R Lab - Day 4
Clustering

MED3007 V24
2024.01.18

Chi Zhang
Oslo Center for Biostatistics and Epidemiology
chi.zhang@medisin.uio.no

Overview

Topics for this morning

Hierarchical clustering

Heatmap

K-means clustering

Exercises, reading time

Clustering

Overview

A type of unsupervised learning technique

Partition the data (without labels) into subgroups, where data within the same group are similar

Hierarchical clustering, K-means are common techniques

Unsupervised: no true outcome labels - **make interpretation with caution**

(in the NCI60 and gene expression examples, we have true labels)

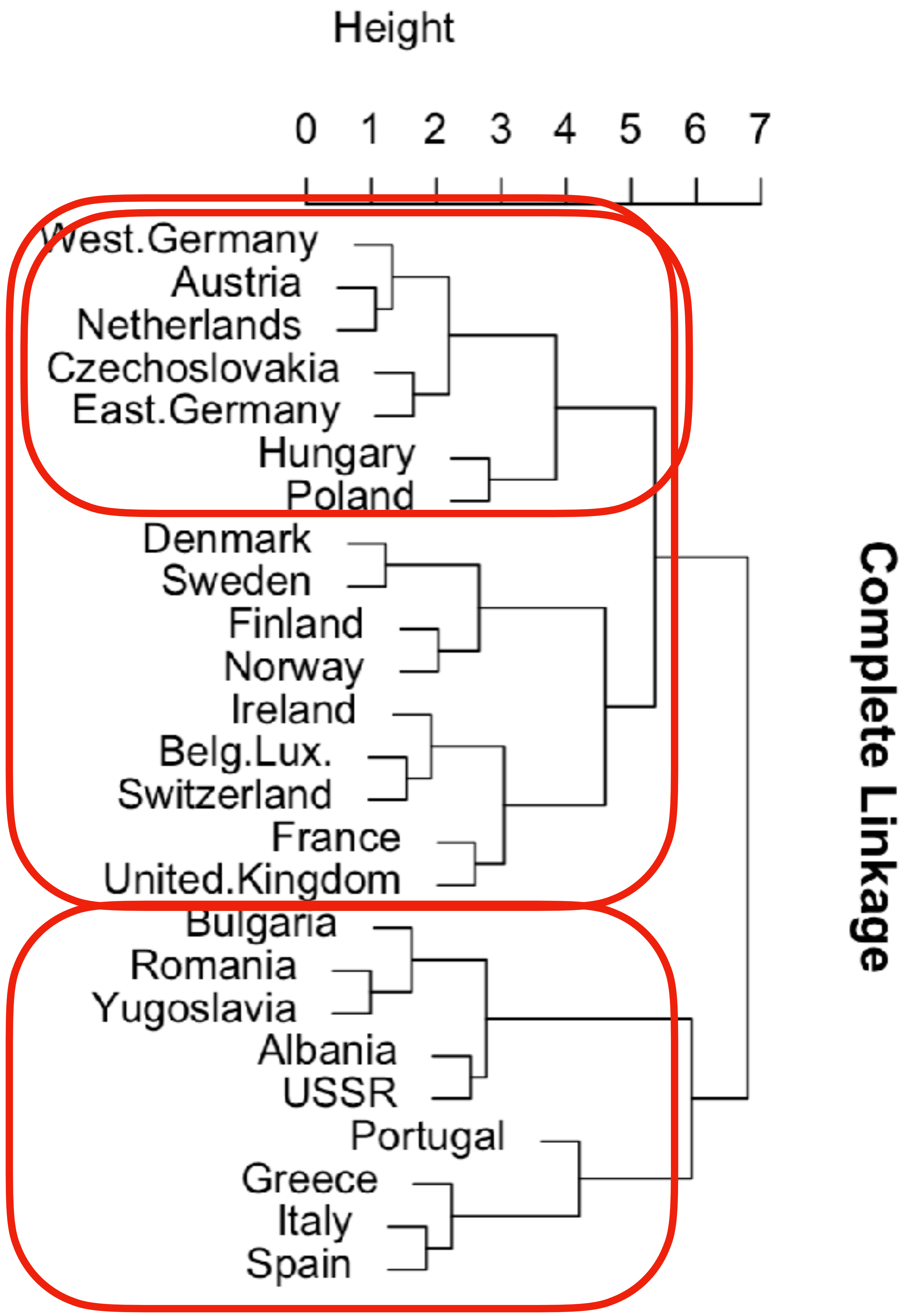
Hierarchical clustering

Overview

Start with **each point** in its own cluster (bottom-up)
Identify the **closest two clusters** and merge them
Repeat
Ends when all points are in a **single cluster**

You do not need to explicitly ask for how many clusters - it computes n = 1 … N clusters

You decide how many clusters you'd like: 2, 3, 4? …



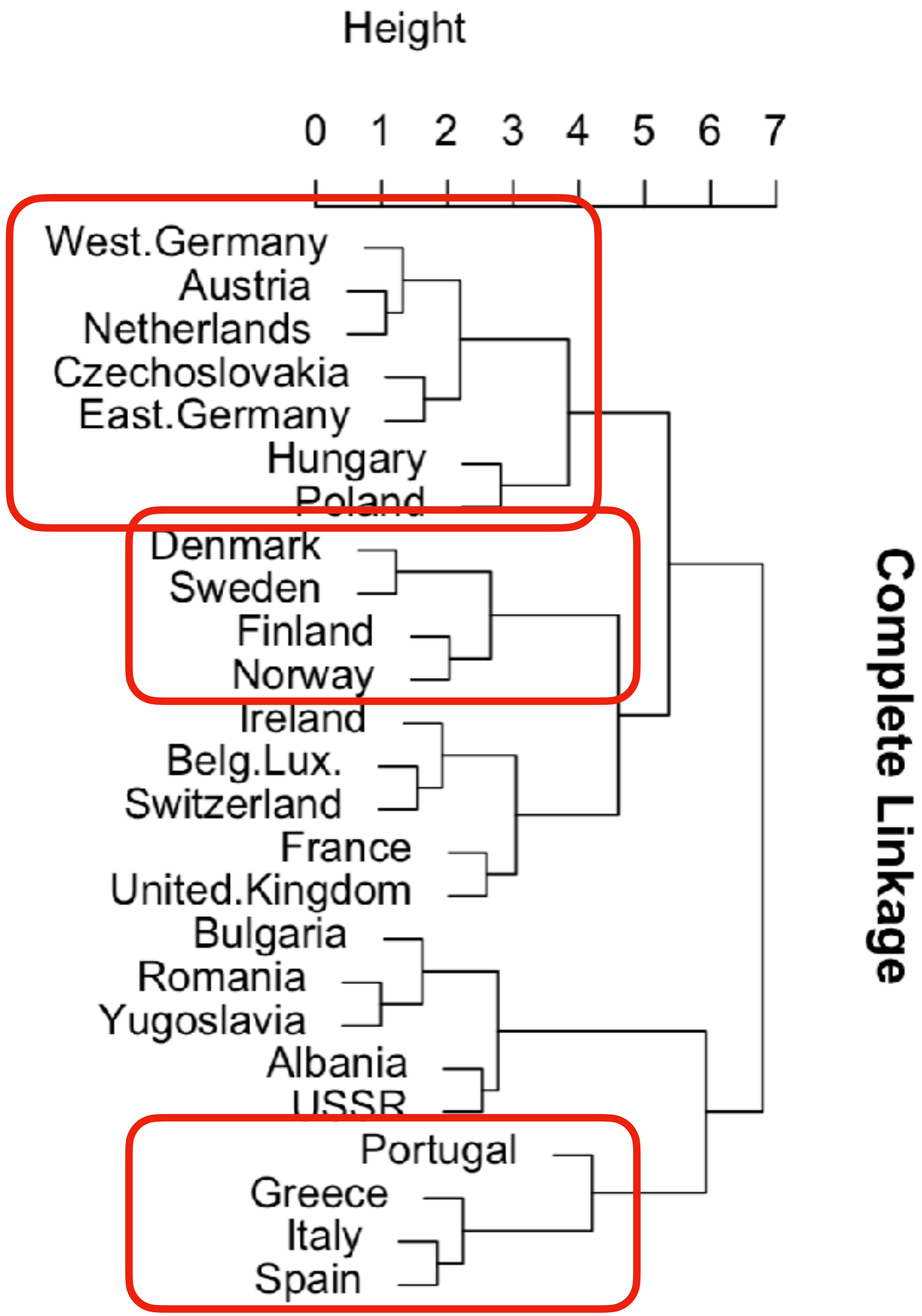
Hierarchical clustering

Overview

Start with **each point** in its own cluster (bottom-up)
Identify the **closest two clusters** and merge them
Repeat
Ends when all points are in a **single cluster**

