kokchun giang

in **unsupervised learning** you find groups of similar characteristics and **cluster** them together

when there is no label you can **cluster** the data points

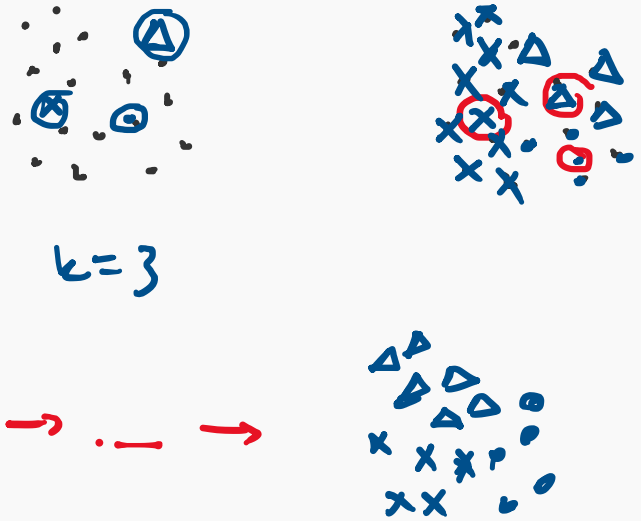
unsupervised learning - there are no labels

- more subjective
- annotation of data cost money

→ clustering

k-means clustering

1. choose k clusters
2. k cluster centers randomized
3. closest pts to centroid is classified as that centroids class
4. compute new cluster center
5. repeat 3-4 until no new class changes

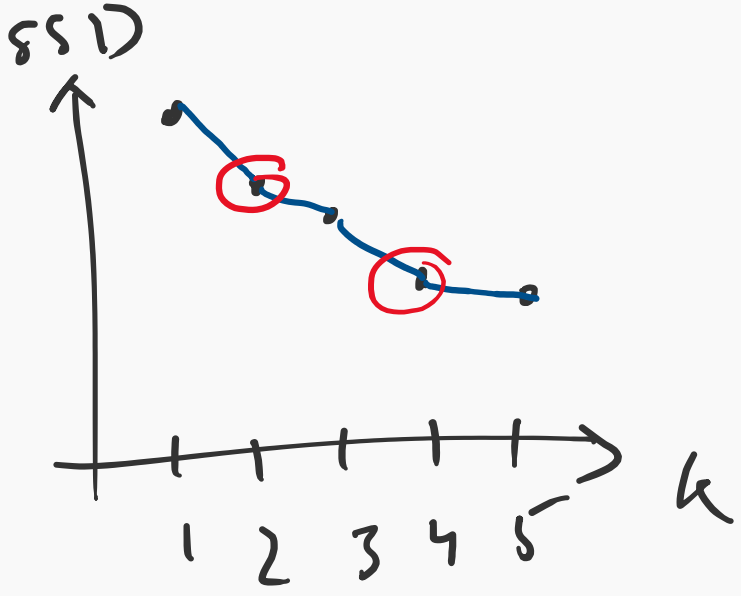


choose k through **elbow** and **silhouette**

measure SSD

- sum of squared distances to cluster center

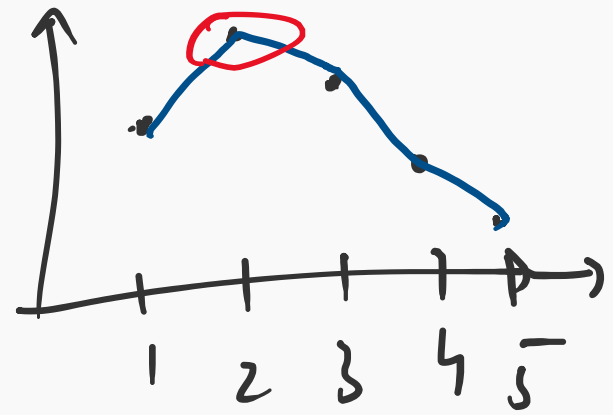
→ do elbow plot



elbows / inflexion pts

- combine w. silhouette plots

measure tightness of clusters



Combine w. domain expertise to check if reasonable & give name to cluster