PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

K-means clustering



Week 7 recap

Kahoot-style warm-up

6 quick questions on PCA.

Not graded.

Use a pseudonym if you want.

Fast and correct wins.



Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/quiz-play.html?quiz=pls206-w8-pca-recap]

[host screen]



1 **Why cluster**
2 The algorithm
3 Many starts
4 Distance & scale
5 Choosing k
6 Colombia

Supervised Learning (Classification Algorithm)


Duck


Duck

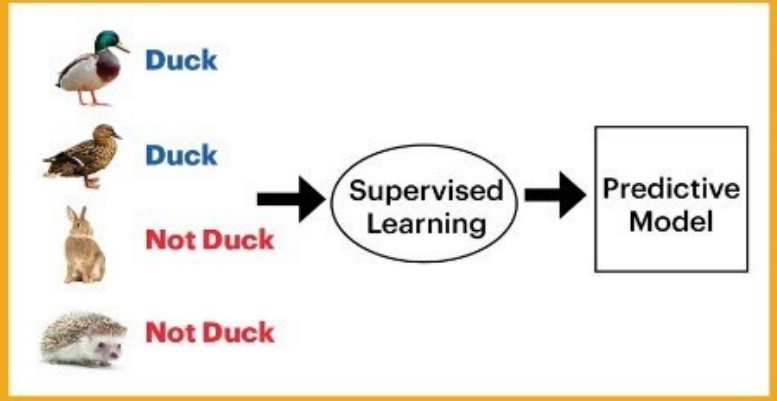

Not Duck

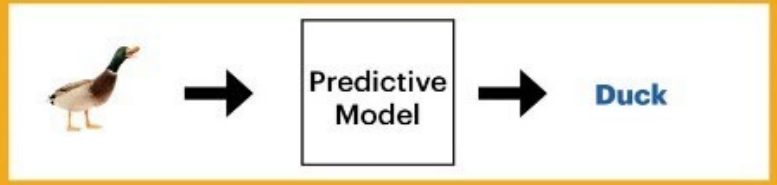

Not Duck

Supervised Learning → Predictive Model

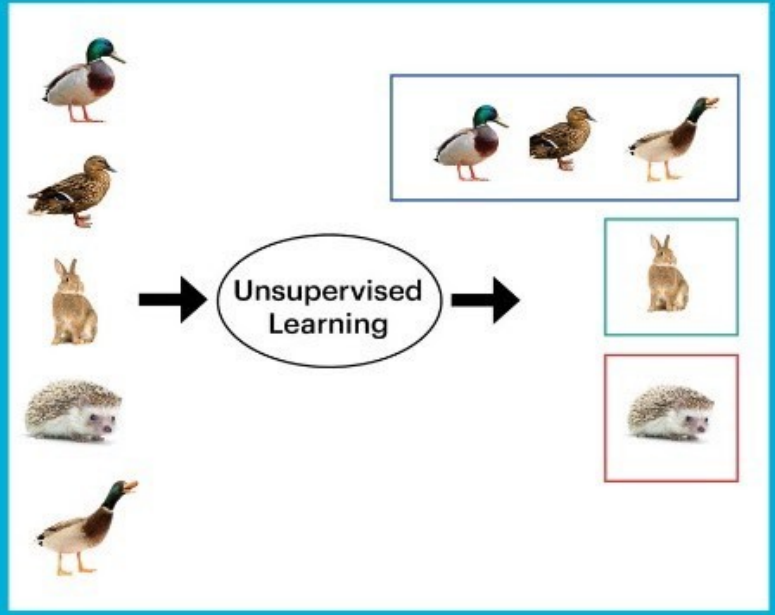

→ Predictive Model → Duck

Supervised Learning (Classification Algorithm)

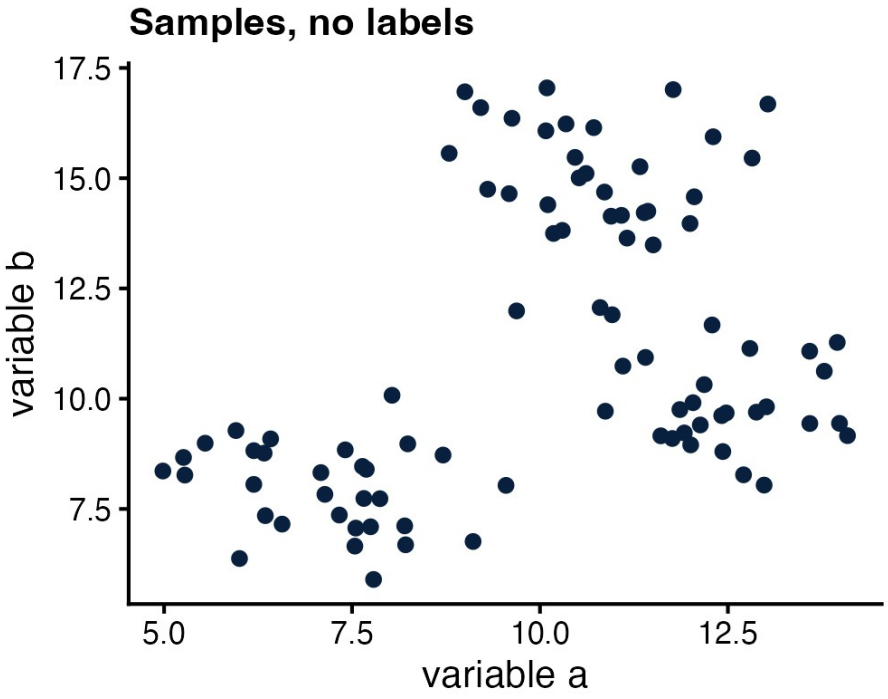


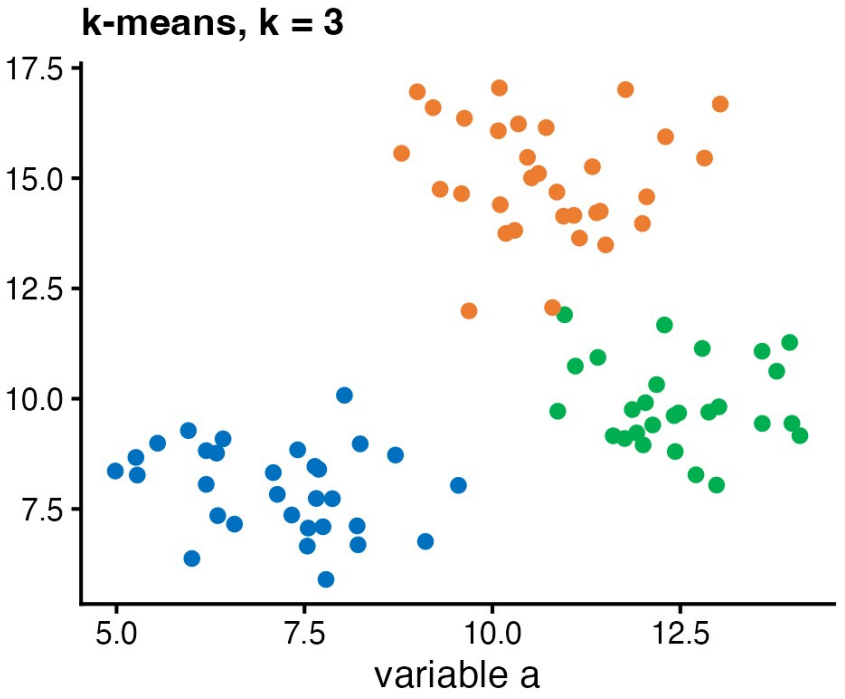


Unsupervised Learning (Clustering Algorithm)



Clustering: groups from the data alone

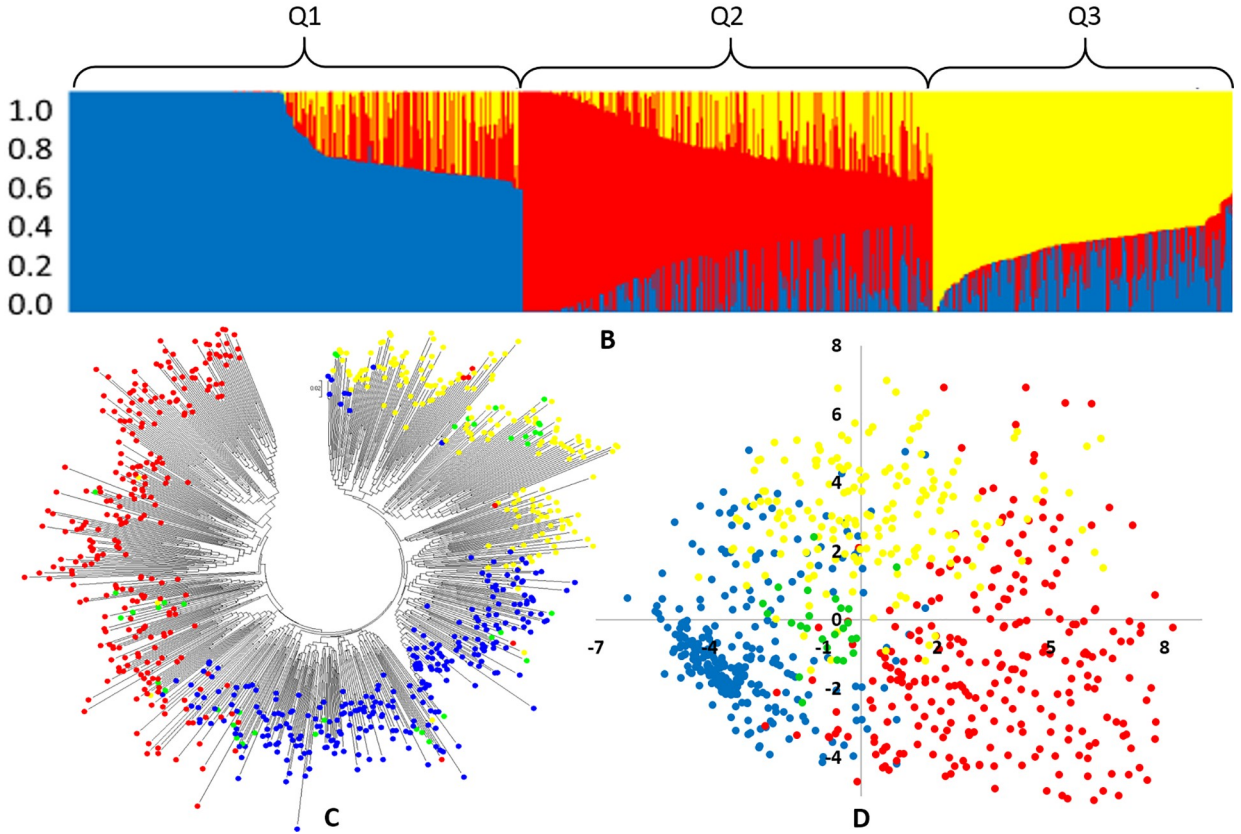




No labels go in. **Groups of similar samples come out.**

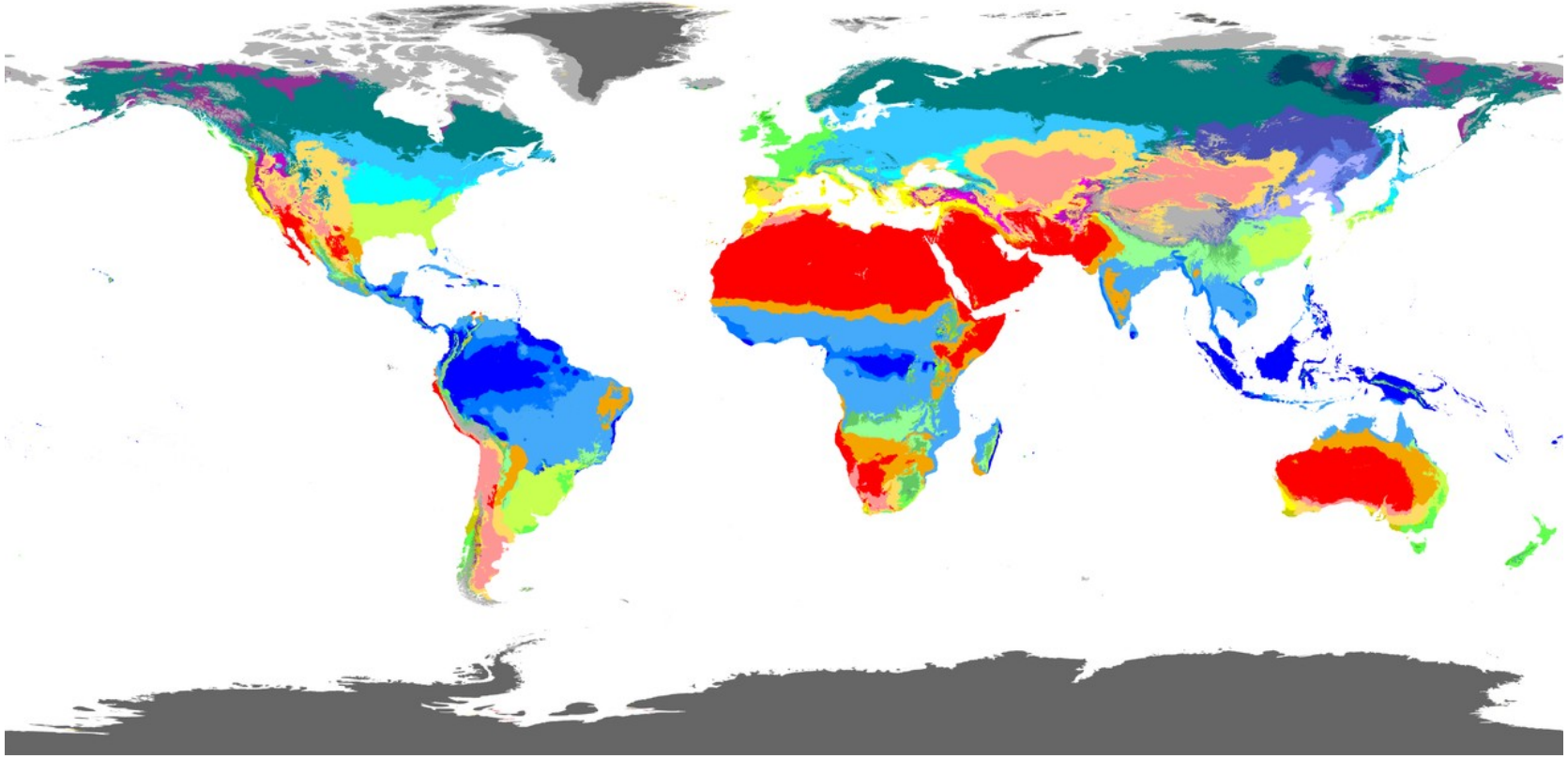
Clustering genotypes into populations by allele data





