COMPUTATIONAL CHALLENGES IN DATA PRIVACY

FAIRNESS IN RISK ASSESSMENT

JULY 06, 2025 | ROHIT VAISH


input
ALGORITHM
output

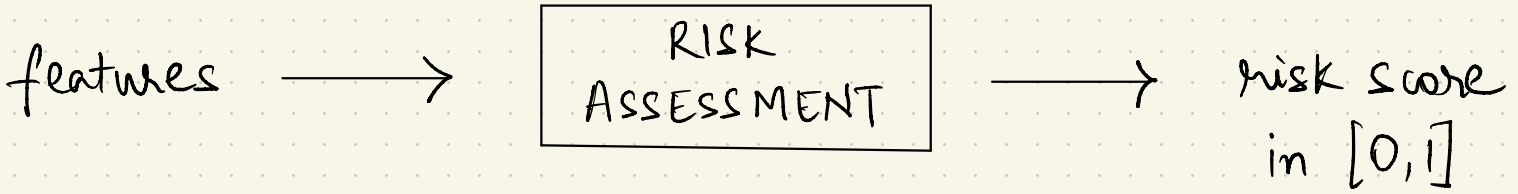

a collection
of numbers
MERGE SORT
sorted list
of numbers

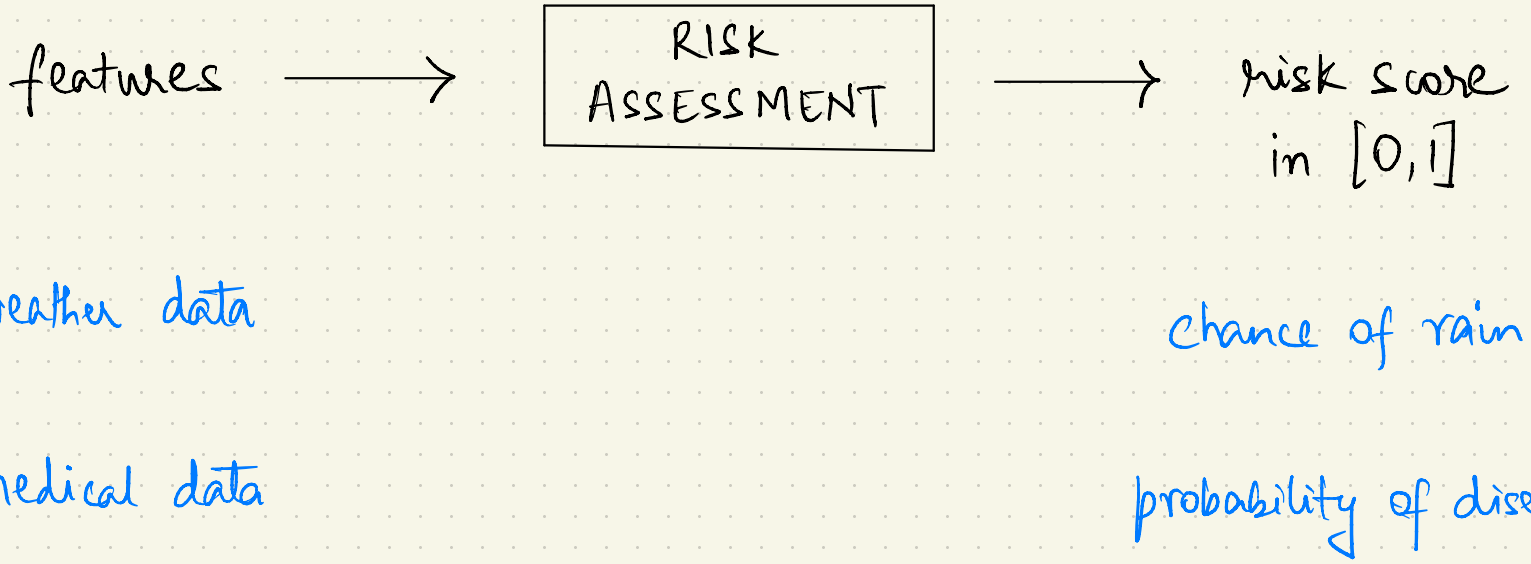

a graph/network with edge weights
DIJKSTRA
shortest path between vertices


browsing histry
or Cookies
RECOMMENDATION
job ads


Credit history
BANKS
loan decision


Criminal records
COURTS
bail decision


features
RISK ASSESSMENT
risk score in [0,1]


features
RISK ASSESSMENT
risk score in [0,1]
weather data
medical data
chance of rain
probability of disease

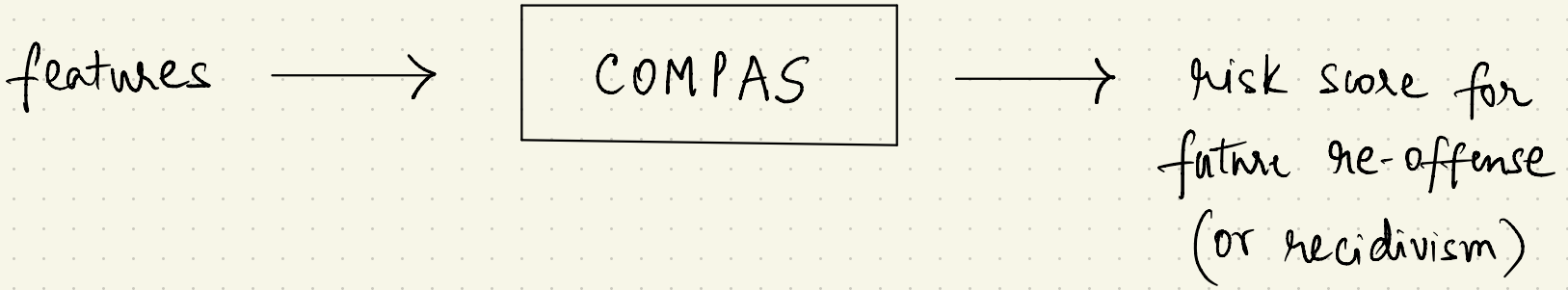
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

