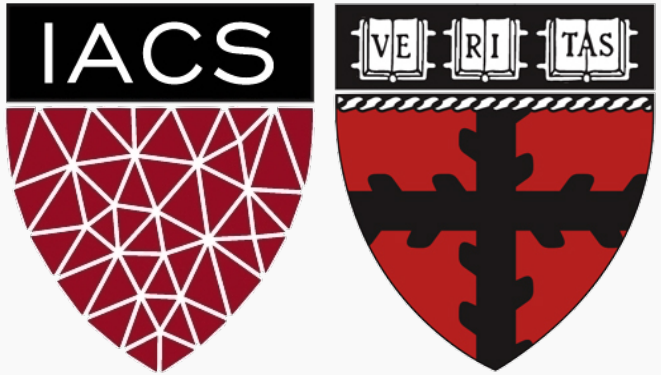
Bagging

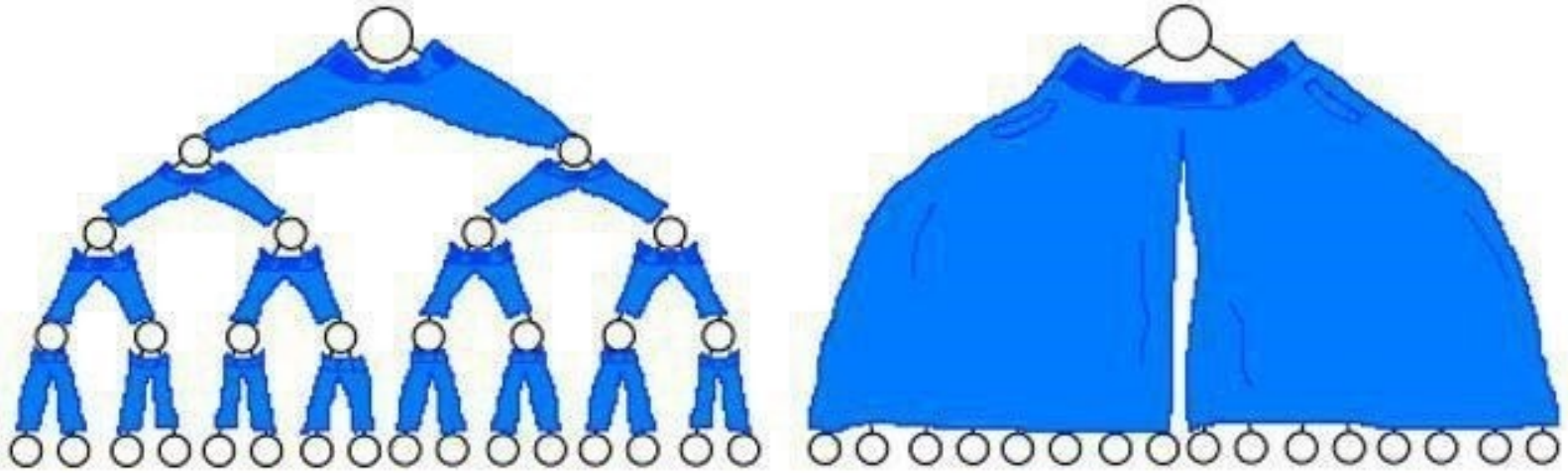
CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai



If a binary tree wore pants would he wear them

like this or like this?



Outline

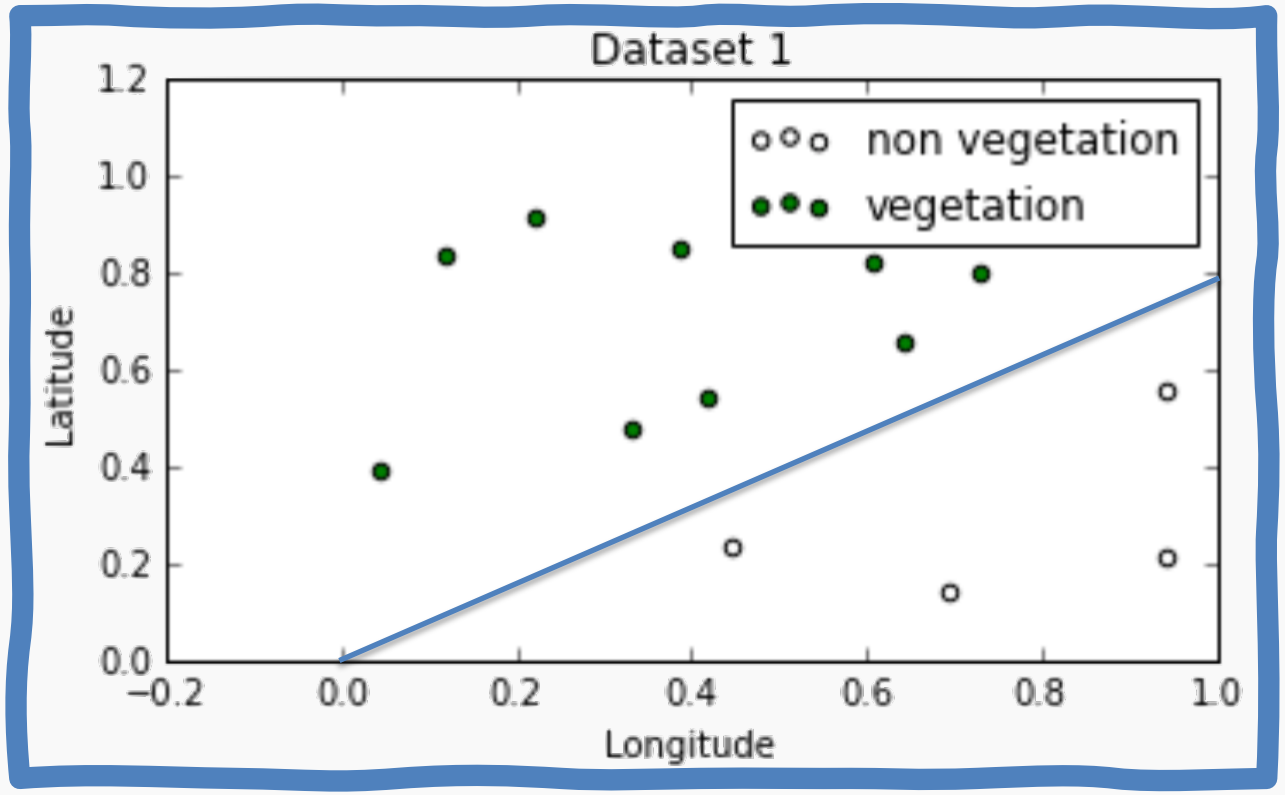
- Review of Decision Trees
- Bagging
- Out of Bag Error (OOB)
- Variable Importance

Outline

- **Review of Decision Trees**
- Bagging
- Out of Bag Error (OOB)
- Variable Importance

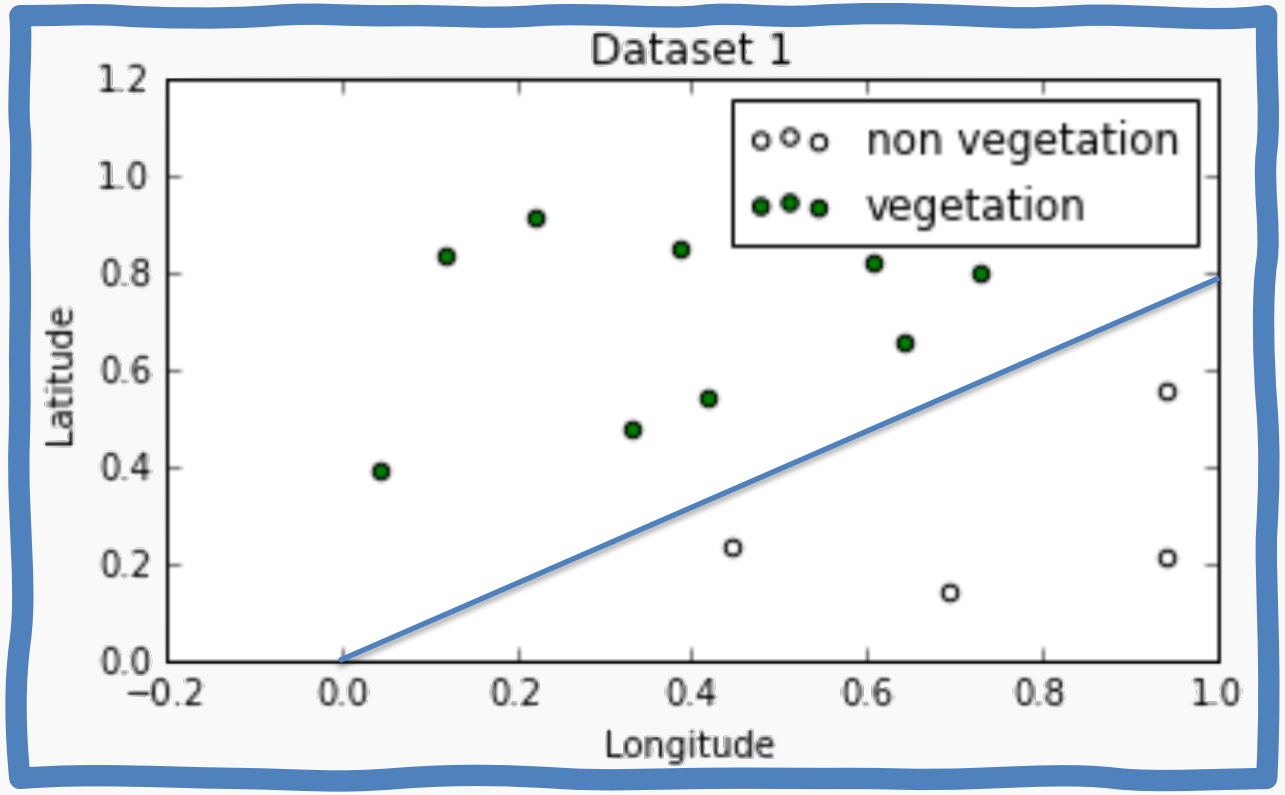
Geometry of Data

Question: Can you guess the equation that defines the decision boundary below?



