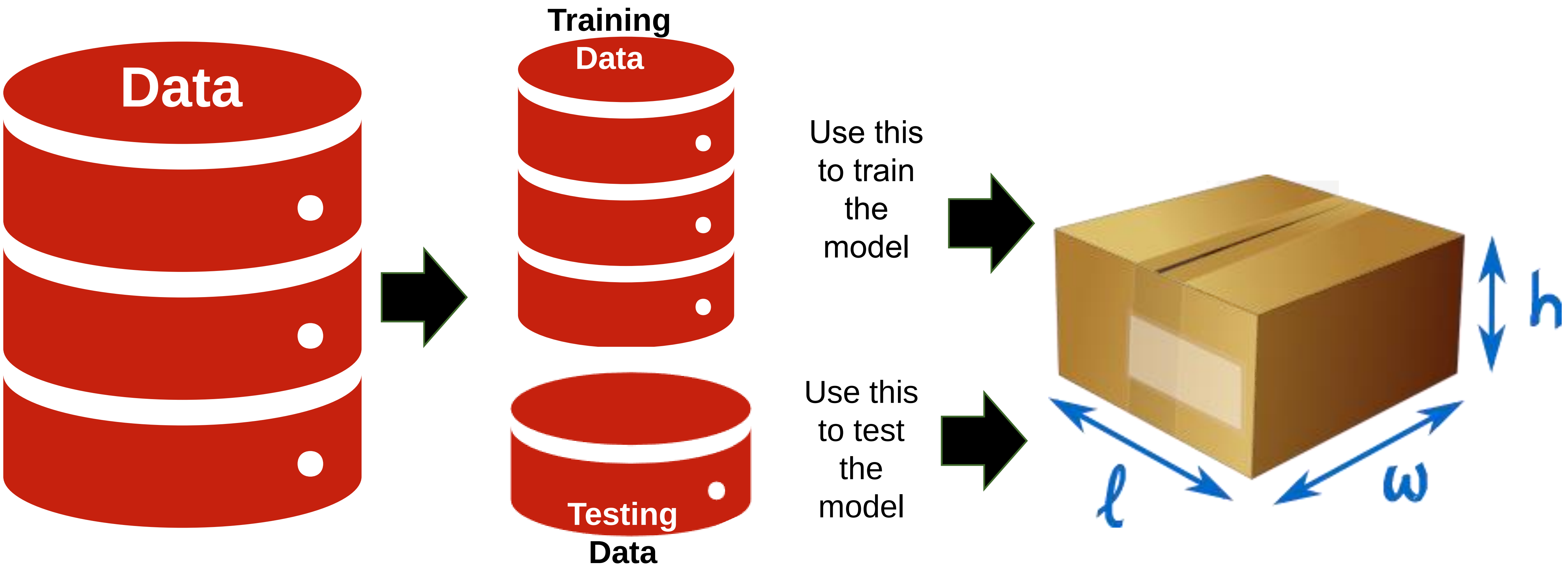
IDENTIFYING AND UNDERSTANDING GROUPS

PART 2 - Informed decision making.

Predictive Modelling.

- Predictive analytics refers to a **series of techniques concerned with making more informed decisions based on an analysis of historical data.**

Dataset.


Data
Training
Data
Testing
Data
Use this
to train
the
model
Use this
to test
the
model
h
ℓ
ω

BACK to the FUTURE.

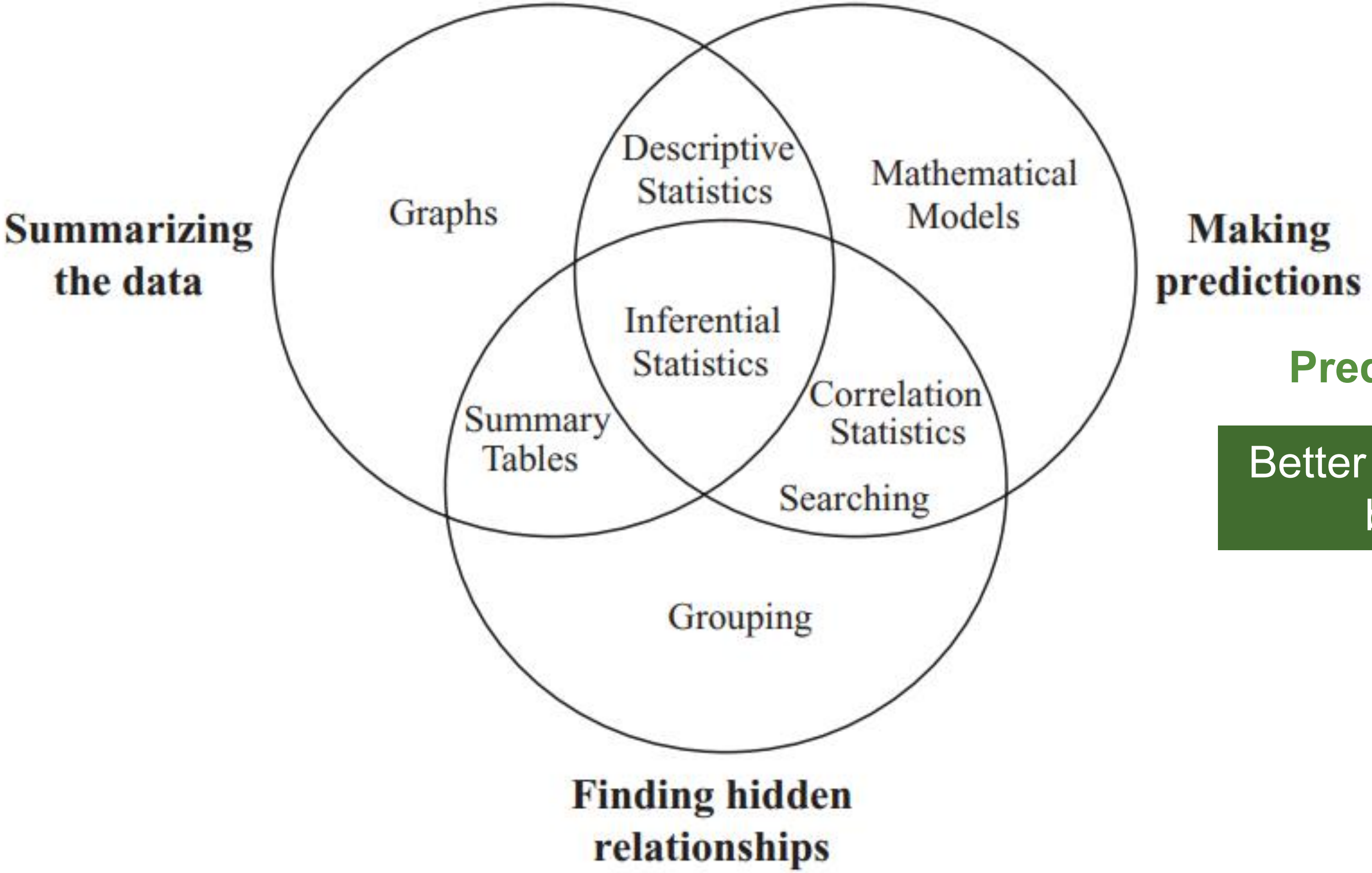


Figure 1.2. Data analysis tasks and methods

Predictive Models

Better equipped to arrive a better decision

LEARNING DECISION TREES FROM DATA.

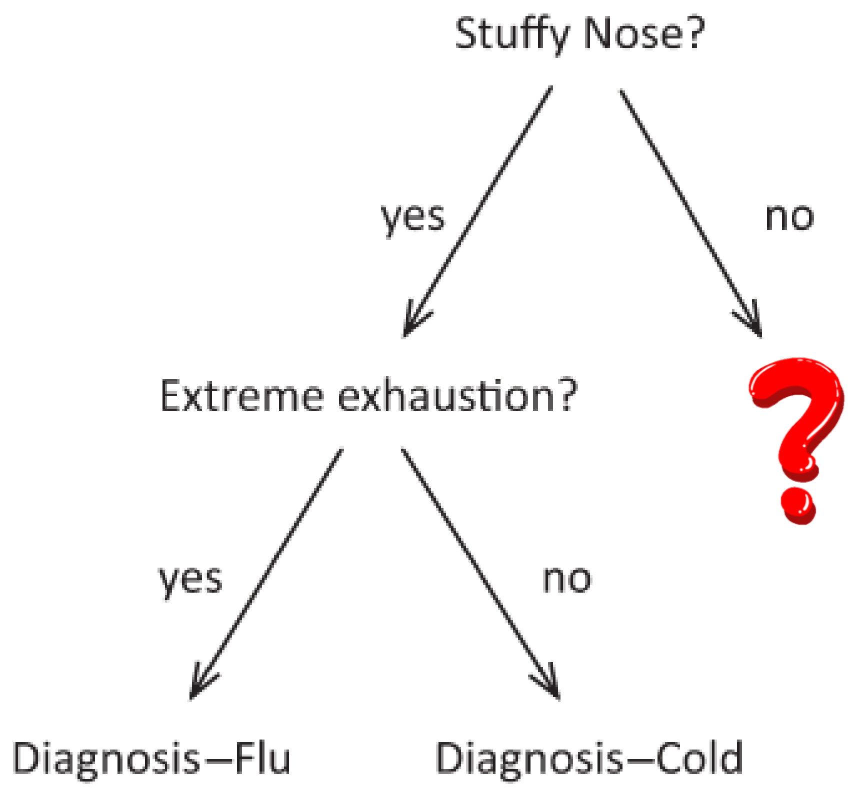
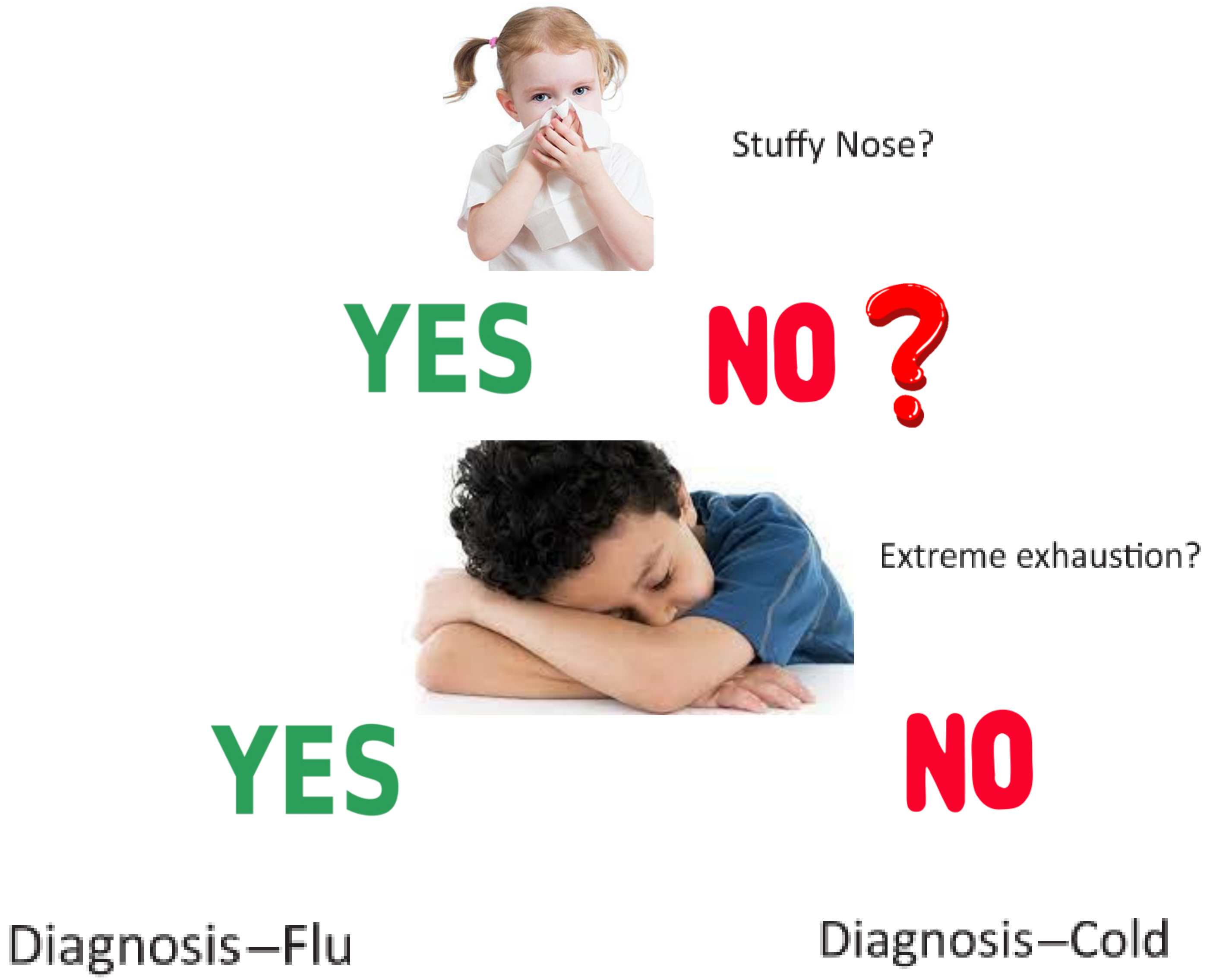
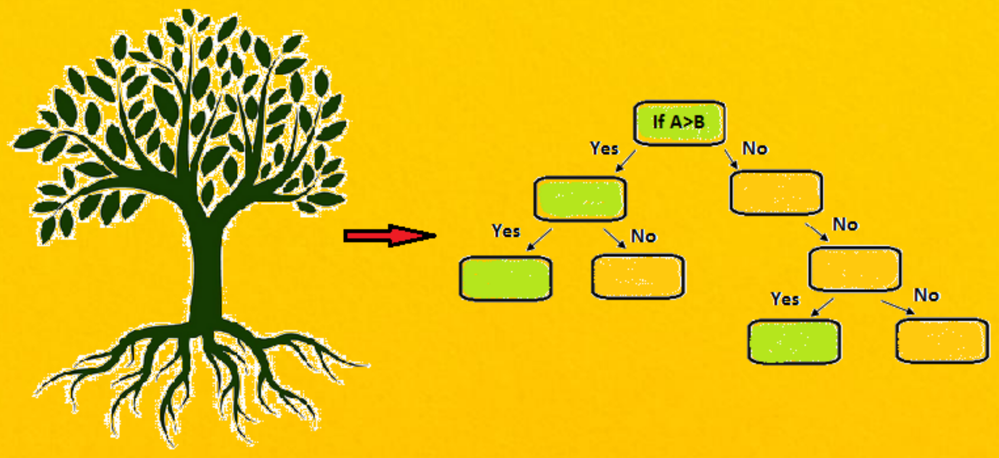


FIGURE 5.30
Decision tree for the diagnosis of colds and flu



It is often necessary to ask a series of questions before coming to a decision?

