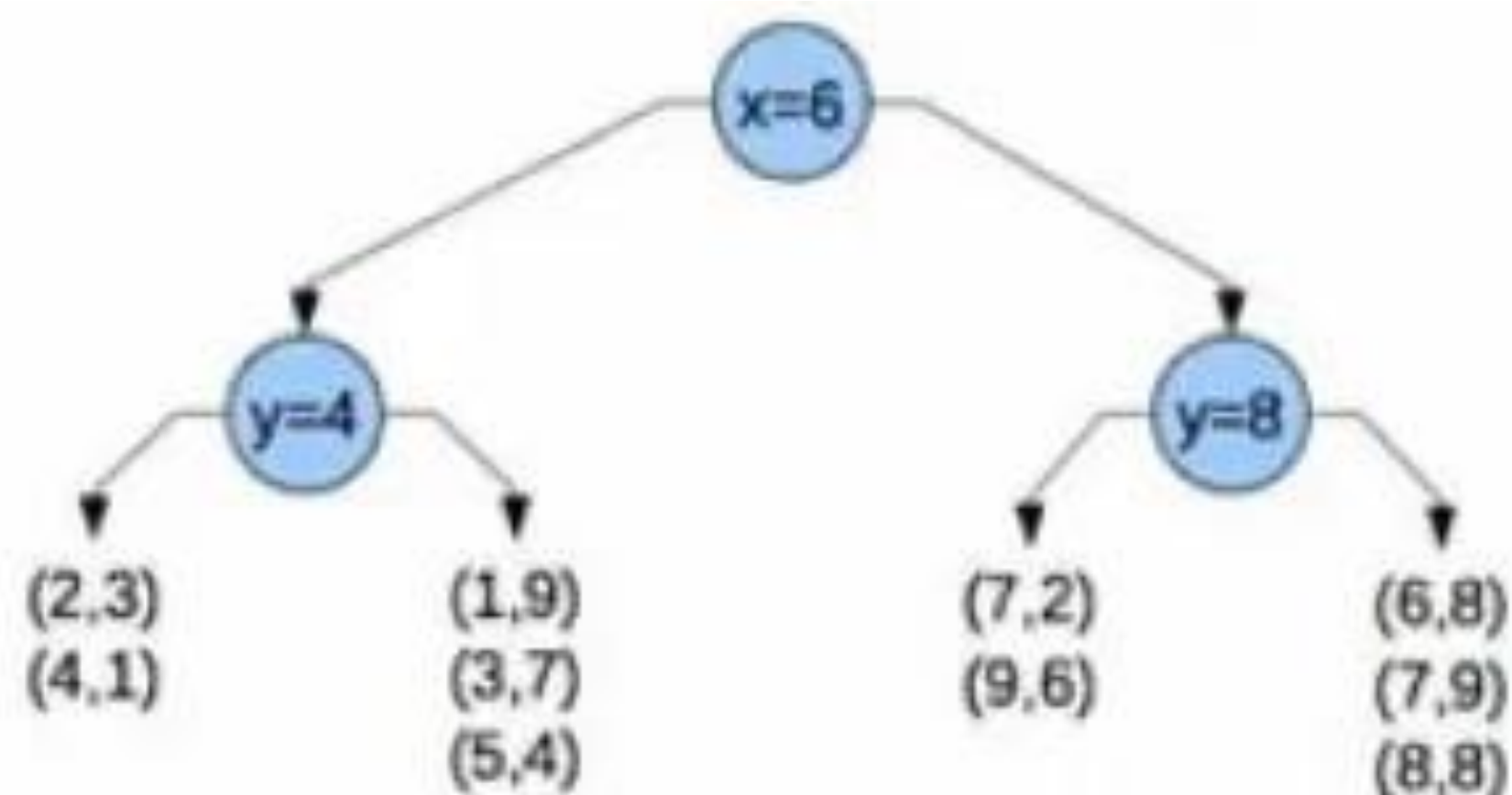
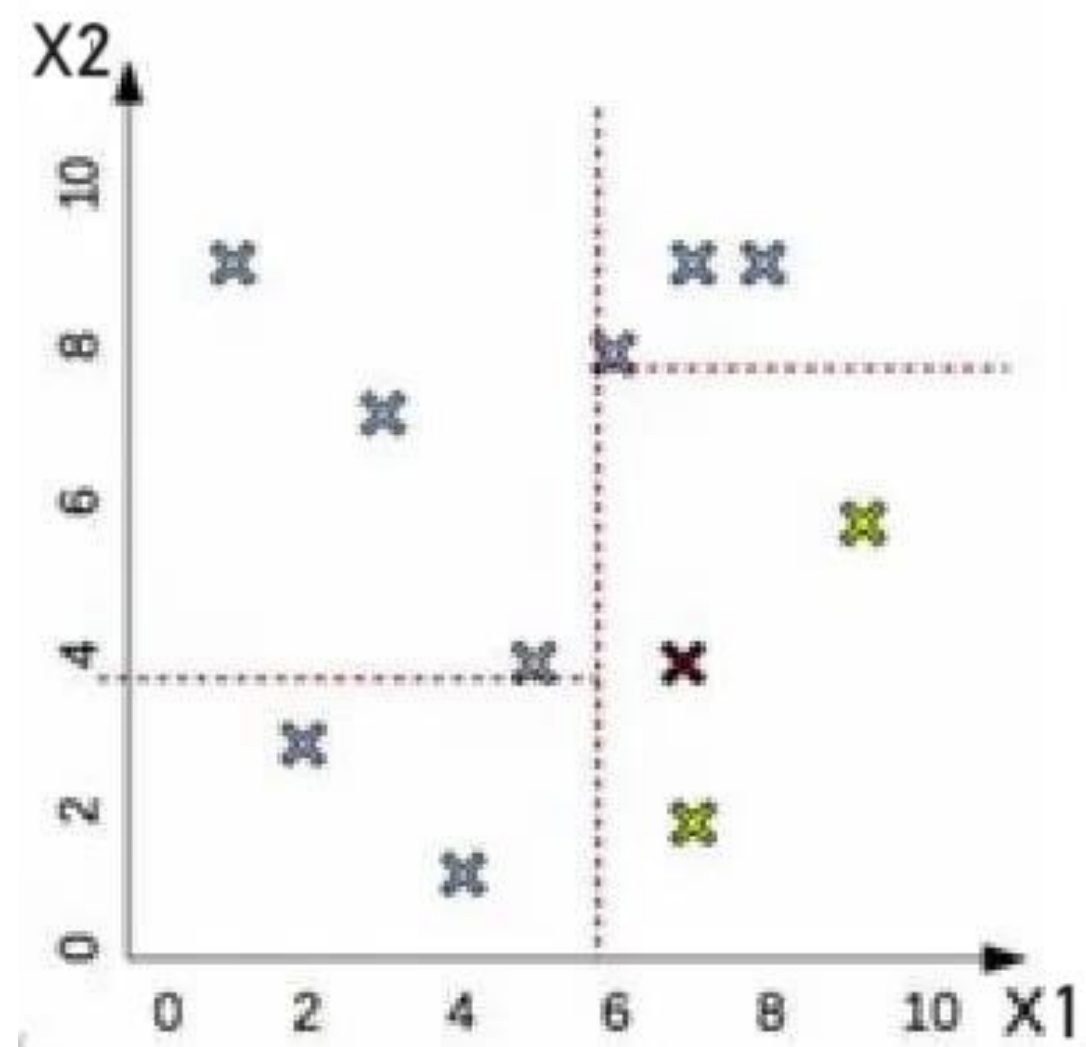
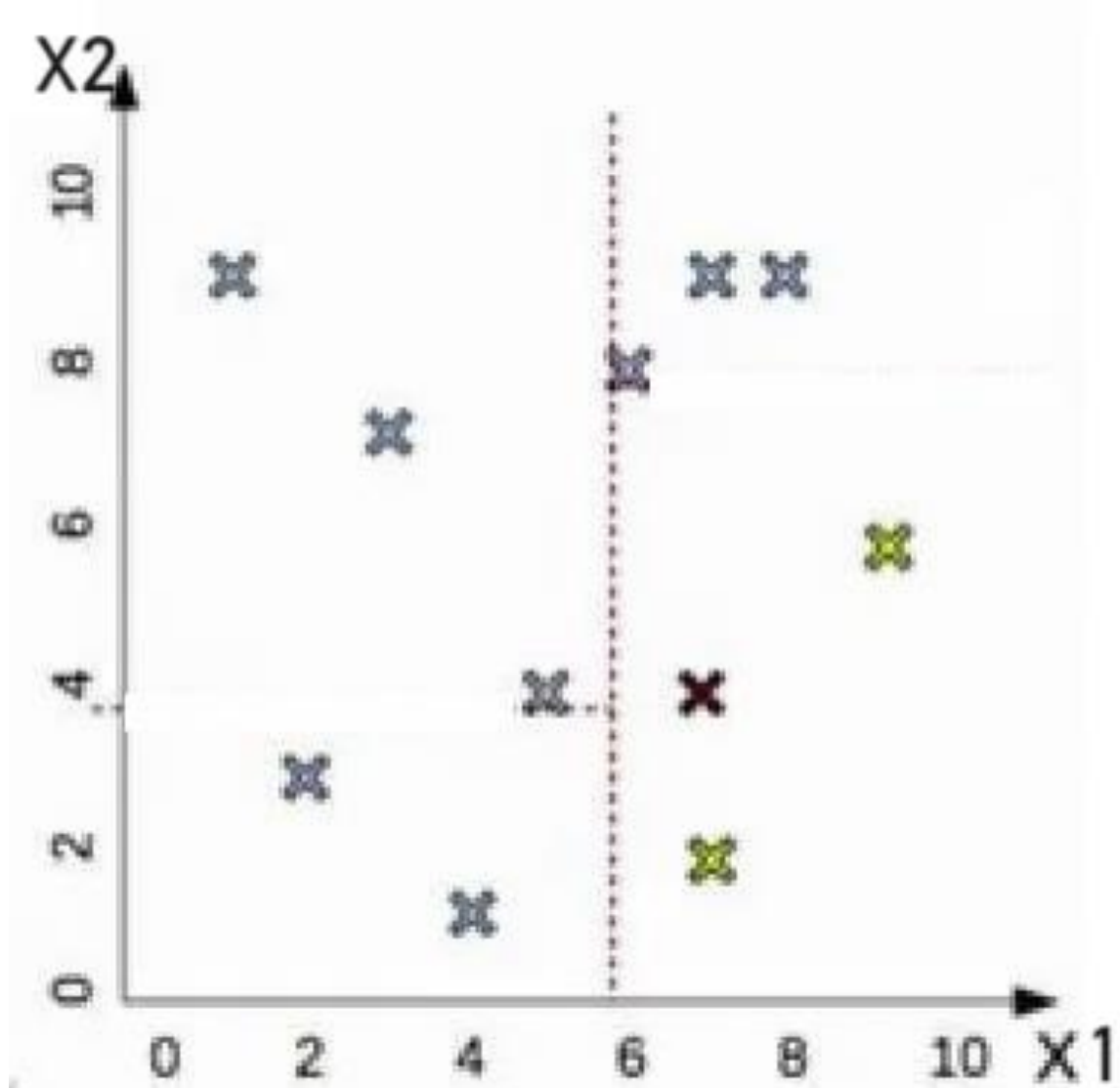
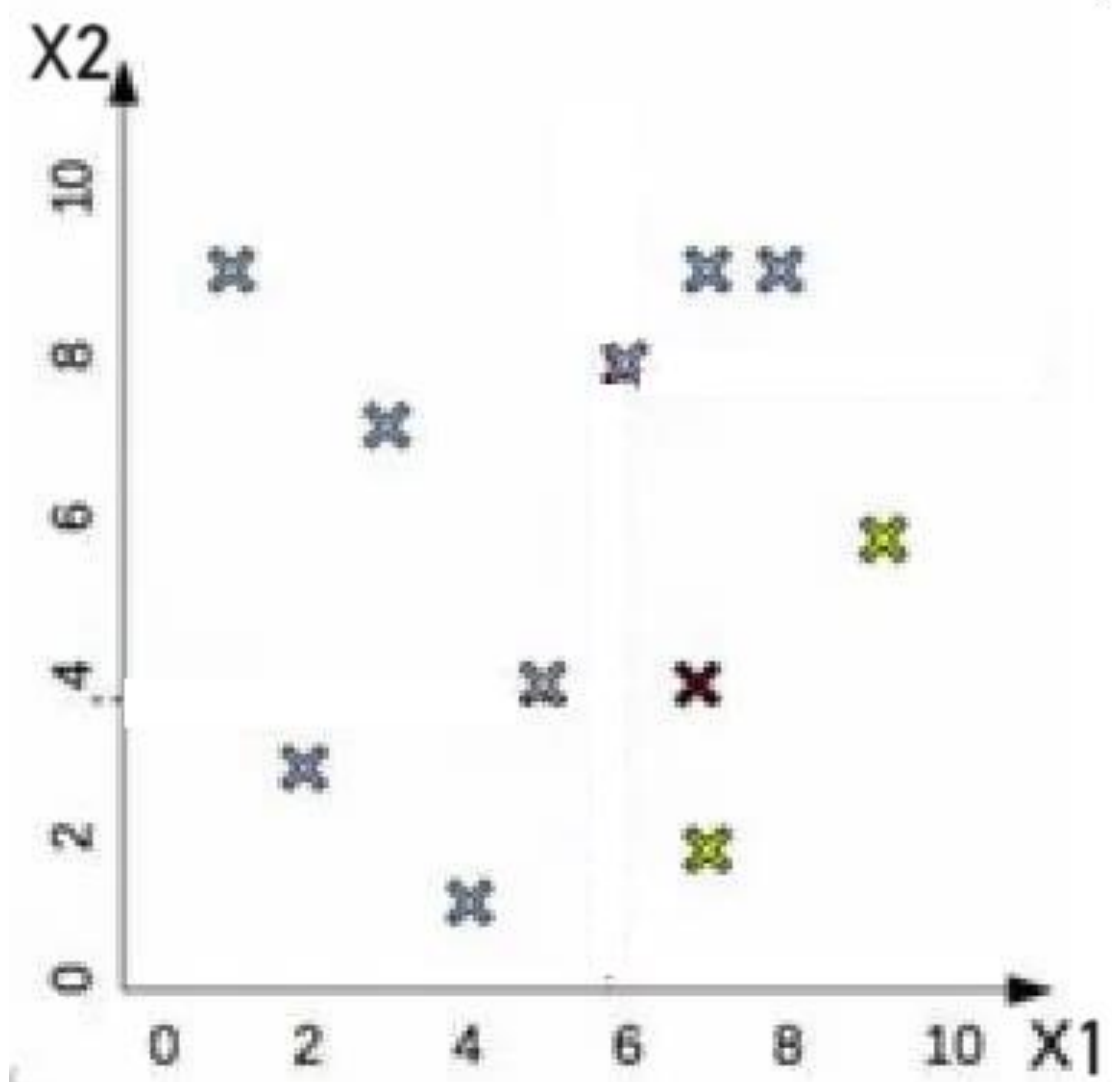




