Data Clustering

Algorithmic Thinking

Luay Nakhleh

Department of Computer Science

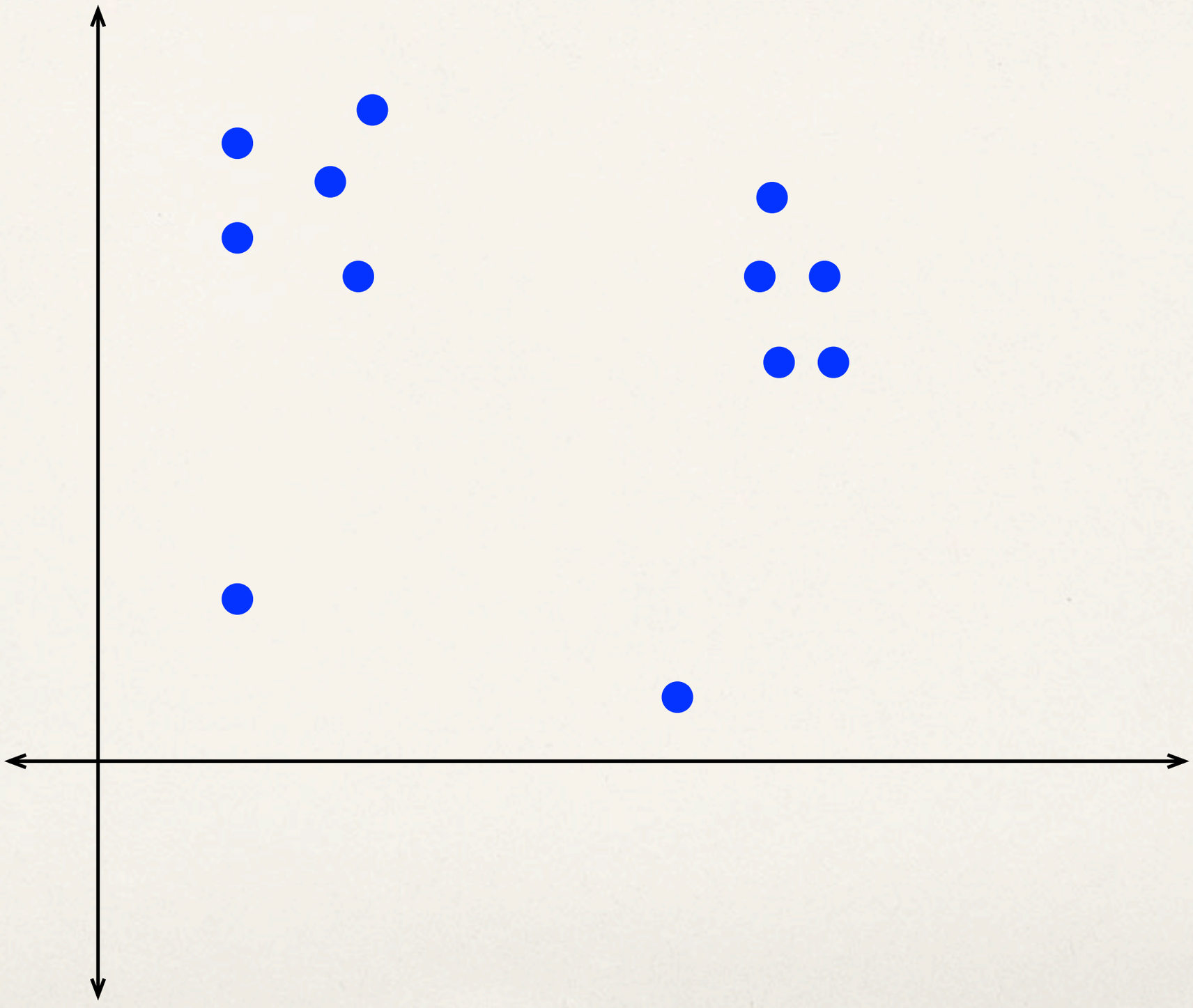
Rice University

- Data clustering is the task of partitioning a set of objects into groups such that the similarity of objects within each group is higher than that of objects across groups.

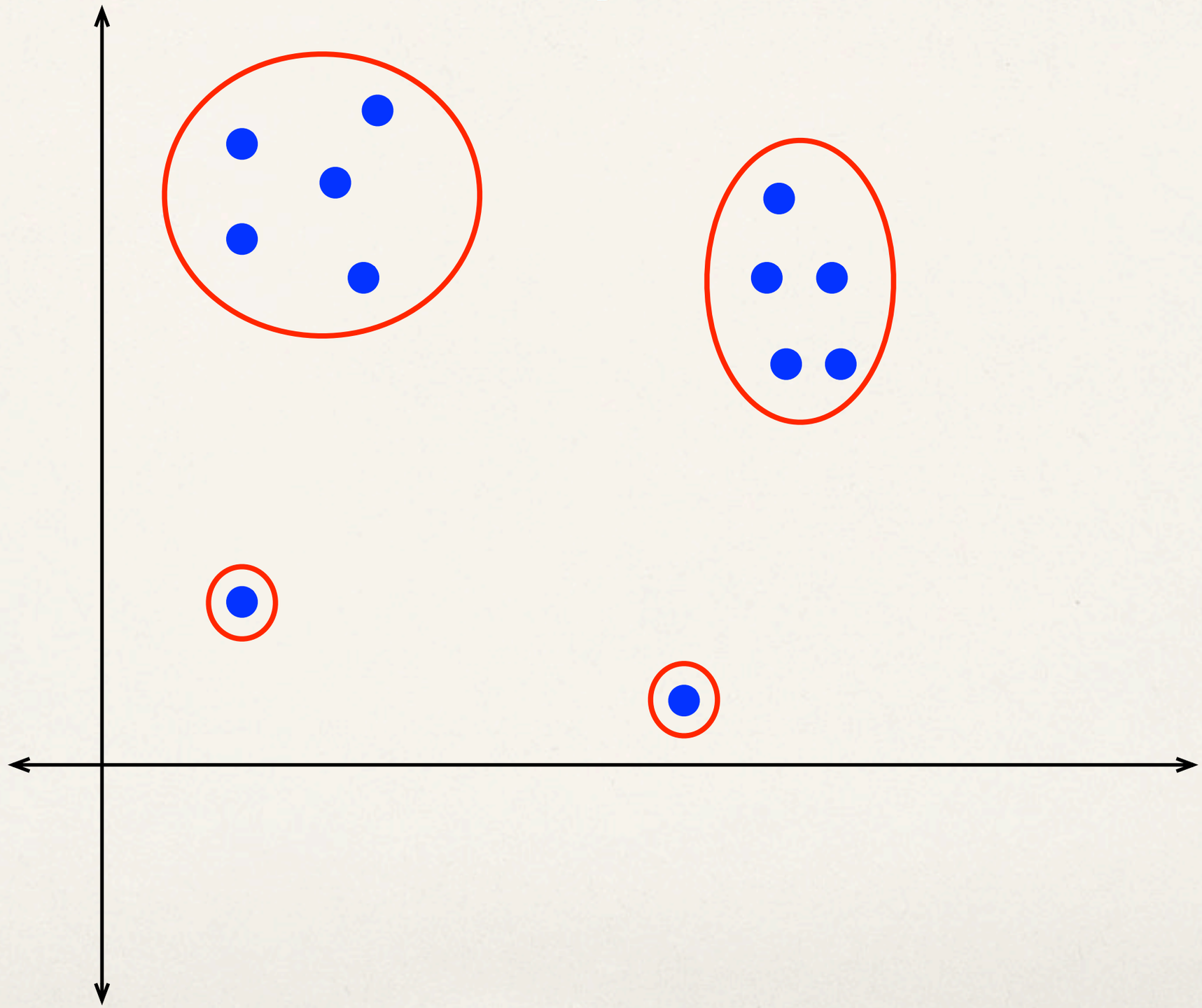
- Clustering helps to reveal hidden patterns in the data and to summarize large data sets.

- It has numerous applications in bioinformatics, image analysis, social networks, ...

12 objects in the 2D space



4 clusters: points within a cluster are closer ("similar")
to each other than to points in other clusters



The Problem

- Input: A set of P of points, a distance measure d, and a positive integer k $(1 \leq k \leq |P|)$

- Output: A partition of P into k subsets, $C1,\ldots,Ck$, each of which we call a cluster, such that the similarity of points within each cluster is higher than that across clusters.

- To cluster the data, we need
  - A distance measure (to quantify how similar or dissimilar two objects are)
  - An algorithm for clustering the data based on the distance measure

- For data in the 2D space, an obvious distance measure is the Euclidian distance.

Two Clustering Algorithms

- Hierarchical clustering
- k-means clustering

Hierarchical Clustering

- Input: A set P of points, and a number of clusters *k*
- Initialization: Put each point in a cluster by itself
- Repeat until *k* clusters:
  - Find two closest clusters and merge them into one

- Q: How do we find a closest pair of clusters?
- A: In our case, using the closest pair algorithm.
- Q: But the closest pair algorithm works on points where each point is given by its (x,y) coordinates and the distance between two points is the Euclidian distance. What do we do about clusters?
- A: Represent each cluster by its center and take the distance between two clusters as the distance between their centers.

