Data Clustering

Danushka Bollegala

UNIVERSITY OF LIVERPOOL

Outline

- Why cluster data?
- Clustering as unsupervised learning
- Clustering algorithms
  - k-means, k-medoids
  - agglomerative clustering
  - Brown's clustering
  - Spectral clustering
- Cluster evaluation measures
  - Purity
  - Normalised Mutual Information
  - Rand Index
  - B-CUBED
  - Precision, Recall, F-score
- Supervised clustering

We look only at topics shown in red here

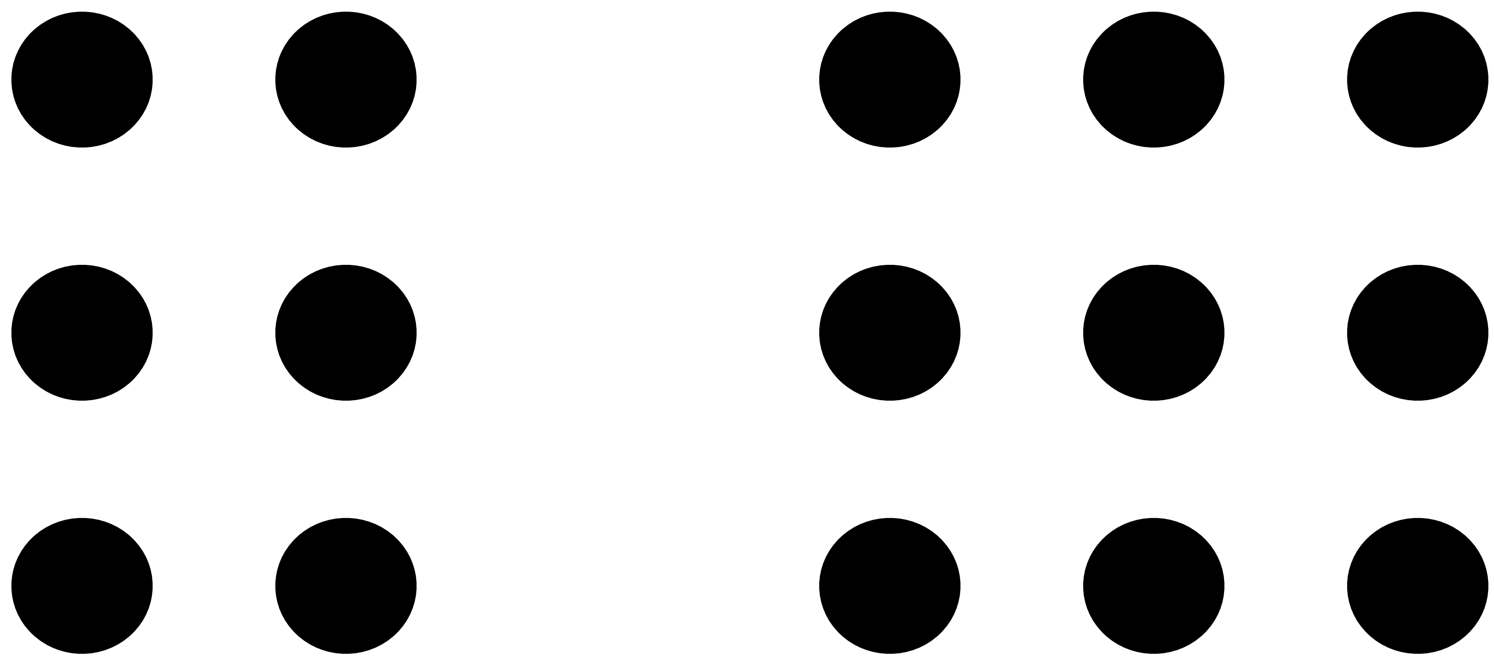
Why cluster data?

- Data Mining has two main objectives
  - Prediction: classification, regression etc.
  - Description: pattern mining, rule extraction, visualisation *clustering*
- Clustering is:
  - Unsupervised learning
    - no label data is required (consider classification algorithms we discussed so far in the lecture which are supervised algorithms)
  - Generalisation / Abstraction of concepts
  - Topic detection
  - Visualisation
  - Outlier detection

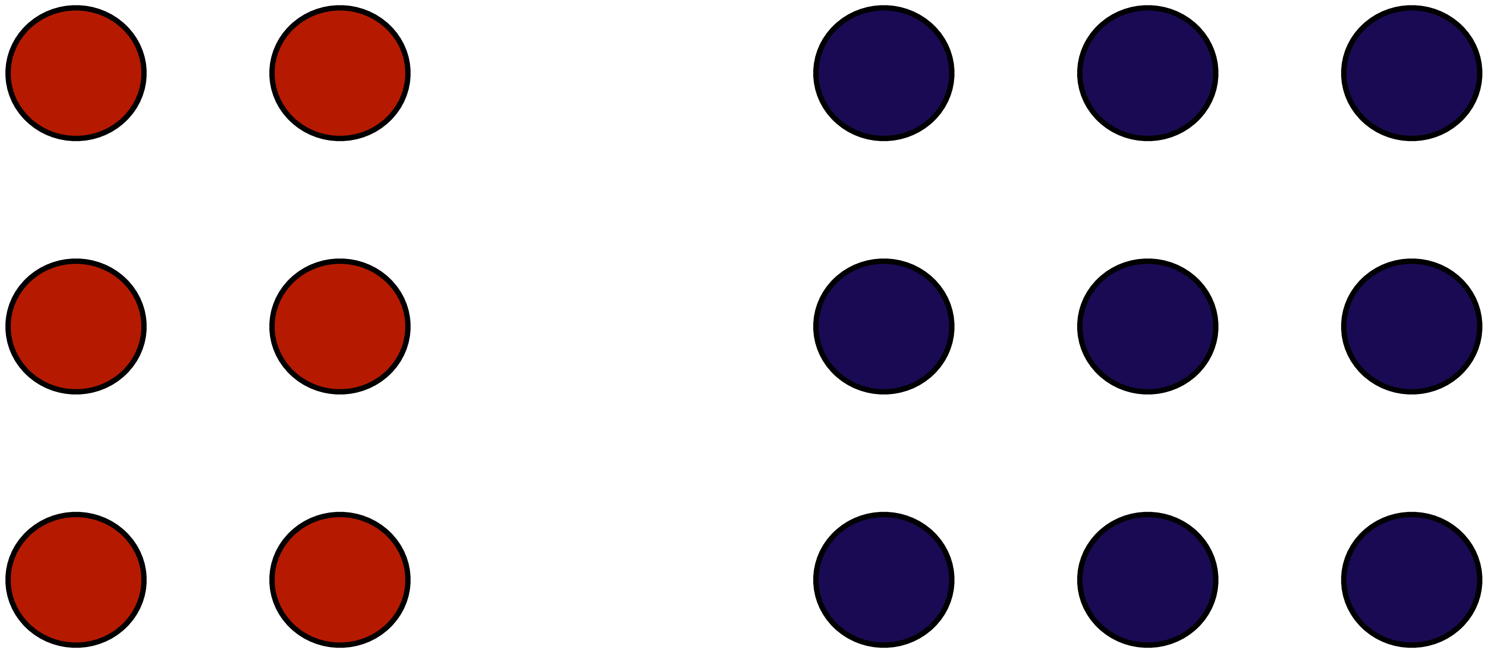
Unsupervised Learning

- Supervised learning
  - labels for training instances are provided
- Unsupervised learning
  - No labels for training instances are provide
- Semi-supervised learning
  - Both labeled and unlabeled training instances are provided
- What can we learn about training data if we do not have any labels?
  - The similarity and distribution of the features can still be learnt and this can be used to create rich feature spaces for supervised learning (if required)

