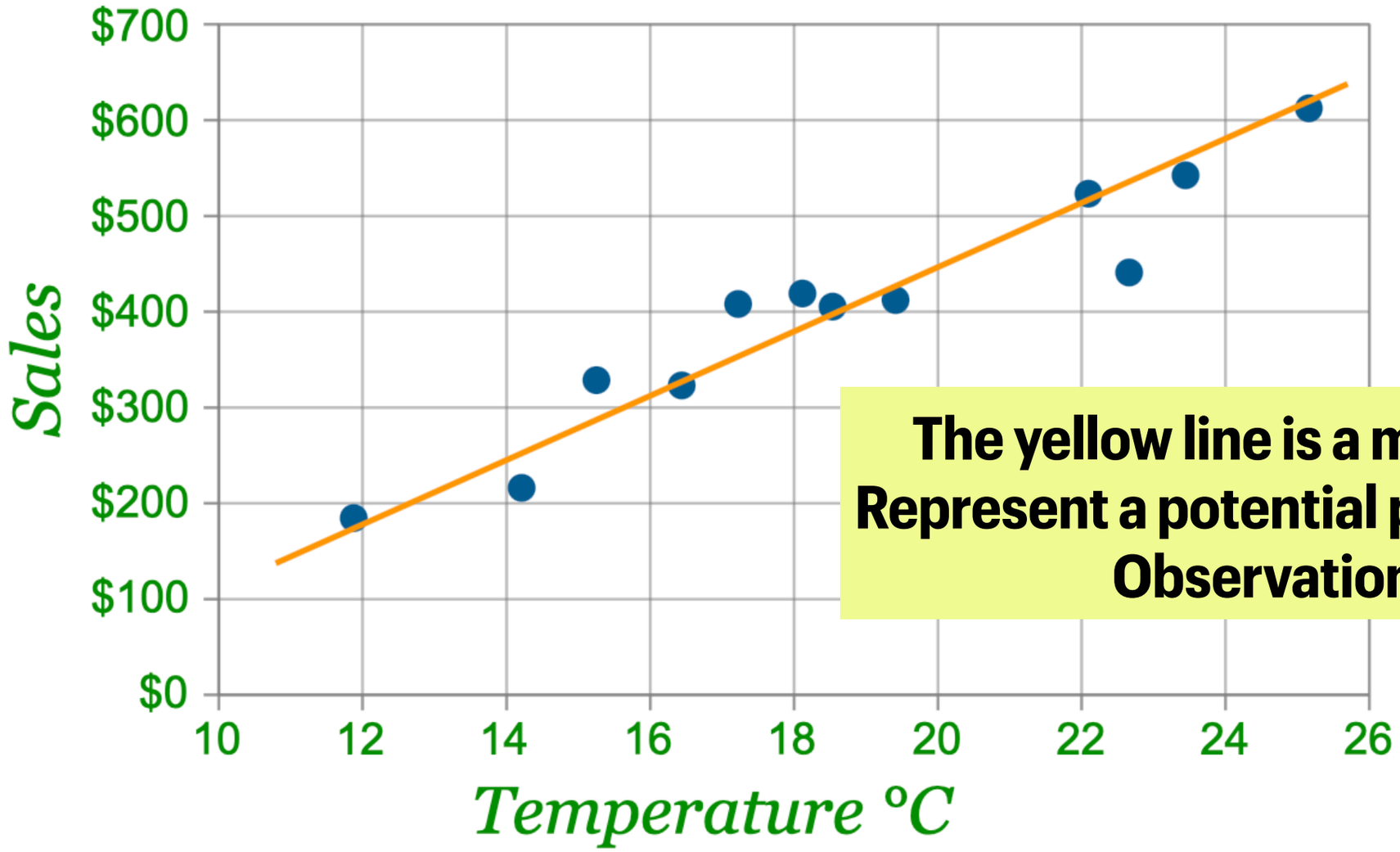
HUMAN LEARNING ABOUT MACHINE LEARNING

DR PAT RYSER-WELCH

SALES OF ICE-CREAMS IN THE SUMMER

If I have captured the sales of screams and the temperature in degree C. Could I possibly predict the sales of ice-cream?

