Classification Accuracy

Vishal Patel

Spring 2022

1. ~~Introduction~~
2. ~~The Data Science Process~~
3. **Supervised Learning: Classification Accuracy Measures**
4. Unsupervised Learning
5. The Grunt Work
6. Wrap Up

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1j} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2j} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3j} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{nj} \end{pmatrix} \quad y = \begin{pmatrix} 1 \\ 1 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix} \quad \hat{p} = \begin{pmatrix} 0.29 \\ 0.90 \\ 0.31 \\ . \\ . \\ . \\ 0.47 \end{pmatrix} \quad \hat{y} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}$$

`predict_proba()` `predict()`

True Class (y)	Predicted Probability ($\hat{p}$)
+	0.29
+	0.90
−	0.31
−	0.89
+	0.95
+	0.71
−	0.24
+	0.12
−	0.59
−	0.47

True Class (y)	Predicted Probability ($\hat{p}$)
+	0.95
+	0.90
−	0.89
+	0.71
−	0.59
−	0.47
−	0.31
+	0.29
−	0.24
+	0.12

True Class (y)	Predicted Probability ($\hat{p}$) ↓	Predicted Class	True / False?	True Positive	False Positive	True Negative	False Negative
+	0.95	+	True Positive	1	0	0	0
+	0.90	+	True Positive	2	0	0	0
−	0.89	+	False Positive	2	1	0	0
+	0.71	+	True Positive	3	1	0	0
−	0.59	+	False Positive	3	2	0	0
−	0.47	−	True Negative	3	2	1	0
−	0.31	−	True Negative	3	2	2	0
+	0.29	−	False Negative	3	2	2	1
−	0.24	−	True Negative	3	2	3	0
+	0.12	−	False Negative	**3**	**2**	**3**	**2**

True Class (y)	Predicted Probability ($\hat{p}$)	Predicted Class	True / False?
+	0.95	+	True Positive
+	0.90	+	True Positive
−	0.89	+	False Positive
+	0.71	+	True Positive
−	0.59	+	False Positive
−	0.47	−	True Negative
−	0.31	−	True Negative
+	0.29	−	False Negative
−	0.24	−	True Negative
+	0.12	−	False Negative

True Positive	False Positive	True Negative	False Negative
3	2	3	2

		Predicted Class	
		+	−
True Class	+	True Positive 3	False Negative 2
	−	False Positive 2	True Negative 3

Confusion matrix

True Class (y)	Predicted Probability ($\hat{p}$) ↓
+	0.95
+	0.90
−	0.89
+	0.71
−	0.59
−	0.47
−	0.31
+	0.29
−	0.24
+	0.12

True Positive	False Positive	True Negative	False Negative
3	2	3	2

ACCURACY $= \dfrac{TP + TN}{\text{Total Population}} = \dfrac{3+3}{10} = 60.0\%$

SENSITIVITY / **RECALL** $= \dfrac{TP}{\text{Total Positives}} = \dfrac{3}{3+2} = 60.0\%$

True Positive Rate

SPECIFICITY $= \dfrac{TN}{\text{Total Negatives}} = \dfrac{3}{3+2} = 60.0\%$

True Negative Rate

True Class (y)	Predicted Probability ($\hat{p}$) ↓
+	0.95
+	0.90
−	0.89
+	0.71
−	0.59
−	0.47
−	0.31
+	0.29
−	0.24
+	0.12

True Positive	False Positive	True Negative	False Negative
3	2	3	2

$$\text{RECALL} = \boxed{\frac{\text{TP}}{\text{Total Positives}}} = 60.0\%$$

$$\text{PRECISION} = \boxed{\frac{\text{TP}}{\text{Predicted Positives}}} = \frac{3}{3+2} = 60.0\%$$

$$\text{F1 SCORE} = \boxed{\frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}} = \frac{1}{\frac{\frac{1}{0.6} + \frac{1}{0.6}}{2}} = 0.60$$

True Class (y)	Predicted Probability ($\hat{p}$) ↓	Predicted Class
+	0.95	+
+	0.90	+
−	0.89	+
+	0.71	+
−	0.59	+
−	0.47	−
−	0.31	−
+	0.29	−
−	0.24	−
+	0.12	−

THRESHOLD = **0.5**

True Positive	False Positive	True Negative	False Negative
3	2	3	2

Measure	Value
Accuracy	0.60
Sensitivity (Recall)	0.60
Specificity	0.60
Precision	0.60
F1 Score	0.60

True Class (y)	Predicted Probability ($\hat{p}$) ↓	Predicted Class
+	0.95	+
+	0.90	+
−	0.89	+
+	0.71	+
−	0.59	−
−	0.47	−
−	0.31	−
+	0.29	−
−	0.24	−
+	0.12	−

THRESHOLD = **0.6**

True Positive	False Positive	True Negative	False Negative
3	1	4	2

Measure	Value
Accuracy	0.70
Sensitivity (Recall)	0.60
Specificity	0.80
Precision	0.75
F1 Score	0.67

PROBABILITY THRESHOLD VALUES

Measure	0.1	0.2	0.3	0.4	**0.5**	0.6	0.7	0.8	0.9
Accuracy	0.50	0.40	0.40	0.50	0.60	**0.70**	**0.70**	0.60	**0.70**
Sensitivity (Recall)	**1.00**	0.80	0.60	0.60	0.60	0.60	0.60	0.40	0.40
Specificity	--	--	0.20	0.40	0.60	0.80	0.80	0.80	**1.00**
Precision	0.50	0.44	0.43	0.50	0.60	0.75	0.75	0.67	**1.00**
F1 Score	**0.67**	0.57	0.50	0.55	0.40	**0.67**	**0.67**	0.50	0.57

1. The values depend on the decision boundary.

Breathalyzers

ASSUMPTIONS:

1. **One in a thousand** drivers is driving drunk.
2. A breathalyzer **never** fails to detect a truly drunk person.
3. However, it displays false drunkenness in **5%** of the cases when the driver is actually sober.
4. A police officer stops a **random** driver, and asks her to take the breathalyzer test.
5. The test indicates that the driver **is** drunk.

What is the probability that the driver is DUI?

(a) 95% **(b)** 90% **(c)** 5% **(d)** 2%

1,000 Tests

One Actual Drunk Driver

The test never fails to detect a truly drunk person.
Hence, the test result for this person would display drunkenness (aka “positive”).

True Positive	False Positive
1	0

999 Responsible Drivers

The tests would show false drunkenness in 5% of these cases.
Hence, ~50 of these tests would display drunkenness (aka “positives”).

True Positive	False Positive
0	50

PRECISION = $\boxed{\dfrac{\textbf{TP}}{\textbf{Predicted Positives}}} = \dfrac{1+0}{1+0+50+0} \approx 2.0\%$

This is the probability that *one* of the drivers among all drivers who tested positive is actually drunk.

ALTERNATIVE SCENARIO:

1,000 Tests

500
Drunk Drivers

The test never fails to detect a truly drunk person.
Hence, the test result for these drivers would display drunkenness (aka “positives”).

True Positive	False Positive
500	0

500
Responsible Drivers

The tests would show false drunkenness in 5% of these cases.
Hence, 25 test results would display drunkenness (aka “positives”).

