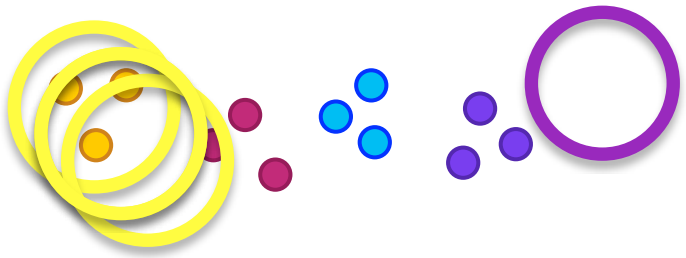
Cluster Analysis I

MENGQIAN LU

Measure dissimilarity between **groups** (cont'd)

- None is perfect, normally choose complete linkage to get started

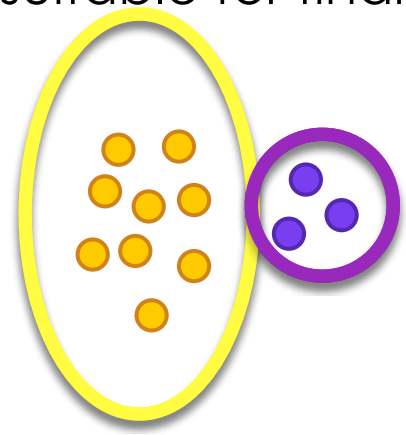
2. *Complete Linkage:* suitable for finding compact but not well separated cluster – "crowd clusters"



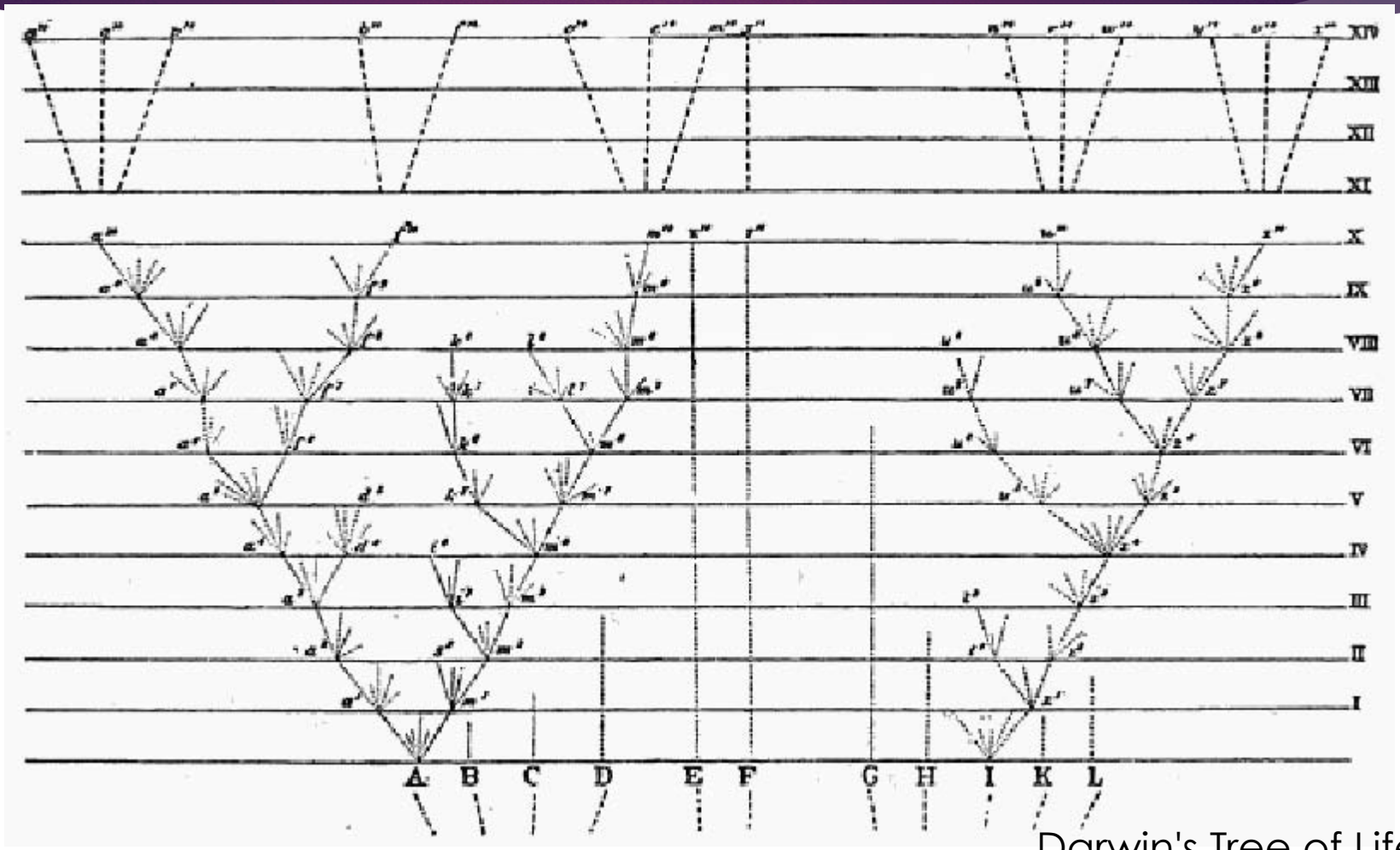
Measure dissimilarity between **groups** (cont'd)

