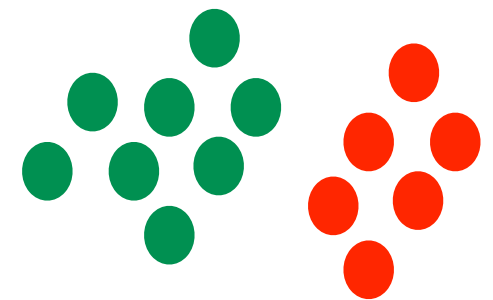
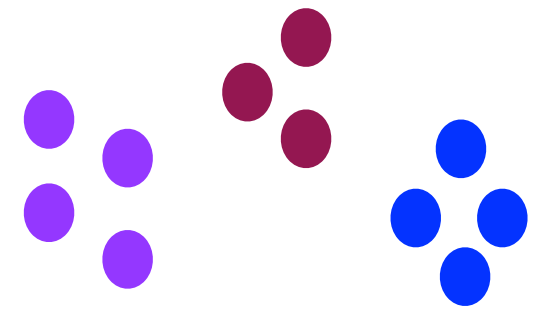
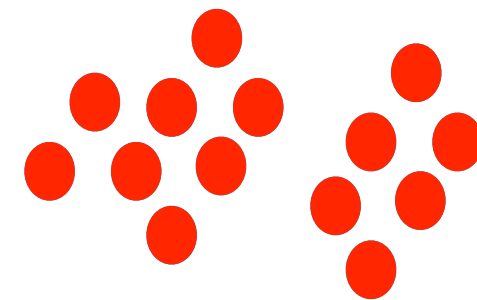
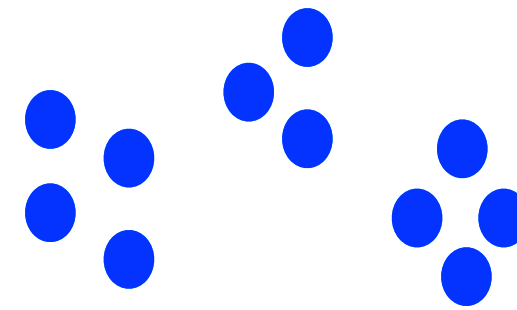


A SET OF POINTS - EACH POINT REPRESENTS A FACEBOOK USER

HERE'S ANOTHER WAY ..



HERE'S ONE WAY OF DOING THIS..



THERE ARE DIFFERENT WAYS TO DEFINE SIMILARITY. FOR EXAMPLE -

- 1. THEY LIKE/FOLLOW THE SAME THINGS**
- 2. THEY ARE FROM THE SAME STATE**
- 3. BOTH OF THE ABOVE**

THE OBJECTIVE OF CLUSTERING IS TO DIVIDE UP THESE USERS INTO GROUPS

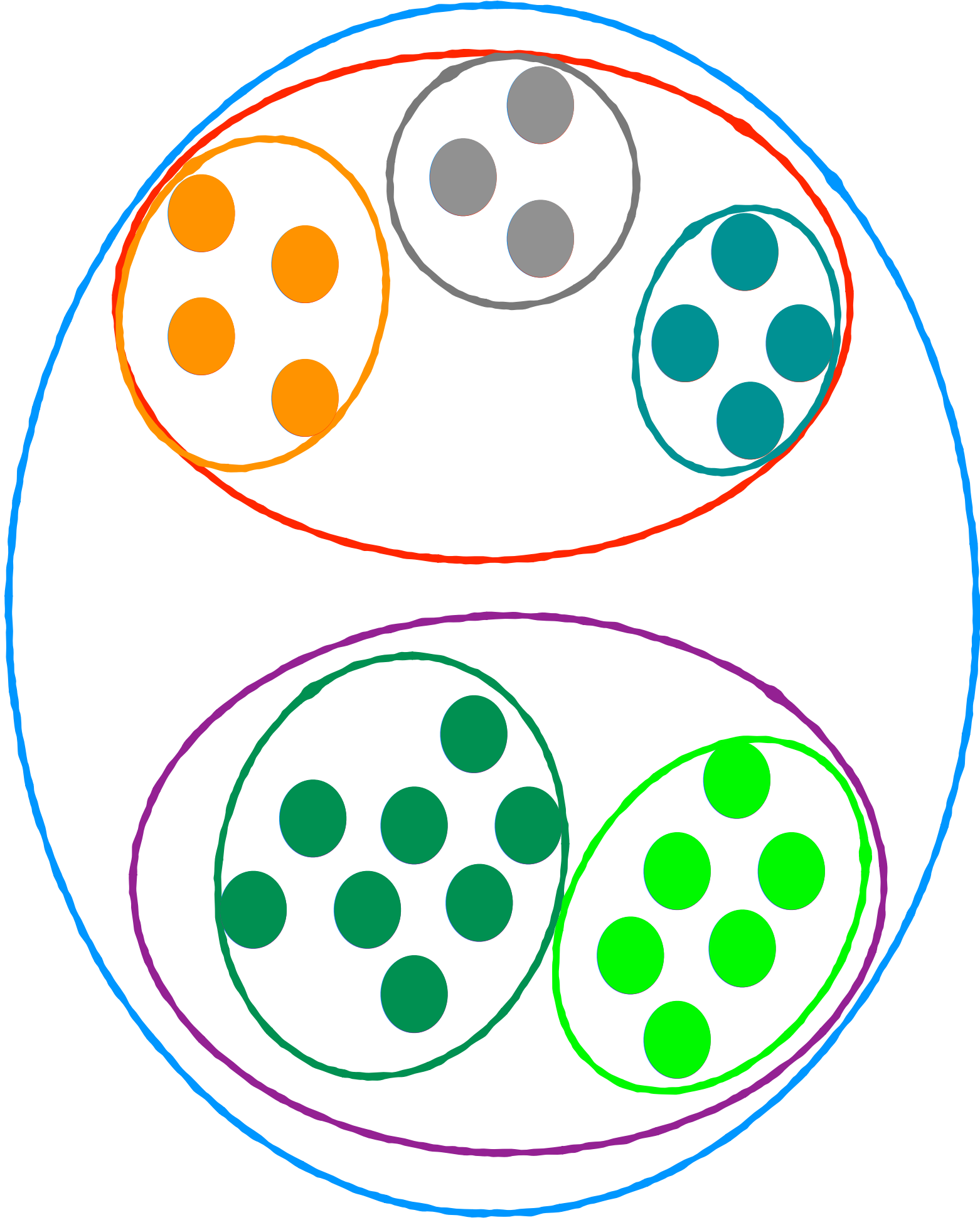
USERS IN A GROUP ARE SIMILAR TO ONE ANOTHER

USERS FROM DIFFERENT GROUPS ARE VERY DIFFERENT FROM EACH OTHER

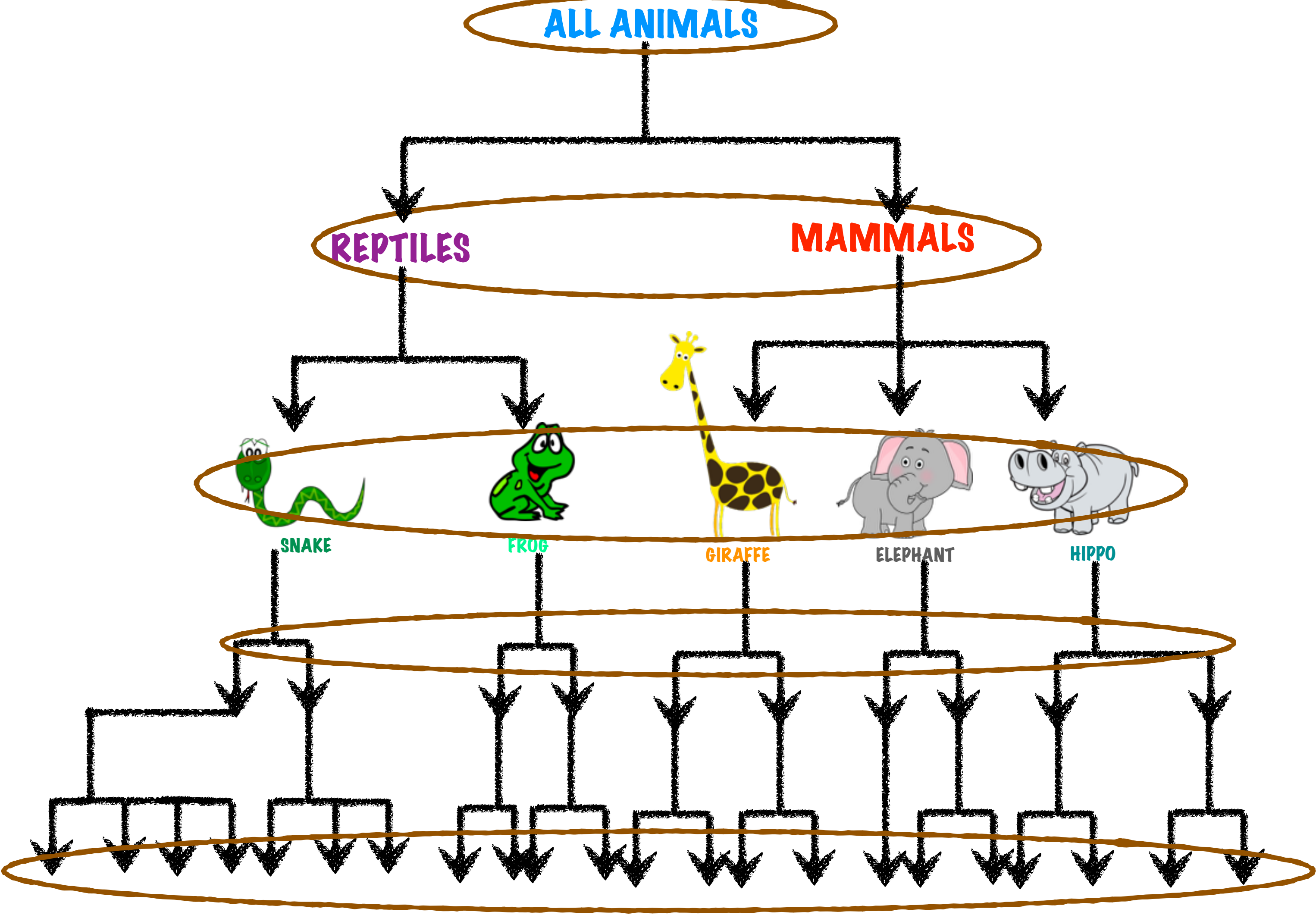
MAXIMIZE INTRACLUSTER SIMILARITY

MINIMIZE INTERCLUSTER SIMILARITY

HIERARCHICAL CLUSTERING



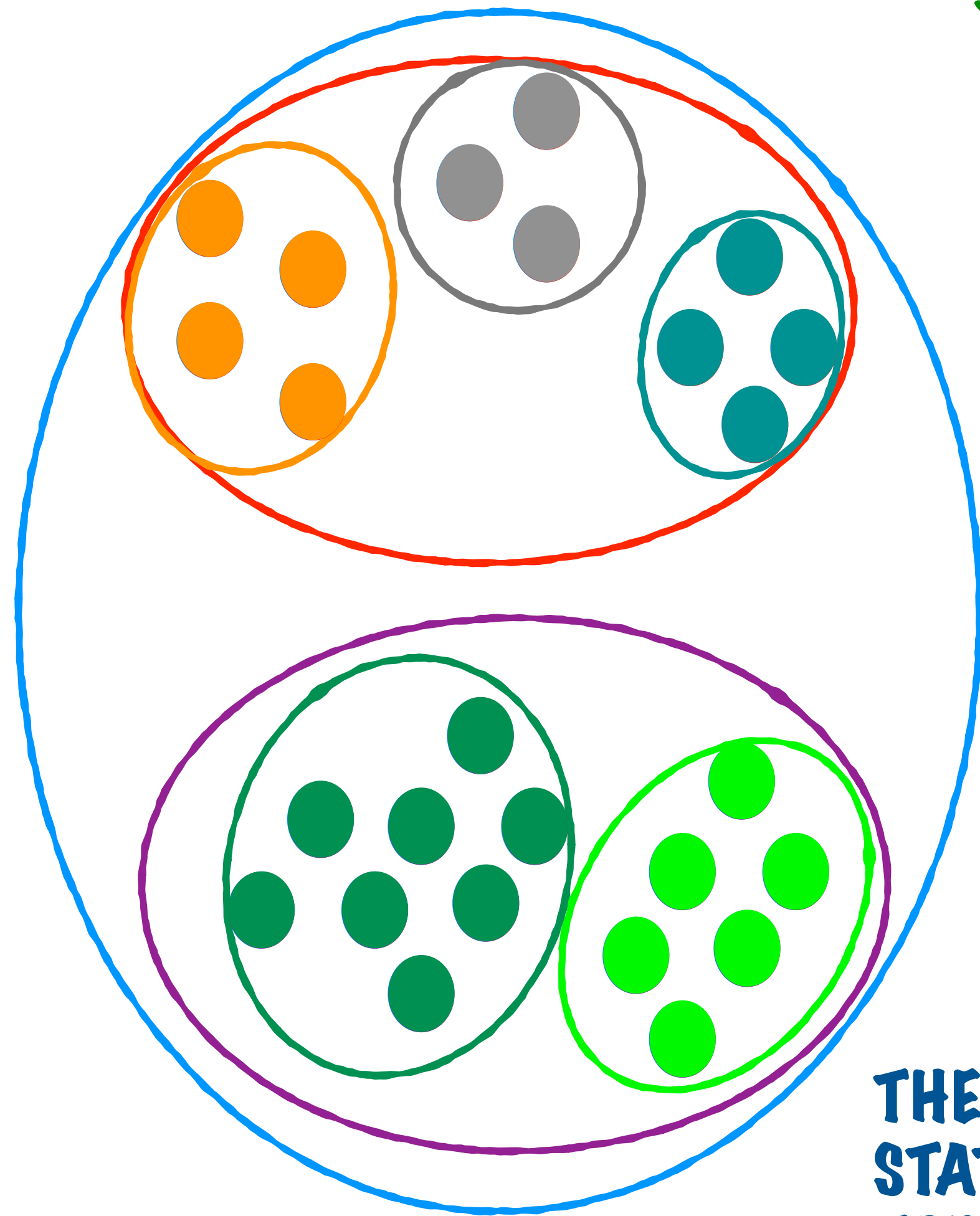
A CROSS SECTION AT EACH LEVEL OF THE HIERARCHY IS A POSSIBLE SET OF CLUSTERS



EACH INDIVIDUAL ANIMAL IS A CLUSTER

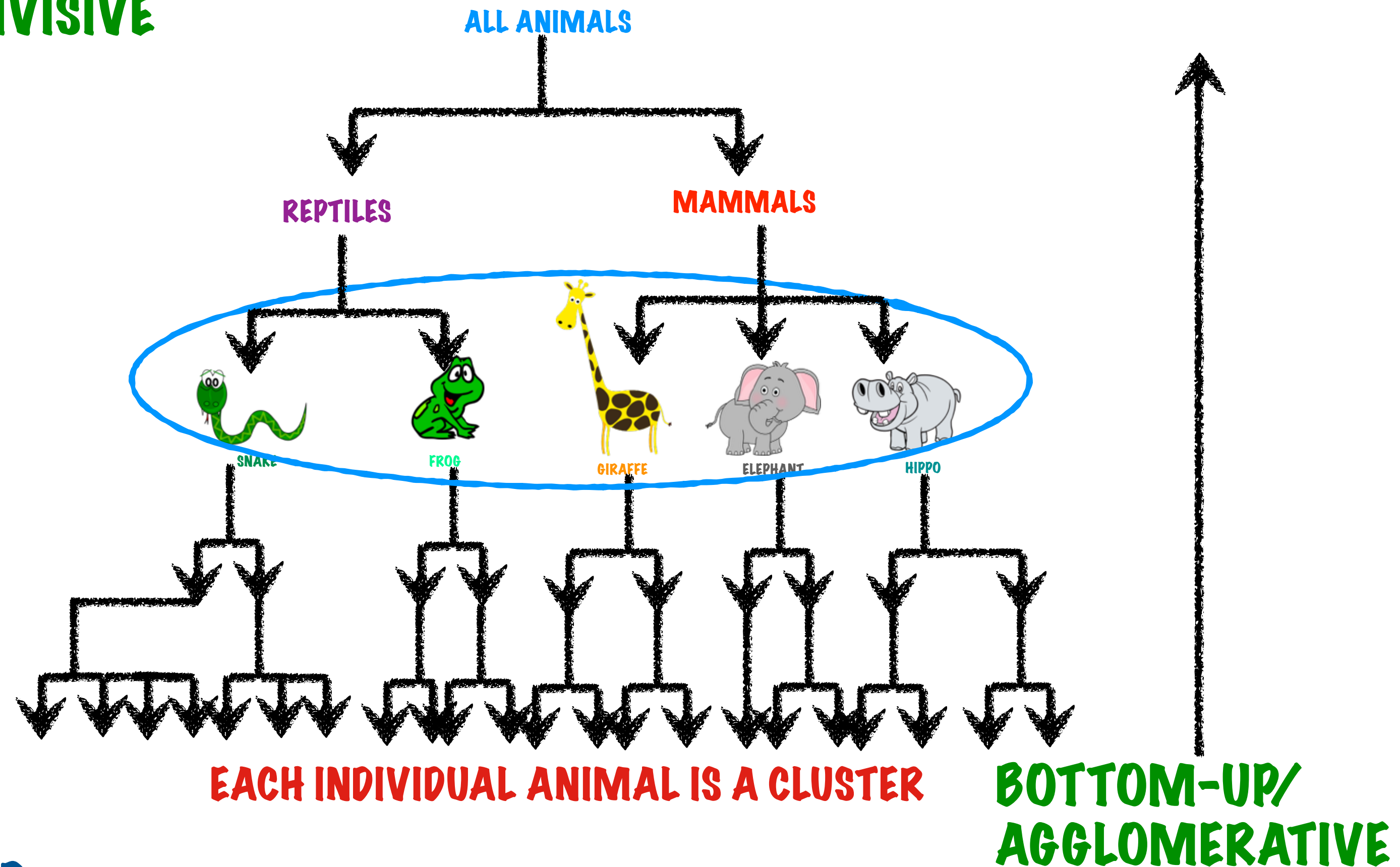
HIERARCHICAL CLUSTERING

A CROSS SECTION AT EACH LEVEL OF THE HIERARCHY IS A POSSIBLE SET OF CLUSTERS



TOP-DOWN/DIVISIVE

THE ULTIMATE END STATE IN BOTH THE TOP-DOWN AND THE BOTTOM-UP IS USELESS



THE ALGORITHM NEEDS A STOPPING CONDITION

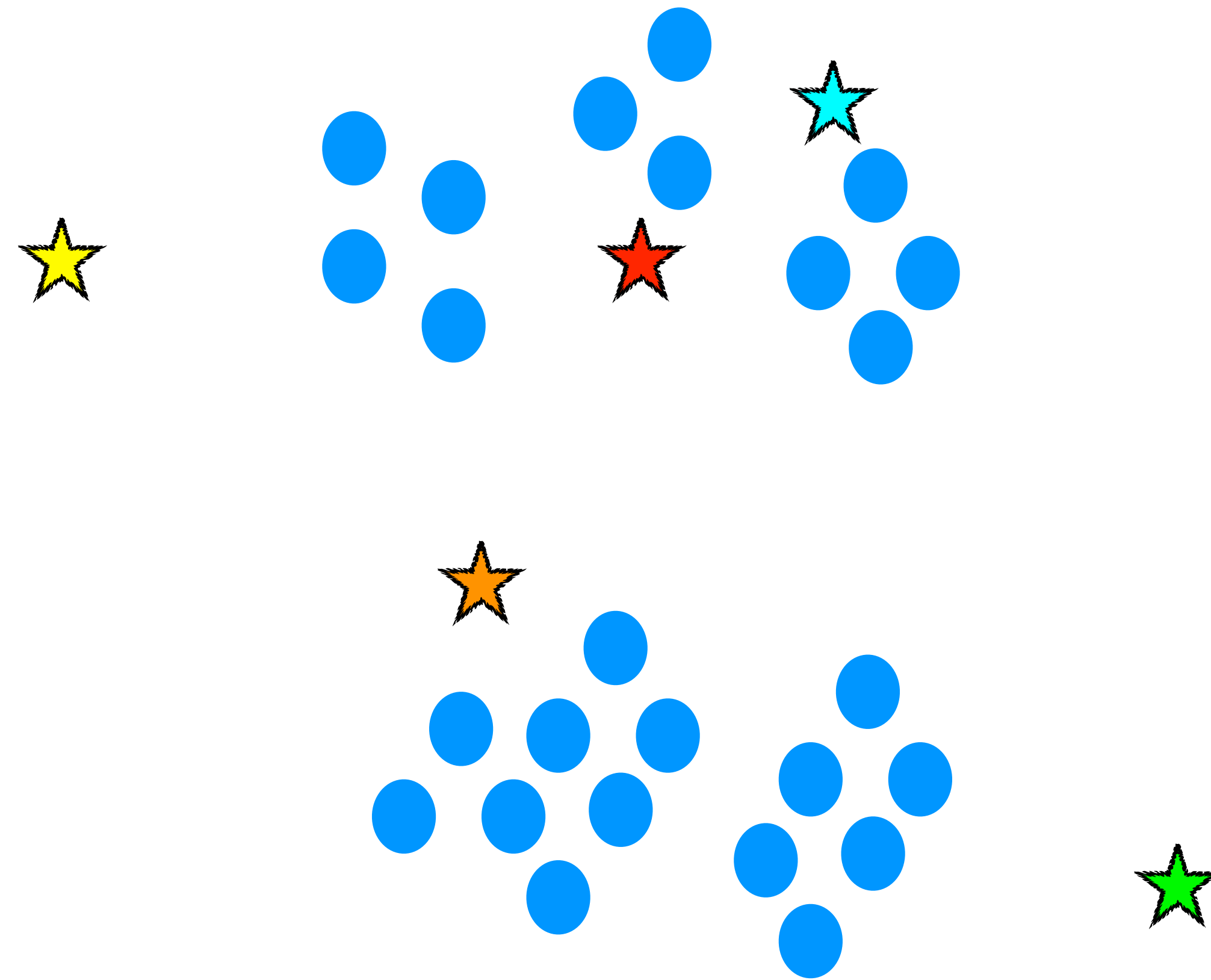
EXAMPLES:

1) NUMBER OF CLUSTERS

2) $\text{AVG}(\text{INTRACLUSTER DISTANCE}) - \text{AVG}(\text{INTERCLUSTER DISTANCE})$

K-MEANS CLUSTERING

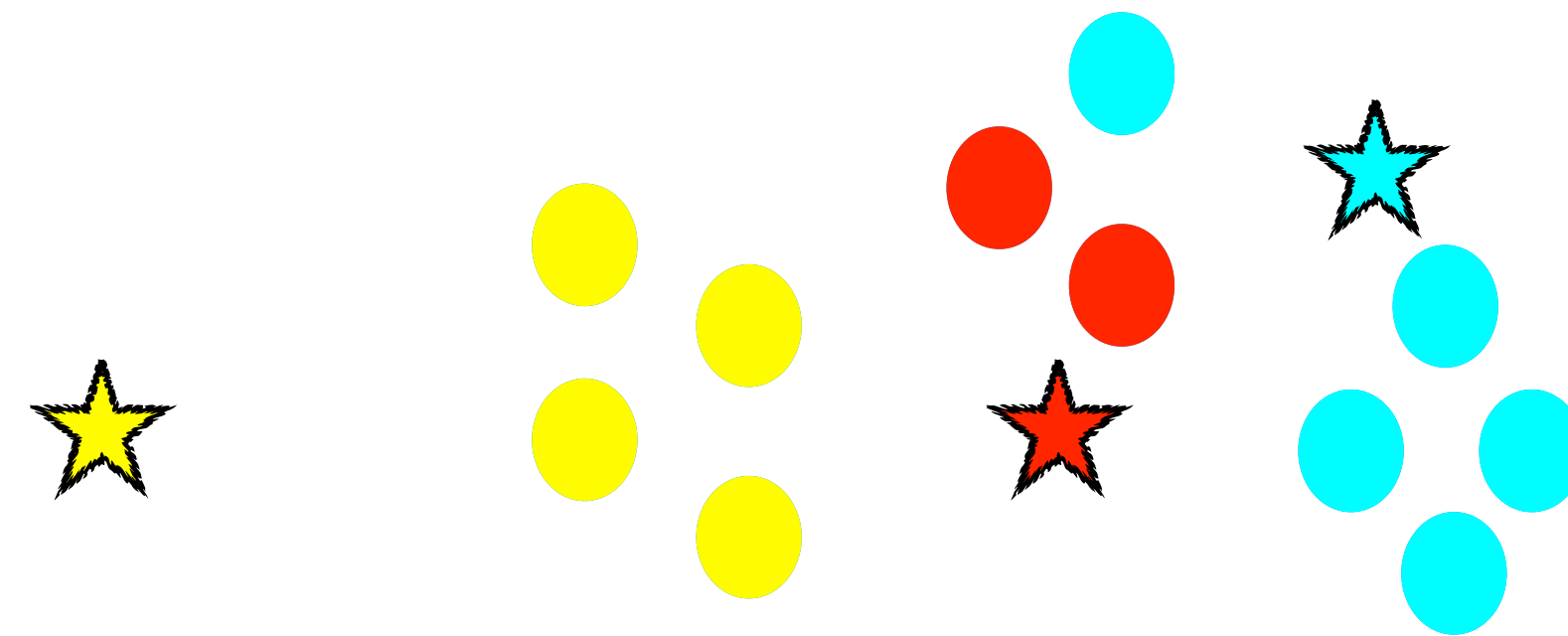
1. INITIALIZE A SET OF POINTS AS THE "K" MEANS
(CENTROIDS OF THE CLUSTERS YOU WANT TO FIND)



YOU WANT TO DIVIDE THIS DATA INTO K CLUSTERS

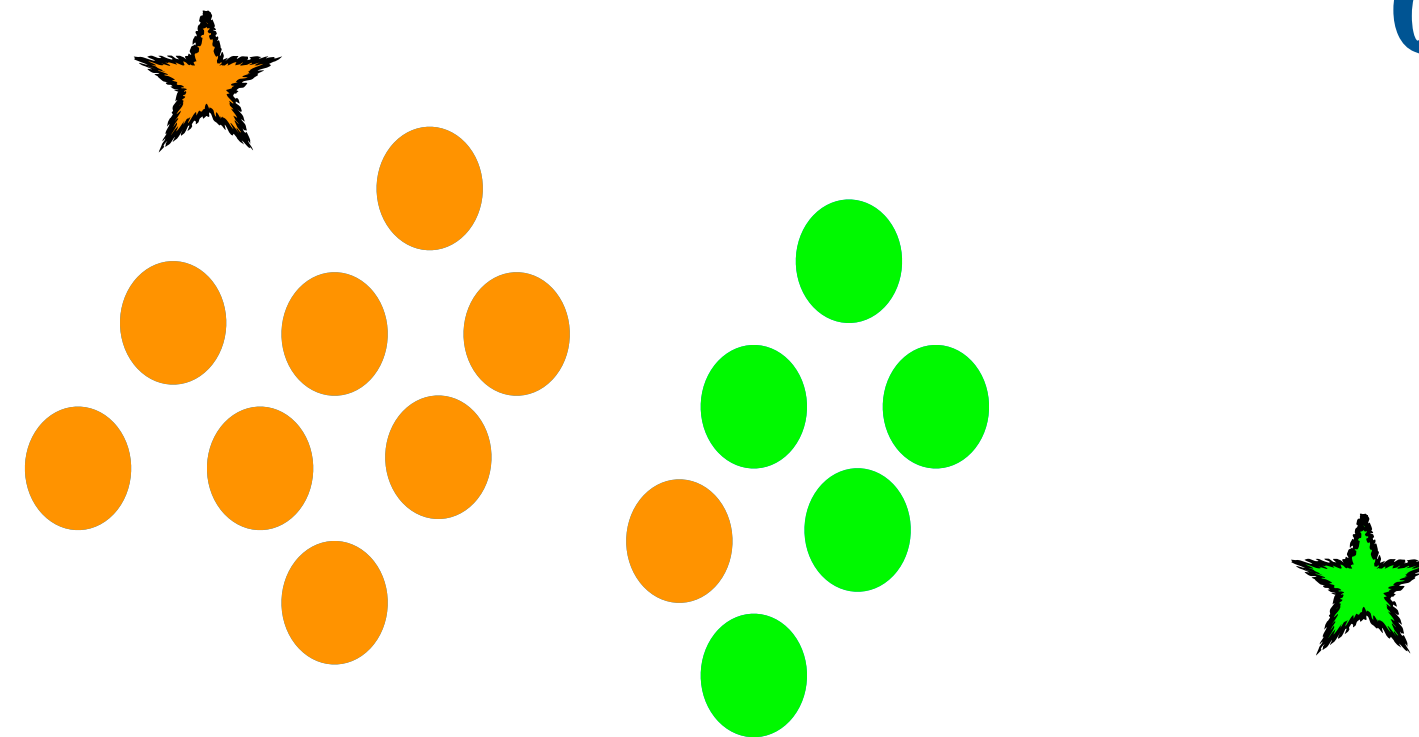
K-MEANS CLUSTERING

1 . INITIALIZE A SET OF POINTS AS THE "K" MEANS
(CENTROIDS OF THE CLUSTERS YOU WANT TO FIND)



2 . ASSIGN EACH POINT TO THE
CLUSTER BELONGING TO THE
NEAREST MEAN

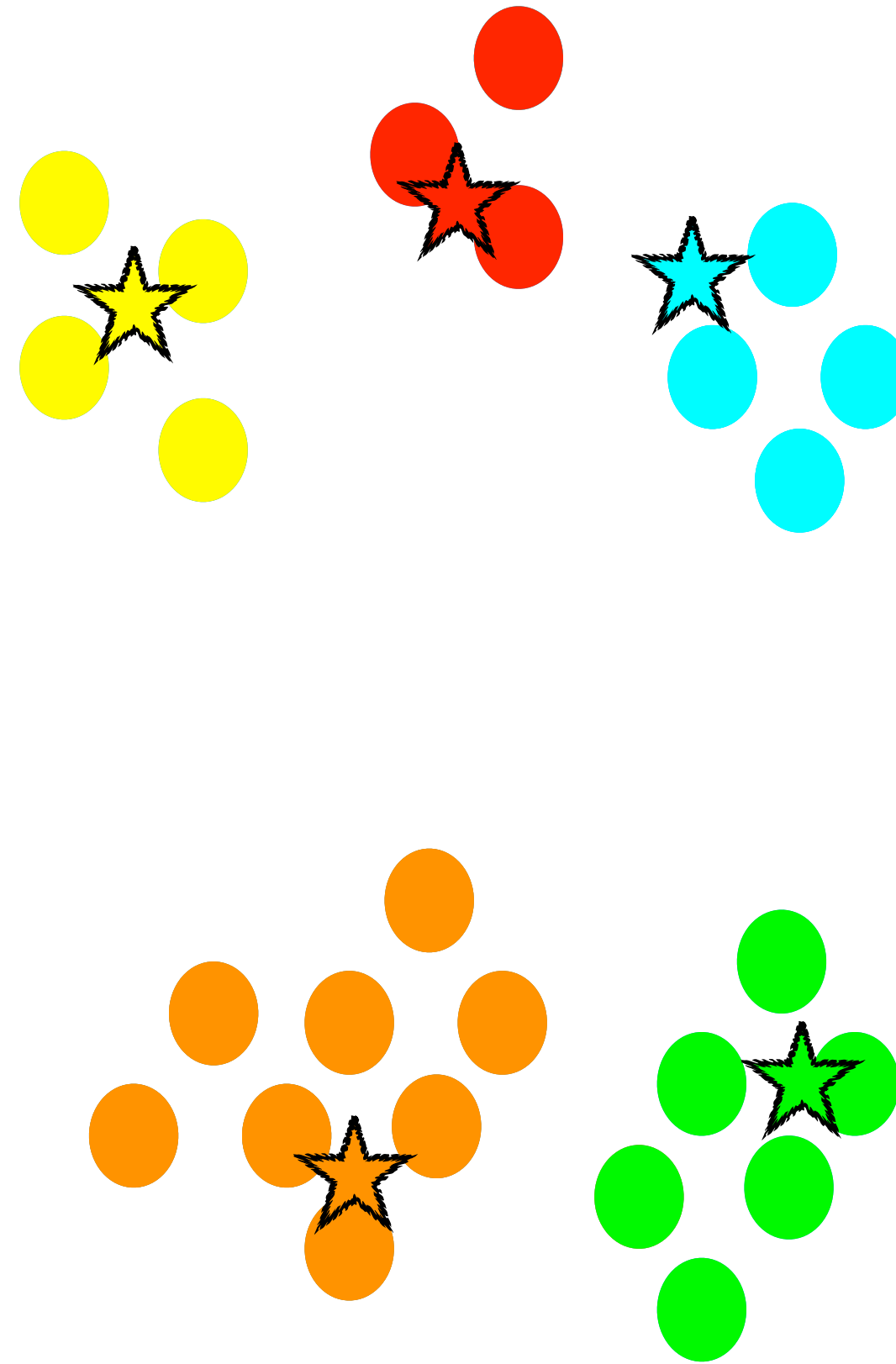
3 . FIND THE NEW MEANS /
CENTROIDS OF THE CLUSTERS



YOU WANT TO DIVIDE THIS DATA INTO K CLUSTERS

K-MEANS CLUSTERING

1 . INITIALIZE A SET OF POINTS AS THE "K" MEANS
(CENTROIDS OF THE CLUSTERS YOU WANT TO FIND)



2 . ASSIGN EACH POINT TO THE
CLUSTER BELONGING TO THE
NEAREST MEAN

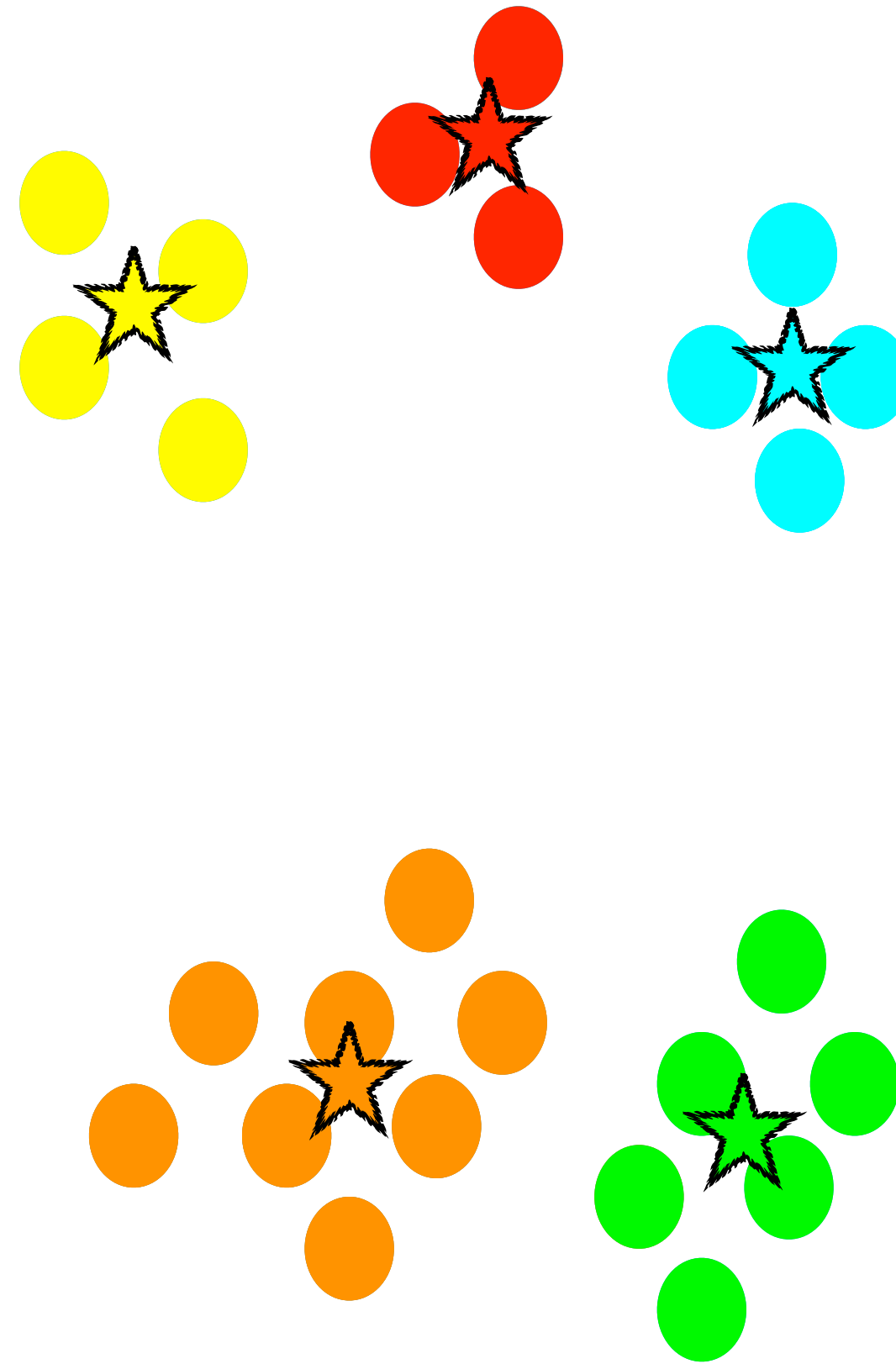
3 . FIND THE NEW MEANS /
CENTROIDS OF THE CLUSTERS

RINSE AND REPEAT
STEPS 2, 3 UNTIL THE MEANS
DON'T CHANGE ANY MORE

YOU WANT TO DIVIDE THIS DATA INTO K CLUSTERS

K-MEANS CLUSTERING

1 . INITIALIZE A SET OF POINTS AS THE "K" MEANS
(CENTROIDS OF THE CLUSTERS YOU WANT TO FIND)



2 . ASSIGN EACH POINT TO THE
CLUSTER BELONGING TO THE
NEAREST MEAN

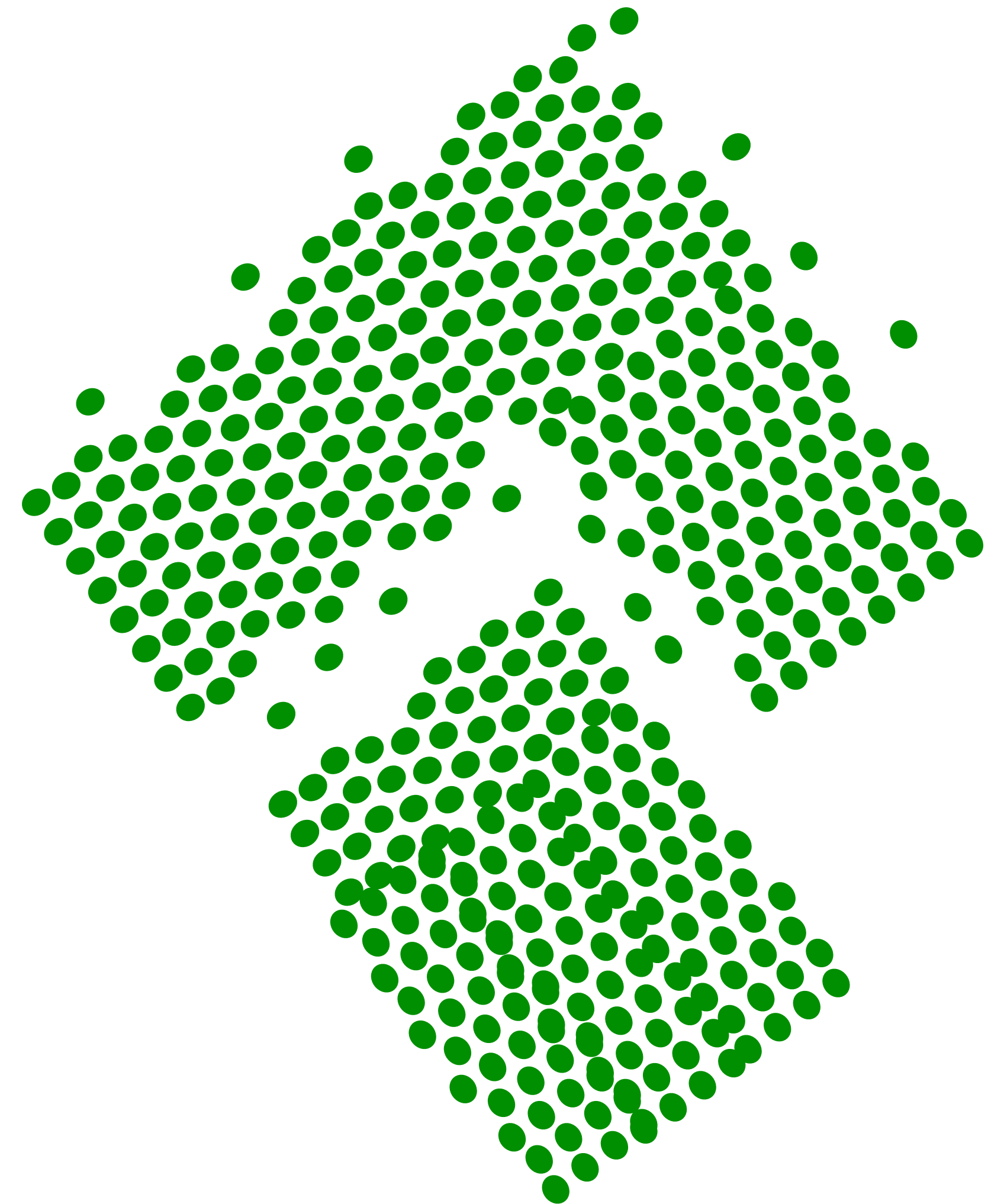
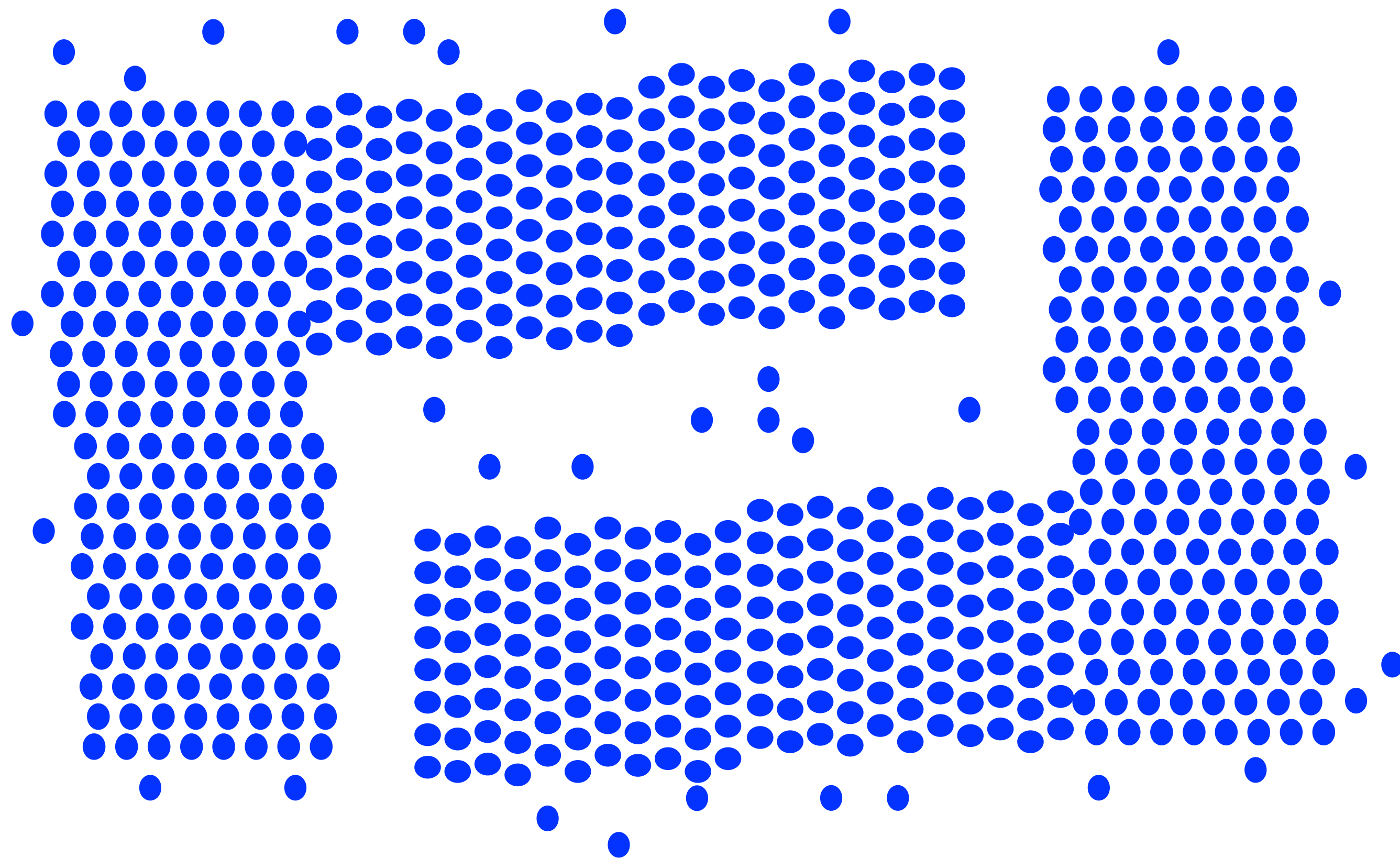
3 . FIND THE NEW MEANS /
CENTROIDS OF THE CLUSTERS

RINSE AND REPEAT
STEPS 2, 3 UNTIL THE MEANS
DON'T CHANGE ANY MORE
CONVERGENCE

YOU WANT TO DIVIDE THIS DATA INTO K CLUSTERS

IN K-MEANS CLUSTERING - THE CLUSTERING IS BASED ON DISTANCE

DISTANCE BASED CLUSTERING DOESN'T REALLY HELP WHEN YOUR DATA LOOKS LIKE THIS...

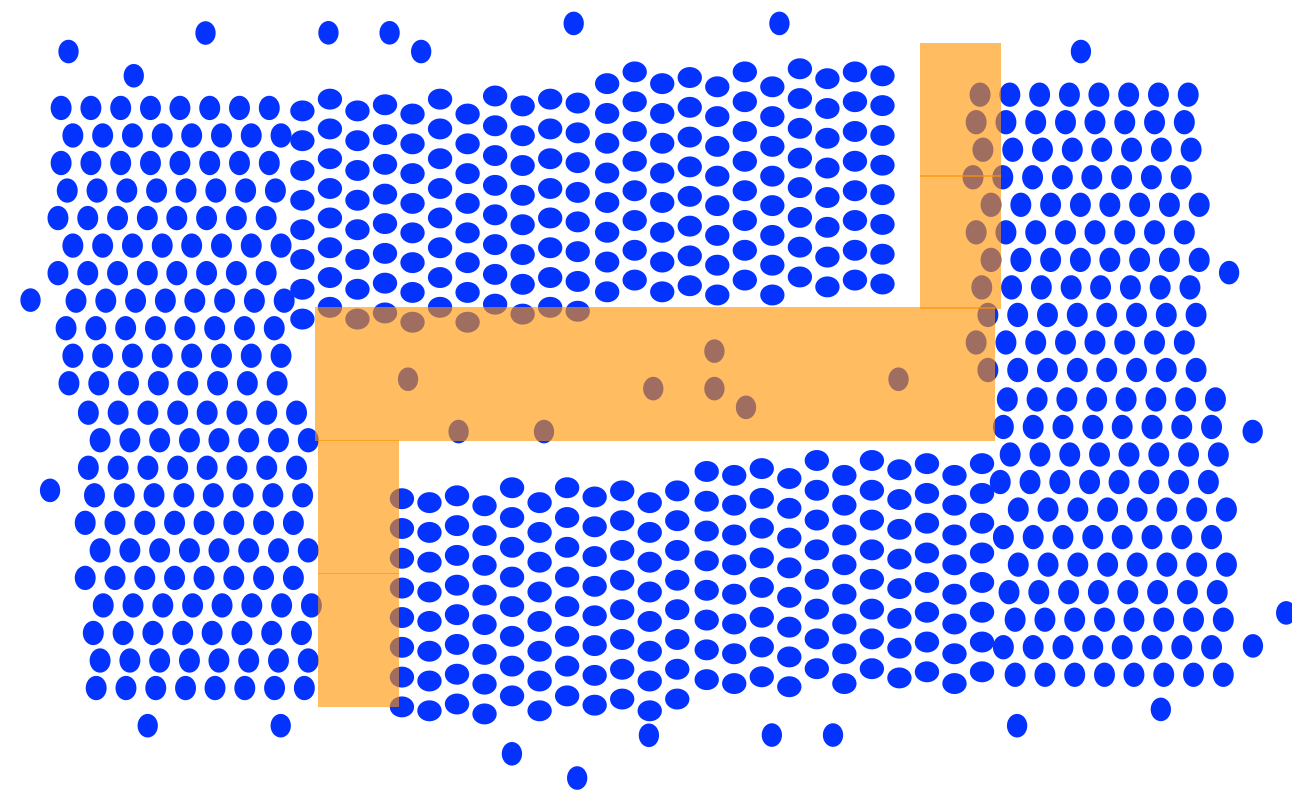


..OR LIKE THIS

DENSITY BASED CLUSTERING

THIS METHOD SCANS FOR DROPS IN DENSITY
AND MARKS THEM AS CLUSTER BOUNDARIES

THE MOST FAMOUS DENSITY BASED
CLUSTERING ALGORITHM IS CALLED



DBSCAN WILL VISIT EACH POINT AND SCAN ITS
NEIGHBOURHOOD

HOW MANY POINTS ARE
AROUND THIS POINT?

WHICH CLUSTER DO THEY
BELONG TO?

IT THEN DECIDES IF THE POINT BELONGS TO
SOME CLUSTER OR IT'S AN OUTLIER

DBSCAN
DENSITY-BASED
SPATIAL
CLUSTERING OF
APPLICATIONS
WITH NOISE

(QUITE A MOUTHFUL :))