Xiong et al

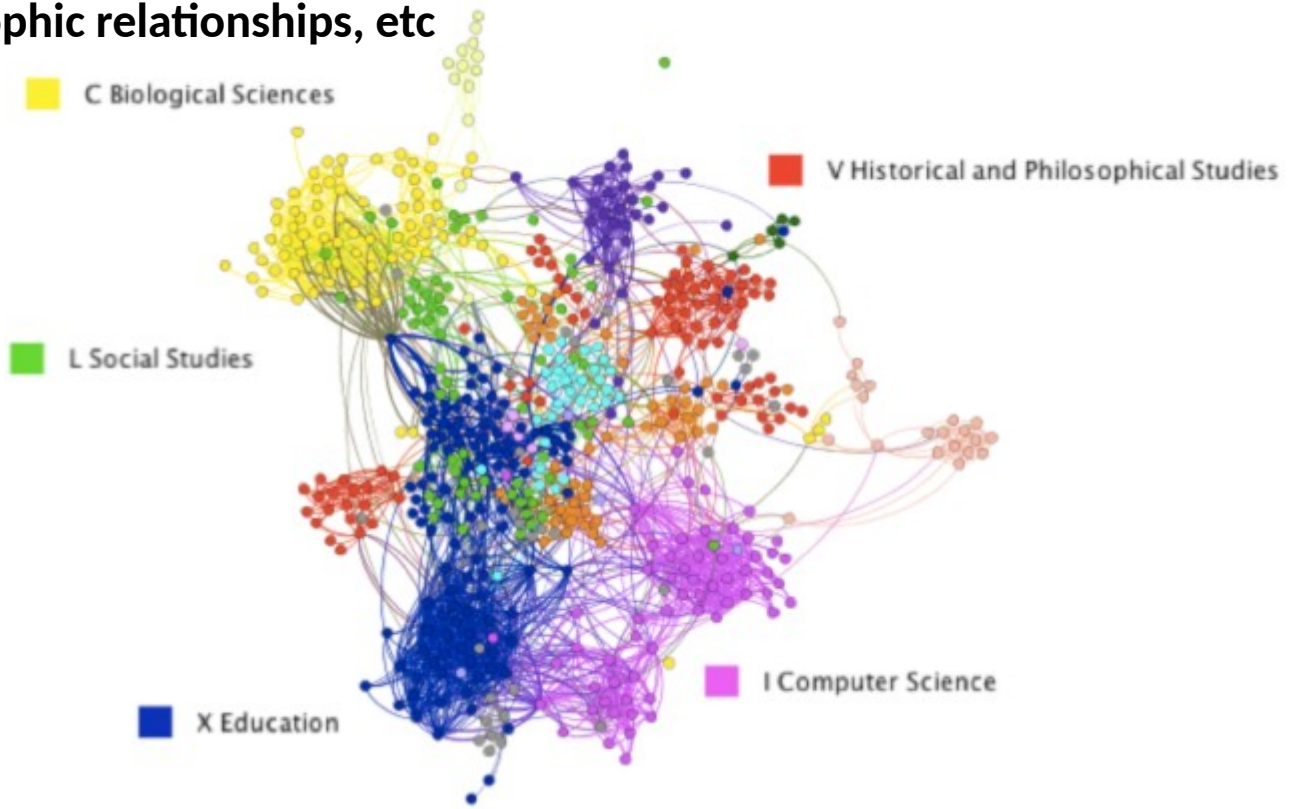
Clustering regions into categories by climate data



Wikipedia

Clustering interacting parts into networks by connection data

Could apply to social interactions in animals, genes in a regulatory network, trophic relationships, etc



ScienceAdvances

Current Issue | First release papers | Archive | About

HOME > SCIENCE ADVANCES > VOL. 6, NO. 22 > ELUCIDATING ECOLOGICAL COMPLEXITY: UNSUPERVISED LEARNING DETERMINES GLOBAL MARINE ECO-...

RESEARCH ARTICLE | OCEANOGRAPHY

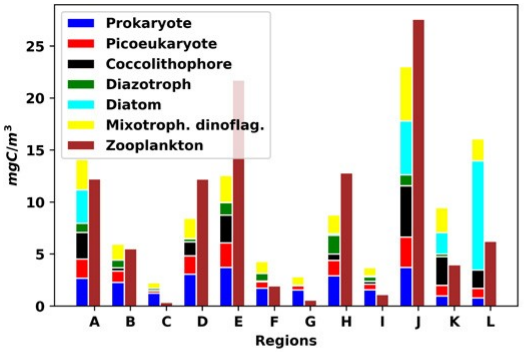
Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces

MAIKE SONNEWALD, STEPHANIE DUTKIEWICZ, CHRISTOPHER HILL, AND GAEL FORGET Authors Info & Affiliations

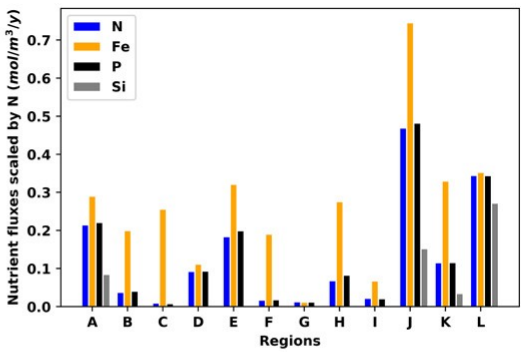
SCIENCE ADVANCES · 29 May 2020 · Vol 6, Issue 22 · DOI: 10.1126/sciadv.aay4740

2,268 27

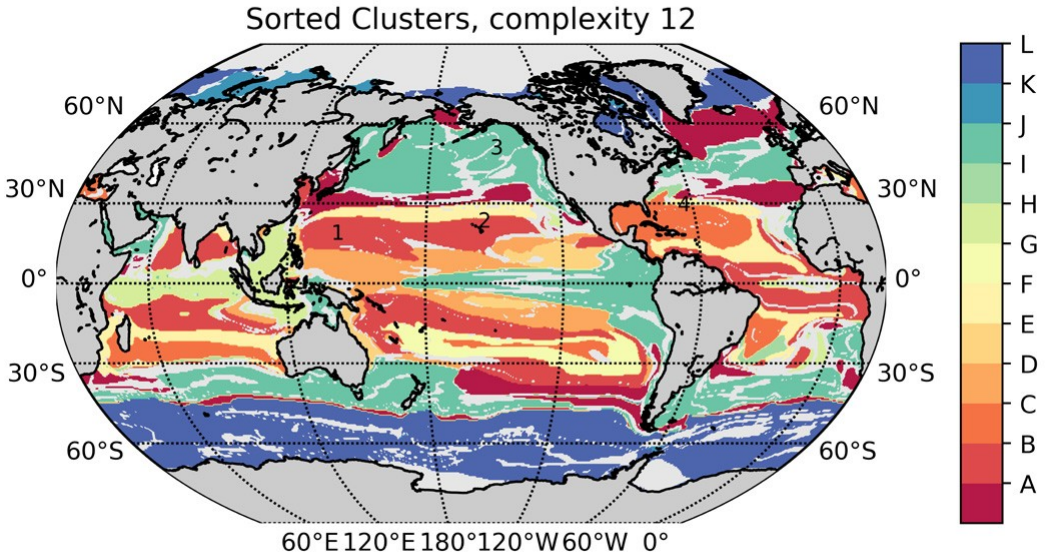
"improve understanding and monitoring of marine ecosystems"



A Ecological ensemble



B Nutrient fluxes



C Global provinces

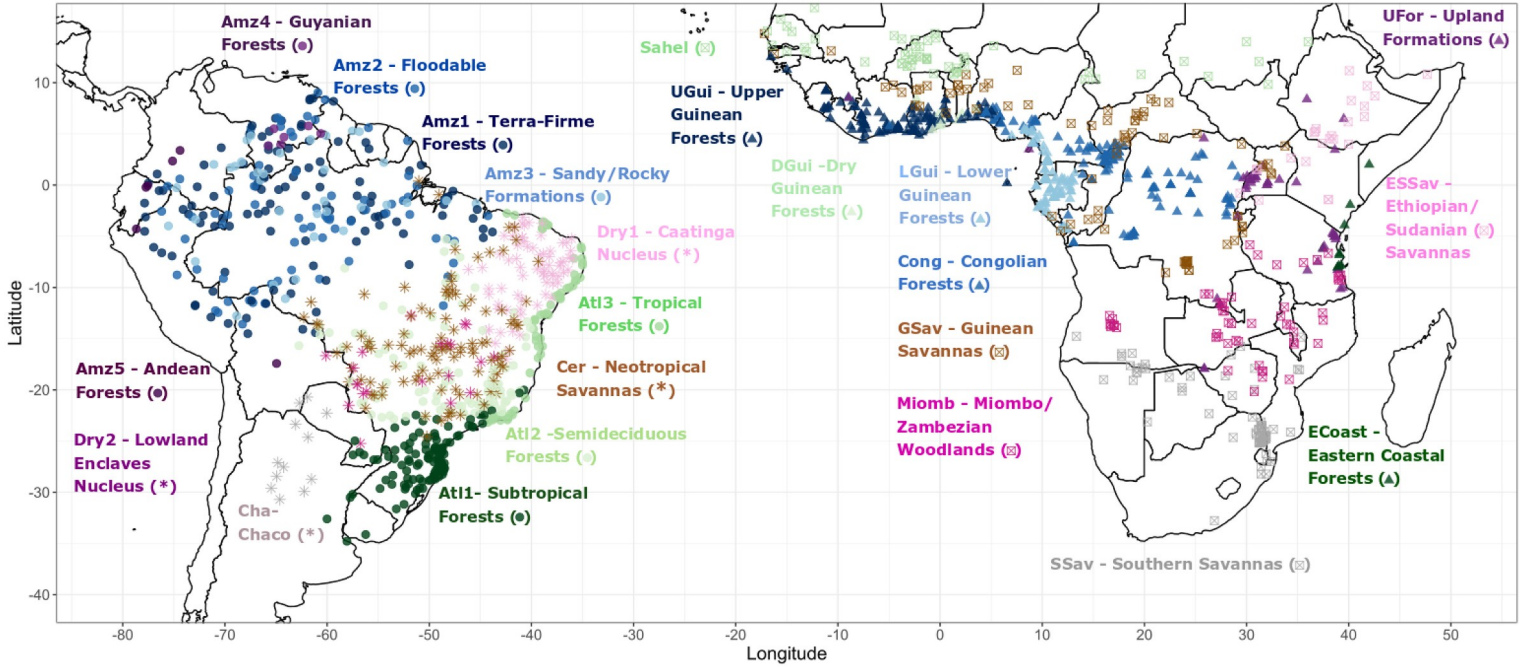
RESEARCH ARTICLE | ECOLOGY |

Dissecting the difference in tree species richness between Africa and South America

Pedro Luiz Silva de Miranda, Kyle G. Dexter, Michael D. Swaine, +2, and Adeline Fayolle Authors Info & Affiliations

Edited by Douglas Schemske, Michigan State University, East Lansing, MI; received July 4, 2021; accepted February 17, 2022

