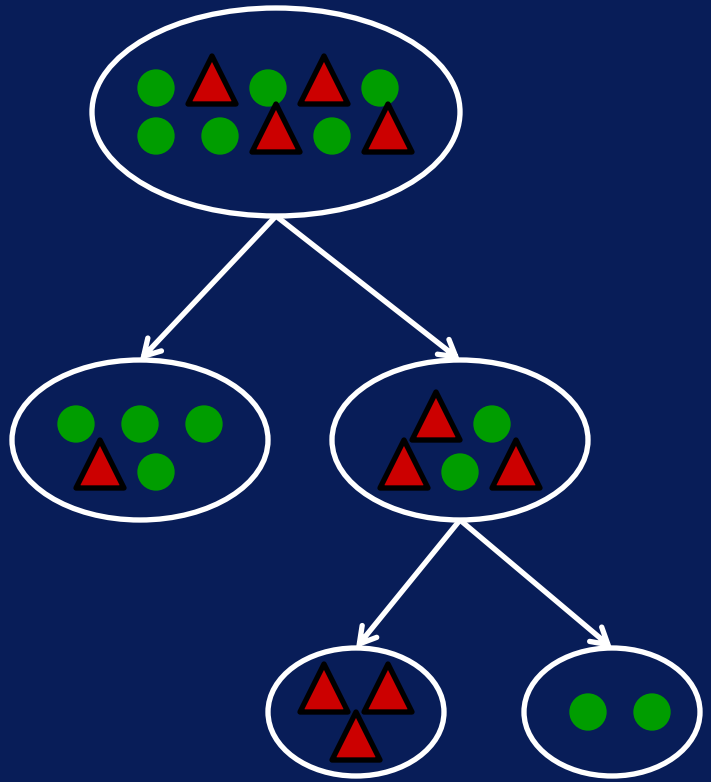
Overfitting in Decision Trees

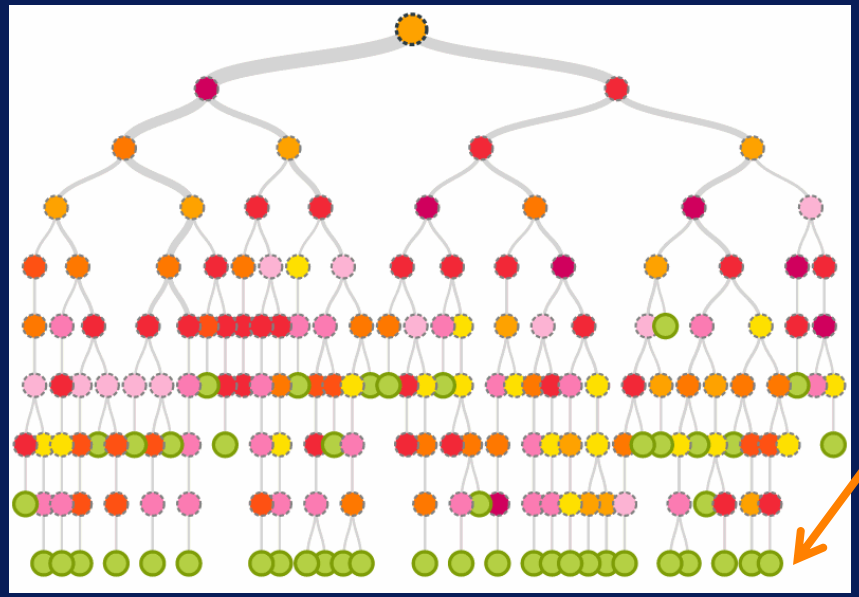
After this video you will be able to..

- Discuss overfitting in the context of decision tree models
- Explain how overfitting is addressed in decision tree induction
- Define pre-pruning and post-pruning

Decision Tree Induction



Overfitting in Decision Tree



If nodes are fitting to noise in training data, model will not generalize well

Source: http://piepdx.org/blog/2013/12/10/which-one-is-is

Avoiding Overfitting in Decision Tree

Pre-Pruning	Post-Pruning
Stop growing tree before fully grown	Grow tree to max size, then prune

Control number of nodes to limit complexity of tree

Pre-Pruning

- Restrictive stopping conditions for growing tree:
  - Stop if number of records < some threshold
  - Stop if improvement in impurity measure < some threshold

Pre-Pruning
Stop growing tree before fully grown

Post-Pruning

- Pruning
  - Remove nodes from bottom up
  - Replace subtree with leaf node if generalization error improves or does not change

Post-Pruning
Grow tree to max size, then prune

Overfitting in Decision Tree

Pre-Pruning	Post-Pruning
Stop growing tree before fully grown	Grow tree to max size, then prune

- Post-pruning used more often
- But is more computational expensive

Tree Pruning to Avoid Overfitting

Pre-Pruning	Post-Pruning
Stop growing tree before fully grown	Grow tree to max size, then prune

Control number of nodes to limit complexity of tree