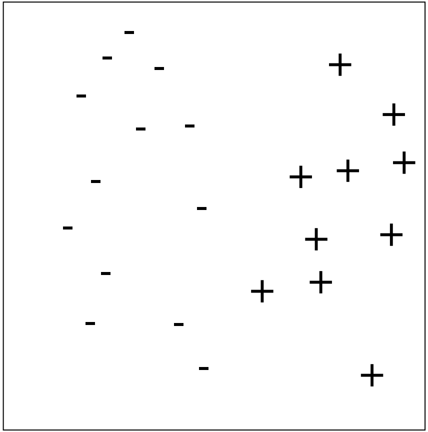
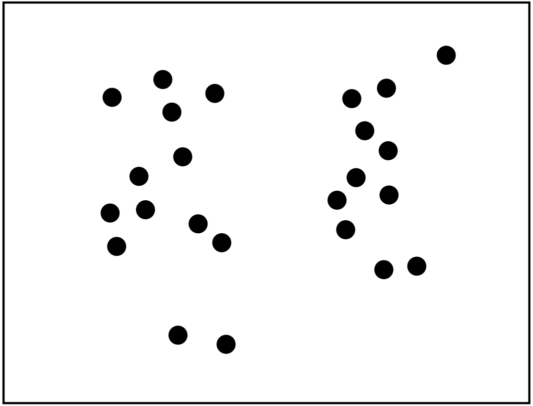
Clustering

Unsupervised learning

Instead of

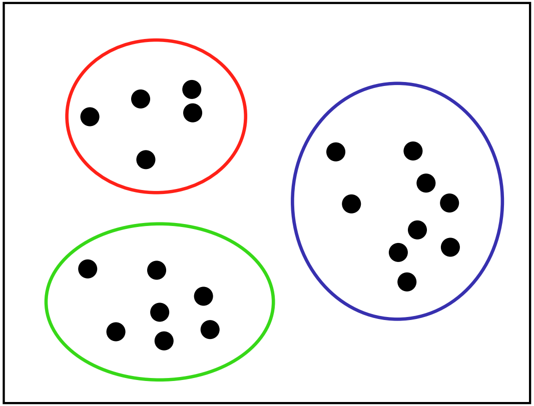


we have no label this time



Clustering

We can split data into different groups



This is called **clustering**.

Clustering

There are several ways to do clustering

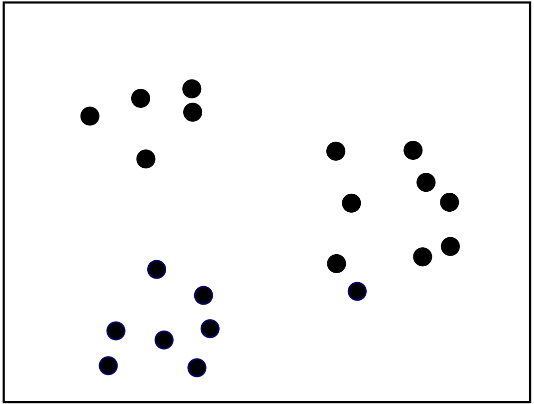
- k-mean clustering
- Gaussian mixture models
- Hierarchical clustering
- Spectral clustering

k-mean clustering

k-mean clustering

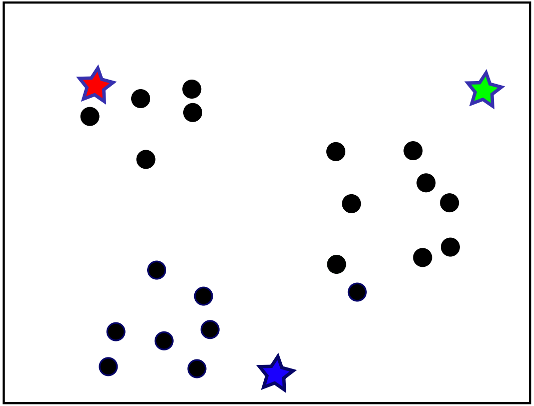
First, choose k, the number of clusters

- Find k points called **centers**
- cluster the points into k groups by the **closest centers**



