Clustering

Matthew J. Graham
CACR

Methods of Computational Science
Caltech, 2009 February 19

matthew graham

- identify groups of 'like' objects:
  - -- to define samples with common features
  - -- to identify outliers
  - -- to partition parameter space
- associate object similarity with:
  - -- proximity in parameter space
  - -- how objects are/can be described

types of clustering

- hierarchical
  - -- agglomerative (bottom-up)
  - -- divisive (top-down)
- partitional
- density-based
- biclustering

distance measures

- Euclidean:
  $D_e(p,q) = \sqrt{\Sigma_{i=1}^{n}(p_i - q_i)^2}$
- Manhattan or taxicab:
  $D_t(p,q) = \Sigma_{i=1}^{n}|p_i - q_i|$
- Mahalanobis (correlations, scale-invariant):
  $D_m(p,q) = \sqrt{(p-q)^T S^{-1}(p-q)}$
- Cosine:
  $D_c(p,q) = 1 - \frac{p \cdot q}{|p||q|}$

how many clusters

- Rule of thumb: $k \sim (n/2)^{\frac{1}{2}}$
- Percentage of variance explained as function of number of clusters - elbow criterion
- Cluster validity, e.g. Davies-Bouldin index
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

k-means

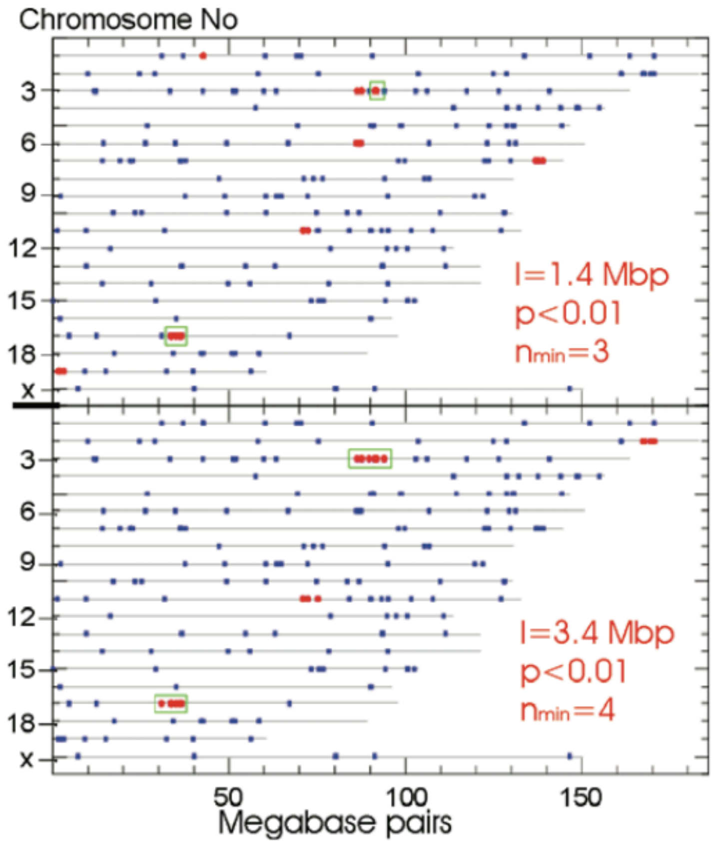
- Choose number of clusters k
- Randomly generate k clusters and determine cluster centers
- Assign each point to the nearest cluster center
- Recompute new cluster centers
- Repeat until convergence criterion is met

friends-of-friends

- Link all pairs of points separated by less than some specified distance
- Each distinct subset of connected points is a group
- At some critical distance, groups percolate: any side of the set of points can be reached from any point (perfect connectivity)
- Danger of bridging or snaking

