



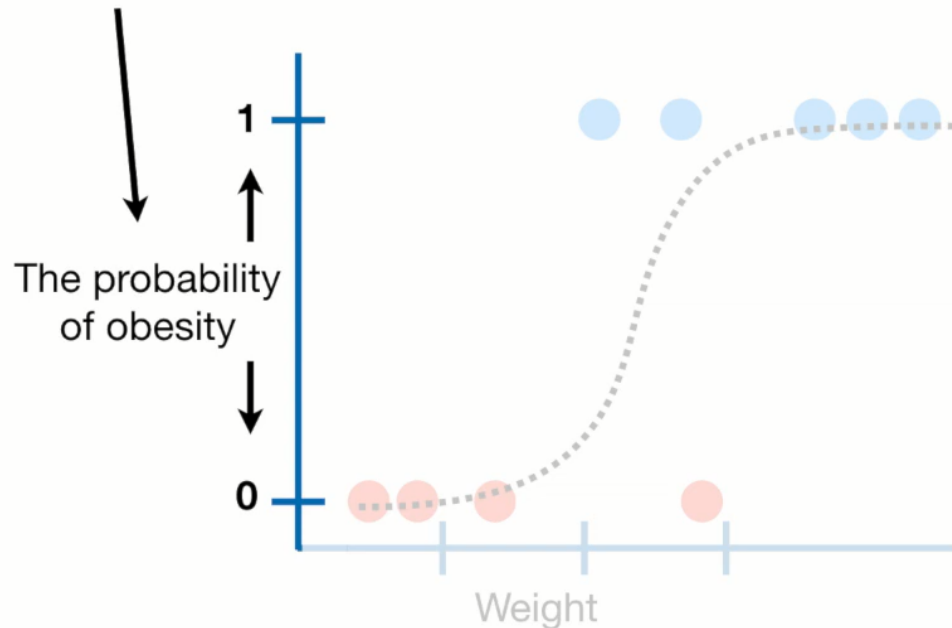
ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA – EDIZIONE 2

Operazione Rif. PA 2019-11596/RER “Anticipare la crescita con le nuove competenze sui Big Data - Edizione 2”, approvata dalla Regione Emilia-Romagna con DGR n° 789 del 20 maggio 2019 e co-finanziata dal Fondo Sociale Europeo PO 2014-2020

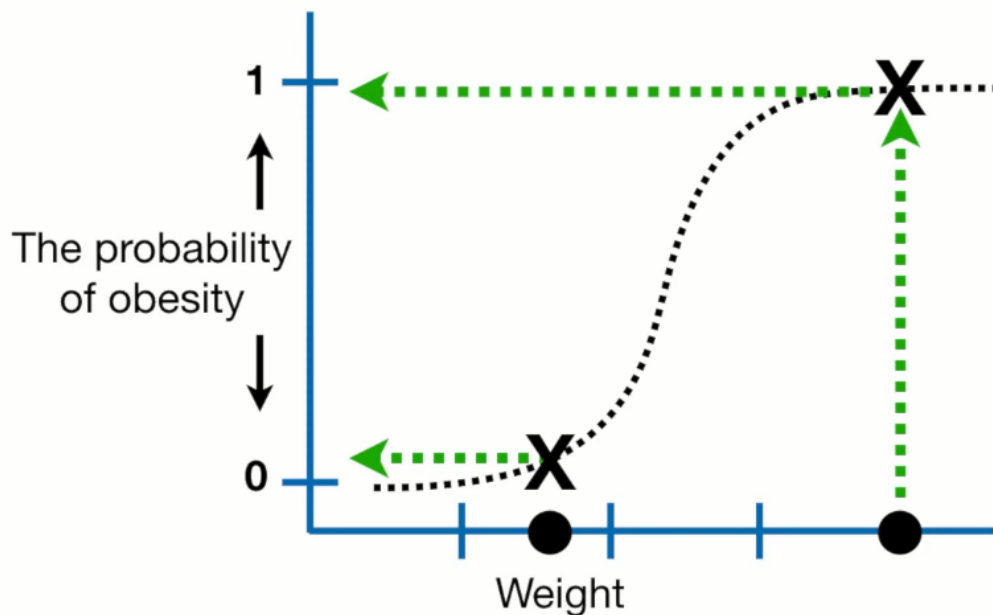
Prog. 4 Ed. 7 Titolo “Tecnologie & Software di Data Science



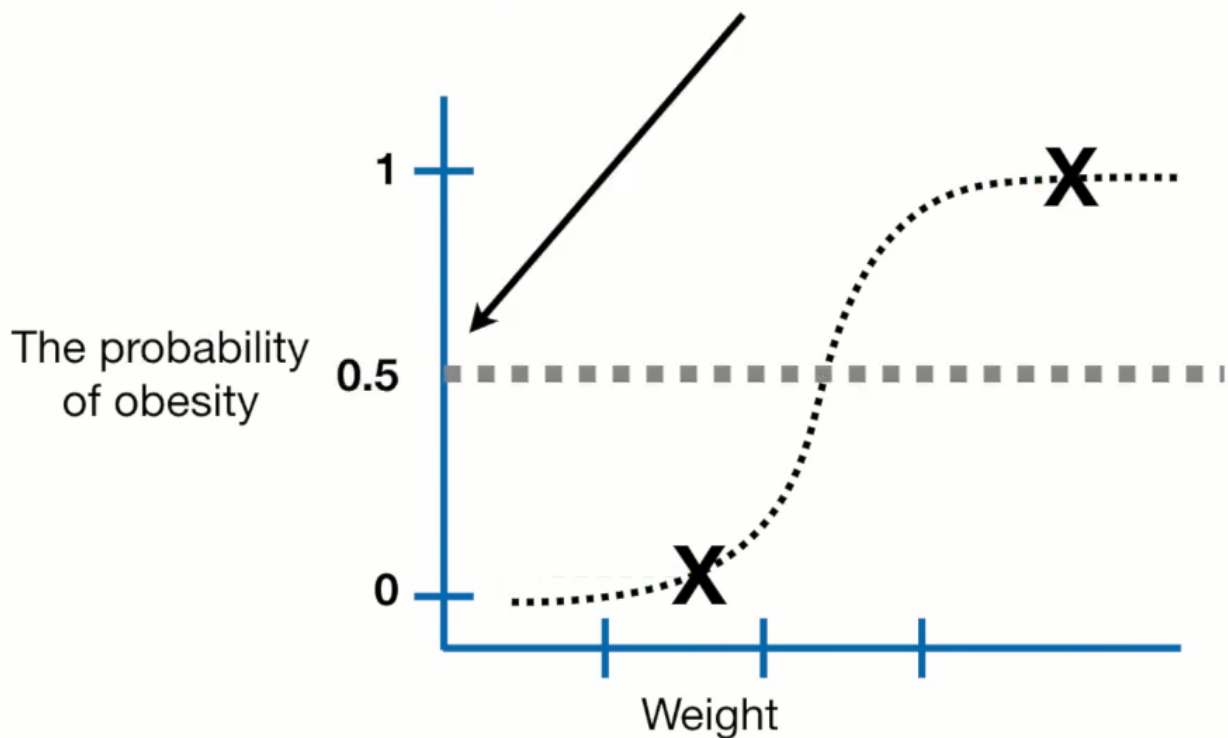
When we're doing Logistic Regression, the y-axis is converted to the probability that a mouse **is obese**.



