Inherent Tradeoffs in Learning Fair Representations

Han Zhao
han.zhao@cs.cmu.edu
Machine Learning Department, Carnegie Mellon University

Incompatibility between Definitions of Fairness

COMPAS (Northpointe):

Recidivism risk assessment tool used in a county in Florida

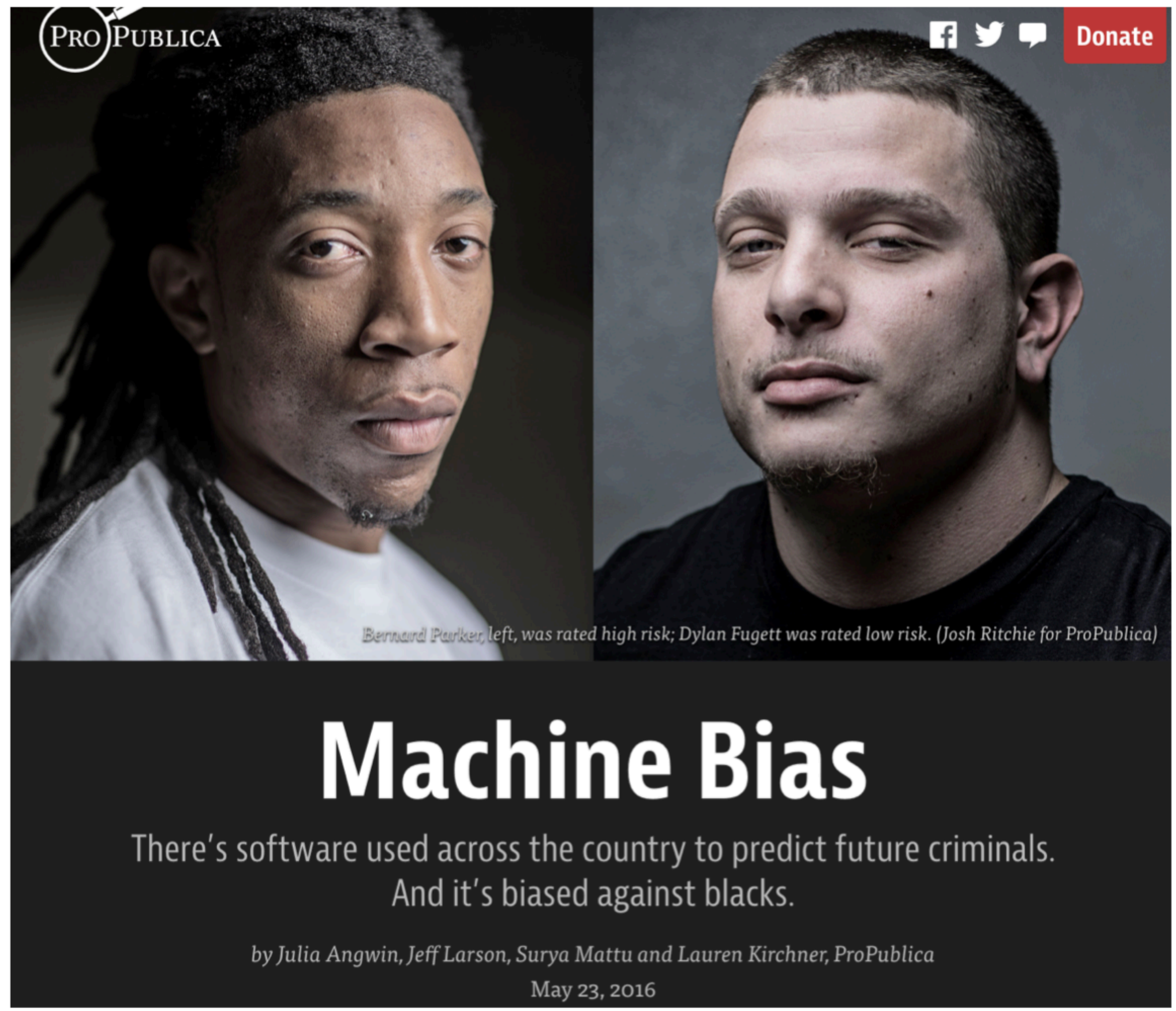
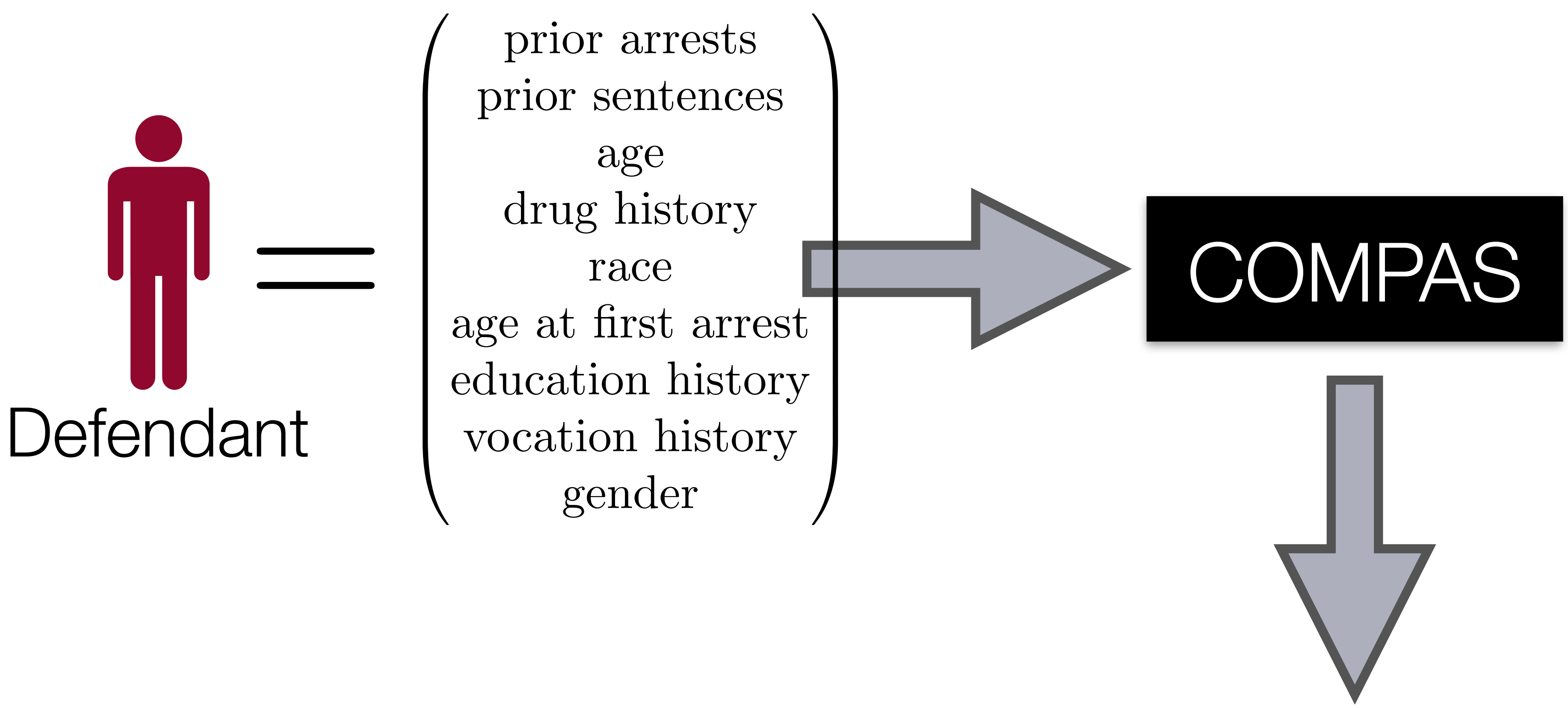


Figure credit: ProPublica, Larson et al., 2016

Incompatibility between Definitions of Fairness

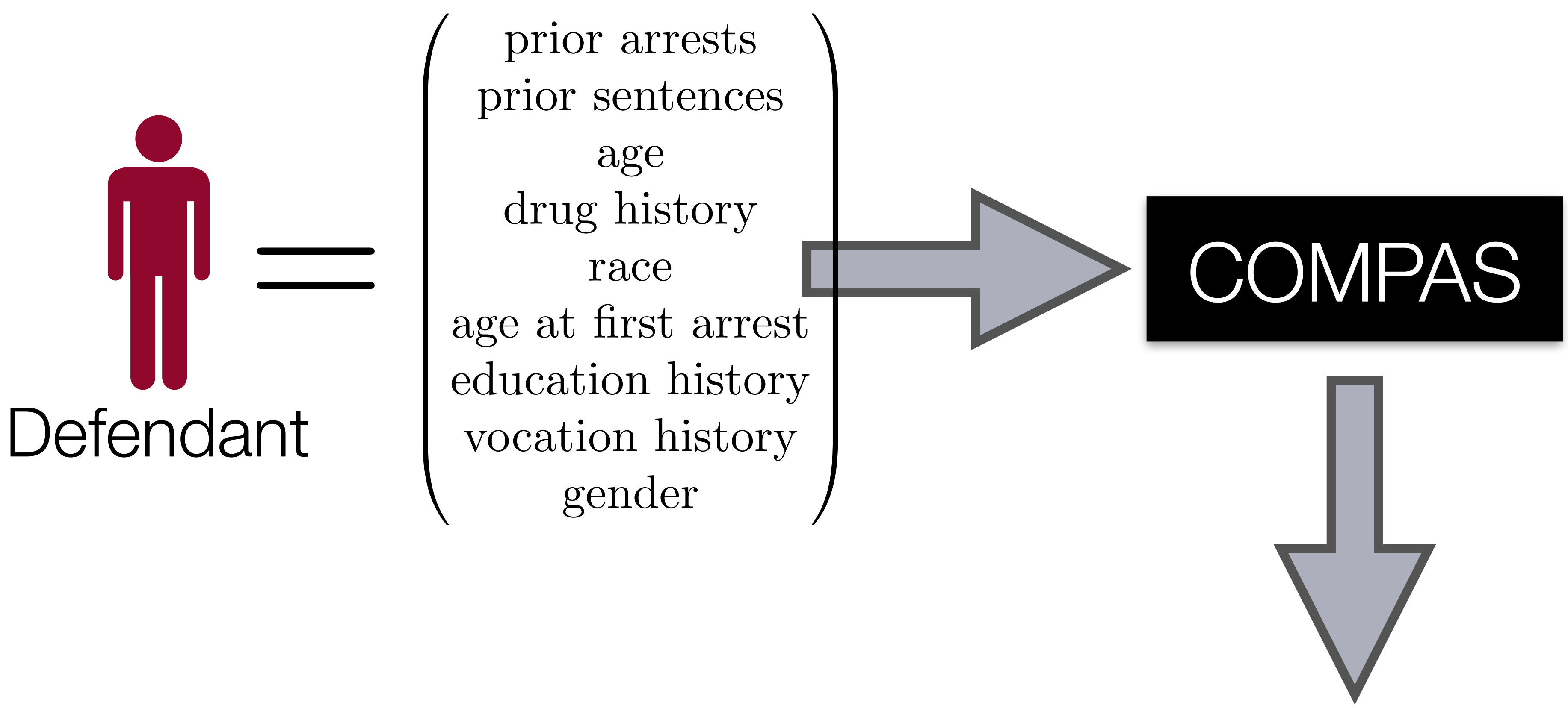
COMPAS (high level):



