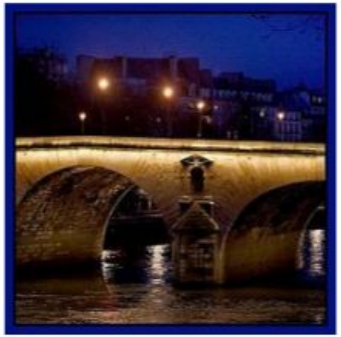
Visual Datasets and Bias

Do datasets provide a *good* representation of the world?

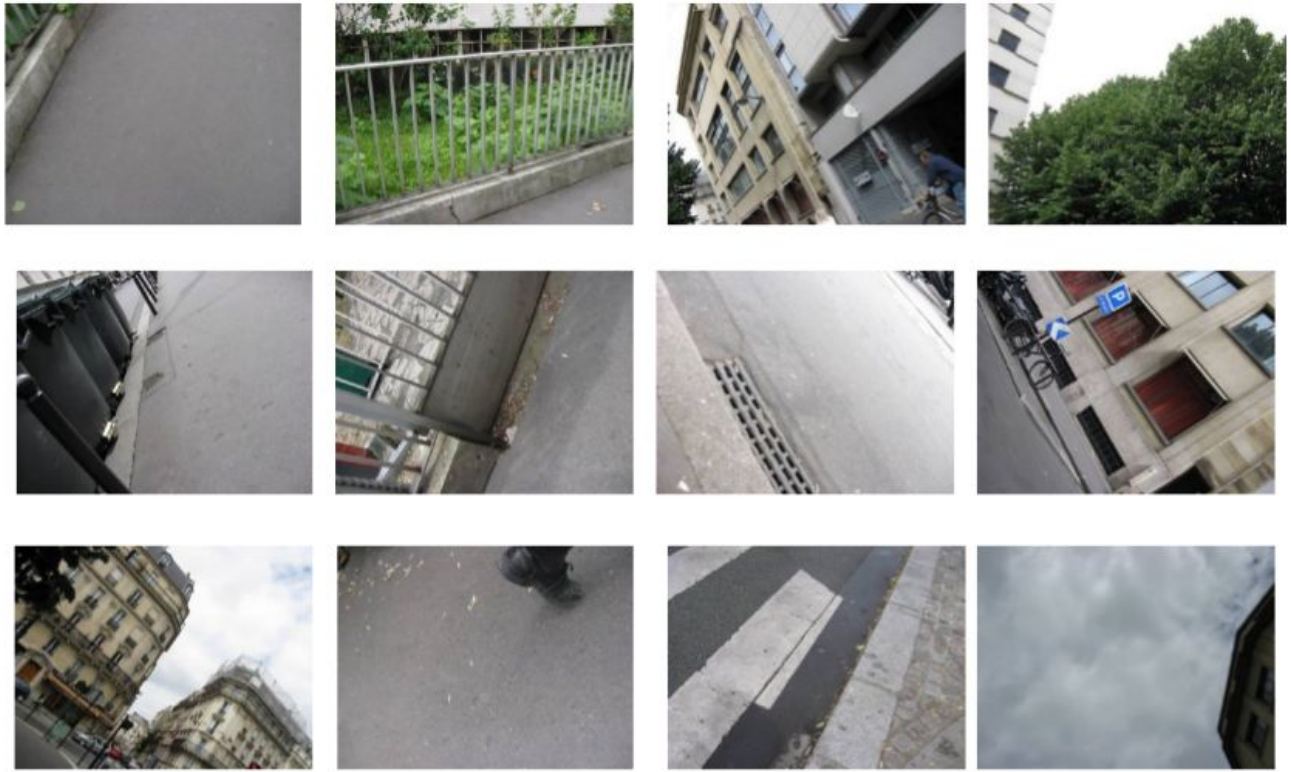
Visual Data is Inherently Biased

- Internet is a tremendous repository of visual data (Flickr, YouTube, Picassa, etc)
- But it's not random samples of visual world

Flickr Paris



Sampled Alyosha Efros's Paris



Sampling Bias

- People like to take pictures on vacation

Photographer Bias

- People want their pictures to be recognizable and/or interesting

vs.