True Positive	False Positive
0	25

$$\textbf{PRECISION} = \frac{\text{TP}}{\text{Predicted Positives}} = \frac{500}{500 + 25} \approx 95.0\%$$

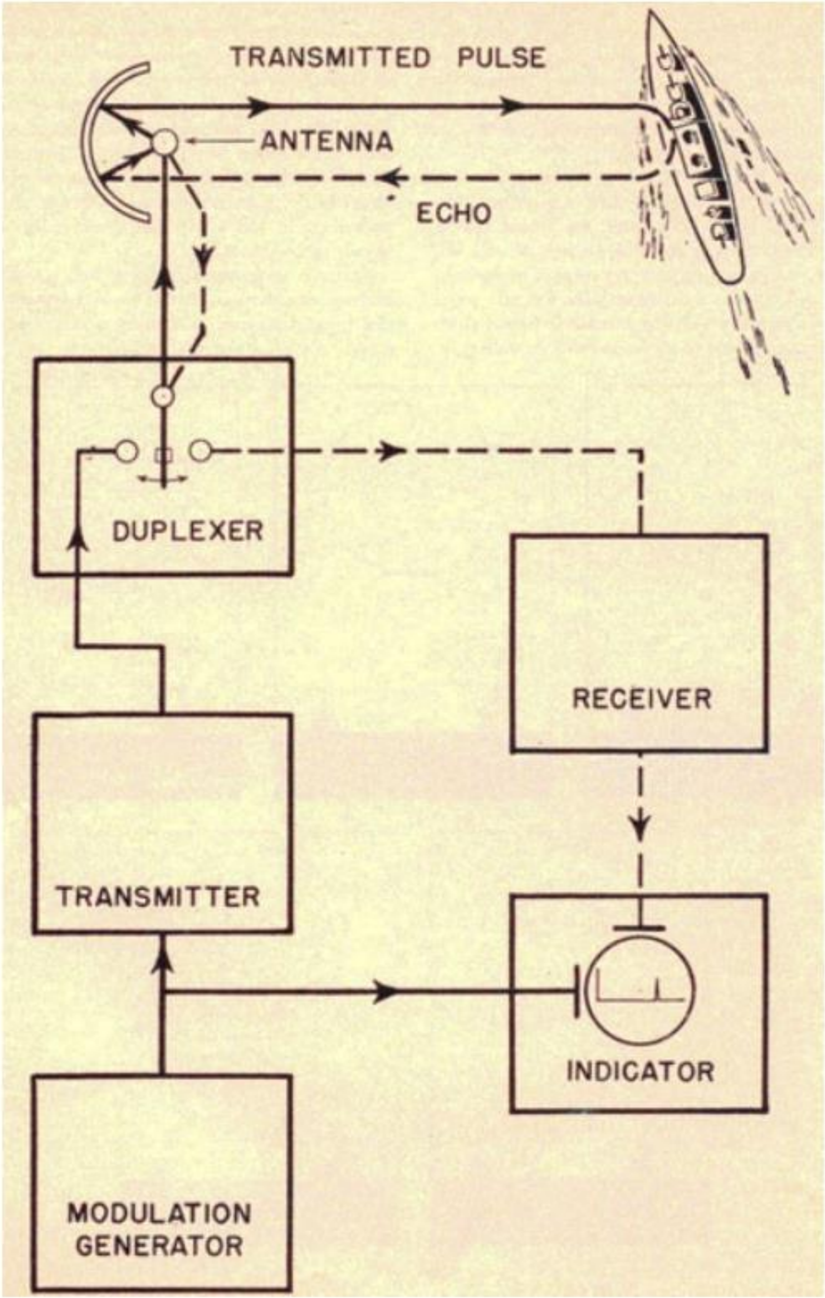
This is the probability that *one* of the drivers among all drivers who tested positive is actually drunk.

1. The values depend on the **decision boundary**.

2. The values depend on the **class probabilities**.

We need a measure that is independent of those two factors.

RADAR OPERATOR'S MANUAL



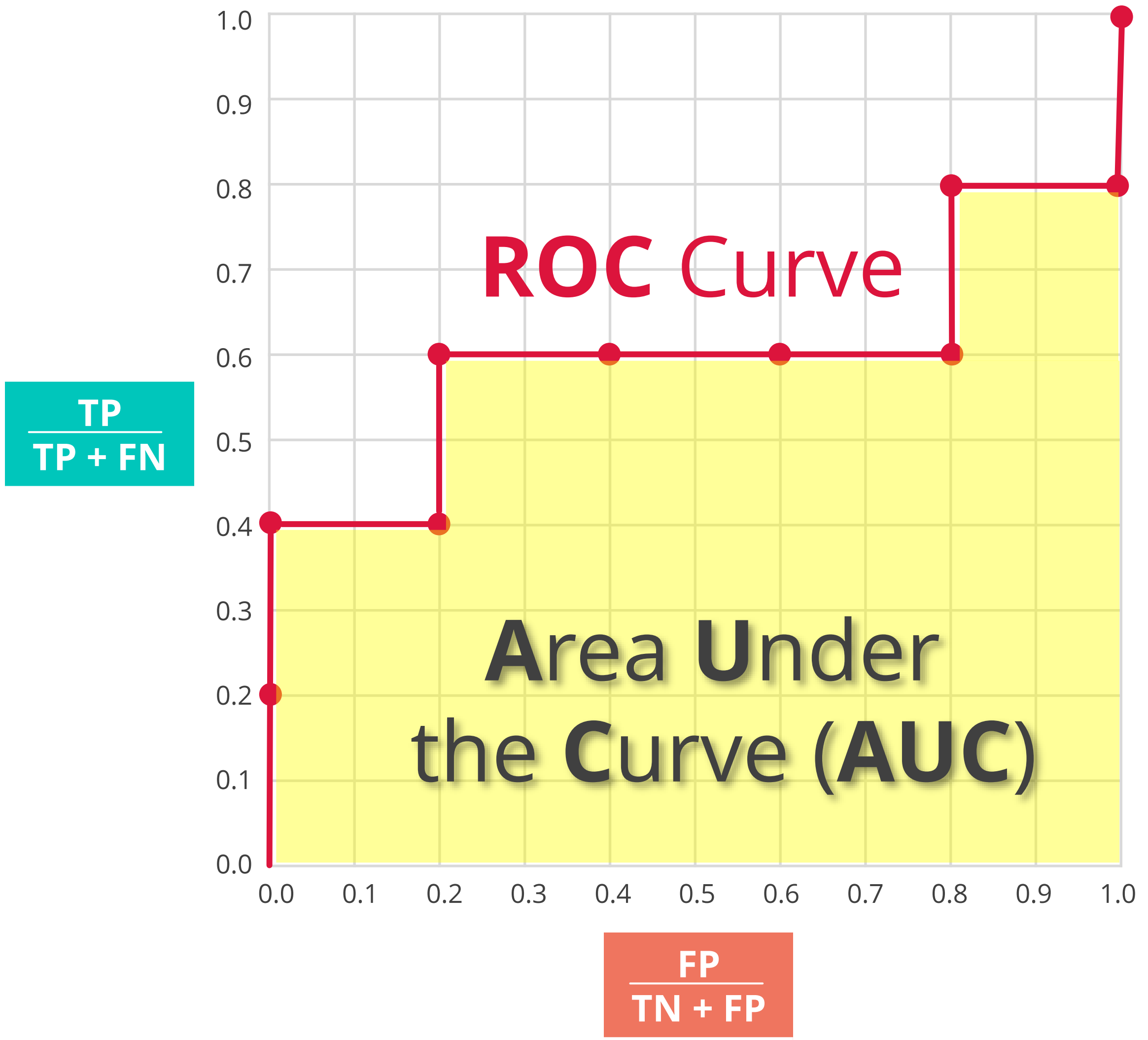
- Following the attack on Pearl Harbor in 1941, the United States army began new research to improve the **predictive accuracy** to correctly detect an enemy aircraft from their radar signals.
- They needed **a new measure** to assess various **radar receiver operators** on their ability to discriminate signal (e.g., enemy aircraft) from noise (e.g., flocks of birds).
- This is a **binary classification** problem!

