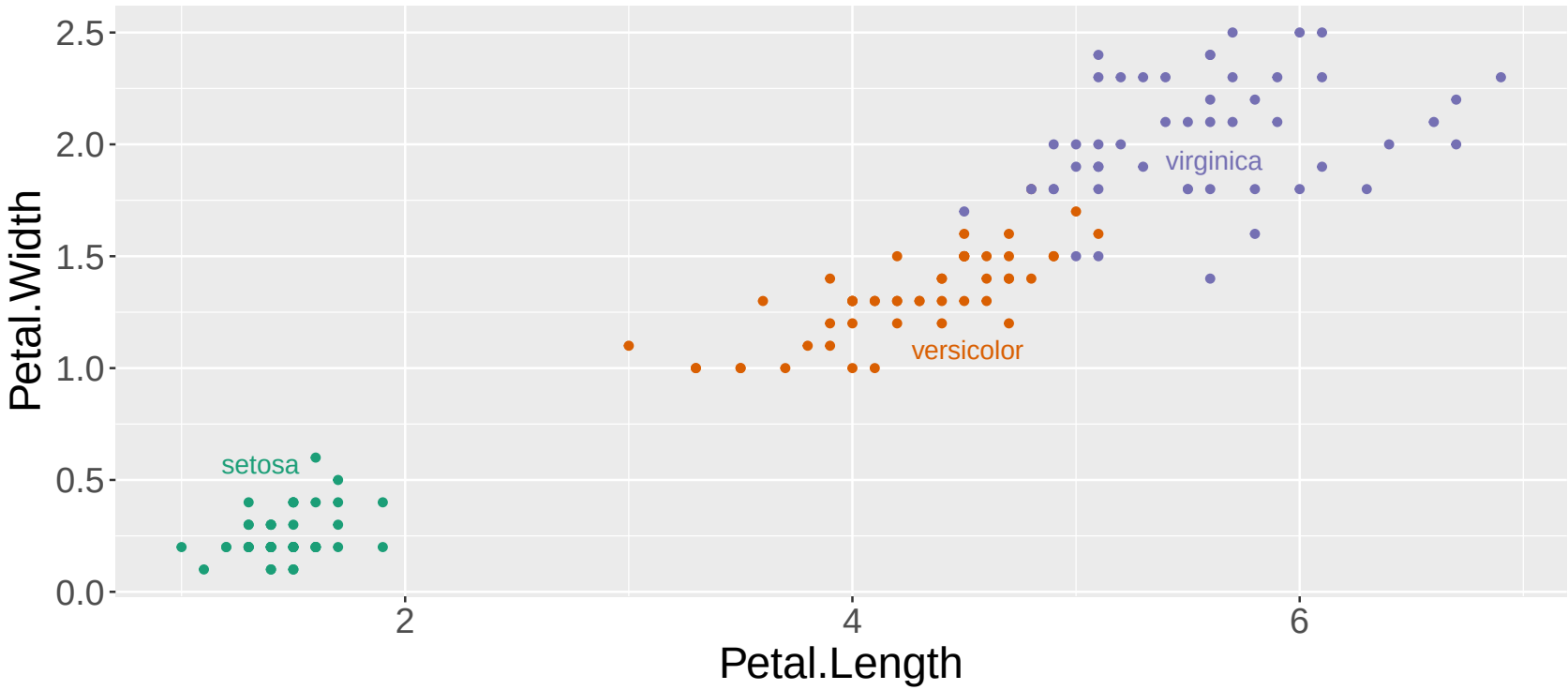
Clustering and k-means

Toby Dylan Hocking

Visualize iris data with labels

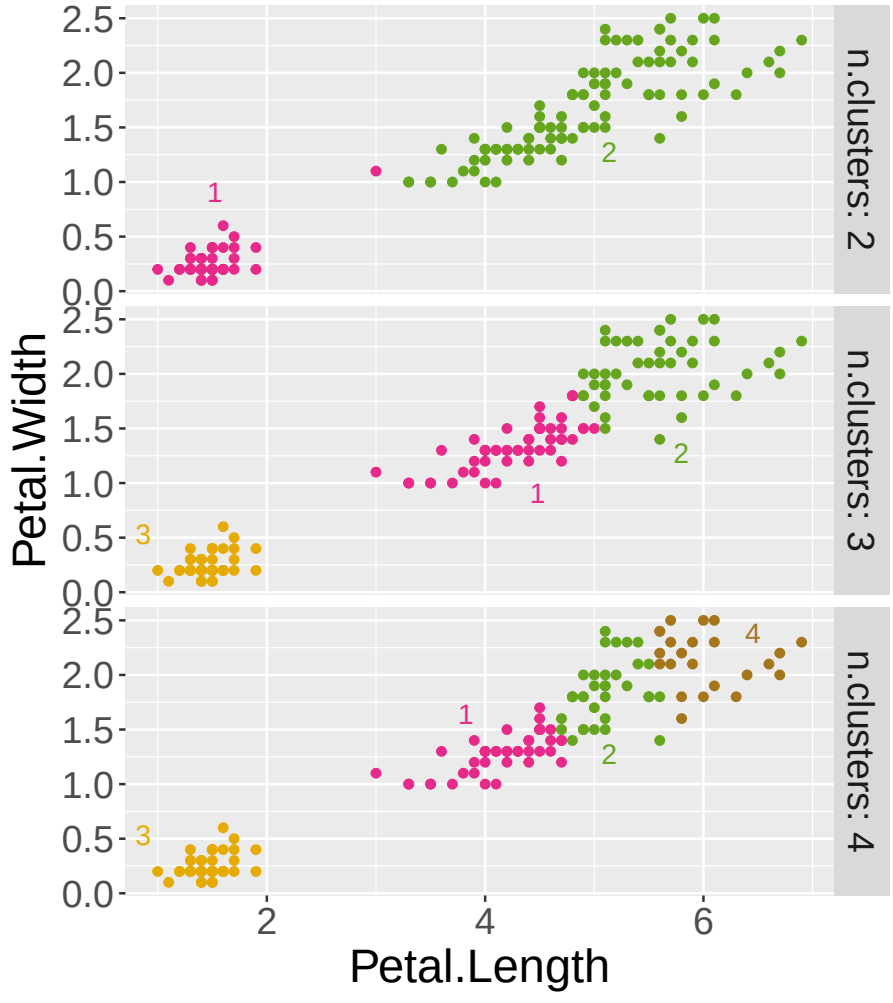


Visualize iris data without labels

- Let $X \in \mathbb{R}^{150\times 2}$ be the data matrix (input for clustering).

```
##       Petal.Width Petal.Length
## [1,]          0.2          1.4
## [2,]          0.2          1.4
## [3,]          0.2          1.3
## [4,]          0.2          1.5
## [5,]          0.2          1.4
## [6,]          0.4          1.7
```

Visualize several clusterings

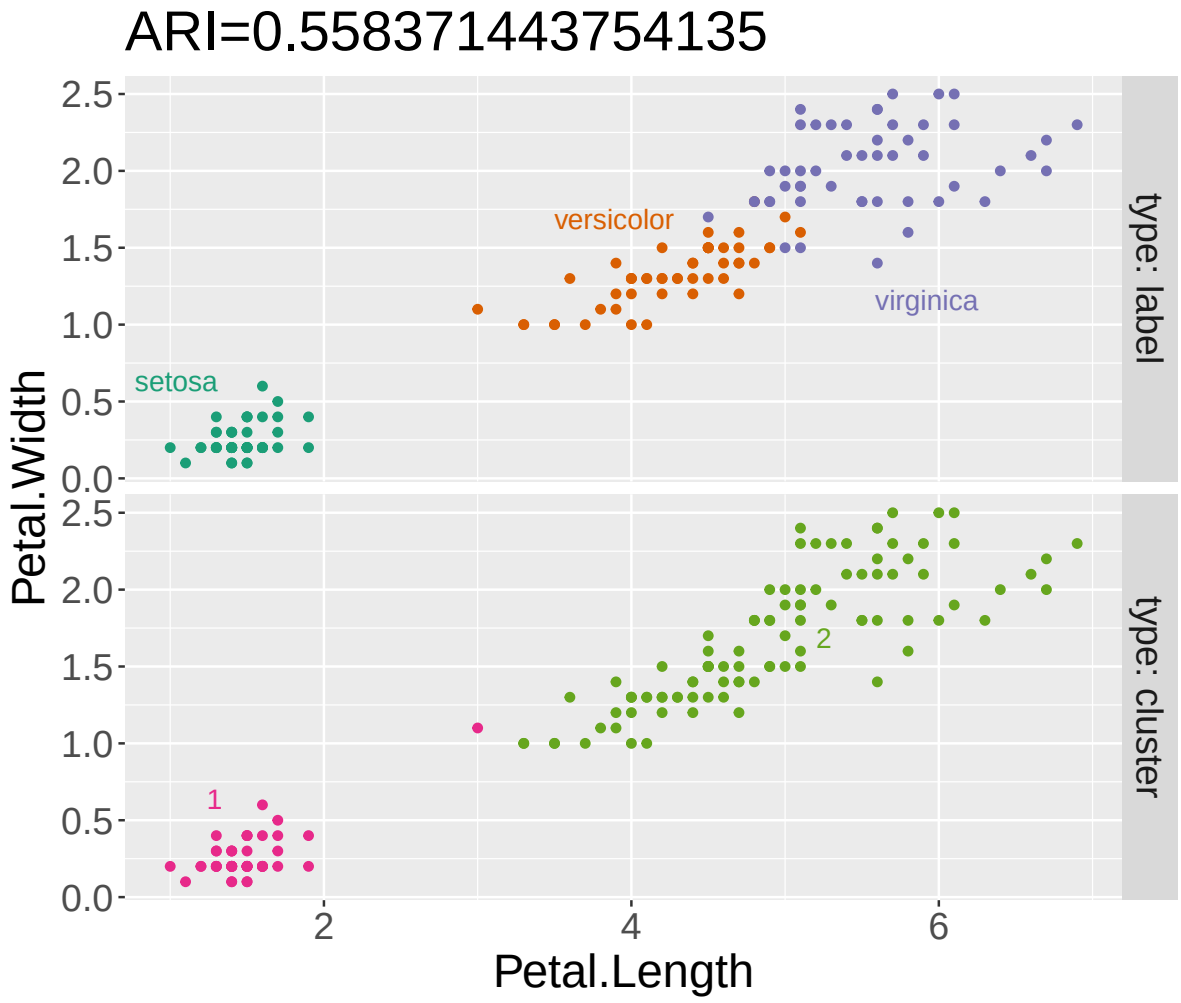


- K-means algorithm (kmeans function in R).
- Which K is best? How to choose number of clusters?

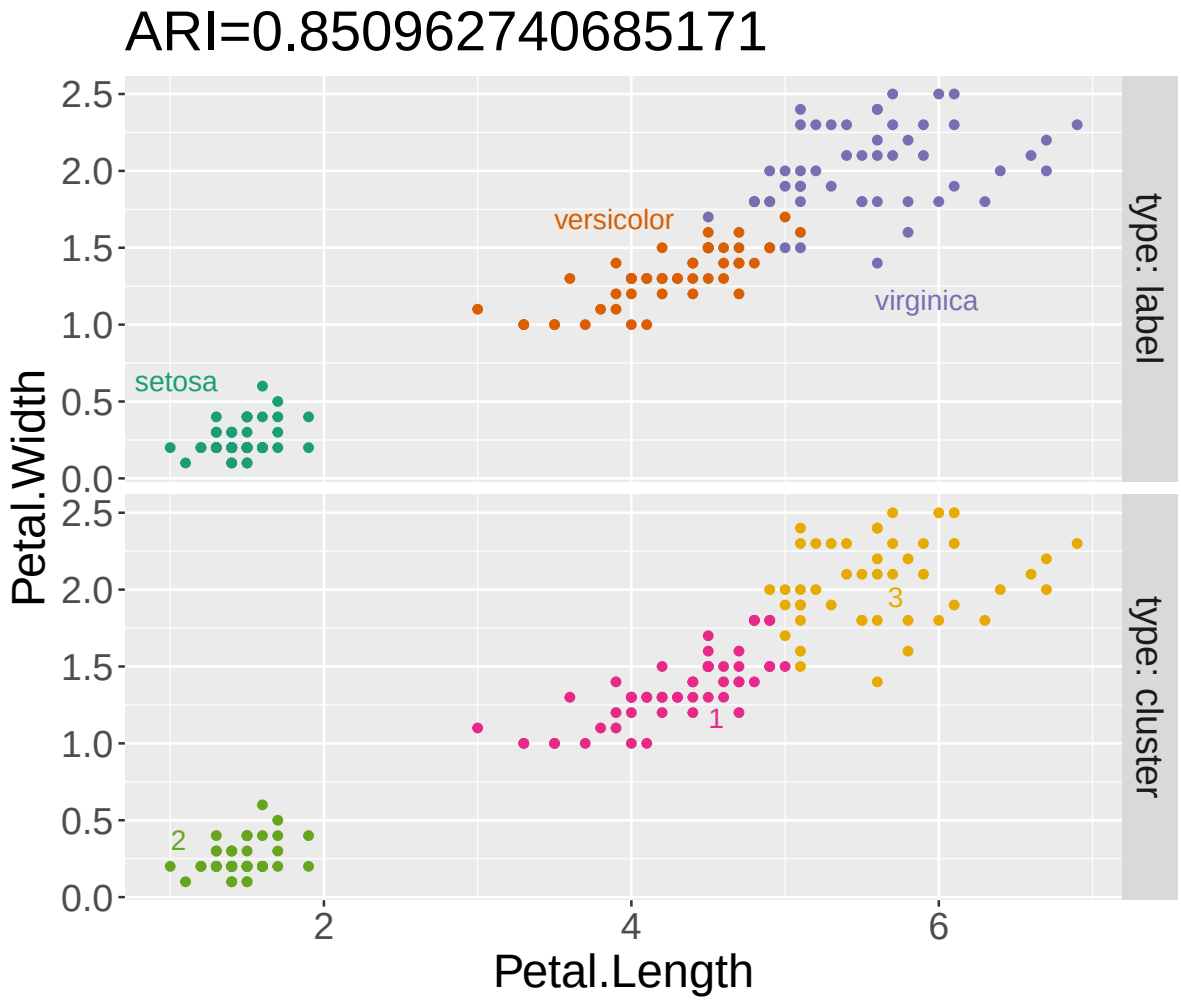
Adjusted Rand Index (ARI)

