Introduction to
Microarray Data Analysis and Gene Networks
Lecture 5

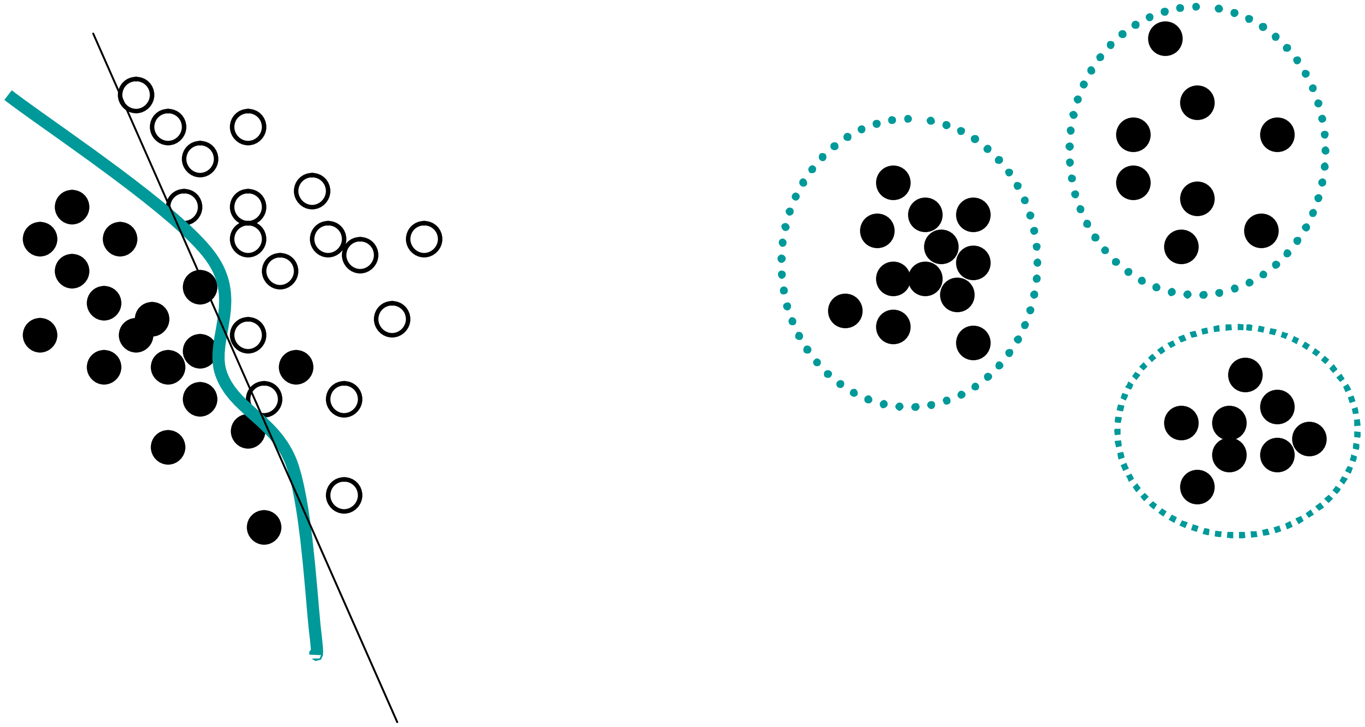
Alvis Brazma

European Bioinformatics Institute

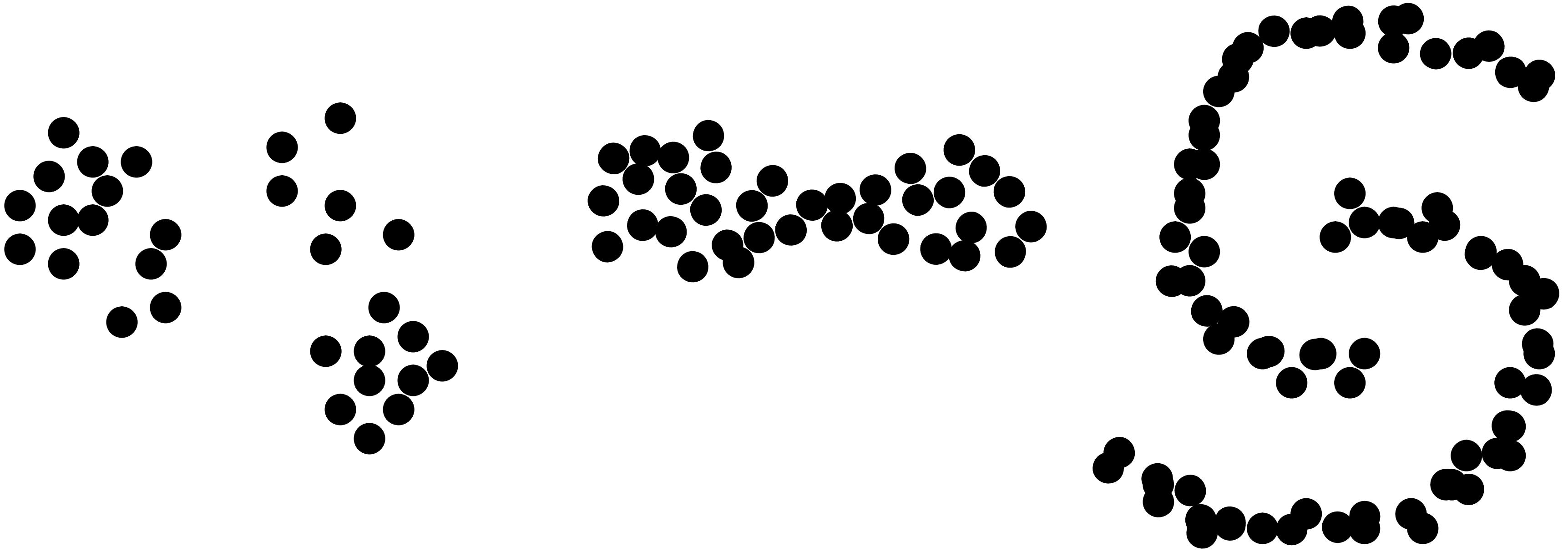
Lecture 5

- Clustering
  - Hierarchical
  - K-means
- A few minutes about representing experimental designs
  - Experiment design graphs, replicates
  - Experimental factors
- A few minutes about supervised learning
- Practical

Supervised vs. unsupervised analysis - class discovery vs. clustering



What is a cluster?



