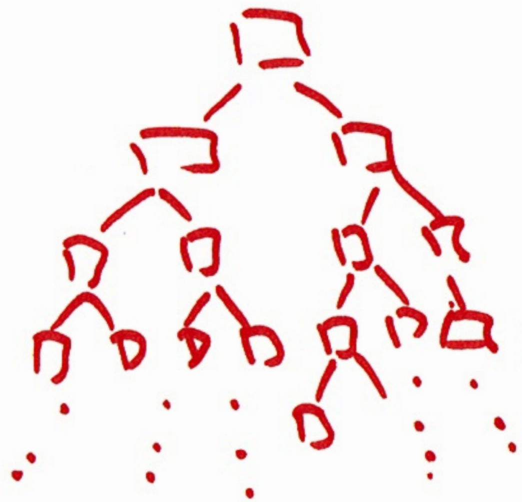
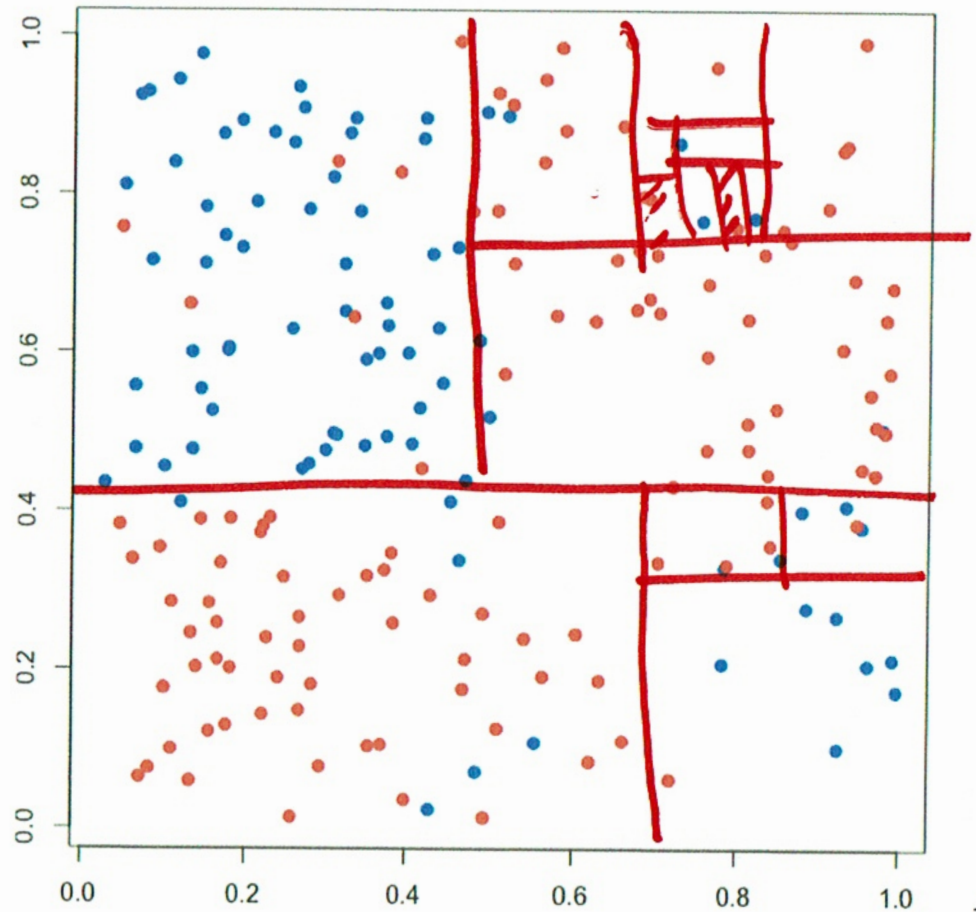
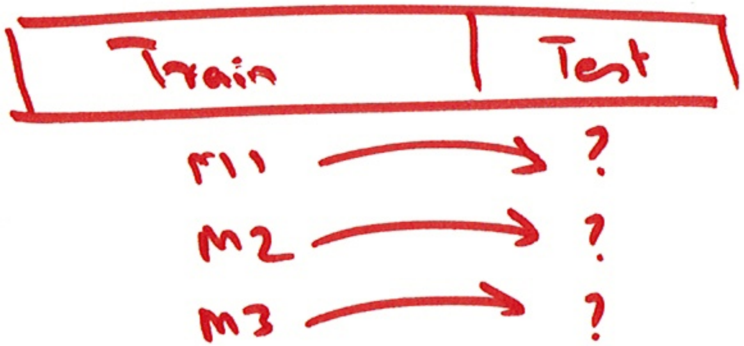
