Yale University

Clustering

S&DS 265 · Introductory Machine Learning · Lecture 15

Zhuoran Yang

Fall 2026

AI usage: figures were generated, and these slides polished, with the help of AI tools.

Outline

Learning Objectives

In this lecture you will learn:

1. Why representation and distance determine what “similar” means before clustering begins;
2. How the K-means objective produces alternating assignment and centroid updates;
3. Why those updates descend yet can stop at different local optima;
4. How elbow, silhouette, stability, and hierarchical views support different choices;
5. When K-means is appropriate, why it misses curved groups, and how a different distance can recover them.

Outline

Motivating Example: Grouping Colleges

The same 777 colleges can be compared by size, selectivity, price, spending, or student outcomes. No column in the table says which colleges belong together.

Before any algorithm runs: which representation and distance make two colleges similar, and for what stated purpose?

Example and data: ISLP Chapter 12 (James et al., 2023).

The College Data

college	Apps	Enroll	Outstate	Expend	Grad.Rate
Abilene Christian	1660	721	7440	7041	60
Adelphi	2186	512	12280	10527	56
Adrian	1428	336	11250	8735	54
Agnes Scott	417	137	12960	19016	59

The same 777-college table as the previous lecture, now asked for groups rather than directions. There is still no observed "correct cluster" column.

Source: First four rows of College.csv from the ISLP resources.

Reading One Row

college	Apps	Enroll	Outstate	Expend	Grad.Rate
Abilene Christian	1660	721	7440	7041	60

One row mixes counts, dollars, and a percentage. Lecture 14 asked which directions such rows vary along. Today's question is easier to state and harder to answer: which other college is most like this one?

The Answer Changes with the Units

Measure the distance from Abilene Christian to each of the other 776 colleges and sort, once in the recorded units and once standardized.

						rank, 1 = nearest	
college	Apps	F.Undergrad	Outstate	Expend	Grad.Rate	raw	standardized
Abilene Christian	1660	2885	7440	7041	60		
Fayetteville State	1455	2632	6806	6972	24	1	287
Freed-Hardeman	895	1174	5500	6078	62	89	1

The Answer Changes with the Units

Measure the distance from Abilene Christian to each of the other 776 colleges and sort, once in the recorded units and once standardized.

						rank, 1 = nearest	
college	Apps	F.Undergrad	Outstate	Expend	Grad.Rate	raw	standardized
Abilene Christian	1660	2885	7440	7041	60		
Fayetteville State	1455	2632	6806	6972	24	1	287
Freed-Hardeman	895	1174	5500	6078	62	89	1

Nothing changed but the units.

The Answer Changes with the Units

Measure the distance from Abilene Christian to each of the other 776 colleges and sort, once in the recorded units and once standardized.

						rank, 1 = nearest	
college	Apps	F.Undergrad	Outstate	Expend	Grad.Rate	raw	standardized
Abilene Christian	1660	2885	7440	7041	60		
Fayetteville State	1455	2632	6806	6972	24	1	287
Freed-Hardeman	895	1174	5500	6078	62	89	1

Nothing changed but the units.

The two winners graduate 24% and 62% of their students.

Why the Raw Distance Ignored Graduation Rate

derive

- Squared distance adds one term per column, $\|x - z\|_2^2 = \sum_j (x_j - z_j)^2$, so each column contributes in proportion to its own squared spread.

Why the Raw Distance Ignored Graduation Rate

derive

- Squared distance adds one term per column, $\|x - z\|_2^2 = \sum_j (x_j - z_j)^2$, so each column contributes in proportion to its own squared spread.
- Dollars and head counts win by construction: Expend (30%), F.Undergrad (26%) and Outstate (18%) supply 74% of all squared distance across the 777 colleges.

Why the Raw Distance Ignored Graduation Rate

derive

- Squared distance adds one term per column, $\|x - z\|_2^2 = \sum_j (x_j - z_j)^2$, so each column contributes in proportion to its own squared spread.
- Dollars and head counts win by construction: Expend (30%), F.Undergrad (26%) and Outstate (18%) supply 74% of all squared distance across the 777 colleges.
- The five percentage columns together supply 0.001%, so the raw ranking never really consulted graduation rate.

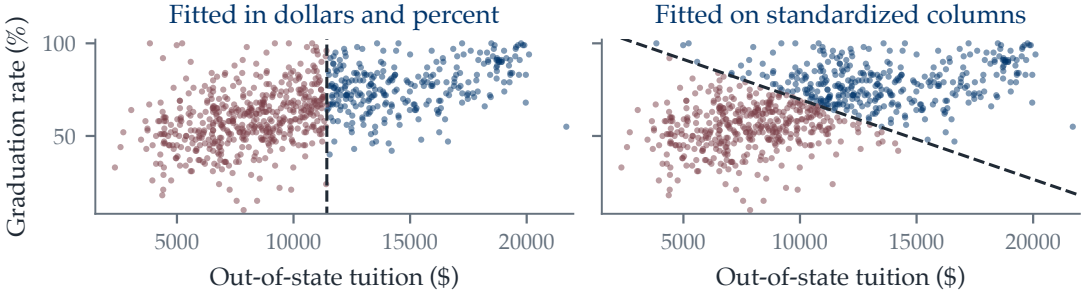
Why the Raw Distance Ignored Graduation Rate

derive

- Squared distance adds one term per column, $\|x - z\|_2^2 = \sum_j (x_j - z_j)^2$, so each column contributes in proportion to its own squared spread.
- Dollars and head counts win by construction: Expend (30%), F.Undergrad (26%) and Outstate (18%) supply 74% of all squared distance across the 777 colleges.
- The five percentage columns together supply 0.001%, so the raw ranking never really consulted graduation rate.

Standardizing is not tidying: it is the decision to let one percentage point matter as much as one dollar.

Units Decide Which Colleges Group Together



Choose two center points and put each college with the closer one; the colors show that split, made with no label. In dollars and percent the boundary is vertical, so graduation rate never enters it. Standardize first and it tilts, moving 109 colleges across.

Course nearest-center fit with $K = 2$ on Outstate and Grad.Rate; both panels are drawn in the original units.

Purposes for Clustering

Summarize	Compress	Organize
discover recurring profiles and exemplars	replace observations or colors with a codebook	form neighborhoods for retrieval or later modeling

The same observations can support different coherent partitions because purpose changes which differences count; there is no universal ground-truth partition to recover.

Distance Declares Which Differences Matter

Euclidean

$$d(x, z) = \|x - z\|_2$$

absolute coordinate differences

Cosine

$$d(x, z) = 1 - \frac{x^\top z}{\|x\|\|z\|}$$

direction, not magnitude

Mahalanobis

$$d_M^2 = (x - z)^\top M (x - z)$$

weighted, correlated directions

Cosine calls a small college and a large one with the same spending pattern similar; Euclidean does not. Choosing a distance is choosing what to ignore, so audit a few nearest neighbors before optimizing a global partition.

From Neighbors to Groups

- Ranking answered a question about one college at a time: who is nearest to this one?

From Neighbors to Groups

- Ranking answered a question about one college at a time: who is nearest to this one?
- A grouping is a stronger object. It must place every college somewhere, and colleges sharing a group are then treated as interchangeable.

From Neighbors to Groups

- Ranking answered a question about one college at a time: who is nearest to this one?
- A grouping is a stronger object. It must place every college somewhere, and colleges sharing a group are then treated as interchangeable.
- So the question becomes: which split of all 777 colleges into K groups is the best one?

