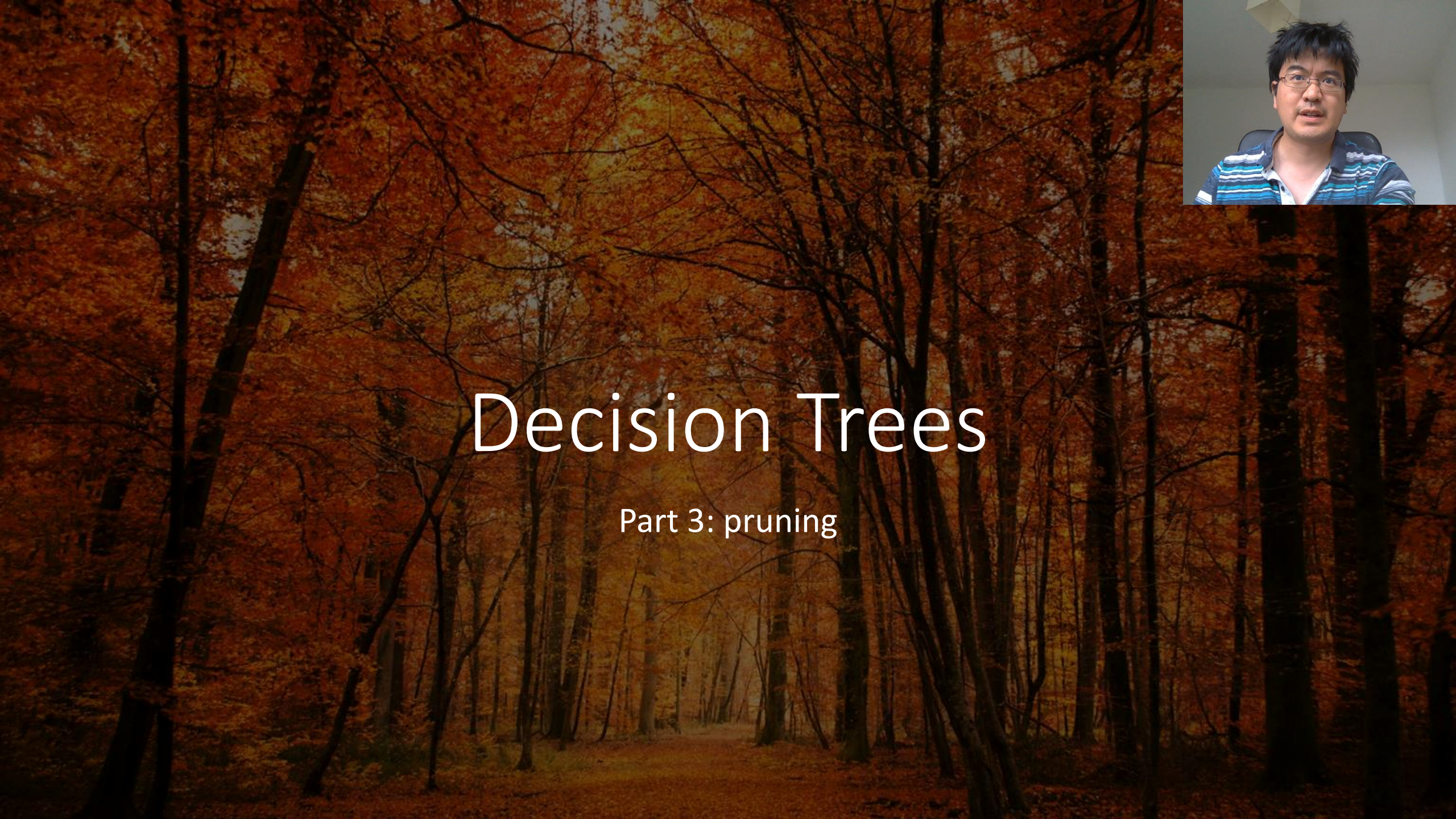
