@S S Roy,9th Sept,2025

Hierarchical clustering

---

1. What hierarchical clustering *is* (precise conceptual statement)

Hierarchical clustering produces a **nested family of partitions** $\{C^{(t)}\}_t$ of a dataset $X = \{x_1, \ldots, x_n\}$ indexed by a scale parameter (height) t. Equivalently it constructs a rooted tree (dendrogram) whose leaves are the individual points and internal nodes represent successive merges (agglomerative) or splits (divisive). The dendrogram induces a **cophenetic (ultrametric) distance** $c(i, j)$: the height at which leaves i and j first share a common ancestor. This cophenetic distance is an ultrametric and encodes the complete hierarchy.

---

2. Two canonical frameworks

- **Agglomerative (bottom-up)** — start with n singleton clusters, repeatedly merge the "nearest" pair until one cluster remains.
- **Divisive (top-down)** — start with one cluster and recursively split (less common in practice).

All agglomerative methods differ only in the **linkage** (how inter-cluster distance is defined) and the **base metric** $d(x_i, x_j)$.

---

3. Distances & preprocessing (critical choices)

- **Metric choices:** squared-Euclidean ($\| x - y \|^2$), Euclidean, Manhattan, cosine distance $1 - \cos(\theta)$, correlation distance $1 - \rho$, Mahalanobis, Gower (mixed data).
- **Scaling:** mandatory for heterogeneous features — standardize or use domain weights. Ward's method assumes Euclidean geometry (uses variances).
- **Categorical / mixed data:** use Gower distance or appropriate dissimilarity matrix before linkage.

Practical rule: choose metric to reflect similarity semantics (cosine for text, Euclidean for geometry/continuous features, Gower for mixed types).

---

4. Linkage criteria — definition & properties (mathematically compact)

Let clusters A, B and a third cluster K. Denote pairwise distances $d(\cdot, \cdot)$. The Lance–Williams **recurrence** gives a unified update:

$$d(A \cup B,\ K) = \alpha_A d(A, K) + \alpha_B d(B, K) + \beta\, d(A, B) + \gamma\ \mid d(A, K) - d(B, K) \mid .$$

Common parameterizations:

Method	Update parameters	Short description / property
Single	$\alpha_A = \alpha_B = \frac{1}{2},\ \beta = 0,\ \gamma = -\frac{1}{2}$	$d(A \cup B, K) = \min(d(A, K), d(B, K))$. Produces chaining; equivalent to MST-based clustering.
Complete	$\alpha_A = \alpha_B = \frac{1}{2},\ \beta = 0,\ \gamma = +\frac{1}{2}$	$d = \max$. Tends to compact clusters.

Method	Update parameters	Short description / property
Average (UPGMA)	(\alpha_A=\tfrac{	A
Centroid (UPGMC)	(\alpha_A=\tfrac{	A
Ward	α and β depend on (	A

Ward merge cost (useful closed form):
merging A and B increases total within-cluster sum of squares by

$$\Delta_{A,B} = \frac{|A|\ |B|}{|A| + |B|} \left\| \mu_A - \mu_B \right\|^2,$$

where μ_A is the centroid of A. Agglomerative Ward merges the pair with smallest Δ.

---

5. Important mathematical properties (short proofs / statements)

- **Ultrametric from the dendrogram:** The cophenetic distance $c(i,j)$ is an ultrametric: for any triple i, j, k,

  $$c(i,j) \leqslant \max\{c(i,k), c(j,k)\}.$$

  Reason: in the tree, the earliest common ancestor of i and j sits at or below the later of ancestors with k.
- **Single linkage ≡ Minimum Spanning Tree (MST):** Building the MST (e.g., Kruskal) and cutting edges above threshold t yields exactly the connected components produced by single linkage at height t. *Sketch:* Kruskal's algorithm adds smallest edges that connect components — identical to single-link merges.
- **Monotonicity (no inversions):** A linkage is *monotone* if the sequence of merge distances is nondecreasing. Single, complete, average and Ward are monotone. Centroid and some median rules can be non-monotone (inversions) — their cophenetic distances may decrease after a merge, complicating dendrogram interpretation.

---

6. Algorithmic complexity & scalable strategies

- **Naïve implementation:** maintain full $n \times n$ distance matrix and search minima each step → worst-case $O(n^3)$ time and $O(n^2)$ memory.

- **Efficient updates (Lance–Williams):** allow $O(n^2)$ time with $O(n^2)$ memory if implemented carefully (nearest-neighbor search + update).
- **Specialized algorithms:** SLINK (Sibson) computes **single linkage** in $O(n^2)$ time and $O(n)$ memory. CLINK does the same for complete linkage. The *nearest-neighbor chain* algorithm gives $O(n^2)$ for many linkages.
- **Practical note:** hierarchical clustering is viable up to a few 10Ks of points with optimized C/Fortran libraries (SciPy, fast implementations). For very large data use hybrid strategies (e.g., BIRCH, sample + refine, or graph-based/DBSCAN or approximate clustering).

