Digital Technologies and Value Creation

Dr. Philippe Blaettchen
Bayes Business School (formerly Cass)

www.bayes.city.ac.uk

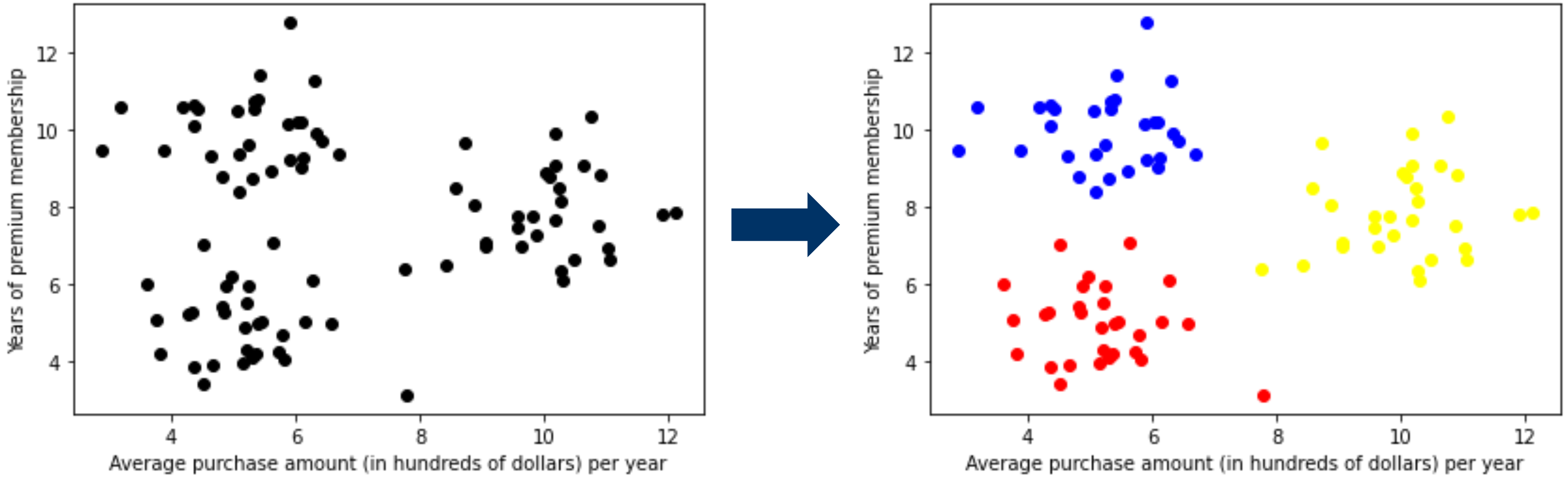
Learning objectives of today

Goals:
- Understand what clustering is
- Understand how K-means clustering works
- Understand how hierarchical clustering works

How will we do this?
- Learn the technical parts and how to use this in Python. In class, we'll make things more intuitive
- "Happiness Index" dataset: cluster countries based on some of their salient characteristics and plot on maps

What is clustering?

What is clustering



Each datapoint corresponds to a customer.

The goal of clustering is to **group together points that have “similar” characteristics**.

(Here, we could cluster to provide, e.g., specific promotional offers.)

Examples of clustering applications

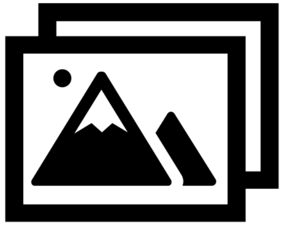
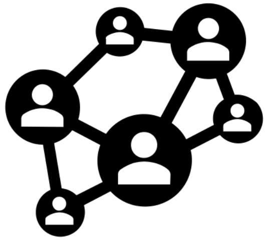


Image segmentation (e.g., background vs foreground)

Detection of cancer (finding lumps in scans)

Segmentation of customers

Credit card or insurance claims fraud



Communities in social networks

Phylogeny (clustering of species)

… and many more

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Clustering – an overview

- Clustering creates groups:
  - **Within the same group:** observations **similar**
  - **Across groups:** observations **dissimilar**

- **Similarity =** two observations are similar if the features of one observation are "close" to the features of the other

- What does "close" mean? → **Many methods for clustering:** we consider two very popular ones:
  - K-means
  - Hierarchical clustering