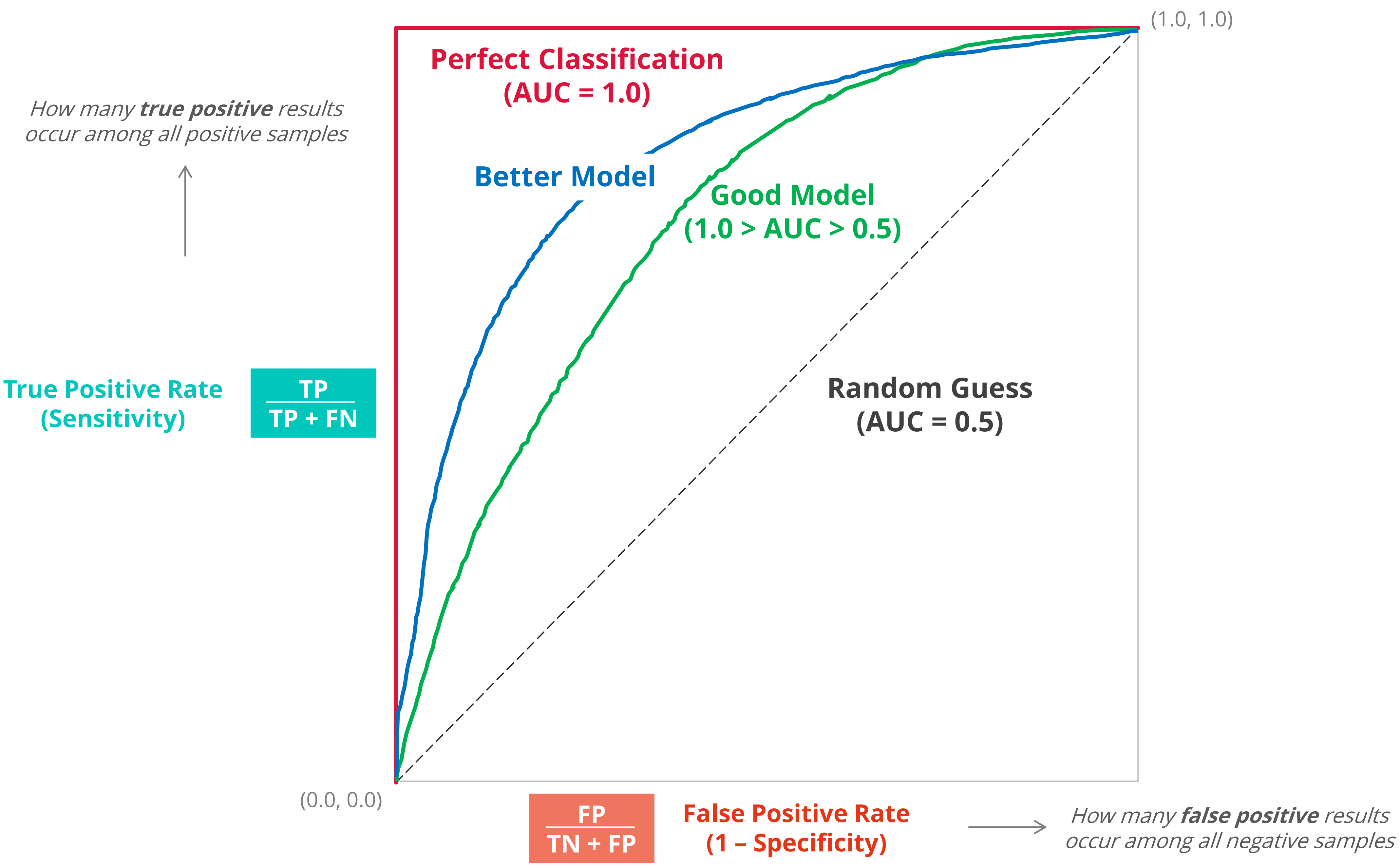
Receiver **O**perating **C**haracteristic

ROC Curve

ROC Curve

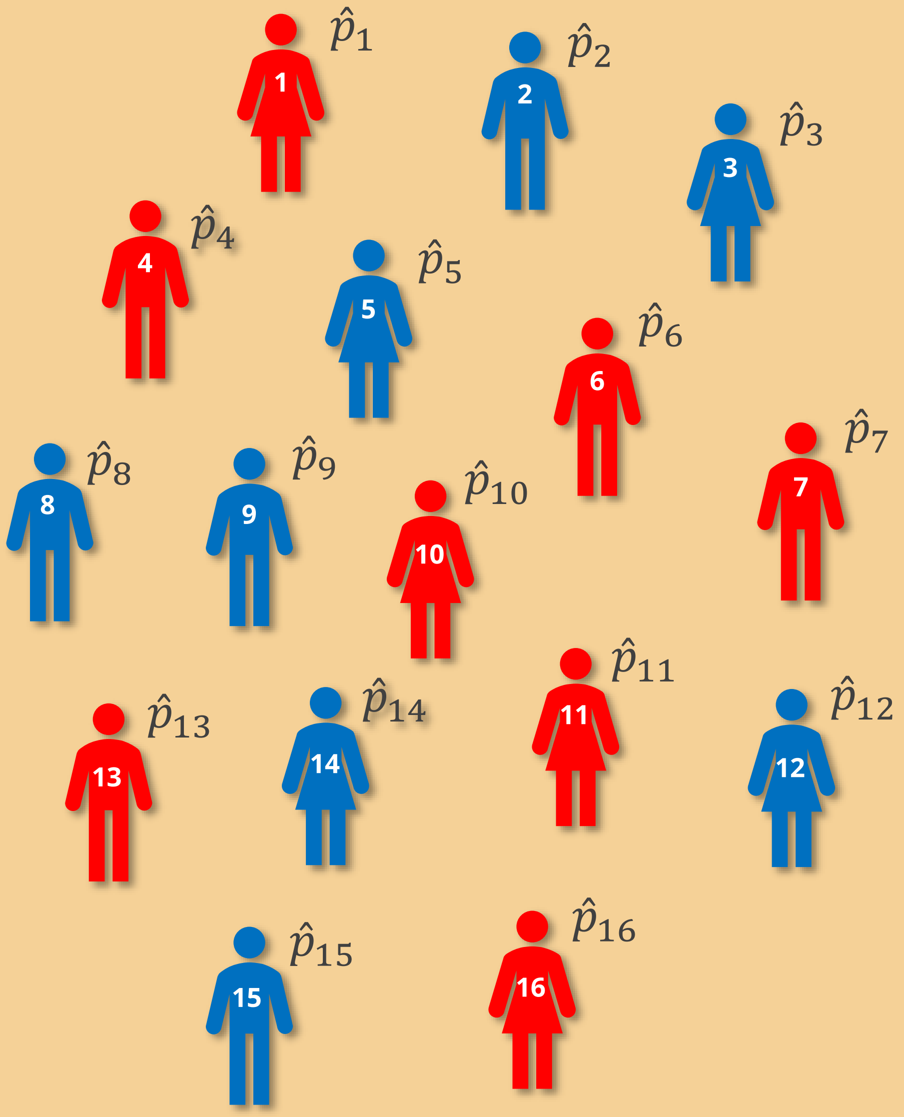
True Class (y)	Predicted Probability ($\hat{p}$) ↓	$\frac{TP}{TP + FN}$	$\frac{FP}{TN + FP}$
+	0.95		
+	0.90		
−	0.89		
+	0.71		
−	0.59		
−	0.47		
−	0.31		
+	0.29		
−	0.24		
+	0.12		