Answers to one question
may lead to
other questions
or
may lead to decision.



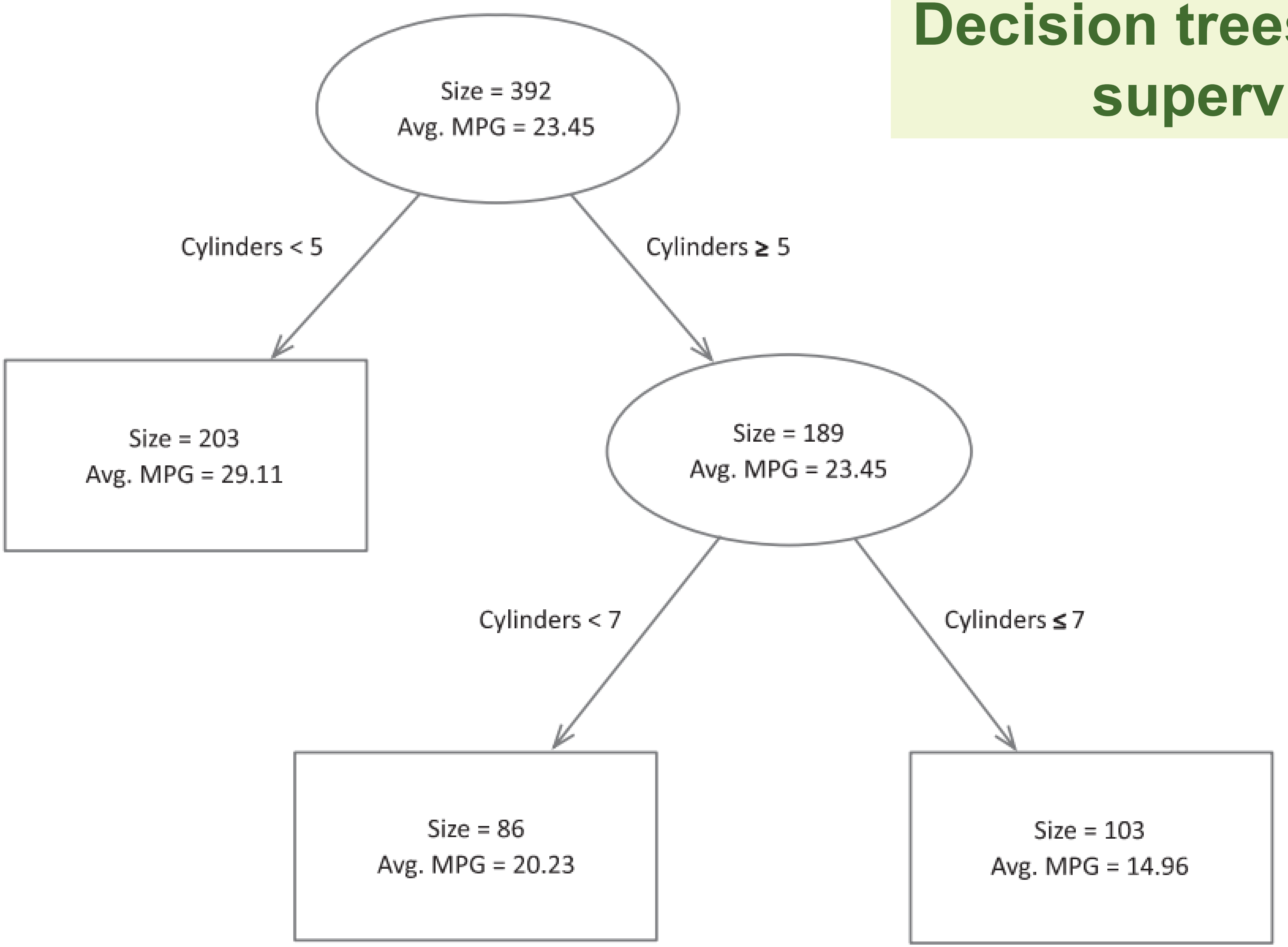


FIGURE 5.31 Decision tree generated from a data set of cars.

Decision trees are an example of a supervised approach.

Supervised methods
An attempt to place (classify) each observation into interesting groups based on a selected variable

These methods **iterate** over a **training set** of observations and adjust parameters as the classifier correctly or incorrectly classifies each observation.

Decision trees.

- Generated by hand to precisely and consistently define a decision-making process
  - Can also be generated automatically from the data.
- Consist of a series of decision points based on certain selected variables.
- Figure 5.31 illustrates a simple decision tree.

- This decision tree generated:
  - Based on a data set of cars that included variables for the number of cylinders (Cylinders) and the car's fuel efficiency (MPG).
  - Uses **number of cylinders (Cylinders)** to attempt to achieve the **goal of classifying the observations** according to their fuel efficiency.
- **Top of the tree** is a node representing **entire data set of 392 observations** (Size = 392).
- The **data set** is initially divided into **two subsets**:
  - Set of 203 cars (i.e., Size = 203) where the number of cylinders is fewer than 5 **(LEFT)**
  - Remaining observations where number of cylinders 5 or greater **(RIGHT)**

Data set was **classified** into groups using the variable MPG.

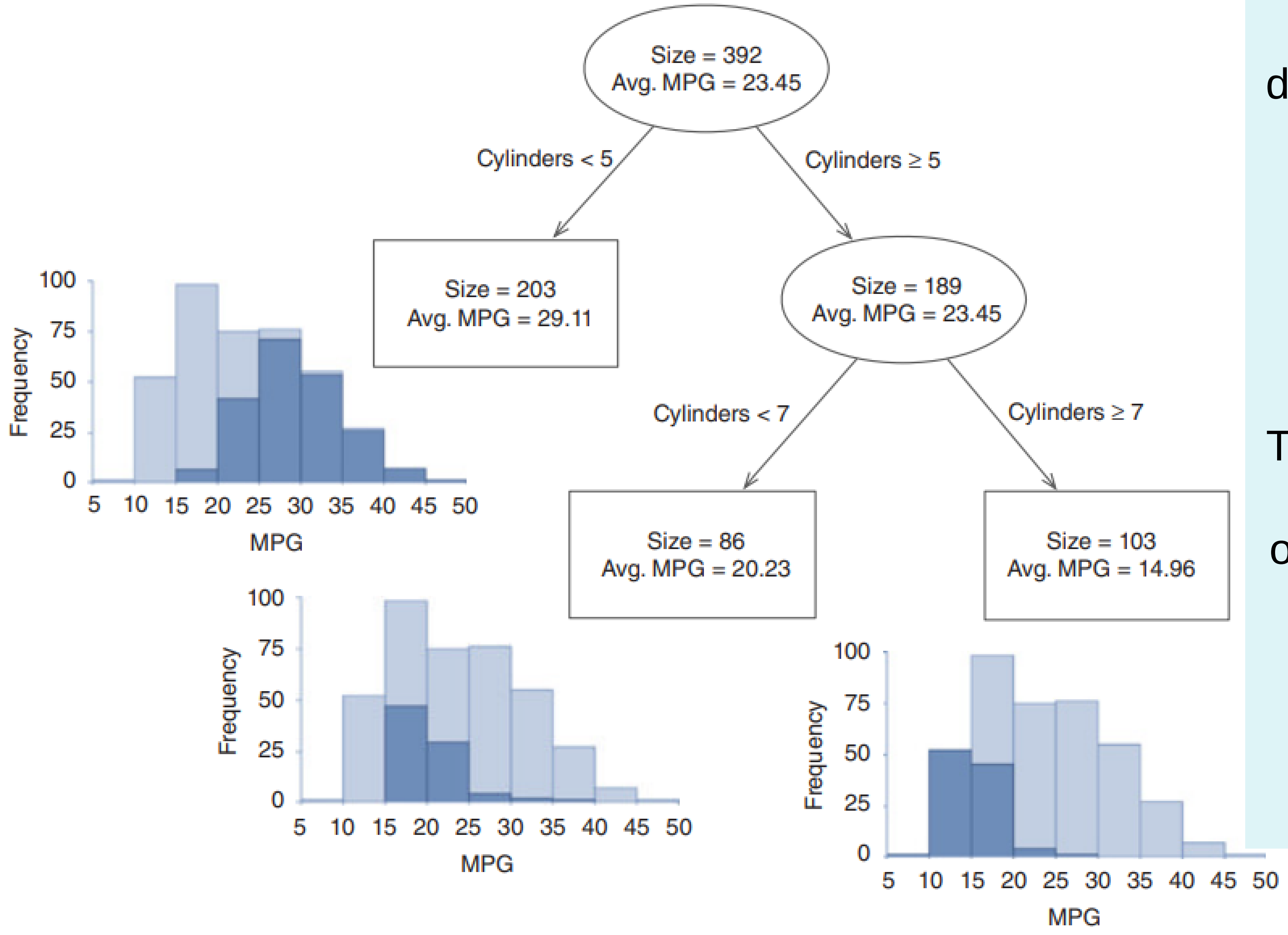


FIGURE 5.32 Decision tree illustrating the use of a response variable MPG to guide the tree generation.

The overall shape of the histograms depicts the frequency distribution for the MPG variable.

The highlighted frequency distribution is the subset within the node.

The frequency distribution for the node containing 203 observations shows a set biased toward good fuel efficiency, whereas for the node of 103 observations it illustrates a set biased toward poor fuel efficiency

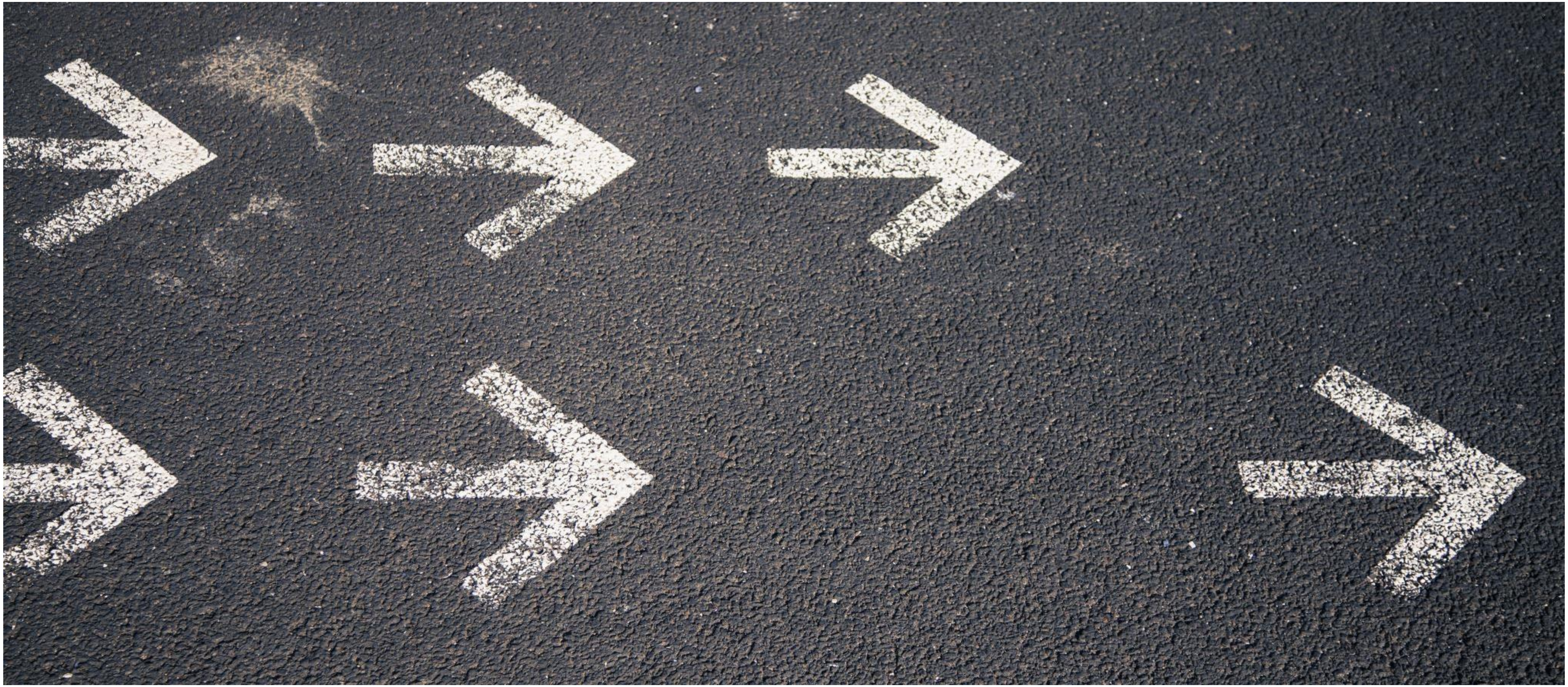
Why use Decision Trees?

1. **Easy to understand** and **use** in explaining how decisions are reached based on multiple criteria.
2. Can handle **categorical and continuous variables** since they partition a data set into distinct regions based on ranges or specific values.

Disadvantages:

1. **Building decision trees** can be **computationally expensive**, particularly when analyzing a **large data set** with many continuous variables
2. **Generating a useful decision tree** automatically can be **challenging**, since large and complex trees can be easily generated;
   1. Trees that are too small may not capture enough information; and
   2. Generating the "best" tree through optimization is difficult.

Decision Trees and **Decision points.**



BREAK FOR NOW.

Tree Generation and Splitting.

