Viviana Acquaviva

MACHINE LEARNING FOR PHYSICS AND ASTRONOMY

PRINCETON UNIVERSITY PRESS

A 15-MINUTE INTRO TO MACHINE LEARNING

vacquaviva@citytech.cuny.edu

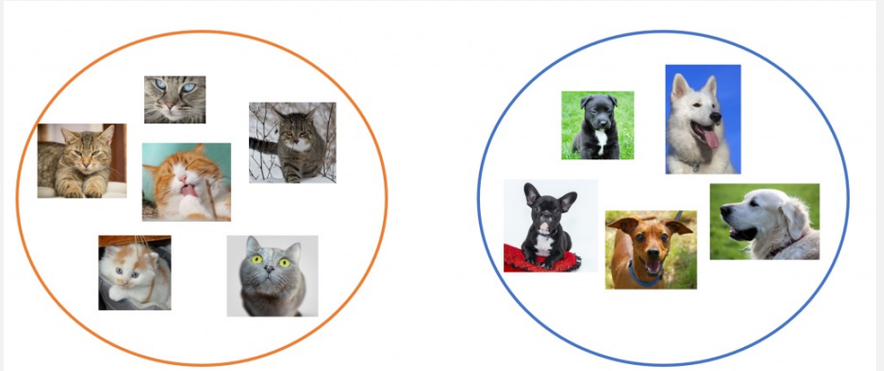
WHAT IS MACHINE LEARNING?

THE PROCESS OF TEACHING A MACHINE TO MAKE DECISIONS

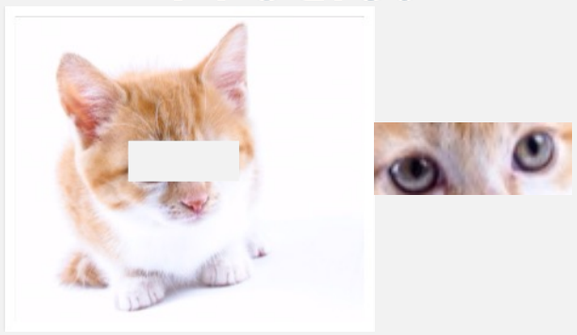
THE PROCESS OF TEACHING A MACHINE TO MAKE DECISIONS

