ONLINE MASTERS IN **DATA SCIENCE**

DSC 257R - UNSUPERVISED LEARNING

HIERARCHICAL CLUSTERING

SANJOY DASGUPTA, PROFESSOR

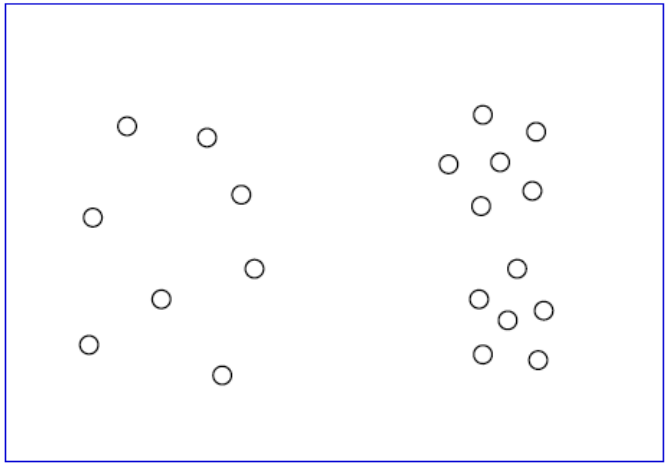
UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Hierarchical Clustering

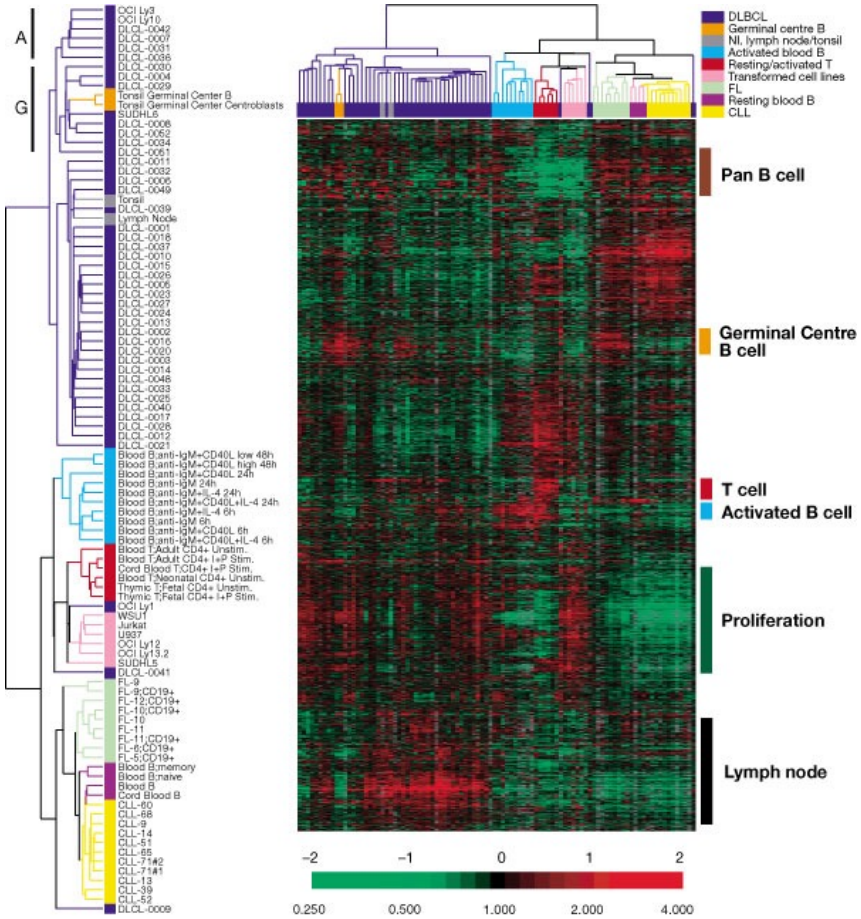
Choosing the number of clusters (k) is difficult.



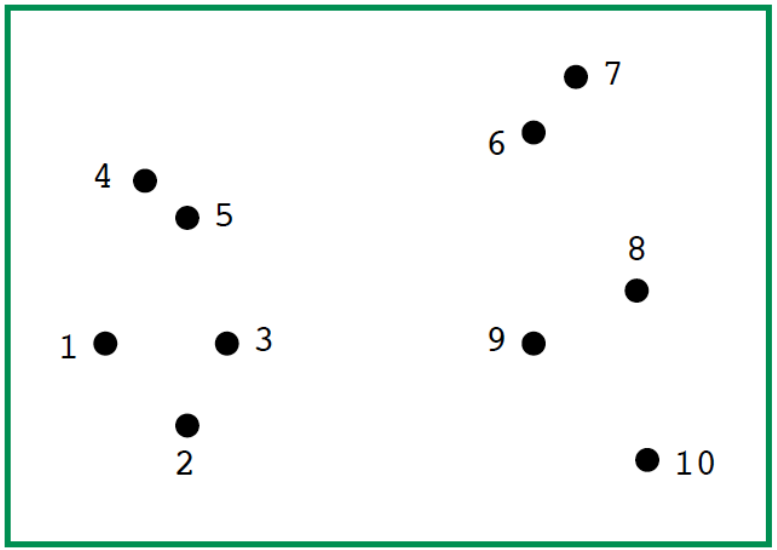
Often: no single right answer, because of multiscale structure.

Hierarchical clustering avoids these problems.

Example: Gene Expression Data



The Single Linkage Algorithm



- Start with each point in its own, singleton, cluster
- Repeat until there is just one cluster:
  - Merge the two clusters with the closest pair of points
- Disregard singleton clusters

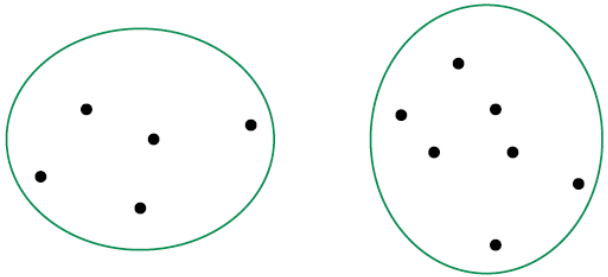
Linkage Methods

- Start with each point in its own, singleton, cluster
- Repeat until there is just one cluster:
  - Merge the two “closest” clusters

Linkage Methods

- Start with each point in its own, singleton, cluster
- Repeat until there is just one cluster:
  - Merge the two "closest" clusters

How to measure distance between two clusters C and C'?



- Single linkage

$$\text{dist}(C, C') = \min_{x \in C, x' \in C'} \|x - x'\|$$

- Complete linkage

$$\text{dist}(C, C') = \max_{x \in C, x' \in C'} \|x - x'\|$$

Average Linkage

Three commonly-used variants:

1. Average pairwise distance between points in the two clusters

$$\text{dist}(C, C') = \frac{1}{|C| \cdot |C'|} \sum_{x \in C} \sum_{x' \in C'} \|x - x'\|$$

2. Distance between cluster centers

$$\text{dist}(C, C') = \|\text{mean}(C) - \text{mean}(C')\|$$

3. Ward's method: the increase in k-means cost occasioned by merging the two clusters

$$\text{dist}(C, C') = \frac{|C| \cdot |C'|}{|C| + |C'|} \|\text{mean}(C) - \text{mean}(C')\|^2$$