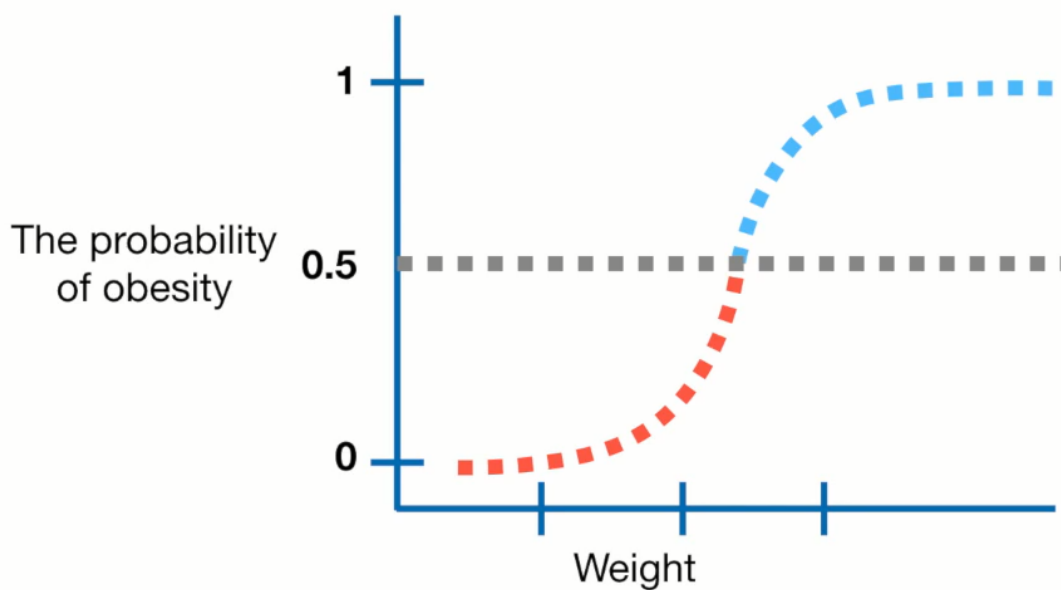
So this Logistic Regression tells us the **probability** that a mouse is **obese** based on its weight.

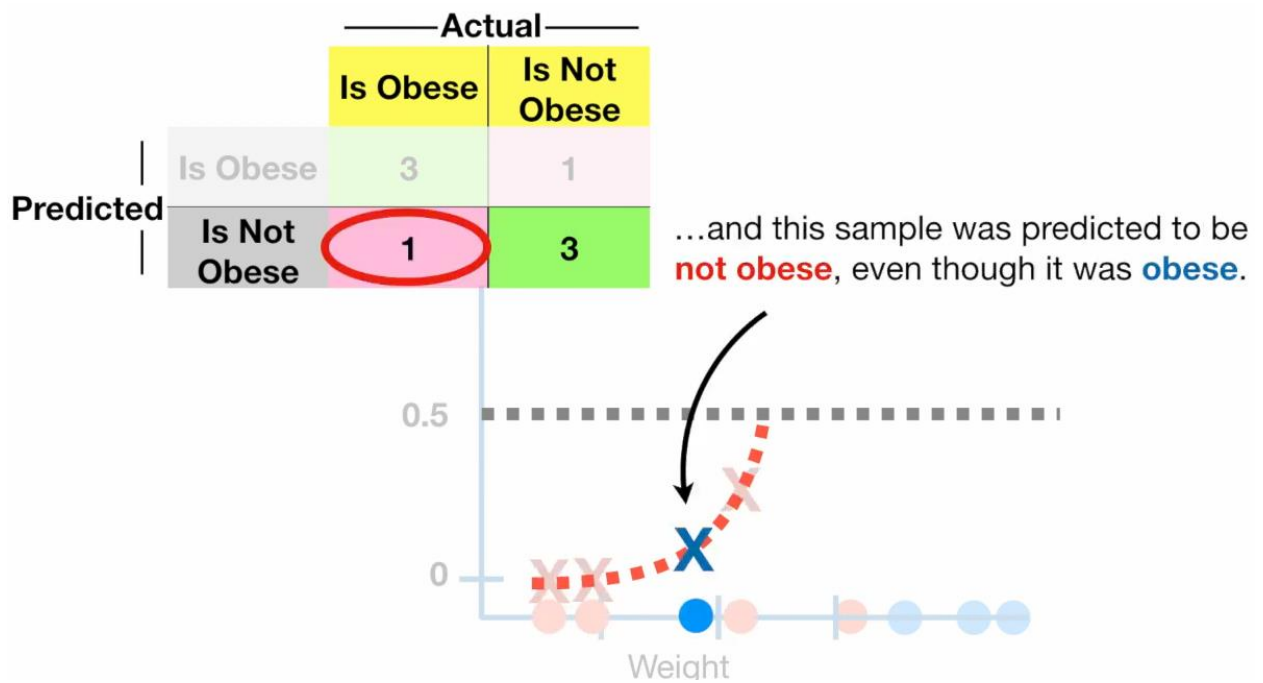
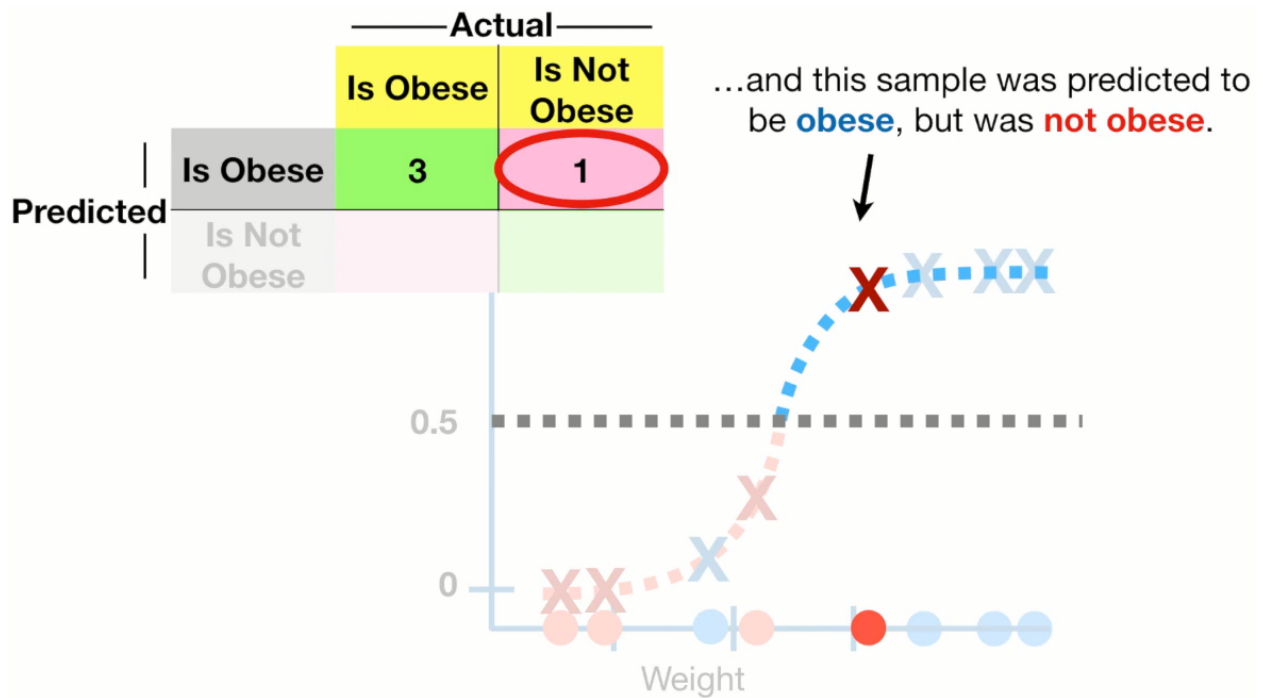