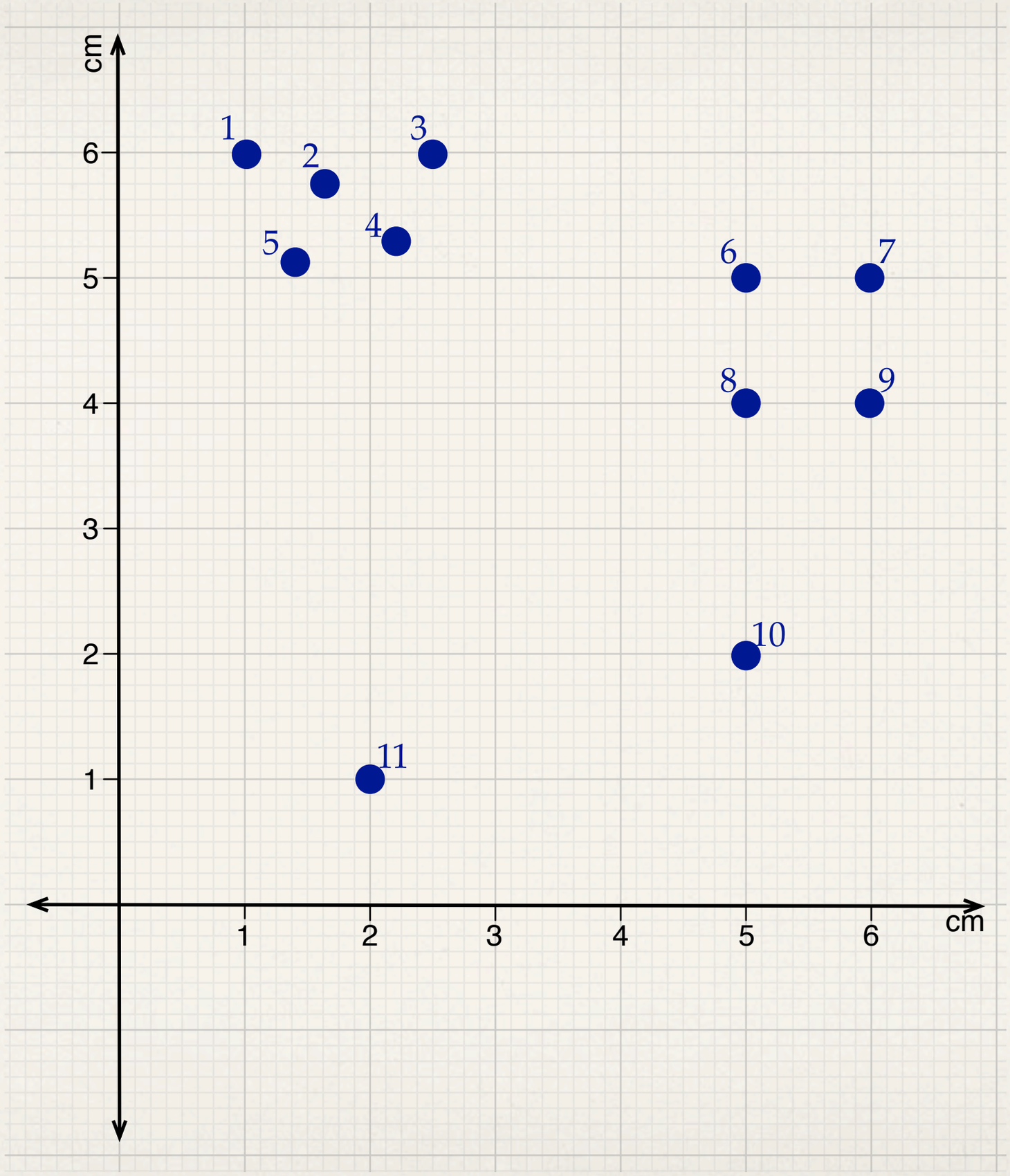
- Let (x_1,y_1), (x_2,y_2),...,(x_m,y_m) be the coordinates of the m points in a cluster C.

- The center of cluster C is given by the point (x',y') where

$$x' = \sum_{i=1}^{m} x_i/m \qquad y' = \sum_{i=1}^{m} y_i/m$$

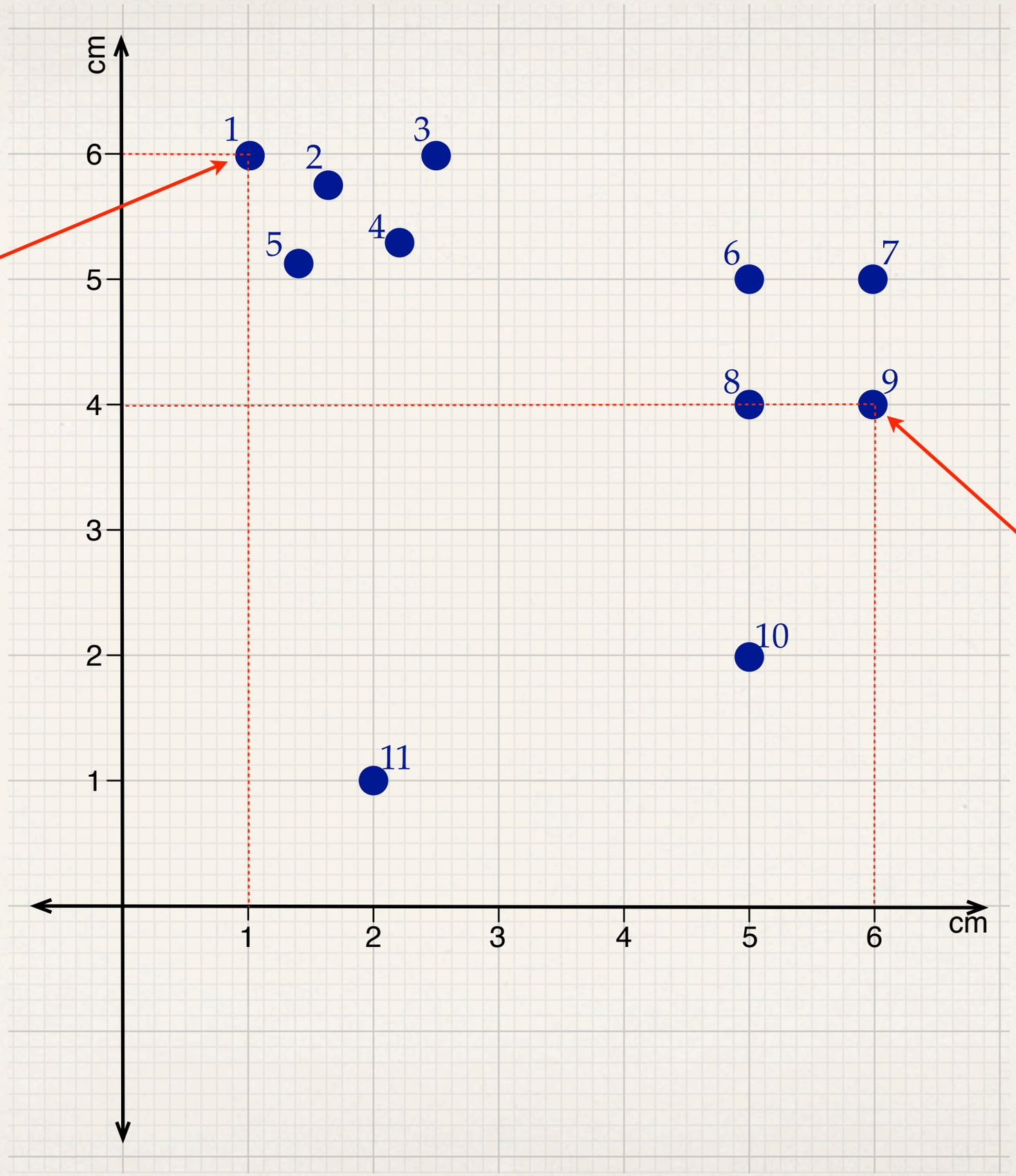
- The distance between two clusters is then the Euclidian distance between their centers.

- Let's consider an example of 11 points (1, 2, ..., 11) that we want to group into 4 clusters.


cm
6
5
4
3
2
1
1
2
3
4
5
6
cm
1
2
3
4
5
6
7
8
9
10
11

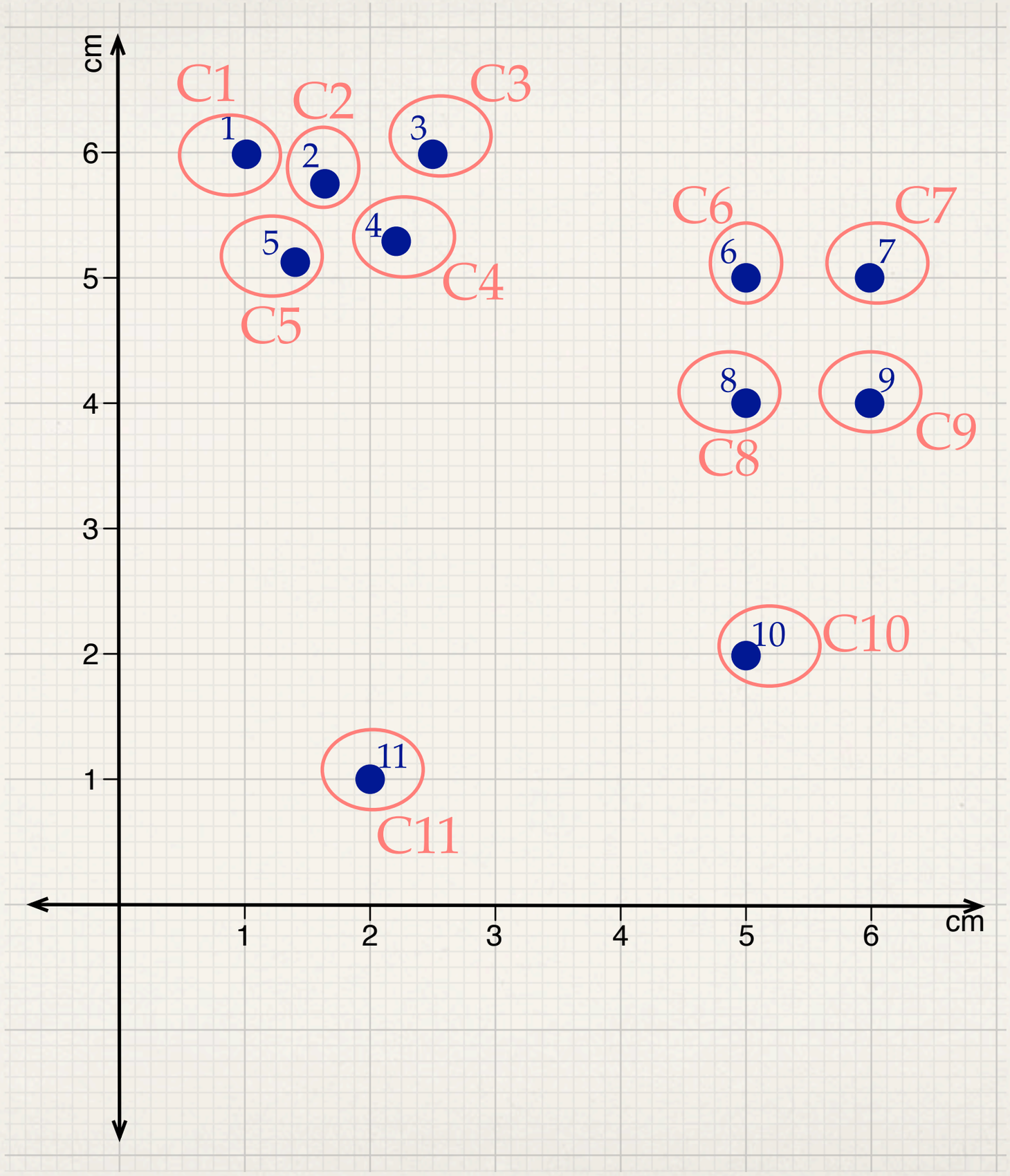
- Each point is given by its (x,y) coordinates.

