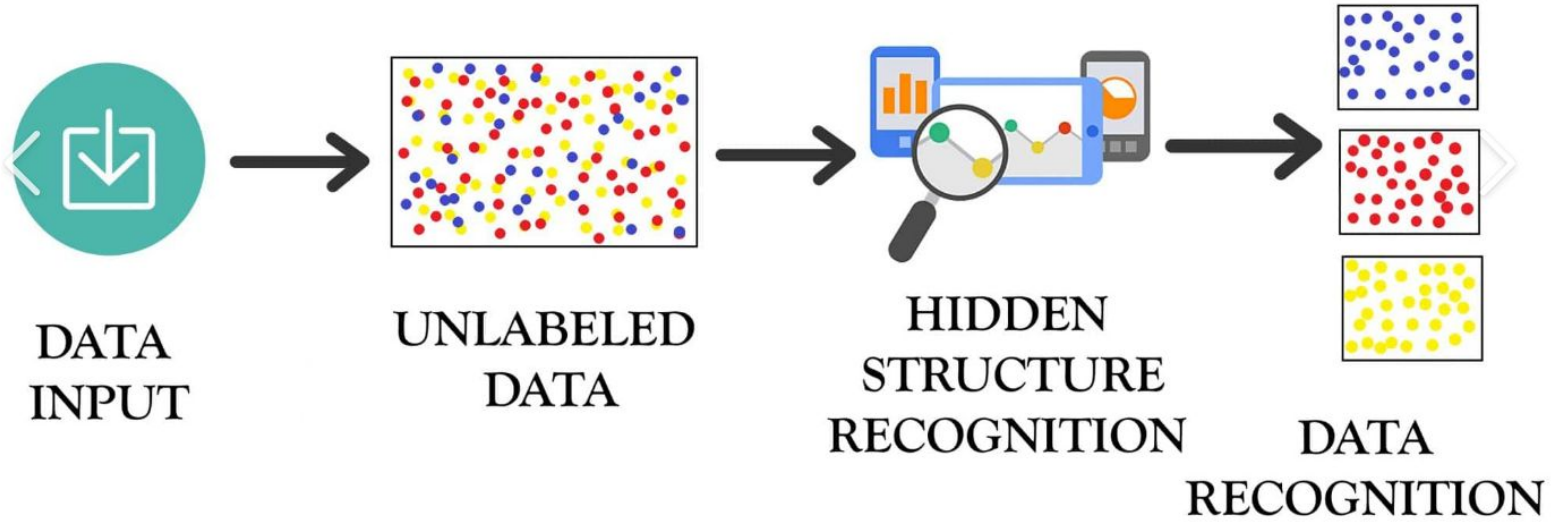
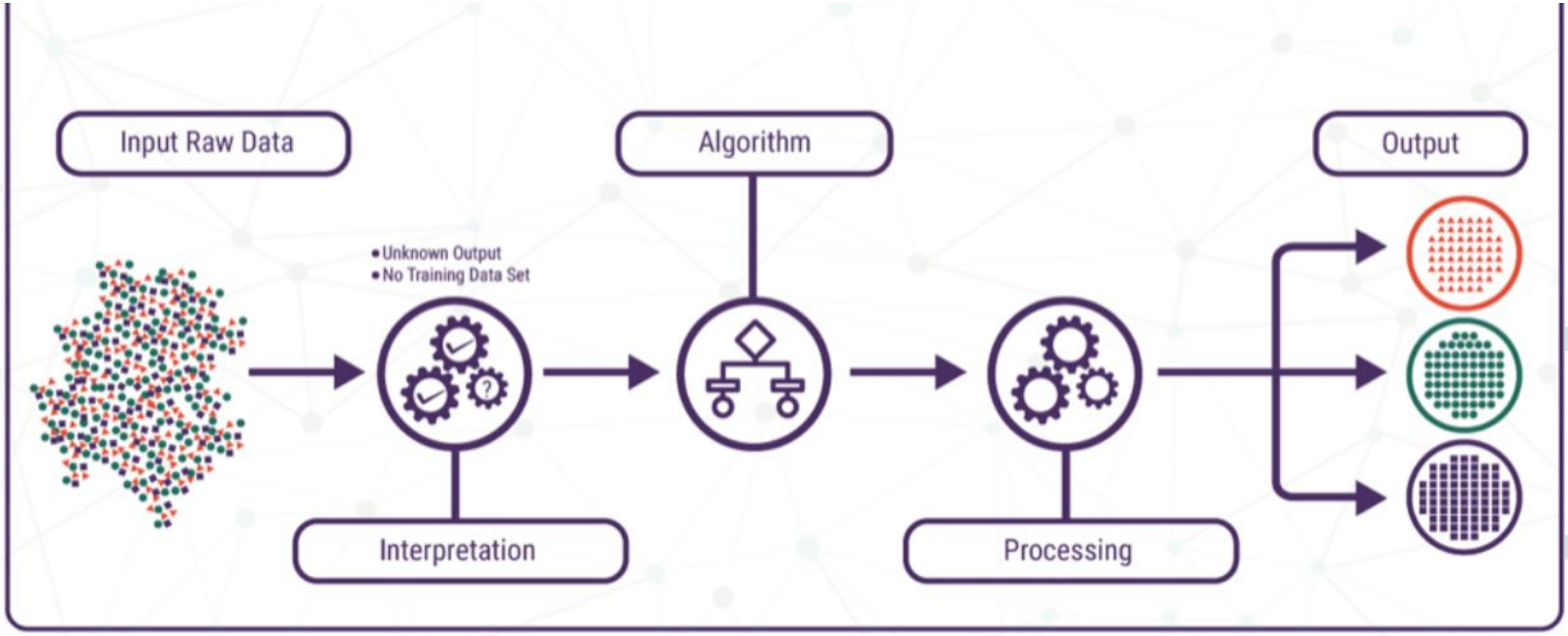
Intro to Unsupervised Learning
Principles of ML & Deep Learning - SAILea

Ananya Raghu and Anisha Raghu

Unsupervised Learning

- *Unsupervised learning*: uses unlabeled data, discovers "structure" or underlying patterns in data
- AI that uses machine learning to analyze unlabeled data and cluster them
  - Identify hidden patterns, population/sample behaviors
  - Help determine relationships
- K Means clustering is one of the simplest and most popular unsupervised learning algorithms

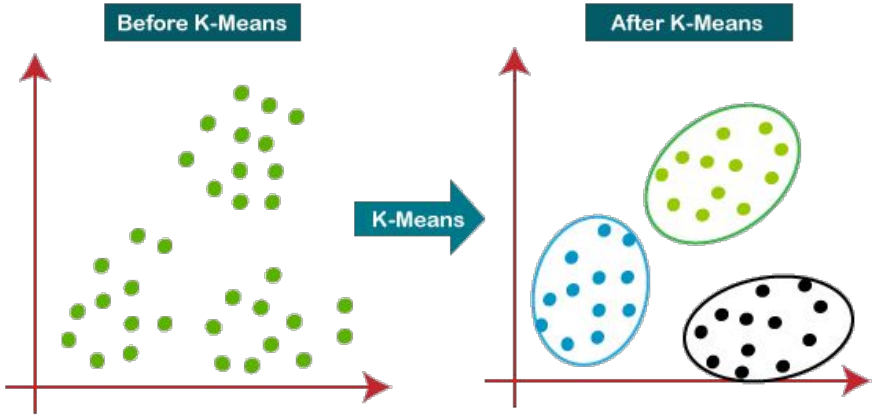

DATA
INPUT
UNLABELED
DATA
HIDDEN
STRUCTURE
RECOGNITION
DATA
RECOGNITION


Input Raw Data
Algorithm
Output
• Unknown Output
• No Training Data Set
Interpretation
Processing

K Means Clustering

- K means clustering works by clustering data points into groups

- Centroid: location of center of cluster (doesn't have to be a data point) → can *specify* number of cluster centers you want



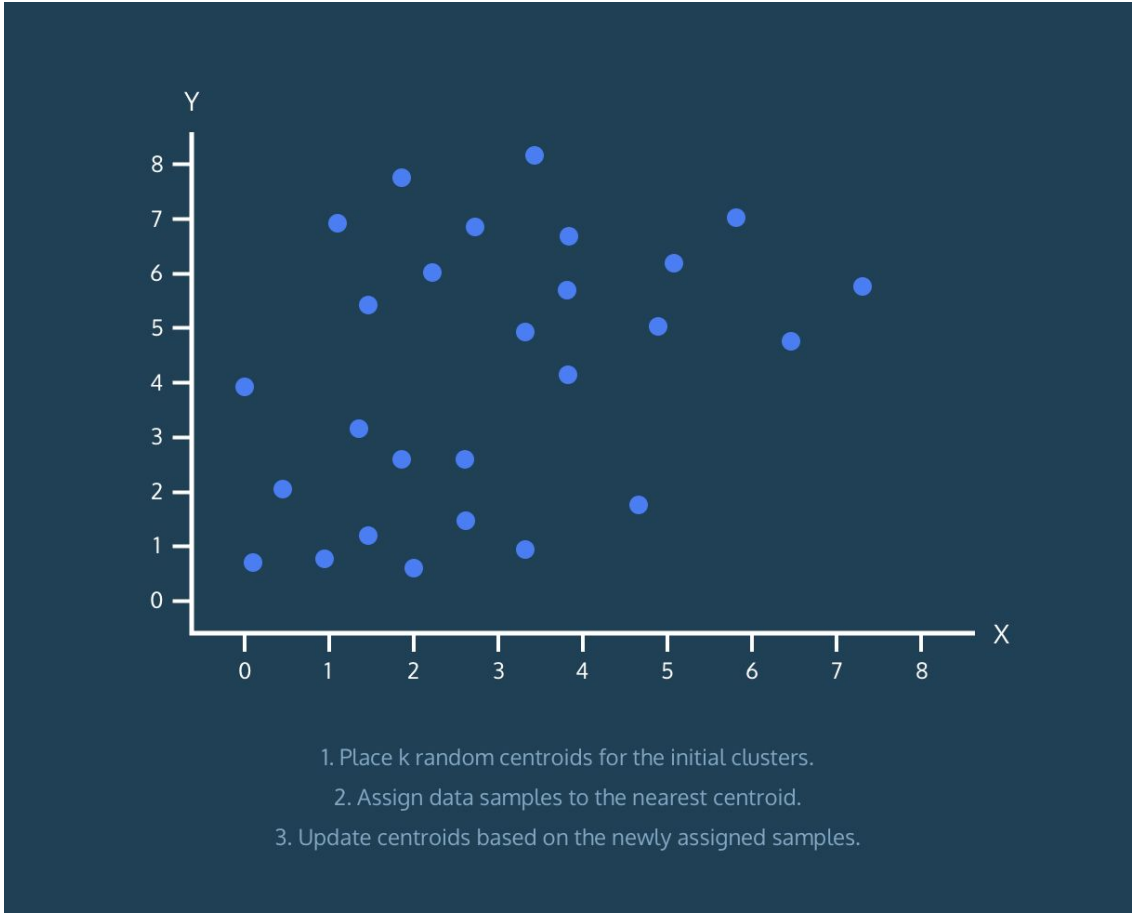
The Algorithm

Step 1: Randomly initialize k cluster centroids

Step 2: Categorize each point as part of the cluster whose centroid it is closest to.

Step 3: Update the mean coordinates after all points per cluster are assigned

Step 4: Repeat the process for a certain number of iterations → have clusters at end


Y
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
X
1. Place k random centroids for the initial clusters.
2. Assign data samples to the nearest centroid.
3. Update centroids based on the newly assigned samples.

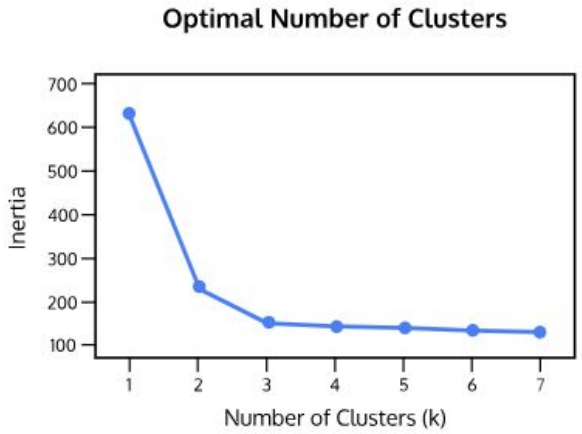
Source

Concept of Inertia

$$\sum_{i=1}^{N}(x_i - C_k)^2$$

- Inertia is a measurement of how well a dataset responds to K-means clustering
- Measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster.
- Ideally: low Inertia, and a small amount of clusters
  - The extreme case is to have zero inertia → but this means that every SINGLE point is its own cluster!
  - *Tradeoff* between minimizing inertia and having a low amount of clusters
- To find the ideal number of clusters we can use the ELBOW method

Elbow Method



Source

- To find the optimal number of clusters we can look for the elbow point in this graph to the right: where the rate of decrease in Inertia begins to slow down
- Not much of an improvement after this “elbow” point
- In the graph to the right this is K=3.