Hyperparameter Tuning for Clustering Models

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Hyperparameter tuning in clustering algorithms

Using ParameterGrid in scikit-learn

Evaluating Clustering Models

Evaluating Clustering Models

Homogeneity	Completeness	V-measure
Adjusted Rand Index (ARI)	Adjusted Mutual Info	Silhouette

Evaluating Clustering Models

Homogeneity	Completeness	V-measure
Adjusted Rand Index (ARI)	Adjusted Mutual Info	Silhouette

Important advantage of Silhouette scoring: Does not require labeled data

Silhouette Score

Defines Silhouette coefficient for each sample

Measure of how similar an object is to objects in its own cluster

And how different it is from objects in other clusters

Overall Silhouette score averages Silhouette coefficient of each sample

No need for labeled data

Silhouette Coefficient

$$s^i = \frac{b^i - a^i}{\max(a^i\,,\,b^i)}$$

a^i = Mean distance of point i from all other points in same cluster

b^i = Mean distance of point i from all points in next nearest cluster

Silhouette Score

$$S = \text{Average}\ (s^i)$$

Overall score is average of coefficients for all points

Silhouette Score

Bounded between -1 (incorrect) and +1 (perfect) clustering

Scores around 0 indicate overlapping clusters

Tend to be higher for dense and well separated clusters

Hyperparameter Tuning for K-means Clustering

Hyperparameters

Model configuration properties that define a model, and remain constant during the training of the model

Understanding Hyperparameters

Model Inputs

Model Parameters

Model Hyperparameters

Understanding Hyperparameters

Model Inputs

Input data points, training dataset

Model Parameters

Model Hyperparameters

Understanding Hyperparameters

Model Inputs

Input data points, training dataset

Model Parameters

Reference vectors, i.e. centroids of each cluster

Model Hyperparameters

Understanding Hyperparameters

Model Inputs

Input data points, training dataset

Model Parameters

Reference vectors, i.e. centroids of each cluster

Model Hyperparameters

Number of clusters, initial values, distance measure

Hyperparameters in K-Means Clustering

Number of clusters

Seeds - initial values

Distance measures

Number of Clusters

K is the most important hyperparameter

Sometimes obvious e.g. 10 in MNIST digit classification

Otherwise, apply standard method to find the "best" value of K

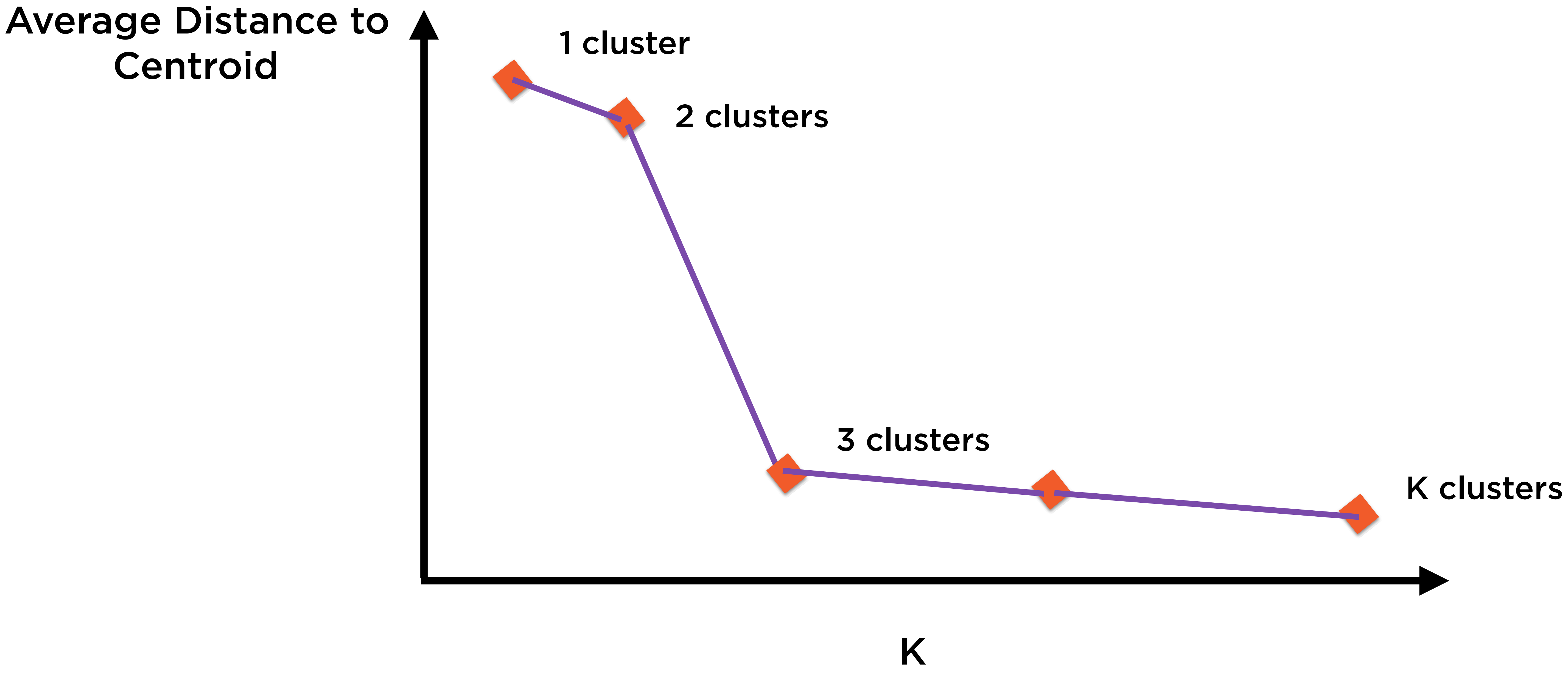
Elbow Method

Pick range of candidate values of K (e.g. 1 to 10)

