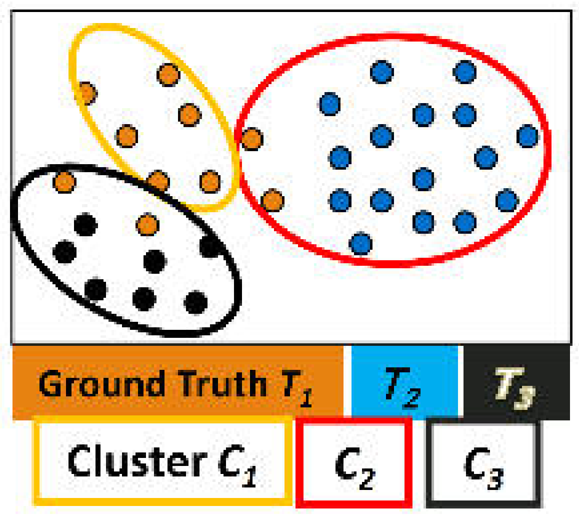
External Measures II: Entropy-Based Measures

Entropy-Based Measures (I): Conditional Entropy

- **Entropy of clustering *C*:** $H(\mathcal{C}) = -\sum_{i=1}^{r} p_{C_i} \log p_{C_i}$ $p_{C_i} = \frac{n_i}{n}$ (i.e., the probability of cluster C_i)



- **Entropy of partitioning *T*:** $H(\mathcal{T}) = -\sum_{j=1}^{k} p_{T_j} \log p_{T_j}$
- **Entropy of *T* with respect to cluster C_i:** $H(\mathcal{T}|C_i) = -\sum_{j=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log\left(\frac{n_{ij}}{n_i}\right)$
- **Conditional entropy of *T* with respect to clustering *C*:** $H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r} \left(\frac{n_i}{n}\right) H(\mathcal{T}|C_i) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log\left(\frac{p_{ij}}{p_{C_i}}\right)$
  - The more a cluster's members are split into different partitions, the higher the conditional entropy
  - For a perfect clustering, the conditional entropy value is 0, where the worst possible conditional entropy value is *log k*

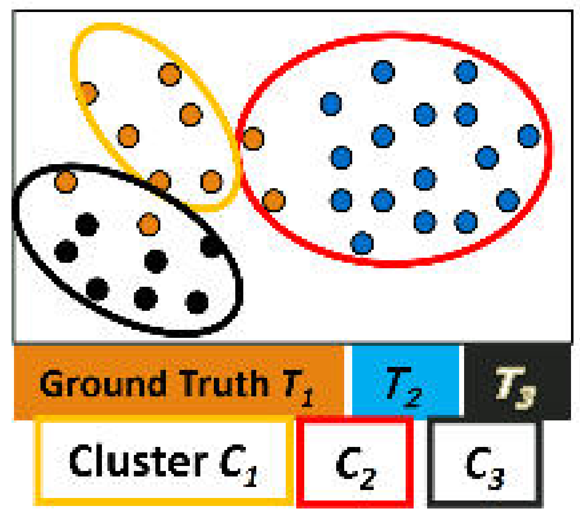
$$H(\mathcal{T}|\mathcal{C}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij}(\log p_{ij} - \log p_{C_i}) = -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log p_{ij} + \sum_{i=1}^{r}(\log p_{C_i} \sum_{j=1}^{k} p_{ij})$$

$$= -\sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log p_{ij} + \sum_{i=1}^{r}(p_{C_i} \log p_{C_i}) = H(\mathcal{C}, \mathcal{T}) - H(\mathcal{C})$$

Entropy-Based Measures (II): Normalized Mutual Information (NMI)

- **Mutual information**:
  - Quantifies the amount of shared info between the clustering *C* and partitioning *T* $I(C,T) = \sum_{i=1}^{r}\sum_{j=1}^{k} p_{ij} \log(\frac{p_{ij}}{p_{C_i} \cdot p_{T_j}})$

    C = T 时 · I(C,T) = H(C) = H(T)
  - Measures the dependency between the observed joint probability p_{ij} of *C* and *T*, and the expected joint probability $p_{Ci} \cdot p_{Tj}$ under the independence assumption
  - When *C* and *T* are independent, $p_{ij} = p_{Ci} \cdot p_{Tj}$, $I(C, T) = 0$. However, there is no upper bound on the mutual information



- **Normalized mutual information** (NMI)

$$NMI(C,T) = \sqrt{\frac{I(C,T)}{H(C)} \cdot \frac{I(C,T)}{H(T)}} = \frac{I(C,T)}{\sqrt{H(C) \cdot H(T)}}$$

  - Value range of NMI: [0,1]. Value close to 1 indicates a good clustering