

15-440
Distributed Systems
Recitation 10:
Kmeans

Slides By: Hend Gedawy
& Previous TAs

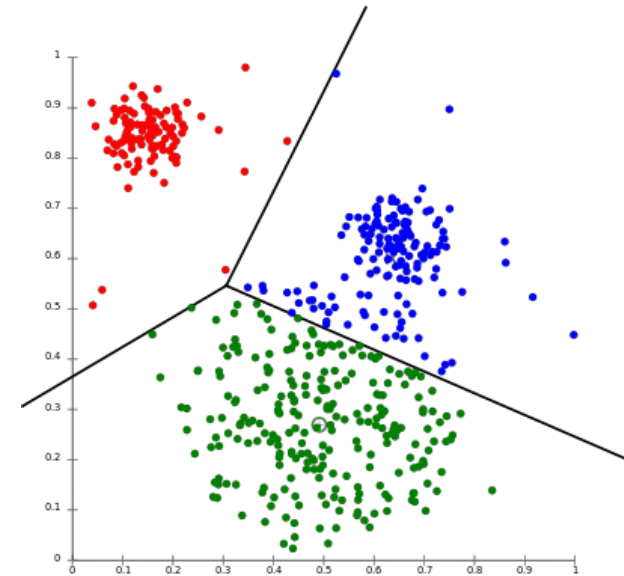


Announcements

- **Quiz II** Next Sunday
- **PS4 Out** Due Nov. 12
- **P3** Due Nov. 16

Outline

- Project 3 Overview & Kmeans Algorithm
- Kmeans – 2D Points
 - Sequential
 - Parallel
- Kmeans – DNA Strands



Project 3 Overview

- **Objective:** apply Message Passing Interface (**MPI**), a library standard for writing message passing programs, to a popular real problem, namely **cluster analysis using the k-Means algorithm**.
- **Apply k-Means clustering to two different applications (datasets);** data points in a 2D plane and DNA strands in biology.
 - You will provide **sequential and parallel implementations of the K-Means algorithm**
 - dataset as input and K centroids as output.
 - Specifically, you **deliverables** are:
 - **Data Generator code** for DNA strands
 - Note: Data Generator for 2D points is given
 - **Sequential clustering** implementation for both data types
 - **Parallel implementation** for both data types
- You will also conduct and analyze some **scalability studies** on various **degrees of parallelism and data set sizes**.

Clustering

- 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

- Clustering is useful for statistical data analysis and machine learning to discover hidden patterns, data structures, relationship between data and also detect anomalies or outliers

What's the difference between clustering and classification?

Clustering Application Examples



Classify various customers according to their interests which helps with targeted marketing.



Identify which message is spam and which is not.... using the sender address, key terms inside the message and other factors



Classify documents and content according to their categories and search terms

K-Means Objective



Maximize intra-cluster similarity and minimize inter-cluster similarity.

What is “k”?

K-Means Algorithm Explained

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

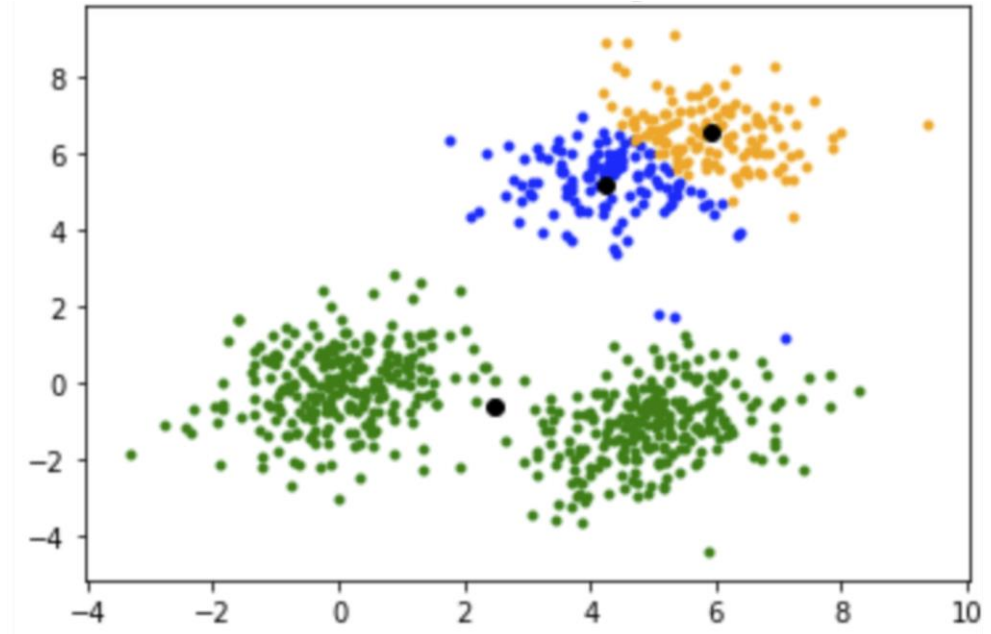
1. Select centroids (center of cluster) for each of the k clusters.
2. Calculate the distance of all data points to the centroids.
3. Assign data points to the closest cluster.
4. Find the new centroids of each cluster by taking the mean of all data points in the cluster.
5. Repeat steps 2,3 and 4 until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

[Let's see how it works!](#)

