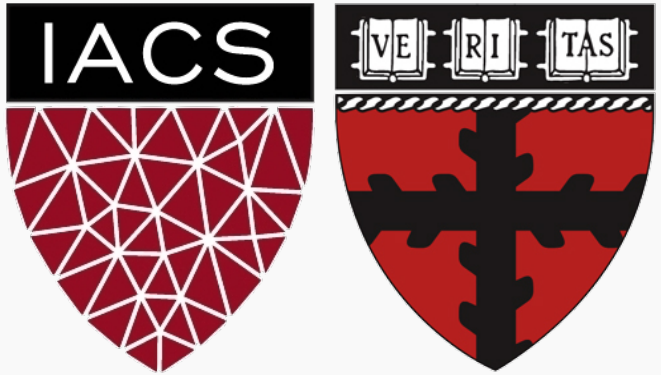
Lecture 16: Bagging and Random Forest

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner



ANNOUNCEMENTS

- Homework 5 (109) due tonight 11:59 pm
- Homework 5 (209) due next Wednesday 11:59 pm, Nov 6
- Homework 6 (109) to be released tomorrow morning
- No 209 for homework 6

ANNOUNCEMENTS

- **Homework 7 individual**

HW1

HW4

HW7

Without OH or ED forum

ANNOUNCEMENTS

- **After CS109B, STAT 139 …**

Outline

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

Outline

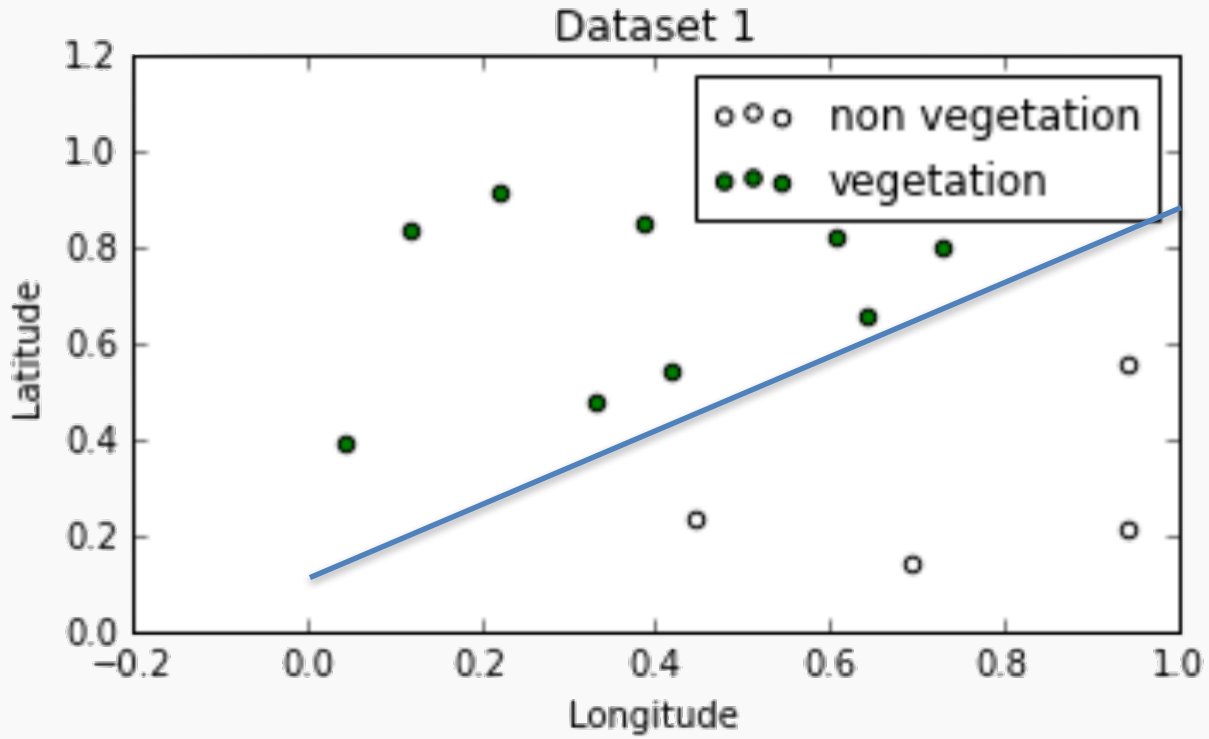
- **Review of Decision Trees**
- Bagging
- Out of Bag Error (OOB)
- Variable Importance
- Random Forests

sklearn bins continuous predictors to try all splits at edges of bins to see which split gives the best 'purity'

Geometry of Data

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

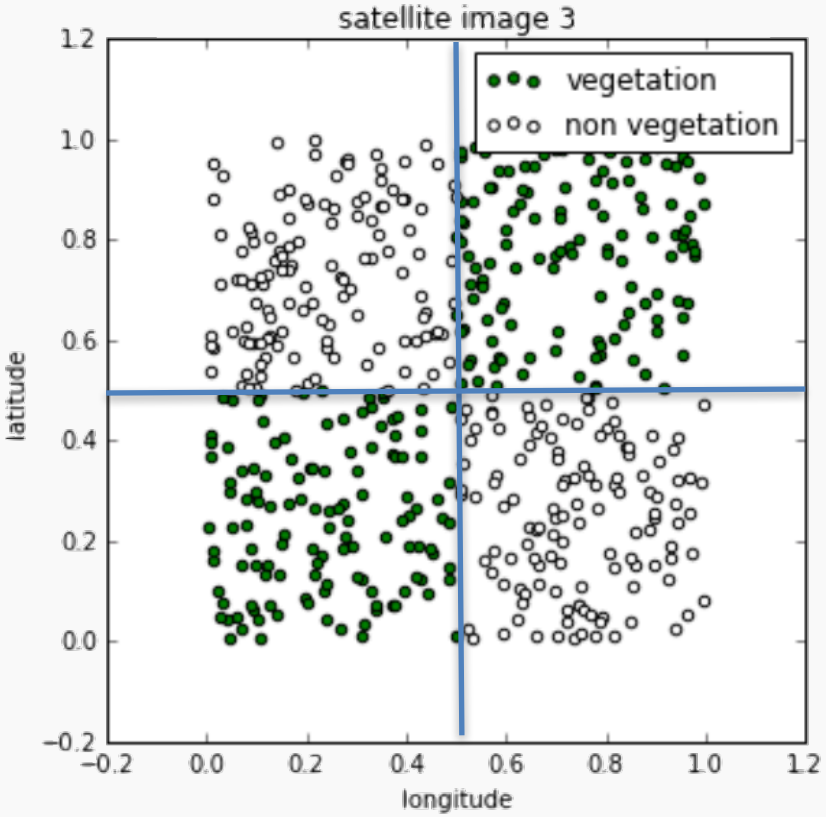
$$-0.8x_1 + x_2 = 0 \implies x_2 = 0.8x_1 \Rightarrow Latitude = 0.8\, Lon$$



Geometry of Data

Complicate decision boundaries can not be explained with LogRegression.

vertical split - how do you choose? all splits equal green and white…bad example



Decision Trees

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

1. Start with an empty decision tree (undivided feature space)
2. Choose the 'optimal' predictor on which to split and choose the 'optimal' threshold value for splitting by applying a ***splitting criterion***
3. Recurse on each new node until ***stopping condition*** is met

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 .

Splitting Criteria

The splitting criteria we've examined each minimize a loss function

A. **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).
B. **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.

Overall loss function to minimize:
for each split, you minimize entropy in a greedy way
you want to minimize cross-entropy of whole tree
entropy is a measure of 'confusion'

Stopping Conditions

Stopping Conditions

Common simple stopping conditions:

- 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.

The appropriate thresholds can be determined by evaluating the model on a held-out data set or, better yet, via cross-validation.

Stopping Conditions

- Compute the gain in purity, information or reduction in entropy of splitting a region R into R_1 and R_2:

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

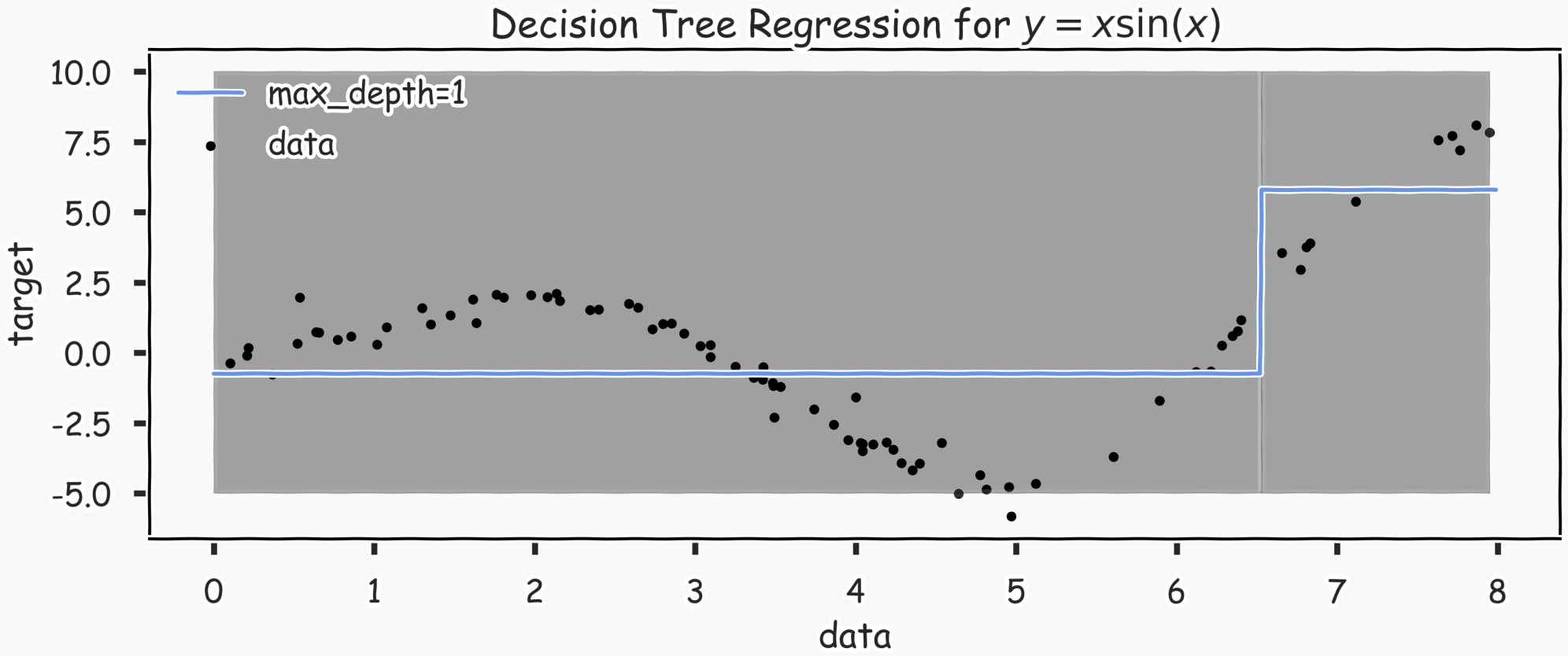
where m is a metric like the Gini Index or entropy. Don't split if the gain is less than some pre-defined threshold (**min_impurity_decrease**).

