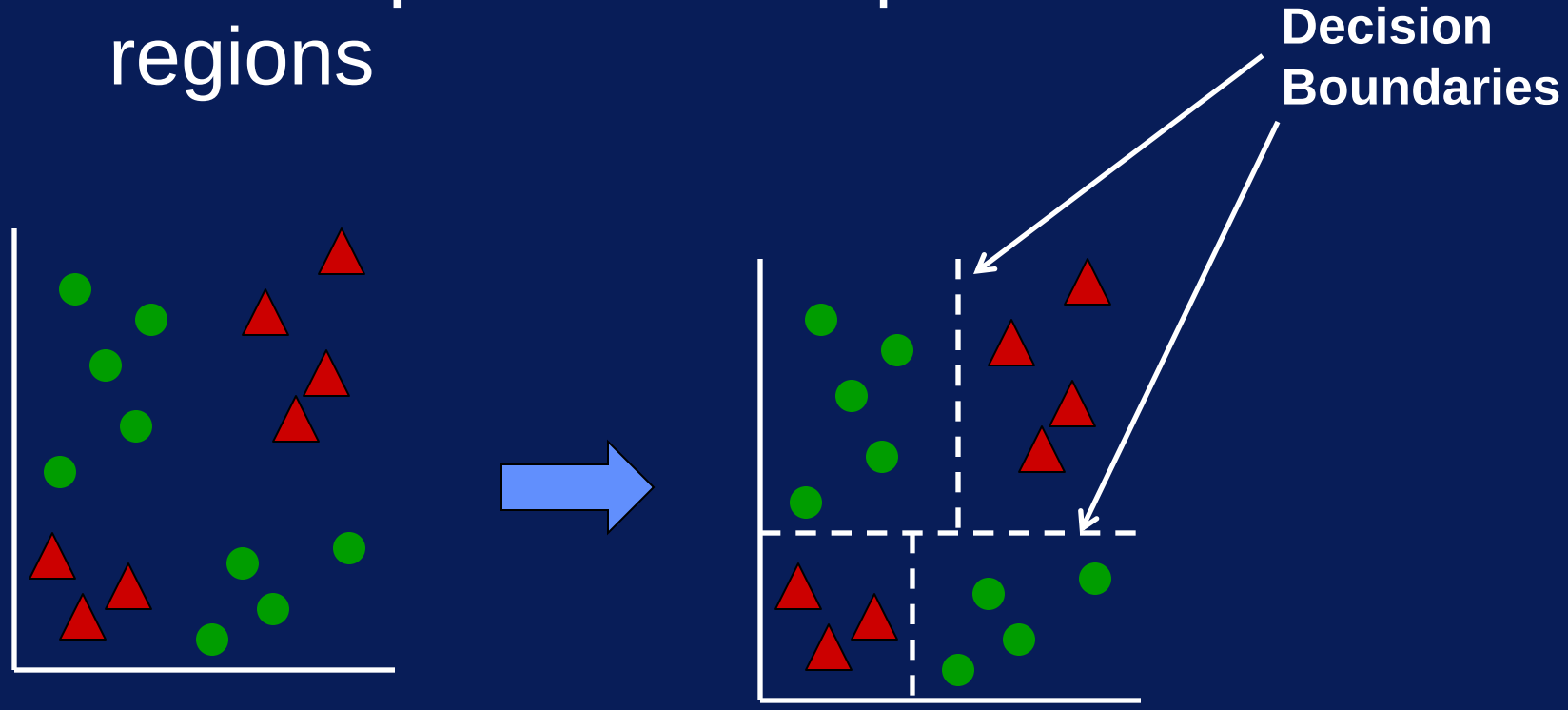
Decision Tree

After this video you will be able to..

