# K-MEANS++ OPTIMAL INITIALIZATION ALGORITHM

An Improved K-means Clustering Method

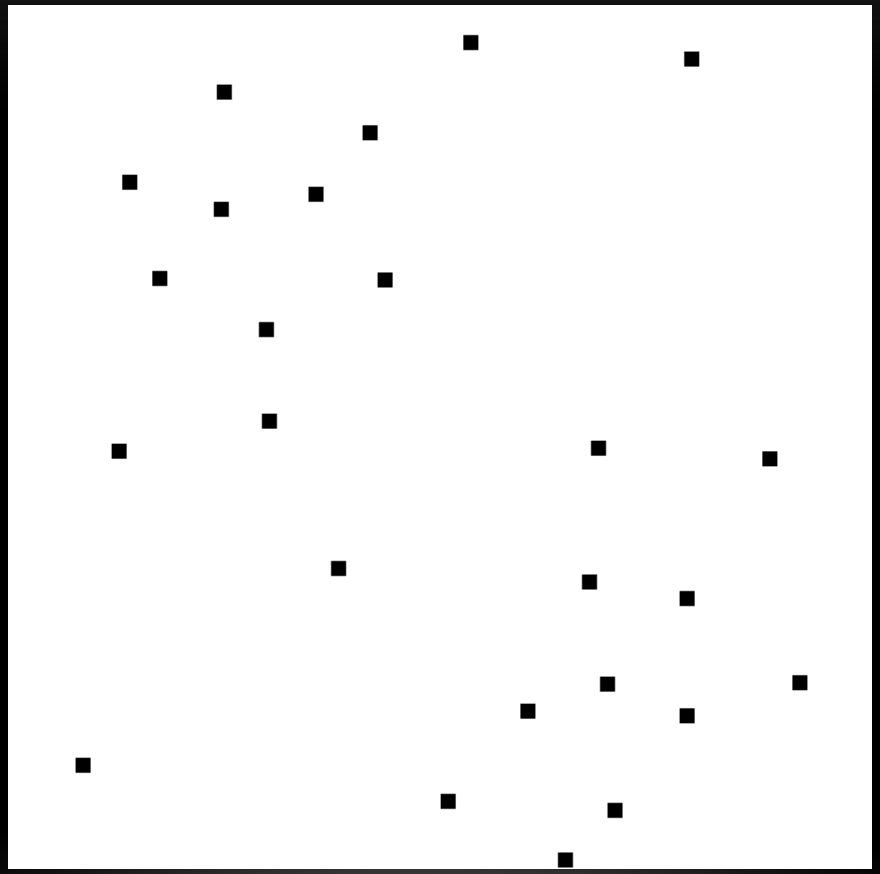
# OVERVIEW

- K-means Clustering Algorithm
- K-means++ Initialization Algorithm
- Experiment
- Datasets
- Conclusion

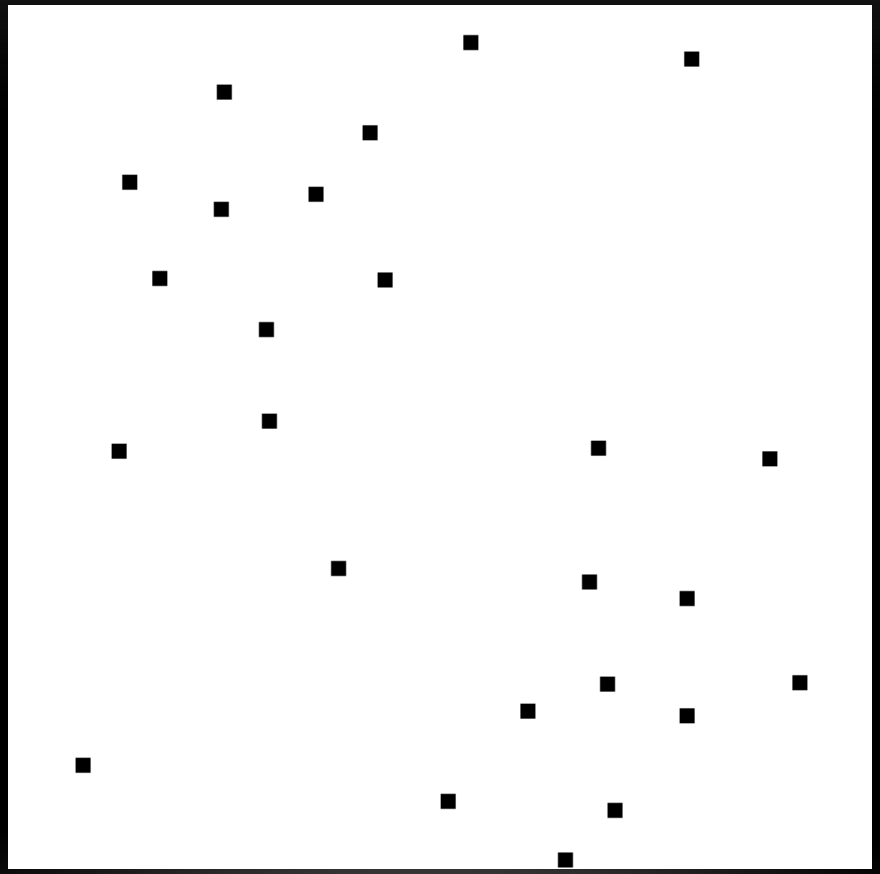
# K-MEANS CLUSTERING ALGORITHM

- A well-known naïve clustering method.
- Designed to find natural clusters in unclassified datasets.
- Only requires a single input parameter - K
- Uses random initialization technique for centroids.
- Uses Euclidean distance to determine instances' cluster assignments.
- Calculates means of finished clusters then starts over.

# CLUSTERING EXAMPLE

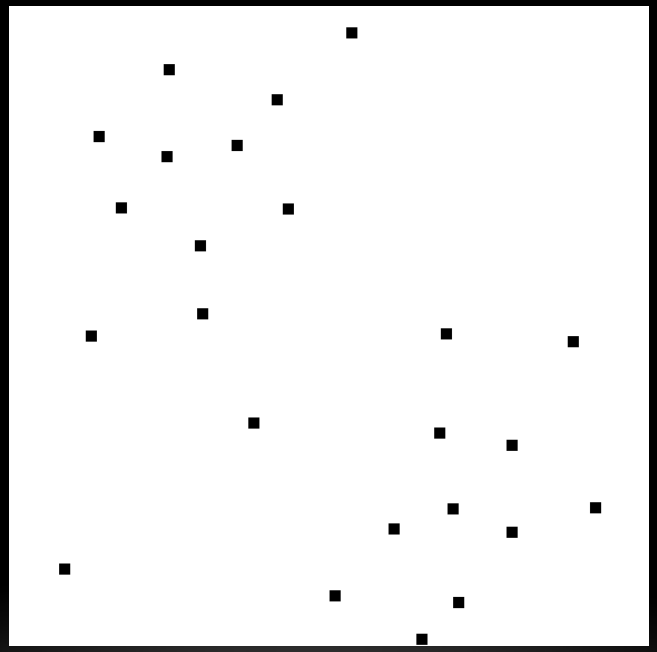


# MEAN CALCULATION AND RE-CLUSTERING



# K-MEANS++ INITIALIZATION ALGORITHM

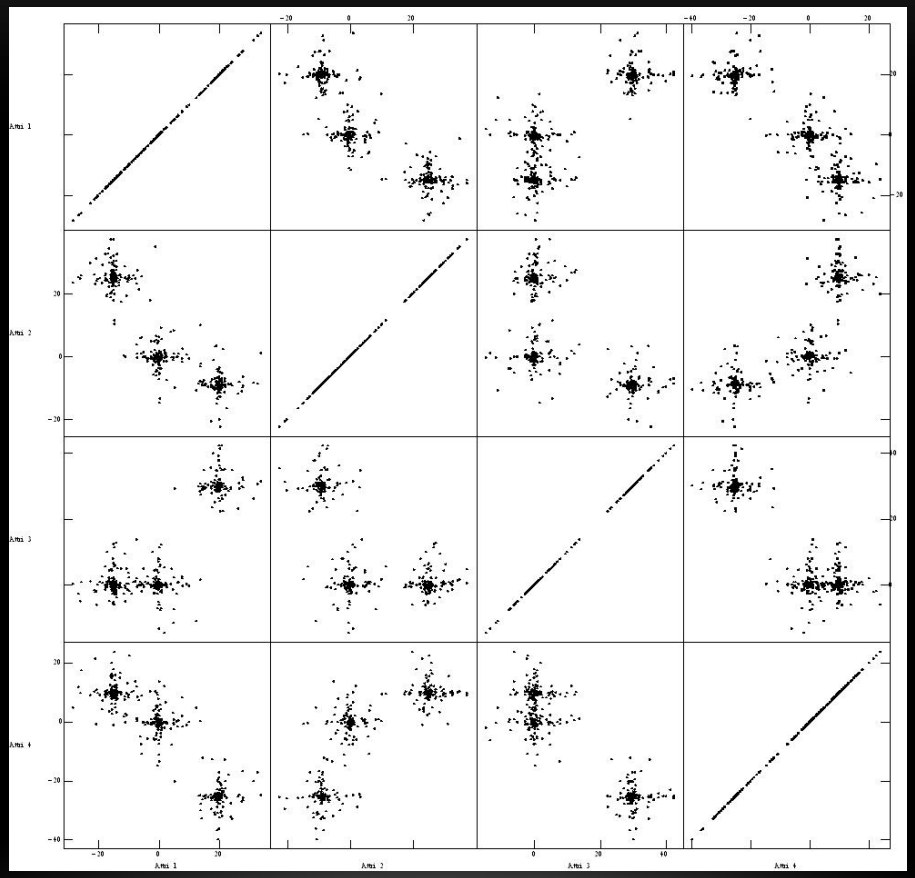
- Arbitrarily selects the first centroid.
- Every other centroids selected based on distance from other centroids.



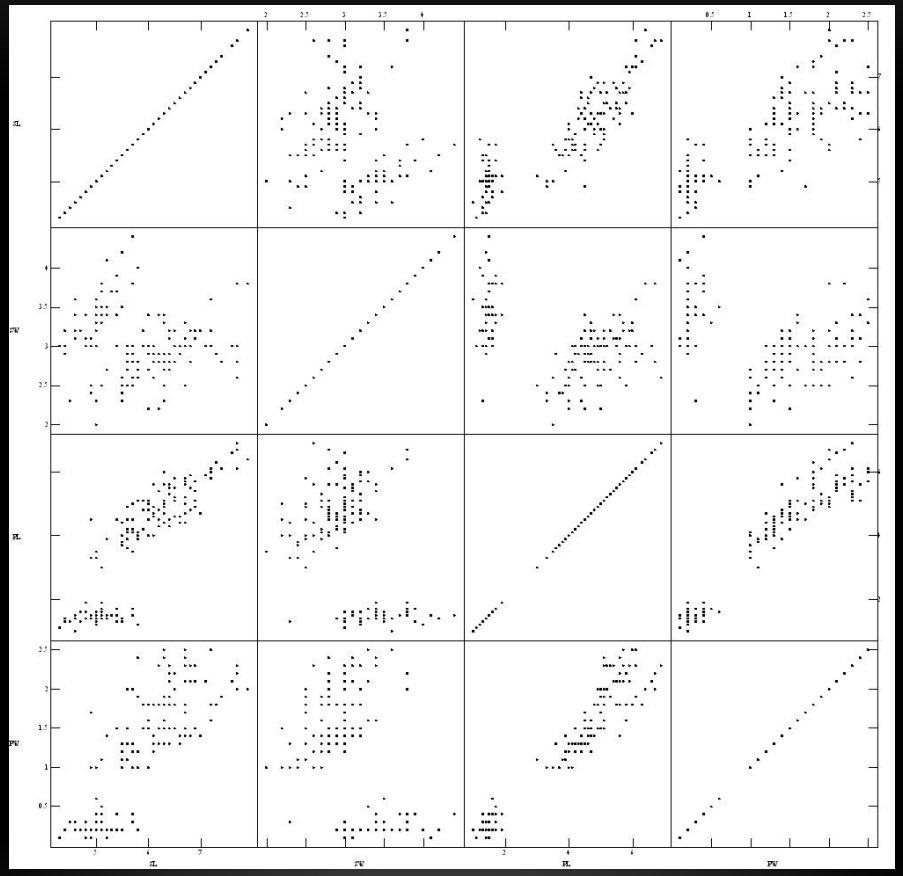
# EXPERIMENT

- Compared standard K-means and K-means++ methods.
- Goal: to discover if either one of them produces better results than the other.
- Setup:
  - Both methods run against 3 datasets with classes – Cluster, Iris, and Wine.
  - Each set has 3 classes which are used to verify the quality of the resulting clusters.
  - Quality in clusters is also determined by majority class
  - Fixed “arbitrary” setup to create a optimal and worst random centroid selection.
  - Both methods run against both centroid setups 3 times with a different K value.
  - Total of 36 trials.

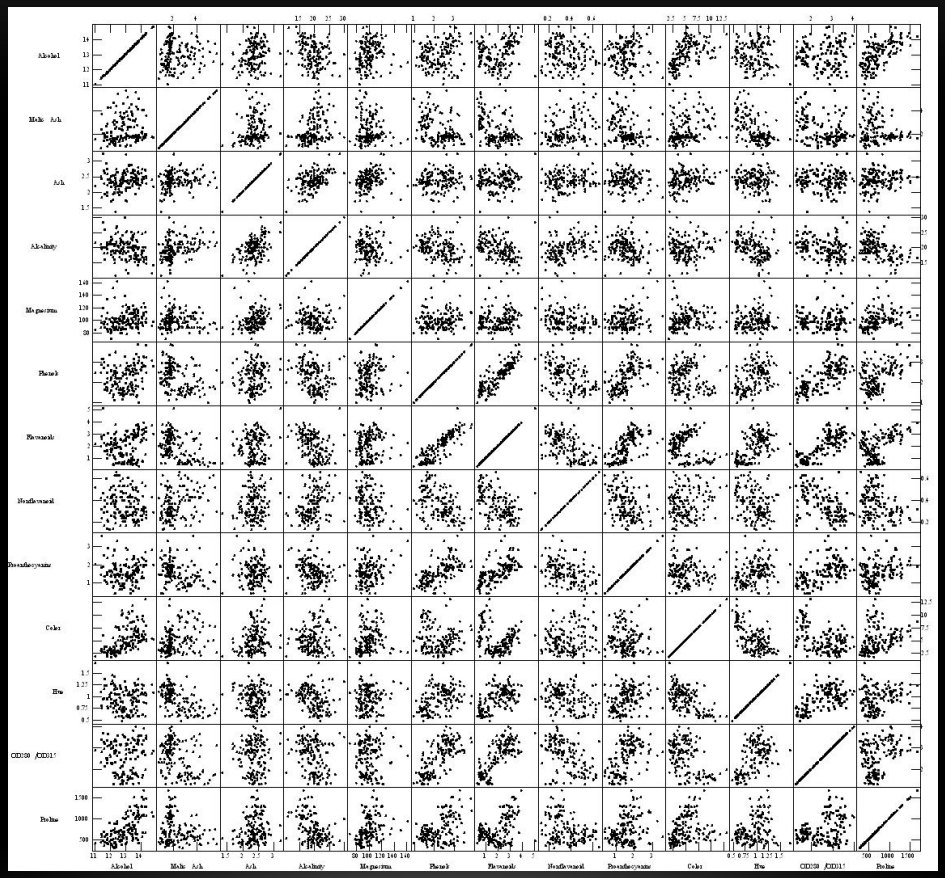
# MULTIDIMENSIONAL DATA - CLUSTER



# MULTIDIMENSIONAL DATA - IRIS



# MULTIDIMENSIONAL DATA - WINE



# RESULTS

- K-means++ proven to be better.
- No reason to use standard K-means.
- Still not perfect.

| K-means | | |
|---|---|---|
| Cluster Dataset | | |
| | Optimal | Worst |
| 3 Clusters | 1 | 1 |
| 5 Clusters | 0 | 1 |
| 7 Clusters | 1 | 1 |
| Iris Dataset | | |
| | Optimal | Worst |
| 3 Clusters | 17 | 18 |
| 5 Clusters | 23 | 24 |
| 7 Clusters | 10 | 17 |
| Wine Dataset | | |
| | Optimal | Worst |
| 3 Clusters | 46 | 53 |
| 5 Clusters | 39 | 44 |
| 7 Clusters | 42 | 43 |

| K-means++ | | |
|---|---|---|
| Cluster Dataset | | |
| | Optimal | Worst |
| 3 Clusters | 1 | 1 |
| 5 Clusters | 1 | 1 |
| 7 Clusters | 1 | 1 |
| Iris Dataset | | |
| | Optimal | Worst |
| 3 Clusters | 17 | 17 |
| 5 Clusters | 19 | 19 |
| 7 Clusters | 3 | 4 |
| Wine Dataset | | |
| | Optimal | Worst |
| 3 Clusters | 45 | 45 |
| 5 Clusters | 25 | 26 |
| 7 Clusters | 42 | 42 |

# IMPORTANT NOTES

- Imperfect simulation of K-means++
- Results could be better.
- Results should give clearer favor to K-means++

# REVIEW

- K-means Clustering Algorithm
- K-means++ Initialization Algorithm
- Comparison Experiment
- Multidimensional Datasets
- Results

# WORKS CITED


- Aleshunas, J. (2013). Cluster Set.
- Alsabti, K., Ranka, S., & Singh, V. (1997). *An effcient k-means clustering algorithm.*
- Arthur, D., & Vassilvitskii, S. (2007). *K-means++: the advantages of careful seeding.* Philadelphia: Society for Industrial and Applied Mathematics Philadelphia.
- Fisher, R. A. (1936). Iris Flower Data Set.
- Forina, M. (1988). Wine Recognition Data. *PARVUS: An extendable package of programs for data exploration, classification and correlation*. Genoa, Italy: Institute of Pharmaceutical and Food Analysis and Technologies.
- Inaba, M., Katoh, N., & Imai, H. (1994). Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. *SCG '94 Proceedings of the tenth annual symposium on Computational geometry* (pp. 332-339). New York: ACM.
- MacKay, D. (2003). An Example Inference Task: Clustering. In D. MacKay, *Information Theory, Inference and Learning Algorithms* (pp. 284-292). Cambridge University Press.
- Shaefer, I. (2013). Cluster Set Modified.