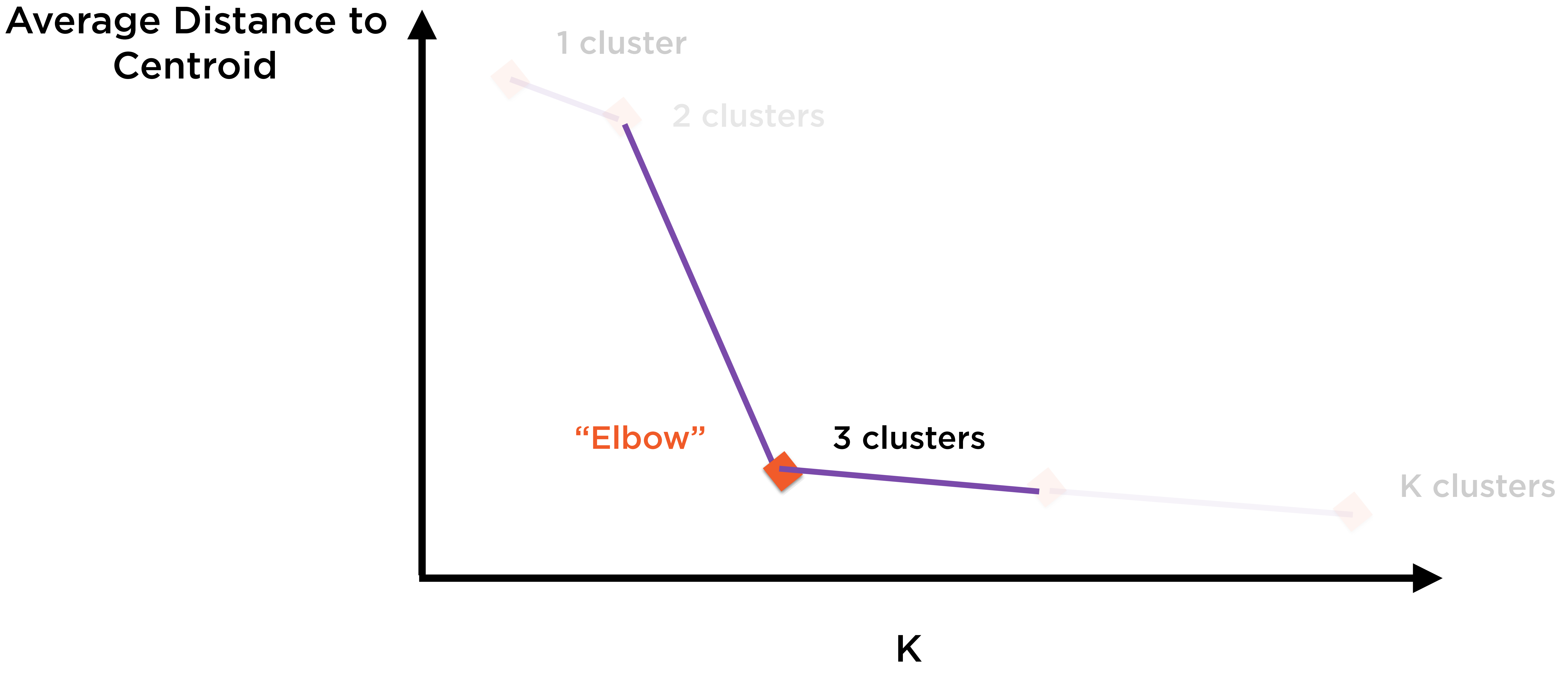
Calculate average distance from centroid for each value

Plot and find “elbow”

Elbow Method



Elbow Method



Silhouette Method

Pick range of candidate values of K (e.g. 1 to 10)