Geometry of Data

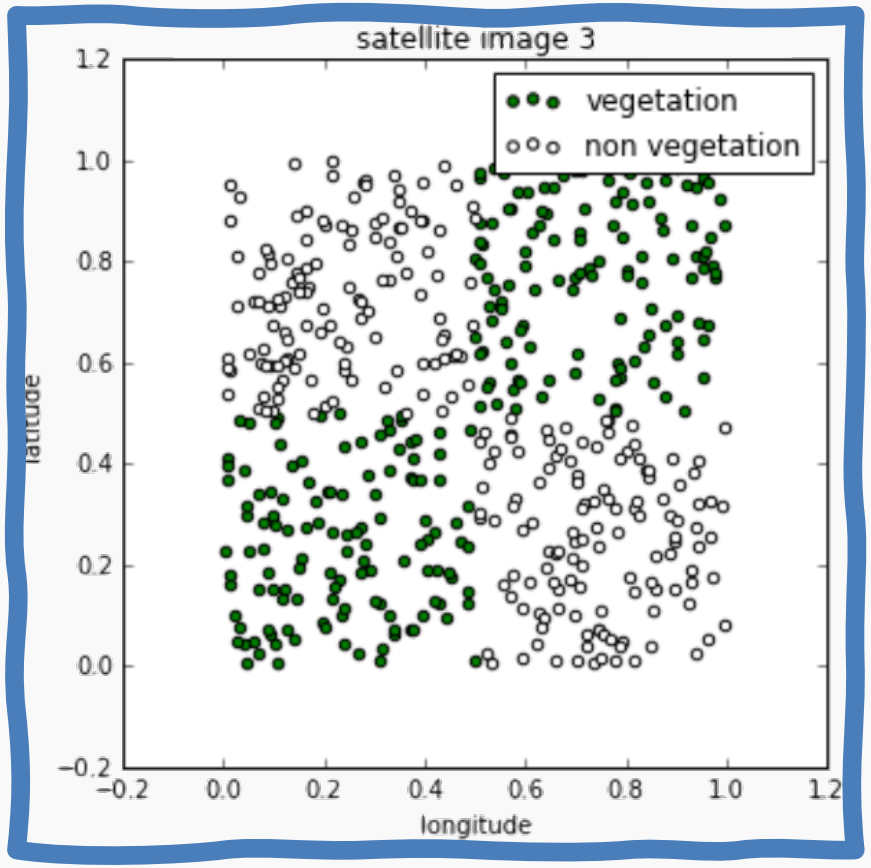
Question: Can you guess the equation that defines the decision boundary below?



$$-0.8x_1 + x_2 = 0 \implies x_2 = 0.8x_1 \Rightarrow \ Latitude = 0.8\ Lon$$

Geometry of Data

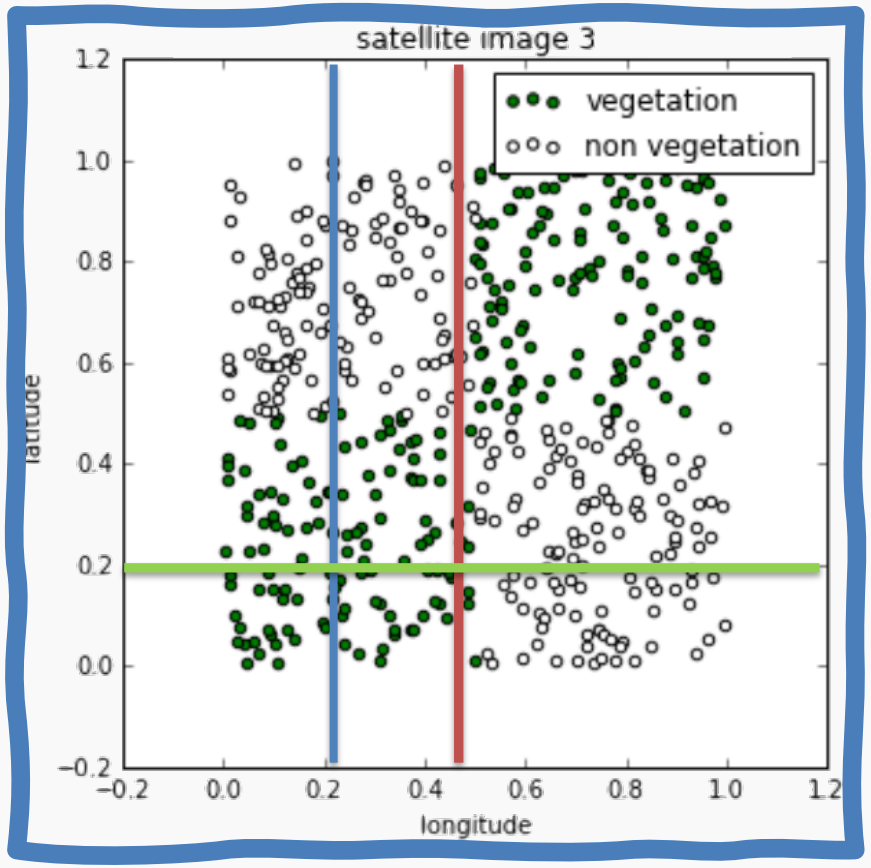
Complicate decision boundaries can not be explained with Log Regression.



Geometry of Data

But can be described using Trees.

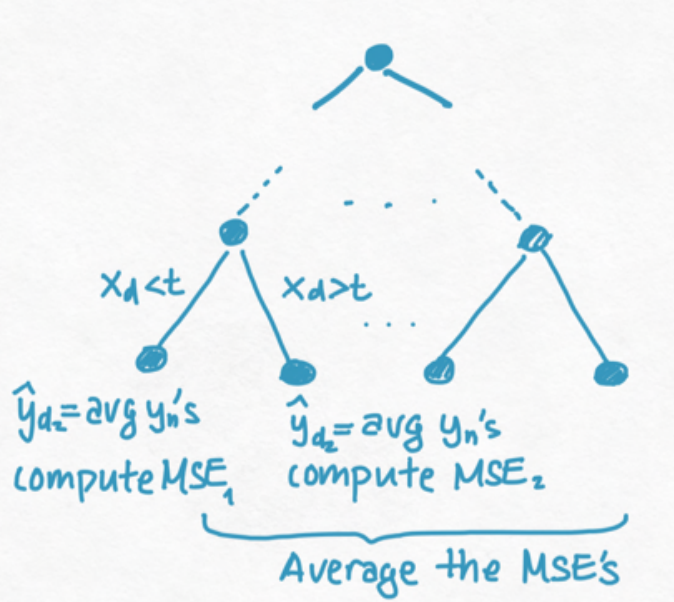
Which one represents the first split of a decision tree?



Decision Trees

To learn a decision tree model, we take a greedy approach:

1. Start with a node containing all the data.
2. If stopping condition is not met:
   A. Choose the 'optimal' predictor and threshold and divide the data in the node into two sets.
3. For each new node, repeat step 2.

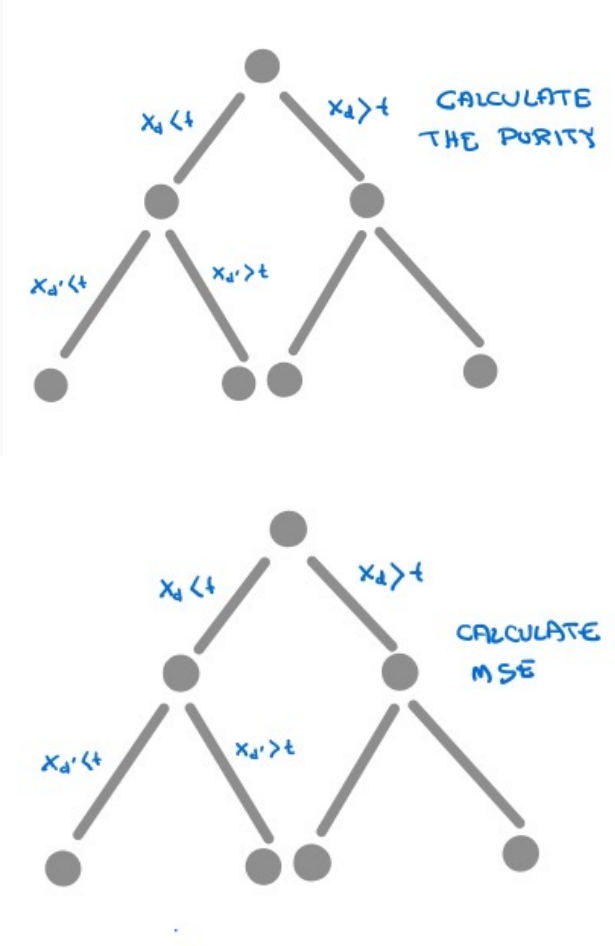


Decision Trees: Splitting Criteria

Splitting Criteria:

For classification, purity of the regions is a good indicator the performance of the model. Entropy as a splitting criterial minimizes the cross-entropy (greedy). Gini is also a splitting criteria.

For regression, we want to select a splitting criterion that promotes splits that improves the predictive accuracy of the model as measured by the MSE

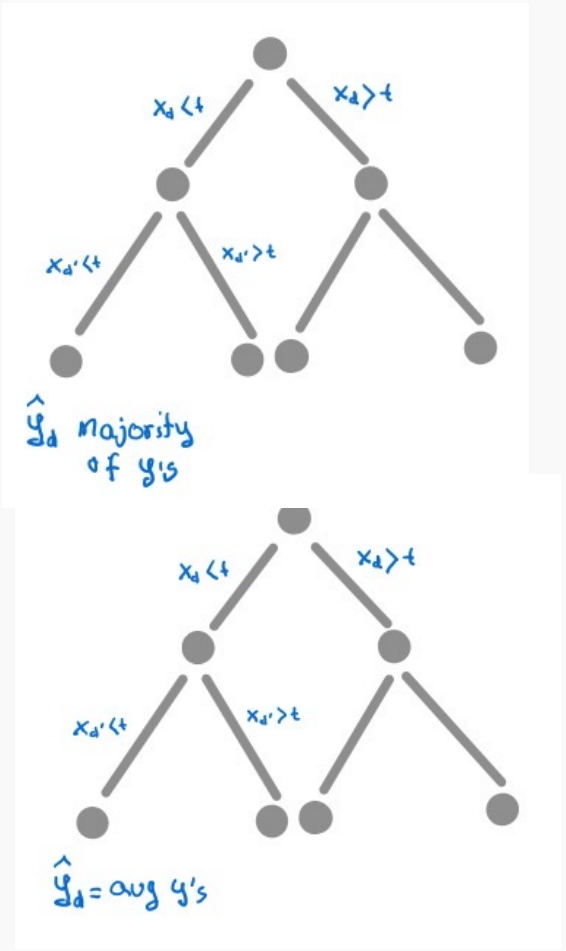


Decision Trees: Prediction

Prediction:

For classification, we label each region in the model with the label of the class to which the **plurality** of the points within the region belong.

For regression, we predict with the **average** of the output values of the training points contained in the region.