One way to classify mice is to set a threshold at **0.5**...



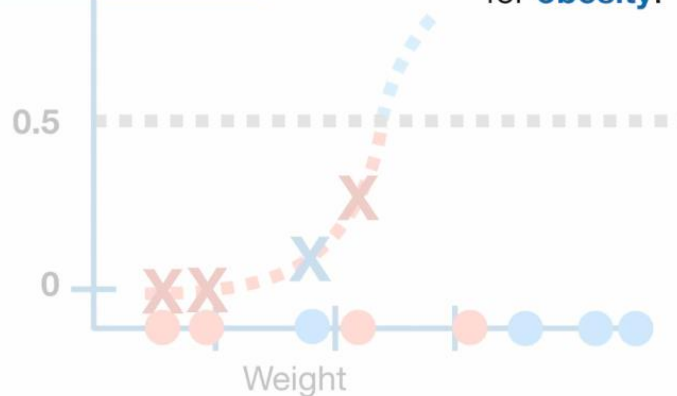
To evaluate the effectiveness of this Logistic Regression, with the classification threshold set to **0.5**, we can test it with mice that we know are **obese** or **not obese**.





		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	3	1
	Is Not Obese	1	3

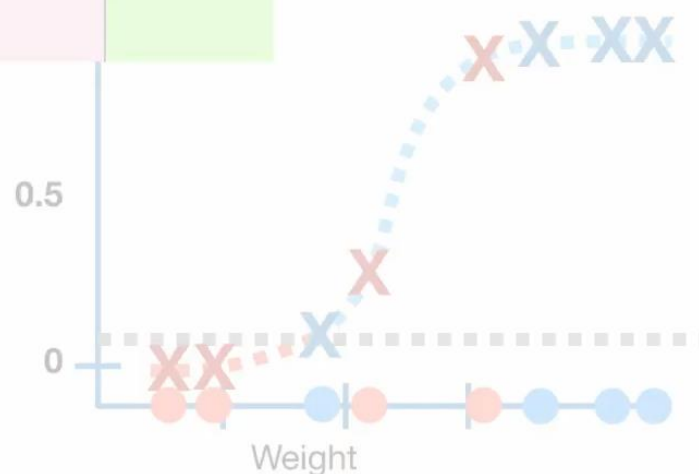
Once the **Confusion Matrix** is filled in, we can calculate **Sensitivity** and **Specificity** to evaluate this Logistic Regression when **0.5** is the threshold for **obesity**.

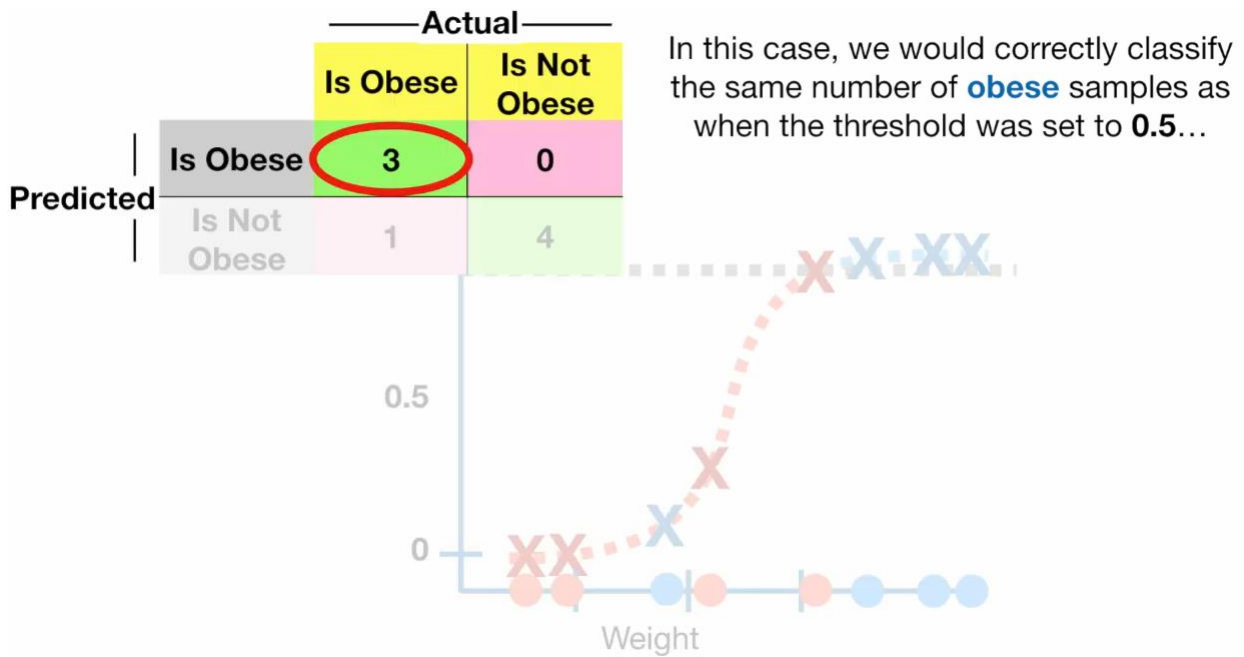
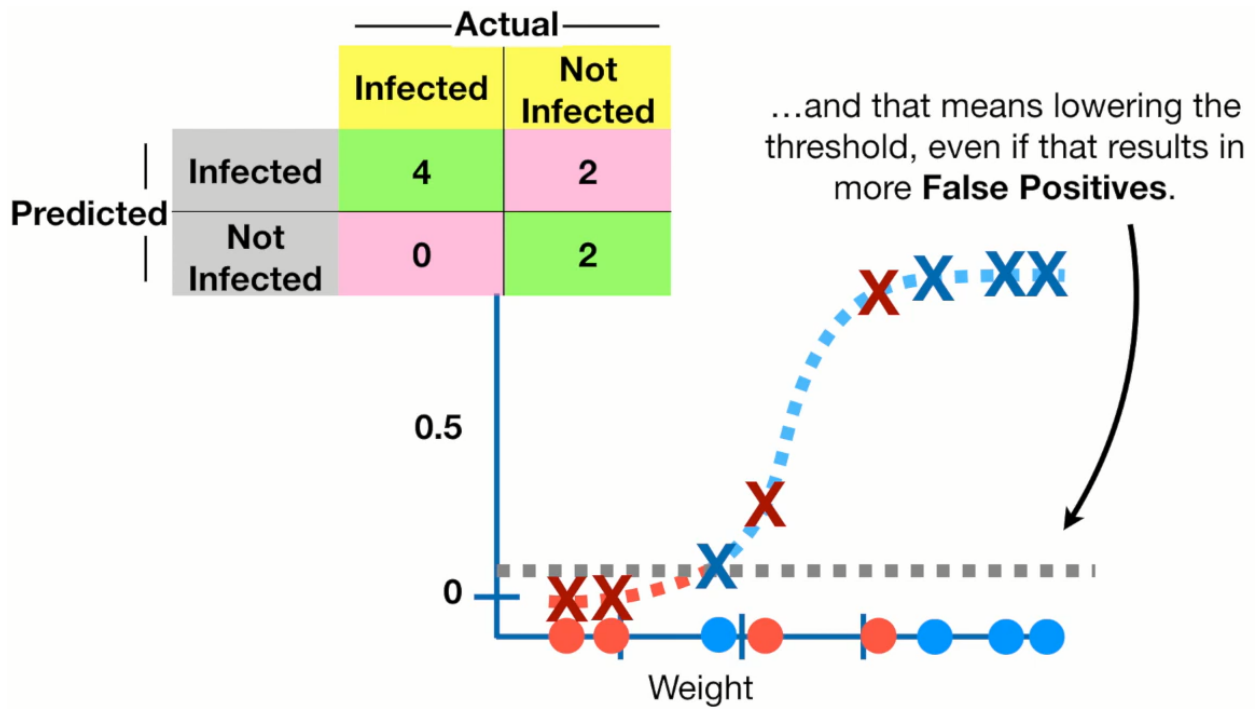