In the place of purity gain, we can instead compute accuracy gain for splitting a region R

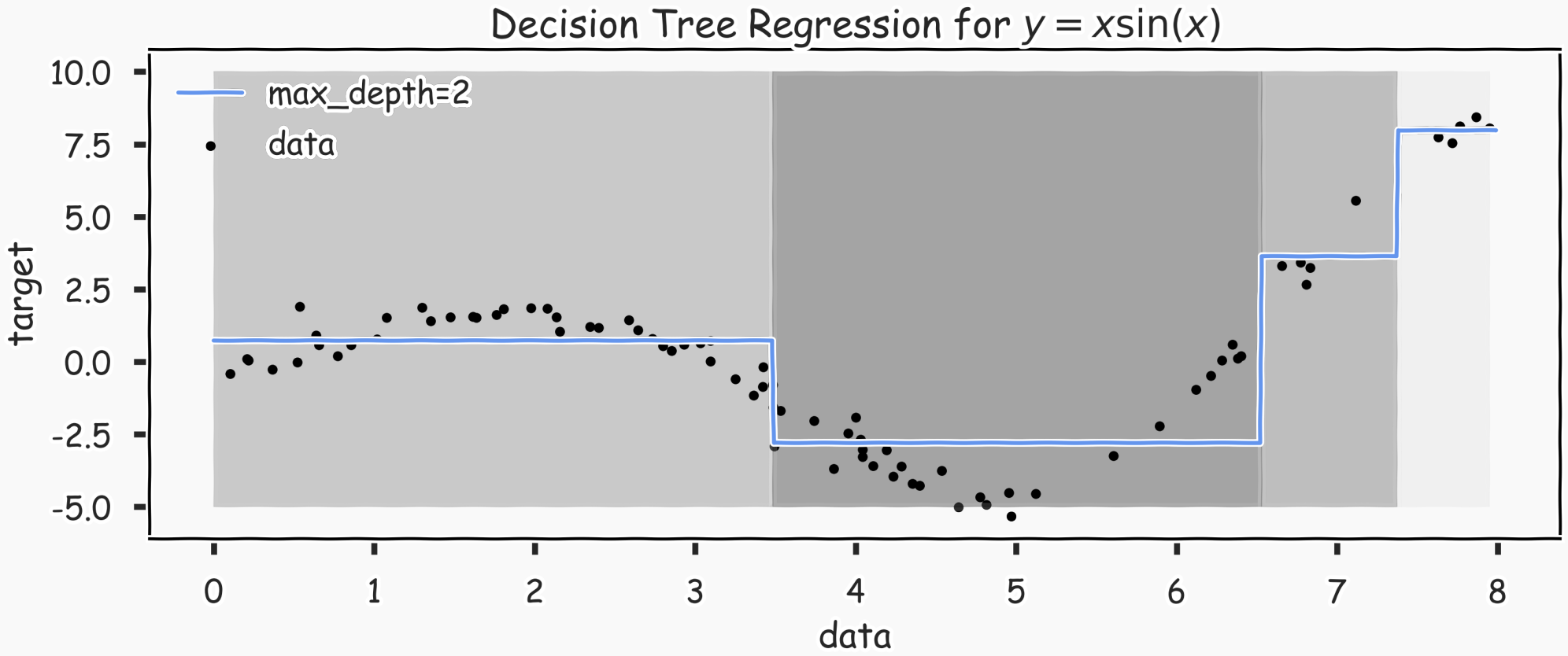
$$\text{Gain}(R) = \Delta(R) = MSE(R) - \frac{N_1}{N} MSE(R_1) - \frac{N_2}{N} MSE(R_2)$$

and stop the tree when the gain is less than some pre-defined threshold

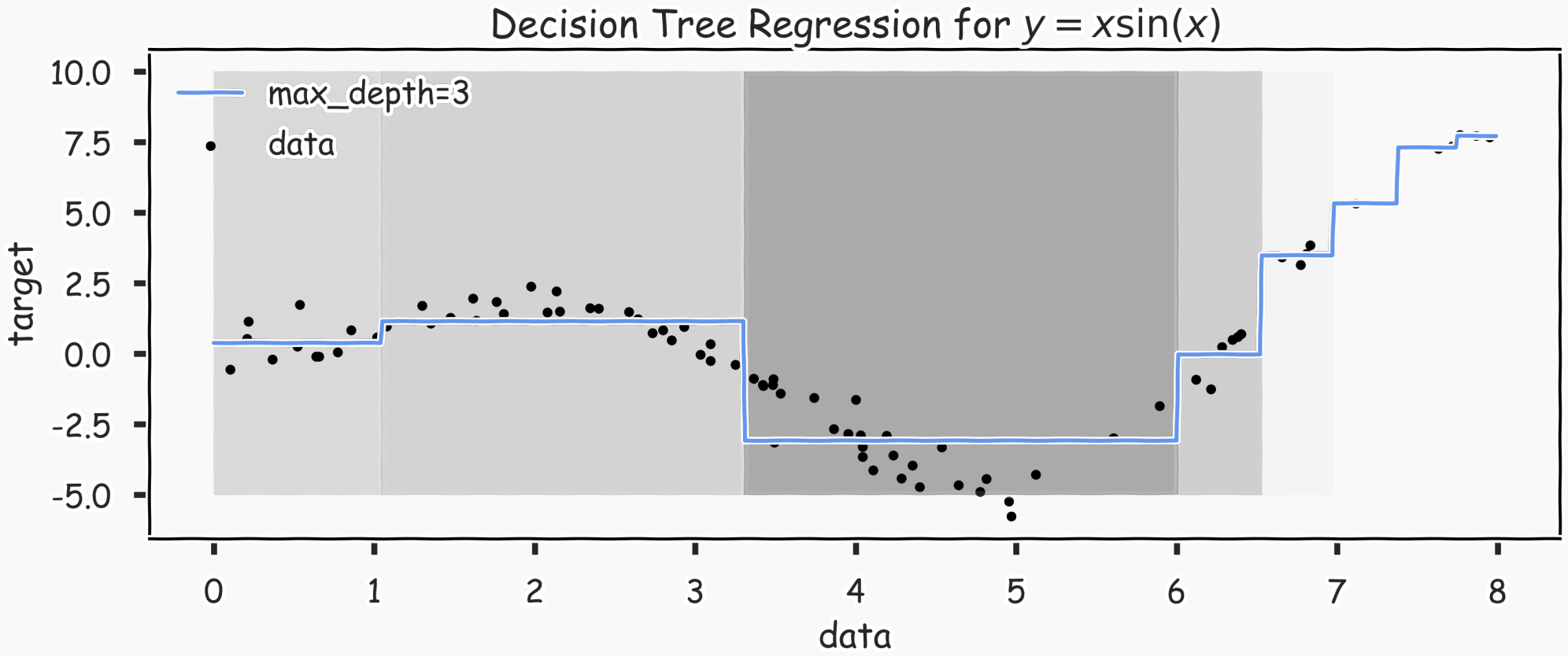
Regression Trees Prediction (grey scale represents MSE)



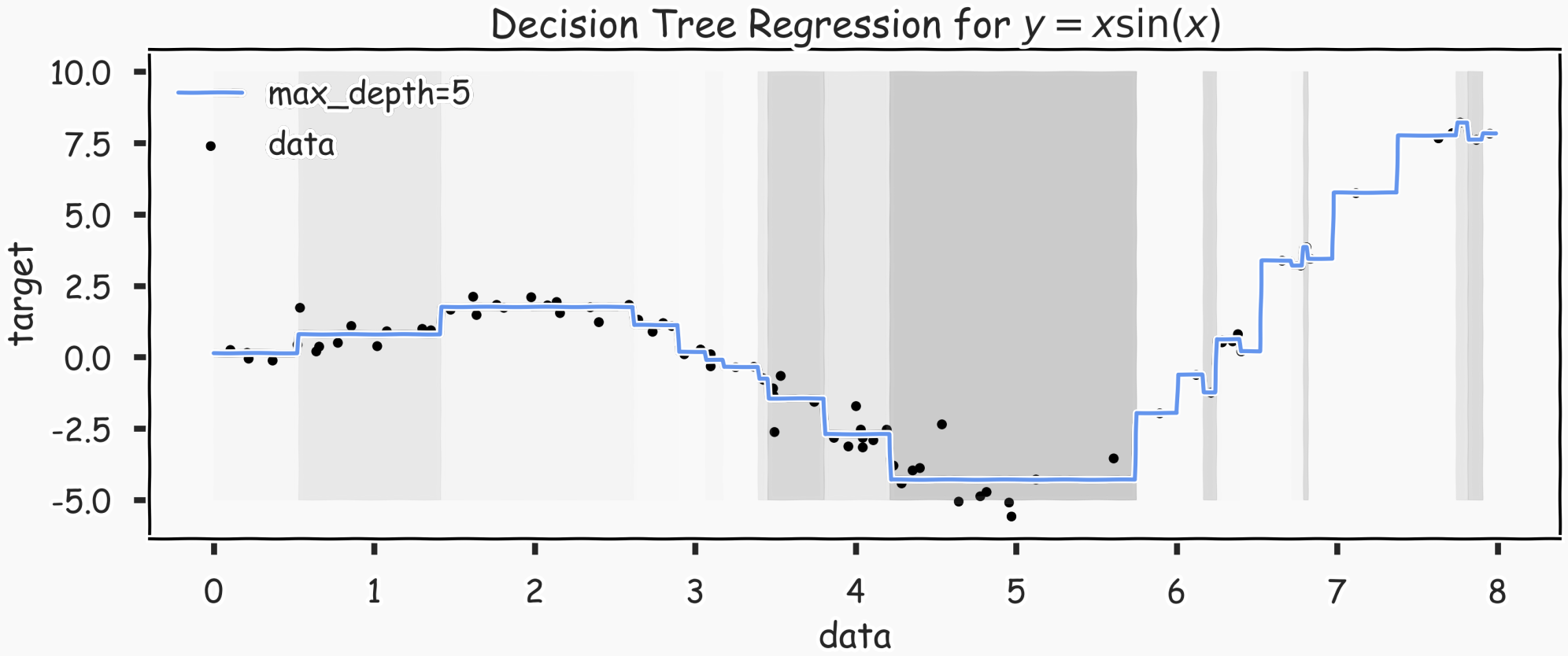
Regression Trees Prediction (grey scale represents MSE)



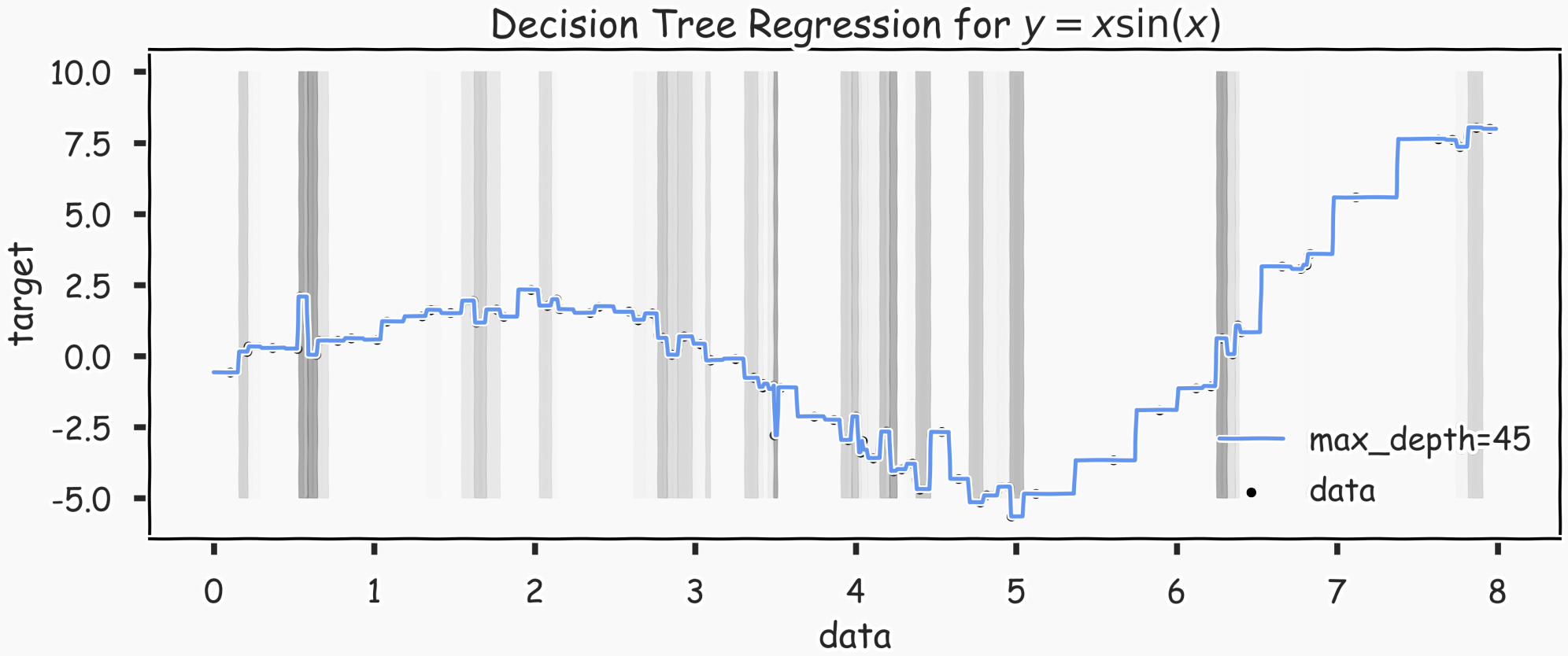
Regression Trees Prediction (grey scale represents MSE)



Regression Trees Prediction (grey scale represents MSE)



