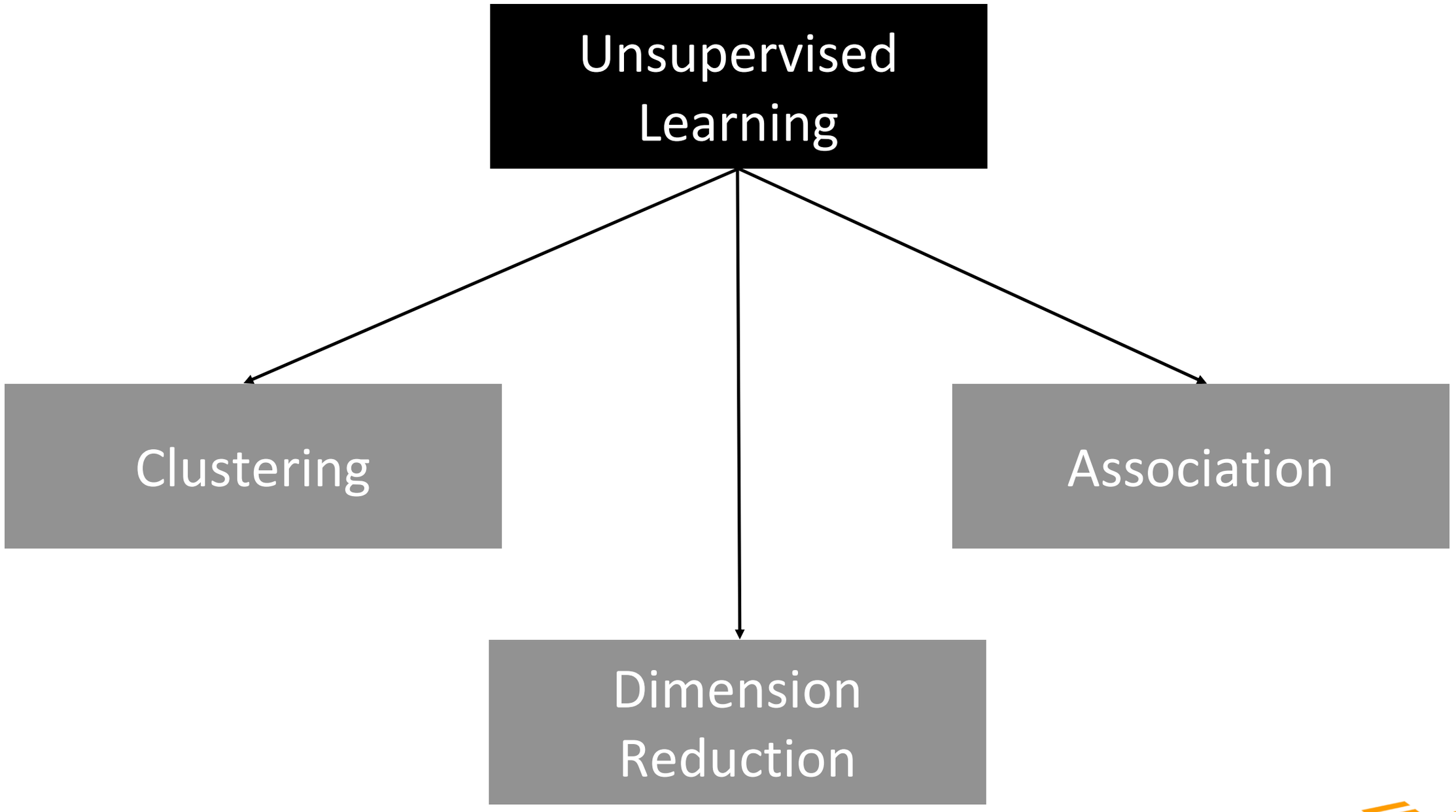
DICE ANALYTICS

DATA SCIENCE & MACHINE LEARNING COURSE

https://www.facebook.com/diceanalytics/
https://pk.linkedin.com/company/diceanalytics

Unsupervised Learning

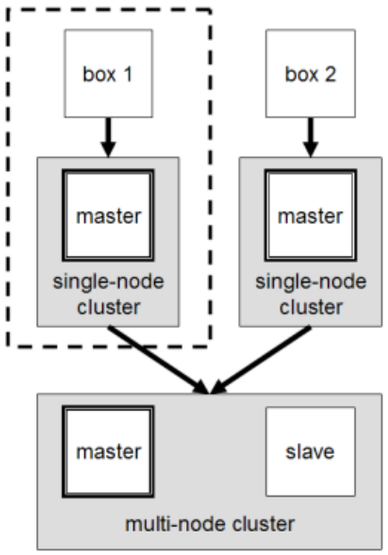
Types of Unsupervised Learning

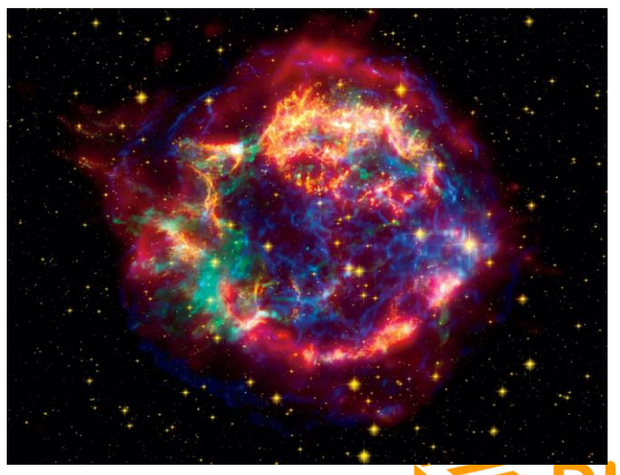


Clustering

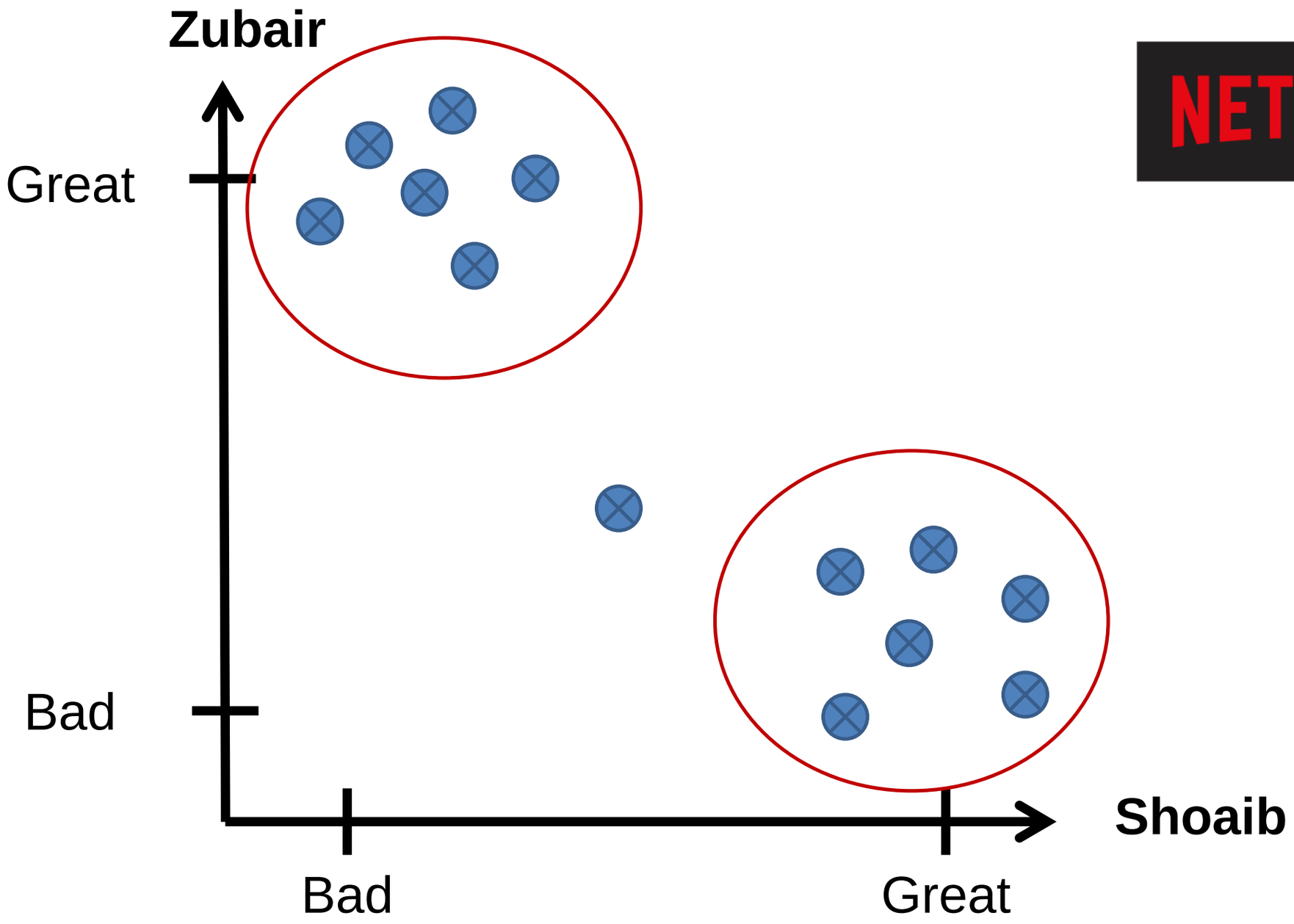
- Finding hidden groups in data.
- Algorithms : K-Means, DBSCAN, Hierarchical etc.
- Examples :

