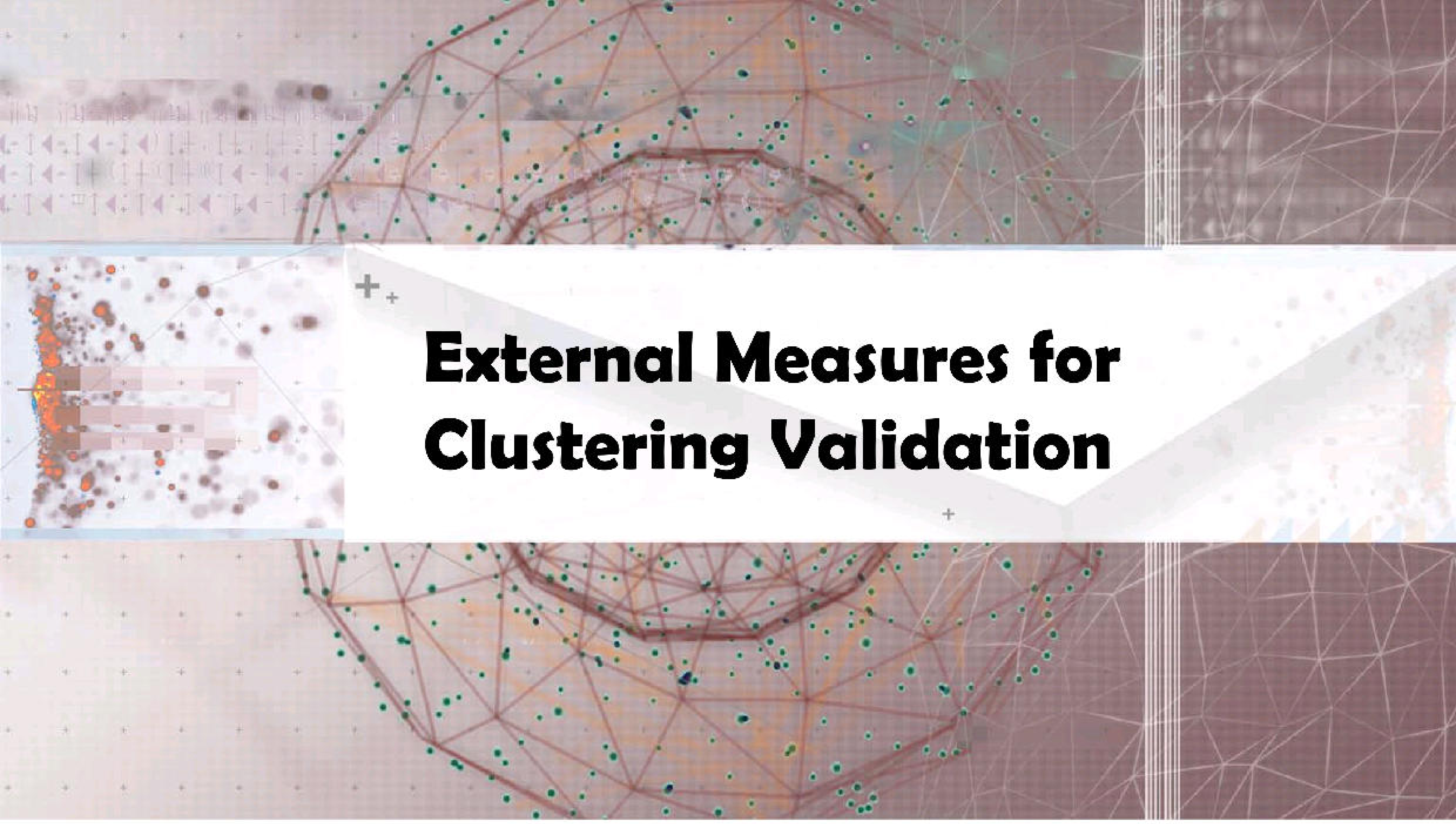
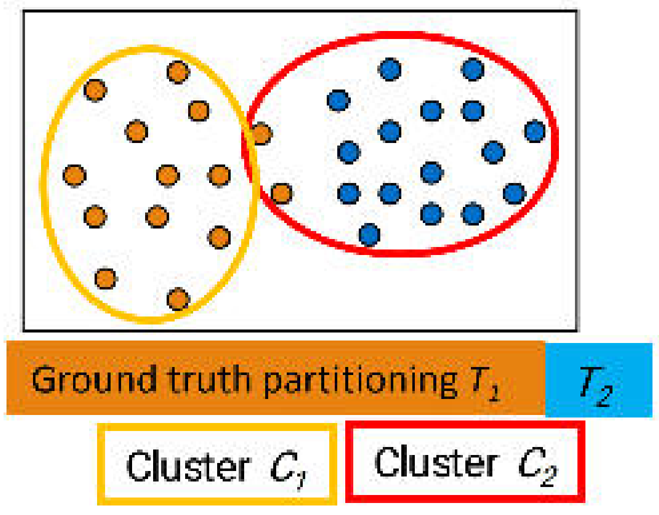


External Measures for Clustering Validation

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