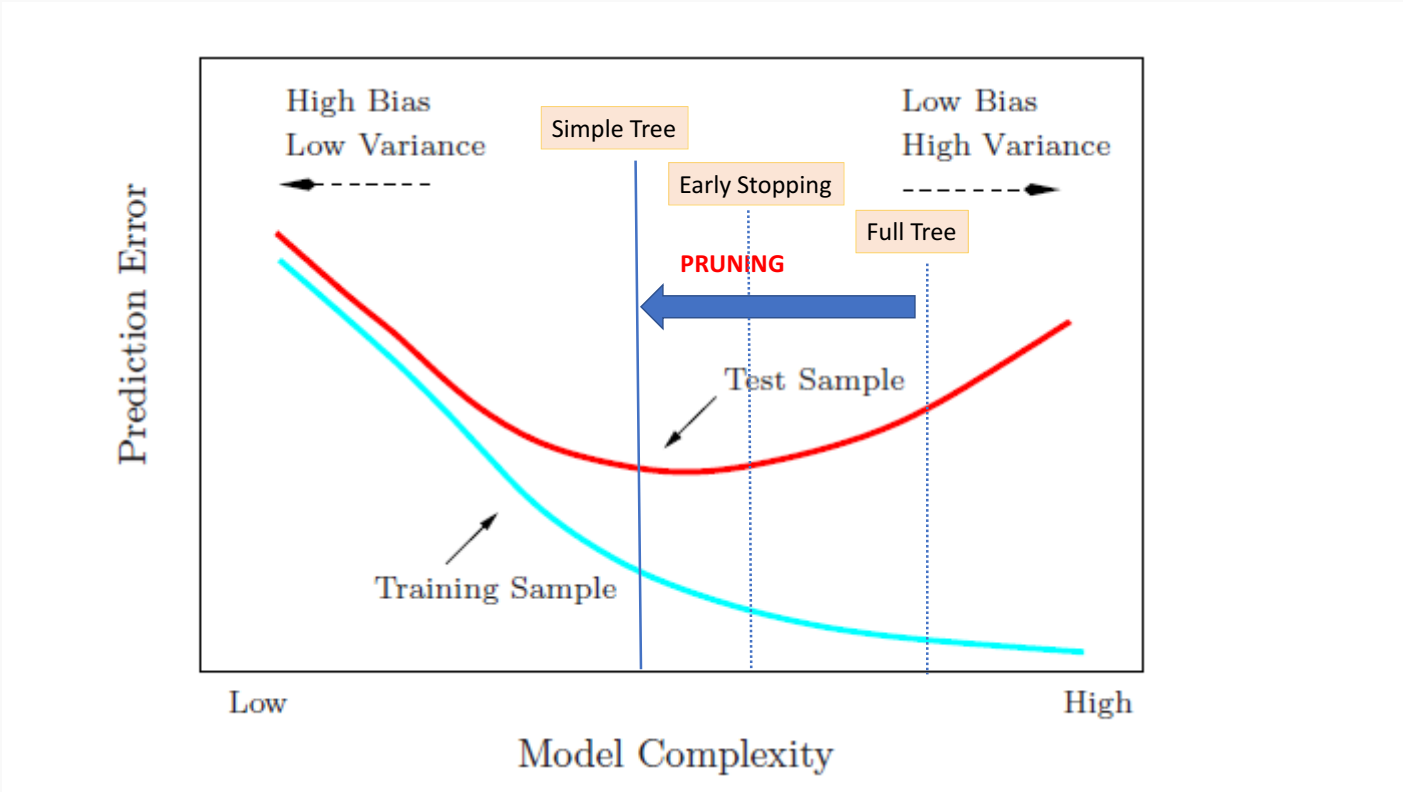
Regression Trees Prediction (grey scale represents MSE)



Overfitting

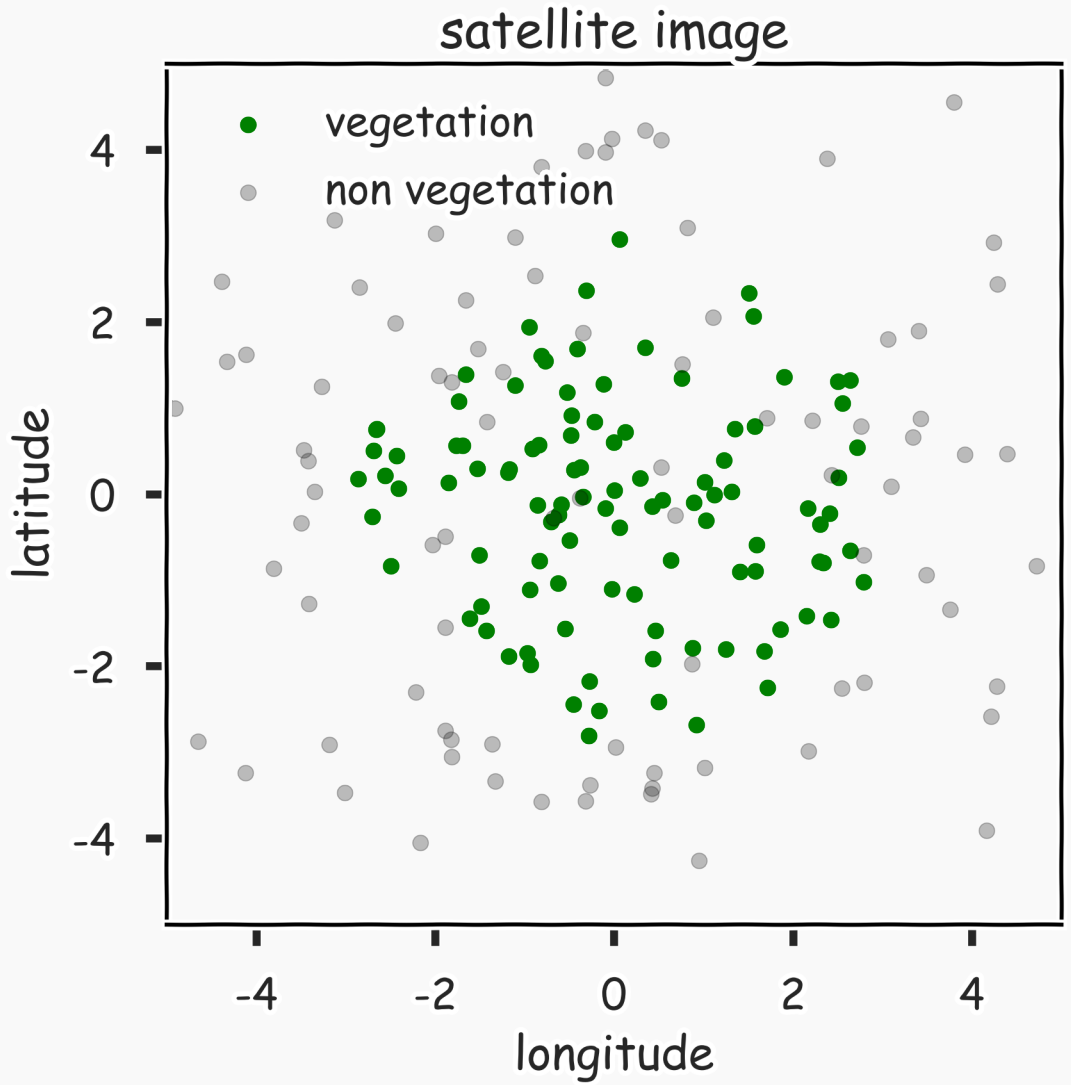
Same issues as with classification trees. Avoid overfitting by pruning or limiting the depth of the tree and using CV.

Like regressions, overfit trees very dependent on the input data - high variance in the model
How do we avoid high variance in the model? Bagging

