

The background features a complex, abstract design. It includes a network of thin, reddish-brown lines connecting various points, some of which are green dots. There are also larger, fainter geometric shapes and patterns in shades of brown and grey. A prominent white, angular shape, resembling a stylized 'V' or a folded piece of paper, is positioned behind the title text. On the left side, there is a small, rectangular inset image showing a cluster of orange and red dots on a light background, with a grid of small '+' symbols overlaid.

# **External Measures for Clustering Validation**



# 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

## ① **Cluster homogeneity**

- The purer, the better

## ② **Cluster completeness**

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

## ③ **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)

## ④ **Small cluster preservation**

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

# Commonly Used External Measures

## ■ Matching-based measures (To be covered)

- Purity, maximum matching, F-measure

## ■ Entropy-Based Measures

- Conditional entropy (To be covered)
- Normalized mutual information (NMI) (To be covered)
- Variation of information

## ■ Pairwise measures (To be covered)

- Four possibilities: True positive (TP), FN, FP, TN
- Jaccard coefficient, Rand statistic, Fowlkes-Mallow measure

## ■ Correlation measures

- Discretized Huber static, normalized discretized Huber static