Stopping Conditions

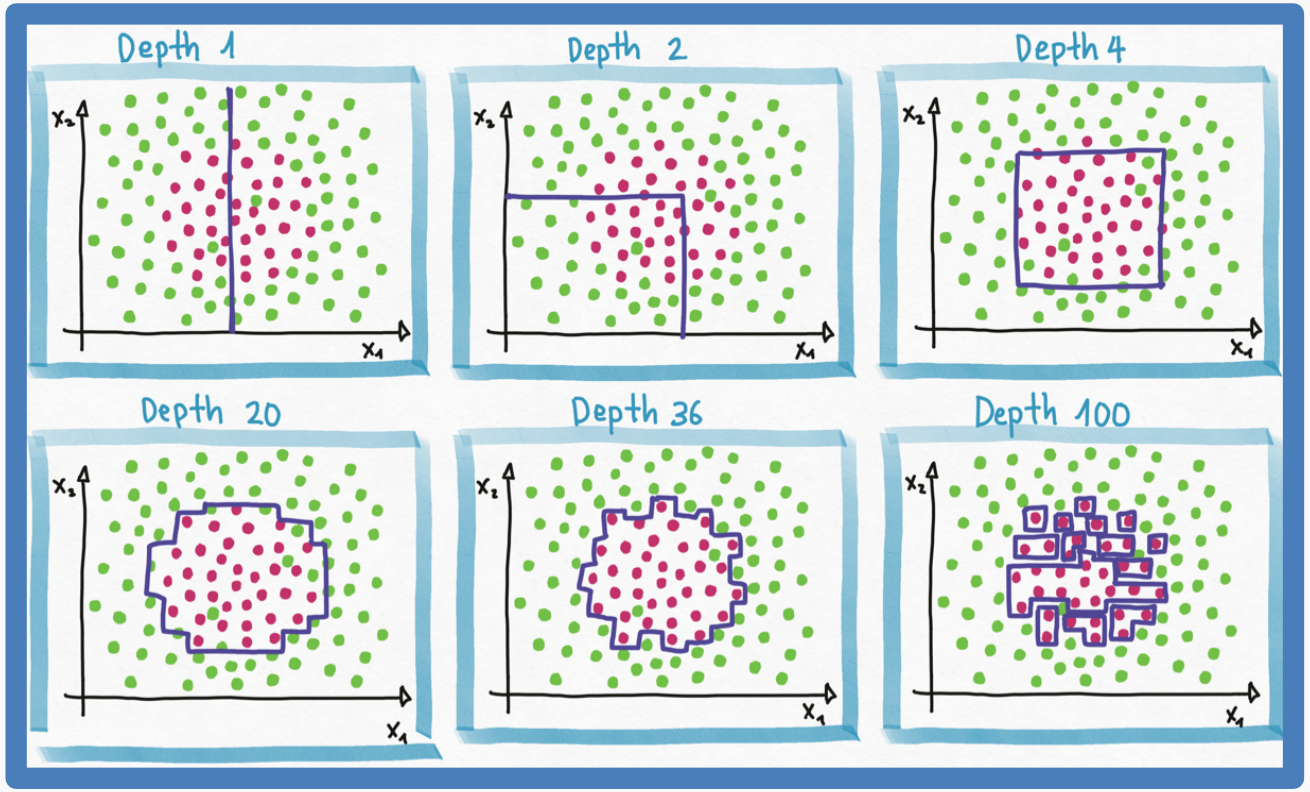
The stopping condition is usually a maximum depth or a minimum MSE.

But others common simple stopping conditions are:

- Don't split a region if all instances in the region belong to the same class.
- Don't split a region if the number of instances in the sub-region will fall below pre-defined threshold (`min_samples_leaf`).
- Don't split a region if the total number of leaves in the tree will exceed pre-defined threshold.
- Don't split if the gain in purity, information, reduction in entropy or MSE of splitting a region R into R_1 and R_2 is less than some pre-defined threshold.

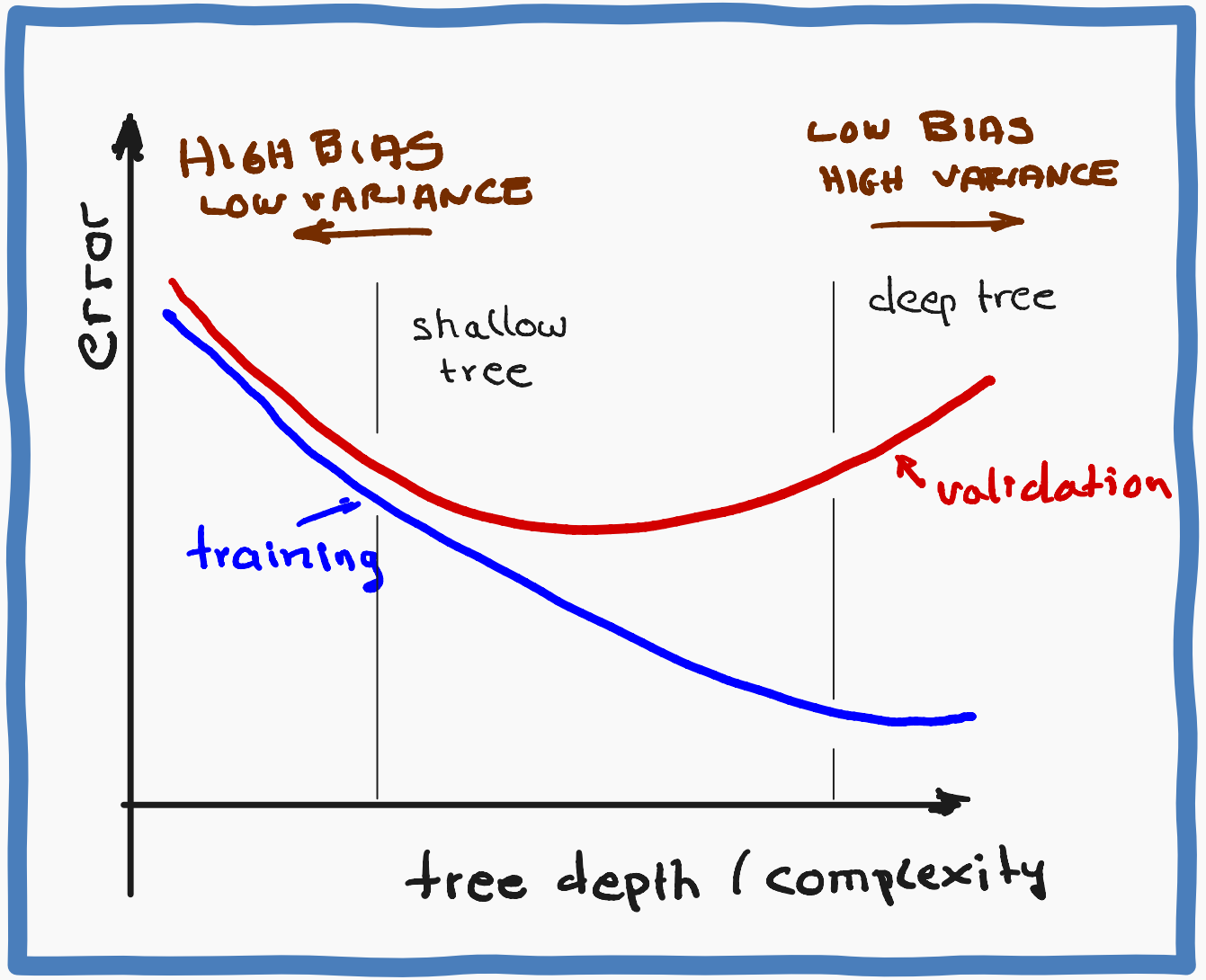
Overfitting

When a tree is too shallow, it cannot divide the input data into enough regions, so the model underfits. When the tree is too deep it cuts the input space into too many regions and fit to the noise of the data -> overfits.

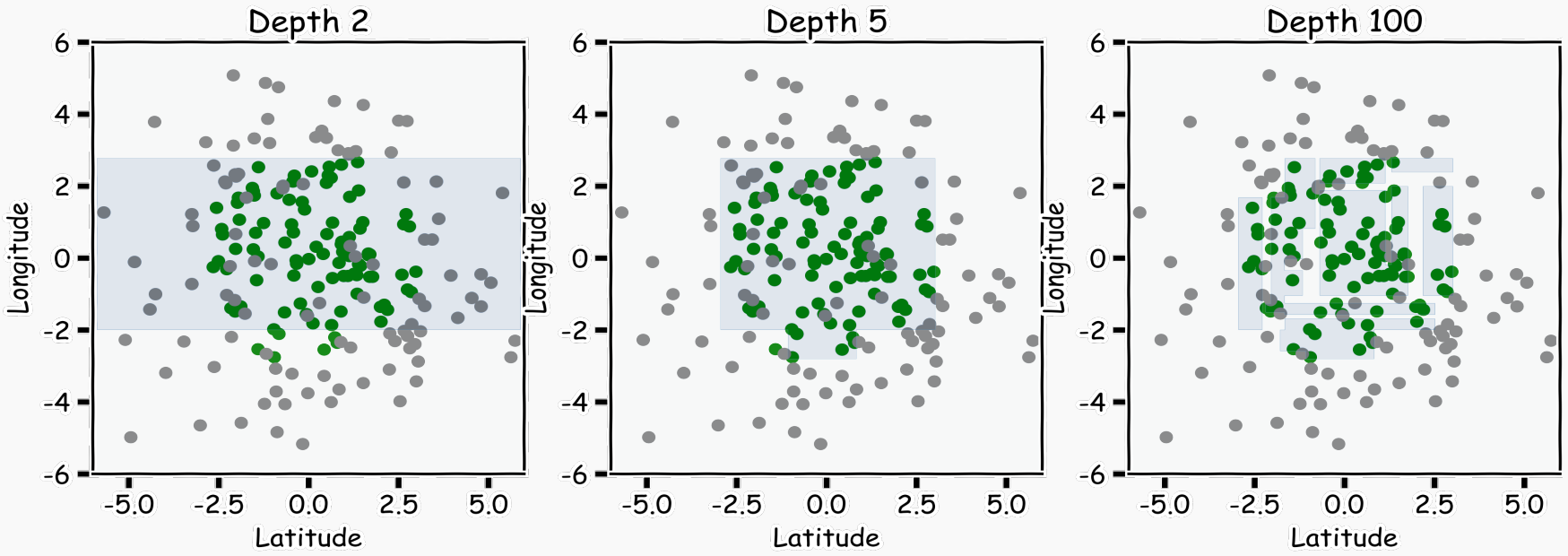


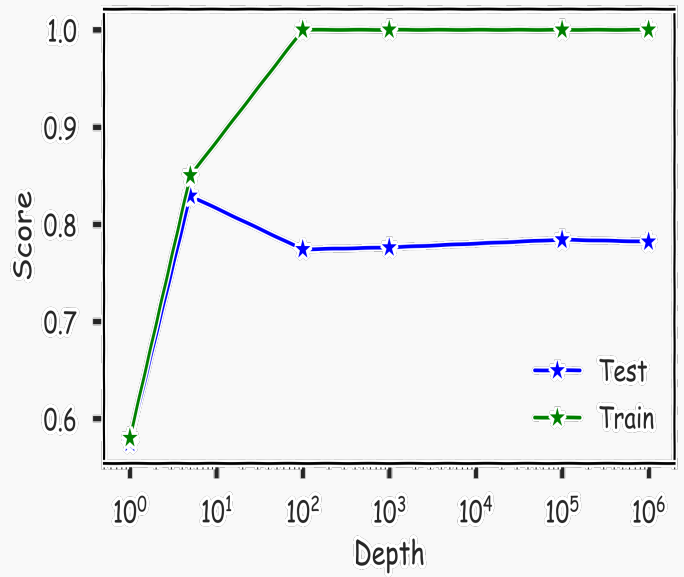
Overfitting

Avoid overfitting by pruning or limiting the depth of the tree and using CV.



