f(t_1(x)) = f(t_2(x)) \ldots = f(t_m(x))$
    - Ej: $t_1(x)$ = asignar x.genero = mujer

Balanceo de datos con generación sintética

ID	Genero	...	Contratar
Juan	M	...	Si
Marta	F	...	No
Luis	M	...	Si
Pepe	M	...	No
...	...	...	...



ID	Genero	...	Contratar
Juan	M	...	Si
Marta	F	...	No
Luis	M	...	Si
Pepe	M	...	No
...	...	...	...
Juan	F	...	Si
Marta	M	...	No
Luis	F	...	Si
Pepe	F	...	No

- Es suficiente con el género?
  - NO: Balancear atributos correlacionados

Campo de investigación activa

- Técnicas no son maduras
- Avances en modelos > avances en prevención de sesgos
- ¿Debemos eliminar sesgos?
  - Distribuciones en el mundo
    - Depende de qué datos
  - Distribuciones según ética/sentido común
    - Depende de cuáles personas
  - Distribución en el modelo
    - Depende de qué empresa