Fairness in Machine Learning

Fairness in Machine Learning

Q: What is **Fairness?**

Fairness in Machine Learning

Q: What is **Fairness?**

A: Understanding and mitigation of ***discrimination*** based on sensitive characteristics.

(Hardt Moritz, CS294, UC Berkley, 2017)

Discrimination

(according to the US labor law)

Discrimination

(according to the US labor law)

Disparate Treatment

Discrimination

(according to the US labor law)

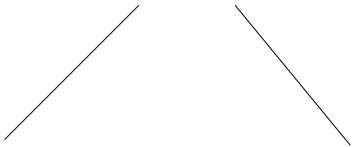
Disparate Treatment

- Formal
- Intentional

Discrimination

(according to the US labor law)

Disparate Treatment



Formal

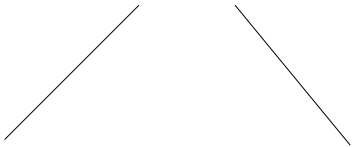
Intentional

‘age’ > 55

Discrimination

(according to the US labor law)

Disparate Treatment



Formal

Intentional

‘age’ > 55

‘experience, years’ > 30

Discrimination

(according to the US labor law)

Disparate Treatment

Formal

‘age’ > 55

Intentional

‘experience, years’ > 30

Disparate Impact

Discrimination

(according to the US labor law)

Disparate Treatment

Formal

‘age’ > 55

Intentional

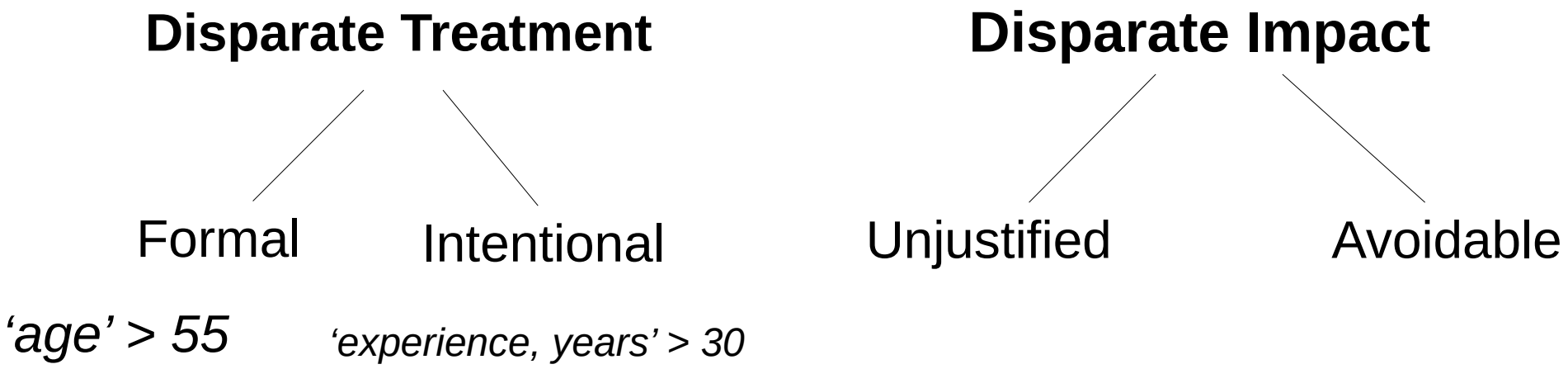
‘experience, years’ > 30

Disparate Impact

Unjustified

Discrimination

(according to the US labor law)



Discrimination

(according to the US labor law)

Disparate Treatment

- Formal: *‘age’ > 55*
- Intentional: *‘experience, years’ > 30*

Disparate Impact

- Unjustified
- Avoidable

PositiveRate(Class A) = 60%

PositiveRate(Protected Class B) = 30%

Concepts of Fairness

Concepts of Fairness

Independence (Demographic parity)

PositiveRate(Class1) = PositiveRate(Class2)

Concepts of Fairness

Independence (Demographic parity)

PositiveRate(Class1) = PositiveRate(Class2)

Separation (Equality of opportunity)

TruePositiveRate(Class1) = TruePositiveRate(Class2)