Lecture 99: KD Trees & Decision Trees

Recap

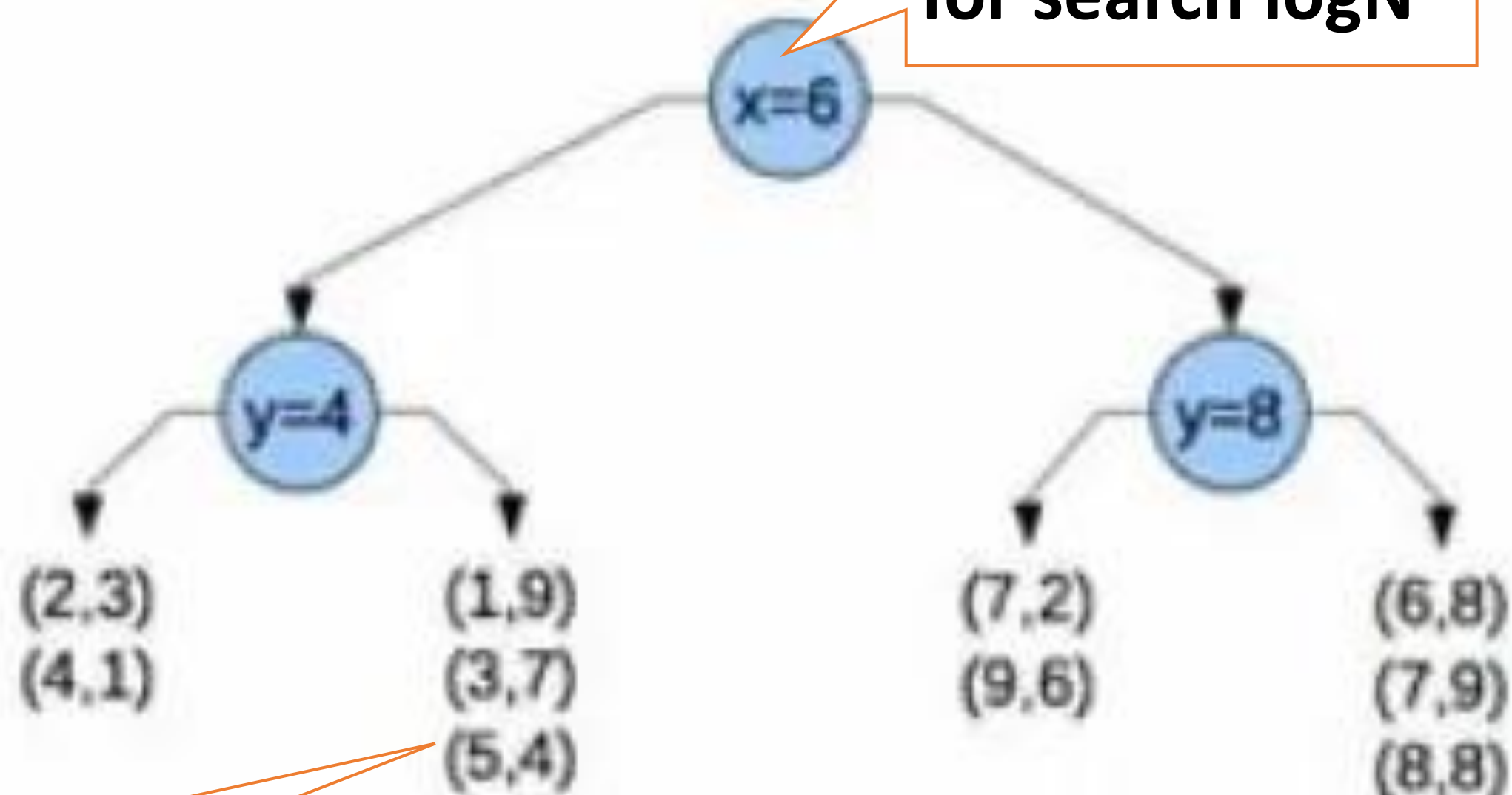
- Hierarchical Clustering
 - Divisive Clustering
 - Agglomerative Clustering
- Linkages
 - Simple, Complete, Average, Centroid, Ward
- Dendrogram



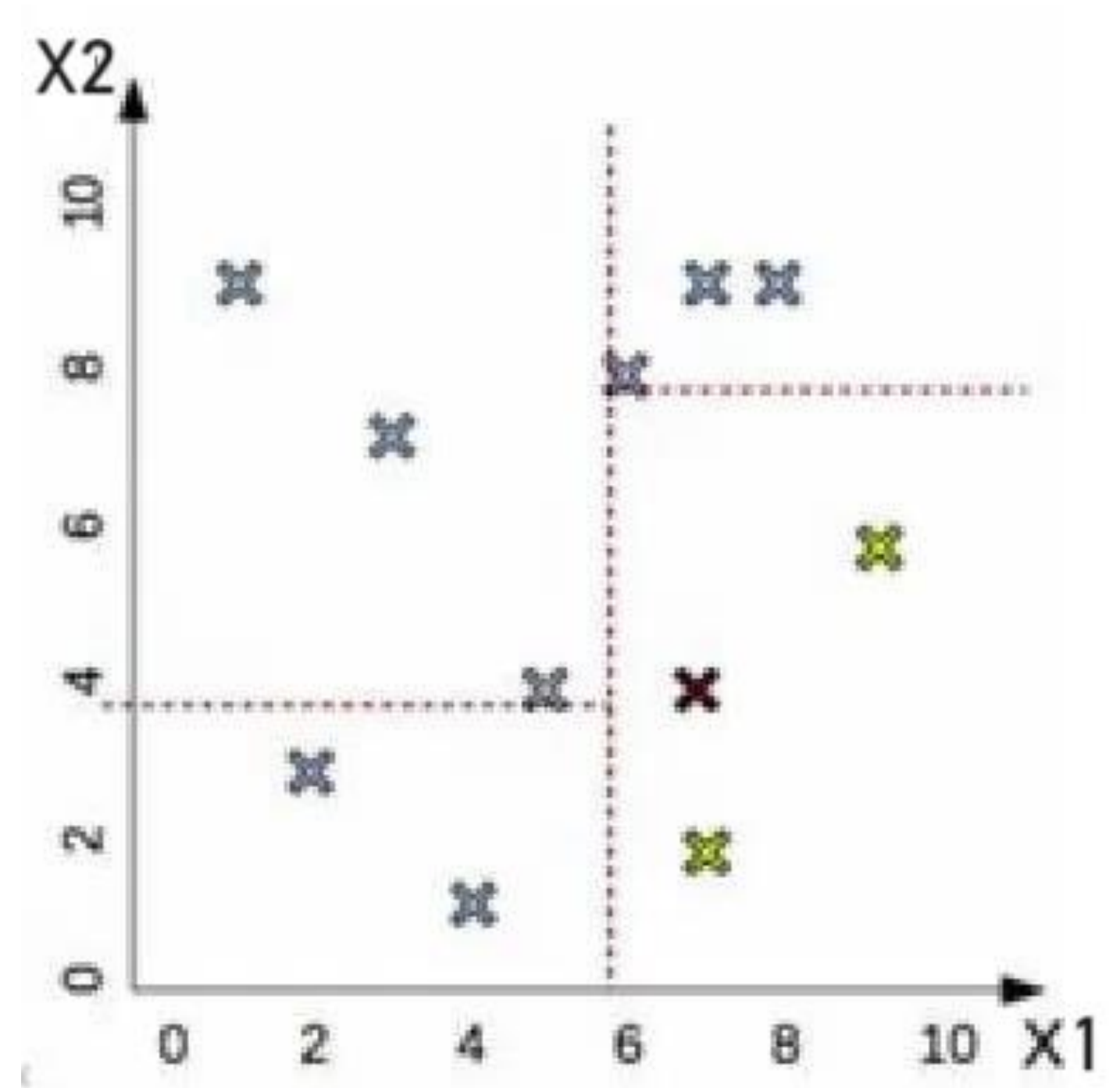
KD Trees - Prediction

- How to search?
 - DFS

Cost of going down the tree for search $\log N$



kNN within a node





Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

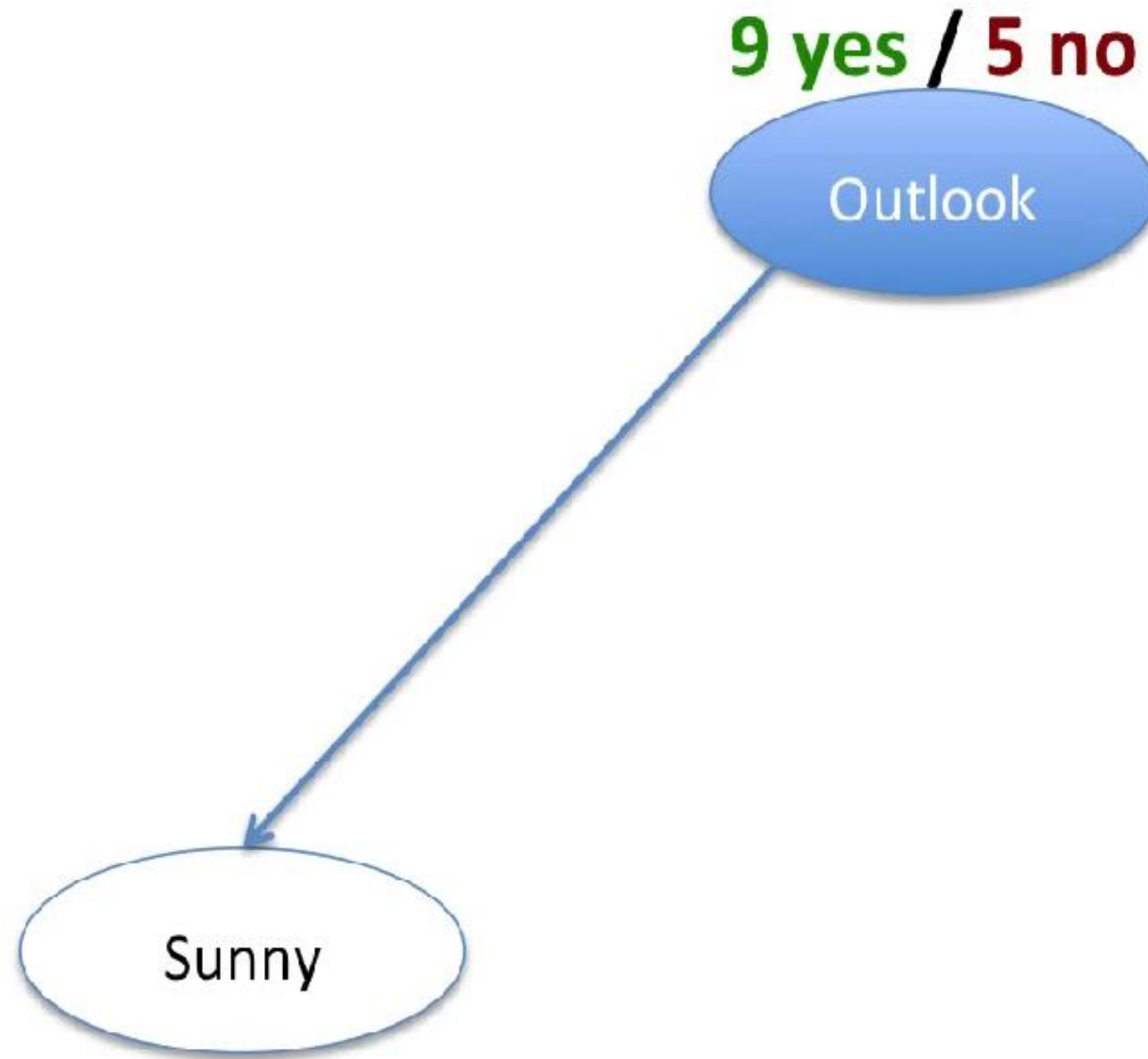
New data:

D15	Rain	High	Weak	?
-----	------	------	------	---

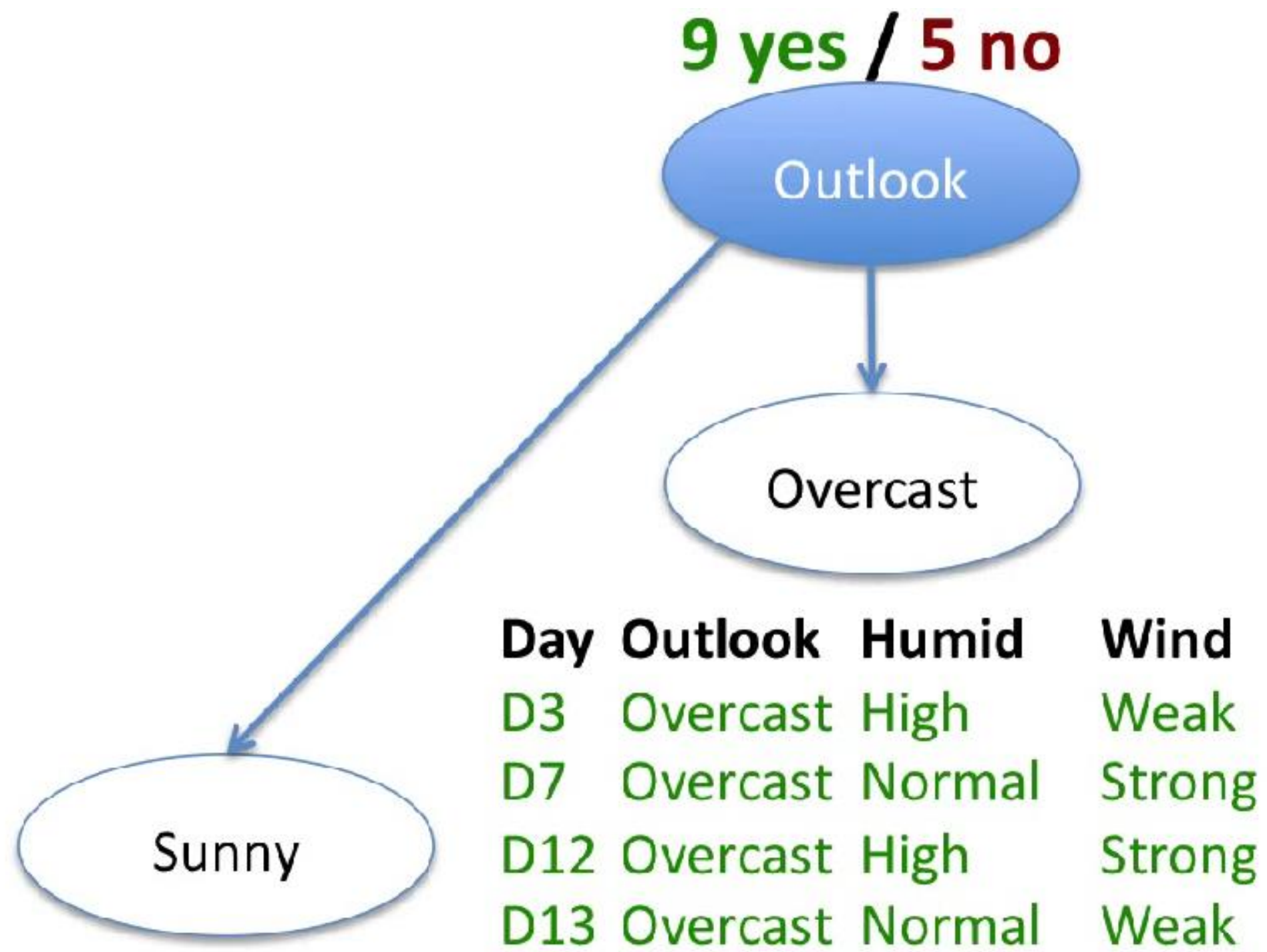
- Training
 - Split on a feature
 - Are subsets pure?
 - If Yes, Stop
 - If no repeat
- Prediction

9 yes / 5 no

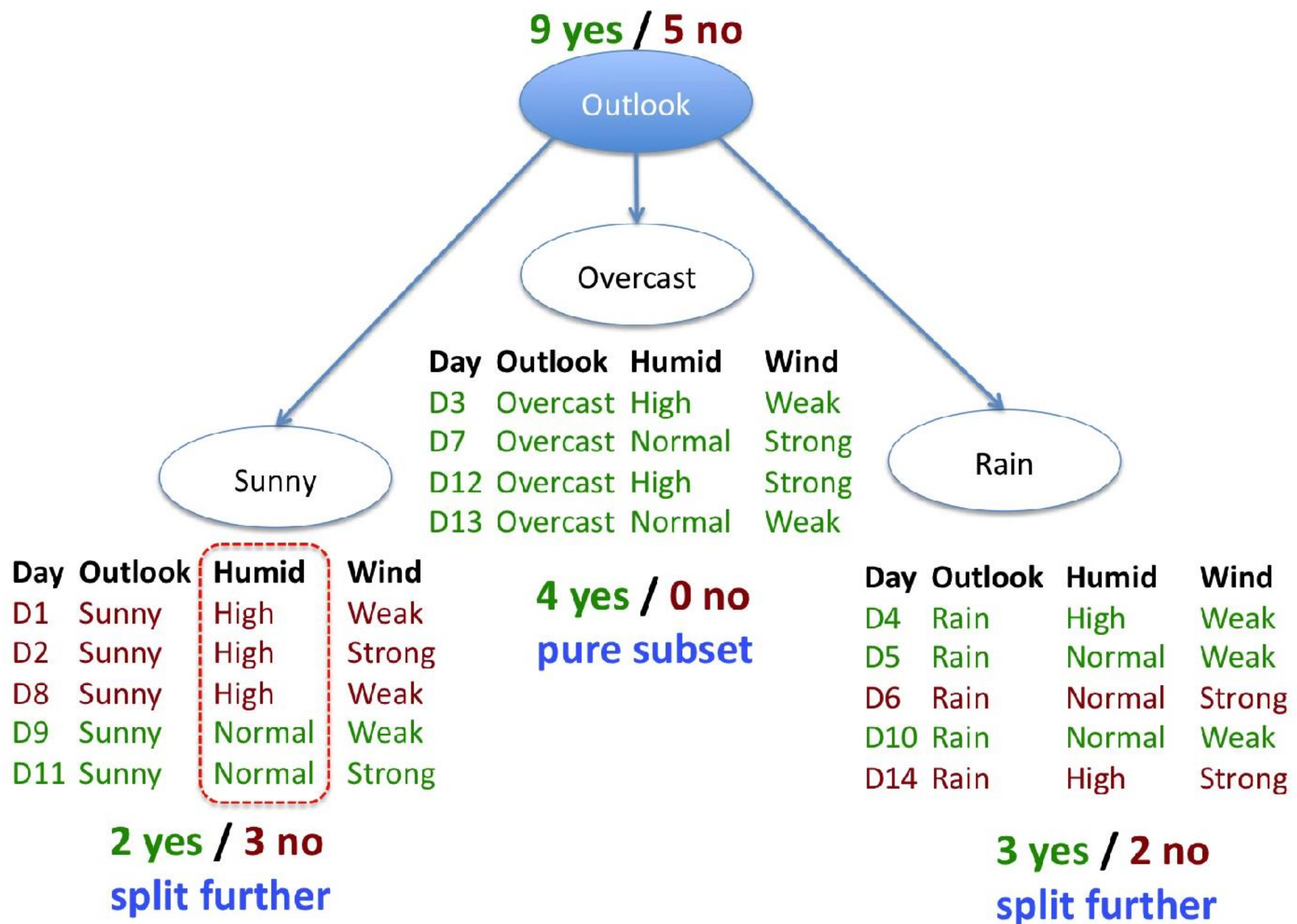


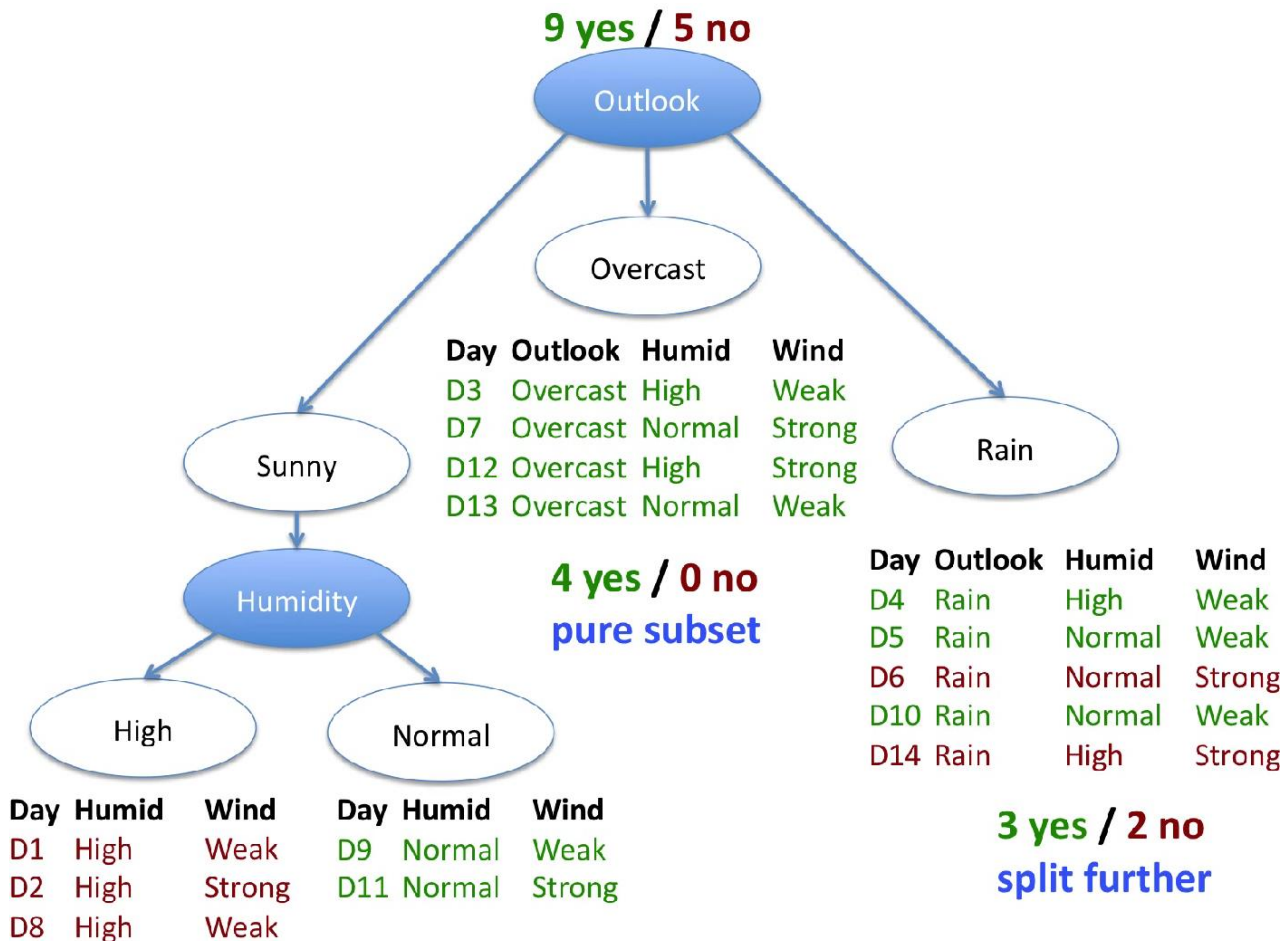


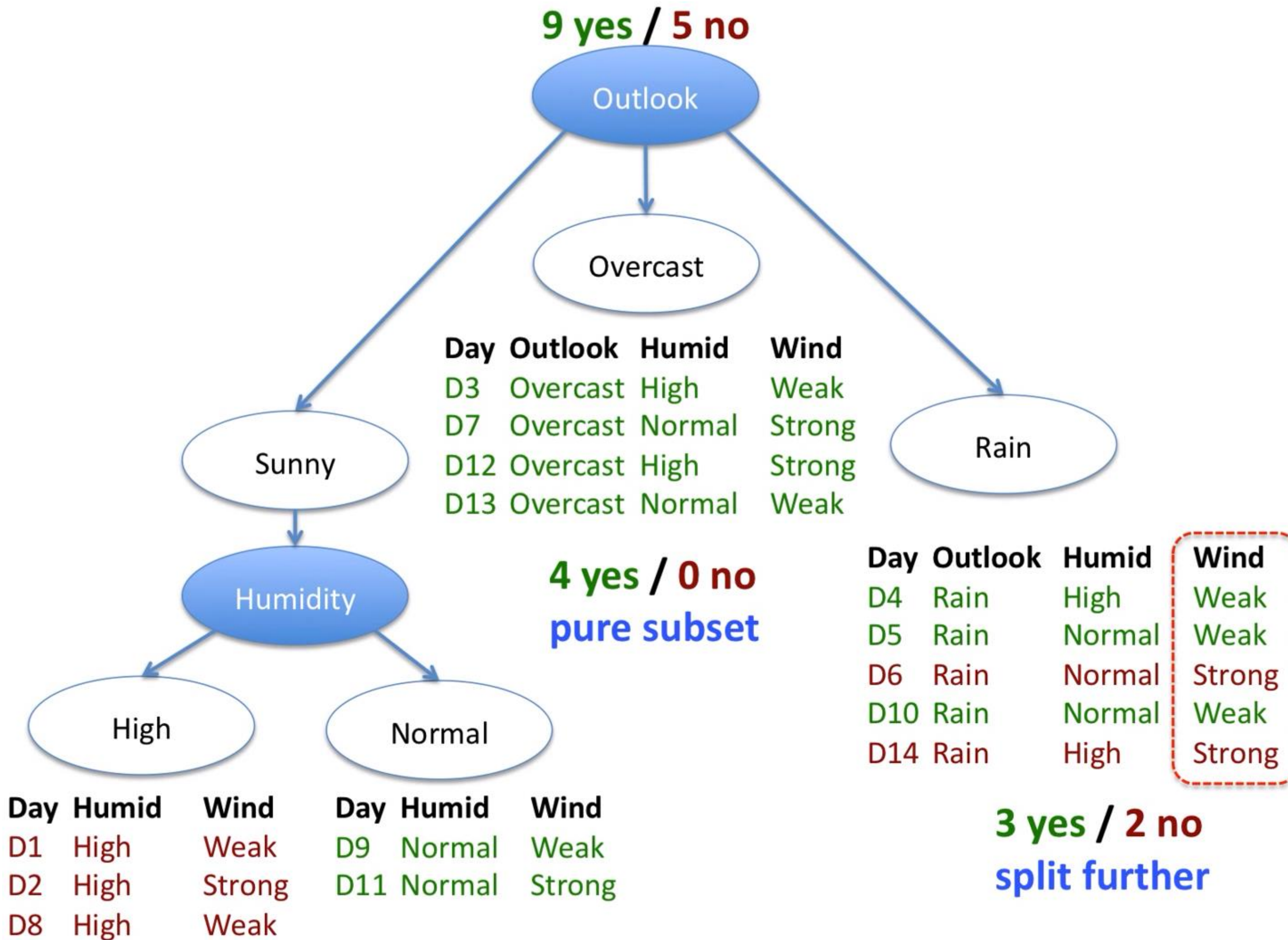
Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