You do not need to explicitly ask for how many clusters - it computes n = 1 … N clusters

(Food.txt example: hierarchical clustering algorithm can identify clusters of dietary preference by geographical regions)



Hierarchical clustering

Linkage

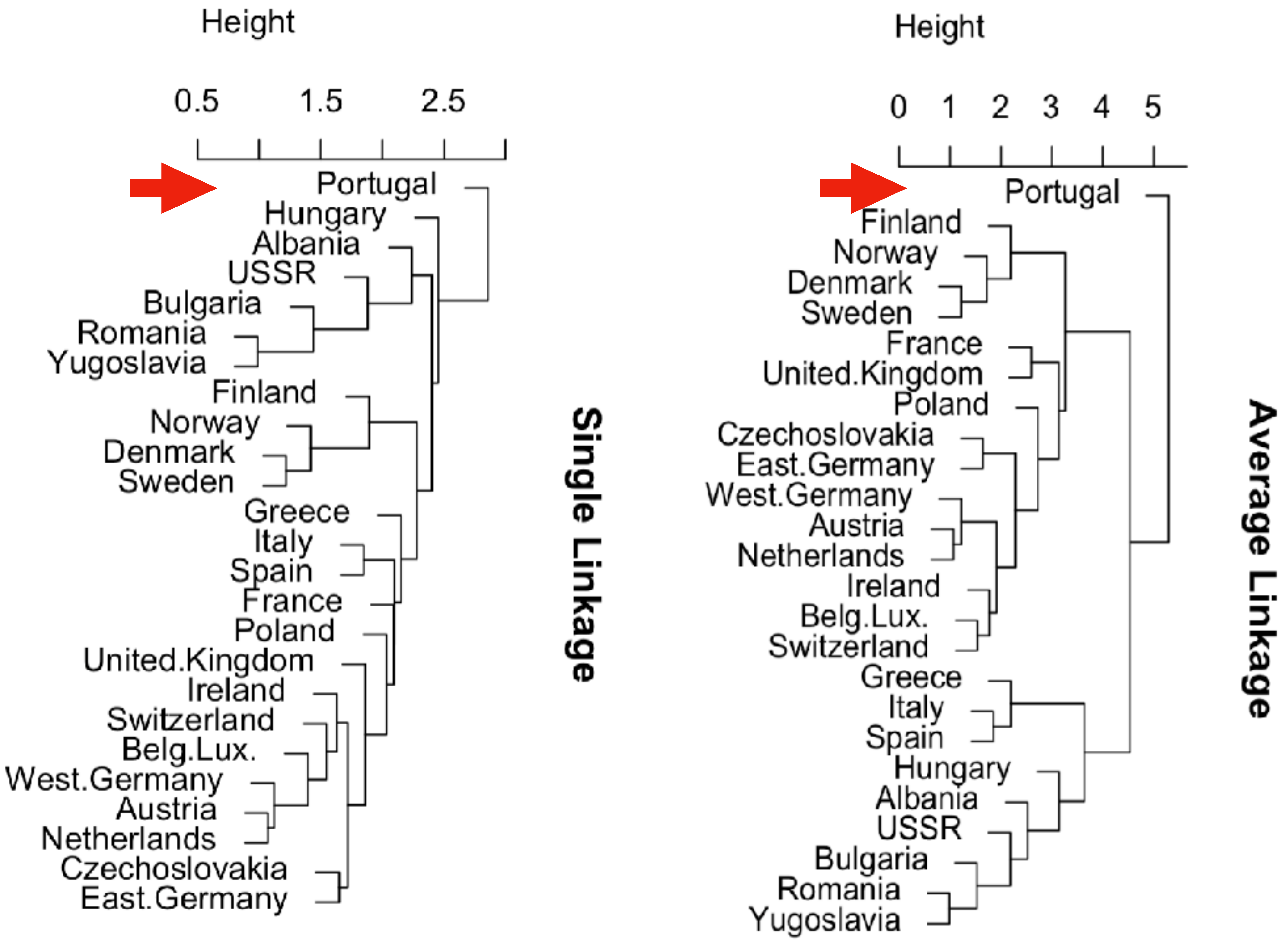
Linkage options: ways to compute the distance between clusters (rather than any two points)

Complete

Single

Average

Each has its own pros and cons, try different ones to see which is better



Hierarchical clustering

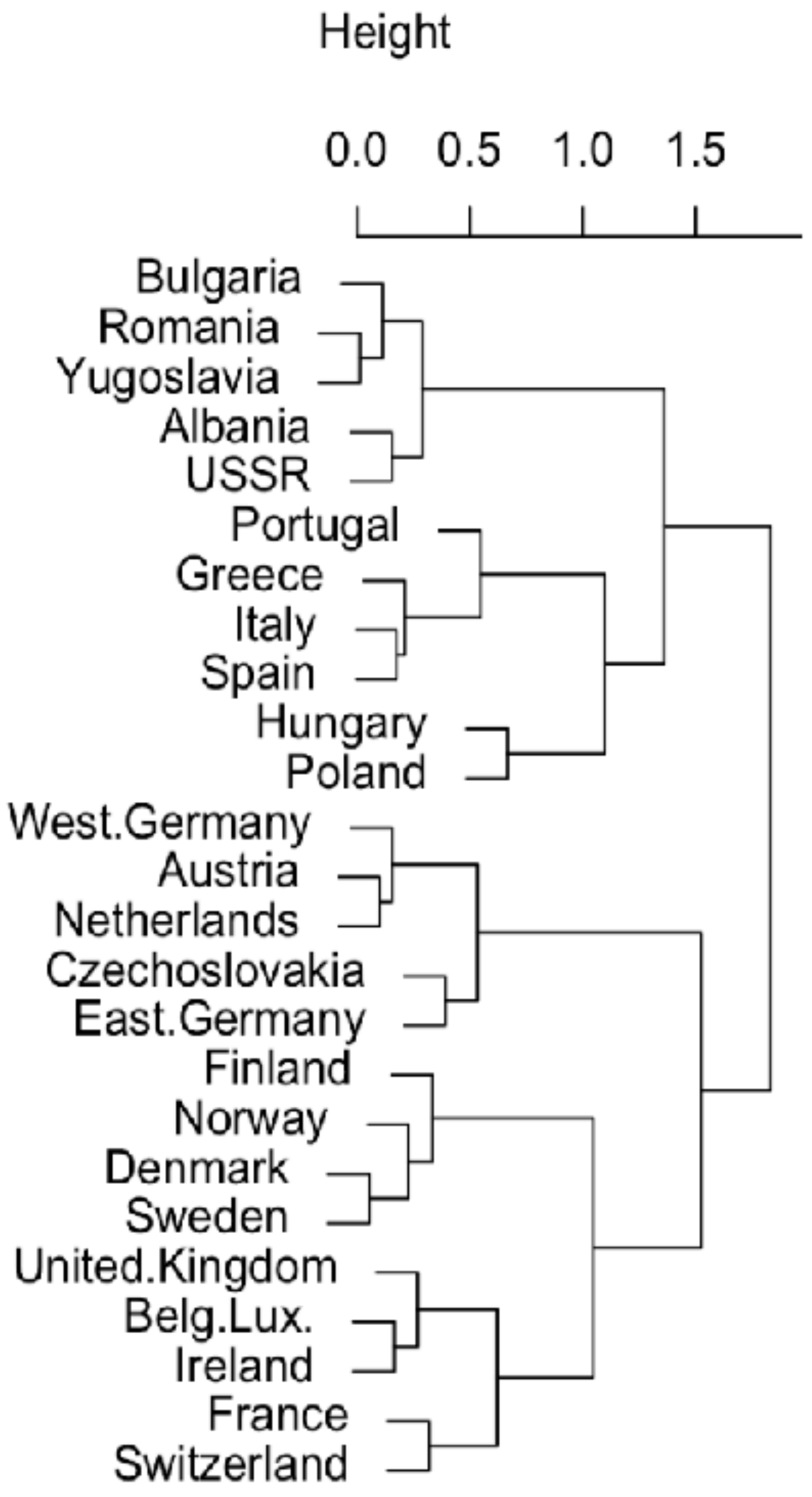
Dissimilarity, scaling

Dissimilarity: how each pair of subjects differ.