March 29, 2022 | 119 (14) e2112336119 | https://doi.org/10.1073/pnas.2112336119





1 Why cluster

2 **The algorithm**

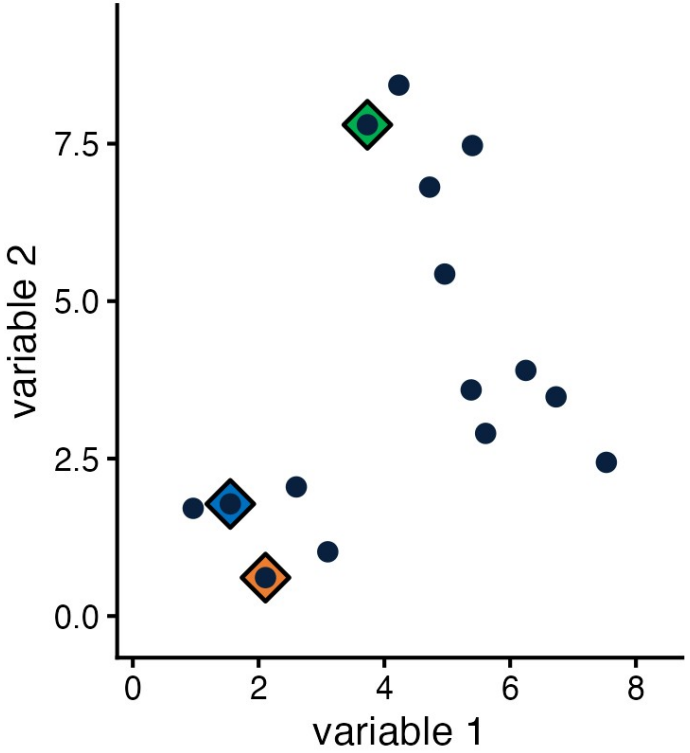
3 Many starts

4 Distance & scale

5 Choosing k

6 Colombia

Step 1: drop k random centres



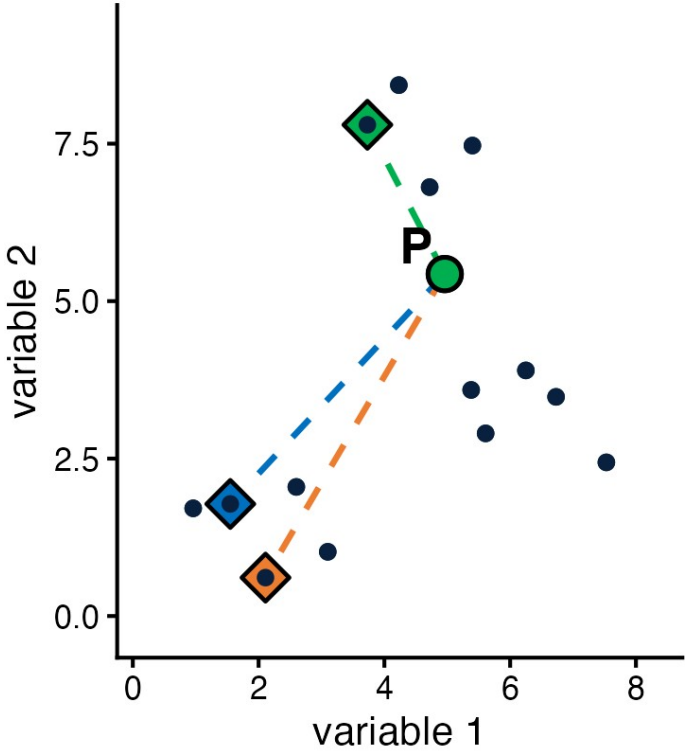
15 samples

2 variables

k = 3

We choose k. **Three random samples become the first centres.**

Step 2: distance to every centre



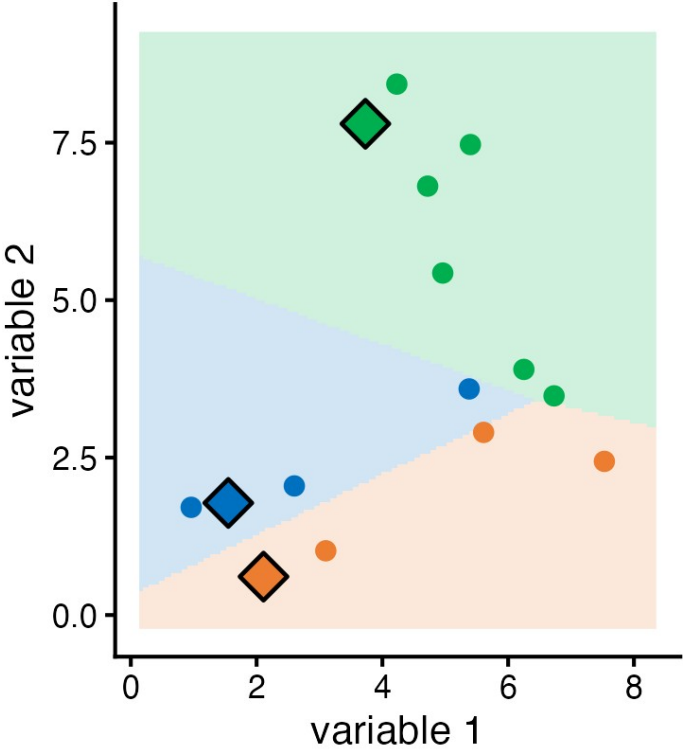
P to blue: **5.0**

P to orange: **5.6**

P to green: **2.7**

P joins the **nearest centre**: green

Step 3: every sample joins its nearest centre

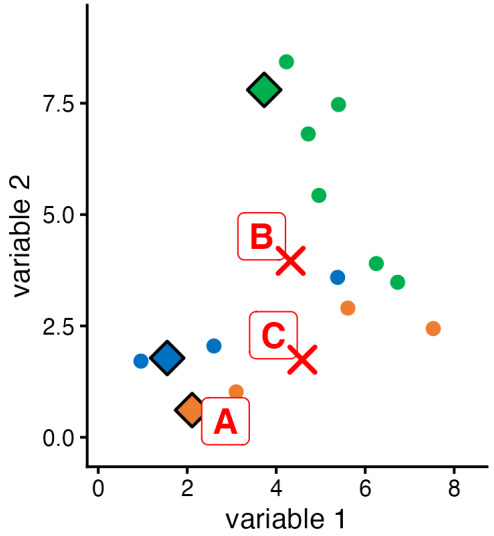


Shaded = each centre's **territory**: everywhere closer to it than to the others

Live poll

Multiple choice — picture

Step 4 moves every centre. Where does the orange centre go?

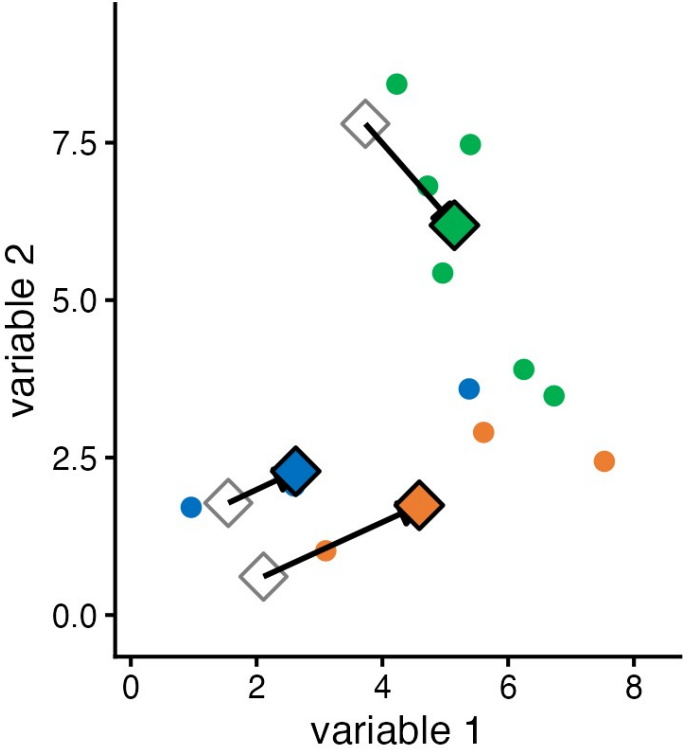


Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w8-next-centre](greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w8-next-centre)

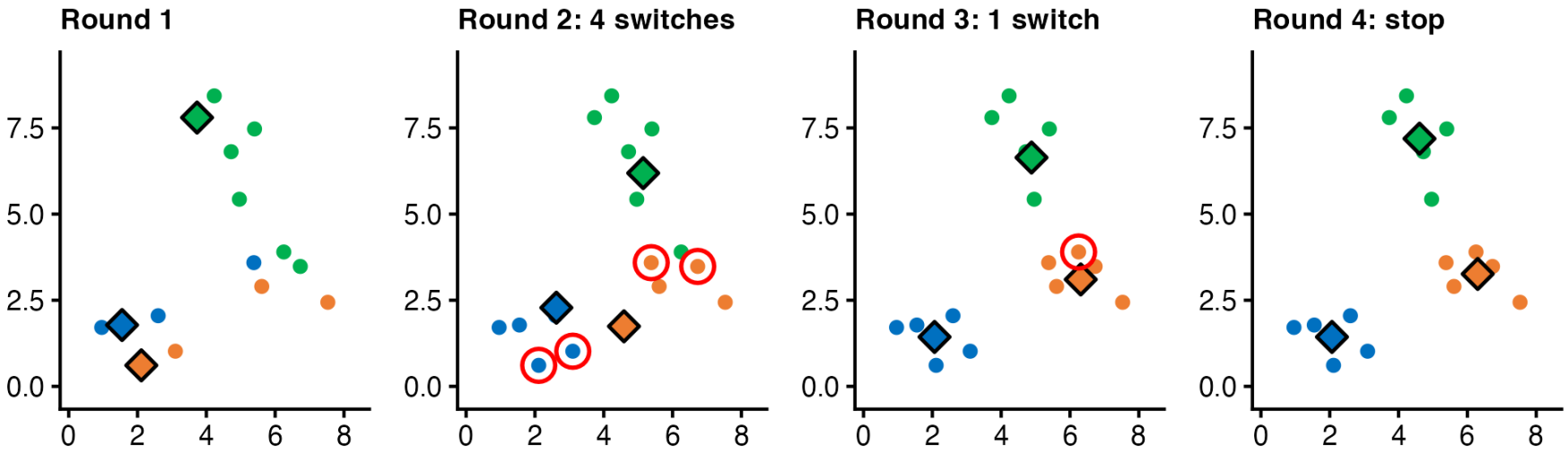
[host screen]

Step 4: move each centre to its cluster's mean



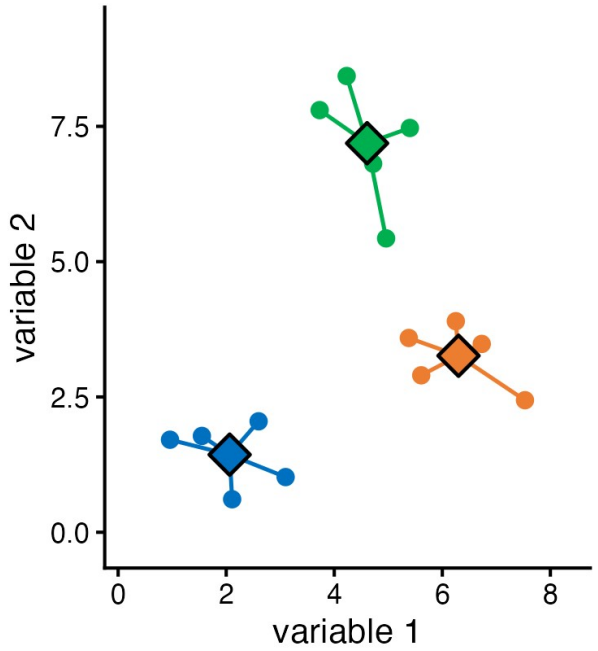
New centre = **the mean** of the samples that joined it. That's the 'means' in k-means.

