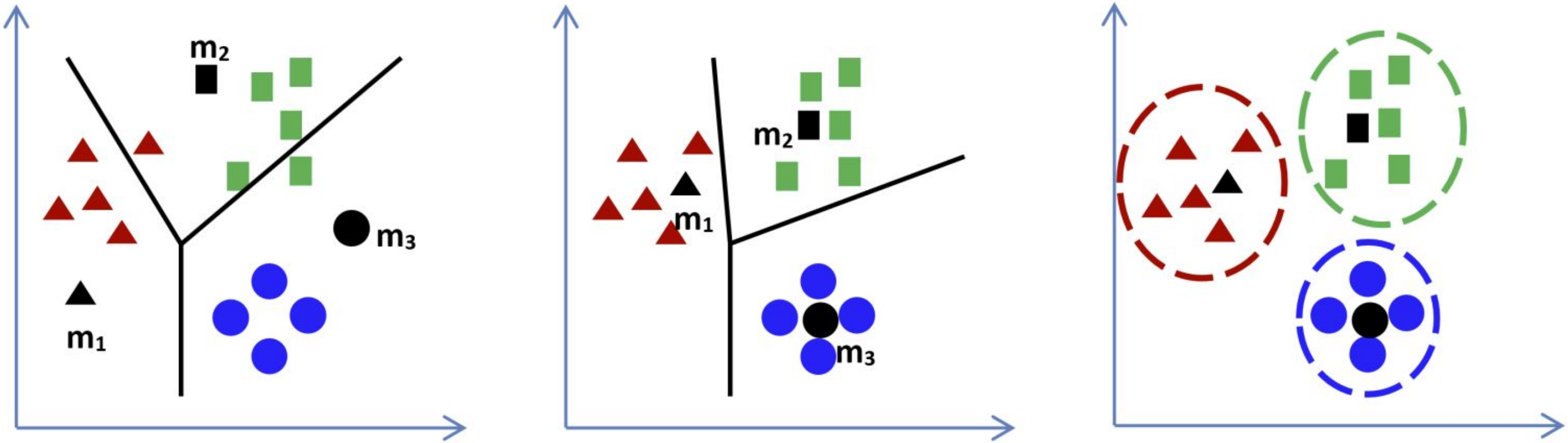
15-440
Distributed Systems
K-Means

Zeinab Khalifa

جامعة كارنيجي ميلون في قطر
Carnegie Mellon University Qatar

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.
  - Visualization



(a) Initial Means. (b) Recalculated Means. (c) Final Clusters.

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.

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

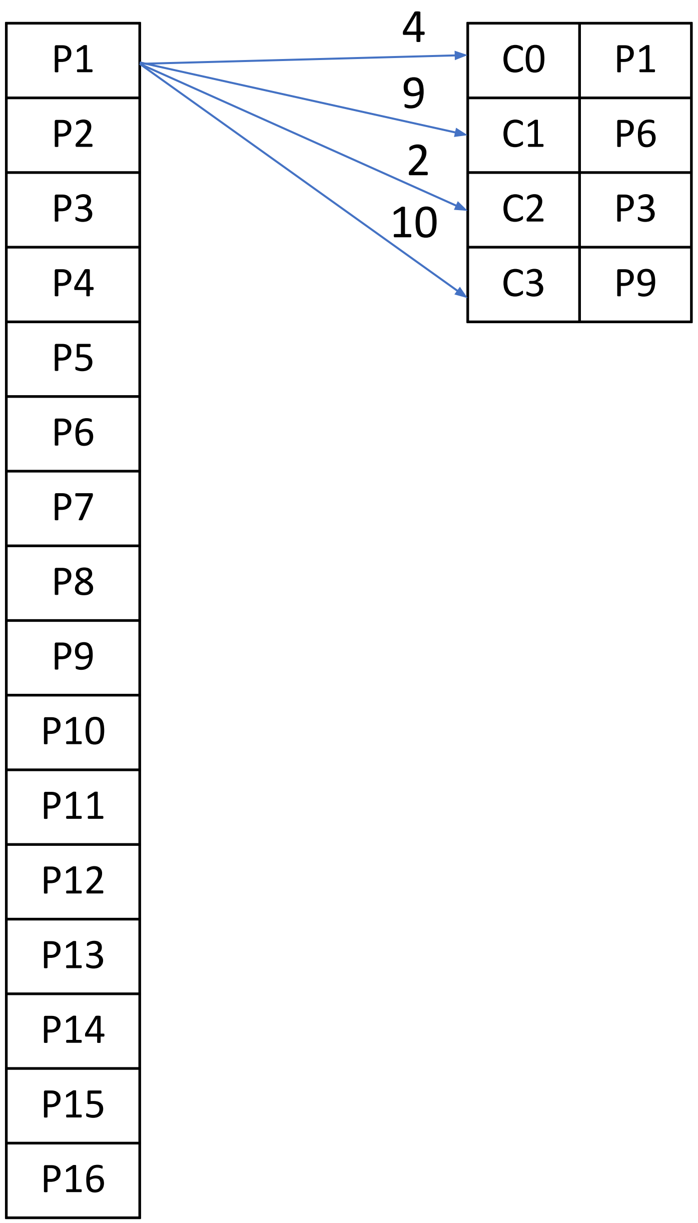
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

