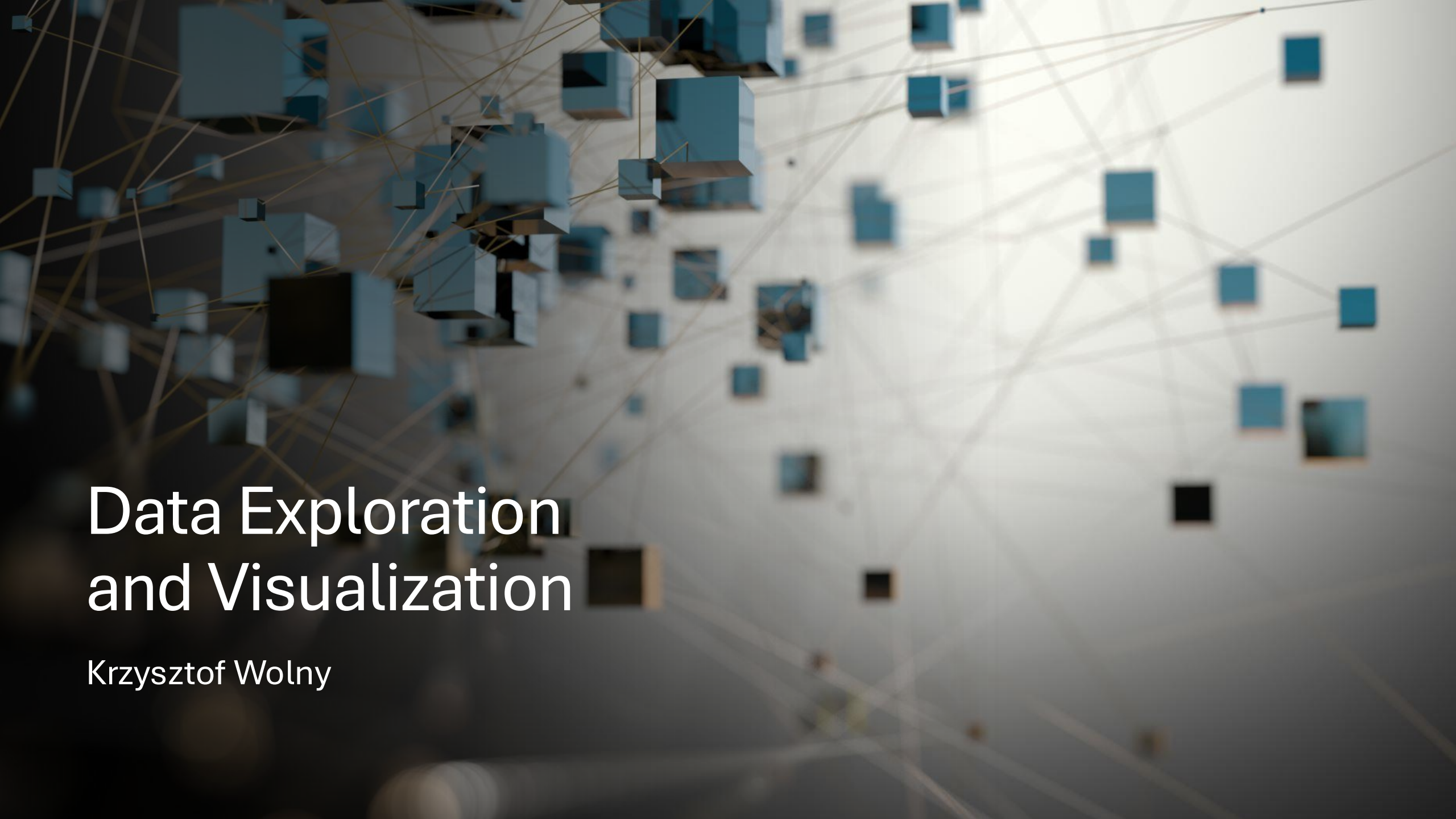


Data Exploration and Visualization

Krzysztof Wolny

Exercise 1

Synthetic Dataset

- Synthetic classification dataset using make_classification from scikit-learn. The dataset was designed with the following characteristics:
  - Total samples: 1,000
  - Features per sample: 20
  - Informative features: 5
  - Noise features: 15
  - Number of classes: 2 (binary classification)

Methodology

Create synthetic dataset

Select top 5 features using:

RFE - Recursive Feature Elimination used with SVM

CS - Correlation-based Selection

DTC - Decision Tree Classifier-based Selection

Evaluate accuracy, precision, recall and F1 score on SVM and Random Forest models on :

Set with all features

Sets with features selected by chosen feature selection methods

Set of only important features

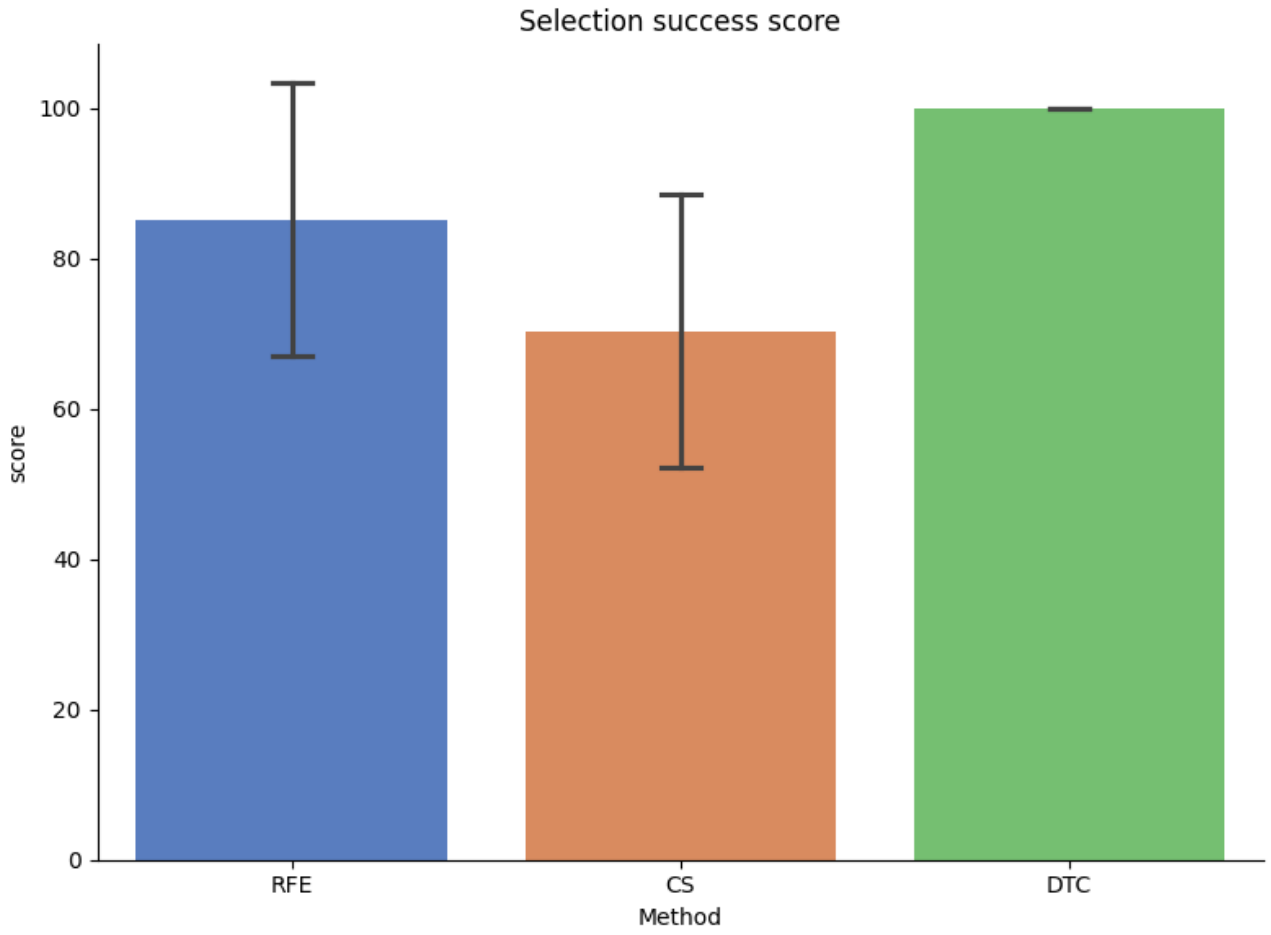
Data projection obtained via:

- PCA,
- Multidimensional scaling,
- TSNE

Repeat whole proccess 6 times in order to get average behaviour

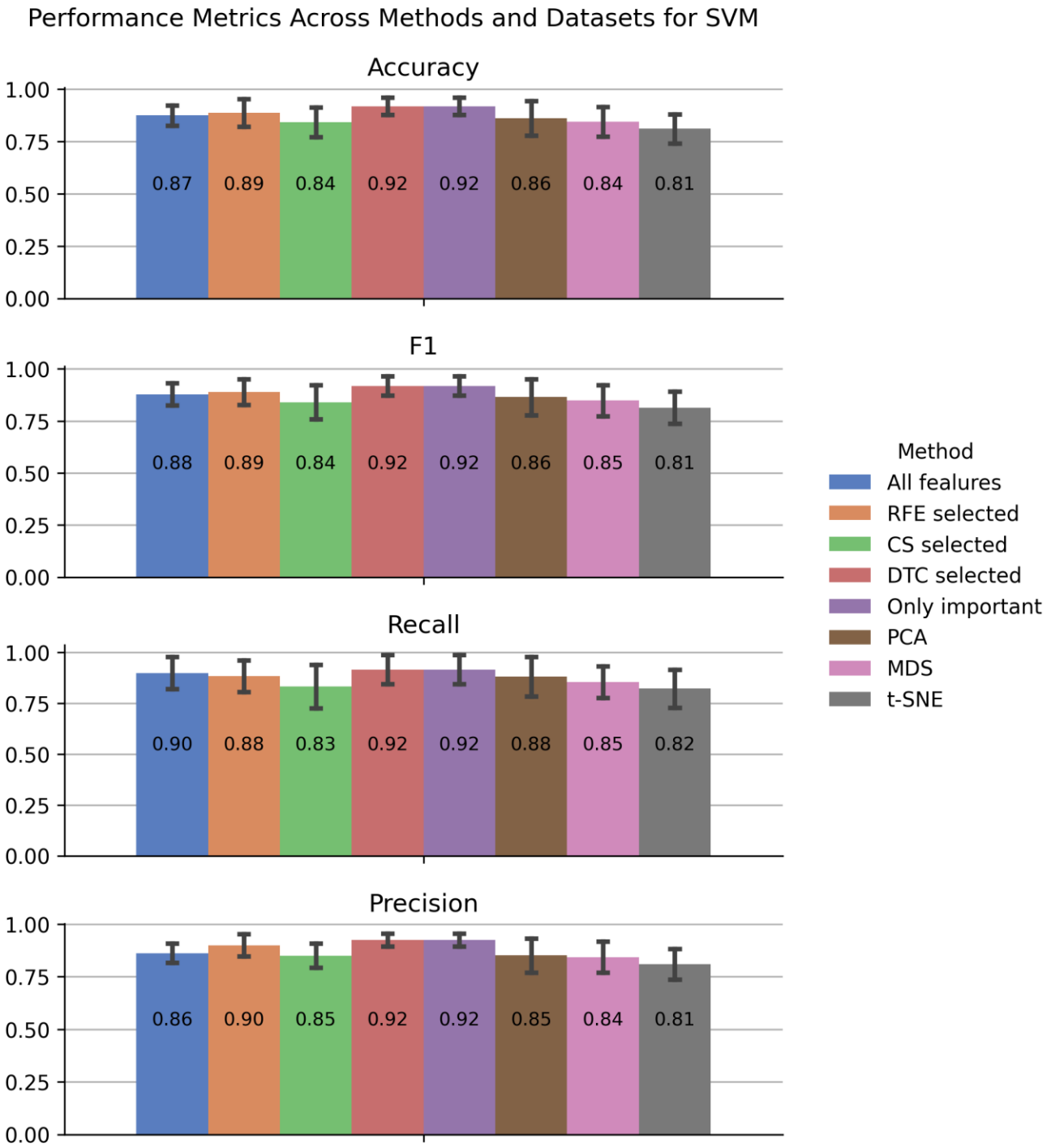
Evaluation of feature selection methods

- Success score suggested in (Bolón-Canedo et. al. 2012)
- Best results were obtained using DTC
- CS had the worst score



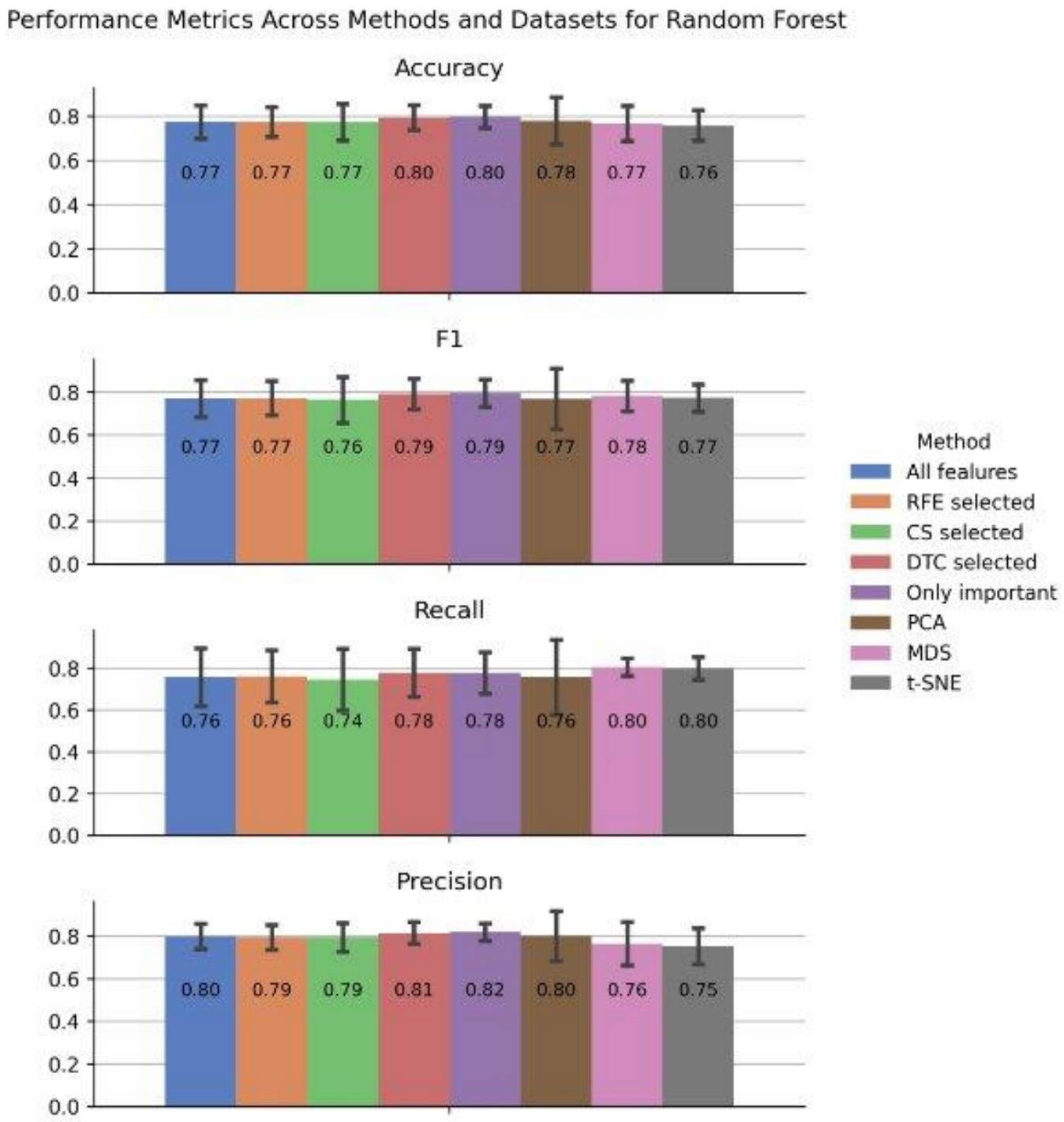
Evaluation for SVM

- Similarly to selection score CS gets worse results than any other selection
- Best results out of dimensionality reduction techniques get PCA
- DTC and only important features get the same score – DTC is selecting only important features



Evaluation for Random Forest

- PCA and CS have higher variability
- Random forest gets worse results then SVM
- CS gets similar accuracy and F1 to RFE



Exercise 2

Datasets

- I have analyzed 16 datasets:
  - a1,
  - a2,
  - a3,
  - aggregation,
  - compound,
  - d31,
  - r15,
  - flame,
  - jain,
  - pathbased,
  - spiral,
  - s1,
  - s2,
  - s3,
  - s4,
  - Unbalance
- For larger datasets I sampled 10 000 observations.

