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

- K-means Clustering

Supervised Learning (Classification Algorithm)


Duck


Duck

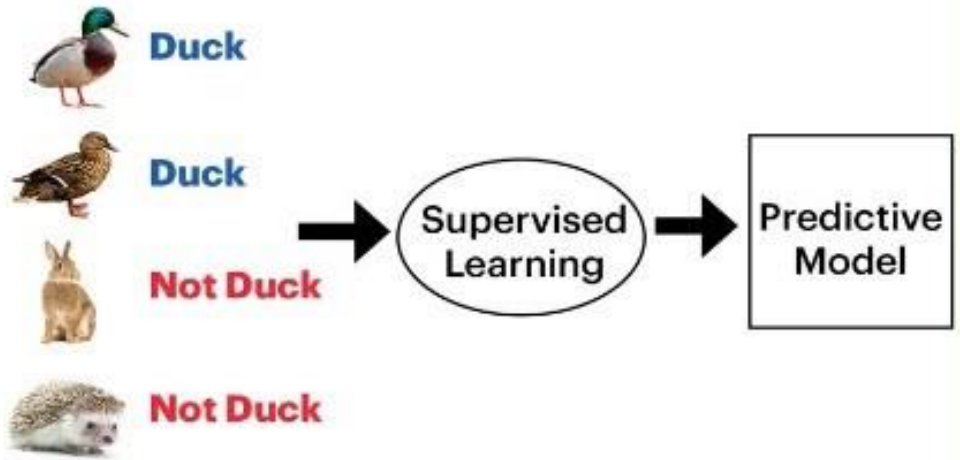

Not Duck

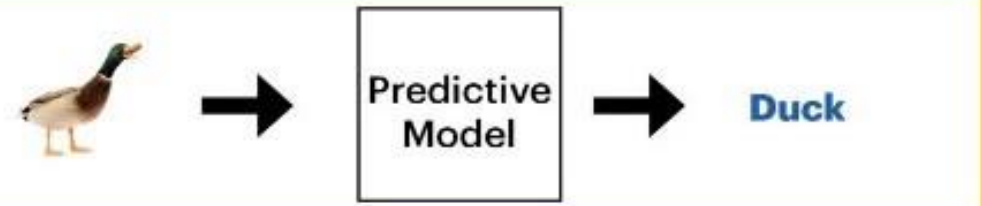

Not Duck

Supervised Learning → Predictive Model

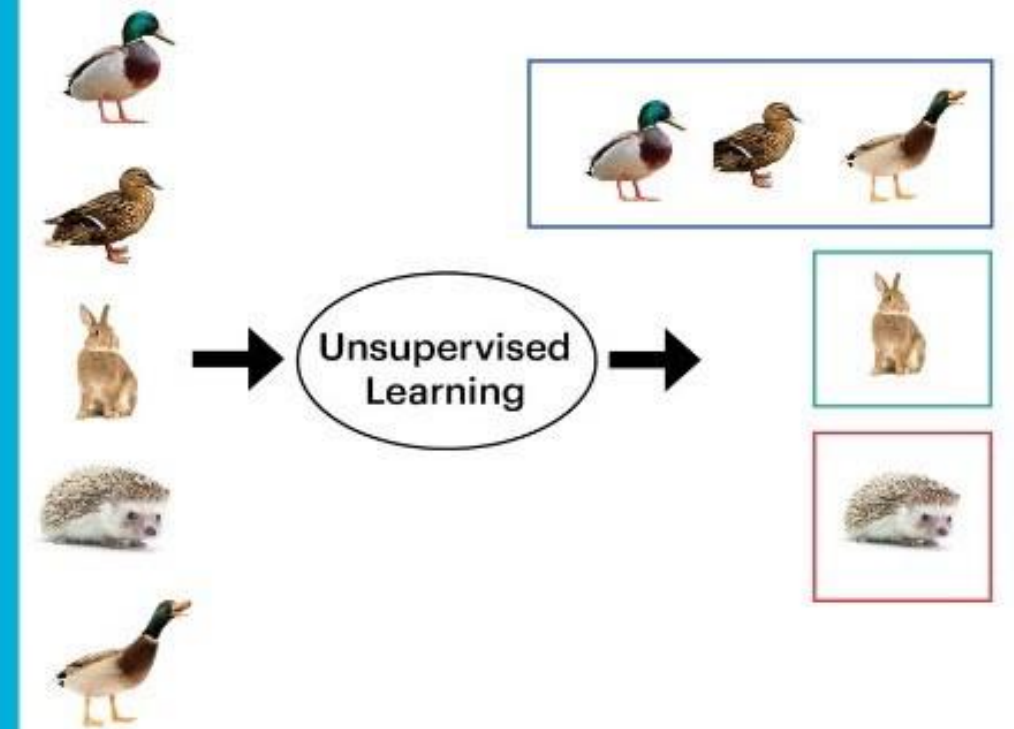

→ Predictive Model → Duck

Supervised Learning (Classification Algorithm)

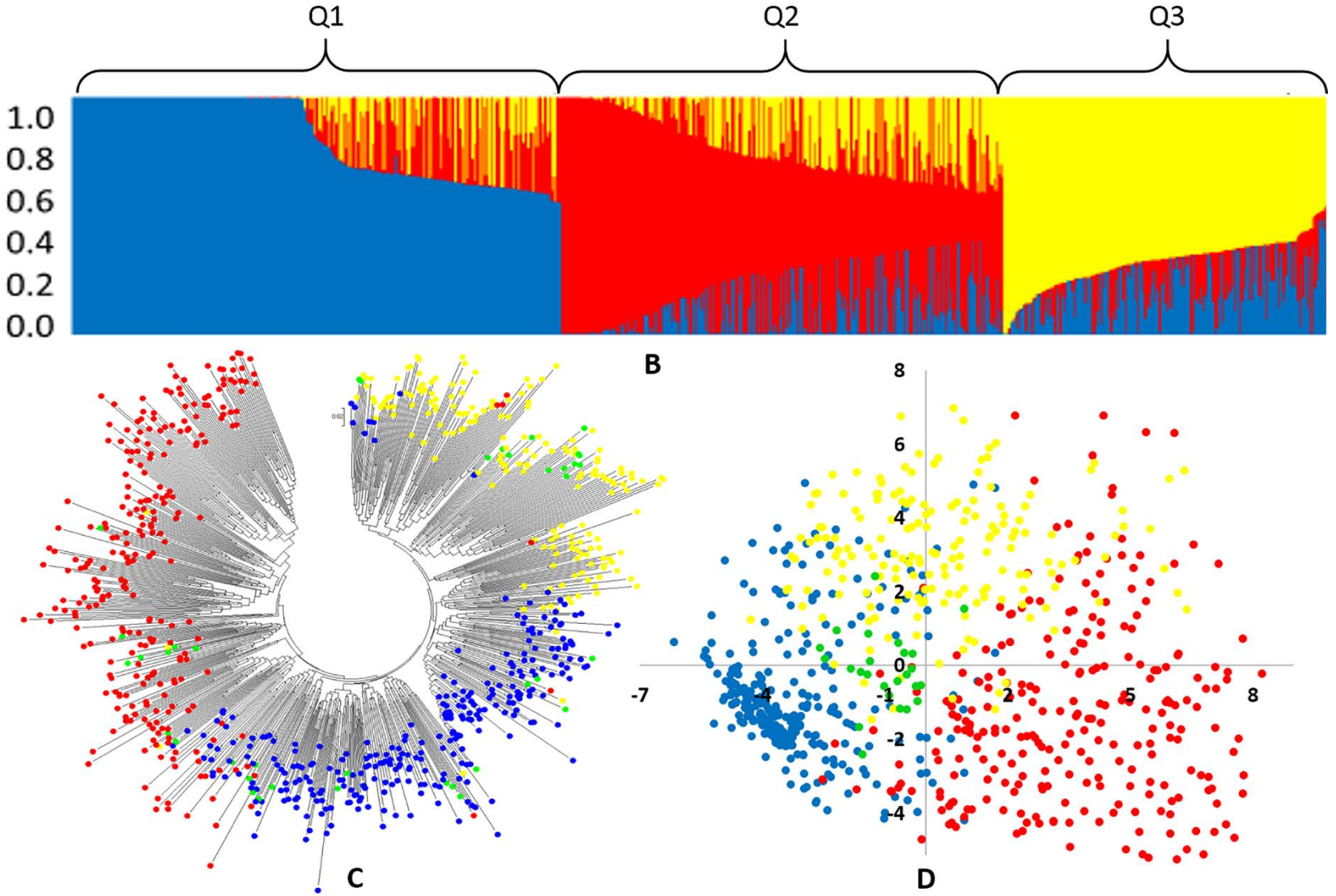




Unsupervised Learning (Clustering Algorithm)

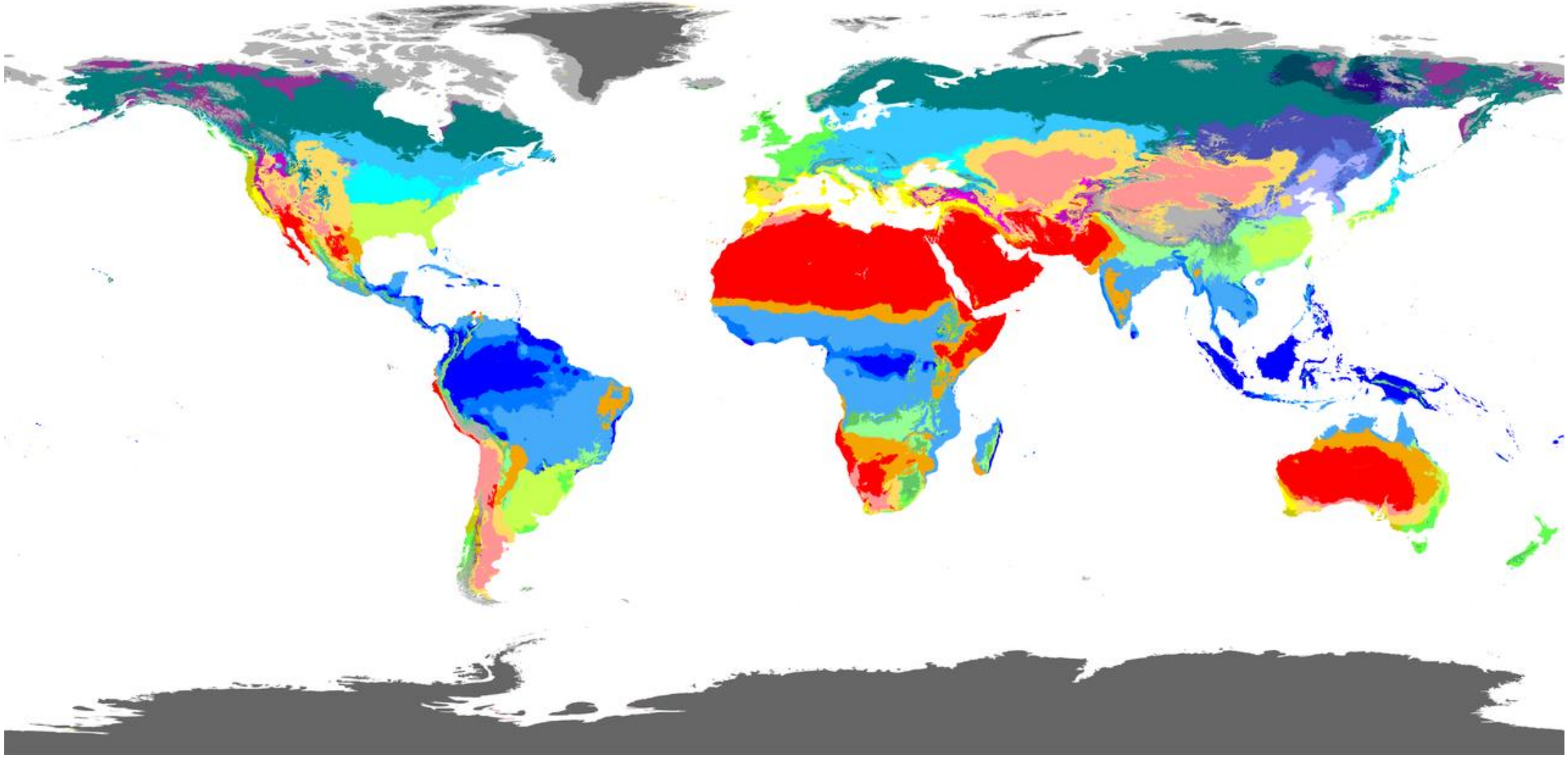


Clustering genotypes into populations by allele data



Xiong et al 2016

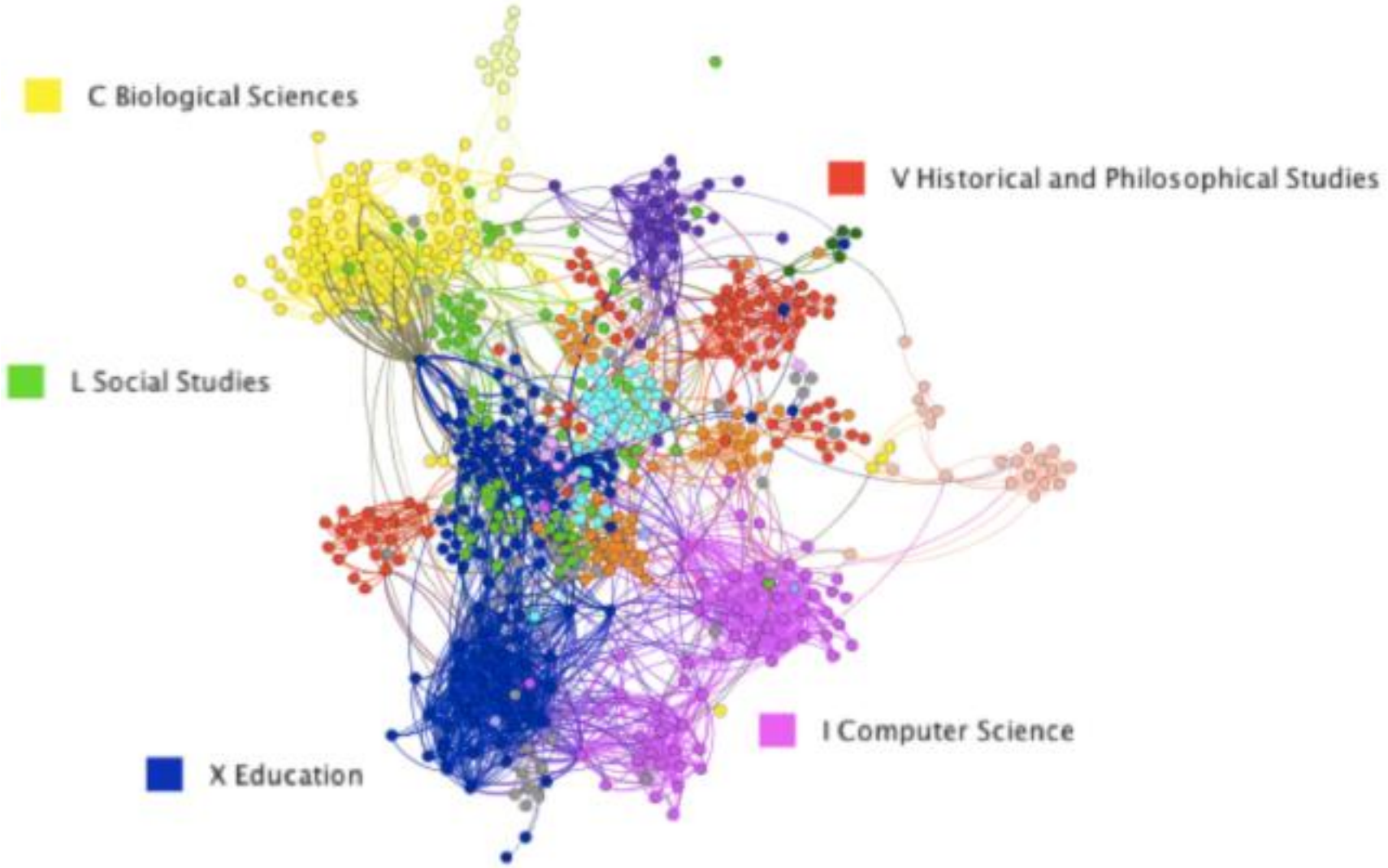
Clustering regions into categories by climate data



Wikipedia commons

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

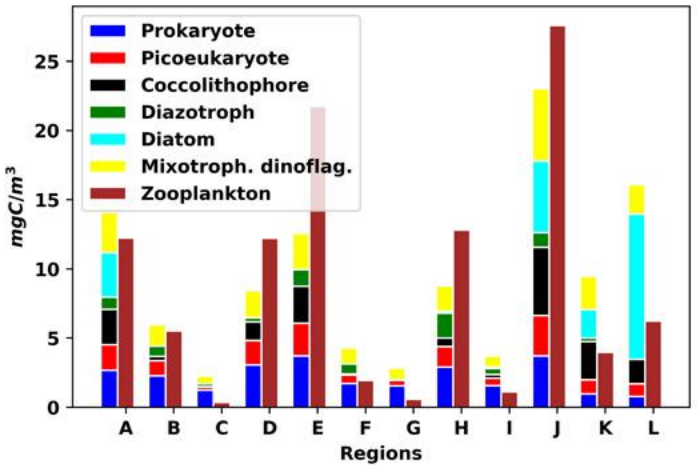
Elucidating ecological complexity: Unsupervised learning determines global marine eco-provinces

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

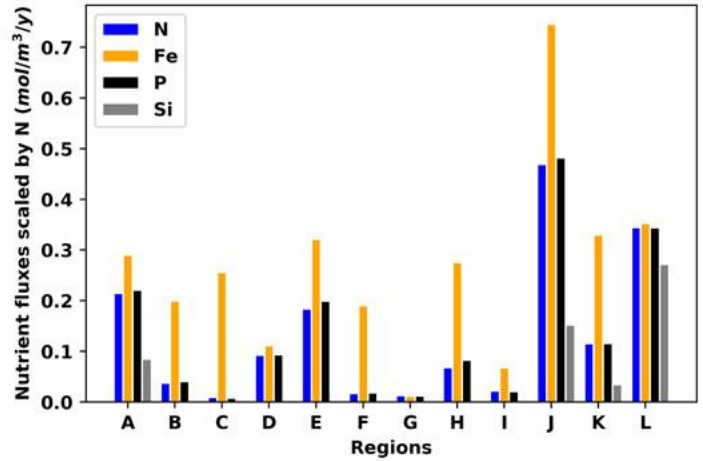
SCIENCE ADVANCES • 29 May 2020 • Vol 6, Issue 22 • DOI: 10.1126/sciadv.aay4740

2,268 27

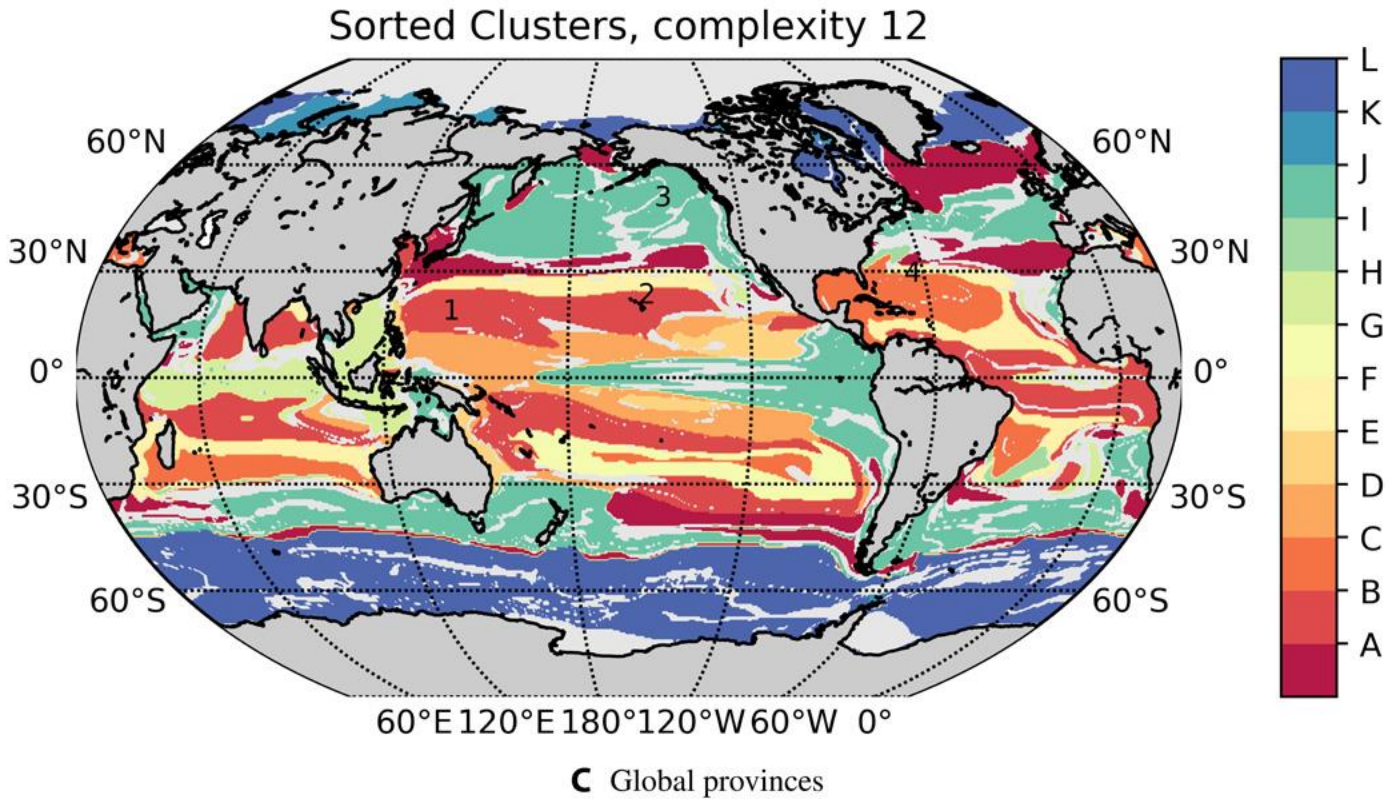
"improve understanding and monitoring of marine ecosystems"