Recognize

Group together

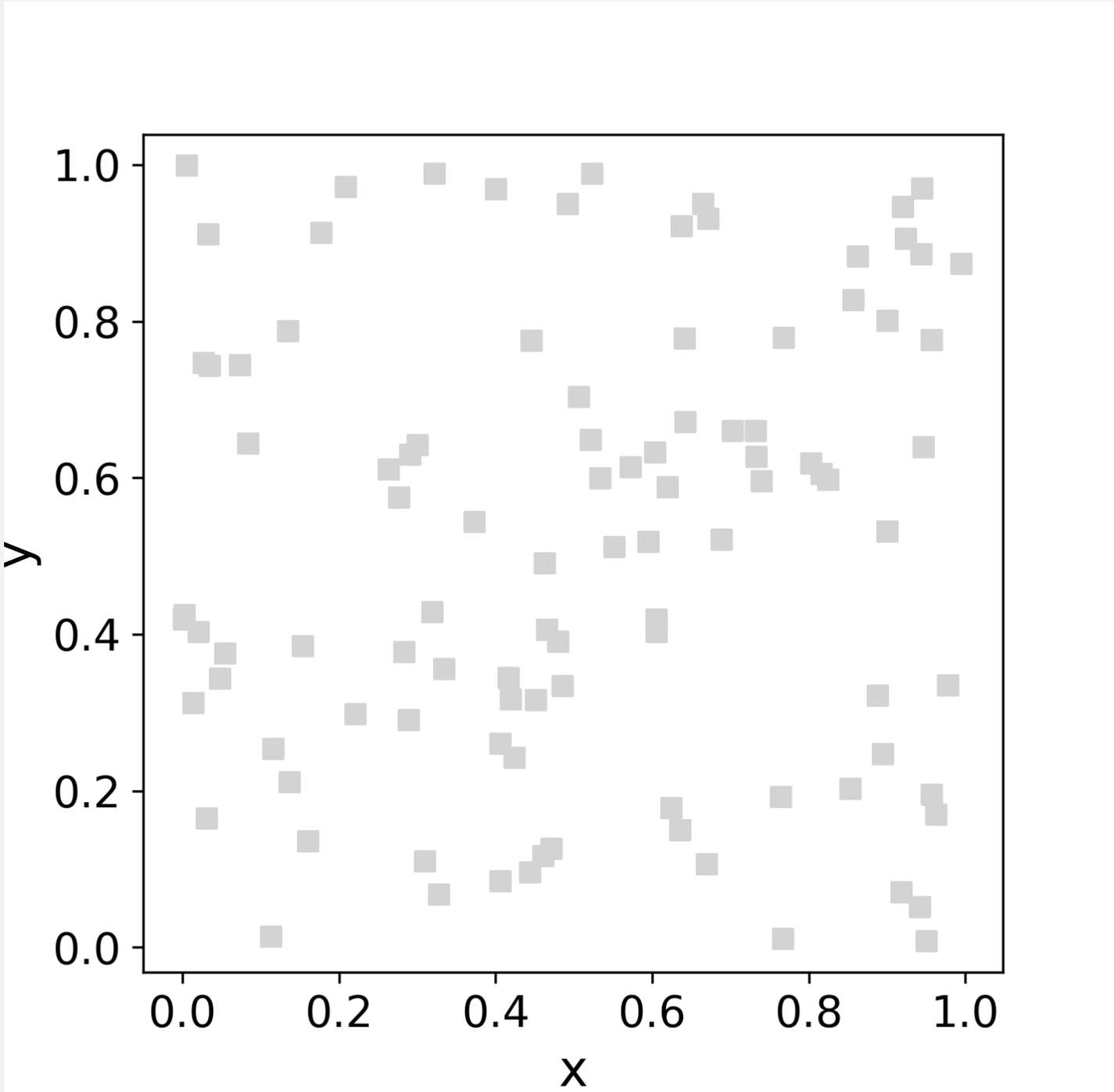


Predict

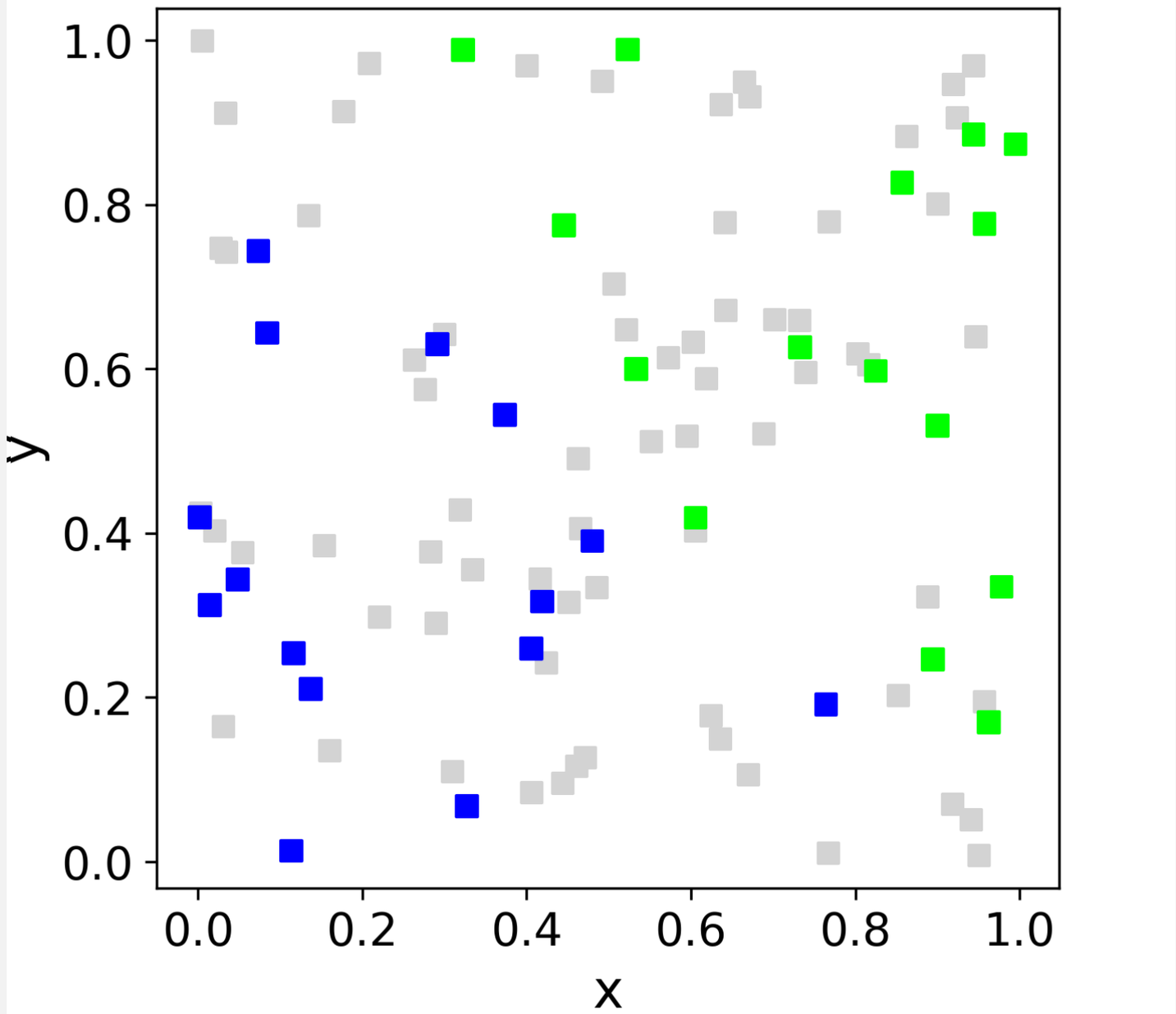


Simplify

OUR BRAIN MACHINE LEARNS



OUR BRAIN MACHINE LEARNS

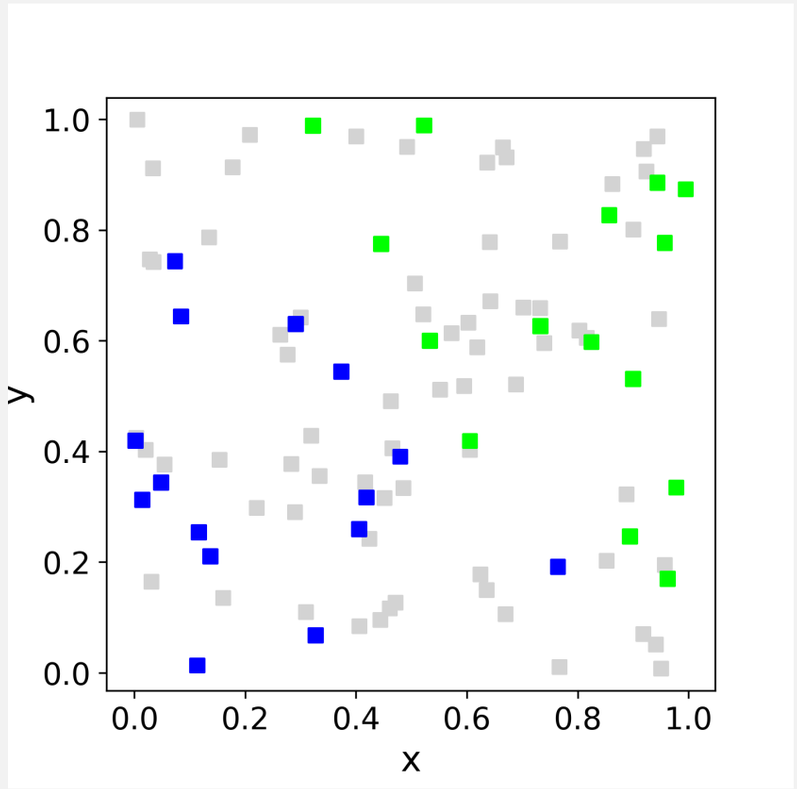


MACHINE LEARNING JARGON

Features are observable quantities, known for all objects (input)

Label or Target is the property that we want to predict (if it exists)

Instances (or examples) are the objects in our data set

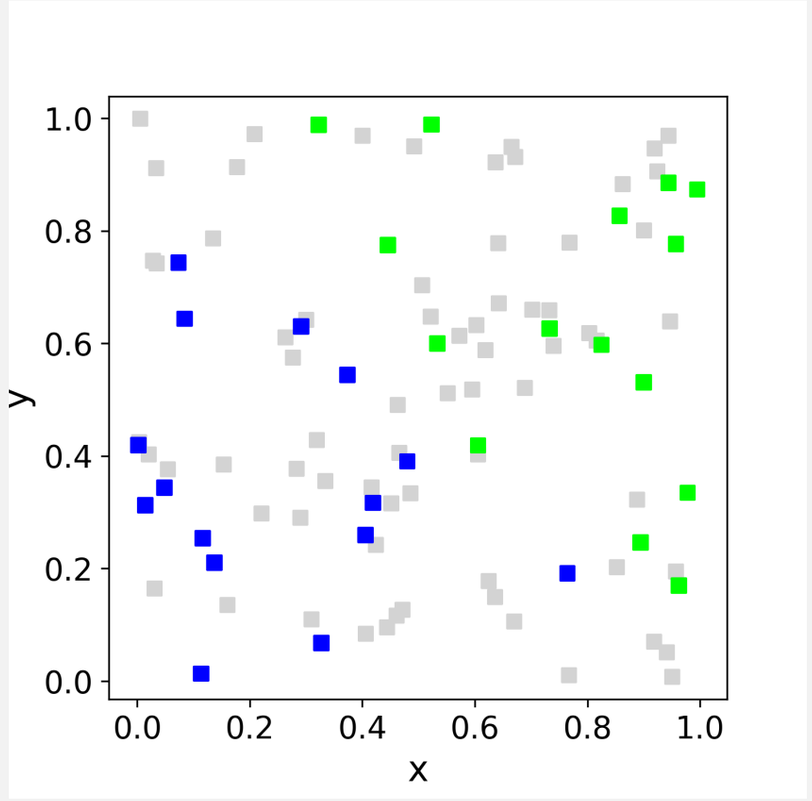


In this case…?

SUPERVISED MACHINE LEARNING

1. The desired output is an unknown quantity (label or target) that we'd like to predict from the input features on an object-by-object basis

2. we learn BY EXAMPLE: we need a set of objects with known labels. This is called the LEARNING SET



How many instances in our learning set?

SIZE AND QUALITY OF THE LEARNING SET ARE VERY IMPORTANT

INPUT	OUTPUT
1	3
2	3
3	?

INPUT	OUTPUT
one	3
two	3
three	5
four	4
five	4
six	?

Data representation

(sometimes called feature engineering)

and determining whether you have enough data to create a good model are crucial.

SUPERVISED LEARNING VISUAL SUMMARY

FEATURES (INPUT)

OUTPUT

M1	M2	M3	M4	M5	....	MN	T
Known	Known	Known	Known	Known	Known	Known	**Known**
Known	Known	Known	Known	Known	Known	Known	**Known**
Known	Known	Known	Known	Known	Known	Known	**Known**
Known	Known	Known	Known	Known	Known	Known	**Known**
Known	Known	Known	Known	Known	Known	Known	**Known**
Known	Known	Known	Known	Known	Known	Known	**Known**
Known	Known	Known	Known	Known	Known	Known	**Known**
Known	Known	Known	Known	Known	Known	Known	**(Known)**
Known	Known	Known	Known	Known	Known	Known	**?**

INSTANCES In LEARNING SET

LABELS/TARGETS

NEW DATA

PREDICTION

REGRESSION VS. CLASSIFICATION

We talk about classification when the target is a discrete variable (or class).

Let's look at this image recognition problem:

There are a finite (how many?) numbers of possible outcomes.

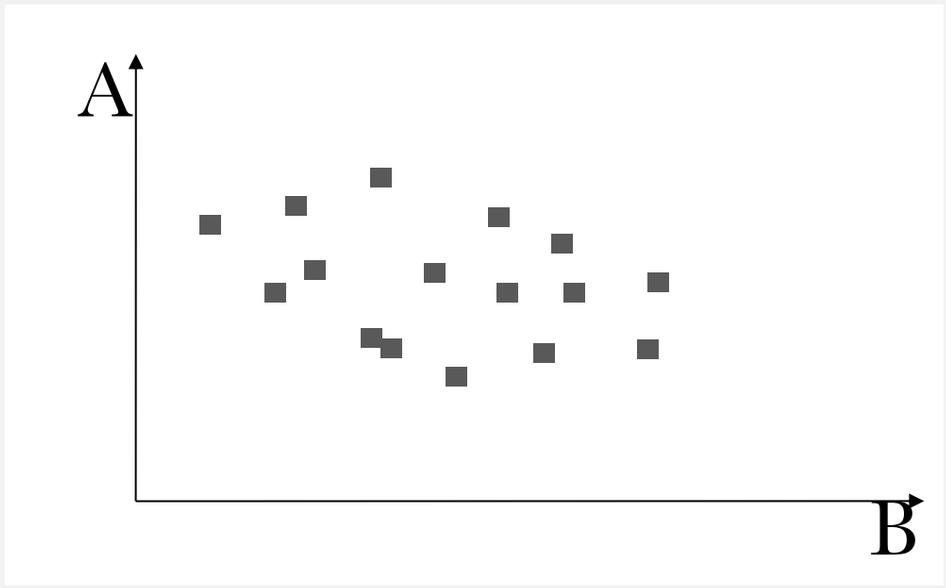
REGRESSION VS. CLASSIFICATION

- We talk about regression if the outcome is a continuous variable.
- For example, if I am trying to predict the probability that it will rain in a hour based on the current weather conditions, the outcome (target) is a continuous variable that can have all values between 0 and 1.

