Decision Trees – Stopping Conditions

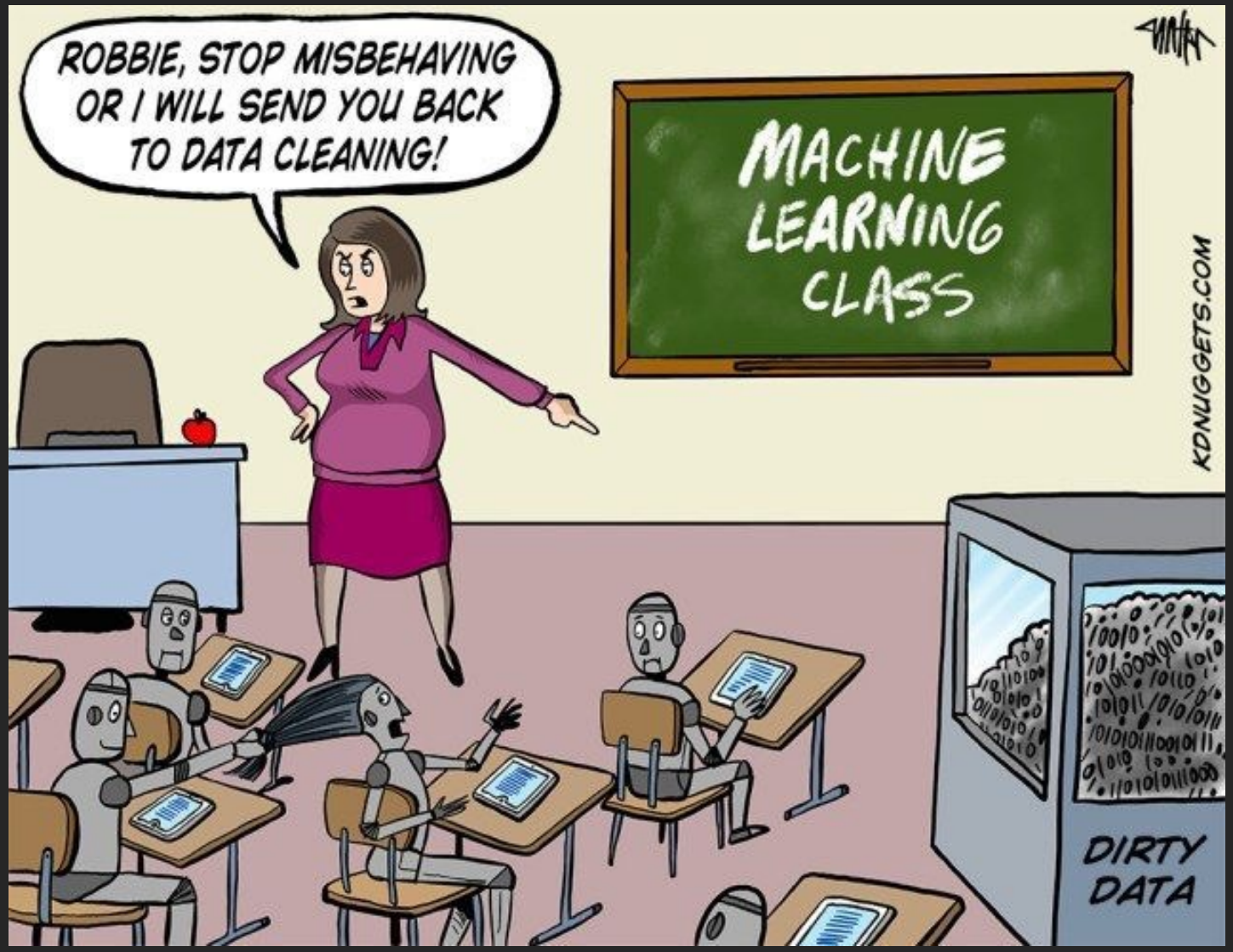
CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Gumb

Jenny Arakaki
Interlaken, Switzerland

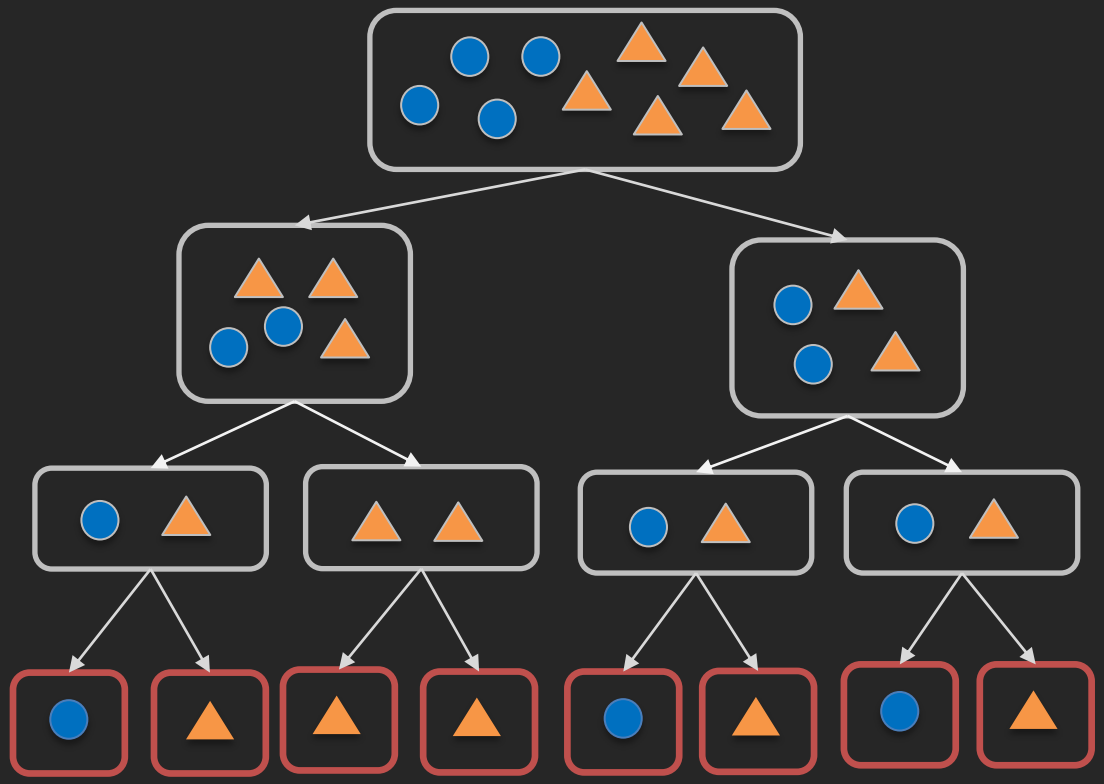
Outline

- Motivation
- Decision Trees – Classification
  - Intuition
  - Predictions
  - Splitting Criteria
  - Stopping Conditions


ROBBIE, STOP MISBEHAVING OR I WILL SEND YOU BACK TO DATA CLEANING!
MACHINE LEARNING CLASS
KDNUGGETS.COM
DIRTY DATA

Stopping Conditions

Question: If we don't terminate the decision tree algorithm manually, what will the leaf nodes of the decision tree look like?



The tree will continue to grow until each region contains **exactly one training point** and the model attains 100% **training** accuracy.

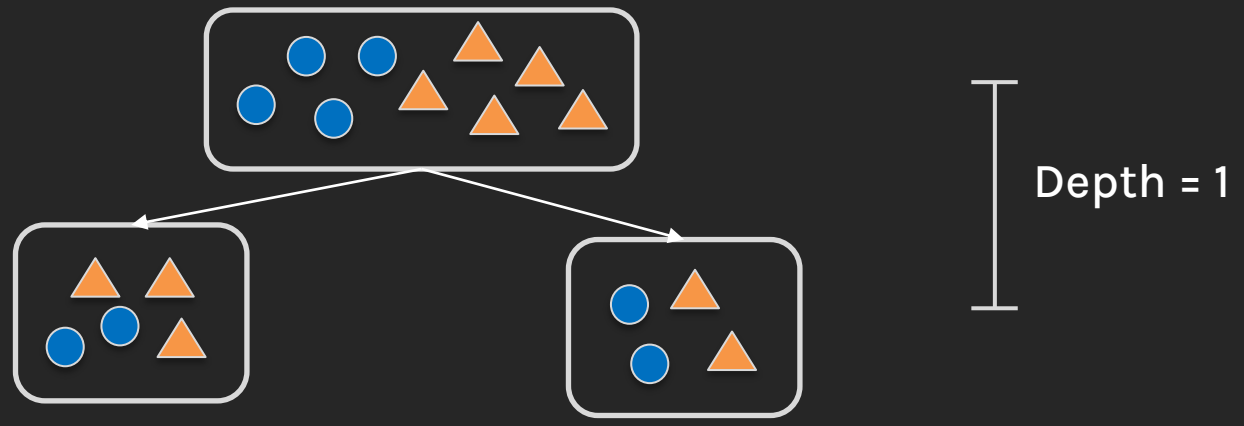
Stopping Conditions

Question: How can we prevent this from happening?

Stopping Conditions

The most common stopping condition is to limit the maximum depth *(max_depth)* of the tree.

