

Indian Institute of Technology Madras
ZANZIBAR CAMPUS

Z3003 · DATA MINING: SCALING TO LARGE DATA

Clustering II: Hierarchical Clustering and Evaluating Clusters

Tuesday 10 November 2026 · 50 minutes

Dr. Innocent Nyalala | BS Semester 5 (DSAI) | IIT Madras Zanzibar

By the End of Today You Can

IIT Madras Zanzibar

1. Build a hierarchical clustering by hand and draw its dendrogram
2. Cut a dendrogram to get any number of clusters
3. Compare single, complete, average and Ward linkage
4. Judge a clustering without labels: silhouette, stability, usefulness
5. Explain why it does not scale, and the remedies

Today's Plan (45 minutes + 5 minutes questions)

IIT Madras Zanzibar

minutes	part	what you get
0 – 5	Clusters inside clusters	the idea of a hierarchy
5 – 20	Agglomerative by hand	5 farms, every merge, the dendrogram
20 – 28	Linkage	four rules; our 3,000 farms; code
28 – 36	Is a clustering good?	silhouette by hand, stability, usefulness
36 – 45	Scale, exercise, summary	n^2 distances
45 – 50	Questions	

Clusters Inside Clusters

IIT Madras Zanzibar

Zanzibar

regions
→ districts
→ villages

Biology

family
→ genus
→ species

Our farms

2 big groups
→ 4 groups
→ individual farms

One tree gives every level. Decide how many groups AFTER looking (like OLAP roll-up, Week 4).

Agglomerative Clustering: Merge the Closest, Repeat

IIT Madras Zanzibar

1	**Start**	every farm is its own cluster
2	**Find**	the two closest clusters
3	**Merge**	join them: one cluster fewer
4	**Recompute**	distances from the new cluster (linkage rule)
5	**Repeat**	until one cluster; record each merge height

Definition: dendrogram = a tree drawing of every merge; the height of each join = the distance at which the groups merged

Worked Example: 5 Farms

IIT Madras Zanzibar

farm	(x_1, x_2)
A	(1, 1)
B	(2, 1)
C	(4, 3)
D	(5, 4)
E	(5, 6)

	A	B	C	D	E
A	—	1.00	3.61	5.00	6.40
B	1.00	—	2.83	4.24	5.83
C	3.61	2.83	—	1.41	3.16
D	5.00	4.24	1.41	—	2.00
E	6.40	5.83	3.16	2.00	—

Smallest distance in the table: A – B = 1.00, so A and B merge first.

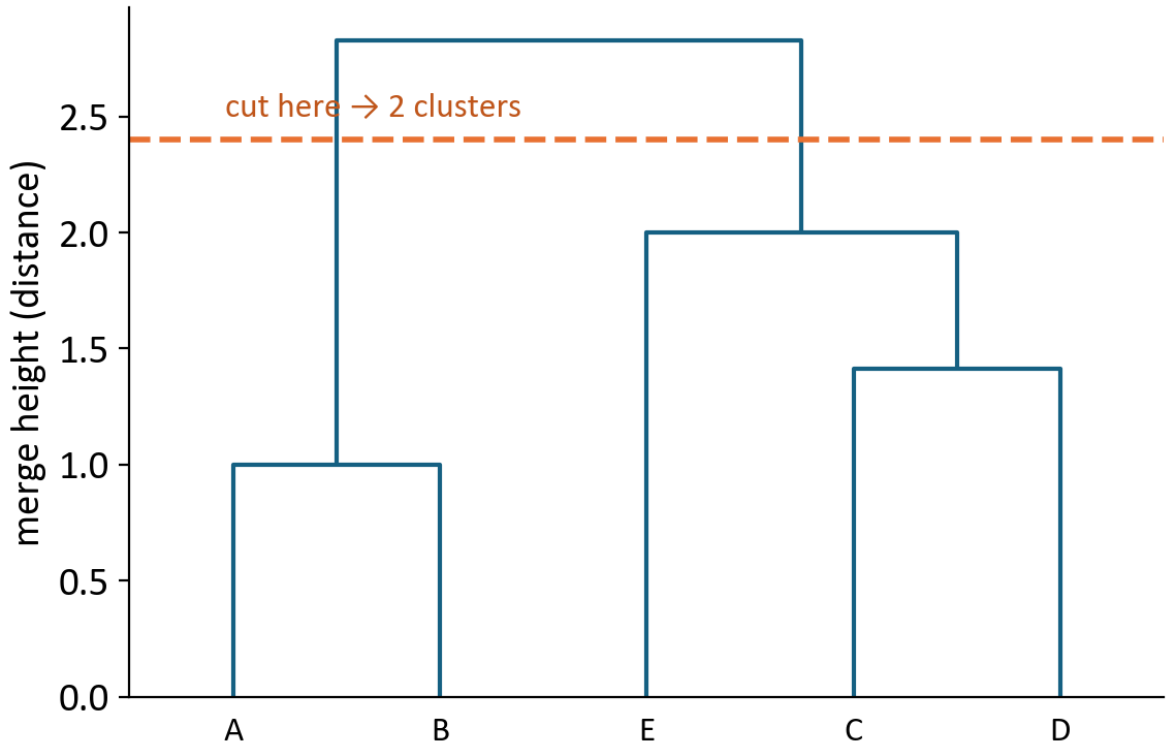
Every Merge, Step by Step (Single Linkage)

IIT Madras Zanzibar

merge	joins	why	height
1	A + B	closest pair	1.00
2	C + D	next closest	1.41
3	{C, D} + E	single link: min(C–E 3.16, D–E 2.00)	2.00
4	{A, B} + {C, D, E}	closest pair across: B – C	2.83

Heights 1.00, 1.41, 2.00, 2.83 are the join heights of the dendrogram.

