Clustering

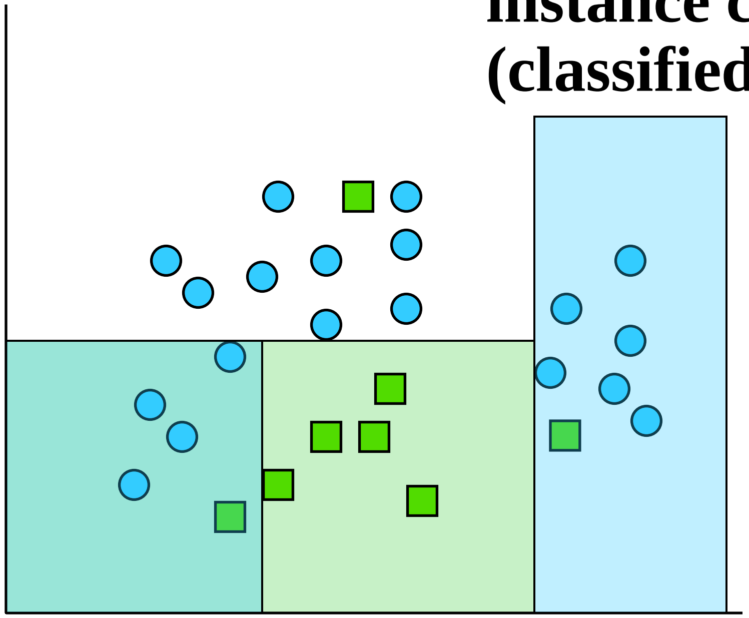
Outline

- Introduction
- K-means clustering
- Hierarchical clustering

Classification vs. Clustering

Classification: Supervised learning:

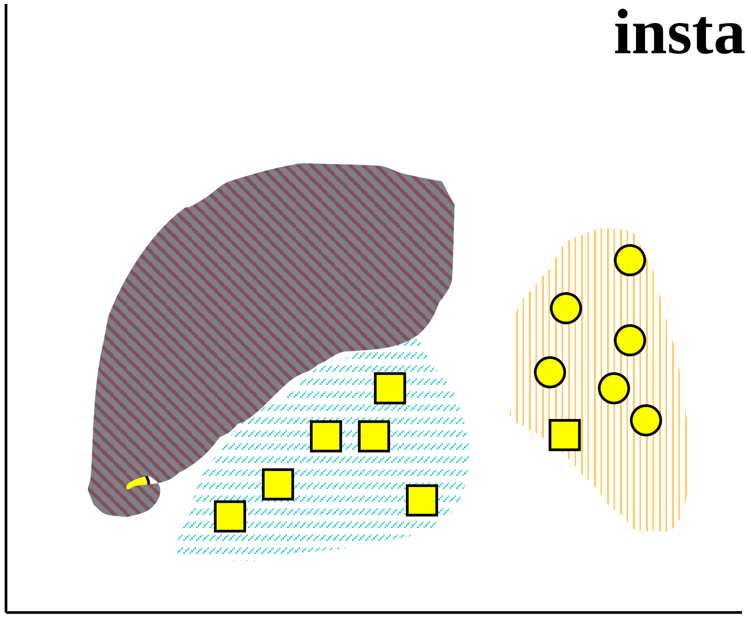
Learns a method for predicting the instance class from pre-labeled (classified) instances



Clustering

Unsupervised learning:

Finds “natural” grouping of instances given un-labeled data



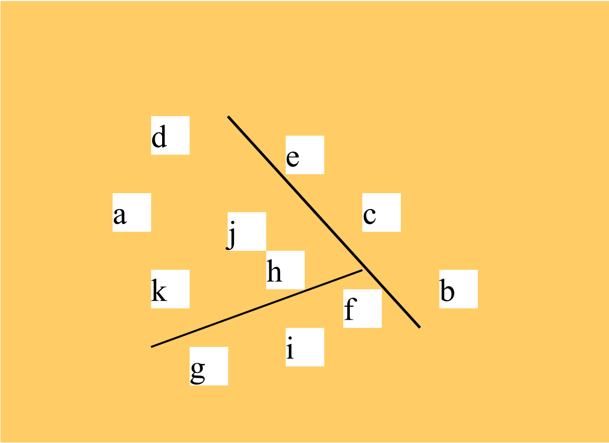
Clustering Methods

- Many different method and algorithms:
  - For numeric and/or symbolic data
  - Deterministic vs. probabilistic
  - Exclusive vs. overlapping
  - Hierarchical vs. flat
  - Top-down vs. bottom-up