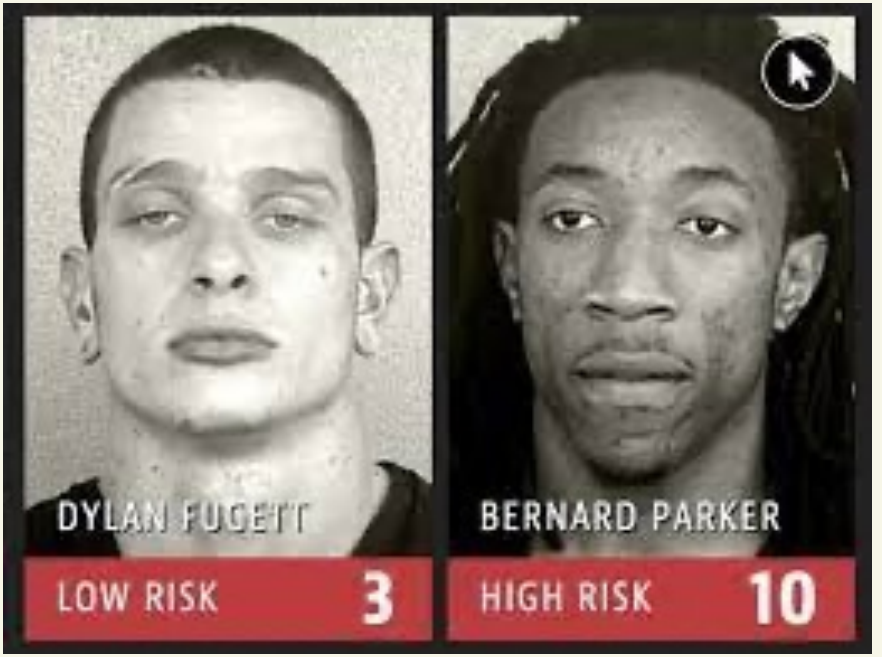
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Correctional Offender Management Profiling for Alternative Sanctions



ProPublica's Concern



ProPublica's Concern

	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%

ProPublica's Concern

	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%

African-American defendants more likely to be incorrectly labeled high-risk

White defendants more likely to be incorrectly labeled low-risk

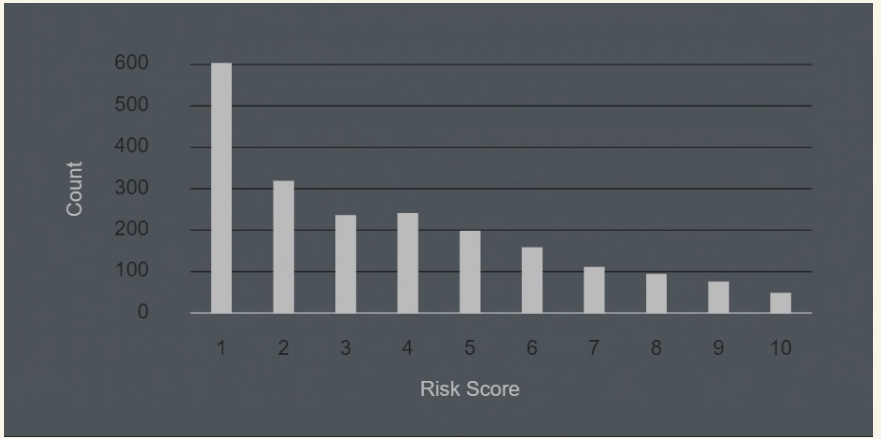
ProPublica's Concern

	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%

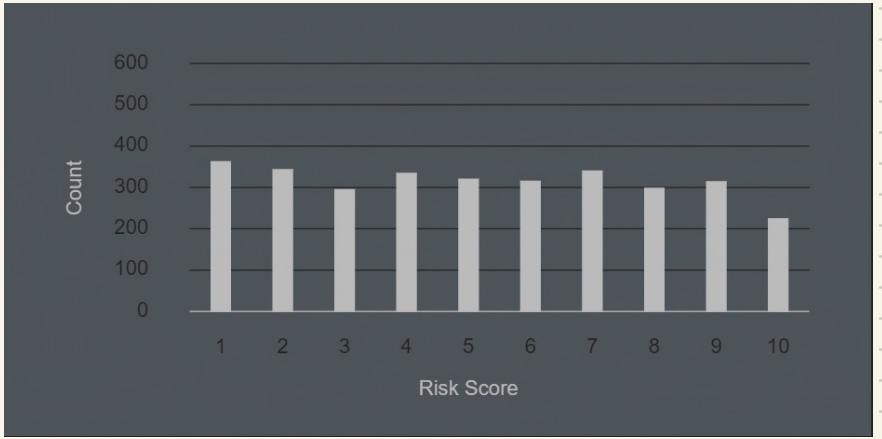
African-American defendants more likely to be incorrectly labeled high-risk

White defendants more likely to be incorrectly labeled low-risk

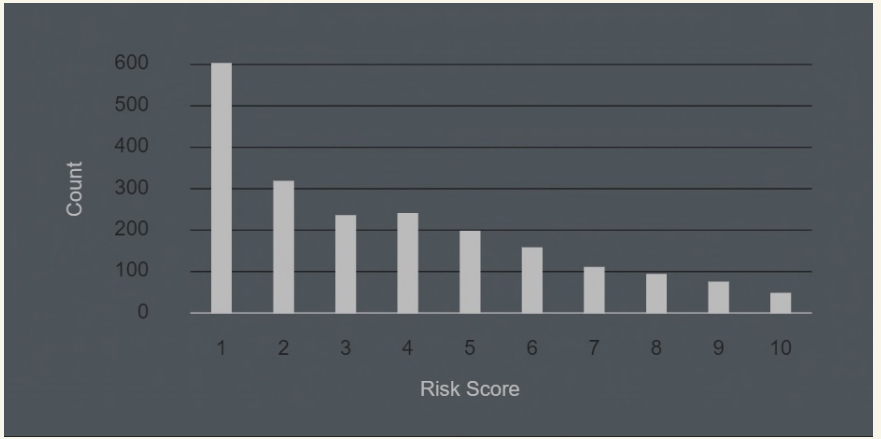
Imbalanced error rates



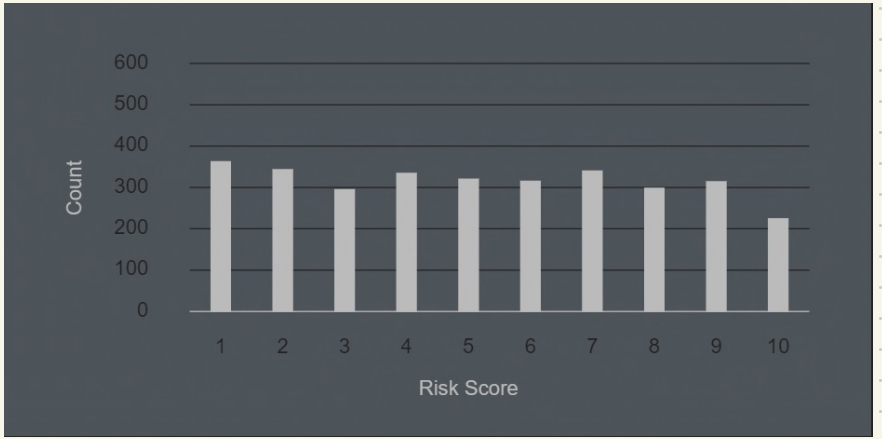
Black or White?