(1.0, 1.0)
Perfect Classification
(AUC = 1.0)
How many **true positive** results occur among all positive samples
Better Model
Good Model
(1.0 > AUC > 0.5)
Random Guess
(AUC = 0.5)
True Positive Rate
(Sensitivity)
$\frac{TP}{TP + FN}$
(0.0, 0.0)
$\frac{FP}{TN + FP}$
False Positive Rate
(1 – Specificity)
How many **false positive** results occur among all negative samples

$\hat{p}$ = Probability of being a Republican



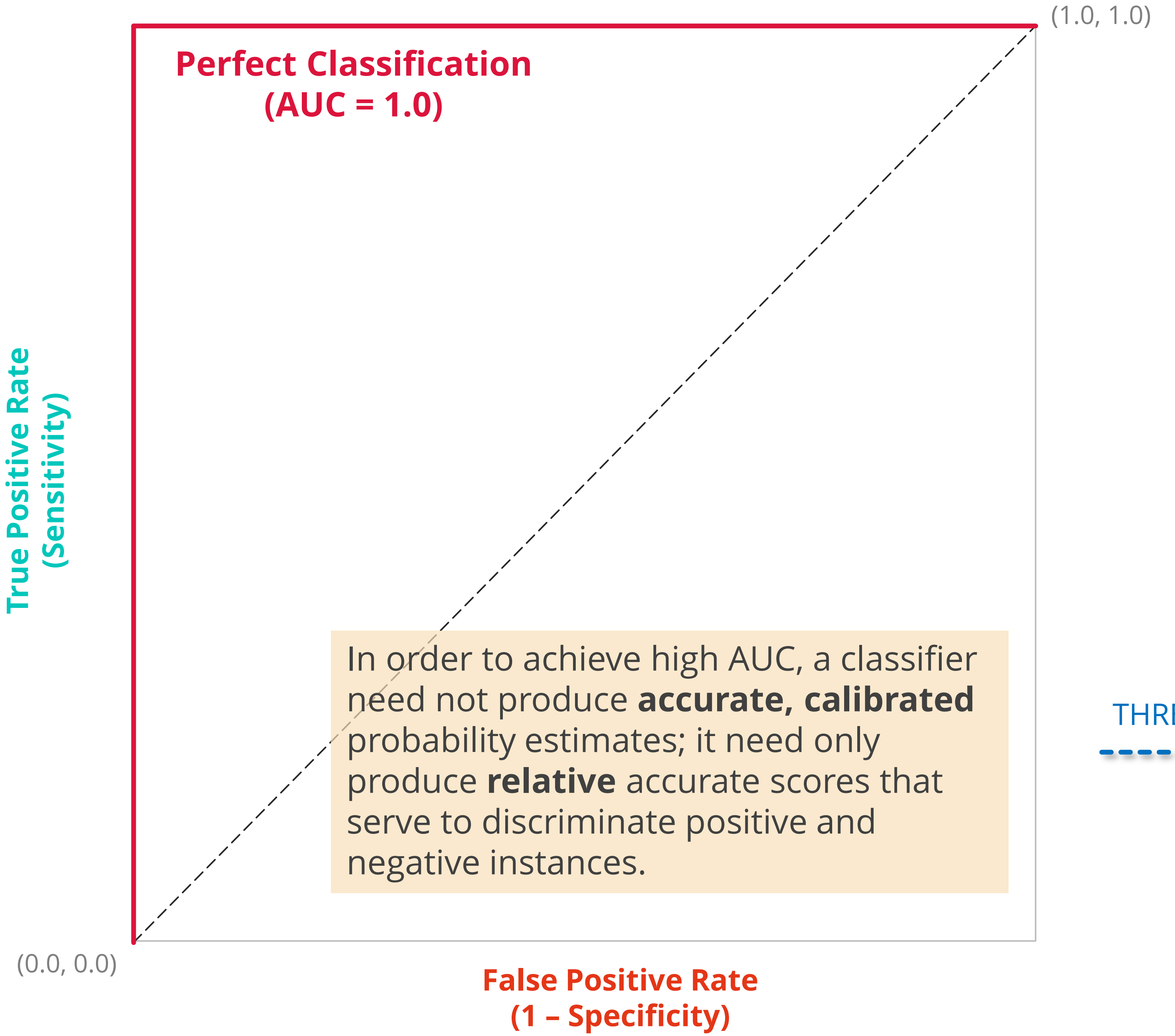
Actual Republican **Actual Democrat**

(Both randomly selected)

What is the **probability** that $\hat{p}_4$ is greater than $\hat{p}_{14}$?

↓

AUC



In order to achieve high AUC, a classifier need not produce **accurate, calibrated** probability estimates; it need only produce **relative** accurate scores that serve to discriminate positive and negative instances.

ACCURACY = 80%!

THRESHOLD = **0.5**

y	$\hat{p}$	$\hat{y}$
1	0.99999	1
1	0.99999	1
1	0.99993	1
1	0.99986	1
1	0.99964	1
1	0.99955	1
0	0.68139	1
0	0.50961	1
0	0.48880	0
0	0.44951	0

http://binf.gmu.edu/mmasso/ROC101.pdf

	MODEL **A**		MODEL **B**	
True Class (y)	**Predicted Probability ($\hat{p}$)**		**Predicted Probability ($\hat{p}$)**	
+	0.29 (−)	✖	0.11 (−)	✖
+	0.90 (+)		0.90 (+)	
−	0.31 (−)		0.31 (−)	
−	0.89 (+)	✖	0.98 (+)	✖
+	0.95 (+)		0.95 (+)	
+	0.71 (+)		0.71 (+)	
−	0.24 (−)		0.24 (−)	
+	0.12 (−)	✖	0.02 (−)	✖
−	0.59 (+)	✖	0.90 (+)	✖
−	0.47 (−)		0.47 (−)	

Which model is better?

	MODEL **A**	MODEL **B**
True Class (y)	**Predicted Probability ($\hat{p}$)**	**Predicted Probability ($\hat{p}$)**
+	0.29 (−)	0.11 (−)
−	0.89 (+)	0.98 (+)
+	0.12 (−)	0.02 (−)
−	0.59 (+)	0.91 (+)

Equivocally wrong

Emphatically wrong

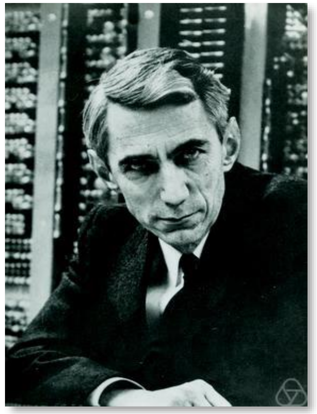
Log Loss / Cross-entropy Loss

$$logLoss = -\frac{1}{N}\sum_{i=1}^{N}\boxed{(y_i(\log(\hat{p}_i))} + \boxed{(1 - y_i)\log(1 - \hat{p}_i))}$$