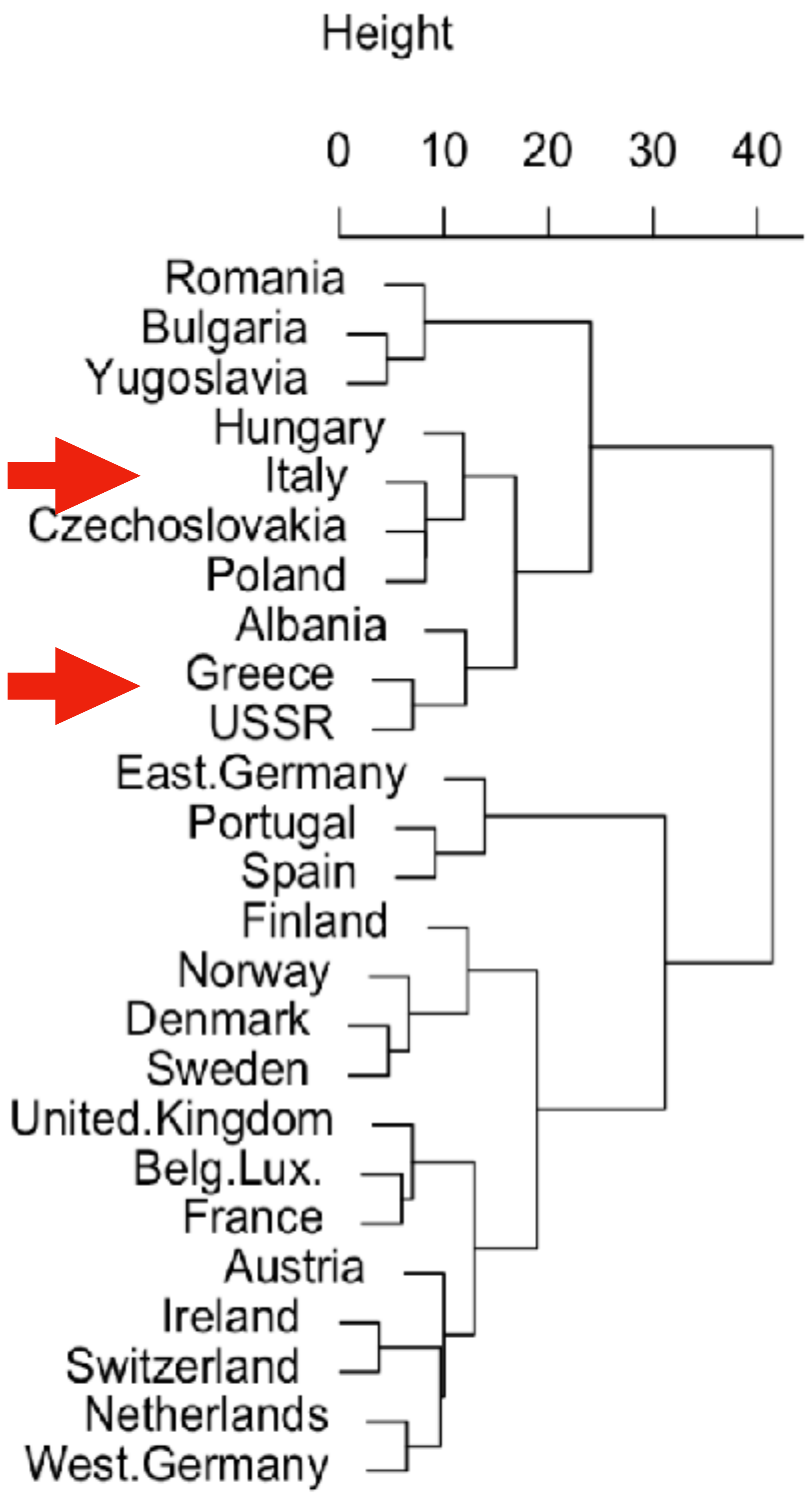
Euclidean distance? Correlation based distance? Or else

Scaling: whether we pre-process the data to have mean 0 and variance 1

Could affect the result if data columns have very different variances (like the food example)



Complete linkage with correlation-based distance



Complete linkage with unscaled features

Heatmap

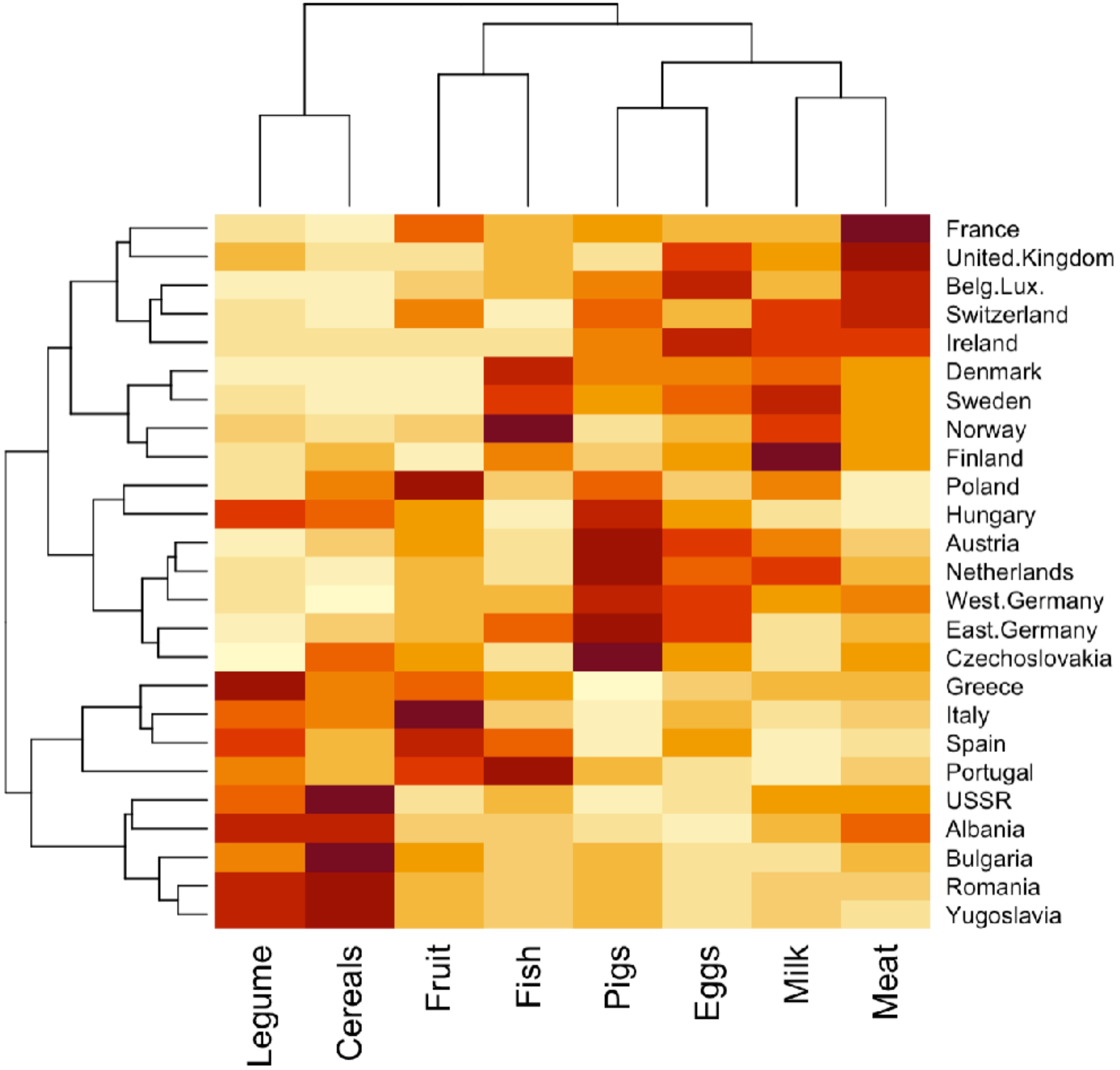
Overview

Heatmap visualizes similar values with similar shades of color

```
heatmap(data)
```

By default, this command carries out **hierarchical clustering** on both columns and rows