Black or White?


600
500
400
300
200
100
0
Count
1
2
3
4
5
6
7
8
9
10
Risk Score

White


600
500
400
300
200
100
0
Count
1
2
3
4
5
6
7
8
9
10
Risk Score

Black

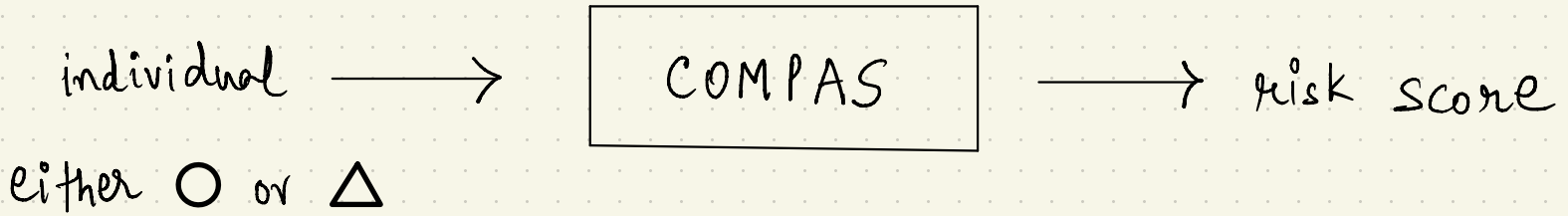
Balanced error rates

BALANCED ERROR RATES

Two types/groups of individuals : ○ and △

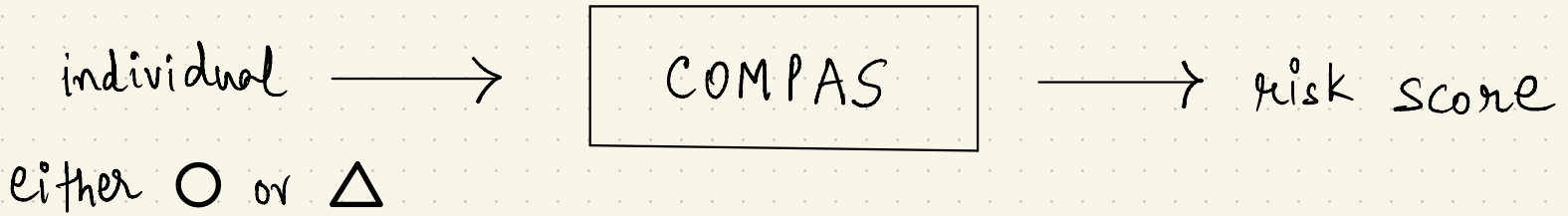
BALANCED ERROR RATES

Two types/groups of individuals : ○ and △



BALANCED ERROR RATES

Two types/groups of individuals : ○ and △



Actual outcomes

○ : ○ reoffends

○ : ○ doesn't reoffend

△ : △ reoffends

△ : △ doesn't reoffend

BALANCED ERROR RATES

Balance for red class :

avg. score of ○s = avg. score of △s

BALANCED ERROR RATES

Balance for red class :

avg. score of ○s = avg. score of △s

subject to reoffending, the score of a typical ○ is same as
" " " " △.

BALANCED ERROR RATES

Balance for red class:

avg. score of ○s = avg. score of △s

subject to reoffending, the score of a typical ○ is same as " " " " △.

Balance for green class:

avg. score of ○s = avg. score of △s

BALANCED ERROR RATES

Balance for red class:

avg. score of ○s = avg. score of △s

subject to reoffending, the score of a typical ○ is same as " " " " △.

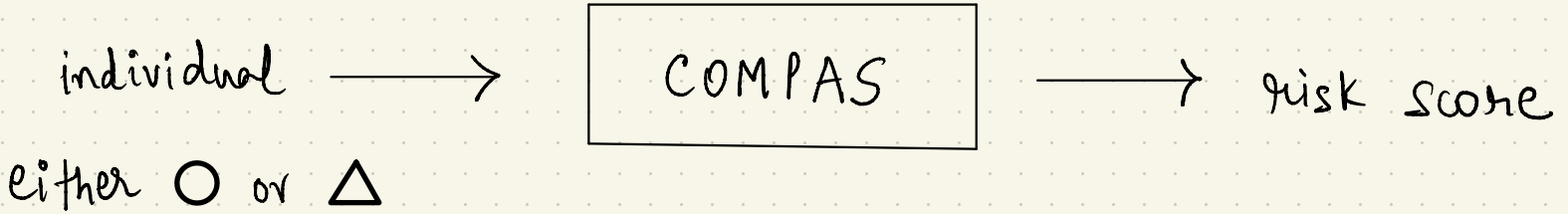
Balance for green class:

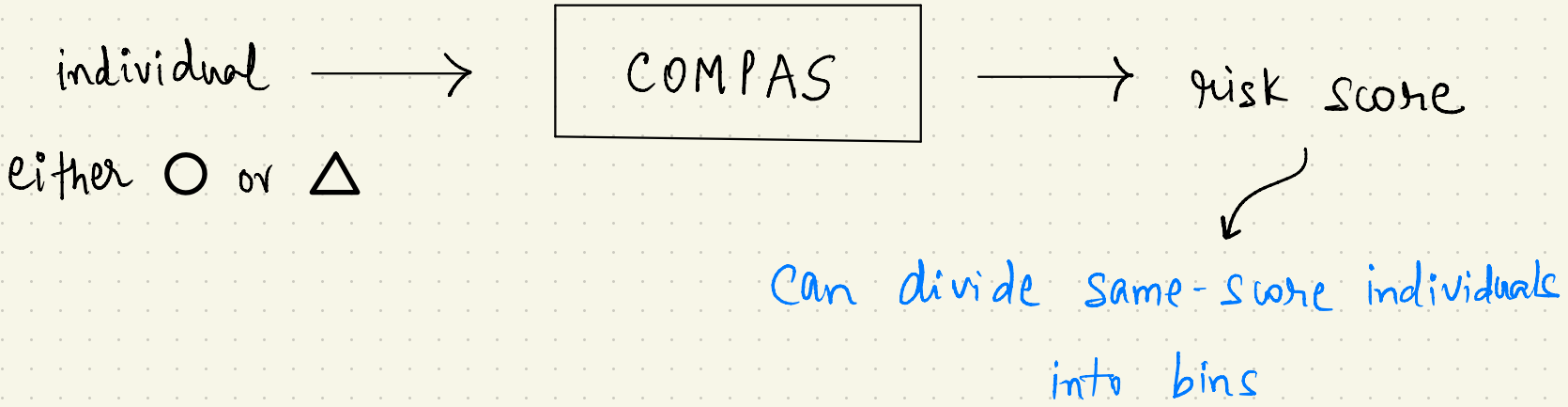
avg. score of ○s = avg. score of △s

subject to not reoffending, the score of a typical ○ is same as " " " " △.

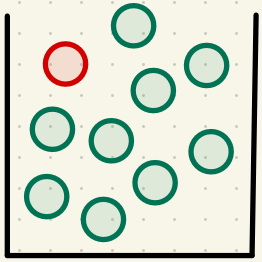
COMPAS' Rebuttal

Calibration

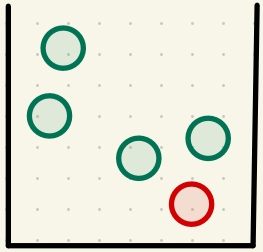

individual
either ○ or △
COMPAS
risk score


individual
either ○ or △
COMPAS
risk score
Can divide same-score individuals into bins

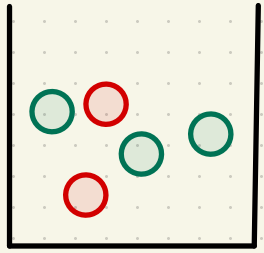
CALIBRATION



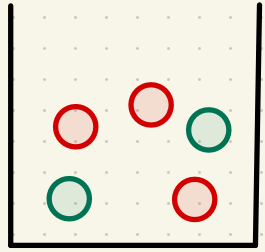
0.1



0.2

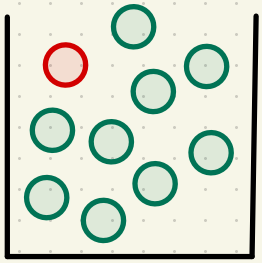


0.4

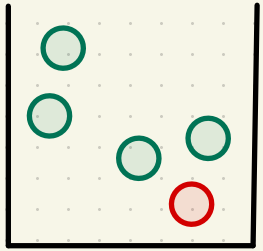


0.6

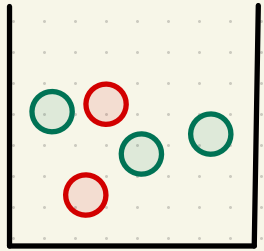
CALIBRATION



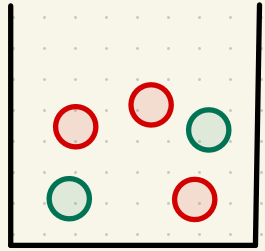
0.1



0.2



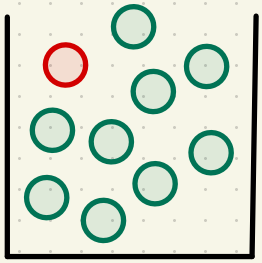
0.4



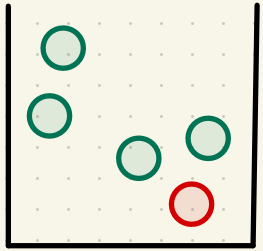
0.6

0.2 means 20% regardless of shape

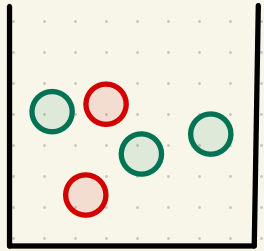
CALIBRATION



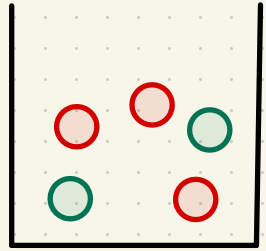
0.1



0.2



0.4

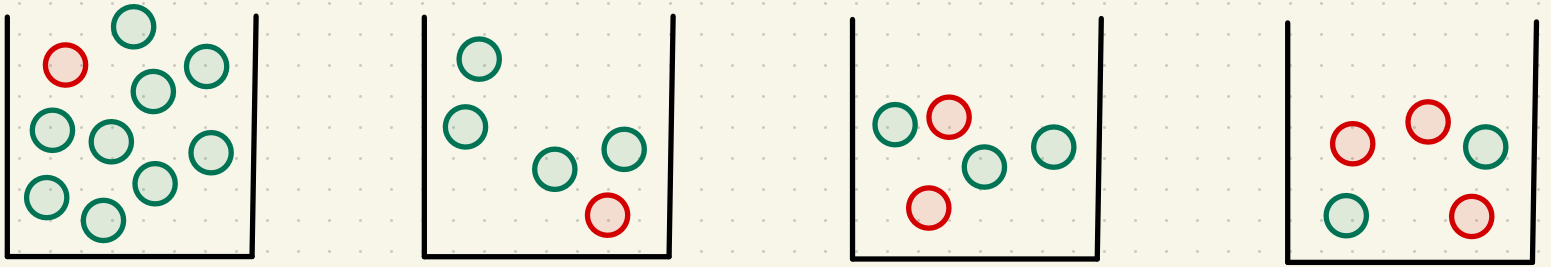


0.6

0.2 means 20% regardless of shape

(same holds for △ △)