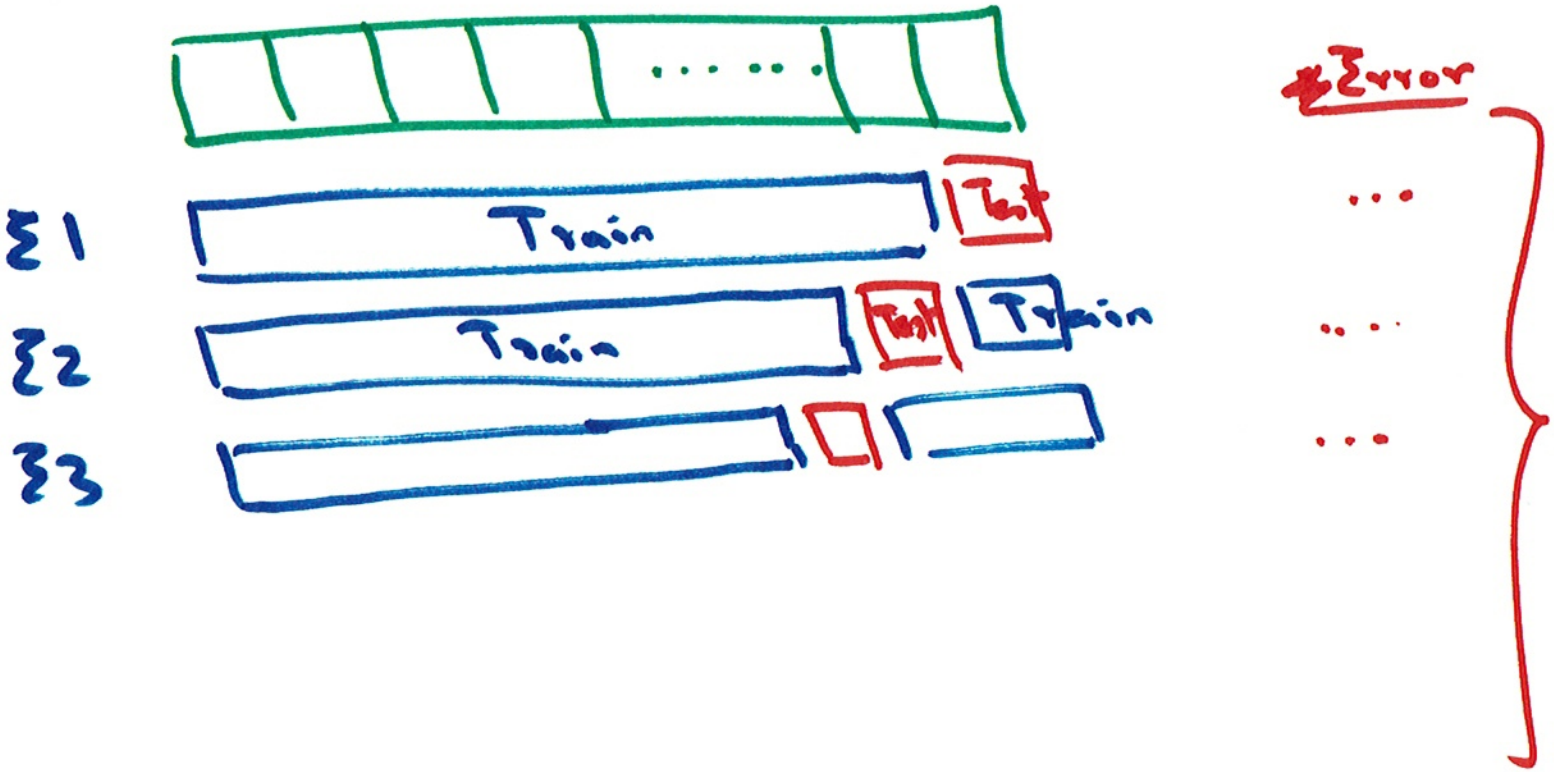
Decision trees are prone to 'overfitting'