coordinates of point 1 are (1,6)



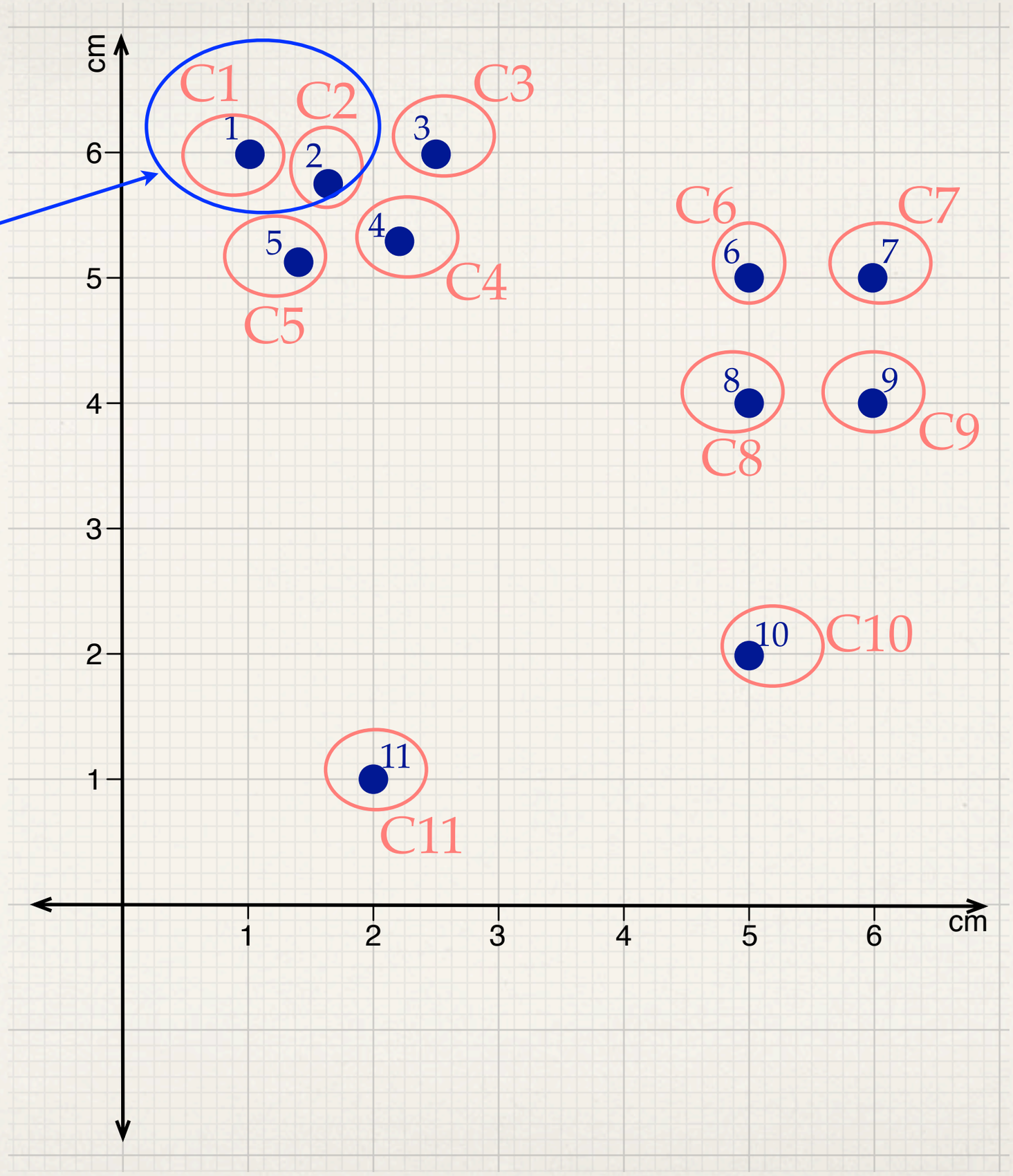
coordinates of point 6 are (6,4)

- (1) Form 11 clusters, C1,C2,...,C11 each of which contains exactly one of the points.

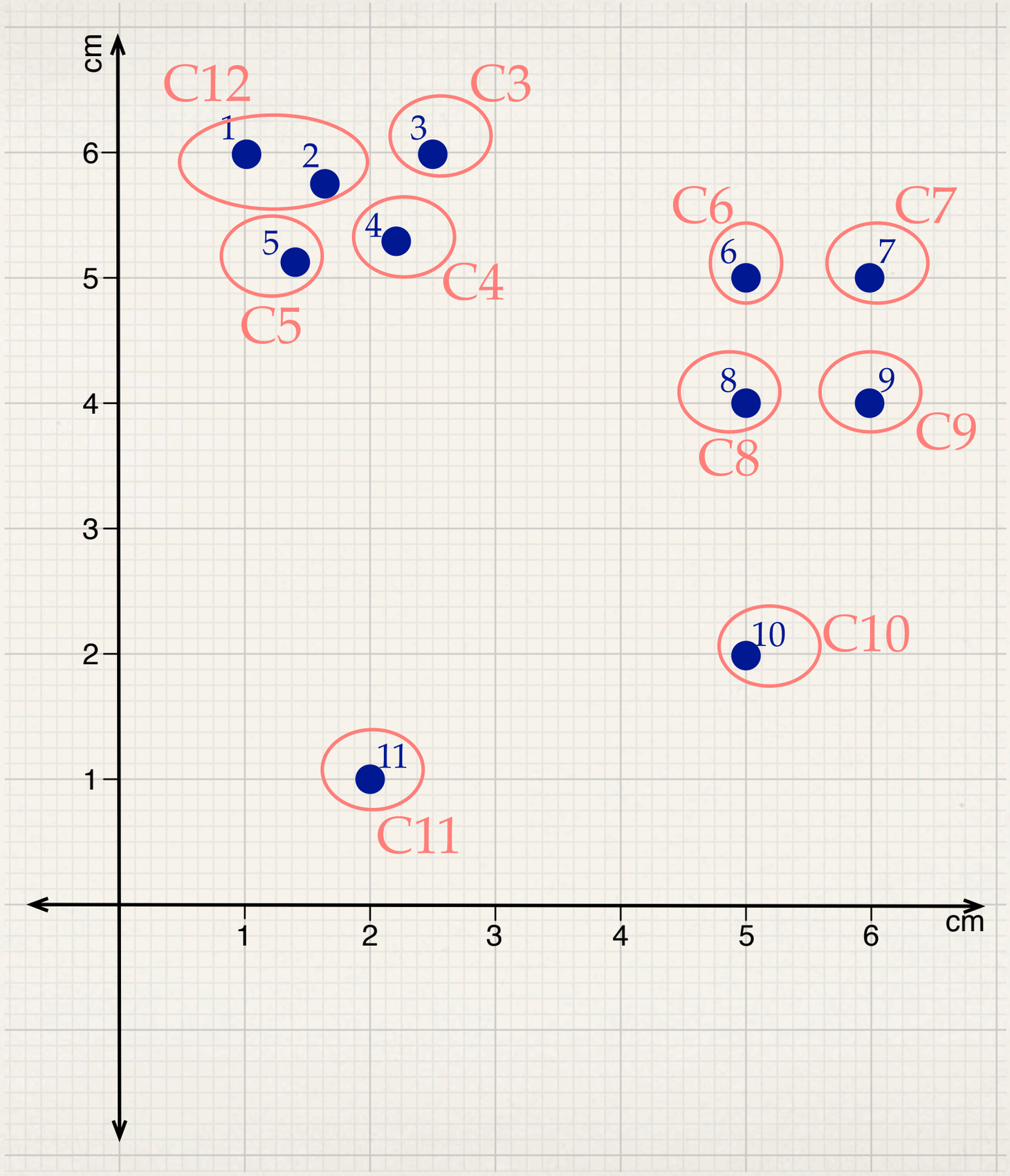

cm
6
5
4
3
2
1
C1
C2
C3
C4
C5
C6
C7
C8
C9
C10
C11
1
2
3
4
5
6
7
8
9
10
11
1
2
3
4
5
6
cm

