Topics in Data Engineering

Session 4

Masaomi Kimura

Clustering

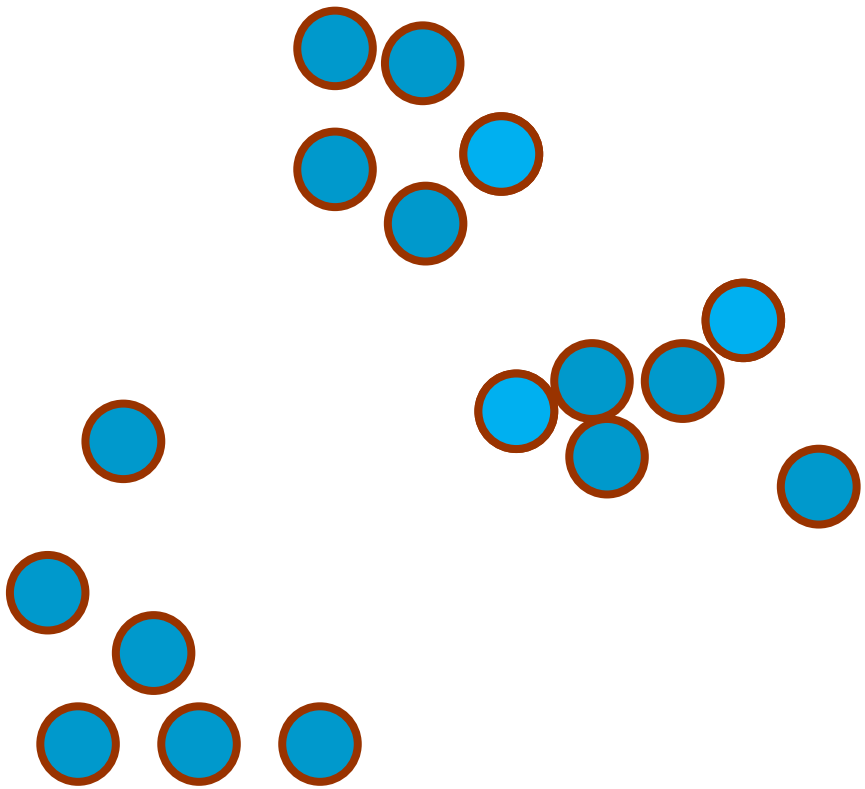
Major types of clustering method

- Cut-based clustering
  - given data are divided into clusters (groups of similar data)
- hierarchical clustering
  - a dendrogram is generated to show the structure of similarity between data

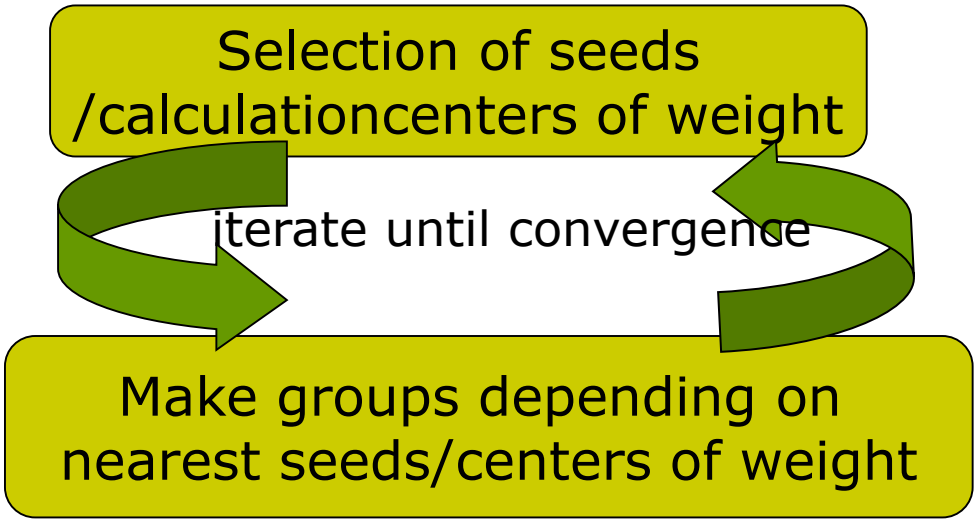
K-means

- developed by J.B.MacQueen in 1967
- a representative cut-based method
- necessary to specify the number of clusters(K)
- target data are vectors whose elements are numbers

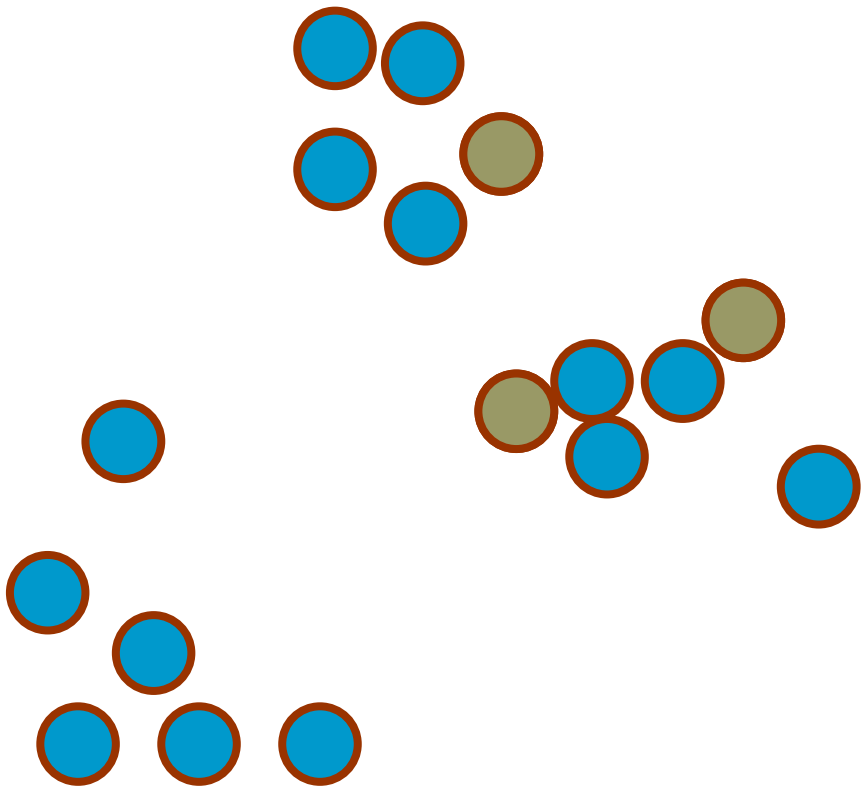
K-means



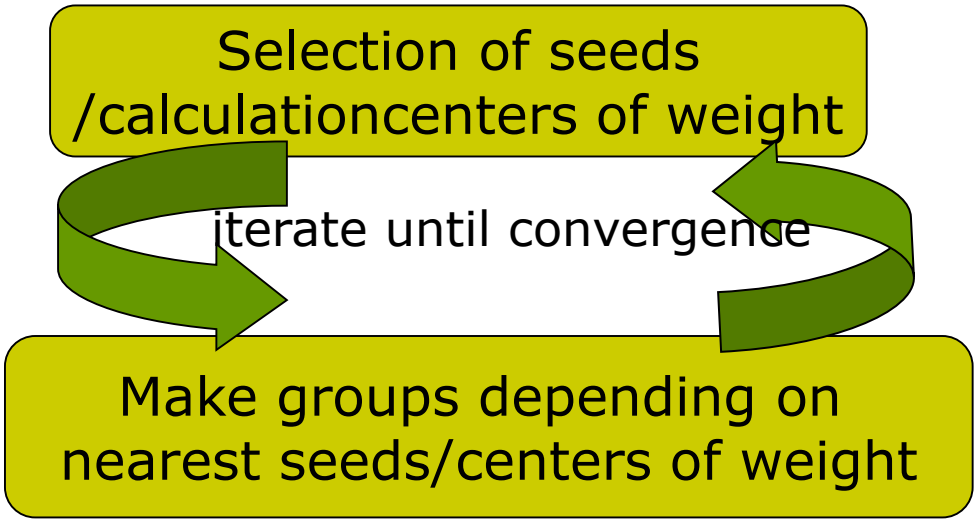
K=the number of clusters



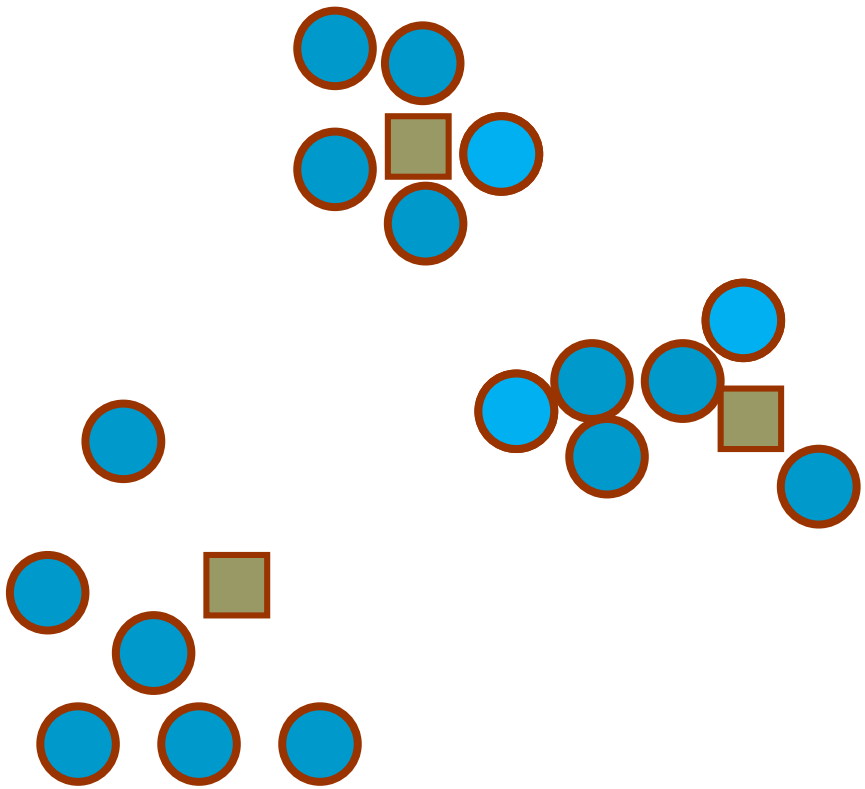
K-means



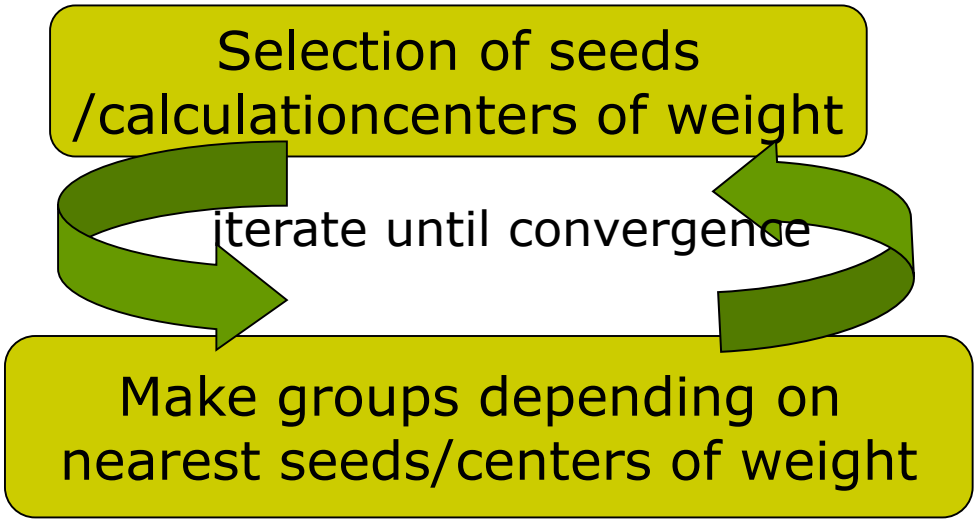
K=the number of clusters



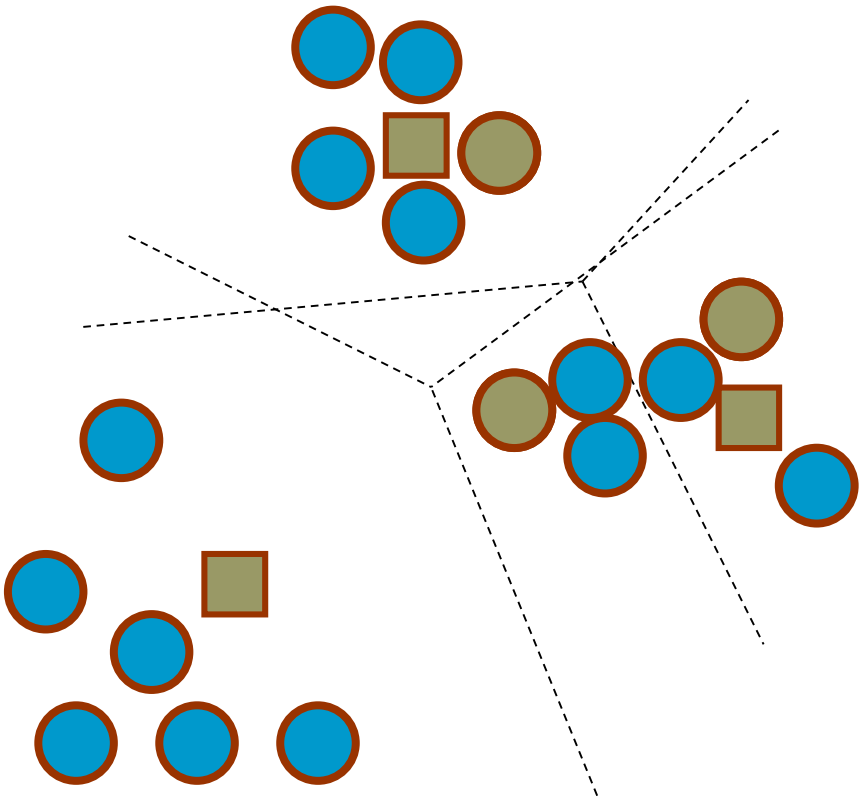
K-means



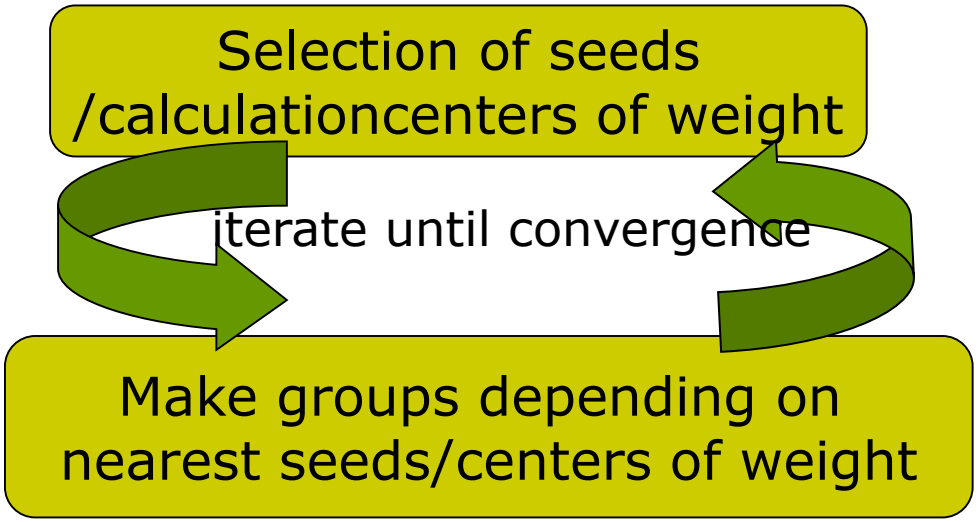
K=the number of clusters



K-means



K=the number of clusters



A procedure of K-means

1. Select K seeds from given data
2. Make groups whose data are grouped depending on which seed is nearest
3. Calculate a center of weight of data in each group

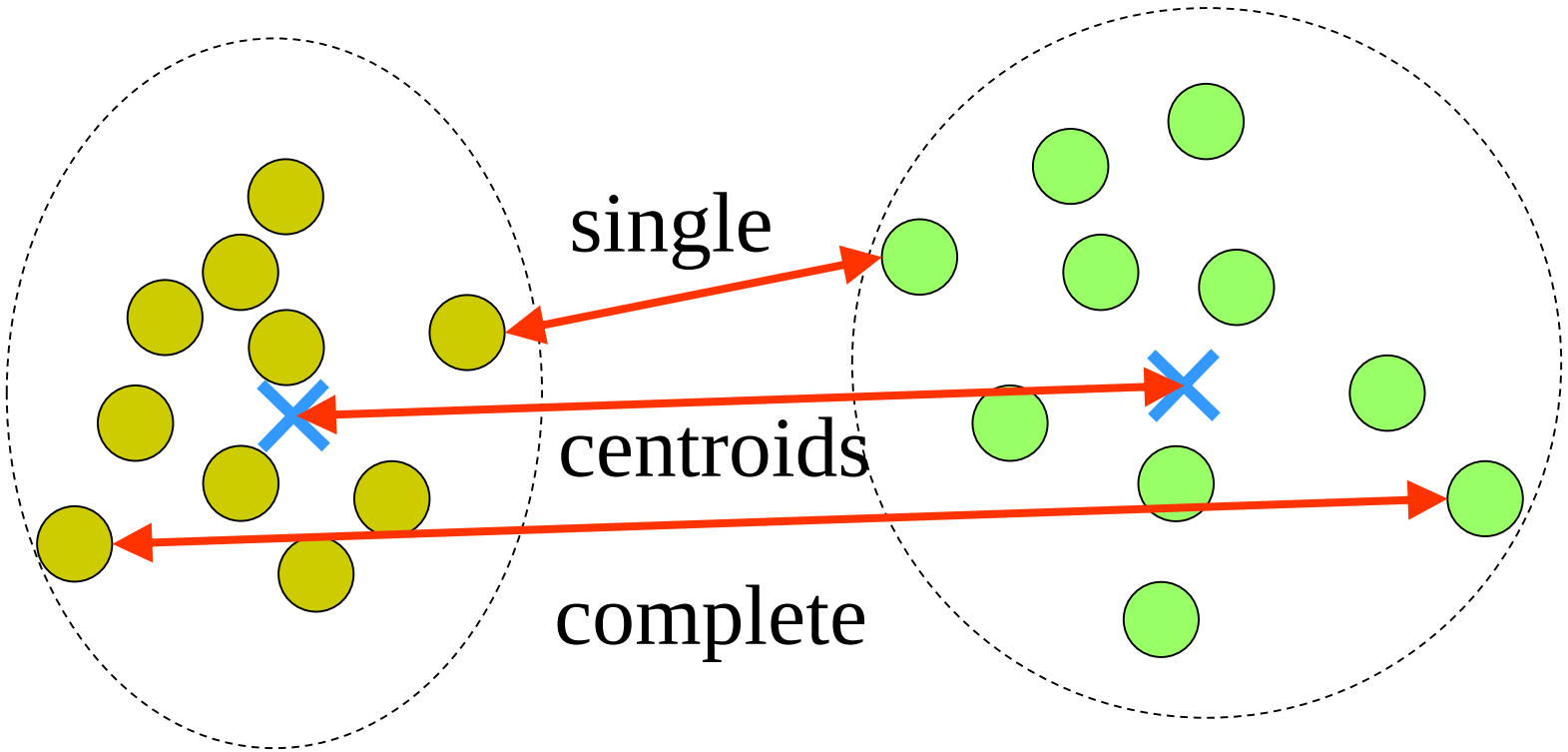
$$\vec{x}_g = \frac{1}{N_c} \sum_{i \in C} \vec{x}_i$$

4. Instead of seeds in Step 2, use the centers of weight
5. Iterate Step 2, Step 3 and Step 4, until the centers of weight converge.

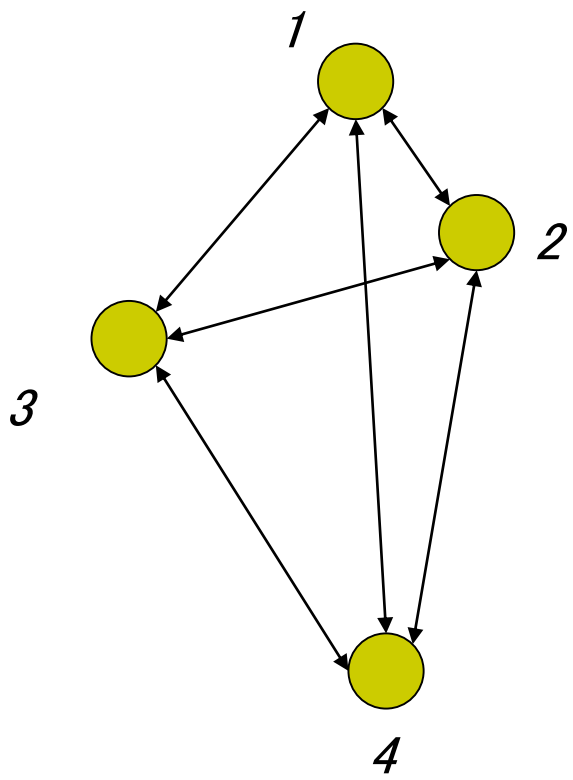
Hierarchical clustering

- Merge data in the similar/near order
  - Use a distance function to measure distances between data
  - Need to define distance between clusters
    - single linkage
    - complete linkage
    - comparison of centroids
  - Output is a graph, dendrogram, to visualize the steps of the merges.

Distances between clusters



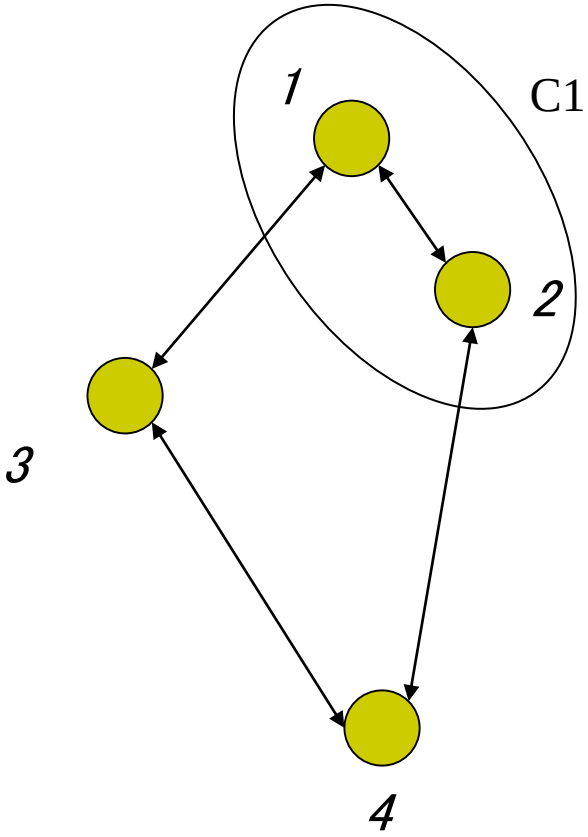
Step1 (single linkage)



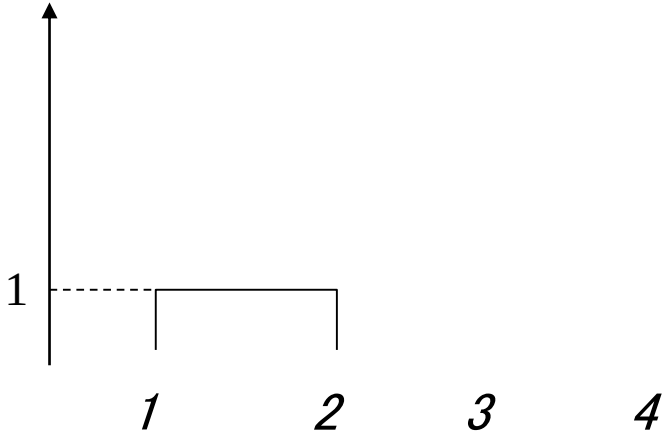
	1	2	3	4
1	0			
2	1	0		
3	2	3	0	
4	6	5	4	0

1 2 3 4

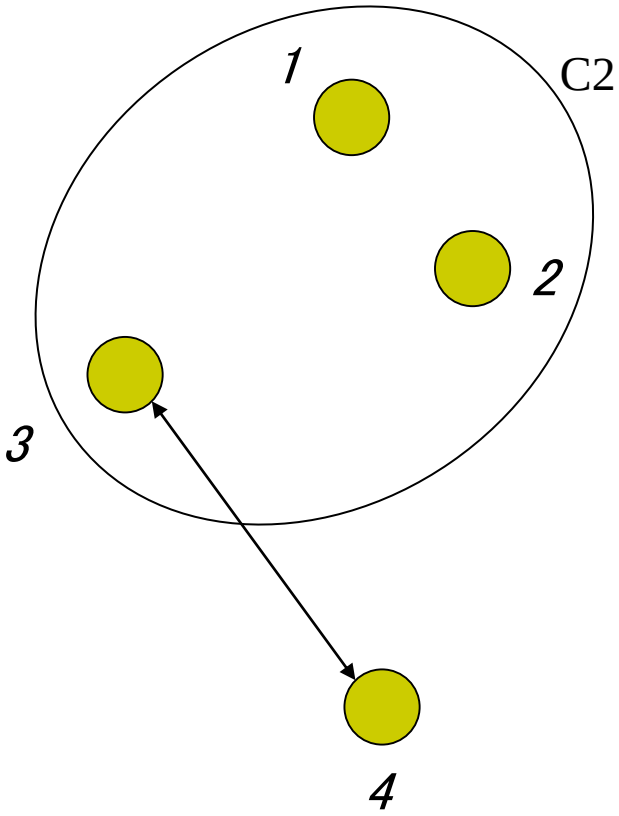
Step2 (single linkage)



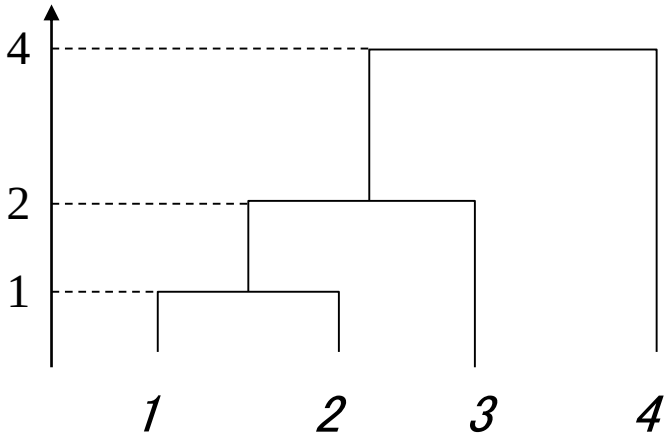
	C1	3	4
C1	0		
3	2	0	
4	5	4	0



Step3 (single linkage)



	C2	4
C2	0	
4	4	0

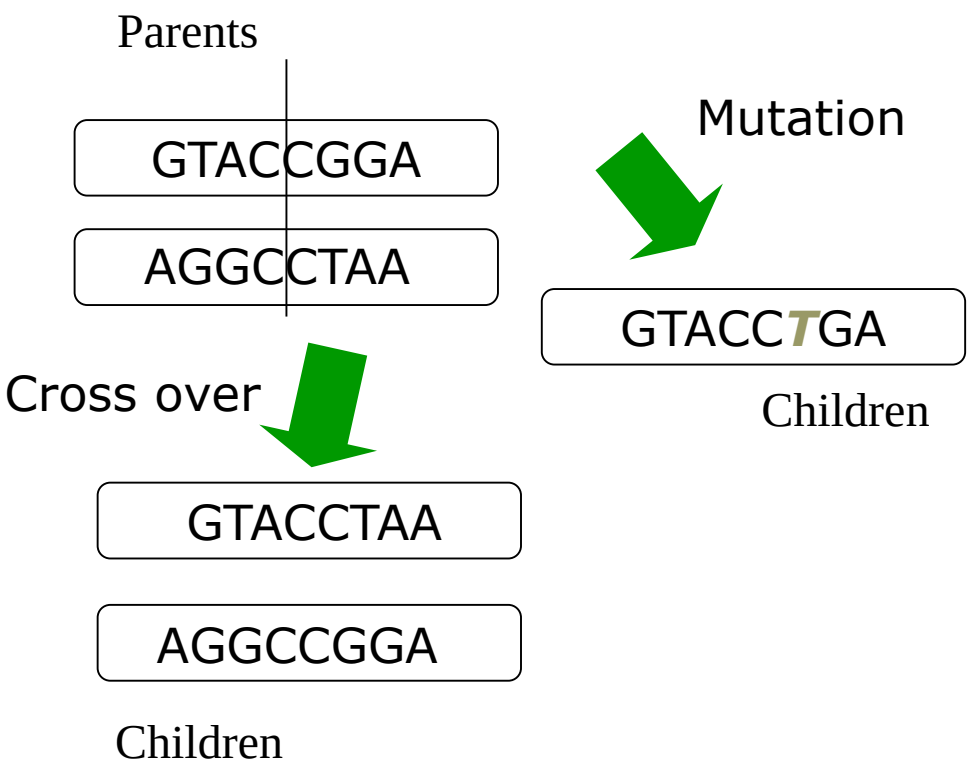


Merits/demerits of clustering

- Merits
  - easy to understand the structure of similarity/dissimilarity of data
  - applicable to any data if we can define distance between them
- Demerits
  - inapplicable if we cannot define data distance
  - results might not be unique depending on the selection of seeds

Genetic algorithm(GA)

- A model of evolution mechanism of species
- One of major optimization method (of evaluation functions)



In concrete calculation, instead of real nucleotides, binary values (1/0) are used.

How do we find an argument x that gives a maximum value y?

$$y = \frac{x}{256}\left(1 - \frac{x}{256}\right)$$

$$0 < x < 256$$

How about this?

$$y = \frac{|x|}{256}\left(1 - \frac{|x|}{256}\right)$$

$$0 < x < 256$$

Step 1

Initialization: generate individuals (x)

- $x = [17]_{10} = [000010001]_2 \rightarrow y = 0.062$
- $x = [80]_{10} = [001010000]_2 \rightarrow y = 0.215$
- $x = [255]_{10} = [01111111\color{red}{1}]_2 \rightarrow y = 0.003$

individuals
(candidate solutions)

chromosomes

Step 2

Cross over: swap a part of individuals

$[000010001]_2$ → $[000011111]_2$ y=0.106

$[011111111]_2$ → $[011110001]_2$ y=0.055

Mutation: randomly flip a bit in chromosomes

$[001010000]_2$ → $[001110000]_2$ y=0.246

Step 3

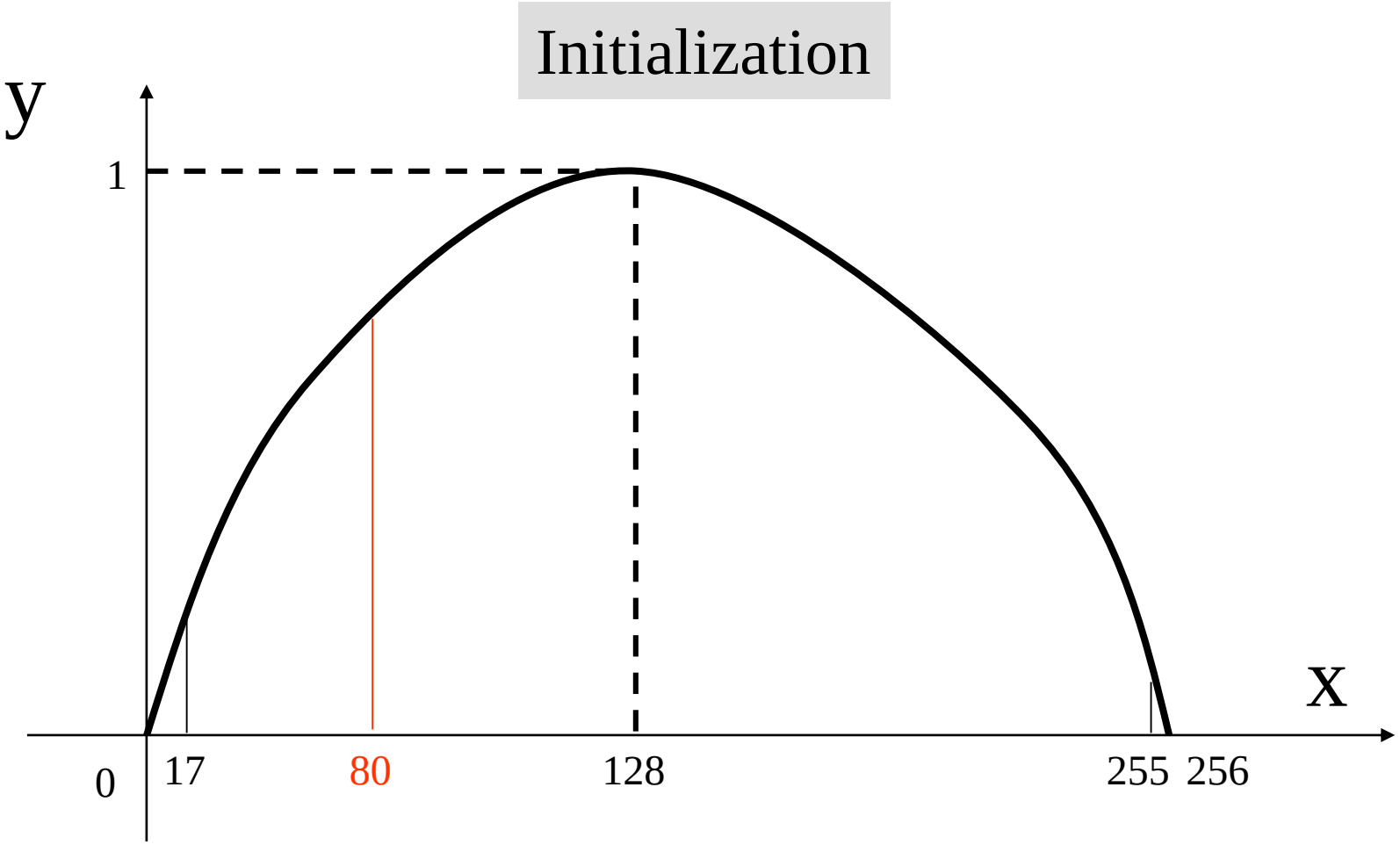
Selection: select individuals who have largest fitness values (= values of the evaluation function)

- $x=[000010001]_2 \rightarrow y=0.062$ (dead)
- $x=[001010000]_2 \rightarrow y=0.215$ (alive)
- $x=[011111111]_2 \rightarrow y=0.003$ (dead)
- $x=[000011111]_2 \rightarrow y=0.106$ (dead)
- $x=[011110001]_2 \rightarrow y=0.055$ (dead)
- $x=[001110000]_2 \rightarrow y=0.246$ (alive)