Clustering Algorithms

- K-means
- Genie (with parameter $g \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$)
- Agglomerative Hierarchical Clustering (with single, average, complete and Ward linkage)
- DBSCAN

DBSCAN parameters selection

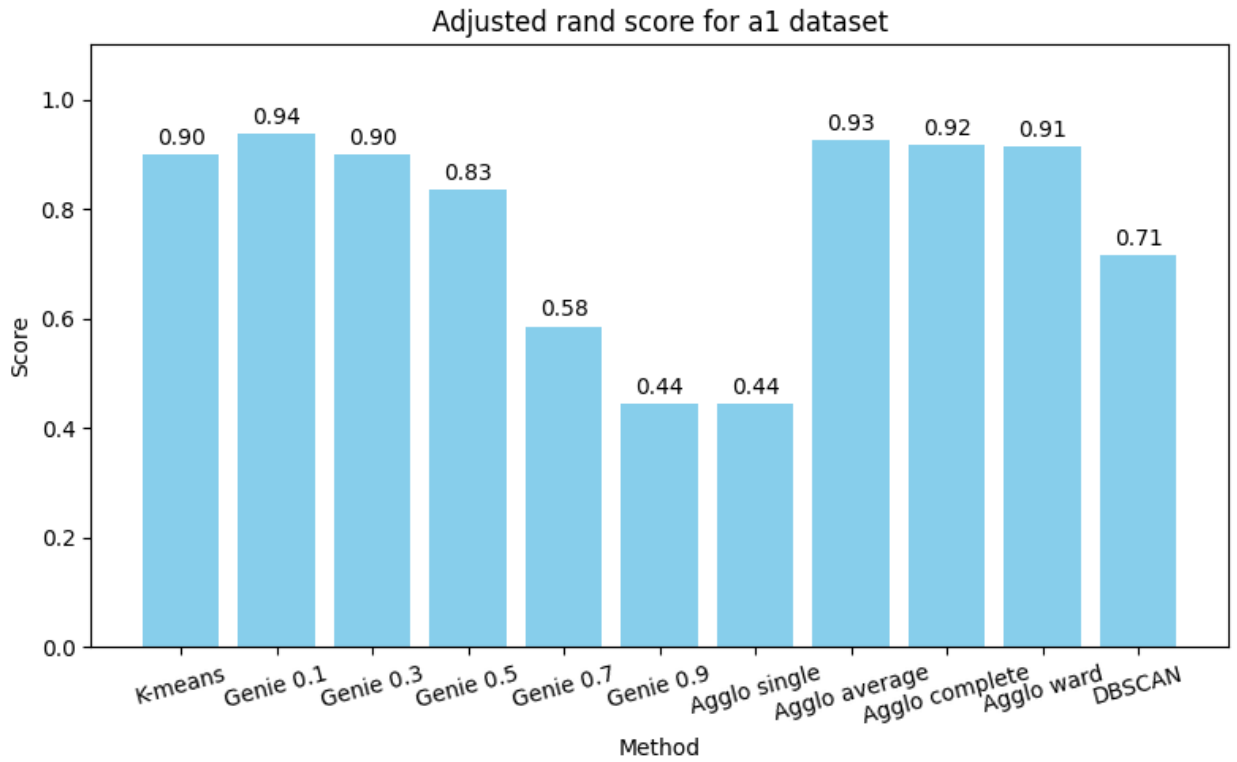
- Searching for the best parametrs
- Checking Min samples = 4 – 10
- Calculating Eps candidate using "elbow" method. Calculating elbow with kneed library
- Check Eps from 0.5 * Eps to 1.5 Eps with step 0.1* Eps
- Fit and predict DBSCAN with each combination.
- Choose best parameters using:
  - Real number of clusters in dataset
  - Number of clusters from DBSCAN
  - Share of outliers in a prediction
  - Adjusted Rand Index of a fit
  - Visualization of clustering

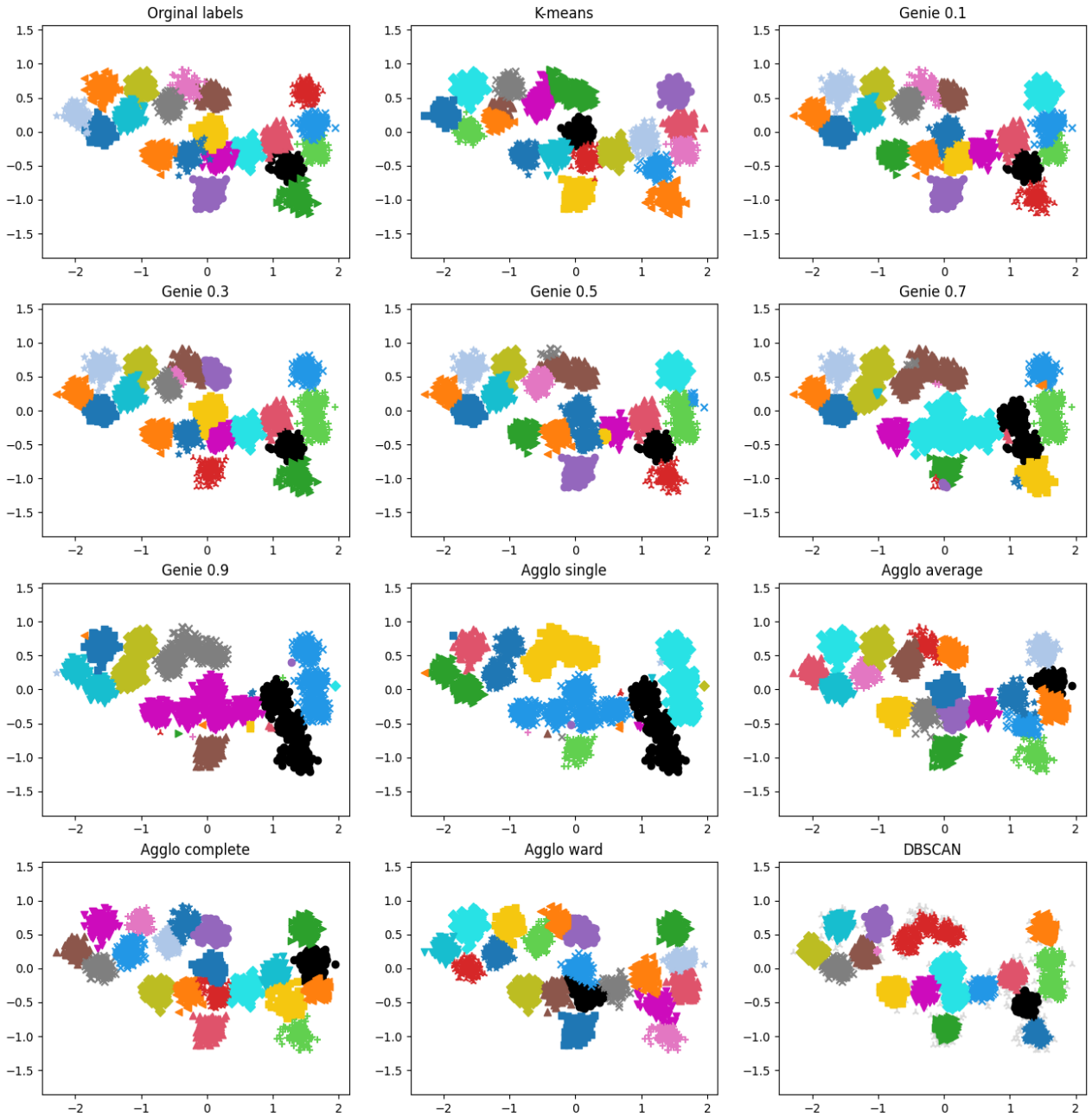
Dataset	Min Samples	Eps
a1	7	0.061111
a2	4	0.047888
a3	10	0.061540
aggregation	7	0.144697
compound	6	0.170477
d31	10	0.085117
r15	9	0.126741
flame	8	0.163551
jain	4	0.235065
pathbased	5	0.248905
spiral	4	0.277380
s1	10	0.113822
s2	9	0.095061
s3	10	0.091968
s4	7	0.092762
unbalance	4	0.120815

Table 1: Final DBSCAN parameters per each dataset

a1

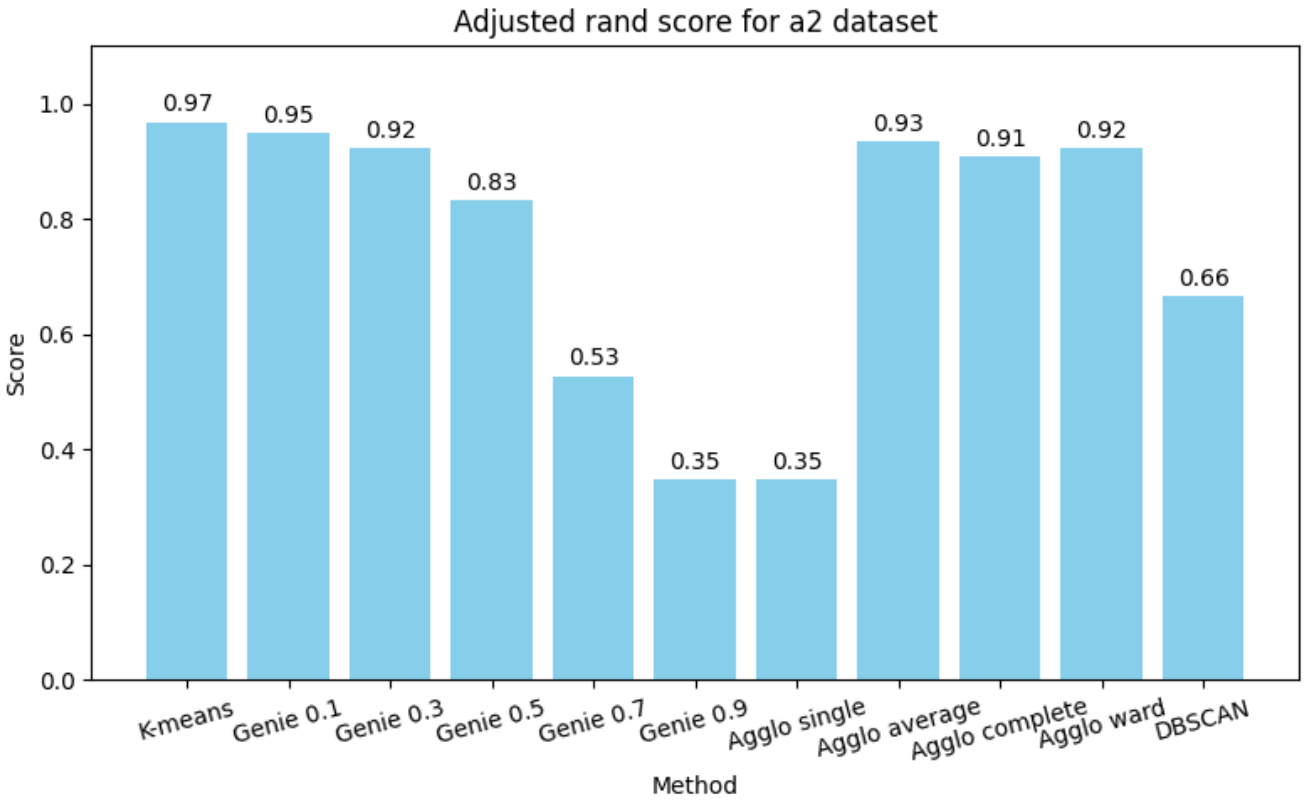
- The best are K-means and Genie (g=.1)
- Genie(g=0.1) and Agglomerative Hierarchical Clustering with single linkage have similar results(it will be repeating

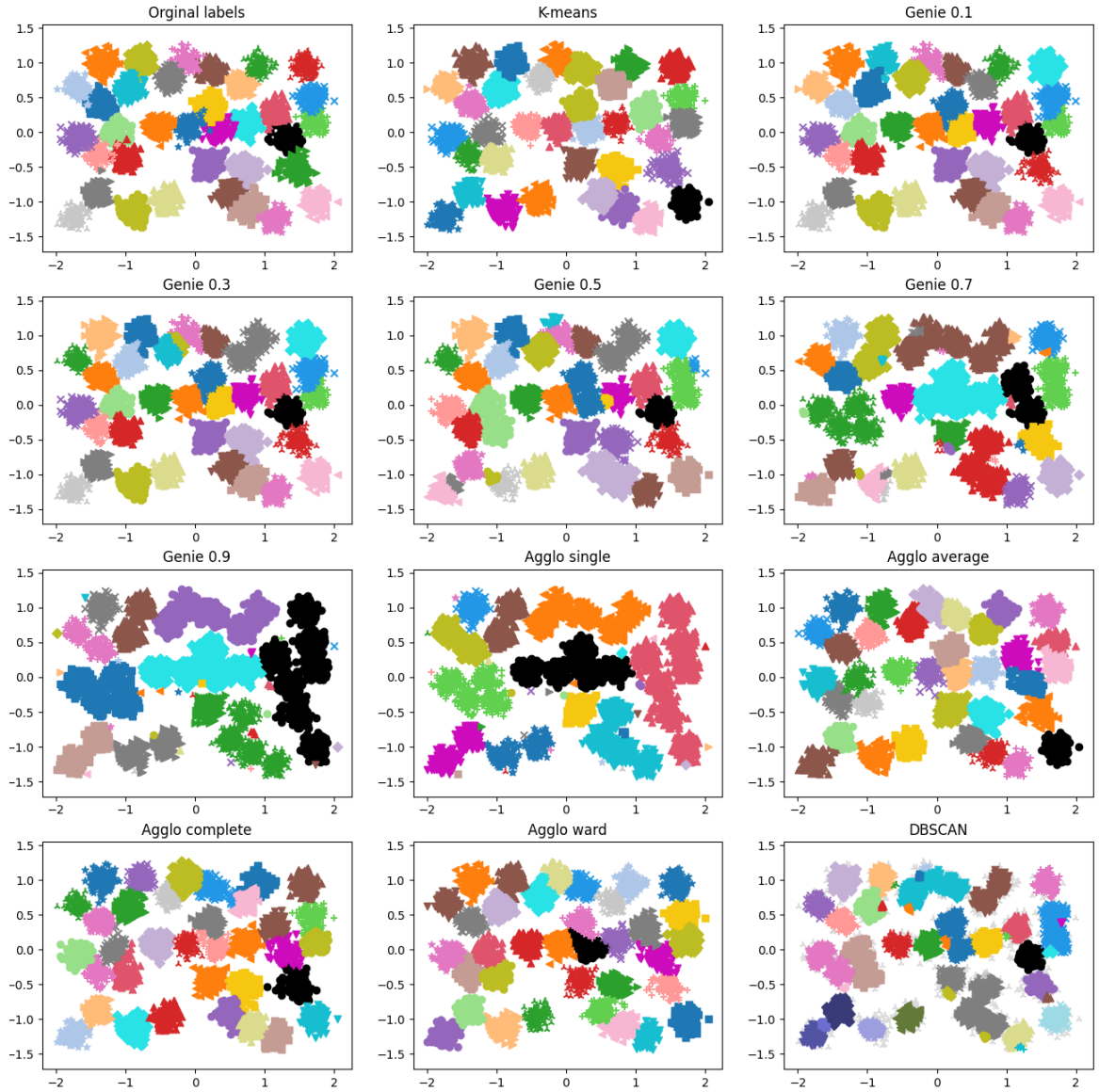




a2

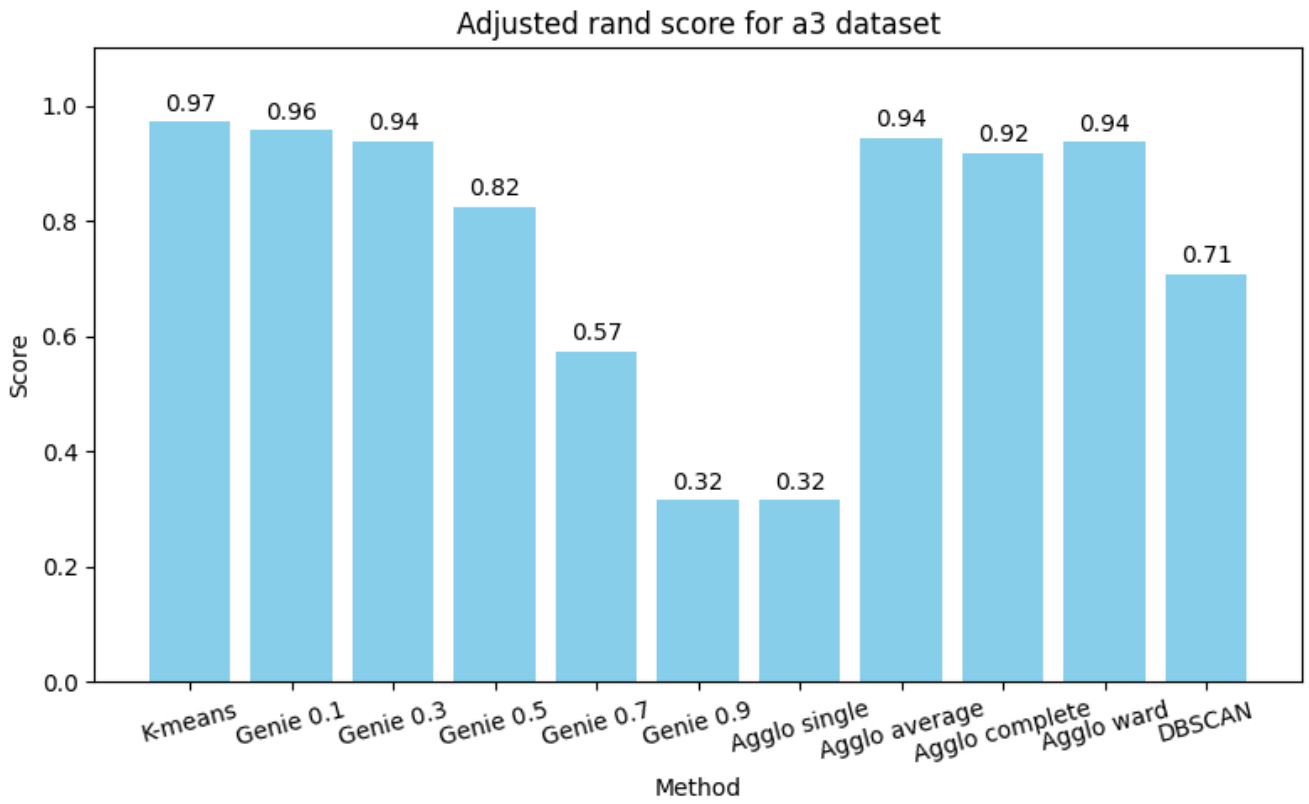
- In this dataset it is easy for cluster to start "eating" next bubbles
- The best with dealing with that is K-means, Genie with small g's and AHC with average, complete and ward linkage