Clusters: exclusive vs. overlapping

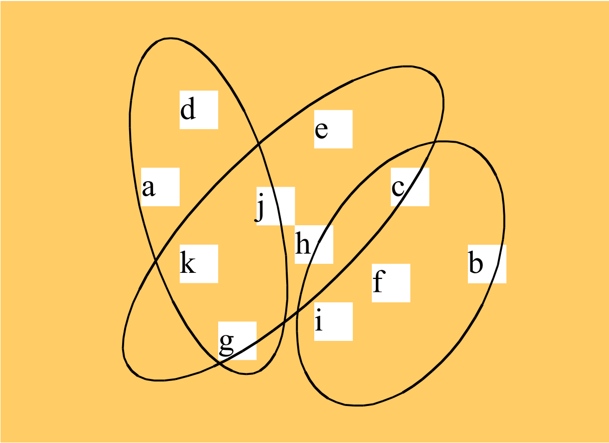
Simple 2-D representation

Non-overlapping



Venn diagram

Overlapping



Clustering Evaluation

- Manual inspection
- Benchmarking on existing labels
- Cluster quality measures
  - distance measures
  - high similarity within a cluster, low across clusters

The distance function

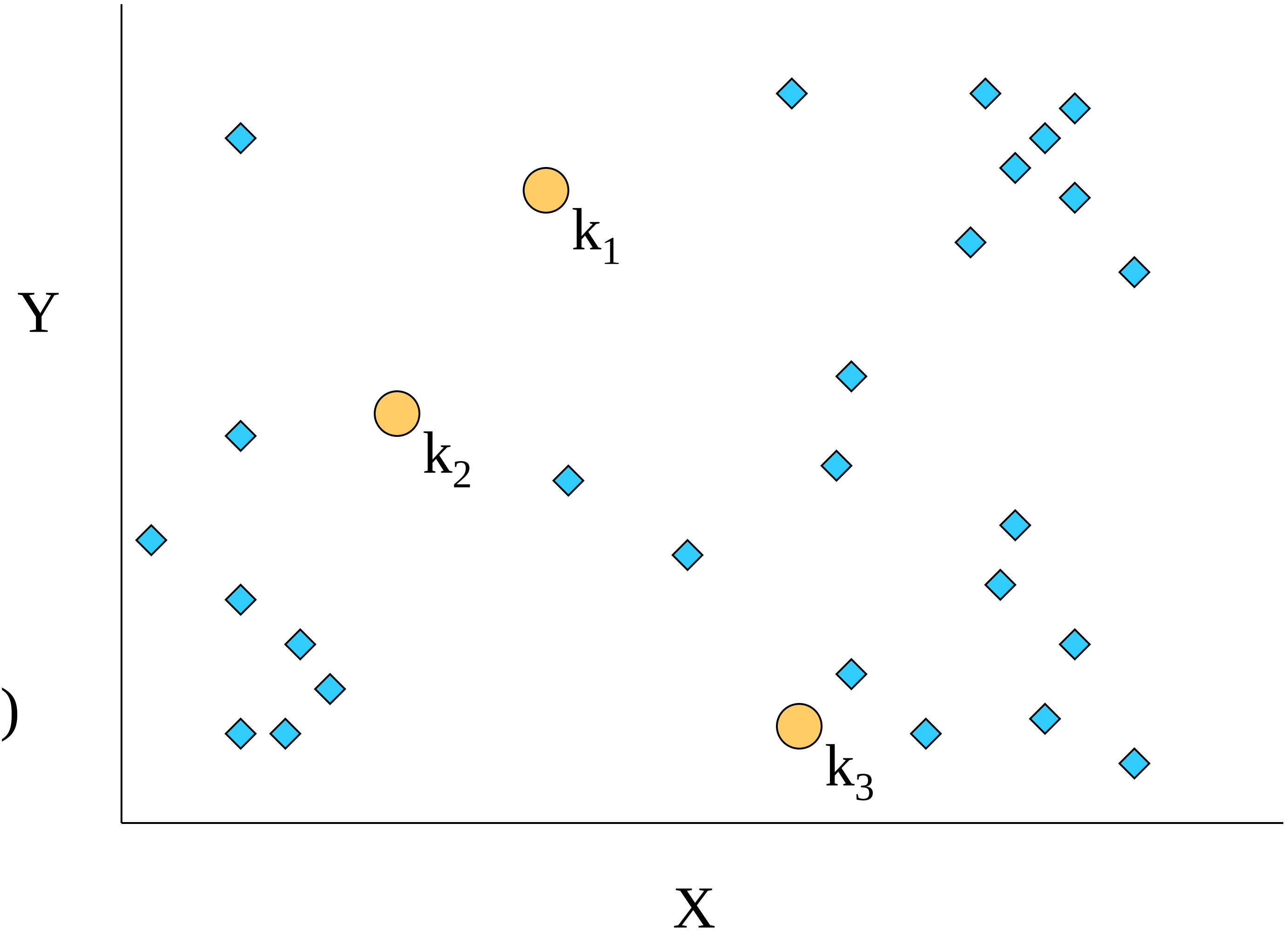
- Simplest case: one numeric attribute A
  - Distance(X,Y) = A(X) – A(Y)
- Several numeric attributes:
  - Distance(X,Y) = Euclidean distance between X,Y
- Nominal attributes: distance is set to 1 if values are different, 0 if they are equal
- Are all attributes equally important?
  - Weighting the attributes might be necessary