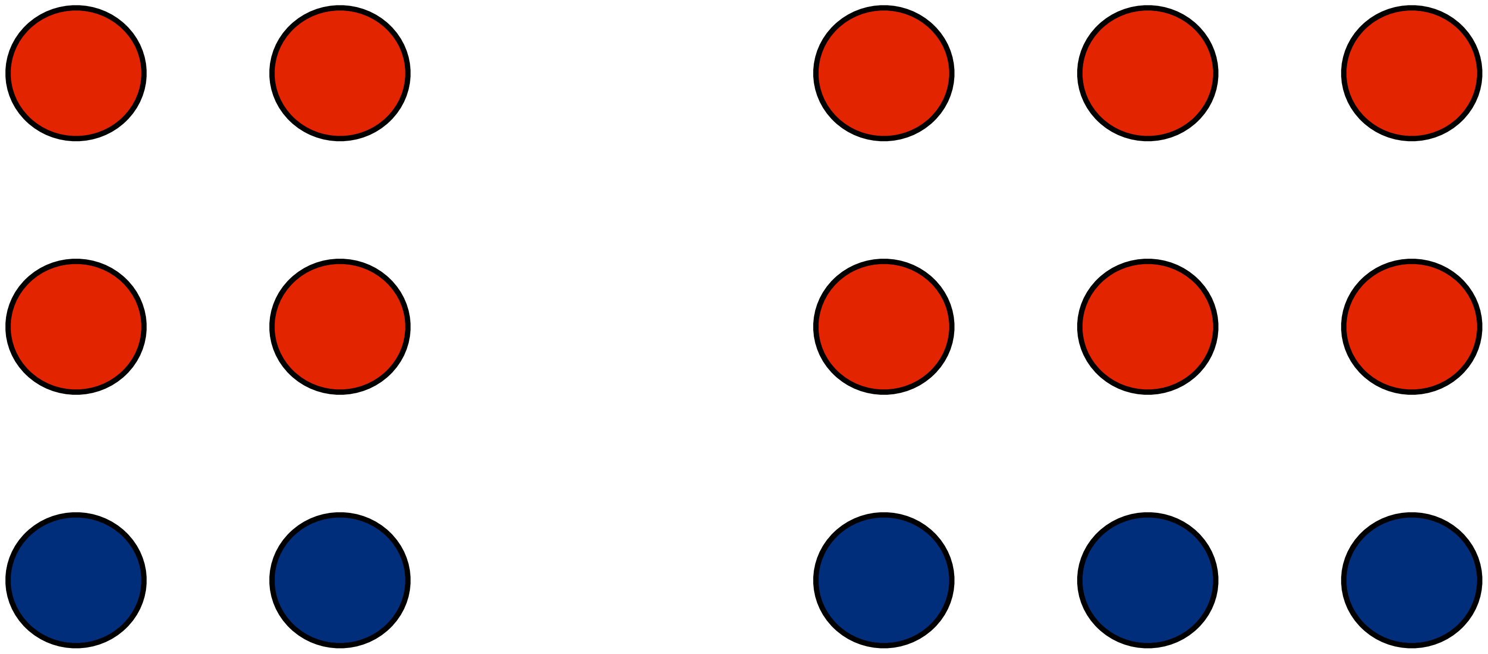
Quiz: Cluster the Following Data



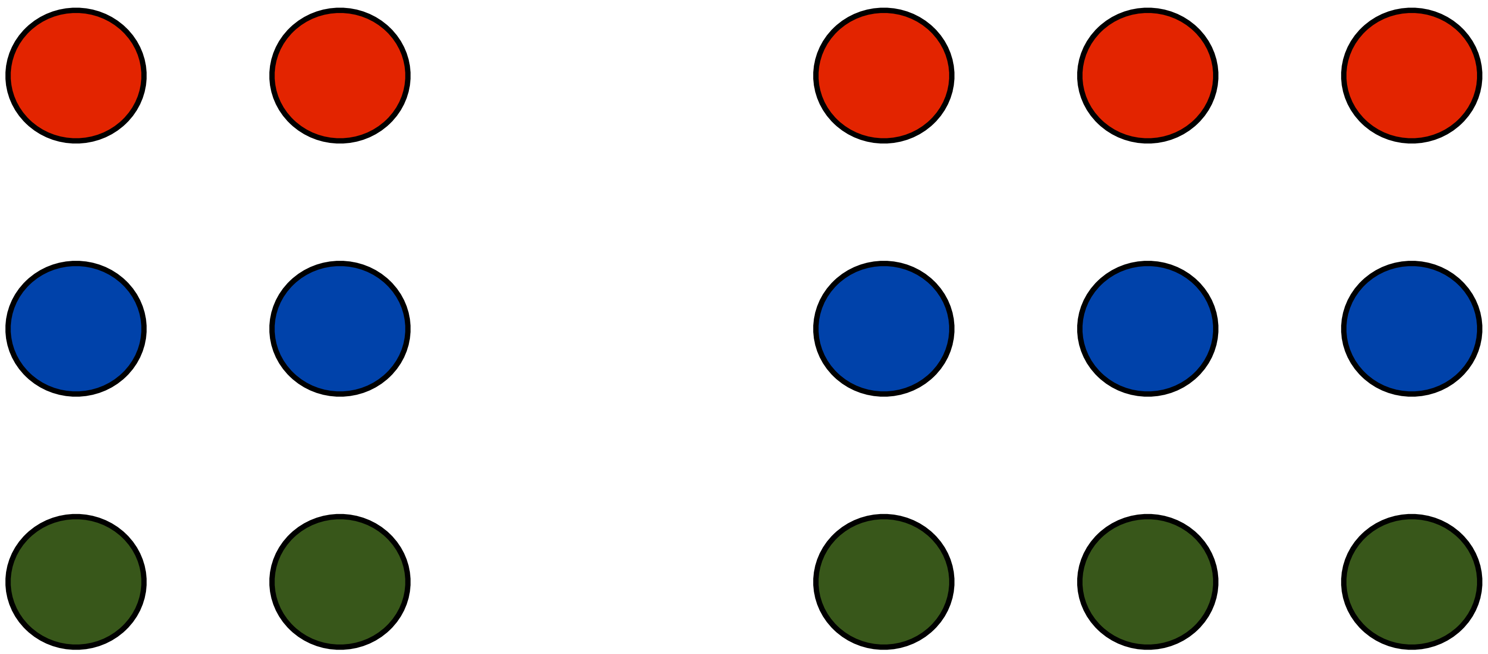
Quiz: Cluster the Following Data



Quiz: Cluster the Following Data



Quiz: Cluster the Following Data



How many clusters?

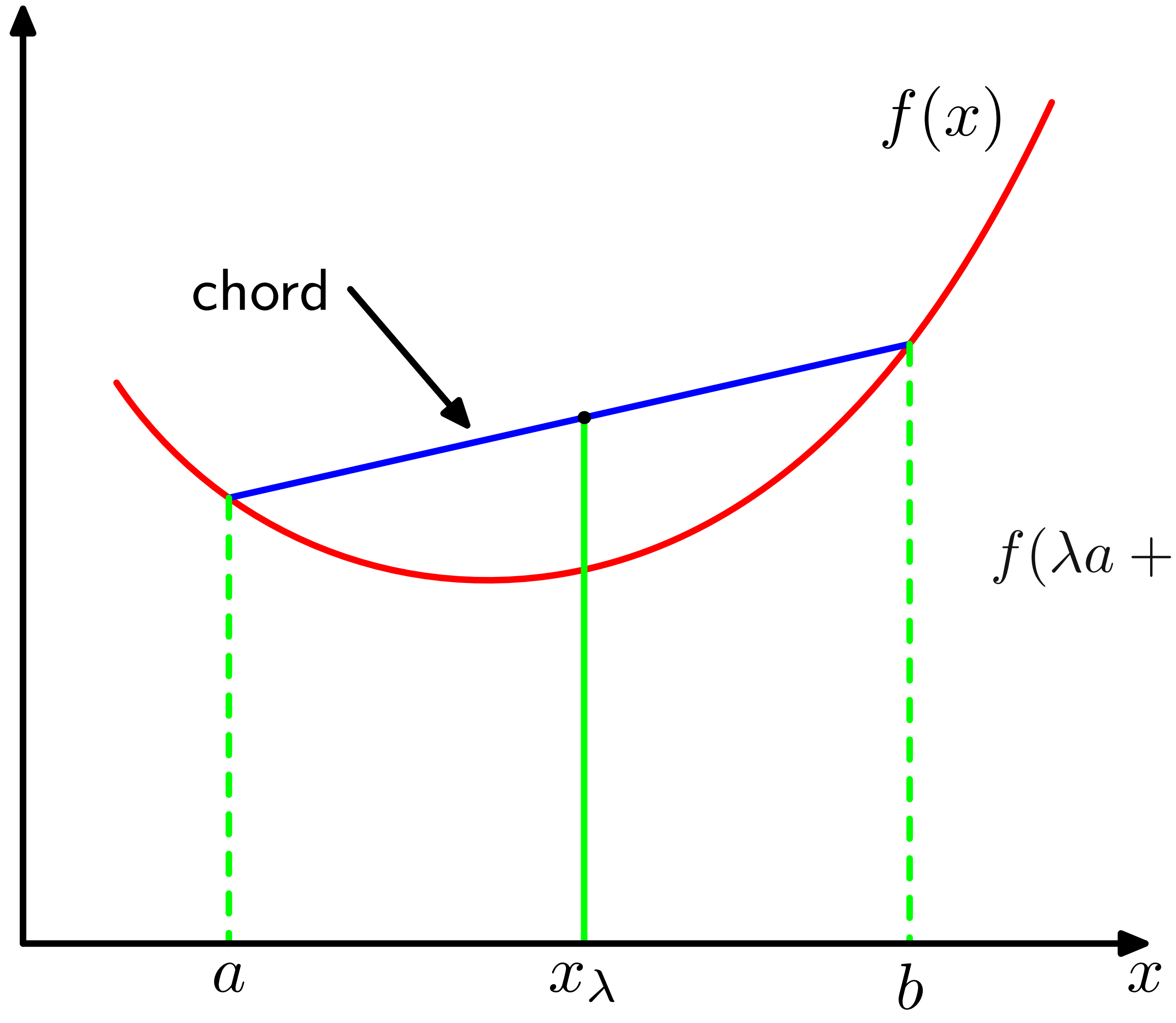
General Remarks

- A single dataset can be clustered into several ways
- There is no single right or wrong clustering
  - Simply different views on the same data
- If so how can we measure the quality of a clustering algorithm?
  - Two ways
    - Compare the clusters produced by a clustering algorithm against some reference (gold standard) set of clusters (direct evaluation)
    - Use the clusters as features for some other (eg. supervised learning) task and measure the difference in the performance of the second task (indirect evaluation)