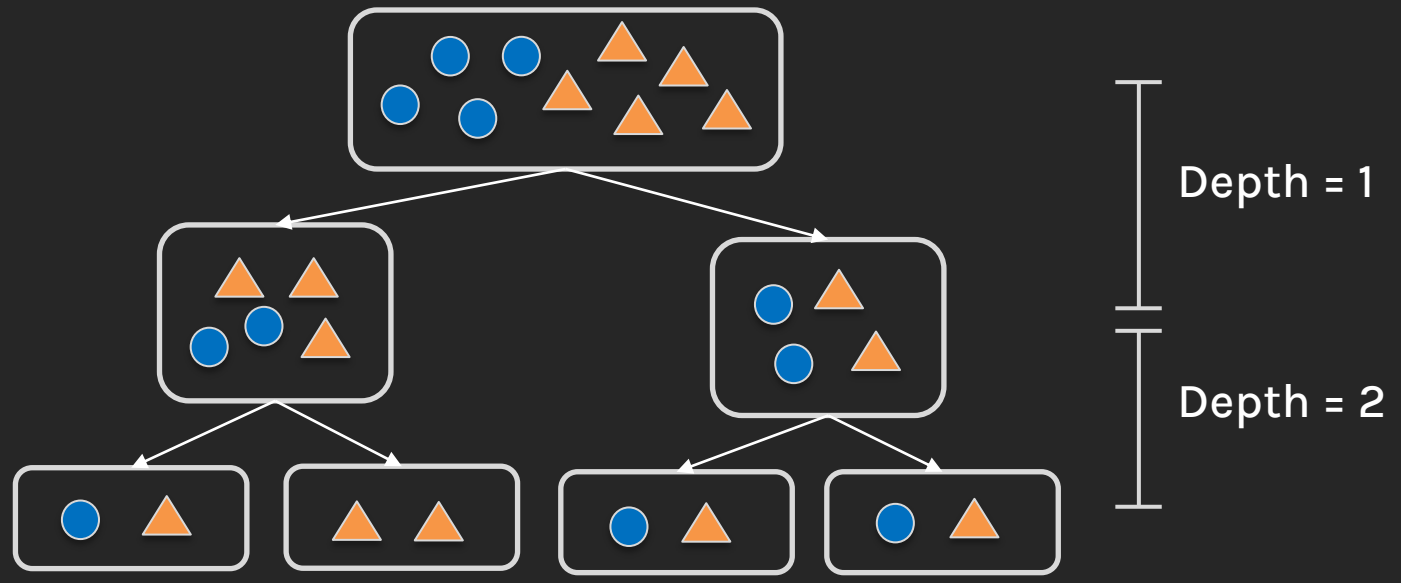
max_depth = 1



Stopping Conditions

The most common stopping condition is to limit the maximum depth *(max_depth)* of the tree.

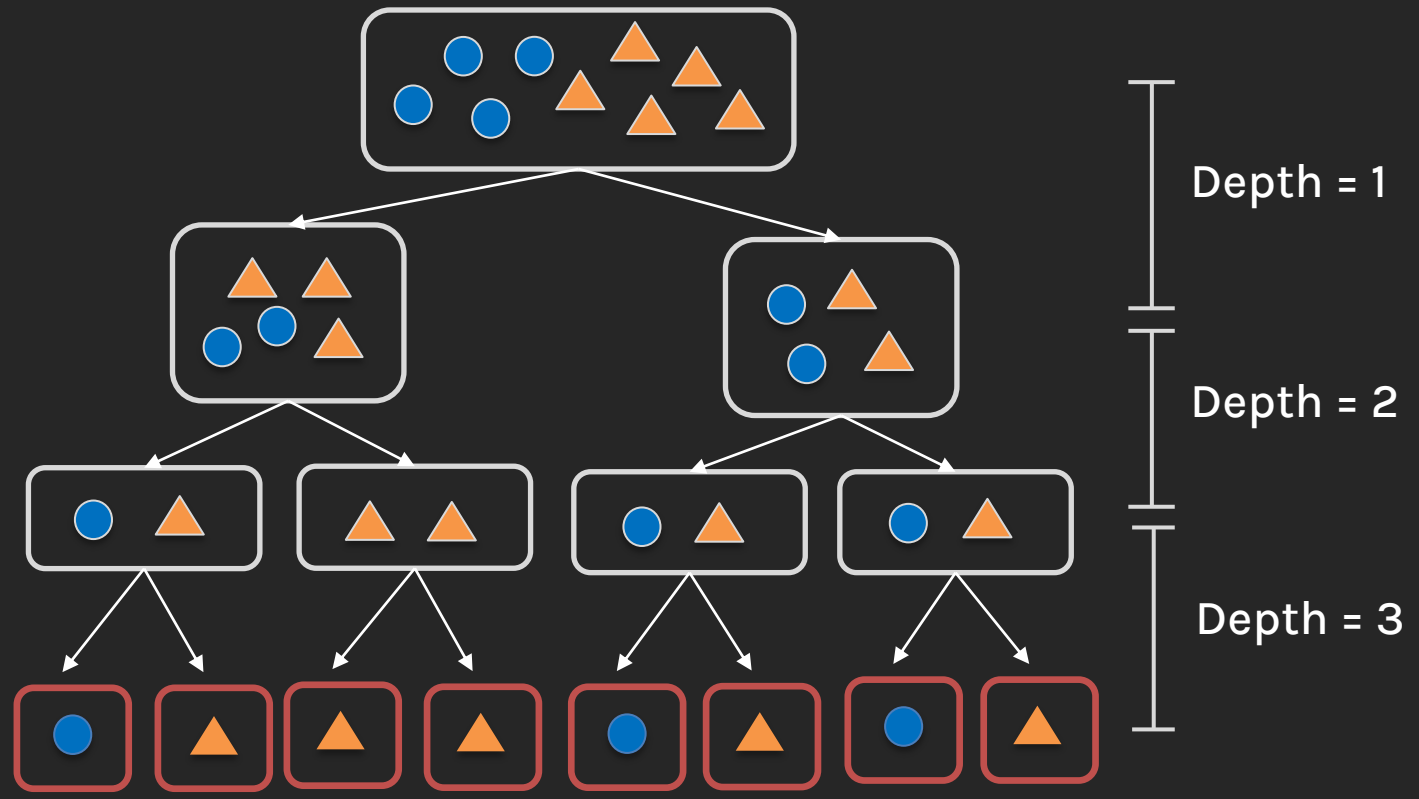
max_depth = 2



Stopping Conditions

The most common stopping condition is to limit the maximum depth *(max_depth)* of the tree.

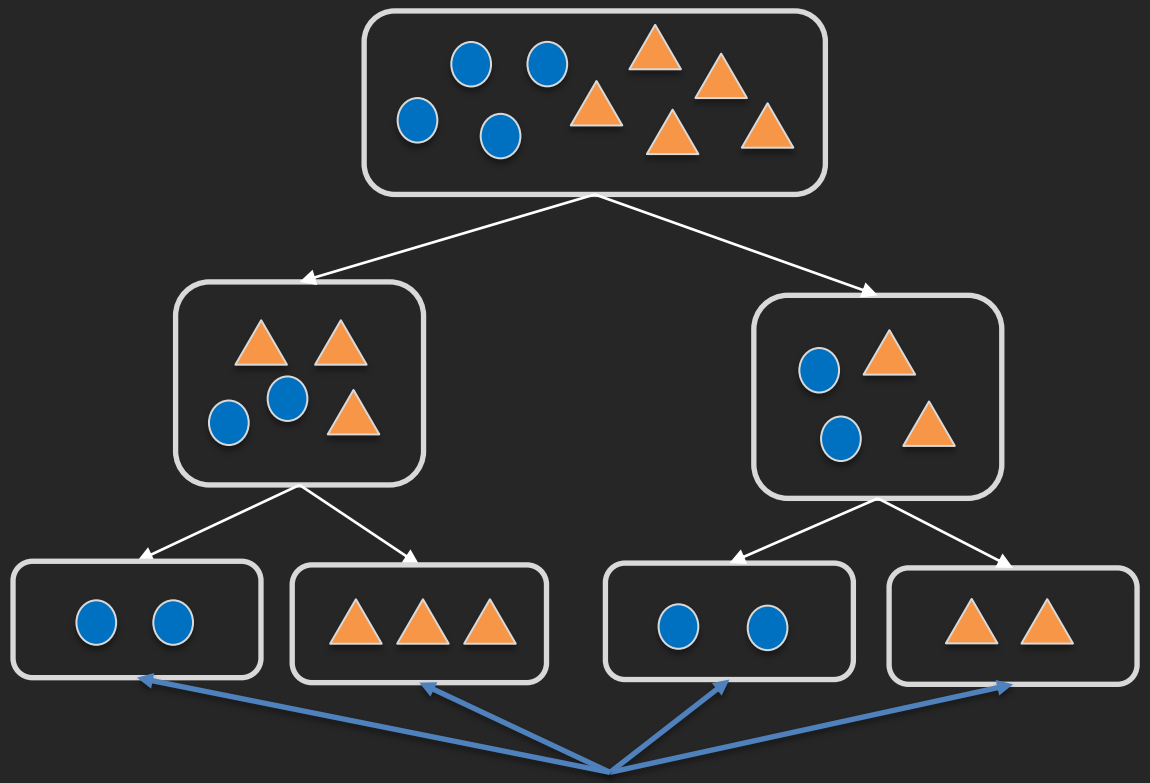
max_depth = 3



Stopping Conditions

Other common simple stopping conditions are:

- Don’t split a region if all instances in the region belong to the same class.

