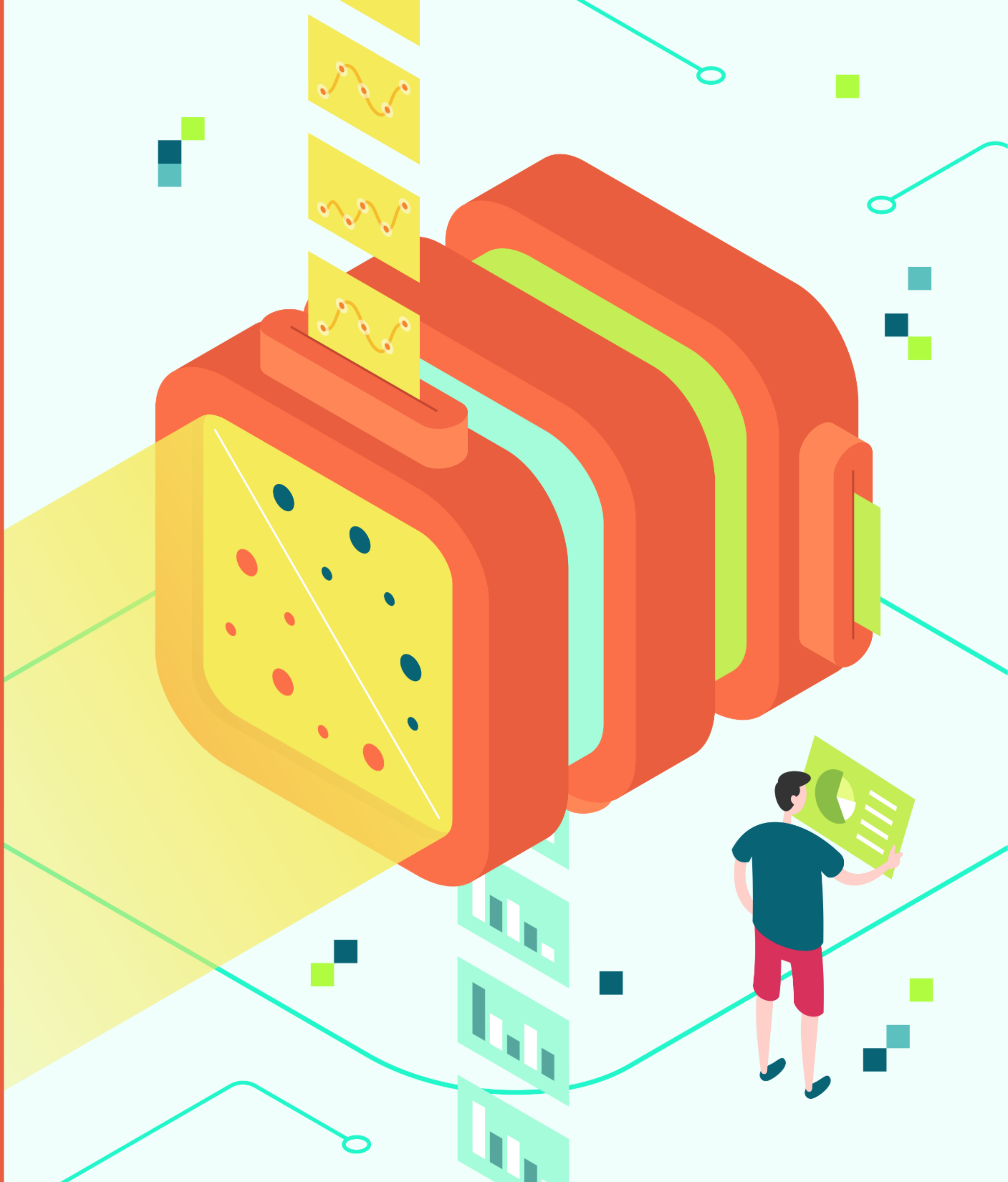
MACHINE LEARNING

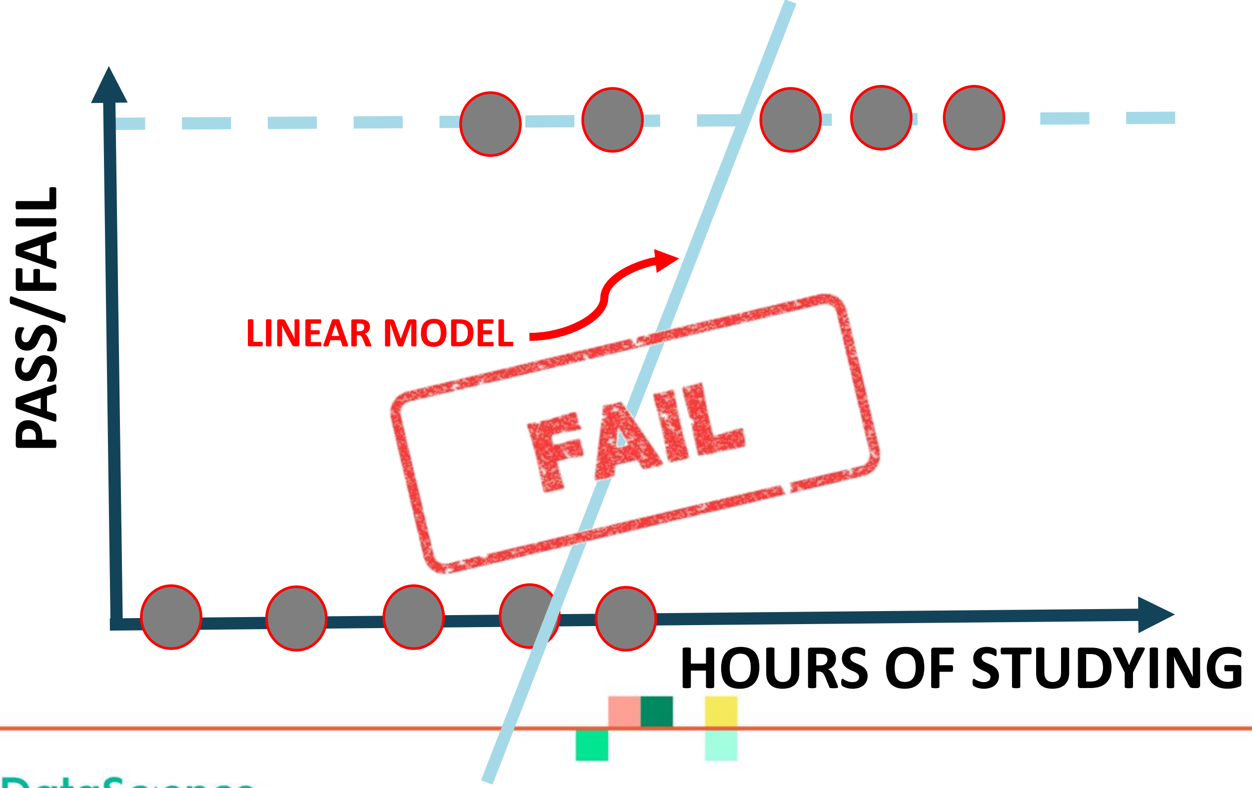
REGRESSION

LOGISTIC REGRESSION



LOGISTIC REGRESSION: INTUITION