Simple Clustering: K-means

Works with numeric data only

1) Pick a number (K) of cluster centers (at random)
2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)
3) Move each cluster center to the mean of its assigned items
4) Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

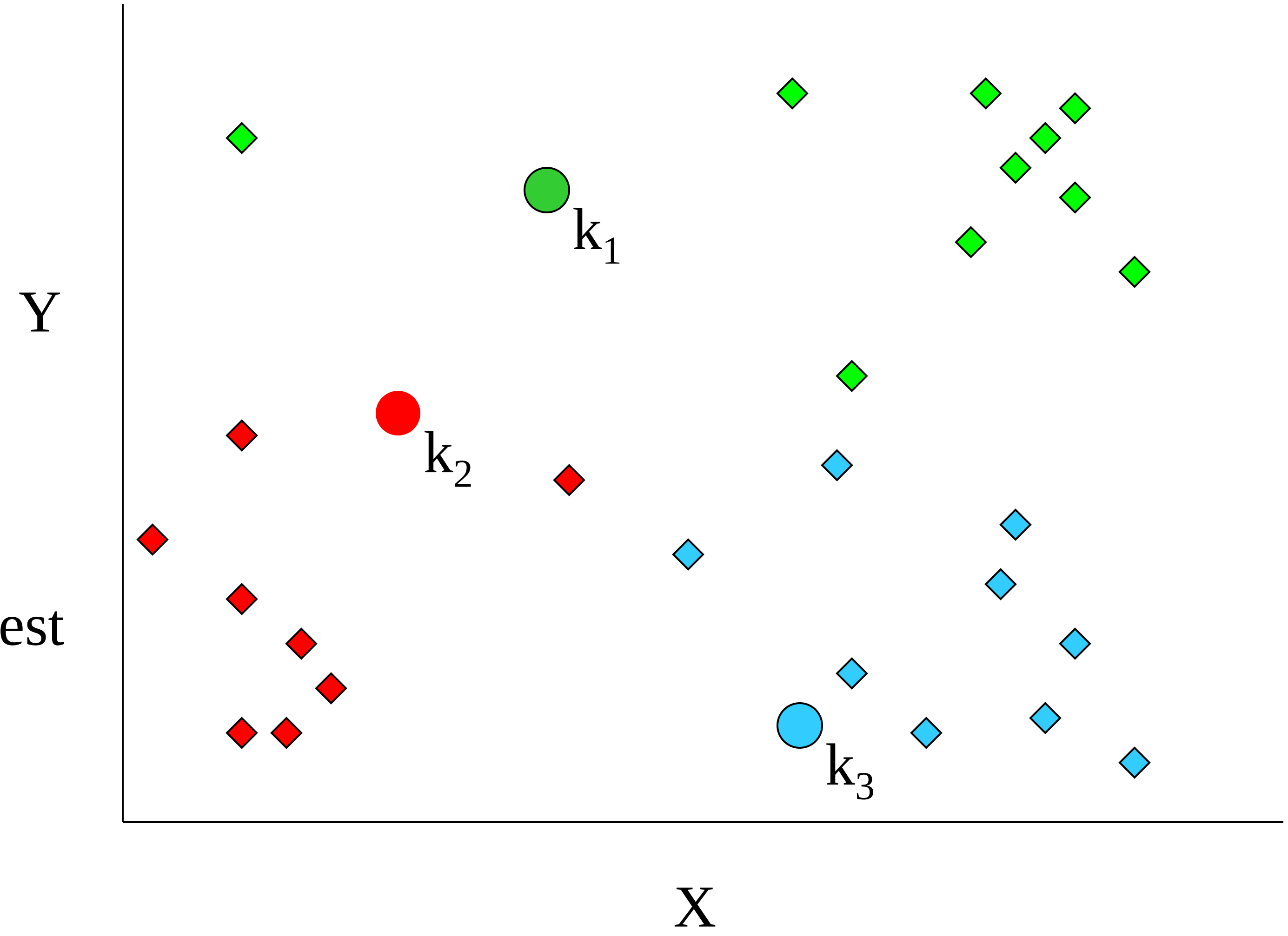
K-means example, step 1



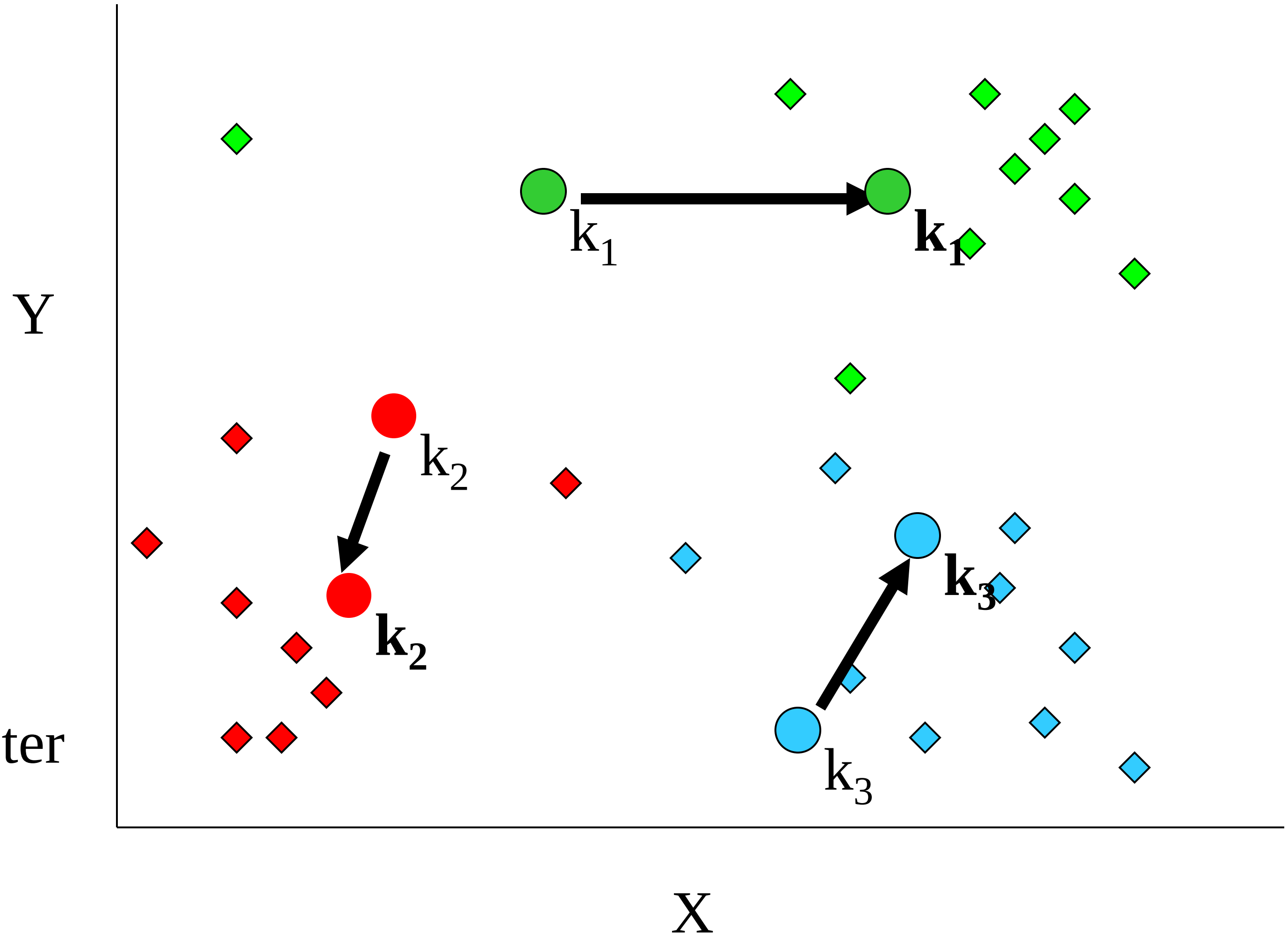
Pick 3 initial cluster centers (randomly)