Repeat until nothing changes



Assign, move, assign, move… **Stop when no sample switches.**

The score: total within-cluster SS



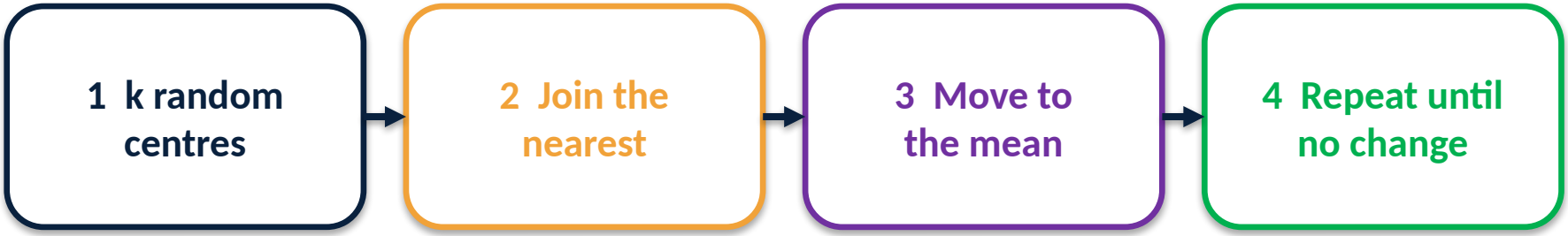
$$\mathrm{SS} = \sum_{\text{samples}} d^2$$

d: distance from a sample to its own centre

Here: **15.6**

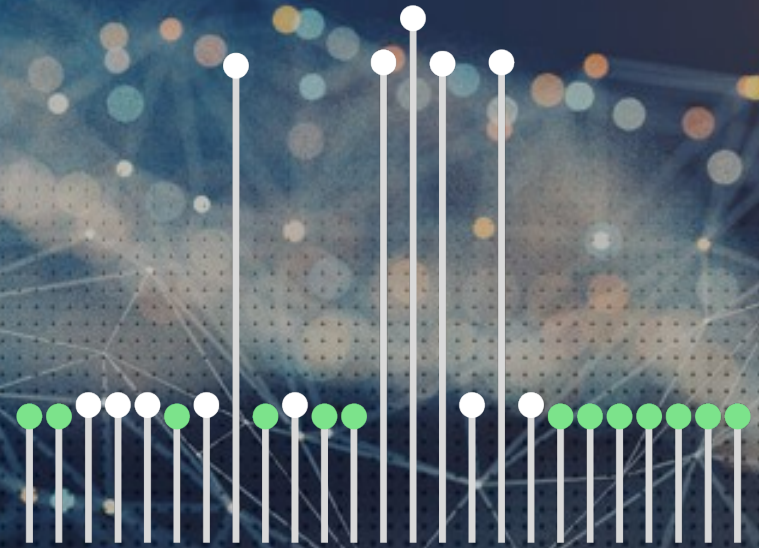
Tight clusters = **small SS**. Same sum of squares as in lm().

k-means in four steps

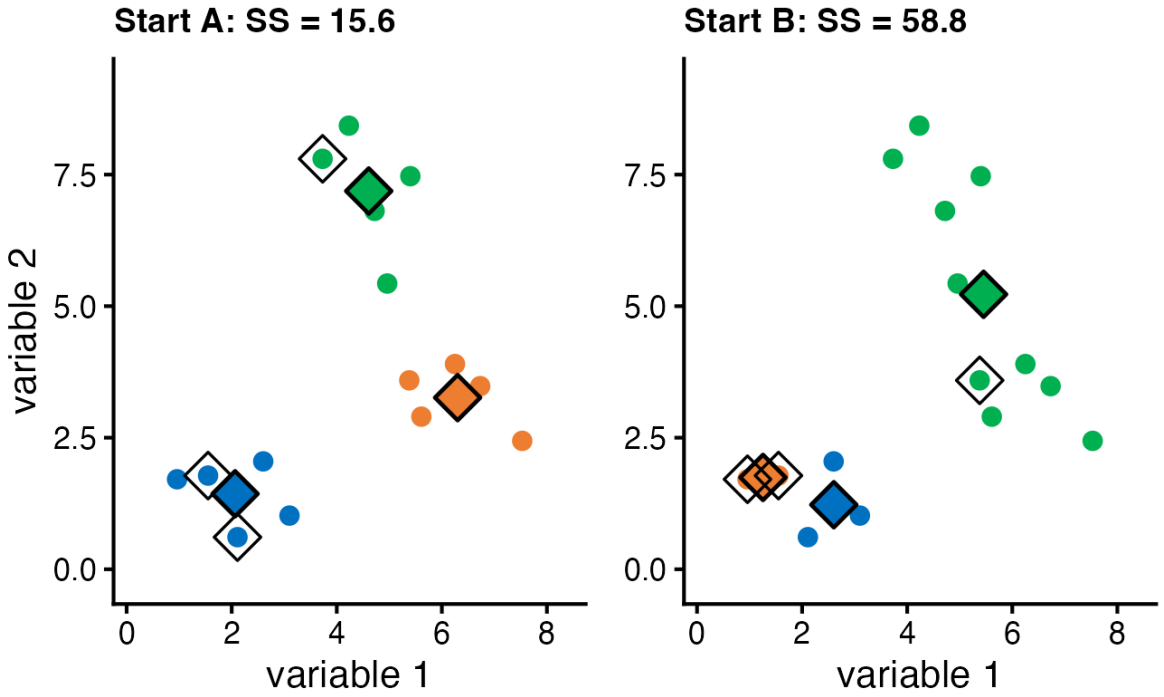


Each round can only lower the SS, so it always stops. **But where it stops depends on the start.**

1 Why cluster
2 The algorithm
3 **Many starts**
4 Distance & scale
5 Choosing k
6 Colombia

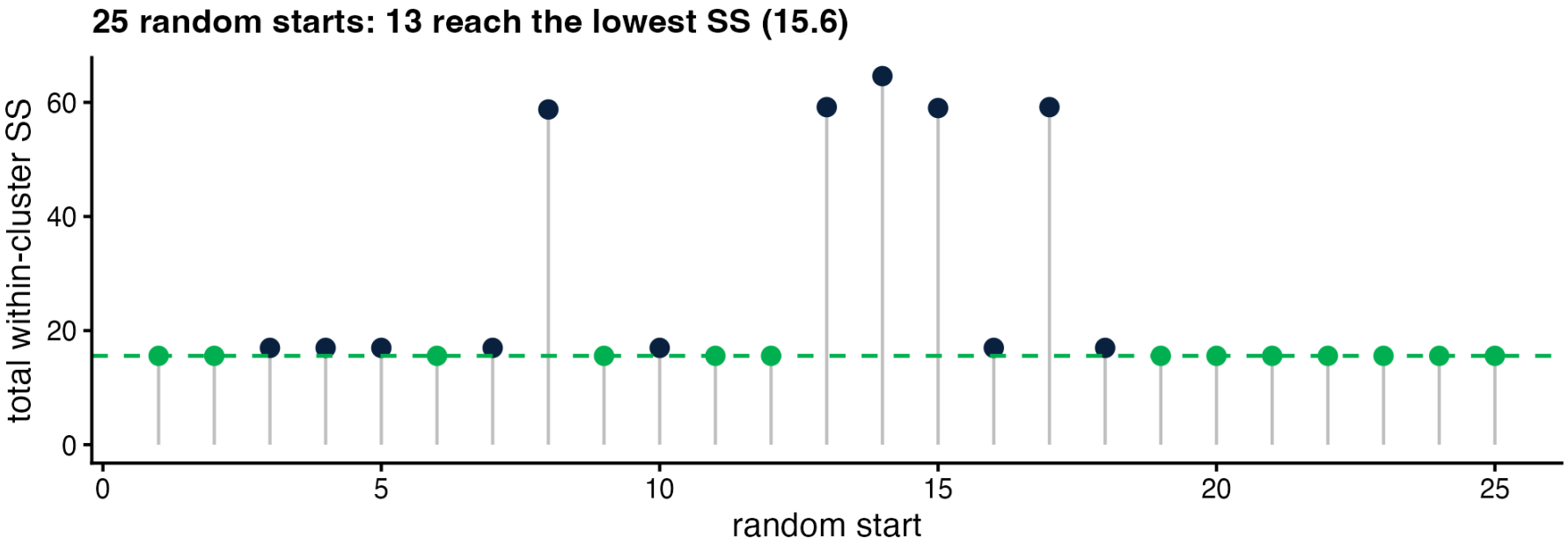


A different start, a worse answer



Dashed diamonds = the starting centres. Start B gets stuck: **SS 58.8 vs 15.6**

Many starts: keep the lowest SS



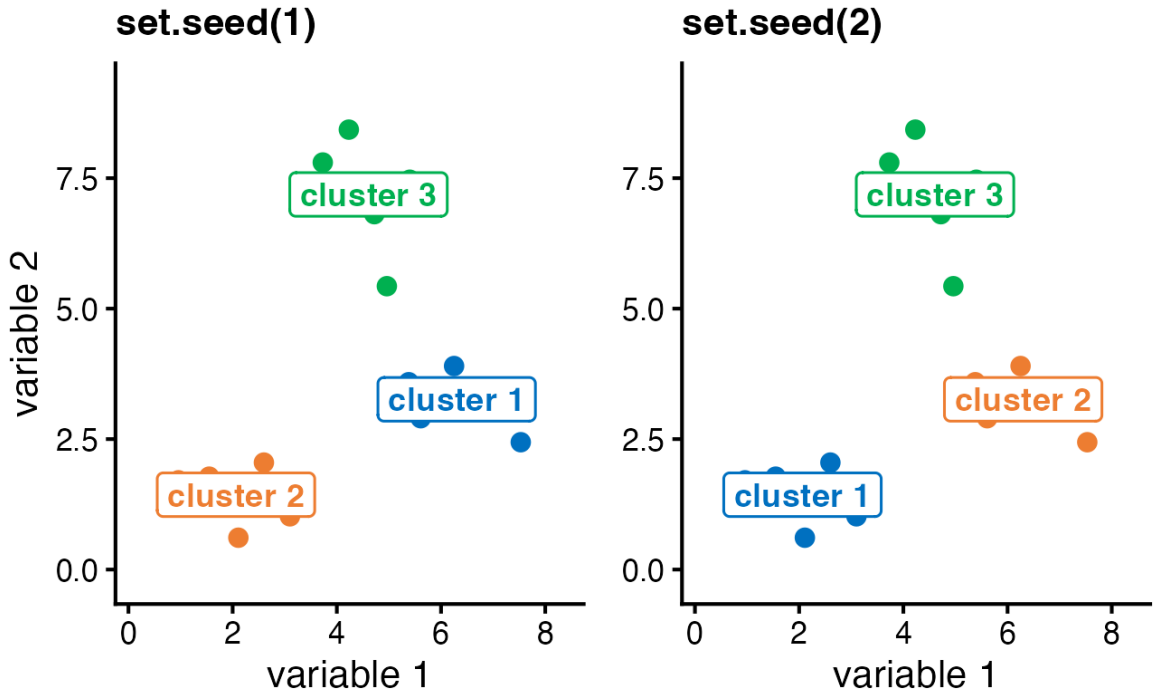
nstart = 25: run 25 random starts, keep the one with the smallest SS

In R: kmeans()

```
> set.seed(1)
> km <- kmeans(pts, centers = 3, nstart = 25)
> km$cluster
 [1] 2 2 2 2 2 1 1 1 1 1 3 3 3 3 3   ← cluster of each sample
> km$centers
     v1    v2
1 6.300 3.262   ← centres = cluster means
2 2.064 1.434
3 4.608 7.188
> km$tot.withinss
[1] 15.57068   ← total within-cluster SS
```

$cluster who is where **$centers** the means **$tot.withinss** the score

Cluster numbers are just names



Same clusters, **different numbers**. Describe a cluster by what's in it, not by its number.



1 Why cluster

2 The algorithm

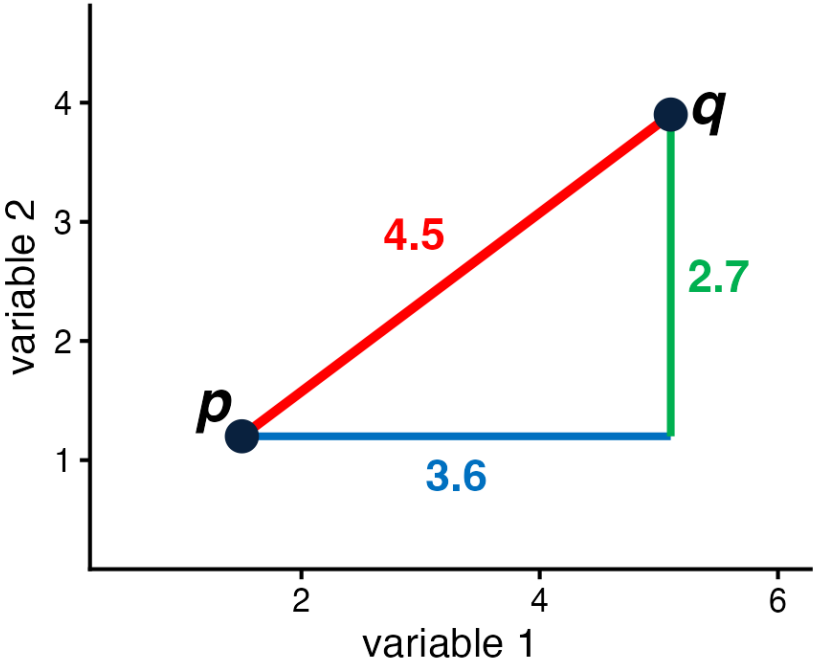
3 Many starts

▶ **4 Distance & scale**

5 Choosing k

6 Colombia

Distance: a straight line, in any number of variables



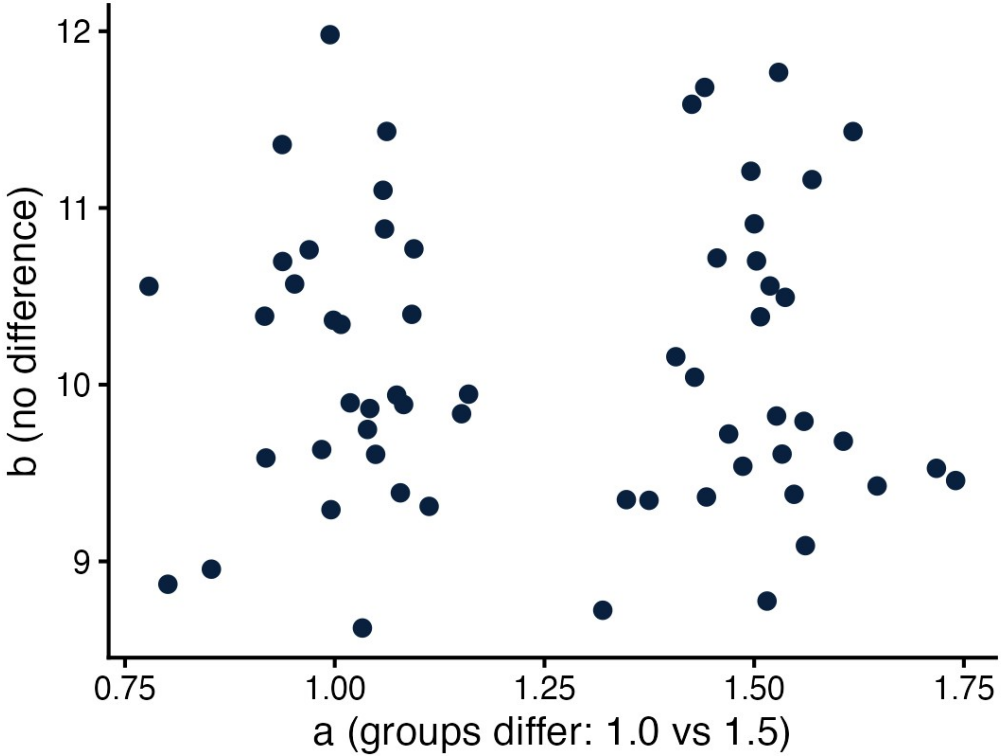
$$d = \sqrt{\Delta_1^2 + \Delta_2^2 + \cdots + \Delta_m^2}$$