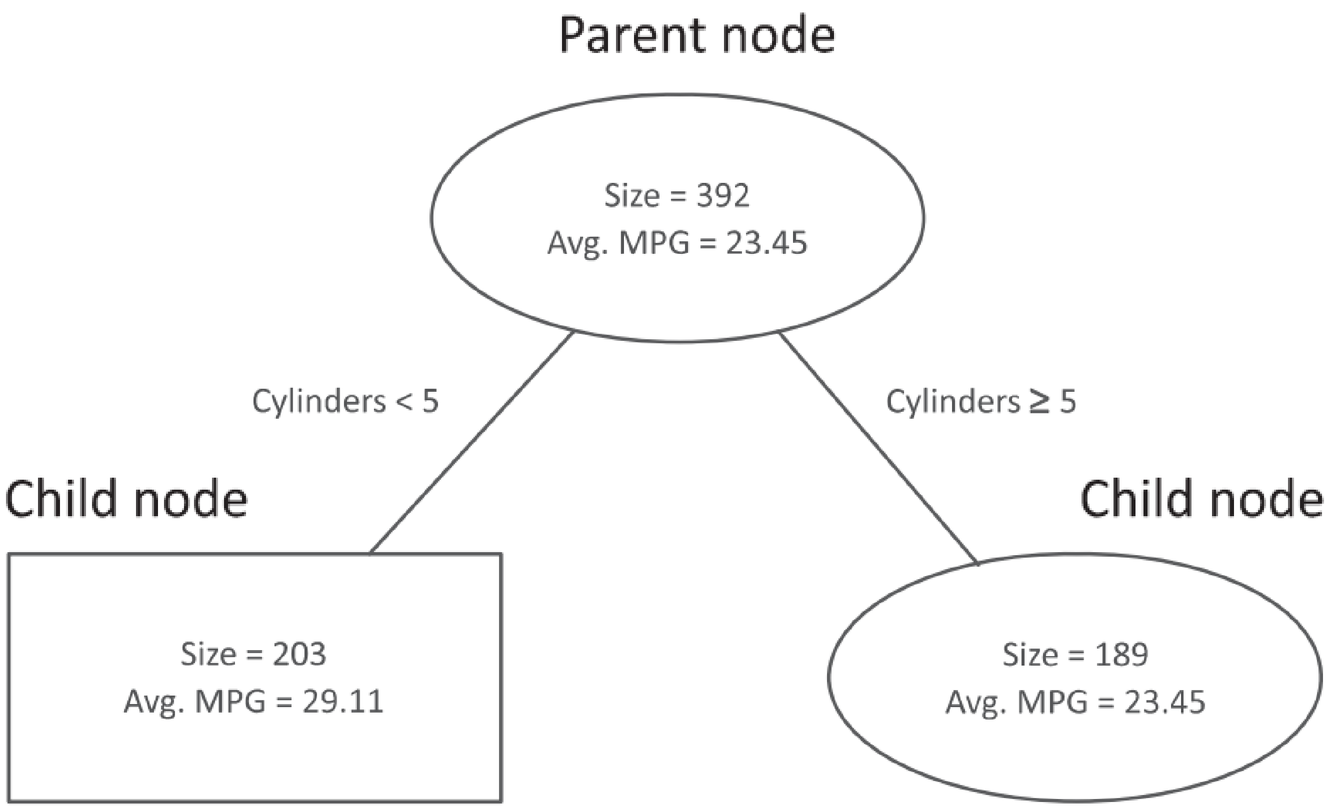
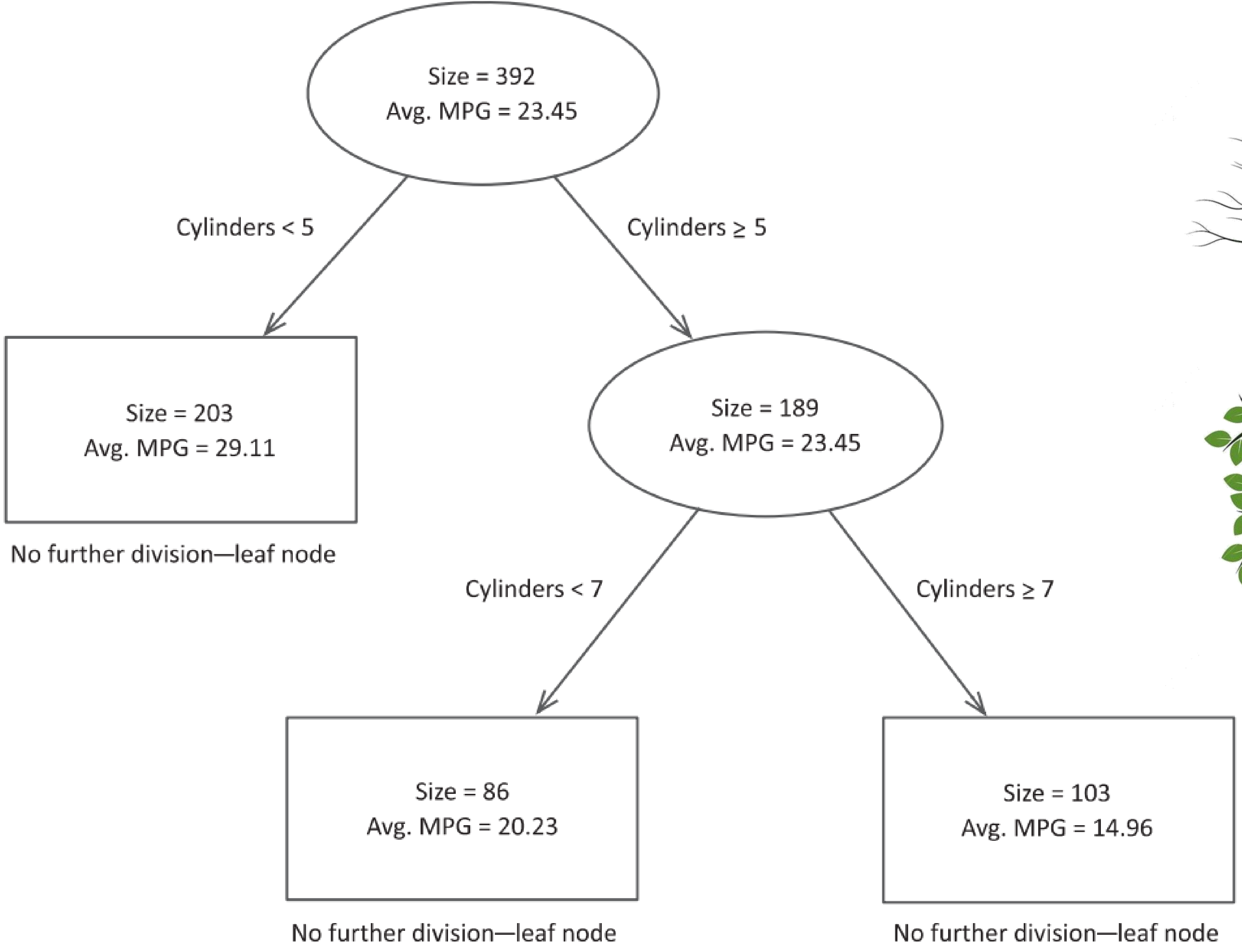
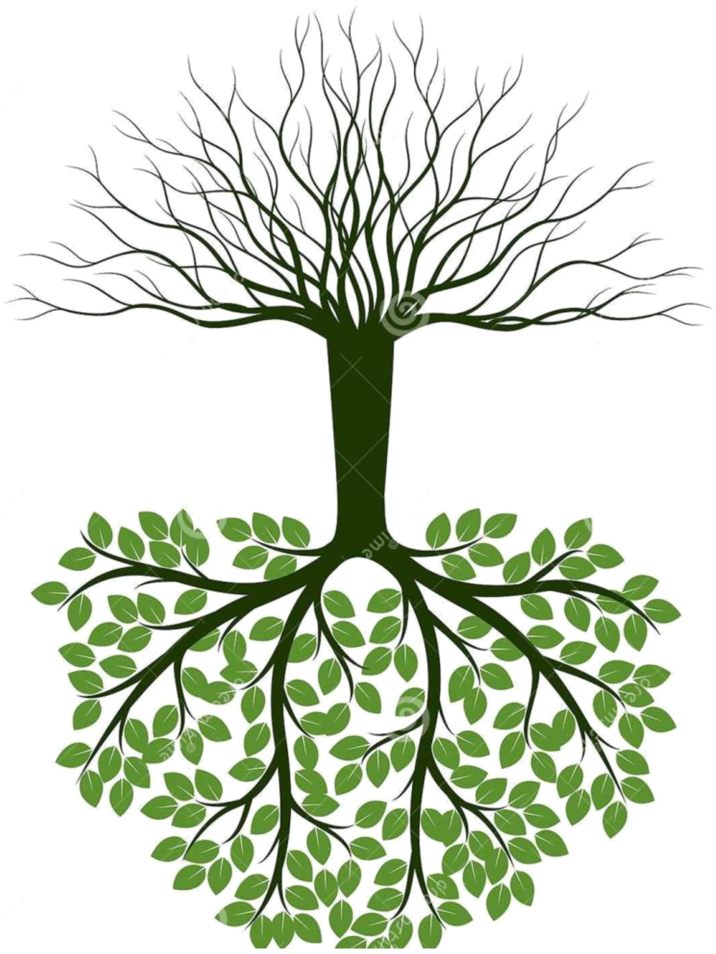


FIGURE 5.33 Relationship between parent and child nodes.

- A tree is **made up** of a series of **decision points**, where the **split** of the entire set of observations or a subset of the observations is based on some criteria.
- **Each point** in the **tree** represents **set of observations** called **node.**
- Relationship between two connected nodes is defined as a **parent–child relationship.**
- The larger set that will be divided into two or more smaller sets is the **parent node.**
- **Nodes** resulting from the **division of the parent** are **child nodes**.
- A **child node with no children** is a **leaf node.**





Inverted tree structure

FIGURE 5.34 Illustration of leaf nodes.

Splitting.

- A **table of data** is **used** to **generate a decision tree** where:
  - **Variables** are used as **potential decision points** (**splitting variables**),
  - **One variable** is used to **guide the construction** of tree **(response variable).**
    - Used to guide which splitting variables are selected and at what value the split is made.

- A decision tree **splits** the **data set** into **increasingly smaller, nonoverlapping subsets.**
- **Topmost node**, or **root** of tree, **contains all observations**.
- **Based on some criteria**, observations are usually **split into two new** nodes, where **each node** represents a **subset of observations.**

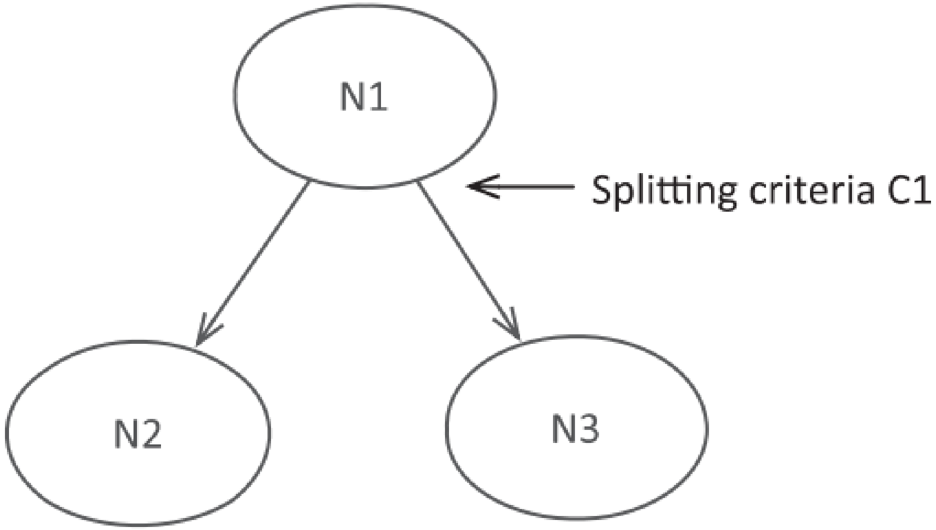


FIGURE 5.35 Node N1 split into two based on the criteria C1.

Splitting.

- **Process of examining variables** to **determine a criterion for splitting** is **repeated** for **all subsequent nodes**.
- Additionally, a **condition** is **needed** to **end** the process.
- For example: Process can stop when size of subset is less than a predetermined value.
- Figure 5.36, each of the two newly created subsets (N2 and N3) is examined in turn to determine if they should be further split or whether the splitting should stop.

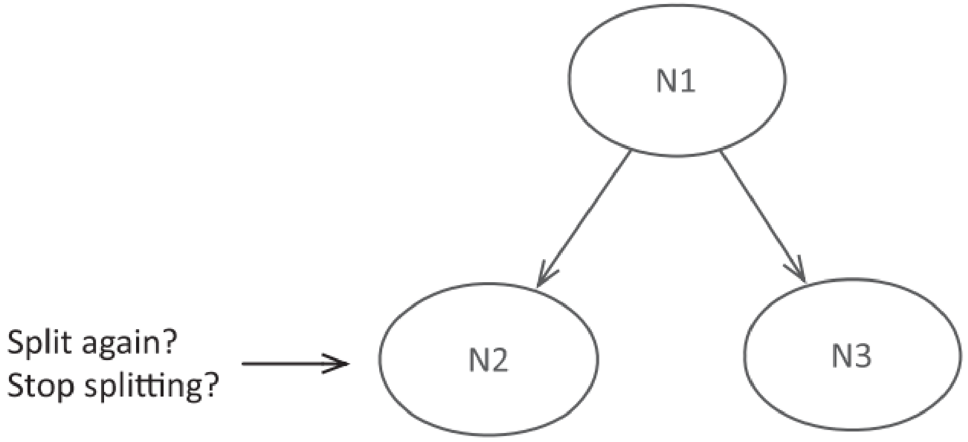


FIGURE 5.36 Evaluation of whether to continue to grow the tree

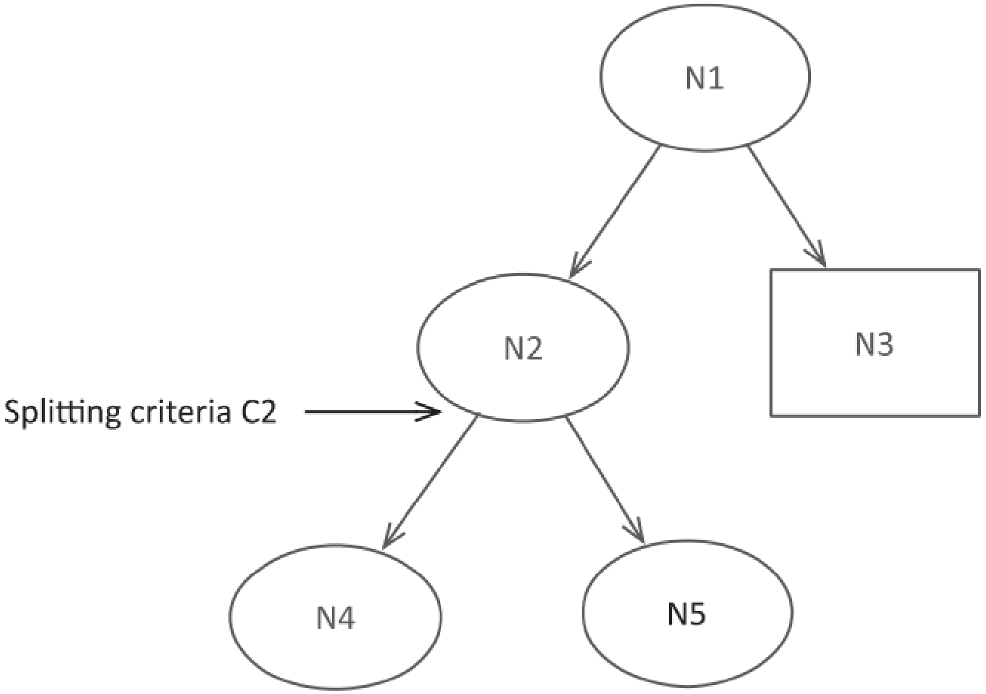


FIGURE 5.37 Tree further divided.

