Clustering

Week 07 - Day 02

Why clustering?

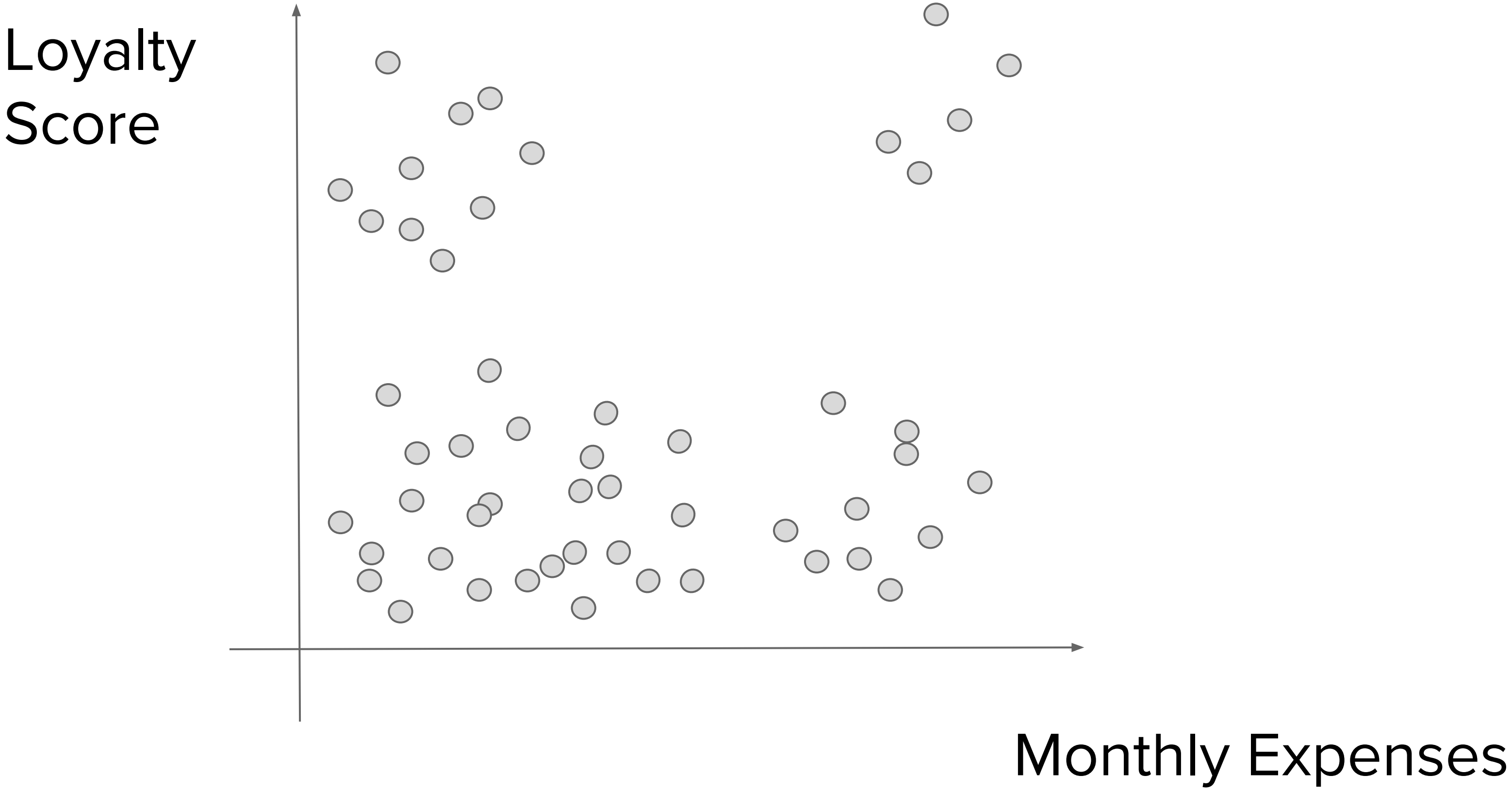