The blue dots match some observed sales against the sales

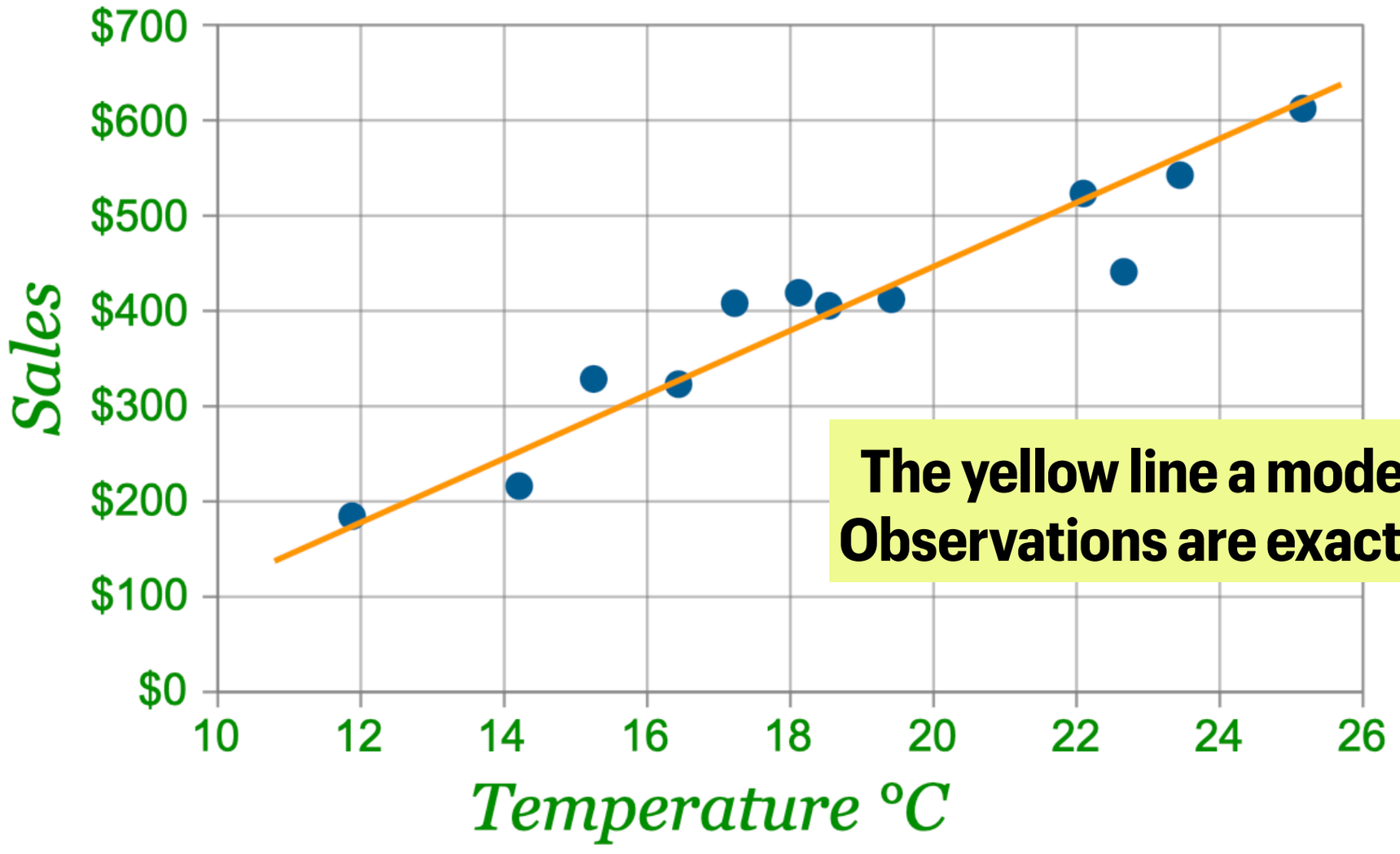


The yellow line is a model , that Represent a potential pattern in the Observations.