Reduce the variance: Depth of the tree





We've seen that large trees have high variance and are prone to overfitting.

Use train/validation or cross validation to estimate the best depth.

Limitations of Decision Tree Models

Decision trees models are highly interpretable and fast to train, using our greedy learning algorithm.

However, to **capture a complex decision boundary** (or approximate a complex function), we need to use a large tree (since each time we can only do axis-aligned splits).

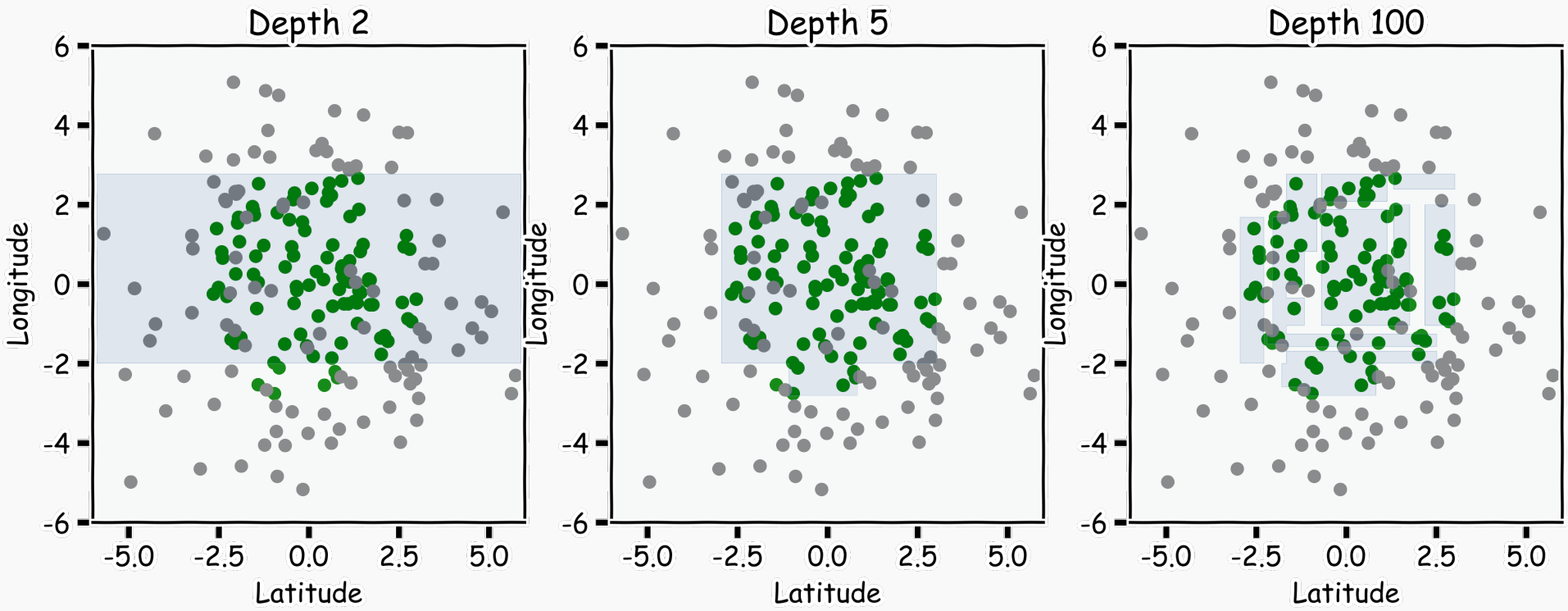
We've seen that large trees have high variance and are prone to overfitting.

For these reasons, in practice, decision tree models often underperform when compared with other classification or regression methods.

Outline

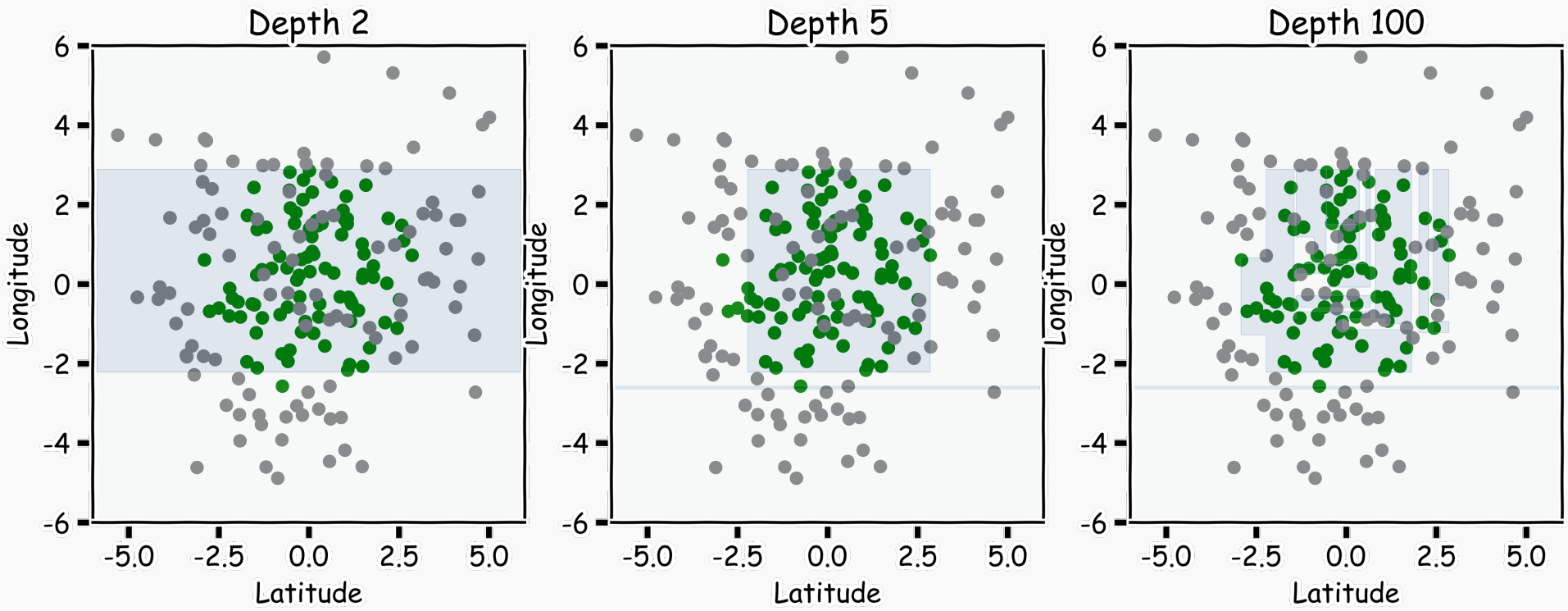
- Review of Decision Trees
- **Bagging**
- Out of Bag Error (OOB)
- Variable Importance

Bagging

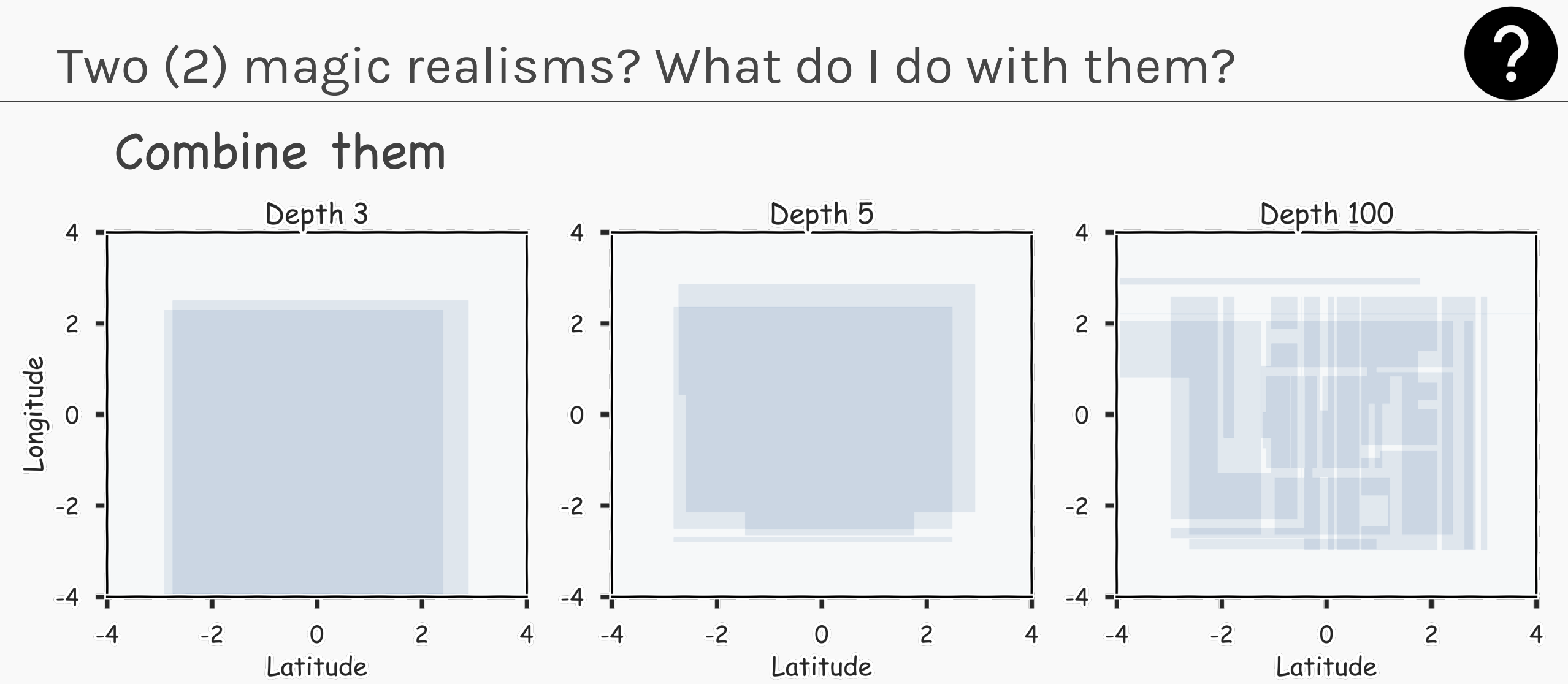

Depth 2
Depth 5
Depth 100
Longitude
Latitude
6
4
2
0
-2
-4
-6
-5.0
-2.5
0.0
2.5
5.0