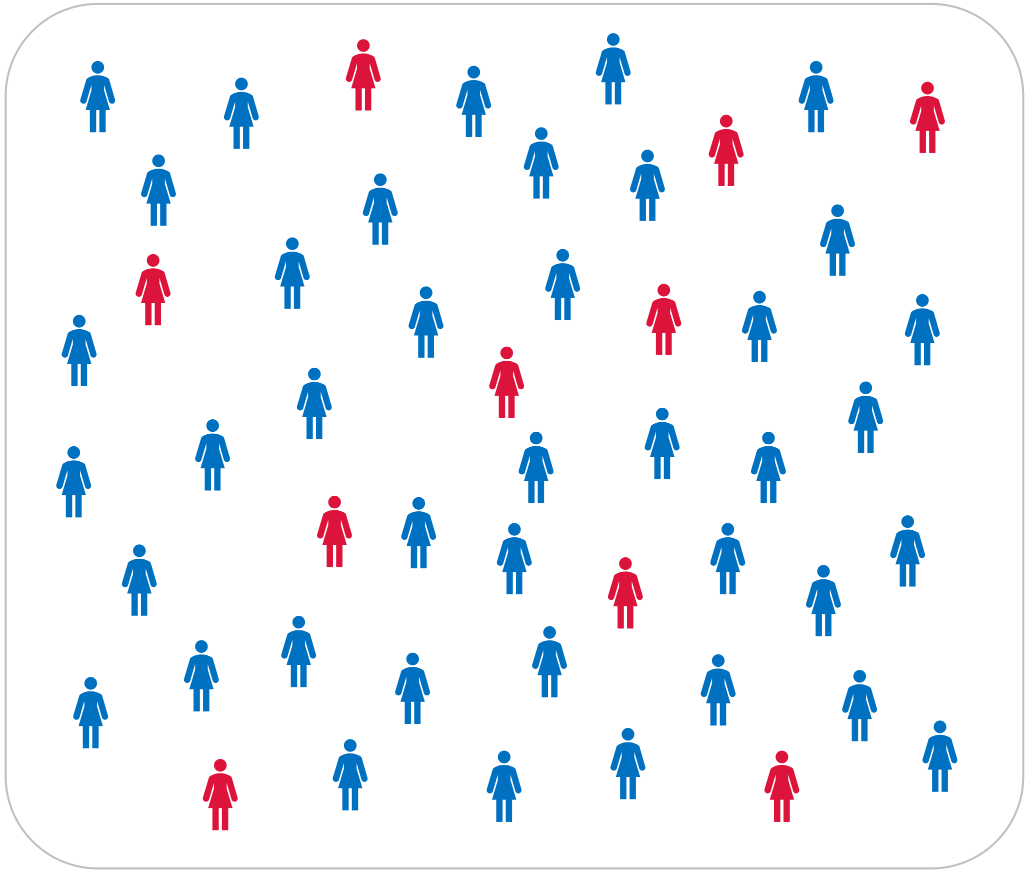
If $y_i = 1$ and $\hat{p}_i$ is high → Good!

If $y_i = 0$ and $\hat{p}_i$ is low → Good!

