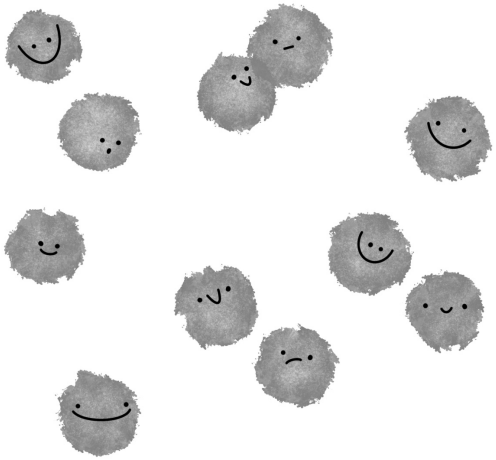
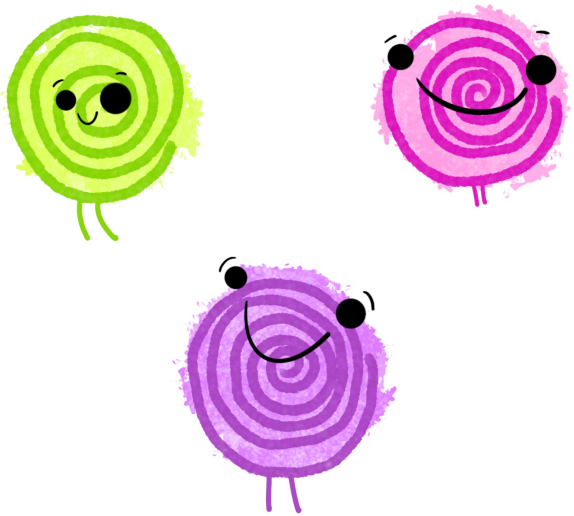
k-means clustering

: assign each observation to one of k clusters based on the nearest cluster centroid.

