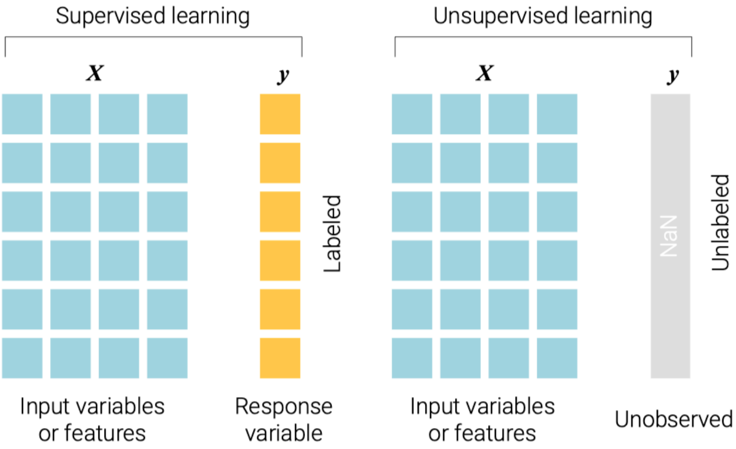
Lecture 9

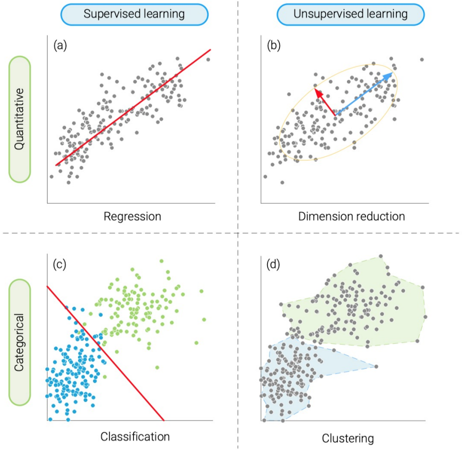
Unsupervised Learning

Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analytics and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

Supervised vs Unsupervised Learning





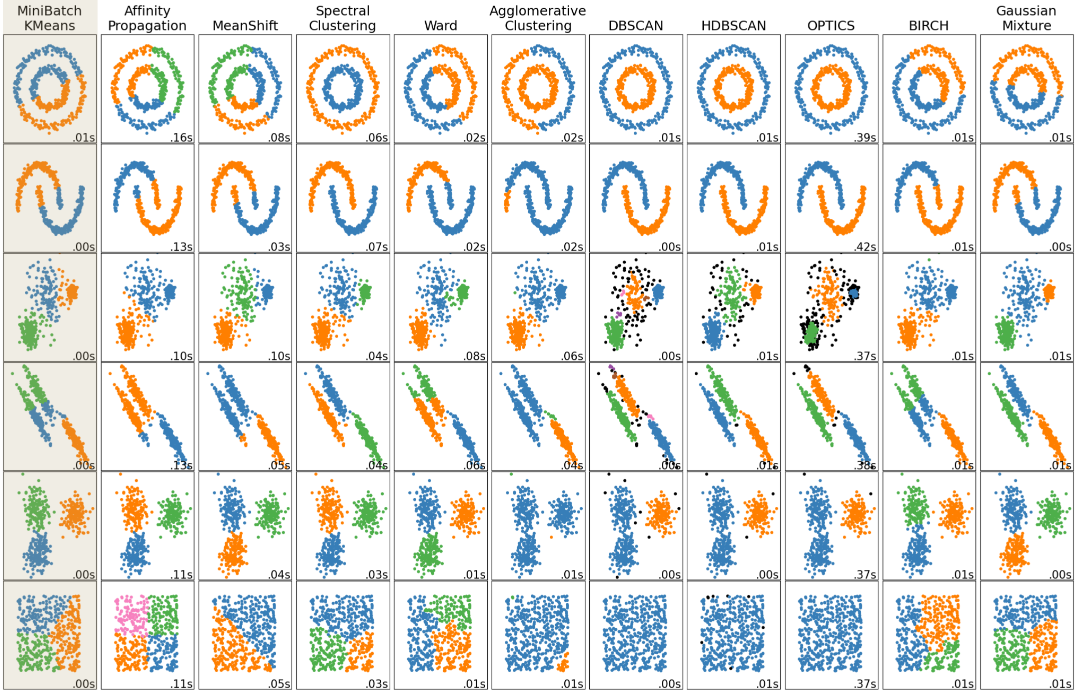
Source:
Visualizations for
Machine Learning
(Iris Series)

Supervised learning: Learning a function that maps inputs to outputs using labeled examples (Bishop, 2006).
Unsupervised learning: Learning hidden structure from unlabeled data (Hastie, Tibshirani & Friedman, 2009).

Clustering

Clustering is one of the most common used tools to recognize the unknown *class* based on some known features.

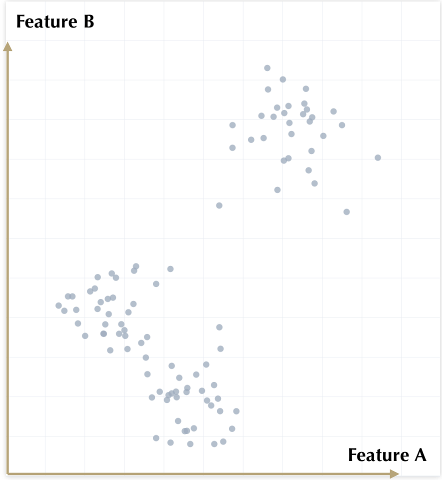
There are many methods for different patterns, and we will introduce k-means, which is the most classical one (often see as the baseline).



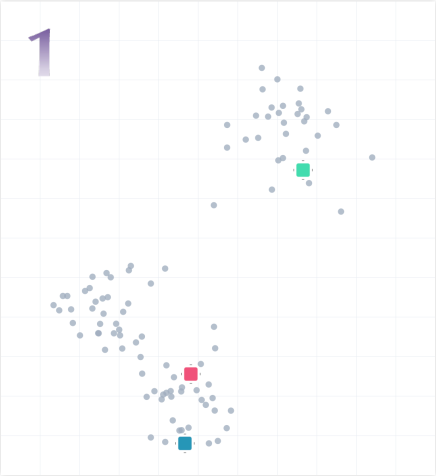
A comparison of the clustering algorithms in scikit-learn. Source: https://scikit-learn.org/stable/modules/clustering.html

k-Means Clustering

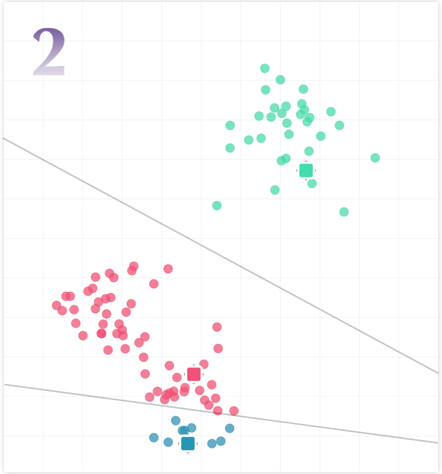
The k-means Process



Users decide the number of clusters (k, hyperparameter).



Randomly guess k cluster center locations (the initial centers matter).

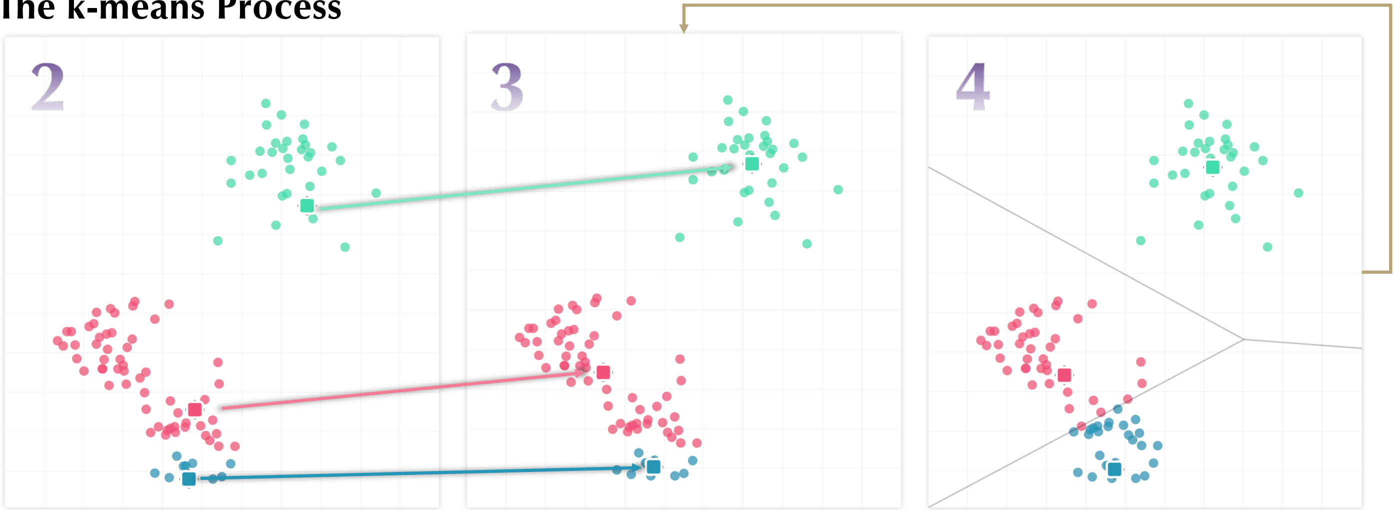


Each datapoint finds out which center it's closest to.

Source: Gemini https://gemini.google.com/share/9a5e4746162b

k-Means Clustering

The k-means Process



Repeat until the centroids stop moving.

Each datapoint finds out which center it's closest to.

Each center finds the centroid of the points it owns and moves there.

Each datapoint finds out which center it's closest to.

Source: Gemini https://gemini.google.com/share/9a5e4746162b

k-Means Clustering

How to Evaluate the Clustering Results

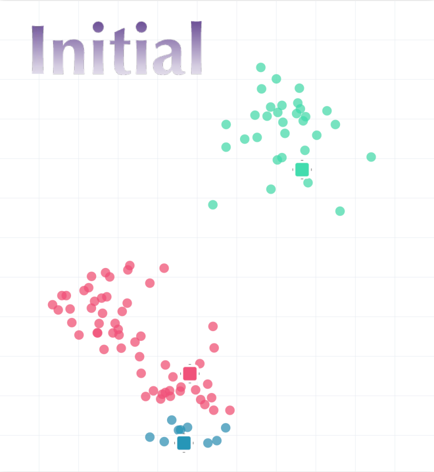
Most common measure is **sum of square error** (AKA **WSS**, within-cluster sum of squares): for each point, the error is the distance to the nearest center.

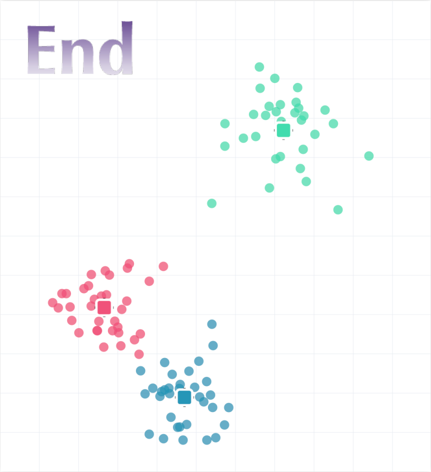
Center of cluster C_i

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \text{distance}^2(m_i, x)$$

Each data point

We prefer the clustering with the smallest error (SSE).



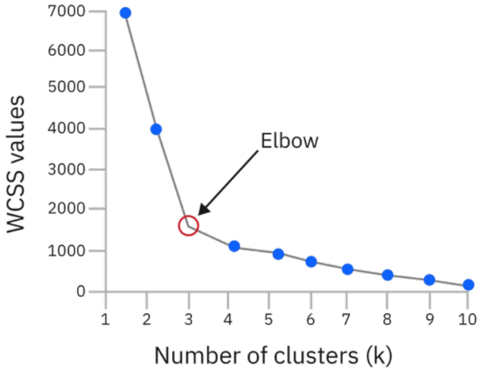


Source: Gemini https://gemini.google.com/share/9a5e4746162b

k-Means Clustering

Some Problems of k-means

- **Normalization**: we are measuring the distances between points. So, normalization is required before training.
- **How to decide on k, a hyperparameter?** Not using cross-validation. But run different k and check the diagrams using Elbow method (there are more methods).
- **How to decide on the initial centers?** Multiple runs **or** K-means++ approach (optional: Computing initial centroids in k-means).
- **K-means has problems when** clusters are of differing sizes, densities or non-globular shapes. Find other clustering approaches. Domain knowledge matters!
- **K-means has problems when** the data contains outliers or redundant features. Remove them. Domain knowledge matters!



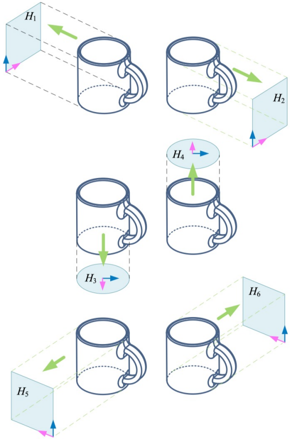
Elbow Method. Source: IBM
https://www.ibm.com/think/topics/k-means-clustering

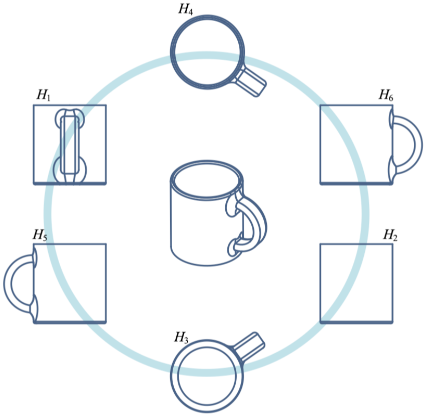
Other Unsupervised Learning Methods

For dimensionality reduction:

Principal Component Analysis (PCA)

Finds the directions of maximum variance in the data and projects the data onto those directions to reduce dimensionality.





Source: Visualizations for Machine Learning (Iris Series)

Thank you!

Haoyu Yue / yohaoyu@washington.edu
Ph.D. Student, Interdisciplinary Urban Design and Planning
University of Washington

RE 519 Real Estate Data Analysis and Visualization
Course Website: www.yuehaoyu.com/data-analytics-visualization/
Autumn 2025

The course was developed based on previous instructors: Christian Phillips, Siman Ning, Feiyang Sun
Cover page credits: Visax