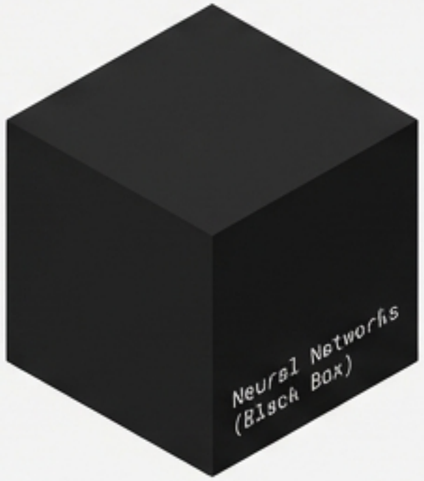
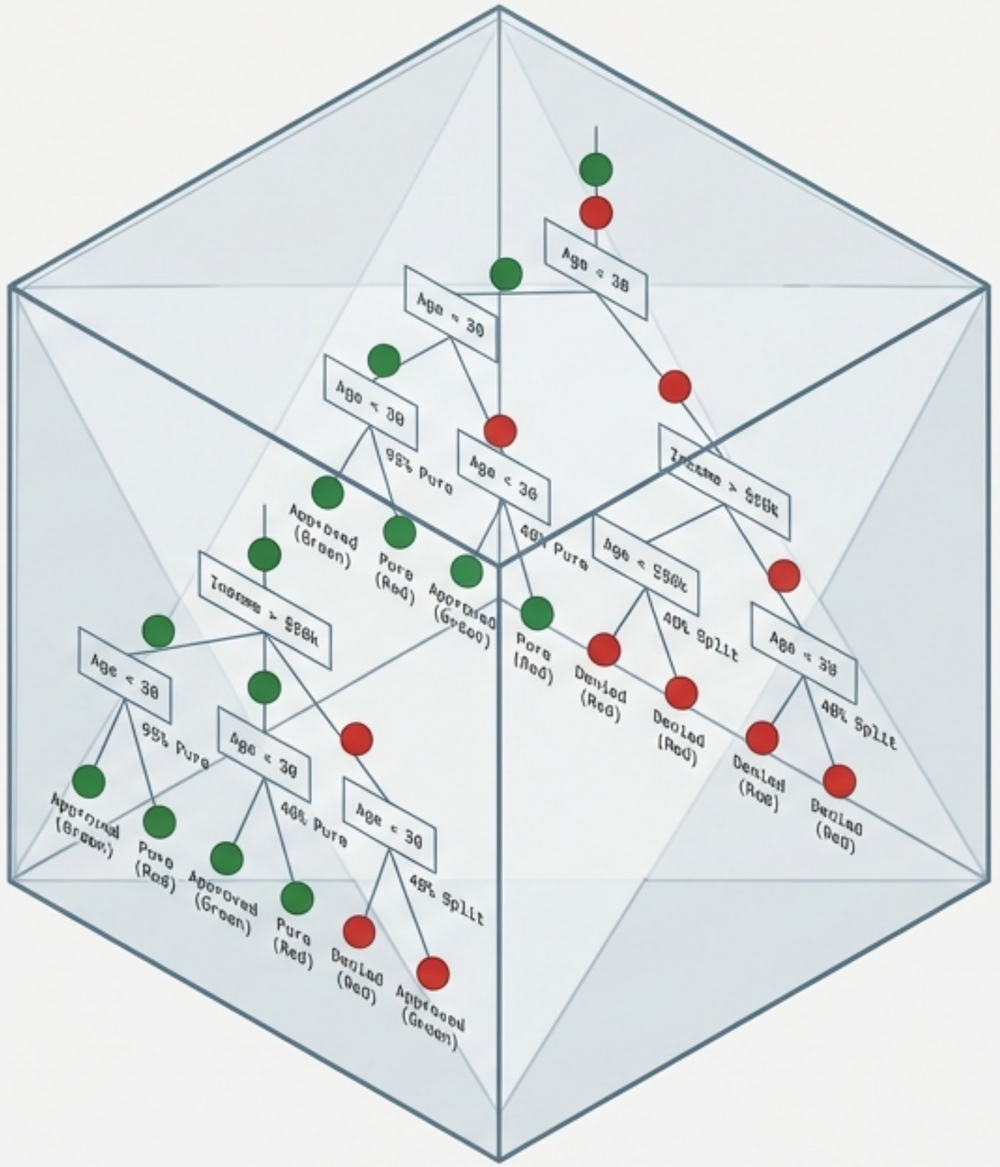
# DECISION TREES