Now let's talk about what happens when we use a different threshold for deciding if a sample is **obese** or **not**.

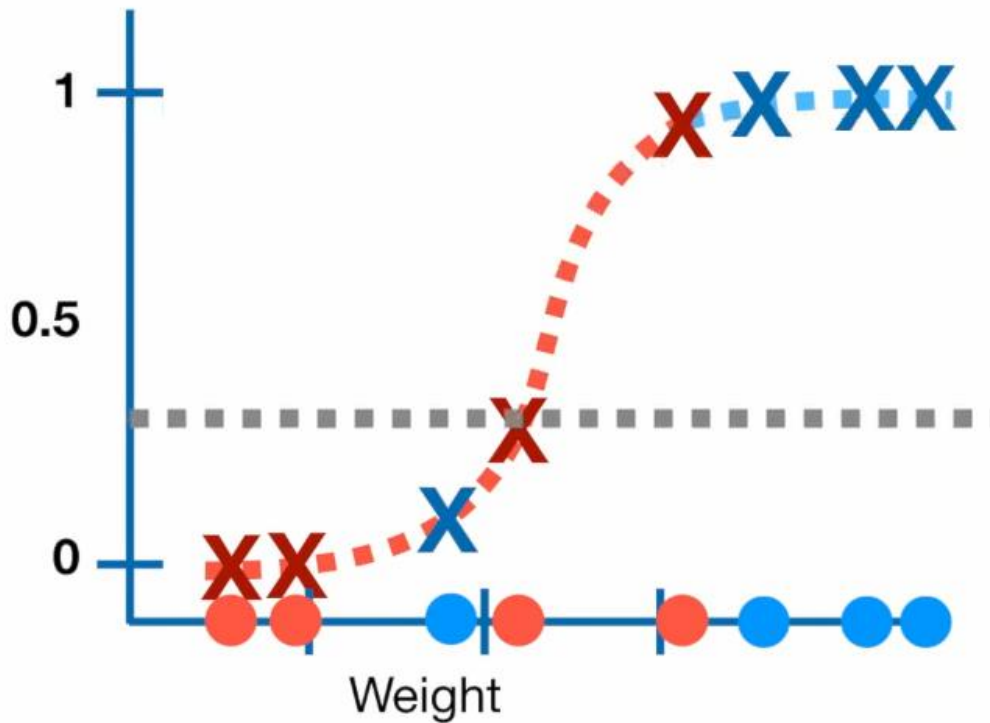
		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	4	2
	Is Not Obese		

...but it would also increase the number of **False-Positives**.

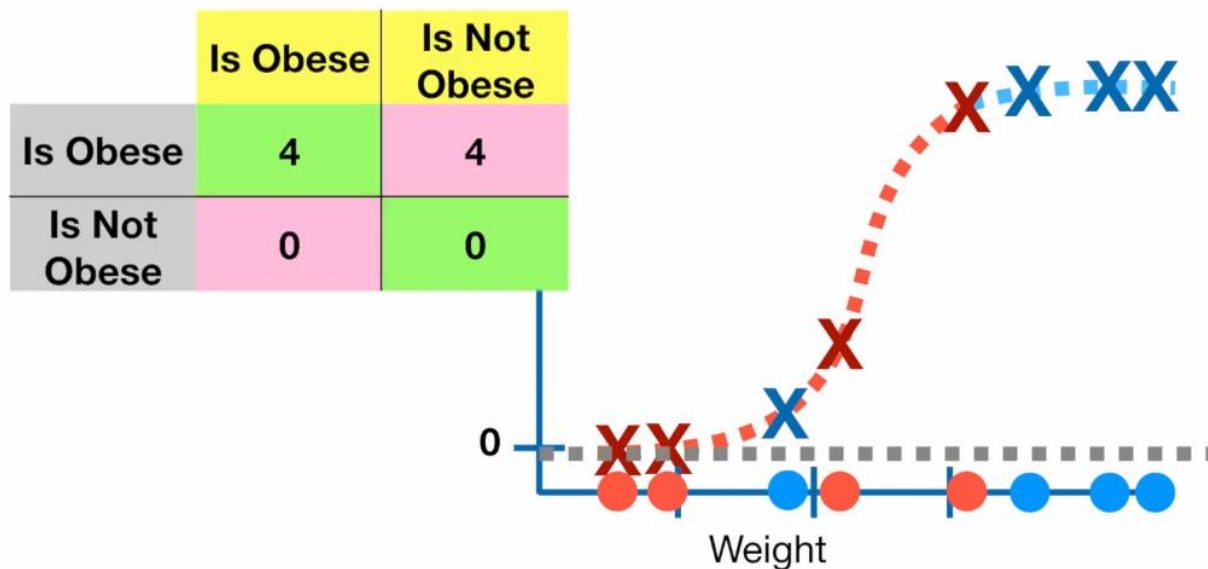




...but the threshold could be set to anything between 0 and 1.



