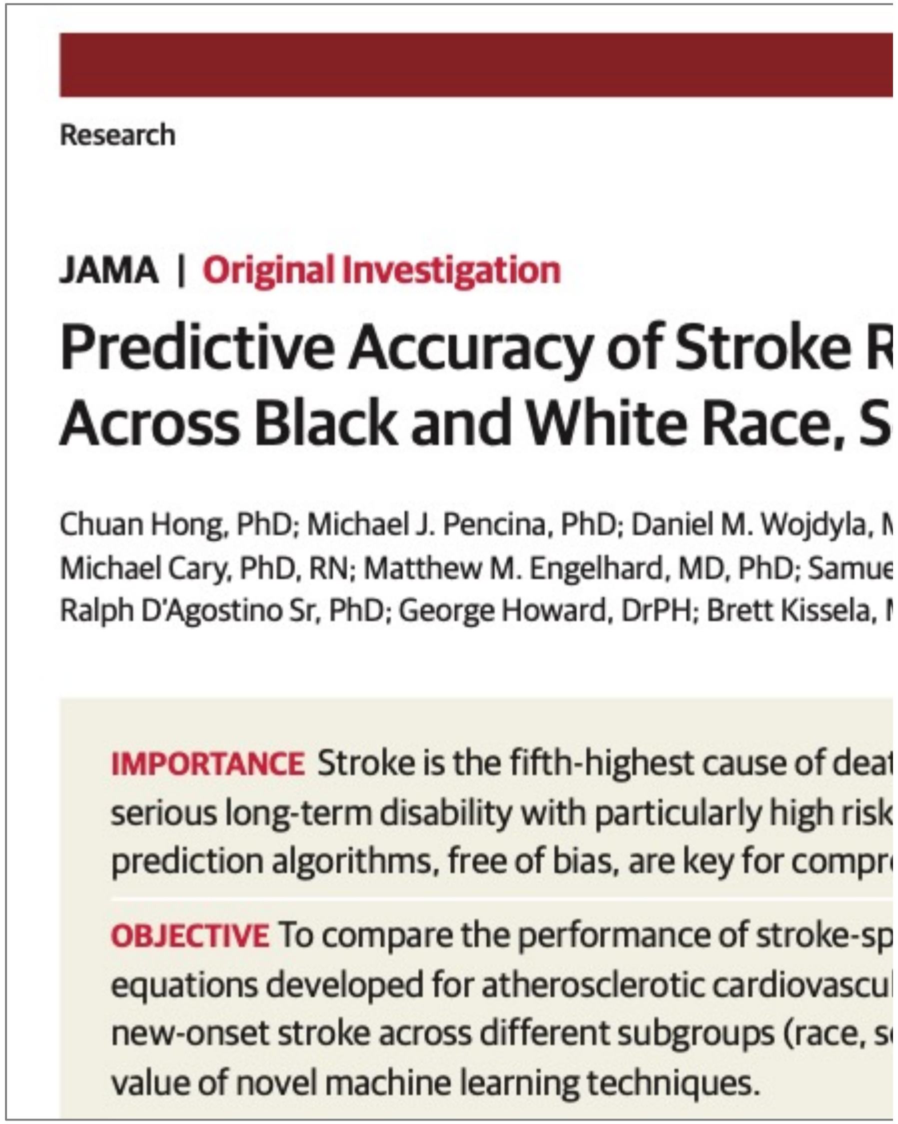
AI in Medicine I

Tutorial

Trustworthy AI: Fairness & Bias

Prof. Christian Wachinger
Lab for AI in Medical Imaging (www.ai-med.de)
Institute of Radiology, Klinikum rechts der Isar, TUM