You need groups

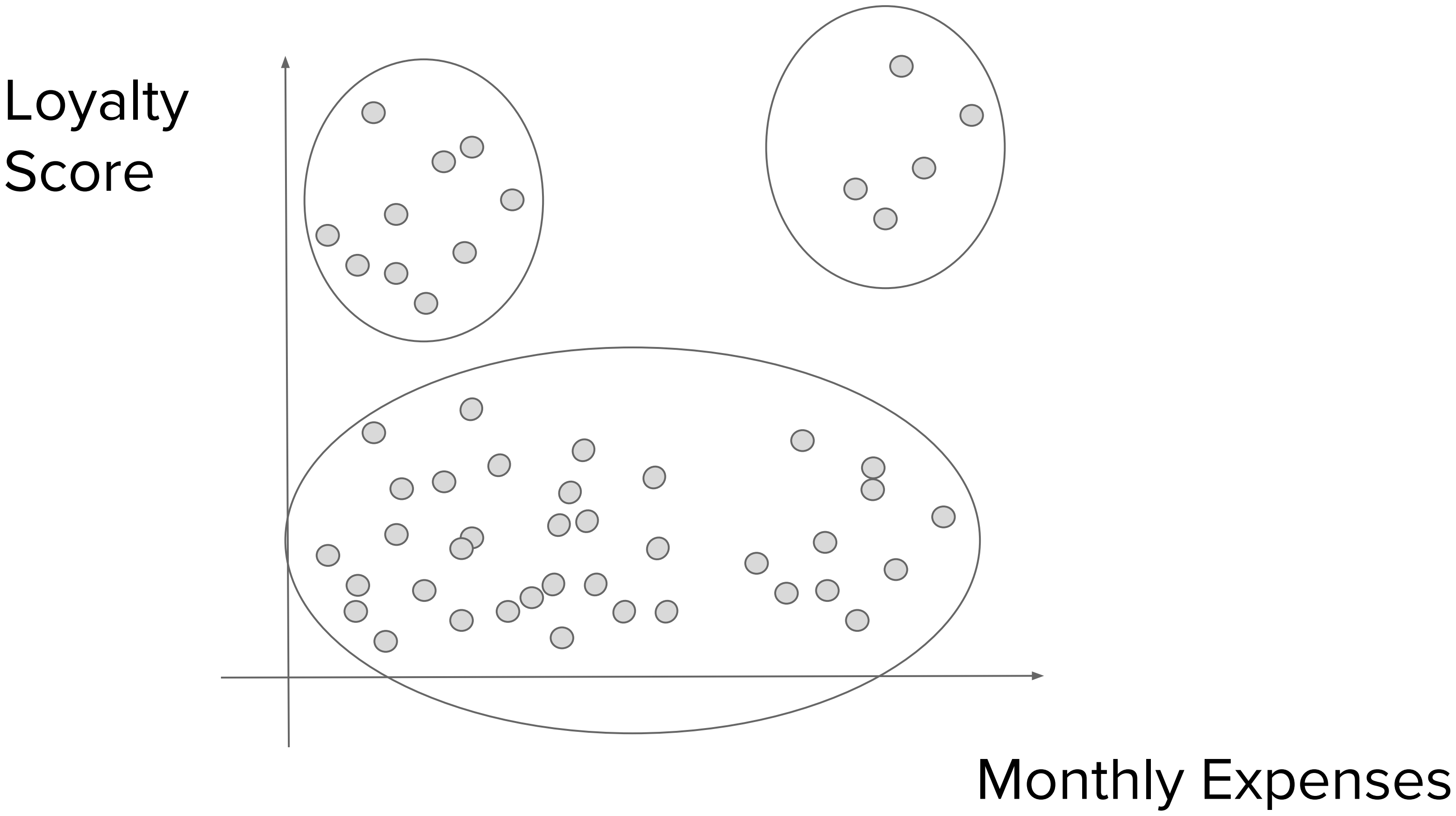
but

you don’t have labels!

