Direct Evaluation of Text Clusters

- Question to answer: How close are the system-generated clusters to the ideal clusters (generated by humans)?
  - "Closeness" can be assessed from multiple perspectives
  - "Closeness" can be quantified
  - "Clustering bias" is imposed by the human assessors
- Evaluation procedure:
  - Given a test set, have humans to create an ideal clustering result (i.e., an ideal partitioning of text objects or "gold standard")
  - Use a system to produce clusters from the same test set
  - Quantify the similarity between the system-generated clusters and the gold standard clusters
  - Similarity can be measured from multiple perspectives (e.g., purity, normalized mutual information, F measure)