Research

JAMA | Original Investigation

Predictive Accuracy of Stroke R
Across Black and White Race, S

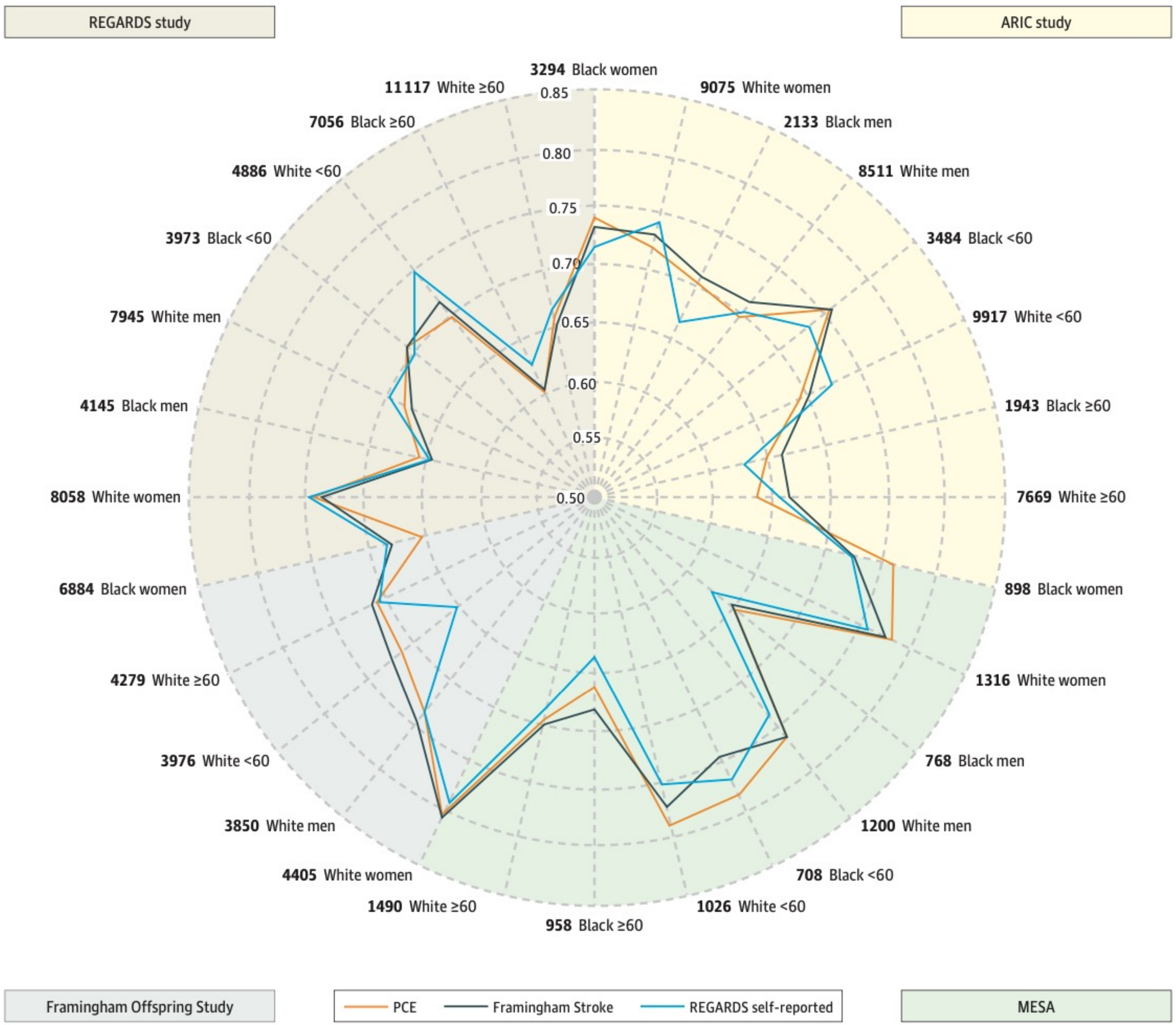
Chuan Hong, PhD; Michael J. Pencina, PhD; Daniel M. Wojdyla, M
Michael Cary, PhD, RN; Matthew M. Engelhard, MD, PhD; Samue
Ralph D'Agostino Sr, PhD; George Howard, DrPH; Brett Kissela, M

IMPORTANCE Stroke is the fifth-highest cause of deat
serious long-term disability with particularly high risk
prediction algorithms, free of bias, are key for compre

OBJECTIVE To compare the performance of stroke-sp
equations developed for atherosclerotic cardiovascul
new-onset stroke across different subgroups (race, se
value of novel machine learning techniques.

JAMA. 2023 Jan 24;329(4):306-317.
doi: 10.1001/jama.2022.24683

Figure. Comparison of C Index for Stroke Risk Prediction by Race, Sex, and Age



Sensitive Attribute *A*

Race:

- A social construct that categorizes individuals into groups based on physical characteristics
- Has changed over time and varies across cultures and societies
- Has no biological or genetic validity. Humans are a single species with a high degree of genetic diversity and variation, but no clear genetic boundaries between racial groups.
- Common racial categories: White, Black, Asian, Native American, Pacific Islander, and mixed-race

Ethnicity:

- A shared cultural heritage, language, nationality, religion that identify a particular group of people.
- An individual can be a member of multiple ethnic groups and may identify with different ethnicities at different times in their life
- Common ethnicities include: African, Arab, Asian, European, Hispanic, Jewish, …

Sensitive Attribute *A*

Sex:

- Biological and physiological characteristics that define males and females, including chromosomes, hormones, and reproductive anatomy
- Typically, people are classified as male or female at birth based on their anatomy and chromosomes
- Biological sex is not always clear-cut. Some individuals are born with intersex conditions

Gender:

- A social construct that refers to the culturally and socially defined roles that a society considers appropriate for men and women
- Understanding of gender as binary has been challenged, as many people identify as non-binary, gender non-conforming, or transgender. Gender as a spectrum.

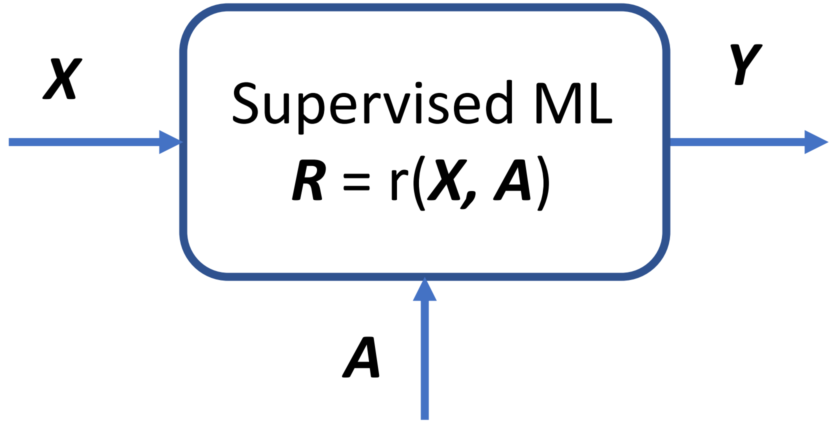
Agenda

1. Fairness criteria
2. Reweighing
3. Loan example
4. Coding: fairness

Summary of fairness criteria

Fairness	**Criteria**
Unawarness	Exclude $\boldsymbol{A}$ in prediction
Demographic parity	$P(\boldsymbol{R}=1\mid\boldsymbol{A}=1)=P(\boldsymbol{R}=1\mid\boldsymbol{A}=0)$
Equality of odds	$P(\boldsymbol{R}=1\mid\boldsymbol{A}=1,\boldsymbol{Y})=P(\boldsymbol{R}=1\mid\boldsymbol{A}=0,\boldsymbol{Y})$
Equal opportunity	$P(\boldsymbol{R}=1\mid\boldsymbol{A}=1,\boldsymbol{Y}=1)=P(\boldsymbol{R}=1\mid\boldsymbol{A}=0,\boldsymbol{Y}=\boldsymbol{1})$

Fairness Criteria



Which fairness criteria does R_1 satisfy?