Choosing K & Initial Centroids

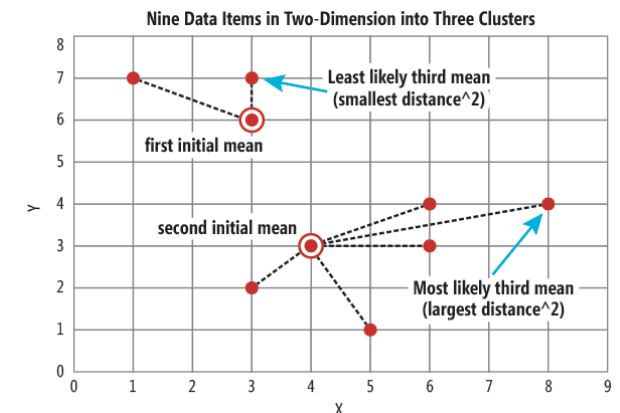
In practice, people often try different values of k and see how their results vary.

A poor choice of the **initial centroids** will take longer to converge or may result in bad clustering.



Choosing Initial Centroids – Example Approaches

- **Random:** pick them randomly from among your data points.
 - Not efficient: it is likely that many of the initial centroids end up in the same cluster.
- **"farthest" heuristic:** initialize the first centroid randomly, then to initialize the j th centroid to the point whose minimum distance to the preceding centroids is largest.
 - centroids are well spread-out from each other. [Let's try it out!](#)
- **k-means++:** works similar to the "farthest" heuristic, by choosing first centroid randomly, but ...
 - Choose the j th centroid to be the point with probability proportional to the square of its distance to the nearest preceding centroid.
 - Instead of getting points that are at the edges of their true clusters (as in the farthest heuristic approach), you're more likely to get ones near the center of their true clusters.



Credit:

<https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>

<https://learn.microsoft.com/en-us/archive/msdn-magazine/2015/august/test-run-k-means-data-clustering>

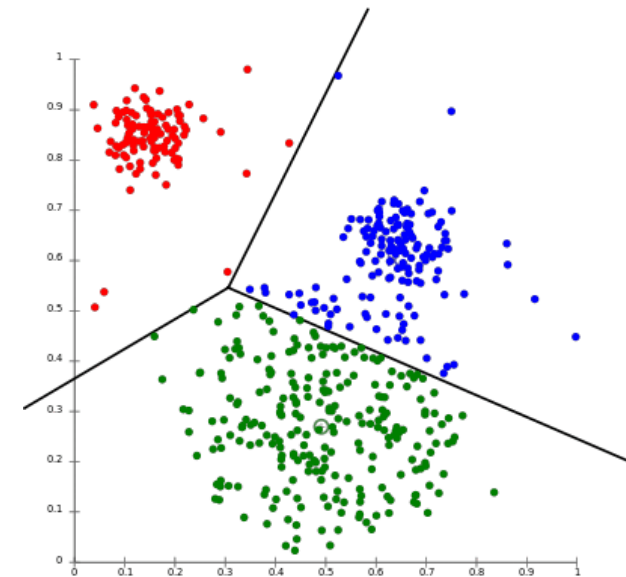
K-Means in Project 3

The main functions you will need are: distance and mean.

- 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?

Outline

- Project 3 Overview & Kmeans Algorithm
- Kmeans – 2D Points
 - Sequential
 - Parallel
- Kmeans – DNA Strands



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

Sequential K-Means: (1) Choosing Initial Centroids

Initial centroids/means

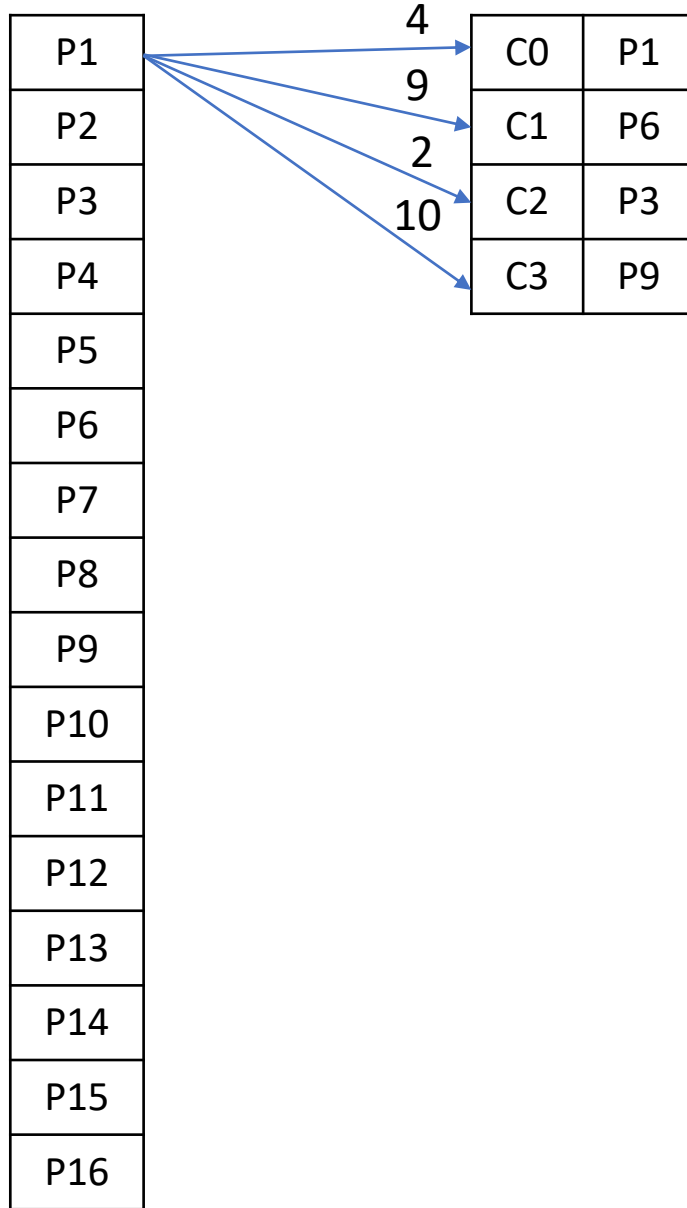
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

