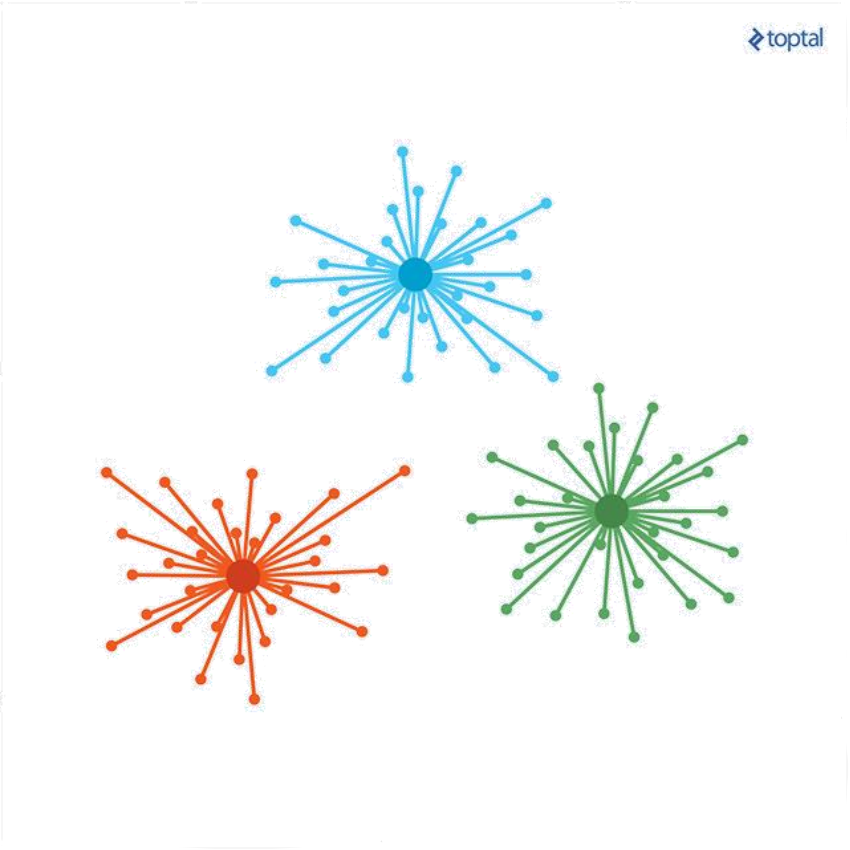
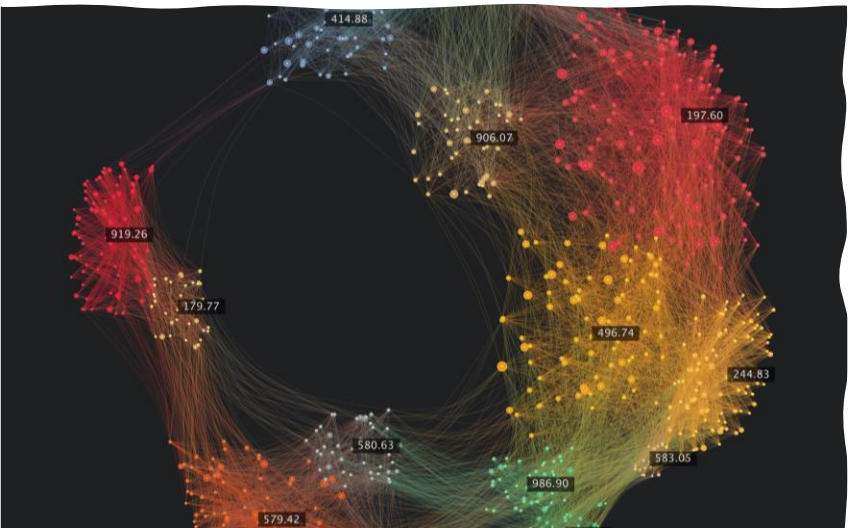
IDENTIFYING AND UNDERSTANDING GROUPS

Chapter 5

Clustering.

Analysing groups





Introduction

- For a given data set, ***not known beforehand,*** what **groups of observations** the **entire data** set is composed of.
- Example: Look at **customer data** collected from a **particular store**
  - Is it possible to **identify and summarize classes** of **customers** **directly from the data**
  - Can you answer questions such as *"What types of customers visit the store?"*
  - ***Clustering*** **groups data into sets of related observations or clusters**, so that **observations within each group** are **more similar** to other **observations within the group than** to observations **within other groups**.

Overview.

- **Clustering** is **a widely used and flexible approach** to analyzing data, in which **observations are automatically organized into groups.**
- Clustering is **an unsupervised method** for grouping:
  - Groups are **not known in advance** and **a goal (a specific variable)** is **not used** to direct **how the grouping** is generated - **all variables are considered** in the analysis.
- **Clustering method** chosen to **subdivide the data into groups** applies an automated procedure to discover the groups based on some criteria and its solution is **extracted from patterns or structure existing in the data.**
- Many clustering methods exist.
- For clustering, there is **no way to measure accuracy** and **the solution** is judged by its “usefulness.”
  - Clustering is **open ended way to explore, understand, and formulate questions** about the **data** in exploratory data analysis.

TABLE 3.1 Portion of a Table Describing Various Animals

Name	Hair	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Toothed	Backbone	Breathes	Venomous	Fins	Legs	Tail
Aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0
Bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1
Boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1
Buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1
Carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1
Catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1
Cheetah	1	0	0	1	0	0	1	1	1	1	0	0	4	1
Chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1
Clam	0	0	1	0	0	0	1	0	0	0	0	0	0	0

- **Cluster analysis** organizes the **data into groups** of.
- Using the **numeric variables** shown in Table 3.1, the data set can be clustered in a number of ways.

- Animals organized into **4** groups representing the following general categories:
  1. **Mammals:** Four-legged animals that produce milk, the fish,
  2. **Fish and amphibians:** Aquatic animals
  3. **Invertebrates:** Six-legged animals with no backbone, and
  4. **Birds:** set of two-legged airborne animals with feathers.

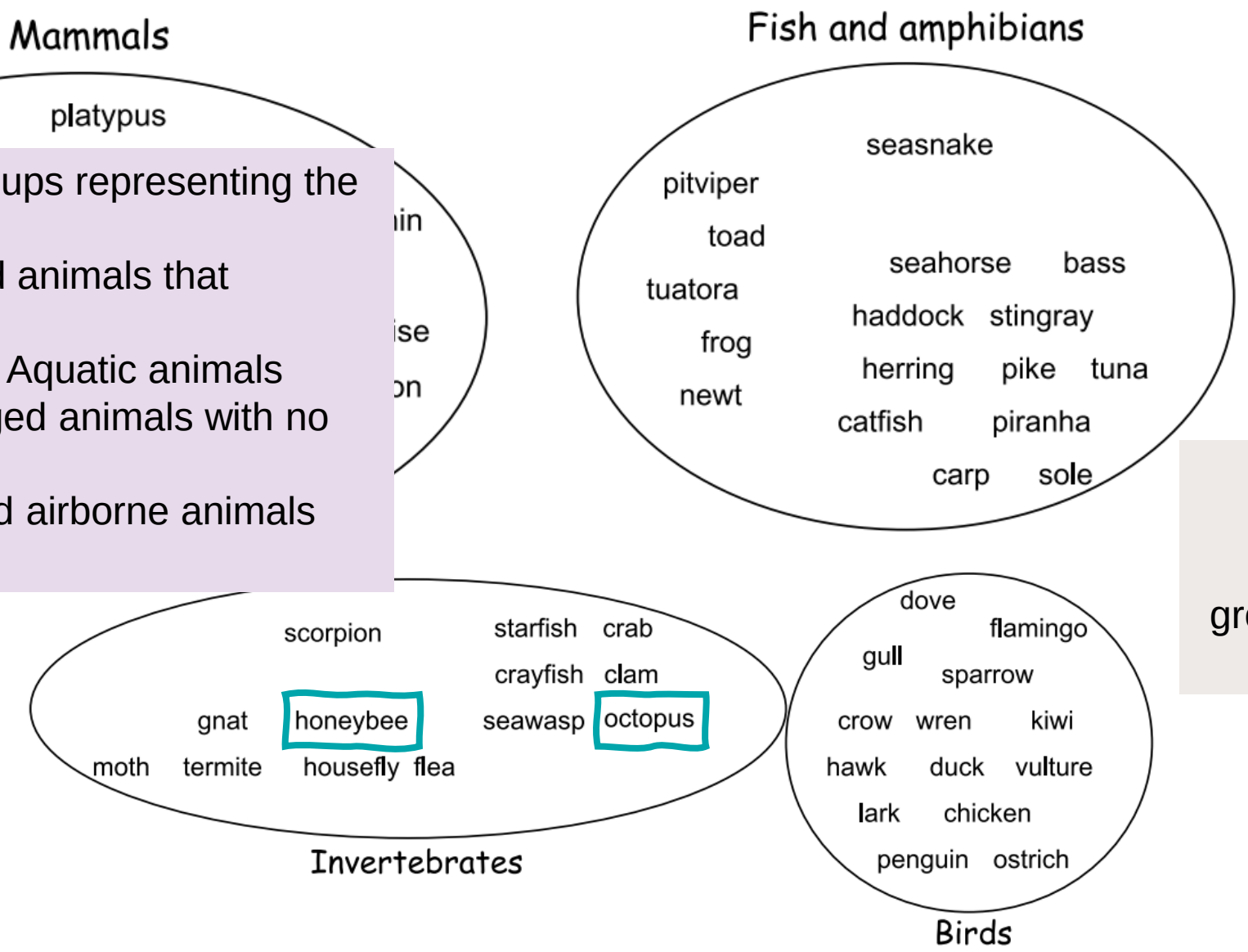


Figure 3.1 shows one possible grouping of this data

Figure 3.1 Four groups of animals in the data set

http://archive.ics.uci.edu/ml/datasets/ Zoo; Murphy and Aha, 1994) Updated: UCI Machine Learning Repository

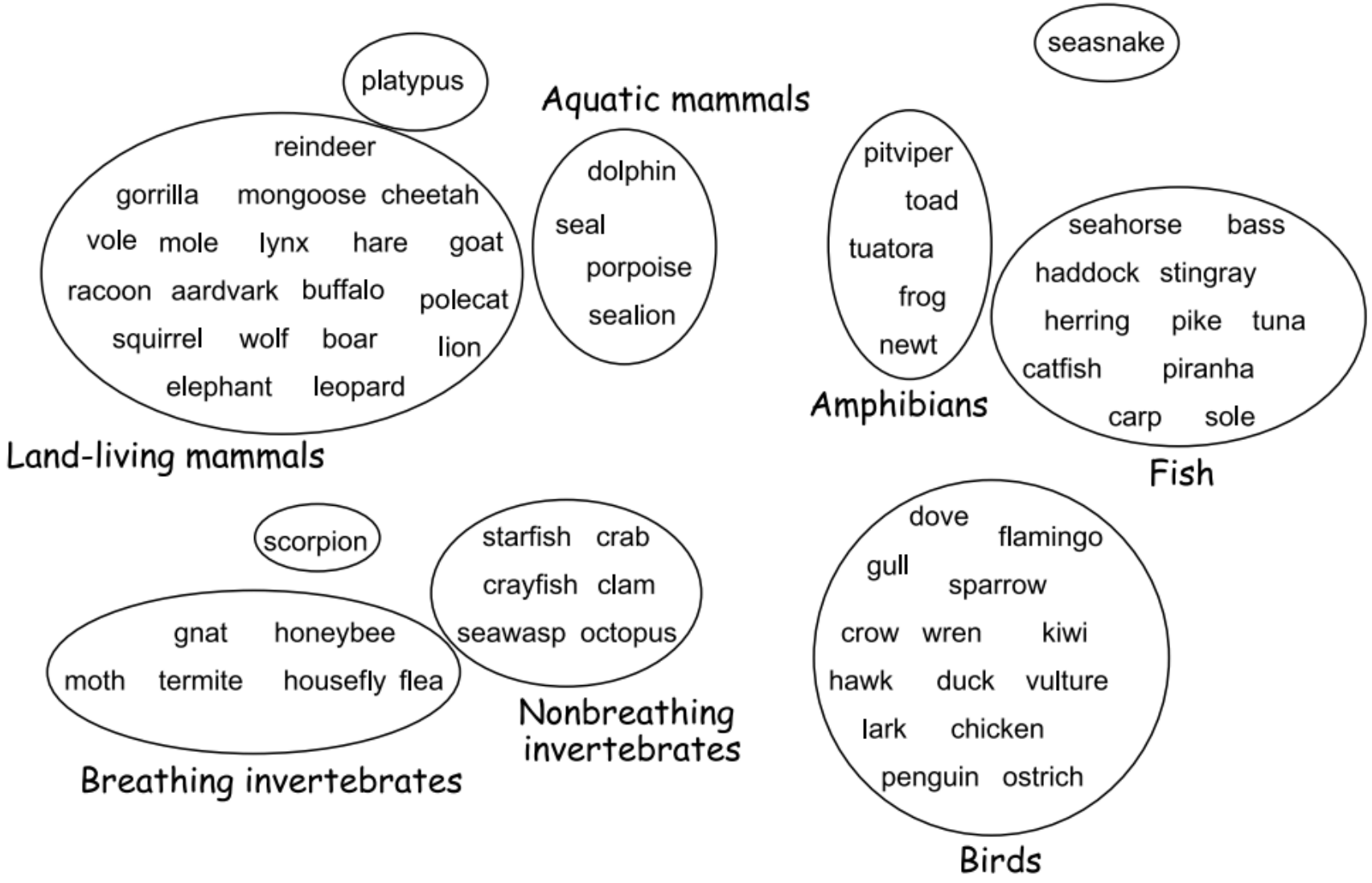


Figure 3.2 Ten groups generated from the animal data set

In Fig. 3.2,

Subsets of larger mammals:

- Land-living mammals group
- Aquatic mammals group
- Platypus

Platypus: Group of its own, since it is both aquatic and lays eggs.

Fish and amphibians and **invertebrates** groups have been divided into smaller groups.