K-means example, step 2

Assign each point to the closest cluster center



K-means example, step 3

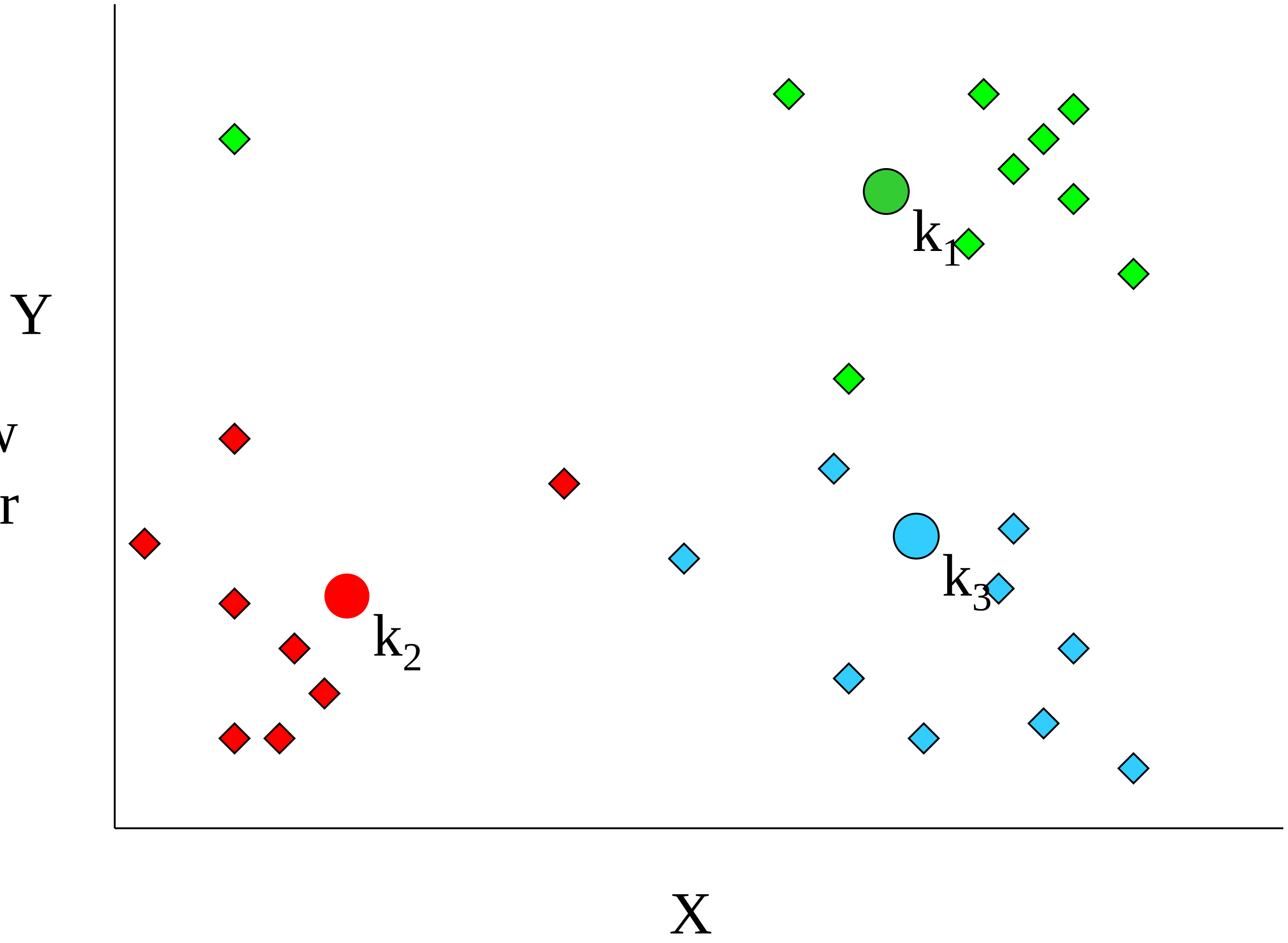


Move
each cluster