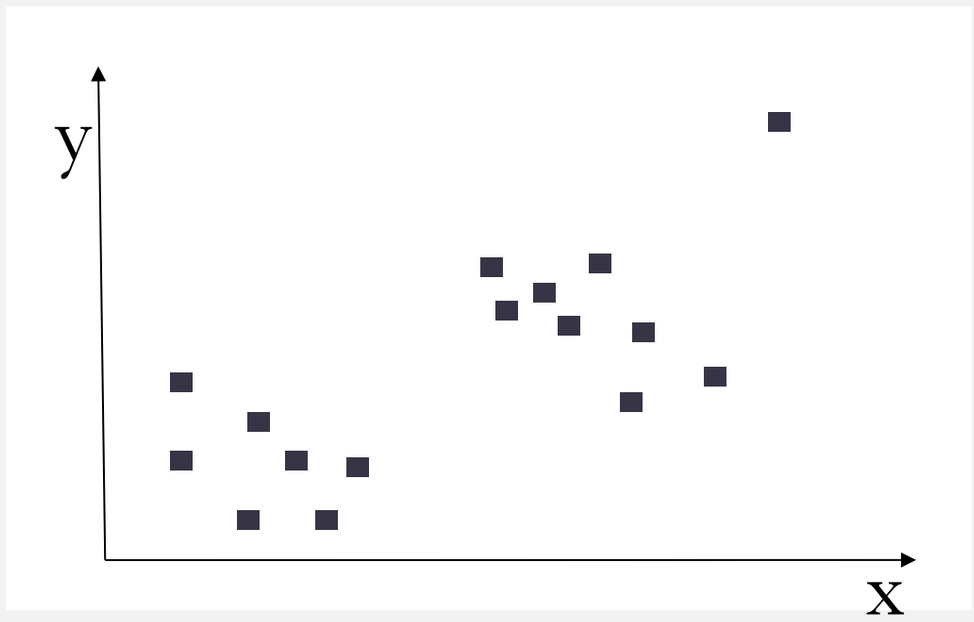
What if I was trying to decide whether I should bring an umbrella?

UNSUPERVISED MACHINE LEARNING

No labeled examples: goal is to find a PATTERN

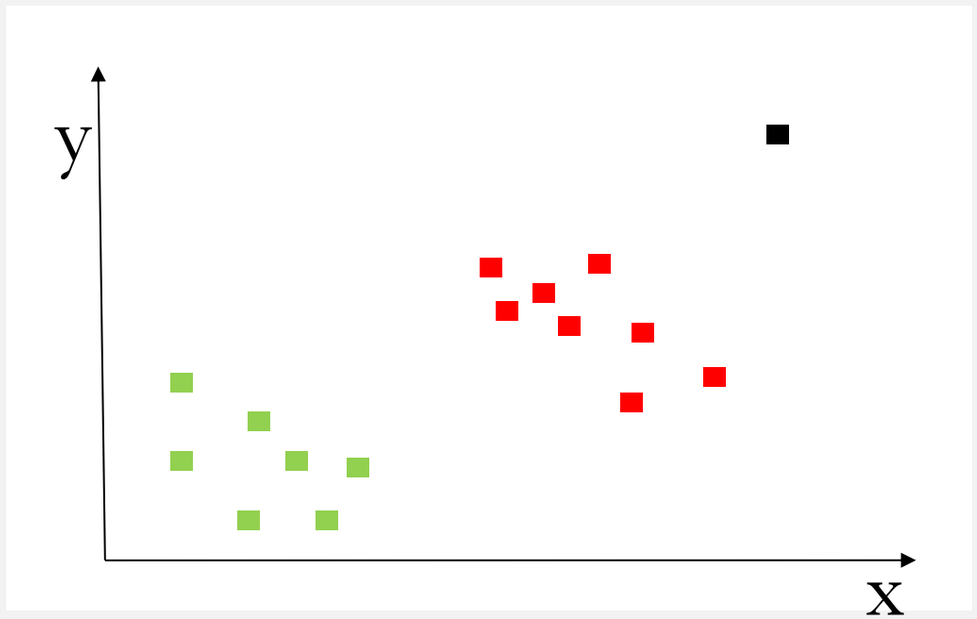


MAP



Useful to group together similar objects, find outliers, or find more efficient representations of data

Can be combined with human input or limited labels to understand the groups



ALGORITHM #1:

DECISION TREES

DECISION TREES

- Work by splitting data on different values of features
- Simplest trees are binary trees
- If categorical features, the split would be on yes/no
- If numerical, the split would be on a certain value (e.g. $x > 100$ or $x < 100$)

EXAMPLE: THIS 2-FEATURE DATA SET.

HOW SHOULD WE SPLIT?

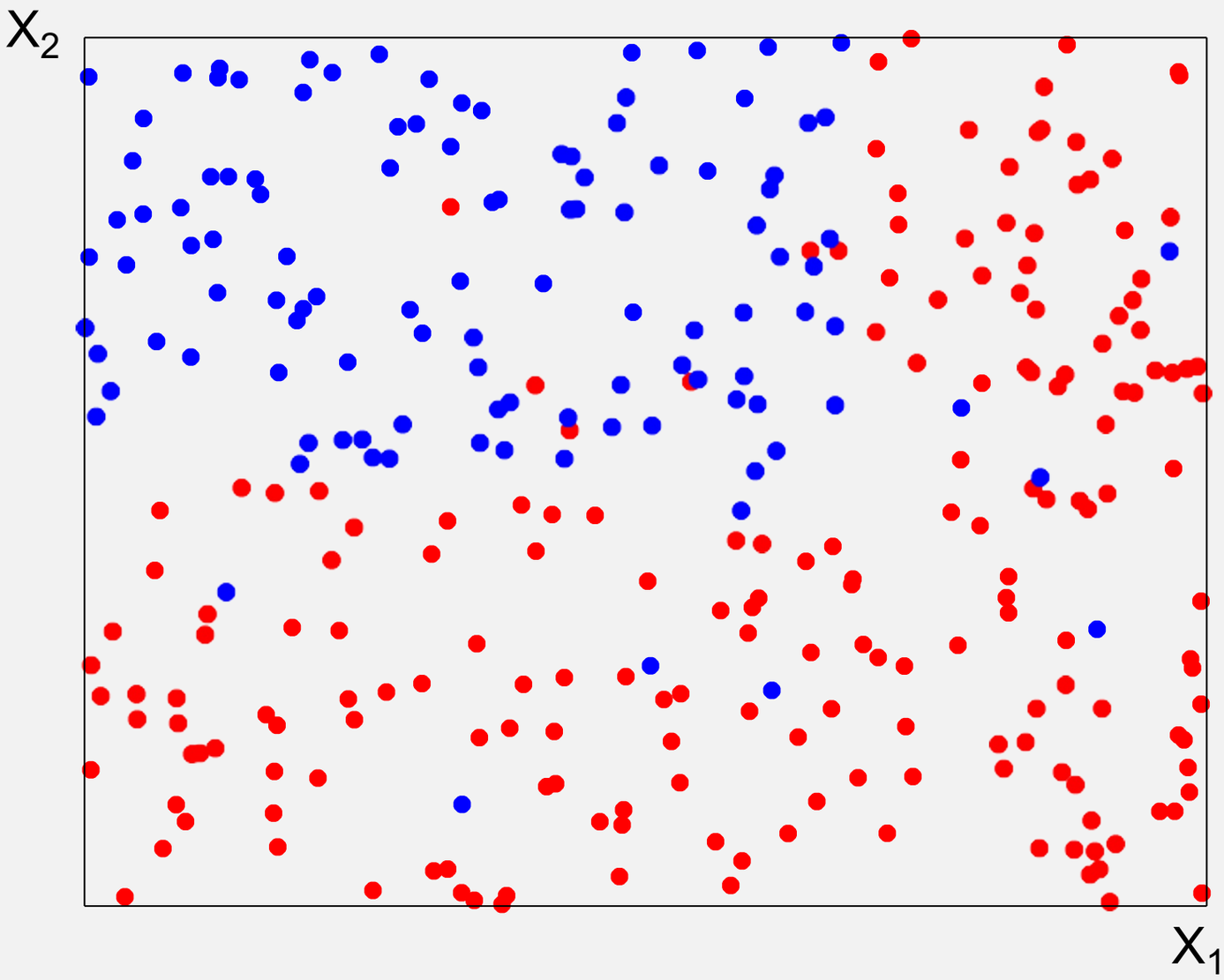


Figure credit:
Gilles Louppe

EXAMPLE: THIS 2-FEATURE DATA SET.

