Text Clustering

- Problem and Motivation

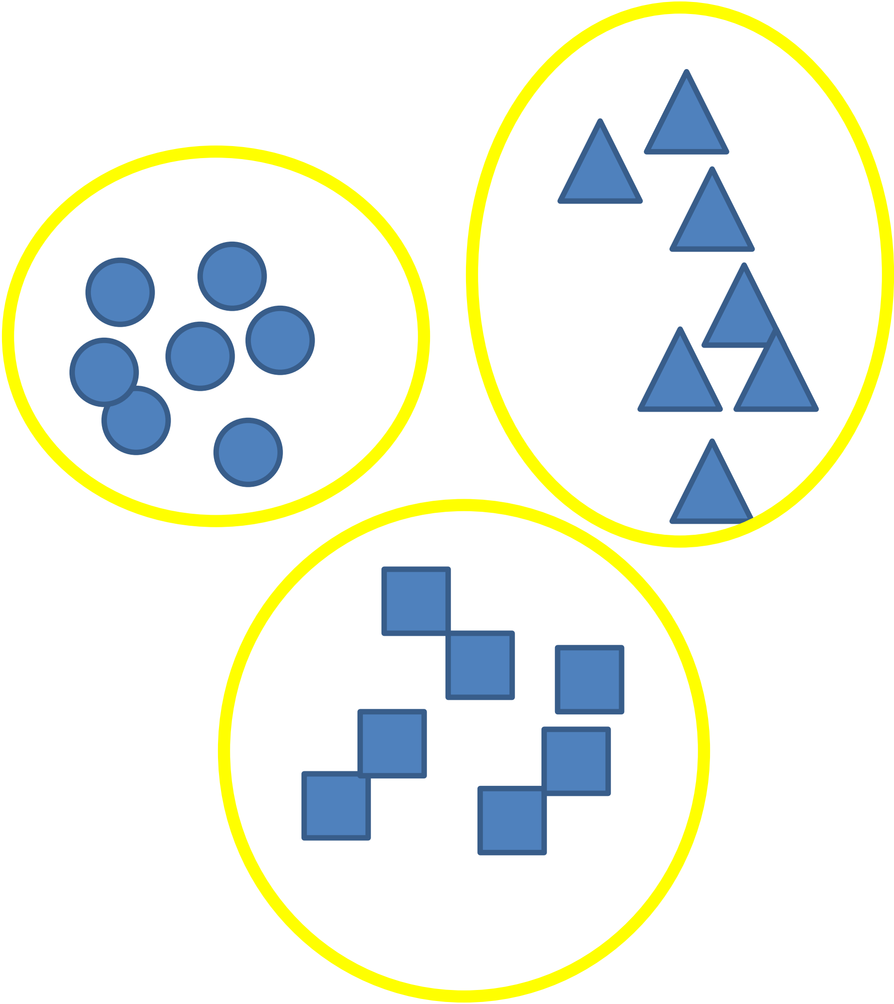
(CPEG 657: Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

Problem Definition

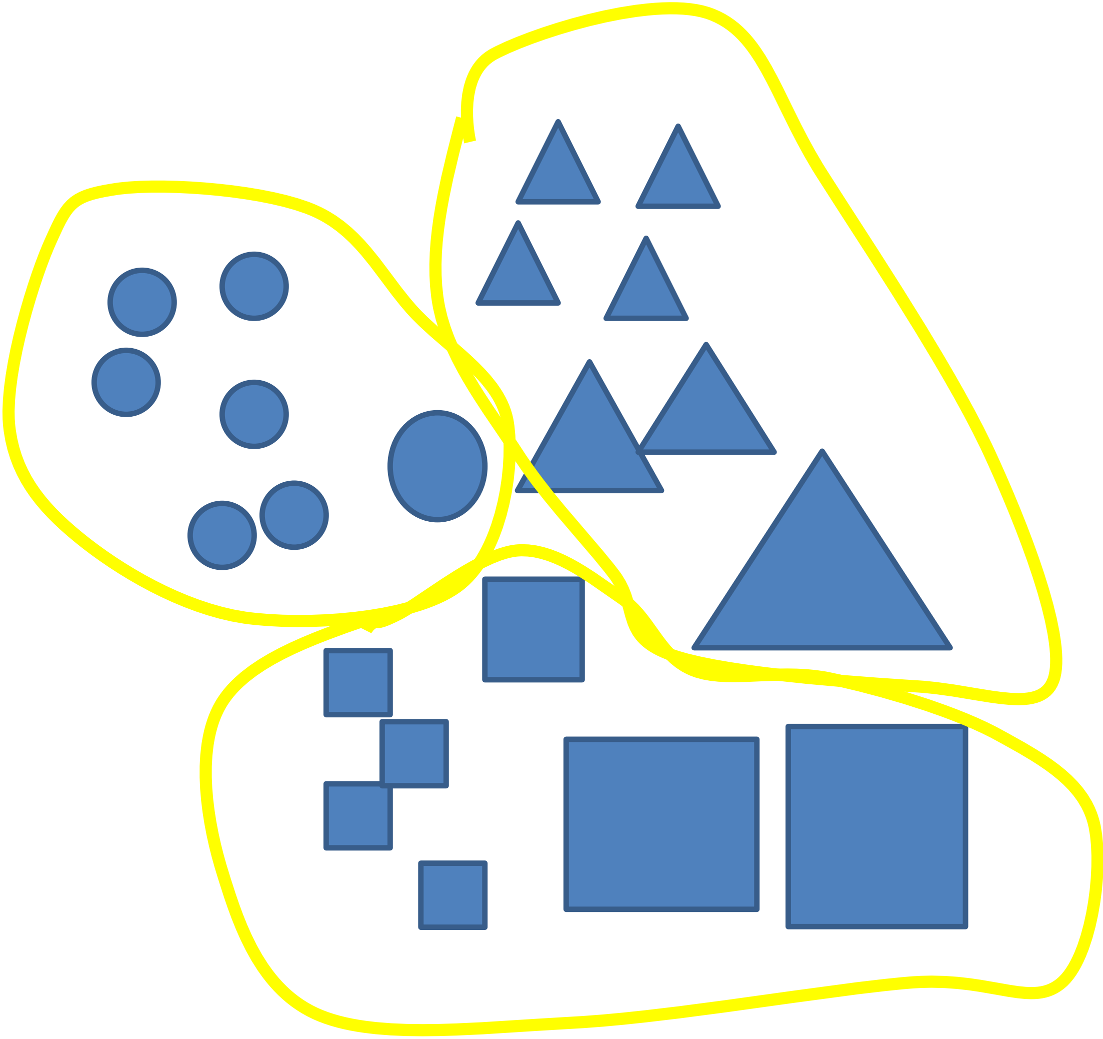
- The process of grouping a set of documents into classes of similar documents
  - Documents within a cluster should be similar.
  - Documents from different clusters should be dissimilar.
- Unsupervised learning
    - learning from raw data

A data set with clear cluster structure



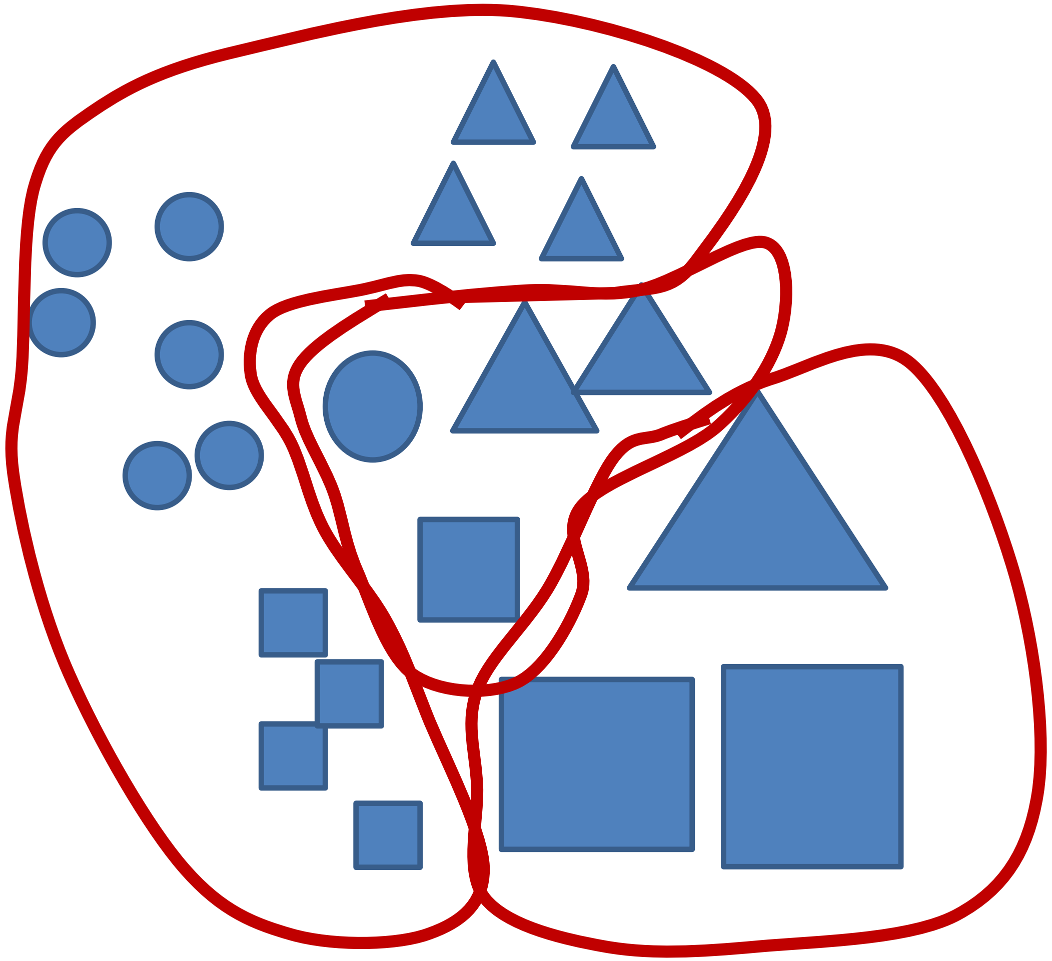
How can we detect the three clusters in this case?

Clustering Bias



- Any two objects can be similar, depending on how you look at them!
- A user must define the similarity in some way

Clustering Bias



- Any two objects can be similar, depending on how you look at them!
- A user must define the similarity in some way

Issues for clustering

- Document representation
- Distance measure
- The number of clusters: fixed vs. dynamic generated

Hard vs. soft clustering

- Hard clustering
  - Each document belongs to exactly one cluster
- Soft clustering
  - A document can belong to more than one cluster.

Clustering Methods

- Similarity-based methods
  - Need a similarity function
  - Construct a partition
  - Typically "hard" clustering
- Model-based methods
  - First compute the model
  - Clusters are obtained easily after having a model
  - Typically "soft" clustering