- (2) Find the two closest clusters

the two closest clusters

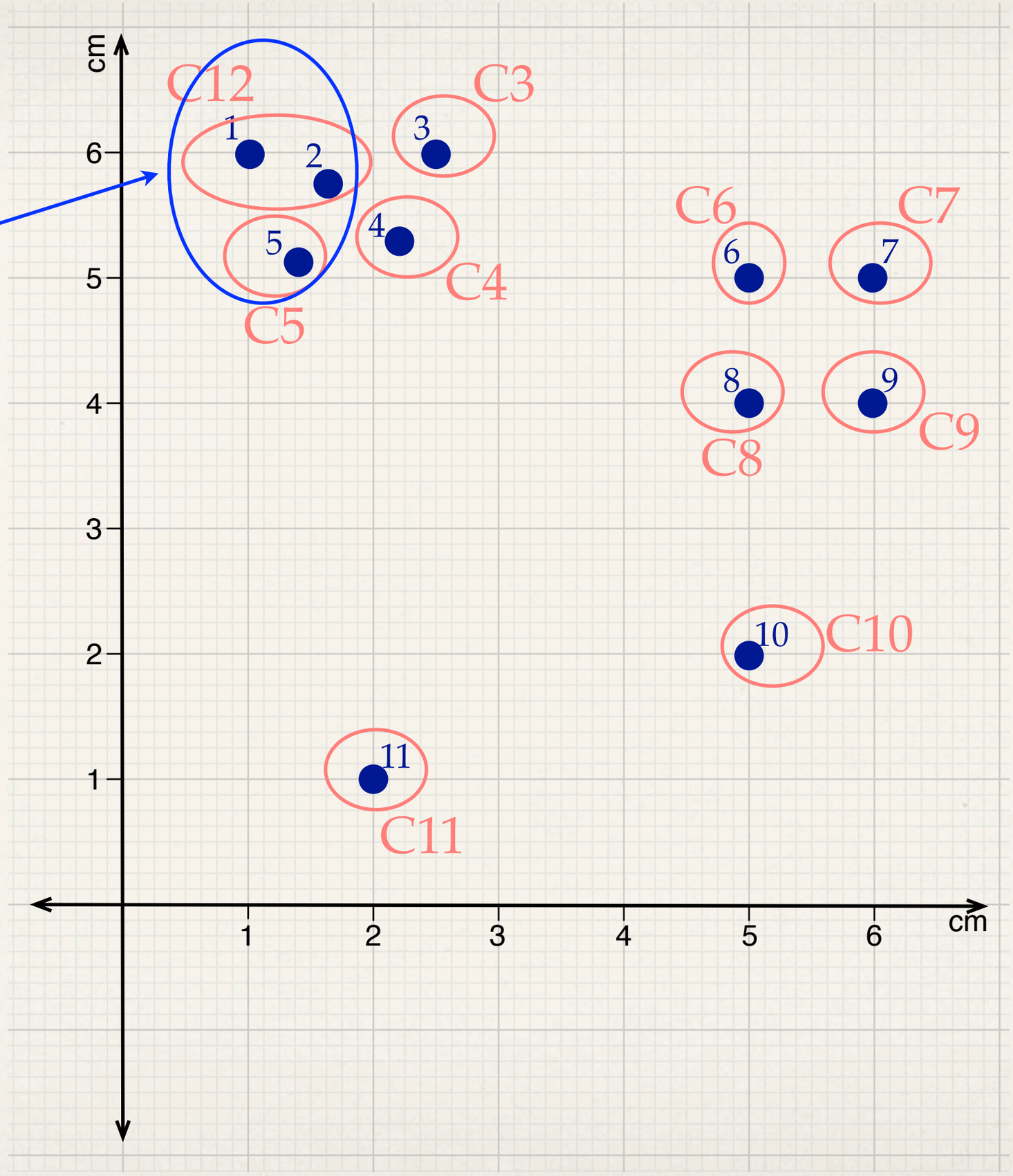


- ... and merge them into one (thus, the number of clusters would decrease by one after this step)

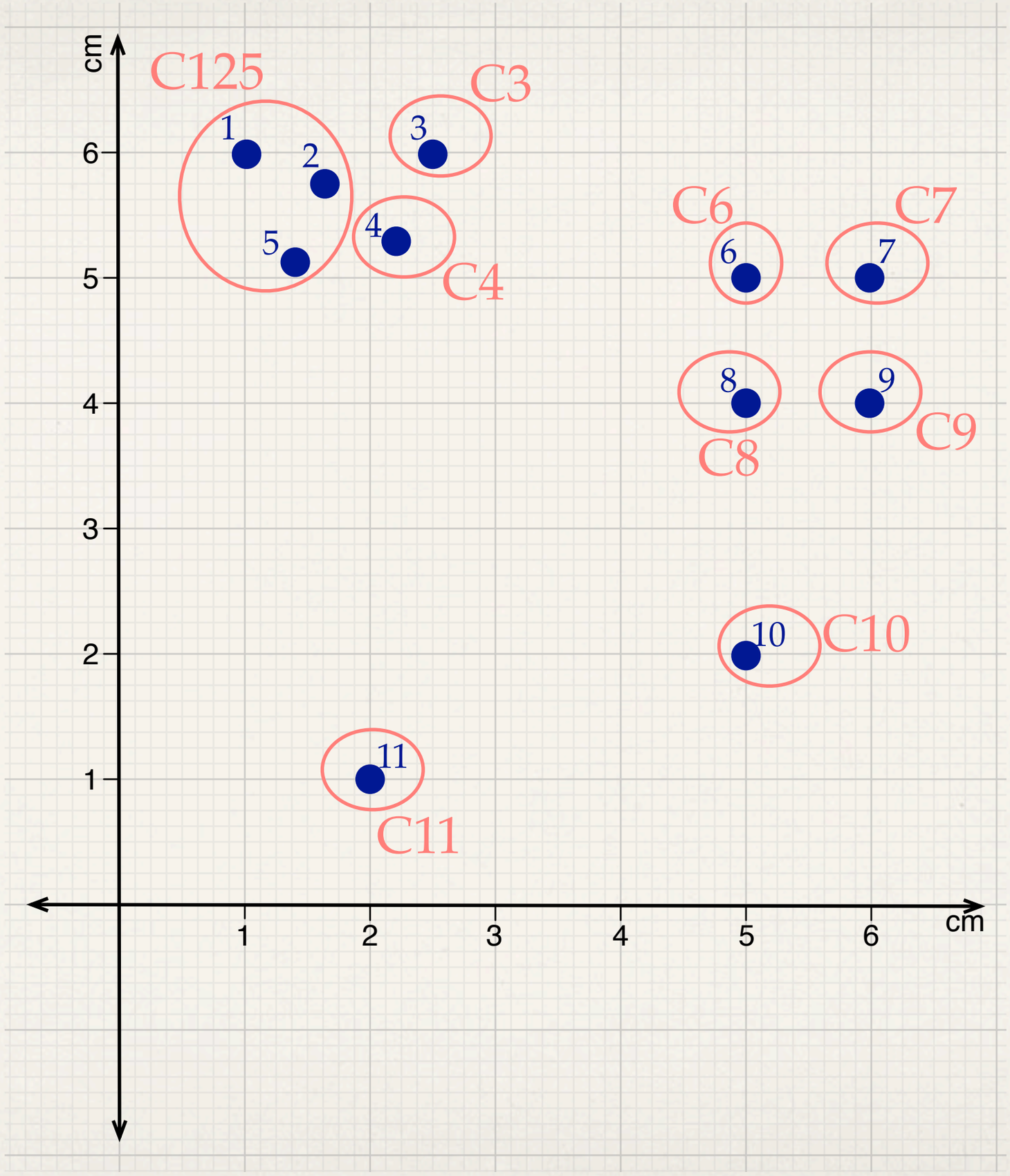

cm
6
5
4
3
2
1
C12
C3
1
2
3
4
5
C4
C5
C6
C7
6
7
8
9
C8
C9
10
C10
11
C11
1
2
3
4
5
6
cm

