Iterative Consensus Clustering: An Algorithm We Can All Agree On

Mindy Hong[1], Robert Pearce[2], Kevin Valakuzhy[3], Carl Meyer[2], Shaina Race[2]

1 EMORY UNIVERSITY

2

3 THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

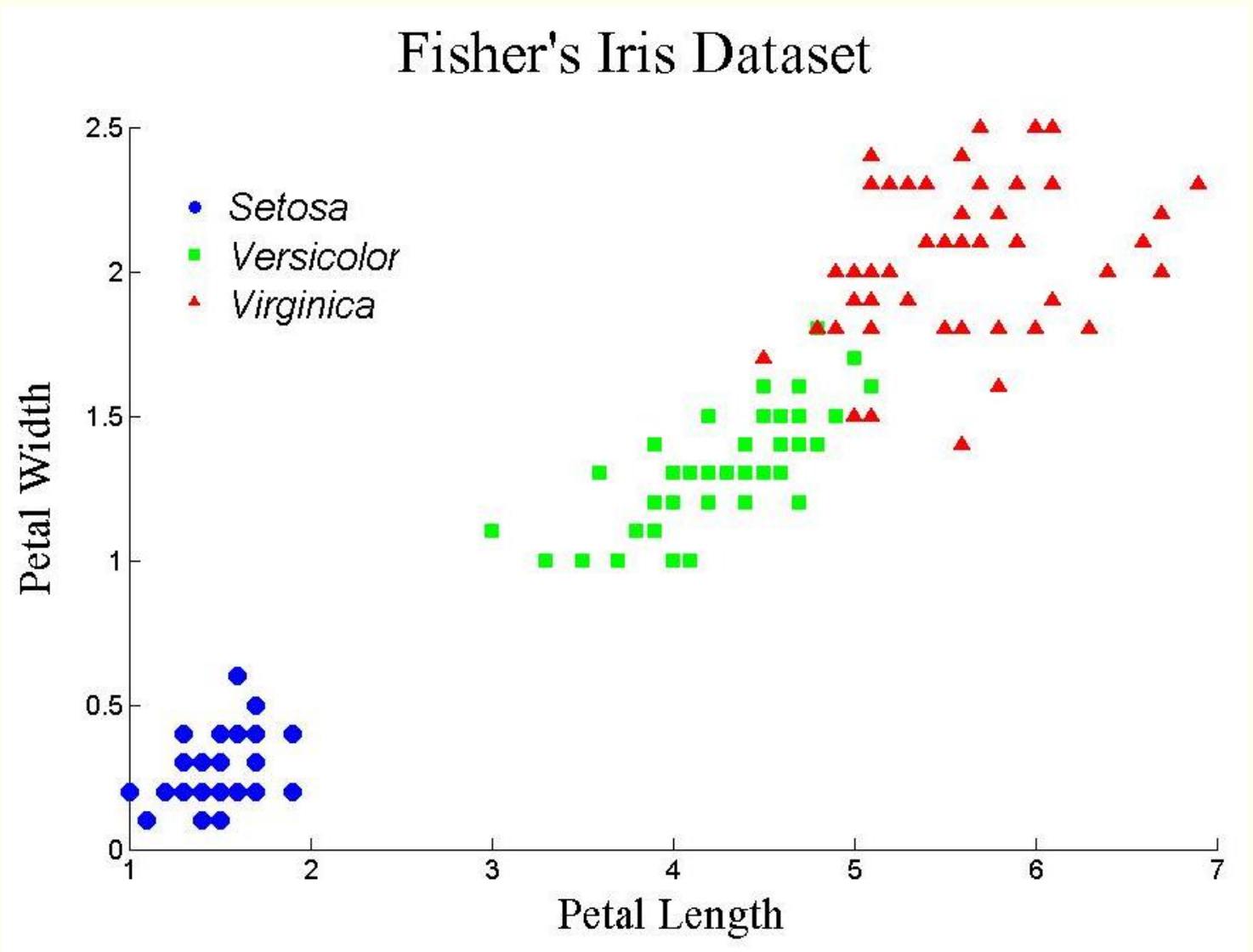
Background Information

Clustering : Grouping data based on a predefined metric of similarity.

Why Cluster?