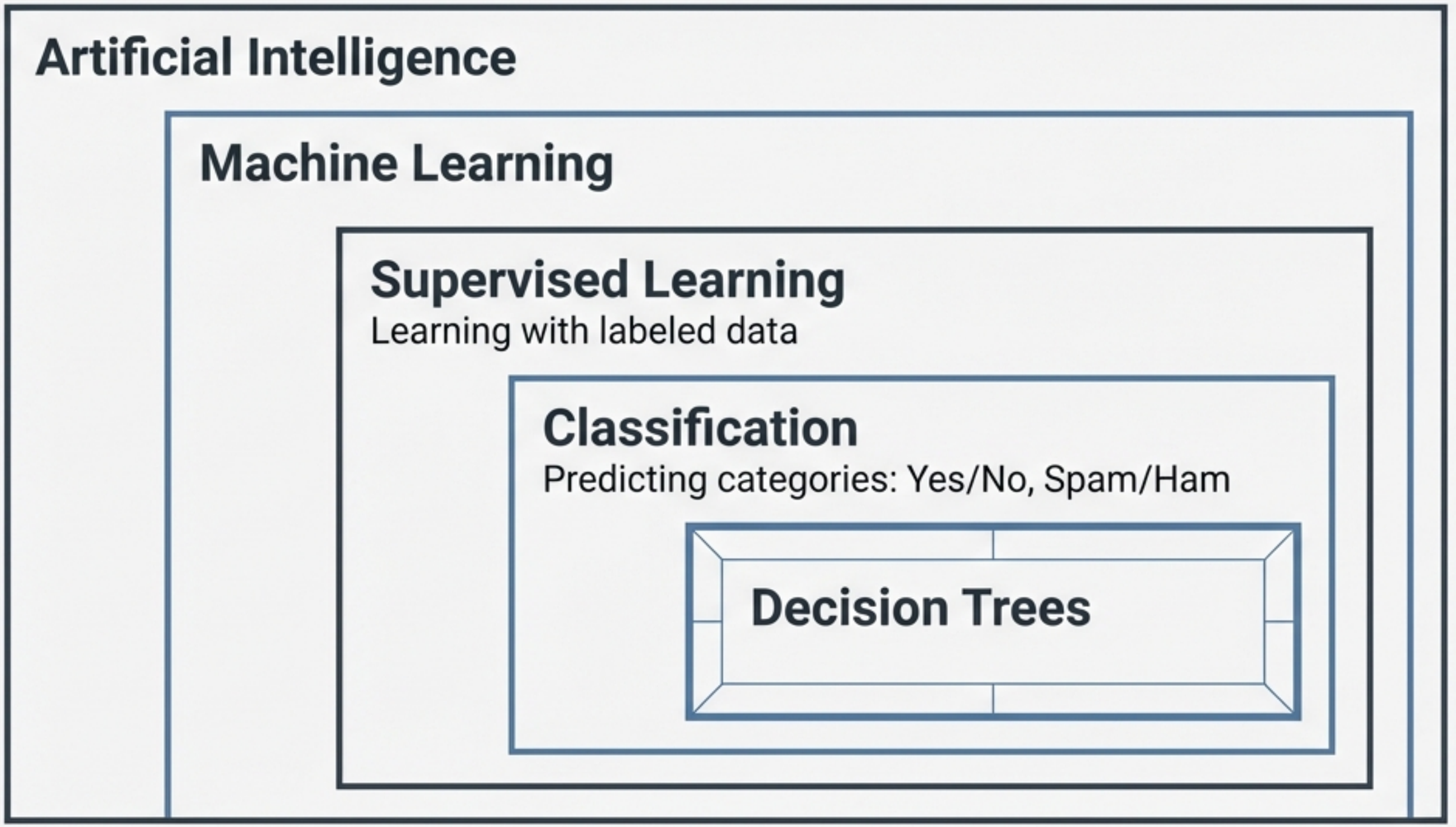
## The Transparent Logic of Machine Learning




From Human Intuition to the Math of Choice

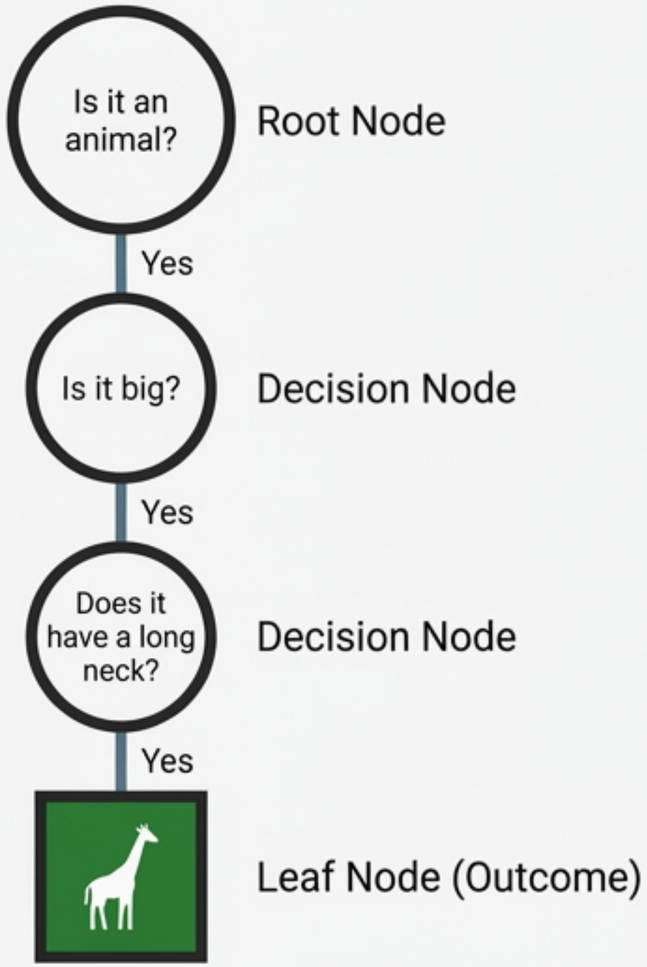
# A Classification Tool for Supervised Learning




THE GOAL: To predict a discrete class label for new, unseen data.

