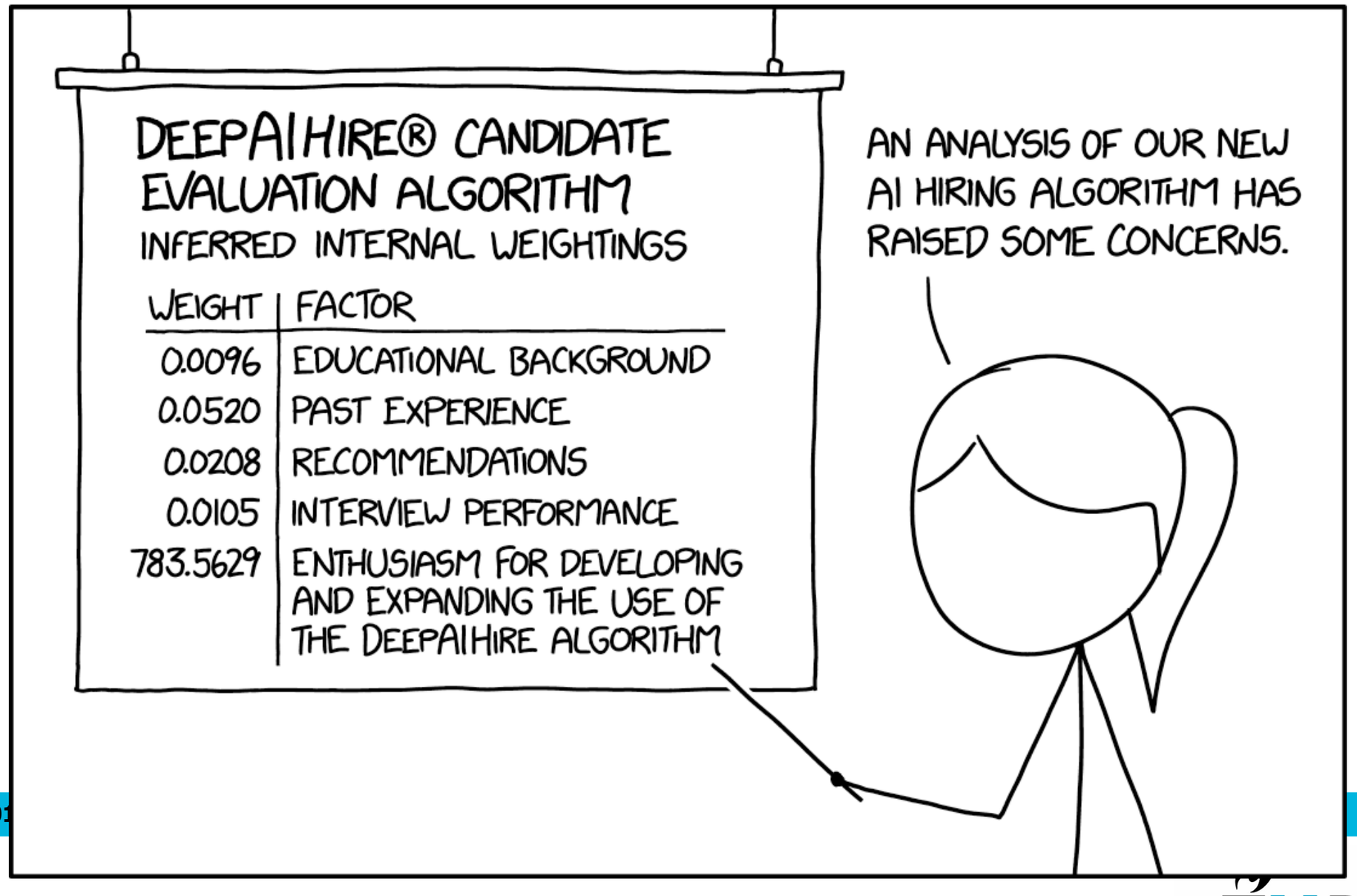
Fairness in ML



Contents

- What means fairness in ML?
- Definitions of fairness:
  - Indendence
  - Separation
  - Sufficiency
- How to enforce fairness?
- Open issues

Job platform biased rating

- Males consistently rated higher than females:

Search Query	Work Experience	Education Experience	Candidate	Xing Ranking
Brand Strategist	146	57	male	1
Brand Strategist	327	0	female	2
Brand Strategist	502	74	male	3
Brand Strategist	444	56	female	4
Brand Strategist	139	25	male	5
Brand Strategist	110	65	female	6
Brand Strategist	12	73	male	7
Brand Strategist	99	41	male	8
Brand Strategist	42	51	female	9
Brand Strategist	220	102	female	10
		. . .		
Brand Strategist	3	107	female	20
Brand Strategist	123	56	female	30
Brand Strategist	3	3	male	40

TABLE I: Top k results on www.xing.com (Jan 2017) for an employer's job search query "Brand Strategist".

Bias in your data

- Some function $\hat{y} = f(\mathbf{x})$ is fitted to some training data:

$$\mathcal{X} = \{(\mathbf{x}_i, y_i); i = 1, ..., N\}$$

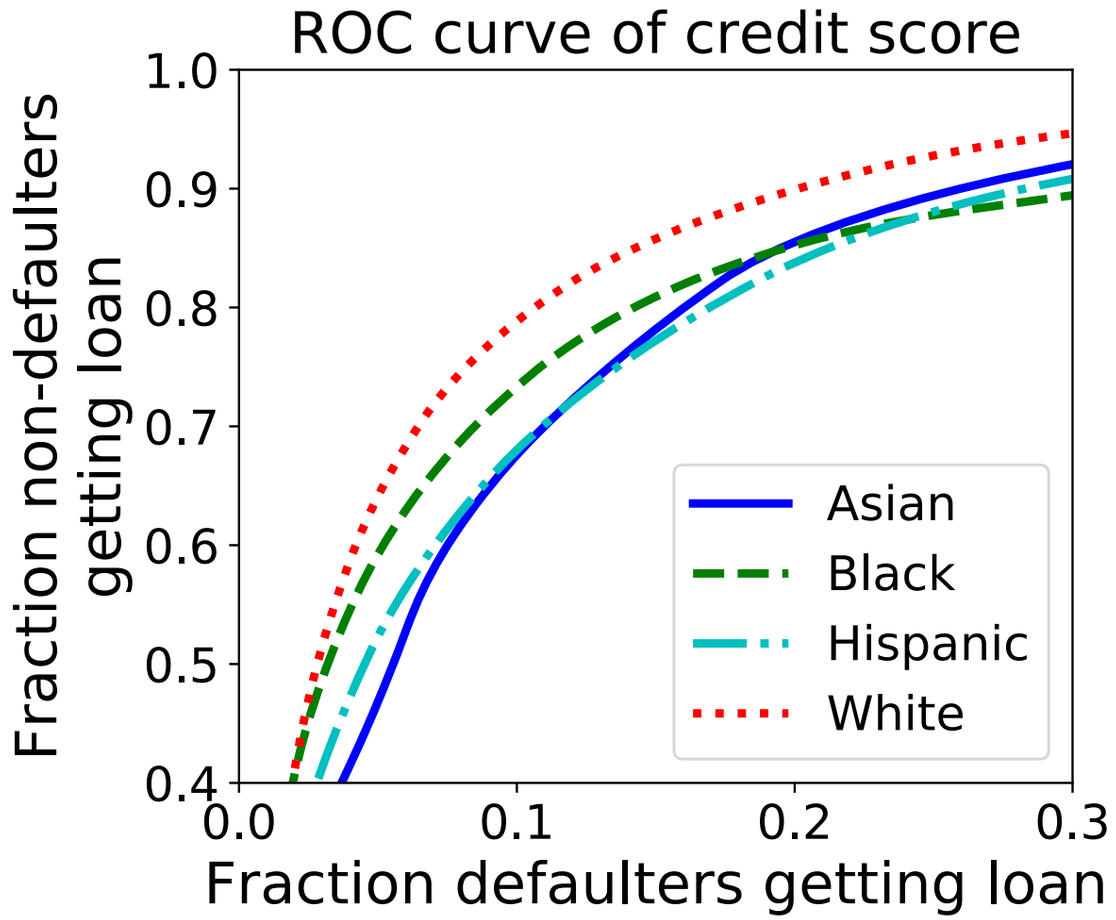
- Training data comes from somewhere (historical? anyhow selected)
- Therefore it often ‘inherits’ the biases of humans

- Assume, we have some attribute (age, gender, race, …) that we don’t want to be biased for. Call it:

attribute A

Examples of attribute A

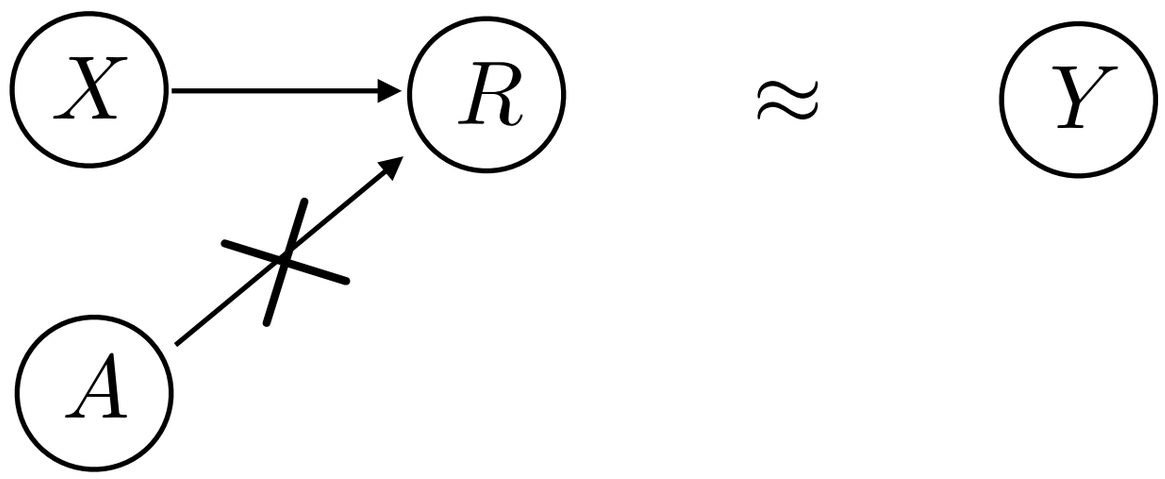
- For hiring people: use age/gender
- Predict wealth: use DNA
- Detect pot holes in roads: use photo's made by citizens



- Evaluate loan application/ mortgage: use address
- …

Target, prediction, sensitive attrib.