But you can choose to not do it (by setting some arguments). See exercise code



Heatmap

Overview

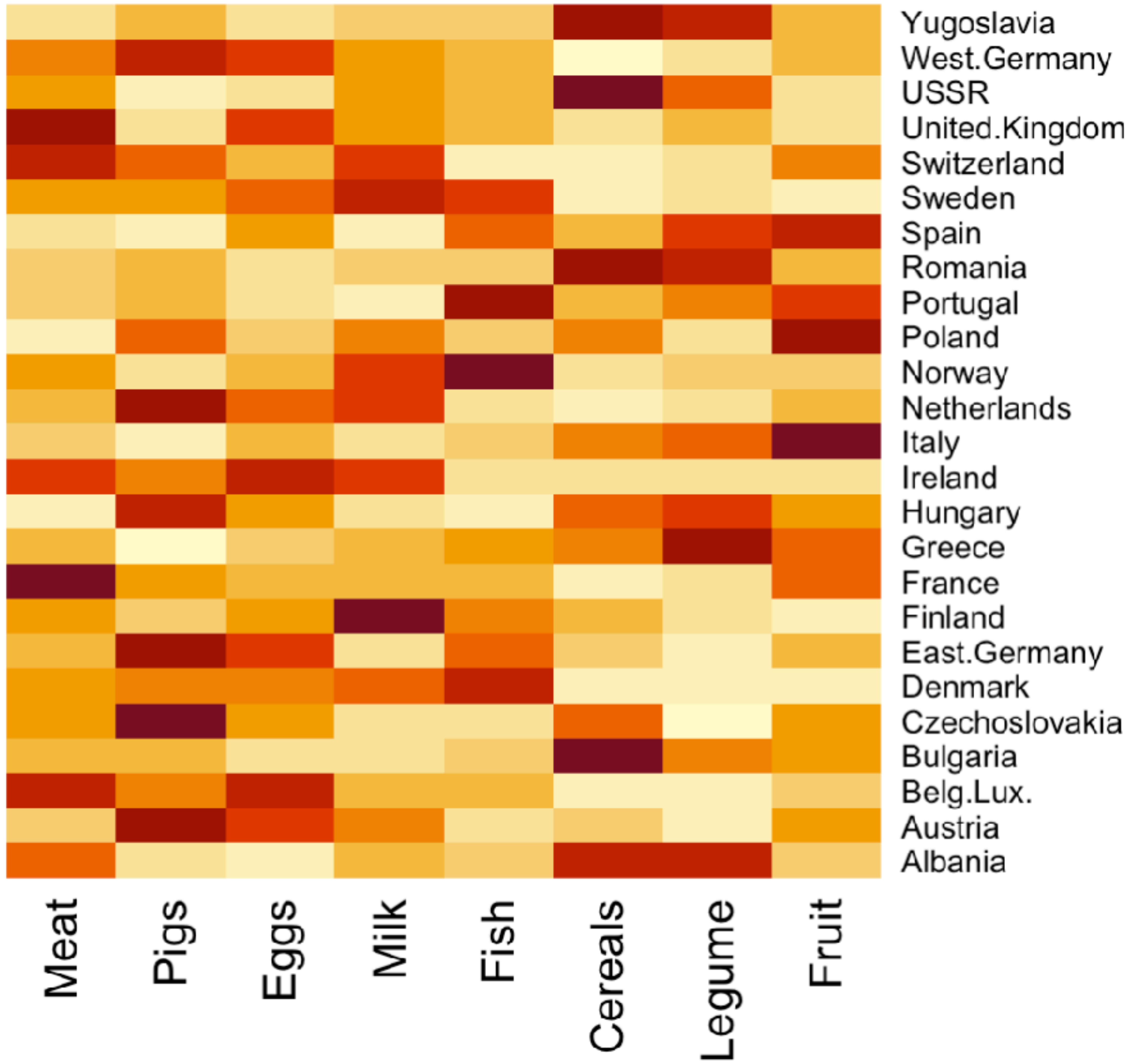
Heatmap visualizes similar values with similar shades of color

```
heatmap(data)
```

By default, this command carries out **hierarchical clustering** on both columns and rows

But you can choose to not do it (by setting some arguments). See exercise code

(Original order of colume and row names)



K-means clustering

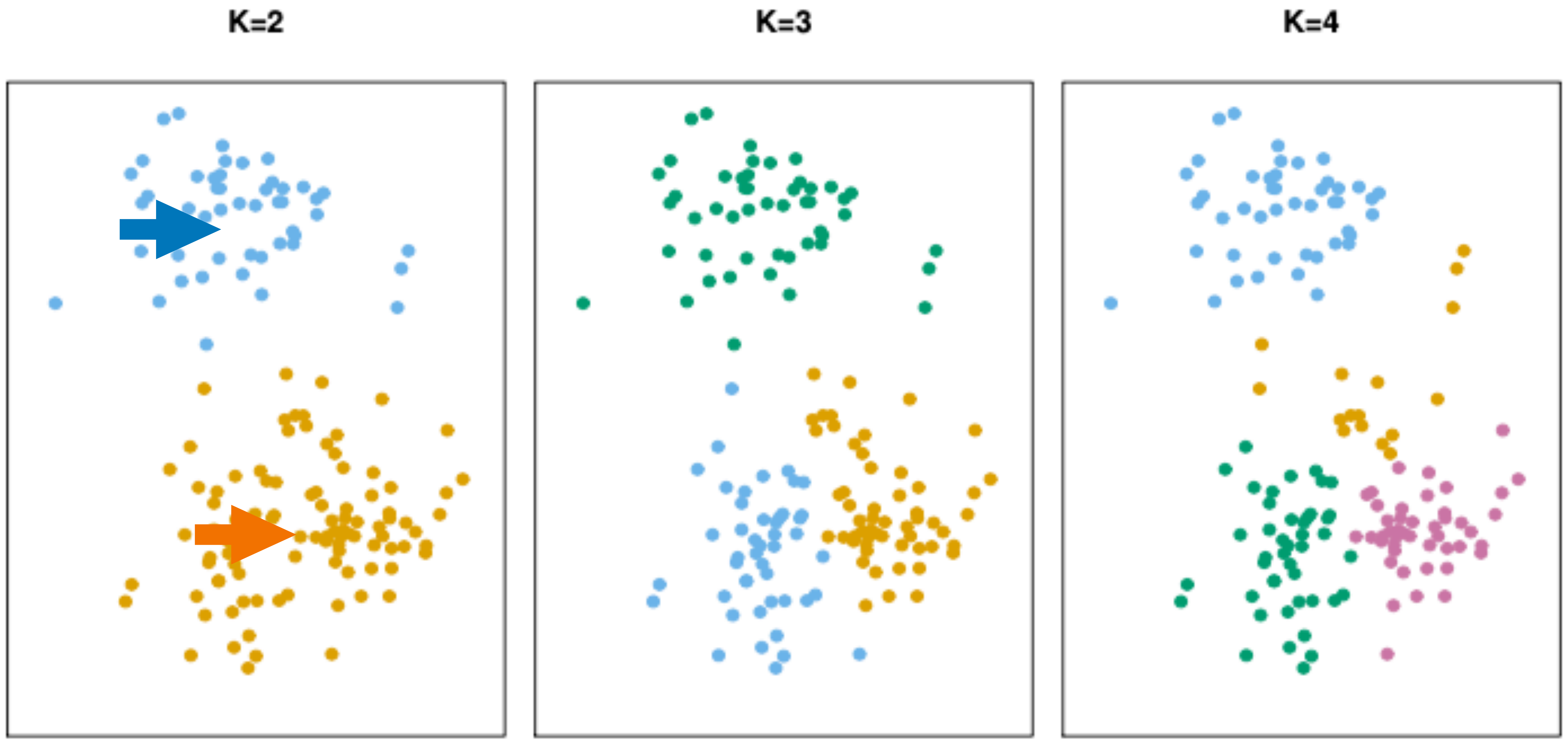
Overview

Partitions the data in **K clusters**, each data point belongs to the cluster wih the nearest mean (center, **centroid**)

Centroid can be thought of as the center of the data cloud

Need to explictly tell the algorithm how many clusters to compute

The results from K-means can be compared with hierarchical clustering, to see if the clusters agree



(Figure 12.7 ISLR book)

NCI 60 example

Hierarchical clustering

64 cancer cell lines, 6830 gene expression measurements

Ignore the cancer types, as clustering is unsupervised - but we can check how well the clustering corresponds to the true label.