Clustering as Optimisation

- Given a dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of N instances represented as *d* dimensional real vectors ($\mathbf{x}_i \in R^d$), partition these N instances into k clusters $S_1,\ldots,S_k$ such that some objective function $f(S_1,\ldots,S_k)$ is minimised.
- Observations
  - k and *f* are given
  - *f* can be the similarity between the clusters (good to create dissimilar clusters as much as possible), information gain, correlation and various other such *goodness* measures (heuristics)
  - Often clustering is an NP hard and a non-convex problem
    - http://rangevoting.org/VattaniKmeansNPC.pdf
    - approximations, relaxations are required in practice

Convex Functions



$$f(\lambda a + (1-\lambda)b) \leqslant \lambda f(a) + (1-\lambda)f(b).$$

Clustering Algorithms

- Partitioning
  - Construct k partitions and iteratively update the partitions
    - k-Means, k-Medoids
- Hierarchical
  - Create a hierarchy of clusters (dendrogram)
    - Agglomerative clustering (bottom-up)
    - Conglomerative clustering (top-down)
- Graph-based clustering
  - Graph-cut algorithms (Spectral Clustering)
- Model-based clustering
  - Mixture of Gaussians
- Other types: Non-parametric Bayesian (Latent Dirichlet Allocation), Expectation Maximisation (EM) algorithm, and many more ...

k-Means Derivation

$$\arg\min_{S_1,\ldots,S_k} \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i} ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2$$

We want to minimize the distance between data instances (x_j) and some cluster centres (μ_i)

$$f(S_1, \ldots, S_k) = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i} ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2$$

This objective function is called the *within cluster sum of squares* (WCSS) objective

$$\frac{\partial f(S_1, \ldots, S_k)}{\partial \mu_i} = 0$$

$$\frac{\partial f(S_1, \ldots, S_k)}{\partial \mu_i} = \sum_{\boldsymbol{x}_j \in S_i} 2(\boldsymbol{x}_j - \boldsymbol{\mu}_i)$$

$$\mu_i = \frac{1}{|S_i|} \sum_{\boldsymbol{x}_j \in S_i} \boldsymbol{x}_j$$

Just compute the centroid (mean) of each cluster and that will give you the cluster centers