---

7. Interpreting and using the dendrogram

- **Cutting strategies:** pick level t (height) → flat partition; or cut to produce k clusters.
- **Statistical criteria:** silhouette score, cophenetic correlation (below), gap statistic, inconsistency coefficients, stability via bootstrap.
- **Cophenetic correlation coefficient (quality of hierarchy):**

$$\mathrm{CCC} = \frac{\sum_{i<j}(d_{ij} - \bar{d})(c_{ij} - \bar{c})}{\sqrt{\sum_{i<j}(d_{ij} - \bar{d})^2}\sqrt{\sum_{i<j}(c_{ij} - \bar{c})^2}},$$

  where d_{ij} are original dissimilarities and c_{ij} are cophenetic distances. Values near 1 indicate the dendrogram preserves the pairwise dissimilarities well.

---

8. Strengths, weaknesses & practitioner's heuristics

- **Strengths:**
  • No need to pre-specify k (full hierarchy).
  • Provides multi-scale view; interpretable dendrogram.
  • Works with any dissimilarity matrix (flexible for non-Euclidean data).
- **Weaknesses / pitfalls:**
  • Sensitive to metric & scaling.
  • Single linkage: *chaining* — joins via noisy points => poor compactness.
  • Centroid/median: can cause non-monotone merges (inversions).
  • Computationally heavy for large n.

- **Heuristics:**
  • For compact spherical clusters use **Ward** (Euclidean).
  • For elongated clusters or when connectivity matters use **single** (but beware chaining).
  • For compromise use **average** linkage.
  • Always standardize features if scales differ. Validate cluster choice with silhouette or stability/resampling.

---

9. Practical pipeline (short)

1. Clean data; impute missing values.
2. Choose/compute dissimilarity d (Gower for mixed). Standardize/weight features.
3. Select linkage (Ward/average/complete/single) according to geometry.
4. Compute condensed distance matrix (store efficiently). Use optimized library (SciPy/fast C implementation).
5. Inspect dendrogram; compute cophenetic correlation.
6. Choose cut (height or k) using silhouette / gap statistic / stability.
7. Validate clusters with domain checks.

---

10. Advanced notes (concise)

- **Constrained clustering:** must-link / cannot-link constraints can be incorporated in some agglomerative variants (modify allowable merges).
- **Bootstrap stability:** repeatedly resample and recompute hierarchy; measure cooccurrence of pairs to identify robust clusters.
- **Graph interpretations:** single linkage connected components ↔ thresholded graph; spectral clustering is an alternative that uses eigenstructure of similarity graph (useful when hierarchy is not desired).
- **Hybrid & large-scale:** use sampling to build dendrogram on representative points, then assign rest (fast but approximate).

---

11. Pocket mathematical references (formulas)

- **Ward merge cost:** $\Delta_{A,B} = \dfrac{|A|\ |B|}{|A| + |B|} \| \mu_A - \mu_B \|^2$.
- **Lance–Williams general update:** see §4.

- **Cophenetic correlation:** see §7 formula.
- **Ultrametric inequality:** $c(i,j) \leqslant \max(c(i,k), c(j,k))$.

---

12. Recommended canonical readings (textbooks & focused references)

(These are the classic/standard sources to study hierarchical clustering and clustering theory.)

- *The Elements of Statistical Learning* — Hastie, Tibshirani & Friedman — (chapter on clustering; Ward and linkage discussion).
- *Pattern Recognition and Machine Learning* — Christopher M. Bishop — (probabilistic view; clustering methods).
- *Algorithms for Clustering Data* — Jain & Dubes — (classical algorithms and properties).
- *Data Mining: Concepts and Techniques* — Han, Kamber & Pei — (practical clustering methods).
- *Modern Multivariate Statistical Techniques* / Murtagh & Contreras — (detailed hierarchical algorithms; SLINK/CLINK/Lance-Williams theory).
- Original algorithmic papers: Sibson (SLINK, 1973), Lance & Williams (update formula).

- **Use Ward + Euclidean** for compact, spherical clusters.
- **Use average/complete** if you want to avoid chaining but still flexible shapes.
- **Use single** only when connectivity/chain structure matters (and be careful of noise).
- **If mixed data:** compute Gower dissimilarities first.
- **For large n:** sample / use BIRCH / approximate algorithms.
- **Check cophenetic correlation** to judge dendrogram faithfulness.