1. Market Segmentation
2. Social Network Analysis
3. Organize Computing
4. Astronomical Data Analysis





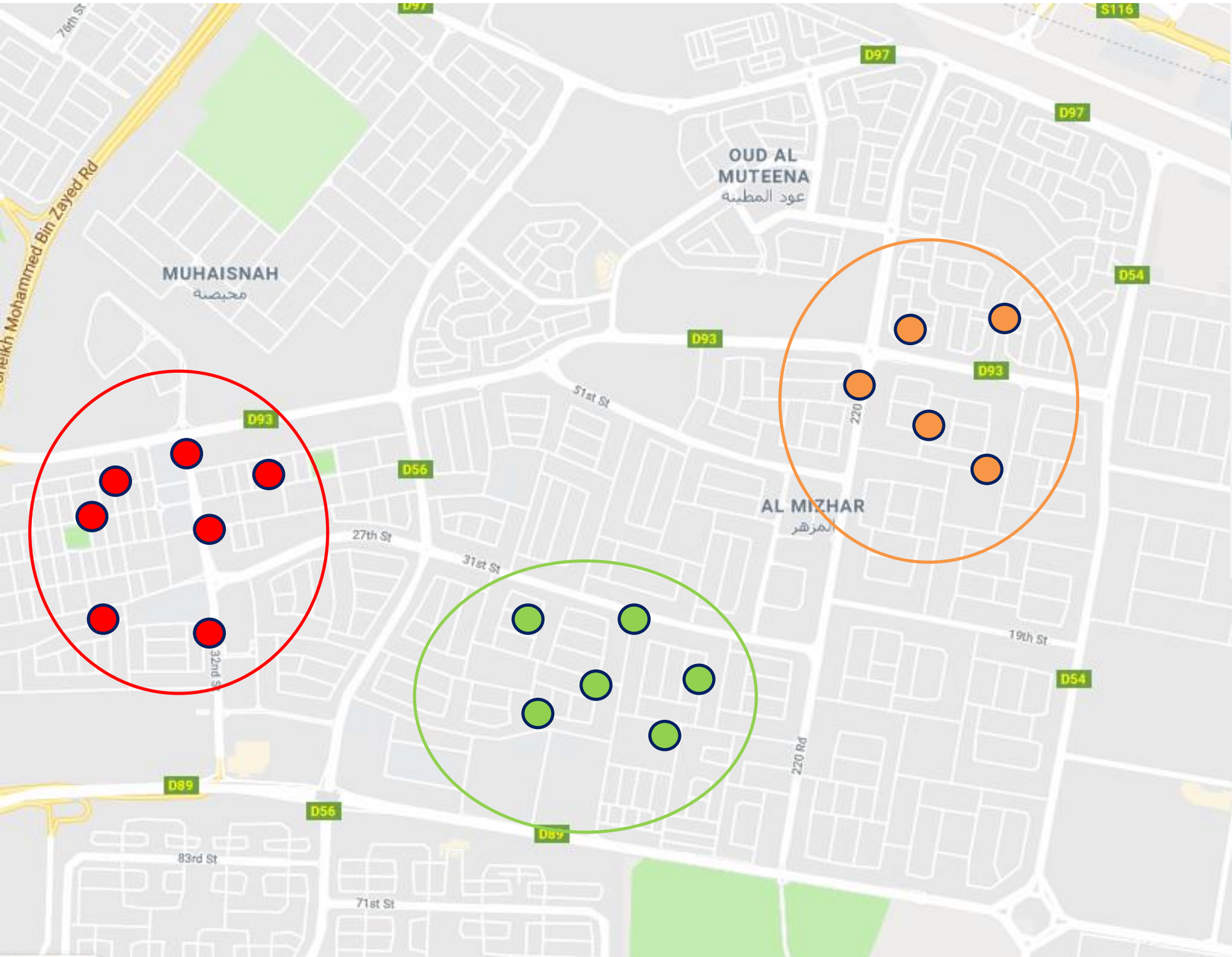
Clustering Movies Example

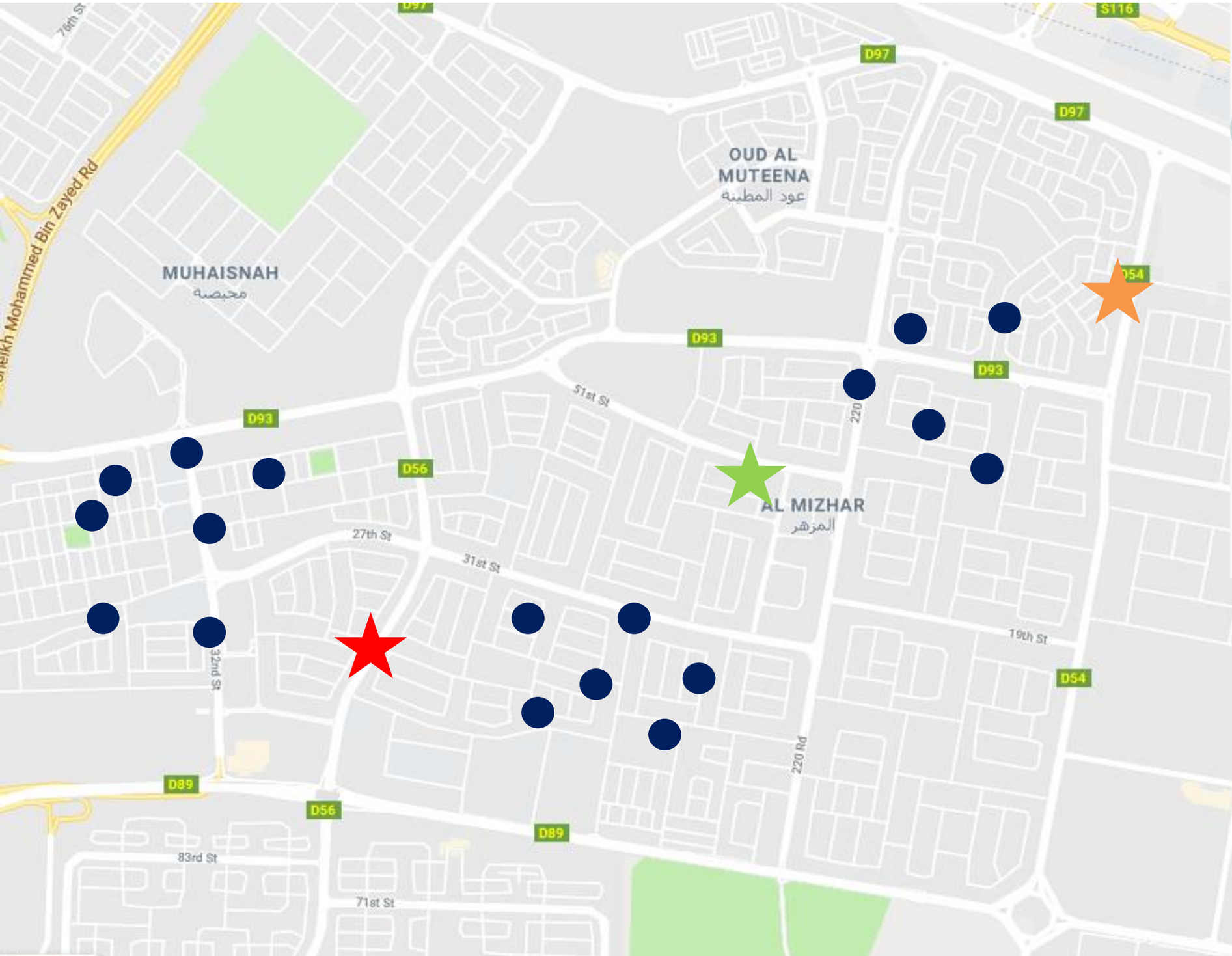


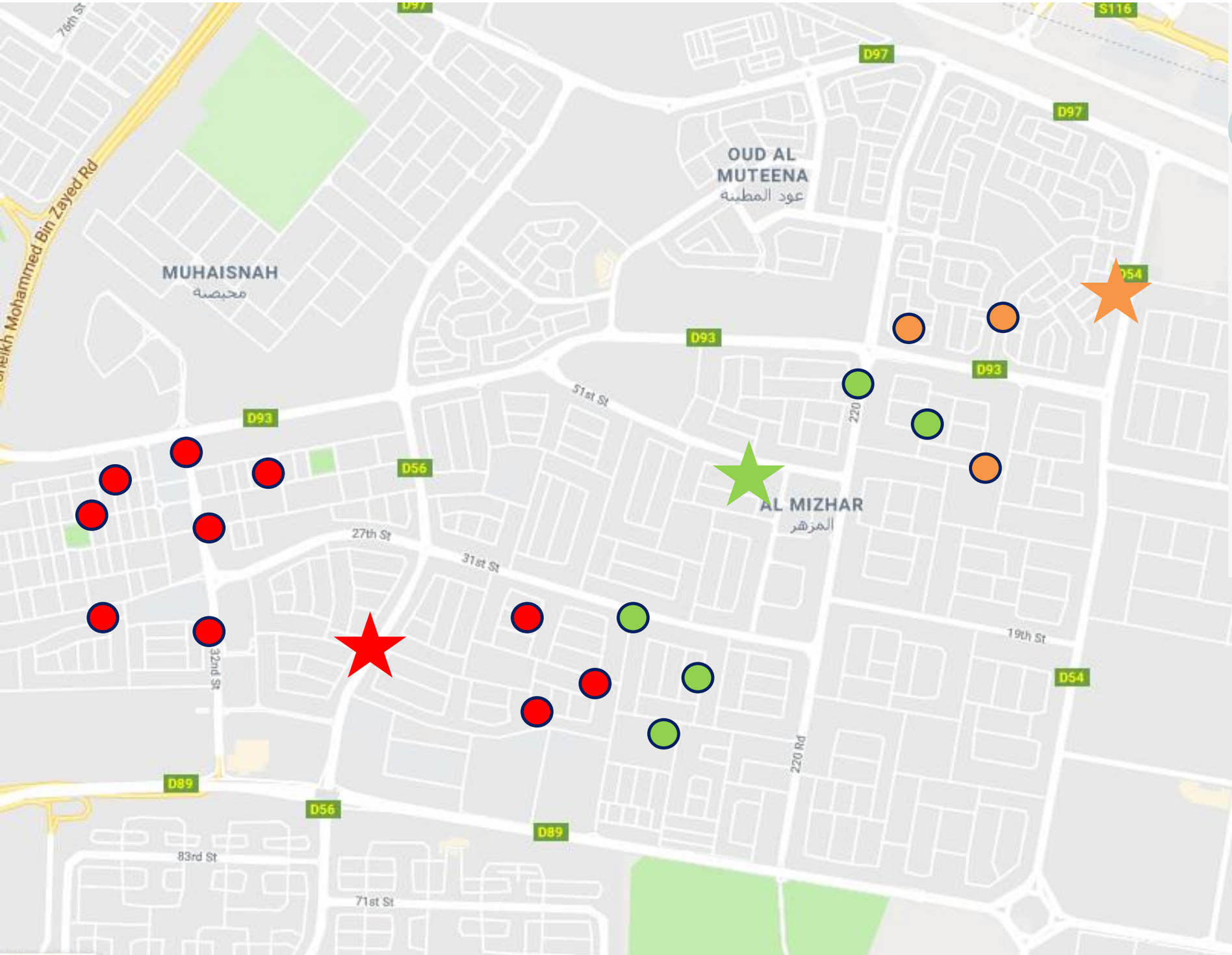
NETFLIX

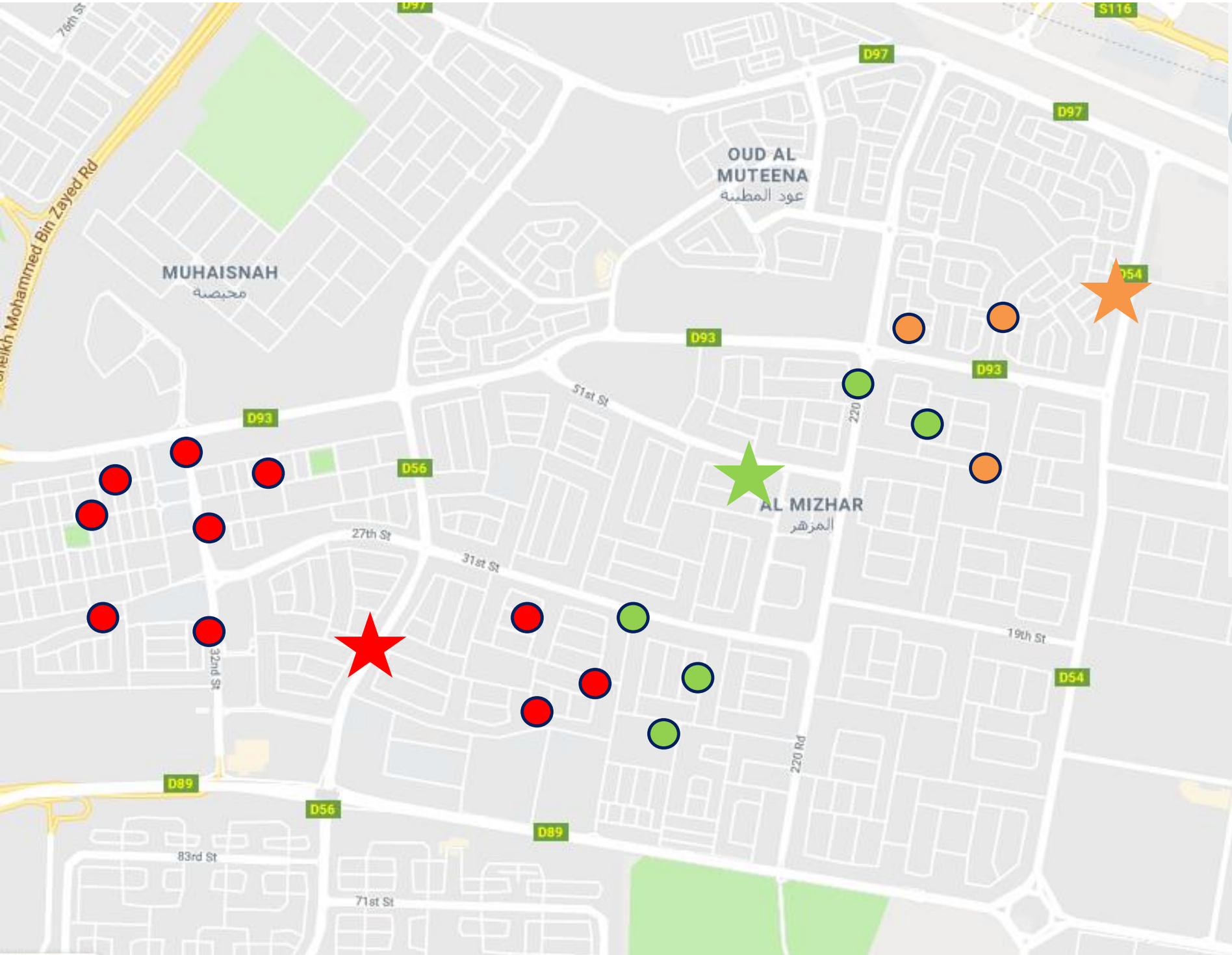
DICE ANALYTICS