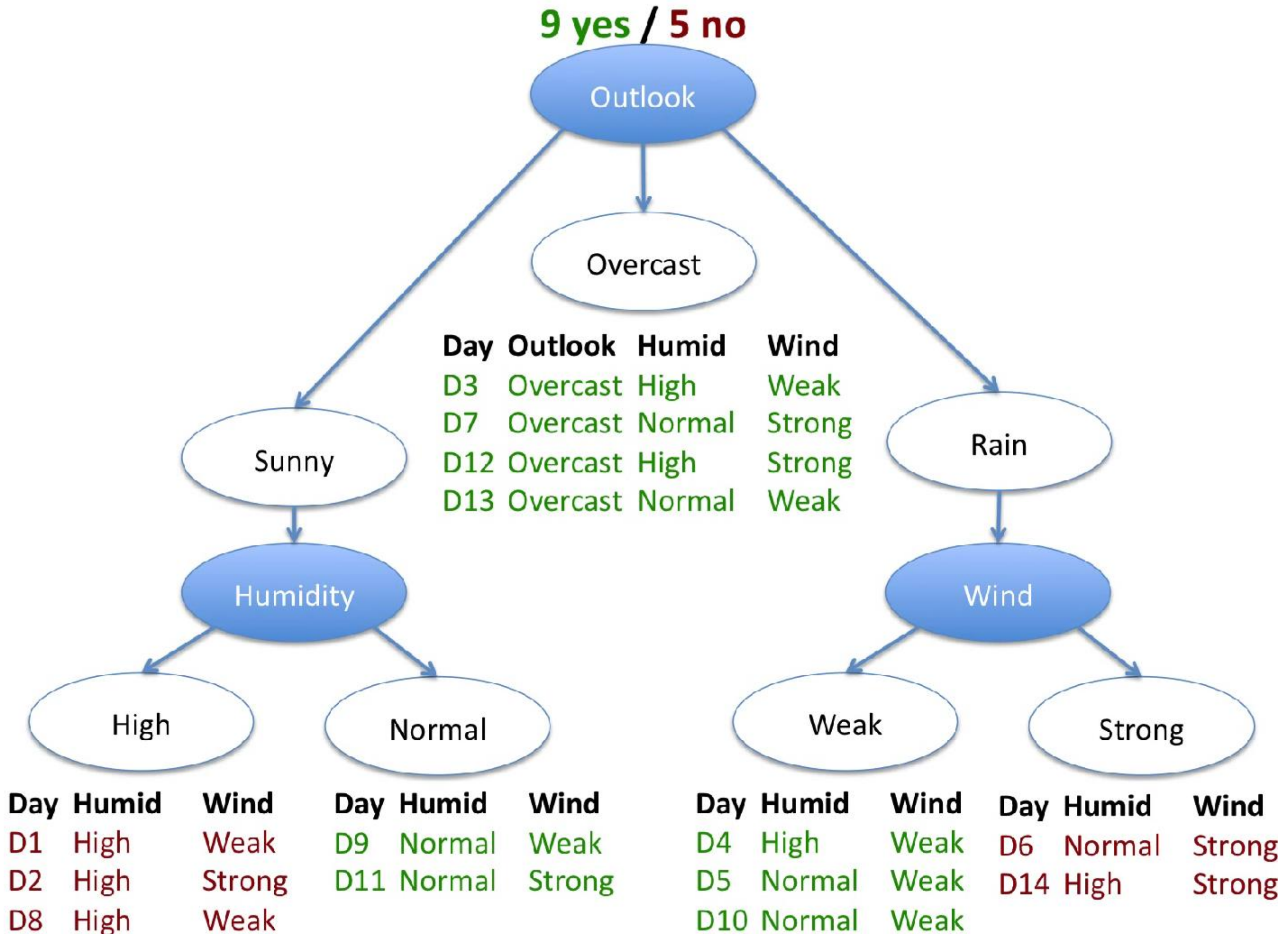


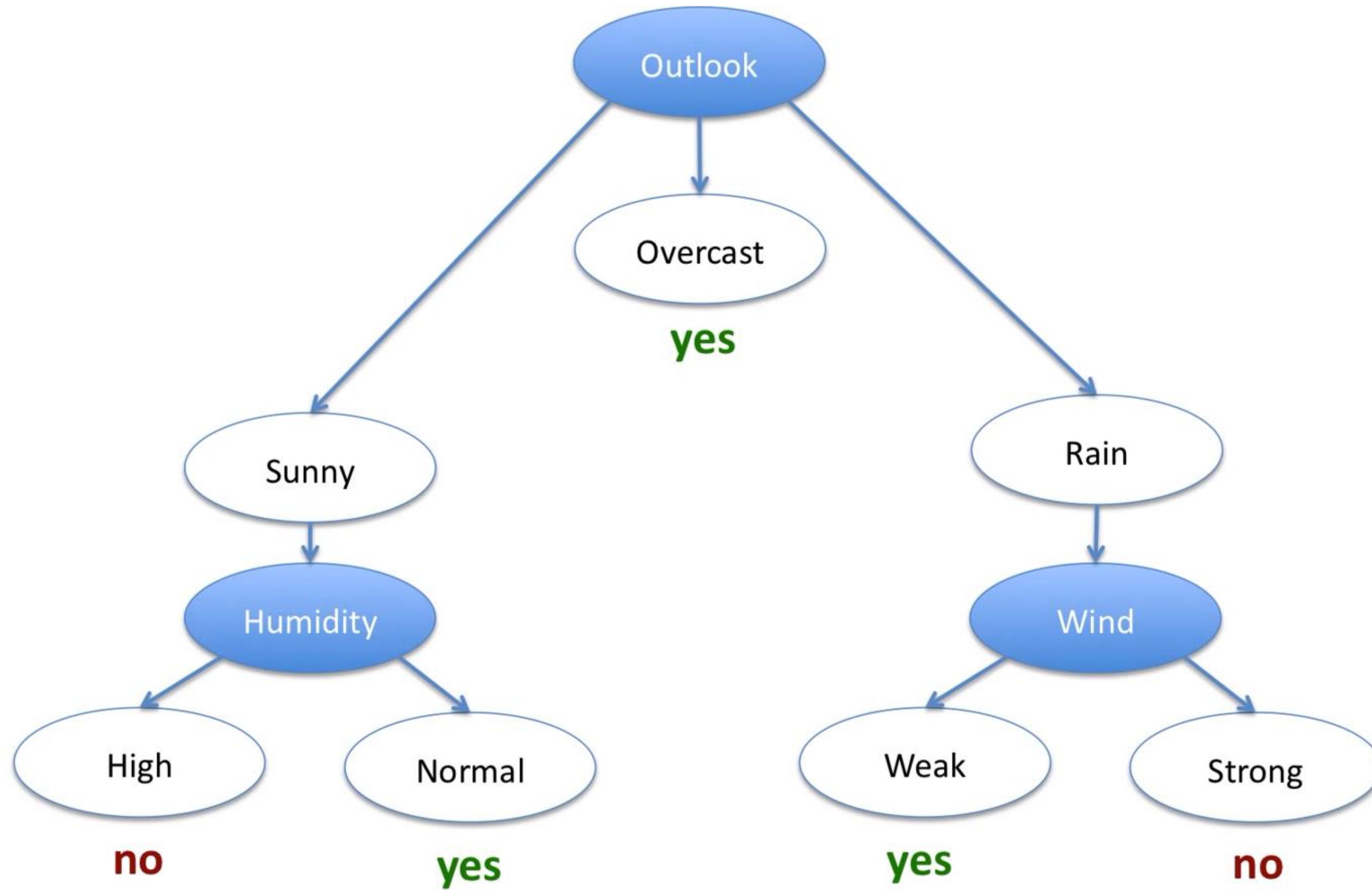
Day	Outlook	Humid	Wind
D1	Sunny	High	Weak
D2	Sunny	High	Strong
D8	Sunny	High	Weak
D9	Sunny	Normal	Weak
D11	Sunny	Normal	Strong

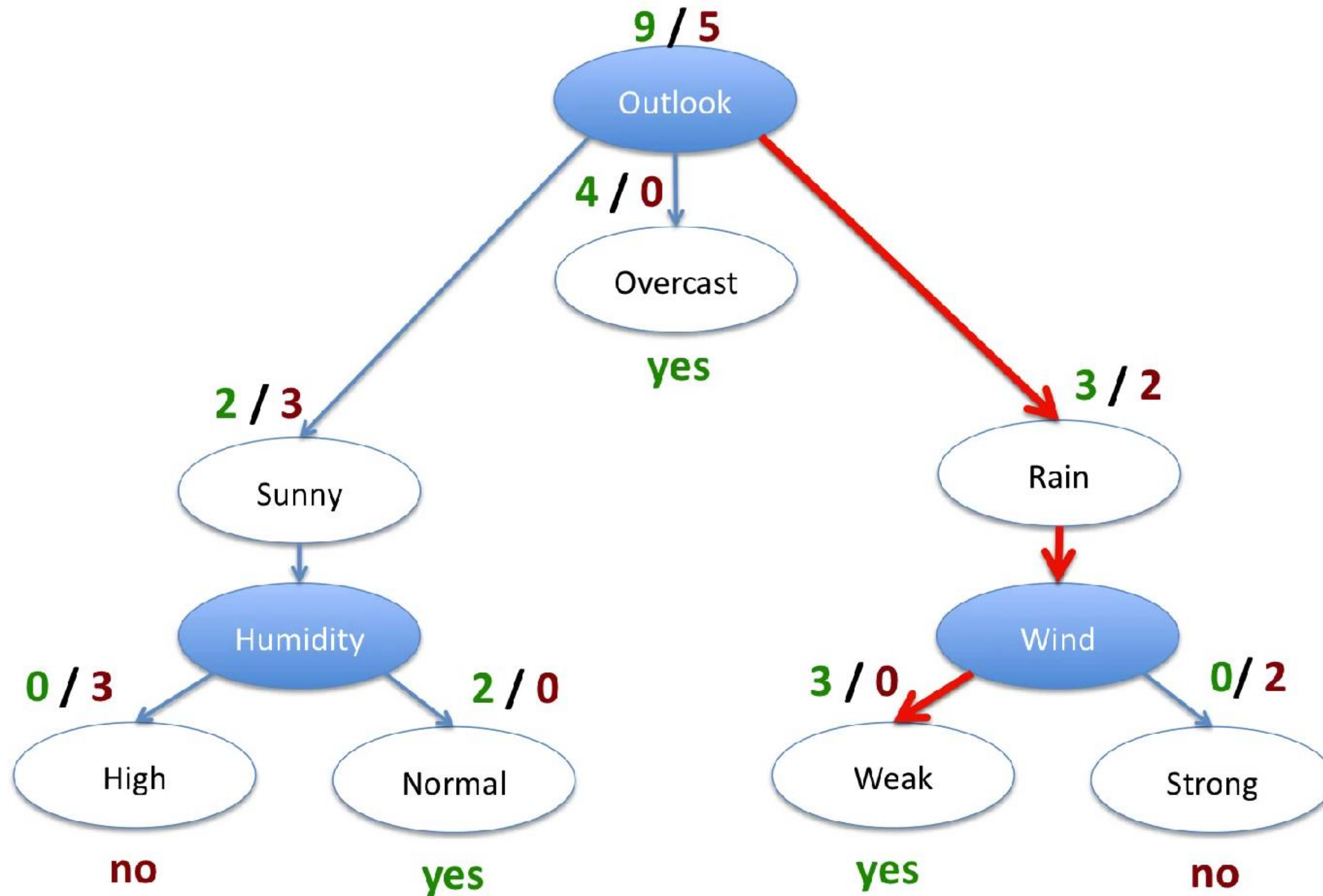












New data:

Day	Outlook	Humid	Wind	
D15	Rain	High	Weak	→ Yes



Decision Tree – ID3 algorithm

ID3 algorithm

- Recursive algorithm
- Operates on node
- Split(data):
 - Pick the best feature x_i to split data (how)
 - x_i becomes decision attribute for this node
 - Create a node
 - Branch on all possible values of x_i (categorical)
 - Split data for each categorical value of x_i
 - Send to each branch
 - If subset is pure, Stop
 - If impure split(data-subset)