Splitting Criteria: Dividing Observations.

- It is **common** for the **split at each level** to be a **two-way split.**
- To **split more than two ways**, care should be taken when using these methods because making **too many splits early** in the construction of the tree may **result** in **missing interesting relationships** that become exposed as tree construction continues: How?
- *Possible because of dividing set into small groups based on single criterion.*

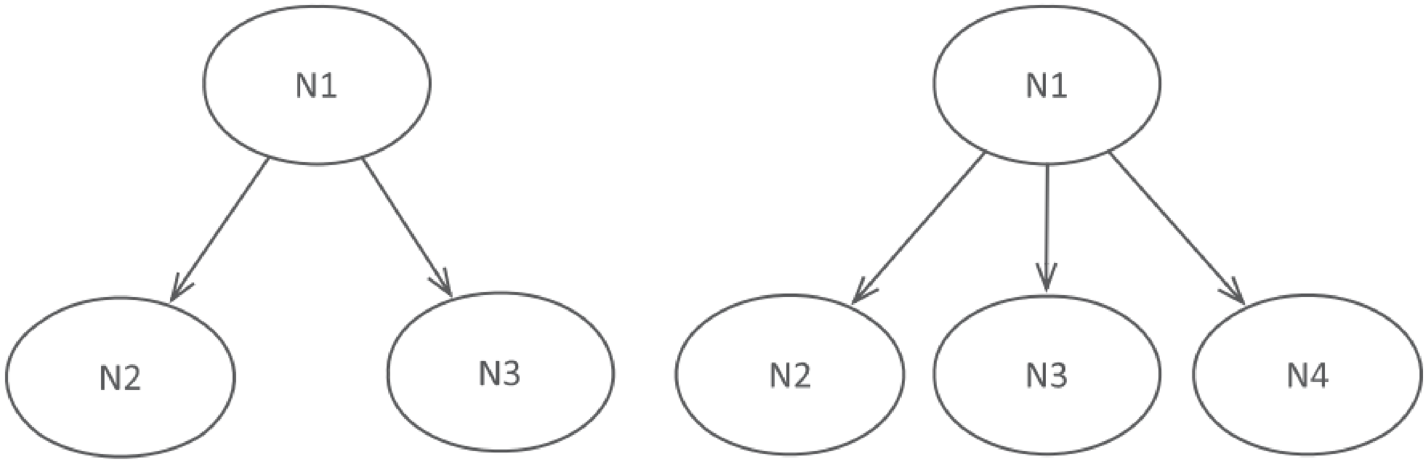


FIGURE 5.38 Alternative splitting of nodes.

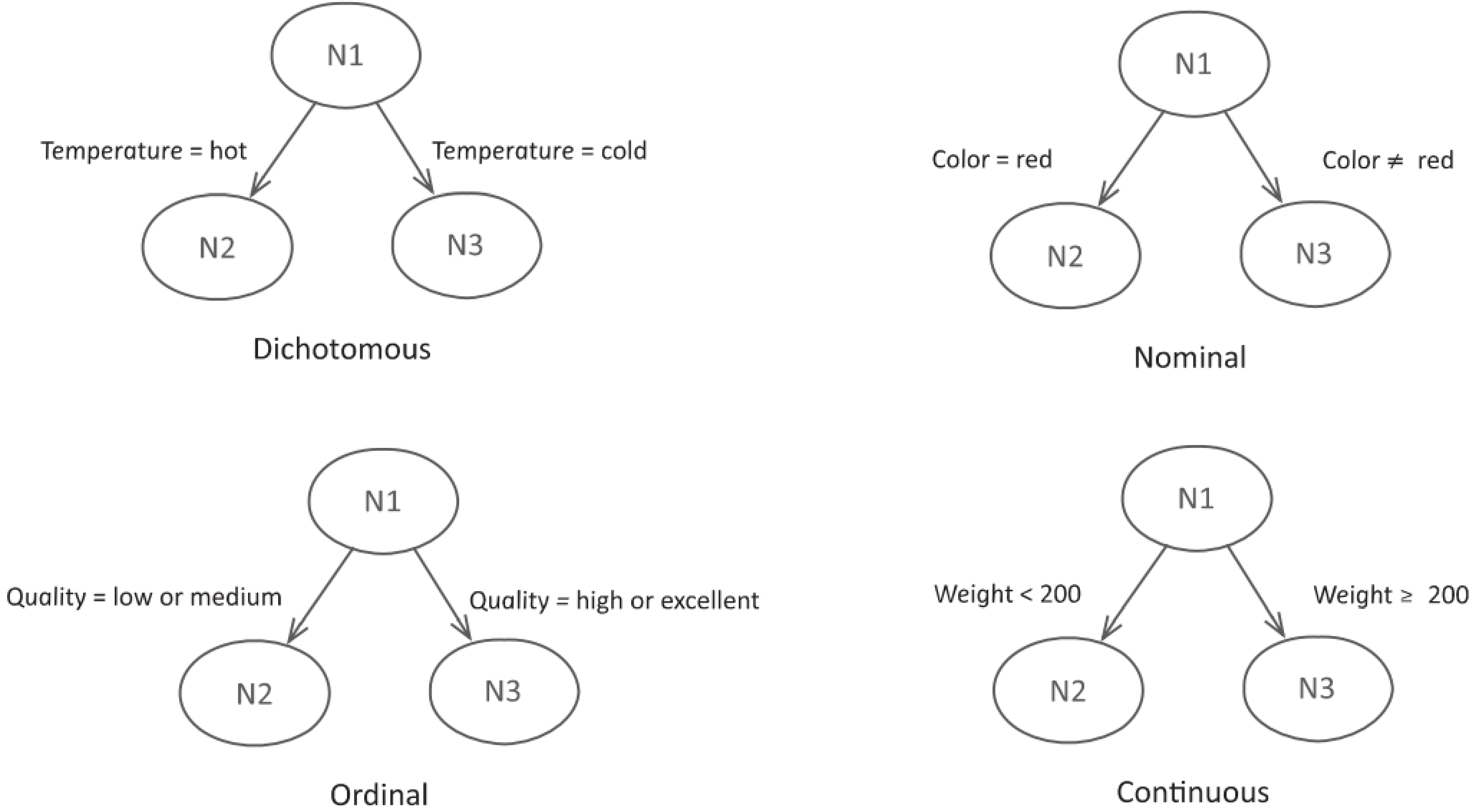


FIGURE 5.39 Splitting examples based on variable type.

Splitting Criteria: Dividing Observations.

Dichotomous	Nominal	Ordinal	Continuous
Temperature may have only two values: **“hot” and “cold**	**Color** can take the values **“red,” “green,” “blue,” and “black”** may be split **two-ways.**	**Quality** with possible values **“low,” “medium,” “high,” and “excellent”** may be split **four ways.**	**Weight** which can take any value **between 0 and 1,000** with a selected **cut-off of 200**

Splitting Criteria: Dividing Observations.

Dichotomous: Variables with two values are the most straightforward to split since each branch represents a specific value.

Nominal: Values are **discrete values with no order**, two-way split is accomplished by one subset being composed of a set of observations that equal a certain value and the other being those observations that do not equal that value.

Ordinal: Variable's **discrete values are ordered**, the resulting subsets may be made up of more than one value, as long as the **ordering is retained.**

Continuous: Values can be **split two ways**. A **specific cut-off value** needs to be determined for observations with **values less than cut-off in left subset** and those with **values greater than or equal to are in right subset**

The **Splitting** Criterion.

1

Variable on which to split

2

Values of that variable to use for the split

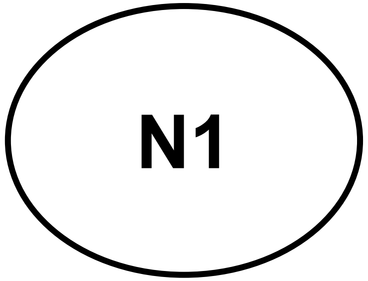
To determine **best split,** a **ranking is made** of all possible splits of all variables using a **score calculated** for each split.

Scoring Splits for Categorical Response Variables

Before split

Temperature:
1 Hot (10 observations)
2 Cold (10 observations)

Even distribution



20 observations

Different criteria are considered for **splitting** these observations

Split a: **Each subset** contains **10 observations**. All 10 observations in N2 have "hot" temperature values and all 10 observations in node N3 are "cold."

Split b: **Each subset (N2 and N3)** contain **10 observations**. However, in this case there is an even distribution of "hot" and "cold" values in each subset.

Split c: Splitting criterion results in two subsets where node **N2** has **9 observations** (1 "hot" and 8 "cold") and node **N3** has **11 observations** (9 "hot" and 2 "cold")

Objective for an **optimal split** is to **create subsets** which **result in observations** with a **single response value**

Scoring Splits for Categorical Response Variables

A **decision tree** simply **partitions** the training data set into **disjoint subsets** so that each **subset** is as **pure** as possible.

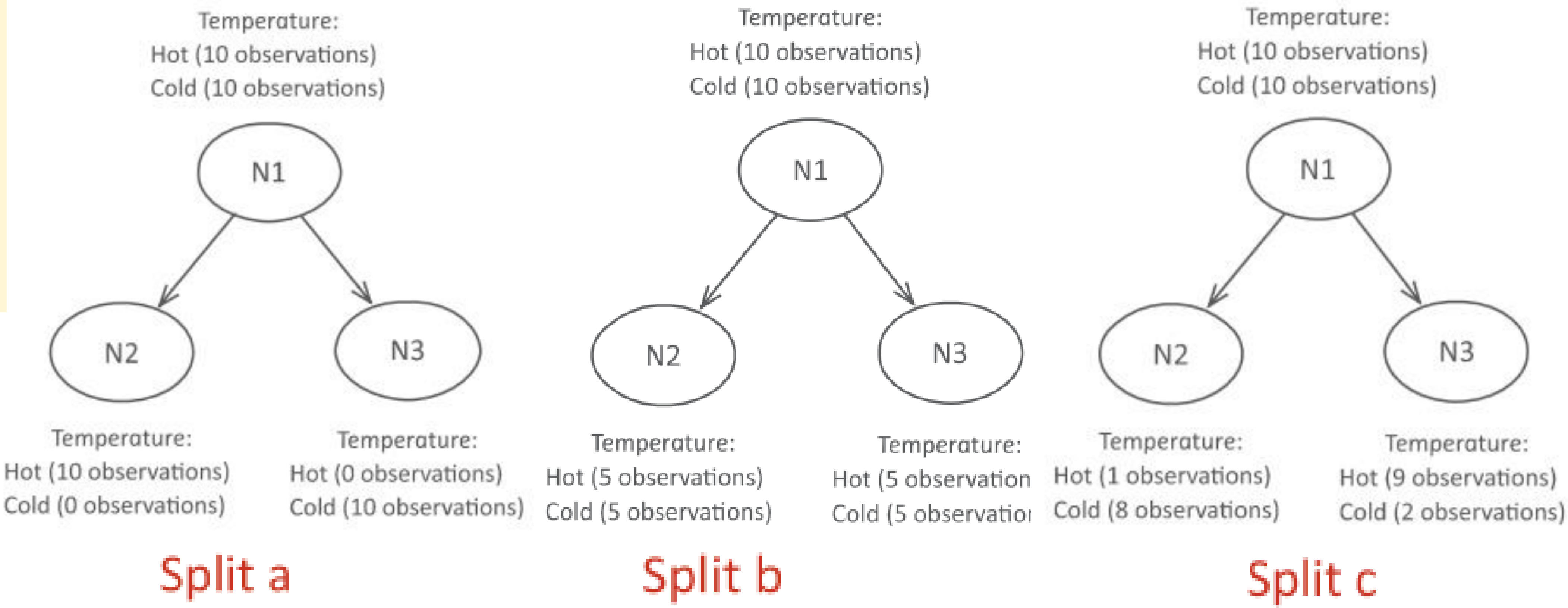
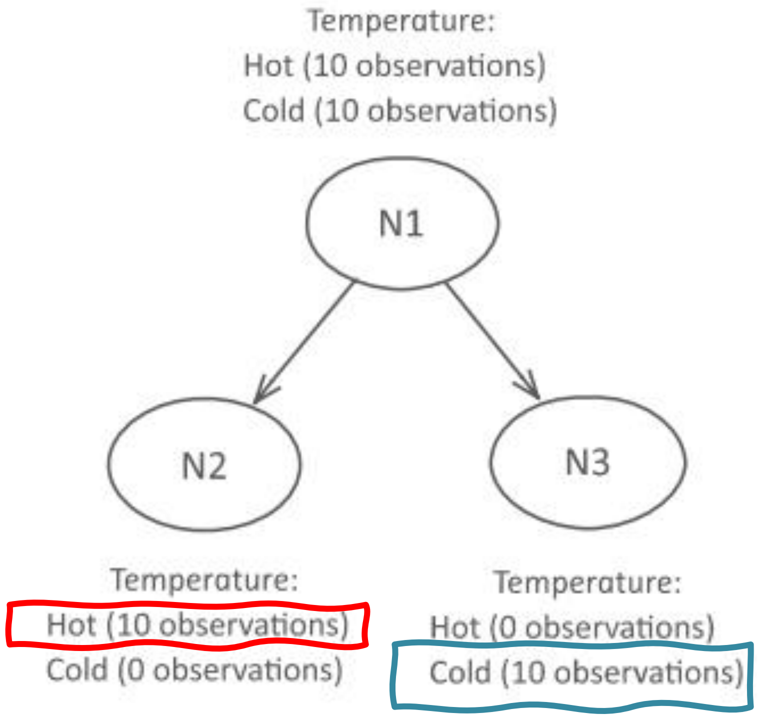


