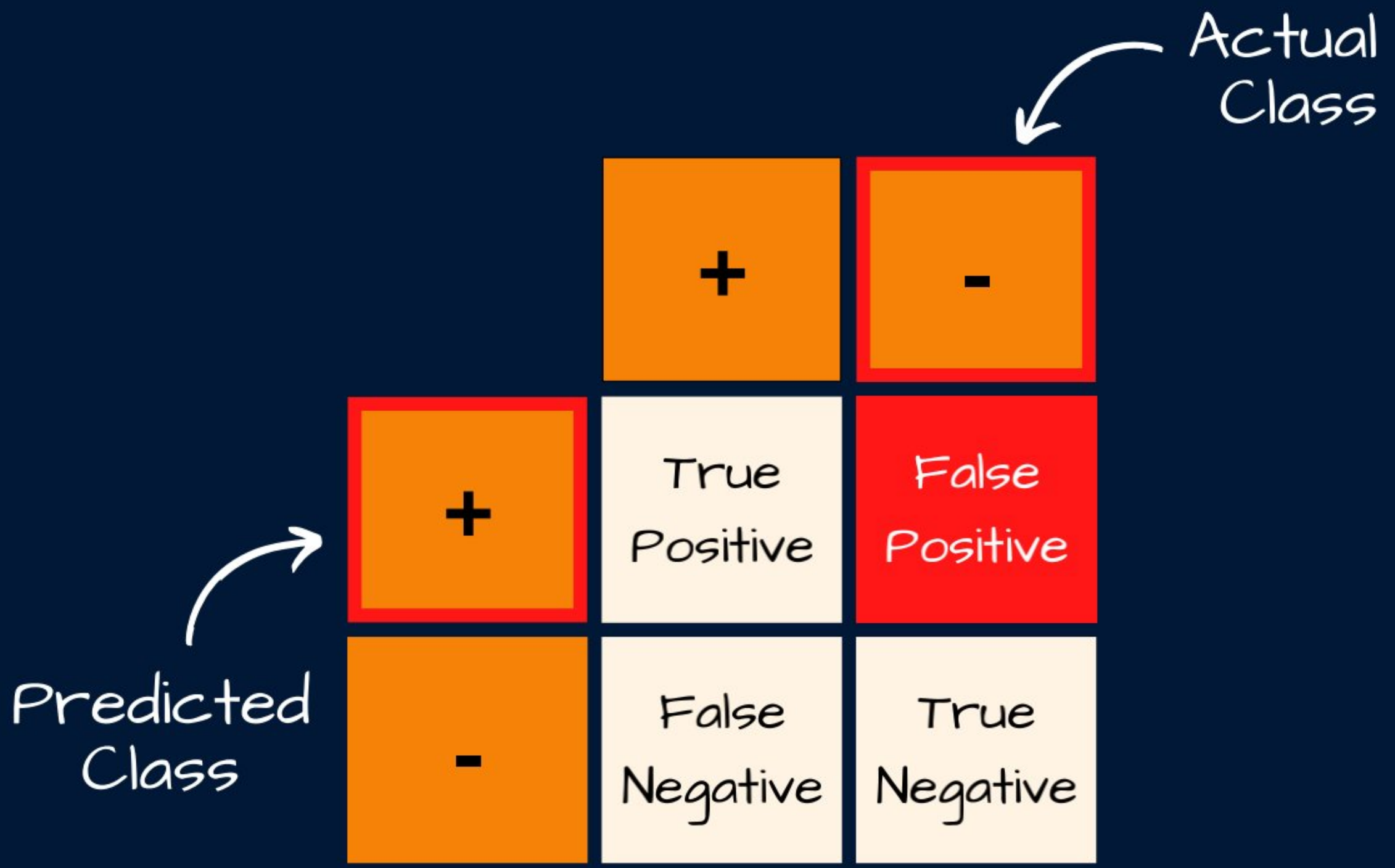
DATA SCIENCE INFINITY

ASSESSING CLASSIFICATION ACCURACY IN MACHINE LEARNING



AN OVERVIEW

DATA SCIENCE INFINITY

Classification Problems

In Machine Learning we will often take on tasks known as **Classification Problems...**

This is where we get an algorithm to learn the patterns in our data - with it's output being a **prediction of a class or label**, so one or zero, yes or no, cat or dog, and so on...