This is a fictitious data

https://www.mathsisfun.com/data/least-squares-regression.html

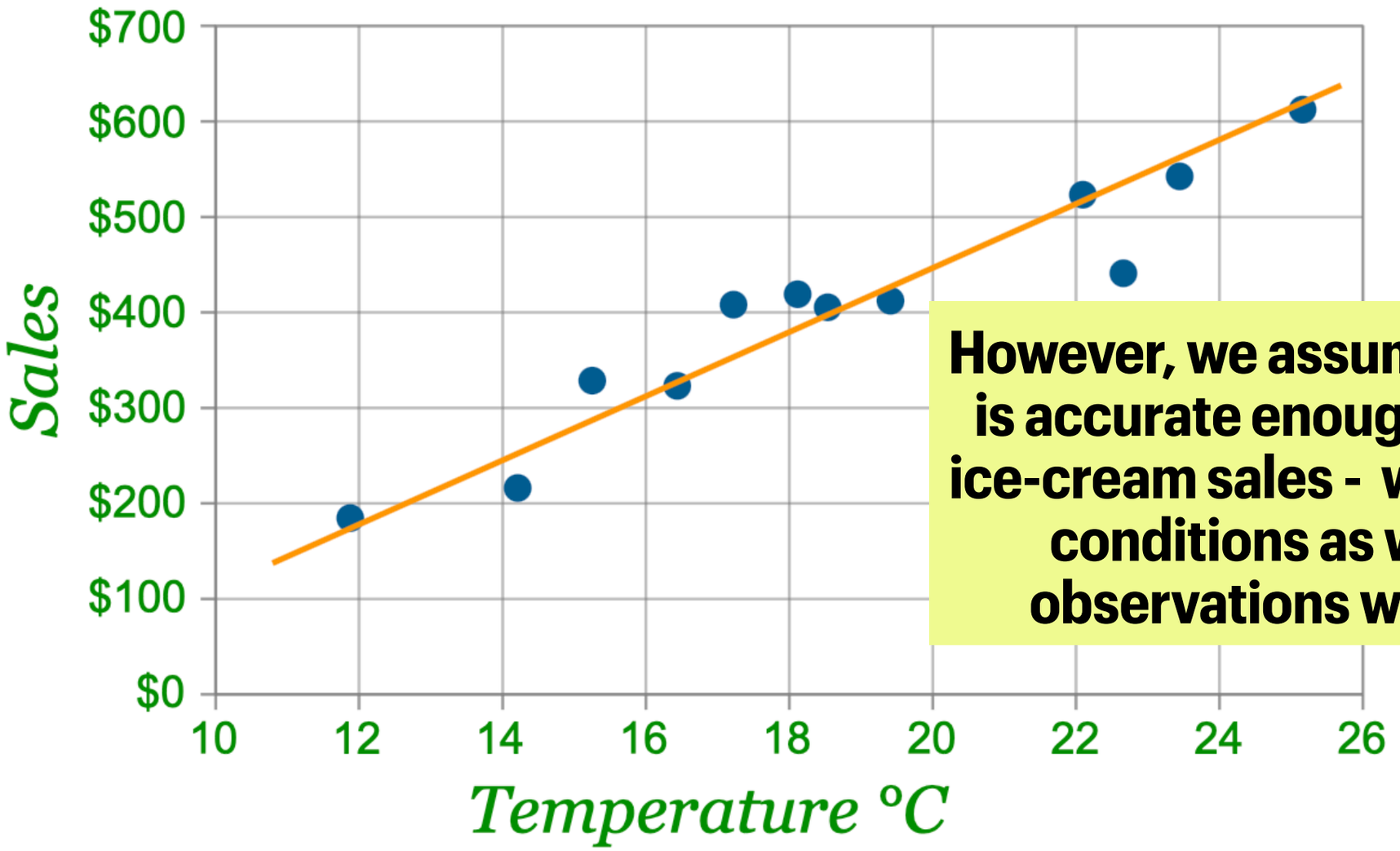
The blue dots match some observed sales against the sales



The yellow line a model . Not all the Observations are exactly on the line.