$$= \sqrt{3.6^2 + 2.7^2} = 4.5$$

Δ_j: difference in variable j m: number of variables

Pythagoras, one term per variable. **Every variable's units count directly.**

Two groups, very different units



a: SD **0.27**

b: SD **0.83**

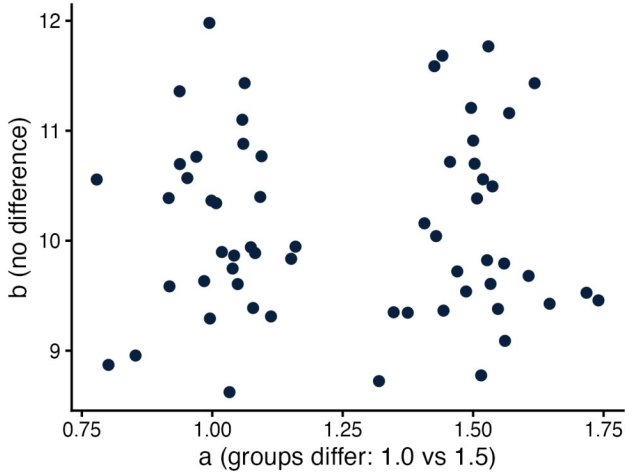
Groups differ in a only

Live poll

Multiple choice — picture

k = 2 on these raw numbers. Which split does k-means find?

A Left vs right **B Top vs bottom**

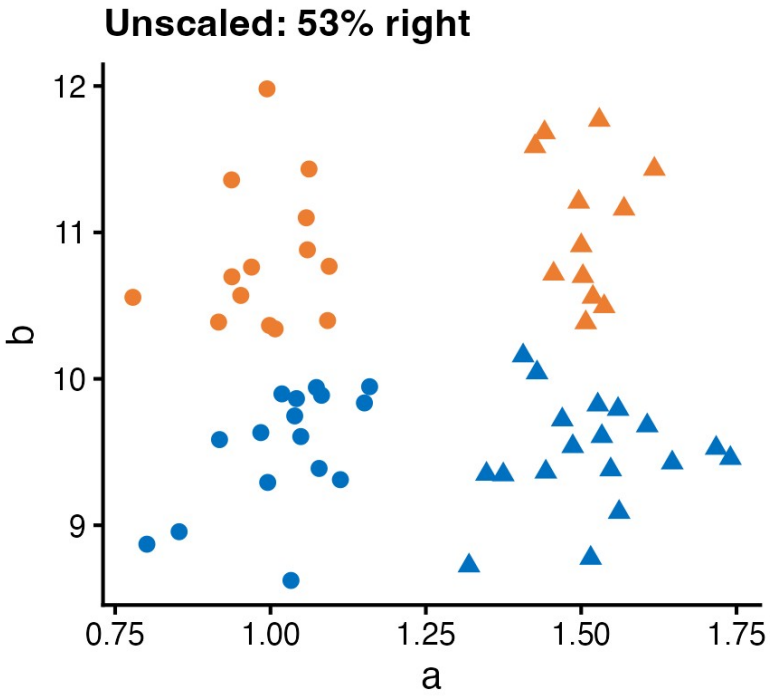


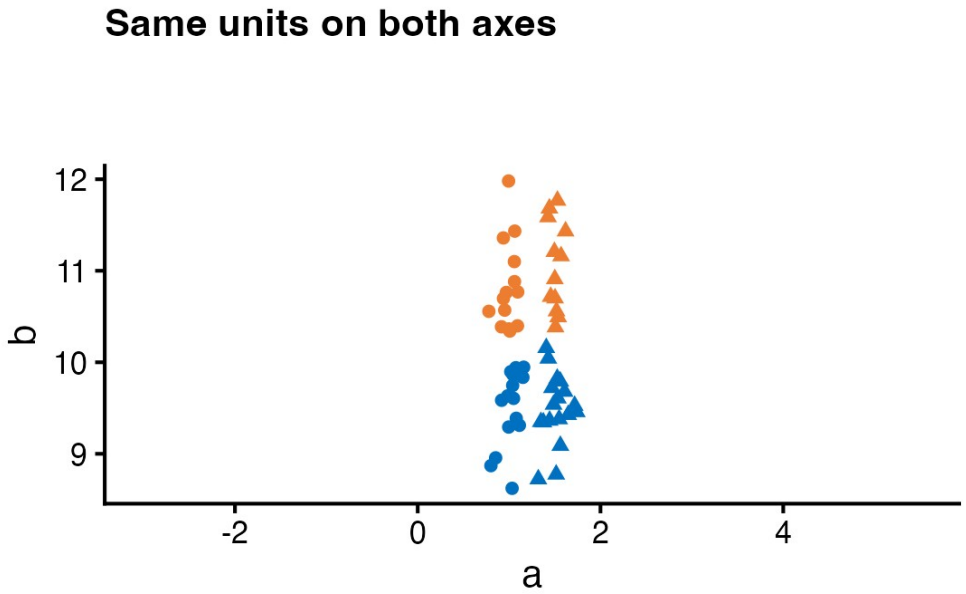
Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w8-scale](greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w8-scale)

[host screen]

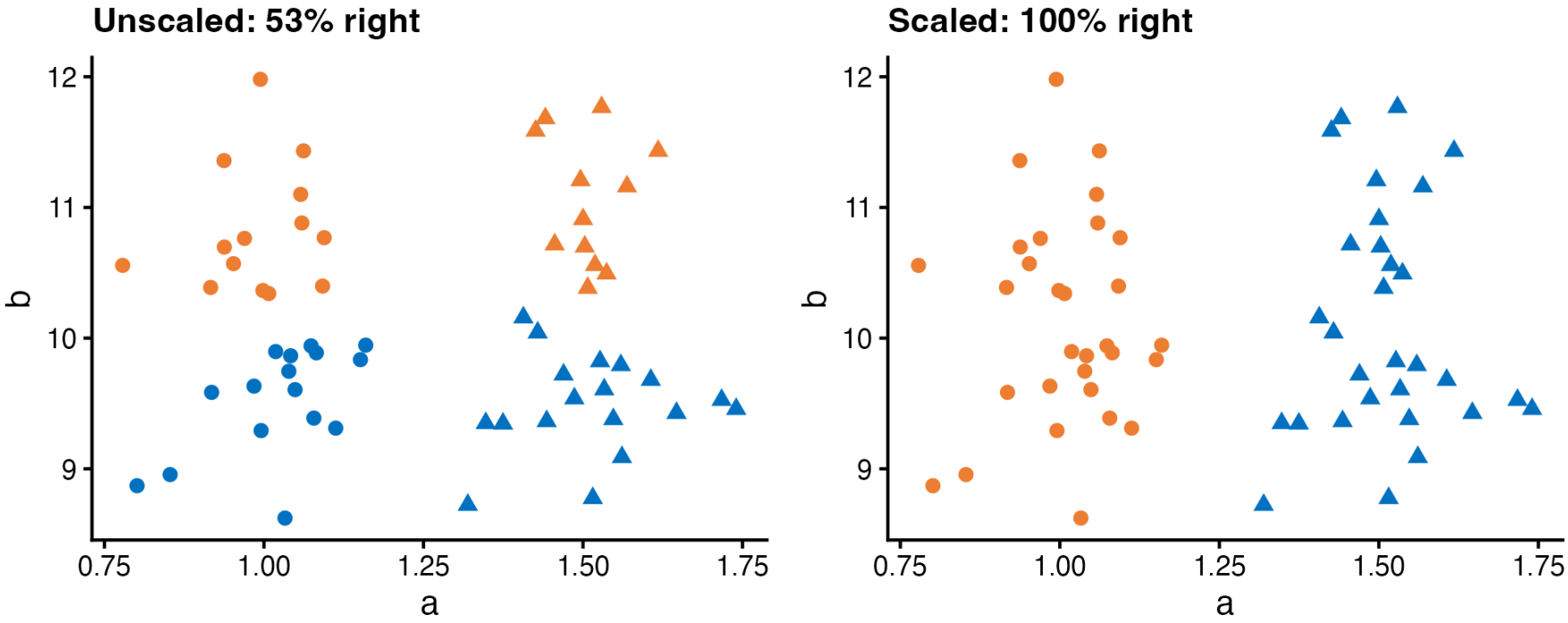
Unscaled: the big numbers win





Shape = true group, colour = cluster. **Only 53% right: a coin flip.**

Scale first: every variable SD 1



kmeans(scale(x), ...): the groups come back, **100%** right

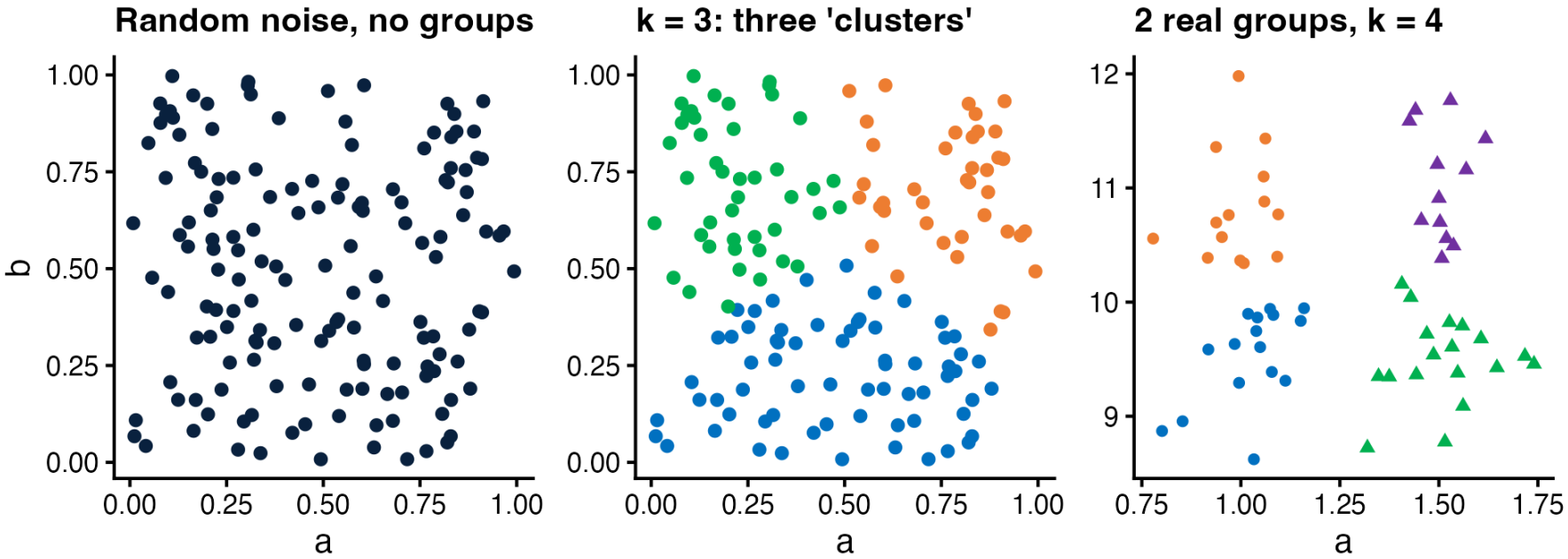
In R: compare with table()

```
> set.seed(1)
> ku <- kmeans(sim2[, c("a", "b")], centers = 2, nstart = 25)
> table(ku$cluster, sim2$group)

   group1 group2
1      16     18  ← unscaled: a coin flip
2      14     12
> set.seed(1)
> ks <- kmeans(scale(sim2[, c("a", "b")]), centers = 2, nstart = 25)
> table(ks$cluster, sim2$group)

   group1 group2
1       0     30  ← scaled: perfect
2      30      0
```