The happiness index

- The **World Happiness Report** is an annual publication of the United Nations Sustainable Development Solutions Network
- It contains **attributes of different countries**, feeding into one index, the happiness index.
- **Goal: cluster these countries**

	Country or region	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	Finland	1.340	1.587	0.986	0.596	0.153	0.393
1	Denmark	1.383	1.573	0.996	0.592	0.252	0.410
2	Norway	1.488	1.582	1.028	0.603	0.271	0.341
3	Iceland	1.380	1.624	1.026	0.591	0.354	0.118
4	Netherlands	1.396	1.522	0.999	0.557	0.322	0.298
...	...	...	...	...	...	...	...
140	Rwanda	0.359	0.711	0.614	0.555	0.217	0.411
141	Tanzania	0.476	0.885	0.499	0.417	0.276	0.147
142	Afghanistan	0.350	0.517	0.361	0.000	0.158	0.025
143	Central African Republic	0.026	0.000	0.105	0.225	0.235	0.035
144	South Sudan	0.306	0.575	0.295	0.010	0.202	0.091

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

K-means clustering

K-means clustering overview

Step 1:
Assign randomly each point to one of the K clusters

Step 2:
For each cluster, compute the centroid of the cluster, i.e., the average of all the points in the cluster

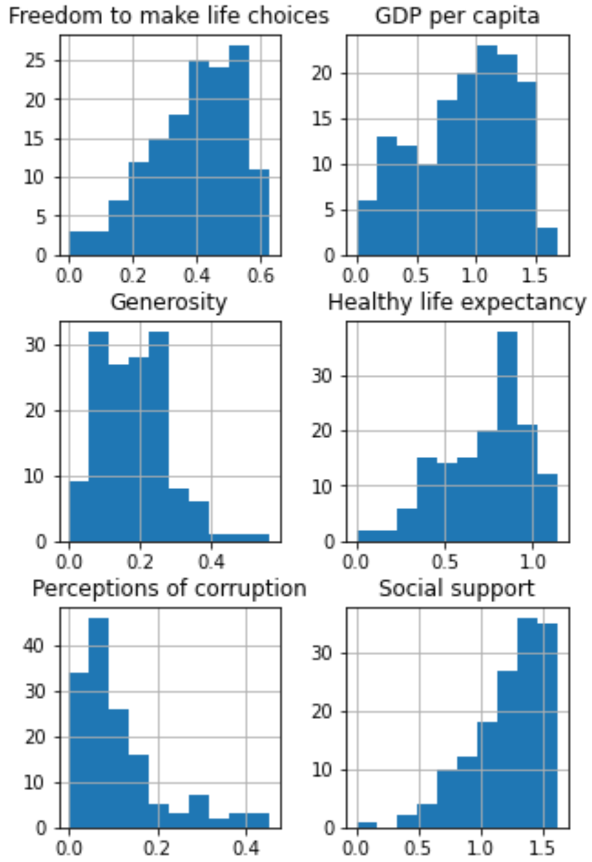
Step 3:
Reassign the clusters based on how close each point is to the centroid

Reiterate until the cluster assignments stop moving or the centroids nearly stop moving