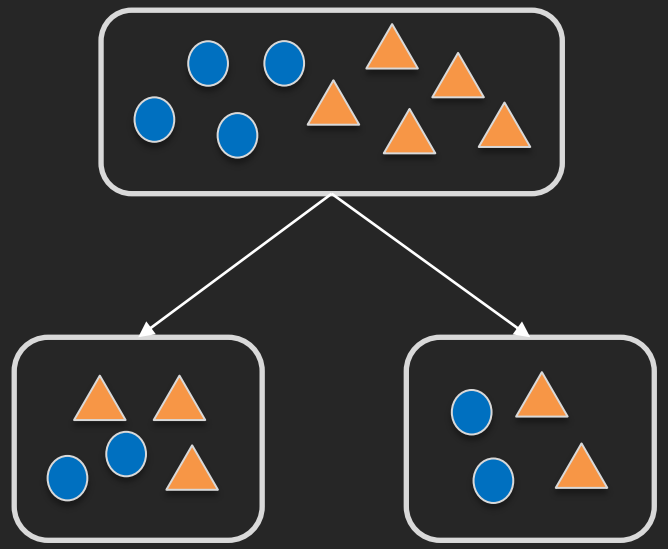


Stopping Conditions

Other common simple stopping conditions are:

- Don't split a region if the number of instances in any of the sub-regions will fall below pre-defined threshold *(min_samples_leaf)*.

min_samples_leaf = 4



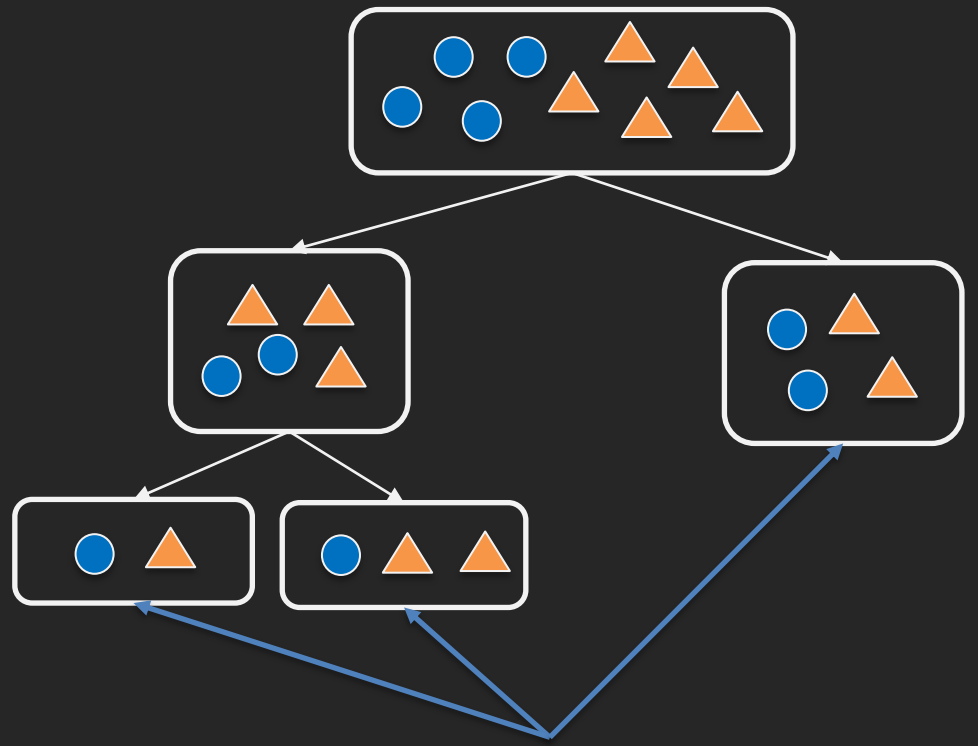
Stopping Conditions

Other common simple stopping conditions are:

- Don't split a region if the total number of leaves in the tree will exceed a pre-defined threshold *(max_leaf_nodes)*.

max_leaf_nodes = 3

Do you see any issue with this?



Stopping Conditions

Normally, Sklearn grows trees in what is called **'level-order'**-fashion until a stopping condition such as *max_depth* is met.

However, if a value for *max_leaf_nodes* is specified, Sklearn will instead grow the tree in a '**best-first**' fashion.

But what do level-order and best-first growth ***mean***?

Stopping Conditions

Normally, Sklearn grows trees in what is called **'level-order'**-fashion until a stopping condition such as `max_depth` is met.

However, if a value for `max_leaf_nodes` is specified, Sklearn will instead grow the tree in a '**best-first**' fashion.

But what do level-order and best-first growth ***mean***?

Level-order is also sometimes called '*breadth*-first'-search but we will use the term 'level-order' so it won't be confused with the similar sounding '*best*-first' search.

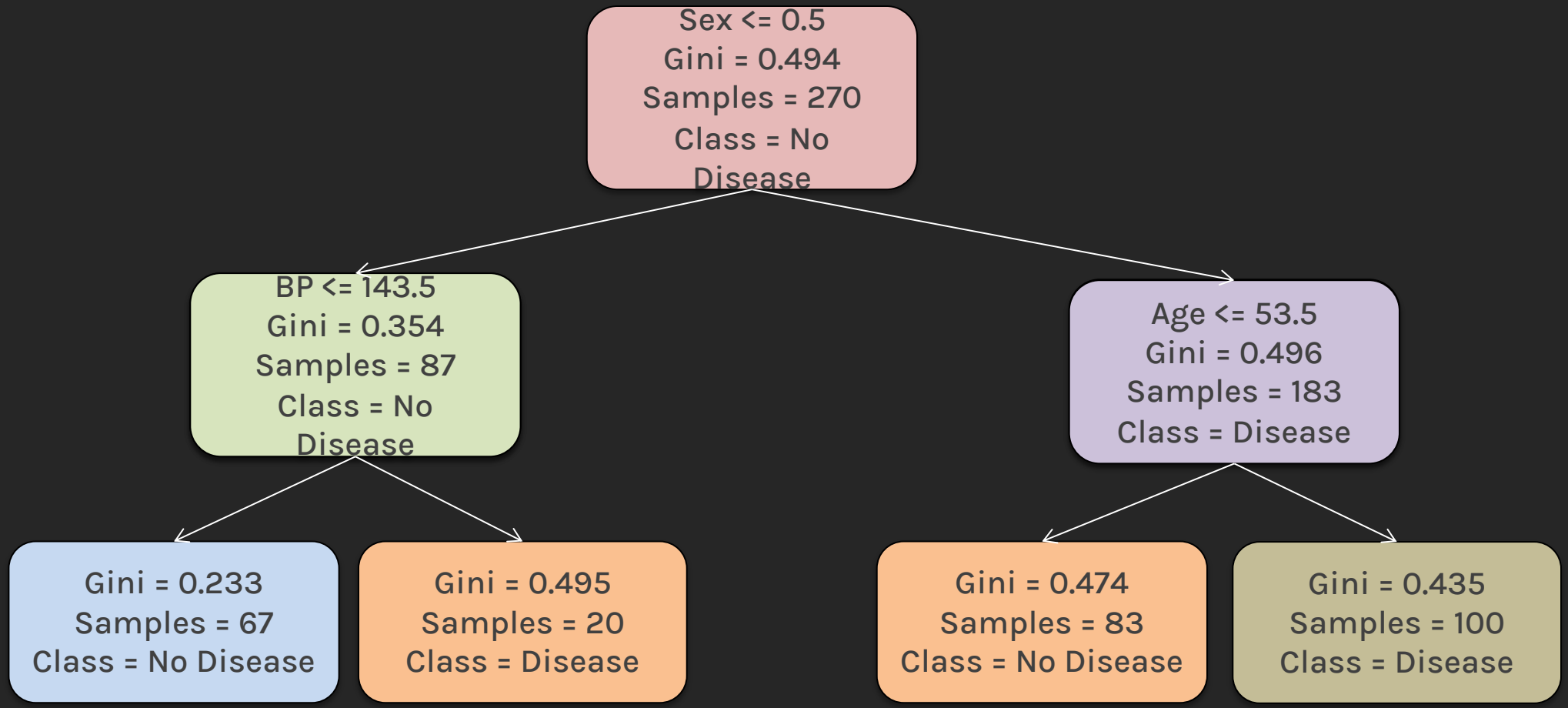
Example 1: Level-Order

Consider the following decision tree with max_depth=2 that predicts if a person has heart disease based on age, sex, BP and cholesterol:

Gini = 0.494
Samples = 270
Class = No Disease

Example 1: Level-Order

Consider the following decision tree with max_depth=2 that predicts if a person has heart disease based on age, sex, BP and cholesterol:



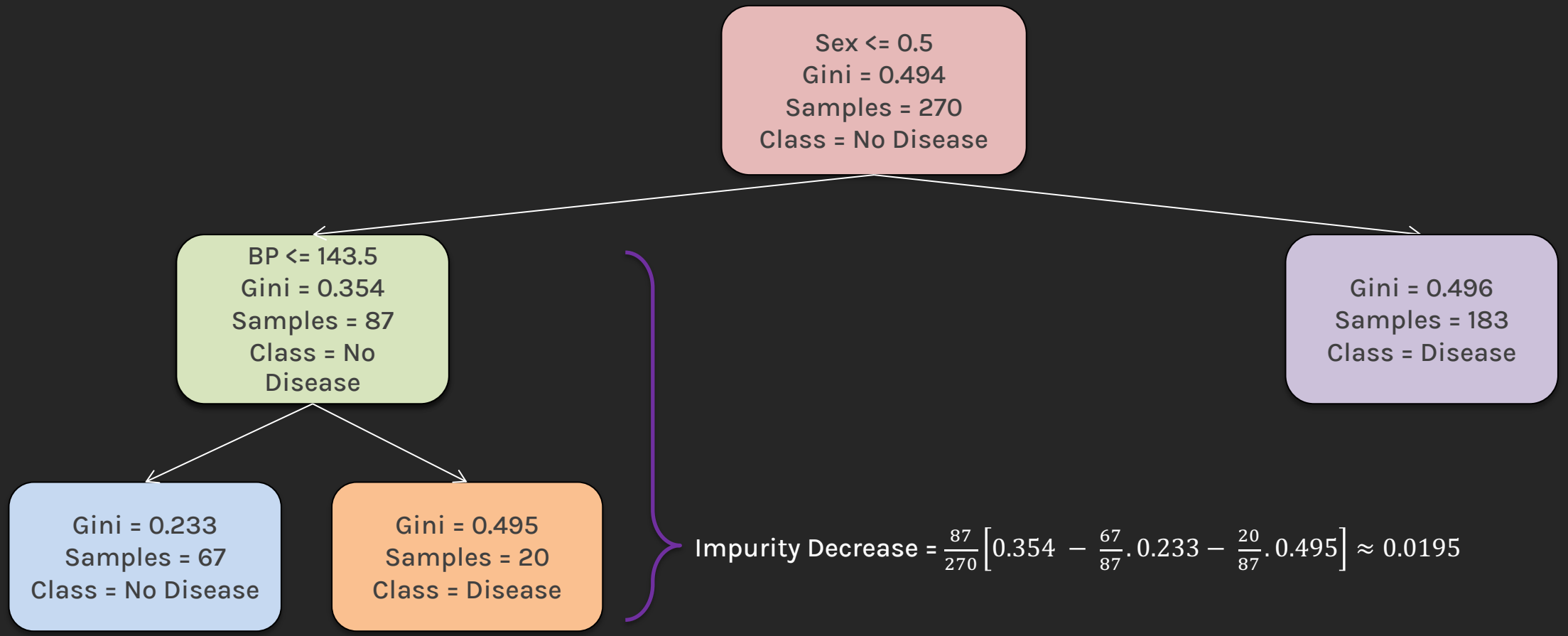
Example 1: Best-first growth

Sklearn determines the best split based on **impurity decrease**. The resulting tree will be the same when fully grown, just the order in which it is built is different.

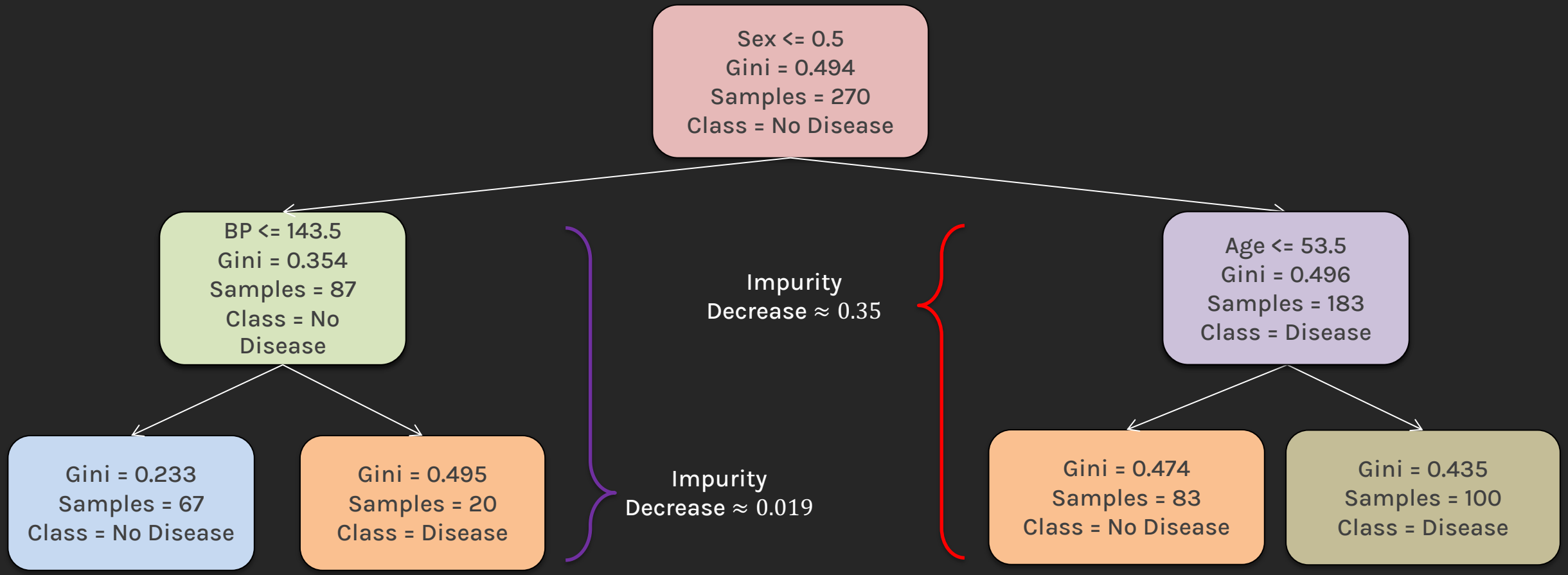
Gini = 0.494
Samples = 270
Class = No Disease

Example 1: Best-first growth

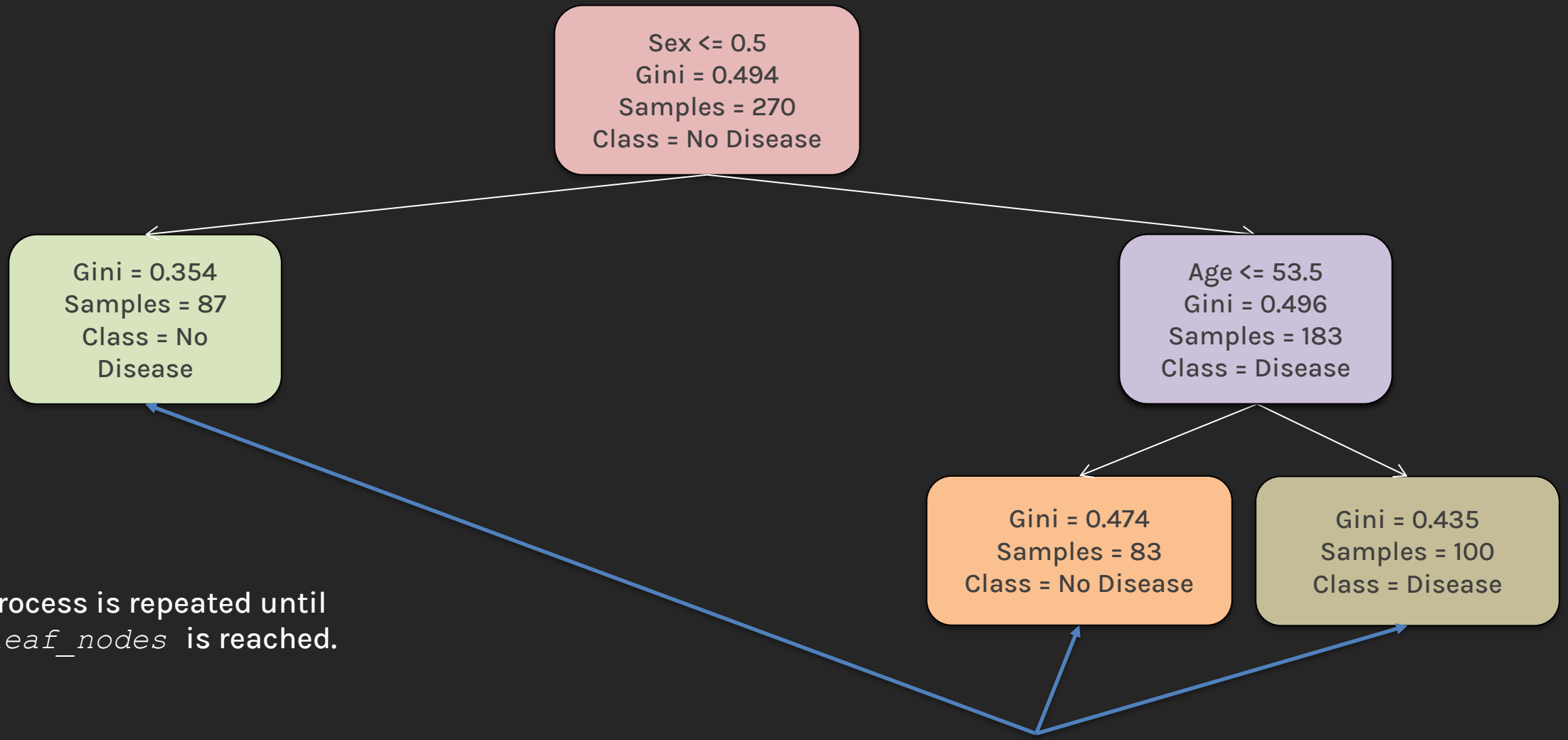
Sklearn determines the best split based on **impurity decrease.** The resulting tree will be the same when fully grown, just the order in which it is built is different.