- Y: true target
- R: prediction $R = f(X)$
- A: sensitive attribute we should be insensitive to



main source: `fairmlbook.org`

My intuition/solution

- Have intermediate representation X’:

$$X \rightarrow X' \rightarrow R \quad \approx \quad Y$$

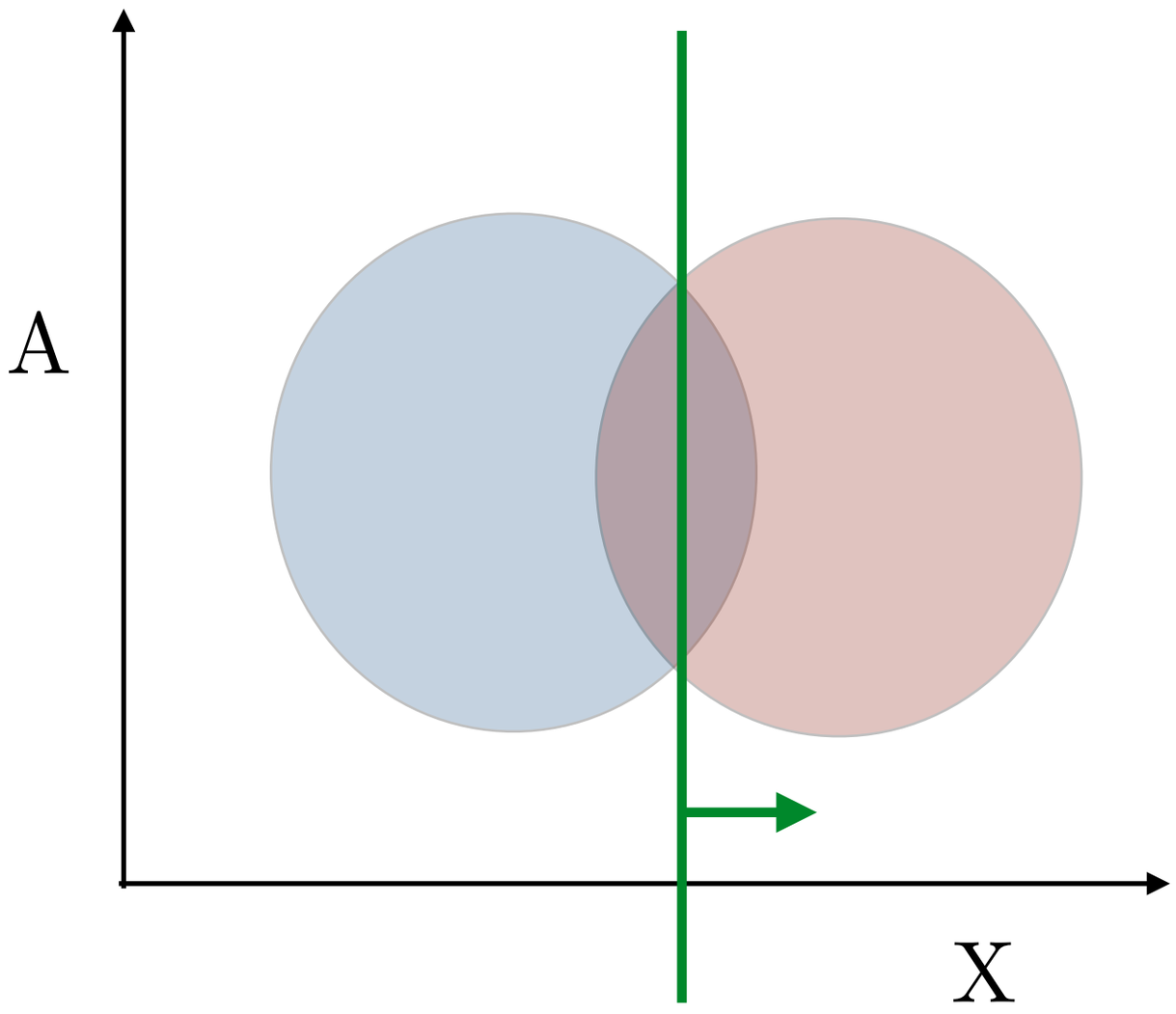
- where the mutual information between X’ and A is 0:

$$I(X';A) = 0$$

$$I(X';A) = D_{KL}(P_{X'A}||P_{X'}P_A) = 0$$

$$\sum_{x'}\sum_{a} P(x',a)\log\frac{P(x',a)}{P(x')P(a)} = 0$$

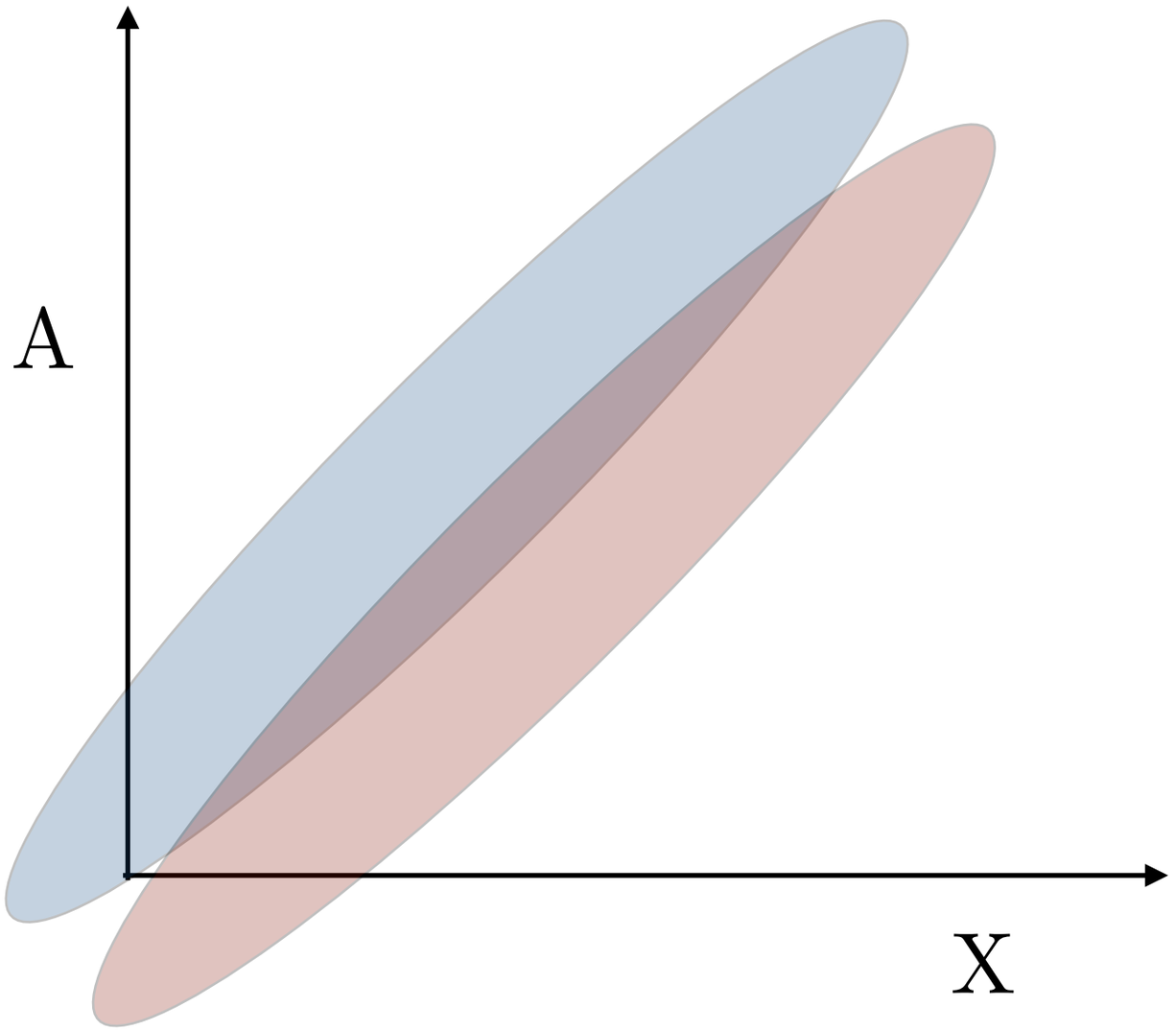
Independence

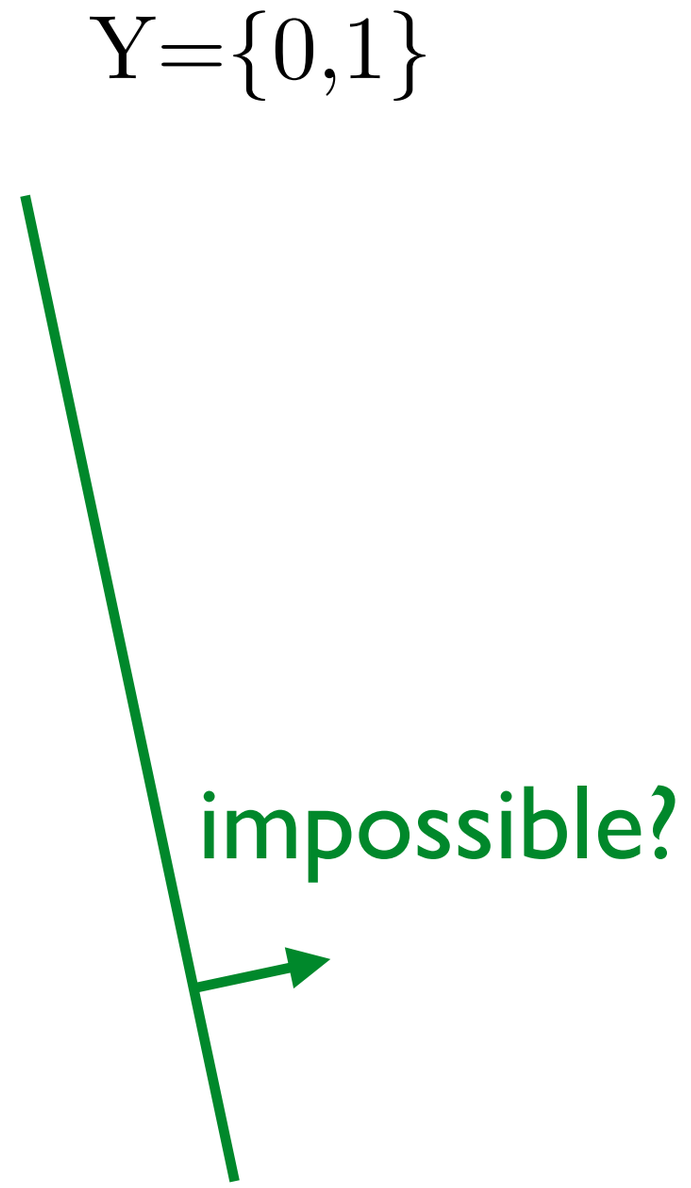


$Y=\{0,1\}$

- A is not informative for prediction R: no issue!