Original group of birds remains the same since it already represented a set of similar animals.

http://archive.ics.uci.edu/ml/datasets/ Zoo; Murphy and Aha, 1994) Updated: UCI Machine Learning Repository

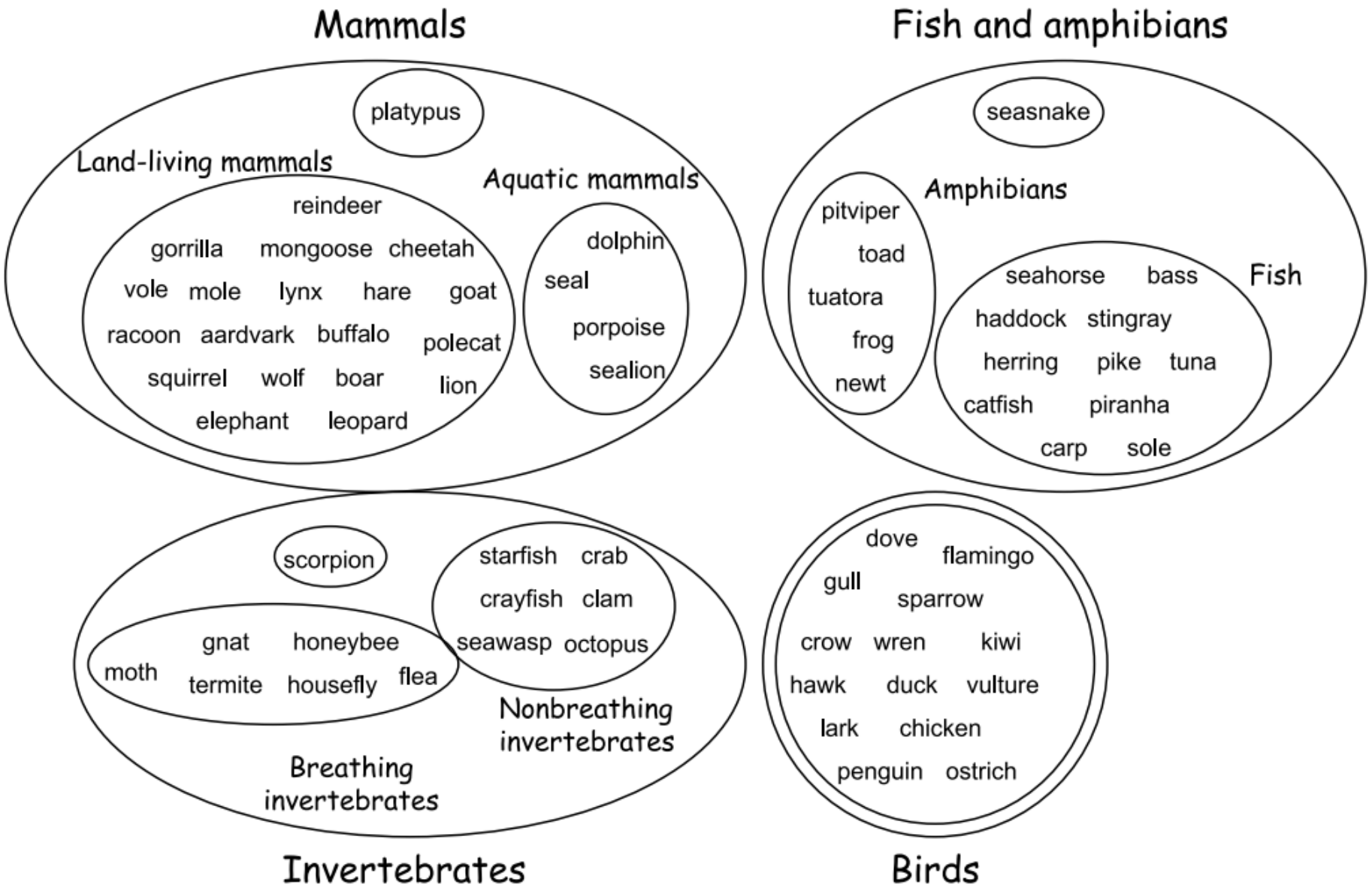


Figure 3.3 The two groupings superimposed

http://archive.ics.uci.edu/ml/datasets/ Zoo; Murphy and Aha, 1994) Updated: UCI Machine Learning Repository

Zoo Data Set

Download: Data Folder, Data Set Description

Abstract: Artificial, 7 classes of animals

Data Set Characteristics:	Multivariate	**Number**
Attribute Characteristics:	Categorical, Integer	**Number**
Associated Tasks:	Classification	**Missing**

Data Set Information:

A simple database containing 17 Boolean-valued attributes. The "type" attribute appea
are 2 instances of "frog" and one of "girl"!)

Class# -- Set of animals:
====== ===
1 -- (41) aardvark, antelope, bear, boar, buffalo, calf, cavy, cheetah, deer, dolphin, elep
oryx, platypus, polecat, pony, porpoise, puma, pussycat, raccoon, reindeer, seal, seali
2 -- (20) chicken, crow, dove, duck, flamingo, gull, hawk, kiwi, lark, ostrich, parakeet, p
3 -- (5) pitviper, seasnake, slowworm, tortoise, tuatara
4 -- (13) bass, carp, catfish, chub, dogfish, haddock, herring, pike, piranha, seahorse,
5 -- (4) frog, frog, newt, toad
6 -- (8) flea, gnat, honeybee, housefly, ladybird, moth, termite, wasp
7 -- (10) clam, crab, crayfish, lobster, octopus, scorpion, seawasp, slug, starfish, worm

Attribute Information:

1. animal name: Unique for each instance
2. hair: Boolean
3. feathers: Boolean
4. eggs: Boolean
5. milk: Boolean
6. airborne: Boolean
7. aquatic: Boolean
8. predator: Boolean
9. toothed: Boolean
10. backbone: Boolean
11. breathes: Boolean
12. venomous: Boolean
13. fins: Boolean
14. legs: Numeric (set of values: {0,2,4,5,6,8})
15. tail: Boolean
16. domestic: Boolean
17. catsize: Boolean
18. type: Numeric (integer values in range [1,7])

Attribute Name	Role	Type	Description	Units	Missing Values
animal_name	Other	Categorical	N/A	N/A	False
hair	Feature	Categorical	N/A	N/A	False
feathers	Feature	Categorical	N/A	N/A	False
eggs	Feature	Categorical	N/A	N/A	False
breathes	Feature	Categorical	N/A	N/A	False
venomous	Feature	Categorical	N/A	N/A	False
domestic	Feature	Categorical	N/A	N/A	False
catsize	Feature	Categorical	N/A	N/A	False
type	Target	Categorical	N/A	N/A	False

```
aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1
antelope,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
bass,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,4
bear,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1
boar,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,1
buffalo,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
calf,1,0,0,1,0,0,0,1,1,1,0,0,4,1,1,1,1
carp,0,0,1,0,0,1,0,1,1,0,0,1,0,1,1,0,4
catfish,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,4
cavy,1,0,0,1,0,0,0,1,1,1,0,0,4,0,1,0,1
cheetah,1,0,0,1,0,0,1,1,1,1,0,0,4,1,0,1,1
chicken,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,2
chub,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,4
clam,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,7
crab,0,0,1,0,0,1,1,0,0,0,0,0,4,0,0,0,7
crayfish,0,0,1,0,0,1,1,0,0,0,0,0,6,0,0,0,7
crow,0,1,1,0,1,0,1,0,1,1,0,0,2,1,0,0,2
deer,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
dogfish,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,1,4
dolphin,0,0,0,1,0,1,1,1,1,1,0,1,0,1,0,1,1
dove,0,1,1,0,1,0,0,0,1,1,0,0,2,1,1,0,2
duck,0,1,1,0,1,1,0,0,1,1,0,0,2,1,0,0,2
elephant,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
flamingo,0,1,1,0,1,0,0,0,1,1,0,0,2,1,0,1,2
flea,0,0,1,0,0,0,0,0,0,1,0,0,6,0,0,0,6
frog,0,0,1,0,0,1,1,1,1,1,0,0,4,0,0,0,5
frog,0,0,1,0,0,1,1,1,1,1,1,0,4,0,0,0,5
fruitbat,1,0,0,1,1,0,0,1,1,1,0,0,2,1,0,0,1
giraffe,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1
```

Class	Count	Mean (petal width (cm))
Iris-setosa	50	0.244
Iris-versicolor	50	1.33
Iris-virginica	50	2.03

FIGURE 5.1 Simple summary table showing how the mean petal width changes for the different classes of flowers.

Question: How a single variable such as petal width varies among different species as illustrated in Figure 5.1?

- Easily extended to help understand the relationships between groups and multiple variables, as illustrated in Figure 5.2 where **three predefined categories** are used to **group the observations**.
- **Summary information on multiple variables** is presented (using the mean value in the example).
- These tables can use summary statistics (e.g., mean, mode, median, and so on) in addition to graphs such as box plots that illustrate the subpopulations.

Class	Count	Mean (sepal length (cm))	Mean (sepal width (cm))	Mean (petal length (cm))	Mean (petal width (cm))
Iris-setosa	50	5.01	3.42	1.46	0.244
Iris-versicolor	50	5.94	2.77	4.26	1.33
Iris-virginica	50	6.59	2.97	5.55	2.03

FIGURE 5.2 The use of a summary table to understand multiple variables for a series of groups.

A **variety of graphs** (e.g., histograms, box plots, and so on) **for each group** can also be shown in a table or grid format known as small multiples that allows comparison.

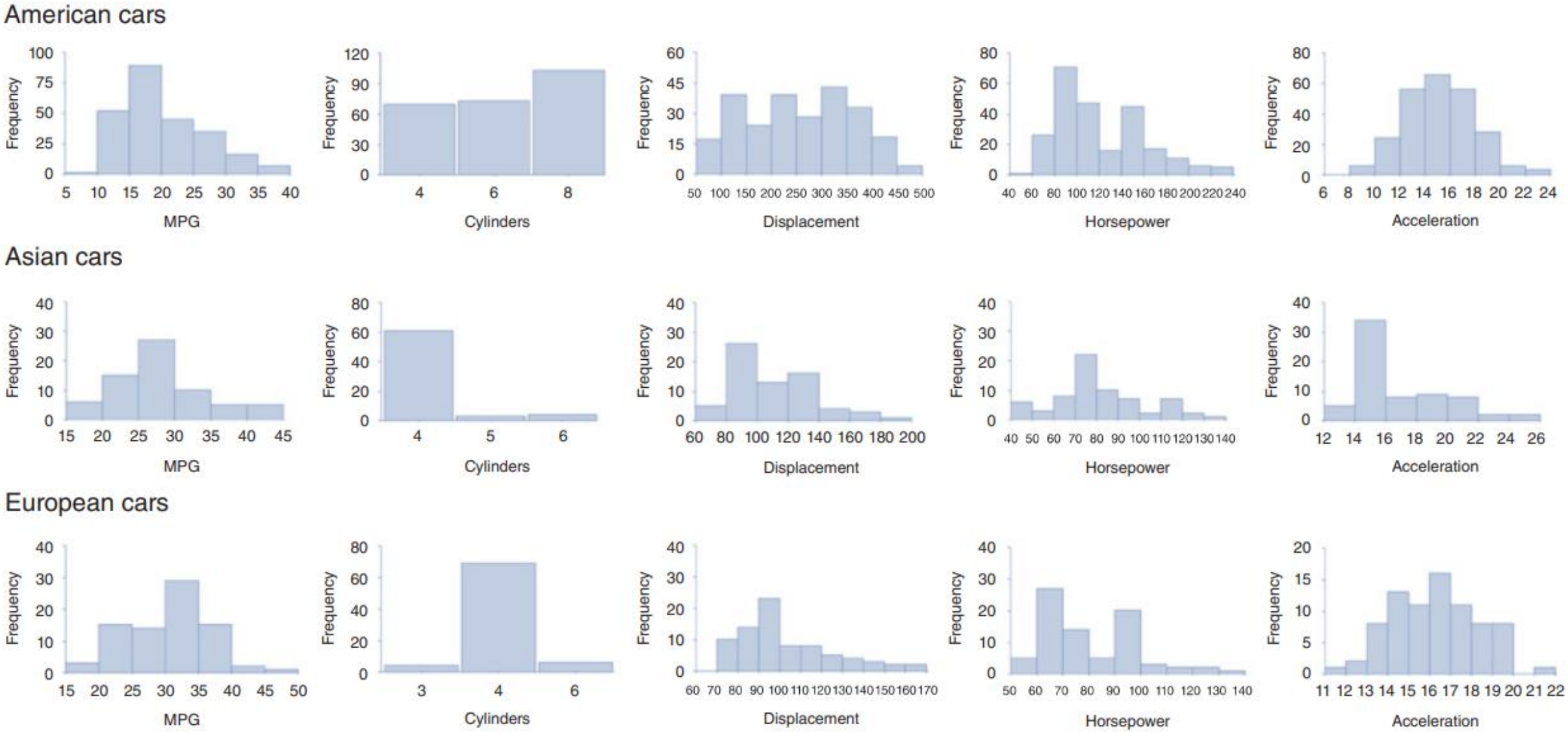


FIGURE 5.3 Matrix showing the frequency distribution for a common set of variables for three groups of cars—American, European, and Asian.

A **variety of graphs** (e.g., histograms, box plots, and so on) **for each group** can also be shown in a table or grid format known as small multiples that allows comparison.

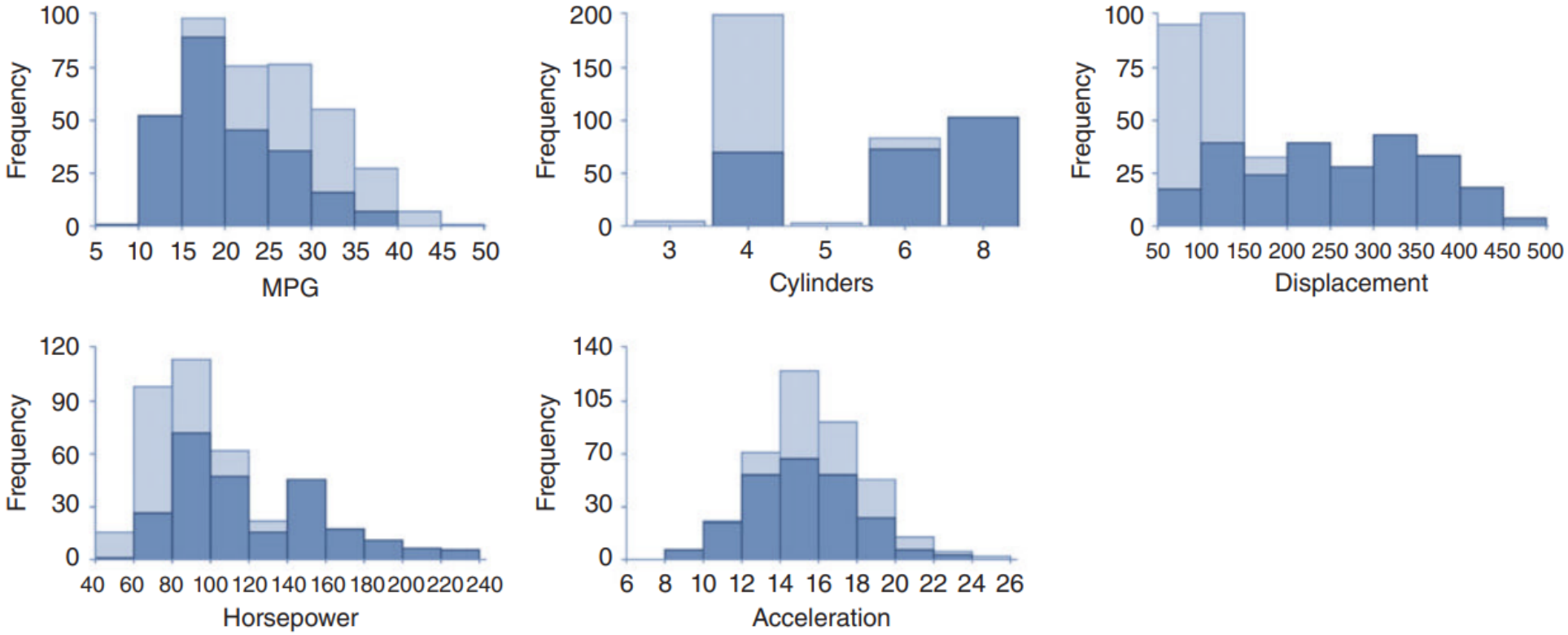


FIGURE 5.4 Highlighting a group of observations using shading.