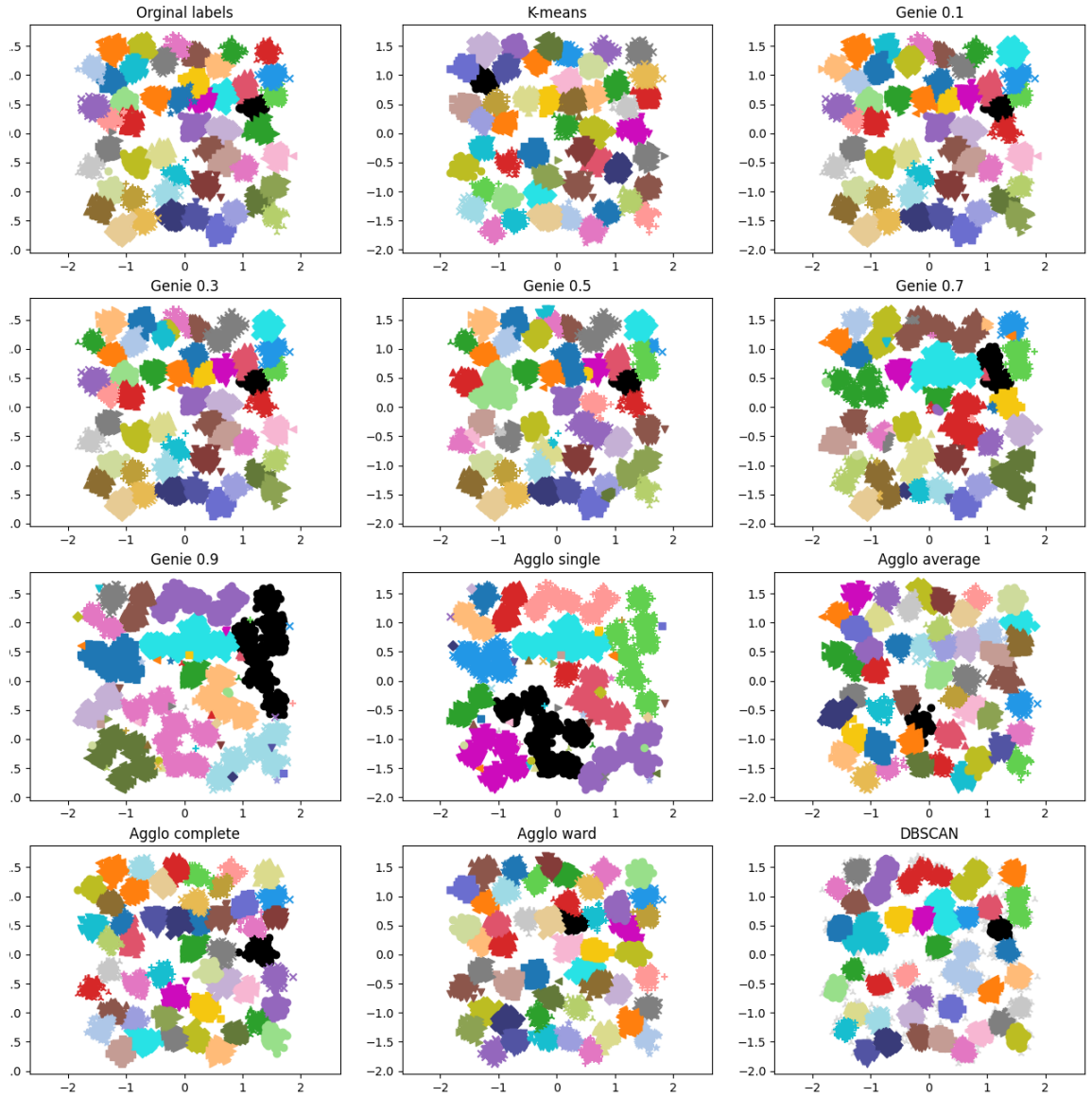




a3

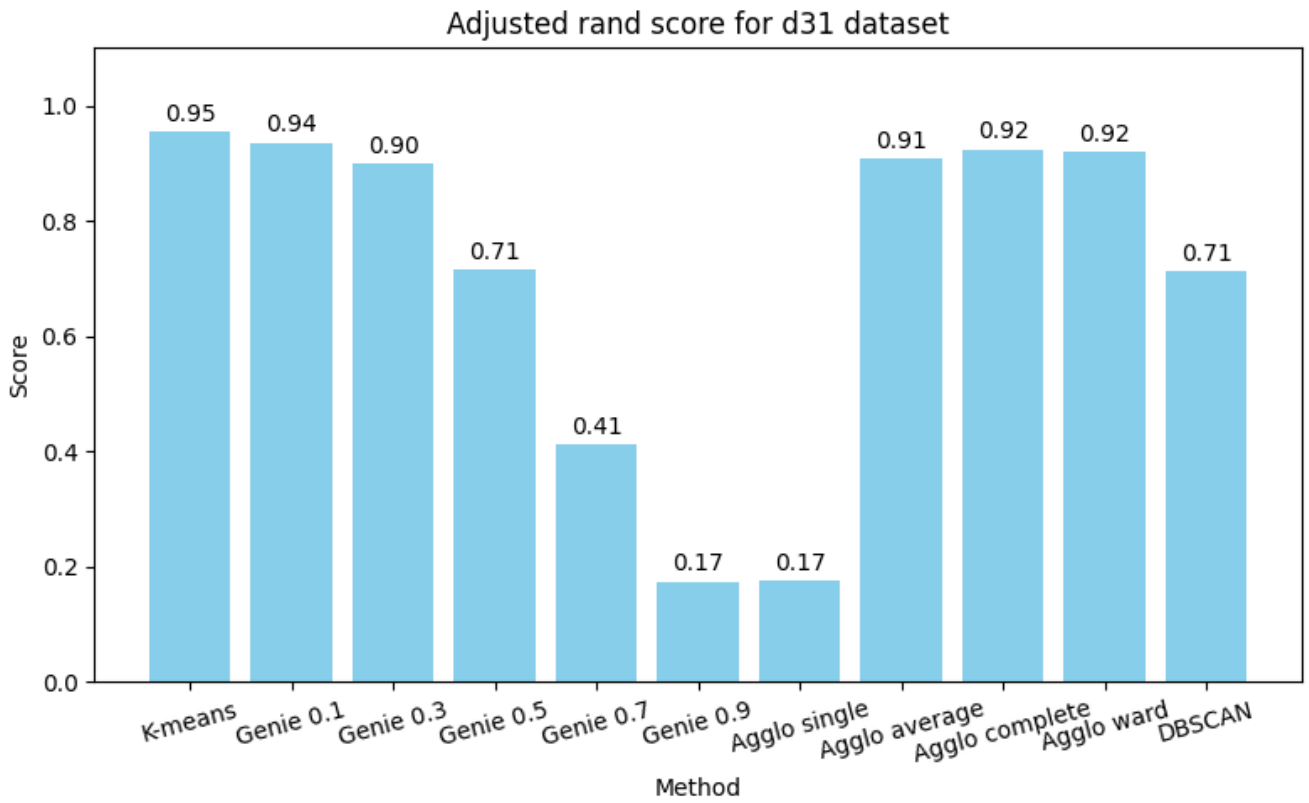
- Similar results like last example

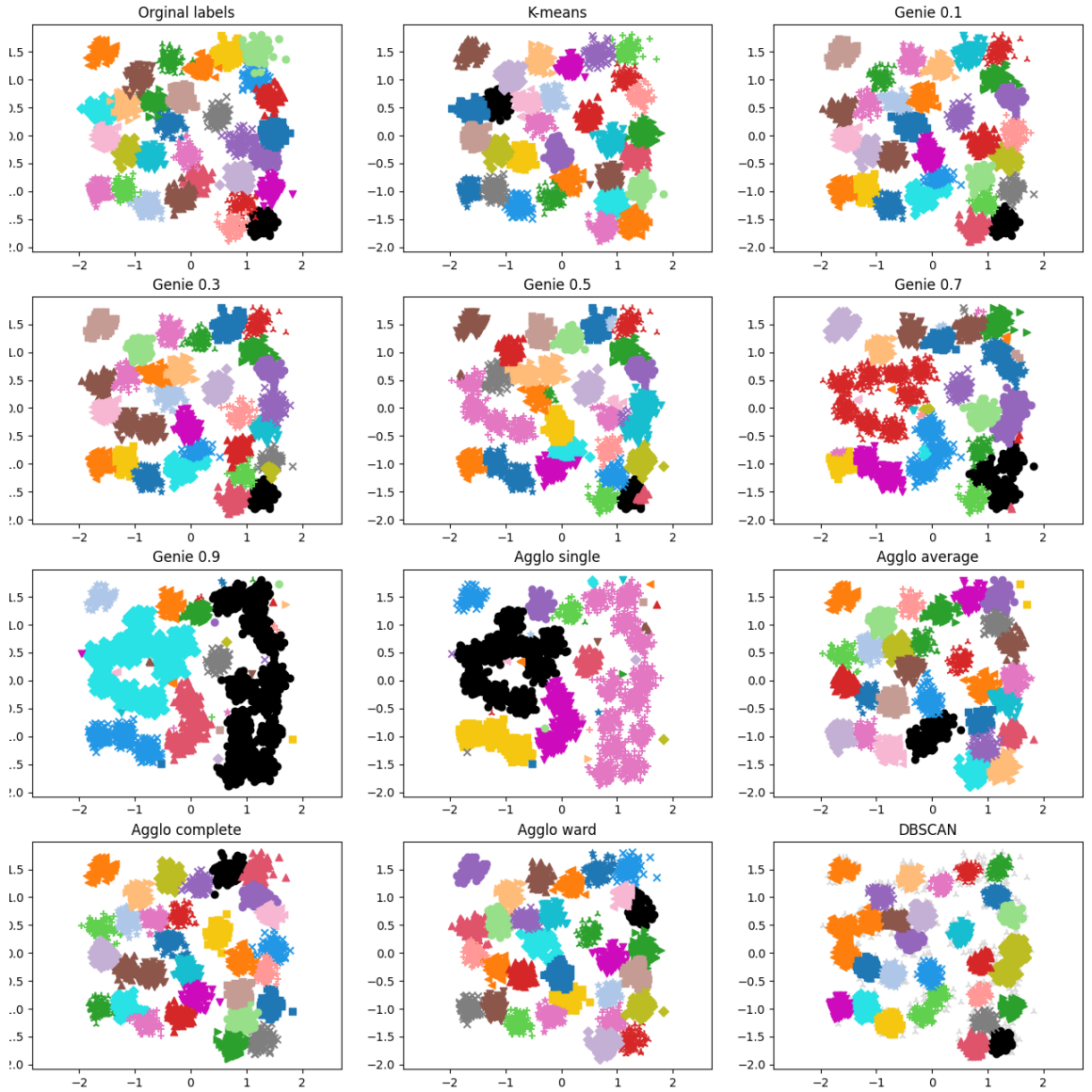




d31

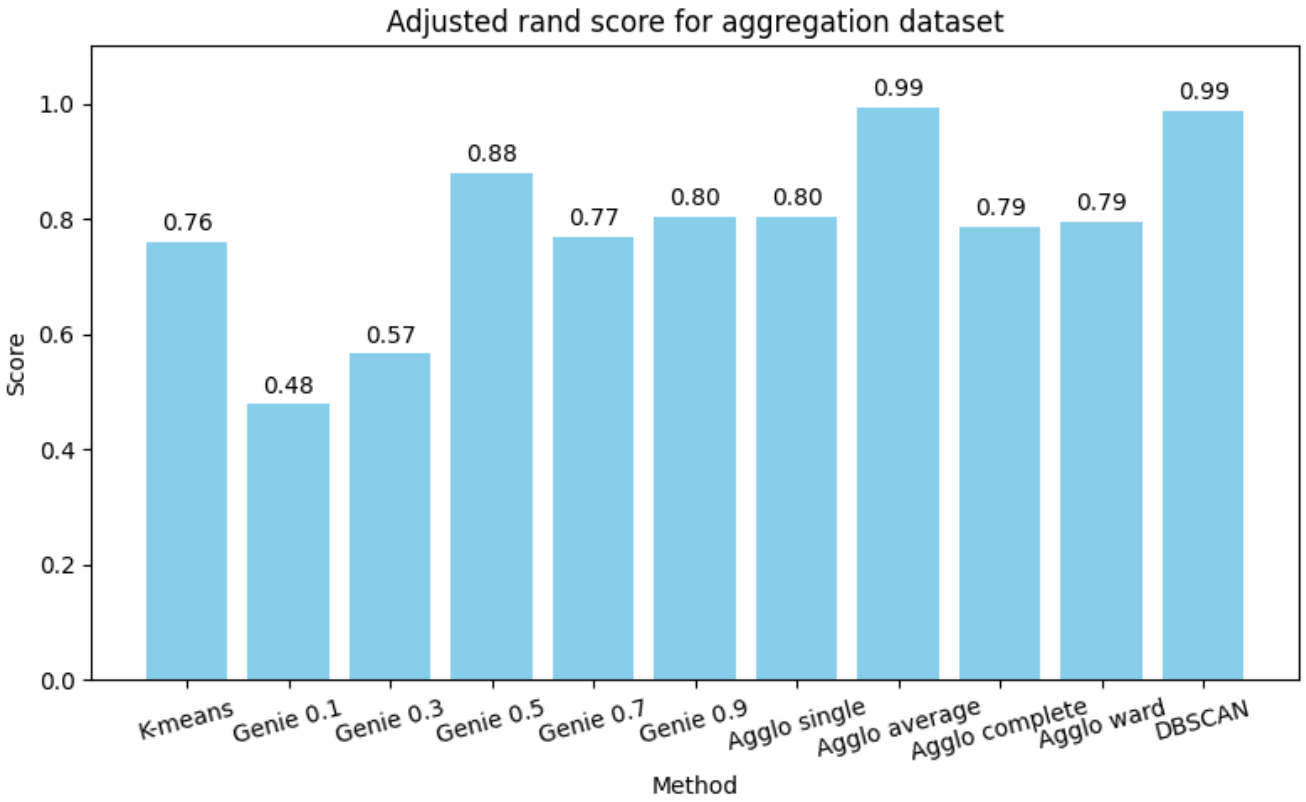
- Similar results like last example

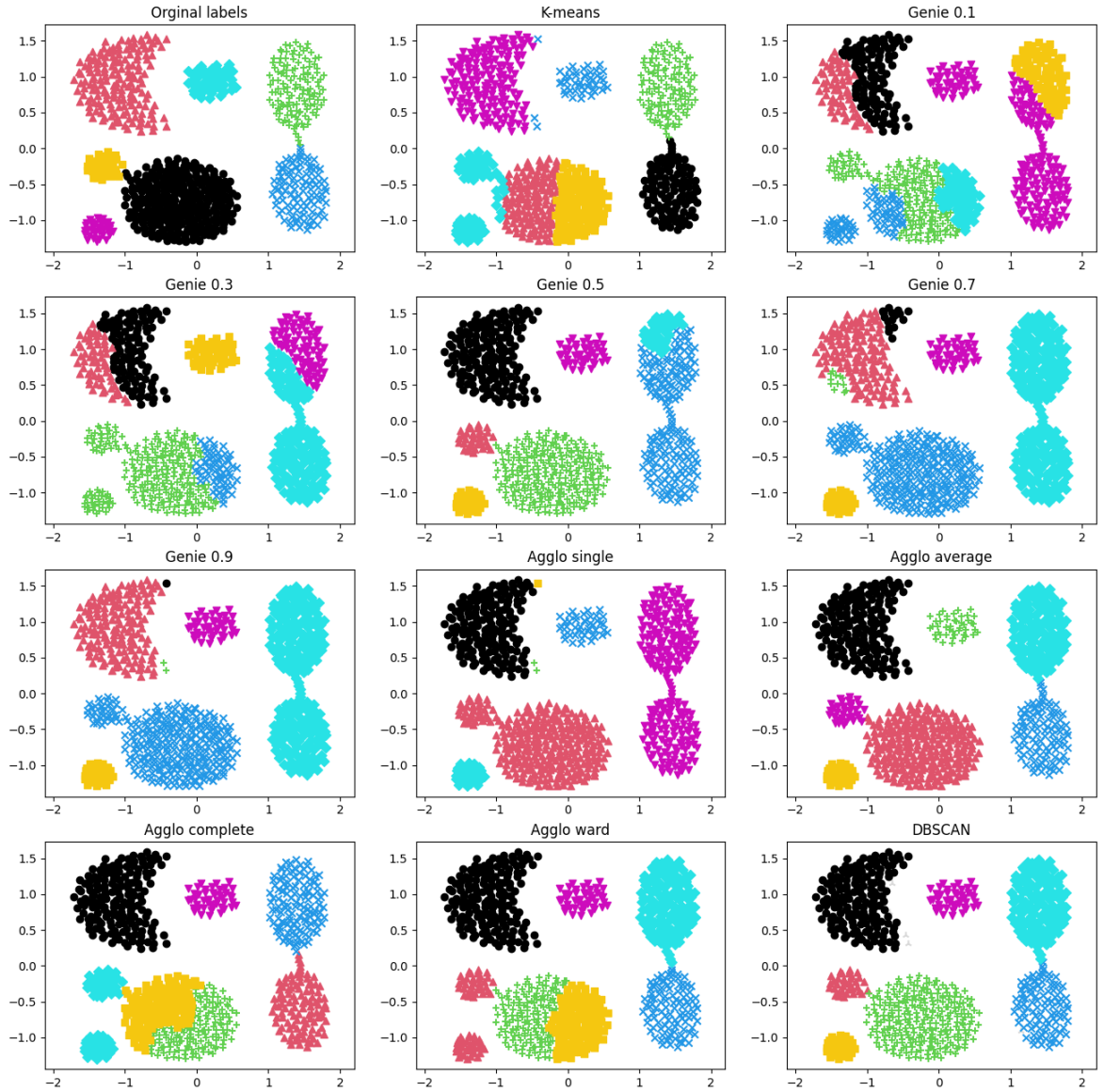




aggregation

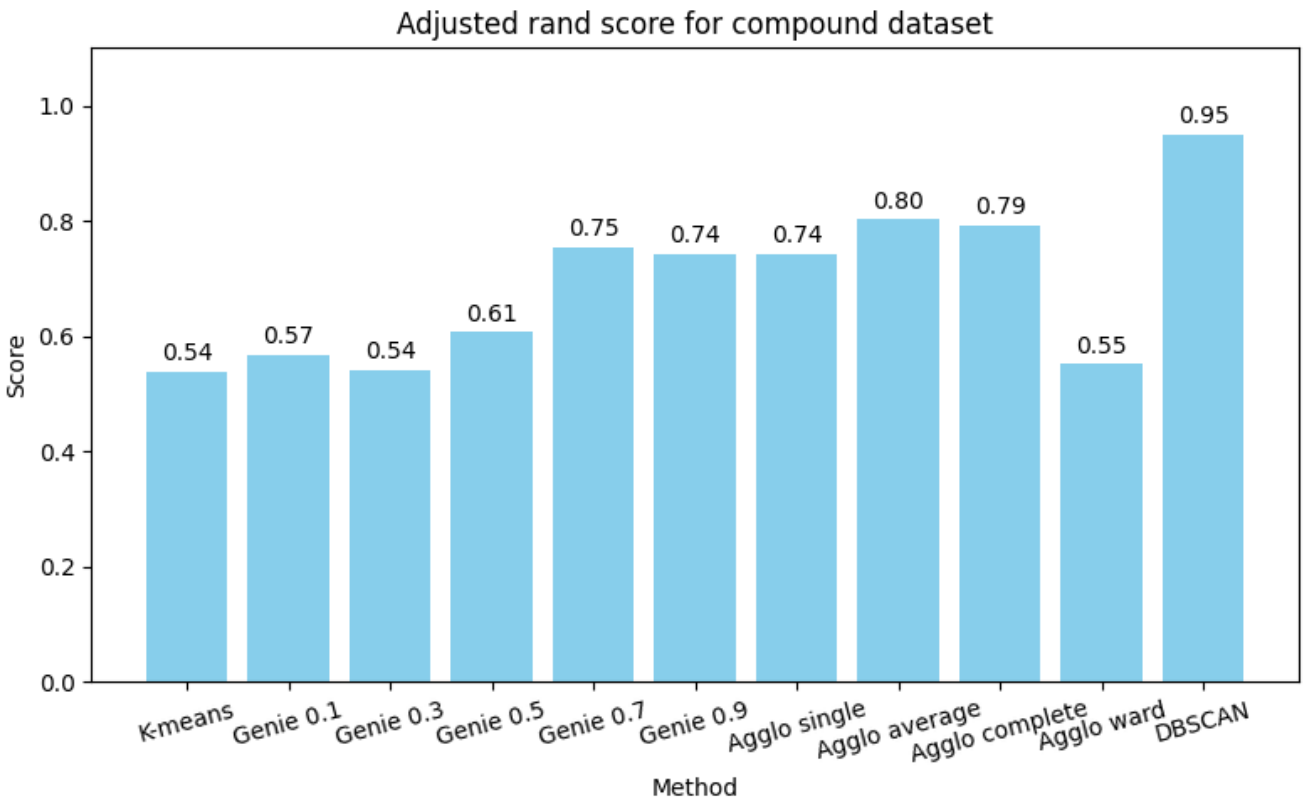
- The best results are from AHC with average linkage and DBSCAN

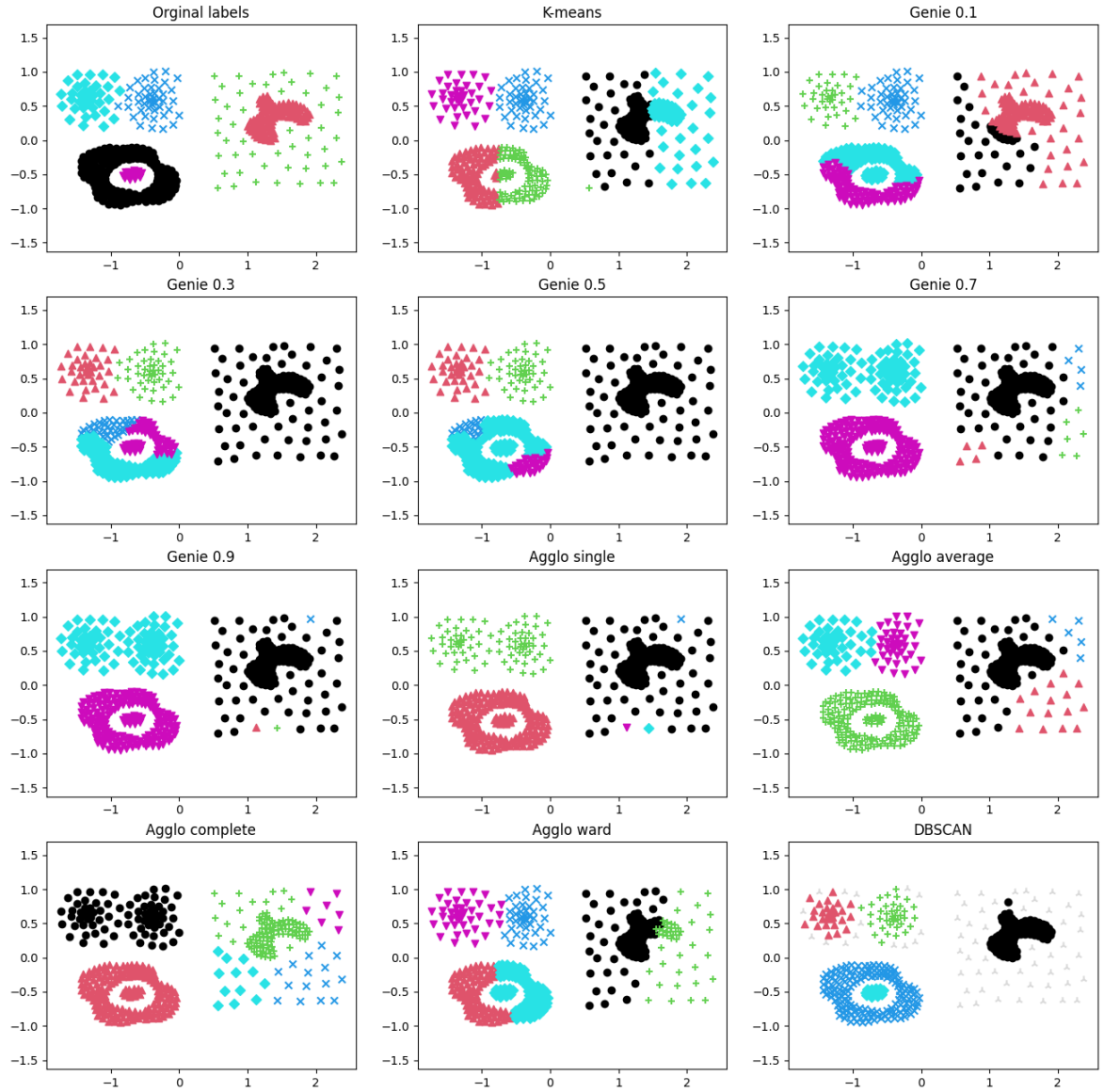




compound

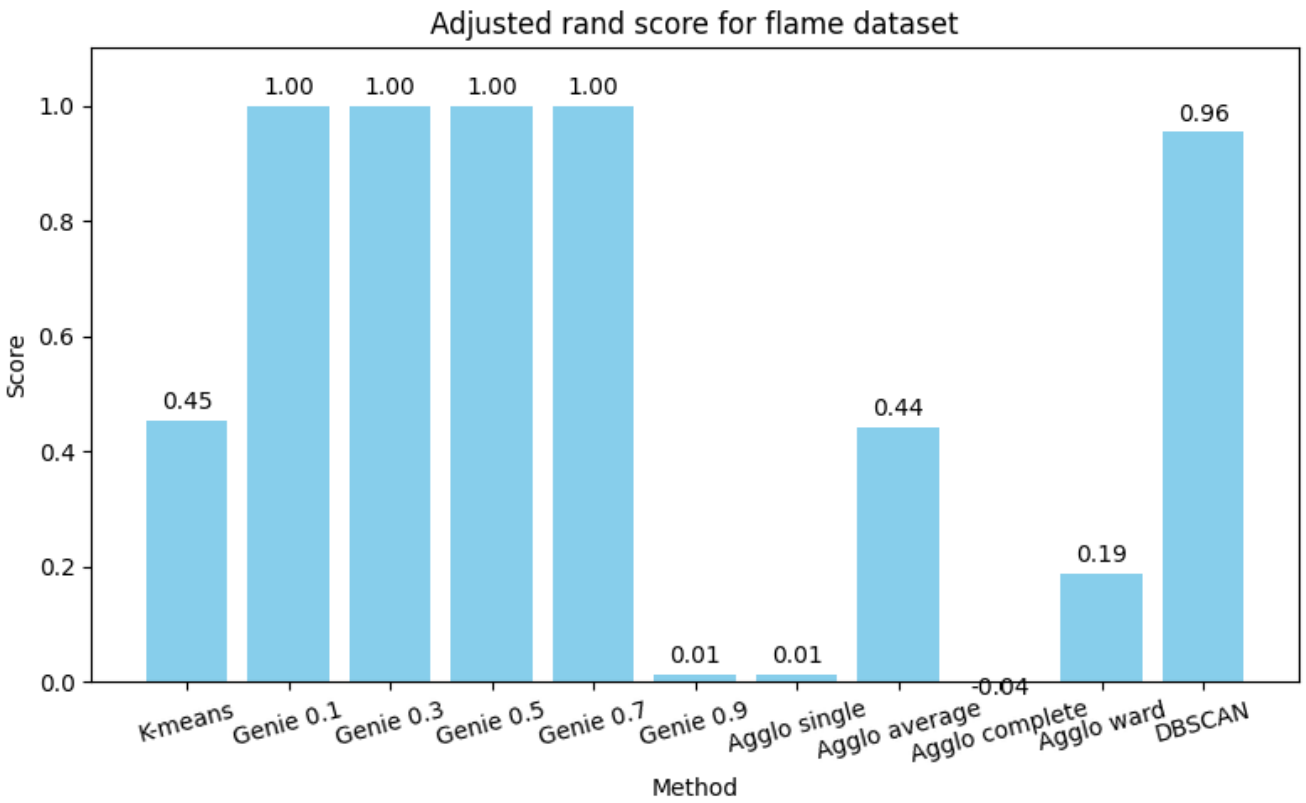
- DBSCAN got best score, because it classified least dense regions as outliers
- None of the algorithms are able to get very good results