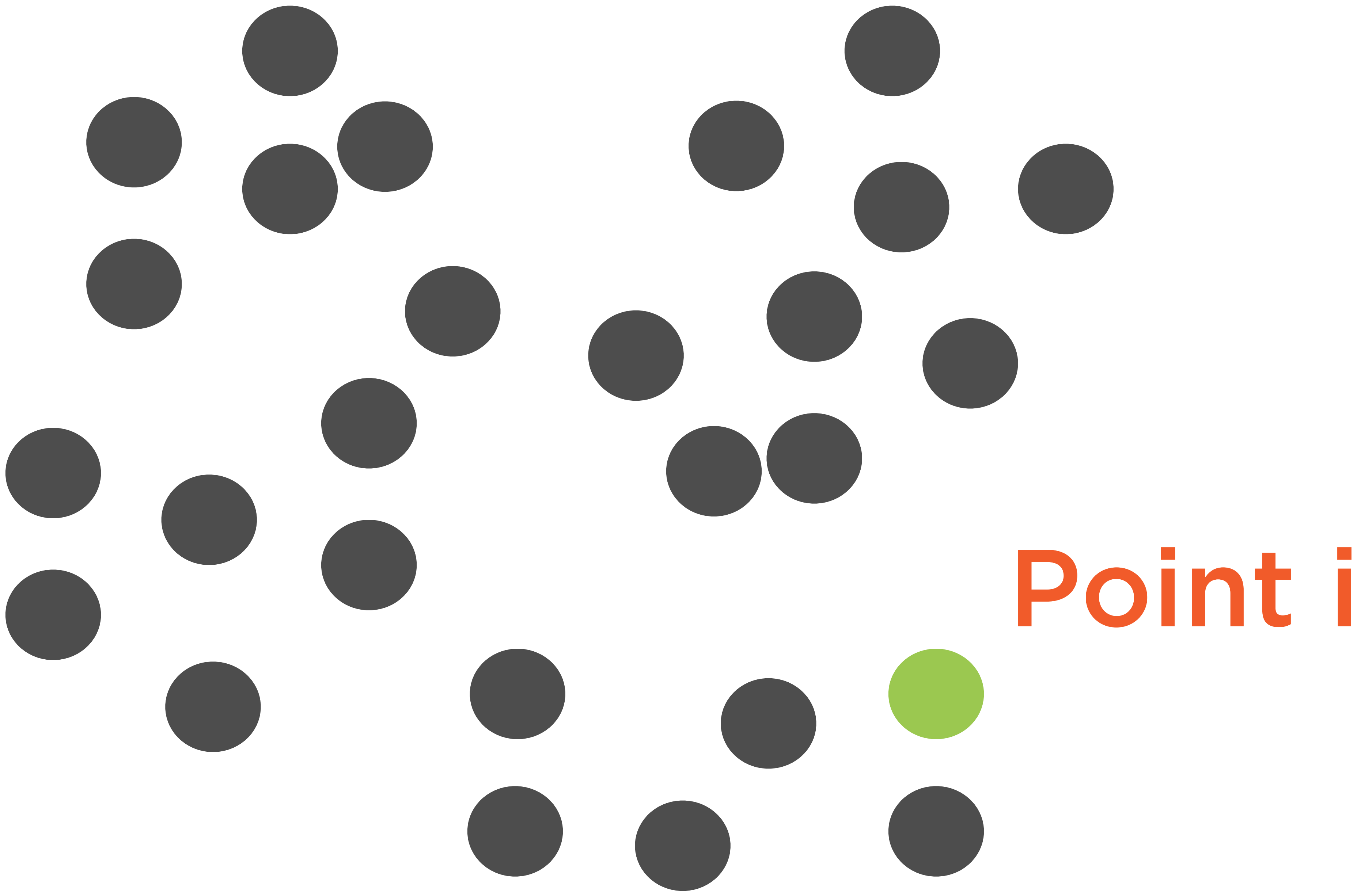
Plot silhouettes for each value of K

Ideal value of silhouette = 1

Worst possible value of silhouette = -1

Silhouette Coefficient

For any point i, calculate silhouette coefficient



Silhouette Coefficient

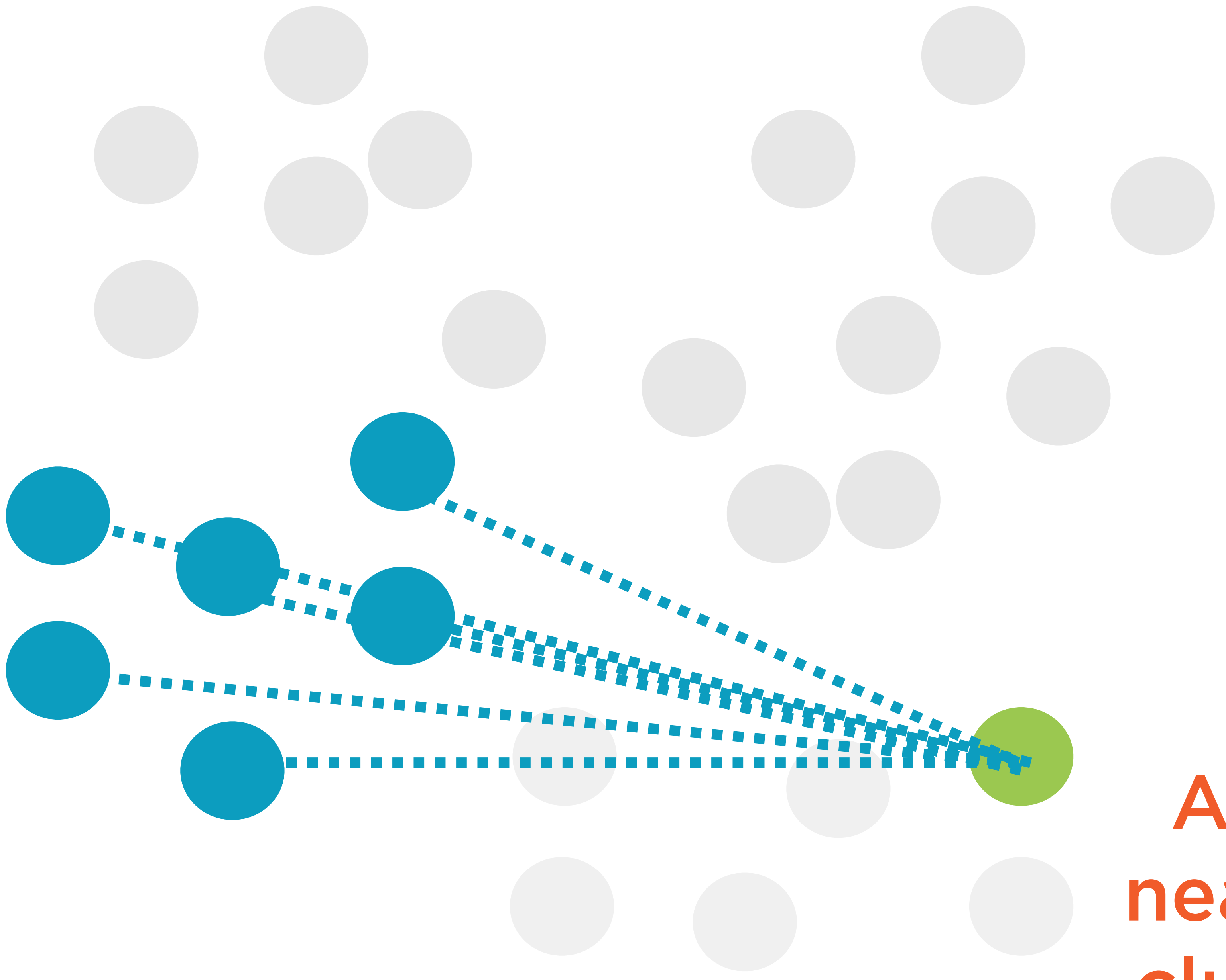
For any point i, calculate silhouette coefficient

Silhouette Coefficient

Find a(i) = average distance of i to other points in same cluster

Silhouette Coefficient

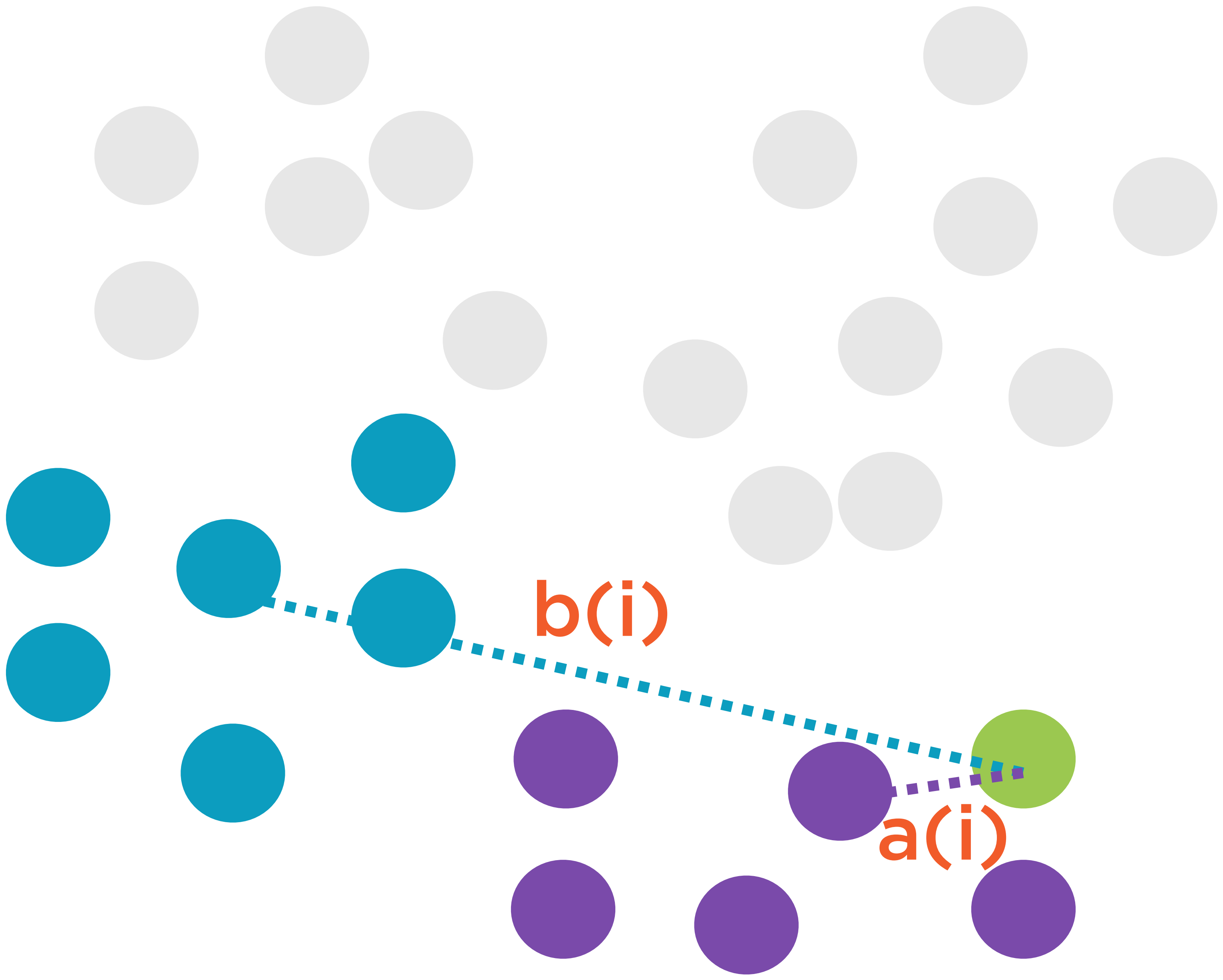
Find b(i) = average distance to nearest other cluster



Average to nearest other cluster = b(i)