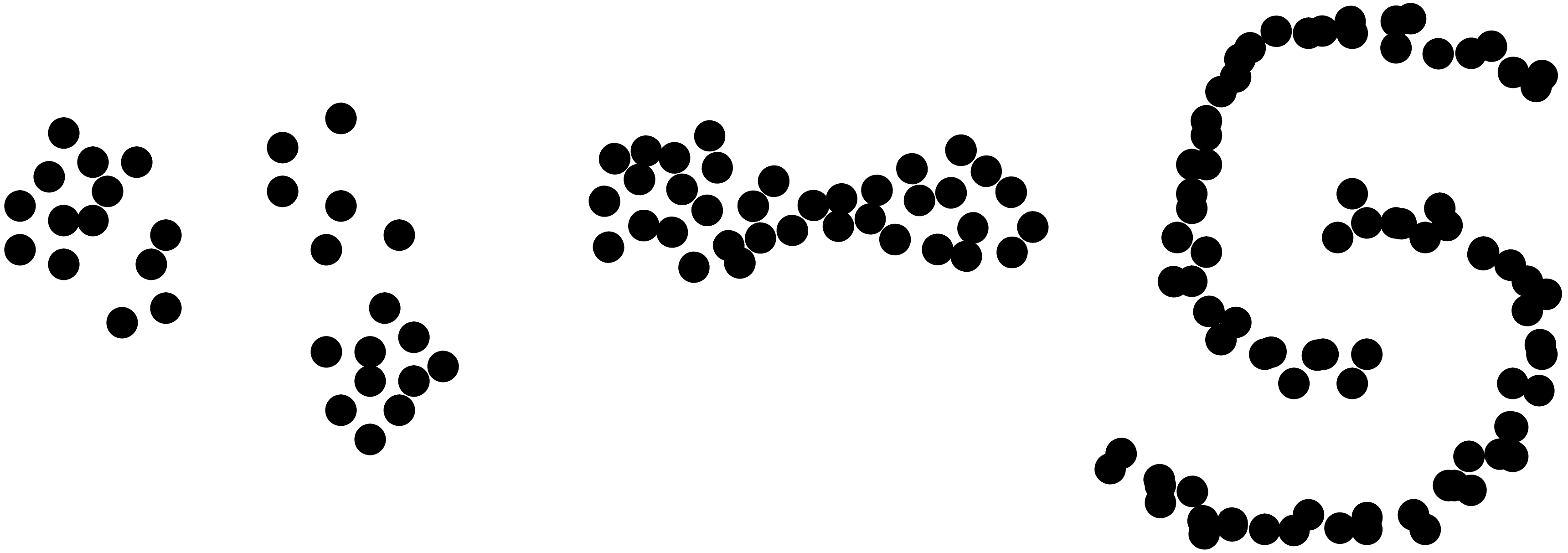
•In a set of elements, subsets of elements that are in some sense closer to each other than 'average'

•Closeness can be defined by a distance measure

•Distance by itself is not sufficient

- •How to measure distance between more than 2 points?
- •Shape of the cluster?
- •Thresholds of closeness which are the same clusters, which are not

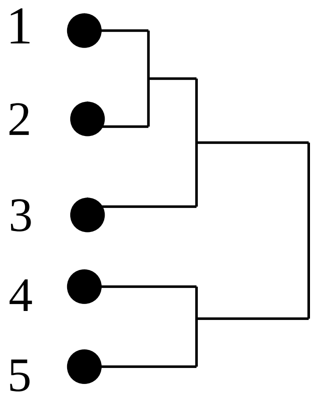
What is a cluster?

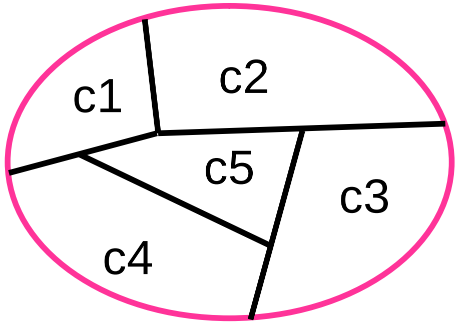


The definition of what is a ‘cluster’ is difficult
In practice it is defined by an algorithm that finds clusters

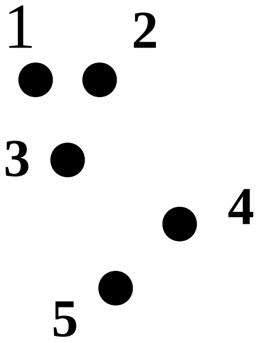
Clustering algorithms

- Hierarchical vs flat
  - Hierarchical clustering builds a hierarchical tree (also called dendrogram) showing the relationship among the elements
  - Flat clustering partitions the set of elements in subsets (nonoverlapping or overlapping)

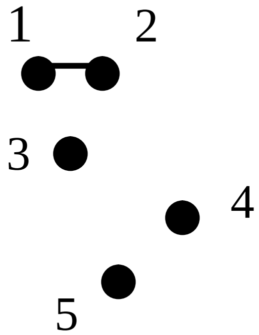




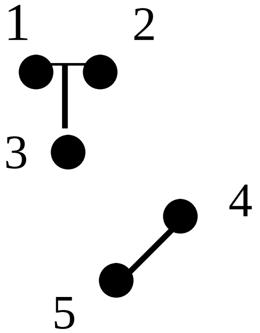
Hierarchical clustering – how does it work?



