Recidivism Case Study

Allen Downey
Olin College

For an editable version of these slides, contact downey@allendowney.com

Part 1: Machine Bias

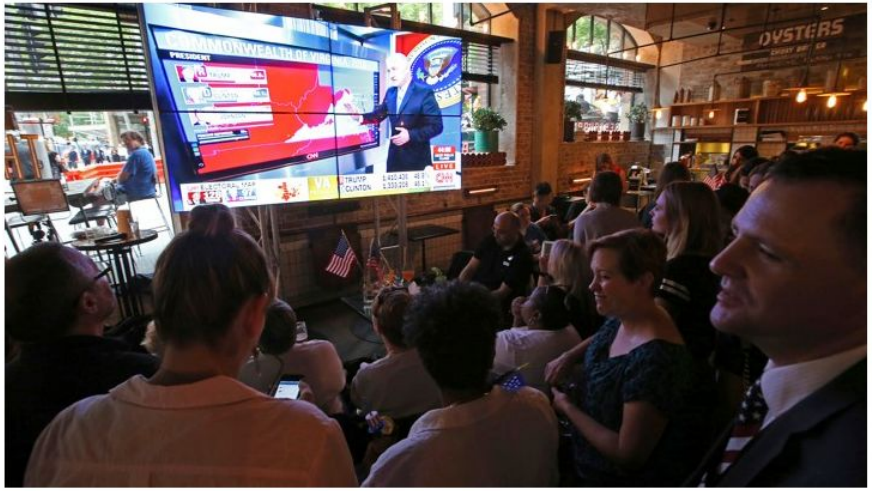
The love affair with data science ended in 2016.

POLITICS

What Went Wrong With the 2016 Polls?

Here's how public-opinion surveys and election forecasters misread the outcome of the presidential race.

VANN R. NEWKIRK II NOVEMBER 9, 2016



RICK RYCROFT / AP

Donald Trump's surprise victory poses the question: How did we get this thing *this* wrong? From the myriad polls and poll aggregators, to the vaunted oracles at Nate Silver's FiveThirtyEight and the *New York Times*'s shiny forecasting interface, most serious predictors completely misjudged Trump's chances of victory.

MORE STORIES

Democrats Should Be Worried About the Latino Vote

CHRISTIAN PAZ

The Iran Plane Crash Is the Big Story

URI FRIEDMAN

The 2020 U.S. Presidential Race: A Cheat Sheet

DAVID A. GRAHAM

CURRENT AFFAIRS
A Magazine of Politics & Culture

DECEMBER 29, 2016

WHY YOU SHOULD NEVER, EVER LISTEN TO NATE SILVER

Part I of our "How The Press Failed You" Series...

by NATHAN J. ROBINSON

MORE CURRENT AFFAIRS

THE CREDIBILITY GAP

WE MUST SAVE AND STRENGTHEN OUR PRECIOUS PUBLIC ASSETS

WE NEED MORE PUBLIC MONUMENTS TO ANIMALS

HOW NOT TO CRITICIZE ELIZABETH WARREN

SOLIDARITY IS THE BEST MEDICINE

Current Affairs

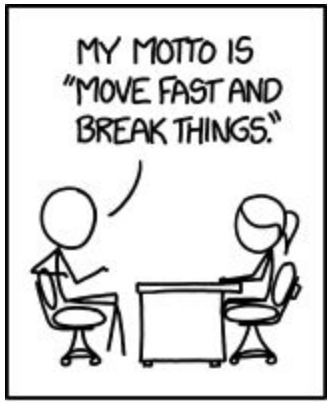
Tweets by @curaffairs

Facebook admits 'malicious actors' spread misinformation during the 2016 U.S. election

It also cites a government report that found Russia played a major role in the presidential race.

By Tony Romm and Kurt Wagner | Apr 28, 2017, 2:50pm EDT

https://www.vox.com/2017/4/28/15476142/facebook-report-trump-clinton-russia-us-presidential-election



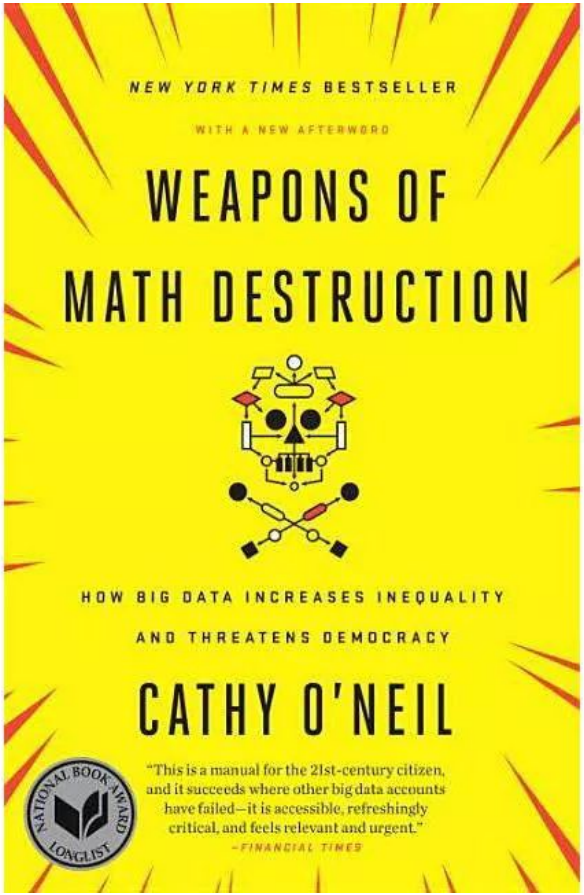
JOBS I'VE BEEN FIRED FROM

FEDEX DRIVER
CRANE OPERATOR
SURGEON
AIR TRAFFIC CONTROLLER
PHARMACIST
MUSEUM CURATOR
WAITER
DOG WALKER
OIL TANKER CAPTAIN
VIOLINIST
MARS ROVER DRIVER
MASSAGE THERAPIST

Used to be cool.

Now it's a punchline.

https://xkcd.com/1428/



https://weaponsofmathdestructionbook.com

"We live in the age of the algorithm. Increasingly, the decisions that affect our lives—where we go to school, whether we get a car loan, how much we pay for health insurance—are being made not by humans, but by mathematical models. In theory, this should lead to greater fairness: Everyone is judged according to the same rules, and bias is eliminated."

"But ... the opposite is true. The models being used today are opaque, unregulated, and uncontestable, even when they're wrong. Most troubling, they reinforce discrimination…

Models are propping up the lucky and punishing the downtrodden, creating a 'toxic cocktail for democracy'. Welcome to the dark side of Big Data"

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

This is a difficult topic to take on, but it gives us a chance to

1. Learn some important technical material in context, and
2. Take on a fundamental challenge to the future of data science.

The ProPublica article by Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner is an example of the power of data journalism.

But it also illustrates some of the hazards.

"We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the same benchmark used by the creators of the algorithm."

"...the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years."

"In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants."

Prediction Fails Differently for Black Defendants

	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.)

Where did these numbers come from, and what do they mean?

The confusion matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Summary metrics

Accuracy

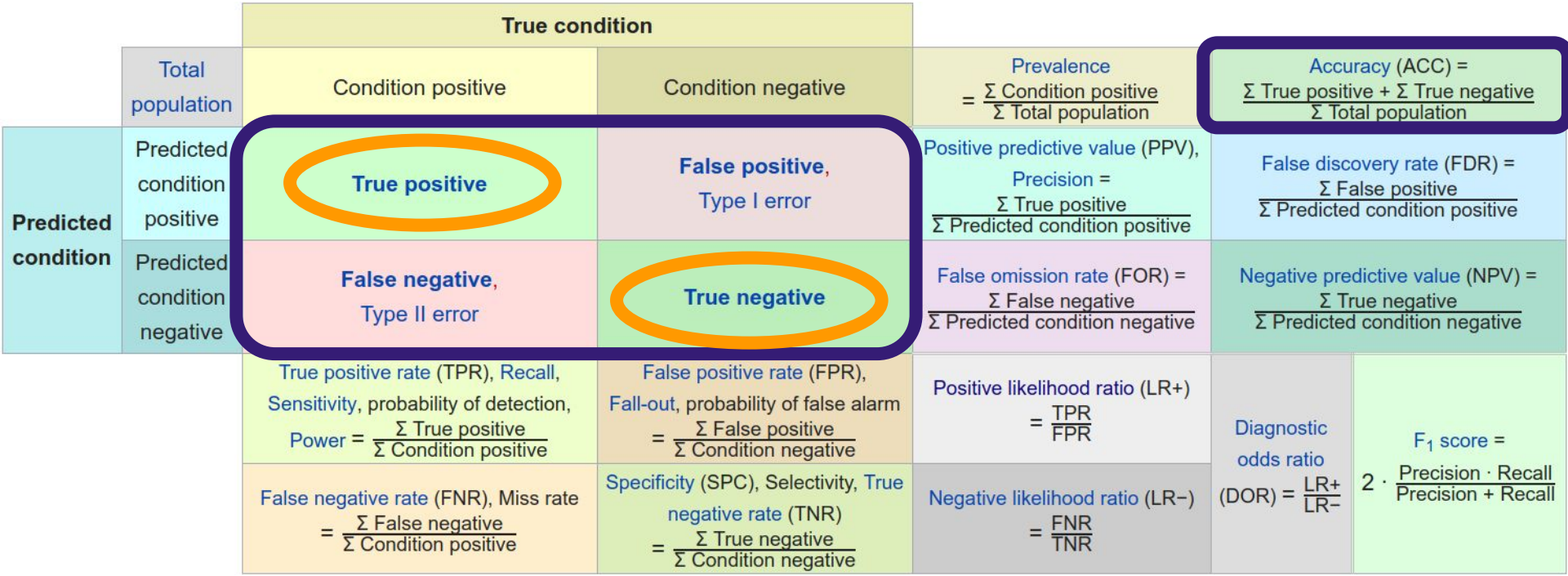
Positive predictive value (PPV)

Negative predictive value (NPV)

Sensitivity (complement of FNR)

Specificity (complement of FPR)

Accuracy



How many cases are classified correctly?

