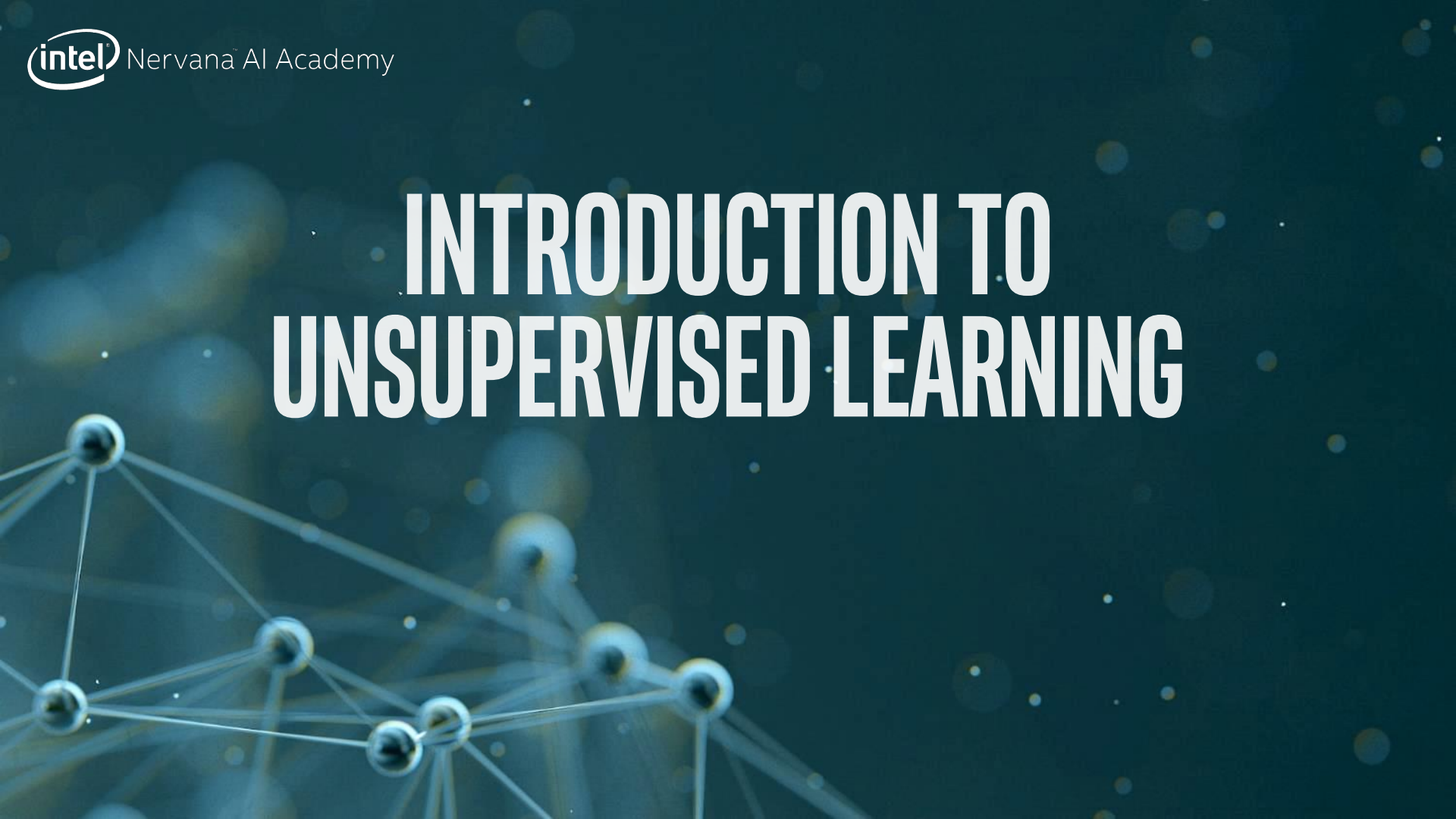

intel Nervana AI Academy
INTRODUCTION TO
UNSUPERVISED LEARNING

TYPES OF MACHINE LEARNING

SUPERVISED — Data points have known outcome

UNSUPERVISED — Data points have unknown outcome

TYPES OF MACHINE LEARNING

SUPERVISED — Data points have known outcome

UNSUPERVISED — Data points have unknown outcome

TYPES OF UNSUPERVISED LEARNING

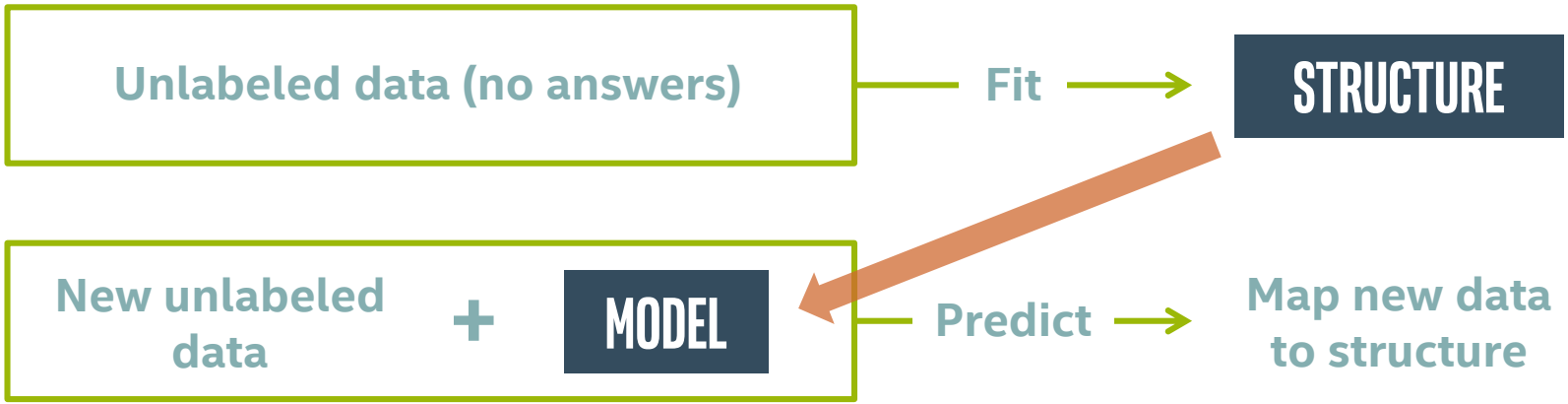
CLUSTERING

Identify unknown structure in data

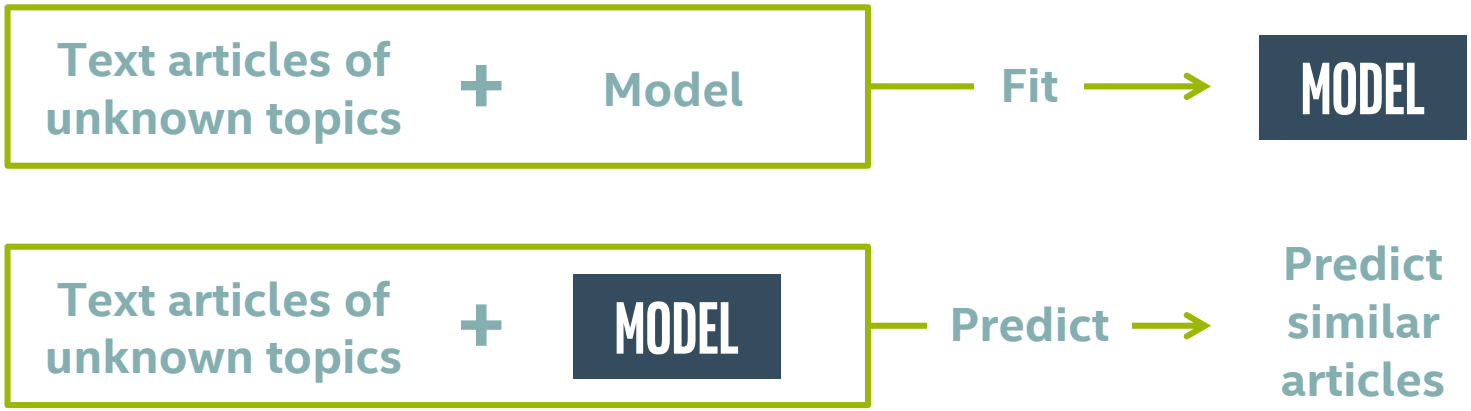
TYPES OF UNSUPERVISED LEARNING

CLUSTERING	Identify unknown structure in data
DIMENSIONALITY REDUCTION	Use structural characteristics to simplify data

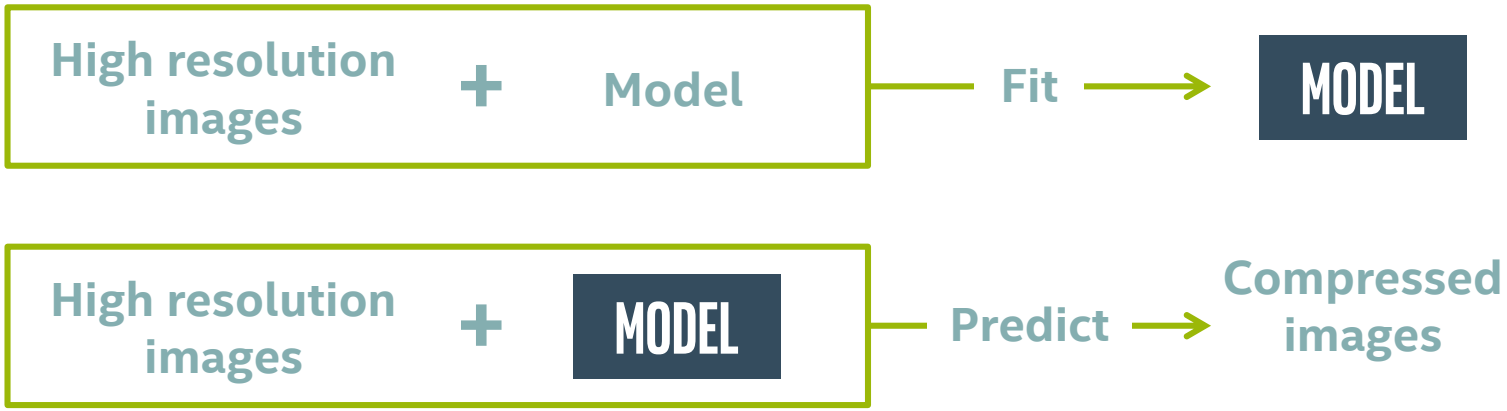
UNSUPERVISED LEARNING OVERVIEW



CLUSTERING: FINDING DISTINCT GROUPS



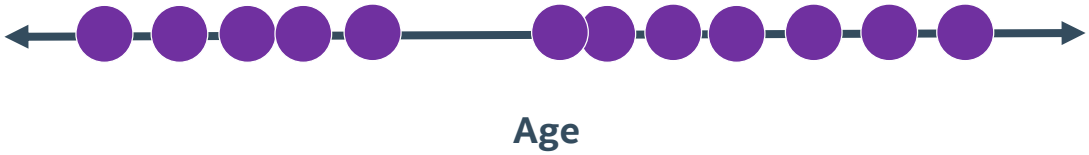
DIMENSIONALITY REDUCTION: SIMPLIFYING STRUCTURE



INTRODUCTION TO UNSUPERVISED LEARNING

Users of a web application:

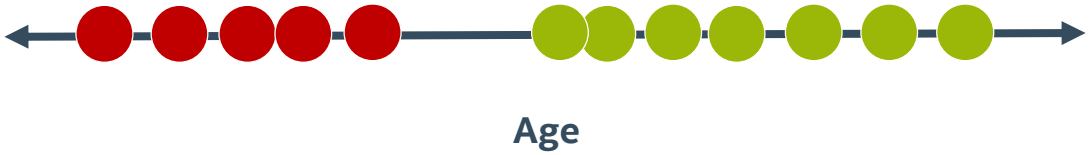
- One feature (age)



INTRODUCTION TO UNSUPERVISED LEARNING

Users of a web application:

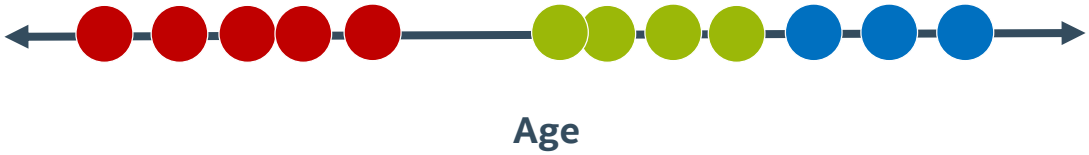
- One feature (age)
- Two clusters



INTRODUCTION TO UNSUPERVISED LEARNING

Users of a web application:

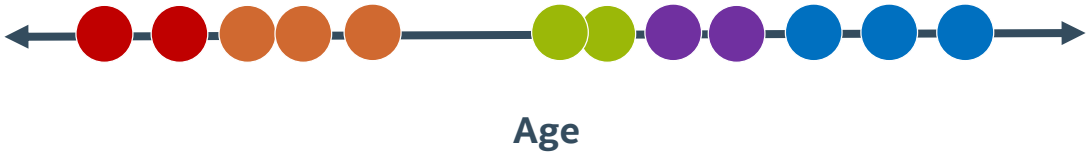
- One feature (age)
- Three clusters



INTRODUCTION TO UNSUPERVISED LEARNING

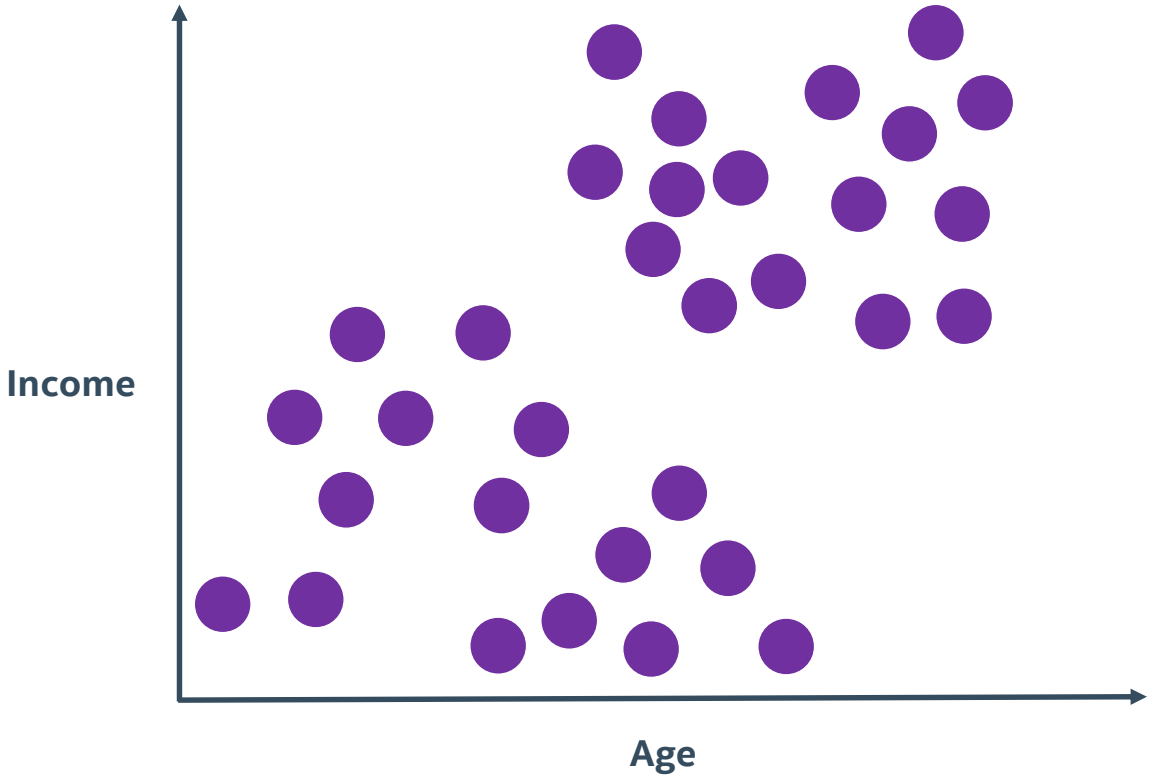
Users of a web application:

- One feature (age)
- Five clusters



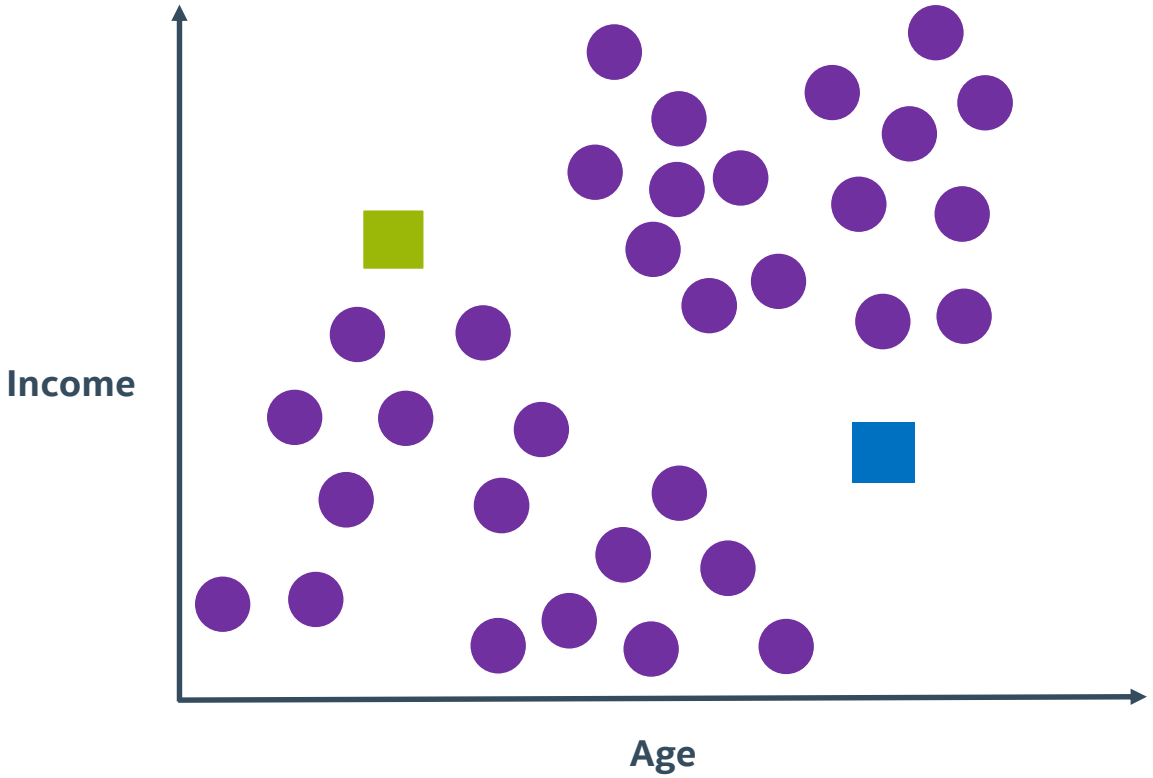
K-MEANS ALGORITHM

K = 2 (find two clusters).