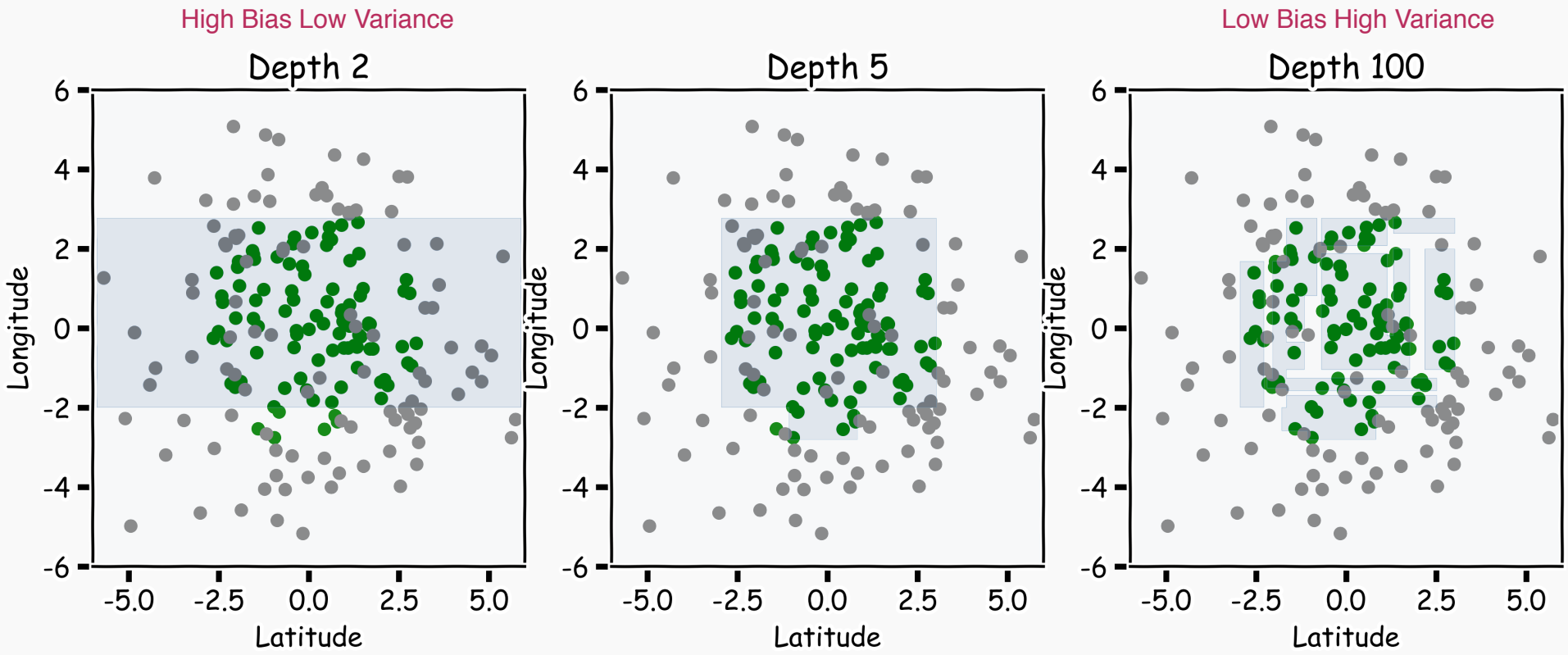


Bagging

Reduce the variance

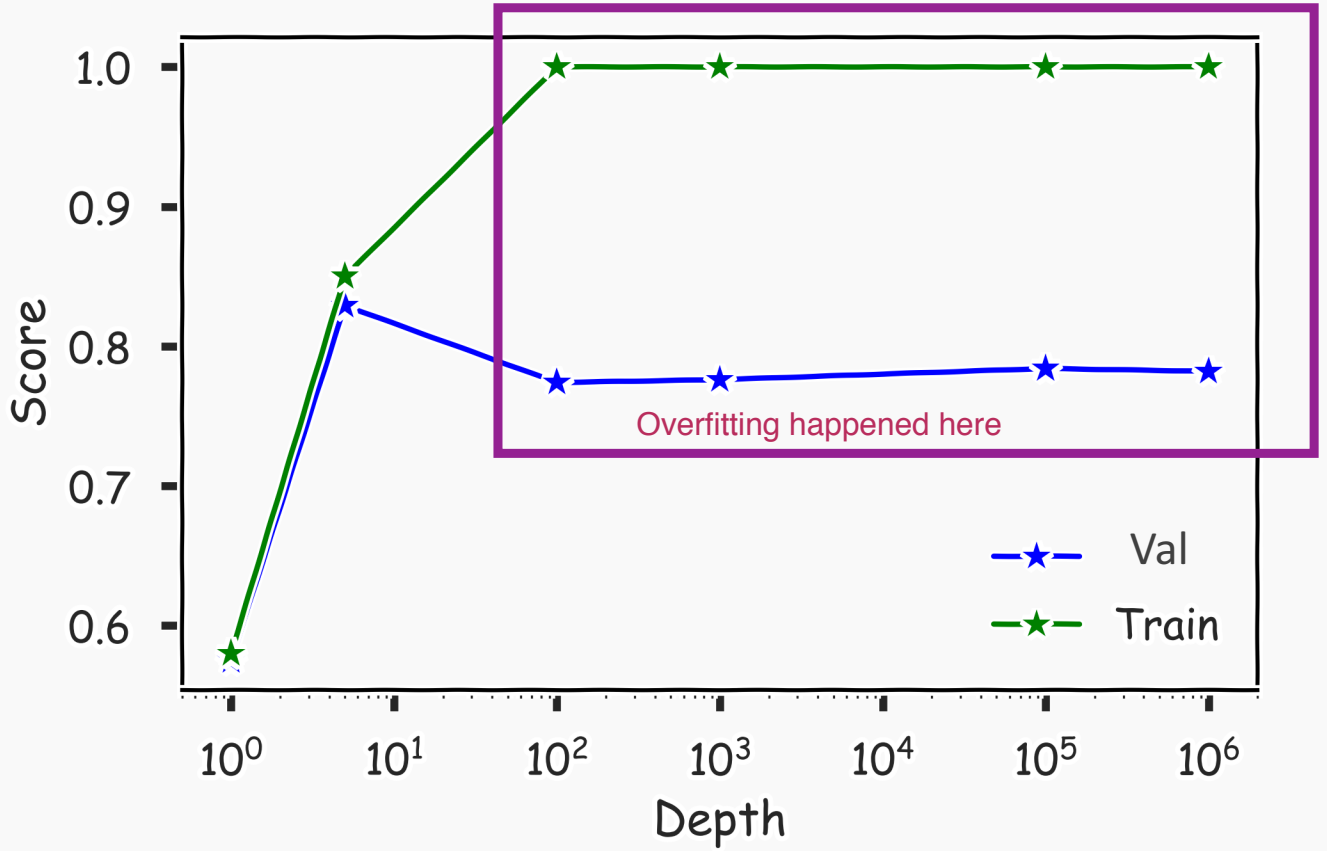


Hyper-parameters: Depth



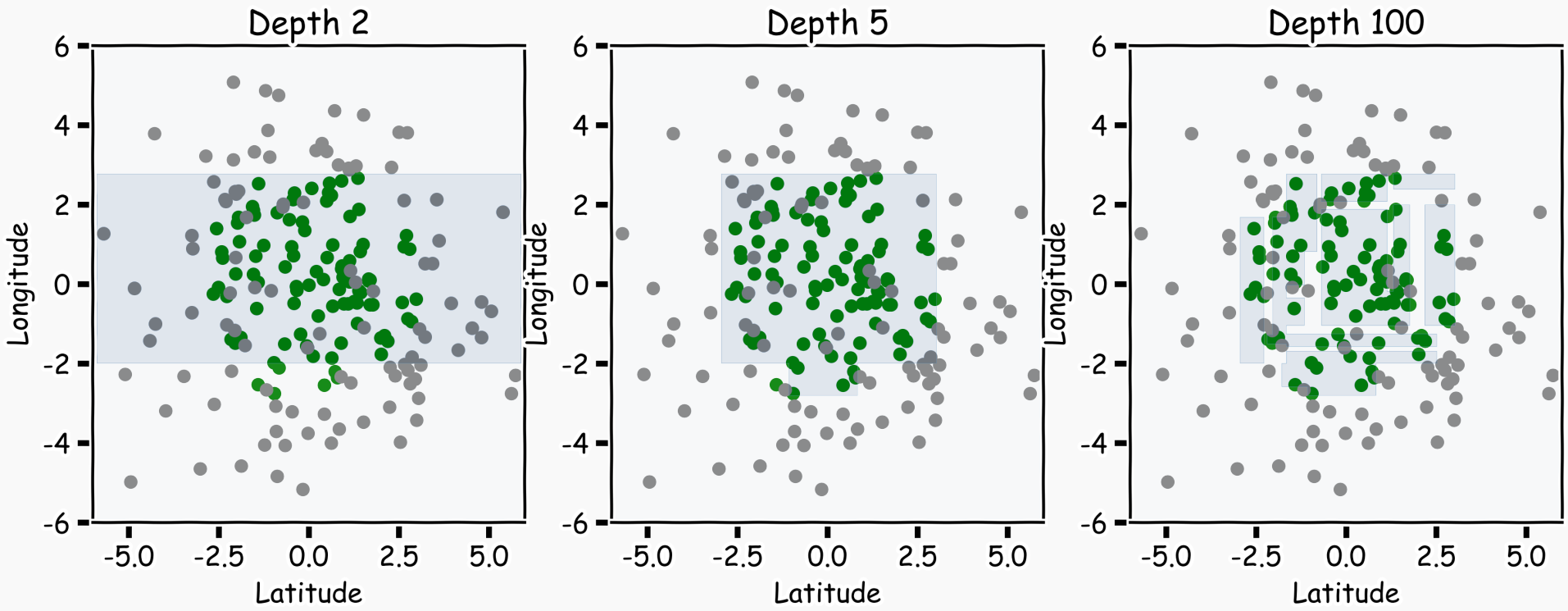
Hyper-parameters: Depth

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



Magic realism: Bootstrap

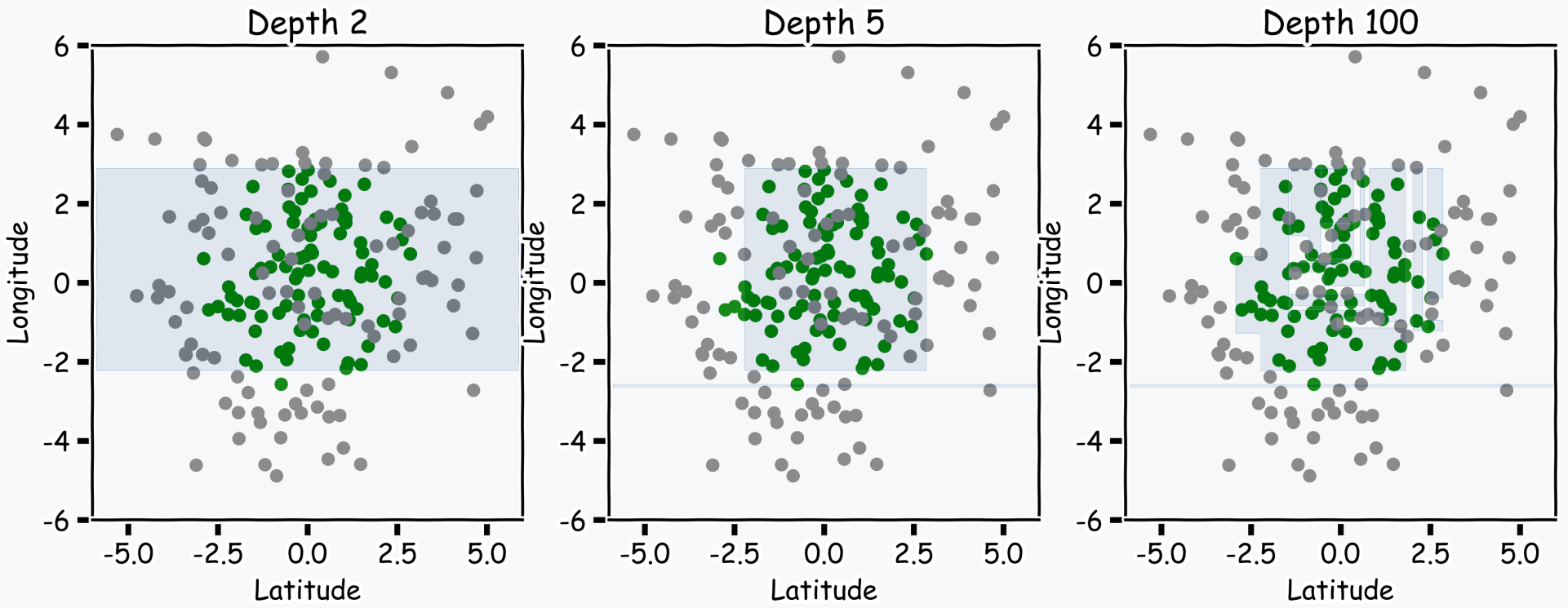
Bootstrap iter 1



Magic realism: Bootstrap

Bootstra
p iter 2

Change boundaries a lot on this next iter



Limitations of Decision Tree Models

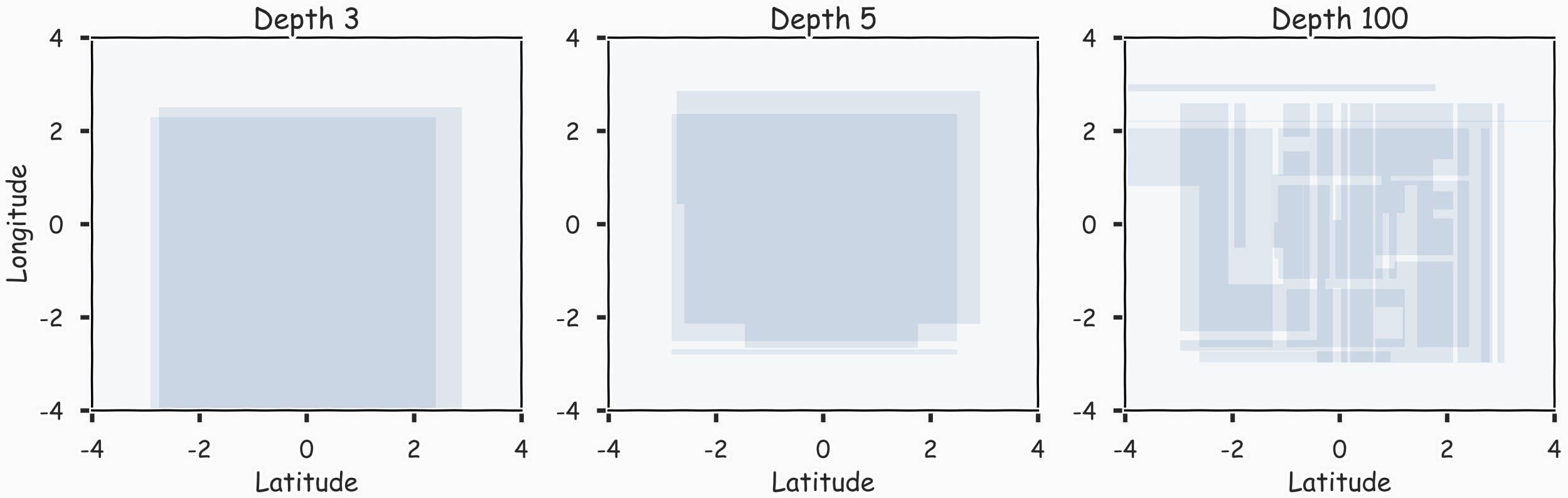
Decision trees models are highly interpretable and fast to train, using our greedy learning algorithm.

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

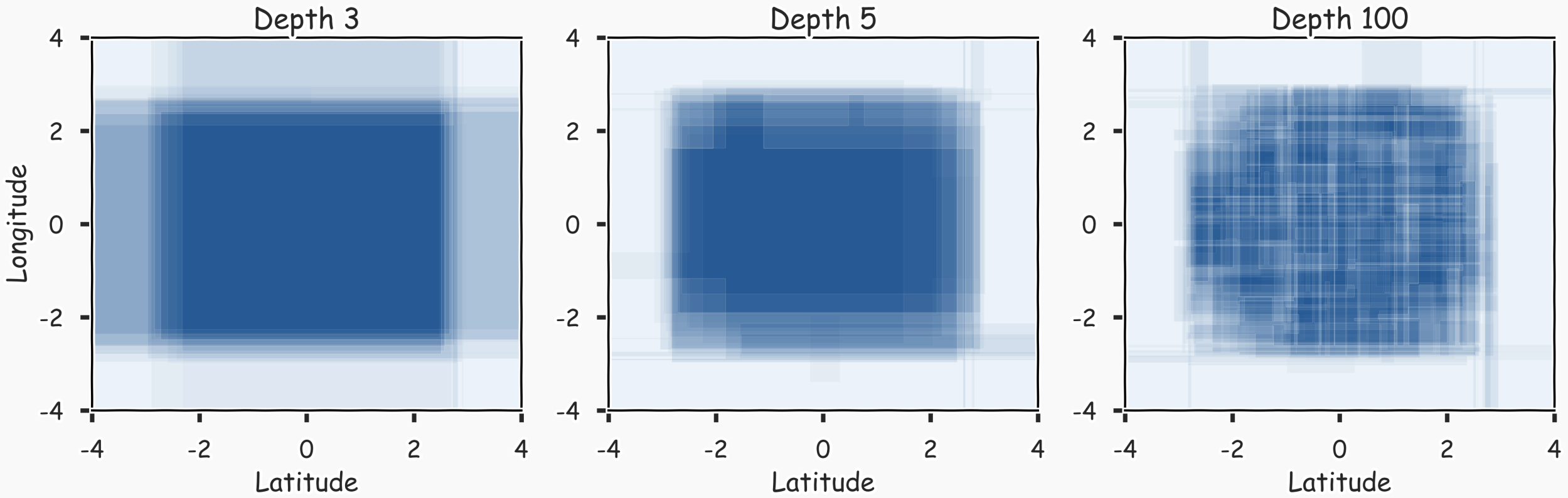
We've seen that large trees have high variance and are prone to overfitting.