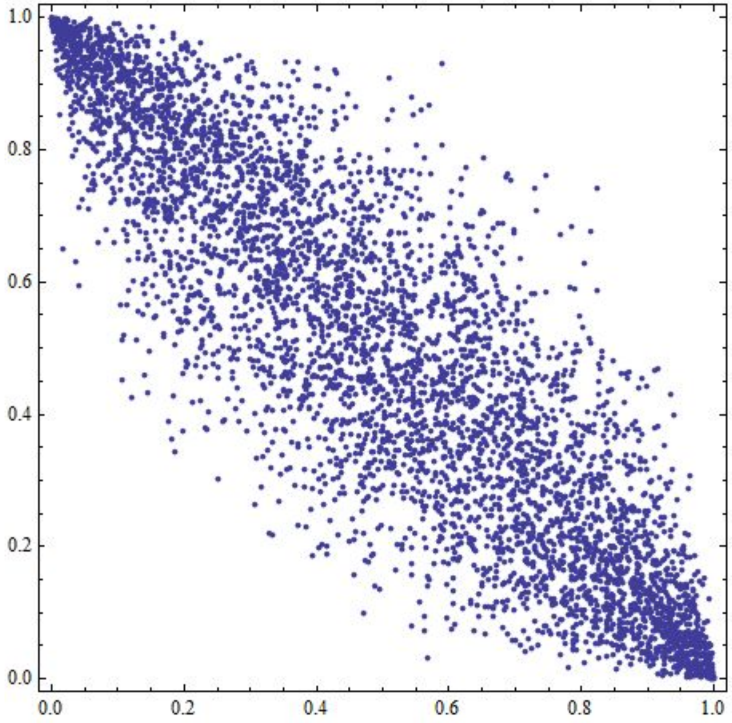
Create your own labels/groups

E.g. Identify customers to send a promotion to


Loyalty Score
Monthly Expenses


Loyalty Score
Monthly Expenses

Often it’s not that easy!



Can be a step during EDA

K-Means Clustering

Algorithms to extract k clusters

- Based on distances
- Iterative
- Random initialization
- K is an input

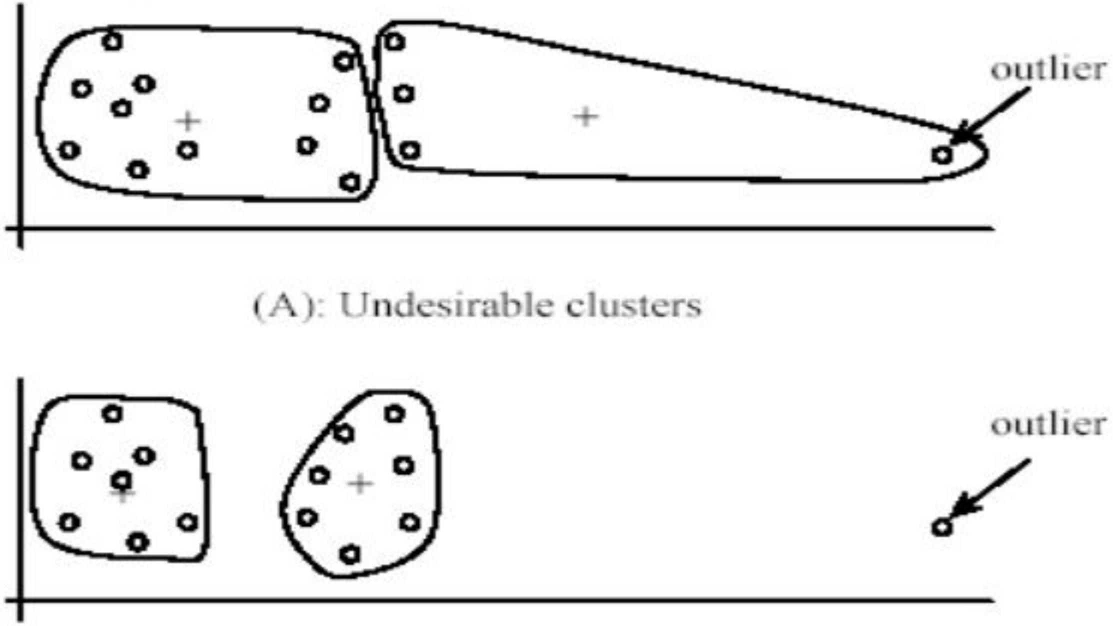
https://upload.wikimedia.org/wikipedia/commons/e/ea/K-means_convergence.gif

