Advanced Discovery: Silhouette Analysis

Measuring How Well Points Fit Their Clusters

The Silhouette Score Formula

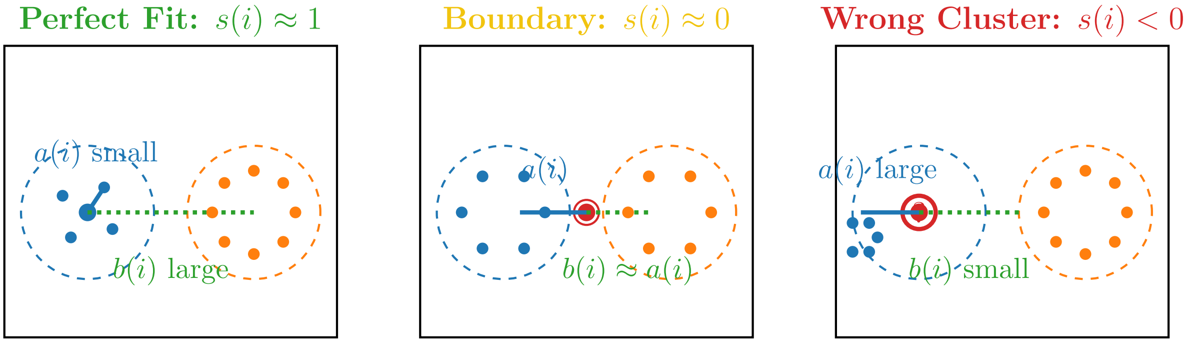
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Range: [-1, 1]

Within-cluster distance

Nearest-cluster distance

Visual Intuition: Three Cases



Perfect Fit: $s(i) \approx 1$

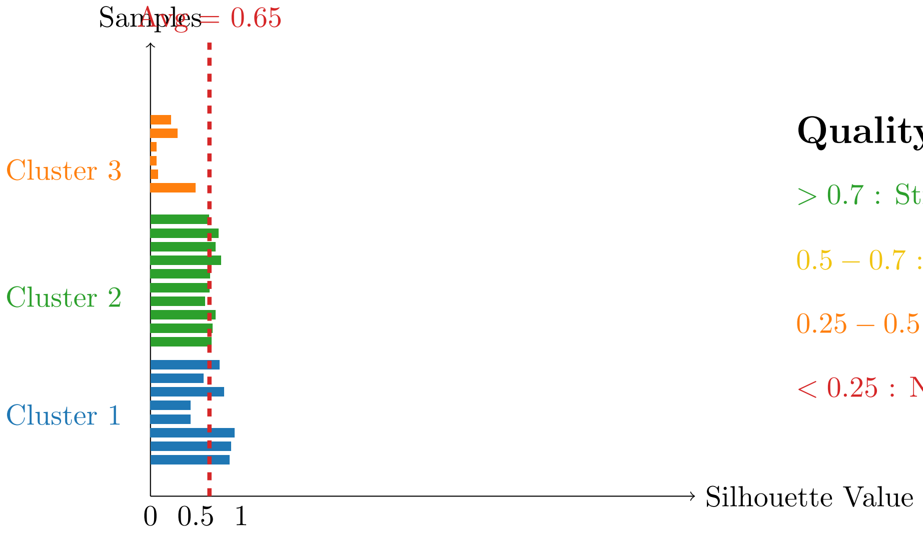
Boundary: $s(i) \approx 0$

Wrong Cluster: $s(i) < 0$

Cluster Quality Metrics Comparison

Metric	Formula	Range	Interpretation
Silhouette	$\frac{b-a}{\max(a,b)}$	$[-1, 1]$	Higher = better
Davies-Bouldin	$\frac{1}{k}\sum \max \frac{\sigma_i+\sigma_j}{d_{ij}}$	$[0, \infty)$	Lower = better
Calinski-Harabasz	$\frac{B/(k-1)}{W/(n-k)}$	$[0, \infty)$	Higher = better
Dunn	$\frac{\min d_{inter}}{\max d_{intra}}$	$[0, \infty)$	Higher = better
Inertia	$\sum \lvert\lvert x_i - c_j \rvert\rvert^2$	$[0, \infty)$	Lower = better

Advanced: Average Silhouette Width



Quality Scale:

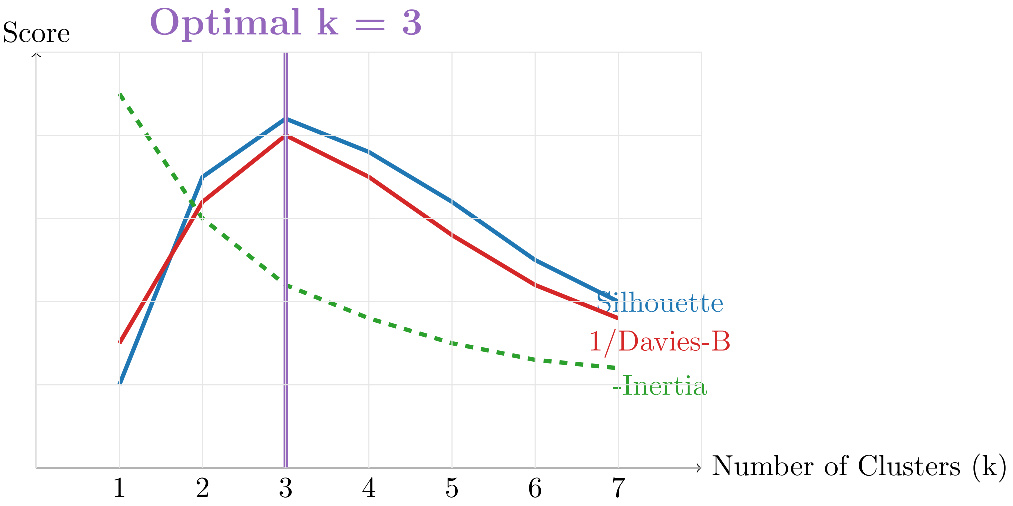
> 0.7 : Strong structure

0.5 – 0.7 : Reasonable

0.25 – 0.5 : Weak

< 0.25 : No structure

Discovery Challenge: Optimize k



Your Investigation

Given 100 points, 4 metrics, how do you decide k?

Silhouette says: k = ____ Davies-B says: k = ____ Inertia says: k = ____

When metrics disagree, which wins? Why?

Next: DBSCAN - When you don't know k at all!