- Measures agreement between two label/cluster vectors (the two vectors must be the same size).
- Number of labels does not have to be equal to the number of clusters.
- Labels may be different from clusters, and not obvious to match.
- Here labels are species names (setosa, virginica, versicolor) whereas clusters are integers (1, 2, 3).
- Best value $= 1$ (perfect agreement).
- Random/constant assignment $= 0$ (clustering meaningless).

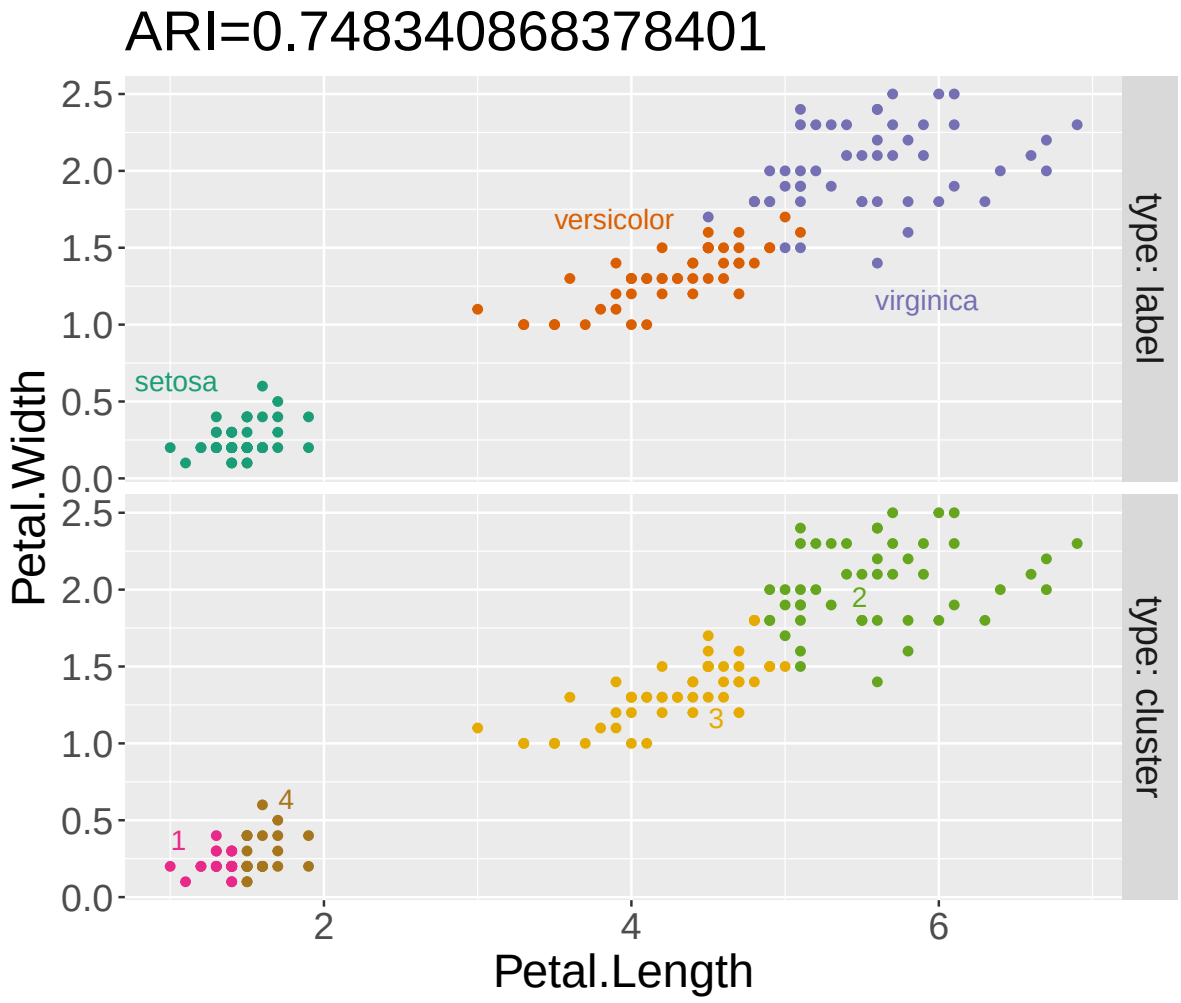
Compare two clusters to labels



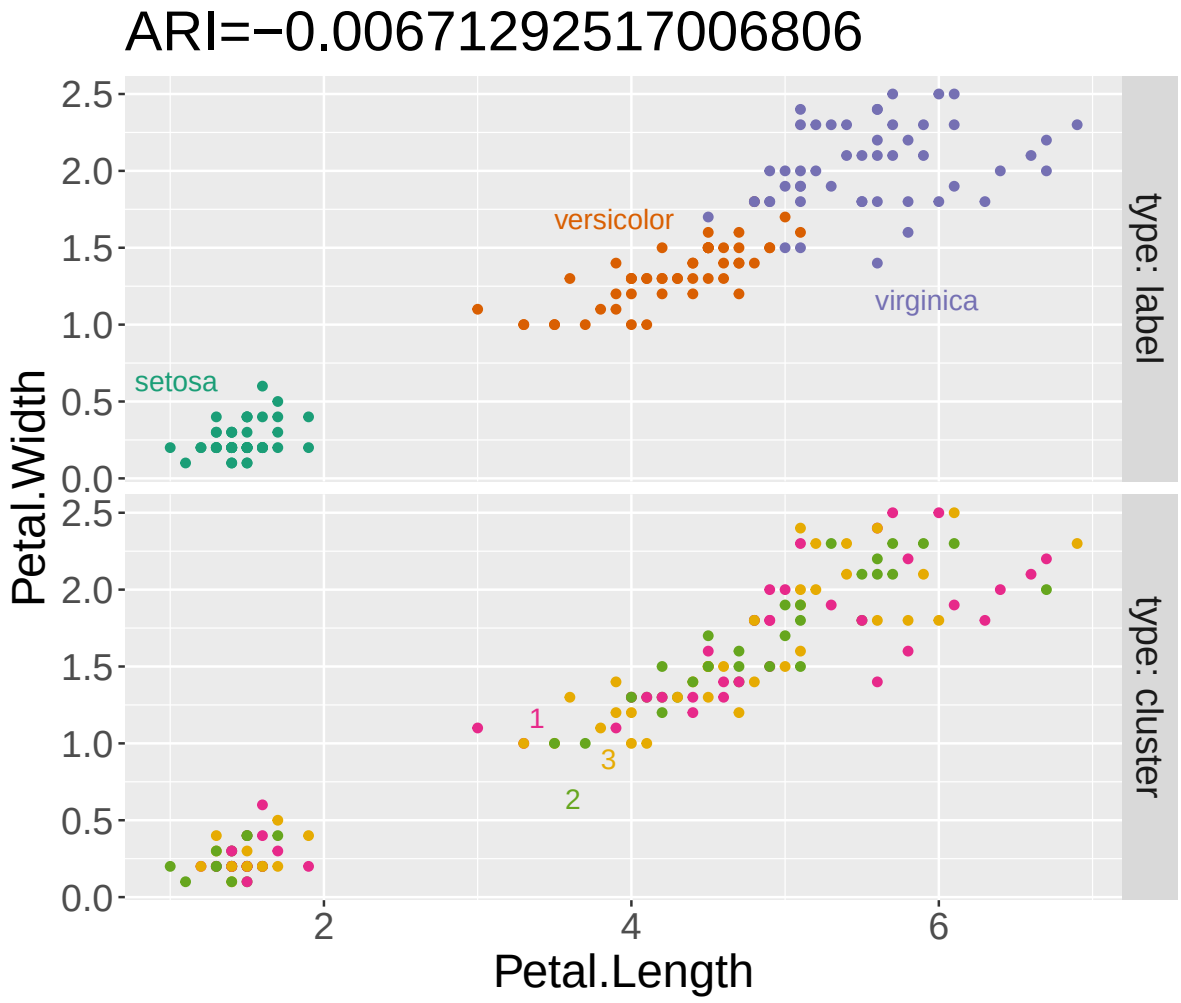
Compare three clusters to labels



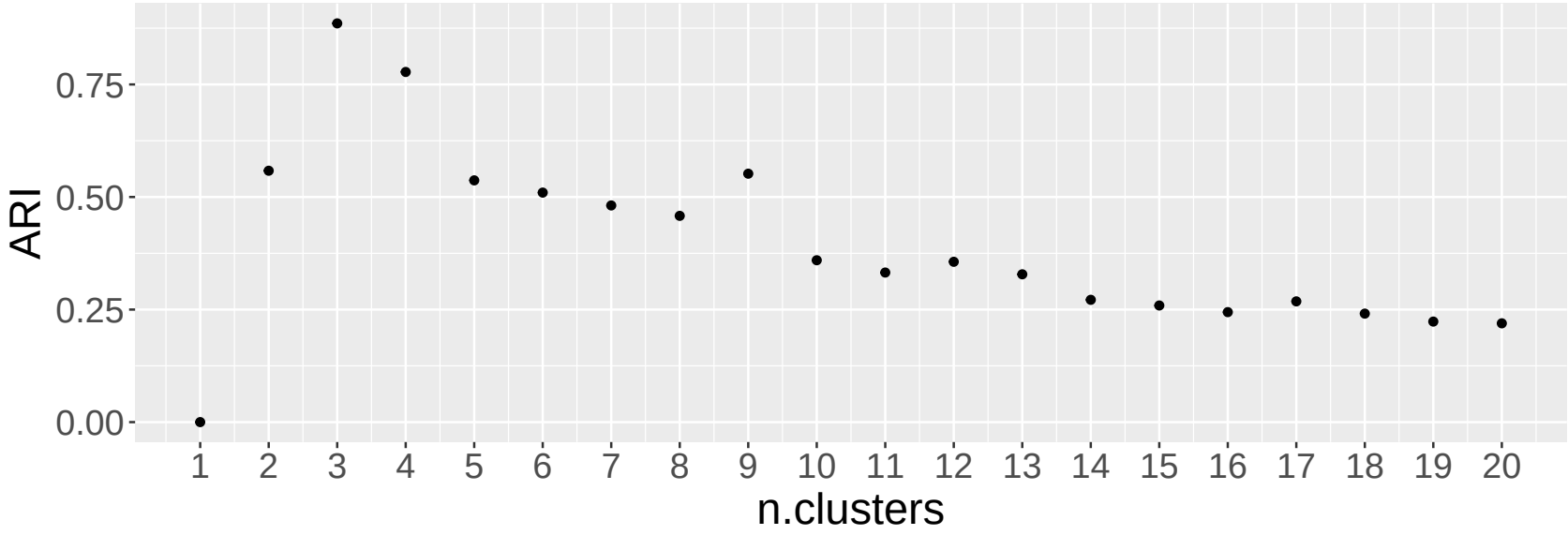
Compare four clusters to labels



Compare random clusters to labels

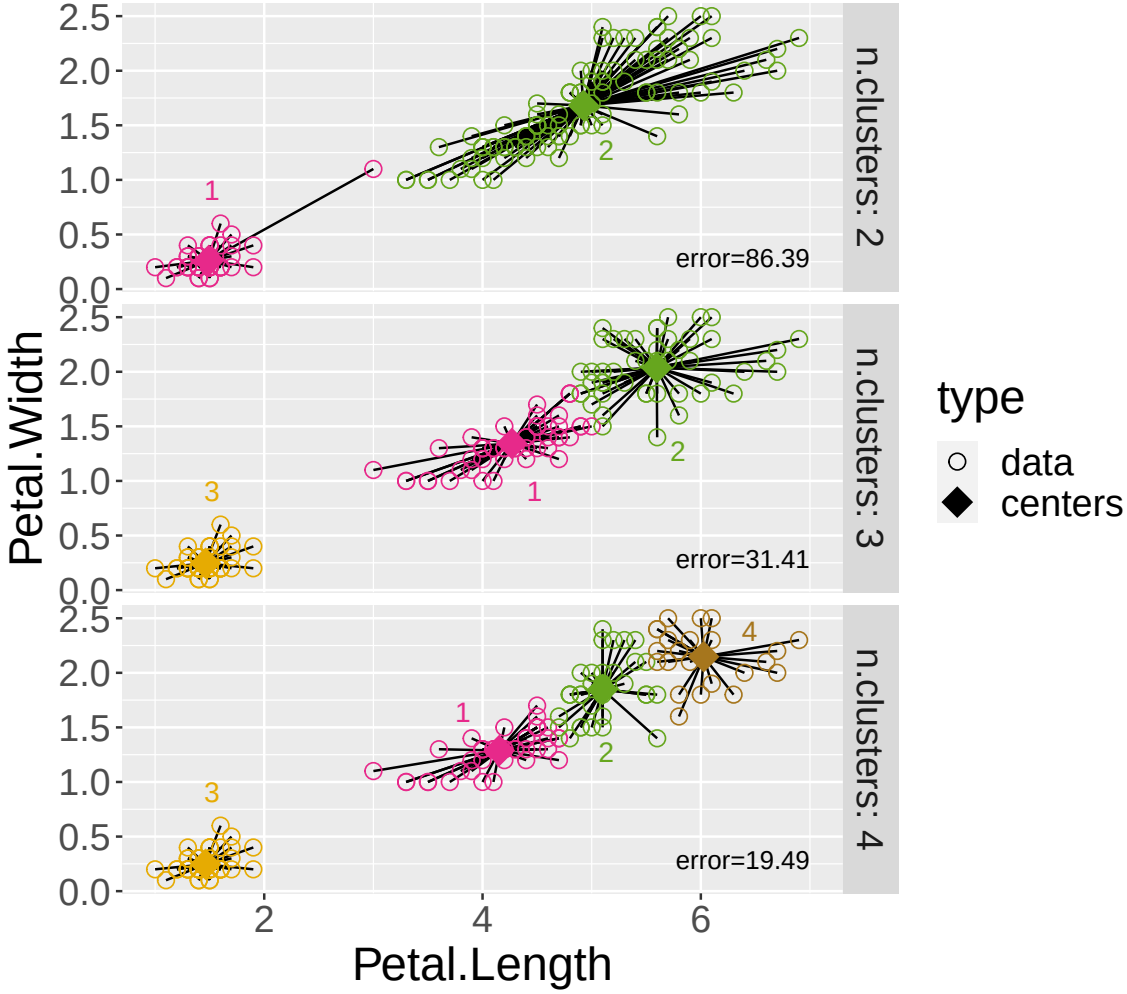


Compute ARI for several clusterings



- Which K is best? Clear peak at 3 clusters, which makes sense since there are three species in these data.