Risk score: $C(x) \in (0, 1)$

Incompatibility between Definitions of Fairness

COMPAS (high level):

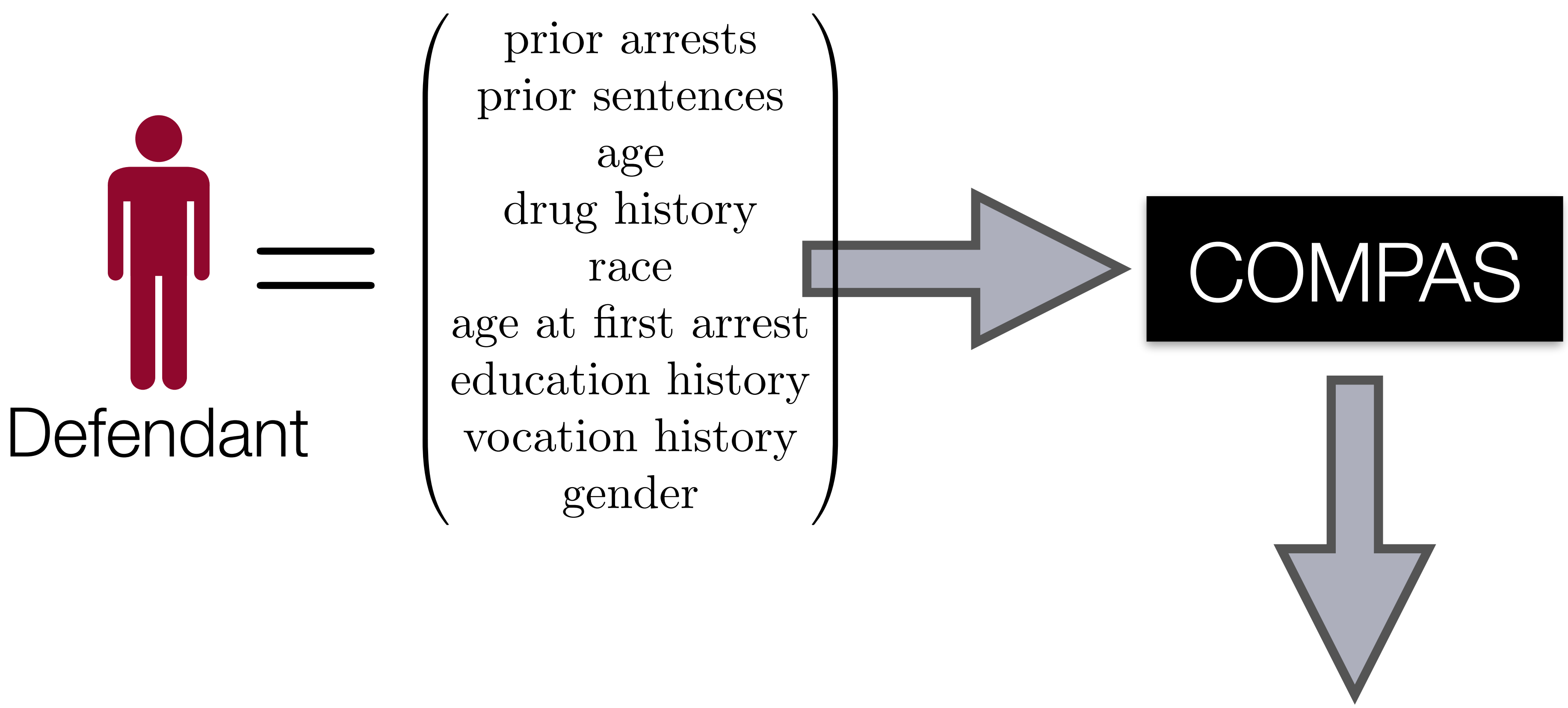


Risk score: $C(x) \in (0, 1)$

- Risk score ~ likelihood of defendant to recidivate

Incompatibility between Definitions of Fairness

COMPAS (high level):

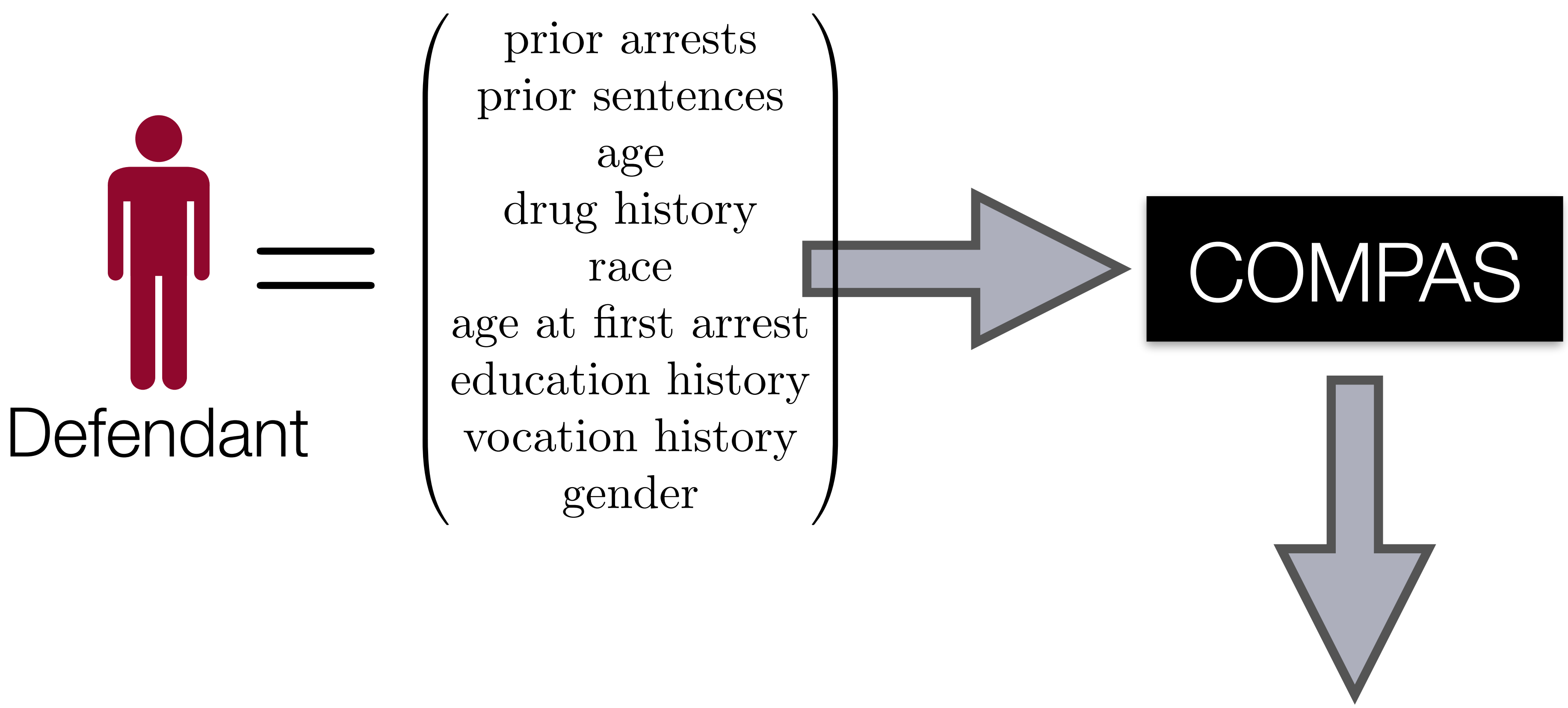


Risk score: $C(x) \in (0, 1)$

- Risk score ~ likelihood of defendant to recidivate
- Inputs have (noisy) true label: 0 (not recidivate) / 1 (will recidivate)

Incompatibility between Definitions of Fairness

COMPAS (high level):



Risk score: $C(x) \in (0, 1)$

- Risk score ~ likelihood of defendant to recidivate
- Inputs have (noisy) true label: 0 (not recidivate) / 1 (will recidivate)
- The risk score + thresholding: 0 (low risk) / 1 (high risk)

Incompatibility between Definitions of Fairness

ProPublica criticism:

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

Source: ProPublica, Larson et al., 2016

Incompatibility between Definitions of Fairness

ProPublica criticism:

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

- Black defendants more likely than white to be **incorrectly** labeled "high risk"

Source: ProPublica, Larson et al., 2016

Incompatibility between Definitions of Fairness

ProPublica criticism:

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

- Black defendants more likely than white to be **incorrectly** labeled "high risk"
- White defendants more likely than black to be **incorrectly** labeled "low risk"

Source: ProPublica, Larson et al., 2016

Incompatibility between Definitions of Fairness

ProPublica criticism:

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

- Black defendants more likely than white to be **incorrectly** labeled "high risk"
- White defendants more likely than black to be **incorrectly** labeled "low risk"

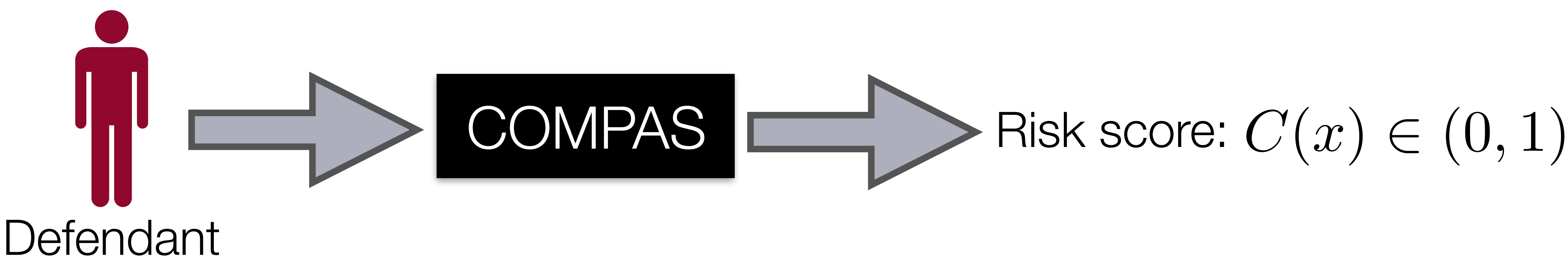
Bias: Disparate FPR/FNR across groups!