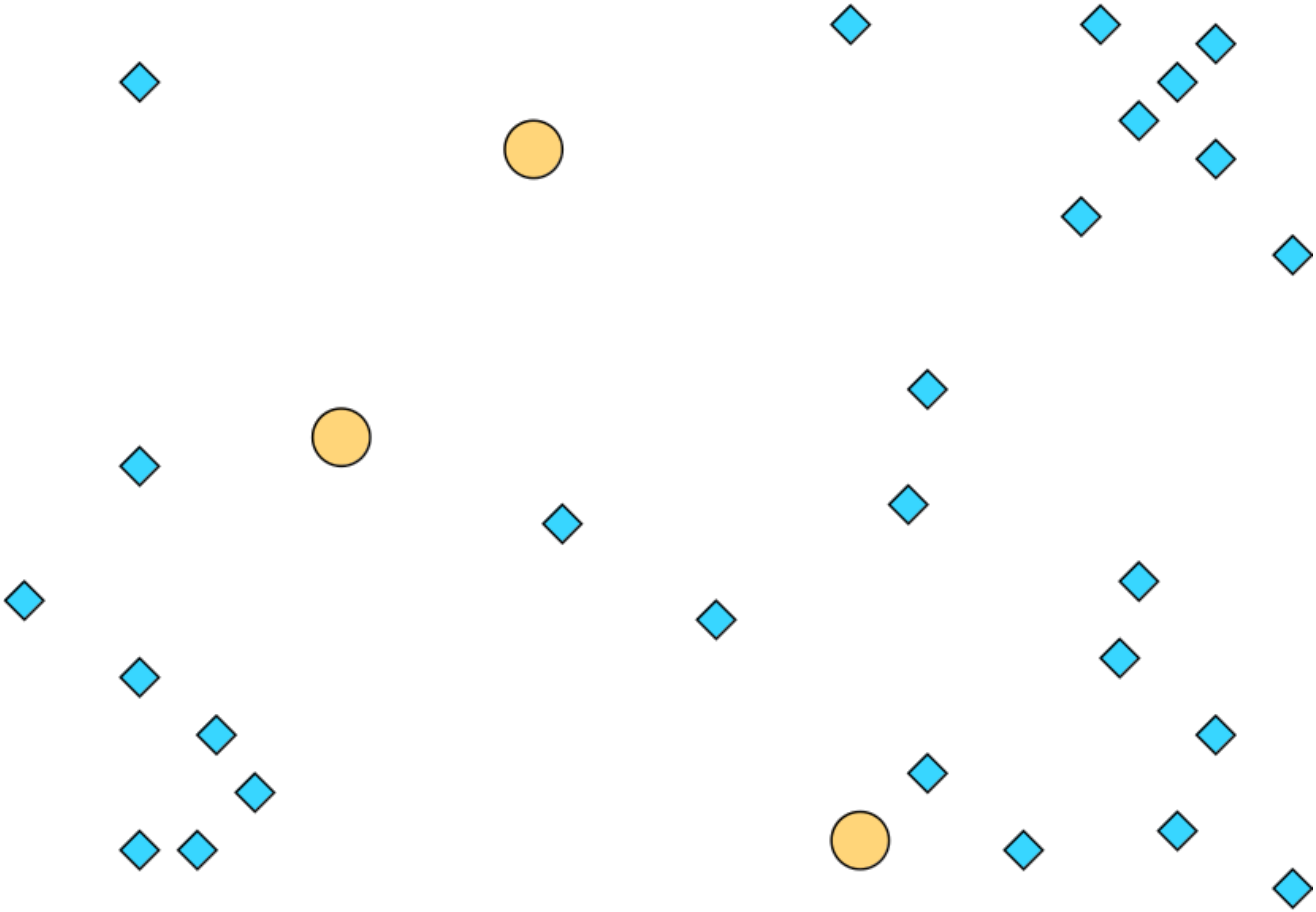
k-Means Clustering

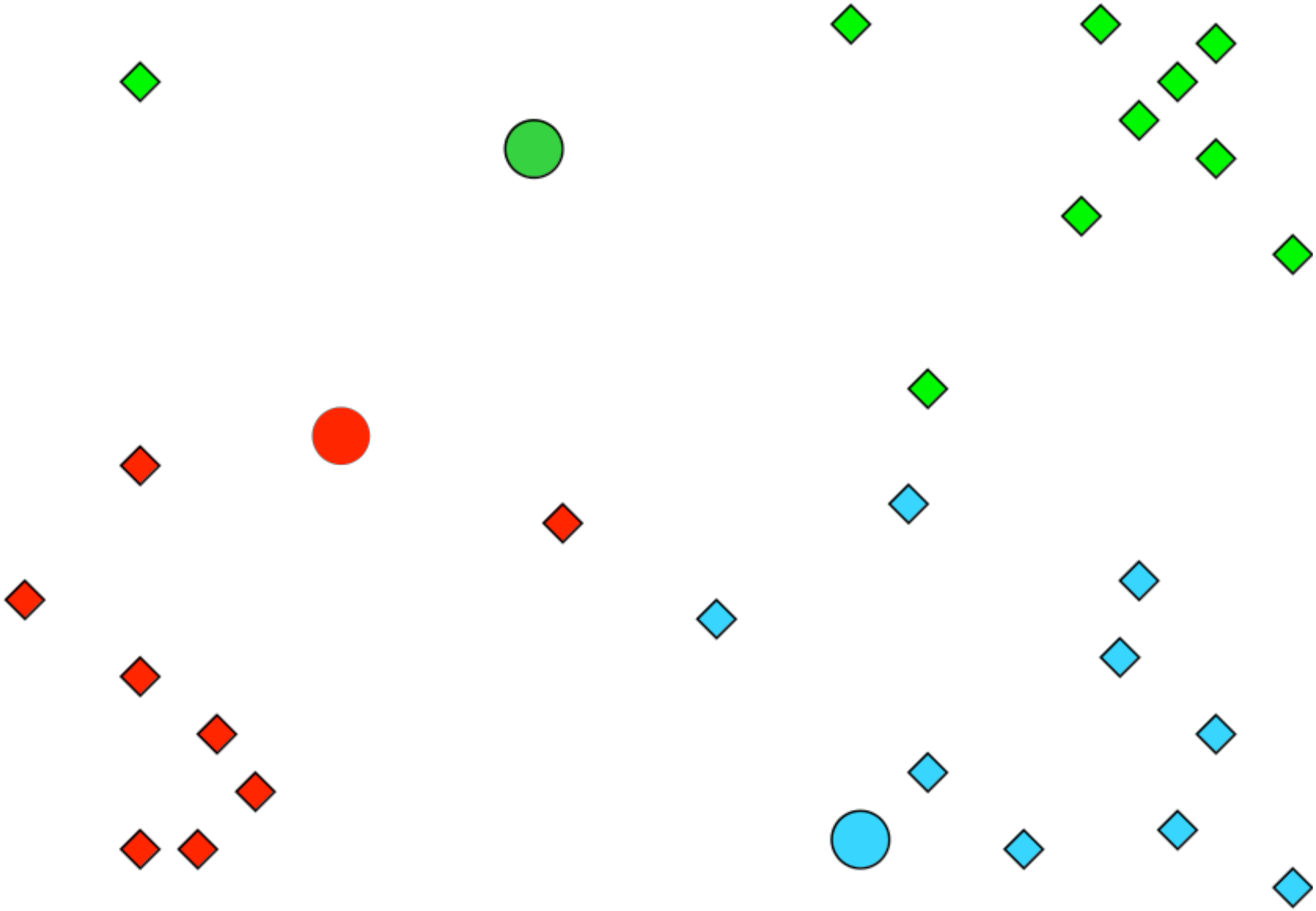
- INPUT
  - The number of clusters *k*
  - Dataset $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ of N instances represented as *d* dimensional real vectors ($\mathbf{x}_i \in R^d$)

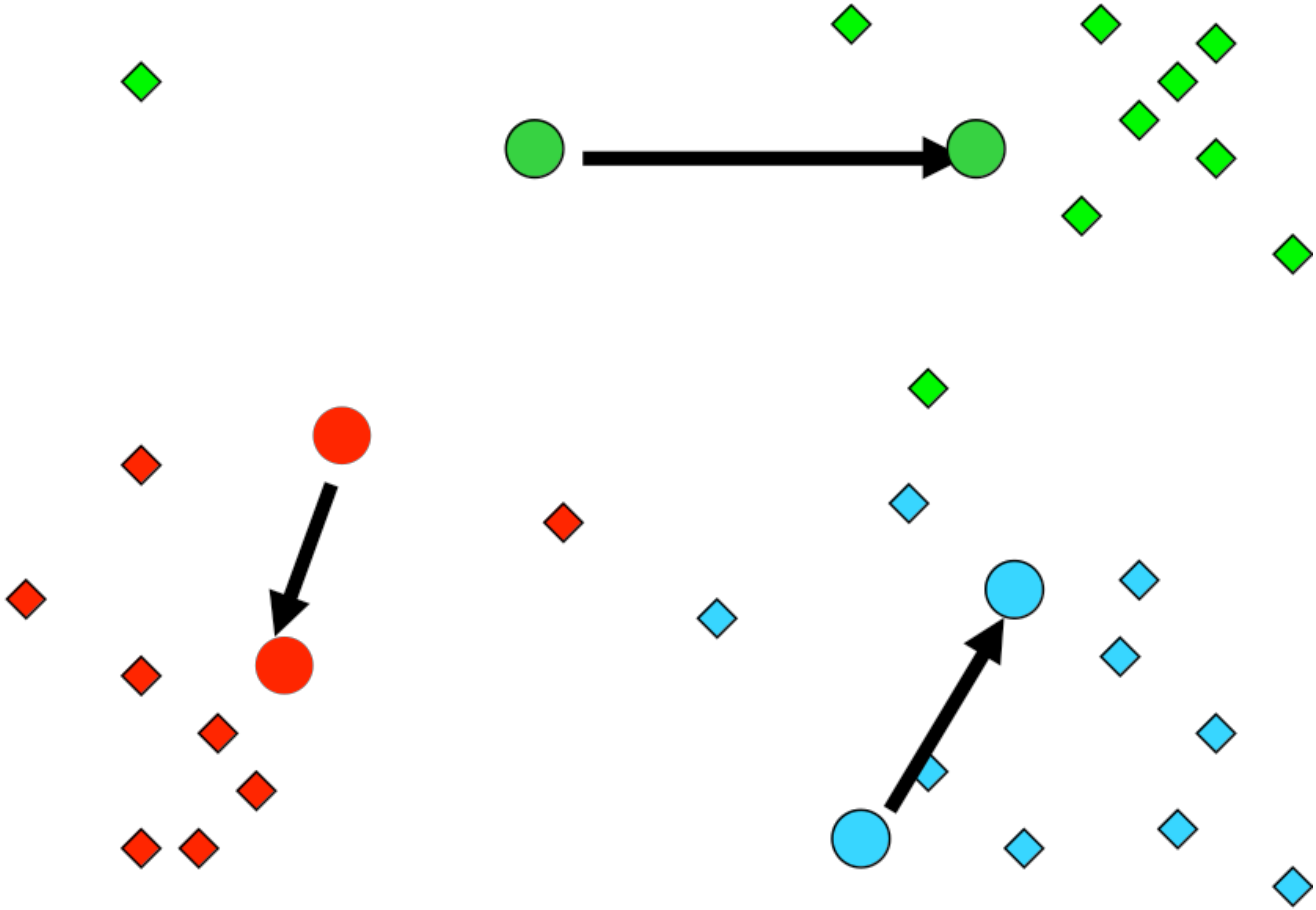
1. Set k instances from the dataset randomly. (initial cluster means/ centers)
2. Assign all other instances to the closest cluster centre.
3. Compute the mean of each cluster
4. until convergence repeat between steps 2 and 3

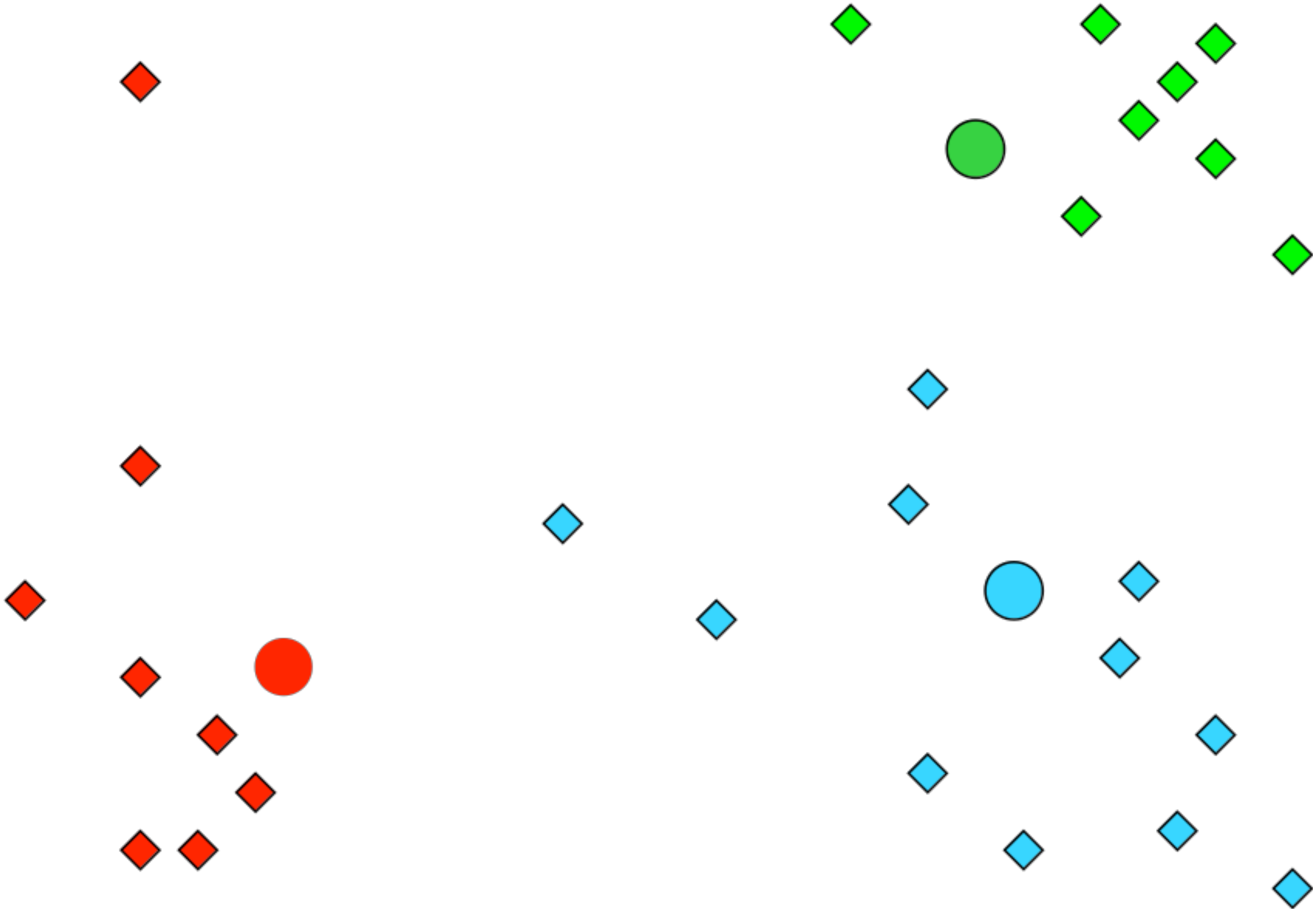
convergence = no instances have moved among clusters