From Neighbors to Groups

- Ranking answered a question about one college at a time: who is nearest to this one?
- A grouping is a stronger object. It must place every college somewhere, and colleges sharing a group are then treated as interchangeable.
- So the question becomes: which split of all 777 colleges into K groups is the best one?

That word, best, has no meaning yet. Supplying one is the work ahead, and every later choice follows from it.

Outline

Problem Setup of K-Means

- A grouping is an assignment: $c_i \in \{1, \dots, K\}$ says which group observation i belongs to.

Problem Setup of K-Means

- A grouping is an assignment: $c_i \in \{1, \dots, K\}$ says which group observation i belongs to.
- Tightness needs something to be close to: give each group a representative μ_k.

Problem Setup of K-Means

- A grouping is an assignment: $c_i \in \{1, \ldots, K\}$ says which group observation i belongs to.
- Tightness needs something to be close to: give each group a representative μ_k.
- Charge each observation the squared distance to its own representative.

Problem Setup of K-Means

- A grouping is an assignment: $c_i \in \{1, \dots, K\}$ says which group observation i belongs to.
- Tightness needs something to be close to: give each group a representative μ_k.
- Charge each observation the squared distance to its own representative.

$$W(c, \mu) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \mu_{c_i}\|_2^2$$


every segment, squared and added

Squared Euclidean distance is a choice here, not a requirement: any distance can stand in its place.

What That Criterion Assumes

Three commitments were made in writing it down.

- **Squared Euclidean distance**, so the cost of being wrong grows quadratically and every column counts as it was scaled.

What That Criterion Assumes

Three commitments were made in writing it down.

- **Squared Euclidean distance**, so the cost of being wrong grows quadratically and every column counts as it was scaled.
- **One representative per group**, so a group is described by a single point and is therefore compact by construction.

What That Criterion Assumes

Three commitments were made in writing it down.

- **Squared Euclidean distance**, so the cost of being wrong grows quadratically and every column counts as it was scaled.
- **One representative per group**, so a group is described by a single point and is therefore compact by construction.
- **A hard assignment**, so a college on the boundary is recorded as fully in one group.

What That Criterion Assumes

Three commitments were made in writing it down.

- **Squared Euclidean distance**, so the cost of being wrong grows quadratically and every column counts as it was scaled.
- **One representative per group**, so a group is described by a single point and is therefore compact by construction.
- **A hard assignment**, so a college on the boundary is recorded as fully in one group.

None of these is discovered from the data. They are the definition of "best" that the next slides then optimize.

Fix One, Minimize the Other

- $W(c, \mu)$ has two unknowns of different kinds: a discrete assignment c and continuous centers μ. Minimizing over both at once is NP-hard.

Fix One, Minimize the Other

- $W(c, \mu)$ has two unknowns of different kinds: a discrete assignment c and continuous centers μ. Minimizing over both at once is NP-hard.
- So fix one and minimize the other, in turn:

$$c^{t+1} \in \arg\min_{c} \; W(c, \mu^t), \qquad \mu^{t+1} \in \arg\min_{\mu} \; W(c^{t+1}, \mu).$$

Fix One, Minimize the Other

- $W(c, \mu)$ has two unknowns of different kinds: a discrete assignment c and continuous centers μ. Minimizing over both at once is NP-hard.
- So fix one and minimize the other, in turn:

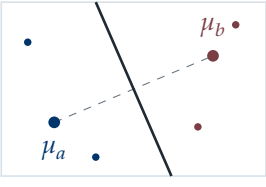
$$c^{t+1} \in \arg\min_{c} \; W(c, \mu^t), \qquad \mu^{t+1} \in \arg\min_{\mu} \; W(c^{t+1}, \mu).$$

- Each of those has a closed-form answer, worked out on the next two slides. Repeat until nothing moves.

Nothing here is specific to clustering; the same move returns at the end of the section.

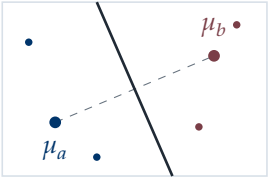
Assignment Step: Enumerate

With the centers $\mu_1, \ldots, \mu_K$ fixed, the sum splits over observations, so each one is settled on its own by enumerating the K candidates.



Assignment Step: Enumerate

With the centers $\mu_1, \dots, \mu_K$ fixed, the sum splits over observations, so each one is settled on its own by enumerating the K candidates.

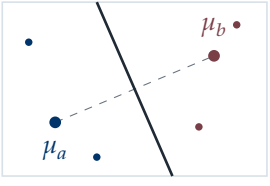


Each observation takes the nearest center, $c_i \leftarrow \arg\min_k \|x_i - \mu_k\|_2^2$, and comparing two centers cancels the shared $\|x\|_2^2$:

$$\|x - \mu_a\|_2^2 \le \|x - \mu_b\|_2^2 \iff 2x^\top(\mu_b - \mu_a) \le \|\mu_b\|_2^2 - \|\mu_a\|_2^2.$$

Assignment Step: Enumerate

With the centers $\mu_1, \dots, \mu_K$ fixed, the sum splits over observations, so each one is settled on its own by enumerating the K candidates.



Each observation takes the nearest center, $c_i \leftarrow \arg\min_k \|x_i - \mu_k\|_2^2$, and comparing two centers cancels the shared $\|x\|_2^2$:

$$\|x - \mu_a\|_2^2 \le \|x - \mu_b\|_2^2 \iff 2x^\top(\mu_b - \mu_a) \le \|\mu_b\|_2^2 - \|\mu_a\|_2^2.$$

The test is linear: the boundary is the bisector above.

The Best Representative Is the Mean

- Hold the assignment $c_1, \ldots, c_n$ fixed and write $C_k = \{i : c_i = k\}$ for the observations that landed in group k.

The Best Representative Is the Mean

- Hold the assignment $c_1, \ldots, c_n$ fixed and write $C_k = \{i : c_i = k\}$ for the observations that landed in group k.
- Each representative is then chosen on its own, by minimizing $\sum_{i \in C_k} \|x_i - \mu\|_2^2$ over μ.

The Best Representative Is the Mean

- Hold the assignment $c_1, \ldots, c_n$ fixed and write $C_k = \{i : c_i = k\}$ for the observations that landed in group k.
- Each representative is then chosen on its own, by minimizing $\sum_{i \in C_k} \|x_i - \mu\|_2^2$ over μ.
- Setting the gradient to zero gives the cluster mean, which is where the method gets its name:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

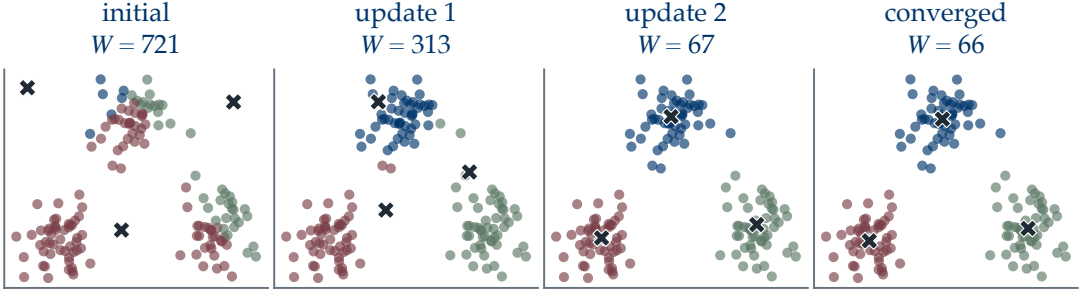
The Best Representative Is the Mean

- Hold the assignment $c_1, \dots, c_n$ fixed and write $C_k = \{i : c_i = k\}$ for the observations that landed in group k.
- Each representative is then chosen on its own, by minimizing $\sum_{i \in C_k} \|x_i - \mu\|_2^2$ over μ.
- Setting the gradient to zero gives the cluster mean, which is where the method gets its name:

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

The mean is special to squared Euclidean loss: absolute distance would give the median instead.