$$H(p,q) = -\sum_{i} p_i \log q_i$$



Model Lift and Gain

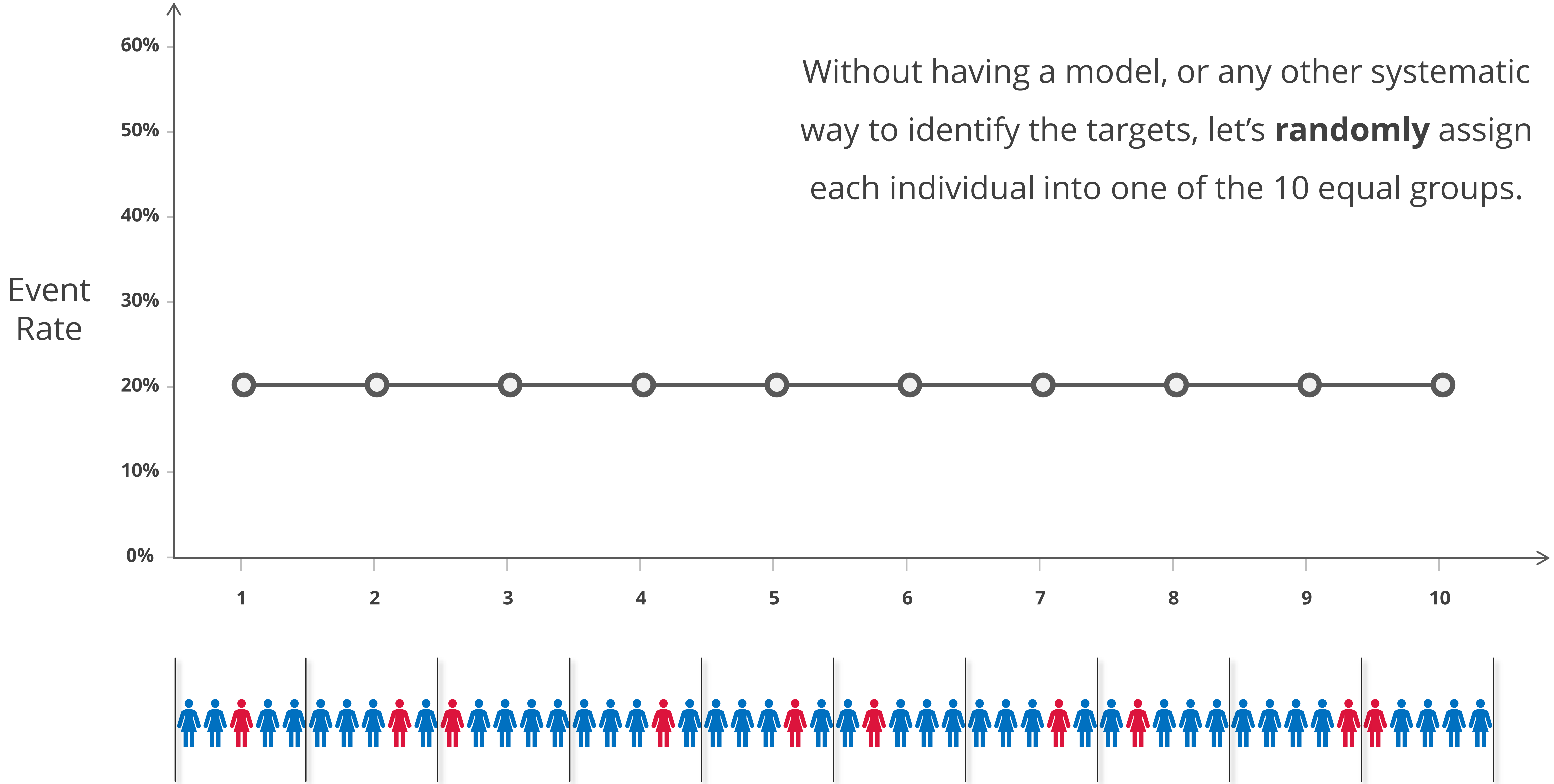


N = 50

Targets / Events = 10

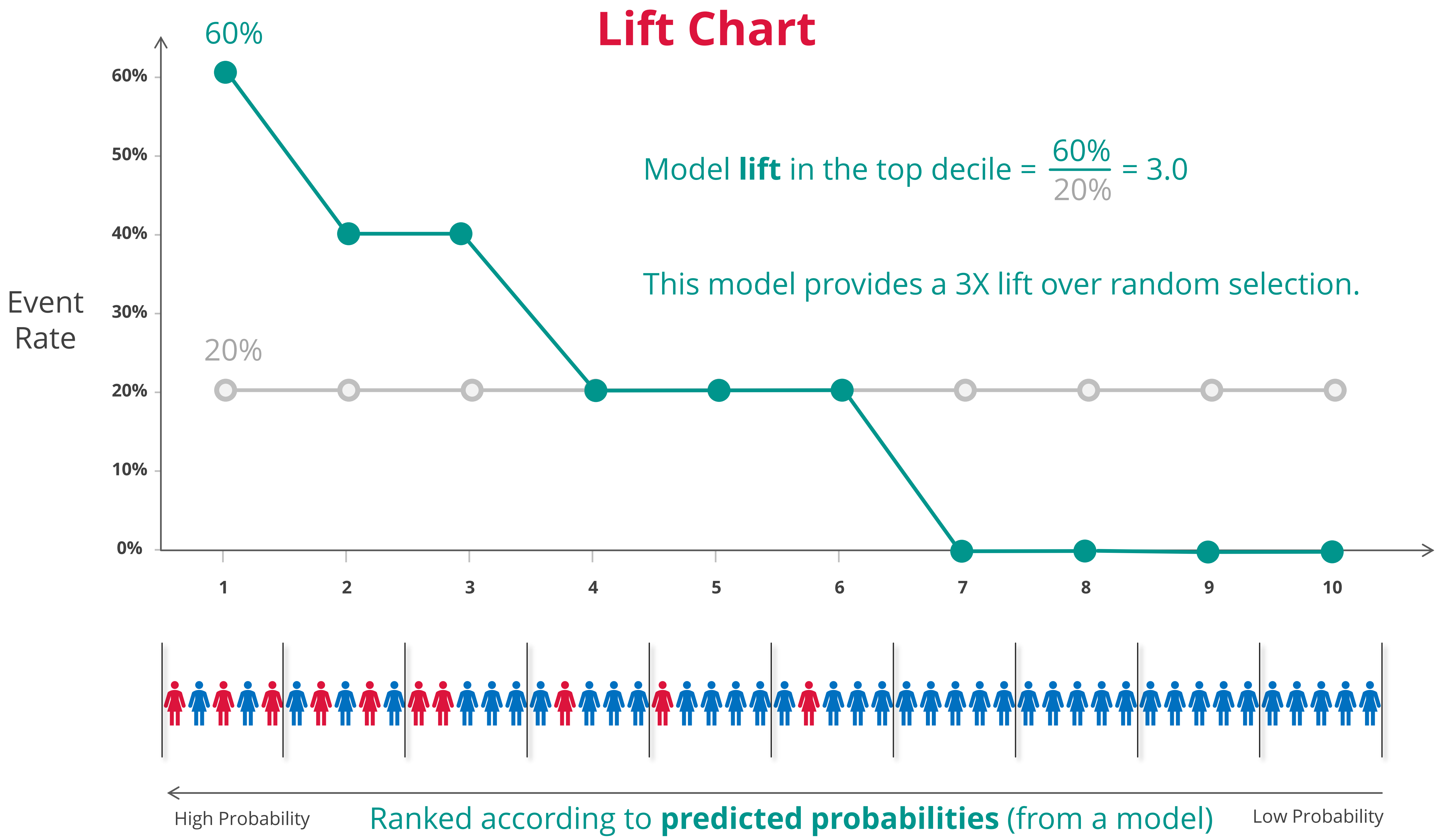
Non-events = 40

Event Rate = 20%

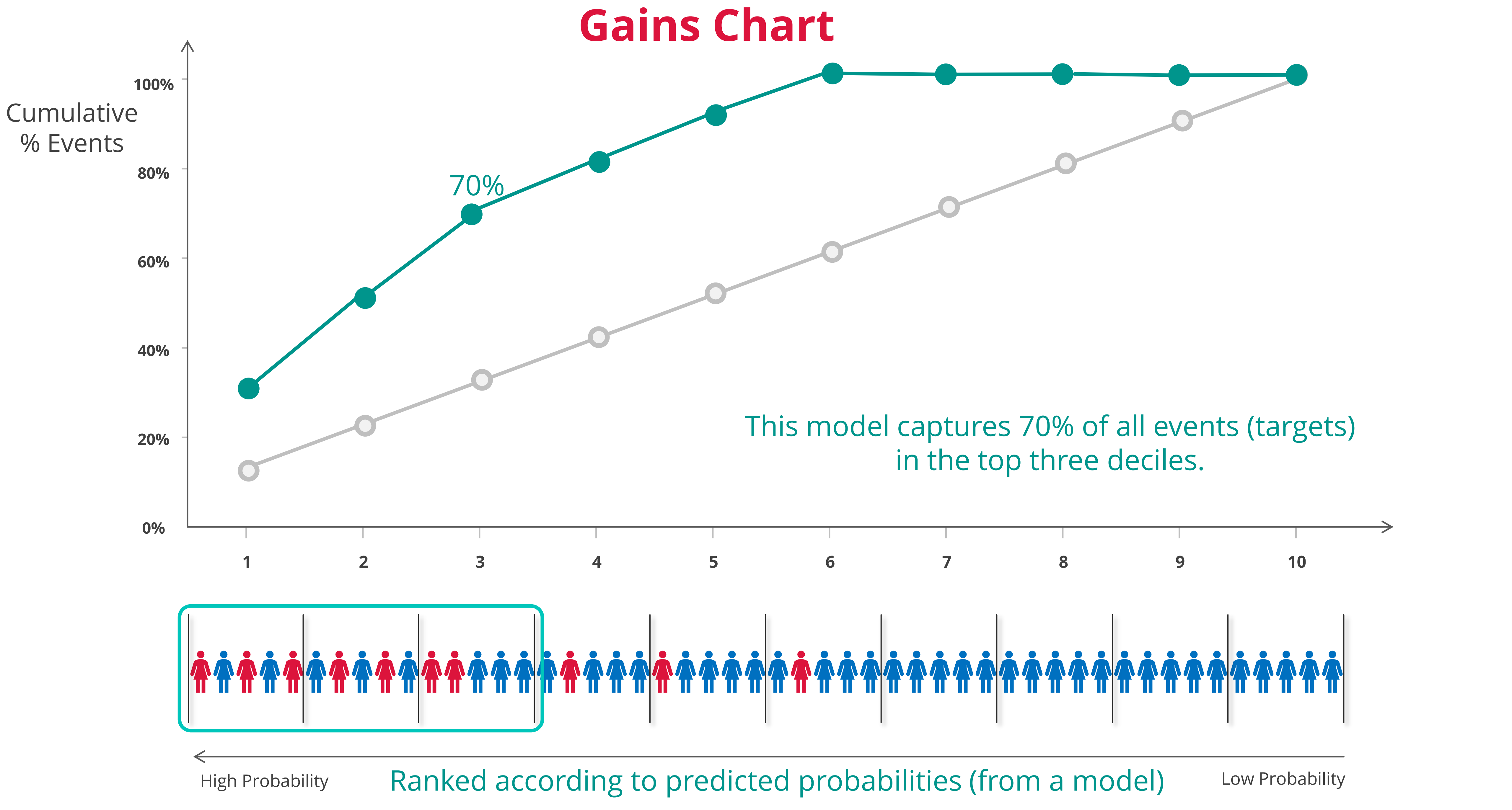


Without having a model, or any other systematic way to identify the targets, let's **randomly** assign each individual into one of the 10 equal groups.