(often after a fixed number of iterations specified by the user)









Issues with k-Means

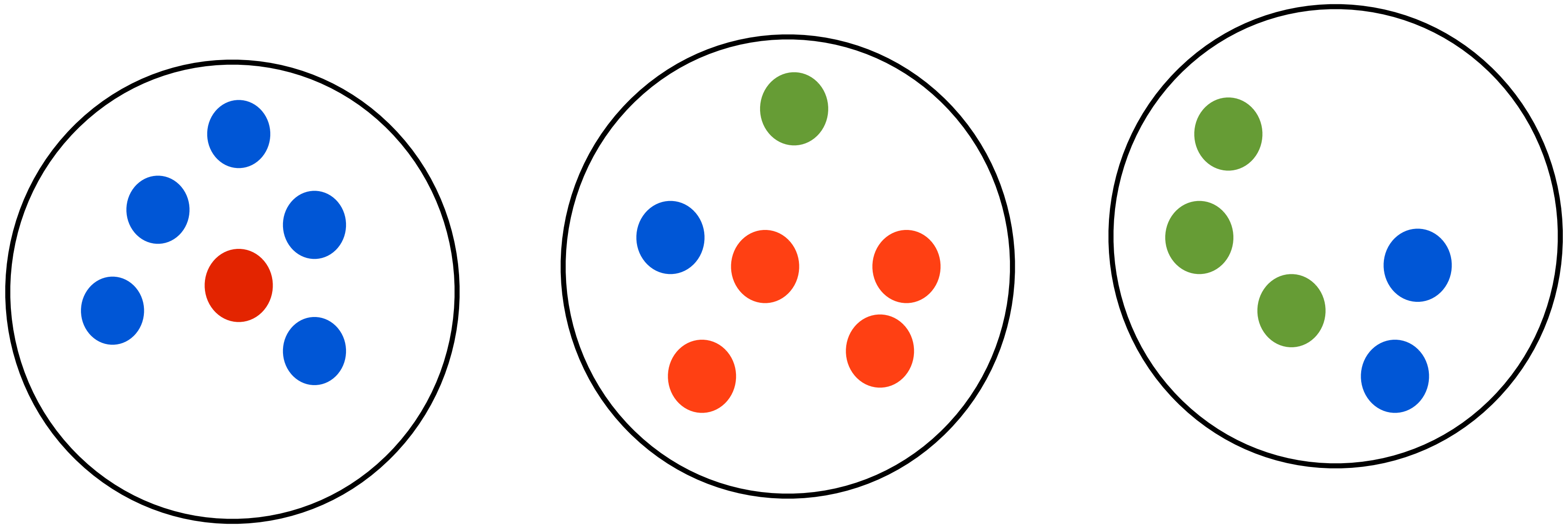
- Results can vary depending on the initial random choices
- Can get trapped in a local minimum that isn't the global optimal solution
    - Repeat the clustering procedure multiple times with different initialisations and select the *best* final clustering
        - *best*? according to what? many heuristics exist.
            - smallest number of iterations before convergence
            - largest total distance between the final cluster means
- Outliers have a larger effect on the mean value, hence cluster centre and the cluster
- cluster centres (means) are not actual instances in the cluster
    - We could pick actual instances as initial cluster centroids.

Evaluating Clustering — Purity

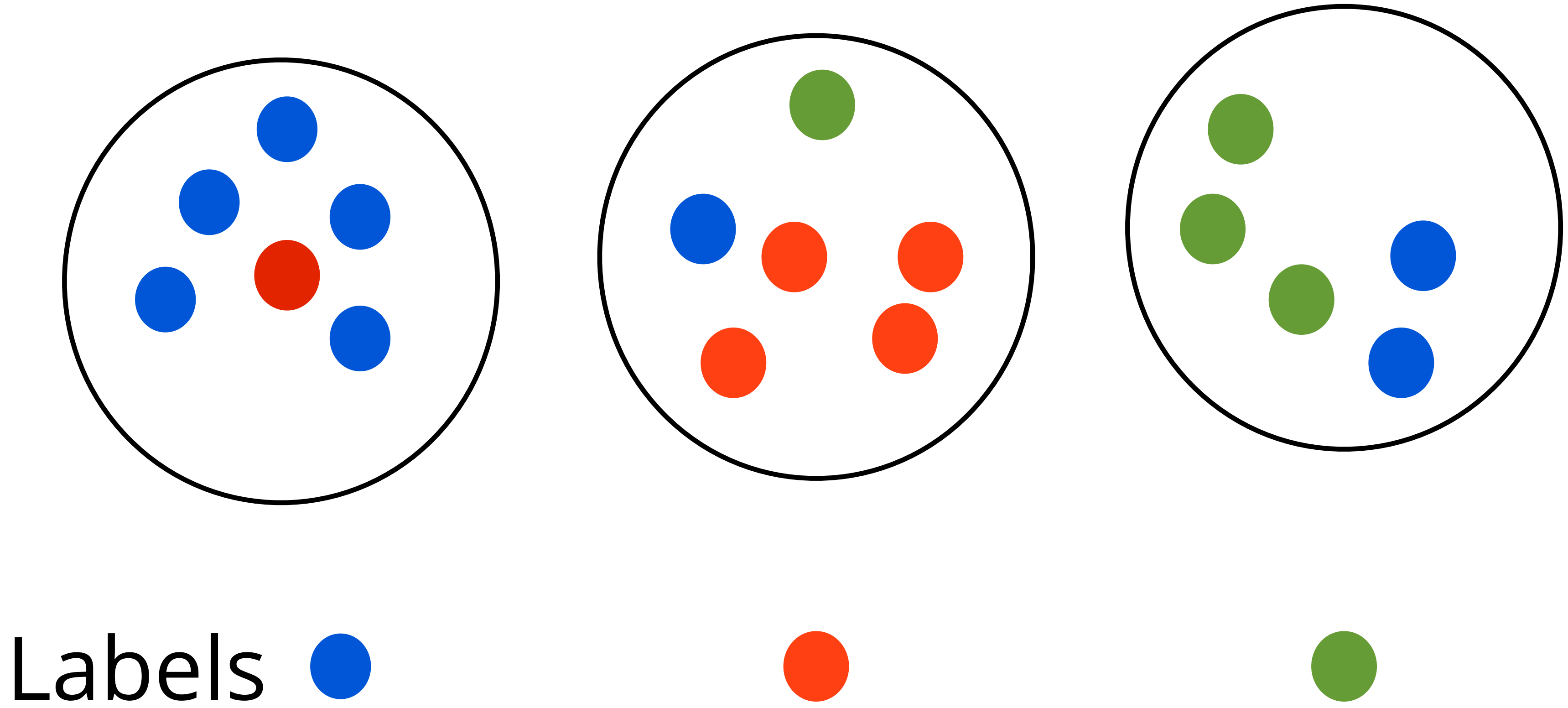
- Let us assume that we have a set $\Omega = \{\omega_1,\ldots,\omega_K\}$ clusters for a set of classes $C = \{c_1,\ldots,c_J\}$
- Purity measures the ratio of the items that are in the cluster with the same class as its own.

$$\text{purity}(\Omega, \mathrm{C}) = \frac{1}{N}\sum_k \max_j \left|\omega_k \cap c_j\right|$$

- Here, N is the total number of items.



Quiz: Compute purity for this clustering.



purity = (5 + 4 + 3) / 17 = 12/17 = 0.71

Purity achieves its maximum value of 1 for singletons (each item is in a cluster containing only that single item)! Obviously this is not good "clustering" and purity does not recognise this.

