Unsupervised Learning

All about input,
no concept of output

Input	Output
Yes	No

NO labeled data
NO target data

Label	Target
No	No

Input data

X1	X2	X3	…	Xn

Algorithm finds pattern

X1	X2	X3	…	Xn	Pattern

We are feeding the data to unsupervised algorithm. The algorithm finds the pattern and produce the output as groups/clusters.

Applications

Sports Analytics	**Cluster pitch map**
Segmentation	**Customer**

Customer Data
Regular Customer
Occasion Customer
Rare Customer

Cluster Patterns

Medical data

Height	Weight	Age
6.5	70	30
5.5	92	37
6.0	40	18
6.4	75	29

Cluster Patterns

Medical data

Height	Weight	Age	Cluster
6.5	70	30	2
5.5	92	37	3
6.0	40	28	1
6.4	75	25	2

Algorithm generates pattern and produces 3 clusters

Employee data

Age	Salary	Phone no.	Qualification	Experience
30	45000	9840012345	B.Tech	5
40	80000	8917364748	B.E	14
20	20000	7874498394	B.Sc	1
31	51000	9955585803	B.Tech	6

Clustering algorithms are used to find similarities between the data

Employee data

Age	Salary	Phone no.	Qualification	Experience	Cluster
30	45000	9840012345	B.Tech	5	2
40	80000	8917364748	B.E	14	1
20	20000	7874498394	B.Sc	1	3
31	51000	9955585803	B.Tech	6	2

Clustering algorithms are used to find similarities between the data

Sample Results with 5 clusters

Pattern 1	Pattern 2	Pattern 3	Pattern 4	Pattern 5
5000 records	10000 records	20000 records	5000 records	10000 Records
Age 80 Sal 10000	Age 60 Sal 20000	Age 70 Sal 25000	Age 20 Sal 30000	Age 35 Sal 50000

Example

A company wants to introduce a new branded product in the market.

Planning to do a marketing campaign

There are **10 million** customers

3 channels for the campaign are used

Channels	Usage per Cost
Letter	$1
Email	$1
Phone call	$1

Cost is **$3** per customer on launching a new product

Selling price is **$10,000** per product

Scenario 1:

Assume that **10,000** customers are going to buy the product

Marketing Expense	10M customer * $3	30M	(-)
Sales Revenue	10,000 customer * $10,000	100M	(+)
Net		70M	

Total Revenue after marketing expense is **70 million dollars**

Scenario 2: Data Scientist involvement

Performs segmentation and provides the report

Segments	Total Customers
Low value customer (Budget Buyers)	4M
Mid range customer (Bargain Hunters)	3M
High value customer (Occasional Explorers)	2M
High premium customer (Loyal Luxury Seekers)	1M

Decision:

Concentrate on High premium loyal customer

Assume **9,000** customers are going to buy the product

Marketing Expense	1M customer * $3	$3M	(-)
Sales Revenue	9,000 customer * $10000	$90M	(+)
Net		$87M	

Total Revenue after marketing expense is **87 million dollars**

Final Result:

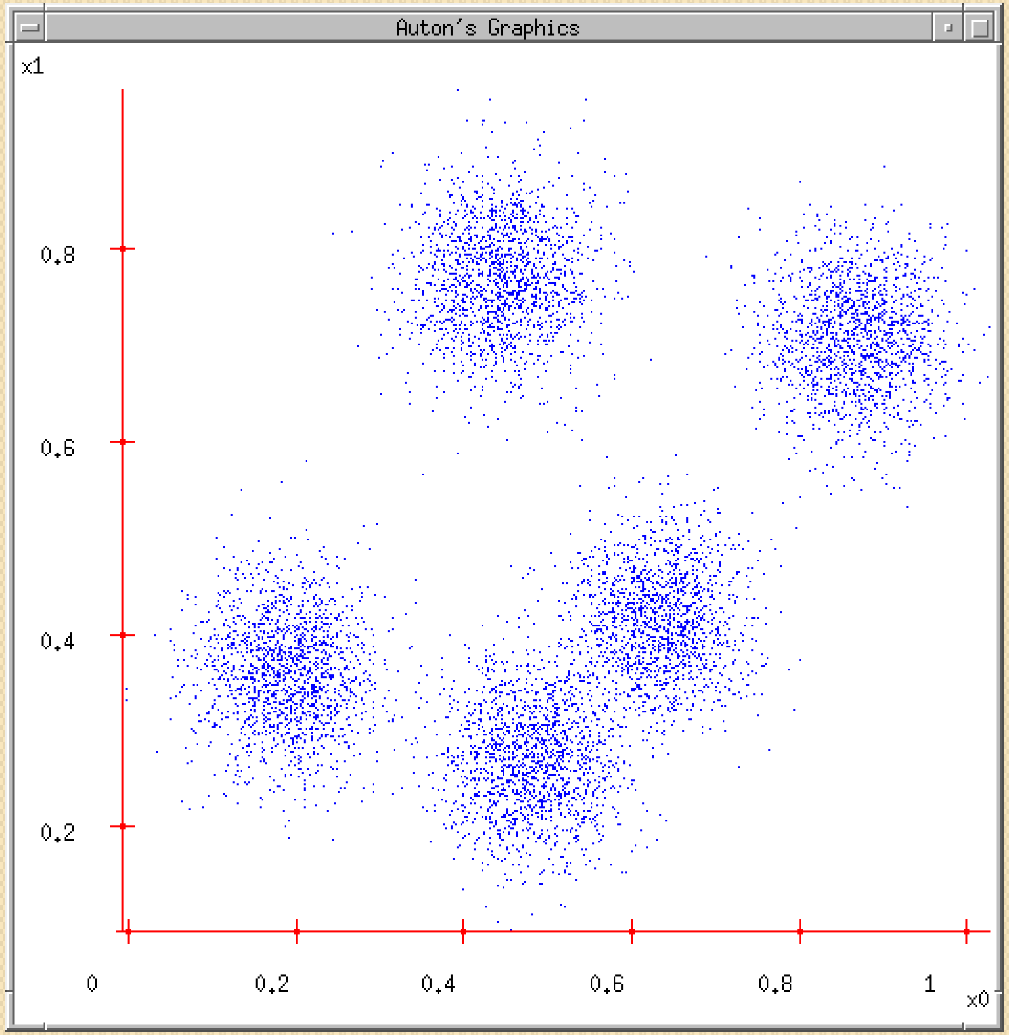
Expense reduced from 30M to 3M as well as resources, time utilization gets reduced