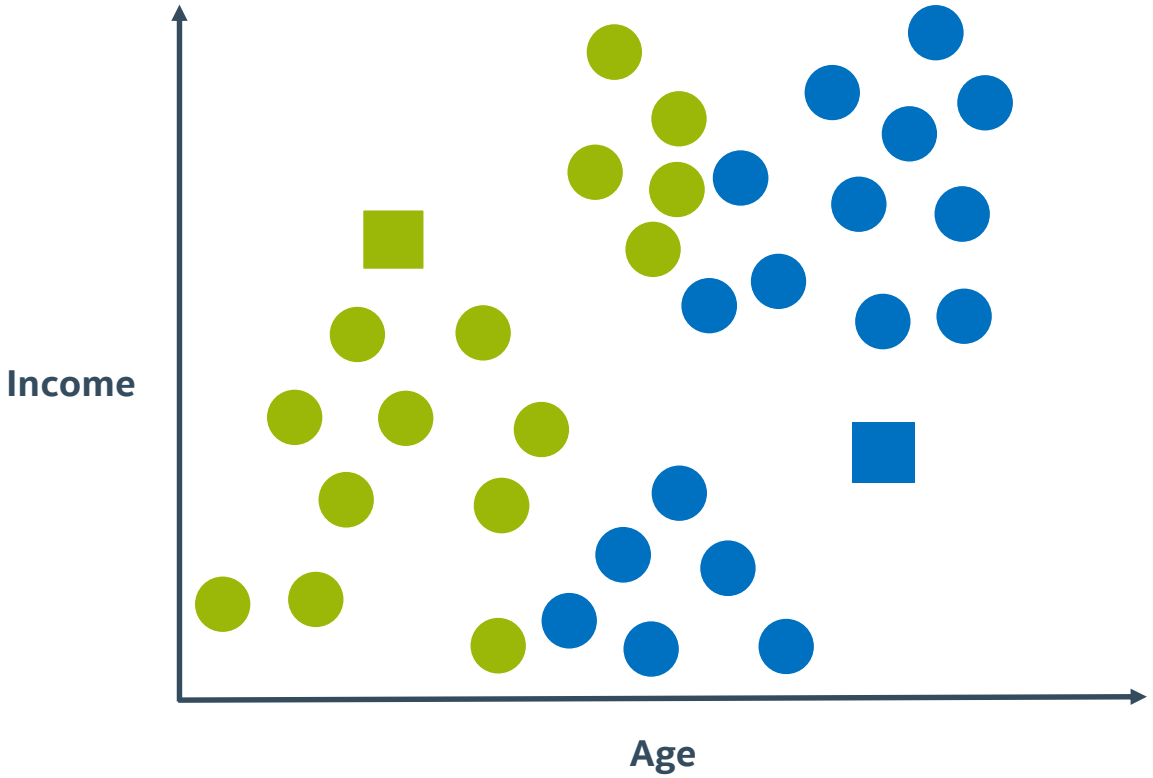
K-MEANS ALGORITHM

K = 2, Randomly assign cluster centers.



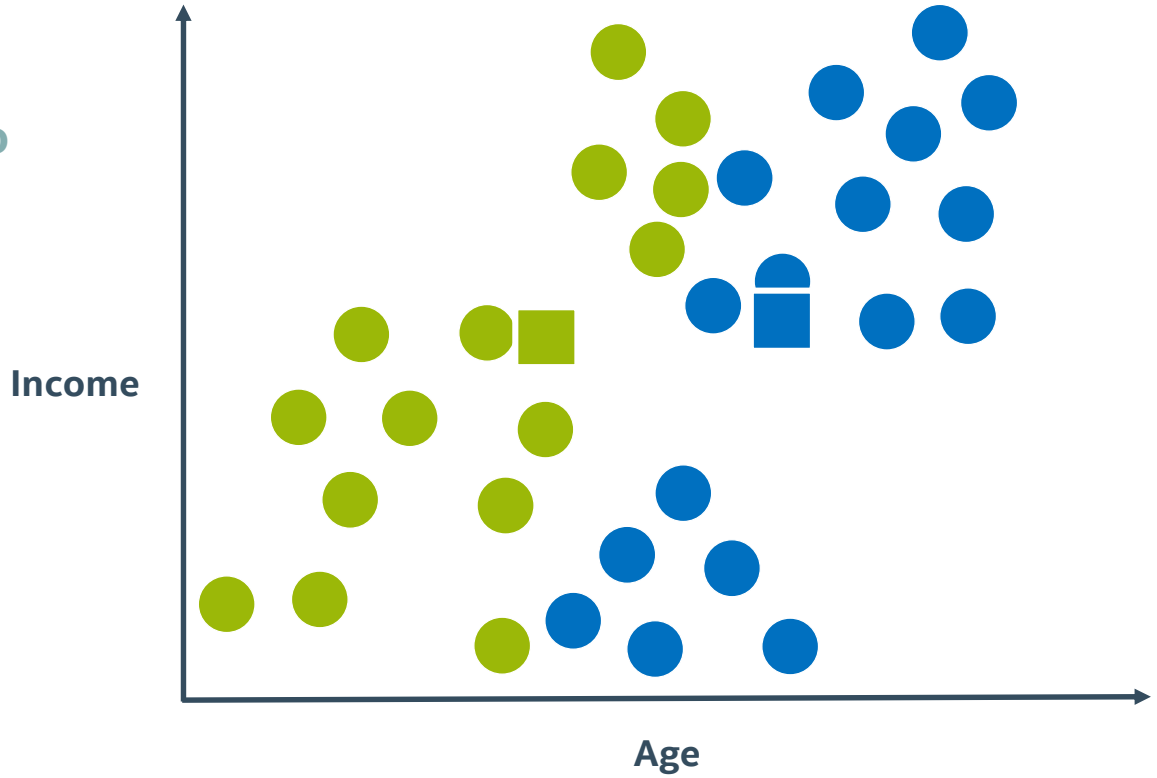
K-MEANS ALGORITHM

K = 2, Each point belongs to closest center.



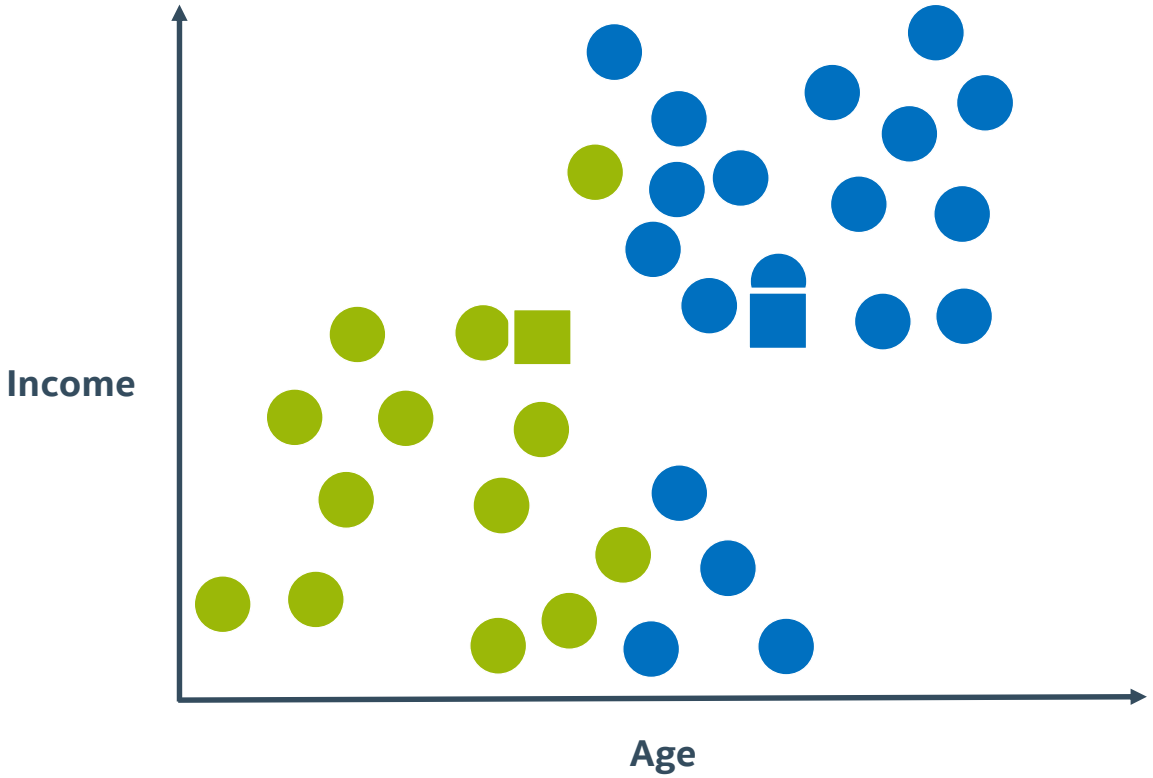
K-MEANS ALGORITHM

K = 2, Move each center to cluster's mean.



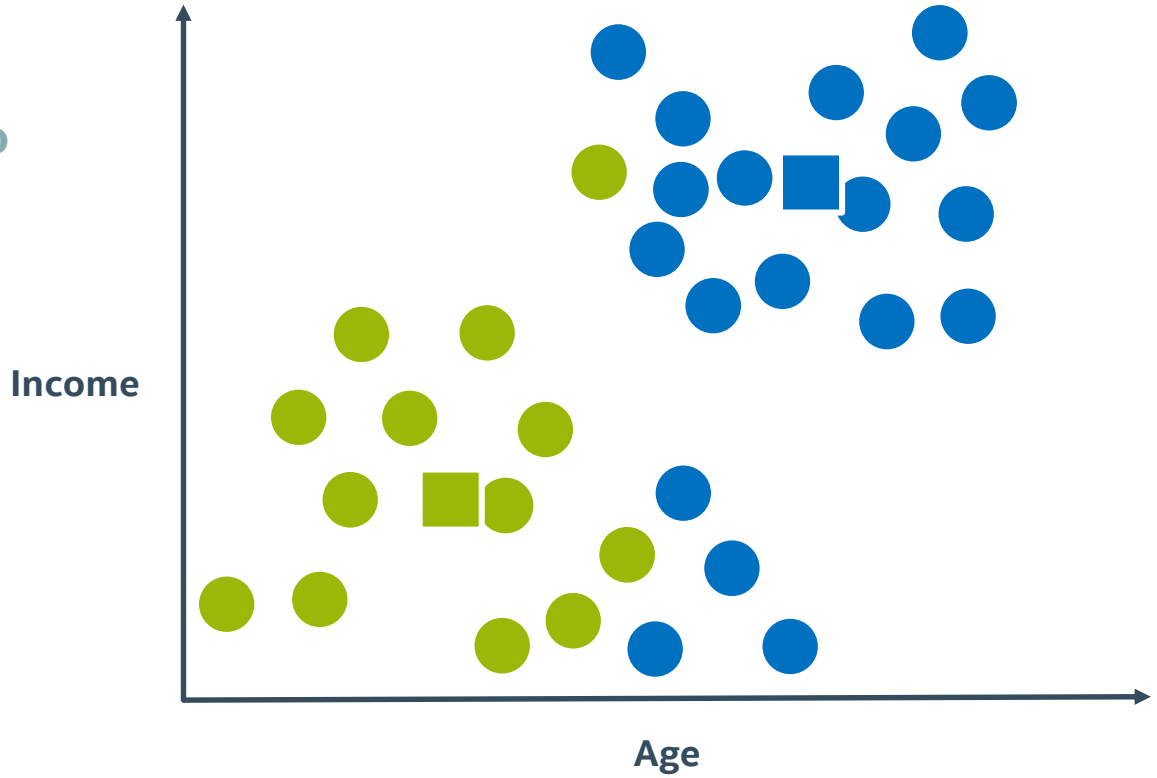
K-MEANS ALGORITHM

K = 2, Each point belongs to closest center.