HOW SHOULD WE SPLIT?

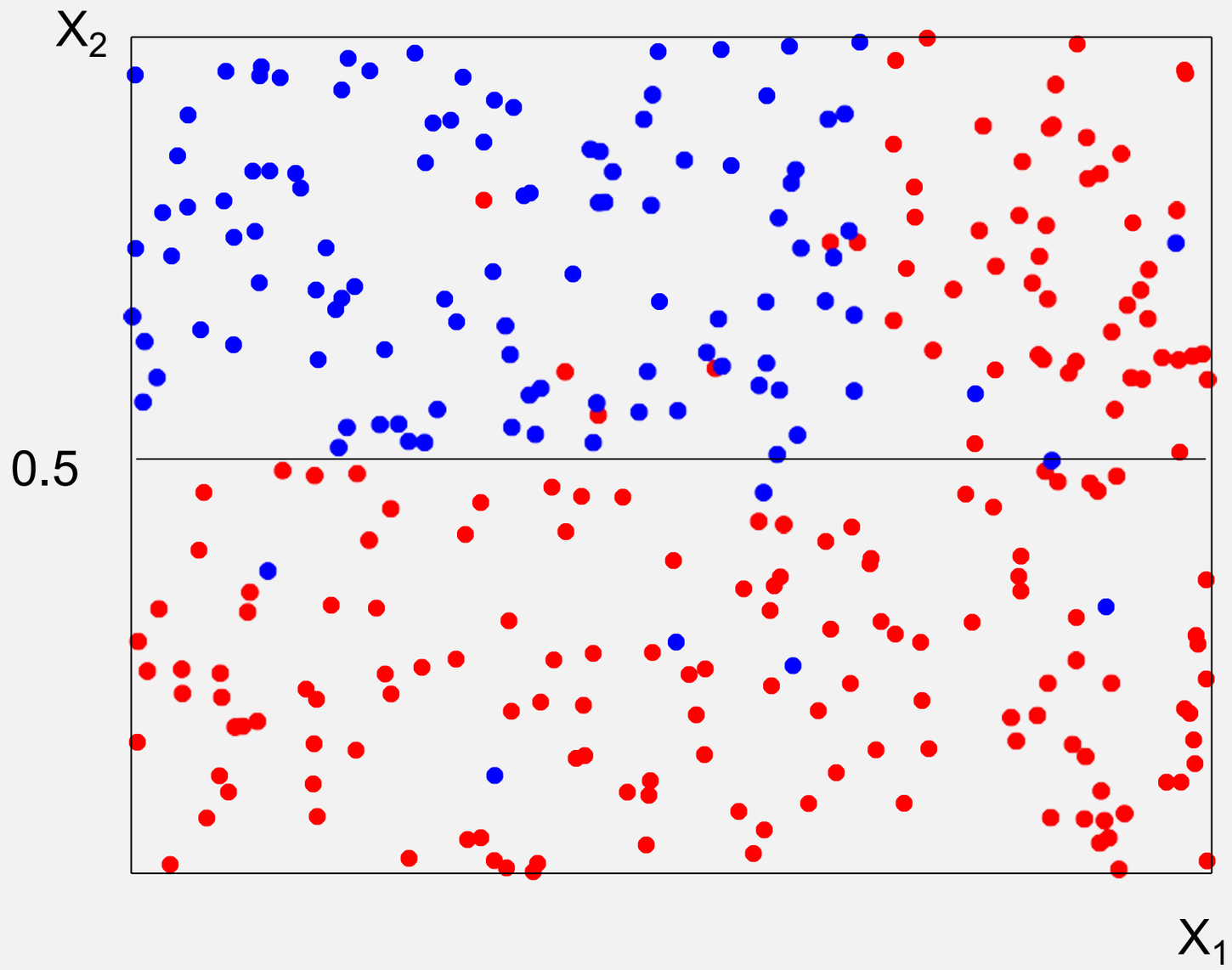


Figure credit:
Gilles Louppe

SHOULD WE STOP?

