Bias & Fairness in ML

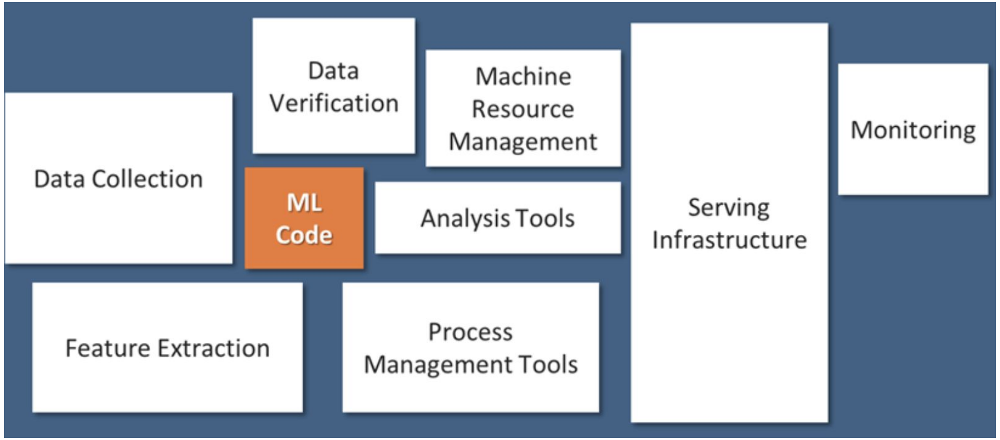
ML systems are hard

- lack of understanding of how exactly the neural nets work (black box,million params)

- dependencies on the training data, evaluation data, hidden biases

- real life context is hard, human users

- support systems (infrastructure, machines, monitoring) - MLOps



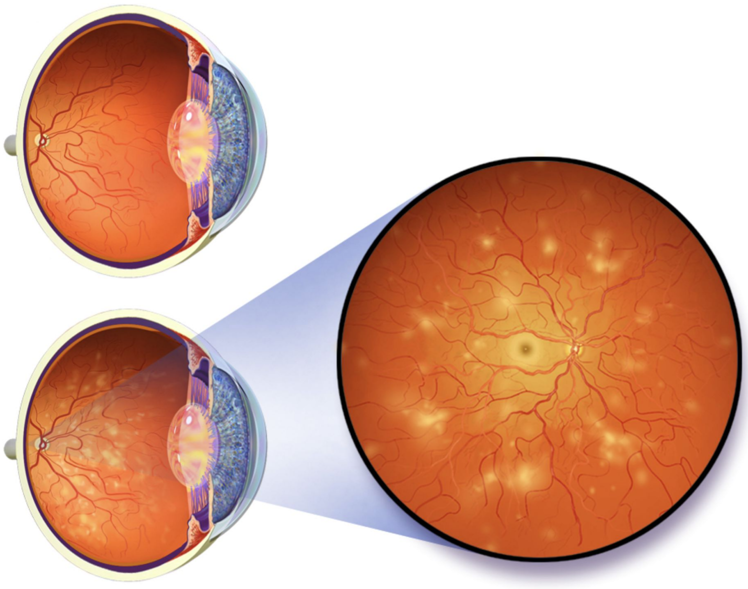
Case Study: Google Health 2020

- diabetic retinopathy - eye disease

- 80% diabetic people could develop it - is the result of damage to the small blood vessels and neurons of the retina.

- one of the leading causes of blindness

- it has no early warning signs, but if detected in time, blocked or leaking blood vessels, high chance that it can be treated (90%)



10 weeks vs 10 minutes

- Thailand Health Ministry: annual screening of 60% of diabetic people for diabetic retinopathy

- roughly 4.5 million patients, only 200 eye specialists, spread around the country

- nurses take photos of patients’ eyes, send them off to specialist (can take up to 10 weeks)

- Google Health developed an ML system with 90% accuracy in identify signs of diabetic retinopathy (“human specialist level”) – result in less than 10 minutes