Social Bias

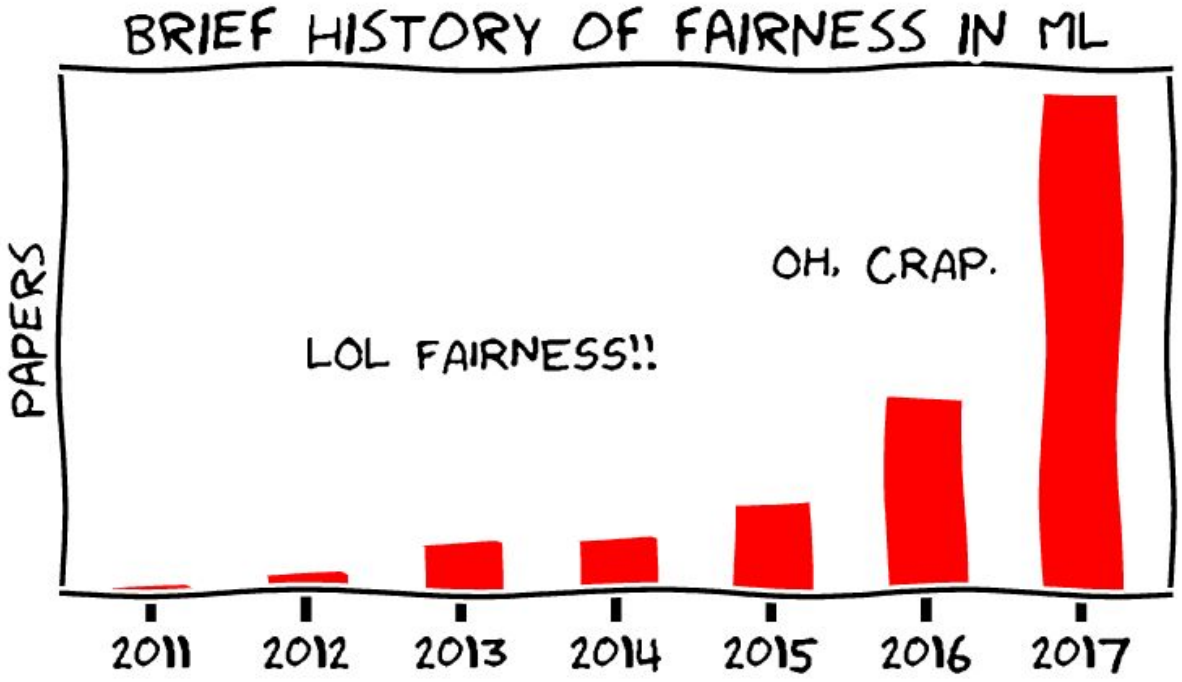
Little Leaguer

Kids with Santa

The Graduate

Newlyweds

“100 Special Moments” by Jason Salavon


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH, CRAP.
2011
2012
2013
2014
2015
2016
2017

Where does bias come from?

Collection:

- Demographic, geographic, behavioral, temporal biases,

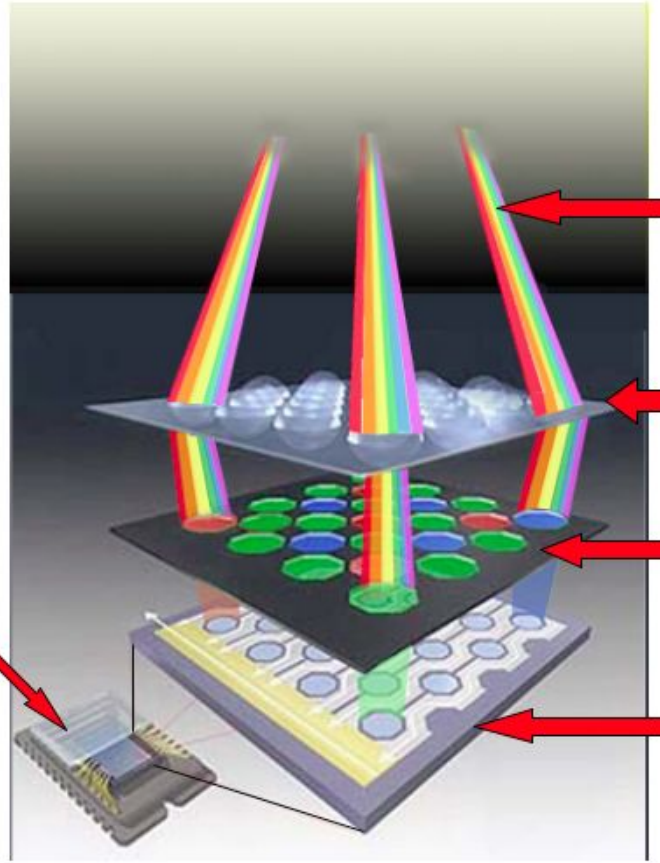
Measurement:

- What do we choose to measure? How do we measure (e.g., *grit*)?

Pre-existing biases

- gender roles in text and images, racial stereotypes in language

RGB Inside the Camera



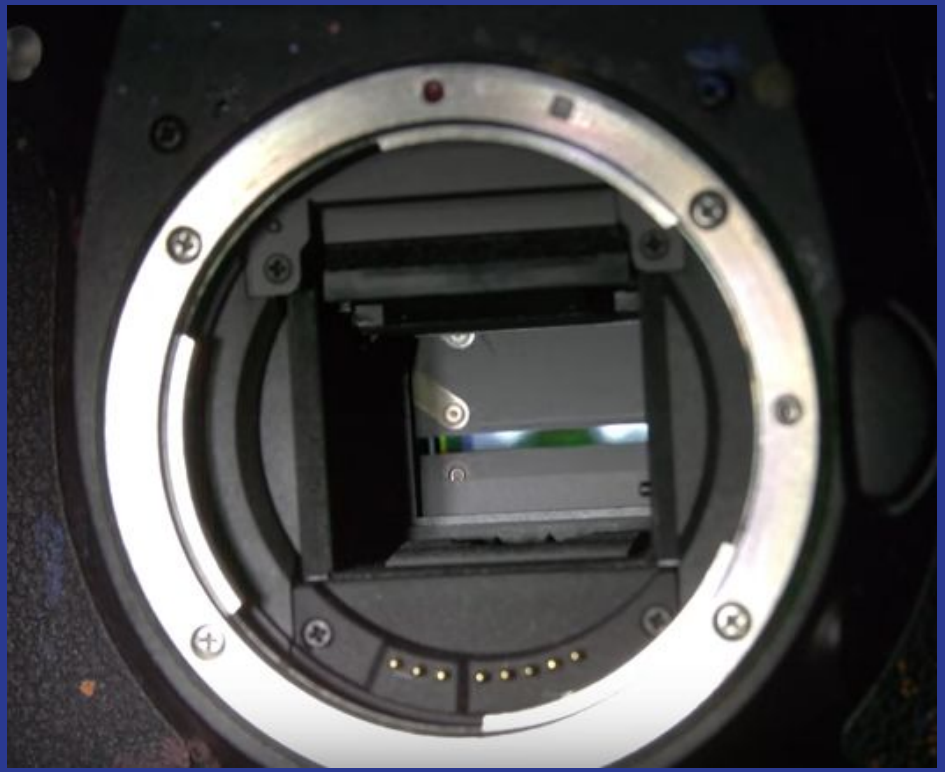
Incoming Visible light

Visible Light passes through IR-Blocking Filter

Color Filters control the color light reaching a sensor

Color blind sensors convert light reaching each sensor into electricity

Millions of light sensors



https://www.youtube.com/watch?v=CmjeCchGRQo&t=361s