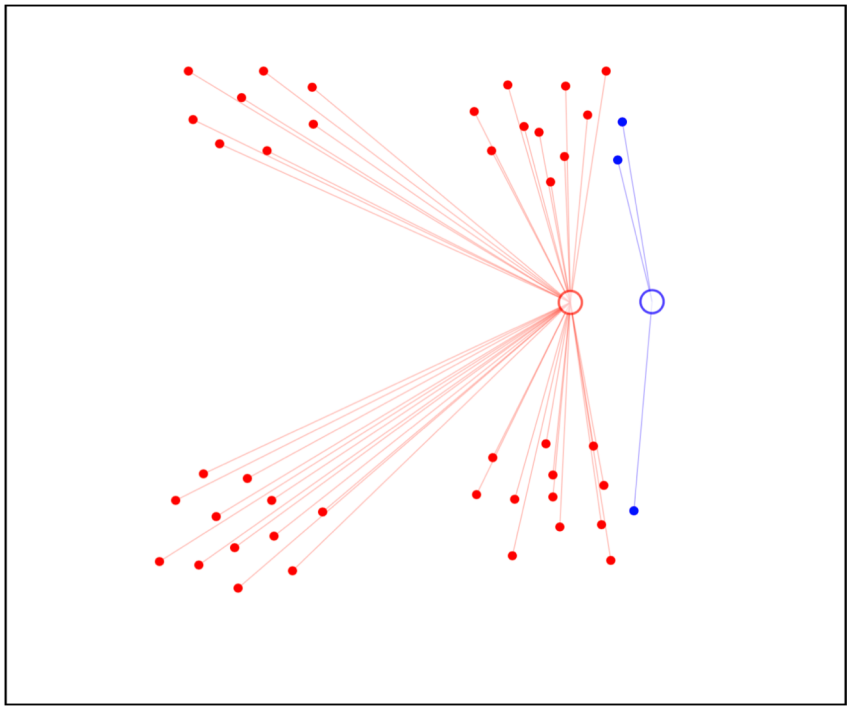
K-means

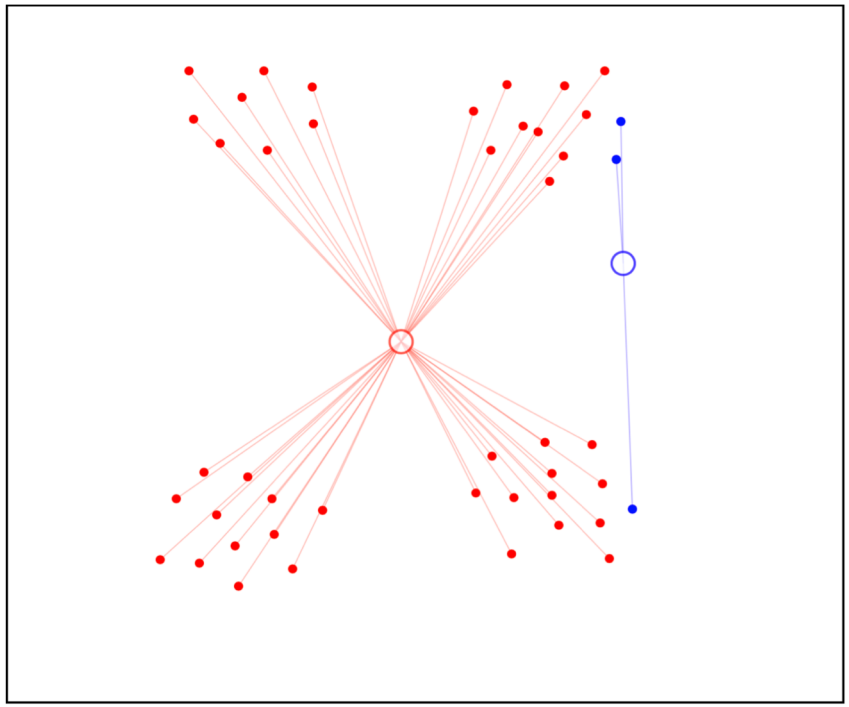
- Centroid-based: describe each cluster by its mean
- Goal: assign data to K.
- Algorithm objective: minimize the within-cluster variances of all clusters.