Goal: find out whether observations (data) cluster into distict types of cancer

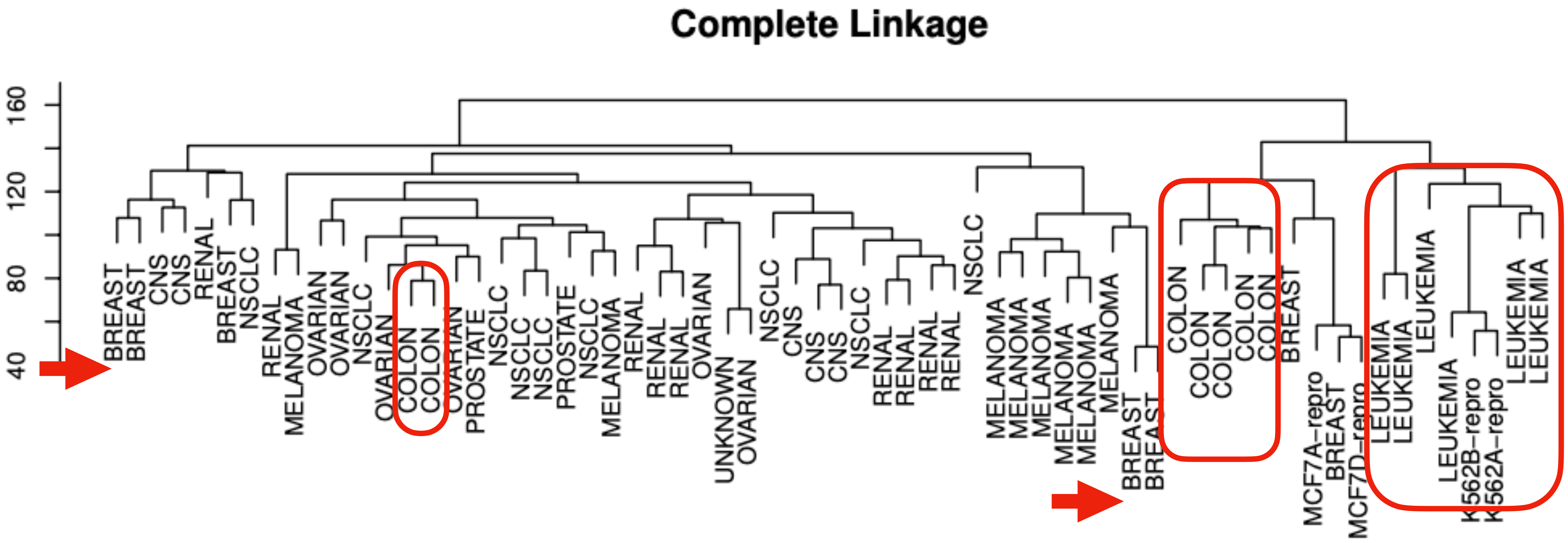
R command for hierarchical clustering:

```
hclust(distance_data, method = ‘which_method_to_use’)
```

Can also plot results into dendrogram.

NCI 60 example

Hierarchical clustering



NCI 60 example

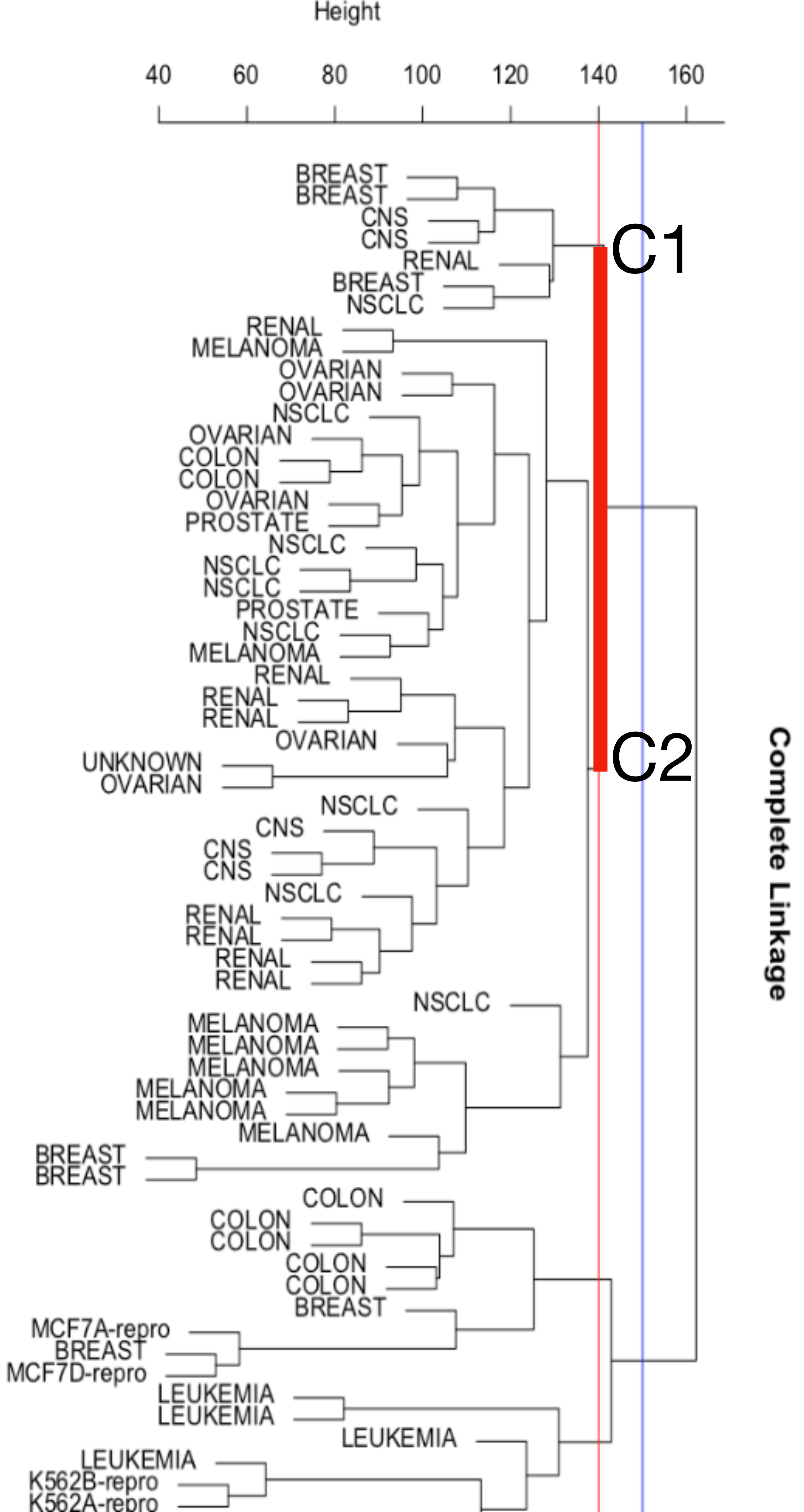
Number of clusters

All the clusters are computed (from 64 clusters - one for each data point; to 1 cluster - all data together)

'Height' in the dendrogram: essentially the **distance between clusters**

e.g. C1 and C2 are merged at around 140 - distance between C1 and C2 is 140

Based on 'height', you can decide whether you want 2, or 4 clusters (or other numbers)



NCI 60 example

Number of clusters

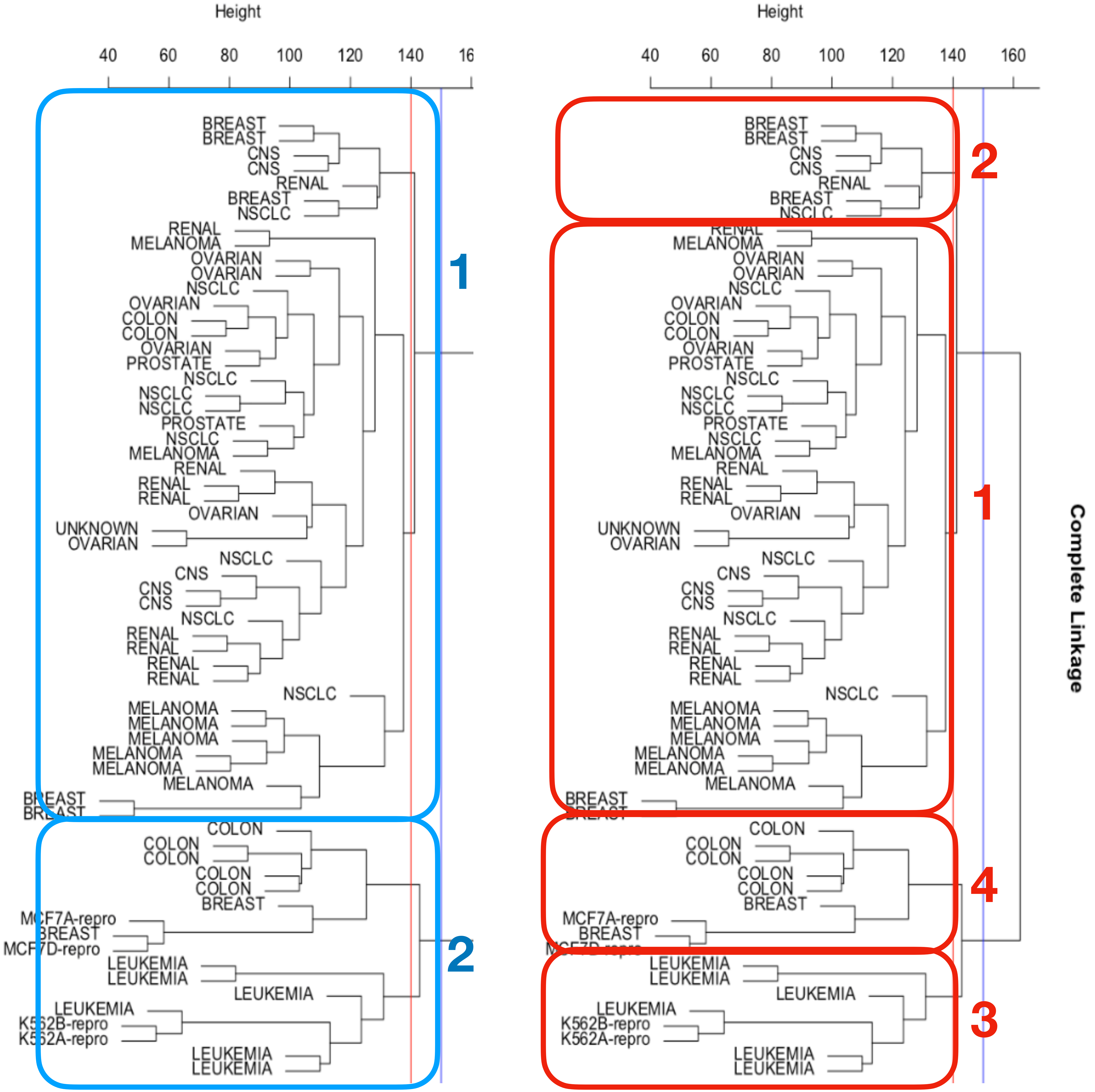
Specify how many clusters you want, and you can check which ones are further divided

```
# Compare 2 clusters and 4 clusters:
hc.clusters <- cutree(hc.complete, c(2, 4))
head(hc.clusters) # print first 6 results
```

```
   2 4
V1 1 1
V2 1 1
V3 1 1
V4 1 1
V5 1 2
V6 1 2
```