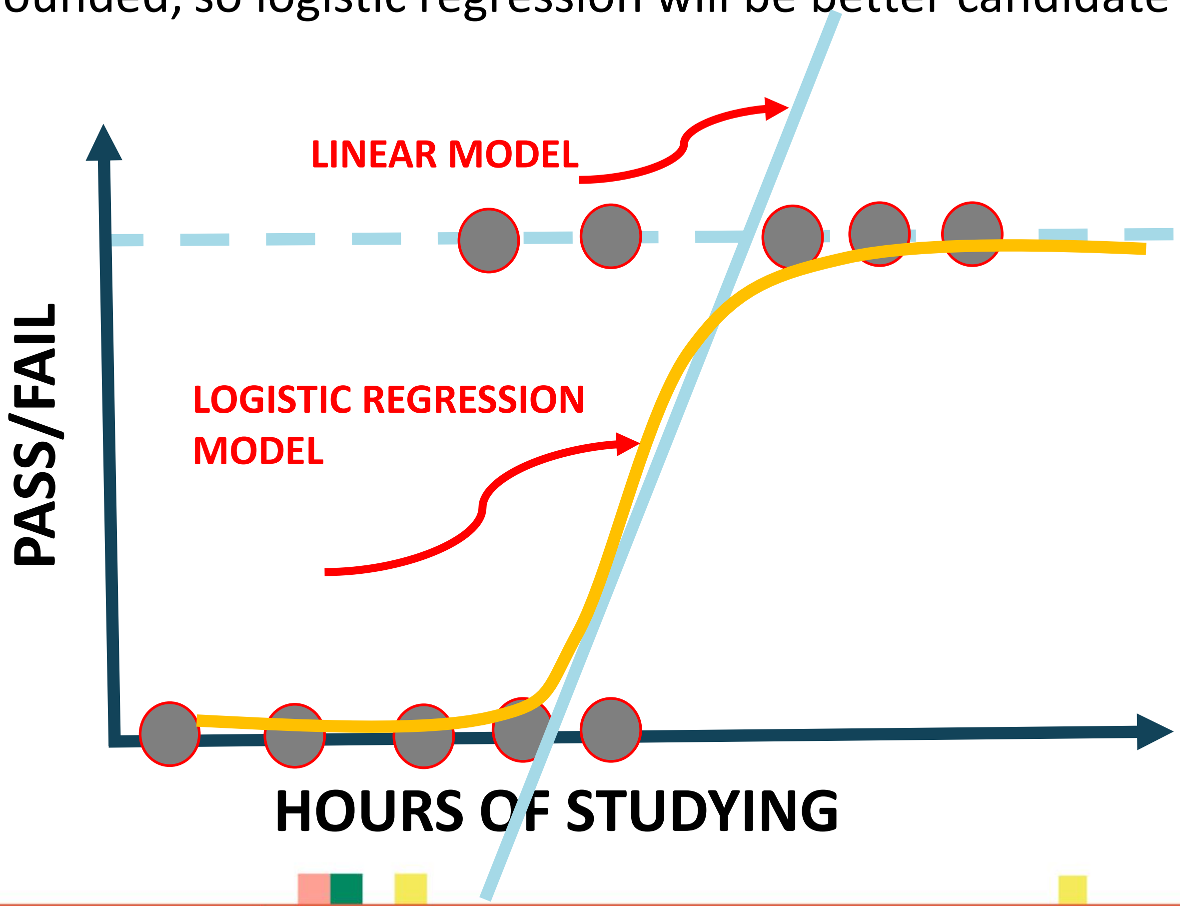
- **Linear regression** is used to predict outputs on a continuous spectrum.
  - Example: predicting revenue based on the outside air temperature.
- **Logistic regression is used to predict binary outputs** with 2 possible values (0 or 1)
  - Logistic model output can be one of two classes: pass/fail, win/lose, healthy/sick