- How to choose number of clusters? We don't have access to labels (here, species) at training time, when we run the clustering algorithm.

Visualization of squared error

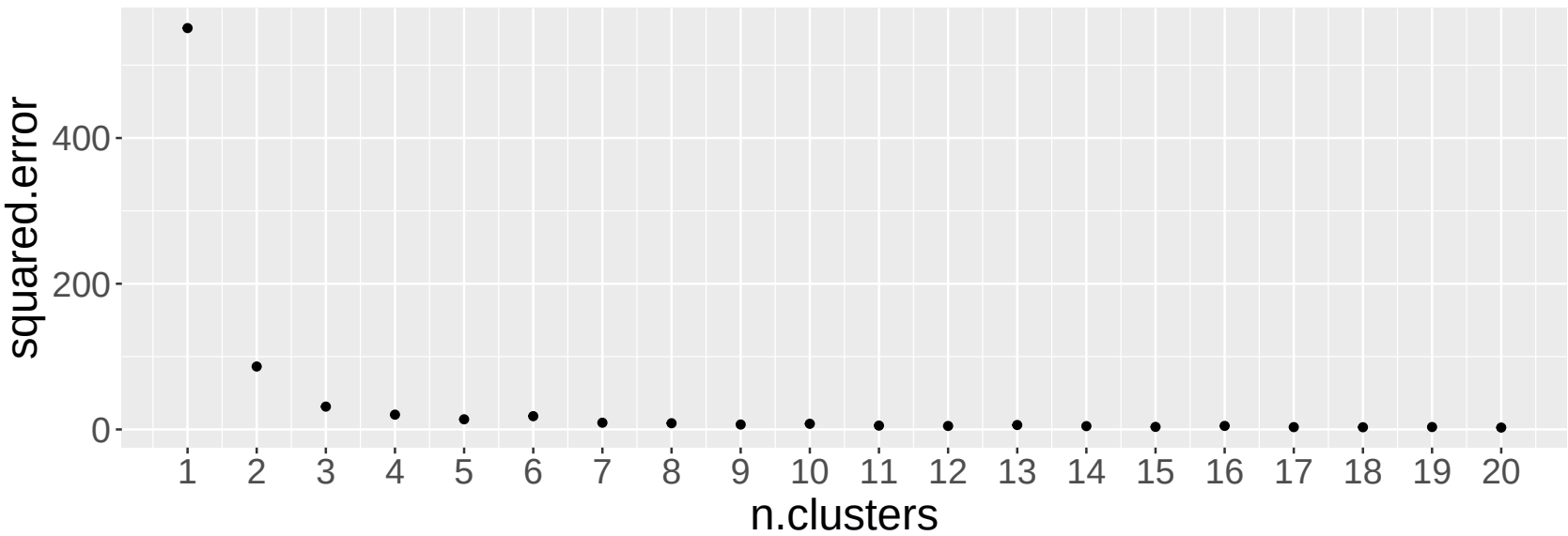


- Black line segments show distance from each data point to its (closest) cluster center.
- This is the distance/error that the K-means algorithm attempts to minimize.

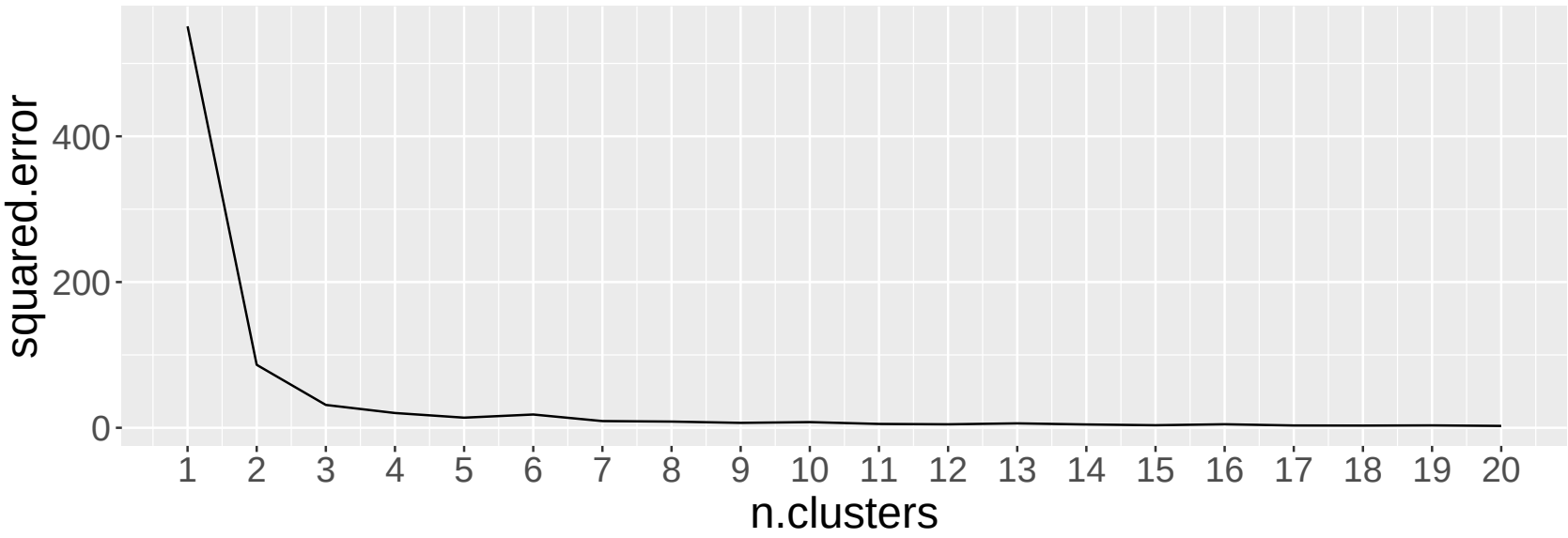
Compute error for several clusterings

- Let $X \in \mathbb{R}^{150\times 2}$ be the data matrix.
- Let K be the number of clusters.
- Let $H \in \mathbb{R}^{150\times K}$ be the matrix which assigns each data point to a cluster (there is a one in every row).
- Let $M \in \mathbb{R}^{K\times 2}$ be the matrix of cluster centers.
- K-means wants to minimize the within-cluster squared error,

$$\min_{H,M} || \underbrace{X}_{\text{data}} - \underbrace{HM}_{\text{center}} ||_2^2$$

