CS434a/541a: Pattern Recognition
Prof. Olga Veksler

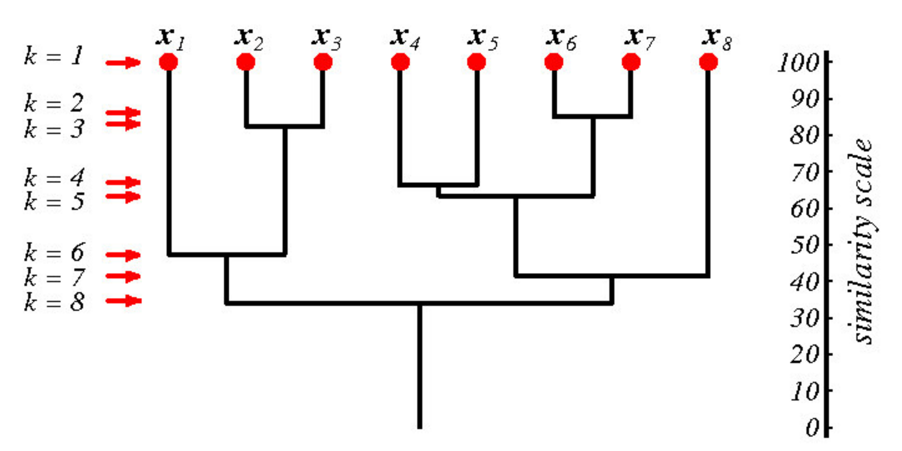
Lecture 16

Today

- Continue Clustering
  - Last Time
    - “Flat Clustring”
  - Today
    - Hierarchical Clustering
      - Divisive
      - Agglomerative
  - Applications of Clustering

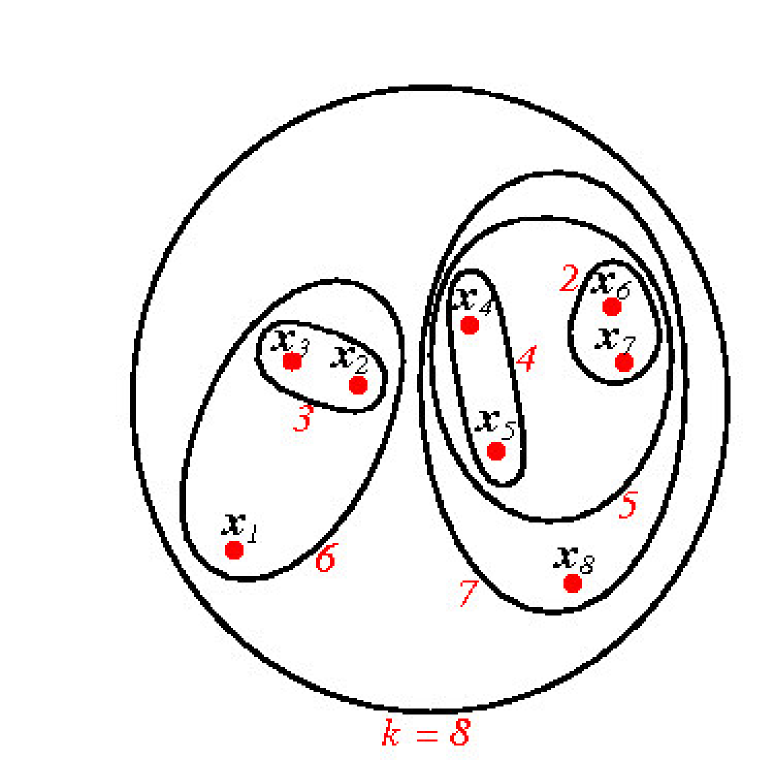
Hierarchical Clustering: Dendogram

- preferred way to represent a hierarchical clustering is a dendrogram
  - Binary tree
  - Level ***k*** corresponds to partitioning with ***n-k+1*** clusters
  - if need ***k*** clusters, take clustering from level ***n-k+1***
  - If samples are in the same cluster at level ***k***, they stay in the same cluster at higher levels
  - dendrogram typically shows the similarity of grouped clusters