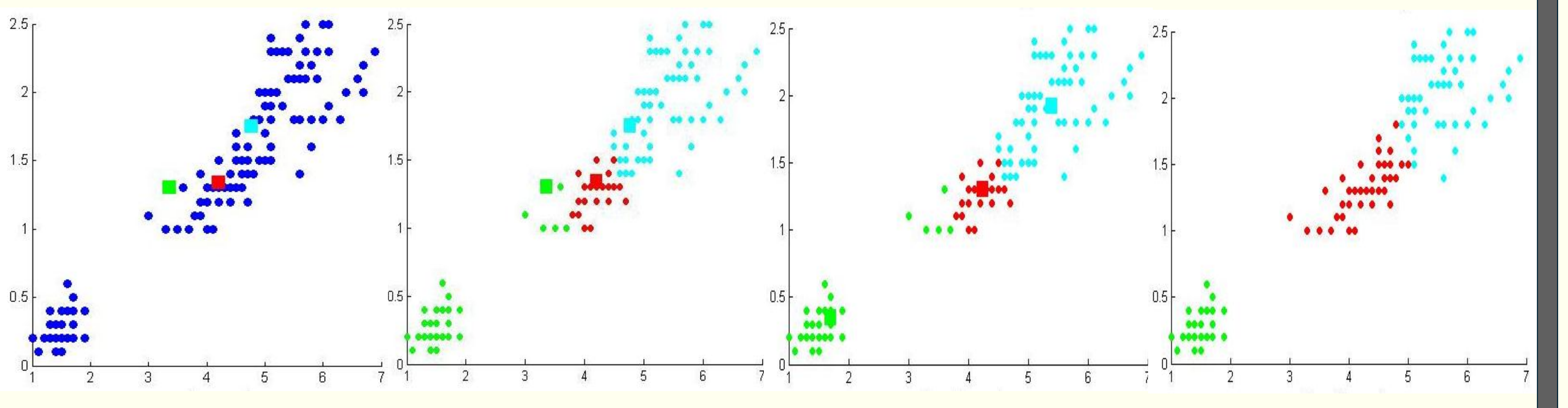
First step in interpreting large amounts of data