But even if we made one confusion matrix for each threshold that mattered, it would result in a confusingly large number of confusion matrices.

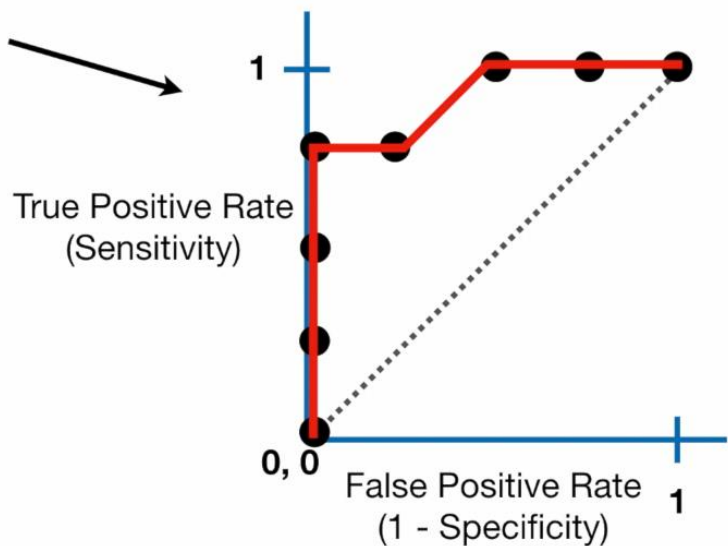


But even if we made one confusion matrix for each threshold that mattered, it would result in a confusingly large number of confusion matrices.

	Is Obese	Is Not Obese		Is Obese	Is Not Obese
Is Obese	4	4	Is Obese	4	2
Is Not Obese	0	0	Is Not Obese	0	2

	Is Obese	Is Not Obese		Is Obese	Is Not Obese
Is Obese	4	3	Is Obese	3	2
Is Not Obese	0	1	Is Not Obese	1	2

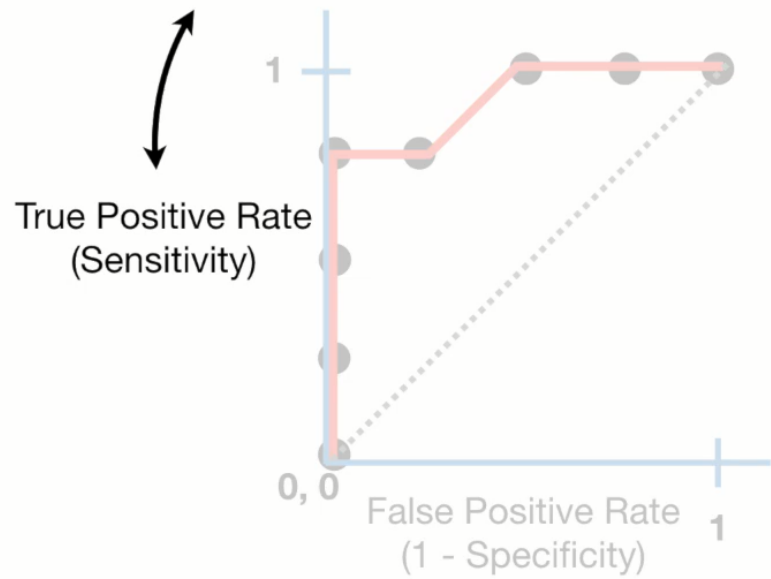
So instead of being overwhelmed with confusion matrices, **Receiver Operator Characteristic (ROC) graphs** provide a simple way to summarize all of the information.



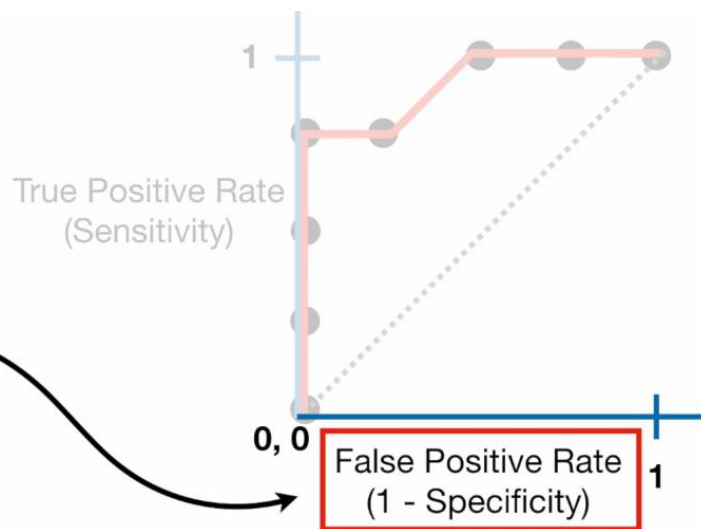
The y-axis shows the **True Positive Rate**, which is the same thing as **Sensitivity**.

True Positive Rate (Sensitivity)

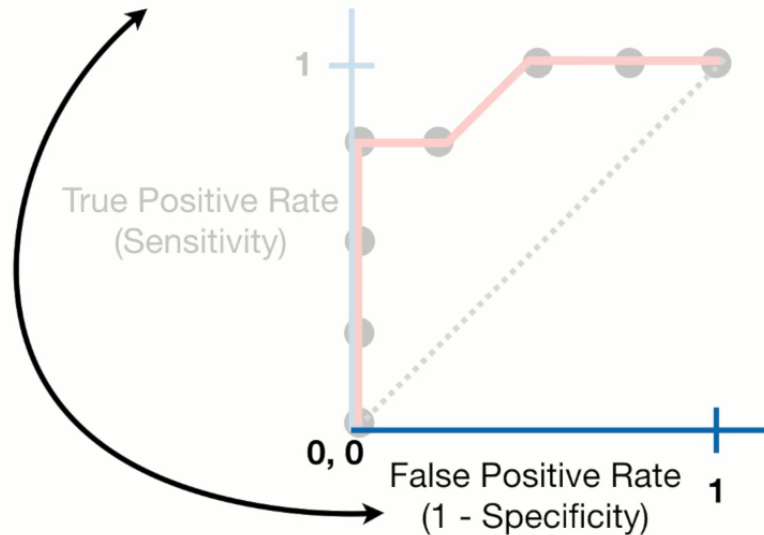
$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$



The x-axis shows the **False Positive Rate**, which is the same thing as **1 - Specificity**.