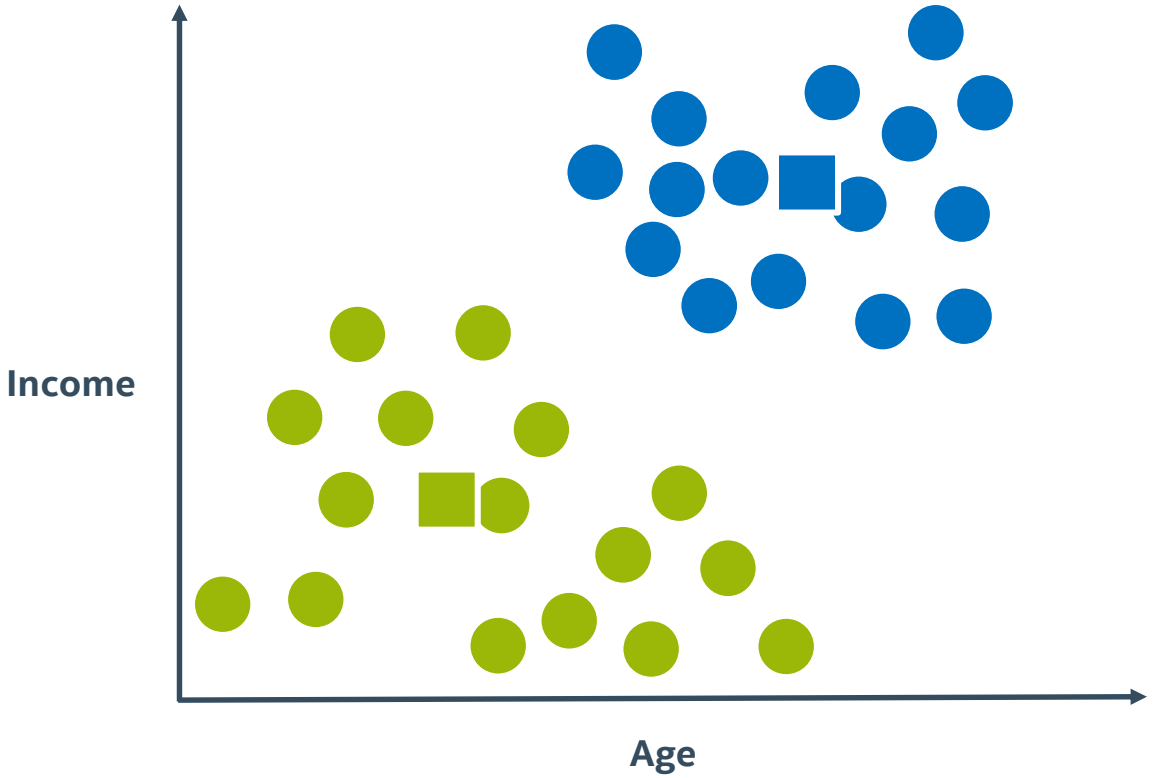
K-MEANS ALGORITHM

K = 2, Move each center to cluster's mean.



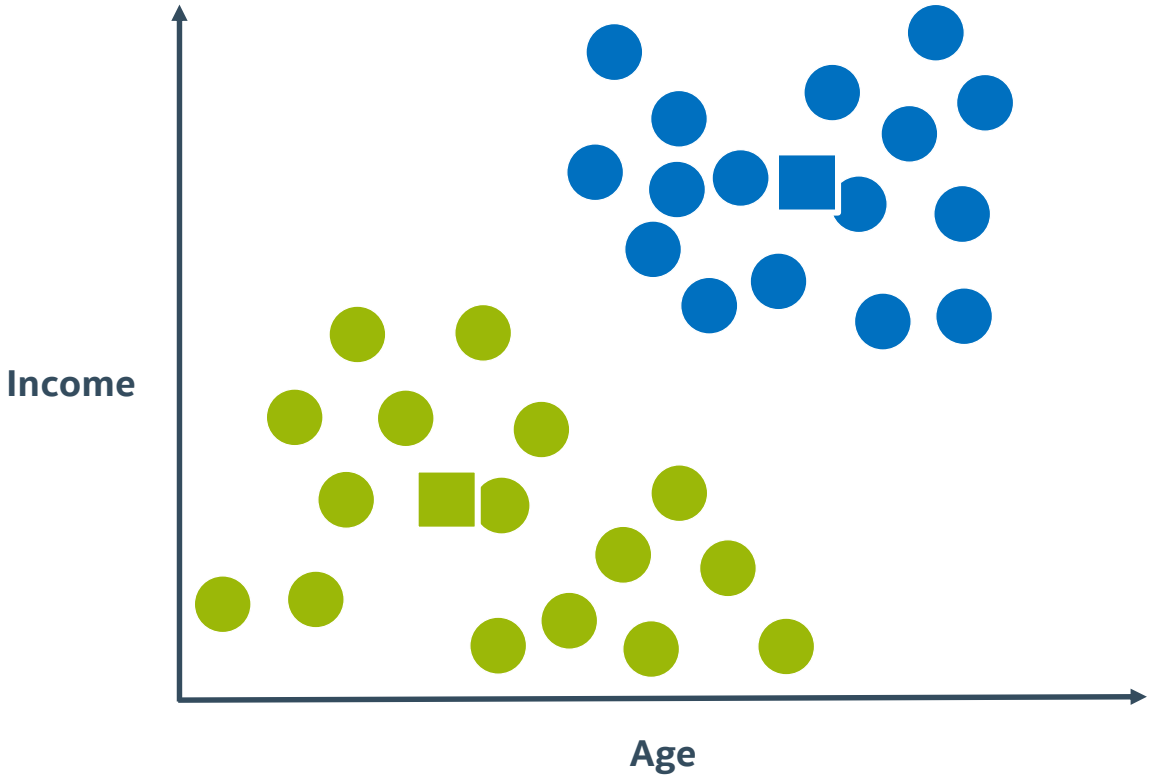
K-MEANS ALGORITHM

K = 2, Points don't change
→ Converged.



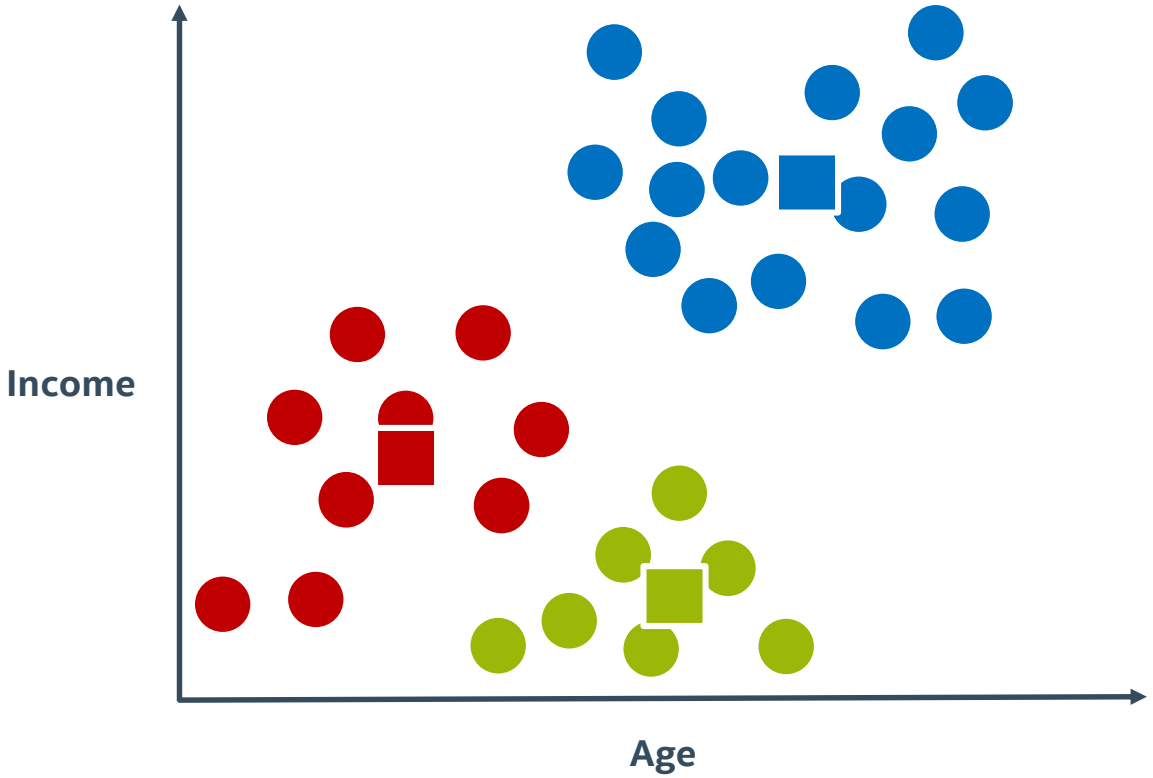
K-MEANS ALGORITHM

K = 2, Each point belongs to closest center.