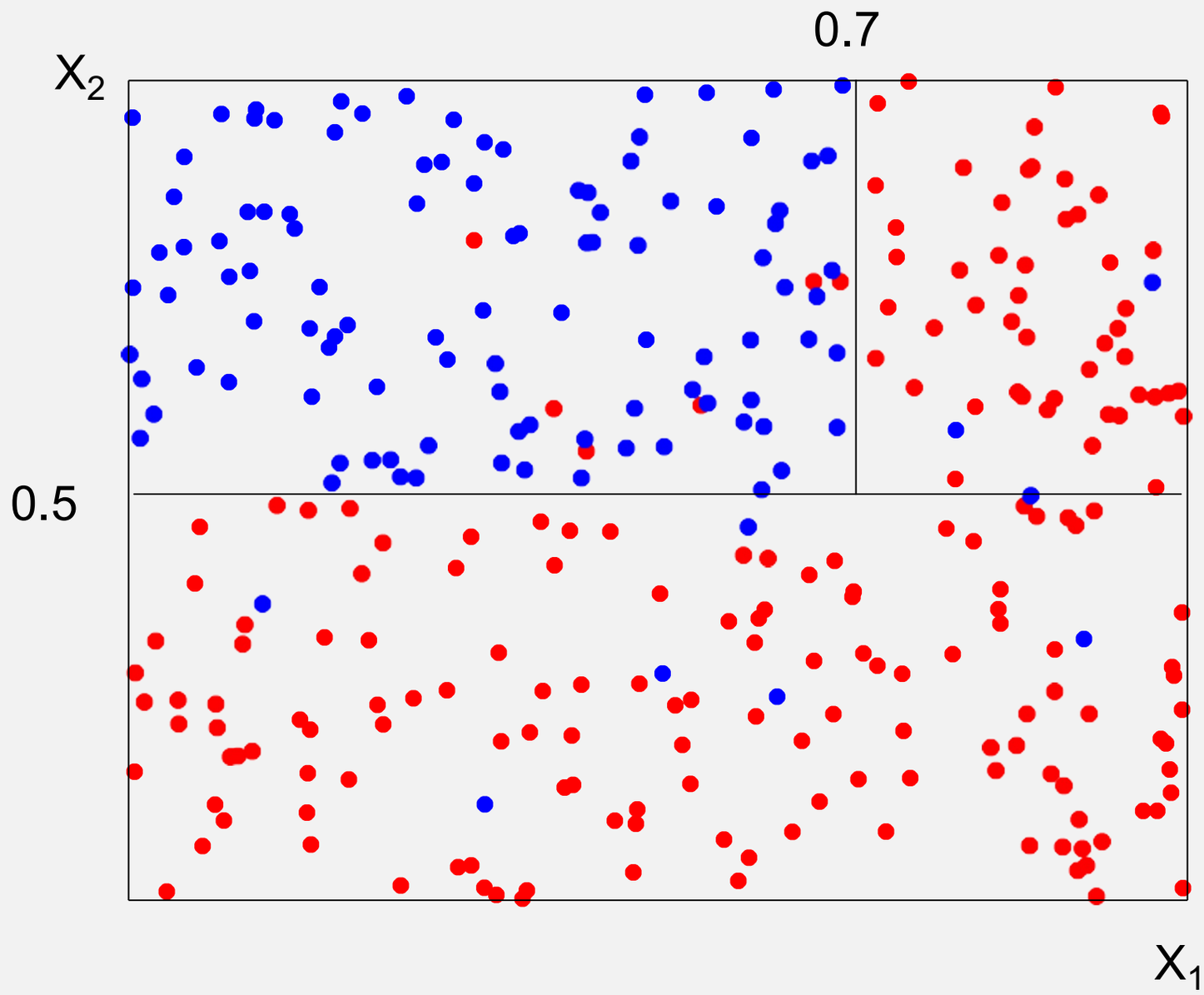
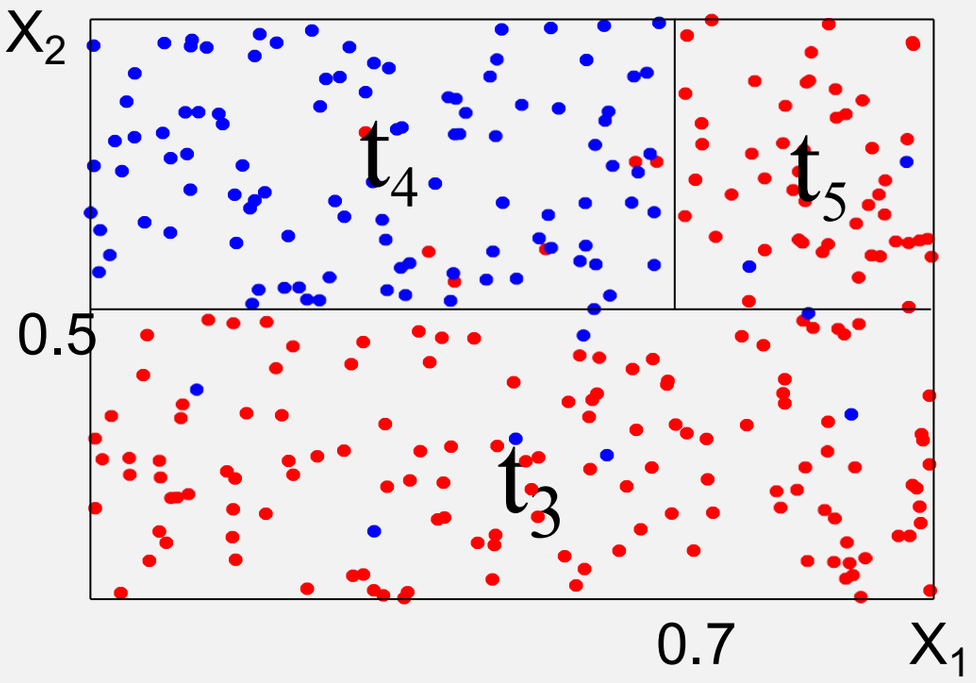
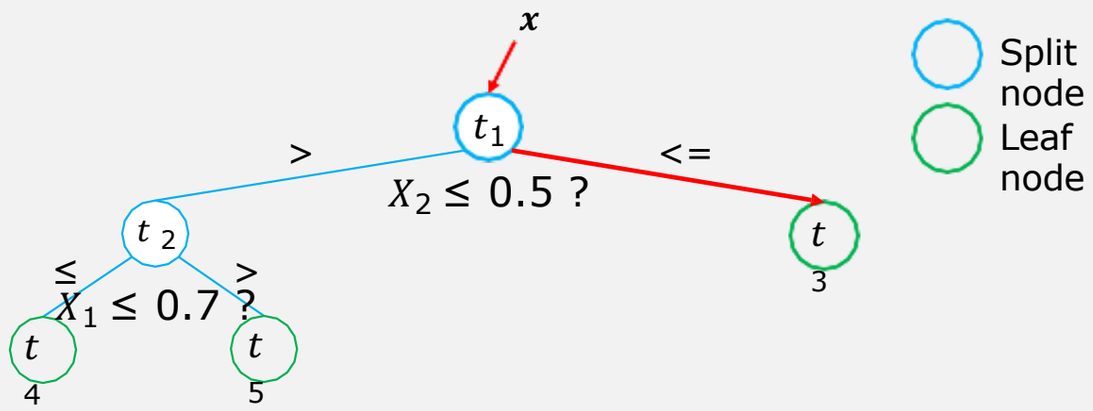


Figure credit:
Gilles Louppe

NODES (SPLITS AND LEAVES) DEFINE THE DECISION TREE





In a terminal node (leaf),
the model is ready to output a classification
(and all objects in that leaf have the same class)

Figure credit:
Gilles Louppe

Important questions:

How do we decide which splits to make, among the many possible ones?

How do we decide whether we should stop?

BUILDING DECISION TREES

Find measure of impurity (e.g. Gini impurity) that we want to minimize.

Find splits that maximize decrease of impurity

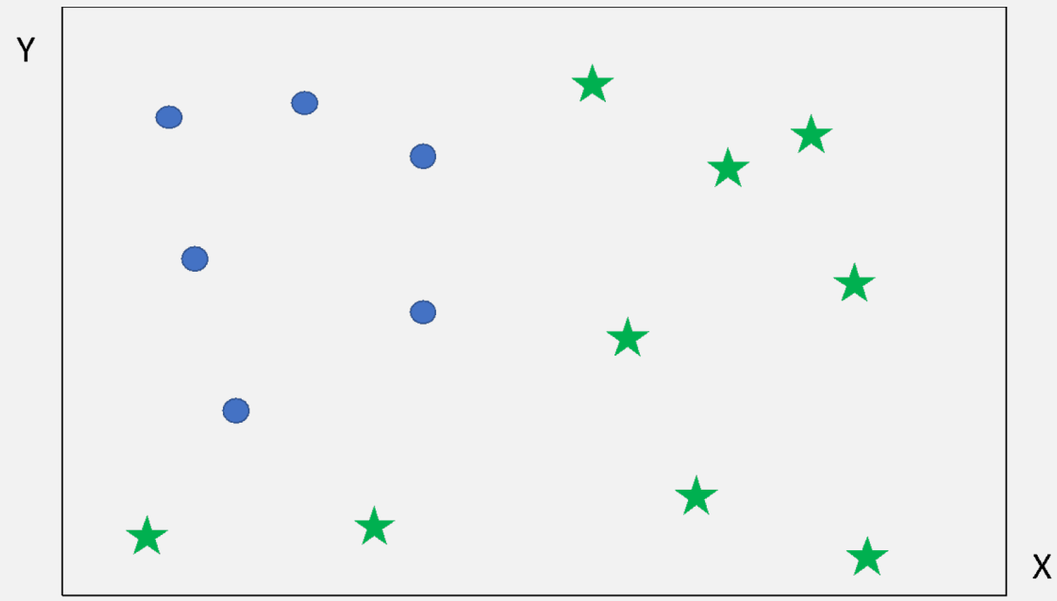
Select stopping criterion as impurity < ε (e.g. , 0)

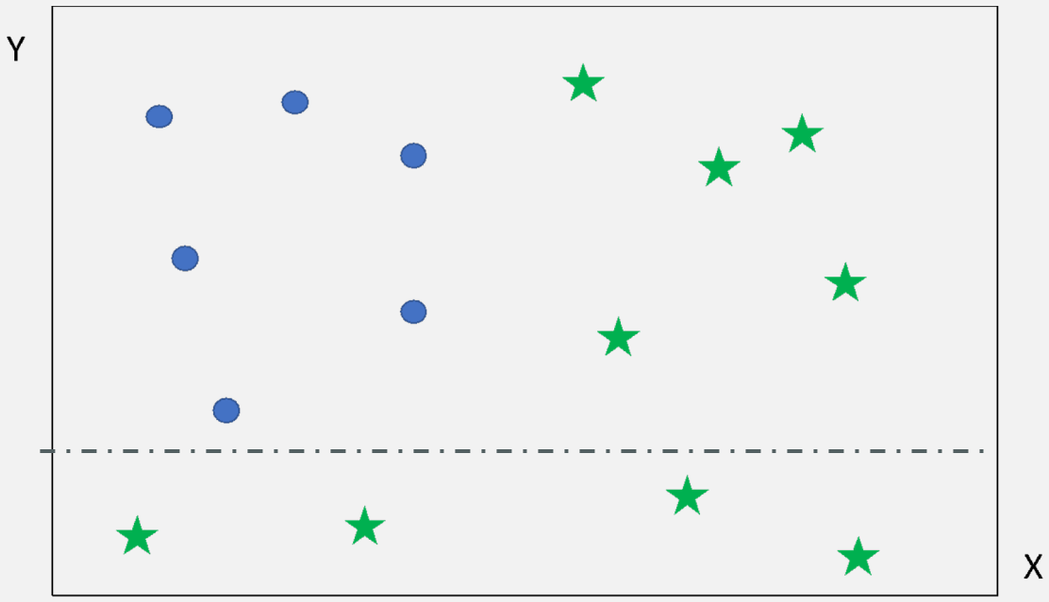
$$\text{Gini (node L)} = 1 - \sum f(i)^2$$

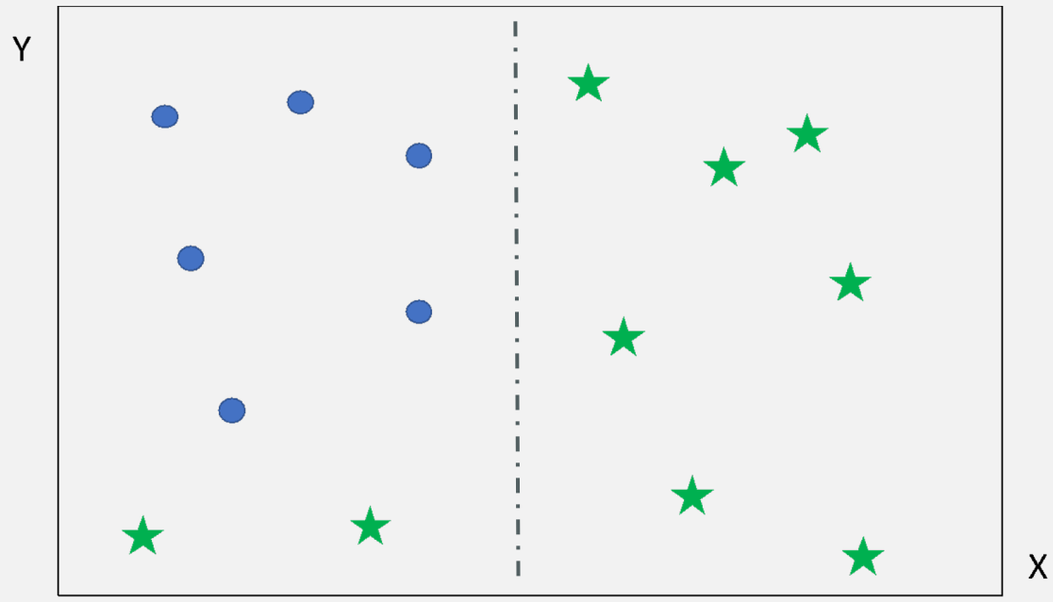
where f(i) is the fractional abundance of the i-eth class

$$L_L/L * (1 - \sum f(i)^2)_L + L_R/L * (1 - \sum f(i)^2)_R$$

L = total # of objects in original split; L_L and L_R = # of objects in each of the new splits



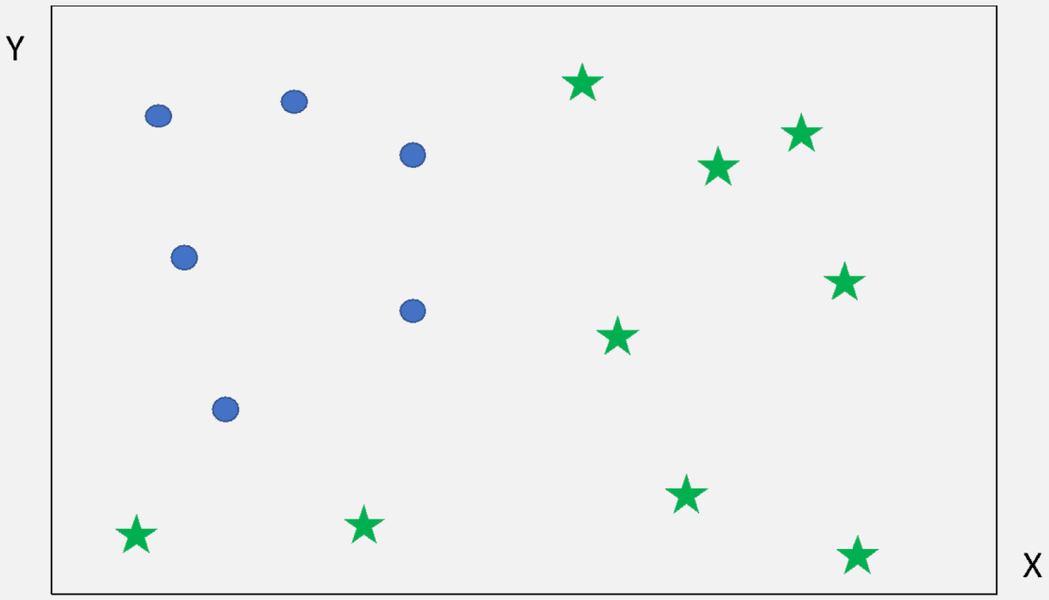




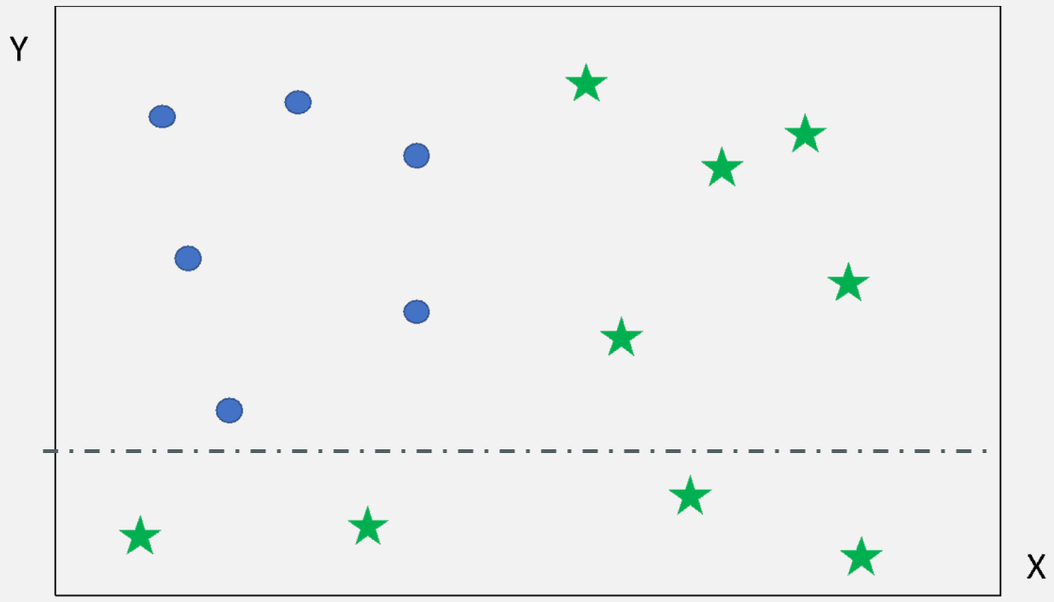
Which split should we do first?

Let's calculate the Gini impurity in the original and each of the two.

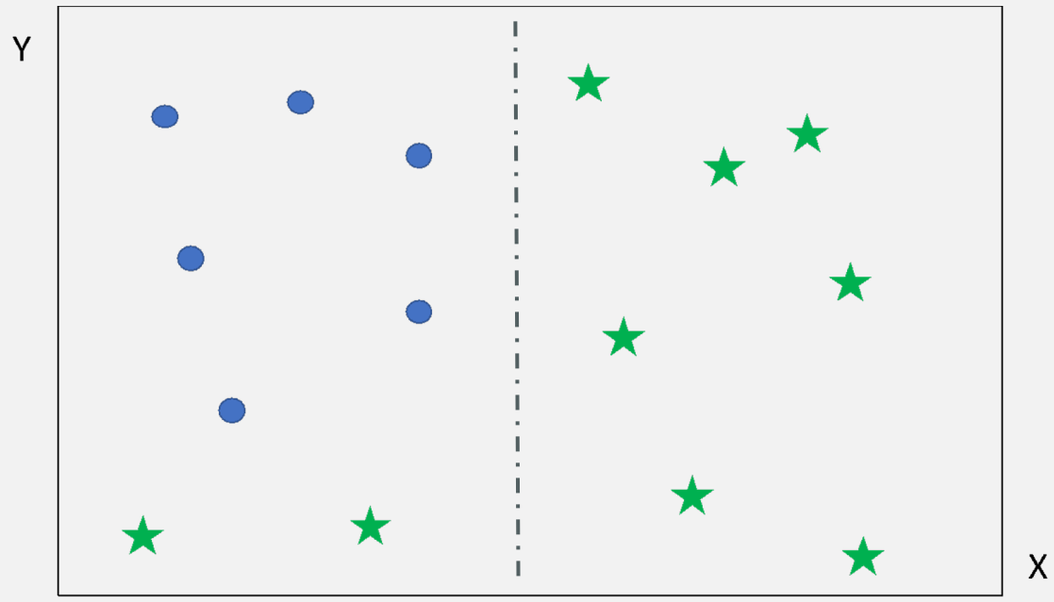
$$L_L/L * (1 - \sum f(i)^2)_L + L_R/L * (1 - \sum f(i)^2)_R$$



$$(1 - \sum f(i)^2) =$$
$$1 – (6/15)^2 – (9/15)^2$$
$$= 0.48$$



$$L_L/L * (1 - \sum f(i)^2)_L + L_R/L * (1 - \sum f(i)^2)_R$$
$$= 4/15 * 0 + 11/15 * (1 – (6/11)^2 – (5/11)^2) = 0.363$$



$$L_L/L * (1 - \sum f(i)^2)_L + L_R/L * (1 - \sum f(i)^2)_R$$
$$= 7/15 * 0 + 8/15 * (1-(2/8)^2 – (6/8)^2) = 0.2$$