- None is perfect, normally choose complete linkage to get started

3. *Average Linkage:* suitable for finding well separated, oval shaped cluster



Hierarchical clustering – Dengrogram



Darwin's Tree of Life.

hclust() in R

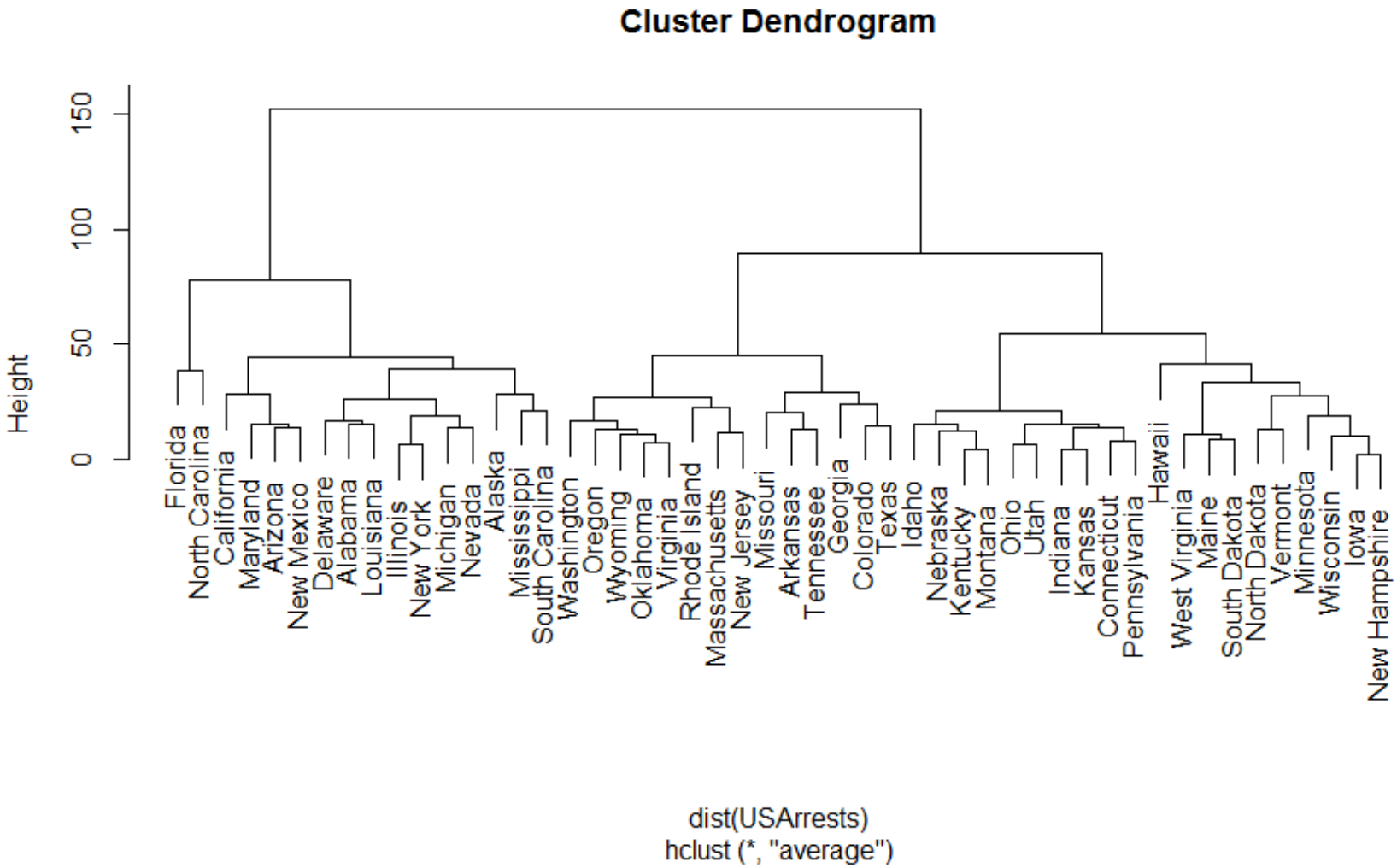
- data(USArrests): numbers are "arrests per 100,000 residents" in the four types of crimes in the 50 US states in 1973.

```
> head(USArrests)
```