Hours Studying	Pass/Fail
1	0
1.5	0
2	0
3	1
3.25	0
4	1
5	1
6	1

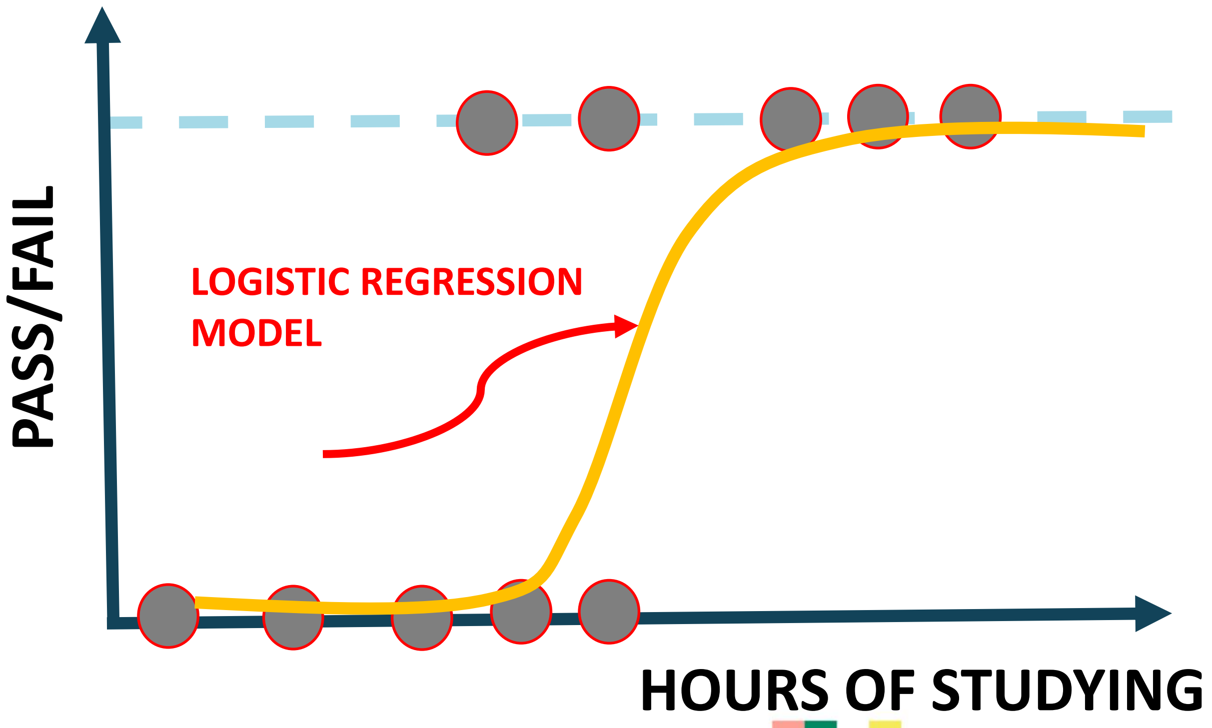
LOGISTIC REGRESSION: INTUITION

- Linear regression is not suitable for classification problem.
- Linear regression is unbounded, so logistic regression will be better candidate in which the output value ranges from 0 to 1.



LOGISTIC REGRESSION: MATH

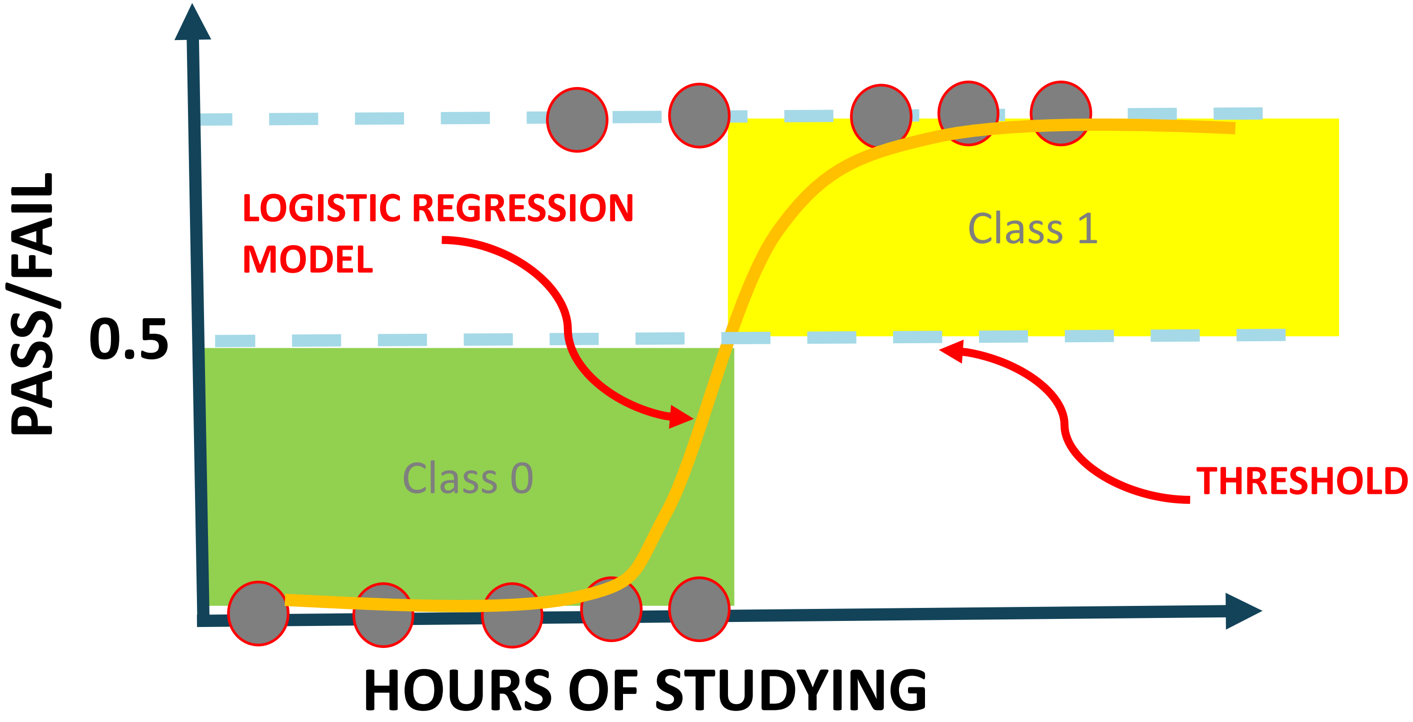
- Linear regression is not suitable for classification problem.
- Linear regression is unbounded, so logistic regression will be better candidate in which the output value ranges from 0 to 1.