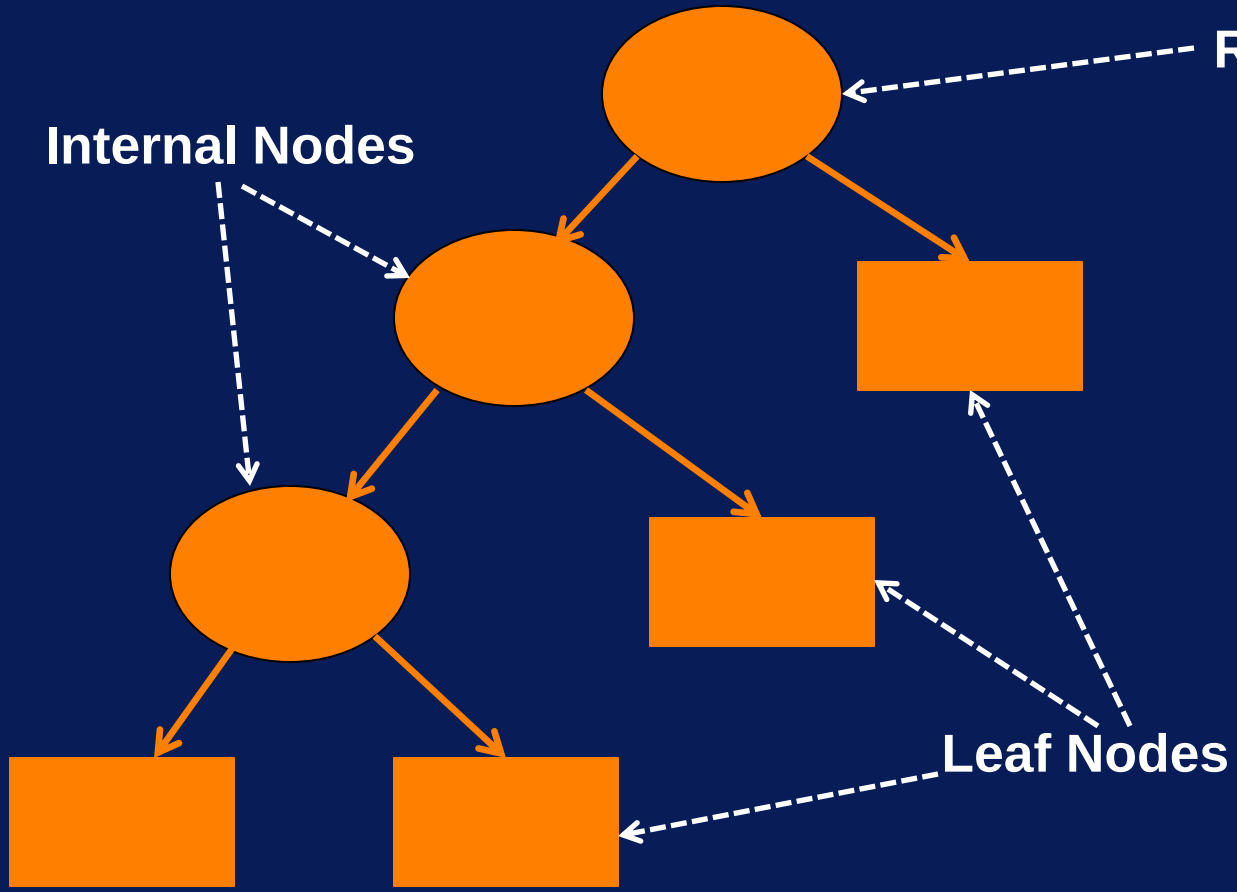
- Explain how a decision tree is used for classification
- Describe the process of constructing a decision tree for classification
- Interpret how a decision tree comes up with a classification decision