This is a fictitious data

The blue dots match some observed sales against the sales



However, we assume the model is accurate enough to predict ice-cream sales - within similar conditions as when the observations were made

This is a fictitious data

DATA = MODEL + ERROR

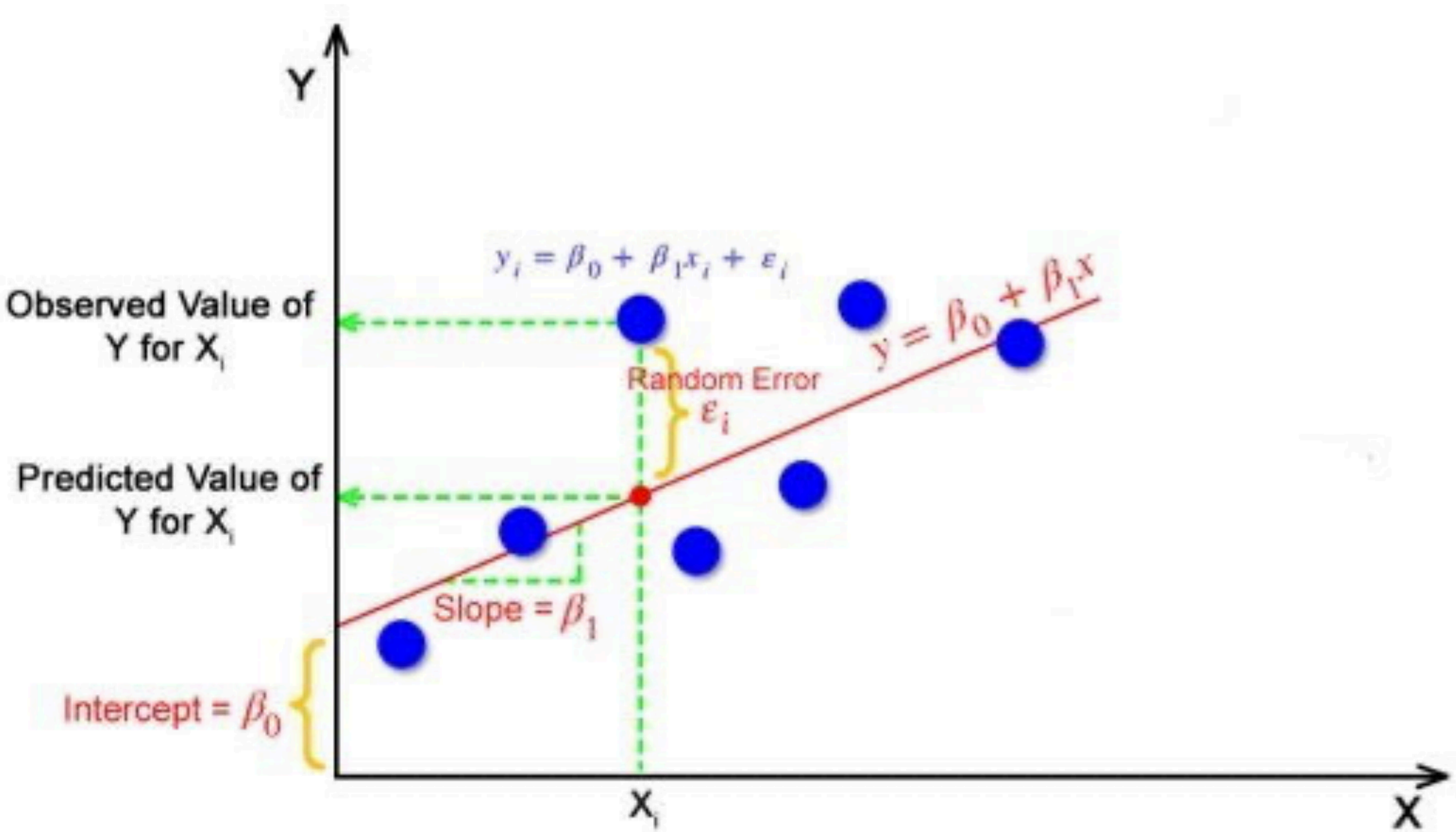
Models

Models capture the nature of some data as simply as possible. The basic structure of a statistical model considers data as the sum of a model and some errors.

$$data = model + error$$

- The model expresses the values we expect the data to be take given our knowledge.
- The error reflects the differences between the model's prediction and the observed data.

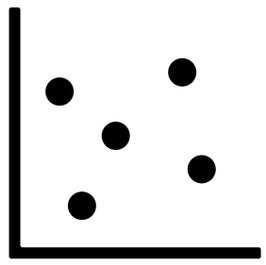