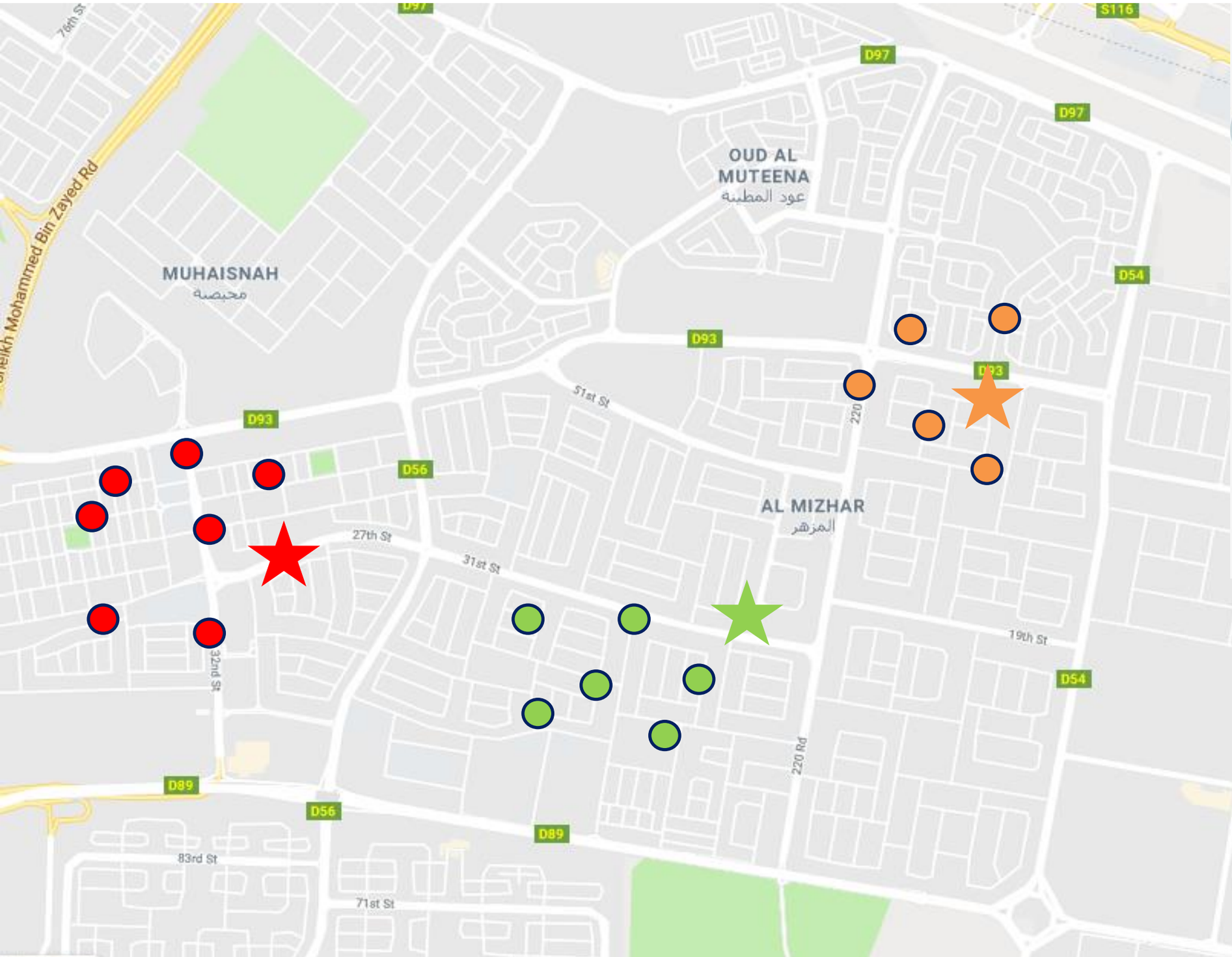

MUHAISNAH
محيصنة
OUD AL MUTEENA
عود المطينه
AL MIZHAR
المزهر
D97
D54
D93
D56
D89
S116
76th St
Sheikh Mohammed Bin Zayed Rd
51st St
220
27th St
31st St
32nd St
19th St
220 Rd
83rd St
71st St


D97
S116
D97
D97
OUD AL
MUTEENA
عود المطينه
MUHAISNAH
محيصنة
D54
D93
D93
51st St
220
D93