Concepts of Fairness

Independence (Demographic parity)

PositiveRate(Class1) = PositiveRate(Class2)

Separation (Equality of opportunity)

TruePositiveRate(Class1) = TruePositiveRate(Class2)

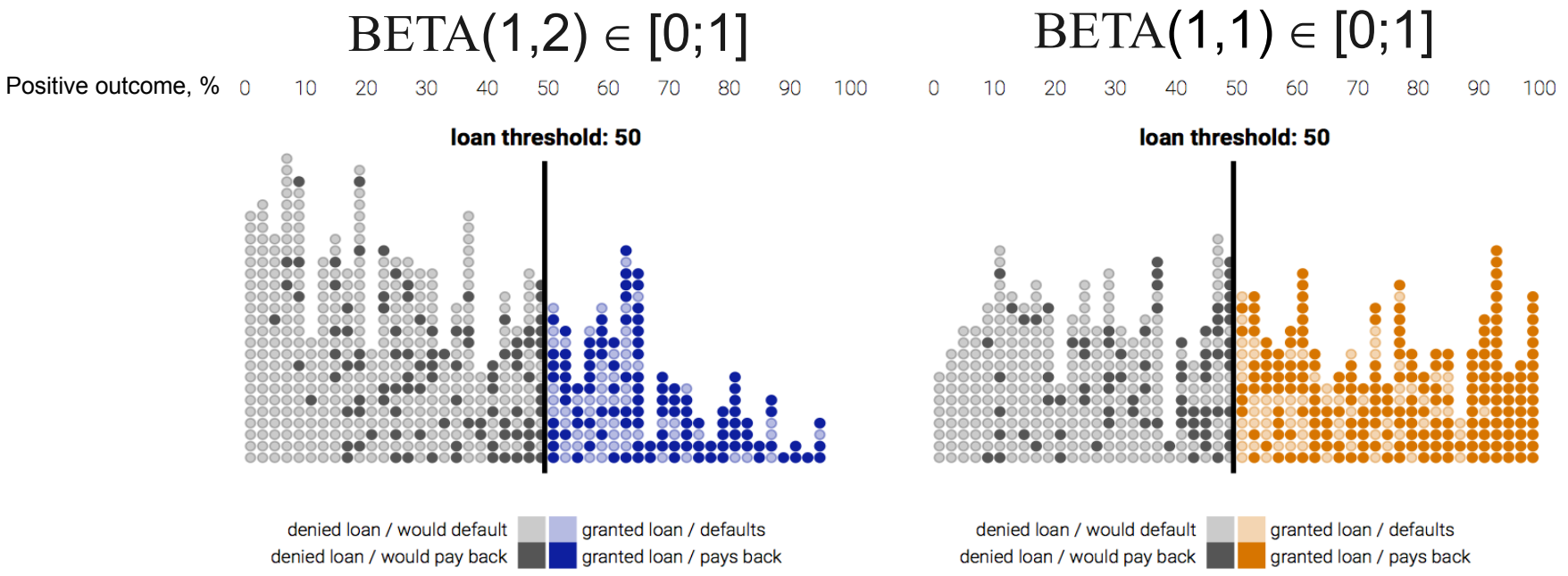
Sufficiency (Group unawareness)

Class membership is not considered

Simulating Loan Decisions

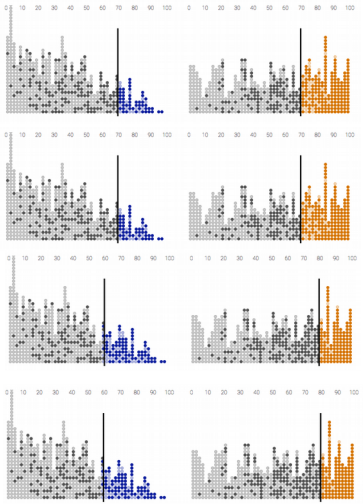
- 2 different groups
- Perfect predictor (Score = P)
- True positive → +3
- False positive → -7