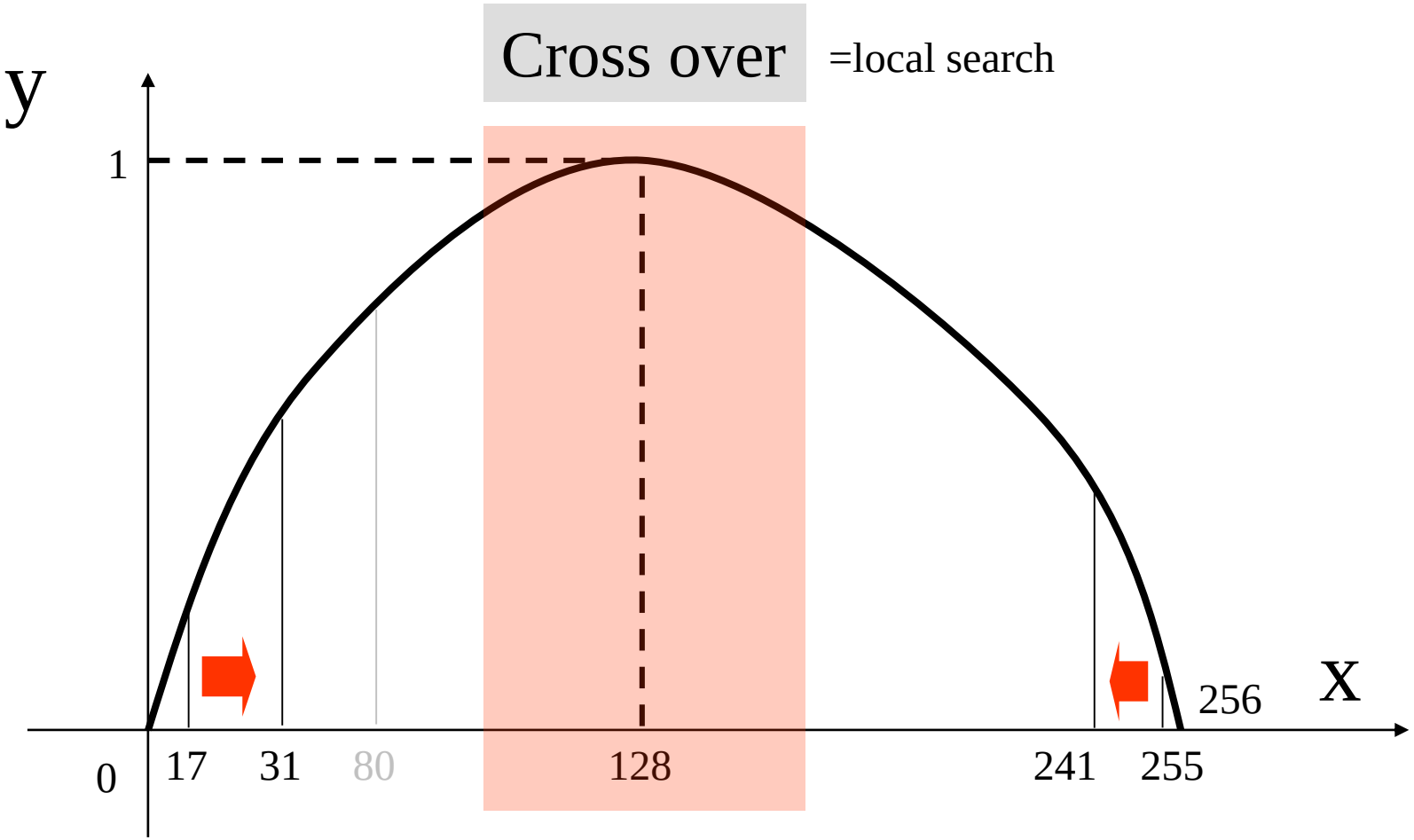
Step 4

Termination: terminate iteration if the fitness values get unchanged

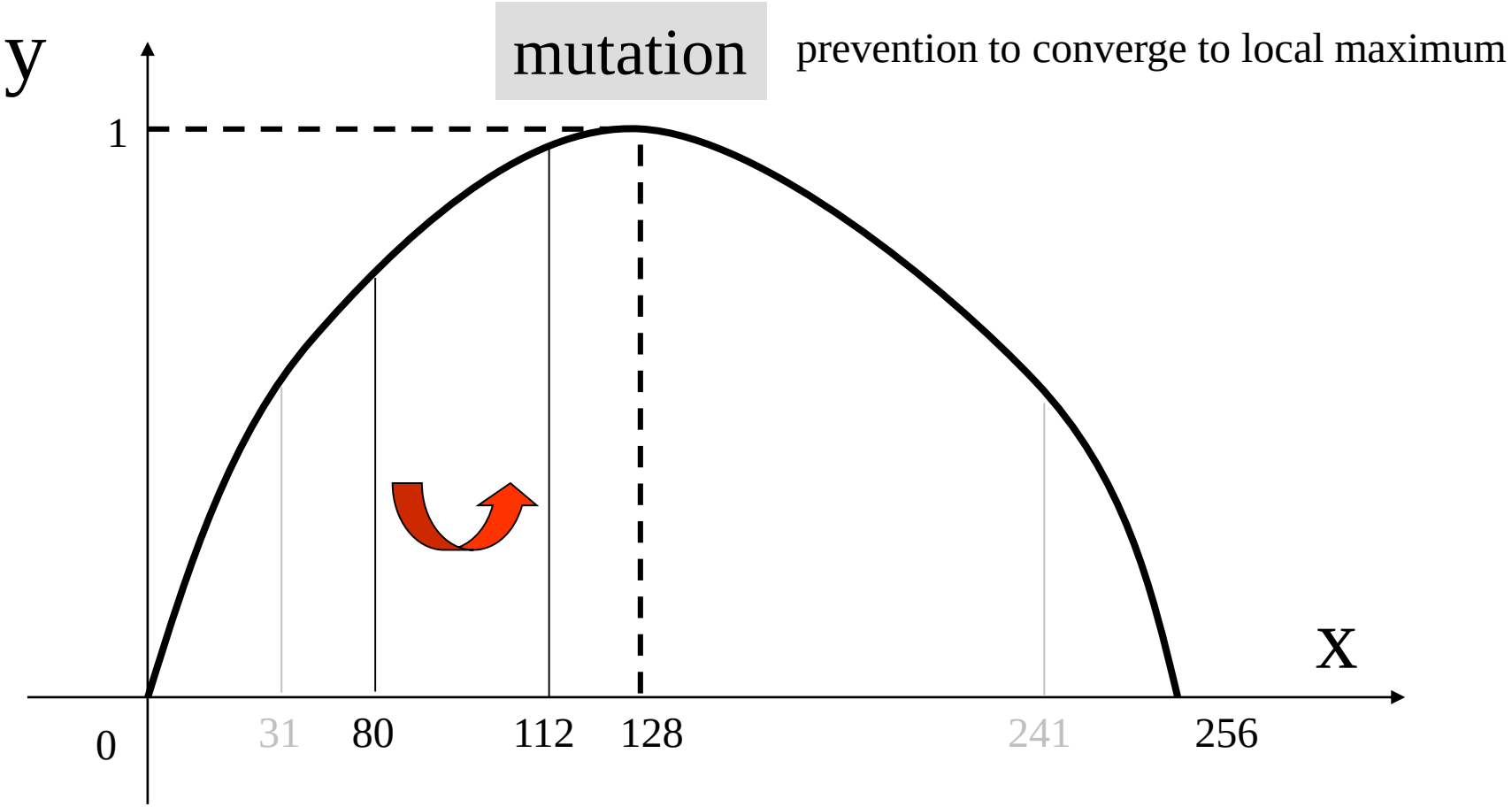
Graphical understanding of GA



Graphical understanding of GA (cont'd)



Graphical understanding of GA (cont'd)



Merits/Demerits of GA

- Merits
  - Easy to understand the algorithm with the analogy of evolution of species
  - Applicable if we have an evaluation function
- Demerits
  - A globally optimal solution is not guaranteed to be obtained by this algorithm