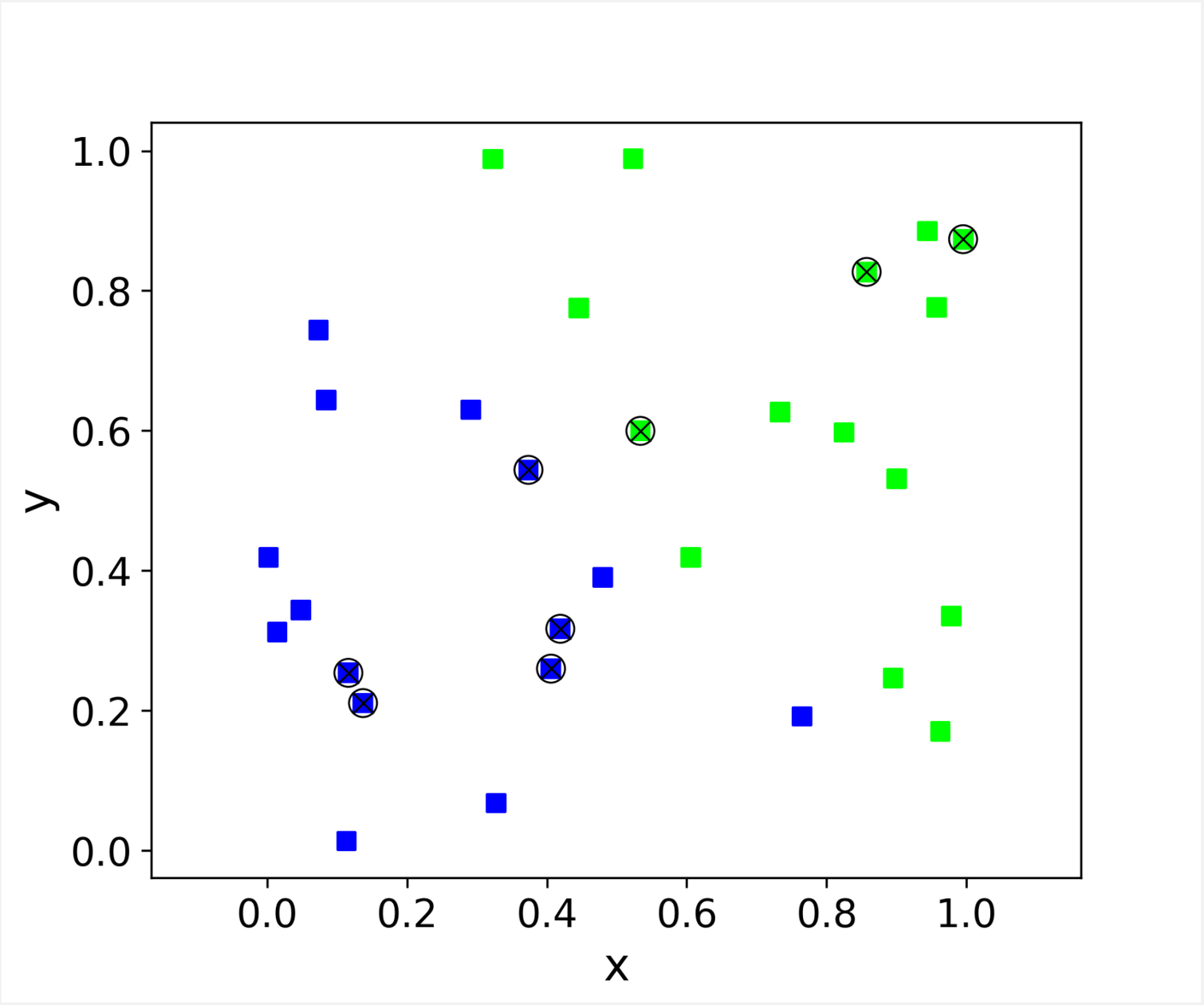
SUPERVISED LEARNING: BUILDING MODELS

IMPORTANT:
YOU SHOULDN'T USE THE ENTIRE LEARNING SET TO BUILD YOUR MODEL.

It is customary to split the learning set into a training set and test set

By building a model on the training set and applying it to the test set, you "mimic" what happens when your model sees new data for which the labels are not known. Otherwise you would be too optimistic! (note: you still are!)

CROSSED POINTS = TEST SET (SELECTED AT RANDOM)



Train set: used to build model

Test set: used to evaluate performance

Train score (error) = Performance on train set

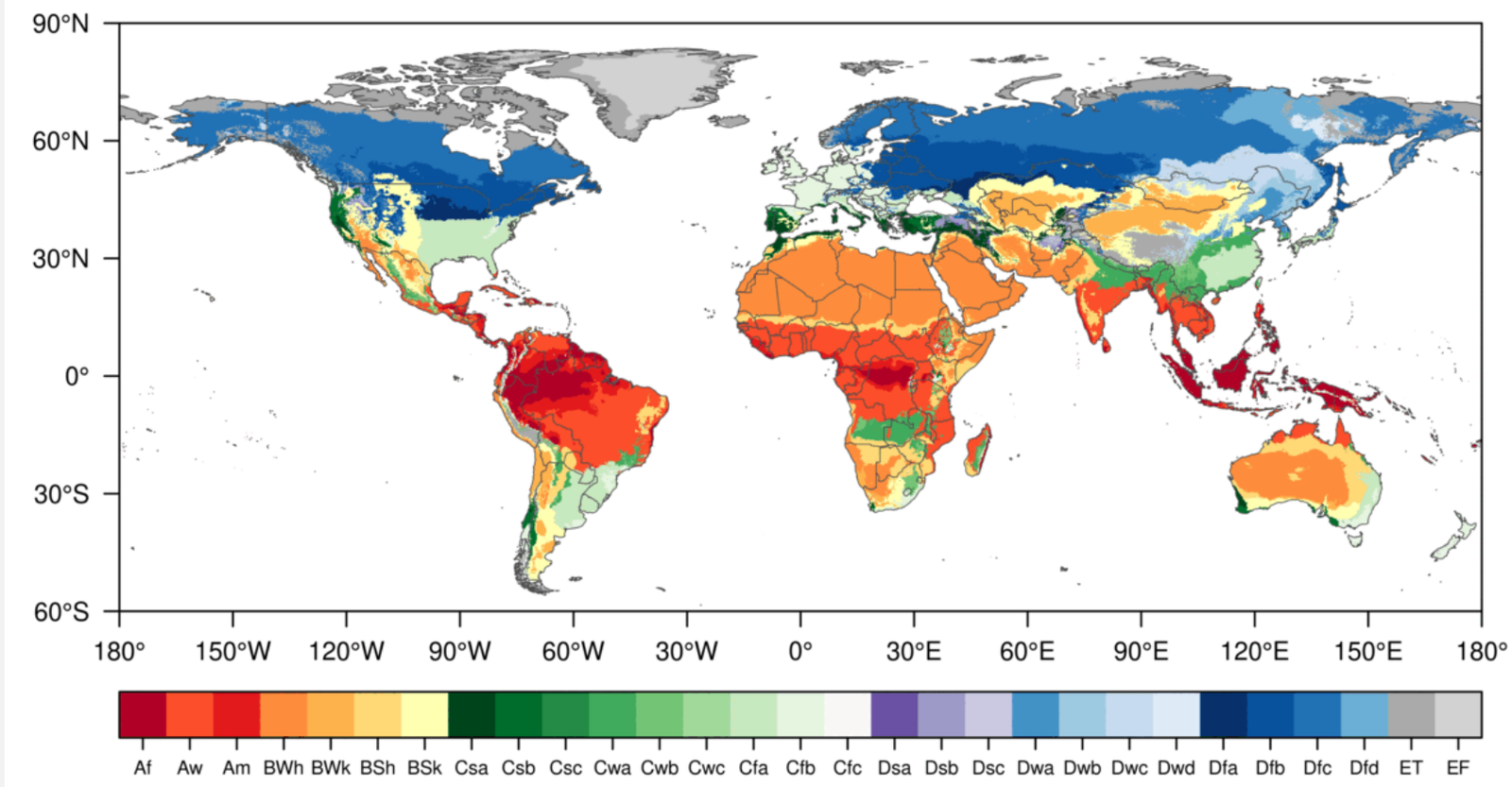
Test score (error) = Performance on test set

Generalization score (error) = Performance on new data

Test error is used as *proxy* for generalization error

We are now ready to build our first tree
(with pen and paper!)