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

Initial centroids/means

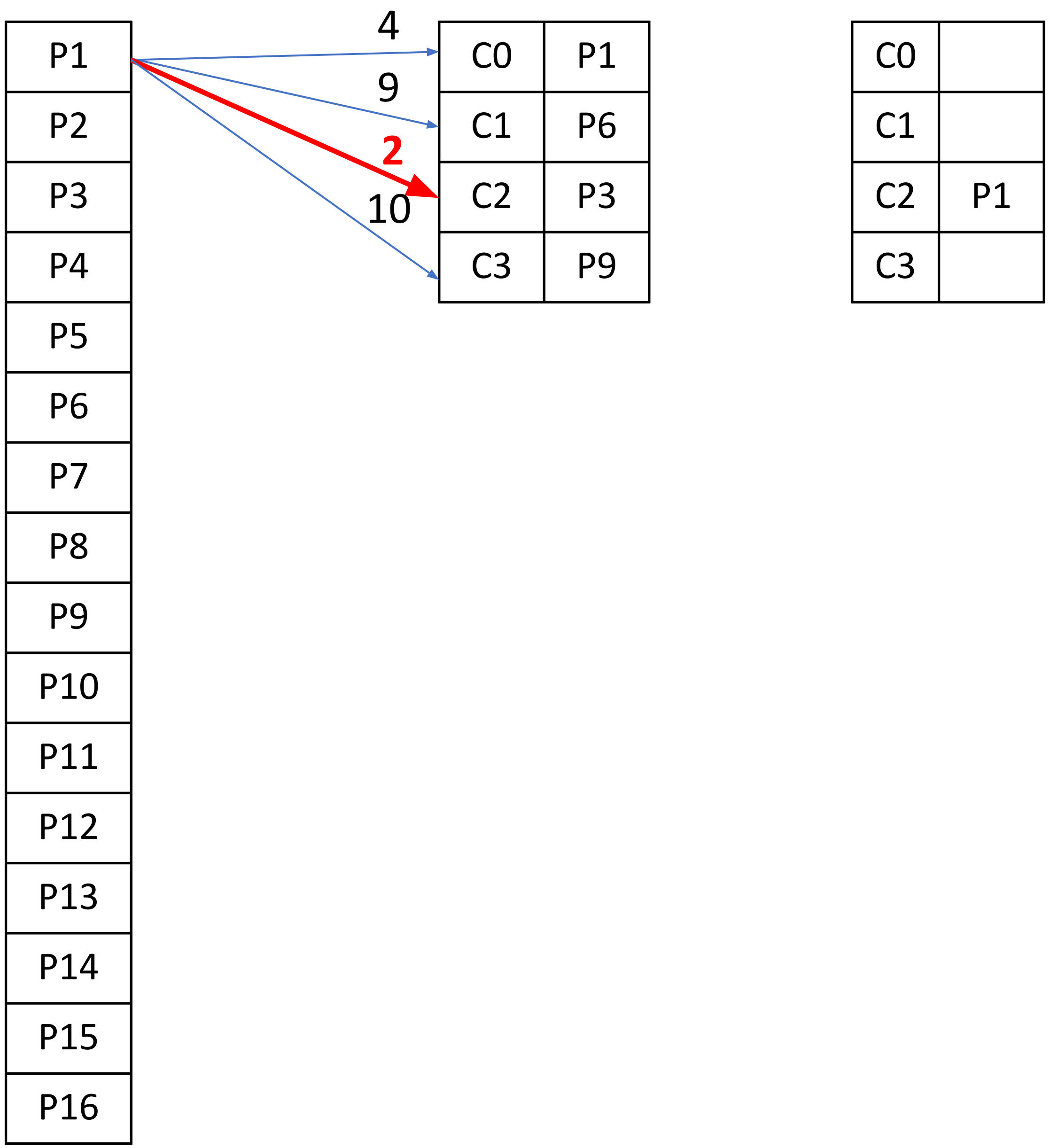


P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	
C1	
C2	
C3	

Initial centroids/means

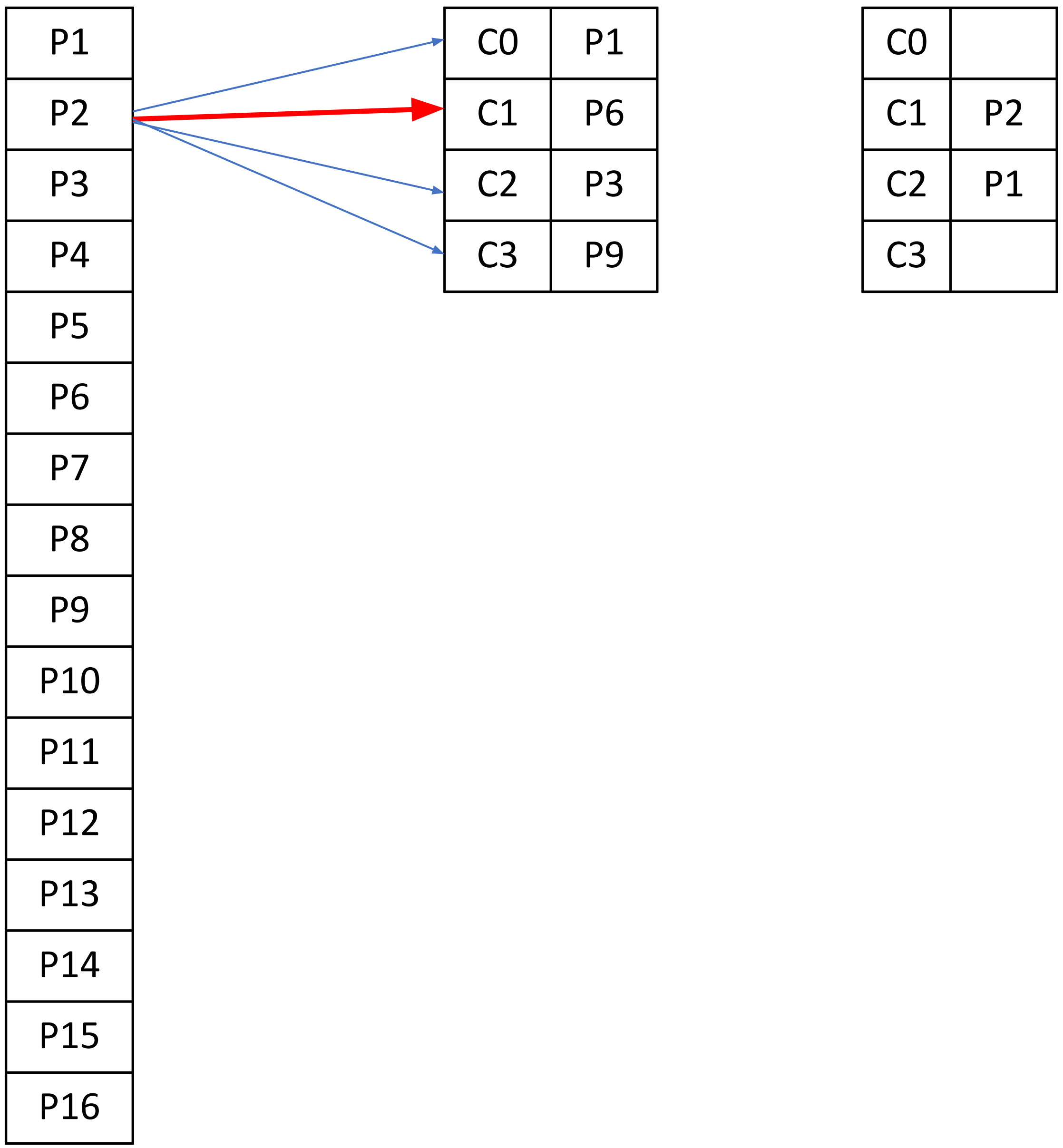


P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	
C1	
C2	P1
C3	

Initial centroids/means

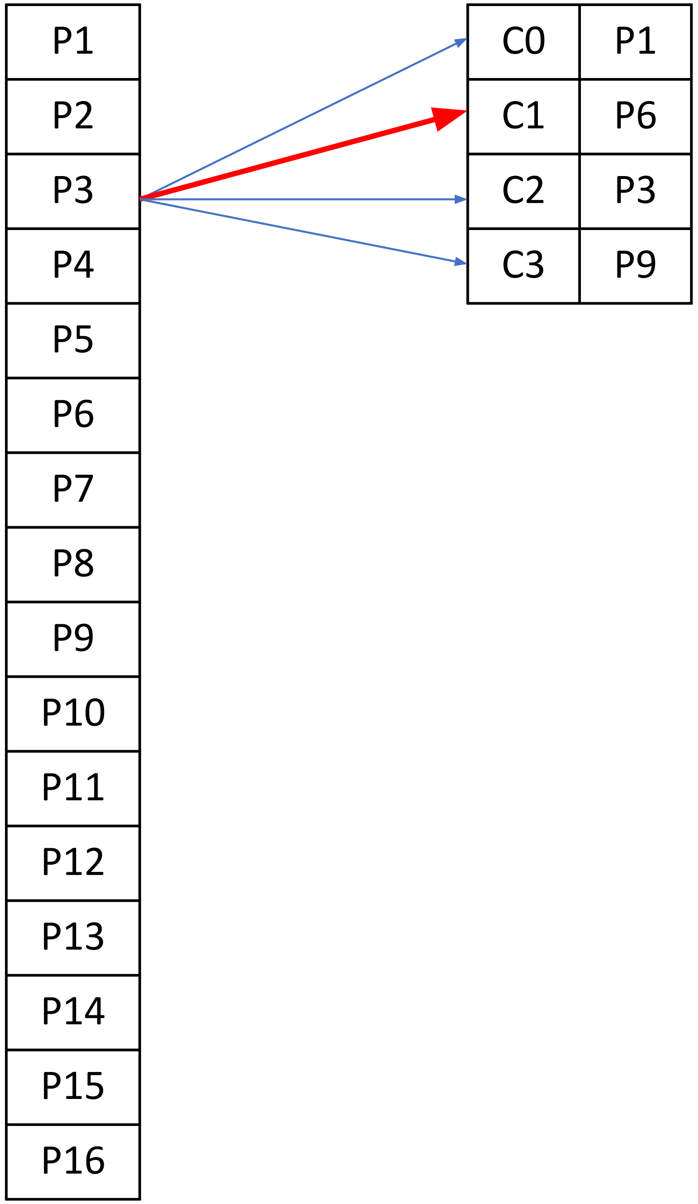


P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	