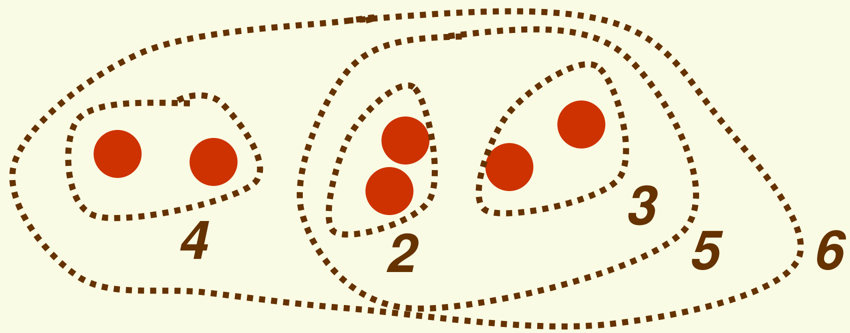
Hierarchical Clustering: Venn Diagram

- Can also use Venn diagram to show hierarchical clustering, but similarity is not represented quantitatively



Hierarchical Clustering

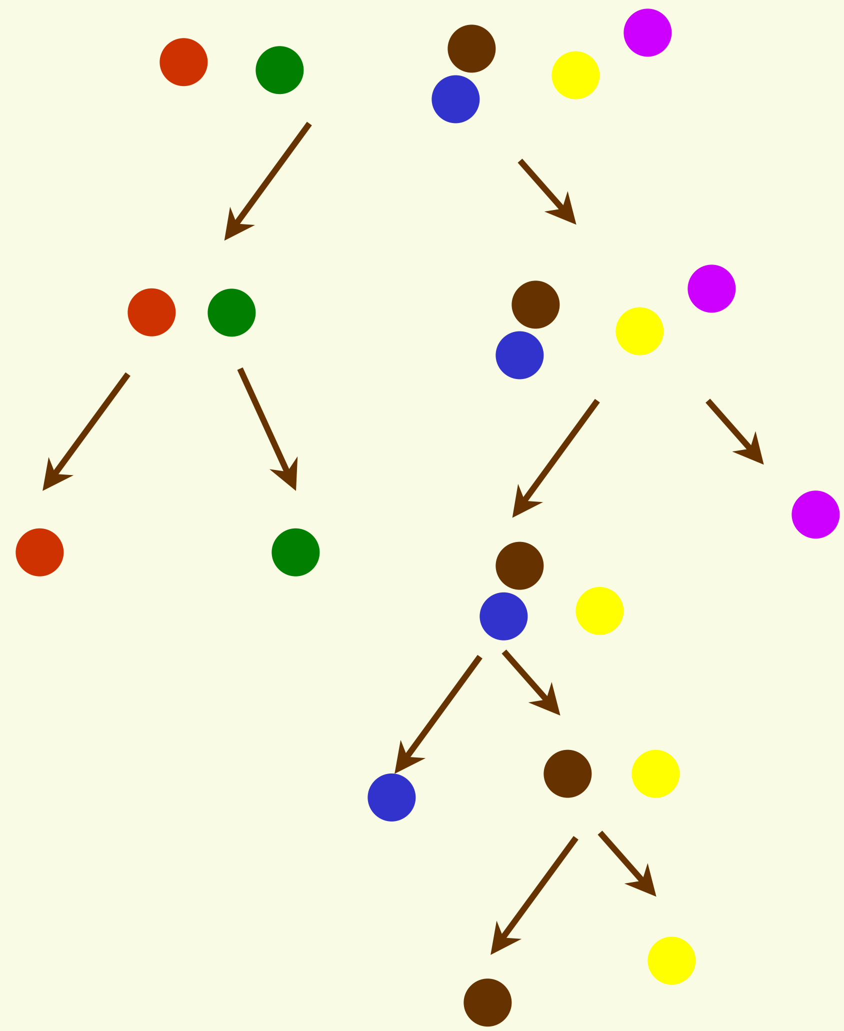
- Algorithms for hierarchical clustering can be divided into two types:
  1. Agglomerative (bottom up) procedures
     - Start with ***n*** singleton clusters
     - Form hierarchy by merging most similar clusters