Source: ProPublica, Larson et al., 2016

Incompatibility between Definitions of Fairness

Northpointes' defense:

Defendants labeled as "high risk" **equally likely** to recidivate, regardless of race



- The COMPAS tool $C(x)$ is **statistically calibrated by group**

Incompatibility between Definitions of Fairness

Northpointes’ defense:

Defendants labeled as “high risk” **equally likely** to recidivate, regardless of race

Risk score: $C(x) \in (0,1)$

Defendant

- The COMPAS tool $C(x)$ is **statistically calibrated by group**
- Let $A \in \{0,1\}$ be the group membership (race), $Y \in \{0,1\}$ be the true label (recidivism), then

$$\forall a \in \{0,1\}, \forall c \in (0,1), \quad \Pr(Y = 1 \mid C(x) = c, A = a) = c$$

Incompatibility between Definitions of Fairness

Northpointes' defense:

Defendants labeled as "high risk" **equally likely** to recidivate, regardless of race

Risk score: $C(x) \in (0, 1)$

Defendant

- The COMPAS tool $C(x)$ is **statistically calibrated by group**
- Let $A \in \{0, 1\}$ be the group membership (race), $Y \in \{0, 1\}$ be the true label (recidivism), then

$$\forall a \in \{0,1\}, \forall c \in (0,1), \quad \Pr(Y = 1 \mid C(x) = c, A = a) = c$$

No Bias: Equal treatment!

Incompatibility between Definitions of Fairness

Fundamental incompatibility between different notions of fairness:

- True label: $Y \in \{0, 1\}$
- Group membership: $A \in \{0, 1\}$
- Probabilistic classifier: $\widehat{Y} \in (0, 1)$ or binary classifier: $\widehat{Y} \in \{0, 1\}$
- Base rate: $\Pr(Y = 1 \mid A = a), \ a \in \{0, 1\}$
- Difference of base rates:

$$\Delta_{\mathrm{BR}} = |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$$

Incompatibility between Definitions of Fairness

Fundamental incompatibility between different notions of fairness:

- True label: $Y \in \{0, 1\}$
- Group membership: $A \in \{0, 1\}$
- Probabilistic classifier: $\widehat{Y} \in (0, 1)$ or binary classifier: $\widehat{Y} \in \{0, 1\}$
- Base rate: $\Pr(Y = 1 \mid A = a), \; a \in \{0, 1\}$
- Difference of base rates:

$$\Delta_{\mathrm{BR}} = |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$$

Theorem (Chouldechova'17, Kleinberg, Mullainathan, Raghavan'16): Statistical calibration and Equalized FPR/FNR cannot hold simultaneously unless $\Delta_{\mathrm{BR}} = 0$ ($A \perp Y$) or $\widehat{Y} = Y$ (perfect prediction).

Incompatibility between Definitions of Fairness

Lesson learned:

Depending on the problem, choose the appropriate criterion

Incompatibility between Definitions of Fairness

Lesson learned:

Depending on the problem, choose the appropriate criterion

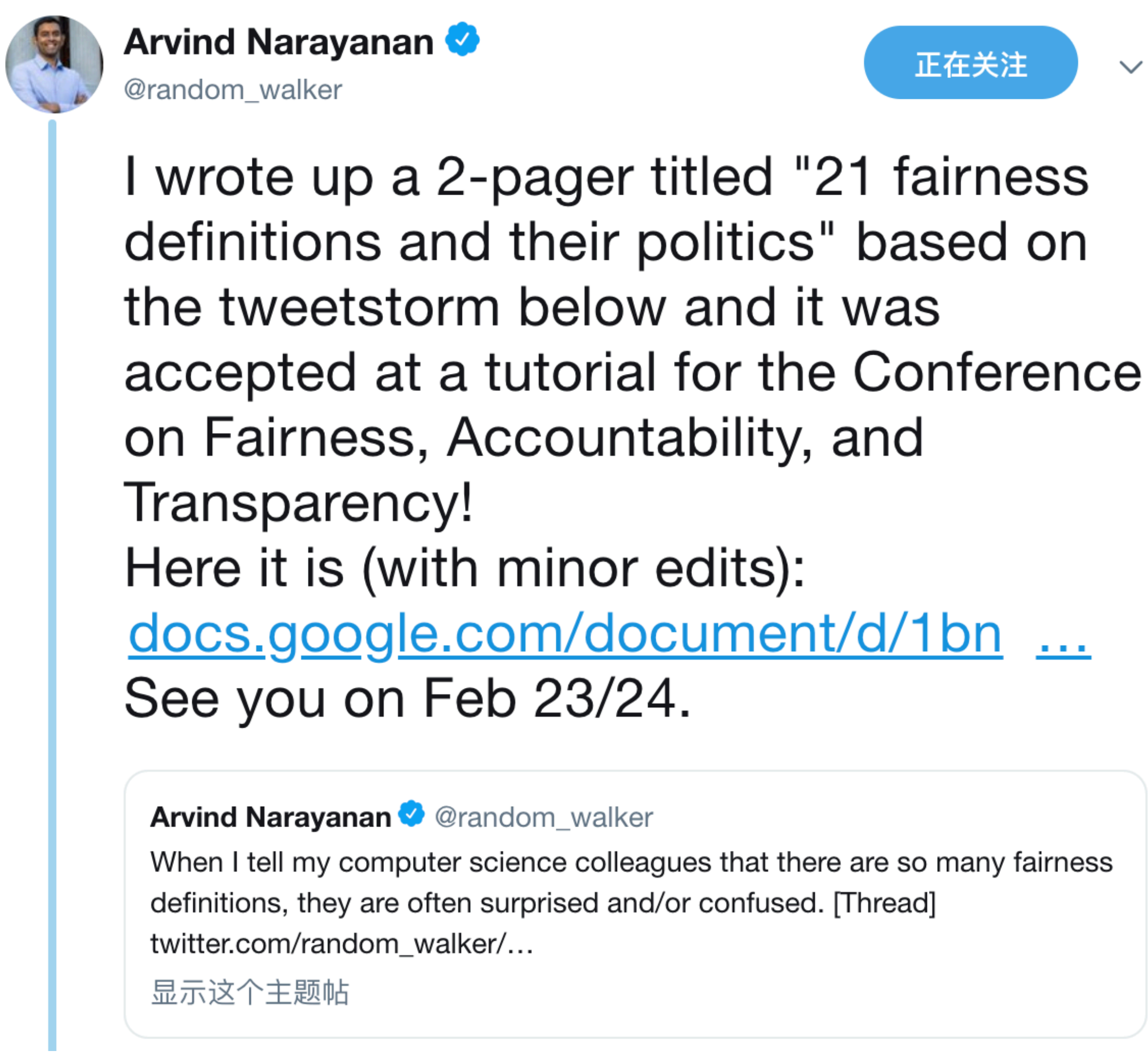
But, there are just too many definitions…

Incompatibility between Definitions of Fairness

Lesson learned:

Depending on the problem, choose the appropriate criterion

But, there are just too many definitions…


Arvind Narayanan
@random_walker
正在关注

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!
Here it is (with minor edits): docs.google.com/document/d/1bn …
See you on Feb 23/24.

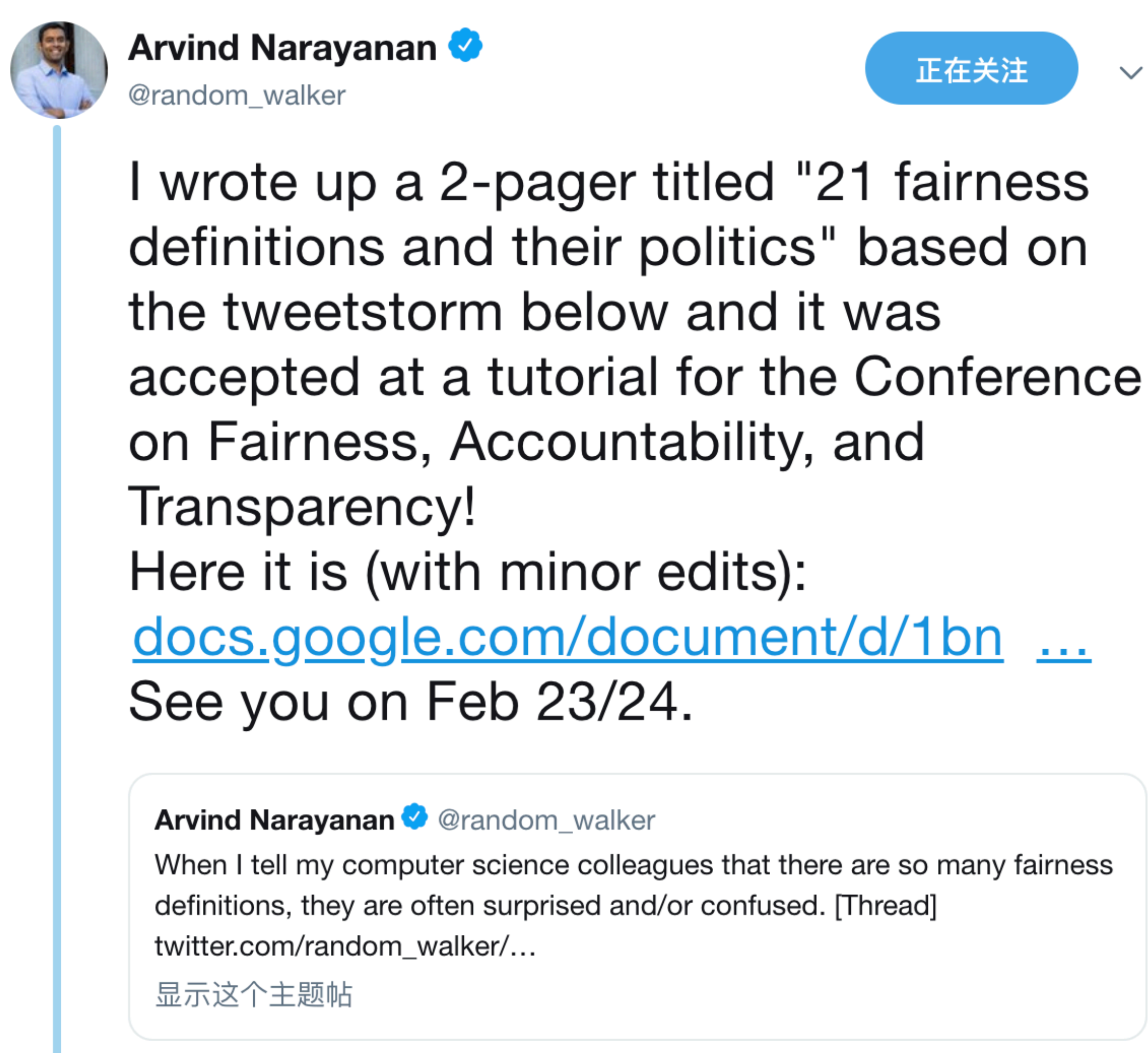
Arvind Narayanan @random_walker
When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread]
twitter.com/random_walker/…
显示这个主题帖