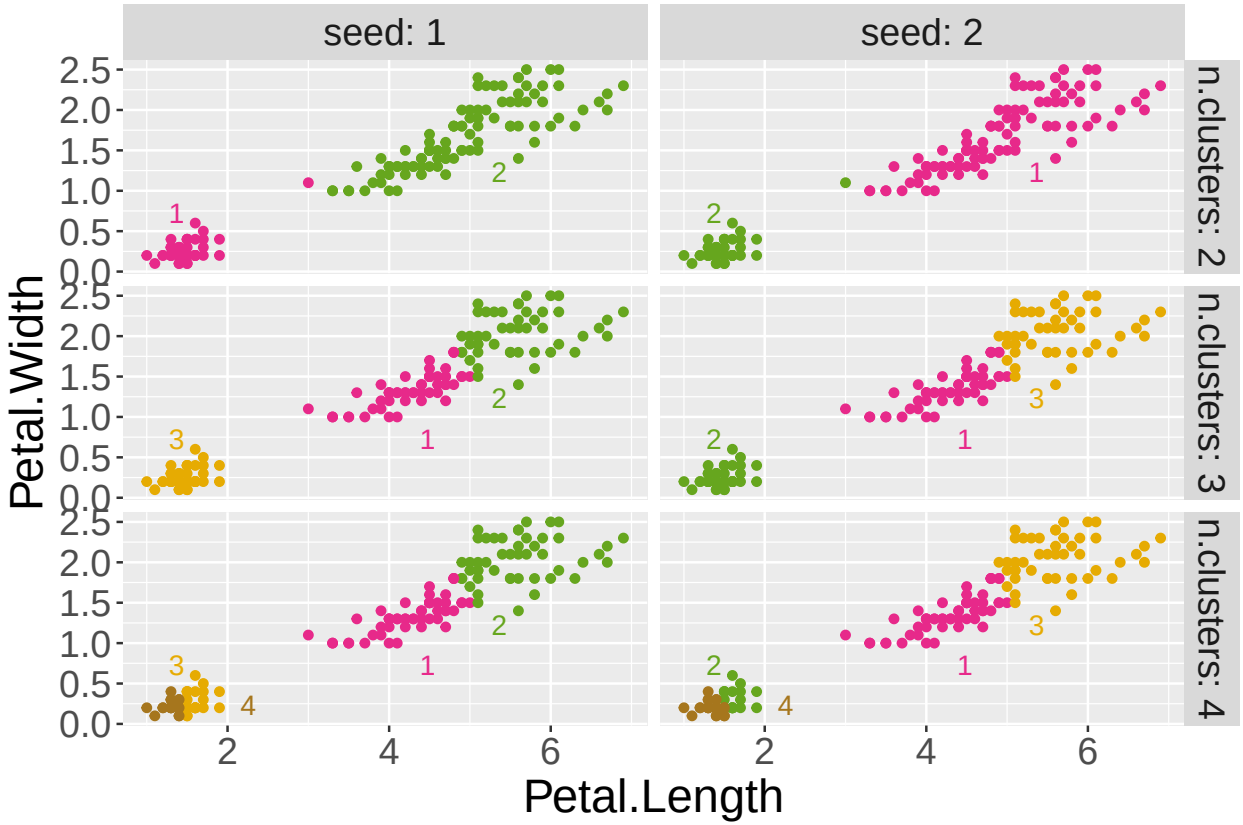


Model selection via error curve analysis



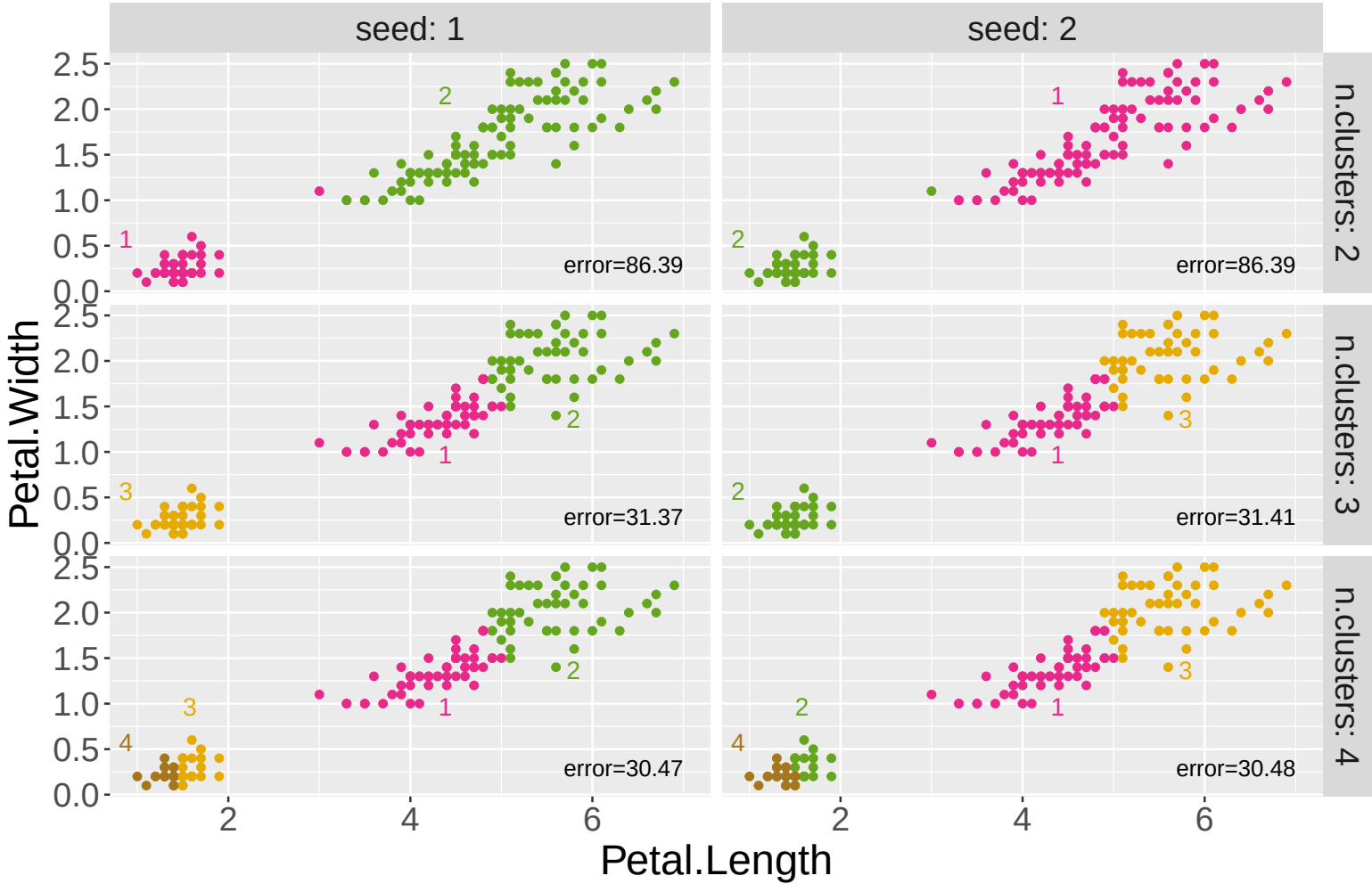
- These error values can be computed using only the input data (labels/outputs are not required).
- The curve stops decreasing rapidly after three clusters.
- In general, for any problem/data set, making this plot and then locating the "kink in the curve" is a good rule of thumb for selecting the number of clusters.

Visualize clusters using two random seeds



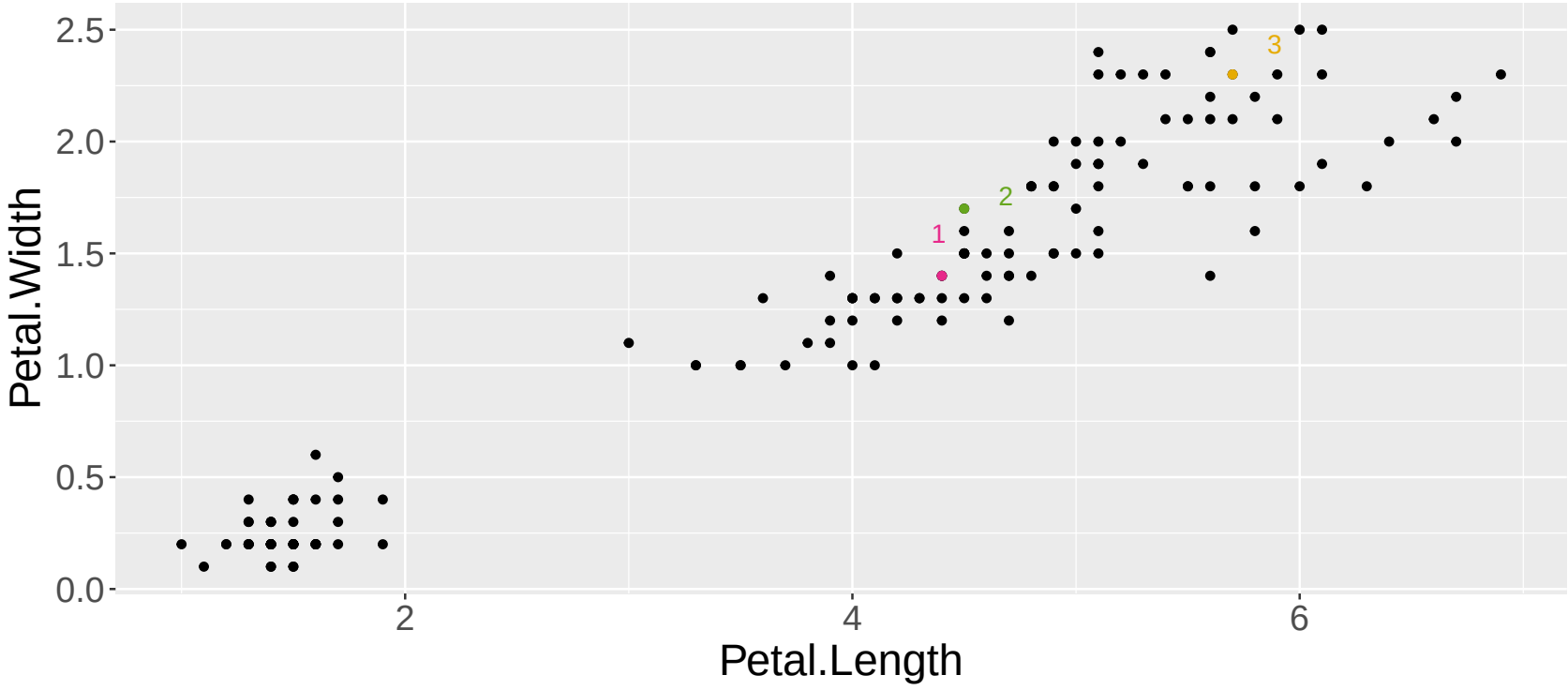
- Goal of K-means is to minimize the squared error.
- Hard non-convex problem due to the 0/1 valued H matrix.
- So not possible to get global (absolute best) minimum in practice. Instead K-means returns a local minimum.
- Result of K-means algorithm, and quality of local minimum, depends on the intialization / random seed.

Choose between seeds using min error

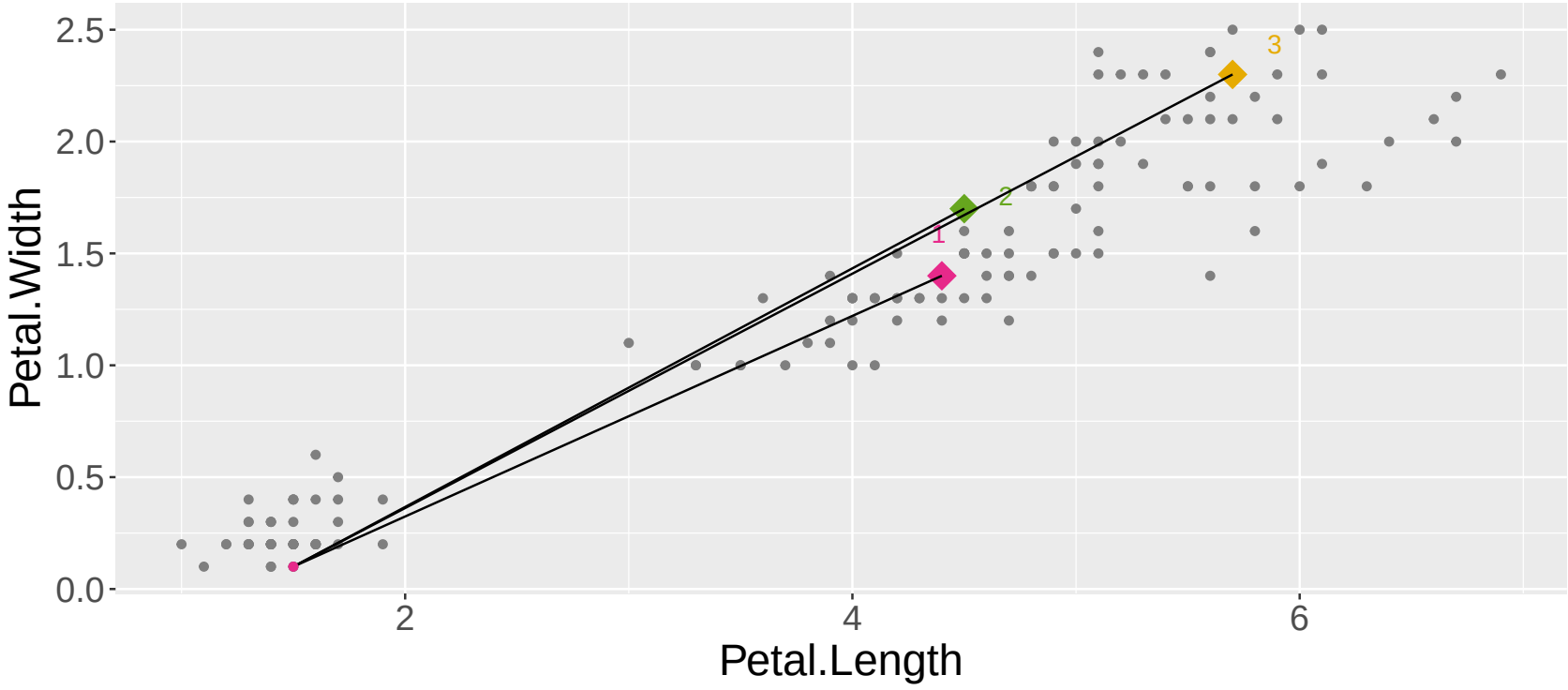


- Try several different random seeds.
- Keep the result with minimum error.