Evaluating Clustering — NMI

- Let us assume that we have a set $\Omega = \{\omega_1, \ldots, \omega_K\}$ clusters for a set of classes $C = \{c_1, \ldots, c_J\}$
- Normalised Mutual Information (NMI) computes the ratio of information that we can know about the classes C given the clusters Ω to the averaged information that is contained in C and Ω.

$$\text{NMI}(\Omega, \mathcal{C}) = \frac{I(\Omega, \mathcal{C})}{[H(\Omega) + H(\mathcal{C})]/2}$$

$$I(\Omega, \mathcal{C}) = \sum_k \sum_j p(\omega_k \cap c_j) \log \left(\frac{p(\omega_k \cap c_j)}{p(\omega_k)p(c_j)} \right) = \sum_k \sum_j \frac{|\omega_k \cap c_j|}{N} \log \left(\frac{N|\omega_k \cap c_j|}{|\omega_k||c_j|} \right)$$

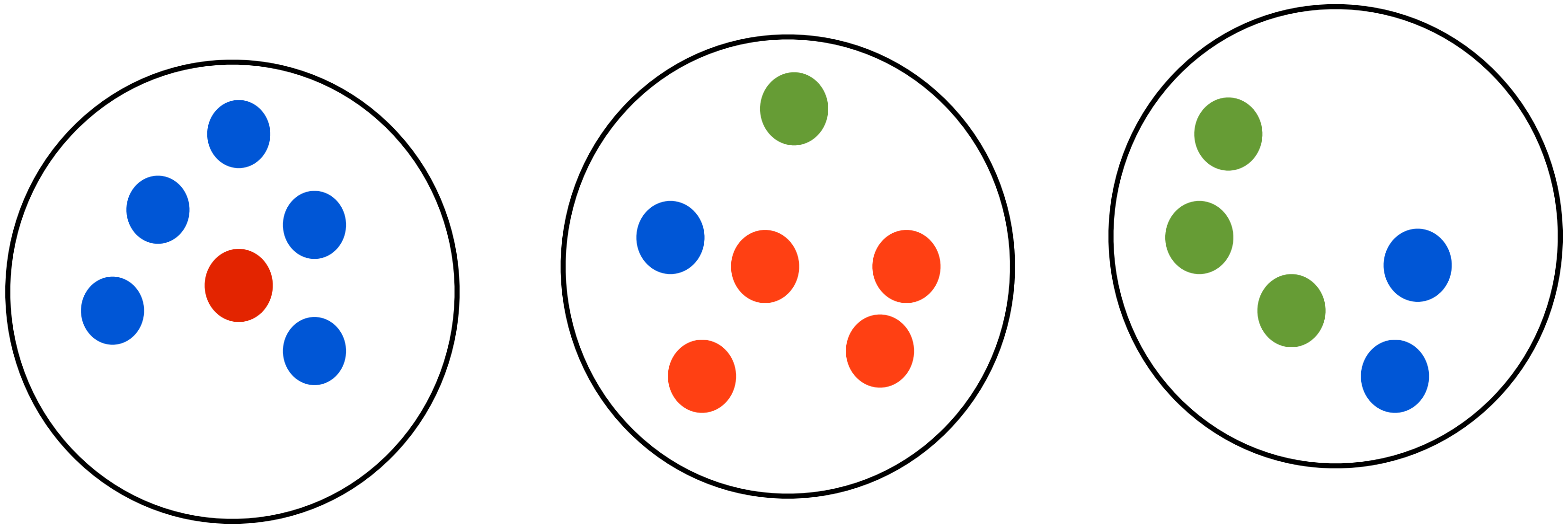
Mutual Information (MI)

$$H(\Omega) = -\sum_k p(\omega_k) \log p(\omega_k) = -\sum_k \frac{|\omega_k|}{N} \log \frac{|\omega_k|}{N}$$

Entropy

Why we do we normalise by the average?

- $I(X,Y) \leq [H(X) + H(Y)]/2$
- Proof (sketch):
  - $I(X,Y) = H[X] - H[X|Y] = H[Y] - H[Y|X]$
  - Add those two and use the fact that (conditional) entropy is nonnegative
    - $H[X|Y] + H[Y|X] \geq 0$



Quiz: Compute NMI for this clustering.

Let C_1 = Blue, C_2 = Red and C_3 = Green.

$p(C_1) = \frac{8}{17}$, $p(C_2) = \frac{5}{17}$, $p(C_3) = \frac{4}{17}$.

$$\therefore H(C) = -\sum_{j=1}^{3} p(C_j) \log p(C_j)$$

$$= -\left[\frac{8}{17}\log\frac{8}{17} + \frac{5}{17}\log\frac{5}{17} + \frac{4}{17}\log\frac{4}{17}\right] = 1.055$$

Likewise,

$p(\omega_1) = \frac{6}{17}$, $p(\omega_2) = \frac{6}{17}$, $p(\omega_3) = \frac{5}{17}$

$$H[\Omega] = -\left[\frac{6}{17}\log\frac{6}{17} + \frac{6}{17}\log\frac{6}{17} + \frac{5}{17}\log\frac{5}{17}\right] = 1.095$$

$P(\omega_1 \cap C_1) = \frac{5}{17}$ $\quad P(\omega_1 \cap C_2) = \frac{1}{17}$ $\quad P(\omega_1 \cap C_3) = \frac{0}{17}$

$P(\omega_2 \cap C_1) = \frac{1}{17}$ $\quad P(\omega_2 \cap C_2) = \frac{4}{17}$ $\quad P(\omega_2 \cap C_3) = \frac{1}{17}$

$P(\omega_3 \cap C_1) = \frac{2}{17}$ $\quad P(\omega_3 \cap C_2) = \frac{0}{17}$ $\quad P(\omega_3 \cap C_3) = \frac{3}{17}$

$$I(\omega, C) = \sum_{k=1}^{2}\sum_{j=1}^{3} P(\omega_k \cap C_j) \log \frac{P(\omega_k \cap C_j)}{P(\omega_k)\, P(C_j)} = 0.4496$$

$$\therefore NMI(\omega, C) = \frac{I(\omega, C)}{(H(\omega) + H(C))/2} = \frac{0.4496}{(1.055 + 1.095)/2} = 0.4182$$

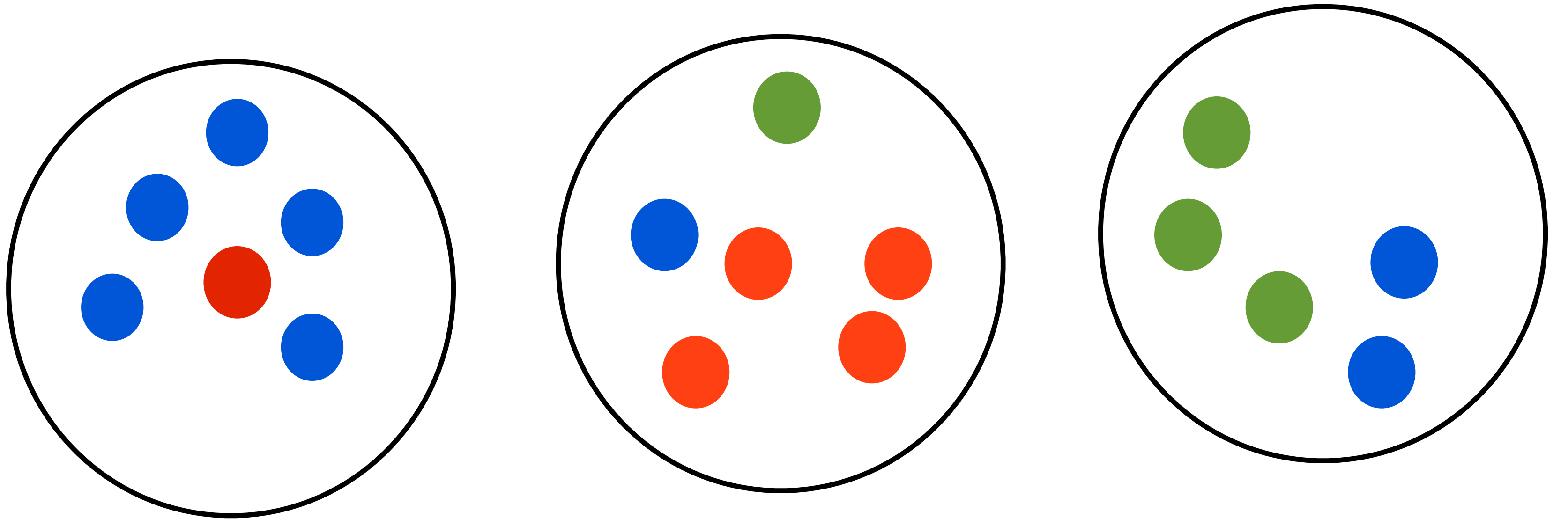
Evaluating Clustering — Rand Index (RI)

