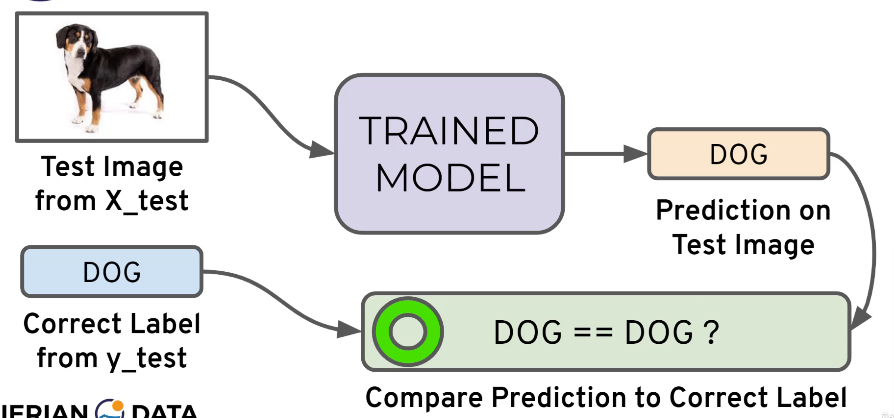
# Classification Metrics

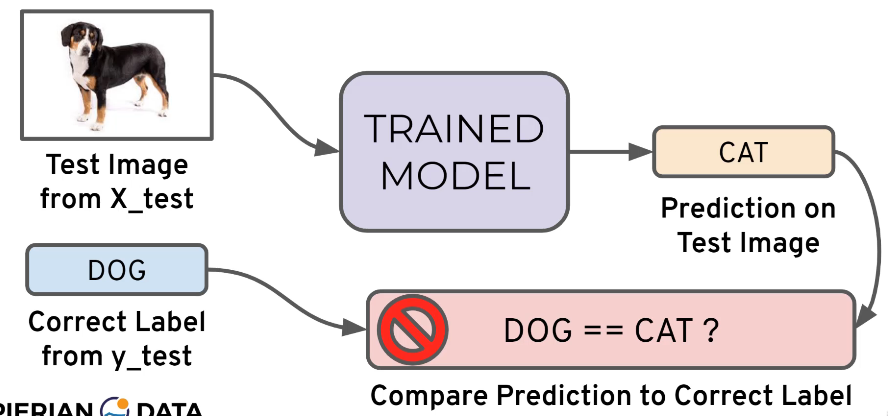
The key classification metrics are:

1. Accuracy
2. Recall
3. Precision
4. F1-Score

## Reason For classification metrics

* Typically, in any classification task the model can only achieve two results:  
  1. Either the model was correct in its prediction  
  2. Or the model was incorrect in its prediction
* Fortunately, incorrect vs correct expands to situations where there are multiple classes
* For the purposes of explaining the metrics, lets imagine a binary classification situation, where there are only two methods available
* In this example, we will attempt to predict if an image is a dog or a cat.
* Since this is a supervised learning, we will first **fit/train** a model on training data, then **test** the model on **testing data.**
* Once we have the model’s predictions from the **X\_test** data, we compare it to the **true y values** (the correct labels).





* In real world, a single metric won’t tell the complete story
* To understand all of this, lets bring back the 4 metrics mentioned and see how they are calculated.
* We could organize our predicted values compared to the real values in a **confusion matrix.**

### Accuracy

* Accuracy in classification problems is the **number of correct predictions** made by the model divided by the **total number of predictions**.
* For example, if the X\_set was 100 images and our model **correctly** predicted 80 images, then we have 80/100 i.e., **0.8 or 80% accuracy**.
* Accuracy is useful when the target classes are well balanced.
* In our example, we would have roughly the same amount of cat images as we have dog images.
* Accuracy is not a good choice with unbalanced classes
* Imagine we had 99 images of dogs and only 1 image of a cat.
* If our model was simply a line that always predicted dog, we would get 99% accuracy. However, since our model is made to predict both dogs and cats, the accuracy in real world would be different.

In this situation we will want to know more about recall and precision.

### Recall

* Ability of a model to find all the relevant cases within a dataset.
* The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives.

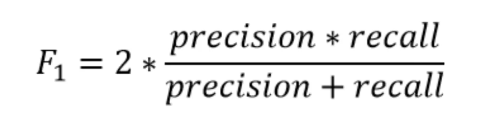
### Precision

* Ability of a classification model to identify only the relevant data points.
* Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

#### Recall and Precision

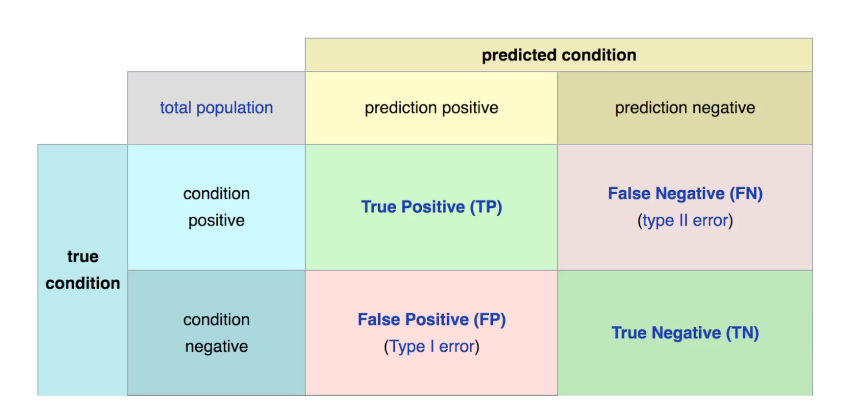
* Often there is a trade-off between Recall and Precision
* While recall expresses the ability to find all the relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant and actually were relevant.

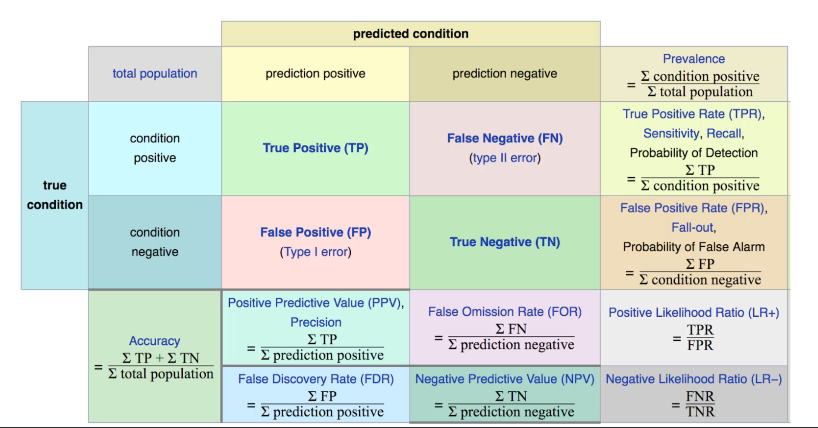
### F1 Score

* In cases where we want to find an optimal blend of precision and recall we can combine the two metrics using what is called the F1 Score.
* The F1 score is the harmonic mean of precision and recall taking both metrics into account in the following equation:  
  
* The reason we use the harmonic mean instead of a simple average is because it punishes the extreme values.
* A classifier with a precision of 1.0 and a recall of 0.0 has a simple average of 0.5 but an F1 score of 0.

## Confusion Matrix

We can also view all correctly classified versus incorrectly classified images in the form of a confusion matrix.





* The main point to remember with the confusion matrix and the various calculated metrics is that they are all fundamentally ways of comparing the predicted values versus the true values.
* What constitutes a “good” metrics will really depend on the specific situation.
* The good enough accuracy depends on the context of the situation.
* Often models are used as quick diagnostic tests to have before having a more invasive test
* Often, we have a precision/recall trade off. We need to decide if the model should focus on fixing the False Positives vs the False Negatives.
* In disease diagnosis, it is probably better to go in the direction of False Positive, so we make sure we correctly classify as many cases of disease as possible.
* All of this is to say, machine learning is not performed in a "vacuum", but instead a collaborative process where we should consult with experts in the domain (e.g., medical doctors)

## Evaluating Regression

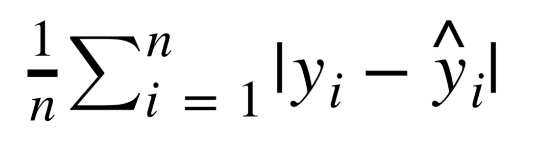
* Regression is a task when a model attempts to predict continuous values (unlike categorical values, which is classification).
* The previous metrics are not much useful for regression problems. We need metrics designed for continuous values.
* For example, attempting to predict the price of a house given its features is a regression task.
* Attempting to predict the country a house is in given its features would be a classification task instead.

Most common evaluation metrics for regression:

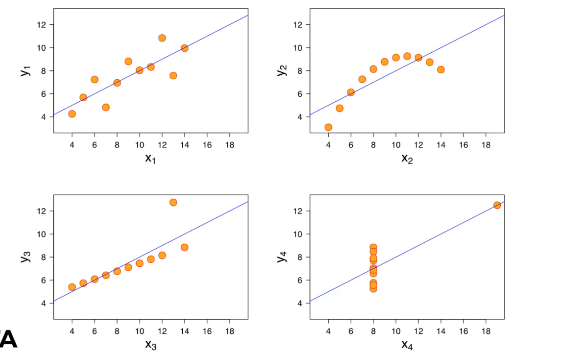
* Mean Absolute Error
* Mean Squared Error
* Root Mean Square Error

### Mean Absolute Error (MAE)

* This is the mean of the absolute value of errors.
* It is easy to understand.

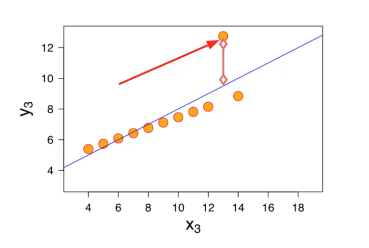


#### MAE won't punish large errors

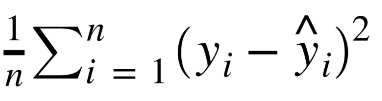


Although the data are varied, the line is the same.

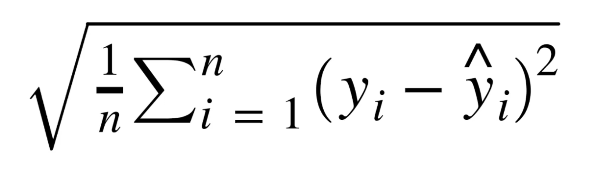
* We want our error metrics to account for cases like the following:



### Mean Square Error (MSE)

* This is the mean of the squared errors
* Larger errors are noted more than that with MAE, making MSE more popular.  
    
  
* One issue that should be considered is that the units also get squared.

### Root Mean Square Error (RMSE)

* This is the root of the mean of the squared errors.
* This is the most popular method utilizing the advantage of MSE while also keeping units in check.  
    
  
* Once again, the value of RMSE being good or not is context specific.
* A RMSE of $10 is a fantastic for predicting the price of a house, but horrible for predicting the price of a candy bar.

So, compare the error metric to the average value of the label in the data set to try to get an intuition of its overall performance. Domain knowledge also plays an important role here.