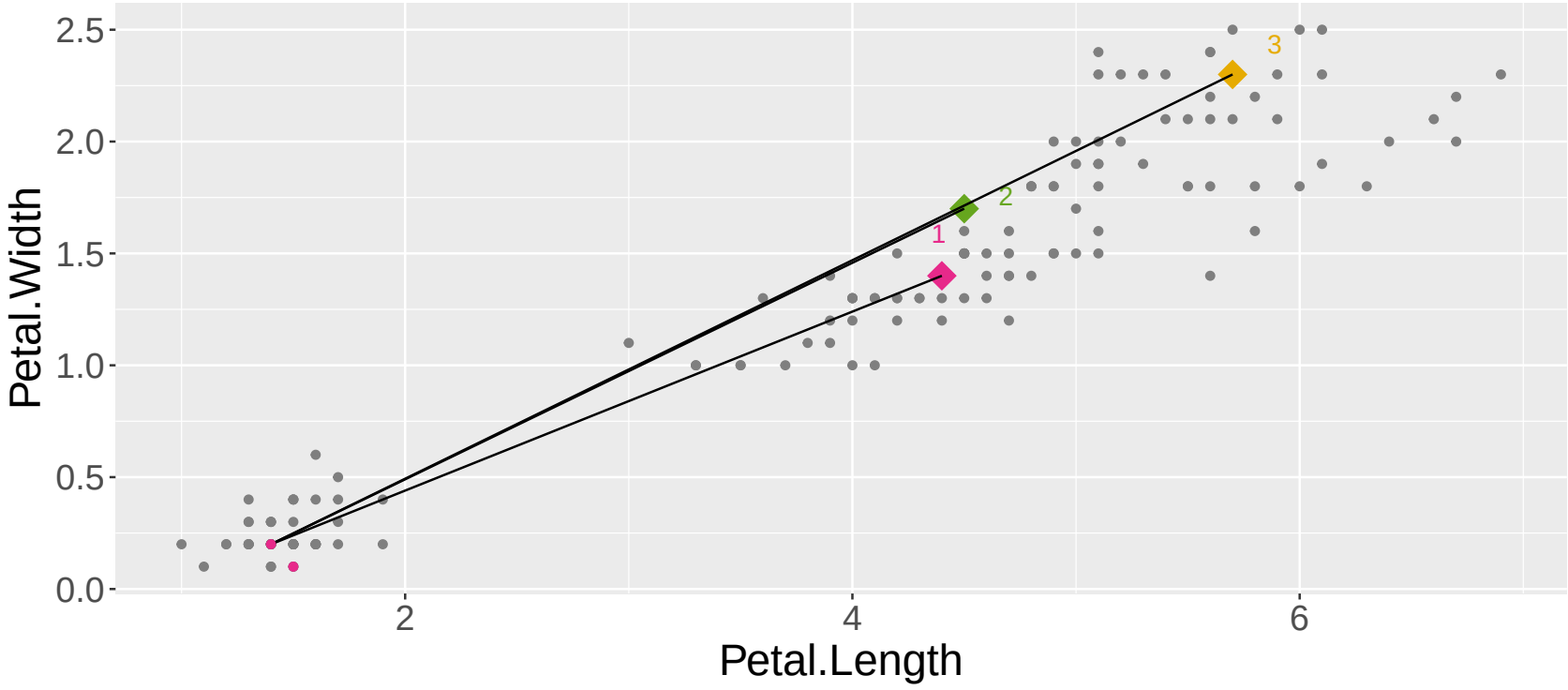
K-means starts with three random cluster centers



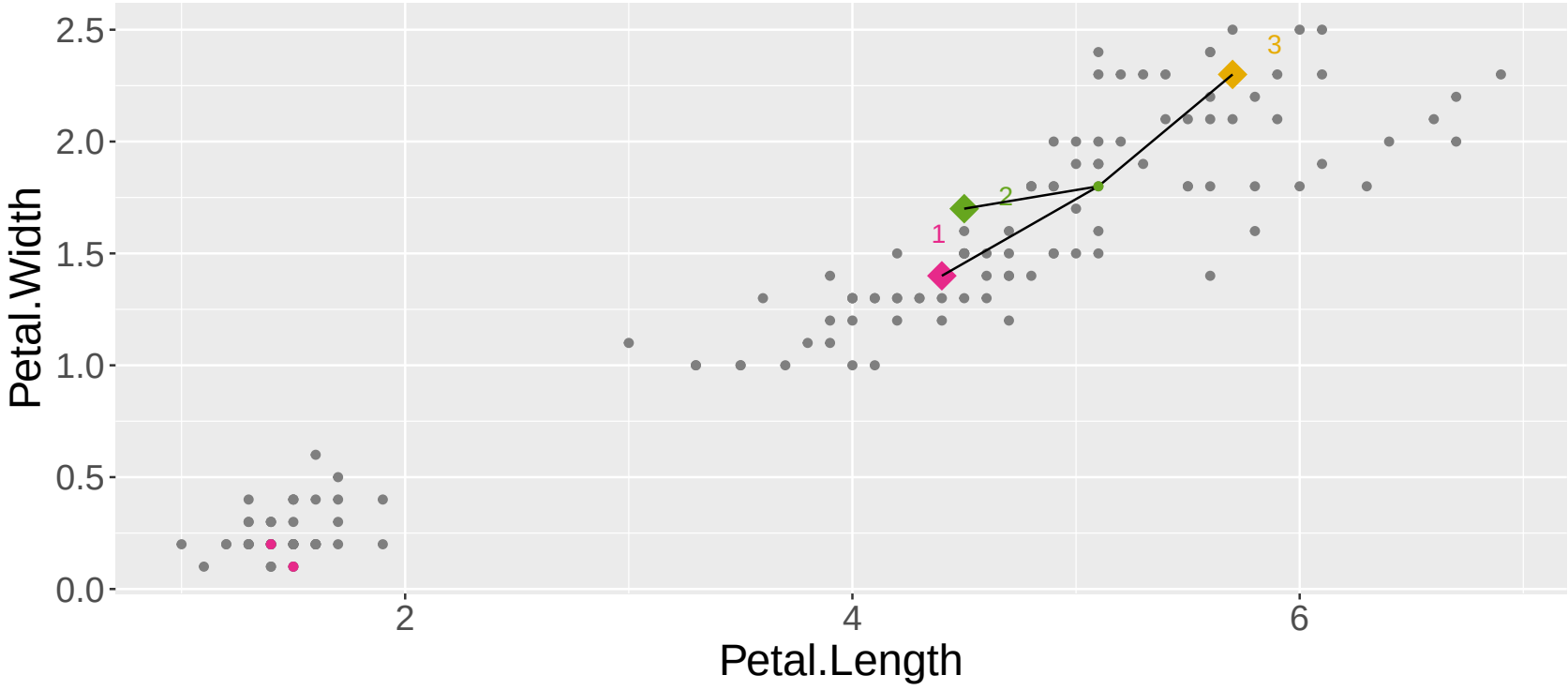
Compute closest cluster center for each data point



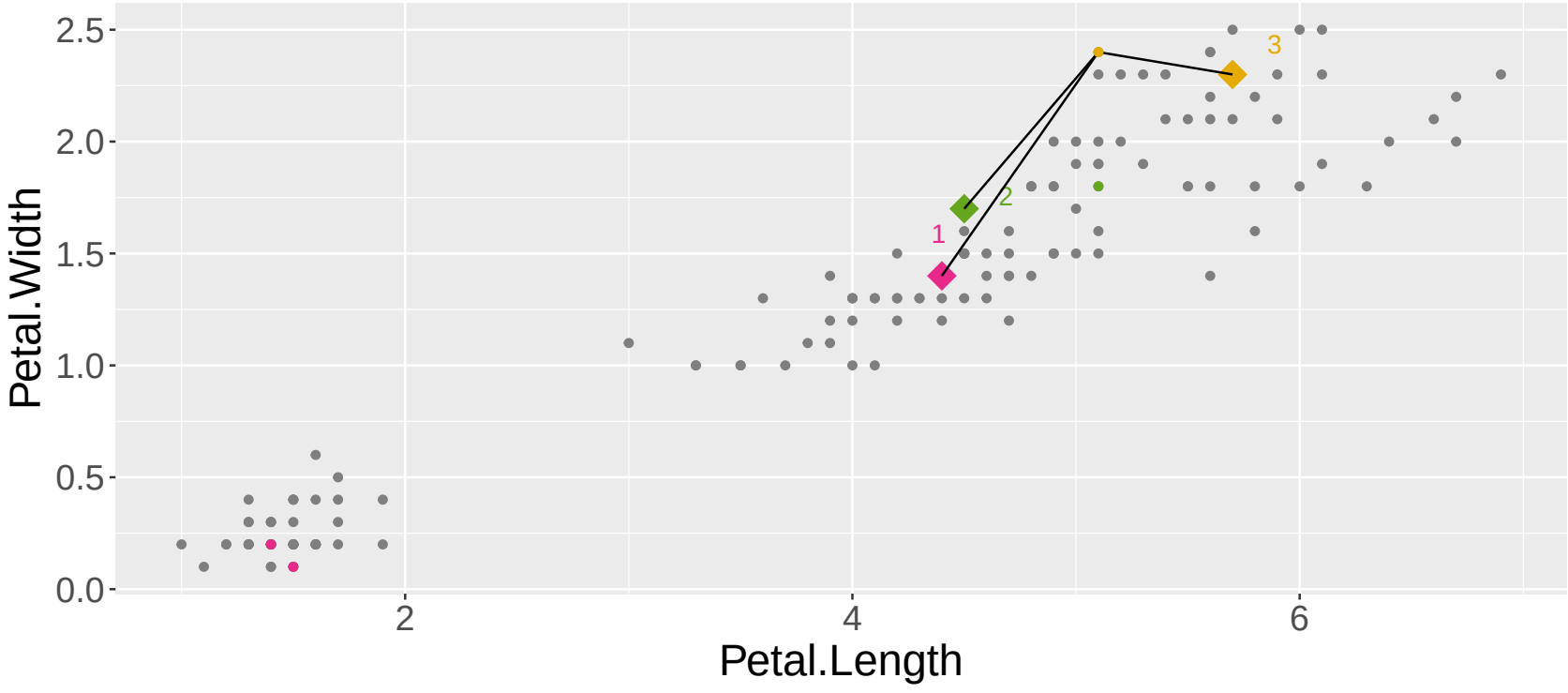
Compute closest cluster center for each data point



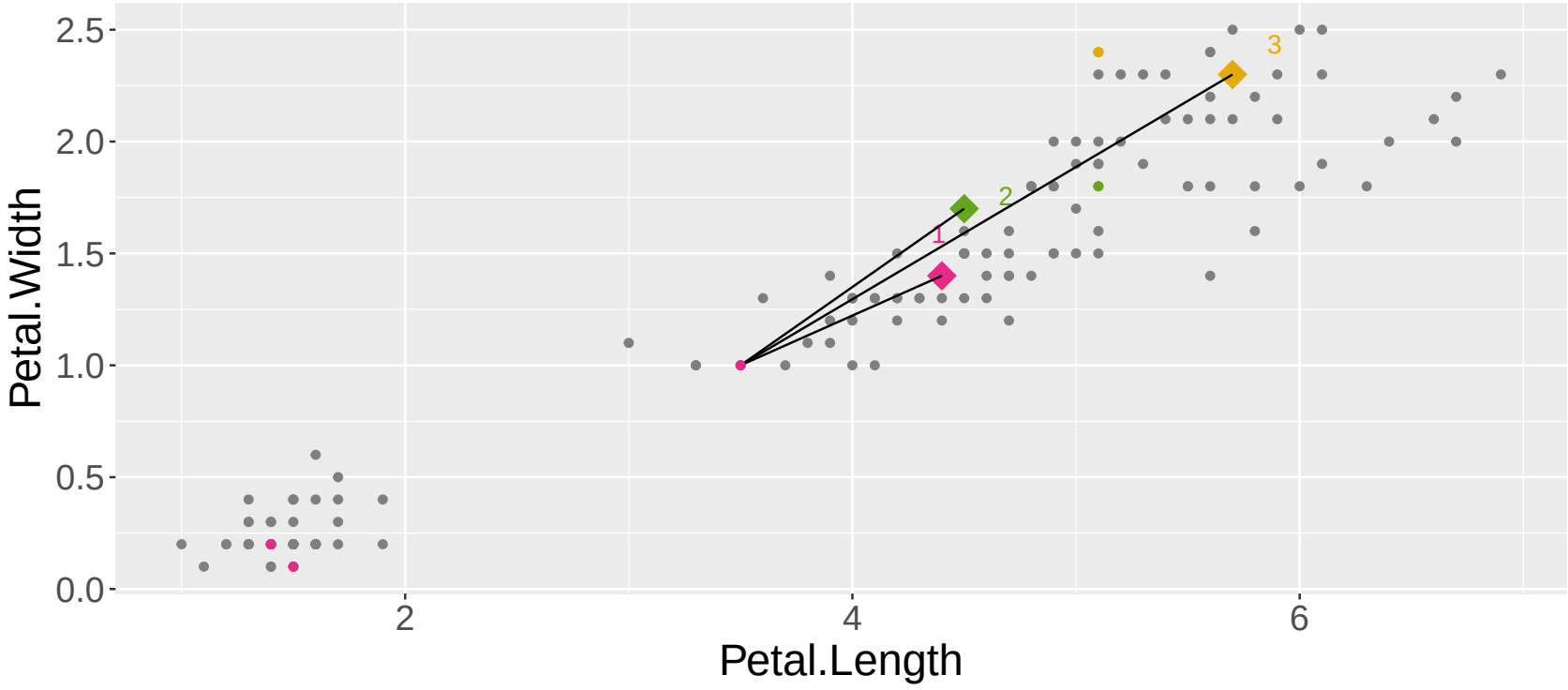
Compute closest cluster center for each data point