- Linear equation:
  - $y = b_0 + b_1 * x$
- Apply Sigmoid function:
  - $P(x) = \ sigmoid\ (y)$
  - $P(x) = \frac{1}{1+e^{-y}}$
  - $P(x) = \frac{1}{1+e^{-(b_0+b_1*x)}}$

LOGISTIC REGRESSION: FROM PROBABILITY TO CLASS

- Now we need to convert from a probability to a class value which is “0” or “1”.

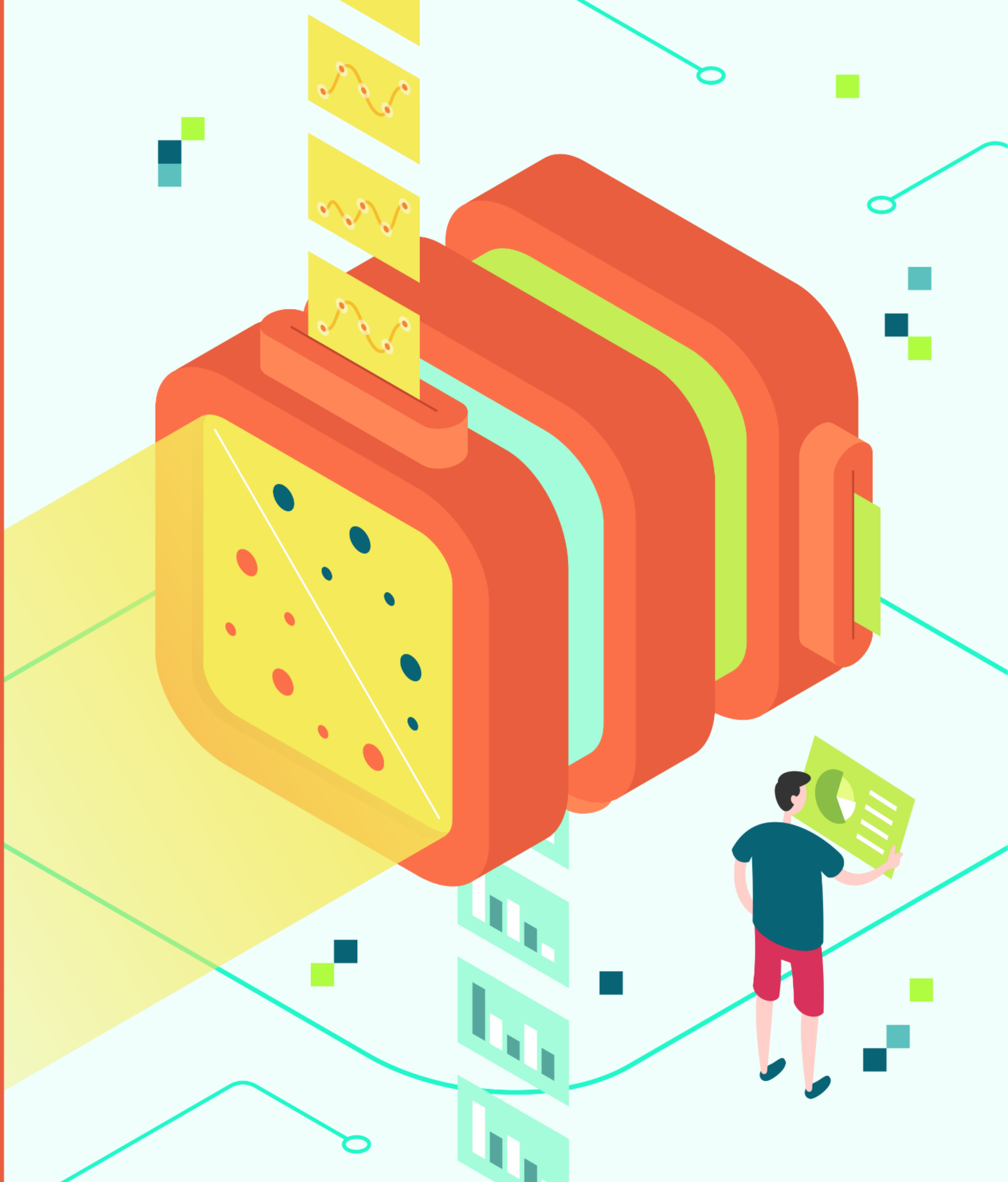


- <u>Linear equation:</u>
  - $y = b_0 + b_1 * x$
- <u>Apply Sigmoid function:</u>
  - $P(x) = \ sigmoid\ (y)$
  - $P(x) = \frac{1}{1+e^{-y}}$

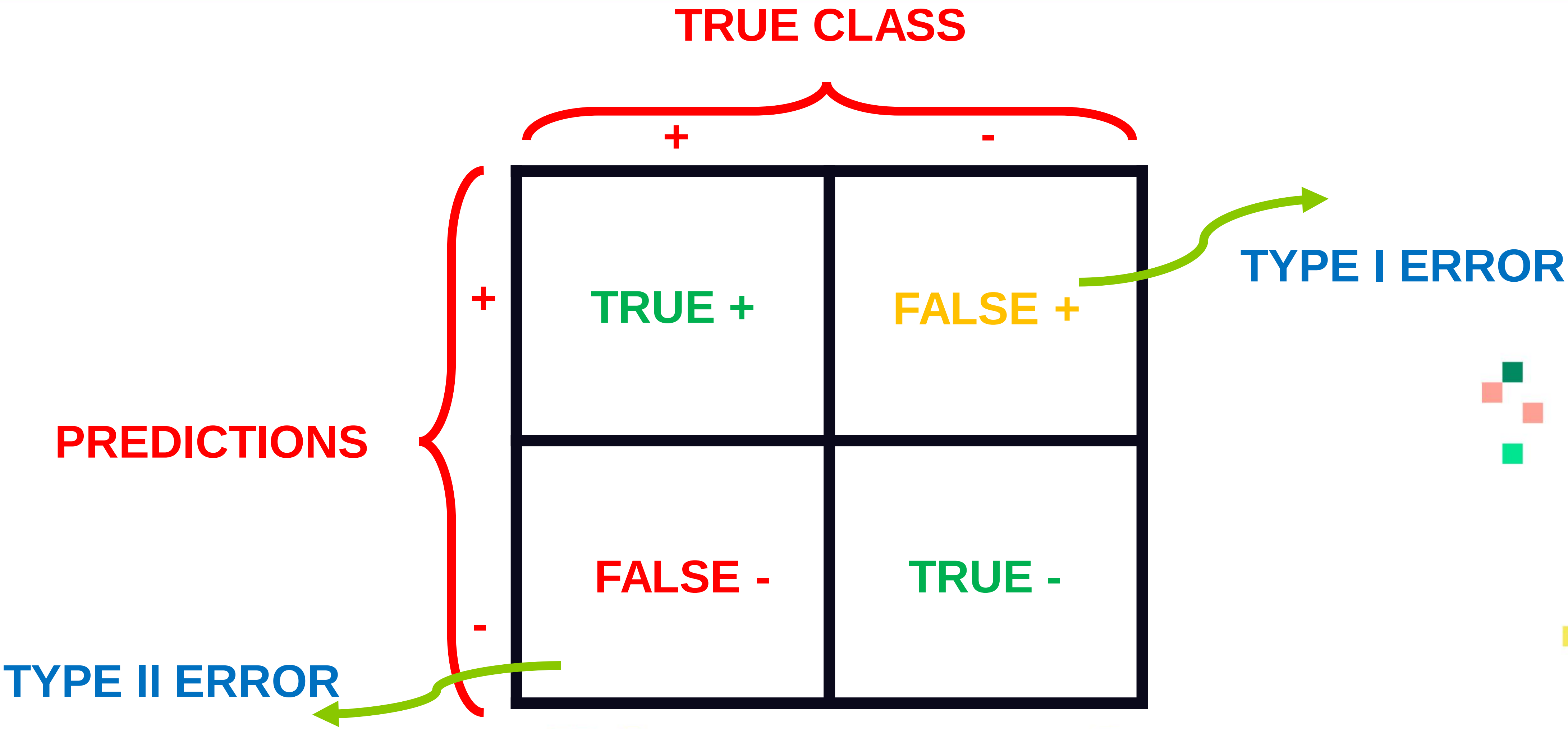
MACHINE LEARNING

REGRESSION

PERFORMANCE ASSESSMENT



CONFUSION MATRIX



CONFUSION MATRIX

- A confusion matrix is used to describe the performance of a classification model:
  - True positives (TP): cases when classifier predicted TRUE (they have the disease), and correct class was TRUE (patient has disease).
  - True negatives (TN): cases when model predicted FALSE (no disease), and correct class was FALSE (patient do not have disease).
  - False positives (FP) (Type I error): classifier predicted TRUE, but correct class was FALSE (patient did not have disease).
  - False negatives (FN) (Type II error): classifier predicted FALSE (patient do not have disease), but they actually do have the disease

KEY PERFORMANCE INDICATORS (KPI)

- Classification Accuracy = (TP+TN) / (TP + TN + FP + FN)
- Misclassification rate (Error Rate) = (FP + FN) / (TP + TN + FP + FN)
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) (When model predicted TRUE class, how often was it right?)
- Recall = TP/ Actual TRUE = TP/ (TP+FN) (when the class was actually TRUE, how often did the classifier get it right?)

KEY PERFORMANCE INDICATORS (KPI)

Accuracy

"How many predictions were correct out of all predictions made."

"Out of everything I guessed, how many did I get right?"

Example: You took a test with 100 questions and got 90 right,
Your accuracy is 90%

Precision

“Out of all the time you predicted ‘positive’, how many were actually positive?.”