Just as a child learns to distinguish a dog from a cat by seeing labeled examples, the algorithm builds a logic map based on training data.

# The Anatomy of a Decision: The ‘20 Questions’ Analogy




ROOT NODE:
The starting point representing the entire population.

SPLITTING:
Dividing a node into two or more sub-nodes based on a condition.

LEAF NODE:
The terminal node where a final classification is made.

# The Core Challenge: Which Question First?

With thousands of features available, the algorithm must calculate which attribute provides the most “clue” to reduce uncertainty.

| Income | Credit History | Age | Loan Amount |
|---|---|---|---|
| 50k | Good | 30 | 10k |
| 75k | Bad | 45 | 20k |
| 100k | Good | 25 | 5k |
| 40k | Bad | 50 | 15k |
| 80k | Good | 35 | 30k |

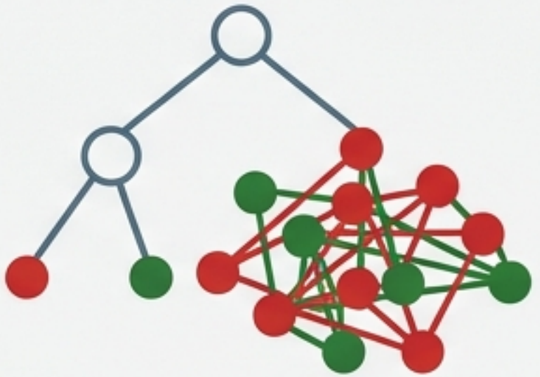
Bank Loan Applicants

Split by Age?

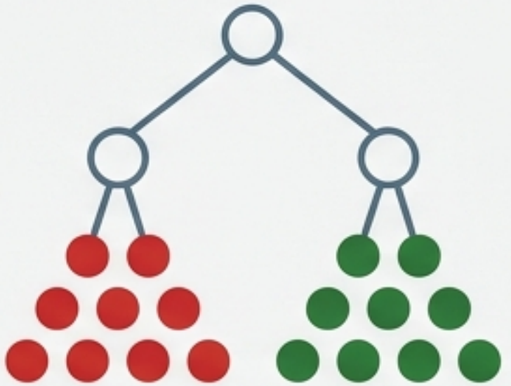
$H_x$

Entropy (H)

Split by Credit History?

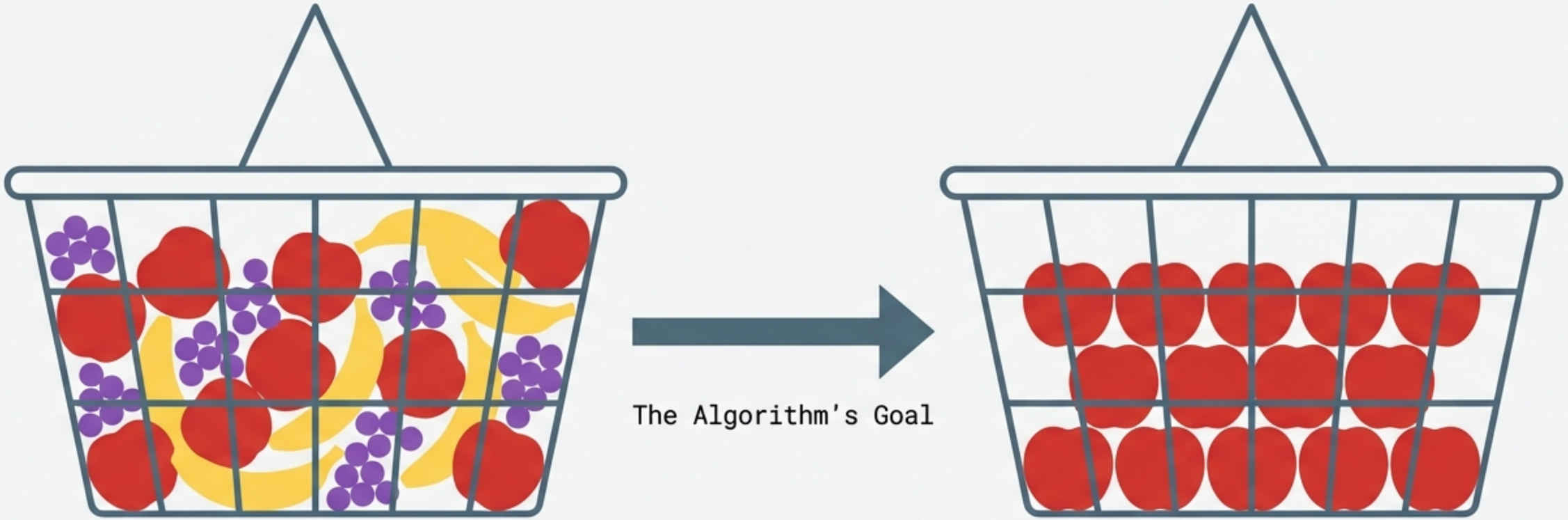


High Uncertainty



Low Uncertainty

# Measuring Impurity: The Fruit Basket Analogy




High Entropy (Impure)

Hard to predict random pick.