$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

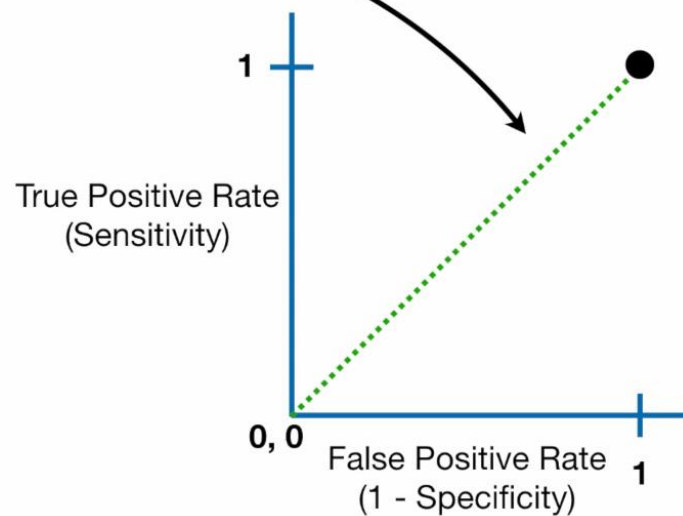


$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

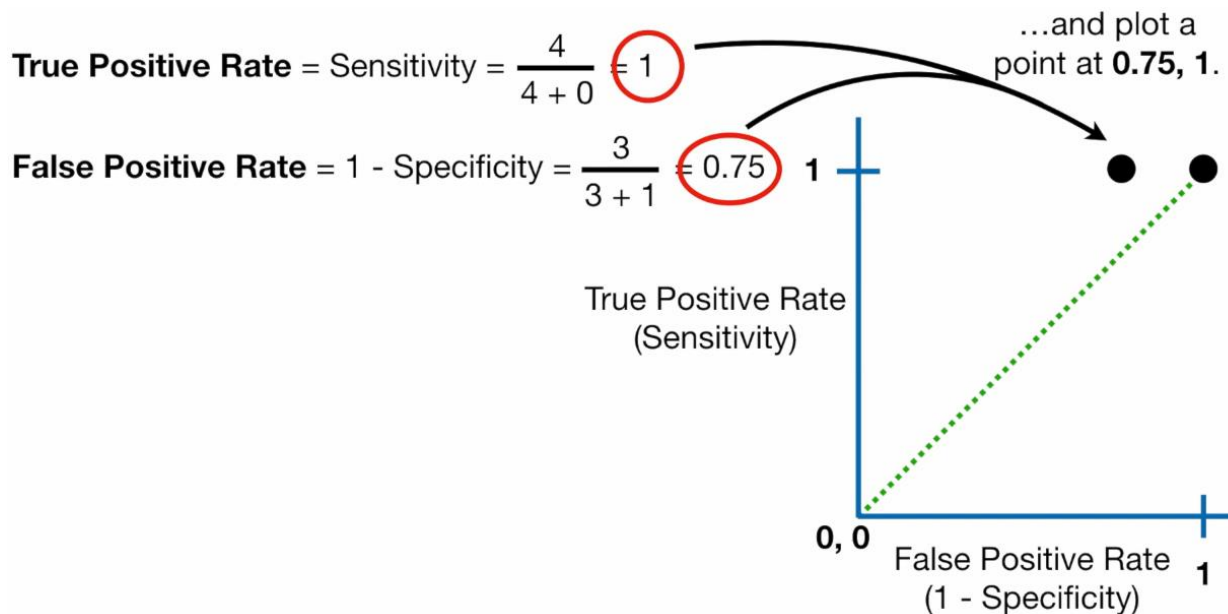
		Actual	
		Is Obese	Is Not Obese
Predicted	Is Obese	True Positives	False Positives
	Is Not Obese	False Negatives	True Negatives

The **False Positive Rate** tells you the proportion of **not obese** samples that were incorrectly classified and are **False Positives**.

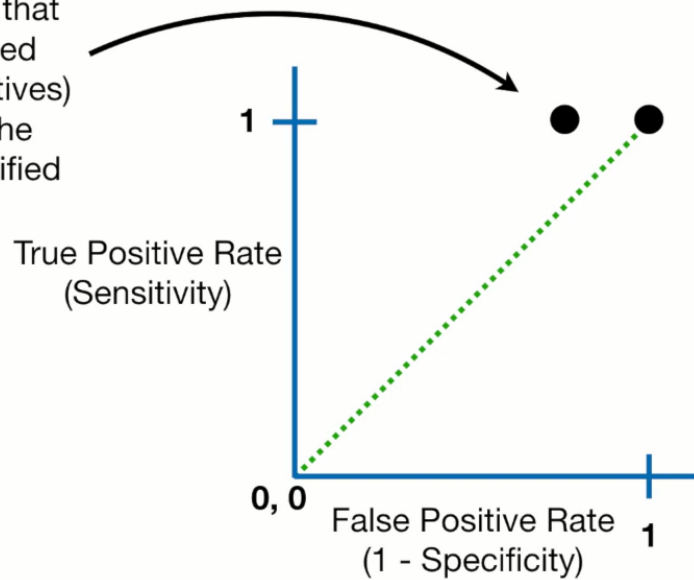
This **green diagonal line** shows where the
True Positive Rate = False Positive Rate



Any point on this **line** means that the
proportion of *correctly* classified obese
 samples is the same as the **proportion of**
incorrectly classified samples that are **not**
obese.

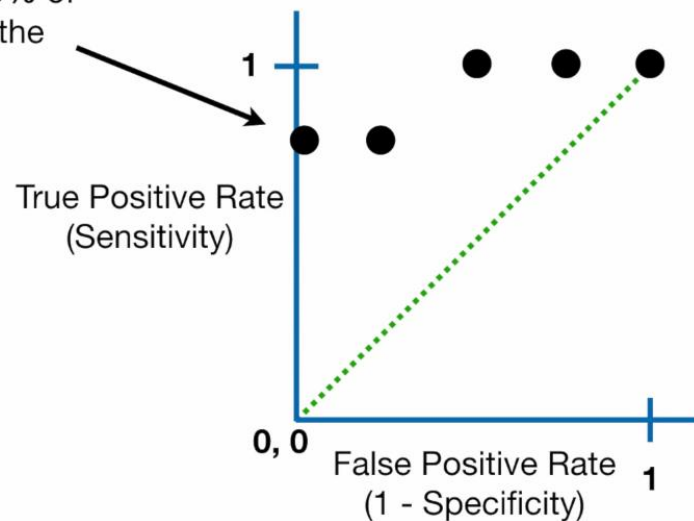


Since the new point **(0.75, 1)** is to the left of the **dotted green line**, we know that the proportion of correctly classified samples that were **obese** (true positives) *is greater* than the proportion of the samples that were incorrectly classified as **obese** (false positives).