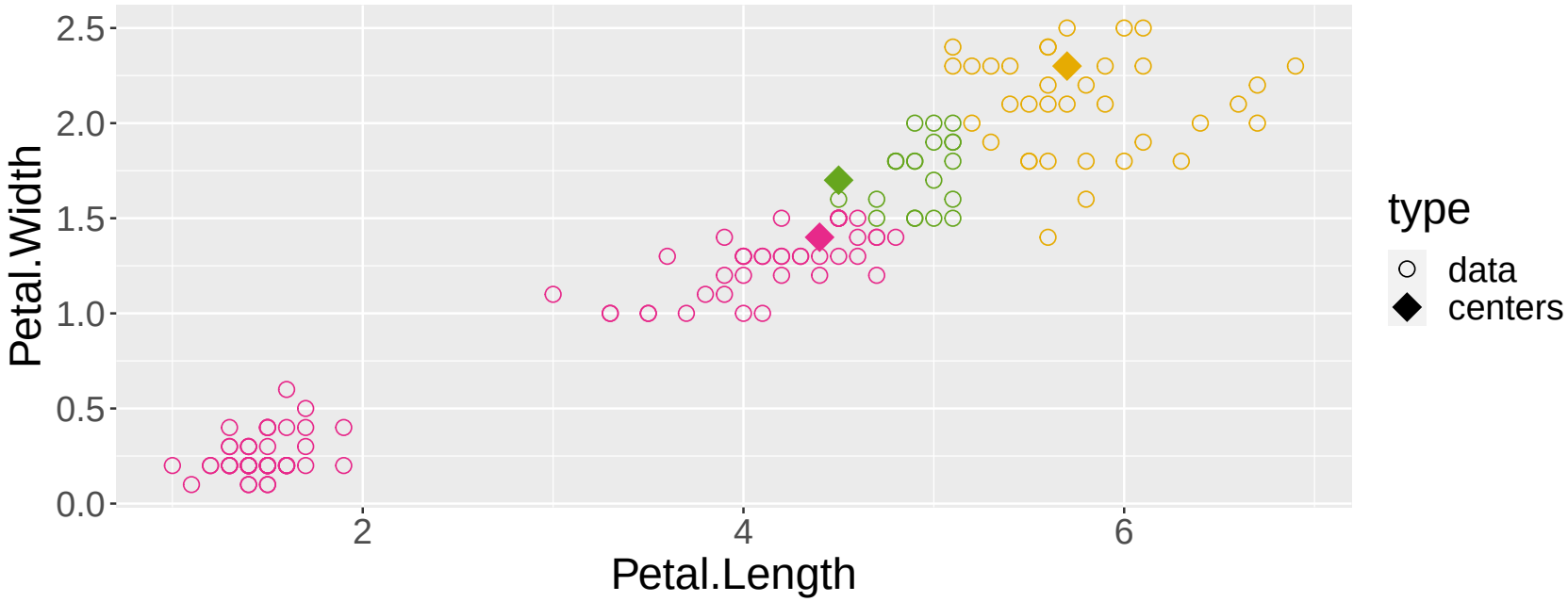
Compute closest cluster center for each data point



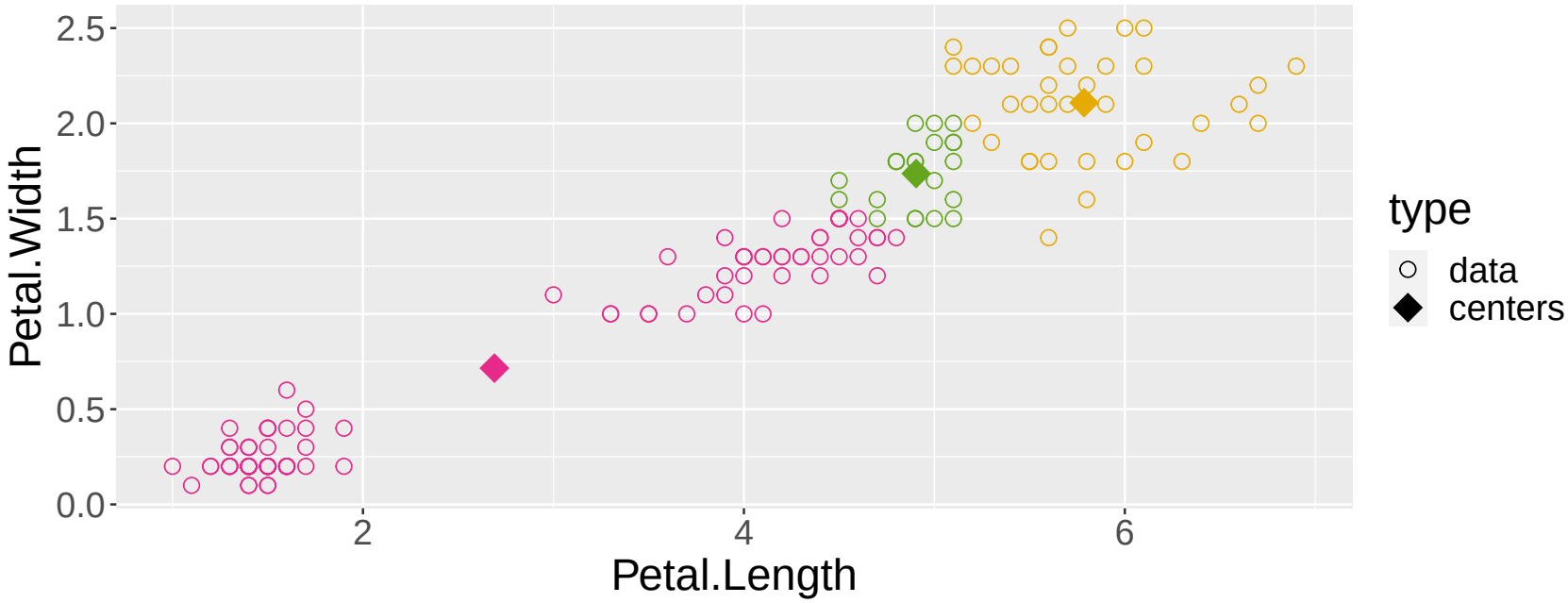
Compute closest cluster center for each data point



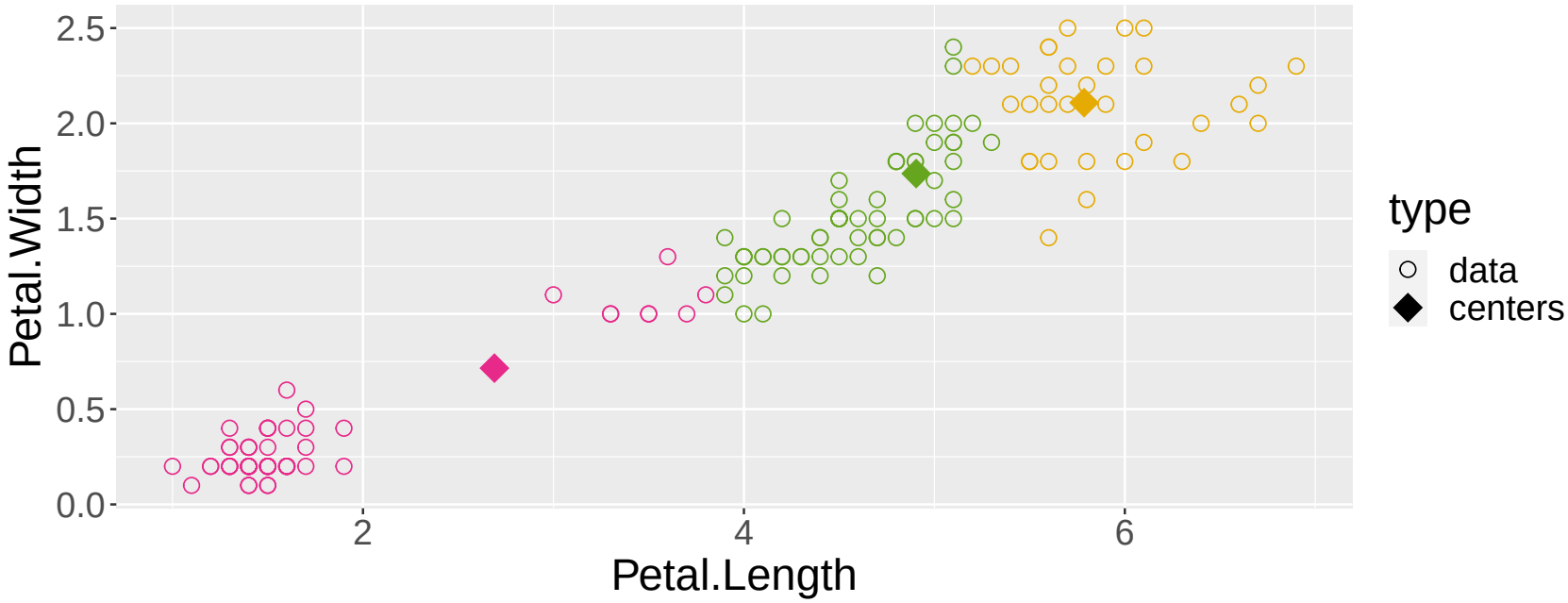
All data points assigned to nearest cluster



