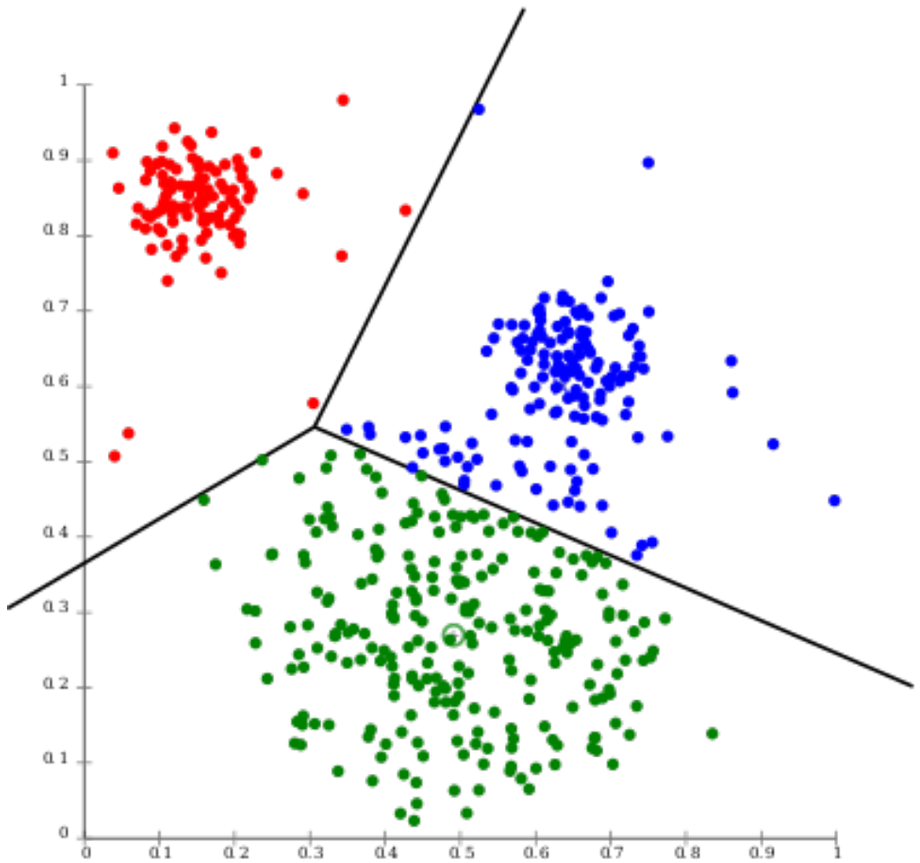
DENSITY BASED CLUSTERING

DECEMBER 12, 2017

MOTIVATION

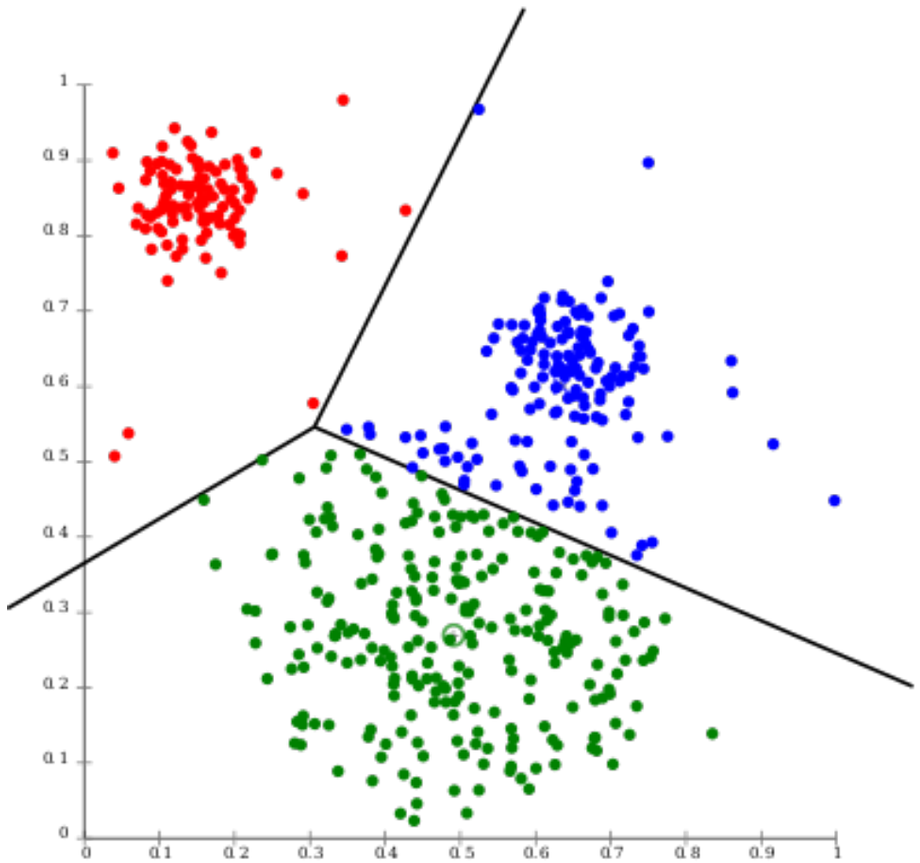
The central idea of **K-means clustering** is partitioning the space based on the **closest mean**.



The blue points are closer to the blue mean and so forth…

MOTIVATION

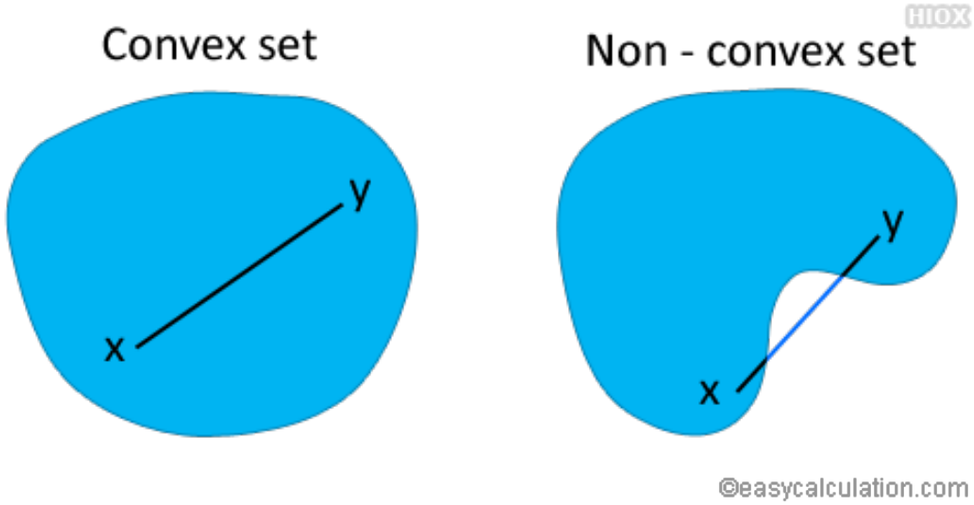
The central idea of **K-means clustering** is partitioning the space based on the **closest mean**.

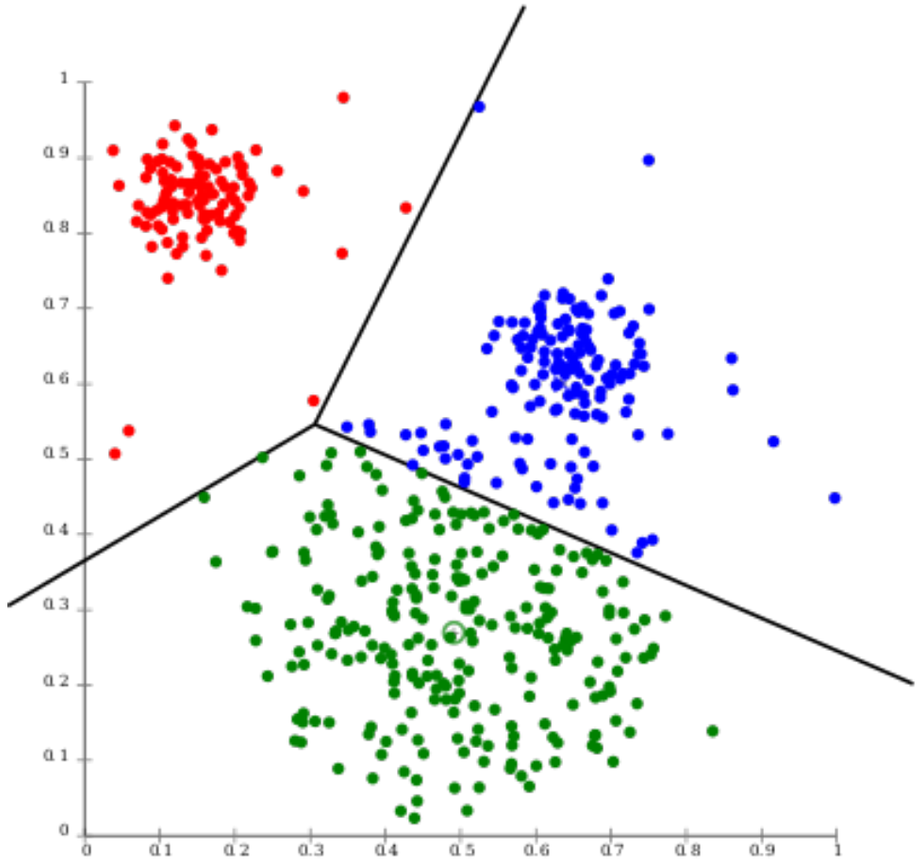


The clusters are all convex regions

MOTIVATION

What if the clusters are not convex?

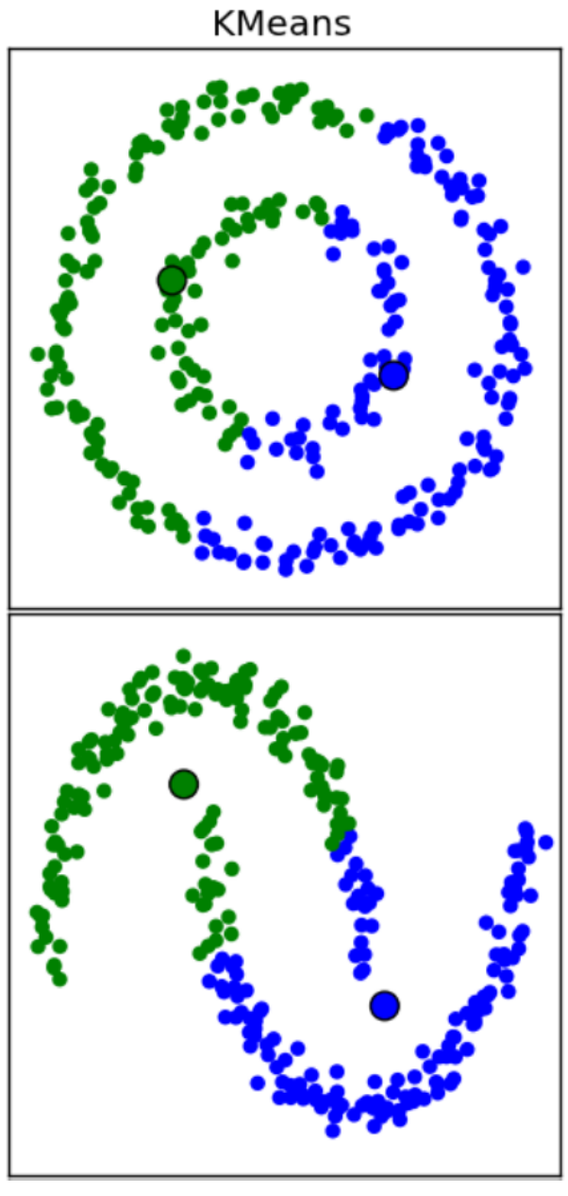




MOTIVATION

What if the clusters are not convex?

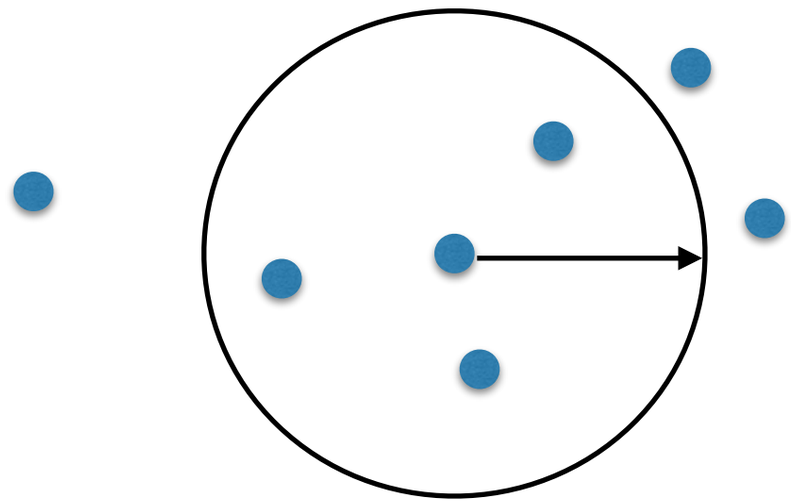
K-means does not handle such examples well



DENSITY-BASED CLUSTERING

The density-based clustering algorithm **(DBSCAN)** overcomes these deficiencies.

It has two hyper-parameters — **radius** and **minpoints**



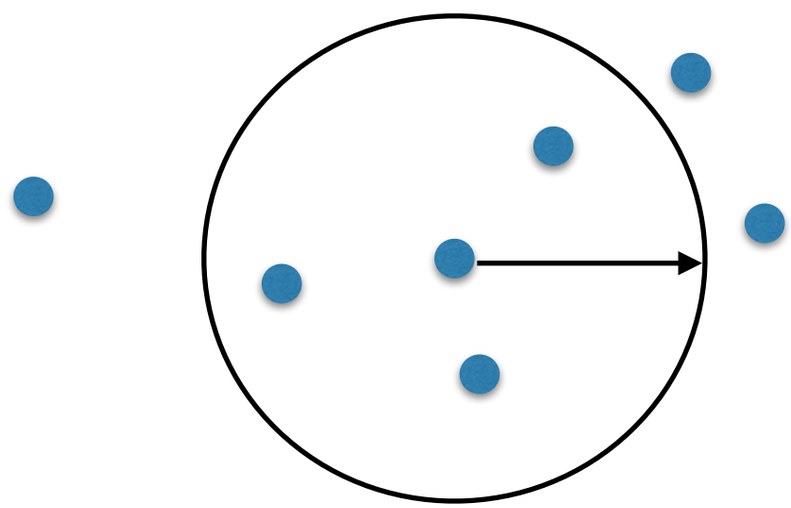
Radius — distance that determines if points are close to each other

DENSITY-BASED CLUSTERING

For this particular radius, we have three **reachable** points

minpoints — number of reachable points needed to say a region is "**dense**"

— a point is said to be a "**core point**" if it has at least minpoints reachable points.

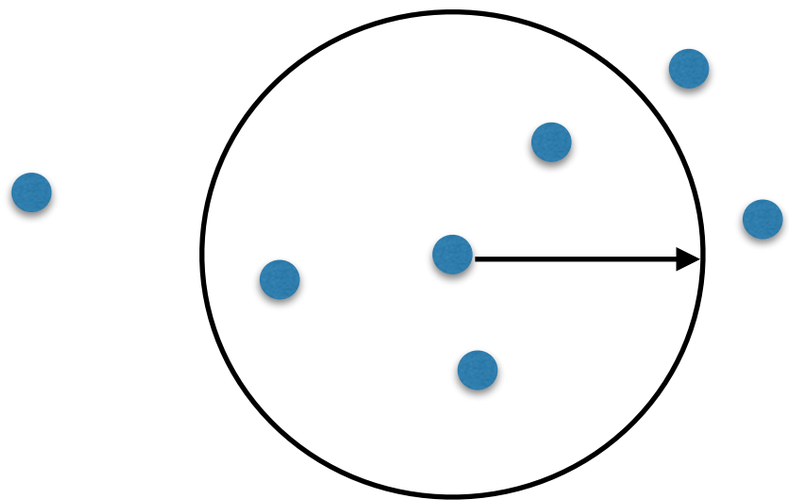


If minpoints = 3, then the centre is a core point.

DENSITY-BASED CLUSTERING

Each core point defines a cluster.

If two core points are reachable from one another, merge clusters.



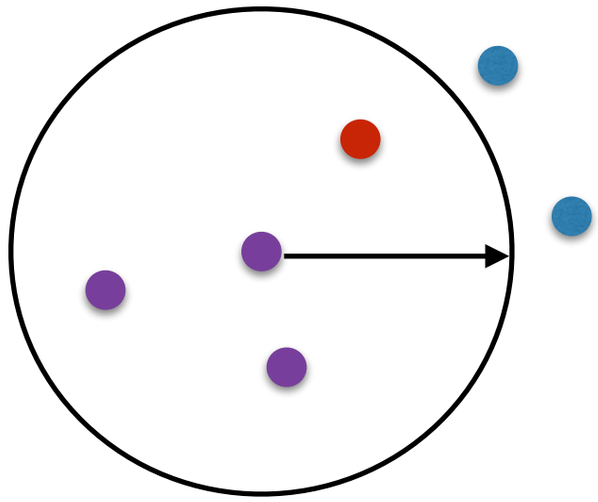
If minpoints = 3, then the centre is a core point.

DBSCAN ALGORITHM

Initialize with an arbitrary point. Check if it is a core point.

If it is, add all reachable points to its cluster.

For each newly-added point, check if it is a core point and add its neighbors. Continue till every point has been checked in the cluster.



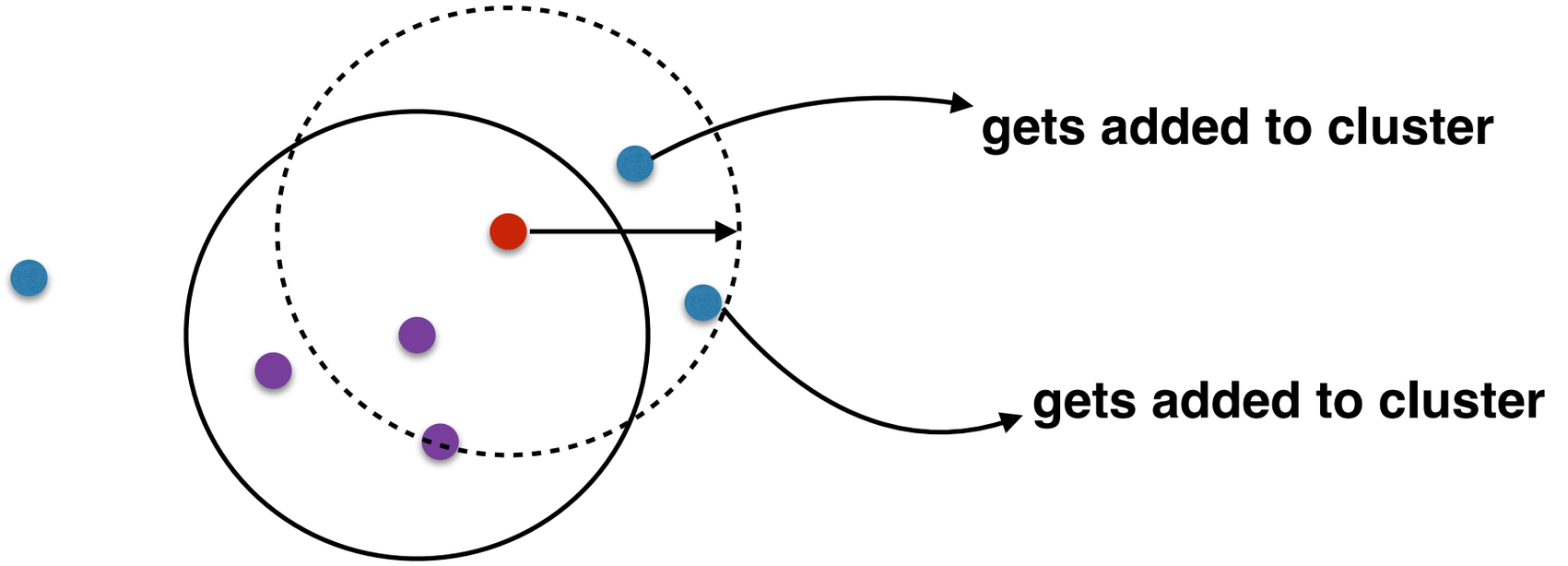
The centre is a core point, and all 3 points get added to cluster.

DBSCAN ALGORITHM

Initialize with an arbitrary point. Check if it is a core point.

If it is, add all reachable points to its cluster.

For each newly-added point, check if it is a core point and add its neighbors. Continue till every point has been checked in the cluster.

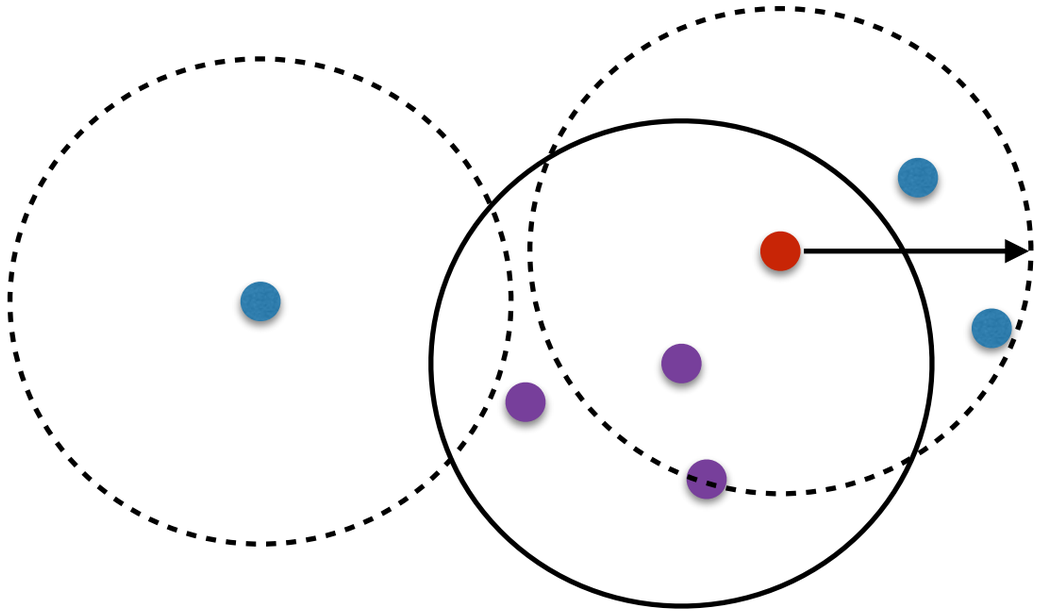


DBSCAN ALGORITHM

Initialize with an arbitrary point. Check if it is a core point.

If it is, add all reachable points to its cluster.

For each newly-added point, check if it is a core point and add its neighbors. Continue till every point has been checked in the cluster.



Not a core point, so no cluster assigned

PROS AND CONS

Pros

- Non-parameteric approach to clustering — no need to pre-specify the number of clusters we are looking for
- Can identify non-convex clusters

Cons

- Some points are not assigned to a cluster
- Computationally expensive to identify which cluster a new observation belongs to

ACKNOWLEDGEMENTS

- *Some of the figures and concepts in this presentation are adapted from "Density Based Clustering", Machine Learning and Data Mining, Mark Schmidt, UBC*