PPV

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Of all positive tests, how many are correct?

NPV

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Of all negative tests, how many are correct?

Sensitivity

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

When the condition is present, how often is it detected?

FNR

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

False negative rate is the complement of specificity.

Specificity

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

When the condition is absent, how often do we get it right?

FPR

		True condition			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$	F_1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

False positive rate is the complement of specificity.

FPR

At least, that's the convention.

But when people hear "false positive rate", they might think:

1. Of all cases, how many are false positives?
2. Of all positive tests, how many are false?
3. Of all cases where the condition is absent,
   how many get a positive test?

I think the authors of the ProPublica article do a good job explaining this to a lay audience.

Even so, there is a danger of misunderstanding.

As a consumer of these metrics:

1. Remember that different metrics answer different questions.
2. Look closely at what is computed and what is reported.

(Also notice that they go by different names in different domains.)

As a producer of these metrics:

1. Choose the metrics that quantify what you are interested in.
2. Present them carefully, in ways appropriate to the audience.
3. Remember that no one metric tells the whole story.

In the notebook I suggest you avoid "false positive rate" because I think it's prone to misinterpretation.

But that's a suggestion, not a rule.

It depends on your audience and the goal.

The authors of the ProPublica article broke my rule.

What do we think of that choice?

Other thoughts and questions?

Coming up...

The Washington Post
Democracy Dies in Darkness

Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

(Rich Pedroncelli/Associated Press)

By **Sam Corbett-Davies**, **Emma Pierson**, **Avi Feller** and **Sharad Goel**

Oct. 17, 2016 at 5:00 a.m. EDT

This past summer, a heated debate broke out about a tool used in courts across the country to help make bail and sentencing decisions. It's a controversy that touches on some of the big criminal justice questions facing our society. And it all turns on an algorithm.

Part 2: "It's actually not that clear."

The Washington Post
Democracy Dies in Darkness

Monkey Cage

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

(Rich Pedroncelli/Associated Press)

By **Sam Corbett-Davies**, **Emma Pierson**, **Avi Feller** and **Sharad Goel**

Oct. 17, 2016 at 5:00 a.m. EDT

This past summer, a heated debate broke out about a tool used in courts across the country to help make bail and sentencing decisions. It's a controversy that touches on some of the big criminal justice questions facing our society. And it all turns on an algorithm.

As we did with the ProPublica article, let's:

1. Understand the arguments in the WaPo article,
2. Replicate their analysis, and
3. Get deeper into the technical details.

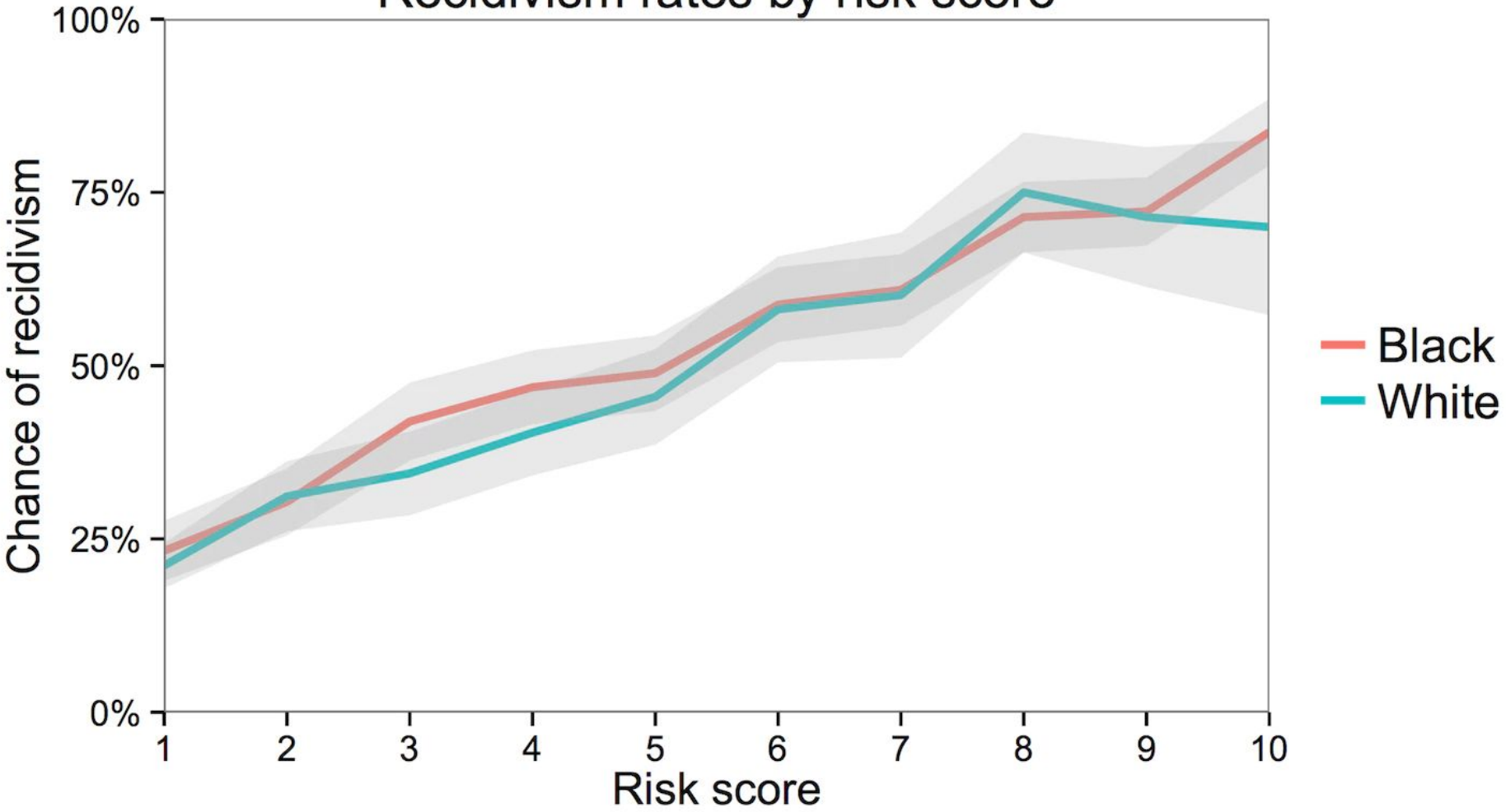
"What does it mean for an algorithm to be fair?

Surprisingly, there is a mathematical limit to how fair any algorithm — or human decision-maker — can ever be."

"Northpointe contends they are indeed fair because scores mean essentially the same thing regardless of the defendant’s race.

...

The plot below shows this approximate equality between white and black defendants holds for every one of Northpointe’s 10 risk levels."


Recidivism rates by risk score
100%
75%
50%
25%
0%
Chance of recidivism
1
2
3
4
5
6
7
8
9
10
Risk score
Black
White

The output of COMPAS is a 10-point risk score,
not a binary classification.

The graph shows that higher score implies higher risk.

And the score is equally "calibrated".