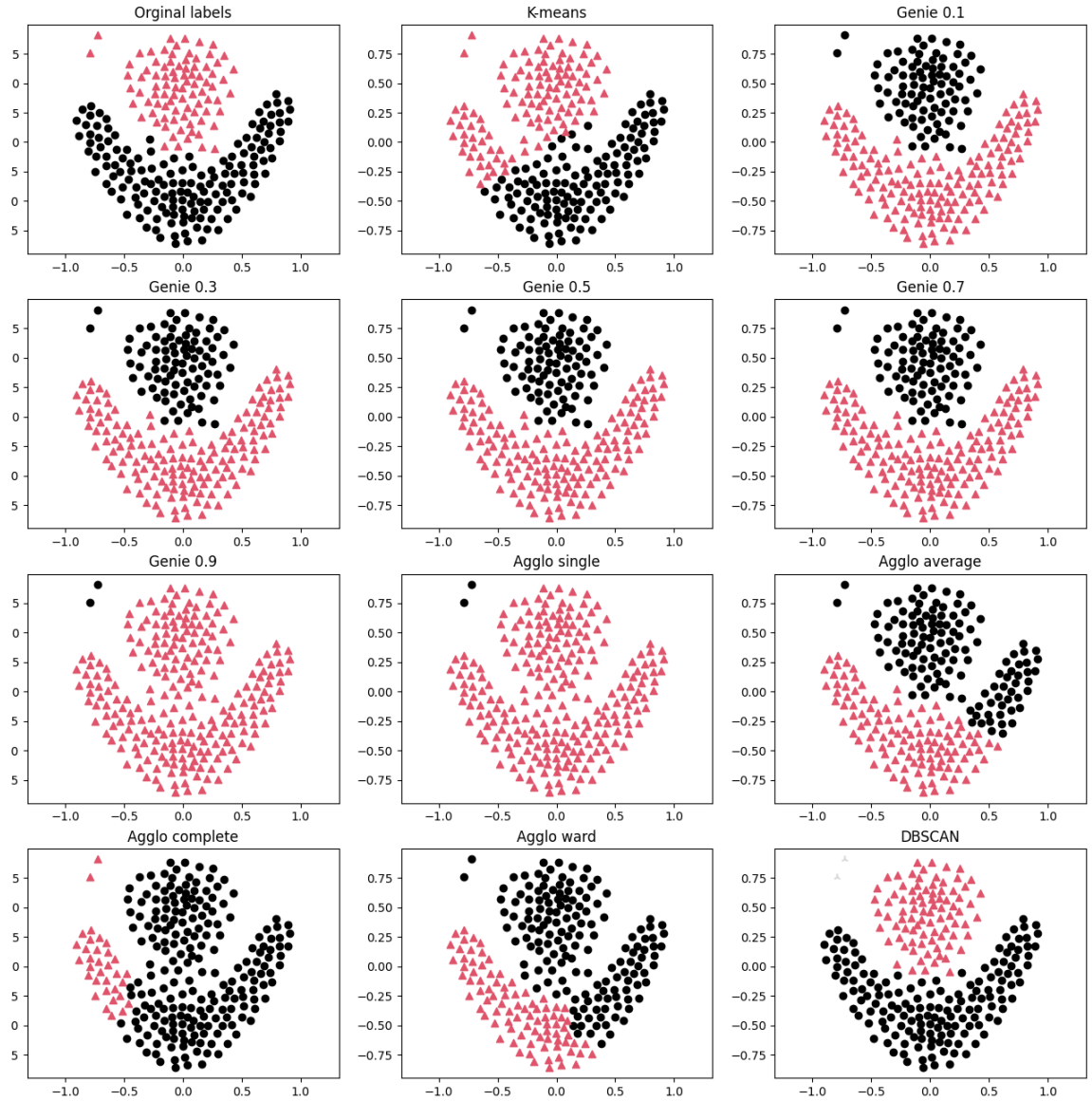




flame

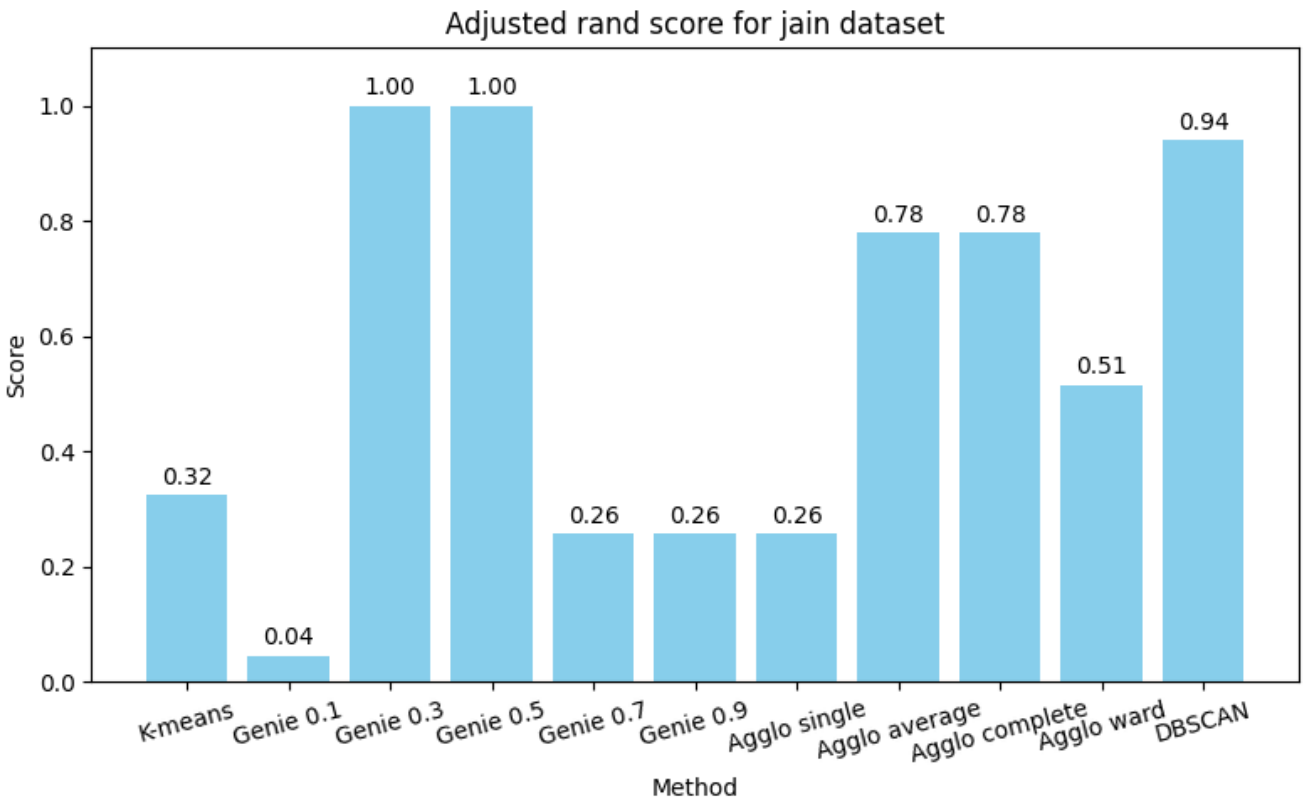
- Best results were given by Genie and DBSCAN
- DBSCAN also is able to detect data that look like outliers

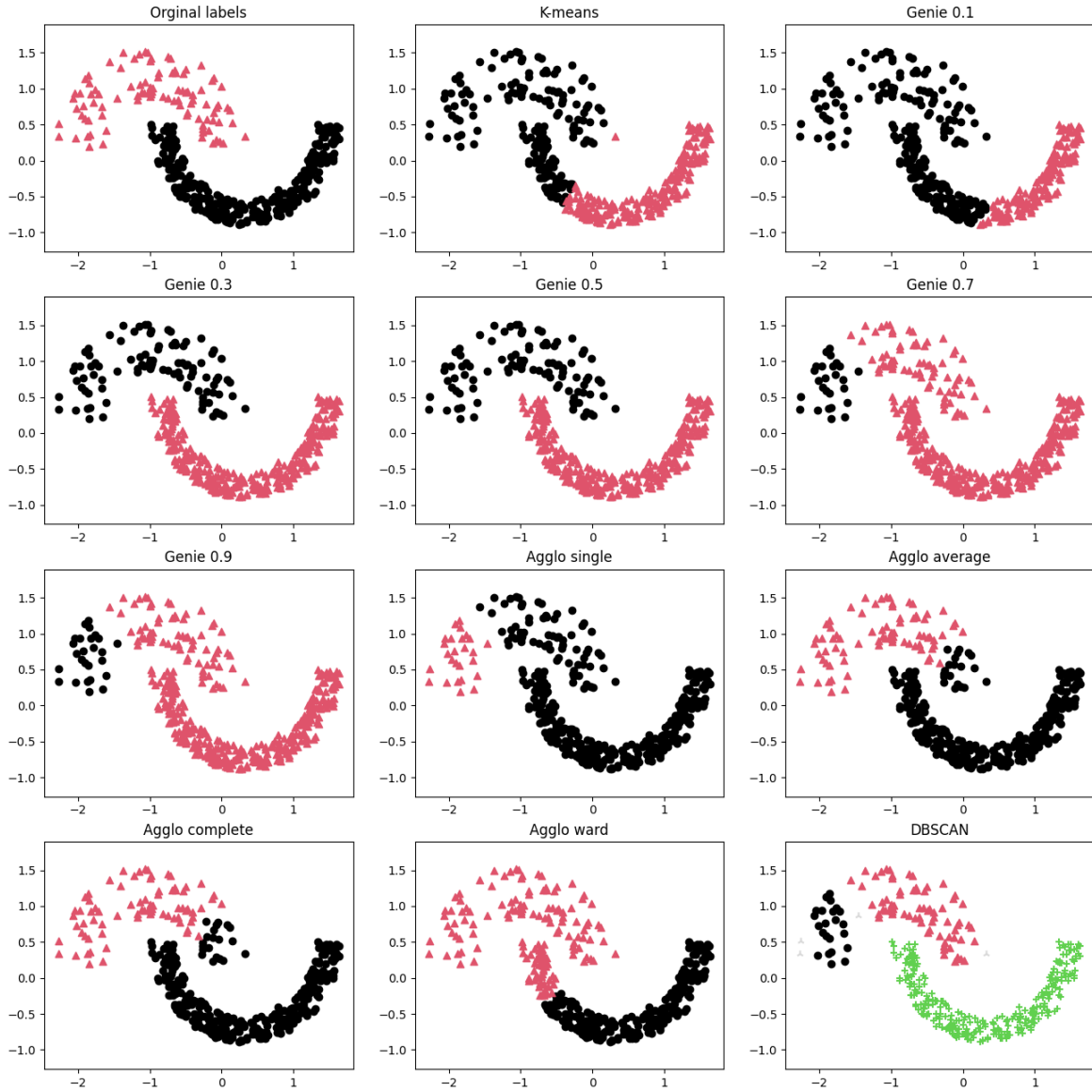




jain

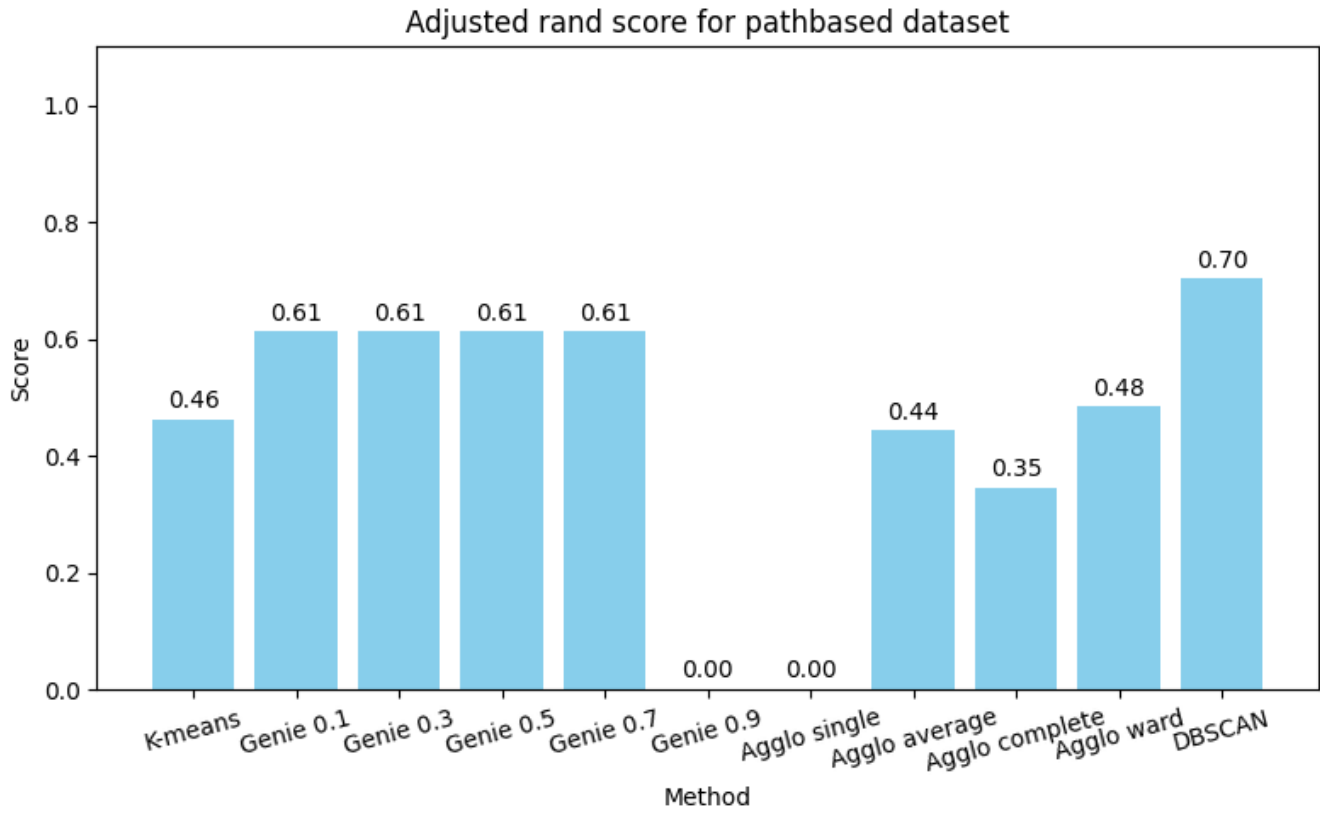
- Genie with smaller g were providing best clustering

