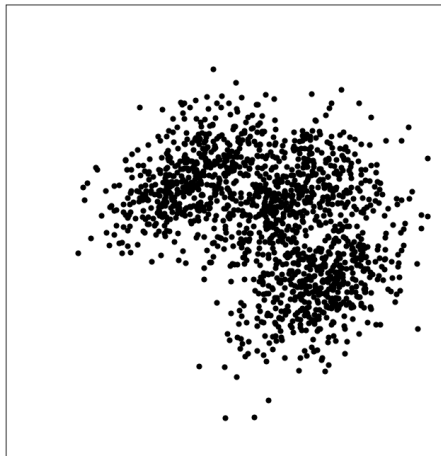# Hierarchical Clustering

Frederik Mallmann-Trenn
6CCS3AIN
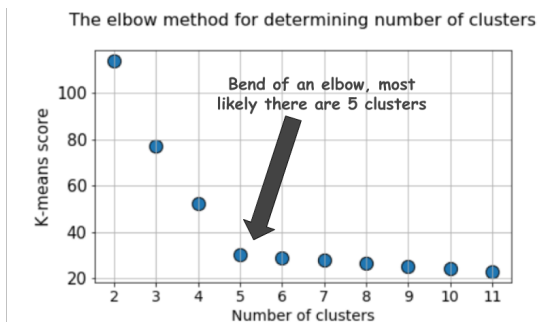
How should we choose k?

How should we choose k?
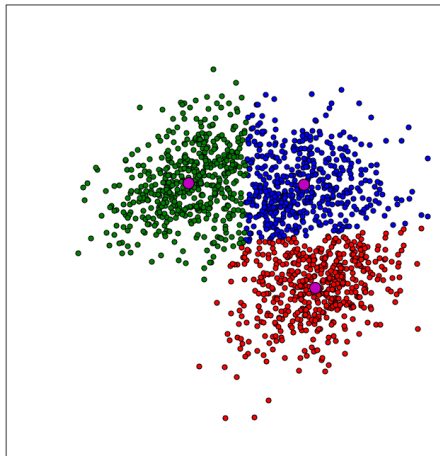
Elbow method

The elbow method for determining number of clusters

Bend of an elbow, most likely there are 5 clusters

Source: https://towardsdatascience.com/

- The more clusters the better the lower the cost/score
- Of course we could have one cluster per datapoint, but that does not really help
- One way of strike a good tradeoff is to use the elbow method

Elbow method

The elbow method for determining number of clusters

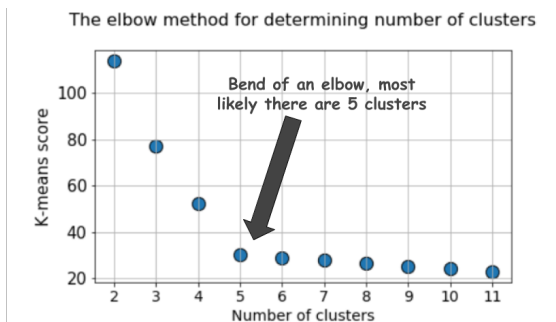Bend of an elbow, most likely there are 5 clusters

Source: https://towardsdatascience.com/

- The more clusters the better the lower the cost/score
- Of course we could have one cluster per datapoint, but that does not really help
- One way of strike a good tradeoff is to use the elbow method
- A natural definition: Pick the $k \in \{2, 3, \ldots, n - 1\}$ that maximises $cost_{k-1}/cost_k$, where $cost_j$ is the cost of the clustering with j clusters
- $k = 1$ and $k = n$ are ruled out as they result in trivial clusters

Flat Clustering

- So far we tried to assign the points to k clusters.
- We haven't assumed any structure among the clusters

Clustering

The problem with flat clustering is that it's flat

Example: Cluster the following news headlines in 3 categories

Clustering

The problem with flat clustering is that it's flat

Example: Cluster the following news headlines in 3 categories

- Deepmind's AlphaBingo wins world championship

- Black holes swallow stars whole according to new study

- Someone didn't dope

- Researcher finally figured out the rules of cricket

Clustering

The problem with flat clustering is that it's flat

Example: Cluster the following news headlines in 3 categories

CS   • Deepmind's AlphaBingo wins world championship

Physics   • Black holes swallow stars whole according to new study

Sports   • Someone didn't dope

Sports   • Researcher finally figured out the rules of cricket

Clustering

The problem with flat clustering is that it's flat

Example: Cluster the following news headlines in 3 categories

Science
- Deepmind's AlphaBingo wins world championship

Science
- Black holes swallow stars whole according to new study

Cycling
- Someone didn't dope

Cricket
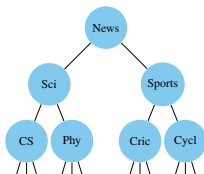- Researcher finally figured out the rules of cricket

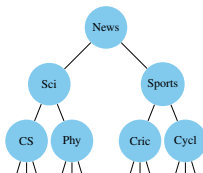Structure is lost ...

Hierarchical Clustering

- Recursive partitioning of data at increasingly finer granularity represented as a tree
- The leaves of the hierarchical cluster tree represent data.
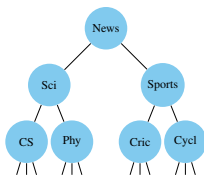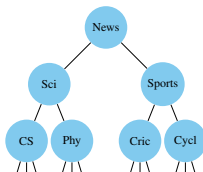
Hierarchical Clustering

Hierarchical Clustering

- Google's AlphaBingo wins world champinship

- Someone finally figured out why neural nets work

- Black holes swallow stars whole according to new study

- Someone didn't dope

- Researcher finally figured out the rules of cricket

Hierarchical Clustering

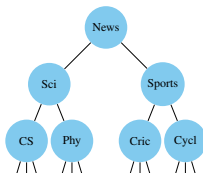Science   •Google's AlphaBingo wins world champinship

Science   •Someone finally figured out why neural nets work

Science   •Black holes swallow stars whole according to new study

Sports   •Someone didn't dope

Sports   •Researcher finally figured out the rules of cricket

Hierarchical Clustering

CS   •Google's AlphaBingo wins world champinship

CS   •Someone finally figured out why neural nets work

Physics   •Black holes swallow stars whole according to new study

Cycling   •Someone didn't dope

Cricket   •Researcher finally figured out the rules of cricket