C1	P2
C2	P1
C3	

Initial centroids/means

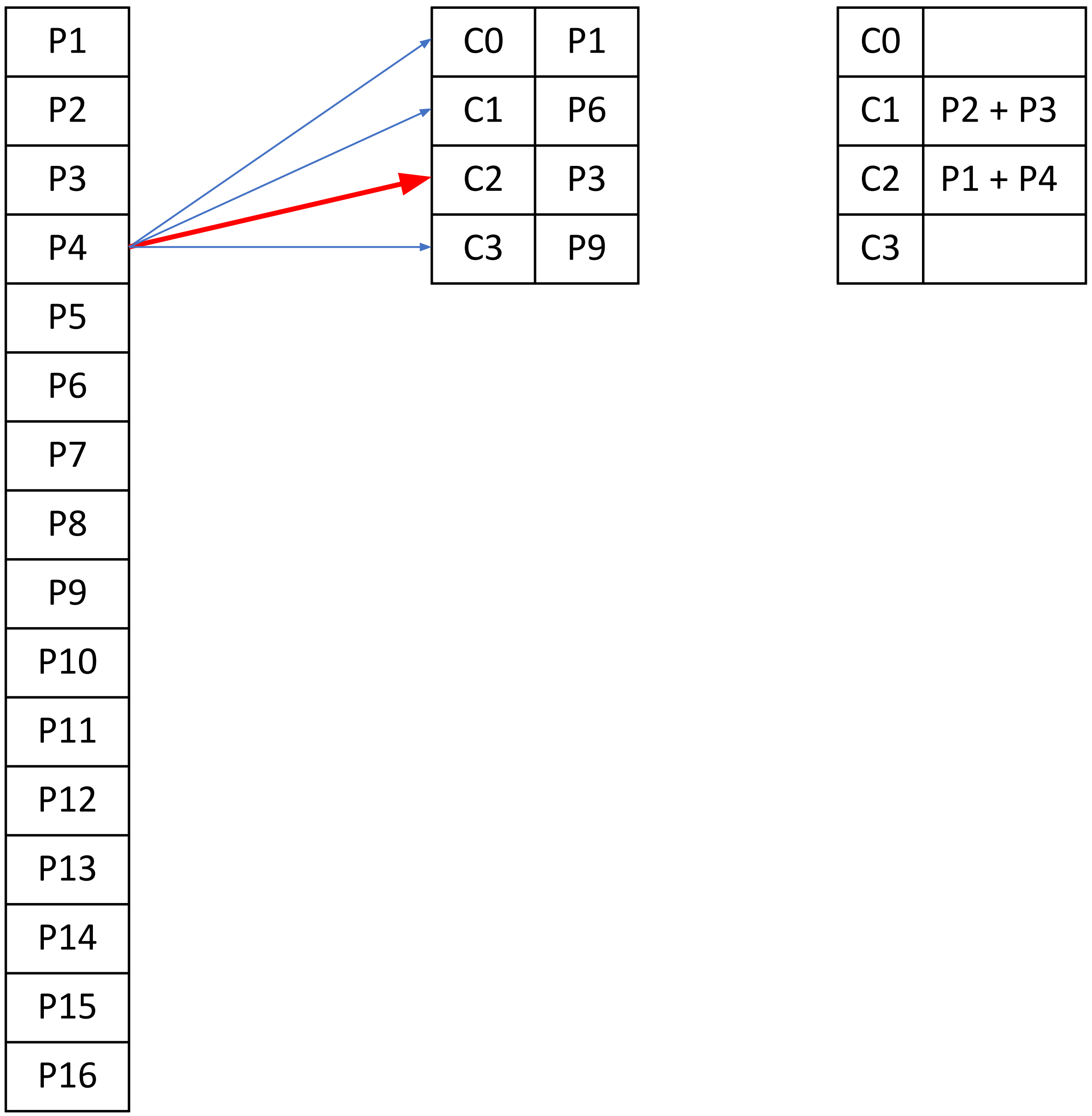


P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

C0	
C1	P2 + P3
C2	P1
C3	

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



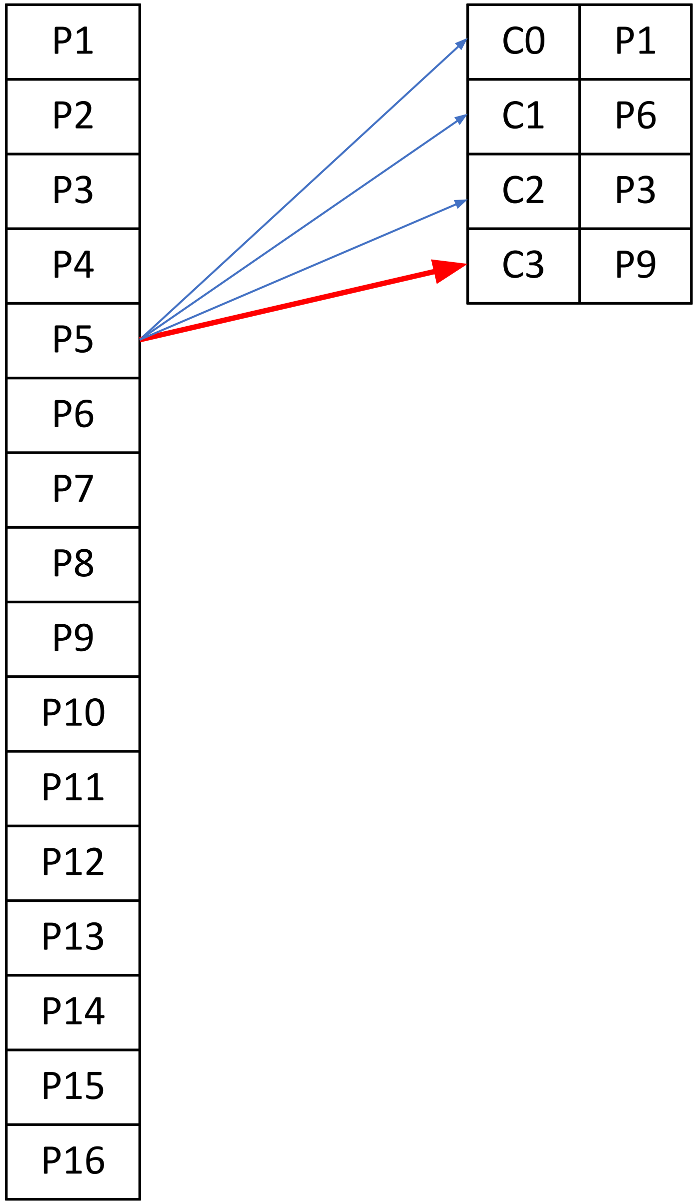
Initial centroids/means

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	
C1	P2 + P3
C2	P1 + P4
C3	

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	
C1	P2 + P3
C2	P1 + P4
C3	P5