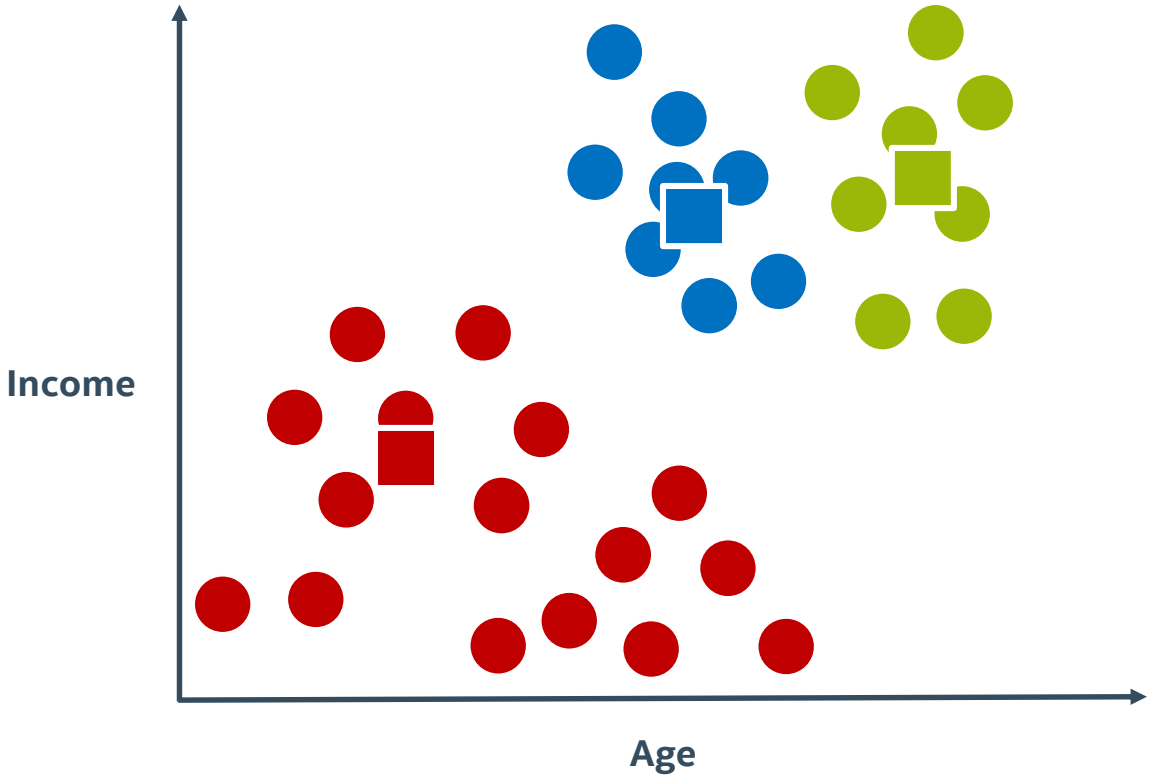
K-MEANS ALGORITHM

K = 3

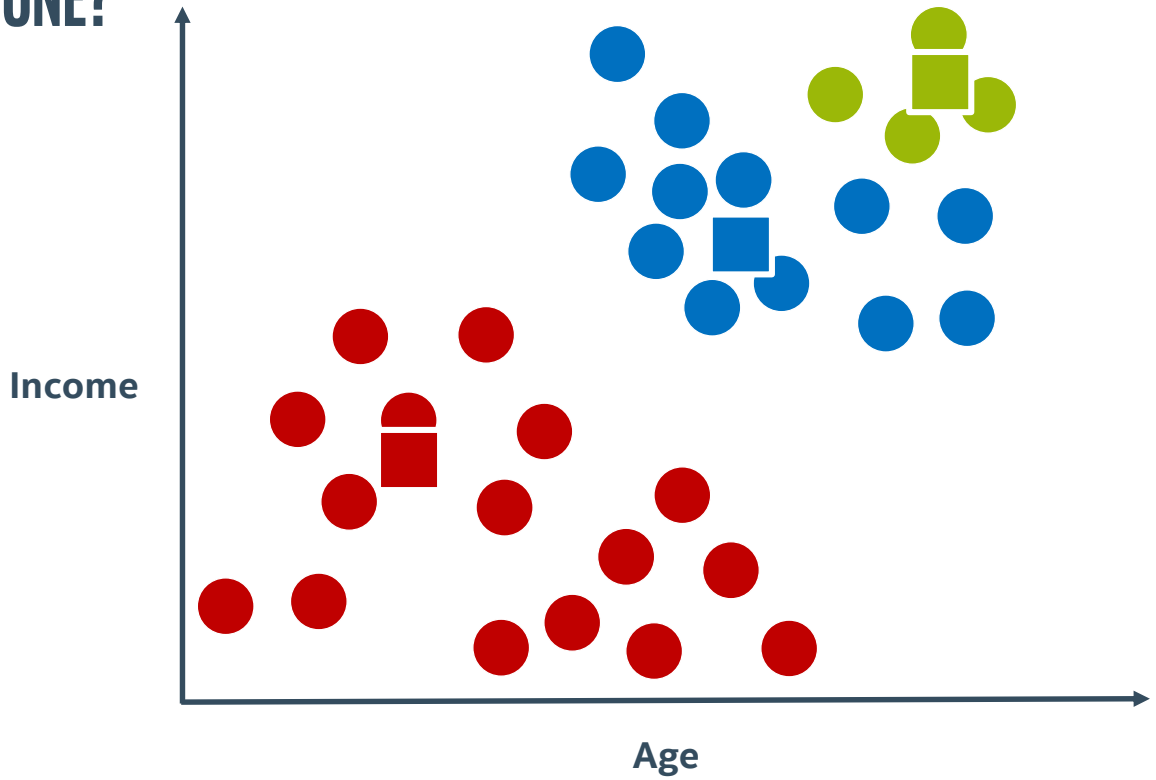


K-MEANS ALGORITHM

K = 3, Results depend on initial cluster assignment.



WHICH MODEL IS THE RIGHT ONE?



WHICH MODEL IS THE RIGHT ONE?

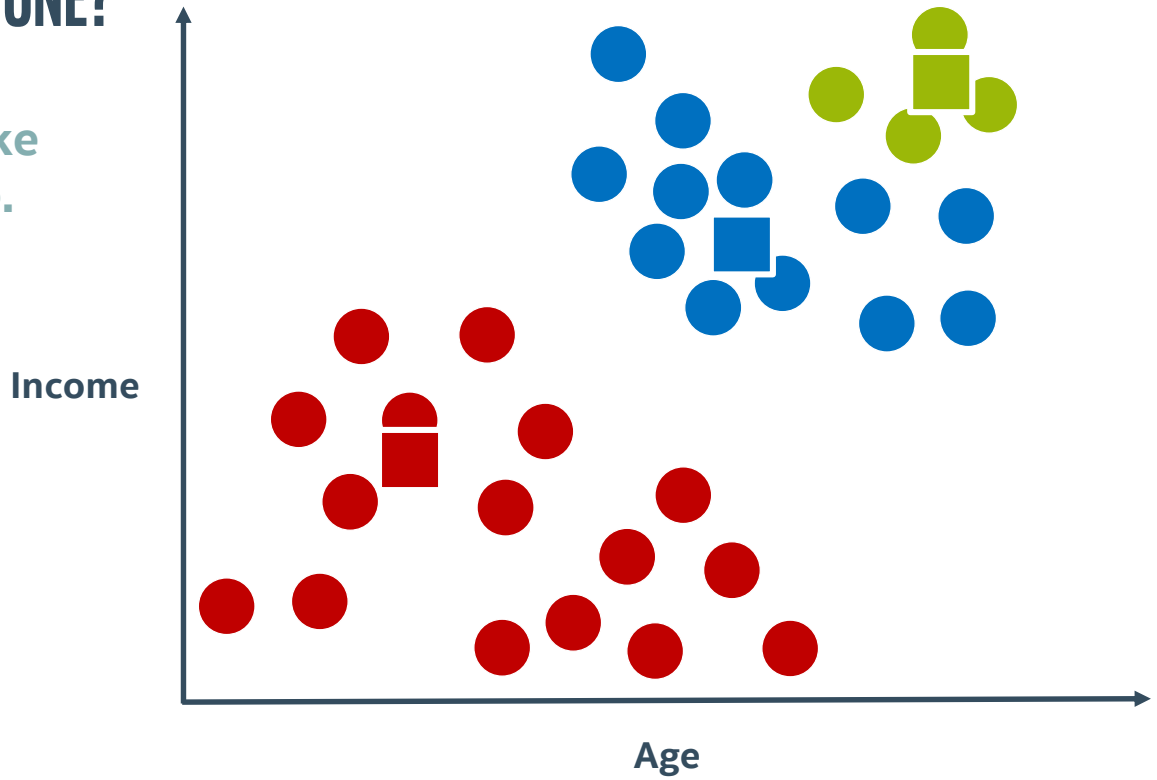
- **Inertia:** sum of squared distance from each point (x_i) to its cluster (C_k)

$$\sum_{i=1}^{n} (x_i - C_k)^2$$

- Smaller value corresponds to tighter clusters
- Other metrics can also be used

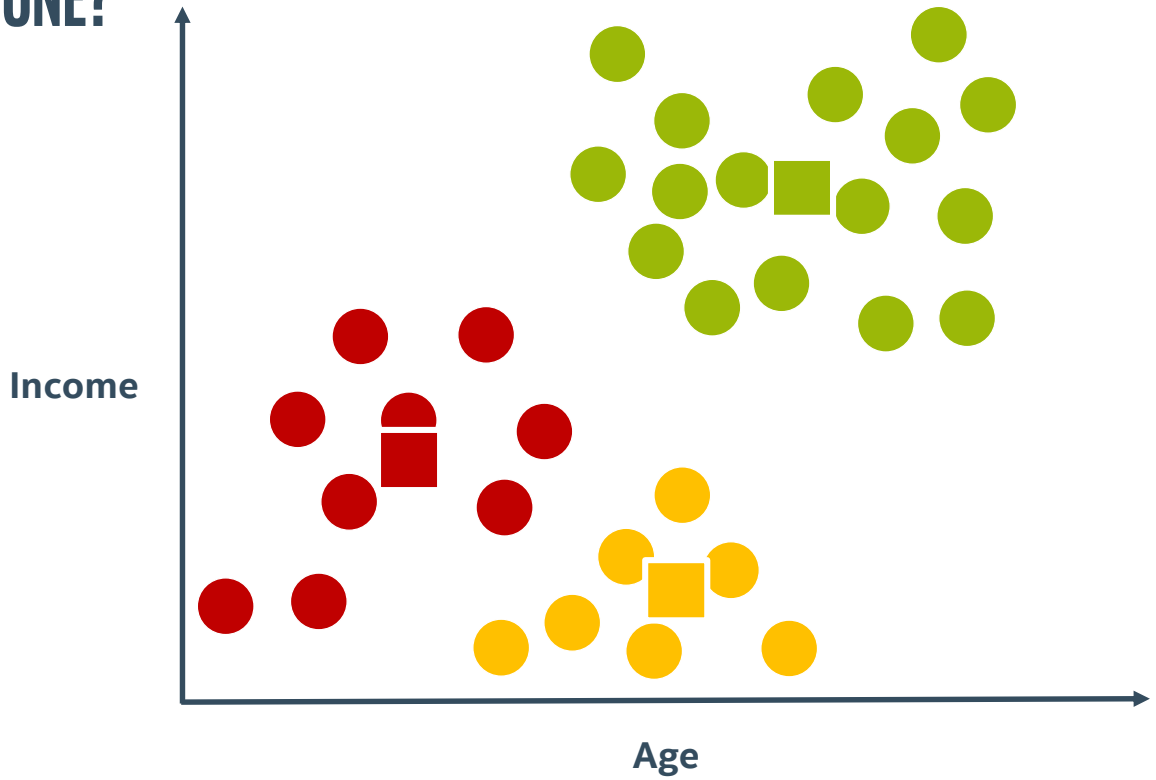
WHICH MODEL IS THE RIGHT ONE?