Sequential K-Means: (1) Choosing Initial Centroids

Initial centroids/means

Assigned Points



C0	
C1	
C2	
C3	

Sequential K-Means:

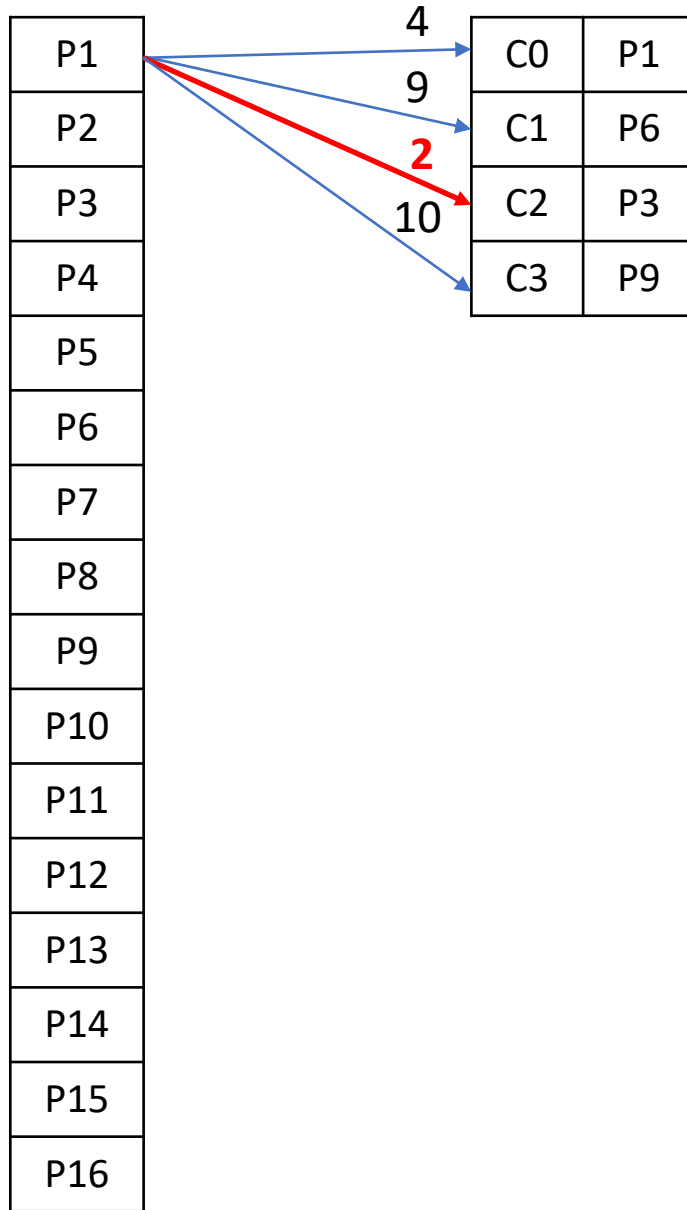
For each datapoint-

(2) Calculate distance to all centroids

(3) Assign it to the closest Cluster

Initial centroids/means

Assigned Points



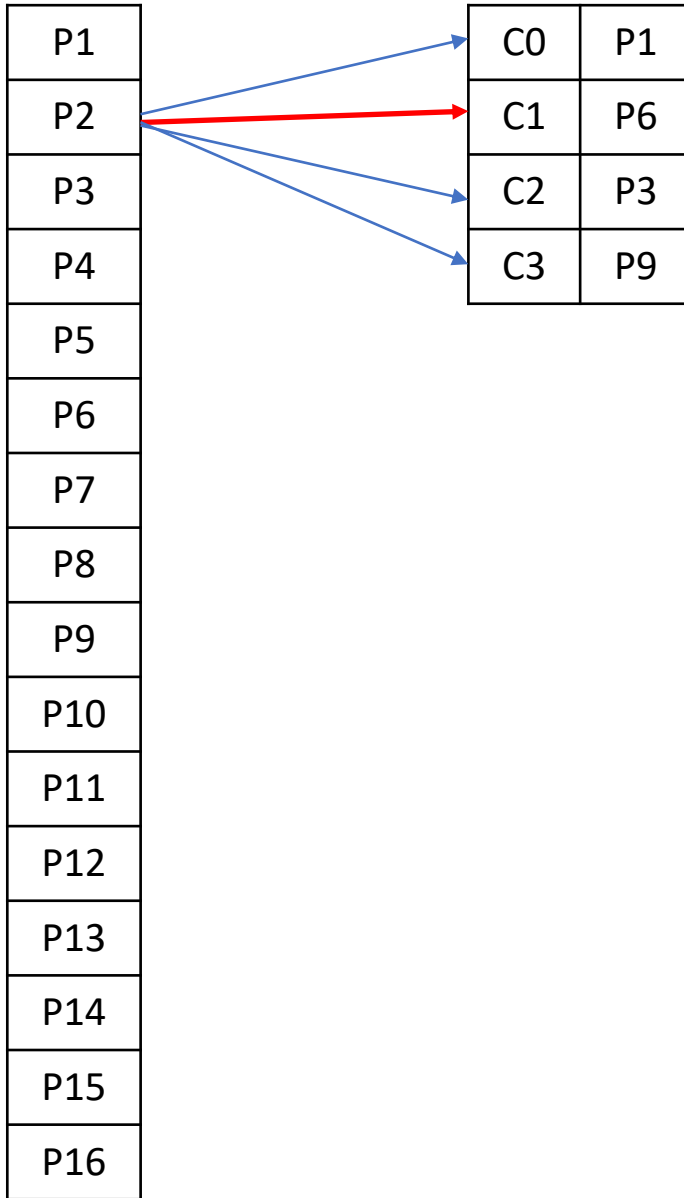
C0	
C1	
C2	P1
C3	

Sequential K-Means:
For each datapoint-
(2) Calculate distance to all centroids
(3) Assign it to the closest Cluster

Initial centroids/means

Assigned Points

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



C0	
C1	P2
C2	P1
C3	

Sequential K-Means:
For each datapoint-
(2) Calculate distance to
all centroids
(3) Assign it to the
closest Cluster

Initial centroids/means

Assigned Points

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

C0	
C1	P2 , P3
C2	P1
C3	

Sequential K-Means:

For each datapoint-

(2) Calculate distance to all centroids

(3) Assign it to the closest Cluster