  2. Divisive (top bottom) procedures
     - Start with all samples in one cluster
     - Form hierarchy by splitting the “worst” clusters

Divisive Hierarchical Clustering

- Any "flat" algorithm which produces a fixed number of clusters can be used
  - set ***c*** = ***2***

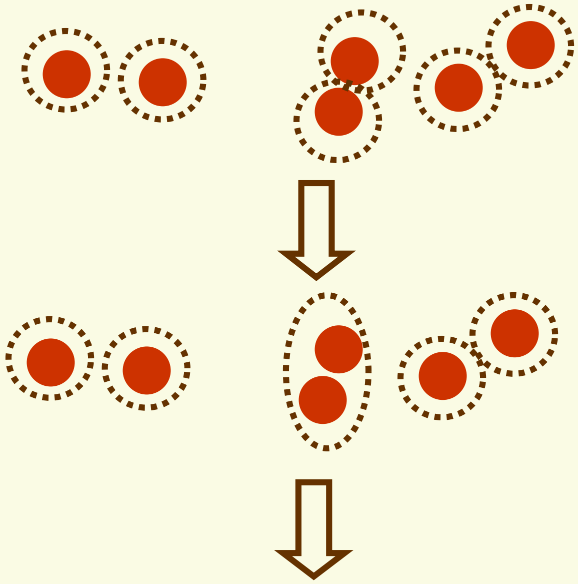


Agglomerative Hierarchical Clustering

initialize with each example in singleton cluster

while there is more than ***1*** cluster

1. find 2 nearest clusters
2. merge them



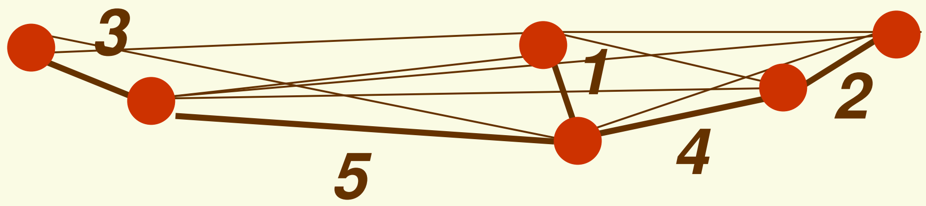
- Four common ways to measure cluster distance

1. minimum distance $$d_{\min}(D_i, D_j) = \min_{x \in D_i, y \in D_j} \| x - y \|$$
2. maximum distance $$d_{\max}(D_i, D_j) = \max_{x \in D_i, y \in D_j} \| x - y \|$$
3. average distance $$d_{avg}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{x \in D_i} \sum_{y \in D_j} \| x - y \|$$
4. mean distance $$d_{mean}(D_i, D_j) = \| \mu_i - \mu_j \|$$

Single Linkage or Nearest Neighbor

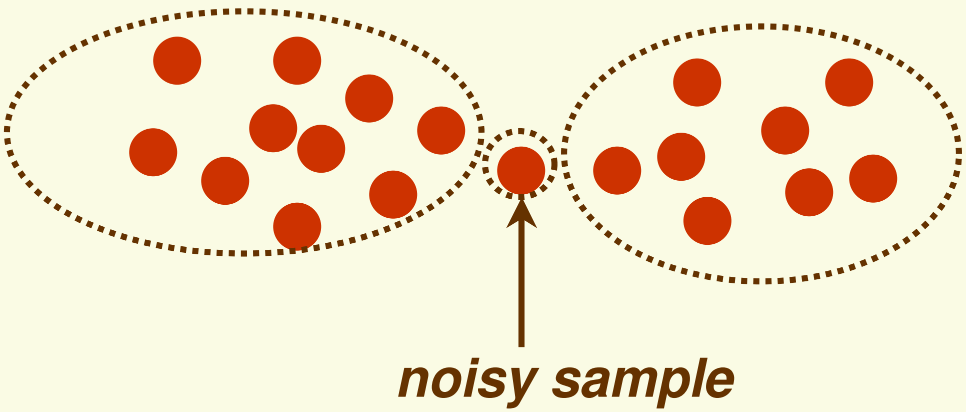
- Agglomerative clustering with minimum distance