FIGURE 5.40 Evaluating splits based on categorical response data. *(Temperature)*

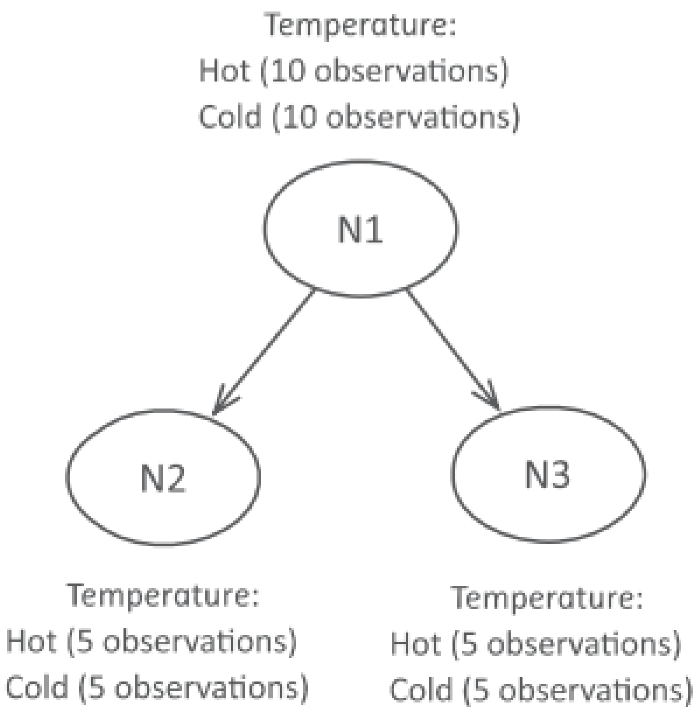
Best split?

Split a



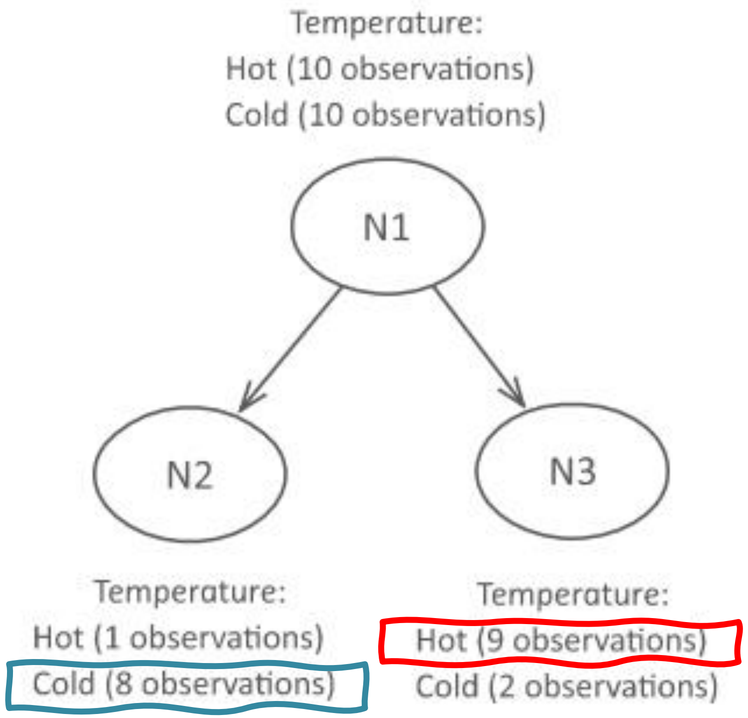
Best split
since each node contains observations where response for each node is of same category.

Split b



Not a good split.
Results in the **same even split** of "hot" and "cold" values (50% "hot," 50% "cold") in each of the resulting nodes (N2 and N3)

Split c



A good split
Not as clean as Split a
since both subsets have a mixture of "hot" and "cold" values. Proportion of "hot" and "cold" values in node N2 is biased toward cold values and in node N3 toward hot values.

“goodness” of splitting criteria

How clean each split is.

Based on the **proportion** of **different categories** of **response variable**

which is a measurement known as **impurity.**

As **tree** is being **generated**

desirable to

decrease level of impurity until

ideally there is **only one category**

at a

terminal node

The “goodness” of splitting criteria

How clean each split is.

This is based on the **proportion** of **different categories** of **response variable**

Proportion is a **measurement** known as **impurity.**

As **tree** is being **generated**

desirable to

decrease level of impurity

until

ideally there is **only one category**

at a

terminal node

Calculating **impurity**

Mis-classification

Gini

Entropy

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Measure of Purity/Disorder

$$\text{Gain} = \text{Entropy}(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$

TABLE 5.16 Entropy Scores According to Different Splitting Criteria

	Response Values		
Scenario	Hot	Cold	Entropy
Scenario 1	0	10	0
Scenario 2	1	9	0.469
Scenario 3	2	8	0.722
Scenario 4	3	7	0.881
Scenario 5	4	6	0.971
Scenario 6	5	5	1
Scenario 7	6	4	0.971
Scenario 8	7	3	0.881
Scenario 9	8	2	0.722
Scenario 10	9	1	0.469
Scenario 11	10	0	0

Cleaner splits result in lower scores

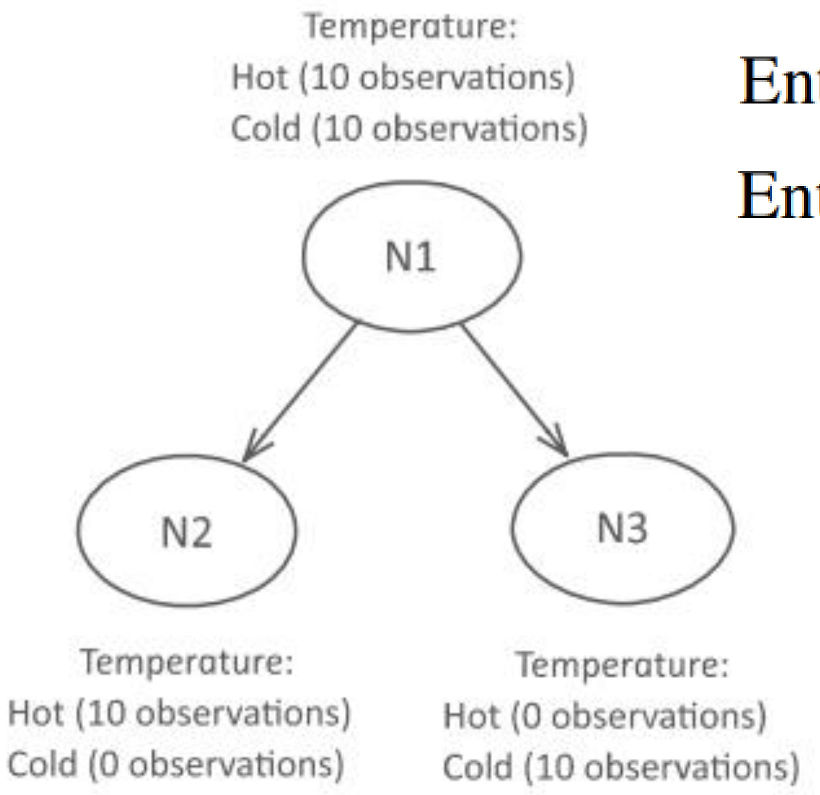
Entropy

$$\text{Entropy}(S) = -\sum^{c} p_i \log_2 p_i$$

Where,
S: set of observations
p$_i$: fraction of the observations that belong to a particular value
c: number of different possible values of response variable.

Set of **100 observations** where **temperature response variable** had **60** observations with "hot" values and 40 with "cold" values, $\mathbf{p}_{\mathbf{hot}} = 0.6$ and $\mathbf{p}_{\mathbf{cold}} = 0.4$.

Values for entropy calculated for each split: Split a



$$\text{Entropy (N1)} = -(10/20)\log_2(10/20) - (10/20)\log_2(10/20) = 1$$

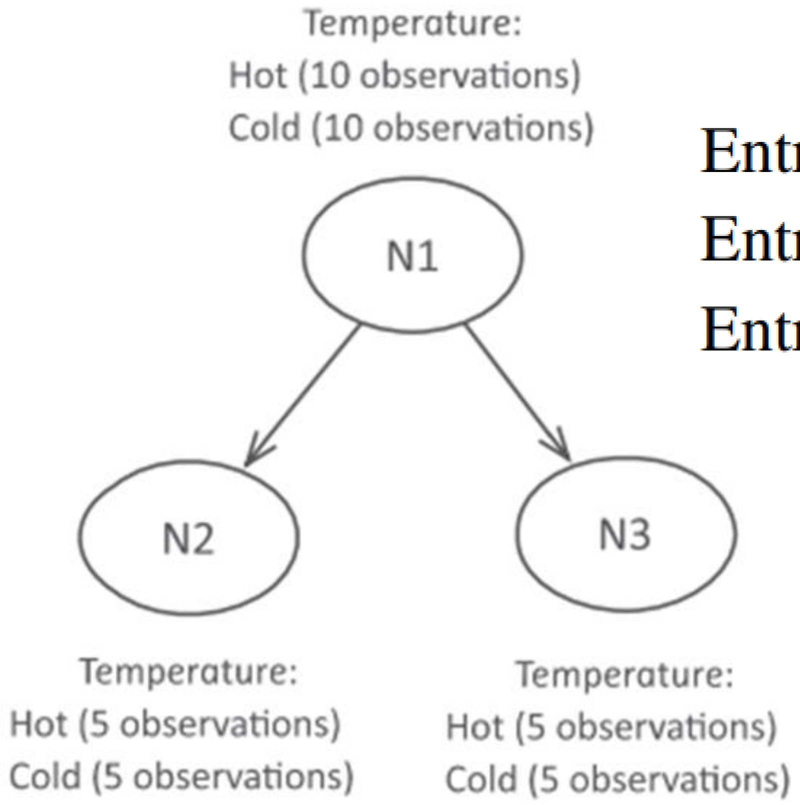
$$\text{Entropy (N2)} = -(10/10)\log_2(10/10) - (0/10)\log_2(0/10) = 0$$

$$\text{Entropy (N3)} = -(0/10)\log_2(0/10) - (10/10)\log_2(10/10) = 0$$

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

When pi = 0, then value for 0 log2 (0) = 0

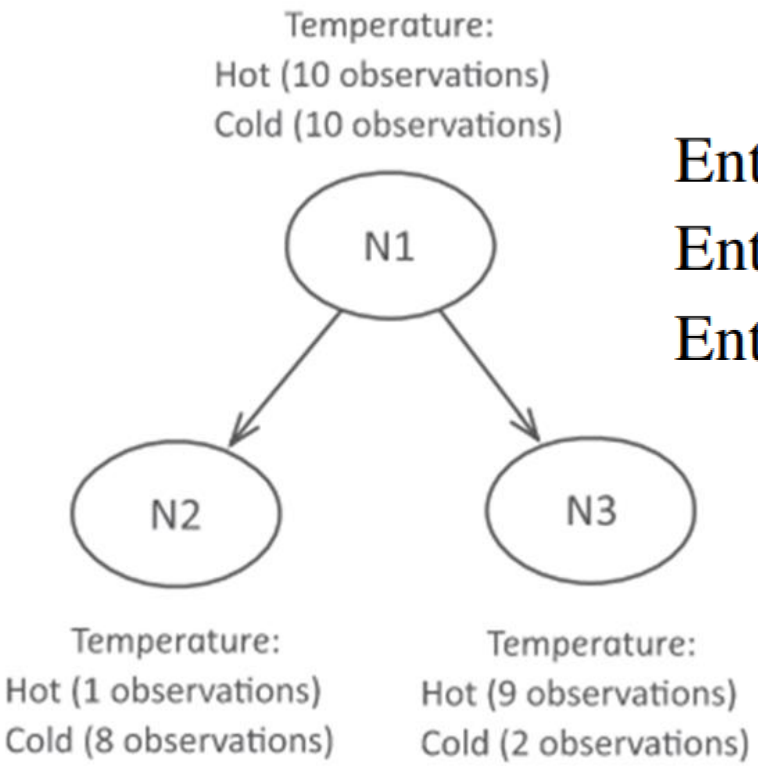
Values for entropy calculated for each split: Split b



$$\text{Entropy (N1)} = -(10/20)\log_2(10/20) - (10/20)\log_2(10/20) = 1$$
$$\text{Entropy (N2)} = -(5/10)\log_2(5/10) - (5/10)\log_2(5/10) = 1$$
$$\text{Entropy (N3)} = -(5/10)\log_2(5/10) - (5/10)\log_2(5/10) = 1$$

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Values for entropy calculated for each split: Split c



$$\text{Entropy (N1)} = -(10/20)\log_2(10/20) - (10/20)\log_2(10/20) = 1$$
$$\text{Entropy (N2)} = -(1/9)\log_2(1/9) - (8/9)\log_2(8/9) = 0.503$$
$$\text{Entropy (N3)} = -(9/11)\log_2(9/11) - (2/11)\log_2(2/11) = 0.684$$

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Which is **best split?**

- In order to **determine the best split**, **calculate a ranking based** on **how cleanly each split separates the response data.**
- This is calculated based on the **impurity before and after the split.**

