UNSUPERVISED MACHINE LEARNING

TYPES OF UNSUPERVISED LEARNING

- Unsupervised transformations of a dataset are algorithms that create a new representation of the data which might be easier for humans or other machine learning algorithms to understand compared to the original representation of the data. Example: recommender.
- Clustering algorithms, on the other hand, partition data into distinct groups of similar items. Example: uploading photos on social media, auto-tagging person.

CHALLENGES IN UNSUPERVISED LEARNING

- A major challenge in unsupervised learning is evaluating whether the algorithm learned something useful.
- Unsupervised learning algorithms are usually applied to data that does not contain any label information, so we don't know what the right output should be.
- Therefore, unsupervised algorithms are used often in an exploratory setting, when a data scientist wants to understand the data better, rather than as part of a larger automatic system.

WHAT IS PRINCIPAL COMPONENT ANALYSIS?

- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
- Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.
- So, to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

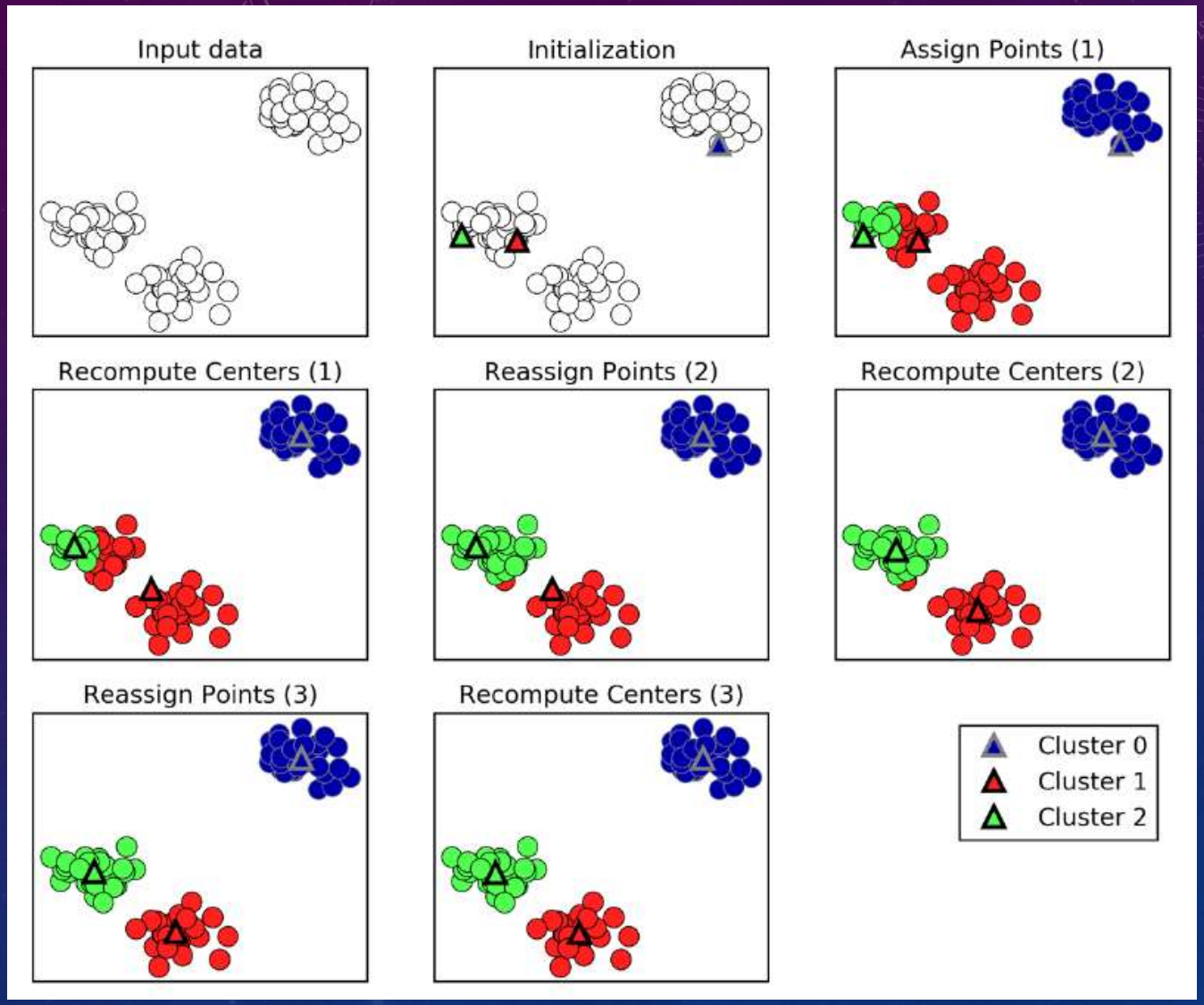
https://builtin.com/data-science/step-step-explanation-principal-component-analysis

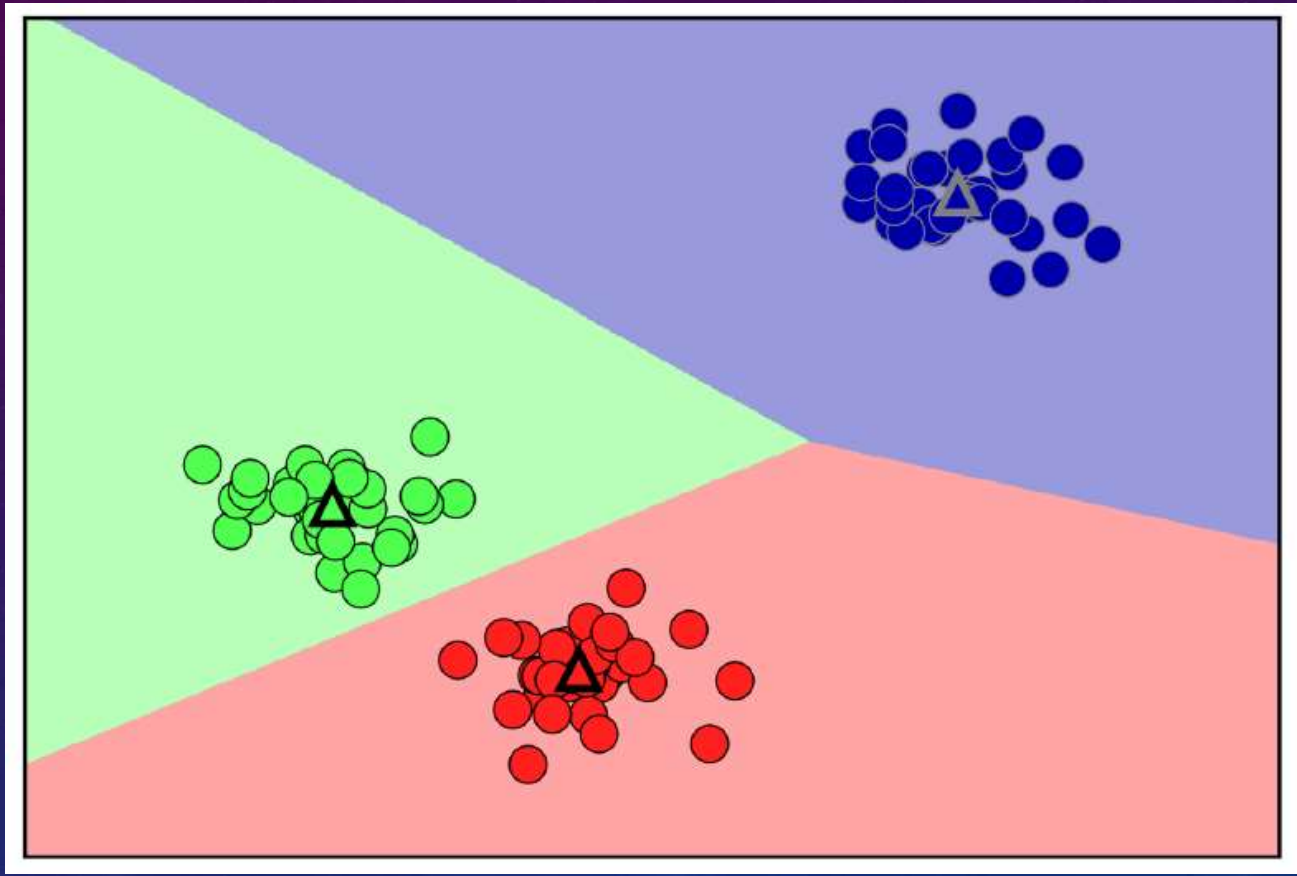
CLUSTERING

- Clustering is the task of partitioning the dataset into groups, called clusters.
- The goal is to split up the data in such a way that points within a single cluster are very similar and points in different clusters are different.
- Similarly, to classification algorithms, clustering algorithms assign (or predict) a number to each data point, indicating which cluster a particular point belongs to.

K-MEANS CLUSTERING

- k-means clustering is one of the simplest and most used clustering algorithms.
- It tries to find cluster centers that are representative of certain regions of the data.
- The algorithm alternates between two steps: assigning each data point to the closest cluster center, and then setting each cluster center as the mean of the data points that are assigned to it.
- The algorithm is finished when the assignment of instances to clusters no longer changes.
- In the next figure;
  - Cluster centers are shown as triangles, while data points are shown as circles.
  - Colors indicate cluster membership.


Input data
Initialization
Assign Points (1)
Recompute Centers (1)
Reassign Points (2)
Recompute Centers (2)
Reassign Points (3)
Recompute Centers (3)
Cluster 0
Cluster 1
Cluster 2



Cluster centers and cluster boundaries found by the k-means algorithm

K-MEANS CLUSTERING

ACTIVITY

THE END