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

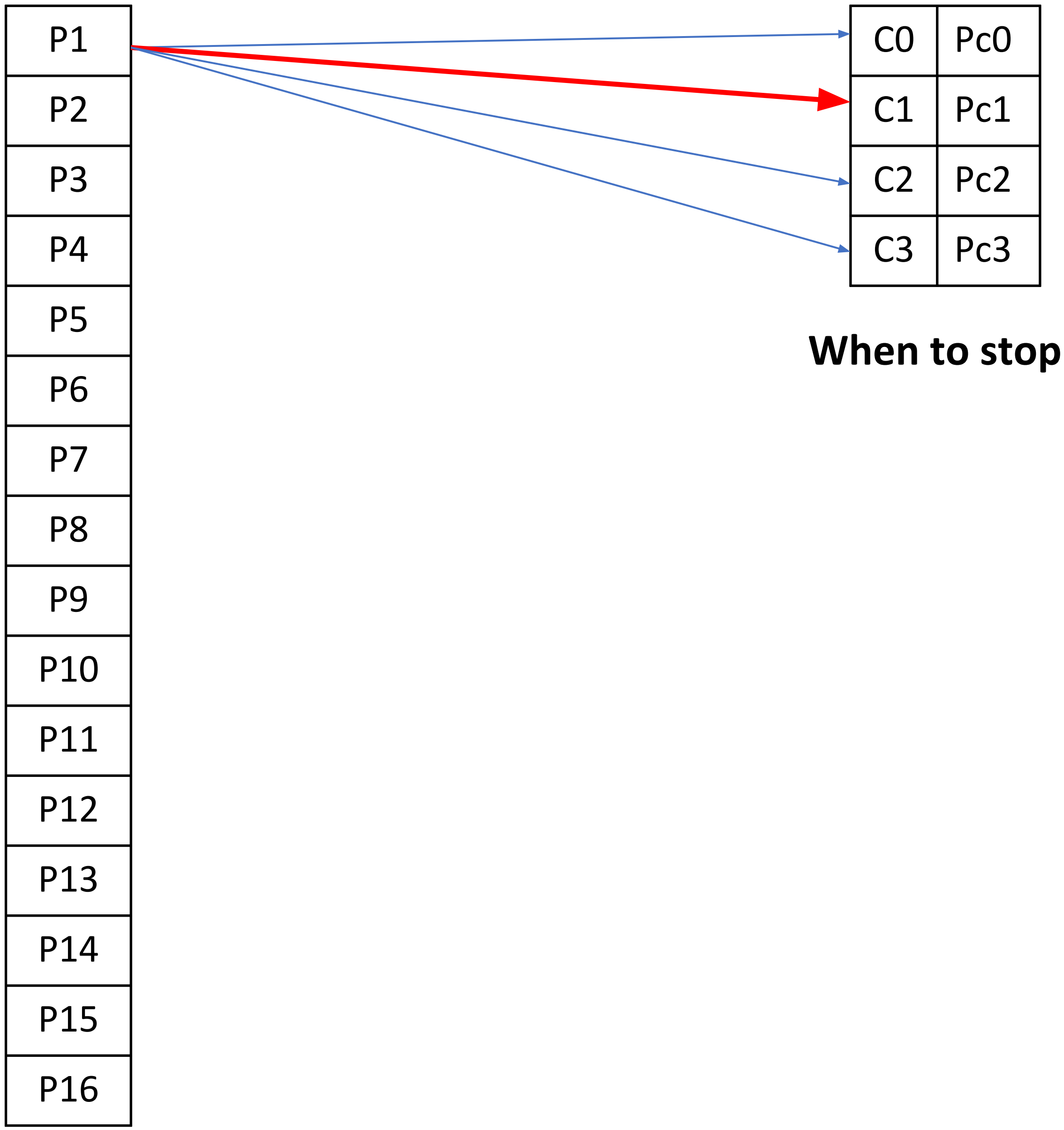
Centroids after iteration 1

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)*



P1

P2

P3

P4

P5

P6

P7

P8

P9

P10

P11

P12

P13

P14

P15

P16

C0	Pc0
C1	Pc1
C2	Pc2
C3	Pc3

When to stop?

Parallel K-Means

How can we parallelize?

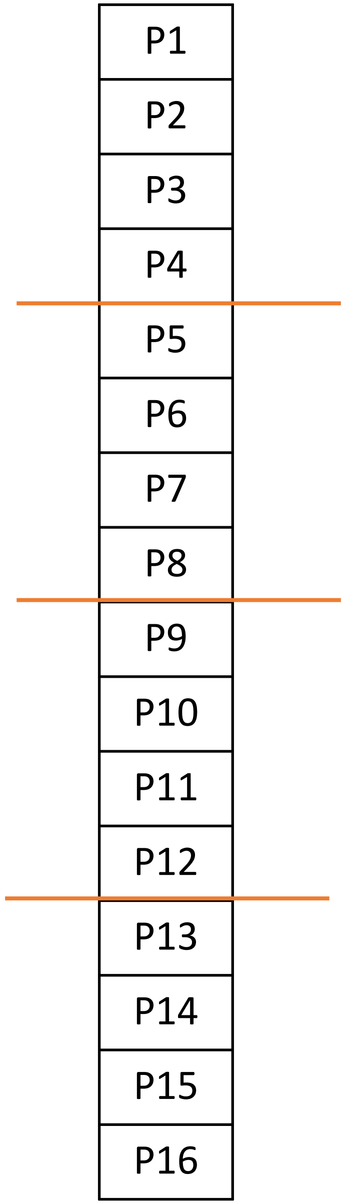
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

How can we parallelize?

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

C0	P1
C1	P6
C2	P3
C3	P9

How can we parallelize?