UNIVERSITY OF BIRMINGHAM

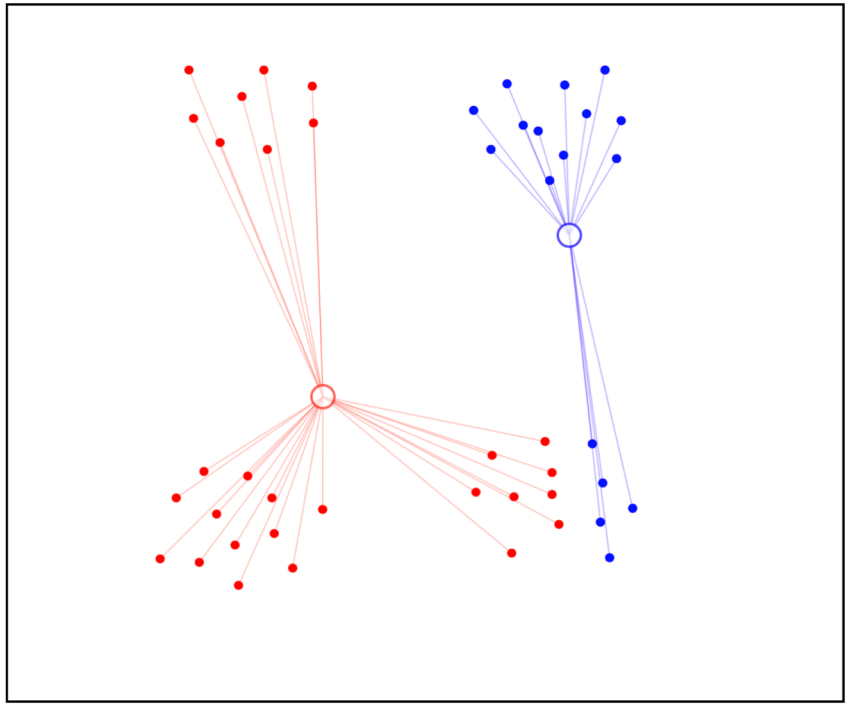
Initialize 2 clusters and assign points to clusters



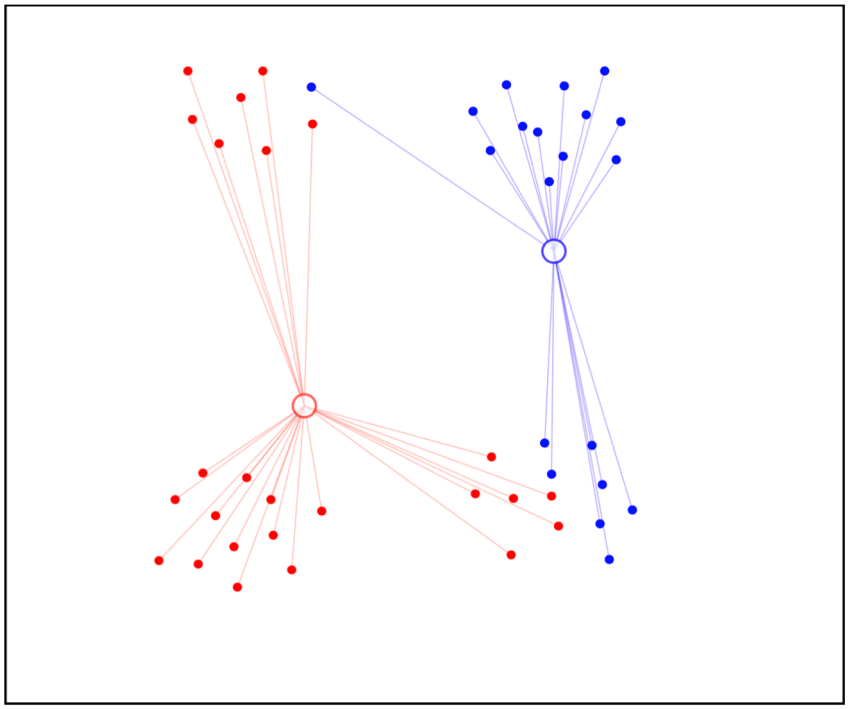
Adjust mean



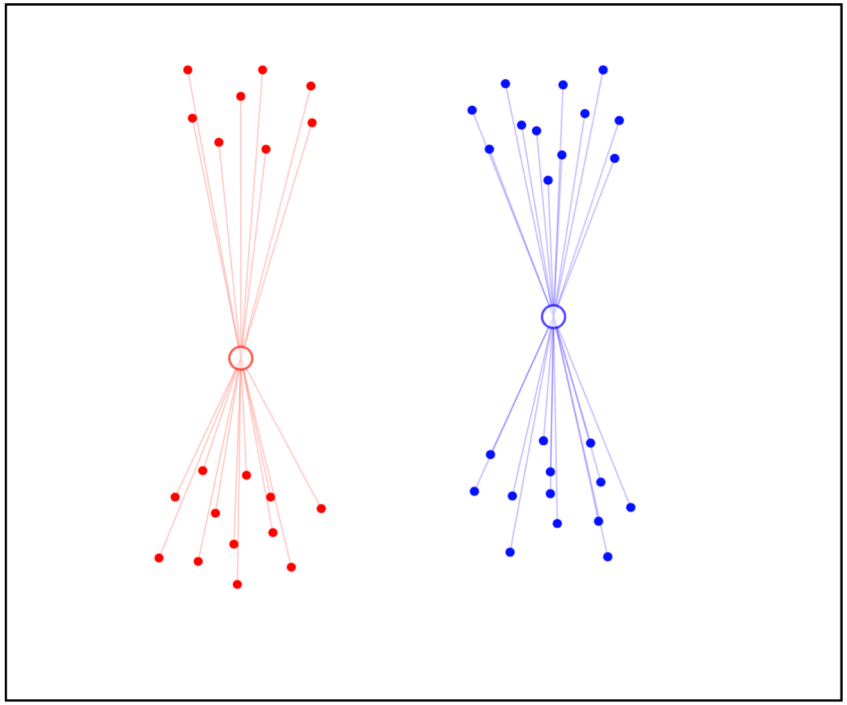
Reassign points to clusters and adjust mean