These Classification Problems fall into an area known as **Supervised Learning** where we train the model or algorithm to learn the relationships between inputs & outputs based upon **known** classes or labels.

This means that when training the model, for each data-point, we have both a predicted label *and* an actual label.

Comparing the predictions against the actuals gives us a view of **how good** our model is!

An example...

Let's say we were training a model to predict whether students would pass an exam or not. After training the model we have the following information...

How did our model do?

	Actual Result	Predicted Result	
	Pass	Pass	
	Pass	Fail	
	Fail	Pass	
	Fail	Fail	
	Pass	Pass	
	Fail	Fail	

As you can see, sometimes our model predicts the result correctly, and sometimes it gets it wrong. Let's summarise this information down to assess our overall model performance!

DATA SCIENCE INFINITY

	Actual Result	Predicted Result
	Pass	Pass
	Pass	Fail
	Fail	Pass
	Fail	Fail
	Pass	Pass
	Fail	Fail

We can summarise our results into a 2x2 table known as a **Confusion Matrix!**

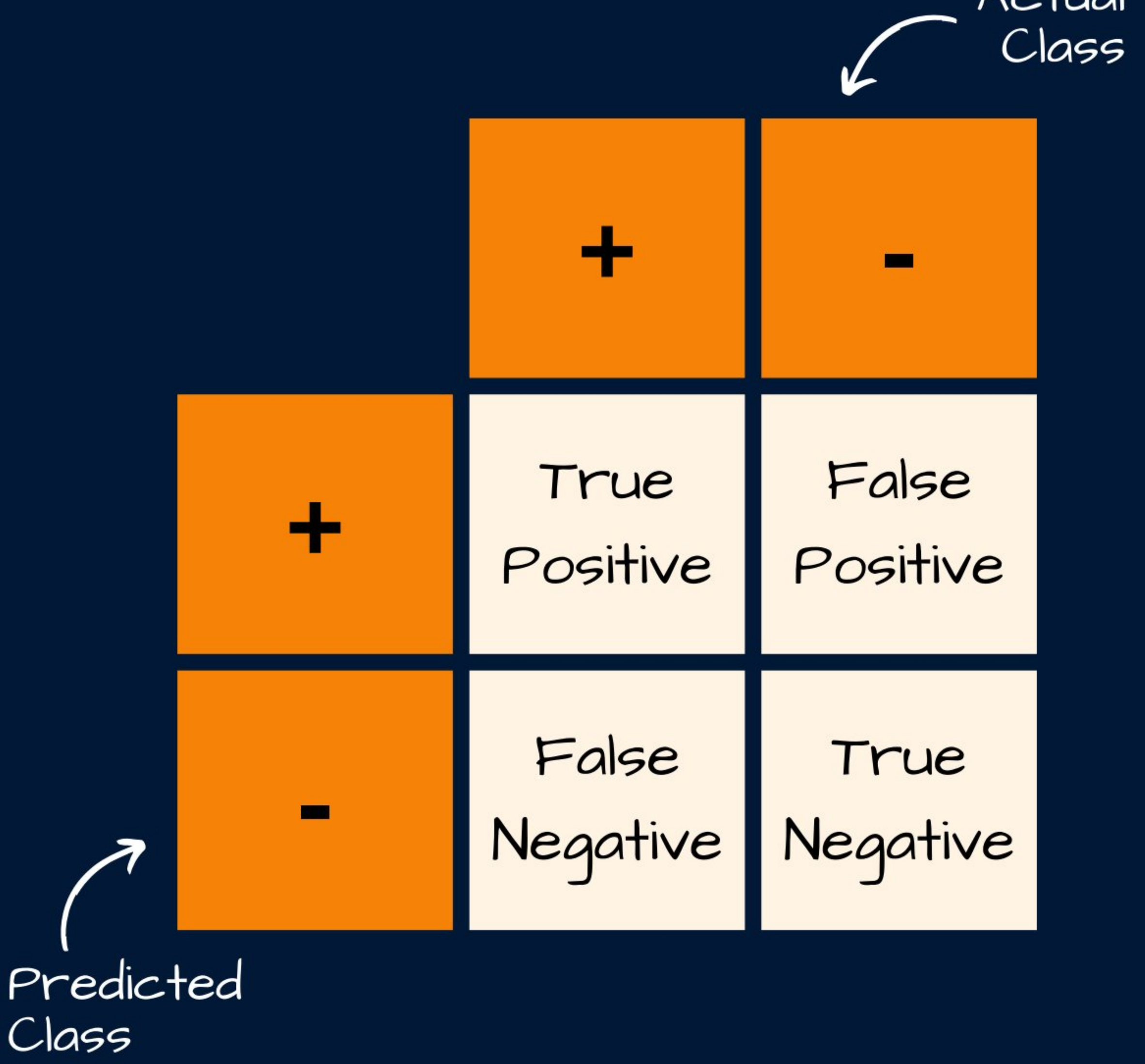
We simply count up the number of data-points by each class, and then split this by what their prediction was - partitioning everything into 4 segments...

Predicted Class	Actual Class: Pass (+)	Actual Class: Fail (-)
Pass (+)	2	1
Fail (-)	1	2

DATA SCIENCE INFINITY

Confusion Matrix...

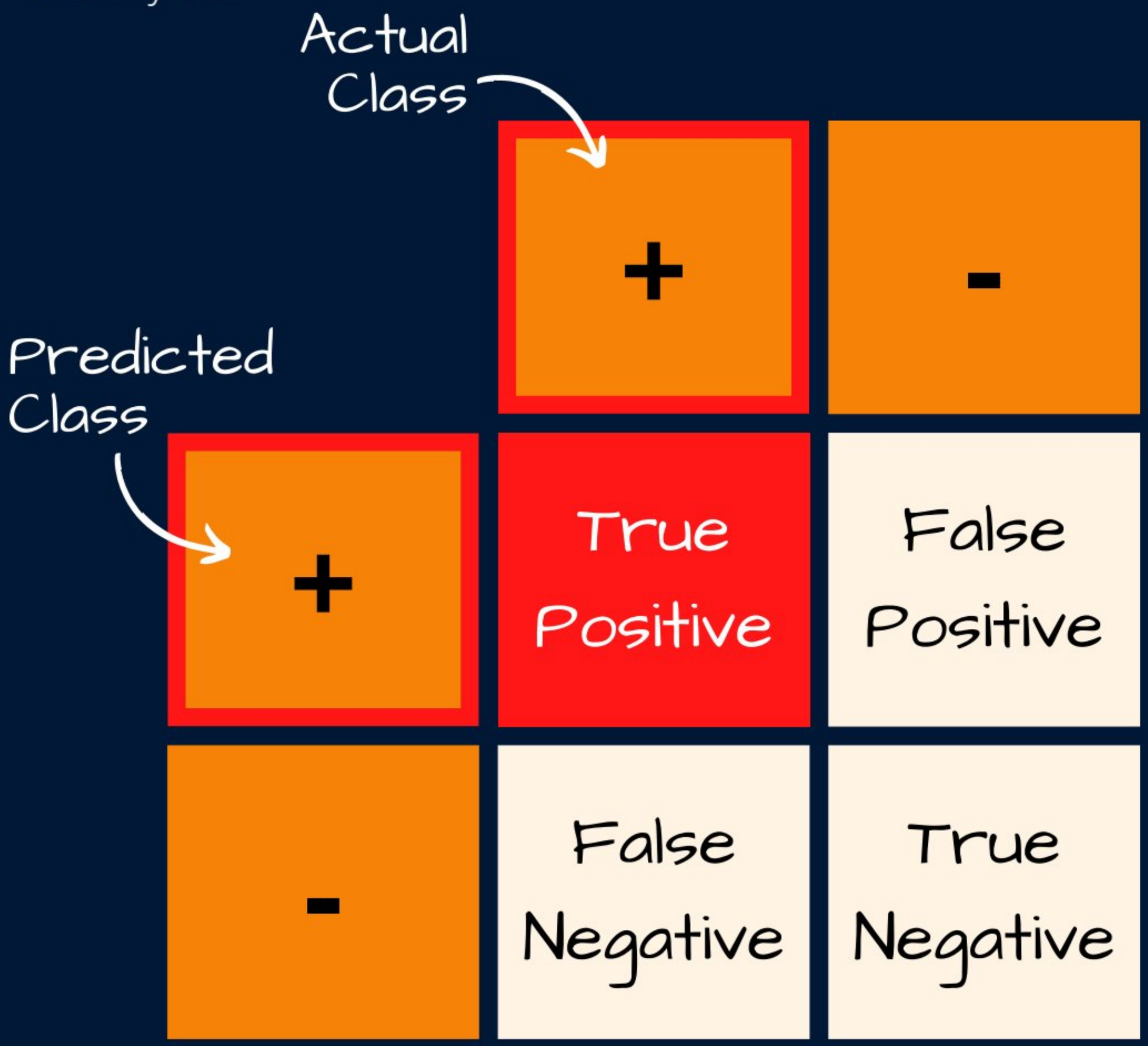
The 4 segments of the Confusion Matrix have names that you might have heard of...



True Positive

A **True Positive** is where we predict a positive result and are **correct** in our prediction

Example: Our model predicts a patient has a disease, and they actually do.

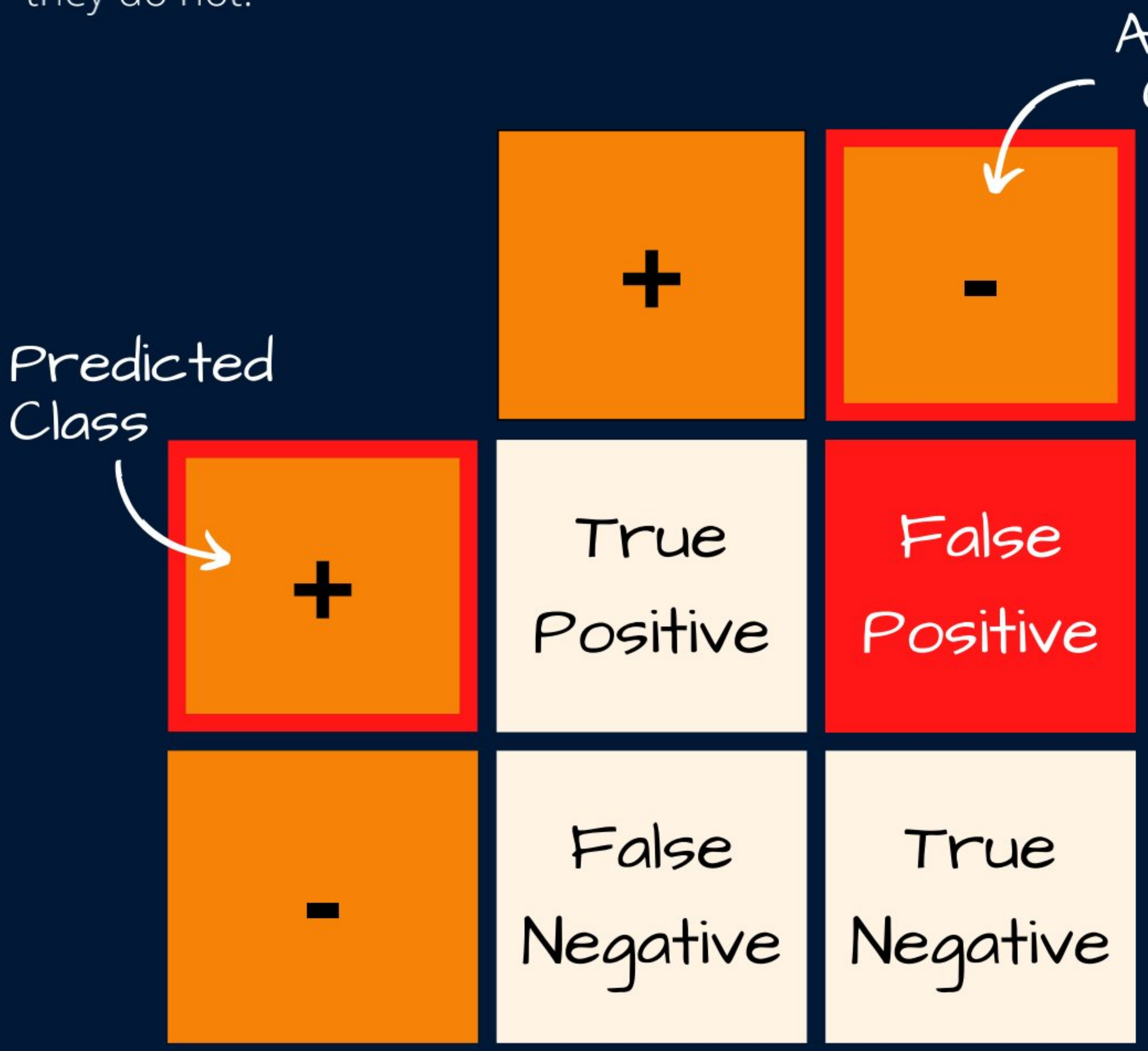


DATA SCIENCE INFINITY

False Positive

A **False Positive** is where we predict a positive result and are **incorrect** in our prediction. This is often called a **Type 1 Error**

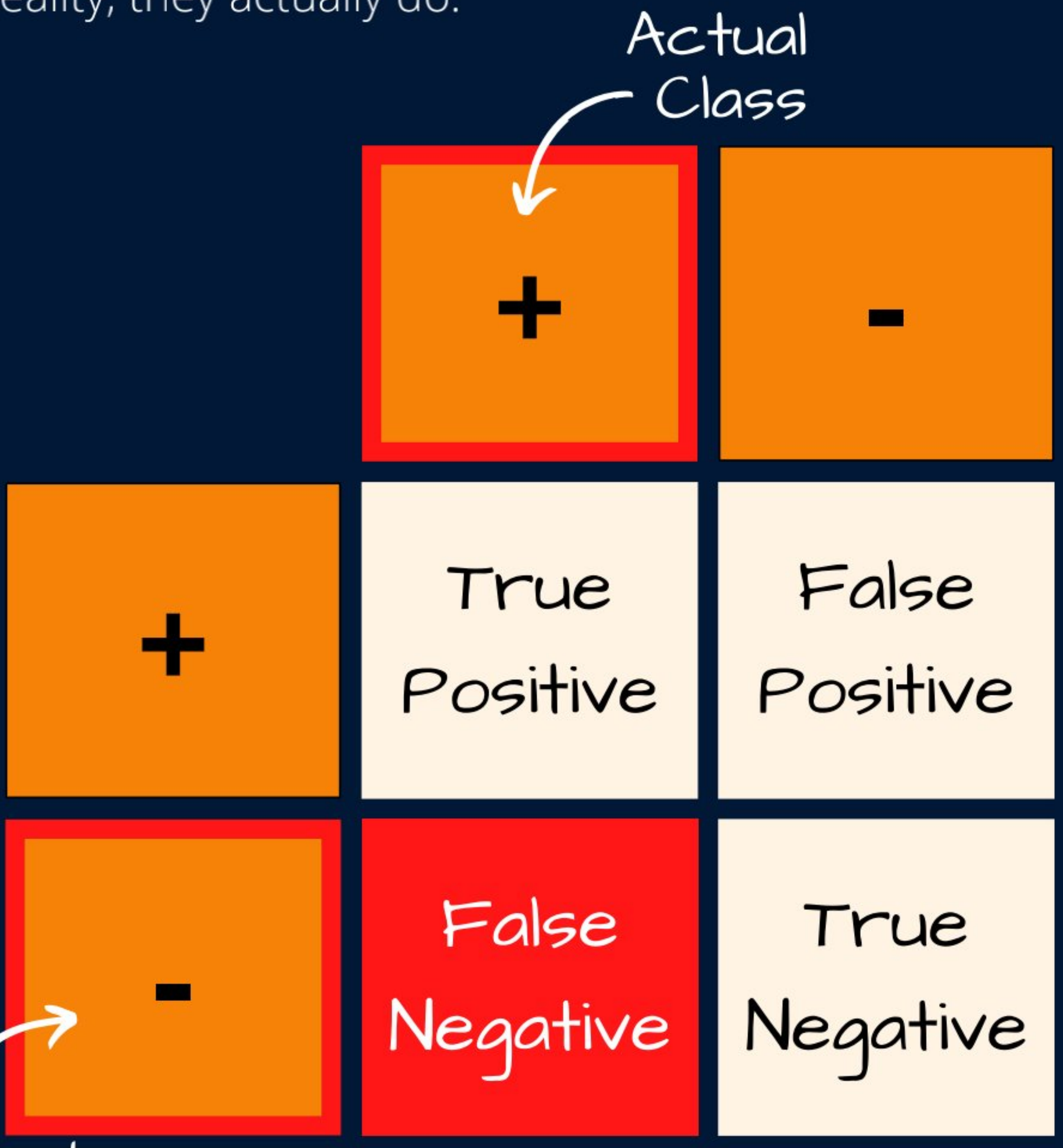
Example: Our model predicts a patient has a disease, but in reality, they do not.



False Negative

A **False Negative** is where we predict a negative result and are **incorrect** in our prediction. This is often called a **Type 2 Error**

Example: Our model predicts a patient does not have a disease, but in reality, they actually do.

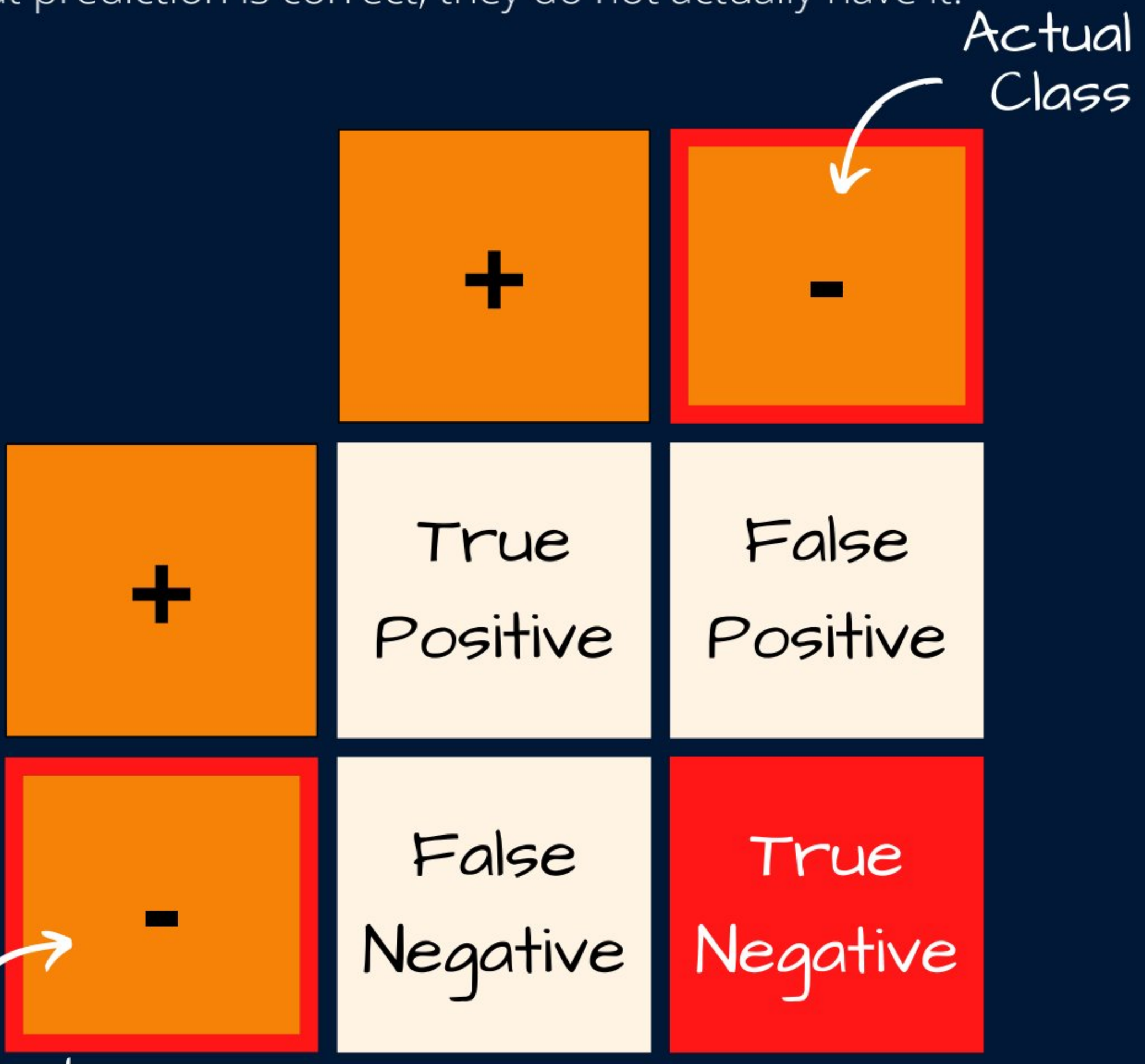


DATA SCIENCE INFINITY

True Negative

A **True Negative** is where we predict a negative result and are **correct** in our prediction.

Example: Our model predicts a patient does not have a disease, and that prediction is correct, they do not actually have it.



DATA SCIENCE INFINITY

Classification Accuracy

Classification Accuracy is very intuitive, it simply measures the number of **correct classification predictions** (across both classes) out of **all predictions** made!

Back to our example...

	Actual Result	Predicted Result
	Pass	Pass
	Pass	Fail
	Fail	Pass
	Fail	Fail
	Pass	Pass
	Fail	Fail

Predicted Result \ Actual Result	Pass (+)	Fail (-)
Pass (+)	2	1
Fail (-)	1	2

Our model correctly predicted the results of 4 students, from the total of 6 - giving a **Classification Accuracy** of...

Correct	4
Incorrect	2
Total	6

4 / 6 = **66.6%**

But be careful...

While **Classification Accuracy** is very intuitive, there are scenarios where it can be misleading. It is important to know when this can happen, and what metrics are more appropriate!

Let us instead consider a scenario where we're building a Classification Model to predict a binary outcome of whether each of 1,000 patients have a **rare** medical disease, or not.

We train the model, and then assess the results using **Classification Accuracy** on our test set of 100 patients. We find that our model has **correctly classified 97 out of the 100** patients, giving a **Classification Accuracy of 97%**.

Classification Accuracy

97 / 100 = 97%

Correct	97
Incorrect	3
Total	100

Wow - we are correctly predicting for 97% of patients!

Before we jump for joy - let's take a look at the Confusion Matrix that gave us this result...

While the 98% Classification Accuracy is correct - we notice that this disease we are predicting for, is indeed **very rare**!

We can see from the Confusion Matrix that **only 2 patients out of our set of 100 actually do have the disease**!

Predicted Result \ Actual Result	Disease	No Disease
Disease	1	2
No Disease	1	96

Correct	97
Incorrect	3
Total	100

Classification Accuracy = 97%

This is what we call **Imbalanced Data**!

Think about this - if we didn't use any fancy classification model, and we simply predicted that **everyone was disease free** - we'd actually get a higher accuracy rate 98%! Unfortunately, now our model doesn't look quite so amazing!

So what can we do? Well, there are a couple of supplementary metrics we can use over and above Classification Accuracy...

Precision

Definition: Of all observations that were predicted, or classified as positive - what proportion actually were positive

In our example: Of all patients we predicted to have the disease - what proportion actually did

We can easily calculate this from our Confusion Matrix by taking the number of True Positives, and dividing that by the sum of True Positives & False Positives

$$\frac{TP}{(TP + FP)}$$

Predicted Result \ Actual Result	Disease	No Disease
Disease	1 TRUE POSITIVE	2 FALSE POSITIVE
No Disease	1 FALSE NEGATIVE	96 TRUE NEGATIVE

$$\frac{1}{(1+2)}$$

$$\frac{1}{3}$$

33%

Of all patients we *predicted* to have the disease, only 33% actually did! We would have patients thinking they were ill, when they were not!

This gives us a whole new way to think about **how good** our model really is!

Recall

Definition: Of all positive observations, how many did we predict as positive

In our example: Of all patients who actually had the disease, how many did we correctly predict

We can easily calculate this from our Confusion Matrix by taking the number of True Positives, and dividing that by the sum of True Positives & False Negatives

$$\frac{TP}{(TP + FN)}$$

Predicted Result \ Actual Result	Disease	No Disease
Disease	1 TRUE POSITIVE	2 FALSE POSITIVE
No Disease	1 FALSE NEGATIVE	96 TRUE NEGATIVE

$$\frac{1}{(1+1)} \rightarrow \frac{1}{2} \rightarrow 50\%$$

Of all patients who *actually* had the disease, our model only predicted 50% of them!

Again, for a disease prediction model, this potentially isn't great...

DATA SCIENCE INFINITY

Precision vs. Recall

While we've seen that **Classification Accuracy** can be a good indicator of model performance - we should always dig deeper and analyse both **Precision** & **Recall** before getting too excited about our accuracy measure, especially in cases where the data is biased heavily towards one of the classes....

The tricky thing with these two, Precision & Recall, is that it's hard to optimise both - it's essentially a zero-sum game.

If you try to improve Precision, the Recall falls and vice-versa.

In some scenarios, it will make more sense to try and elevate one of them, in spite of the other.

In the case of the rare disease classification model like in our example, perhaps it's more important to improve Recall as we want to classify as many cases as possible, but then again, we don't want to just classify everyone as having the disease - as that isn't a great solution either...

So - there is one more metric we can use...

DATA SCIENCE INFINITY

F1-Score

The **F1-Score** is the harmonic mean of **Precision** & **Recall**.

A good F1-Score comes when there is a **balance** between Precision & Recall, **rather than a disparity** between them.

$$\textbf{F1-Score} = \frac{2 * (\textbf{Recall} * \textbf{Precision})}{(\textbf{Recall} + \textbf{Precision})}$$

In our example we had...

Precision = 0.33
Recall = 0.50

$$\rightarrow \frac{2 * (0.33 * 0.5)}{(0.33 + 0.5)}$$

$$\rightarrow \frac{0.33}{0.833}$$

$$\rightarrow \textbf{0.396}$$

We now have several metrics that give us a much better understanding of our predictions, and of our model!

DATA SCIENCE INFINITY

Bonus! Precision & Recall Cheat-Sheet!

It can be hard to remember what Precision & Recall each mean, and which way around the go.

Below is a cheat-sheet with a high level overview of what values of each would mean!

Precision	Recall	Meaning
High	High	The model is differentiating between classes well
High	Low	The model is struggling to detect the class, but when it does it is very trustworthy
Low	High	The model is identifying most of the class, but is also incorrectly including a high number of data points from another class
Low	Low	The model is struggling to differentiate between classes