Profit increased by **17 million dollars**

Clustering in ML

k-Means Clustering

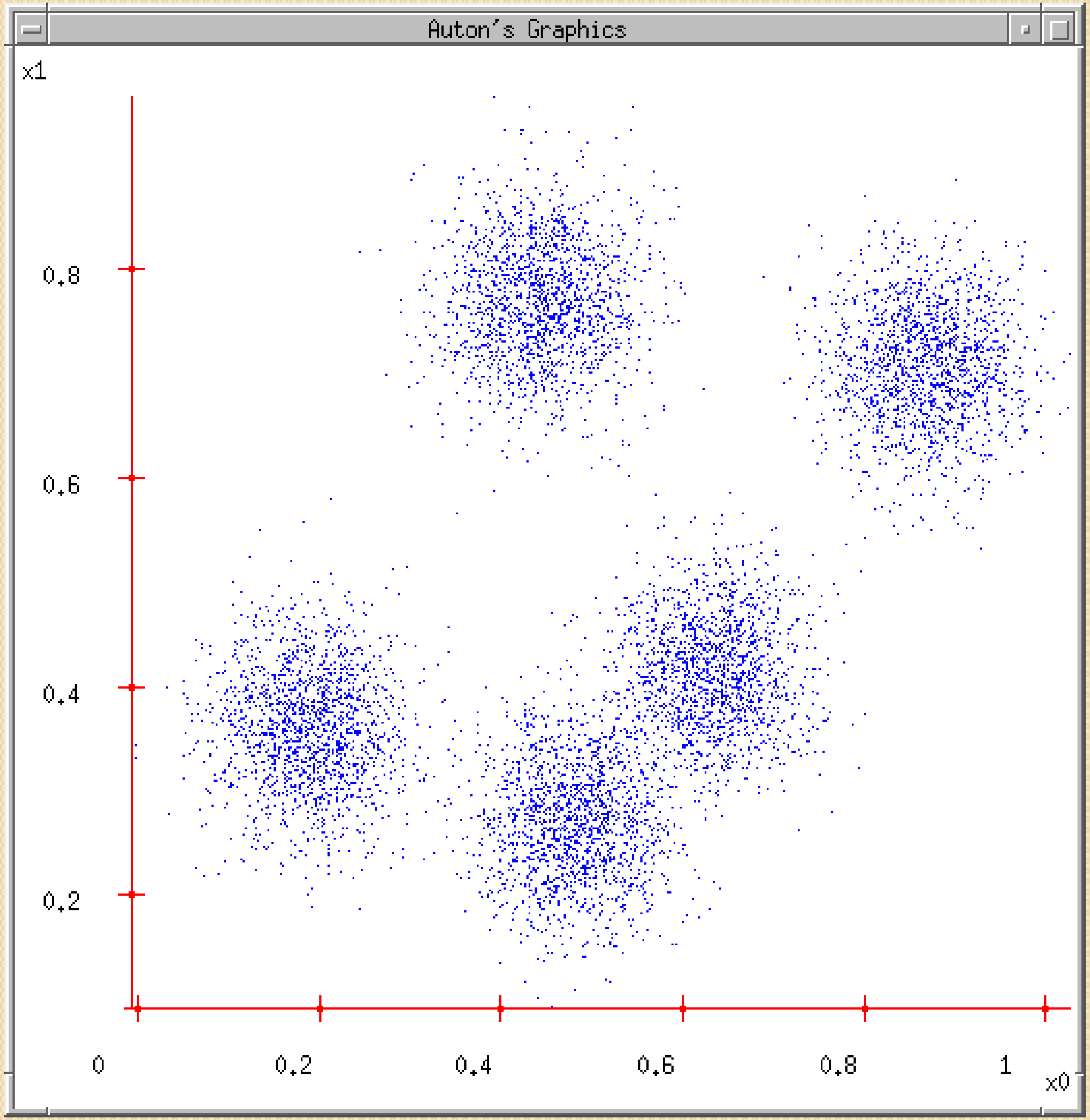
Data



K-means

1. Ask how many clusters?

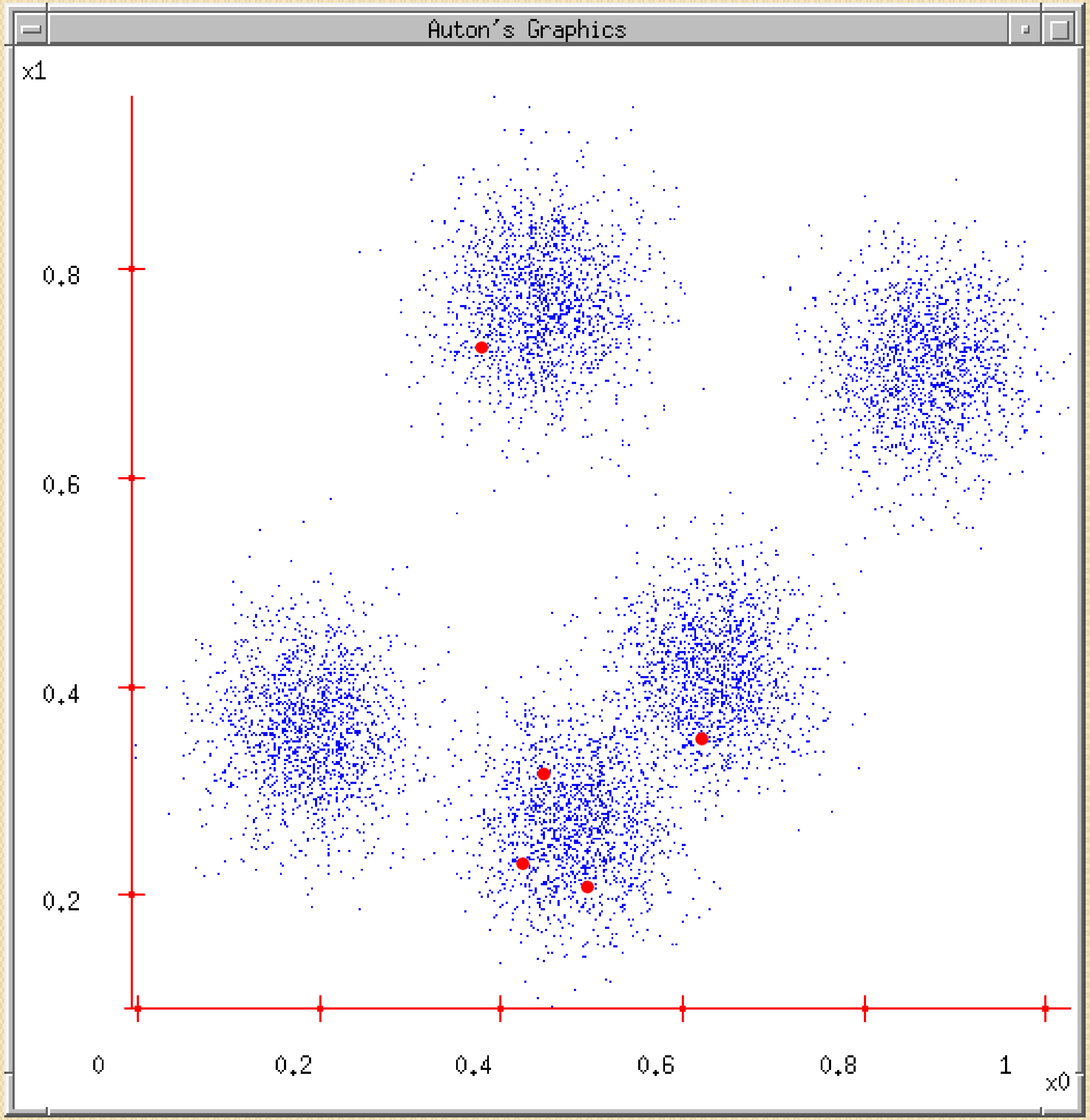
   (e.g. k=5)