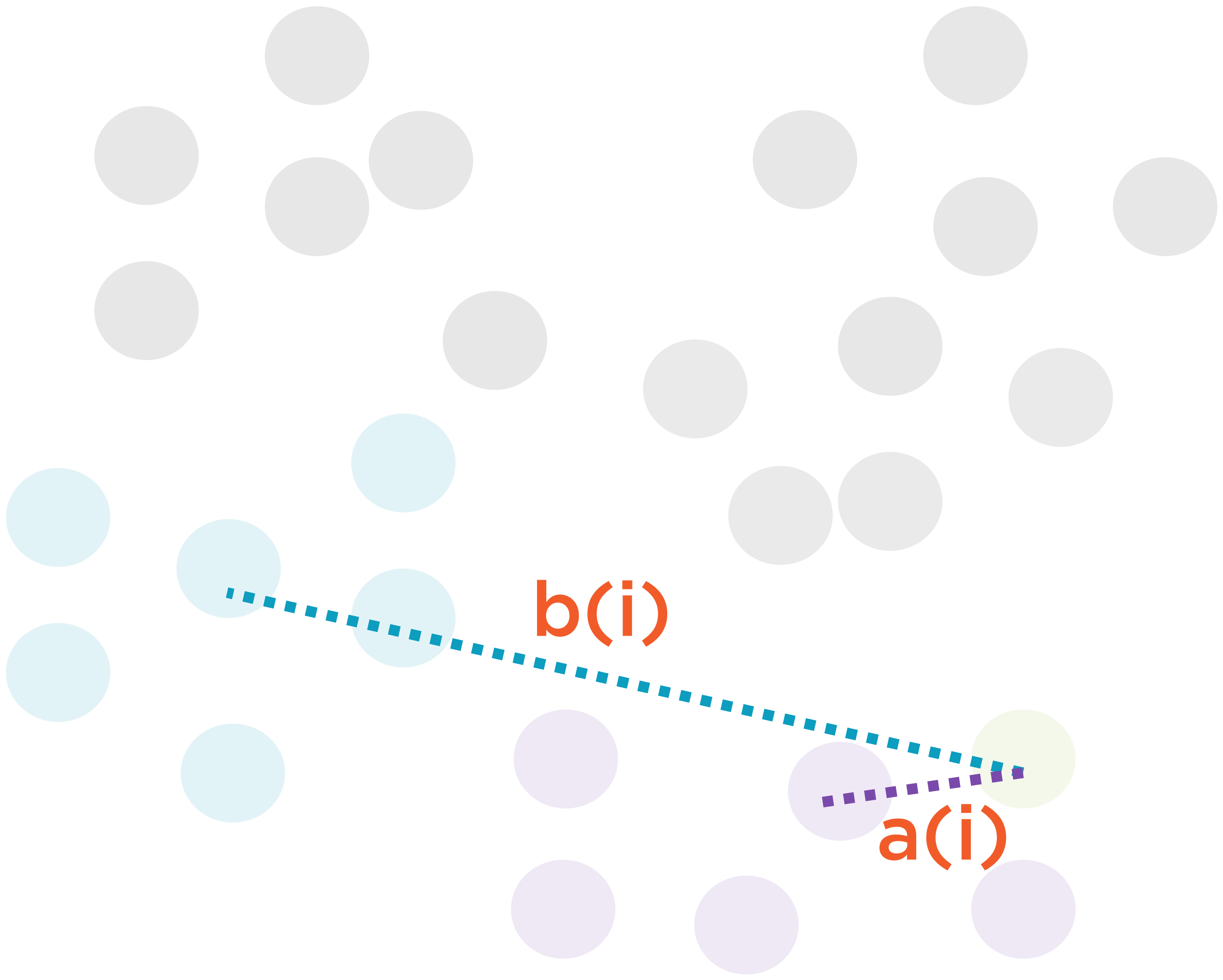
Silhouette Coefficient

Ideally, a(i) << b(i)



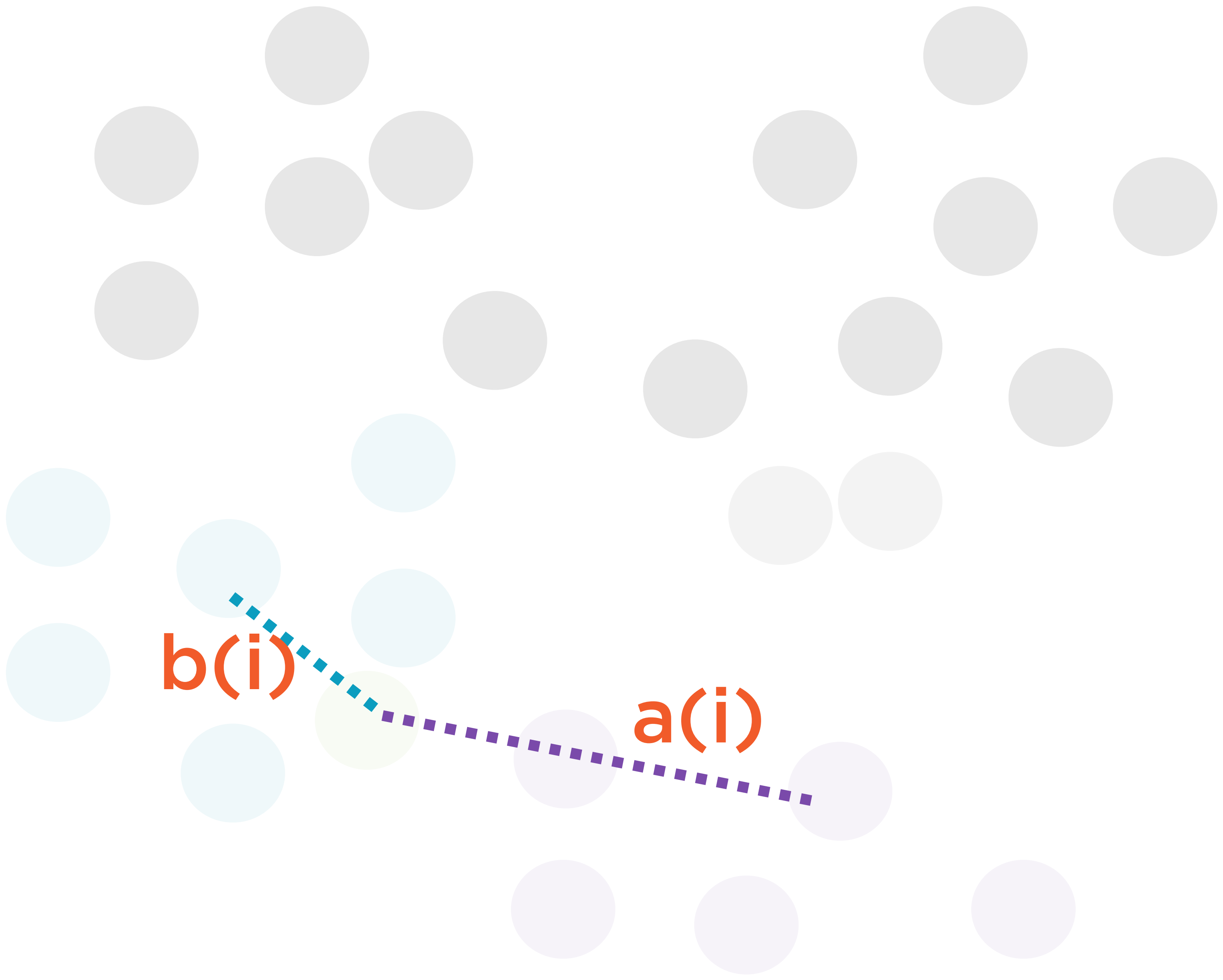
Silhouette Coefficient

Ideally, a(i) << b(i)