center
to the mean
of each cluster

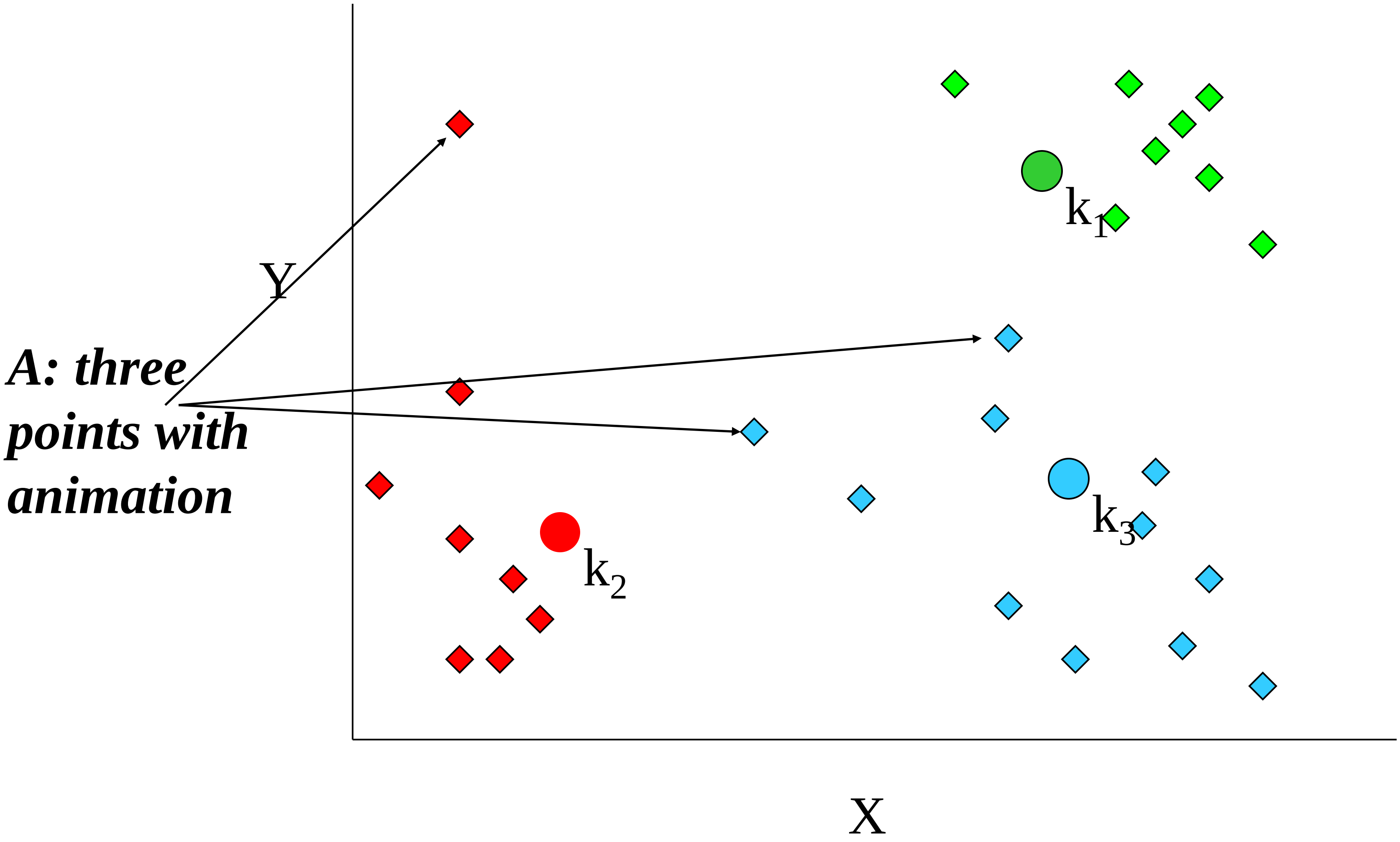
K-means example, step 4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?