Independence

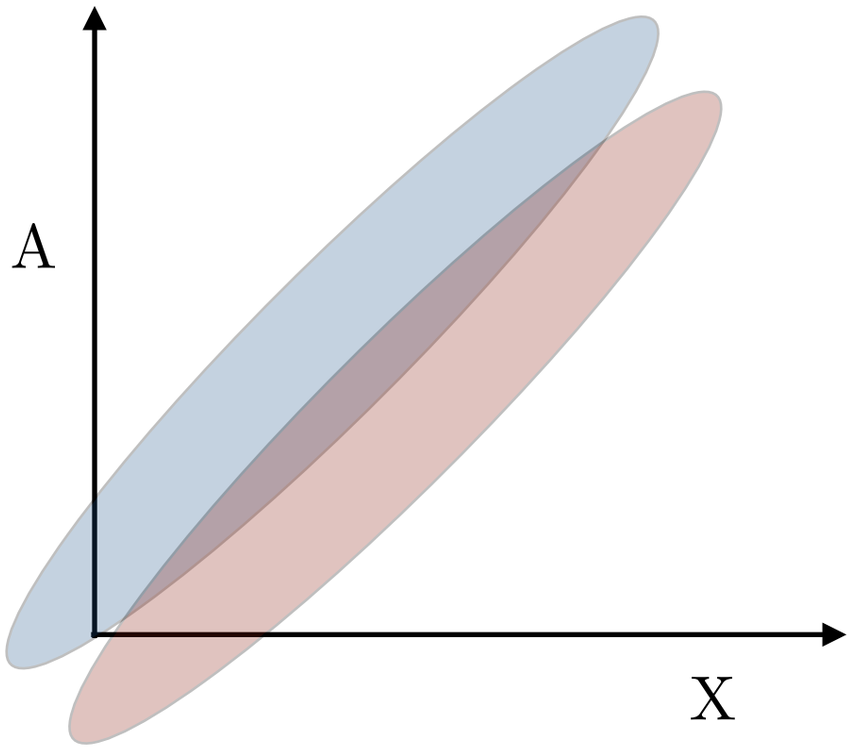


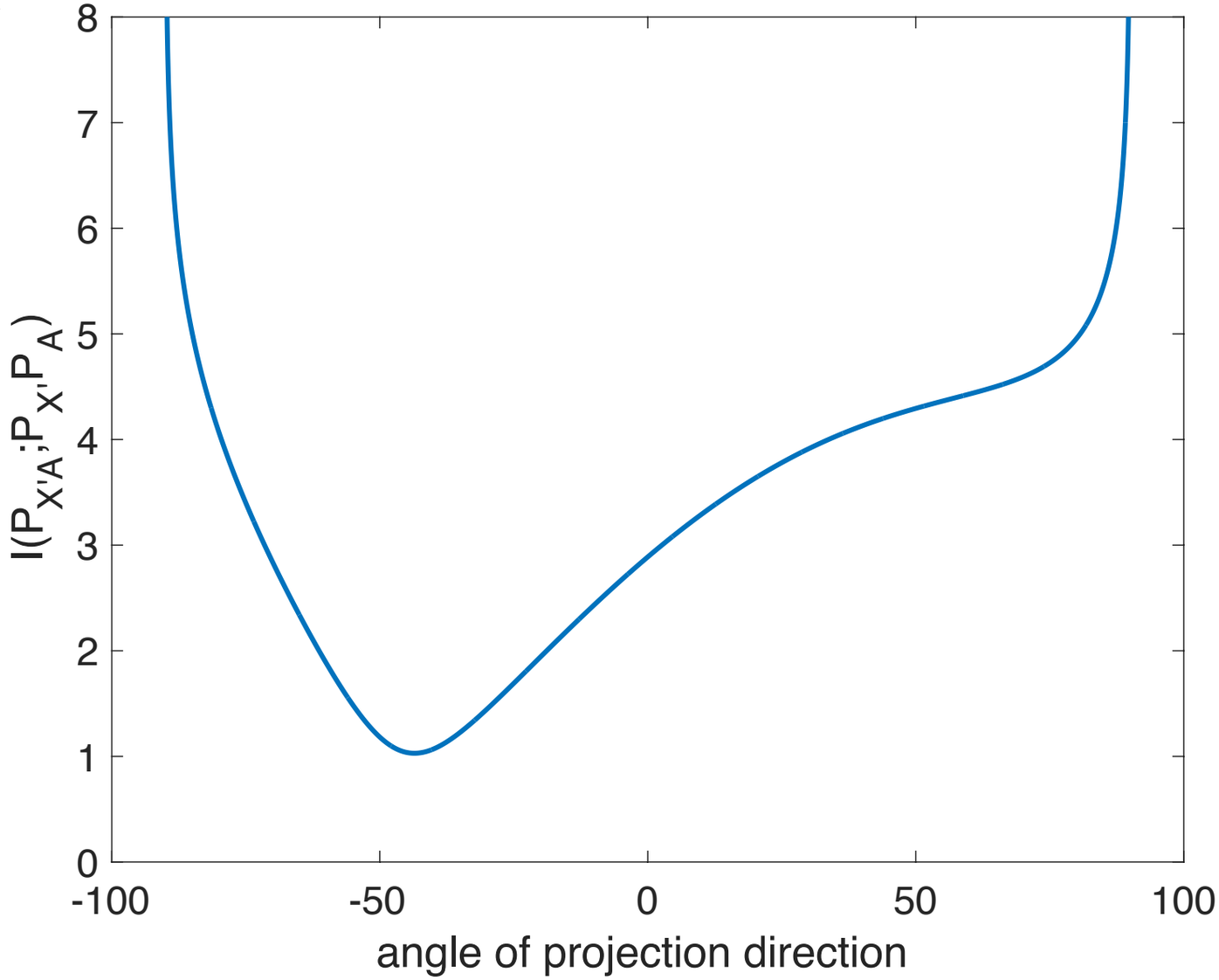


- A is correlated with X: how to define R?

Independence

$$Y=\{0,1\}$$

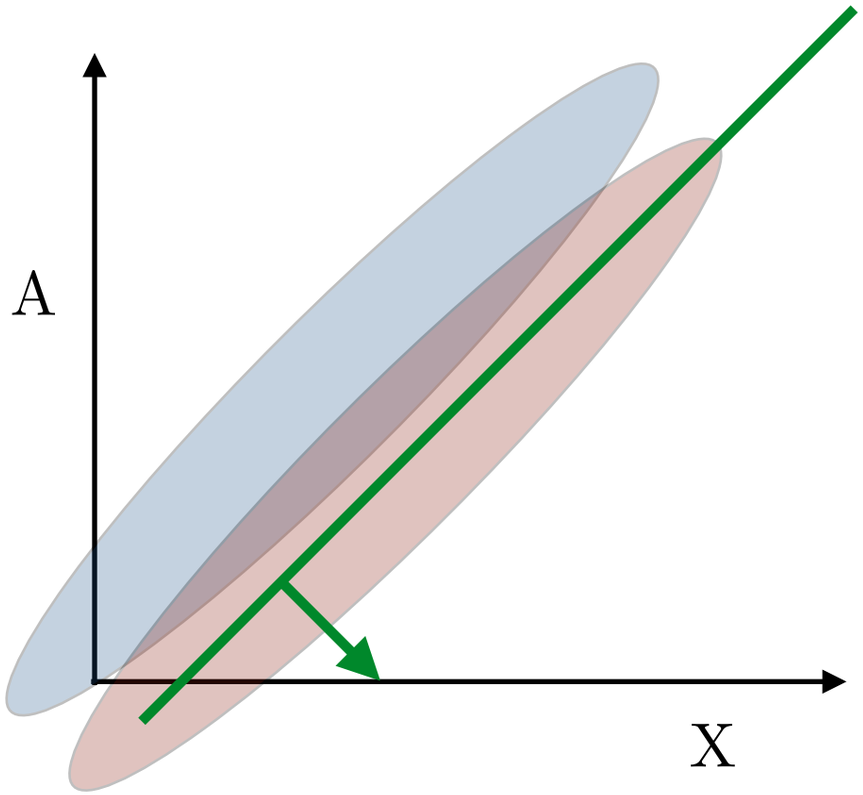


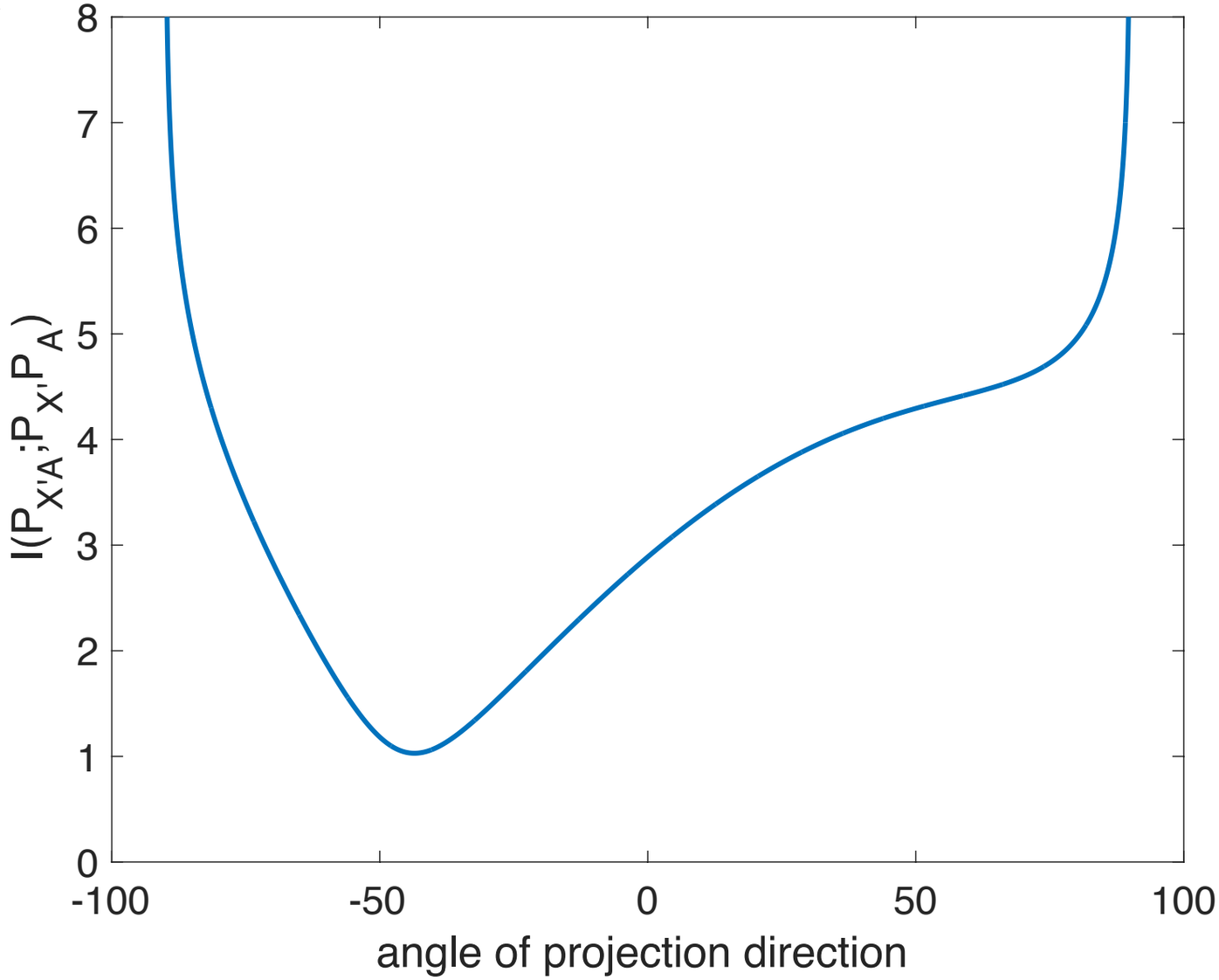


- Optimal R is **not** independent of A!