D56
AL MIZHAR
المزهر
27th St
31st St
19th St
D54
32nd St
220 Rd
D89
D56
D89
83rd St
71st St

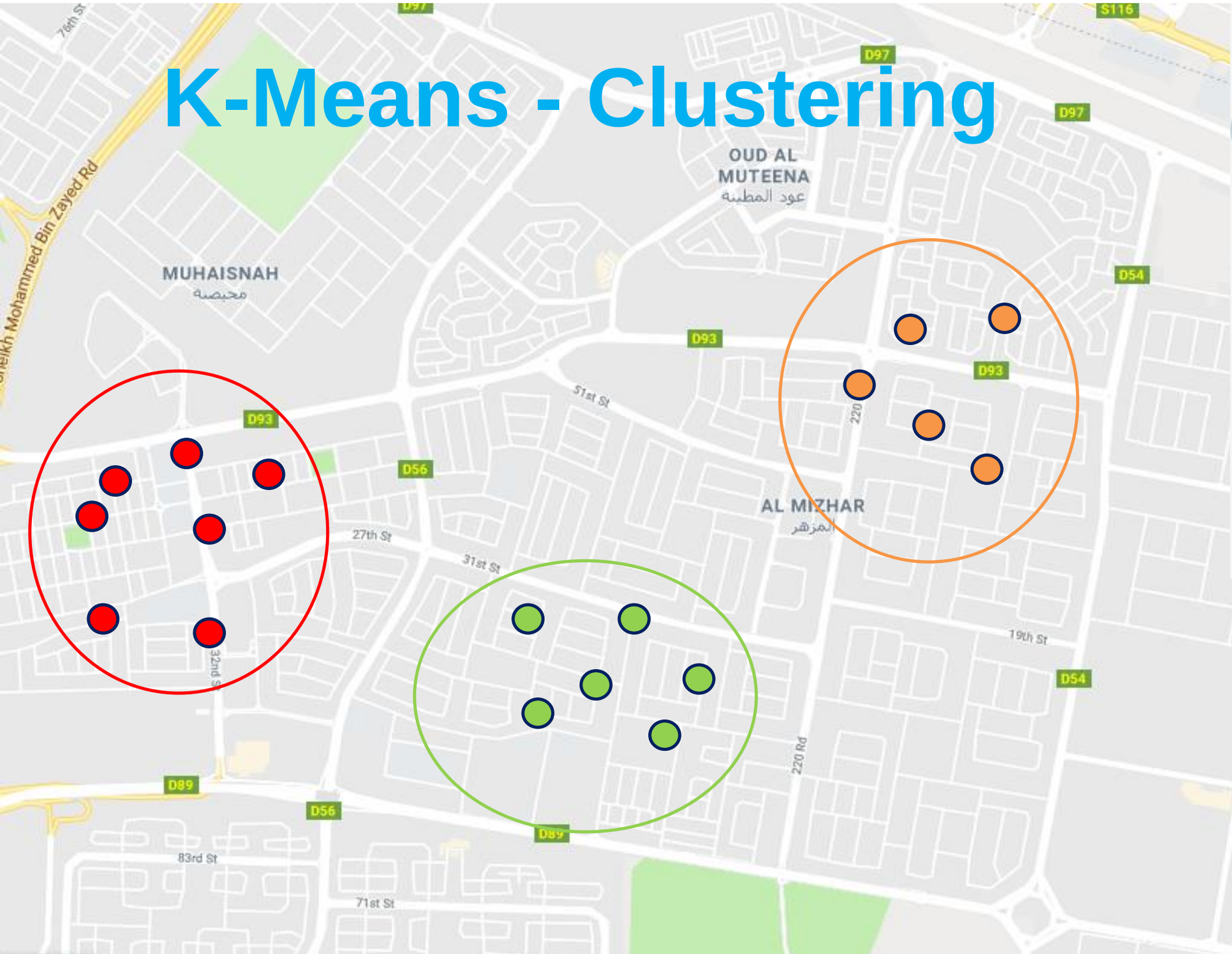



MUHAISNAH
محيصنة
OUD AL MUTEENA
عود المطينه
AL MIZHAR
المزهر
Sheikh Mohammed Bin Zayed Rd
76th St
51st St
27th St
31st St
32nd St
19th St
220 Rd
83rd St
71st St
D97
D93
D56
D89
D54
S116