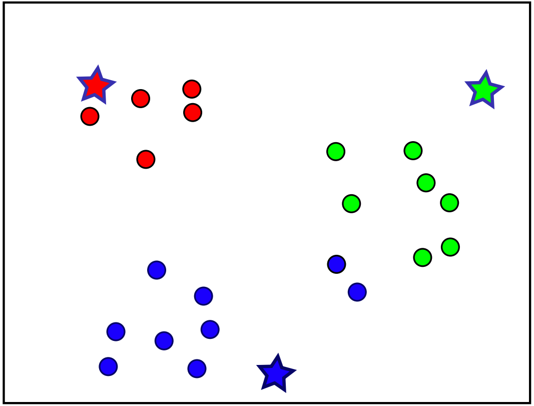
The algorithm

Randomly choosing the initial centers (in this case, $k = 3$)



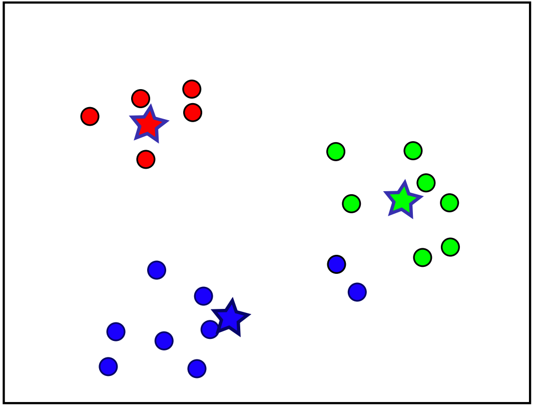
The algorithm

Assign points to their closest centers



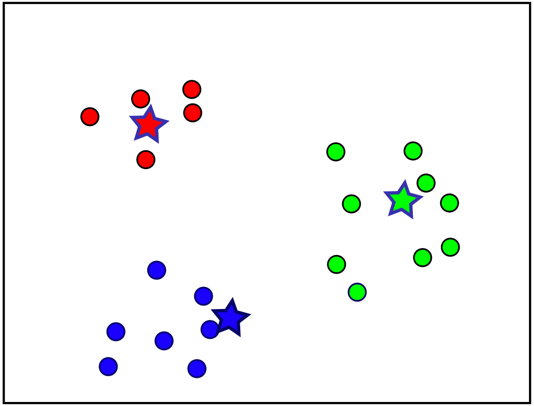
The algorithm

New centers are the averages of each color



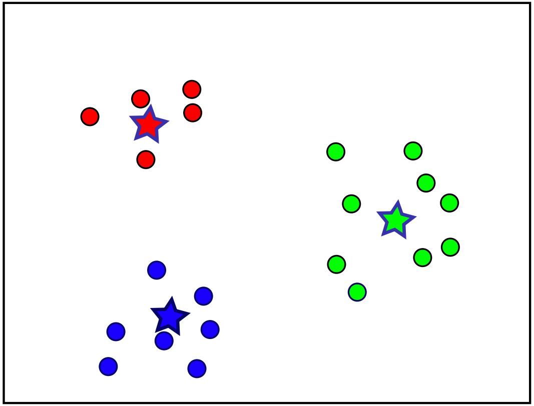
The algorithm

- Repeat until the centers stop moving



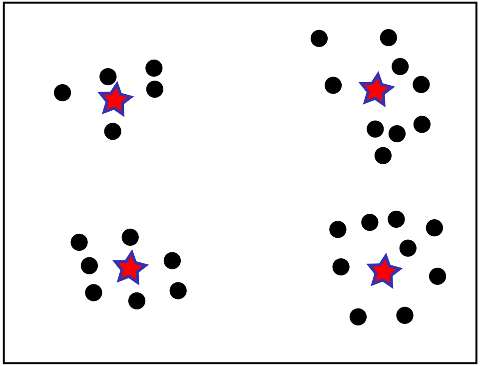
The algorithm

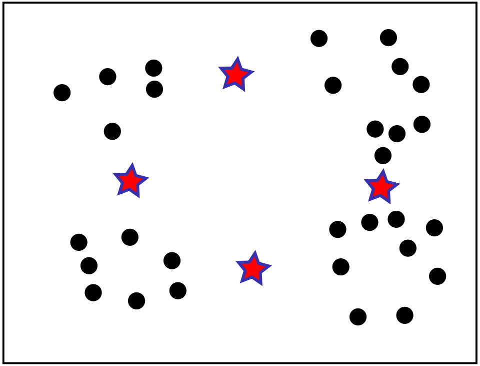
- Repeat until the centers stop moving



Initialization

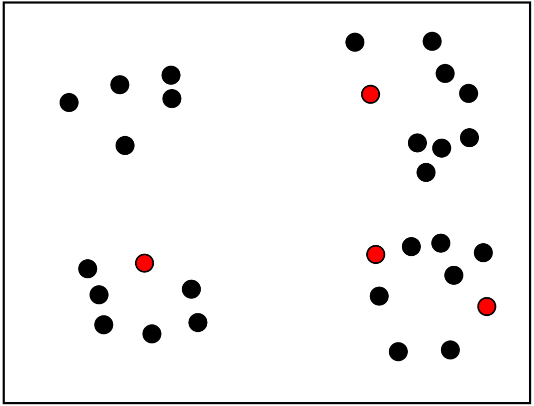
Initialization matters





Initialization

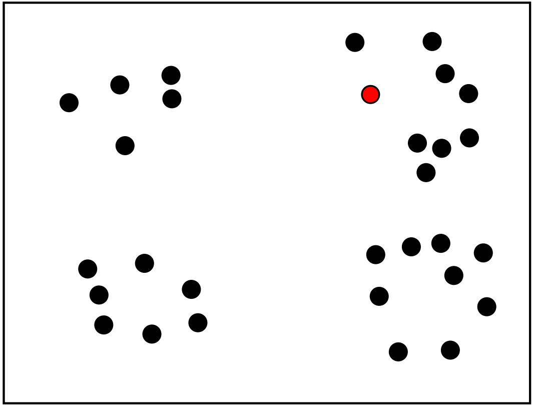
How to choose the initial centers?