Gain

$$\text{Gain} = \text{Entropy}(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$

Entropy

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Discussing **metrics** used to **train decision trees**

One of them is **information gain**

Information Gain

$$\text{Gain} = \text{Entropy(parent)} - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$

Where,

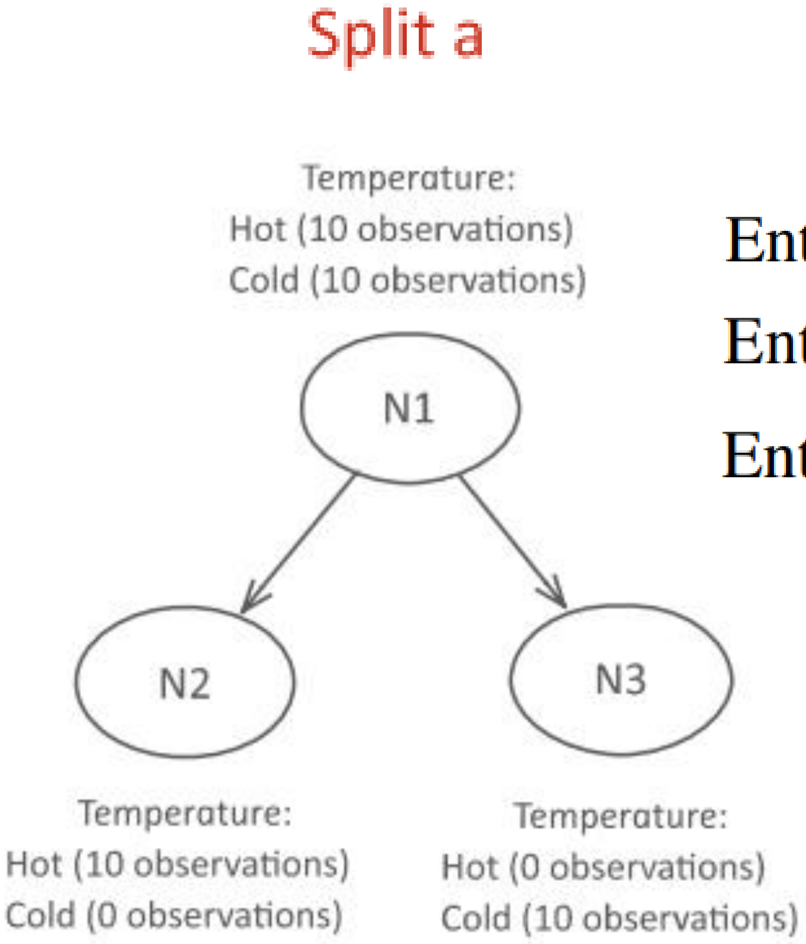
N: Number of observations in the **parent** node,

k: Number of **possible resulting nodes**

N(v_j): Number of observations for each of the j child nodes

V_j: Set of observations for the j^{th} node.

$$\text{Gain} = \text{Entropy}(\text{parent}) - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$



$$\text{Entropy (N1)} = -(10/20)\log_2(10/20) - (10/20)\log_2(10/20) = 1$$

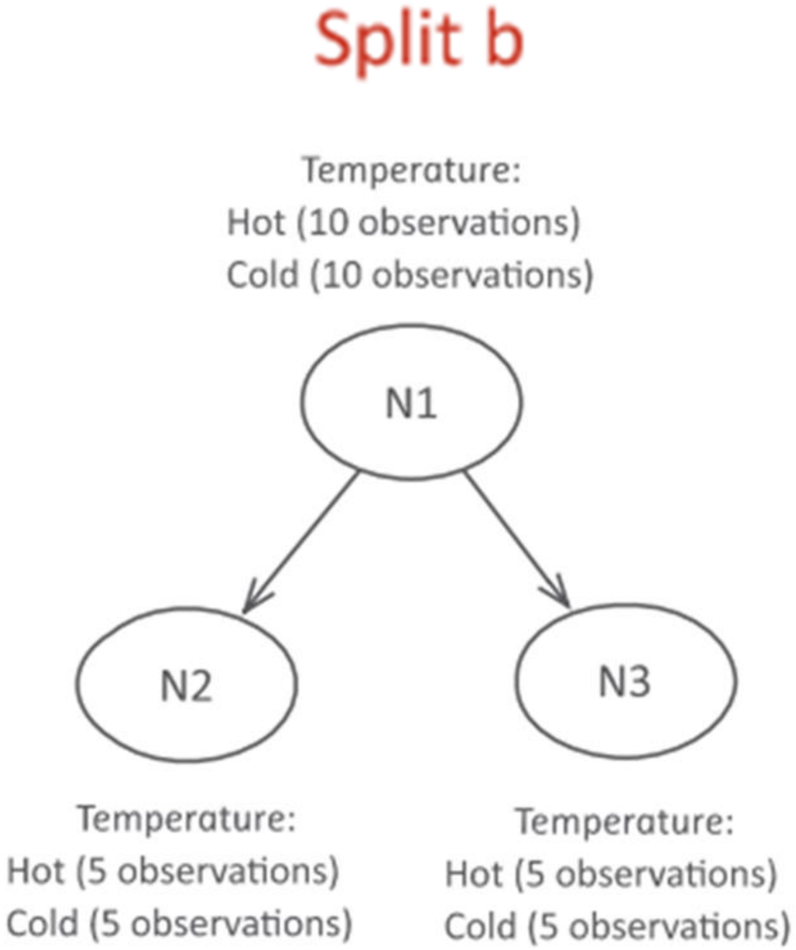
$$\text{Entropy (N2)} = -(10/10)\log_2(10/10) - (0/10)\log_2(0/10) = 0$$

$$\text{Entropy (N3)} = -(0/10)\log_2(0/10) - (10/10)\log_2(10/10) = 0$$

$$\text{Gain(Splita)} = 1 - (((10/20)\,0) + ((10/20)\,0)) = 1$$

FIGURE 5.41 Calculation of gain for each split.

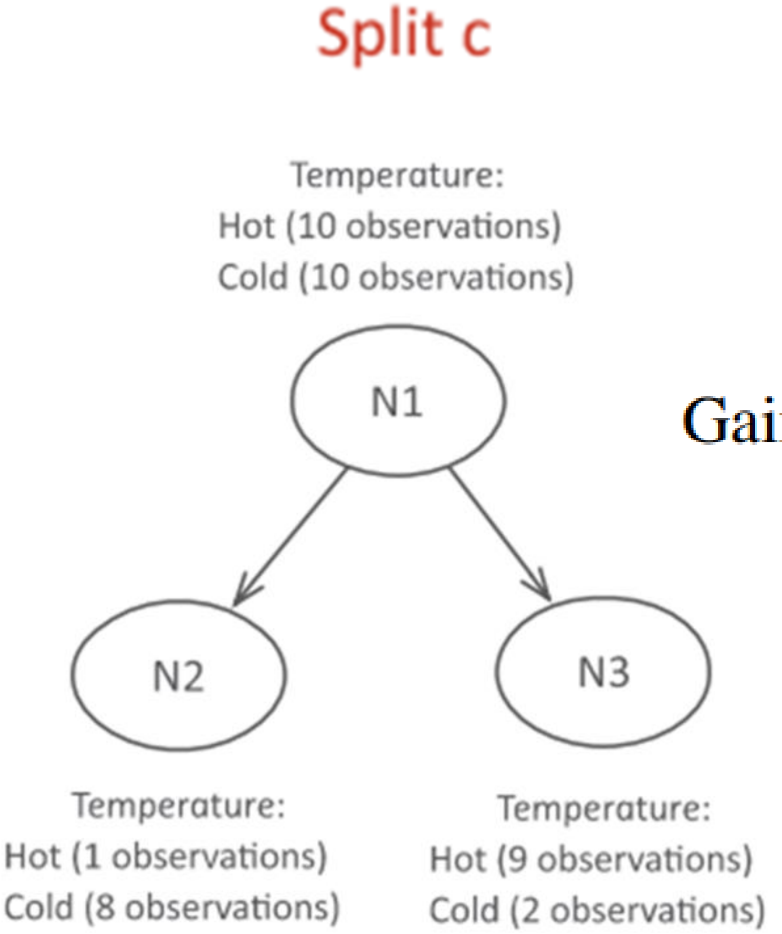
$$\text{Gain} = \text{Entropy(parent)} - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$



$$\text{Gain(Splitb)} = 1 - (((10/20)\ 1) + ((10/20)\ 1)) = 0$$

FIGURE 5.41 Calculation of gain for each split.

$$\text{Gain} = \text{Entropy(parent)} - \sum_{j=1}^{k} \frac{N(v_j)}{N} \text{Entropy}(v_j)$$



$$\text{Gain(Split c)} = 1 - (((9/20)\,0.503) + ((11/20)\,0.684)) = 0.397$$

FIGURE 5.41 Calculation of gain for each split.

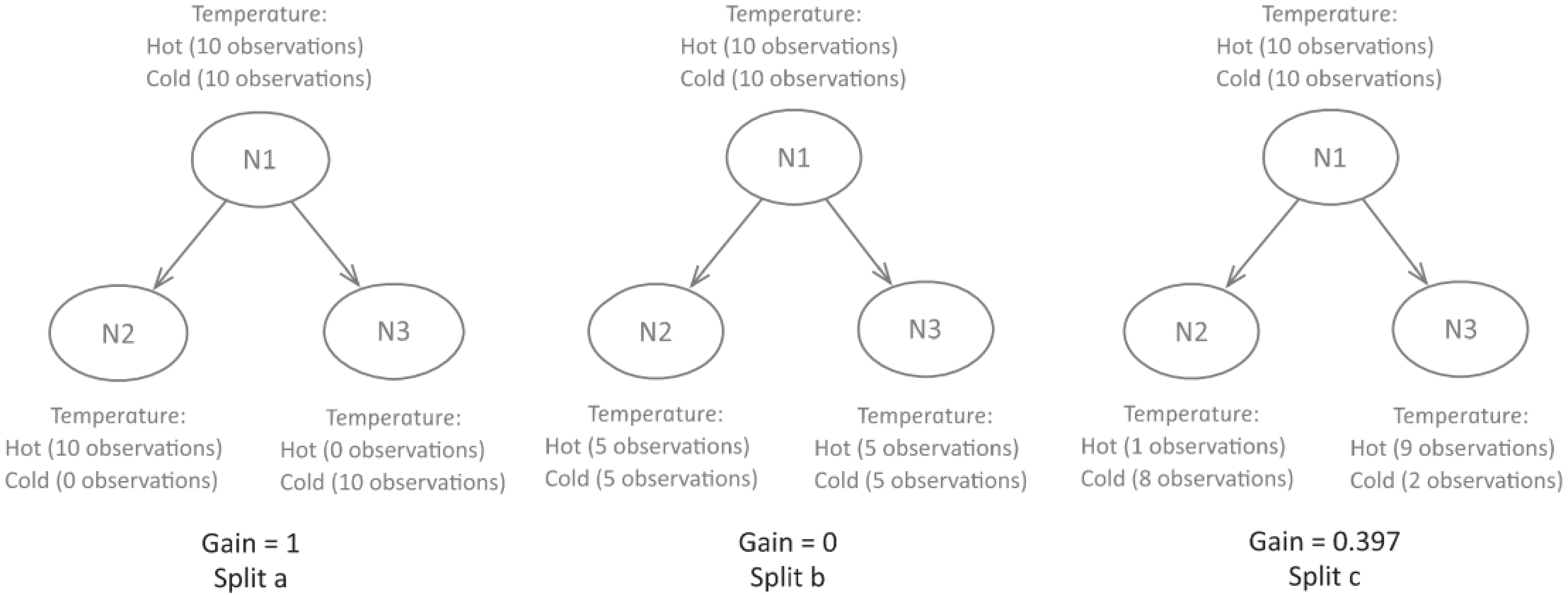
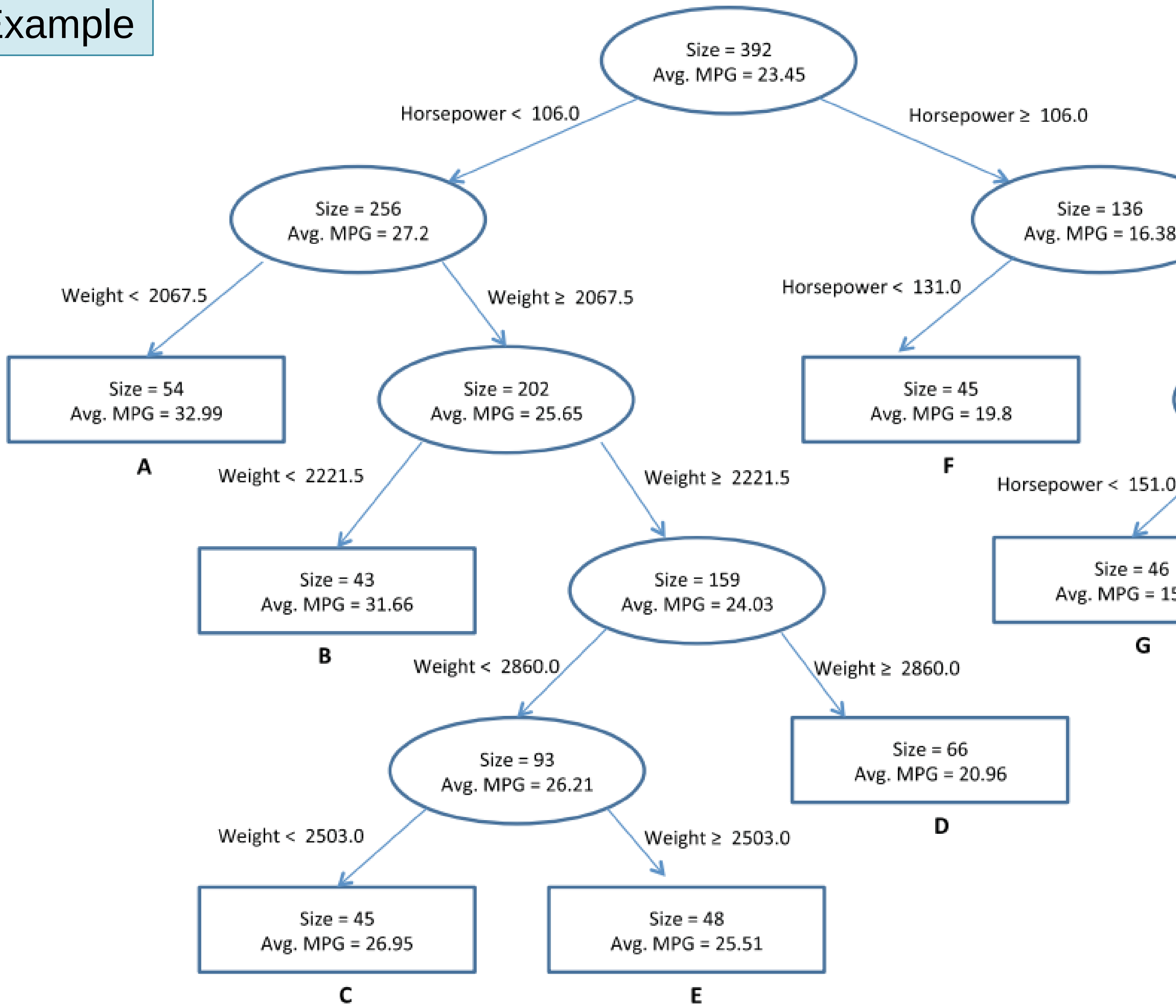


FIGURE 5.41 Calculation of gain for each split.