Initiate multiple times, take model with the best score.



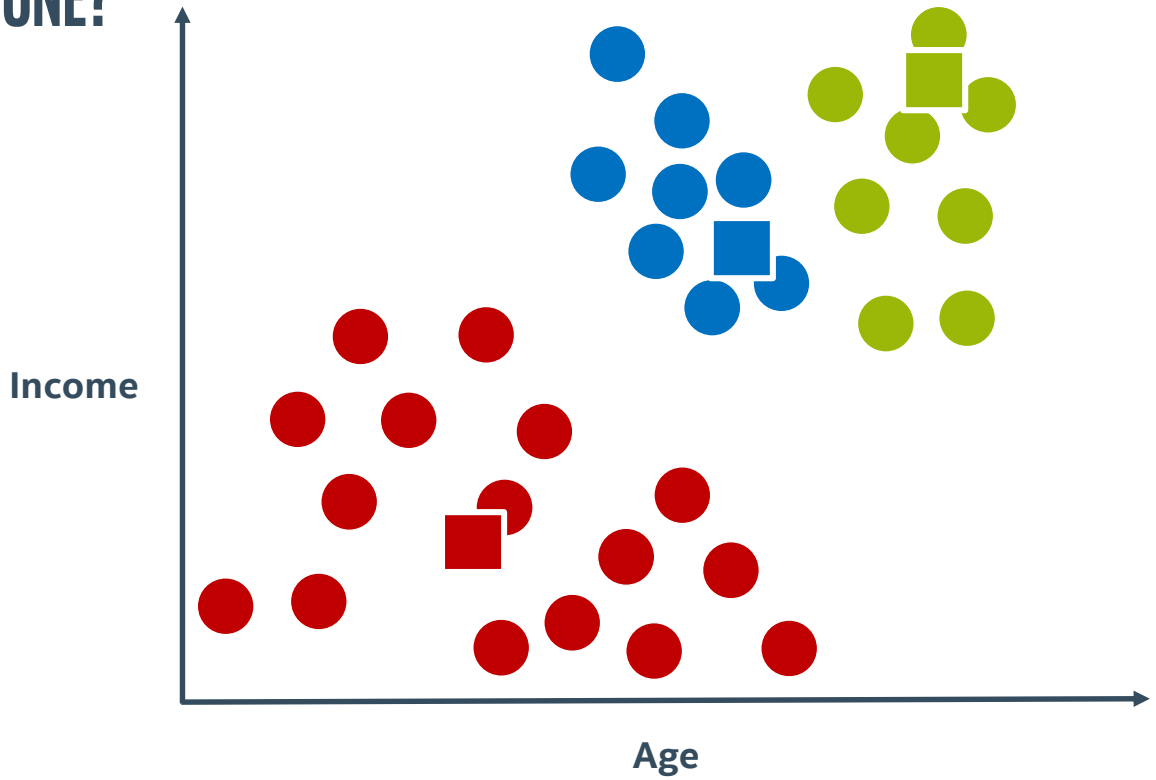
WHICH MODEL IS THE RIGHT ONE?

Inertia = 12.645



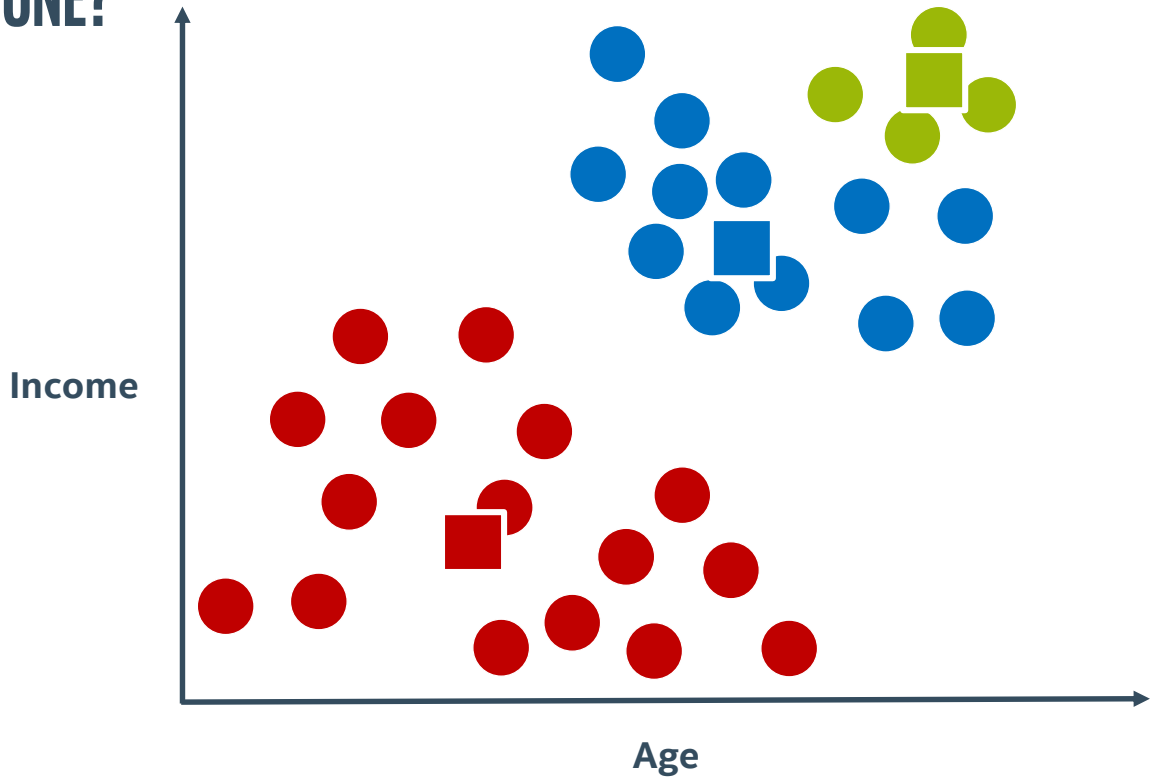
WHICH MODEL IS THE RIGHT ONE?

Inertia = 12.943

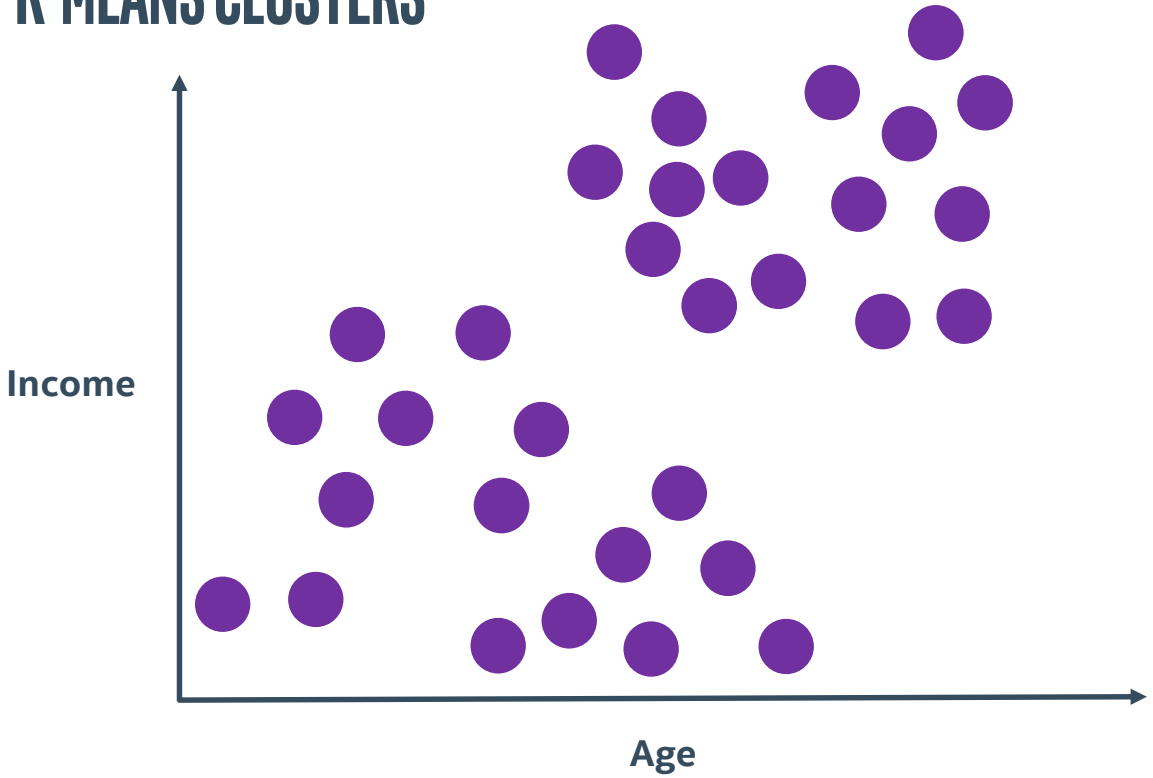


WHICH MODEL IS THE RIGHT ONE?