- (3) Find the two closest clusters

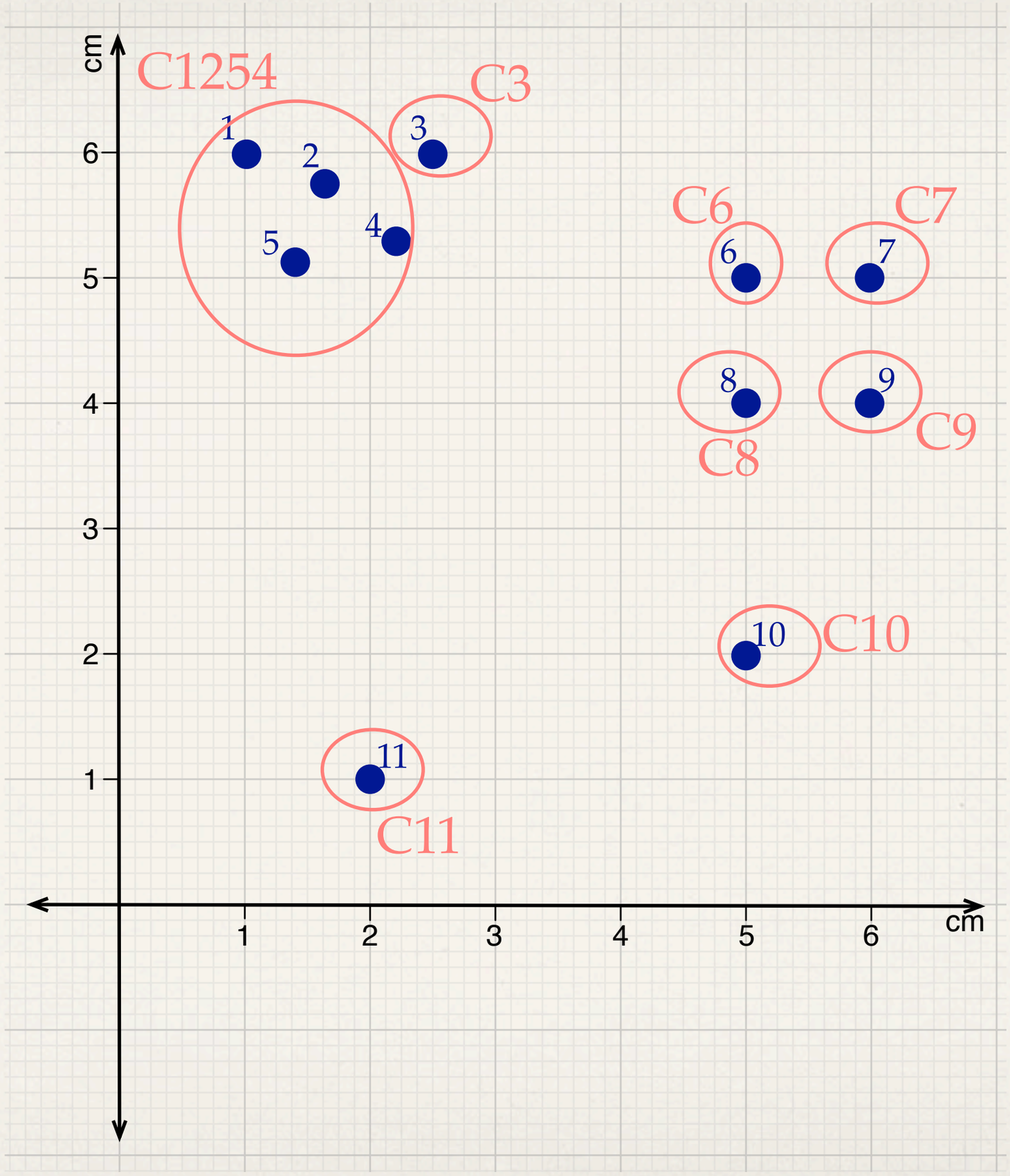
the two closest clusters



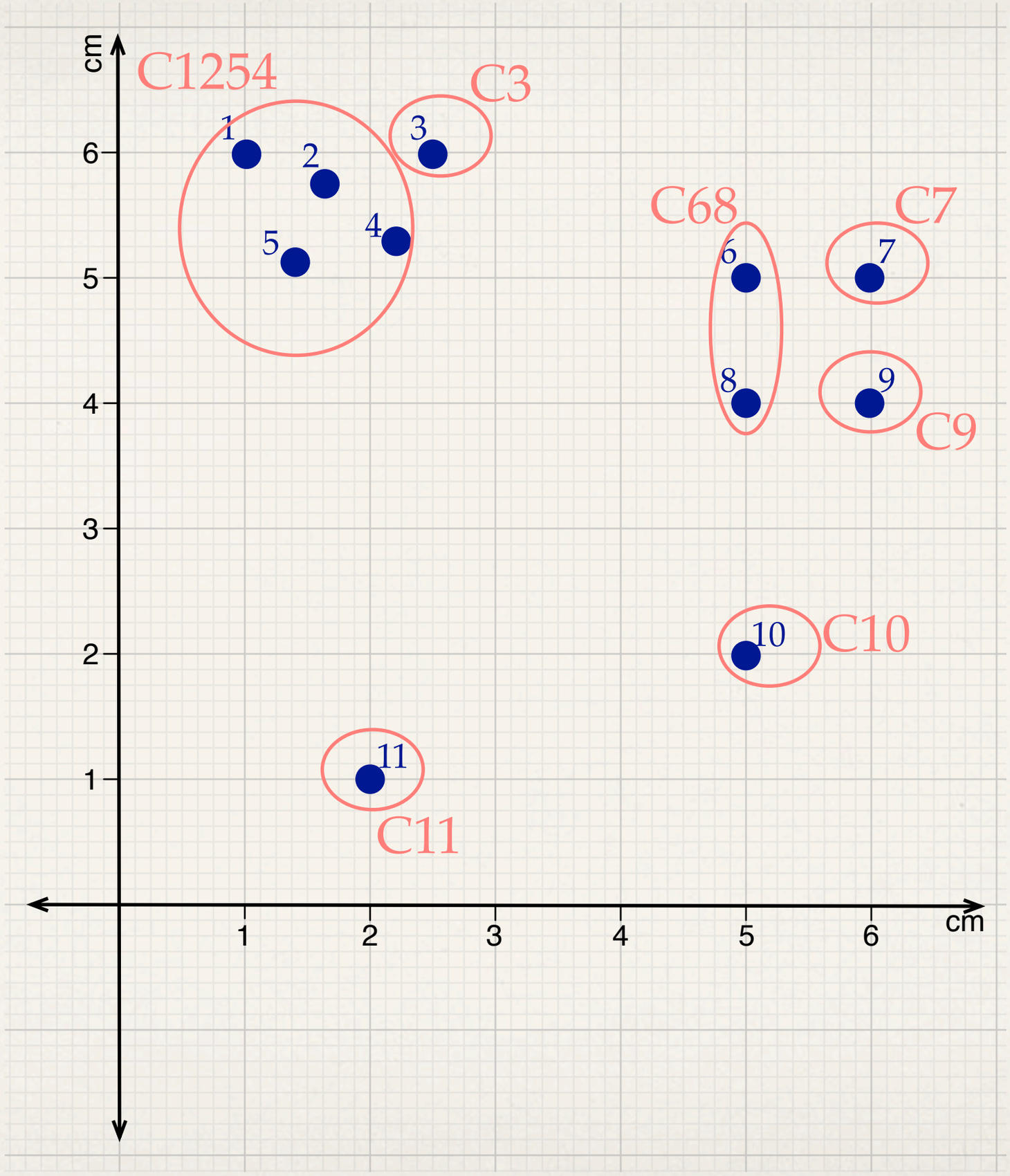
- ... and merge them into one (thus, the number of clusters would decrease by one after this step)


cm
C125
C3
C4
C6
C7
C8
C9
C10
C11
1
2
3
4
5
6
7
8
9
10
11
1
2
3
4
5
6
cm

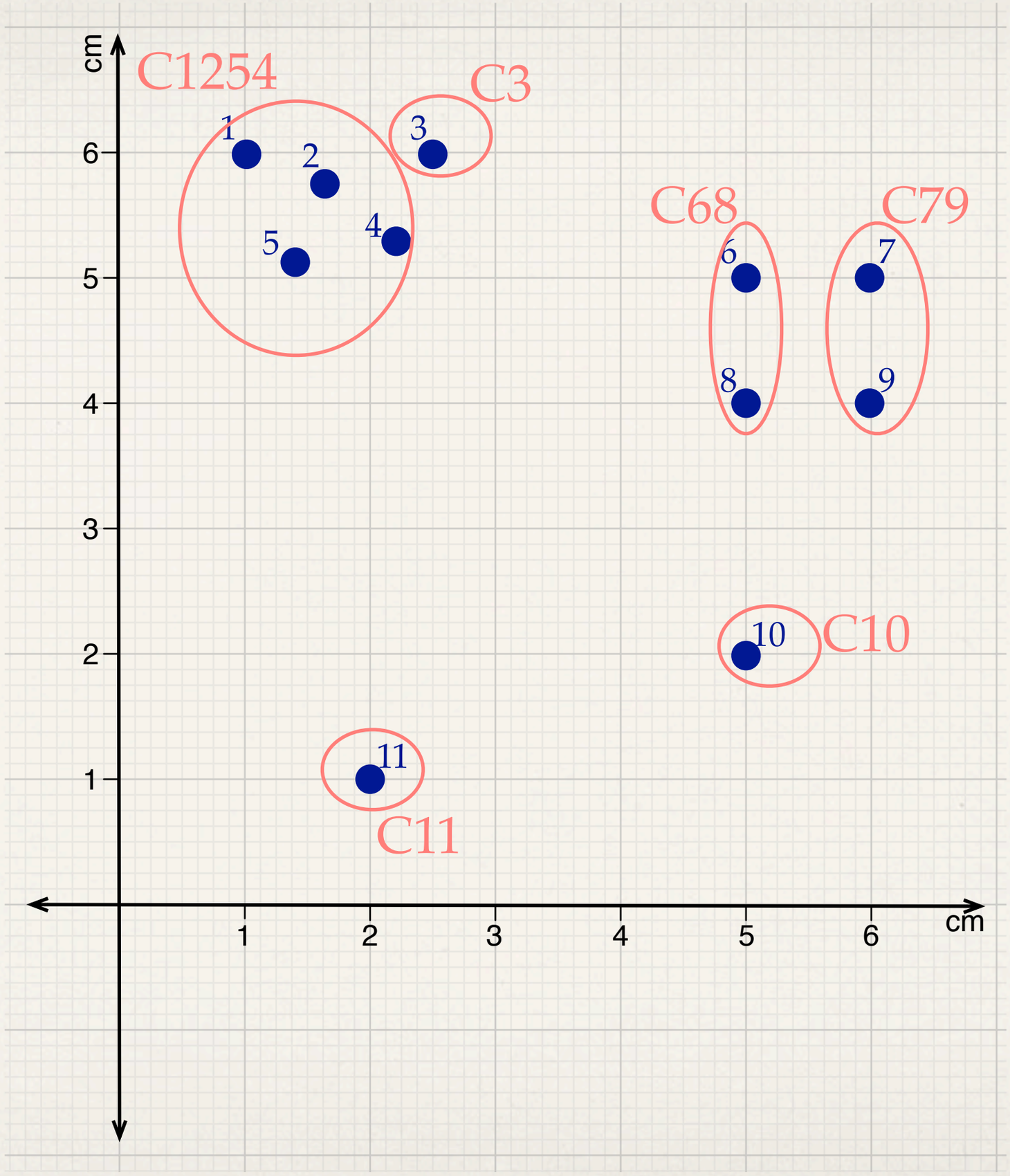
- (4) Find the two closest clusters and merge them into one (thus, the number of clusters would decrease by one after this step)


cm
C1254
C3
1
2
3
4
5
C6
C7
6
7
8
9
C8
C9
10
C10
11
C11
1
2
3
4
5
6
cm

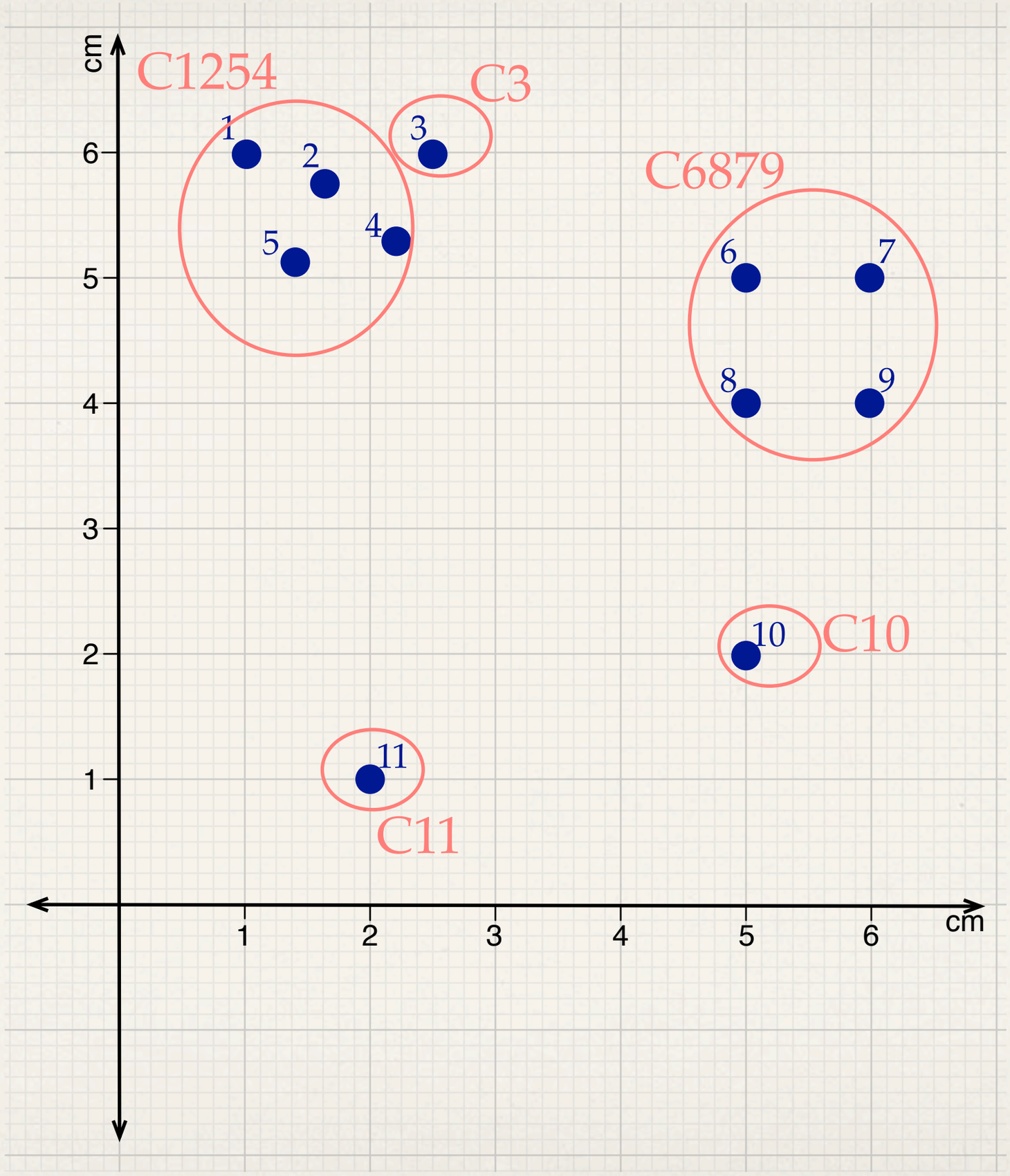
- (5) Find the two closest clusters and merge them into one (thus, the number of clusters would decrease by one after this step)