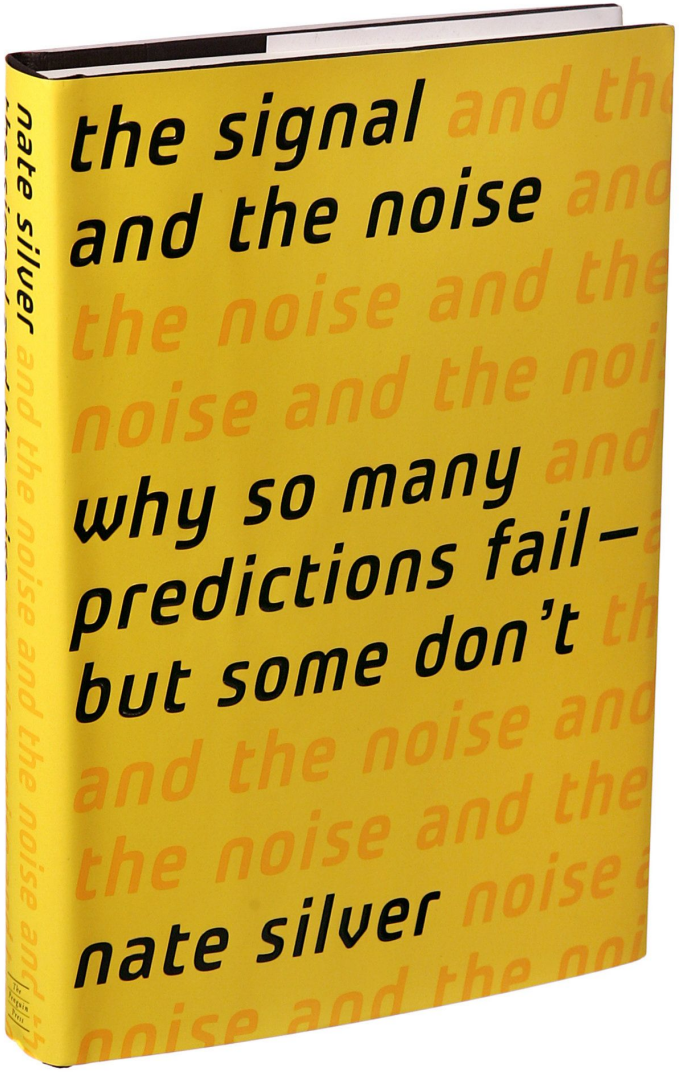

the signal
and the noise
why so many
predictions fail—
but some don't
nate silver
nate silver