Alternating Assignment and Update



Black crosses are the centers μ_k; color is the current assignment c_i. Each panel recolors every point to its nearest cross, then moves each cross to the mean of its own color. The cost falls $721 \to 313 \to 67 \to 66$ and then stops.

Original course simulation, seed 26518; the same three groups throughout.

Monotone Descent

Write W for the criterion. The assignment step gives

$$W(c^{t+1}, \mu^t) \leq W(c^t, \mu^t),$$

and the centroid update gives

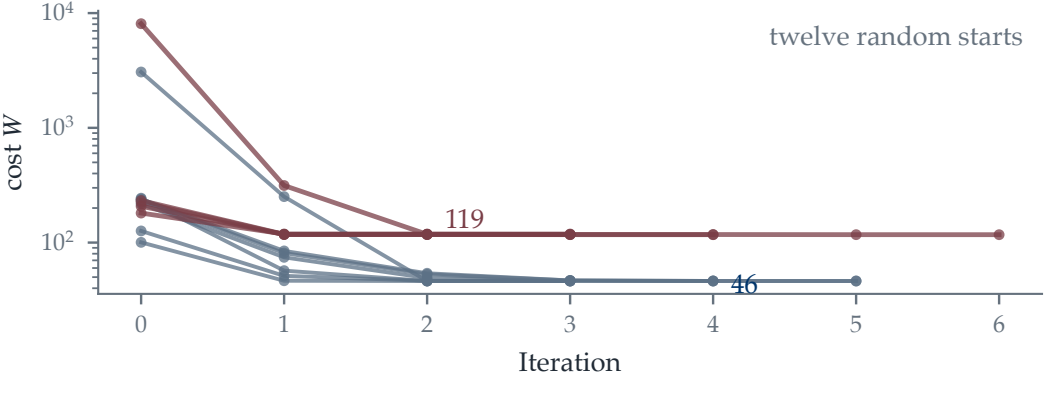
$$W(c^{t+1}, \mu^{t+1}) \leq W(c^{t+1}, \mu^t).$$

Hence

$$W(c^{t+1}, \mu^{t+1}) \leq W(c^t, \mu^t).$$

The cost never rises, and only finitely many assignments exist, so the procedure stops after finitely many steps. It stops at a partition that neither update can improve on its own, which need not be the one with the lowest cost.

Descending, but Not to the Same Place



Twelve random starts on one data set. Every curve falls and none rises, exactly as the algebra promised, and all settle within a few iterations. Seven stop at $W = 46$ and five at $W = 119$: same criterion, same data, two different answers.

Original course simulation, seed 26518; cost on a logarithmic scale.

One-Dimensional Worked Example

For points $0, 2, 8, 10$ and initial centers $1, 7$:

$$C_1 = \{0, 2\}, \qquad C_2 = \{8, 10\}.$$

Updating gives

$$\mu_1 = (0 + 2)/2 = 1, \qquad \mu_2 = (8 + 10)/2 = 9.$$

The converged objective is

$$(0 - 1)^2 + (2 - 1)^2 + (8 - 9)^2 + (10 - 9)^2 = 4.$$

The second center moves even though the first assignment was already stable.

Computational Cost

- The assignment step compares every observation with every center: nK distances in d dimensions, or $O(nKd)$. Over T iterations, $O(nKdT)$.

Computational Cost

- The assignment step compares every observation with every center: nK distances in d dimensions, or $O(nKd)$. Over T iterations, $O(nKdT)$.
- The answer depends on where the centers started, so the whole run is repeated from R different **random starts** and the partition with the lowest W is kept, at $O(RnKdT)$.

Computational Cost

- The assignment step compares every observation with every center: nK distances in d dimensions, or $O(nKd)$. Over T iterations, $O(nKdT)$.
- The answer depends on where the centers started, so the whole run is repeated from R different **random starts** and the partition with the lowest W is kept, at $O(RnKdT)$.
- At scale, **mini-batch** updates move the centers using a sample of observations rather than all of them, trading exact descent for speed.

Those R starts are not just an implementation detail. How much they disagree is what we look at next.

Aside: Alternating Minimization

- Strip away the clustering and a general strategy is left. To minimize $f(u, v)$ when the joint problem is hard but each block is easy, alternate:

$$u \leftarrow \arg\min_{u} f(u, v), \qquad v \leftarrow \arg\min_{v} f(u, v).$$

Aside: Alternating Minimization

- Strip away the clustering and a general strategy is left. To minimize $f(u, v)$ when the joint problem is hard but each block is easy, alternate:

$$u \leftarrow \arg\min_{u} f(u, v), \qquad v \leftarrow \arg\min_{v} f(u, v).$$

- With more than two blocks, cycling through them one at a time is **cyclic coordinate descent**.

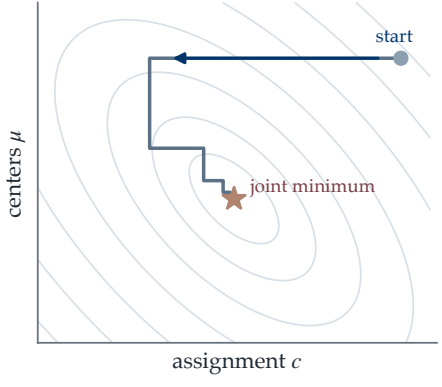
Aside: Alternating Minimization

- Strip away the clustering and a general strategy is left. To minimize $f(u, v)$ when the joint problem is hard but each block is easy, alternate:

$$u \leftarrow \arg\min_u f(u, v), \qquad v \leftarrow \arg\min_v f(u, v).$$

- With more than two blocks, cycling through them one at a time is **cyclic coordinate descent**.
- It reappears throughout the course: in EM next lecture, and wherever a model has parameters of two different kinds.

One Block at a Time

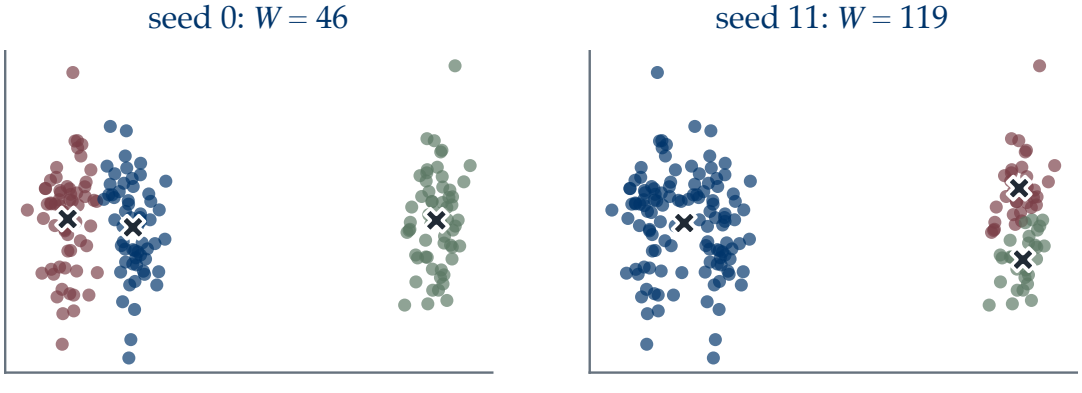


Each move minimizes exactly, but only along one block, so the path turns a corner every step. It cannot go uphill, and it stops when neither block wants to move. That stopping point need not be the joint minimum.