cm
C1254
C3
1
2
3
4
5
C68
C7
6
7
8
9
C9
C10
10
11
C11
1
2
3
4
5
6
cm

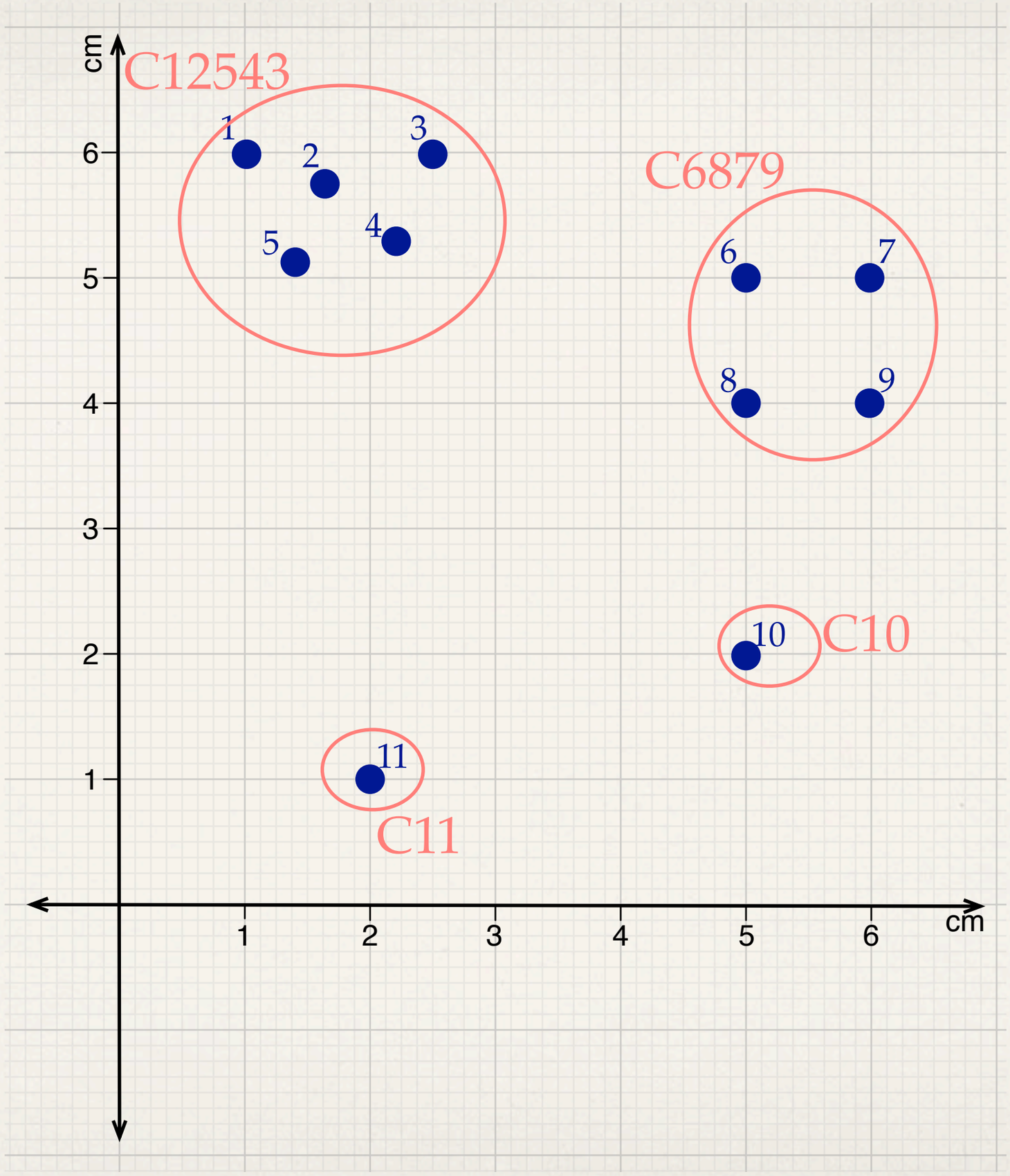
- (6) Find the two closest clusters and merge them into one (thus, the number of clusters would decrease by one after this step)


cm
C1254
C3
1
2
3
4
5
C68
C79
6
7
8
9
10
C10
11
C11
1
2
3
4
5
6
cm

- (7) Find the two closest clusters and merge them into one (thus, the number of clusters would decrease by one after this step)


cm
C1254
C3
C6879
C10
C11
1
2
3
4
5
6
7
8
9
10
11
1
2
3
4
5
6
cm

- (8) Find the two closest clusters and merge them into one (thus, the number of clusters would decrease by one after this step)

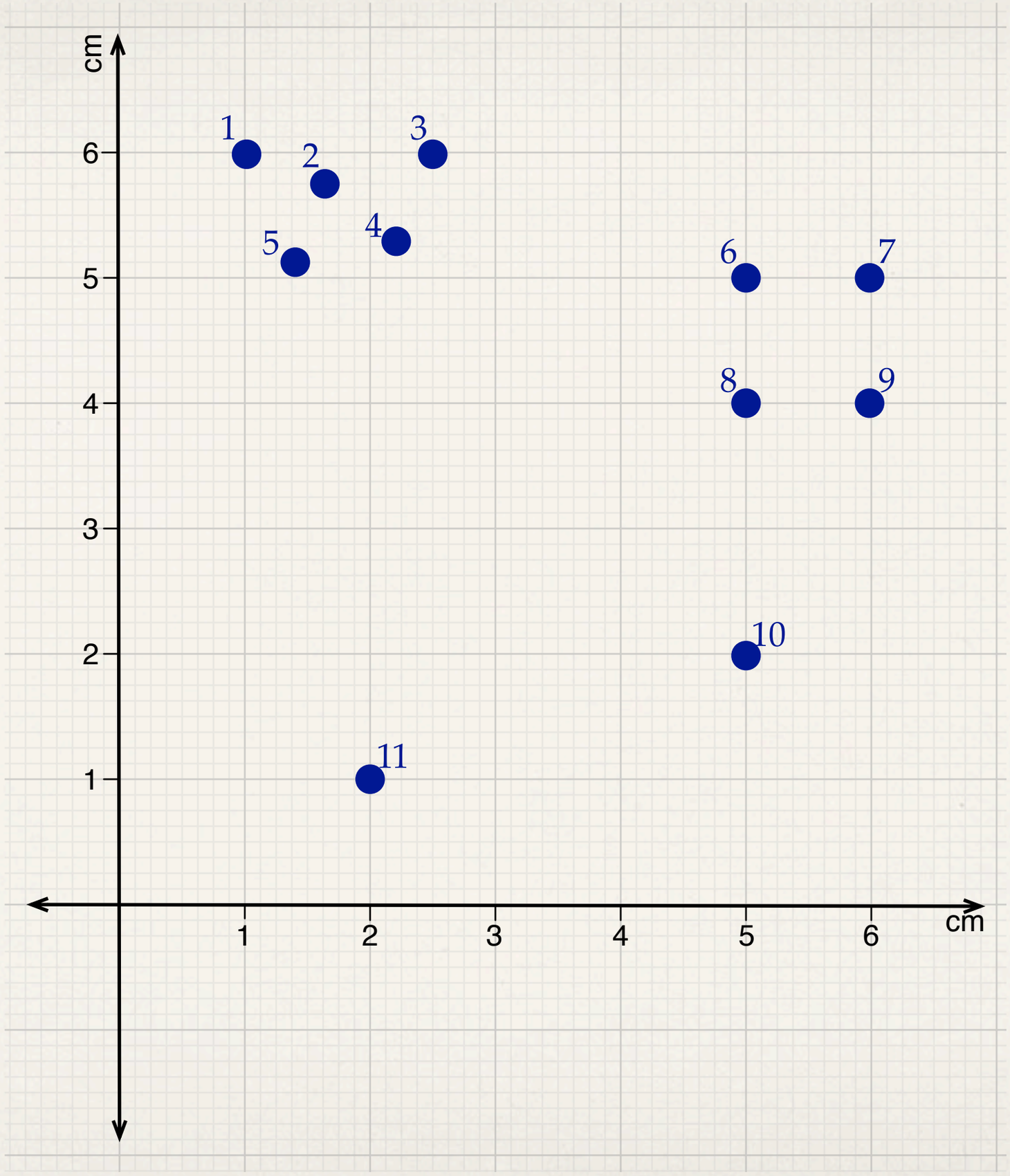

cm
C12543
1
2
3
4
5
C6879
6
7
8
9
10
C10
11
C11
1
2
3
4
5
6
cm

- (9) The pre-specified number of clusters (4) has been reached, so we return the four clusters, each as a set containing the points in it:
  - {1,2,3,4,5}
  - {6,7,8,9}
  - {10}
  - {11}