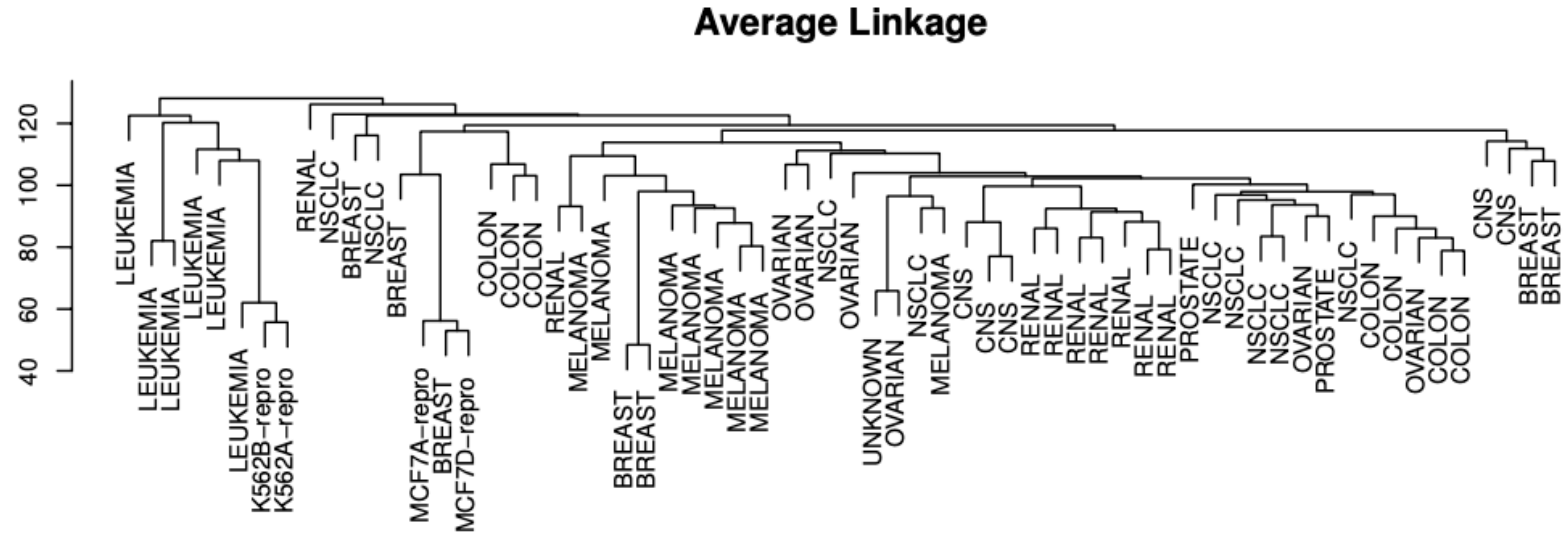
```
# cross tabulation
table(hc.clusters[,"2"], hc.clusters[,"4"])
```

```
     1  2  3  4
  1 40  7  0  0
  2  0  0  8  9
```