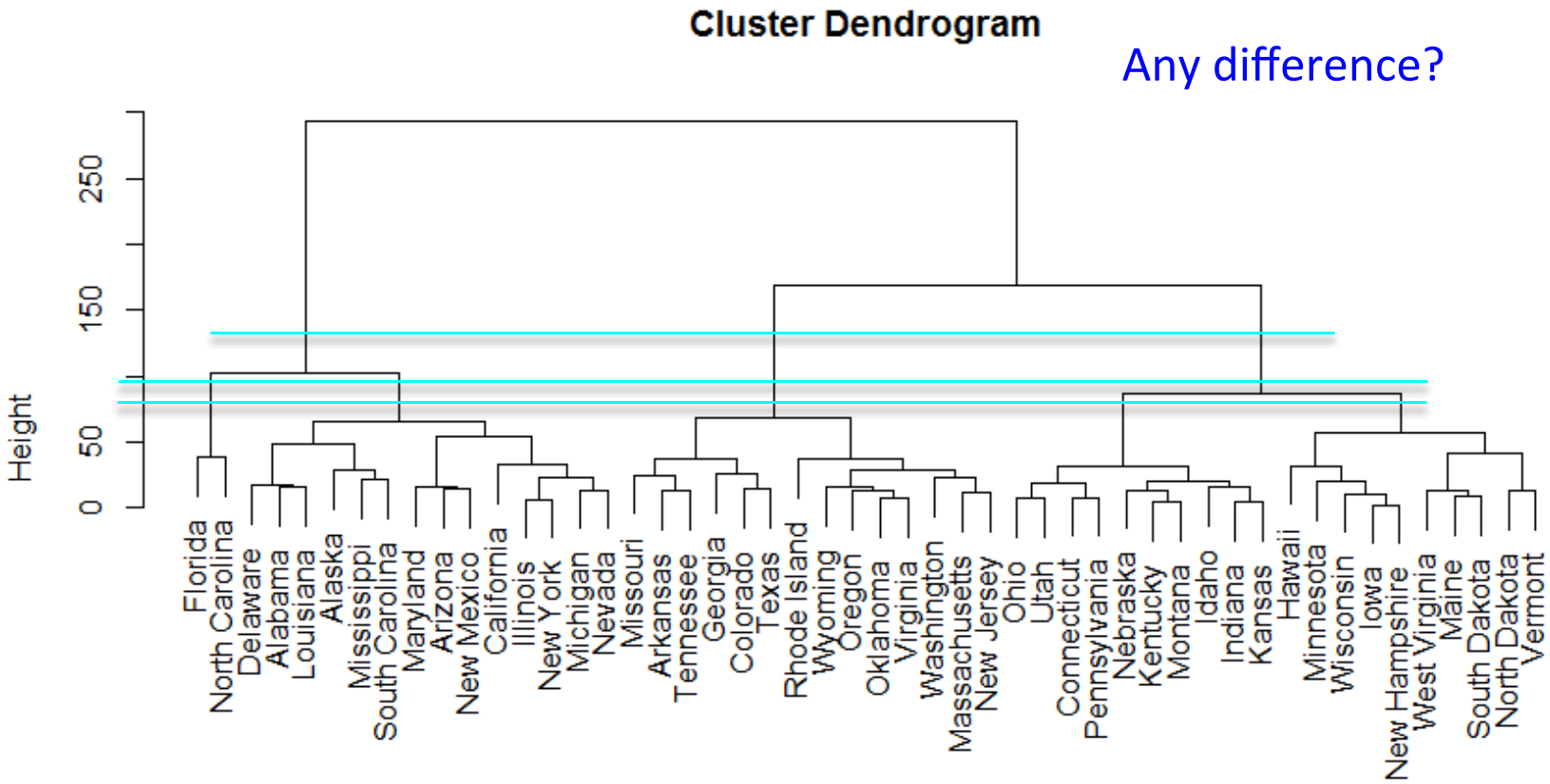
	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10.0	263	48	44.5
Arizona	8.1	294	80	31.0
Arkansas	8.8	190	50	19.5
California	9.0	276	91	40.6
Colorado	7.9	204	78	38.7

```
hc = hclust(dist(USArrests), "ave")
plot(hc)
```