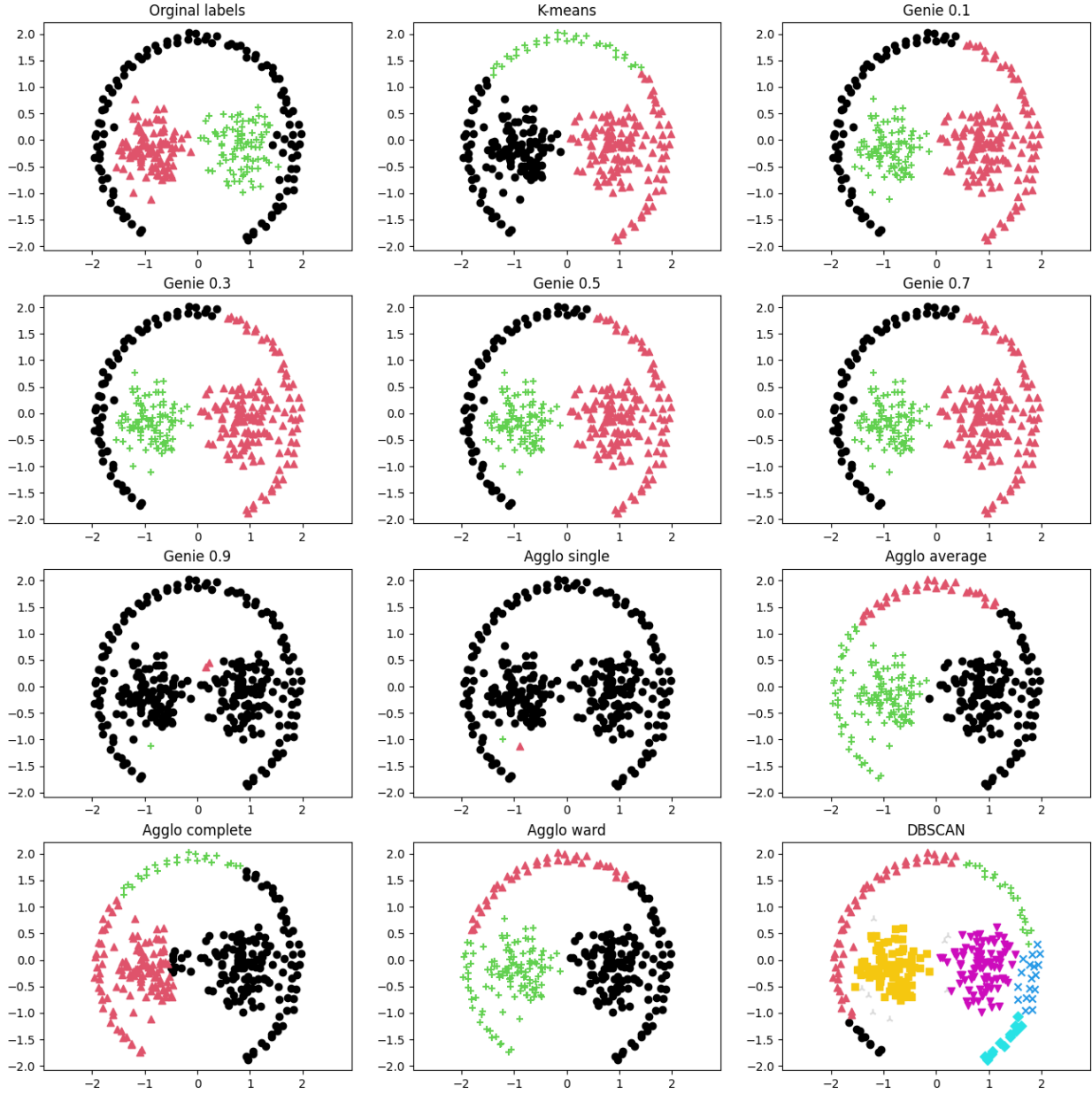




pathbased

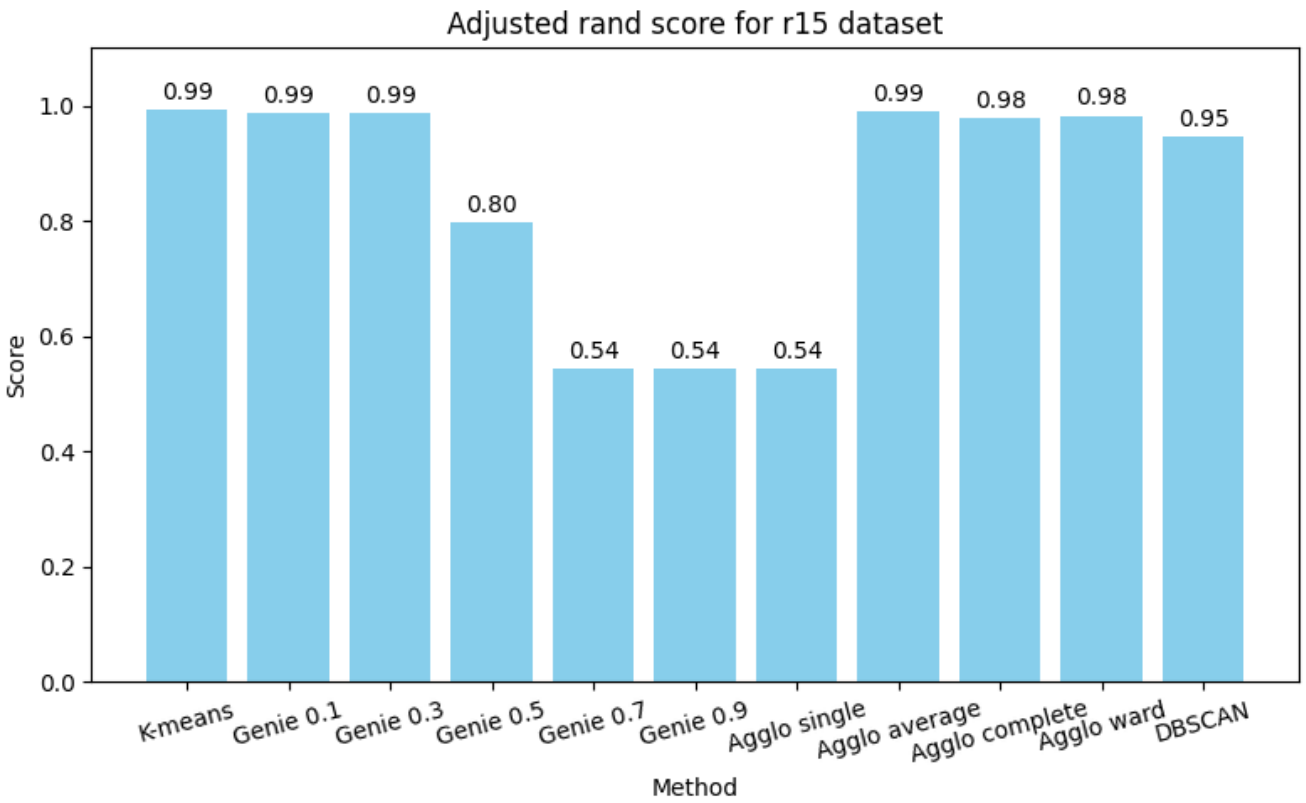
- None of the algorithms were able to obtain results like in original dataset

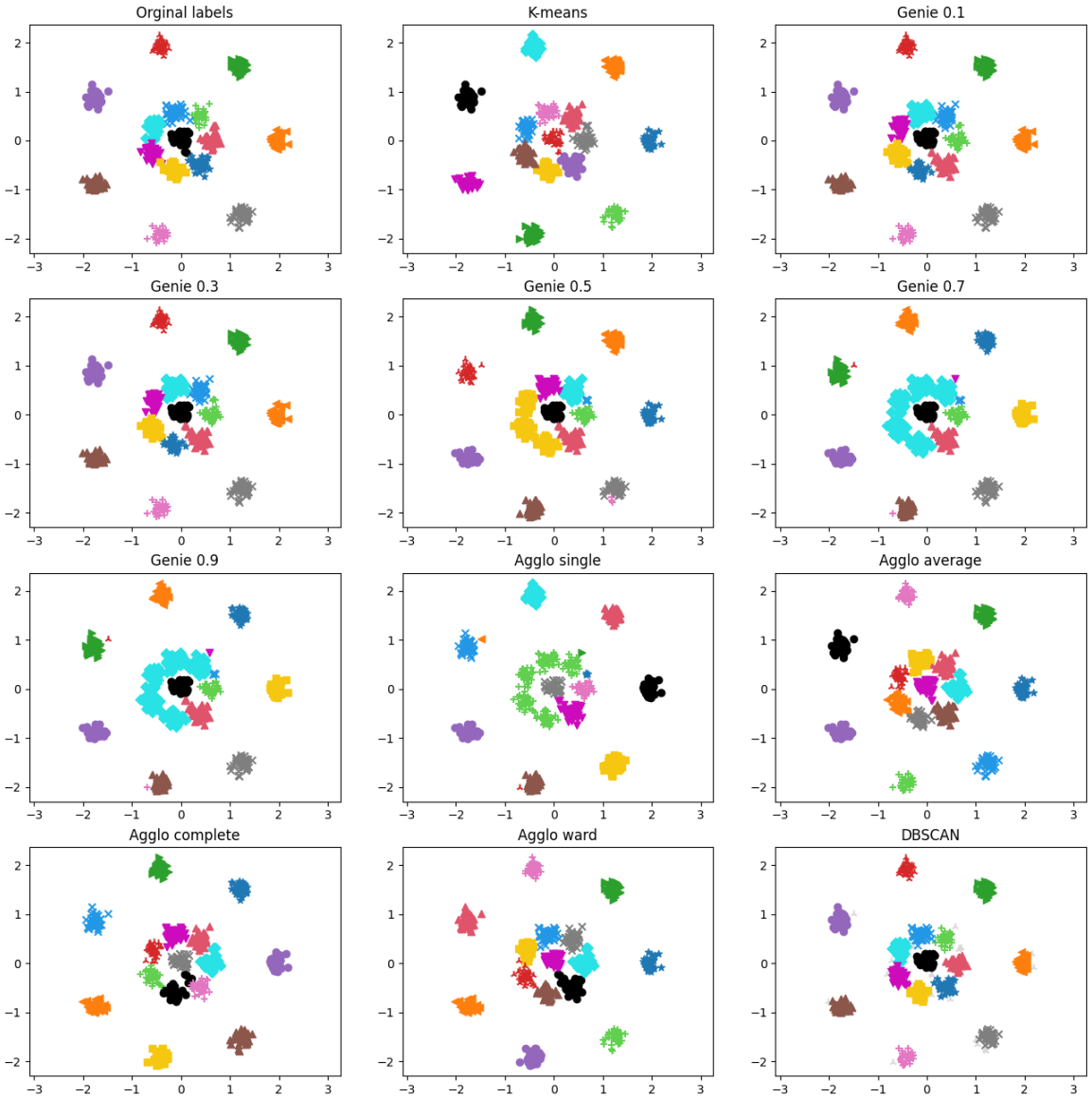




r15

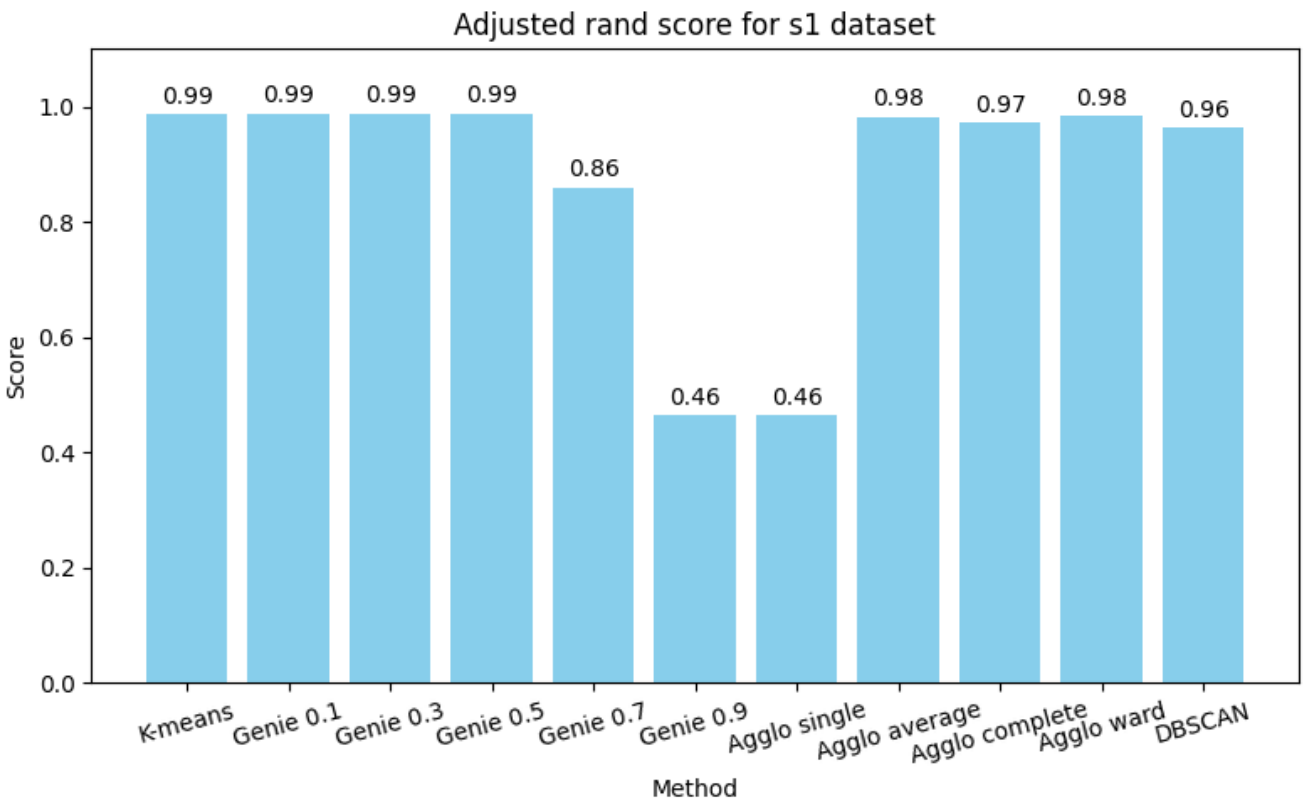
- Goods fit for majority of the models.
- Genie with higher g and AHC with single linkage

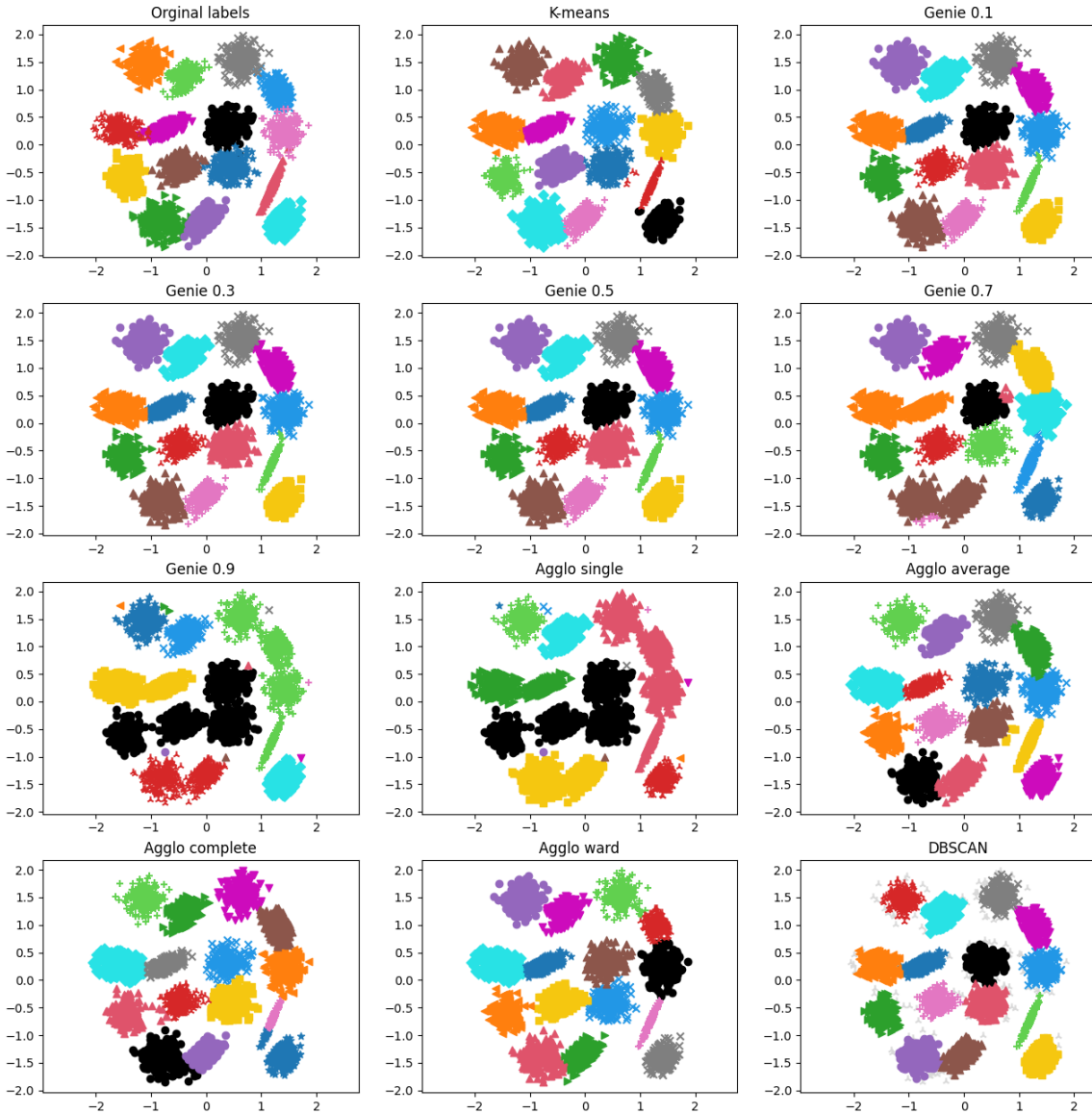




s1

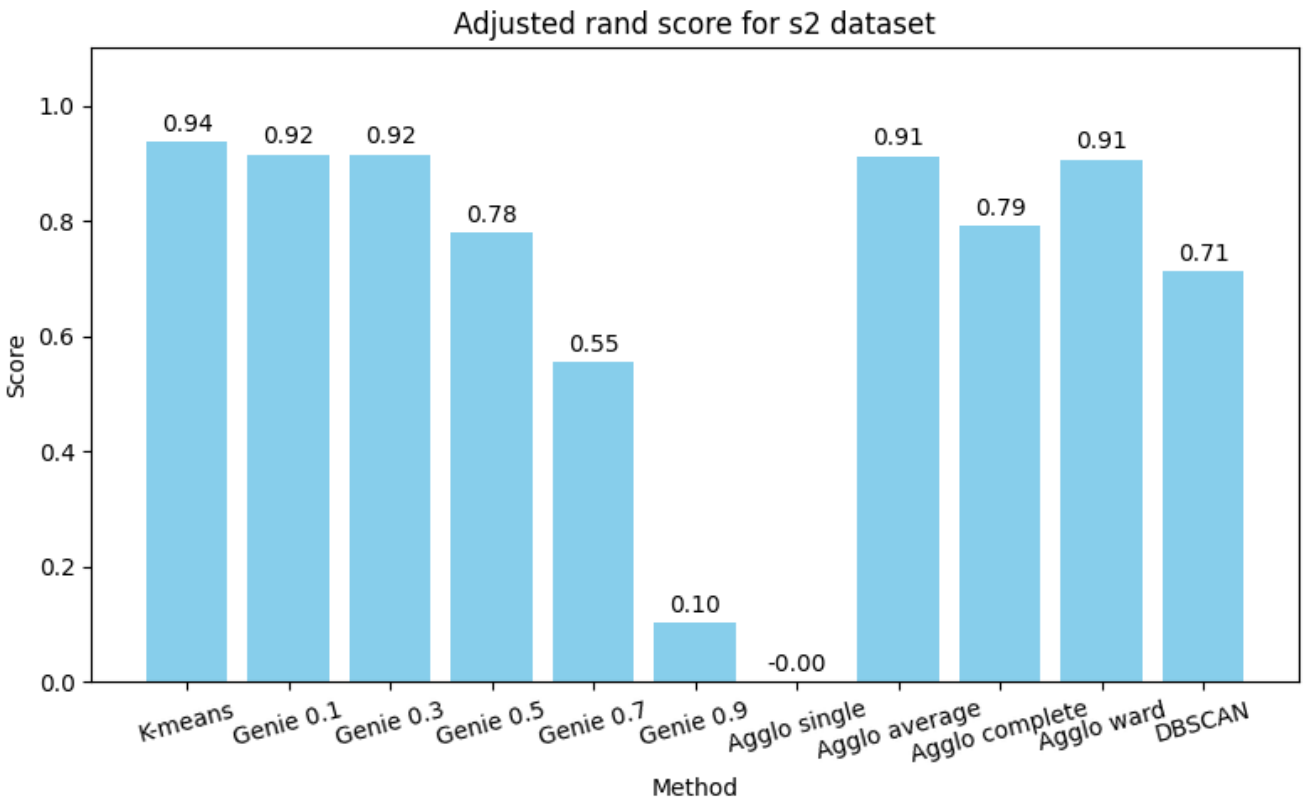
- Only Genie with higher g and AHC with single linkage is unable to fit the data