Incompatibility between Definitions of Fairness

Lesson learned:

Depending on the problem, choose the appropriate criterion

But, there are just too many definitions…



Definition	Paper	Citation #
Group fairness or statistical parity	[12]	208
Conditional statistical parity	[11]	29
Predictive parity	[10]	57
False positive error rate balance	[10]	57
False negative error rate balance	[10]	57
Equalised odds	[14]	106
Conditional use accuracy equality	[8]	18
Overall accuracy equality	[8]	18
Treatment equality	[8]	18
Test-fairness or calibration	[10]	57
Well calibration	[16]	81
Balance for positive class	[16]	81
Balance for negative class	[16]	81

Incompatibility between Definitions of Fairness

Lesson learned:

Depending on the problem, choose the appropriate criterion

But, there are just too many definitions…

Arvind Narayanan @random_walker 正在关注

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!
Here it is (with minor edits): docs.google.com/document/d/1bn …
See you on Feb 23/24.

> **Arvind Narayanan** @random_walker
> When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread] twitter.com/random_walker/…
> 显示这个主题帖

Definition	Paper	Citation #
Group fairness or statistical parity	[12]	208
Conditional statistical parity	[11]	29
Predictive parity	[10]	57
False positive error rate balance	[10]	57
False negative error rate balance	[10]	57
Equalised odds	[14]	106
Conditional use accuracy equality	[8]	18
Overall accuracy equality	[8]	18
Treatment equality	[8]	18
Test-fairness or calibration	[10]	57
Well calibration	[16]	81
Balance for positive class	[16]	81
Balance for negative class	[16]	81

Incompatibility between Definitions of Fairness

Lesson learned:

Depending on the problem, choose the appropriate criterion

But, there are just too many definitions…

Arvind Narayanan @random_walker 正在关注

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!
Here it is (with minor edits): docs.google.com/document/d/1bn …
See you on Feb 23/24.

> **Arvind Narayanan** @random_walker
> When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread] twitter.com/random_walker/…
> 显示这个主题帖

Definition	Paper	Citation #
Group fairness or statistical parity	[12]	208
Conditional statistical parity	[11]	29
Predictive parity	[10]	57
False positive error rate balance	[10]	57
False negative error rate balance	[10]	57
Equalised odds	[14]	106
Conditional use accuracy equality	[8]	18
Overall accuracy equality	[8]	18
Treatment equality	[8]	18
Test-fairness or calibration	[10]	57
Well calibration	[16]	81
Balance for positive class	[16]	81
Balance for negative class	[16]	81

Incompatibility between Definitions of Fairness

Lesson learned:

Depending on the problem, choose the appropriate criterion

But, there are just too many definitions…

Arvind Narayanan @random_walker

正在关注

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!
Here it is (with minor edits): docs.google.com/document/d/1bn …
See you on Feb 23/24.

> **Arvind Narayanan** @random_walker
> When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread] twitter.com/random_walker/…
> 显示这个主题帖

Definition	Paper	Citation #
Group fairness or statistical parity	[12]	208
Conditional statistical parity	[11]	29
Predictive parity	[10]	57
False positive error rate balance	[10]	57
False negative error rate balance	[10]	57
Equalised odds	[14]	106
Conditional use accuracy equality	[8]	18
Overall accuracy equality	[8]	18
Treatment equality	[8]	18
Test-fairness or calibration	[10]	57
Well calibration	[16]	81
Balance for positive class	[16]	81
Balance for negative class	[16]	81

Fairness vs Utility

Statistical parity (demographic parity):

$$\widehat{Y} \perp A$$

The prediction given by an algorithm shouldn't take the sensitive attribute A into account

- College admission: affirmative action
- Movie recommendation
- …

Fairness vs Utility

Statistical parity (demographic parity):

$$\widehat{Y} \perp A$$

The prediction given by an algorithm shouldn't take the sensitive attribute A into account

- College admission: affirmative action
- Movie recommendation
- …

How to achieve statistical parity while preserving utility?

Fairness vs Utility

Statistical parity (demographic parity): $\widehat{Y} \perp A$

$$\widehat{Y} \perp A \Leftrightarrow I(\widehat{Y}; A) = 0 \Leftrightarrow \Pr(\widehat{Y} \mid A = 0) = \Pr(\widehat{Y} \mid A = 1)$$

So it suffices if we could learn invariant representations Z that is independent of A, then any predictor $\widehat{Y}$ upon Z should be independent of A as well

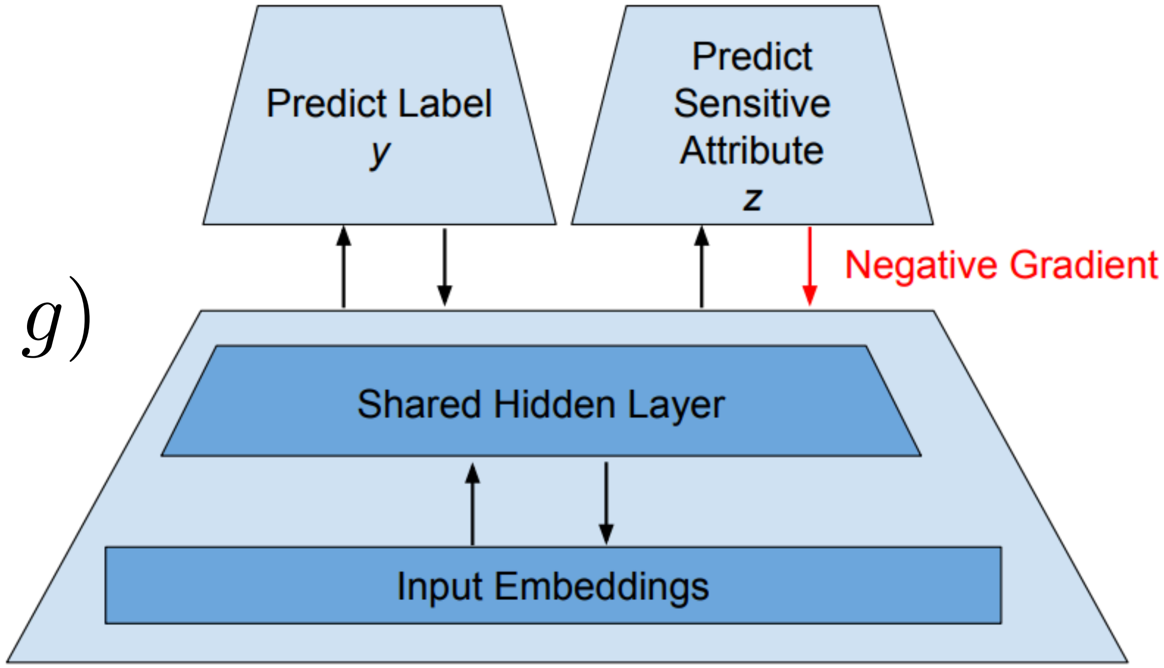
Fairness vs Utility

Statistical parity (demographic parity): $\widehat{Y} \perp A$

$$\widehat{Y} \perp A \Leftrightarrow I(\widehat{Y}; A) = 0 \Leftrightarrow \Pr(\widehat{Y} \mid A = 0) = \Pr(\widehat{Y} \mid A = 1)$$

So it suffices if we could learn invariant representations Z that is independent of A, then any predictor $\widehat{Y}$ upon Z should be independent of A as well

$$\min_{h,g} \max_{h'} \quad \varepsilon_Y(h \circ g) - \lambda \cdot \varepsilon_A(h' \circ g)$$



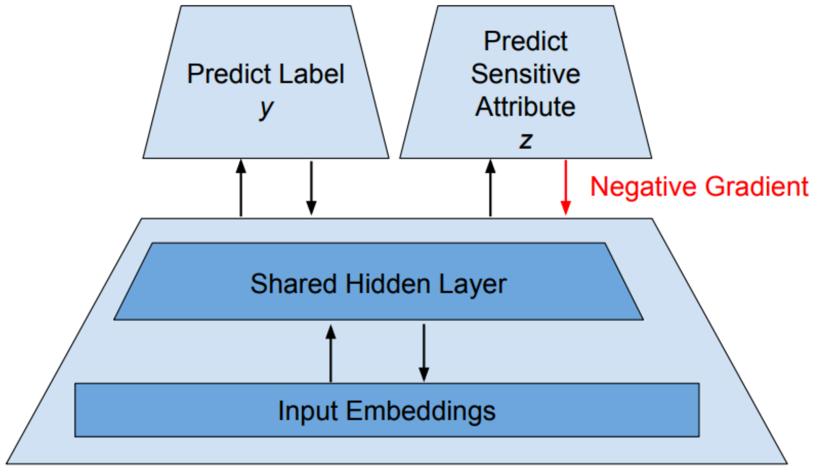
Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations, Beutel et al. FAT/ML 2017

Fairness vs Utility

Minimax optimization formulation:

$$\min_{h,g} \max_{h'} \quad \varepsilon_Y(h \circ g) - \lambda \cdot \varepsilon_A(h' \circ g)$$

In practice, the loss function $\varepsilon(\cdot)$ is often chosen as the cross-entropy loss



