15-440 Distributed Systems K-Means

Zeinab Khalifa

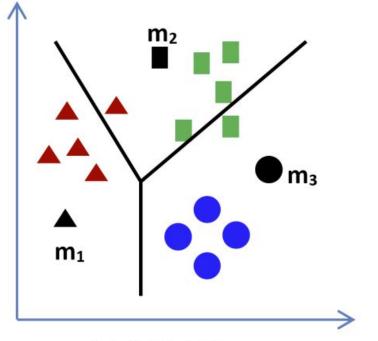


K-Means

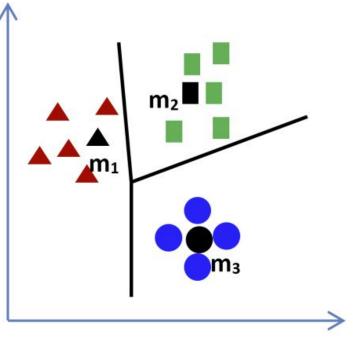
- Clustering algorithm
- an iterative algorithm that attempts to find *K* similar groups in a given data set via minimizing a mean squared distance function
- Applications:
- Data mining
- Statistical data analysis: machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.

Carnegie Mellon University Oataı

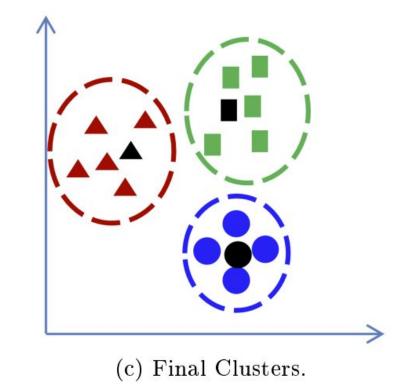
- Visualization



(a) Initial Means.



(b) Recalculated Means.





Explained in plain English, **k-Means** roughly follows this approach:

- 1. We start by deciding how many clusters we would like to form from our data. We call this value k. The value of k is generally a small integer, such as 2, 3, 4, or 5, but may be larger.
- 2. Next, we select k points to be the centroids of k clusters which at present have no members. The list of centroids can be selected by any method (e.g., randomly from the set of data points). It is usually better to pick centroids that are far apart.
- 3. We then compute the *Euclidean distance* (the similarity function with a data set of data points) from each data point to each centroid. A data point is assigned to a cluster such that its distance to that cluster is the smallest among all other distances.
- 4. After associating every data point with one of k clusters, each centroid is recalculated so as to reflect the true mean of its constituent data points.
- 5. Steps 3. and 4. are repeated for a number of times (say, μ); essentially until the centroids start varying very little.

ة کارنىدى مىلون فى قطر **Carnegie Mellon University Qatar**

K-Means

You need to define similarities and recalculate the centroids

- What is the similarity between two data points?
- What is the similarity between two DNA strands?
- How to recalculate the data points centroids?
- How to recalculate the DNA centroids?