Initial centroids/means

Assigned Points

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

C0	
C1	P2 , P3
C2	P1 , P4
C3	

Sequential K-Means:
 For each datapoint-
 (2) Calculate distance to
 all centroids
 (3) Assign it to the
 closest Cluster

Initial centroids/means

Assigned Points

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

C0	
C1	P2 , P3
C2	P1 , P4
C3	P5

Sequential K-Means:
For each datapoint-
(2) Calculate distance to
all centroids
(3) Assign it to the
closest Cluster

Initial centroids/means

Assigned Points

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

Sequential K-Means:

For each datapoint-

(2) Calculate distance to all centroids

(3) Assign it to the closest Cluster

Initial centroids/means

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

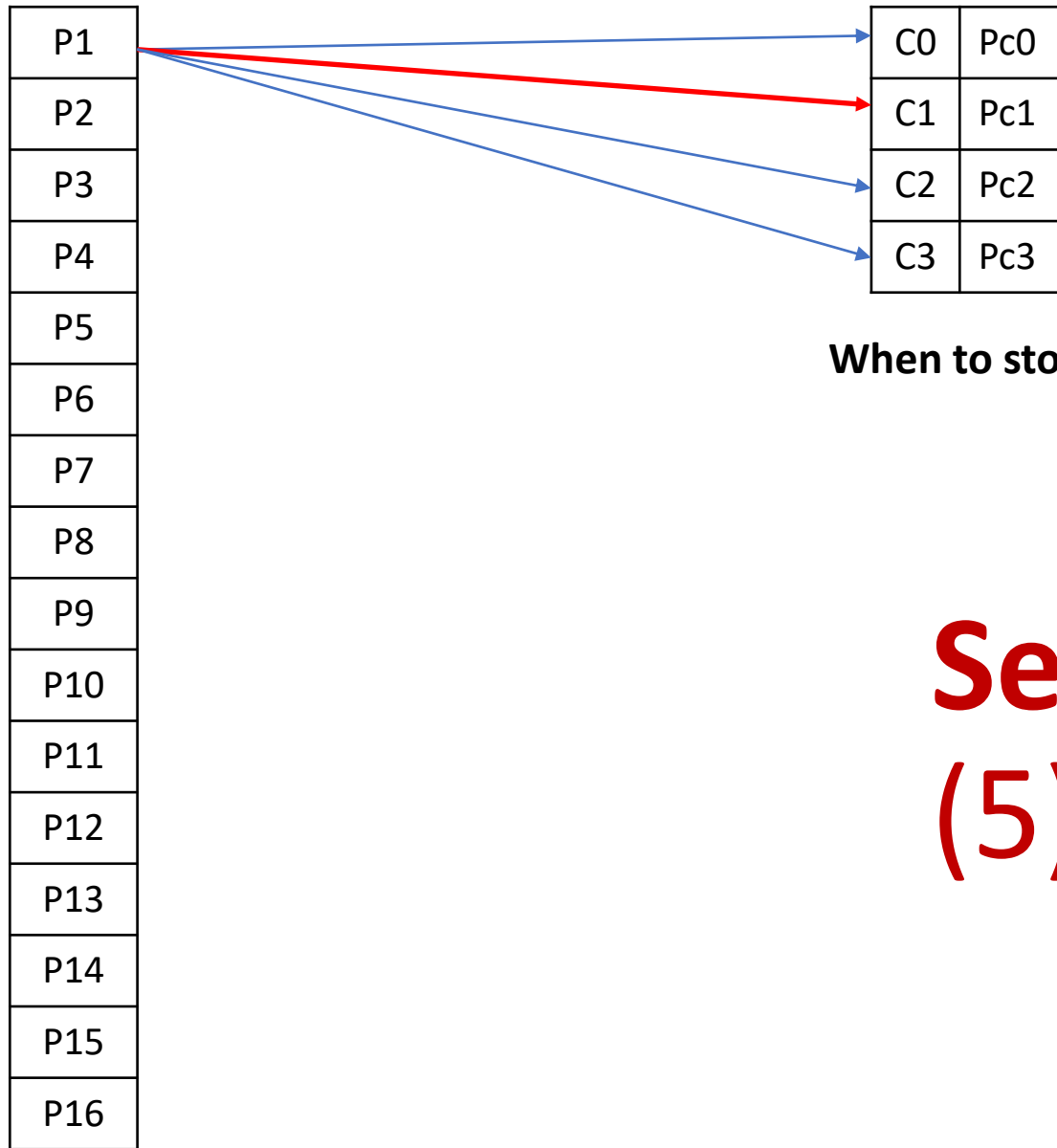
C0	$(P6 + P8 + P10 + P13)/4$
C1	$(P2 + P3 + P7 + P11)/4$
C2	$(P1 + P4 + P12 + P15 + P16)/5$
C3	$(P5 + P9 + P14)/3$