“When I say someone has the disease, how often am I right?”

Example: If a model predicted 10 people had a disease, but only 7 really did, precision is 7/10 = 70%

Recall (Sensitivity or True Positive Rate)

“Out of all the people who actually had the disease, how many did you successfully find?”

“How good am I at catching the real positives?”

Example: If 10 people had the disease and your model correctly predicted 7 of them, recall is 7/10 = 70%

F1 Score

It is a balance between precision and recall

"Let's find a good middle ground between being right when I say 'positive' (precision) and not missing real positives (recall)"

Example: If precision = 70% and recall = 70%, the F1 Score is also 70%. If one is high and the other is low, the F1 Score will be closer to the lower value.

KEY PERFORMANCE INDICATORS (KPI)

Metric	Question it answers	When it's useful
Accuracy	"How often am I right overall?"	Balanced datasets
Precision	"When I predict positive, am I right?"	Avoiding **false alarms** (e.g., spam detection)
Recall	"Did I catch all the actual positives?"	Avoiding **misses** (e.g., cancer detection)
F1 Score	"Is there a balance between both?"	When precision and recall both matter

PRECISION Vs. RECALL EXAMPLE

TRUE CLASS

PREDICTIONS		+	-
	+	TP = 1	FP = 1
	-	FN = 8	TN = 90

- Classification Accuracy = (TP+TN) / (TP + TN + FP + FN) = 91%
- Precision = TP/Total TRUE Predictions = TP/ (TP+FP) = ½=50%
- Recall = TP/ Actual TRUE = TP/ (TP+FN) = 1/9 = 11%