

Today

- Ensemble methods
 - Bagging
 - Random forest

Random forest

Algorithm

for many repetitions

- randomly select m predictor variables

- resample the data (rows) with replacement

- train the tree model

- record prediction

final prediction = mean of predictions

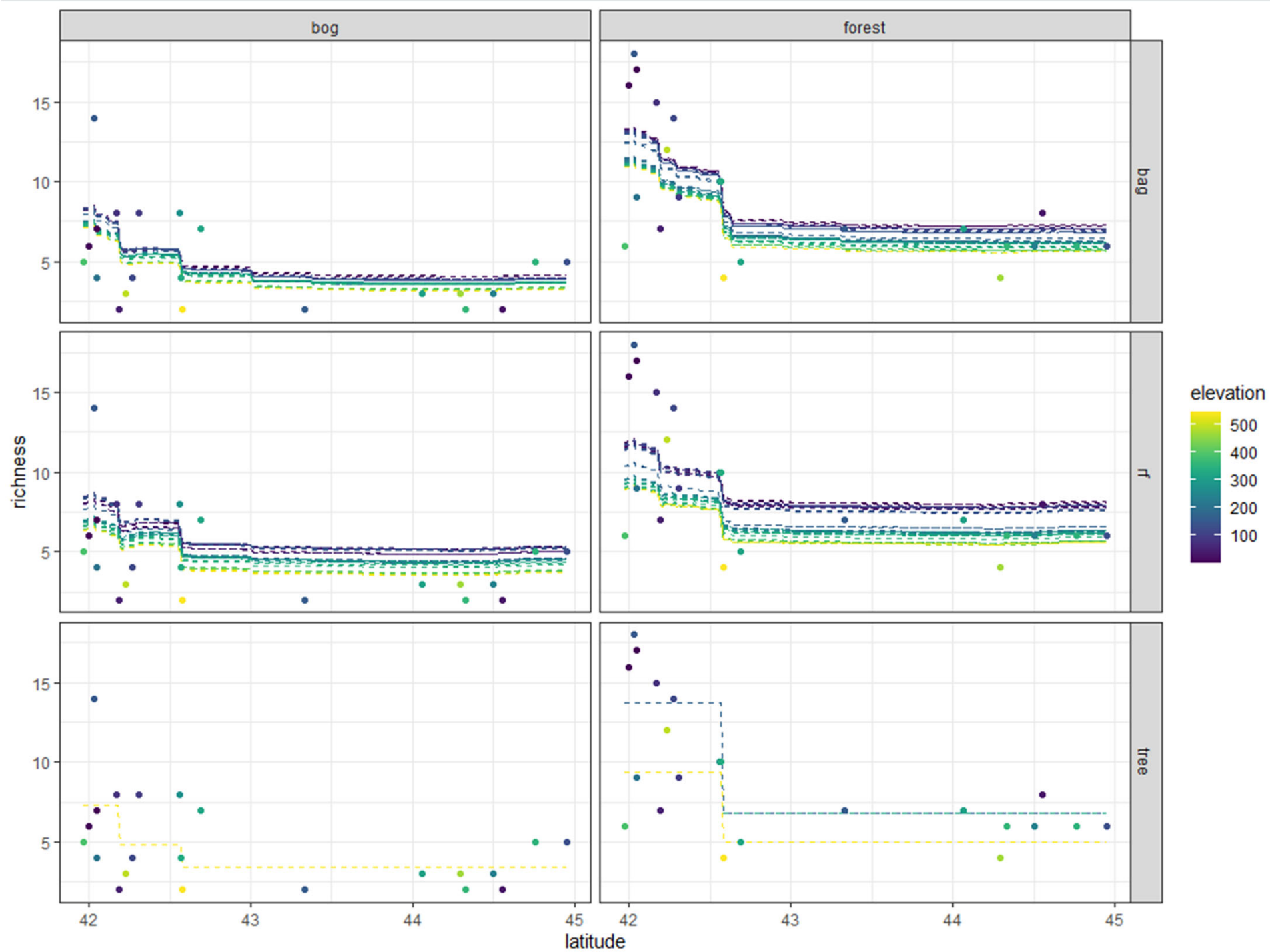
Random forest as R

```
# Parameters
m <- 2 #Number of predictors to sample at each iteration
boot_reps <- 500

# Setup
n <- nrow(ants)
c <- ncol(ants)
nn <- nrow(grid_data)
boot_preds <- matrix(rep(NA, nn*boot_reps), nrow=nn, ncol=boot_reps)

# Main algorithm
for ( i in 1:boot_reps ) {
  # randomly select m predictor variables
  predictor_indices <- sample(2:c, m)
  boot_data <- ants[,c(1,predictor_indices)]
  # resample the data (rows) with replacement
  boot_indices <- sample(1:n, n, replace=TRUE)
  boot_data <- boot_data[boot_indices,]
  # train the tree model
  boot_fit <- tree(richness ~ ., data=boot_data)
  # record prediction
  boot_preds[,i] <- predict(boot_fit, newdata=grid_data)
}
rf_preds <- rowMeans(boot_preds)
```