A = Ethnicity, ***Y*** = Hired (1:yes, 0:no)

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	Yes	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Demographic parity for predictor R_1

$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Hisp}) =$

$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Whi}) =$

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Demographic parity for predictor R_1

$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Hisp}) = 2/3$

$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Whi}) = 2/3$

✅ Demographic parity

$P(\boldsymbol{R} = 1 | \boldsymbol{A} = 1) = P(\boldsymbol{R} = 1 | \boldsymbol{A} = 0)$

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Excercise

- Equality of odds ?
- Equal opportunity ?

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Equality of odds for predictor R_1

$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{yes}) =$

$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{yes}) =$

$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{no}) =$

$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{no}) =$

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Equality of odds for predictor R_1

$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{yes}) = 1$
$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{yes}) = 1/2$

$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{no}) =$
$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{no}) =$

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Equality of odds for predictor R_1

$$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{yes}) = 1$$
$$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{yes}) = 1/2$$
$$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{no}) = 0$$
$$P(\boldsymbol{R_1} = 1 | \boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{no}) = 1$$

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Equality of odds for predictor R_1

$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{yes}) = 1$$
$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{yes}) = 1/2$$
$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{no}) = 0$$
$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{no}) = 1$$

❌ Equality of odds

$$P(\boldsymbol{R} = 1|\boldsymbol{A} = 1, \boldsymbol{Y}) = P(\boldsymbol{R} = 1|\boldsymbol{A} = 0, \boldsymbol{Y})$$

Ethnicity *A*	Skills	Experience	Loves tacos	Hired? *Y*	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

Equal opportunity for predictor R_1

$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{yes}) = 1$$
$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{yes}) = 1/2$$
$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Hisp}, \boldsymbol{Y} = \text{no}) = 0$$
$$P(\boldsymbol{R_1} = 1|\boldsymbol{A} = \text{Whi}, \boldsymbol{Y} = \text{no}) = 1$$

❌ Equal opportunity

$$P(\boldsymbol{R} = 1|\boldsymbol{A} = 1, \boldsymbol{Y} = 1) = P(\boldsymbol{R} = 1|\boldsymbol{A} = 0, \boldsymbol{Y} = \mathbf{1})$$

Ethnicity A	Skills	Experience	Loves tacos	Hired? Y	Predictor R_1
Hispanic	Python	1	Yes	No	0
Hispanic	C++	5	Yes	Yes	1
Hispanic	Java	1	No	Yes	1
White	Java	2	No	Yes	0
White	C++	3	No	Yes	1
White	C++	0	No	No	1

2) Reweighing

Example: Loan

Expected: P(short) = P(tall) = 0.5
P(loan) = 0.3
P(no loan) = 0.7

P observed	**Short**	**Tall**
Loan	0.25	0.2
No loan	0.3	0.25

Please compute all 4 weights.

2) Reweighing

Example: Loan

Expected: P(short) = P(tall) = 0.5
P(loan) = 0.3
P(no loan) = 0.7

P observed	**Short**	**Tall**
Loan	0.25	0.2
No loan	0.3	0.25

Please compute all 4 weights.

Short loan:

$$w = \frac{0.5 \ * 0.3}{0.25} = 0.6$$

Short no-loan:

$$w = \frac{0.5 \ * 0.7}{0.3} = 1.17$$

Tall loan: 0.75

Tall no-loan: 1.4

0,5	0,3	0,25	0,6
0,5	0,7	0,3	1,166666667
0,5	0,3	0,2	0,75
0,5	0,7	0,25	1,4

Loan example

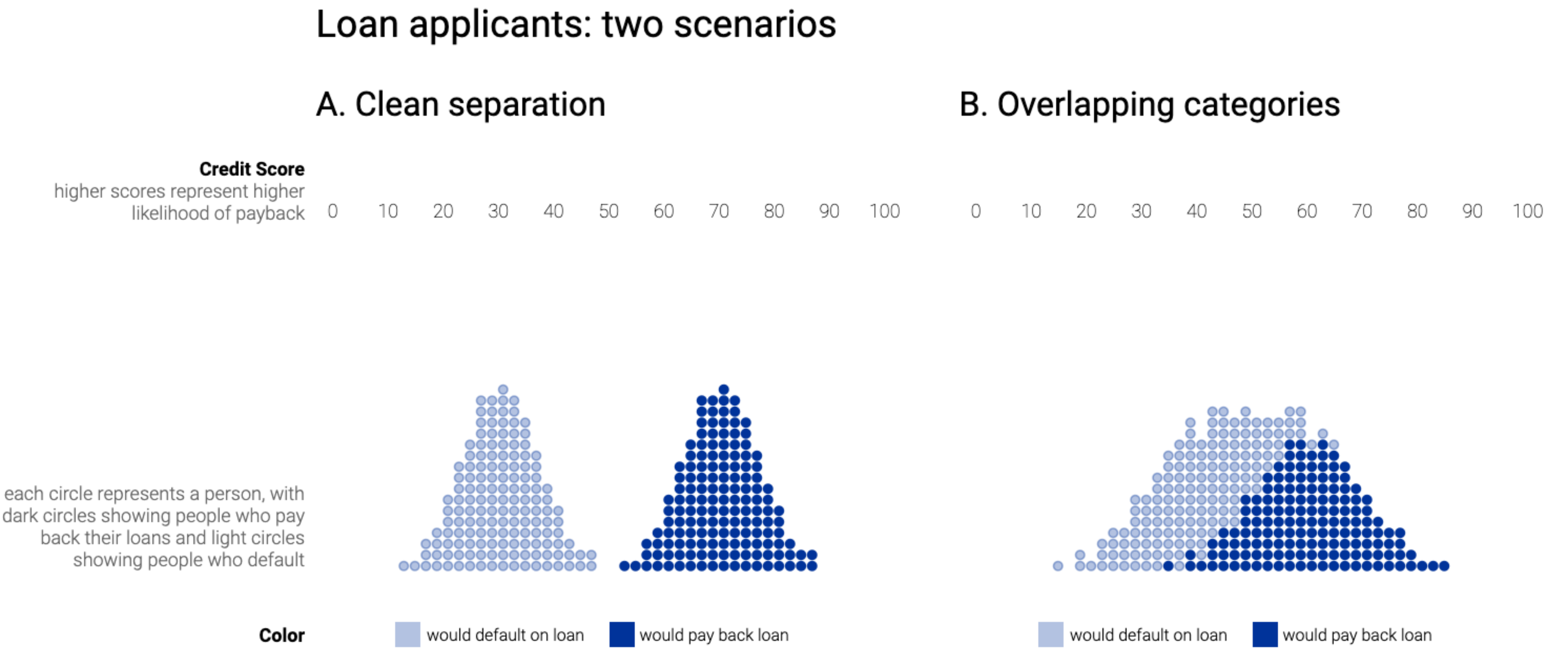
https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Simulating loan thresholds:

- Threshold with most correct decisions?
- Threshold that is most profitable?

Simulating loan decisions for different groups

- Which loan strategy would you choose and why?

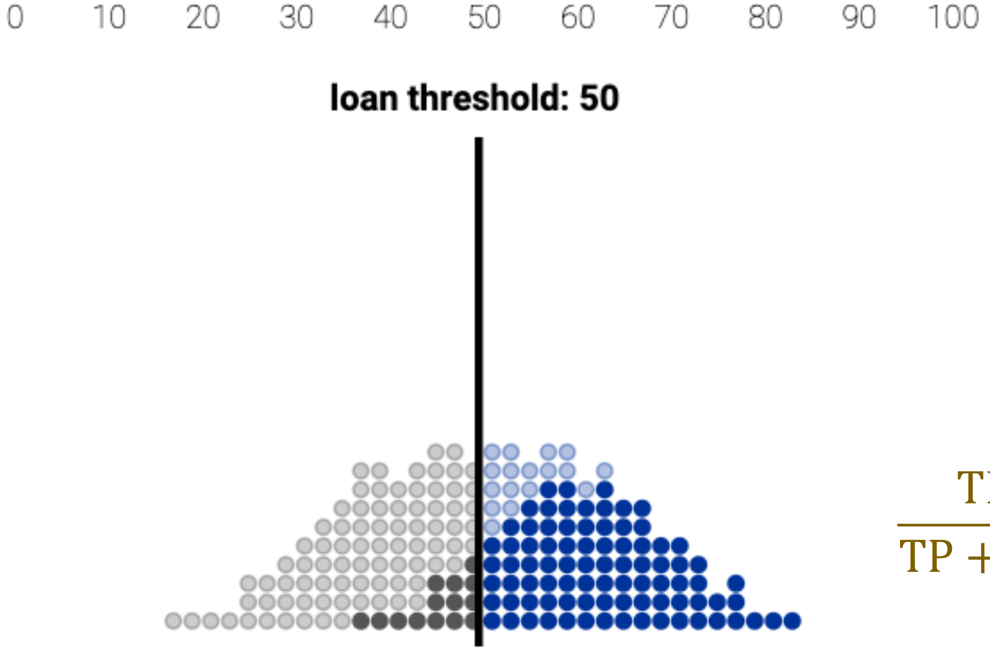

Loan applicants: two scenarios
A. Clean separation
B. Overlapping categories
Credit Score
higher scores represent higher likelihood of payback
0
10
20
30
40
50
60
70
80
90
100
each circle represents a person, with dark circles showing people who pay back their loans and light circles showing people who default
Color
would default on loan
would pay back loan
would default on loan
would pay back loan

Simulating loan thresholds

Drag the black threshold bars left or right to change the cut-offs for loans.