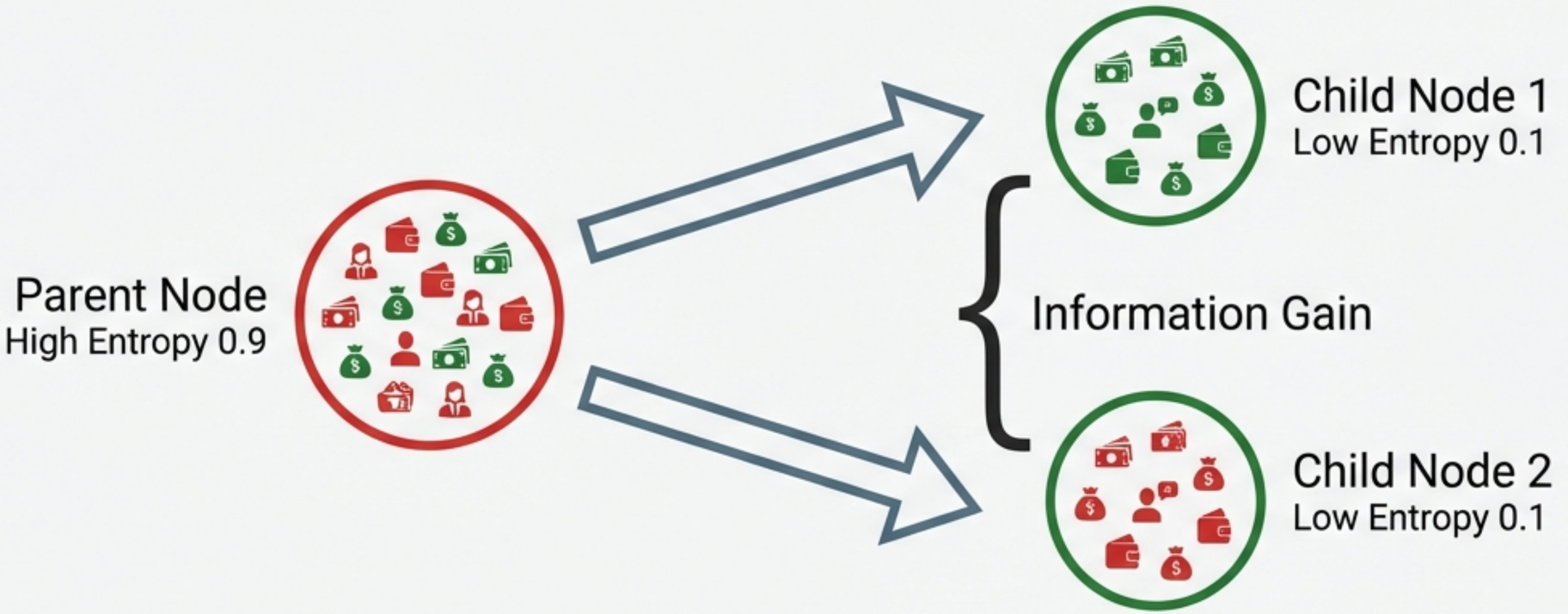
Low Entropy (Pure)

Easy to predict random pick.

Entropy is a mathematical measure of randomness. The algorithm seeks splits that turn High Entropy mixed buckets into Low Entropy pure buckets.

# Information Gain: The Math of Choice

```
Information Gain = Entropy(Parent) - Weighted Avg[Entropy(Children)]
```




The algorithm calculates this "Gain" for every single feature (Income, Age, Debt).
The feature that yields the HIGHEST Information Gain is selected as the next decision node.