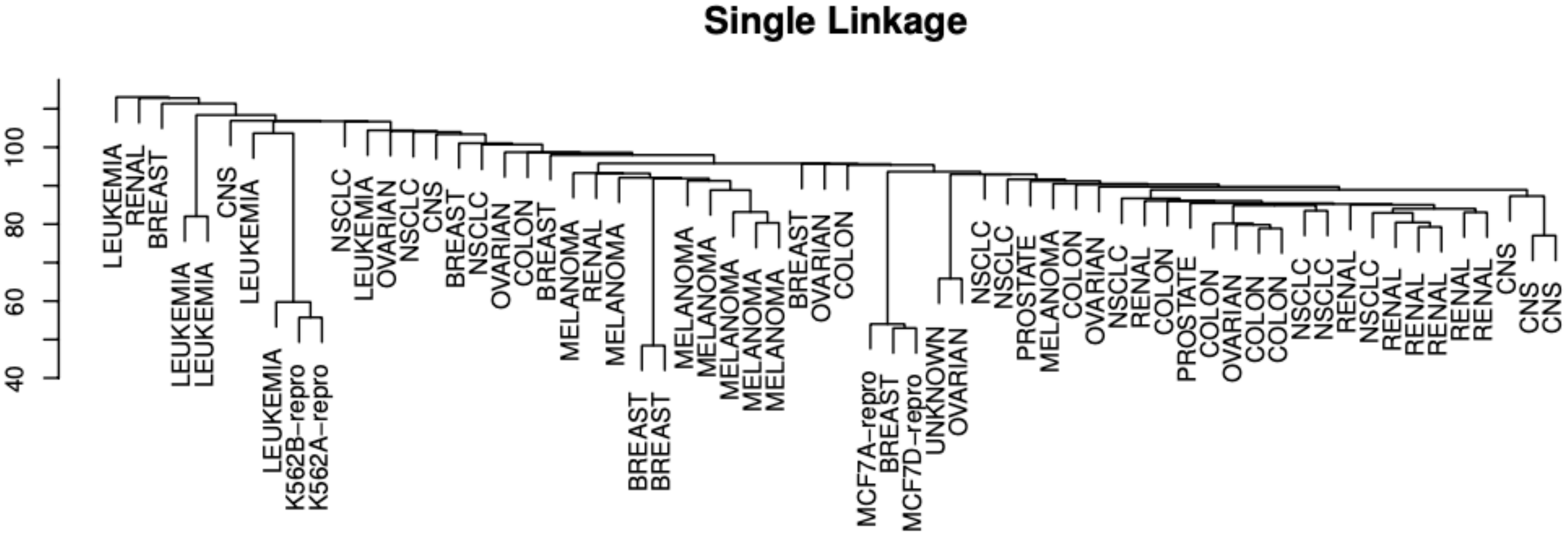
NCI 60 example

Different options for HC



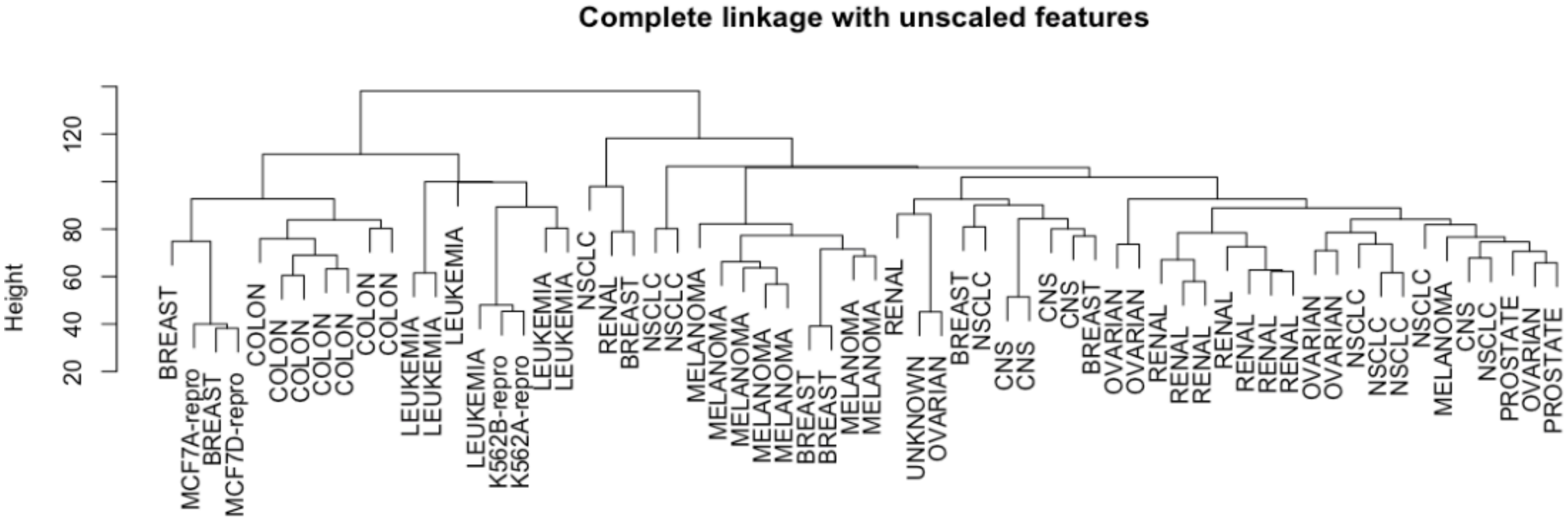
NCI 60 example

Different options for HC



NCI 60 example

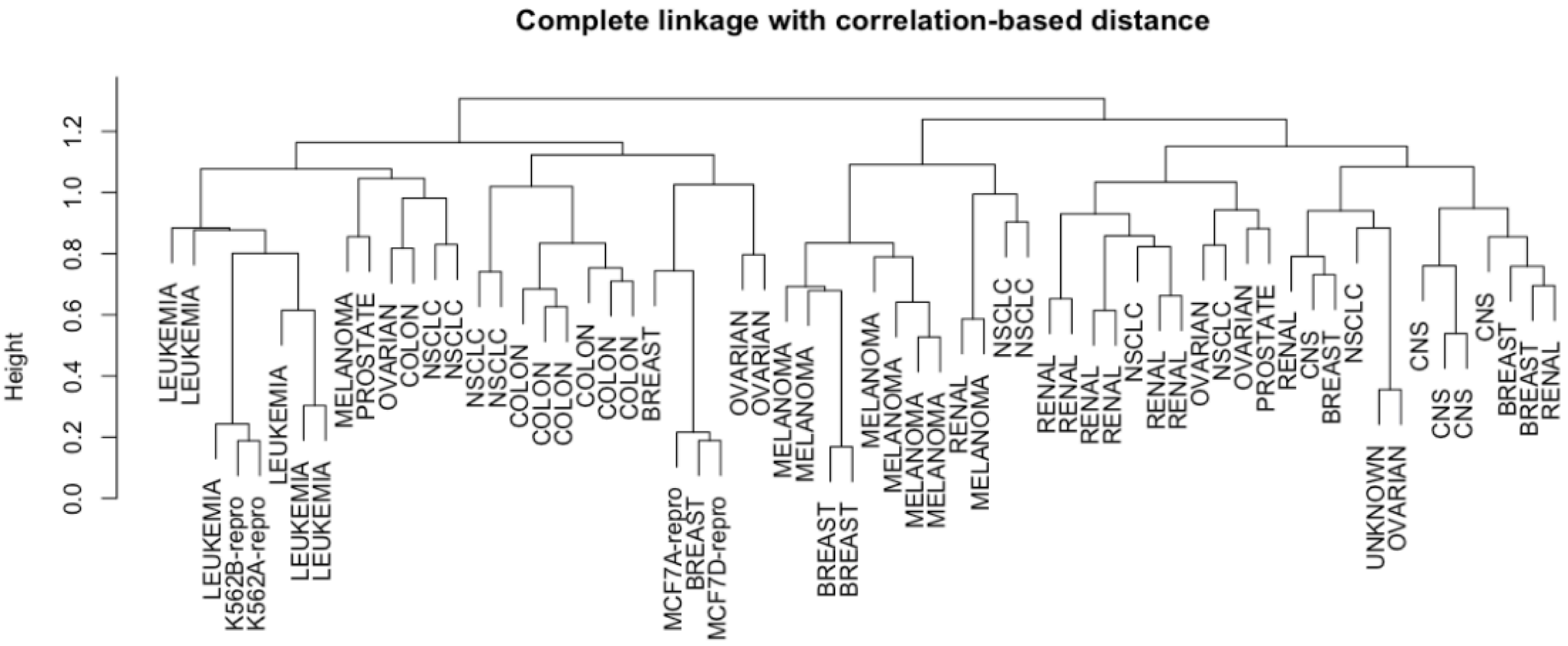
Different options for HC



Complete linkage with unscaled features

NCI 60 example

Different options for HC



NCI 60 example

K-means clustering

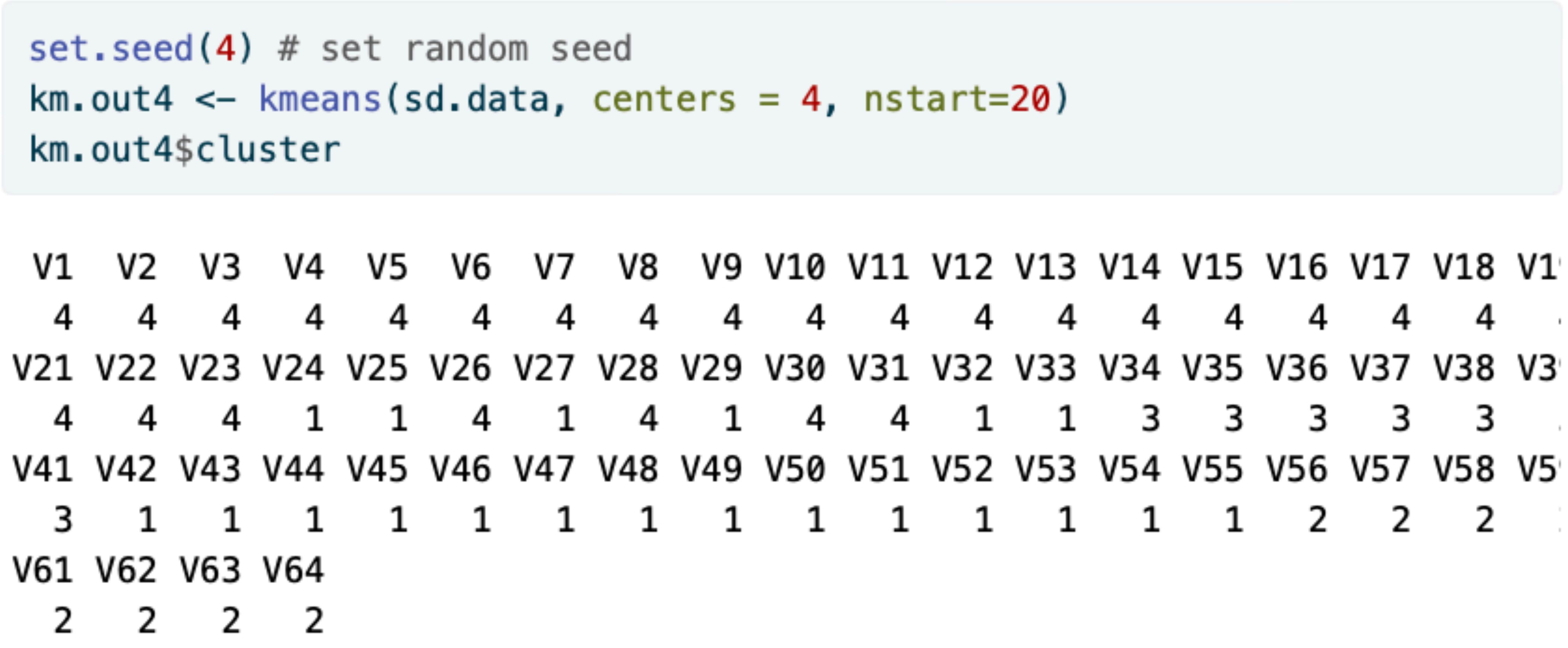
For K-means, you need to specify **number of clusters**.