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

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

Sequential K-Means: (4) Recalculating Centroids

Centroids after iteration 1



When to stop?

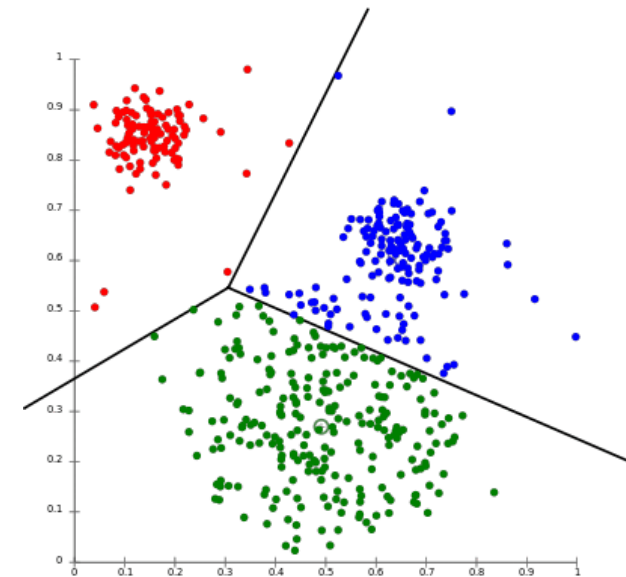
**Sequential K-Means:
(5) Repeat....**

When to Stop?

- Centroids of newly formed clusters don't change much
- Points remain in the same cluster
- Reach a Maximum number of iterations

Outline

- Project 3 Overview & Kmeans Algorithm
- Kmeans – 2D Points
 - Sequential
 - Parallel
- Kmeans – DNA Strands



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

How can we parallelize?