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

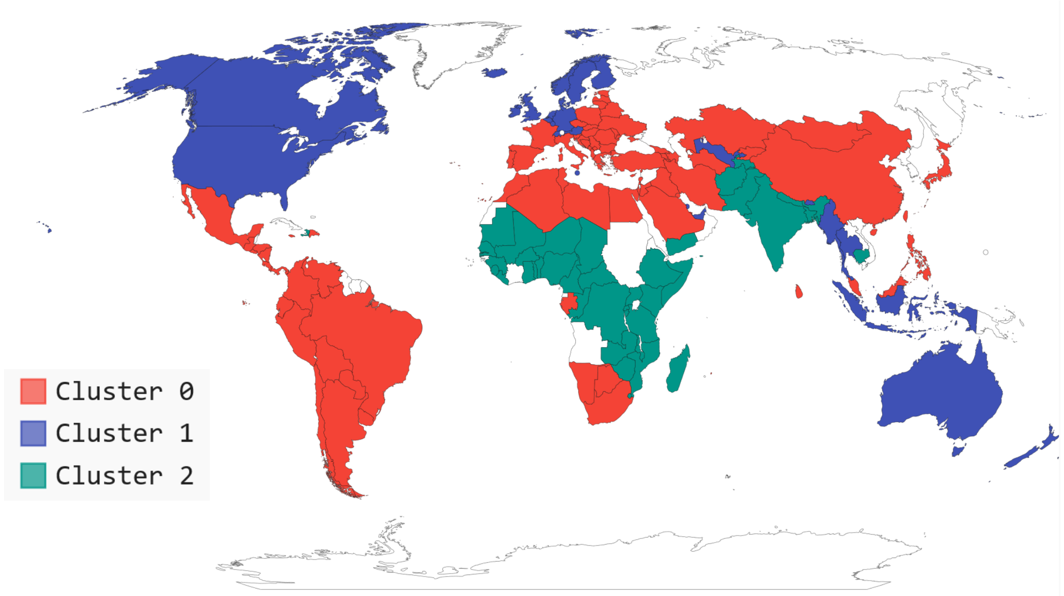
K-means clustering in Python

The importance of scaling



Without scaling, clustering is affected…

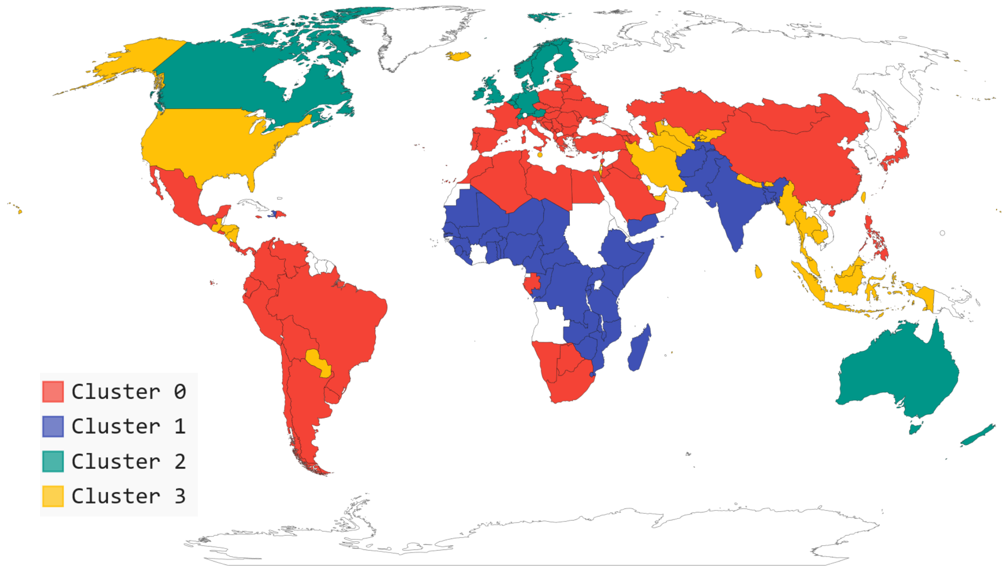
- More by GDP/ Life expectancy/Social support
- Less by generosity/corruption/freedom