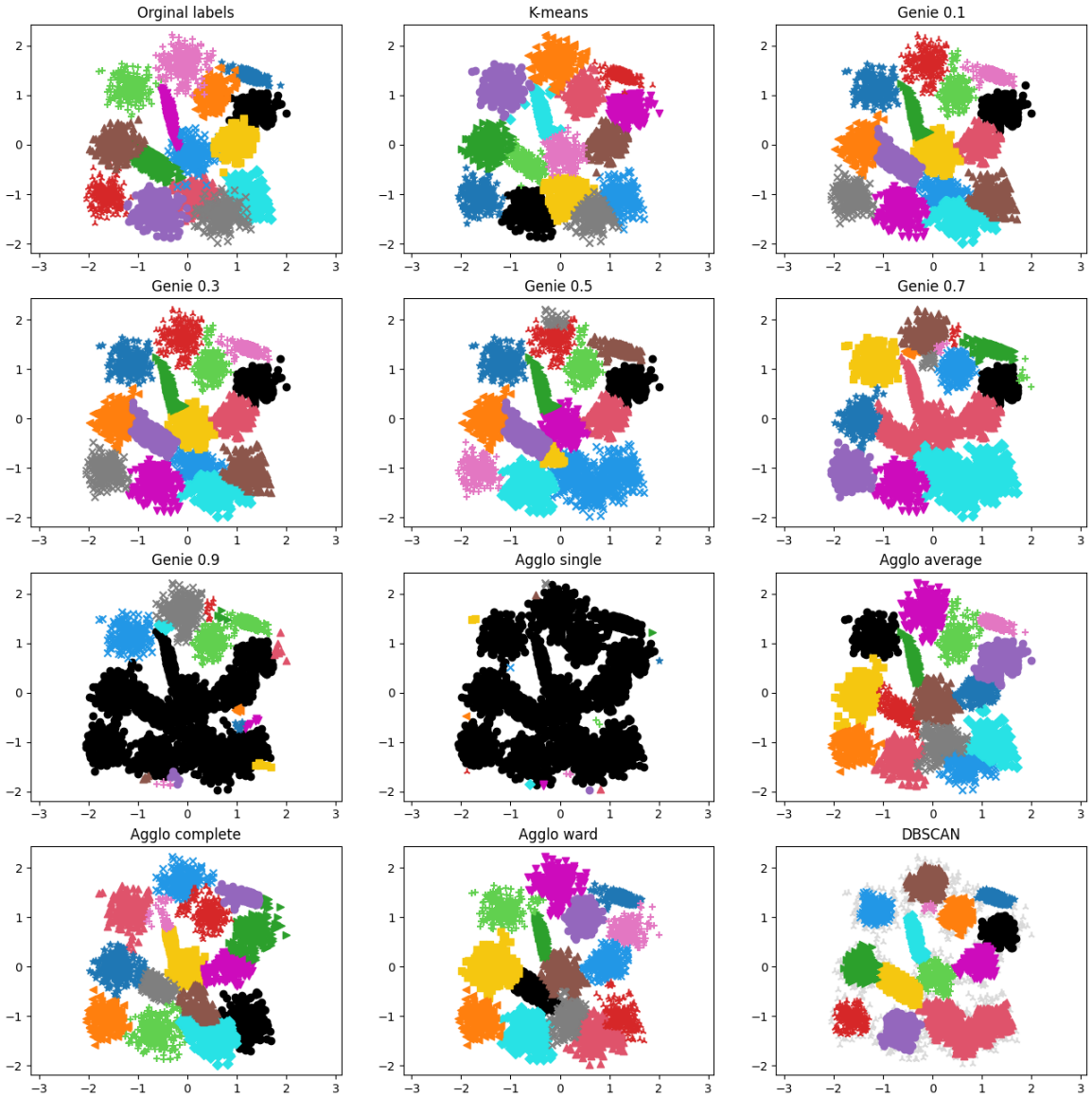




s2

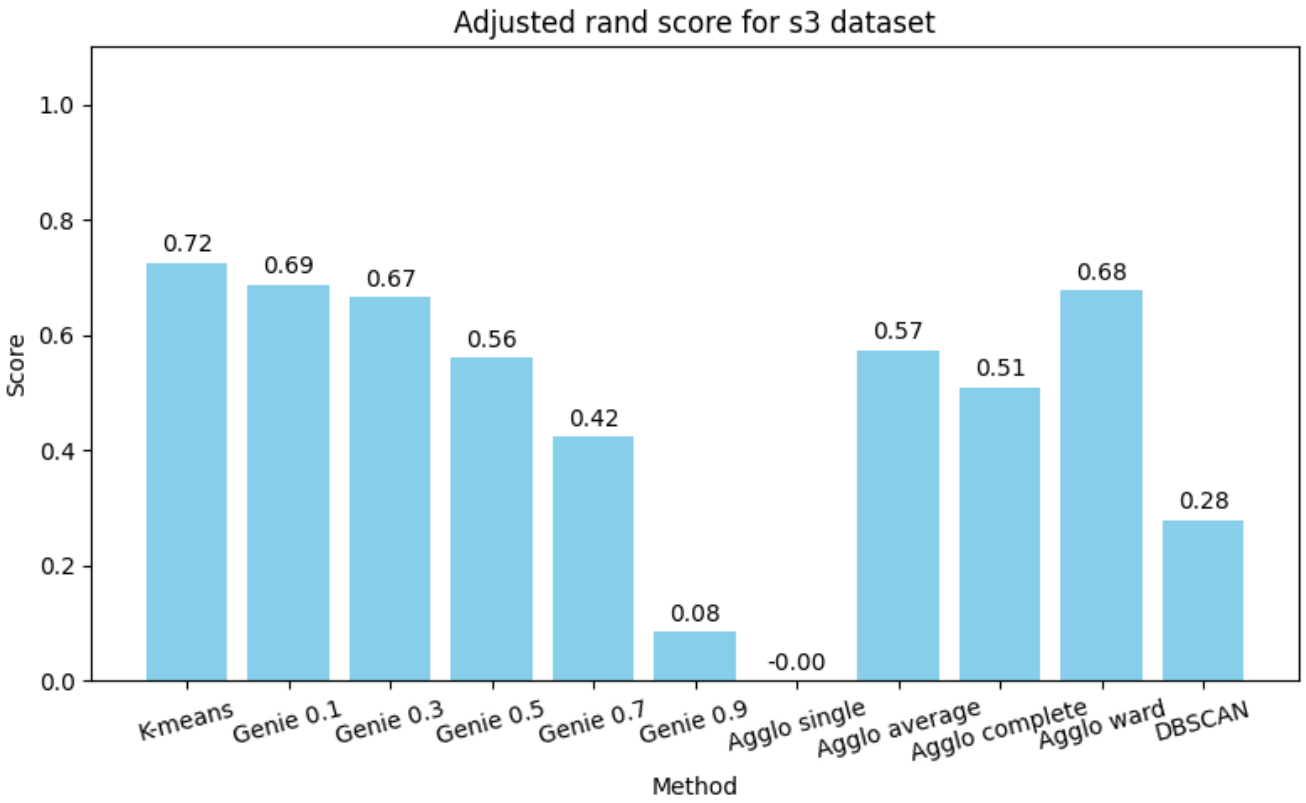
- With more overlapped data it is harder to find good clusters to AHC with single linkage
- The best naturally are methods with balancing capabilities

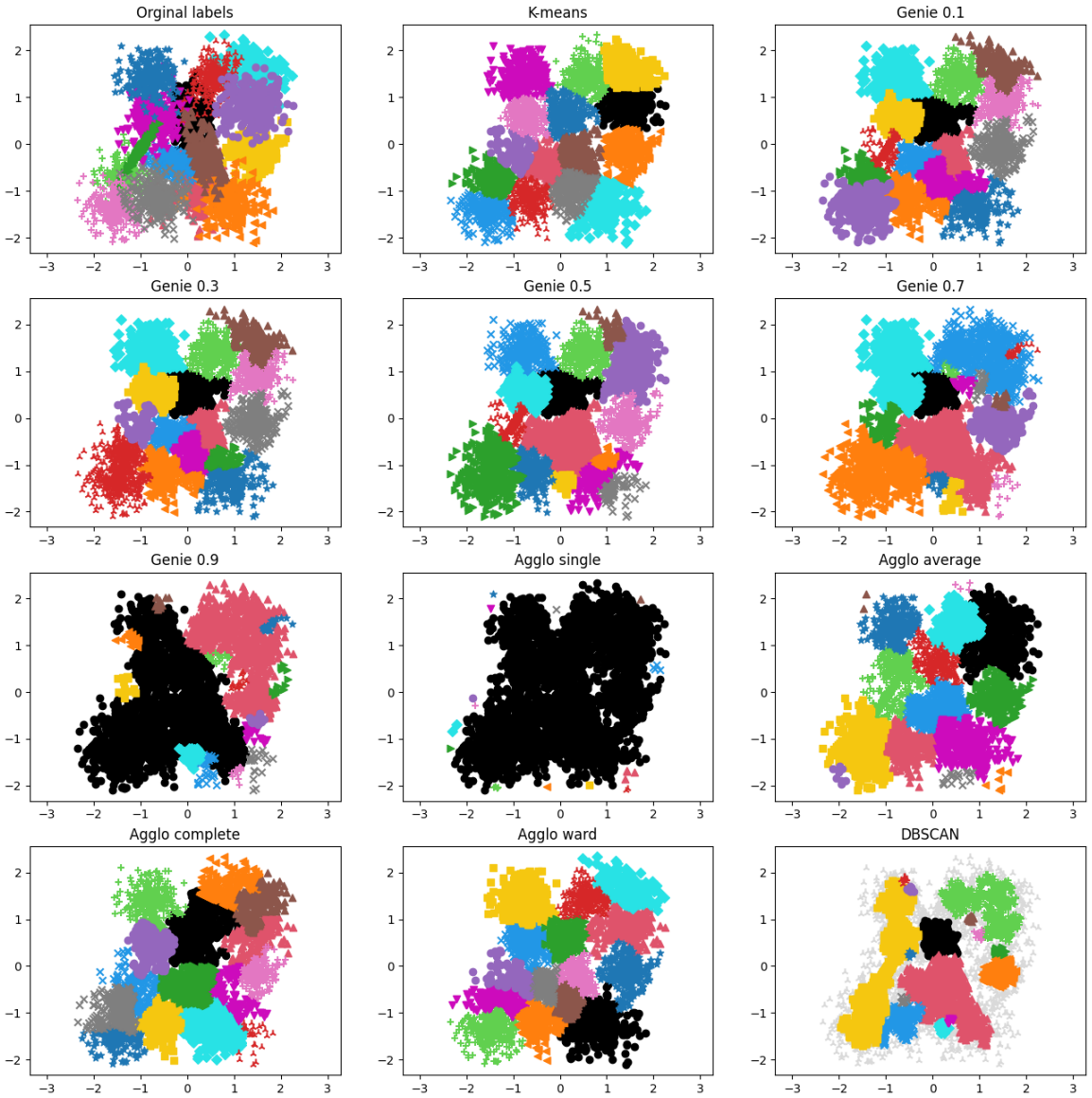




s3

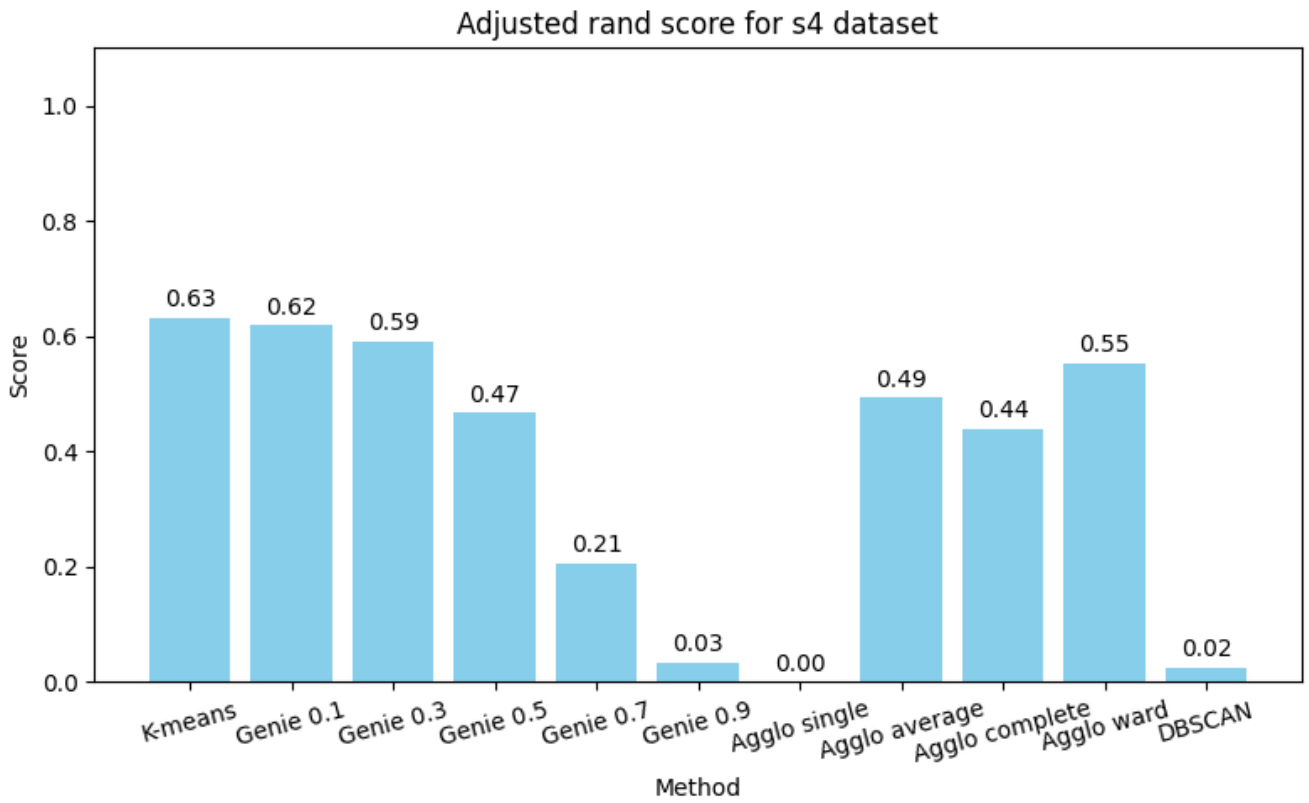
- Overlaping clusters make it harder to get a good adjusted rank score