MUHAISNAH
محيصنة
OUD AL MUTEENA
عود المطينه
AL MIZHAR
المزهر
Sheikh Mohammed Bin Zayed Rd
76th St
51st St
27th St
31st St
32nd St
19th St
220 Rd
83rd St
71st St
D97
D93
D56
D89
D54
S116

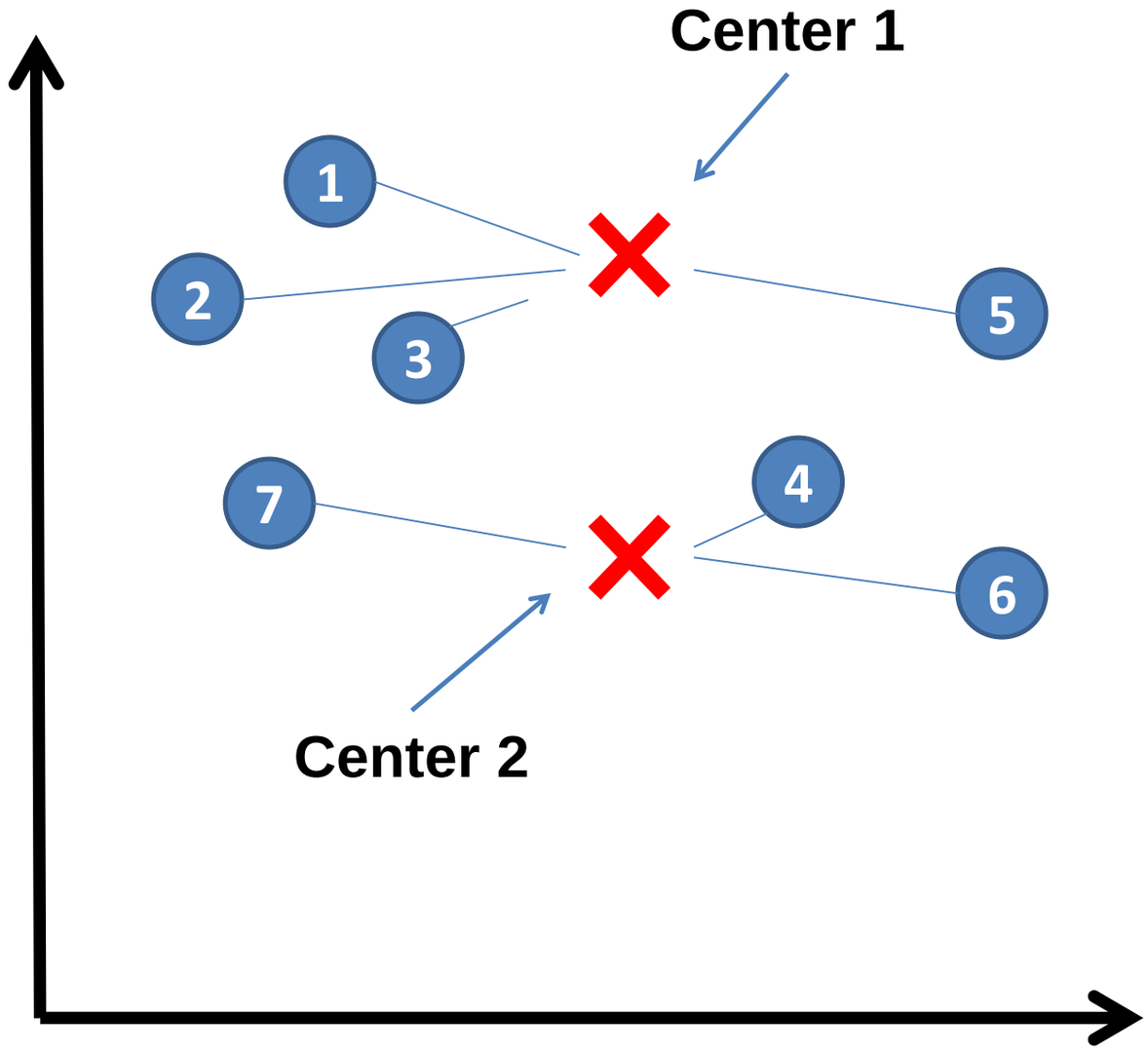


K-Means - Clustering



DICE ANALYTICS

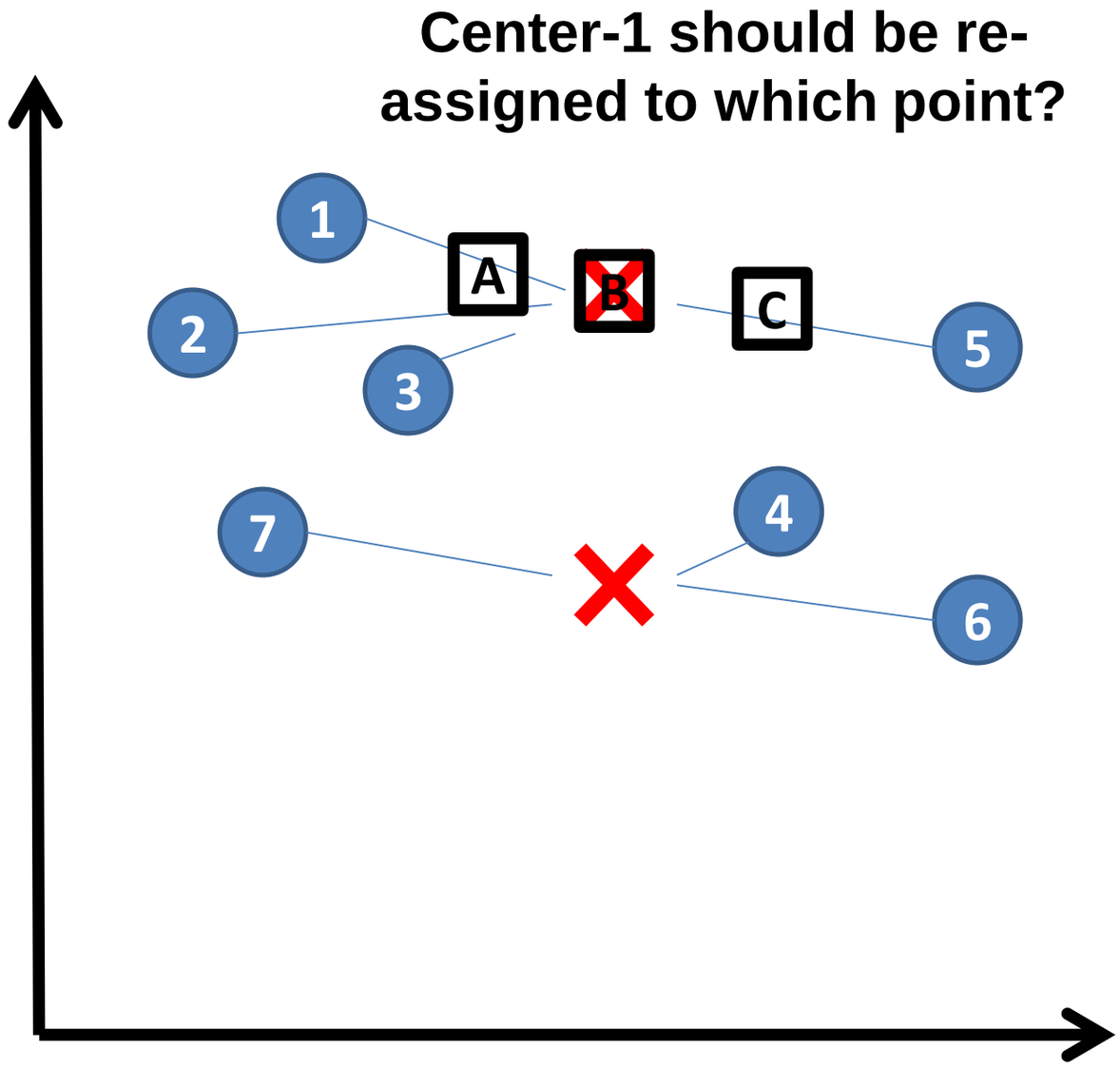