K-means

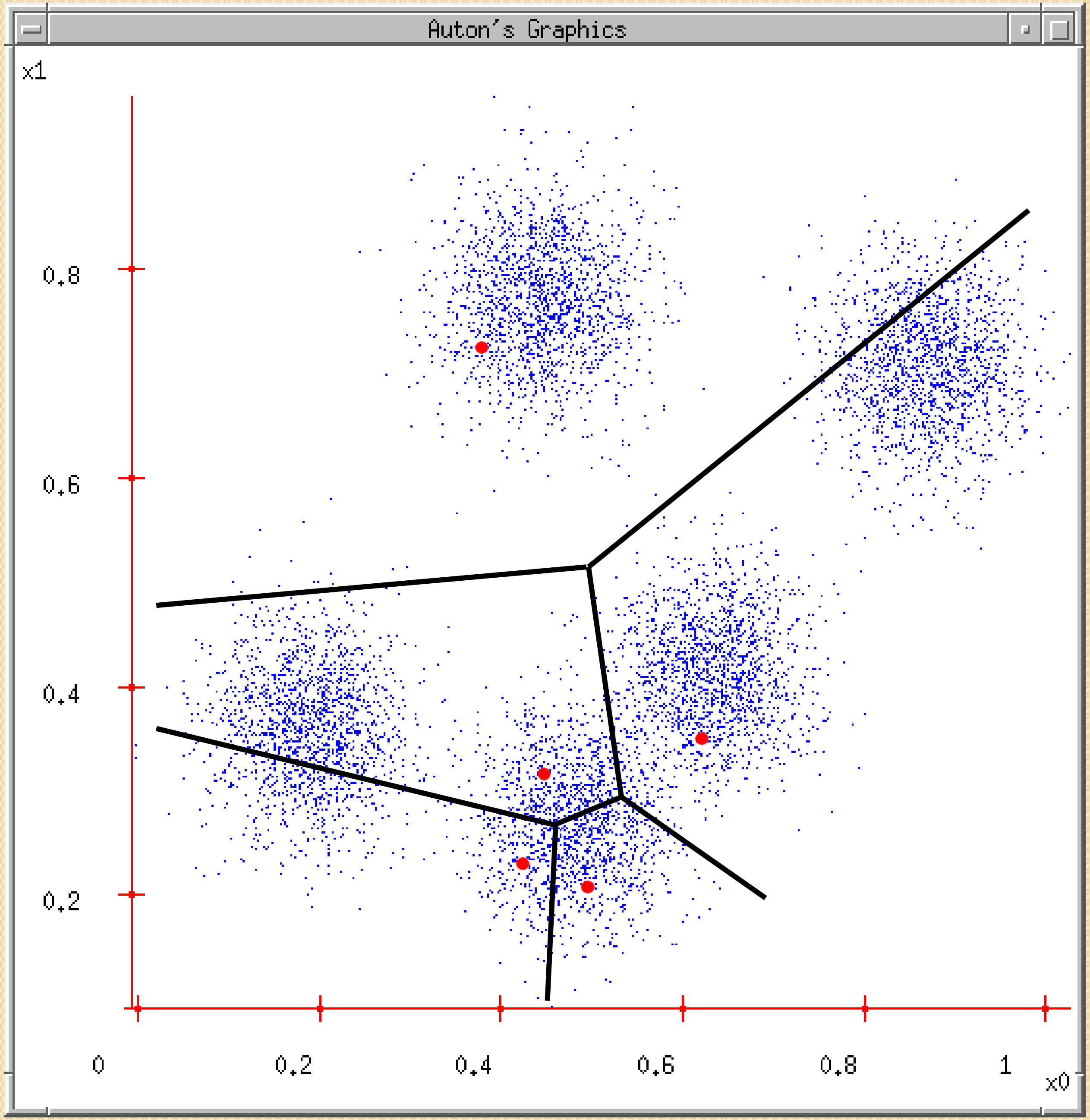
1. Ask how many clusters ? *(e.g. k=5)*

2. Randomly guess k cluster Center locations



K-means

1. Ask how many clusters ? *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.



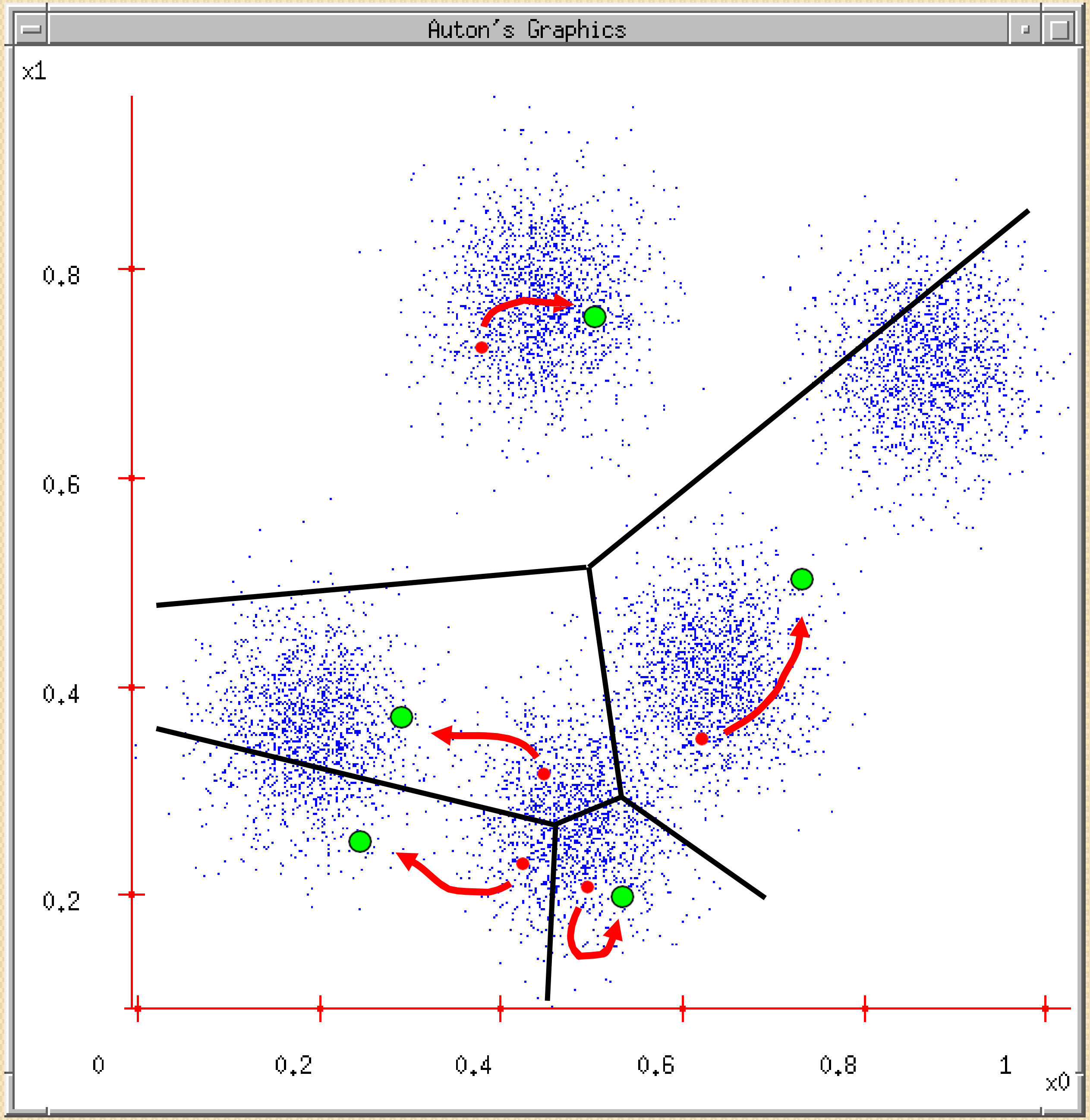
K-means

1. Ask how many clusters? *(e.g. k=5)*

2. Randomly guess k cluster Center locations