Schematic; the picture is smooth so both moves can be drawn, while the assignment block is really discrete.

Outline

What the Two Answers Look Like



The best and worst of 40 random starts. At $W = 46$ the two nearby groups are separated and the far group keeps one center; at $W = 119$ the nearby pair has been merged and the far group split in two. Only the starting centers differed.

Original course simulation, seed 26518; crosses are the fitted centers.

K-Means++ Initialization

Every center is one of the observations; only the sampling changes.

(i) Pick the first center uniformly at random among them.

(ii) For each observation x, let $D(x)$ be its distance to the nearest center chosen so far.

(iii) Pick the next center, again an observation, with probability

$$\Pr(x) = \frac{D(x)^2}{\sum_j D(x_j)^2},$$

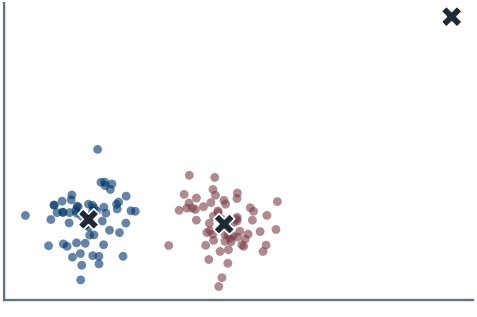
so a point far from every chosen center is the likely pick.

(iv) Repeat until K centers are chosen, then alternate as before.

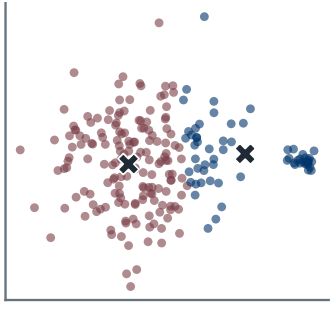
Source: Arthur and Vassilvitskii (2007).

Two Ways the Criterion Misreads the Data

one stray point, $K = 3$



one big group, one small, $K = 2$



Left: a single stray point takes a whole group to itself, so at $K = 3$ one of the three clusters has one member. Right: the small tight group is swallowed by its neighbors while the big diffuse group is cut in two, giving sizes 63 and 162.

Original course simulation, seed 26518; best of 30 starts in each panel.

Why, and What Helps

- Both follow from the cost being a sum of squared distances: a far point counts enormously, and every group is charged alike whatever its size.

Why, and What Helps

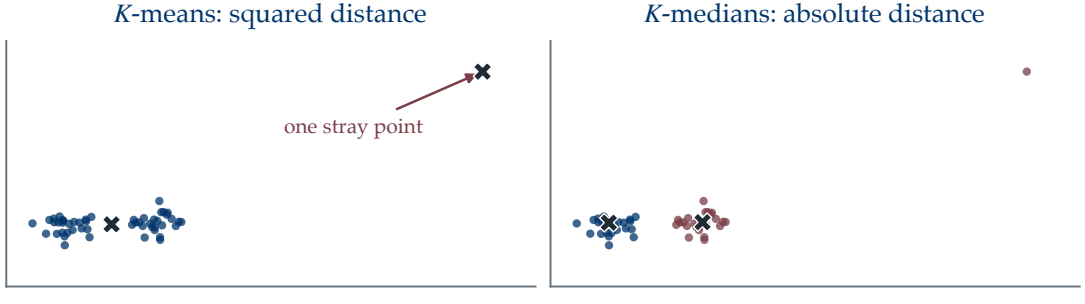
- Both follow from the cost being a sum of squared distances: a far point counts enormously, and every group is charged alike whatever its size.
- Against stray points, use a **robust distance**: absolute distance gives K-medians, and requiring the representative to be an observation gives K-medoids.

Why, and What Helps

- Both follow from the cost being a sum of squared distances: a far point counts enormously, and every group is charged alike whatever its size.
- Against stray points, use a **robust distance**: absolute distance gives K-medians, and requiring the representative to be an observation gives K-medoids.
- Against imbalance, allow the groups to differ. A mixture model gives each group its own spread, which is where the next lecture begins.

Check cluster sizes before reading anything into the profiles, and report what the implementation does with an empty cluster.

Squared Distance against Absolute Distance



The same 50 observations and one stray point, both fits with $K = 2$. Squared distance makes the stray so expensive that K-means spends a whole center on it and merges the two real groups. Absolute distance costs far less, so K-medians keeps the groups and leaves the stray where it is.

Original course simulation, seed 26518; best of 40 starts for each method.

Outline

The Fitted Cost Cannot Select K

More centers can never do worse:

$$W(K+1) \leq W(K),$$

because any K-center solution is still available at $K+1$, with one center left unused. So the fitted cost falls no matter what, and choosing the K with the smallest W would always answer $K = n$.

The Fitted Cost Cannot Select K

More centers can never do worse:

$$W(K+1) \leq W(K),$$

because any K-center solution is still available at $K+1$, with one center left unused. So the fitted cost falls no matter what, and choosing the K with the smallest W would always answer $K = n$.

Two criteria are used instead. The elbow asks where the cost stops falling quickly; the silhouette asks whether the groups are separated at all.

The Elbow

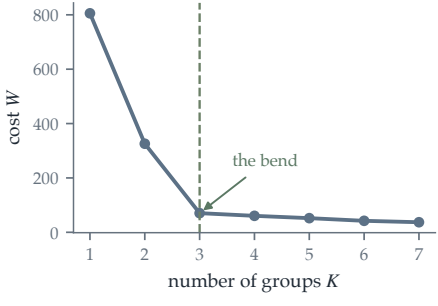
- Plot $W(K)$ and look at how fast it falls, not how low the loss goes.

The Elbow

- Plot $W(K)$ and look at how fast it falls, not how low the loss goes.
- While K is below the number of real groups, a new center splits a genuine group and the cost drops sharply.

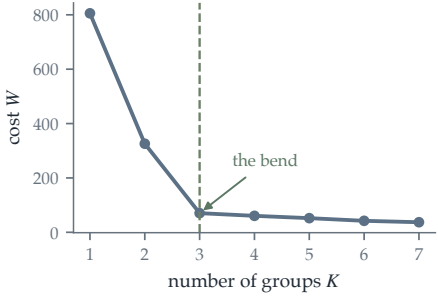
The Elbow

- Plot $W(K)$ and look at how fast it falls, not how low the loss goes.
- While K is below the number of real groups, a new center splits a genuine group and the cost drops sharply.
- Past that point a new center only subdivides a group that was already tight, and the curve flattens.



The Elbow

- Plot $W(K)$ and look at how fast it falls, not how low the loss goes.
- While K is below the number of real groups, a new center splits a genuine group and the cost drops sharply.
- Past that point a new center only subdivides a group that was already tight, and the curve flattens.



Here $806 \to 326 \to 70$, then $61, 52, 42$: the bend at $K = 3$ is clear. On real data it usually is not.

The Silhouette Score

- For point i, let $a(i)$ be its mean distance to its own group and $b(i)$ the smallest mean distance to any other group.

