Hierarchical Clustering
Lecture 15

David Sontag
New York University
Agglomerative Clustering

- **Agglomerative clustering:**
  - First merge very similar instances
  - Incrementally build larger clusters out of smaller clusters

- **Algorithm:**
  - Maintain a set of clusters
  - Initially, each instance in its own cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
    - Stop when there’s only one cluster left

- Produces not one clustering, but a family of clusterings represented by a **dendrogram**
Agglomerative Clustering

- How should we define “closest” for clusters with multiple elements?
Agglomerative Clustering

• How should we define “closest” for clusters with multiple elements?

• Many options:
  – Closest pair  
    (single-link clustering)
  – Farthest pair  
    (complete-link clustering)
  – Average of all pairs

• Different choices create different clustering behaviors
Agglomerative Clustering

• How should we define “closest” for clusters with multiple elements?

- Closest pair (single-link clustering)
- Farthest pair (complete-link clustering)

[Pictures from Thorsten Joachims]
Clustering Behavior

Average  Farthest  Nearest

Mouse tumor data from [Hastie et al.]
Agglomerative Clustering

When can this be expected to work?

**Closest pair**
(single-link clustering)

![Diagram showing closest pair clustering]

**Strong separation** property:
*All points are more similar to points in their own cluster than to any points in any other cluster*

Then, the true clustering corresponds to some **pruning** of the tree obtained by single-link clustering!

Slightly weaker (stability) conditions are solved by average-link clustering

(Balcan et al., 2008)