Clustering.

To illustrate the process of clustering, **a set of observations** are shown on the scatterplot.

- Observations plotted using two hypothetical dimensions
- **Similarity** between observations is **proportional to the physical distance** between the observations.
- **Two clear regions** that can be considered as clusters:
  - Cluster A and Cluster B, since many of the observations are contained within these two regions on the scatterplot
- Clustering not only assists in **identifying groups of related observations**, can also **locate outliers (**observations that are not similar to others) since they fall into groups of their own

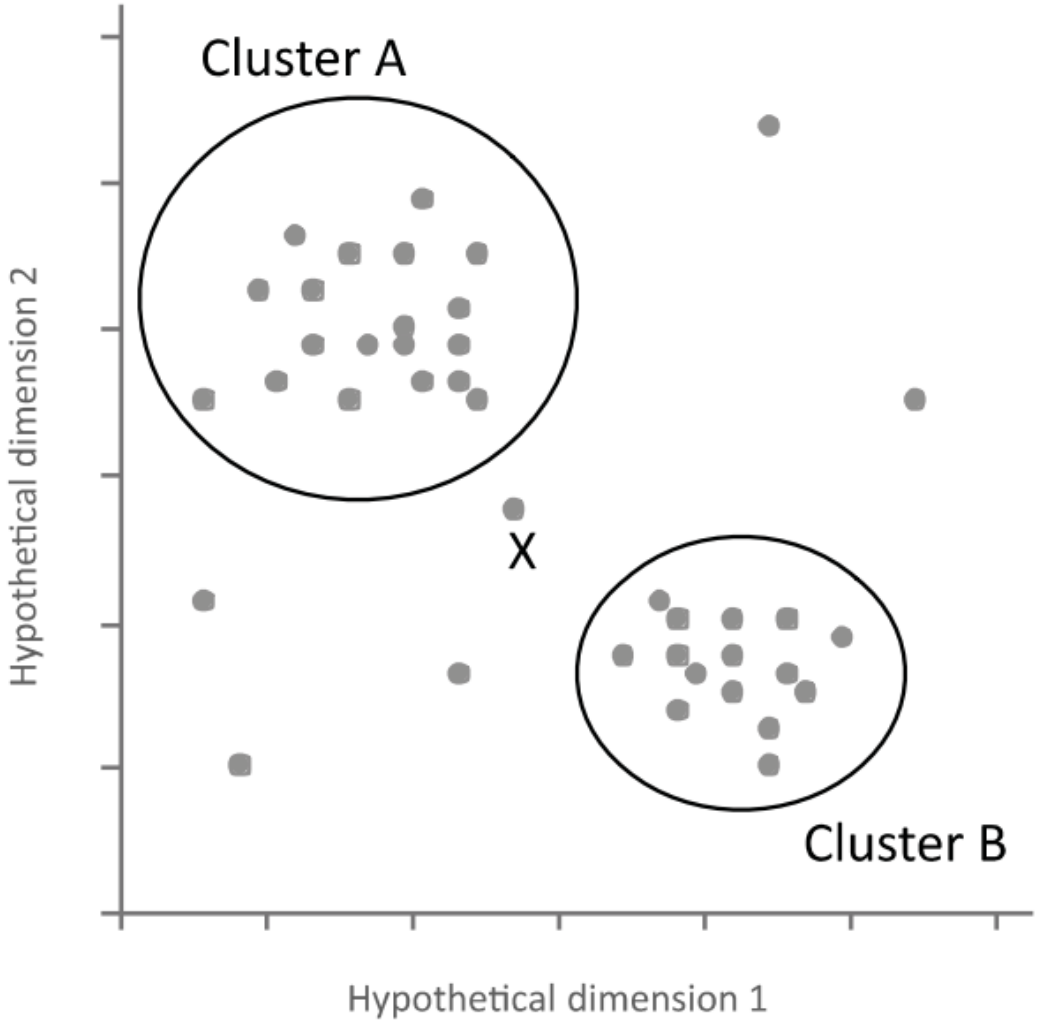


FIGURE 5.5 Illustration of clusters and outliers

Clustering.

- **Cluster analysis** requires **neither a prior understanding** of the data nor **assumptions** to be made about it.
- For example, it is **not necessary** that **variables** follow a **normal distribution**, nor is it necessary to **identify independent** and **response variables**
- **Care** should be taken to **select the appropriate variables,** including only those **relevant** to the **grouping exercise**.
- **Results** from a cluster analysis are **groups of observations** that **share common characteristics**.

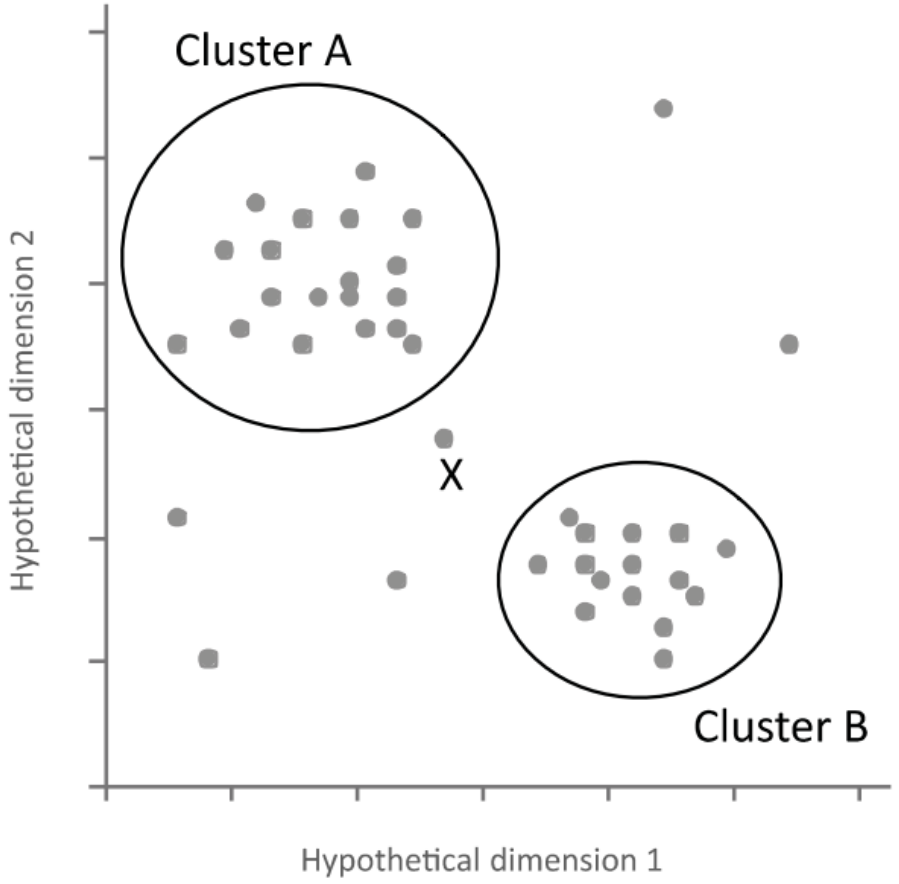


FIGURE 5.5 Illustration of clusters and outliers.

The way to get started is to quit talking and begin doing.

Walt Disney

Clustering.

- Is a **flexible approach for grouping**.
- For example, based on the **criteria for clustering** the observations, observation **X** was **not** determined to be a **member of cluster A**.
- However, if a more **relaxed criterion** was used, **X may** have been **included** in **cluster A.**
- In Figure 5.5, there are **six observations** that **do not fall within** cluster **A or B**, observed to be **outliers.**

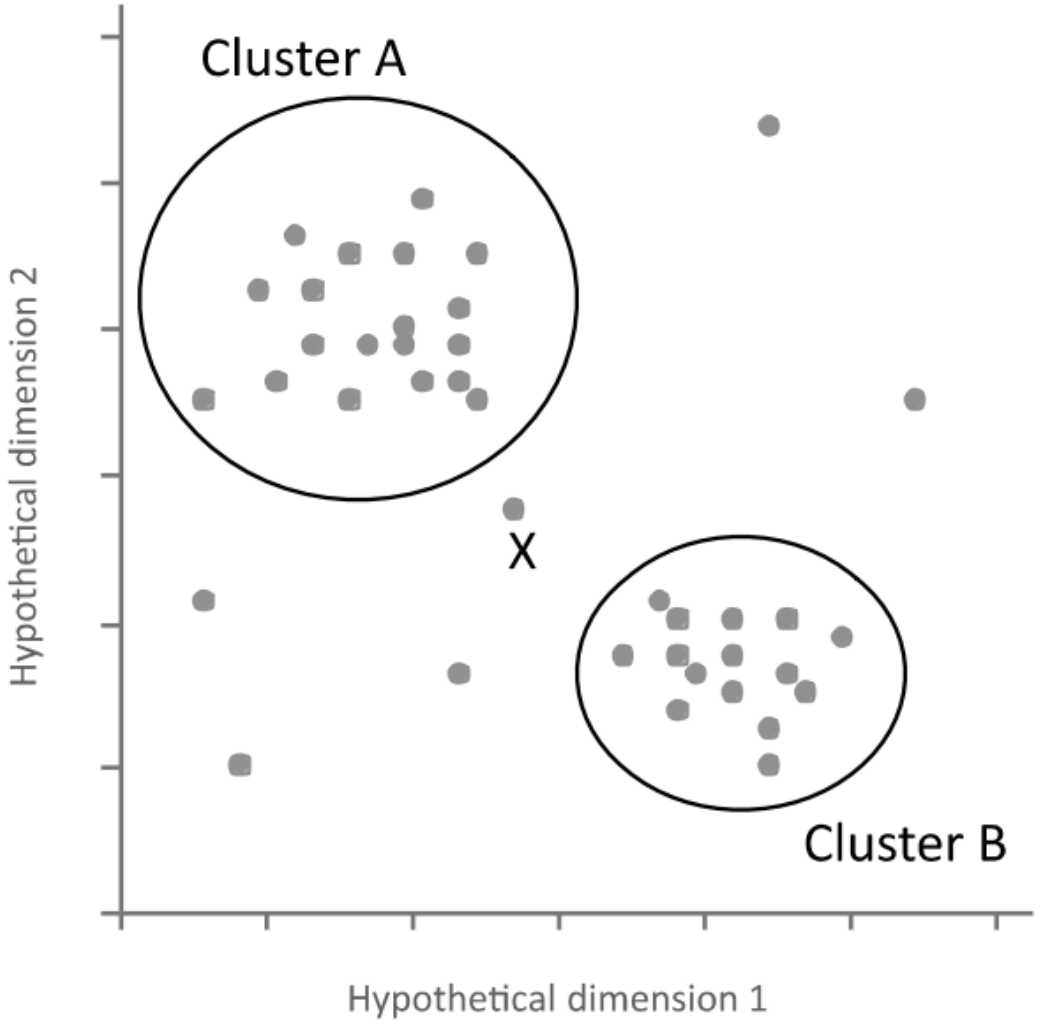


FIGURE 5.5 Illustration of clusters and outliers.

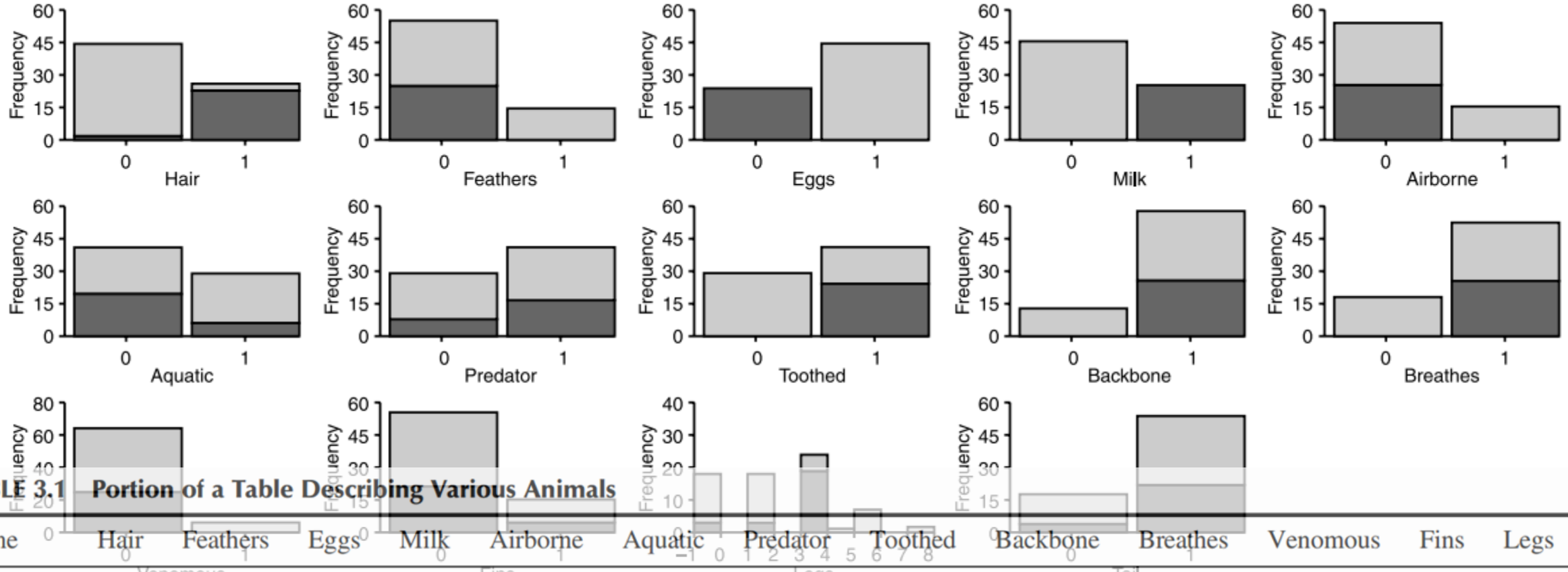


Figure 3.4 Histogram matrix for the animals in the mammals groups compared to the total data set

TABLE 3.1 Portion of a Table Describing Various Animals

Name	Hair	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Toothed	Backbone	Breathes	Venomous	Fins	Legs	Tail
Aardvark	1	0	0	1	0	0	1	1	1	1	0	0	4	0
Bass	0	0	1	0	0	1	1	1	1	0	0	1	0	1
Boar	1	0	0	1	0	0	1	1	1	1	0	0	4	1
Buffalo	1	0	0	1	0	0	0	1	1	1	0	0	4	1
Carp	0	0	1	0	0	1	0	1	1	0	0	1	0	1
Catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1
Cheetah	1	0	0	1	0	0	1	1	1	1	0	0	4	1
Chicken	0	1	1	0	1	0	0	0	1	1	0	0	2	1
Clam	0	0	1	0	0	0	1	0	0	0	0	0	0	0

Clustering methods.