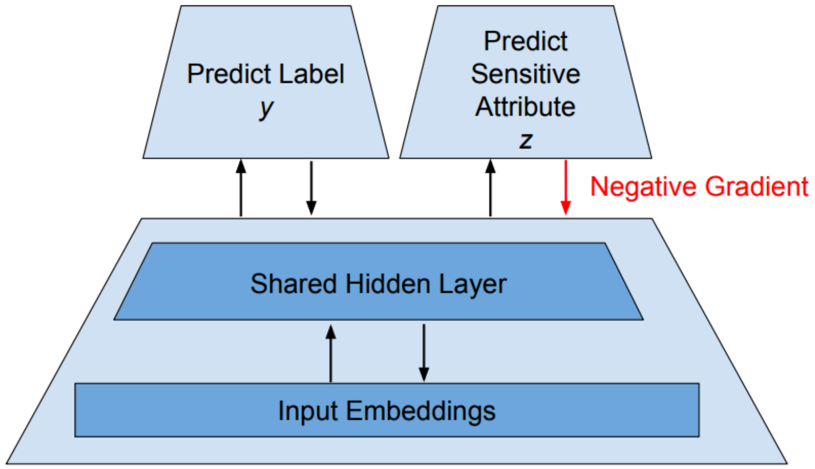
Fairness vs Utility

Minimax optimization formulation:

$$\min_{h,g} \max_{h'} \quad \varepsilon_Y(h \circ g) - \lambda \cdot \varepsilon_A(h' \circ g)$$

In practice, the loss function $\varepsilon(\cdot)$ is often chosen as the cross-entropy loss

- Shared representations $Z = g(X)$



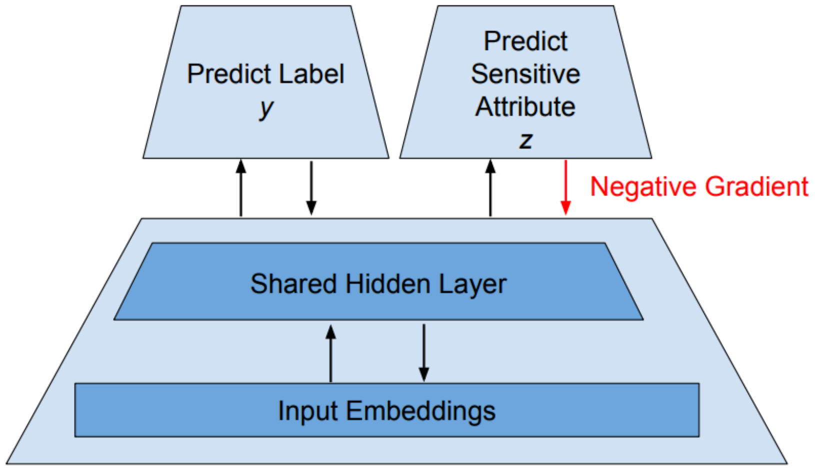
Fairness vs Utility

Minimax optimization formulation:

$$\min_{h,g} \max_{h'} \quad \varepsilon_Y(h \circ g) - \lambda \cdot \varepsilon_A(h' \circ g)$$

In practice, the loss function $\varepsilon(\cdot)$ is often chosen as the cross-entropy loss

- Shared representations $Z = g(X)$
- For any fixed $Z = g(X)$, the optimal h, h' is given by the corresponding conditional distribution



Fairness vs Utility

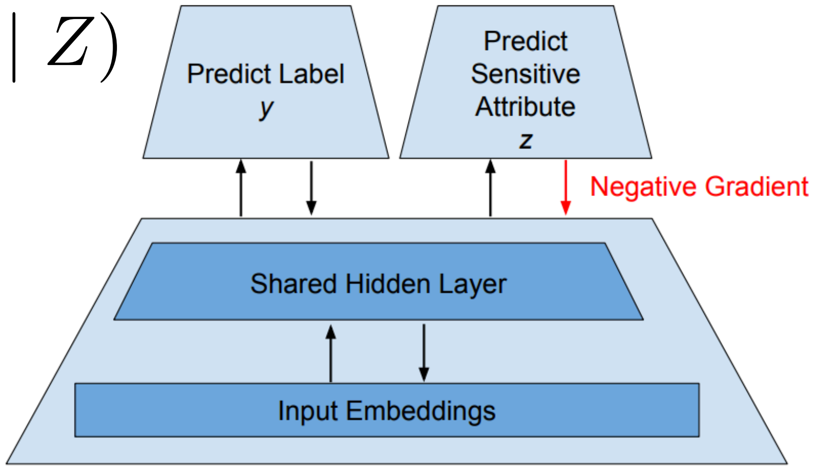
Minimax optimization formulation:

$$\min_{h,g} \max_{h'} \quad \varepsilon_Y(h \circ g) - \lambda \cdot \varepsilon_A(h' \circ g)$$

In practice, the loss function $\varepsilon(\cdot)$ is often chosen as the cross-entropy loss

- Shared representations $Z = g(X)$
- For any fixed $Z = g(X)$, the optimal h, h' is given by the corresponding conditional distribution

$$h(Z) = \Pr(Y = 1 \mid Z); \quad h'(Z) = \Pr(A = 1 \mid Z)$$



Fairness vs Utility

Simplified optimization:

$$\min_{Z=g(X)} \quad H(Y \mid Z) - \lambda \cdot H(A \mid Z)$$

Clearly, the optimal solution depends on the "coupling" between A, Y:

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Simplified optimization:

$$\min_{Z=g(X)} \quad H(Y \mid Z) - \lambda \cdot H(A \mid Z)$$

Clearly, the optimal solution depends on the "coupling" between A, Y:

- If $A = Y, \ a.s.,$ then we cannot hope to find a good tradeoff

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Simplified optimization:

$$\min_{Z=g(X)} \quad H(Y \mid Z) - \lambda \cdot H(A \mid Z)$$

Clearly, the optimal solution depends on the “coupling” between A, Y:

- If $A = Y,\ a.s.,$ then we cannot hope to find a good tradeoff

- If $A \perp Y$, then we can preserve information about Y while filtering out information related to A

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS’19

Fairness vs Utility

Simplified optimization:

$$\min_{Z=g(X)} \quad H(Y \mid Z) - \lambda \cdot H(A \mid Z)$$

Clearly, the optimal solution depends on the "coupling" between A, Y:

- If $A = Y, \ a.s.$, then we cannot hope to find a good tradeoff
- If $A \perp Y$, then we can preserve information about Y while filtering out information related to A

In general, tradeoff exists between fairness and utility

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Theorem: If $\widehat{Y} = (h \circ g)(X)$ satisfies statistical parity, then

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \Delta_{\mathrm{BR}}$$

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Theorem: If $\widehat{Y} = (h \circ g)(X)$ satisfies statistical parity, then

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \Delta_{\mathrm{BR}}$$

- $\mathrm{Err}_a(h \circ g)$ is the true binary classification error conditioned on group $A = a$

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Theorem: If $\widehat{Y} = (h \circ g)(X)$ satisfies statistical parity, then

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \Delta_{\mathrm{BR}}$$

- $\mathrm{Err}_a(h \circ g)$ is the true binary classification error conditioned on group $A = a$
- Recall $\Delta_{\mathrm{BR}} = |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$ measures the difference of the base rates

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Theorem: If $\widehat{Y} = (h \circ g)(X)$ satisfies statistical parity, then

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \Delta_{\mathrm{BR}}$$

- $\mathrm{Err}_a(h \circ g)$ is the true binary classification error conditioned on group $A = a$
- Recall $\Delta_{\mathrm{BR}} = |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$ measures the difference of the base rates
- Interpretation: cannot simultaneously minimize errors on both groups, has to sacrifice accuracy on one of the (minority) group if we enforce statistical parity

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Theorem: If $\widehat{Y} = (h \circ g)(X)$ satisfies statistical parity, then

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \Delta_{\mathrm{BR}}$$

- $\mathrm{Err}_a(h \circ g)$ is the true binary classification error conditioned on group $A = a$
- Recall $\Delta_{\mathrm{BR}} = |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$ measures the difference of the base rates
- Interpretation: cannot simultaneously minimize errors on both groups, has to sacrifice accuracy on one of the (minority) group if we enforce statistical parity
- If $A = Y$, then $\Delta_{\mathrm{BR}} = 1$, meaning $\max\{\mathrm{Err}_0(h \circ g), \mathrm{Err}_1(h \circ g)\} \geq 0.5$

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Theorem: If $\widehat{Y} = (h \circ g)(X)$ satisfies statistical parity, then

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \Delta_{\mathrm{BR}}$$

- $\mathrm{Err}_a(h \circ g)$ is the true binary classification error conditioned on group $A = a$
- Recall $\Delta_{\mathrm{BR}} = |\Pr(Y = 1 \mid A = 0) - \Pr(Y = 1 \mid A = 1)|$ measures the difference of the base rates
- Interpretation: cannot simultaneously minimize errors on both groups, has to sacrifice accuracy on one of the (minority) group if we enforce statistical parity
- If $A = Y$, then $\Delta_{\mathrm{BR}} = 1$, meaning $\max\{\mathrm{Err}_0(h \circ g), \mathrm{Err}_1(h \circ g)\} \geq 0.5$
- If $A \perp Y$, then $\Delta_{\mathrm{BR}} = 0$, lower bound gracefully degrades to 0, i.e., no constraint on utility

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Approximate version exists as well, consider:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

Then the following lower bounds hold:

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \Delta_{\mathrm{BR}} - \Delta_{\mathrm{DP}}(\widehat{Y})$$

where

$$\Delta_{\mathrm{DP}}(\widehat{Y}) = \left|\Pr(\widehat{Y} = 1 \mid A = 0) - \Pr(\widehat{Y} = 1 \mid A = 1)\right|$$

is an approximate version of statistical parity (demographic parity)

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Approximate version exists as well, consider:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

Define f-divergence between distribution $\mathcal{P}$ and $\mathcal{Q}$:

$$D_f(\mathcal{P} \mid\mid \mathcal{Q}) = \mathbb{E}_{\mathcal{Q}}\left[f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right]$$

where $f(1) = 0$ and is convex

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Approximate version exists as well, consider:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