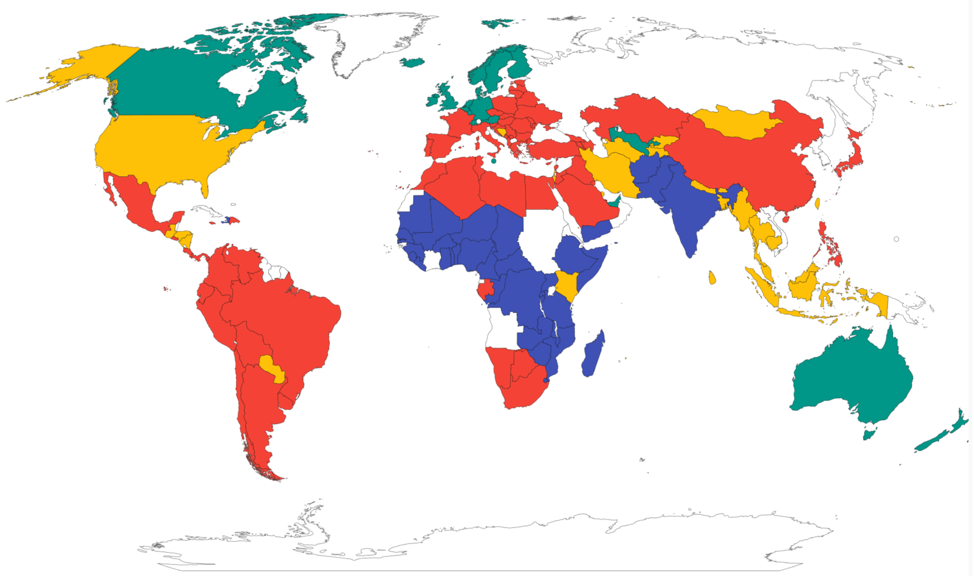


	Country or region	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Code
27	Saudi Arabia	1.403	1.357	0.795	0.439	0.08	0.132	sa
14	United Kingdom	1.333	1.538	0.996	0.45	0.348	0.278	gb

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

The impact of random sampling





Exactly the same code is run both times, but we get different outputs.

Why?

Due to **Step 1:** each country is put into a cluster at random.
When this changes, the end-result can also change.

How to mitigate this effect?

The impact of random sampling

Idea:

- Run K-means **many times** (n_init is set to 10 by default)

```
kmeans = KMeans(n_clusters=4,n_init=1).fit(happiness_quant)
```

- Among all possible models obtained, pick the **one that is best** and work with that one.

What do we mean by “best”?

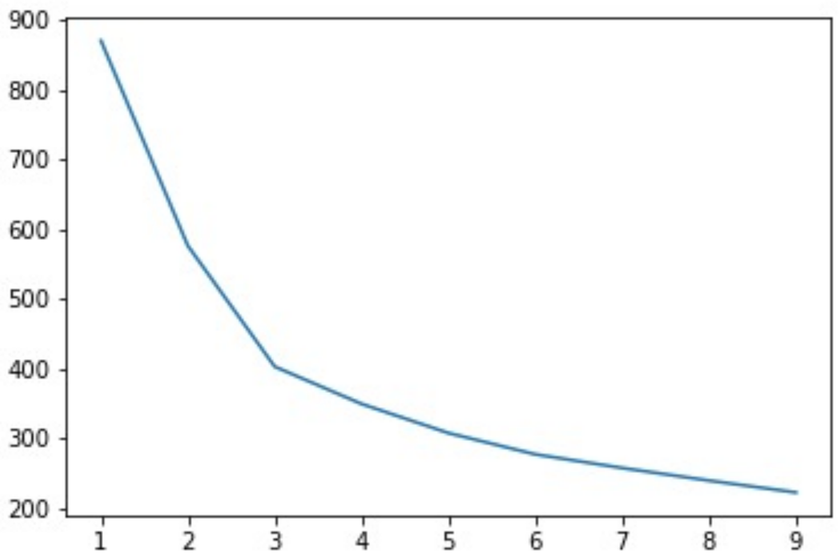
- For each cluster, compute the **distance of each point in the cluster to the centroid of the cluster.**
- **Add** all these distances **squared** together: **inertia of the model**
- The best model is the one which has the **smallest** inertia

How to pick K?

In all these examples, K was given to us. **How to pick K?**

- For different values of K, run K-means and retain the inertia each time
- Plot the inertia as a function of K: **elbow plot**

```
inertia_K=[]
K = range(1,10)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(happiness_quant)
    inertia_K.append(kmeanModel.inertia_)
```



Choose what is the "elbow" of the graph: here, K=3.

"Best return for investment"