friends-of-friends example

Up- or down-regulated genes in the mouse genome



linkage

single: $min\{d(x,y) : x \in A, y \in B\}$

-- Results easily in snake-like clusters even if they don't exist

complete: $max\{d(x,y) : x \in A, y \in B\}$

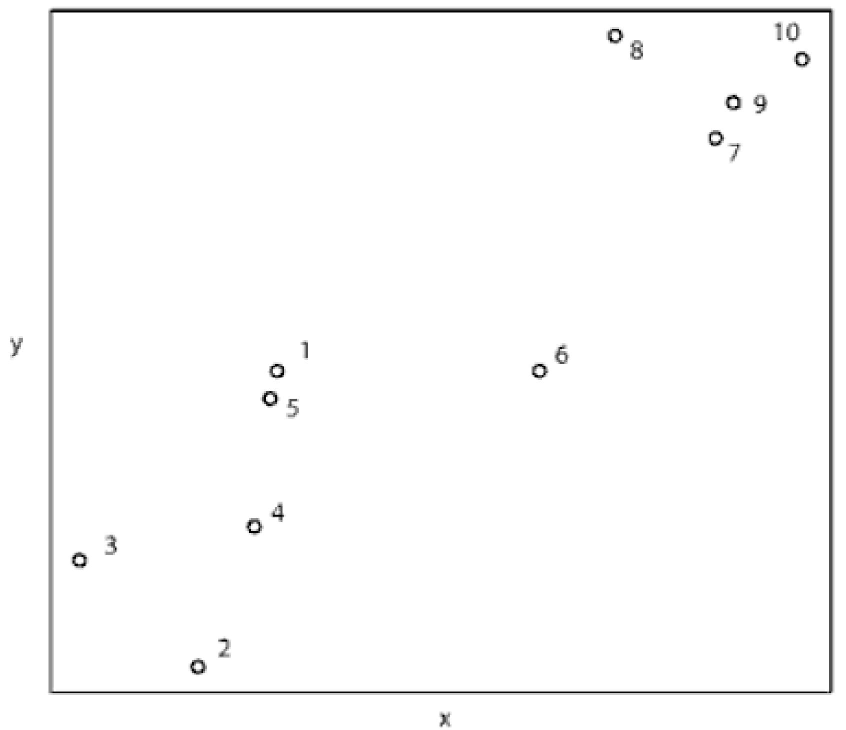
-- Eliminates the snake formation but sometimes produces puzzling configurations between tight and loosely formed clusters.

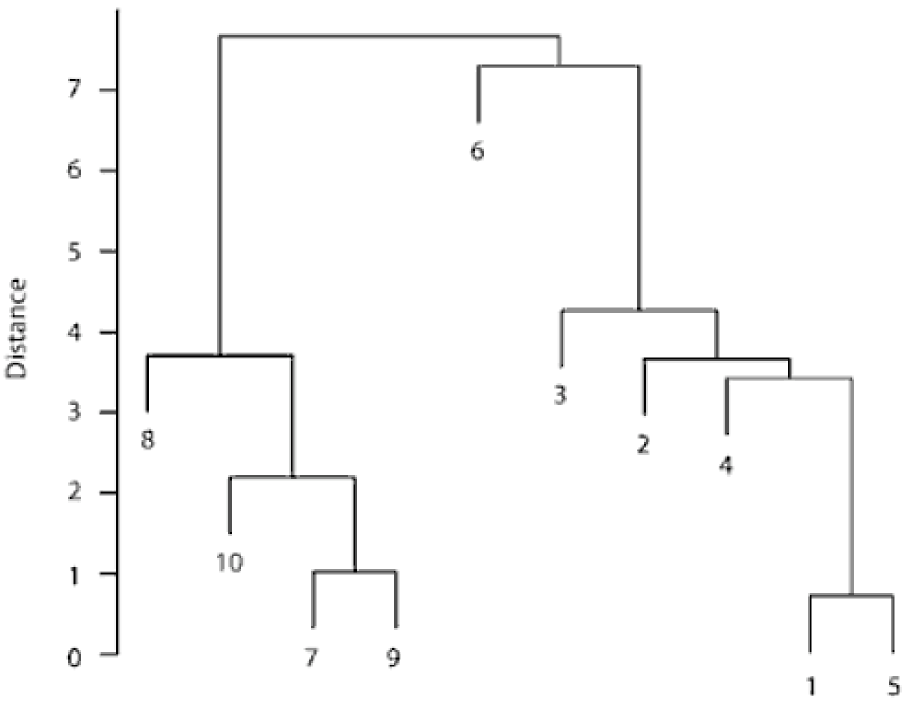
average: $\frac{1}{|A| \cdot |B|} \Sigma_{x \in A} \Sigma_{y \in B} d(x,y)$

-- Joins clusters with smallest average distances
-- Not as outlier sensitive
-- Tends to form clusters with small within-cluster variation
-- Biased to form clusters with approximately the same variance