Example 1: Best-first growth



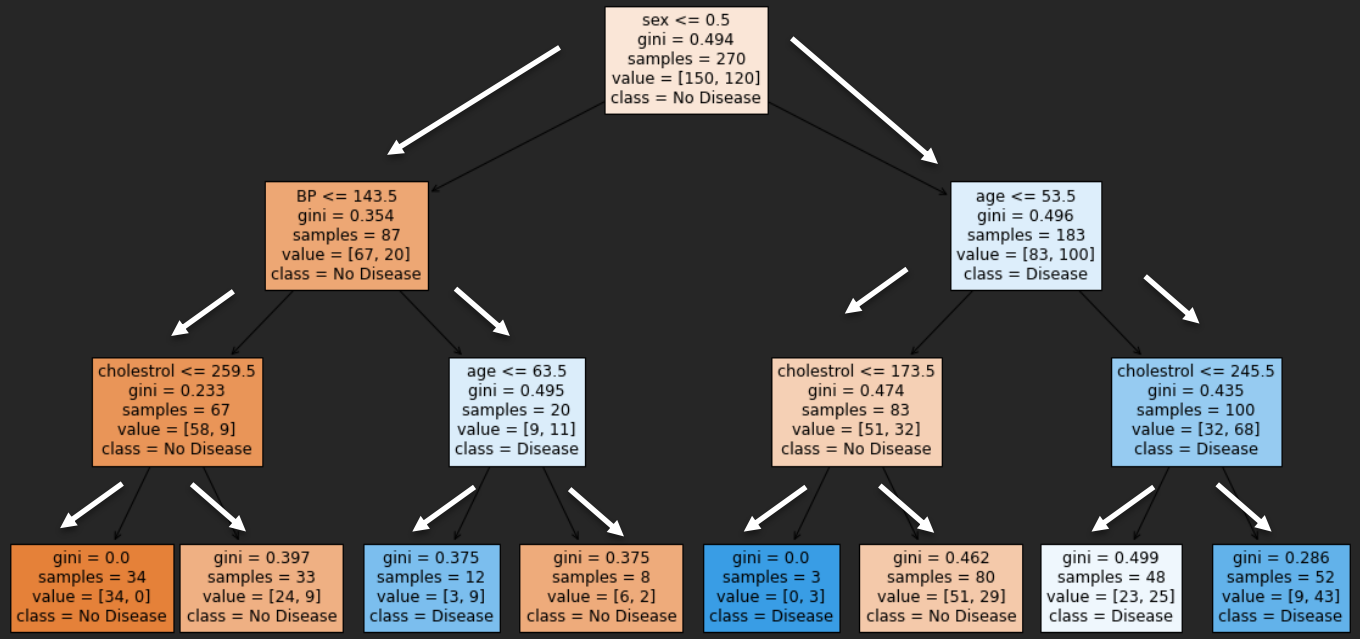
Example 1: Best-first growth



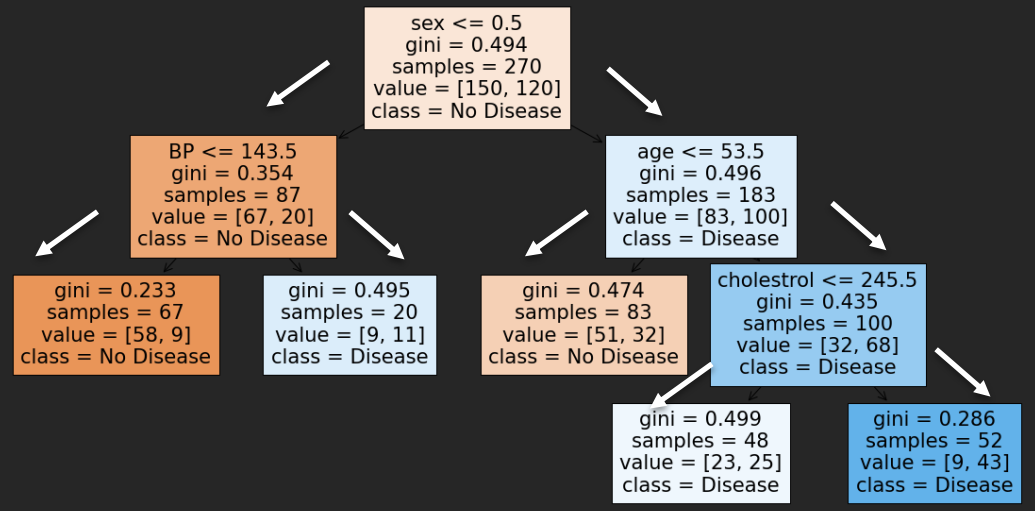
This process is repeated until *max_leaf_nodes* is reached.

Now we compare the impurity decrease when splitting these 3 nodes!

Example 2: Level-order vs Best-first growth



max_depth = 3



max_leaf_nodes = 5

Stopping Conditions

A more restrictive stopping condition is:

Compute the gain in purity of splitting a region R into R_1 and R_2:

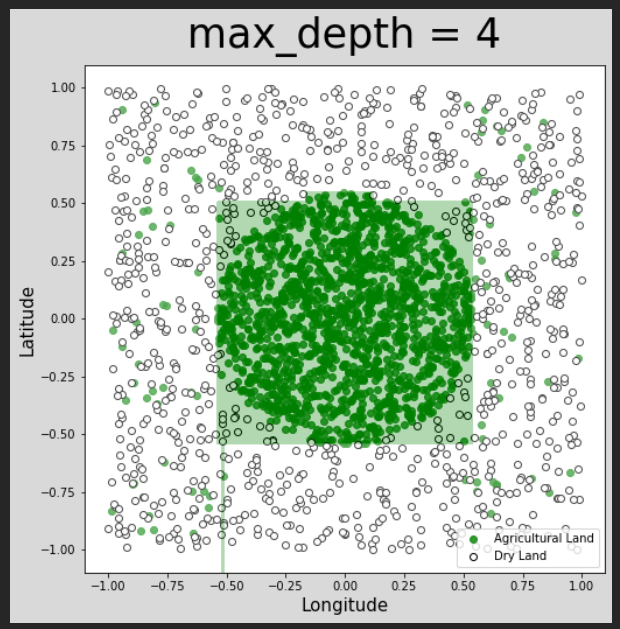
$$Gain(R) = \Delta(R) = m(R) - \frac{N_1}{N} m(R_1) - \frac{N_2}{N} m(R_2)$$

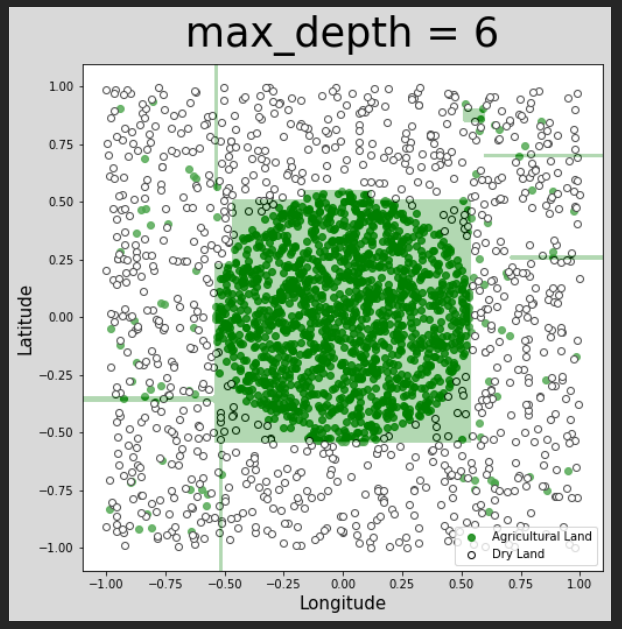
Classification Error/Gini Index/Entropy

Don't split if the gain is less than some pre-defined threshold (`min_impurity_decrease`).

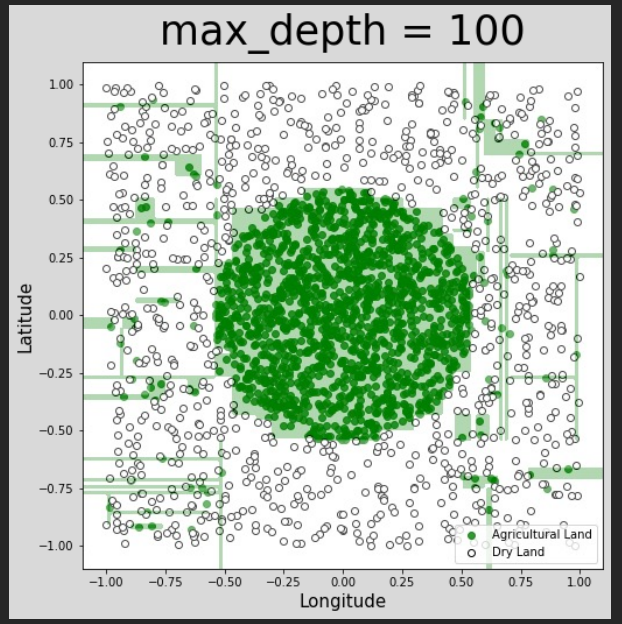
How do we decide what is the appropriate stopping condition or stopping method?

