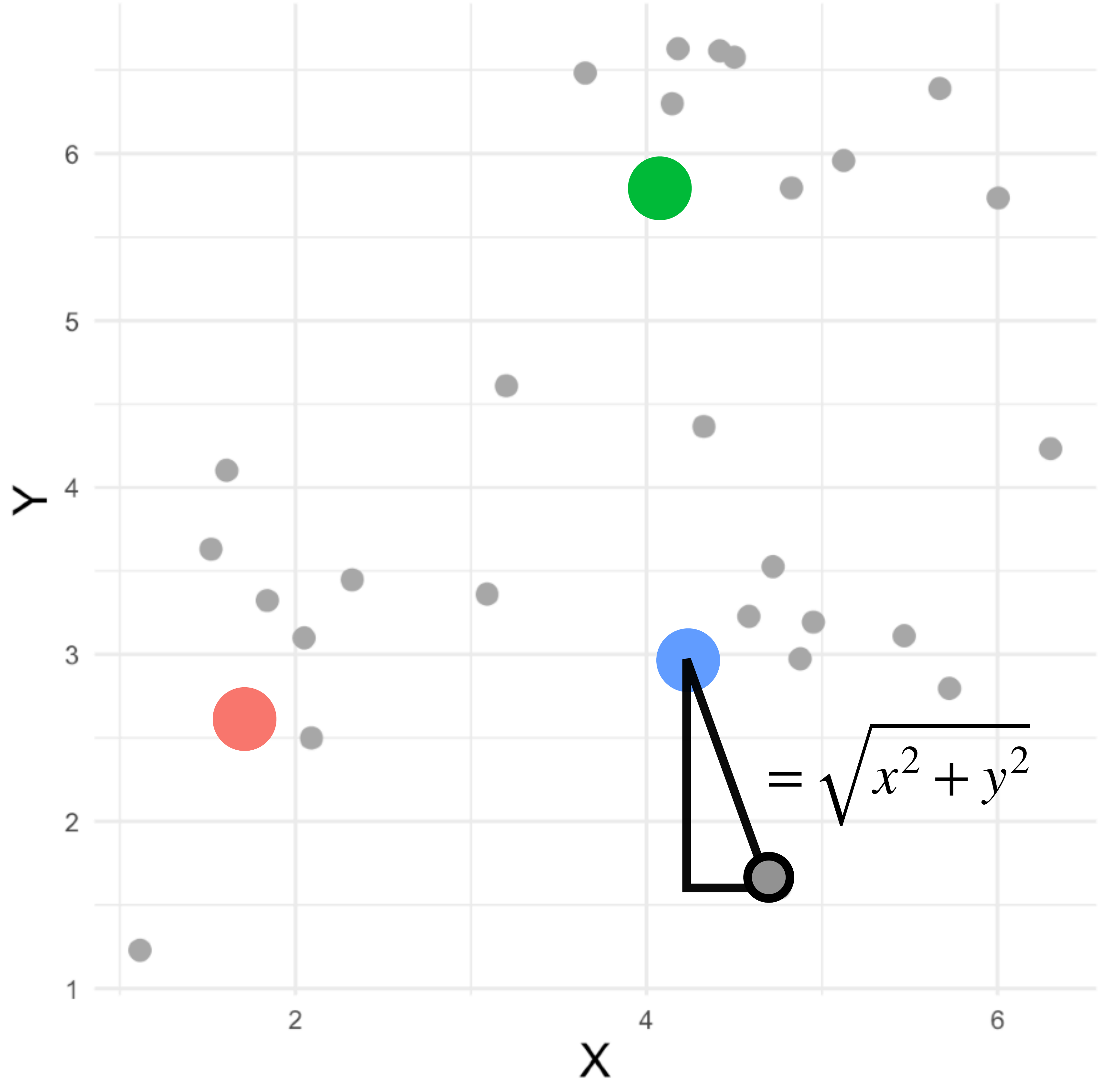
K-Means

$$C_1 \cup C_2 \cup \ldots \cup C_K = \{1,\ldots,n\}$$

$$C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$$

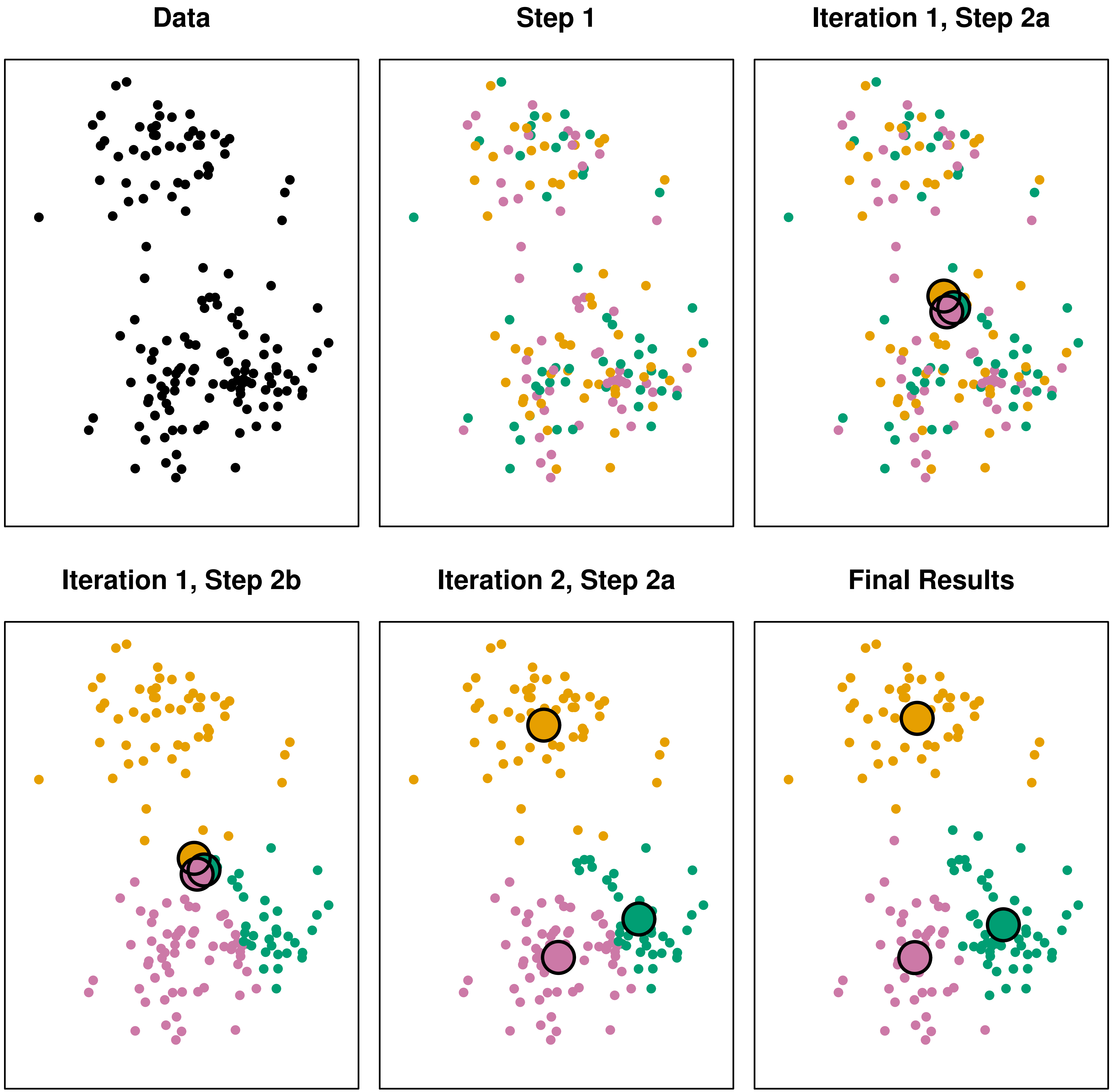
$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \underbrace{W(C_k)}_{\text{within cluster variance}} \right\}$$

$$\text{where } W(C_k) = \frac{1}{|C_k|} \underbrace{\sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}_{\text{squared Euclidean distance}}$$



K-Means: Algorithm

1. Choose **k** random points as cluster centers
2. For each data point, assign it the cluster whose centroid is the closest
3. Using these assignments, recalculate the centers
4. Reiterate from step (2) until **convergence**:
   - cluster membership does not change
   - center only changes very very little



[source: ISLR2]