Method 1: pick centers randomly



Initialization

Method 2: k-**means++** (Arthur & Vassilvitskii, 2006)

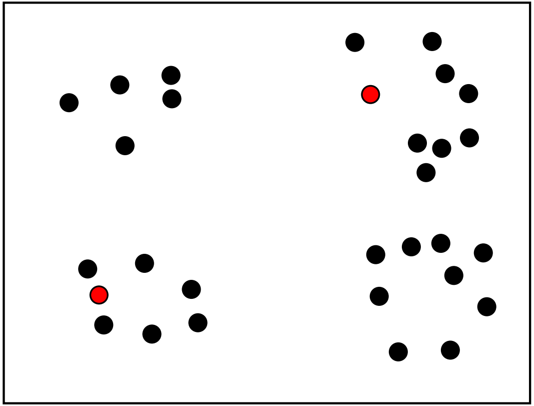
- Pick the first point randomly from the data as the first center
- Pick the next centers with **higher chance of picking a point that is far away from the previous centers**



Initialization

Method 2: k-**means++** (Arthur & Vassilvitskii, 2006)

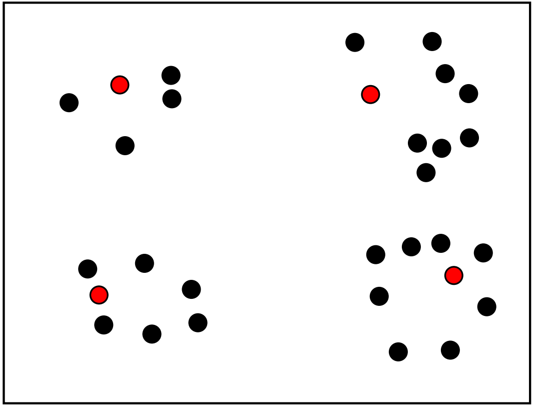
- Pick the first point randomly from the data as the first center
- Pick the next centers with **higher chance of picking a point that is far away from the previous centers**



Initialization

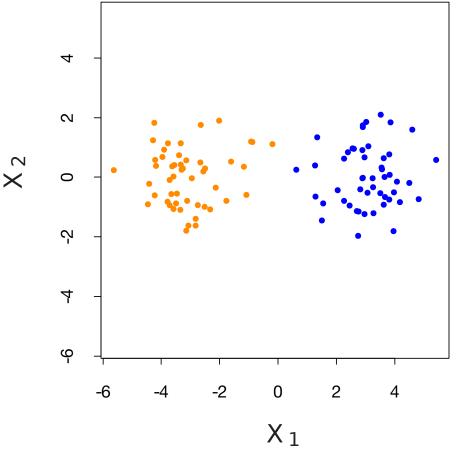
Method 2: k-**means++** (Arthur & Vassilvitskii, 2006)