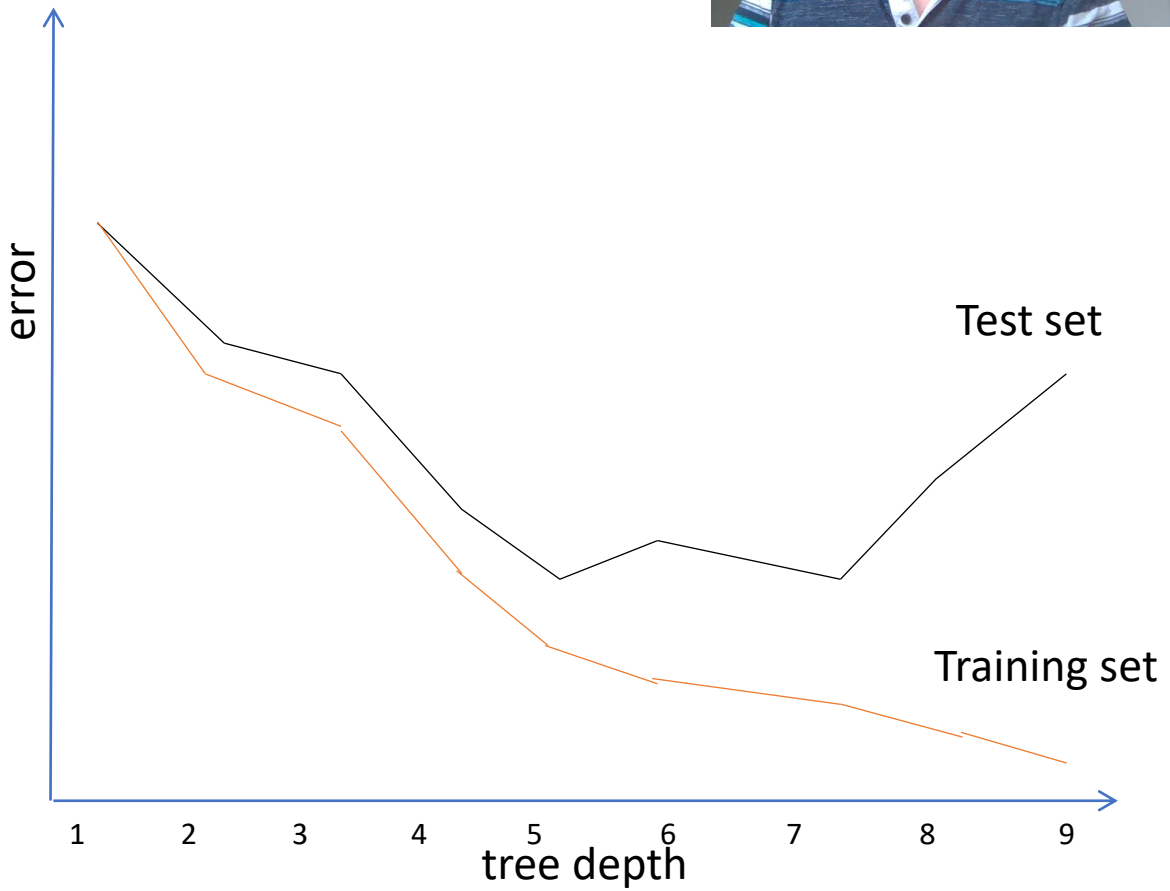