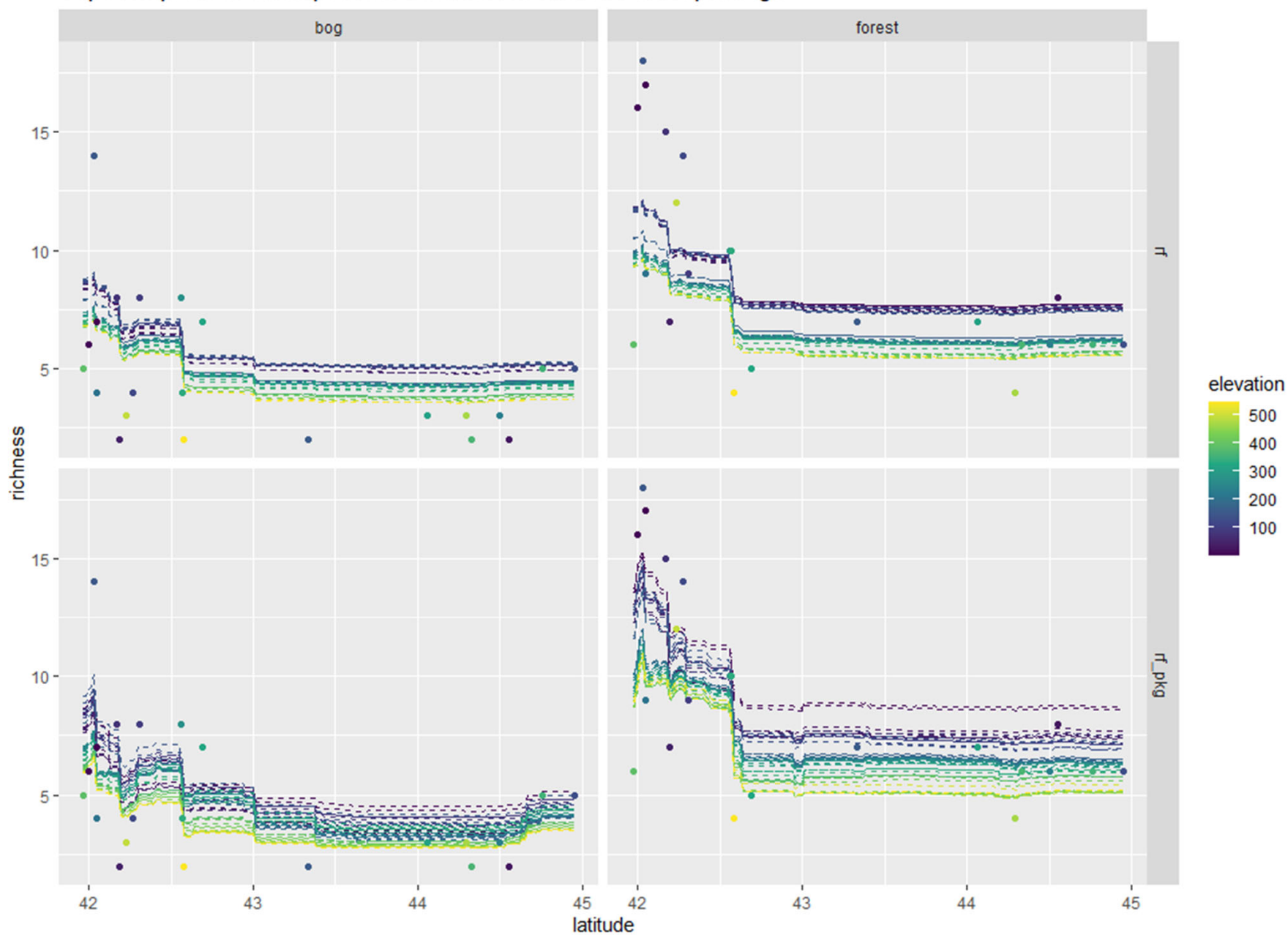
} the new bit



R packages

- randomForest
 - original Breimen (2001) algorithm
 - Fortran
- ranger
 - fast implementation
 - C++

