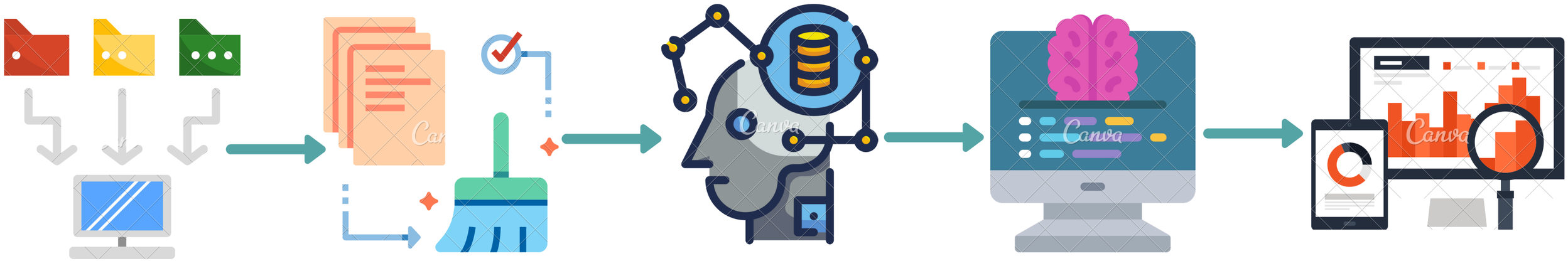
Pipeline

DATA CLEANING

TRAIN MODEL



DATA COLLECTION

FEATURE EXTRACTION

VISUALIZE

Performance Measures

The accuracy of a classification method is the ability of the method to correctly determine the class of a randomly selected data instance.

The most obvious criterion to use for estimating the performance of a classifier is predictive accuracy.

Performance Measures

A more difficult trade-off occurs when the classes are severely unbalanced. Suppose we are considering investing in one of the leading companies quoted on a certain stock market.

Can we predict which companies will become bankrupt by the next two years (so we can avoid investing in them)?

- **The proportion of such companies is obviously small, lets say 0.02, so on average out of every 100 companies 2 will become bankrupt.**
- **Call these “bad” and “good” companies.**
- **If we have a very trusting classifier that always predicts “good” under all circumstances its predictive accuracy will be 98 %, a very high value.**
- **Looked at only in terms of predictive accuracy this is a very successful classifier.**