- **Different** types of **clustering methods** exist. **Selection** of a **method** influenced by:
  - **Complexity**
  - **Time** it takes to generate the clusters, and
  - **Number of observations** it can process.
- Certain methods require **number of clusters** to be generated **chosen prior** to the analysis; Other methods this **number determined** based on an **analysis** of the results.
- Clustering is **dependent** on the **data set being analyzed:**
  - Different data sets result in different groups of observations, even if data sets contain similar content.
  - Some cases: adding single new observation to existing set can have substantial affect on the groupings.
  - Ties in similarity scores can affect groupings, and reordering the observations may give different results.
- Number of **clustering methods**, the most common:

Hierarchical | Partitioned | Fuzzy

Hierarchical.

Organize observations **hierarchically.**

For example, **animal dataset** can be organized by **four high level** categories which are then subsequently organized into **sub-categories,**

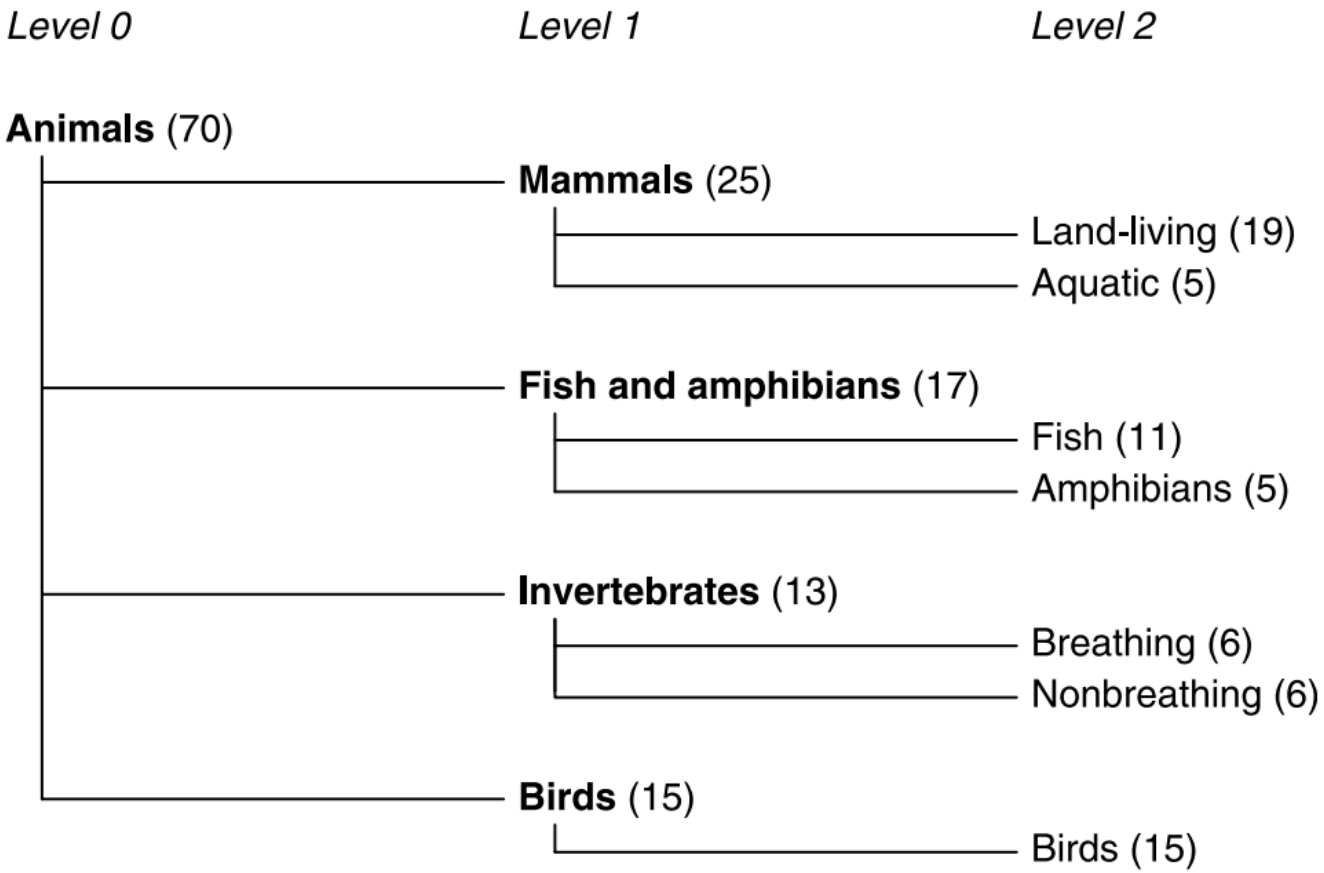


Figure 3.5 Hierarchical relationships for the animal data set

Partitioned.

Assigns observations to a **single cluster** and number of clusters is set prior to any cluster **analysis.**

For example, Fig. 3.1 generated as result of clustering animal data set where **number of clusters** was preset to four.

Does not provide a **flexible hierarchical organization,** is generally faster to compute and can handle larger data sets

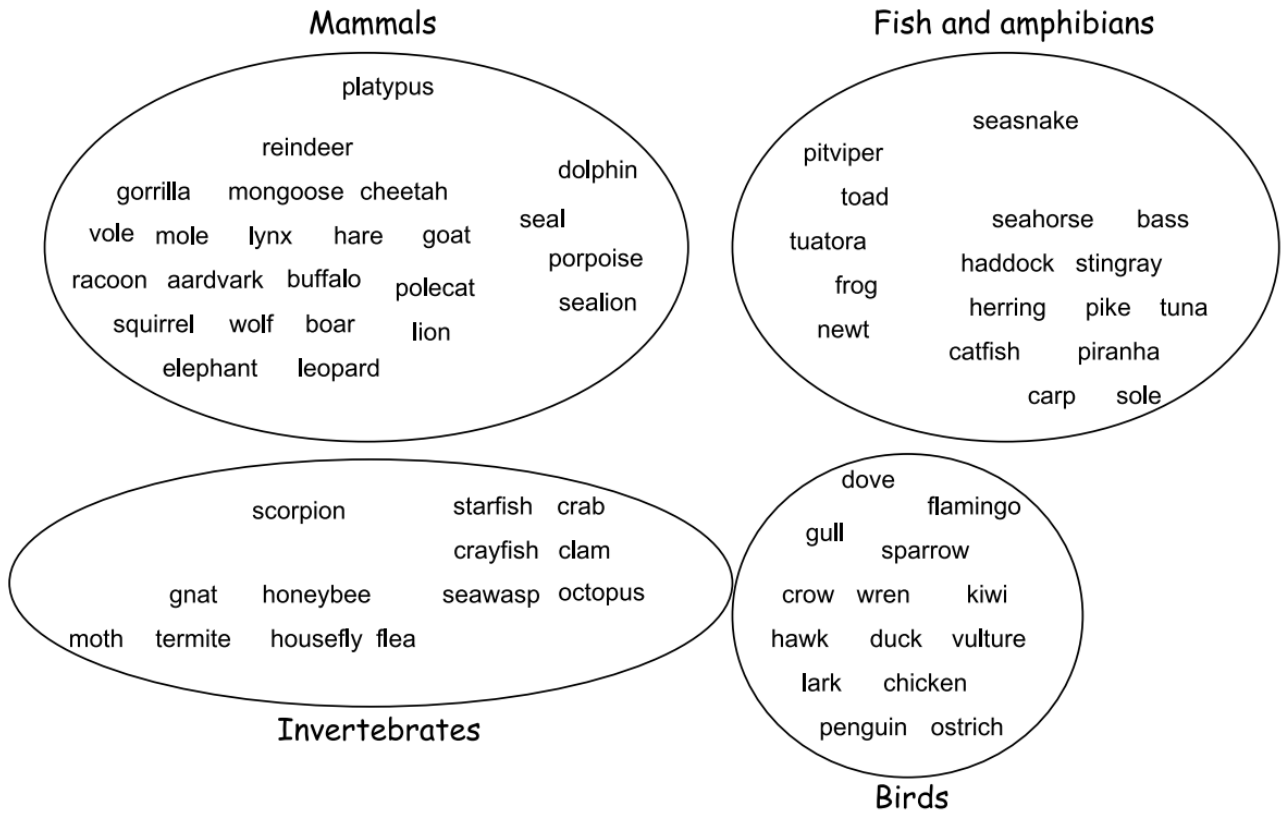


Figure 3.1 Four groups of animals in the data set

Distance measures.

- A **requirement** for any **clustering exercise** is **calculating the distance *between two observations***.
- **All clustering approaches require** a **formal approach** to defining how **similar two observations** are to each other **as measured** by the **distance between them**.

Distance: Method of clustering to **measure how similar observations are to each other**

Distances.

To **calculate similarity**, ***compute*** **the** **distance between observations.**

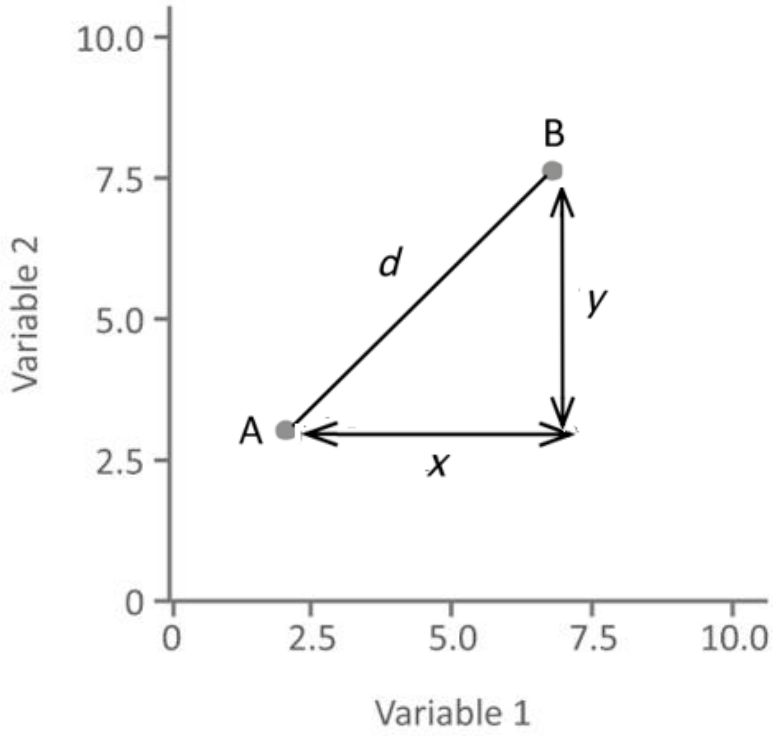


FIGURE 5.6 Distance between two observations A and B.

TABLE 5.1 Table Showing Two Observations A and B

Observation ID	Variable 1	Variable 2
A	2	3
B	7	8

Physical distance

Distances.

TABLE 3.3 Information on Customers

Customer	Geographic	Income	Employer information	Age	Children	Gender	Electronics	Groceries	Clothes
A	Urban	85,000	Technology	31	0	Male	15	1	0
B	Suburbs	38,000	Services	43	3	Female	1	33	8
C	Suburbs	150,000	Legal	51	2	Female	5	0	0
D	Suburbs	32,000	Services	49	1	Female	2	29	9
E	Urban	104,000	Technology	36	0	Male	8	0	1

- **Starting point** for **any cluster analysis** is **data table or data matrix.**
- Example: Retail company gathered information about its customers: addresses, income, occupations, along with purchase categories information those customers made in the year.
- Table 3.3: Five customer entries. **Cluster analysis can only operate on numerical data**.
- **Any nonnumeric values** should be **converted** to **numbers.**

TABLE 3.4 Information on Customers Converted to Numbers

Transformed values

Customer	Urban	Income	Technology	Services	Legal	Age	Children	Female	Electronics	Groceries	Clothes
A	1	85,000	1	0	0	31	0	0	15	1	0
B	0	38,000	0	1	0	43	3	1	1	33	8
C	0	150,000	0	0	1	51	2	1	5	0	0
D	0	32,000	0	1	0	49	1	1	2	29	9
E	1	104,000	1	0	0	36	0	0	8	0	1

Distances.

TABLE 3.5 Information on Customers, Normalized to the Range 0 to 1

Customer	Urban	Income	Technology	Services	Legal	Age	Children	Female	Electronics	Groceries	Clothes
A	1	0.449	1	0	0	0	0	0	1	0.03	0
B	0	0.051	0	1	0	0.6	1	1	0	1	0.89
C	0	1.000	0	0	1	1	0.67	1	0.29	0	0
D	0	0.000	0	1	0	0.9	0.33	1	0.071	0.88	1
E	1	0.610	1	0	0	0.25	0	0	0.5	0	0.11

- Common to **normalize the data** to standard ranges, such as **between zero and one** or using **z-score** (the number of standard deviations above or below the mean), to ensure a variable is not considered more important as a consequence of the range on which it is measured.
- Table 3.5: Variables: income, age, children, electronics, groceries, and clothes have all been normalized (range 0-1).
- After deciding which variables to use, a **method for determining the distance** must be selected. Many methods to choose from: **Euclidean, Manhattan**.