The Silhouette Score

- For point i, let $a(i)$ be its mean distance to its own group and $b(i)$ the smallest mean distance to any other group.
- The **silhouette** contrasts the two on a fixed scale:

$$\boxed{s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]}$$

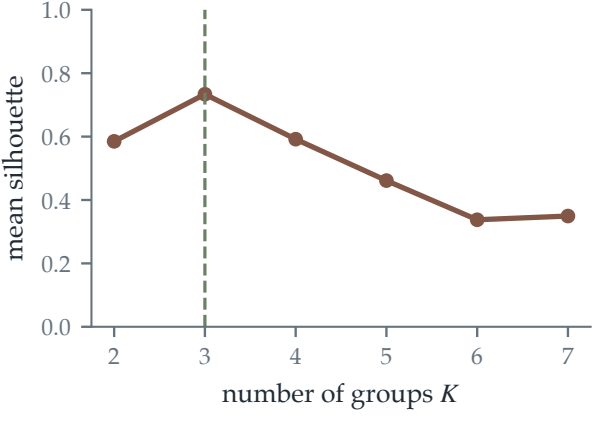
The Silhouette Score

- For point i, let $a(i)$ be its mean distance to its own group and $b(i)$ the smallest mean distance to any other group.
- The **silhouette** contrasts the two on a fixed scale:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1]$$

- Near $+1$ the point sits comfortably inside its group, near 0 it is on a boundary, below 0 it is closer to a different one.

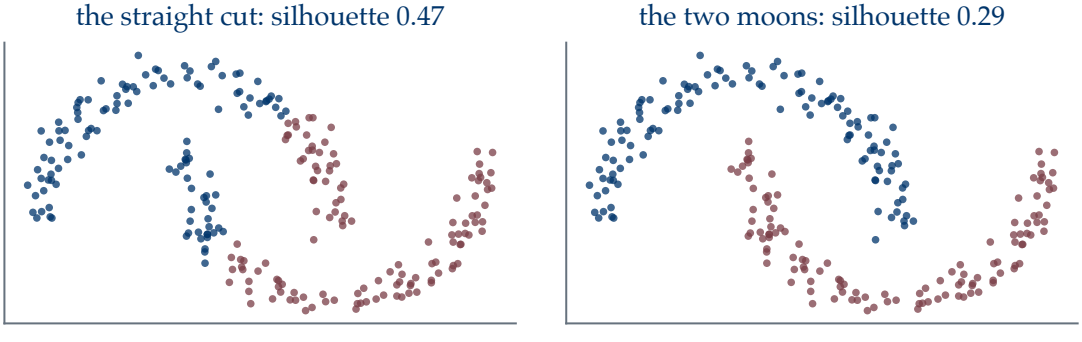
Silhouette across K



Averaging $s(i)$ over all observations gives one number per K, and unlike the cost it does not have to improve with K: here it rises to 0.73 at $K = 3$ and falls away after.

Each fit is the best of eight starts.

What the Silhouette Rewards



The same curved data under two groupings. The score prefers the straight cut at 0.47 over the two moons at 0.29, because $a(i)$ and $b(i)$ are straight-line distances: a moon is long, so its own members are far away, while the other moon passes close by.

Original course simulation, seed 26518; mean silhouette over all 240 observations.

So What Is It Evidence For?

The agreement with K-means is structural, not independent support. Both are built from straight-line distances: W rewards points close to their own center, and $s(i)$ rewards points close to their own group and far from the next one.

So What Is It Evidence For?

The agreement with K-means is structural, not independent support. Both are built from straight-line distances: W rewards points close to their own center, and $s(i)$ rewards points close to their own group and far from the next one.

So the score is highest on exactly the shapes a nearest-center rule produces: evidence that the groups are compact and separated, not evidence about how many groups exist.

So What Is It Evidence For?

The agreement with K-means is structural, not independent support. Both are built from straight-line distances: W rewards points close to their own center, and $s(i)$ rewards points close to their own group and far from the next one.

So the score is highest on exactly the shapes a nearest-center rule produces: evidence that the groups are compact and separated, not evidence about how many groups exist.

Two curved groups that interleave will score badly at every K, however real they are — a case we meet at the end of the lecture.

Outline

Decide K in Hindsight

K-means has to be told K before it starts, and the elbow and the silhouette are attempts to guess it from outside.

Decide K in Hindsight

K-means has to be told K before it starts, and the elbow and the silhouette are attempts to guess it from outside.

Hierarchical clustering turns the order around. Build the whole tree of groupings, from every observation alone at the bottom to one single group at the top, and only then cut it, at whatever height can be defended.

Decide K in Hindsight

K-means has to be told K before it starts, and the elbow and the silhouette are attempts to guess it from outside.

Hierarchical clustering turns the order around. Build the whole tree of groupings, from every observation alone at the bottom to one single group at the top, and only then cut it, at whatever height can be defended.

Two questions follow, and the rest of the section is those two: how is the tree built, and where should it be cut?

Building the Tree

(i) Start with n groups, each holding one observation.

(ii) Merge the two closest groups, and record the distance at which they merged. That number is the **merge height**, and the tree is drawn against it.

(iii) Repeat until one group is left, after $n - 1$ merges.

Building the Tree

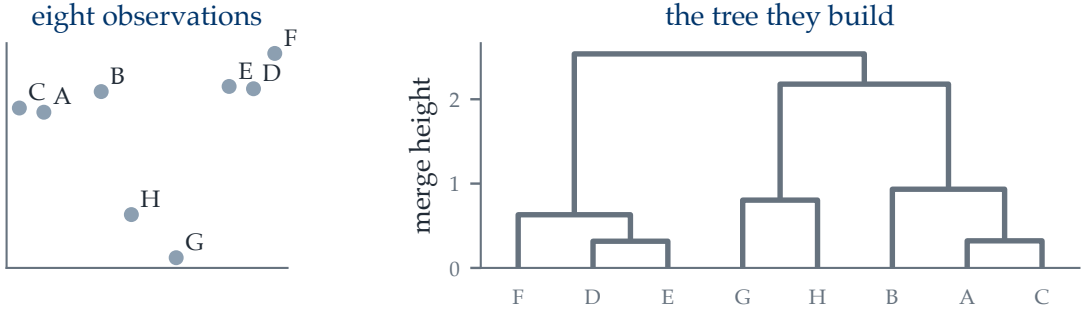
(i) Start with n groups, each holding one observation.

(ii) Merge the two closest groups, and record the distance at which they merged. That number is the **merge height**, and the tree is drawn against it.

(iii) Repeat until one group is left, after $n - 1$ merges.

Step (ii) needs a distance between groups, and the distance between points does not supply one. That choice is called the linkage.

How to Read a Dendrogram



Read it by height: a merge is drawn at the distance between the two groups it joins, so A and C, the closest pair, merge lowest. The left-to-right order is not information, because the two branches under any merge can be swapped without changing the tree: E and G stand side by side yet are only joined at the very top.

Average linkage on eight points.

Linkage: the Distance between Two Groups

Merging needs a distance between groups. Three rules build one from point distances; a fourth differs.

rule	$d(A, B)$	what it does
single	$\min_{a \in A,\, b \in B} d(a, b)$	follows chains: one bridge joins all
complete	$\max_{a \in A,\, b \in B} d(a, b)$	the worst pair rules, so groups stay compact
average	$\frac{1}{\lvert A \rvert \lvert B \rvert} \sum_{a \in A} \sum_{b \in B} d(a, b)$	a compromise between the two
Ward	the rise in the K-means cost	next slide

Ward Linkage: Merge Where the Cost Rises Least

Write $W_S = \sum_{i \in S} \|x_i - \mu_S\|_2^2$ for the cost of one group, the same quantity K-means minimizes. Merging A and B can only raise it, and **Ward linkage** calls that rise the distance:

$$d(A, B) = W_{A \cup B} - W_A - W_B = \frac{|A|\,|B|}{|A| + |B|} \|\mu_A - \mu_B\|_2^2.$$

Ward Linkage: Merge Where the Cost Rises Least

Write $W_S = \sum_{i \in S} \|x_i - \mu_S\|_2^2$ for the cost of one group, the same quantity K-means minimizes. Merging A and B can only raise it, and **Ward linkage** calls that rise the distance:

$$d(A, B) = W_{A \cup B} - W_A - W_B = \frac{|A|\,|B|}{|A| + |B|} \|\mu_A - \mu_B\|_2^2.$$

Reading the right-hand side: merges are cheap when the two centers are close, and expensive when both groups are large.