FIGURE 4-7: NATIONAL WEATHER SERVICE CALIBRATION

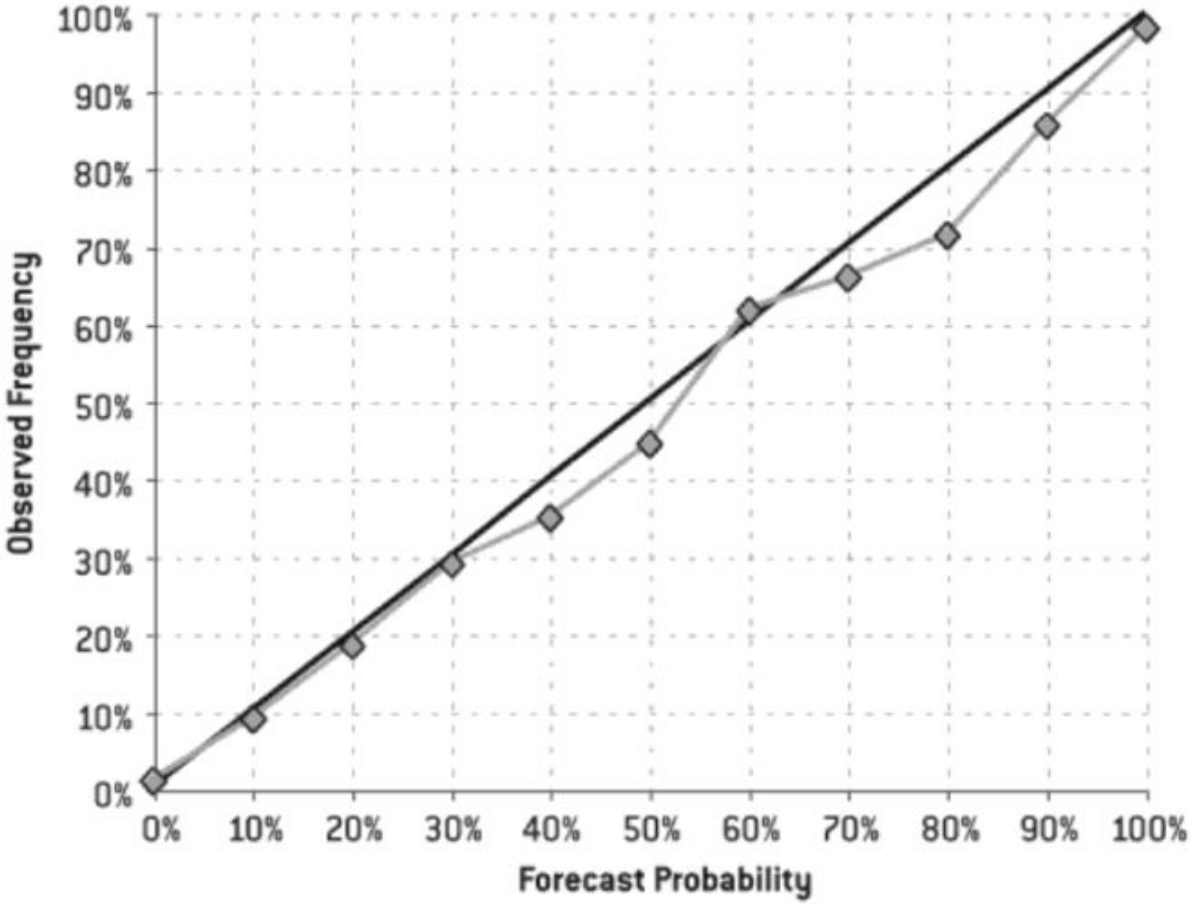


FIGURE 4-8: THE WEATHER CHANNEL CALIBRATION

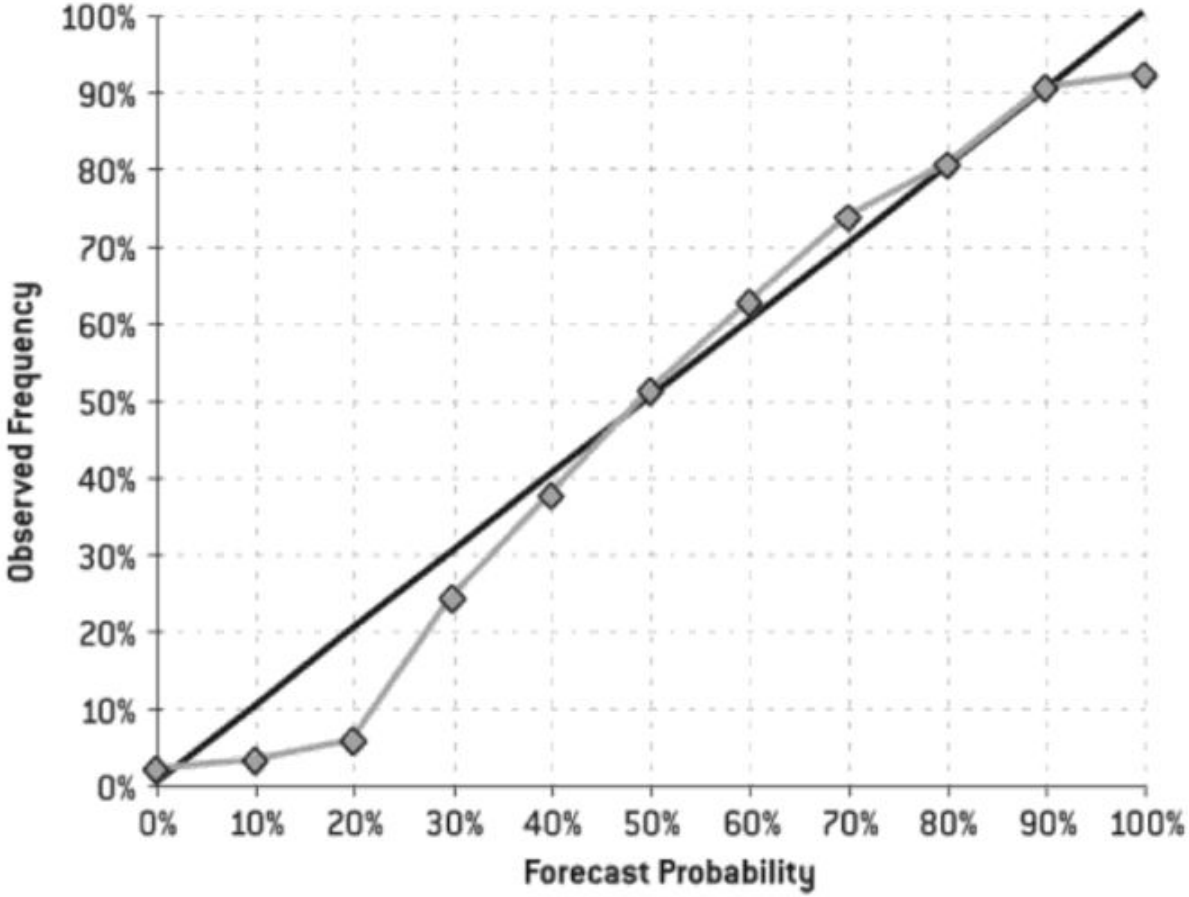
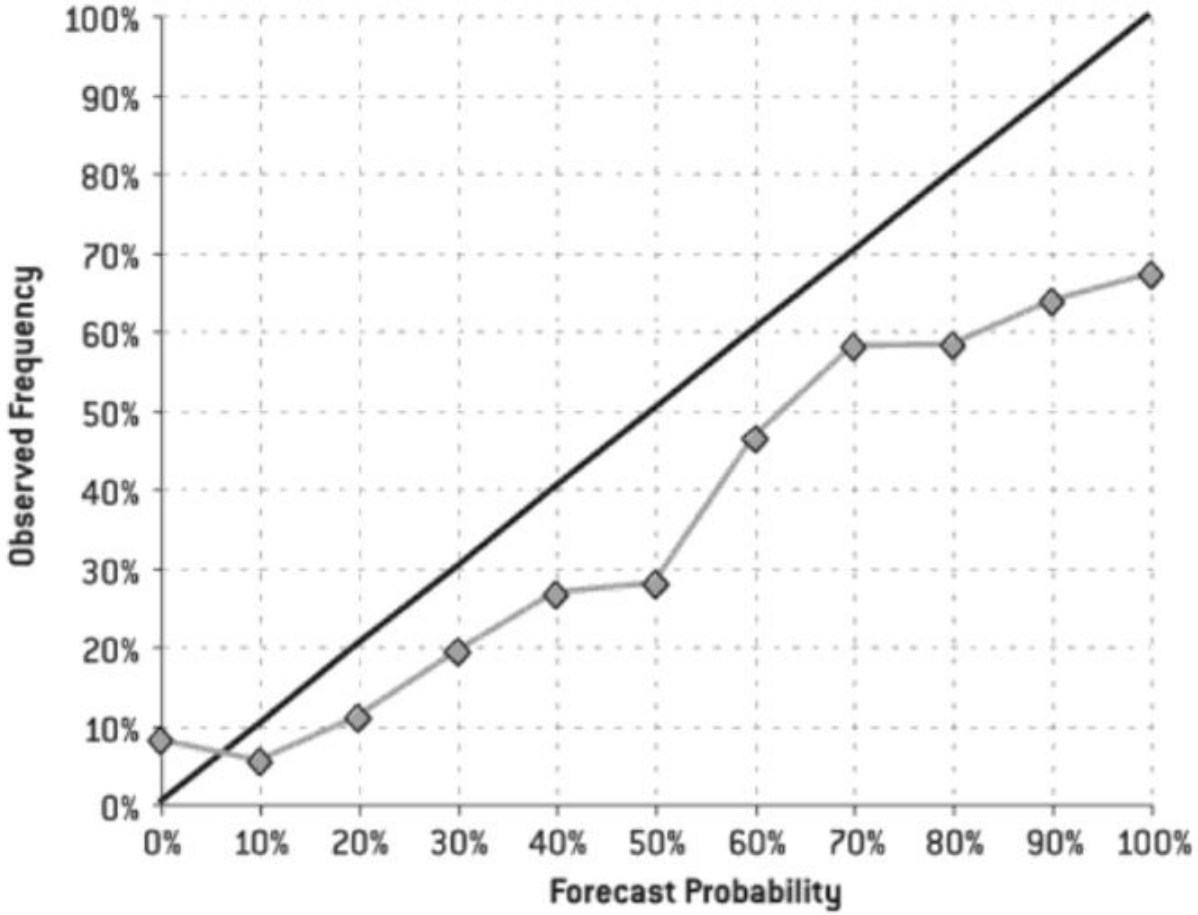
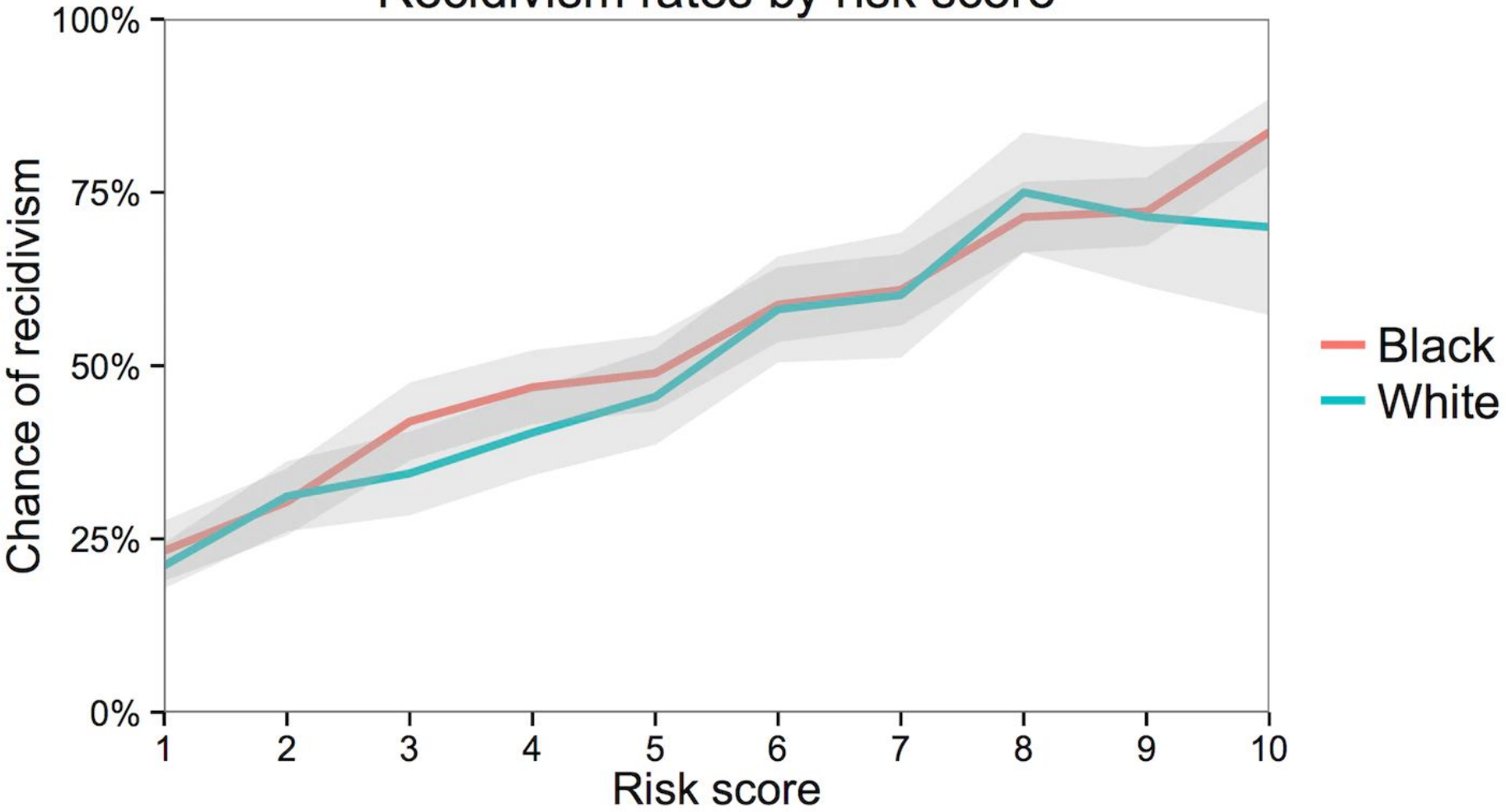


FIGURE 4-9: LOCAL TV METEOROLOGIST CALIBRATION



When I called the curve in the WaPo article a calibration curve, I was speaking loosely.


Recidivism rates by risk score
100%
75%
50%
25%
0%
Chance of recidivism
1
2
3
4
5
6
7
8
9
10
Risk score
Black
White

Nevertheless it shows that risk scores implies
the same probability of recidivism
for black and white defendants.

Which implies that PPV and NPV are the same.

Equal calibration implies equal predictive values.

But not necessarily the other way around.

"But ProPublica points out that among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be classified as medium or high risk …

ProPublica argues that a fair algorithm cannot make these serious errors more frequently for one race group than for another."

WaPo: "Here’s the problem: it’s actually impossible
for a risk score to satisfy both fairness criteria at the same time."

