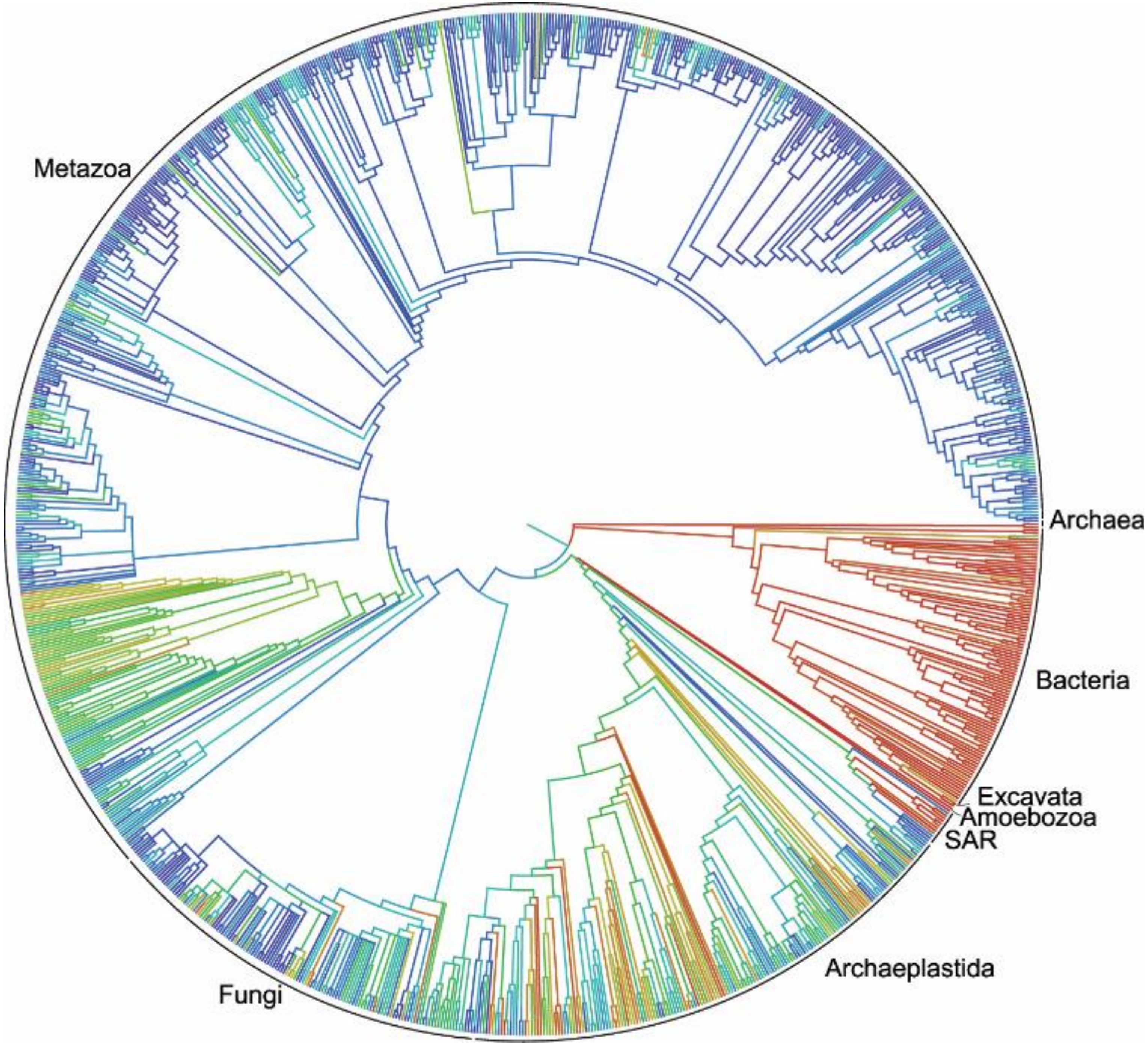
Hierarchical Clustering



[https://phys.org/news/2015-09-tree-life-million-species.html]

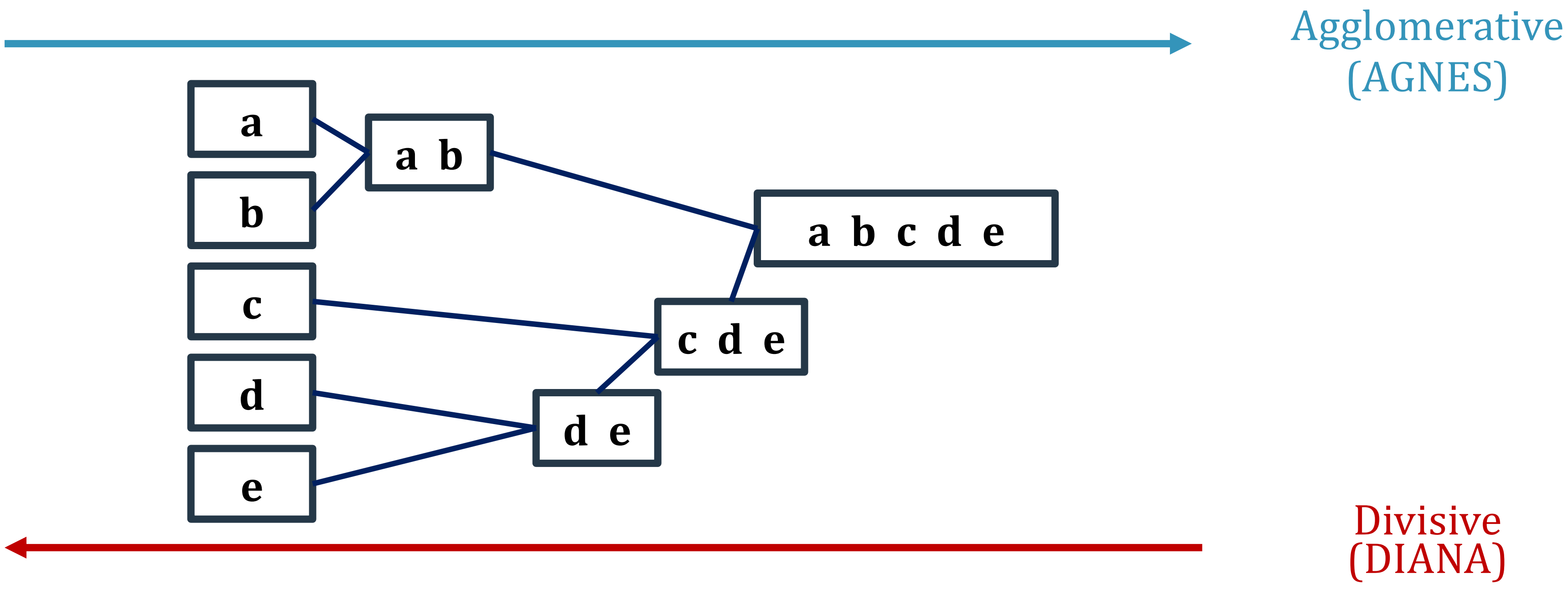
Hierarchical Clustering Overview

Hierarchical clustering (HC) clusters a dataset into a **hierarchy** of clusters (order relation is set containment).

There are two main strategies:

- **Bottom-up** (agglomerative)
  initially, each observation starts in its own **separate** cluster
  clusters are **merged** as the hierarchy is climbed
  after the last merge, all observations are in the **same** cluster
- **Top-down** (divisive)
  initially, each observation starts in the **same** cluster
  clusters are **split** as the hierarchy is descended down
  after the last split, each observation ends in its own **separate** cluster

Bottom-up HC is significantly faster than Top-down HC (poly. vs. exp.)


Agglomerative
(AGNES)
a
b
c
d
e
a b
d e
c d e
a b c d e
Divisive
(DIANA)

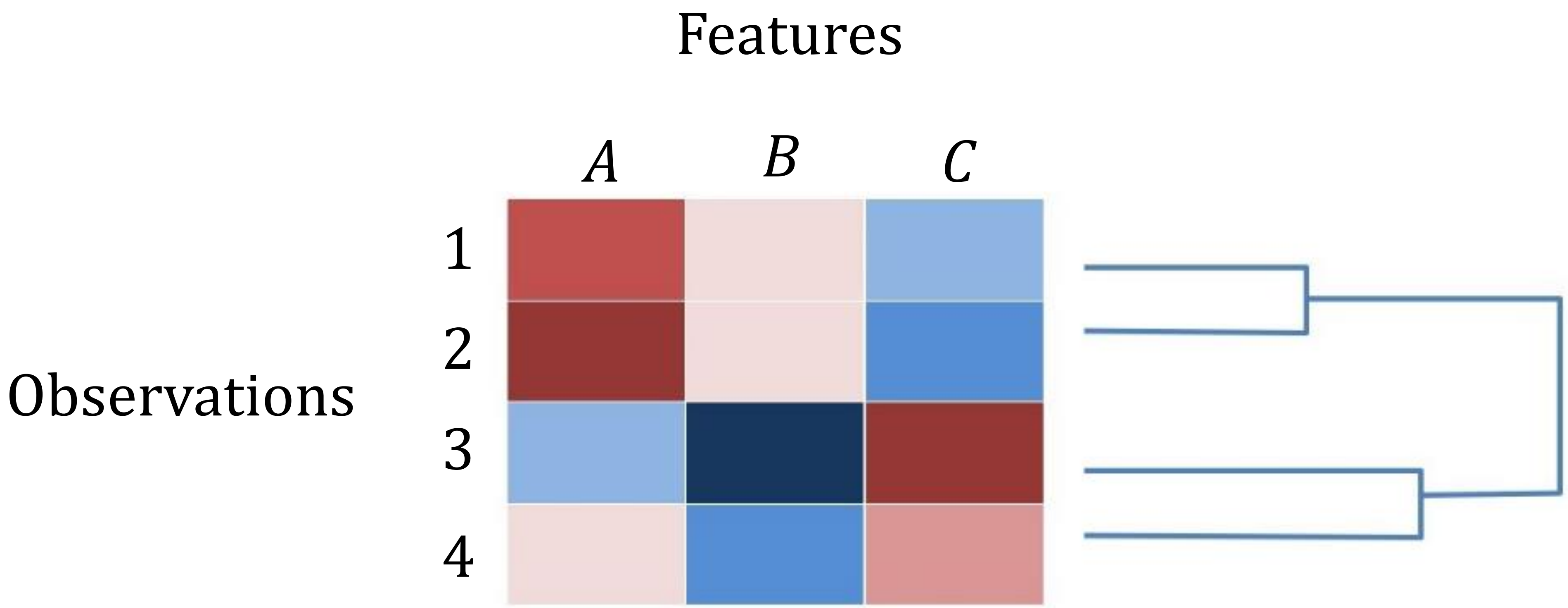
Hierarchical Clustering Overview

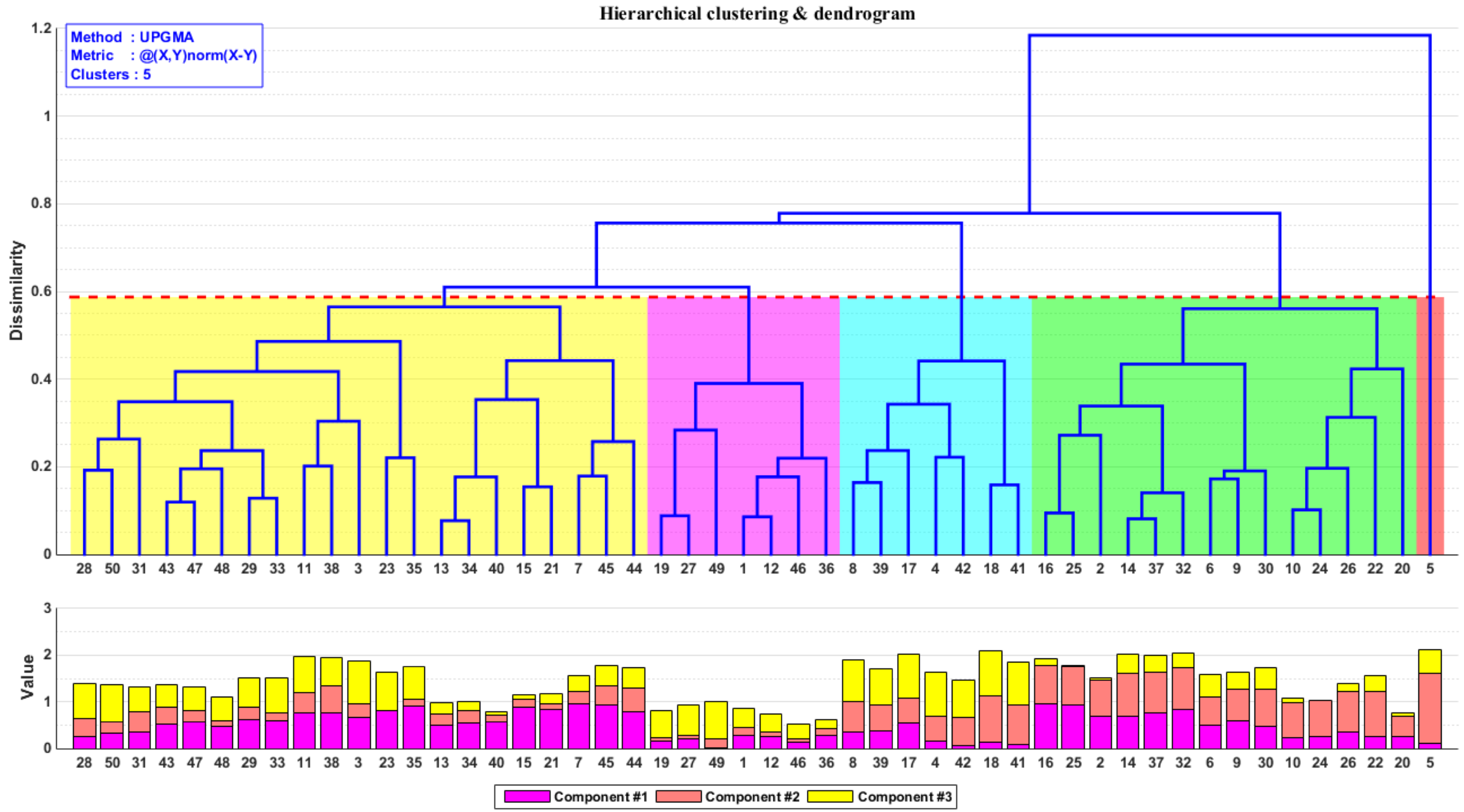
The main question: how to split, or how to merge, clusters?

This requires the notion of a distance between clusters (**linkage**).

- in Bottom-up HC, **nearest pairs** of clusters are merged up the hierarchy (requires only computing distances between pairs)
- in Top-down HC, a cluster must be **optimally split** into sub-clusters down the hierarchy (much harder, computationally)

Another issue: at what level do we **report** the clustering scheme? When do we stop climbing or descending the hierarchy?


Features
A
B
C
Observations
1
2
3
4



[author unknown]

Back to Knowledge Discovery

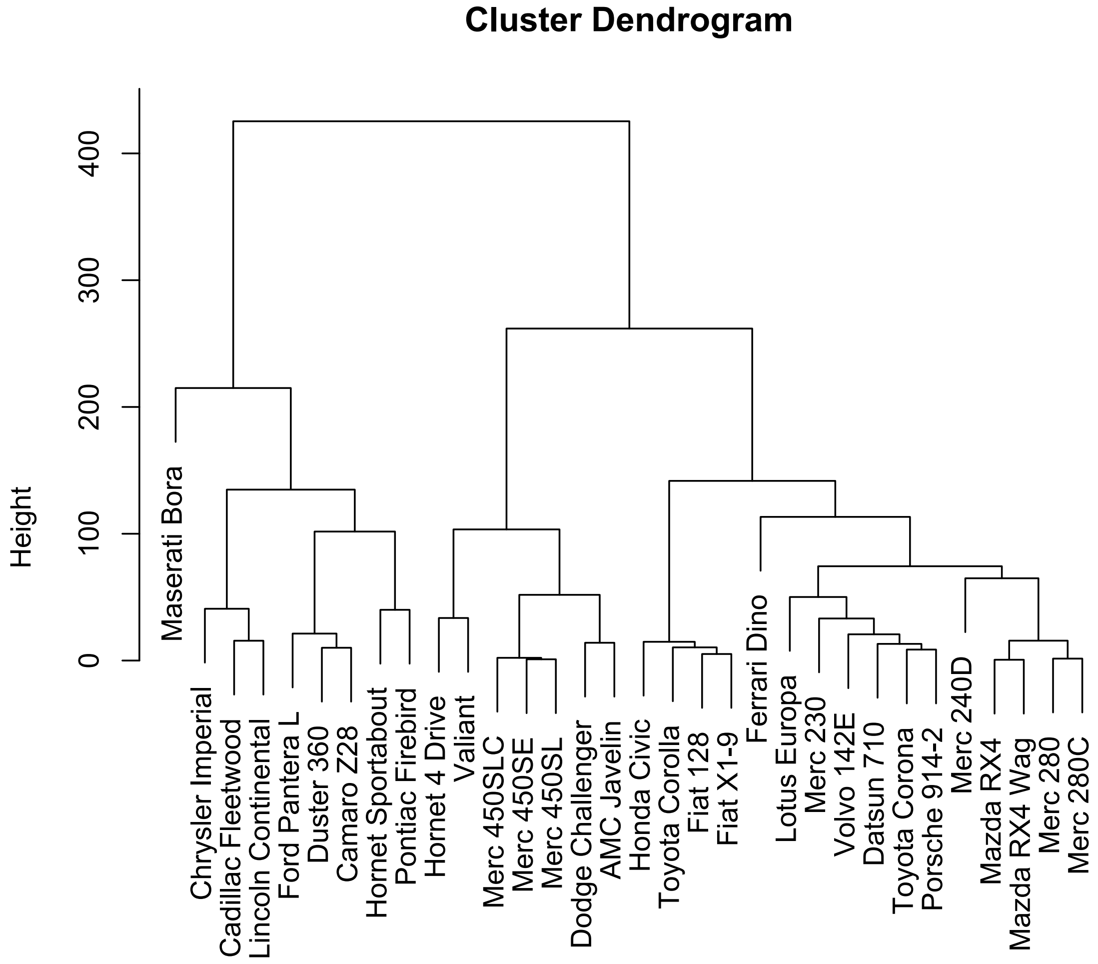
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1

More unsupervised learning: what underlying structures can we discover in this data?

[mtcars dataset]

Back to Knowledge Discovery

What do you notice in this diagram, structure-wise?

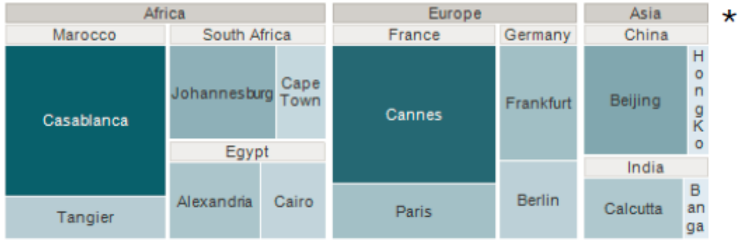


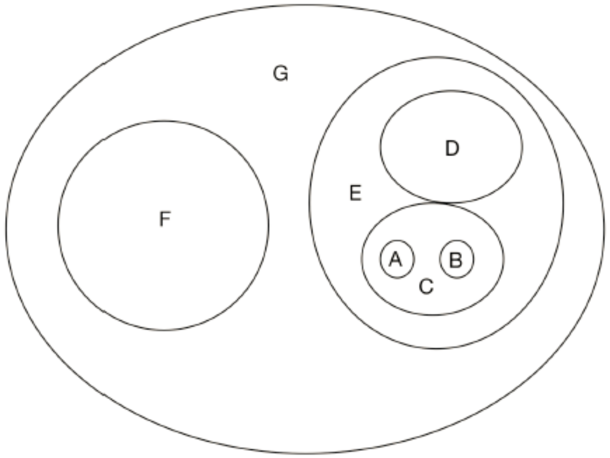
In a nutshell, **hierarchical systems** are ordered sets where elements and/or subsets are organized in a given relationship to one another, both among themselves and within the whole.

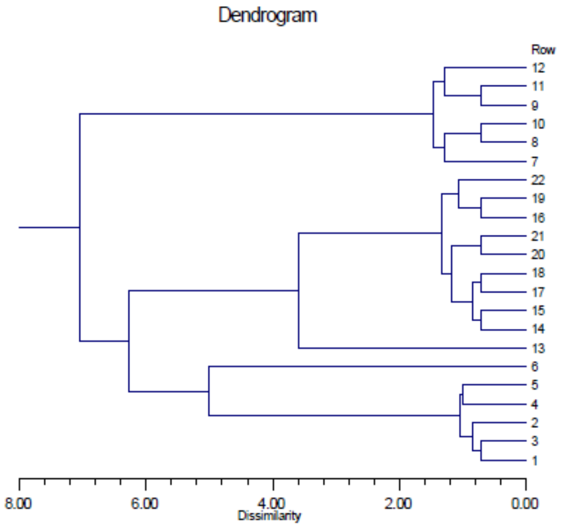
Relationships vary according to the field domain and type of system, but in general, we can describe them by the properties of elements and the laws that govern them (e.g., how they are shared and/or related).

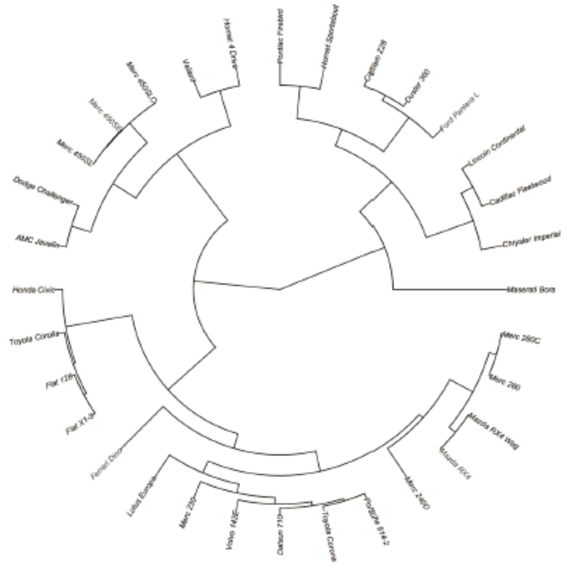
– I. Meireilles, *Design for Information*

Visualizing Hierarchy





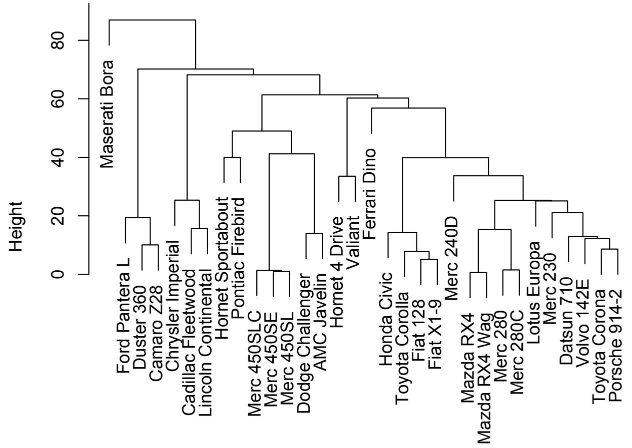


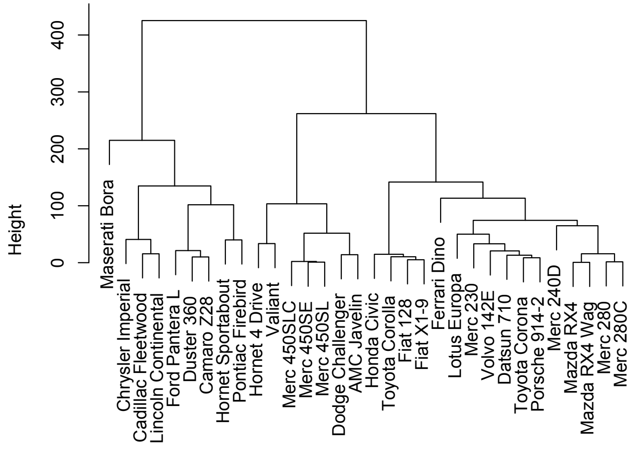


(mammals(primates (apes(orangutan,human)),(monkeys(pygmy marmoset)),(lemurs(ruffed lemur))),(cetacea (whales(long-finned pilot whale,southern right whale)),(dolphins(striped dolphin, bottle-nose dolphin))))

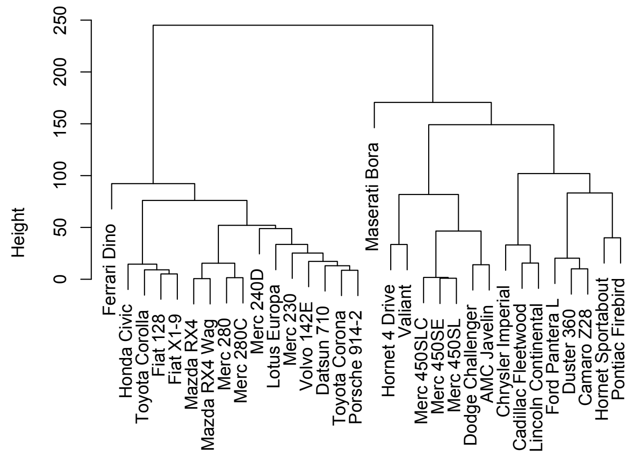
(*treemap from: TIBCO Spotfire Documentation: What is a Treemap?)

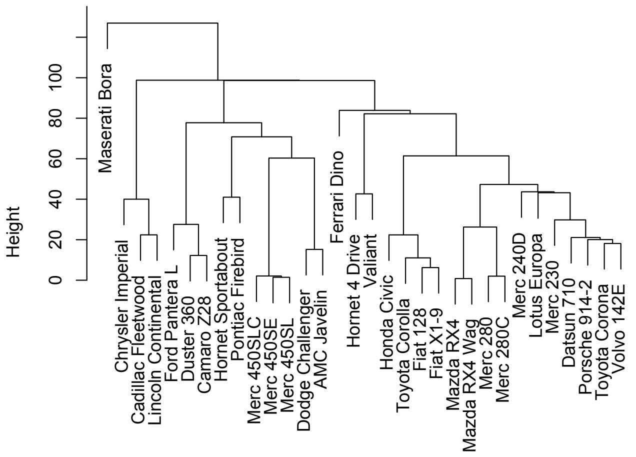
Button Press Hierarchical Clustering





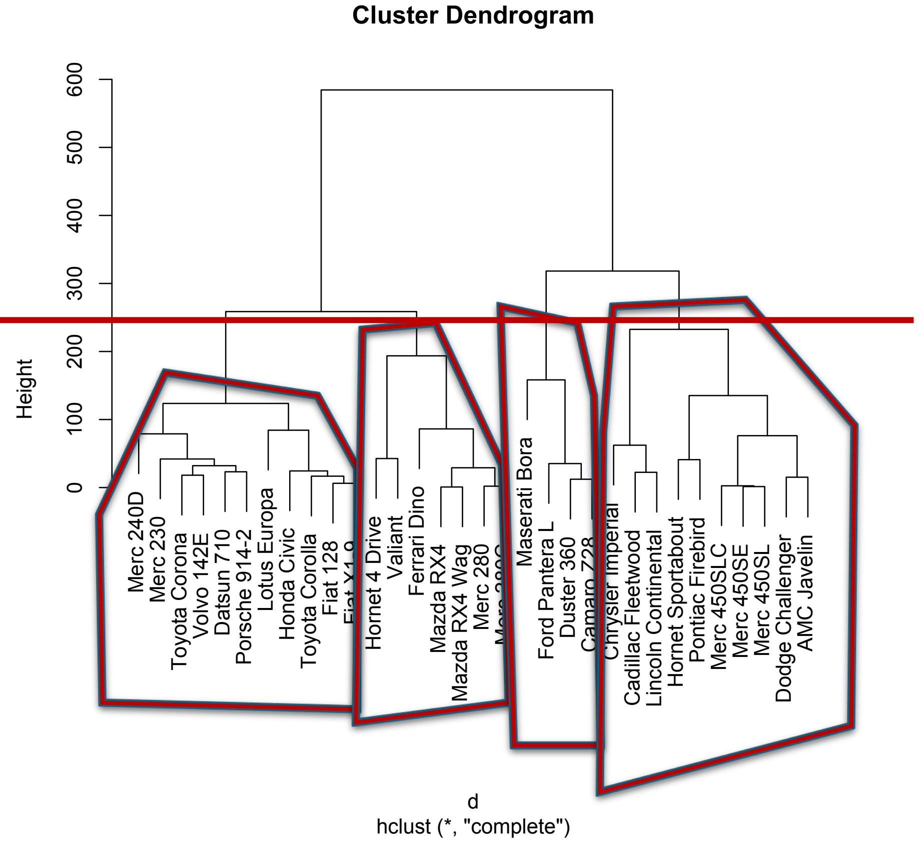
What can we say about the data structure if presented with a **cluster dendrogram**?

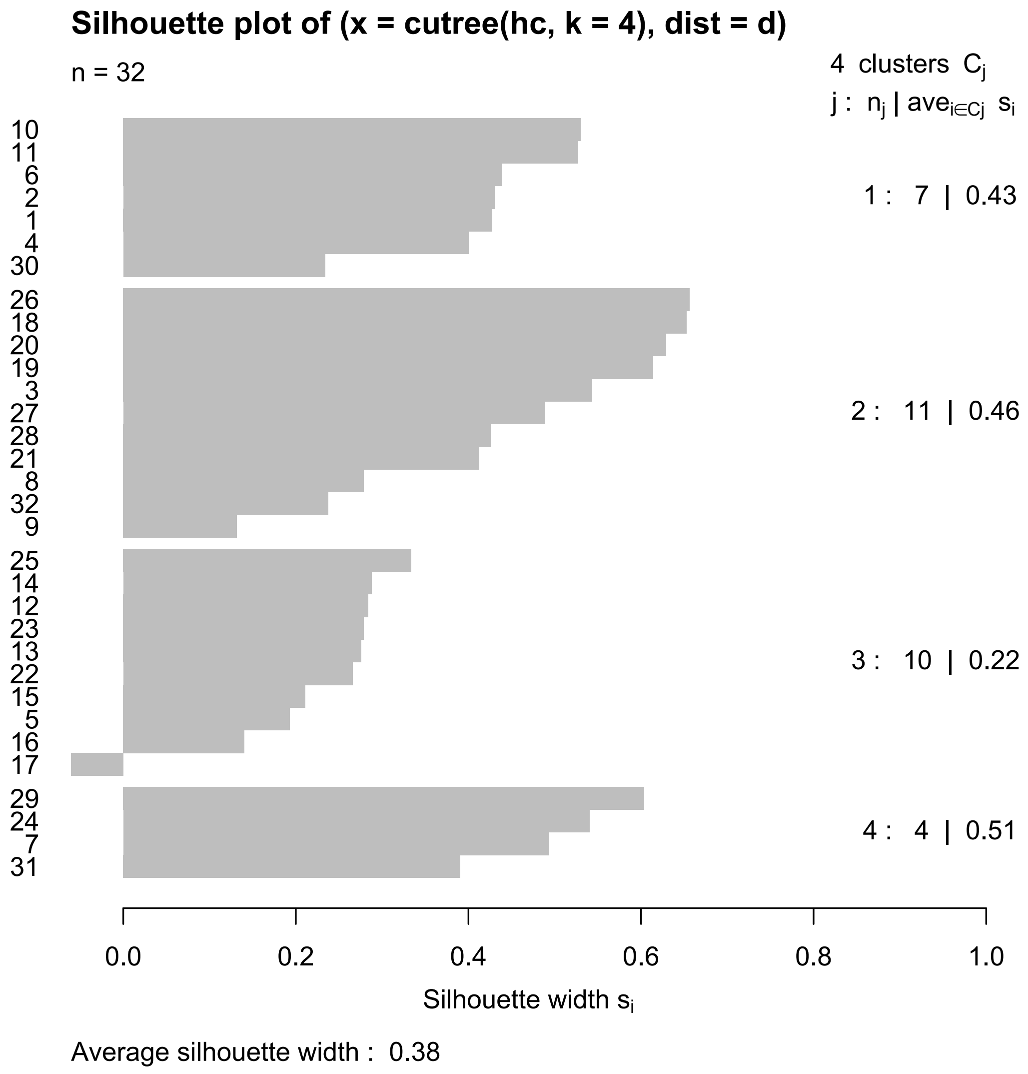




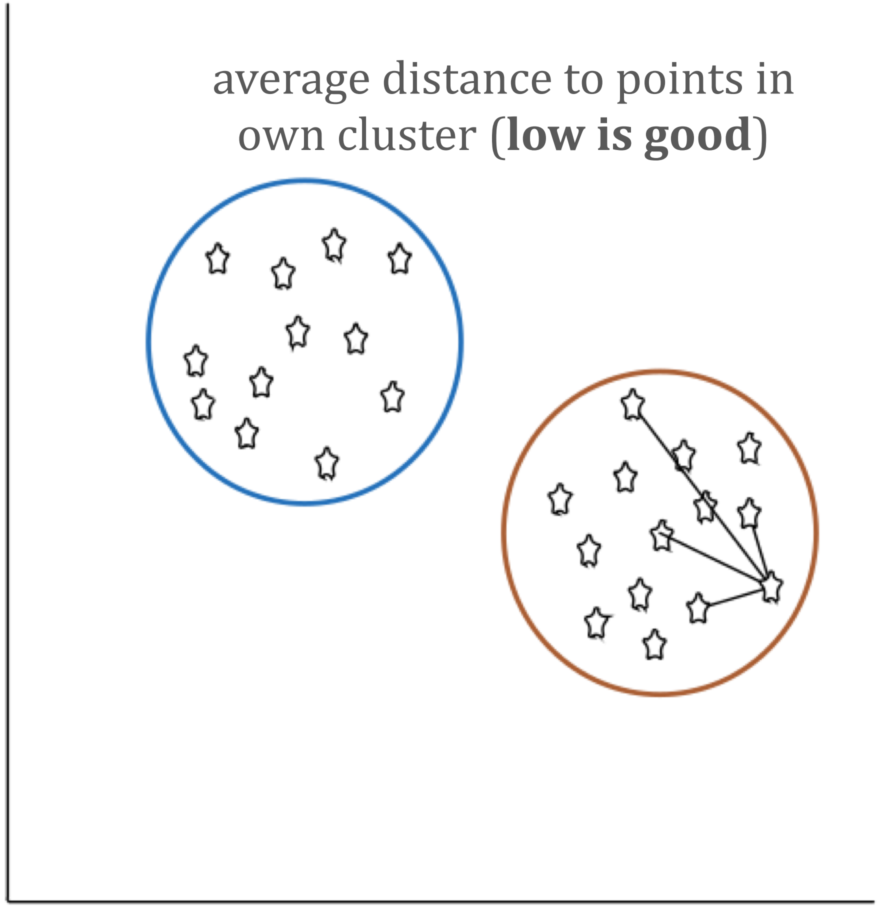
Same data clustered using different parameter settings: which one is **optimal**?

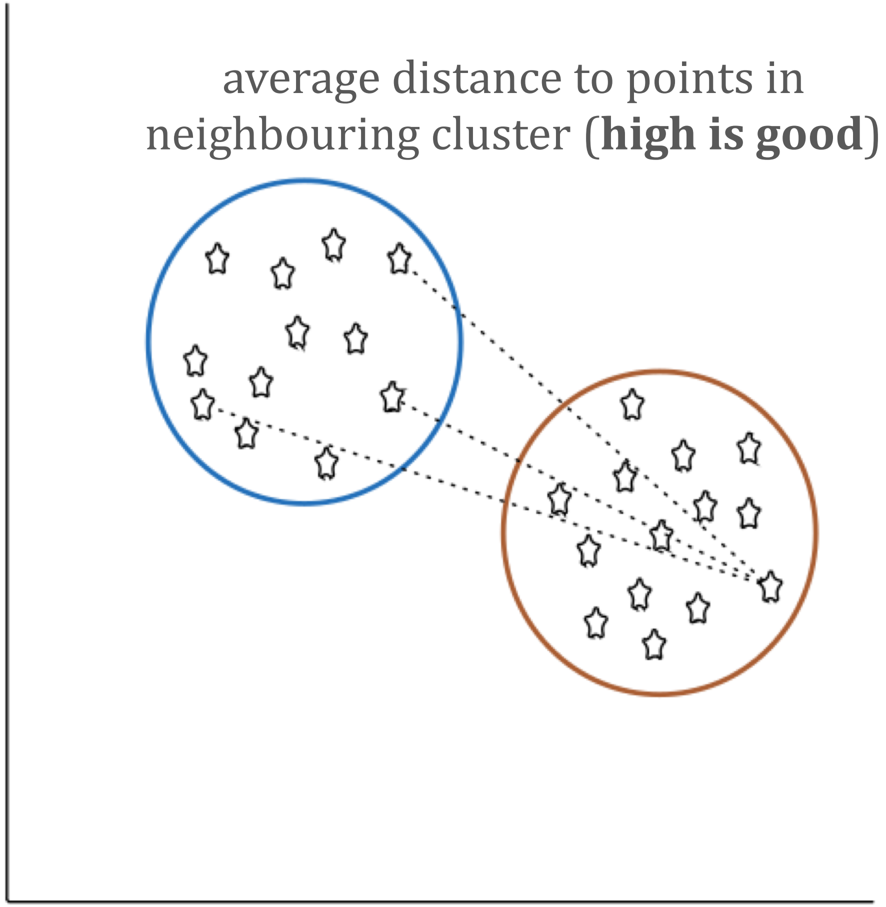
Evaluating the Results





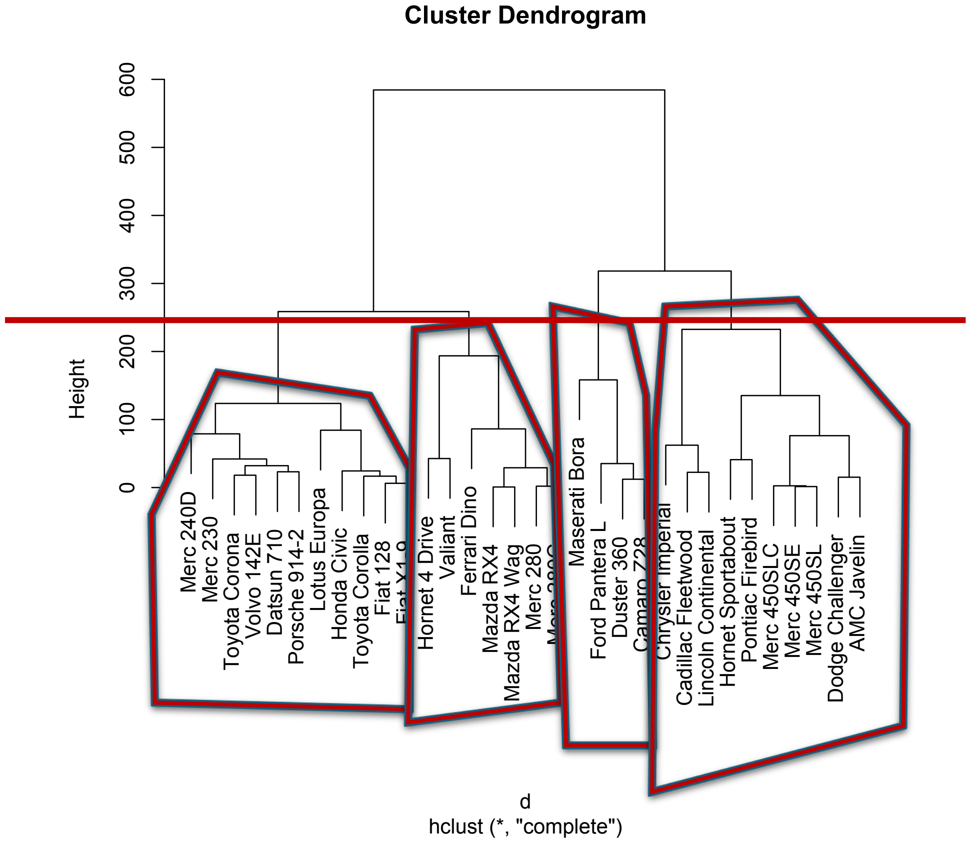
The Silhouette Metric

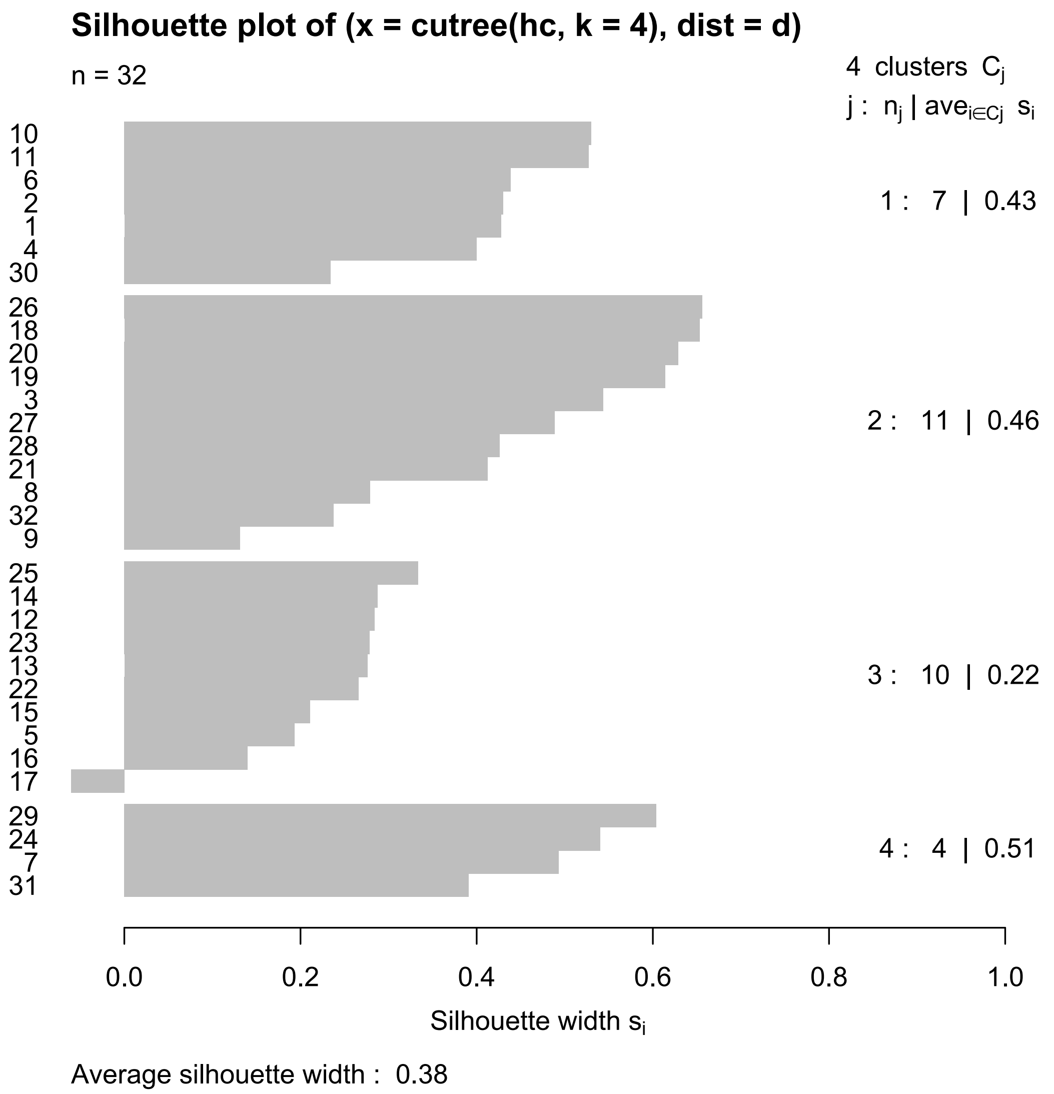




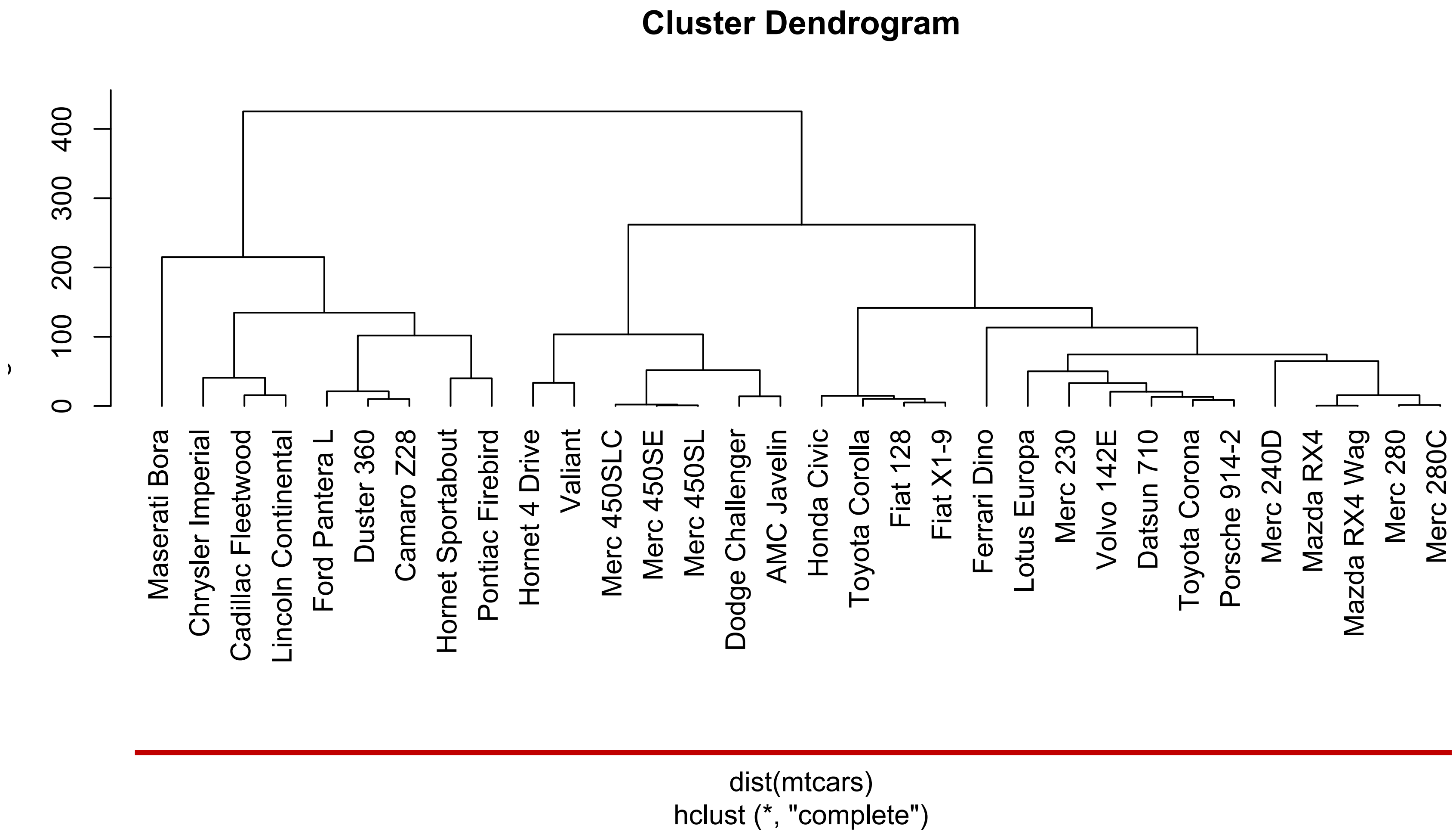
$$\text{silhouette metric for a point} = \frac{\text{average dissimilarity with neighbouring cluster} - \text{average dissimilaity with own cluster}}{\text{maximum dissimilarity value (own or neighbour)}}$$

Evaluating the Results

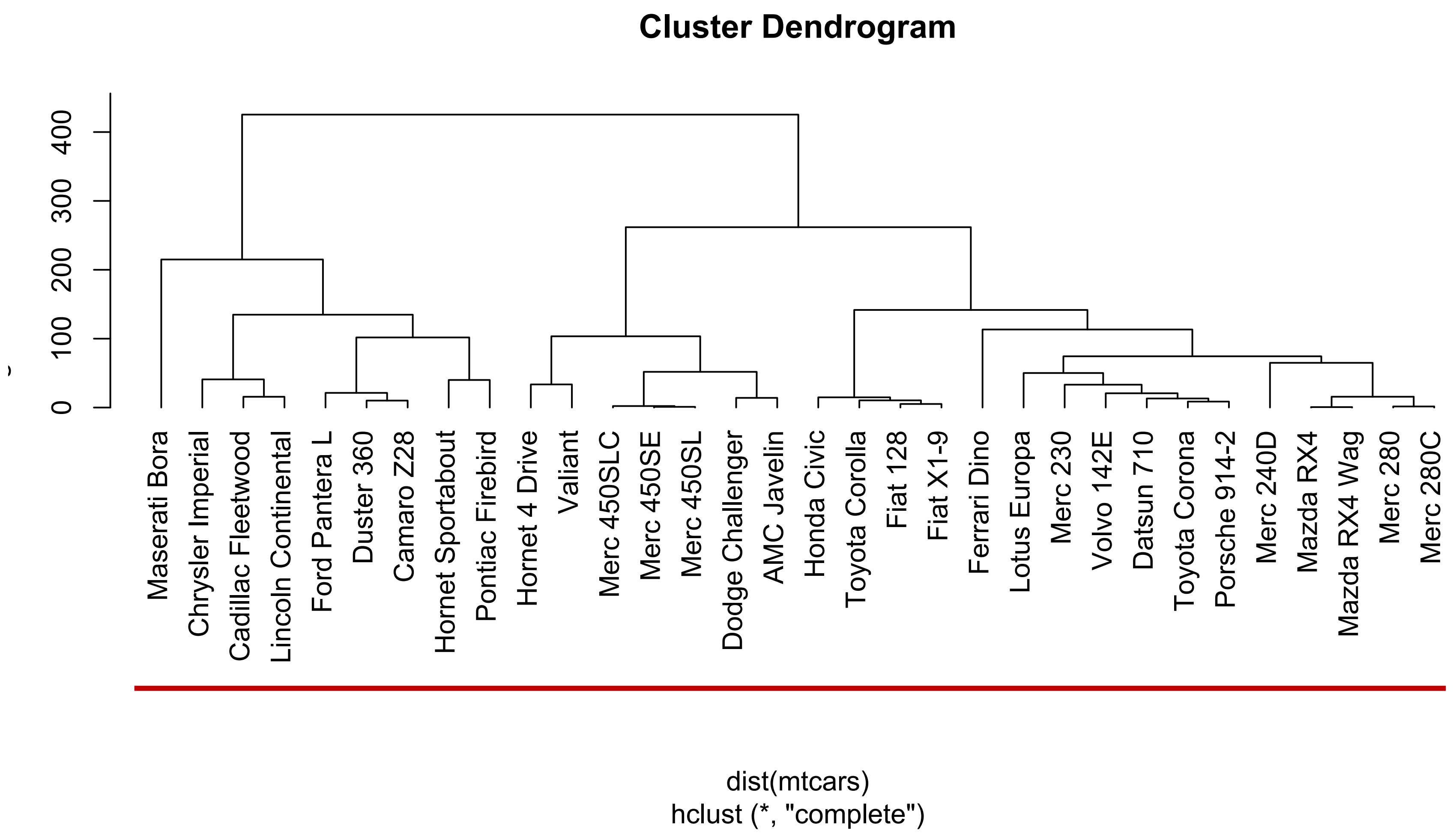




Hierarchical Clustering Algorithm
An Illustration with mtcars

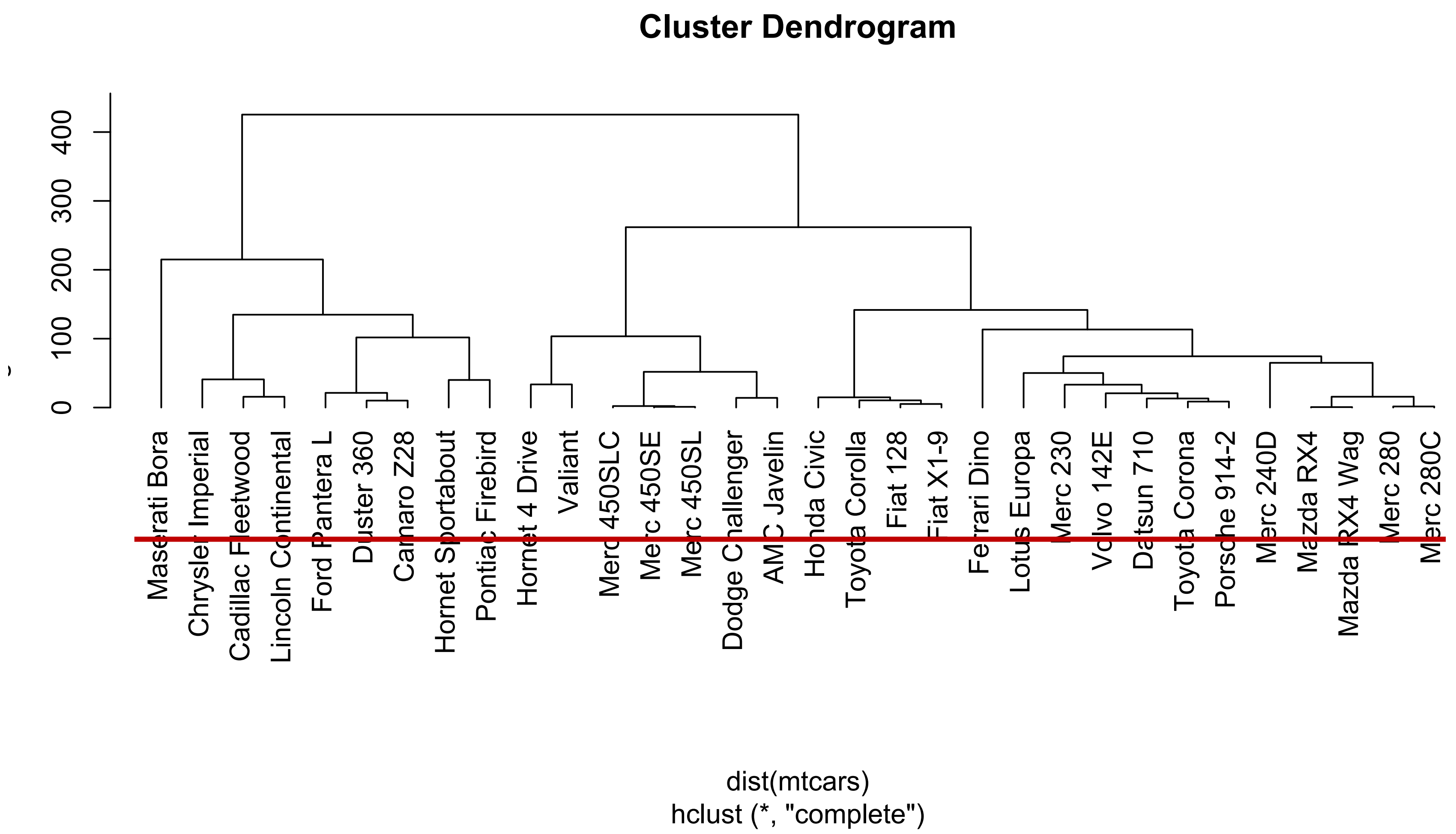


Hierarchical Clustering Algorithm
An Illustration with mtcars

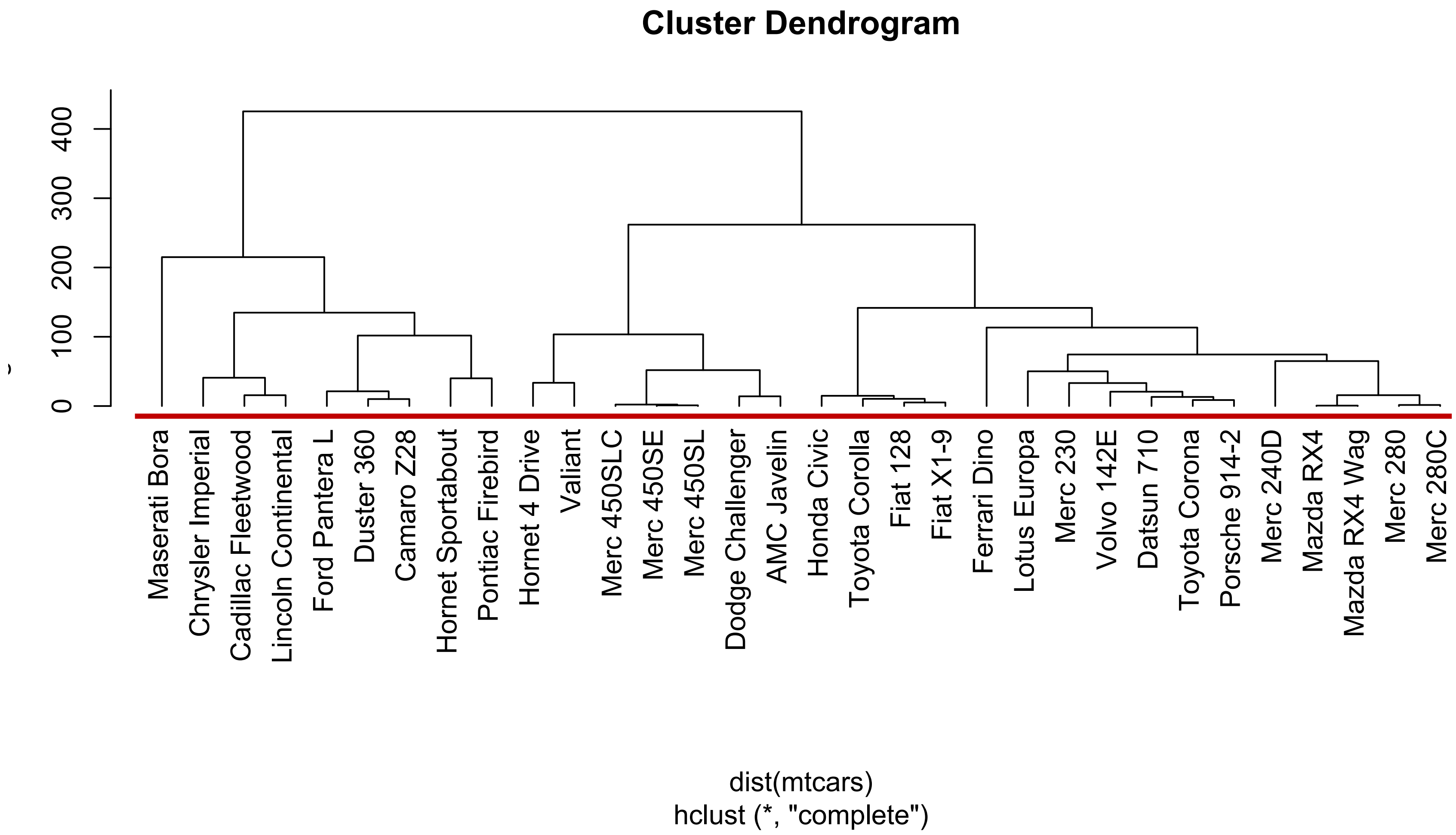


Hierarchical Clustering Algorithm

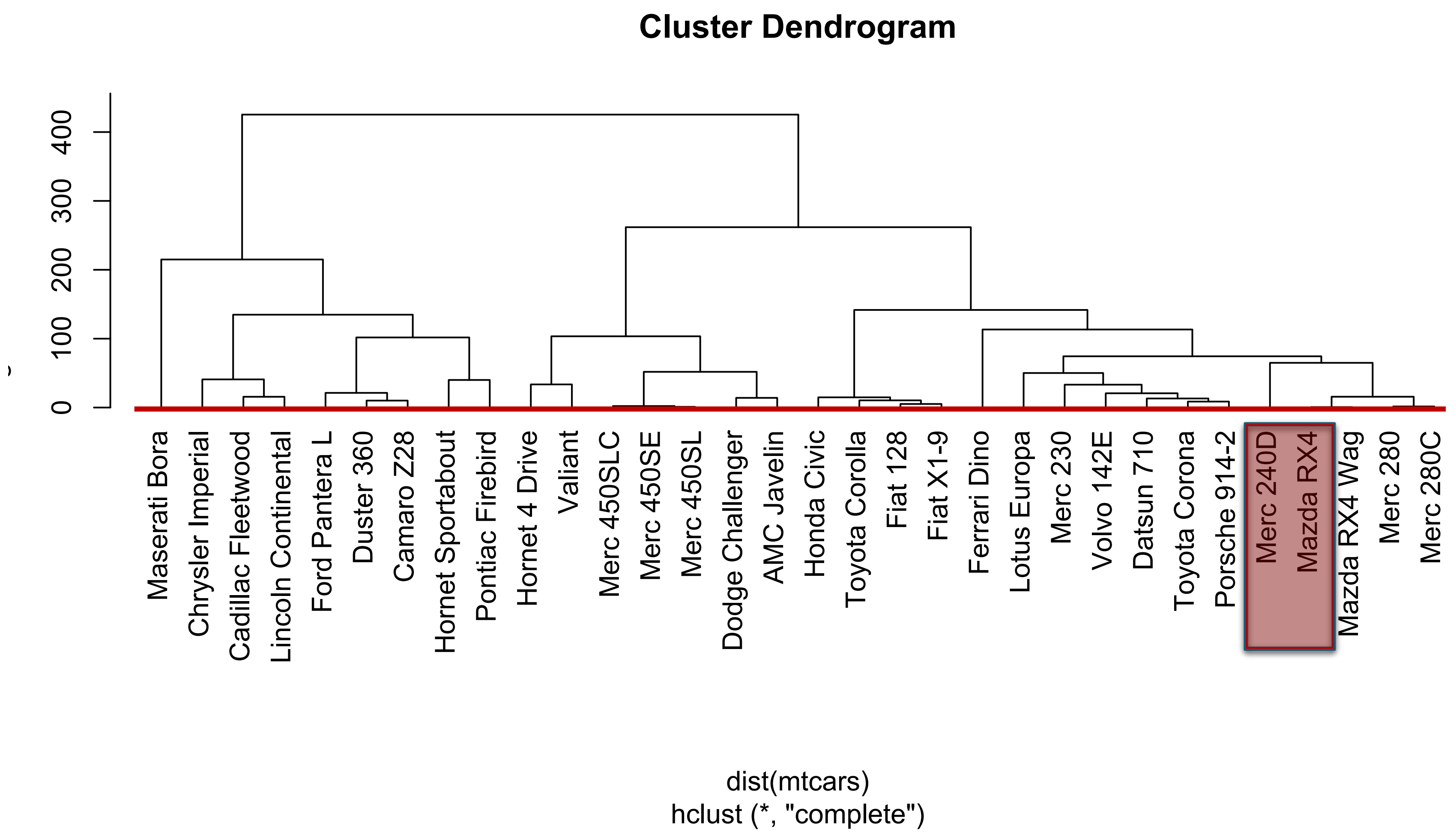
An Illustration with mtcars