K-Means Algorithm



KMeans Steps

1) **Assign**
2) **Optimize**

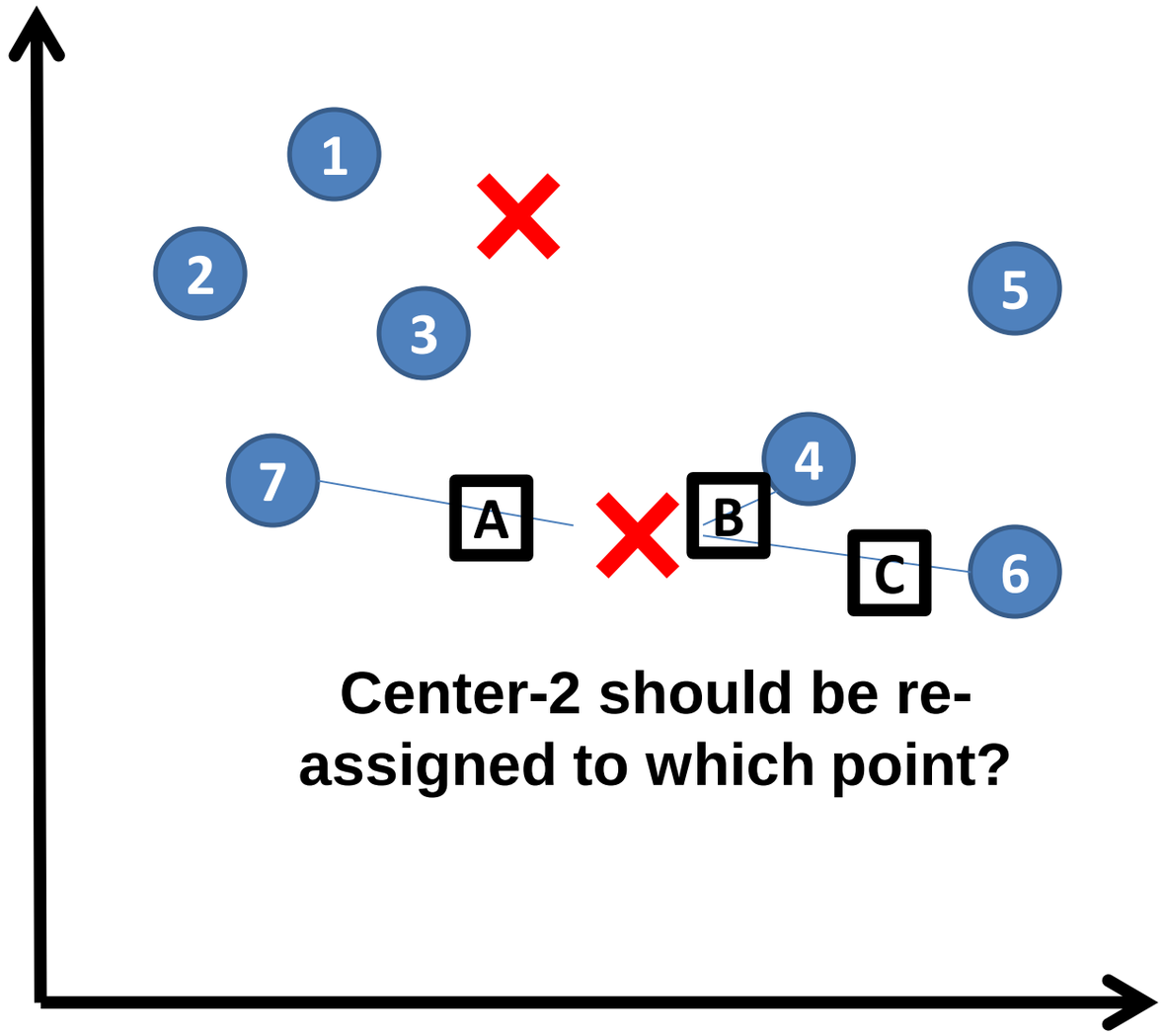
K-Means Algorithm



KMeans Steps

1) Assign
2) Optimize
3) Re-Assign

K-Means Algorithm

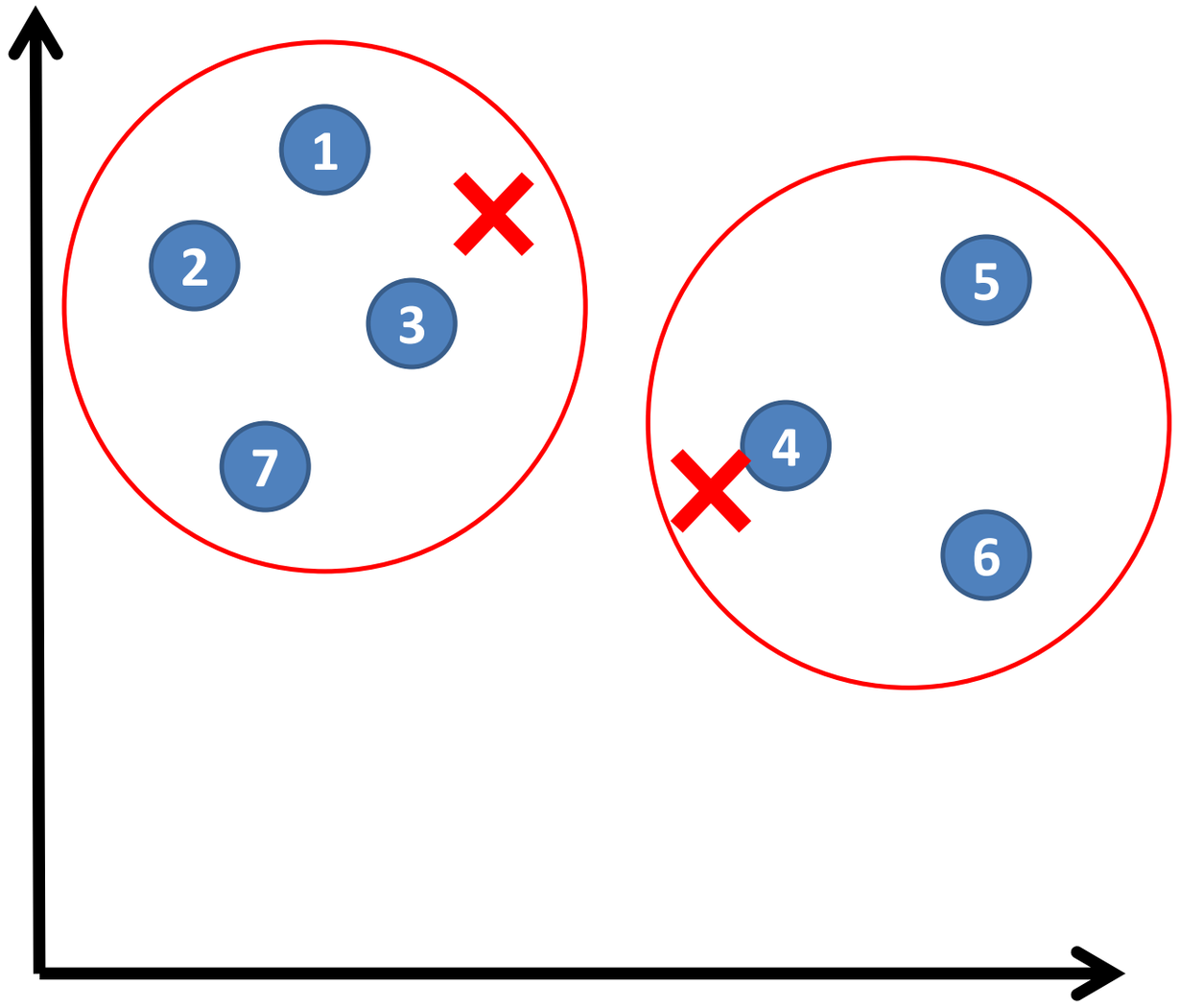


Center-2 should be re-assigned to which point?