Begin with
all data

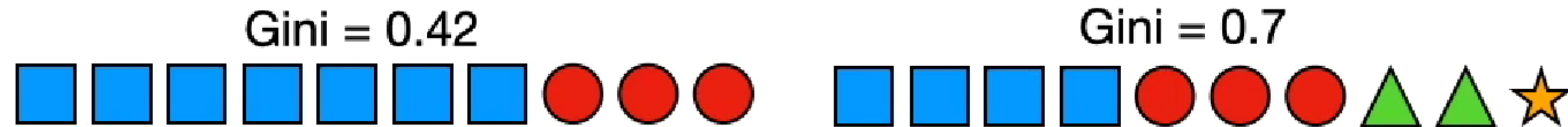
Discovered
independently
in 1983/84 by
Quinlan and
Breimanetal

Performance
depends on
attribute
picked for split



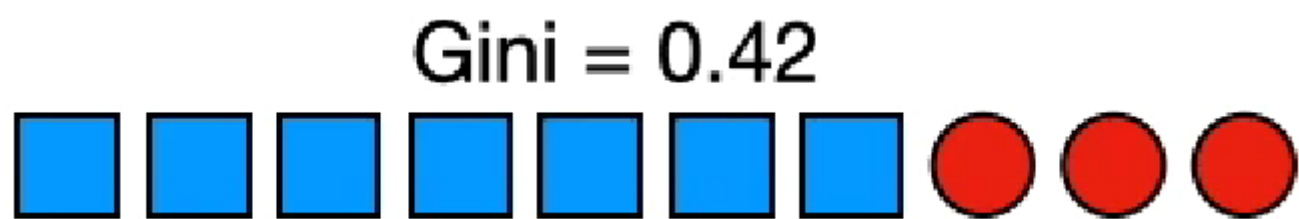
Gini Impurity Intuition

- Measures the mixture of categorical variable



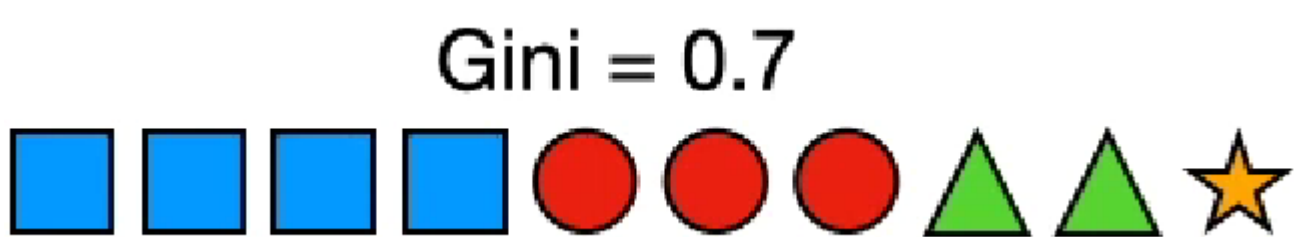
- Which is more diverse?
- How to measure diversity/impurity?

Gini Impurity Intuition - 2



		Same
		Different
		Different
		Same
		Same
		Different
		Same
		Same
		Different
		Same

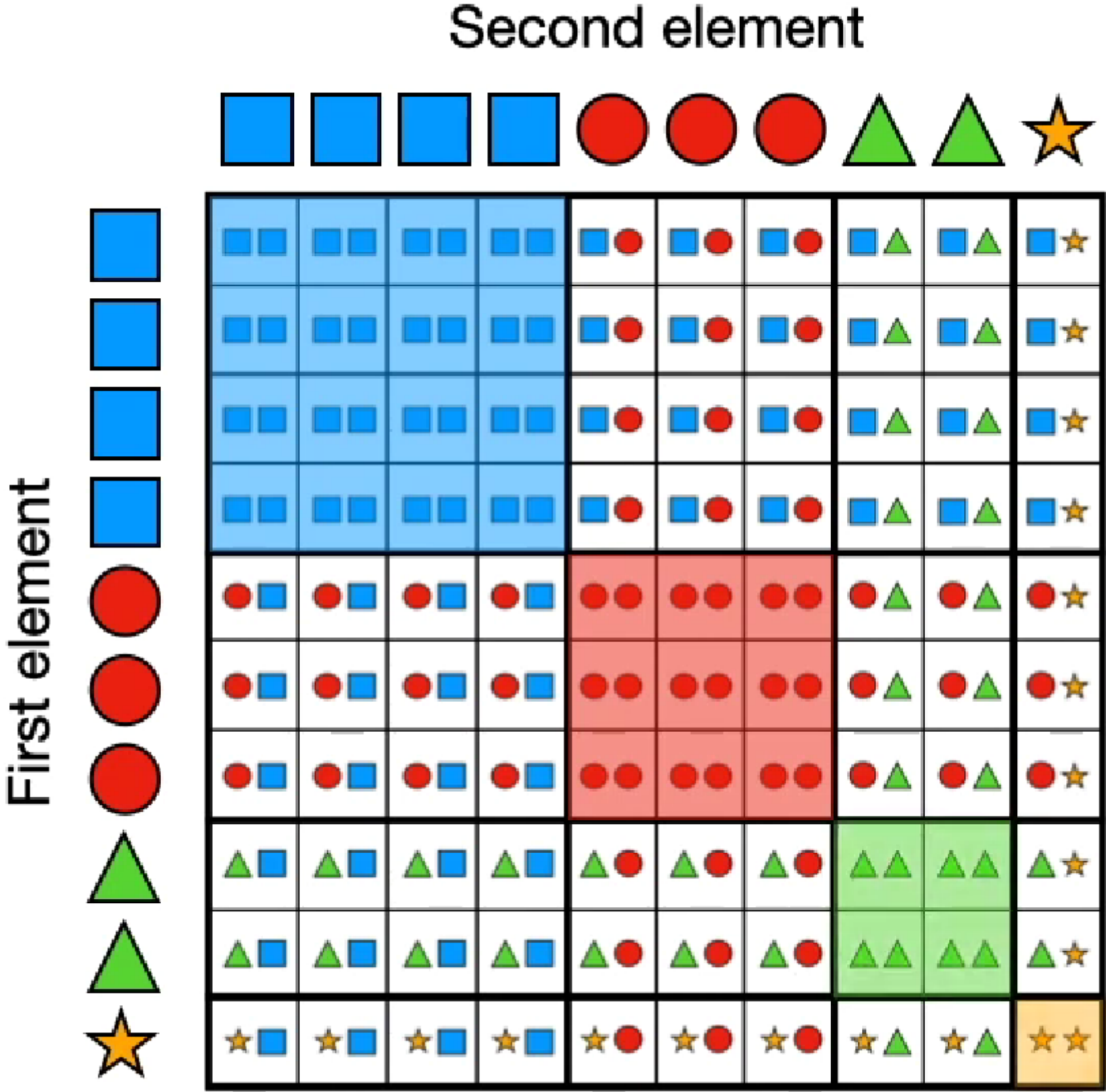
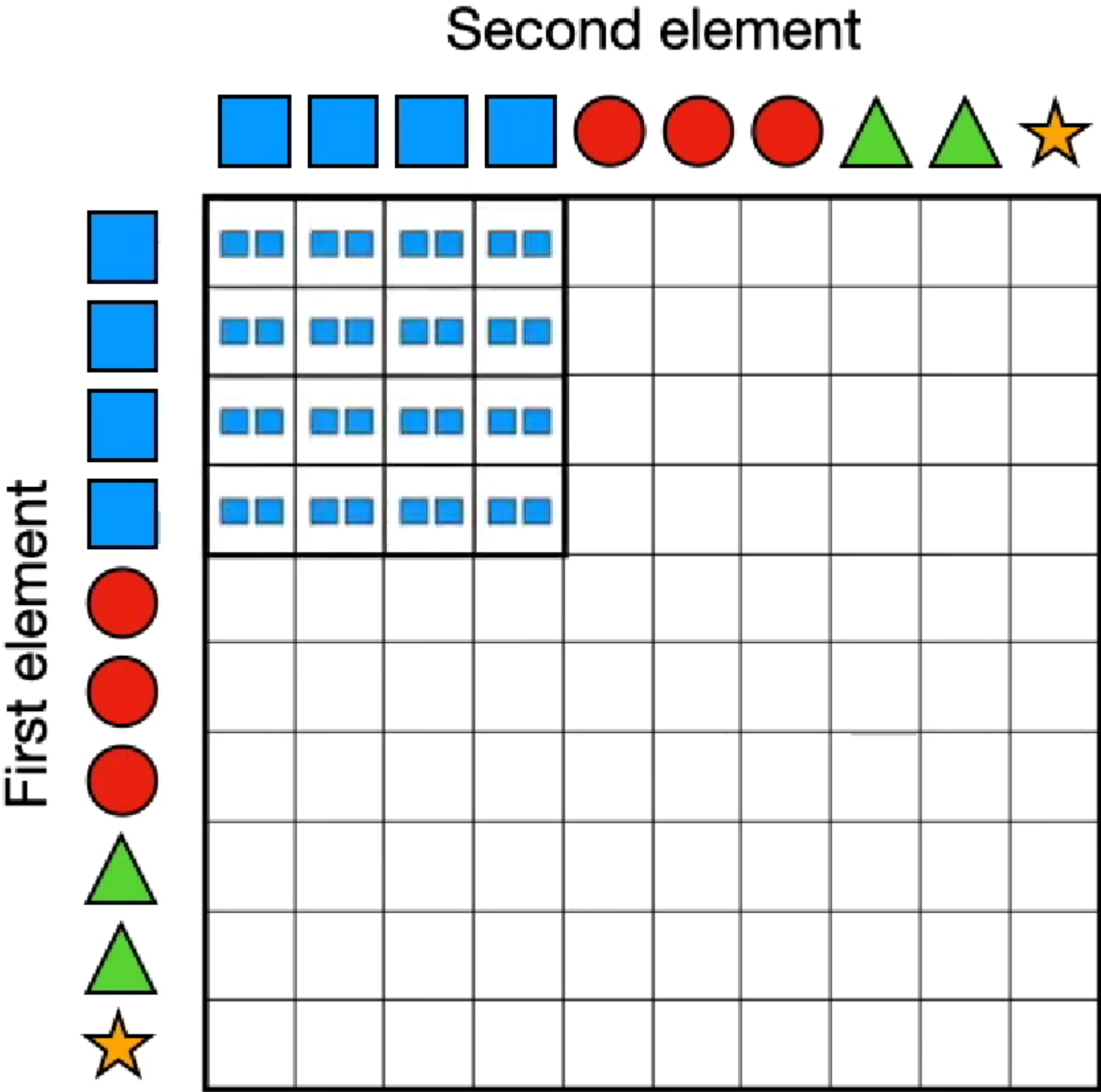
Different:
4 out of 10



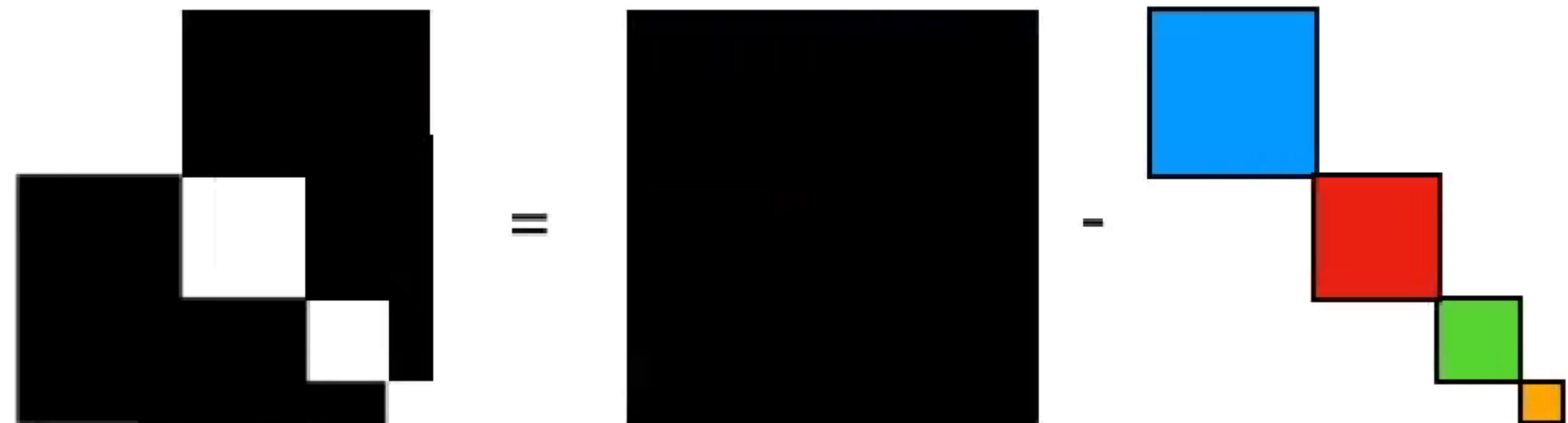
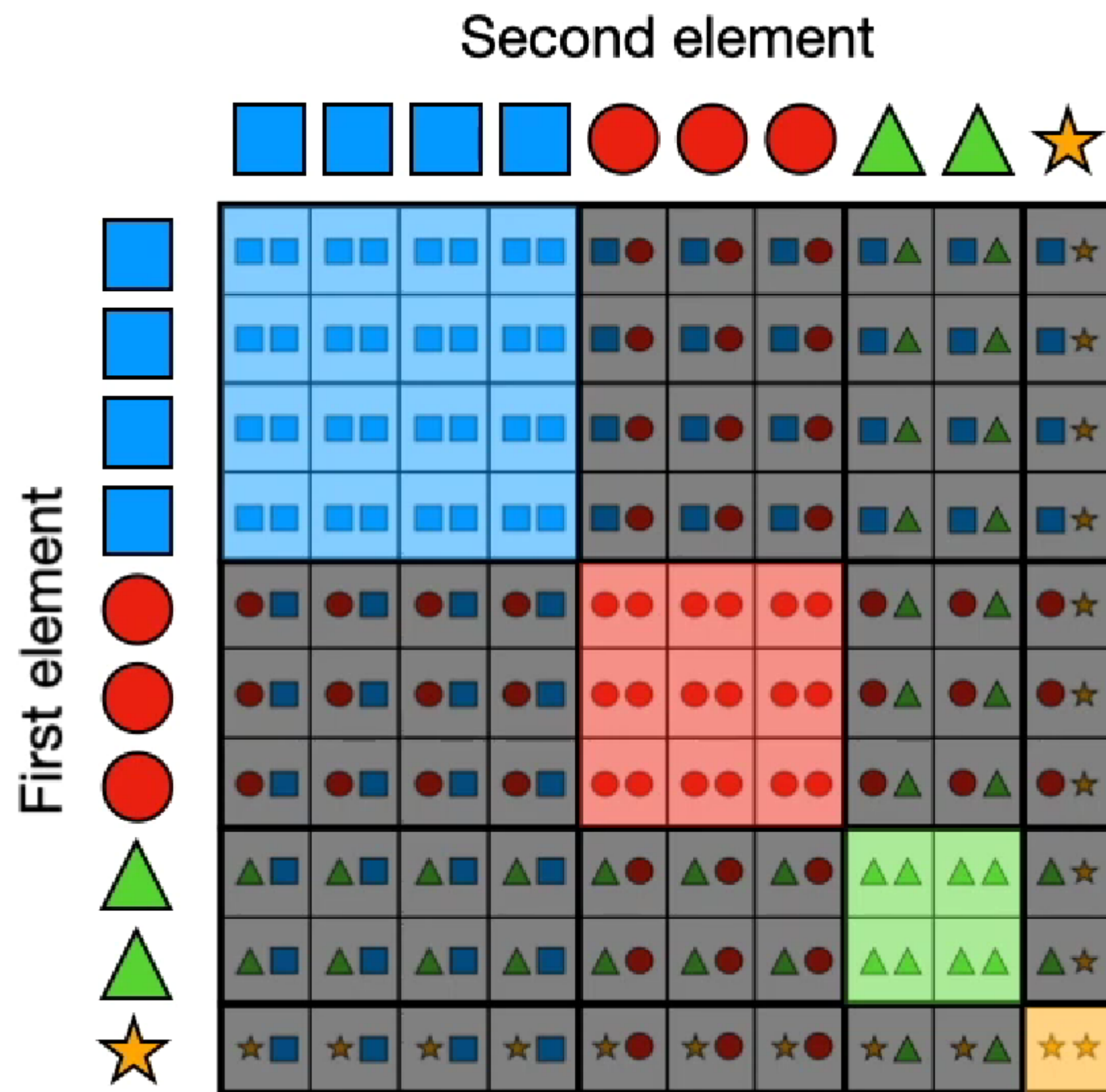
		Different
		Same
		Different
		Different
		Different
		Same
		Same
		Different
		Different
		Different