Distances.

- Distance measures generally **share certain properties**:
  1. **Any distance measure between two observations ≥ 0**
  2. **Distance** between **observations A and B** is the **same** as the distance between **observations B and A.**
  3. If **distance is zero**, there is **no difference** between the two observations, that is, the two observations have the **same values** for **all variables.**
  4. Distance between A and B ($d_{A,B}$) satisfies the following assumptions with respect to a third observation (C), based on the distance between A and C ($d_{A,C}$) and the distance between B and C ($d_{B,C}$):

$$d_{A,B} \leq d_{A,C} + d_{B,C}$$


After the break.

Distances.

- Clustering methods generally require a **distance matrix** or **dissimilarity matrix (D)** as **input**.
  - **Square n-by-n matrix**, where **n** is number of observations in the data set to be clustered.
- This distance matrix has the following format:

Diagonal values are **all zero** distances since there **is no distance between two identical observations**

$$D = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ \\ n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & & n \end{array} \\ \begin{bmatrix} 0 & d_{1,2} & d_{1,3} & \dots & d_{1,n} \\ d_{2,1} & 0 & d_{2,3} & \dots & d_{2,n} \\ d_{3,1} & d_{3,2} & 0 & \dots & d_{3,n} \\ \dots & \dots & \dots & \dots & \dots \\ d_{n,1} & d_{n,2} & d_{n,3} & \dots & 0 \end{bmatrix} \end{array}$$

Where:
$d_{i,j}$: Distance value between pairs of observations;
$d_{1,2}$: Distance between first and second observation..

1.Numeric Distance Measures.

Numeric Distance Measures.

- Distance measures **"compute a number"** for **any pair** of **observations.**
- These **measures compare** the **values for each of the variables** in the **two observations** and **compute a distance** based on some **function relating to the differences** between these values

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

Table is limited to six normalized variables (income, age, children, electronic, groceries, and clothes

Distances: Calculation

Euclidean.

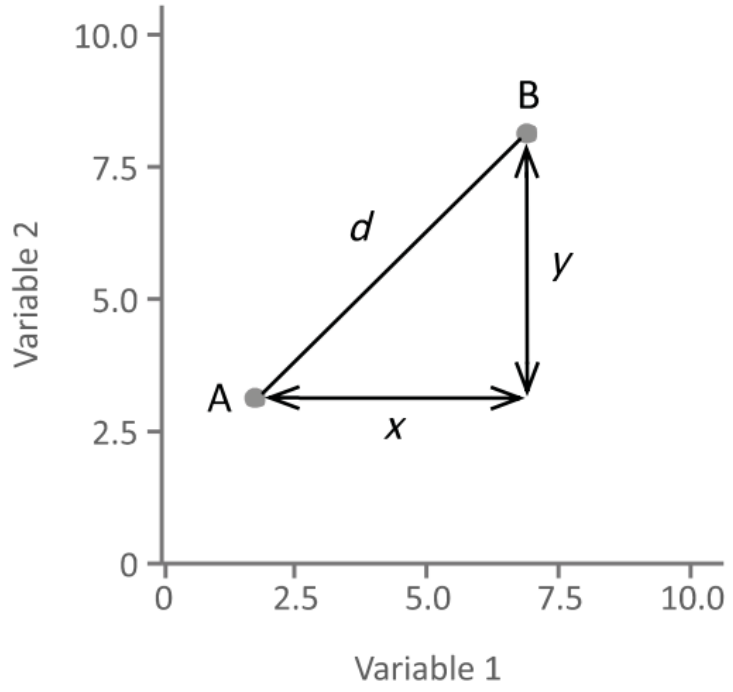
- **Distance** between the **two observations** is calculated using **simple trigonometry:**

$$x = 7 - 2 = 5$$

$$y = 8 - 3 = 5$$

$$d = \sqrt{x^2 + y^2} = \sqrt{25 + 25} = 7.07$$

- **Euclidean distance** is one of most common distance functions.
- If a data set only has **two variables**, then Euclidean distance would **calculate the physical distance** between the two points plotted on a scatterplot.



IGURE 5.6 Distance between two observations A and B.

Euclidean Distance (d).

The following formula calculates the **Euclidean distance** between **two observations (p and q),** measured over **n** variables:

$$d_{p,q} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

TABLE 5.2 Three Observations with Values for Five Variables

ID	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
A	0.7	0.8	0.4	0.5	0.2
B	0.6	0.8	0.5	0.4	0.2
C	0.8	0.9	0.7	0.8	0.9

The Euclidean distance between A and B is

$$d_{A-B} = \sqrt{(0.7-0.6)^2 + (0.8-0.8)^2 + (0.4-0.5)^2 + (0.5-0.4)^2 + (0.2-0.2)^2}$$

$$d_{A-B} = 0.17$$

The Euclidean distances between A and C is

$$d_{A-C} = \sqrt{(0.7-0.8)^2 + (0.8-0.9)^2 + (0.4-0.7)^2 + (0.5-0.8)^2 + (0.2-0.9)^2}$$

$$d_{A-C} = 0.83$$

The Euclidean distance between B and C is

$$d_{B-C} = \sqrt{(0.6-0.8)^2 + (0.8-0.9)^2 + (0.5-0.7)^2 + (0.4-0.8)^2 + (0.2-0.9)^2}$$

$$d_{B-C} = 0.86$$

The distance between **A and B is 0.17,** whereas distance between **A and C is 0.83**, which indicates that there is **more similarity between observations A and B than A and C**. **C is not closely related to either A or B.**

Euclidean Distance (d).

- Figure 5.7 where the **values for each variable** are plotted along the **horizontal axis** and the **height** of the bar measured against the v**ertical axis** represents the **data value.**
- The **shape of histograms A and B** are **similar,** whereas the shape of histogram **C is not similar to A or B.**

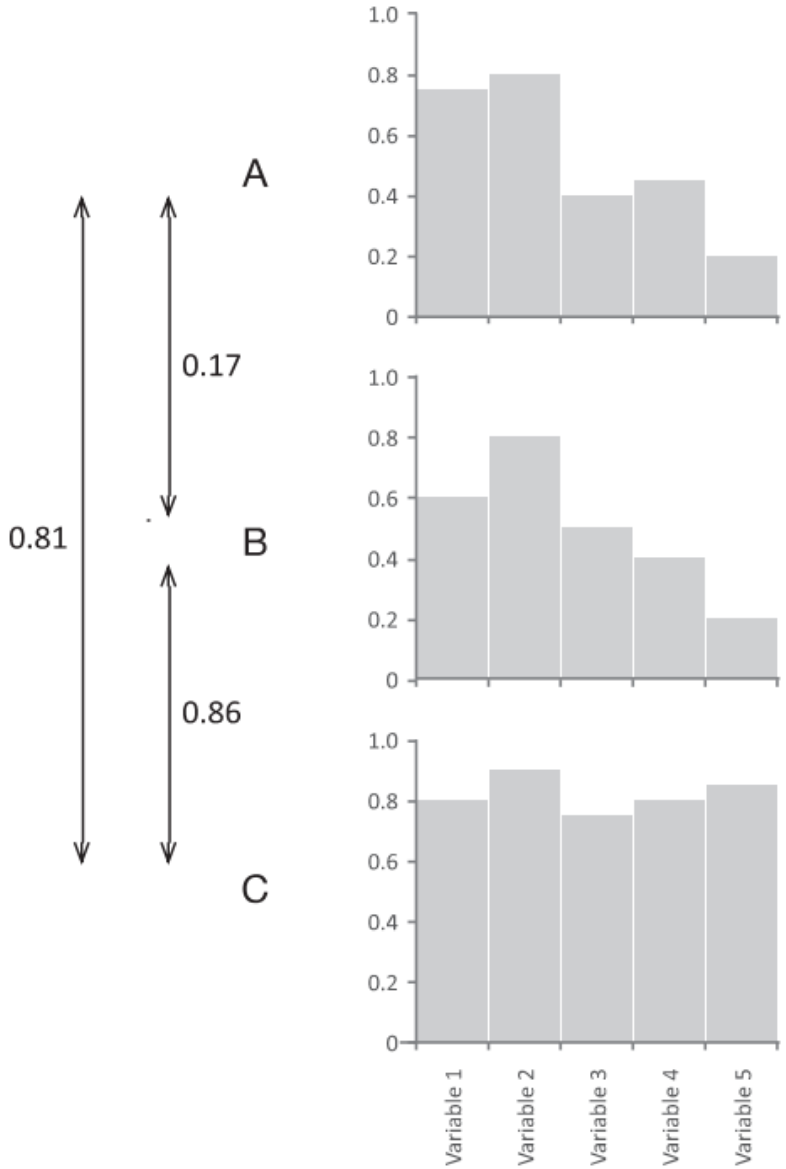


FIGURE 5.7 Distances between three observations: A–B, B–C, and A–C.

Euclidean Distance (d).

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

$$d_{p,q} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

The Euclidean distance between A and B is

$$d_{A,B} = \sqrt{\begin{aligned}&(0.449 - 0.051)^2 + (0 - 0.6)^2 + (0 - 1)^2 \\ &+ (1 - 0)^2 + (0.03 - 1)^2 + (0 - 0.89)^2\end{aligned}}$$

$$d_{A,B} = 2.061$$

Square Euclidean.

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

- The **square Euclidean** is the **sum of the squares** of the difference between the two observations, and it is calculated using the following formula:

$$d_{p,q} = \sum_{i=1}^{n} (p_i - q_i)^2$$

- To **calculate the square euclidean** distance between observations A and B using the data in Table 3.6, the following calculation is made:

$$d_{A,B} = (0.449 - 0.051)^2 + (0 - 0.6)^2 + (0 - 1)^2 + (1 - 0)^2$$

$$+ (0.03 - 1)^2 + (0 - 0.89)^2$$

$$d_{A,B} = 4.249$$

Manhattan Distance.

TABLE 3.6 Six Numeric Variables Describing Five Customers

Customer	Income	Age	Children	Electronics	Groceries	Clothes
A	0.449	0	0	1	0.03	0
B	0.051	0.6	1	0	1	0.89
C	1.000	1	0.67	0.29	0	0
D	0.000	0.9	0.33	0.071	0.88	1
E	0.610	0.25	0	0.5	0	0.11

Manhattan distance, also called the **city block distance**, is the **sum of the absolute distances** between the variables, which is **always a positive value** representing the difference. The formula is

$$d_{p,q} = \sum_{i=1}^{n} |p_i - q_i|$$

$$d_{A,B} = |0.449 - 0.051| + |0 - 0.6| + |0 - 1| + |1 - 0| + |0.03 - 1| + |0 - 0.89|$$

$$d_{A,B} = 4.856$$

Binary distance measures.

Binary Distance Measures.

- Distances calculated for **categorical values** because they can be easily **converted to binary dummy variables**.
- **For each variable**, the values in the **two observations are compared** to determine whether they are **the same or different.**

TABLE 3.8 Four Alternatives (*A*, *B*, *C*, and *D*) for Comparing Two Binary Values

		q_i	
		1	0
p_i	1	A	B
	0	C	D

TABLE 3.7 Distances between all Pairs of Observations

	Euclidean	Square euclidean	Manhatten
A,*B*	2.061	4.249	4.856
A,*C*	1.502	2.258	2.962
A,*D*	1.924	3.705	4.459
A,*E*	0.593	0.351	1.052
B,*C*	1.744	3.043	3.857
B,*D*	0.754	0.569	1.321
B,*E*	1.813	3.29	4.187
C,*D*	1.714	2.939	3.526
C,*E*	1.103	1.217	2.131
D,*E*	1.628	2.651	3.789

Binary Distance Measures.

- **Similarity measures** explain how **alike two observations** are **to each other**, with **high similarity values** representing situations when the **two observations are alike**.
- This contrasts with **distance (or dissimilarity)** measures, where **low values** indicate the **observations are alike.**
- The similarity and distance calculations for binary variables are based on the **number of common and different values in the four situations** summarized in Table 3.8 **for two observations p and q, over all variables**

TABLE 3.8 Four Alternatives (*A*, *B*, *C*, and *D*) for Comparing Two Binary Values

		q_i	
		1	0
p_i	1	A	B
	0	C	D

a: Number of variables where the value for both p and q is one (A)
b: Number of variables where the value for p is one and the value for q is zero (B)
c: Number of variables where the value for p is zero and the value for q is one (C)
d: Number of variables where the value for both p and q is zero (D)

Binary Distance Measures.

- Take **five binary variables**: **urban, technology, services, legal, and female.**
- Using Table 3.9, the values for a, b, c, and d for customers B and C are calculated as:

TABLE 3.9 Five Customers with Data for Five Binary Variables

Customer	Urban	Technology	Services	Legal	Female
A	1	1	0	0	0
B	0	0	1	0	1
C	0	0	0	1	1
D	0	0	1	0	1
E	1	1	0	0	0

a: 1 (since female is both one in observations B and C);
b: 1 (since services is one in observation B and zero in observation C);
c: 1 (since legal is zero in observation B and one in observation C);
d: 2 (since urban and technology are both zero in observations B and C).

Since these **four counts cover all possible situations**, summing a, b, c, and d should **equal the total number of variables** selected, which is five in this example.

Simple matching.

- Calculates **number of common ones or zeros** as **a proportion of all variables**
- Formula is used to calculate the **similarity coefficients ($s_{p,q}$)**

$$s_{p,q} = \frac{a+d}{a+b+c+d}$$

TABLE 3.9 Five Customers with Data for Five Binary Variables

Customer	Urban	Technology	Services	Legal	Female
A	1	1	0	0	0
B	0	0	1	0	1
C	0	0	0	1	1
D	0	0	1	0	1
E	1	1	0	0	0

The corresponding distance calculation is:

$$d_{p,q} = 1 - \frac{a+d}{a+b+c+d}$$

Using Table 3.9, the simple distance between *B* and *C* is: 0.6

$$d_{B,C} = 1 - \frac{1+2}{1+1+1+2} \qquad d_{B,C} = 0.4$$

Jaccard.

- Calculates **the proportion of common ones against the total number of values** that are one in either or both observations.
- Five Customers with Data for **Five Binary Variables Customer, Urban, Technology, Services, Legal** and **Female** the method does **not incorporate d** (the number of values where the variables are both zero).