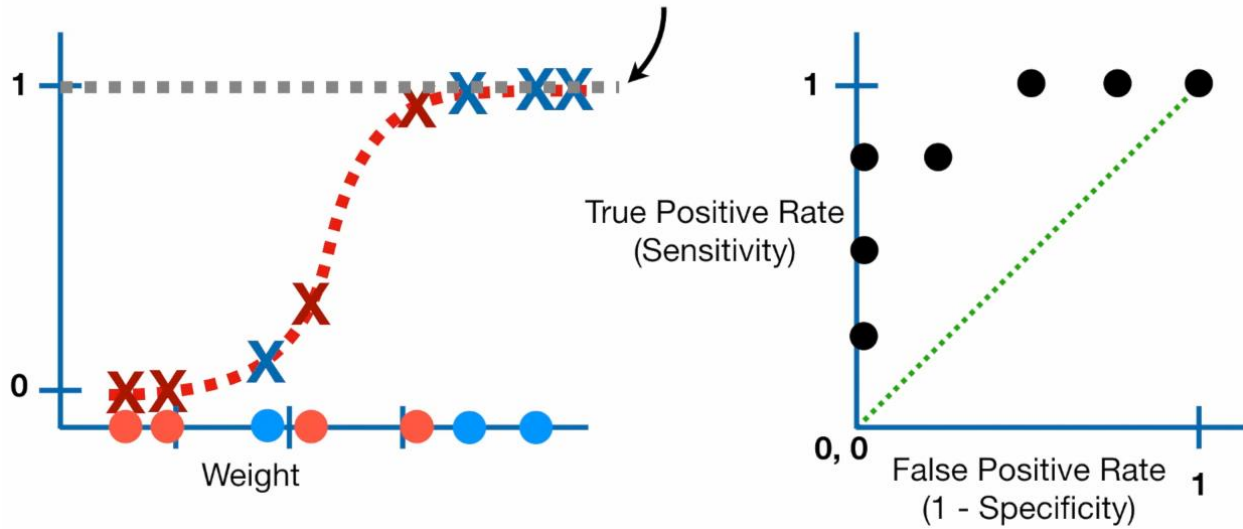


In other words, the new threshold for deciding if a sample is **obese** or **not** is better than the first one.

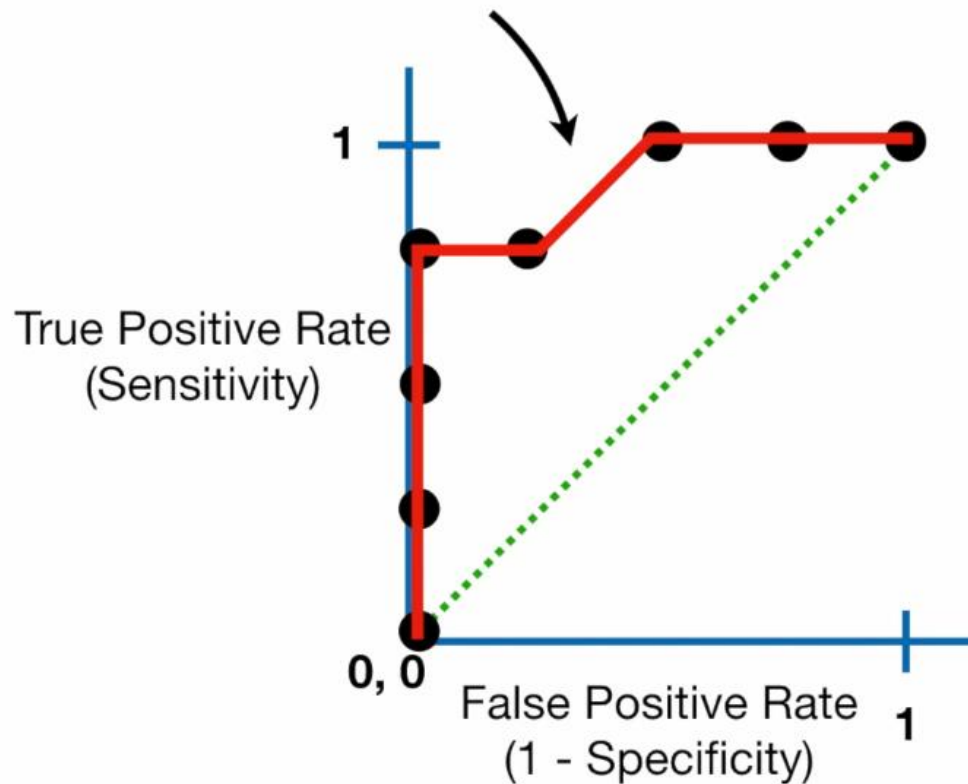
The threshold represented by the new point **(0, 0.75)** correctly classified **75%** of the **obese** samples and **100%** of the samples that were **not obese**.



Lastly, we choose a threshold that classifies all of the samples as **not obese**...

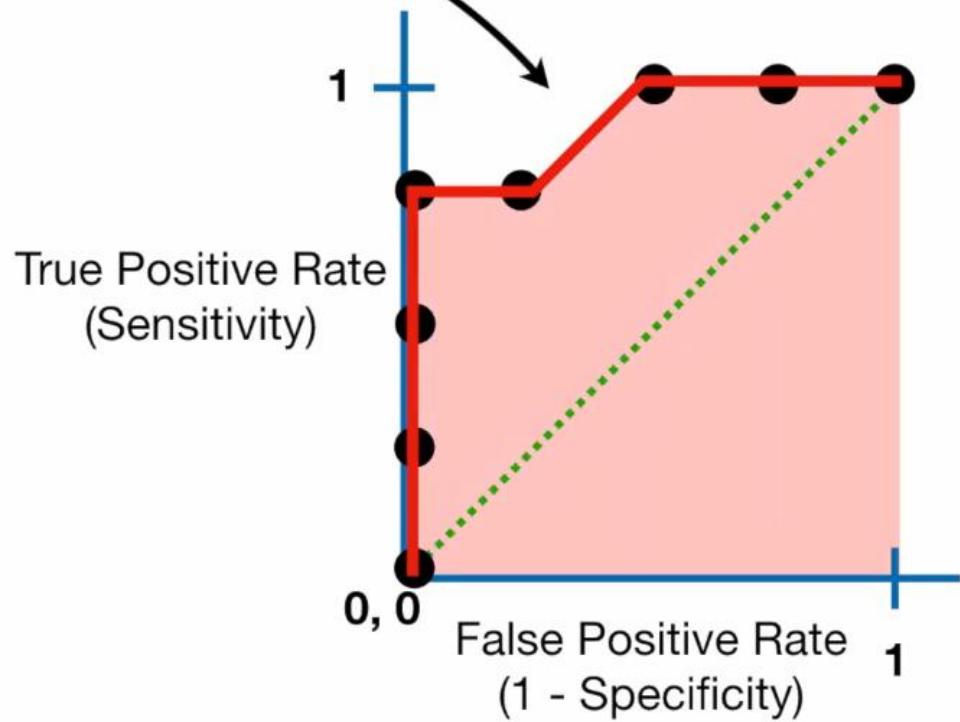


If we want, we can connect the dots...