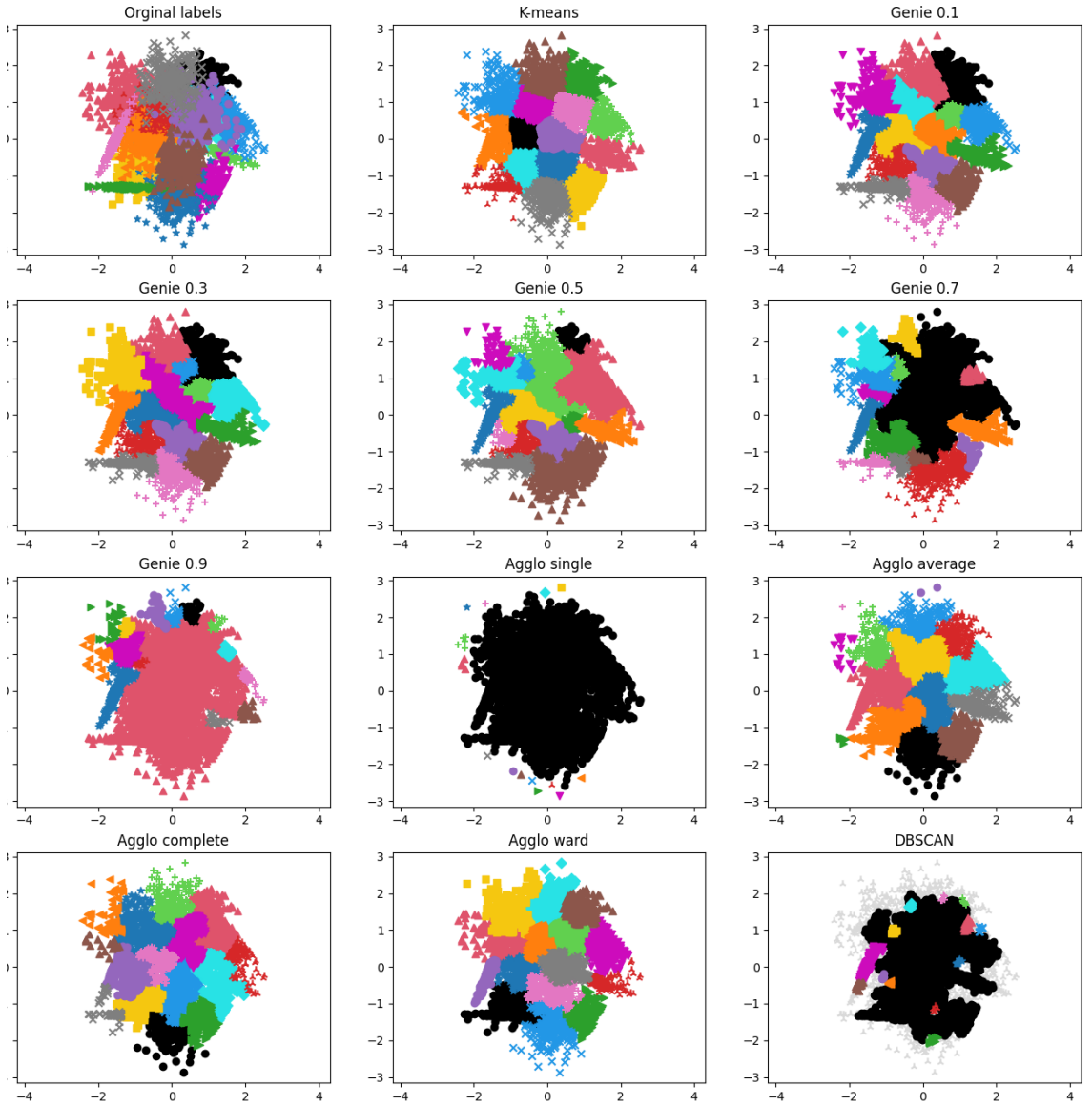




s4

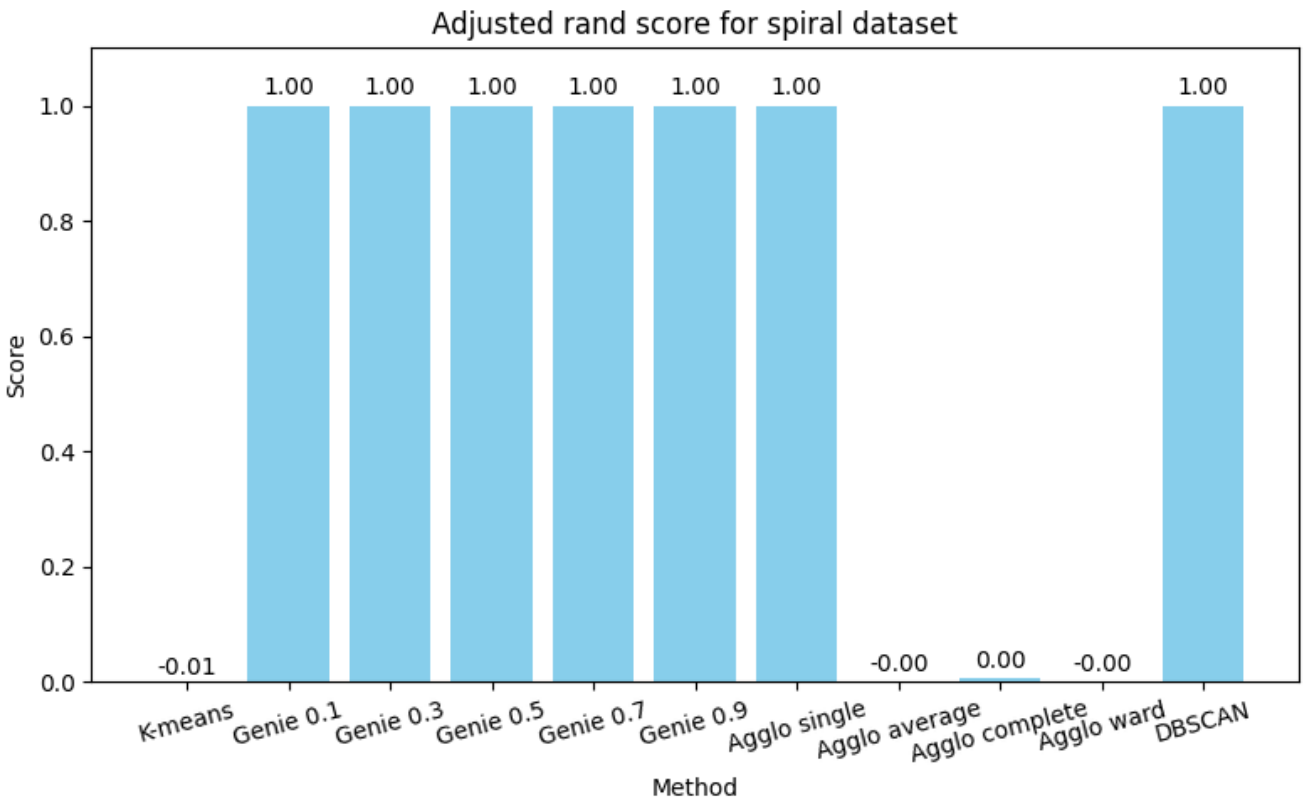
- It is very hard in DBSCAN to find right parameters, because of density of the data

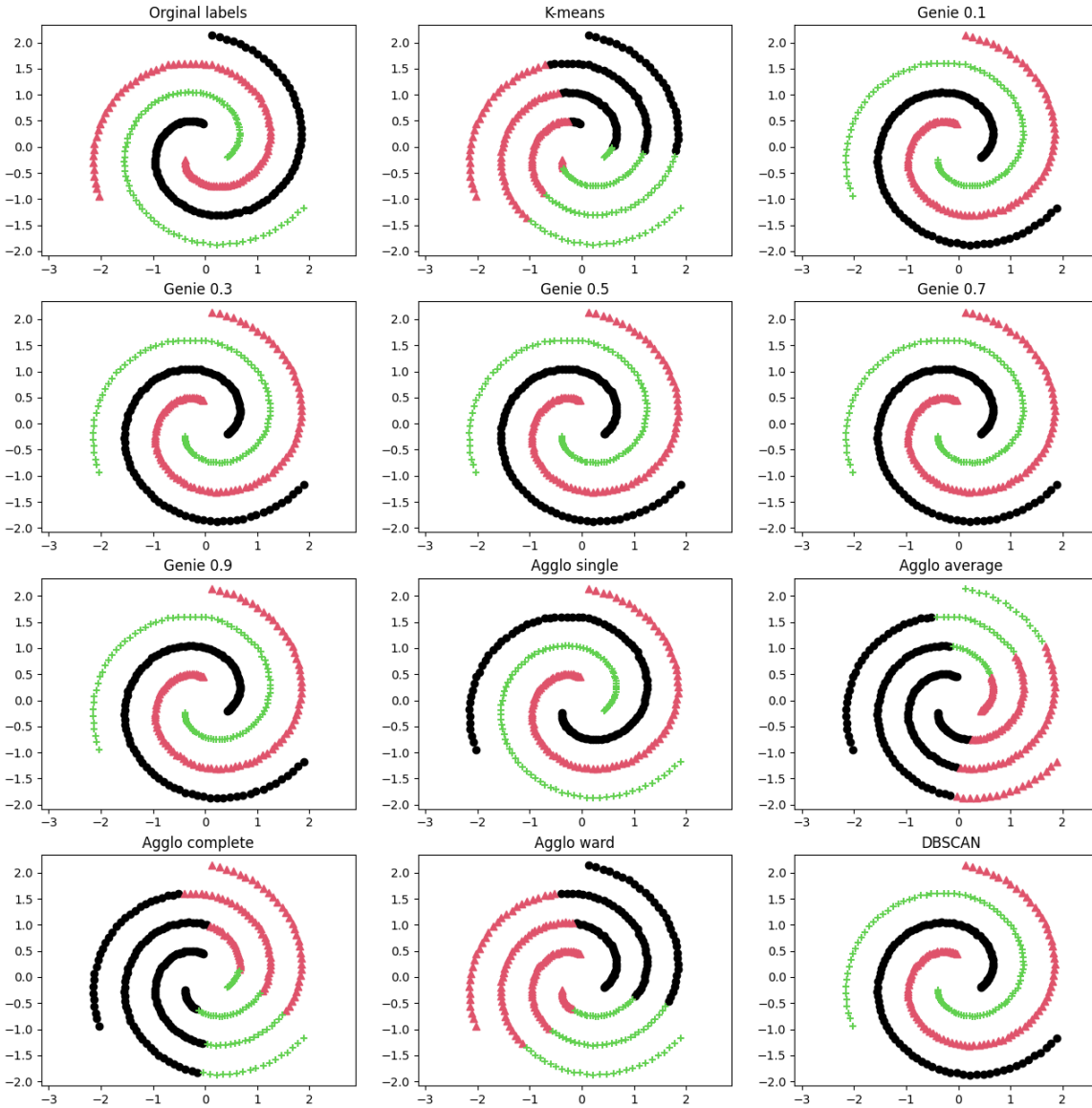




spiral

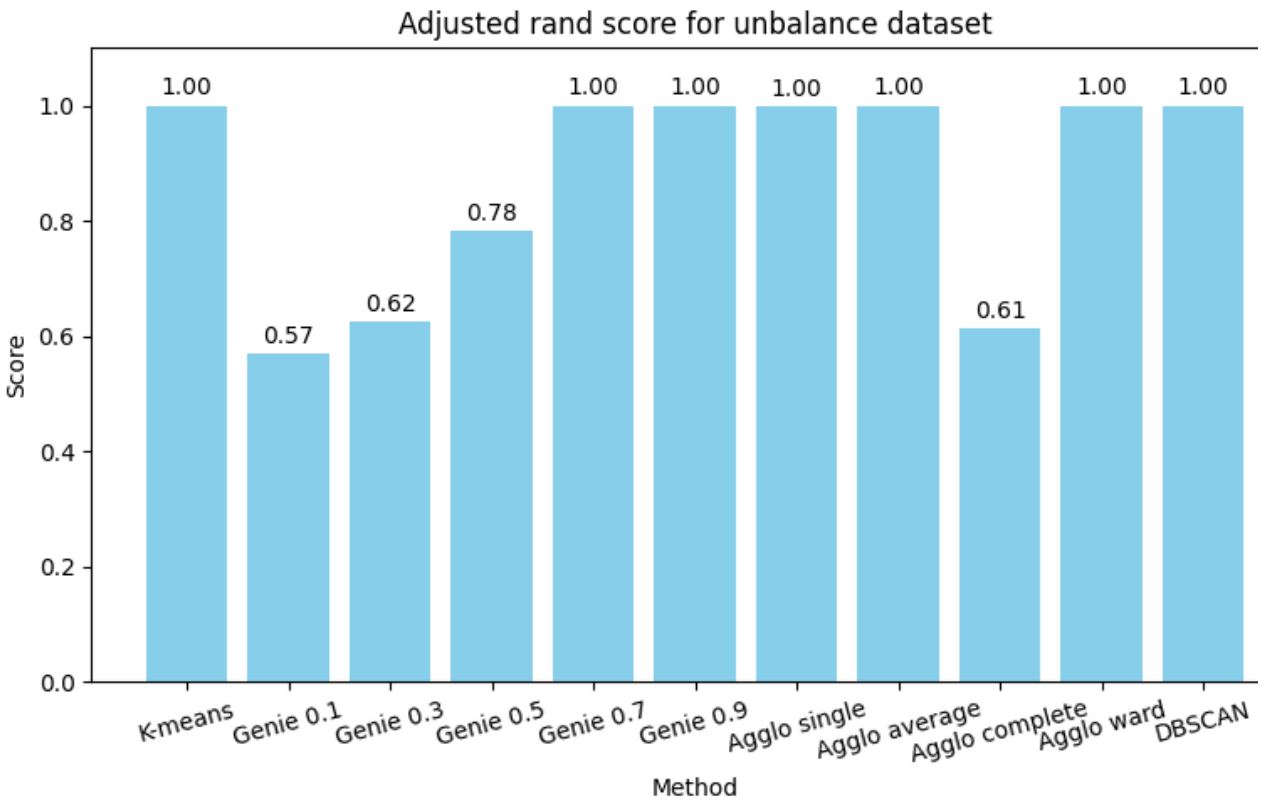
- Depending on the algorithm it either get it perfectly or performs poorly

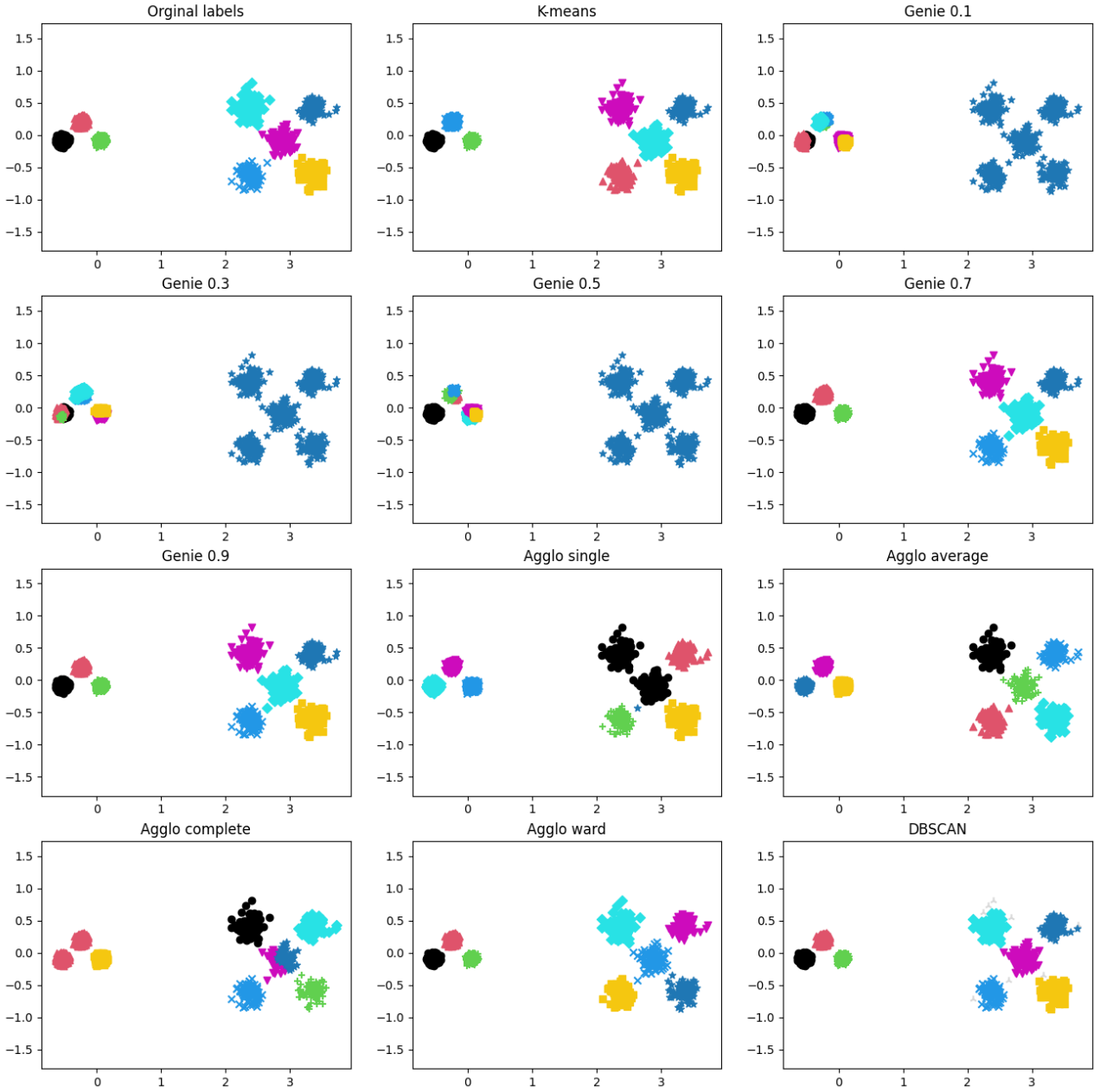




unbalance

- Unlike in r15 that looks to be similar problem Genie with higher g gets better results





References

- [Gagolewski, M., A framework for benchmarking clustering algorithms, SoftwareX 20, 101270, 2022, DOI:10.1016/j.softx.2022.101270, URL:https://clustering-benchmarks.gagolewski.com/
- Charu C. Aggarwal, *Data Mining: The Textbook*, Springer, 2015.
- Bolón-Canedo et. al. 2012

Thank you