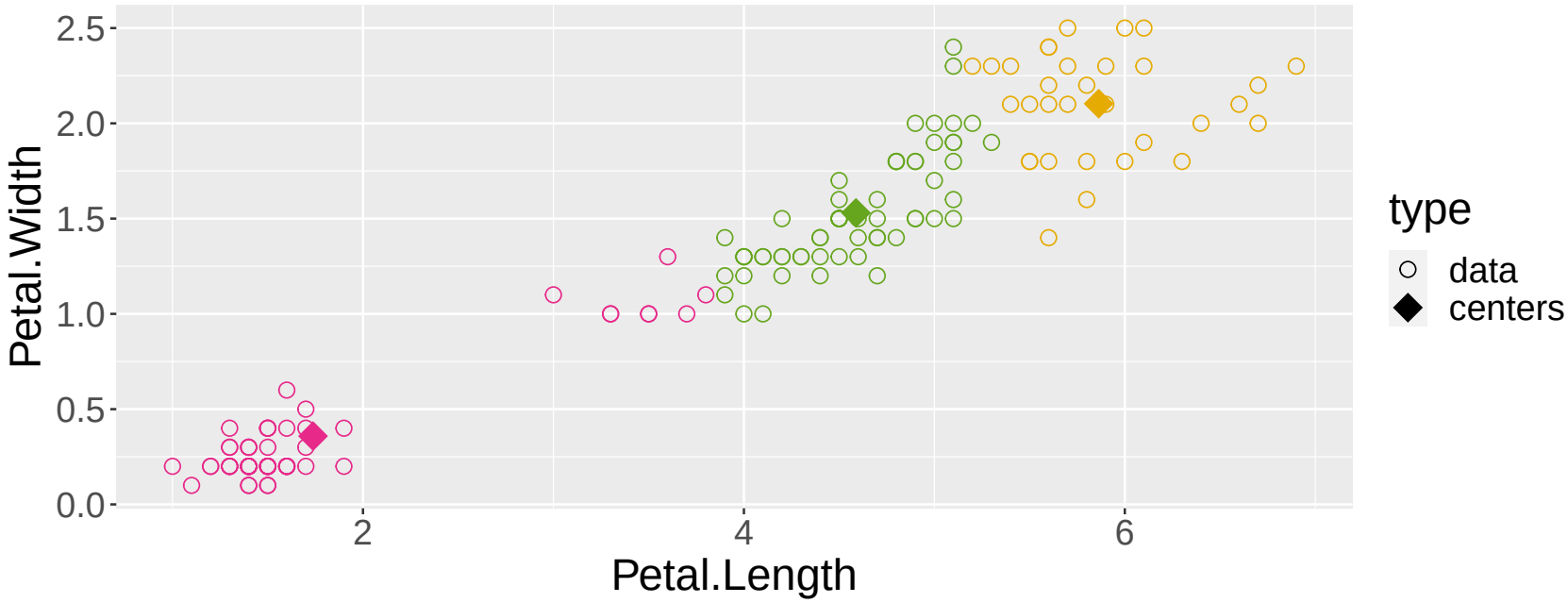
Cluster centers updated



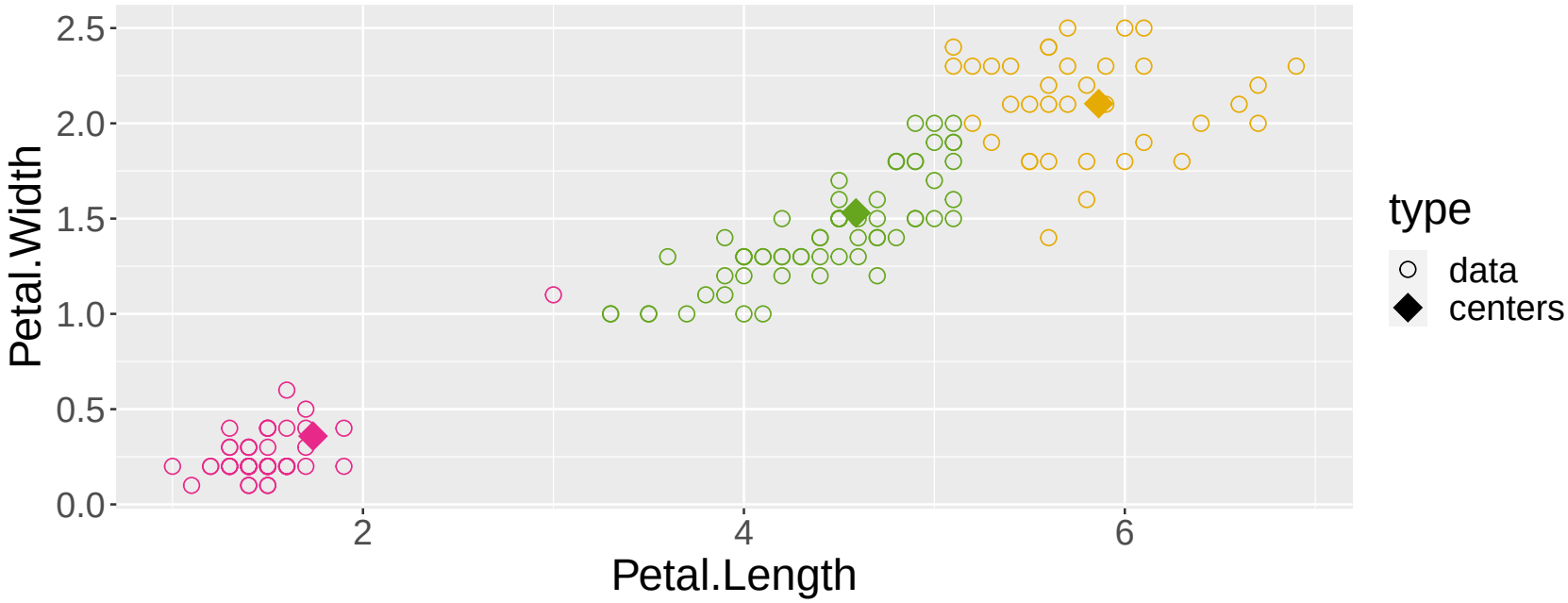
Compute new assignments



Compute new centers



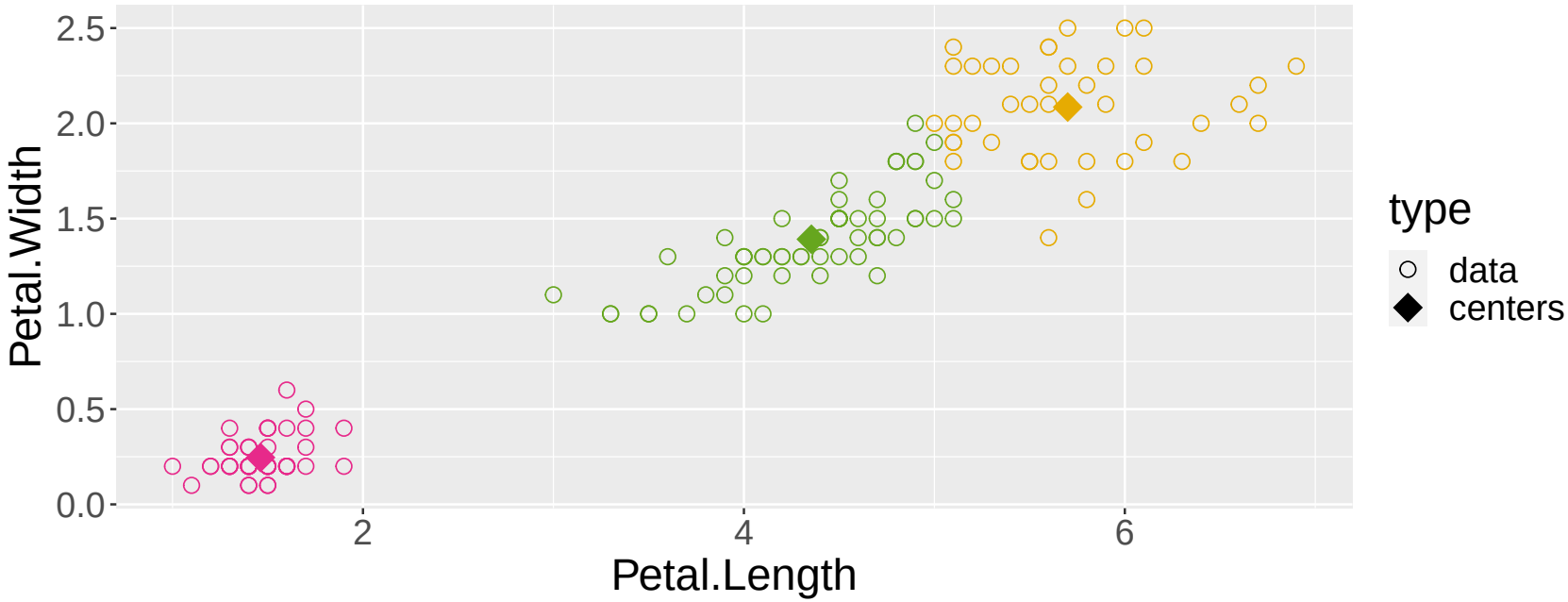
Compute assignments iteration 3



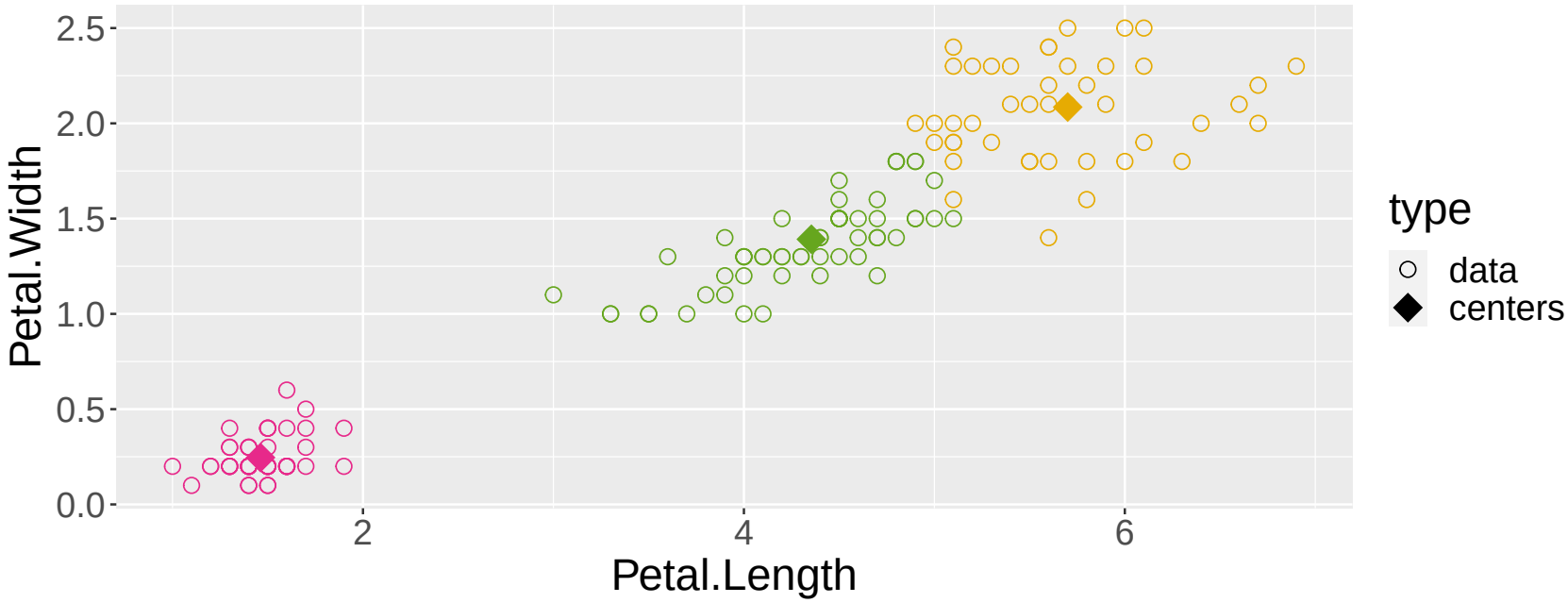
Compute centers iteration 3

Compute assignments iteration 4

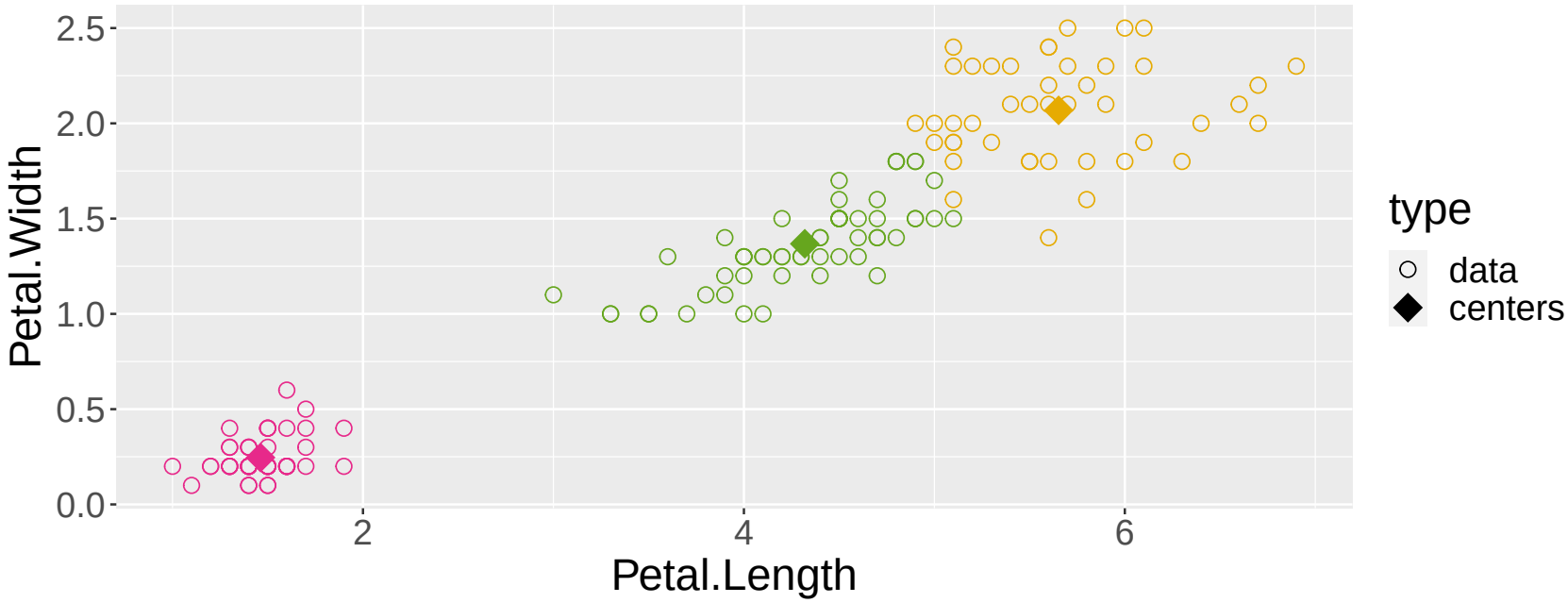
Compute centers iteration 4