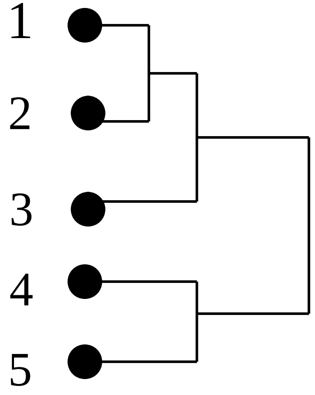
	1	2	3	4	5
1		**1**	2	5	6
2			2	4	5
3				3	3
4					2
5					

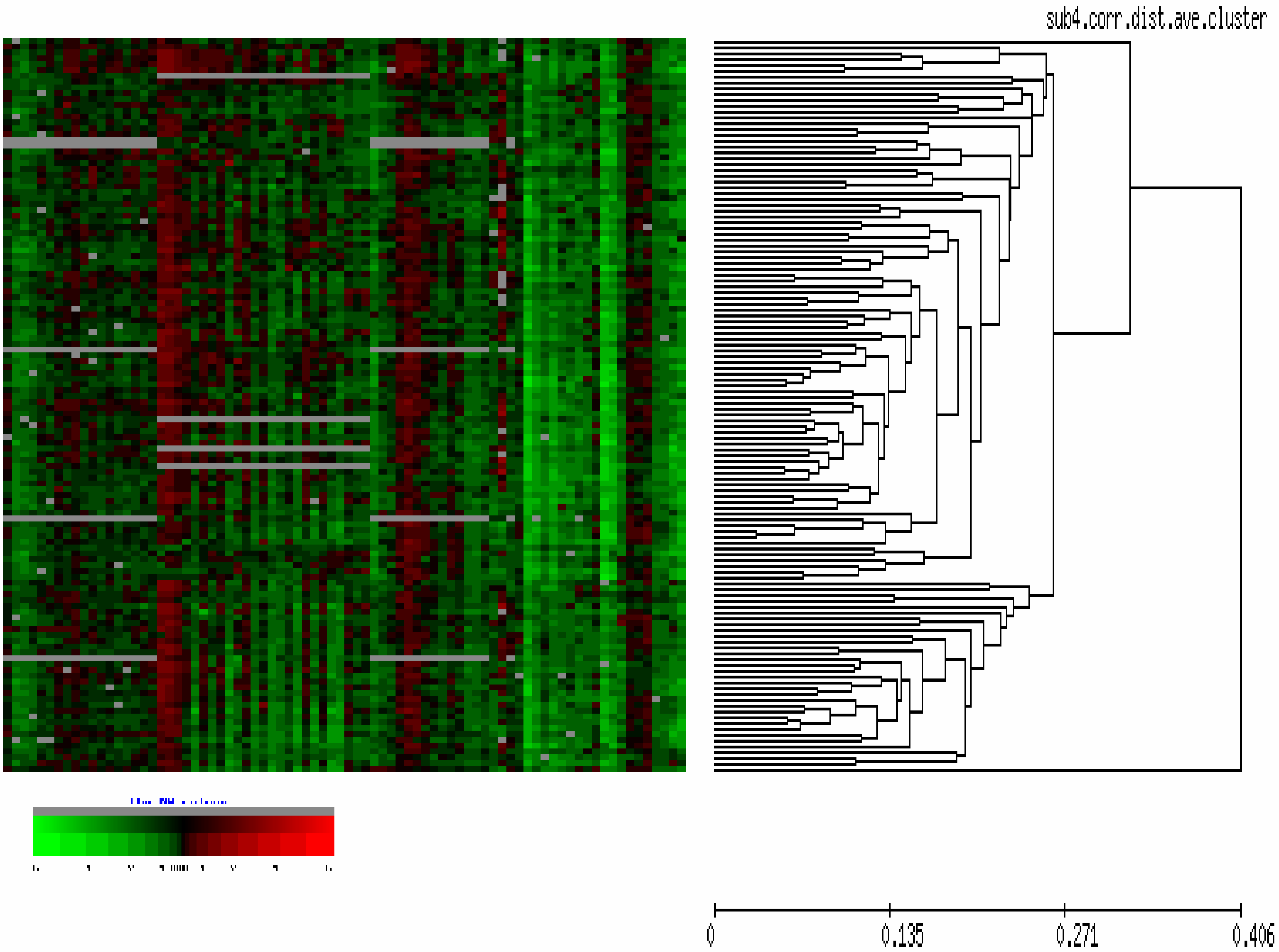


	1,2	3	4	5
1,2		**2**	4.5	5.5
3			3	3
4				**2**
5				

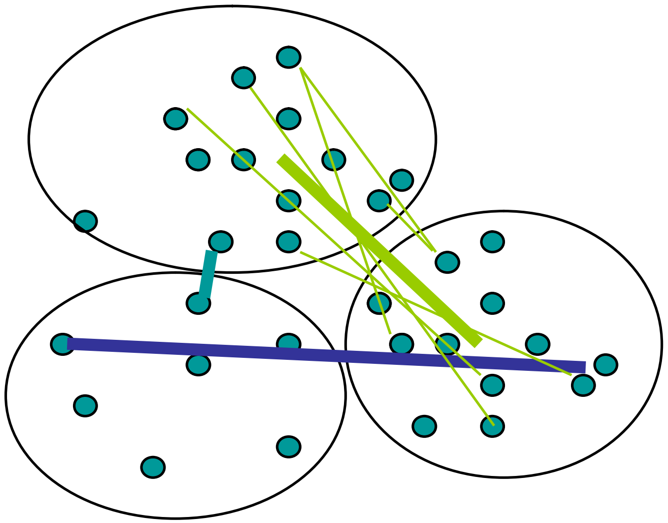


	1,2	3	4	5
1,2		**2**	4.5	5.5
3			3	3
4				**2**
5				




sub4.corr.dist.ave.cluster
0
0.135
0.271
0.406

Different linkages



Keep joining together two closest clusters by using the:

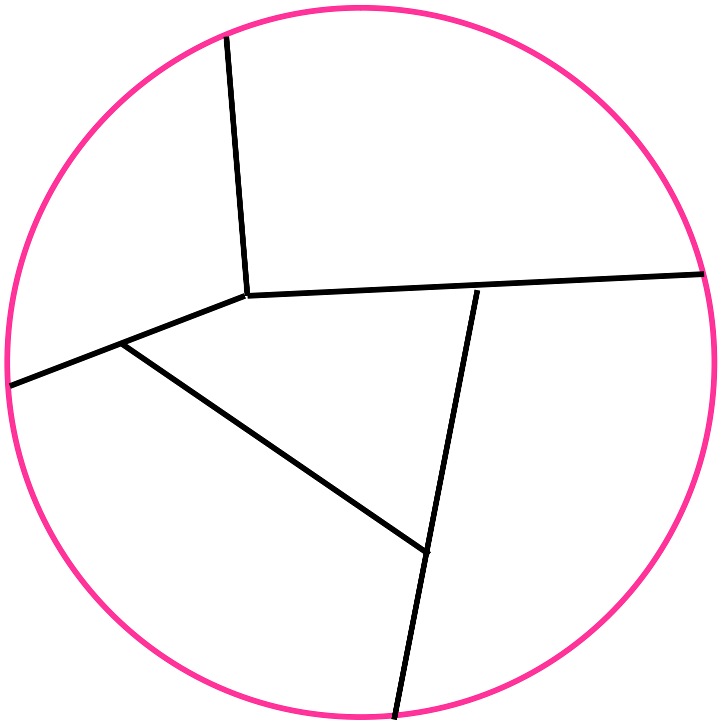
▬ Minimum distance
=> **Single linkage**

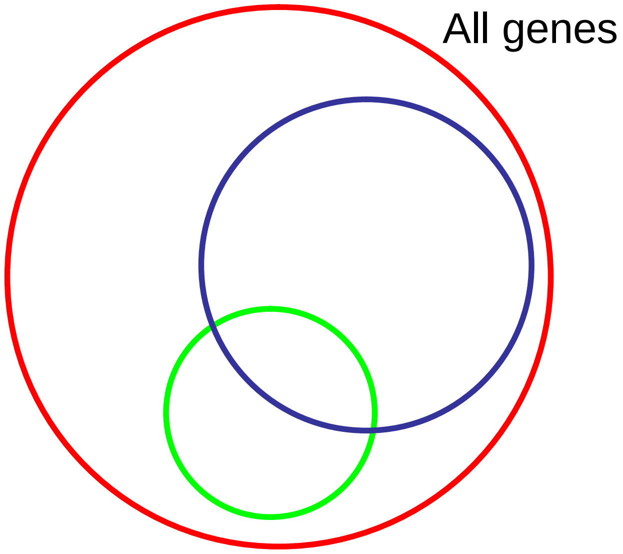
▬ Maximum distance
=> **Complete linkage**