10 equal groups, aka **deciles**



Lift Chart
60%
60%
50%
40%
30%
20%
10%
0%
Event Rate
20%
Model **lift** in the top decile = $\frac{60\%}{20\%}$ = 3.0
This model provides a 3X lift over random selection.
1
2
3
4
5
6
7
8
9
10
High Probability
Ranked according to **predicted probabilities** (from a model)
Low Probability


Gains Chart
Cumulative % Events
100%
80%
60%
40%
20%
0%
70%
This model captures 70% of all events (targets) in the top three deciles.
1
2
3
4
5
6
7
8
9
10
High Probability
Ranked according to predicted probabilities (from a model)
Low Probability

When a measure becomes a target,

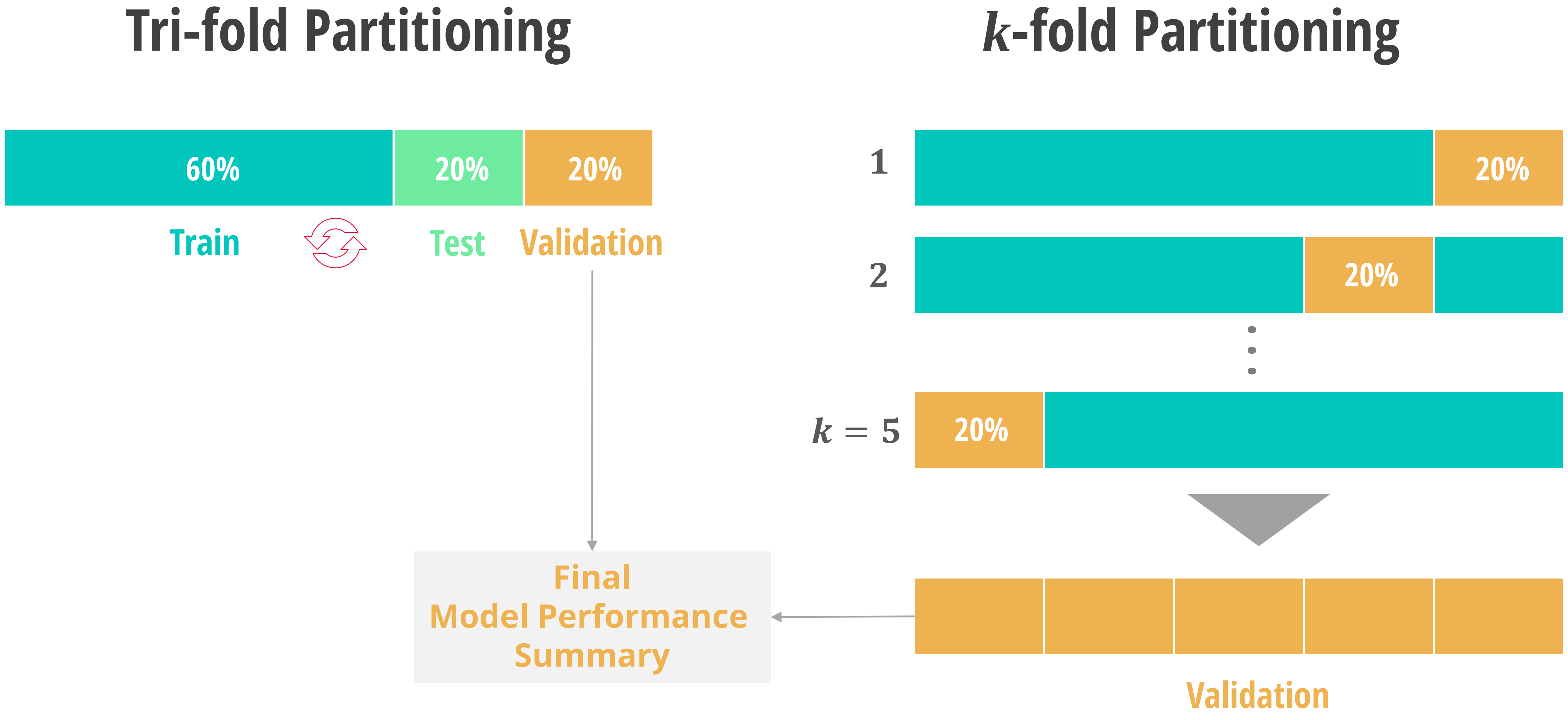
it ceases to be a good measure.