KMeans Steps

1) **Assign**
2) **Optimize**
3) **Re-Assign**

K-Means Algorithm



$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} \left\| x_n - \mu_k \right\|^2$$

KMeans Steps

1) **Assign**
2) **Optimize**
3) **Re-Assign**
4) **Re-Optimize**

Continued Till Minimum Cost Function

K-Means Playground

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

K-Means in scikit-learn

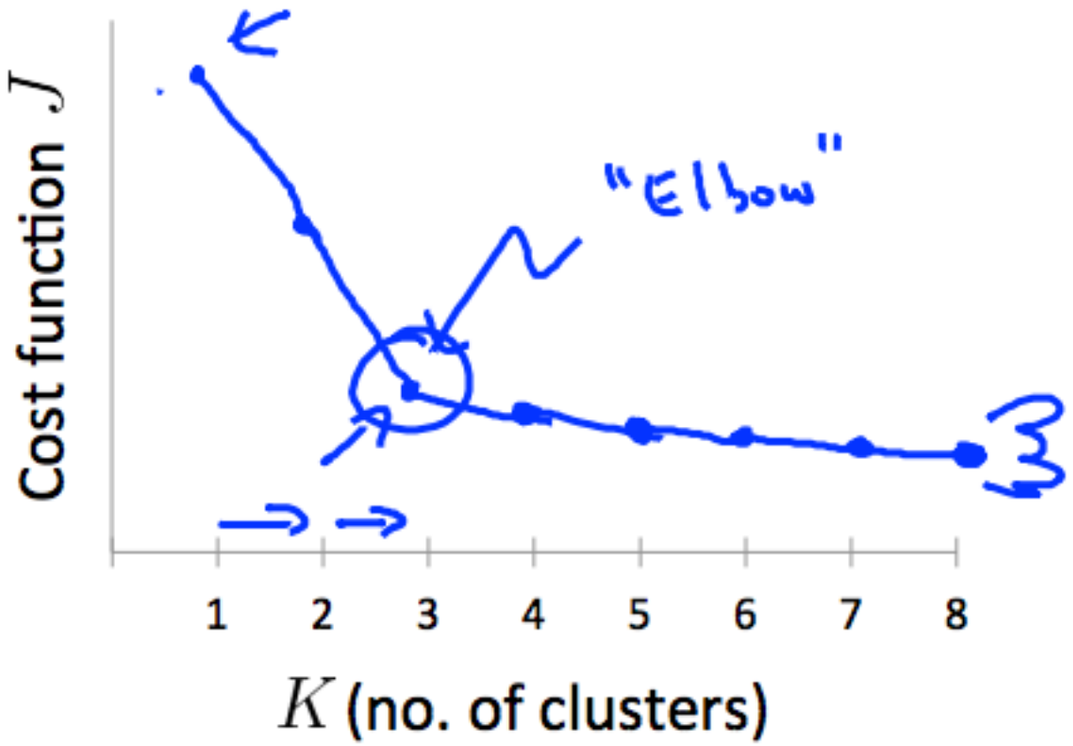
class `sklearn.cluster.` **`KMeans`** (*n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=None, algorithm='auto'*)

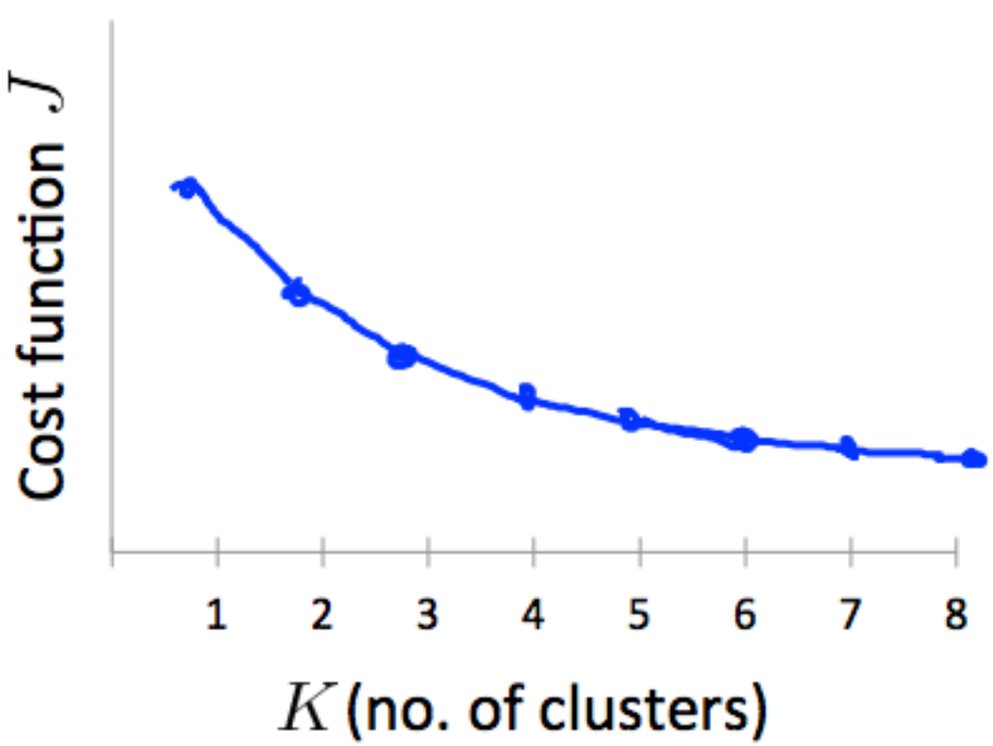
```
>>> from sklearn.cluster import KMeans
>>> import numpy as np
>>> X = np.array([[1, 2], [1, 4], [1, 0],
...               [10, 2], [10, 4], [10, 0]])
>>> kmeans = KMeans(n_clusters=2, random_state=0).fit(X)
>>> kmeans.labels_
array([1, 1, 1, 0, 0, 0], dtype=int32)
>>> kmeans.cluster_centers_
array([[10.,  2.],
       [ 1.,  2.]])
```

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

How much K? (Elbow)

Elbow method:





by Andrew Ng

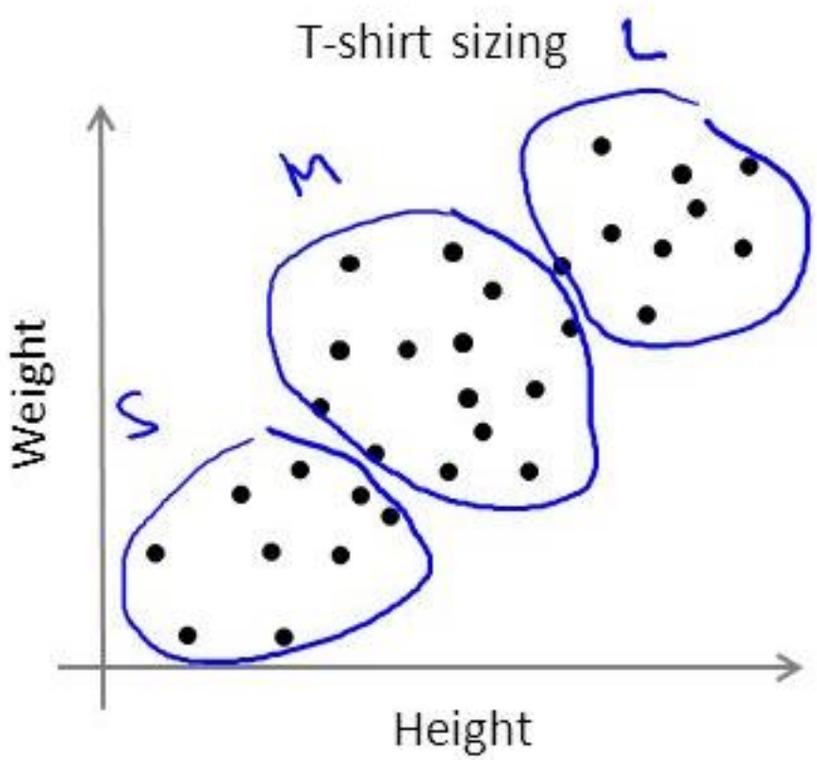
How much K? (Elbow)