# The ‘Play Outside’ Example: The Training Data

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

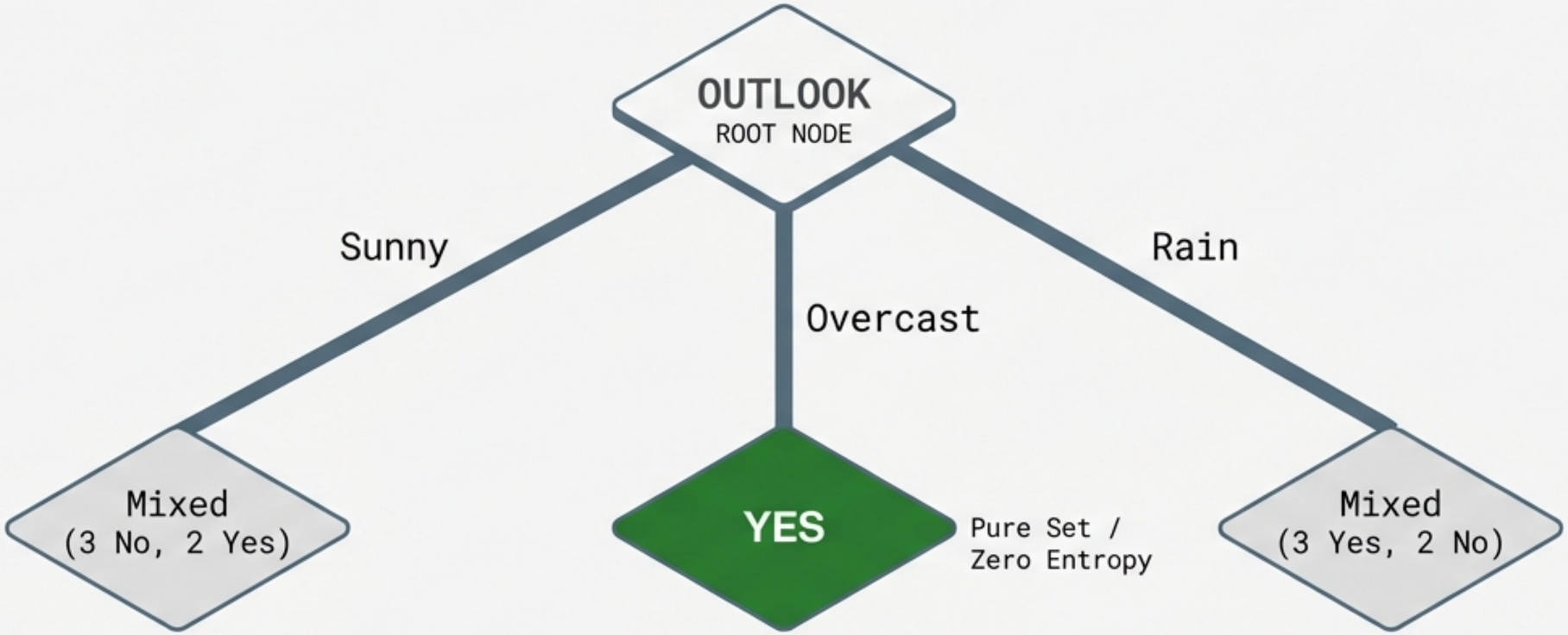
Total Days: 14

Decision YES: 9

Decision NO: 5

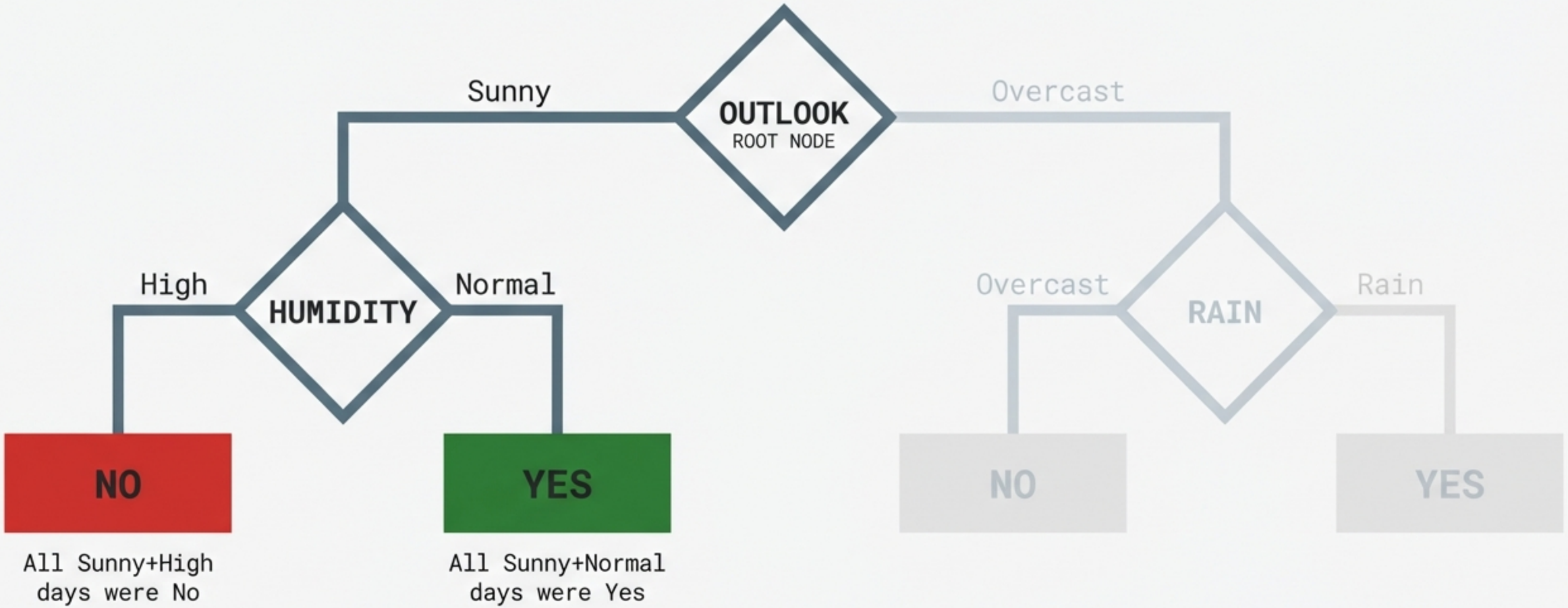
Goal: Learn the rules of weather that predict the decision.