Decision Trees

Part 3: pruning

What is Pruning?

Modifying a decision tree so that is has **poorer** performance on a training data set so that it has **better** performance on a validation/test set



Pre-pruning (Early Stopping)

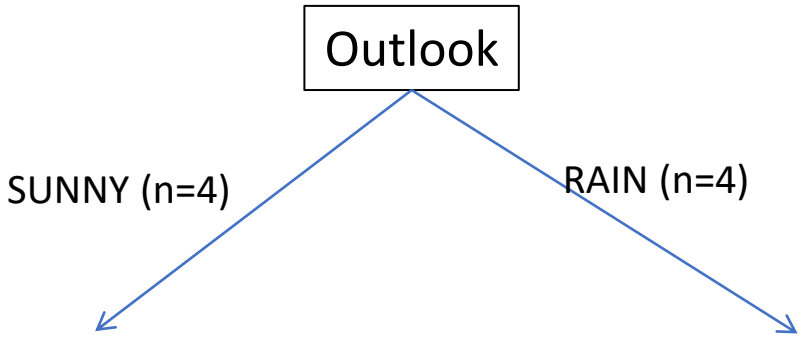
Idea – as we train the decision tree, decide whether it is ‘worth’ adding in an extra decision

Early stopping using Information Gain

- Alternative: stop splitting when information gain is below a given threshold
- Example:

Sunny:

humidity	wind	Tennis?
High	Low	yes
high	Low	No
normal	High	Yes
normal	high	No



Early stopping using Information Gain

For the ‘sunny’ branch:

P(tennis| sunny): $\frac{4}{8} = 0.5$,

P(tennis|sunny, high): $\frac{1}{2} = 0.5$

P(tennis|sunny, normal): $\frac{1}{2} = 0.5$

Entropy before: $0.5(0.5\log(0.5)) + 0.5(0.5\log(0.5))$

Entropy after ‘high’: $0.5(0.5\log(0.5)) + 0.5(0.5\log(0.5))$

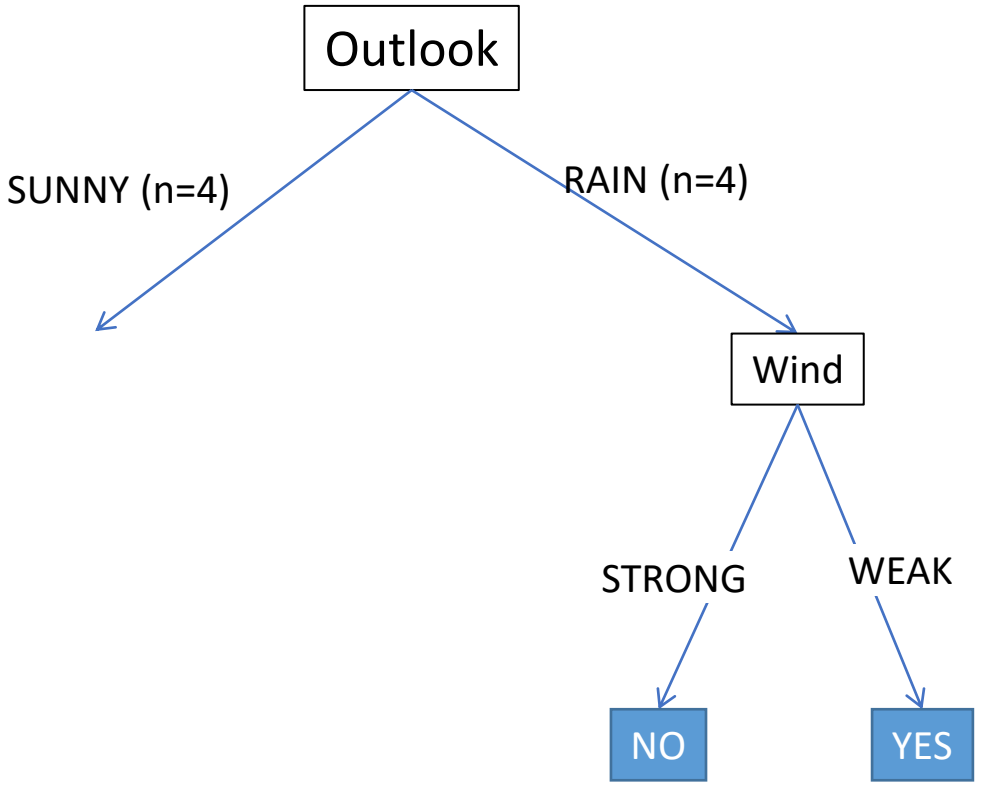
Information gain: 0

Early stopping using Information Gain

- Extra split is non-informative, so end this branch

Sunny:

humidity	wind	Tennis?
High	Low	yes
high	Low	No
normal	High	Yes
normal	high	No



Pros and Cons

Pros

- Does not require validation data
- Fast to train – no unnecessary branches are created

