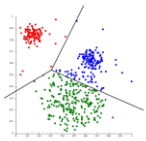
15-440 Distributed Systems Kmeans

March 24, 2022

Laila Elbeheiry



K-Means at a High Level

- Clustering Algorithm

> **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters).
> - Wikipedia

- Applications:
  - Data mining
  - Statistical data analysis: machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics
  - Visualization
  - Detecting anomalies or outliers

What is "k"?

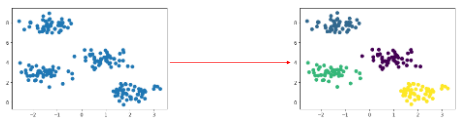


Image source: https://towardsdatascience.com/k-means-clustering-explained-4528df86a120

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

K-Means Explained

K-means is an iterative process that works by executing the following steps:

1. Select centroids (center of cluster) for each of the k clusters. 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.
2. Calculate the distance of all data points to the centroids.
3. Assign data points to the closest cluster.
4. Find the new centroids of each cluster by taking the mean of all data points in the cluster.
5. Repeat steps 2,3 and 4 until all points converge and cluster centers stop moving.

Let's see how it works!

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

K-Means in Project 3

You need to define a distance function and a mean function.

- How to calculate the distance between points in a 2D plane?
- How to calculate the distance between DNA strands?
- How to find the mean of points in a 2D plane?
- How to find the mean of DNA strands?

Sequential Kmeans

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Initial centroids/means

C0	P1
C1	P6
C2	P3
C3	P9

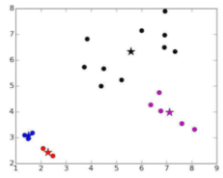
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9



The blue and red starts 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:

1. 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.

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Initial centroids/means

P1	4	C0 P1
P2	9	C1 P6
P3	2	C2 P3
P4	10	C3 P9
P5		
P6		
P7		
P8		
P9		
P10		
P11		
P12		
P13		
P14		
P15		
P16		

C0	
C1	
C2	
C3	

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

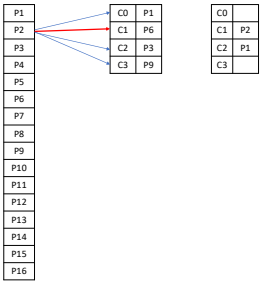
Initial centroids/means

P1	4	C0 P1
P2	9	C1 P6
P3	2	C2 P3
P4	10	C3 P9
P5		
P6		
P7		
P8		
P9		
P10		
P11		
P12		
P13		
P14		
P15		
P16		

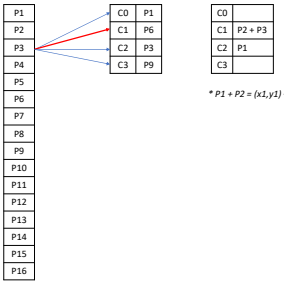
C0	
C1	
C2	P1
C3	

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

Initial centroids/means

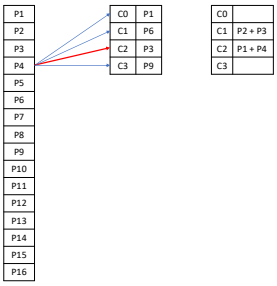


Initial centroids/means

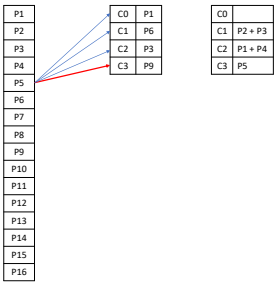


** P1 + P2 = (x1,y1) + (x2,y2) = (x1+x2, y1+y2)*

Initial centroids/means



Initial centroids/means



Initial centroids/means

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

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

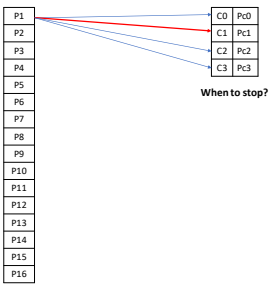
Centroids after iteration 1

P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	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

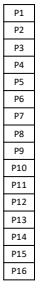
** P/N = (x/N,y/N)*



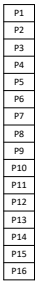
When to stop?

Parallel K-Means

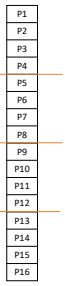
How can we parallelize?