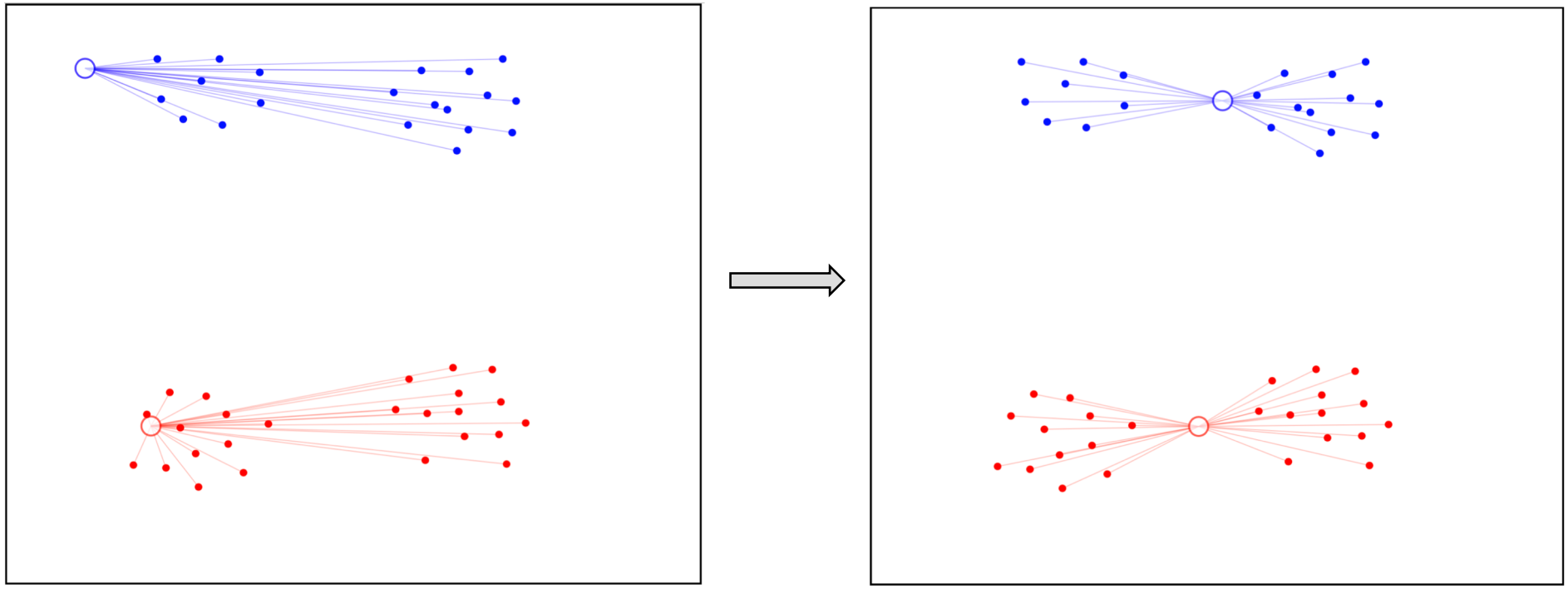
Reassign points to clusters and adjust mean



Repeat this, until no cluster changes



If we have a different starting point



Initial clusters

Final clusters

K-means

- A non-deterministic method
- Finds a local optimal result (multiple restarts are often necessary)

UNIVERSITY OF BIRMINGHAM

Algorithm description

1. Initialization
   - Data are $\mathbf{x}_{1:N}$
   - Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).
2. Repeat
   1. Assign each data point to its closest mean

$$z_n = \arg\min_{i \in \{1,\ldots,k\}} d(\mathbf{x}_n, \mathbf{m}_i)$$

Euclidean distance

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

   2. Compute each cluster mean to be the coordinate-wise average over data points assigned to that cluster,

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\{n\,:\,z_n = k\}} \mathbf{x}_n$$

For each dimension j of x_i in cluster k:

$$(\sum_i x_{i,j})/N_k$$

3. Until assignments $\mathbf{z}_{1:N}$ do not change

UNIVERSITY OF BIRMINGHAM

K-means: finding optimal k

- Plot the cost for each k and find the “Elbow”