P1
P2
P3
P4

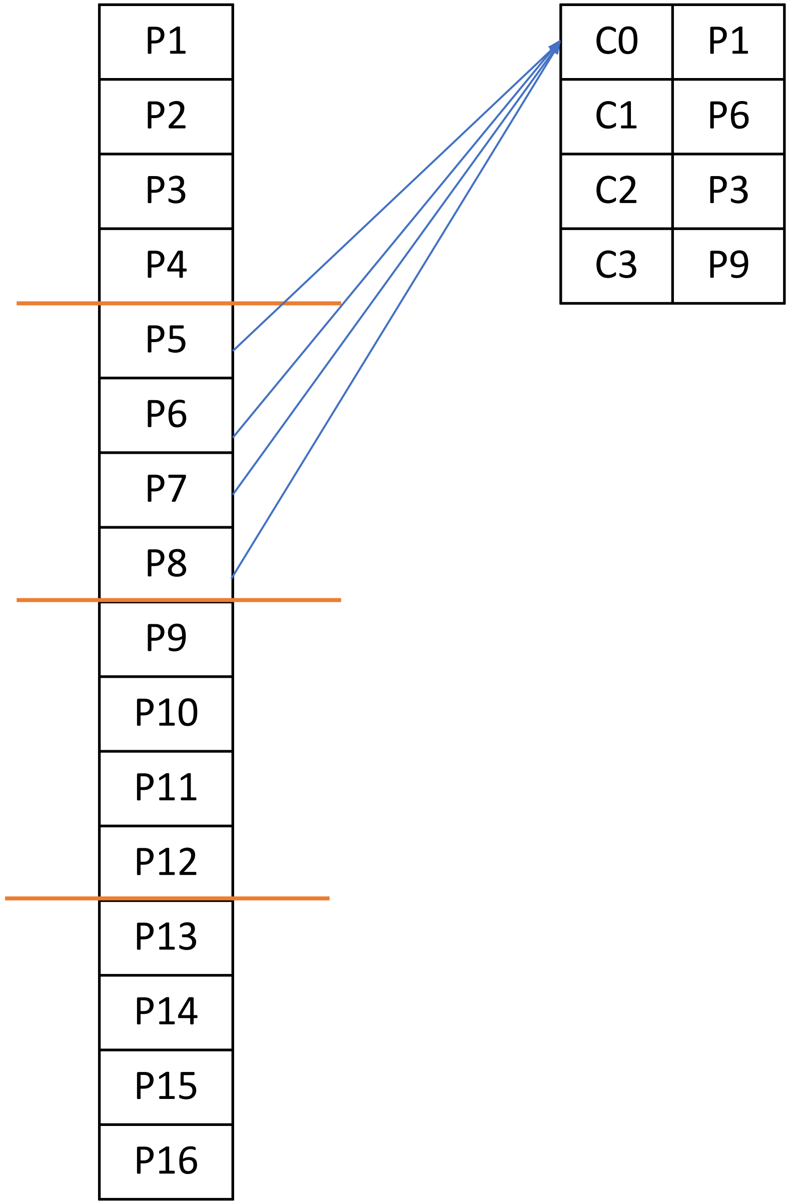
P5
P6
P7
P8

P9
P10
P11
P12

P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

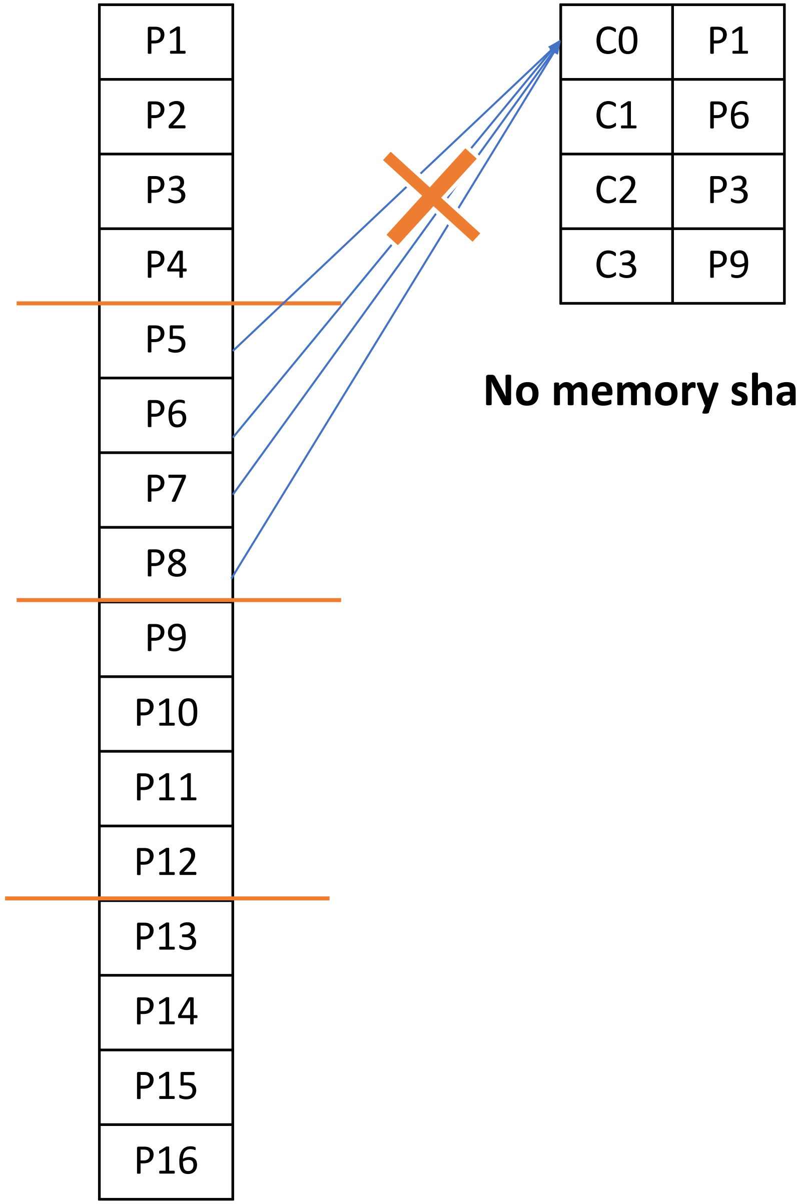
How can we parallelize?



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

C0	P1
C1	P6
C2	P3
C3	P9

How can we parallelize?

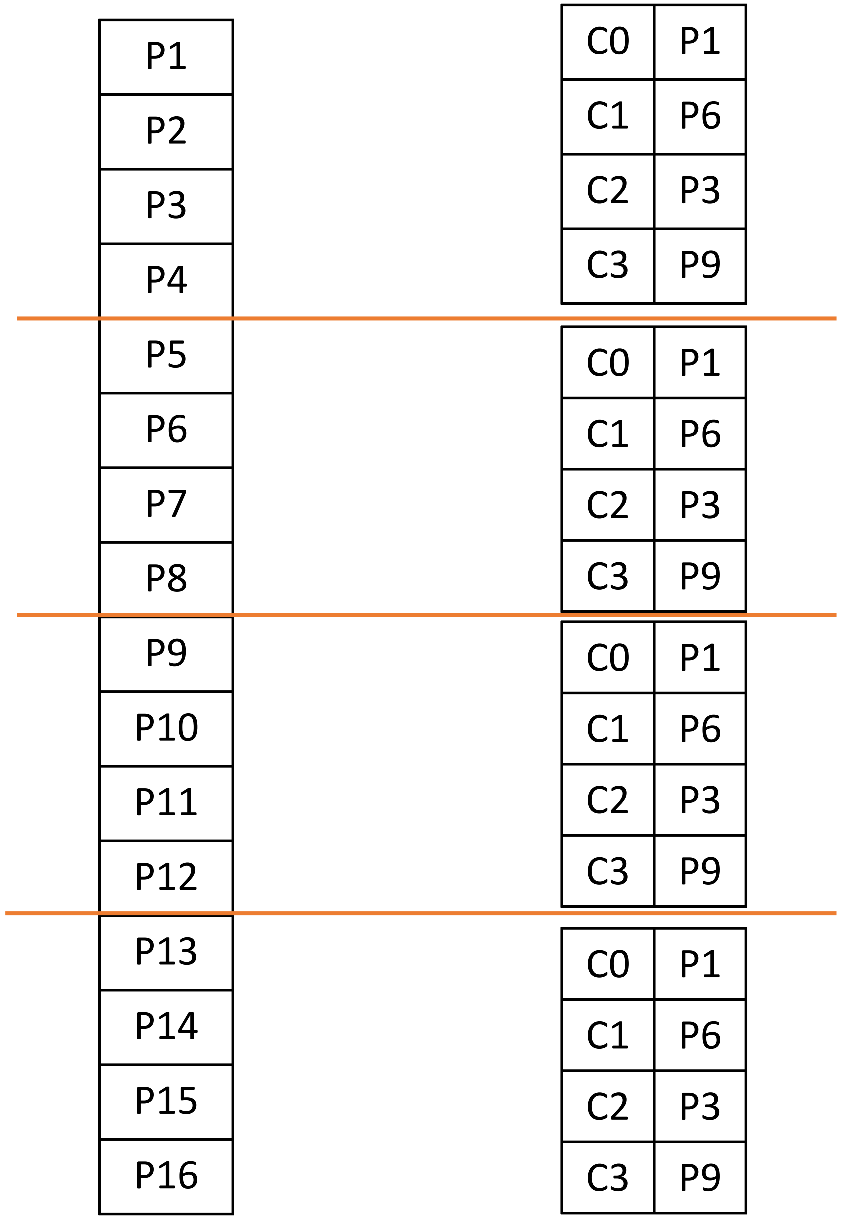


P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

No memory sharing

How can we parallelize?