# Step 1: Selecting the Root Node (Outlook)

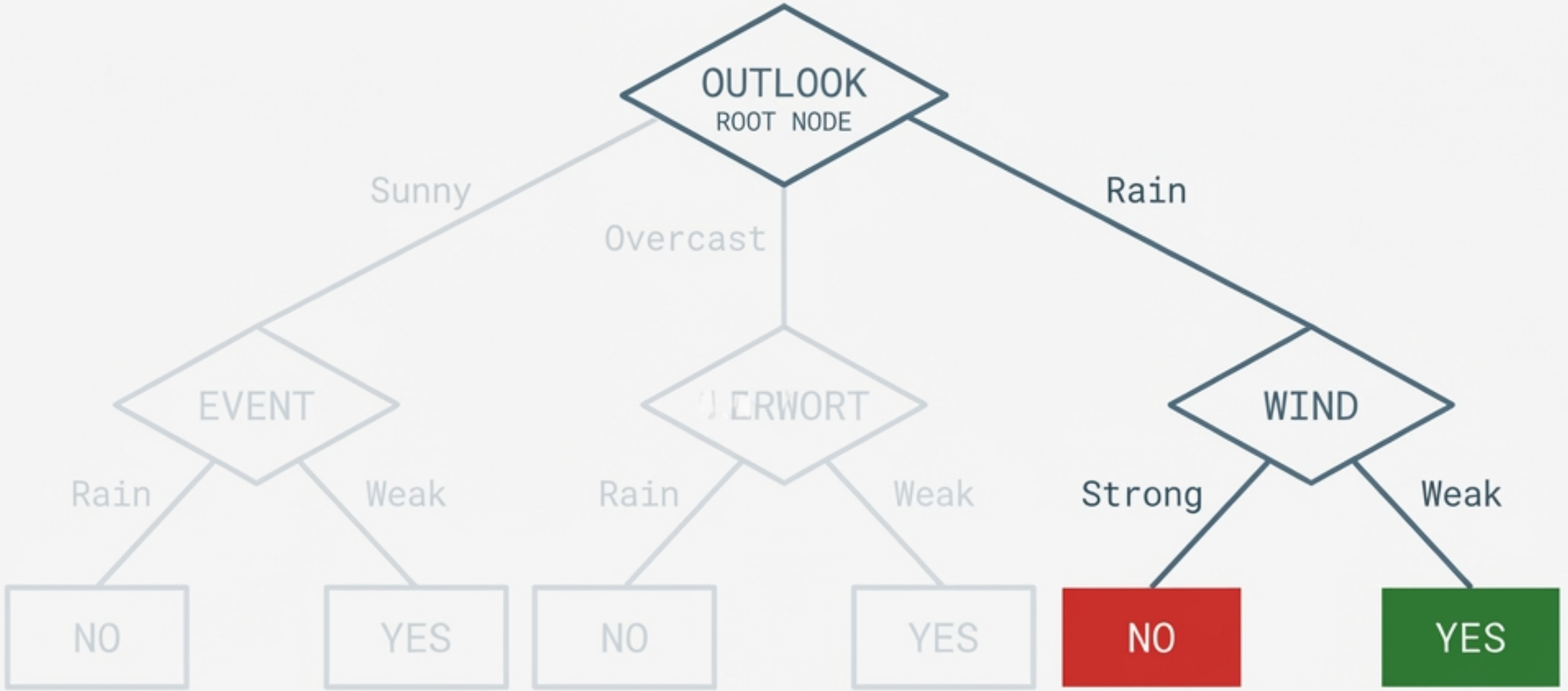



Why Outlook? It provides the highest Information Gain. Notice "Overcast" is already a pure result—always play on overcast days.