▬ Average distance
=> **Average linkage**

Alternative – maintain a *centroid* in each cluster and use it for linking

Flat clusterings



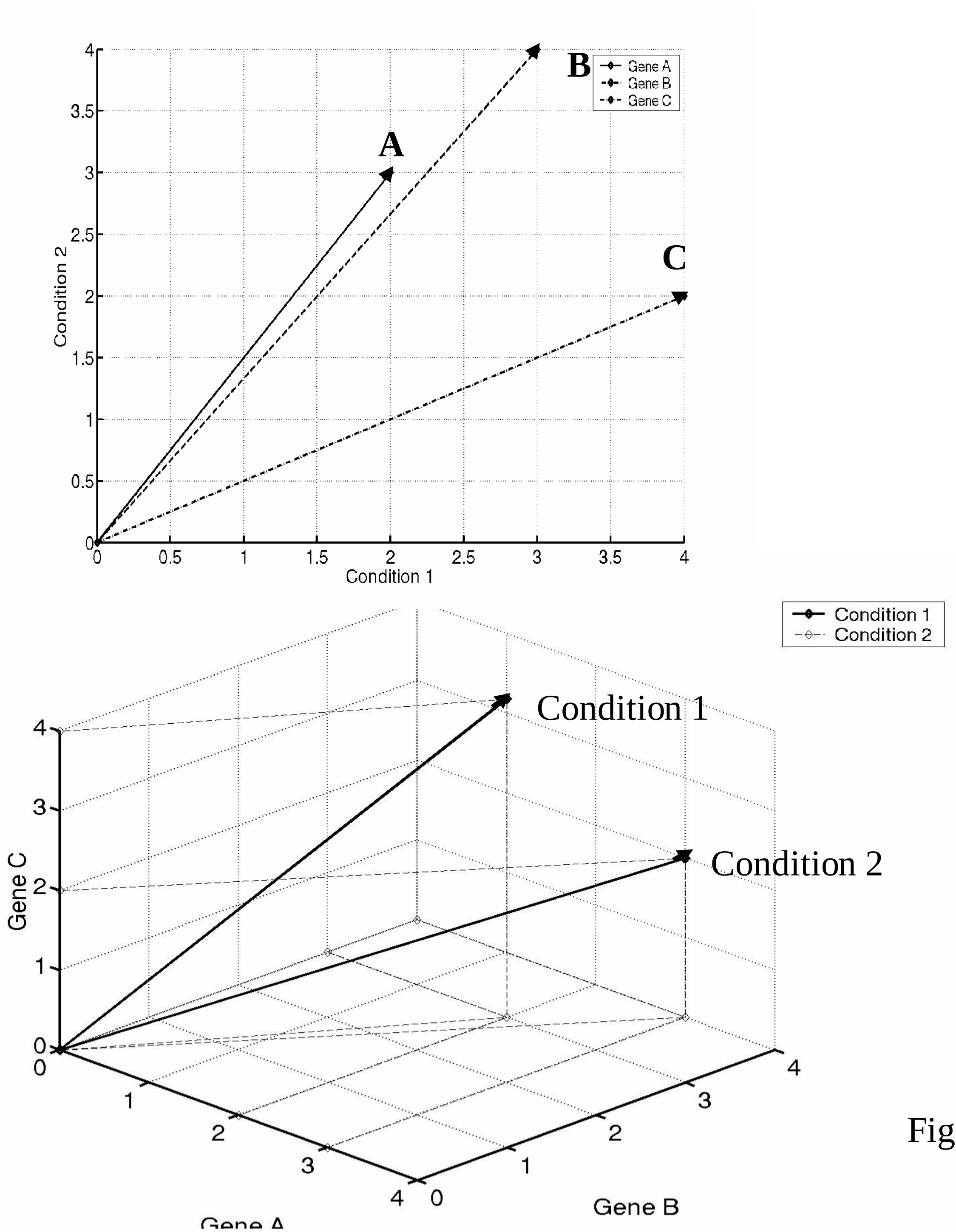


Clustering genes and smaples

- When does it make sense to cluster samples?

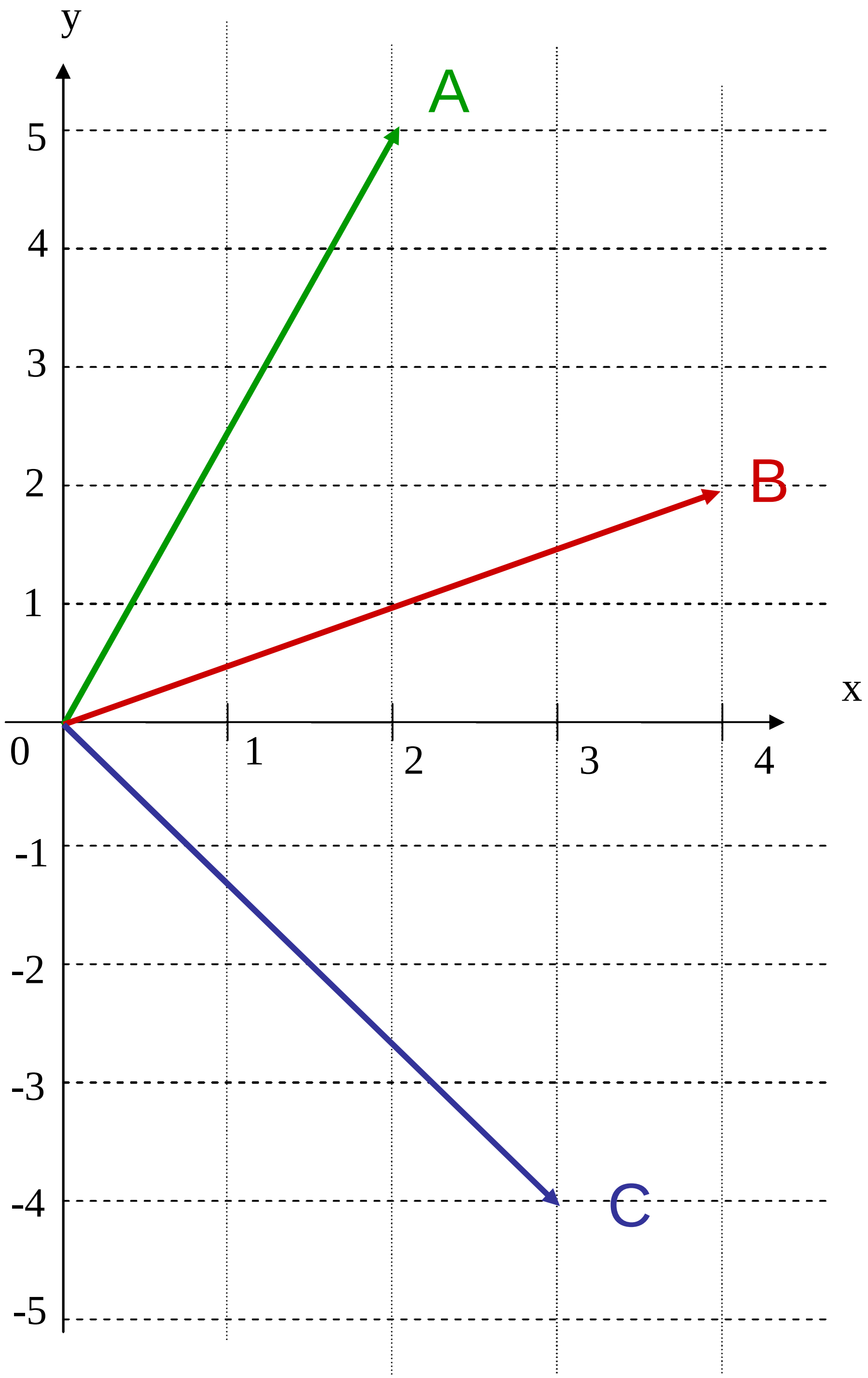
K means clutering

- K stands for number of clusters one wants to obtain – K has to be guessed
- We need a notion of a gravity center – in n dimensional Euclidean space the gravity center of vectors (each of weight 1) is defined as the vector of mean coordinates along each dimension separately