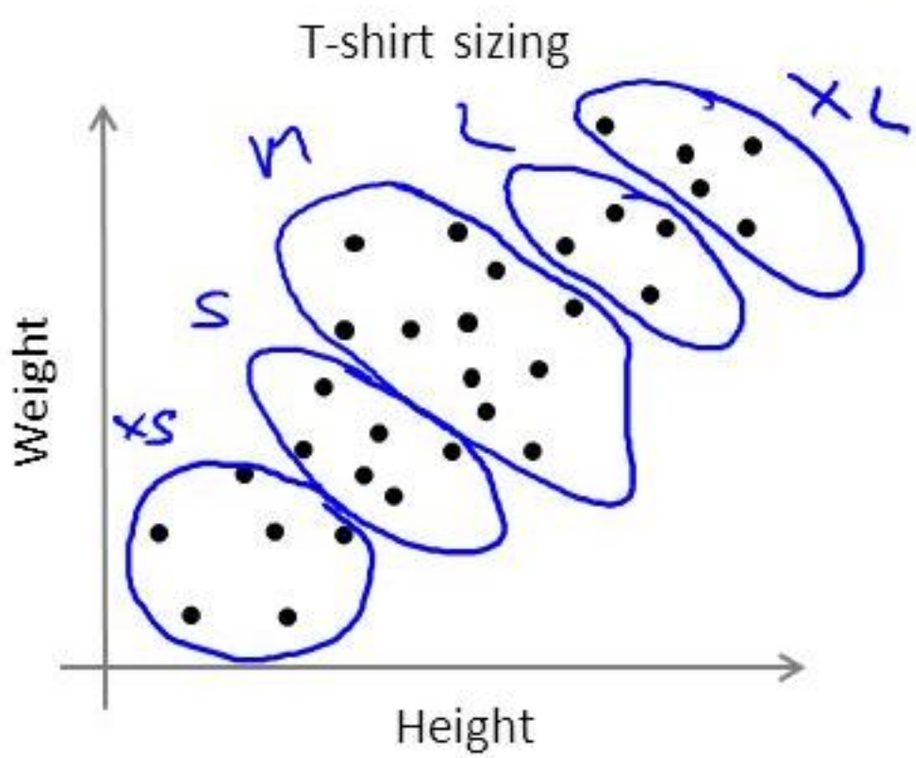
Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

K=3 S, M, L

K=5 XS, S, M, L, XL

E.g.





by Andrew Ng

Cluster Validation – Internal Indices

Measure the fit between data and the structure using only the data

Metric	Range	Available in sklearn
Silhouette Index	[-1,1]	Yes
Calinski-Harabasz		No
BIC		No
Dunn Index		No

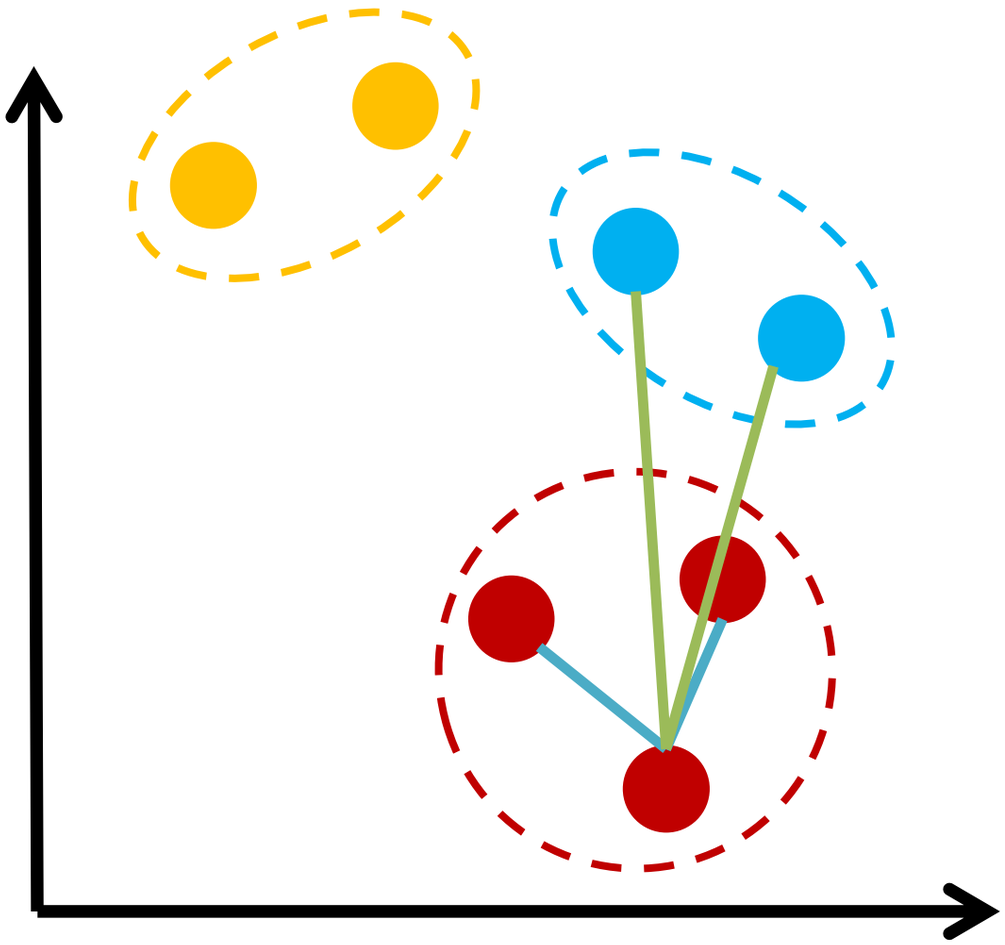
Source: Udacity Machine Learning Nano-Degree

Cluster Validation – Internal Indices

Silhouette Coefficient
Between -1 and 1

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

Clustering Result



a = average distance to other samples in same cluster

b = average distance to other samples in closest neighboring cluster

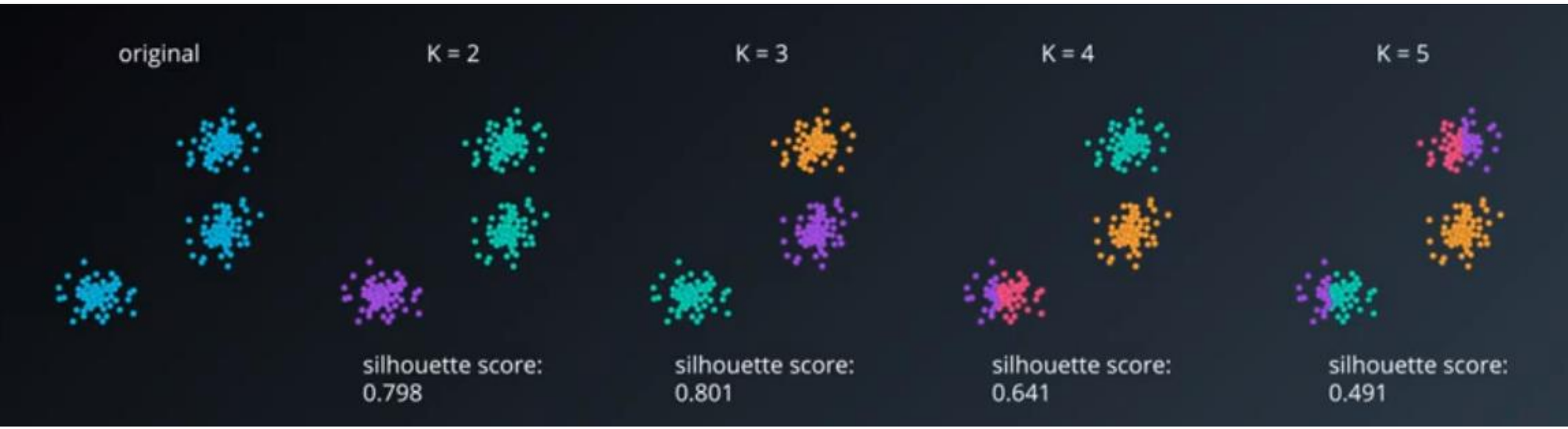
$$S = \text{Average}(S_1, S_2, S_3\ldots., S_n)$$

Source: Udacity Machine Learning Nano-Degree

DICE ANALYTICS

Cluster Validation – Internal Indices

Silhouette Coefficient – Finding K



Source: Udacity Machine Learning Nano-Degree

DICE ANALYTICS