You can't have

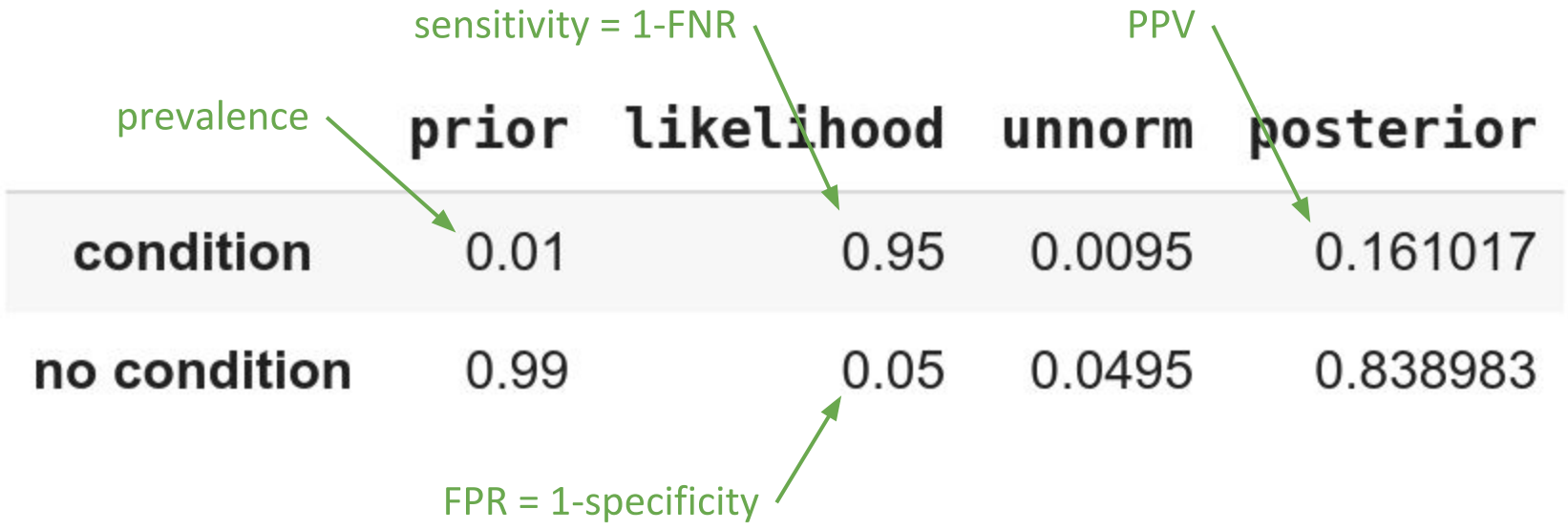
1. Equal predictive values (PPV and NPV) and
2. Equal error rates (FPR and FNR, specificity and sensitivity)

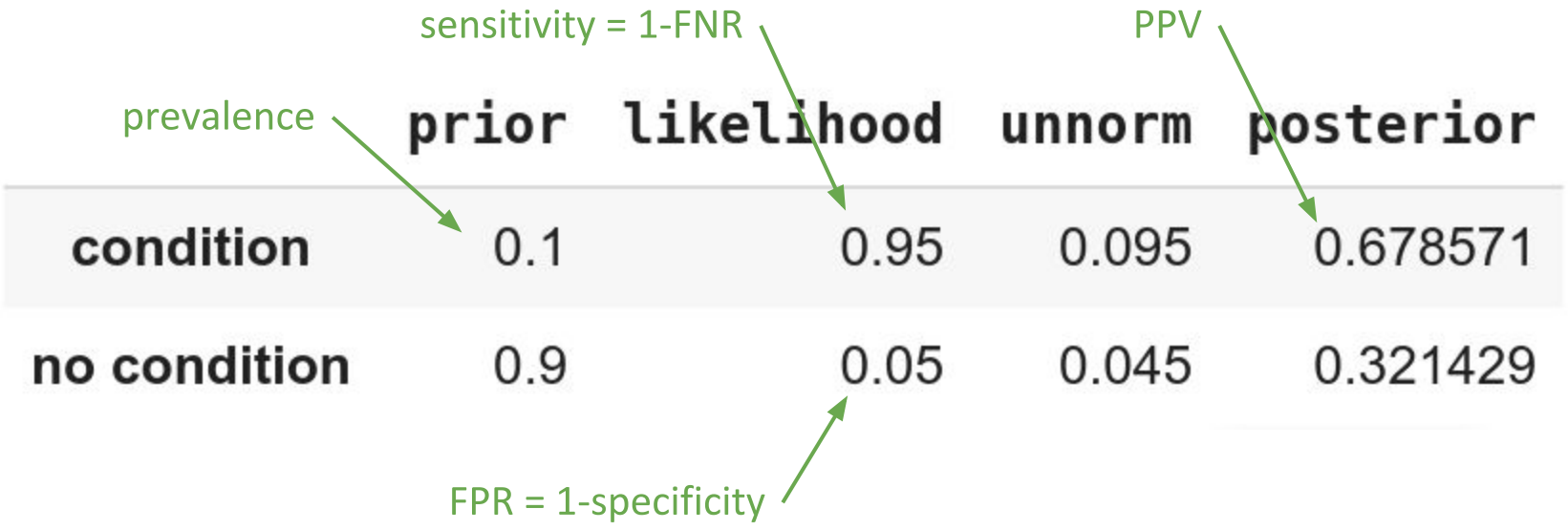
[If prevalence is not equal.]

The WaPo article tries to explain without getting too technical.

We can do better.

Remember Bayesville?


sensitivity = 1-FNR
PPV
prevalence
prior
likelihood
unnorm
posterior
condition
0.01
0.95
0.0095
0.161017
no condition
0.99
0.05
0.0495
0.838983
FPR = 1-specificity


sensitivity = 1-FNR
PPV
prevalence
prior
likelihood
unnorm
posterior
condition
0.1
0.95
0.095
0.678571
no condition
0.9
0.05
0.045
0.321429
FPR = 1-specificity

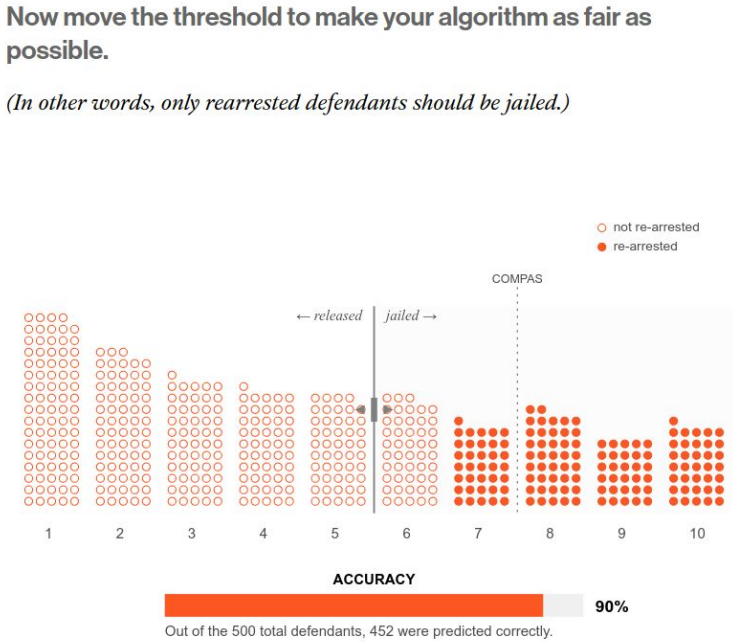
Given prevalence, FPR, and FNR, we can compute PPV.

We can also use Bayes to compute NPV.

And with a little algebra, we can go the other way.

You can see the details in the fourth notebook.

Another way to understand this "uncertainty principle of fairness" is to interact with it.



https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/

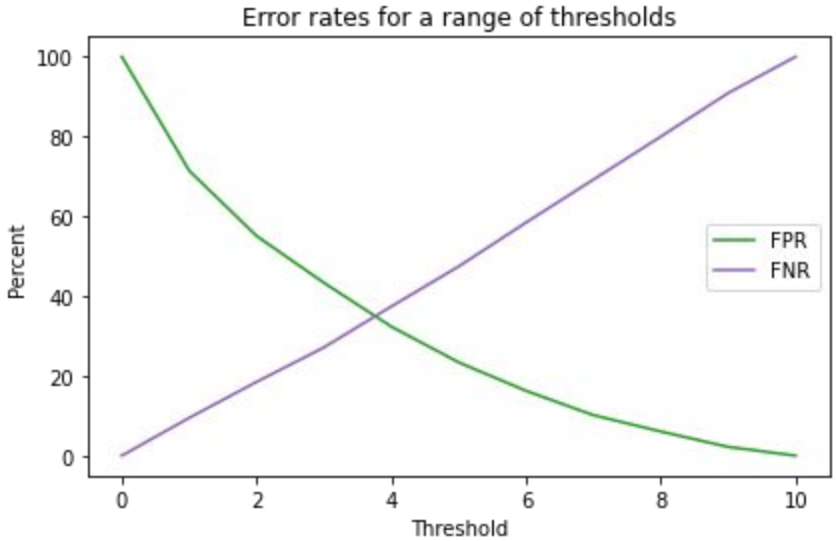
In the notebook I take a different approach.

ProPublica chose an arbitrary threshold for "low" and "high" risk.