OUR FIRST EXAMPLE WILL BE A SUPERVISED CLASSIFICATION PROBLEM,
IN WHICH WE ATTEMPT TO PREDICT
THE KÖPPEN-GEIGER CLIMATE
CLASSIFICATION OF A LOCATION
BASED ON SOME INFORMATION
ON TEMPERATURE AND PRECIPITATION.

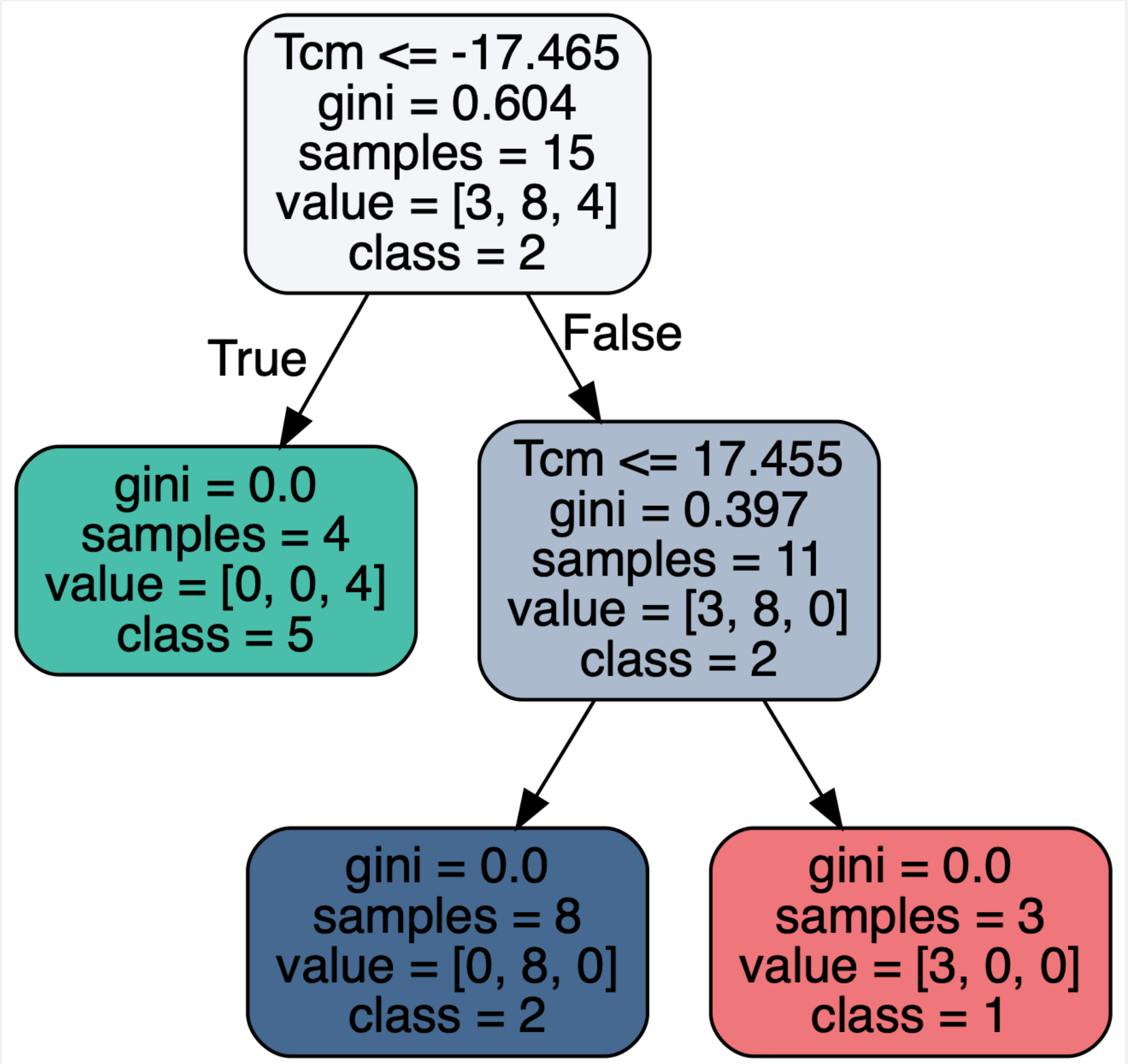


First	Second	Third	Description	Criterion
A			Tropical	Not (B) & $T_{cold} \geq 18$
	f		– Rainforest	$P_{dry} \geq 60$
	m		– Monsoon	Not (Af) & $P_{dry} \geq 100 - MAP/25$
	w		– Savannah	Not (Af) & $P_{dry} < 100 - MAP/25$
B			Arid	$MAP < 10 \times P_{threshold}$
	W		– Desert	$MAP < 5 \times P_{threshold}$
	S		– Steppe	$MAP \geq 5 \times P_{threshold}$
		h	– Hot	$MAT \geq 18$
		k	– Cold	$MAT < 18$
C			Temperate	Not (B) & $T_{hot} > 10$ & $-3 < T_{cold} < 18$
	w		– Dry winter	$P_{wdry} < P_{swet}/10$
	s		– Dry summer	Not (w) & $P_{sdry} < 40$ & $P_{sdry} < P_{wwet}/3$
	f		– Without dry season	Not (s) or (w)
		a	– Hot summer	$T_{hot} \geq 22$
		b	– Warm summer	Not (a) & $T_{mon10} \geq 4$
		c	– Cold summer	Not (a or b) & $1 \leq T_{mon10} < 4$
D			Boreal	Not (B) & $T_{hot} > 10$ & $T_{cold} \leq -3$
	w		– Dry winter	$P_{wdry} < P_{swet}/10$
	s		– Dry summer	Not (w) & $P_{sdry} < 40$ & $P_{sdry} < P_{wwet}/3$
	f		– Without dry season	Not (s) or (w)
		a	– Hot summer	$T_{hot} \geq 22$
		b	– Warm summer	Not (a) & $T_{mon10} \geq 4$
		c	– Cold summer	Not (a), (b) or (d)
		d	– Very cold winter	Not (a) or (b) & $T_{cold} < -38$
E			Polar	Not (B) & $T_{hot} \leq 10$
	T		– Tundra	$T_{hot} > 0$
	F		– Frost	$T_{hot} \leq 0$

gif from http://glass.umd.edu/KGClim; table from Cui et al 2021

T (wettest month) (C)	T (coldest month) (C)	Ptot	kc
35.55	21.88	944	1
25.15	14.01	353	2
4.43	-19.24	334	5
-5.49	-34.30	534	5
24.28	-8.93	421	2
34.49	15.84	34	2
23.80	19.07	1044	1
32.27	13.61	147	2
25.03	-10.76	216	2
2.35	-30.90	137	5
34.96	13.15	17	2
11.57	-15.30	50	2
8.56	-34.95	214	5
27.59	22.90	1439	1
23.06	-15.69	97	2
22.98	-7.31	282	2
27.37	11.52	66	2
27.27	23.58	1712	1
27.58	6.41	262	2
18.20	-9.11	37	2

THE OPTIMAL SOLUTION



RESOURCES

- All the data I used to make this exercise, together with the code and some classroom activity ideas, are here:

  https://github.com/vacquaviva/LEAP2025_ClimateClass/

- I wrote a book! You can buy it of course (yay), but you can also get all the slides and code for free (go to the "Resources" page):

  https://press.princeton.edu/books/paperback/9780691206417/machine-learning-for-physics-and-astronomy

- We built free online courses on ML for Physics and Astronomy in English and Spanish, sponsored by Flatiron's CCA. They follow the textbook, but buying the book is not required!

  https://openlearning.flatironinstitute.org/

- They have videos, quizzes, and notebooks with learning check-ins; thanks to team "Javioli" (w/ Olga Privman, and Jake Postiglione) and "Astromaquinarios" (w/ Genaro Suarez, Rosario Cecilio-Flores, Manuel Pichardo Marcano, and Lucia Perez)