A
B
C
Gene A
Gene B
Gene C
Condition 1
Condition 2
0
0.5
1
1.5
2
2.5
3
3.5
4
Condition 1
Condition 2
Condition 1
Condition 2
Gene A
Gene B
Gene C
0
1
2
3
4

Figure 4.2



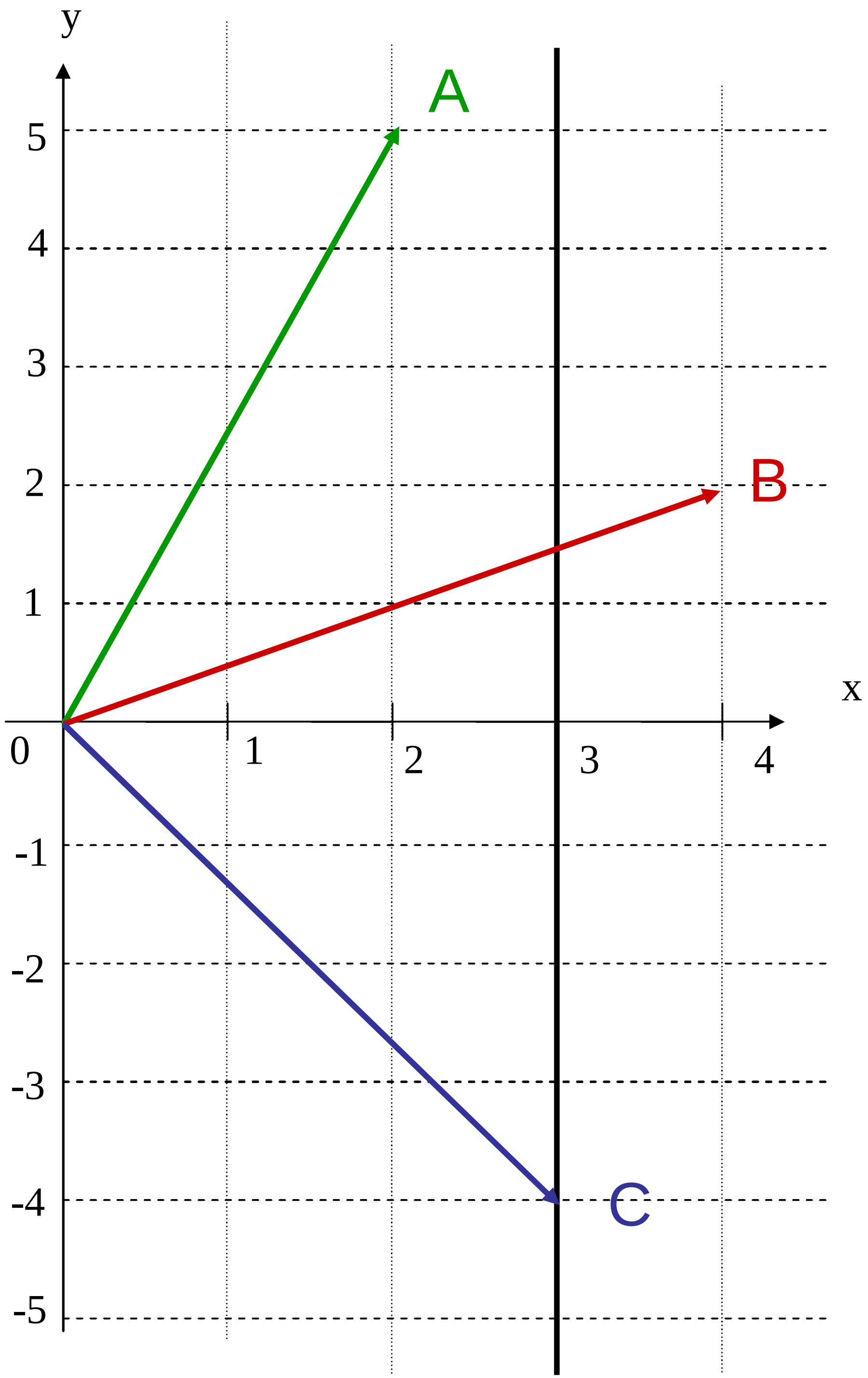
$A = (2,5)$

$B = (4,2)$

$C = (3,-3)$