OBSERVATIONS

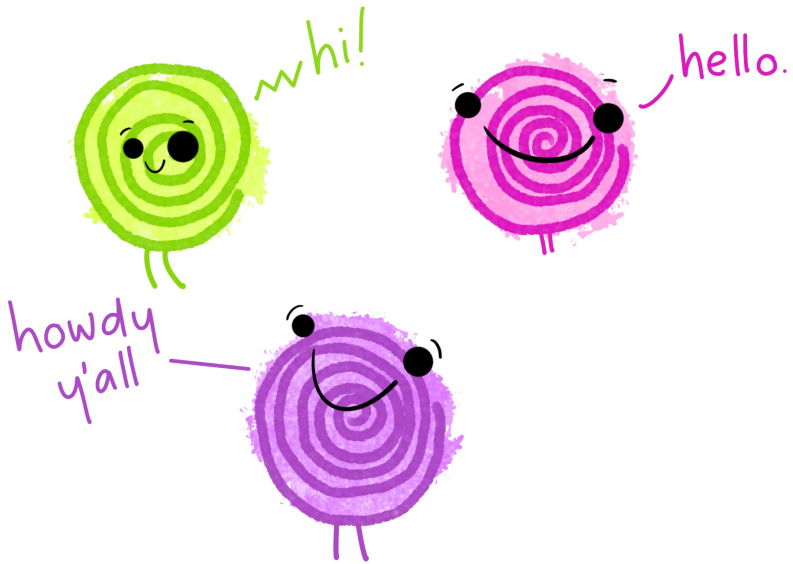


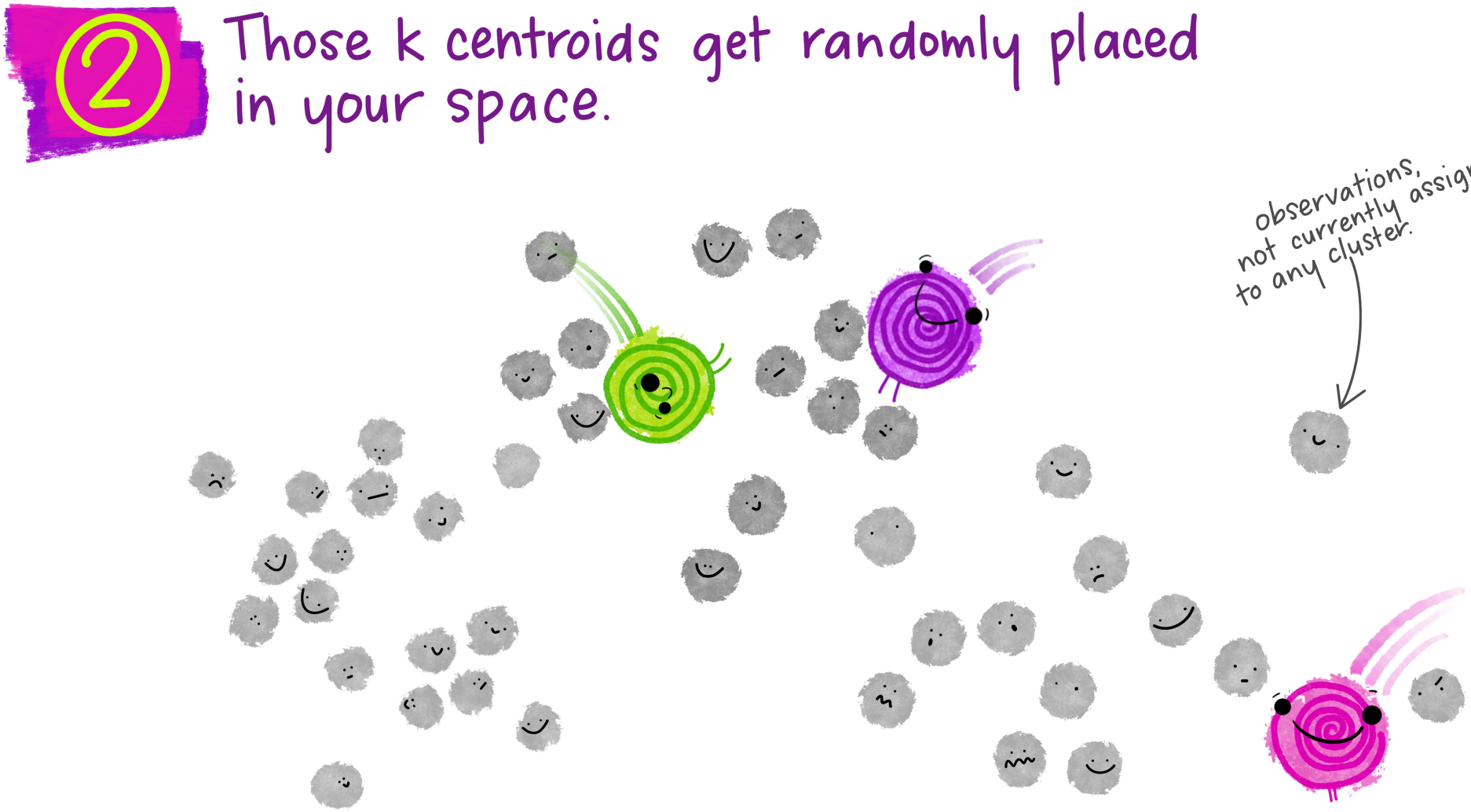
cluster CENTROIDS



1 Specify the number of clusters (in this example, k = 3).

Then imagine k cluster centroids are created.




2
Those k centroids get randomly placed in your space.
observations, not currently assigned to any cluster.
@allison_horst

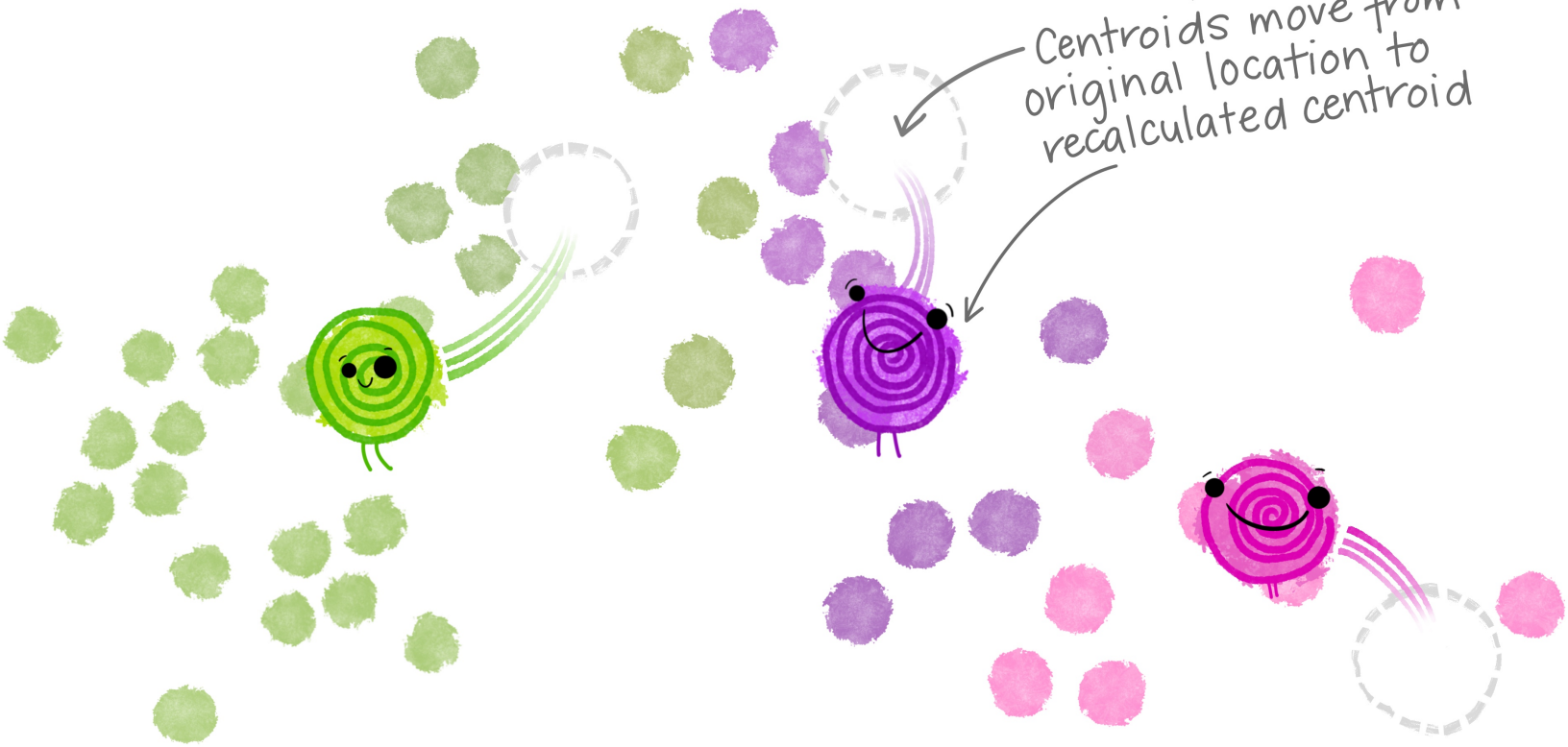
Each observation gets temporarily "assigned" to its closest centroid.

(e.g. by Euclidean distance)

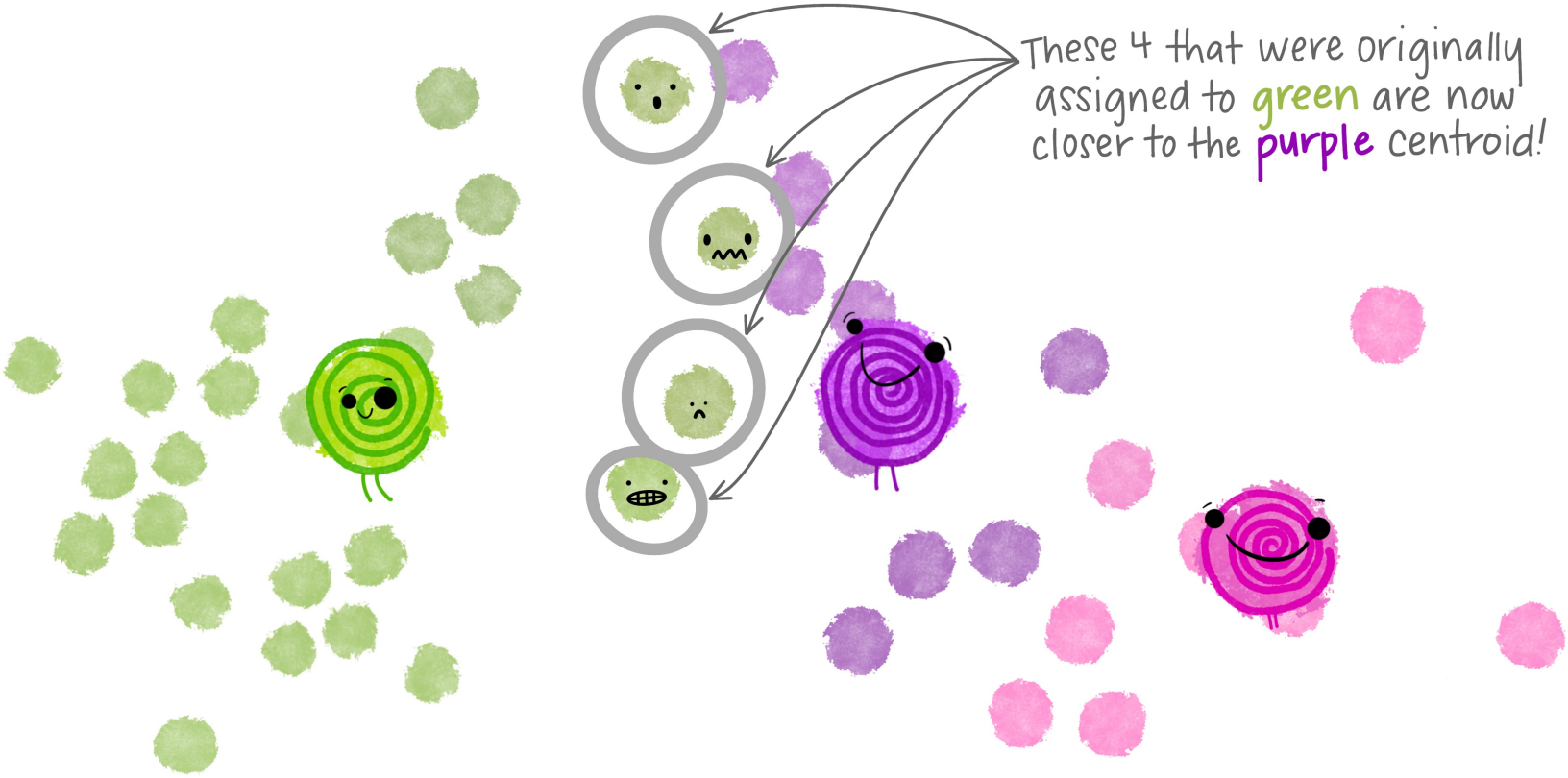
Note: observations stay put, but their color updates to nearest centroid!

4

Then the centroid of each cluster is calculated based on all observations assigned to that cluster...



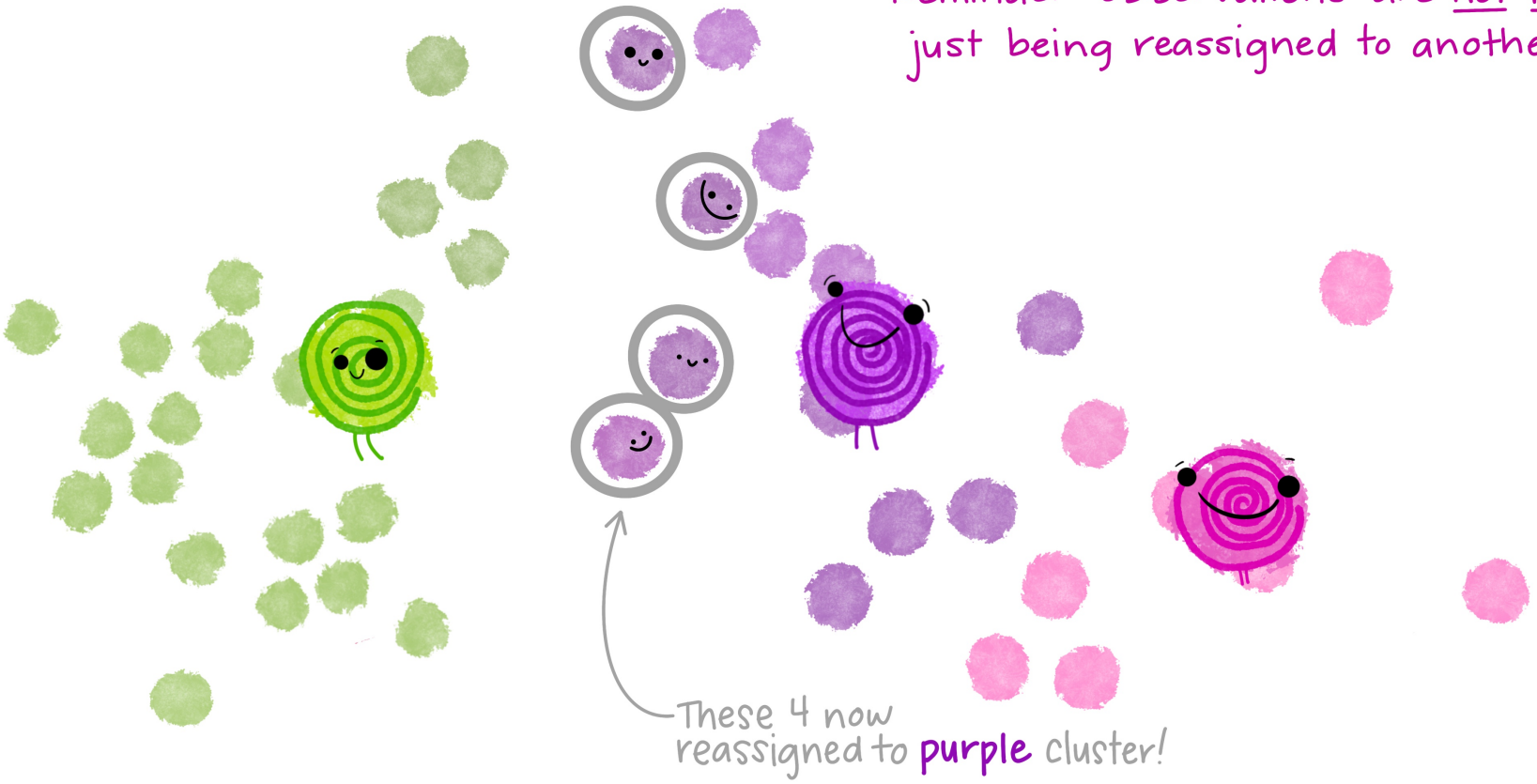
UH OH. Now that the cluster centroids have moved, some of the observations are now closer to a different centroid!



5

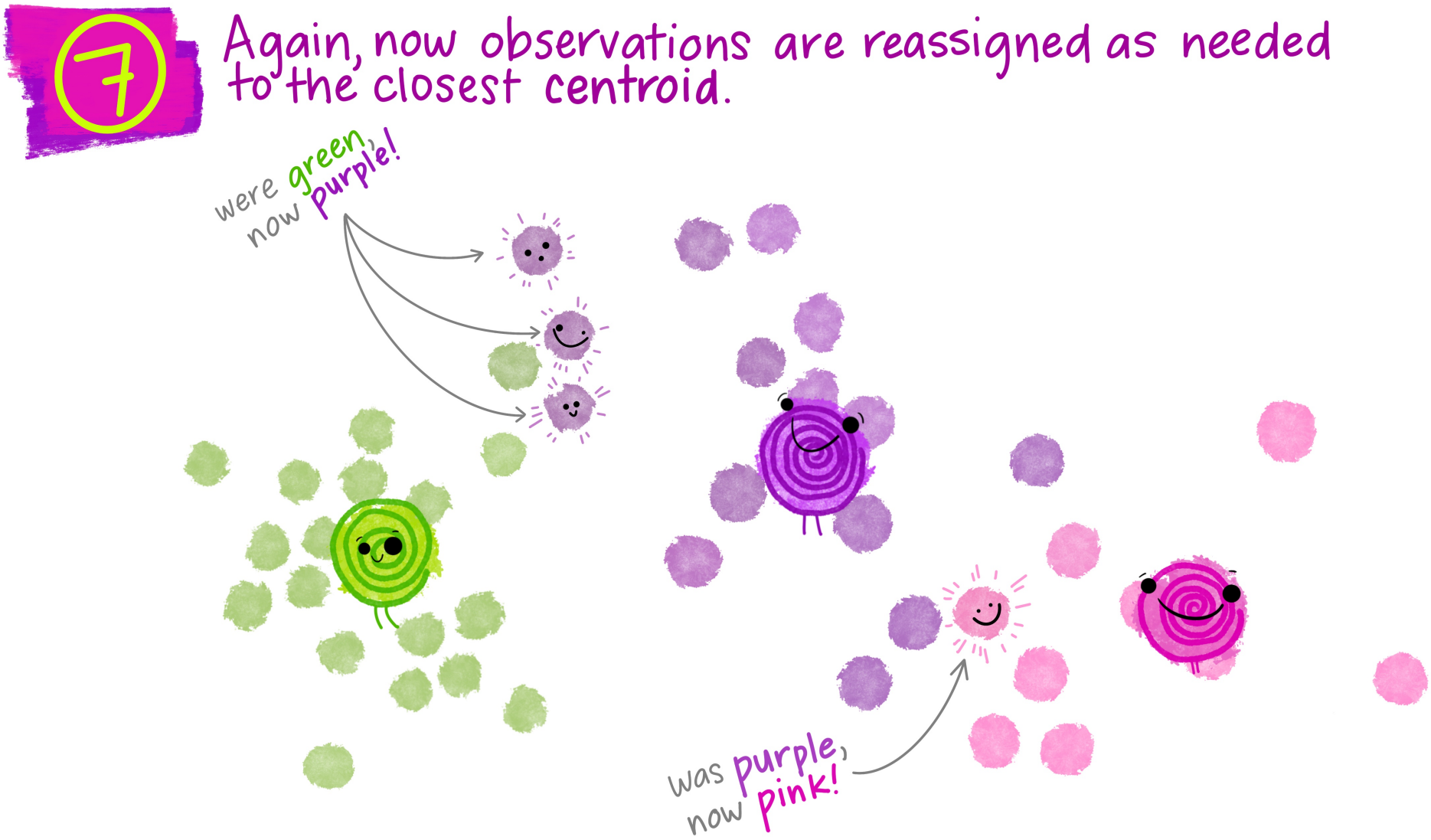
NO PROBLEM!
Observations get reassigned* to a different cluster based on the recalculated centroid.

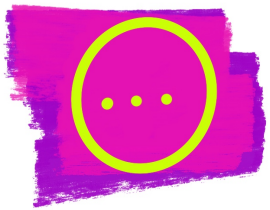
*Reminder: observations are not moving, just being reassigned to another cluster.



6

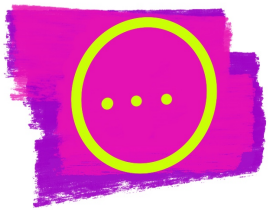
But now that observations have been reassigned, the centroids need to move again [recalculate centroids from updated clusters]


7
Again, now observations are reassigned as needed to the closest centroid.
were green, now purple!
was purple, now pink!
@allison_horst



Then the centroid for each cluster is recalculated...

...which means observations will be reassigned...



That iterative process of

Recalculate cluster centroids
↳ Reassign observations to nearest centroid
↳ Recalculate cluster centroids
↳ Reassign observations to nearest centroid
↳ Recalculate cluster centroids
↳ Reassign observations to nearest centroid

↓

Continues until nothing is moving or being reassigned anymore!

Which means the iteration is done and each observation is assigned to its final cluster.