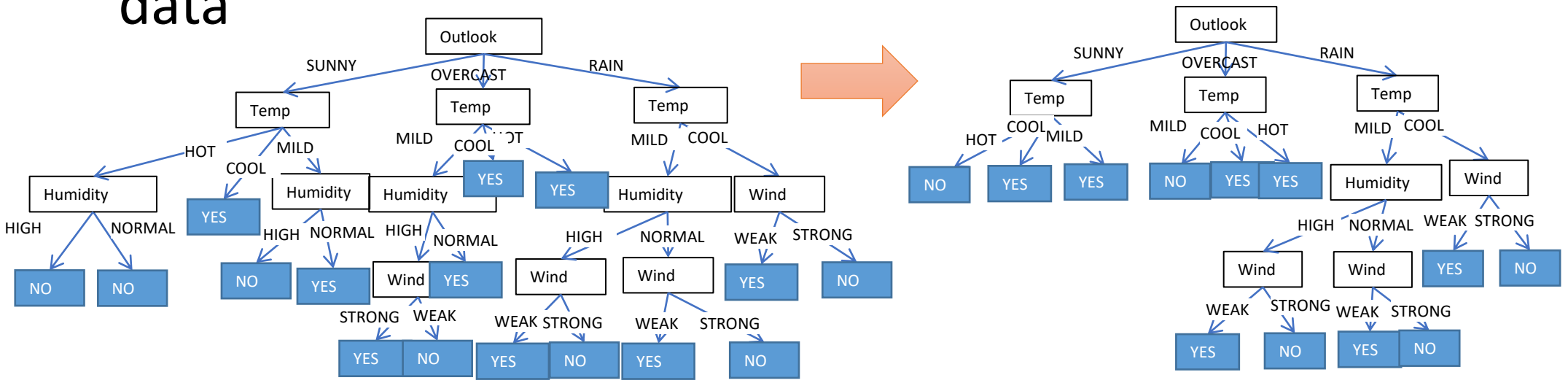
Cons

- No guarantee that we get the best answer
  - Possible for one split to have low information gain, but subsequent splits to have high information gain.

Post-pruning via cross validation

Idea:

- Allow the decision tree to overfit to the training data (perfect classifier)
- Prune back the tree so that it still performs well on validation data