Decision Tree Overview

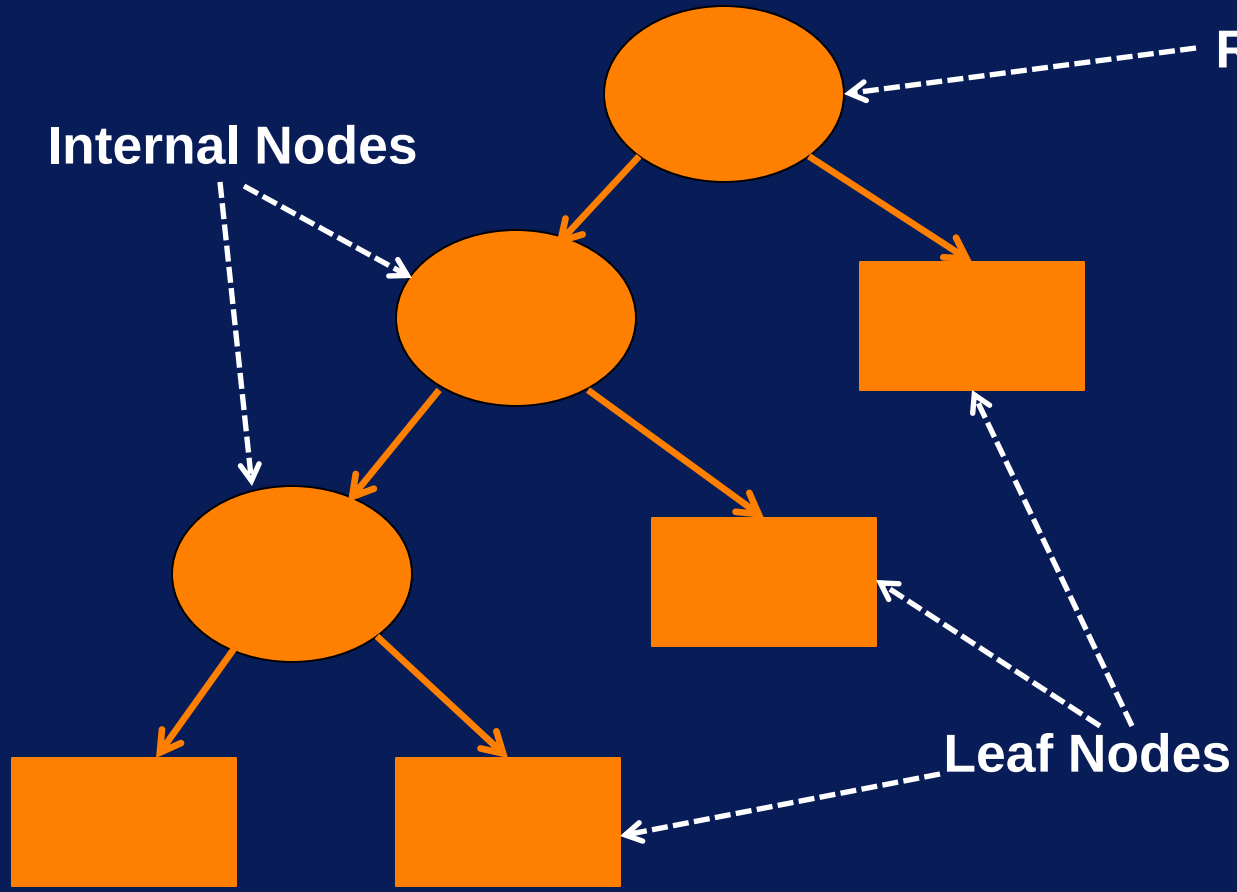
- Idea: Split data into “pure” regions



Classification Using Decision Tree

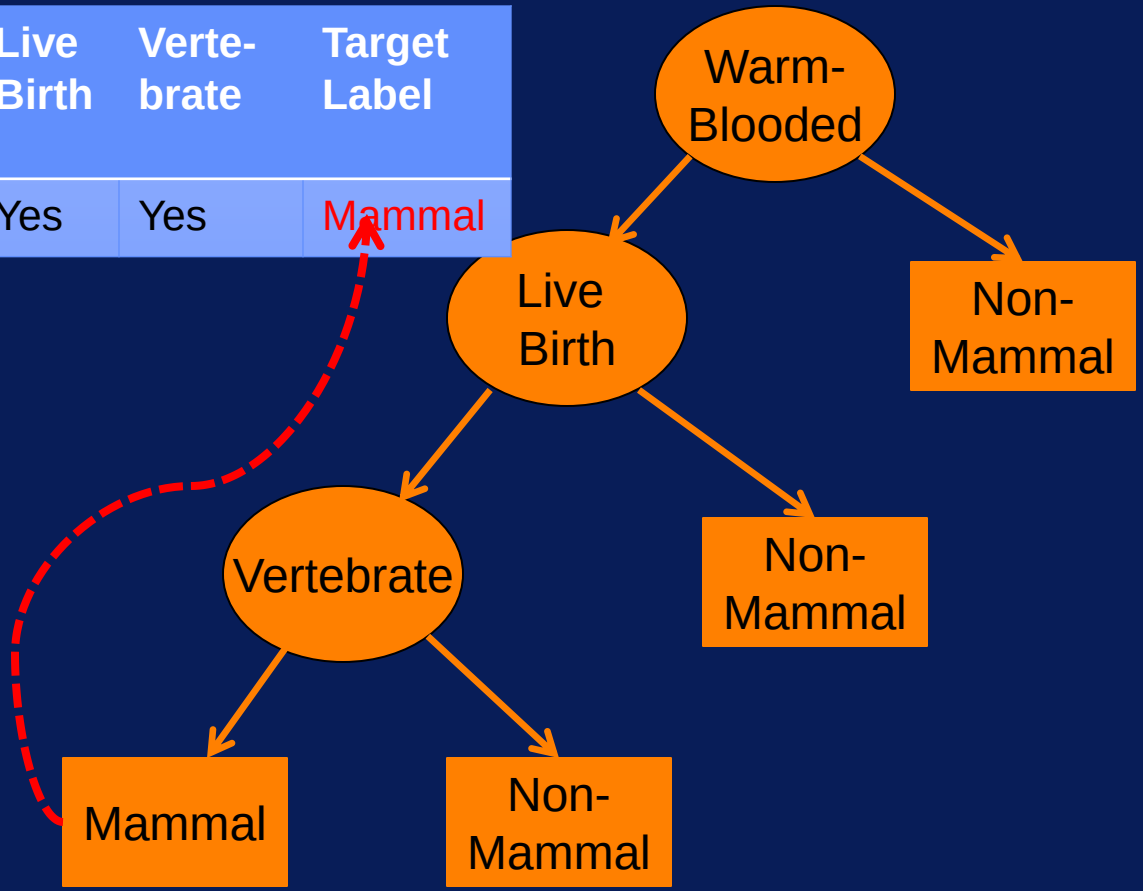


Classification Using Decision Tree



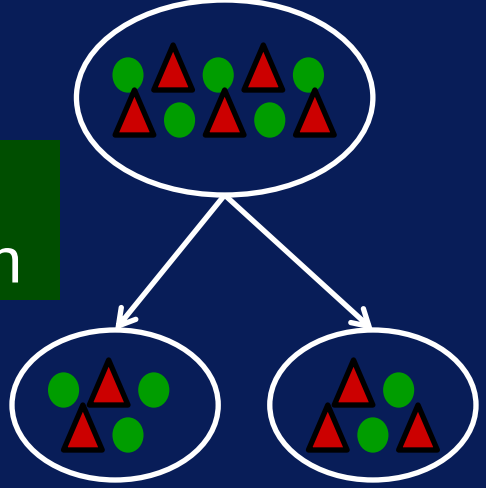
Example Decision Tree

Warm-Blooded	Live Birth	Verte-brate	Target Label
Yes	Yes	Yes	Mammal