TABLE 3.9 Five Customers with Data for Five Binary Variables

Customer	Urban	Technology	Services	Legal	Female
A	1	1	0	0	0
B	0	0	1	0	1
C	0	0	0	1	1
D	0	0	1	0	1
E	1	1	0	0	0

Formula is used to calculate the similarity coefficients ($s_{p,q}$):

$$s_{p,q} = \frac{a}{a+b+c}$$

$$S_{B,C} = 0.33$$

The corresponding distance calculation is:

$$d_{p,q} = \frac{b+c}{a+b+c}$$

Using Table 3.9, the simple distance between *B* and *C* is: $d_{B,C} = \frac{1+1}{1+1+1}$ $d_{B,C} = 0.67$


AFTER THE
BREAK.

Why grouping?

- Finding hidden relationships: organize observations in different ways.
  - a data set of retail transactions is grouped and these groups are used to find nontrivial associations, such as customers who purchase doormats often purchase umbrellas at the same time
- Becoming familiar with the data: types of observations are present in the data.
  - a database of medical records will be used to create a general model for predicting a number of medical conditions; data set is characterized by grouping the observations; both male and female patients with a variety of conditions; young female patients having flu.
- Segmentation:Techniques for grouping data may lead to divisions that simplify the data for analysis.

Grouping Approaches.

- Supervised and Unsupervised
- Types of variables
- Data set size limit
- Interpretable and actionable
- Overlapping group/mutually exclusive

Grouping Approaches.

- **Supervised versus unsupervised:**
- **One distinction** between the different methods is whether they u**se the response variable** to **guide how the groups are generated**.
- Methods that **do not use any data to guide** how the groups are generated are called **unsupervised methods,** whereas methods that make use of the response variable to **guide group generation** are called **supervised methods.**
- For example,
  - A data set of cars could be grouped using an unsupervised method. The groups generated would be based on general classes of cars.
  - Alternatively, we could group the cars using car fuel efficiency to direct the grouping.

Good Clustering

- A **good clustering method** will produce high quality clusters with
  - High intra-class similarity (within groups)
  - Low inter-class similarity (between groups)
- Clustering methods ***(unsupervised)*** can be
  - **Hierarchical**
  - **Partitioning**
  - Others (Density based, grid based, model based)

Given dataset, no class label

Hierarchical decomposition
Divisive or bottom-up
CHAMELEON and BIRCH

Centroid based techniques
k-means
k-medoids
K: no. of partitions
PAM, CLARA, CLARANS

1. AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Example of a hierarchical method for grouping observations.
- Uses a **"bottom-up" approach** to clustering as it **starts with each observation** and **progressively creates cluster** by **merging observations** together until all are a **member** of a **final single cluster.**
- The **major limitation** of agglomerative hierarchical clustering is that it is normally **limited to data sets with fewer than 10,000 observatio**ns because the **computational cost** to generate the hierarchical tree can be **high,** especially for larger numbers of observations/expensive.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

TABLE 5.4 Table of Observations to Cluster

	Name	Variable 1	Variable 2	Variable 3	Variable 4	Variable 5
1	A	7.9	8.6	4.4	5.0	2.5
2	B	6.8	8.2	5.2	4.2	2.2
3	C	8.7	9.6	7.5	8.9	9.8
4	D	6.1	7.3	7.9	7.3	8.3
5	E	1.5	2.0	5.1	3.6	4.2
6	F	3.7	4.3	5.4	3.3	5.8
7	G	7.2	8.5	8.6	6.7	6.1
8	H	8.5	9.7	6.3	5.2	5.0
9	I	2.0	3.4	5.8	6.1	5.6
10	J	1.3	2.6	4.2	4.5	2.1
11	K	3.4	2.9	6.5	5.9	7.4
12	L	2.3	5.3	6.2	8.3	9.9
13	M	3.8	5.5	4.6	6.7	3.3
14	N	3.2	5.9	5.2	6.2	3.7

Measured over 5 variables

Measured on same scale

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

1. **Distance** between **all pairs** of observations is calculated.

   Method for calculating the distance along with the variables to include in the calculation should be set **prior to clustering** (***Euclidean distance*** across all continuous variables shown in Table 5.4.)

   Distances between **all combinations of observations** are **summarized** in a **distance matrix**, as illustrated in Table 5.5.

   - Distances between four observations are shown (A, B, C, D) and each value in the table shows the distance between two indexed observations.
   - Diagonal values are excluded, since these pairs are of the same observation. It should be noted that a **distance matrix** is usually **symmetric** about the diagonal as the distance between, for example, A and B is the same as the distance between B and A.

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

TABLE 5.6 Calculated Distances Between All Pairs of Observations

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
A		0.282	1.373	1.2	1.272	0.978	1.106	0.563	1.178	1.189	1.251	1.473	0.757	0.793
B	0.282		1.423	1.147	1.113	0.82	1.025	0.56	1.064	1.065	1.144	1.44	0.724	0.7
C	1.373	1.423		0.582	1.905	1.555	0.709	0.943	1.468	1.995	1.305	1.076	1.416	1.378
D	1.2	1.147	0.582		1.406	1.092	0.403	0.808	0.978	1.543	0.797	0.744	1.065	0.974
E	1.272	1.113	1.905	1.406		0.476	1.518	1.435	0.542	0.383	0.719	1.223	0.797	0.727
F	0.978	0.82	1.555	1.092	0.476		1.191	1.039	0.57	0.706	0.595	1.076	0.727	0.624
G	1.106	1.025	0.709	0.403	1.518	1.191		0.648	1.163	1.624	1.033	1.108	1.148	1.051
H	0.563	0.56	0.943	0.808	1.435	1.039	0.648		1.218	1.475	1.169	1.315	0.984	0.937
I	1.178	1.064	1.468	0.978	0.542	0.57	1.163	1.218		0.659	0.346	0.727	0.553	0.458
J	1.189	1.065	1.995	1.543	0.383	0.706	1.624	1.475	0.659		0.937	1.344	0.665	0.659
K	1.251	1.144	1.305	0.797	0.719	0.595	1.033	1.169	0.346	0.937		0.64	0.774	0.683
L	1.473	1.44	1.076	0.744	1.223	1.076	1.108	1.315	0.727	1.344	0.64		0.985	0.919
M	0.757	0.724	1.416	1.065	0.797	0.727	1.148	0.984	0.553	0.665	0.774	0.985		0.196
N	0.793	0.7	1.378	0.974	0.727	0.624	1.051	0.937	0.458	0.659	0.683	0.919	0.196	

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

Two closest observations are identified **(M and N)** and are merged into a **single cluster.**

These **two observations** from now on will be considered a **single group**

TABLE 5.5 Distance Matrix Format

	A	B	C	D	...
A		$d_{A,B}$	$d_{A,C}$	$d_{A,D}$	...
B	$d_{B,A}$		$d_{B,C}$	$d_{B,D}$	...
C	$d_{C,A}$	$d_{C,B}$		$d_{C,D}$	...
D	$d_{D,A}$	$d_{D,B}$	$d_{D,C}$		...
...	...	...	...	...	

2. **All** observations (minus M & N merged into a cluster) along with the newly created cluster are compared to **see which observation or cluster** should be joined into the next cluster.

Analyse individual observations and clusters**, so, a joining or linkage rule** is needed to determine the distance between an observation and a cluster of observations. This joining/linkage rule should be set prior to clustering.

JOINING OR LINKAGE RULE.

Average linkage	Single linkage	Complete linkage
Distance between **all members of cluster** (a, b, and c) and **observation** under consideration (X) are determined.	**Distance between all members of cluster** (a, b, and c) and **observation** under consideration (X) are determined.	**Distance between all members of cluster** (a, b, and c) and **observation** under consideration X) are determined.
Average is calculated.	**Smallest is selected.**	**Highest is selected**

STEPS for AGGLOMERATIVE HIERARCHICAL CLUSTERING

- **Distances** between **all combinations of groups** and **observations** are considered and the **smallest distance** is selected.
- Consider **distance between two clusters**, the **linkage/joining concept is extended to the joining of two clusters**, as illustrated in Figure 5.10.

7. **Process of assessing all pairs of observations/clusters, then combining the pair with the smallest distance is repeated until there are no more clusters or observations to join together since only a single cluster remains**.

- Figure 5.11 illustrates this process for some steps based on the observations shown in Table 5.6

Linkage rules for considering an observation and a cluster & two clusters.

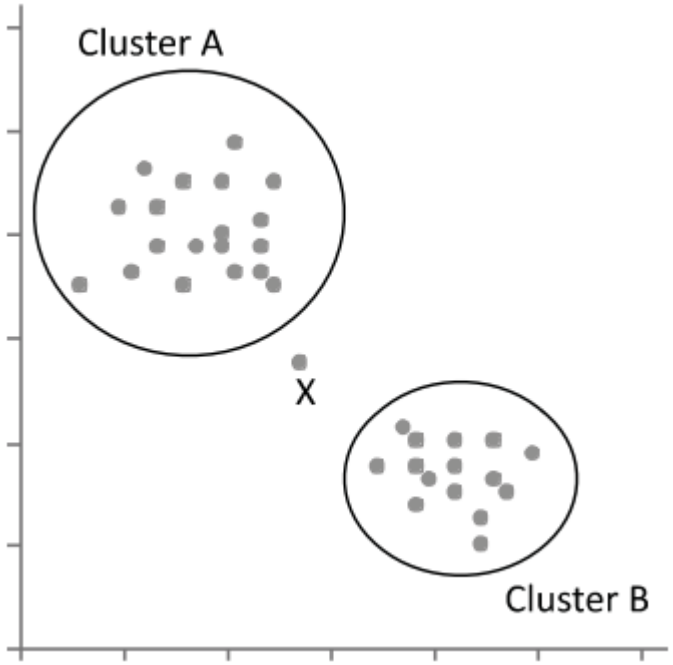


FIGURE 5.8 Comparing observation X with two clusters A and B.

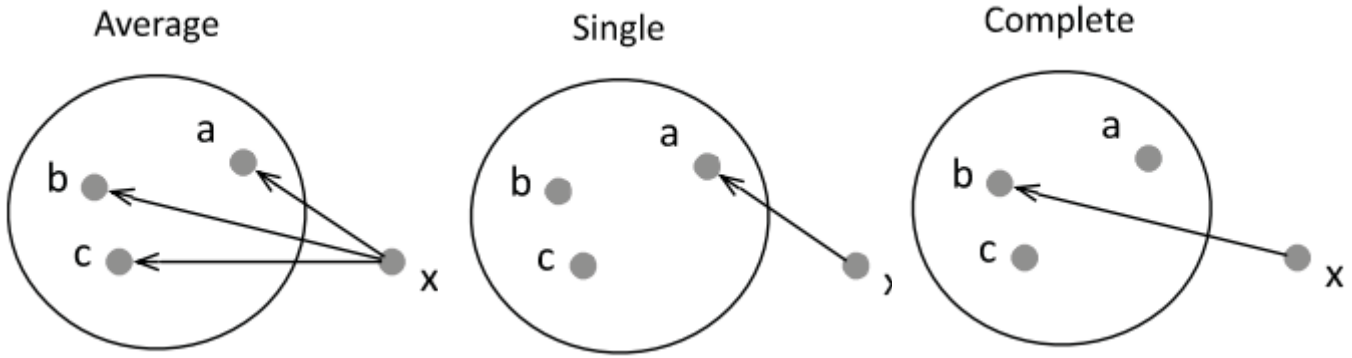


FIGURE 5.9 Different linkage rules for considering an observation and a cluster.

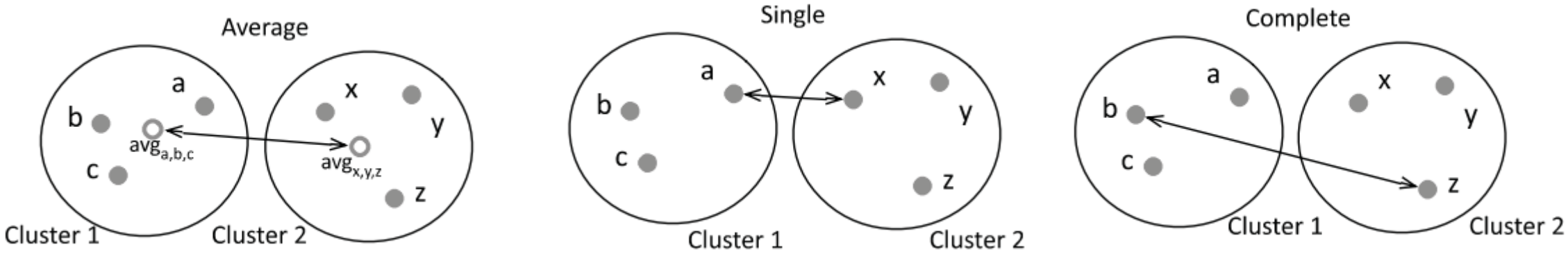


FIGURE 5.10 Different linkage rules for considering two clusters.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

Horizontal length of the lines joining M and N reflects **distance at which cluster** was formed

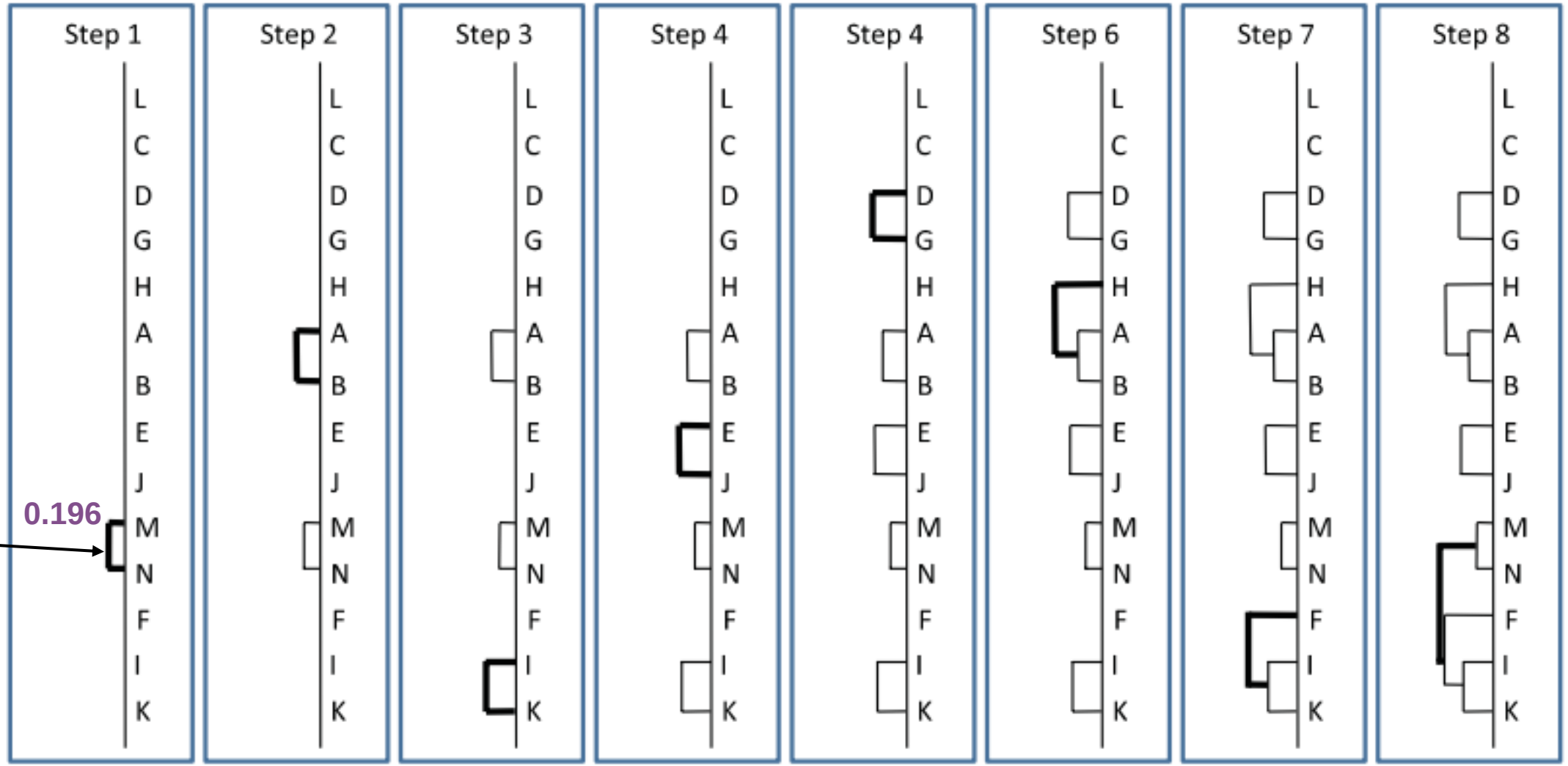


FIGURE 5.11 Steps 1 through 8 of the clustering process.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

- Figure 5.12 shows **completed hierarchical clustering** for all 14 observations.
- When clustering completes, **a tree called a dendrogram** is generated showing **similarity between observations and clusters.**

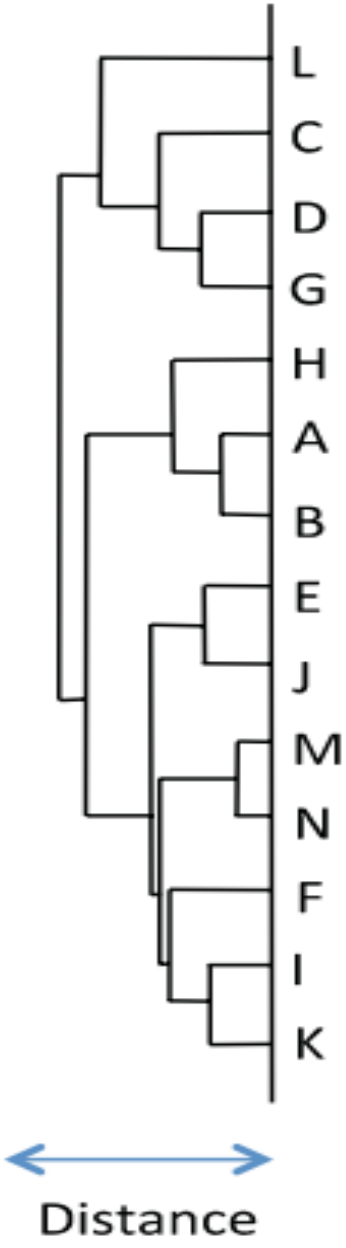


FIGURE 5.12 Completed hierarchical clustering for the 14 observations.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

- To **divide a data set into a series of distinct clusters**, **select a distance** at which the clusters are to be created.
- **Where this distance intersects with a horizontal line on the tree** (shown by the circles)**, a cluster is formed.**
- Three different distances (i, j, k) are used to divide the tree into clusters:
- Where this vertical line intersects with the tree **at distance i: 2** clusters formed: {L,C,D,G} and {H,A,B,E,J,M,N,F,I,K};
- **At distance j: 4** clusters are formed: {L}, {C,D,G}, {H,A,B}, and {E,J,M,N,F,I,K};
- **At distance k: 9** clusters are formed: {L}, {C}, {D,G}, {H}, {A,B}, {E,J}, {M,N}, {F}, and {I,K}.

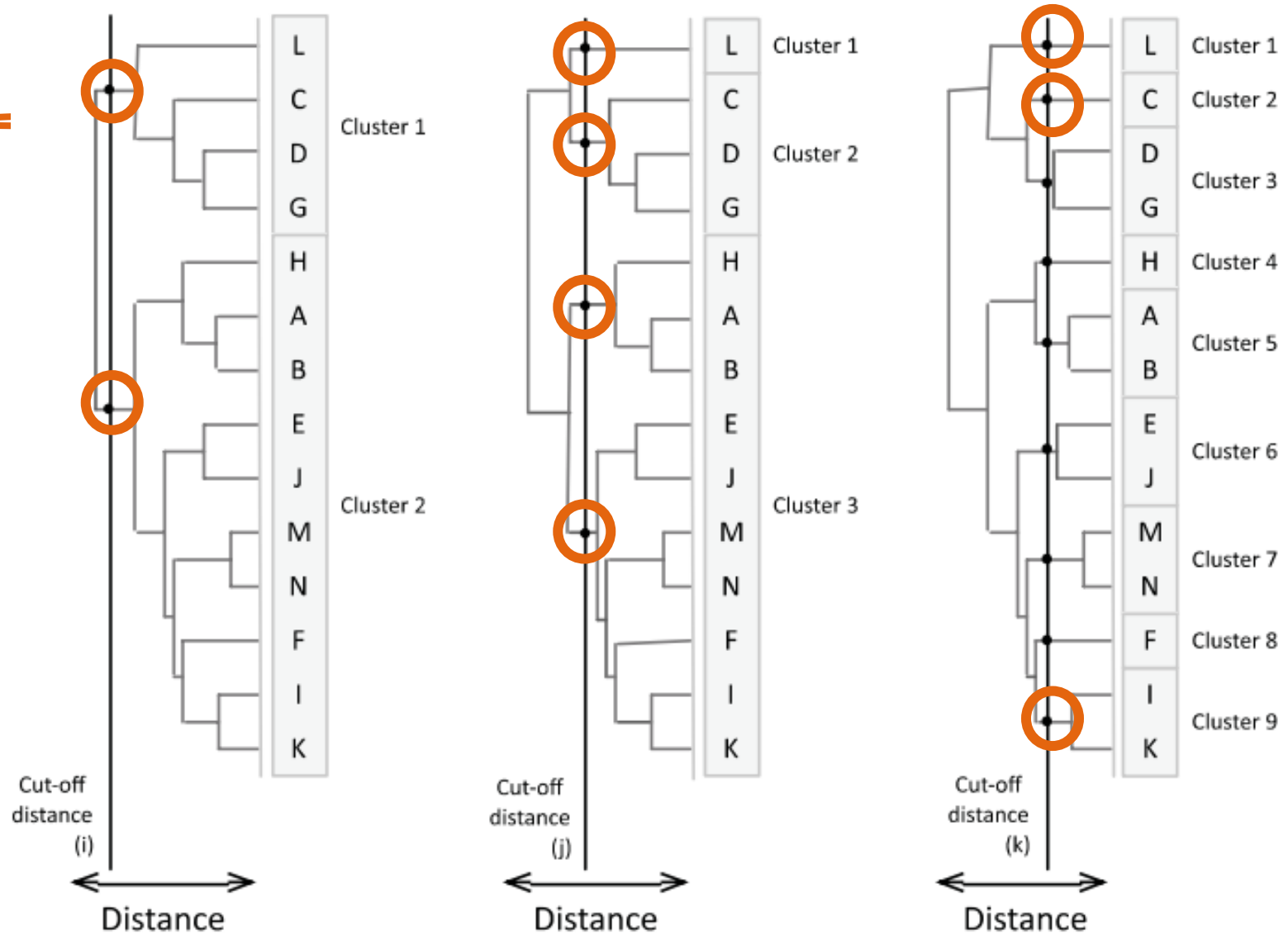


FIGURE 5.13 Cluster generation using three distance cut-offs.

AGGLOMERATIVE HIERARCHICAL CLUSTERING LINKAGE RULE choice

- **Different joining/linkage rules change** how the final hierarchical clustering is presented.
- Figure 5.14 shows the hierarchical clustering of the same set of observations using the average linkage, single linkage, and complete linkage rules.
- Since the **barrier for merging observations** and clusters is **lowest** with **single linkage approach**, the clustering dendrogram may contain **chains of clusters as well as clusters that are spread out.**
- **Barrier** to joining clusters is **highest** with **complete linkage**; it is possible that an observation is closer to observations in other clusters than the cluster to which it has been assigned.

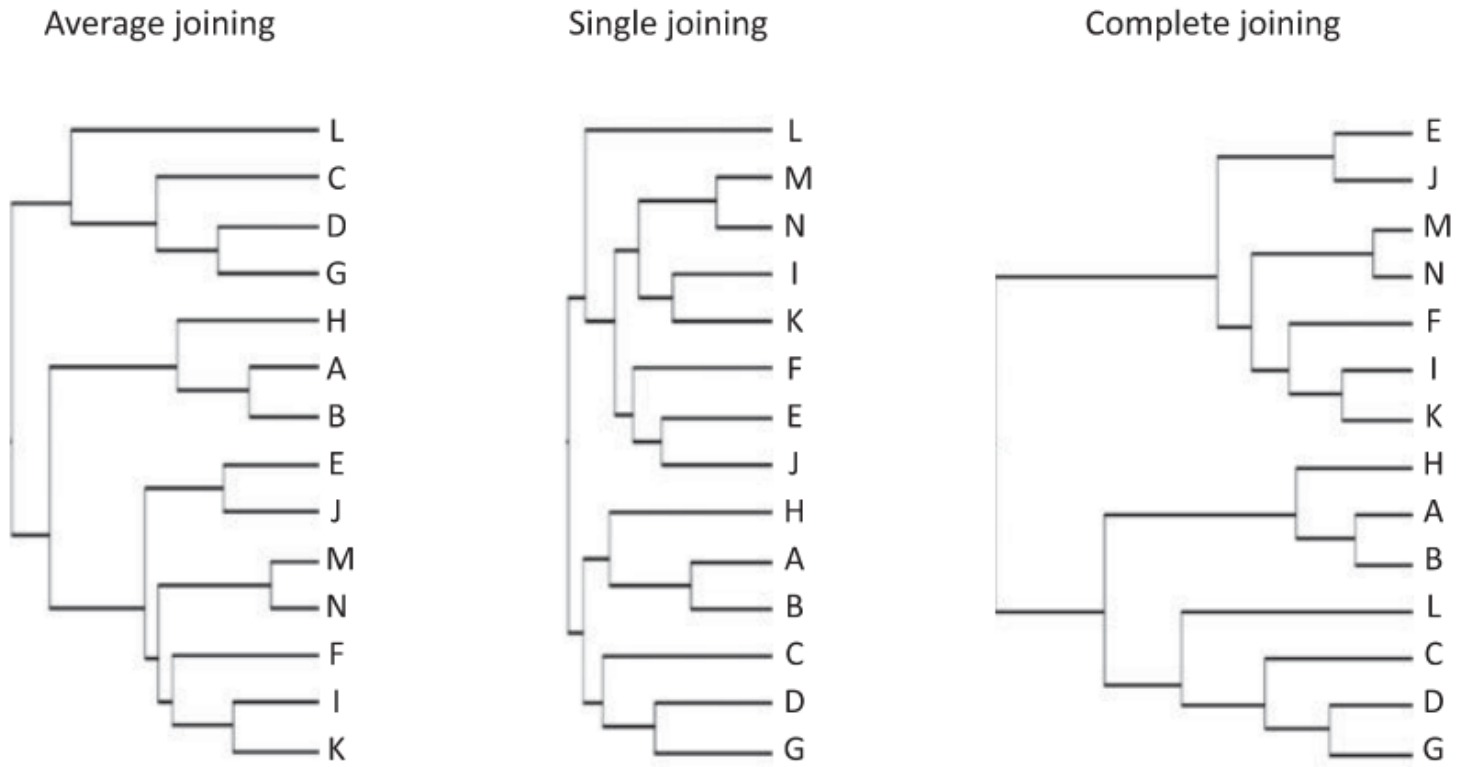


FIGURE 5.14 Different results using three different methods for joining clusters.

The **average linkage approach moderates** the tendencies of the single or complete linkage approaches

Example: AGGLOMERATIVE HIERARCHICAL CLUSTERING

Data set of 392 cars explored using hierarchical agglomerative clustering.