python™

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Hierarchical clustering

Hierarchical clustering overview

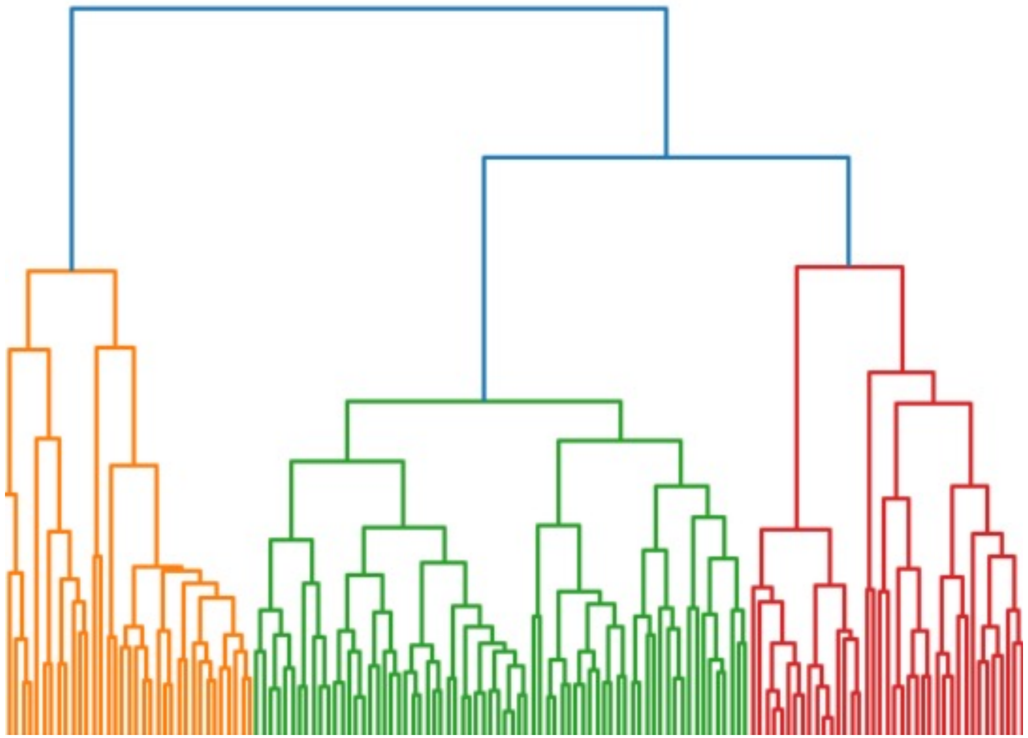
Step 1:
Make a table of the distances between all pairs of points. Each point is considered as its own cluster.

Step 2:
Fuse the two closest clusters into a new cluster.
Draw the corresponding dendrogram.

Step 3:
Recompute the pairwise distance between clusters.

Reiterate until all clusters are merged

A dendogram



Hierarchical clustering in Python

- we use scipy instead of scikit-learn (dendrogram easier to obtain)
- Scaling is important for hierarchical clustering

```
Z = linkage(happiness_quant,method='average')

dendrogram(Z) #be sure to scroll down all the way to the end to see the dendrogram
```

```
labels=fcluster(Z1, 3, criterion='maxclust')
labels
```

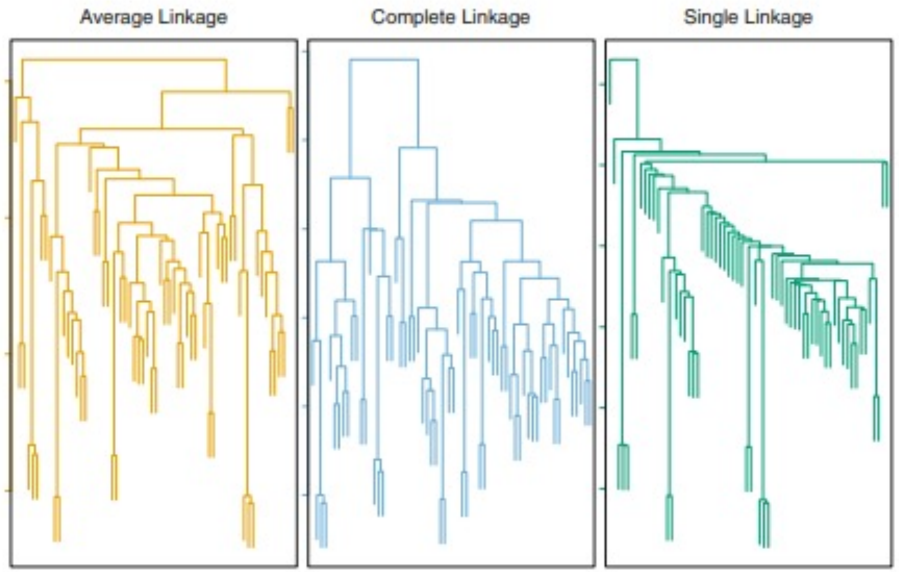
Different linkages

- We build the dendogram based on “closeness” of clusters
- It’s easy to measure the distance between two observations
- But when are clusters containing multiple observations “close” to each other?
    - Single-linkage: shortest distance between any pair of observations within the two clusters
    - Complete-linkage: distances between the farthest of any pair of observations within the two clusters
    - Average-linkage: distances between pairs of observations within the two clusters are averaged
    - Ward linkage: weighted distance between the centers of each cluster
    - …