$$d_{\min}(D_i, D_j) = \min_{x \in D_i, y \in D_j} \| x - y \|$$

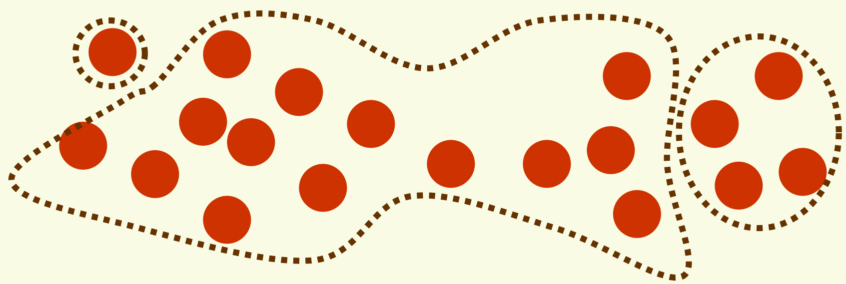


- generates minimum spanning tree
- encourages growth of elongated clusters
- disadvantage: very sensitive to noise

*what we want at level with **c=3***



*what we get at level with **c=3***

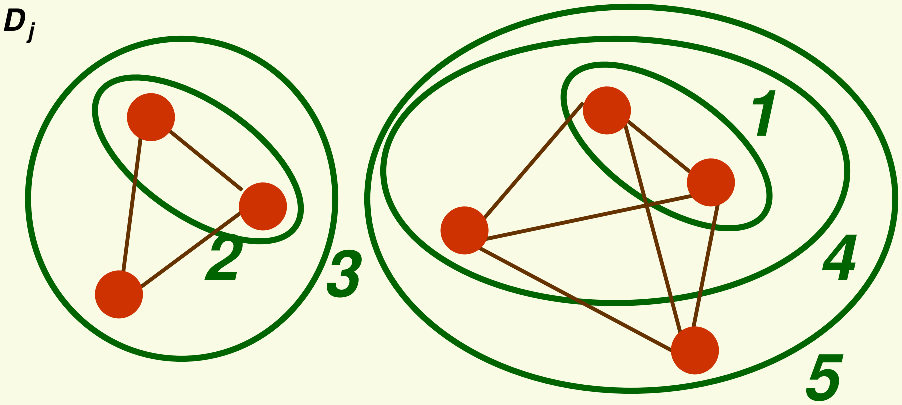


Complete Linkage or Farthest Neighbor

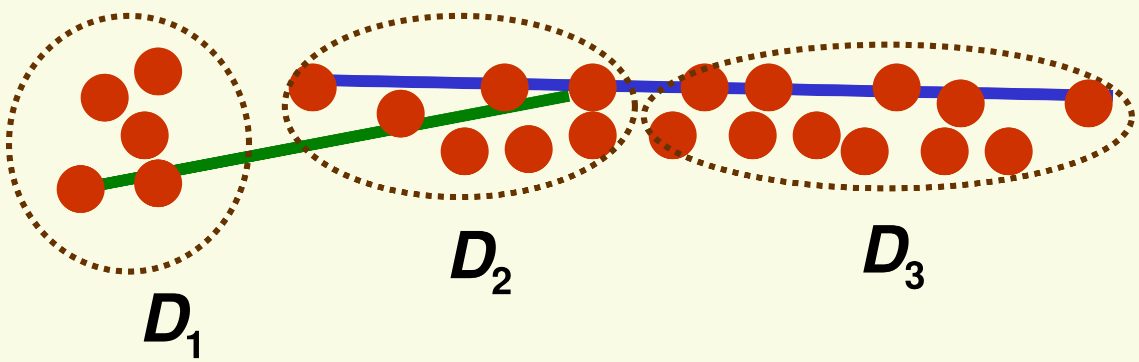
- Agglomerative clustering with maximum distance

$$d_{\max}(D_i, D_j) = \max_{x \in D_i, y \in D_j} \| x - y \|$$

- encourages compact clusters



- Does not work well if elongated clusters present



  - $d_{\max}(D_1, D_2) < d_{\max}(D_2, D_3)$
  - thus D_1 and D_2 are merged instead of D_2 and D_3