AN ILLUSTRATION

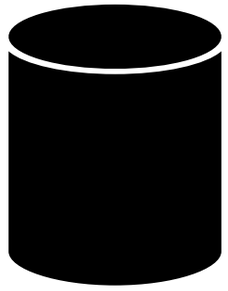


https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/

HOW CAN WE LEARN A MODEL?

DATA OR SAMPLE





LEARNING ALGORITHM

Neural AI

Regression

Supervised learning

Model fitting

Non-deterministic Methodologies

Gradient-descent

MODEL AND ERROR

Neural network

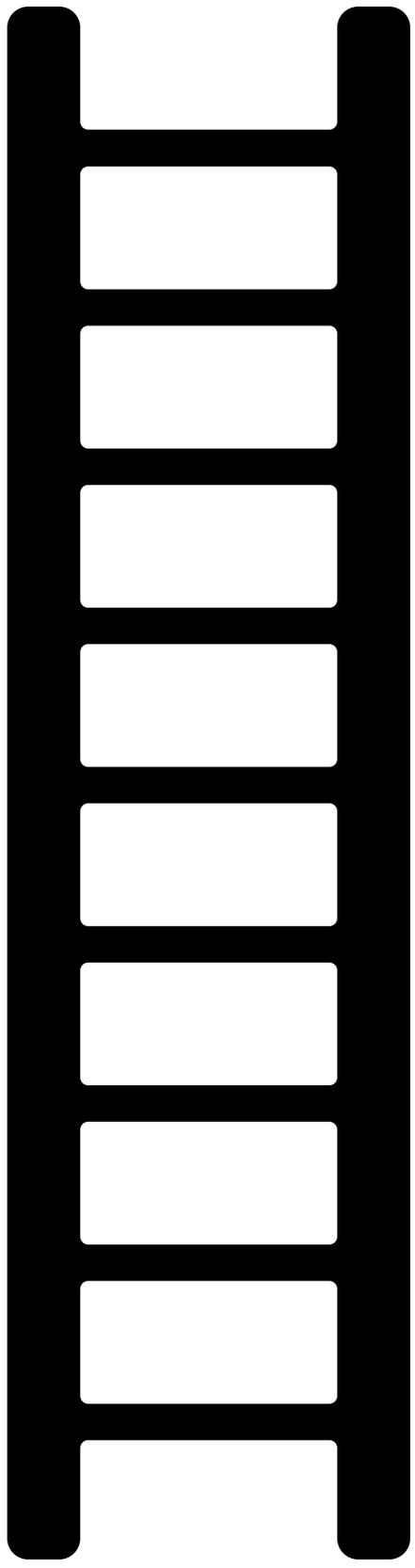
Decision Tree

Regression line

General linear model

Probabilistic model

Statistical model



STEP 4 - LEARNING OR MODEL FITTING

STEP 3 - PREPARE DATA FOR LEARNING

STEP 2- CLEAN AND PREPARE DATA

STEP 1 - EXPLORE THE DATA

HOW CAN MEASURE THE QUALITY OF A MODEL

1. Apply the model on a dataset

2 . Compare the predicted values against known expected values.