3. Each datapoint finds out which Center it’s closest to.

4. Each Center finds the centroid of the points it owns



K-means

1. Ask how many clusters? *(e.g. k=5)*

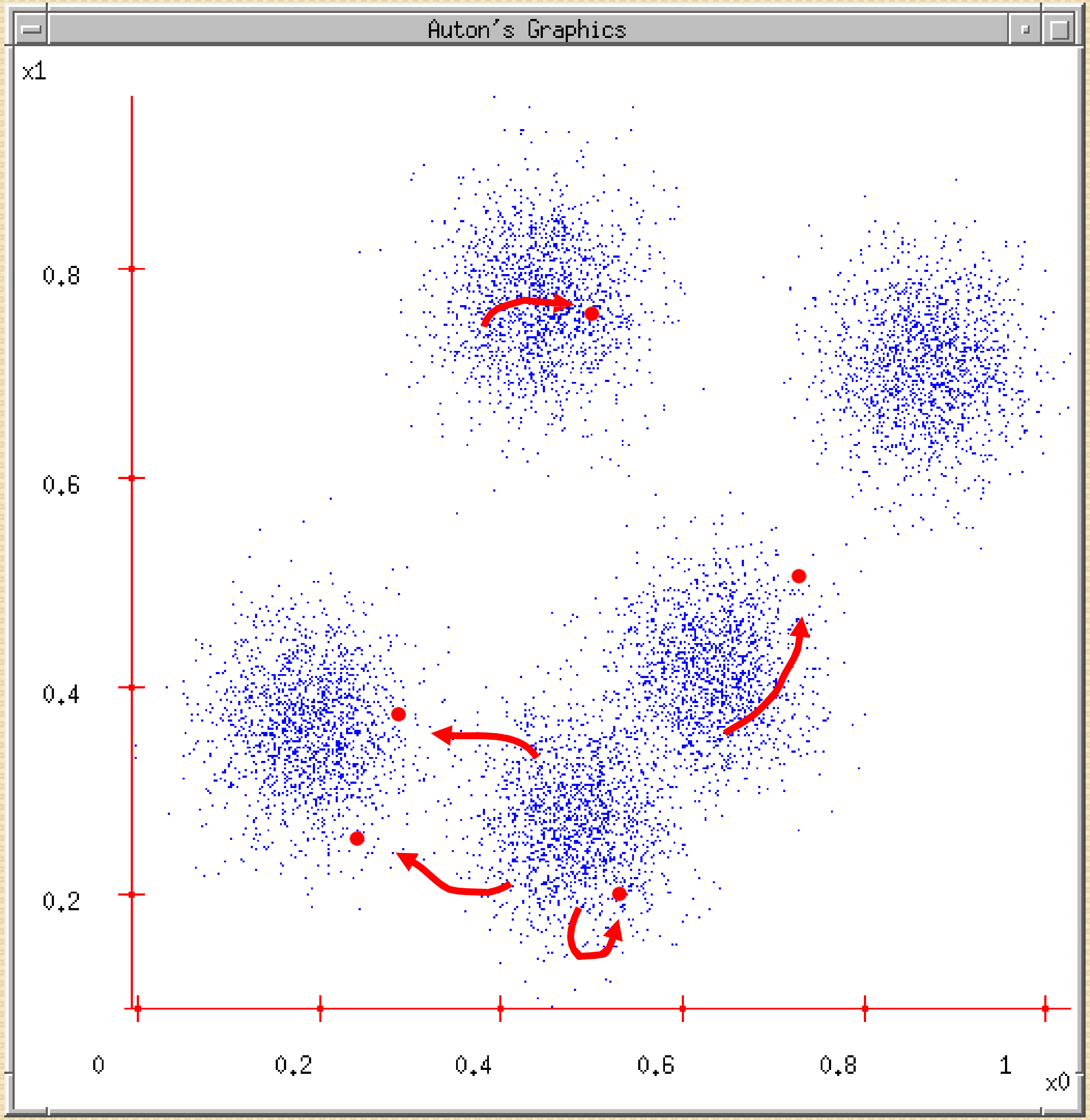
2. Randomly choose k Center points

3. Each datapoint finds out which Center it's closest to.

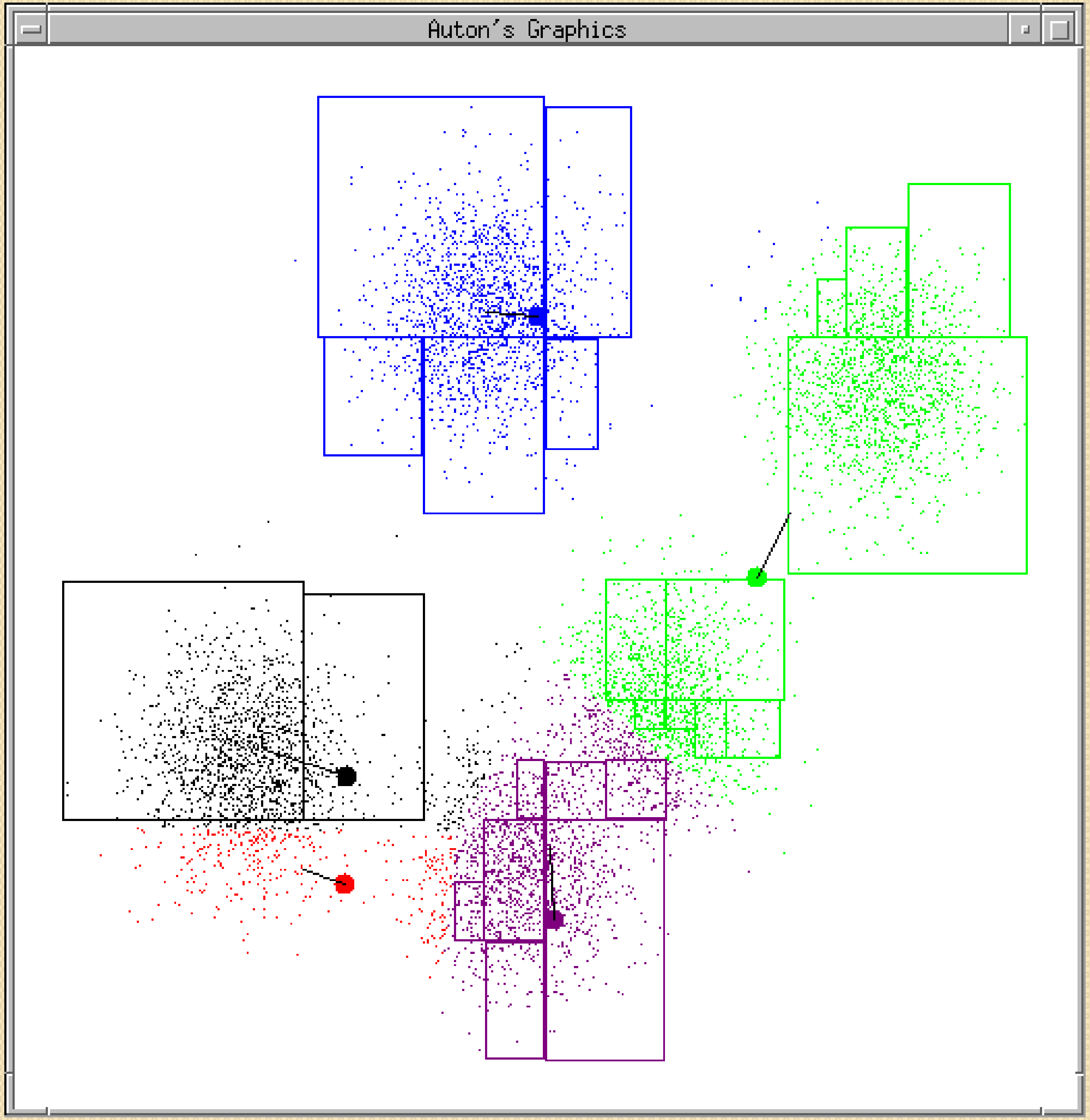
4. Each Center finds the centroid of the points it owns…