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

Parallel K-Means: (1) Choosing Initial Centroids

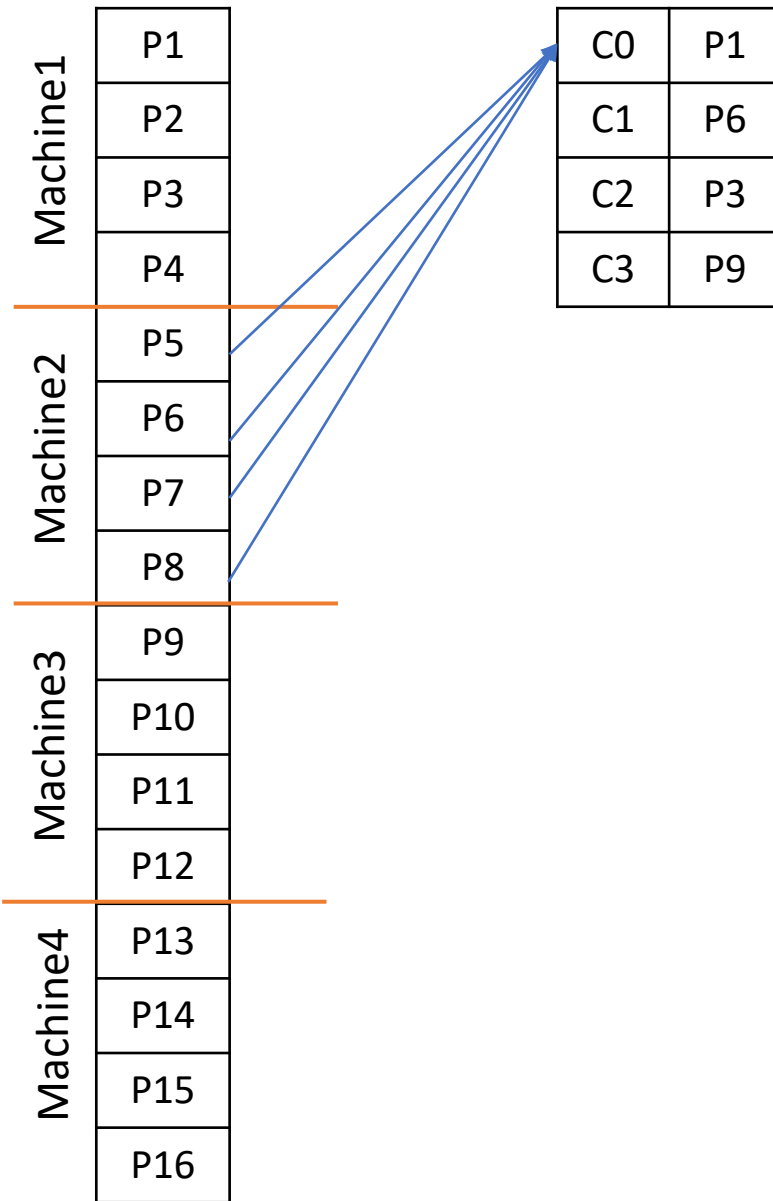
Initial centroids/means

C0	P1
C1	P6
C2	P3
C3	P9

Machine1	P1
	P2
	P3
	P4
Machine2	P5
	P6
	P7
	P8
Machine3	P9
	P10
	P11
	P12
Machine4	P13
	P14
	P15
	P16

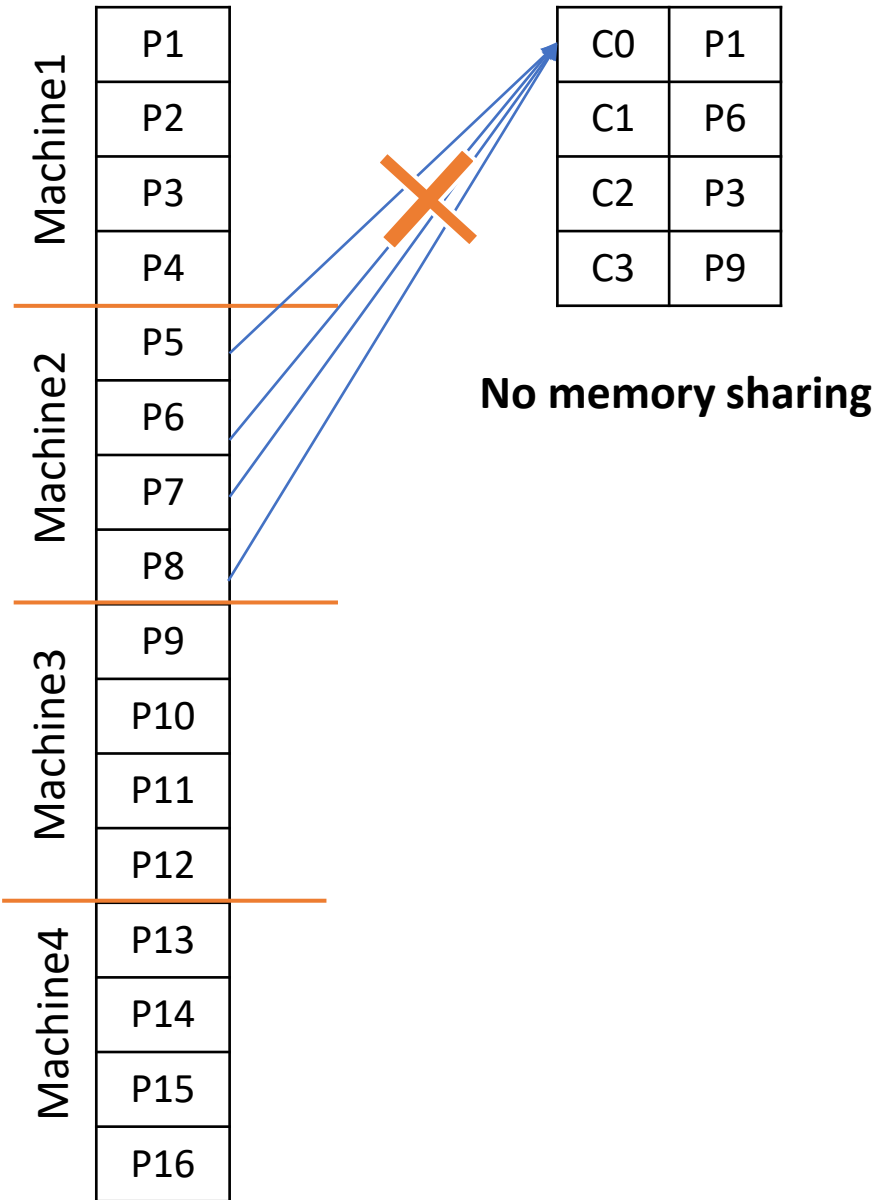
Parallel K-Means: (2) Split & Distribute Points to Multiple machines

Initial centroids/means



Is this sufficient?