- Decision Tree is a powerful algorithm that can adapt well and capture various patterns in the data
- If allowed to grow fully, they become over-complex & tend to fit even the noise
- Thus, a fully grown tree may not 'generalize' well on test or new unseen data






Train
Test
M1
M2
M3
?
?
?


Σ Error
Σ1
Σ2
Σ3
Train
Test
Train
Test
Train

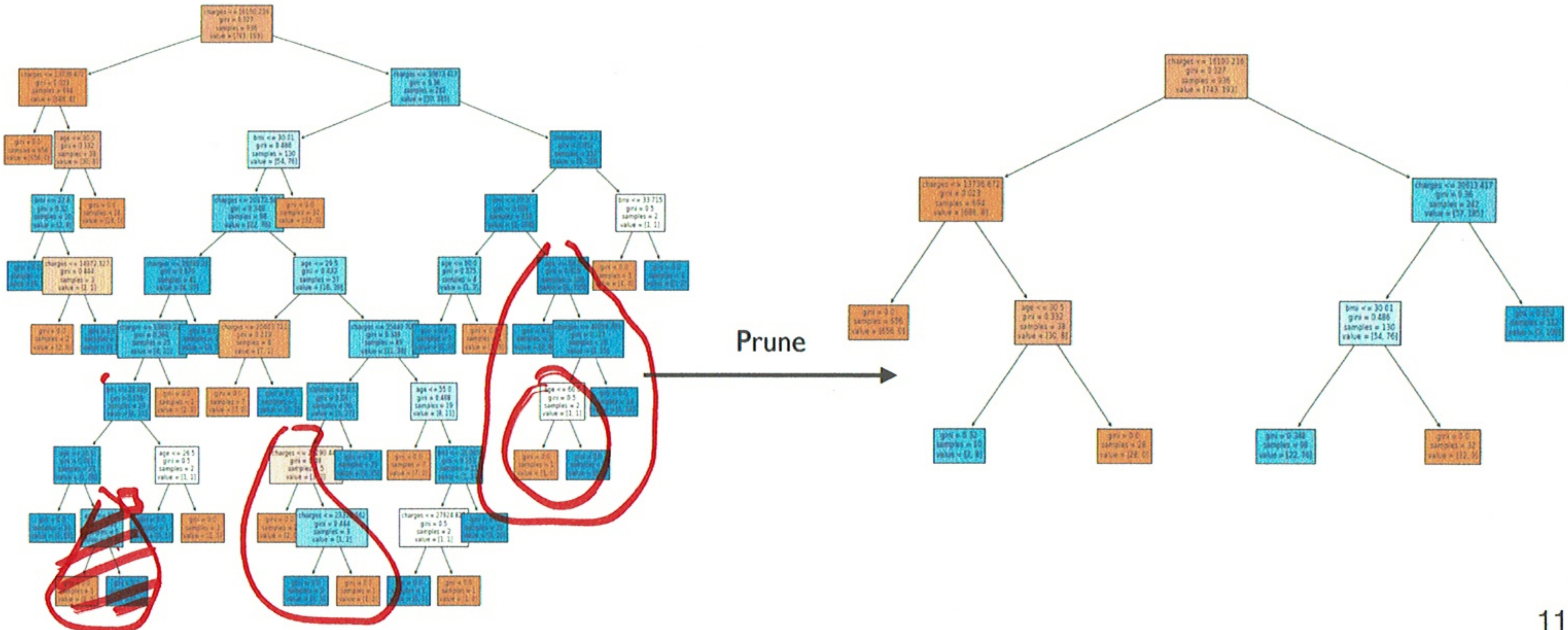
Post-Pruning: Cost-complexity pruning

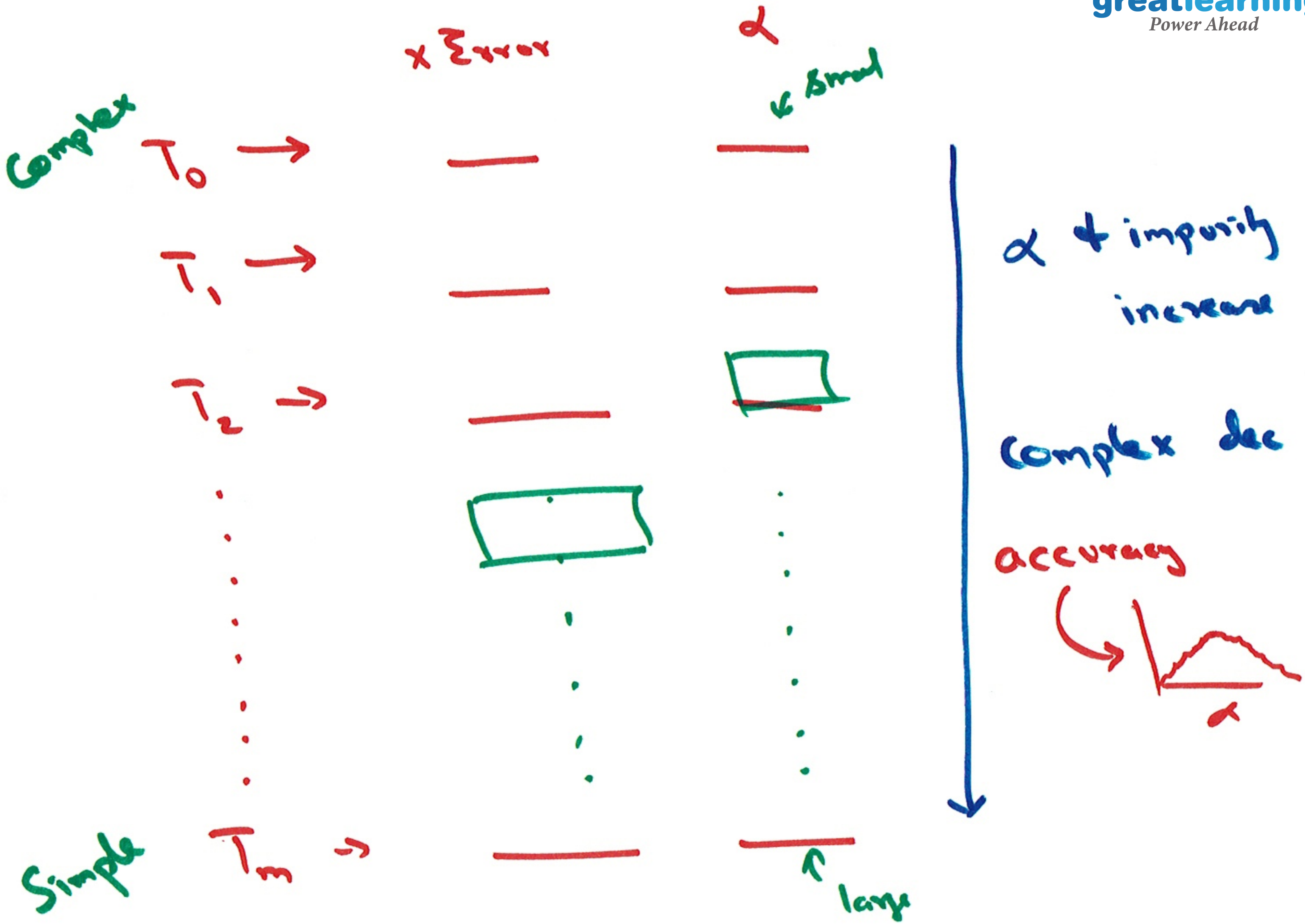
T_0

- Starting from the Full tree, create a sequence of trees that are sequentially smaller (pruned)
- At each step the algorithm
  - try removing each possible subtree
  - find the 'relative error decrease per node' for that subtree - Complexity parameter, α
  - And remove the subtree with the minimum α
- With the list of subtrees, one usually reverts back to using cross-validation errors to find the best final pruned tree