Lab accuracy != Real life outcomes

Assumptions

- training data: high quality images

- better to reject, than give lower accuracy result

- not relying on nurse judgement

- high bandwidth internet connection

Reality

- test data: low quality images - different set up, poor lighting, dozens of patients per hour (1 of 5 rejected)

- rejected patients - unnecessary travel, missed work, anxiety, no car

- frustrated nurses, the algorithm rejected scans showing no signs of disease (unnecessary follow-up) - time wasted retaking/editing images

- patients/nurses expected instant results, but slow internet connection sometimes made them wait hours instead of minutes

Lessons learned

user-centered design process

- understand how AI tools are going to work for people in a context
- there is more to health care than algorithms
- people’s needs, expectations, trust

- reality vs AI hype (COVID coughing, tongue, X-ray)

- when it works well, huge benefit - unstoppable nurse 1000 patients scanned on her own

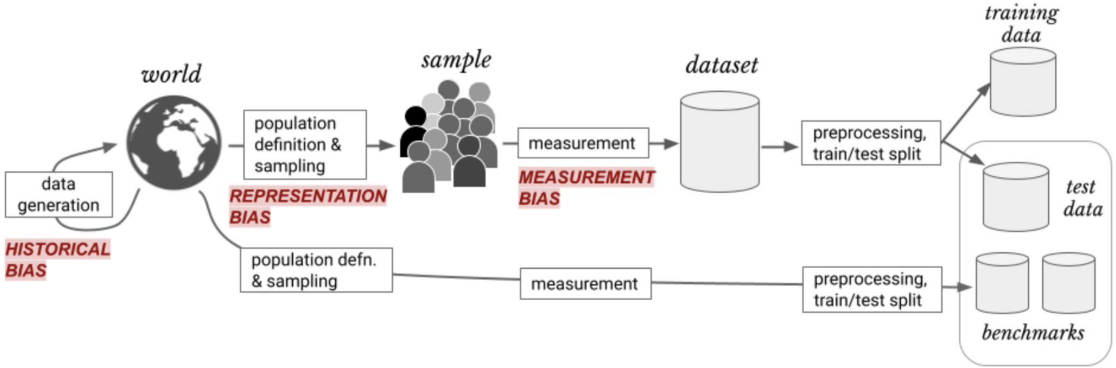
- patients didn’t care who read images (AI or human) - they cared more about their experience - waiting time, follow-up

- is it useful comparing AI to human? human doctors disagree all the time - AI tools need process to discuss uncertainties, rather than simply reject

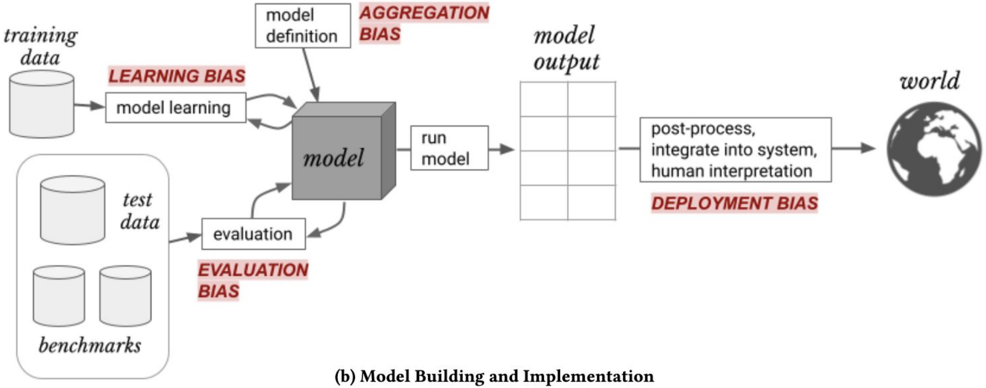
Bias - unintended consequences

- where does it come from?

- how to address it?