Independence

$Y=\{0,1\}$

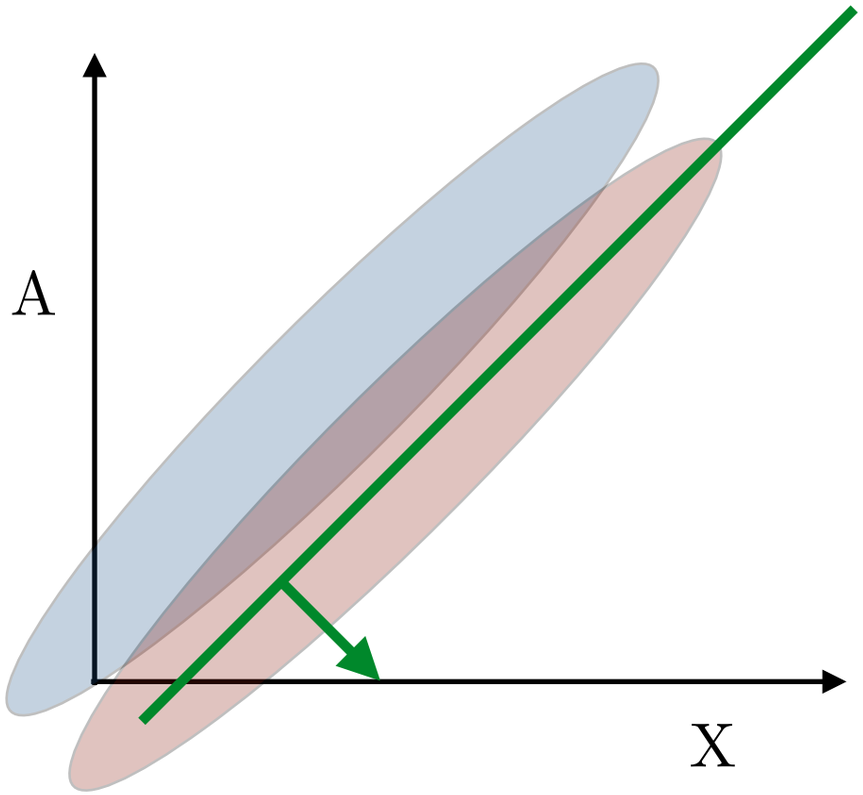


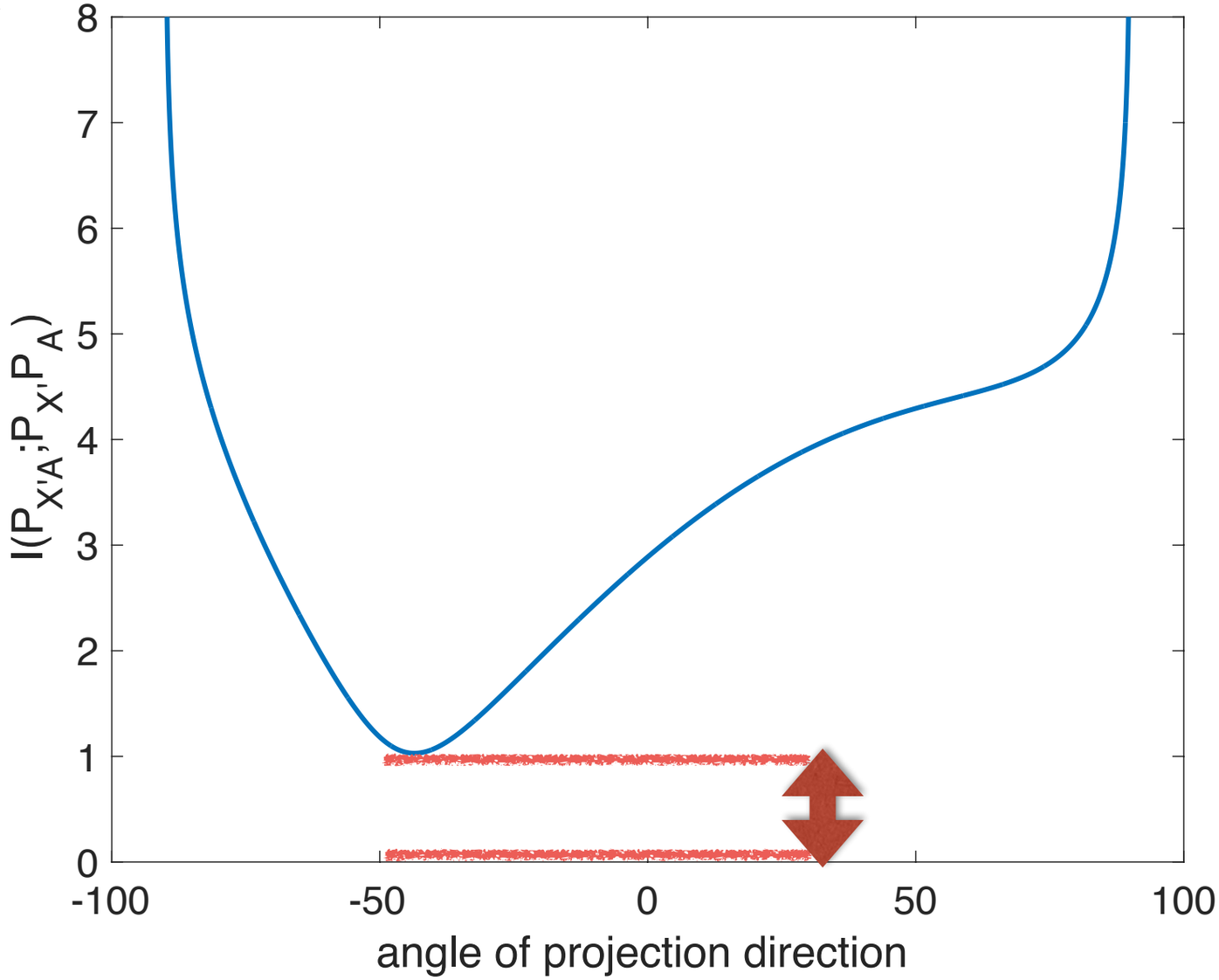


- Optimal R is **not** independent of A!

Independence

$$Y=\{0,1\}$$





- Optimal R is **not** independent of A!

Maybe counterintuitive finding

- In order to check how fair a method is, you NEED the sensitive attribute A!

- Often datasets are ‘too much polluted’, that no fair classifier can be created

- Need other ‘intuitive’/approximate measures

Criteria from literature

Table 6: List of demographic fairness criteria

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Non-discrimination criteria

- Three possible criteria are introduced:
  1. Independence
  2. Separation
  3. Sufficiency

- **Independence**: $R \perp A$

$$P(R|A) = P(R)$$

- Consequence: $\mathbb{P}\{R = 1 \mid A = a\} = \mathbb{P}\{R = 1 \mid A = b\}$

Non-discrimination criteria

- Three possible criteria are introduced:
  1. Independence
  2. Separation
  3. Sufficiency

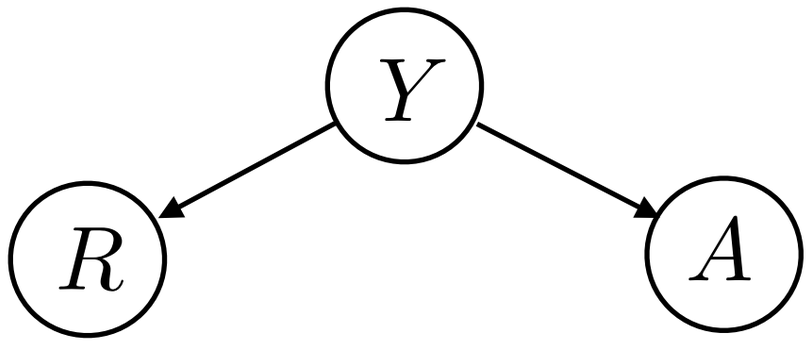
- **Independence**: $R \perp A$

$$P(R|A) = P(R)$$

- Consequence: $\mathbb{P}\{R = 1 \mid A = a\} = \mathbb{P}\{R = 1 \mid A = b\}$
- Problem: when A is informative for Y, we probably loose much performance!