$X=(2+4+3)/3=3$

$Y=(5+2-4)/3=1$

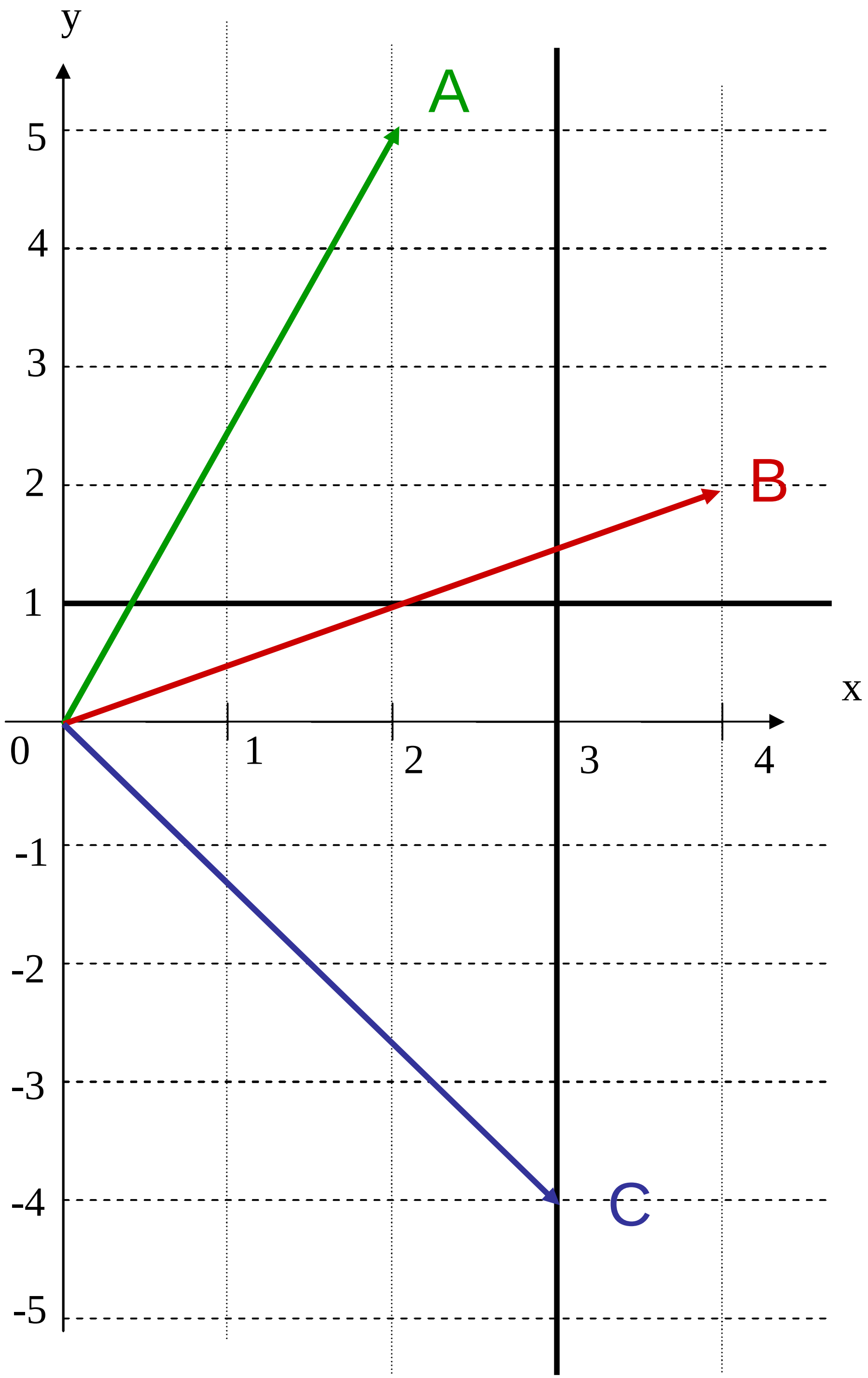


$A = (2,5)$

$B = (4,2)$

$C = (3,-3)$

$X=(2+4+3)/3=3$



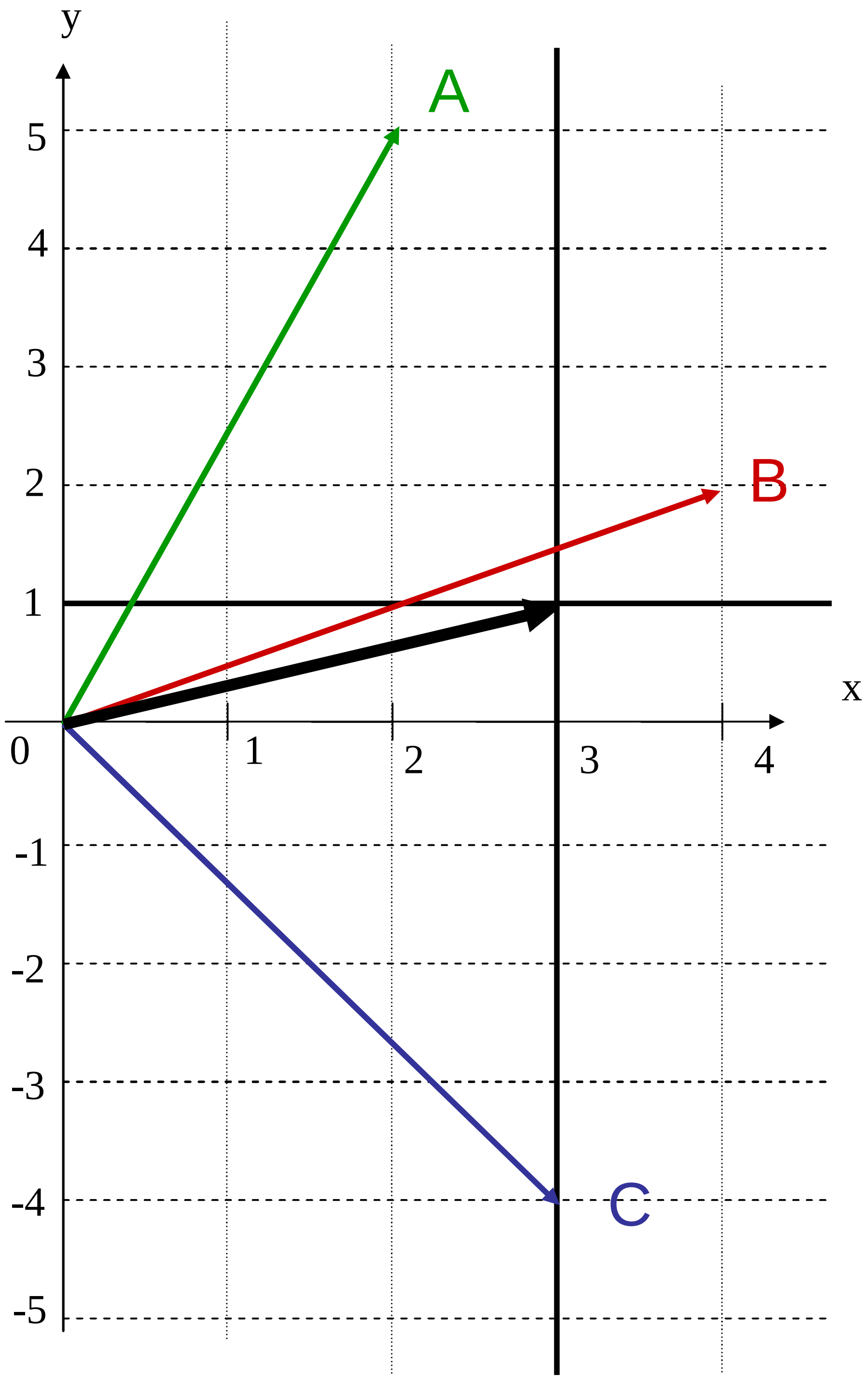
$A = (2,5)$

$B = (4,2)$

$C = (3,-3)$

$X=(2+4+3)/3=3$

$Y=(5+2-4)/3=1$



A = (2,5)