TABLE 5.7 Data Table Containing Automobile Observations

Car Name	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Chevrolet Chevelle Malibu	18	8	307	130	3,504	12	70	American
Buick Skylark 320	15	8	350	165	3,693	11.5	70	American
Plymouth Satellite	18	8	318	150	3,436	11	70	American
Amc rebel sst	16	8	304	150	3,433	12	70	American
Ford Torino	17	8	302	140	3,449	10.5	70	American
Ford Galaxie 500	15	8	429	198	4,341	10	70	American

- This data set was **clustered using the Euclidean distance method** and the **complete linkage rule.**
- The following variables were used in the clustering: **Displacement, Horsepower, Acceleration, and MPG (miles per gallon)**

Example: AGGLOMERATIVE HIERARCHICAL CLUSTERING

TABLE 5.7 Data Table Containing Automobile Observations

Car Name	MPG	Cylinders	Displacement	Horsepower	Weight	Acceleration	Model Year	Origin
Chevrolet Chevelle Malibu	18	8	307	130	3,504	12	70	American
Buick Skylark 320	15	8	350	165	3,693	11.5	70	American
Plymouth Satellite	18	8	318	150	3,436	11	70	American
Amc rebel sst	16	8	304	150	3,433	12	70	American
Ford Torino	17	8	302	140	3,449	10.5	70	American
Ford Galaxie 500	15	8	429	198	4,341	10	70	American

The dendrogram of the generated clusters shows relationships between observations based on similarity of the four selected variables. Each horizontal line at the right represents a single automobile and the order of the observations is related to how similar each car is to its neighbours

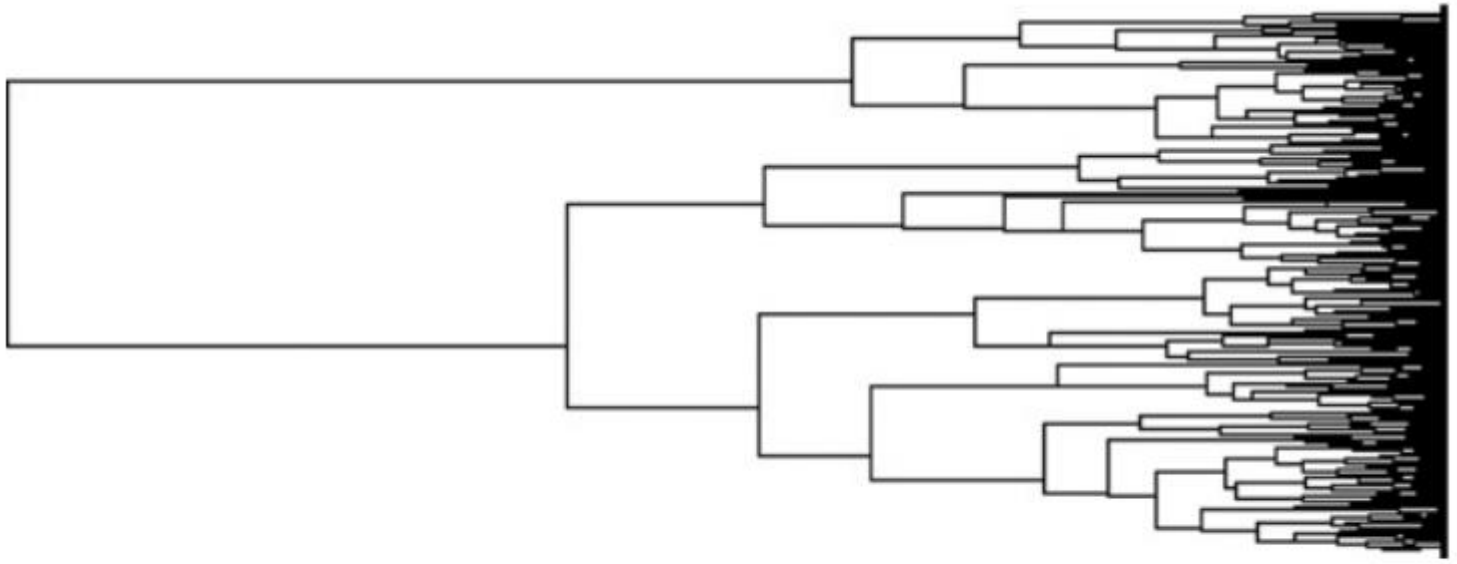
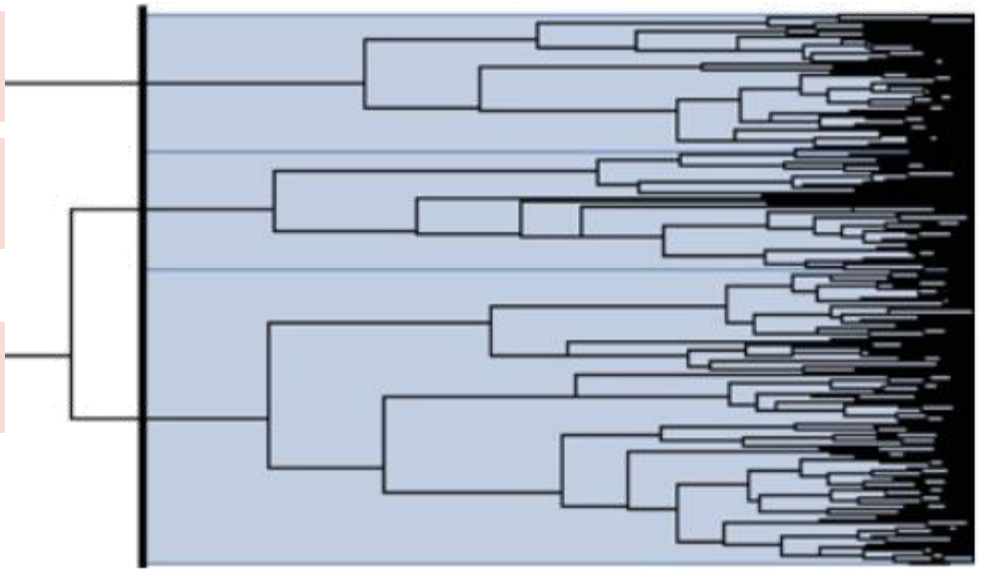


FIGURE 5.15 Hierarchical agglomerative clustering dendrogram generated for the automobile data set.

low fuel efficiency, low acceleration values, higher values for horsepower and displacement.

good fuel efficiency and acceleration as well as low horsepower and displacement

Average fuel efficiency and acceleration as well as few high values for displacement or horsepower



Cluster 1 (97 observations)

Cluster 2 (85 observations)

Cluster 3 (210 observations)

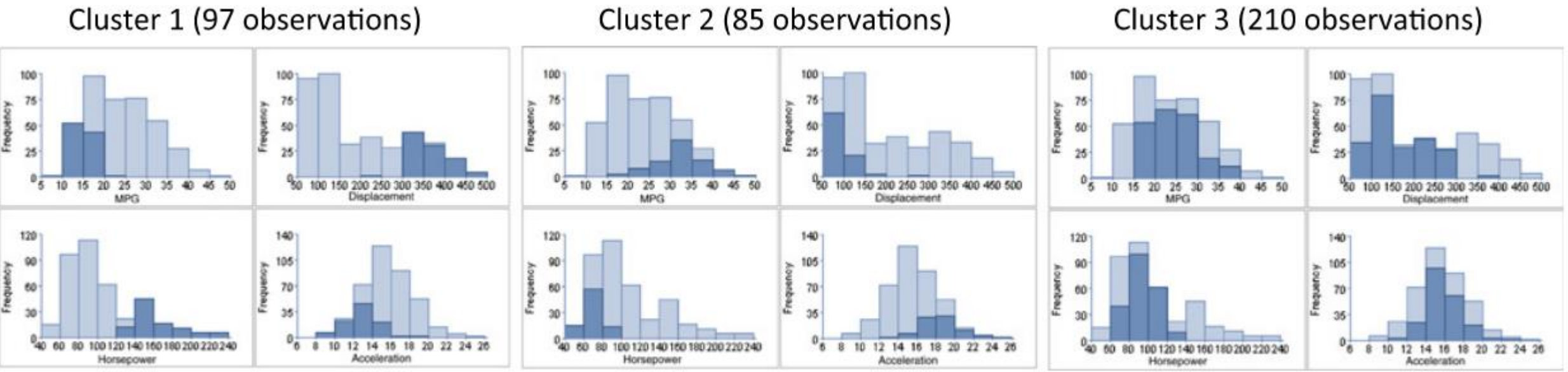


FIGURE 5.16 Automobile data set with three clusters identified.

distance was set to create nine clusters

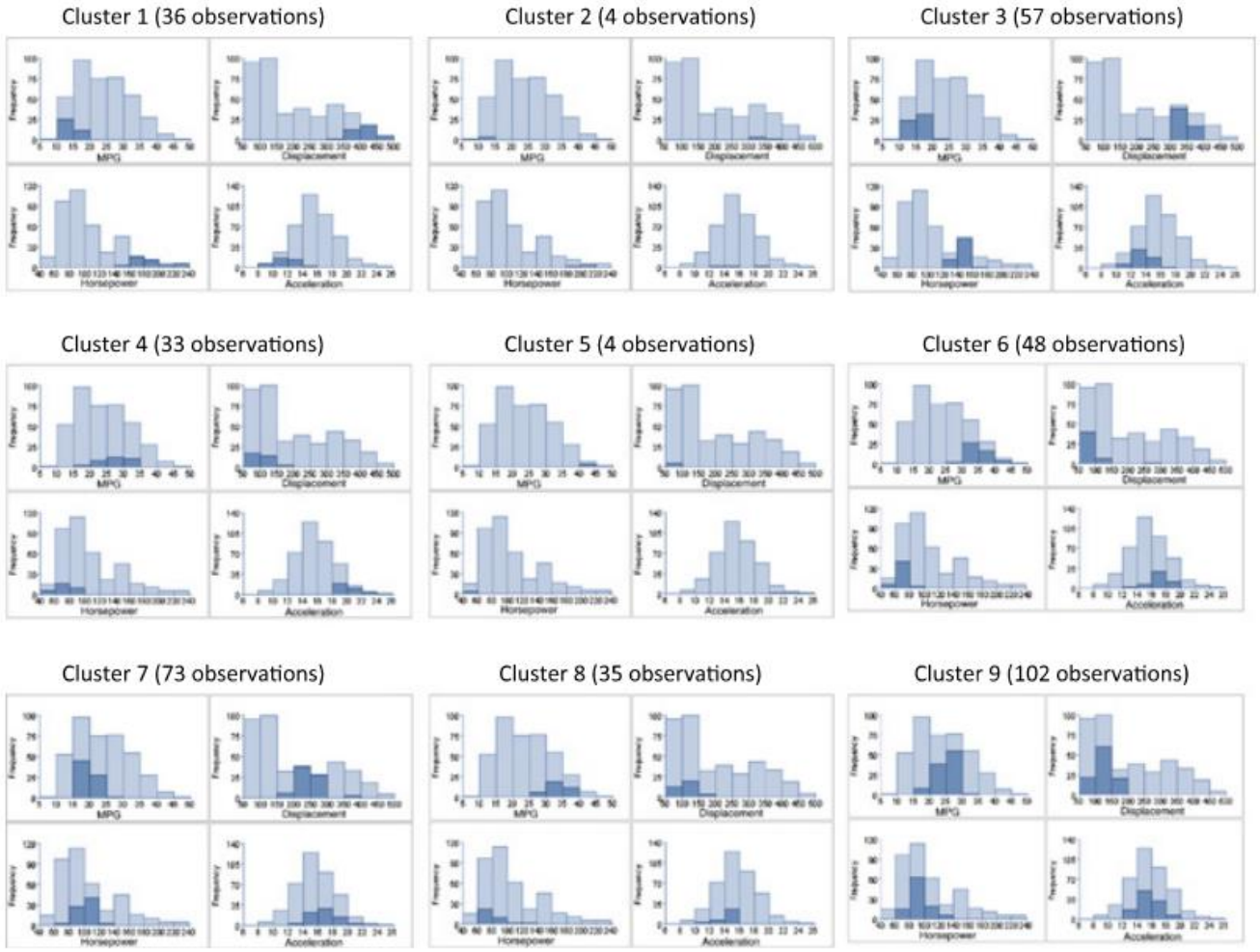


FIGURE 5.17 Automobile data set cluster and split into nine groups.

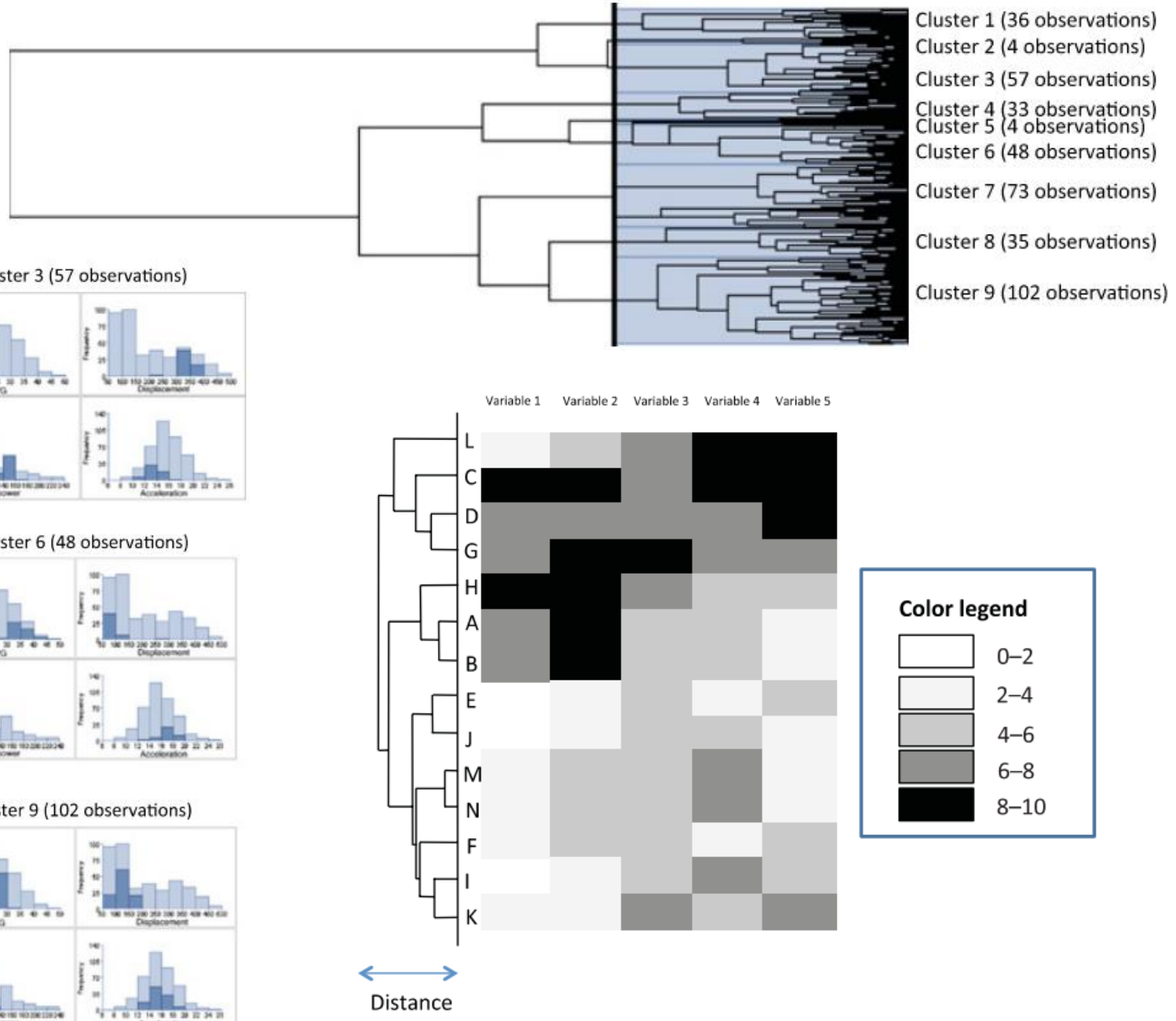


FIGURE 5.18 Clustering dendrogram coupled with a colored heatmap.

Enough for today.

Next: K-means clustering