Ward Linkage: Merge Where the Cost Rises Least

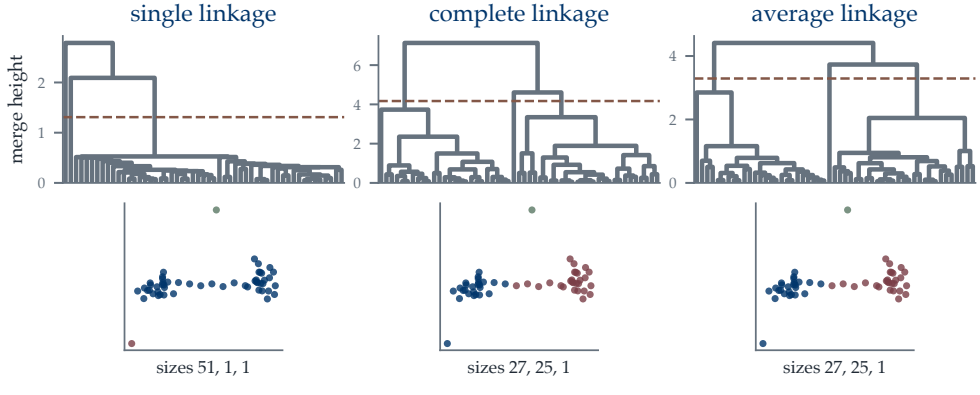
Write $W_S = \sum_{i \in S} \|x_i - \mu_S\|_2^2$ for the cost of one group, the same quantity K-means minimizes. Merging A and B can only raise it, and **Ward linkage** calls that rise the distance:

$$d(A, B) = W_{A \cup B} - W_A - W_B = \frac{|A|\,|B|}{|A| + |B|} \|\mu_A - \mu_B\|_2^2.$$

Reading the right-hand side: merges are cheap when the two centers are close, and expensive when both groups are large.

So Ward resists absorbing a big group into another big one, and tends to return groups of similar size. Appendix A.3 derives the identity.

The Rule Decides the Answer



Above, the tree each rule builds, with the height that leaves three groups marked; below, the grouping that cut produces. Single linkage chains along the bridge, so its tree is a ladder and the cut leaves $51, 1, 1$. Complete and average split the bridge and leave $27, 25, 1$.

Original course simulation, seed 7; each panel cut to give three groups.

Where to Cut

- A horizontal cut at height h keeps the groups that merged below h, so one tree supplies every K without refitting.

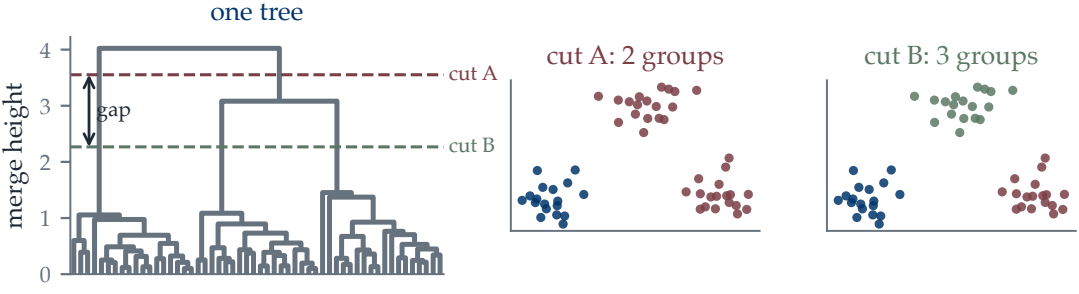
Where to Cut

- A horizontal cut at height h keeps the groups that merged below h, so one tree supplies every K without refitting.
- Cut where the tree has a tall gap: everything below joined cheaply, and the next merge was expensive.

Where to Cut

- A horizontal cut at height h keeps the groups that merged below h, so one tree supplies every K without refitting.
- Cut where the tree has a tall gap: everything below joined cheaply, and the next merge was expensive.
- Or, when K is decided by the user, take the height h that leaves exactly K groups. The tree supports any choice; it does not make the choice.

One Tree, Two Cuts



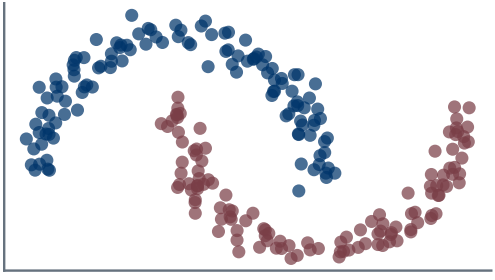
The same tree, cut twice. Above the second merge, cut A leaves two groups; below it, cut B leaves three. The tall gap between those two merge heights is what makes cut B defensible: the three groups joined cheaply, and joining them cost far more.

Original course simulation, seed 5; average linkage on 54 points.

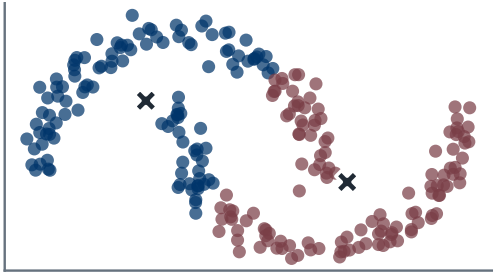
Outline

K-Means Cannot Follow Curved Groups

Connected moons



K-means Voronoi cells



Original course simulation, seed 26518; the same points grouped by connectivity and by nearest center.

When K-Means Works—and When It Does Not

A strong choice when

- Each group is summarized well by one center;
- Groups are compact, roughly round, and have comparable spread;
- Scaled Euclidean distance matches the intended notion of similarity.

A poor choice when

- A group is curved, ring-shaped, or otherwise non-convex;
- Groups have very different sizes or densities, or contain outliers;
- Irrelevant or badly scaled coordinates dominate distance.

With Euclidean distance, nearest-center assignments form convex Voronoi cells. The method is fast and scalable, but no such cell can trace one of the moons above.

Keep the Partitioning Idea, Change What "Close" Means

Straight-line distance makes points across the empty gap look close, even when they lie on different moons. Replace it by a **curved distance**: the length of a route made of short hops between neighboring observations.

Keep the Partitioning Idea, Change What "Close" Means

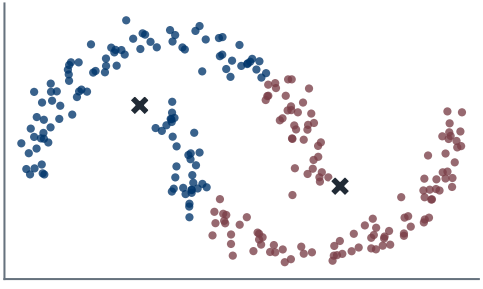
Straight-line distance makes points across the empty gap look close, even when they lie on different moons. Replace it by a **curved distance**: the length of a route made of short hops between neighboring observations.