Separation

- Defined as $R \perp A|Y$: given Y, R and A are independent



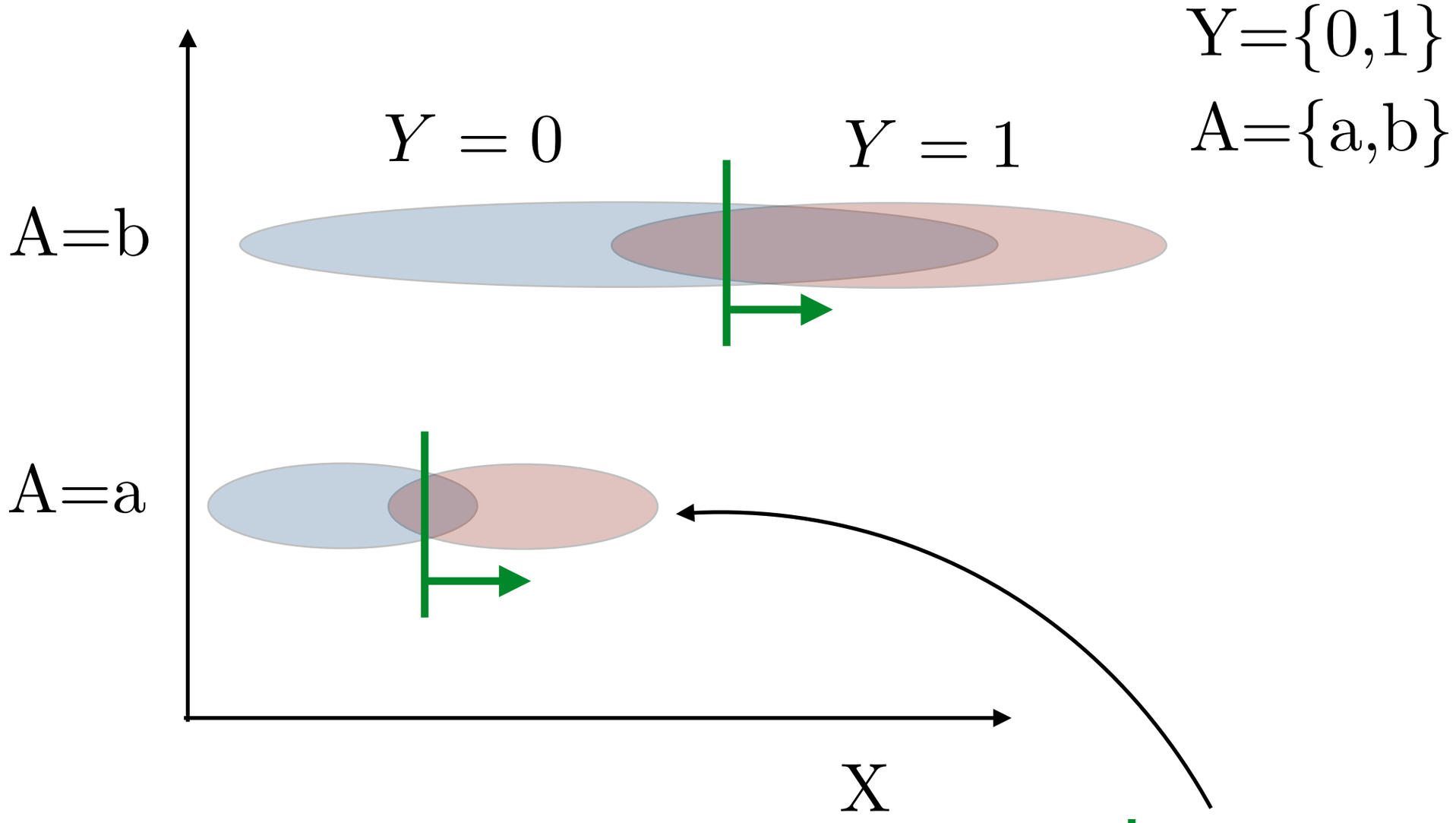
$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}.$$

- For binary variables:
  true positive rate group a=true positive rate group b
  false positive rate group a=false positive rate group b

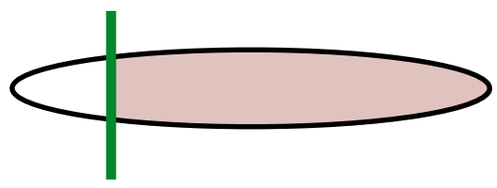
Confusion matrix

		predicted positive	predicted negative		
truth	positive	TP	FN	$\rightarrow \frac{TP}{TP+FN}$	true positive rate
	negative	FP	TN	$\rightarrow \frac{FP}{TP+FN}$	false positive rate

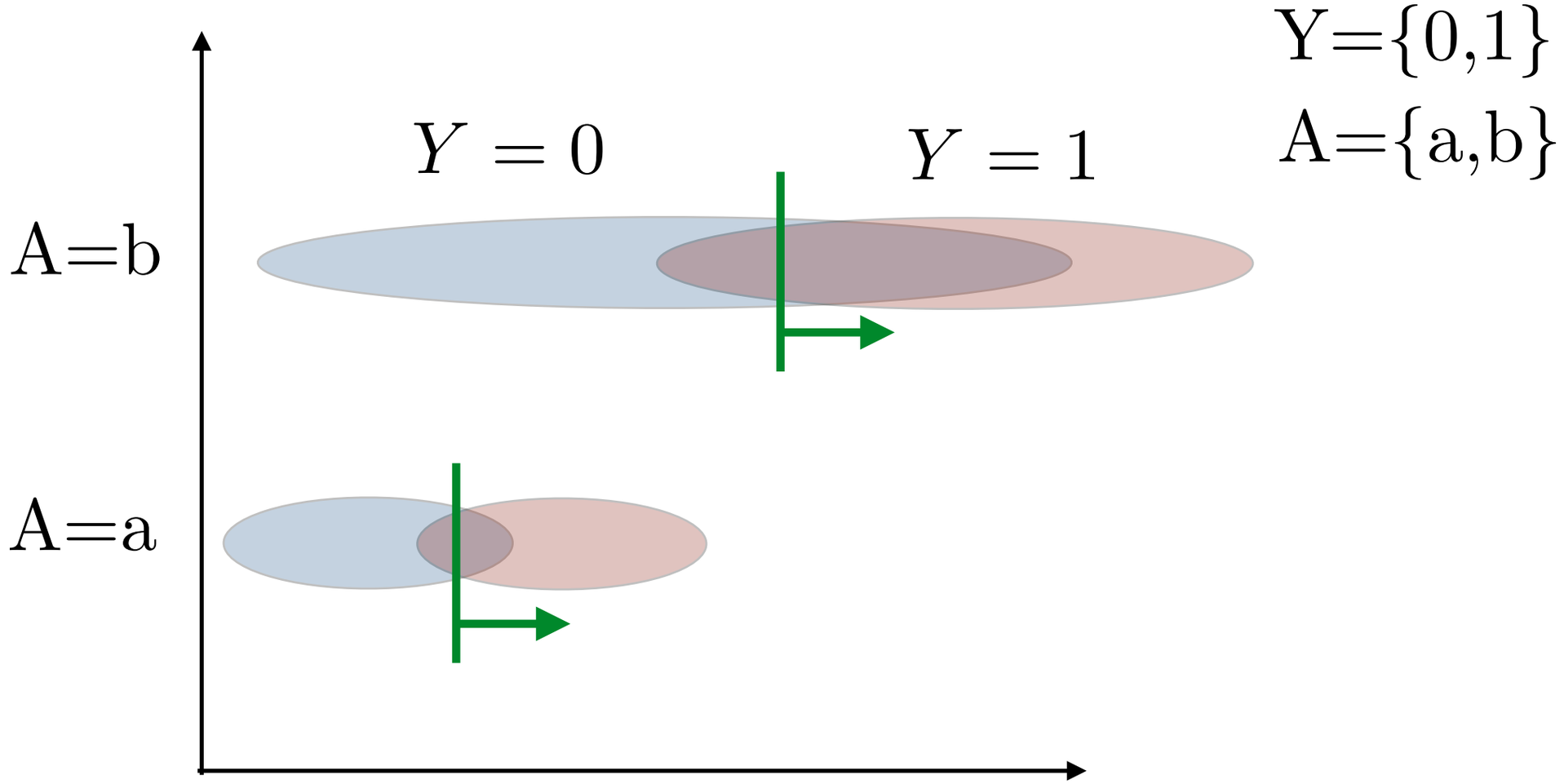
Separation



- $\mathbb{P}(R=1|Y=1,A=a)$ is the red/whole:



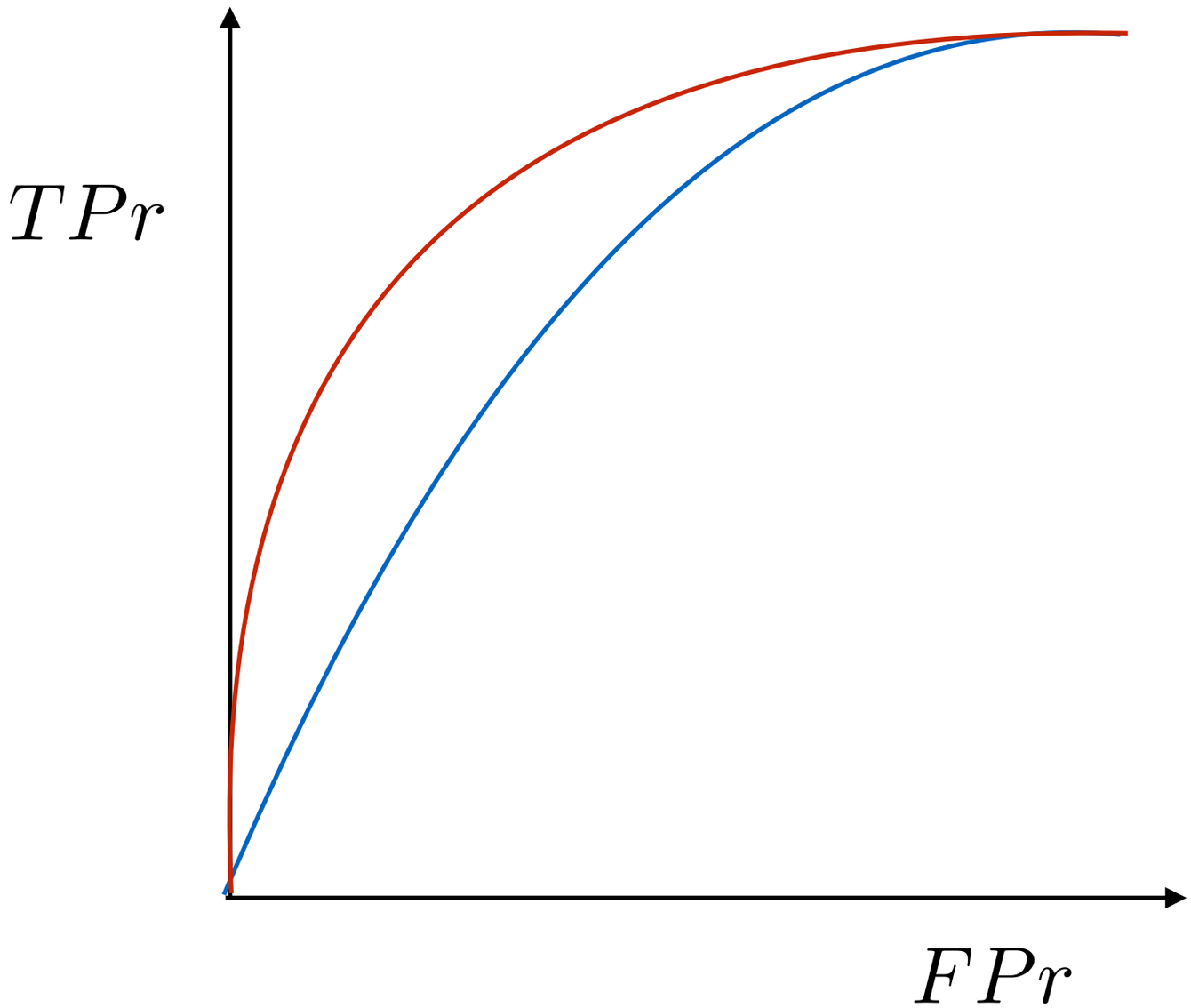
Separation



- These ratios should be the same:

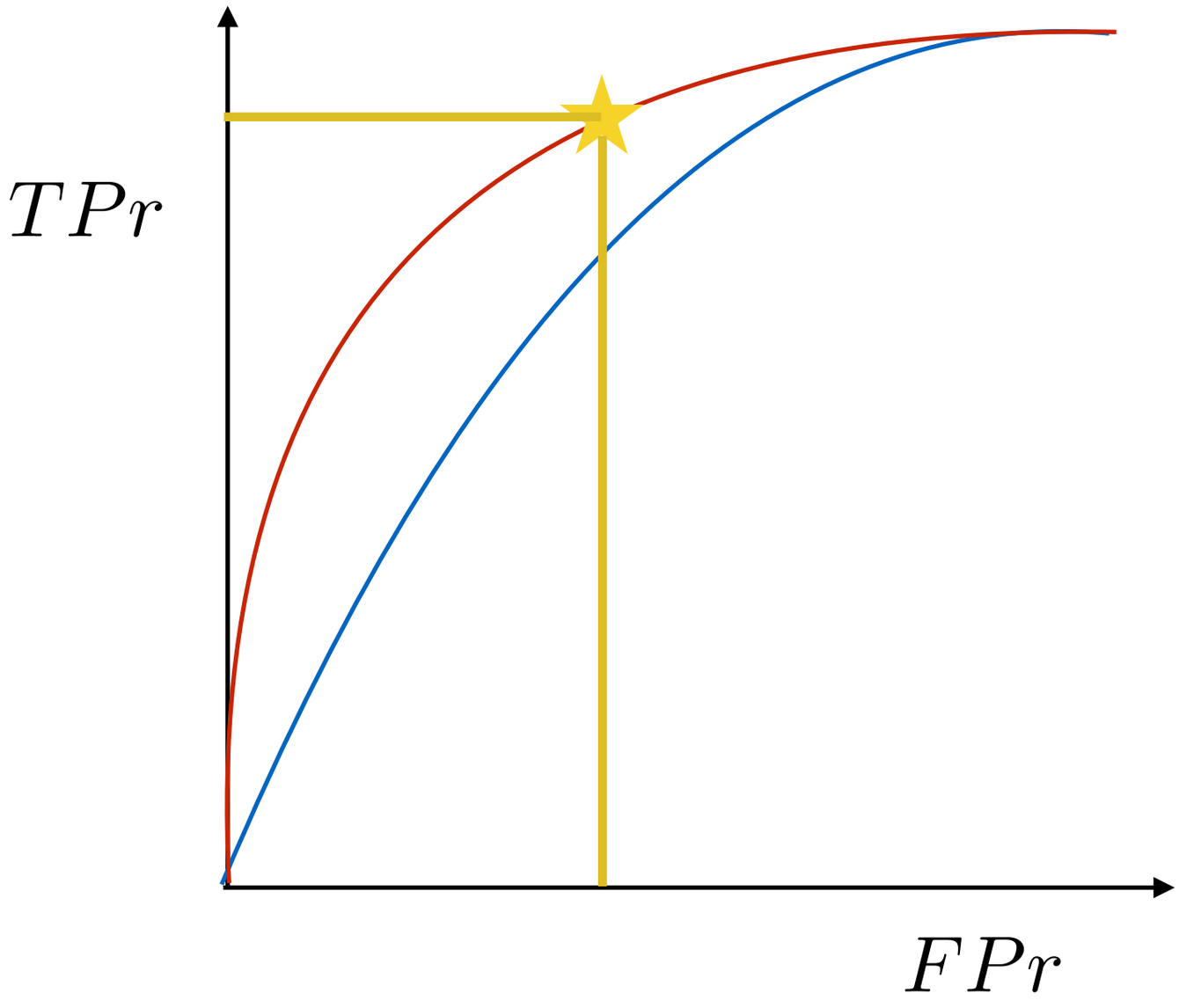
In terms of ROC curve

- ROC curves of the two groups:



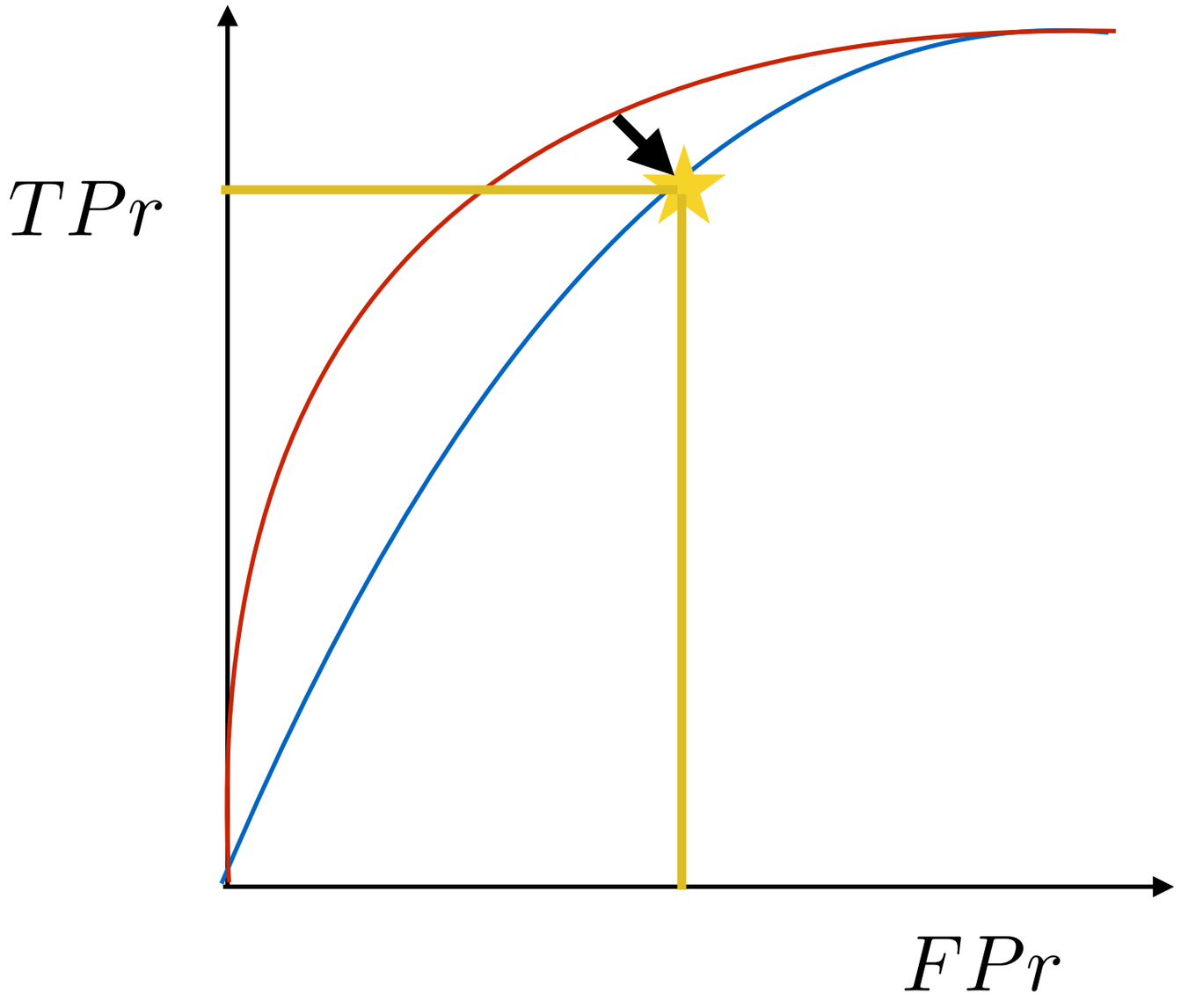
In terms of ROC curve

- ROC curves of the two groups
- This operating point is impossible for the blue group:



In terms of ROC curve

- This operating point is possible for both groups
- But we have to (artificially) deteriorate the performance for red:



Separation

- Relation with independence?

$$P(R=1|A=a) = \sum_y P(R=1, Y=y|A=a)$$

$$= \sum_y P(R=1|Y=y, A=a)P(y|A=a)$$

$$= P(R=1|Y=0, A=a)P(Y=0|A=a)$$

$$+ P(R=1|Y=1, A=a)P(Y=1|A=a)$$

Separation

- Relation with independence?

$$P(R=1|A=a) = \sum_y P(R=1, Y=y|A=a)$$

independence =

$$= \sum_y P(R=1|Y=y, A=a)P(y|A=a)$$

$$= P(R=1|Y=0, A=a)P(Y=0|A=a) + P(R=1|Y=1, A=a)P(Y=1|A=a)$$

$$P(R=1|A=b) = P(R=1|Y=0, A=b)P(Y=0|A=b) + P(R=1|Y=1, A=b)P(Y=1|A=b)$$

Separation

- In separation:

$$P(R=1|A=a) = \sum_y P(R=1, Y=y|A=a)$$

$$= \sum_y P(R=1|Y=y, A=a)P(y|A=a)$$

$$= P(R=1|Y=0, A=a)P(Y=0|A=a)$$
$$+ P(R=1|Y=1, A=a)P(Y=1|A=a)$$

$$P(R=1|A=b) = P(R=1|Y=0, A=b)P(Y=0|A=b)$$
$$+ P(R=1|Y=1, A=b)P(Y=1|A=b)$$

Separation

- When Y and A are independent:

$$P(Y = 0|A = a) = P(Y = 0|A = b)$$
$$P(Y = 1|A = a) = P(Y = 1|A = b)$$

then Separation = Independence

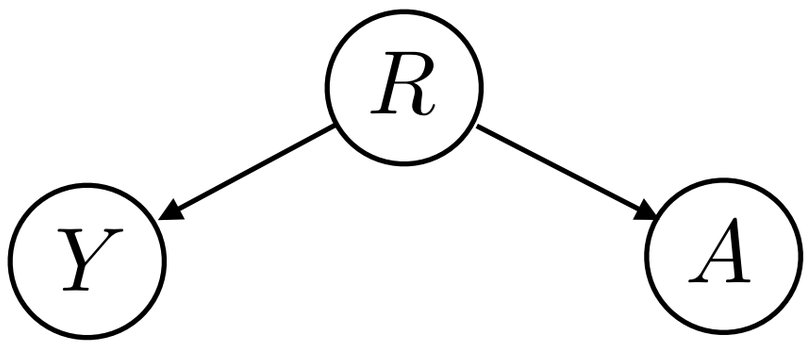
- When Y and A are heavily correlated, then the ratio's

$$\frac{P(Y = 0|A = a)}{P(Y = 0|A = b)} \qquad \frac{P(Y = 1|A = a)}{P(Y = 1|A = b)}$$

should be 'fixed':
artificially deteriorate performance on one group(?)