# Step 2: Resolving the “Sunny” Branch

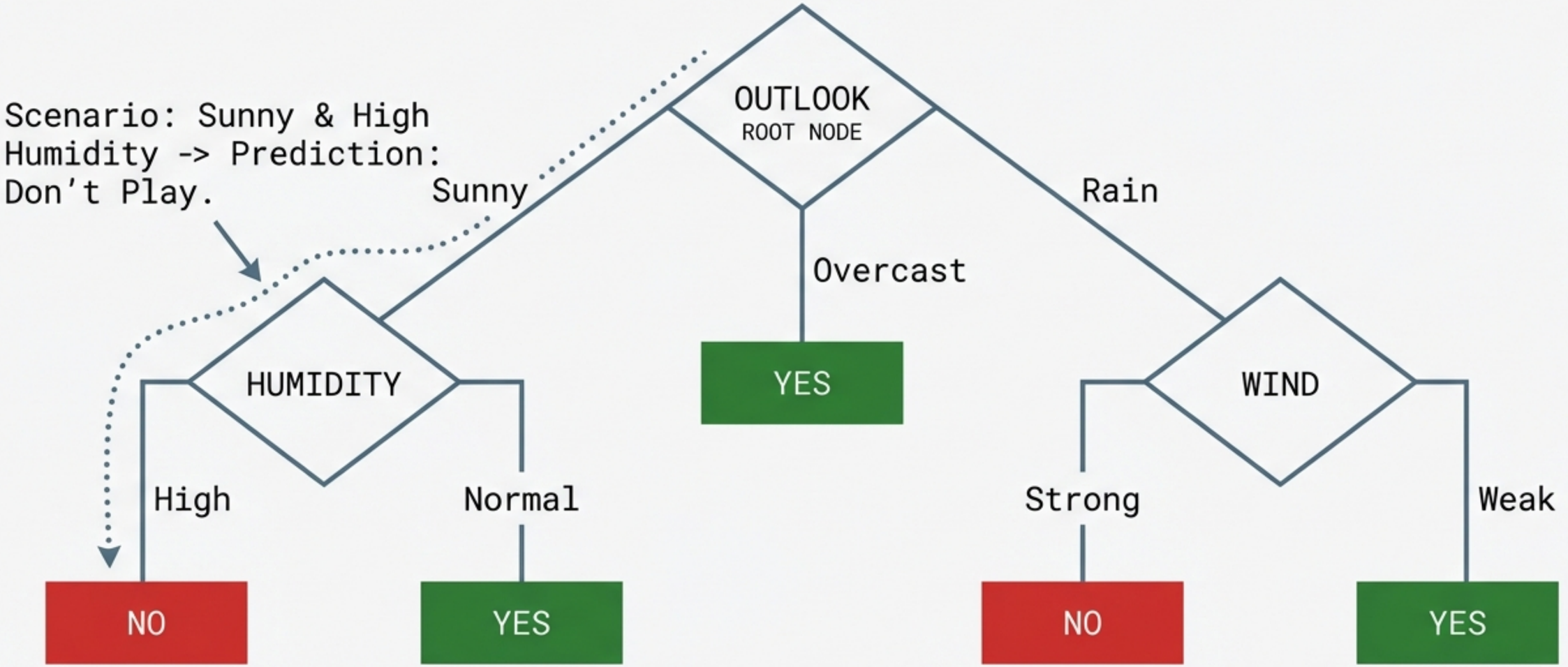


# Step 3: Resolving the ‘Rain’ Branch



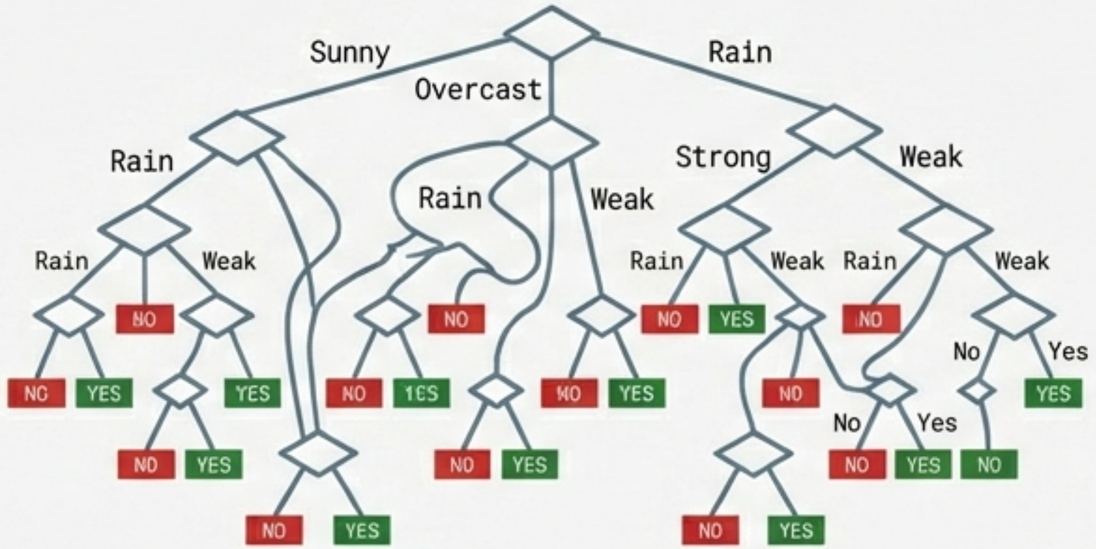

The machine learns: Rain alone isn’t a dealbreaker, but Rain + Strong Wind is.