5. Move to the new position

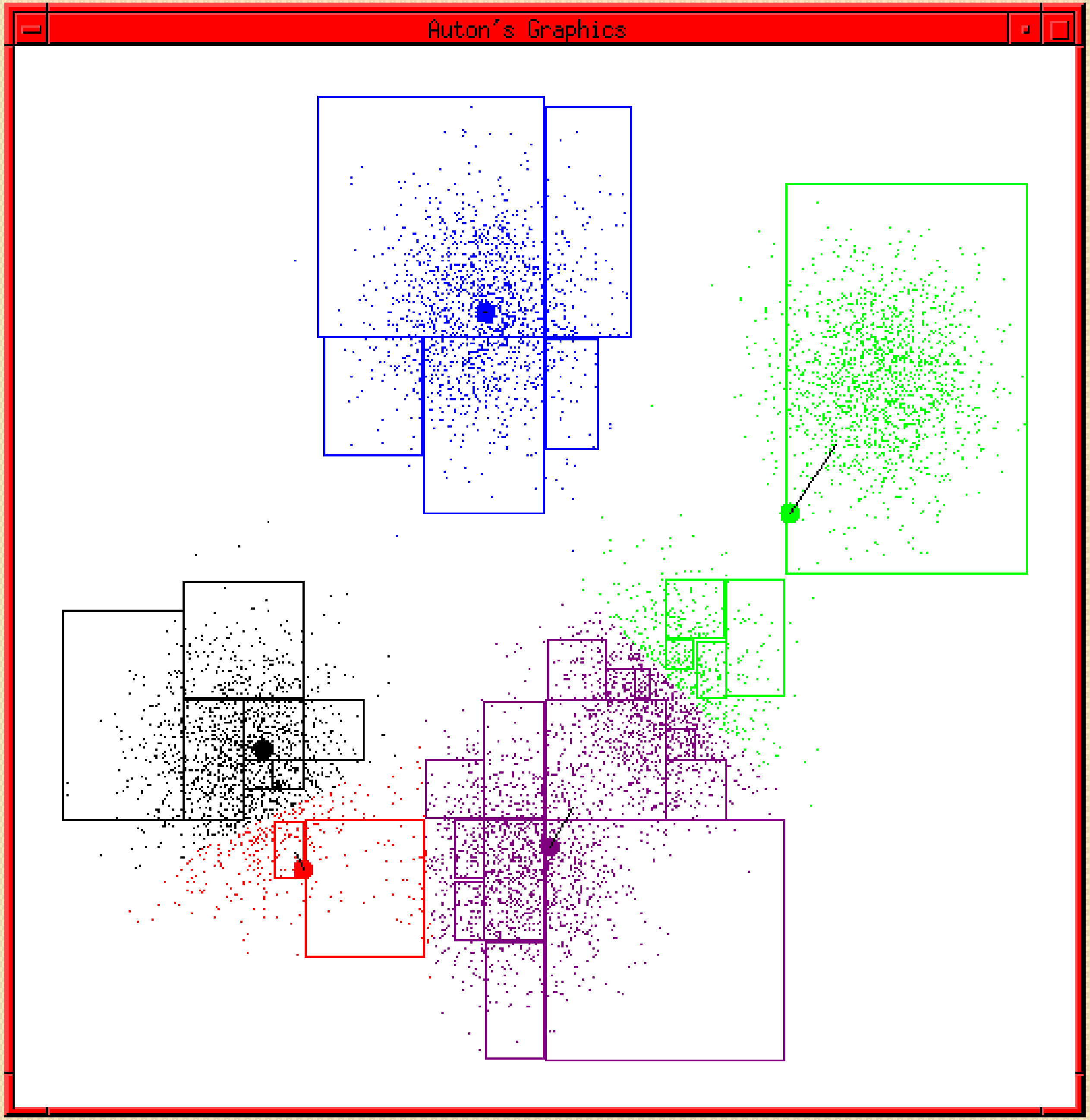
6. Repeat



K-means continues …

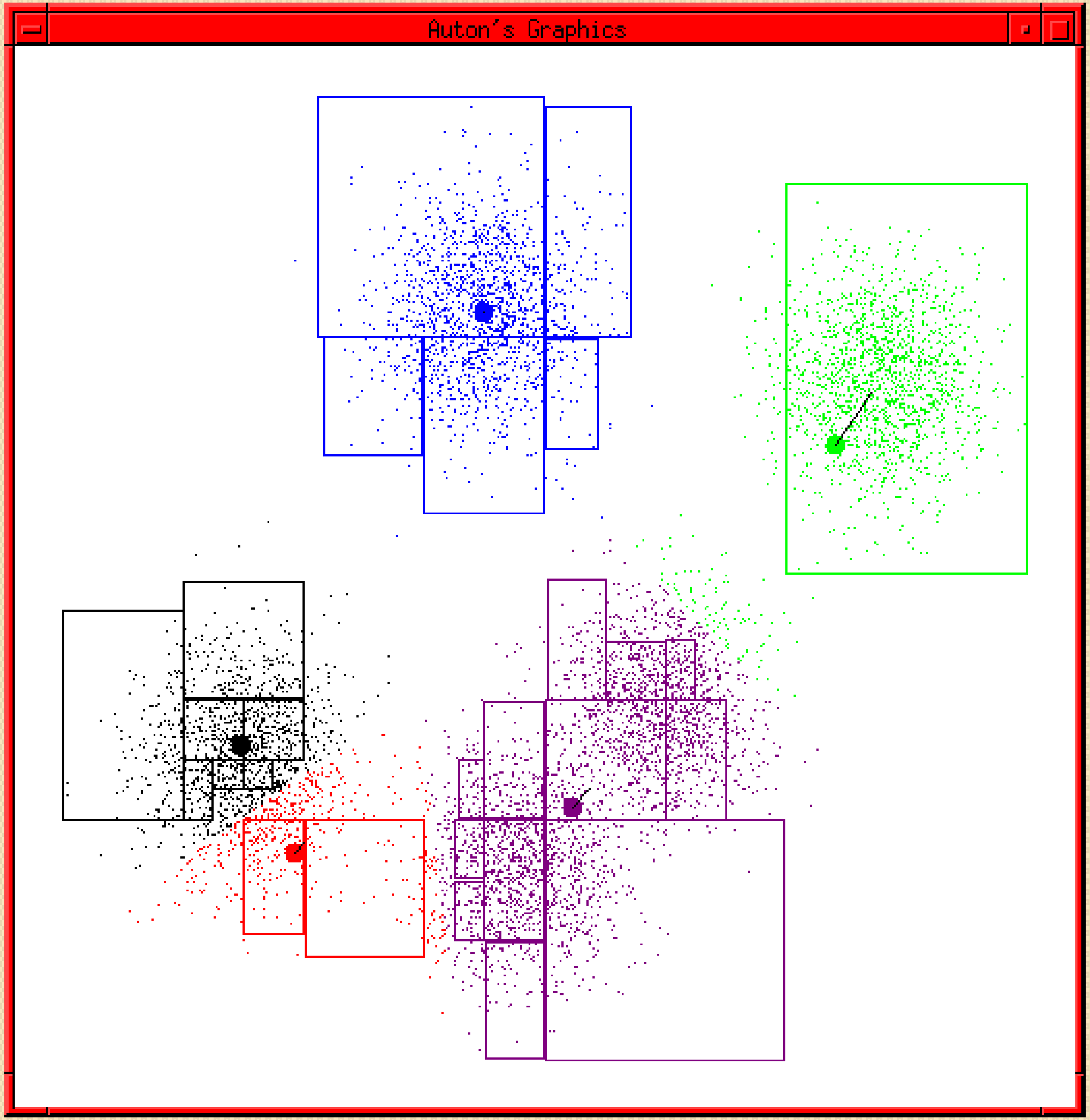


K-means continues …

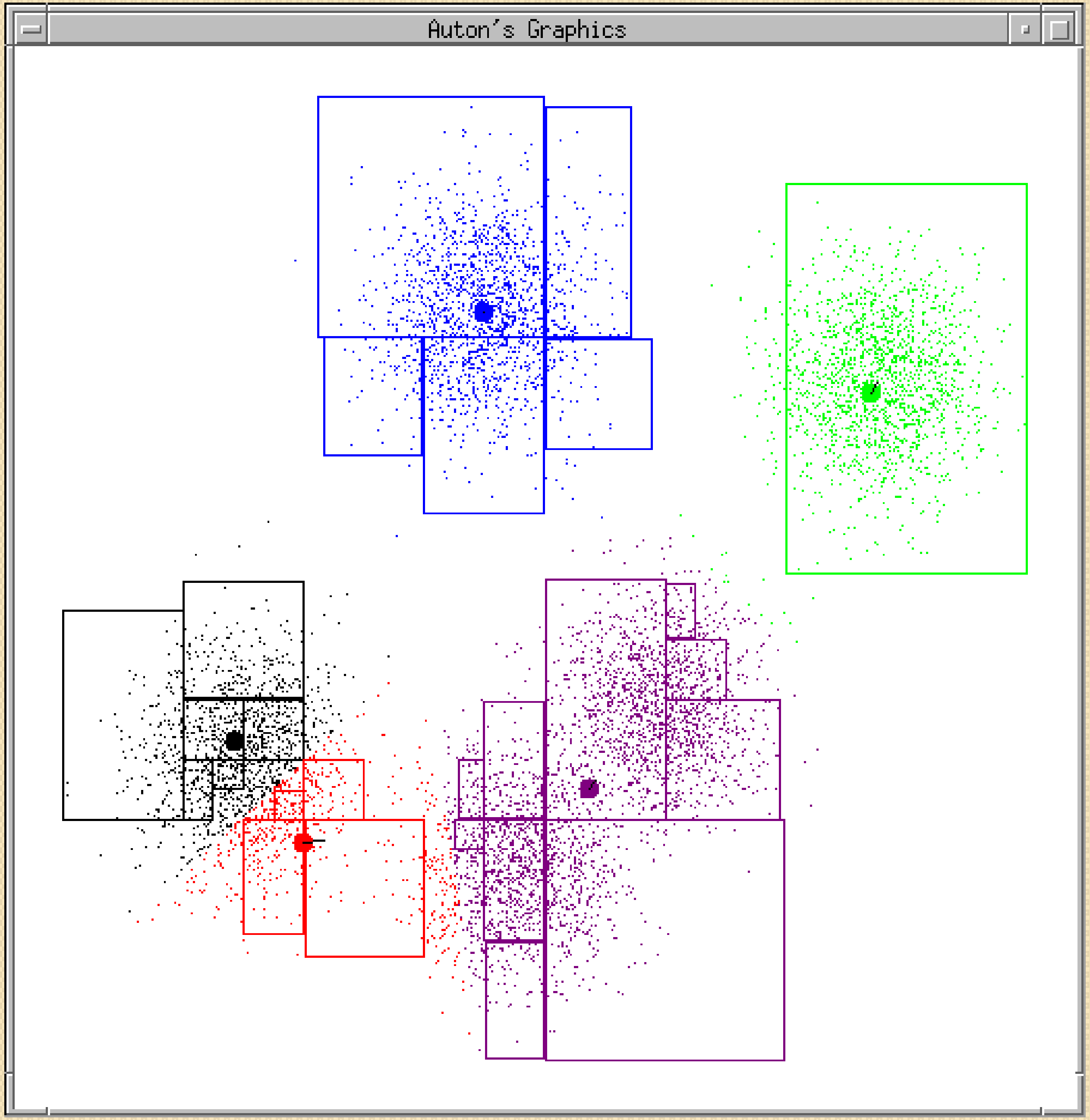


K-means continues