The criterion used in **Split a** is selected as the best splitting criteria.

Example



Example: A set of 392 cars is analyzed using a decision tree.

Two variables were used to split nodes in the tree: **Horsepower**, and **Weight**. **MPG** (miles per gallon) is response variable guiding generation of the tree.

Leaf nodes of the tree can be interpreted using a series of rules

Decision points that are traversed in getting to the node are the rule conditions.

Average MPG value for the leaf nodes interpreted here as:
low (less than 22)
medium (22–26)
high (greater than 26).

FIGURE 5.43 Decision tree generated using Horsepower and Weight as splitting values and guided by MPG.

Example

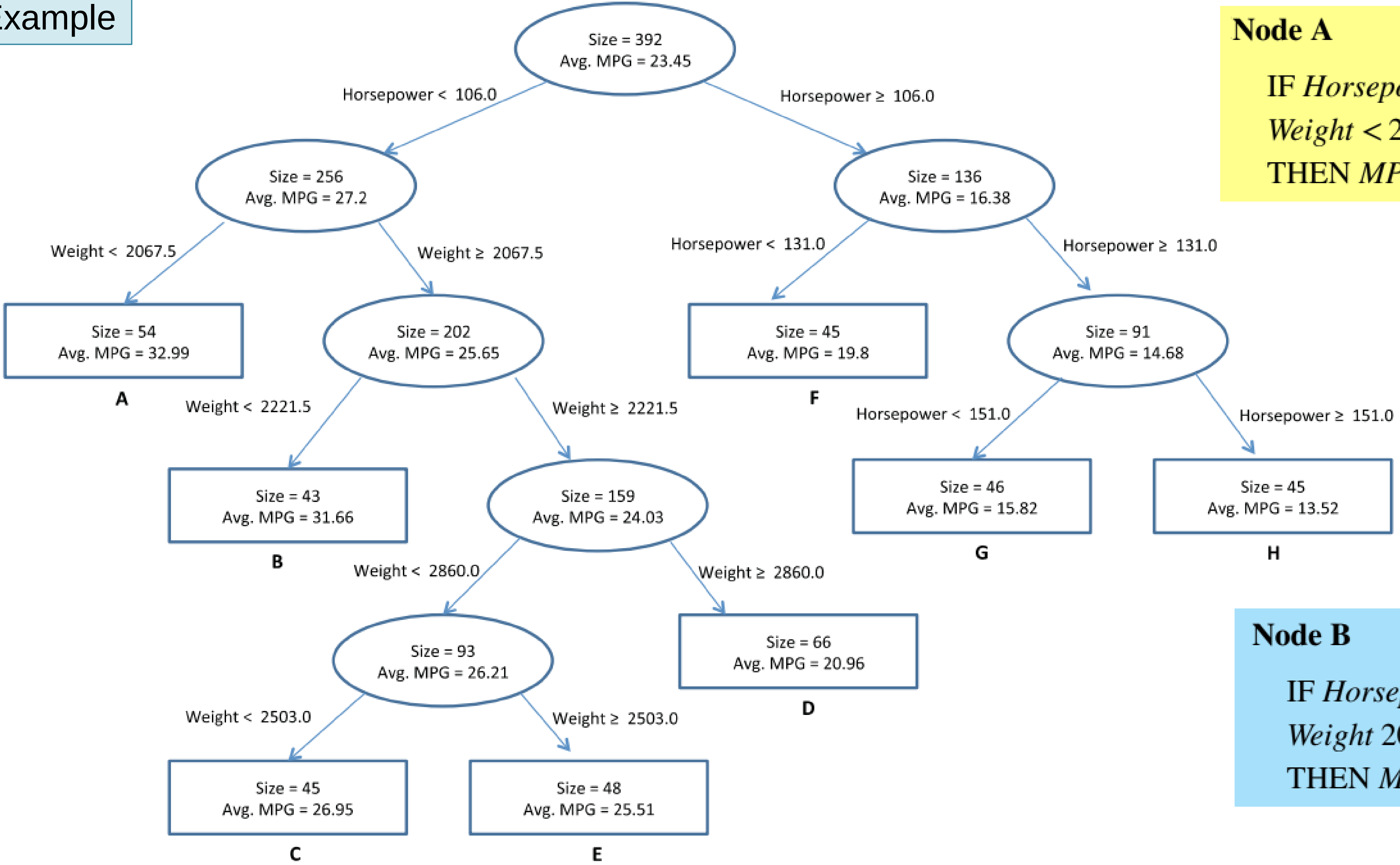


FIGURE 5.43 Decision tree generated using Horsepower and Weight as splitting values and guided by MPG.

Node A

IF *Horsepower* <106 AND
Weight < 2067.5
THEN *MPG* is high

Node B

IF *Horsepower* <106 AND
Weight 2067.5 – 2221.5
THEN *MPG* is high

Scoring Splits for **Continuous Response** Variables

TABLE 5.17 Table of Eight Observations with Values for Two Variables

Observations	Weight	MPG
A	1,835	26
B	1,773	31
C	1,613	35
D	1,834	27
E	4,615	10
F	4,732	9
G	4,955	12
H	4,741	13

The formula for *SSE* is

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

For subset of **n** observations, **SSE value** is computed where,
y_i: Individual value for response variable
$\bar{y}$: Average value for the subset

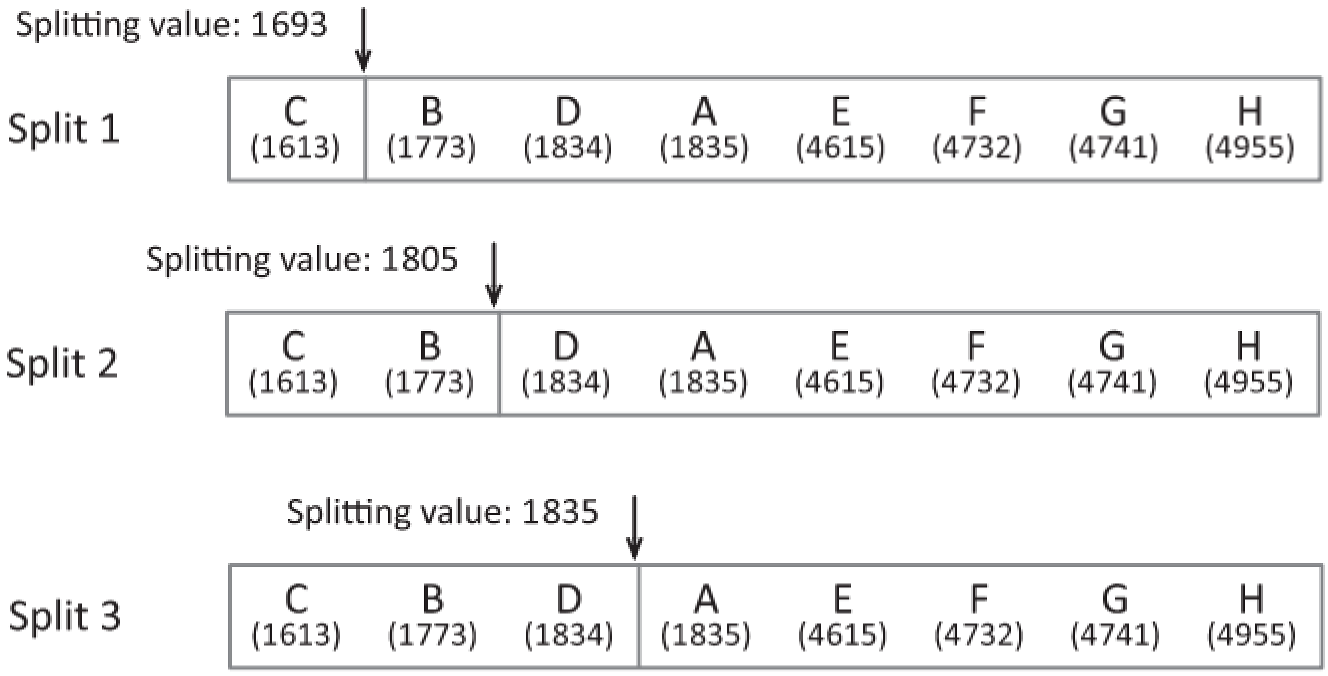
- Use **sum of squares of error (SSE)**
- Resulting split has sets where **response values** are **close** to **mean of group.**
- **Lower a group's SSE** value is**, closer that group's values to mean** of the set.
- For each potential split, a ***SSE*** value is calculated for each resulting node.
- A **score for the split** is calculated by **summing the SSE values** of each resulting node.
- Once all splits for all variables are computed, then split with the **lowest score is selected**

Determine **best split.**

Weight: Splitting variable
MPG: Response variable

TABLE 5.17 Table of Eight Observations with Values for Two Variables

Observations	Weight	MPG
A	1,835	26
B	1,773	31
C	1,613	35
D	1,834	27
E	4,615	10
F	4,732	9
G	4,955	12
H	4,741	13



A **series of values** is used to split the **variable Weight**: 1,693, 1,805, 1,835, 3,225, 4,674, 4,737, 4,955.

These **values** are **midpoint between each pair of values** (**after sorting**) and were selected because they **divided** the data set into **all possible two-ways splits.**

Determine **best split.**

TABLE 5.17 Table of Eight Observations with Values for Two Variables

Observations	Weight	MPG
A	1,835	26
B	1,773	31
C	1,613	35
D	1,834	27
E	4,615	10
F	4,732	9
G	4,955	12
H	4,741	13

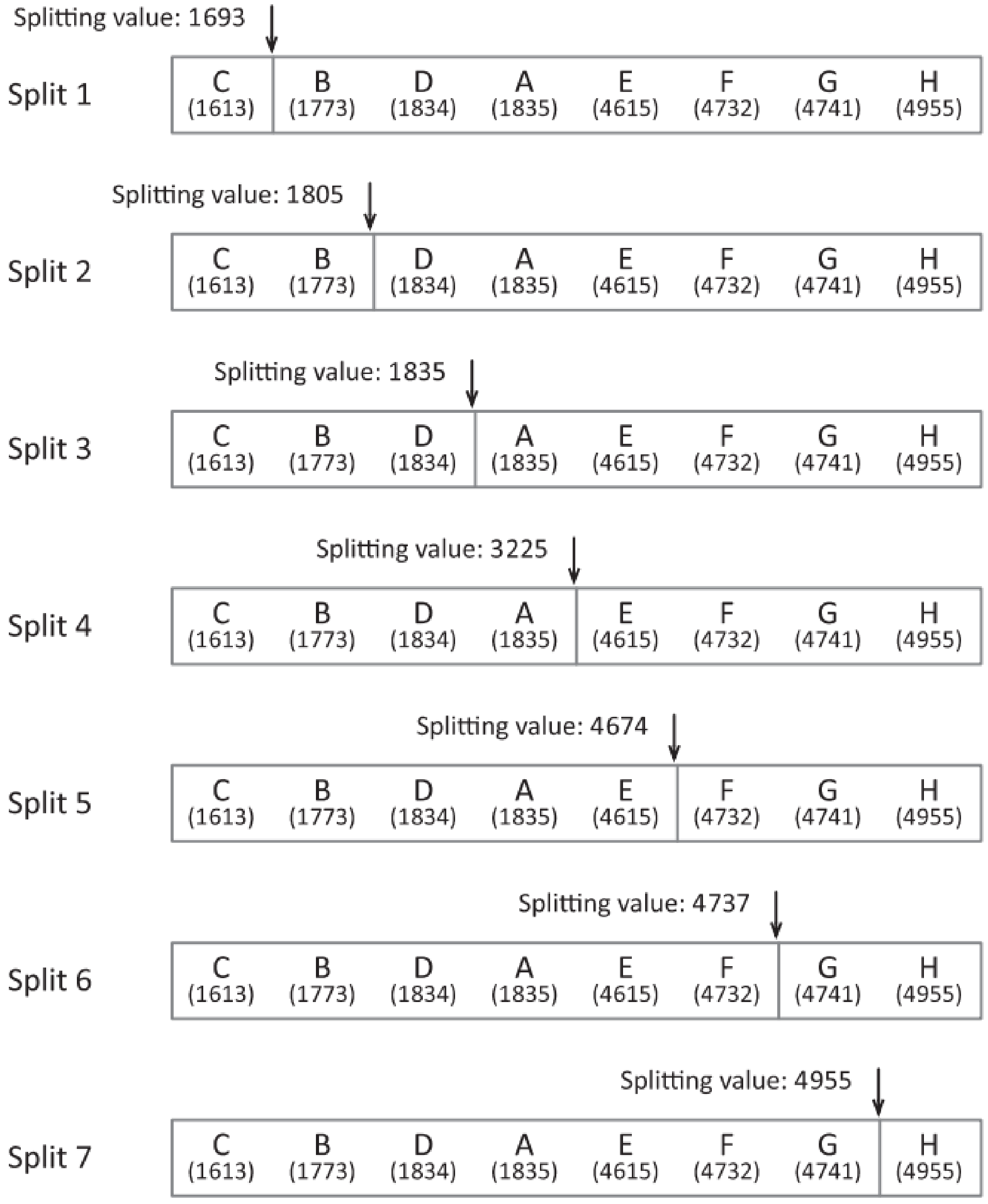


FIGURE 5.42 Illustration of splitting values.