Sequential Kmeans

جامعه کارنی جی میلوں فی قطر **Carnegie Mellon University Qatar**

P1
P2
Р3
P4
P5
P6
Ρ7
P8
P9
P10
P11
P12
P13
P14
P15
P16

جامعة کارنيجي ميلود في قطر Carnegie Mellon University Qatar

P1
P2
Р3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

CO	P1	
C1	P6	
C2	P3	
C3	P9	

جامعة کارنيجي ميلود في قطر Carnegie Mellon University Qatar

Ρ1

P6

Ρ3

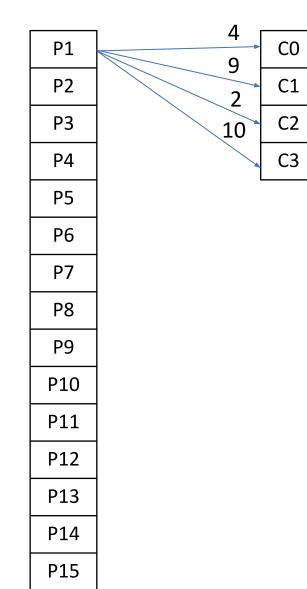
P9

C0

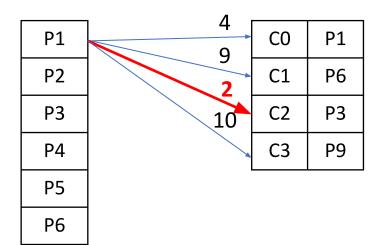
C1

C2

C3



جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar



Ρ7

P8

Ρ9

P10

P11

P12

P13

P14

P15

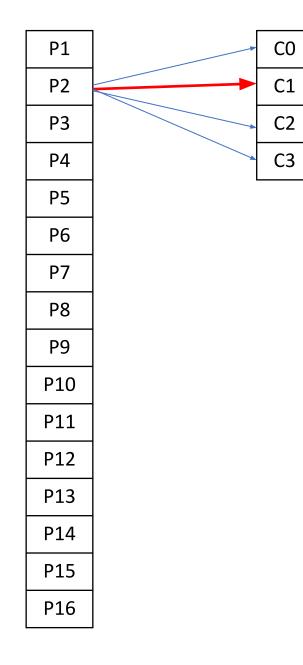
C0 C1 C2 Ρ1 C3

جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

Ρ1

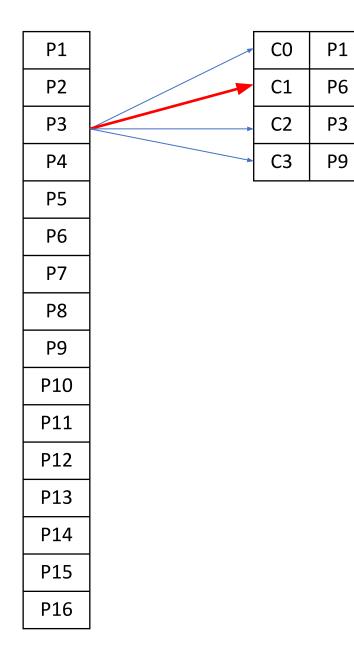
P6

Ρ3



C0	
C1	P2
C2	P1
C3	

جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar



C0	
C1	P2 + P3
C2	P1
C3	

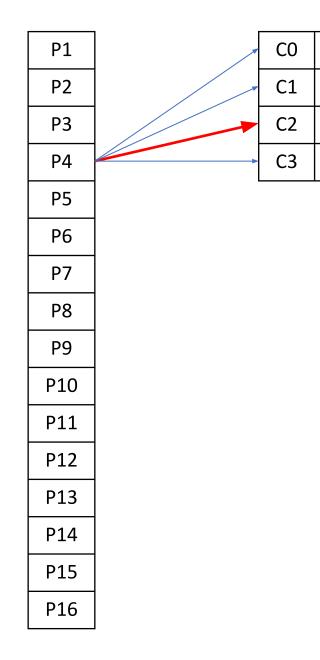
* P1 + P2 = (x1,y1) + (x2,y2) = (x1+x2, y1+y2)

جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

Ρ1

P6

Ρ3



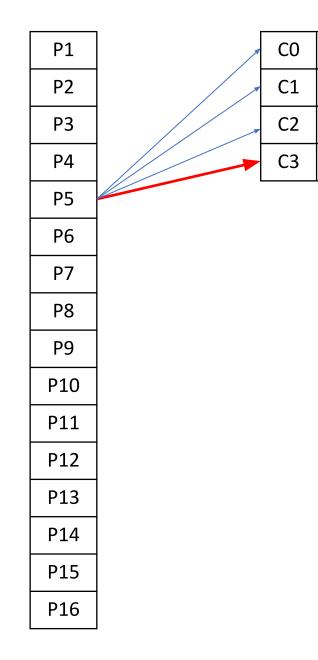
C0	
C1	P2 + P3
C2	P1 + P4
C3	

جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

Ρ1

P6

Ρ3



C0	
C1	P2 + P3
C2	P1 + P4
C3	P5

جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15

C0	P1	
C1	P6	
C2	P3	
C3	P9	

C0	P6 + P8 + P10 + P13
C1	P2 + P3 + P7 + P11
C2	P1 + P4 + P12 + P15 + P16
C3	P5 + P9 + P14

جامعة کارنیجی میلود فی قطر Carnegie Mellon University Qatar

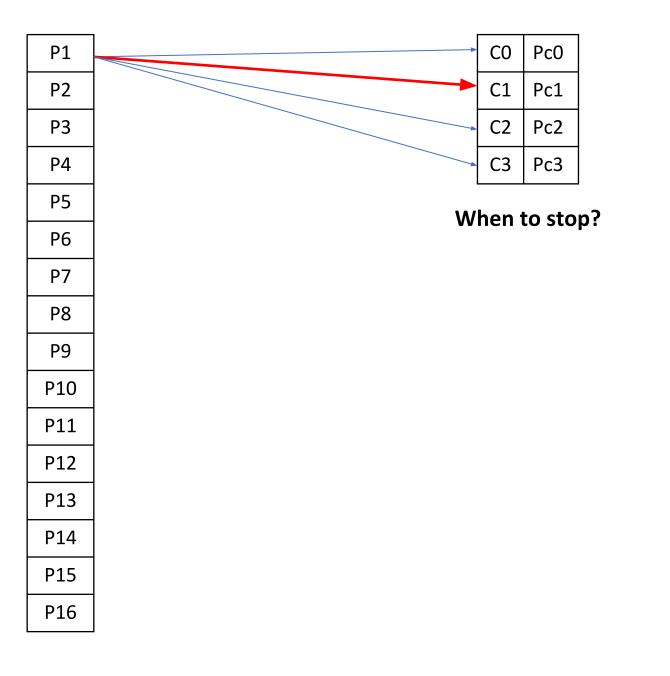
Centroids after iteration 1

P1
P6
P3
P9

C0	(P6 + P8 + P10 + P13)/4
C1	(P2 + P3 + P7 + P11)/4
C2	(P1 + P4 + P12 + P15 + P16) /5
C3	(P5 + P9 + P14)/3

* *P/N* = (*x/N*,*y/N*)

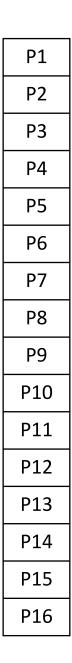
جامعة کارنیجی میلود فی قطر Carnegie Mellon University Qatar



جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

Parallel K-Means





جامعة کارنیجی میلود فی قطر Carnegie Mellon University Qatar

P1
P2
Р3
P4
Р5
P6
P7
P8
Р9
P10
P11
P12
P13
P14
P15
P16

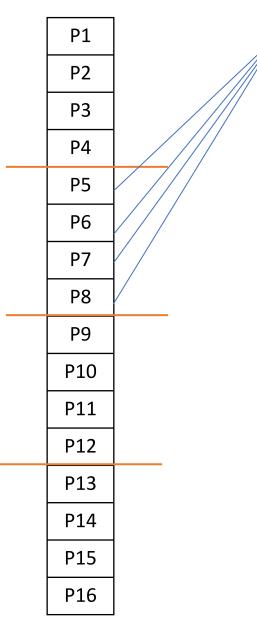
_			
	C0	P1	
	C1	P6	
	C2	Р3	
	C3	P9	

جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

Ρ1 Ρ2 Ρ3 Ρ4 Ρ5 P6 Ρ7 P8 Ρ9 P10 P11 P12 P13 P14 P15 P16

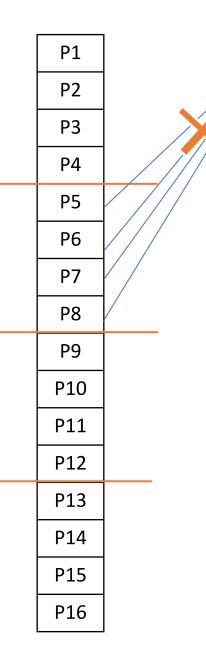
_			_
	C0	P1	
	C1	P6	
	C2	Р3	
	C3	P9	

جامعة کارنیجی میلون فی قطر Carnegie Mellon University Qatar



C0 Ρ1 C1 P6 Ρ3 C2 C3 Ρ9

جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar



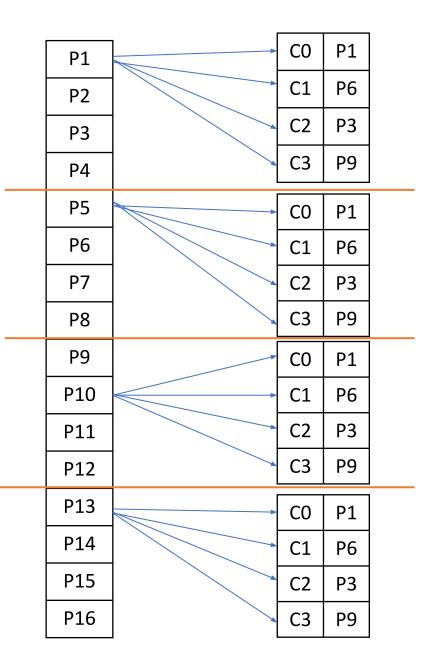
CO	P1	
C1	P6	
C2	P3	
C3	P9	

No memory sharing

جامعۃ کارنیجی میلوں فی قطر Carnegie Mellon University Qatar

	r		
P1		C0	P1
P2		C1	P6
P3		C2	Ρ3
P4		C3	P9
P5		C0	P1
P6		C1	P6
P7		C2	P3
P8		C3	Р9
P9		C0	P1
P9 P10		C0 C1	P1 P6
P10		C1	P6
P10 P11		C1 C2	Р6 Р3
P10 P11 P12		C1 C2 C3	P6 P3 P9
P10 P11 P12 P13		C1 C2 C3 C0	P6 P3 P9 P1
_	P2 P3 P4 P5 P6 P7 P8	P2 P3 P4 P5 P6 P7 P8	P1 C1 P2 C1 P3 C2 P4 C3 P5 C0 P6 C1 P7 C2 P8 C3

جامعة کارنیجی میلود فی قطر Carnegie Mellon University Qatar



جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

				r			
P1		C0	P1		C0	P2 + P3	
P2		C1	P6		C1	0	
P3	_	C2	P3		C2	P1	
P4		С3	P9		C3	P4	
P5		C0	P1		C0	P5	
P6		C1	P6		C1	P7	
P7		C2	P3		C2	P8	
P8	_	C3	P9		C3	P6	
P9		C0	P1		C0	0	
P10		C1	P6		C1	P12	
P11		C2	P3		C2	P10 + P11	
P12		С3	P9		C3	Р9	
P13		C0	P1		C0	P13 + P14 + P16	
P14		C1	P6		C1	0	
							1
P15		C2	P3		C2	0	
	_	C2 C3	P3 P9		C2 C3	0 P15	جامہۃ کارنی جے میلوں فی قطر

						
[P1	C0	P1	C0	P2 + P3 /2	
ŀ	P2	C1	P6	C1	0	
	P3	C2	P3	C2	P1/1	
-	P4	C3	P9	C3	P4/1	
	P5	СО	P1	СО	P5	
-	P6	C1	P6	C1	P7	
ľ	P7	C2	P3	C2	P8	
	P8	C3	P9	C3	P6	
Ī				1		
	P9	CO	P1	CO	0	
-	P9 P10	C0 C1	P1 P6	C0 C1	0 P12	
-						
-	P10	C1	P6	C1	P12	
-	P10 P11	C1 C2	P6 P3	C1 C2	P12 P10 + P11	
-	P10 P11 P12	C1 C2 C3	P6 P3 P9	C1 C2 C3	P12 P10 + P11 P9	
-	P10 P11 P12 P13	C1 C2 C3 C0	P6 P3 P9 P1	C1 C2 C3 C0	P12 P10 + P11 P9 P13 + P14 + P16	
-	P10 P11 P12 P13 P14	C1 C2 C3 C0 C1	P6 P3 P9 P1 P6	C1 C2 C3 C0 C1	P12 P10 + P11 P9 P13 + P14 + P16 O	جامعـ ۃ کارنیـ جـ ی میلوں فی قـطر

				ſ			
P1		C0	P1		C0	P2 + P3	
P2		C1	P6		C1	0	
P3	_	C2	P3		C2	P1	
P4		С3	P9		C3	Р4	
P5		C0	P1		C0	P5	
P6		C1	P6		C1	P7	
P7		C2	P3		C2	P8	
P8	_	C3	P9		C3	P6	
P9		C0	P1		C0	0	
P10		C1	P6		C1	P12	
P11		C2	P3		C2	P10 + P11	
P12		С3	P9		C3	Р9	
P13		C0	P1		C0	P13 + P14 + P16	
P14		C1	P6		C1	0	
							1
P15		C2	P3		C2	0	
	_	C2 C3	P3 P9		C2 C3	0 P15	جامہۃ کارنی جے میلوں فی قطر