- Build a contingency table considering pairs of items in each cluster
  - Positive = same cluster
  - Negative = different clusters
  - True = same class
  - False = different classes
- TP = No. of item pairs that are in the same cluster and belong to the same class
- FP = No. of item pairs that are in the same cluster but belong to different classes
- TN = No. of item pairs that are in different clusters and belong to different classes
- FN = No. of item pairs that are in different clusters but belong to the same class
-

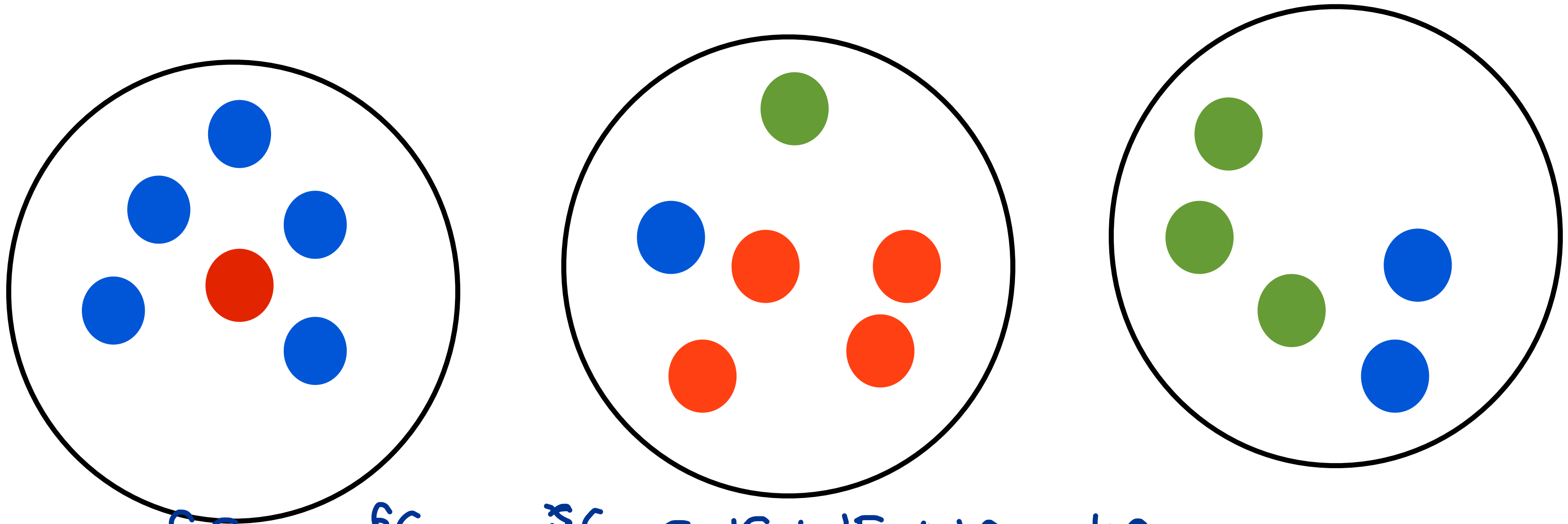
contingency table	same cluster	different clusters
same class	TP	FN
different classes	FP	TN

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}$$

(accuracy of the clustering)



Quiz: Compute RI for this clustering.



$$TP + FP = {}^6C_2 + {}^6C_2 + {}^5C_2 = 15 + 15 + 10 = 40.$$

$$\left[{}^nC_r = \frac{n!}{r!\,(n-r)!} \right]$$

$${}^6C_2 = \frac{6!}{2!\,4!} = \frac{6 \times 5}{2} = 15 \qquad {}^5C_2 = \frac{5!}{2!\,3!} = \frac{5 \times 4}{2} = 10.$$

$$TP = {}^5C_2 + {}^4C_2 + {}^3C_2 + {}^2C_2 = 10 + 6 + 3 + 1 = 20$$

$${}^4C_2 = \frac{4!}{2!\,2!} = \frac{4 \times 3}{2} = 6$$

$$\therefore FP = 40 - 20 = 20.$$

$$FN = (5 \times 1) + (1 \times 4) + (5 \times 2) + (1 \times 2) + (1 \times 3)$$
$$= 5 + 4 + 10 + 2 + 3 = 24$$

$$TN + FN = (5+1)(1+1+4) + (5+1)(3+2) + (1+1+4)(3+2)$$
$$= 6 \times 6 + 6 \times 5 + 6 \times 5 = 36 + 30 + 30 = 96.$$

$$\therefore TN = 96 - 24 = 72$$

	same cluster	different clusters
same class	20	24
different classes	20	72

RI = (20+72) / (20+24+20+72)
= 0.676

Evaluating Clustering — P/R/F

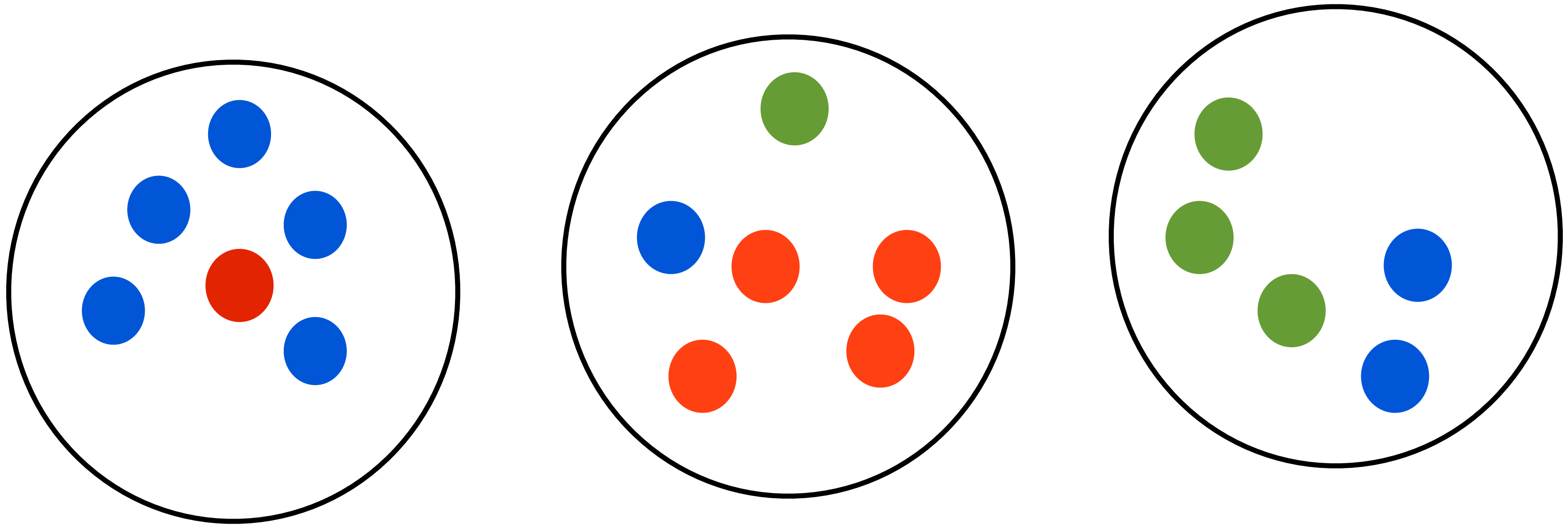
- We can use Precision (P), Recall (R), and F-measure (F) at to evaluate the accuracy of a clustering.
- For this purpose we must first create the contingency table as we did for RI and then compute P, R, F as follows