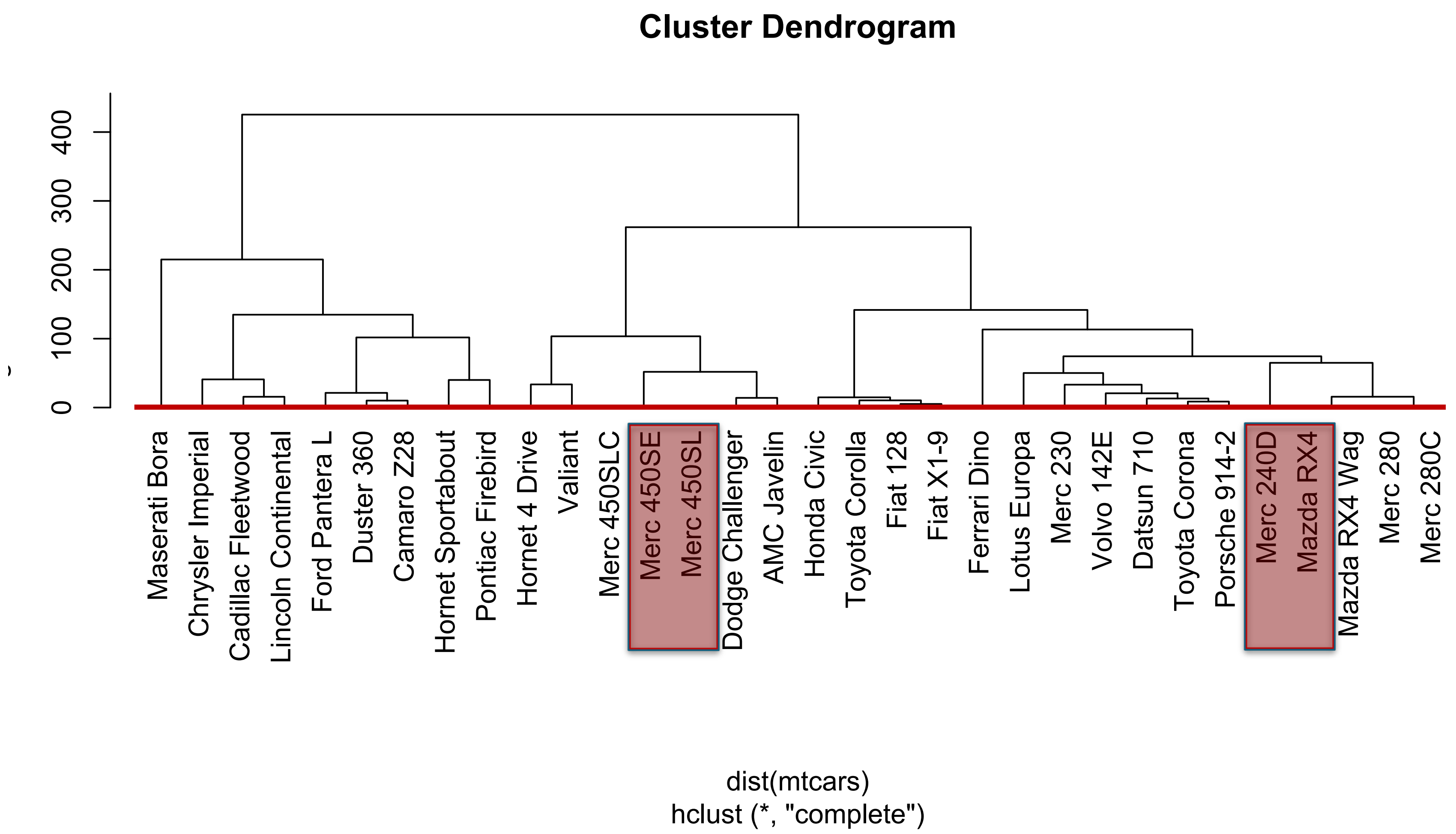
Hierarchical Clustering Algorithm
An Illustration with mtcars



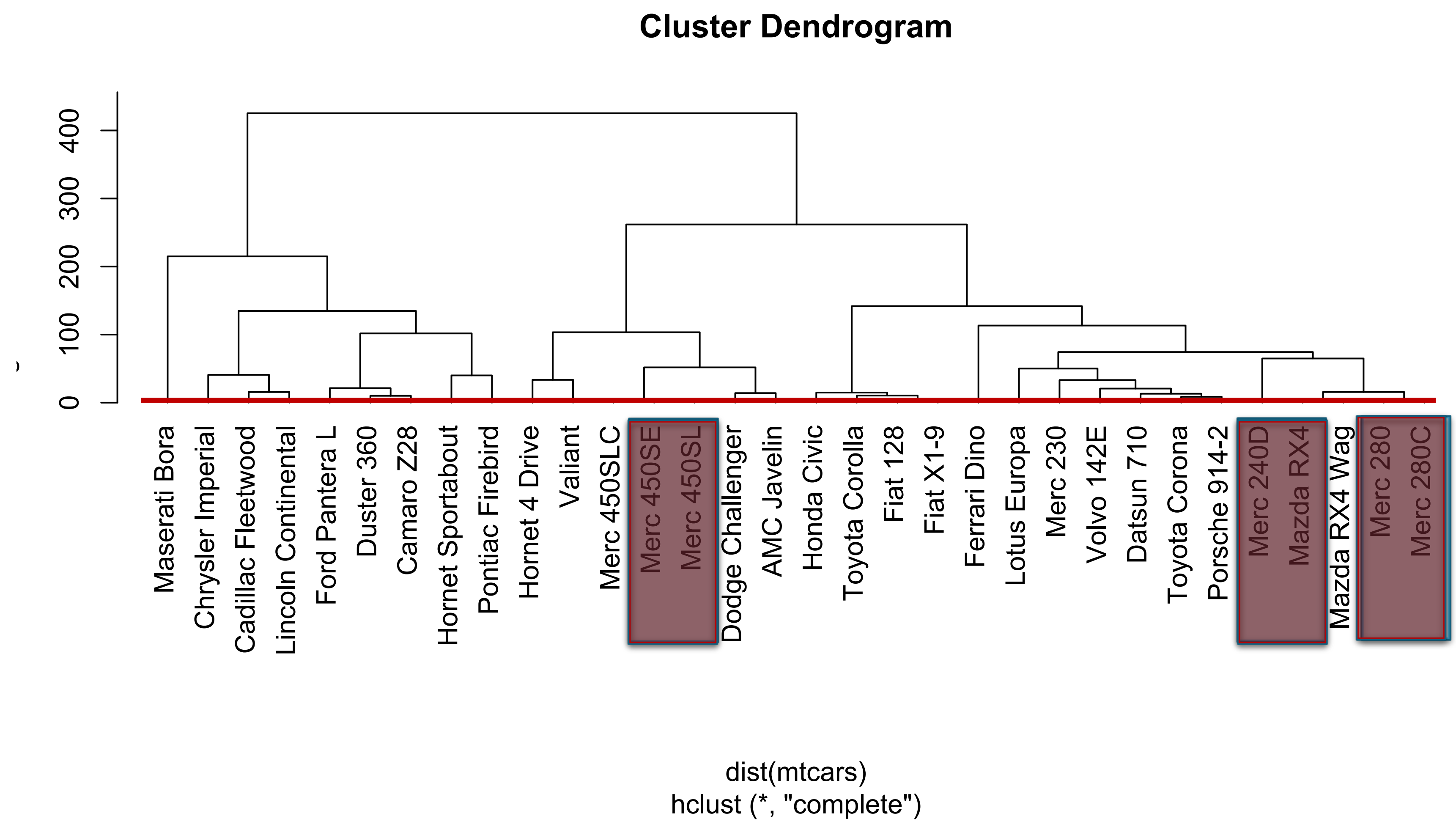
Hierarchical Clustering Algorithm
An Illustration with mtcars



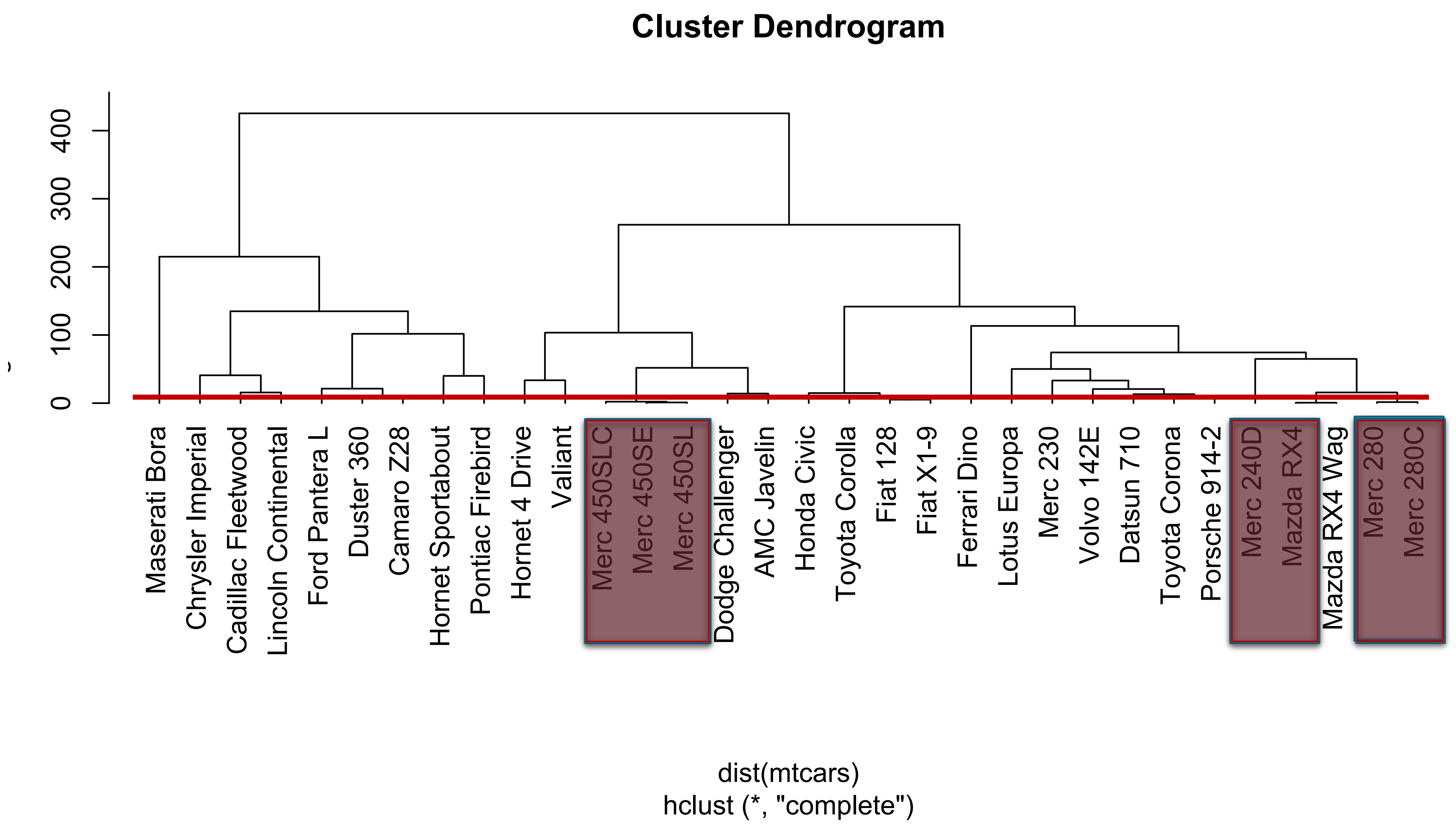
Hierarchical Clustering Algorithm
An Illustration with mtcars



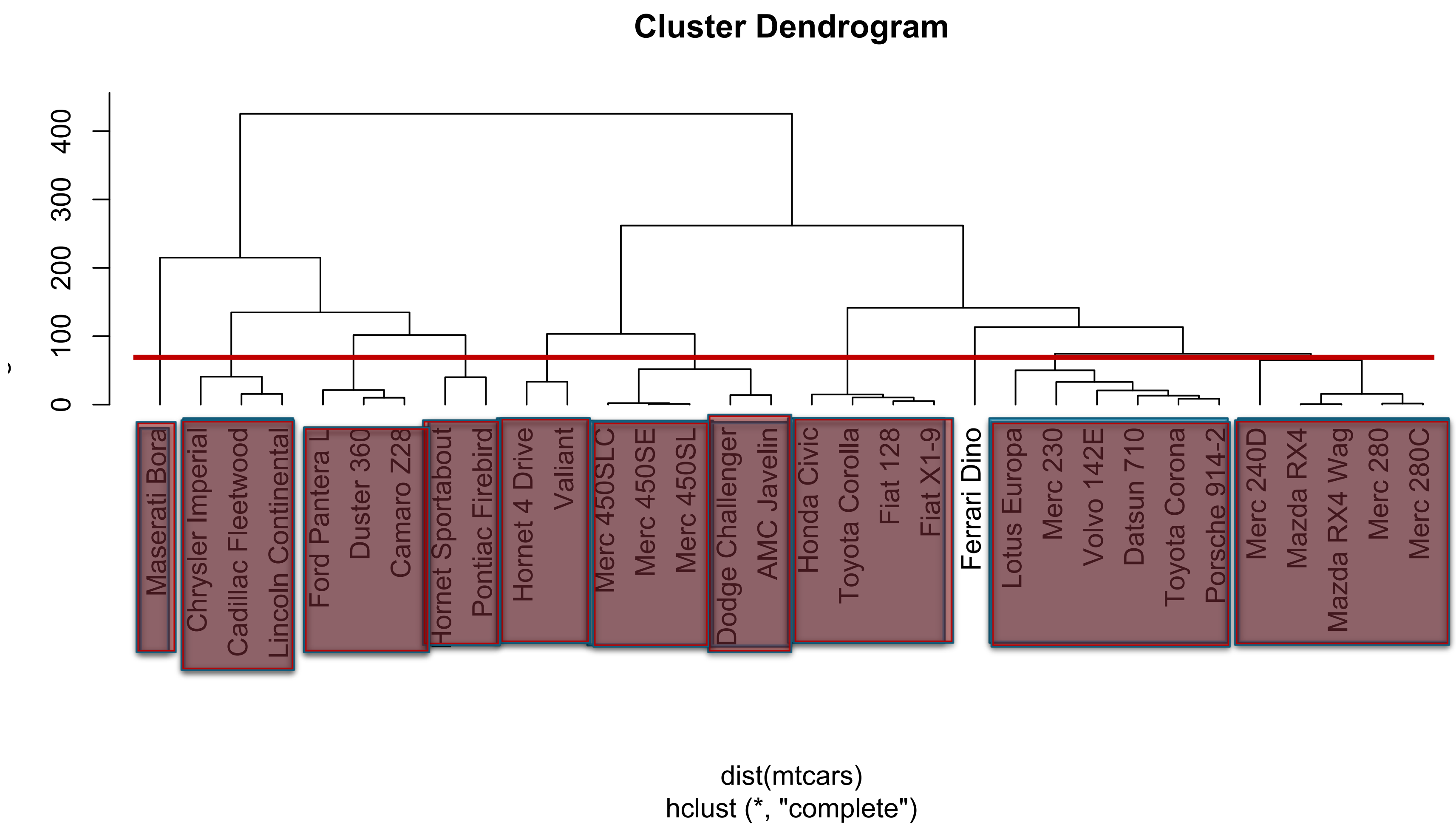
Hierarchical Clustering Algorithm
An Illustration with mtcars