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

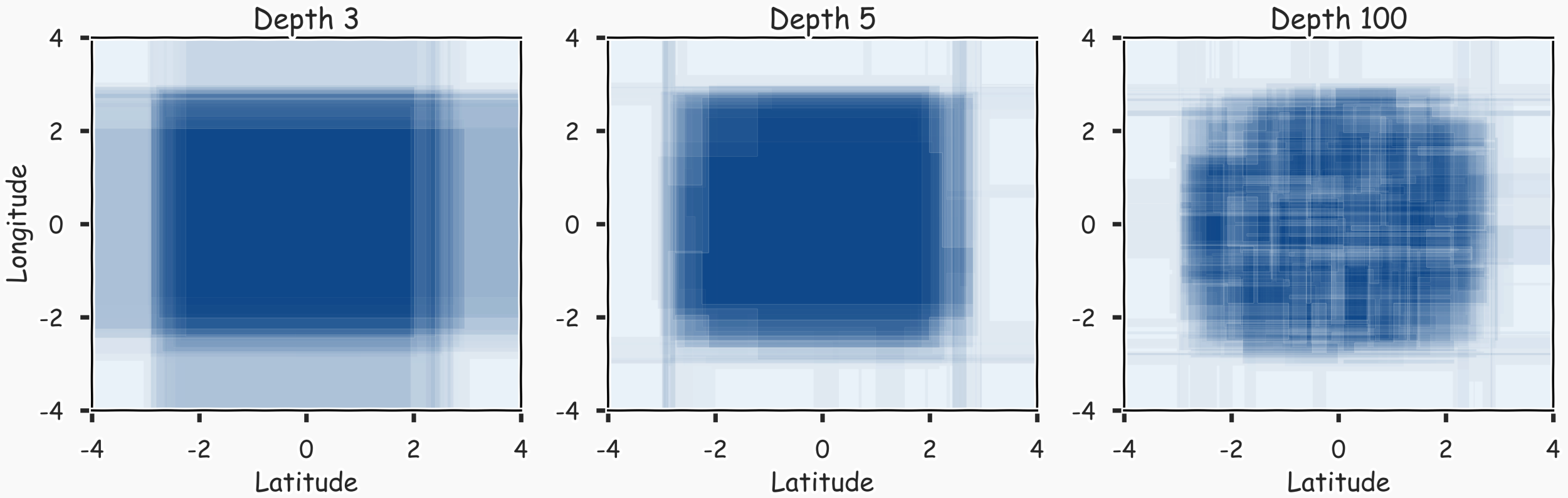
Combine them? 2 magic realisms



Combine them? 20 magic realisms



Combine them? 100 magic realisms

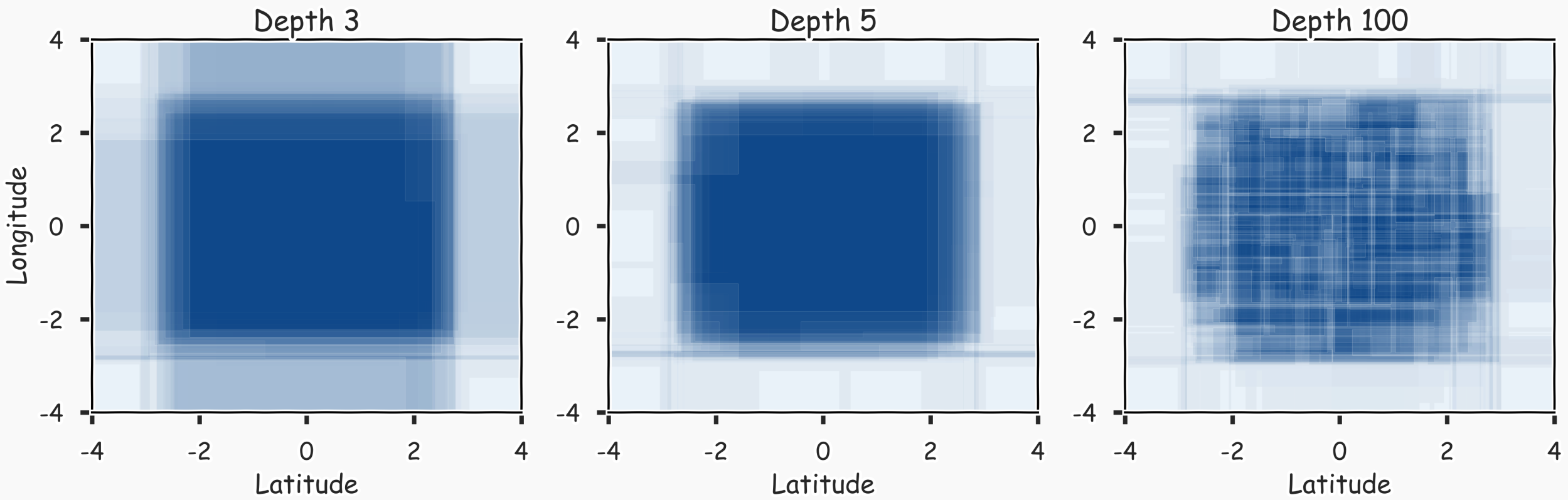


Combine them? 300 magic realisms

Where do all the decision trees agree?

Square

Round-ish? We capture complex boundaries



with enough bootstrapping and averaging, you can wash out the high variance of overfitting and pull out the important features
but you also don't want to do a crazy amount of bootstrapping, (you might repeat samples), so you should limit your tree depth to something reasonable

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 full 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.

bigger trees, not necessarily full

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.

This method is called ***Bagging*** (Breiman, 1996), short for, of course, Bootstrap Aggregating.

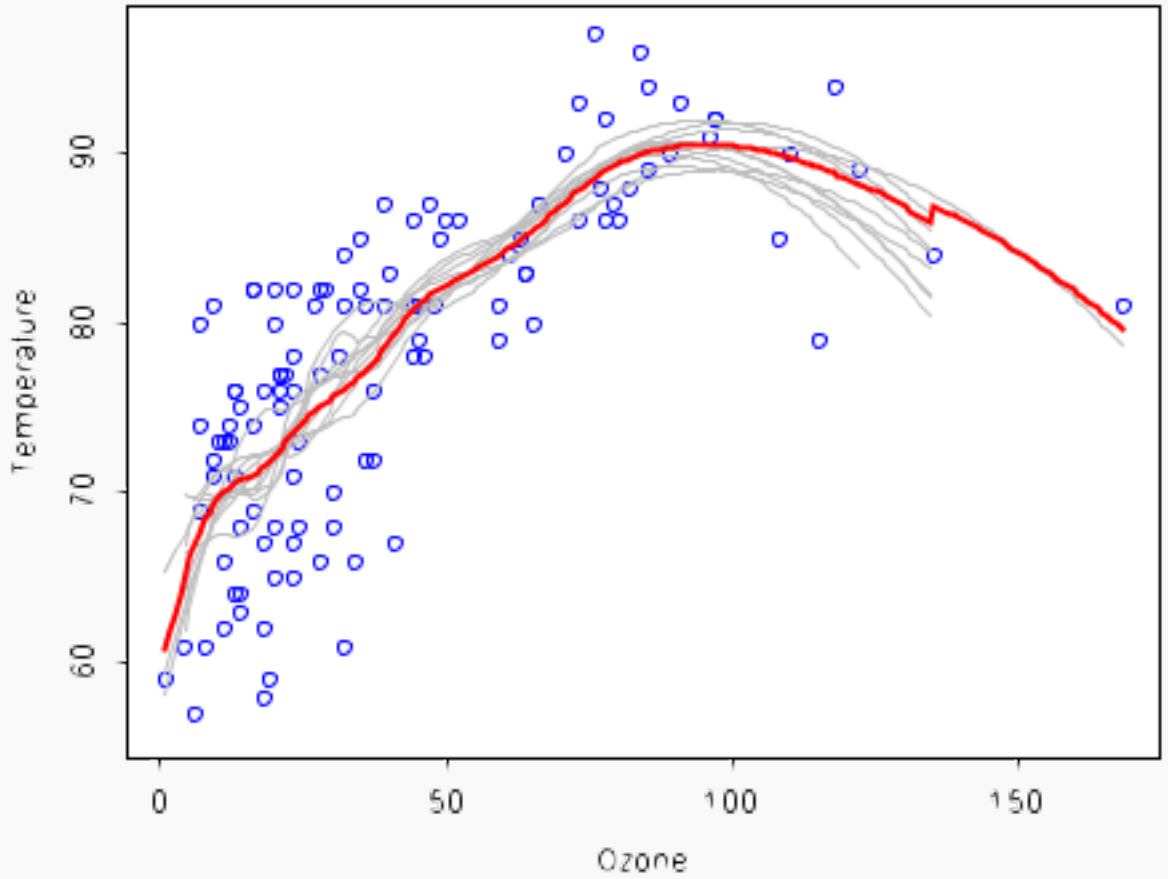
Bagging

Note that bagging enjoys the benefits of:

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

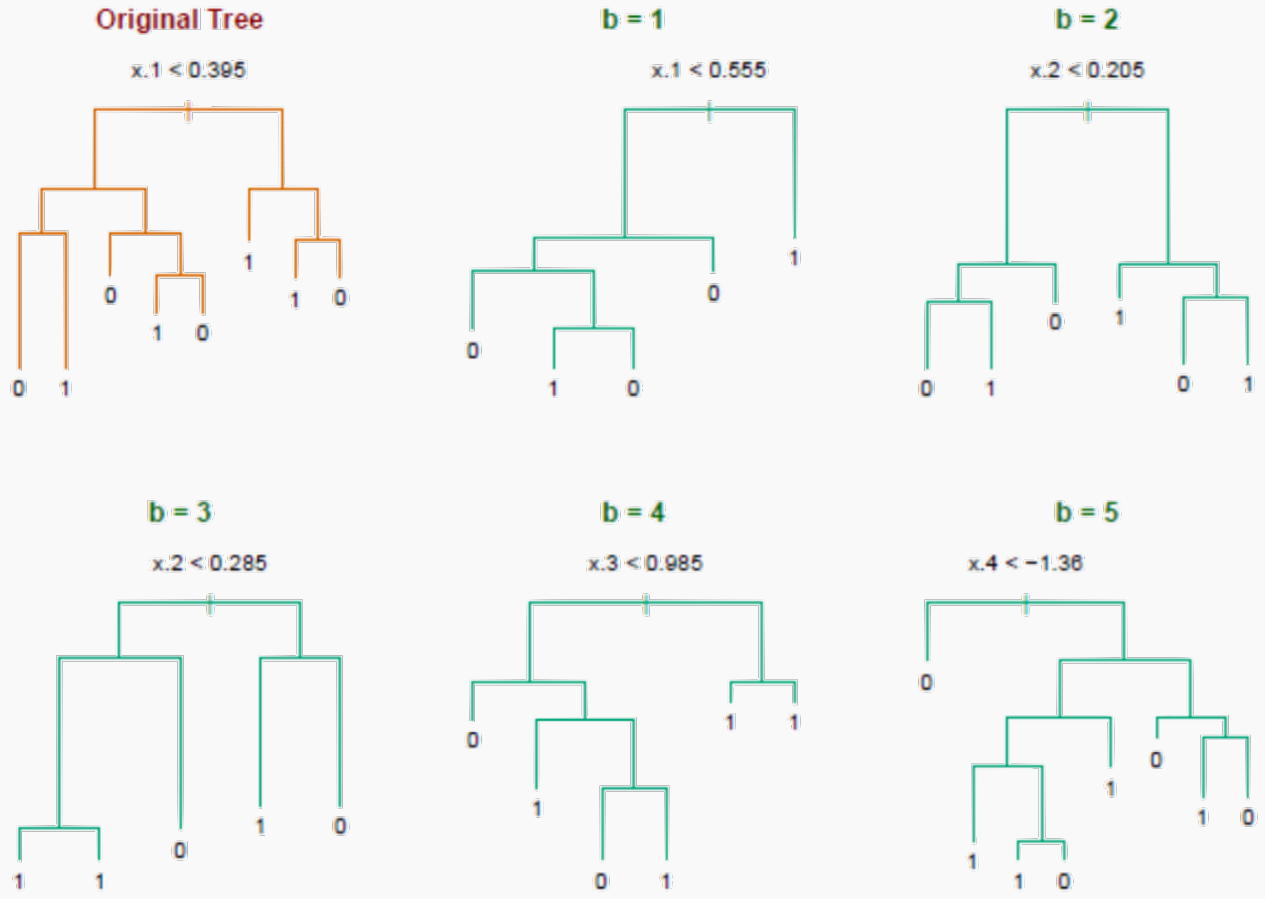
2. Low variance - averaging the prediction of all the models reduces the variance in the final prediction, assuming that we choose a sufficiently large number of trees.