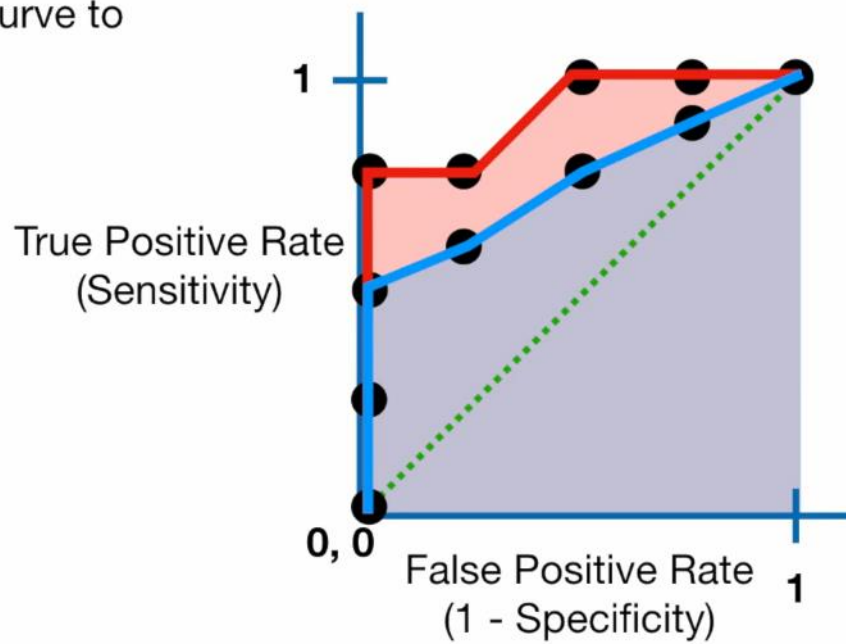


The **ROC** graph summarizes all of the confusion matrices that each threshold produced.

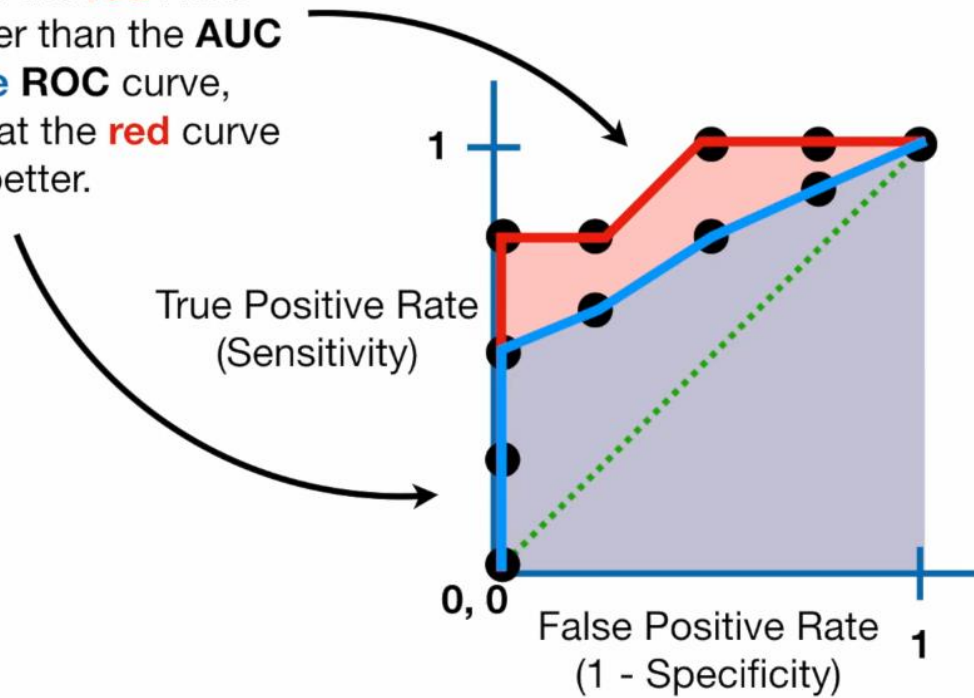
The **AUC** (Area Under the Curve) is **0.9**



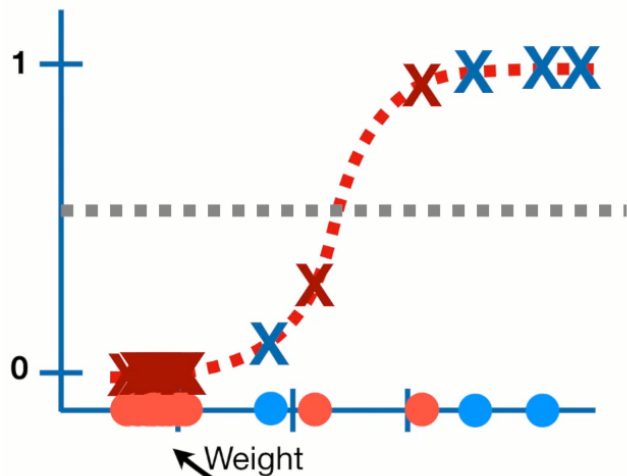
The **AUC** makes it easy to compare one **ROC** curve to another.



The **AUC** for the **red ROC** curve is greater than the **AUC** for the **blue ROC** curve, suggesting that the **red** curve is better.



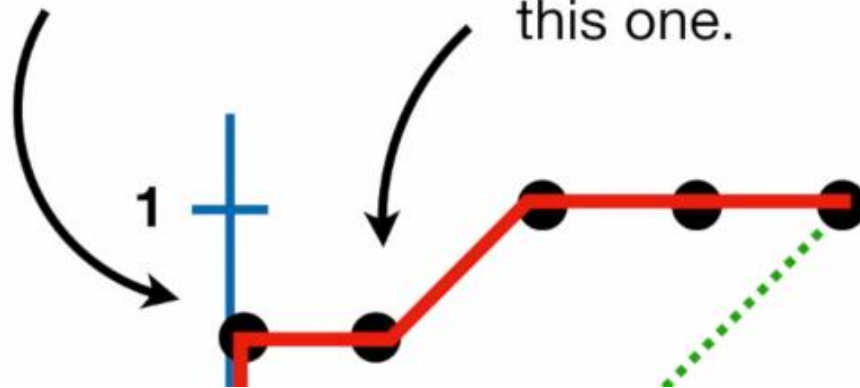
$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$



In practice, this sort of imbalance occurs when studying a rare disease. In this case, the study will contain many more people without the disease than with the disease.

ROC curves make it easy to identify the best threshold for making a decision...

This threshold... ...is better than this one.



...and the **AUC** can help you decide which categorization method is better.