Speeding up Clustering

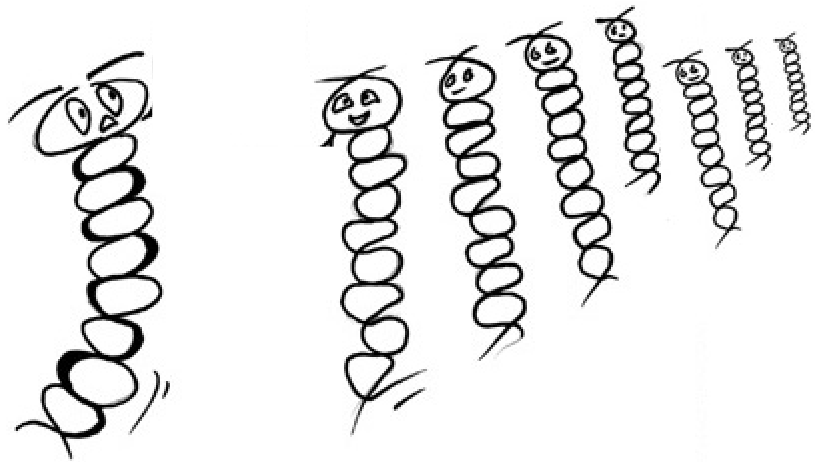
Meelis Kull

Theory Days in Veskisilla

3 Oct, 2004

Outline

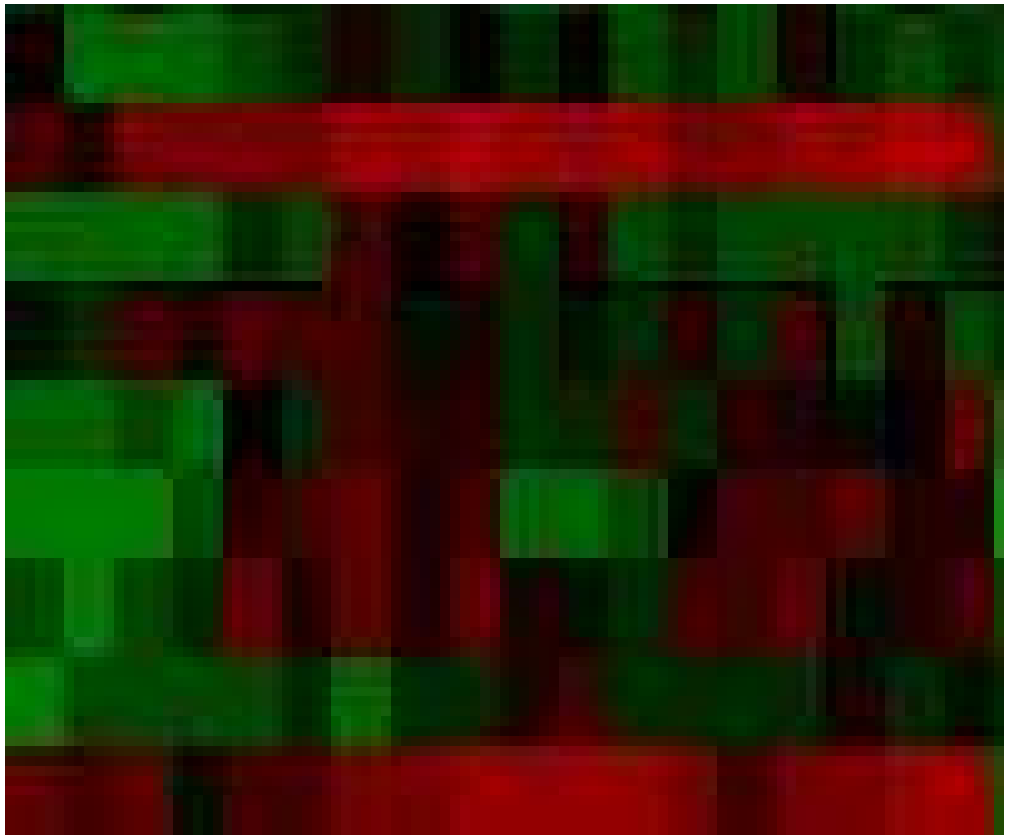
- Why?
- Hierarchical Clustering
- Approximate Hierarchical Clustering
- Finding similar pairs
- Pivots and similarity join
- EGO



Gene expression data

Samples

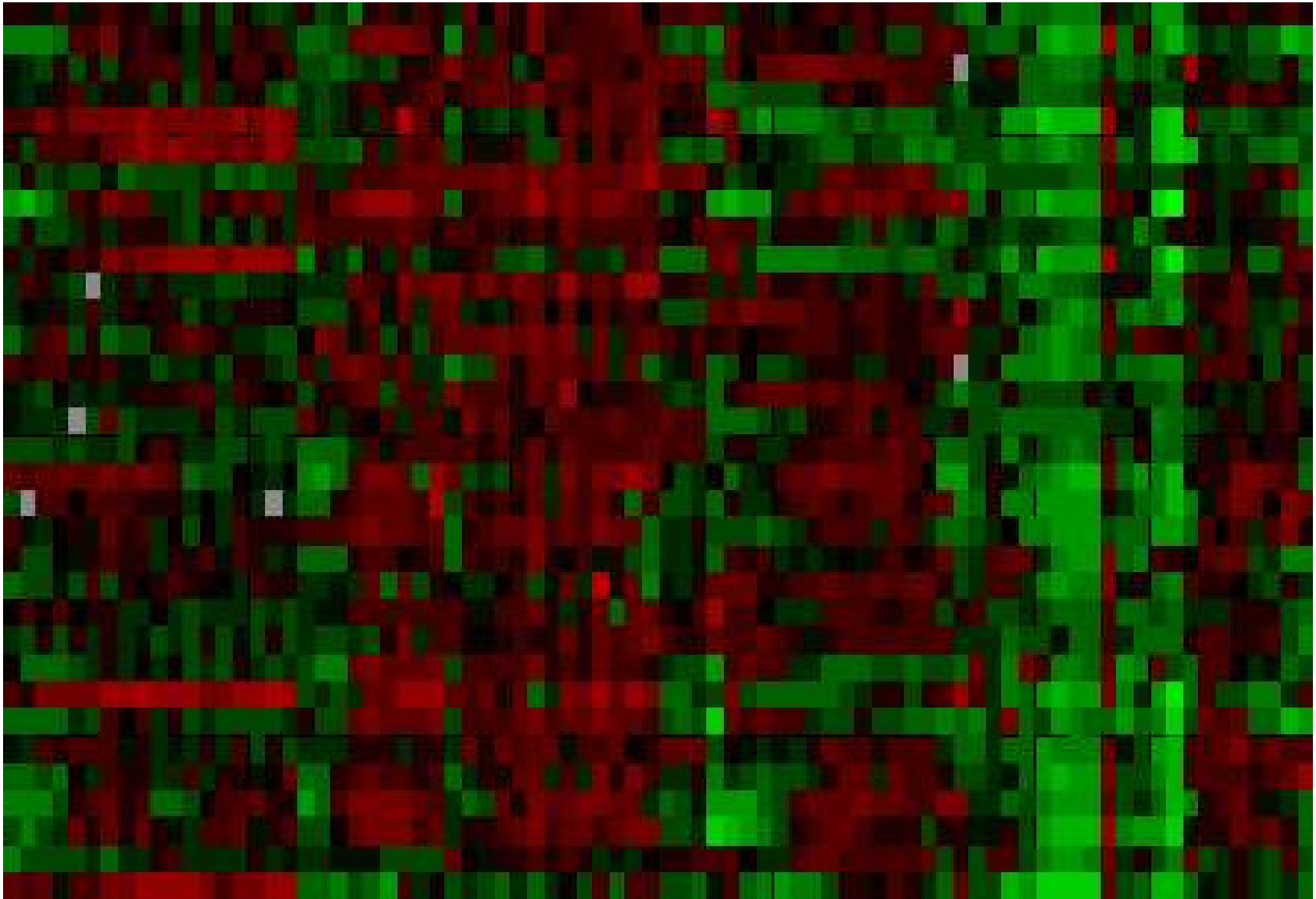
Gene 1
Gene 2
Gene 3
Gene 4
Gene 5
Gene 6
Gene 7
Gene 8
Gene 9