Bagging (regression)



Bagging (classification)

To classify a point, you take the plurality of the outcomes of each of the trees (regression = average). Basically, you put your point thru each bootstrapped tree and see what falls out and take the most common vote



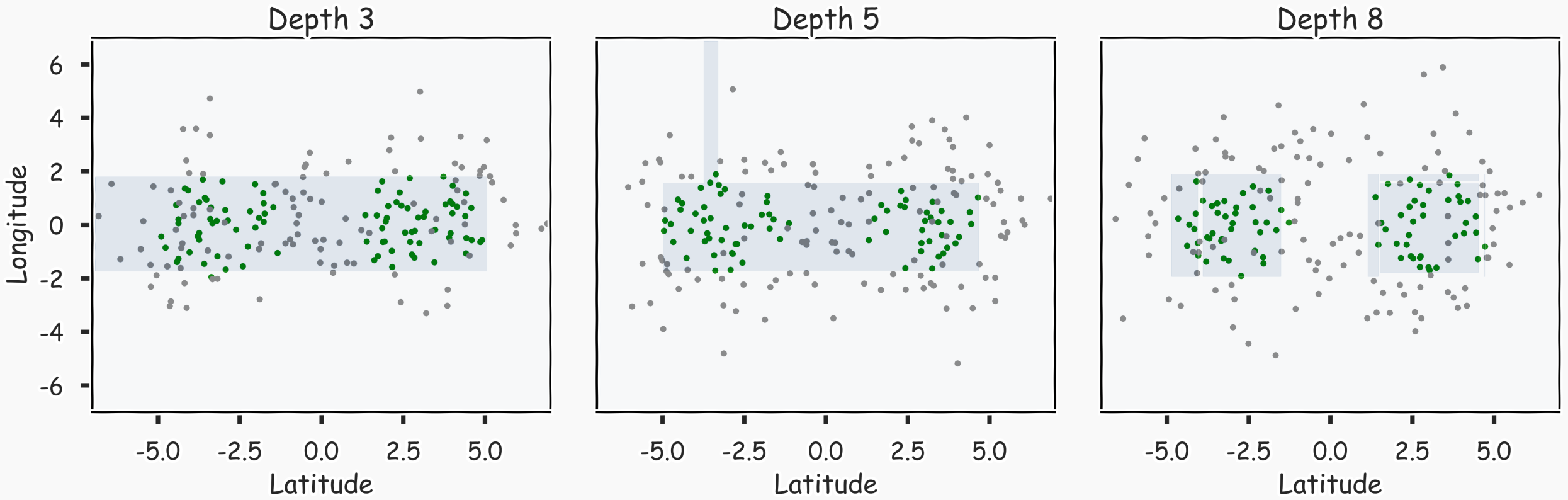
Bagging

Question: Do you see any problems?

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

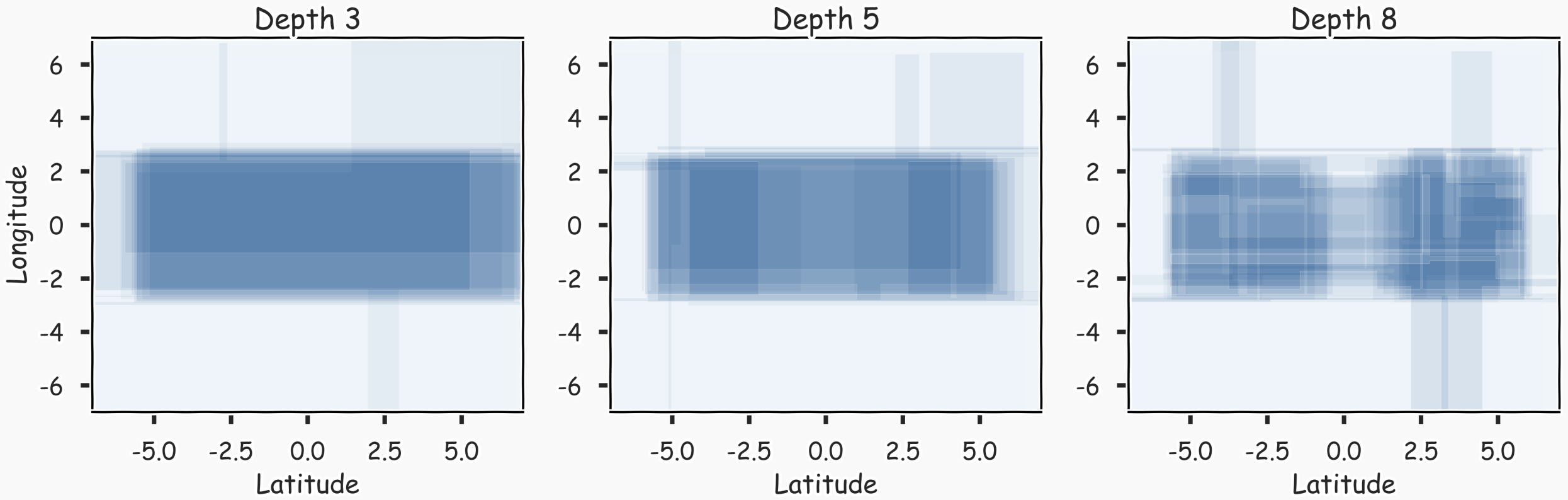
  The **major drawback** of bagging (and other ***ensemble methods*** that we will study) is that the averaged model is no longer easily interpretable - i.e. one can no longer trace the 'logic' of an output through a series of decisions based on predictor values!

Case of underfitting



Case of underfitting

Always underfitting - no matter how many bootstraps



Bagging

Question: Do you see any problems?

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

you are doing CV on the 'BAGGING' model now not just a single tree

Cross Validations

Out-of-Bag Error

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

b/c you sample with replacement, some data do not appear in the bootstrap

Bagging

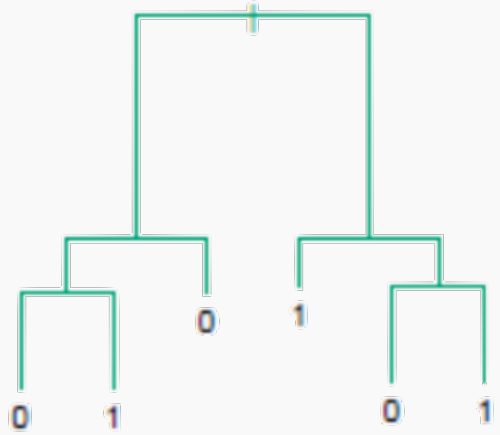
Original Data

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

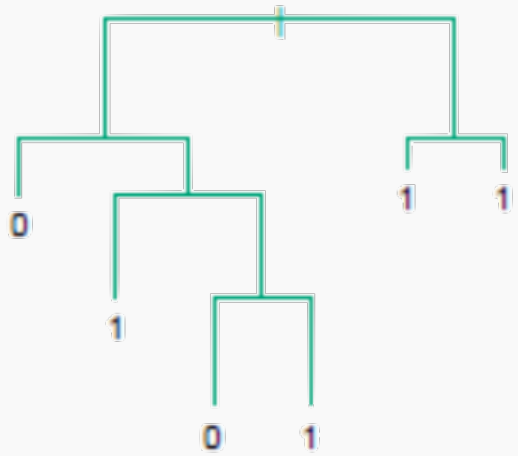
Original Data

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

let's recycle our unused data points!

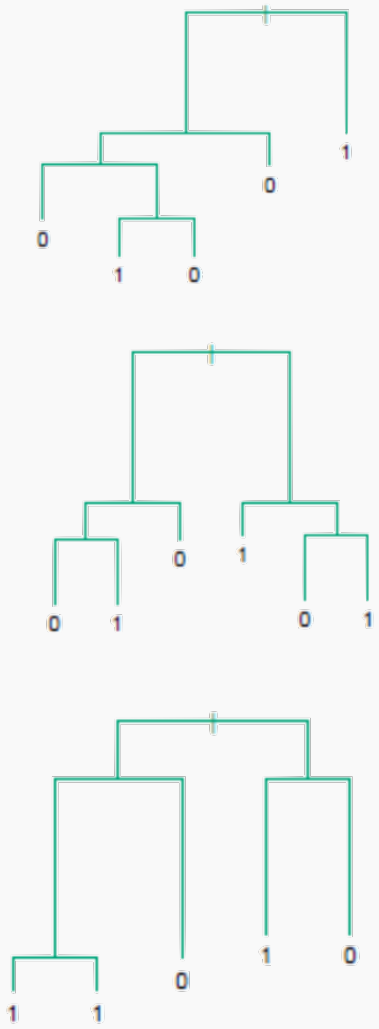
Point-wise out-of-bag error

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

$$\hat{y}_{i,pw} = \text{majority}(\hat{y}_i)$$

make prediction on unused point

misclassification error on unused point

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

Regression

$$\hat{y}_{i,pw} = \sum_{j \in B} \hat{y}_{i,j}$$

MSE on unused point

$$e_i = (y_i - \hat{y}_{i.pw})^2$$

OOB Error

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

Classification

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

Regression

$$Error_{OOB} = \sum_{i}^{n} e_i = \sum_{i}^{n} \left(y_i - \hat{y}_{i,pw}\right)^2$$

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?

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

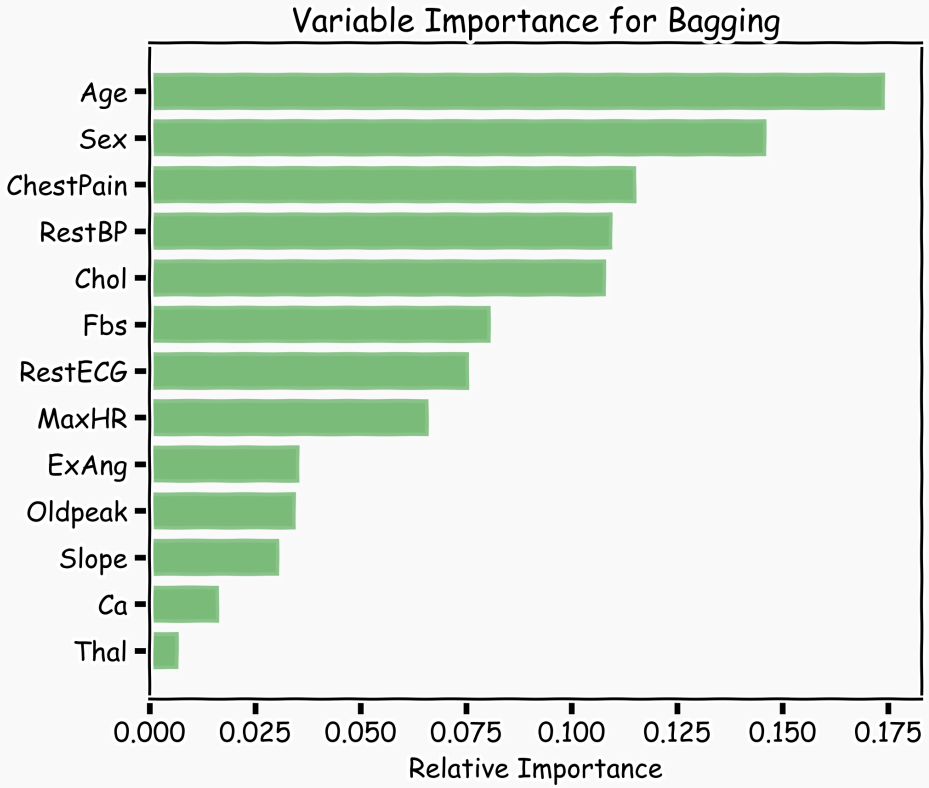
you also run the risk of getting the exact sample draws when you do too many bootstrap samples you just get the same decision tree over and over - pointless

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!

Variable Importance for Bagging

Bagging improves prediction accuracy at the expense of interpretability.