P1
P2
P3
P4

C0	P1
C1	P6
C2	P3
C3	P9

P5
P6
P7
P8

C0	P1
C1	P6
C2	P3
C3	P9

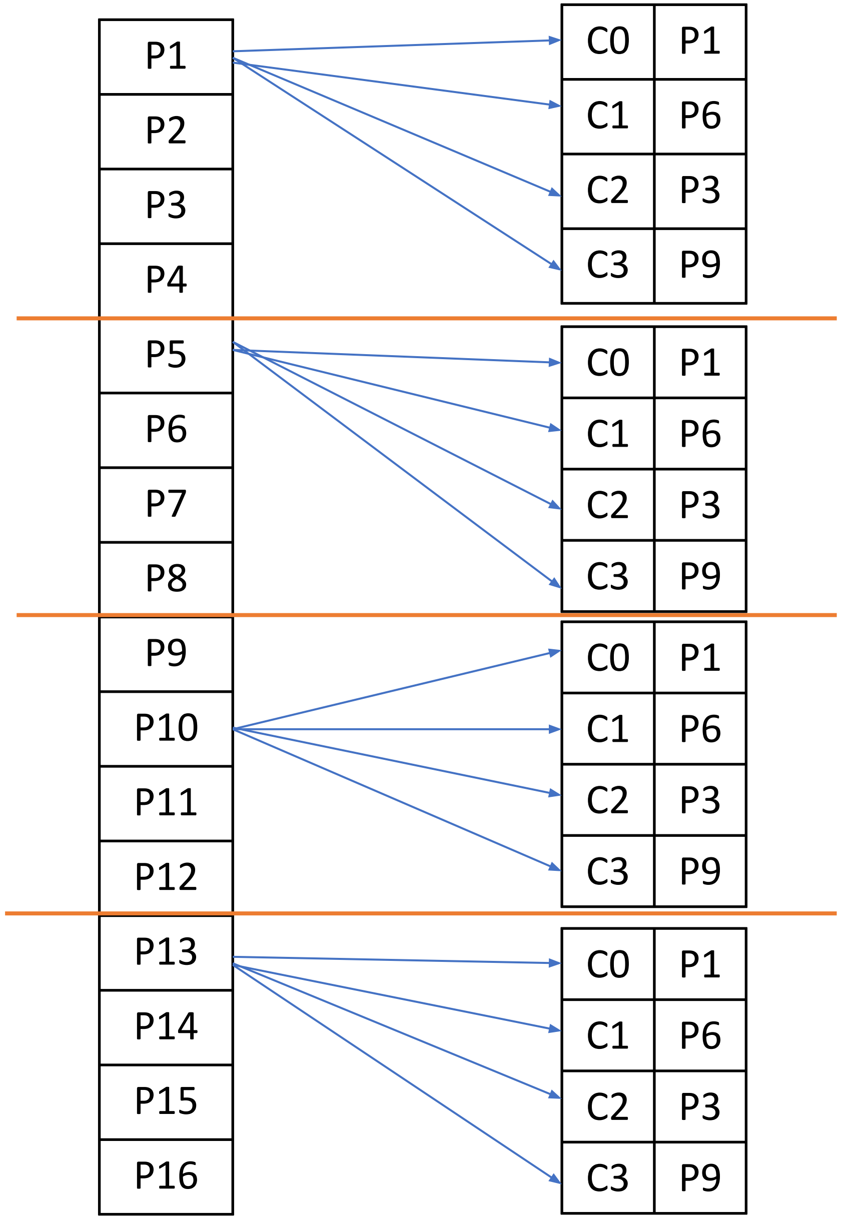
P9
P10
P11
P12

C0	P1
C1	P6
C2	P3
C3	P9

P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

How can we parallelize?



How can we parallelize?

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	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
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

How can we parallelize?

P1	C0	P1	C0	P2 + P3 /2 ✗
P2	C1	P6	C1	0
P3	C2	P3	C2	P1 /1
P4	C3	P9	C3	P4 /1
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	C3	P9	C3	P9
P13	C0	P1	C0	P13 + P14 + P16
P14	C1	P6	C1	0
P15	C2	P3	C2	0
P16	C3	P9	C3	P15

How can we parallelize?

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	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
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

How can we parallelize?

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	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

C0	P1
C1	P6
C2	P3
C3	P9

C0	P1
C1	P6
C2	P3
C3	P9

C0	P1
C1	P6
C2	P3
C3	P9

How can we parallelize?