hclust() in R (cont'd)



hclust() in R (cont'd)



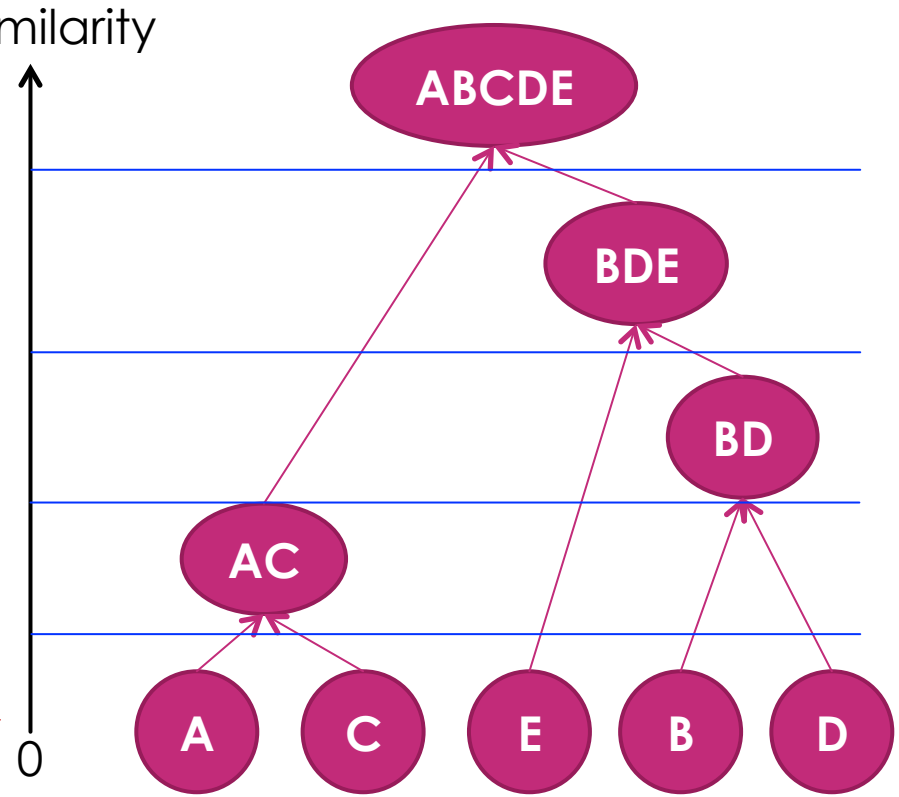
hc = hclust(dist(USArrests), "complete")
plot(hc)

dist(USArrests)
hclust (*, "complete")