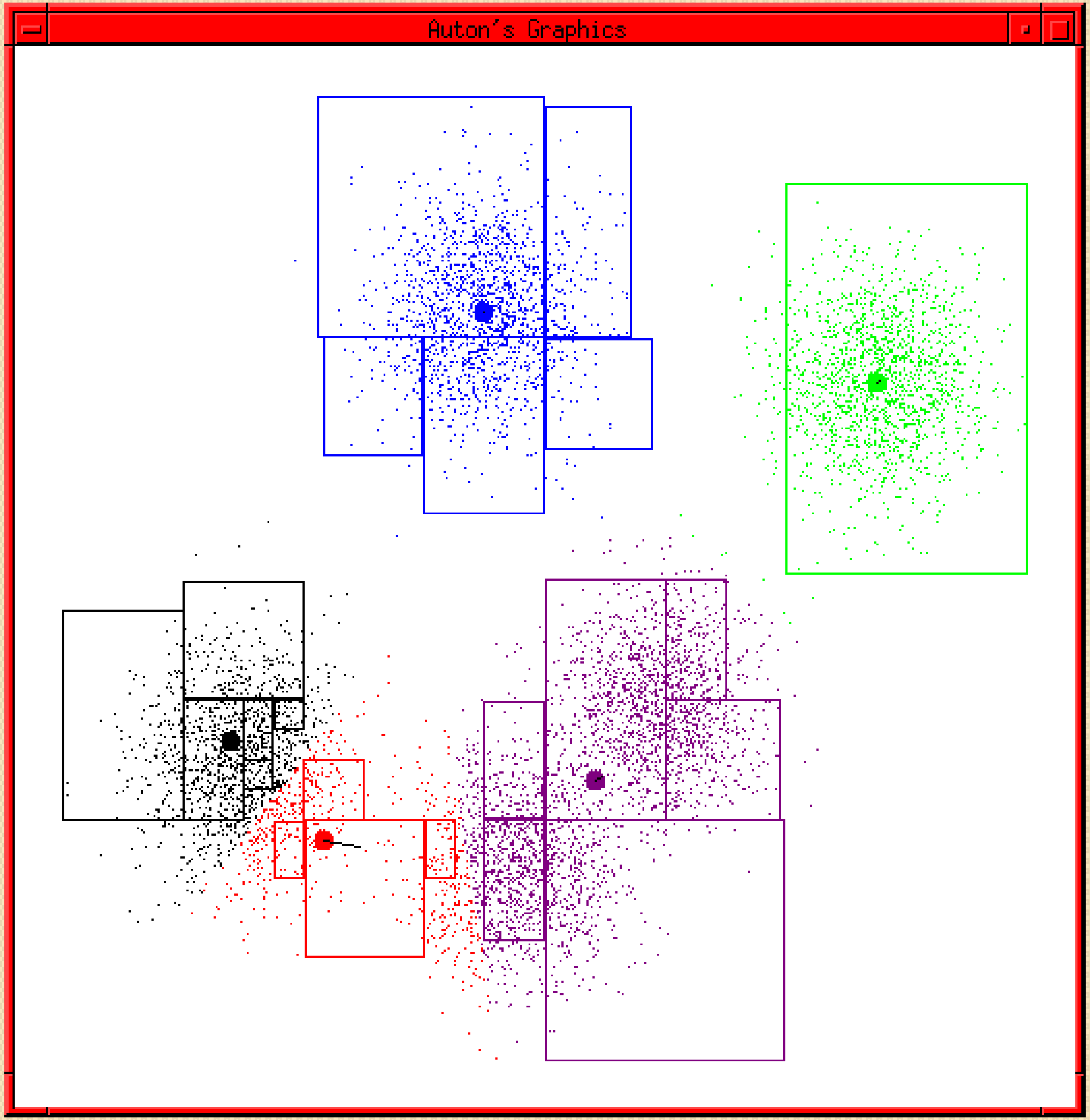
…



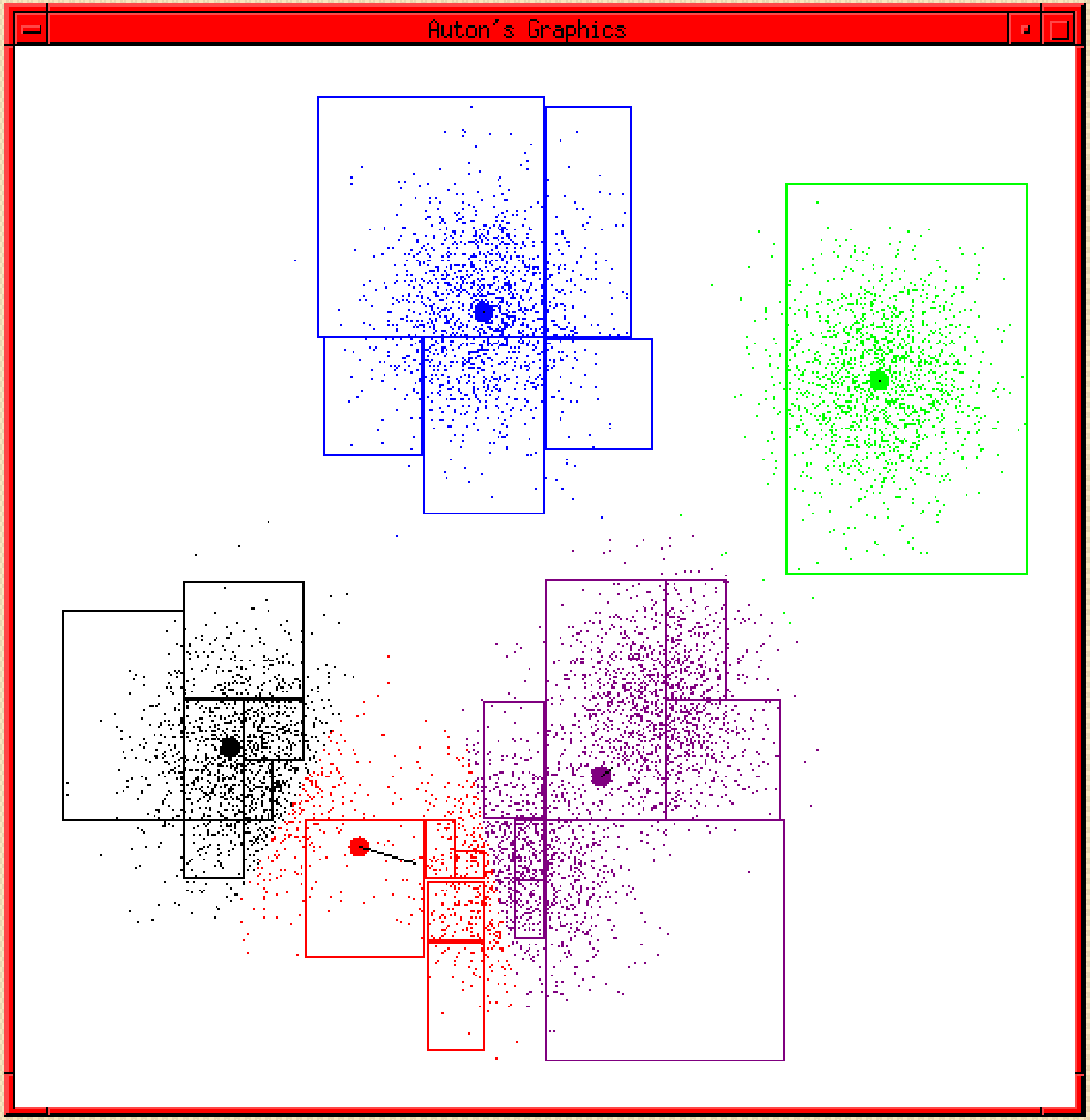
K-means continues …



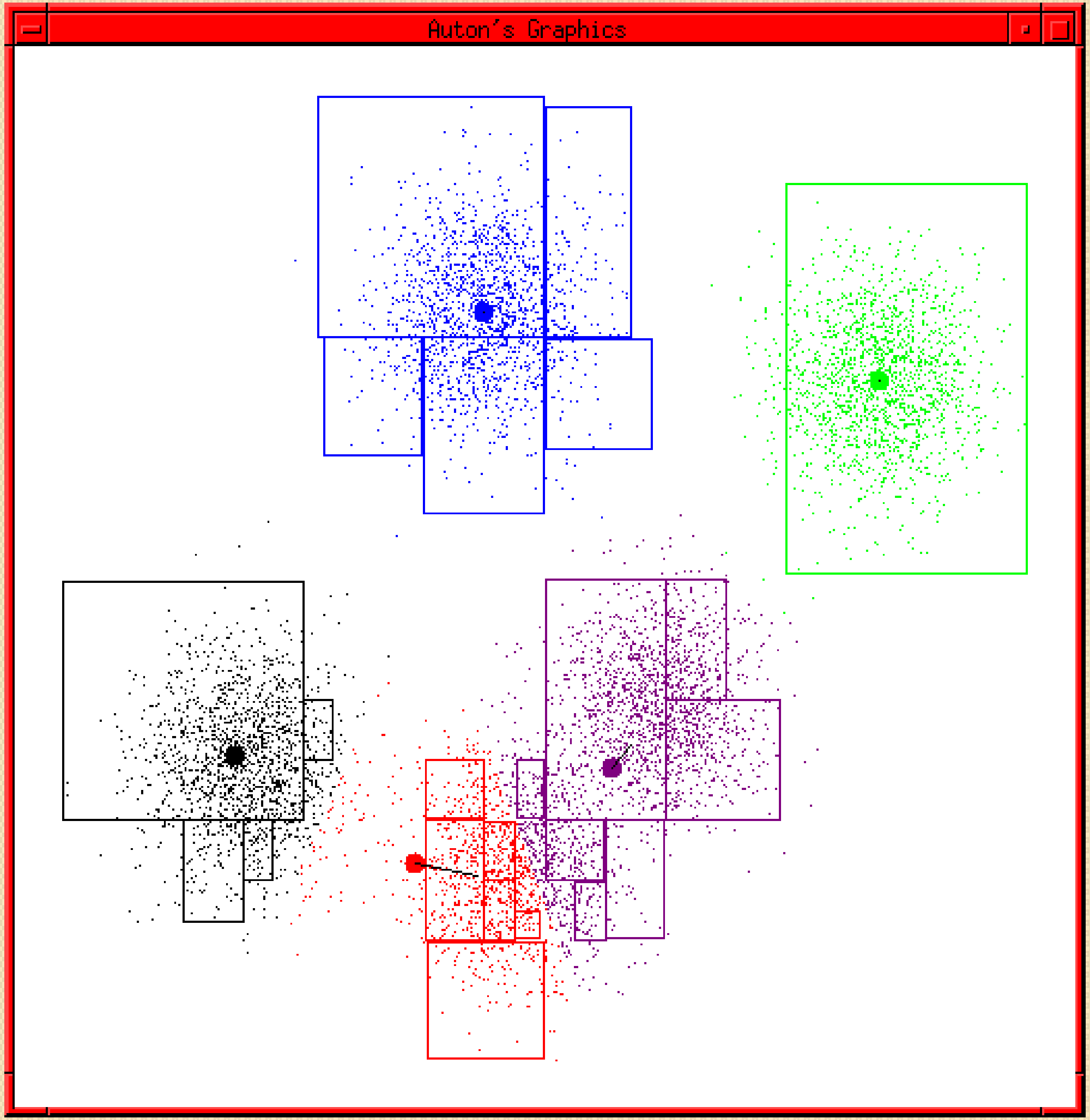
K-means continues …



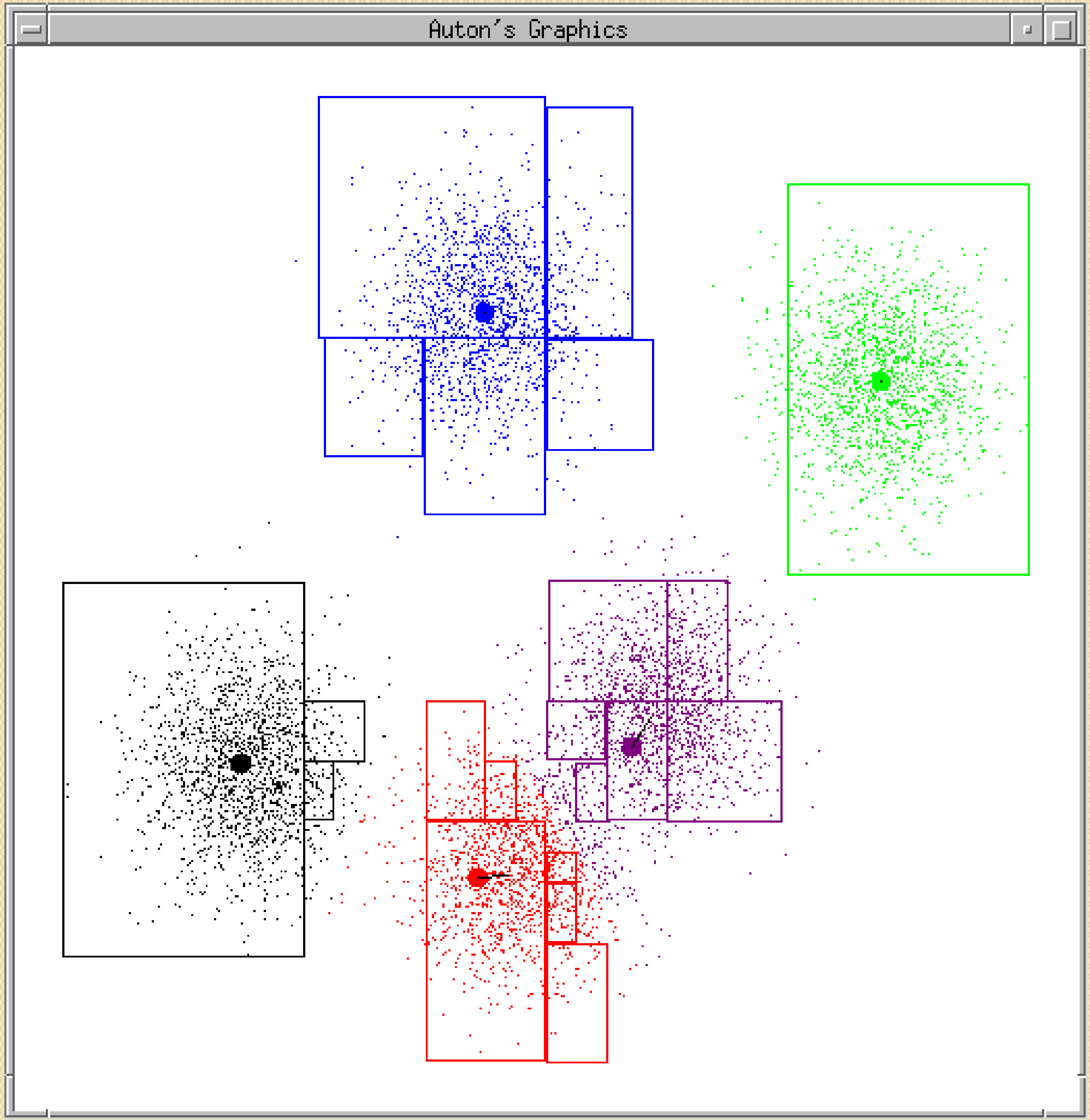
K-means continues …



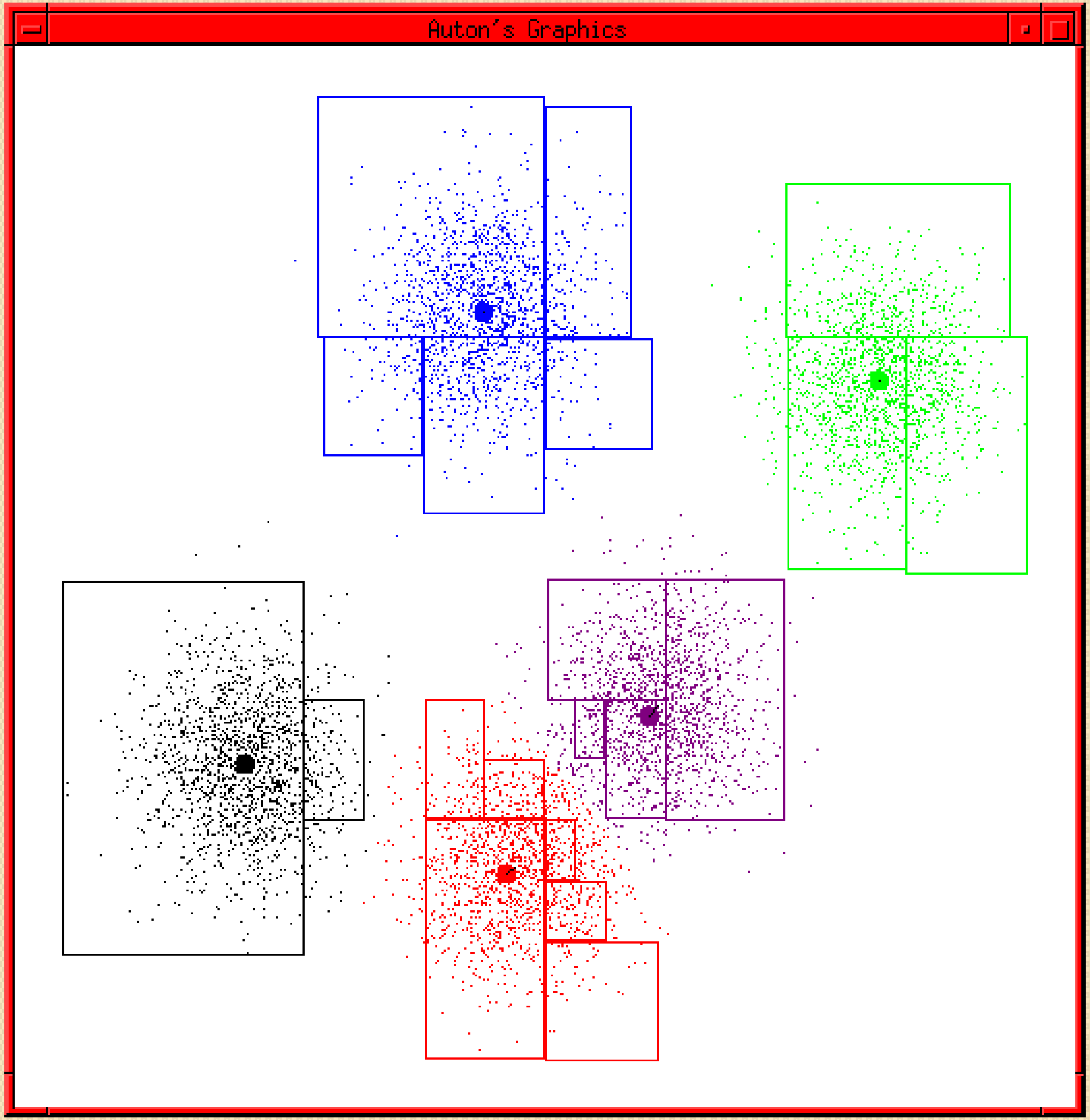
K-means continues …



K-means continues …



K-means stops



Steps

Step 1: Randomly choose 5 centroid points since k=5

Step 2: Find distance between each centroid and all the data points, then choose the nearest centroid