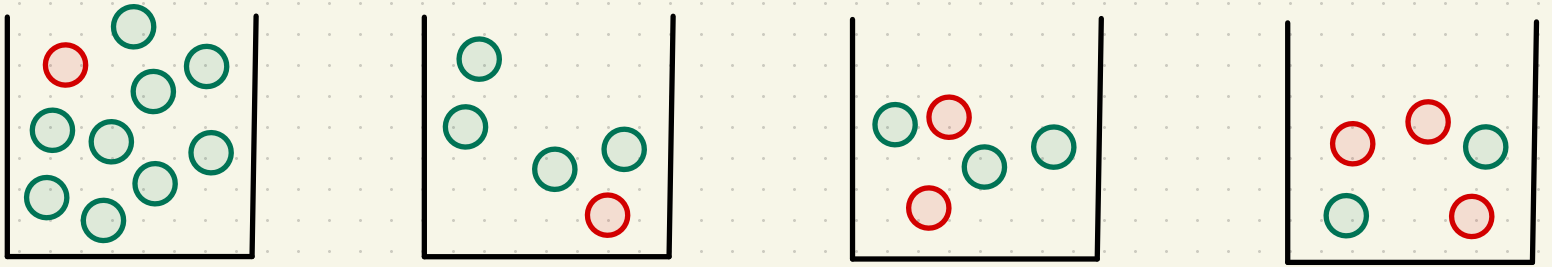
CALIBRATION



0.1 0.2 0.4 0.6

Why is calibration meaningful?

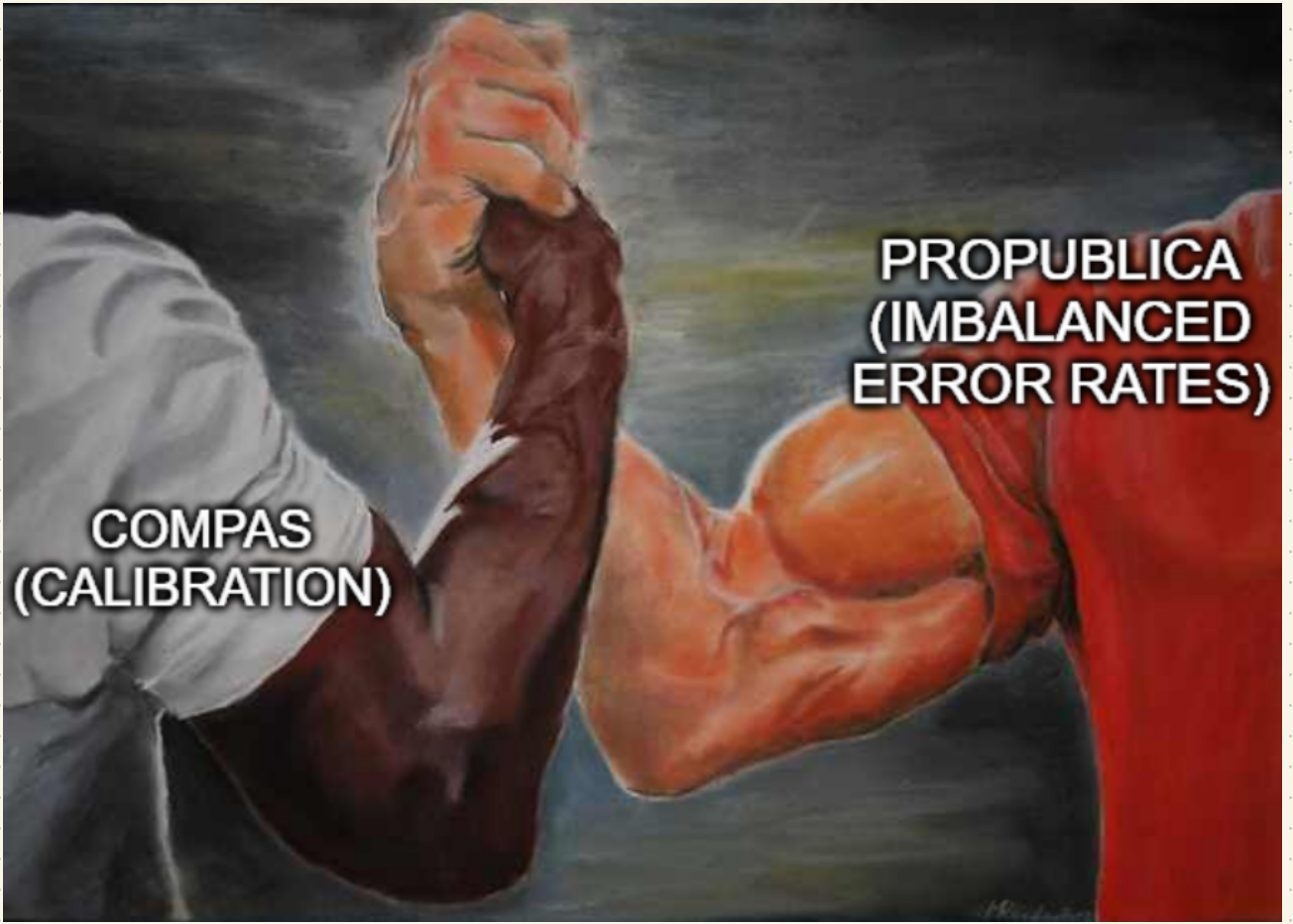
CALIBRATION



0.1 0.2 0.4 0.6

Why is calibration meaningful?

Treat same-score individuals similarly regardless of shape.

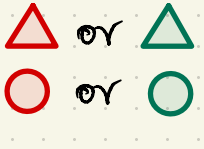

PROPUBLICA
(IMBALANCED
ERROR RATES)
COMPAS
(CALIBRATION)

Can we have the best of both worlds?

Balanced error rates + calibration

BEST OF BOTH WORLDS?

BEST OF BOTH WORLDS?

Perfect prediction : Can correctly predict **only**  from the features.

BEST OF BOTH WORLDS?

Perfect prediction : Can correctly predict **only**  from the features.

Calibration ?

BEST OF BOTH WORLDS?

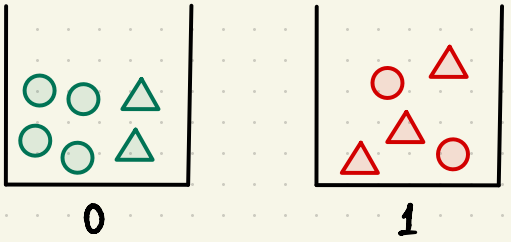
Perfect prediction : Can correctly predict **only**  from the features.

Calibration ? Yes !

BEST OF BOTH WORLDS?

