Measuring Clustering Quality: External Methods

❑ Given the **ground truth** T, $Q(C, T)$ is the **quality measure** for a clustering C

❑ $Q(C, T)$ is good if it satisfies the following **four** essential criteria

  ❑ **Cluster homogeneity**

     ❑ The purer, the better

  ❑ **Cluster completeness**

     ❑ Assign objects belonging to the same category in the ground truth to the same cluster

  ❑ **Rag bag better than alien**

     ❑ Putting a heterogeneous object into a pure cluster should be penalized more than putting it into a *rag bag* (i.e., "miscellaneous" or "other" category)

  ❑ **Small cluster preservation**

     ❑ Splitting a small category into pieces is more harmful than splitting a large category into pieces