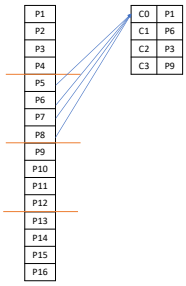
How can we parallelize?



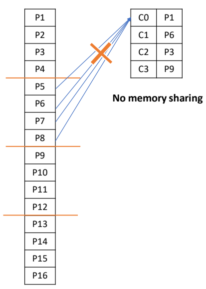
How can we parallelize?



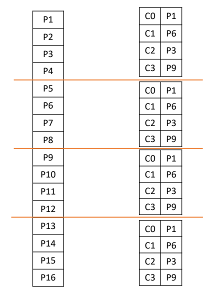
How can we parallelize?



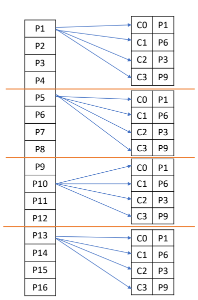
How can we parallelize?



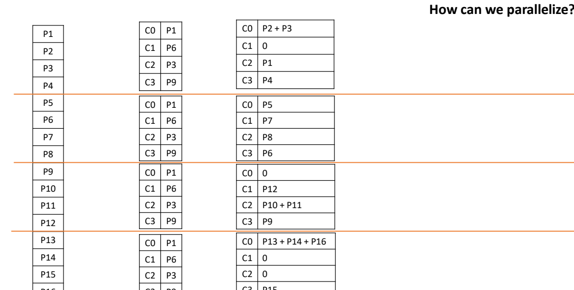
How can we parallelize?

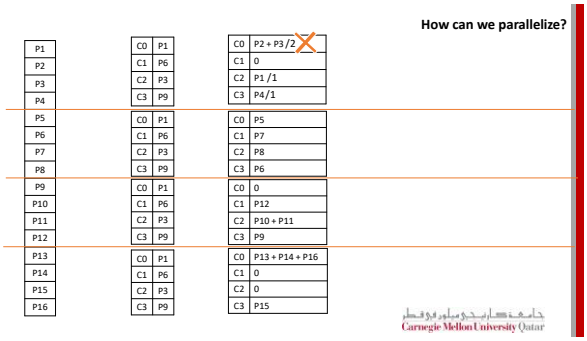


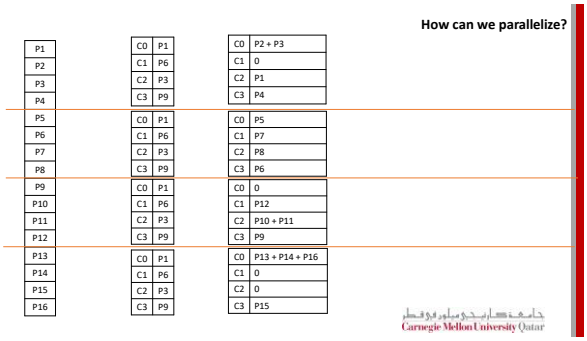
How can we parallelize?

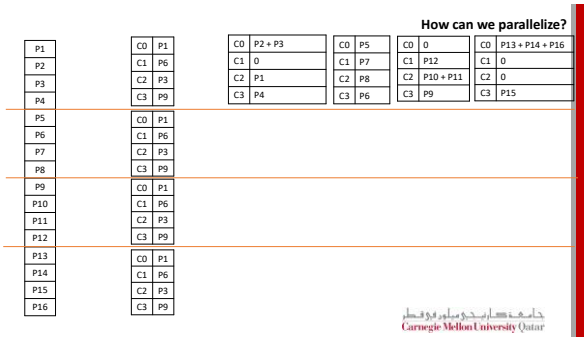


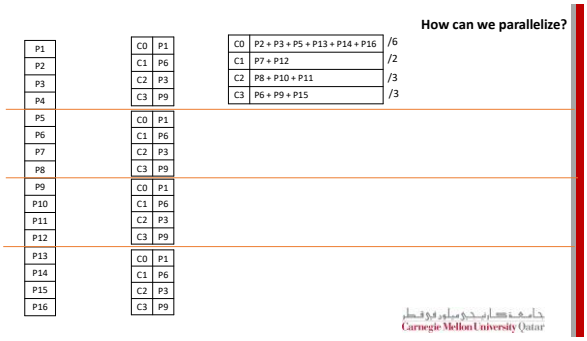
How can we parallelize?