3 . Compute some metrics

EXPECTED VALUE

PREDICTED VALUE

For binary classification (True/False) such as logistic regression four possible events can occur.

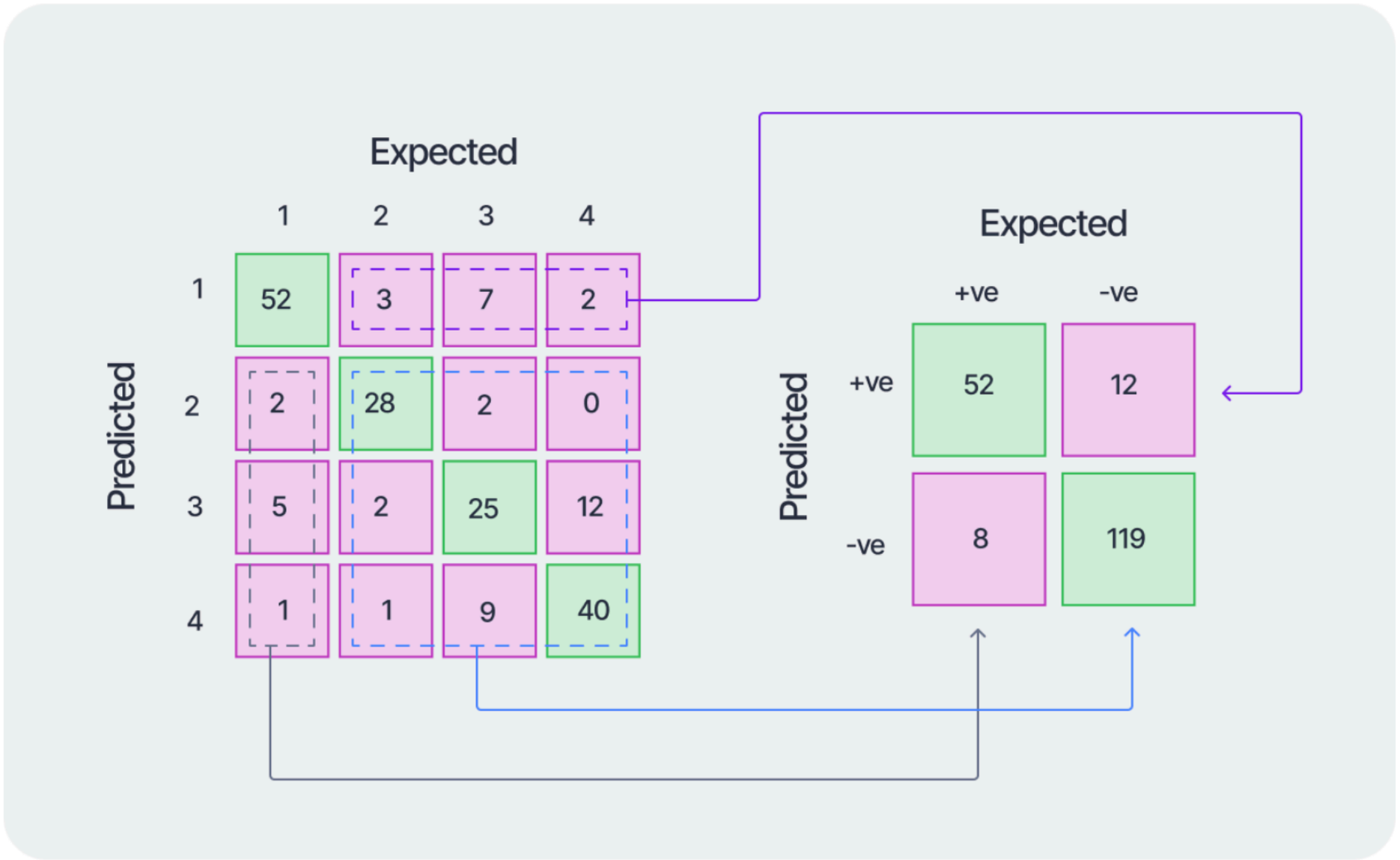
- **True positives (TP):** The number of correct predictions for the *true* class; i.e., the number of predicted True class that are known to be true.
- **True negatives (TN):** The number of correct predictions for the *false* class; i.e., the number of predicted False class that are known to be False.
- **False positives (FP):** The number of erroneous predictions for the True class; i.e., the number of predicted True class that are known to be False.
- **True negatives (TN):** The number of erroneous predictions for the False class; i.e., the number of predicted False class that are known to be True.