Sufficiency

- Defined as $Y \perp A|R$: given R, Y and A are independent



$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}\{Y = 1 \mid R = r, A = b\}$$

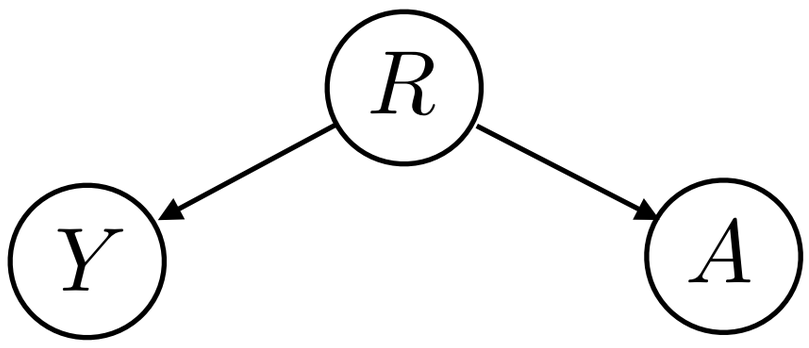
- For binary variables:
  pos. predictive value group a=pos. predictive value group b
  neg. predictive value group a=neg. predictive value group b

Confusion matrix

		predicted positive	predicted negative
truth	positive	TP	FN
	negative	FP	TN
		$\downarrow$ $\frac{TP}{TP+FP}$	$\downarrow$ $\frac{TN}{FN+TN}$
		positive predictive value	negative predictive value

Sufficiency

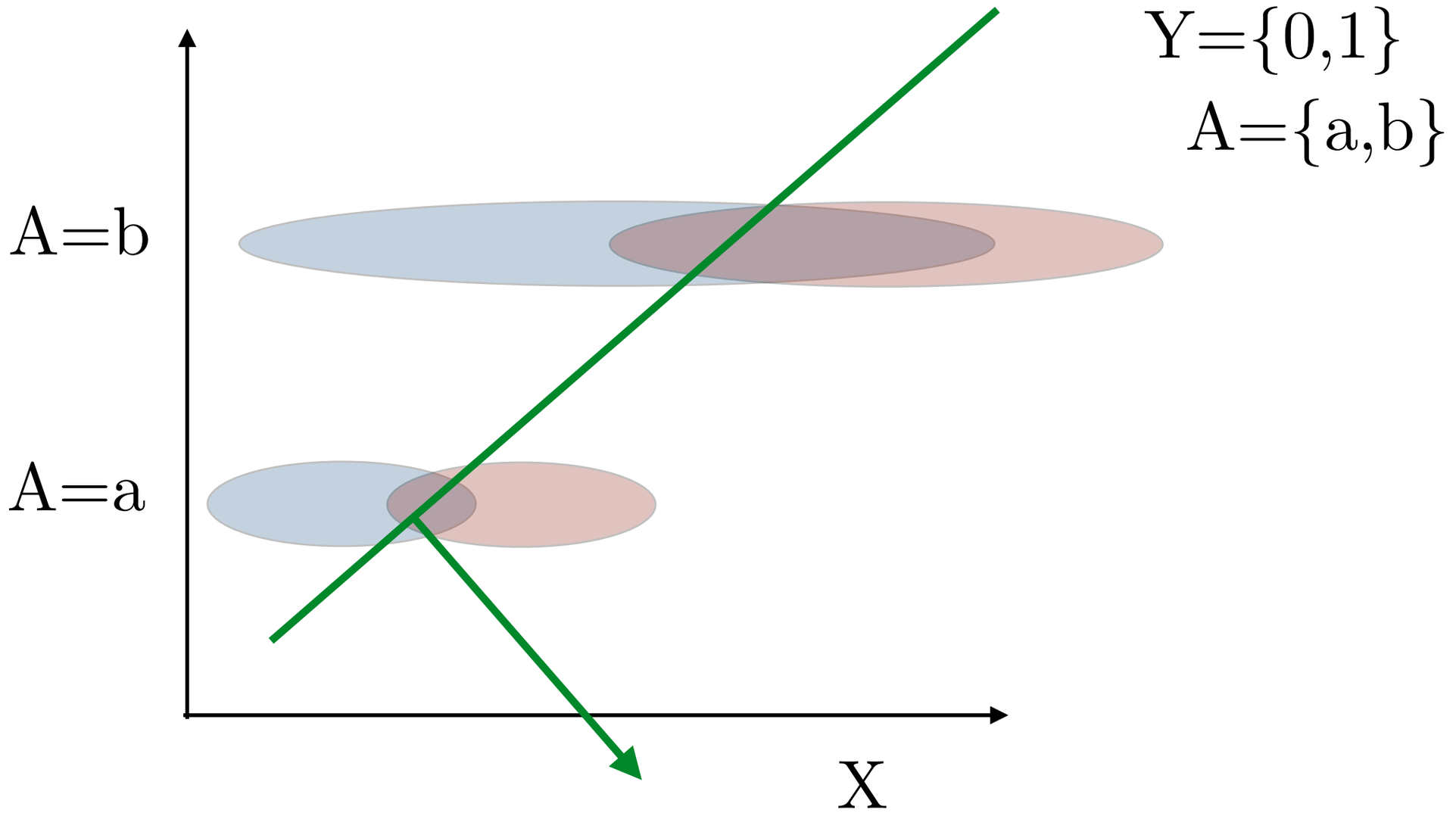
- Defined as $Y \perp A|R$: given R, Y and A are independent



$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}\{Y = 1 \mid R = r, A = b\}$$

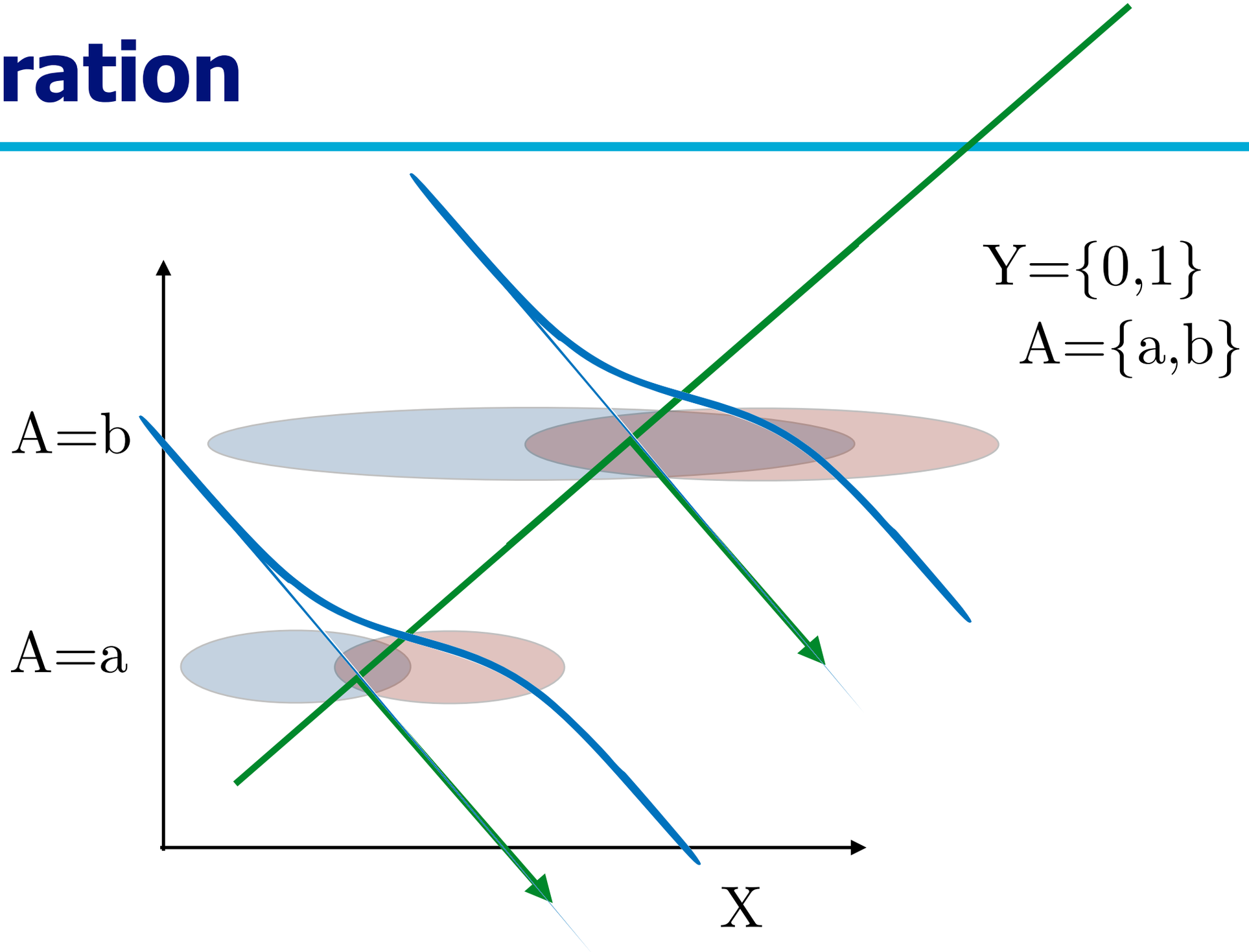
- When you predict a certain R=r, then the probability for Y=1 is the same, regardless of the group

Separation



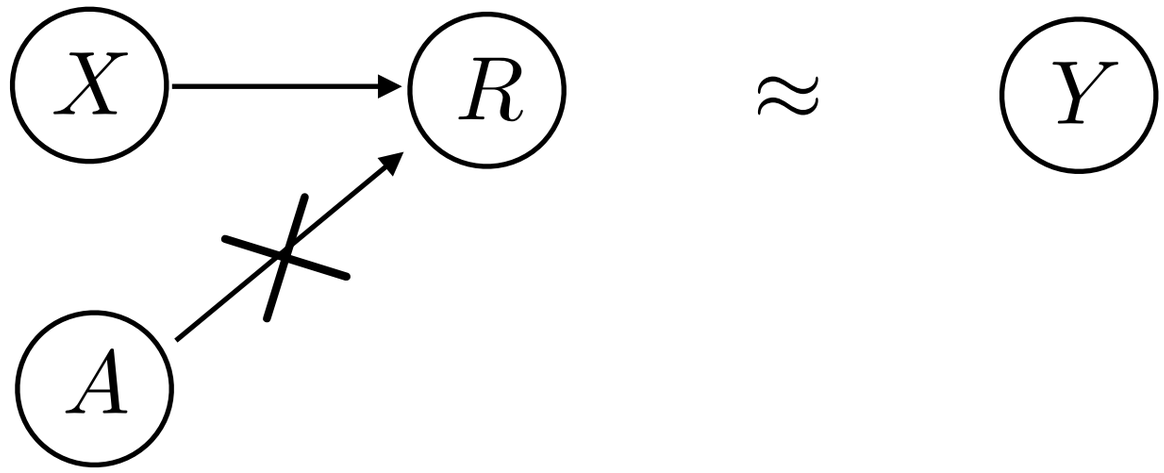
- For given output R, probability Y=1 is the same for all groups