(a) Data Generation



(b) Model Building and Implementation

Representation bias

- training data is not representative (mostly light skinned man, vs only 4% dark skinned women)

- real life data, on which the system is used is vastly different, more diverse

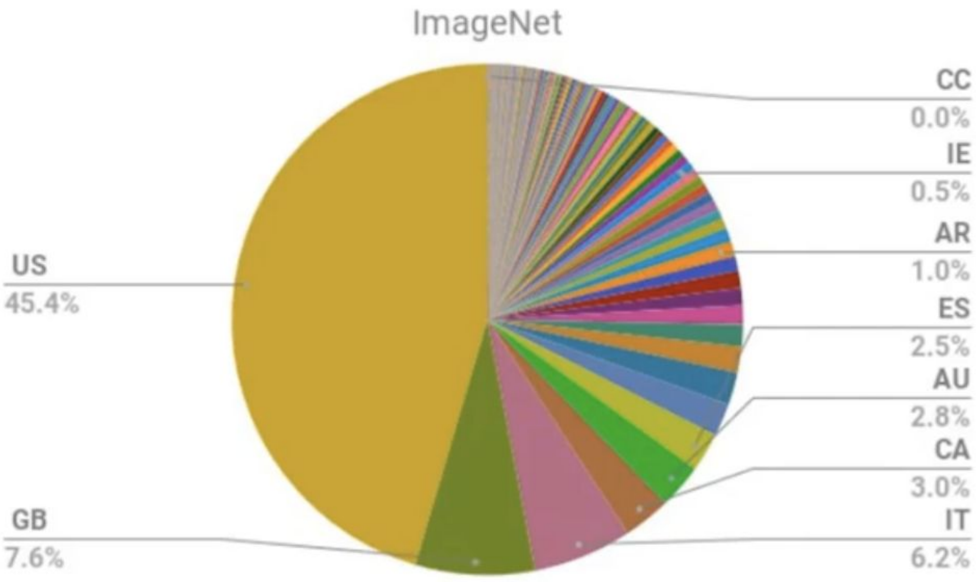
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Joy Buolamwini & Timnit Gebru, gendershades.org

Representation bias

- training data is mostly images from the western world/culture - not representative

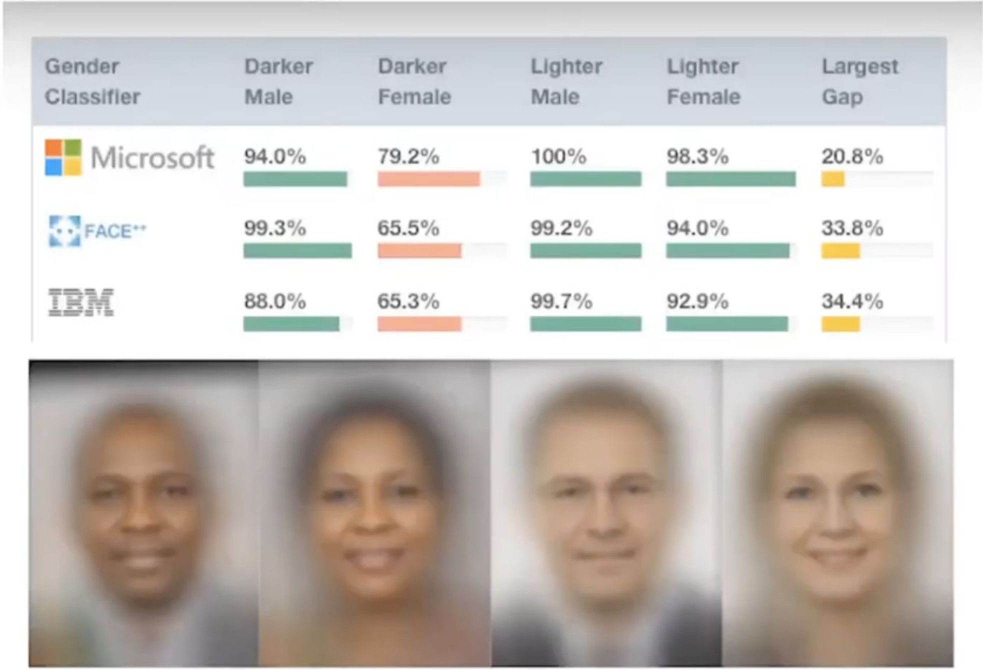
- real life data - algorithm errors - for wedding bride/groom from india/egypt is mislabelled by algorithms