# The Final Model: A Map for Prediction



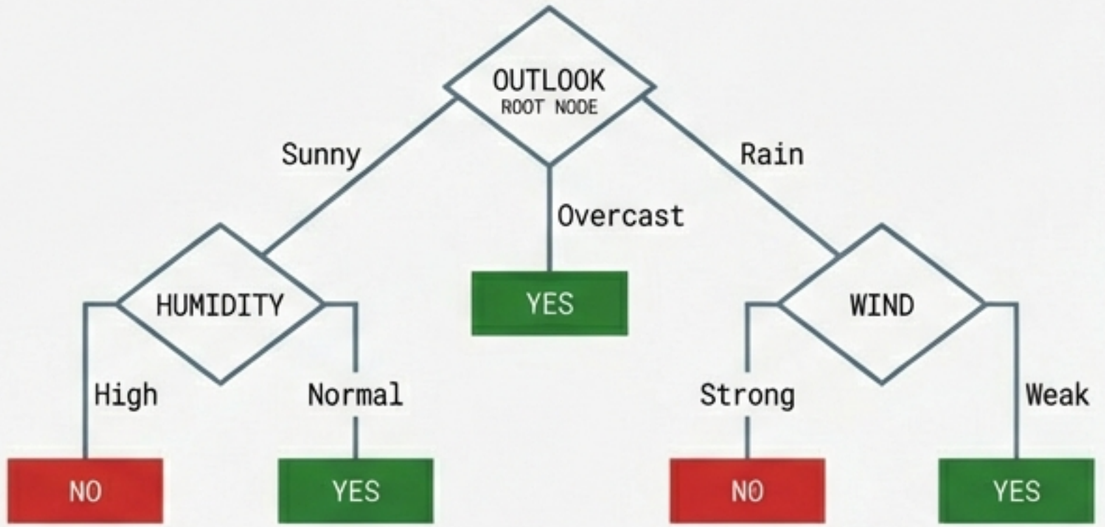
# The Danger of Memorization (Overfitting)

## Overfitted




Memorizes noise. Fails on new data.

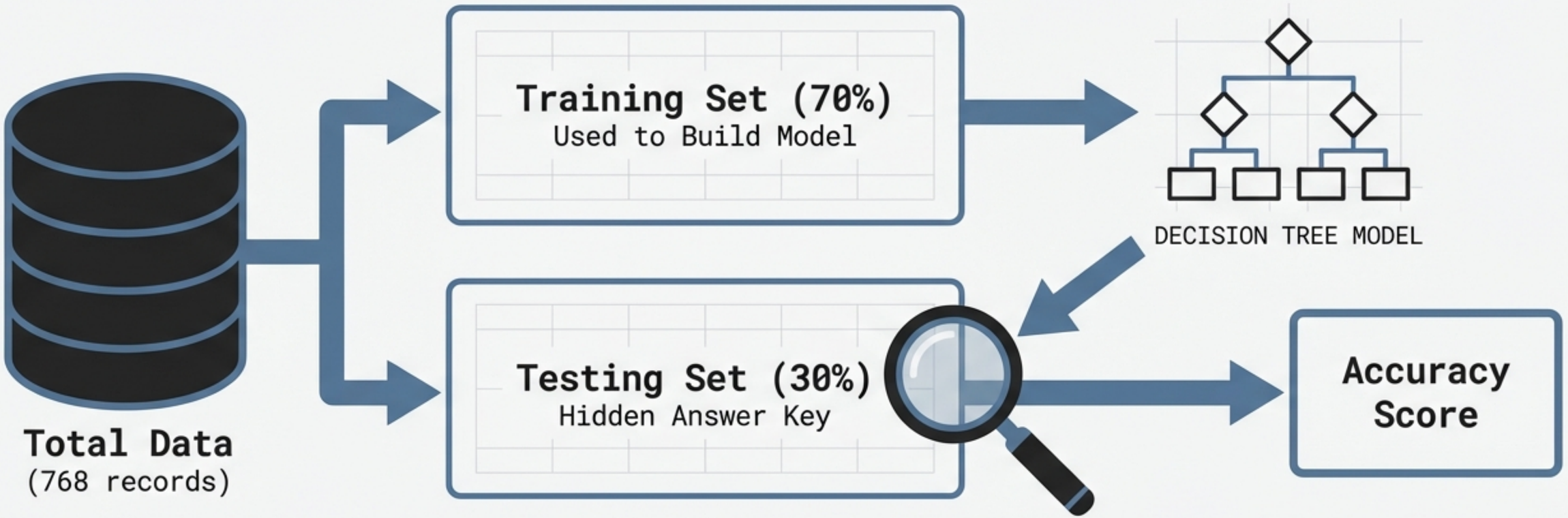
## Pruned / Generalized




Learns patterns. Adapts to new data.

Overfitting occurs when a tree grows so complex it captures errors and noise in the training data. “Pruning” cuts back these branches to improve general performance.

# Evaluation: The “Exam” Concept




We cannot test the student with the same questions used in the lecture. We hide 30% of the data to evaluate if the model truly learned.

# Beyond Accuracy: The Confusion Matrix

Visualizing the Cost of Error in Diabetes Prediction


**High Risk Error:** The model said "Healthy" but the patient has diabetes. In healthcare, we prioritize minimizing this box.

# Summary: The Glass Box Advantage

## PROS

- **Interpretable:** Logic is transparent to stakeholders.
- **Flexible:** Handles numerical and categorical data.
- **Non-Linear:** Captures complex relationships.

## CONS

- **High Variance:** Sensitive to small data changes.
- **Overfitting:** Prone to complexity without pruning.

**Decision Trees act as the fundamental building blocks for powerful ensemble methods like Random Forests, bridging human logic and machine power.**