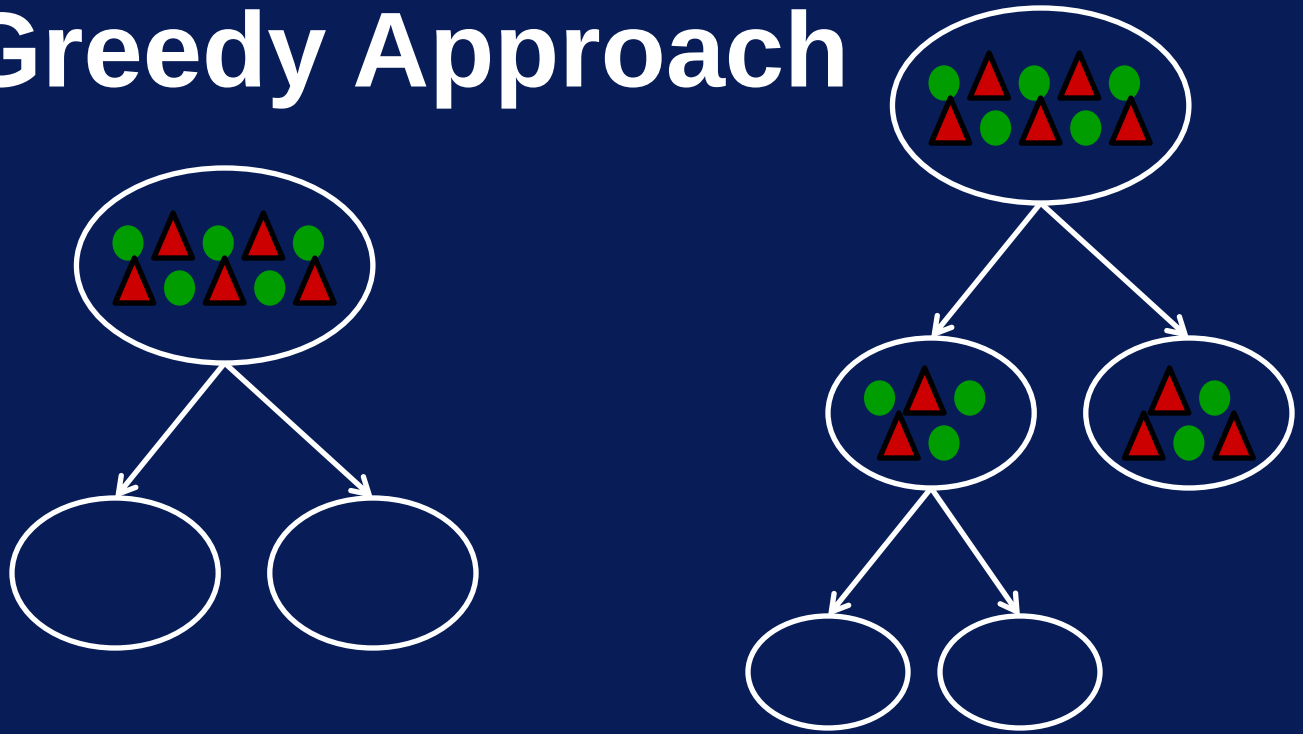
Constructing Decision Tree

Tree Induction



- Start with all samples at a node.
- Partition samples based on input to create purest subsets.
- Repeat to partition data into successively purer subsets.

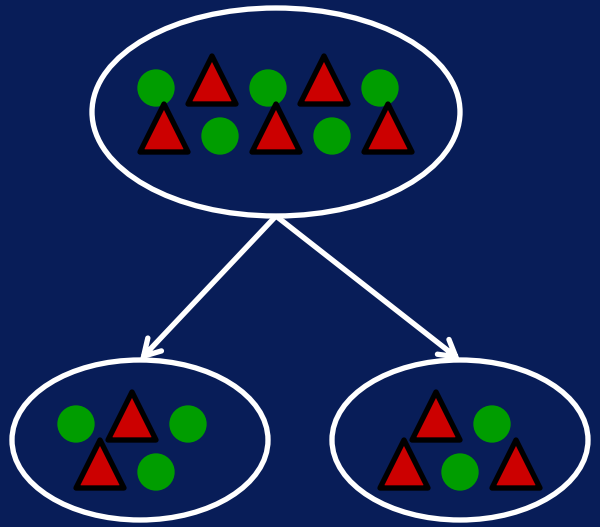
Greedy Approach



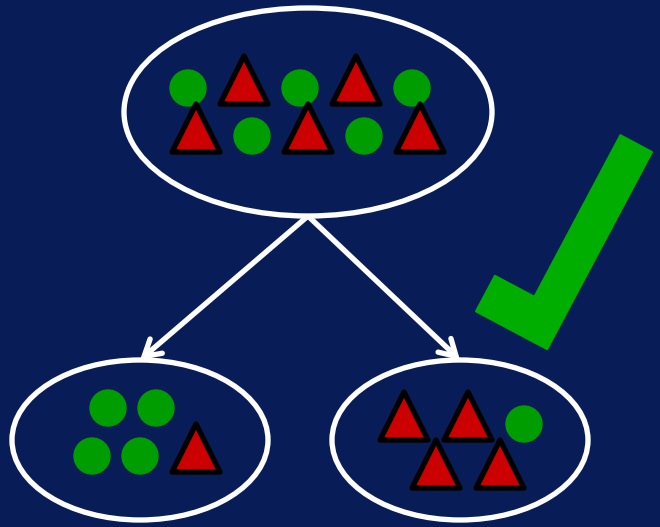
What's the best way to split the current node?