Different:
7 out of 10

Gini Impurity Intuition - 3

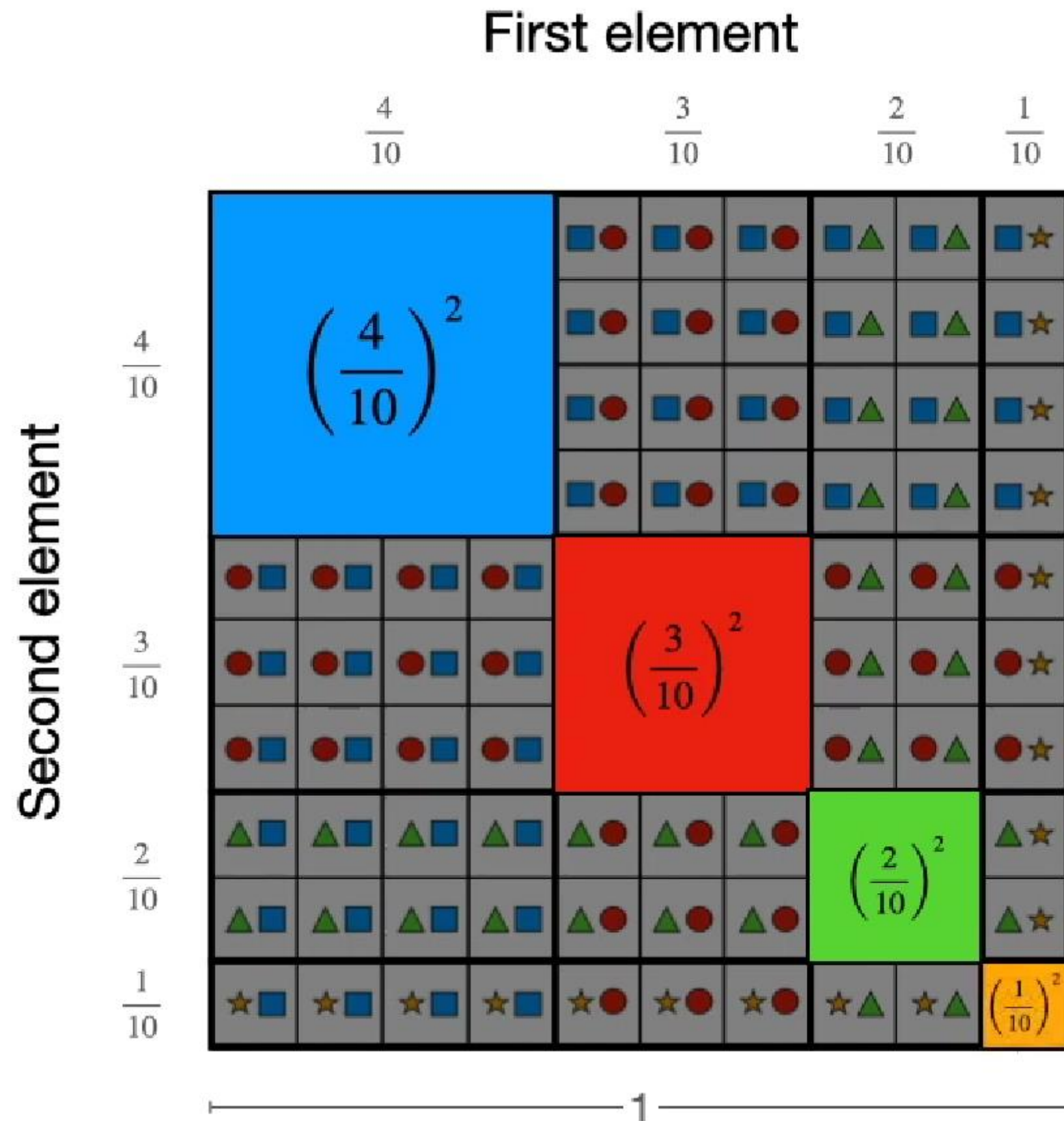


Gini Impurity Intuition - 4



$$\begin{aligned}
 P(\text{Both different}) &= P(\text{Anything}) - P(\text{Both equal}) \\
 &= 1 - P(\text{Both blue}) - P(\text{Both red}) - P(\text{Both green}) - P(\text{Both yellow})
 \end{aligned}$$

Gini Impurity Intuition - 5



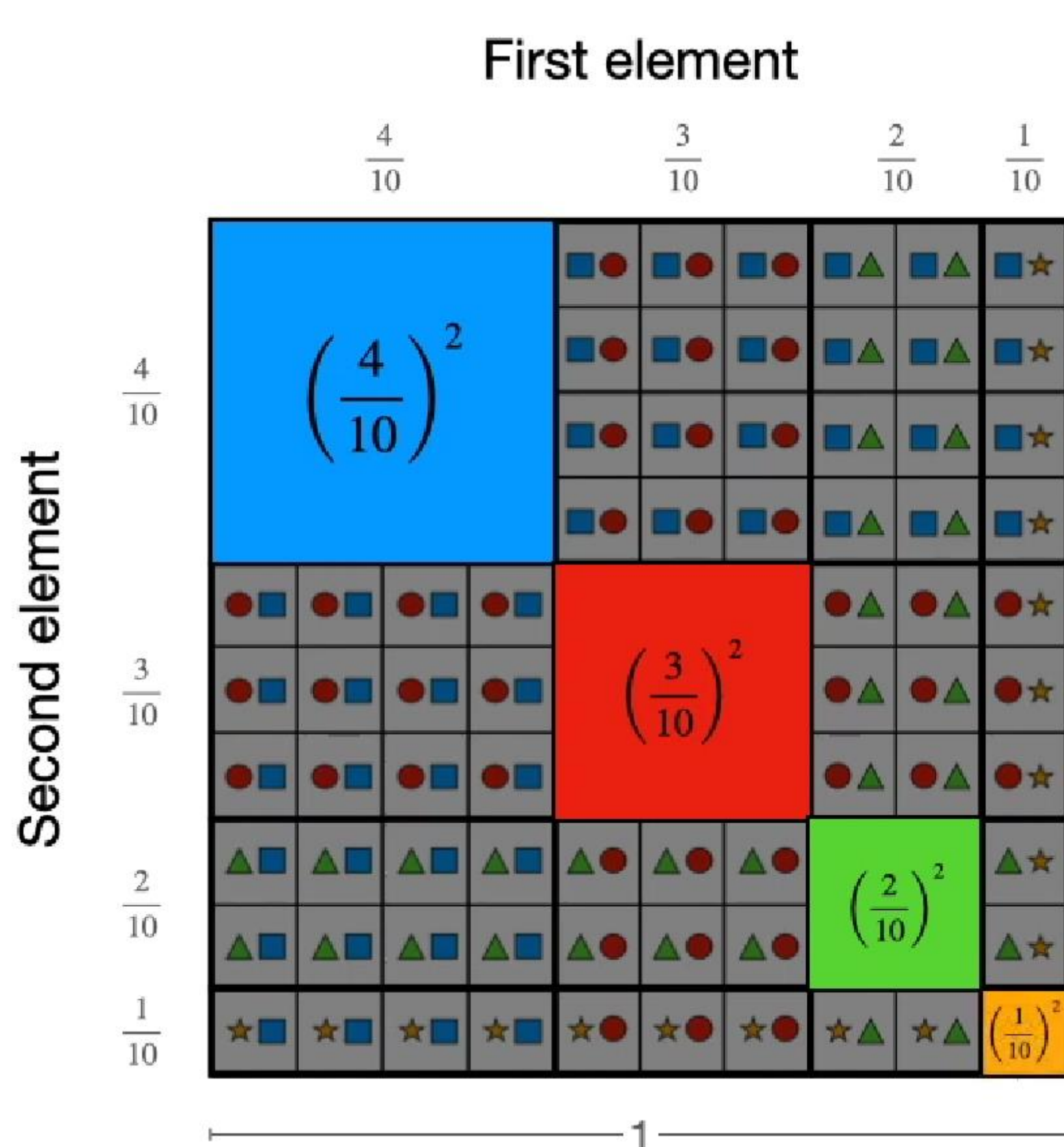
$$P(\text{Both different}) = P(\text{Anything}) - P(\text{Both equal})$$

$$= 1 - P(\text{Both blue}) - P(\text{Both red}) - P(\text{Both green}) - P(\text{Both yellow})$$

$$= 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{3}{10}\right)^2 - \left(\frac{2}{10}\right)^2 - \left(\frac{1}{10}\right)^2$$

$$GiniImpurity = 1 - \sum_{i=1}^k p_i^2$$

Gini Impurity Formula



$$GiniImpurity = 1 - \sum_{i=1}^k p_i^2$$

$$p_1 + p_2 + \dots + p_k = 1$$

$$GiniImpurity = p_1 + p_2 + \dots + p_k - \sum_{i=1}^k p_i^2$$

$$= \sum_{i=1}^k p_i - \sum_{i=1}^k p_i^2 = \sum_{i=1}^k p_i - p_i^2$$

$$= \sum_{i=1}^k p_i(1 - p_i)$$

Gini Impurity of dataset

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

- Calculate Gini Impurity of dataset target variable

$$1 - (P(\text{play} = \text{Yes})^2 + P(\text{play} = \text{No})^2)$$

$$1 - \left(\left(\frac{9}{14} \right)^2 + \left(\frac{5}{14} \right)^2 \right)$$

$$1 - (0.41 + 0.13) = 1 - 0.54 = 0.46$$

$$G(S) = 0.46$$

$$S = \{(x_1, y_1), \dots, (x_n, y_n)\} \quad y_i \in \{1, \dots, k\}$$

Gini Impurity of dataset

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

- Gini Impurity of dataset target variable

$$1 - (P(\text{play} = \text{Yes})^2 + P(\text{play} = \text{No})^2)$$

$$G(S) = 0.46$$

$$S = \{(x_1, y_1), \dots, (x_n, y_n)\} \quad y_i \in \{1, \dots, k\}$$

- Gini Impurity of target variable when split on an attribute

$$G(S|\text{Outlook} = \text{Sunny}) = ?$$

$$G(S|\text{Outlook} = \text{Overcast}) = ?$$

$$G(S|\text{Outlook} = \text{Rainy}) = ?^{27}$$

Selecting attribute for split using Gini

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

- Calculate Gini Impurity for each attribute split

$$G(S)_{\text{Outlook-split}}$$

$$G(S)_{\text{Humidity-split}}$$

$$G(S)_{\text{Wind-split}}$$

- Split on attribute with minimum resulting impurity

Gini Impurity after split

- Recall Total Probability – A and B are random variables
 - B takes two values 0 and 1

$$P(A) = P(A|B = 0)P(B = 0) + P(A|B = 1)P(B = 1)$$

- Recall Total Expectation – A and B are random variables

$$E(A) = E(A|B = 0)P(B = 0) + E(A|B = 1)P(B = 1)$$

- Gini Impurity can be interpreted similarly

$$G(S)_{Outlook} =$$

$$\begin{aligned} &G(S|Outlook = Sunny)P(Outlook = Sunny) + \\ &G(S|Outlook = Overcast)P(Outlook = Overcast) + \\ &G(S|Outlook = Rainy)P(Outlook = Rainy) \end{aligned}$$