k-means Clustering

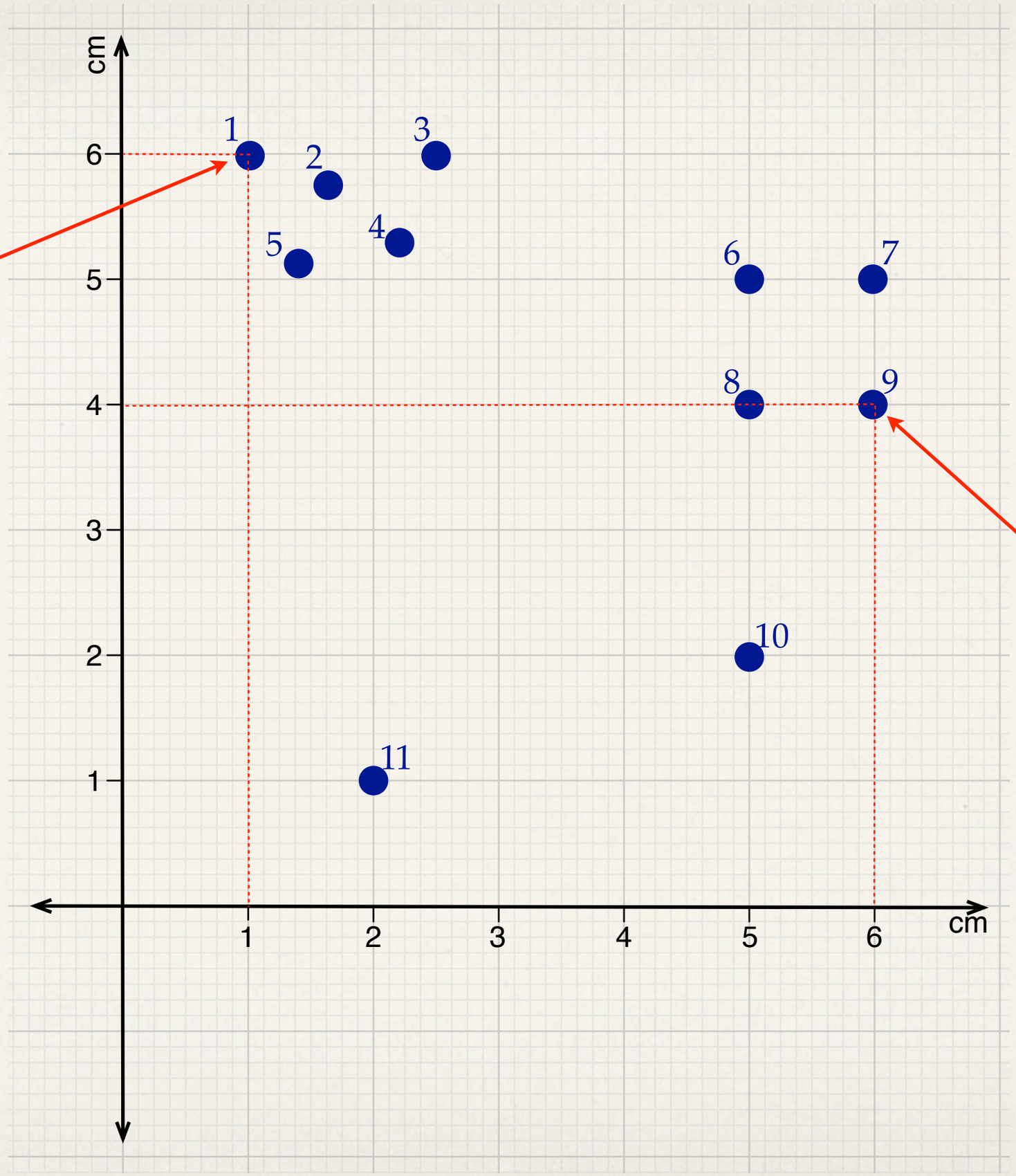
- Input: A set of P of points, a number of clusters k, and a number of iterations q
- Initialize k centers
- Iterate q times:
  - Initialize k empty clusters
  - Add each of the n points to the cluster with the closest center
  - Recompute the clusters' centers using its new set of points
- Return the set of k clusters

- Let's consider an example of 11 points (1, 2, ..., 11) that we want to group into 4 clusters.


cm
6
5
4
3
2
1
1
2
3
4
5
6
cm
1
2
3
4
5
6
7
8
9
10
11

- Each point is given by its (x,y) coordinates.

coordinates of point 1 are (1,6)

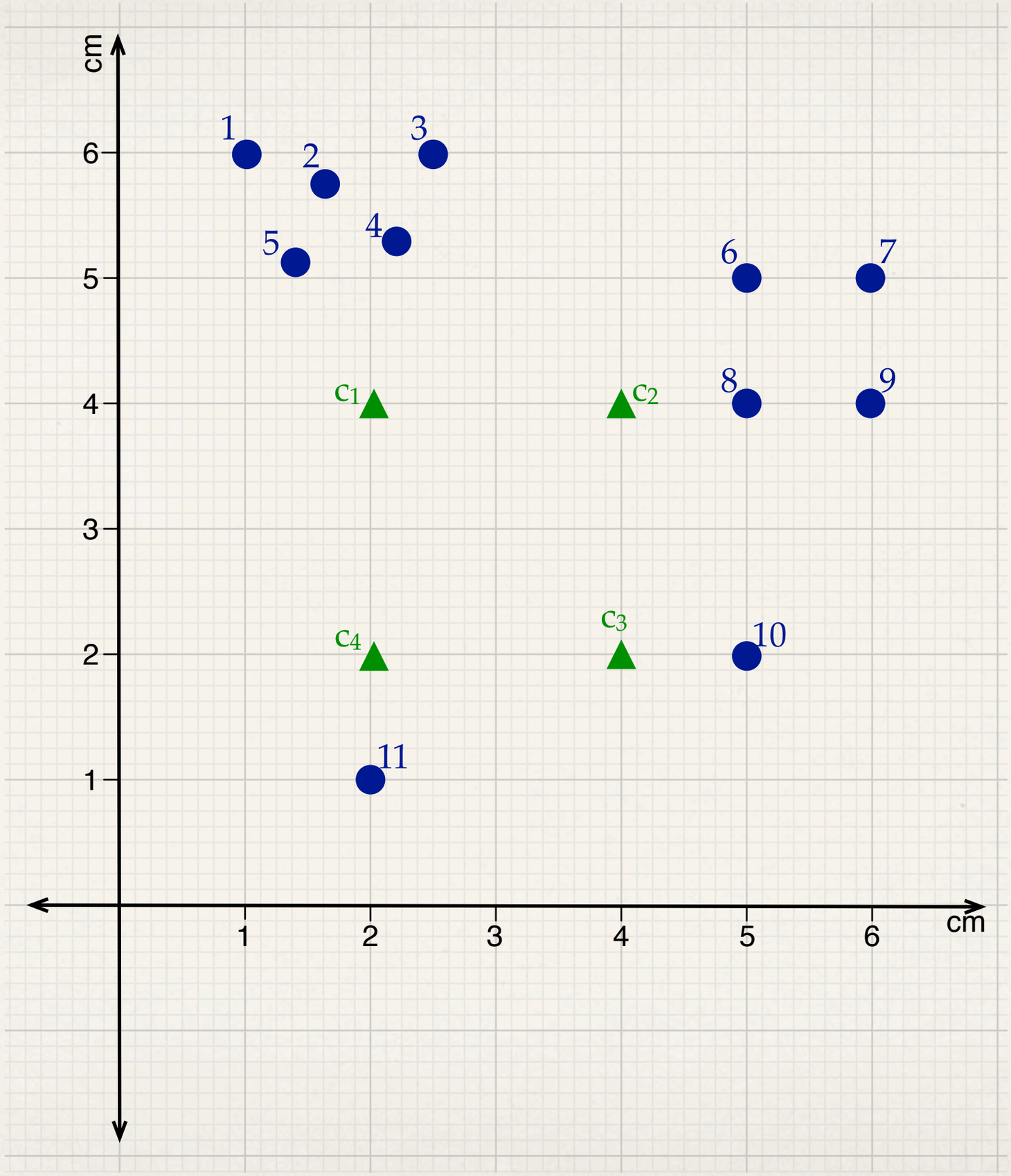


coordinates of point 6 are (6,4)

- First step: Choose 4 center points for the four clusters

Four centers c_1, c_2, c_3, c_4 each with its coordinates

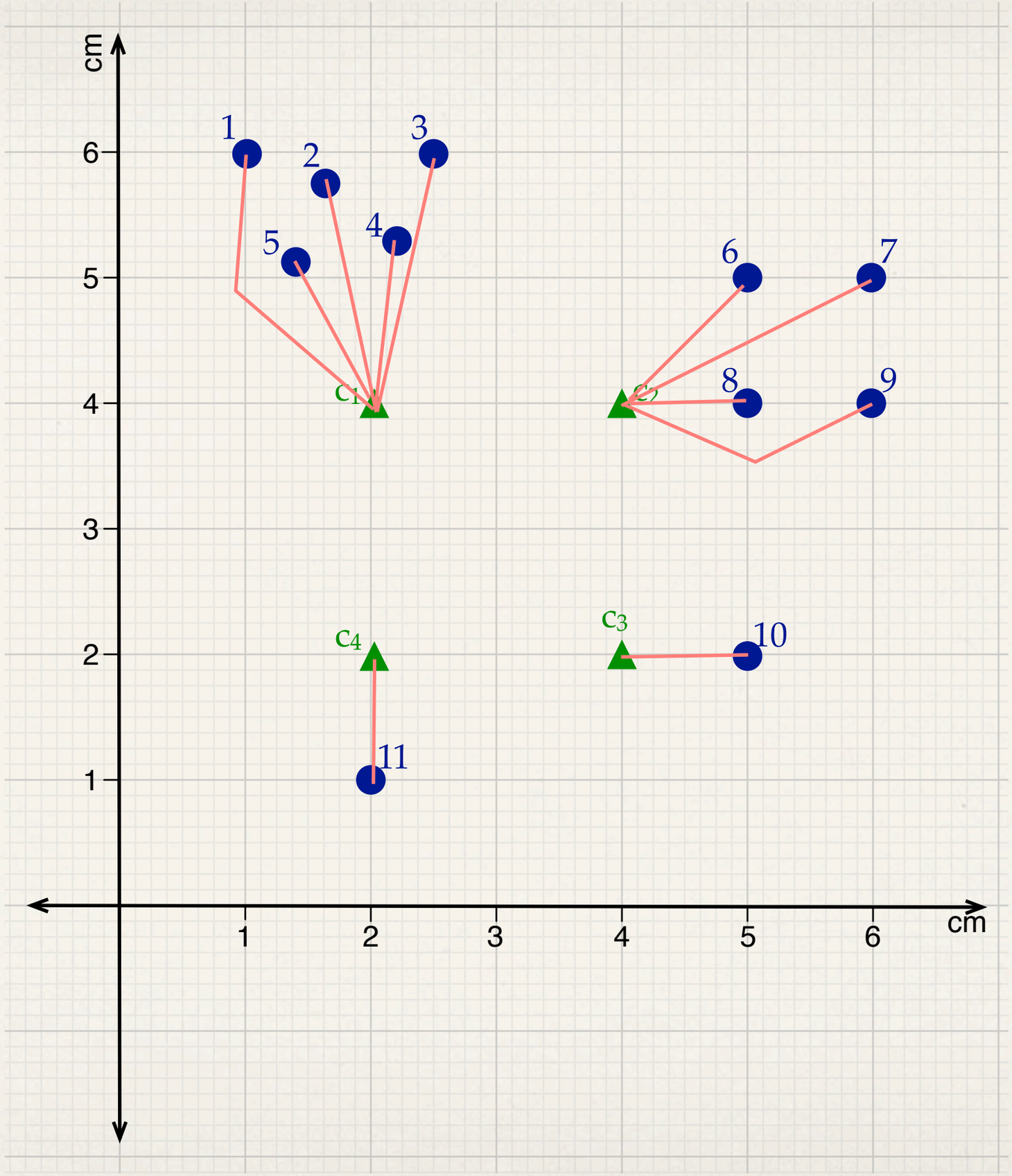
coordinates of c_1 are (2,4)