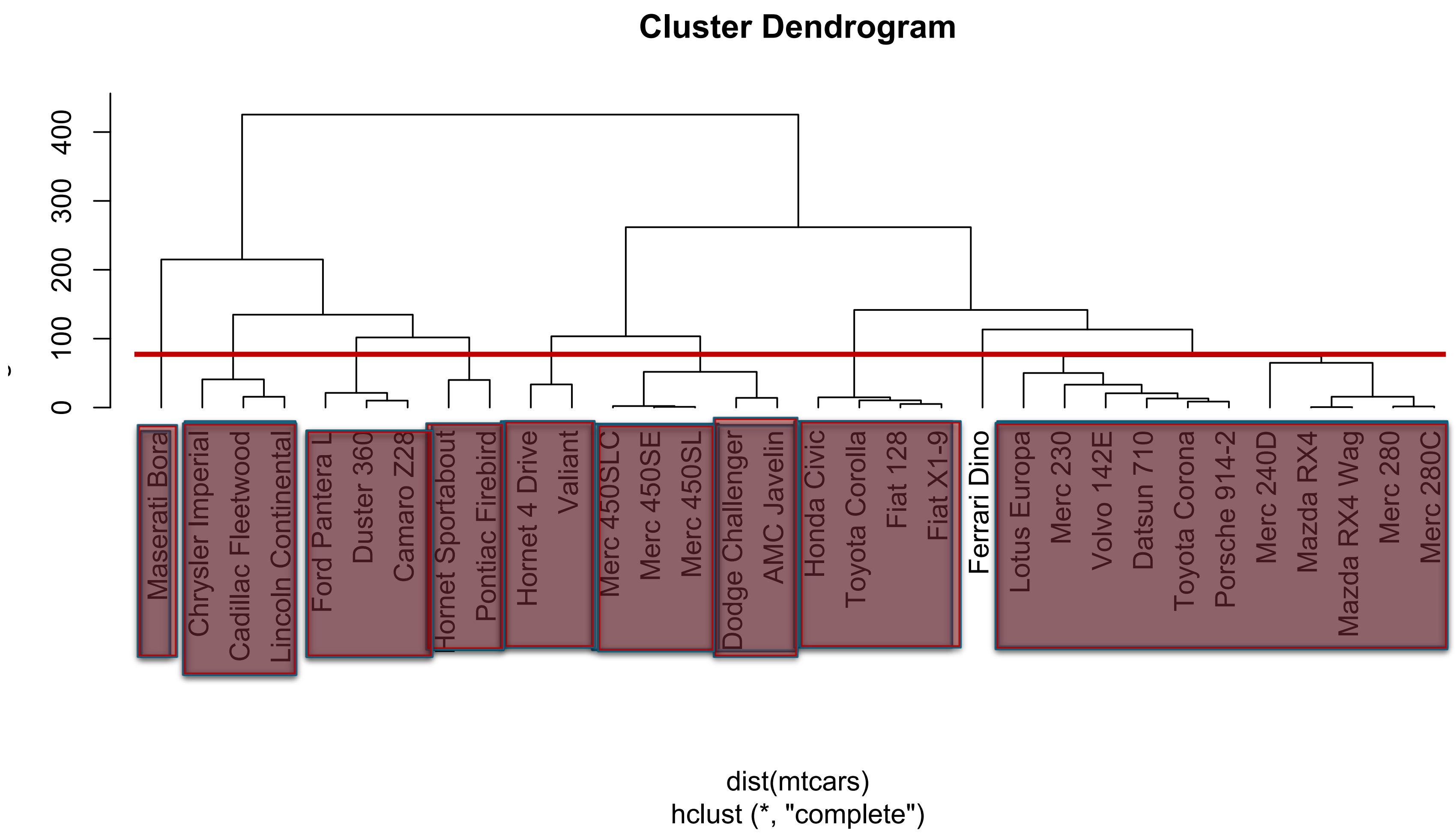
Hierarchical Clustering Algorithm
An Illustration with mtcars



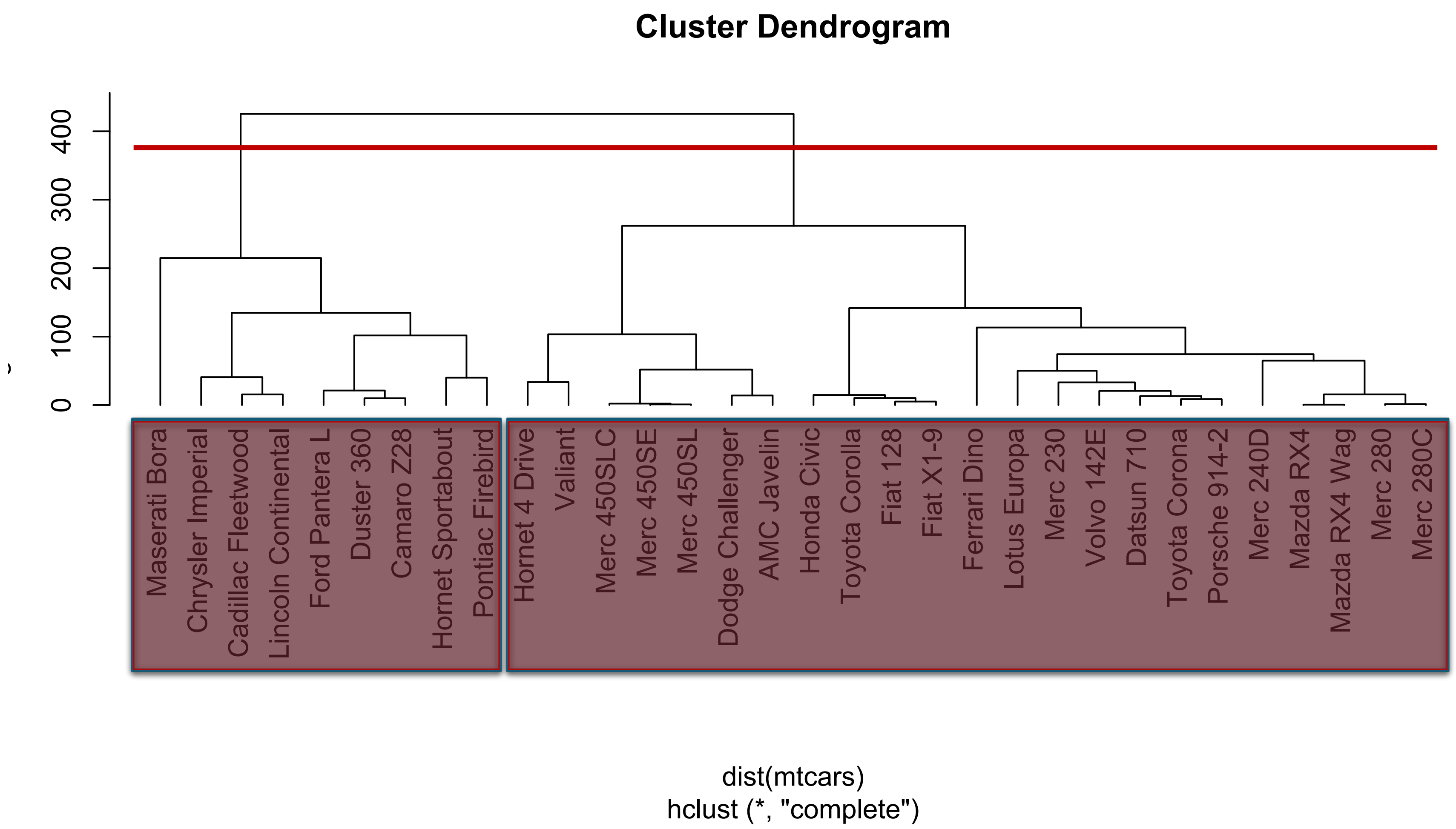
Hierarchical Clustering Algorithm
An Illustration with mtcars



Hierarchical Clustering Algorithm
An Illustration with mtcars

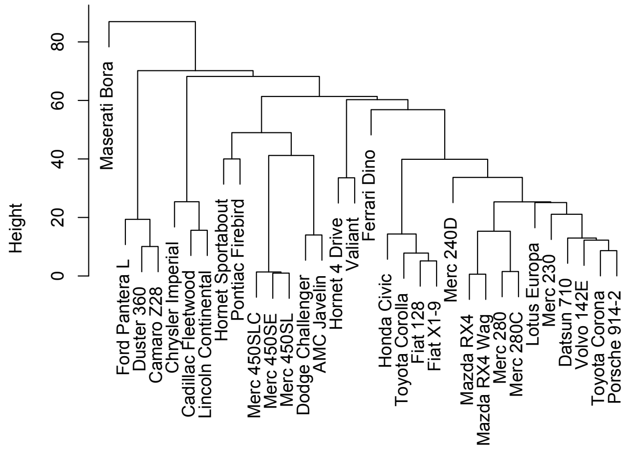


Hierarchical Clustering Algorithm
An Illustration with mtcars

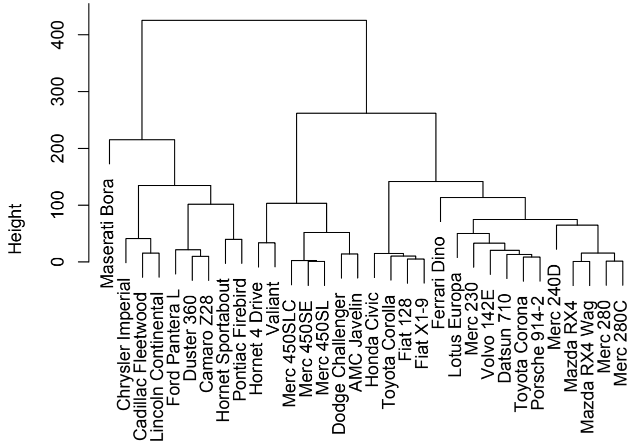


Hierarchical Clustering Algorithm

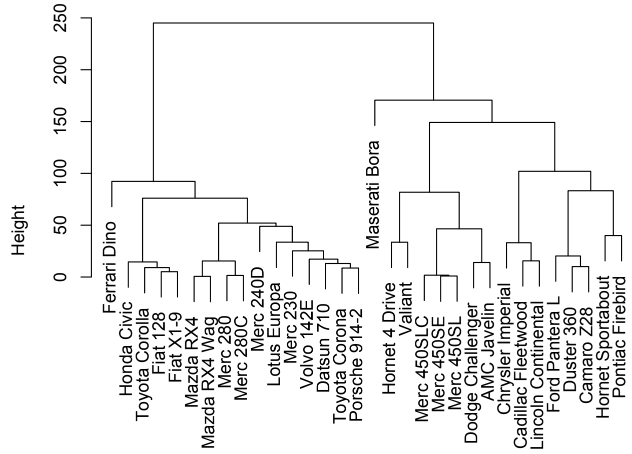
Same Cluster, Different Parameters



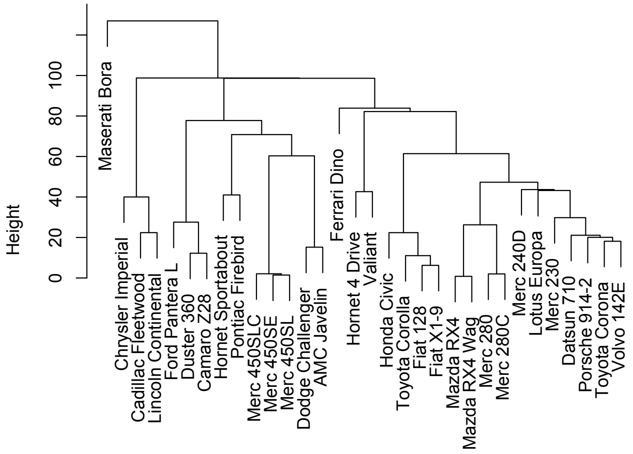
Euclidean Distance Single Link



Euclidean Distance Complete Link



Euclidean Distance Average Link

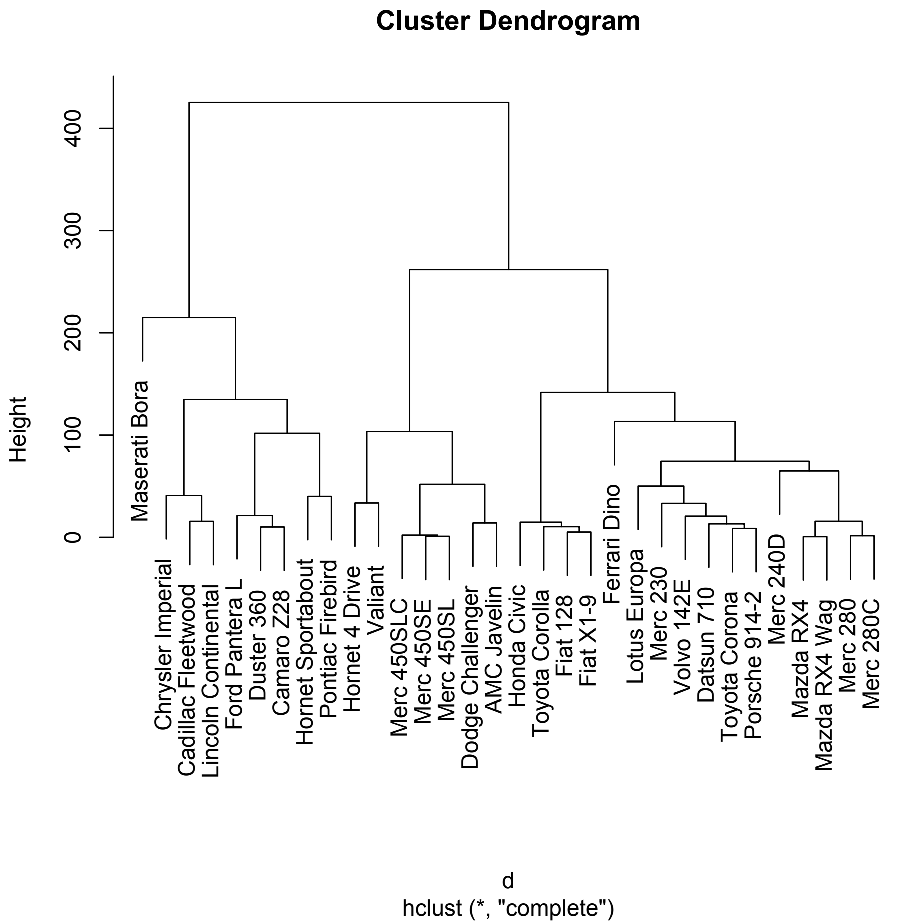


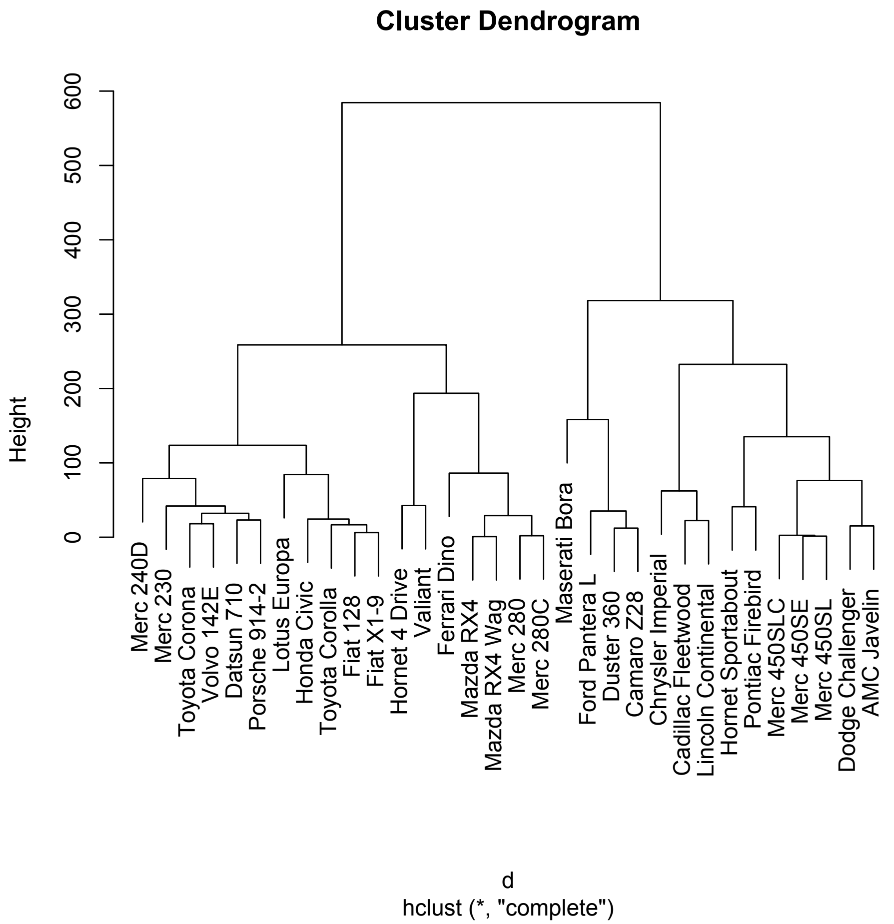
Manhattan Distance Single Link

Same data clustered using different parameter settings:

- **distance** metric
- **linkage** strategy

Hierarchical Clustering Algorithm
Parameters: Distance Metric





Same data clustered using two different distance metrics (euclidean, manhattan)

Hierarchical Clustering Algorithm

Similarity-Dissimilarity

Compare objects

```
> mtcars
                     mpg cyl  disp  hp drat    wt
Mazda RX4           21.0   6 160.0 110 3.90 2.620
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875
Datsun 710          22.8   4 108.0  93 3.85 2.320
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440
Valiant             18.1   6 225.0 105 2.76 3.460
Duster 360          14.3   8 360.0 245 3.21 3.570
Merc 240D           24.4   4 146.7  62 3.69 3.190
Merc 230            22.8   4 140.8  95 3.92 3.150
Merc 280            19.2   6 167.6 123 3.92 3.440
Merc 280C           17.8   6 167.6 123 3.92 3.440
Merc 450SE          16.4   8 275.8 180 3.07 4.070
Merc 450SL          17.3   8 275.8 180 3.07 3.730
Merc 450SLC         15.2   8 275.8 180 3.07 3.780
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250
Lincoln Continental 10.4   8 460.0 215 3.00 5.424
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345
Fiat 128            32.4   4  78.7  66 4.08 2.200
Honda Civic         30.4   4  75.7  52 4.93 1.615
Toyota Corolla      33.9   4  71.1  65 4.22 1.835
Toyota Corona       21.5   4 120.1  97 3.70 2.465
Dodge Challenger    15.5   8 318.0 150 2.76 3.520
AMC Javelin         15.2   8 304.0 150 3.15 3.435
Camaro Z28          13.3   8 350.0 245 3.73 3.840
Pontiac Firebird    19.2   8 400.0 175 3.08 3.845
Fiat X1-9           27.3   4  79.0  66 4.08 1.935
Porsche 914-2       26.0   4 120.3  91 4.43 2.140
Lotus Europa        30.4   4  95.1 113 3.77 1.513
Ford Pantera L      15.8   8 351.0 264 4.22 3.170
Ferrari Dino        19.7   6 145.0 175 3.62 2.770
Maserati Bora       15.0   8 301.0 335 3.54 3.570
Volvo 142E          21.4   4 121.0 109 4.11 2.780
```

Compare values

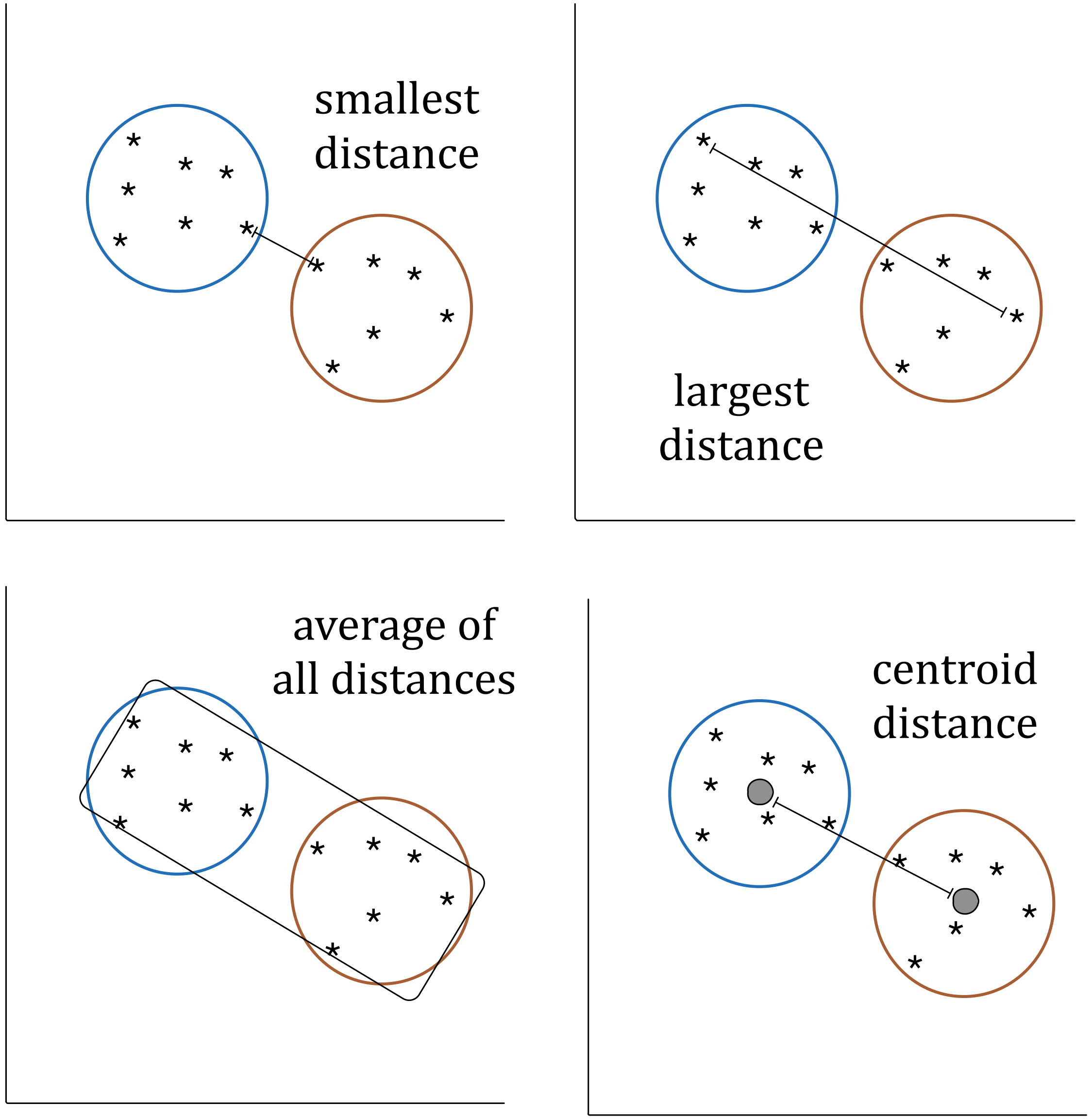
Compare groups

Compare variables


```
> mtcars
                     mpg cyl  disp  hp drat    wt
Mazda RX4           21.0   6 160.0 110 3.90 2.620
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875
Datsun 710          22.8   4 108.0  93 3.85 2.320
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440
Valiant             18.1   6 225.0 105 2.76 3.460
Duster 360          14.3   8 360.0 245 3.21 3.570
Merc 240D           24.4   4 146.7  62 3.69 3.190
Merc 230            22.8   4 140.8  95 3.92 3.150
Merc 280            19.2   6 167.6 123 3.92 3.440
Merc 280C           17.8   6 167.6 123 3.92 3.440
Merc 450SE          16.4   8 275.8 180 3.07 4.070
Merc 450SL          17.3   8 275.8 180 3.07 3.730
Merc 450SLC         15.2   8 275.8 180 3.07 3.780
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250
Lincoln Continental 10.4   8 460.0 215 3.00 5.424
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345
Fiat 128            32.4   4  78.7  66 4.08 2.200
Honda Civic         30.4   4  75.7  52 4.93 1.615
Toyota Corolla      33.9   4  71.1  65 4.22 1.835
Toyota Corona       21.5   4 120.1  97 3.70 2.465
Dodge Challenger    15.5   8 318.0 150 2.76 3.520
AMC Javelin         15.2   8 304.0 150 3.15 3.435
Camaro Z28          13.3   8 350.0 245 3.73 3.840
Pontiac Firebird    19.2   8 400.0 175 3.08 3.845
Fiat X1-9           27.3   4  79.0  66 4.08 1.935
Porsche 914-2       26.0   4 120.3  91 4.43 2.140
Lotus Europa        30.4   4  95.1 113 3.77 1.513
Ford Pantera L      15.8   8 351.0 264 4.22 3.170
Ferrari Dino        19.7   6 145.0 175 3.62 2.770
Maserati Bora       15.0   8 301.0 335 3.54 3.570
Volvo 142E          21.4   4 121.0 109 4.11 2.780
```

Hierarchical Clustering Algorithm

Parameters: Linkage



The chosen linkage algorithm affects which clusters are merged, and the shape of the resulting clusters (e.g. tighter, looser)

Strengths and Limitations – Linkages

Single Linkage (smallest distance)

- can handle non-blob shapes
- sensitive to noise and outliers
- produces elongated clusters

Complete Linkage (largest distance)

- balanced clusters, with similar diameters
- not overly sensitive to noise
- tends to split large clusters
- all clusters tend to have similar diameters

Strengths and Limitations – Linkages

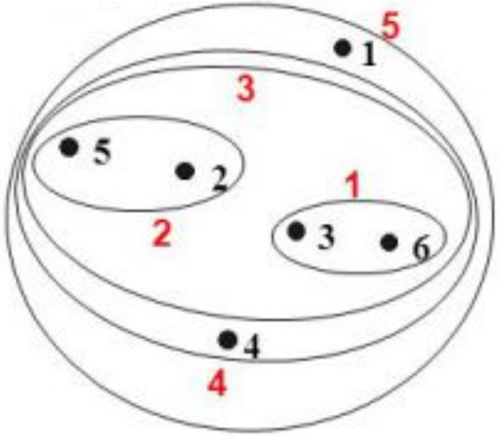
Average Linkage (average distance)

- compromise between single and complete linkages
- not too sensitive to noise and outliers
- tends to produce blob-shaped clusters

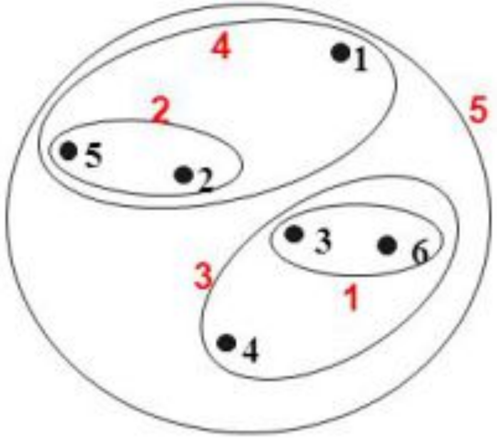
Complete Linkage (largest distance)

- balanced clusters, with similar diameters
- not overly sensitive to noise
- tends to split large clusters
- all clusters tend to have similar diameters

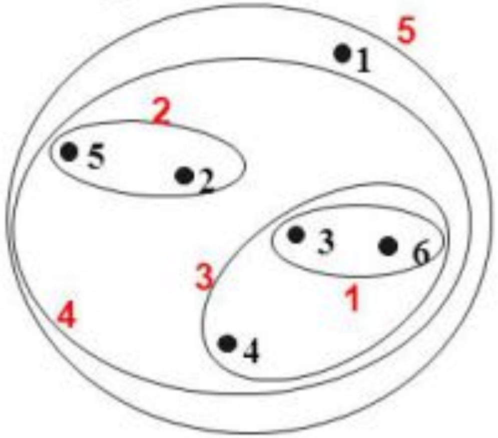
Simple Link



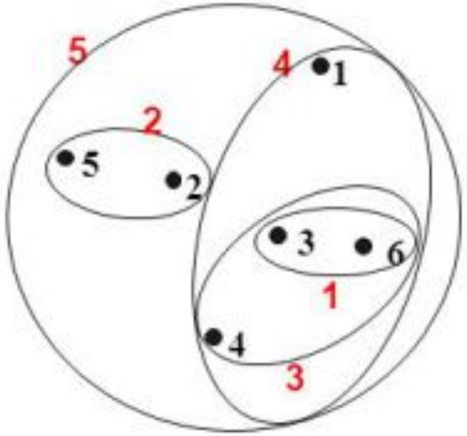
Complete Link



Average Link



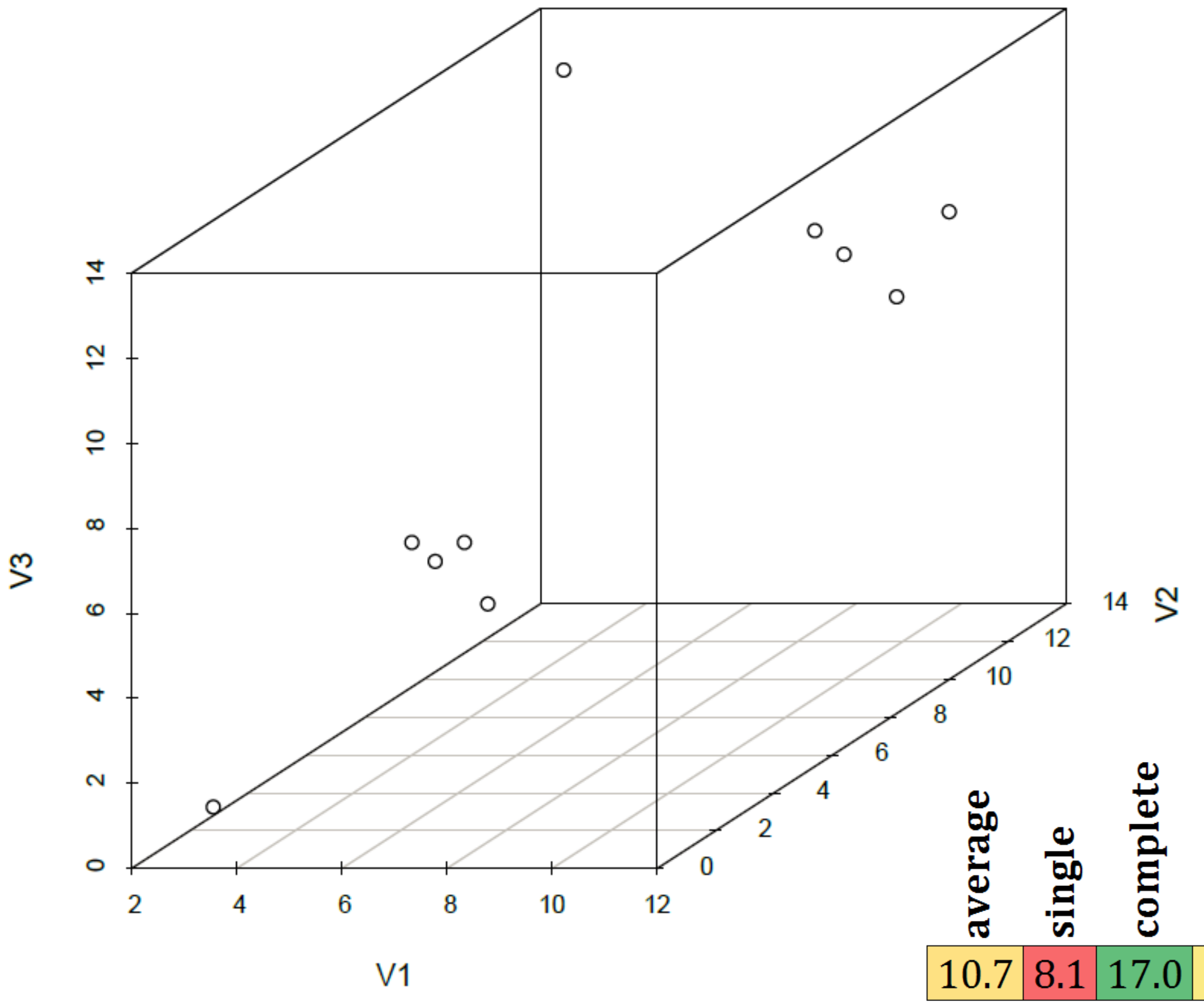
Centroid Link



[https://dataaspirant.com/2018/01/08/hierarchical-clustering-r/]

Hierarchical Clustering Algorithm

Parameters: Linkage – Example



Data

Cluster A	Cluster B
1: (5,5,5) 2: (5,6,5) 3: (4,6,5) 4: (6,5,4) 5: (3,1,1)	6: (10,10,10) 7: (11,10,9) 8: (12,10,11) 9: (10,9,11) 10: (13,13,13)
c: (4.6,4.6,4)	c: (11.2,10.4,10.8)