Perfect prediction : Can correctly predict **only**  from the features.

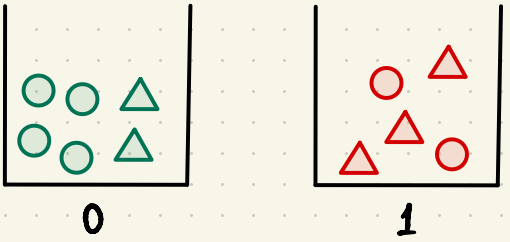
Calibration ? Yes !



BEST OF BOTH WORLDS?

Perfect prediction : Can correctly predict **only**  from the features.

Calibration ? Yes !

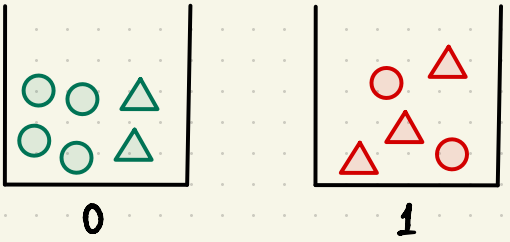


Balanced error rates?

BEST OF BOTH WORLDS?

Perfect prediction : Can correctly predict **only**  from the features.

Calibration ? Yes !

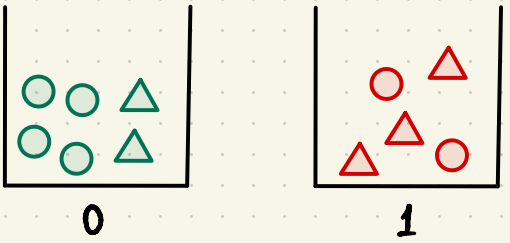


Balanced error rates? Yes !

BEST OF BOTH WORLDS?

Perfect prediction: Can correctly predict **only** △ or △ ○ or ○ from the features.

Calibration? Yes!



Balanced error rates? Yes!

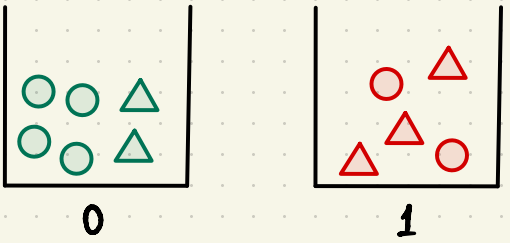
avg score of △s = avg score of ○s

" " △s = " " ○s

BEST OF BOTH WORLDS?

Perfect prediction ✓ : Can correctly predict **only** △ or △ / ○ or ○ from the features.

Calibration? Yes!



Balanced error rates? Yes!

avg score of △s = avg score of ○s

" " △s = " " ○s

BEST OF BOTH WORLDS?

Equal base rates: when $\frac{\#\,\triangle s}{\#\,\triangle s} = \frac{\#\,\bigcirc s}{\#\,\bigcirc s} = p(\text{say})$

BEST OF BOTH WORLDS?

Equal base rates: when $\frac{\#\,\triangle s}{\#\,\triangle s} = \frac{\#\,\bigcirc s}{\#\,\bigcirc s} = p(\text{say})$

Calibration?

BEST OF BOTH WORLDS?

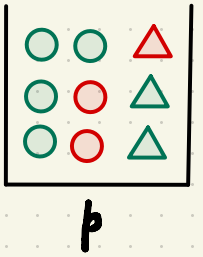
Equal base rates: when $\frac{\#\,\triangle\text{s}}{\#\,\triangle\text{s}} = \frac{\#\,\bigcirc\text{s}}{\#\,\bigcirc\text{s}} = p(\text{say})$

Calibration? Yes!

BEST OF BOTH WORLDS?

Equal base rates: when $\frac{\#\,\triangle\text{s}}{\#\,\triangle\text{s}} = \frac{\#\,\bigcirc\text{s}}{\#\,\bigcirc\text{s}} = p(\text{say})$

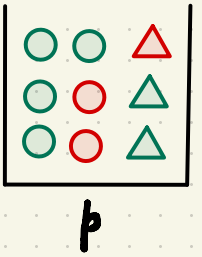
Calibration? Yes!



BEST OF BOTH WORLDS?

Equal base rates: when $\frac{\#\,\triangle\text{s}}{\#\,\triangle\text{s}} = \frac{\#\,\bigcirc\text{s}}{\#\,\bigcirc\text{s}} = p(\text{say})$

Calibration? Yes!

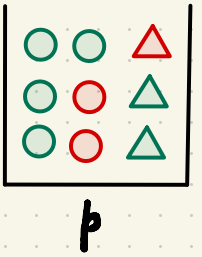


Balanced error rates?

BEST OF BOTH WORLDS?

Equal base rates: when $\frac{\#\,\triangle\text{s}}{\#\,\triangle\text{s}} = \frac{\#\,\bigcirc\text{s}}{\#\,\bigcirc\text{s}} = p(\text{say})$

Calibration? Yes!

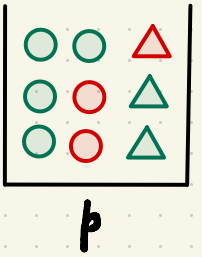


Balanced error rates? Yes!

BEST OF BOTH WORLDS?

Equal base rates: when $\frac{\#\,\triangle\text{s}}{\#\,\triangle\text{s}} = \frac{\#\,\bigcirc\text{s}}{\#\,\bigcirc\text{s}} = p\,(\text{say})$

Calibration? Yes!



p

Balanced error rates? Yes!

avg score of △s = avg score of ○s = p

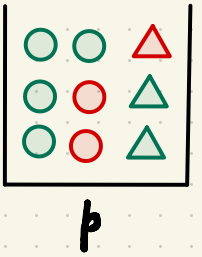
" " △s = " " ○s = p

BEST OF BOTH WORLDS?

✔

Equal base rates: when $\frac{\#\,\triangle\text{s}}{\#\,\triangle\text{s}} = \frac{\#\,\bigcirc\text{s}}{\#\,\bigcirc\text{s}} = p\,(\text{say})$

Calibration? Yes!



Balanced error rates? Yes!

avg score of $\triangle$s = avg score of $\bigcirc$s = p

" " $\triangle$s = " " $\bigcirc$s = p

BEST OF BOTH WORLDS?

	Calibration	Balanced error rates
Perfect prediction	✓	✓
Equal base rates	✓	✓

BEST OF BOTH WORLDS?

	Calibration	Balanced error rates
Perfect prediction	✓	✓
Equal base rates	✓	✓

Any others?

BEST OF BOTH WORLDS?

	Calibration	Balanced error rates
Perfect prediction	✓	✓
Equal base rates	✓	✓

Any others?

NO!

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Theorem: Any risk assessment method that satisfies calibration and balanced error rates must either be perfect or have equal base rates.

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Theorem: Any risk assessment method that satisfies calibration and balanced error rates must either be perfect or have equal base rates.

* Not about computational power

* Impossibility of "balancing averages" by assigning scores

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

individual $\longrightarrow$ [risk assessment method (eg., COMPAS)] $\longrightarrow$ risk score

either ○ or △

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

$n_{\bigcirc} = \#\bigcirc + \#\bigcirc$

$n_{\triangle} = \#\triangle + \#\triangle$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch: By calibration:

Bin with score b_1:

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch: By calibration:

Bin with score b_1: $b_1 \times [\#\Delta s \text{ in bin } b_1] = \#\Delta s \text{ in bin } b_1$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch: By calibration:

Bin with score b_1: $b_1 \times [\#\Delta s \text{ in bin } b_1] = \#\Delta s \text{ in bin } b_1$

Bin with score b_2: $b_2 \times [\#\Delta s \text{ in bin } b_2] = \#\Delta s \text{ in bin } b_2$

$\vdots$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch: By calibration:

Bin with score b_1: $b_1 \times [\#\Delta s \text{ in bin } b_1] = \#\Delta s \text{ in bin } b_1$

Bin with score b_2: $b_2 \times [\#\Delta s \text{ in bin } b_2] = \#\Delta s \text{ in bin } b_2$

$\vdots$

Summing over bins: Total score of all $\Delta s = \#\Delta s$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch: By calibration:

Bin with score b_1: $b_1 \times [\#\,\Delta s \text{ in bin } b_1] = \#\,\Delta s \text{ in bin } b_1$

Bin with score b_2: $b_2 \times [\#\,\Delta s \text{ in bin } b_2] = \#\,\Delta s \text{ in bin } b_2$

$\vdots$

Summing over bins: Total score of all Δs $=$ $\#\,\Delta s$

Similarly, Total score of all $\circ s$ $=$ $\#\,\circ s$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch: By calibration:

Bin with score b_1: $b_1 \times [\#\,\Delta s \text{ in bin } b_1] = \#\,\Delta s \text{ in bin } b_1$

Bin with score b_2: $b_2 \times [\#\,\Delta s \text{ in bin } b_2] = \#\,\Delta s \text{ in bin } b_2$

$\vdots$

Summing over bins: Total score of all Δs $=$ $\#\,\Delta s$

Similarly, Total score of all $\circ s$ $=$ $\#\,\circ s$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

r = avg score of all red shapes

g = " " " green "

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch: r = avg score of all red shapes

g = " " " green "

By balanced error rates:

avg score of △s = avg score of ○s = r

avg score of △s = avg score of ○s = g

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

Total score of all Δs = Total score of Δs + Total score of Δs

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

Total score of all Δs = Total score of Δs + Total score of Δs

$= \#\Delta s \cdot r + \#\Delta s \cdot g$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

Total score of all $\triangle$s = Total score of $\triangle$s + Total score of $\triangle$s

$$= \#\triangle s \cdot r + \#\triangle s \cdot g$$

$$= \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

Total score of all $\triangle$s = Total score of $\triangle$s + Total score of $\triangle$s

$$= \#\triangle s \cdot r + \#\triangle s \cdot g$$

$$= \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$= \#\triangle s$$ by ☺

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

Similarly,

$$g = \frac{\#\circ s \ (1-r)}{n_\circ - \#\circ s}$$

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

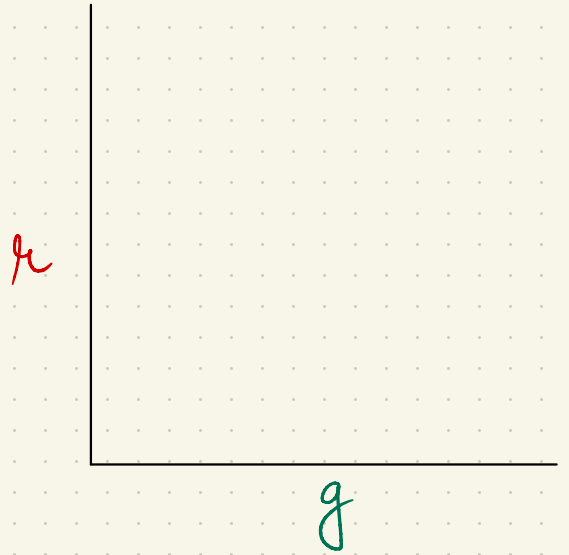
Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

Similarly,

$$g = \frac{\#\circ s \ (1-r)}{n_\circ - \#\circ s}$$



[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

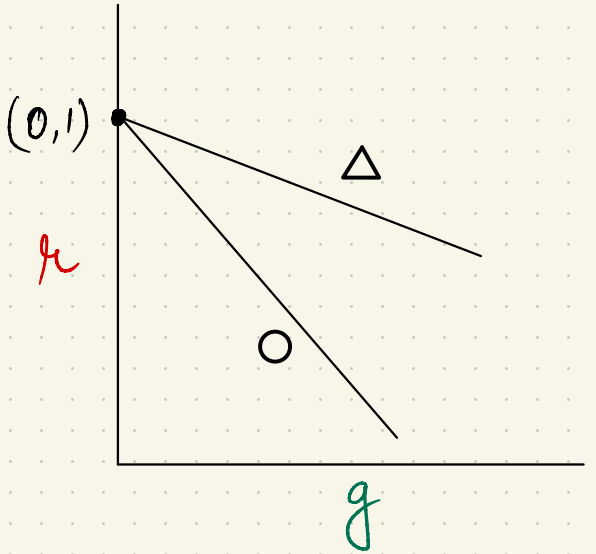
Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

Similarly,

$$g = \frac{\#\circ s \ (1-r)}{n_\circ - \#\circ s}$$



[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

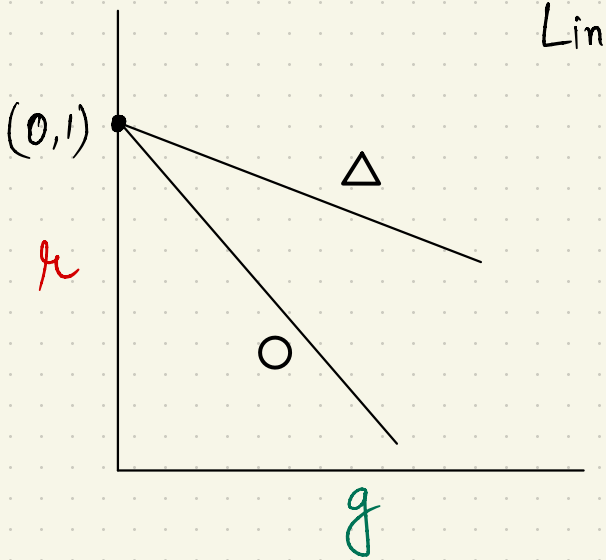
Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

Similarly,

$$g = \frac{\#\circ s \ (1-r)}{n_\circ - \#\circ s}$$



[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

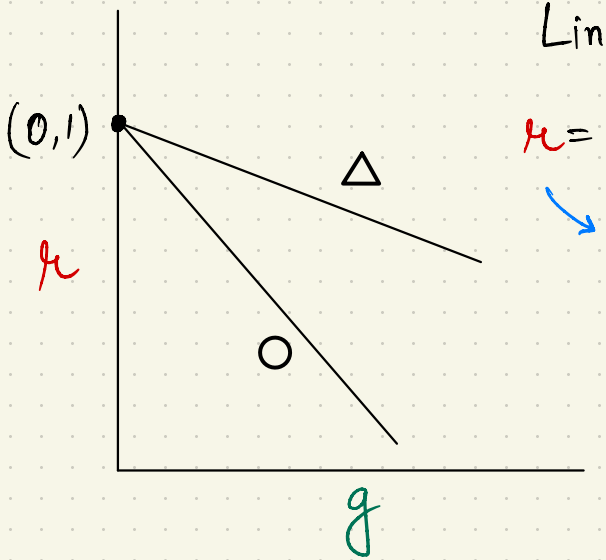
Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

Similarly,

$$g = \frac{\#\circ s \ (1-r)}{n_\circ - \#\circ s}$$



Lines intersect when

$r = 1, g = 0$

↘ perfect prediction

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

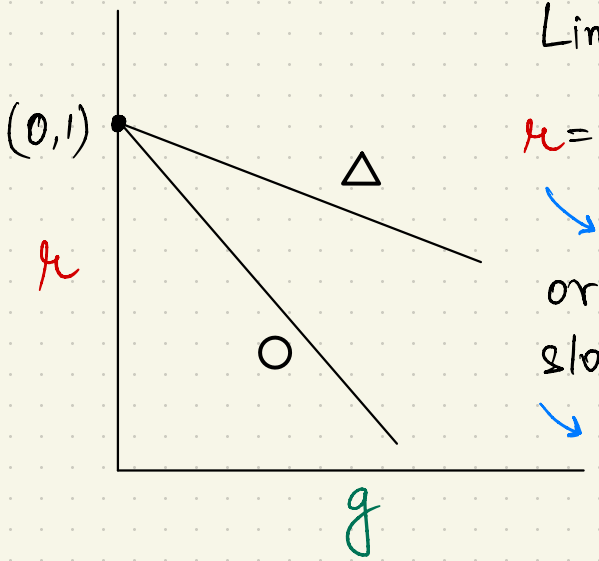
Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

Similarly,

$$g = \frac{\#\circ s \ (1-r)}{n_\circ - \#\circ s}$$



Lines intersect when $r=1, g=0$

↘ perfect prediction

or

slopes are equal

↘ equal base rates

[Kleinberg, Mullainathan, Raghavan ITCS 2017]

Thm: Calibration and balanced error rates $\Rightarrow$ perfect prediction or equal base rates

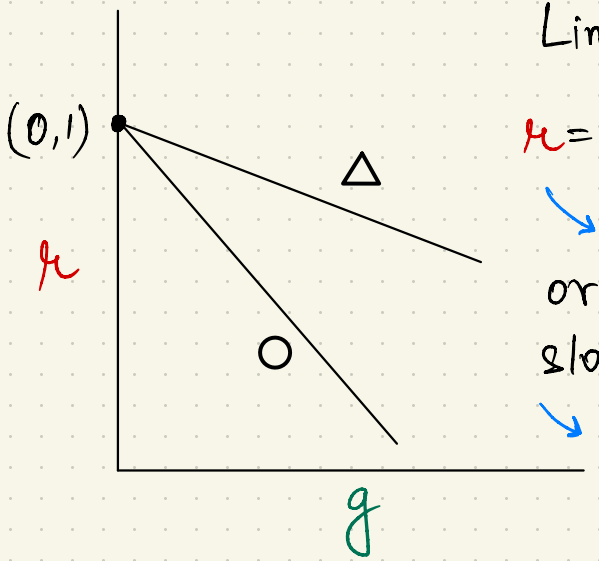
Proof sketch:

$$\#\triangle s = \#\triangle s \cdot r + (n_\triangle - \#\triangle s) \cdot g$$

$$\Rightarrow g = \frac{\#\triangle s \ (1-r)}{n_\triangle - \#\triangle s}$$

Similarly,

$$g = \frac{\#\circ s \ (1-r)}{n_\circ - \#\circ s}$$



Lines intersect when $r=1$, $g=0$

↘ perfect prediction

or

slopes are equal

↘ equal base rates

∎

EXTENSIONS

* Approximate calibration and approx. balanced error rates also in conflict.

EXTENSIONS

* Approximate calibration and approx. balanced error rates also in conflict.

* Binary classification [Chouldechova 2017; Corbett-Davies, Pierson, Feller, Goel 2016]

EXTENSIONS

* Approximate calibration and approx. balanced error rates also in conflict.

* Binary classification [Chouldechova 2017; Corbett-Davies, Pierson, Feller, Goel 2016]

* Workarounds
  - balanced error rates (w/o calibration) [Hardt, Price, Srebro 2016]
  - Calibration and one-sided error only

FIND OUT MORE AT

"Fairness and Machine Learning: Limitations and Oppurnities"

Book by Solon Barocas, Moritz Hardt and Arvind Narayanan

https://fairmlbook.org