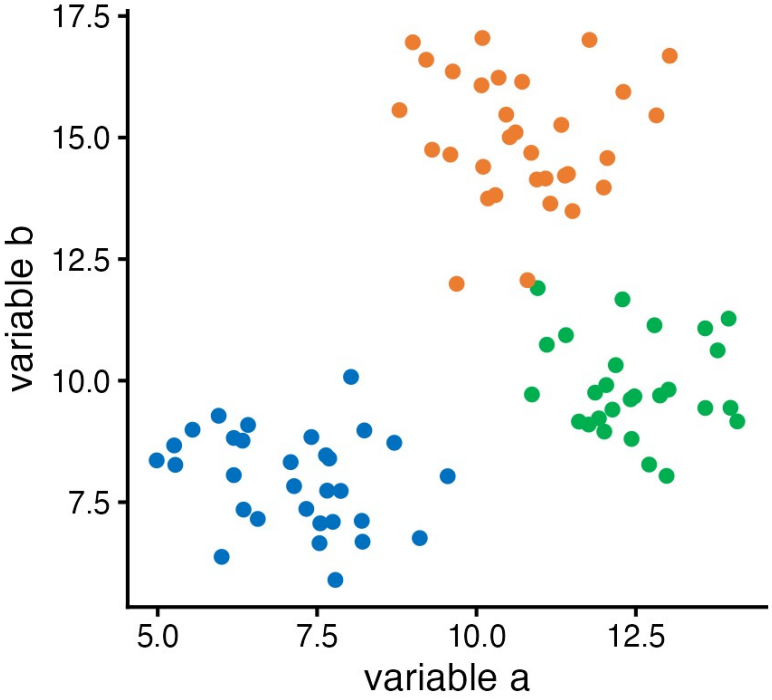
1 Why cluster
2 The algorithm
3 Many starts
4 Distance & scale
5 Choosing k
6 Colombia

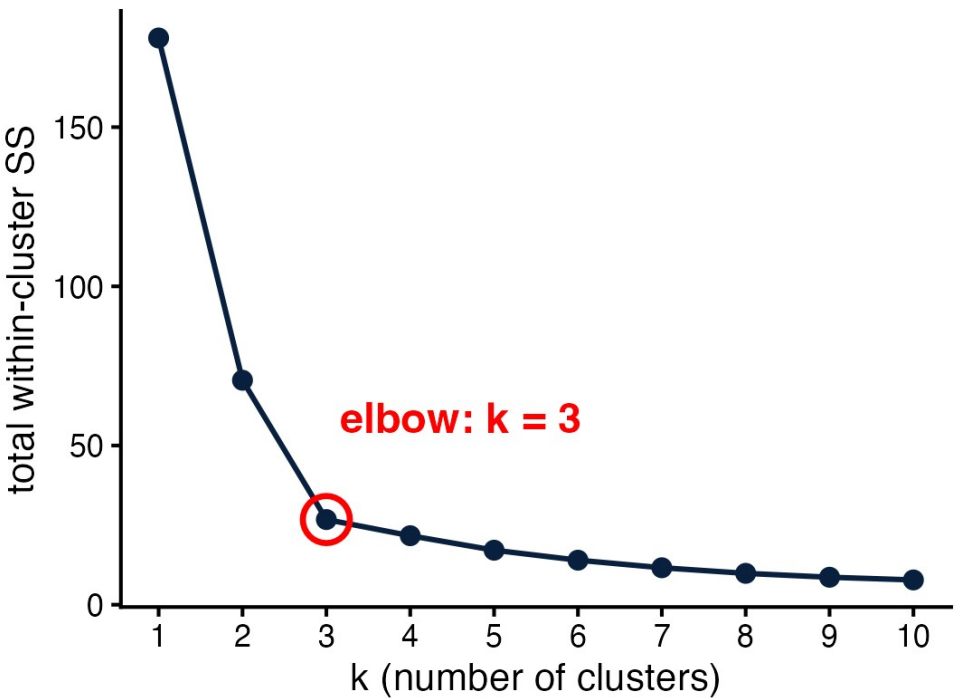
k-means always finds k clusters



Ask for k, get k, **whether or not the groups are real**

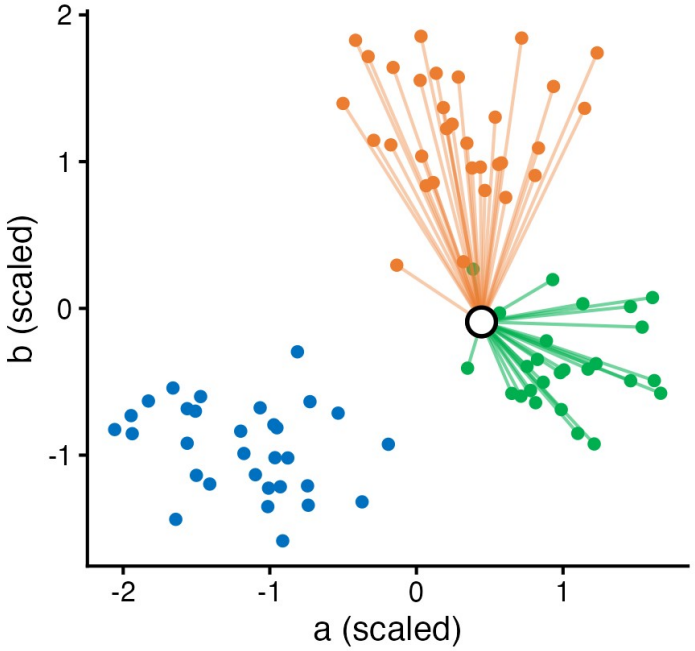
Elbow: where more clusters stop paying





SS always falls as k grows. **Look for the bend.**

Silhouette: is this sample in the right cluster?



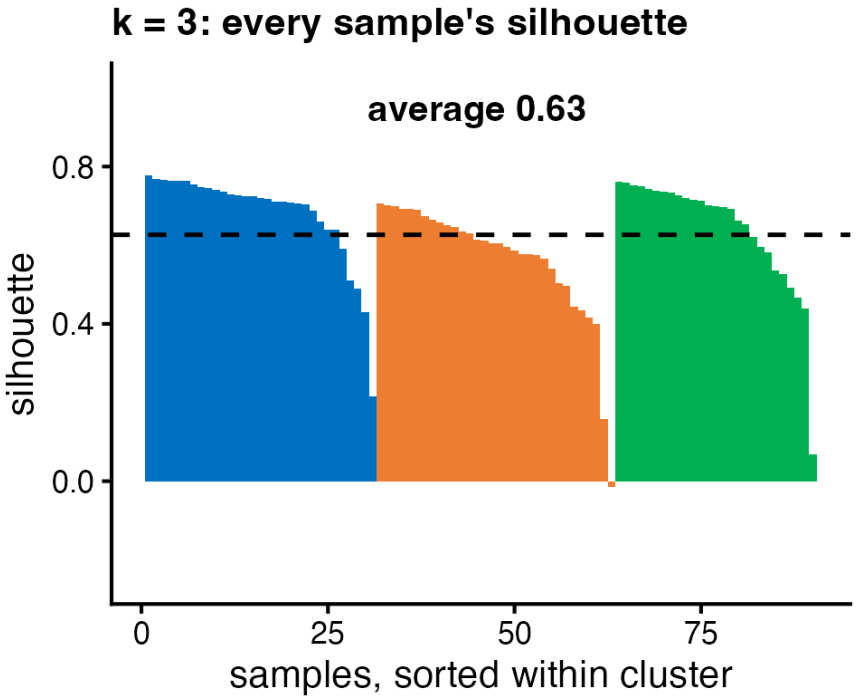
$$s = \frac{b - a}{\max(a, b)}$$

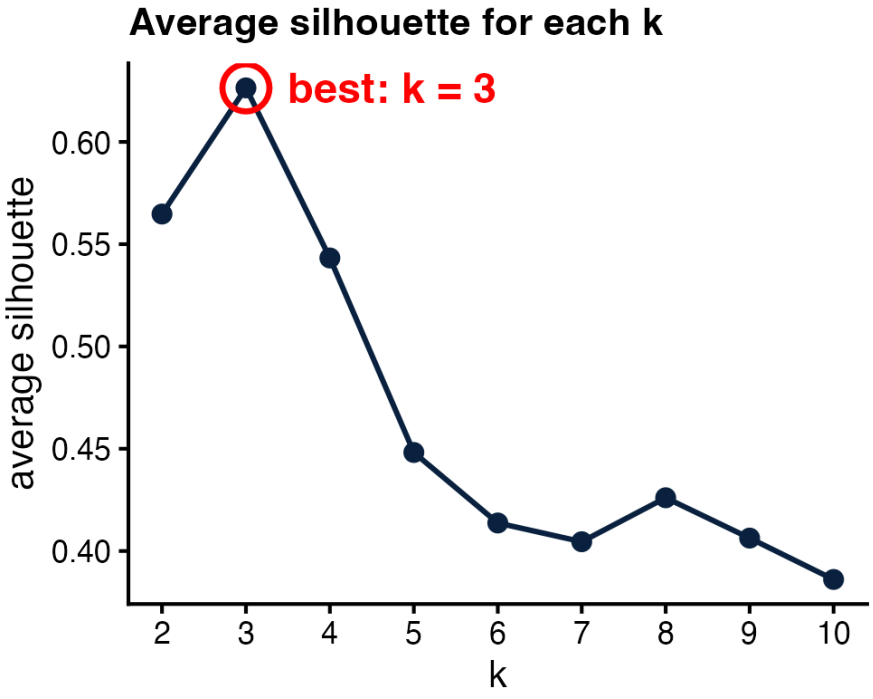
$$= \frac{1.38 - 0.74}{1.38} = 0.47$$

near 1: well inside near 0: on a border

below 0: wrong cluster?

Average silhouette: pick the highest



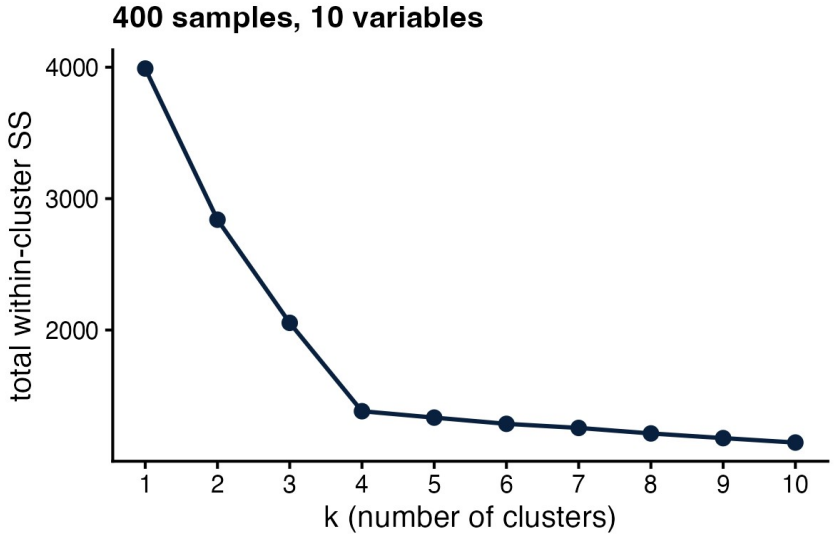


Silhouette and elbow agree: **k = 3**

Live poll

Multiple choice — picture

400 samples, 10 variables. How many clusters does this elbow suggest?

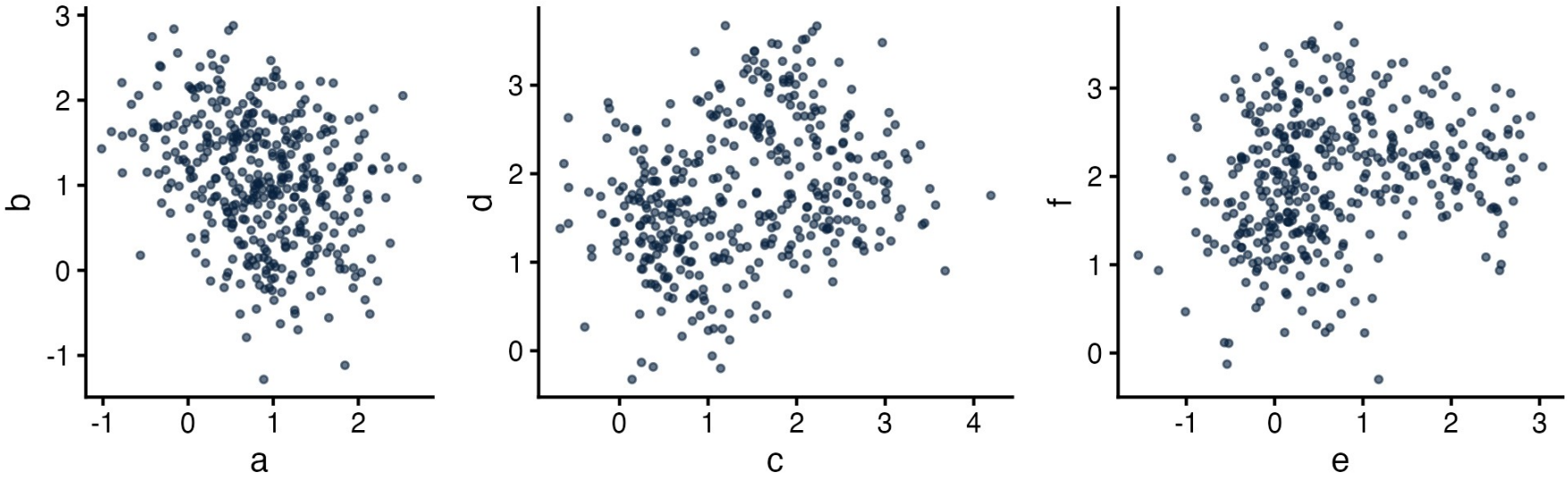


Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w8-elbow](greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w8-elbow)

