- Random forest models have multiple hyper-parameters to tune
  - the number of predictors to randomly select at each split
  - the total number of trees in the ensemble
  - the minimum leaf node size
- Generally tuned through cross validation ($\implies$ data and problem dependent)
- Use **out-of-bag errors** to evaluate model's predictive accuracy
  - cease training once the out-of-bag error stabilizes
  - if sample large enough, estimate is approximately LOO-CV error for bagging
- When the number of predictors is large, but the number of relevant predictors is small, random forests can perform poorly
- Increasing number of trees in ensemble does not increase risk of overfitting
  - but the trees in the ensemble may become more correlated, increase the variance.

# Tuning the Random Forest

# Tuning the Random Forest

- Random forest models have multiple hyper-parameters to tune
  - the number of predictors to randomly select at each split
  - the total number of trees in the ensemble
  - the minimum leaf node size
- Generally tuned through cross validation ($\implies$ data and problem dependent)
- Use **out-of-bag errors** to evaluate model's predictive accuracy
  - cease training once the out-of-bag error stabilizes
  - if sample large enough, estimate is approximately LOO-CV error for bagging
- When the number of predictors is large, but the number of relevant predictors is small, random forests can perform poorly
- Increasing number of trees in ensemble does not increase risk of overfitting
  - but the trees in the ensemble may become more correlated, increase the variance.

# Out-Of-Bag Errors (OOB Errors)



data  bootstrapped samples  decision trees  unused data (OOB)

bag 1

bag 2

bag 3

point-wise OOB error over full training set:
- classification → majority
- regression → average