- Thresholds are set to maximize utility
- Fairness constraints



Simulating Loan Decisions

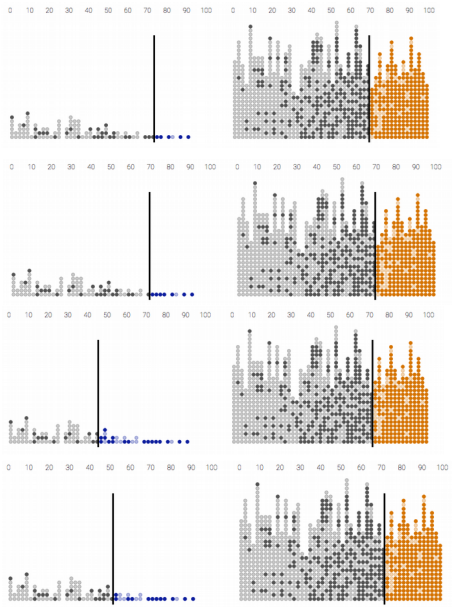
- Groups are of equal size (n=1000000)



	profit	total discrate	profit (X1, X2)		thresholds (X1, X2)		positiveRate (X1, X2)		TruePositiveRate (X1, X2)	
Maximize Profit	539674	0.598	89814	449860	70	69	0.091	0.311	0.217	0.525
Group Unaware	539633	0.607	89814	449819	**70**	**70**	0.091	0.301	0.217	0.511
Demographic Parity	415232	0.634	23910	391322	57	81	**0.186**	**0.191**	0.397	0.345
Equal Opportunity	445025	0.621	34715	410310	58	79	0.177	0.211	**0.382**	**0.377**

Simulating Loan Decisions

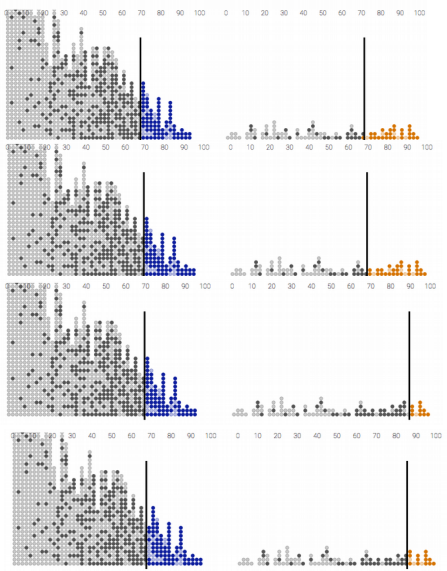
- Group 1 size = 10^5
- Group 2 size = 10^6



	profit	total discrate	profit (X1, X2)		thresholds (X1, X2)		positiveRate (X1, X2)		TruePositiveRate (X1, X2)	
Maximize Profit	461503	0.511	8985	452518	74	70	0.067	0.3	0.167	0.511
Group Unaware	461478	0.508	8960	452518	**70**	**70**	0.089	0.3	0.214	0.511
Demographic Parity	434527	0.514	-15536	450063	47	72	**0.281**	**0.28**	0.542	0.482
Equal Opportunity	443108	0.518	-6955	450063	51	72	0.24	0.28	**0.481**	**0.482**

Simulating Loan Decisions

- Group 1 size = 10^6
- Group 2 size = 10^5



	profit	total discrate	profit (X1, X2)		thresholds (X1, X2)		positiveRate (X1, X2)		TruePositiveRate (X1, X2)	
Maximize Profit	133095	0.748	88992	44103	70	71	0.09	0.291	0.216	0.494
Group Unaware	133045	0.746	88992	44053	**70**	**70**	0.09	0.301	0.216	0.508
Demographic Parity	114126	0.776	88903	25223	69	90	**0.096**	**0.101**	0.229	0.191
Equal Opportunity	118310	0.758	87605	30705	68	87	0.102	0.131	**0.241**	**0.244**

Simulating Loan Decisions

- Conclusion:
  - Discrimination is domain-specific
  - Outcome of a fairness concept application depends on the problem, group sizes, underlying distributions and other parameters.

Designing for Fairness (Google’s recommendations)

- Consider the problem
- Ask experts
- Train the models to account for bias
- Interpret outcomes
- Publish with context

https://developers.google.com/machine-learning/fairness-overview

https://developers.google.com/machine-learning/crash-course/fairness