Threshold Decision

Credit Score
higher scores represent higher likelihood of payback



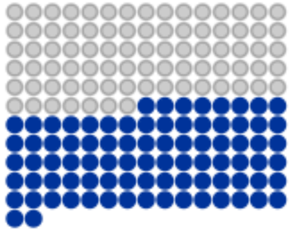
each circle represents a person, with dark circles showing people who pay back their loans and light circles showing people who default

Color

	TN	FP	
denied loan / would default			granted loan / defaults
denied loan / would pay back			granted loan / pays back
	FN	TP	

Outcome

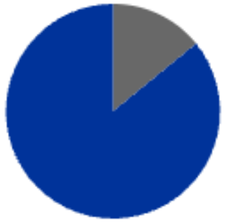
Correct 84%
loans granted to paying applicants and denied to defaulters



Incorrect 16%
loans denied to paying applicants and granted to defaulters

$$\frac{TP}{TP + FN}$$

True Positive Rate 86%
percentage of paying applications getting loans



Profit: **13600**

Positive Rate 52%
percentage of all applications getting loans

$$\frac{TP + FP}{\text{all}}$$

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

Same fractions blue / orange loans

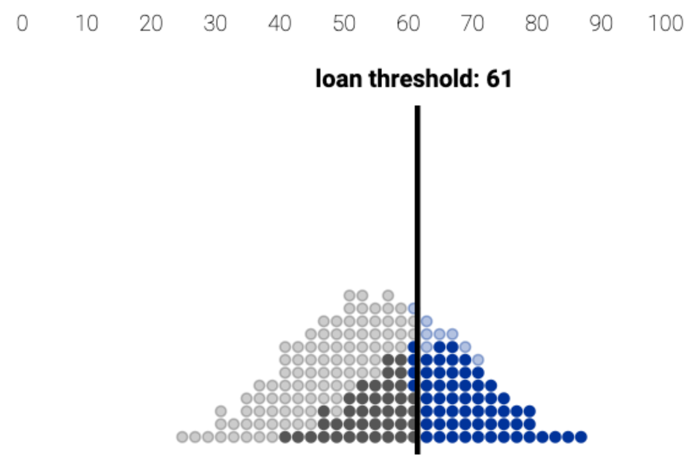
EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off

Max Profit

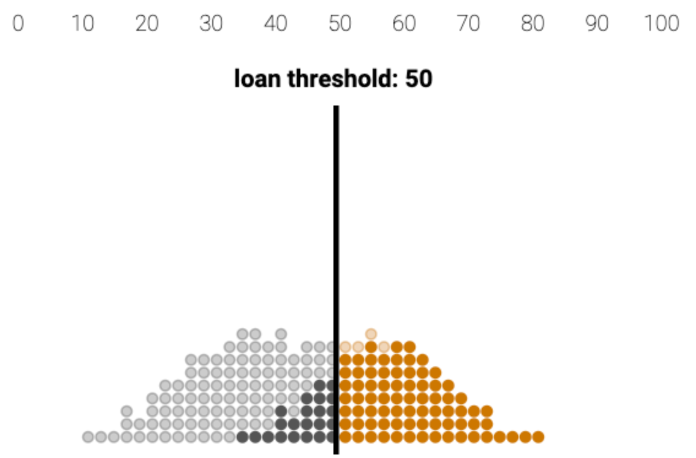
The most profitable, since there are no constraints. But the two groups have different thresholds, meaning they are held to different standards.

Blue Population



denied loan / would default | granted loan / defaults
denied loan / would pay back | granted loan / pays back

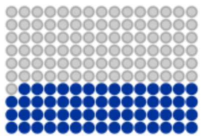
Orange Population



denied loan / would default | granted loan / defaults
denied loan / would pay back | granted loan / pays back

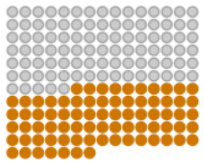
Total profit = 32400

Correct 76%
loans granted to paying applicants and denied to defaulters



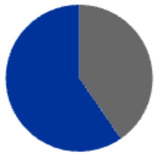
Incorrect 24%
loans denied to paying applicants and granted to defaulters