					 						-
P1	СО	P1	СО	P2 + P3	C0	Р5		0	0	C0	P13 + P14 + P16
P2	C1	P6	C1	0	C1	Р7	(21	P12	C1	0
P3	C2	P3	C2	P1	C2	P8		22	P10 + P11	C2	0
P4	C3	P9	C3	P4	C3	P6		23	P9	C3	P15
		1]	•							
P5	CO	P1									
P6	C1	P6									
P7	C2	P3									
P8	C3	P9									
P9	СО	P1									
P10	C1	P6									
P11	C2	P3									
P12	C3	P9									
P13	СО	P1]								
P14	C1	P6									
P15	C2	P3									
P16	C3	P9							م ا م ا		
			-				1	Car	ی میلود فی ق negie Mellon	Unive	ersity Qatar

	P1	C0	P1	C0	P2 + P3 + P5 + P13 + P14 + P16	/6
	P2	C1	P6	C1	P7 + P12	/2
	P3	C2	P3	C2	P8 + P10 + P11	/3
	P4	C3	P9	C3	P6 + P9 + P15	/3
	P5	C0	P1			
	P6	C1	P6			
	P7	C2	P3			
	P8	C3	P9			
	P9	C0	P1			
	P10	C1	P6			
	P11	C2	P3			
	P12	C3	P9			
	P13	C0	P1			
	P14	C1	P6			
	P15	C2	Р3			
	P16	C3	P9			جامعح کارنیجی میلوں فی قطر
•						بالمحط كارتيجين ميتون فكر

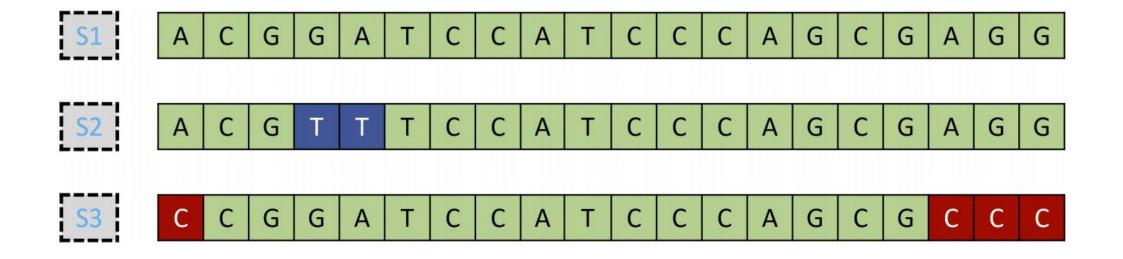
Carnegie Mellon University Qatar

			1		
	P1) P	°c0	
	P2	C1	. P	°c1	
	P3	C2	: P	°c2	
,	P4	C3	P	°c3	
	P5				
,	P6				
	P7				
	P8				
	P9				
	P10				
	P11				
	P12				
	P13				
	P14				
	P15				
	P16				جامہۃ کارنیجی میلوں فی قطر

P1	СО	Pc0	
P2	C1	Pc1	
P3	C2	Pc2	
P4	C3	Pc3	
P5	CO	Pc0	
P6	C1	Pc1	
P7	C2	Pc2	
P8	C3	Pc3	
P9	CO	Pc0	
P10	C1	Pc1	
P11	C2	Pc2	
P12	C3	Pc3	
P13	CO	Pc0	
P14	C1	Pc1	
P15	C2	Pc2	
P16	C3	Pc3	
			جامعہ کارنیجی میلوں فی قطر Carnegie Mellon University Qatar