B = (4,2)

C = (3,-3)

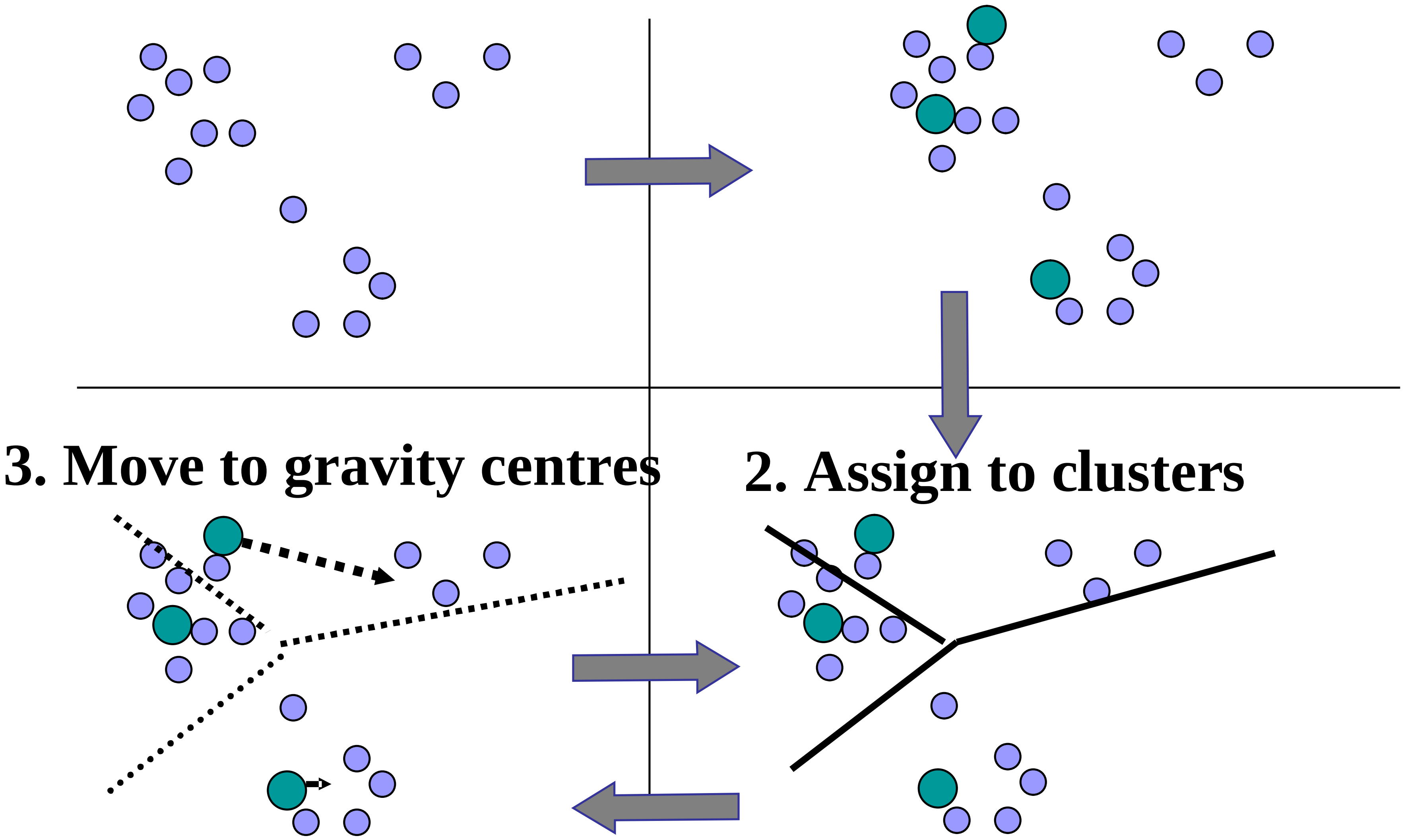
X=(2+4+3)/3=3

Y=(5+2-4)/3=1

G = (3,1)

K means clustering

1. Select K points (vectors) called centers in the space somehow (at random, or more intelligently so that they are far a way)
2. For each vector in the universe that you want to cluster, calculate the distance between it and all the K centers, and assign it to the center which is the closest - In this way K clusters are defined.
3. In each cluster define the new center as its gravity center
4. Repeat steps 2-3 until the gravity centers do not move any more, or after some fixed number of steps



1. Guess K centres

2. Assign to clusters

3. Move to gravity centres

K means clustering

1. Select K points (vectors) called centers in the space somehow (at random, or more intelligently so that they are far a way)
2. For each vector in the universe that you want to cluster, calculate the distance between it and all the K centers, and assign it to the center which is the closest - In this way K clusters are defined.
3. In each cluster define the new center as its gravity center
4. Repeat steps 2-3 until the gravity centers do not move any more, or after some fixed number of steps

Other clustering methods

- Kohonen’s self organising maps
- Self organising trees (Dopazo)
- Probability distribution based clustering
- Two way clustering
- Fuzzy clustering
- Cluster comparison