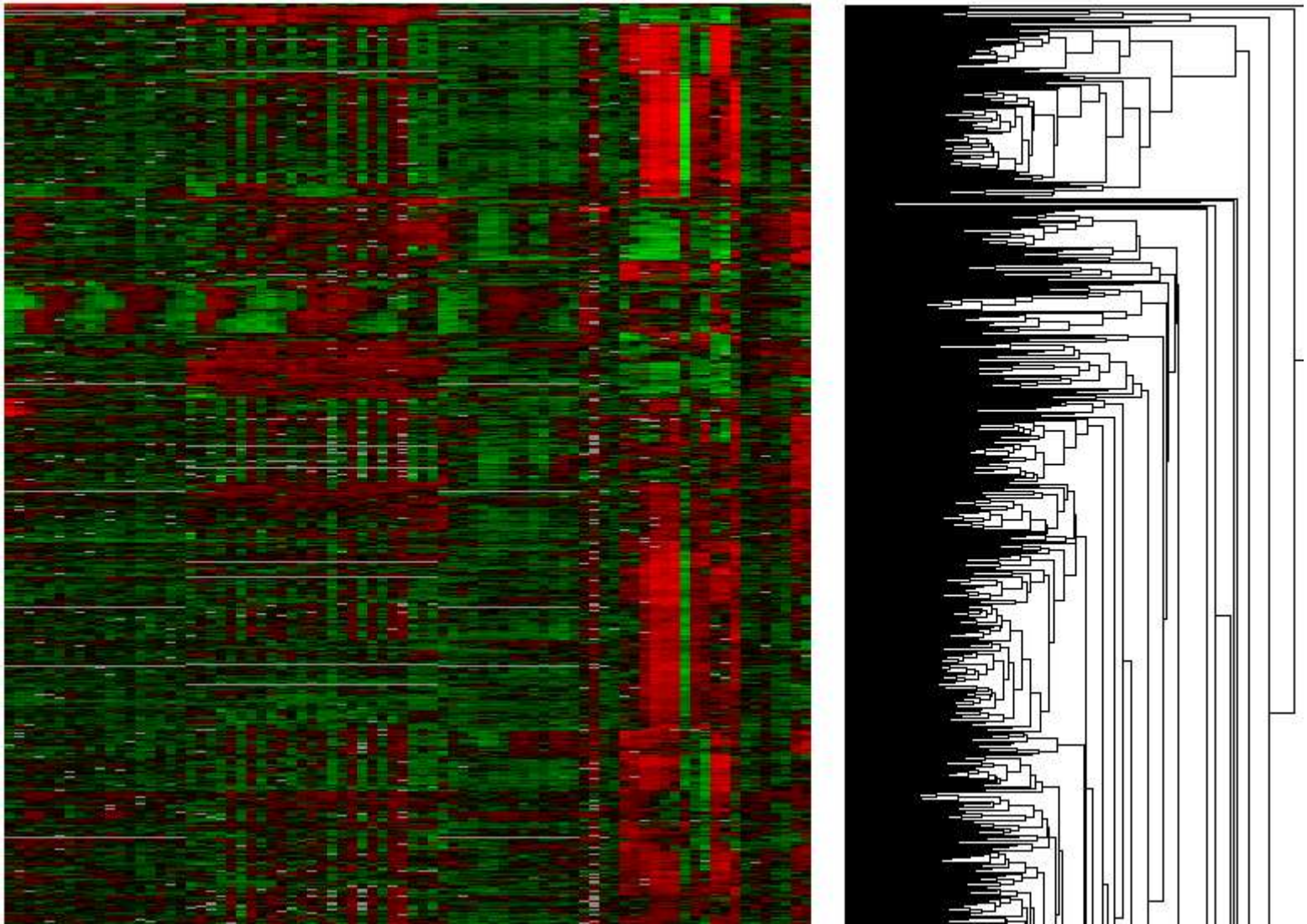
Gene is highly expressed

Gene is lowly expressed

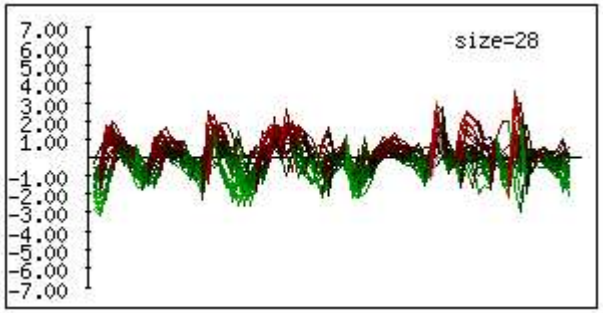
Gene expression data

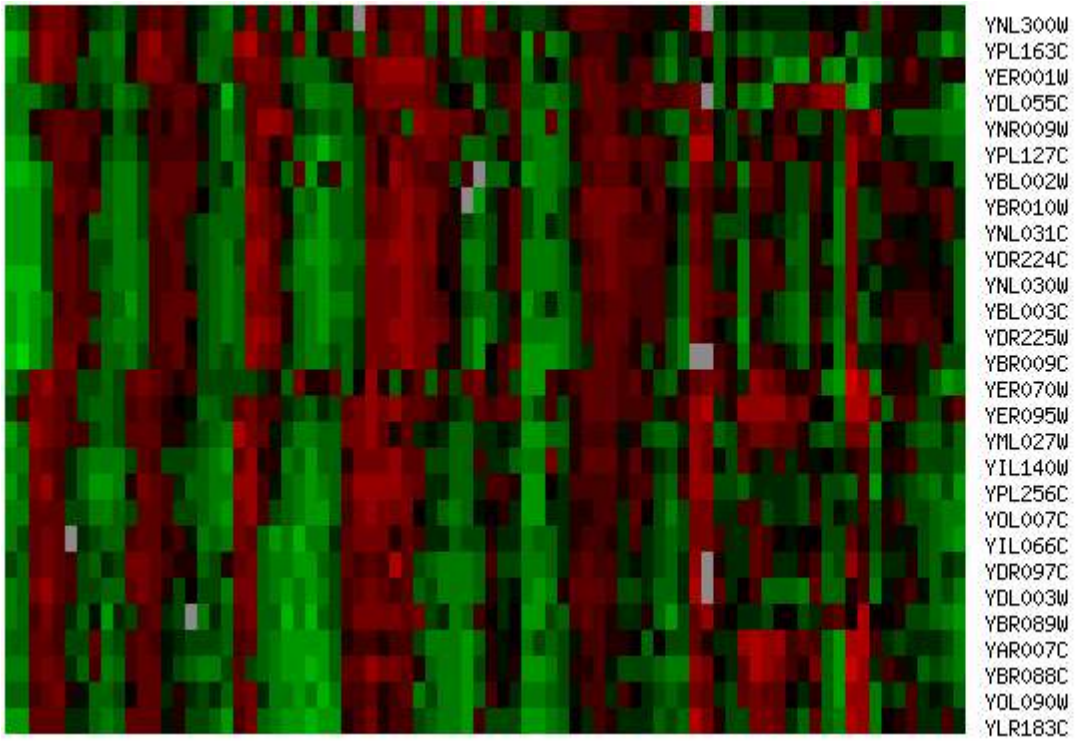


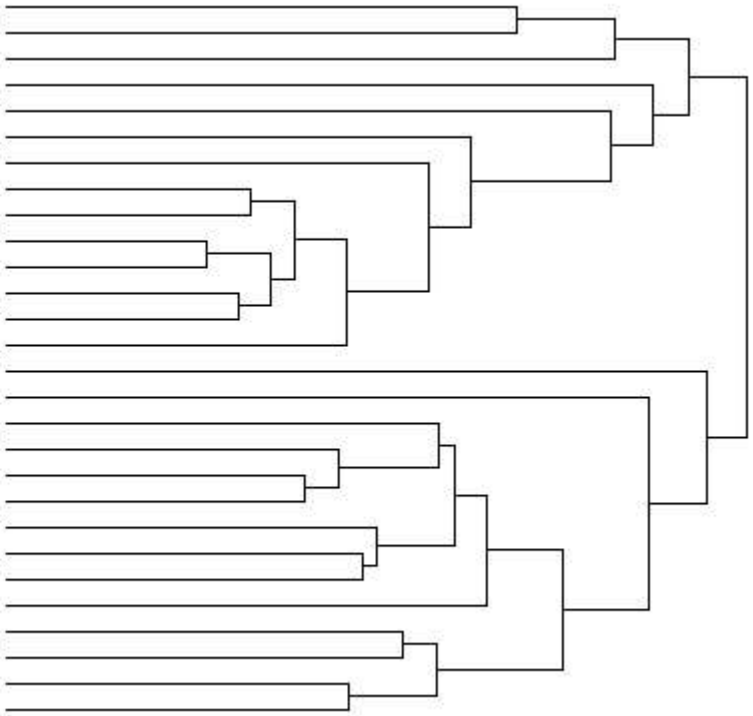
The Biologist's Dream



Dream Zoomed In







Assumptions

$x_1, \ldots, x_n$ – data objects

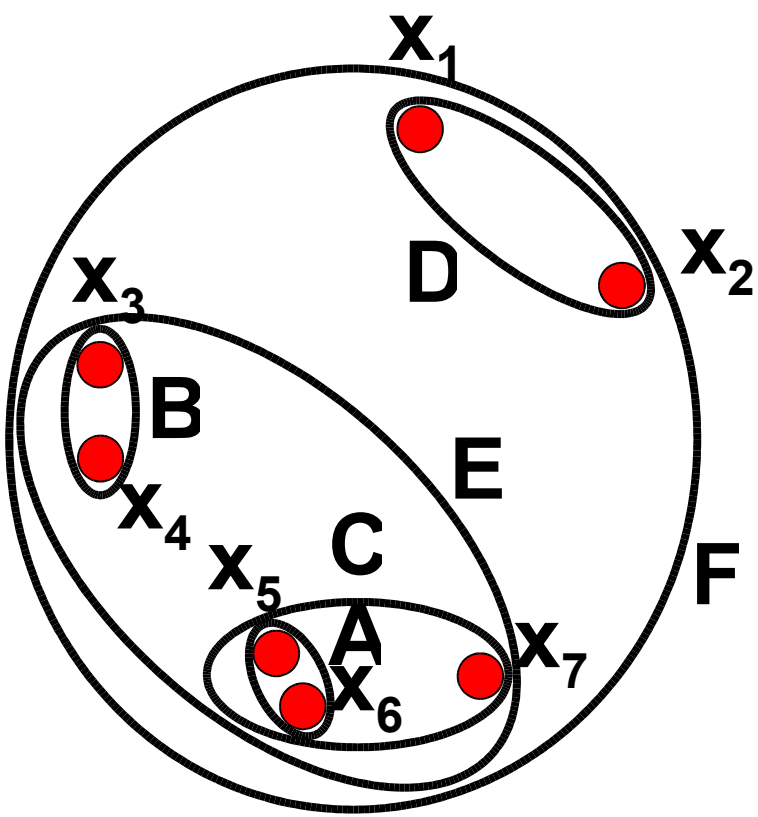
d – distance metric

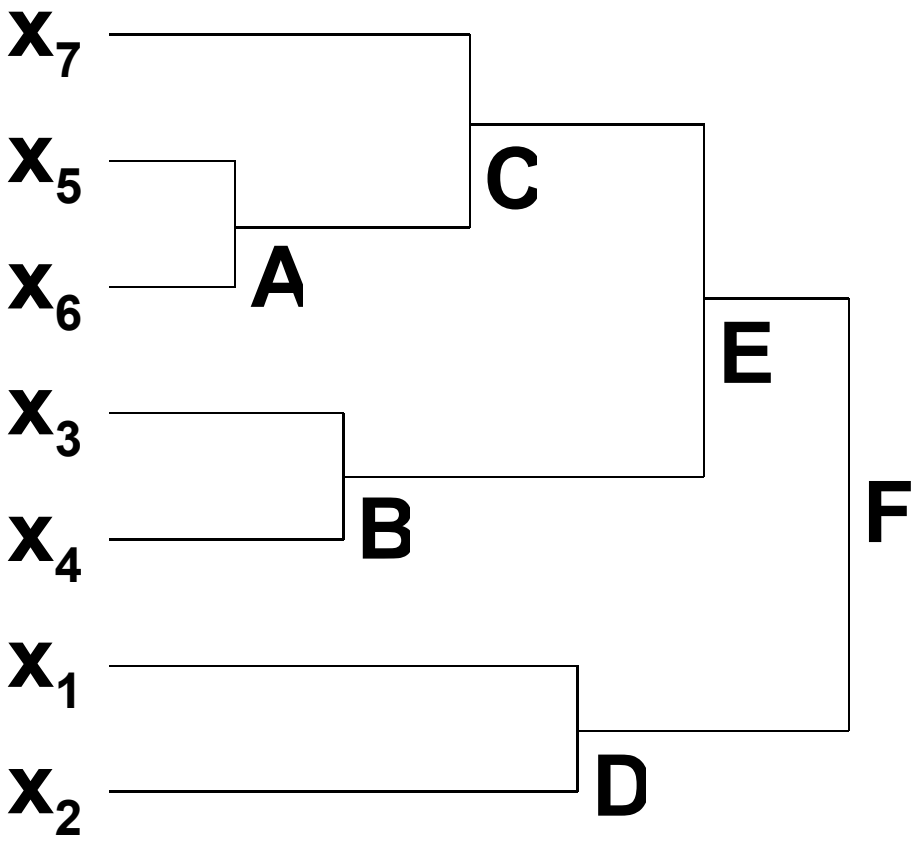
1) $d(x_i, x_j) = 0 \quad \Leftrightarrow \quad x_i = x_j$

2) $d(x_i, x_j) = d(x_j, x_i)$

3) $d(x_i, x_j) \le d(x_i, x_k) + d(x_k, x_j)$

Hierarchical Clustering



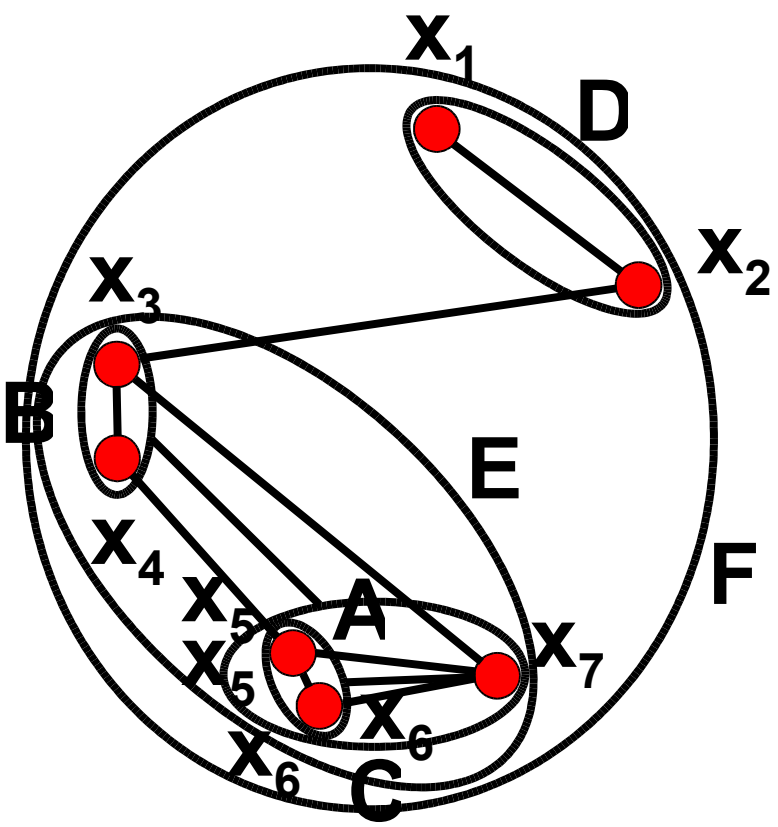


$O(n^2)$

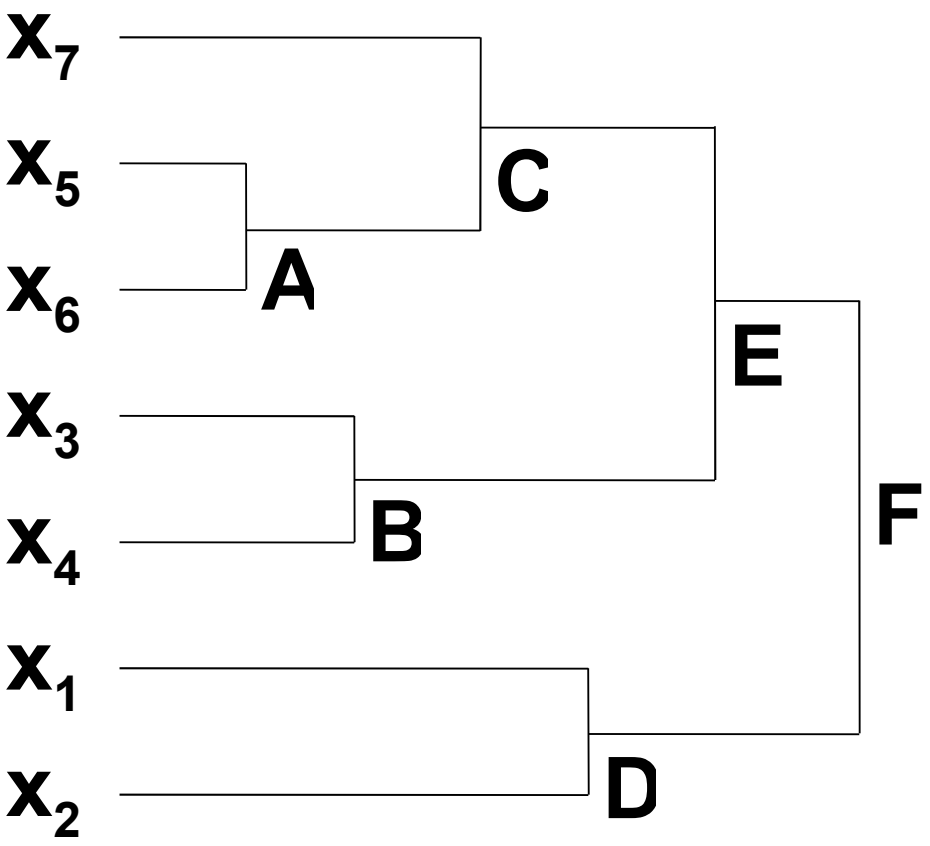
dendrogram

Approximate Hierarchical Clustering

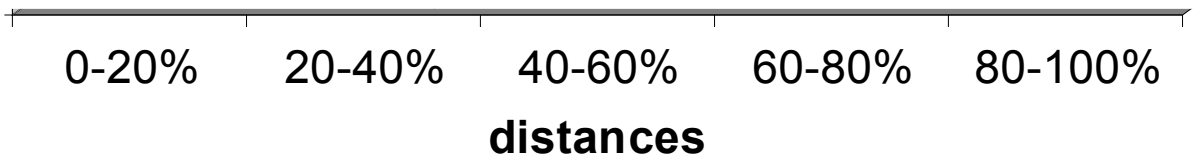
$$O(n \log n \log m + m)$$



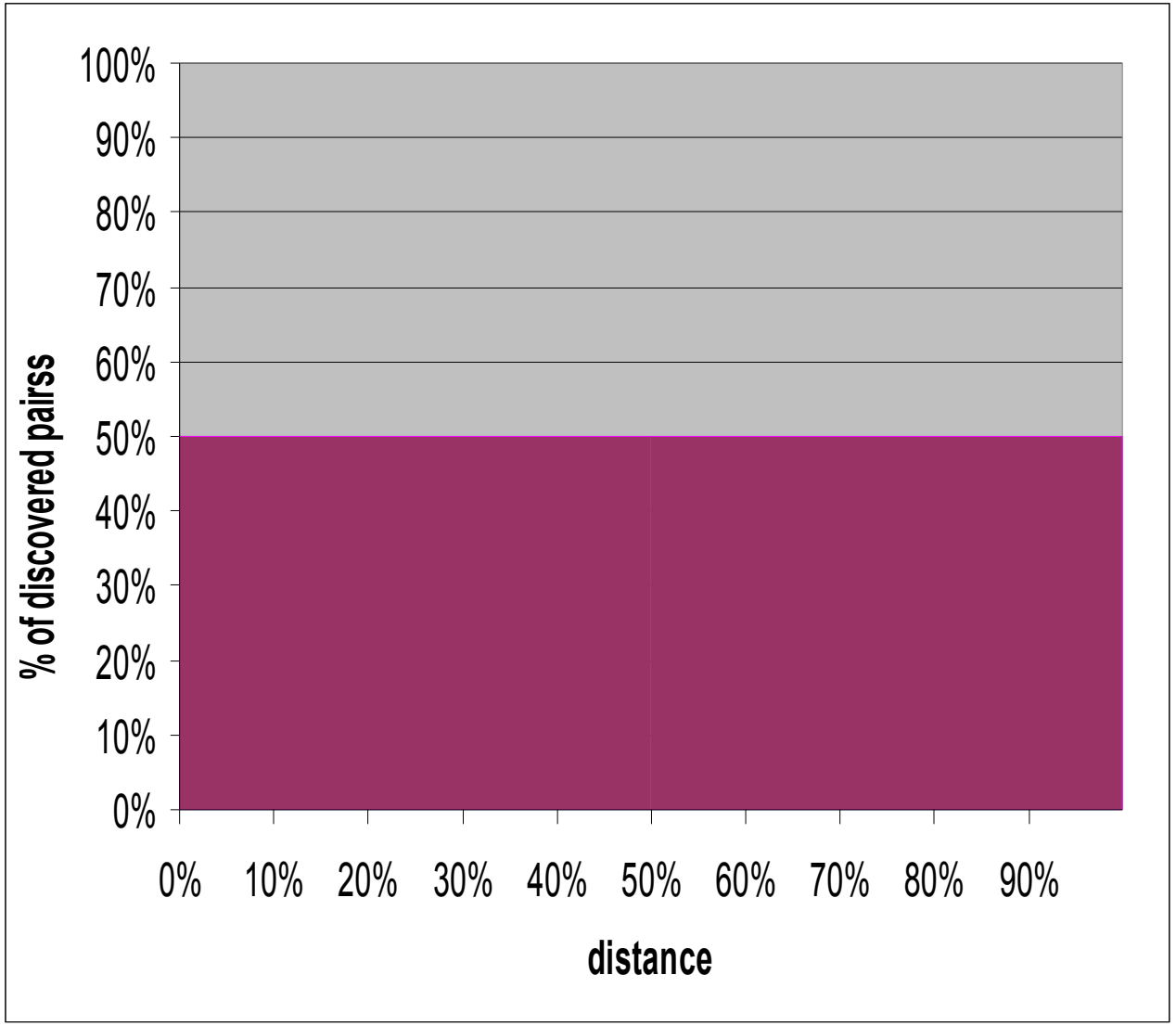
1/3 of the distances calculated



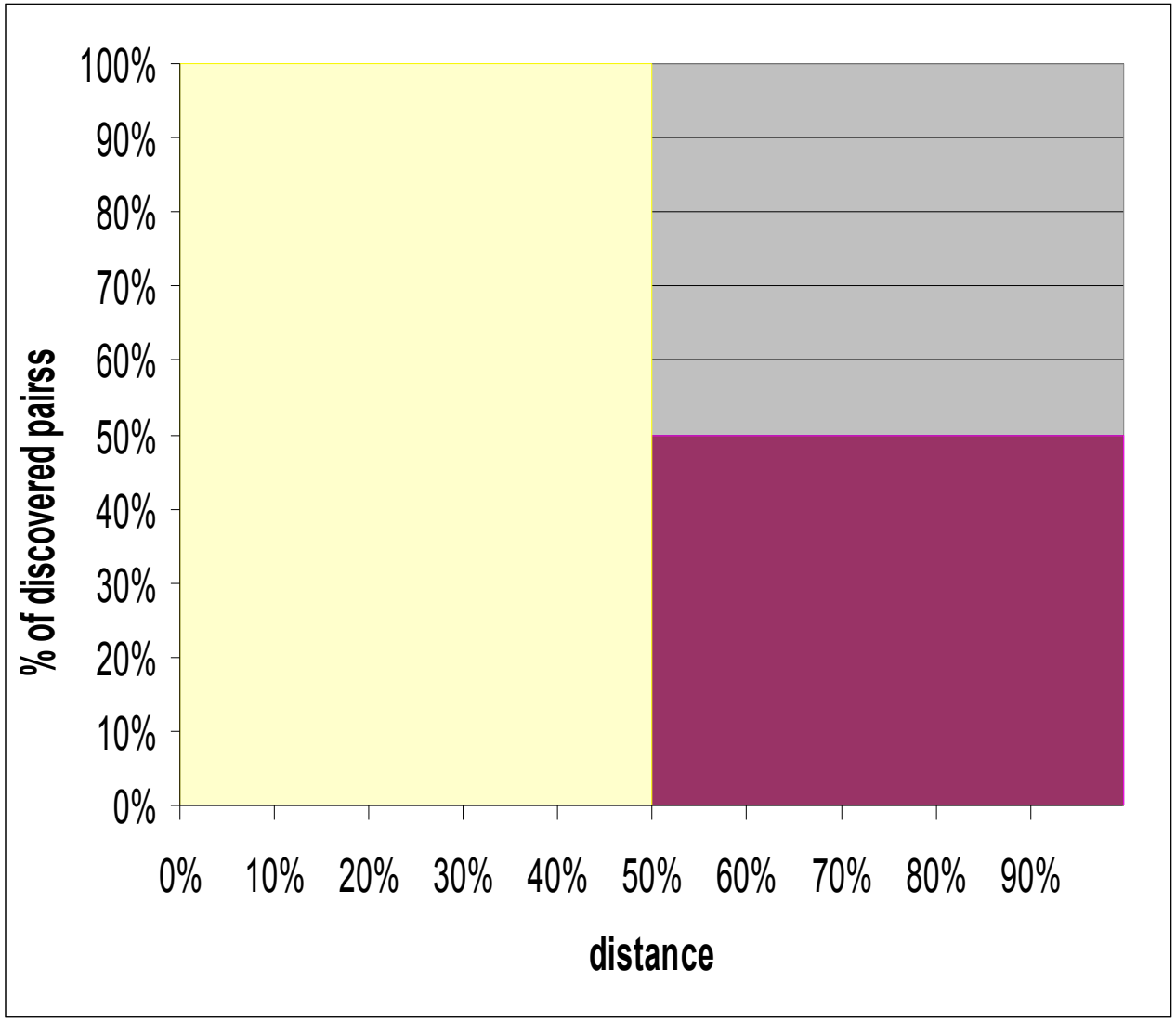
Calculating distances



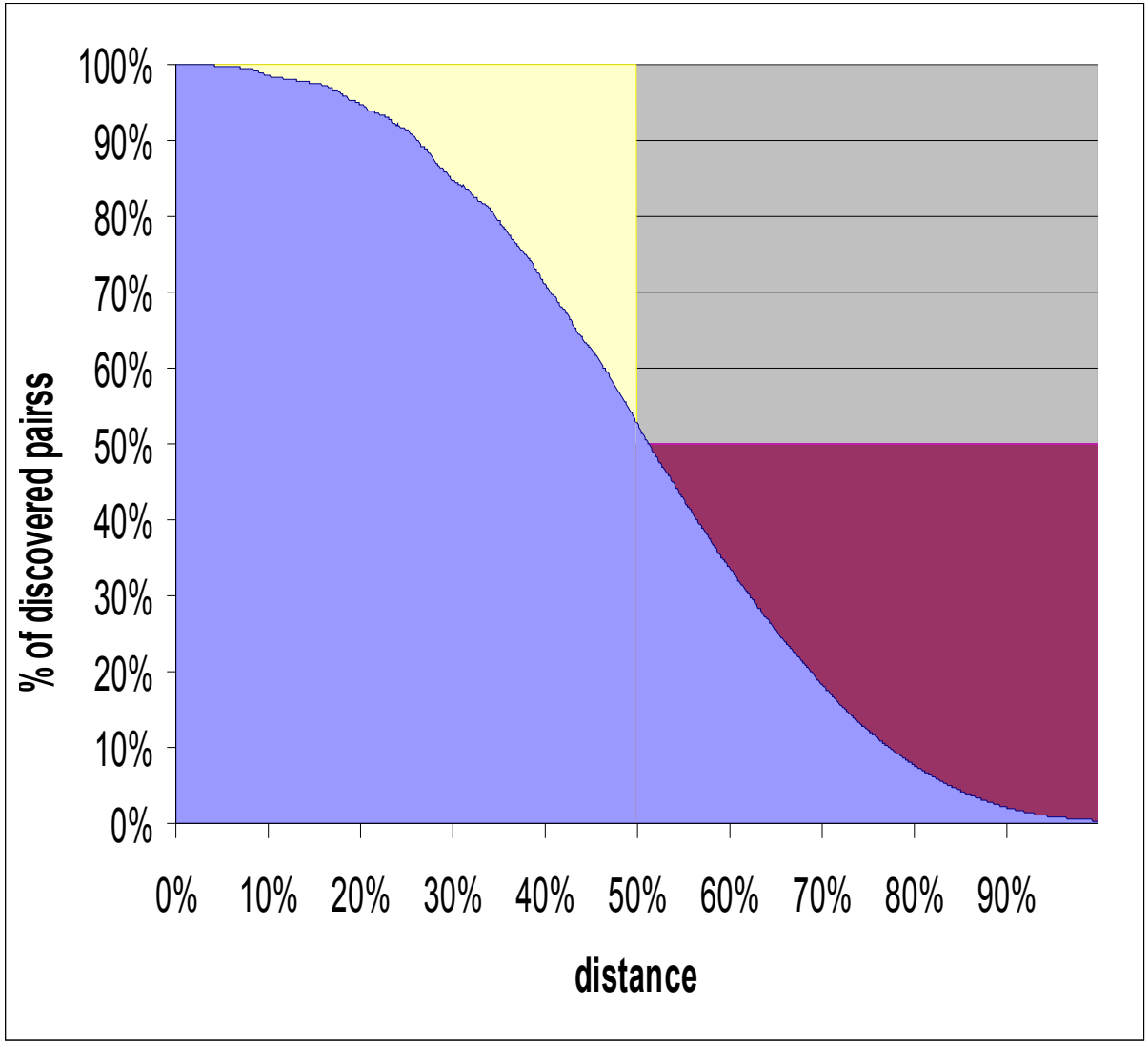
Calculating distances



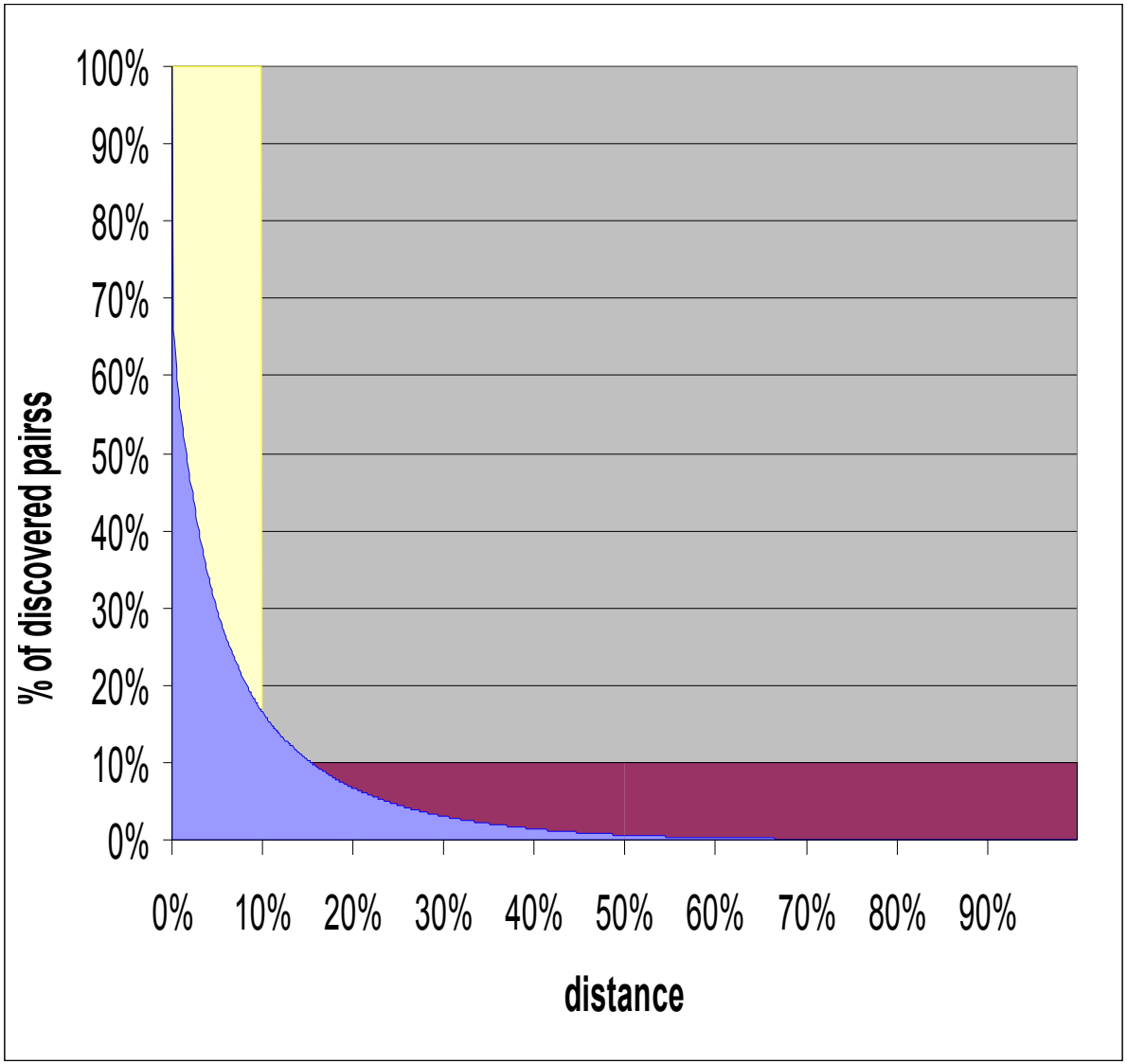
Calculating distances



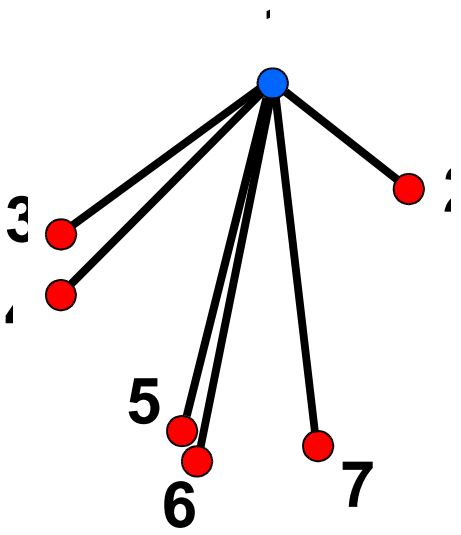
Calculating distances



Calculating distances



How to find all similar pairs without calculating all the distances?



Triangle inequality: $d(x,z) \leq d(x,y)+d(y,z)$