Choosing the number of clusters

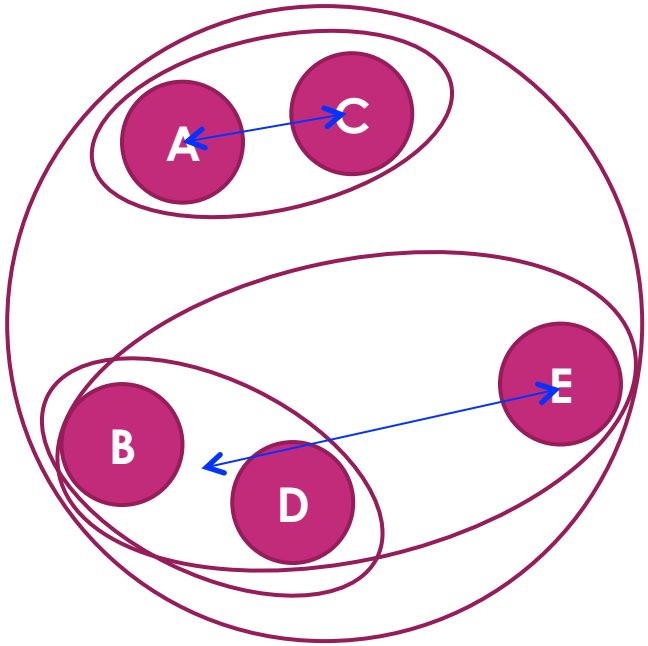
- Get cluster solutions by cutting the tree
- 1 cluster: ABCDE
- 2 clusters: AC – BDE
- 3 clusters: AC – E – BD
- 4 clusters: AC – E – B – D
- 5 clusters: A – C – E – B – D

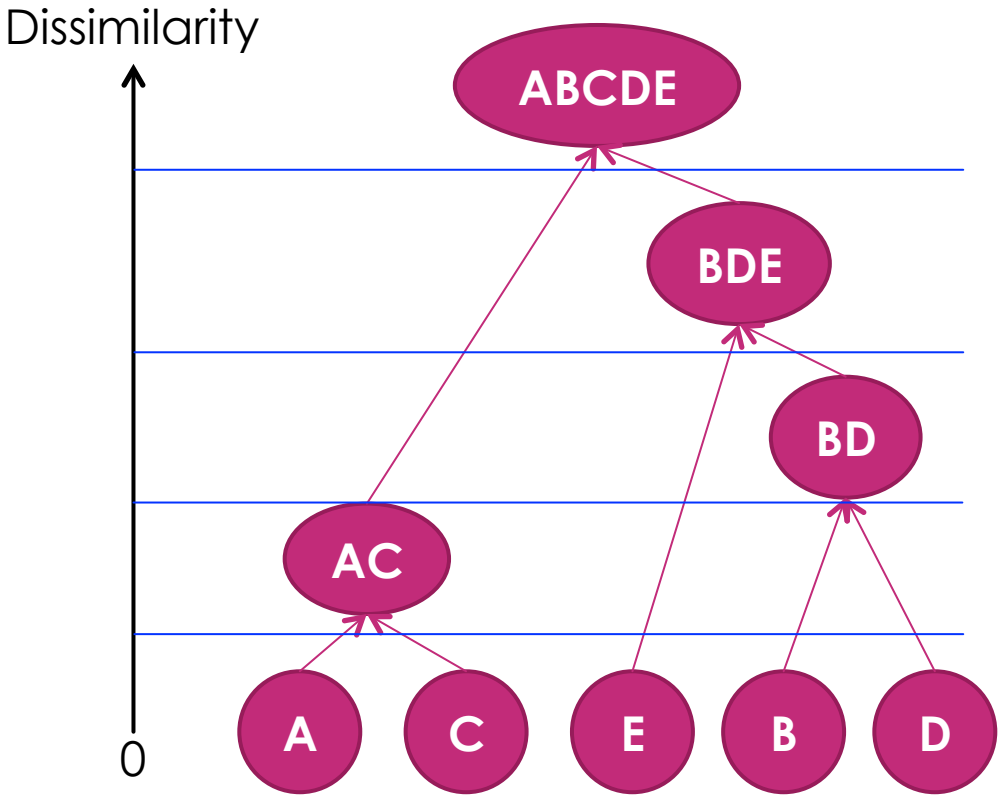
General rule: Find the largest vertical "drop" in the tree