**Expected Value of
G(S) when split
on outlook**

Gini Impurity for Outlook split

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

- Expected Value of Outlook split

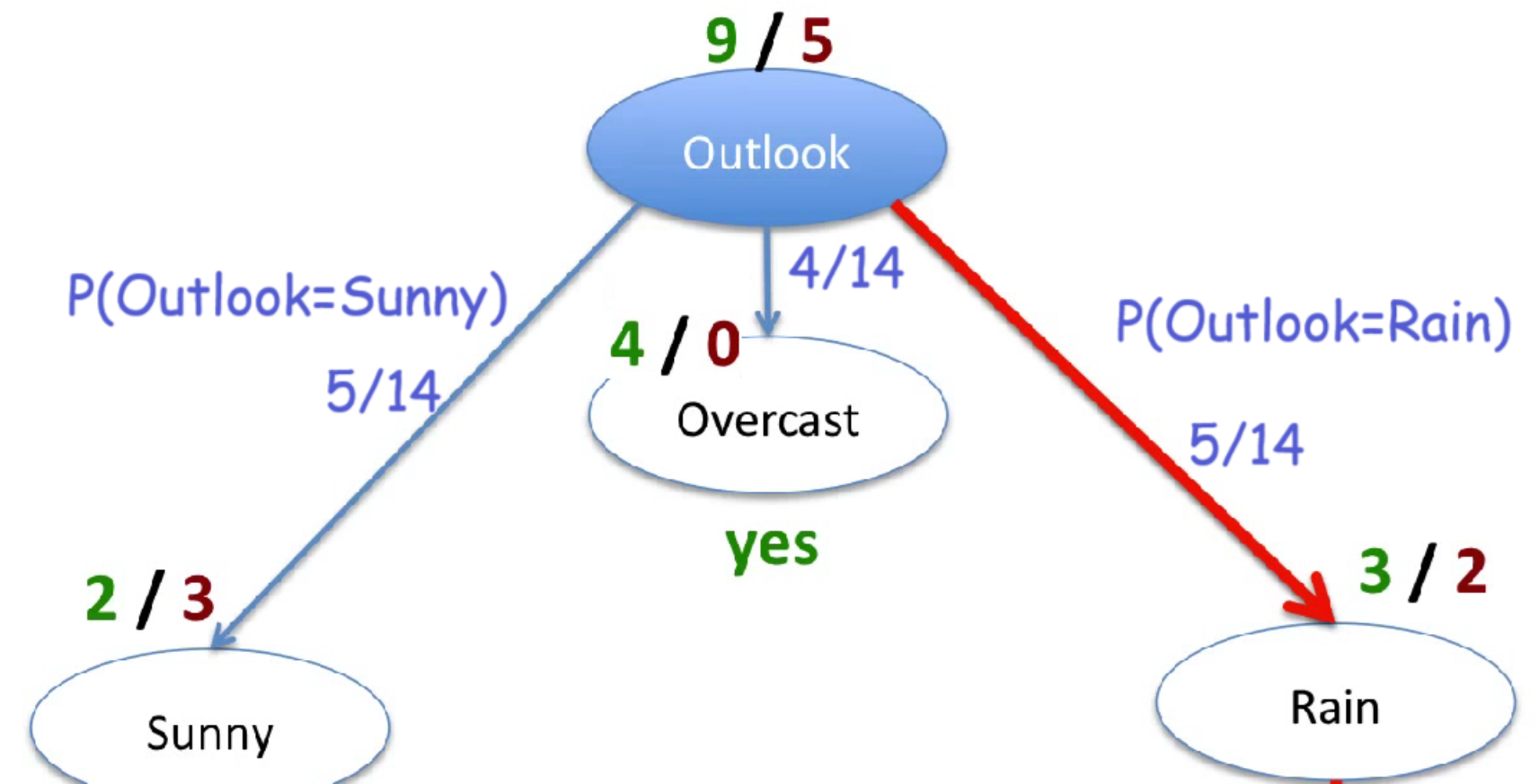
$$G(S)_{Outlook} =$$

$$G(S|O = Sunny)P(O = Sunny) + G(S|O = Overcast)P(O = Overcast) + G(S|O = Rainy)P(O = Rainy)$$

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes

Training examples: **9 yes / 5 no**

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No



$$\begin{aligned}
 G(S|O = \text{Sunny}) &= 1 - ((2/5)^2 + (3/5)^2) \\
 &= 1 - (0.16 + 0.36) = 0.48
 \end{aligned}$$

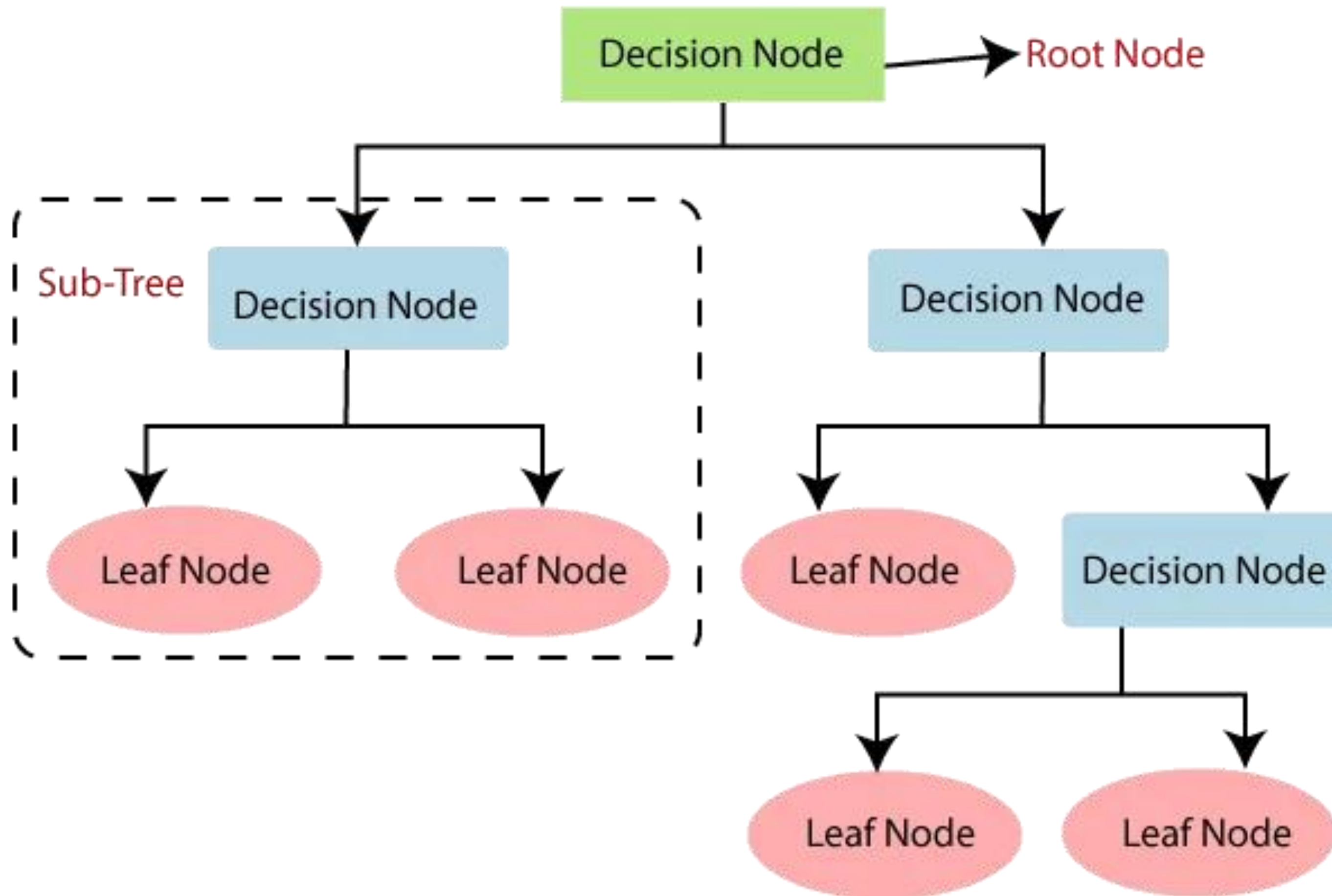
$$\begin{aligned}
 G(S|O = \text{Overcast}) &= 1 - ((4/4)^2 + (0/4)^2) \\
 &= 1 - (1 + 0) = 0
 \end{aligned}$$

$$\begin{aligned}
 G(S|O = \text{Rain}) &= 1 - ((2/5)^2 + (3/5)^2) \\
 &= 1 - (0.16 + 0.36) = 0.48
 \end{aligned}$$

$$G(S)_{\text{Outlook}} = 0.48 * 5/14 + 0 * 4/14 + 0.48 * 5/14$$

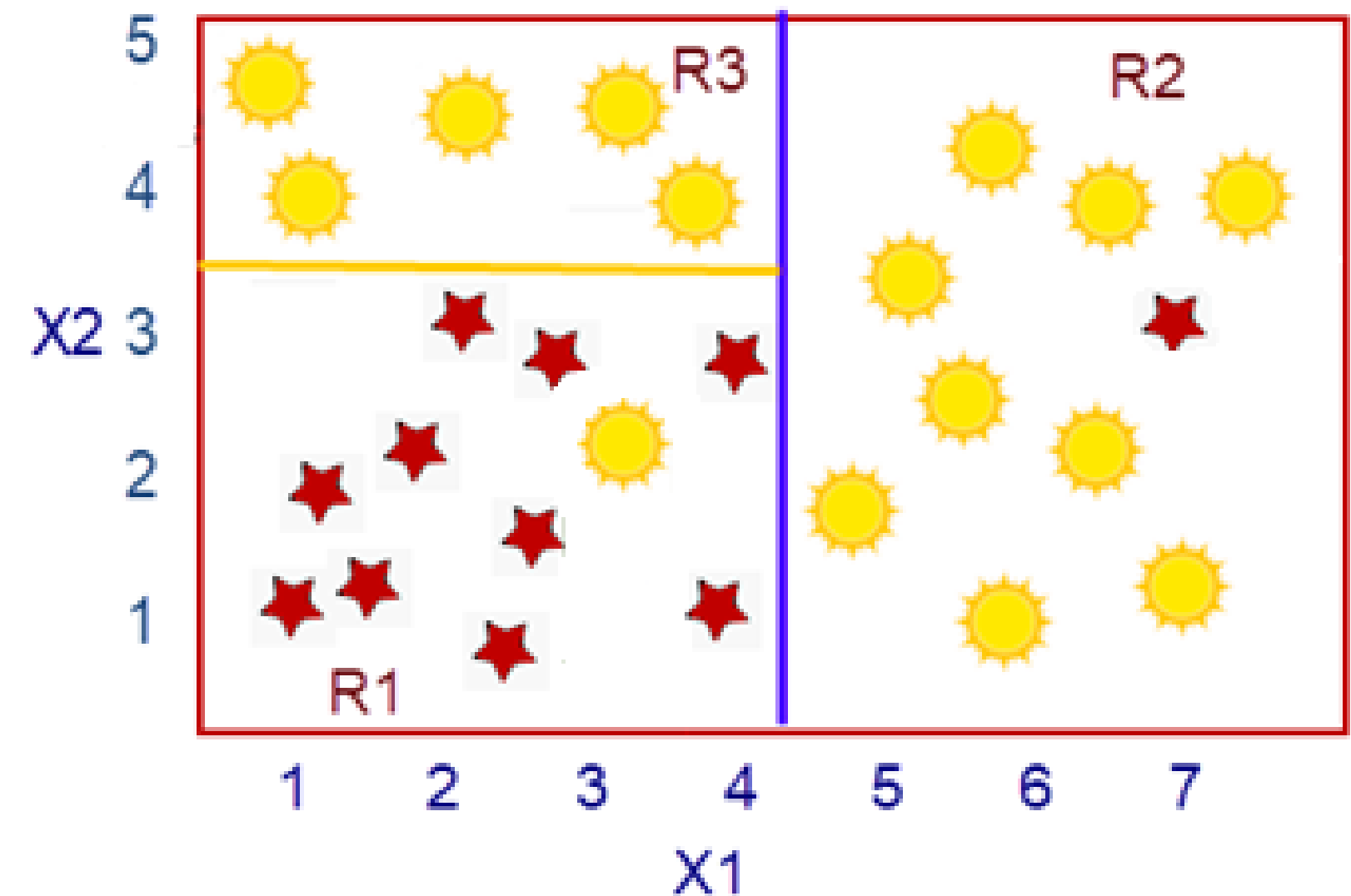
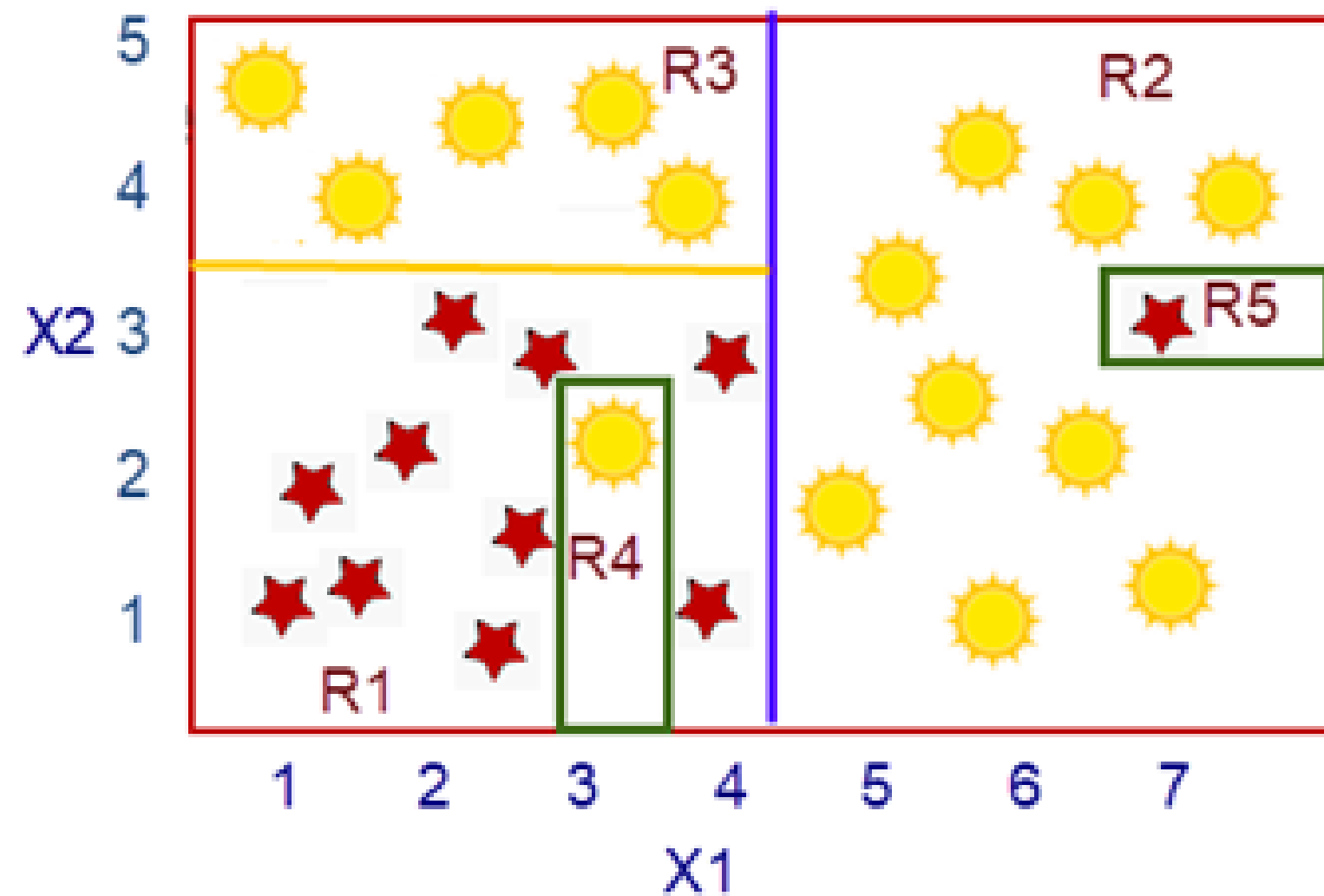