First rule of ModSim:
see a free parameter, sweep it!

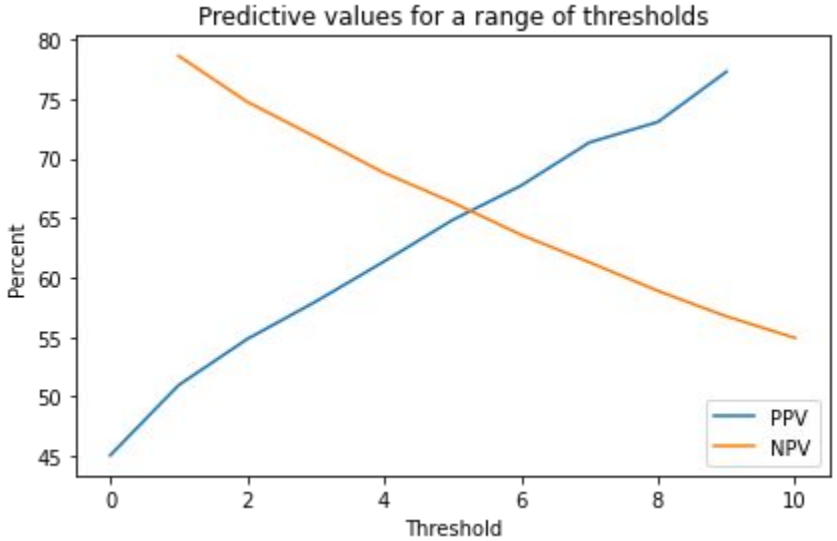
Low threshold,
everyone in jail, high FPR.

High threshold,
everyone goes free, high FNR.



Low threshold, "low" risk means very low, so high NPV.

High threshold, "high" risk means very high, so high PPV.



If risk scores are calibrated and
we use the same threshold for everyone,
we get:

- Equal predictive values (PPV and NPV).
- Unequal error rates (FPR and FNR).

If we use different thresholds, we can get:

- Equal error rates,
- Unequal predictive values.

With respect to black and white defendants, COMPAS has

- Equal predictive values (PPV and NPV),
- Unequal error rates (FPR and FNR).

And as we'll see in the next notebook,
with respect to male and female defendants,
COMPAS has

- Equal error rates,
- Unequal predictive values.

Can't tell if that's by design (because it's opaque).

If so, not clear why.

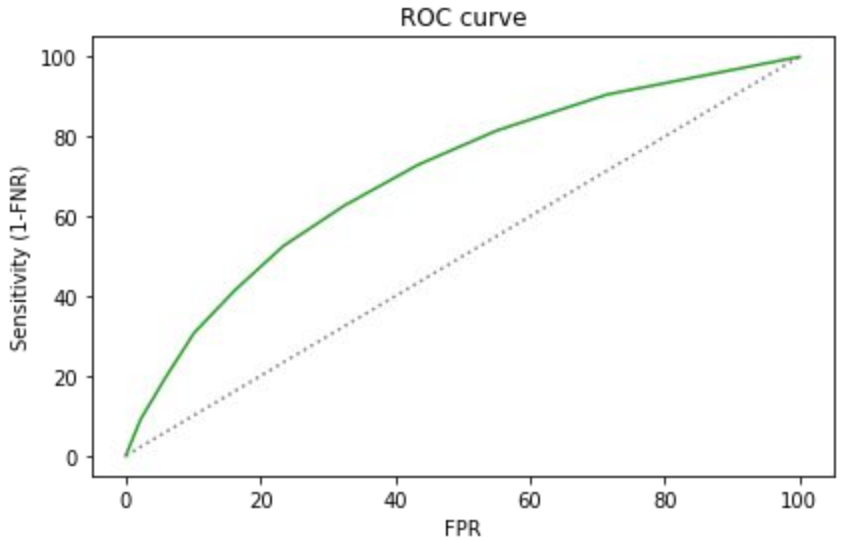
Which is worse?

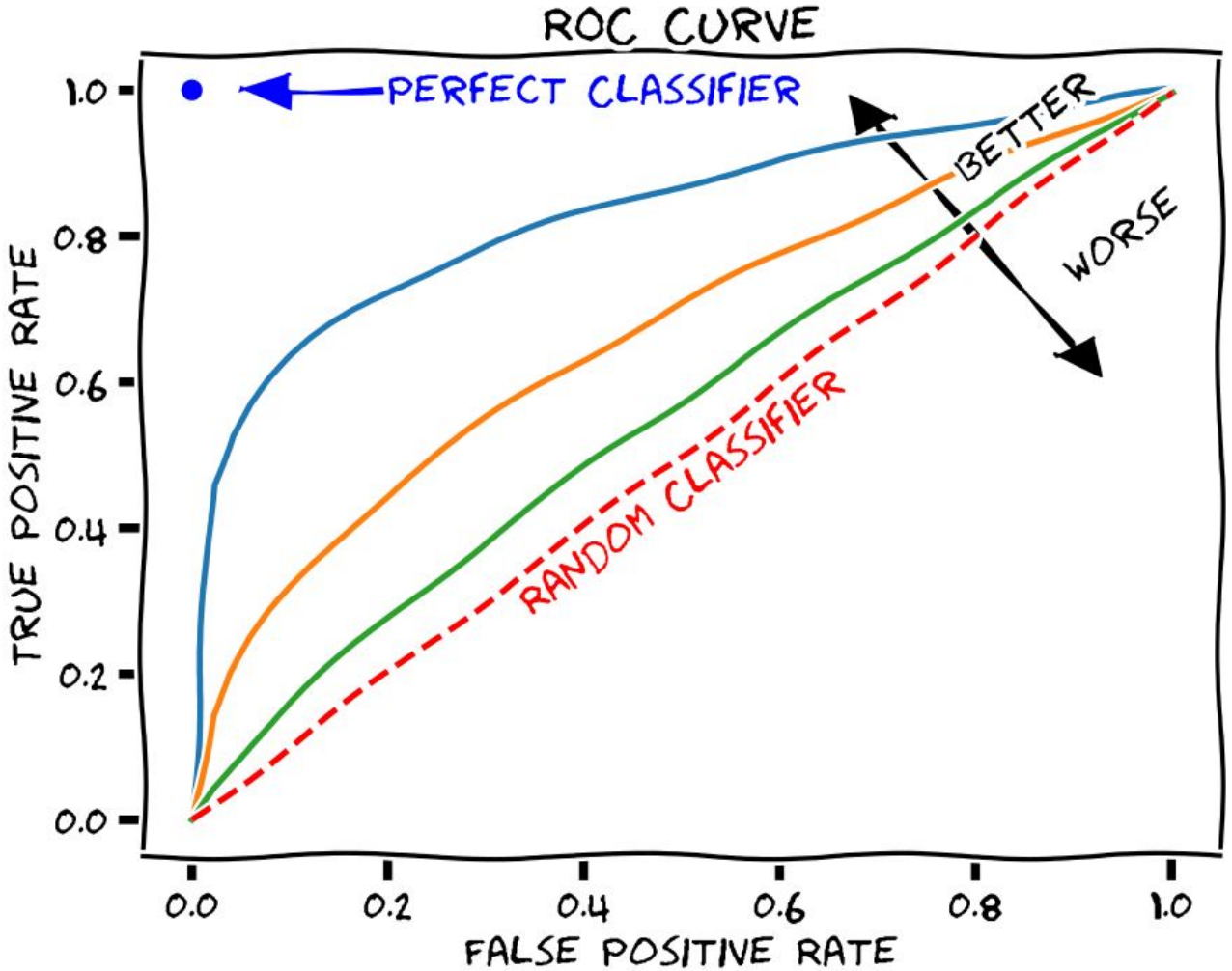
While we're sweeping the threshold...

Receiver operating characteristic (ROC) curve.

Sweep threshold and plot sensitivity vs FPR.

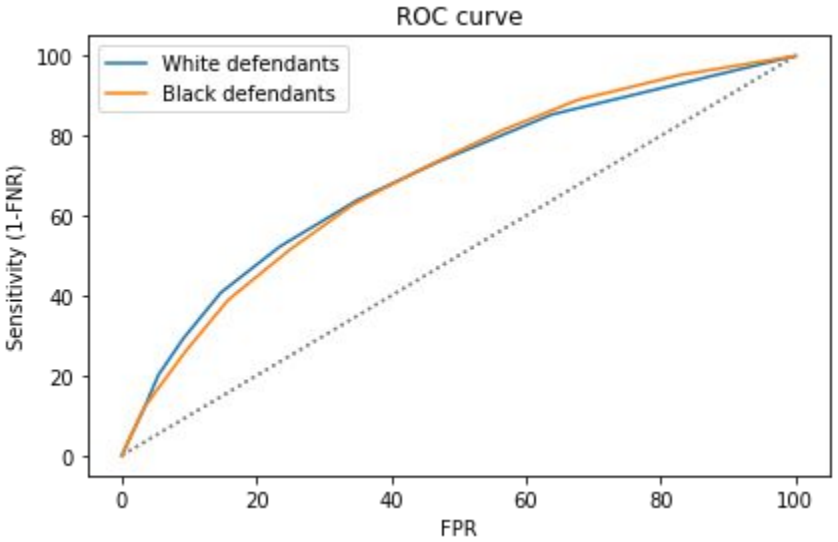
(So it plots from right to left.)