Average and Mean Agglomerative Clustering

- Agglomerative clustering is more robust under the average or the mean cluster distance

$$d_{avg}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{x \in D_i} \sum_{y \in D_j} \| x - y \|$$

$$d_{mean}(D_i, D_j) = \| \mu_i - \mu_j \|$$

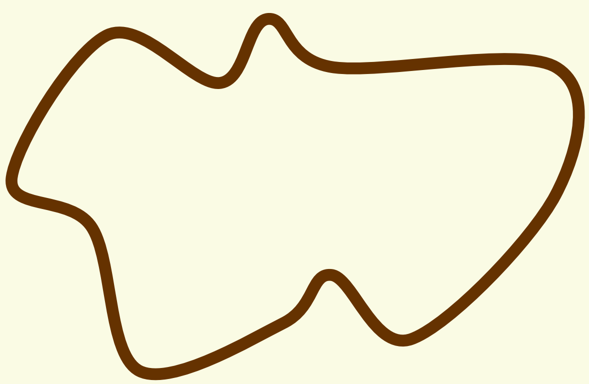
- mean distance is cheaper to compute than the average distance
- unfortunately, there is not much to say about agglomerative clustering theoretically, but it does work reasonably well in practice

Agglomerative vs. Divisive

- Agglomerative is faster to compute, in general
- Divisive may be less “blind” to the global structure of the data

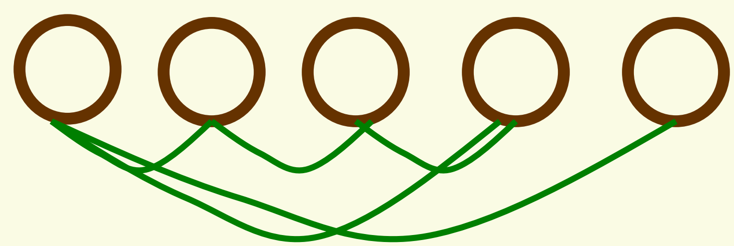
Divisive

when taking the first step (split), have access to all the data; can find the best possible split in 2 parts

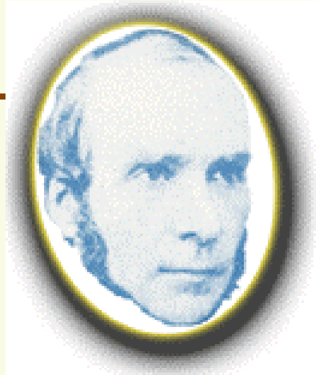


Agglomerative

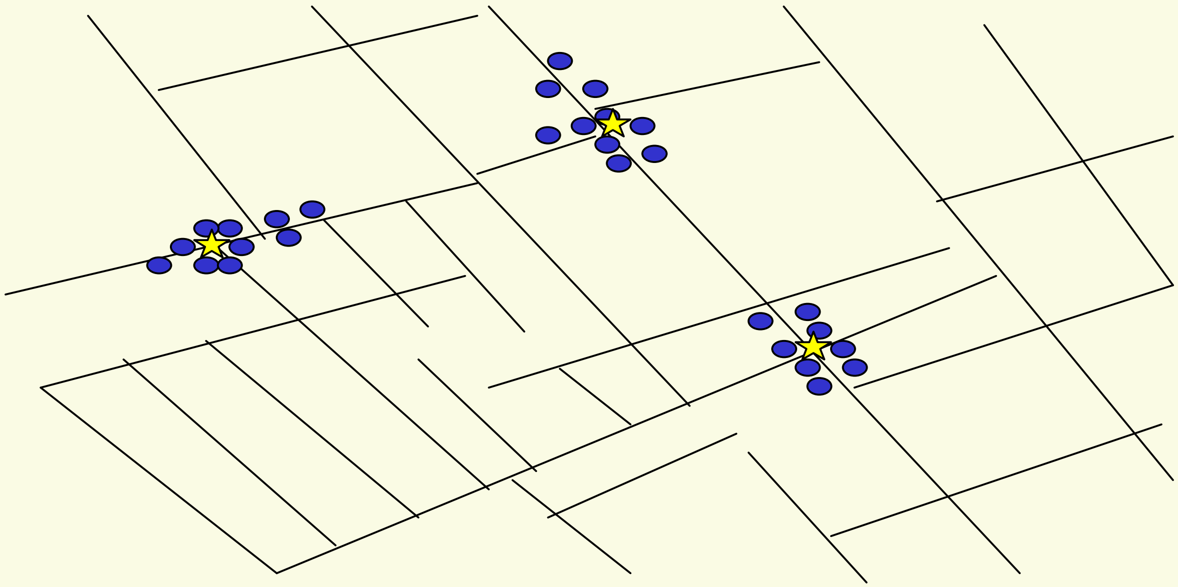
when taking the first step merging, do not consider the global structure of the data, only look at pairwise structure