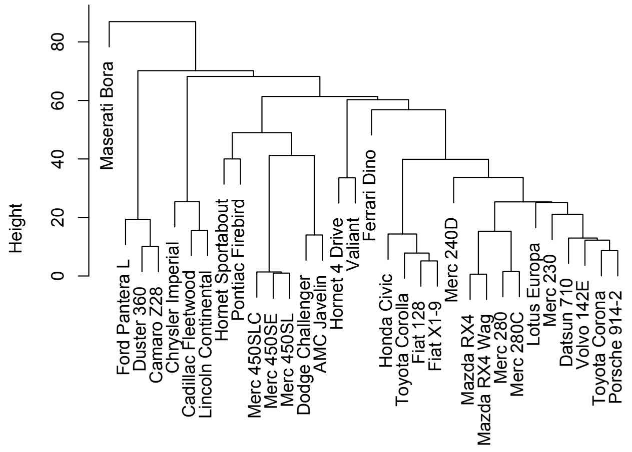
Distance matrix

Cluster *B* \ Cluster *A*	**1**	**2**	**3**	**4**	**5**
6	8.7	8.1	8.8	8.8	14.5
7	8.8	8.2	9.0	8.7	14.5
8	10.5	10.0	10.8	10.5	16.2
9	8.8	8.4	9.0	9.0	14.6
10	11.5	10.8	10.7	12.4	17.0

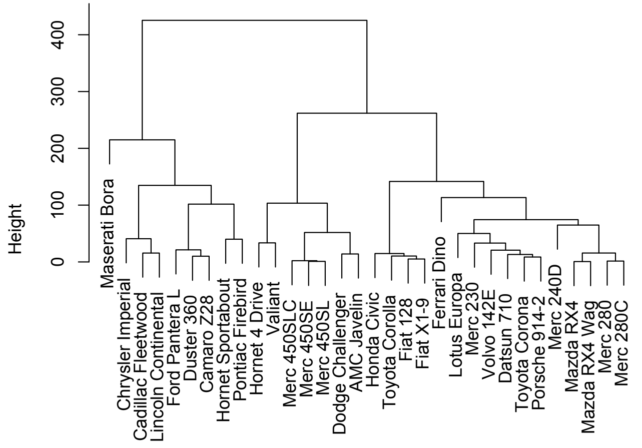
average	**single**	**complete**	**centroid**
10.7	8.1	17.0	11.1

Linkage

Hierarchical Clustering Algorithm
Returning to Our Clustering Results



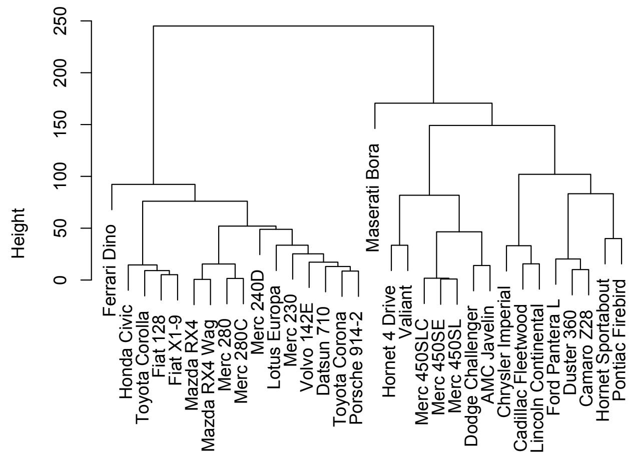
Euclidean Distance Single Link



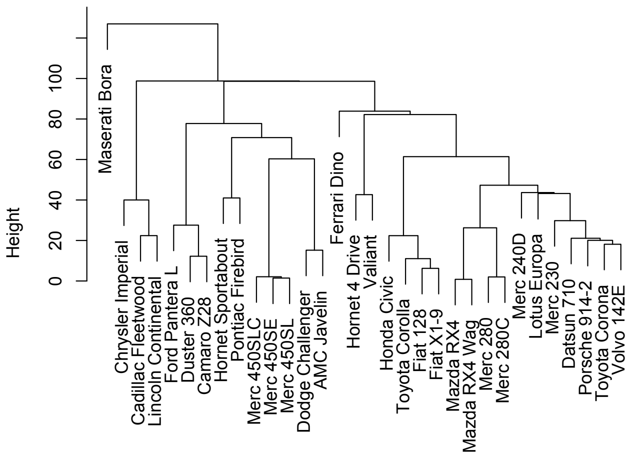
Euclidean Distance Complete Link

It is often said: choose the distance metric that is most **meaningful** for your data.

Can evaluation metrics also inform this choice?



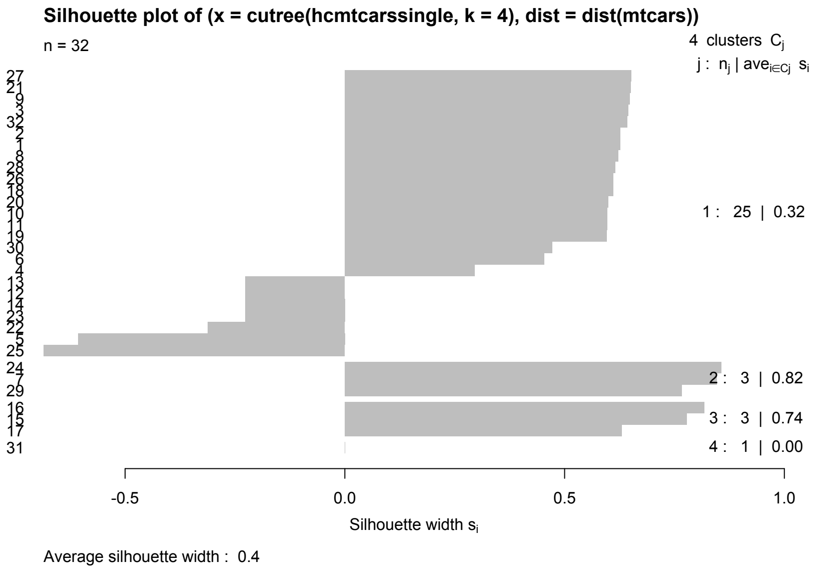
Euclidean Distance Average Link

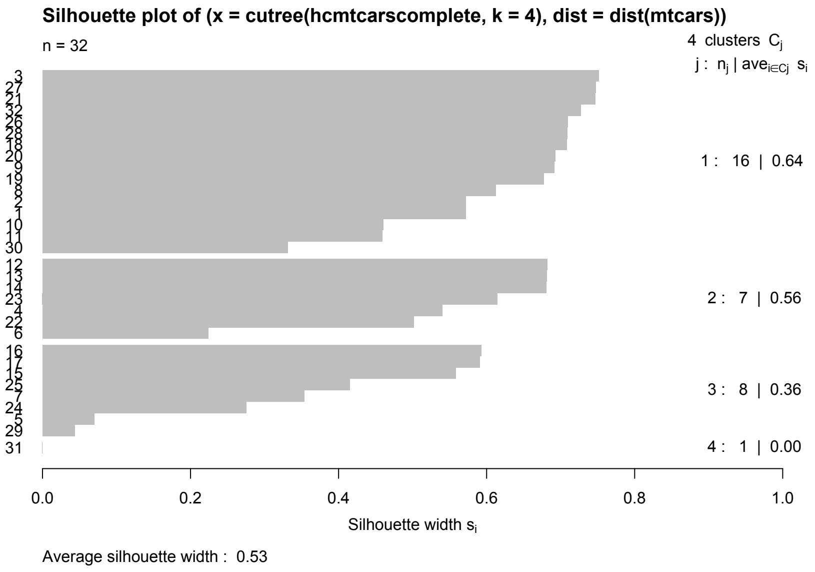


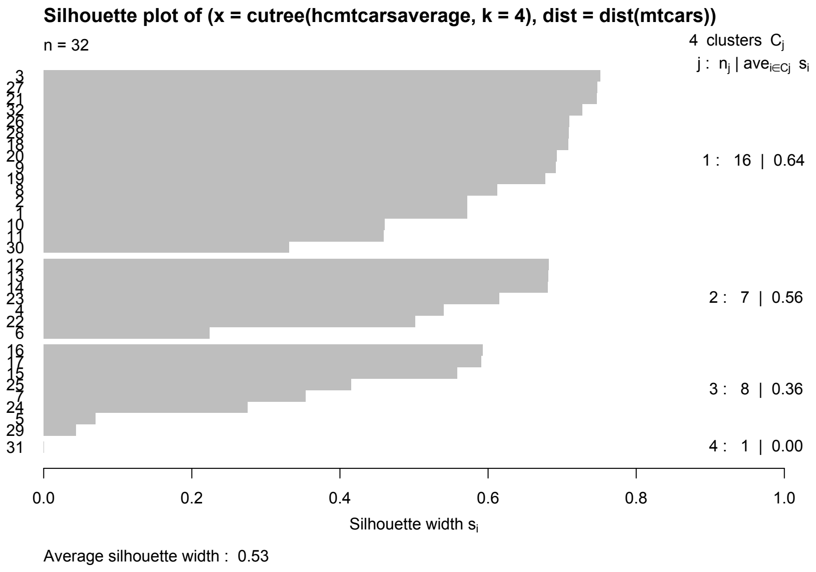
Manhattan Distance Single Link

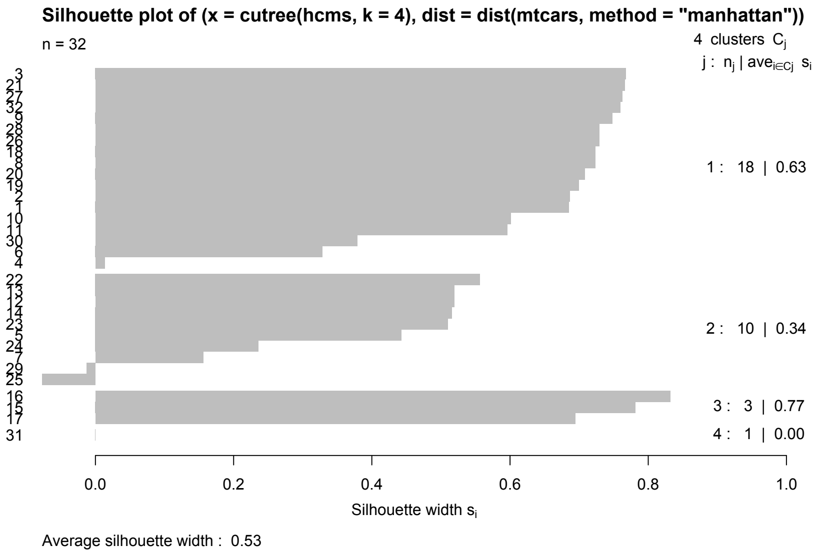
Hierarchical Clustering Algorithm

Silhouette of Clustering Results







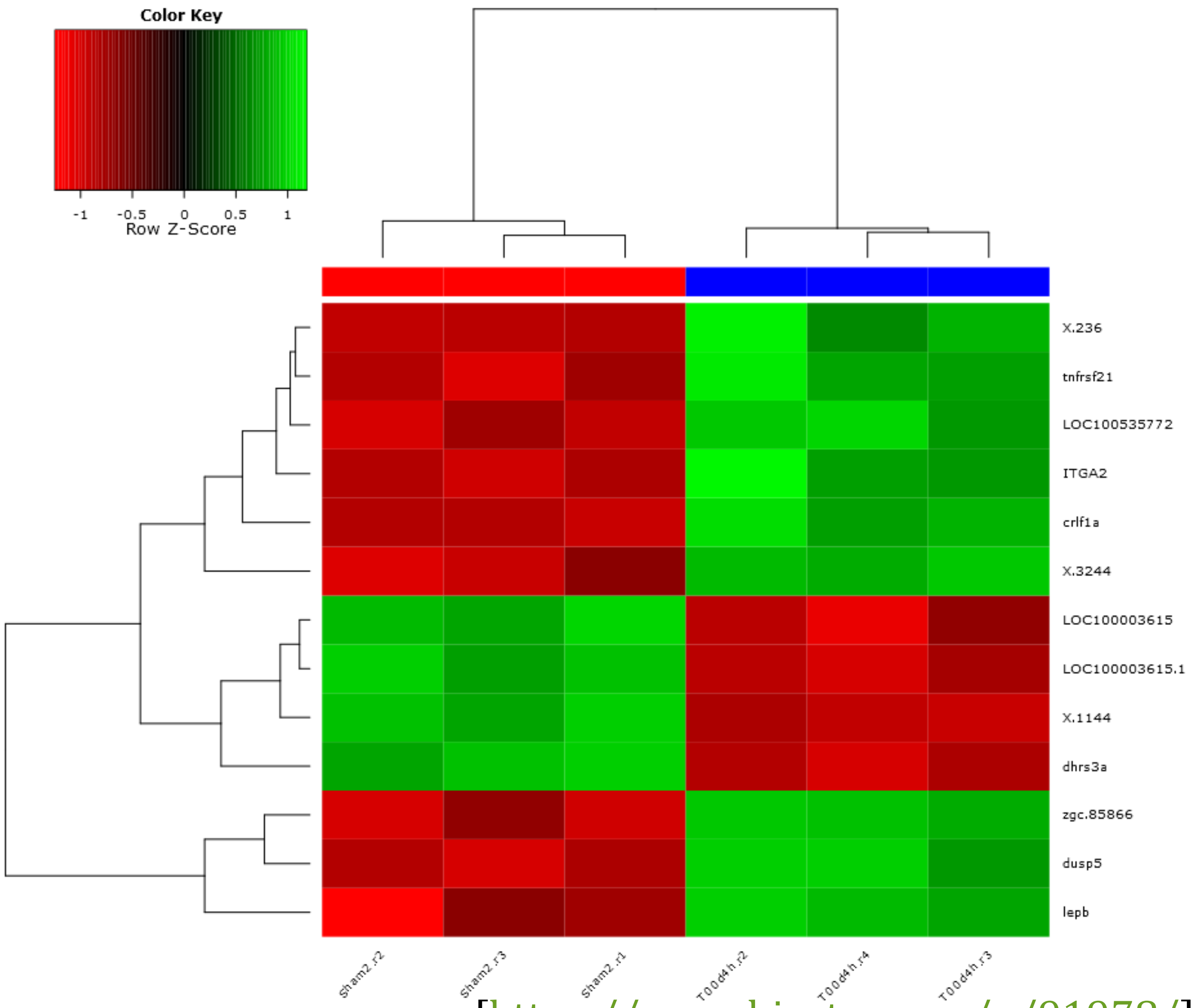


Hierarchical Clustering Notes

HC is deterministic, for a given choice of metric and linkage.

Space and time requirements do make HC unattractive for medium-to-large datasets.

Various linkage strategies: be sure to check out **Wald's method**!



[https://www.biostars.org/p/91978/]

HC Examples and Case Studies
Clustering Myths

Comparative Mythology

- Studying myths from different cultures to understand their similarities and possibly shared origins
- Many myths have splintered off and evolved from common sources

Julien d'Huy (2016): Used a variety of data mining techniques, including hierarchical clustering, to trace the evolution of myths.

Collection of myths broken down into common story elements.

A myth across cultures: the hunter in the sky

J.d'Huy [2016], *Scientists Trace Society's Myths to Primordial Origins*, Scientific American (Online)

HC Examples and Case Studies
Clustering Myths

Myths categorized based on presence/ absence of elements

Myths are clustered based on this categorization.

Result shows myths clustering together – could this suggest a possible common origin for these myths?

Remember, **clustering is knowledge discovery**!

A myth across cultures: the hunter in the sky

J.d'Huy [2016], *Scientists Trace Society's Myths to Primordial Origins*, Scientific American (Online)

HC Examples and Case Studies

Complex building's energy system operation patterns analysis using bag of words representation with hierarchical clustering

A Comparison of Antioxidant, Antibacterial, and Anticancer Activity of the Selected Thyme Species by Means of Hierarchical Clustering and Principal Component Analysis

Use of hierarchical cluster analysis to classify prisons in Ireland into mutually exclusive drug-use risk categories

Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets