Applied Property Testing For Drug Repurposing

Name: Omer Mualem Advisor: Dr. Sarel Cohen

Workshop Track: Research Project Number: 15006801

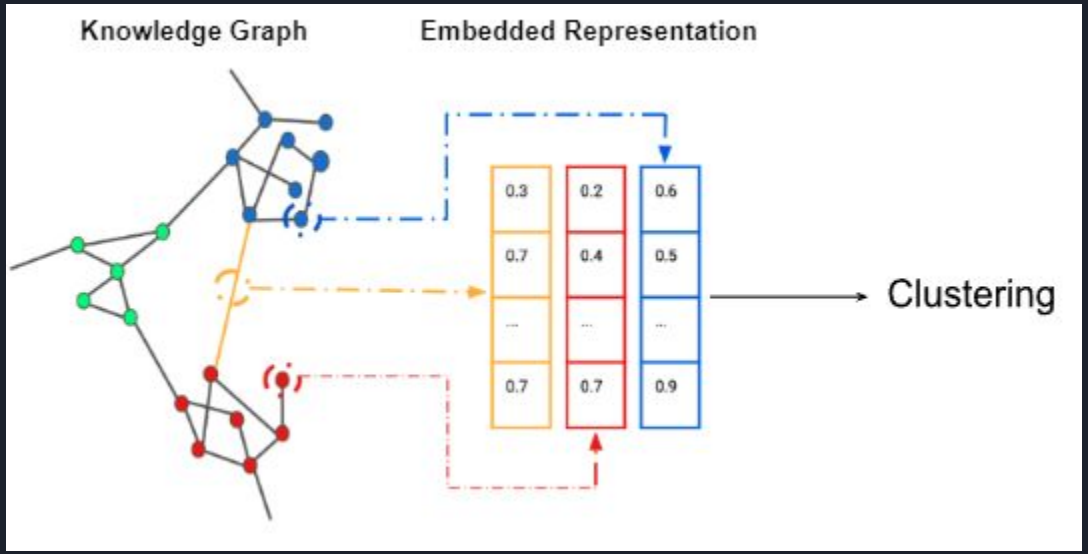
BackGround

- In the face of a global viral outbreak, the immediate need for effective treatments becomes paramount
- My project aims to circumvent these challenges by leveraging existing FDA-approved drugs through a process known as drug repurposing
- Test a new clustering algorithm

Solution

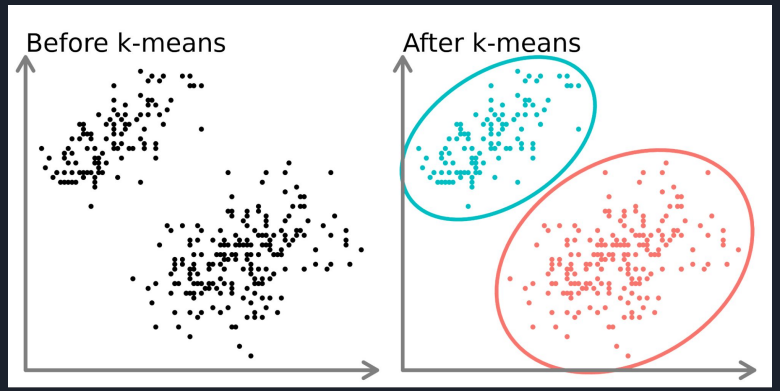
- Relying on recent work on drug repurposing by Cohen et al[PLOS ONE 2023]
- Change clustering algorithm used
- Test for improved result

Drug Repurposing Model Architecture



Current Clustering Algorithm

- K-means
- Minimize the variance within each cluster and maximize the variance between clusters



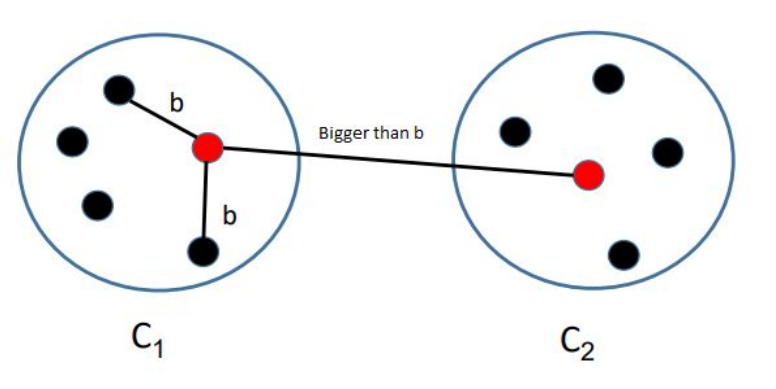
New Clustering Algorithm

- first suggested in a paper by Michal Pernas, Dar, Ron and Alon in 2003
- Property testing
- Utilize the power of randomness

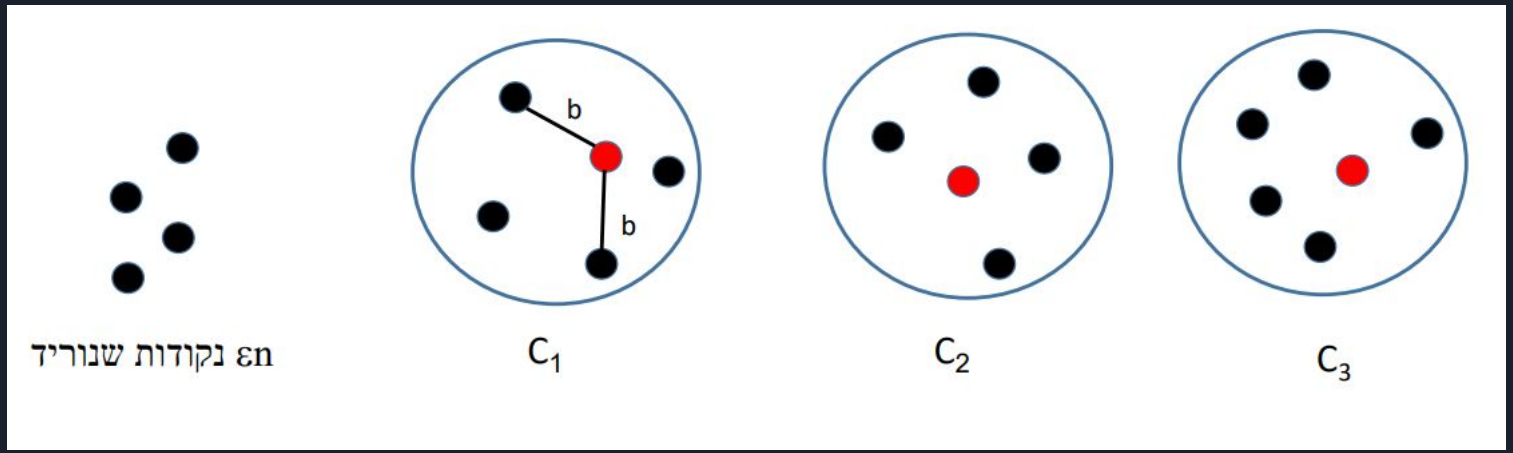
Parameters

- In the new algorithm we introduce new parameters
  - epsilon - outlier percentage parameter
  - b - cluster max radius
  - k - max number of clusters

Parameters Meaning



Parameters Meaning



New Clustering Algorithm

Input: Set of points X, parameters k, b and a outlier parmeter epsilon

Output: Is X (k,b)-clusterable or epsilon-far from being (k,2b)-clusterable

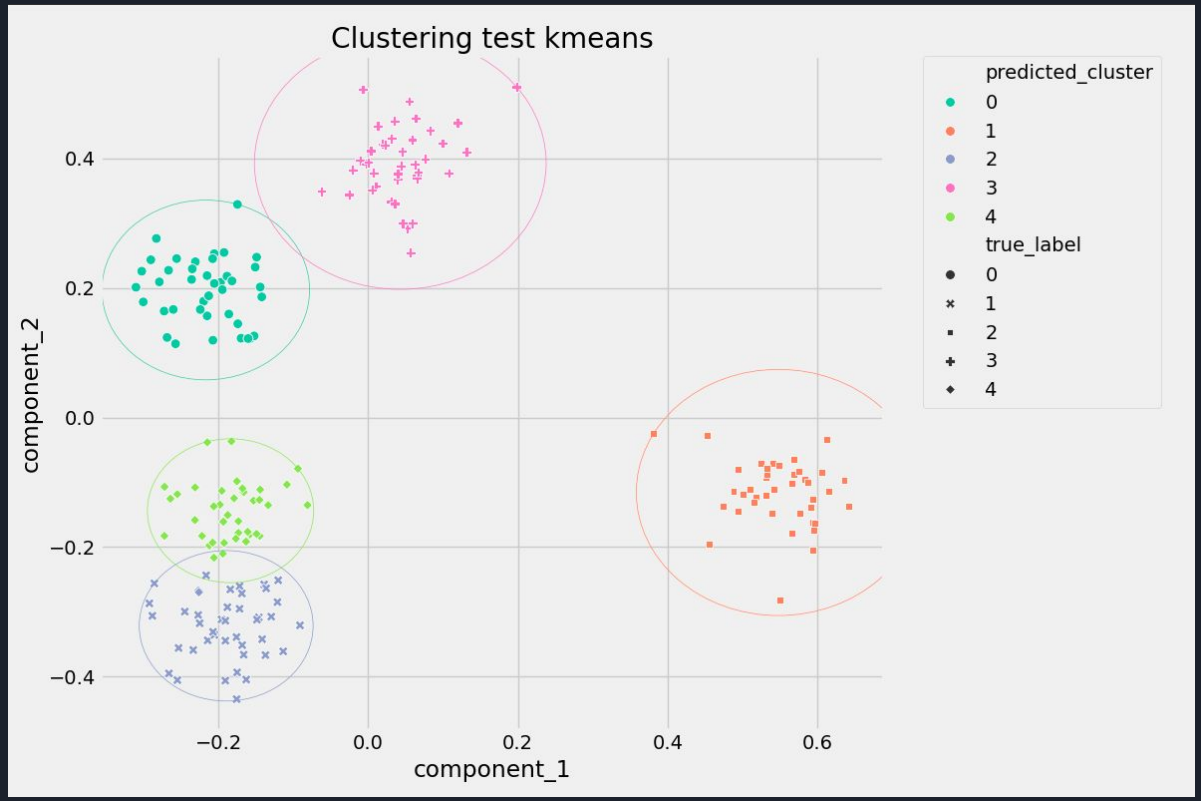
Steps:

1. Let Rep_1 be a random points from X
2. i = 1, Find_New_Rep = True
3. For i < k + 1 and Find_New_Rep = True:
    3.1. Choose ln(3k) / epsilon points from a uniform distribution on X
    3.2. If a point x in the sample has > b distance from all current representors
        then Rep_i+1 = x and i = i + 1
    3.3. Else Find_New_Rep = False
4. If i <= k we accept the representors and parameters, else we decline

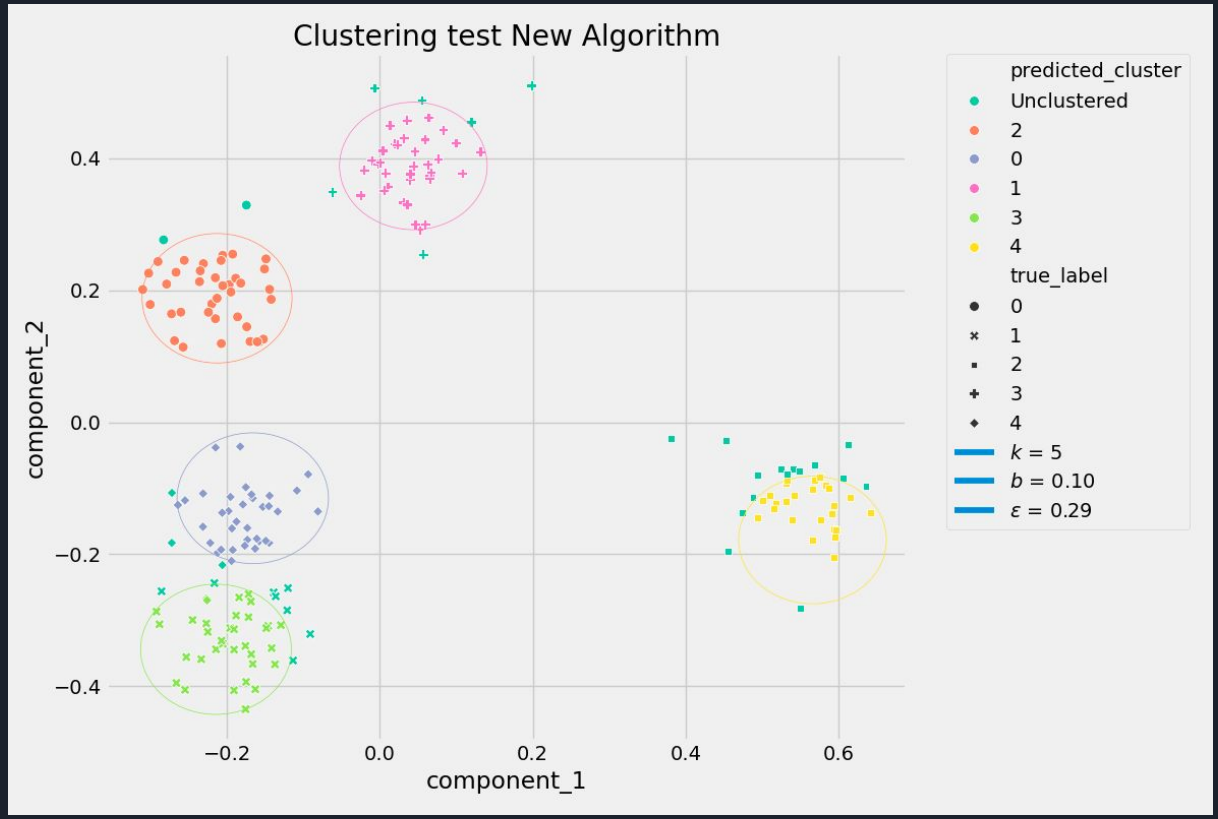
Current Progress

- Created a general pipeline for clustering algorithms using tensorboard to save outputs
- Tested the pipeline with the k-means algorithm on synthetic data
- Implemented the new clustering algorithm
- Tested the new clustering algorithm on synthetic data
- Tested the new clustering algorithm on SIFT 1M dataset training set including 100,000 samples with dimension 128

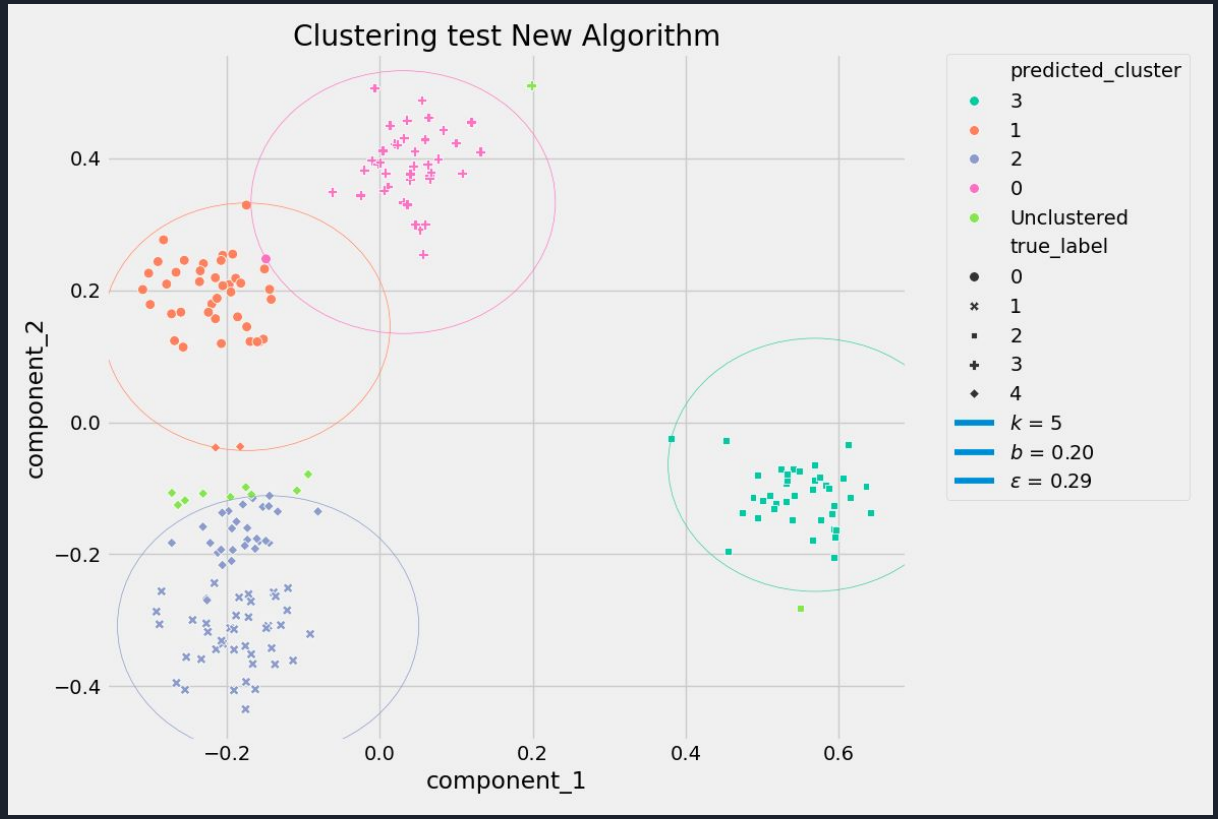
K-means Example On Synthetic Data



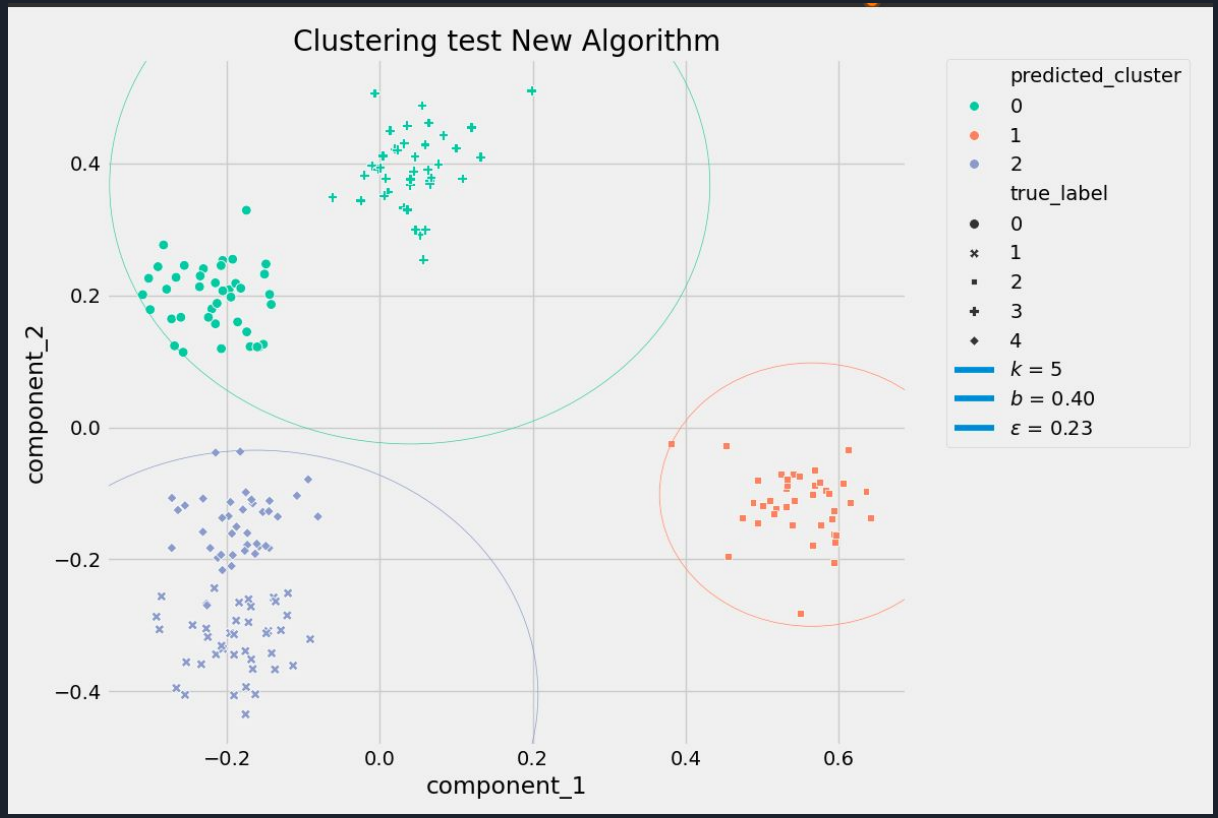
New Algorithm Example On Synthetic Data



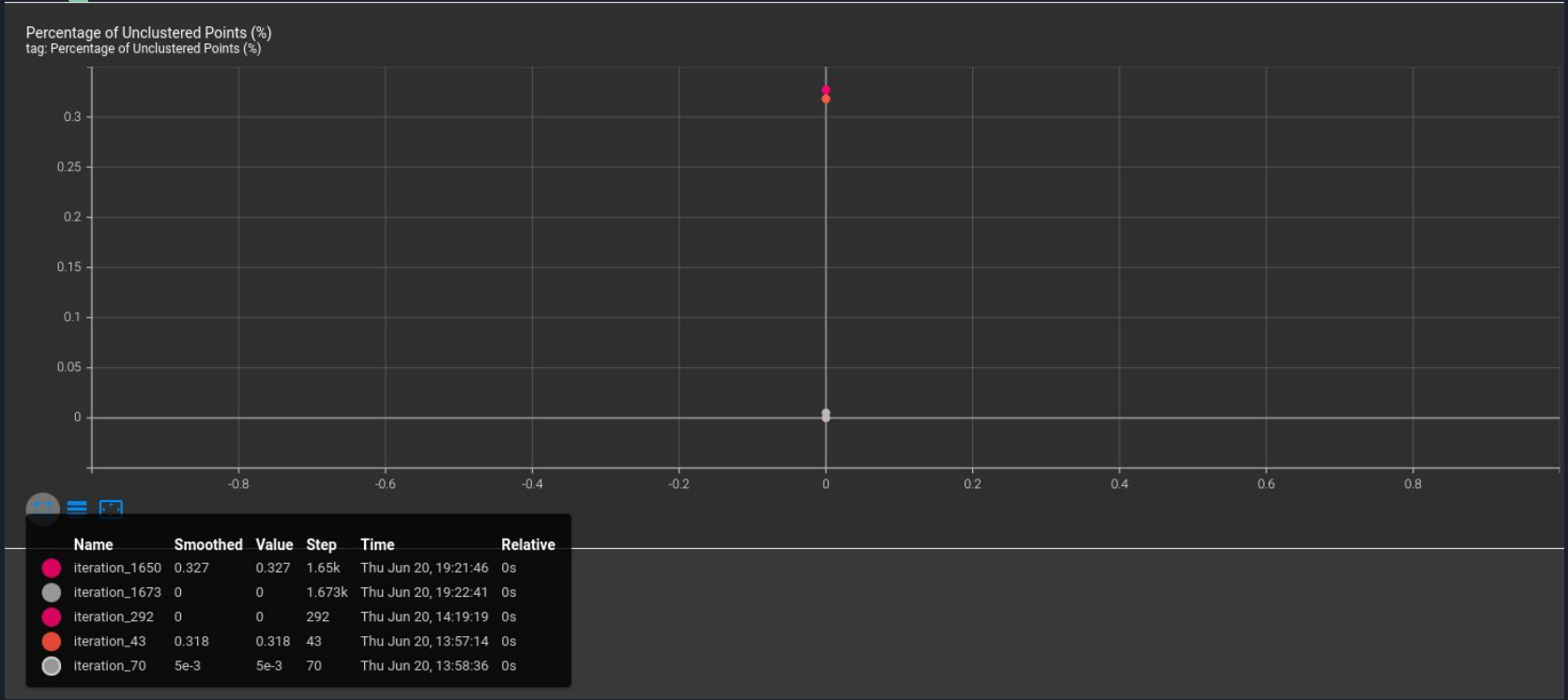
New Algorithm Example On Synthetic Data



New Algorithm Example On Synthetic Data

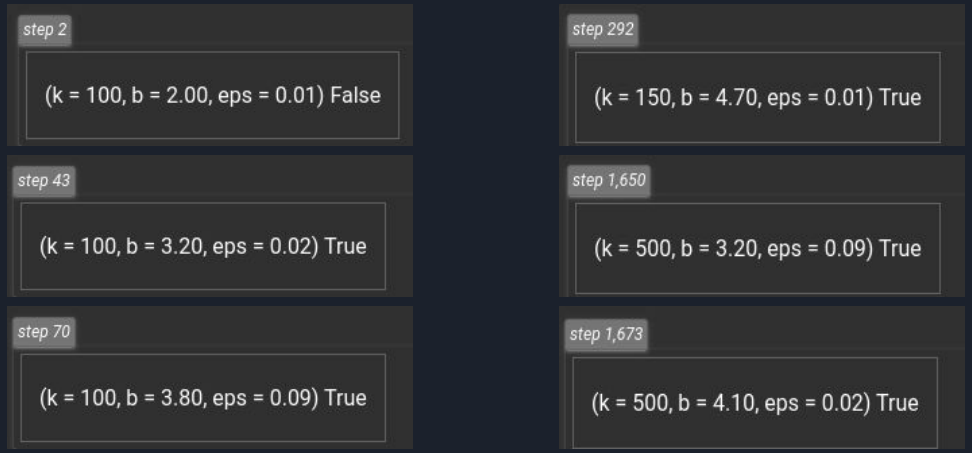


New Algorithm On SIFT Dataset



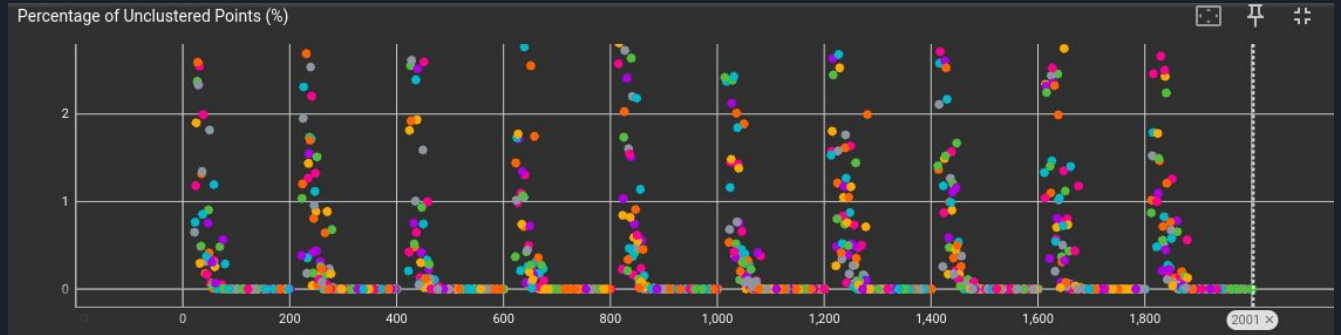
Name	Smoothed	Value	Step	Time	Relative
iteration_1650	0.327	0.327	1.65k	Thu Jun 20, 19:21:46	0s
iteration_1673	0	0	1.673k	Thu Jun 20, 19:22:41	0s
iteration_292	0	0	292	Thu Jun 20, 14:19:19	0s
iteration_43	0.318	0.318	43	Thu Jun 20, 13:57:14	0s
iteration_70	5e-3	5e-3	70	Thu Jun 20, 13:58:36	0s

New Algorithm On SIFT Dataset



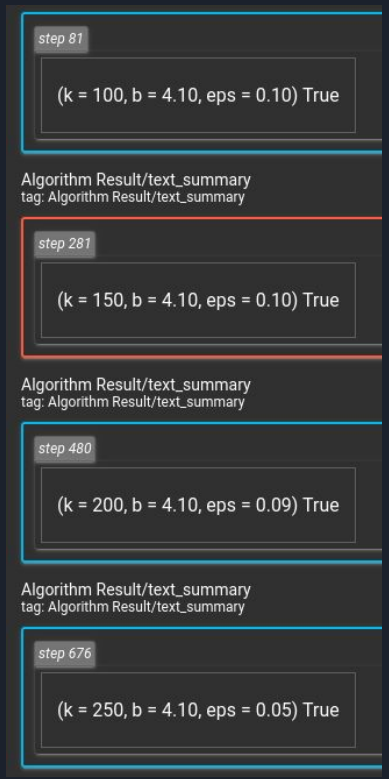
Name	Smoothed	Value	Step	Time	Relative
iteration_1650	0.327	0.327	1.65k	Thu Jun 20, 19:21:46	0s
iteration_1673	0	0	1.673k	Thu Jun 20, 19:22:41	0s
iteration_292	0	0	292	Thu Jun 20, 14:19:19	0s
iteration_43	0.318	0.318	43	Thu Jun 20, 13:57:14	0s
iteration_70	5e-3	5e-3	70	Thu Jun 20, 13:58:36	0s

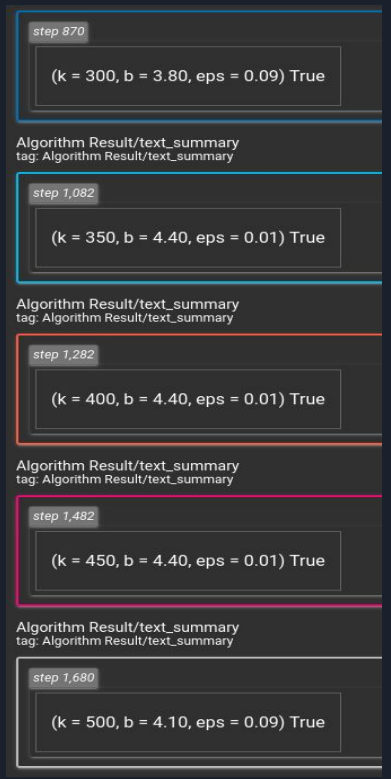
New Algorithm On SIFT Dataset

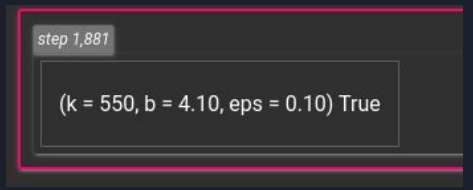


Run ↑	Smoothed	Value	Step	Relative
iteration_100	0	0	100	0
iteration_1000	0	0	1,000	0
iteration_1001	0	0	1,001	0
iteration_101	0	0	101	0
iteration_1014	2.419	2.419	1,014	0
iteration_1015	3.534	3.534	1,015	0
iteration_1016	3.303	3.303	1,016	0
iteration_1017	2.372	2.372	1,017	0
iteration_1018	7.231	7.231	1,018	0
iteration_1019	3.21	3.21	1,019	0
iteration_102	0	0	102	0
iteration_1020	3.808	3.808	1,020	0
iteration_1021	5.16	5.16	1,021	0
iteration_1022	0.533	0.533	1,022	0

New Algorithm On SIFT Dataset







Conclusions And Moving Forward

- Testing the new algorithm provides us with deeper insights into the data structure
- If the drug repurposing dataset is structured as clusters of circles we can possibly improve current clustering by using a different technique
- Next we need to test our clustering on SIFT 1M test dataset
- Finally use the new algorithm on our drug repurposing problem