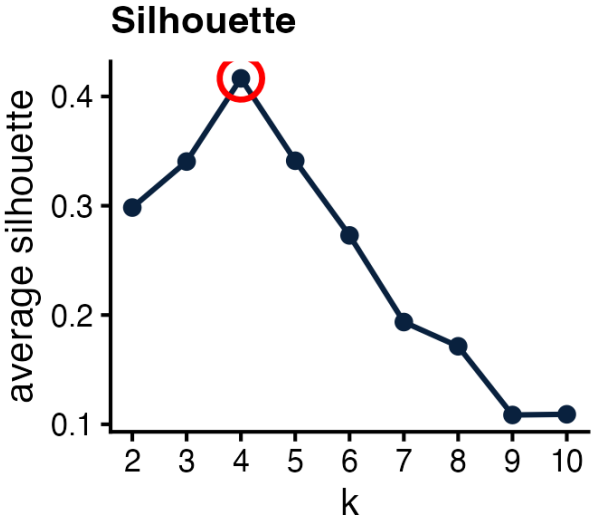
[host screen]

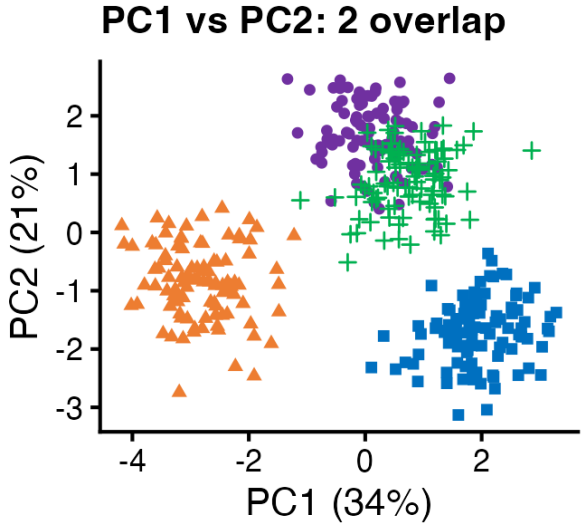
10 variables: no clusters in any pair

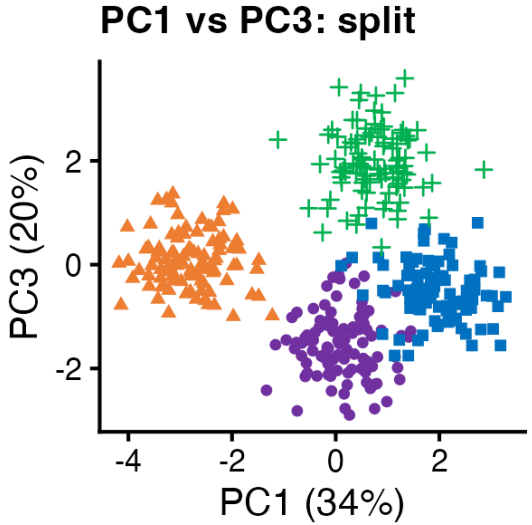


Three of the 45 scatterplots. **No groups visible.**

k-means finds 4; PCA needs two views to show them







Colour = cluster, shape = true group: **100% match**. k-means used all 10 variables at once.

Choosing k: no single right answer

Elbow where the SS stops falling fast

Silhouette highest average

Knowledge what you know about the system

Purpose how many groups you can use

1 Why cluster

2 The algorithm

3 Many starts

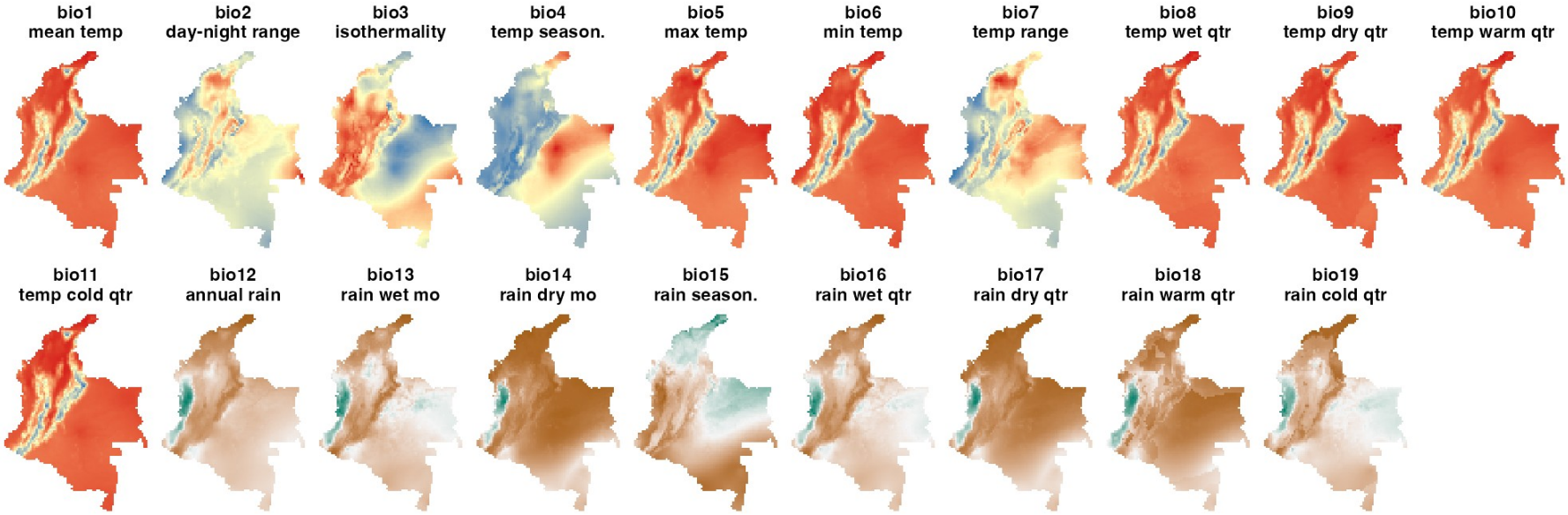
4 Distance & scale

5 Choosing k

▶ 6 Colombia



19 climate variables, 3571 places



WorldClim 2.1, cells about 18 km across. **Which places share a climate?**

Live poll

Multiple choice

k-means with k = 2 on Colombia's 19 scaled climate variables. What will the split be?

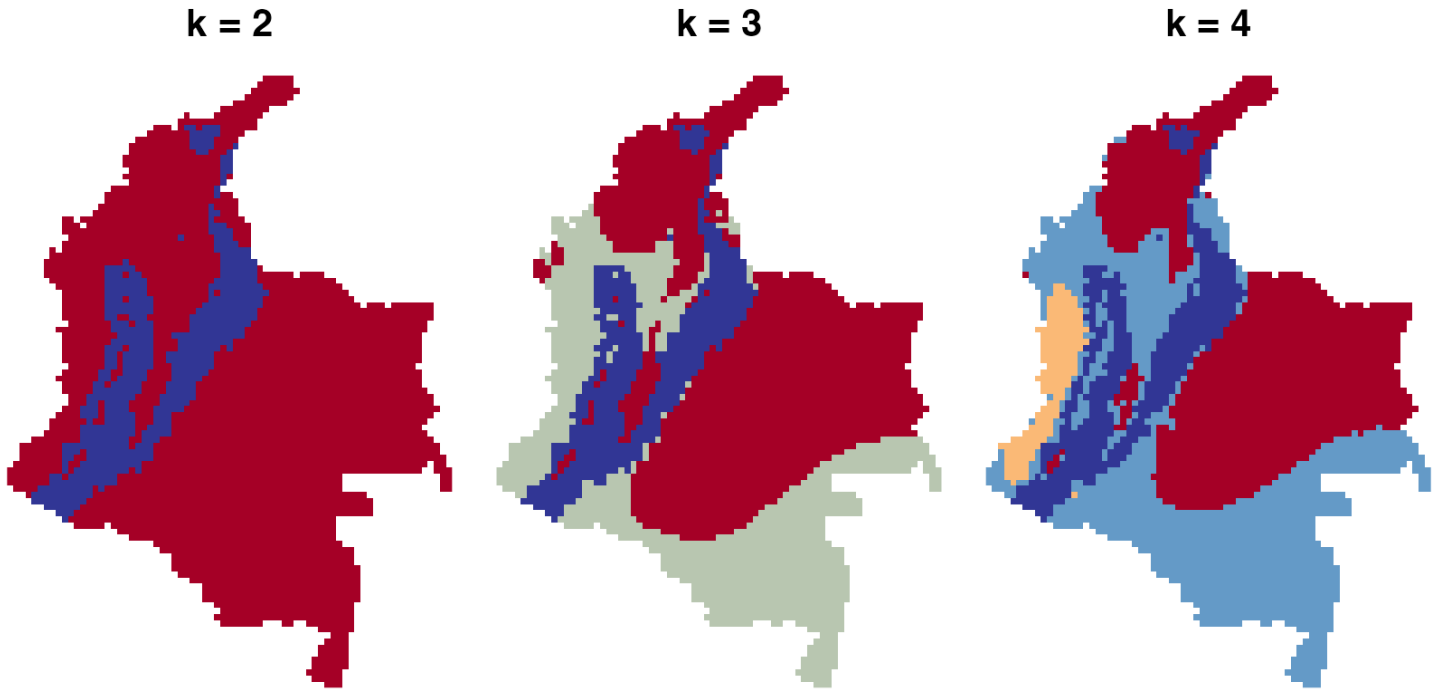
A North vs south
B East vs west
C Cool vs warm
D Wet vs dry

Scan with your phone

Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w8-colombia-k2

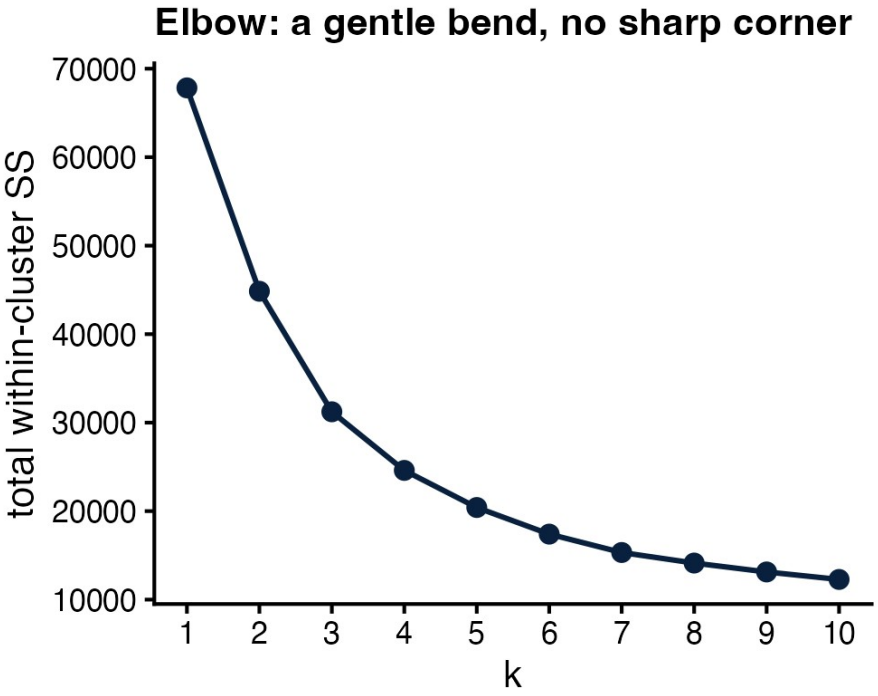
host screen

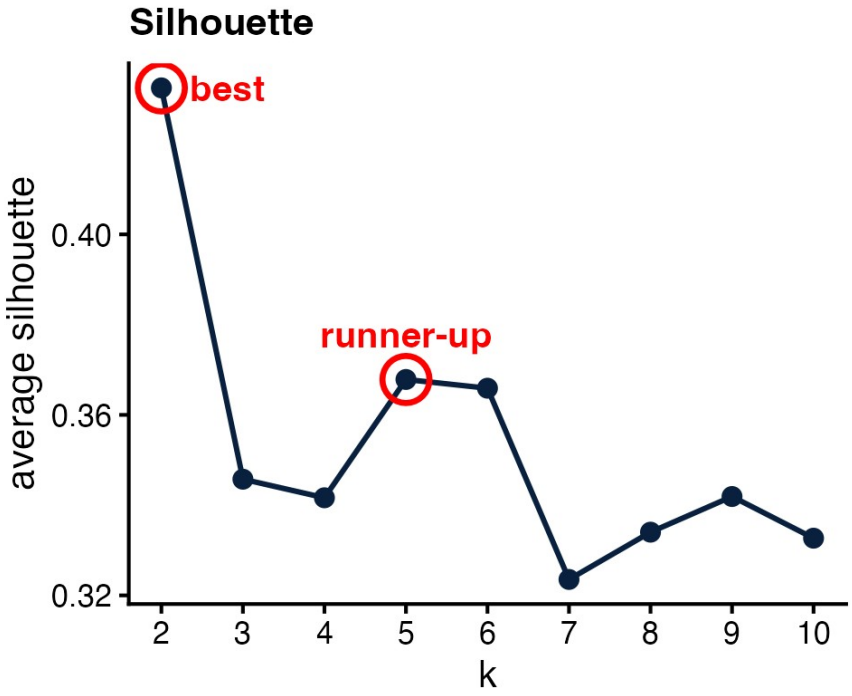
k = 2, 3, 4



Colours run **coldest (blue)** to **warmest (red)**. Each map adds a split.