- Physical Observations
- Gene Expression
- Term Frequencies



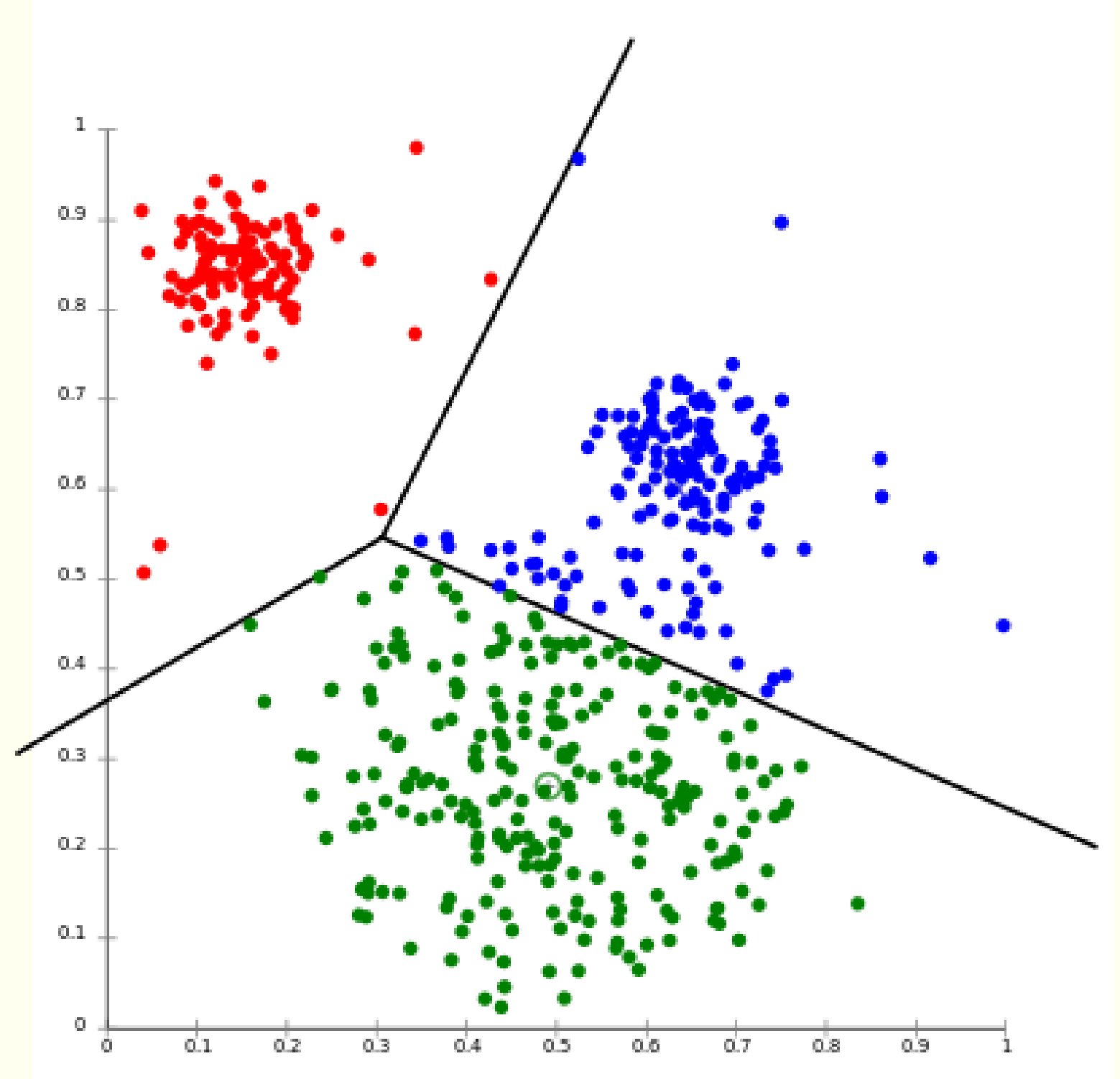
Example Algorithm : K-Means

1. Randomize centroids for each cluster
2. Cluster each point with its nearest centroid
3. Move centroid to mean of its cluster
4. Repeat steps 2 and 3 until equilibrium



Step 1: Initialize Step 2: Group Step 3: Center Step 4: Repeat

Problems



Picture by Chire

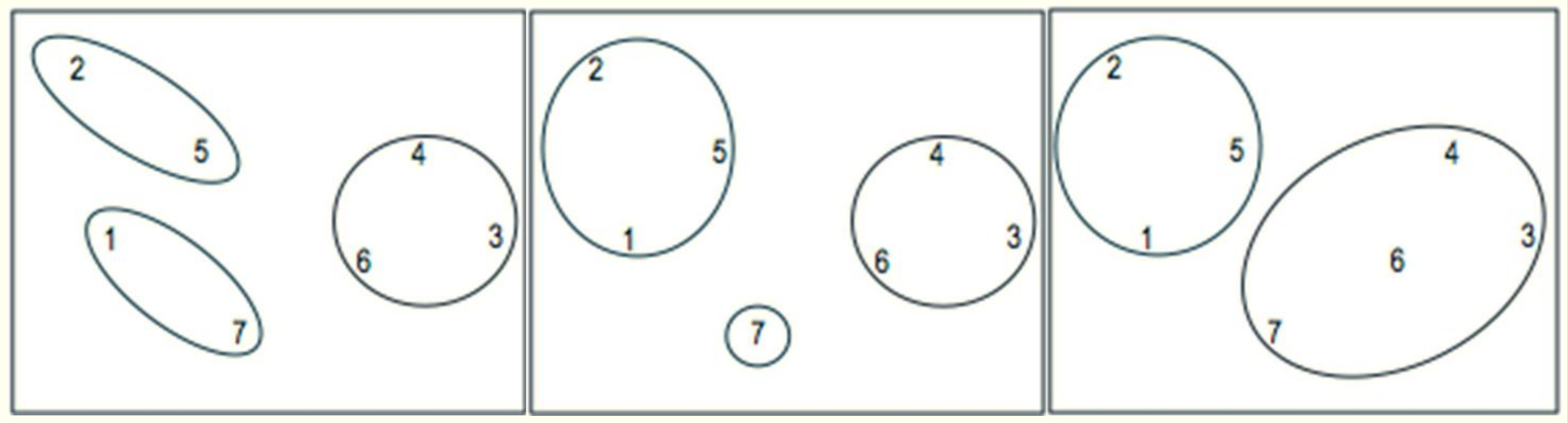
- Fundamental Problem of Clustering
  "There does not exist a best method, that is, one which is superior to all other methods" (Kogan).
- Determining the number of clusters, also known as *k*

Objectives

- Determining an accurate value for *k*
- Develop a technique that uses multiple algorithms to reach a consensus on a final clustering