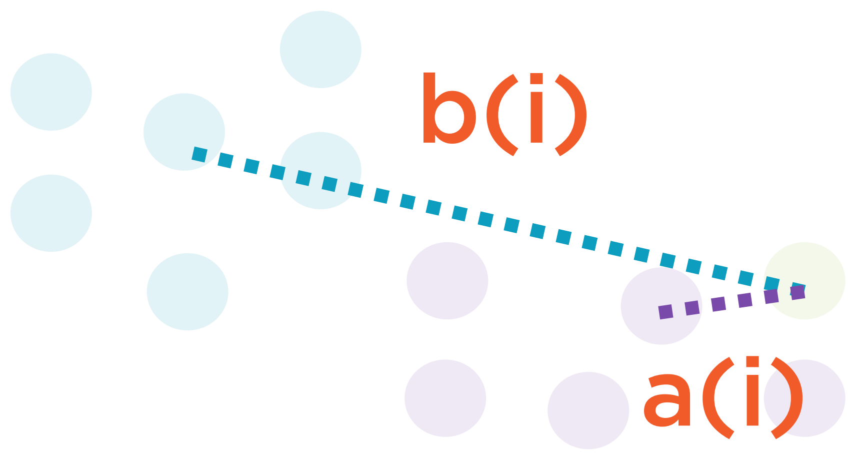


Silhouette Coefficient

If a(i) > b(i), i is likely misclassified



Silhouette Coefficient



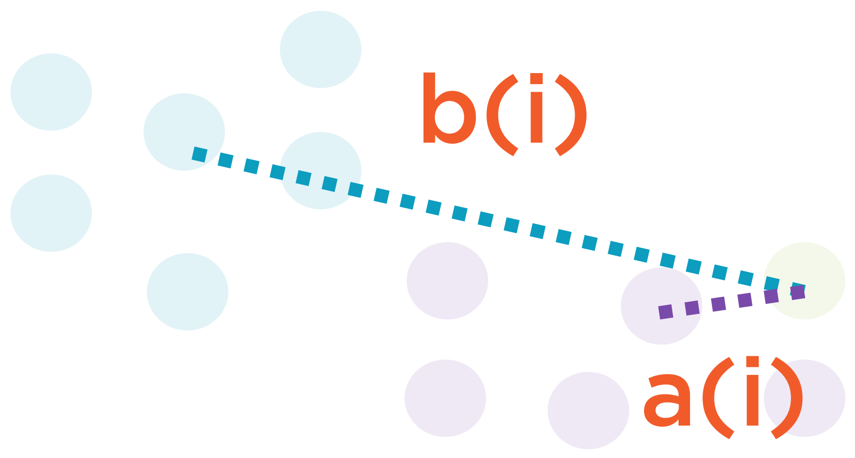
For any point i

$$s(i) = \frac{b(i) - a(i)}{\text{Larger of } b(i) \text{ and } a(i)}$$

a(i) = Average distance inside cluster

b(i) = Average distance to nearest other cluster