A Ecological ensemble



B Nutrient fluxes



C Global provinces

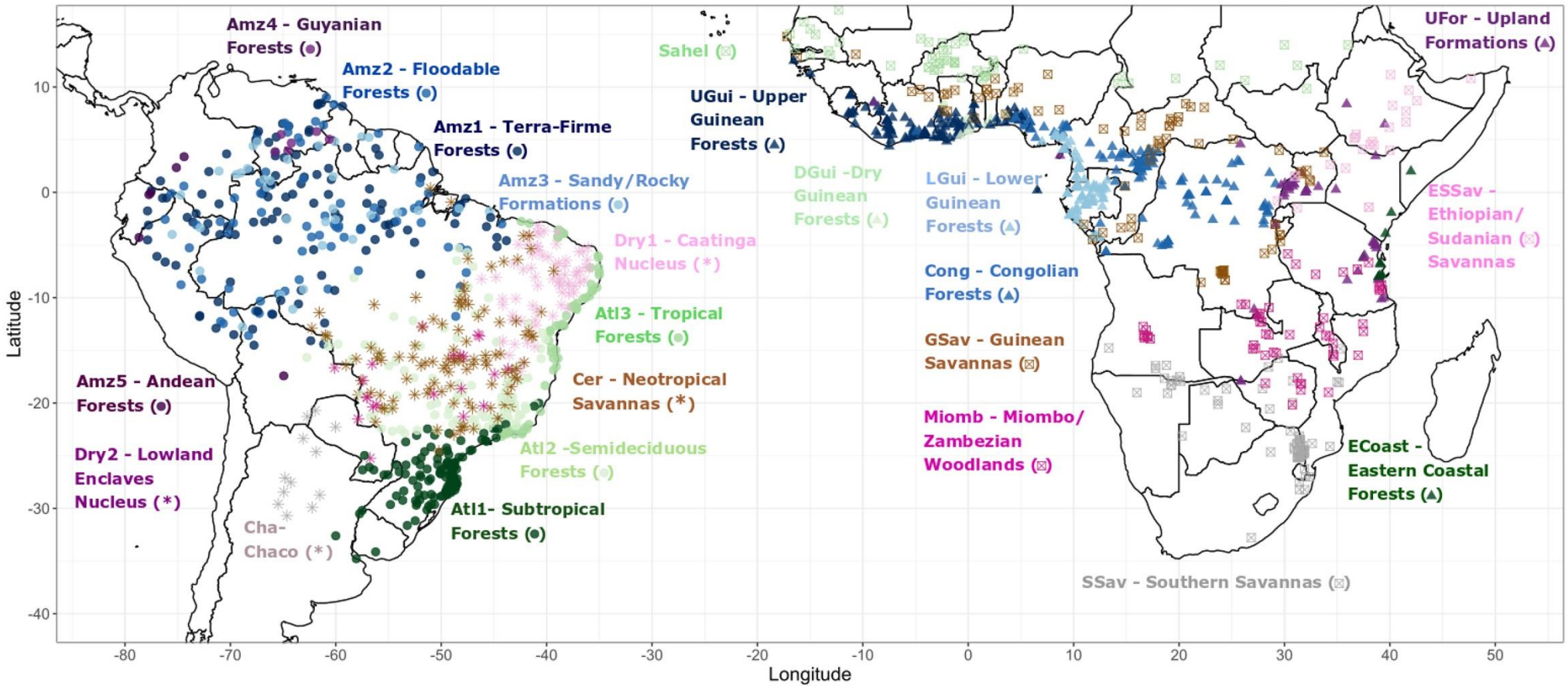
RESEARCH ARTICLE | ECOLOGY |

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

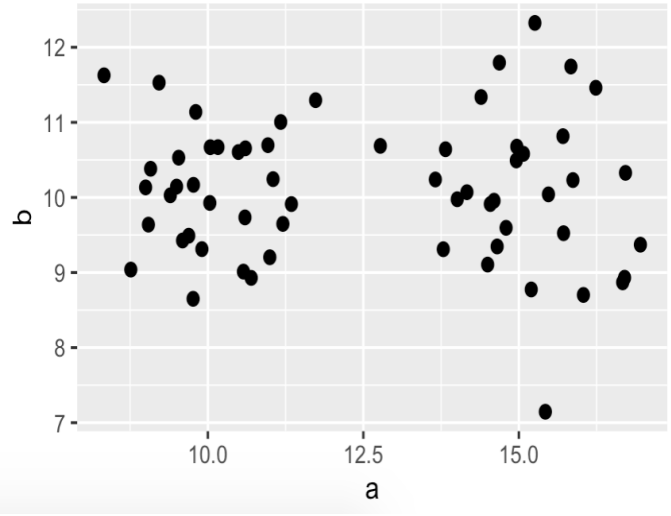
Edited by Douglas Schemske, Michigan State University, East Lansing, MI; received July 4, 2021; accepted February 17, 2022

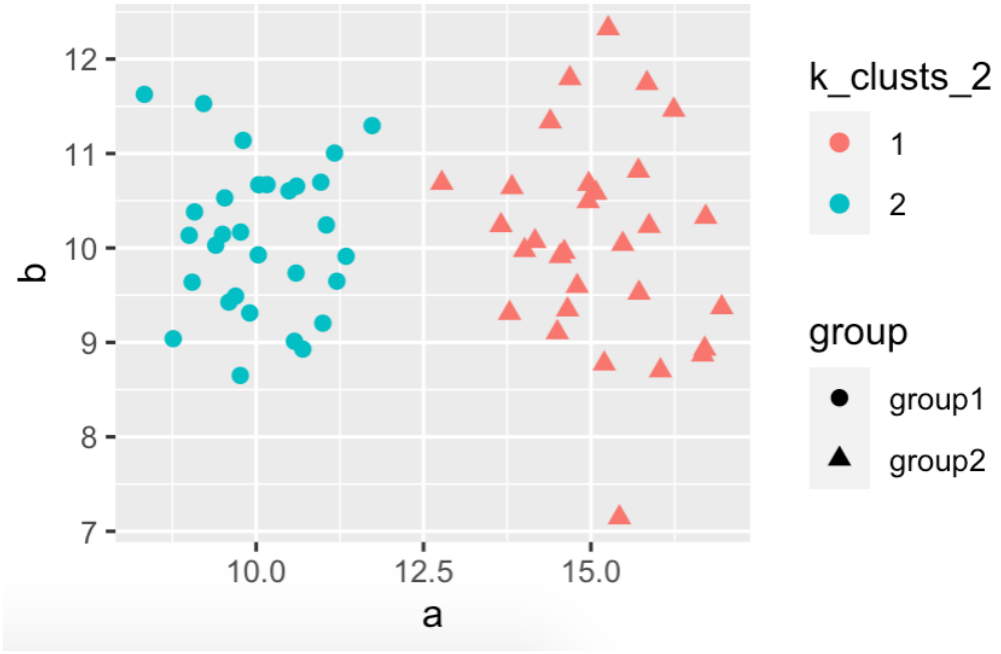
March 29, 2022 | 119 (14) e2112336119 | https://doi.org/10.1073/pnas.2112336119

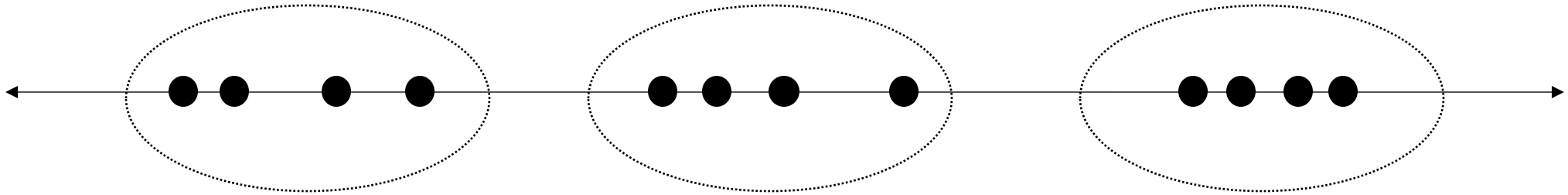


Today

- K-means clustering
  - Intuitive clustering method
  - Tries to minimize the variance within clusters
  - Requires user set number of clusters (k)
  - Not-deterministic
  - Makes clusters even if they don't exist
  - Choosing k
  - K-means to identify hidden clusters
  - K-means to classify climates

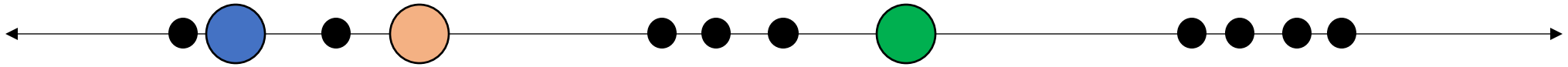


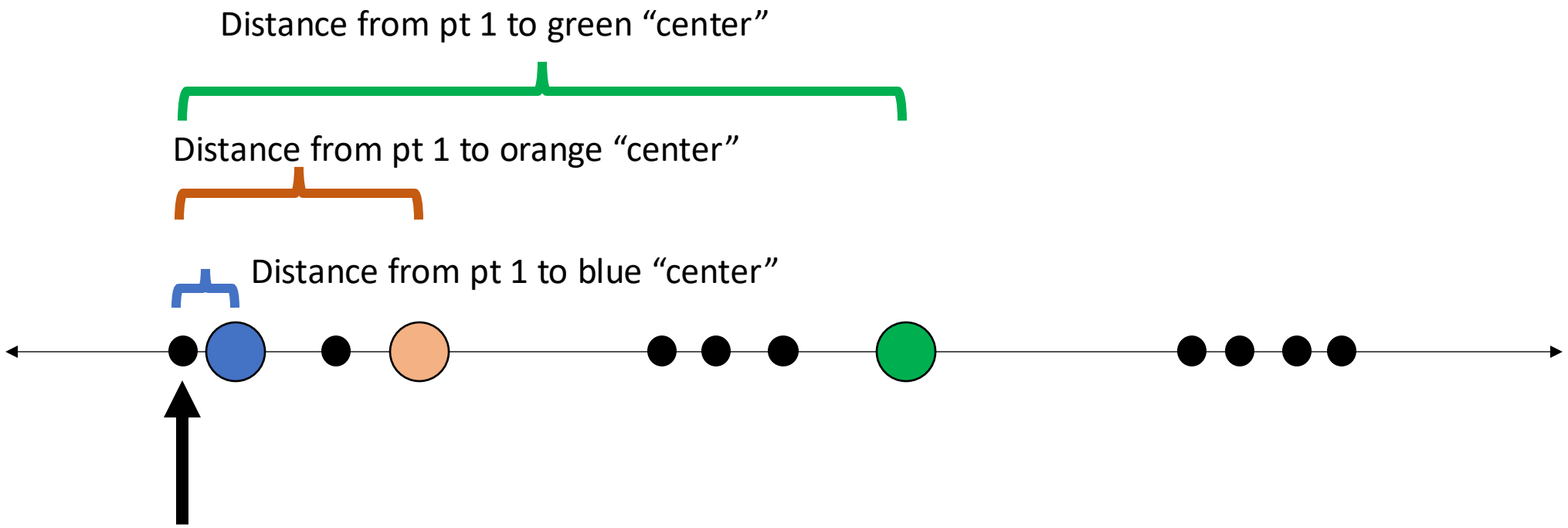




(K = 3)

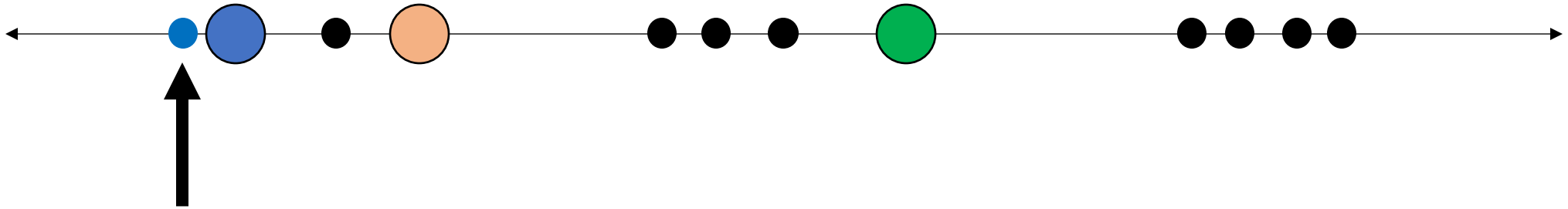
1. Randomly pick 3 points to start as “centers” for clusters



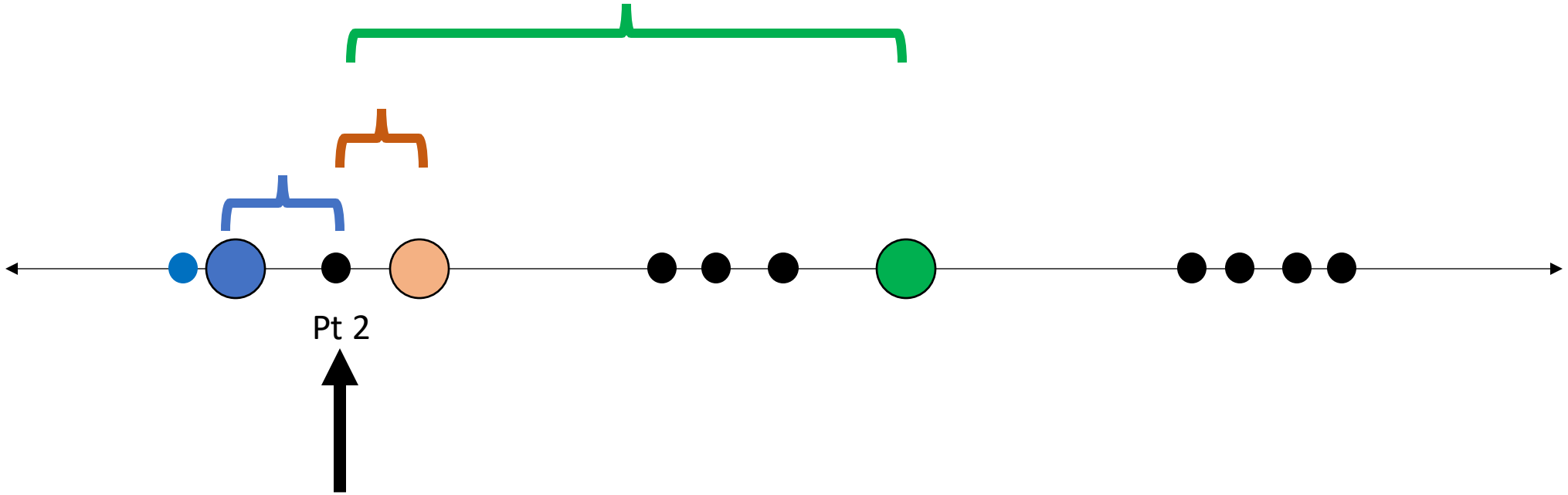


Calculate distance from each point to the centers

Pt 1 is assigned to ”blue” since it is the closest

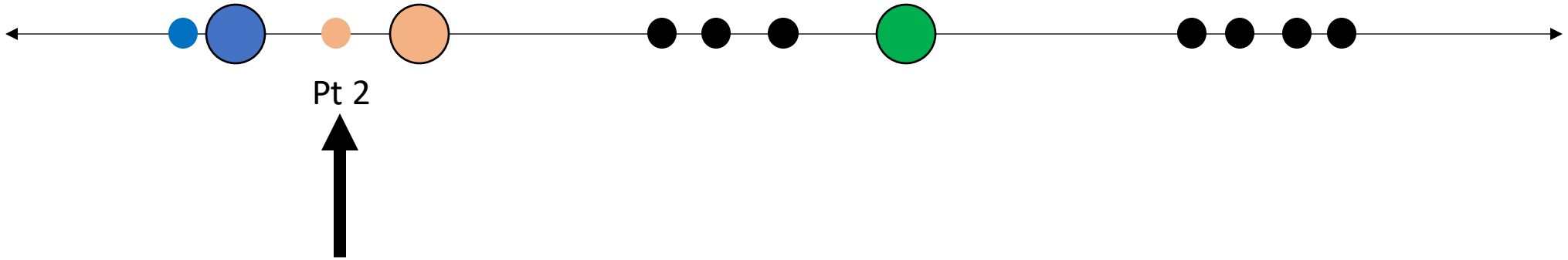


Assign points to closest center

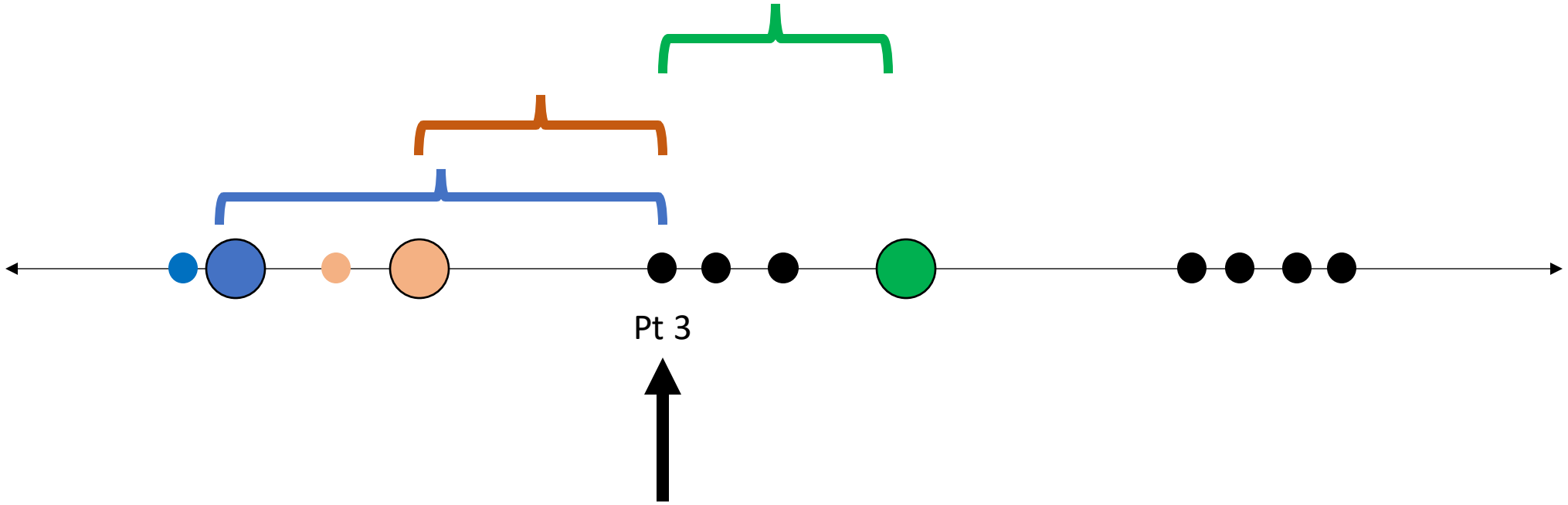


Assign points to closest center

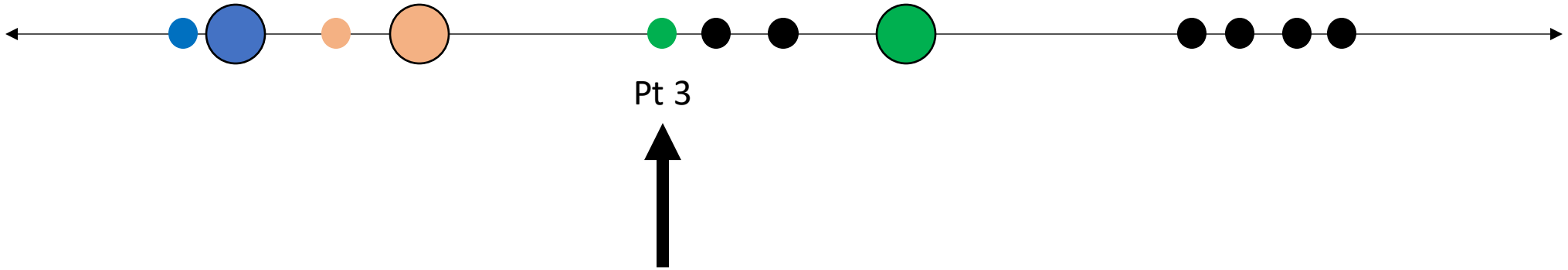
Pt 2 is assigned to ”orange” since it is the closest



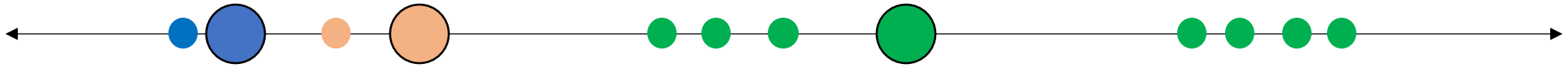
Assign points to closest center



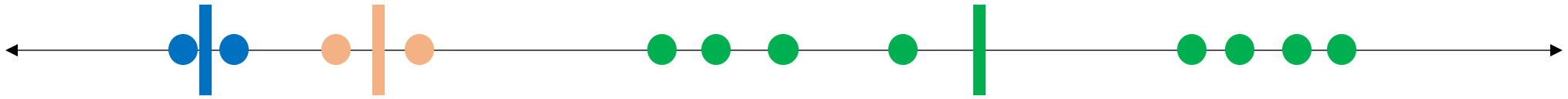
Assign points to closest center



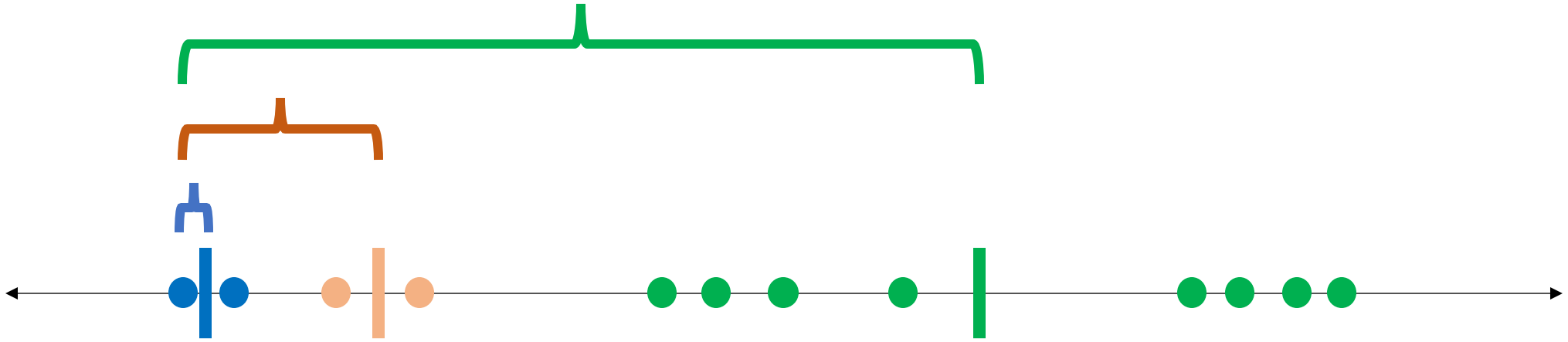
Assign points to closest center



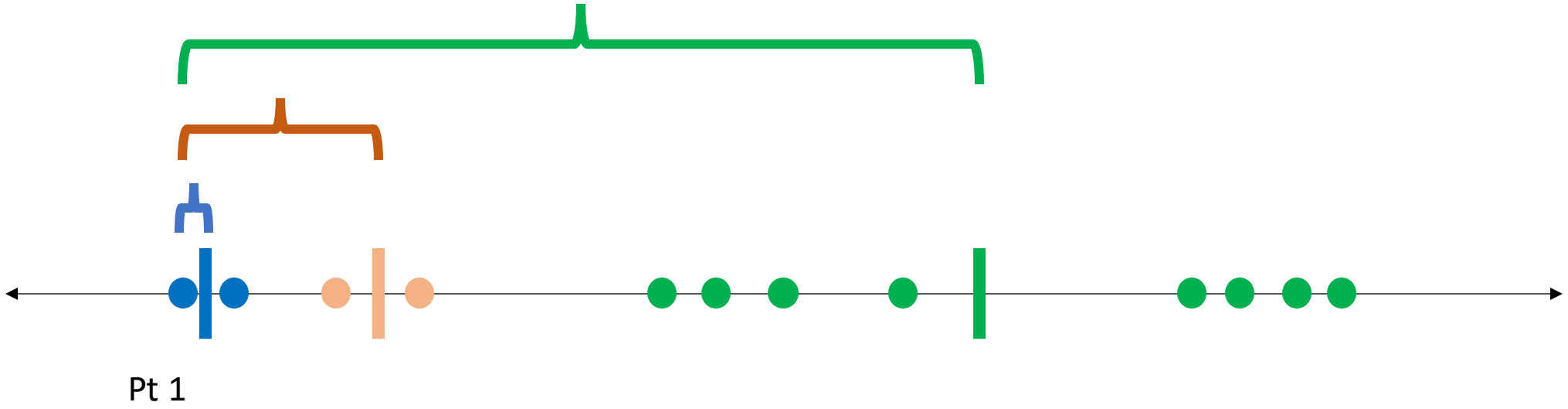
Assign points to closest center



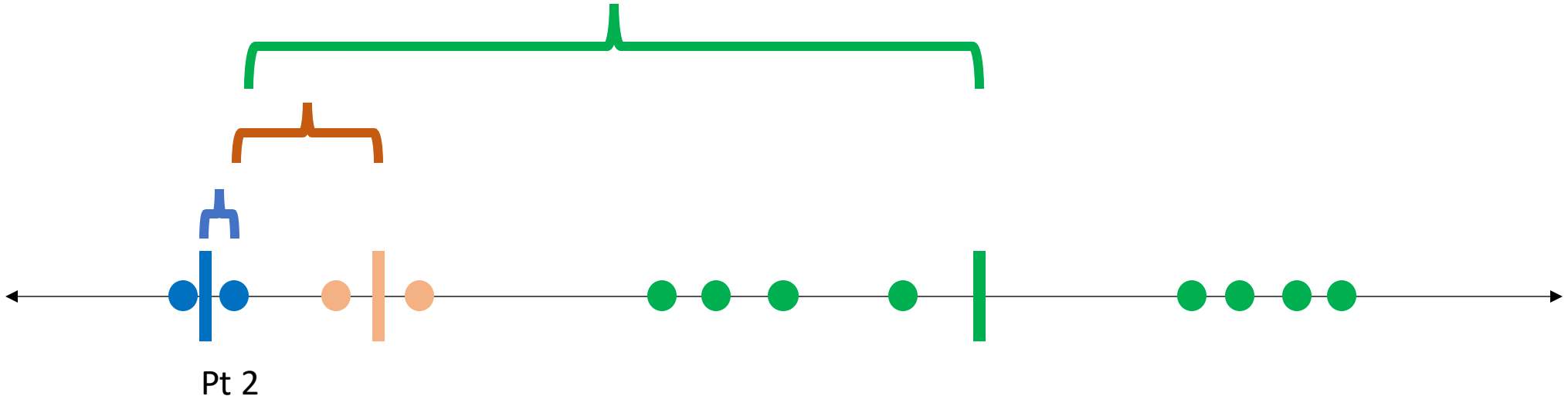
Calculate mean of each cluster



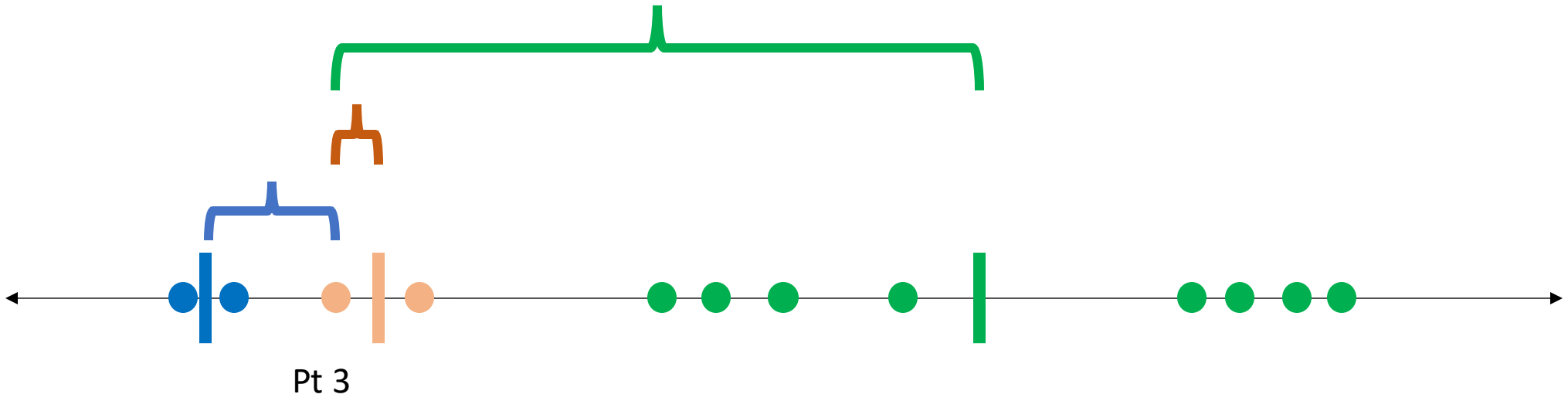
Calculate distance of each point to cluster means and assign to closest



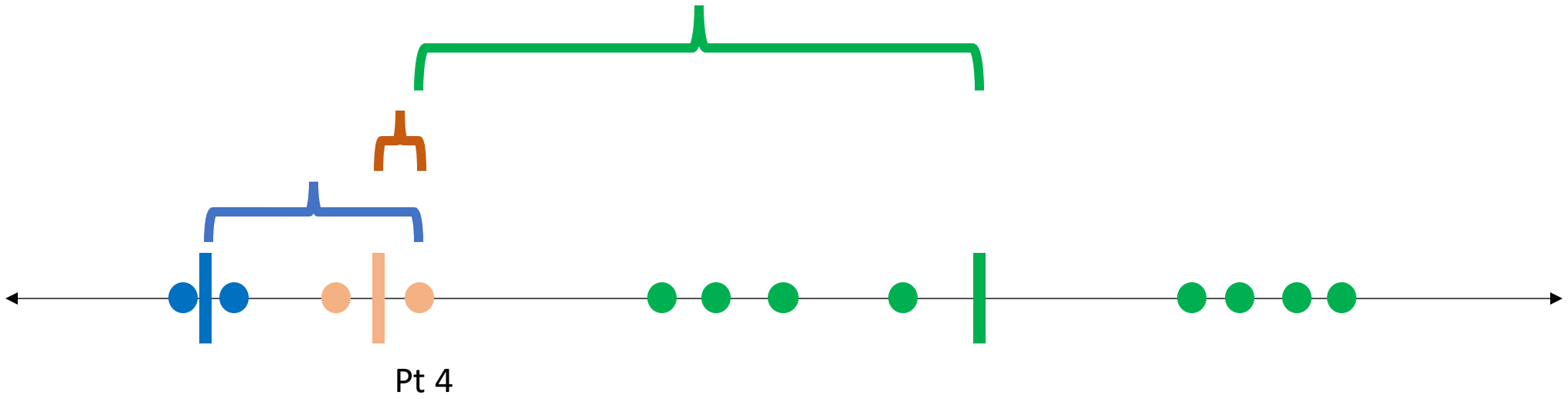
Calculate distance of each point to cluster means and assign to closest



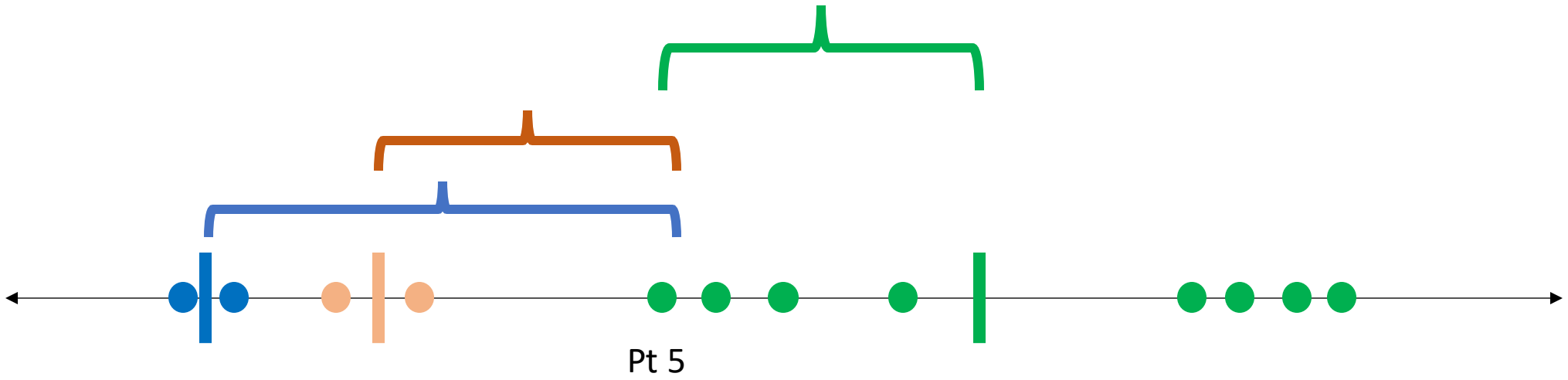
Calculate distance of each point to cluster means and assign to closest



Calculate distance of each point to cluster means and assign to closest

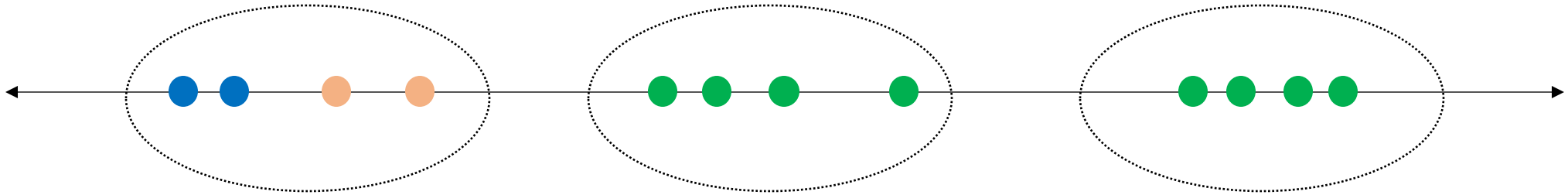


Calculate distance of each point to cluster means and assign to closest

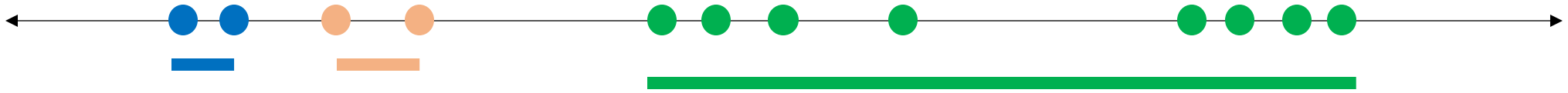


Calculate distance of each point to cluster means and assign to closest

The points did not change so we stop: these are our clusters



(Don’t worry yet)

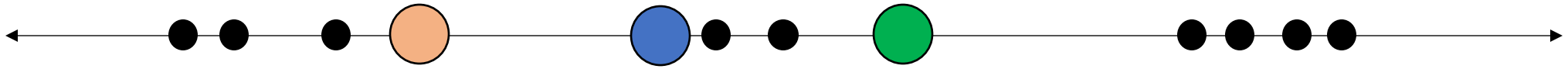


Variance within each cluster calculated

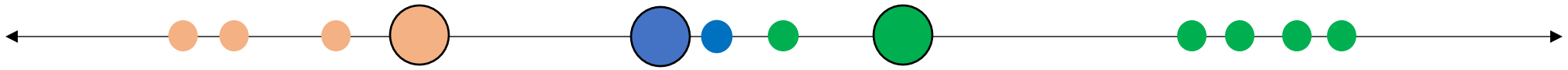
Total variance of clustering results from first run:

(K = 3)

1. Randomly pick 3 new points to start as “centers” for clusters

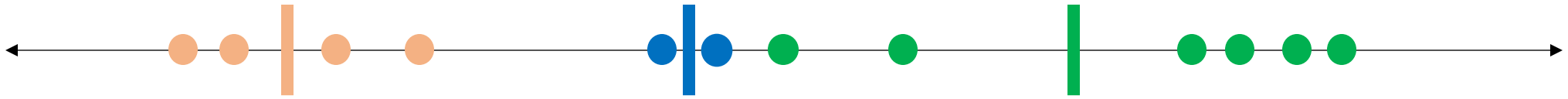


Total variance of clustering results from first run:



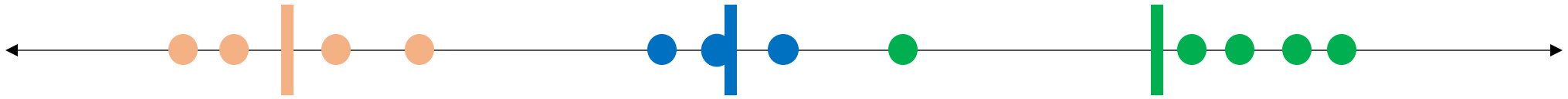
Assign points to closest center

Total variance of clustering results from first run:



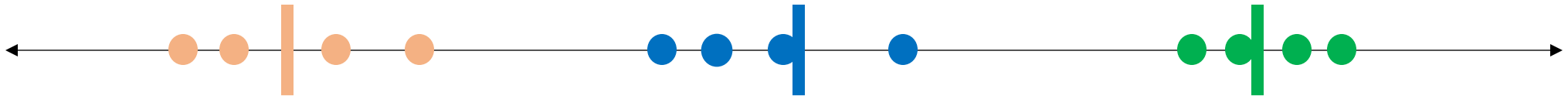
Calculate mean of each cluster

Total variance of clustering results from first run:



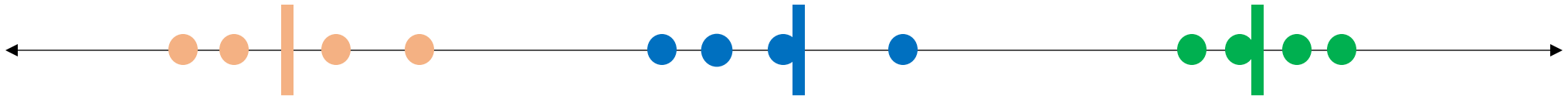
Assign points to closest center, calculate new mean…

Total variance of clustering results from first run:



Assign points to closest center, calculate new mean…

Total variance of clustering results from first run:



Assign points to closest center… doesn’t change so these are our clusters

Total variance of clustering results from first run:

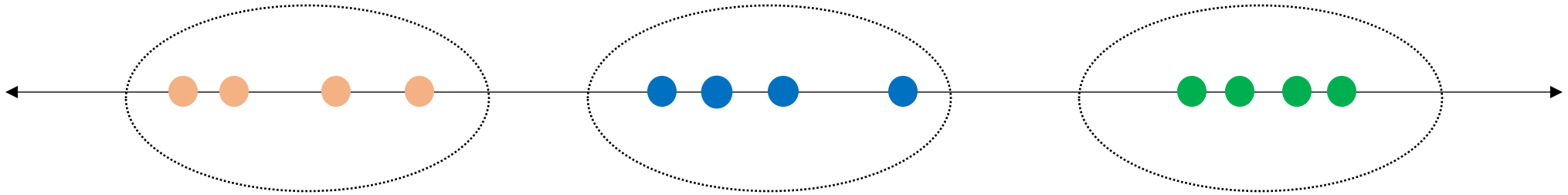
Calculate total within cluster variance

Total variance of clustering results from first run:

Compare that to other runs…

Total variance of clustering results from second run:

Total variance of clustering results from first run:



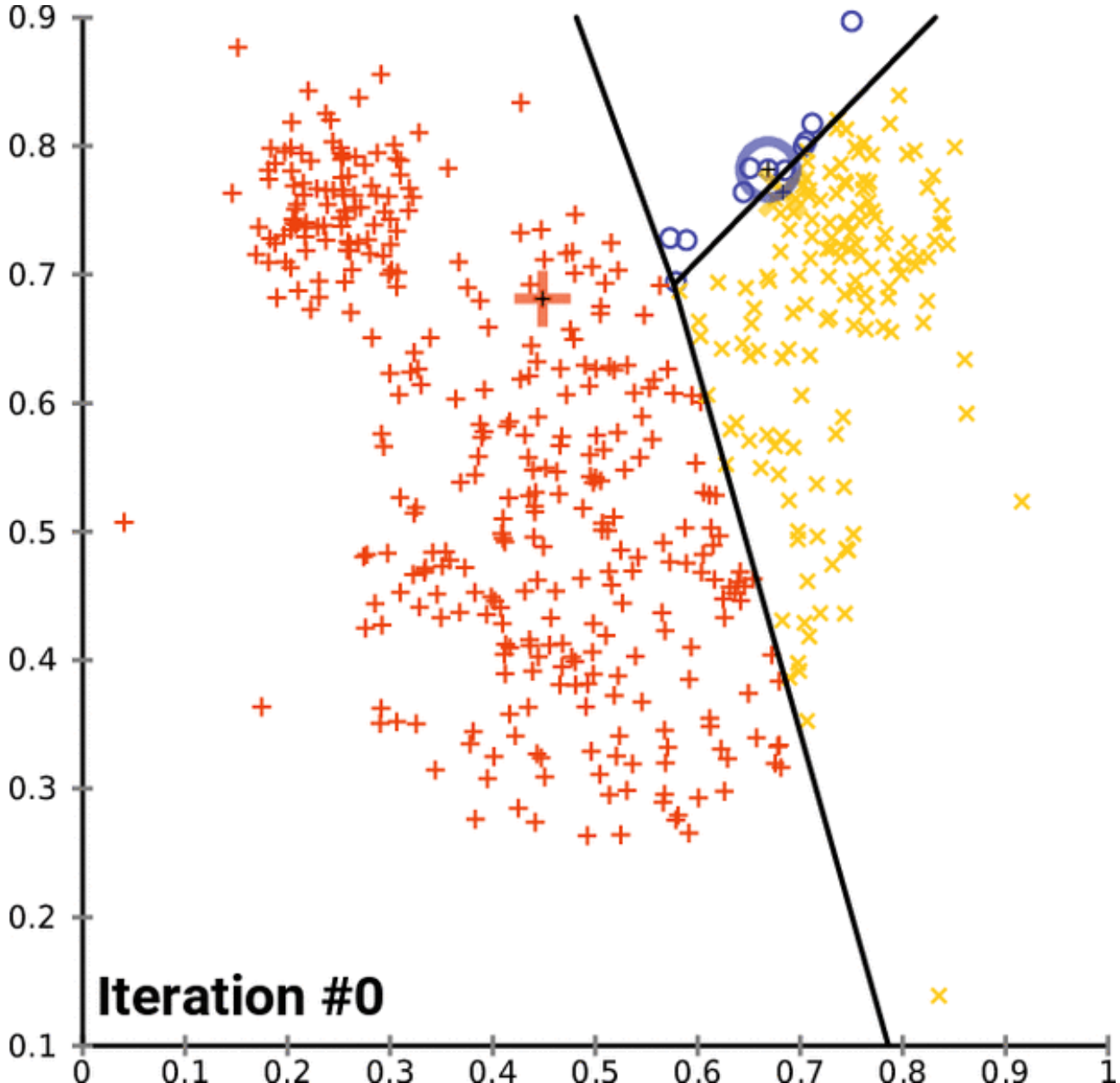
Compare that to other runs…

Total variance of clustering results from second run:

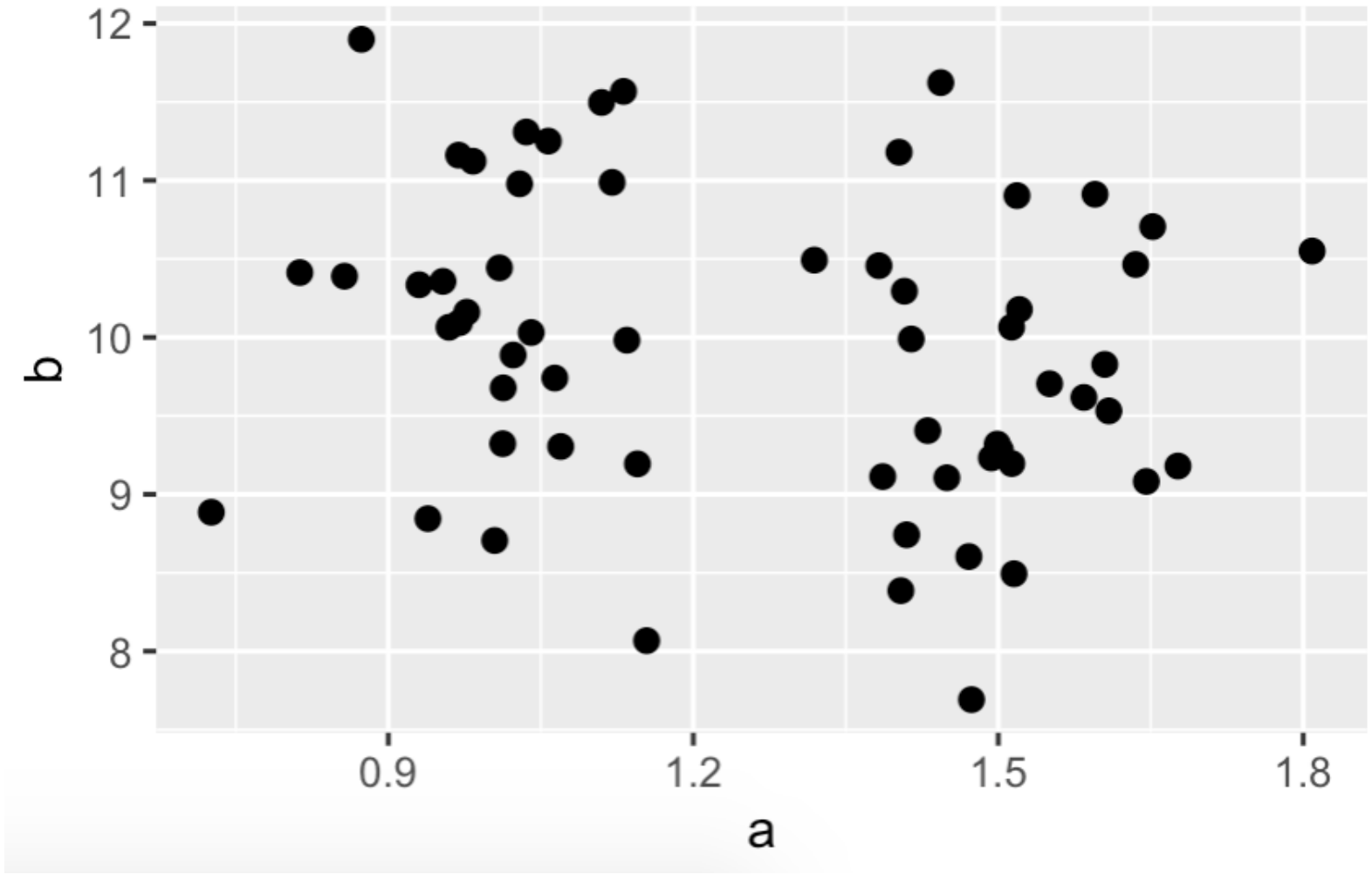
Total variance of clustering results from first run:

Repeat process any number of times. Final clusters are whichever has the smallest in cluster variance

Underlying algorithm for K-means clustering is same for multivariate data

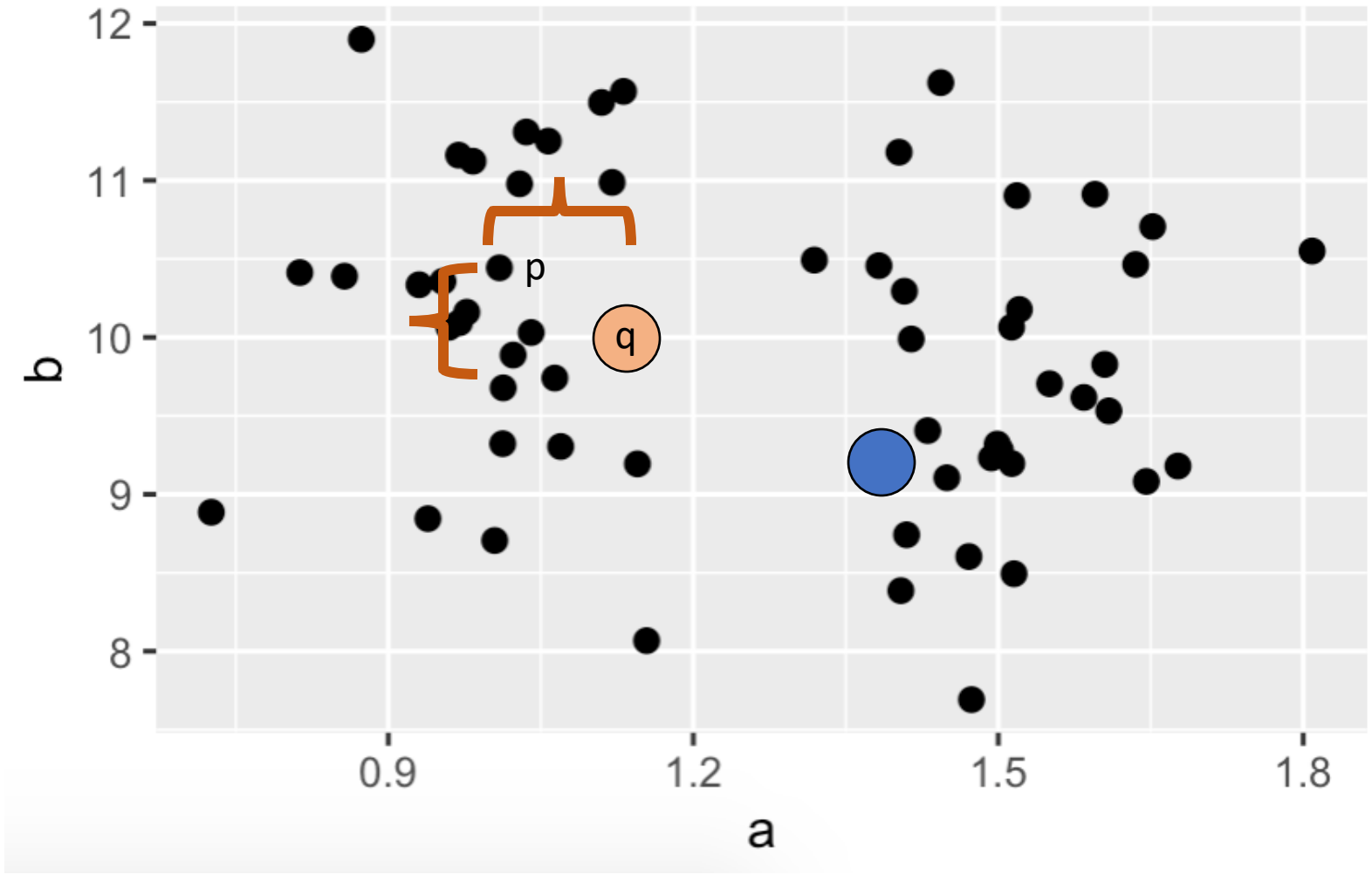


Example



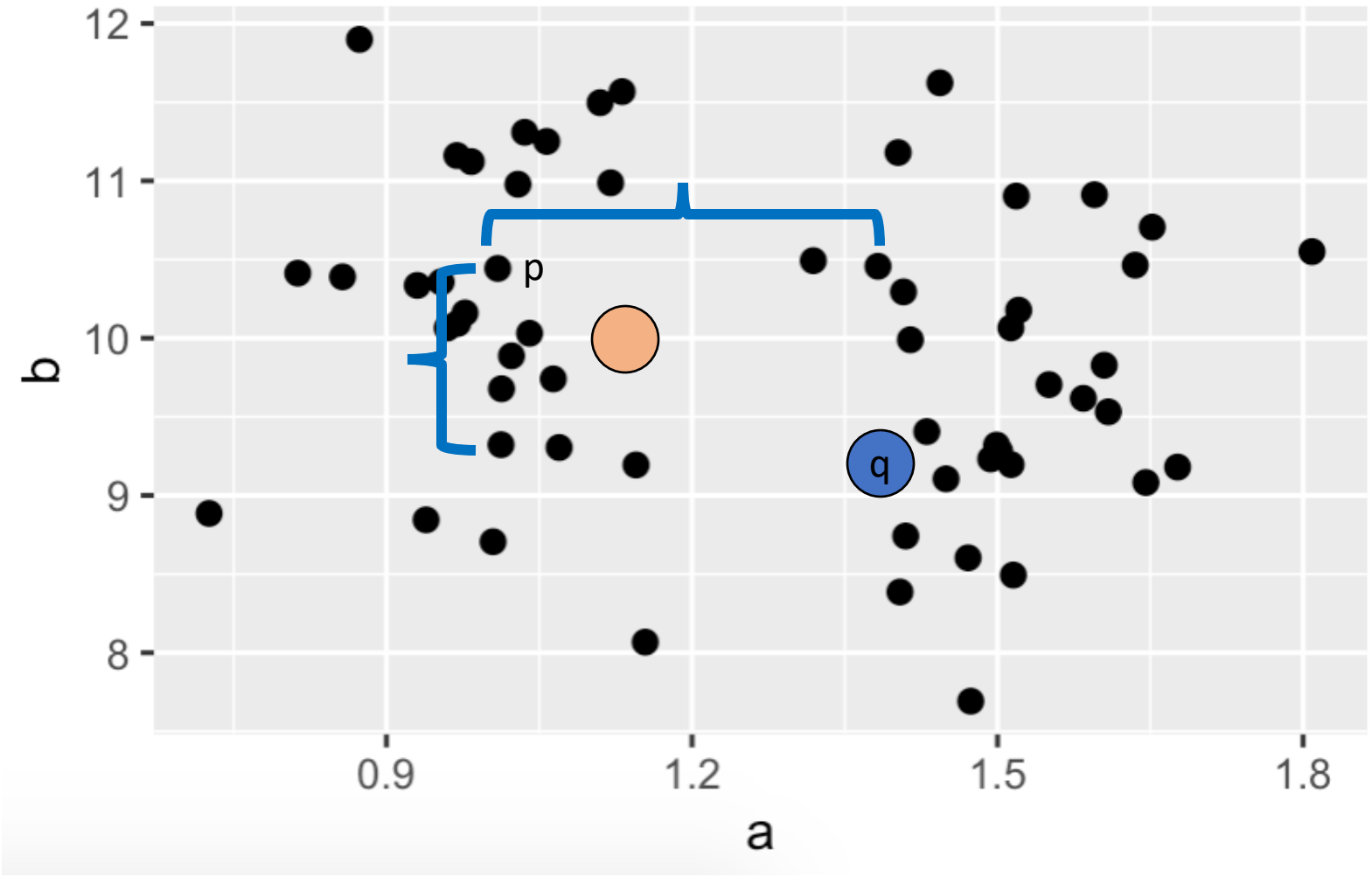
Same principles apply to multivariate data but with Euclidean distance

$$d(p,q) = \sqrt{(p-q)^2}$$

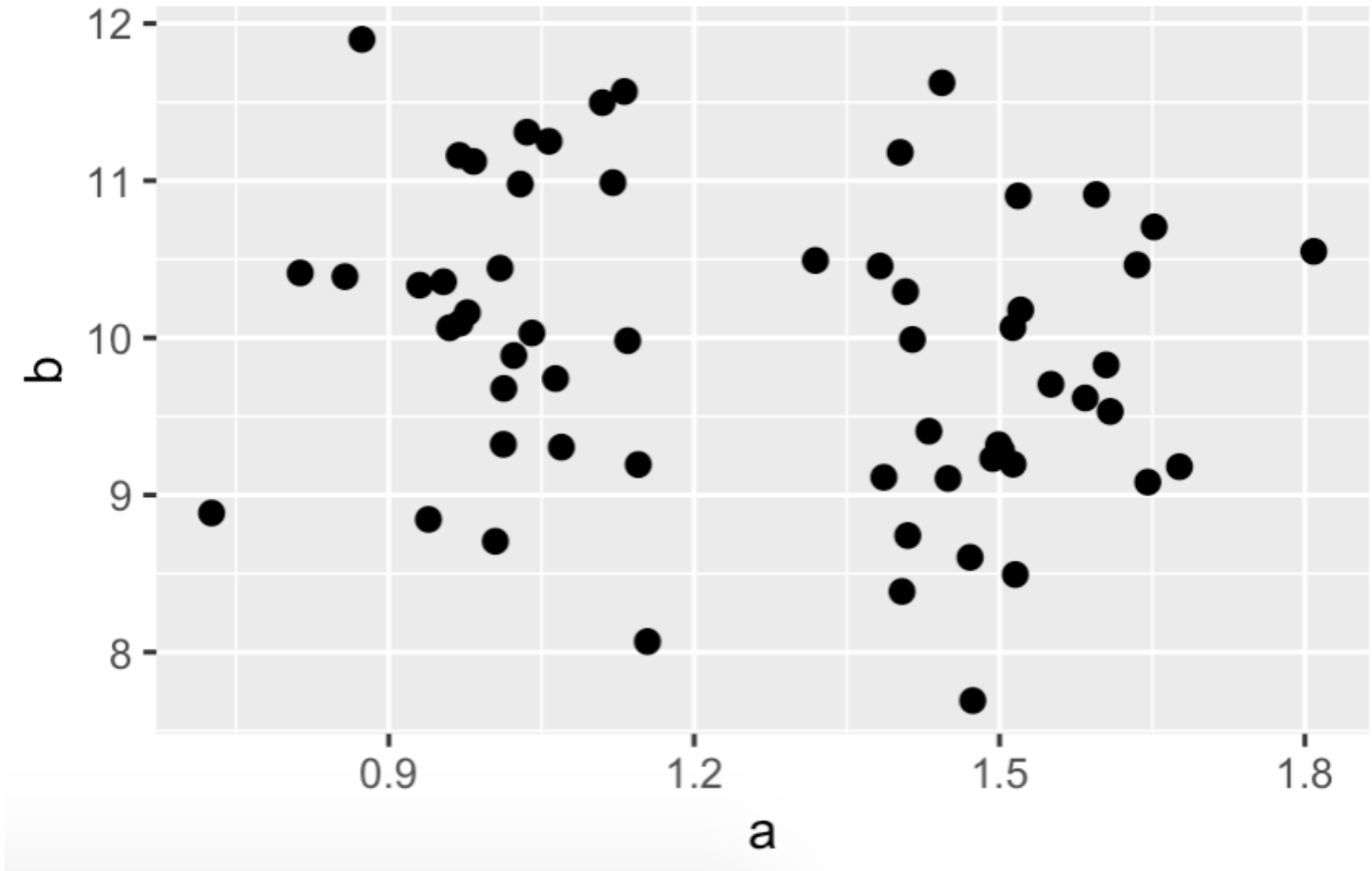


Same principles apply to multivariate data but with Euclidean distance

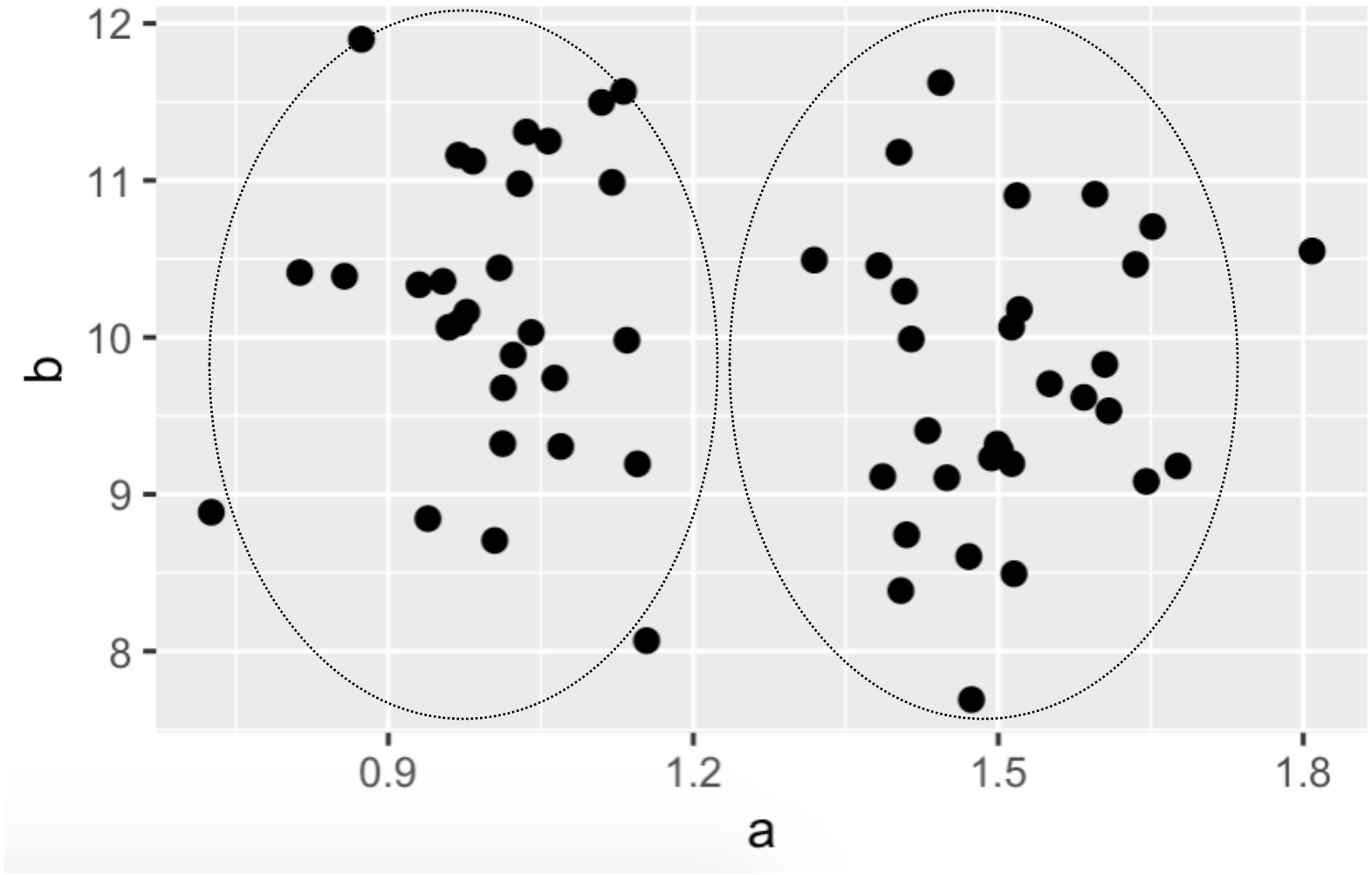
$$d(p,q) = \sqrt{(p-q)^2}$$



K=2

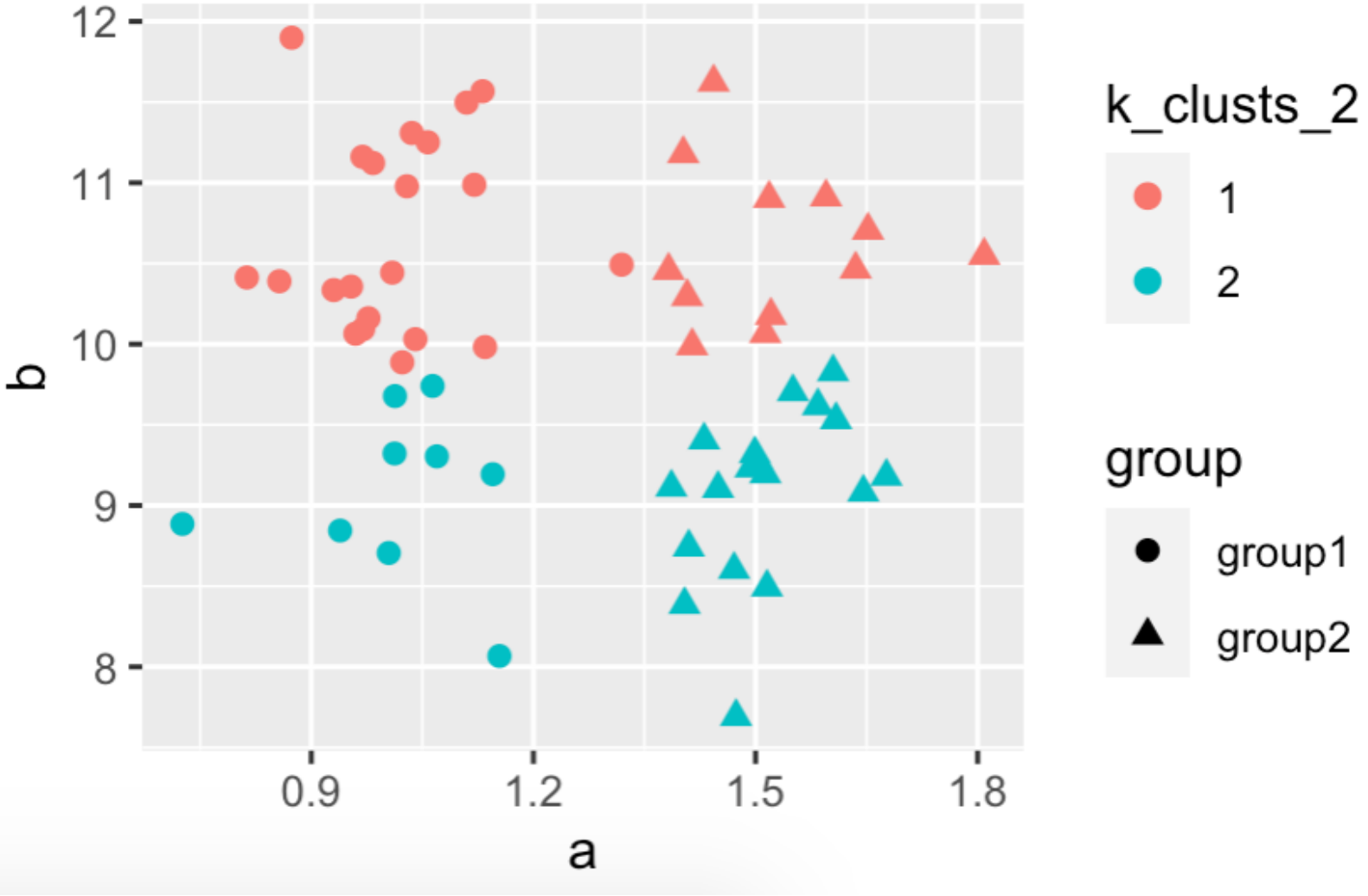


K=2



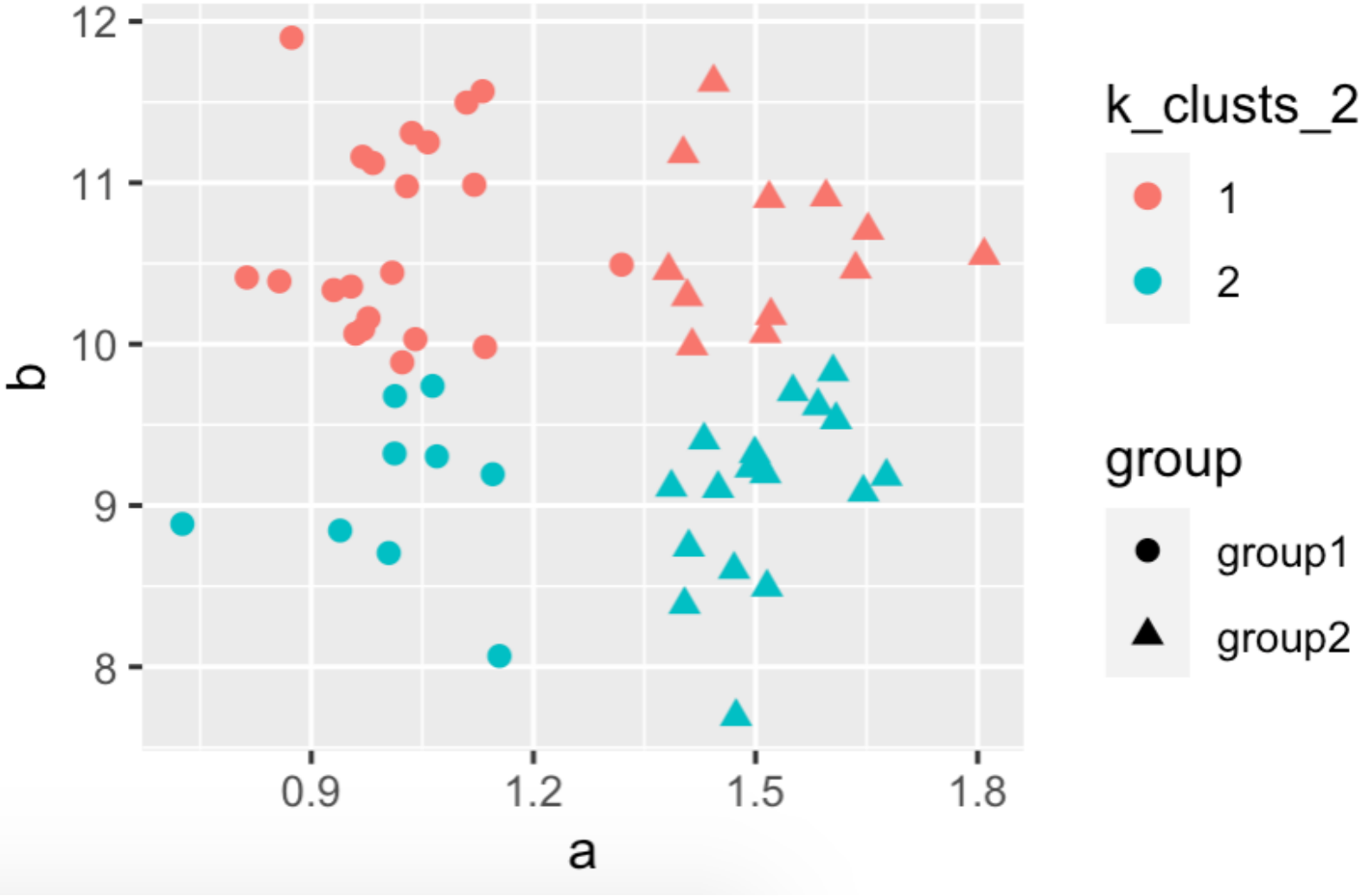
K-means (like other clustering, PCA, etc) often needs scaled data because it relies on Euclidean distance and is optimized by variance

K=2

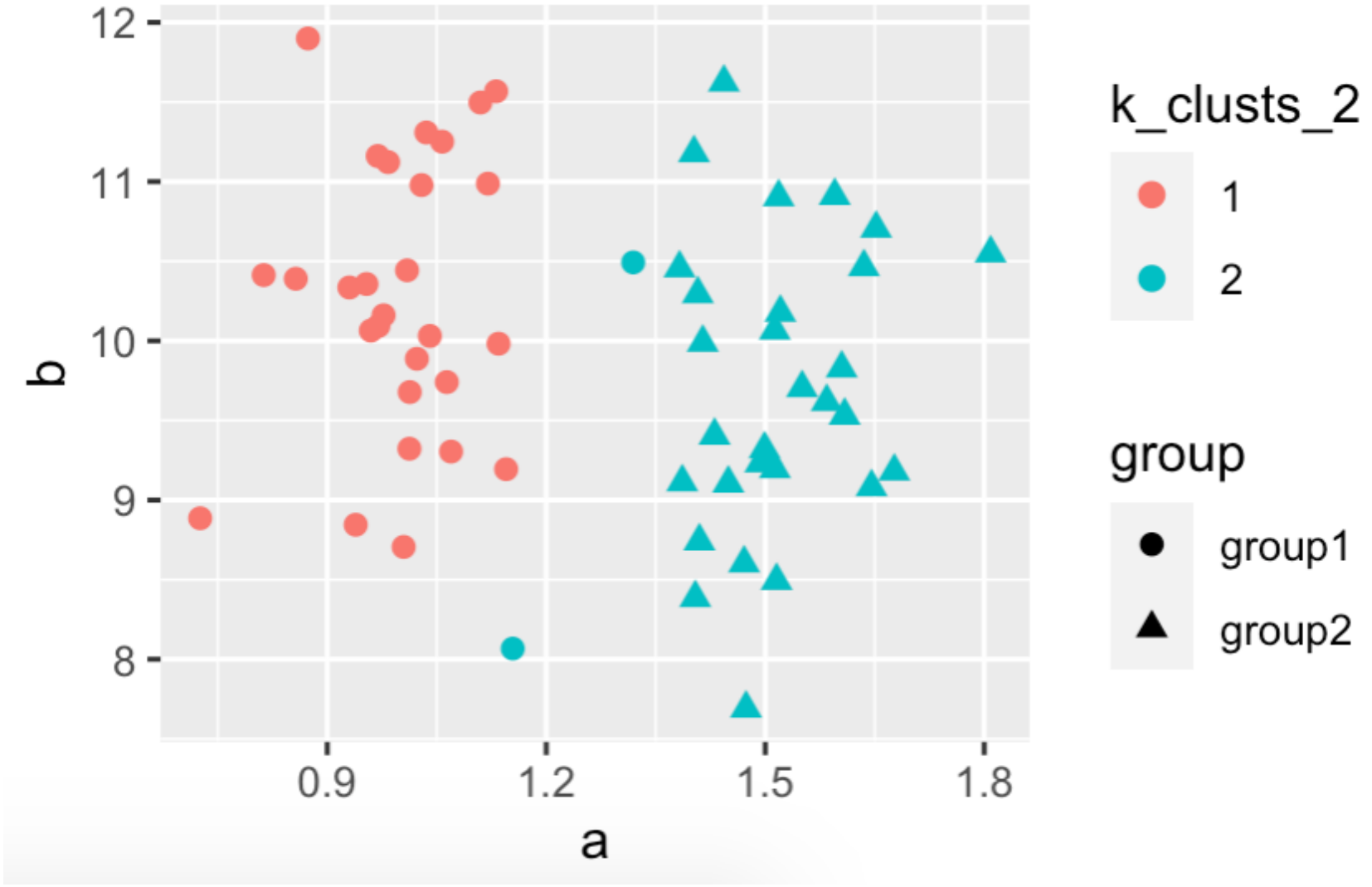


K-means (like other clustering, PCA, etc) often needs scaled data because it relies on Euclidean distance and is optimized by variance

K=2

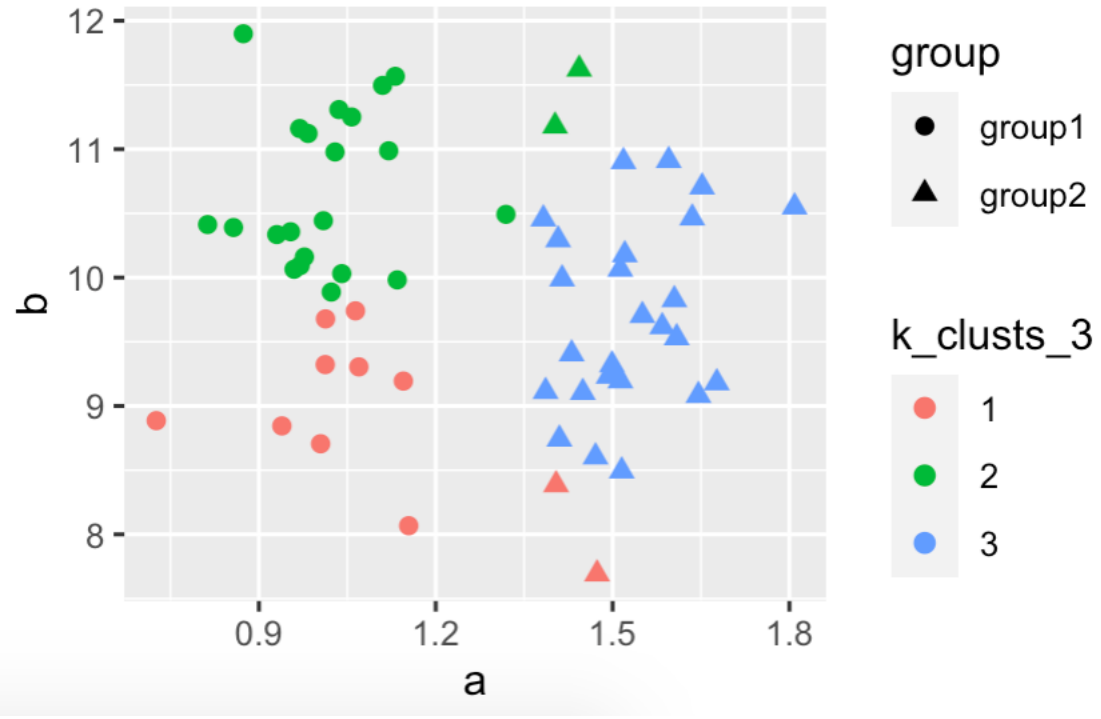


Scaled (z-score) identifies expected groups

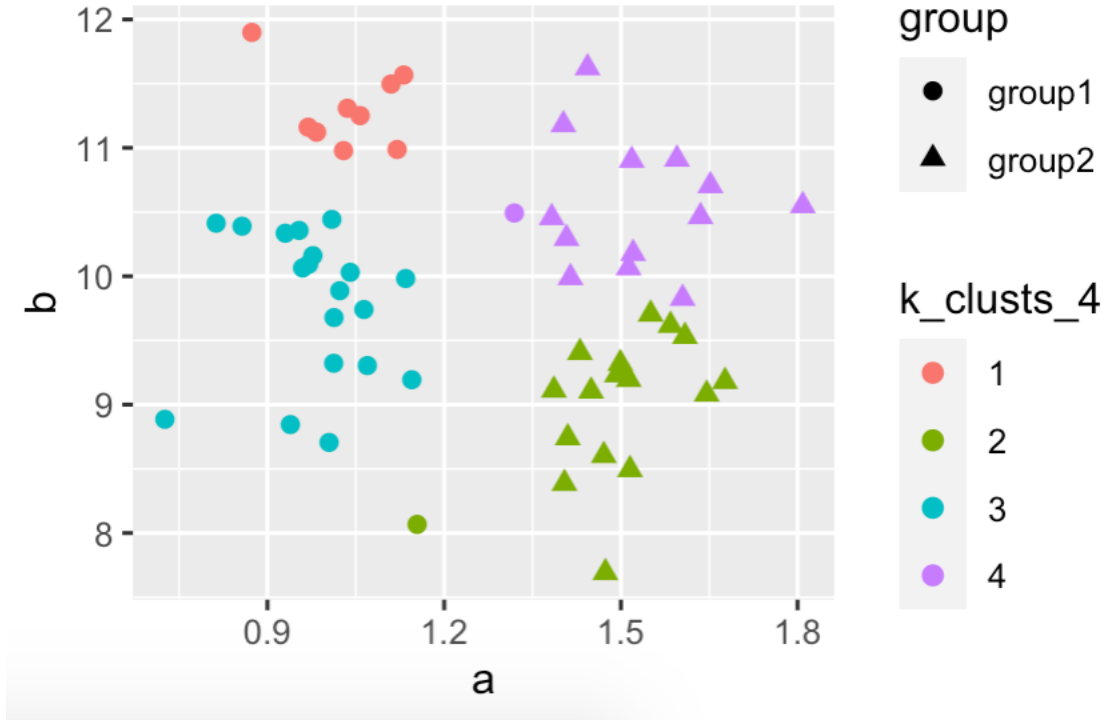


K-means clustering will find however many clusters you tell it to

K=3

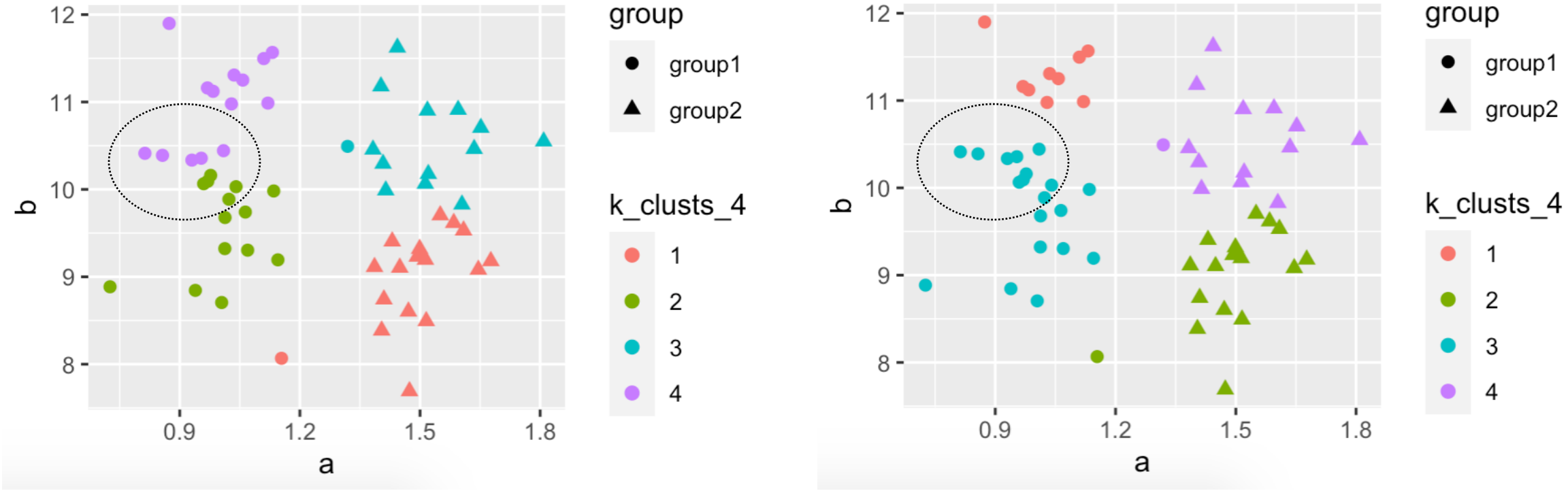


K=4

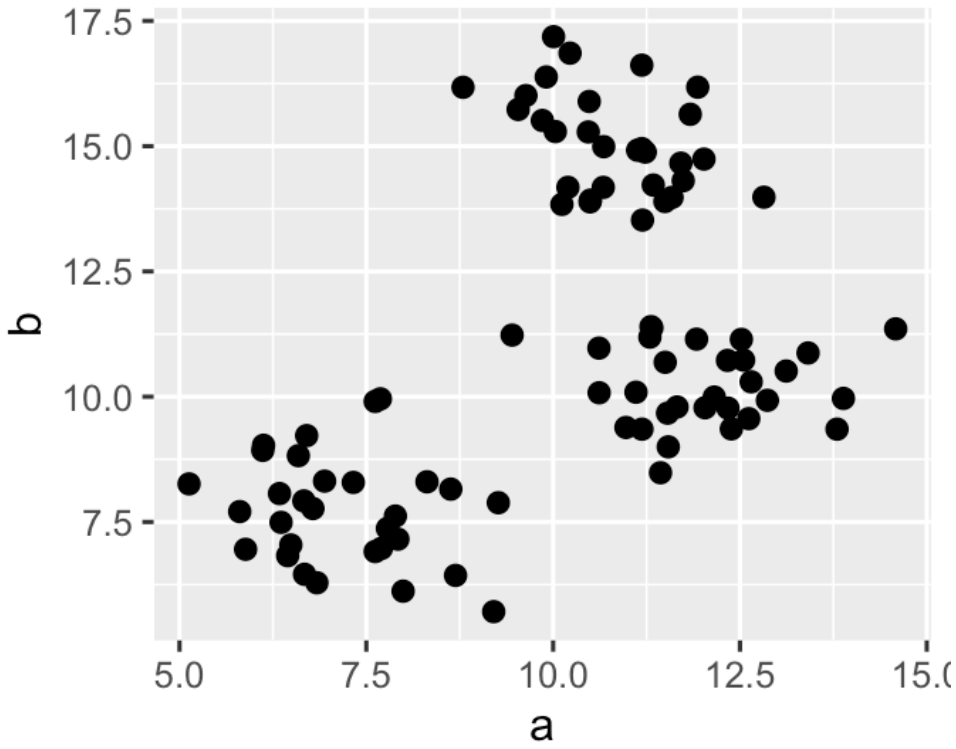


K-means clustering is not deterministic

Group assignments (e.g. 1-4) are arbitrary
Can yield different results (chances of same result greater by increasing number of starts)



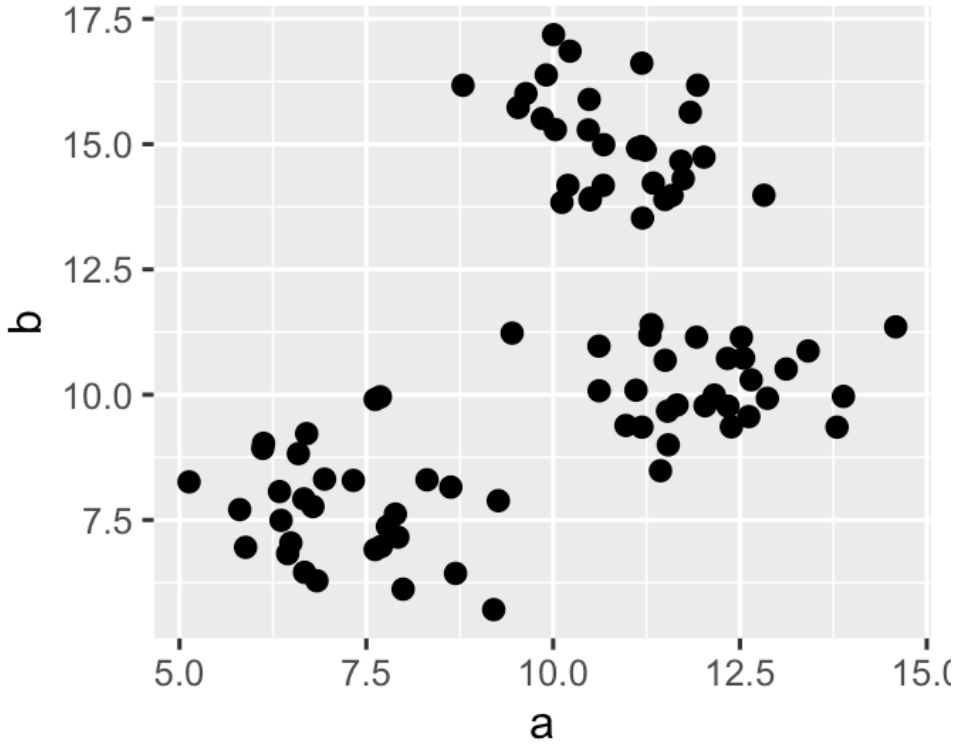
Choosing K



There is no universal method for choosing K!

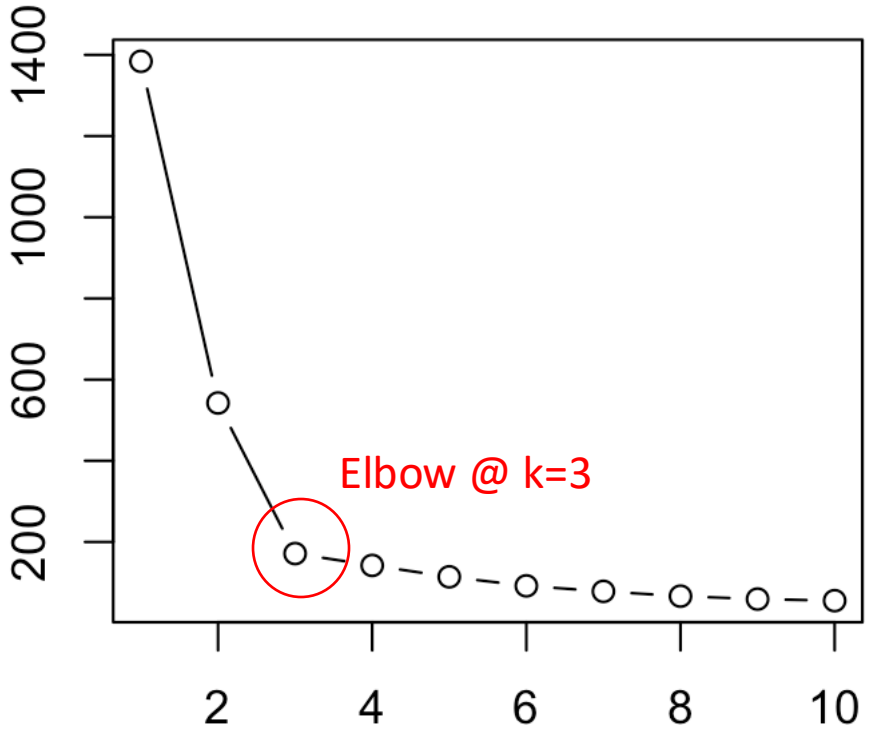
Sometimes its obvious from plotting the data, based on field specific knowledge, or chosen for practical purposes ("I want to identify 4 types of environments to target for breeding")

Choosing K



Total within group sum of squares

Elbow method



K

Silhouette score

Silhouette Score = (b-a)/max(a,b)

Where

a= average intra-cluster distance i.e the average distance between each point within a cluster.

b= average inter-cluster distance i.e the average distance between all clusters.

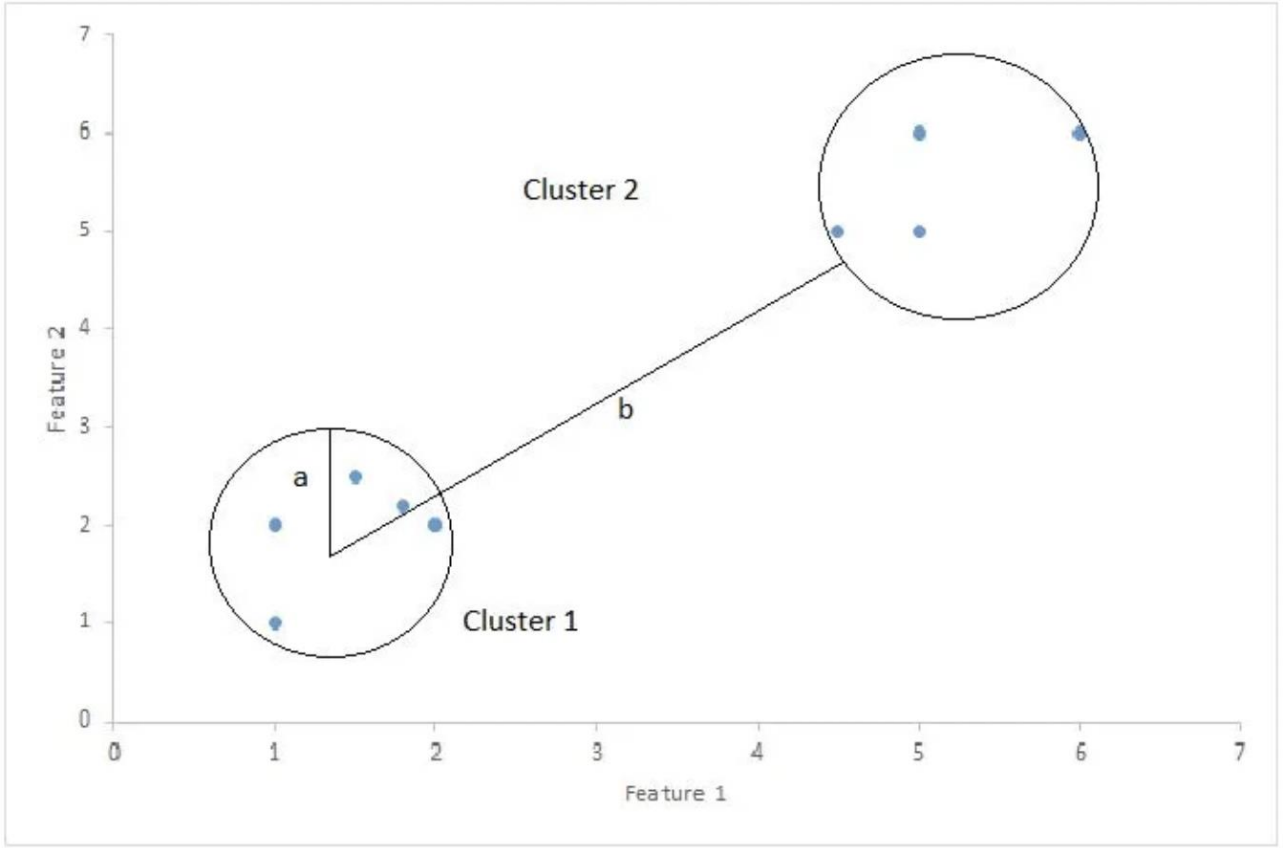
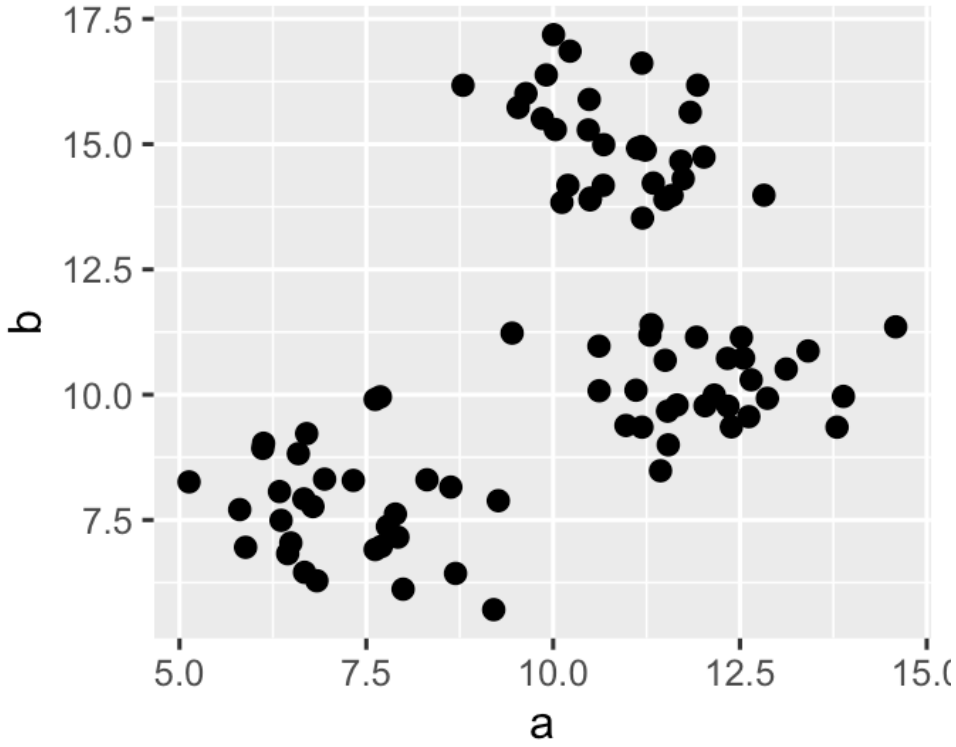


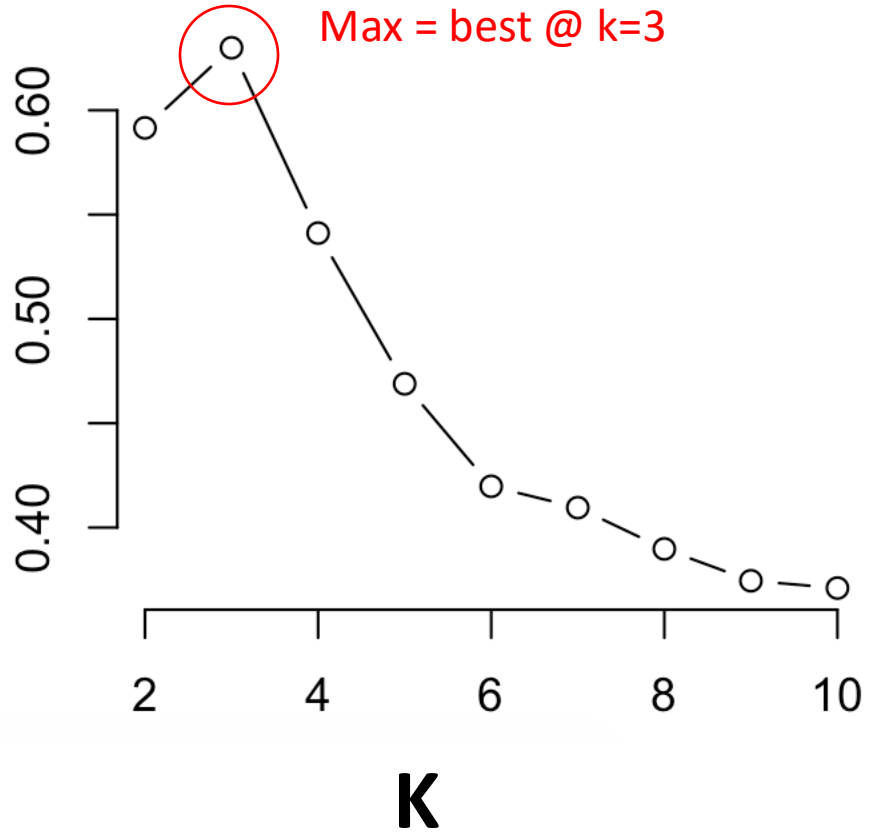
Image by author

Choosing K

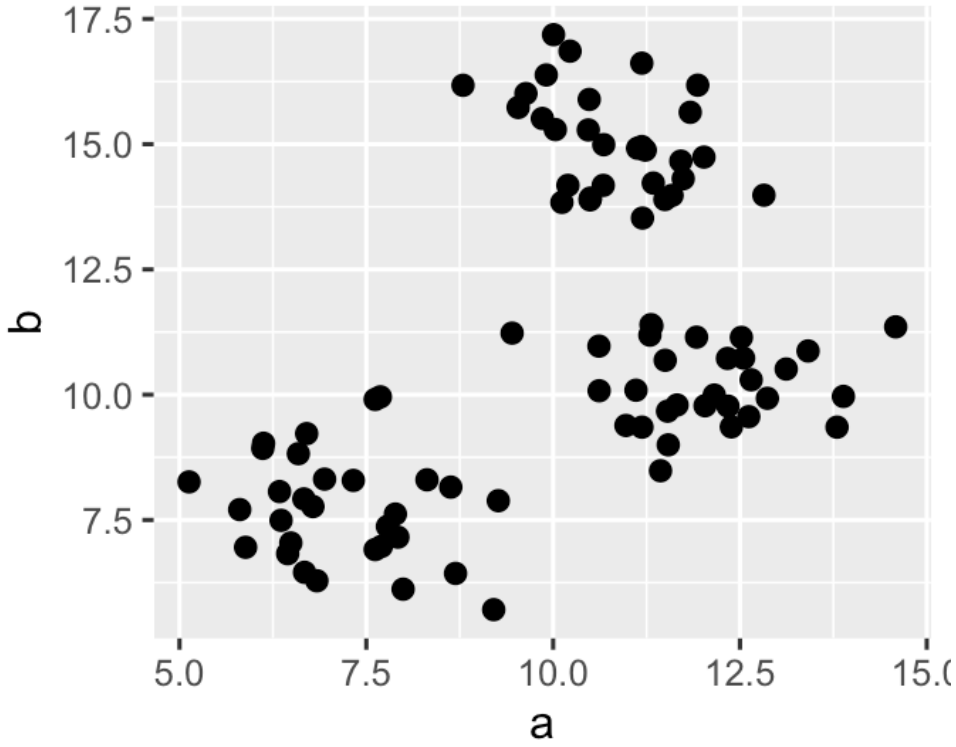


Average silhouette score

Silhouette score method

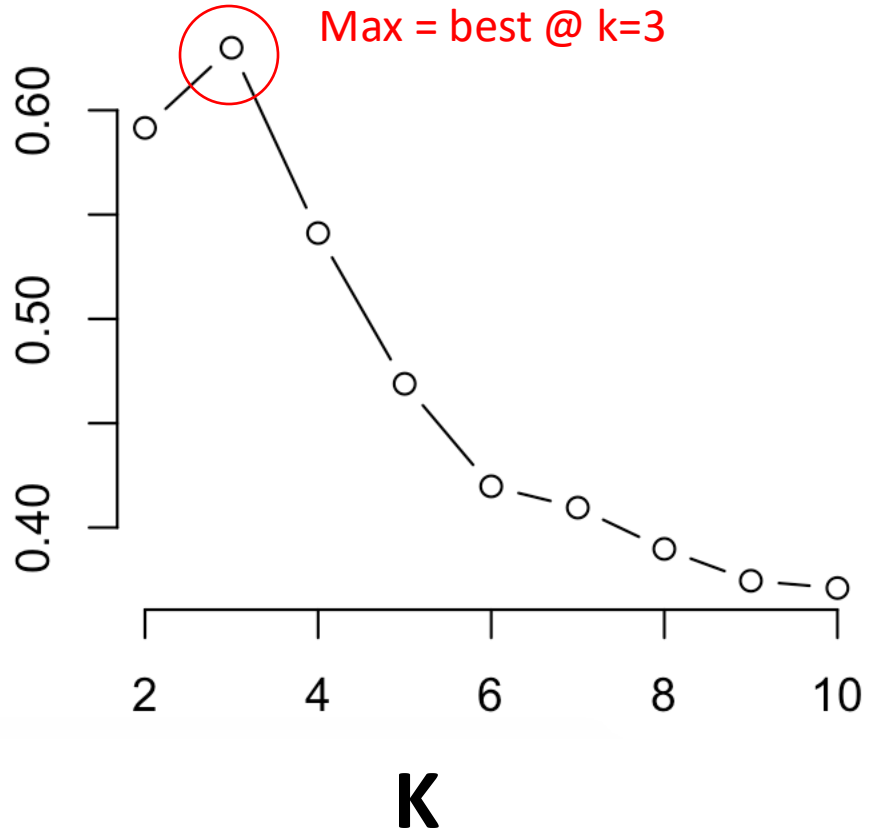


Choosing K



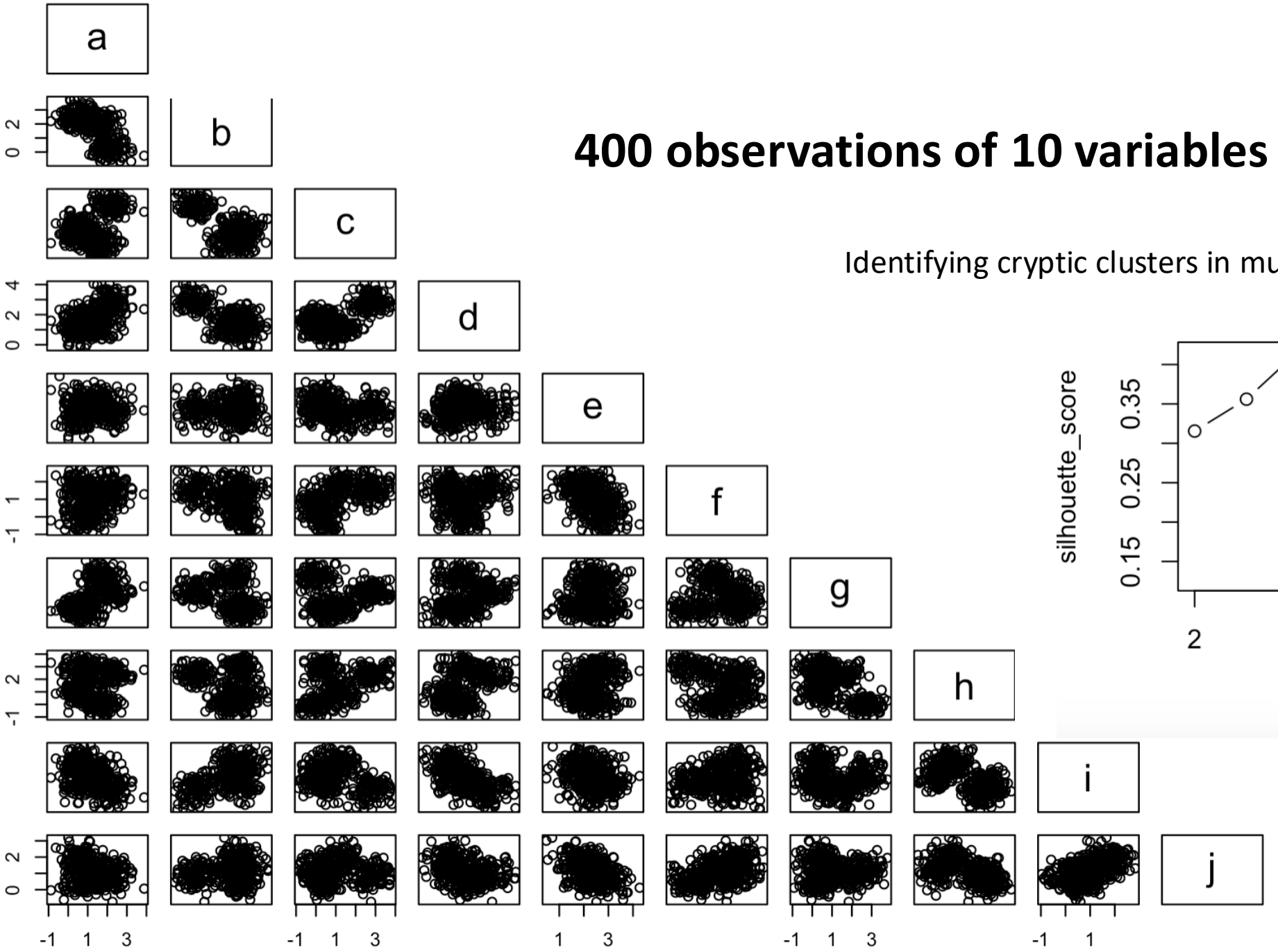
Average silhouette score

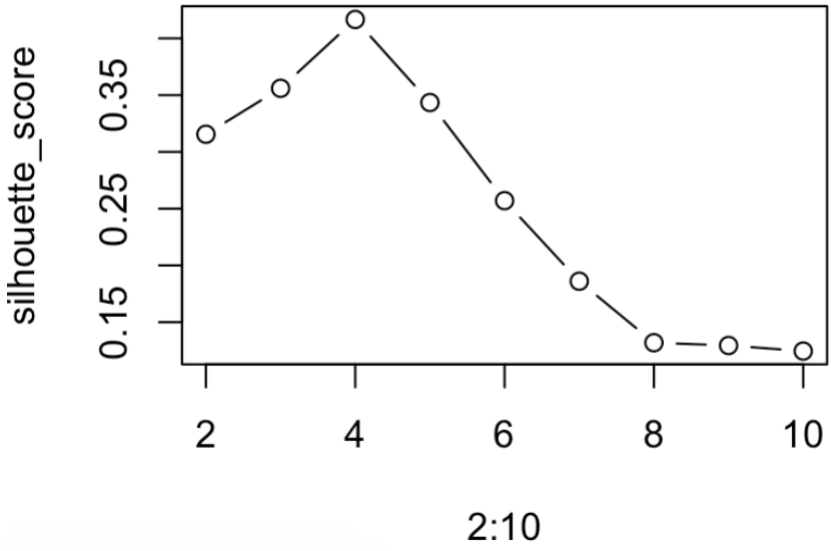
Silhouette score method



400 observations of 10 variables (“a”-”j”)

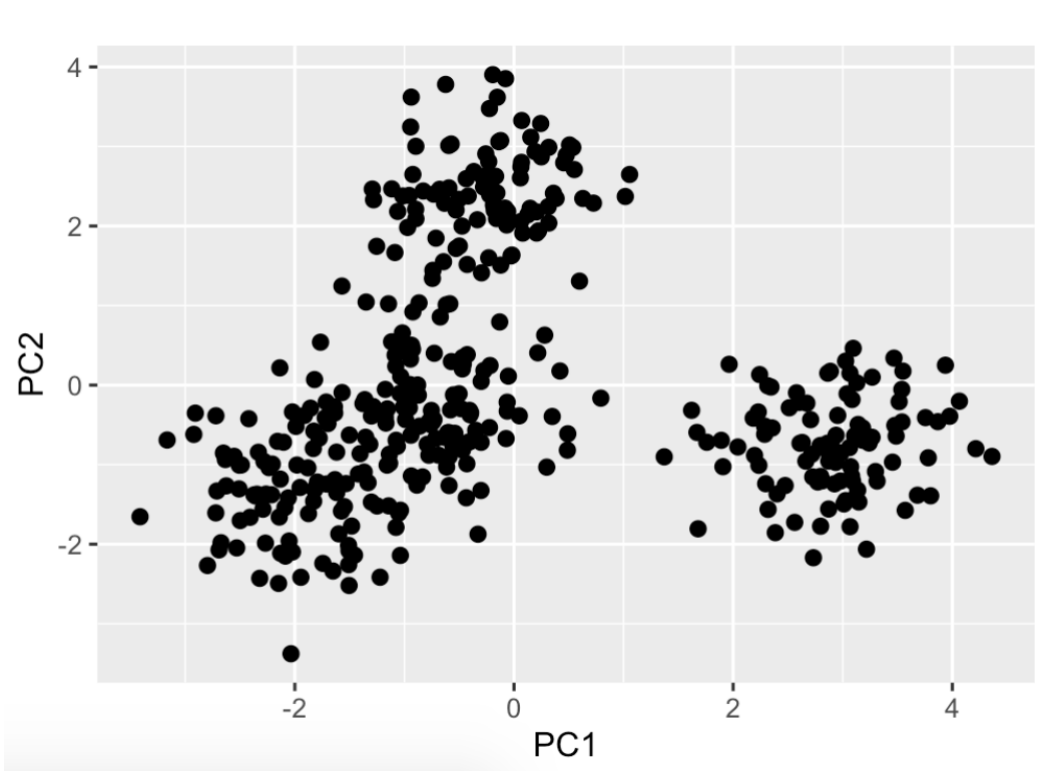
Identifying cryptic clusters in multivariate data

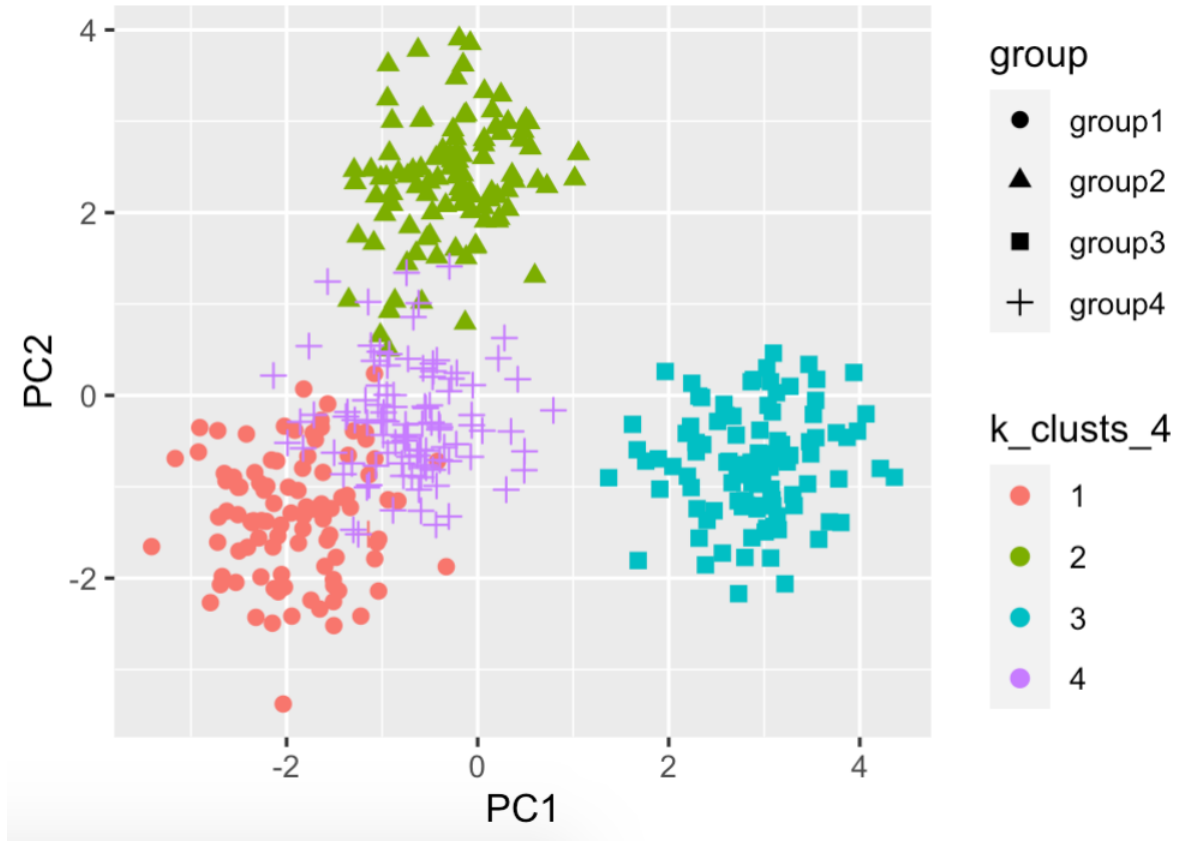




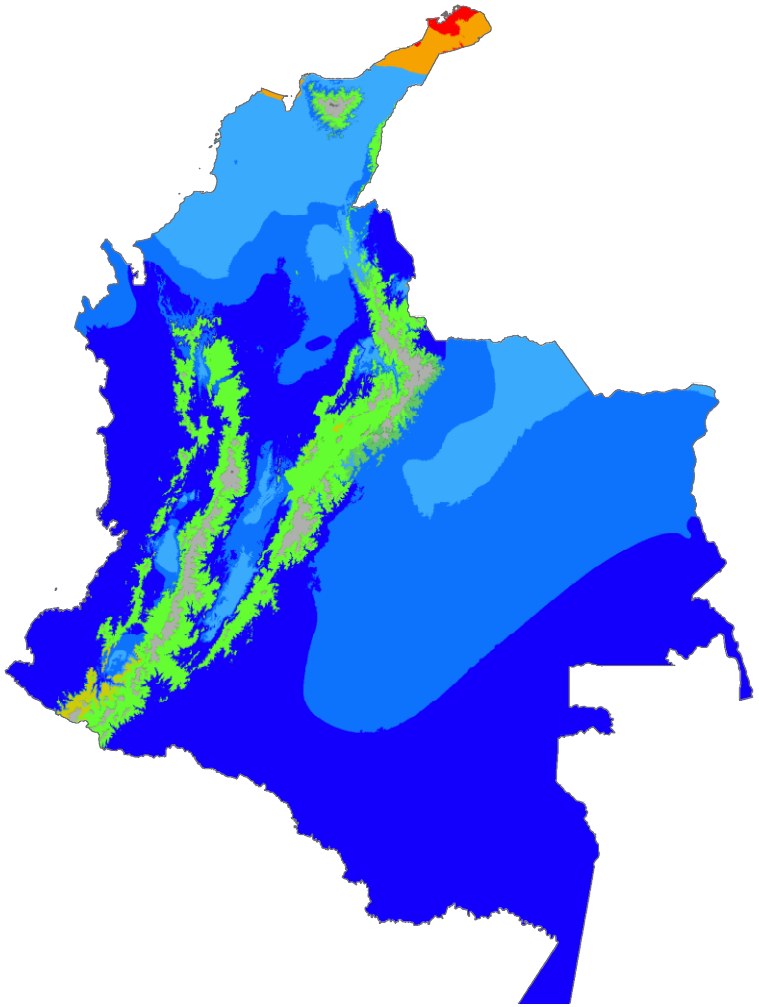
Unsupervised learning!

K-means almost perfectly identifies the 4 clusters





Köppen climate types of Colombia

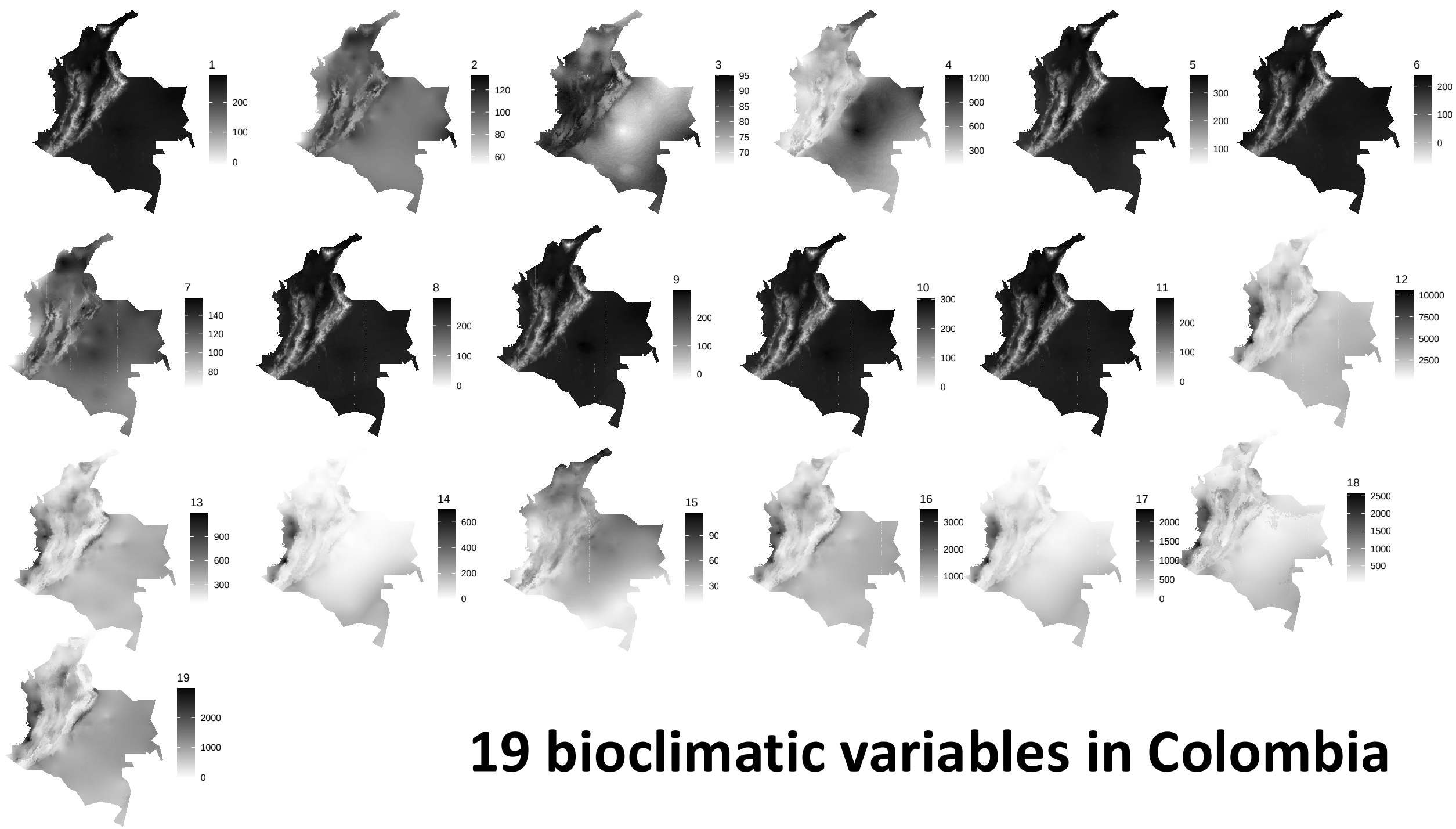


Köppen climate type

- *Af (Rainforest)*
- *Am (Monsoon)*
- *Aw (Savanna)*
- *BWh (Hot desert)*
- *BSh (Hot semi-arid)*
- *Csb (Warm-summer mediterranean)*
- *Csc (Cold-summer mediterranean)*
- *Cwb (Subtropical highland)*
- *Cwc (Cold-summer subtropical highland)*
- *Cfb (Oceanic)*
- *Cfc (Subpolar oceanic)*
- *ET (Tundra)*
- *EF (Ice-cap)*

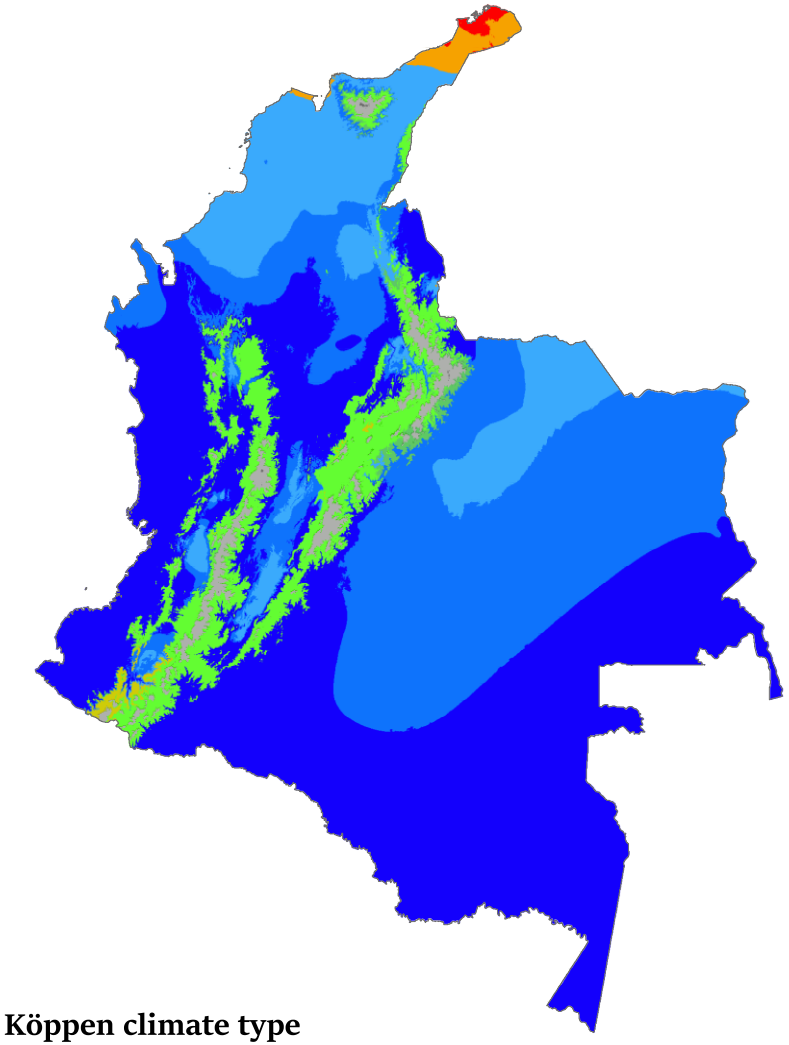
**Isotherm used to separate temperate (C) and continental (D) climates is -3°C*