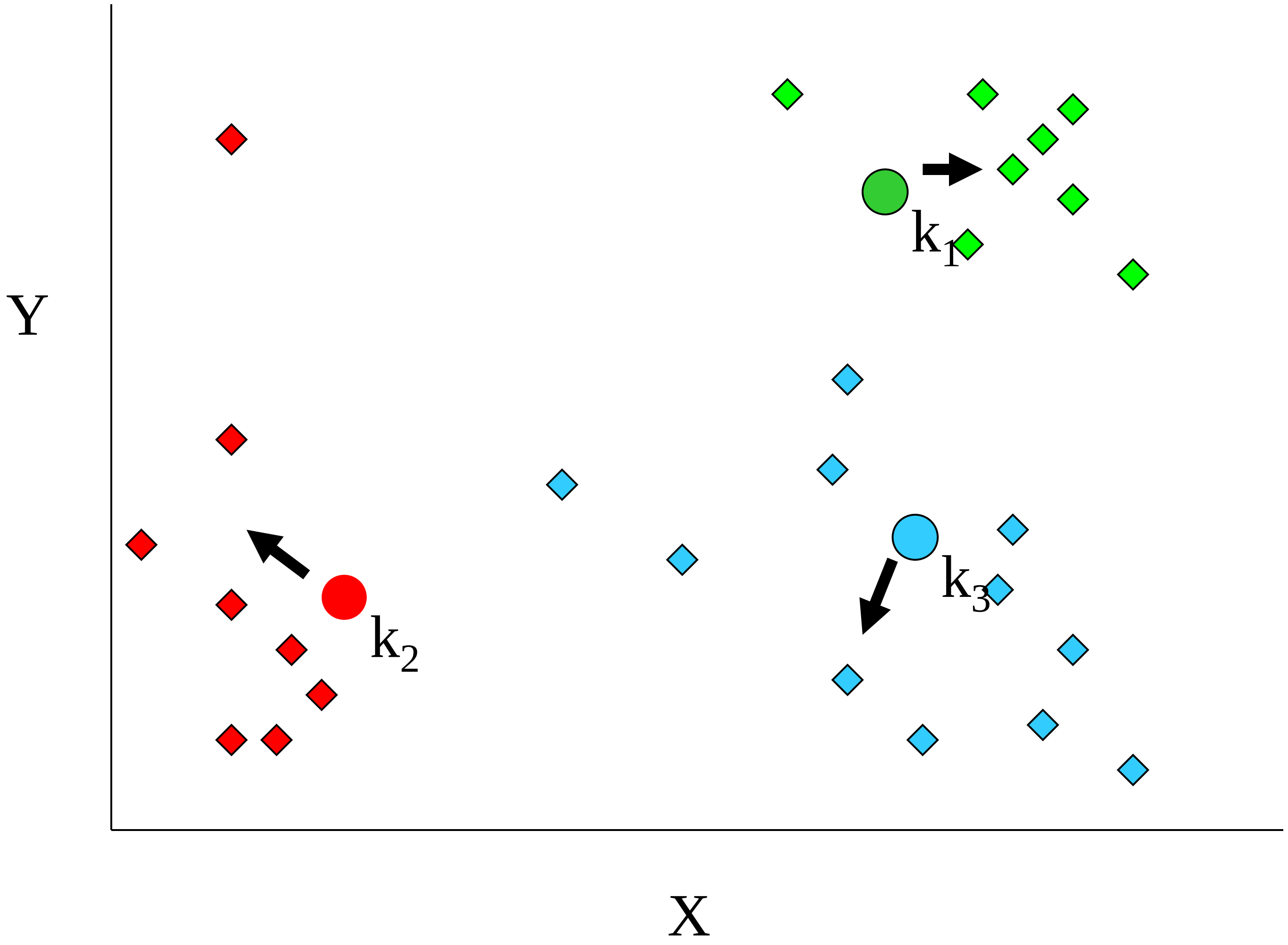


K-means example, step 4 …



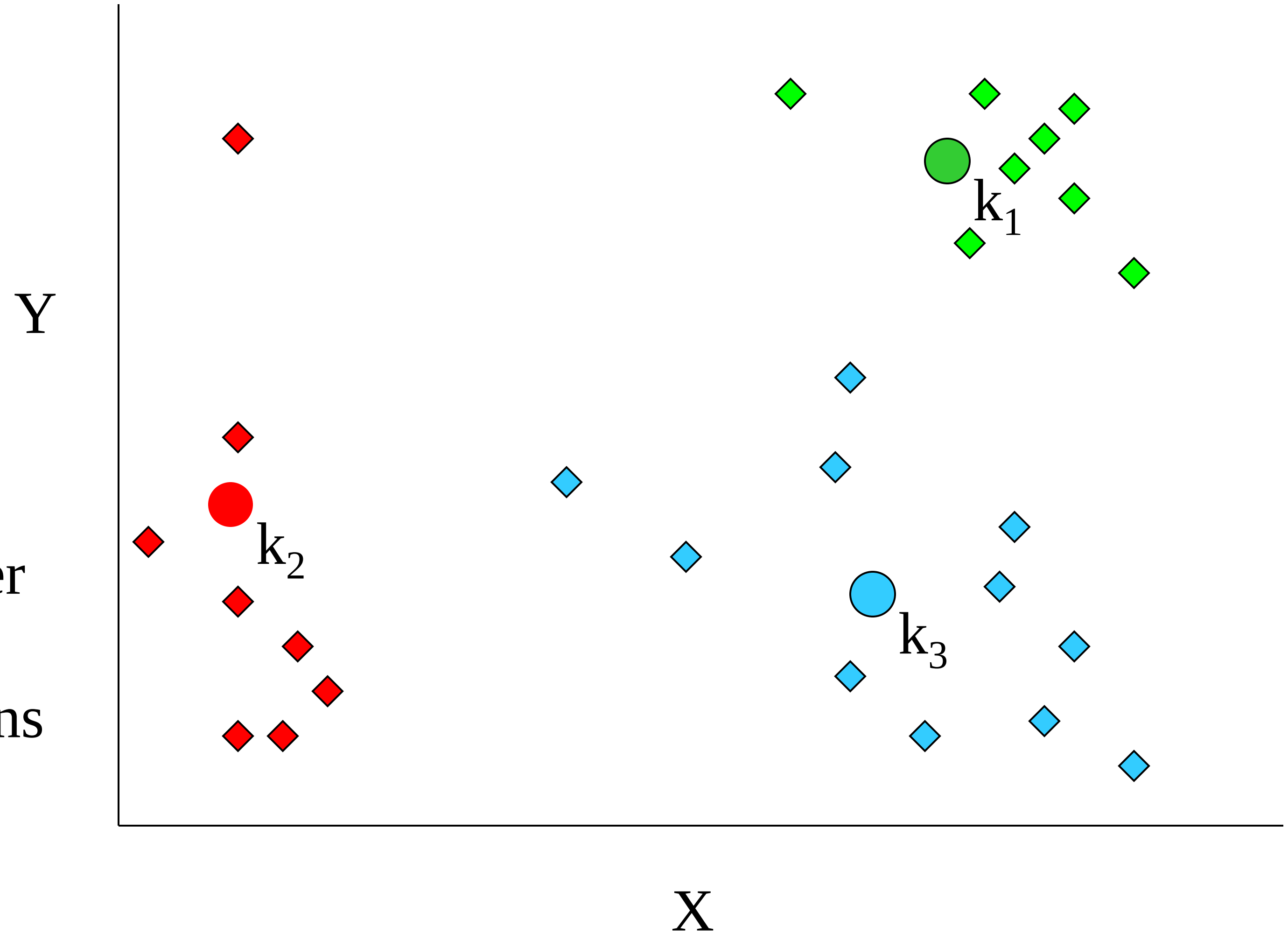
A: three points with animation

K-means example, step 4b



re-compute
cluster
means

K-means example, step 5



move cluster centers to cluster means

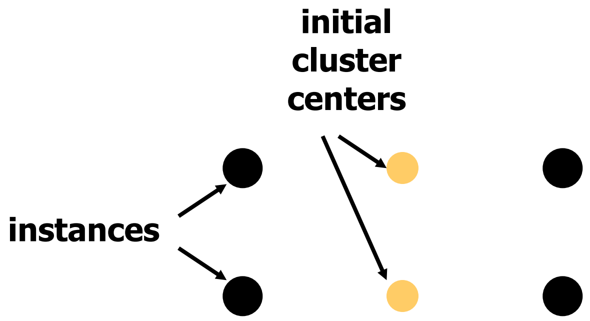
Discussion, 1

What can be the problems with

K-means clustering?

Discussion, 2

- Result can vary significantly depending on initial choice of seeds (number and position)
- Can get trapped in local minimum
  - Example:



- Q: What can be done?

Discussion, 3

A: To increase chance of finding global optimum: restart with different random seeds.

K-means clustering summary

Advantages

- Simple, understandable
- items automatically assigned to clusters

Disadvantages

- Must pick number of clusters before hand
- All items forced into a cluster
- Too sensitive to outliers

K-means clustering - outliers ?

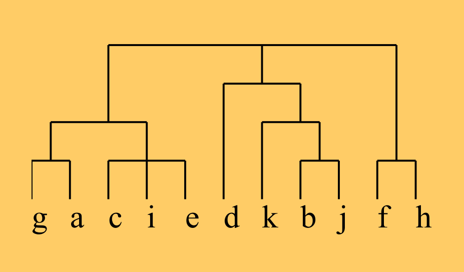
What can be done about outliers?

K-means variations

- **K-medoids** – instead of mean, use medians of each cluster
  - Mean of 1, 3, 5, 7, 9 is **5**
  - Mean of 1, 3, 5, 7, 1009 is **205**
  - Median of 1, 3, 5, 7, 1009 is **5**
  - Median advantage: not affected by extreme values
- For large databases, use sampling

*Hierarchical clustering

- Bottom up
  - Start with single-instance clusters
  - At each step, join the two closest clusters
  - Design decision: distance between clusters
    - E.g. two closest instances in clusters
      vs. distance between means
- Top down
  - Start with one universal cluster