Corollary: $d(y,z) \geq |d(x,y)-d(x,z)|$

$d(\mathbf{5},\mathbf{2}) \geq |d(\mathbf{1},\mathbf{5})-d(\mathbf{1},\mathbf{2})|$

$d(\mathbf{5},\mathbf{6}) \geq |d(\mathbf{1},\mathbf{5})-d(\mathbf{1},\mathbf{6})|$

Finding Similar Pairs

1. Choose 1-20 pivot objects

$$p_1, \ldots, p_q$$

2. Calculate distances from x_i to p_k

$$x_i \rightarrow \left(d(x_i, p_1), d(x_i, p_2), \ldots, d(x_i, p_q) \right)$$

3. Find all pairs of objects which are at similar distances from all pivots

$$(x_i, x_j) \quad s.t. \quad \max_k \left| d(x_i, p_k) - d(x_j, p_k) \right| < \varepsilon$$

Finding Similar Pairs

We have N points in the q-dimensional space,

$$d_1 = (d_{11}, \ldots , d_{1q})$$

$$d_N = (d_{N1}, \ldots , d_{Nq})$$

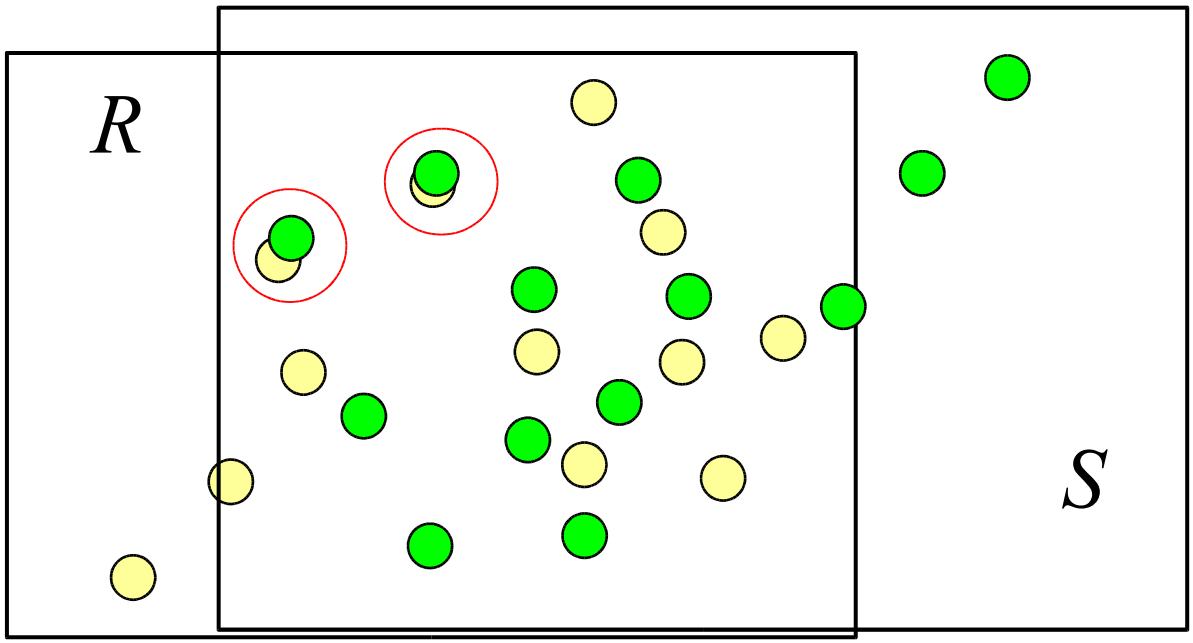
Find pairs (d_i , d_j) s.t. $L_\infty(d_i , d_j) < \varepsilon$ where