DNA stranding

جامعه کارنیجی میلود فی قطر Carnegie Mellon University Qatar



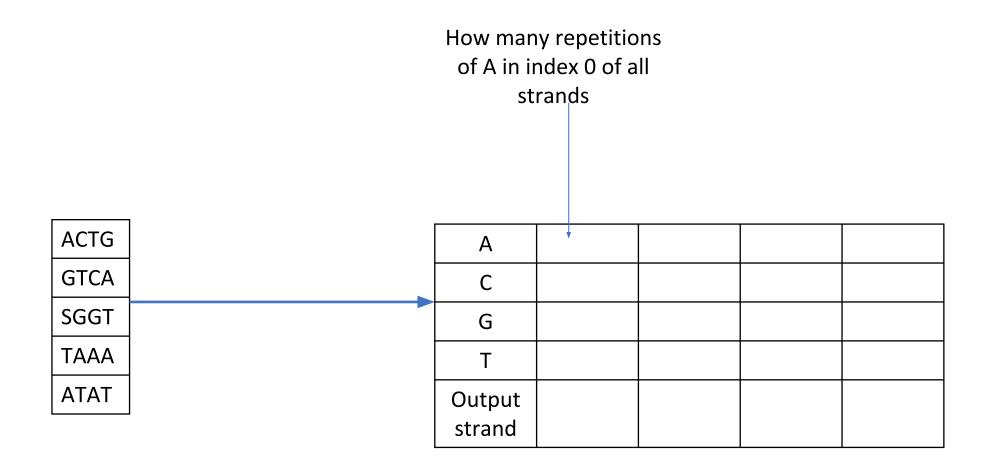
جامعہ کارنیجی ہیلوں فی قطر Carnegie Mellon University Qatar

ACTG
GTCA
SGGT
ΤΑΑΑ
ATAT

جامعة کارنیجی میلود فی قطر Carnegie Mellon University Qatar

	How to get the
ACTG	centroid of these DNA
GTCA	strands?
SGGT	
TAAA	
ATAT	





جامعة کارنيجي ميلون في قطر Carnegie Mellon University Qatar

ACTG
GTCA
SGGT
ΤΑΑΑ
ATAT

А	2		
С			
G			
Т			
Output strand			

جامعة کارنيجي ميلور في قطر Carnegie Mellon University Qatar

ACTG
GTCA
SGGT
ΤΑΑΑ
ATAT

А	2	1	2	1
C	0	1	1	0
G	1	1	1	1
Т	1	2	1	2
Output strand				

جامعة کارنيجي ميلور في قطر Carnegie Mellon University Qatar

ACTG GTCA SGGT TAAA ATAT

А	2	1	2	1
C	0	1	1	0
G	1	1	1	1
Т	1	2	1	2
Output strand				

Get the mean or the median (sort the values and select the middle one)



ACTG
GTCA
SGGT
ΤΑΑΑ
ATAT

А	2	1	2	1
C	0	1	1	0
G	1	1	1	1
Т	1	2	1	2
Output strand	т	G	С	А

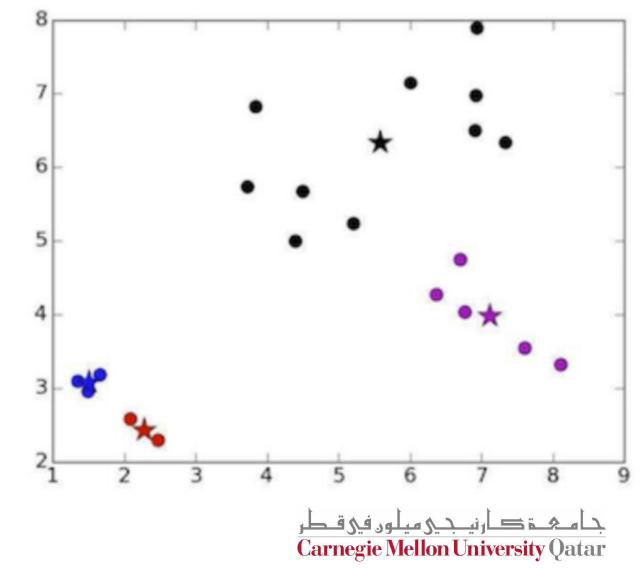
جامعة کارنيجي ميلور في قطر Carnegie Mellon University Qatar

ACTG
GTCA
SGGT
ΤΑΑΑ
ATAT

А	2	1	2	1
C	0	1	1	0
G	1	1	1	1
Т	1	2	1	2
Output strand	Т	G	С	А

جامعة کارنيجي ميلور في قطر Carnegie Mellon University Qatar

Bad Clustering



Bad Clustering

The blue and red stars are called unlucky centroids (*)

A poor choice of the initial centroids will take longer to converge or may result in bad clustering. You can handle this in:

- Your data generators (generate first k points to be far apart and pick them in your implementation)
- 2. Try different sets of random centroids, and choose the best set.