Goodhart's law

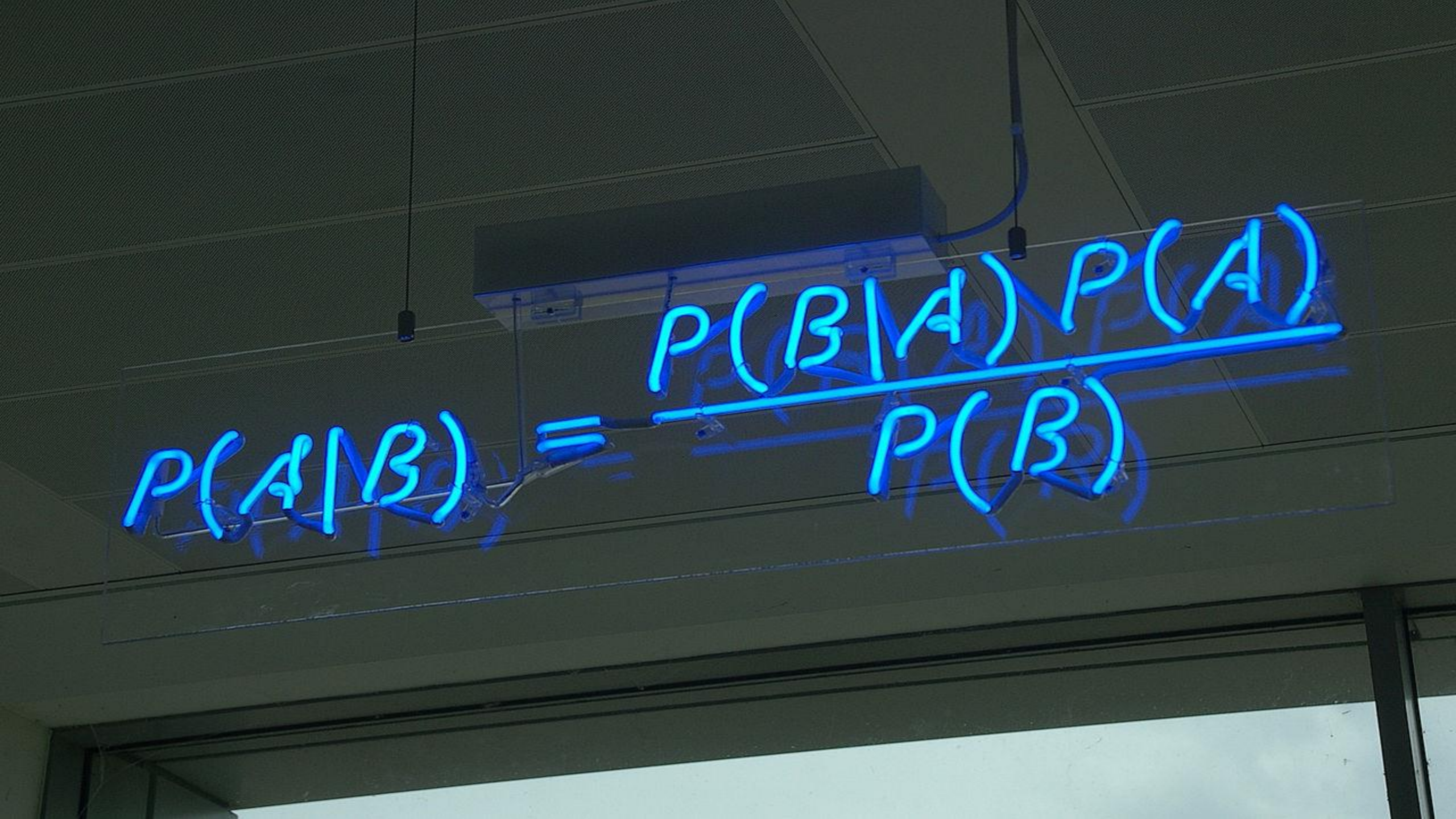
Tri-fold Partitioning

1. Split the modeling dataset into three random partitions: **train**, **test**, and **validation**.
   a) There are no hard and fast rules about the proportions of those three partitions. Suggested proportions are: 60/20/20, 60/30/10, 70/15/15, 70/20/10, or equal split.
   b) If the sample size is small, two-fold partitioning or k-fold cross-validation can be used.
2. **Fit** (train) your models on the **training** set.
3. **Assess** the model accuracy, using one or more metrics, on the **test** set.
   a) Choose metrics that align with the business objectives.
4. Repeat steps 2 and 3 to **refine** your models, e.g., tune hyper-parameters, select a smaller set of predictors.
5. **Select** the best model based on its performance on the **test** set.
6. Once the model is finalized, **measure** the accuracy of the final model using the **validation** (aka the 'hold-out') set.

k-fold Cross Validation



Bayes’ Theorem


$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Thomas Bayes
(1701 – 1761)
English statistician, philosopher, and Presbyterian minister

Bayes' theorem is the law of probability governing the strength of evidence – the rule saying how much to revise our probabilities (change our minds) when we learn a new fact or observe new evidence.

You may want to learn about Bayes' rule if you are:

- A professional who uses statistics,
- A computer programmer working in machine learning;
- A human being.†

† From Steven Pinker's lecture series on 'Bayesian Reasoning'.

A Bayesian Problem

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result has the disease, assuming you know nothing about the person's symptoms or signs?

- Most popular answer: **95%**
- Average answer: **56%**
- Correct answer: **2%** (selected by 18% of doctors!)

Credence in a hypothesis

Posterior

Is it credible to begin with?

Prior

Is it more likely to generate the observed data?

Likelihood

$$p(\text{Hypothesis} \mid \text{Data}) = \frac{p(\text{Hypothesis}) \times p(\text{Data} \mid \text{Hypothesis})}{p(\text{Data})}$$

Marginal

Is the evidence (data) unlikely in general?

Credence
in a diagnosis

Is the disease
common?

Does the disease usually
have those symptoms?

$$p(\text{Disease} \mid \text{Symptoms}) = \frac{p(\text{Disease}) \times p(\text{Symptoms} \mid \text{Disease})}{p(\text{Symptoms})}$$

Are those symptoms unusual across the board?

If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result has the disease, assuming you know nothing about the person's symptoms or signs?

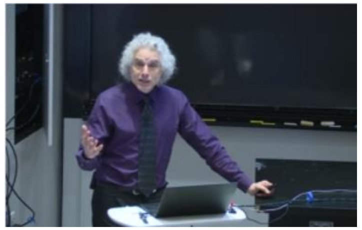
Prevalence = **0.001**

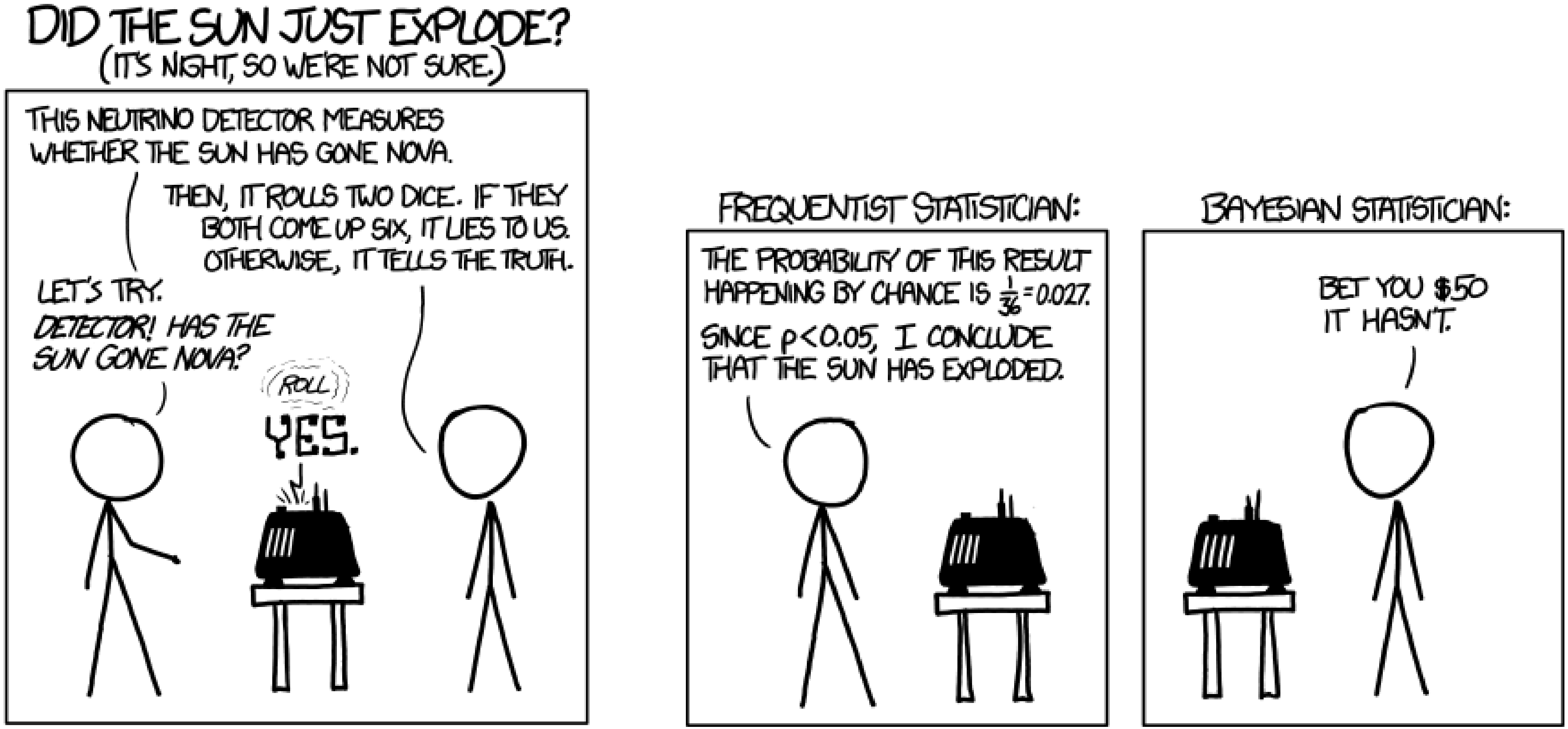
Sensitivity = **1.0** (assume)

$$p(\text{Disease} \mid \text{Test positive}) = \frac{p(\text{Disease}) \times p(\text{Test positive} \mid \text{Disease})}{p(\text{Test positive})} = \mathbf{0.0195} \approx \mathbf{2\%}$$

$$p(\text{Test positive \& Disease}) + p(\text{Test positive \& Healthy})$$

(**1.0 * 0.001**) + (**0.05 * 0.999**)




DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}=0.027$.
SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.