First (?) Application of Clustering



- John Snow, a London physician plotted the location of cholera deaths on a map during an outbreak in the 1850s.
- The locations indicated that cases were clustered around certain intersections where there were polluted wells -- thus exposing both the problem and the solution.



From: Nina Mishra HP Labs

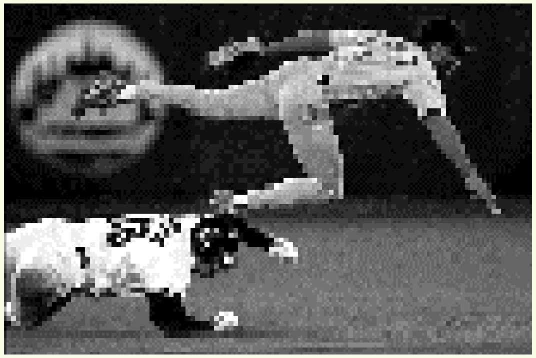
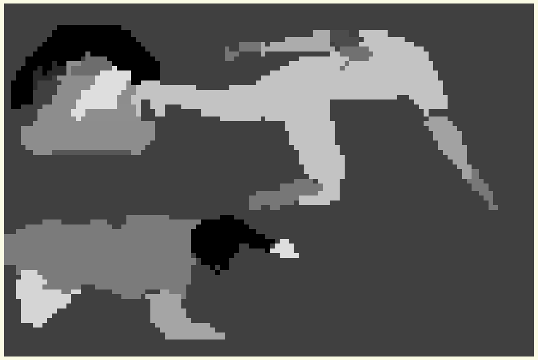
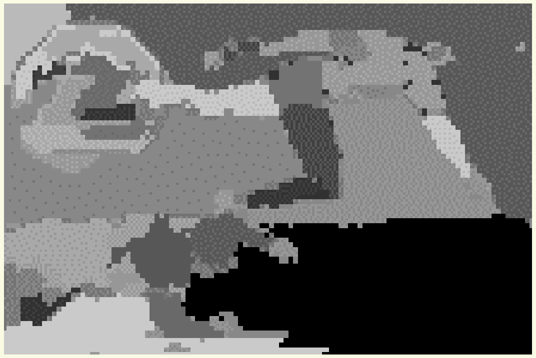
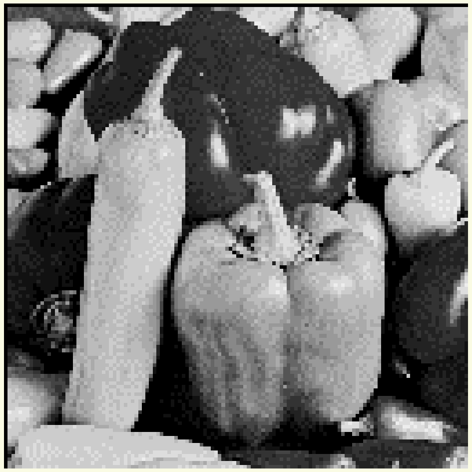
Application of Clustering

- Astronomy
  - SkyCat: Clustered $2x10^9$ sky objects into stars, galaxies, quasars, etc based on radiation emitted in different spectrum bands.