$$L_\infty(d_i, d_j) = \max_k | d_{ik} - d_{jk} |$$

EGO – Epsilon Grid Order

Find pairs (d_i, d_j) s.t. $L_\infty(d_i, d_j) < \varepsilon$ where

$$L_\infty(d_i, d_j) = \max_k | d_{ik} - d_{jk} |$$



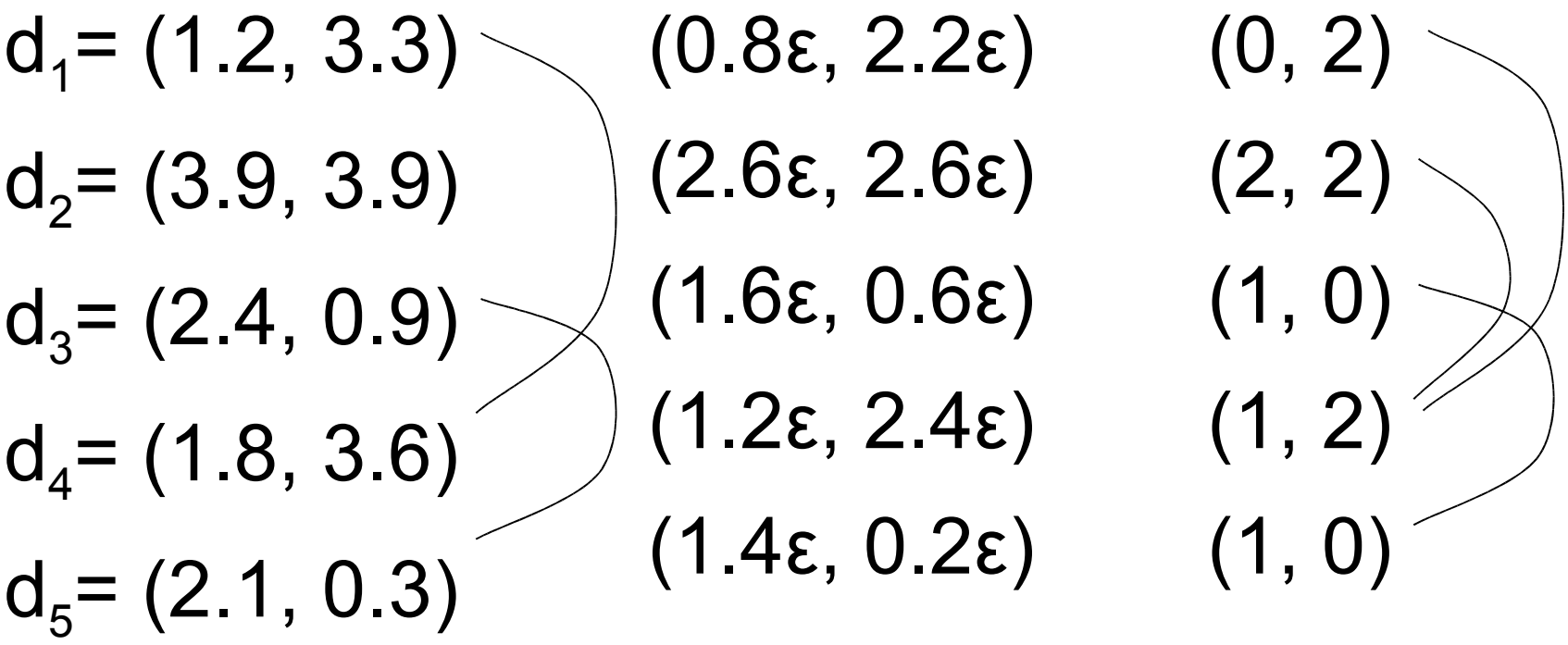
$$R \underset{\varepsilon}{\bowtie} S := \{(r_i, s_j) \in R \times S : \|r_i - s_j\| \leq \varepsilon\}$$

EGO – Epsilon Grid Order

Find pairs (d_i, d_j) s.t. $L_\infty(d_i, d_j) < \varepsilon$ where

$$L_\infty(d_i, d_j) = \max_k | d_{ik} - d_{jk} |$$

$\varepsilon = 1.5$

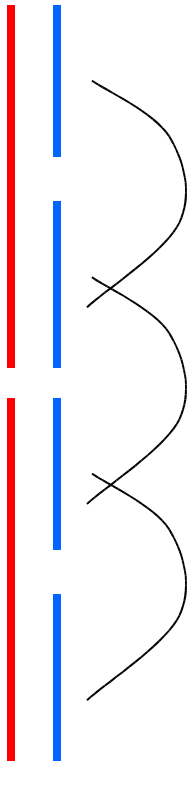


EGO – Epsilon Grid Order

Find pairs (d_i, d_j) s.t. $L_\infty(d_i, d_j) < \varepsilon$ where

$| d_{ik} - d_{jk} | \in \{-1,0,1\}$ for each k

(0, 0, 1, 2)
(0, 2, 3, 2)
(1, 2, 1, 2)
(1, 3, 1, 0)
(2, 0, 0, 0)
(2, 2, 2, 1)
(3, 0, 2, 1)
(3, 1, 0, 2)



join(X,Y) =

join(X.tophalf, Y.tophalf) U
join(X.tophalf, Y.bottomhalf) U
join(X.bottomhalf, Y.tophalf) U
join(X.bottomhalf, Y.bottomhalf)

Summary

- Similar pairs can be found quite fast
- This is useful for speeding up clustering