Confusion matrix

A confusion matrix counts the number True positives, False positives, True negatives, and False negatives. It represents in a table the actual values against the predicted values of a testing dataset. The correct predictions are shown in green with white font. The errorneous predictions in black and orange background.

		Predicted	
		False	True
Actual	False	False Positives	False negatives
	True	True negatives	True Positives

CONFUSION MATRIX



https://www.v7labs.com/blog/confusion-matrix-guide

Measures of quality

The measures of quality for a predictive model are expressed using two performance metric referred as *precision, recall, and accuracy.* These metrics are probabilities computed using the following mathematical formulae.

Accuracy describes how the model perform across *all* classes (True and False). It adds the diagonal values of a confusion matrix and divide by the total of possible outcomes.

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

Precision measures the model's accuracy in prediction the True class as a sample. It reflects how reliable the model is in classifying samples as positive. It is a probability based on the class columns the confusion matrix (green cells). The precision can be computed for each class.

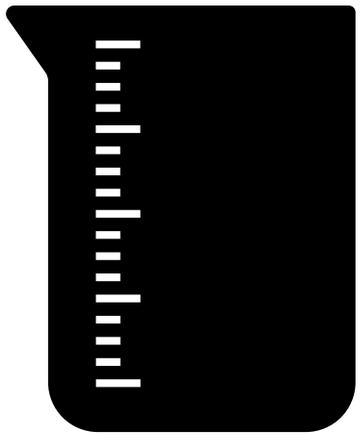
$$Precision = \frac{TP}{TP + FP}$$

Recall is the probability to detect positive classes. It can be calculated for each class. It is a probability is obtained by dividing the true positive by the sum of the true positive and false negative; i.e., the class row of the confusion matrix.

$$Recall = \frac{TP}{TP + FN}$$

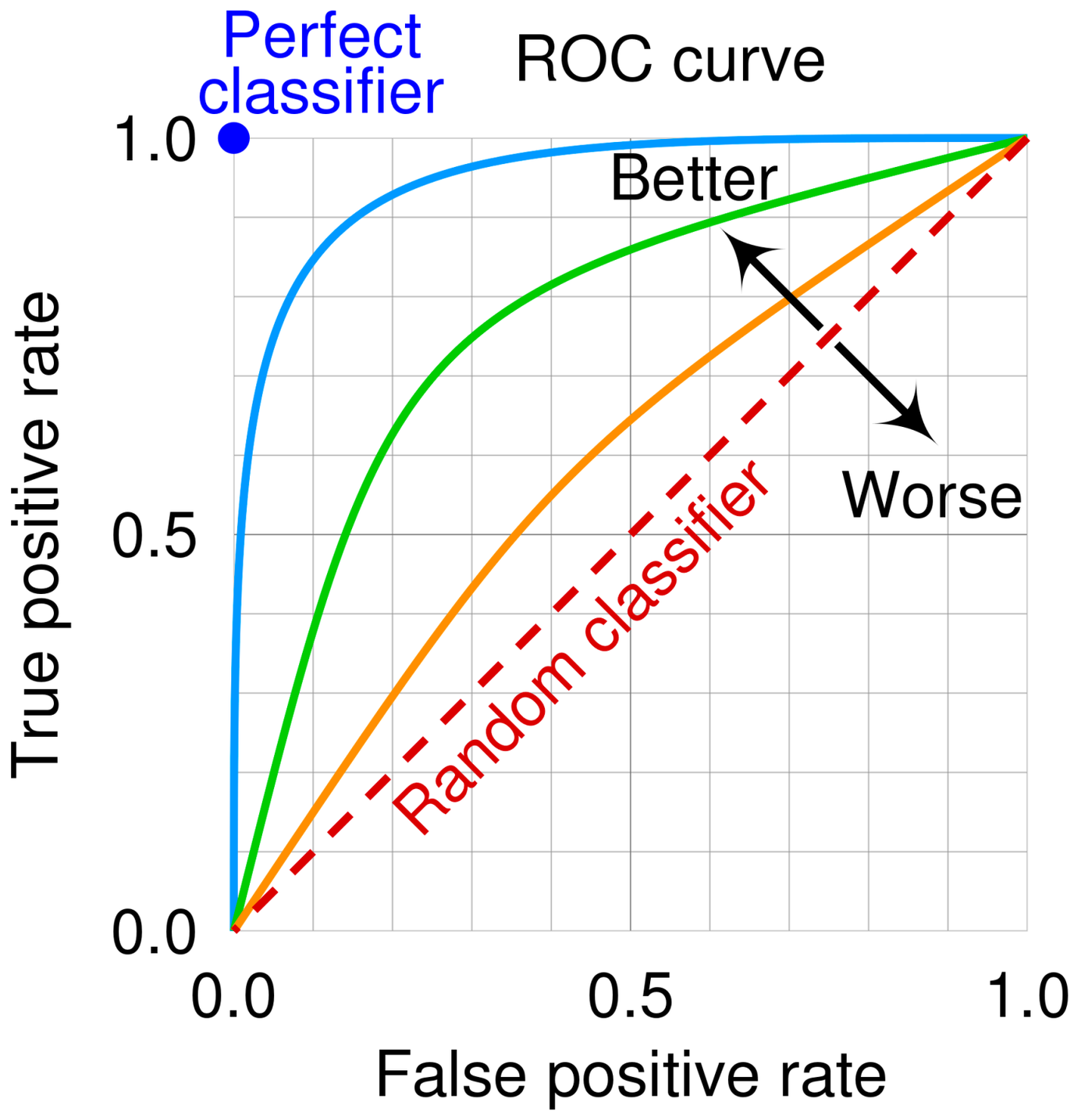
THREE DATASETS:

TRAINING



VALIDATION

TEST


Perfect classifier
ROC curve
1.0
Better
Worse
0.5
Random classifier
True positive rate
0.0
0.0
0.5
1.0
False positive rate

DATA AND LARGE AMOUNT OF IT SUPPORTS MACHINE LEARNING

TWO EXAMPLES

Predicting inspection outcome - Chicago food inspection dataset

The data has some repeated patterns

The data appears to have some dependent statistical variables.

Many of the data may have some clusters or strong relationship between them.

We achieve 100 percent accuracy.

Find the notebook

Predicting surviving the Titanic disaster

The data no clear repeated patterns.

The data appears to have some observations with a lot of complexity.

The data is quite small.

It is hard to achieve more than 80% in models, with machine learning.

Find the notebook

https://www.kaggle.com/code/patriciaryserwelch/data-preparation-for-ml-techniques