Variance vs Bias

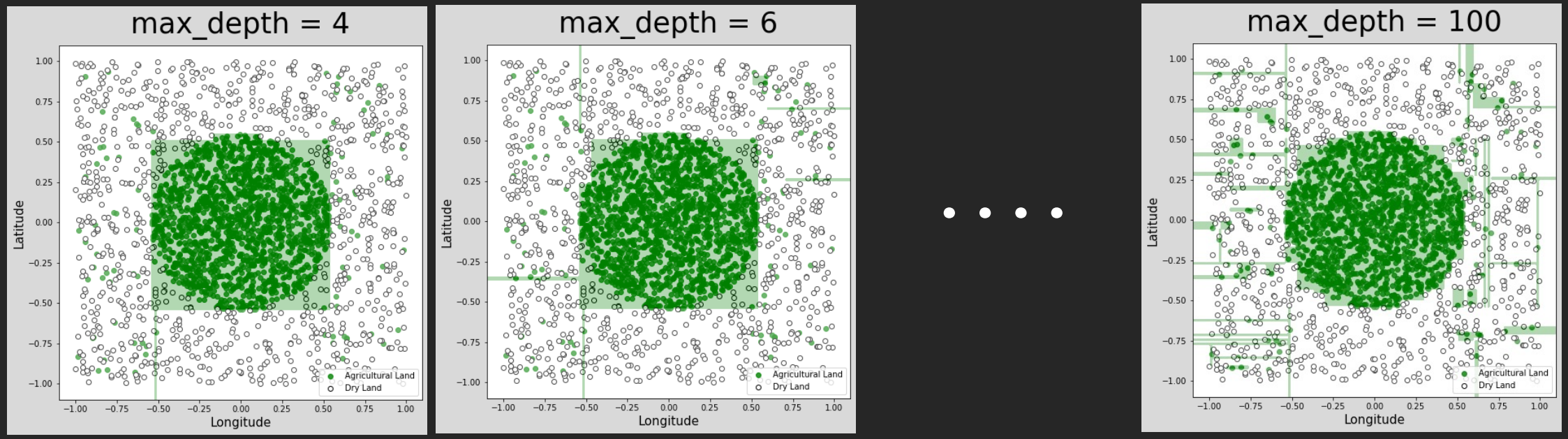




• • • •

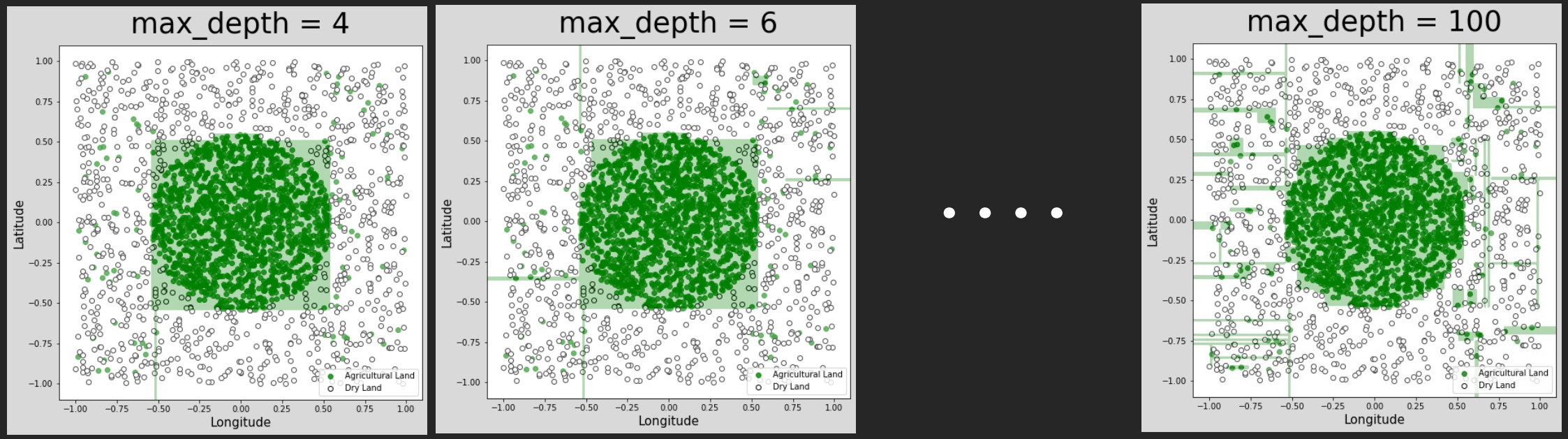


Variance vs Bias



Bias decreases (can overfit)

Variance vs Bias

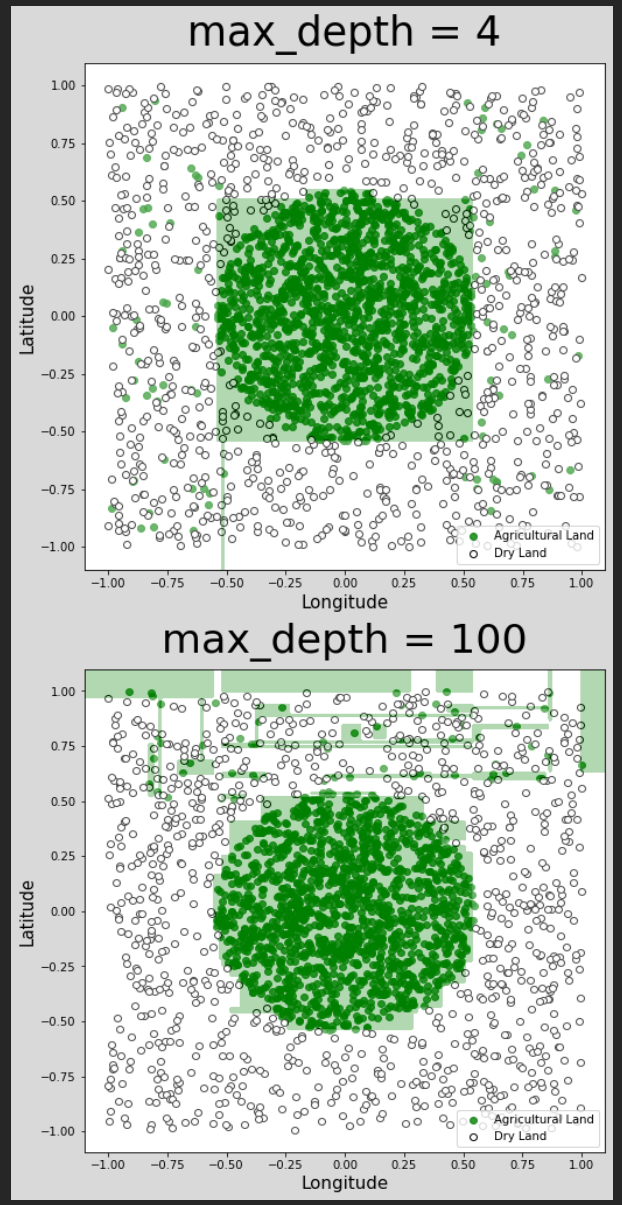


Bias decreases (can overfit)

Variance decreases (can underfit)

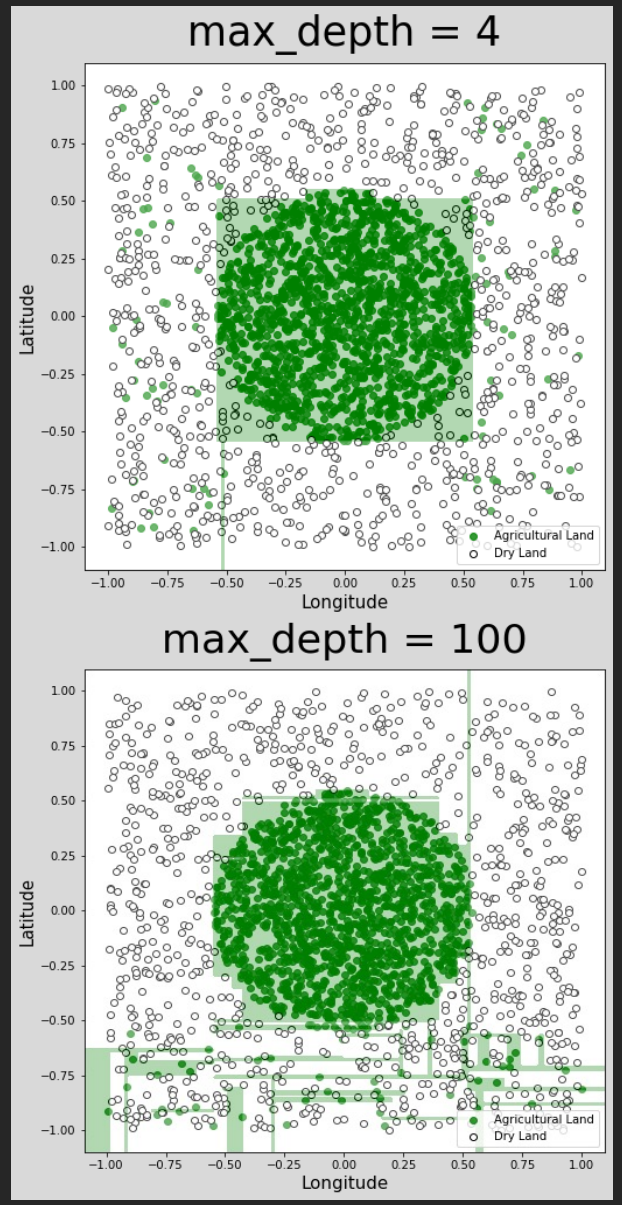
Complex trees are also harder to interpret and more computationally expensive to train.