Initial centroids/means



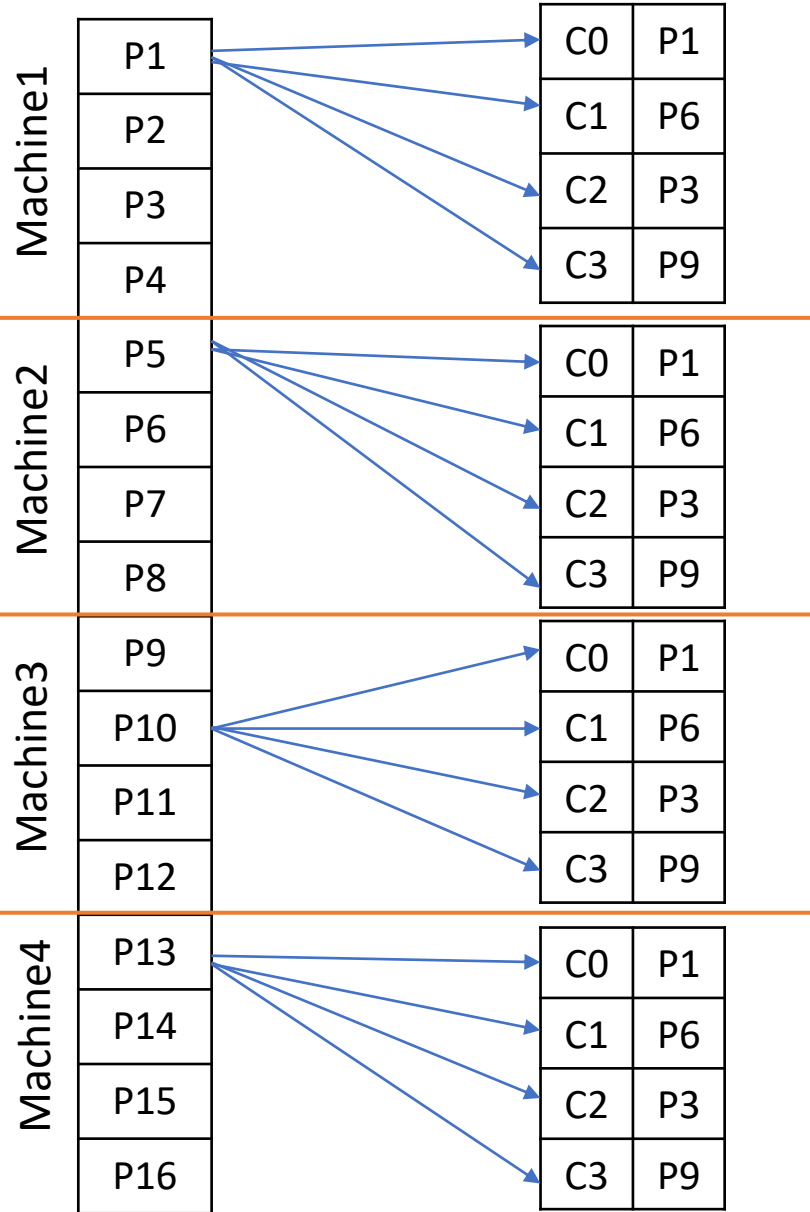
Is this sufficient?

Initial centroids/means

Machine1	P1	C0	P1
	P2	C1	P6
	P3	C2	P3
	P4	C3	P9
Machine2	P5	C0	P1
	P6	C1	P6
	P7	C2	P3
	P8	C3	P9
Machine3	P9	C0	P1
	P10	C1	P6
	P11	C2	P3
	P12	C3	P9
Machine4	P13	C0	P1
	P14	C1	P6
	P15	C2	P3
	P16	C3	P9

Parallel K-Means:
(3) Distribute centroids
to all machines

Initial centroids/means



Parallel K-Means:
(4) At each machine:
calculate distance of each
point to each centroid,
and assign it to the closest
centroid

Initial centroids/means

Assigned Points

Now each machine clustered its assigned points. **How to proceed?**

Machine	Initial centroids/means	Assigned Points
Machine1	P1	C0 P1
	P2	C1 P6
	P3	C2 P3
	P4	C3 P9
Machine2	P5	C0 P1
	P6	C1 P6
	P7	C2 P3
	P8	C3 P9
Machine3	P9	C0 P1
	P10	C1 P6
	P11	C2 P3
	P12	C3 P9
Machine4	P13	C0 P1
	P14	C1 P6
	P15	C2 P3
	P16	C3 P9

Initial centroids/means

Calculating new Centroids



Should each machine calculate new centroids?

Machine1

P1
P2
P3
P4

C0	P1
C1	P6
C2	P3
C3	P9

C0	$(P2 + P3)/2$
C1	0
C2	P1
C3	P4

Machine2

P5
P6
P7
P8

C0	P1
C1	P6
C2	P3
C3	P9

C0	P5
C1	P7
C2	P8
C3	P6

Machine3

P9
P10
P11
P12

C0	P1
C1	P6
C2	P3
C3	P9

C0	0
C1	P12
C2	$(P10 + P11)/2$
C3	P9

Machine4

P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	$(P13 + P14 + P16)/3$
C1	0
C2	0
C3	P15

Initial centroids/means

Assigned Points

		Initial centroids/means		Assigned Points	
Machine1	P1	C0	P1	C0	P2 , P3
	P2	C1	P6	C1	0
	P3	C2	P3	C2	P1
	P4	C3	P9	C3	P4
Machine2	P5	C0	P1	C0	P5
	P6	C1	P6	C1	P7
	P7	C2	P3	C2	P8
	P8	C3	P9	C3	P6
Machine3	P9	C0	P1	C0	0
	P10	C1	P6	C1	P12
	P11	C2	P3	C2	P10 , P11
	P12	C3	P9	C3	P9
Machine4	P13	C0	P1	C0	P13 , P14 , P16
	P14	C1	P6	C1	0
	P15	C2	P3	C2	0
	P16	C3	P9	C3	P15

Initial centroids/means

Assigned Points
Machine 1

Assigned Points
Machine 2

Assigned Points
Machine 3

Assigned Points
Machine 4

Machine1

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

Machine2

P5
P6
P7
P8

C0	P1
C1	P6
C2	P3
C3	P9

Machine3

P9
P10
P11
P12

C0	P1
C1	P6
C2	P3
C3	P9

Machine4

P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

Parallel K-Means:
(5) Each machine reports to the master its clustering assignments