From: Nina Mishra HP Labs

Applications of Clustering

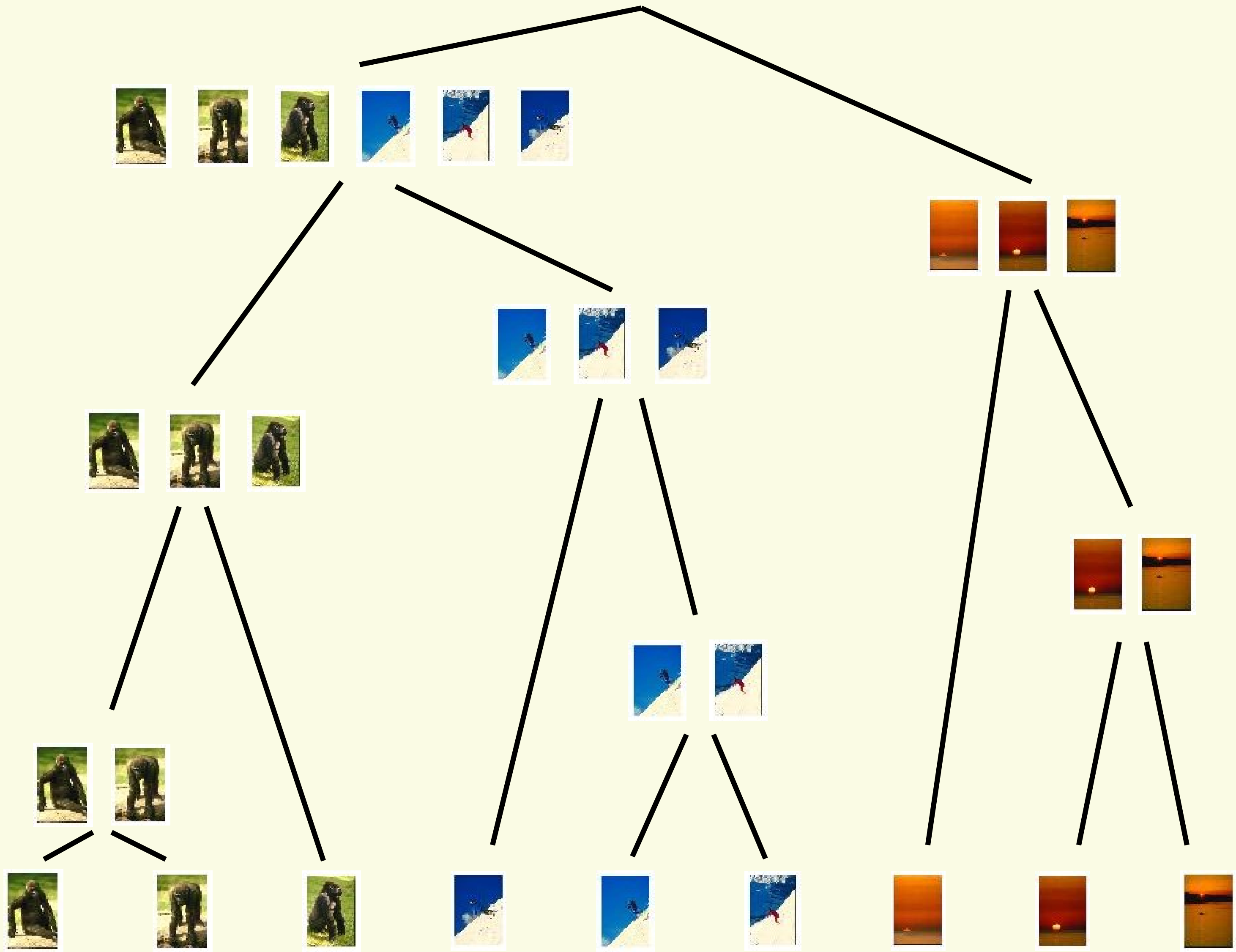
- Image segmentation
  - Find interesting “objects” in images to focus attention at