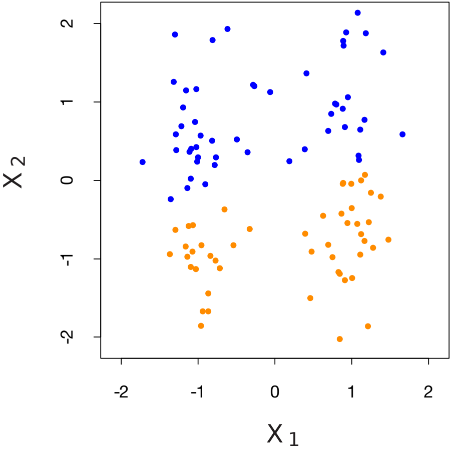
- Pick the first point randomly from the data as the first center
- Pick the next centers with **higher chance of picking a point that is far away from the previous centers**



k-mean clustering and normalization

2-mean clustering before and after normalization





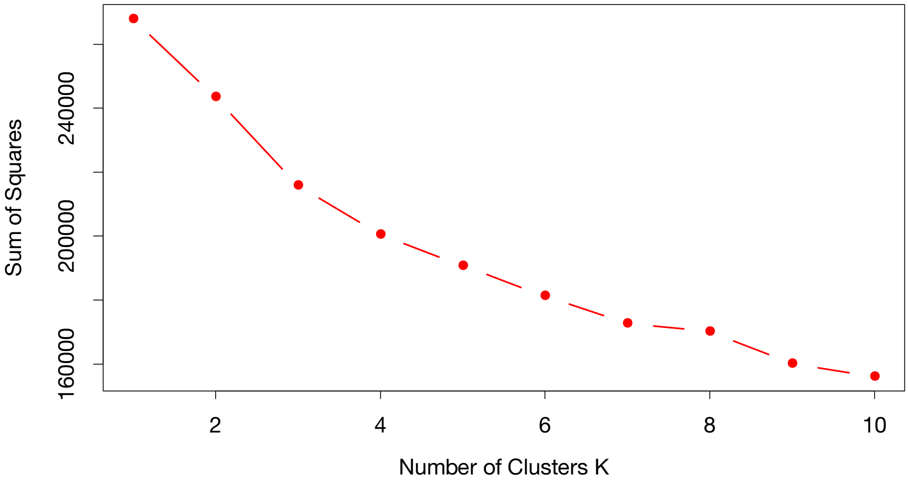
Choosing k

- Data: $x_1, x_2, \ldots, x_n$. Clusters: $C_1, C_2, \ldots, C_k$
- Centers: $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$
- Look at the **total within-cluster sum of squares**:

$$\begin{aligned} W_k &= \frac{1}{2} \sum_{\ell=1}^{k} \sum_{i,j \in C_\ell} \|x_i - x_j\|^2 \\ &= \sum_{\ell=1}^{k} |C_\ell| \sum_{i \in C_\ell} \|x_i - \bar{x}_\ell\|^2, \end{aligned}$$

where $|C_\ell|$ is the number of points in cluster C_ℓ

Plot of W as k increases



Application: Unsupervised Classification

- **Problem 1**: Want to classify the pictures of handwritten numbers, but the data has no labels (probably from budget issues...)

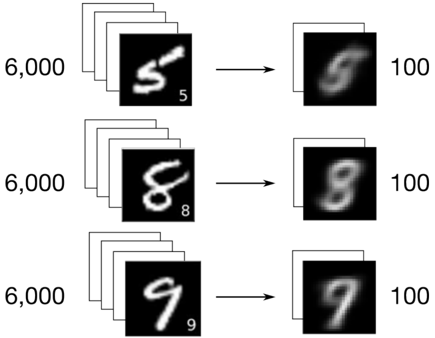
- We can do 10-mean clustering on the data.

10 centers of the clustering

Application: Learning with time/memory constraint

- **Problem 2**: We have 60,000 images with labels, but it's taking too long to train all of them (for example SVM with RBF kernel requires computing $\approx 60000^2$ pairwise distances!)