Hierarchical clustering in Python

Different linkages give different results

Same dataset with different types of linkages can give completely different results:



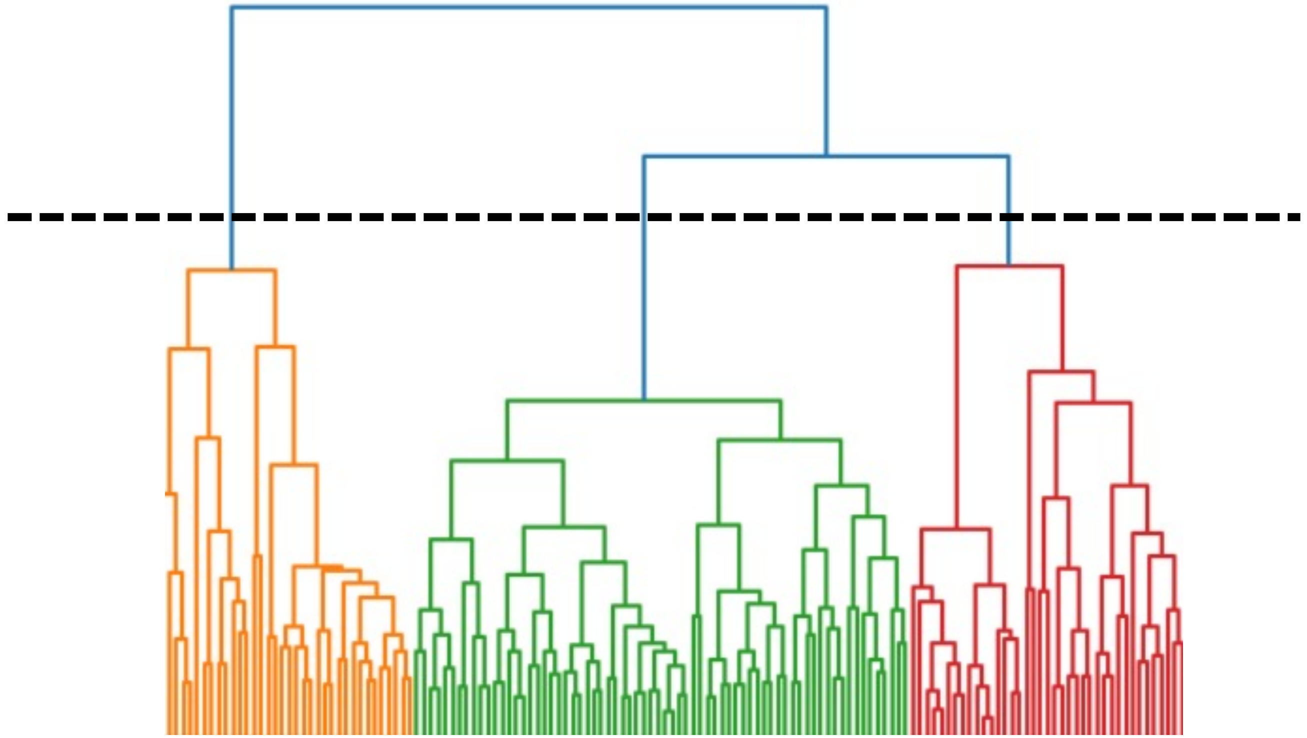
Source: James et al., 2013

In general:

- **Complete, average and Ward** linkage tend to yield **evenly sized** clusters
- **Single linkage tends** to yield extended clusters to which **single leaves** are fused one by one
- **Rule of thumb: complete or average linkage or Ward**. Go with the dendrogram.

How to get clusters from dendograms?

- Cut the final dendrogram at some level:



- In general: try and cut where there's a "jump" on the dendrogram. There may be multiple options and the best answer depends on the context.

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Further study

- Part 3 of the video-exercise notebook contains another example of clustering, this time on newspaper articles
- It also highlights the aspect of interpreting clusters

See you in class!