Msc BDA Part 1

DBSCAN

026| UMAR MUKADAM |2309035

023| ROHAN DSOUZA |2309031

01| JADEN VARGHESE |2309001

What is clustering ?

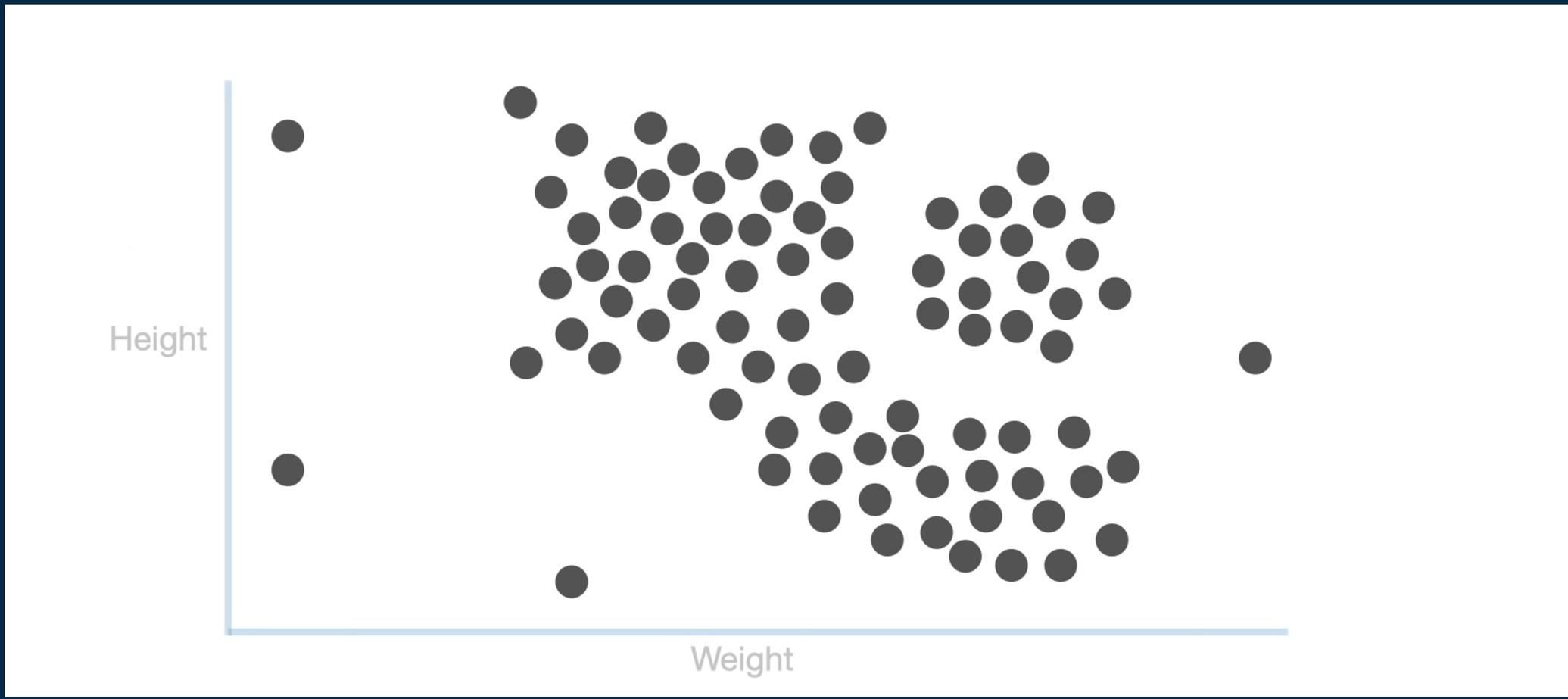
- Technique partitioning a dataset into groups (clusters) with similar data points.
- Clustering is an unsupervised learning method, learning patterns autonomously.
- Widely used in customer segmentation, market analysis, image segmentation, anomaly detection, and pattern recognition.
- Algorithms rely on distance metrics like Euclidean or Manhattan to gauge similarity/dissimilarity.

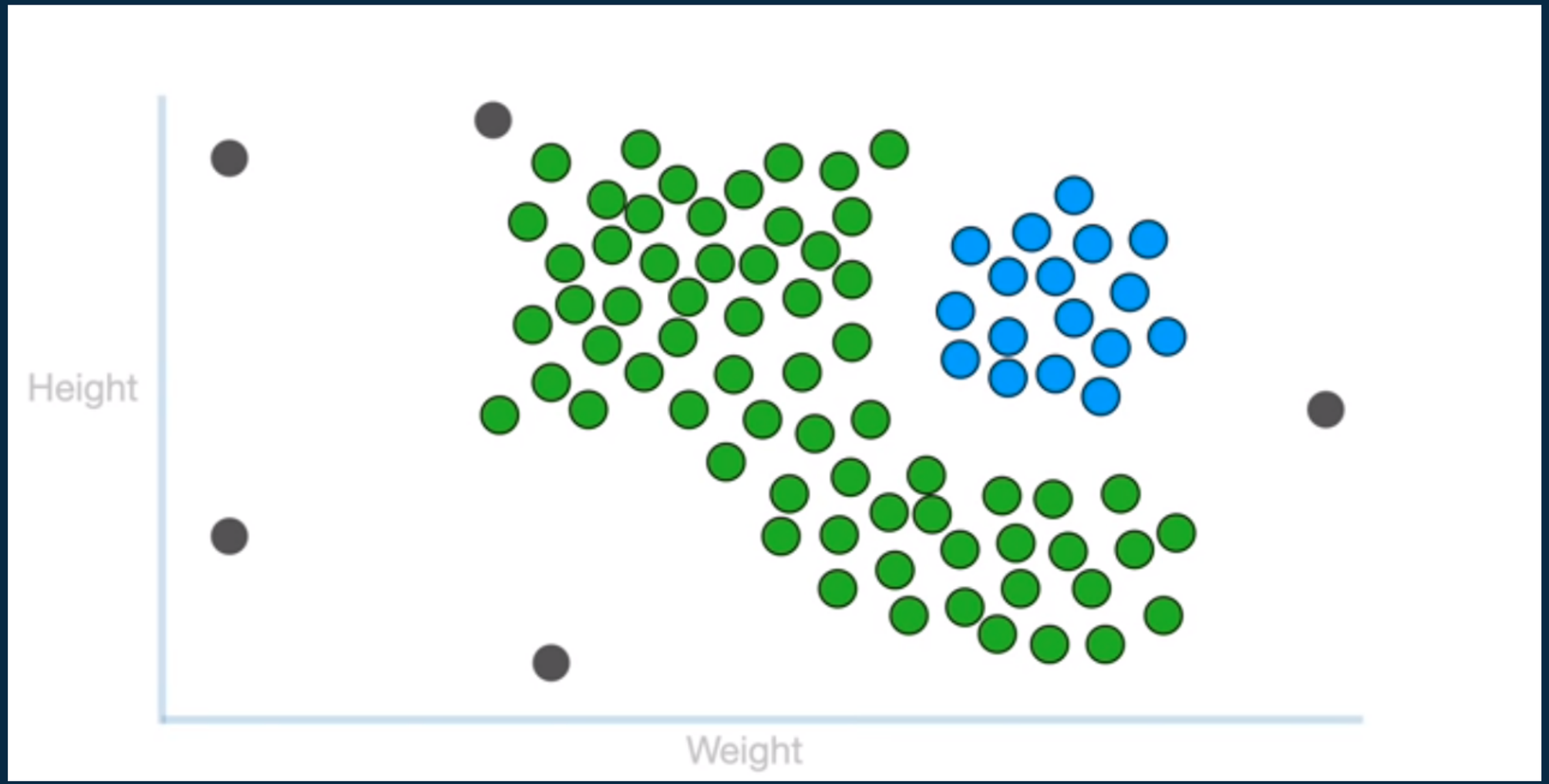
What is DBSCAN ?

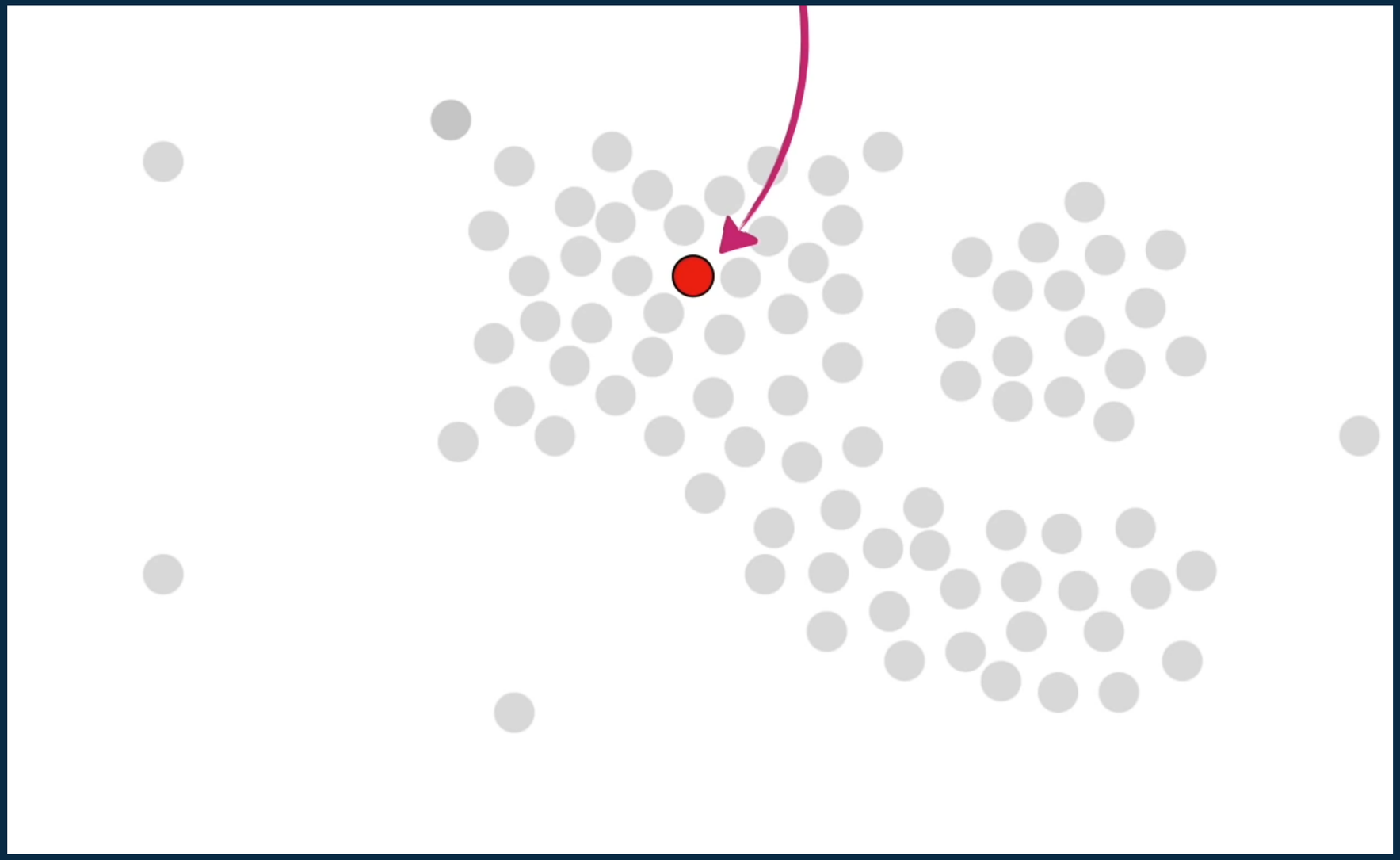
- Density-Based Spatial Clustering of Applications with Noise
- Identifies clusters based on the density of data points
- Does not require specifying the number of clusters in advance
- Classifies points as core, border, or noise based on density reachability.
- Utilizes two parameters: epsilon (ε) and minimum points (MinPts).
- Robust to outliers and capable of identifying clusters of arbitrary shapes.

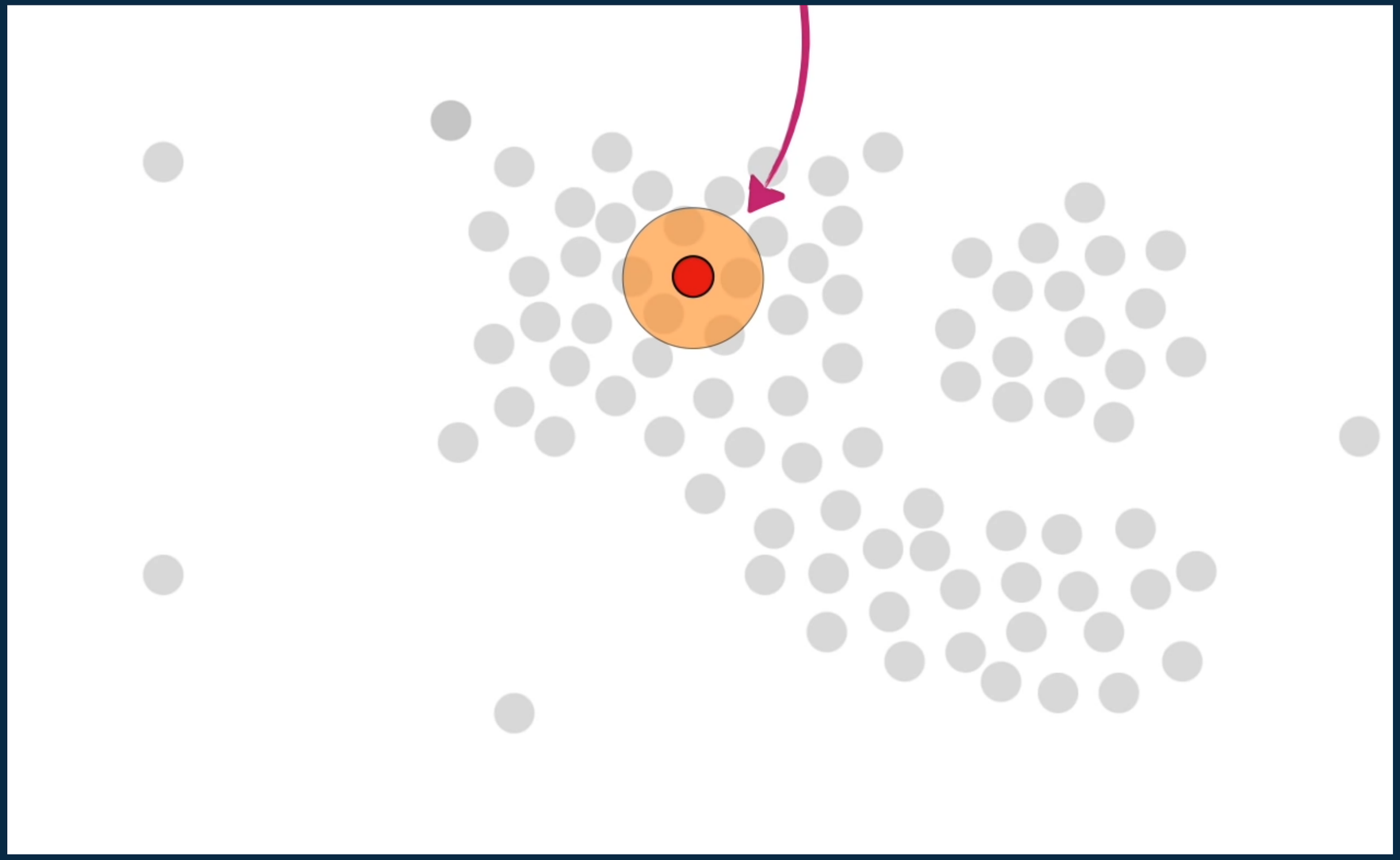
Parameters

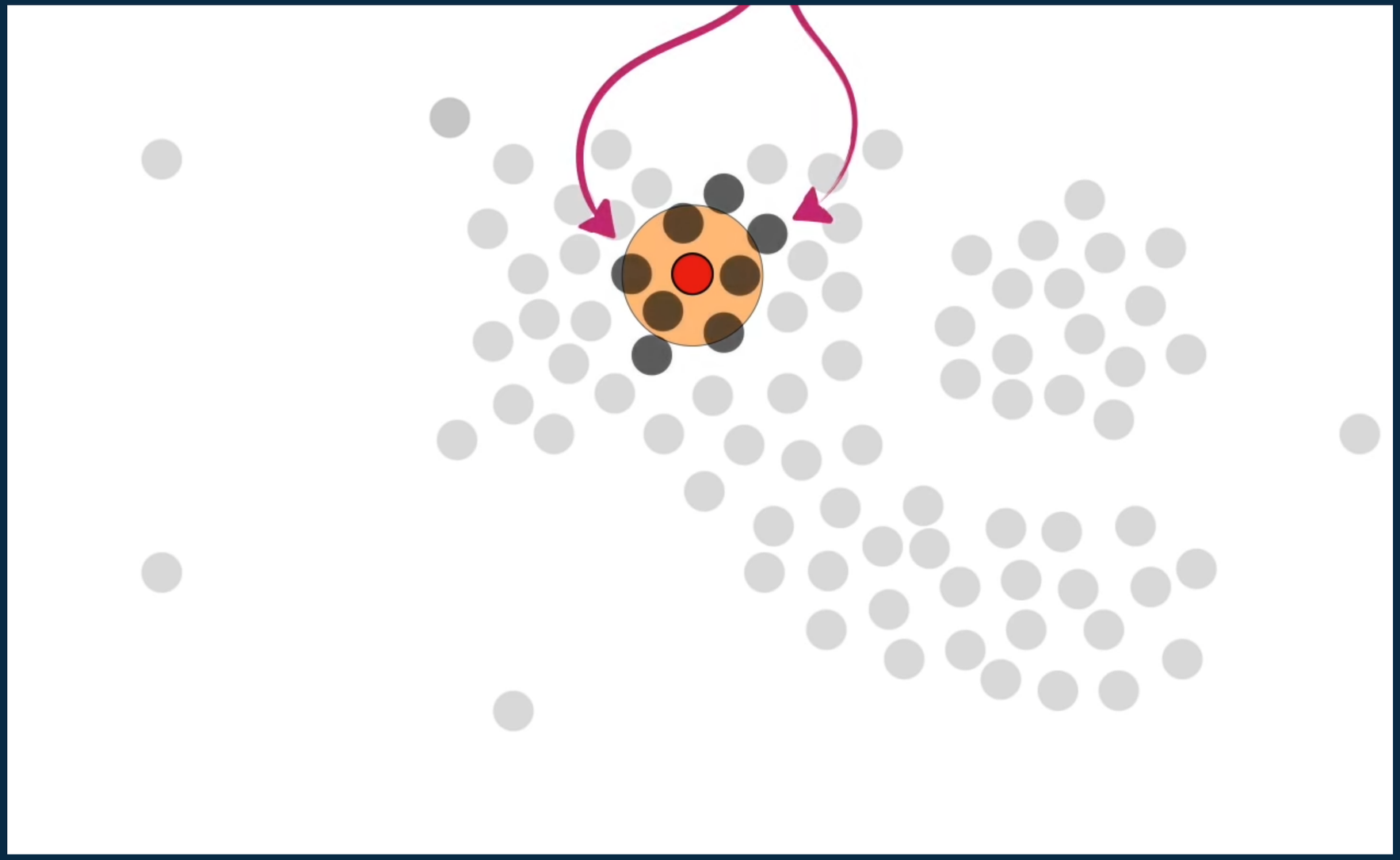
- Epsilon (ε): The maximum radius of the neighborhood.
- Minimum Points (MinPts): The minimum number of points required to form a dense region.
- Core Points: Densely packed data points within clusters, defining the core structure.
- Noise Points: Outliers or isolated data points that do not belong to any cluster.
- Border Points: Data points on the edge of clusters, adjacent to core points but not dense enough to be considered core themselves.

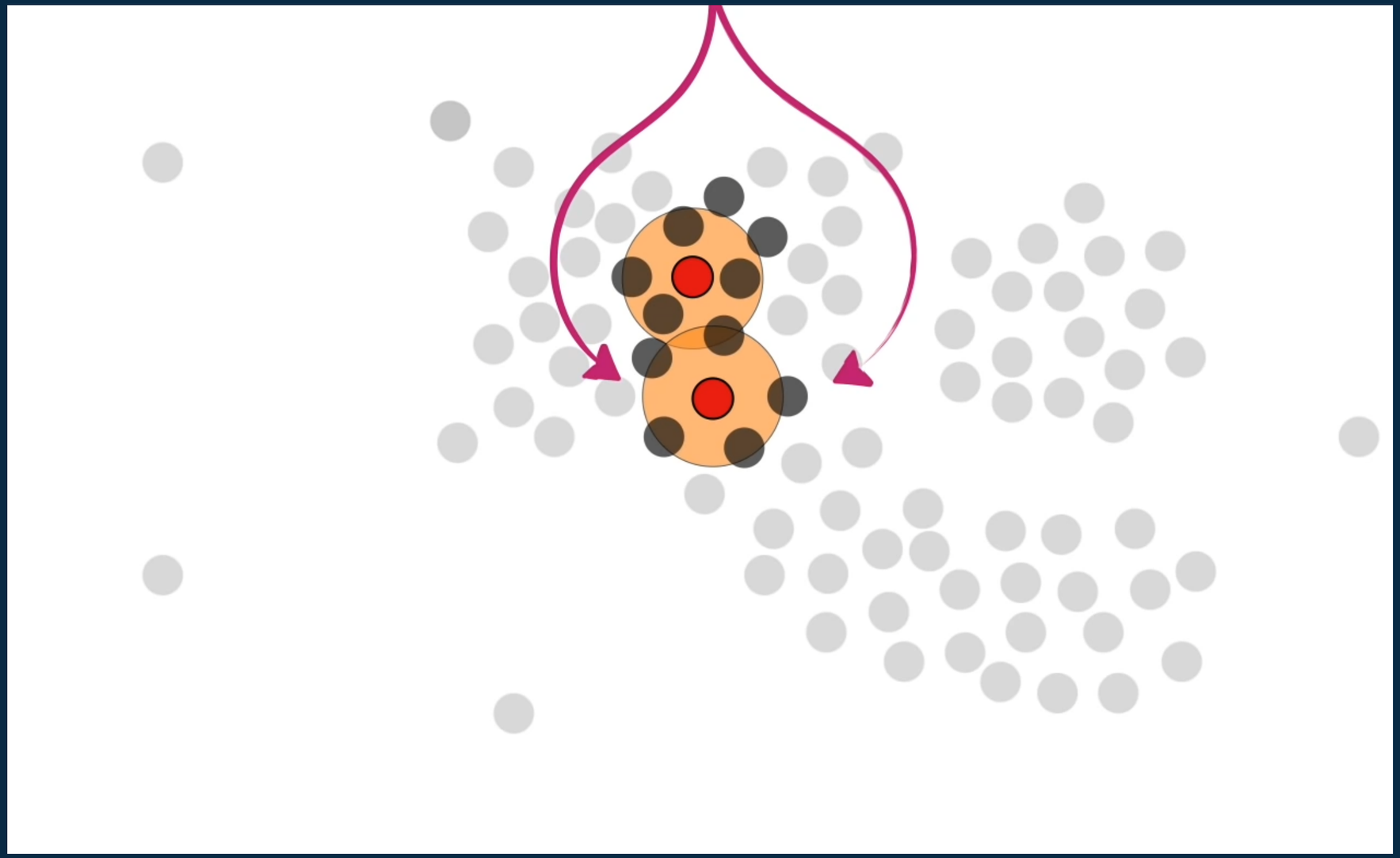

Height
Weight

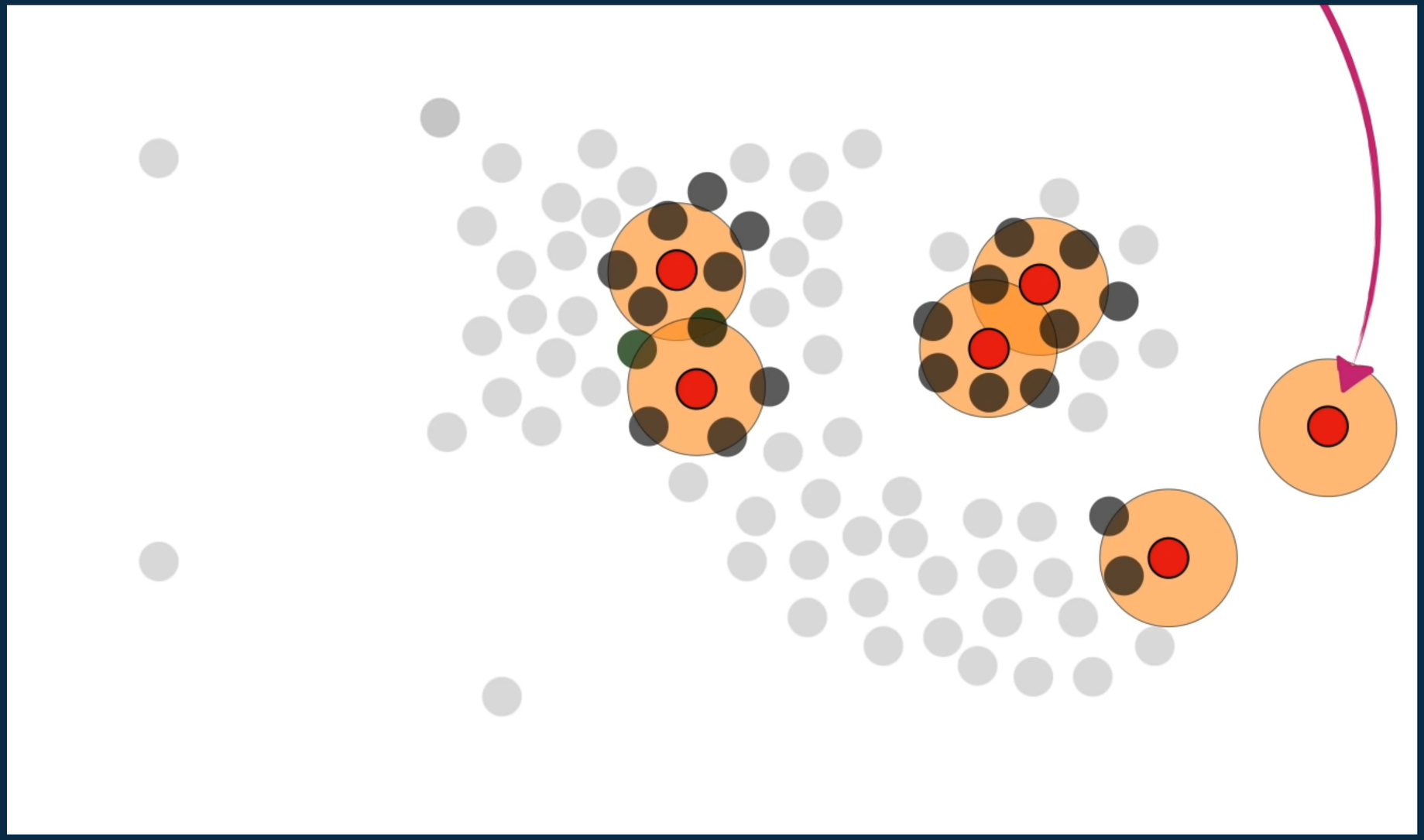

Height
Weight

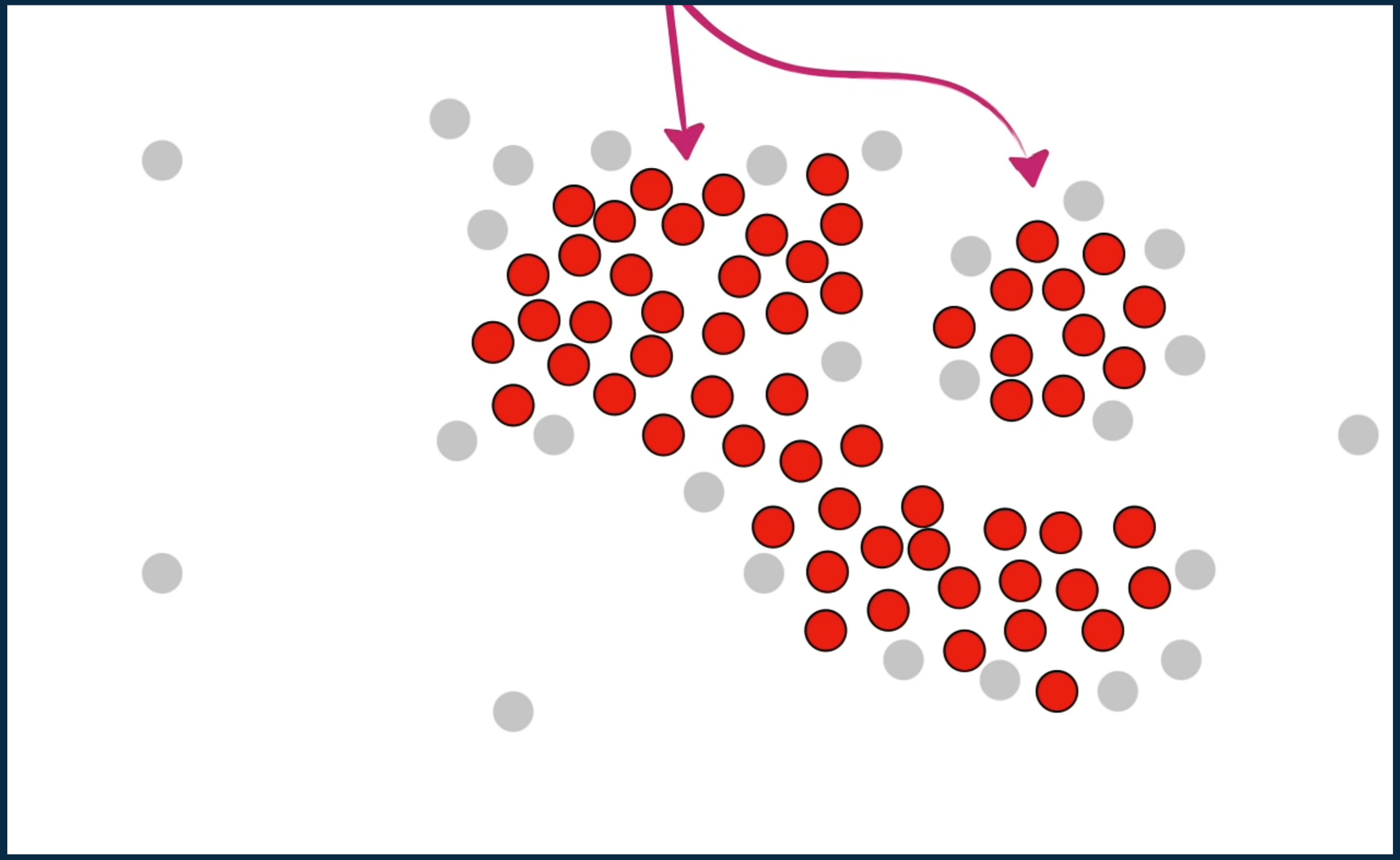


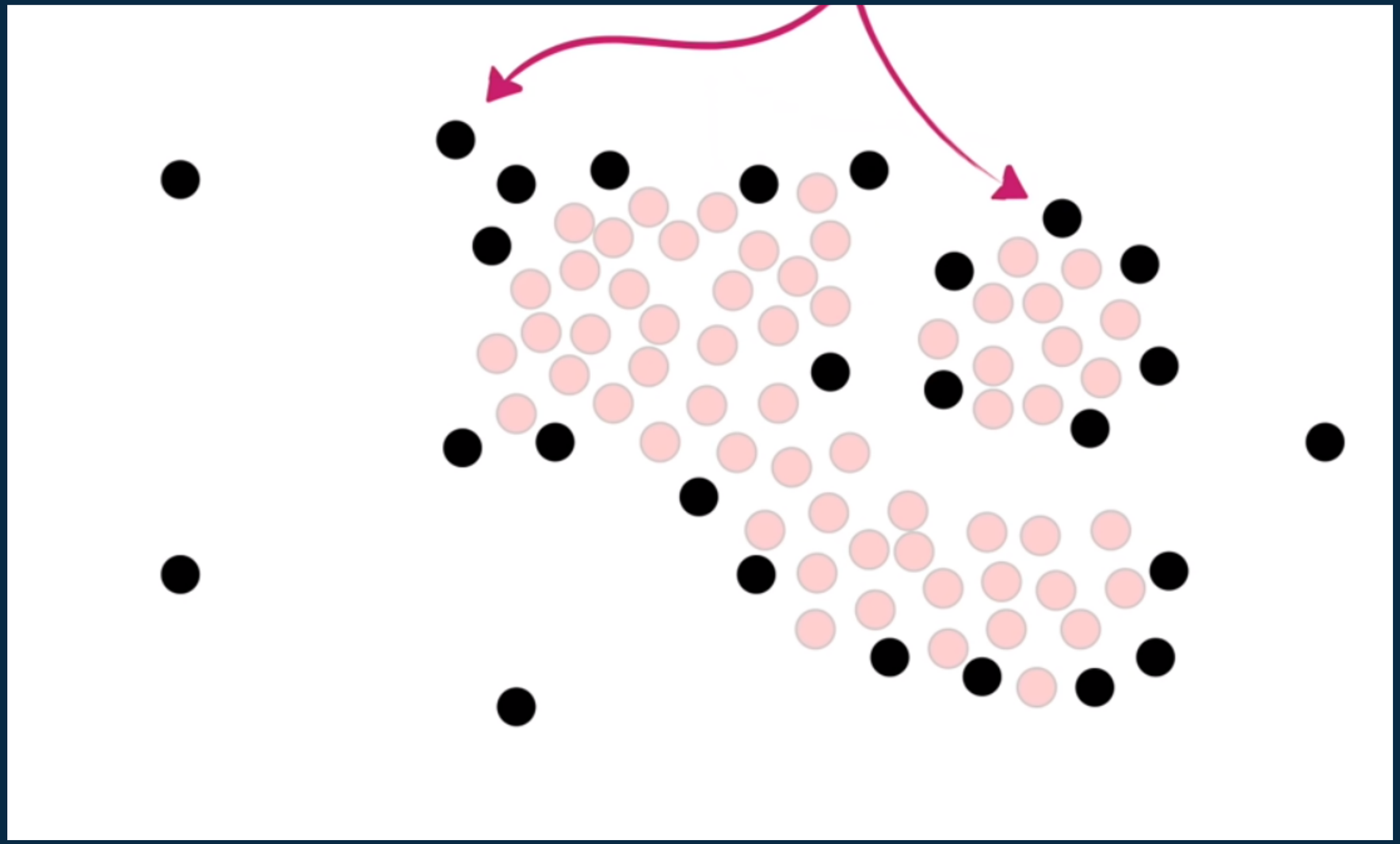


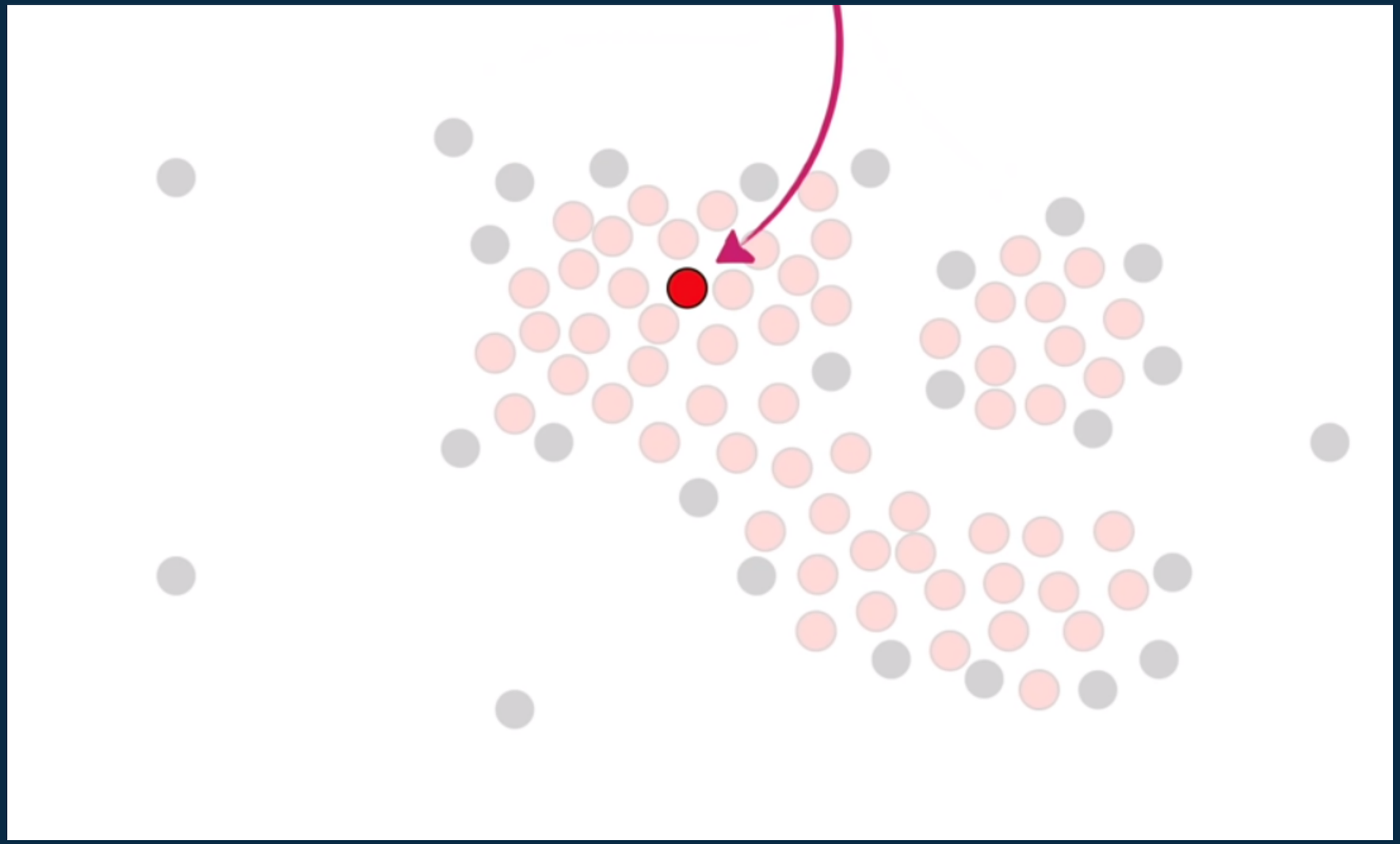


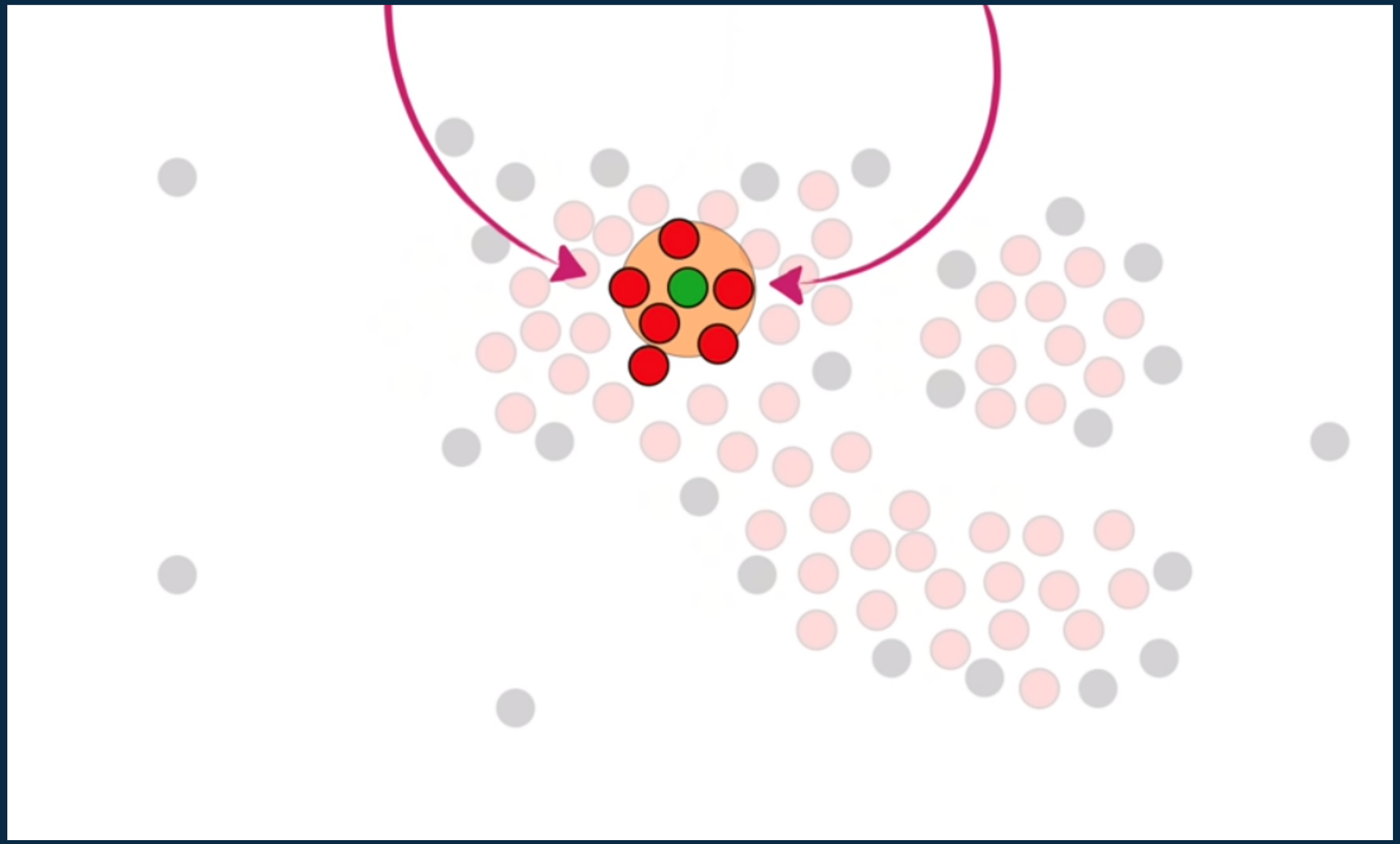


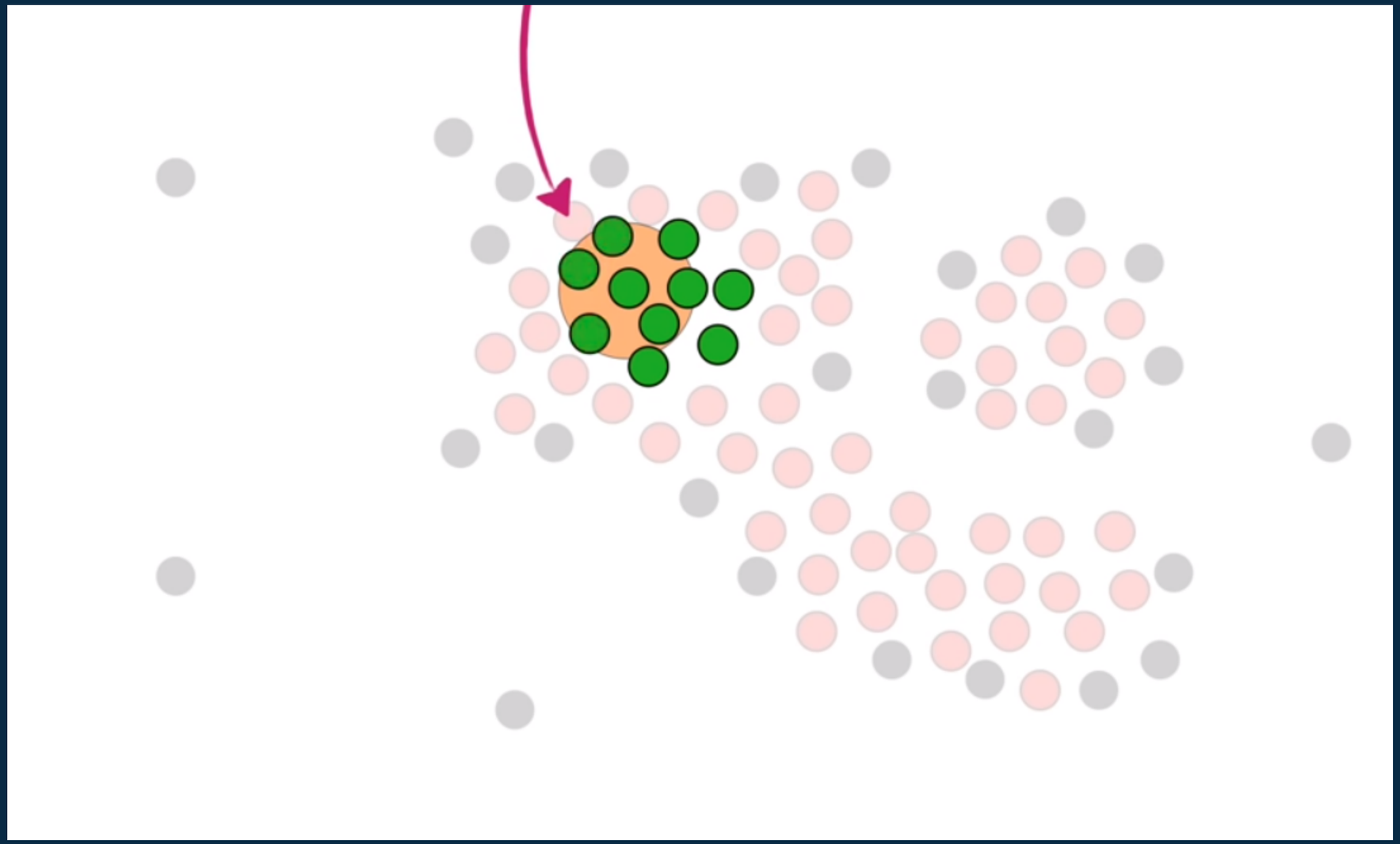


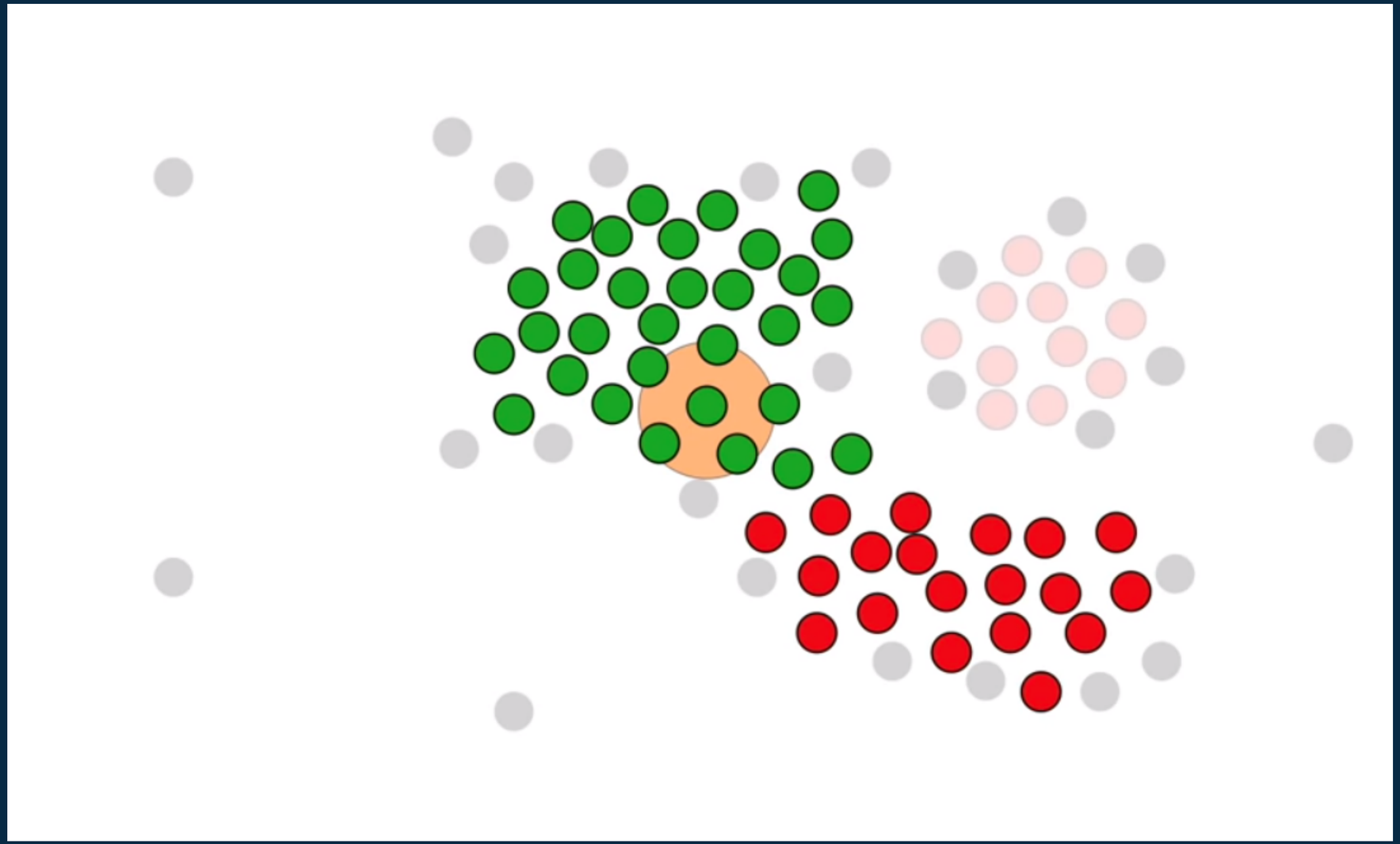


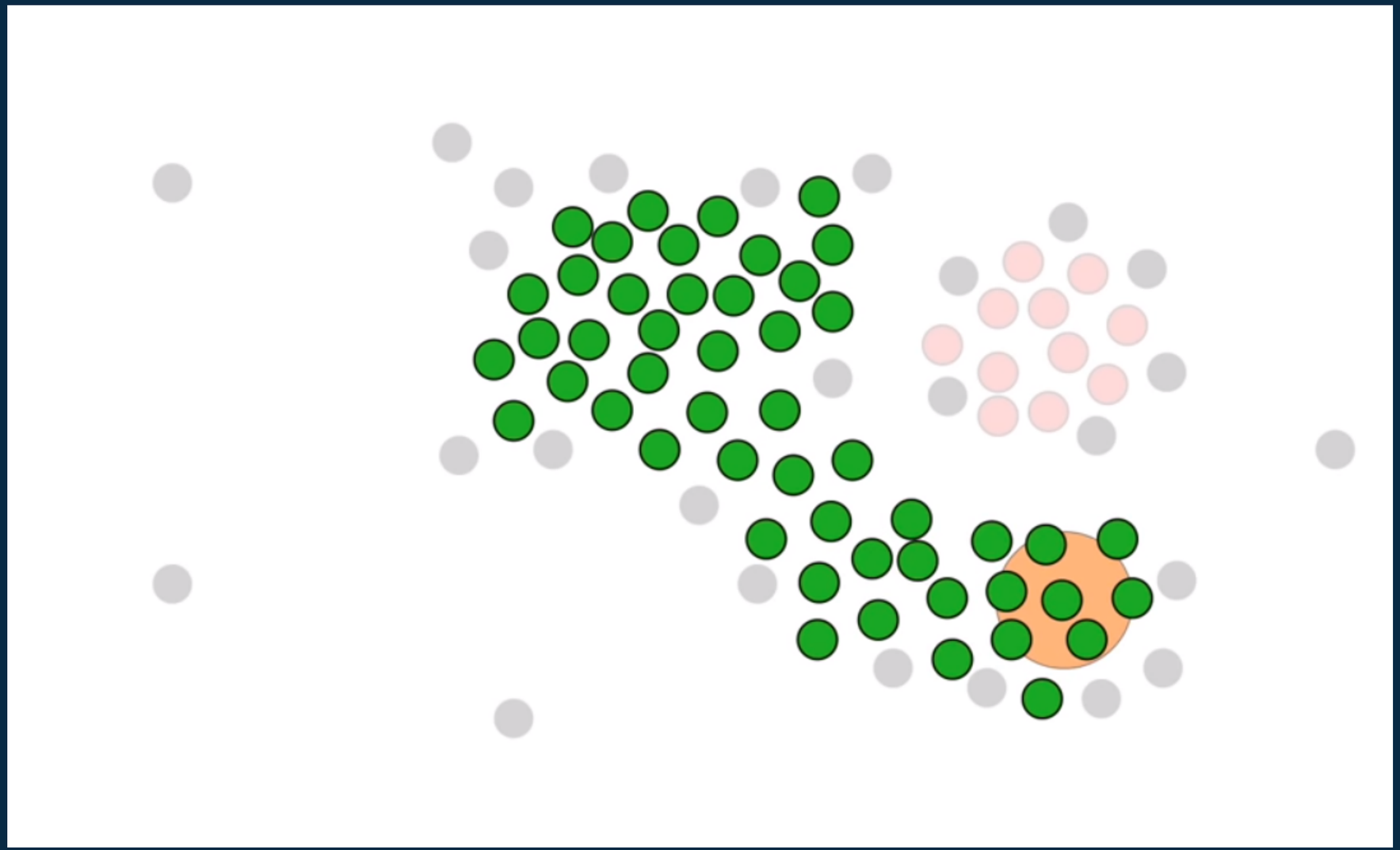


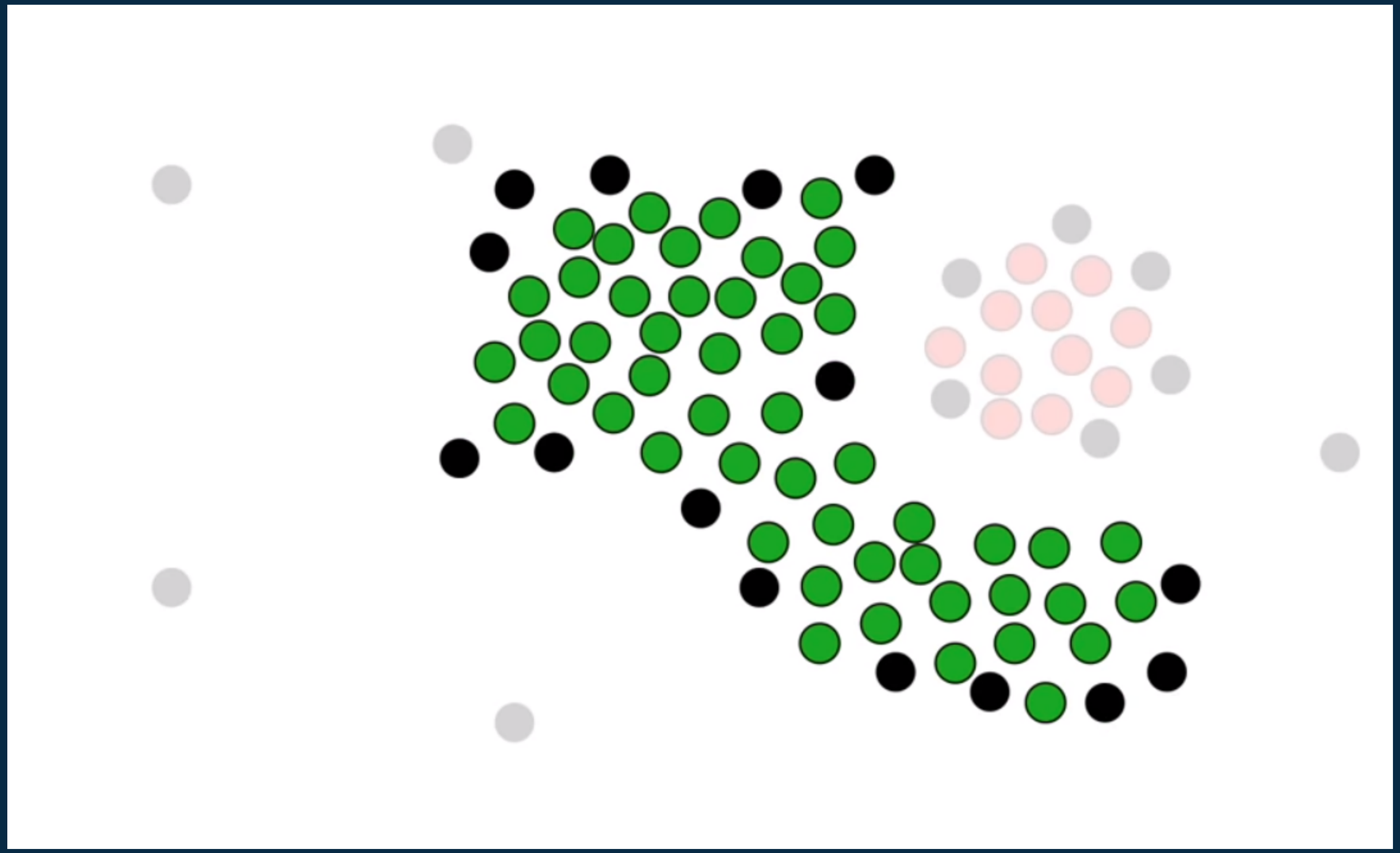


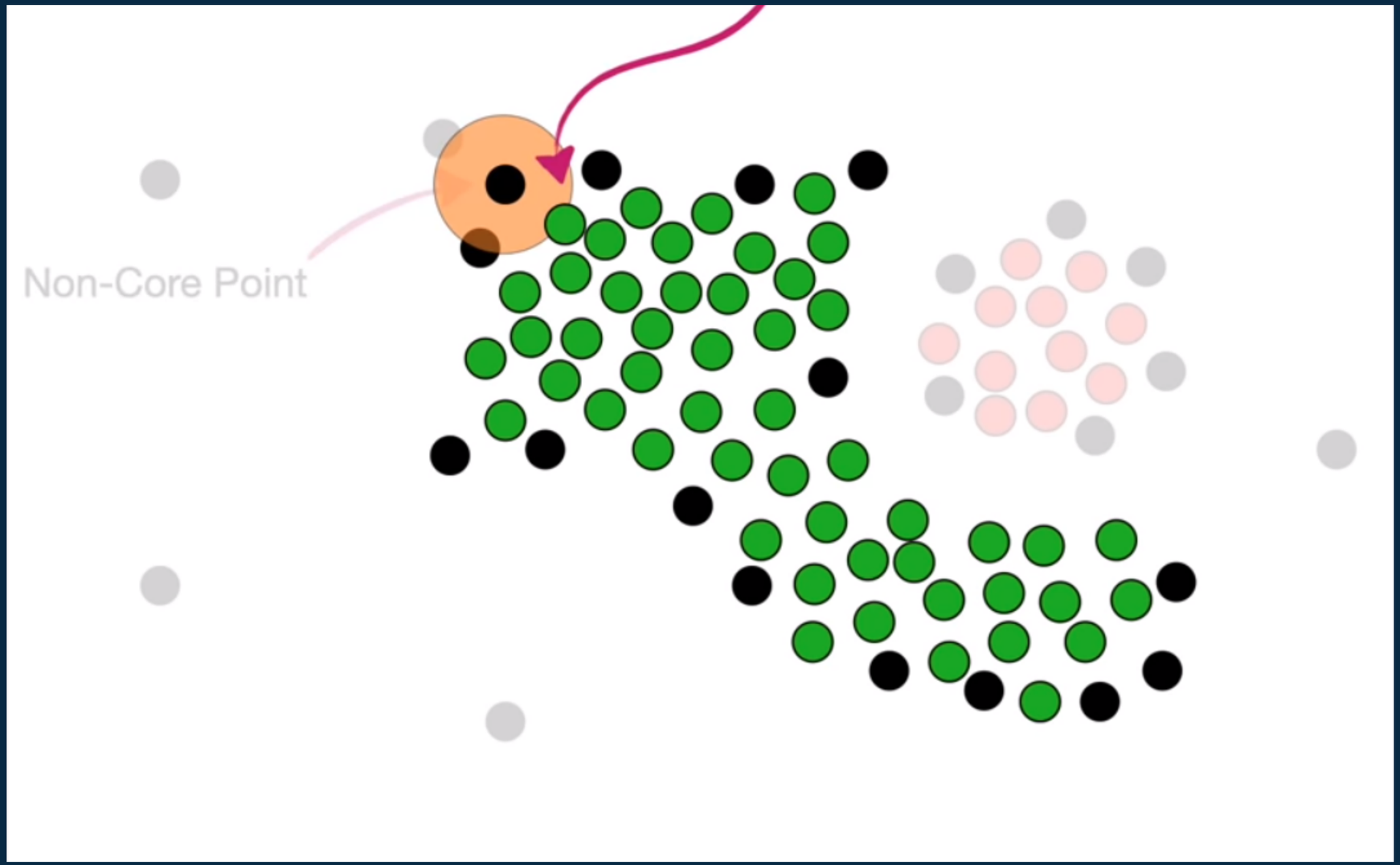


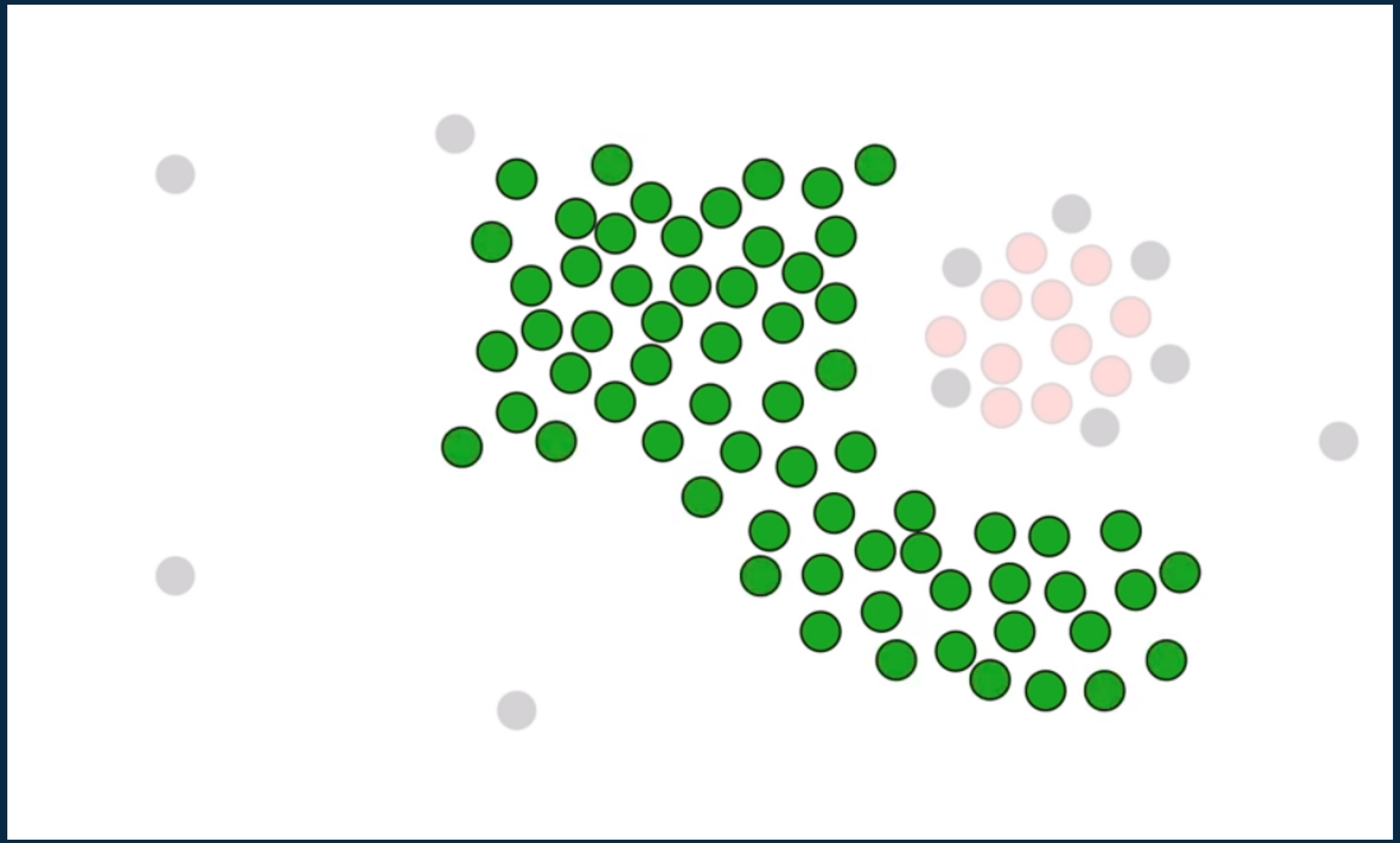


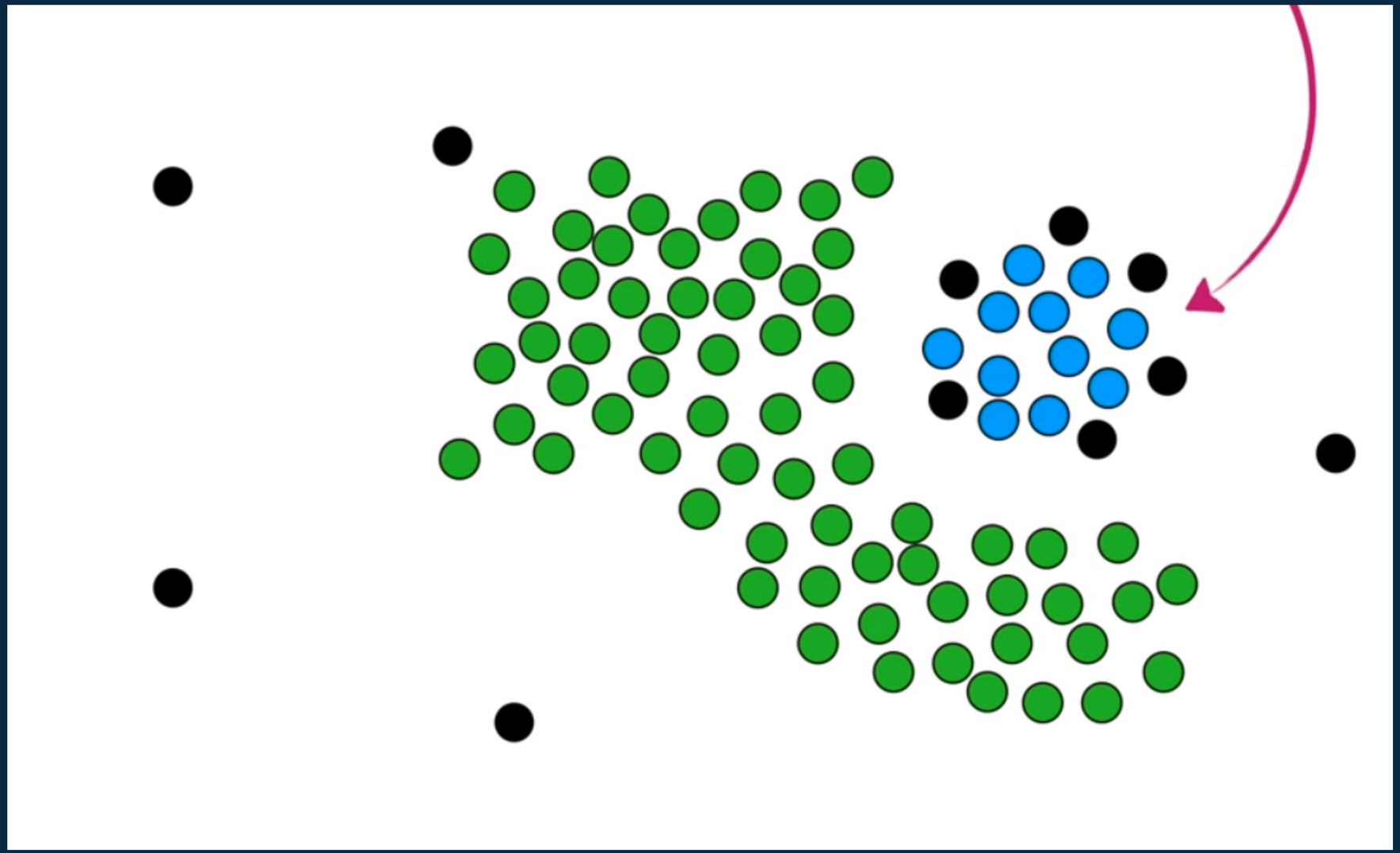


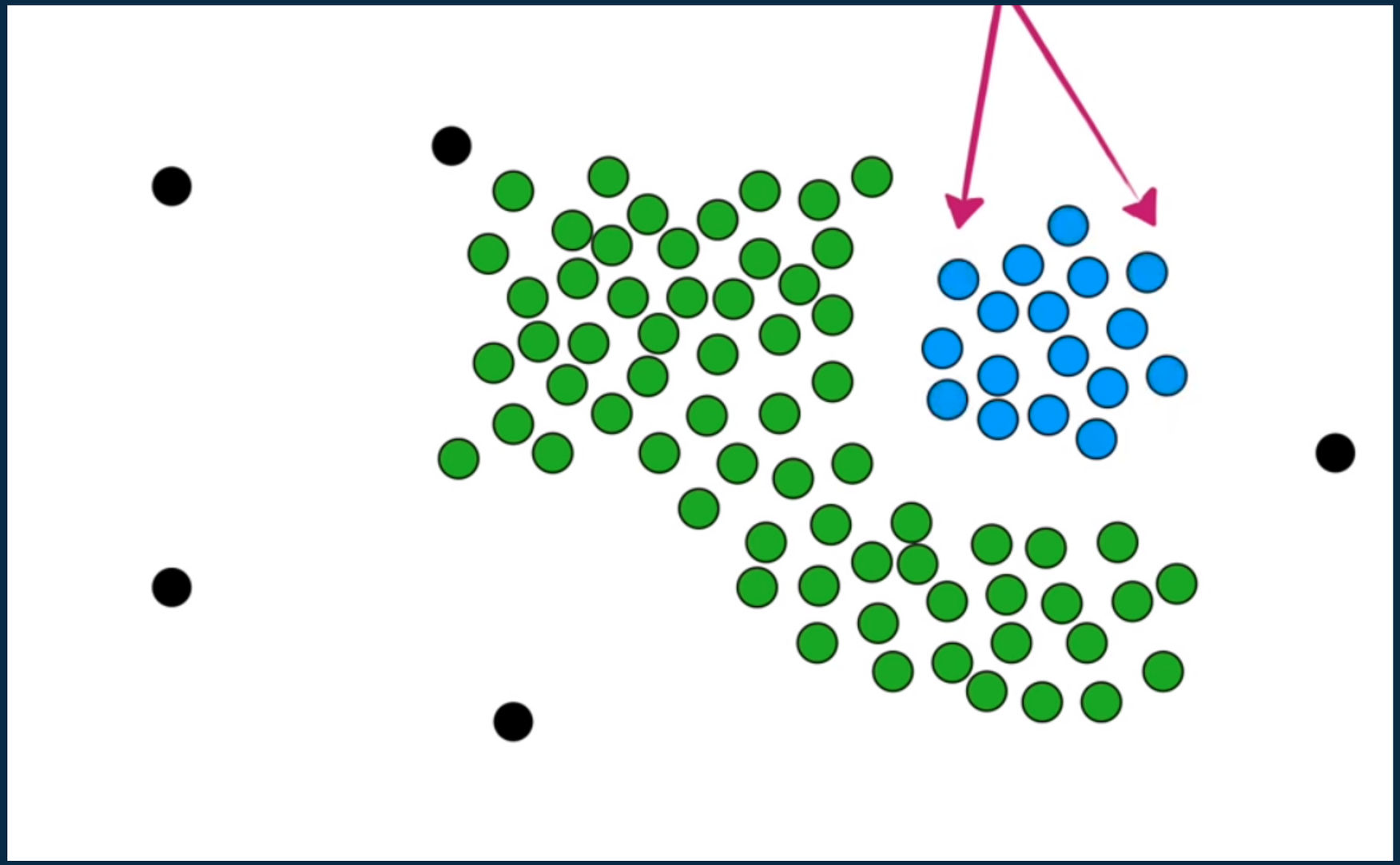


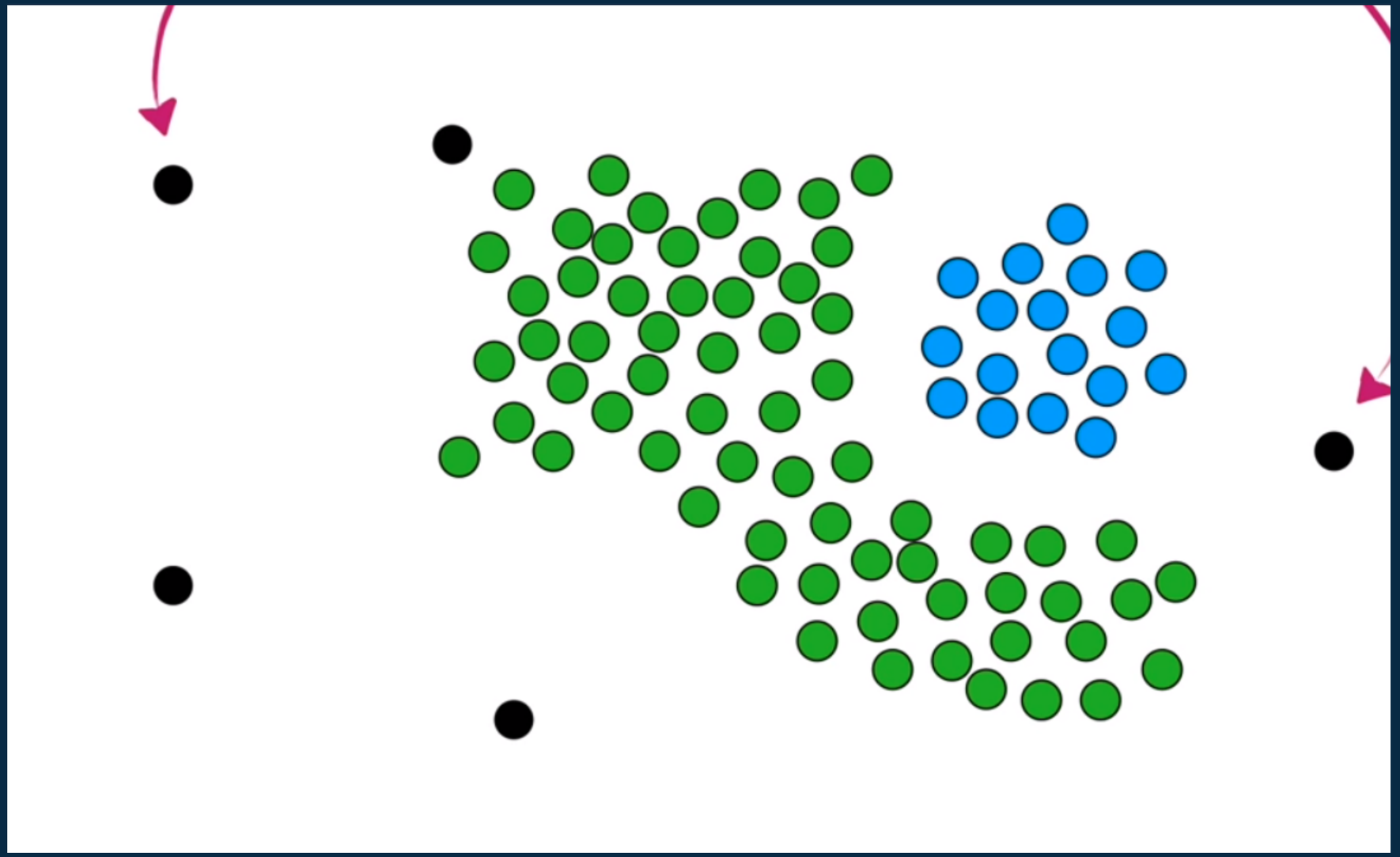


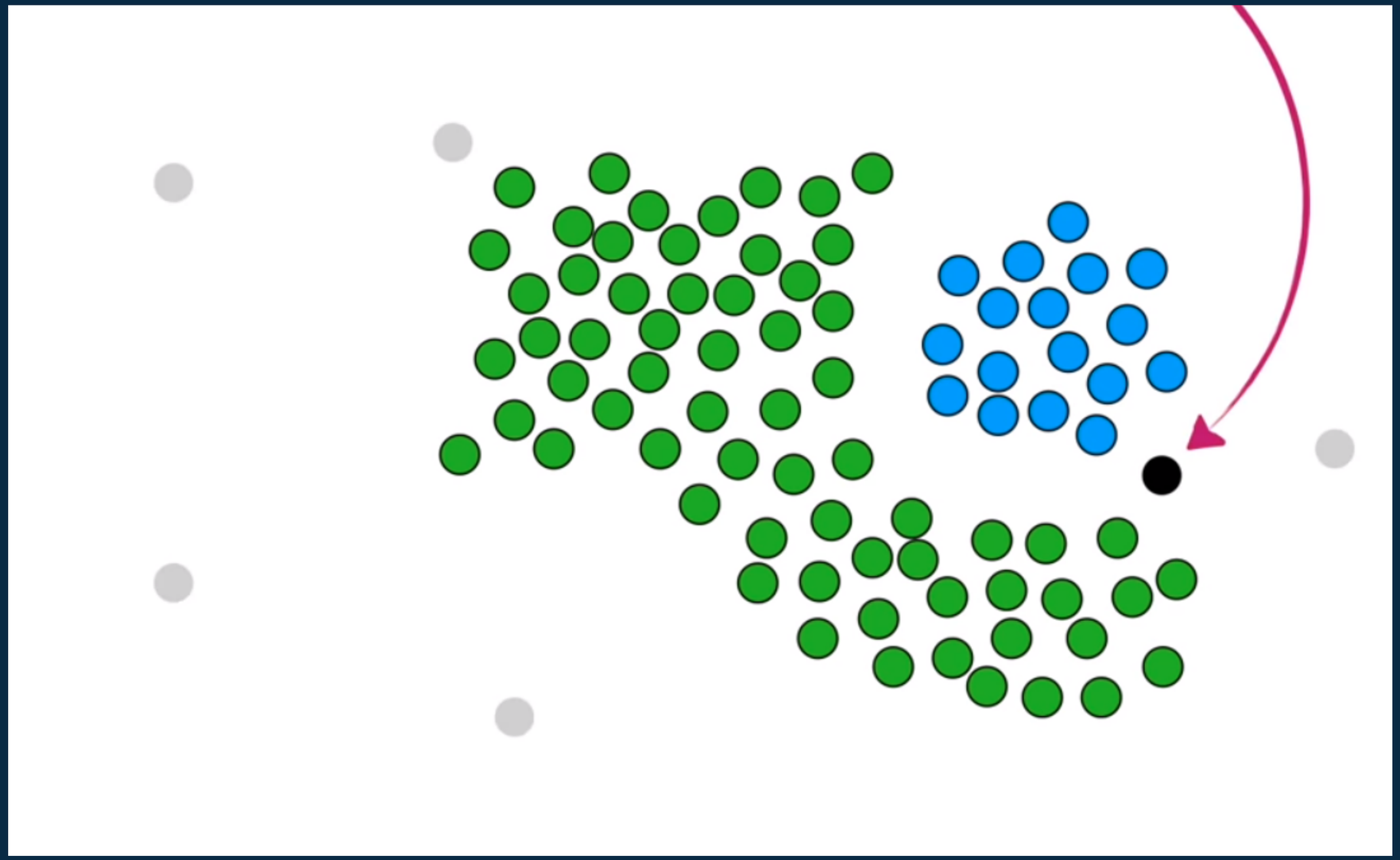

Non-Core Point

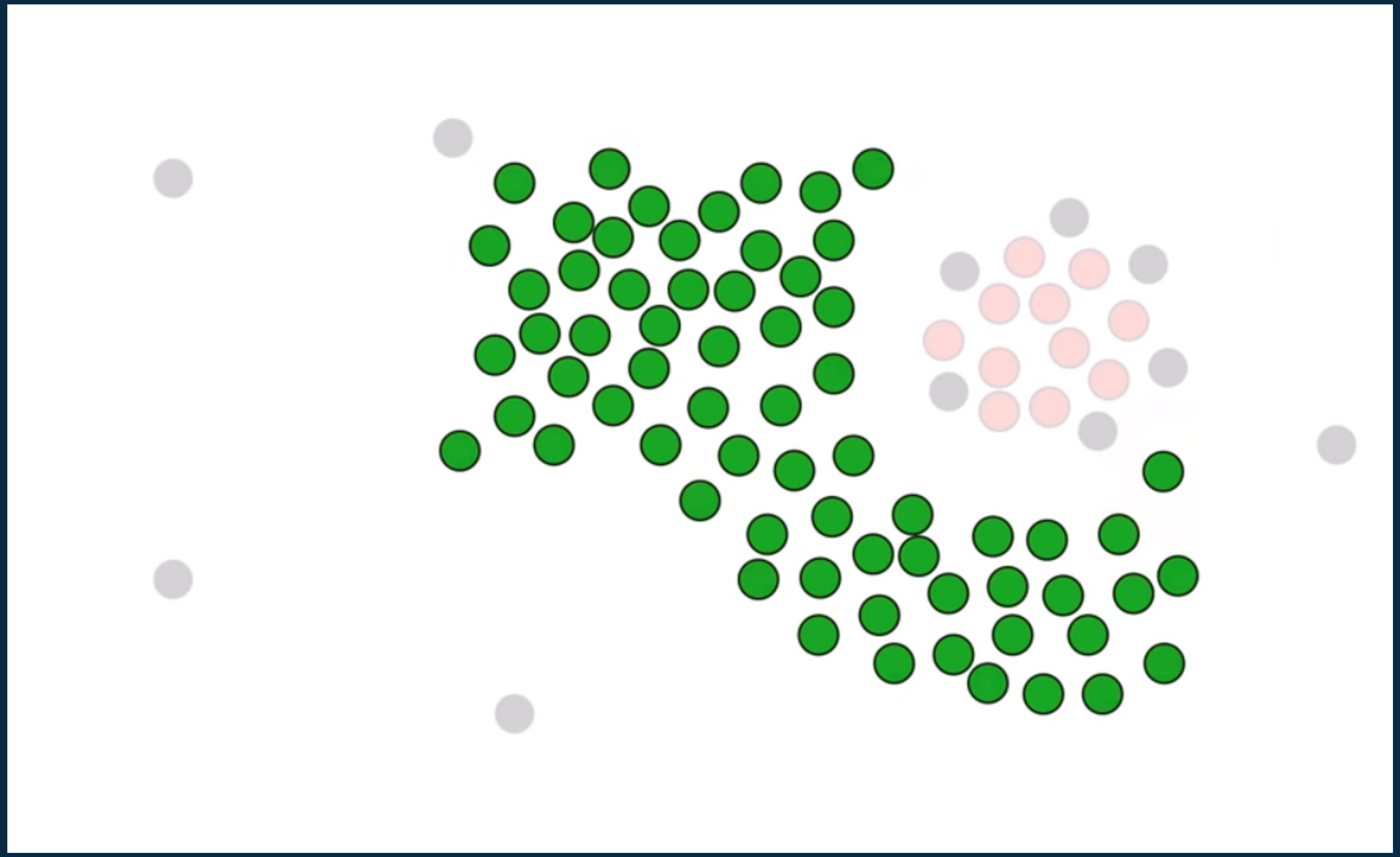


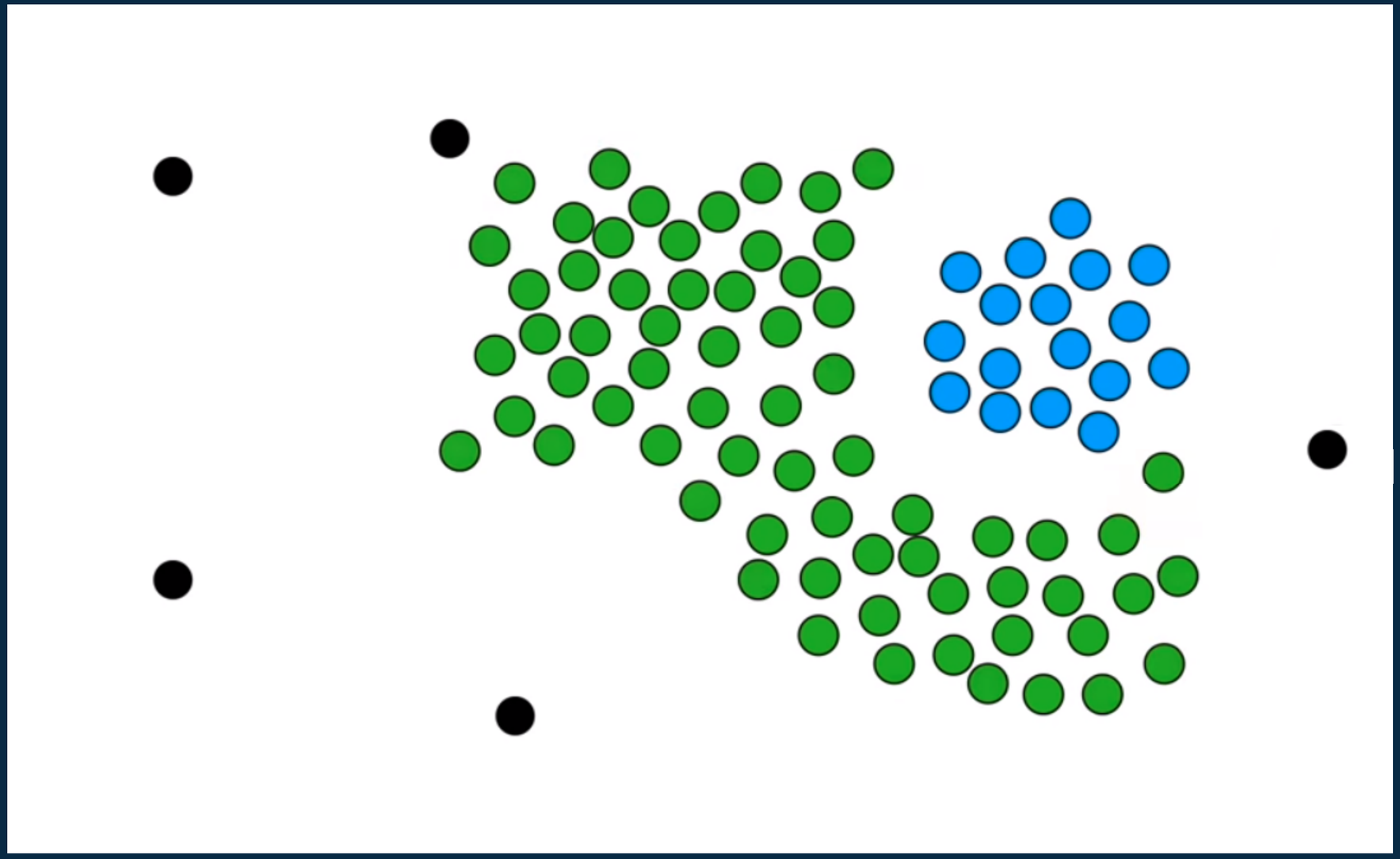






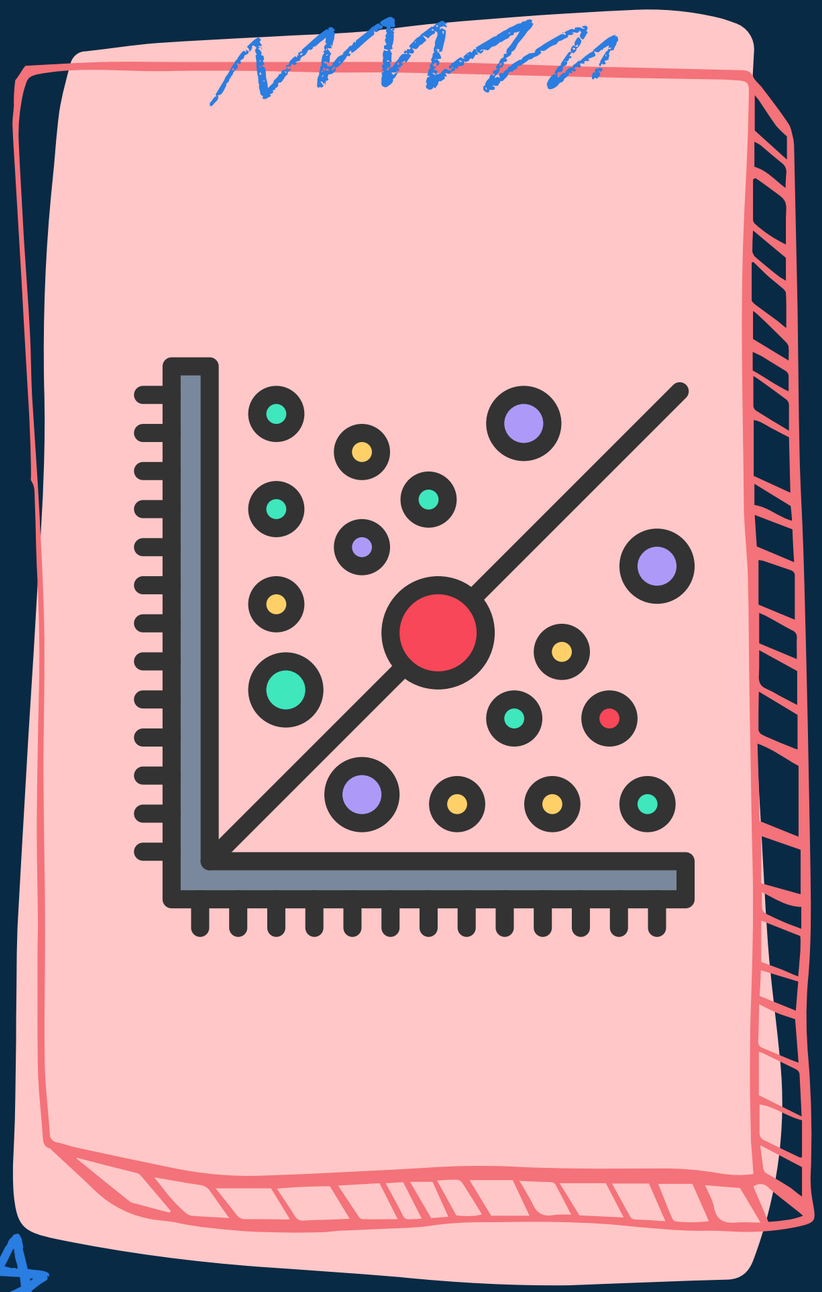






KEY CHARACTERISTICS OF DBSCAN

- Identifies clusters based on the density of data points rather than predefined shapes or distances.
- Doesn't require specifying the number of clusters beforehand, making it flexible and adaptable to different datasets.
- Capable of detecting clusters of various shapes and sizes, including irregular and non-convex clusters.
- Can effectively distinguish noise points from dense regions, making it suitable for datasets with irregularities.



ADVANTAGES OF DBSCAN

- Pattern Discovery
- Unsupervised Learning
- Data Understanding
- Scalability
- Flexibility
- Anomaly Detection

LIMITATIONS OF DBSCAN

- Sensitivity to parameter settings
- Difficulty with clusters of varying densities
- Not suitable for high-dimensional data
- Requires a dense region to define clusters
- Limited to Euclidean distance

APPLICATIONS OF DBSCAN

- Spatial data analysis
- Image segmentation
- Anomaly detection
- Customer segmentation in marketing
- Traffic flow analysis

CONCLUSION

- DBSCAN IS A WIDELY UTILIZED CLUSTERING ALGORITHM IN DATA MINING AND MACHINE LEARNING.
- ITS STRENGTH LIES IN IDENTIFYING CLUSTERS BASED ON LOCAL DENSITY WITHOUT NEEDING THE NUMBER OF CLUSTERS IN ADVANCE.
- DESPITE ITS ADVANTAGES, CAREFUL PARAMETER TUNING IS NECESSARY, AND IT MAY STRUGGLE WITH VARYING CLUSTER DENSITIES.
- UNDERSTANDING DBSCAN'S LIMITATIONS IS CRUCIAL FOR EFFECTIVE REAL-WORLD APPLICATION.
- DBSCAN'S CAPACITY TO HANDLE NOISE AND DETECT CLUSTERS OF ARBITRARY SHAPES MAKES IT INVALUABLE ACROSS DIVERSE DOMAINS.