Bagging

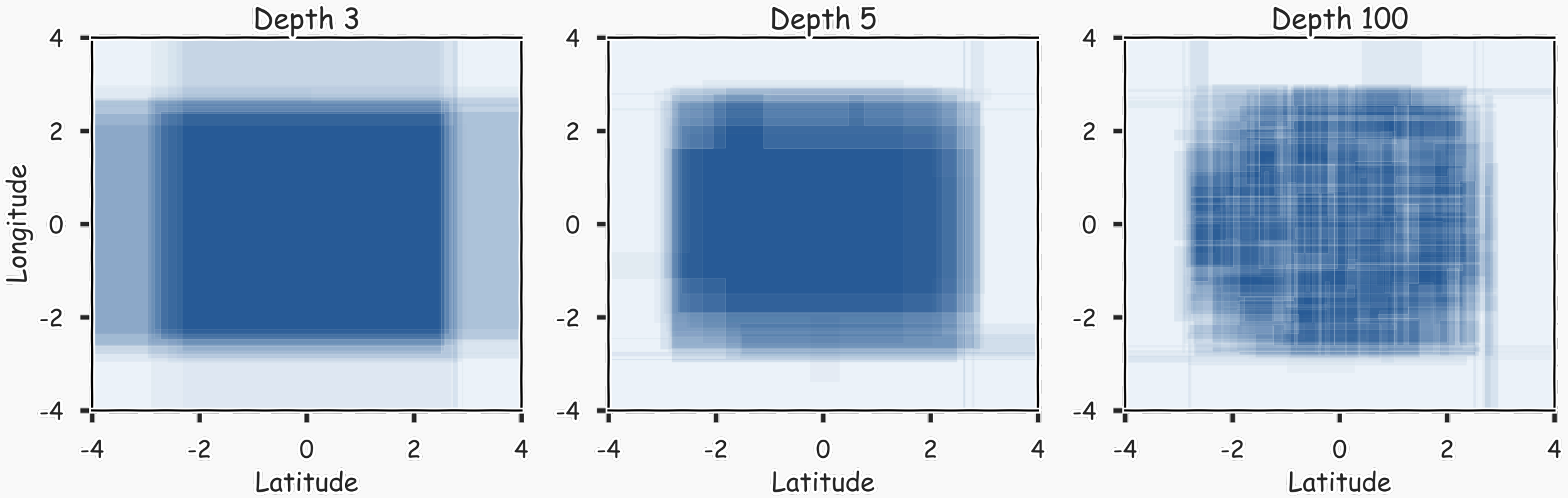
My favorite reality: magic realism -> bootstrap



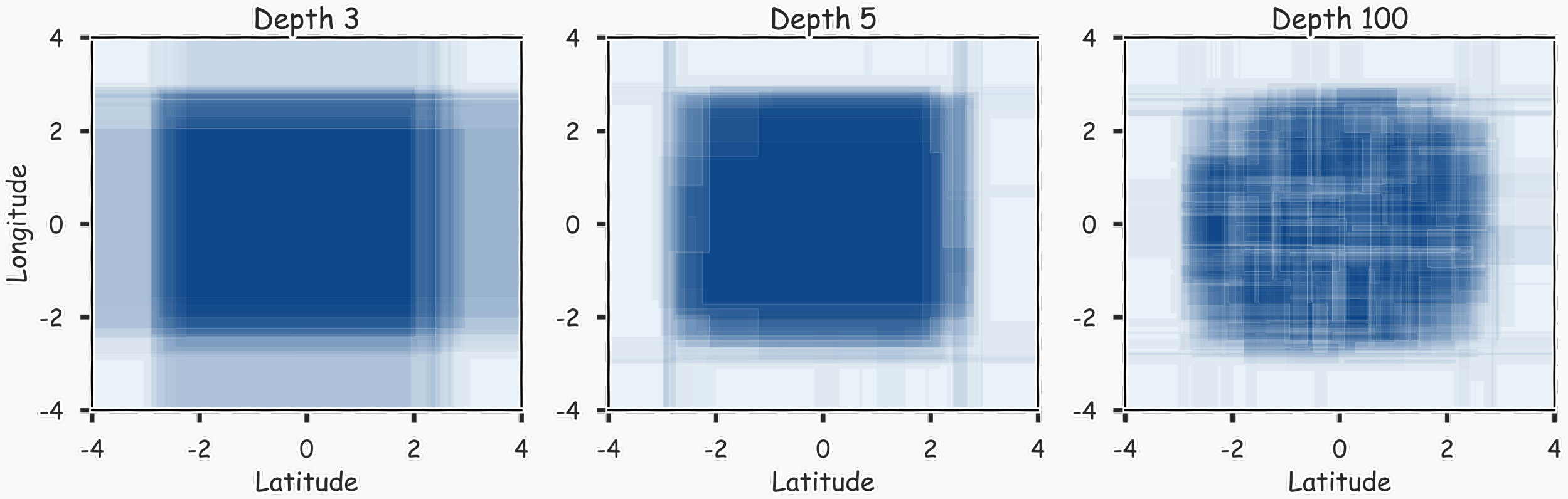
Two (2) magic realisms? What do I do with them?



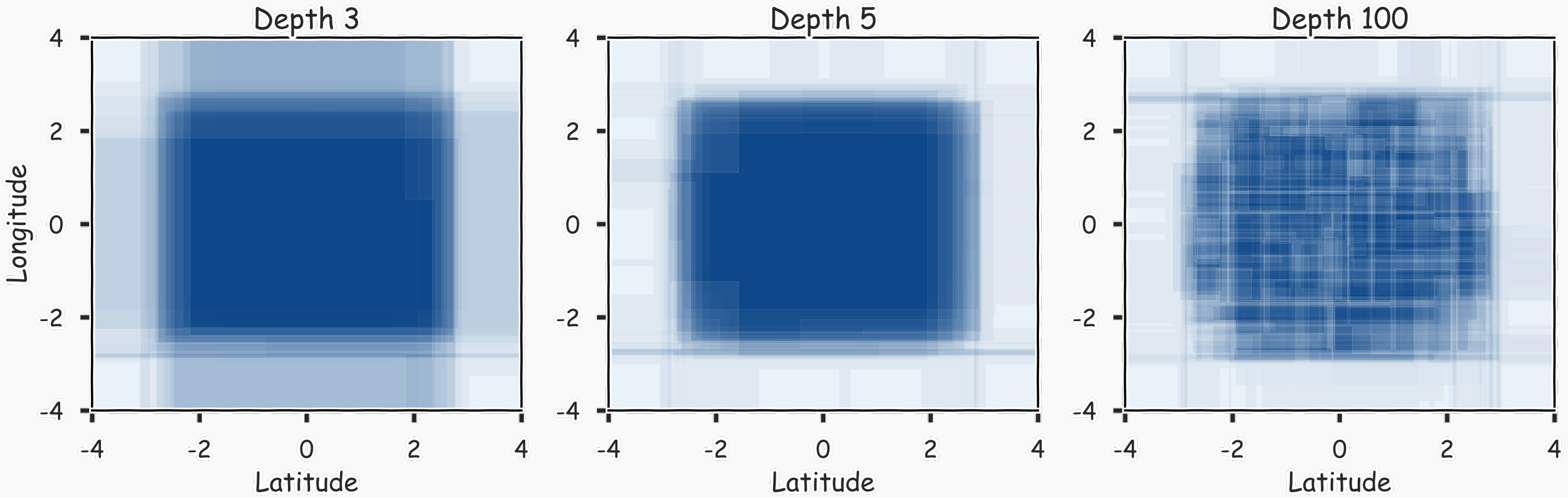
20 magic realisms



100 magic realisms



300 magic realisms



Bagging

One way to adjust for the high variance of the output of an experiment is to perform the experiment multiple times and then average the results.

The same idea can be applied to high variance models:

1. **Bootstrap:** we generate multiple samples of training data, via bootstrapping. We train a deeper decision tree on each sample of data.
2. **Aggregate:** for a given input, we output the averaged outputs of all the models for that input.

This method is called ***Bagging*** (Breiman, 1996), short for, of course, **B**ootstrap **Agg**regating.

For classification, we return the class that is outputted by the plurality of the models. For regression we return the average of the outputs for each tree.

Bagging

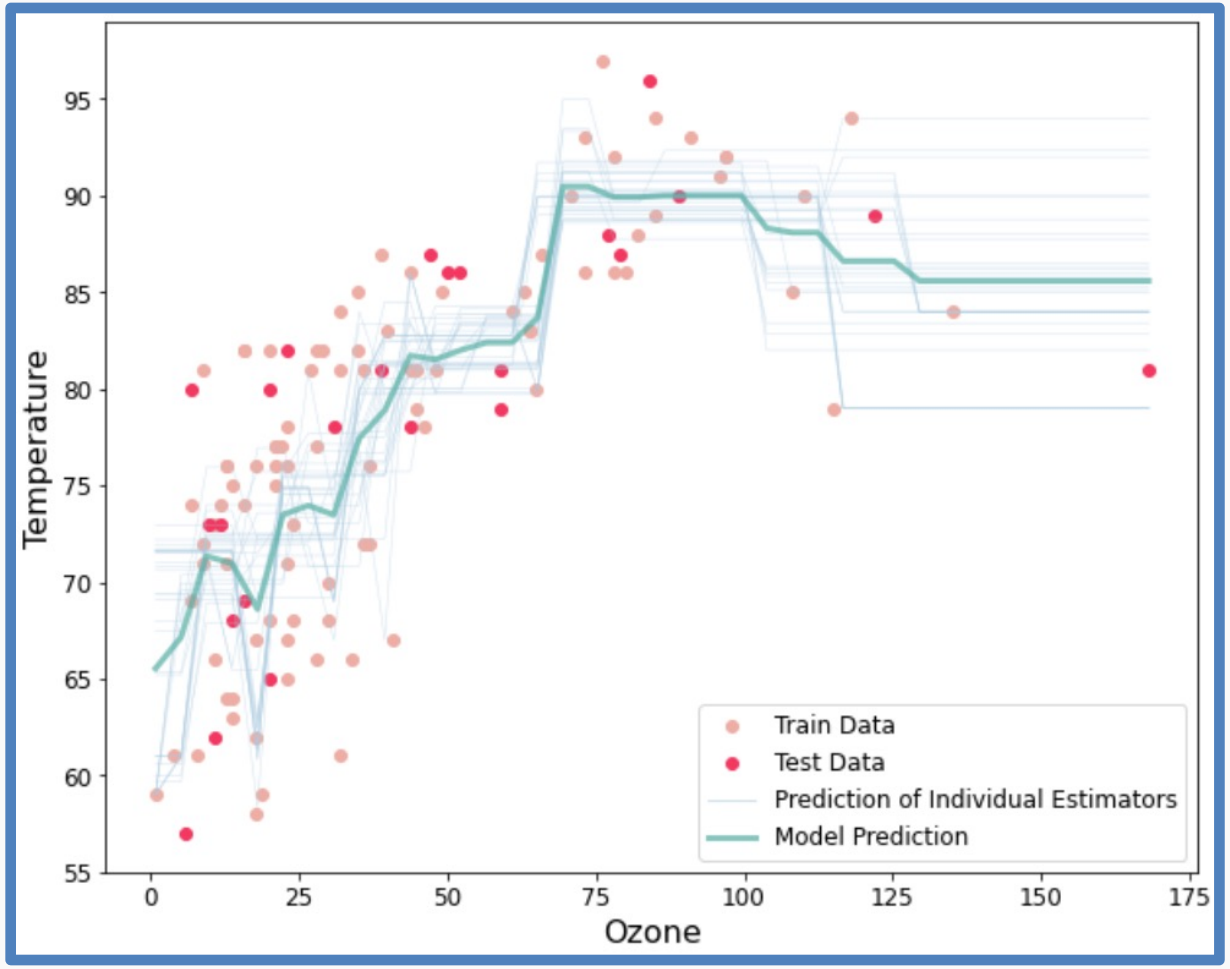
Bagging enjoys the benefits of:

1. High expressiveness - by using deeper trees each model is able to approximate complex functions and decision boundaries.

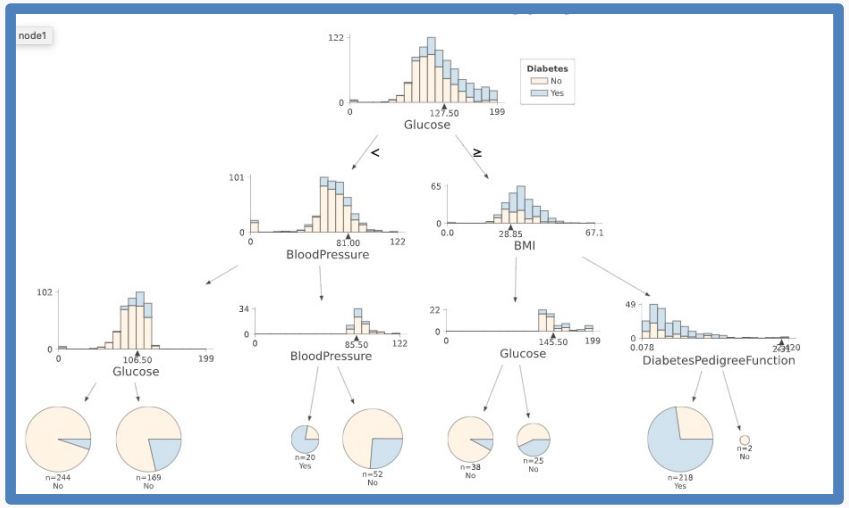
2. Low variance - averaging the prediction of all the models reduces the variance in the final prediction, assuming that we choose a sufficiently large number of trees.

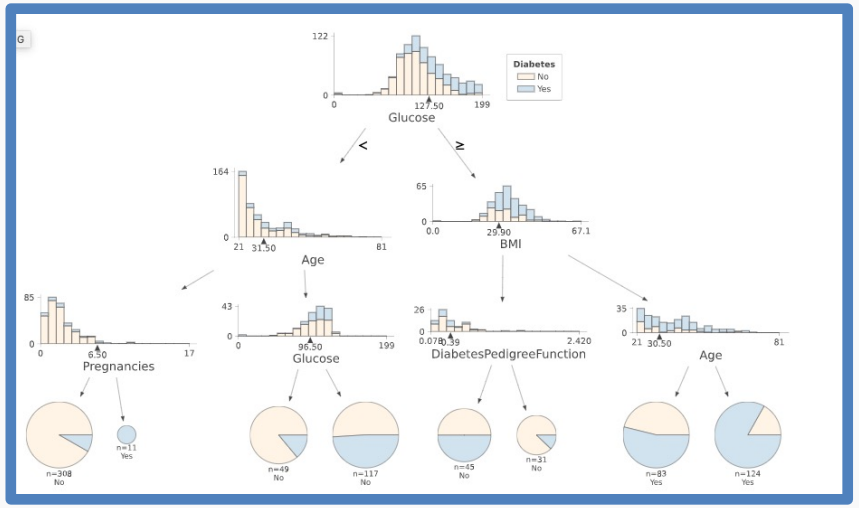
Bagging (regression)

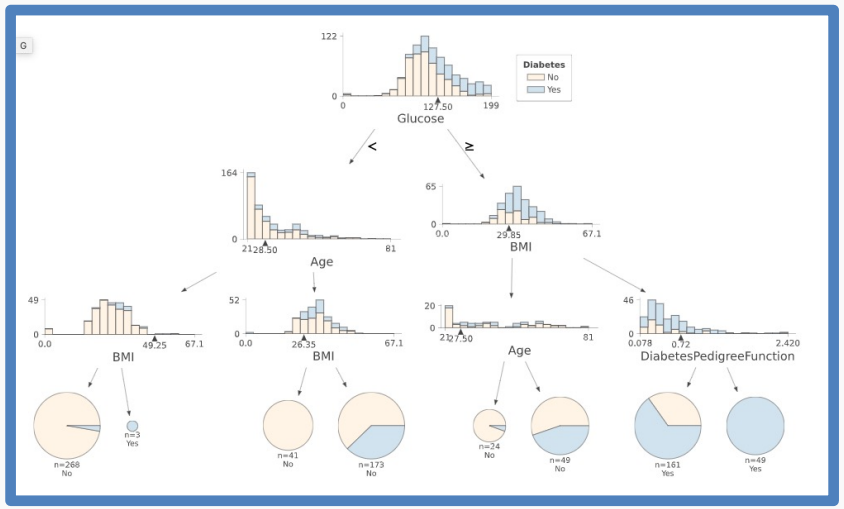
The resulting tree is the average of all tree (estimators).

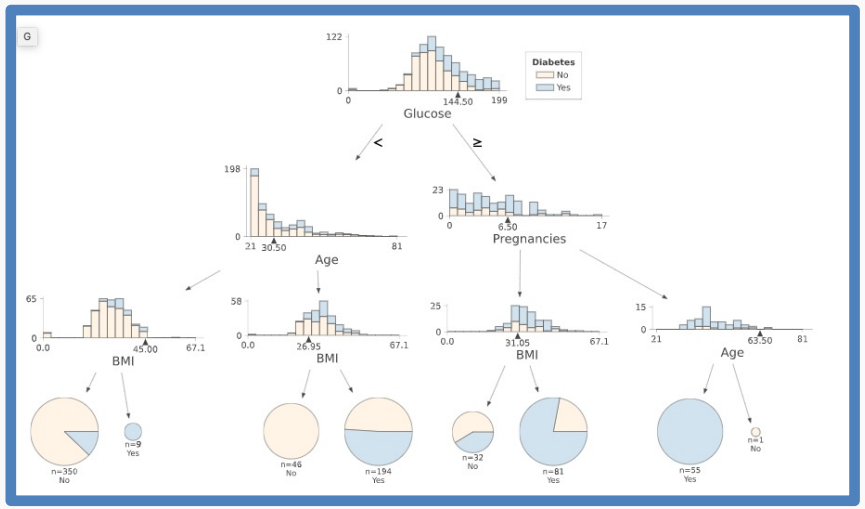


Bagging (classification)









For each bootstrap, we build a decision tree. The results is a combination (majority) of the predictions from all trees.

Bagging

Question: Do you see any problems?

Bagging

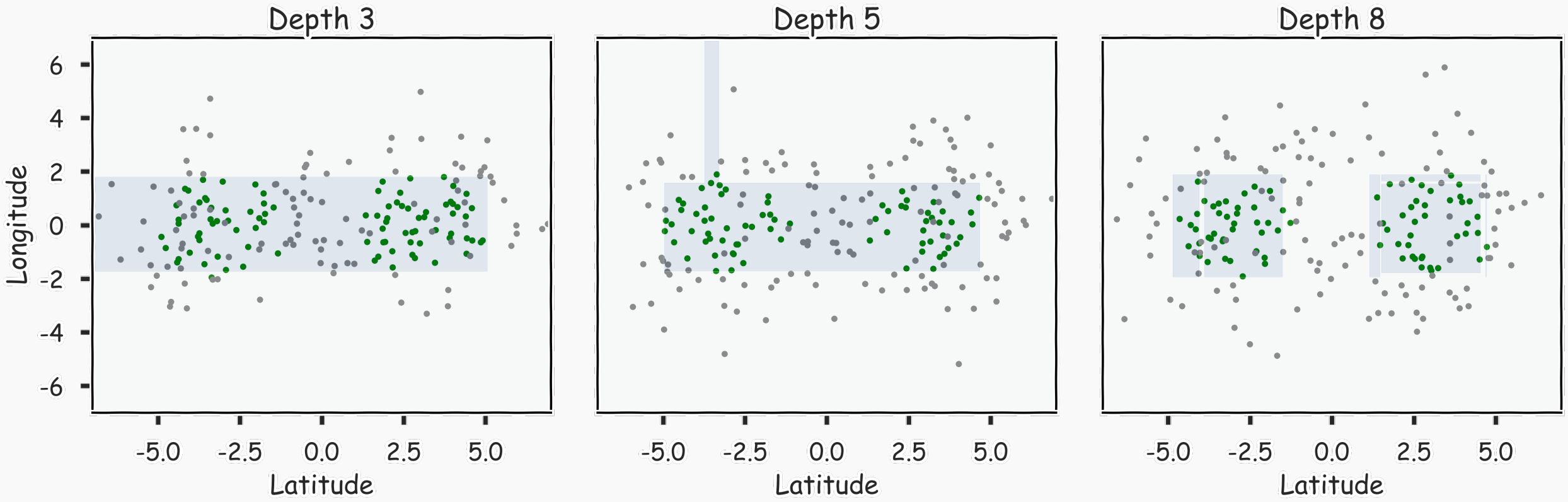
Question: Do you see any problems?

- If trees are too shallow it can still underfit.
- Still some overfitting if the trees are too large.
- Interpretability:

  The **major drawback** of bagging (and other ***ensemble methods*** that we will study) is that the averaged model is no longer easily interpretable - i.e. one can no longer trace the 'logic' of an output through a series of decisions based on predictor values!