Variance vs Bias



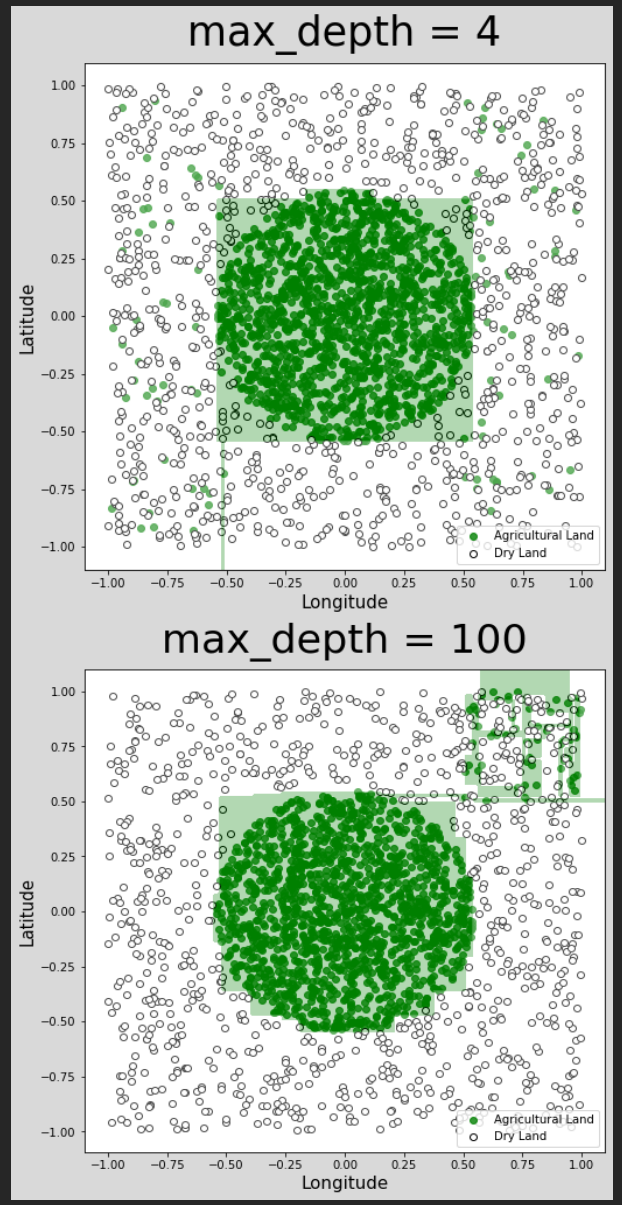
- **High Bias:** Trees of low depth are not a good fit for the training data - it's unable to capture the nonlinear boundary separating the two classes.
- **Low Variance:** Trees of low depth are robust to slight perturbations in the training data - the square carved out by the model is stable if you move the boundary points a bit.
- **Low Bias:** With a high depth, we can obtain a model that correctly classifies all points on the boundary (by zig-zagging around each point).

Variance vs Bias

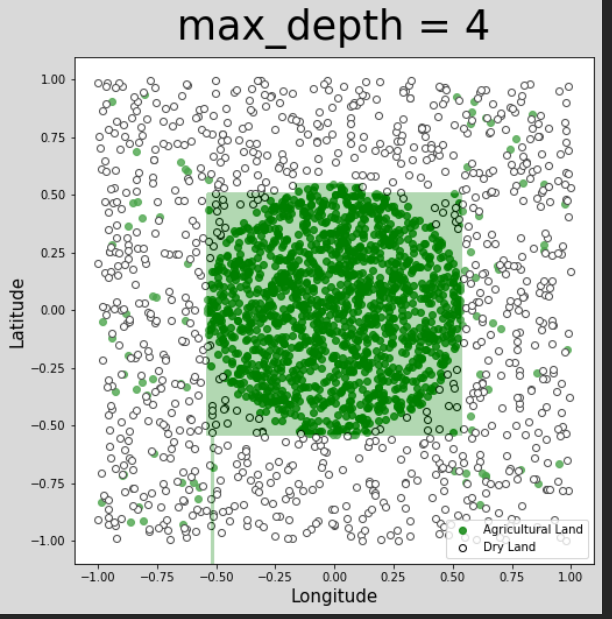


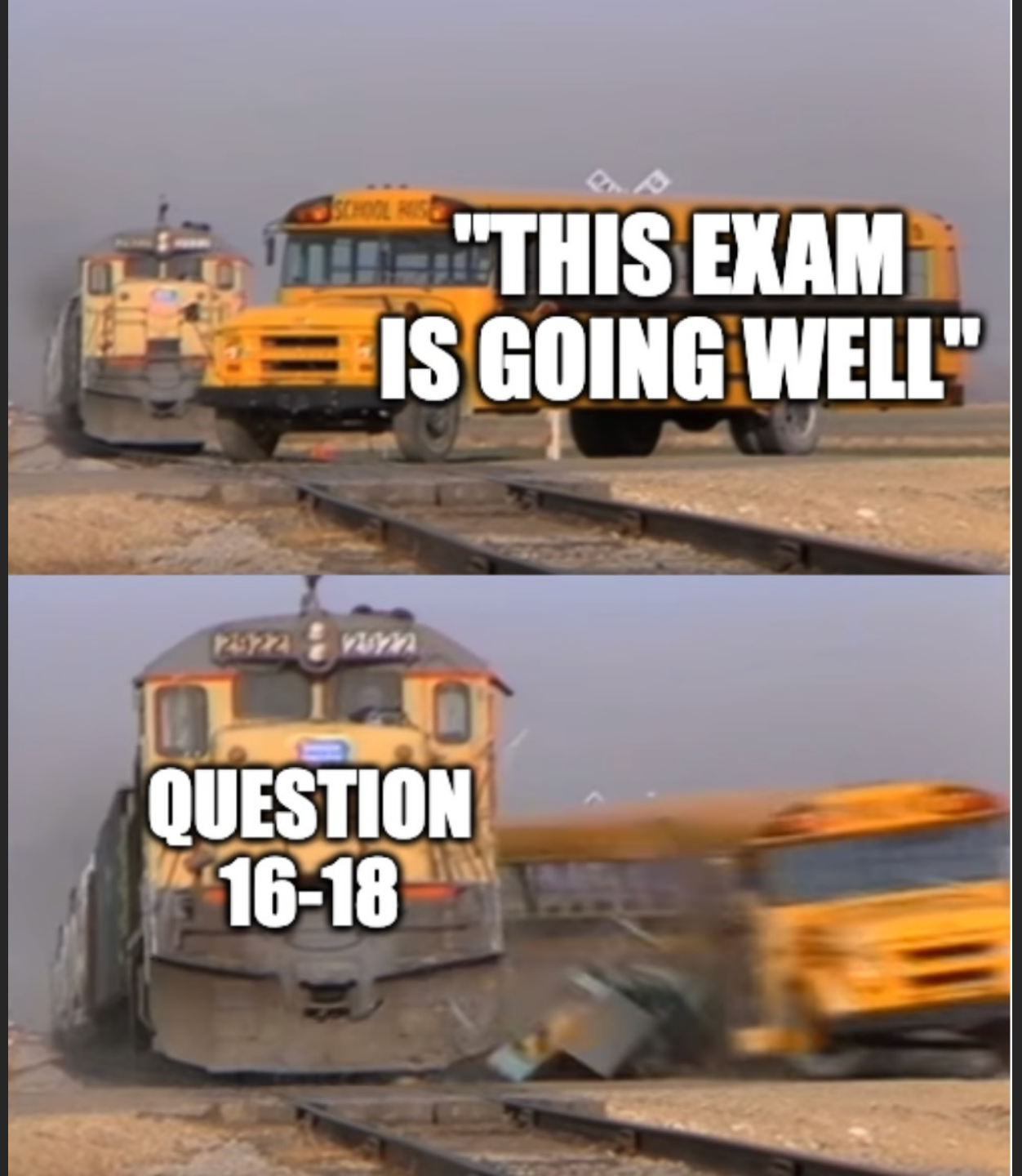
- **High Bias:** Trees of low depth are not a good fit for the training data - it's unable to capture the nonlinear boundary separating the two classes.
- **Low Variance:** Trees of low depth are robust to slight perturbations in the training data - the square carved out by the model is stable if you move the boundary points a bit.
- **Low Bias:** With a high depth, we can obtain a model that correctly classifies all points on the boundary (by zig-zagging around each point).