Reading and Cutting the Dendrogram



Bottom	the individual farms
Join height	distance at which groups merged
Cut	horizontal line: each branch crossed = a cluster
Where	across a long vertical gap

One tree, every number of clusters: choose k AFTER looking, unlike k-means.

Linkage: Four Ways to Measure Distance Between Groups

IIT Madras Zanzibar

Single

closest pair of members
→ long chains

villages close if ANY houses close

Complete

farthest pair of members
→ compact; outlier-sensitive

close only if EVERY house is close

Average

average over all pairs
→ a compromise

Ward

smallest increase in spread
→ balanced, round

like k-means

Linkage Choices on 3,000 of Our Farms

IIT Madras Zanzibar

linkage	4 cluster sizes	silhouette	verdict
single	1,585 · 1,004 · 261 · 150	0.54	chained; uneven
complete	1,682 · 731 · 454 · 133	0.18	uneven
average	1,585 · 1,004 · 261 · 150	0.54	same as single here
Ward	**1,084 · 1,004 · 501 · 411**	**0.45**	**balanced: usable**

Single has the best silhouette (0.54) but the least useful groups. Never judge by one number.

Hierarchical Clustering in Python

IIT Madras Zanzibar

```
1  from scipy.cluster.hierarchy import linkage, dendrogram, fcluster
2  from sklearn.cluster import AgglomerativeClustering
3  L = linkage(Z3000, method="ward")                 # every merge, 3,000 farms
4  dendrogram(L, truncate_mode="lastp", p=20)       # draw the top of the tree
5  labels = fcluster(L, t=4, criterion="maxclust")  # cut into 4 clusters
6  labels2 = AgglomerativeClustering(n_clusters=4, linkage="ward").fit_predict(Z3000)
7  print(silhouette_score(Z3000, labels2))         # ≈ 0.45
```

line	what it does
3	all merges (Ward)
4	draw dendrogram
5	cut: 4 clusters
6	sklearn, one line

Change method= to 'single', 'complete' or 'average' to compare linkages.

How Do We Know a Clustering Is Good?

IIT Madras Zanzibar

Internal

silhouette, from data alone
ours (k = 3): 0.52

Stability

rerun on new samples or starts:
do the same groups return?

External

agreement with known labels:
ARI vs farm type = 0.13

Usefulness

can someone describe and act?
'small, no inputs' → subsidies

The Silhouette, Computed for Two Farms

IIT Madras Zanzibar

	working	value
A: a	own cluster {B}: 1.00	1.00
A: b	(3.61 + 5.00 + 6.40) ÷ 3	5.00
A: s	(5.00 − 1.00) ÷ 5.00	0.80 deep inside
C: a	(1.41 + 3.16) ÷ 2	2.29
C: b	(3.61 + 2.83) ÷ 2	3.22
C: s	(3.22 − 2.29) ÷ 3.22	0.29 near border

Why Hierarchical Clustering Does Not Scale

IIT Madras Zanzibar

pairs of farms = n × (n – 1) ÷ 2 (8 bytes each)

farms n	pairwise distances	memory
10,000	50 million	≈ 0.4 GB
100,000	5 billion	≈ 40 GB
1,000,000	500 billion	≈ 4 TB

10 × more farms → 100 × more memory. Hierarchical is for thousands, not millions.

What to Do Instead on Big Data

IIT Madras Zanzibar

1	**Sample**	hierarchical on 3,000 farms → read the dendrogram → choose k
2	**Then k-means**	run k-means with that k on ALL the farms
3	**Or summarise**	BIRCH: compress first, cluster the summaries (Wednesday)

Hierarchical Clustering in the Real World

IIT Madras Zanzibar

Biology

family trees of species from DNA

Genomics

gene heat maps with dendrograms

Libraries

topics inside topics

Marketing

segments inside segments

Agriculture

farm types at several levels

Social networks

communities within communities

When groups are nested, a tree shows every level at once.

Your Turn (5 minutes, pairs)

IIT Madras Zanzibar

Four farms on one attribute: P = 1, Q = 2, R = 6, S = 9

1. Write all six pairwise distances
2. Run single-linkage clustering: each merge and its height
3. Draw the dendrogram; cut it into 2 clusters
4. Compute the silhouette of R in the 2-cluster result

Solution, Step by Step

IIT Madras Zanzibar

step	working	answer
distances	PQ 1 · PR 5 · PS 8 · QR 4 · QS 7 · RS 3	
merge 1	smallest: P – Q	height 1
merge 2	next: R – S	height 3
merge 3	{P,Q} + {R,S}: closest pair Q – R	height 4
2 clusters	cut between 3 and 4	{P, Q}, {R, S}
silhouette R	a = 3; b = (5 + 4) ÷ 2 = 4.5; (4.5 − 3) ÷ 4.5	0.33

Summary

IIT Madras Zanzibar

1	**Hierarchy**	merge closest repeatedly; dendrogram; cut to choose k
2	**Linkage**	single chains, complete compacts, average, Ward balances
3	**Evaluate**	silhouette, stability, external (ARI), usefulness
4	**Scale**	n^2 distances: 1 million farms ≈ 4 TB → sample or summarise

Next: clustering at scale: mini-batch k-means, BIRCH and DBSCAN.

Questions (5 minutes)

IIT Madras Zanzibar

Q: Is hierarchical better than k-means?

A: Different: every level, no k in advance; but it does not scale.

Q: Can a merge be undone?

A: No: it is greedy, like a decision tree.

Q: Do I need to z-score?

A: Yes, it is distance-based, exactly like k-means.

Q: What is divisive clustering?

A: Top-down: split one cluster repeatedly; rarer and costlier.