Calculate score for splits which result in **three or more** observations

Split 3

For the subset where *Weight* is less than 1835 (C, B, D):

$$\text{Average} = (35 + 31 + 27)/3 = 31$$

$$\text{SSE} = (35 - 31)^2 + (31 - 31)^2 + (27 - 31)^2 = 32$$

For the subset where *Weight* is greater than or equal to 1835 (A, E, F, H, G):

$$\text{Average} = (26 + 10 + 9 + 13 + 12)/5 = 14$$

$$\text{SSE} = (26 - 14)^2 + (10 - 14)^2 + (9 - 14)^2 + (13 - 14)^2 + (12 - 14)^2 = 190$$

Split score = 32 + 190 = 222

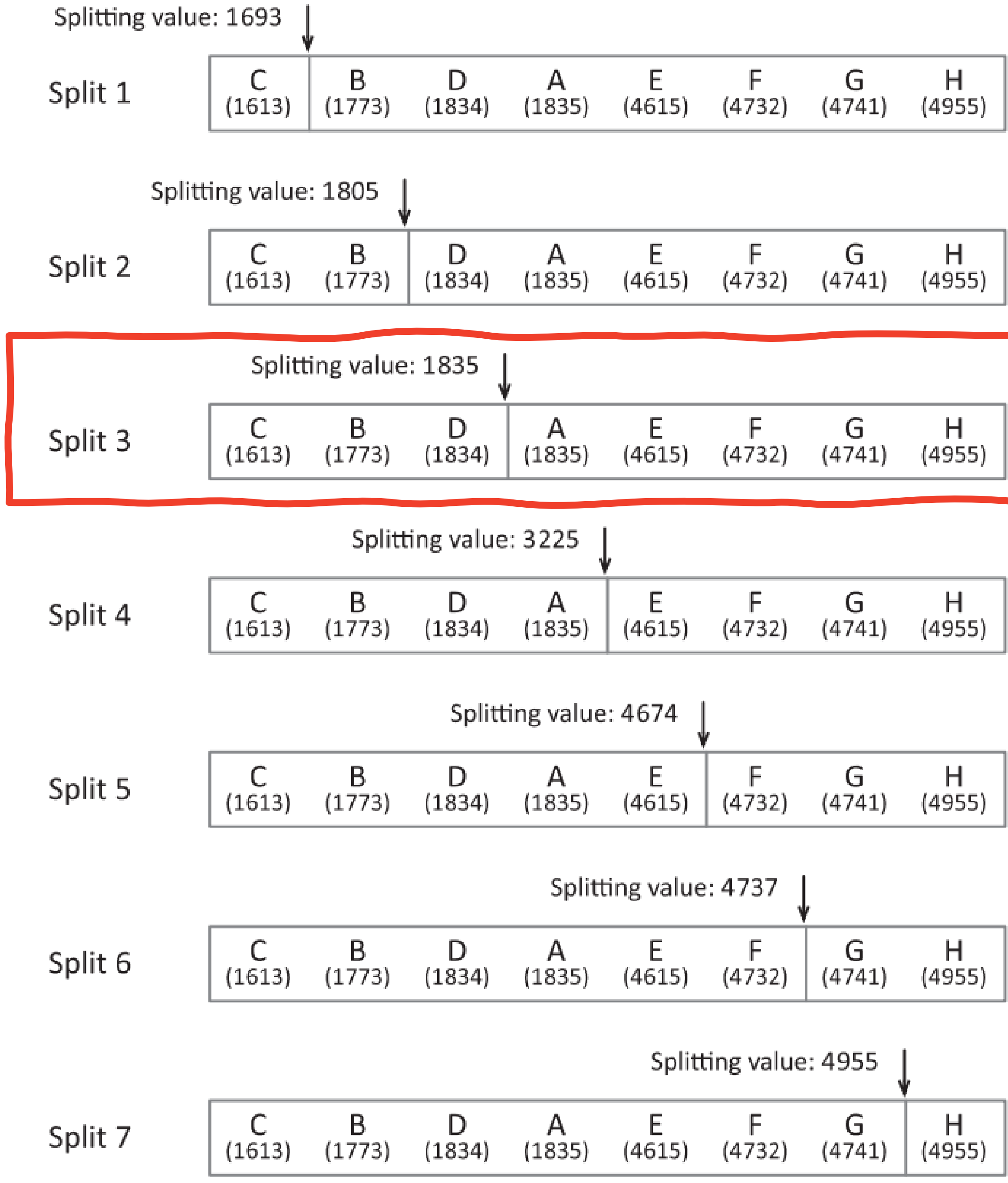


FIGURE 5.42 Illustration of splitting values.

Calculate score for splits which result in **three or more** observations

Split 4

For the subset where *Weight* is less than 3225 (C, B, D, A):

$$\text{Average} = (35 + 31 + 27 + 26)/4 = 29.75$$

$$\text{SSE} = (35 - 29.75)^2 + (31 - 29.75)^2 + (27 - 29.75)^2 + (26 - 29.75)^2 = 50.75$$

For the subset where *Weight* is greater than or equal to 3225 (E, F, H, G):

$$\text{Average} = (10 + 9 + 13 + 12)/4 = 11$$

$$\text{SSE} = (10 - 11)^2 + (9 - 11)^2 + (13 - 11)^2 + (12 - 11)^2 = 10$$

Split score = 50.75 + 10 = 60.75

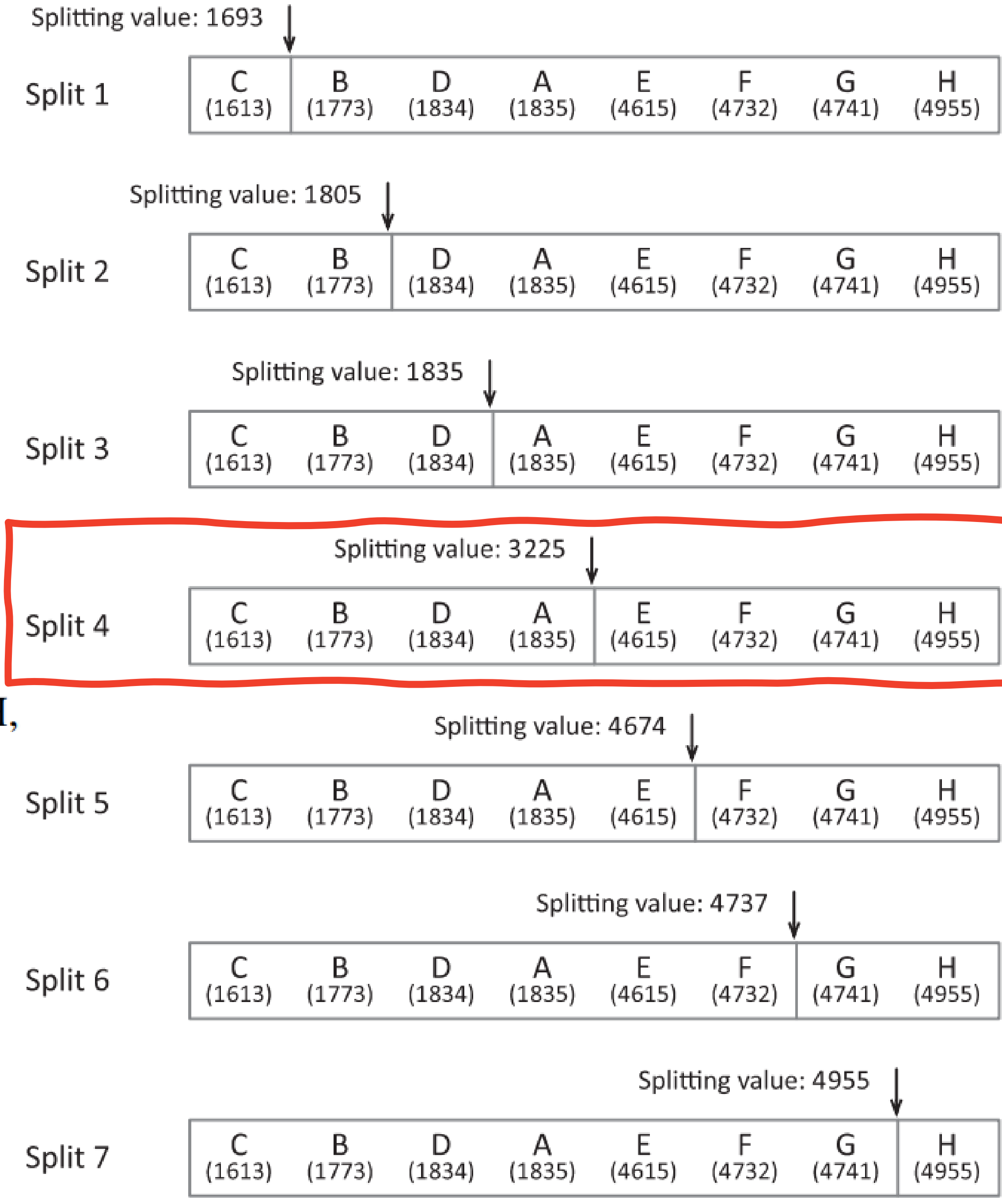


FIGURE 5.42 Illustration of splitting values.

Calculate score for splits which result in **three or more** observations

Split 5

For the subset where *Weight* is less than 4674 (C, B, D, A, E):

$$\text{Average} = (35 + 31 + 27 + 26 + 10)/5 = 25.8$$
$$\text{SSE} = (35 - 25.8)^2 + (31 - 25.8)^2 + (27 - 25.8)^2 + (26 - 25.8)^2 + (10 - 25.8)^2 = 362.8$$

For the subset where *Weight* is greater than or equal to 4674 (F, H, G):

$$\text{Average} = (9 + 13 + 12)/3 = 11.33$$
$$\text{SSE} = (9 - 11.33)^2 + (13 - 11.33)^2 + (12 - 11.33)^2 = 8.67$$

$$\text{Split score} = 362.8 + 8.67 = 371.$$

Split 4 has lowest score and would be selected as best split

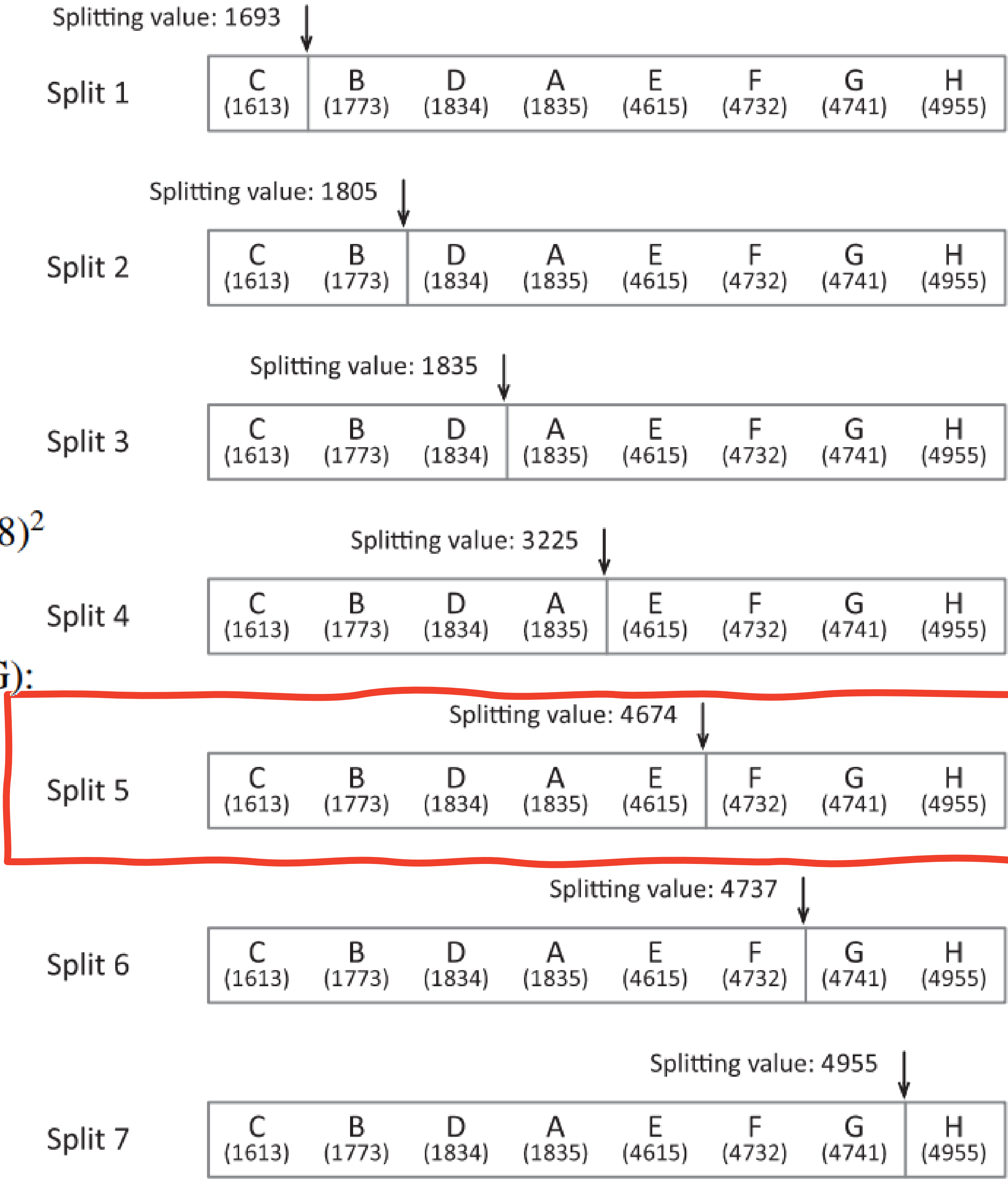


FIGURE 5.42 Illustration of splitting values.

The end.