Separation

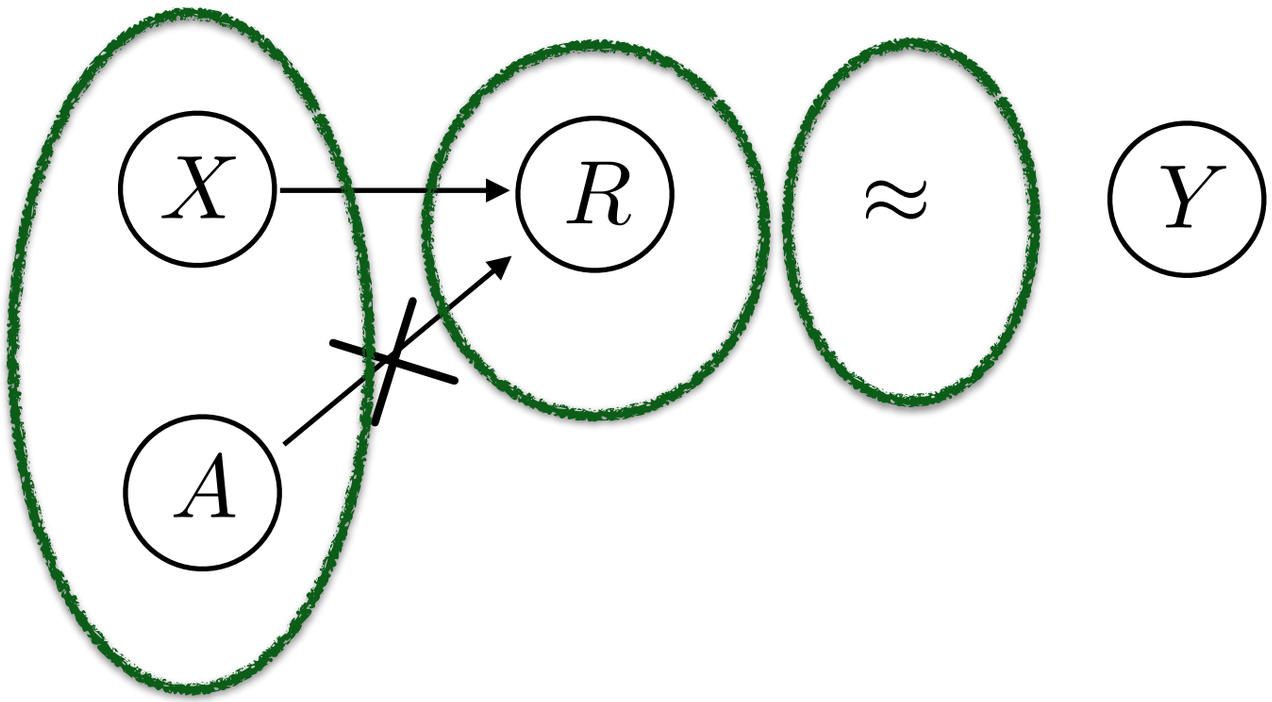


- Claimed that most classifiers do this ‘automatically’

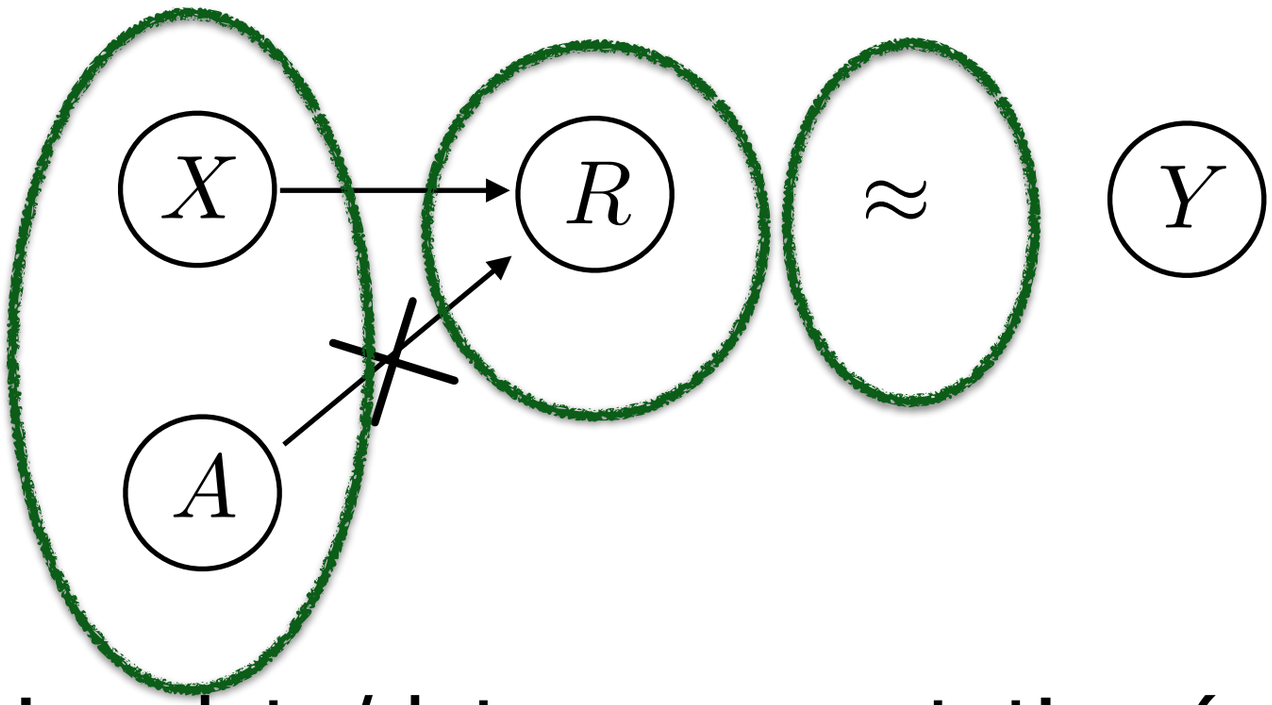
Where to enforce fairness?



Where to enforce fairness?

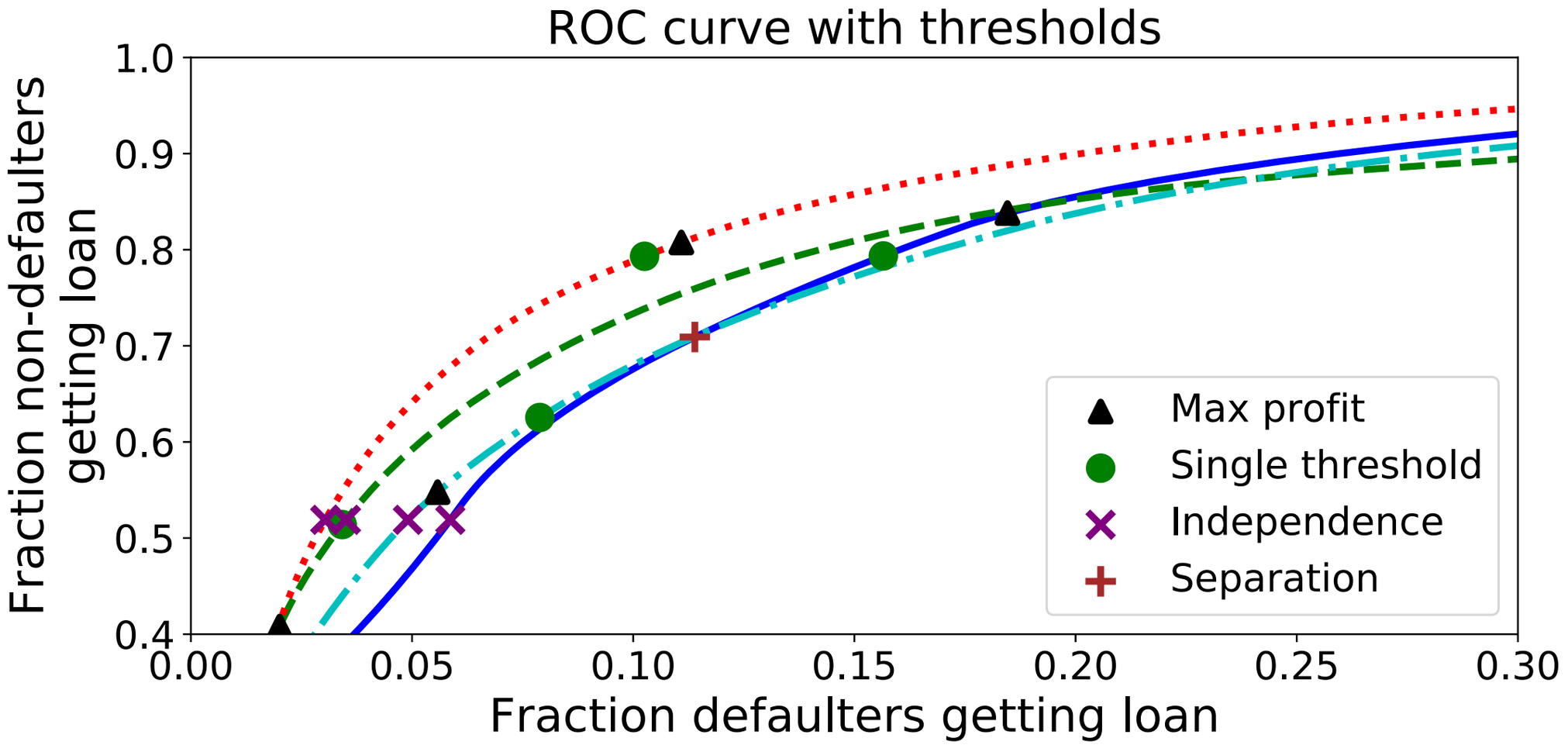


Where to enforce fairness?



- Fix the training data/data representation (resampling/ reweighing)
- Constrain the model/training procedure (constrained optimisation)
- Post hoc fix the predictor score/threshold (rescale output)

Application to credit score



- *Maximum profit:* Pick possibly group-dependent score thresholds in a way that maximizes profit.
- *Single threshold:* Pick a single uniform score threshold for all groups in a way that maximizes profit.
- *Separation:* Achieve an equal true/false positive rate in all groups. Subject to this constraint, maximize profit.
- *Independence:* Achieve an equal acceptance rate in all groups. Subject to this constraint, maximize profit.

Many Issues

- Is there an objective way to define the 'true target' Y? History/human prejudice may have polluted it?
- Can we assume that the sensitive attribute A is known?
- Can we assume that the sensitive attribute A is given?
- If A is correlated with Y, how can we ignore A?
- A is typically correlated with other X's: removing A from the input does not per se help. How to fix this?
- …

Conclusions

- To see if an algorithm is fair, you need to define (and measure!) the sensitive attribute A

- Tricky business; different criteria/constraints exist
- When features are correlated with A, solutions seem sometimes impossible/very suboptimal in terms of classification performance
- Need relaxations of the constraints
- (How to do the tradeoff? Is there a criterion for that?)

TU Delft