Step 3: Move to the new centroid

Step 4: With new centroid, perform Step 2

Step 5: Repeat centroid finding and repeat Step 2

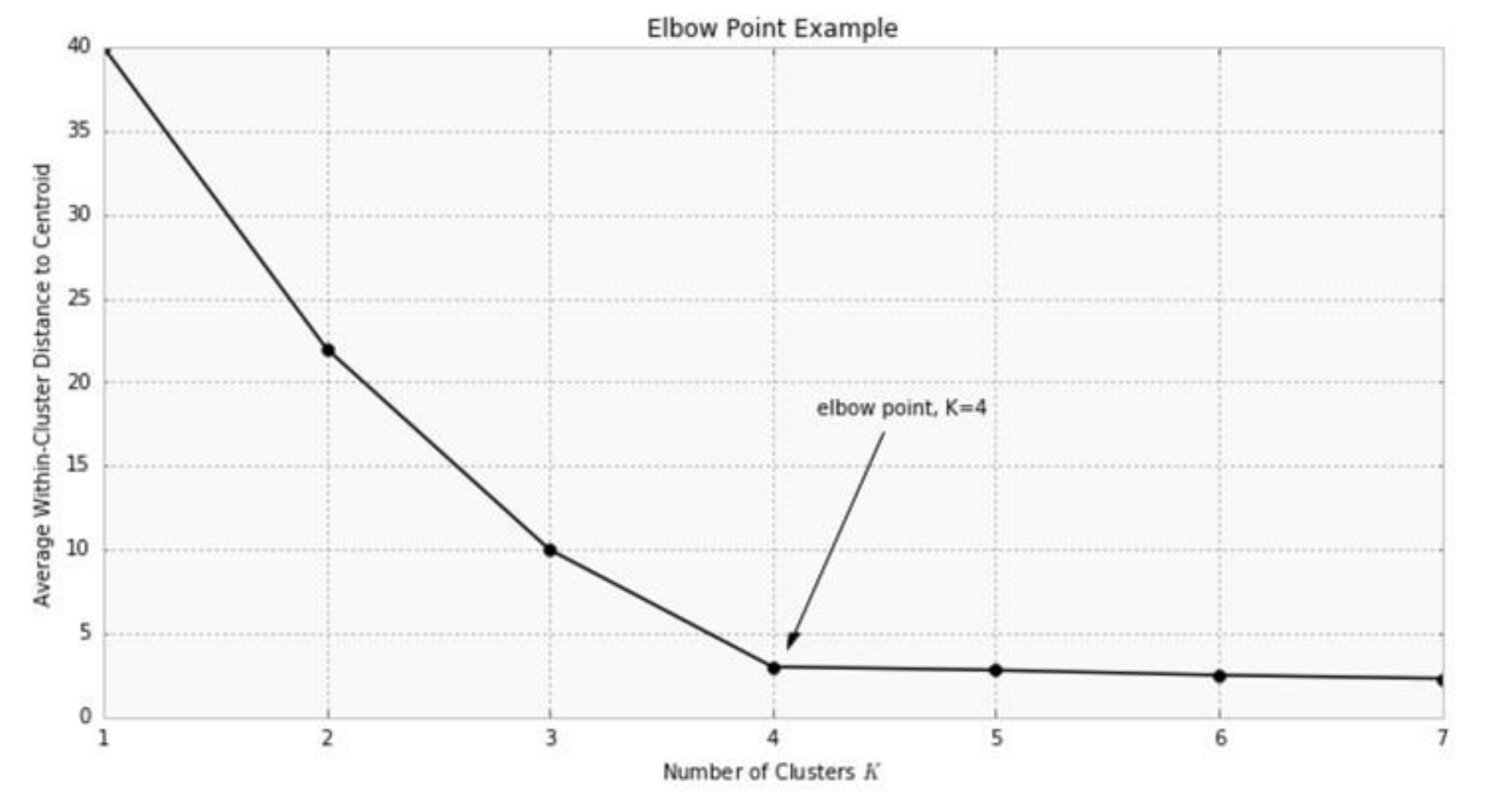
Drawbacks:

Choosing the number of clusters

Solution:

Elbow Method

Choosing K


Elbow Point Example
Average Within-Cluster Distance to Centroid
0
5
10
15
20
25
30
35
40
Number of Clusters K
1
2
3
4
5
6
7
elbow point, K=4

Drawbacks:

Based on initial data points you choose randomly, there is a uncertainty in the formation of cluster

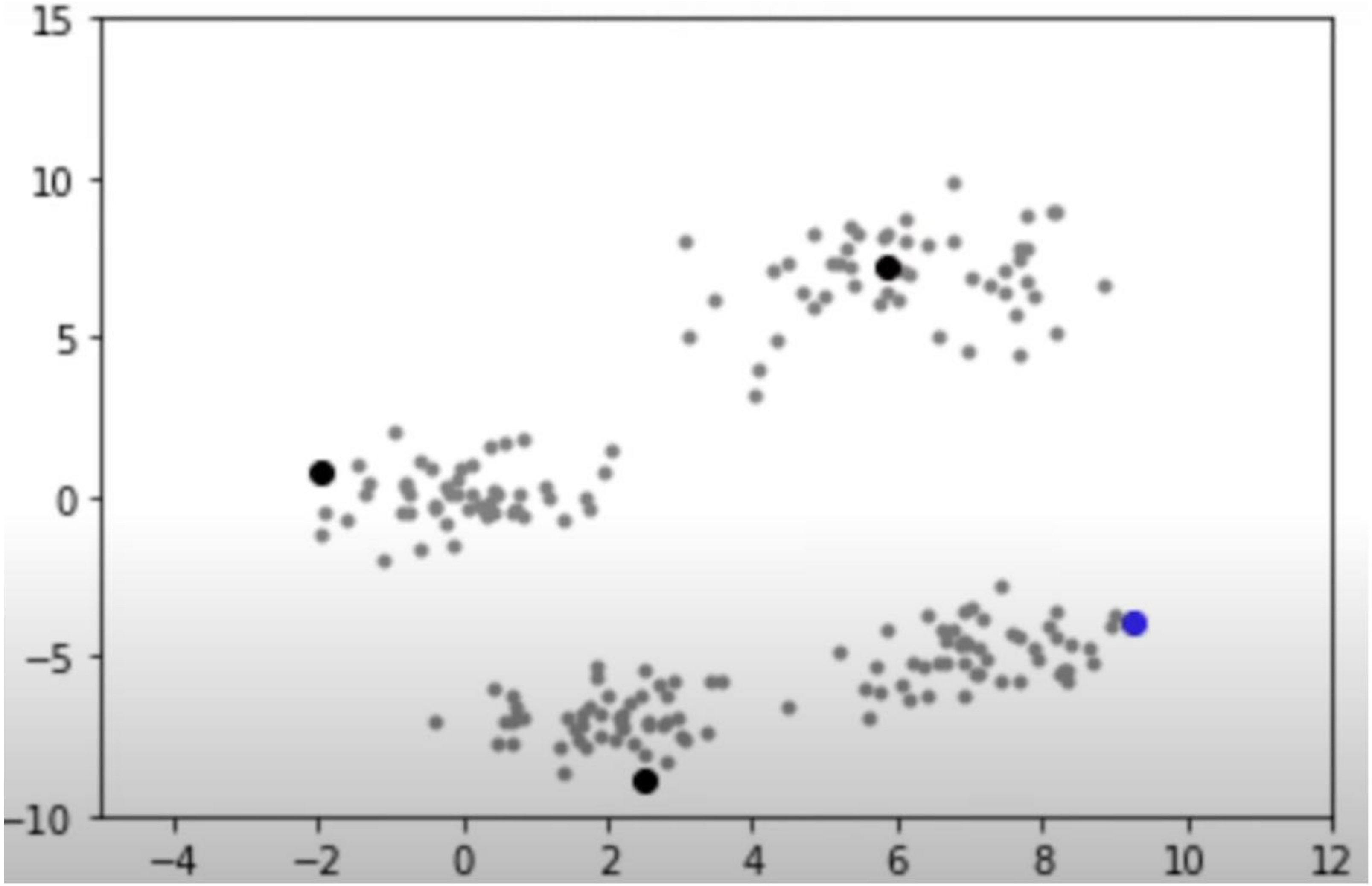
Solution:

K-Means++

K-Means++

First point is always random

But the second and foremost data point is orthogonally far from the first point chosen and so on.


15
10
5
0
−5
−10
−4
−2
0
2
4
6
8
10
12