With different **random seeds**, the results can be slightly different

(Change the number in `set.seed()`, for example, to 3 or 20)

```
kmeans(data, number_of_cluster)
```

Note: input is not distance (as in HC)



```
set.seed(4) # set random seed
km.out4 <- kmeans(sd.data, centers = 4, nstart=20)
km.out4$cluster
```

```
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V1
  4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4   4
V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38 V3
  4   4   4   1   1   4   1   4   1   4   4   1   1   3   3   3   3   3
V41 V42 V43 V44 V45 V46 V47 V48 V49 V50 V51 V52 V53 V54 V55 V56 V57 V58 V5
  3   1   1   1   1   1   1   1   1   1   1   1   1   1   1   2   2   2
V61 V62 V63 V64
  2   2   2   2
```

64 data points (of 6830 dimensions) are grouped into 4 clusters

The assignment (1,2,3,4) do not have meaning; just to distinguish different clusters

NCI 60 example

Visualize clusters

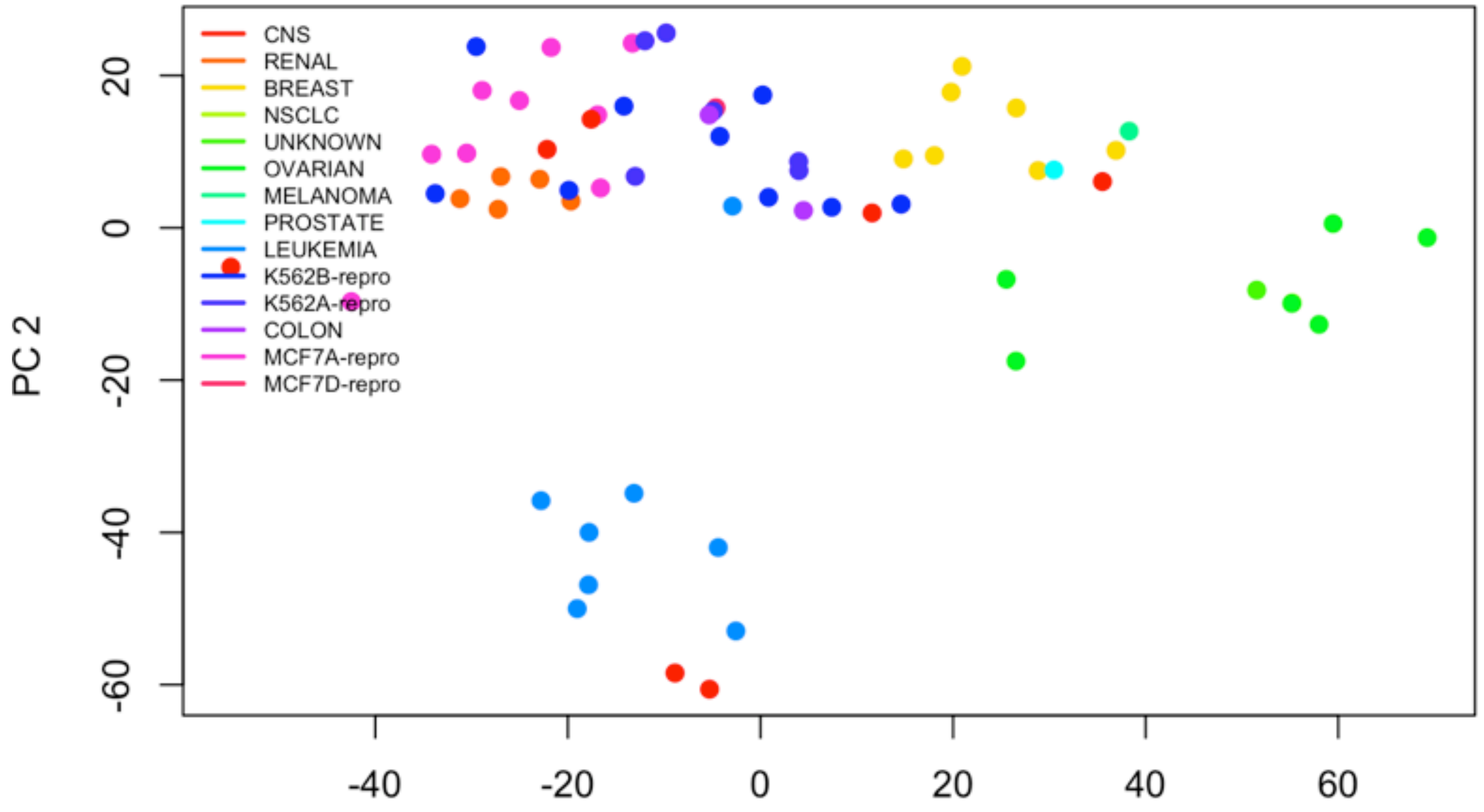
To visualize the clusters (from either K-means, or HC), you might need to combine **principal components** from PCA

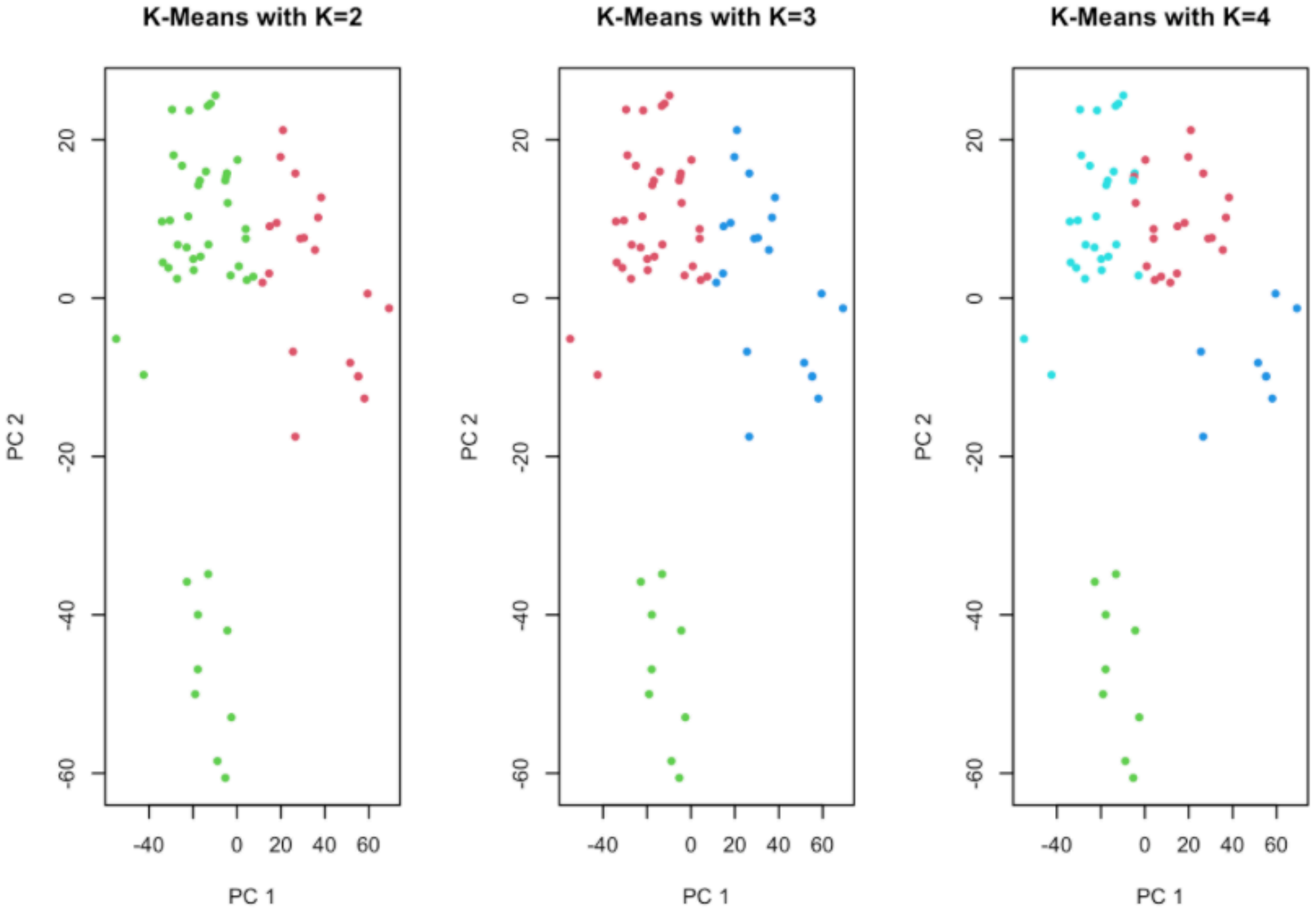
This is especially the case for high dimensional data

Use different colors to distinguish different clusters

Top right: color is the **true label** for cancer (unknown for a real unsupervised problem)

Bottom right: color is **cluster labels**


CNS
RENAL
BREAST
NSCLC
UNKNOWN
OVARIAN
MELANOMA
PROSTATE
LEUKEMIA
K562B-repro
K562A-repro
COLON
MCF7A-repro
MCF7D-repro
PC 2


K-Means with K=2
K-Means with K=3
K-Means with K=4
PC 2
PC 1