linkage example

Distribution of archaeological Bronze Age pottery finds





minimal spanning tree

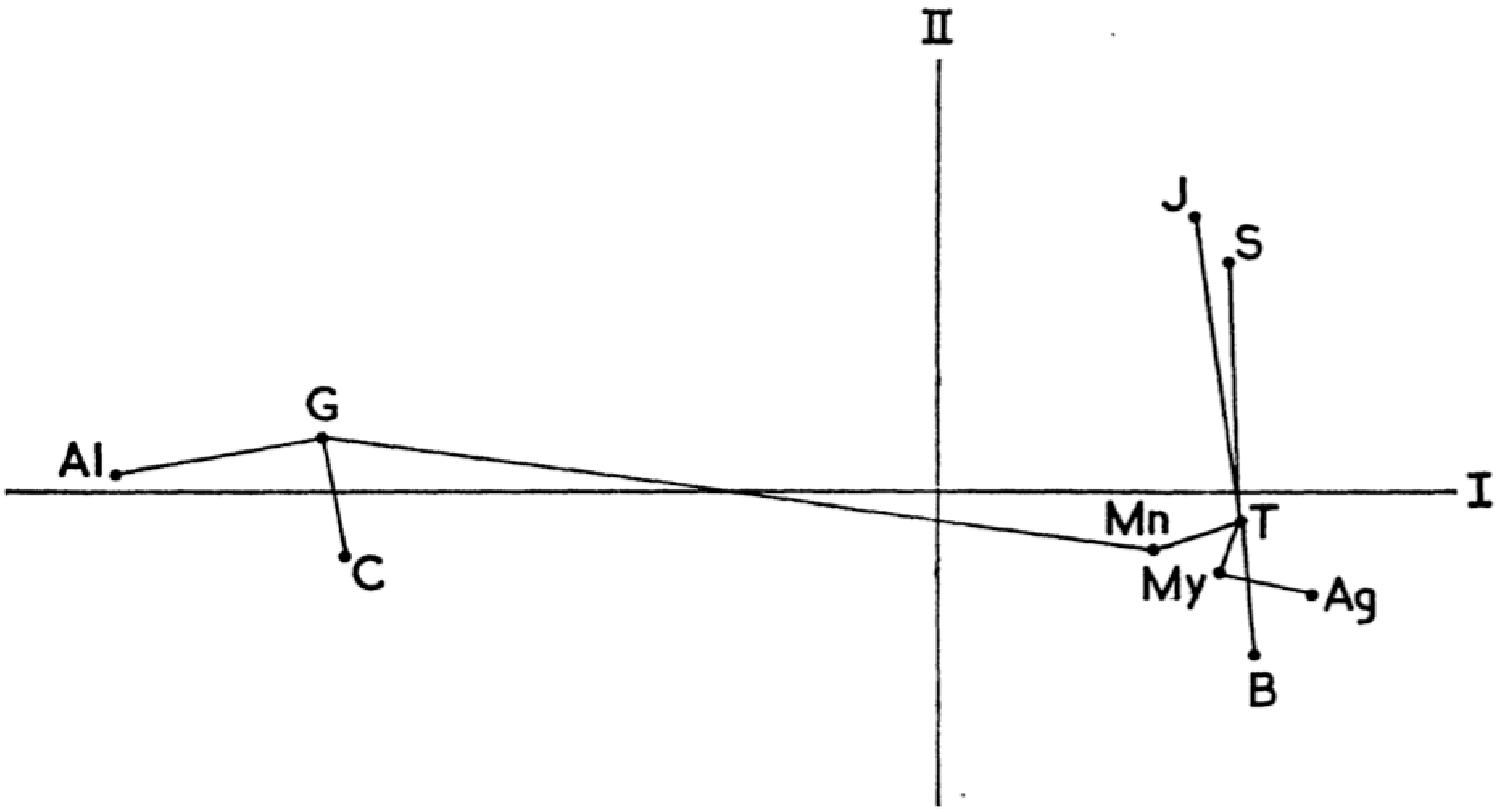
- Consider a set of straight line segments (edges) joining pairs of points such that:

-- no closed loops occur
-- each point is visited by at least one line
-- there is a sequence of edges between any pair of points (connected)
-- the sum of the edge lengths is minimised