Data source: Climate types calculated from data from WorldClim.org



19 bioclimatic variables in Colombia

Köppen climate types of Colombia

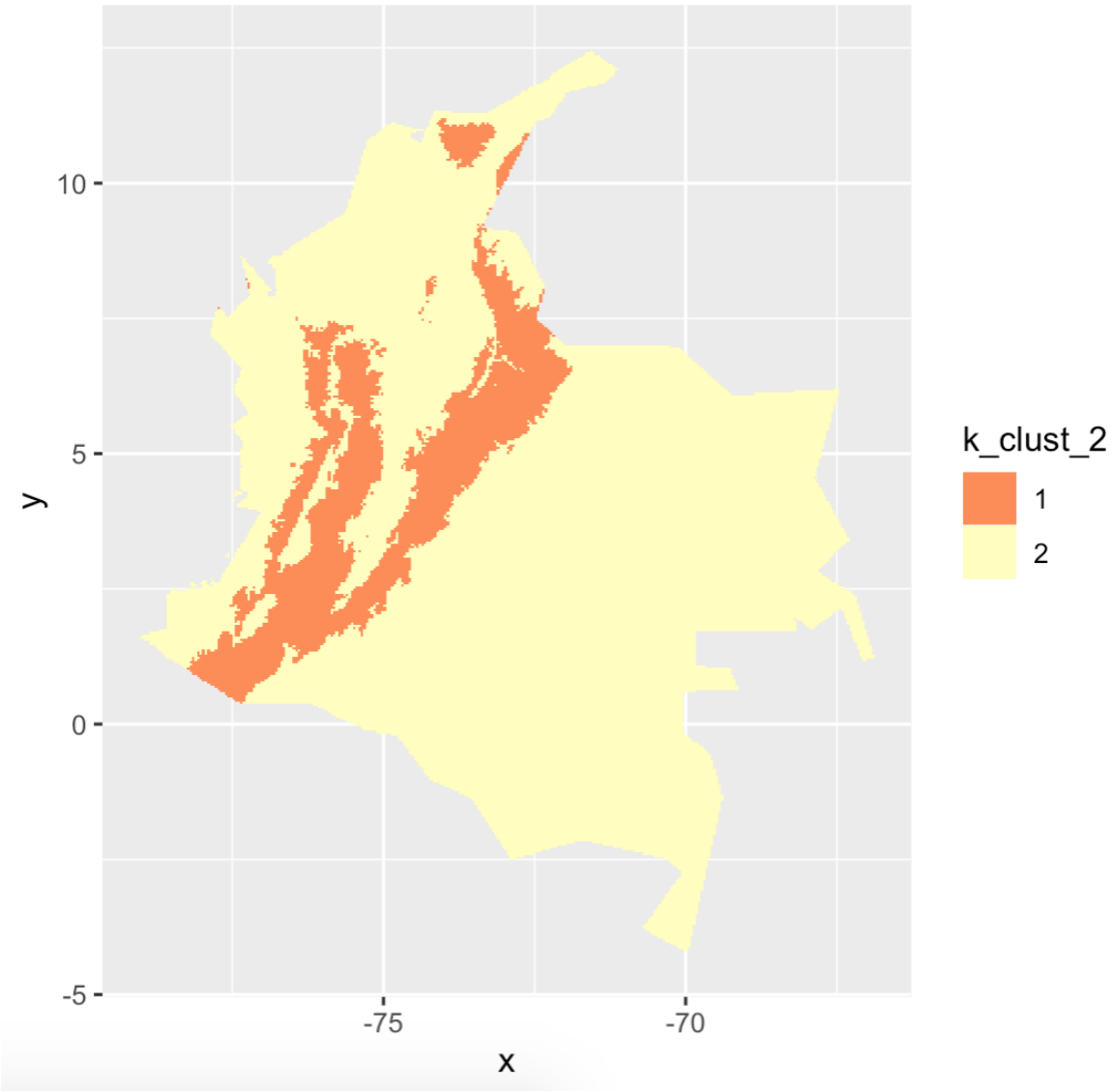


Köppen climate type

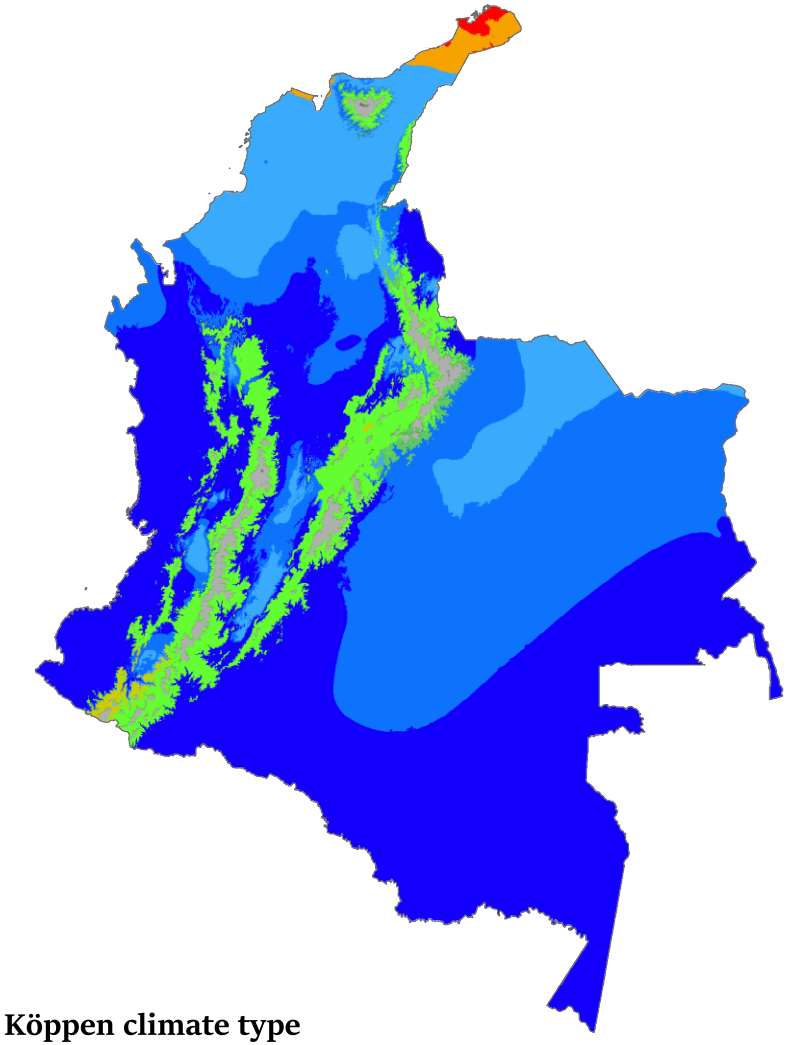
- *Af (Rainforest)*
- *Am (Monsoon)*
- *Aw (Savanna)*
- *BWh (Hot desert)*
- *BSh (Hot semi-arid)*
- *Csb (Warm-summer mediterranean)*
- *Csc (Cold-summer mediterranean)*
- *Cwb (Subtropical highland)*
- *Cwc (Cold-summer subtropical highland)*
- *Cfb (Oceanic)*
- *Cfc (Subpolar oceanic)*
- *ET (Tundra)*
- *EF (Ice-cap)*

**Isotherm used to separate temperate (C) and continental (D) climates is -3°C*
Data source: Climate types calculated from data from WorldClim.org

K=2



Köppen climate types of Colombia

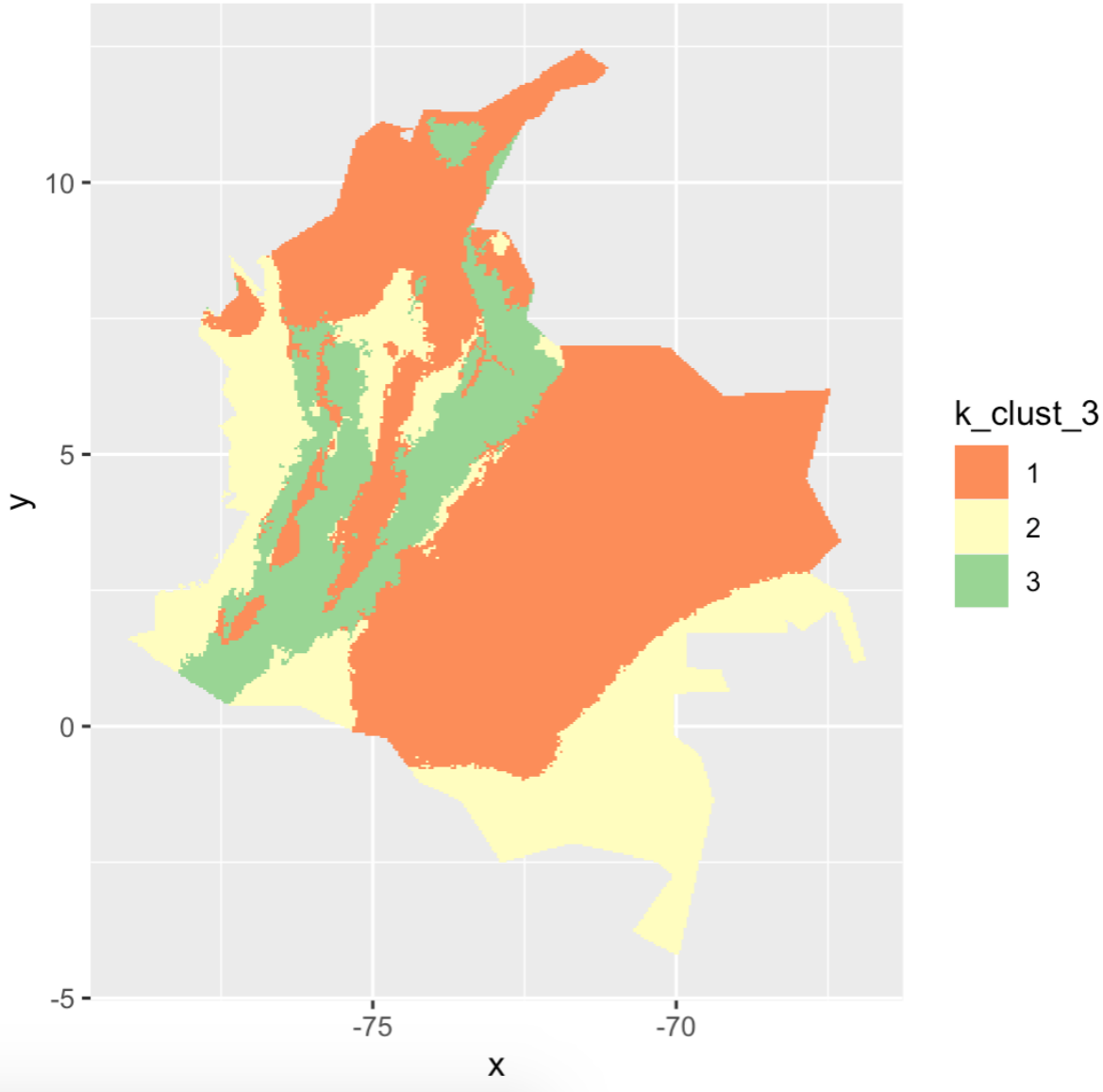


Köppen climate type

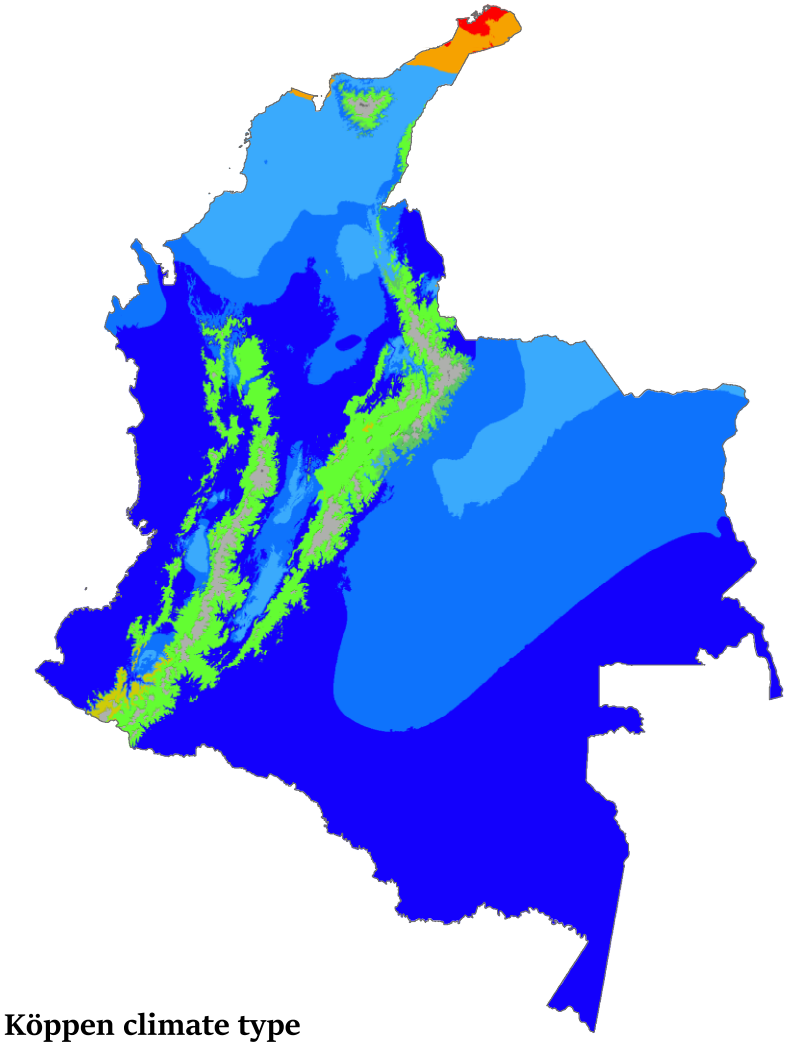
- *Af (Rainforest)*
- *Am (Monsoon)*
- *Aw (Savanna)*
- *BWh (Hot desert)*
- *BSh (Hot semi-arid)*
- *Csb (Warm-summer mediterranean)*
- *Csc (Cold-summer mediterranean)*
- *Cwb (Subtropical highland)*
- *Cwc (Cold-summer subtropical highland)*
- *Cfb (Oceanic)*
- *Cfc (Subpolar oceanic)*
- *ET (Tundra)*
- *EF (Ice-cap)*

*Isotherm used to separate temperate (C) and continental (D) climates is -3°C
Data source: Climate types calculated from data from WorldClim.org

K=3



Köppen climate types of Colombia

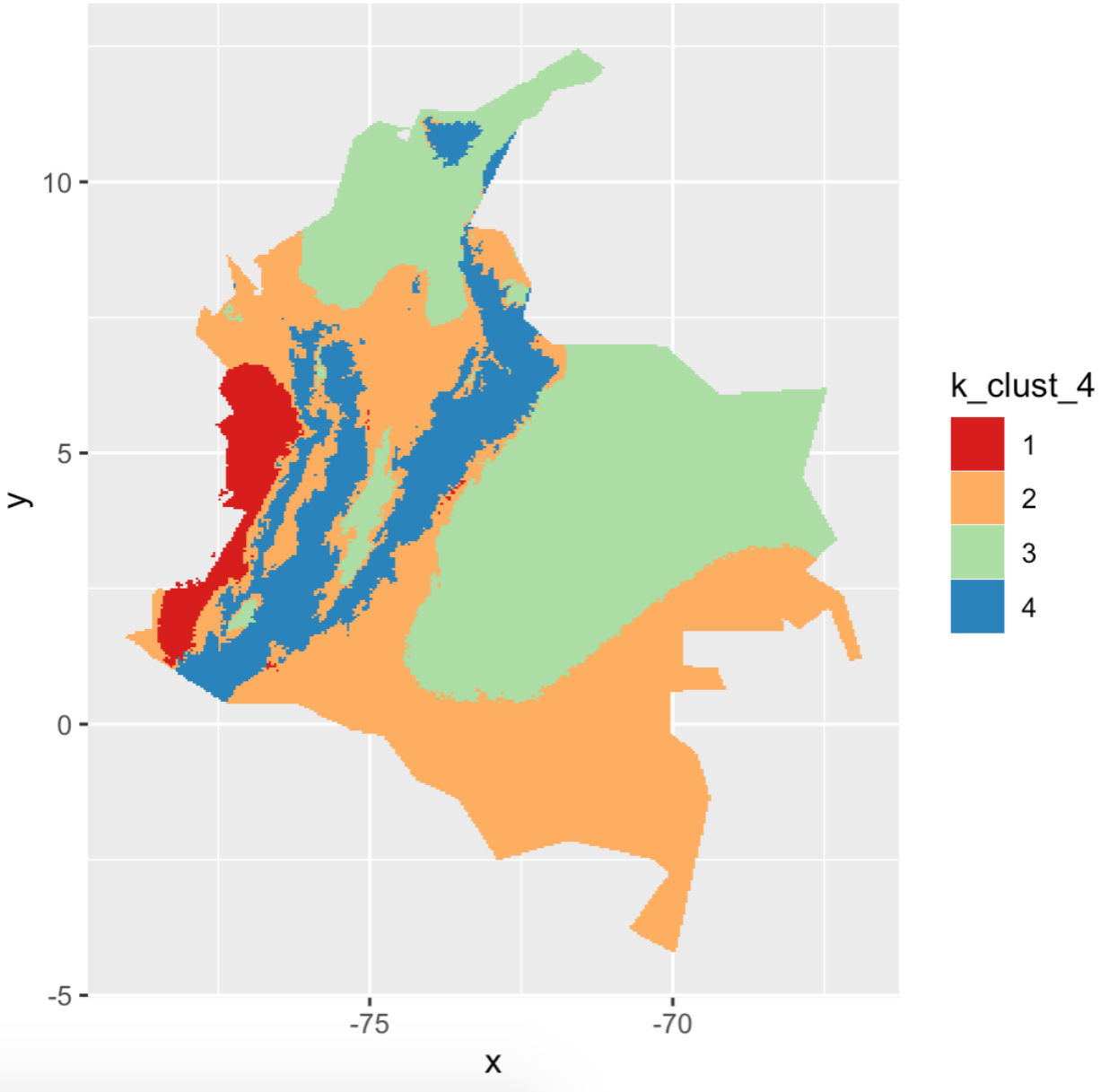


Köppen climate type

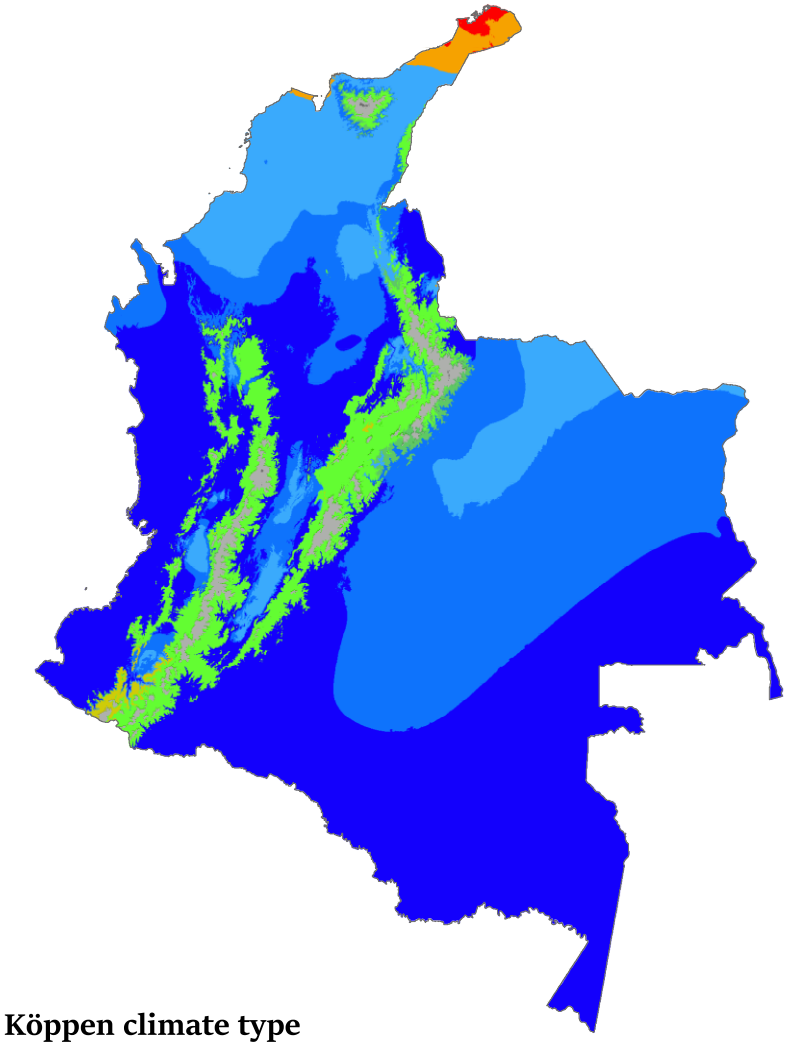
- *Af (Rainforest)*
- *Am (Monsoon)*
- *Aw (Savanna)*
- *BWh (Hot desert)*
- *BSh (Hot semi-arid)*
- *Csb (Warm-summer mediterranean)*
- *Csc (Cold-summer mediterranean)*
- *Cwb (Subtropical highland)*
- *Cwc (Cold-summer subtropical highland)*
- *Cfb (Oceanic)*
- *Cfc (Subpolar oceanic)*
- *ET (Tundra)*
- *EF (Ice-cap)*

**Isotherm used to separate temperate (C) and continental (D) climates is -3°C*
Data source: Climate types calculated from data from WorldClim.org

K=4



Köppen climate types of Colombia



Köppen climate type

- *Af (Rainforest)*
- *Am (Monsoon)*
- *Aw (Savanna)*
- *BWh (Hot desert)*
- *BSh (Hot semi-arid)*
- *Csb (Warm-summer mediterranean)*
- *Csc (Cold-summer mediterranean)*
- *Cwb (Subtropical highland)*
- *Cwc (Cold-summer subtropical highland)*
- *Cfb (Oceanic)*
- *Cfc (Subpolar oceanic)*
- *ET (Tundra)*
- *EF (Ice-cap)*

*Isotherm used to separate temperate (C) and continental (D) climates is -3°C
Data source: Climate types calculated from data from WorldClim.org

K=10