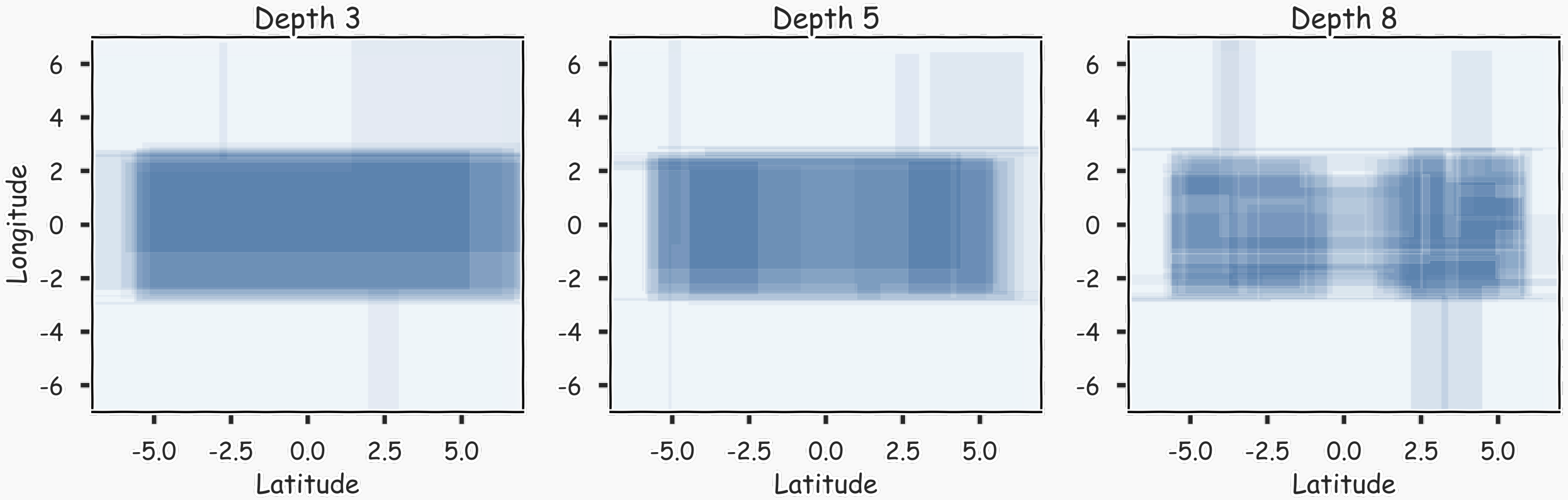
Case of underfitting

Consider the dataset below. To capture the pattern we need deeper tree.



Case of underfitting

Here we fit 100 trees using bootstrapped samples. Even with multiple estimators, the shallow tree will not be able to capture the real pattern.



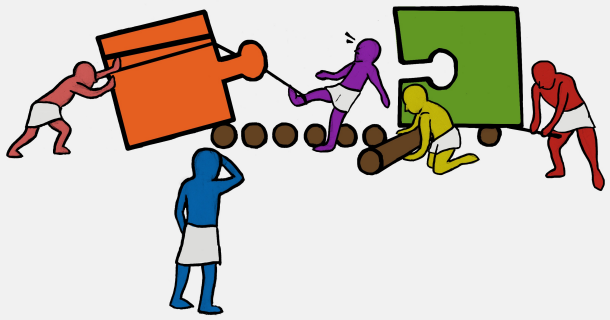
Bagging

Question: Do you see any problems?

- If trees are too shallow it can still underfit.
- Still some overfitting if the trees are too large.

Question: How do we decide on the complexity of the model?

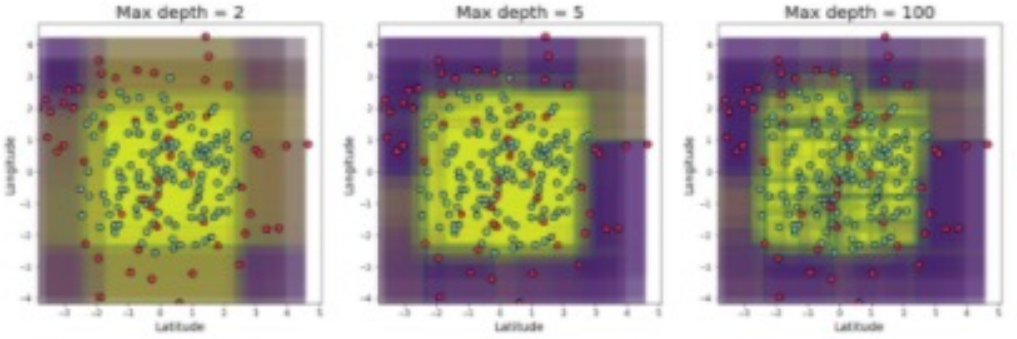
Cross Validation



👩‍🏫 Exercise: Bagging Classification with Decision Boundary

The goal of this exercise is to use **Bagging** (Bootstrap Aggregated) to solve a classification problem and visualize the influence on Bagging on trees with varying depths.

Your final plot will resemble the one below.



Outline

- Review of Decision Trees
- Bagging
- **Out of Bag Error (OOB)**
- Variable Importance

Bagging

Original Data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bootstrap Sample 1

X	Y
X_4	y_4
X_{14}	y_{14}
X_{11}	y_{11}
X_2	y_2
X_{35}	y_{35}
⋮	⋮
X_k	y_k

Decision Tree 1

Used and unused data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bagging

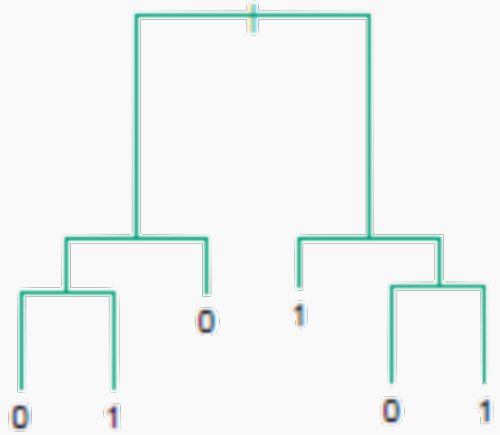
Original Data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bootstrap Sample 2

X	Y
X_5	y_5
X_3	y_3
X_{12}	y_{12}
X_{43}	y_{43}
X_1	y_1
⋮	⋮
X_k	y_k

Decision Tree 2



Used and unused data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bagging

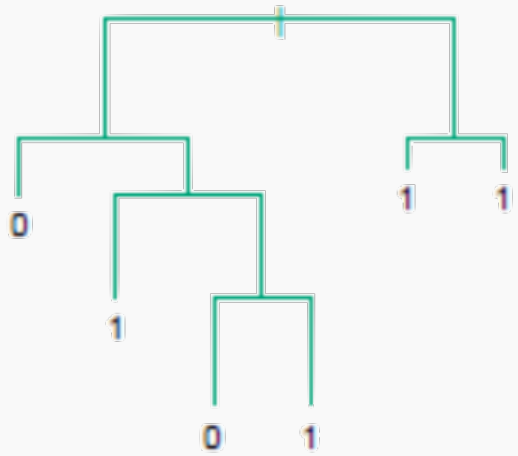
Original Data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

Bootstrap Sample 3

X	Y
X_9	y_9
X_4	y_4
X_1	y_1
X_1	y_1
X_{65}	y_{65}
⋮	⋮
X_k	y_k

Decision Tree 3



Used and unused data

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
X_4	y_4
X_5	y_5
⋮	⋮
X_n	y_n

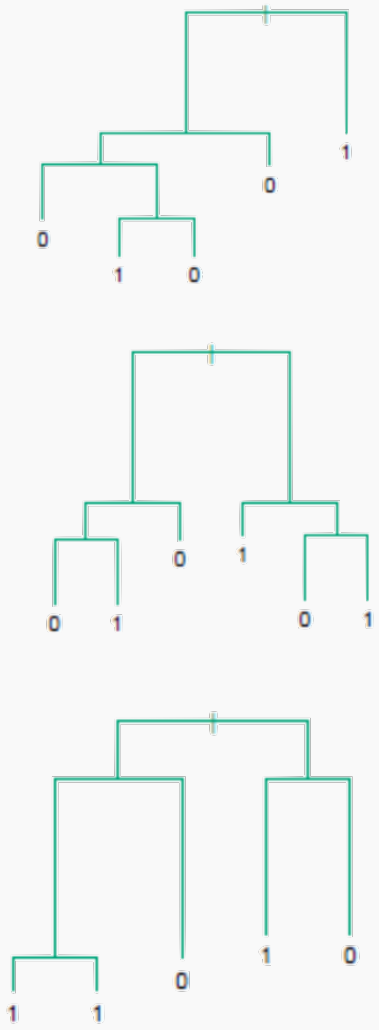
Point-wise out-of-bag error

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
⋮	⋮
X_i	y_i
⋮	⋮
X_n	y_n

Point-wise out-of-bag error

B Trees that did not see $\{X_i, y_i\}$

X	Y
X_1	y_1
X_2	y_2
X_3	y_3
⋮	⋮
X_i	y_i
⋮	⋮
X_n	y_n



Classification

$$\hat{y}_{i,pw} = \text{majority}(\hat{y}_i) \qquad e_i = \mathbb{I}(\hat{y}_{i,pw} \neq y_i)$$

Regression

$$\hat{y}_{i,pw} = \sum_{j \in B} \hat{y}_{i,j} \qquad e_i = (y_i - \hat{y}_{i,pw})^2$$

OOB Error

We average the point-wise out-of-bag error over the full training set.

Classification

$$Error_{OOB} = \frac{1}{B}\sum_{i}^{B} e_i = \frac{1}{B}\sum_{i}^{B} \mathbb{I}(\hat{y}_{i,pw} \neq \mathrm{y_i})$$

Regression

$$Error_{OOB} = \frac{1}{B}\sum_{i}^{B} e_i = \frac{1}{B}\sum_{i}^{B} (y_i - \hat{y}_{i,pw})^2$$

Out-of-Bag Error

Bagging is an example of an ***ensemble method***, a method of building a single model by training and aggregating multiple models.

With ensemble methods, we get a new metric for assessing the predictive performance of the model, the ***out-of-bag error***.

Given a training set and an ensemble of models, each trained on a bootstrap sample, we compute the ***out-of-bag error*** of the averaged model by

1. For each point in the training set, we average the predicted output for this point over the models whose bootstrap training set excludes this point. We compute the error or squared error of this averaged prediction. Call this the point-wise out-of-bag error.

2. We average the point-wise out-of-bag error over the full training set.

Bagging

Question: Do you see any problems?