- If no edge-lengths are equal then the MST is unique

minimal spanning tree example

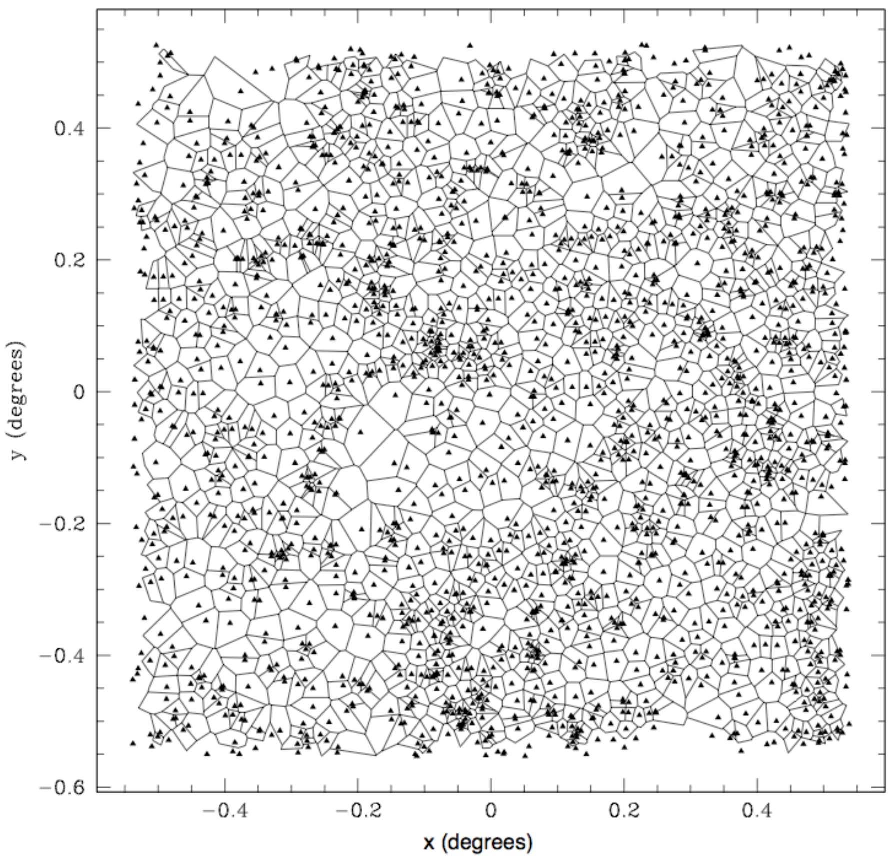
Canonical variate means of skull measurements of white-toothed shrews

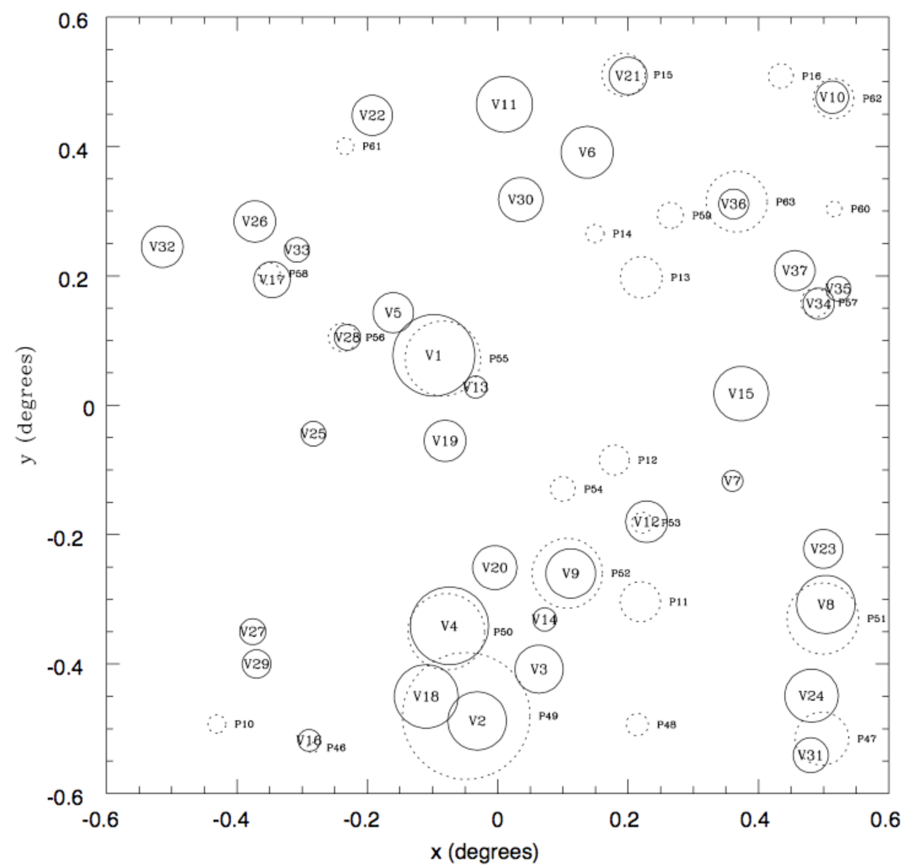


- Consider S as a set of points in a space
- For (almost) any point x in the space, there is one point of S closest to x
- The set of all points closer to a point c of S than to any other point of S is the interior of a convex polytope (**Voronoi cell**) for c
- The set of such polytopes tessellates the whole space and is the **Voronoi tessellation** for set S

voronoi example

Galaxies





kd tree

- A kd-tree is a binary tree constructed on a set of points in k-dimensional space with leaf nodes and non-leaf nodes
- Every non-leaf node generates a splitting hyperplane that divides the space into two subspaces
- Points left of the hyperplane represent left subtree of that node and points to the right the right subtree
- Hyperplanes are always perpendicular to one of k-dimension axes and are cycled through with successive non-leaf nodes

kd tree example

Consider (2,3), (5,4), (9,6), (4,7), (8,1) and (7,2):