Top row: proof of concept code. Bottom row: randomForest package



```
randomForest(richness ~ ., data=ants, ntree=500, mtry=2)
```

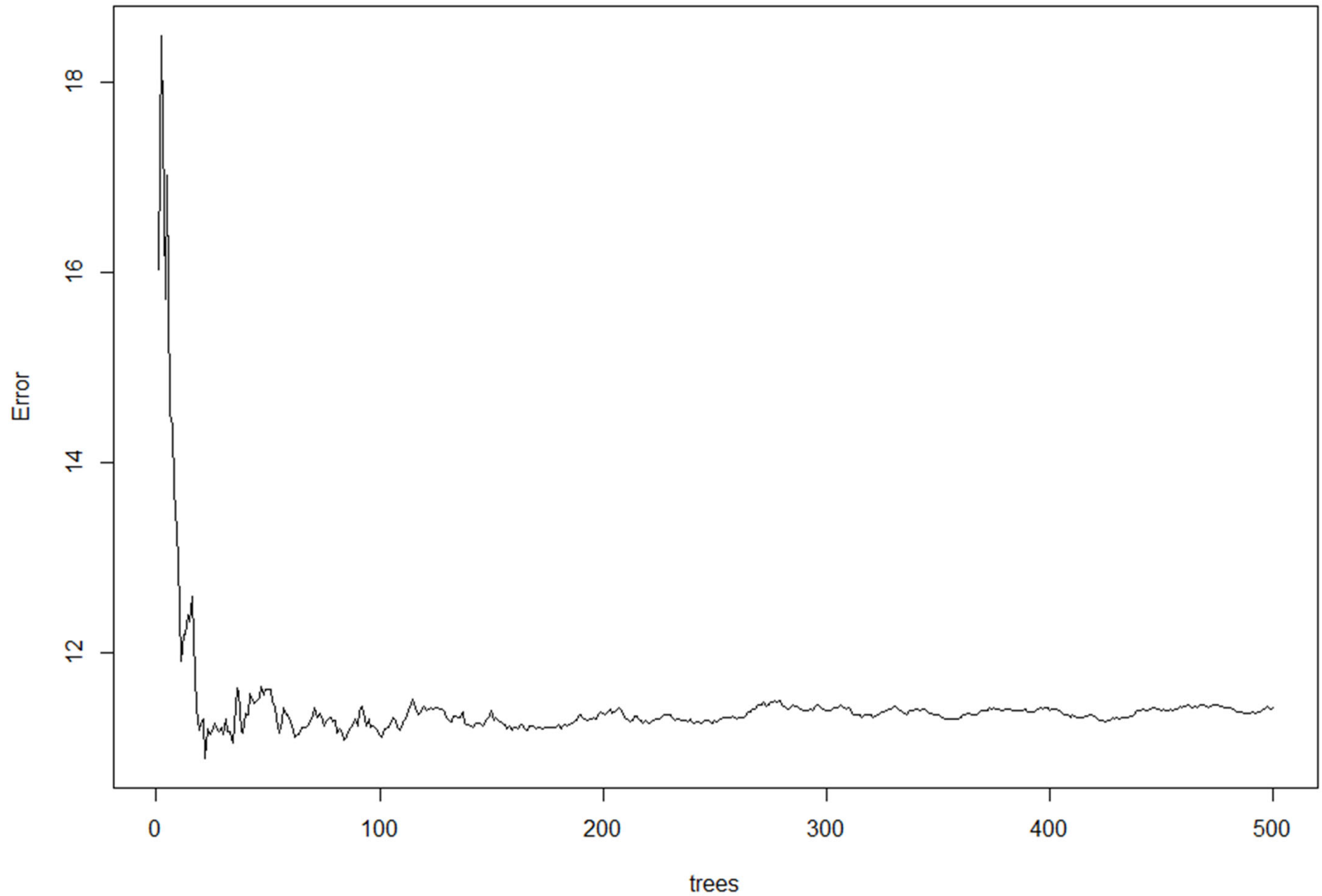
Inference algorithm

- k-fold CV
 - most general strategy
 - expensive, as we've seen
 - use to compare with other models
- Out-of-bag estimate
 - specific to bagging and random forest
 - can use for tuning number of trees and number of models to try