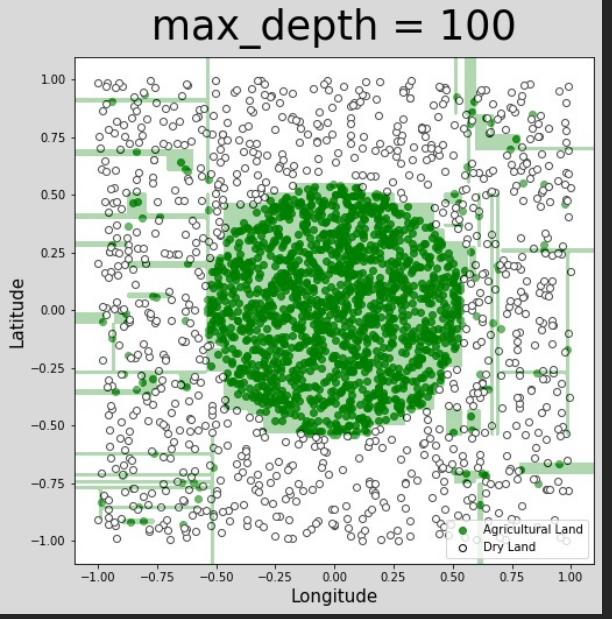
Variance vs Bias

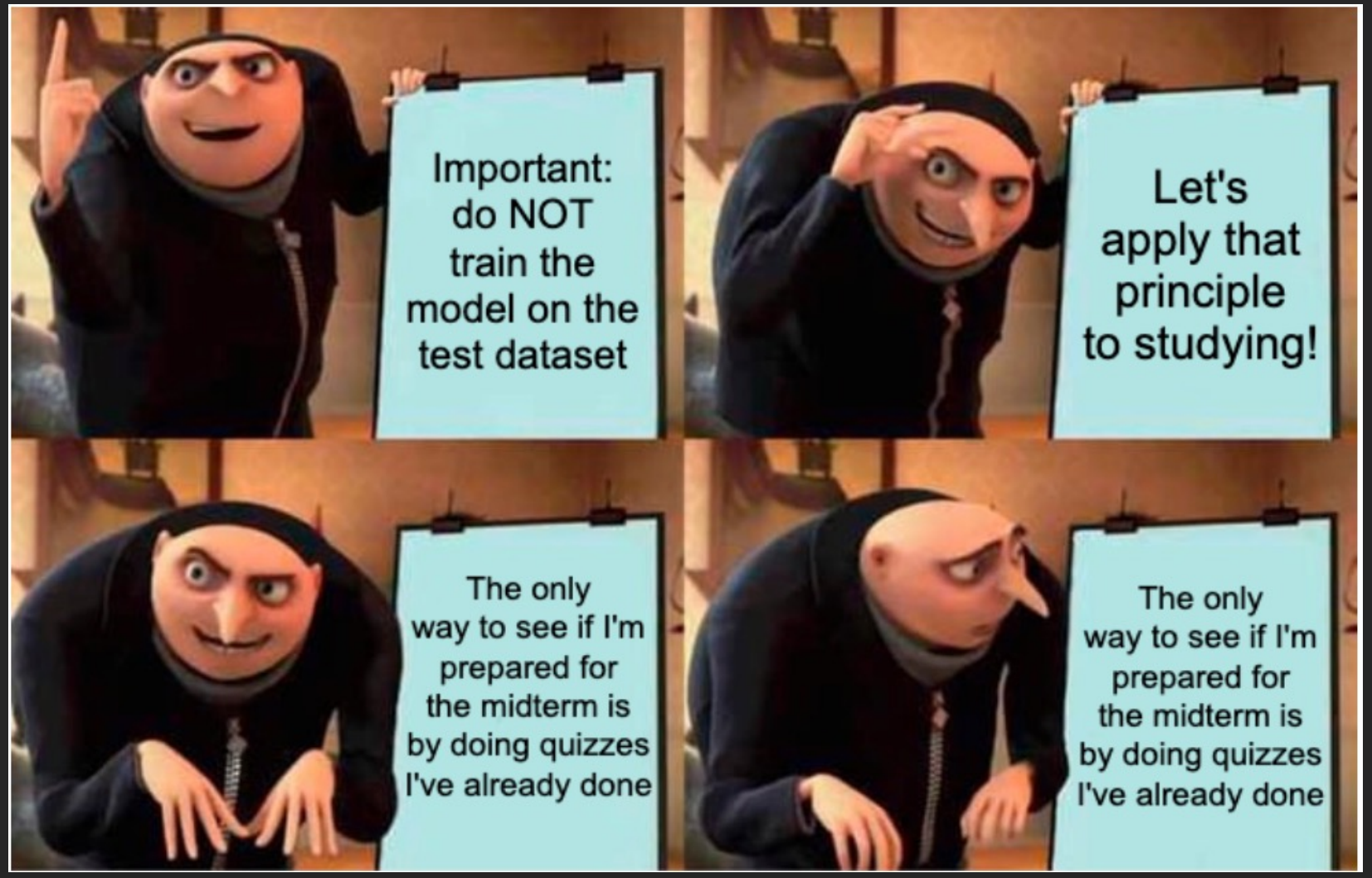


- **High Bias:** Trees of low depth are not a good fit for the training data - it's unable to capture the nonlinear boundary separating the two classes.
- **Low Variance:** Trees of low depth are robust to slight perturbations in the training data - the square carved out by the model is stable if you move the boundary points a bit.
- **Low Bias:** With a high depth, we can obtain a model that correctly classifies all points on the boundary (by zig-zagging around each point).
- **High Variance:** Trees of high depth are sensitive to perturbations in the training data, especially to changes in the boundary points.


max_depth = 4
Latitude
Longitude
Agricultural Land
Dry Land


"THIS EXAM
IS GOING WELL"
QUESTION
16-18


max_depth = 100
Latitude
Longitude
Agricultural Land
Dry Land


Important: do NOT train the model on the test dataset
Let's apply that principle to studying!
The only way to see if I'm prepared for the midterm is by doing quizzes I've already done
The only way to see if I'm prepared for the midterm is by doing quizzes I've already done

Stopping Conditions

max_depth | *min_samples_leaf* | *max_leaf_nodes* | min_impurity_decrease

How can we determine the appropriate hyperparameters?

cross-validation

Summary

Explain the concept of overfitting in decision trees. How can stopping conditions help prevent overfitting?

Overfitting occurs when the decision tree becomes excessively complex, memorizing the training data but failing to generalize to new data. Stopping conditions limit tree growth, preventing it from becoming too specific to the training data.

Describe two common stopping conditions used in decision tree learning and explain how they limit tree growth.

Two examples are max_depth, which limits the maximum depth of the tree, and min_samples_leaf, which sets a minimum number of samples required to form a leaf node. Both prevent the tree from growing too deep and overfitting.

Differentiate between level-order and best-first growth in decision trees. When would you prefer one method over the other?

Level-order growth builds the tree level by level, while best-first prioritizes splits with the highest impurity decrease. Best-first is advantageous when seeking the most informative splits early on, potentially leading to smaller trees with good performance.

Summary

Explain the trade-off between bias and variance in decision trees. How does tree depth impact bias and variance?

Shallow trees have high bias (simplistic model) and low variance (robust to data fluctuations). Deep trees have low bias (complex model) but high variance (sensitive to data changes). The optimal depth balances this trade-off.

What is the role of cross-validation in determining the optimal hyperparameters for a decision tree model?

Cross-validation helps estimate a model's performance on unseen data. By evaluating different hyperparameter combinations (e.g., tree depth, stopping criteria) using cross-validation, we can select those that lead to the best generalization performance.

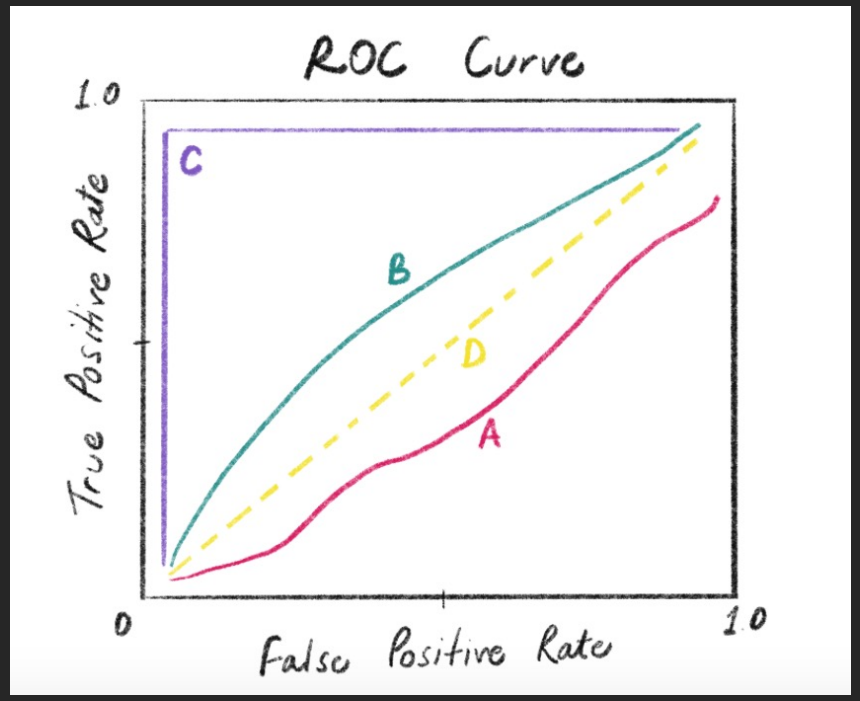
Game time

Consider the ROC plot below. A, B, C, and D represent 4 different binary classification models. Arrange the model names such that they correspond to the sequence of statements below:

Put them in order of the best to worst model

Options

1. A, B, C, D
2. C, B, A, D
3. C, A, B, D
4. B, A, C, D



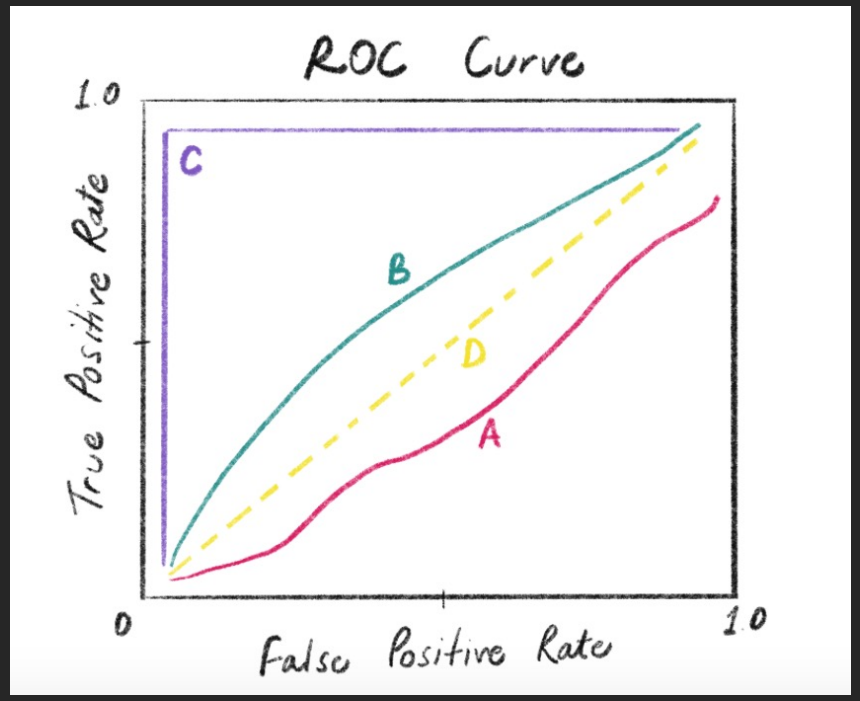
Game time

Consider the ROC plot below. A, B, C, and D represent 4 different binary classification models. Arrange the model names such that they correspond to the sequence of statements below:

Put them in order of the best to worst model

Options

1. A, B, C, D
2. C, B, A, D
3. C, A, B, D
4. B, A, C, D



Thank you