P1
P2
P3
P4

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

P5
P6
P7
P8

C0	P1
C1	P6
C2	P3
C3	P9

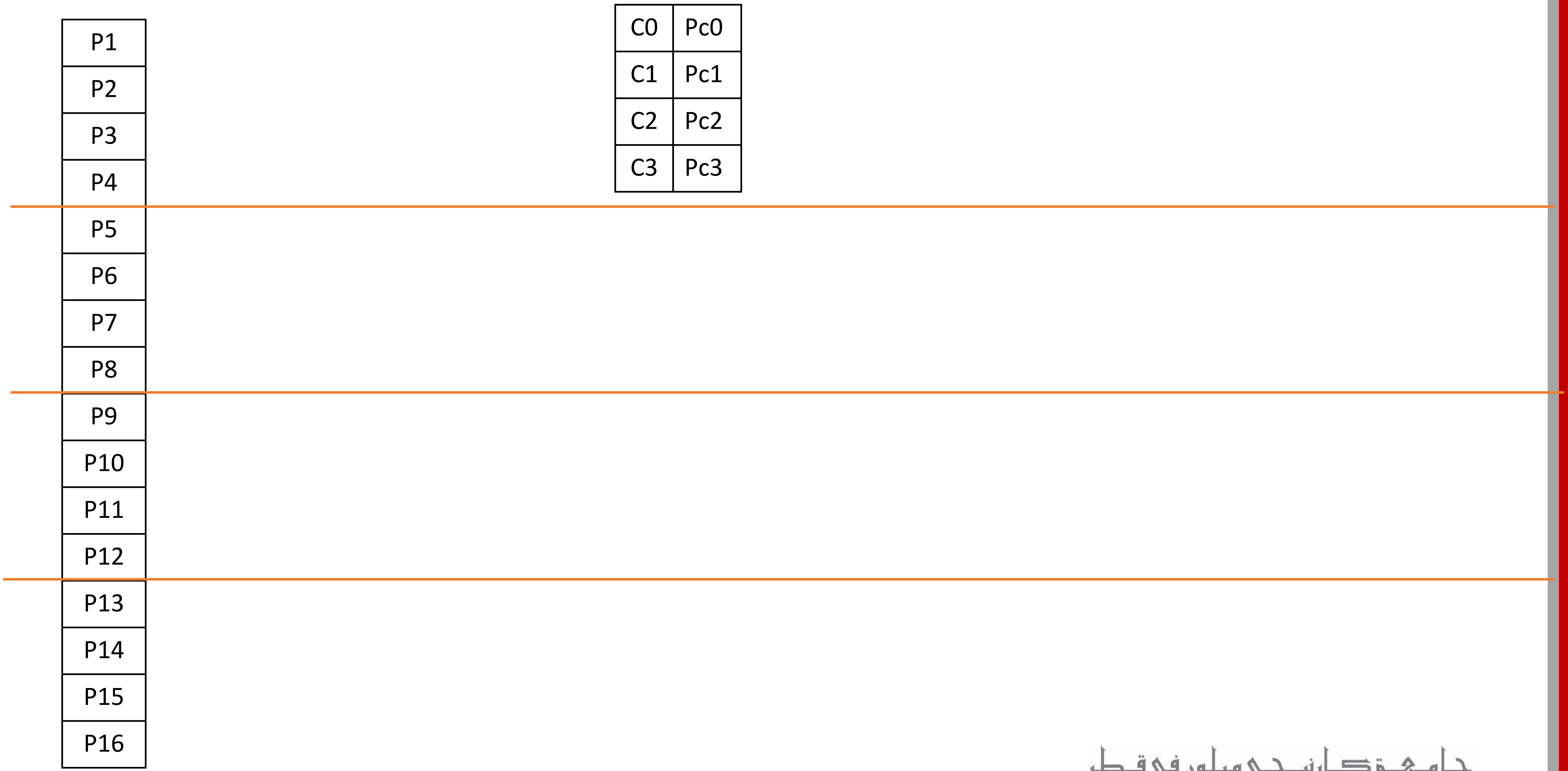
P9
P10
P11
P12

C0	P1
C1	P6
C2	P3
C3	P9

P13
P14
P15
P16

C0	P1
C1	P6
C2	P3
C3	P9

How can we parallelize?



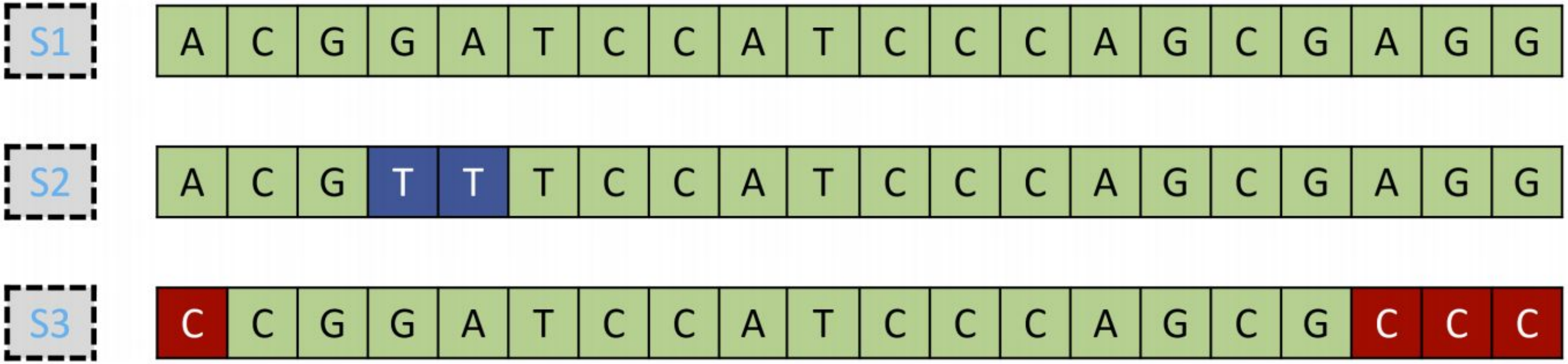
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16

C0	Pc0
C1	Pc1
C2	Pc2
C3	Pc3

How can we parallelize?

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

DNA stranding

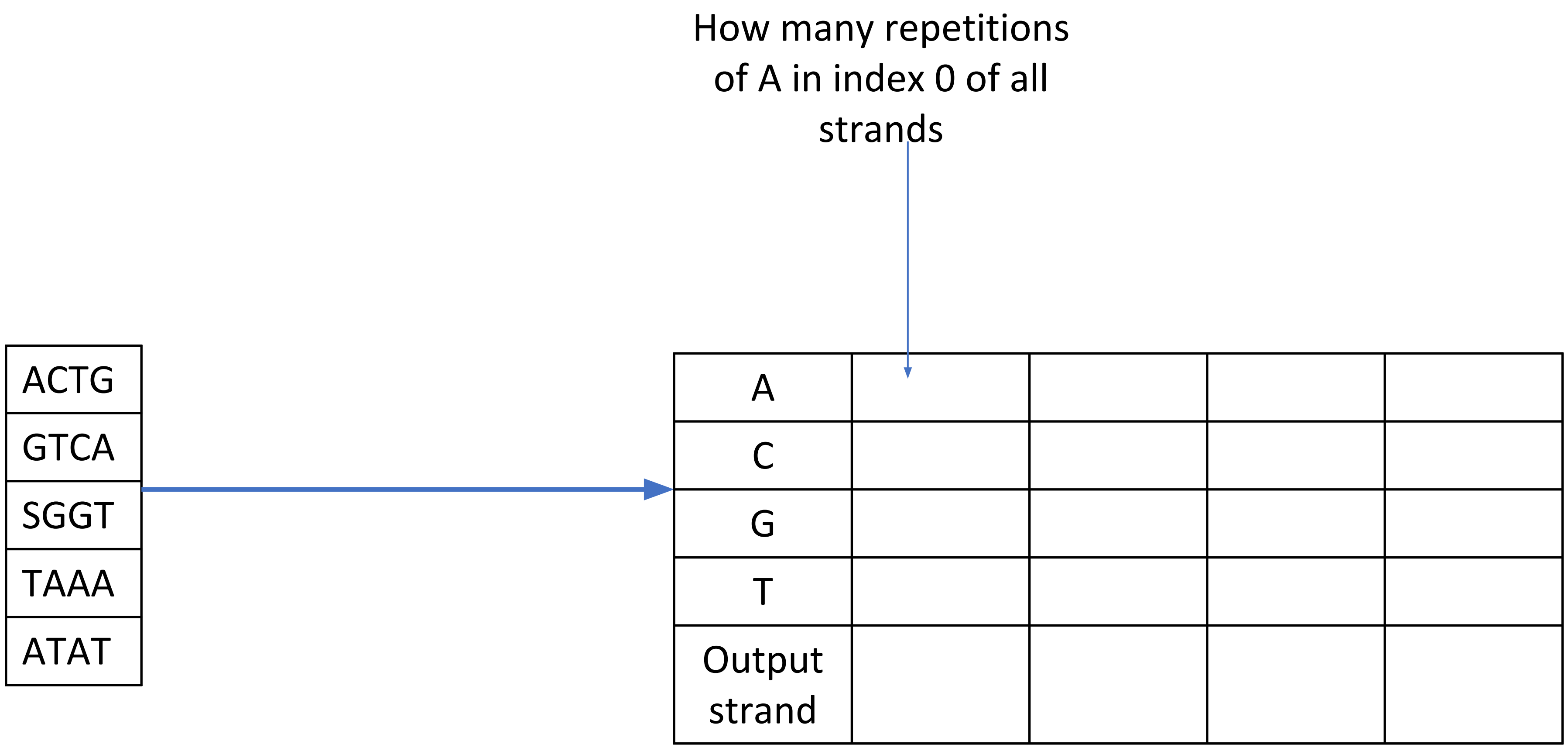


S1	A	C	G	G	A	T	C	C	A	T	C	C	C	A	G	C	G	A	G	G
S2	A	C	G	T	T	T	C	C	A	T	C	C	C	A	G	C	G	A	G	G
S3	C	C	G	G	A	T	C	C	A	T	C	C	C	A	G	C	G	C	C	C

ACTG
GTCA
SGGT
TAAA
ATAT

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)