How to Determine Best Split?

Want subsets to be as homogeneous as possible



Less homogeneous = More pure



More homogeneous = More pure

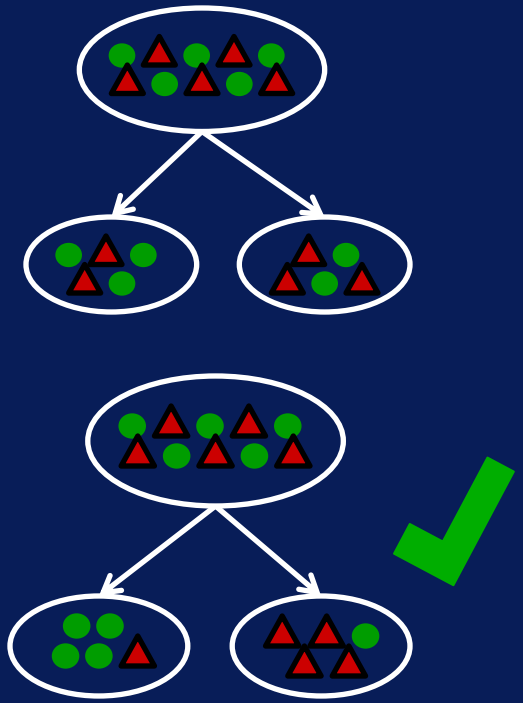
Impurity Measure

- To compare different ways to split data in a node

Gini Index

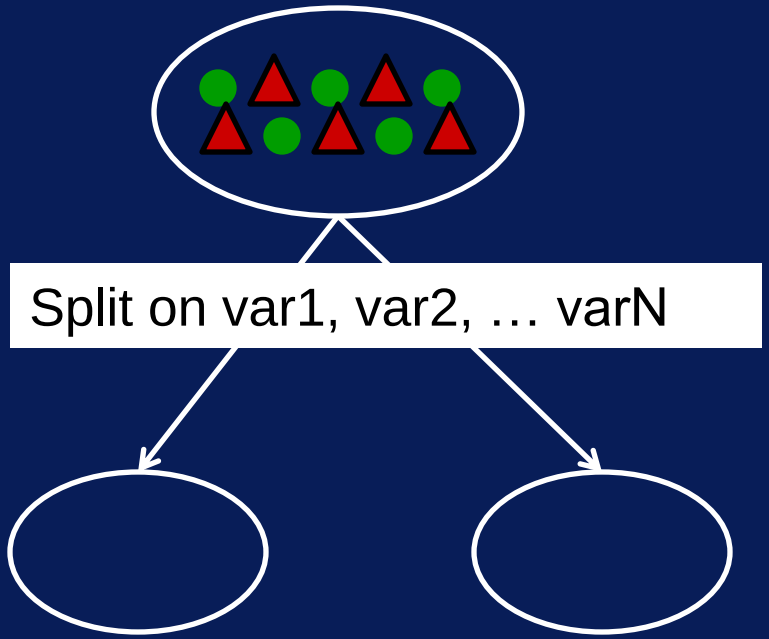
Higher = Less pure

Lower = More pure



What Variable to Split On?