Evaluation bias

- same benchmark dataset used by many companies to develop algorithms

- biases replicated at scale



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Joy Buolamwini & Timnit Gebru, gendershades.org

Historical bias

- fundamental, structural issue with the first step

- can exist even with perfect sampling and feature selection

- Amazon’s HR dream tool 2015: given 100 résumés, spit out top five

- trained on resumes submitted in the previous 10 year, mostly from men, a reflected of male dominance across the tech industry

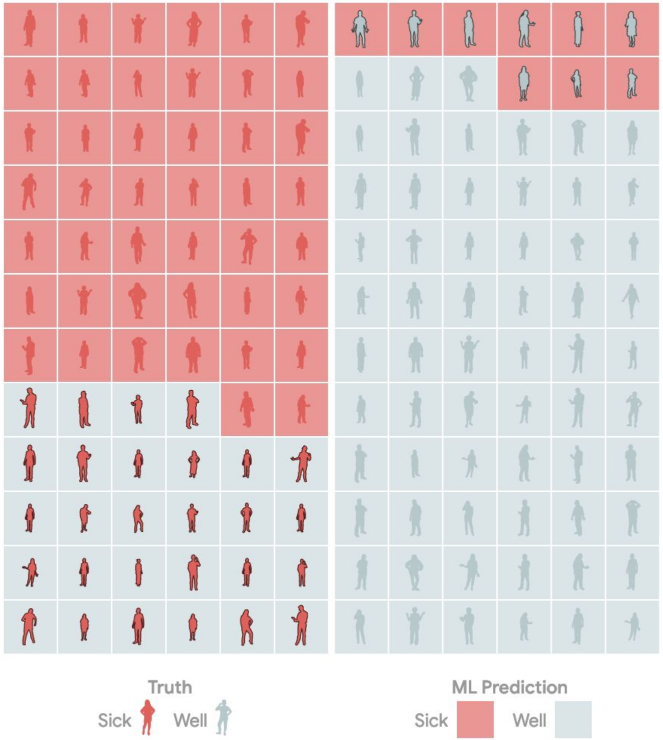
Fairness - it’s harder than you think

- no universal definition

- different ways to measure accuracy

- first problem: quantify/prioritize False Positives vs False Negative

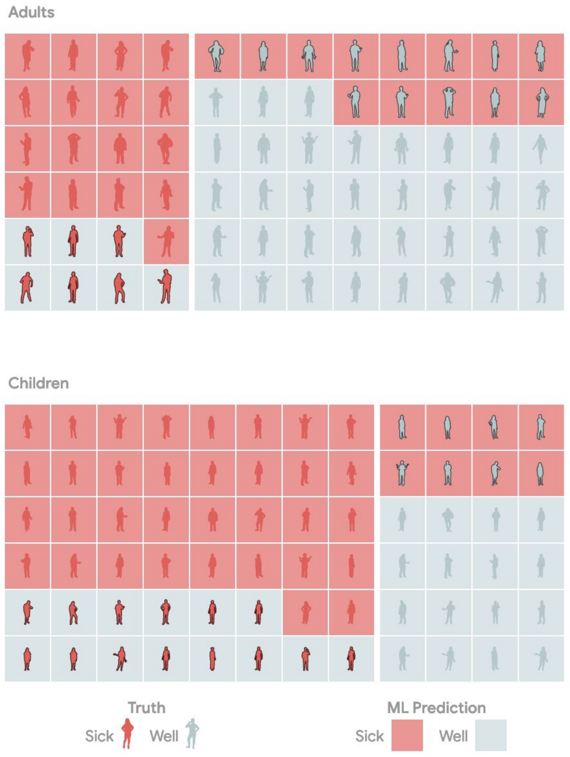
- frequency of screening, cost of follow-up test, treatment/drog supply, risk of delayed detection