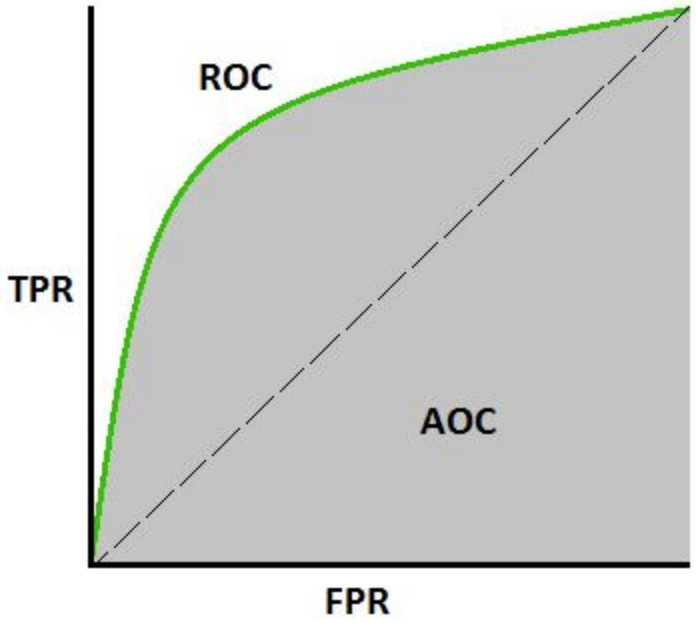


https://commons.wikimedia.org/wiki/File:Roc-draft-xkcd-style.svg

Equal ROC curve implies that risk scores have the same predictive value in each group.



The area under the ROC curve is called the "area under the curve".



https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

AUC is "concordance", which is the
probability you rank two people correctly.

More generally, it gives you a sense of
how well you can classify,
independent of threshold.

(The other metrics all depend on the threshold.)

Next time: male and female defendants.

Part 3: Which kind of fair?

In the Broward County data, about 81% of defendants are male.

Male recidivism* rate is substantially higher (47% vs 36%).

Based on breakdown by race,
we expect higher FPR, lower FNR,
for male defendants.

Nope.

Male defendants	Percent
FPR	32.420091
FNR	37.086814
PPV	63.536317
NPV	66.989977
Prevalence	47.310534

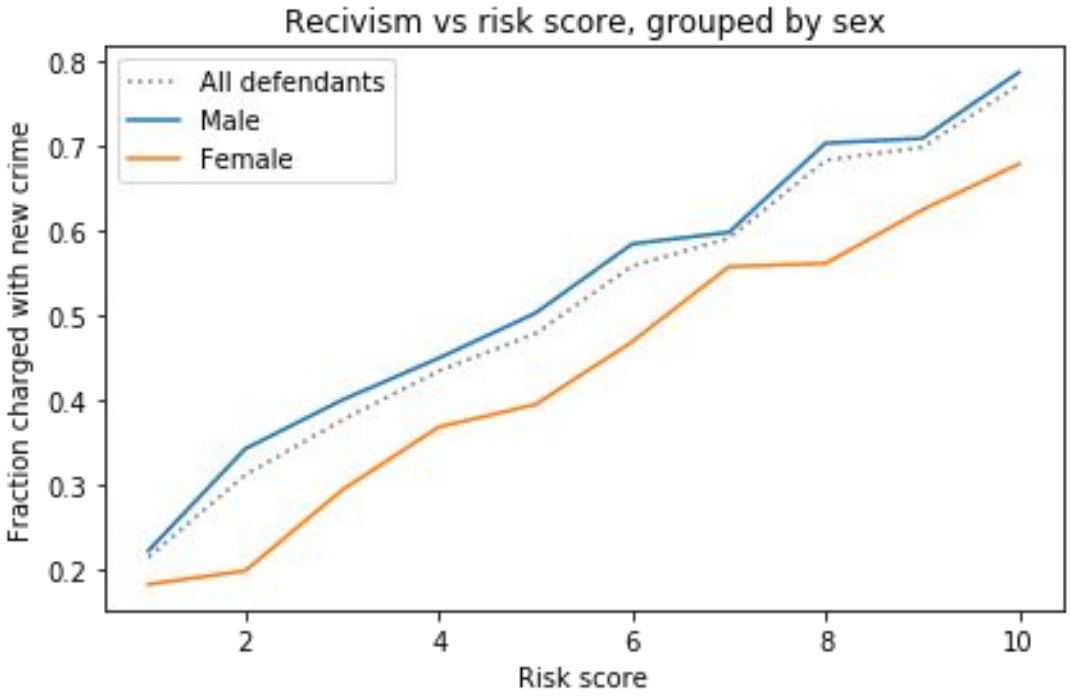
=

≠

Female defendants	Percent
FPR	32.107023
FNR	39.156627
PPV	51.269036
NPV	75.746269
Prevalence	35.698925

That's consistent with the calibration curve.

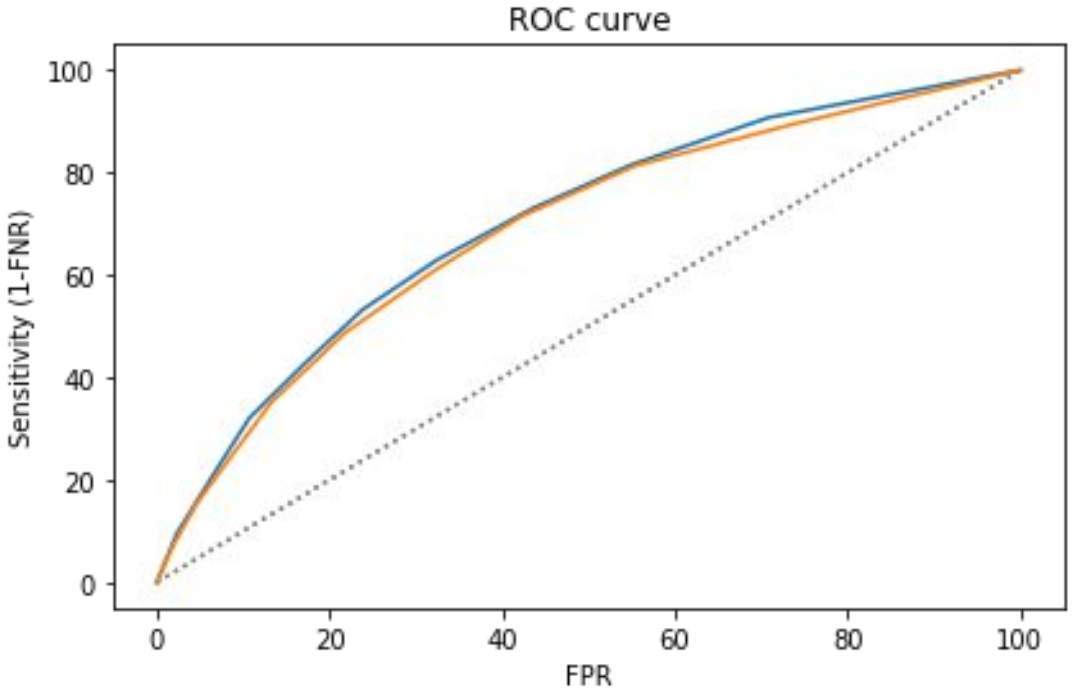
At every risk level, women are substantially less likely to be charged with another crime.



The ROC curves are the same, which means the tradeoff is the same.

That is, COMPAS *could* be calibrated.

But it's not.



For black and white defendants:

- Same predictive values.
- Different error rates.

For male and female defendants:

- Same error rates.
- Different predictive values.

Is that deliberate? If so, why?

We have no way to know.

Which is, as others have observed, a problem in itself.

Stakeholders

Northpointe makes and sells COMPAS.

Government agencies buy COMPAS.

Judges and parole officers use ratings to guide decisions.

Defendants are affected by those decisions, individually.

The general public is affected, in aggregate.

For each stakeholder, there are overt and ulterior motives.

Northpointe

Maximize accuracy and fairness.

Maximize the criteria customers care about.

Maintain competitive advantage.

Avoid blame and negative publicity.

Make money.

The County

Reduce human biases.

Reduce crime.

Minimize time and cost.

Maintain perception of fairness.

Keep prisons full?

The Judge

Aligned with the county?

The defendant

The general public

Costs

For each outcome, there are costs and benefits.