$$P = TP / (TP + FP)$$

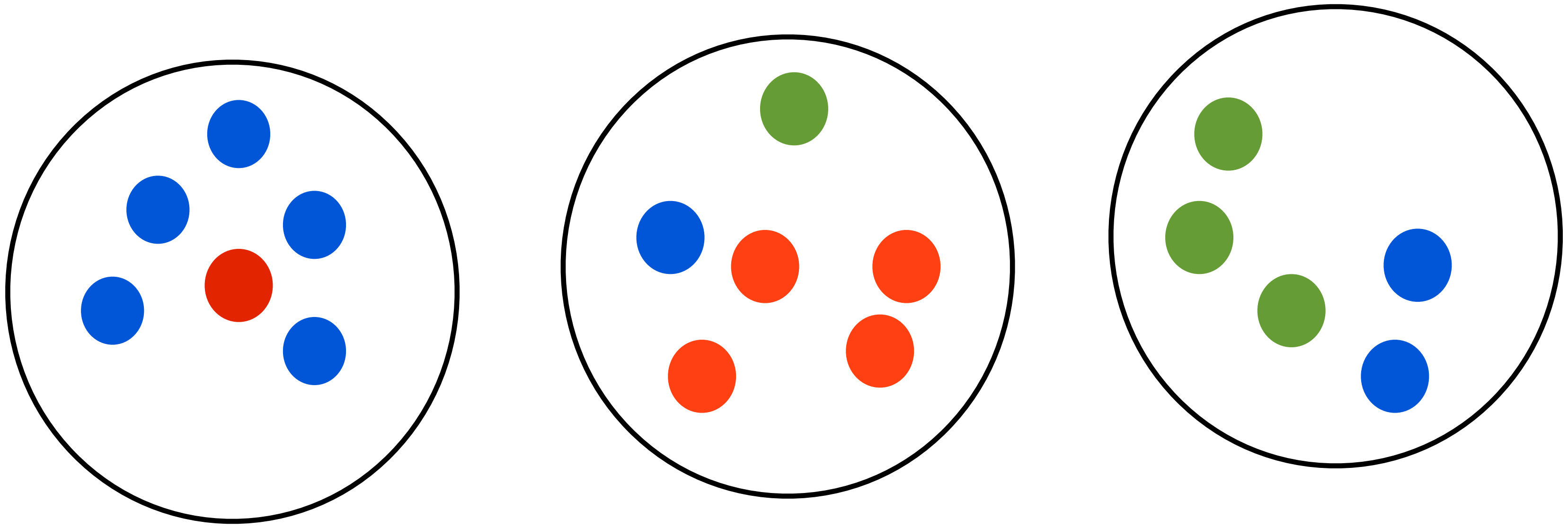
$$R = TP / (TP + FN)$$

$$F = 2PR / (P + R)$$

Ref: https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html



Quiz: Compute P/R/F for this clustering.



	same cluster	different clusters
same class	TP=20	FN=24
different classes	FP=20	TN=72

$$P = TP / (TP + FP) = 20 / (20+20) = 0.5$$

$$R = TP / (TP + FN) = 20 / (20 + 24) = 0.45$$

$$F = 2PR / (P + R) = 0.47$$

B-CUBED Measure

- Proposed in (Bagga B. Baldwin = B^3)
    - A. Bagga and B. Baldwin. Entity-based cross document coreference resolution using the vector space model, In Proc. of 36th COLING-ACL, pages 79--85, 1998.
- We would like to evaluate clustering without labelling any clusters.

$$\text{precision}(x) = \frac{\text{No. of items in C(x) with A(x)}}{\text{No. of items in C(x)}}$$

$$\text{recall}(x) = \frac{\text{No. of items in C(x) with A(x)}}{\text{Total no. of items with A(x)}}$$

$$\text{C(x): The ID of the cluster that x belongs to}$$

$$\text{A(x): label of x}$$

B-CUBED Measure

- Compute the average over all the items (instances) that appear in all clusters (N)

$$\text{Precision} = \frac{1}{N} \sum_{p \in DataSet} \text{Precision}(p)$$

$$\text{Recall} = \frac{1}{N} \sum_{p \in DataSet} \text{Recall}(p)$$

$$F\text{--Score} = \frac{1}{N} \sum_{p \in DataSet} \text{F}(p)$$

Hierarchical Clustering

- Sometimes we might want to organise the data into a hierarchy of subsuming concepts for visualisation (abstraction) purposes
- Two methods exists
  - Conglomerative clustering
    - Start from one big cluster with all data instances and repeatedly partition it
    - Top-down approach
  - Agglomerative clustering
    - Start singletons (clusters with exactly one instance) and iteratively merge the most *similar* two clusters
      - Bottom-up approach
      - computationally more efficient (O(log*n*) merges required)

Merging two clusters

- Single linkage
  - Distance between two clusters A and B is the smallest distance between any instance a ∈ A and b ∈ B

$$D(\mathcal{A},\mathcal{B}) = \min_{a\in\mathcal{A},b\in\mathcal{B}} dist(a,b)$$

- Complete linkage
  - Distance between two clusters A and B is the largest distance between any instance a ∈ A and b ∈ B

$$D(\mathcal{A},\mathcal{B}) = \max_{a\in\mathcal{A},b\in\mathcal{B}} dist(a,b)$$

- Average linkage (Group-Average)
  - Average of all the pairs selected from each cluster

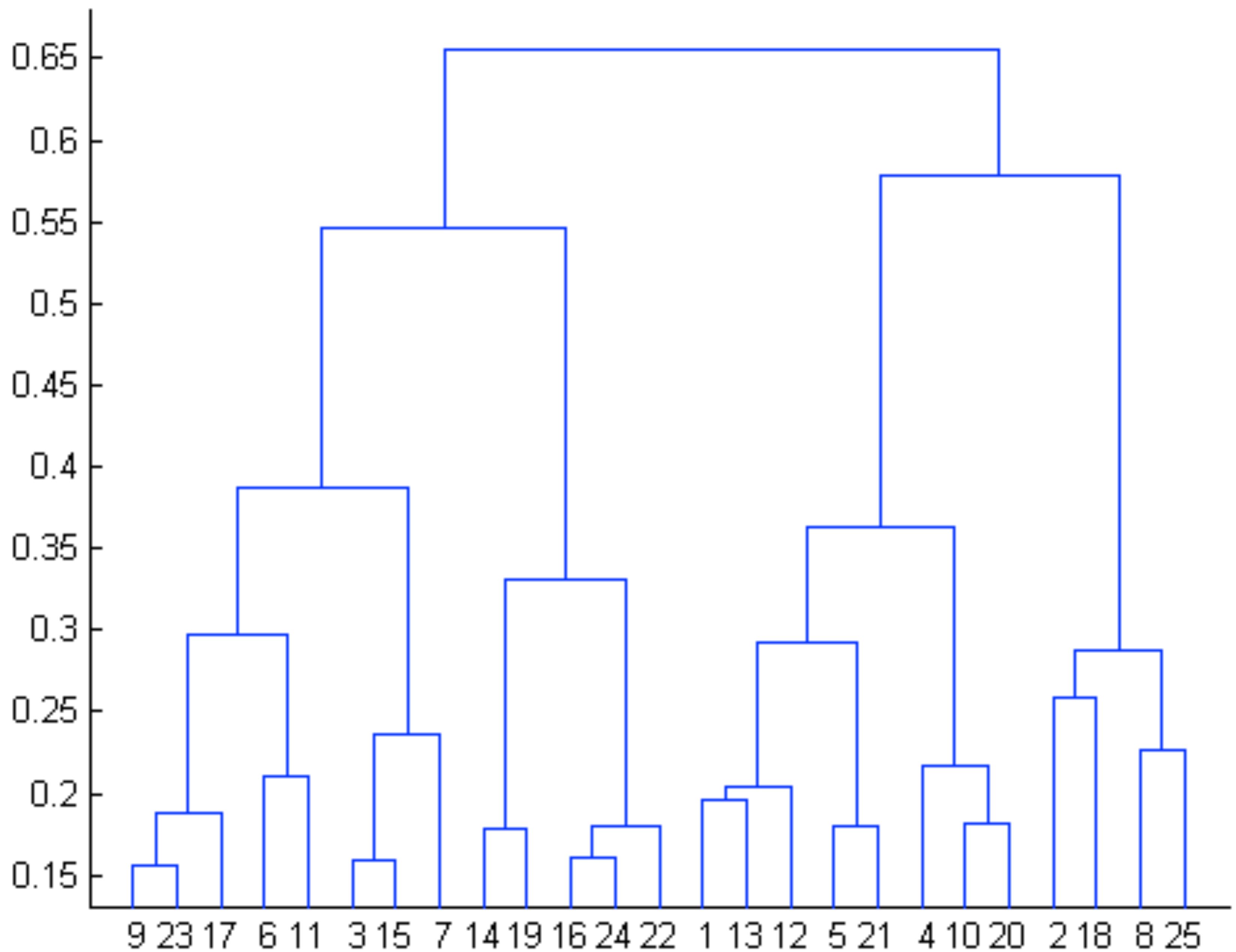
$$D(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{A}||\mathcal{B}|} \sum_{a\in\mathcal{A},b\in\mathcal{B}} dist(a,b)$$

Quiz: Let us assume that in the 2D space there are two clusters {A,B,C} and {P,Q,R}. Which of the distances correspond to the single link and complete link distances between the shown clusters?

Group-Average Agglomerative Clustering

- INPUT:
  - A set of N data instances $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, Number of clusters *k*
- Initialise
  - Create singletons $S_i = \{\mathbf{x}_i\}$ for i = 1, ..., N
- Repeat until only we are left with one cluster
  - Merge the two clusters S_i and S_j with the minimum distance (cf. maximum similarity)
- $$D(\mathcal{S}_i, \mathcal{S}_j) = \frac{1}{|\mathcal{S}_i||\mathcal{S}_j|} \sum_{a \in \mathcal{S}_i, b \in \mathcal{S}_j} dist(a, b)$$

Dendrogram



0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
9 23 17 6 11 3 15 7 14 19 16 24 22 1 13 12 5 21 4 10 20 2 18 8 25