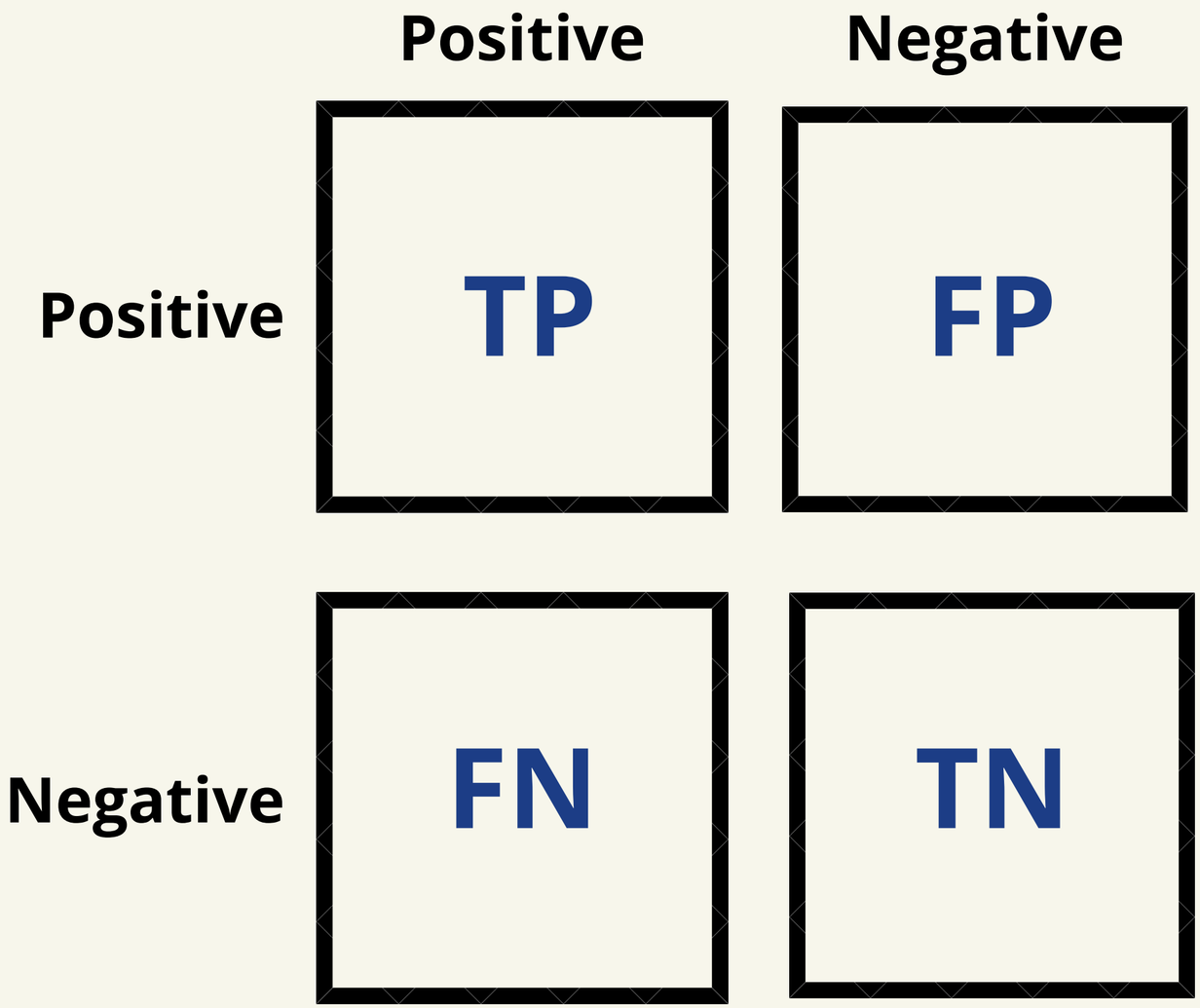
A “confusion matrix” is sometimes used to represent the result of testing in more detail.

The advantage of using this matrix is that it not only tells us how many got misclassified but also what misclassifications occurred.

When there are two classes, positive (+) and negative (-), the confusion matrix consists of four cells, i.e., TP, FP, FN and TN.

Confusion matrix

	Positive	Negative
Positive		
Negative		


Positive
Negative
Positive
TP
FP
Negative
FN
TN

Scoring

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The Perfect Classifier

- A: The Perfect Classifier
  - Here every instance is correctly classified. TP=P, TN=N and following is its Confusion Matrix

		Predicted Class	
		+	-
Actual Class	+	**P**	**0**
	-	**0**	**N**

The Worst Possible Classifier

- B: The Worst Possible Classifier
  - Here every instance is wrongly classified. TP=0, TN=0 and following is its Confusion Matrix

		Predicted Class	
		+	-
Actual Class	+	**0**	**P**
	-	**N**	**0**

The Ultra-Liberal Classifier

- C: The Ultra-Liberal Classifier
  - This Classifier always predicts the positive class. The TP rate = 1, but so is the FP rate.

		Predicted Class	
		+	-
Actual Class	+	**P**	**0**
	-	**N**	**0**

The Ultra-Conservative Classifier

- D: The Ultra-Conservative Classifier
  - This Classifier always predicts the negative class. The FP rate = 0, but so is the TP rate.

		Predicted Class	
		+	-
Actual Class	+	**0**	**P**
	-	**0**	**N**