1. Choose k random centroids (i.e. points not in the data)
2. Assign data points to closer centroids
3. Recalculate centroids using the mean of each cluster
4. Repeat till stable

K-means is stochastic

(i.e. different runs = different results)

K-means is sensitive to outliers

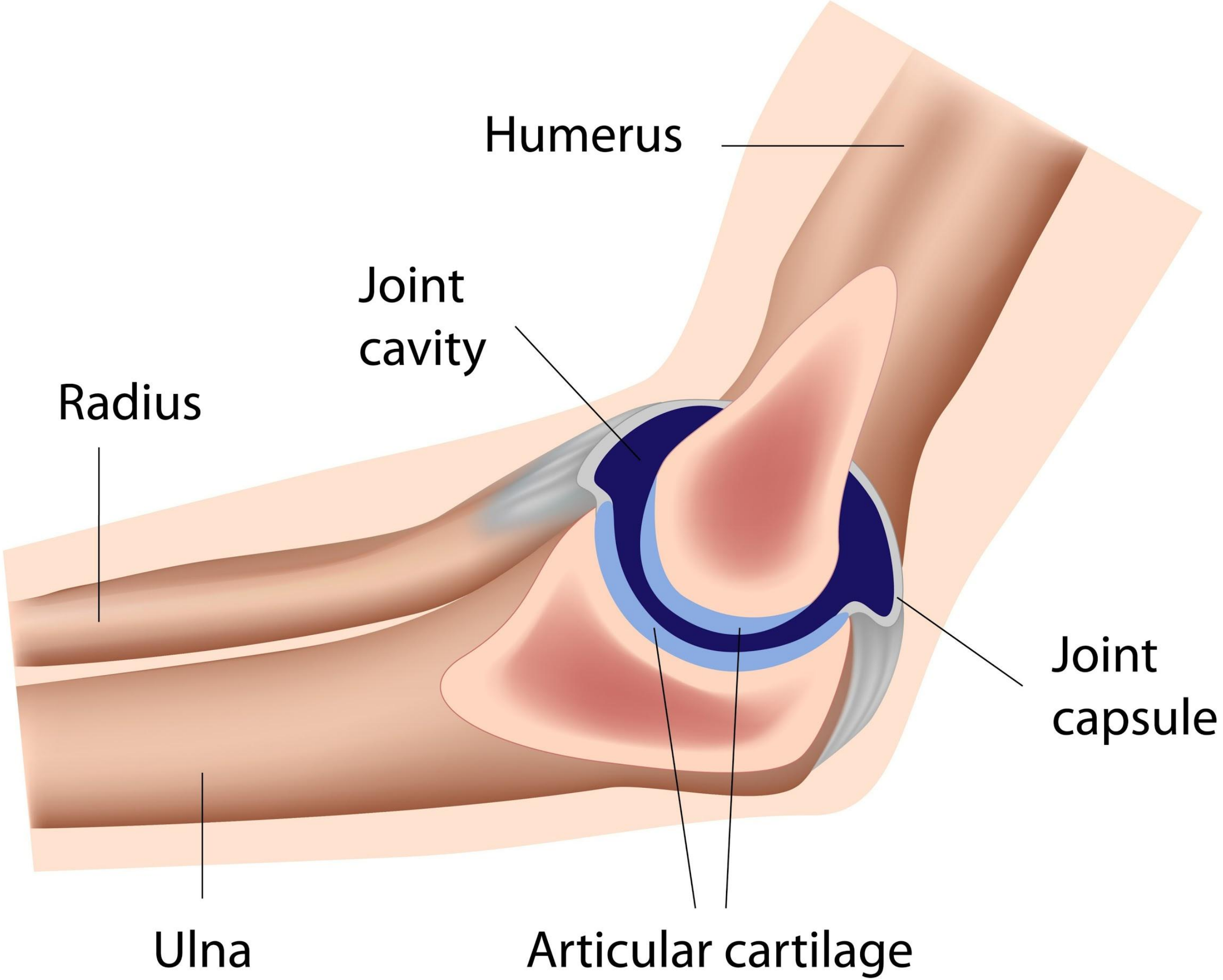


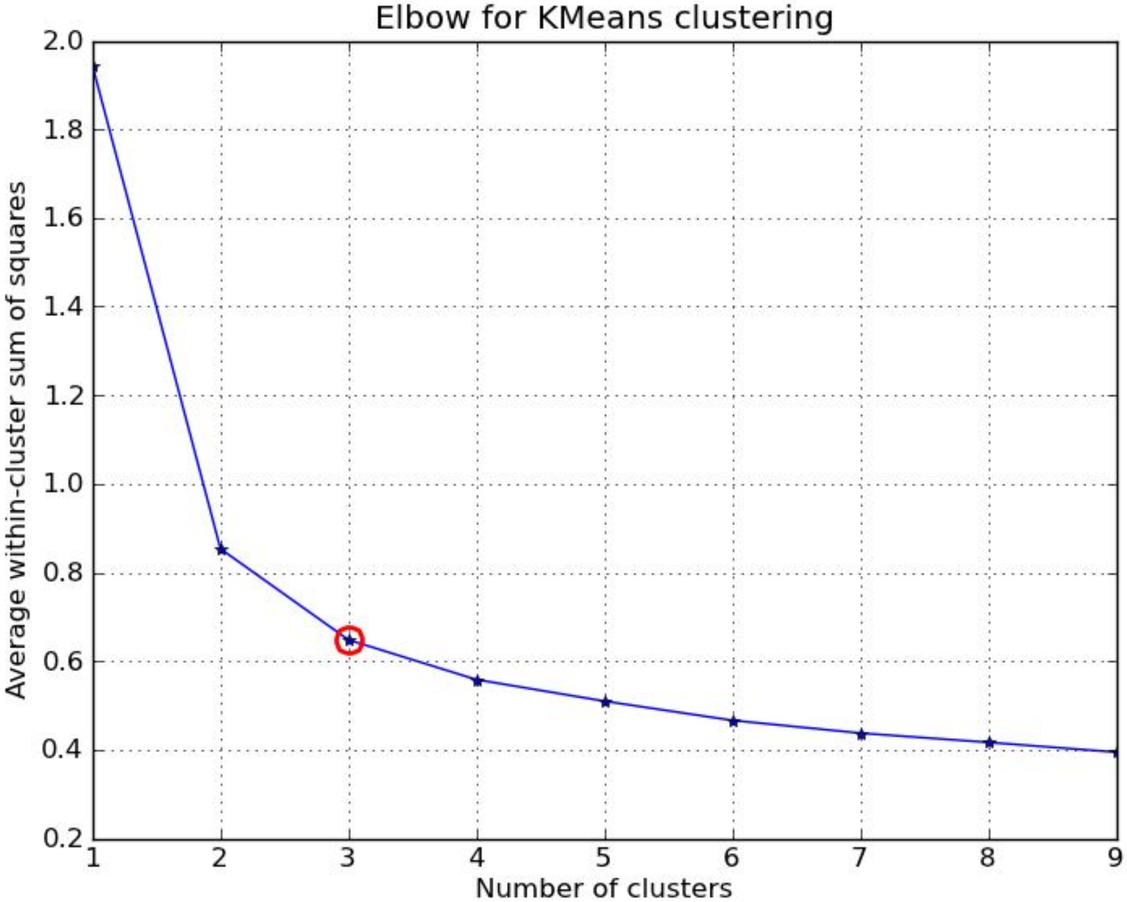
(A): Undesirable clusters

(B): Ideal clusters

Scale your features!

How to choose k


Humerus
Joint cavity
Radius
Joint capsule
Ulna
Articular cartilage


Elbow for KMeans clustering
Average within-cluster sum of squares
Number of clusters
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8
9

Lesson on thursday

Metrics for Clusters

Inertia

Sum of squared distance point-centroid

- Low Inertia = dense clusters
- Values from 0 to infinite

Silhouette

The measure of how far apart clusters are from each other.

- High silhouette score = clusters are well separated
- Values from -1 to +1

Lesson on thursday