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.



important variables will appear in the majority of the bootstrapped trees

100 trees, max_depth=10

Improving on Bagging

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

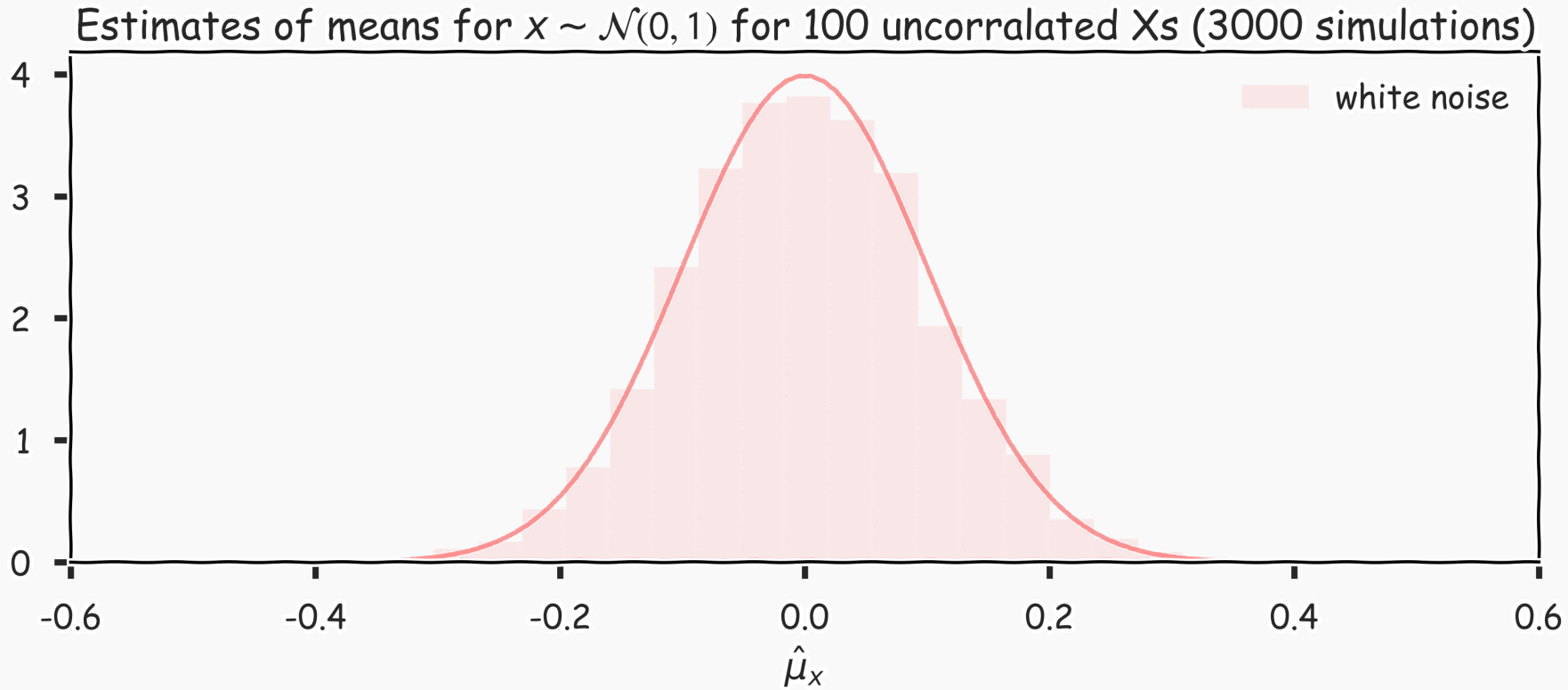
Suppose we have an extremely strong predictor, x_j , in the training set amongst moderate predictors. Then the greedy learning algorithm ensures that most of the models in the ensemble will choose to split on x_j in early iterations.

That is, each tree in the ensemble is 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 :

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

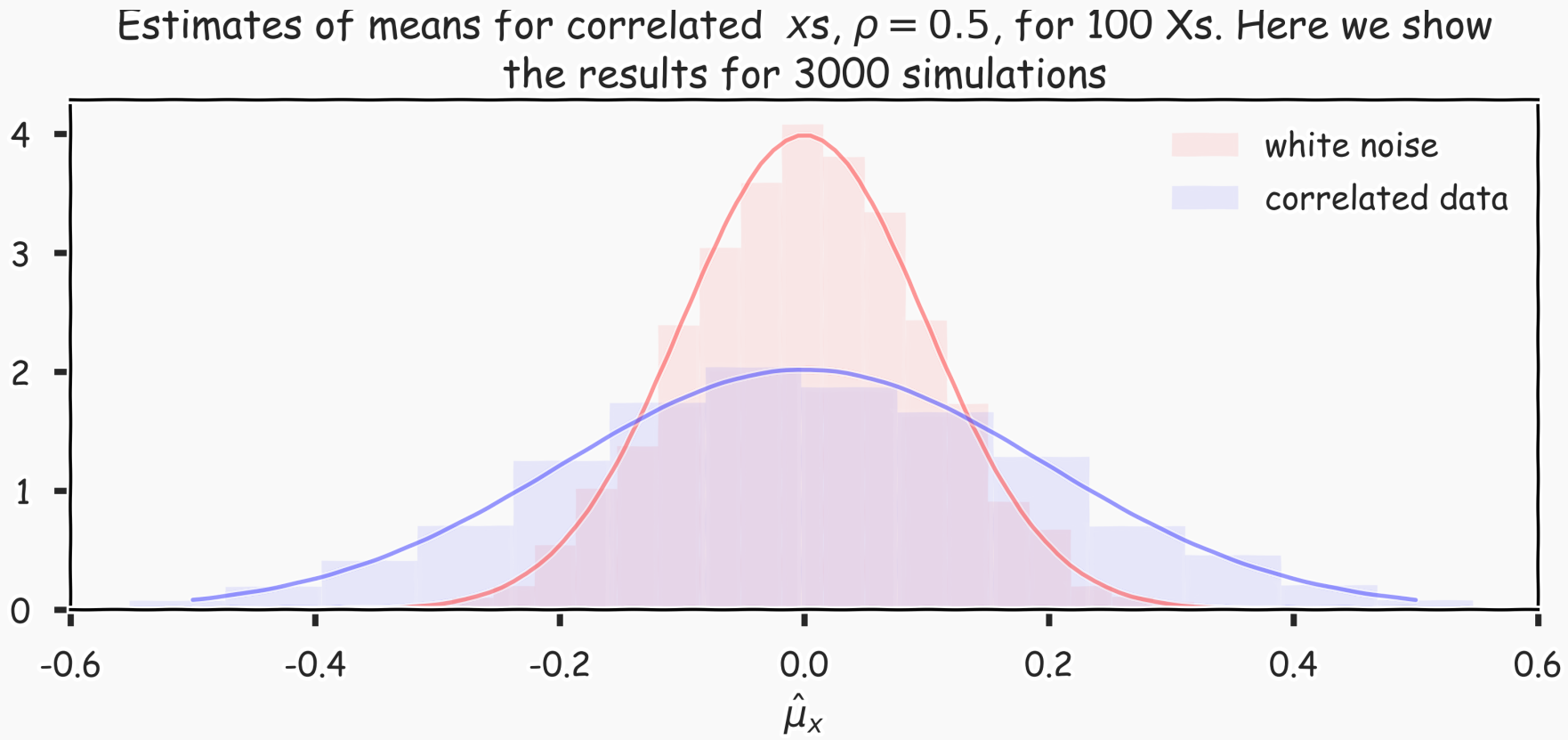


Improving on Bagging

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

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

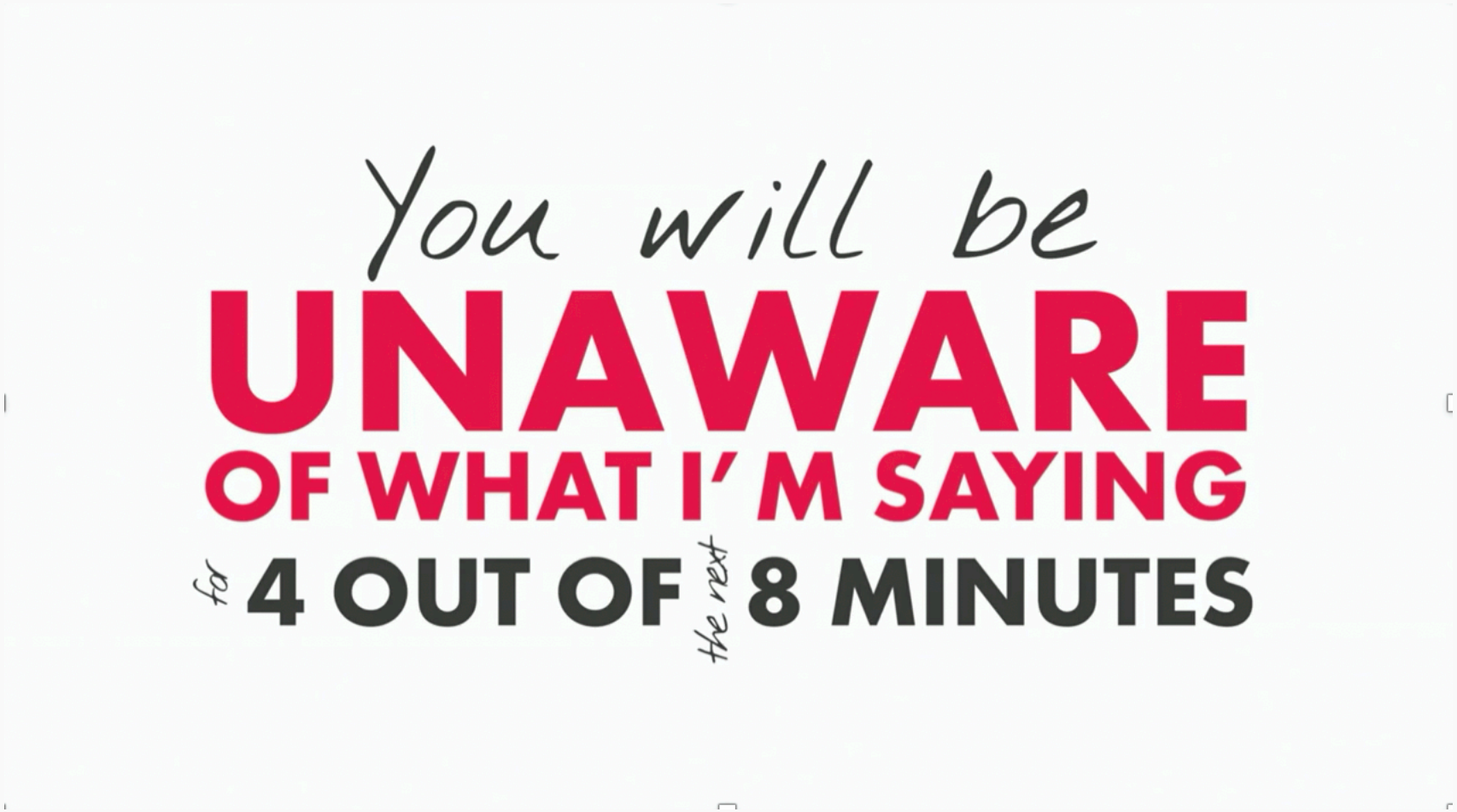
You don't shrink the distribution as well with more bootstraps when data is correlated



Bagging

Question: Do you see any problems?

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



Zeenat Potia

Random Forests

Random Forests

We have a problem of correlated data because the trees look the same due to highly impt predictors

Random Forest is a modified form of bagging that creates ensembles of independent decision trees.

To de-correlate the trees, we:

1. train each tree on a separate bootstrap sample of the full training set (same as in bagging)

2. for each tree, at each split, we ***randomly*** select a set of J' predictors from the full set of predictors.

From amongst the J' predictors, we select the optimal predictor and the optimal corresponding threshold for the split.

Tuning Random Forests

Random forest models have multiple hyper-parameters to tune:

1. the number of predictors to randomly select at each split

2. the total number of trees in the ensemble

   Number of trees doesn't add to complexity of model - no overfitting, you just plateau in how much you minimize variance

3. the minimum leaf node size

In theory, each tree in the random forest is full, but in practice this can be computationally expensive (and added redundancies in the model), thus, imposing a minimum node size is not unusual.