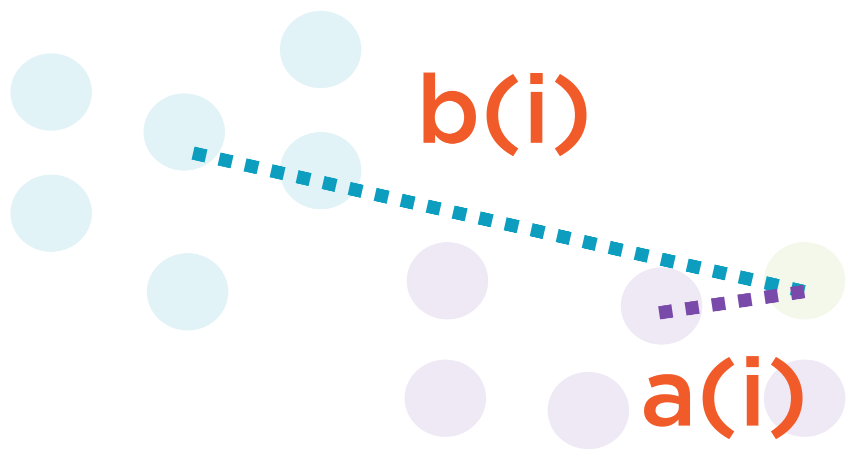
Ideally s(i) = 1



Ideally, a(i) = 0, b(i) = Infinity

$$s(i) = \frac{b(i) - a(i)}{\text{Larger of } b(i) \text{ and } a(i)} = 1$$

Worst-case s(i) = -1

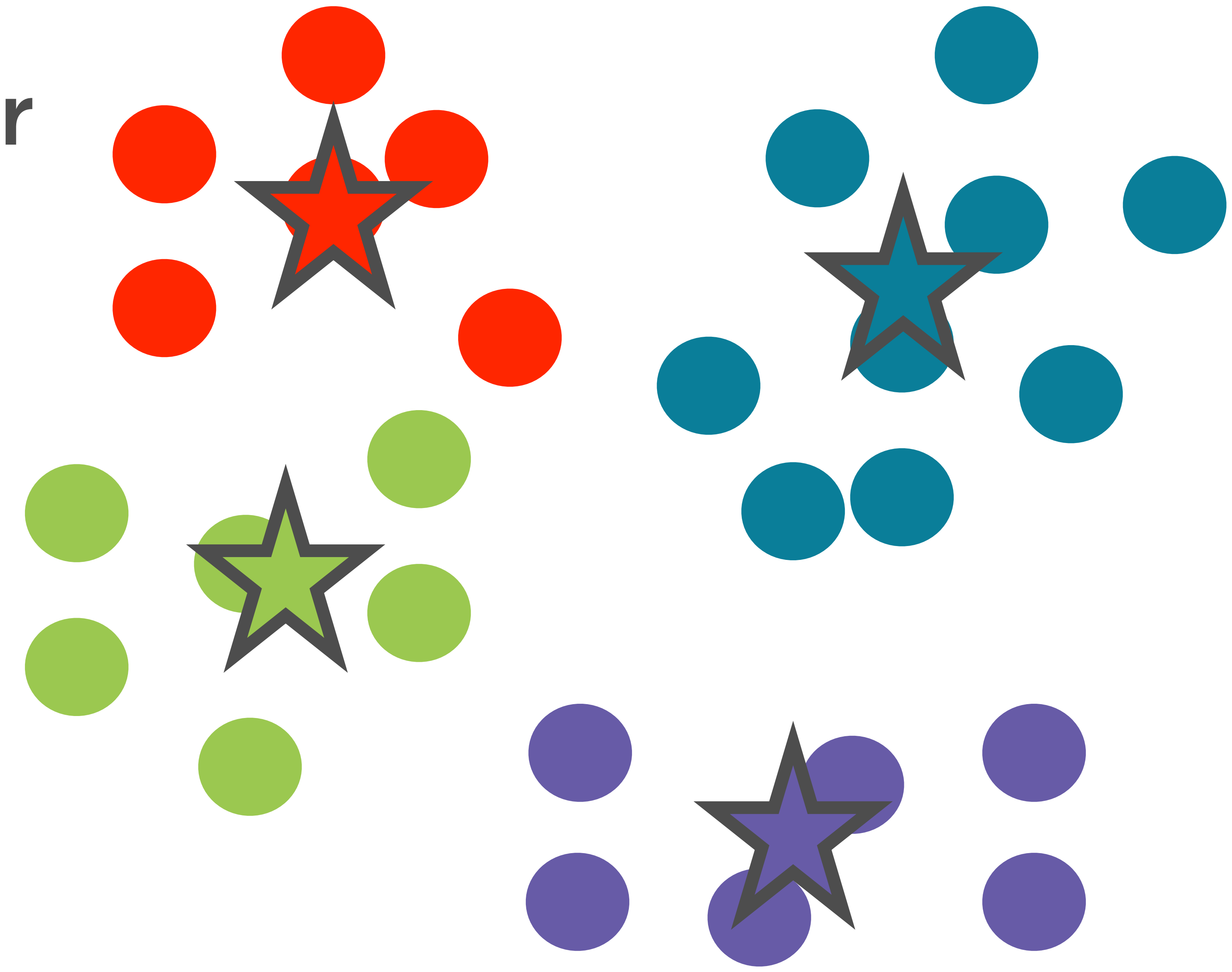


Worst case, a(i) = Infinity, b(i) = 0

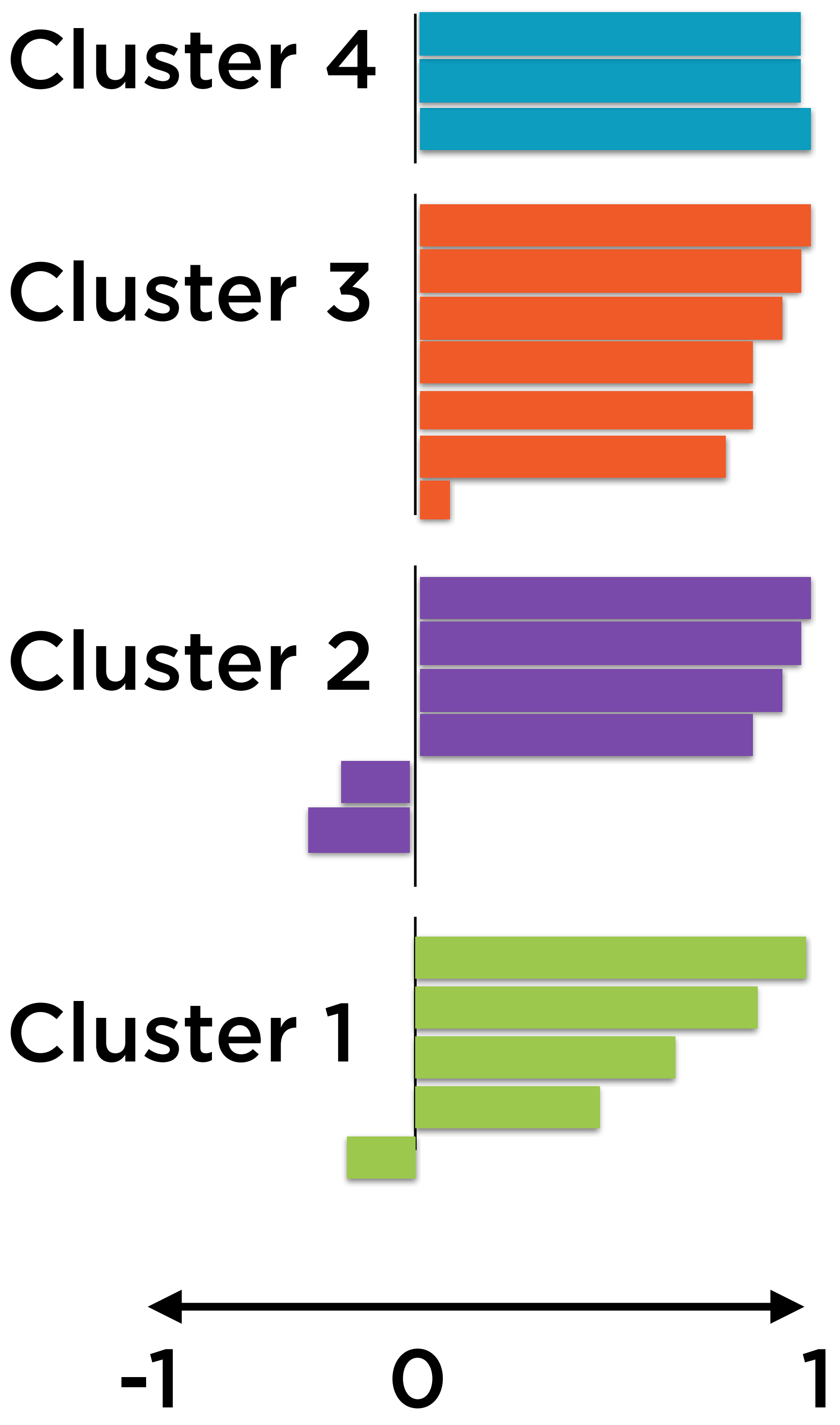
$$s(i) = \frac{b(i) - a(i)}{\text{Larger of } b(i) \text{ and } a(i)} = -1$$

