Clustering
What is clustering?

- A form of unsupervised learning (no labels) used to separate a large group of objects into smaller subgroups of similar objects
- Some examples of clustering:
  - Topic modeling: clustering web pages by subject matter (politics, sports, education, etc.)
  - Identifying hot spots of police violence in urban areas
  - Image segmentation, gene expression, etc.
- Can be seen as a part of exploratory data analysis: clustering helps identify outliers, for example
**k-means clustering**

- *k*-means clustering is the most common type of clustering
- The *k* in *k*-means refers to the number of clusters
- **Pros:**
  - It is easy to understand and to implement, so it can produce quick and dirty results
- **Cons:**
  - You must specify *k* in advance; if *k* is too large, it will find clusters where there are none; if *k* is too small, it will miss “real” clusters
Remember this?

A middle school dance

$k$-means clustering requires that you specify how many clusters you want to group the dance attendees into.

Let’s pick three clusters, so $k = 3$. 

Image and example from John W. Foreman’s *Data Smart* book
A Middle School Dance

- *k*-means clustering starts with three initial points (cluster centers), one per cluster, spread out across the dance floor
- Dancers are assigned to the cluster that’s nearest to them
- The algorithm then slides the cluster centers and their corresponding clusters around until it finds a good fit
The algorithm is initialized with three centers (three black circles).

Each data point is then assigned to the nearest center.

In effect, this operation divides the space into three clusters, which are depicted here as variously shaded regions.
Cluster centers are then recomputed.

Observe how they move towards the data points.

And the process continues (until some stopping criterion is met).
Final form!

Note the locations of the cluster centers and the divisions between clusters.
What do clusters mean?

- It is never a good idea to take an algorithm’s word for it. We must always apply human insight to interpret an algorithm’s output.
- In the case of clustering, we ask what the clusters might signify?
- For a middle school dance, they could be cliques. The kids might be too timid to dance with kids outside their comfort zone!
- $k$-means allows us to cluster data, but we cannot accept a clustering if we cannot attribute meaning to the clusters. We must be able to understand the $why$ behind the assignment.
How does it work?

Step 1: Choose a desired number of clusters, $k$

Step 2: Randomly assign each data point to an initial cluster

Step 3: Compute cluster centroids (this is the $k$-centroids algorithm right?)

Step 4: Re-assign each point to the closest cluster centroid

Step 5: Re-compute cluster centroids

Step 6: Repeat steps 4 and 5 until a stopping criterion is met
We’re still missing something key!

- Goal: to group “similar” objects into meaningful subgroups, called clusters
- Objects are similar if the distance between them is small
- This only begs the question: how do we define the distance between two objects?
What is a good clustering?

- The intra-cluster similarity is high
- The inter-cluster similarity is low

The quality of a clustering depends on the choice of similarity metric: i.e., the definition of distance.