Define f-divergence between distribution $\mathcal{P}$ and $\mathcal{Q}$:

$$D_f(\mathcal{P} \,||\, \mathcal{Q}) = \mathbb{E}_{\mathcal{Q}}\left[f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right]$$

where $f(1) = 0$ and is convex

Name	$D_f(\mathcal{P} \,\|\, \mathcal{Q})$	**Generator** $f(t)$	**Symm.**	**Tri.**
Kullback-Leibler	$D_{\text{KL}}(\mathcal{P} \,\|\, \mathcal{Q})$	$t \log t$	✗	✗
Reverse-KL	$D_{\text{KL}}(\mathcal{Q} \,\|\, \mathcal{P})$	$-\log t$	✗	✗
Jensen-Shannon	$D_{\text{JS}}(\mathcal{P}, \mathcal{Q}) := \frac{1}{2}(D_{\text{KL}}(\mathcal{P}\|\mathcal{M}) + D_{\text{KL}}(\mathcal{Q}\|\mathcal{M}))$	$t \log t - (t+1)\log(\frac{t+1}{2})$	✓	✗
Squared Hellinger	$H^2(\mathcal{P}, \mathcal{Q}) := \frac{1}{2}\int(\sqrt{d\mathcal{P}} - \sqrt{d\mathcal{Q}})^2$	$(1-\sqrt{t})^2/2$	✓	✗
Total Variation	$d_{\text{TV}}(\mathcal{P}, \mathcal{Q}) := \sup_E \lvert\mathcal{P}(E) - \mathcal{Q}(E)\rvert$	$\lvert t-1\rvert/2$	✓	✓

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Approximate version exists as well, consider:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

We can also measure the tradeoff in terms of invariant representations:

(Informal) If $g_\sharp \mathcal{D}_0$ and $g_\sharp \mathcal{D}_1$ are sufficient close to each other, then:

Total variation lower bound:

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq d_{\mathrm{TV}}\left(\mathcal{D}_0(Y), \mathcal{D}_1(Y)\right) - d_{\mathrm{TV}}\left(g_\# \mathcal{D}_0, g_\# \mathcal{D}_1\right)$$

Jensen-Shannon lower bound:

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \left(d_{\mathrm{JS}}\left(\mathcal{D}_0(Y), \mathcal{D}_1(Y)\right) - d_{\mathrm{JS}}\left(g_\# \mathcal{D}_0, g_\# \mathcal{D}_1\right)\right)^2 / 2$$

Hellinger lower bound:

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \left(H\left(\mathcal{D}_0(Y), \mathcal{D}_1(Y)\right) - H\left(g_\sharp \mathcal{D}_0, g_\sharp \mathcal{D}_1\right)\right)^2 / 2$$

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS'19

Fairness vs Utility

Approximate version exists as well, consider:

$$X \xrightarrow{g} Z \xrightarrow{h} \hat{Y}$$

We can also measure the tradeoff in terms of invariant representations:

(Informal) If $g_\sharp \mathcal{D}_0$ and $g_\sharp \mathcal{D}_1$ are sufficient close to each other, then:

Total variation lower bound:

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq d_{\mathrm{TV}}\left(\mathcal{D}_0(Y), \mathcal{D}_1(Y)\right) - d_{\mathrm{TV}}\left(g_{\#}\mathcal{D}_0, g_{\#}\mathcal{D}_1\right)$$

Jensen-Shannon lower bound:

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \left(d_{\mathrm{JS}}\left(\mathcal{D}_0(Y), \mathcal{D}_1(Y)\right) - d_{\mathrm{JS}}\left(g_{\#}\mathcal{D}_0, g_{\#}\mathcal{D}_1\right)\right)^2 / 2$$

Hellinger lower bound:

$$\mathrm{Err}_0(h \circ g) + \mathrm{Err}_1(h \circ g) \geq \left(H\left(\mathcal{D}_0(Y), \mathcal{D}_1(Y)\right) - H\left(g_\sharp\mathcal{D}_0, g_\sharp\mathcal{D}_1\right)\right)^2 / 2$$

The more invariant the representations, the worse the joint error

Inherent Tradeoffs in Learning Fair Representations, Zhao et al., NeurIPS’19

Experiments

Income Prediction: Adult dataset

- Train/Test: 30,162/15,060 adults information collected in a 1994 census
- Target variable: $Y = 1$ iff annual income > 50K
- Sensitive variable: $A = 0/1$ = Male/Female
- Other attributes: age, education, etc.

- Base rates are different across groups:

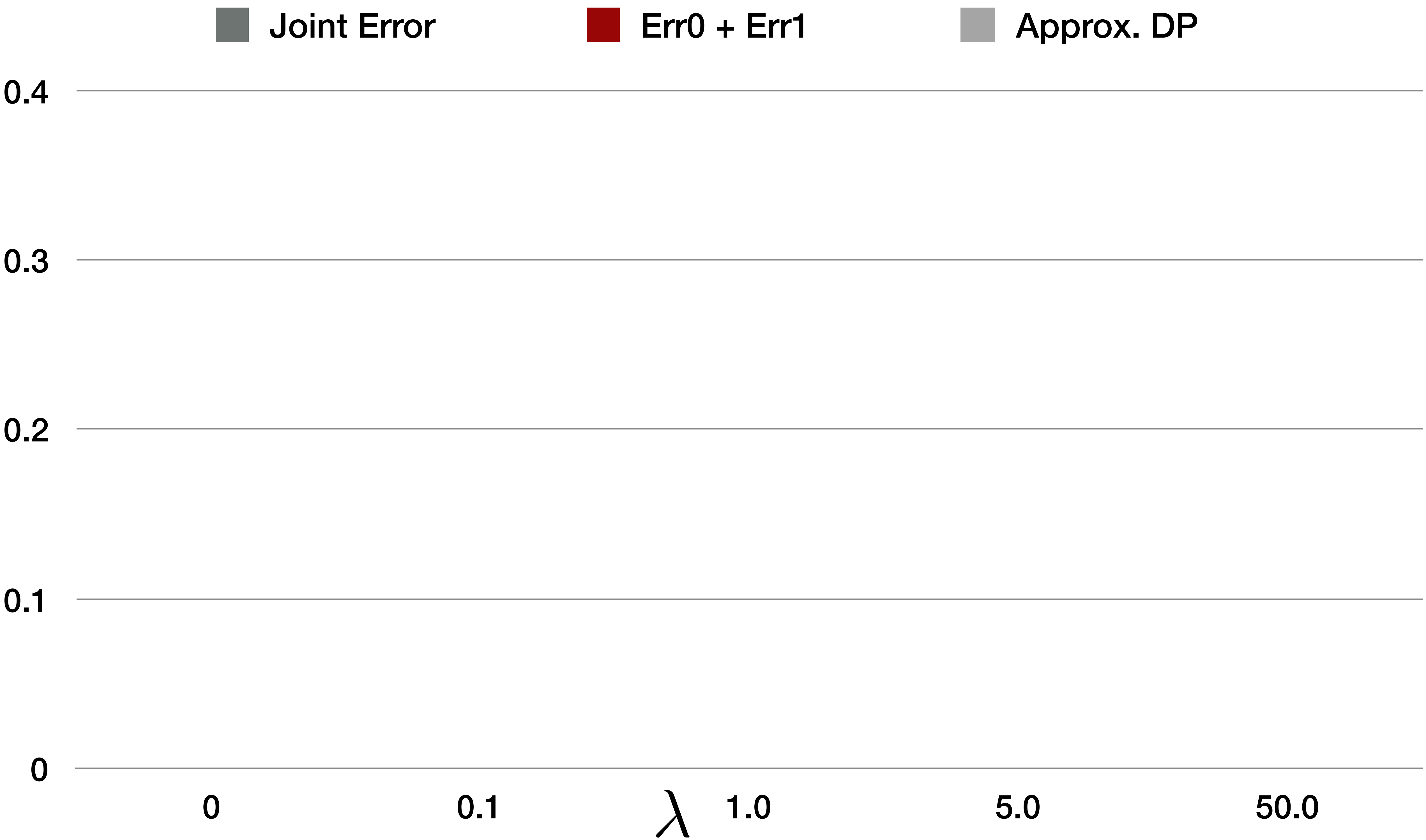
$$\Pr(Y = 1 \mid A = 0) = 0.310 \qquad \Pr(Y = 1 \mid A = 1) = 0.113$$

- Imbalanced marginal distribution:

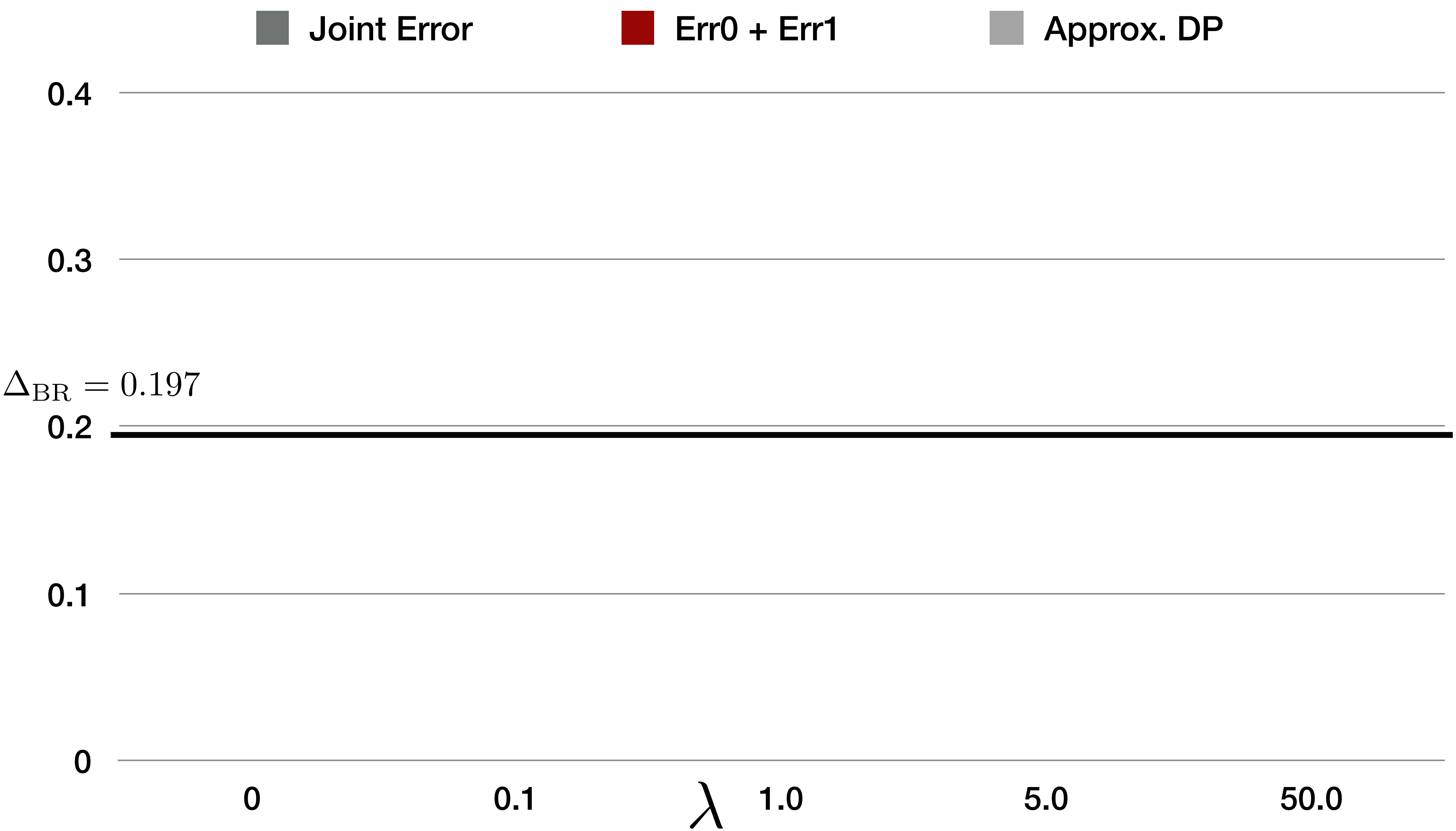
$$\Pr(A = 0) = 0.673$$

Experiments

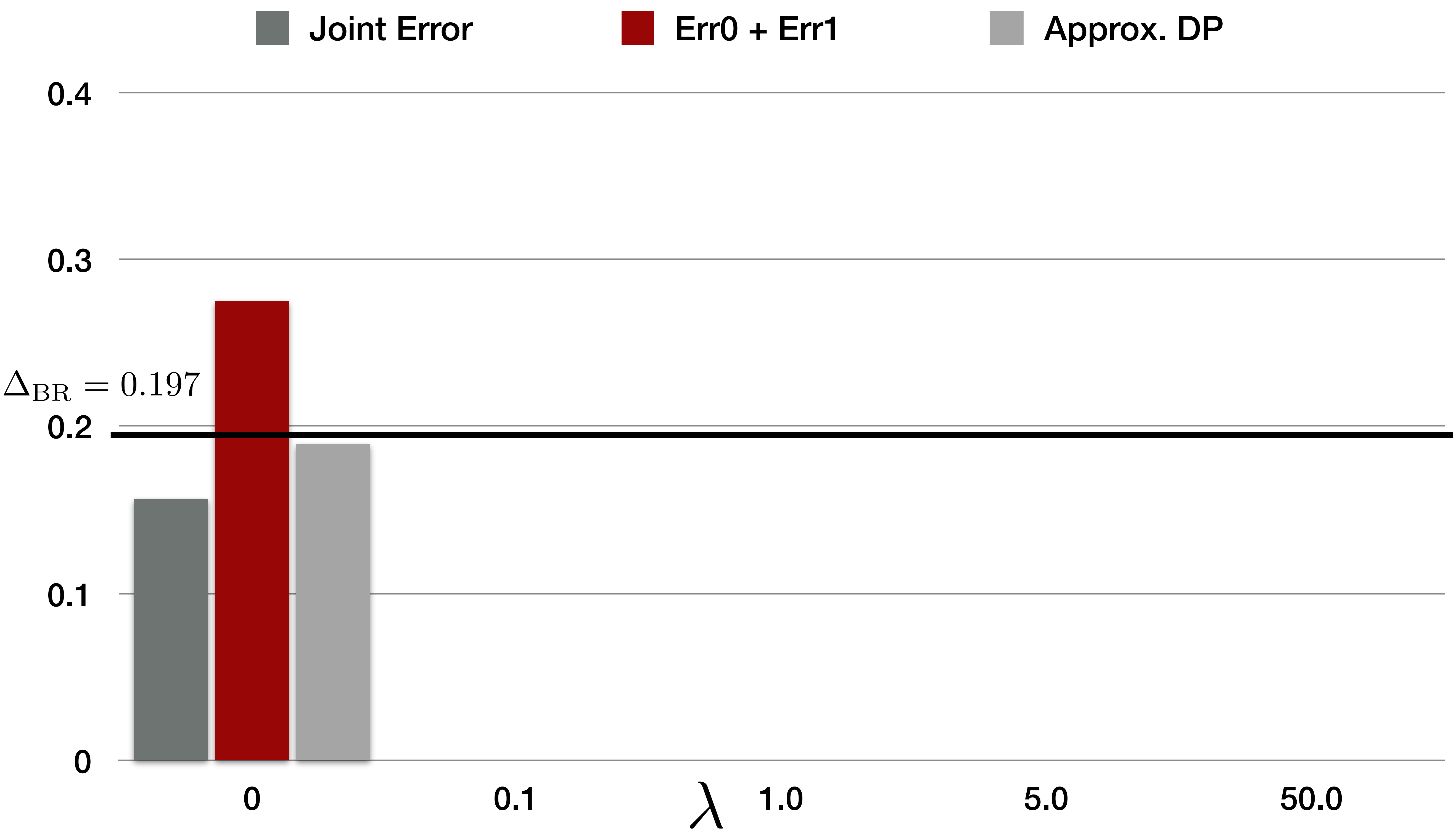
Experiments



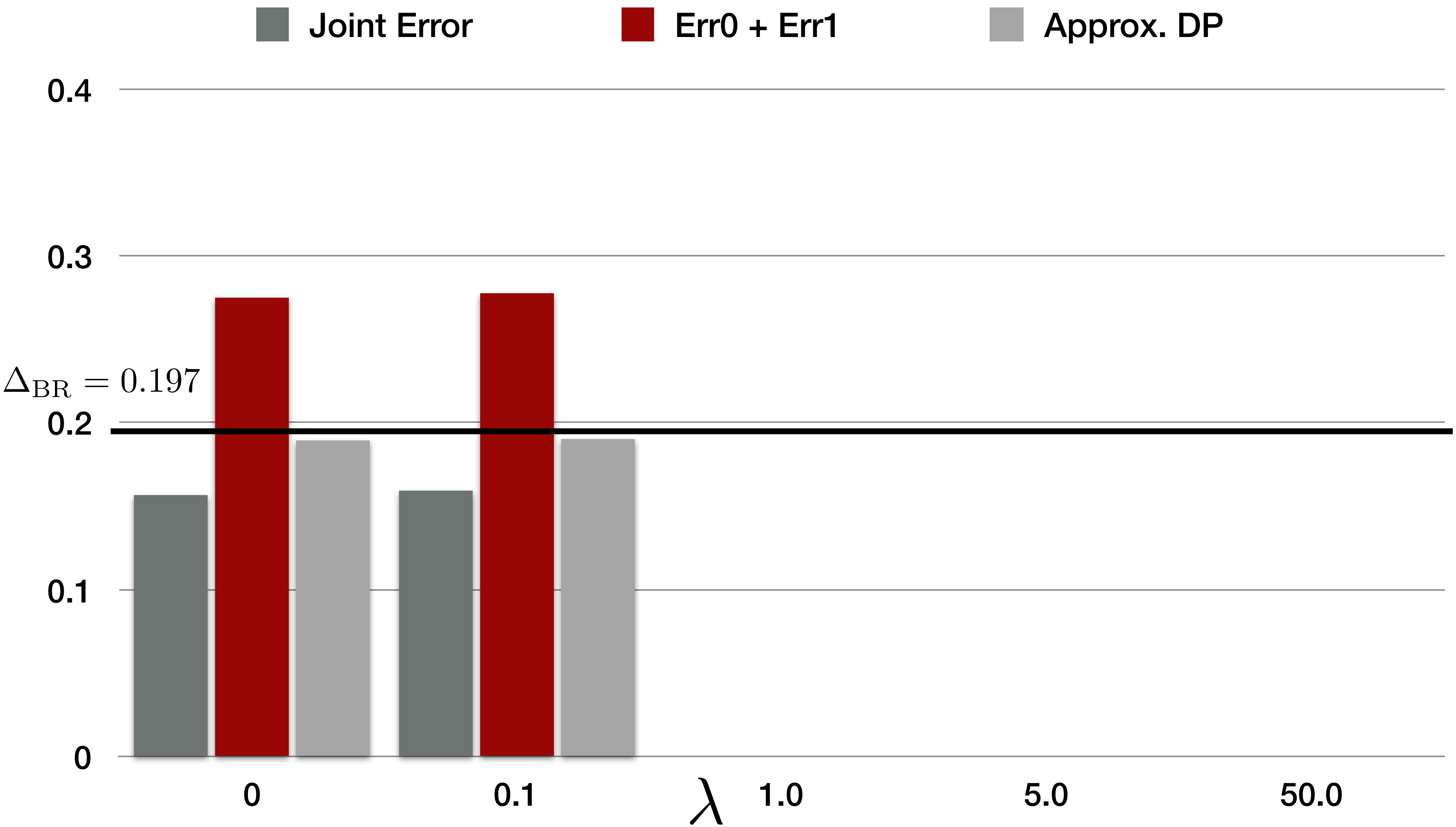
Experiments



