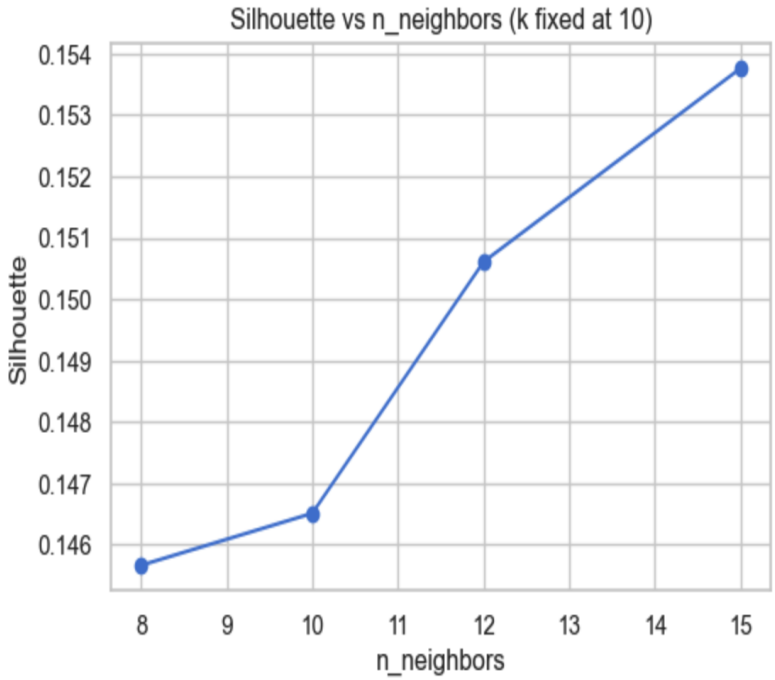
Clustering Handwritten Digits with Spectral Clustering

Dataset & Preparation

- Dataset: sklearn digits (1,797 images, 8×8 pixels, 64 features, labels 0–9)
- StandardScaler used to normalize pixel features
- PCA to 40 components (denoise & speed up graph construction)

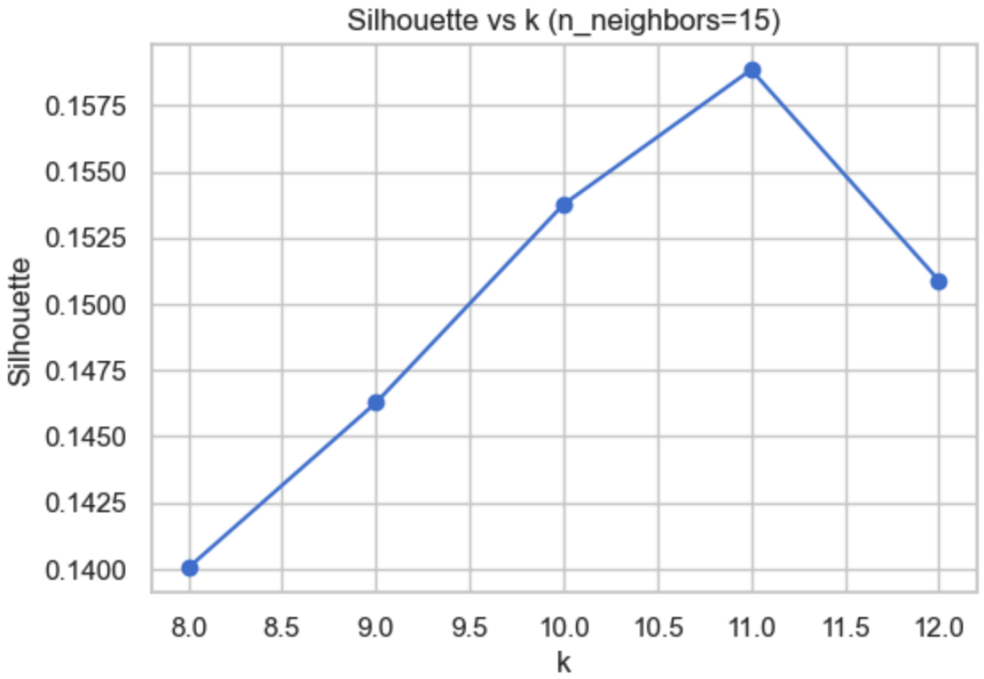
n_neighbors

- Affinity = nearest_neighbors (builds a k-NN graph)
- Sweep n_neighbors ∈ {8, 10, 12, 15} with k fixed at 10
- Visualization: Silhouette vs n_neighbors
- Best: n_neighbors = 15 (highest silhouette in the sweep)



Pick k (clusters)