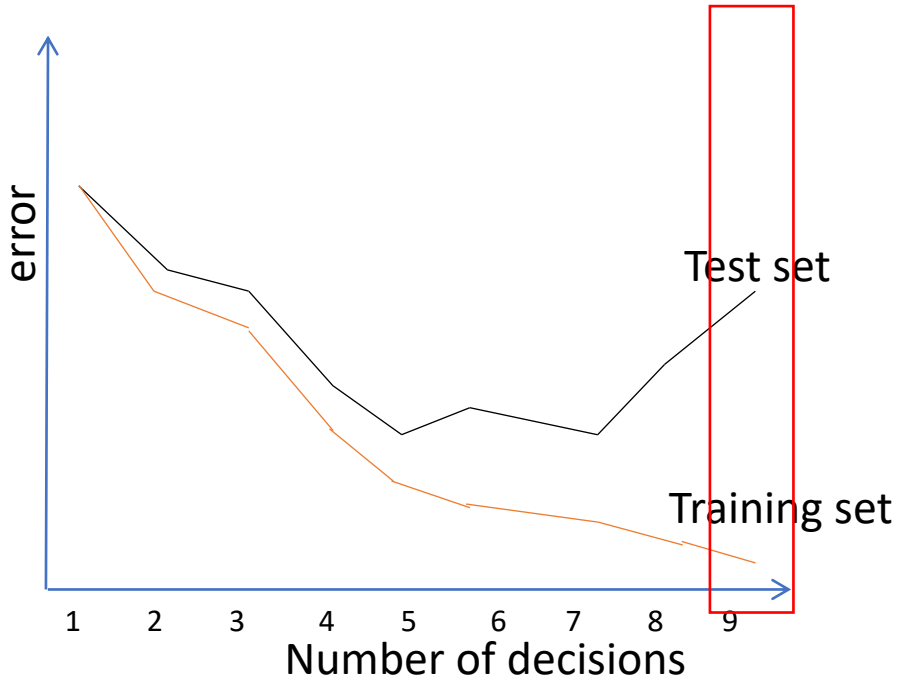


Post-pruning

Train	Validate	Test

1.) Build tree using training data
2.) Keep adding decisions until all tree is as accurate as possible

Result: complex tree, with high accuracy on the training set, but low accuracy on a validation set

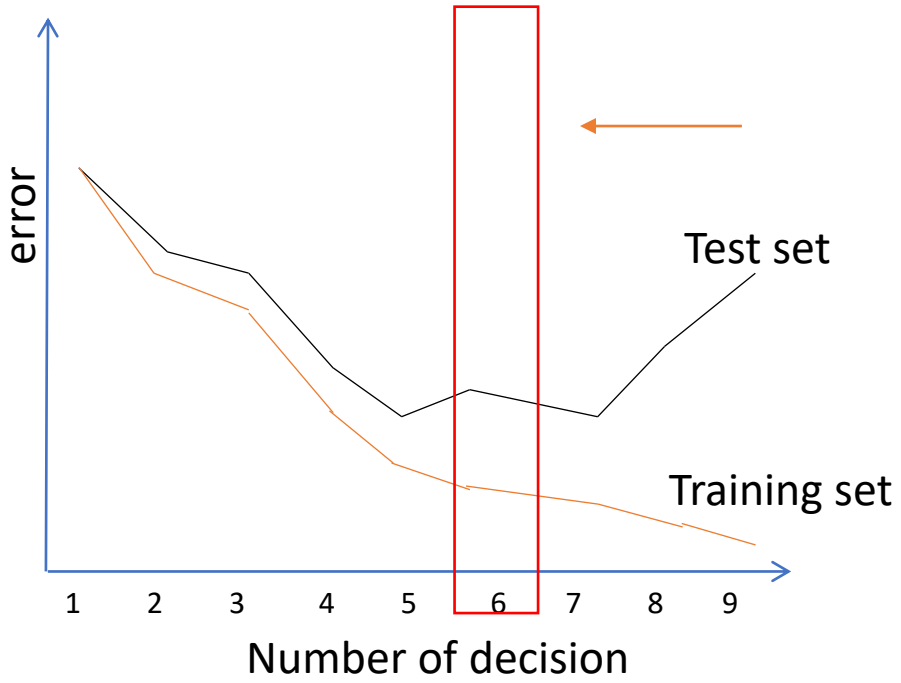


Post-pruning

Train	Validate	Test

1.) Remove a branch, starting at the bottom of the tree
2.) Calculate validation set error
3.) If validation set error has decreased, goto step 1
4.) Otherwise,

Result: complex tree, with high accuracy on the training set, but low accuracy on a validation set



Post-pruning

Pros

- Possible to guarantee the ‘best’ answer (but still very difficult for large trees)

Cons

- Two-stage process, so slower than pre-pruning
- Typically requires validation data (which may mean that it is not possible for smaller datasets)

Summary

- Pruning – method of editing a decision tree so that it gives better generalisability

- Two main types of methods
  - Pre-pruning (early stopping)
    - Fast
    - Works well in most cases (but not always)
  - Post-pruning
    - Analagous to hyperparameter optimisation for other ML classifiers
    - Requires computation of ‘unnecessary’ branches