Correct 87%
loans granted to paying applicants and denied to defaulters

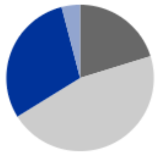


Incorrect 13%
loans denied to paying applicants and granted to defaulters

True Positive Rate 60%
percentage of paying applications getting loans

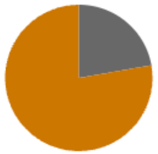


Positive Rate 34%
percentage of all applications getting loans



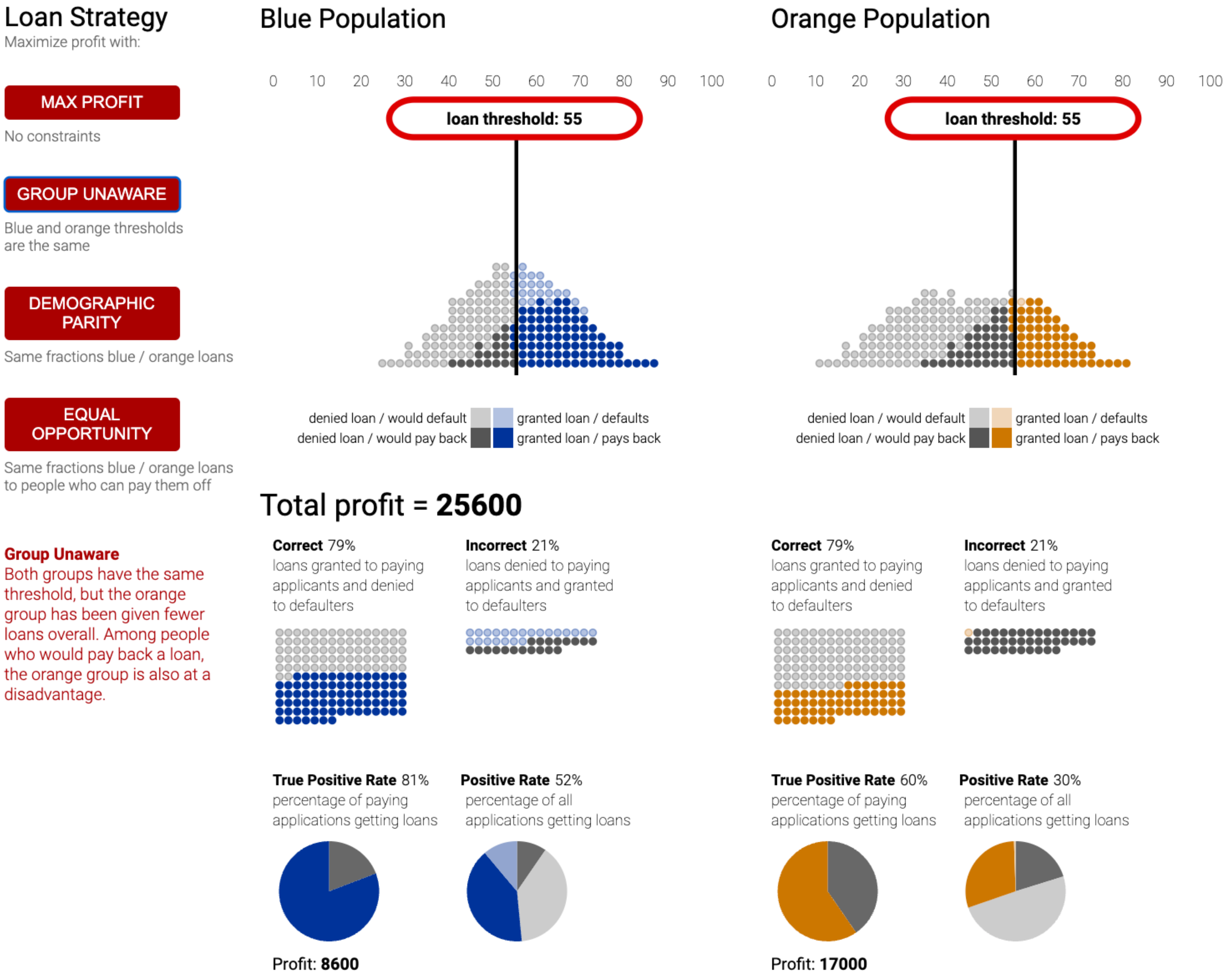
Profit: **12100**

True Positive Rate 78%
percentage of paying applications getting loans

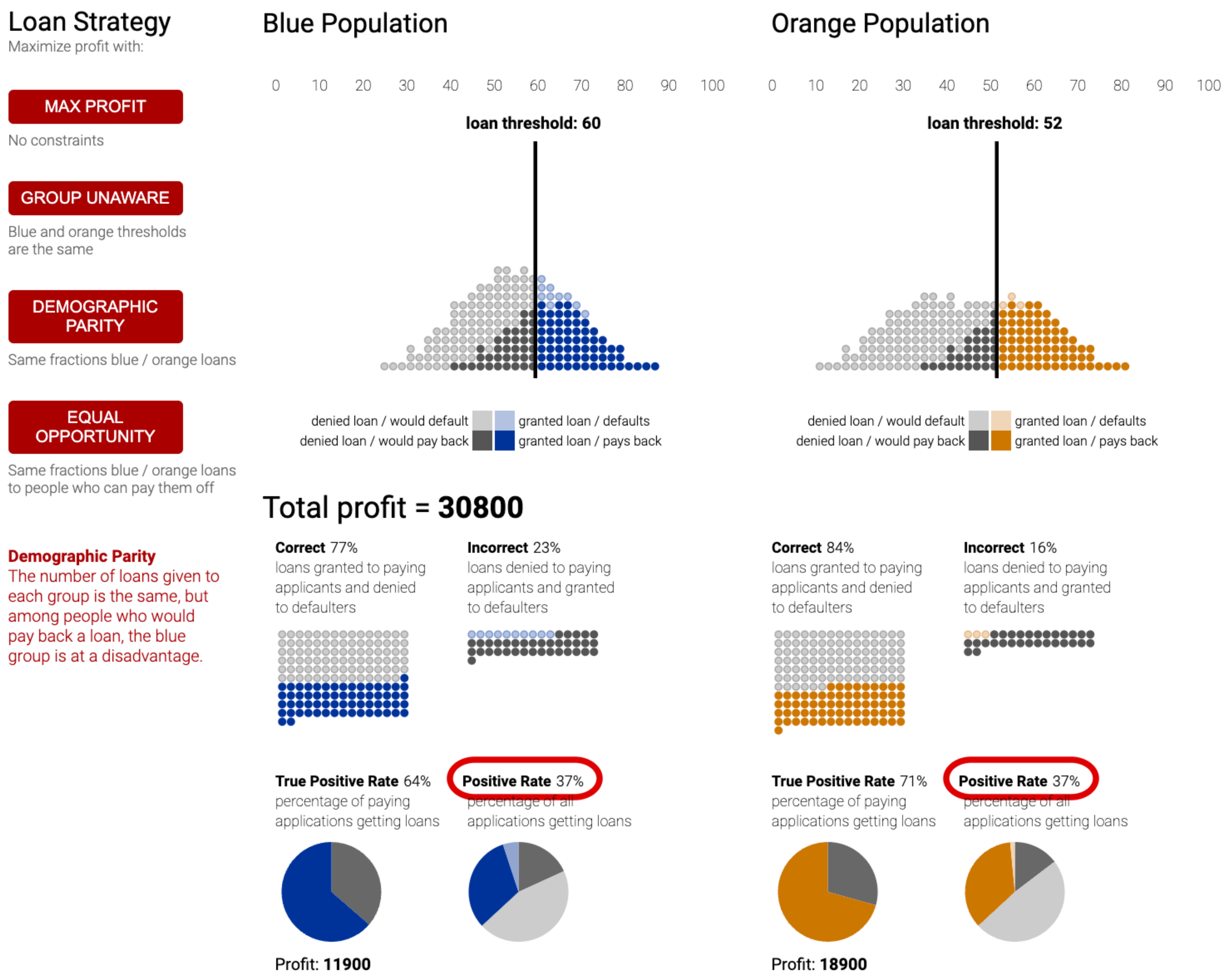


Positive Rate 41%
percentage of all applications getting loans

Profit: **20300**

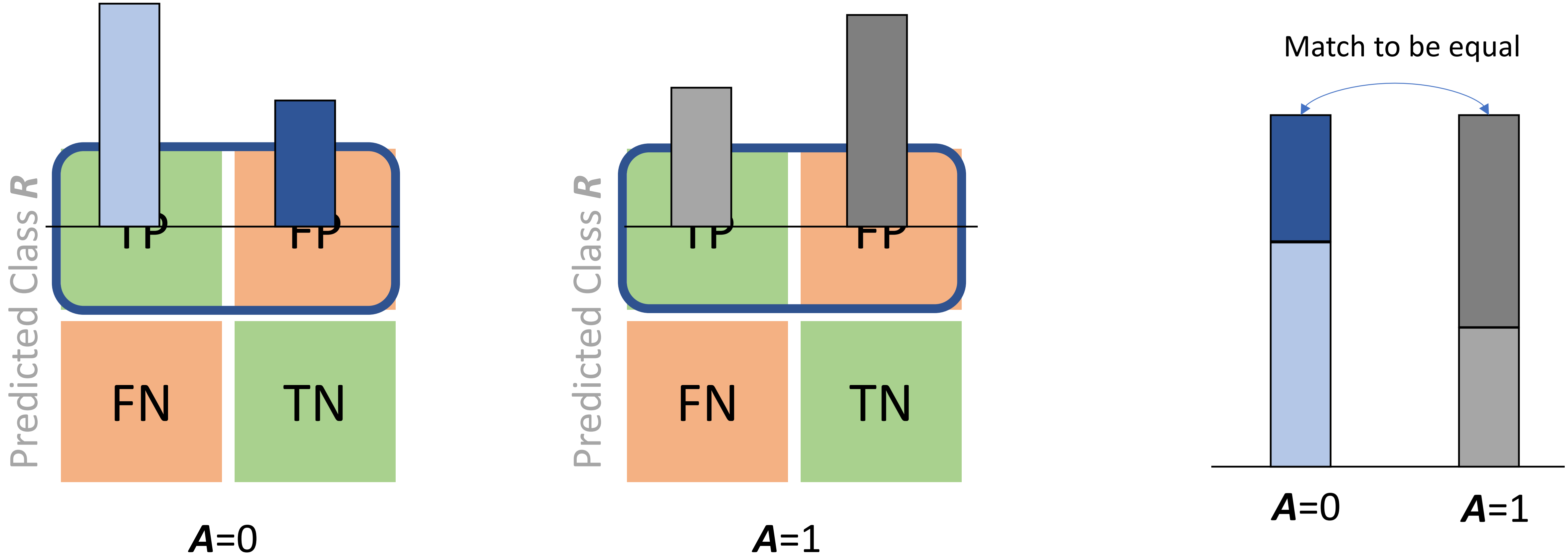


Loan Strategy
Maximize profit with:
MAX PROFIT
No constraints
GROUP UNAWARE
Blue and orange thresholds are the same
DEMOGRAPHIC PARITY
Same fractions blue / orange loans
EQUAL OPPORTUNITY
Same fractions blue / orange loans to people who can pay them off
Group Unaware
Both groups have the same threshold, but the orange group has been given fewer loans overall. Among people who would pay back a loan, the orange group is also at a disadvantage.
Blue Population
0 10 20 30 40 50 60 70 80 90 100
loan threshold: 55
denied loan / would default
denied loan / would pay back
granted loan / defaults
granted loan / pays back
Orange Population
0 10 20 30 40 50 60 70 80 90 100
loan threshold: 55
denied loan / would default
denied loan / would pay back
granted loan / defaults
granted loan / pays back
Total profit = 25600
Correct 79%
loans granted to paying applicants and denied to defaulters
Incorrect 21%
loans denied to paying applicants and granted to defaulters
True Positive Rate 81%
percentage of paying applications getting loans
Positive Rate 52%
percentage of all applications getting loans
Profit: 8600
Correct 79%
loans granted to paying applicants and denied to defaulters
Incorrect 21%
loans denied to paying applicants and granted to defaulters
True Positive Rate 60%
percentage of paying applications getting loans
Positive Rate 30%
percentage of all applications getting loans
Profit: 17000