- If trees are too shallow it can still underfit.
- Still some overfitting if the trees are too large.
- Interpretability:

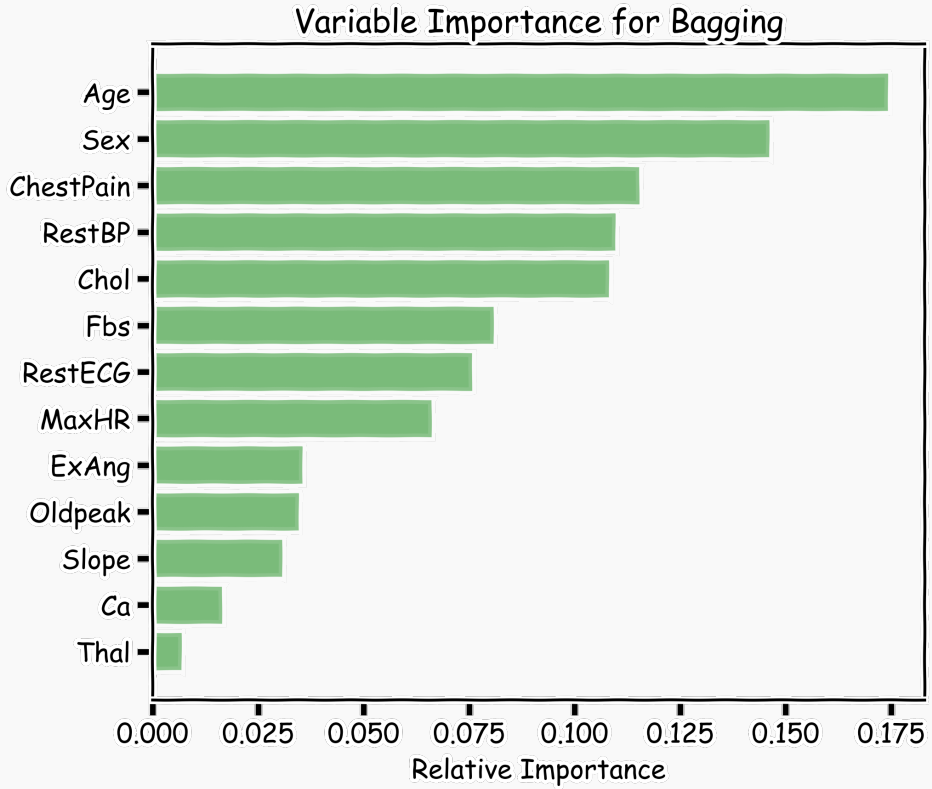
  The **major drawback** of bagging (and other ***ensemble methods*** that we will study) is that the averaged model is no longer easily interpretable - i.e. one can no longer trace the 'logic' of an output through a series of decisions based on predictor values!

Outline

- Review of Decision Trees
- Bagging
- Out of Bag Error (OOB)
- **Variable Importance**

Variable Importance for Bagging

Calculate the total amount that the MSE (for regression) or Gini index (for classification) is decreased due to splits over a given predictor, averaged over all B trees.



100 trees, max_depth=10

Improving on Bagging

In practice, the ensembles of trees in Bagging tend to be highly correlated.

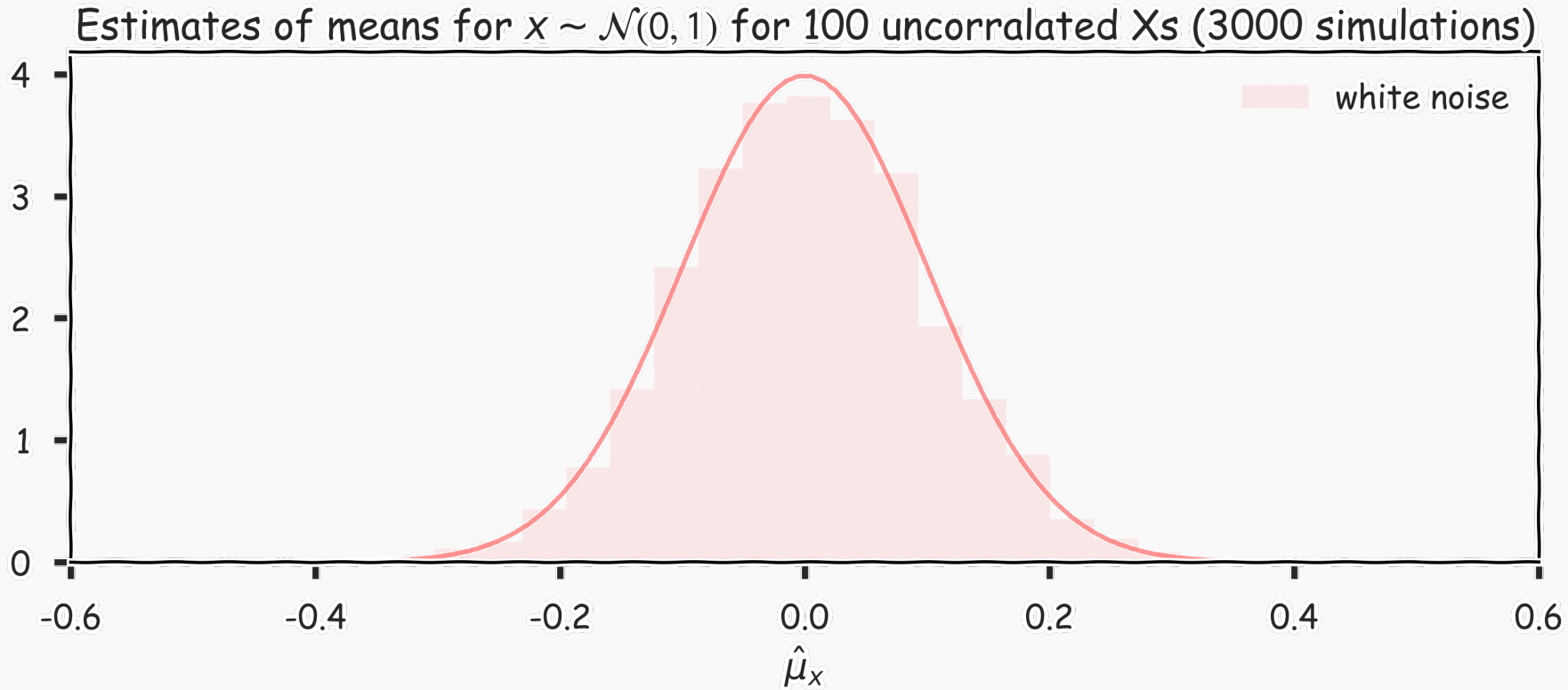
Suppose we have an extremely strong predictor, x_j , in the training set amongst moderate predictors. Then the greedy learning algorithm ensures that most of the models in the ensemble will choose to split on x_j in early iterations.

However, we assumed that each tree in the ensemble is independently and identically distributed, with the expected output of the averaged model the same as the expected output of any one of the trees.

Improving on Bagging

Recall, for B number of identically and independently distributed variable, X, with variance σ^2, the variance of the estimate of the mean is :

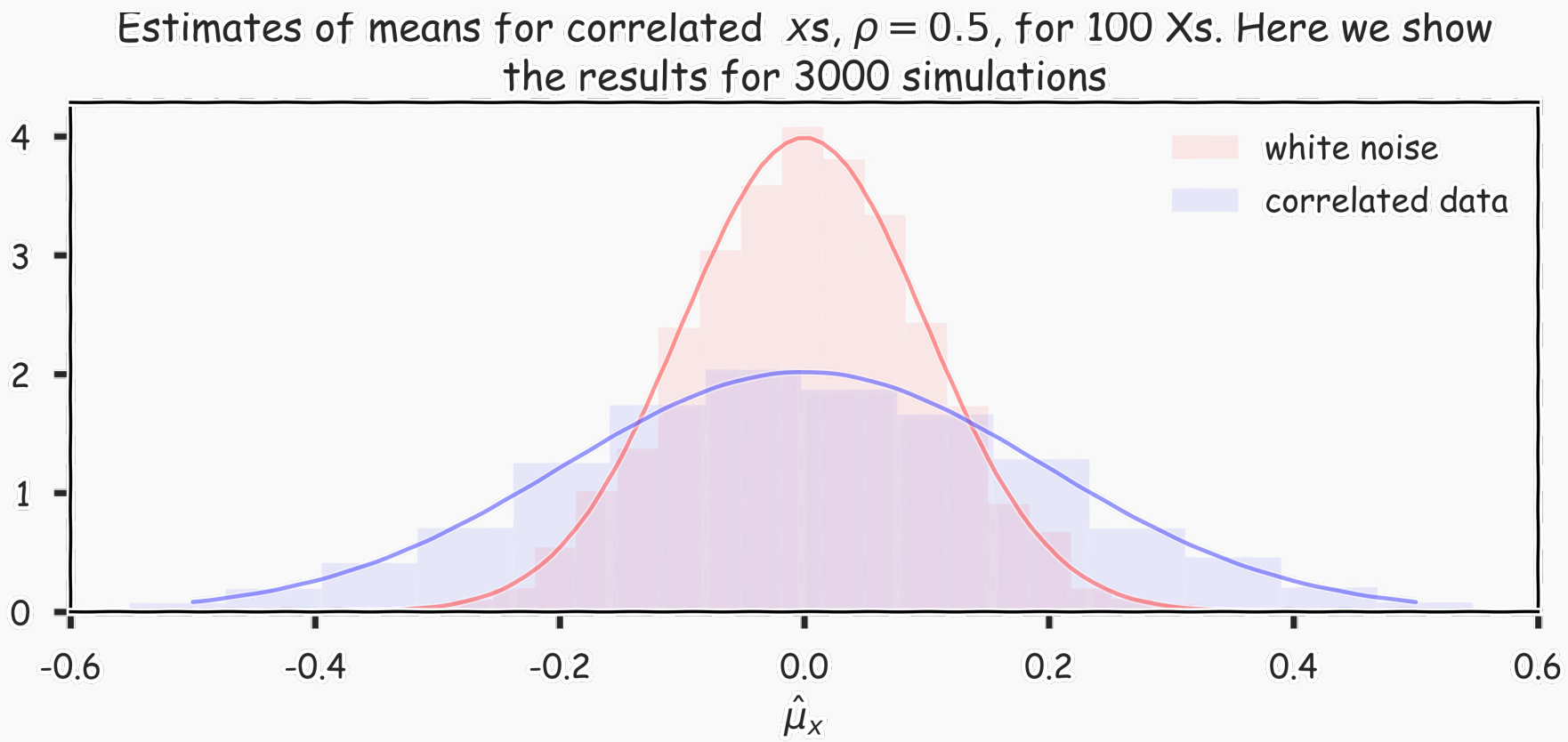
$$\text{var}(\hat{\mu}_x) = \frac{\sigma^2}{B}$$



Improving on Bagging

For B number of identically but not independently distributed variables with pairwise correlation ρ and variance σ^2, the variance of their mean is

$$\text{var}(\hat{\mu}_x) \propto \sigma^2(1+\rho^2)/B$$



Another cliff hanger in CS109A