- Splits on all variables are tested

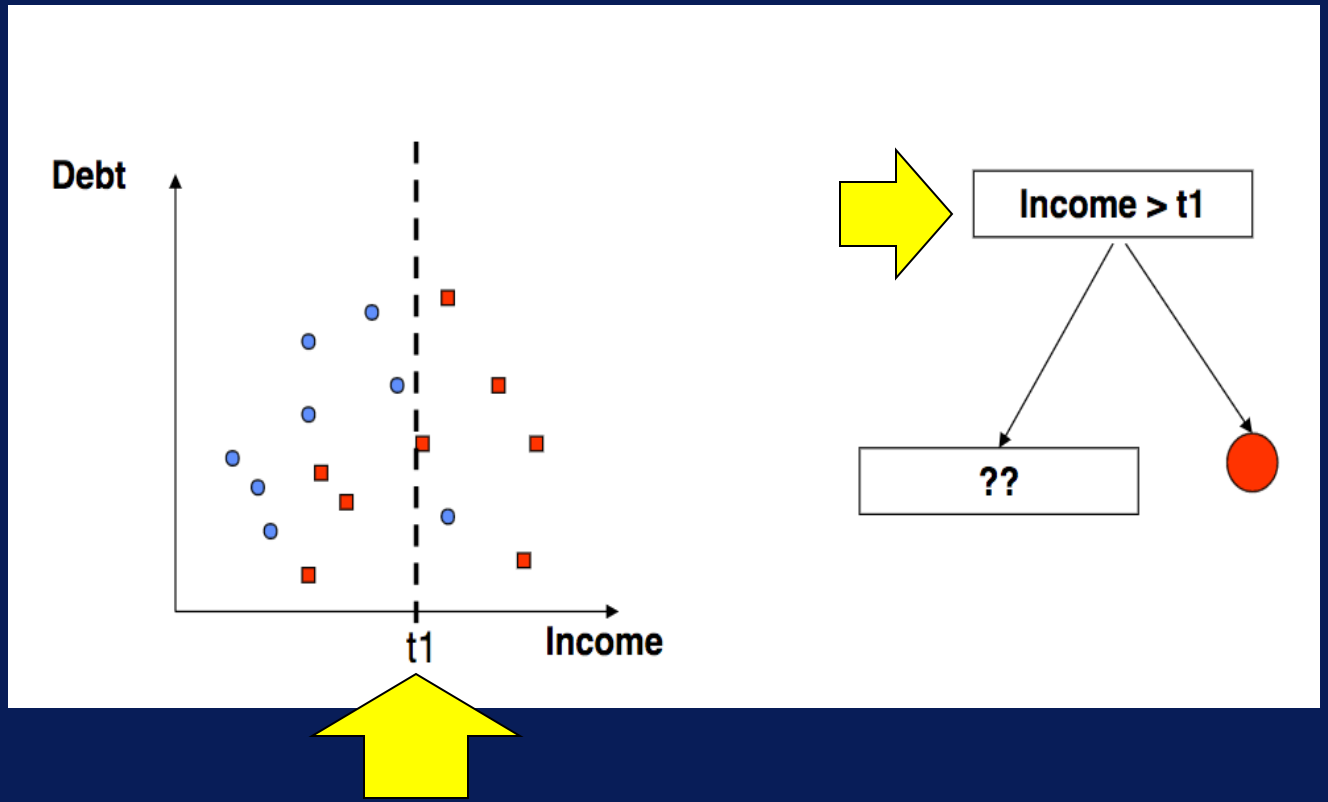


When to Stop Splitting a Node?

- All (or X% of) samples have same class label
- Number of samples in node reaches minimum
- Change in impurity measure is smaller than threshold
- Max tree depth is reached
- Others…

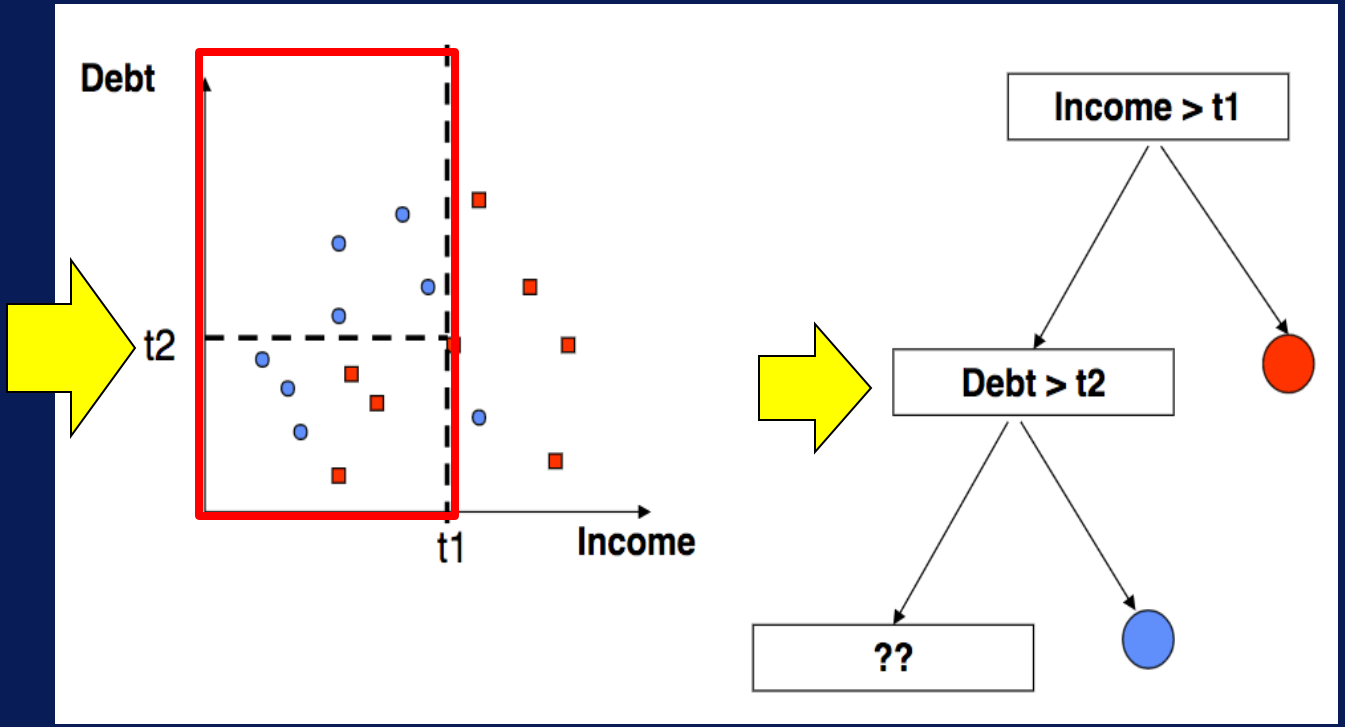
Tree Induction Example

- Split 1



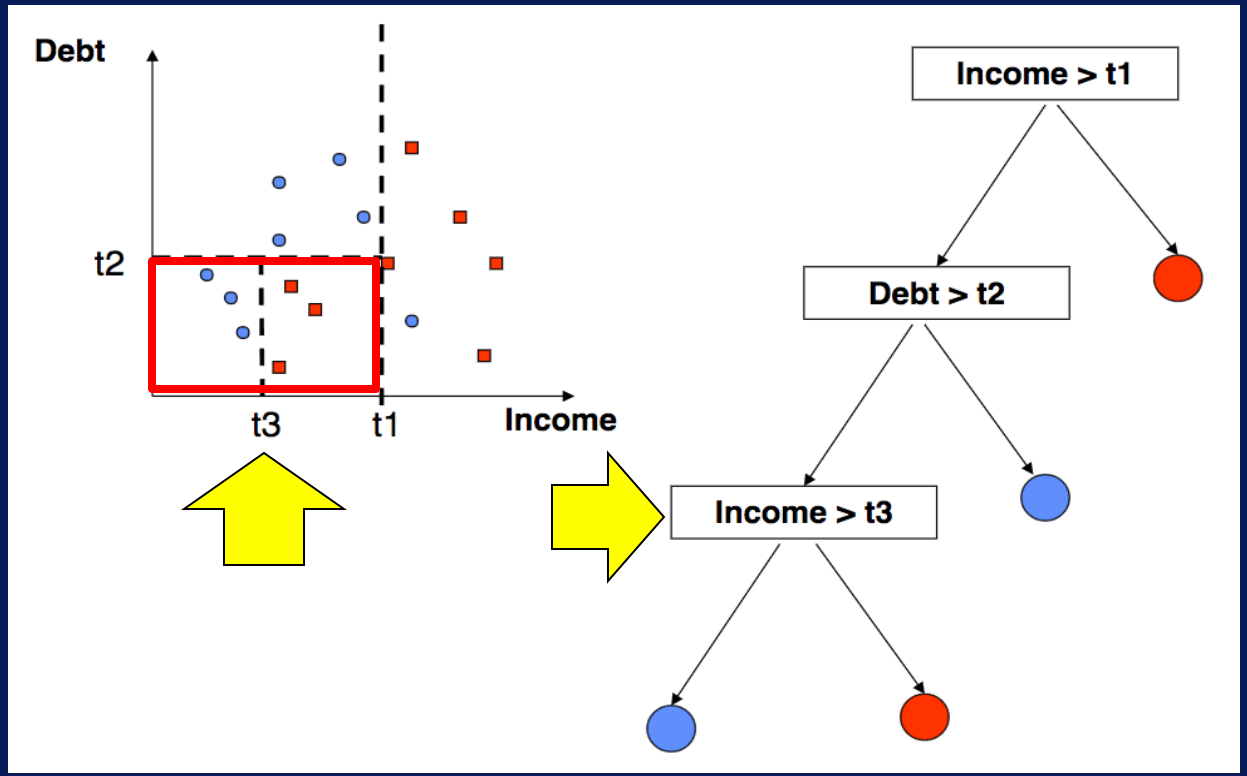
Tree Induction Example

- Split 2



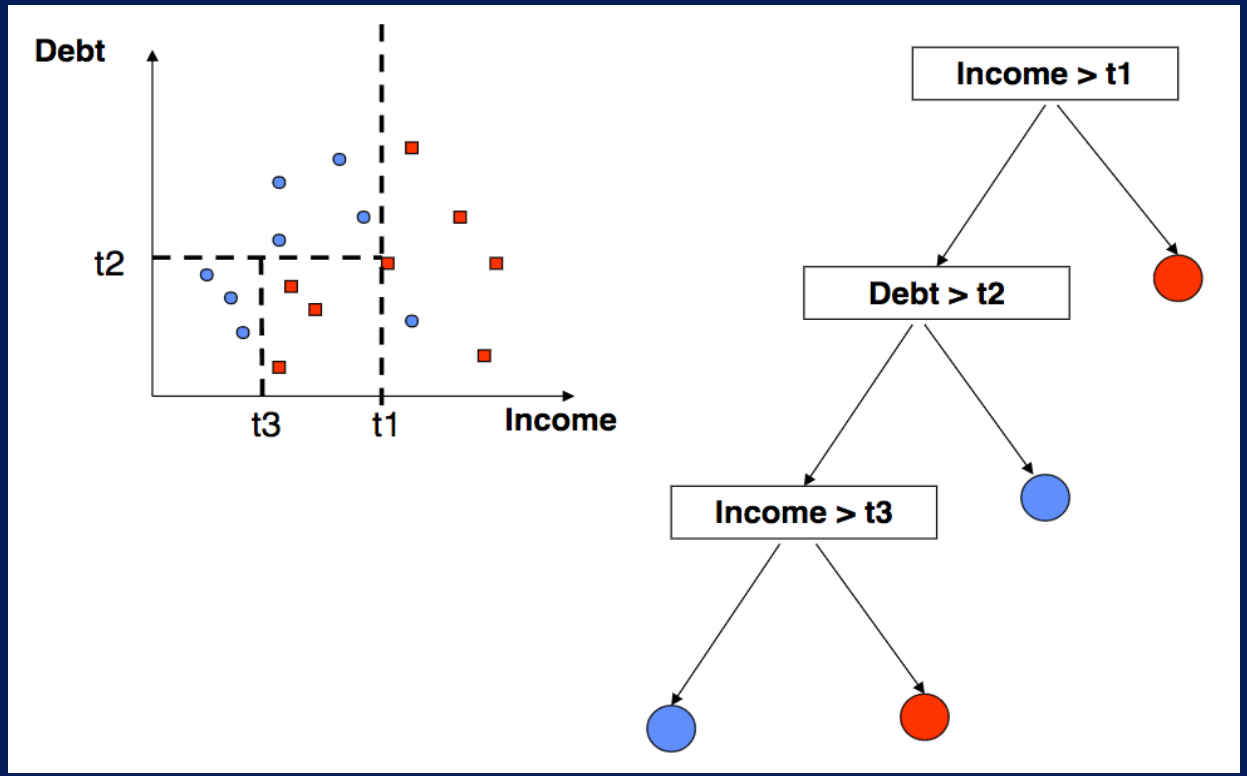
Tree Induction Example

- Split 3



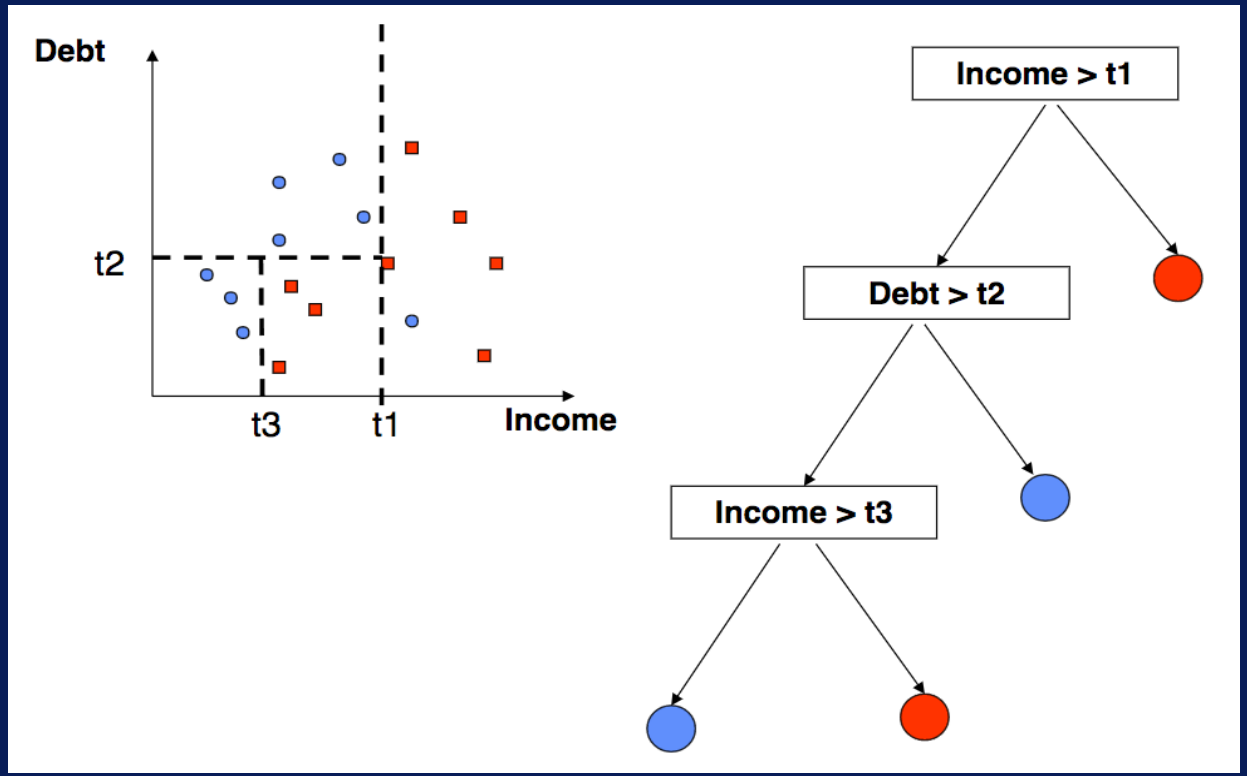
Tree Induction Example

- Resulting model



Decision Boundaries

- Rectilinear = Parallel to axes



Decision Tree for Classification

- Resulting tree is often simple and easy to interpret
- Induction is computationally inexpensive
- Greedy approach does not guarantee best solution
- Rectilinear decision boundaries

Decision Tree