Decision Tree overfitting



How long should split be done?

- Split till all entries in a node are pure.
 - 0 Gini impurity
 - Or stop early?



Measuring Overfitting

- Low training error, falling & increasing validation error

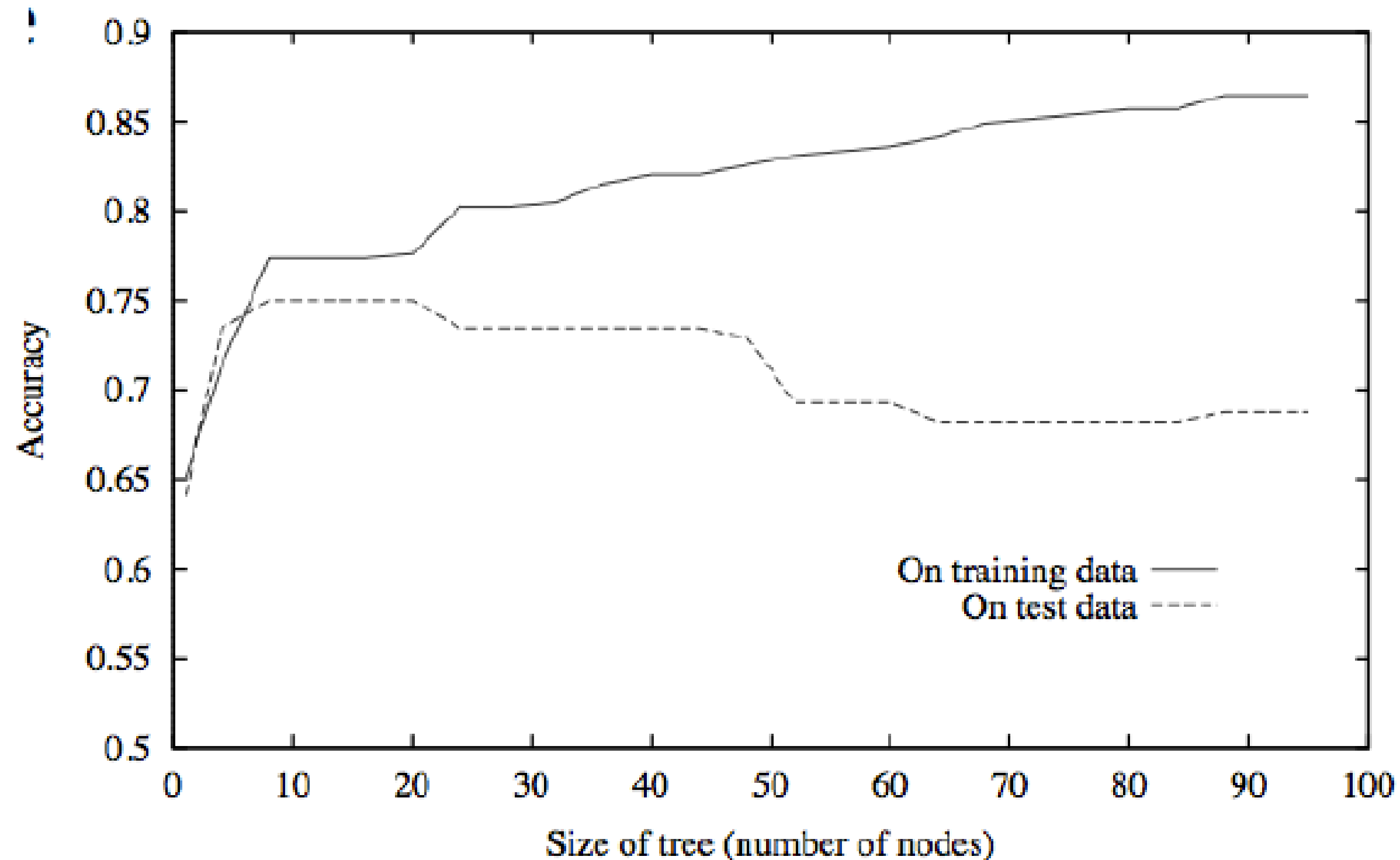


Figure credit: Tom Mitchell, 1997

Pre-pruning

- Done before training a decision tree
- Control Max-Depth of the tree
- Min samples required to split at a node
- Min samples that should be present at a leaf node

Post pruning

- Done after the Decision Tree is overfit
- Reduced Error Pruning
 - Remove a node and the subtree.
 - Check if the resulting tree performs better on validation set
 - If yes, the retain the pruning, else abstain from pruning
- Rule based pruning

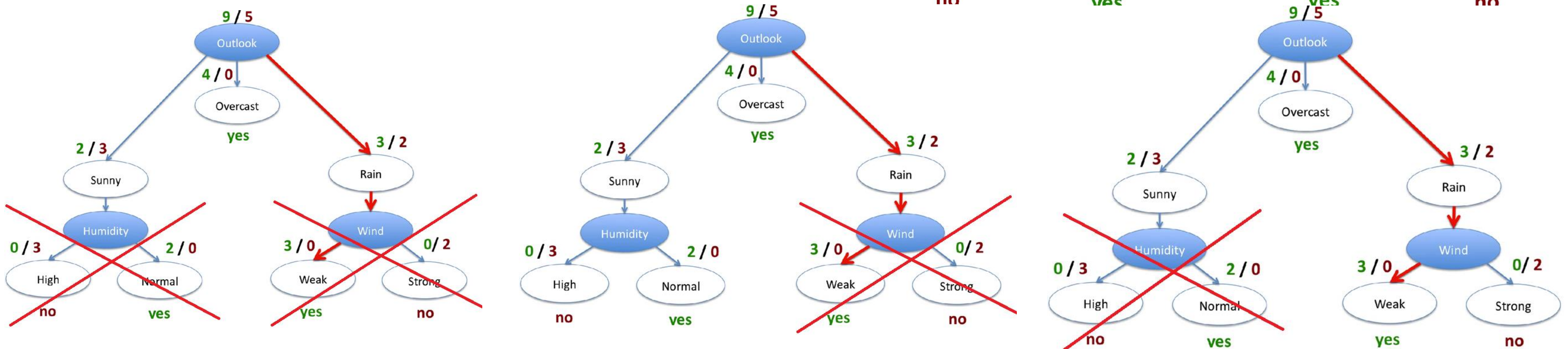
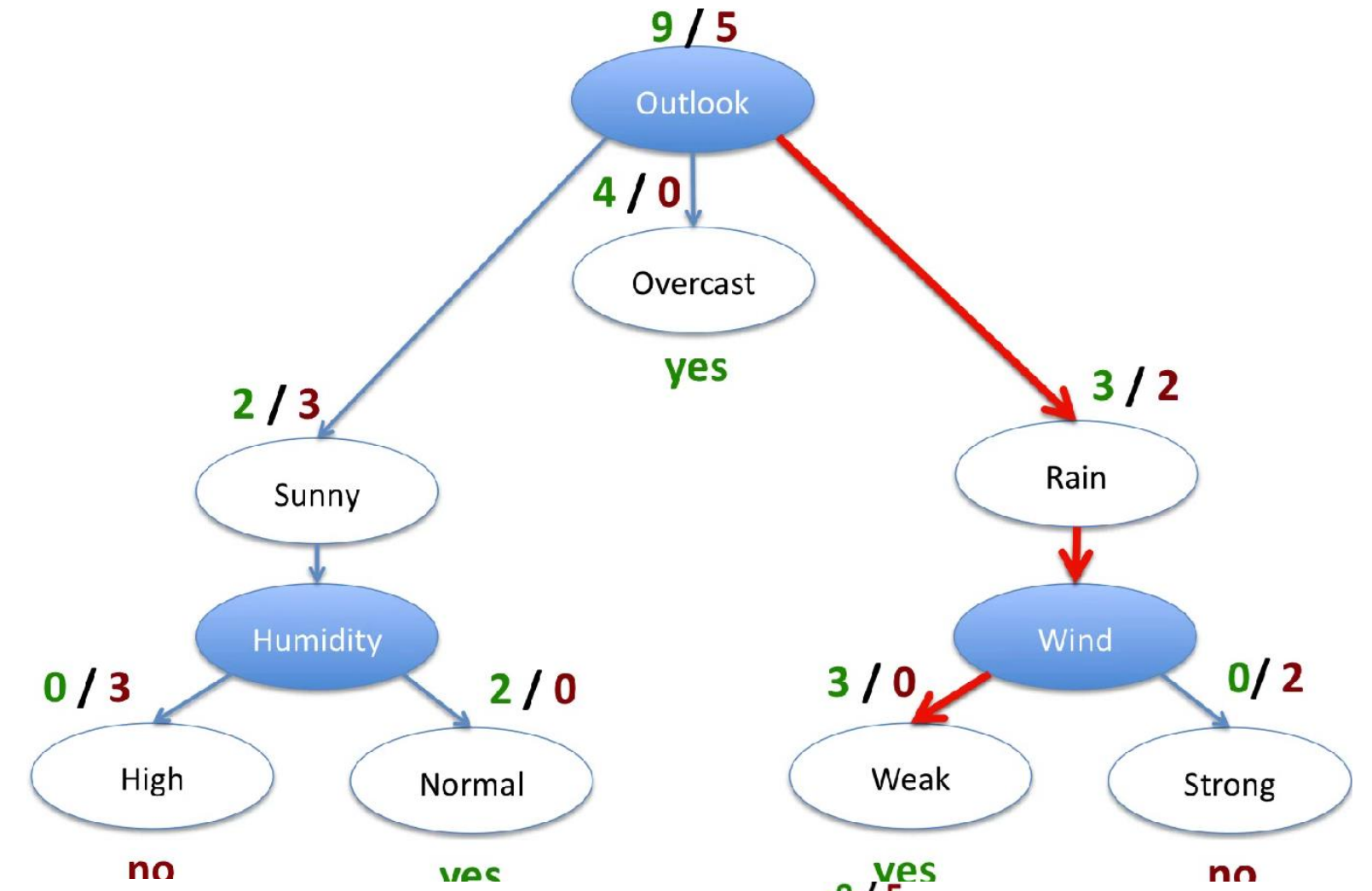
Post Pruning - Cost Complexity Pruning

- Total Loss of a tree with **ALL** data

$$\mathcal{L} = \sum_{i \in \text{features}} G(S)_{\text{feature}_i}$$

- Total loss after adding penalty

$$\mathcal{L} = \sum_{i \in \text{features}} G(S)_{\text{feature}_i} + \alpha|T|$$



Better methods than post pruning

- Post pruning is of theoretical interest these days
- Overfitting mitigated with RandomForest instead of pruning



QUESTIONS