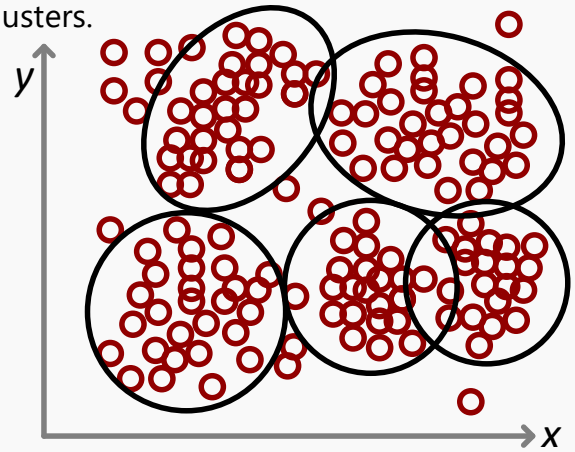
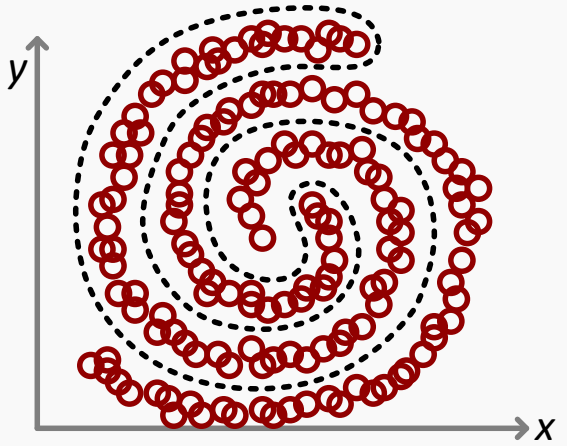
module6 · clustering and recommenders

unsupervised learning clustering

Unsupervised Learning implies the training data has no actual labels to learn from.
The goal in **Clustering** is to group the datapoint with unknown labels into something similar.

In the illustration to the right, the data exhibits five physical clusters. The algorithm applied needs to automatically be able to identify these clusters. This can become very difficult when the clusters are less obvious like in the alternate example. There are two clusters present in the alternate example and would be harder to identify with an **unsupervised clustering learning algorithm**. There is an algorithm that will handle each problem well, but each algorithm cannot handle the example it is not designed to be applied to.





Clustering Applications

- Automatically group documents/webpages into topics
  - Grouping daily news stories into categories
- Clustering large numbers of products
  - Etsy products or Ebay listings
- Clustering customers into those with similar behavior

K-means Clustering

One of the most widely used clustering methods in practice.

- Input number of clusters k, randomly initialize centers
- Assign all points to the nearest cluster k_i center
- Change cluster k_i's centers to centralize datapoints
- Repeat until convergence