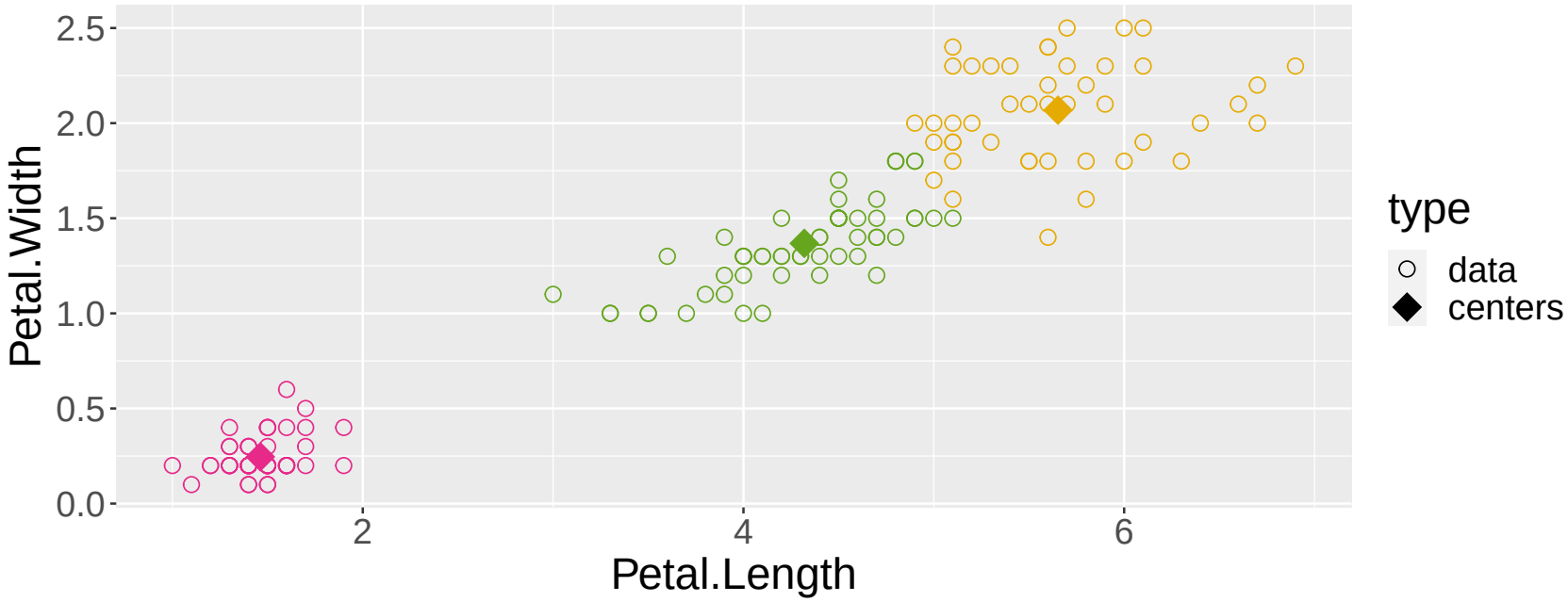
Compute assignments iteration 5



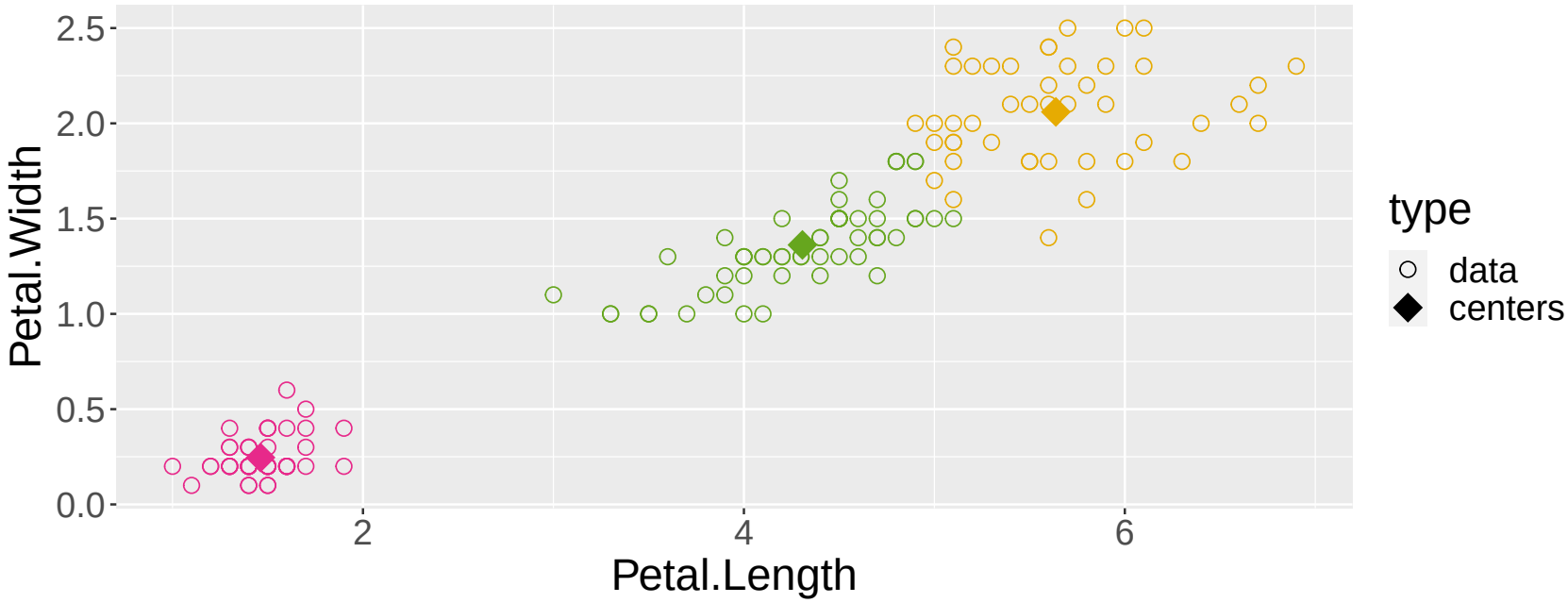
Compute centers iteration 5



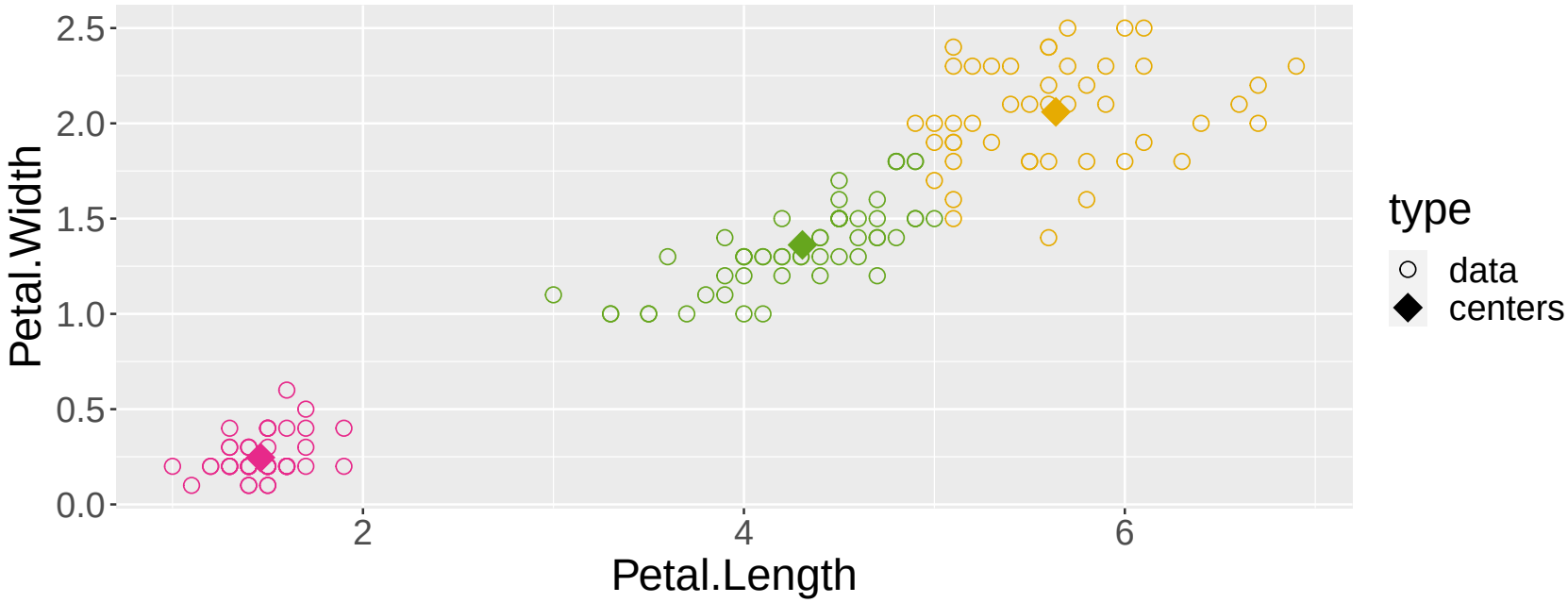
Compute assignments iteration 6



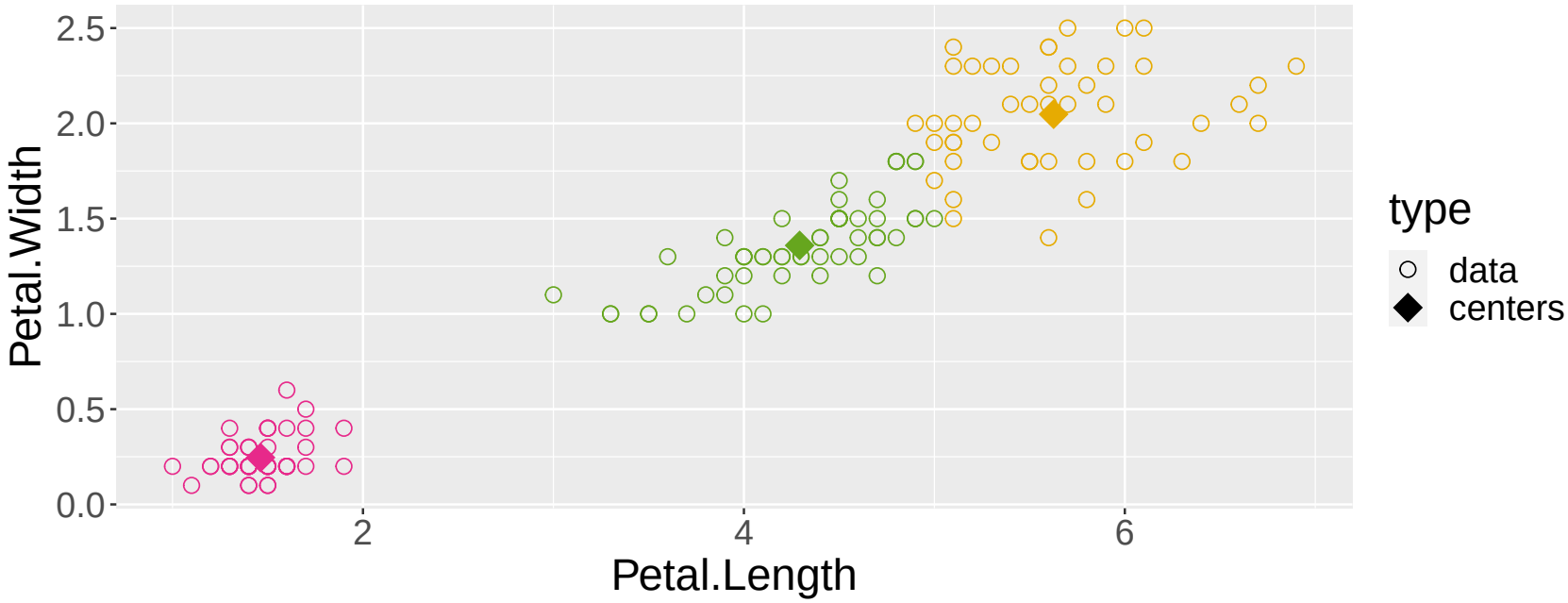
Compute centers iteration 6



Compute assignments iteration 7



Compute centers iteration 7



Compute assignments iteration 8 (no change = stop)