Inertia = 13.112

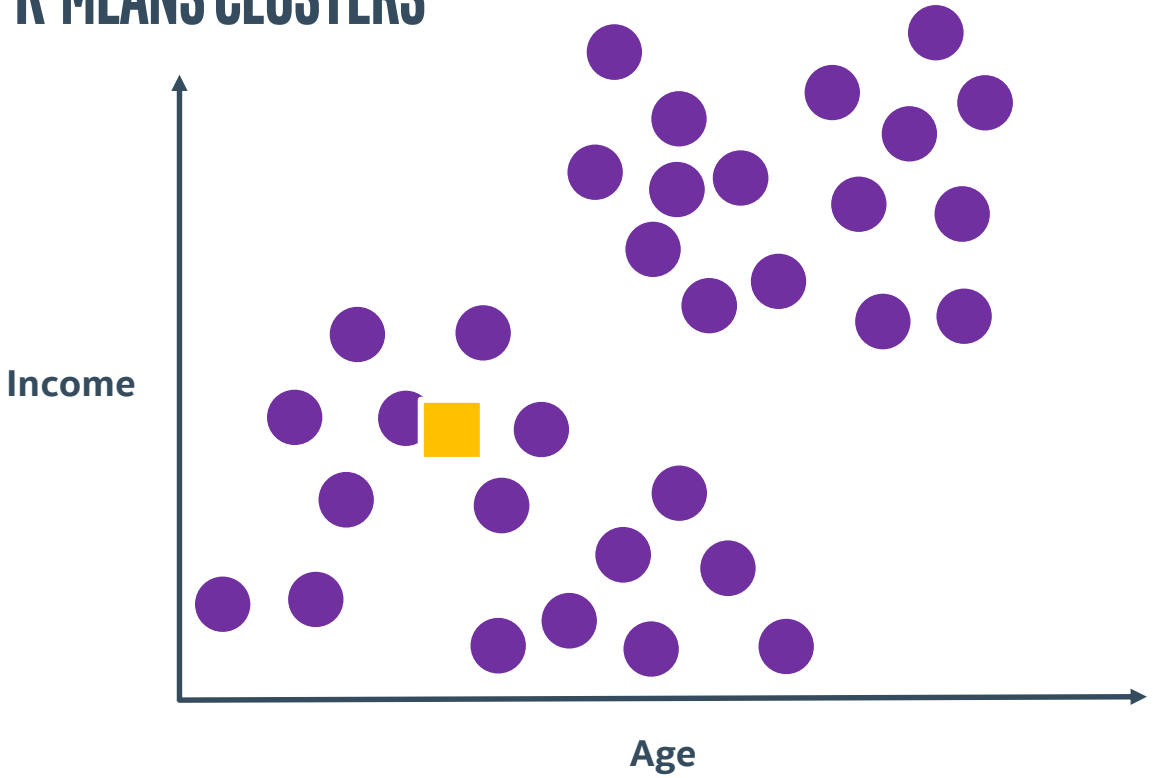


SMARTER INITIALIZATION OF K-MEANS CLUSTERS



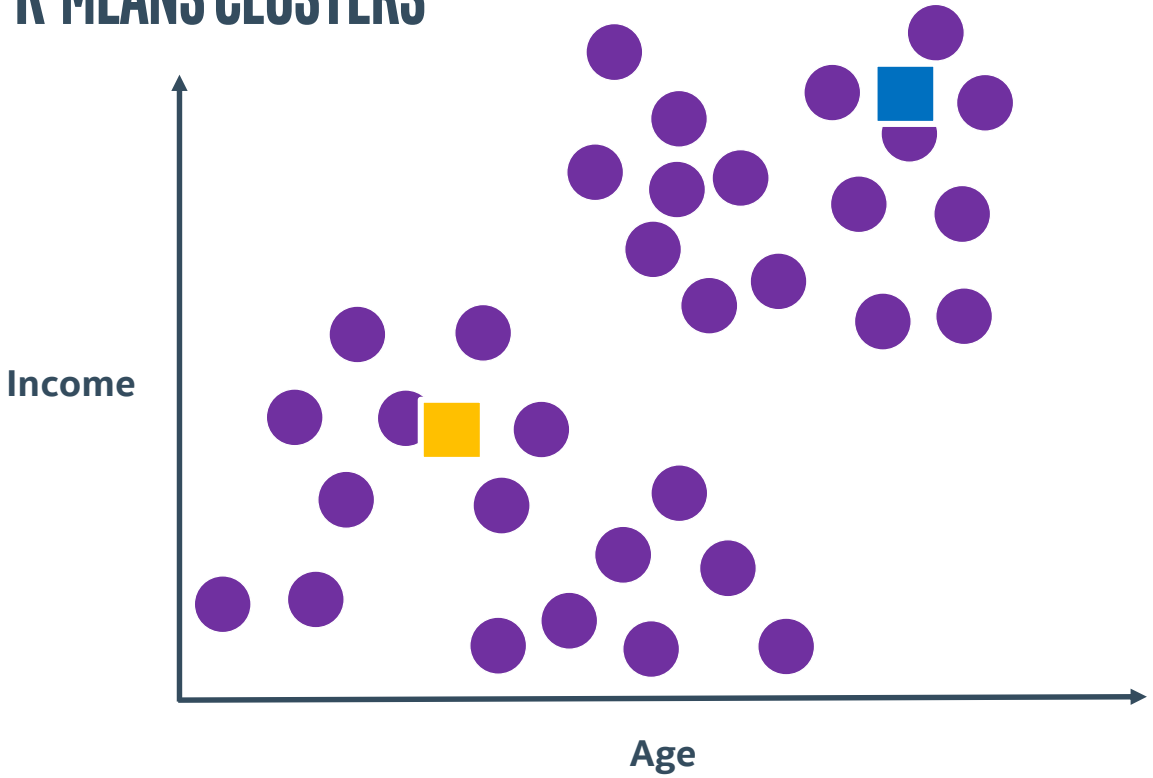
SMARTER INITIALIZATION OF K-MEANS CLUSTERS

Pick one point at random as initial point.



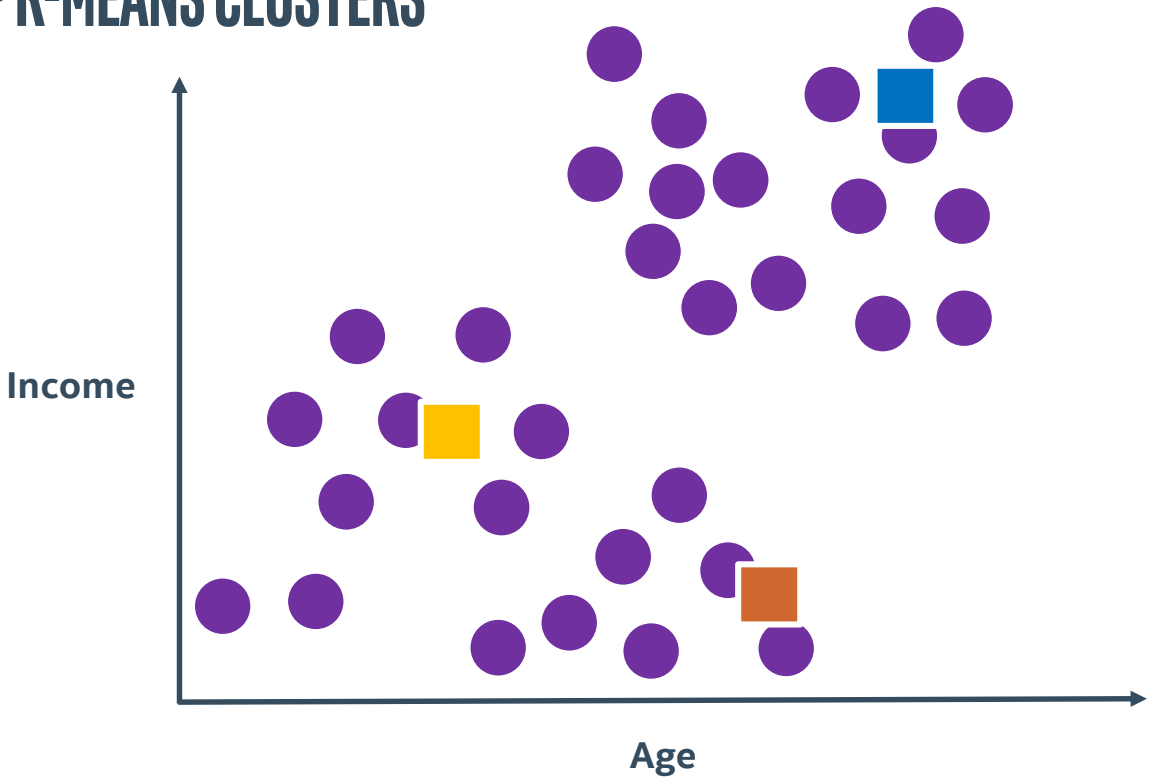
SMARTER INITIALIZATION OF K-MEANS CLUSTERS

Pick next point with
1/distance^2 probability.



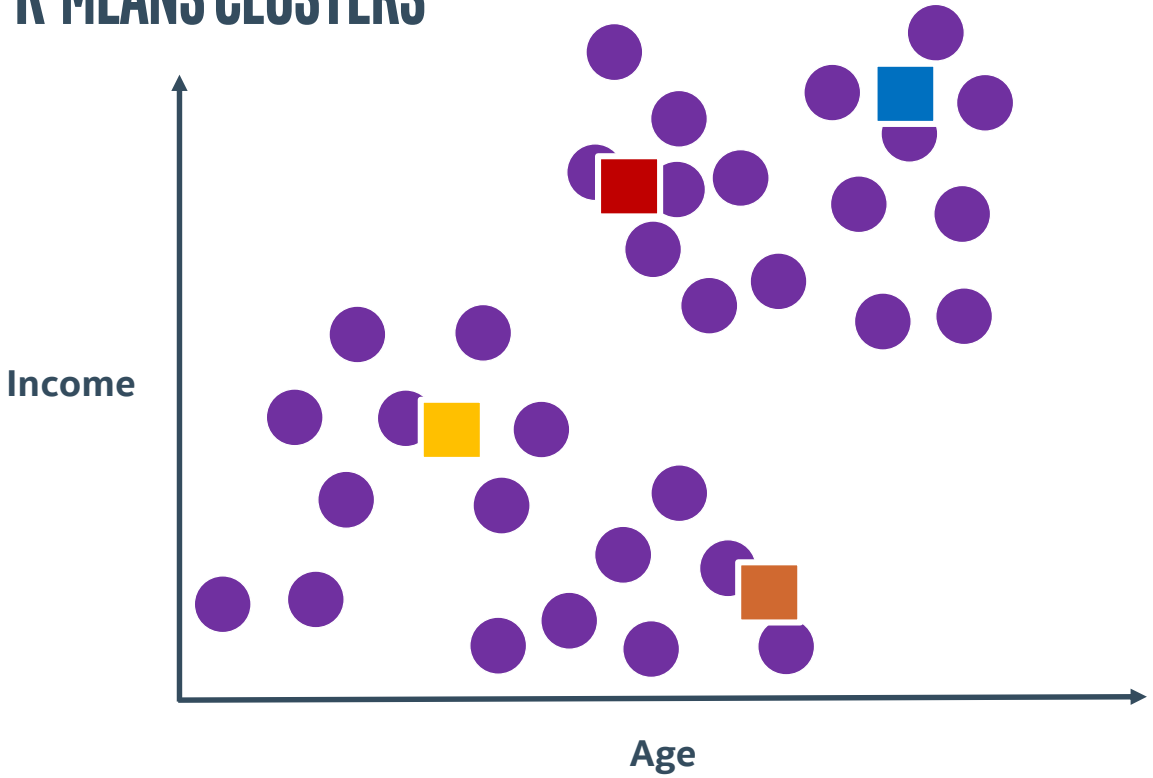
SMARTER INITIALIZATION OF K-MEANS CLUSTERS

Pick next point with
1/distance^2 probability.