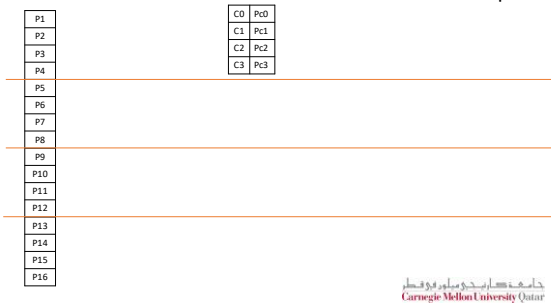

How can we parallelize?
P1
P2
P3
P4
C0 P1
C1 P6
C2 P3
C3 P9
C0 P2 + P3/2
C1 0
C2 P1 /1
C3 P4/1
P5
P6
P7
P8
C0 P1
C1 P6
C2 P3
C3 P9
C0 P5
C1 P7
C2 P8
C3 P6
P9
P10
P11
P12
C0 P1
C1 P6
C2 P3
C3 P9
C0 0
C1 P12
C2 P10 + P11
C3 P9
P13
P14
P15
P16
C0 P1
C1 P6
C2 P3
C3 P9
C0 P13 + P14 + P16
C1 0
C2 0
C3 P15
جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar


How can we parallelize?
P1
P2
P3
P4
C0 P1
C1 P6
C2 P3
C3 P9
C0 P2 + P3
C1 0
C2 P1
C3 P4
P5
P6
P7
P8
C0 P1
C1 P6
C2 P3
C3 P9
C0 P5
C1 P7
C2 P8
C3 P6
P9
P10
P11
P12
C0 P1
C1 P6
C2 P3
C3 P9
C0 0
C1 P12
C2 P10 + P11
C3 P9
P13
P14
P15
P16
C0 P1
C1 P6
C2 P3
C3 P9
C0 P13 + P14 + P16
C1 0
C2 0
C3 P15
جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

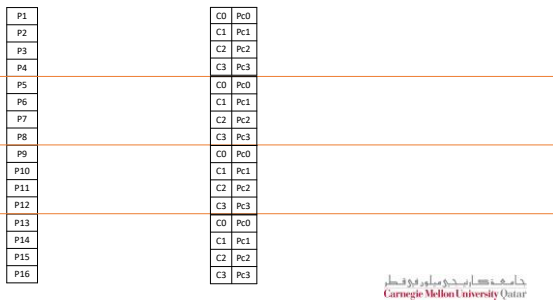

How can we parallelize?
P1
P2
P3
P4
C0 P1
C1 P6
C2 P3
C3 P9
C0 P2 + P3
C1 0
C2 P1
C3 P4
C0 P5
C1 P7
C2 P8
C3 P6
C0 0
C1 P12
C2 P10 + P11
C3 P9
C0 P13 + P14 + P16
C1 0
C2 0
C3 P15
P5
P6
P7
P8
C0 P1
C1 P6
C2 P3
C3 P9
P9
P10
P11
P12
C0 P1
C1 P6
C2 P3
C3 P9
P13
P14
P15
P16
C0 P1
C1 P6
C2 P3
C3 P9
جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar


How can we parallelize?
P1
P2
P3
P4
C0 P1
C1 P6
C2 P3
C3 P9
C0 P2 + P3 + P5 + P13 + P14 + P16
C1 P7 + P12
C2 P8 + P10 + P11
C3 P6 + P9 + P15
/6
/2
/3
/3
P5
P6
P7
P8
C0 P1
C1 P6
C2 P3
C3 P9
P9
P10
P11
P12
C0 P1
C1 P6
C2 P3
C3 P9
P13
P14
P15
P16
C0 P1
C1 P6
C2 P3
C3 P9
جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

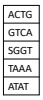
How can we parallelize?



How can we parallelize?



DNA stranding



ACTG
GTCA
SGGT
TAAA
ATAT

How to get the centroid of these DNA strands?

How many repetitions of A in index 0 of all strands

ACTG
GTCA
SGGT
TAAA
ATAT

A				
C				
G				
T				
Output strand				

ACTG
GTCA
SGGT
TAAA
ATAT

A	2			
C				
G				
T				
Output strand				

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand				

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand				

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

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand	**T**	**G**	**C**	**A**

ACTG
GTCA
SGGT
TAAA
ATAT

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2
Output strand	**T**	**G**	**C**	**A**