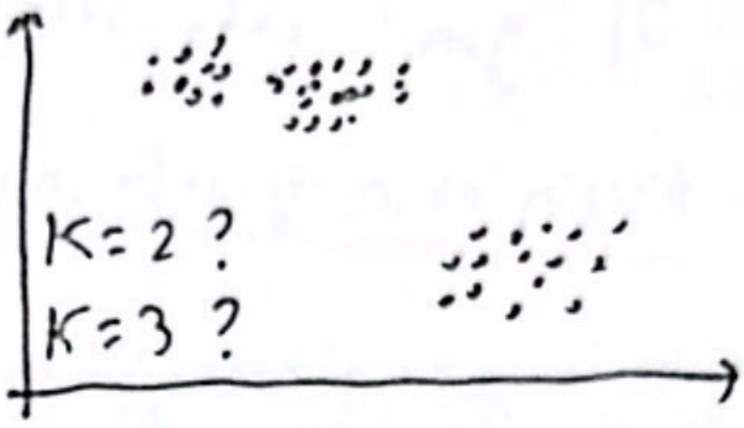
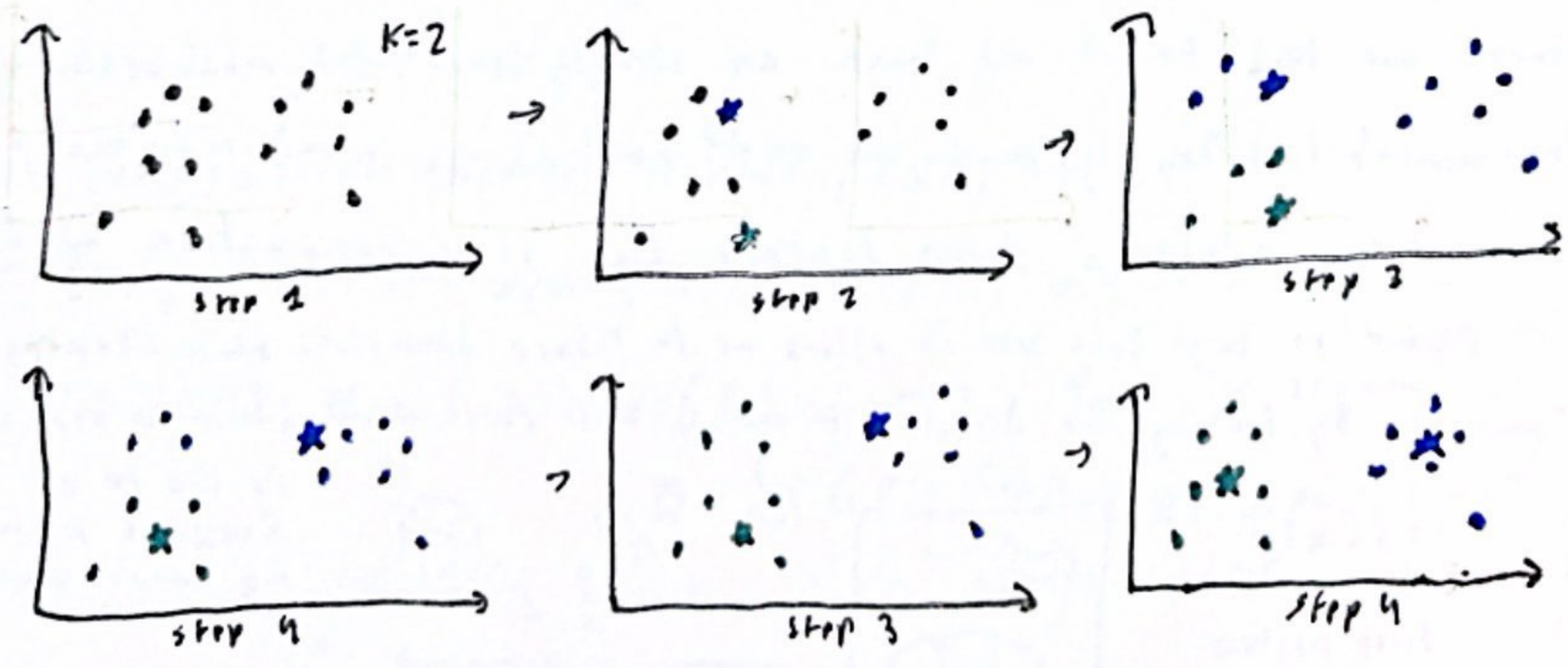
K-Means Clustering (unsupervised clustering)

- The most famous unsupervised clustering algorithm used to find clusters in data, like KNN K is a hyper parameter and stands for the number of Clusters we are looking for. finding the right number of K is a art and depends on problem and Trail and error



! (• No labels in training
• can use labels after to evaluate clusters if labels given)

- K-Means: pick K (step 2), you start by randomly selecting centers for K=2 for your K clusters (in ex K=2) (step 2). Then you assign all data points to the cluster center closest to them (step 3). You then recalculate the cluster centers ie the true center of the cluster based on datapoints assigned to them, you can see the centers moving closer to the actual clusters (step 4). you then assign data points again to new cluster centers (step 3). then you recalculate the cluster center again based on current datapoint assigned to them (step 4). you repeat until centers of clusters have stabalized meaning there is not much re centering or new centers anymore (true center found)



- Note some algos like DBScan or hierarchical clusters dont need you to specify K the find it them selfs