Loan Strategy
Maximize profit with:
MAX PROFIT
No constraints
GROUP UNAWARE
Blue and orange thresholds are the same
DEMOGRAPHIC PARITY
Same fractions blue / orange loans
EQUAL OPPORTUNITY
Same fractions blue / orange loans to people who can pay them off
Demographic Parity
The number of loans given to each group is the same, but among people who would pay back a loan, the blue group is at a disadvantage.
Blue Population
0 10 20 30 40 50 60 70 80 90 100
loan threshold: 60
denied loan / would default
denied loan / would pay back
granted loan / defaults
granted loan / pays back
Orange Population
0 10 20 30 40 50 60 70 80 90 100
loan threshold: 52
denied loan / would default
denied loan / would pay back
granted loan / defaults
granted loan / pays back
Total profit = 30800
Correct 77%
loans granted to paying applicants and denied to defaulters
Incorrect 23%
loans denied to paying applicants and granted to defaulters
True Positive Rate 64%
percentage of paying applications getting loans
Positive Rate 37%
percentage of all applications getting loans
Profit: 11900
Correct 84%
loans granted to paying applicants and denied to defaulters
Incorrect 16%
loans denied to paying applicants and granted to defaulters
True Positive Rate 71%
percentage of paying applications getting loans
Positive Rate 37%
percentage of all applications getting loans
Profit: 18900

Independence (demographic parity)

$\boldsymbol{R} \perp \boldsymbol{A}$: $\quad P(\boldsymbol{R}=1|\boldsymbol{A}=1) = P(\boldsymbol{R}=1|\boldsymbol{A}=0)$ $\quad$ Positive rate is the same for both groups



Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

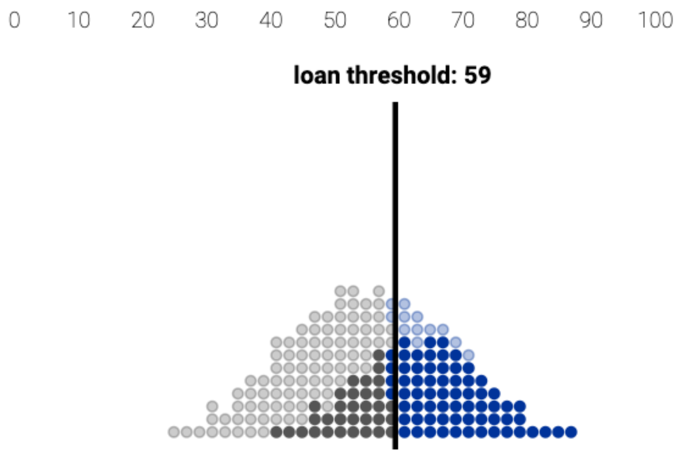
Same fractions blue / orange loans

EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off

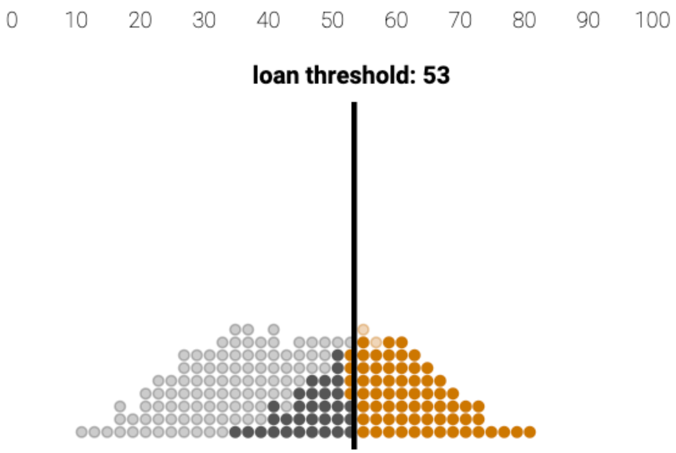
Equal Opportunity
Among people who would pay back a loan, blue and orange groups do equally well. This choice is almost as profitable as demographic parity, and about as many people get loans overall.

Blue Population



denied loan / would default — granted loan / defaults
denied loan / would pay back — granted loan / pays back

Orange Population

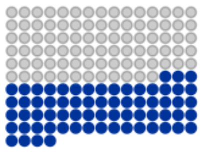


denied loan / would default — granted loan / defaults
denied loan / would pay back — granted loan / pays back

Total profit = **30400**

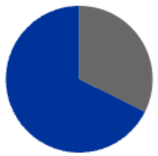
Blue

Correct 78%
loans granted to paying applicants and denied to defaulters



Incorrect 22%
loans denied to paying applicants and granted to defaulters

True Positive Rate 68%
percentage of paying applications getting loans

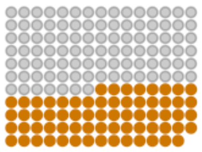


Positive Rate 40%
percentage of all applications getting loans

Profit: **11700**

Orange

Correct 83%
loans granted to paying applicants and denied to defaulters



Incorrect 17%
loans denied to paying applicants and granted to defaulters

True Positive Rate 68%
percentage of paying applications getting loans

Positive Rate 35%
percentage of all applications getting loans

Profit: **18700**

CODING ☺

Open the Notebook:

https://colab.research.google.com/drive/1Se_QrtIheSdXB-T02hj24ABEMXmJyU31?usp=sharing

Coding tasks

1. Study the correlation of features. Do you see something that is interesting or potentially problematic?

2. Implement “Fairness through Unawareness”

3. Discuss the results of the different mitigation strategies wrt:
   - Prediction accuracy
   - Statistical parity / equal opportunity

Coding tasks

Implementation for “Fairness through Unawareness”:

```
lred = LogisticRegression(solver='liblinear')
X_train_red = X_train.drop(['sex_Female','sex_Male','race_African-American','race_Caucasian'],axis=1)
X_test_red = X_test.drop(['sex_Female','sex_Male','race_African-American','race_Caucasian'],axis=1)
lred.fit(X_train_red, y_train)
y_pred_lred = lred.predict(X_test_red)
```