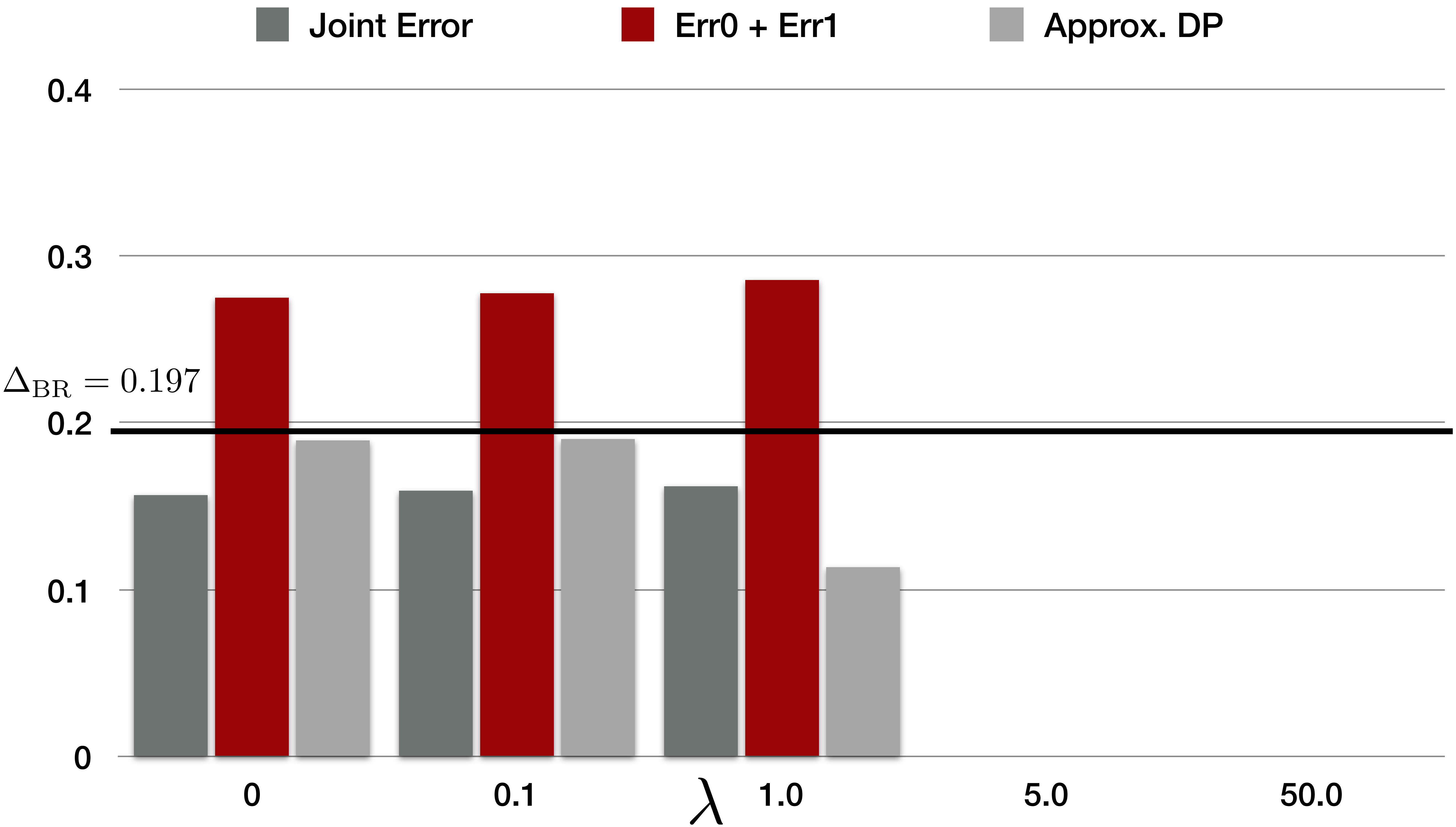
Experiments



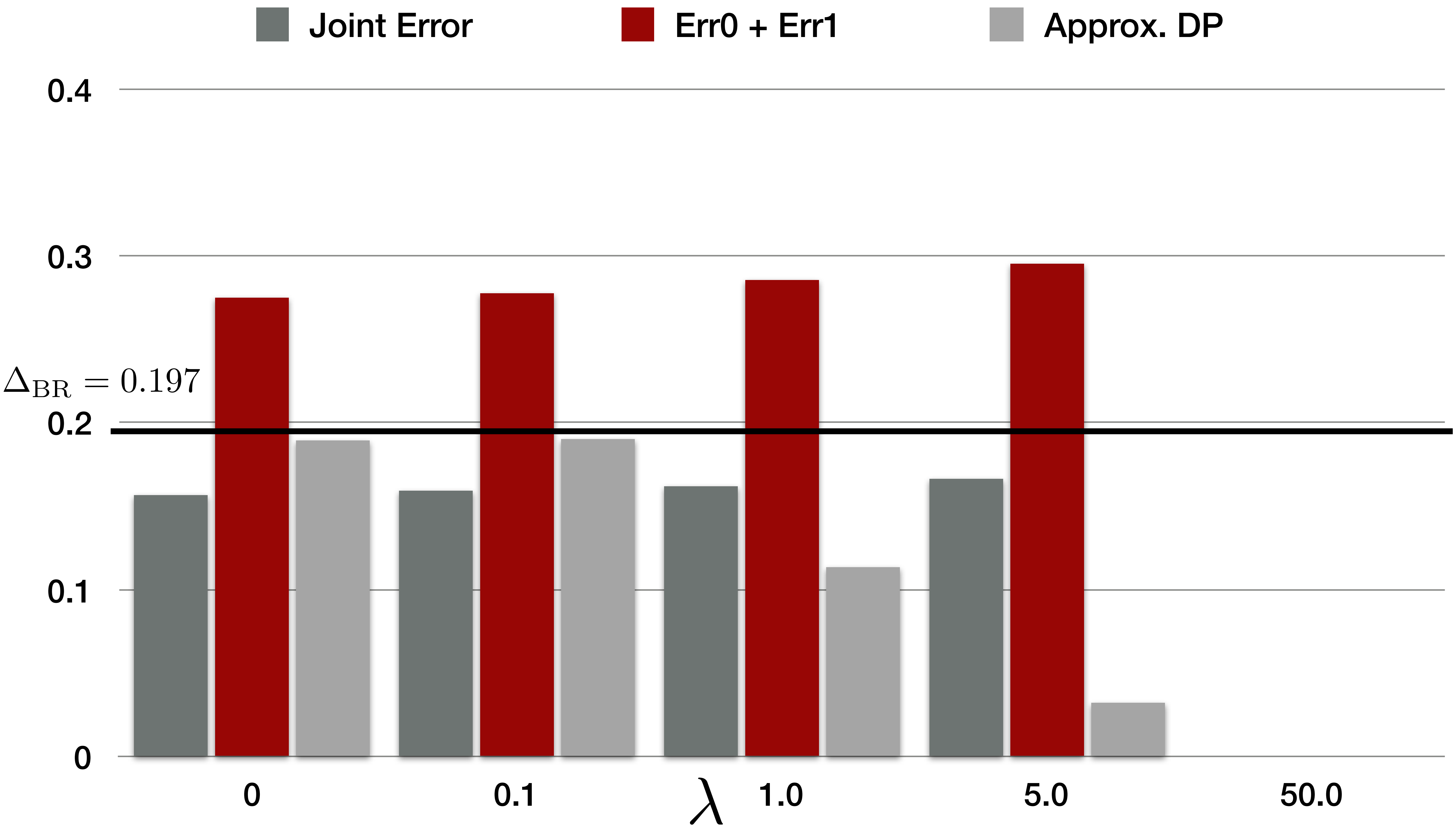
Experiments



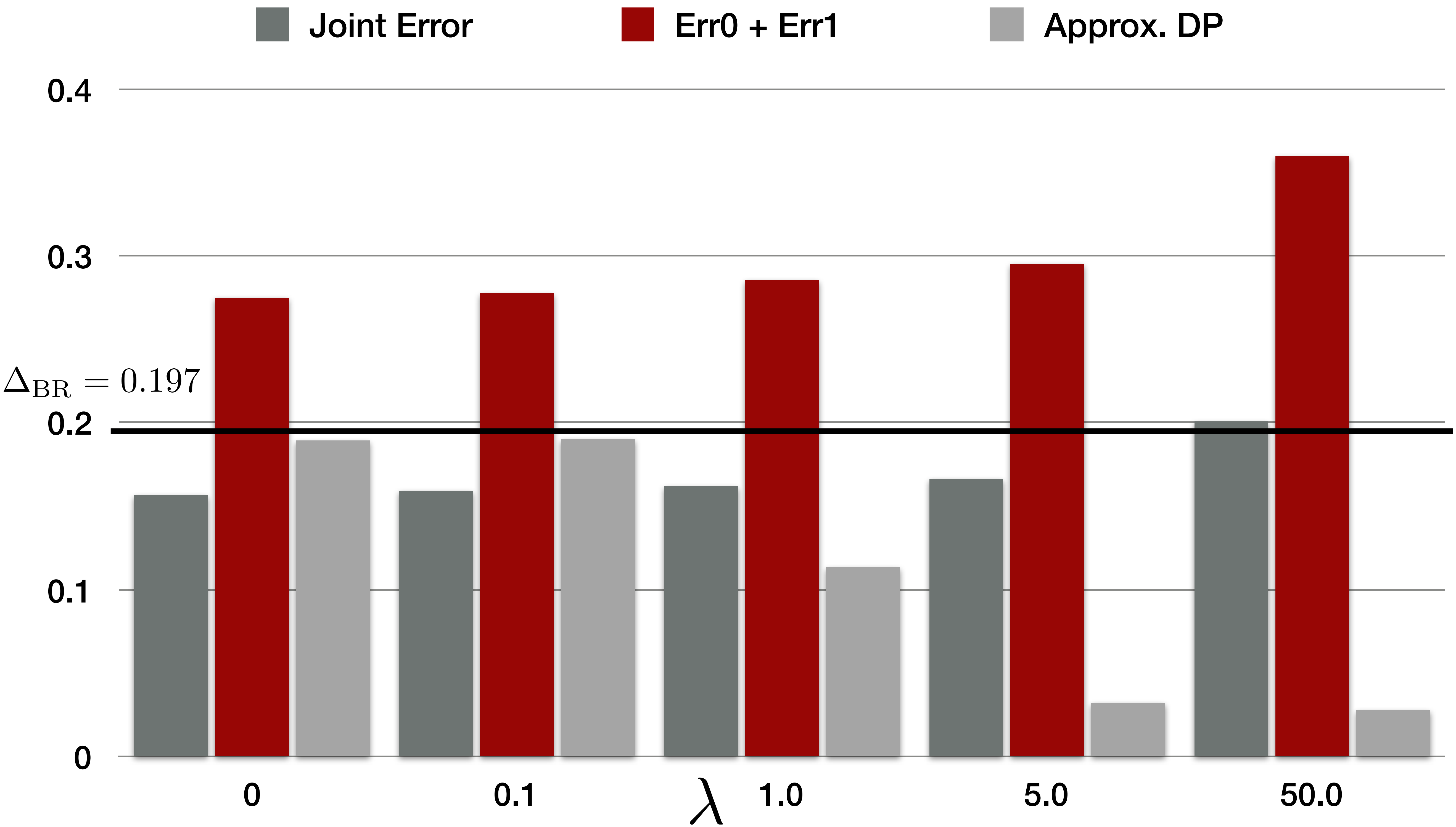
Experiments



Experiments



Experiments



Thanks

Q & A