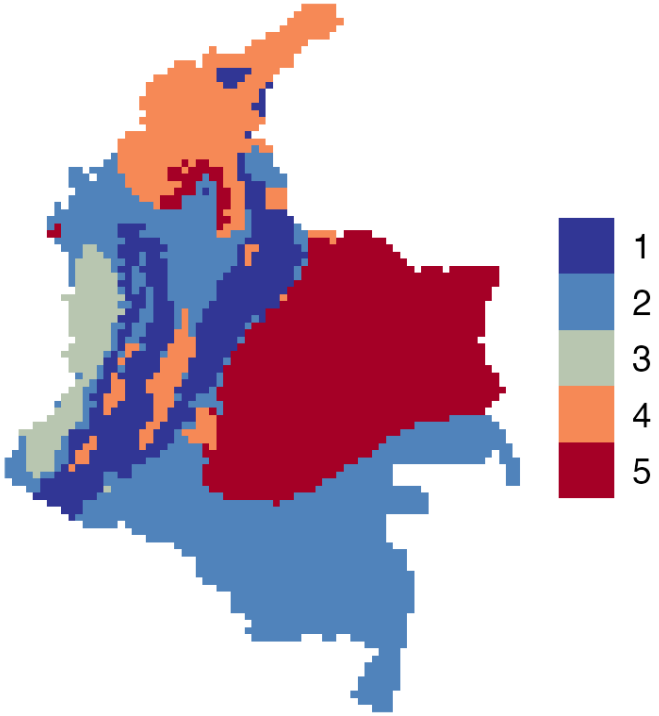
How many climate types?

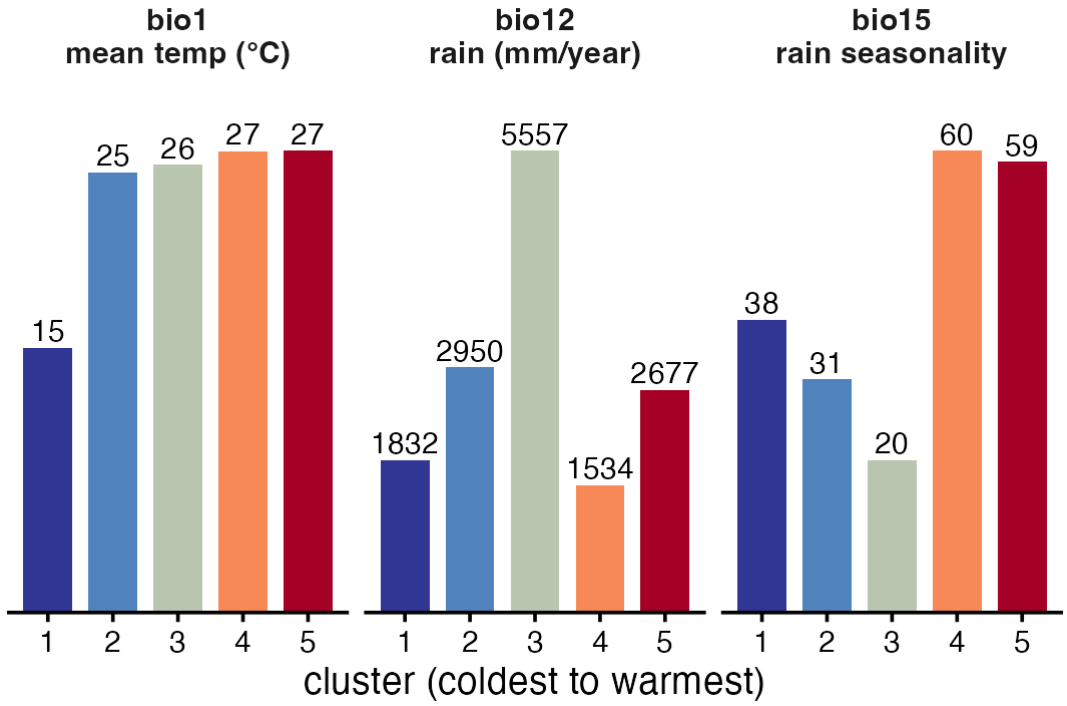




Best **k = 2**: highlands vs lowlands. Runner-up **k = 5** tells us more: **purpose decides**

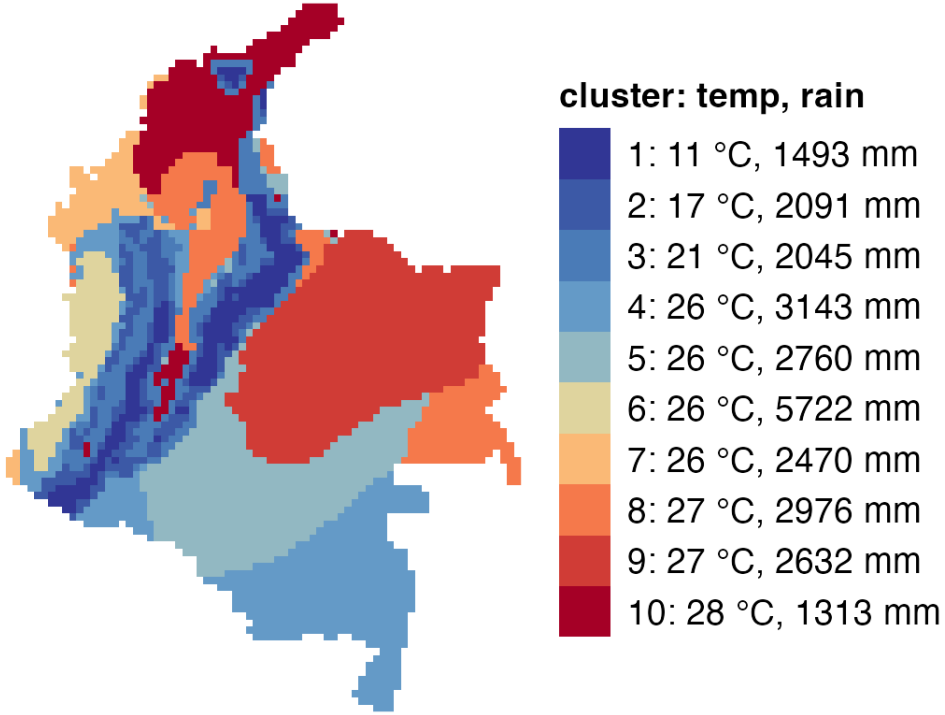
k = 5: what each cluster is





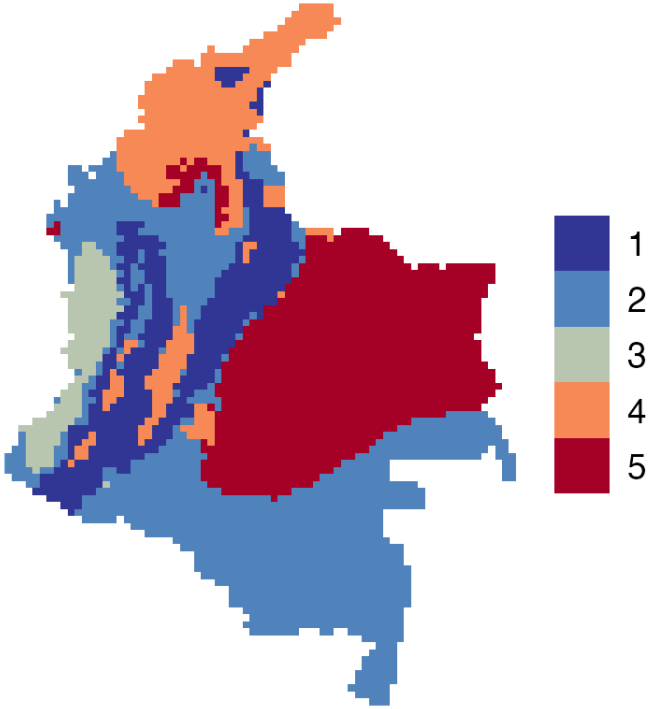
Name clusters by their **means**, not their numbers

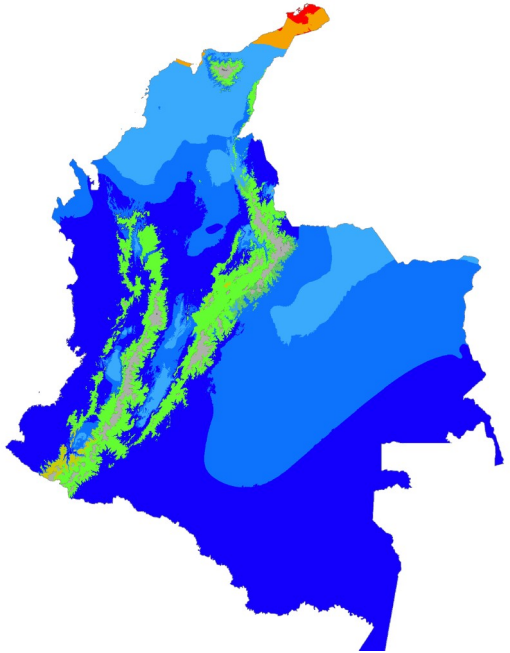
k = 10, coldest to warmest



More k = **finer climate zones**, from 11 to 28 °C

k-means vs the Köppen climate map

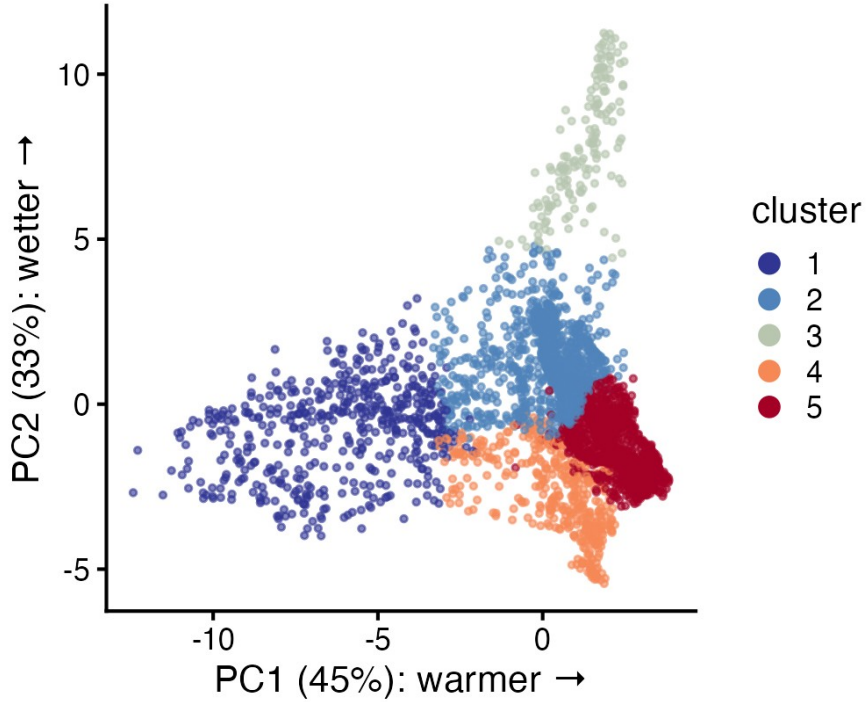






Köppen: hand-drawn thresholds. k-means: **19 numbers, no labels, similar zones**

The clusters on the PCA of the same data



PC1 ≈ temperature
r = 0.96 with bio1

PC2 ≈ rainfall
r = 0.87 with bio12

k-means in practice

1. scale() the variables
2. `kmeans(x, centers = k, nstart = 25)`
3. Elbow + silhouette over a range of k
4. Plot the clusters: a map, a PCA
5. Describe each cluster by its means

Live coding: k-means



`Week8_live_1_kmeans.R`

Simulations, then Colombia

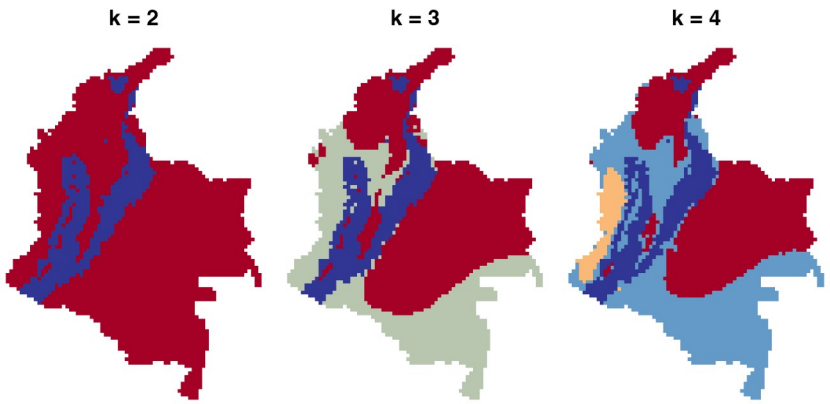
kmeans(nstart = 25)

scale() and table()

Elbow and silhouette

Colombia: maps and cluster means

Wednesday: hierarchical clustering



k-means: choose k up front

Hierarchical: build the whole tree, cut it anywhere