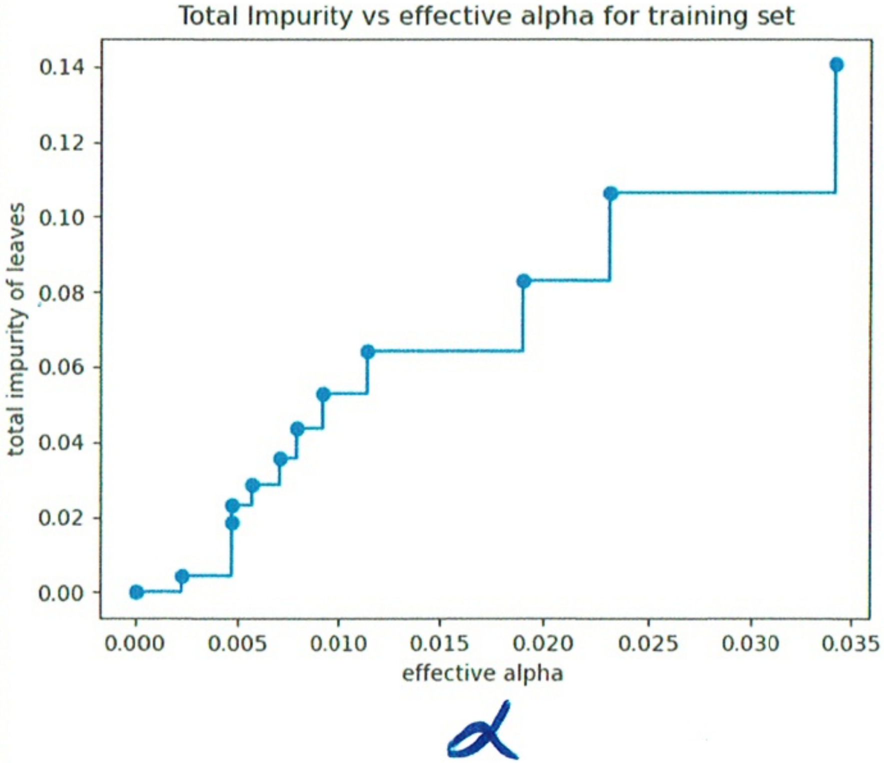
Pruning

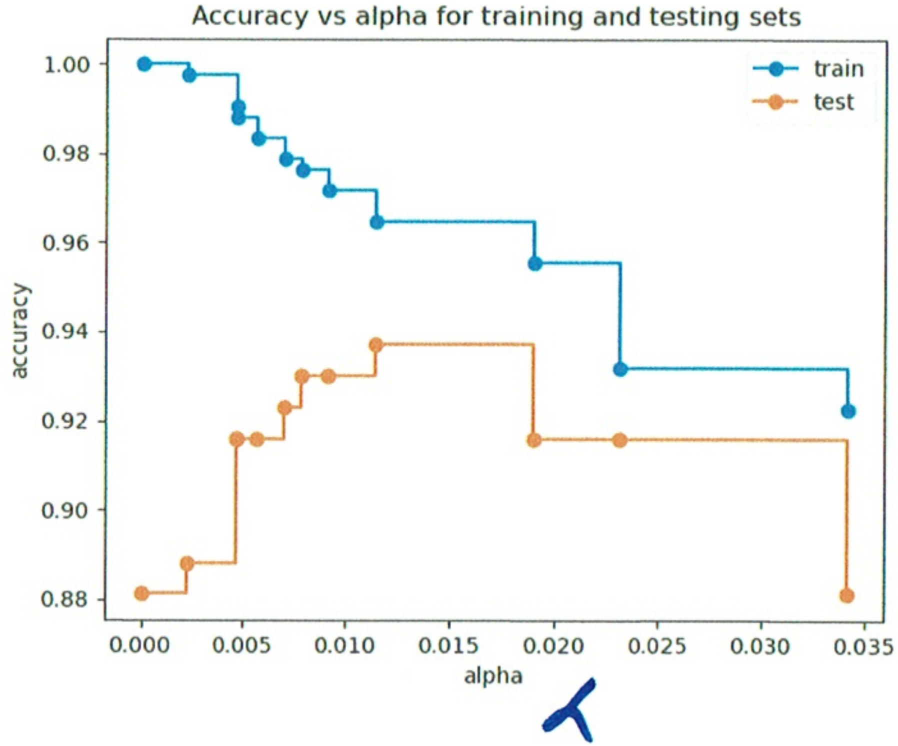
- Ideally we would like a tree that does not over-fit the given data
- One popular and simple way to prune a decision tree is by limiting the depth of the tree to avoid over fitting.
- For example the tree on the right below is generated with a max depth of 2 while the tree on the left has no depth restriction (and hence overfits the data)




x Σ Error
α
Small
Complex
T_0
T_1
T_2
T_m
Simple
large
α ↑ impurity increase
complex dec
accuracy
α

$$\alpha = \frac{\Sigma \text{Error (Pruned)} - \Sigma \text{Error (Org)}}{\text{no. of nodes reduced}}$$


Total Impurity vs effective alpha for training set
total impurity of leaves
effective alpha
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035


Accuracy vs alpha for training and testing sets
accuracy
alpha
train
test
0.88
0.90
0.92
0.94
0.96
0.98
1.00
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035

Impurity Measures in Decision Trees

greatlearning
Power Ahead

	GINI INDEX	ENTROPY	INFORMATION GAIN	VARIANCE
When to use	Classification	Classification	Classification	Regression
Formula	$1 - \Sigma p_i^2$	$-\Sigma p_i \log(p_i)$	$\underbrace{E(Y)}_{P} - \underbrace{E(Y\|X)}_{C}$	$\Sigma(x - \bar{x})^2/N$
Range	0 to 0.5 0 = most pure 0.5 = most impure	0 to 1 0 = most pure 1 = most impure	0 to 1 0 = less gain 1 = more gain	>=0
Characteristics	Easy to compute Non-additive	Computationally intensive Additive	Computationally intensive	The most common measure of spread