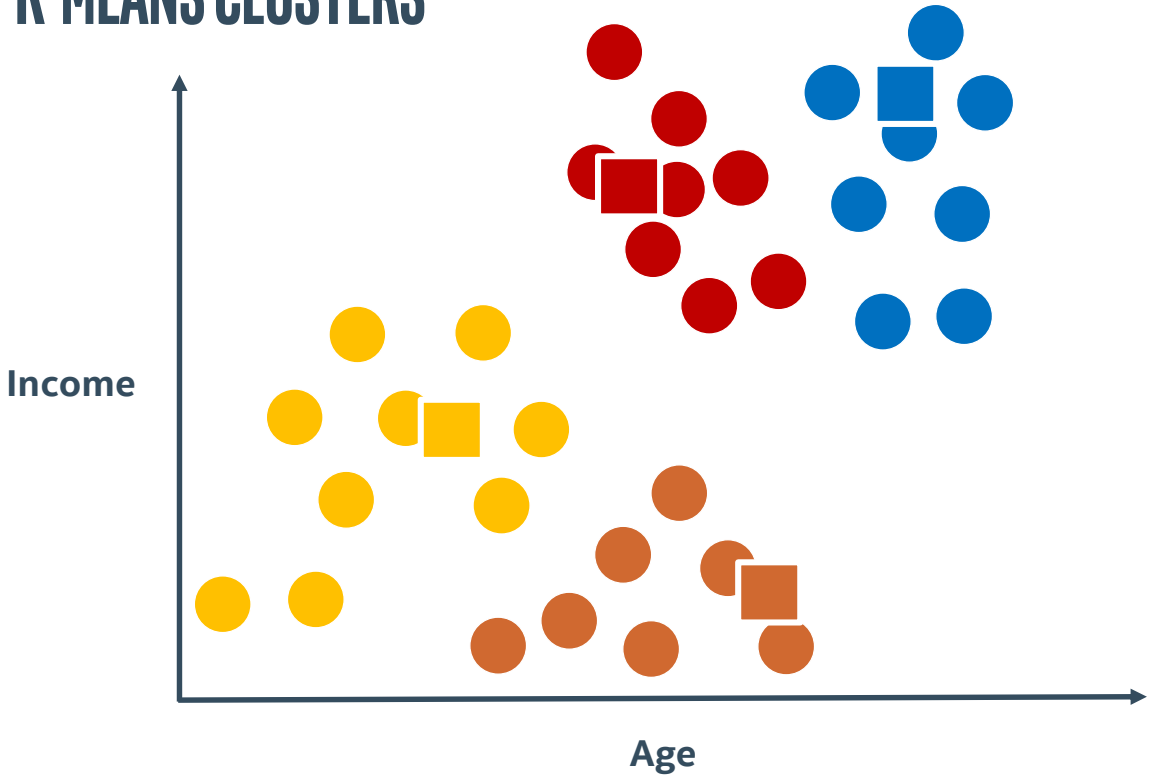
SMARTER INITIALIZATION OF K-MEANS CLUSTERS

Pick next point with
1/distance^2 probability.



SMARTER INITIALIZATION OF K-MEANS CLUSTERS

Assign clusters.



CHOOSING THE RIGHT NUMBER OF CLUSTERS

CHOOSING THE RIGHT NUMBER OF CLUSTERS

- Sometimes the question has a K

CHOOSING THE RIGHT NUMBER OF CLUSTERS

- Sometimes the question has a K
- Clustering similar jobs on 4 CPU cores (K=4)

CHOOSING THE RIGHT NUMBER OF CLUSTERS

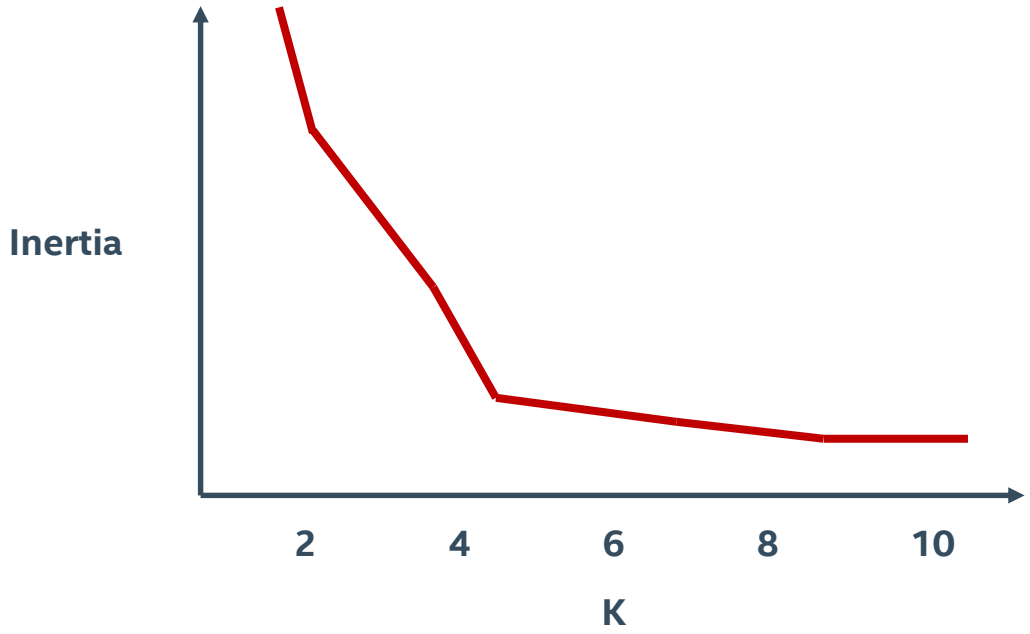
- Sometimes the question has a K
- Clustering similar jobs on 4 CPU cores (K=4)
- A clothing design in 10 different sizes to cover most people (K=10)

CHOOSING THE RIGHT NUMBER OF CLUSTERS

- Sometimes the question has a K
- Clustering similar jobs on 4 CPU cores (K=4)
- A clothing design in 10 different sizes to cover most people (K=10)
- A navigation interface for browsing scientific papers with 20 disciplines (K=20)

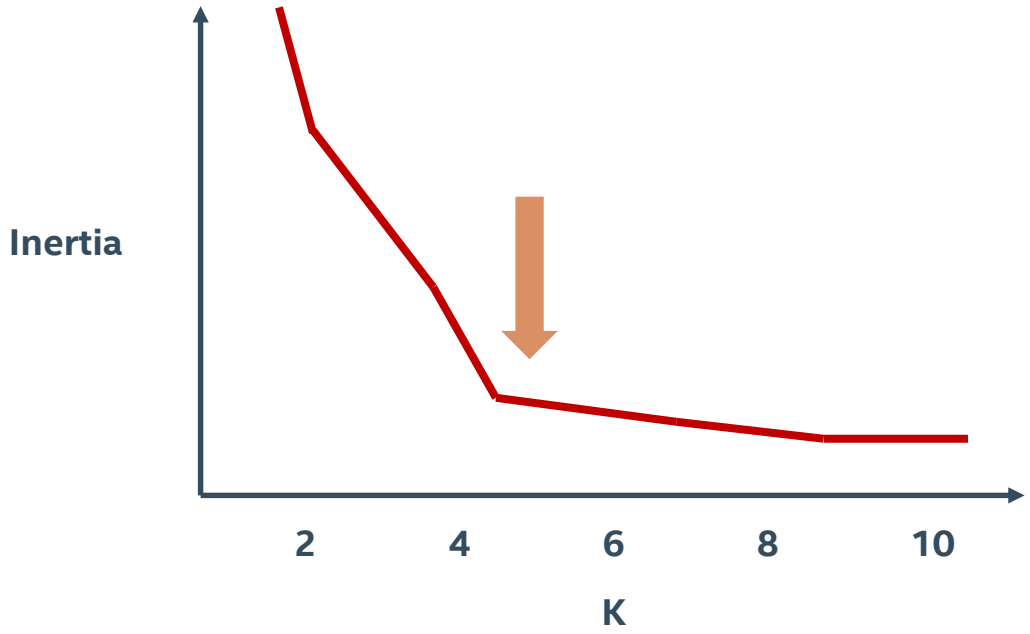
CHOOSING THE RIGHT NUMBER OF CLUSTERS

- Inertia measures distance of point to cluster



CHOOSING THE RIGHT NUMBER OF CLUSTERS

- Inertia measures distance of point to cluster
- Value decreases with increasing K as long as cluster density increases



K-MEANS: THE SYNTAX

Import the class containing the clustering method.

```
from sklearn.cluster import KMeans
```

K-MEANS: THE SYNTAX

Import the class containing the clustering method.

```
from sklearn.cluster import KMeans
```

Create an instance of the class.

```
kmeans = KMeans(n_clusters=3,
                init='k-means++')
```

K-MEANS: THE SYNTAX

Import the class containing the clustering method.

```
from sklearn.cluster import KMeans
```

Create an instance of the class.

```
kmeans = KMeans(n_clusters=3,
                init='k-means++')
```

final number of clusters

K-MEANS: THE SYNTAX

Import the class containing the clustering method.

```
from sklearn.cluster import KMeans
```

Create an instance of the class.

```
kmeans = KMeans(n_clusters=3,
                init='k-means++')
```

kmeans++ cluster initiation

K-MEANS: THE SYNTAX

Import the class containing the clustering method.

```
from sklearn.cluster import KMeans
```

Create an instance of the class.

```
kmeans = KMeans(n_clusters=3,
                init='k-means++')
```

Fit the instance on the data and then predict clusters for new data.

```
kmeans = kmeans.predict(X1)
y_predict = kmeans.predict(X2)
```

K-MEANS: THE SYNTAX

Import the class containing the clustering method.

```
from sklearn.cluster import KMeans
```

Create an instance of the class.

```
kmeans = KMeans(n_clusters=3,
                init='k-means++')
```

Fit the instance on the data and then predict clusters for new data.

```
kmeans = kmeans.predict(X1)
y_predict = kmeans.predict(X2)
```

Can also be used in batch mode with **MiniBatchKMeans**.


intel
Nervana™ AI Academy