ARC

A BETTER CLUSTERING ALGORITHM

Robert Carter
Conor French
Noah Armsworthy

Group 26

The Problem

How many clusters?

Most clustering algorithms require a human selected number of clusters, or running the model over a range to find best estimated match.

What about arbitrary data shapes?

Cluster models need to be selected based on how they perform with specific data shapes.

Implementation Details

LANGUAGE

We opted to use Python so we could utilize libraries like matplotlib and numpy.

DATA STRUCTURE

We organized the clustering algorithm around the distance matrix of all points, computing it once to get O(n^2) time.

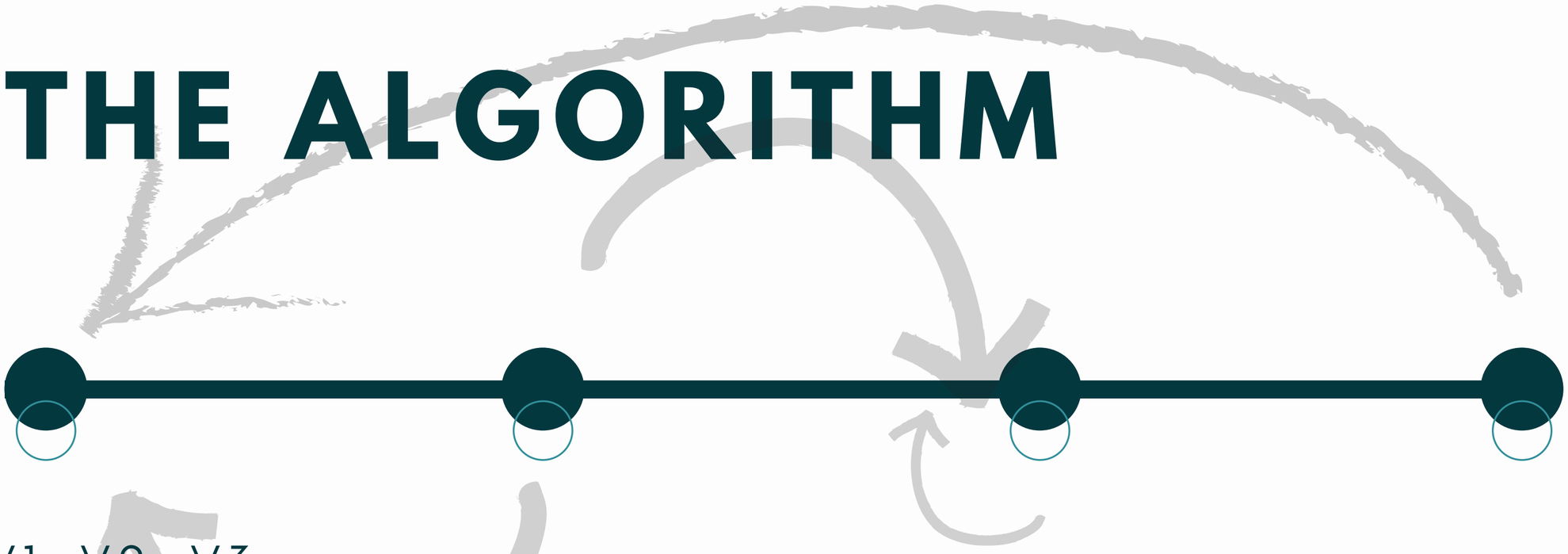
SYNTHETIC DATA

During the programming and testing phase, we generated our data synthetically so that we could measure performance on multiple arbitrary shapes and densities.

TODO: REAL-WORLD DATA

Online Shoppers Purchasing Intention: 18 Attributes representing 12,330 shopping sessions from distinct users.

THE ALGORITHM



V1, V2, V3

Formula 4, 5, 6

In short: V1 is the least distant point from all other points. V2, V3 are the nearest points to V1.

OVERCOME BAD INIT

Formula 7 - if its conditions don't hold, V1 is an outlier!

EXPAND TO E-RADIUS NEIGHBORS

Formula 2 or 3
This step is repeated until there are no new neighbors

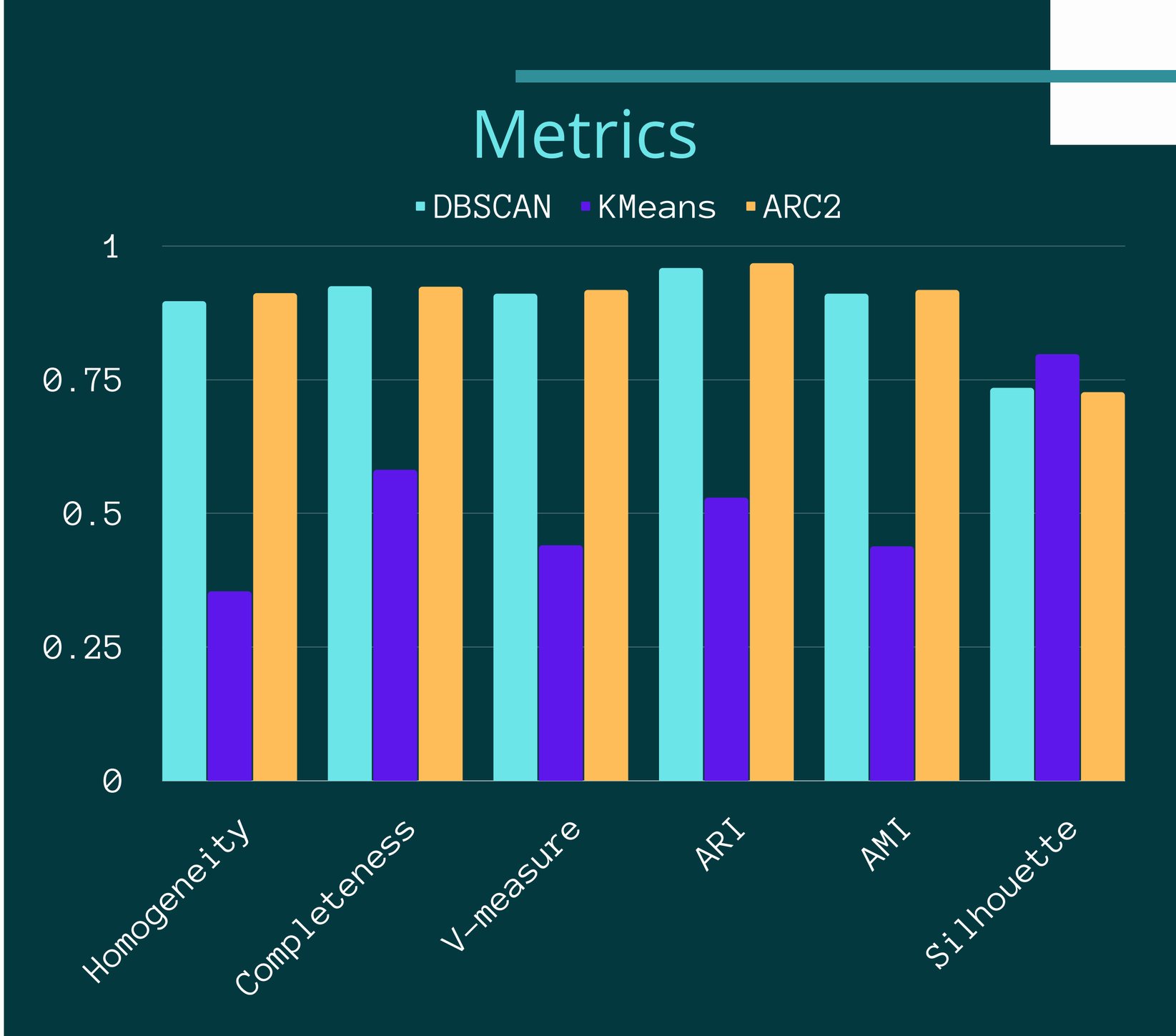
REPEAT WITH REST OF DATA

The process repeats to identify all possible clusters and outliers.

EXPERIMENTAL RESULTS

SUMMARY

- Preliminary synthetic gaussian clusters (Dense & Sparse)
- KMeans manually set to 2 clusters.
  - Looping through a range of clusters retuned increasing score
  - KMeans aims to make only circle-shaped clusters.
  - Aims to minimize inertia (cluster spread)
- DBSCAN stats misleading
  - Predicted 1 cluster
  - Metric calculations still considered -1 as a cluster label.



DBSCAN vs ARC Error

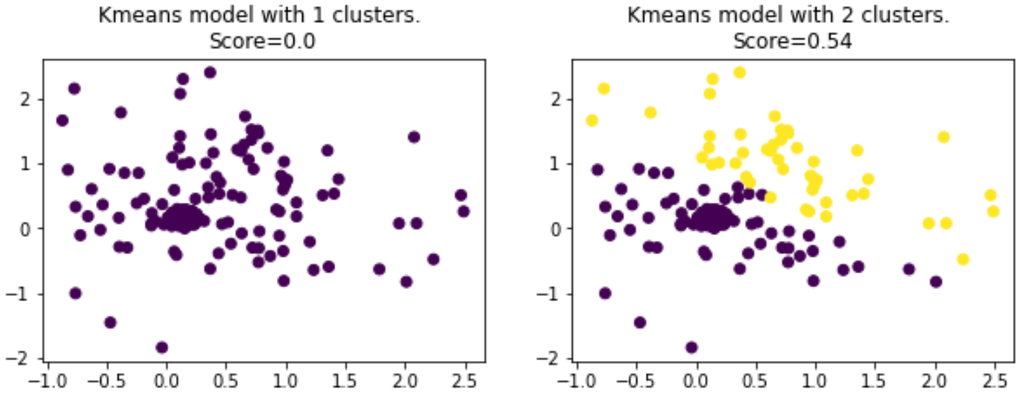
100/600(16.6%)

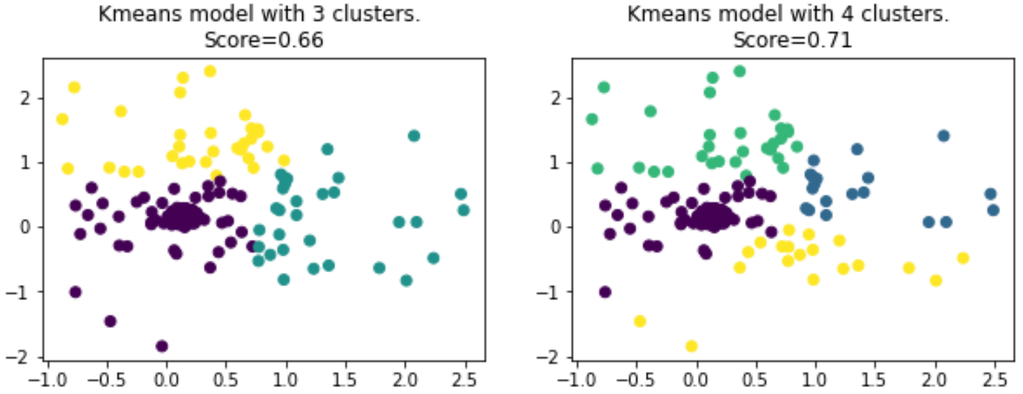
- DBSCAN INCORRECT PREDICTIONS.
- ENTIRE CLUSTER CLASSIFIED AS OUTLIER (SPARSE)

4/600 (0.6%)

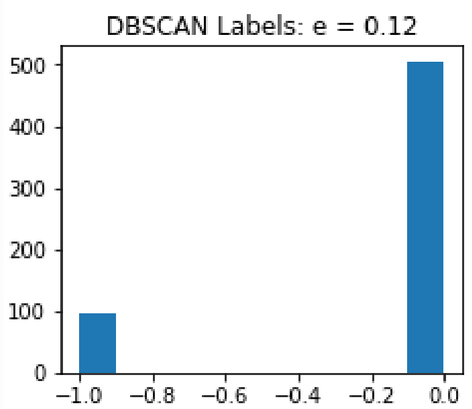
- ARC INCORRECT PREDICTION

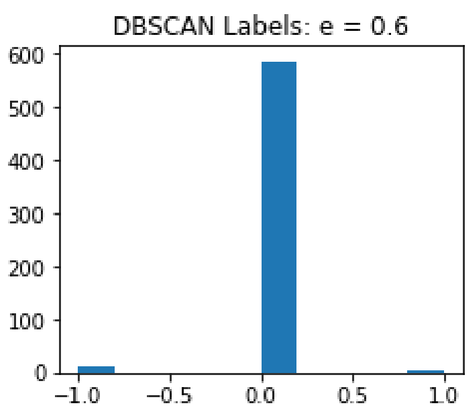
KMeans

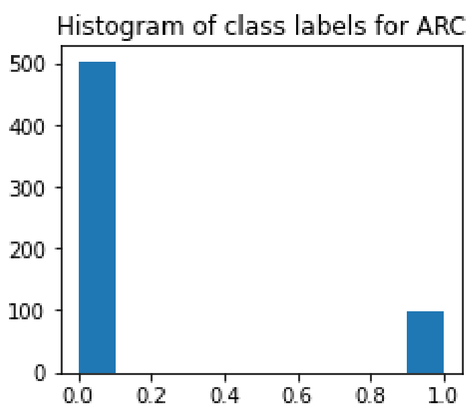


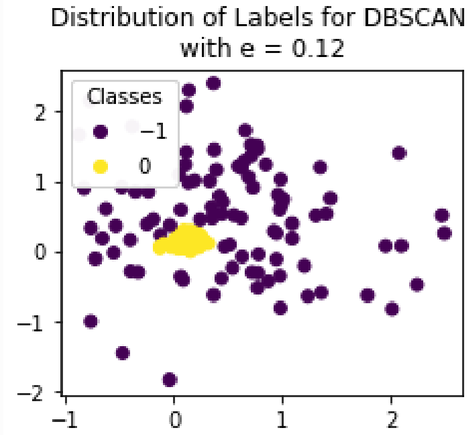


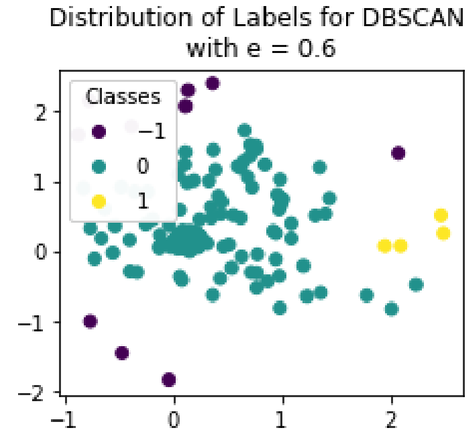
DBScan & ARC

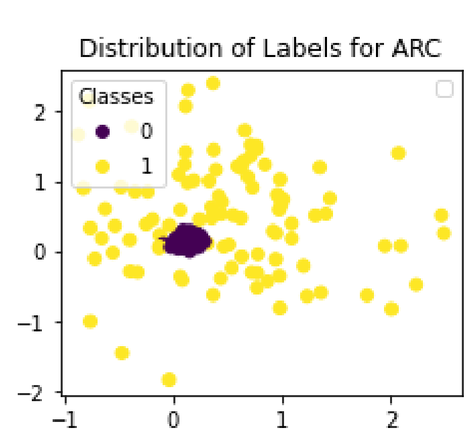


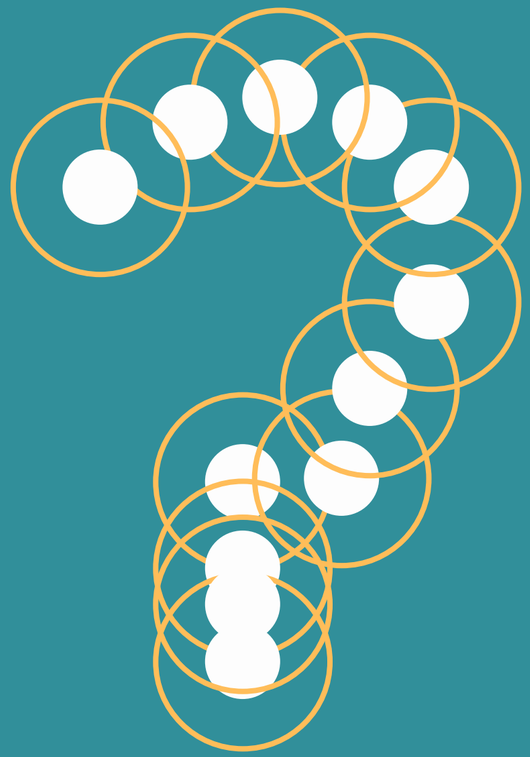












Any Questions?

Reference Paper

T. Vo-Van, A. Nguyen-Hai, M. V. Tat-Hong, T. Nguyen-Trang, "A New Clustering Algorithm and Its Application in Assessing the Quality of Underground Water", Scientific Programming, vol. 2020, Article ID 6458576, 12 pages, 2020. https://doi.org/10.1155/2020/6458576