Tuning Random Forests

There are standard (default) values for each of random forest hyper-parameters recommended by long time practitioners, but generally these parameters should be tuned through **OOB** (making them data and problem dependent).

e.g. number of predictors to randomly select at each split:

- $\sqrt{N_j}$ for classification
- $\frac{N}{3}$ for regression

Using out-of-bag errors, training and cross validation can be done in a single sequence - we cease training once the out-of-bag error stabilizes

Variable Importance for RF

Same as with Bagging:

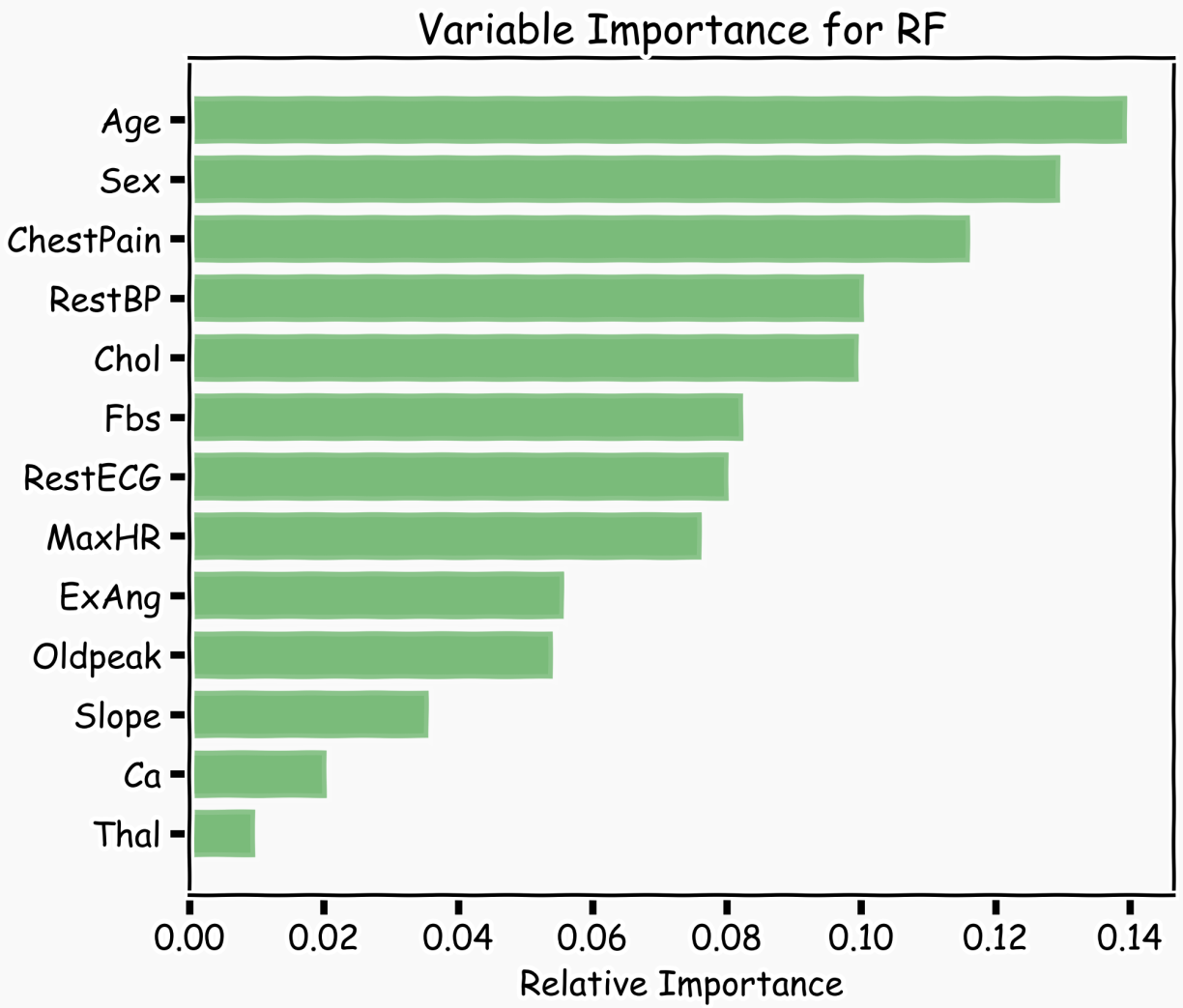
Calculate the total amount that the RSS (for regression) or Gini index (for classification) is decreased due to splits over a given predictor, averaged over all B trees.

Variable Importance for RF

Alternative: Pavlos' method - no one cares?

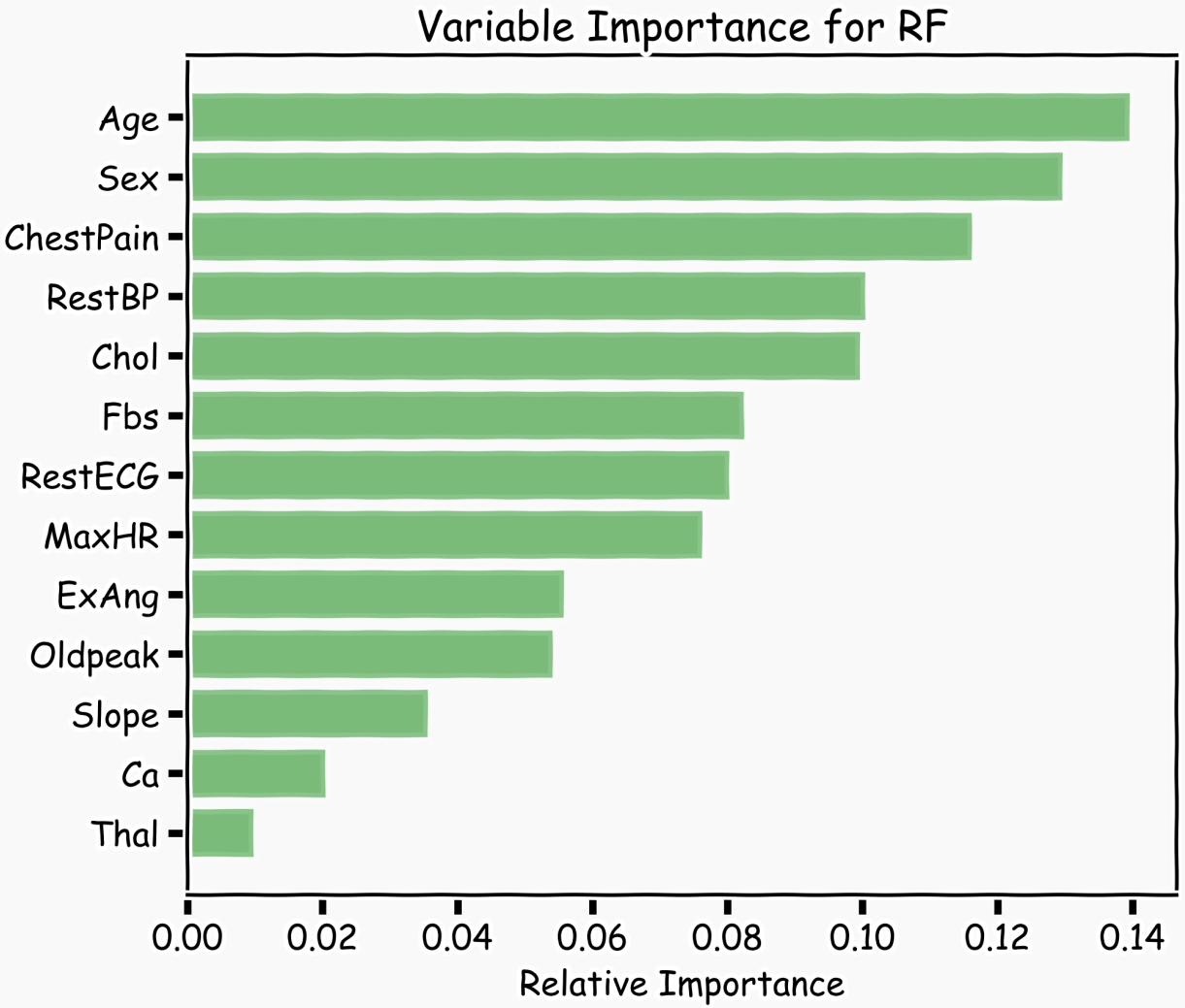
- Record the prediction accuracy on the *oob* samples for each tree.
- Randomly permute the data for column j in the *oob* samples the record the accuracy again.
- The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable j in the random forest.

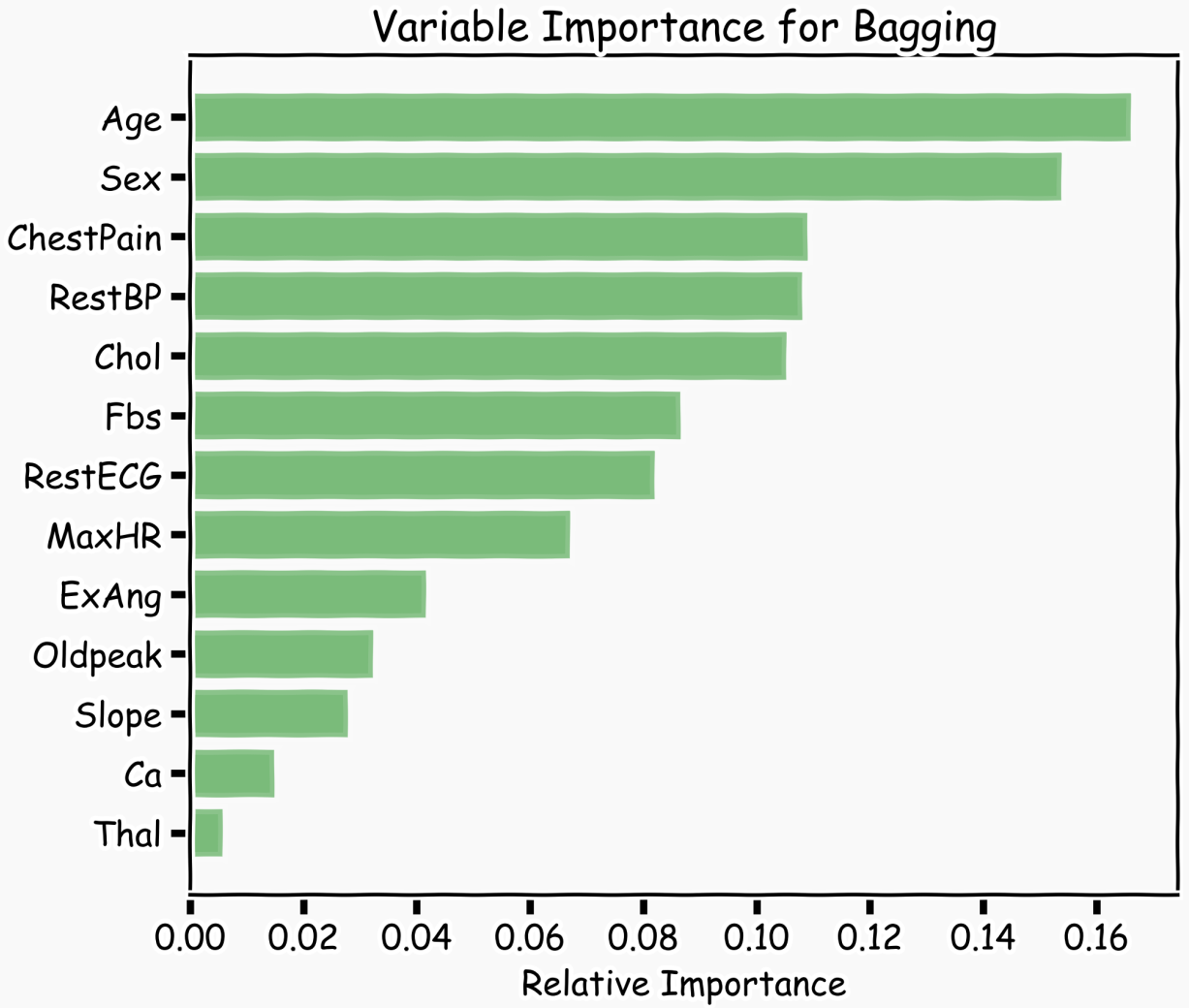
Variable Importance for RF



100 trees, max_depth=10

Variable Importance for RF





with random forests, variable importance is diminished because not all predictors show up in every tree

100 trees, max_depth=10

Final Thoughts on Random Forests

When the number of predictors is large, but the number of relevant predictors is small, random forests can perform poorly.

Question: Why?

In each split, the chances of selected a relevant predictor will be low and hence most trees in the ensemble will be weak models.

Final Thoughts on Random Forests (cont.)

Increasing the number of trees in the ensemble generally does not increase the risk of overfitting.

Again, by decomposing the generalization error in terms of bias and variance, we see that increasing the number of trees produces a model that is at least as robust as a single tree.

However, if the number of trees is too large, then the trees in the ensemble may become more correlated, increase the variance.

Final Thoughts on Random Forests (cont.)

Probabilities:

- Random Forrest Classifier (and bagging) can return probabilities.
- **Question**: How?

Next Lecture

- Unbalance dataset
- Weighted samples
- Categorical data
- Missing data
- Different implementations

AND BOOSTING