- Second step: For each of 11 points, find the closest center (of the four) to it.

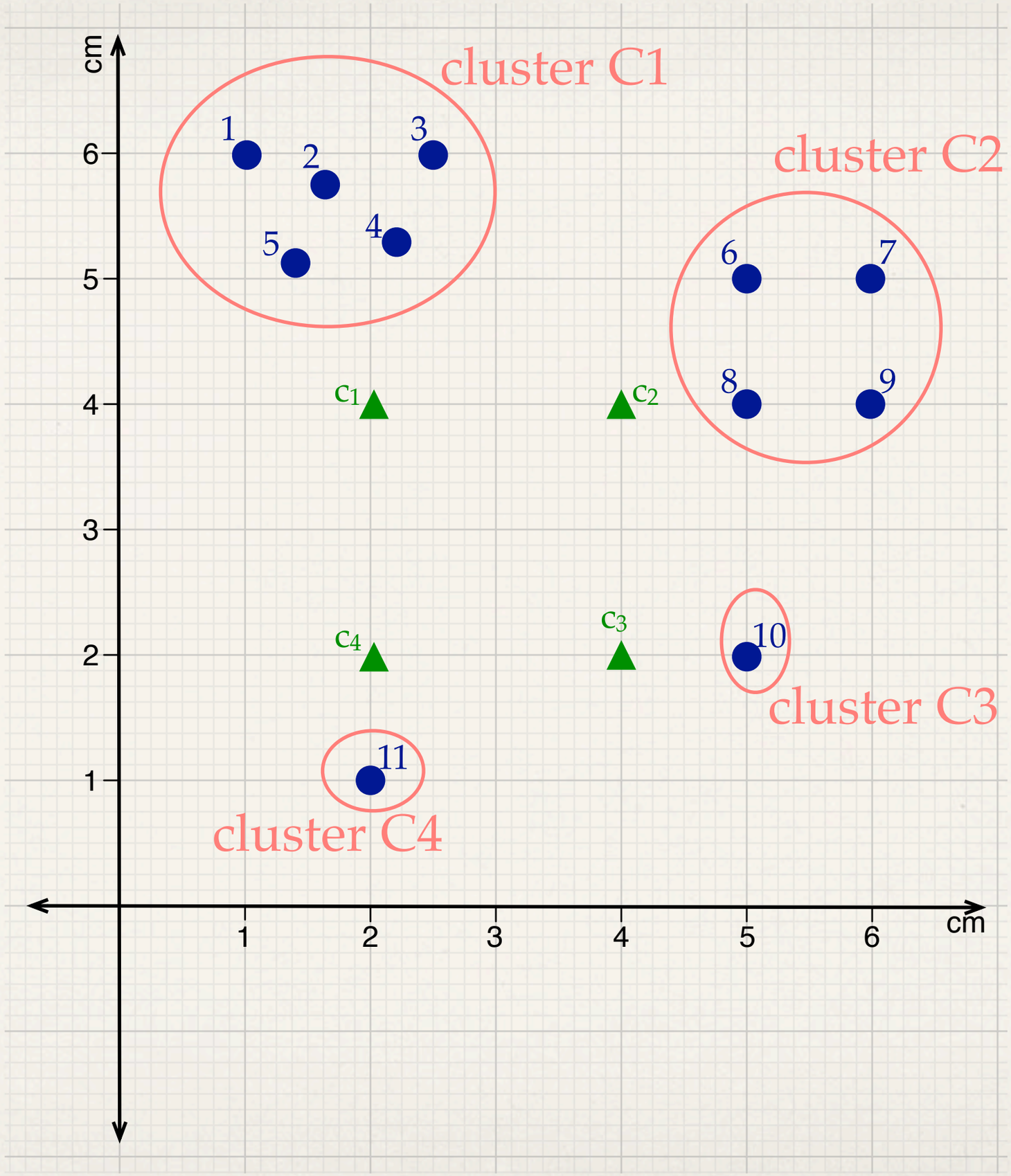
Four centers c_1, c_2, c_3, c_4 each with its coordinates

lines show closest center to each point



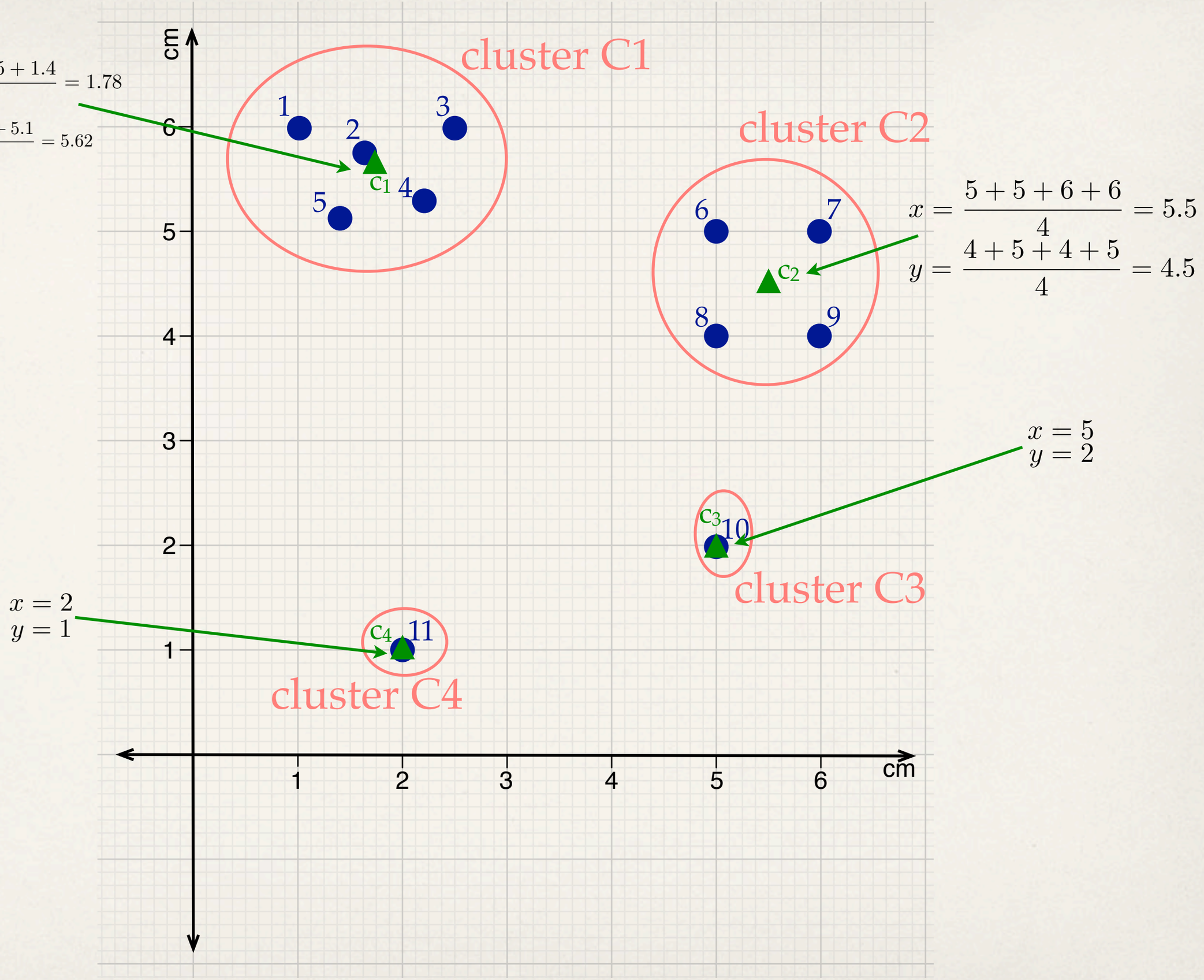
- Third step: Form a cluster for each set of points that "belong" to the same center (thus, forming four clusters).

Four centers c_1, c_2, c_3, c_4 each with its coordinates



- Fourth step: For each cluster, recompute its center as shown previously (to get the x and y coordinates of a center, average the x coordinates and average the y coordinates, respectively, of all points that belong to the center's cluster)

Four centers c_1, c_2, c_3, c_4 each with its coordinates



$$x = \frac{1 + 1.75 + 2.5 + 2.25 + 1.4}{5} = 1.78$$

$$y = \frac{6 + 5.75 + 6 + 5.25 + 5.1}{5} = 5.62$$

$$x = \frac{5 + 5 + 6 + 6}{4} = 5.5$$

$$y = \frac{4 + 5 + 4 + 5}{4} = 4.5$$

$x = 5$
$y = 2$

$x = 2$
$y = 1$

- Now, repeat from the second step: For each point, find its closest center (of the new four centers), form four clusters, and continue...

- After running the procedure for a (pre-specified) number of iterations, we return the four clusters, each as a set containing the points in it:
  - {1,2,3,4,5}
  - {6,7,8,9}
  - {10}
  - {11}