Use this curved distance to assign each point and to re-choose each representative. For a general distance, the representative is the most central observed point—a **medoid**—rather than a mean.

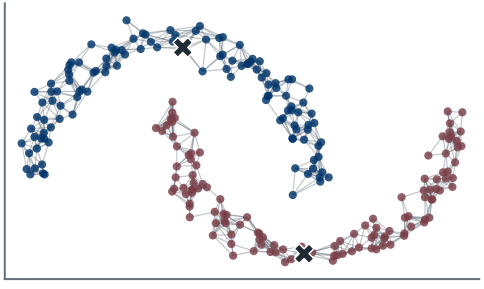
Source: Graph distance and the K-medoids update in the appendix.

Same Moons, Different Distance

Euclidean K-means: 72/240 crossings



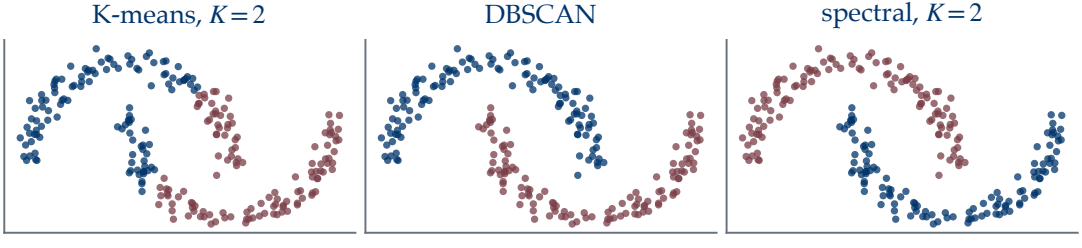
Graph-distance K-medoids: 0/240 crossings



Euclidean *K*-means makes a straight cut and crosses the generating moons for 72 of 240 points. Graph-distance *K*-medoids follows the arcs and crosses them for 0 of 240. The observations did not change; only the geometry did.

Original course simulation, seed 26518; mutual 10-neighbor graph and exact graph shortest paths.

Other Algorithms, Other Shapes



The same two curved groups. *K*-means cuts them straight down the middle, because a nearest-center rule always draws straight boundaries. **DBSCAN** grows groups through dense neighborhoods, and **spectral clustering** partitions a nearest-neighbor graph; both follow the curves.

Original course simulation, seed 26518; DBSCAN with radius 0.22, spectral clustering on a 10-neighbor graph.

An Application: Storing an Image with Few Colors

An RGB image stores three bytes per pixel: this one uses 14,810 distinct colors across 73,080 pixels.

An Application: Storing an Image with Few Colors

An RGB image stores three bytes per pixel: this one uses 14,810 distinct colors across 73,080 pixels.

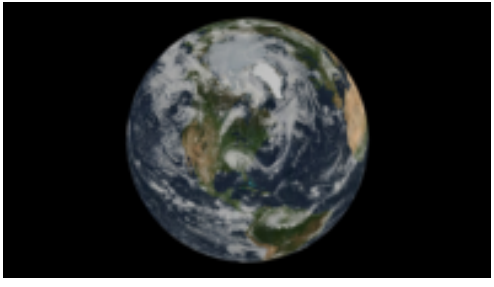
The task: choose a palette of K colors and repaint each pixel with the nearest one. That is K-means, with

- One pixel as a data point in $\mathbb{R}^3$, one palette color as a center;
- W as the total squared color error of the repaint.

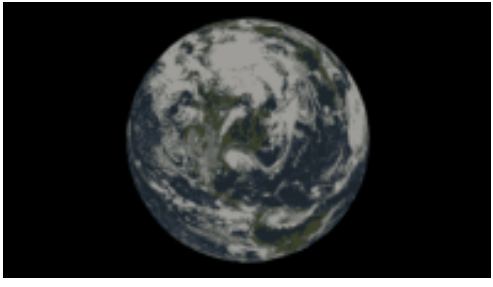
At $K = 8$ a pixel needs 3 bits instead of 24: an eightfold saving, plus the palette.

Eight Colors Are Enough to Recognize the Earth

Original RGB image



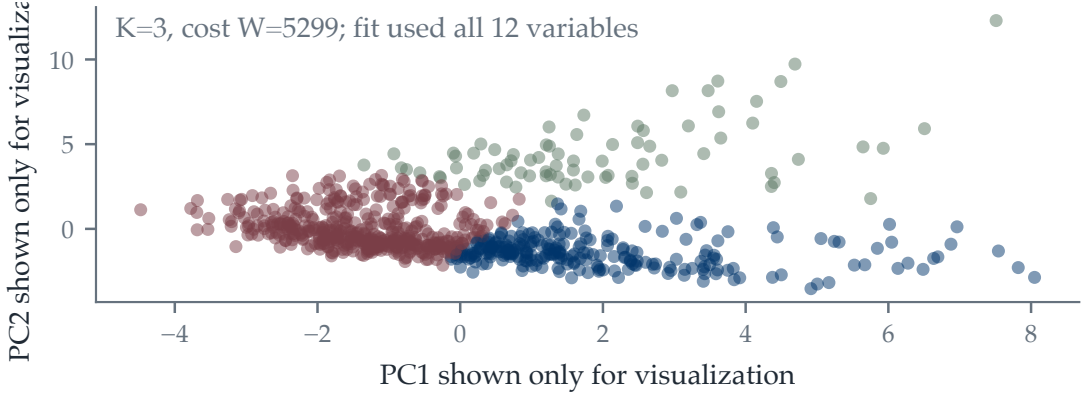
Eight learned colors



The right image uses only the eight centers fitted by K-means on a sample of pixels. Ocean, cloud, land, and ice survive; what is lost is the gradation inside each — the shading of shallow water and the soft edges of cloud.

NASA Blue Marble 2002, resized to 360 × 203; course RGB K-means, 8 centers fitted on 7,000 sampled pixels.

Three College Groups in Standardized Space



Course K-means fit on all 12 standardized variables; PCA scores are used only for the two-dimensional display.

Reading a Cluster Back into the Variables

A partition by itself is a list of labels. To say what a group is, average each variable inside it:

$$\text{profile of cluster } k = (\bar{x}_{k1}, \dots, \bar{x}_{kd}), \qquad \bar{x}_{kj} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}.$$

Reading a Cluster Back into the Variables

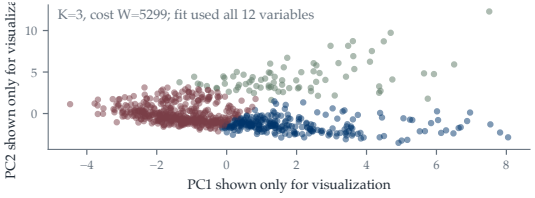
A partition by itself is a list of labels. To say what a group is, average each variable inside it:

$$\text{profile of cluster } k = (\bar{x}_{k1}, \dots, \bar{x}_{kd}), \qquad \bar{x}_{kj} = \frac{1}{n_k} \sum_{i \in C_k} x_{ij}.$$

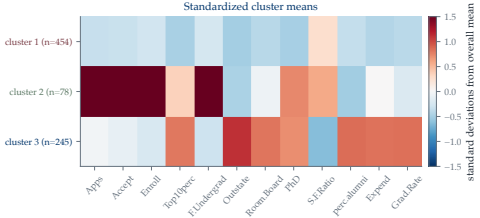
The variables are standardized, so each entry reads as "so many standard deviations above or below the average college" — and the row becomes a description one can argue with.