Clustering tree = Dendrogramm

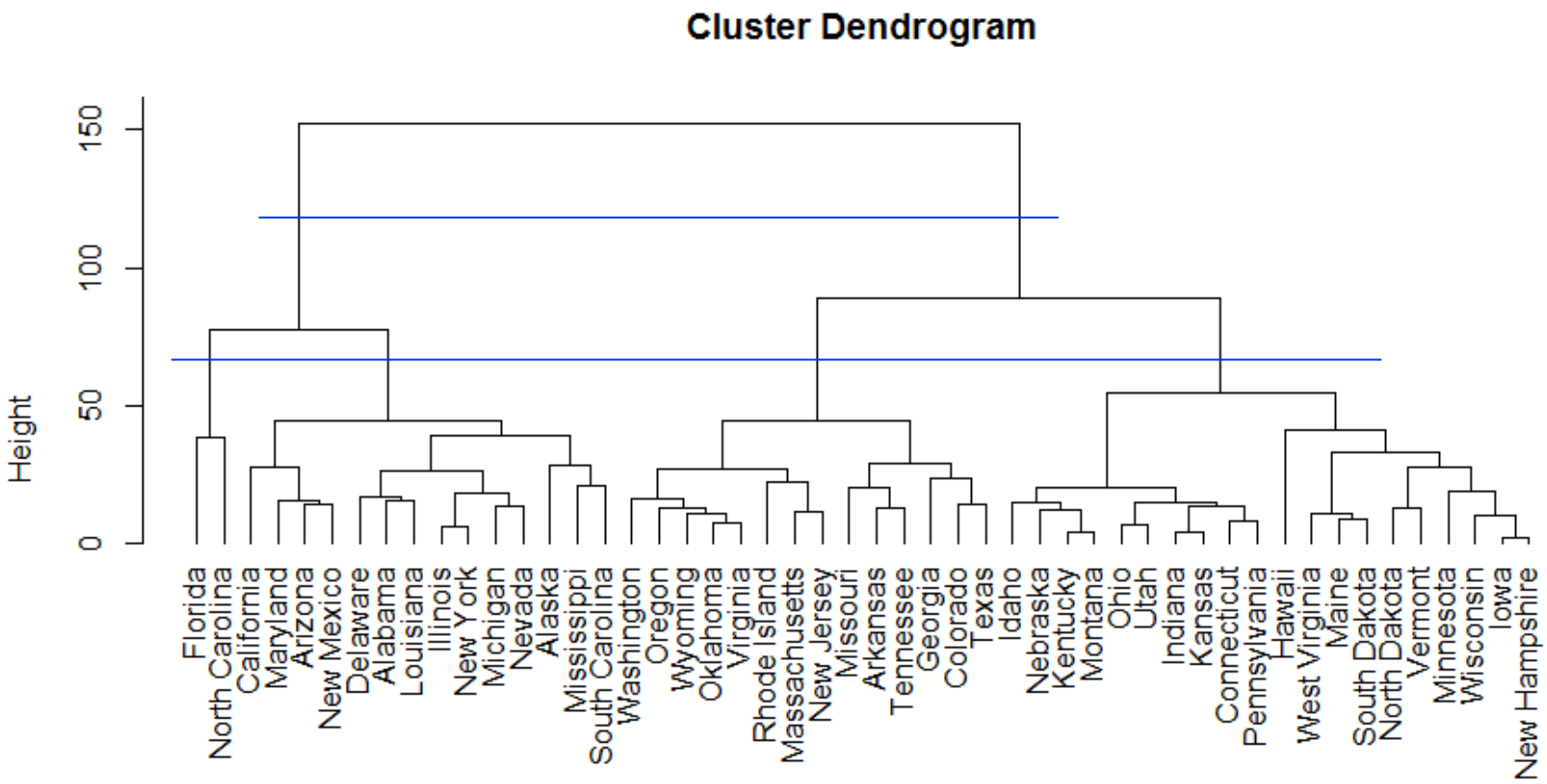
Choosing the number of clusters (cont'd)





Clustering tree = Dendrogramm

Choosing the number of clusters (cont'd)




```
hc = hclust(dist(USArrests), "complete")
plot(hc, hang = -1)
```

Choosing the number of clusters (cont'd)
`cutree()`

```
cutree(hc, k = 1:5)
cutree(hc, h = 250)

## Compare the 2 and 4 grouping:
g24 = cutree(hc, k = c(2,4))
table(grp2 = g24[,"2"], grp4 = g24[,"4"])
```

```
> table(grp2 = g24[,"2"], grp4 = g24[,"4"])
          grp4
grp2  1  2  3  4
1    14  0  0  2
2     0 14 20  0
```

```
> head(cutree(hc, k = 1:5))
           1 2 3 4 5
Alabama    1 1 1 1 1
Alaska     1 1 1 1 1
Arizona    1 1 1 1 1
Arkansas   1 2 2 2 2
California 1 1 1 1 1
Colorado   1 2 2 2 2
```