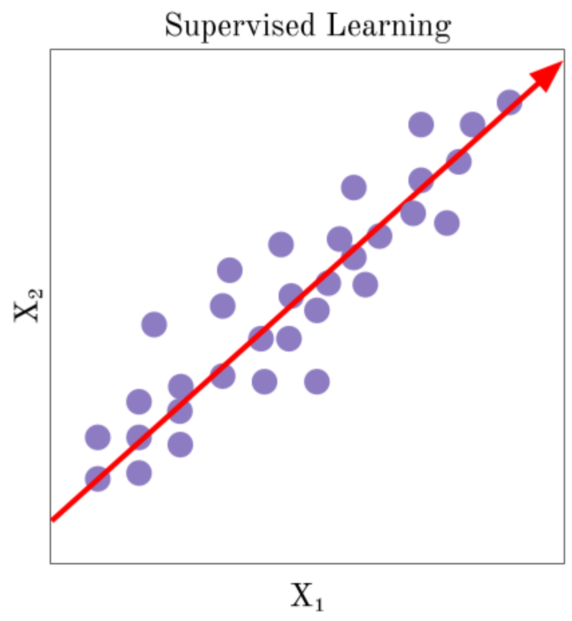
Introduction to Unsupervised Clustering

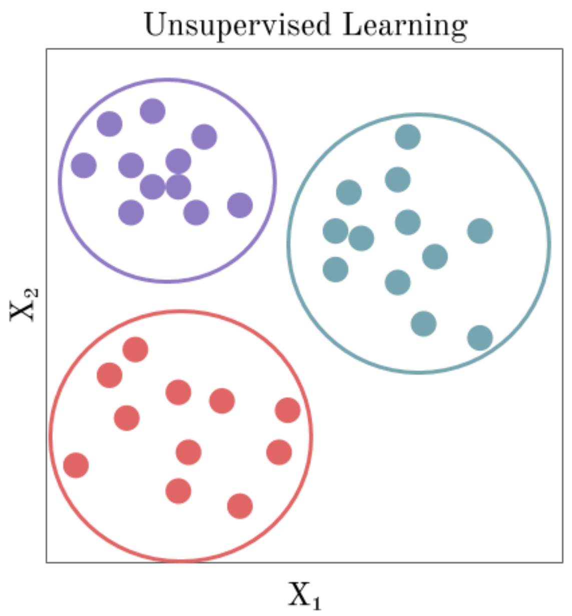
k-means, DBSCAN, and spectral

Alyssa W. Zhang
DRP Symposium
Spring 2024

Unsupervised Learning

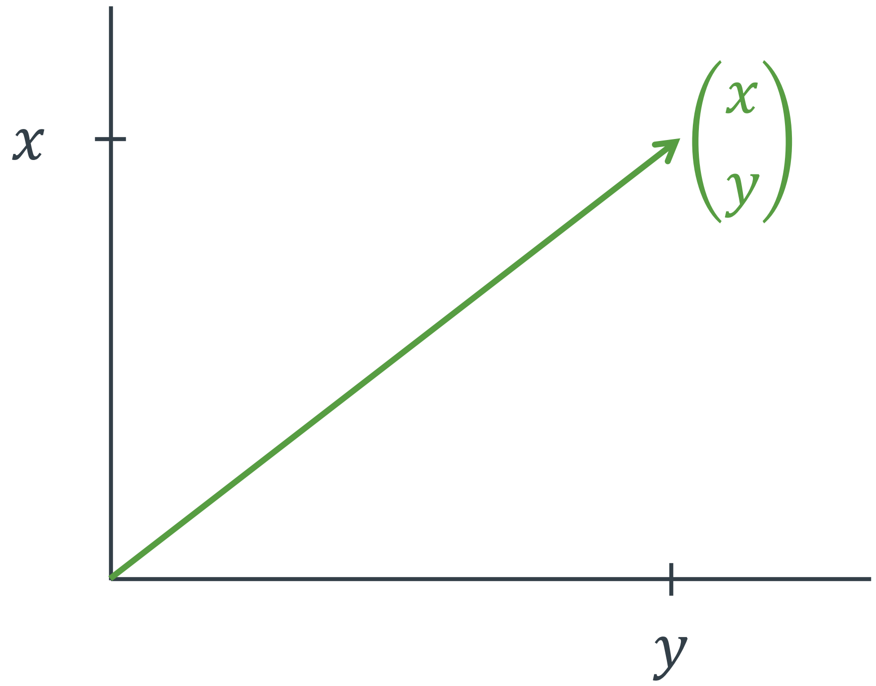
- *Supervised learning* refers to learning from labeled data.
    - The objective is prediction or inference.
    - Regression and classification are forms of supervised learning.
- *Unsupervised learning* refers to learning from unlabeled data.
    - The objective is to “create” labels by grouping data based on similarity.
    - Clustering is a form of unsupervised learning.





Data as High-Dimensional Vectors

- The typical idea of a data point is (x, y), which can be represented as a 2-dimensional vector with elements x and y.



- What if we have more than 2 dimensions?

Data as High-Dimensional Vectors (cont.)

- Suppose we have a large dataset on UT Austin students with names, majors, classifications, emails, and 100 other features.

Name	Major	Classification	Email	…
Zhang, Alyssa	Mathematics	Senior	***@utexas.edu	…
Duncan, Addie	Mathematics	Graduate	***@gmail.com	…
…	…	…	…	…

- We can represent each data point as a *high-dimensional vector*, where each vector element represents a feature.

$$\begin{pmatrix} Name \\ Major \\ Classification \\ Email \\ \dots \end{pmatrix}$$

Measures of Similarity

- *Distance metrics* are used to measure similarity between high-dimensional vectors (data points).

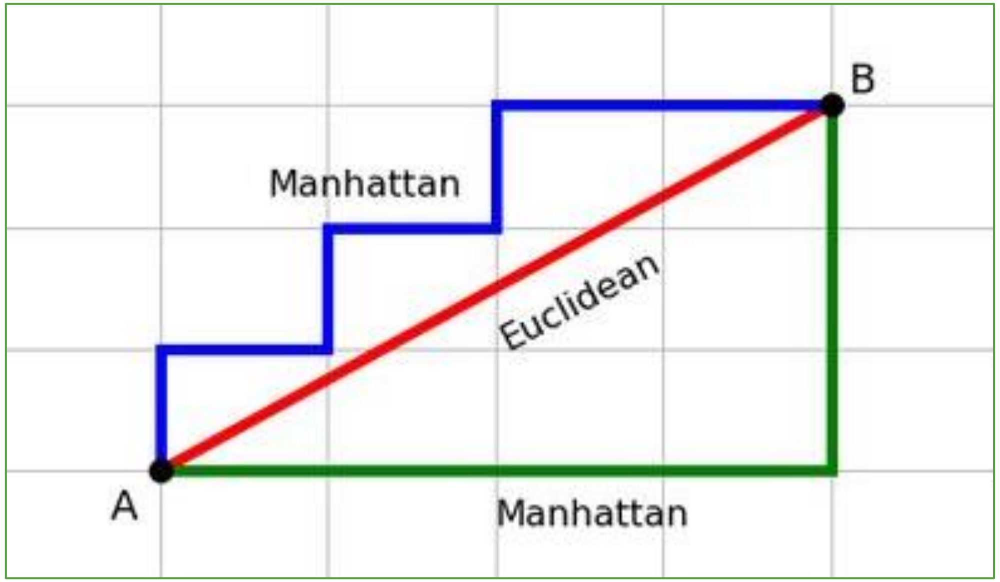
- Because there are various ways of measuring distance, there are various methods of clustering—each with their own benefits and drawbacks.

Euclidean Distance

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

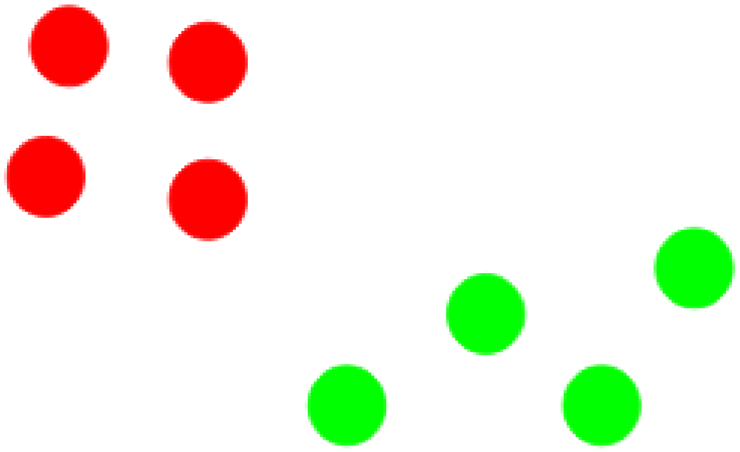
Manhattan Distance

$$d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} |q_i - p_i|$$



k-means Clustering Algorithm

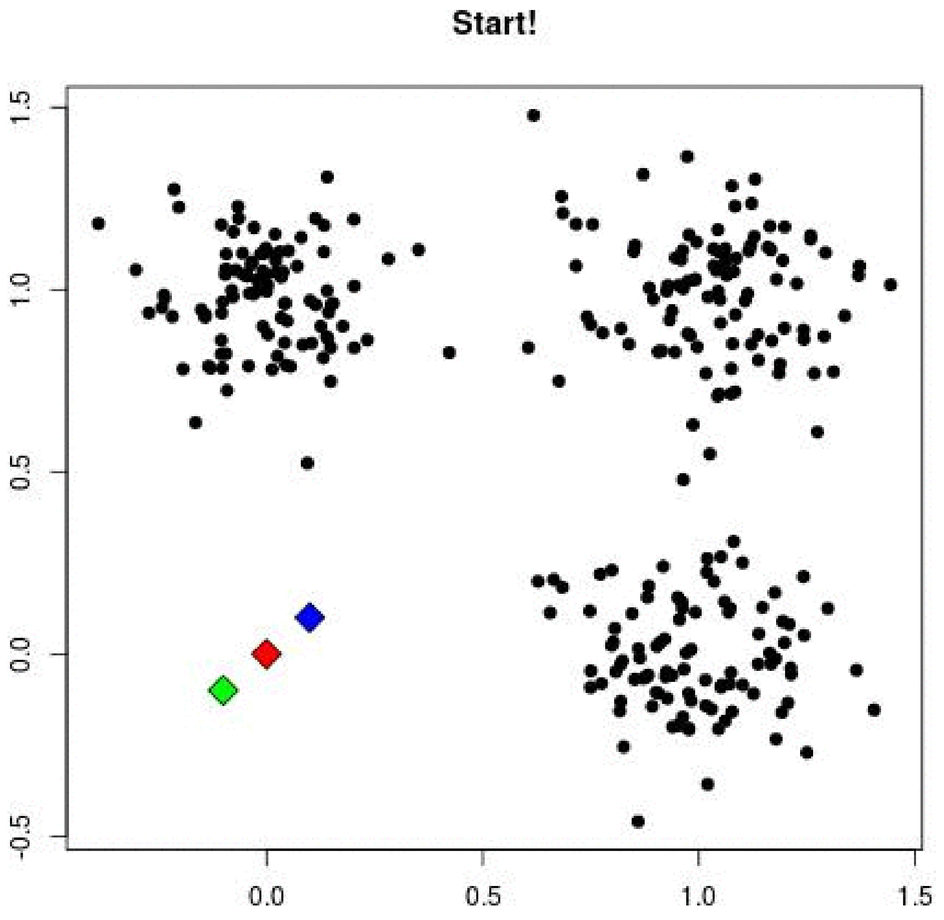
1. Choose the number of clusters (k).
2. Select random centroids.
3. Assign points to the closest centroid.
4. Recompute centroids of newly formed clusters.
5. Repeats steps 3 and 4 until stopping criterion is met.
   - Stopping criterion: When centroids do not change, points remain in the same cluster, or we reach a predetermined maximum number of iterations.



k-means Clustering Pros and Cons

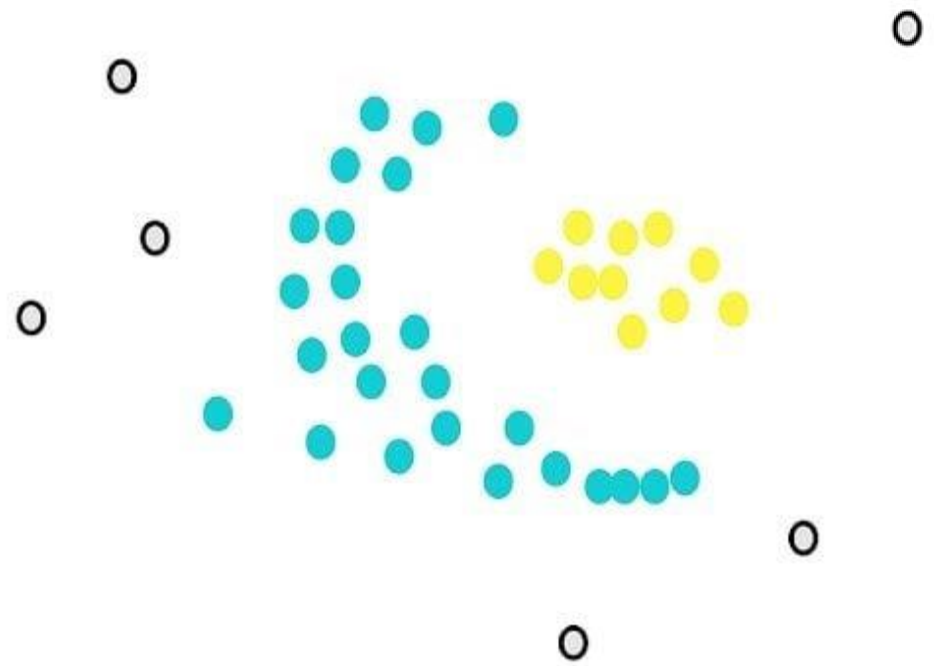
- Pros
  - Relatively simple to implement.
  - Good for linearly separable, convex data.

- Cons
  - k value must be chosen manually, which can lead to over/under clustering.
  - Sensitive to clustering data of varying sizes and density.
  - Sensitive to outliers.



DBSCAN Clustering Algorithm

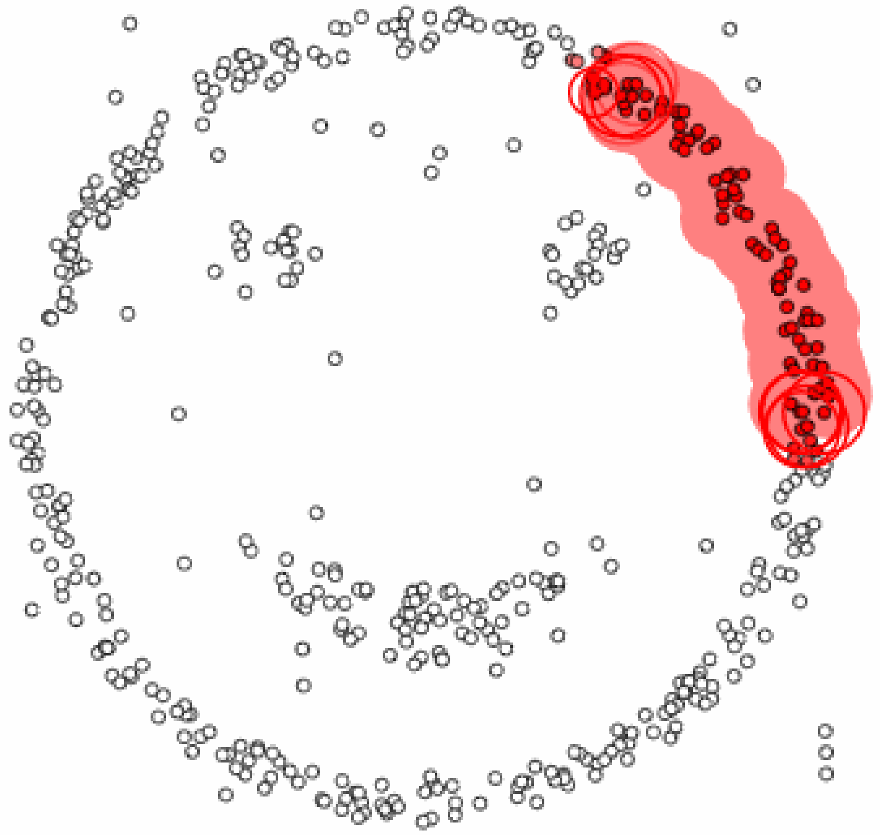
1. Count the number of points close to each point within a radius ε.
2. Identify core points (points with at least m close neighbors).
3. Start with a core point and assign it to a cluster.
4. Expand the cluster by adding neighboring core points and their neighbors.
5. Repeat until no more core points can be added to the cluster.
6. Assign remaining core points to new clusters or noise.



DBSCAN Pros and Cons

- Pros
  - Does not require specifying the number of clusters.
  - Capable of identifying clusters of any shape (incl. non-convex).
  - Effectively handles noise and outliers.

- Cons
  - Affected by choice of m and ε.
  - Limited by the curse of dimensionality.



epsilon = 1.00
minPoints = 4

Concepts for Spectral Clustering

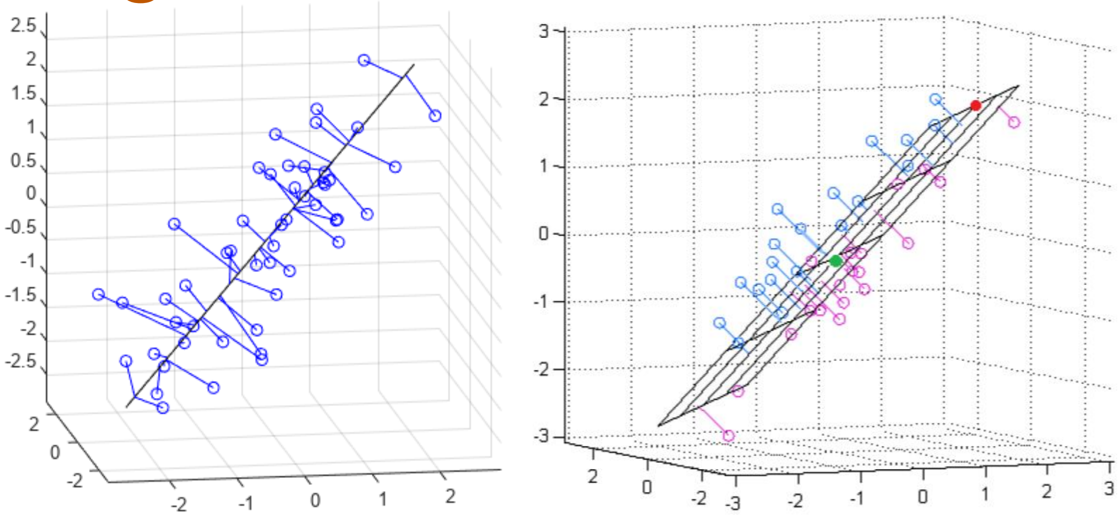
- *Curse of dimensionality* refers to various phenomena that occur with data in high-dimensional spaces that does not occur in low-dimensional spaces.
  - For sufficiently large dimensions, the distances between all pairs of data points will be essentially the same.

- When distances between all pairs of data points are the same, DBSCAN may merge all points into the same cluster.
  - Situation would benefit from projecting data to a lower-dimensional subspace.

Concepts for Spectral Clustering

- *Singular Value Decomposition (SVD)* is a way to find the best-fitting k-dimensional subspace for a data matrix A.
  - Like a k-dimensional line of best fit

- SVD says that any A can decompose into three matrices with special characteristics.
  - The columns of V are right singular vectors, which represent the directions A must transform to best fit its subspace.



Singular Value Decomposition (SVD)

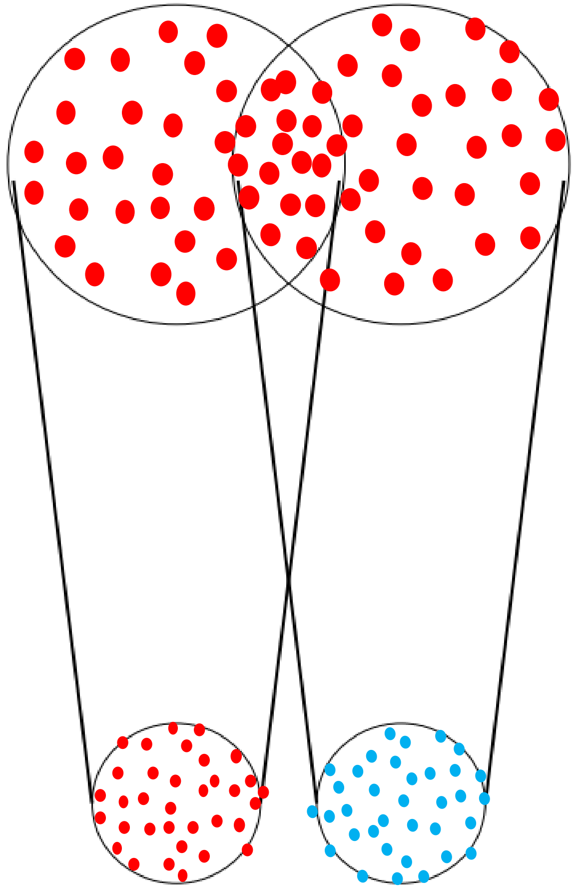
$$A = UDV^T$$

U contains left singular vectors.
D is diagonal and contains singular values.
V contains right singular vectors.

Spectral Clustering Algorithm

1. Find space V spanned by the top k right singular vectors of A.
2. Project data points to V (lower dimensional subspace of A).
3. Cluster the projected points (various methods).

- Spectral clustering is often a pre-processing step that refines other methods of clustering.
  - Ex. k-means



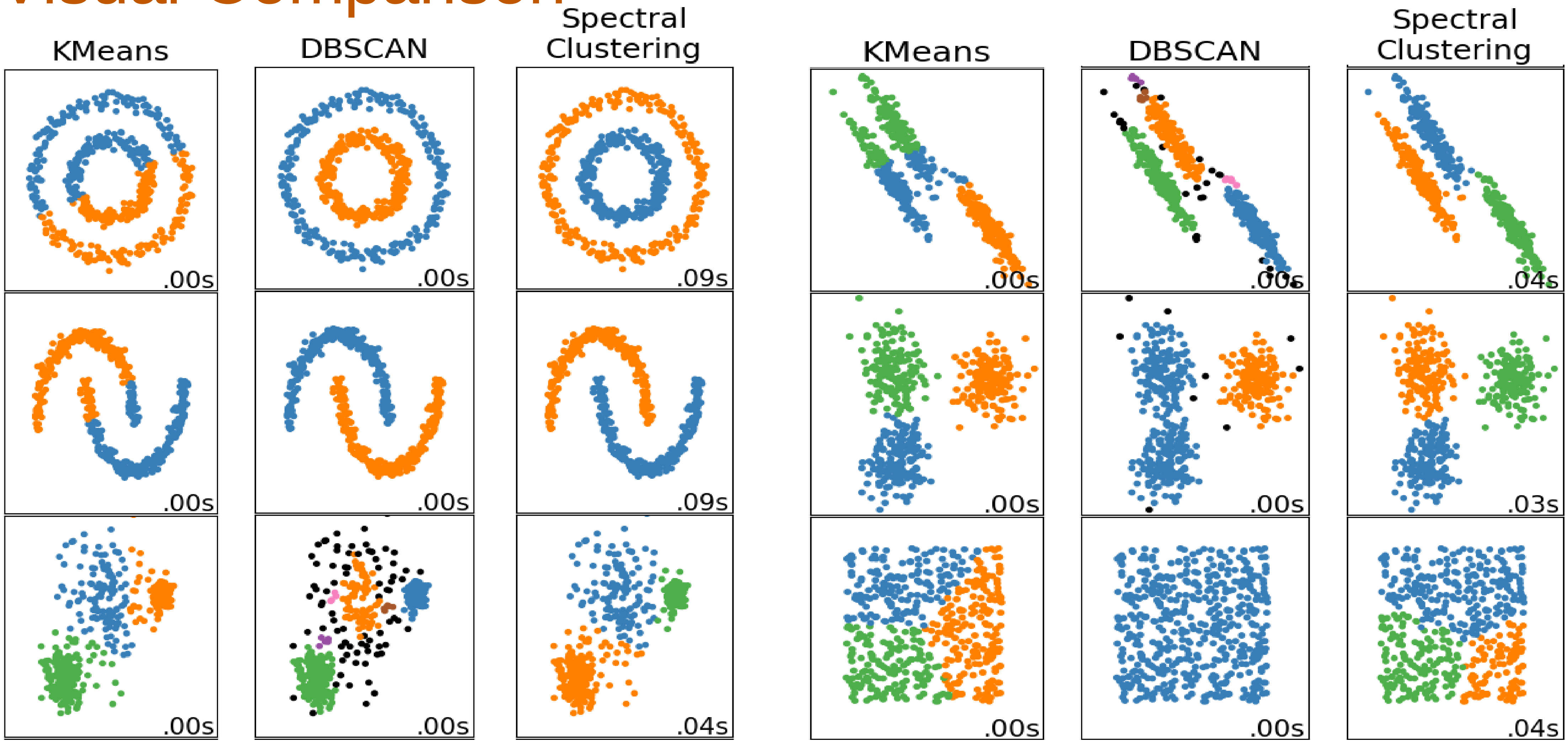
Spectral Clustering Pros and Cons

- Pros
  - No assumptions about cluster shape; good for irregularly shaped data.
  - Introduces linear separability.

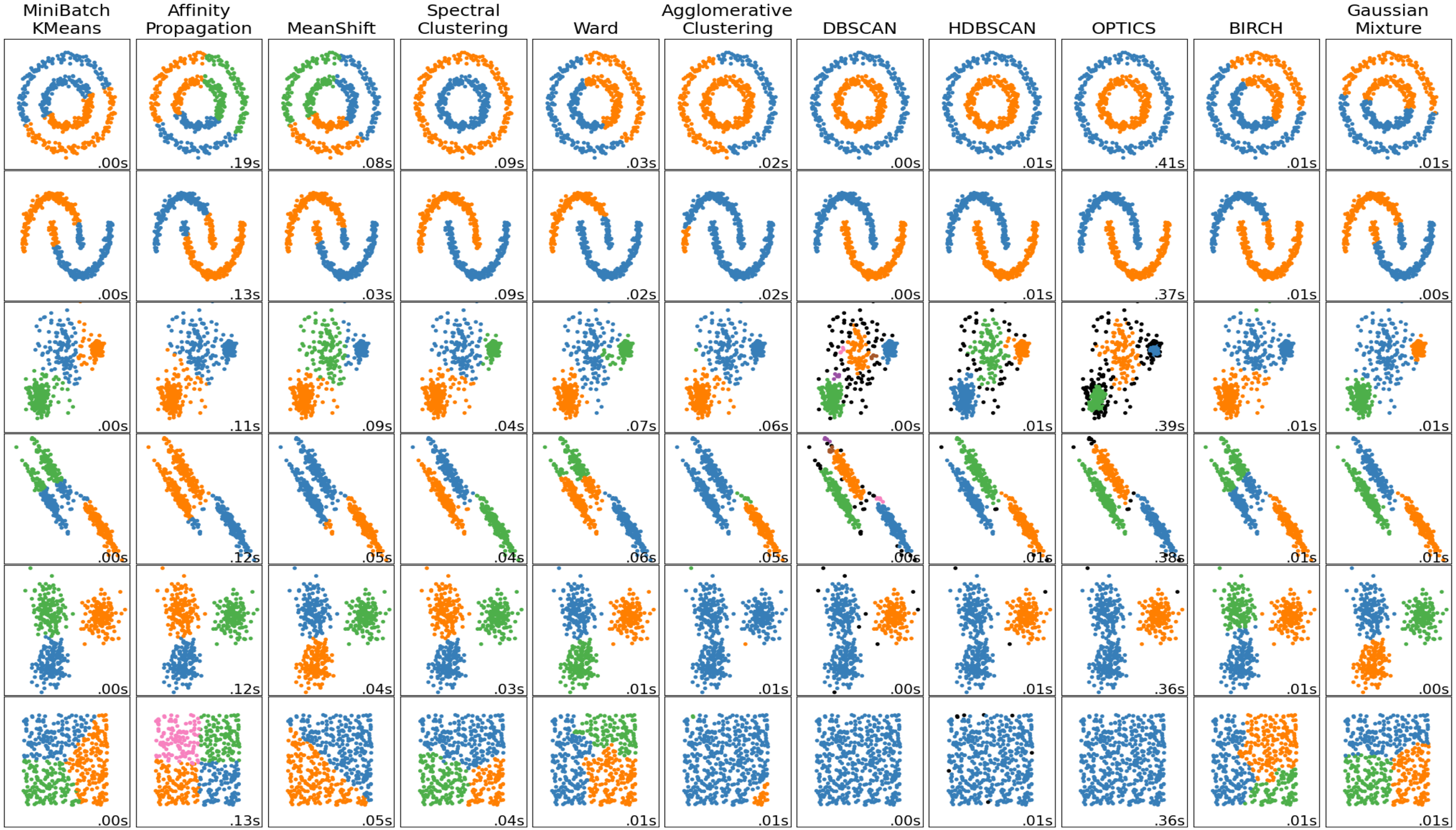
- Cons
  - Sensitive to the choice of rank (k).
  - Project may result in loss of interpretability.

$$\begin{pmatrix} Name \\ Major \\ Classification \\ Email \\ \dots \end{pmatrix} \xrightarrow{project} \begin{pmatrix} ?\,Feature\ 1 \\ ?\,Feature\ 2 \\ ?\,Feature\ 3 \end{pmatrix}$$

Visual Comparison



Visual Comparison (cont.)



Summary

- Clustering is a form of unsupervised learning.
- There are many different methods of clustering, including k-means, DBSCAN, and spectral.
- Deciding on which clustering method to use depends on data.
  - The images throughout this presentation are 2D and 3D, but in reality…