From: Image Segmentation by Nested Cuts, O. Veksler, CVPR2000

Applications of Clustering

- Image Database Organization
  - for efficient search



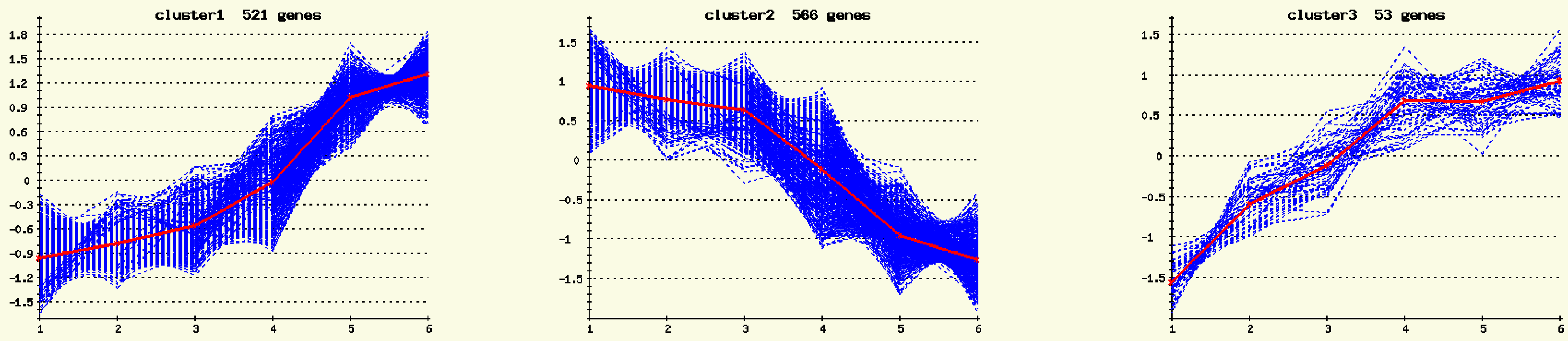
Applications of Clustering

- Data Mining
  - Technology watch
    - Derwent Database, contains all patents filed in the last 10 years worldwide
    - Searching by keywords leads to thousands of documents
    - Find clusters in the database and find if there are any emerging technologies and what competition is up to
  - Marketing
    - Customer database
    - Find clusters of customers and tailor marketing schemes to them

Applications of Clustering

- gene expression profile clustering
    - similar expressions , expect similar function

U18675 4CL -0.151 -0.207 0.126 0.359 0.208 0.091 -0.083 -0.209
M84697 a-TUB 0.188 0.030 0.111 0.094 -0.009 -0.173 -0.119 -0.136
M95595 ACC2 0.000 0.041 0.000 0.000 0.000 0.000 0.000 0.000
X66719 ACO1 0.058 0.155 0.082 0.284 0.240 0.065 -0.159 -0.010
U41998 ACT 0.096 -0.019 0.070 0.137 0.089 0.038 0.096 -0.070
AF057044 ACX1 0.268 0.403 0.679 0.785 0.565 0.260 0.203 0.252
AF057043 ACX2 0.415 0.000 -0.053 0.114 0.296 0.242 0.090 0.230
U40856 AIG1 0.096 -0.106 -0.027 -0.026 -0.005 -0.052 0.054 0.006
U40857 AIG2 0.311 0.140 0.257 0.261 0.158 0.056 -0.049 0.058
AF123253 AIM1 -0.040 0.002 -0.202 -0.040 0.077 0.081 0.088 0.224
X92510 AOS 0.473 0.560 0.914 0.625 0.375 0.387 0.019 0.141



From:De Smet F., Mathys J., Marchal K., Thijs G., De Moor B. & Moreau Y. 2002.
Adaptive Quality-based clustering of gene expression profiles, Bioinformatics, **18**(6), 735-746.

Applications of Clustering

- Profiling Web Users
  - Use web access logs to generate a feature vector for each user
  - Cluster users based on their feature vectors
  - Identify common goals for users
    - Shopping
    - Job Seekers
    - Product Seekers
    - Tutorials Seekers
  - Can use clustering results to improving web content and design

Summary

- Clustering (nonparametric unsupervised learning) is useful for discovering inherent structure in data
- Clustering is immensely useful in different fields
- Clustering comes naturally to humans (in up to 3 dimensions), but not so to computers
- It is very easy to design a clustering algorithm, but it is very hard to say if it does anything good
- General purpose clustering does not exist, for best results, clustering should be tuned to application at hand