  - Find two clusters
  - Proceed recursively on each subset
  - Can be very fast
- Both methods produce a *dendrogram*



Agglomerative Hierarchical Clustering

… you will find the description and an example here:

https://www.learndatasci.com/glossary/hierarchical-clustering/

Other Clustering Approaches

- EM – probability based clustering
- Bayesian clustering
- SOM – self-organizing maps
- …

Discussion

- Can interpret clusters by using supervised learning
  - learn a classifier based on clusters
- Decrease dependence between attributes?
  - pre-processing step
  - E.g. use *principal component analysis*
- Can be used to fill in missing values
- Key advantage of probabilistic clustering:
  - Can estimate likelihood of data
  - Use it to compare different models objectively

Examples of Clustering Applications

- **Marketing:** discover customer groups and use them for targeted marketing and re-organization
- **Astronomy:** find groups of similar stars and galaxies
- **Earth-quake studies:** Observed earth quake epicenters should be clustered along continent faults
- **Genomics:** finding groups of gene with similar expressions
- …

Clustering Summary

- Unsupervised
- Many approaches
  - K-means – simple, sometimes useful
    - K-medoids is less sensitive to outliers
  - Hierarchical clustering – works for symbolic attributes
- Evaluation is a problem

Principal Component Analysis (**PCA**)

… you will find (detailed) description here:

http://setosa.io/ev/principal-component-analysis/

https://youtu.be/_UVHneBUBW0

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf