INFERRING GENE FUNCTIONALITY WITH CLUSTERING

By Juacall Bernard
Advisor: Miao,Heh

OUTLINE

- Introduction
- Goal and Objectives
- Clustering Overview
- Algorithms & Implementation
- Result Evaluation
- Conclusion

INTRODUCTION

- Viewing and analyzing vast amounts of biological data as a whole set can be difficult.

- An easier way to interpret the data is to partition the data set into clusters.

GOAL

- In this project I will attempt to infer gene functions and determine the type of a known or unknown gene by way of clustering.

OBJECTIVES

- Determine the optimal number of clusters.
- Choose a better algorithm.

CLUSTERING REVIEW

What is clustering?

What is clustering useful for?

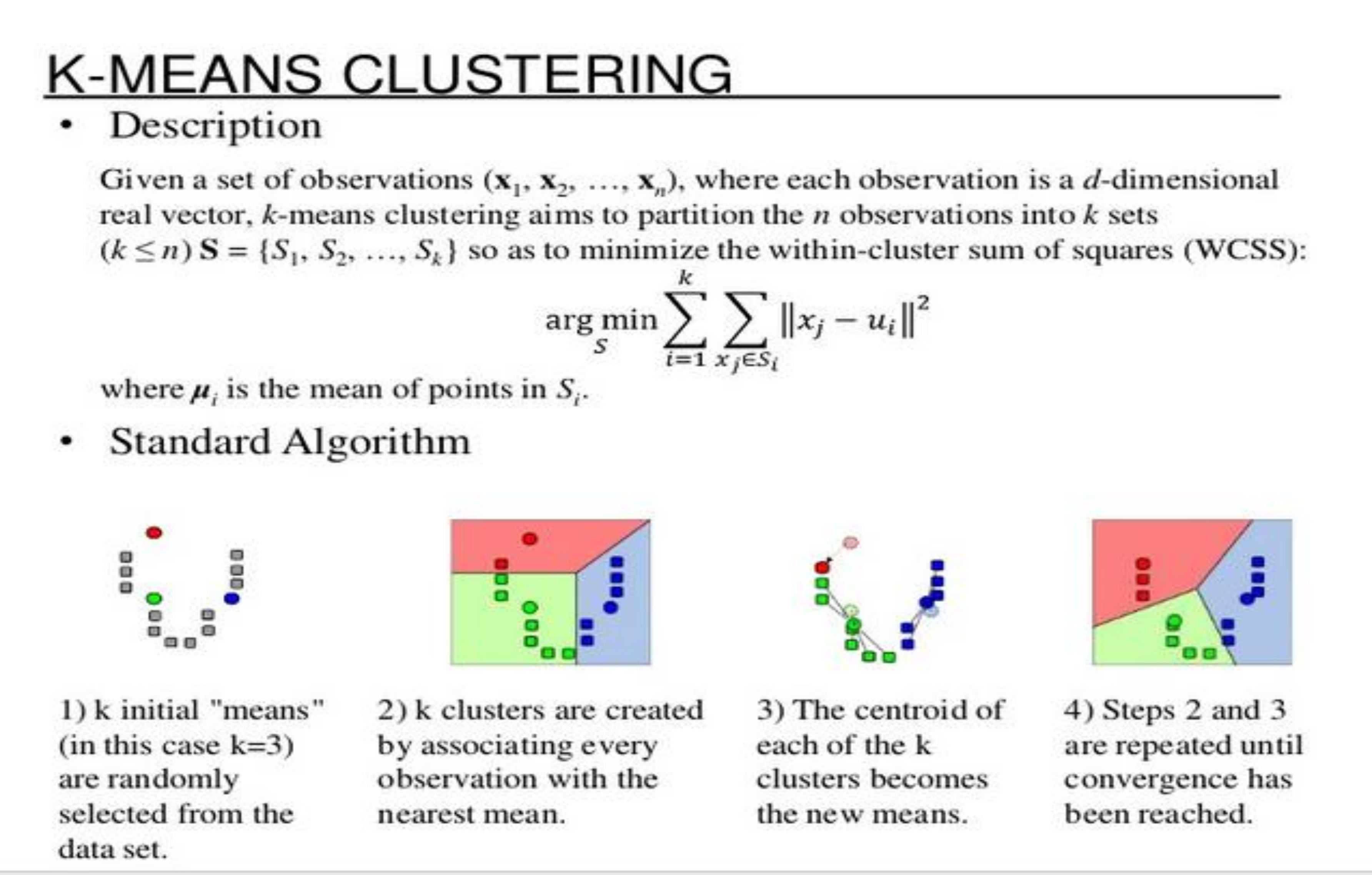
Are there any problems with clustering?

ALGORITHMS & IMPLEMENTATION

- K-Means Clustering
- Spectral Clustering

K-MEANS CLUSTERING

- How does it work?
- Input
- Disadvantages



K-MEANS CLUSTERING

- Description

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \leq n$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - u_i \right\|^2$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

- Standard Algorithm

1) k initial "means" (in this case k=3) are randomly selected from the data set.

2) k clusters are created by associating every observation with the nearest mean.

3) The centroid of each of the k clusters becomes the new means.

4) Steps 2 and 3 are repeated until convergence has been reached.

SIMILARITY MATRIX

- Ɛ-neighborhood graph
- K-nearest graph
- Fully connected graph

DISADVANTAGES

- Sensitive to outliers
- Fixed K values
- Less effective with non globular clusters

SPECTRAL CLUSTERING

- What is Spectral Clustering
- Laplacian Matrix

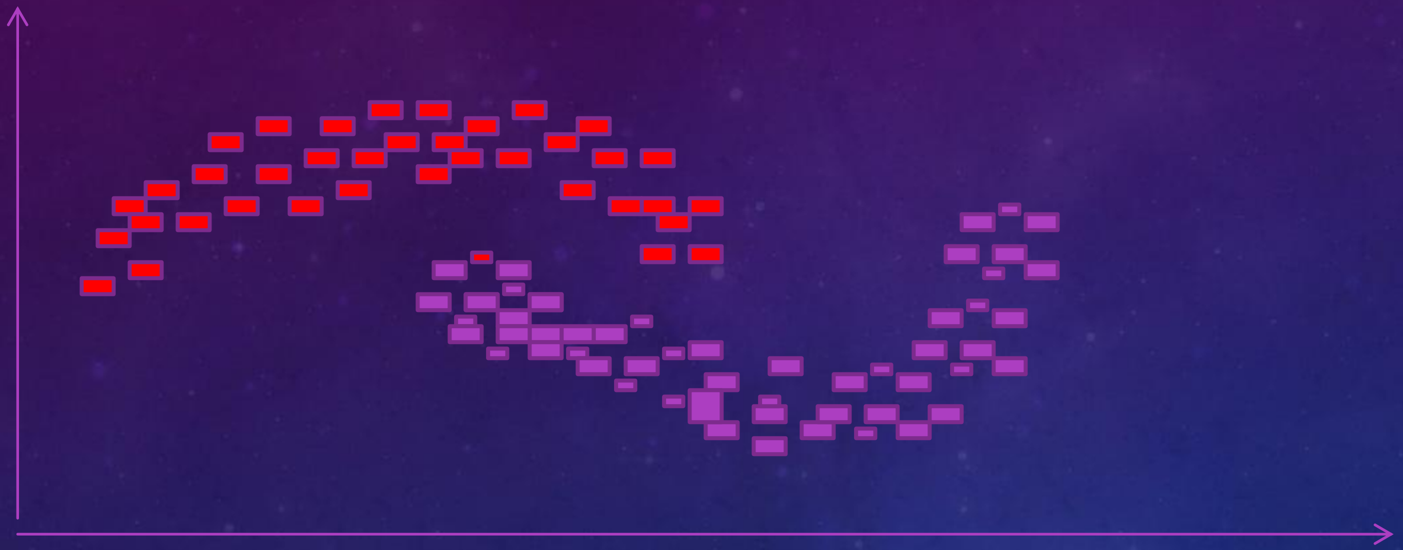
WHAT IS SPECTRAL CLUSTERING

- A Principal Component Analysis Method
- Principal component analysis is a mathematical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- Spectral Clustering can find optimal partition of the data set.

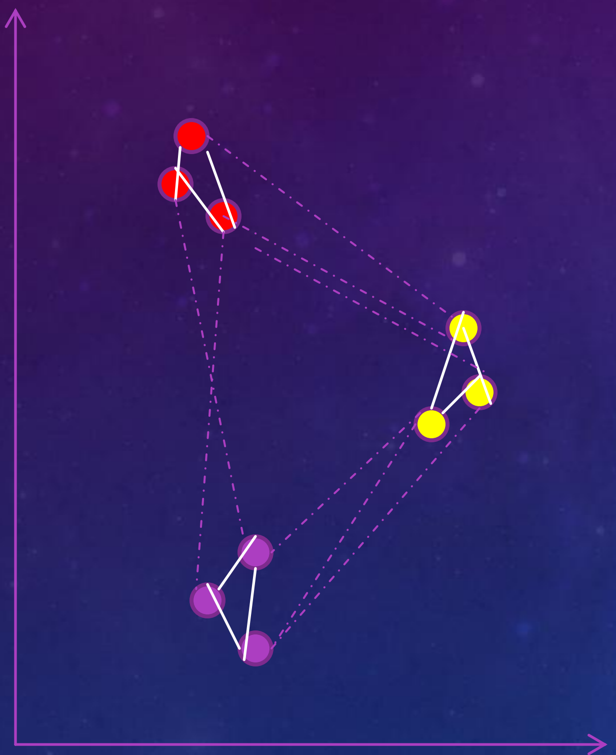
LAPLACIAN MATRIX

- A matrix representation of a graph.
- Main tools for spectral clustering

SPECTRAL CLUSTERING (AFFINITY-BASED CLUSTERING)



Similarity Matrix



X	X	X	…	…	…	…	…	…
X	X	X	…	…	…	…	…	…
X	X	X	…	…	…	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	…	…	…	X	X	X
…	…	…	…	…	…	X	X	X
…	…	…	…	…	…	X	X	X

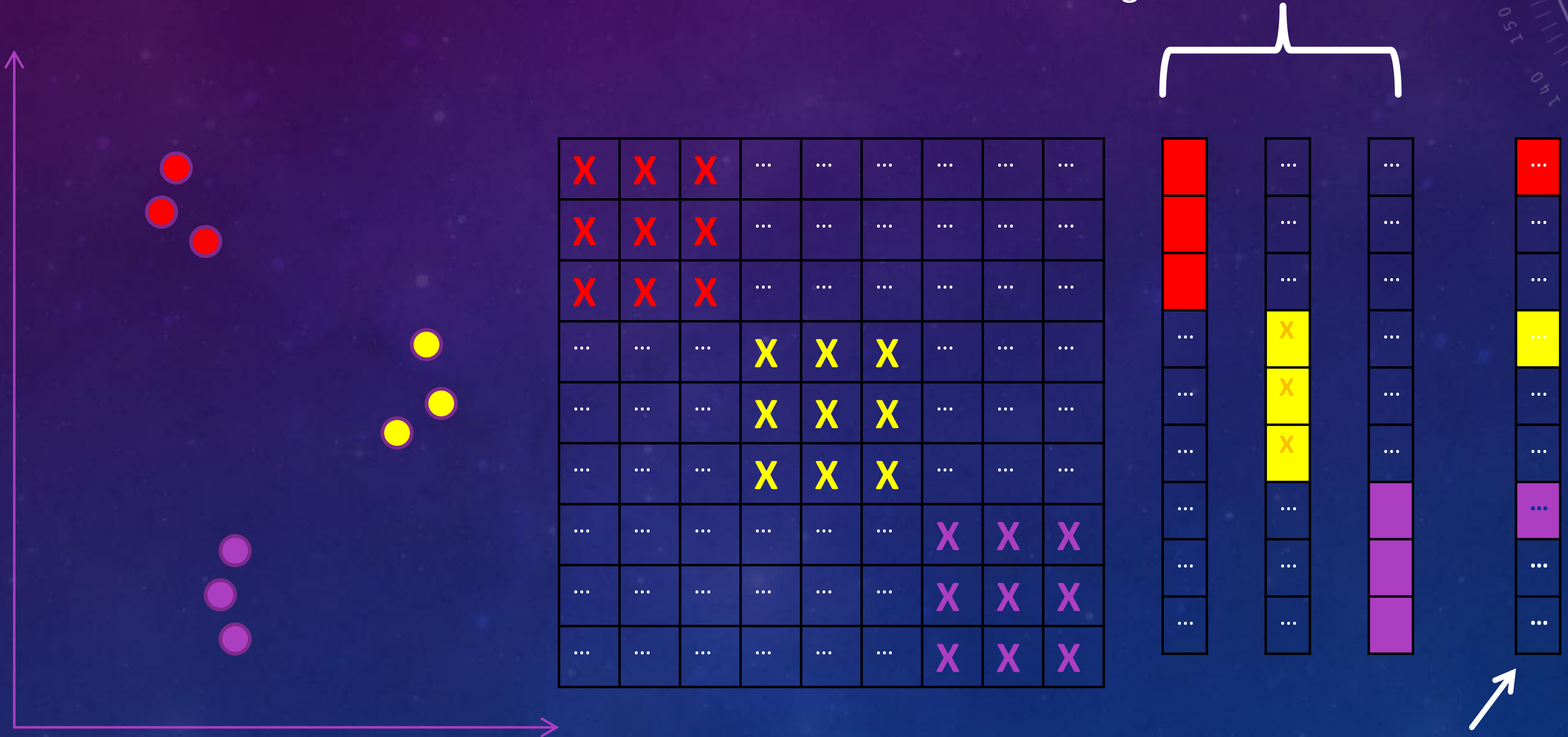
EIGENVECTORS AND EIGENVALUES

- Are used to identify uncorrelated vectors
- Each eigenvector has an eigenvalue that represents how prevalent the eigenvector is in the original data set.
- Eigenvectors of different eigenvalues are orthogonal. They form the dimensionality of a data space, and hence are very useful in clustering data set.

AFFINITY MATRIX AND AFFINITY VECTORS

Eigenvectors with large eigenvalues are most dominant affinity vectors

Eigenvectors with large eigenvalues

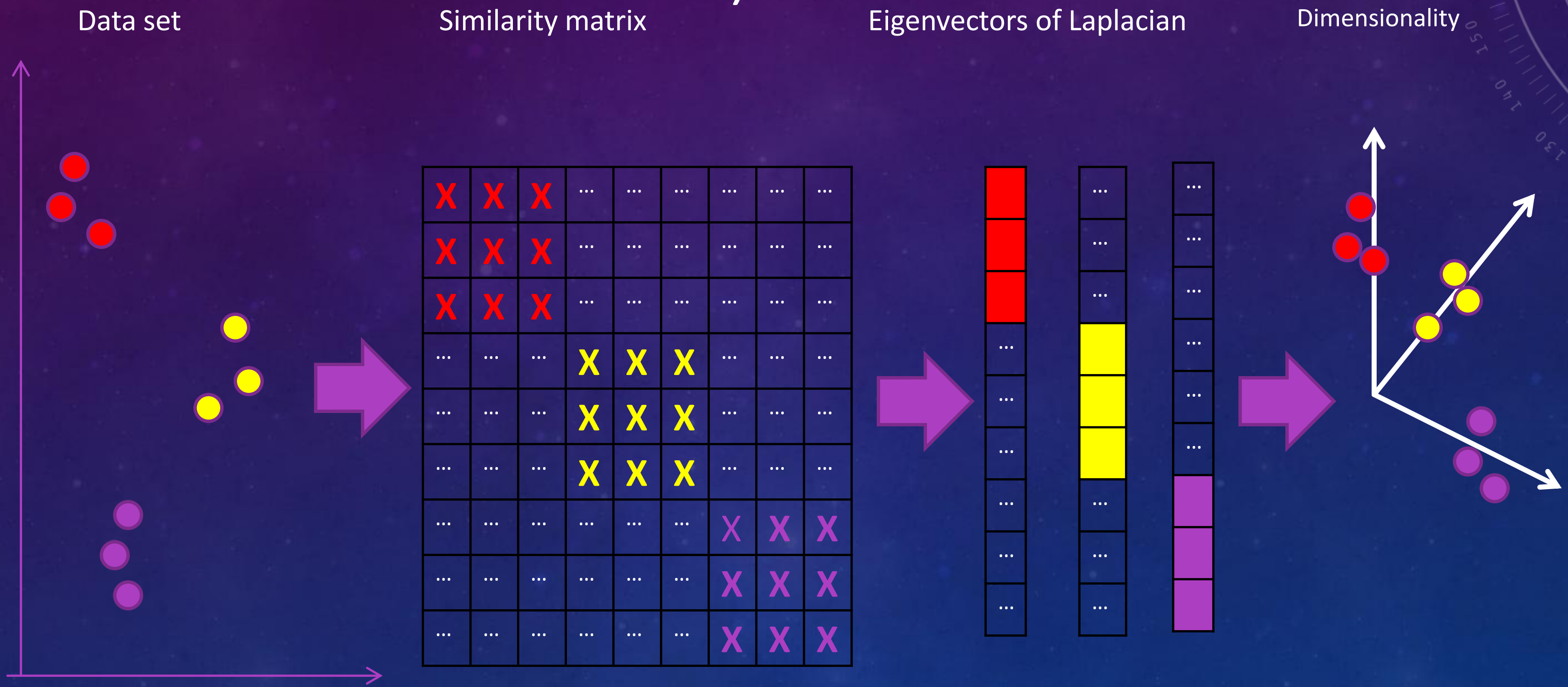


X	X	X	…	…	…	…	…	…
X	X	X	…	…	…	…	…	…
X	X	X	…	…	…	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	…	…	…	X	X	X
…	…	…	…	…	…	X	X	X
…	…	…	…	…	…	X	X	X

Eigenvectors with small eigenvalues less prevalent affinity vectors (noises)

Eigenvectors with small eigenvalue

AFFINITY BASED CLUSTERING (SPECTRAL CLUSTERING)



Data set

Similarity matrix

Eigenvectors of Laplacian

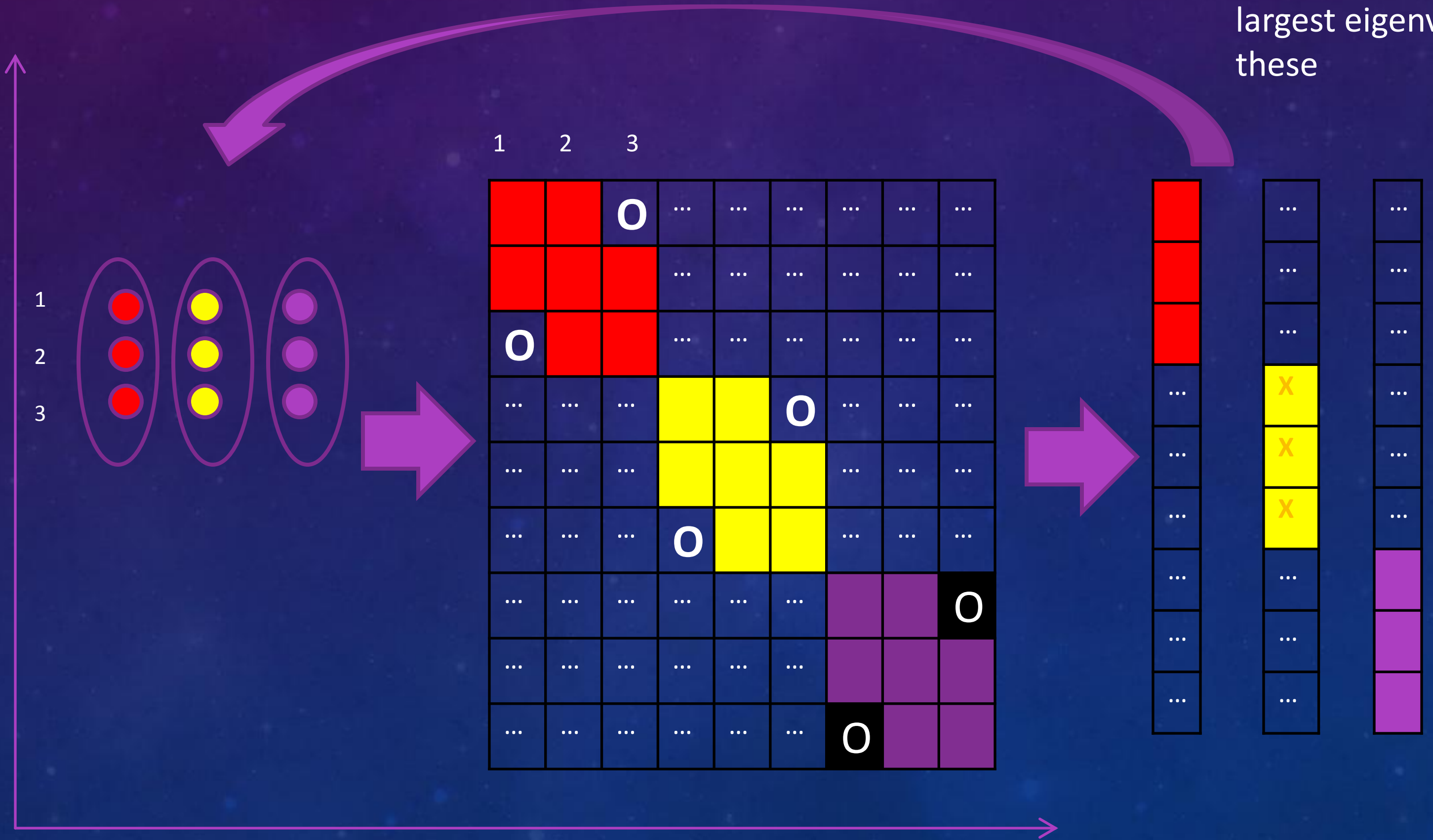
Dimensionality

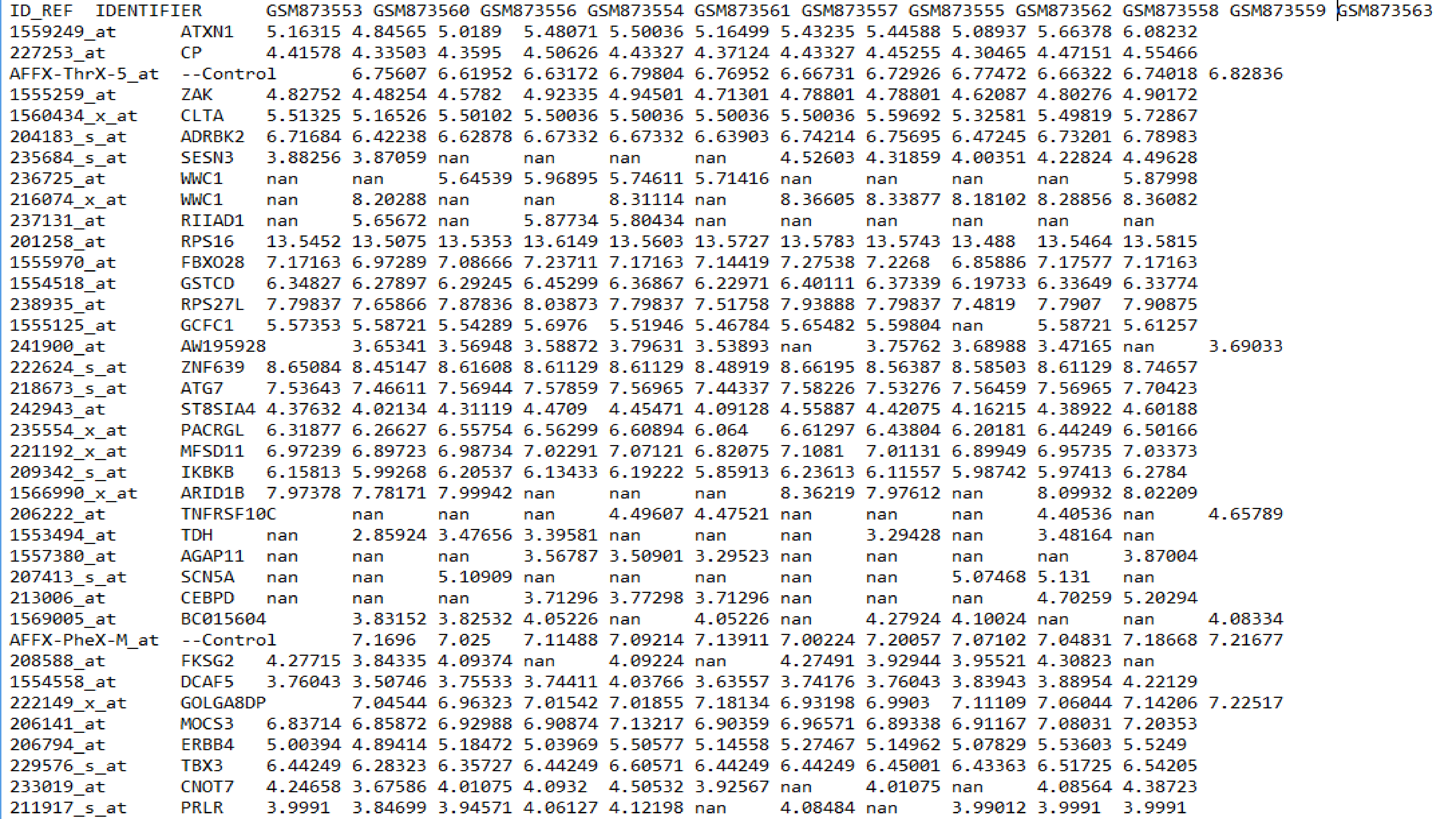
X	X	X	…	…	…	…	…	…
X	X	X	…	…	…	…	…	…
X	X	X	…	…	…	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	X	X	X	…	…	…
…	…	…	…	…	…	X	X	X
…	…	…	…	…	…	X	X	X
…	…	…	…	…	…	X	X	X

WHAT’S THE DIFFERENCE?

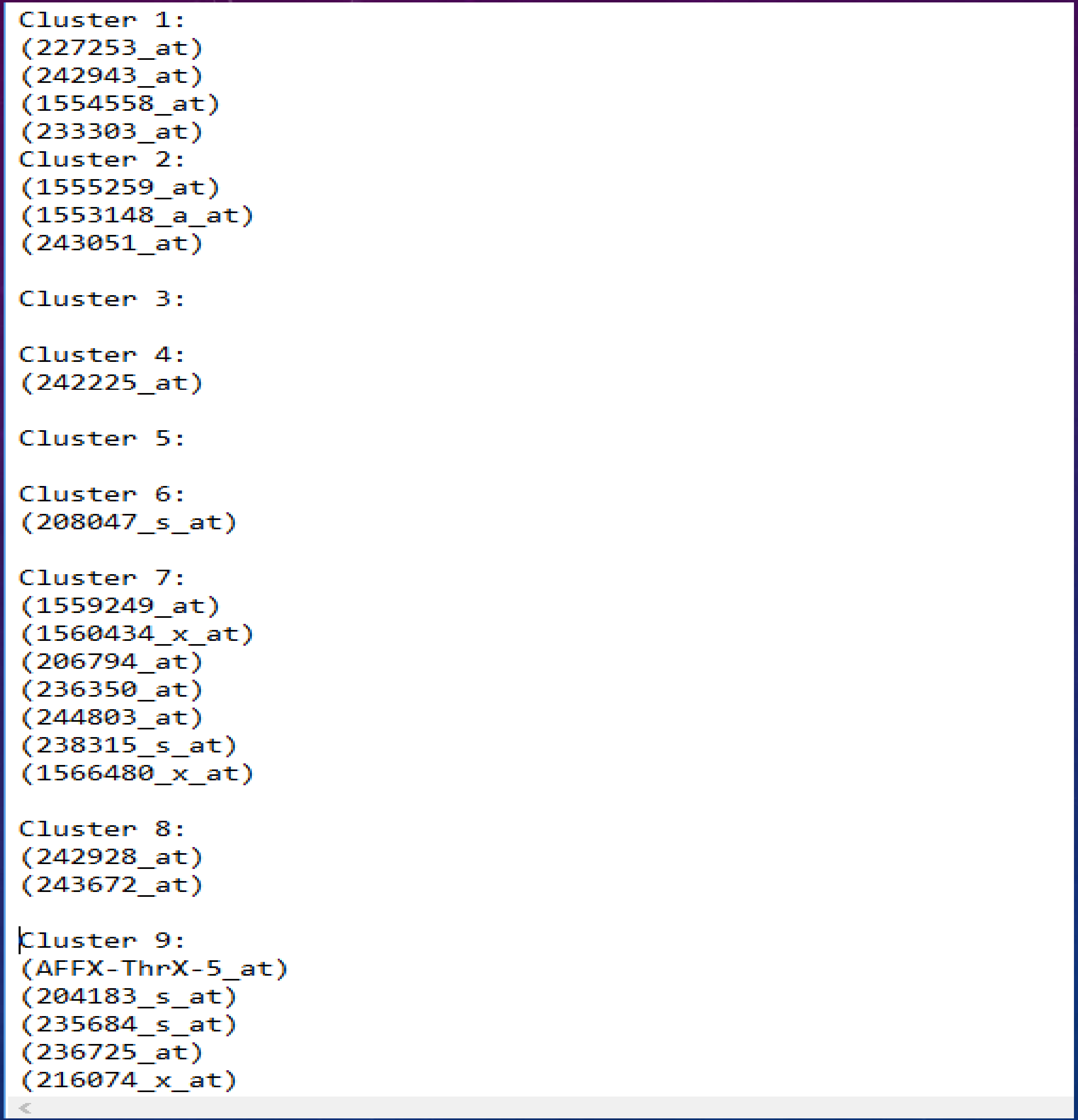
- With Spectral Clustering, the data set can be easily assigned to the shown clusters, which would not be quite the case in traditional clustering techniques.

In spite of the holes, the 3 largest eigenvectoers are still these





```
ID_REF          IDENTIFIER      GSM873553 GSM873560 GSM873556 GSM873554 GSM873561 GSM873557 GSM873555 GSM873562 GSM873558 GSM873559 GSM873563
1559249_at      ATXN1   5.16315 4.84565 5.0189  5.48071 5.50036 5.16499 5.43235 5.44588 5.08937 5.66378 6.08232
227253_at       CP      4.41578 4.33503 4.3595  4.50626 4.43327 4.37124 4.43327 4.45255 4.30465 4.47151 4.55466
AFFX-ThrX-5_at  --Control       6.75607 6.61952 6.63172 6.79804 6.76952 6.66731 6.72926 6.77472 6.66322 6.74018 6.82836
1555259_at      ZAK     4.82752 4.48254 4.5782  4.92335 4.94501 4.71301 4.78801 4.78801 4.62087 4.80276 4.90172
1560434_x_at    CLTA    5.51325 5.16526 5.50102 5.50036 5.50036 5.50036 5.50036 5.59692 5.32581 5.49819 5.72867
204183_s_at     ADRBK2  6.71684 6.42238 6.62878 6.67332 6.67332 6.63903 6.74214 6.75695 6.47245 6.73201 6.78983
235684_s_at     SESN3   3.88256 3.87059 nan     nan     nan     nan     4.52603 4.31859 4.00351 4.22824 4.49628
236725_at       WWC1    nan     nan     5.64539 5.96895 5.74611 5.71416 nan     nan     nan     nan     5.87998
216074_x_at     WWC1    nan     8.20288 nan     nan     8.31114 nan     8.36605 8.33877 8.18102 8.28856 8.36082
237131_at       RIIAD1  nan     5.65672 nan     5.87734 5.80434 nan     nan     nan     nan     nan     nan
201258_at       RPS16   13.5452 13.5075 13.5353 13.6149 13.5603 13.5727 13.5783 13.5743 13.488  13.5464 13.5815
1555970_at      FBXO28  7.17163 6.97289 7.08666 7.23711 7.17163 7.14419 7.27538 7.2268  6.85886 7.17577 7.17163
1554518_at      GSTCD   6.34827 6.27897 6.29245 6.45299 6.36867 6.22971 6.40111 6.37339 6.19733 6.33649 6.33774
238935_at       RPS27L  7.79837 7.65866 7.87836 8.03873 7.79837 7.51758 7.93888 7.79837 7.4819  7.7907  7.90875
1555125_at      GCFC1   5.57353 5.58721 5.54289 5.6976  5.51946 5.46784 5.65482 5.59804 nan     5.58721 5.61257
241900_at       AW195928        3.65341 3.56948 3.58872 3.79631 3.53893 nan     3.75762 3.68988 3.47165 nan     3.69033
222624_s_at     ZNF639  8.65084 8.45147 8.61608 8.61129 8.61129 8.48919 8.66195 8.56387 8.58503 8.61129 8.74657
218673_s_at     ATG7    7.53643 7.46611 7.56944 7.57859 7.56965 7.44337 7.58226 7.53276 7.56459 7.56965 7.70423
242943_at       ST8SIA4 4.37632 4.02134 4.31119 4.4709  4.45471 4.09128 4.55887 4.42075 4.16215 4.38922 4.60188
235554_x_at     PACRGL  6.31877 6.26627 6.55754 6.56299 6.60894 6.064   6.61297 6.43804 6.20181 6.44249 6.50166
221192_x_at     MFSD11  6.97239 6.89723 6.98734 7.02291 7.07121 6.82075 7.1081  7.01131 6.89949 6.95735 7.03373
209342_s_at     IKBKB   6.15813 5.99268 6.20537 6.13433 6.19222 5.85913 6.23613 6.11557 5.98742 5.97413 6.2784
1566990_x_at    ARID1B  7.97378 7.78171 7.99942 nan     nan     nan     8.36219 7.97612 nan     8.09932 8.02209
206222_at       TNFRSF10C       nan     nan     nan     4.49607 4.47521 nan     nan     nan     4.40536 nan     4.65789
1553494_at      TDH     nan     2.85924 3.47656 3.39581 nan     nan     nan     3.29428 nan     3.48164 nan
1557380_at      AGAP11  nan     nan     nan     3.56787 3.50901 3.29523 nan     nan     nan     nan     3.87004
207413_s_at     SCN5A   nan     nan     5.10909 nan     nan     nan     nan     nan     5.07468 5.131   nan
213006_at       CEBPD   nan     nan     nan     3.71296 3.77298 3.71296 nan     nan     nan     4.70259 5.20294
1569005_at      BC015604        3.83152 3.82532 4.05226 nan     4.05226 nan     4.27924 4.10024 nan     nan     4.08334
AFFX-PheX-M_at  --Control       7.1696  7.025   7.11488 7.09214 7.13911 7.00224 7.20057 7.07102 7.04831 7.18668 7.21677
208588_at       FKSG2   4.27715 3.84335 4.09374 nan     4.09224 nan     4.27491 3.92944 3.95521 4.30823 nan
1554558_at      DCAF5   3.76043 3.50746 3.75533 3.74411 4.03766 3.63557 3.74176 3.76043 3.83943 3.88954 4.22129
222149_x_at     GOLGA8DP        7.04544 6.96323 7.01542 7.01855 7.18134 6.93198 6.9903  7.11109 7.06044 7.14206 7.22517
206141_at       MOCS3   6.83714 6.85872 6.92988 6.90874 7.13217 6.90359 6.96571 6.89338 6.91167 7.08031 7.20353
206794_at       ERBB4   5.00394 4.89414 5.18472 5.03969 5.50577 5.14558 5.27467 5.14962 5.07829 5.53603 5.5249
229576_s_at     TBX3    6.44249 6.28323 6.35727 6.44249 6.60571 6.44249 6.44249 6.45001 6.43363 6.51725 6.54205
233019_at       CNOT7   4.24658 3.67586 4.01075 4.0932  4.50532 3.92567 nan     4.01075 nan     4.08564 4.38723
211917_s_at     PRLR    3.9991  3.84699 3.94571 4.06127 4.12198 nan     4.08484 nan     3.99012 3.9991  3.9991
```

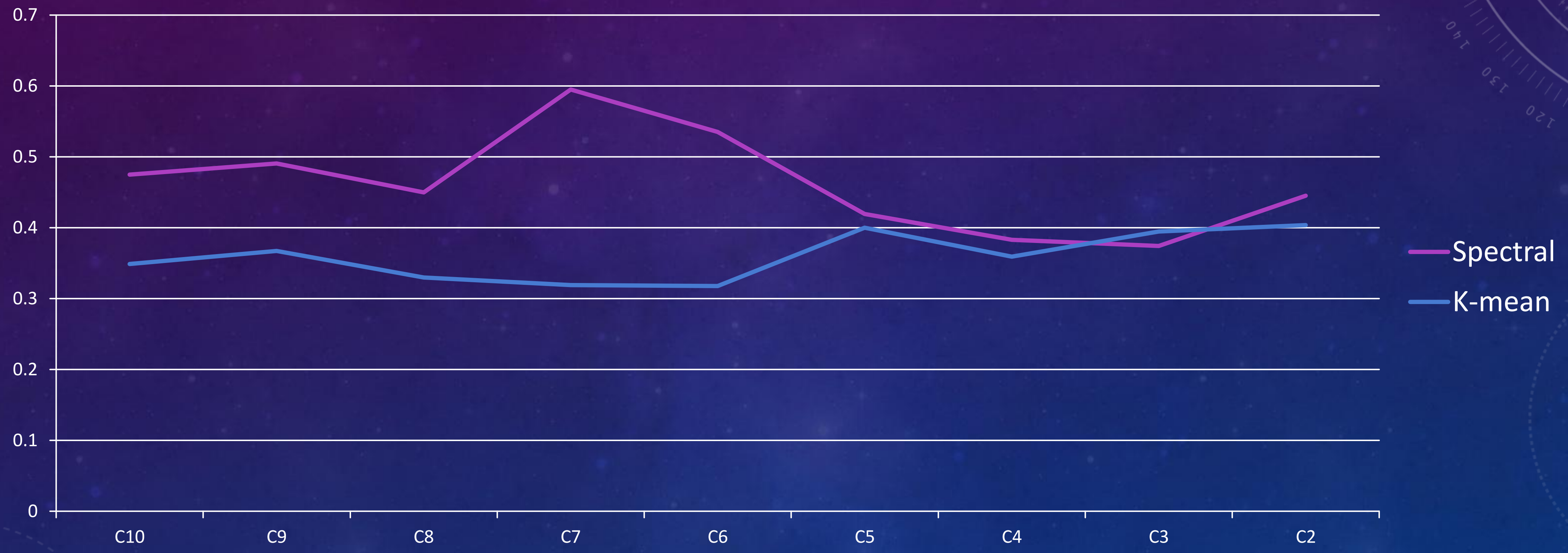


```
Cluster 1:
(227253_at)
(242943_at)
(1554558_at)
(233303_at)
Cluster 2:
(1555259_at)
(1553148_a_at)
(243051_at)

Cluster 3:

Cluster 4:
(242225_at)

Cluster 5:

Cluster 6:
(208047_s_at)

Cluster 7:
(1559249_at)
(1560434_x_at)
(206794_at)
(236350_at)
(244803_at)
(238315_s_at)
(1566480_x_at)

Cluster 8:
(242928_at)
(243672_at)

Cluster 9:
(AFFX-ThrX-5_at)
(204183_s_at)
(235684_s_at)
(236725_at)
(216074_x_at)
```

ABOUT SILHOUETTE

- Method of cluster validation.
- The silhouette can validate in cluster genes simalirty.
- Can also validate overall clustering output.

RESULTS



CONCLUSION

- Clustering is a faster more efficient way of finding similarities then sequencing.
- Spectral Clustering out performs k-means in most cases.