Initial centroids/means

Centroids after iteration 1

Machine1	P1
	P2
	P3
	P4
Machine2	P5
	P6
	P7
	P8
Machine3	P9
	P10
	P11
	P12
Machine4	P13
	P14
	P15
	P16

C0	P1
C1	P6
C2	P3
C3	P9
C0	P1
C1	P6
C2	P3
C3	P9
C0	P1
C1	P6
C2	P3
C3	P9
C0	P1
C1	P6
C2	P3
C3	P9

C0	$(P2 + P3 + P5 + P13 + P14 + P16)/6$
C1	$(P7 + P12)/2$
C2	$(P8 + P10 + P11)/3$
C3	$(P6 + P9 + P15)/3$

Parallel K-Means:
(6) Master node
calculates new
centroids

Centroids after iteration 1

Machine1	P1
	P2
	P3
	P4
Machine2	P5
	P6
	P7
	P8
Machine3	P9
	P10
	P11
	P12
Machine4	P13
	P14
	P15
	P16

C0	Pc0
C1	Pc1
C2	Pc2
C3	Pc3

Parallel K-Means:
(6) Master node
calculates new
centroids

Centroids after iteration 1

Machine1	P1	C0	Pc0
	P2	C1	Pc1
	P3	C2	Pc2
	P4	C3	Pc3
Machine2	P5	C0	Pc0
	P6	C1	Pc1
	P7	C2	Pc2
	P8	C3	Pc3
Machine3	P9	C0	Pc0
	P10	C1	Pc1
	P11	C2	Pc2
	P12	C3	Pc3
Machine4	P13	C0	Pc0
	P14	C1	Pc1
	P15	C2	Pc2
	P16	C3	Pc3

Parallel K-Means:
(7) Repeat.. Master node distributes new centroids to all machines

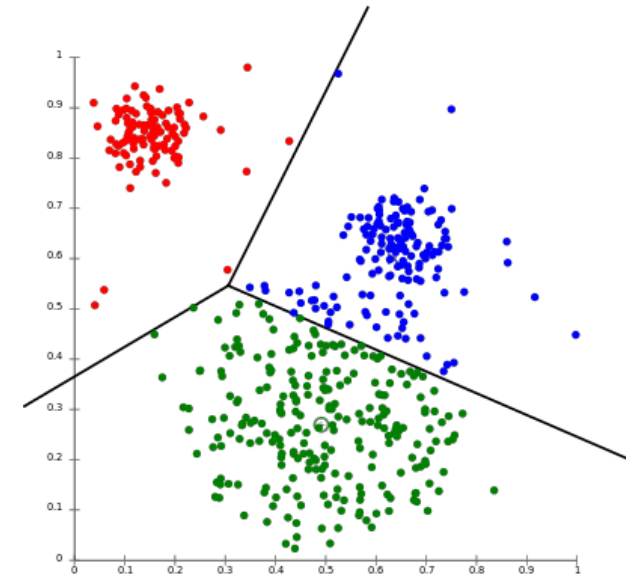
K-Means in Project 3

The main functions you will need are: distance and mean.

- **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?

Outline

- Project 3 Overview & Kmeans Algorithm
- Kmeans – 2D Points
 - Sequential
 - Parallel
- Kmeans – DNA Strands



K-Means: DNA Strands

A strand of DNA consists of a string of molecules called bases, where the possible bases are adenine (A), guanine (G), cytosine (C), and thymine (T).

ACTG
GTCA
SGGT
TAAA
ATAT

Given a list of strands,
How to cluster them using K-means?

Specifically,

How to find the centroid (mean) of a set of points?

How to calculate the distance of a point to a centroid?

K-Means: DNA Strands

Point-Centroid Distance

ACTG
GTCA
SGGT
TAAA
ATAT

The distance between two strands is the number of changes required to turn one strand into the other.


E.g. distance between ACTG and ATAT= 3

K-Means: DNA Strands

Calculating Centroid (mean)

ACTG
GTCA
SGGT
TAAA
ATAT

How to get the
centroid (mean) of
these DNA strands?



K-Means: DNA Strands

Calculating Centroid (mean)

How many repetitions of A in
index 0 of all strands

ACTG
GTCA
SGGT
TAAA
ATAT

A				
C				
G				
T				
Output strand				

K-Means: DNA Strands

Calculating Centroid (mean)

A	C	T	G
A	T	C	G
G	T	A	C
T	A	G	C
A	T	A	T

A	2			
C				
G				
T				
Output strand				

K-Means: DNA Strands

Calculating Centroid (mean)

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				

K-Means: DNA Strands

Calculating Centroid (mean)

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
(For each Index, sort the values and select the
middle one)

K-Means: DNA Strands

Calculating Centroid (mean)

ACTG
GTCA
SGGT
TAAA
ATAT

Unsorted

A	2	1	2	1
C	0	1	1	0
G	1	1	1	1
T	1	2	1	2

Get the mean or the median
(For each Index,
sort the values and select
the middle one)

Index 0		Index 1		Index 2		Index 3	
C	0	A	1	C	1	C	0
G	1	C	1	G	1	A	1
T	1	G	1	T	1	G	1
A	2	T	2	A	2	T	2

Median Strand: GCGA