- We can instead train on 1,000-mean clustering on the data

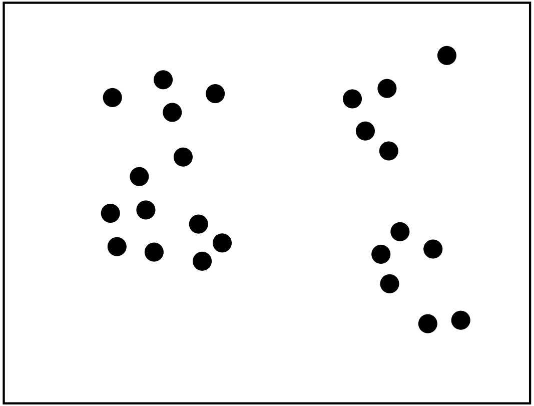


Application: Image Compression

Hierarchical clustering

Hierarchical clustering

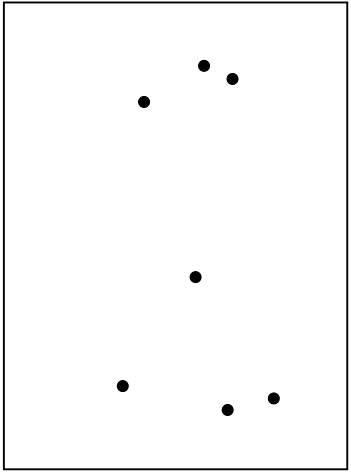
Sometimes we want to be flexible about choosing (k). For example



Cluster with $k = 2$ and 3?

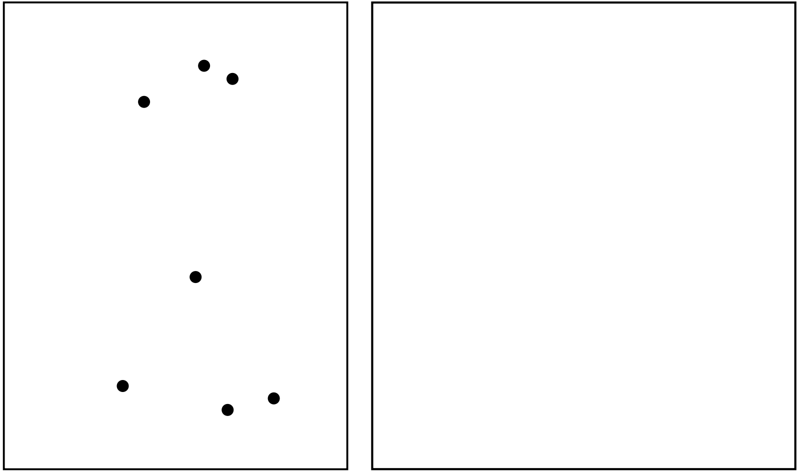
Hierarchical Clustering

Suppose we have the following data



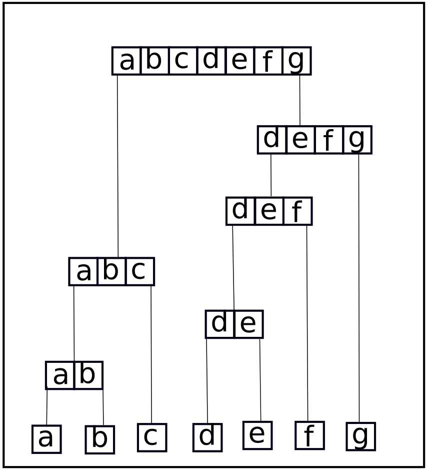
Hierarchical clustering

Step 1: Start from a closest points. Make a **dendrogram** at the same time

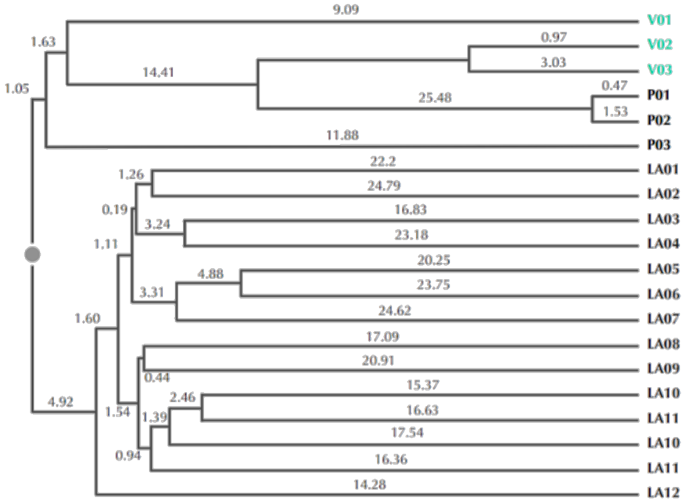


Hierarchical clustering

Step 2: Make a **cut** where you want the actual clustering



Tracking HIV outbreaks



Metzker et al. (2002), Molecular evidence of HIV-1 transmission in a criminal case

Comparison between clustering algorithms