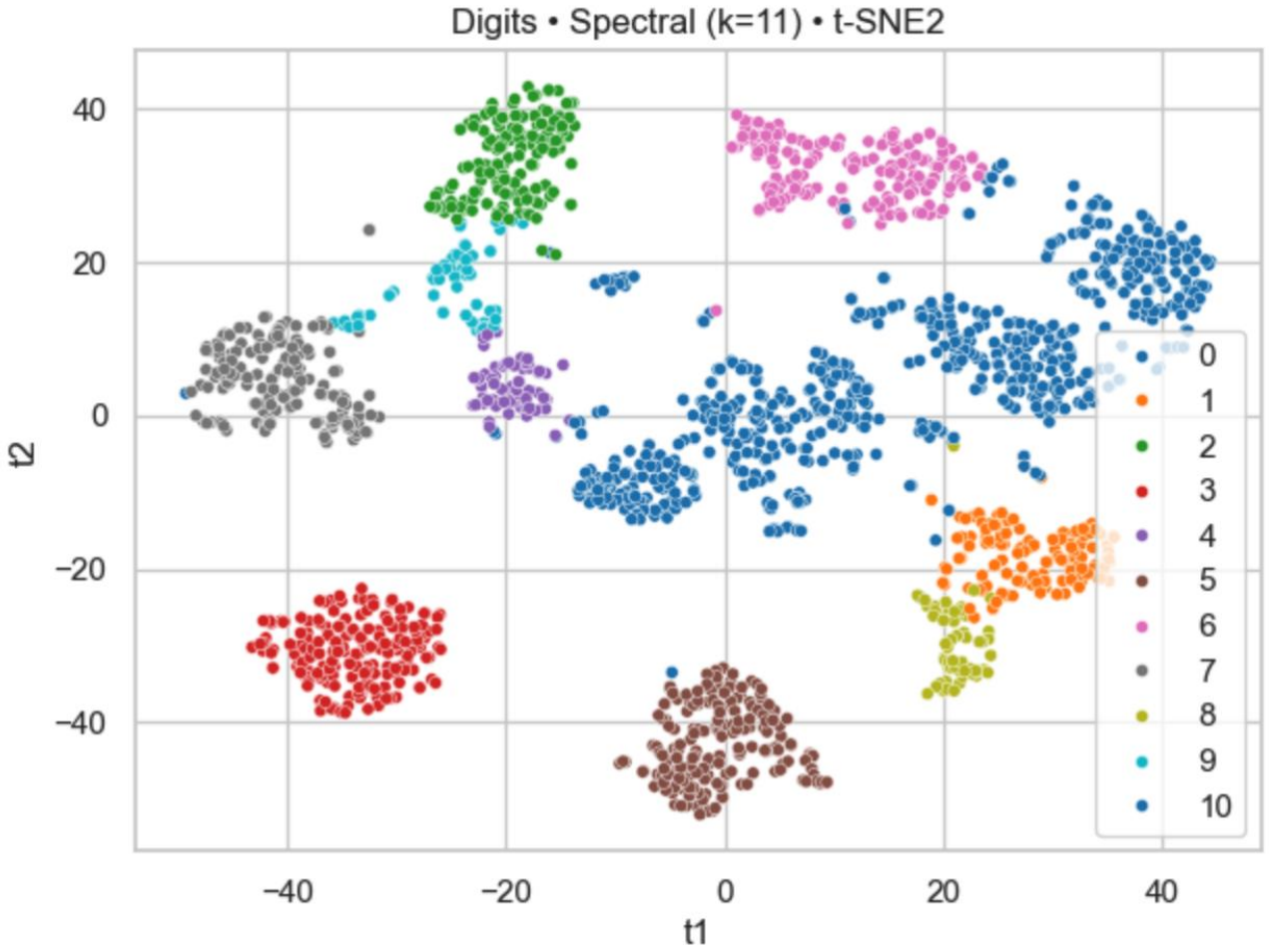
- With n_neighbors fixed at 15, sweep k ∈ [8..12]
- Best k in sweep: 11 (peak silhouette)



Final Spectral Model

- SpectralClustering(affinity='nearest_neighbors', n_neighbors=15, assign_labels='kmeans')
- n_clusters = 11 (from silhouette sweep)
- Fit on PCA(40) space to stabilize distances

Visualization — t-SNE (2D)



Visualization — t-SNE (2D)

- Project PCA(40) features with t-SNE to 2D
- Visualization: t-SNE scatter colored by spectral cluster label
- Shows clear groups; a few bridges/outliers are expected in digits

Evaluation

- Silhouette ≈ 0.159 (reasonable for high-dimensional digits)
- Adjusted Rand Index (ARI) ≈ 0.642 (good agreement with true digits)
- Normalized Mutual Information (NMI) ≈ 0.788 (strong label-invariant overlap)

What the Plots Tell Us

- Silhouette vs n_neighbors: graph connectivity matters; 15 gave the cleanest separation
- Silhouette vs k: peak around 11 shows subtle class splits (some digits form sub-clusters)
- t-SNE scatter: clusters match digit shapes; some overlap where digits look alike (e.g., 4/9)

Takeaways

- Don’t guess parameters — select n_neighbors and k by silhouette curves
- Use PCA before spectral to denoise; t-SNE for human-friendly visuals
- Report label-invariant metrics (ARI/NMI) to quantify alignment with true digits