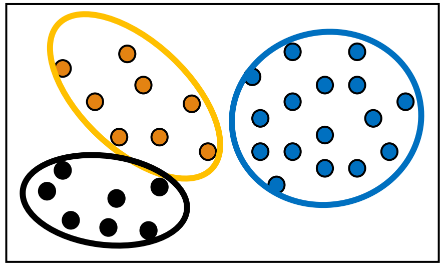
Cluster Stability



- Clusterings obtained from several datasets sampled from the same underlying distribution as ***D*** should be similar or "stable"
- Typical approach:
  - Find good parameter values for a given clustering algorithm
- Example: Find a good value of *k*, the correct number of clusters
- A **bootstrapping approach** to find the best value of *k* (judged on stability)
  - Generate *t* samples of size *n* by sampling from ***D*** with replacement
  - For each sample $\boldsymbol{D}_i$, run the same clustering algorithm with *k* values from 2 to k_{max}
  - Compare the distance between all pairs of clusterings $C_k(\boldsymbol{D}_i)$ and $C_k(\boldsymbol{D}_j)$ via some distance function
    - Compute the expected pairwise distance for each value of *k*
  - The value k^* that exhibits the least deviation between the clusterings obtained from the resampled datasets is the best choice for *k* since it exhibits the most stability