Silhouette Plot

Calculate s(i) for each point



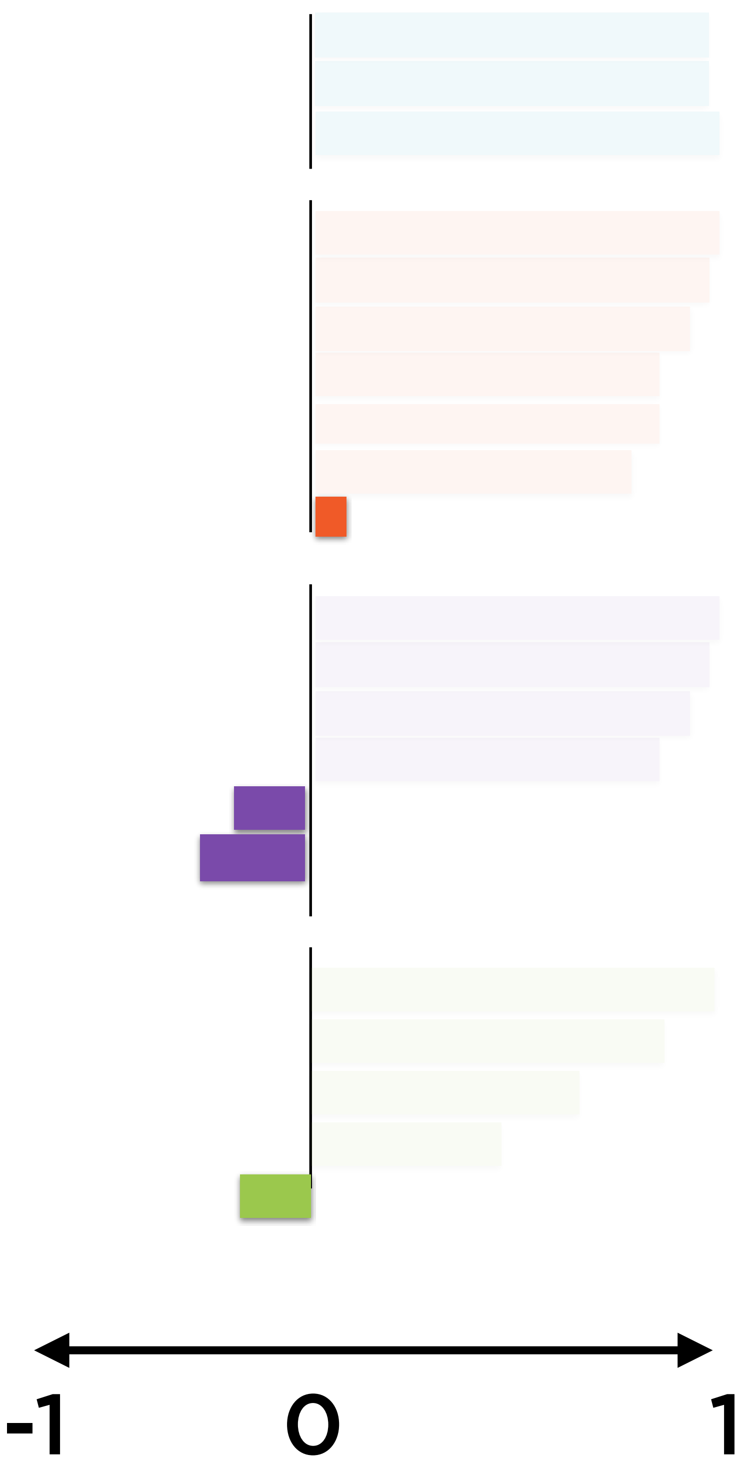
Silhouette Plot



Calculate s(i) for each point

Plot value of s(i) to identify outliers

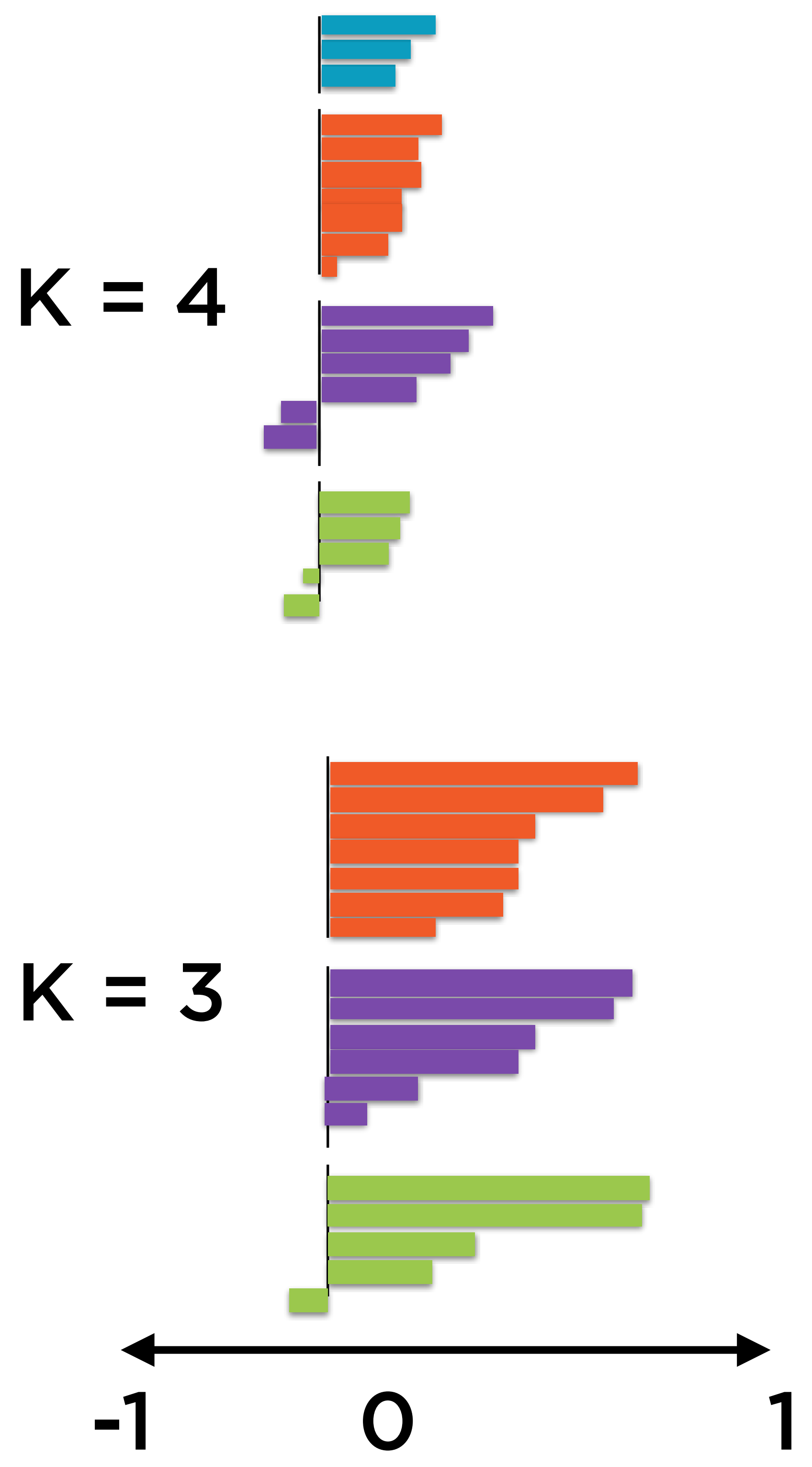
Outliers



Ideally, s(i) = 1

So, s(i) < 0 indicates outliers

“Best” K

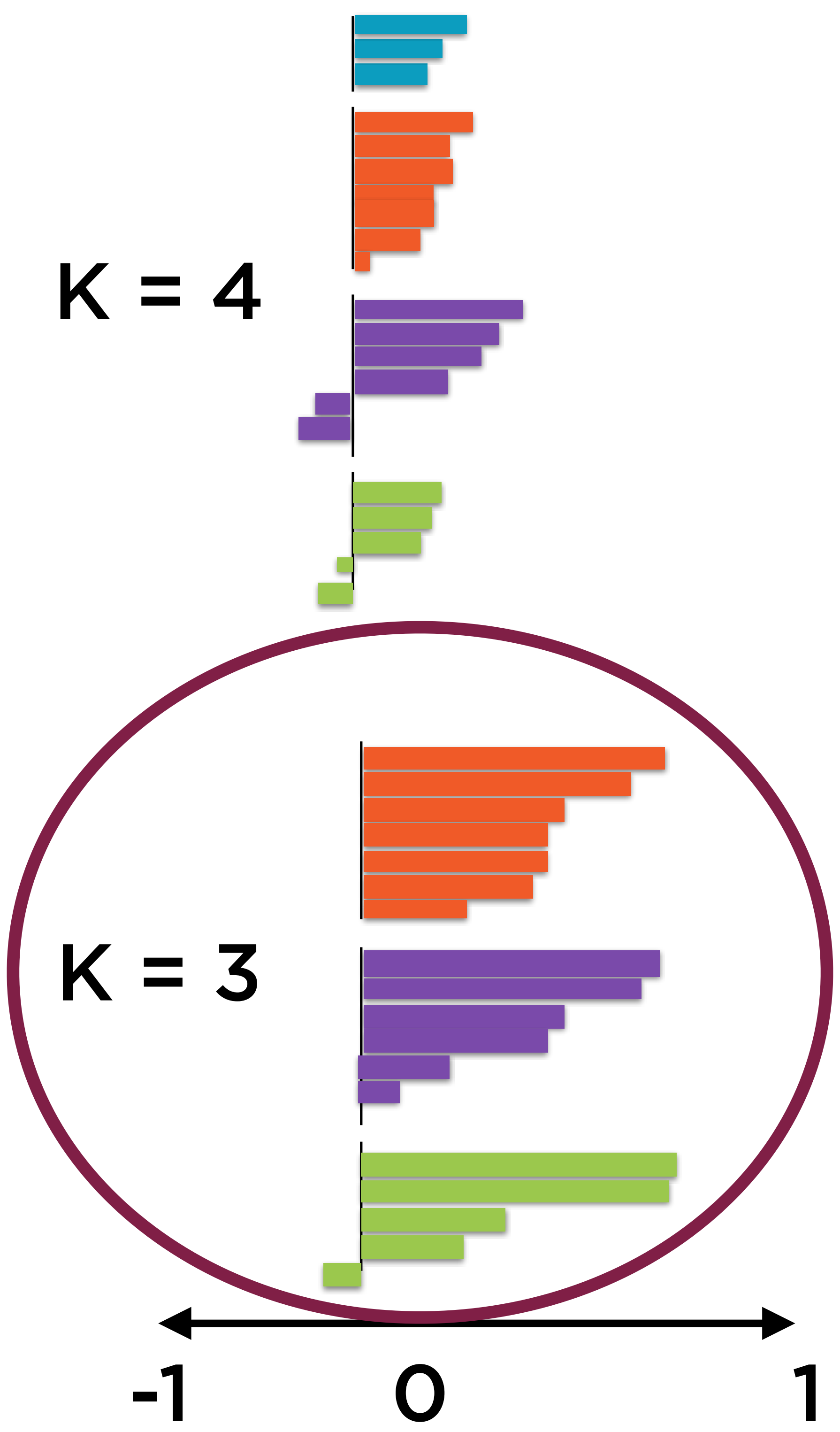


Extend the same idea

Replicate plot for different values of K

Pick K where average silhouette is closest to 1

“Best” K



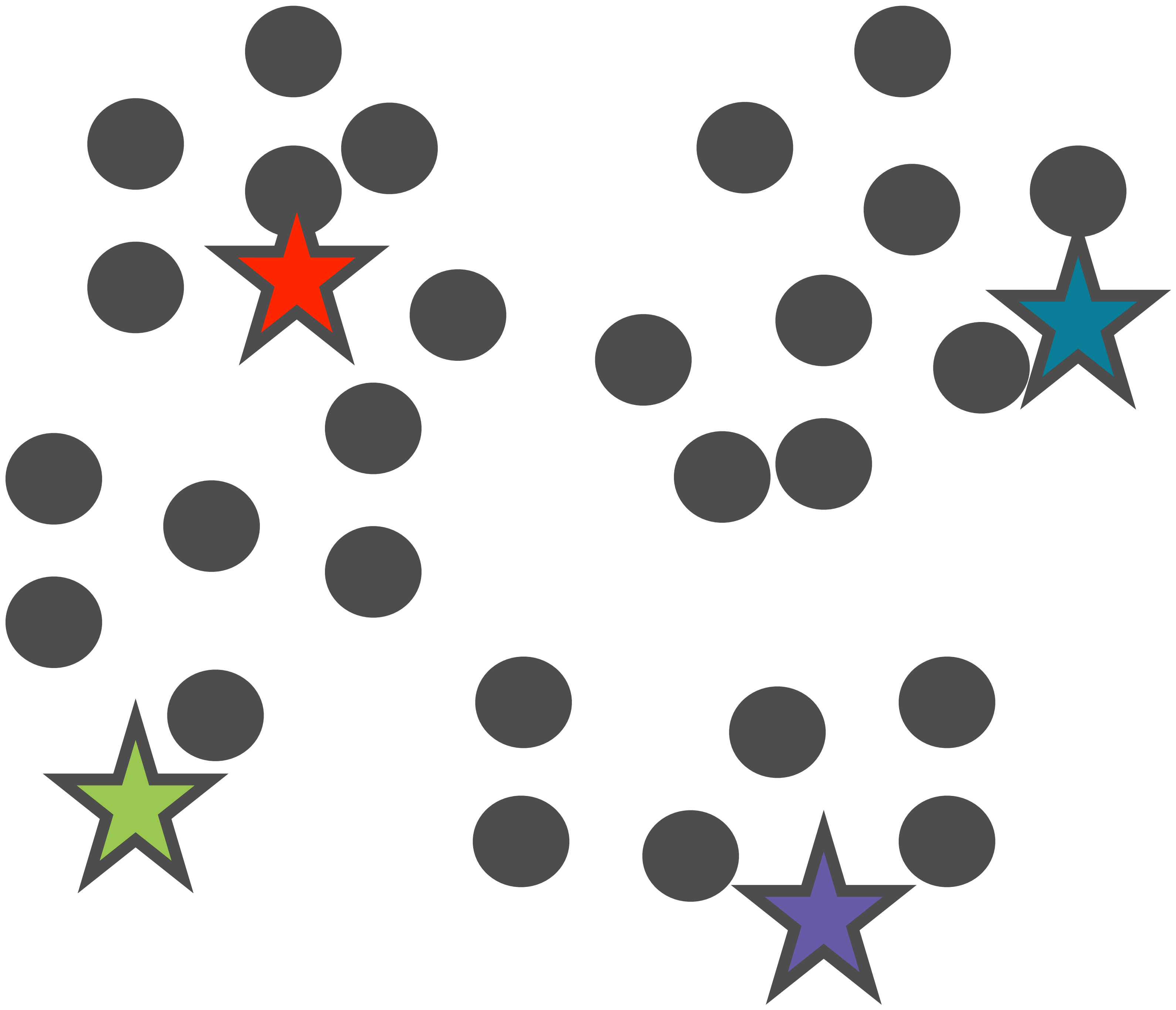
Here K = 3 is noticeably better than K = 4

K = 3 has noticeably larger positive values

Seeds

Final reference vector values sensitive to initial values

Random initialization might not work - examine data carefully



Seeds

Final reference vector values sensitive to initial values

Random initialization might not work - examine data carefull

- Can perform PCA of data
- Divide range of normalized PCs into K
- Take average of each

Distance Measures

Can choose multiple distance measures:

- Euclidean distance - centroid might not be actual data point
- Mahalanobis distance - normalize each dimension to have equal variance
- Cosine distance - cosine of angle between point and centroid

Demo

Hyperparameter tuning for K-means clustering, DBSCAN clustering and mean-shift clustering

Summary

Hyperparameter tuning in clustering algorithms

K-means clustering, DBSCAN, mean-shift clustering

Using ParameterGrid in scikit-learn