- True positive: dangerous person off the streets.
- True negative: low risk person spared jail.
- False negative: freed defendant commits crime.
- False positive: low risk person in jail.

Who gets the benefits?

Who pays the costs?

Which ones are quantifiable?

Which ones make the news?

Not a headline

Florida man classified low risk, goes home to family,
keeps job, contributes to community,
saves tax-payers tens of thousands of dollars.

Florida Inmate Released Amid Pandemic Killed Someone the Next Day, Officials Say

Joseph Edward Williams shot and killed a man in Tampa, Fla., on March 20, the authorities said, one day after he was among more than 160 inmates released from Hillsborough County jails.

By Michael Levenson and Alan Yuhas

April 15, 2020

It was an effort, like many across the country, to try to slow the spread of the coronavirus in jails. But one day after officials in Hillsborough County, Fla., released more than 160 inmates, one of them shot and killed a man in Tampa, the authorities said.

History repeats?

The New York Times

Prison Furloughs in Massachusetts Threaten Dukakis Record on Crime

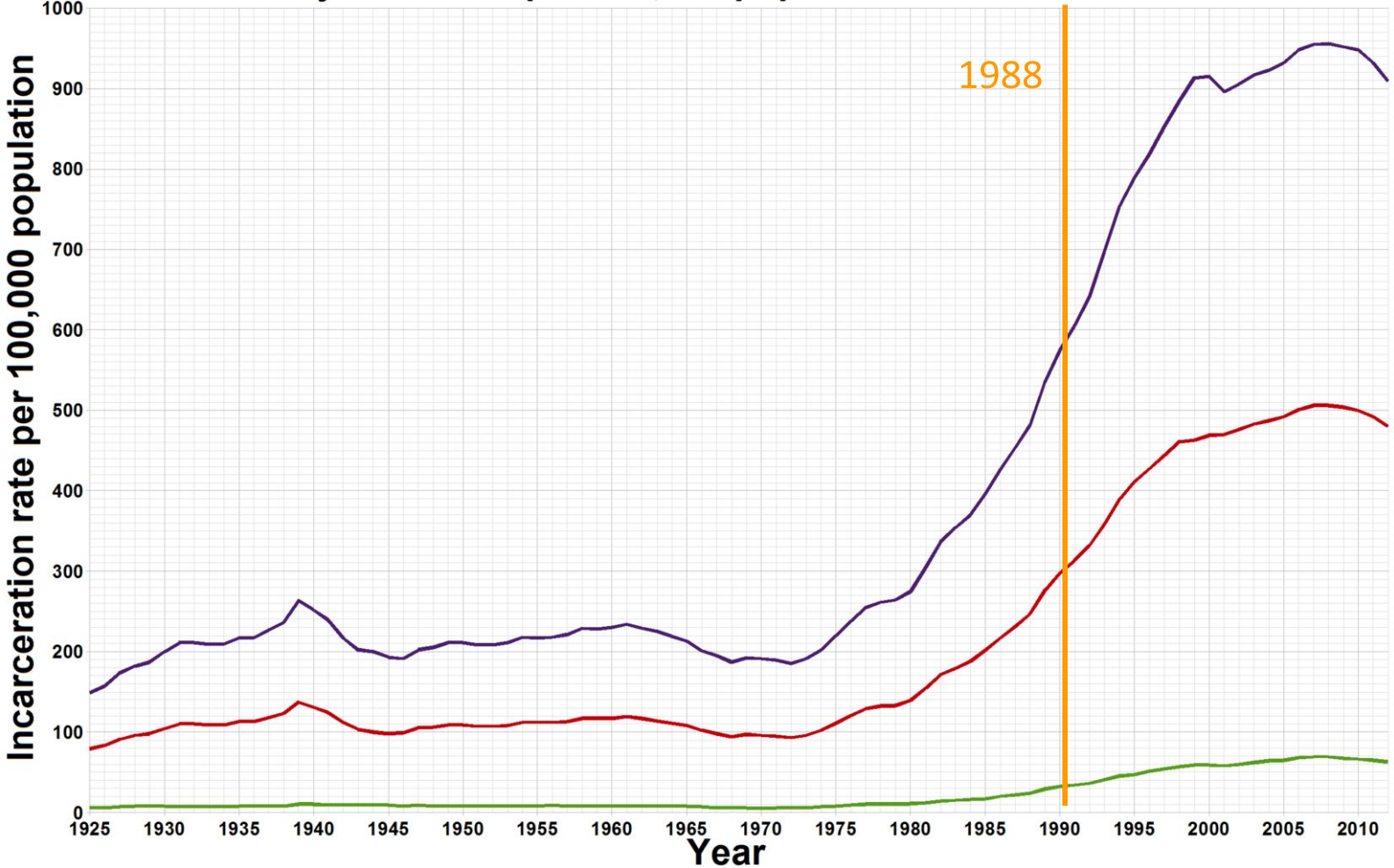
By Robin Toner

July 5, 1988

Angela and Clifford Barnes, a young couple who live in a picture-book home in a suburb of Washington, say they had no desire to become an issue in the 1988 Presidential campaign.

But what happened to the Barneses on the night of April 3, 1987, is now at the center of the effort by Vice President Bush's campaign to portray Gov. Michael S. Dukakis as soft on crime. On that night William R. Horton, a convicted murderer who had been furloughed from the Massachusetts prison system for a weekend and did not go back, broke into the couple's home, bound and stabbed Mr. Barnes and raped his wife.

Incarceration rate of inmates incarcerated under state and federal jurisdiction per 100,000 population 1925-2014



https://en.wikipedia.org/wiki/United_States_incarceration_rate#/media/File:U.S._in

What should we do?

Should we use COMPAS or not?

Either way, should we try to:

1. Calibrate it to achieve equal predictive power.
2. Achieve equal error rates?
3. Temper it to balance these tradeoffs?

COMPAS or not

You are a judge with all the data.

You can feed it into COMPAS and get a risk score.

Or make a decision based on the data.

Let's assume we can't make radical reforms today.

What should you do?

What if you're the defendant:

- Would you want the judge to use COMPAS
  or just squint at the data?
- If you were black, would you want equal error rates?
- If you were female, would you want equal predictive values?

Notebook 4 shows:

- Given prevalence and error rates,
  we can compute predictive values.
- Given prevalence and predictive values,
  we can compute error rates.

It uses SymPy to do the algebra.

Notebook 5 shows that the challenge of defining fairness between groups gets harder as we consider more groups.

Notebook 6 explores "the other calibration curve",
the probability of being classified high risk as a function of
the probability of recidivism.

This is a version of the reference class problem.

What is your probability of false positive if you are:

- White
- White male
- Old white male
- Old white male, married
- Old white male, married with children...

Caucasian	**Female**	**1 Younger than 25**	70.000000	3.703704	38.235294	94.736842	31.034483	87
		2 Between 25 - 45	29.213483	41.221374	59.689922	70.000000	42.394822	309
		3 Older than 45	13.076923	75.609756	37.037037	78.472222	23.976608	171
	Male	**1 Younger than 25**	39.568345	30.487805	67.455621	62.686567	54.125413	303
		2 Between 25 - 45	26.056338	48.735632	60.107817	66.455696	43.369890	1003
		3 Older than 45	8.474576	67.261905	61.111111	76.985743	28.915663	581
Hispanic	**Female**	**1 Younger than 25**	NaN	NaN	NaN	NaN	NaN	17
		2 Between 25 - 45	7.142857	80.952381	57.142857	69.642857	33.333333	63
		3 Older than 45	NaN	NaN	NaN	NaN	NaN	23
	Male	**1 Younger than 25**	52.941176	37.288136	57.812500	52.173913	53.636364	110
		2 Between 25 - 45	21.052632	55.263158	56.043956	70.422535	37.500000	304
		3 Older than 45	13.829787	76.923077	31.578947	80.198020	21.666667	120

FPR

```
('Caucasian', 'Female', '1 Younger than 25') 70.0
('Other', '3 Older than 45') 3.2
```

If you are a young white female who will not recidivate, there's a 70% chance you will be classified high risk.

FNR

```
('Other', '3 Older than 45') 86.4
('Caucasian', 'Female', '1 Younger than 25') 3.7
```

If you are old, mixed race, and will recidivate,
there's an 86% chance you will be classified low risk.

PPV

```
('African-American', 'Male', '1 Younger than 25') 70.3
('Hispanic', '3 Older than 45') 28.6
```

If you are a young black male and classified high risk, there's a 70% chance you will recidivate*.

NPV

```
('Caucasian', 'Female', '1 Younger than 25') 94.7
('Other', 'Male', '1 Younger than 25') 48.4
```

If you are a young, white female and are classified low risk, there is a 95% you will not be charged with another crime.

Even if we decide which metric should be kept constant, we probably can't keep it constant for all subgroups.

But just as a reminder,
human decision makers have the same problem.

I've heard that some of them can be biased.

And they are just as opaque.