“Out of bag” error estimate

- Out of bag samples
 - data not included in a bootstrap sample
 - on average ca $1/3$ are out of bag
- Each bootstrap, use the out of bag samples to gauge prediction error
 - average error across trees within a forest
- Approx equal to LOOCV ($n_trees = large$)
- Computationally efficient
 - part of the bagging step

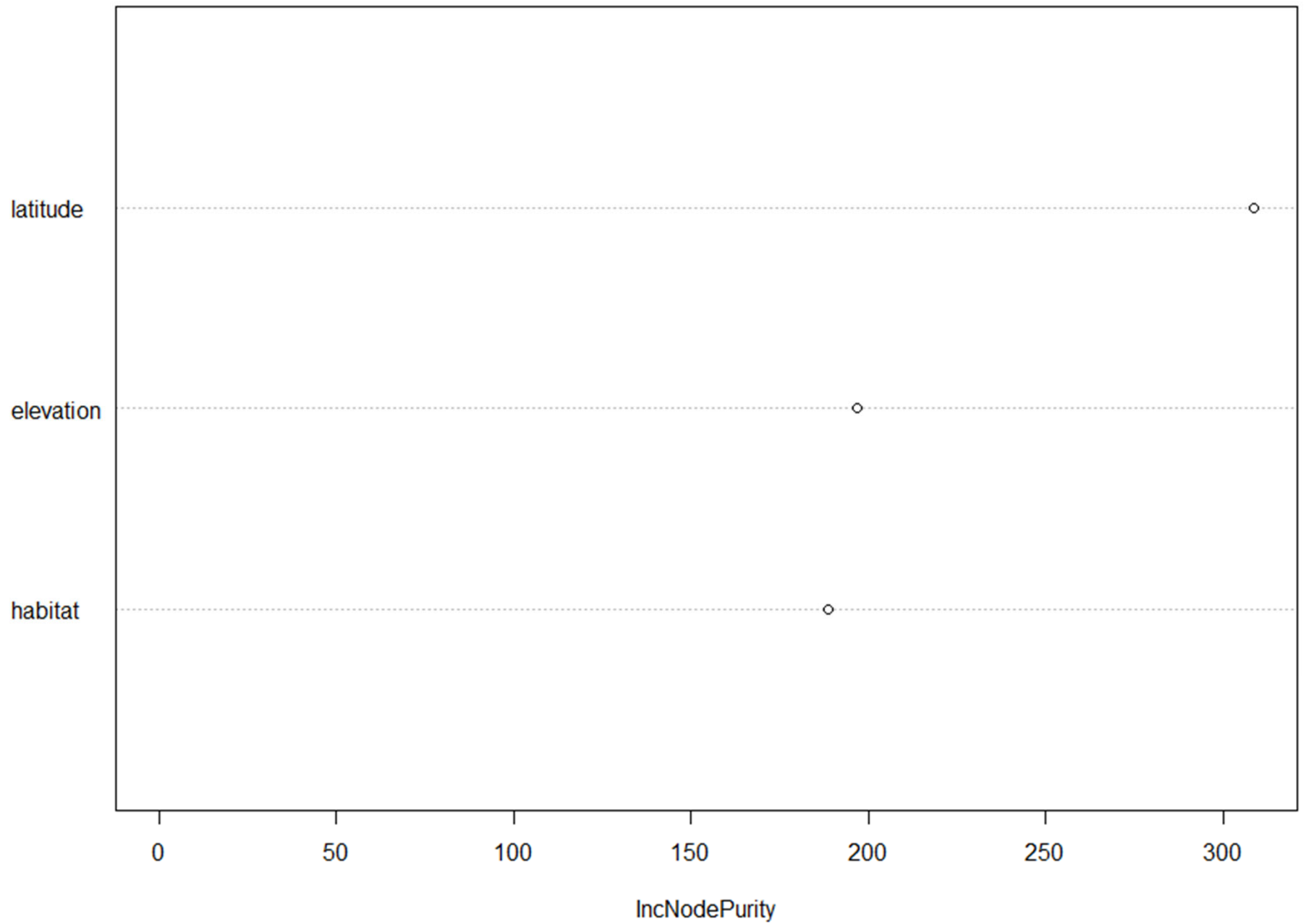
Tuning number of trees with OOB error: ants data



Variable importance

- Average RSS decrease over branch splits for each variable
- Interpretation:
 - more reduction in RSS = more “important”
- Similar to regression concept of “explaining more variation”
- Advantage: explainable machine learning

Variable importance: ants data



Bagging is a special case

bagging: $mtry = \text{number of variables}$