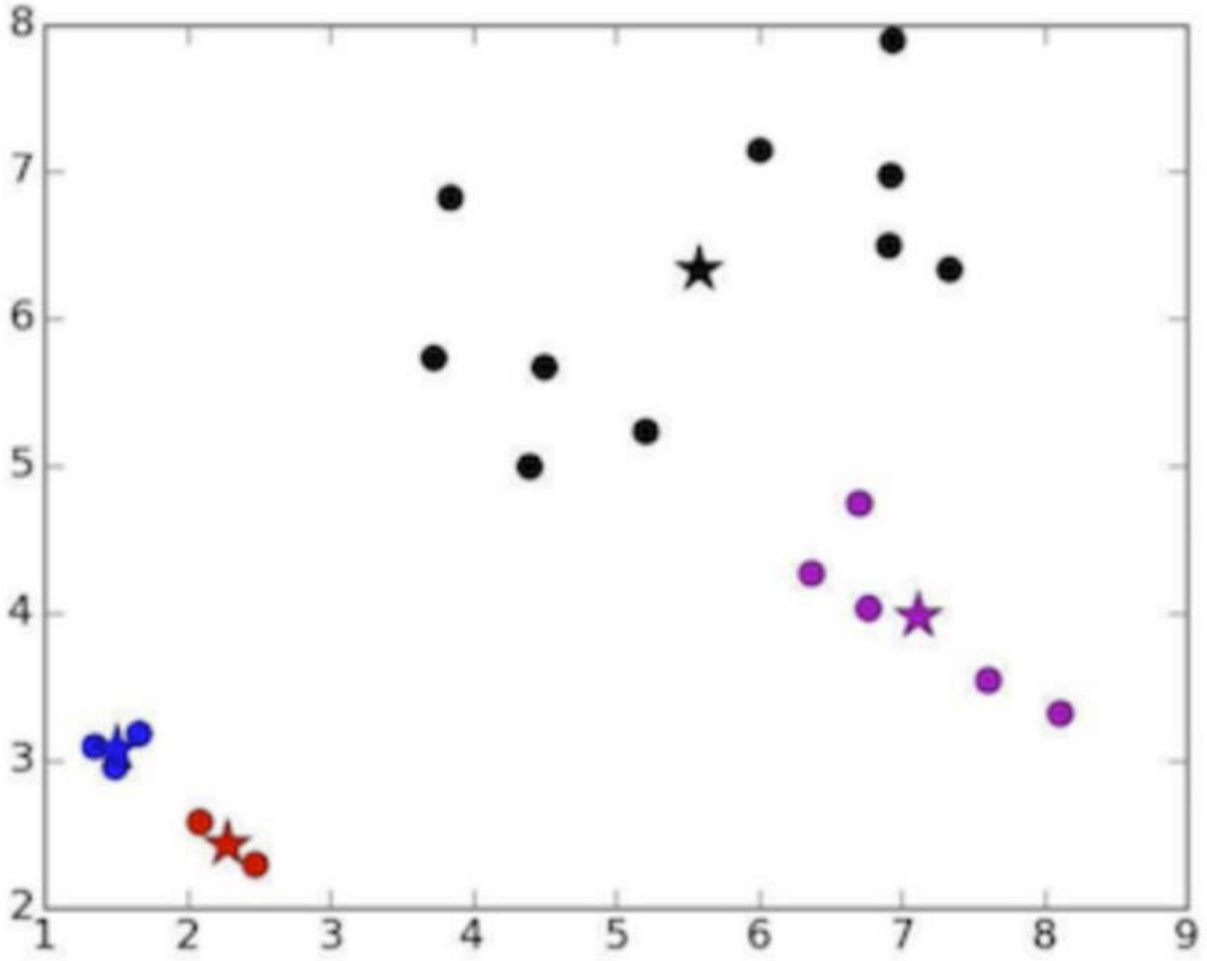
ACTG
GTCA
SGGT
TAAA
ATAT

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**

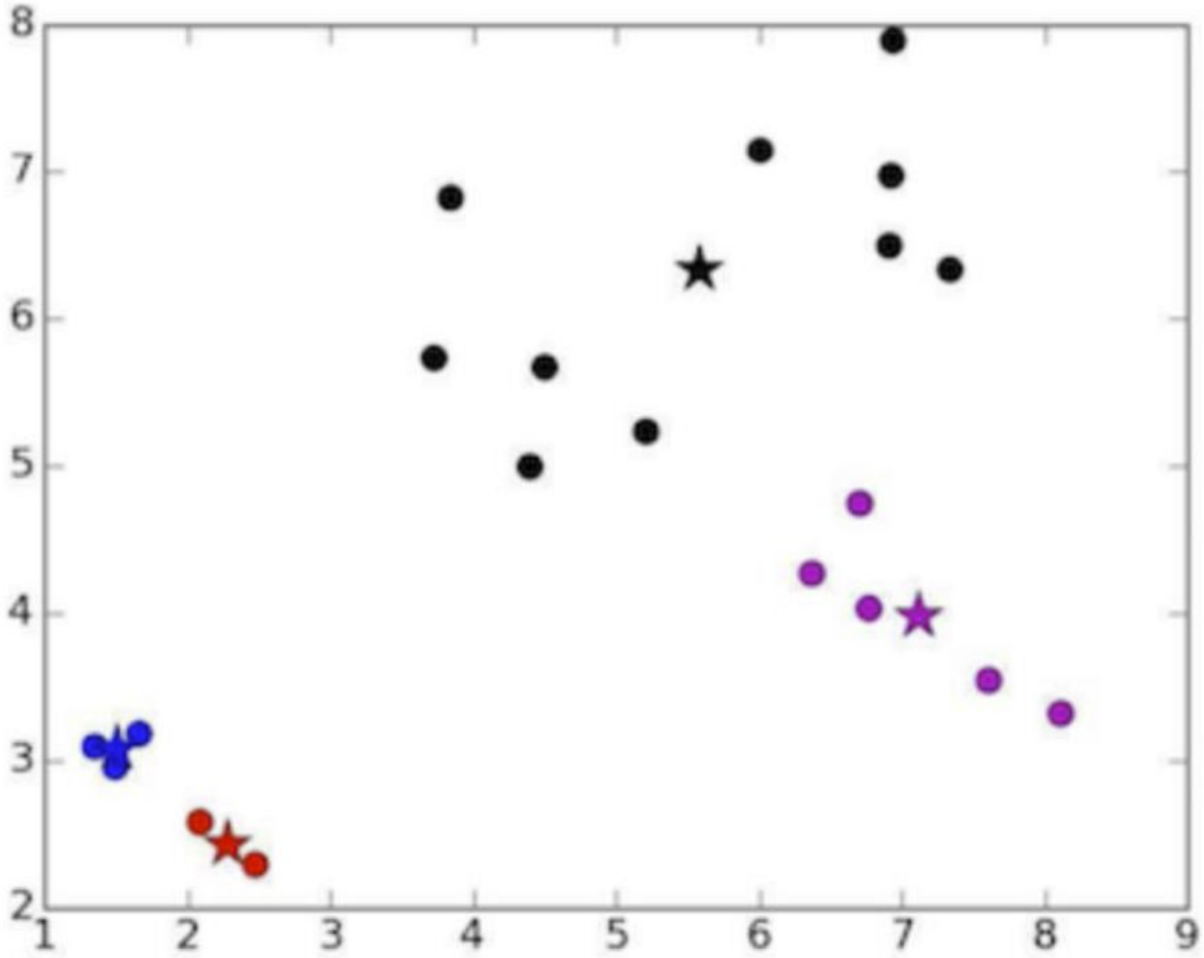
Bad Clustering



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:

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.



8
7
6
5
4
3
2
1
2
3
4
5
6
7
8
9