Fundamentals of Artificial Intelligence

COMP307/AIML420

Tutorial 3: Ensembles & Clustering

Dr. Heitor Murilo Gomes
heitor.gomes@vuw.ac.nz
http://www.heitorgomes.com

Information

- Assignment 1 (due on week 5 - 27 March 2024)

- Submission system is open!

- Helpdesks as available daily (Monday to Friday, 3pm to 4pm) on **CO242B**

- Next week starts the 2 hours helpdesks (from Thursdays onwards)

Ensemble learning

- **Diversity**, **combination** and **base learner**

- **Several reasons to use them** (statistical, computational and representational)

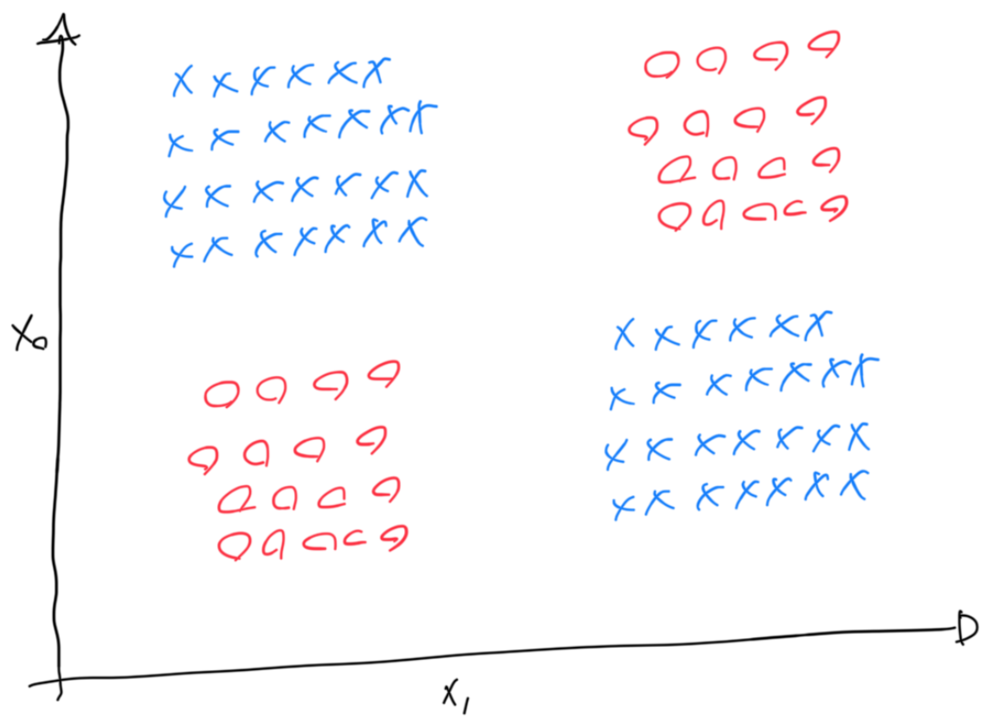
- **Bagging** and **Random Forest**

Representational

Several simple classifiers can approximate complex classification boundaries.

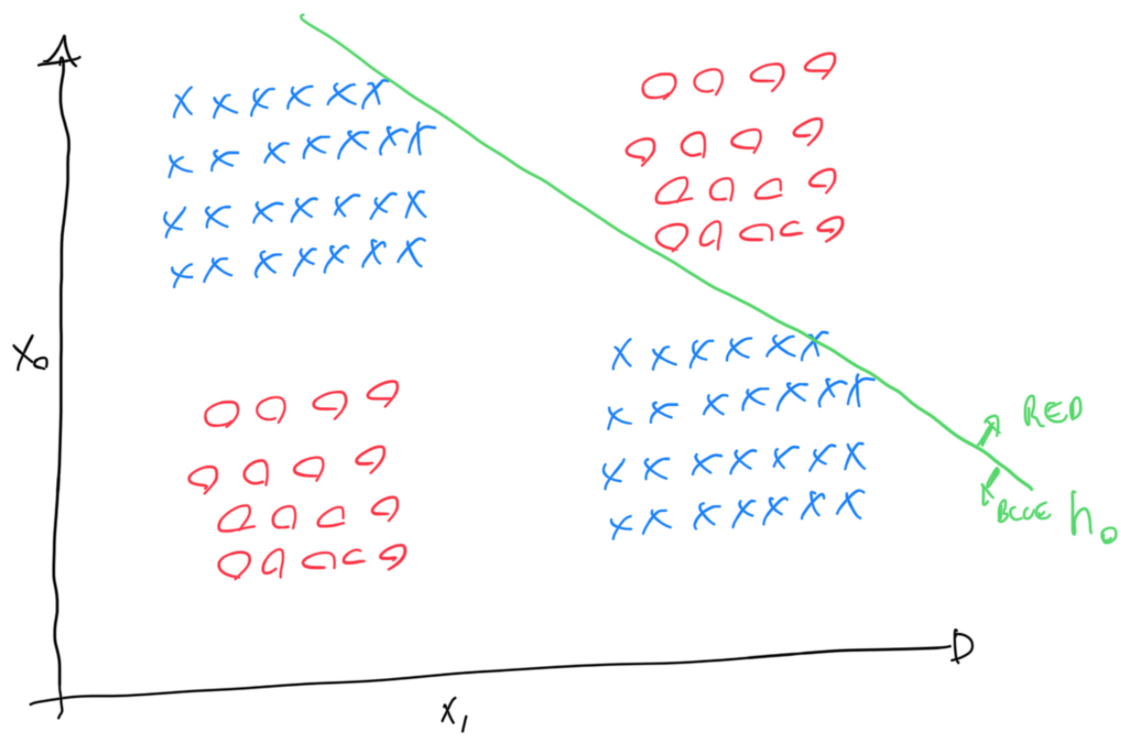
Representational

Several simple classifiers can approximate complex classification boundaries.



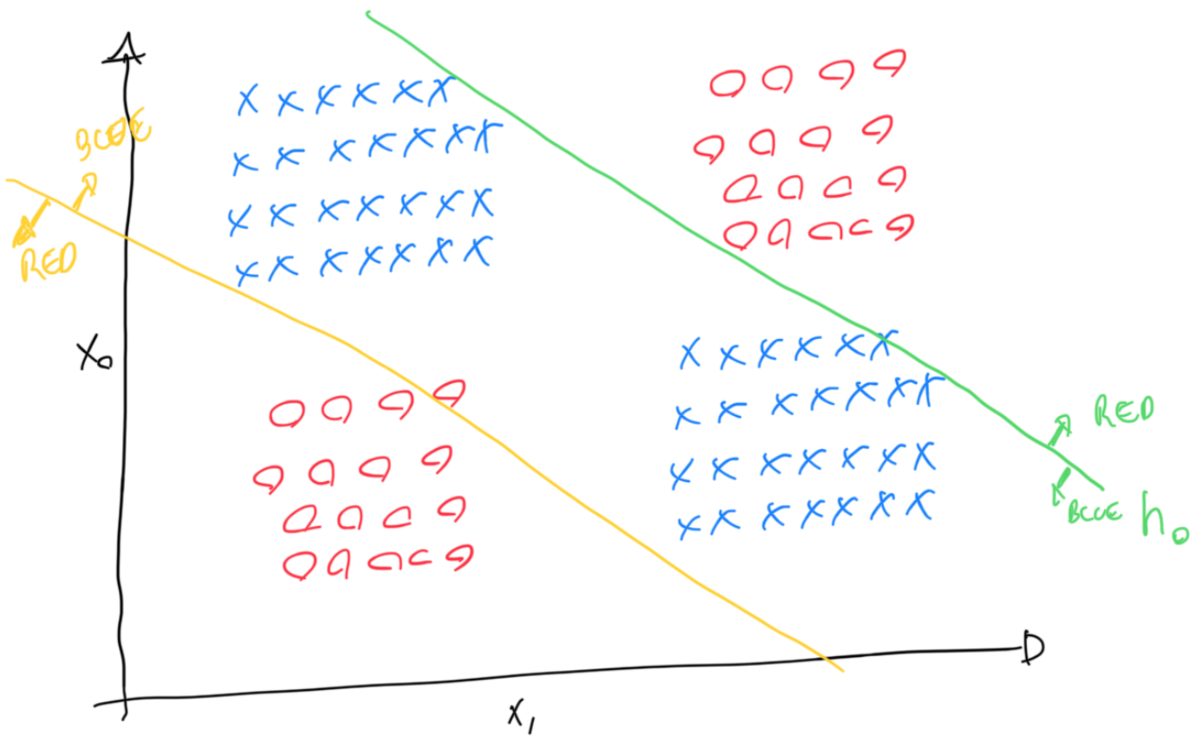
Representational

Several simple classifiers can approximate complex classification boundaries.



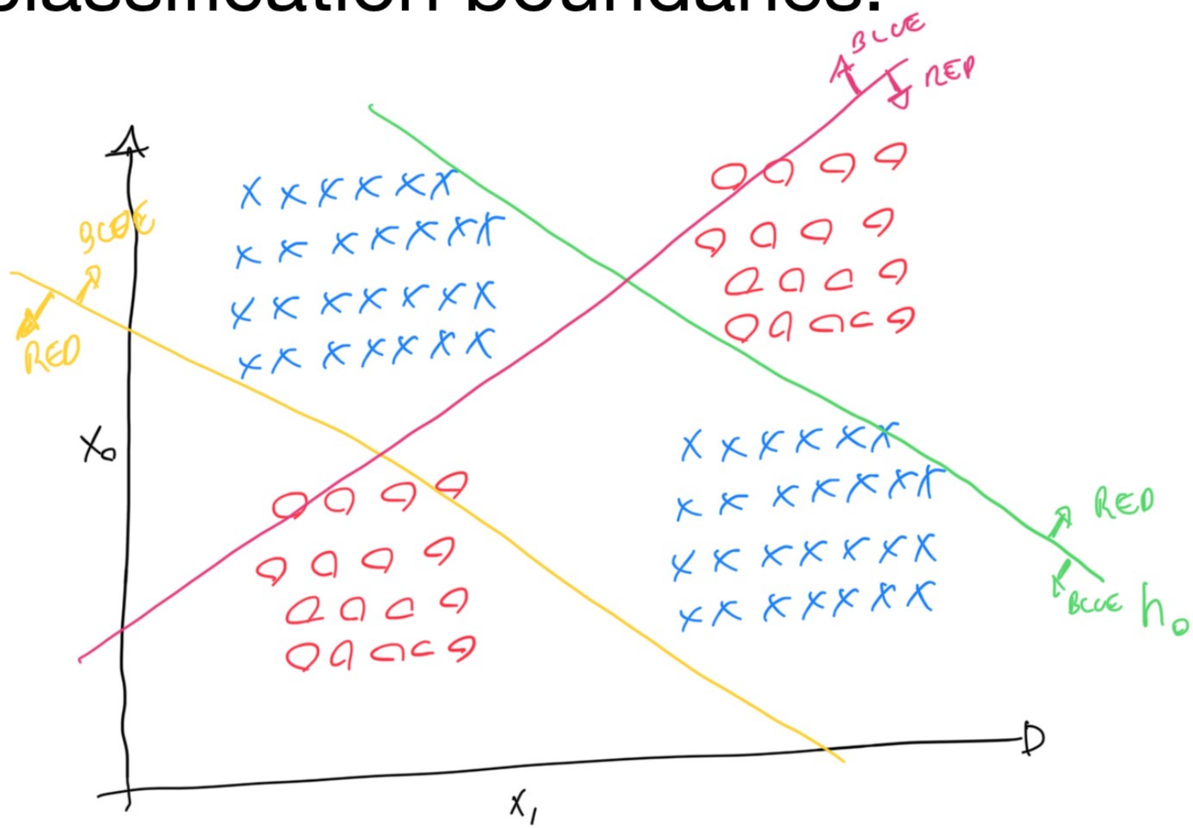
Representational

Several simple classifiers can approximate complex classification boundaries.



Representational

Several simple classifiers can approximate complex classification boundaries.

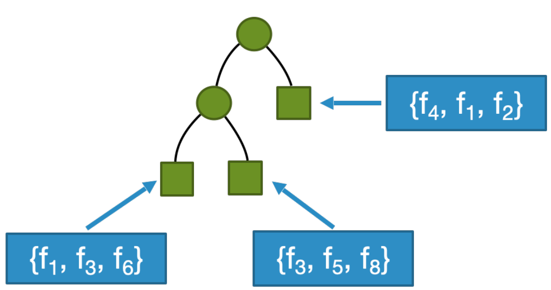


Bagging and Random forest

- **Bagging** train learners on different subsets of instances (bootstrapping)
- **Random forest** besides training on different subsets of instances, also randomizes the subsets of features used for split decisions

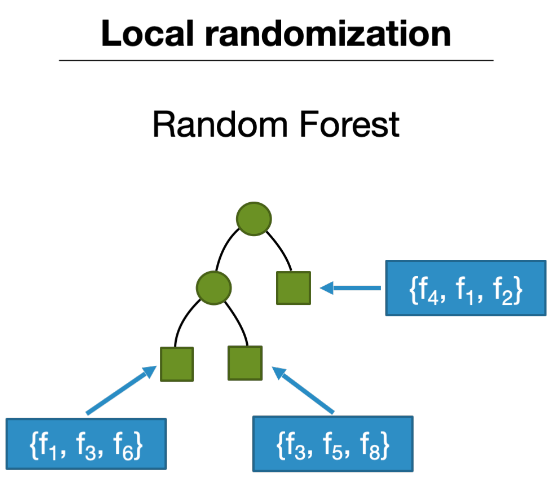
Local randomization

Random Forest



Bagging and Random forest

- **Bagging** train learners on different subsets of instances (bootstrapping)
- **Random forest** besides training on different subsets of instances, also randomizes the subsets of features used for split decisions



Why? Create a diverse set of base learners

Base learner must be **unstable**

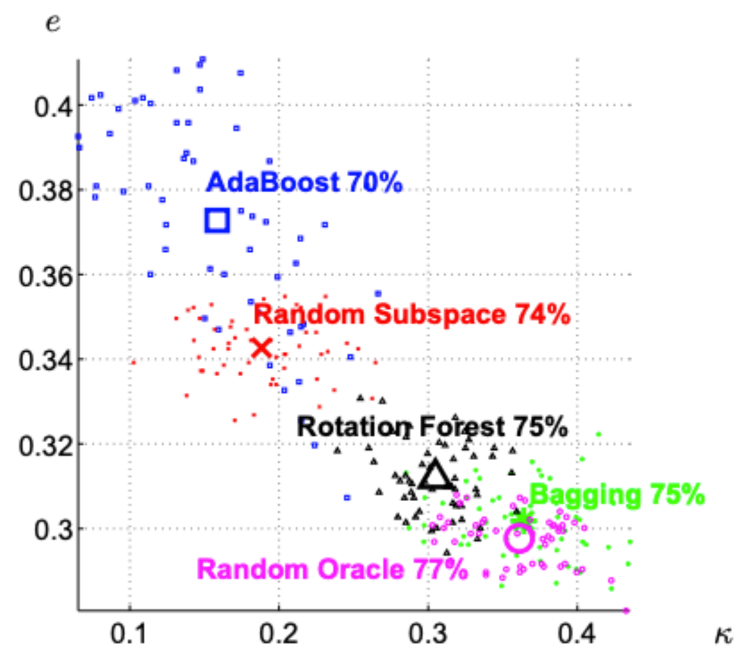
Measuring diversity?

- Interrater agreement measure **Kappa κ** (see [1]) and **Kappa-error diagrams** (see [2])

Pairwise individual error (y-axis) vs pairwise diversity (x-axis)

$\kappa = 1$ means identical classifiers, $\kappa = 0$ indicates independent classifiers

We can use these diagrams to prune the ensemble



Adapted from [1]

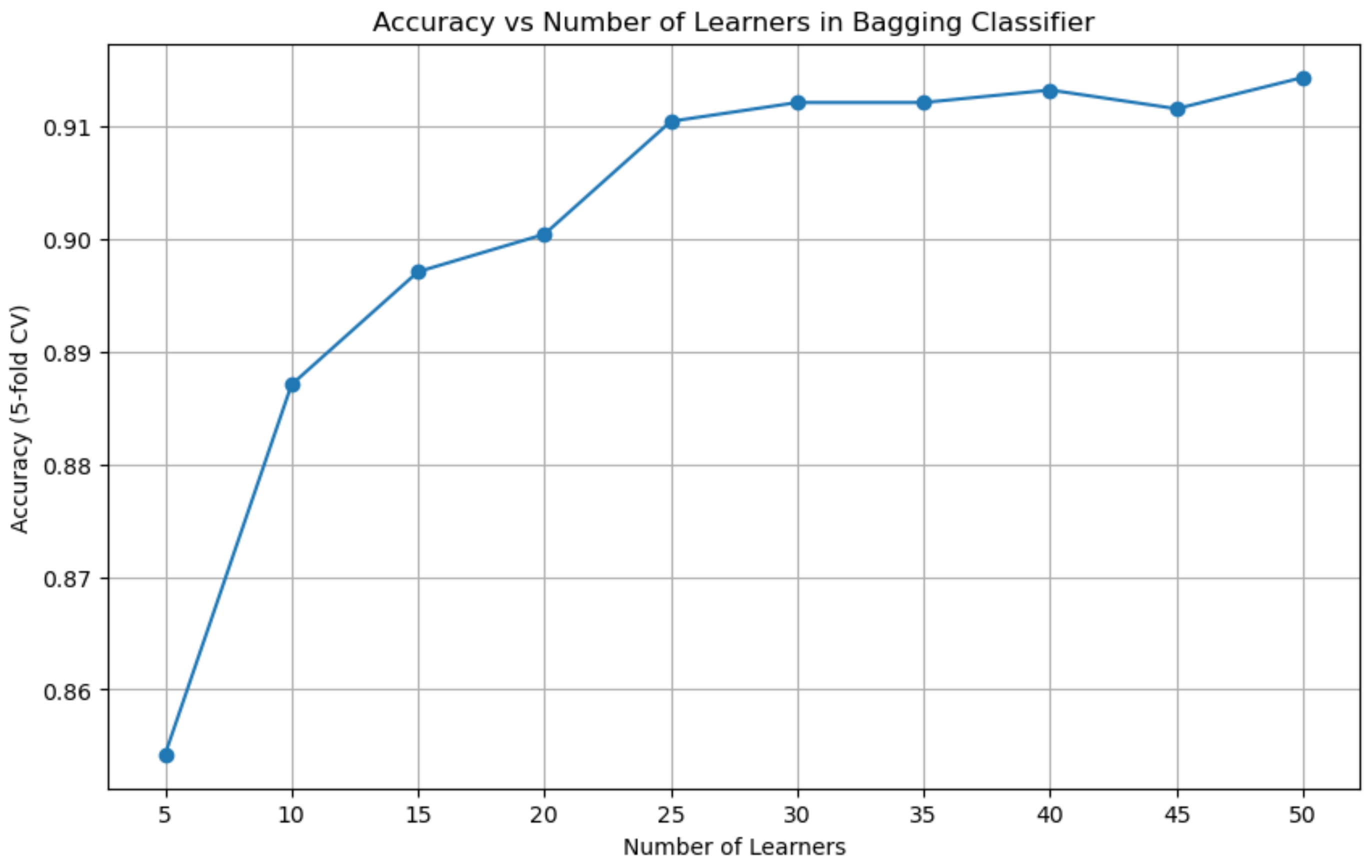
[1] Kuncheva, Ludmila I. "A bound on kappa-error diagrams for analysis of classifier ensembles." *IEEE TKDE*, 2011

[2] D. D. Margineantu and T. G. Dietterich. Pruning adaptive boosting. ICML, 1997

Bagging*

- Impact of hyperparameters

  number of base learners: 5 to 50

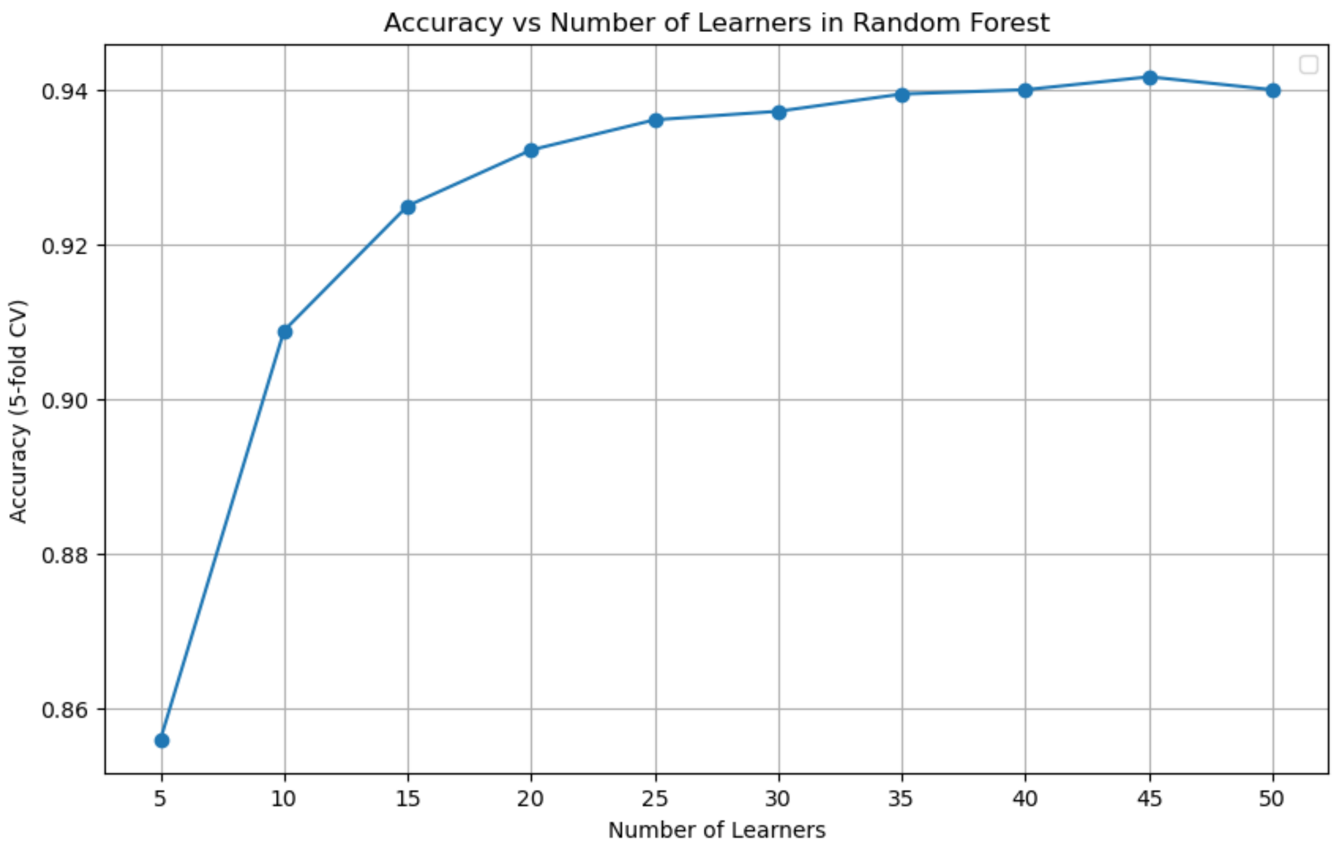


* Digits dataset

Random forest*

- Impact of hyperparameters

  number of base learners: 5 to 50 learners

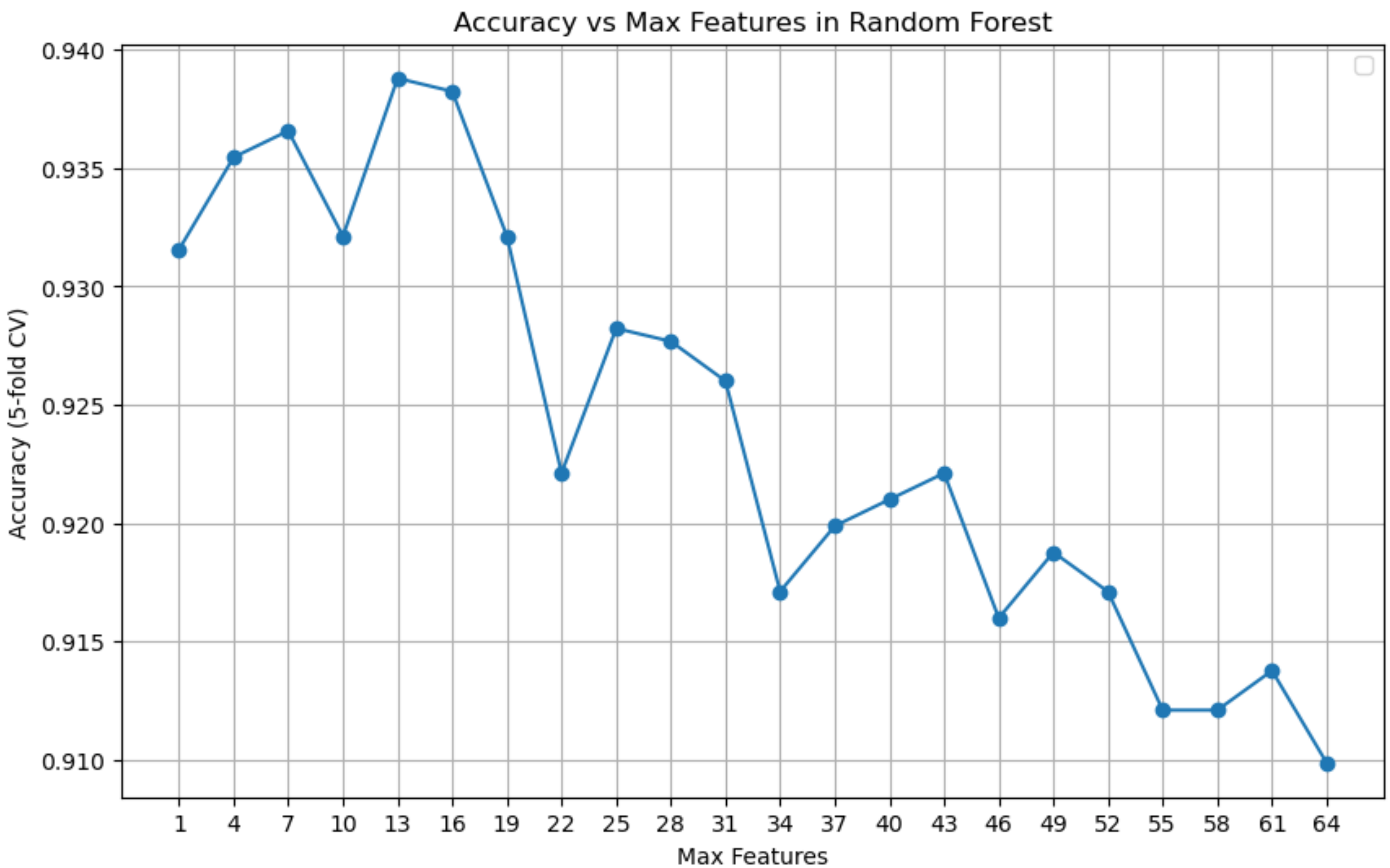


* Digits dataset

Random forest*

- Impact of hyperparameters

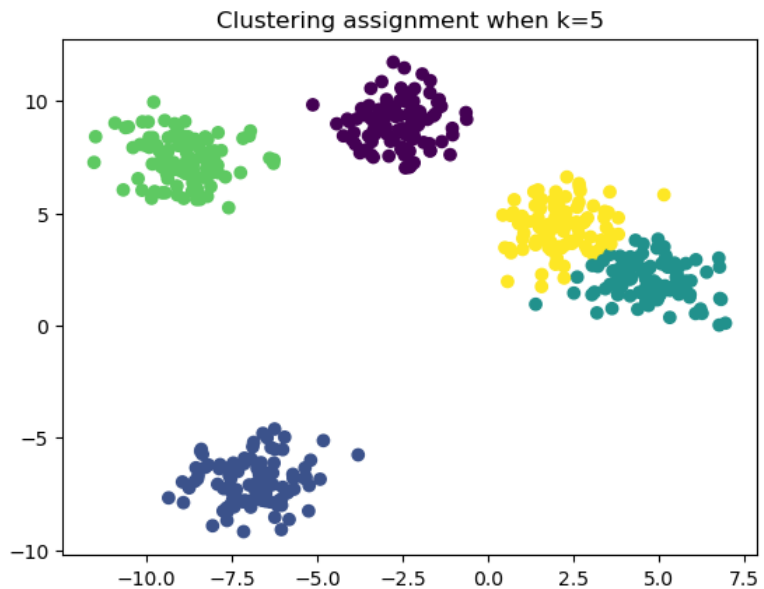
  subspace size (i.e. max features): 1 to 64



* Digits dataset

Clustering

- **Goal:** identify patterns or **structures** in the data that are not **immediately apparent**

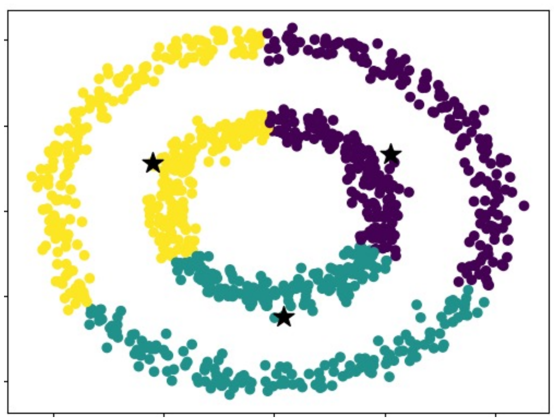


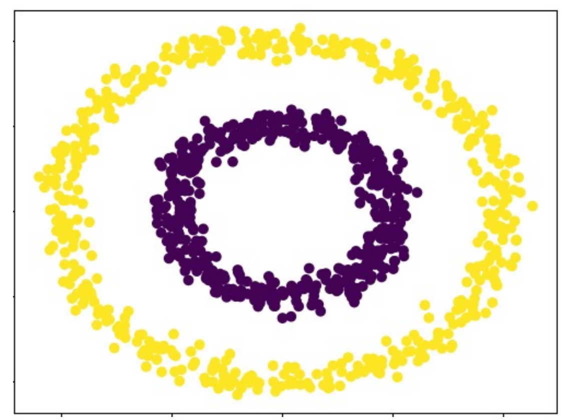
Clustering applications

- **Image segmentation** (segment an image into multiple regions, each of which corresponds to a distinct object or part of the image)
- **Customer/Market/Product segmentation** (identify groups to guide market research efforts)
- **Document clustering** (organize search results, topic identification, preprocessing for text classification, …)
- **Anomaly detection** (identifying rare or unexpected events)

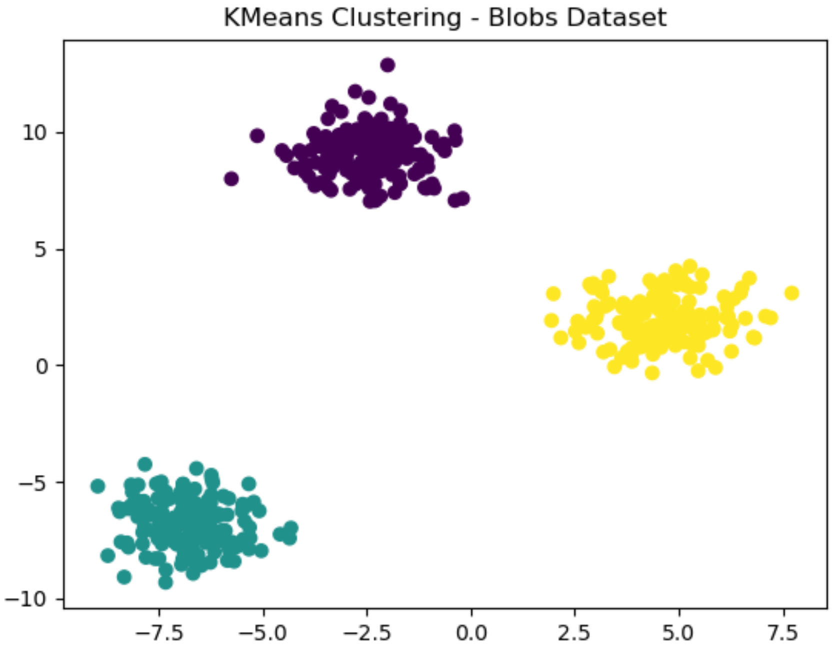
K-means & DBSCAN

- K-means
  - **K**: number of clusters
  - Centroid-based
- DBSCAN
  - **eps:** radius of the neighborhood
  - **min_points:** minimum number of points in a neighborhood
  - Density-based

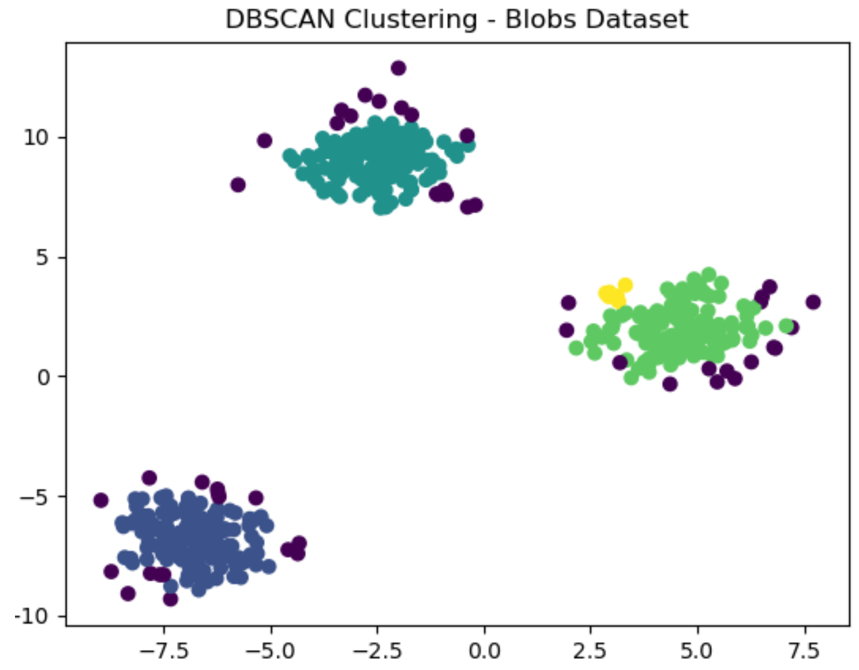

K-means


DBSCAN

More examples – Blobs dataset

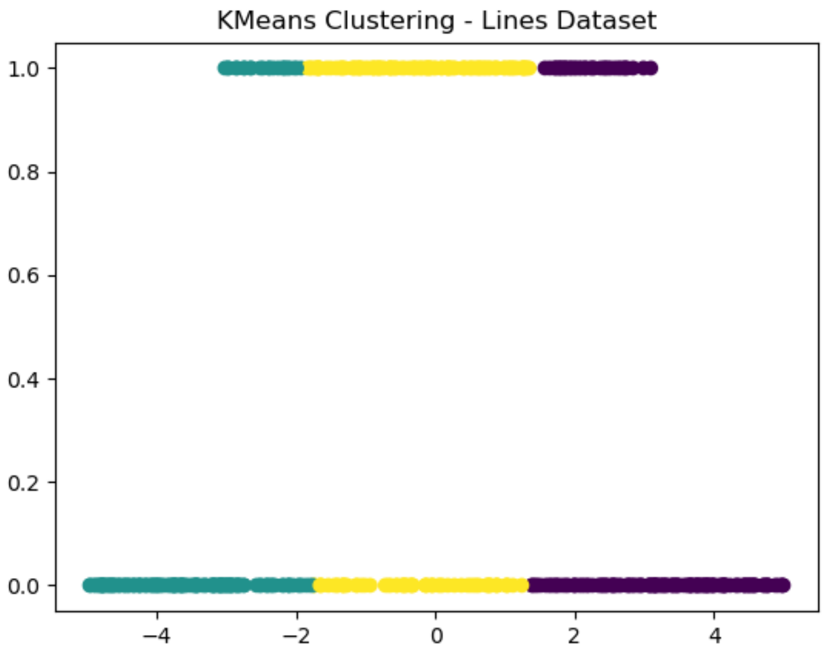


K-means

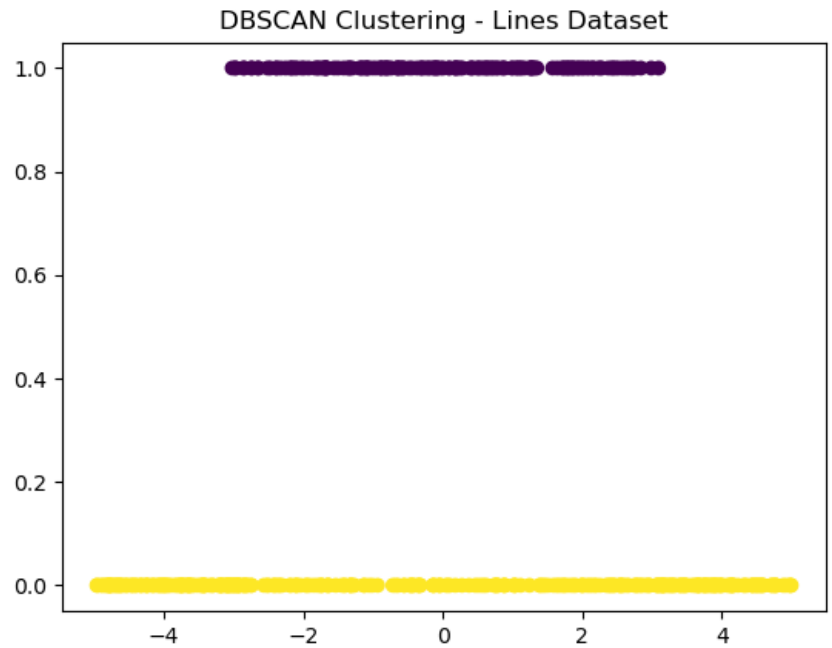


DBSCAN

More examples – Lines dataset

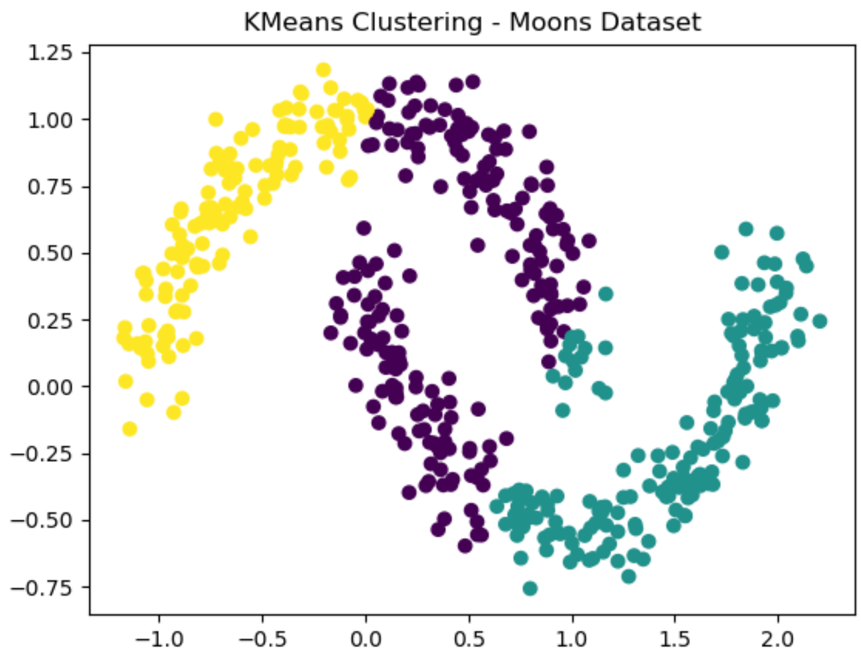


K-means

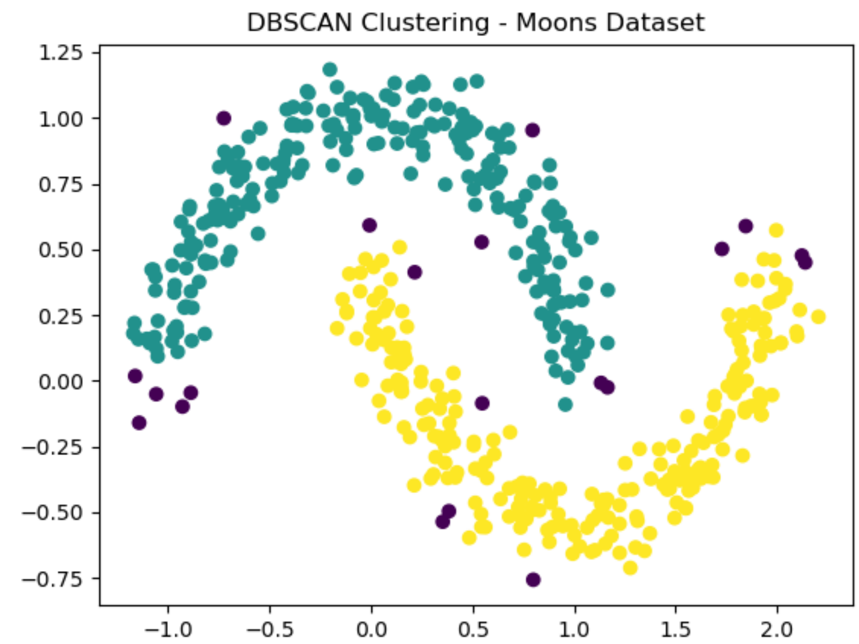


DBSCAN

More examples – Moons dataset



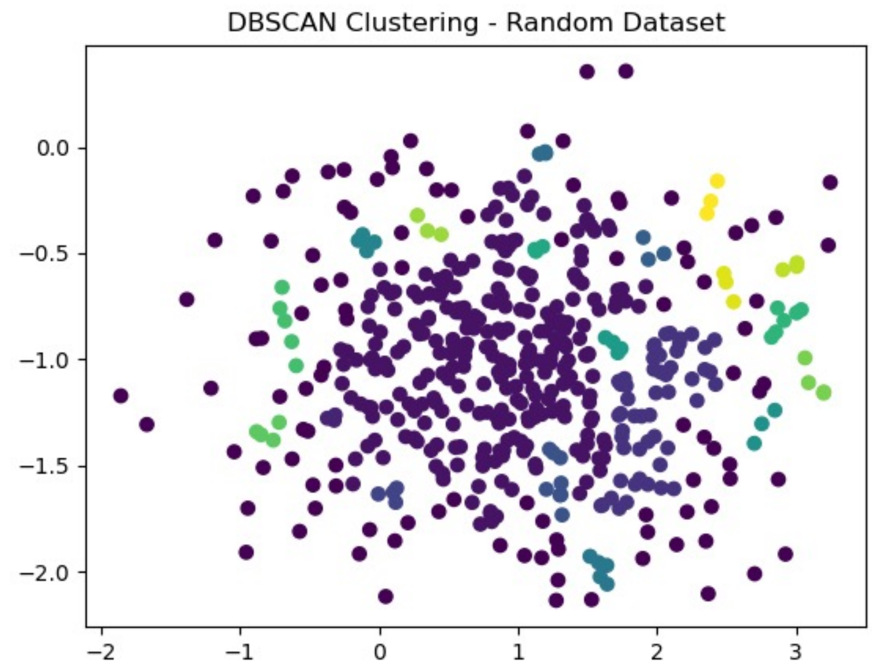
K-means



DBSCAN

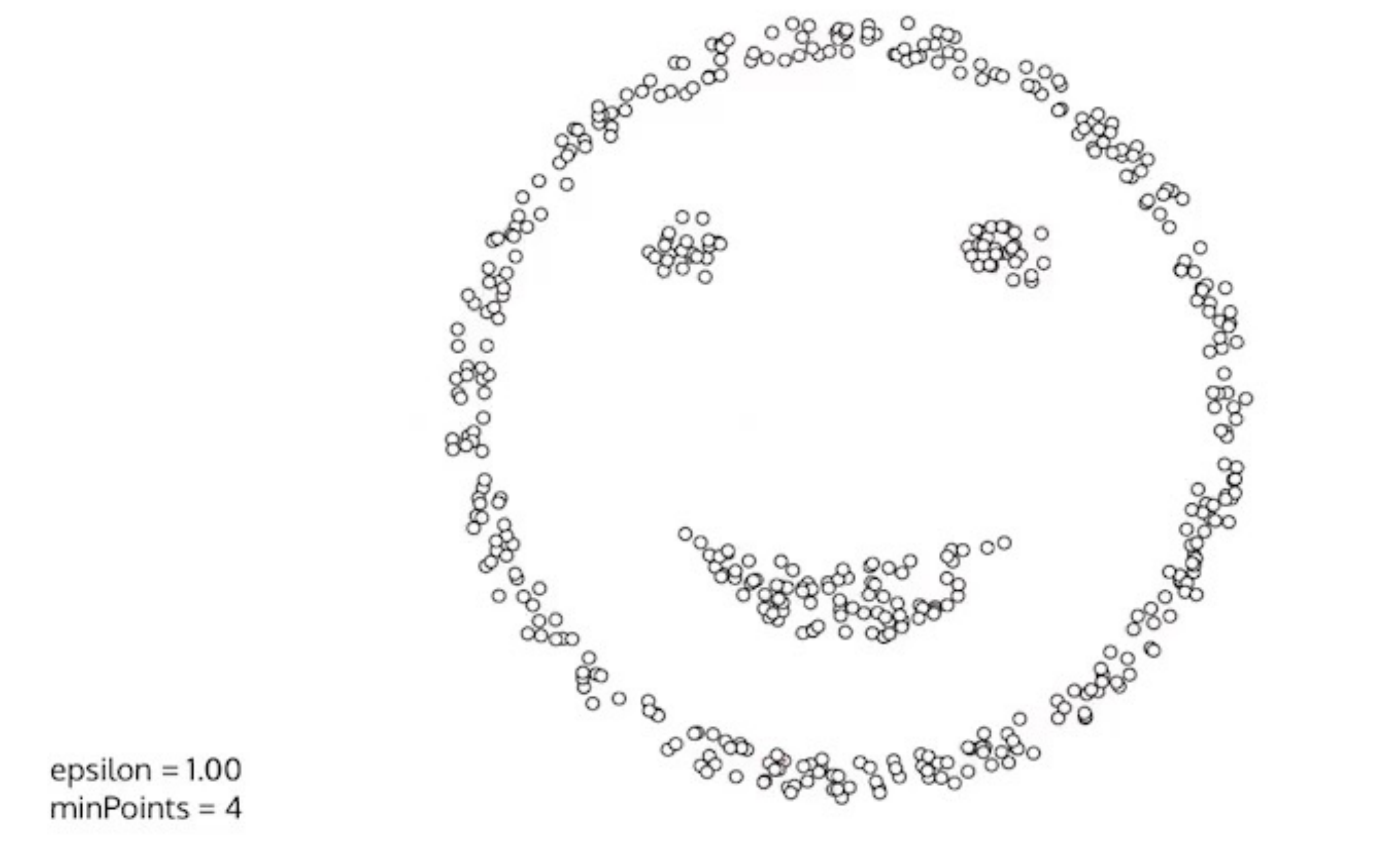
More examples – Random dataset

K-means



DBSCAN

DBSCAN example



Source: naftaliharris.com/blog/visualizing-dbscan-clustering/

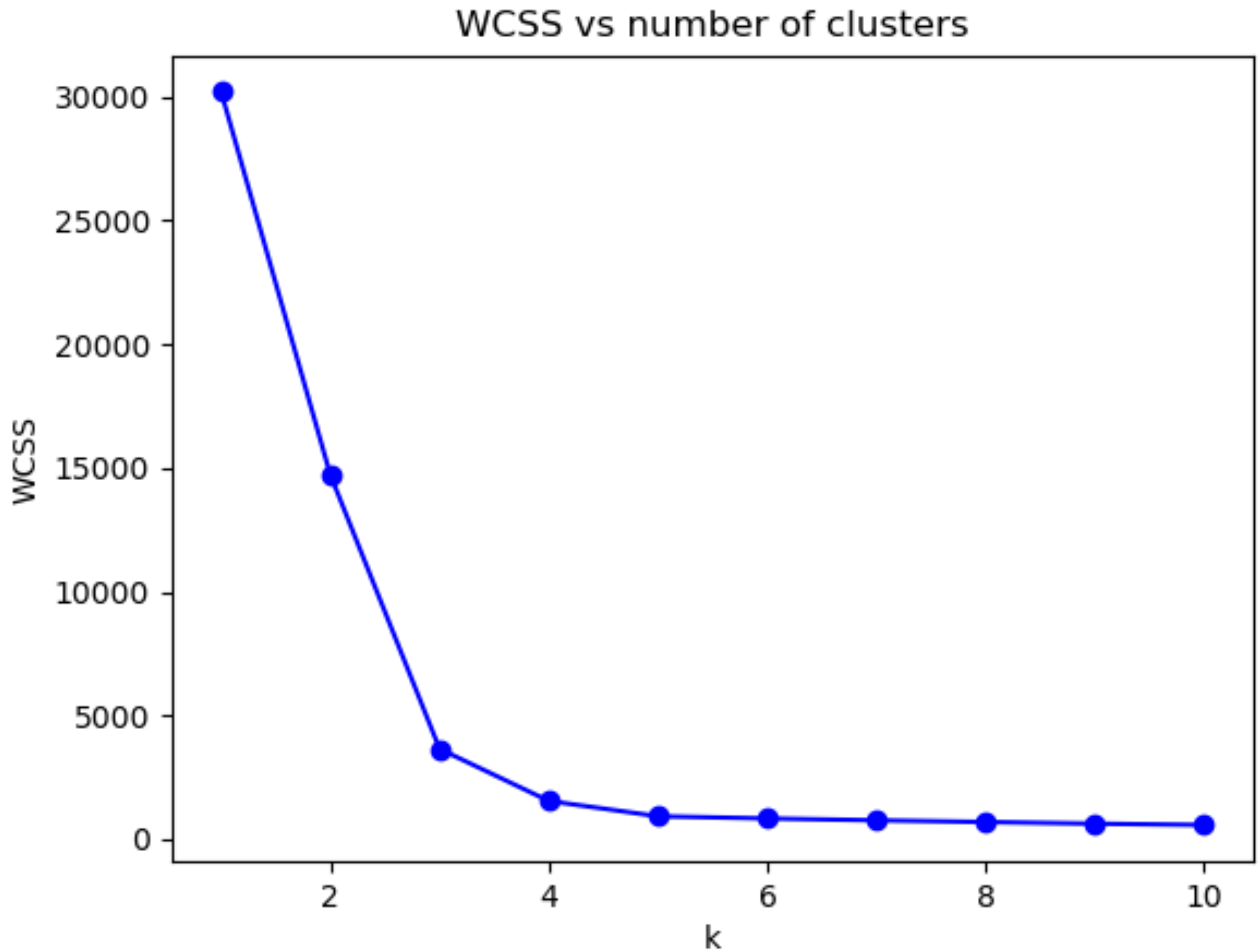
Elbow method for k-means

1. Use a **clustering quality measure** to assess the quality of different clustering executions
2. Plot such measure **varying k**
3. Where we find the "**elbow**" is the number of appropriate clusters

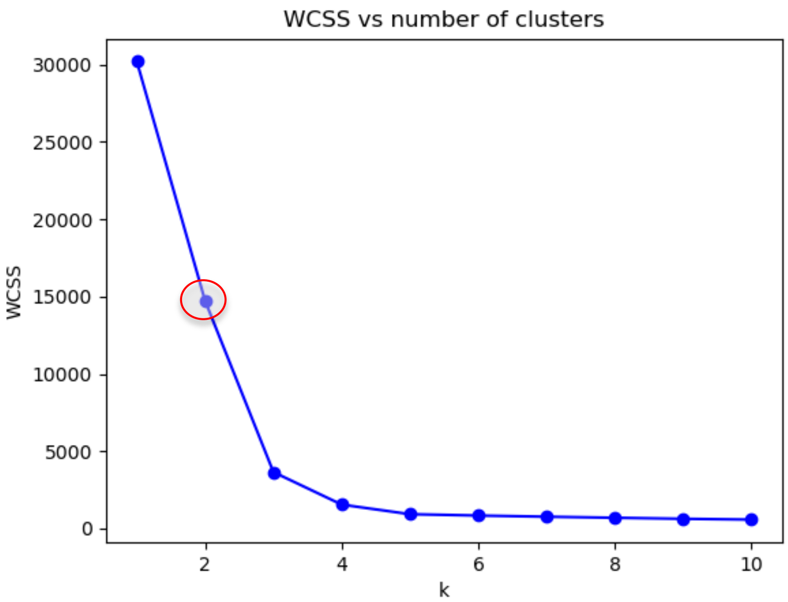
Elbow method for k-means

1. Use a **WCSS** to assess the quality of different clustering executions
2. Plot **WCSS varying k**
3. Where we find the "**elbow**" is the number of appropriate clusters

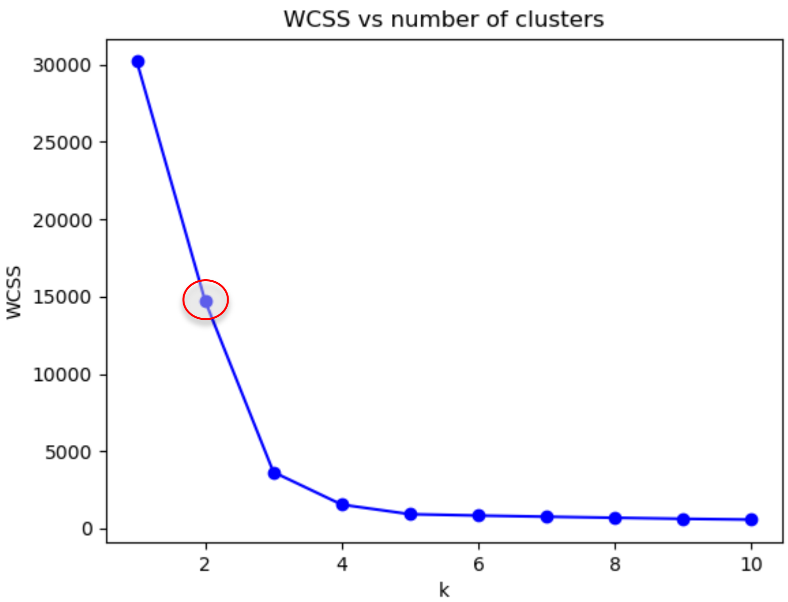
Elbow method for k-means

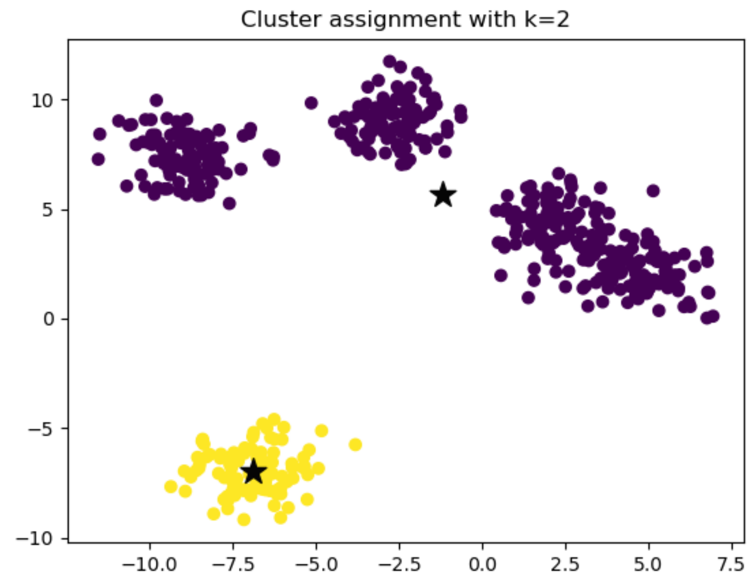


Elbow method for k-means

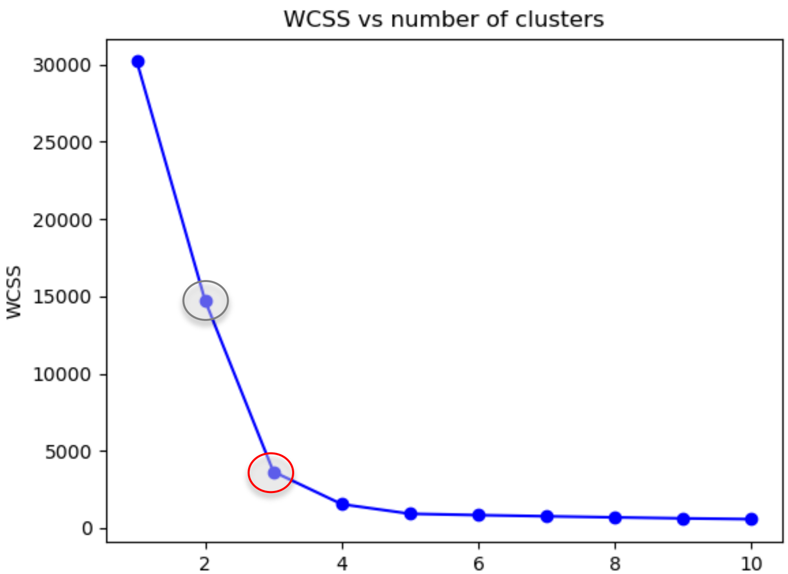


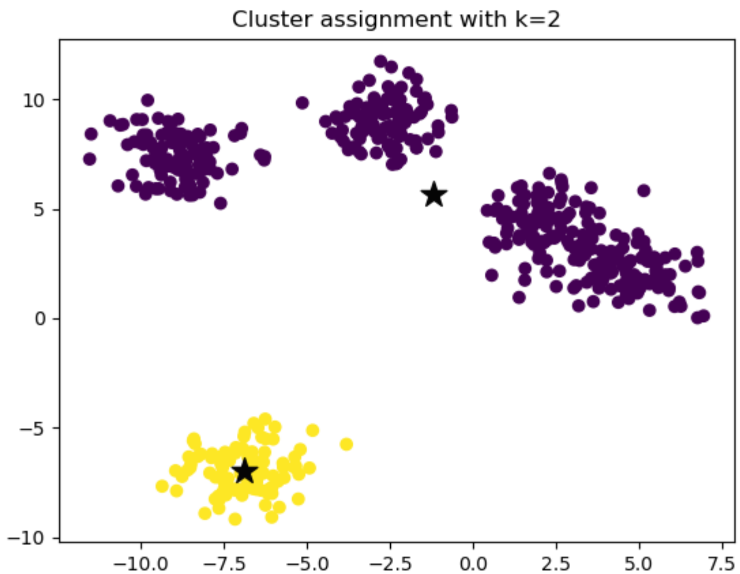
Elbow method for k-means

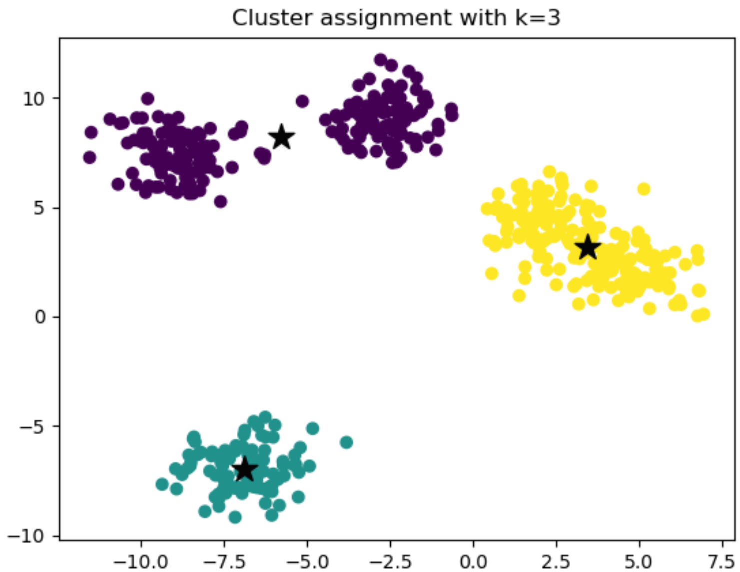




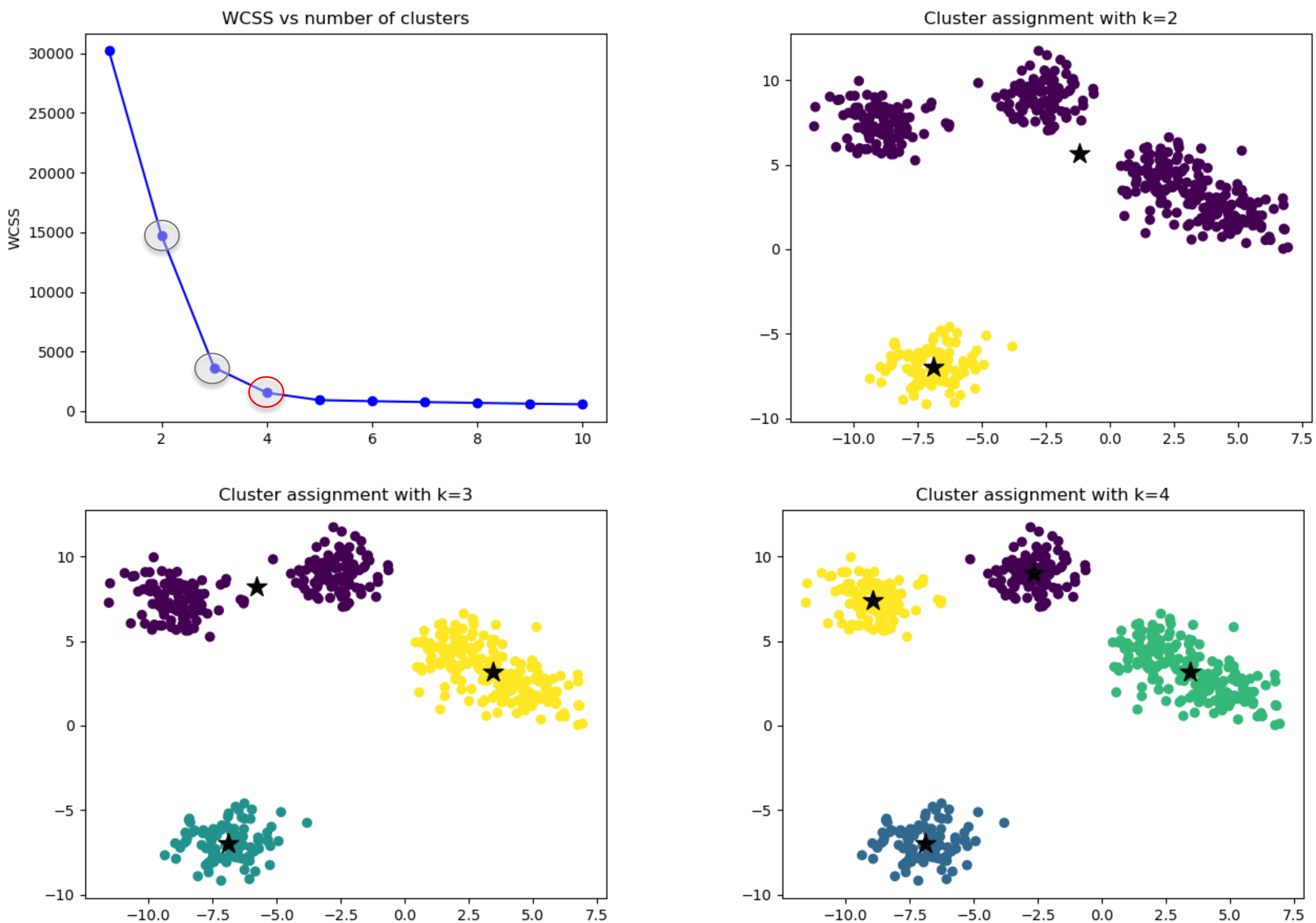
Elbow method for k-means



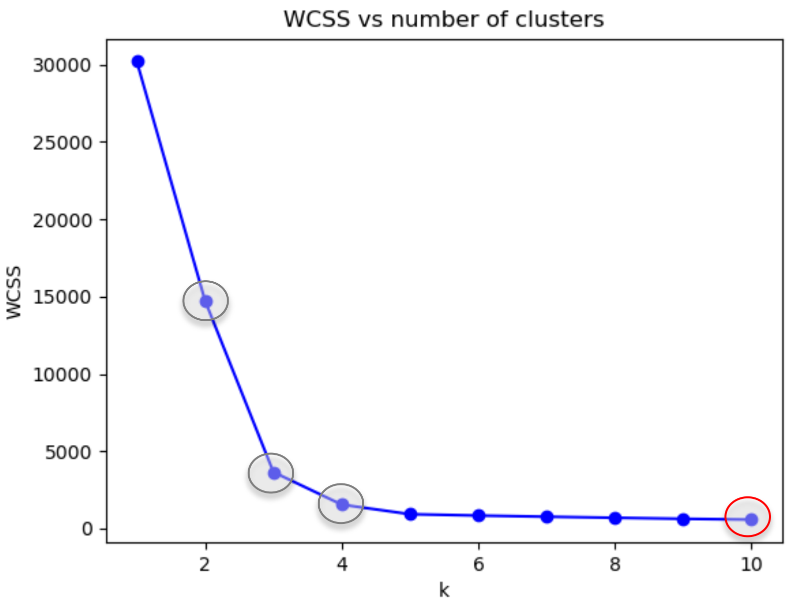


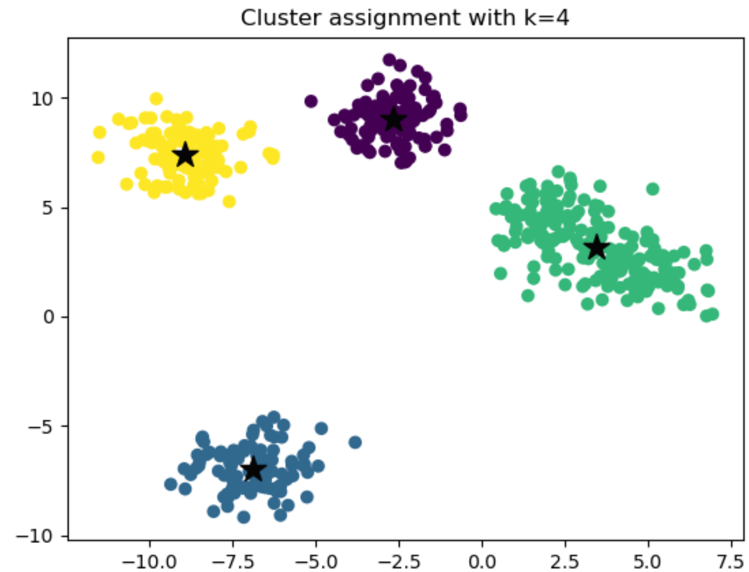


Elbow method for k-means



Elbow method for k-means





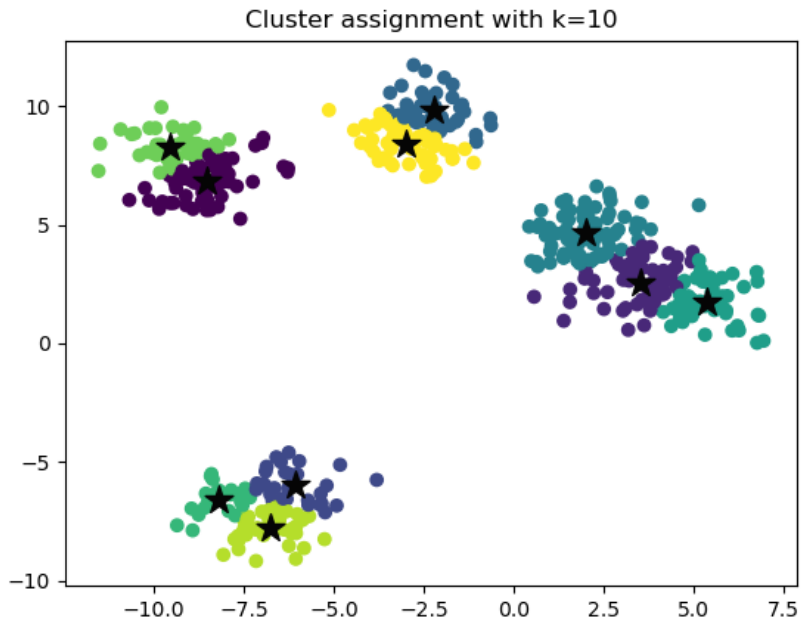


Image Segmentation

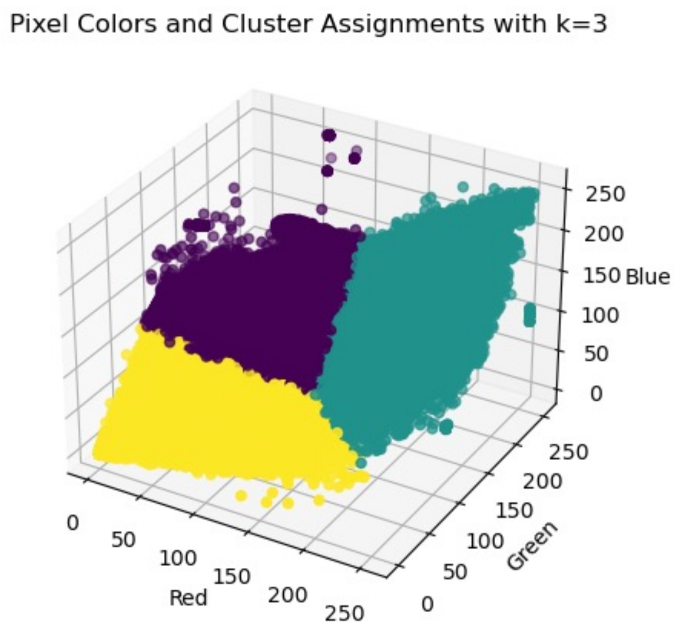
Image generated with DALL-E

Image Segmentation

Pseudo-code

1. Load the image
2. Create an array where each pixel is represented by 3 values (RGB)
3. Apply k-means on the array
4. Use the cluster assignments to “paint” the image and observe the segments

Image Segmentation Examples



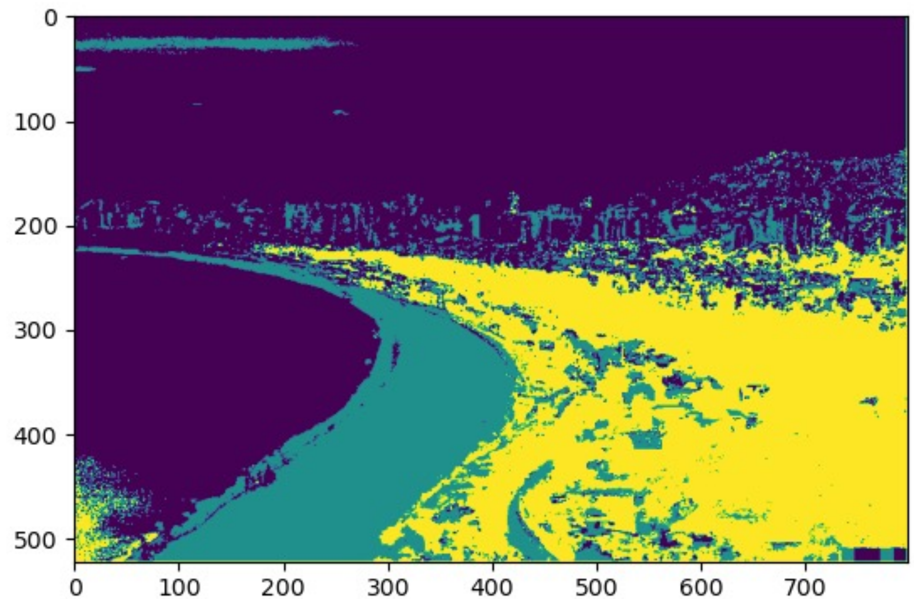
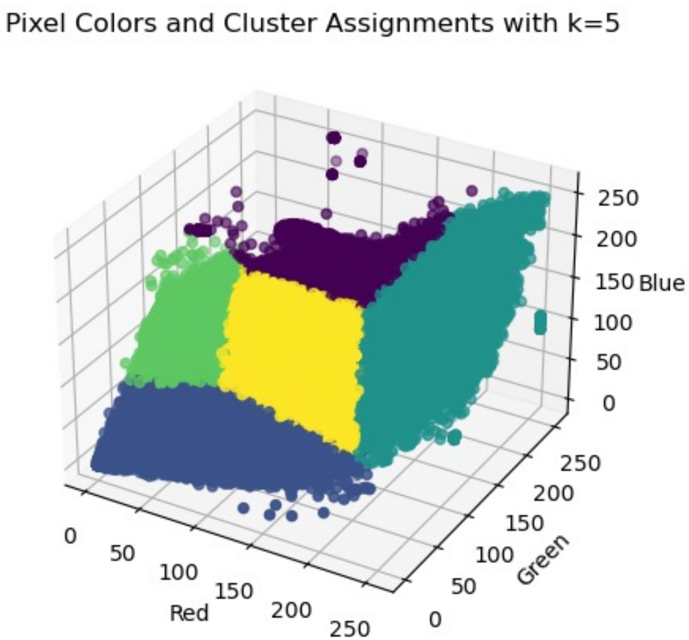


Image Segmentation Examples



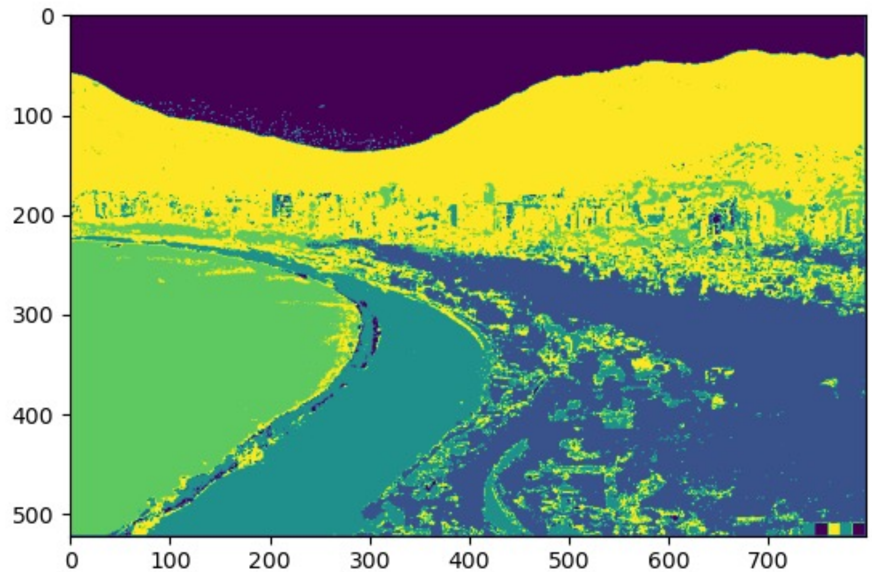
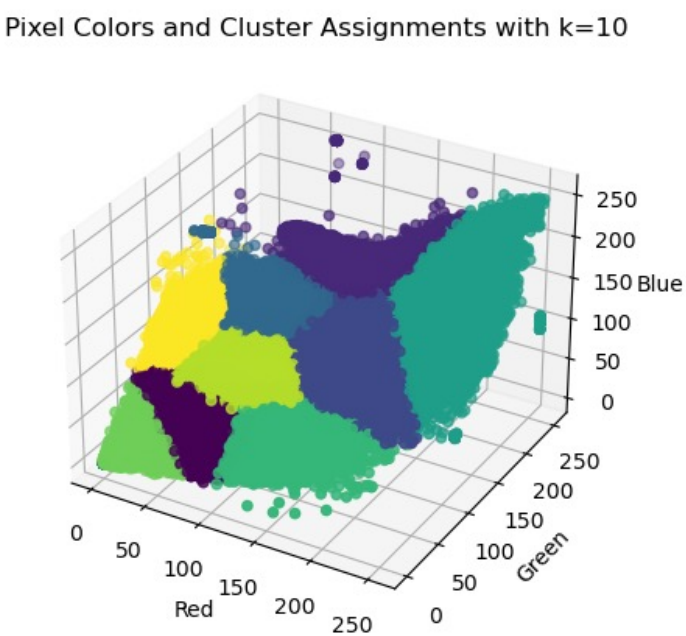
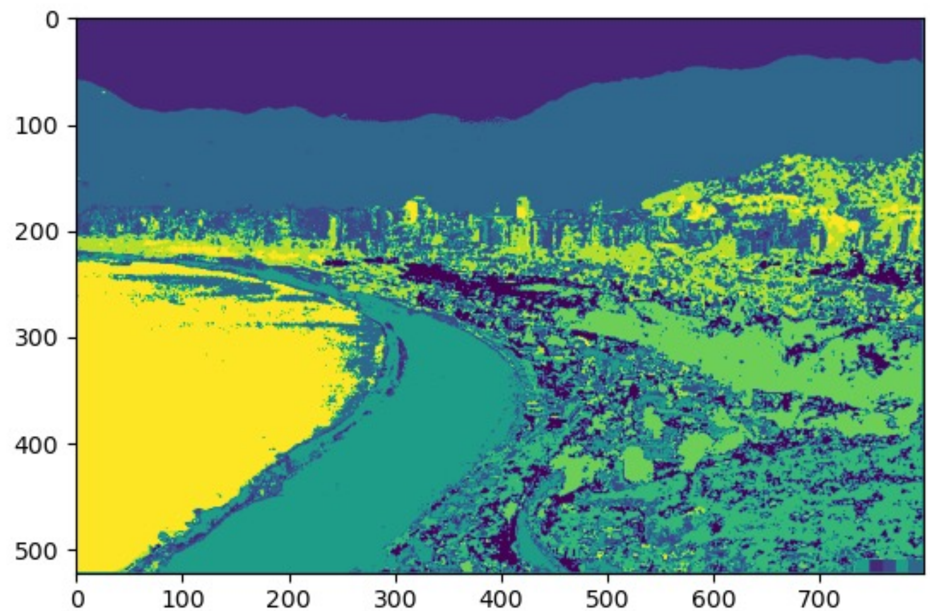


Image Segmentation Examples



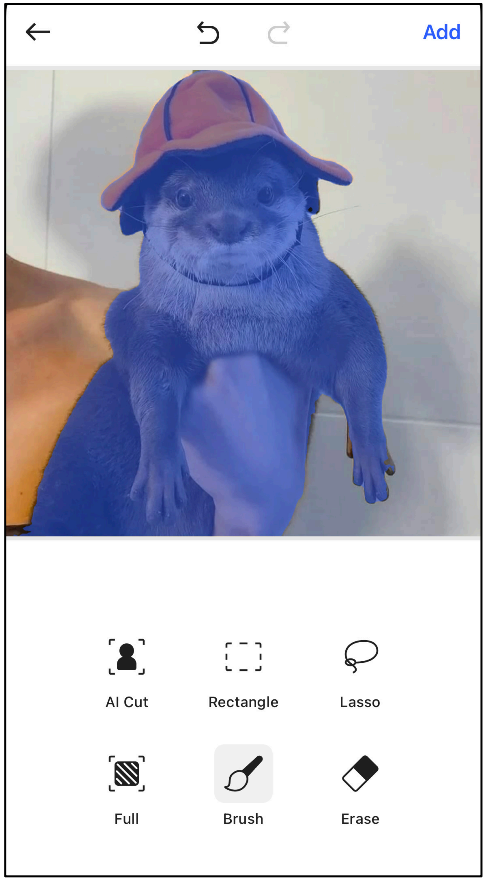


Summary

- Some more examples of **k-means** and **DBSCAN**
- Ensemble methods (diversity, combination, base learners) + measuring diversity and experimentation
- Elbow method and image segmentation example
- High dimensionality & other challenges

Coming up next…

- Search



Sticker.ly app