Fairness - across groups

- differences/disparity between groups - different prevalence/base rate

- second problem: quantify/prioritize accuracy differences between groups



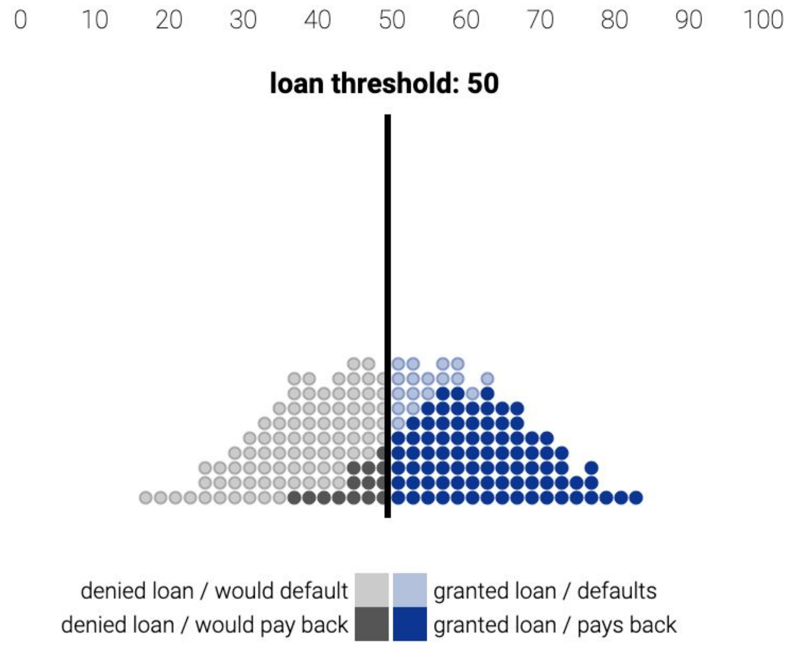
Selecting a fair threshold

- given a scoring system, is it possible to find a “fair” threshold?

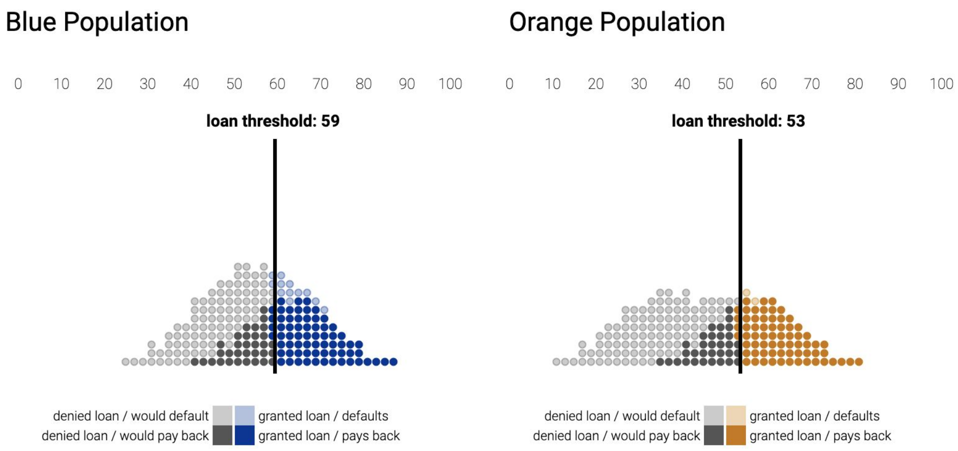
- loan application

- credit score - number of factors, income, payment history, promptness in paying debts

- decide who gets a loan and who doesn’t



Possible strategies



Maximize Profit
- different standards between groups

Group Unaware
- same threshold applied
- different percentages get loan

Demographic Parity
- same percentage of loans given (whole group focus)
- difference in loans paid back
- fewer qualified people given loans in one group

Equal opportunity
- same percentage of qualified people given loans
- qualified sub-groups in focus

Fairness - Conclusions

- it is harder than you think
- it may not be possible to satisfy every definition of fairness
- focus on the notions of fairness that make sense for your use case/context of your model
- rely on domain experts, use a multidisciplinary approach

References

- Google Health’s medical AI fails
- Fast.ai course - Bias and Fairness
- Article - Understanding source of bias
- PAIR - Measuring Fairness
- Google Research - Attacking discrimination