Interpreting Three Groups in the College Dataset

K-means clustering result



Profiles in the original variables



- **Group 1 (red;** $n = 454$**)** is below average across most variables; **Group 2 (green;** $n = 78$**)** is the large public-college group, about 2.5 SD above average in enrollment and undergraduate population.
- **Group 3 (blue;** $n = 245$**)** is expensive and selective: tuition is about 1.1 SD above average, with high alumni giving and graduation rates.

Clustering in the Deep-Learning Era

- Raw features rarely give a useful distance, as the previous slide showed. Modern practice clusters learned embeddings, where a trained network has already decided what counts as similar.
- The same objective reappears as **vector quantization**: replace a vector by the nearest of K codebook entries. That is the assignment step, and the codebook update is the centroid step.
- A **VQ-VAE** does this inside a network, learning encoder and codebook together, turning a continuous representation into discrete tokens.

Clustering in the Deep-Learning Era

- Raw features rarely give a useful distance, as the previous slide showed. Modern practice clusters learned embeddings, where a trained network has already decided what counts as similar.
- The same objective reappears as **vector quantization**: replace a vector by the nearest of K codebook entries. That is the assignment step, and the codebook update is the centroid step.
- A **VQ-VAE** does this inside a network, learning encoder and codebook together, turning a continuous representation into discrete tokens.

Lecture 17 builds that model, on representations rather than raw measurements.

Summary: K-Means

$$\min_{c,\mu} \sum_i \|x_i - \mu_{c_i}\|_2^2, \qquad \mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i.$$

1. Representation and scaling define the geometry before optimization begins.
2. Assignment and centroid updates reduce W; multiple starts diagnose local minima.
3. Euclidean K-means fits compact centered groups; curved groups need a different distance, representation, or method.

Summary: Evidence for a Partition

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad W(K+1) \leq W(K).$$

1. Internal criteria, seed stability, sample stability, and profiles answer different questions.
2. K, distance, linkage, and algorithm are modeling decisions, not facts in the table.
3. Use-specific validation determines whether a grouping is useful.

Next Lecture

A hard assignment records a college on the boundary as fully inside one group, and the fitted partition carries no trace of how close the call was.

Lecture 16 — Latent Variables and Expectation–Maximization treats membership as an unobserved variable, replaces the label by a probability, and alternates inference about that variable with parameter updates.

Reading for this lecture: ISLP Chapter 12.

A.1 Why the Centroid Is the Mean

derive

For a fixed nonempty cluster C_k, minimize $J_k(\mu) = \sum_{i \in C_k} \|x_i - \mu\|_2^2$.

$$\nabla_\mu J_k(\mu) = -2 \sum_{i \in C_k} (x_i - \mu) = -2 \sum_{i \in C_k} x_i + 2|C_k|\mu.$$

Setting the gradient to zero gives

$$\boxed{\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i}.$$

The mean is special to squared Euclidean loss; other losses imply other representatives.

A.2 Reading the Silhouette Formula

For point i, let $a(i)$ be mean distance to its own cluster and $b(i)$ the smallest mean distance to another cluster. We want a scale-free contrast with range $[-1, 1]$:

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

If $b \geq a$, then $s = 1 - a/b \in [0, 1]$; if $a > b$, then $s = b/a - 1 \in [-1, 0)$.

Near 1 means cohesive and separated, near 0 means boundary-like, and below 0 means closer on average to another cluster. The criterion favors separated compact groups.

A.3 Ward's Identity

derive

For any group S and any point z, expanding the square and using $\sum_{i \in S}(x_i - \mu_S) = 0$ gives

$$\sum_{i \in S} \|x_i - z\|_2^2 = W_S + |S| \, \|\mu_S - z\|_2^2.$$

Apply it to A and to B at the merged mean $\mu = \frac{|A|\mu_A + |B|\mu_B}{|A| + |B|}$:

$$W_{A \cup B} - W_A - W_B = |A| \, \|\mu_A - \mu\|_2^2 + |B| \, \|\mu_B - \mu\|_2^2.$$

Since $\mu_A - \mu = \frac{|B|}{|A|+|B|}(\mu_A - \mu_B)$ and $\mu_B - \mu = -\frac{|A|}{|A|+|B|}(\mu_A - \mu_B)$, the right-hand side collapses:

$$\frac{|A| \, |B|^2 + |B| \, |A|^2}{(|A| + |B|)^2} \|\mu_A - \mu_B\|_2^2 = \frac{|A| \, |B|}{|A| + |B|} \|\mu_A - \mu_B\|_2^2.$$

A.4 Why the Groups Are Bounded by Straight Lines

derive

Fix the centers. Observation x joins group k exactly when $\|x - \mu_k\|_2^2 \leq \|x - \mu_j\|_2^2$ for every j. Expanding both sides, the term $\|x\|_2^2$ cancels:

$$2x^\top(\mu_j - \mu_k) \leq \|\mu_j\|_2^2 - \|\mu_k\|_2^2.$$

Each condition is linear in x, so it describes a half-space, and group k is the intersection of the $K - 1$ half-spaces that beat the other centers. An intersection of half-spaces is a convex region with flat faces: the cells of a **Voronoi diagram** of the centers.

So every group K-means can produce is convex. A curved group such as a moon is out of reach whatever the centers are, which is what the two-moons example shows.

A.5 The Curved Distance Used for the Moons

Construct a **mutual 10-nearest-neighbor graph**:

- Observations x_i and x_j are linked only when each is among the other's ten nearest Euclidean neighbors;
- The length of that edge is the short straight step $\|x_i - x_j\|_2$.

The distance between two observations is the shortest route through this graph:

$$d_G(x_i, x_j) = \min_{\pi : i \rightsquigarrow j} \sum_{(u,v) \in \pi} \|x_u - x_v\|_2.$$

In the displayed simulation the graph has exactly two connected components, one per moon. A route follows a moon; there is no edge that jumps the empty gap.

A.6 K-Medoids with the Graph Distance

A mean is not defined for a graph distance, so each representative is an observed point m_k. Alternate two steps:

$$c_i \leftarrow \arg\min_k d_G(x_i, m_k), \qquad m_k \leftarrow \arg\min_{x_j : c_j = k} \sum_{i : c_i = k} d_G(x_i, x_j).$$

The update selects the member with the smallest total distance to the other members: the **medoid**.

For disconnected pairs, the course visualization uses a penalty equal to twice the largest finite graph distance. The labels generating the moons are used only to report the crossing count, never to fit the partition.

A.7 DBSCAN: a Group Is a Crowded Region

Fix a radius ε and a minimum count m. Let $N_\varepsilon(x_i) = \{x_j : \|x_i - x_j\|_2 \le \varepsilon\}$.

- x_i is a **core point** when $|N_\varepsilon(x_i)| \ge m$.
- Starting from a core point, keep adding nearby core points and their neighbors. Everything reached belongs to the same group.
- A non-core point reached from a core point is a border point; an unreached point is labeled noise.

No value of K is required, and a group can bend as long as a chain of crowded neighborhoods continues. The difficulty is choosing (ε, m) when groups have different densities.

A.8 Spectral Clustering: Partition a Graph

Build a neighbor graph with affinity matrix A, degrees $D_{ii} = \sum_j A_{ij}$, and graph Laplacian

$$L = D - A.$$

Its eigenvectors with the smallest eigenvalues vary slowly across strong graph connections. Use K such eigenvectors as new coordinates for each observation, then run K-means in that coordinate system.

A good split keeps many links inside each group and cuts few links between groups. This follows curved groups, but it still requires K, a graph-building scale, and an eigenvector computation.