Methods

Consensus Clustering

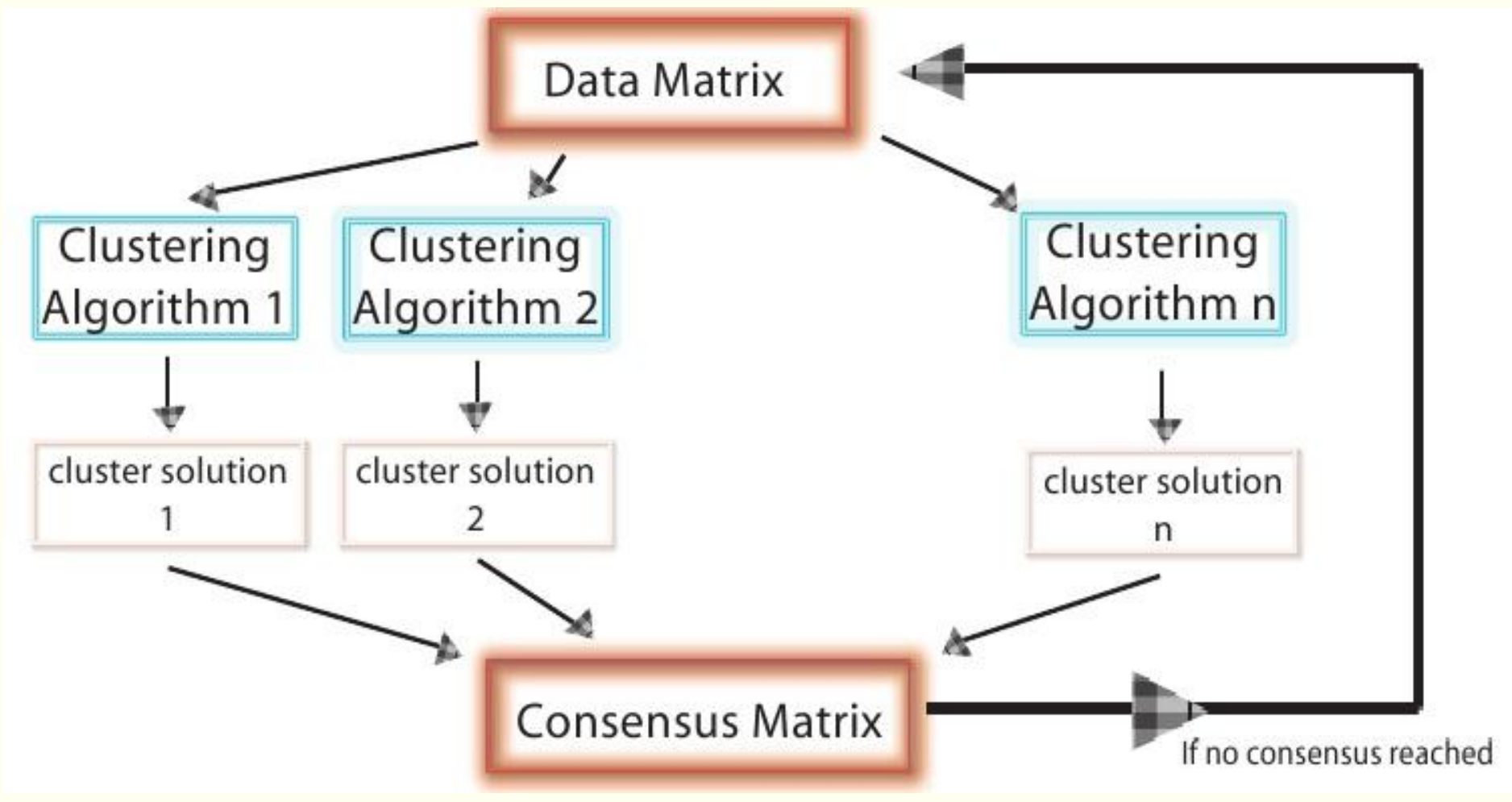


$$\begin{array}{c|ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \hline 1 & 0 & 2 & 0 & 0 & 2 & 0 & 1 \\ 2 & 2 & 0 & 0 & 0 & 3 & 0 & 0 \\ 3 & 0 & 0 & 0 & 3 & 0 & 3 & 1 \\ 4 & 0 & 0 & 3 & 0 & 0 & 3 & 1 \\ 5 & 2 & 3 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 3 & 3 & 0 & 0 & 1 \\ 7 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \end{array}$$

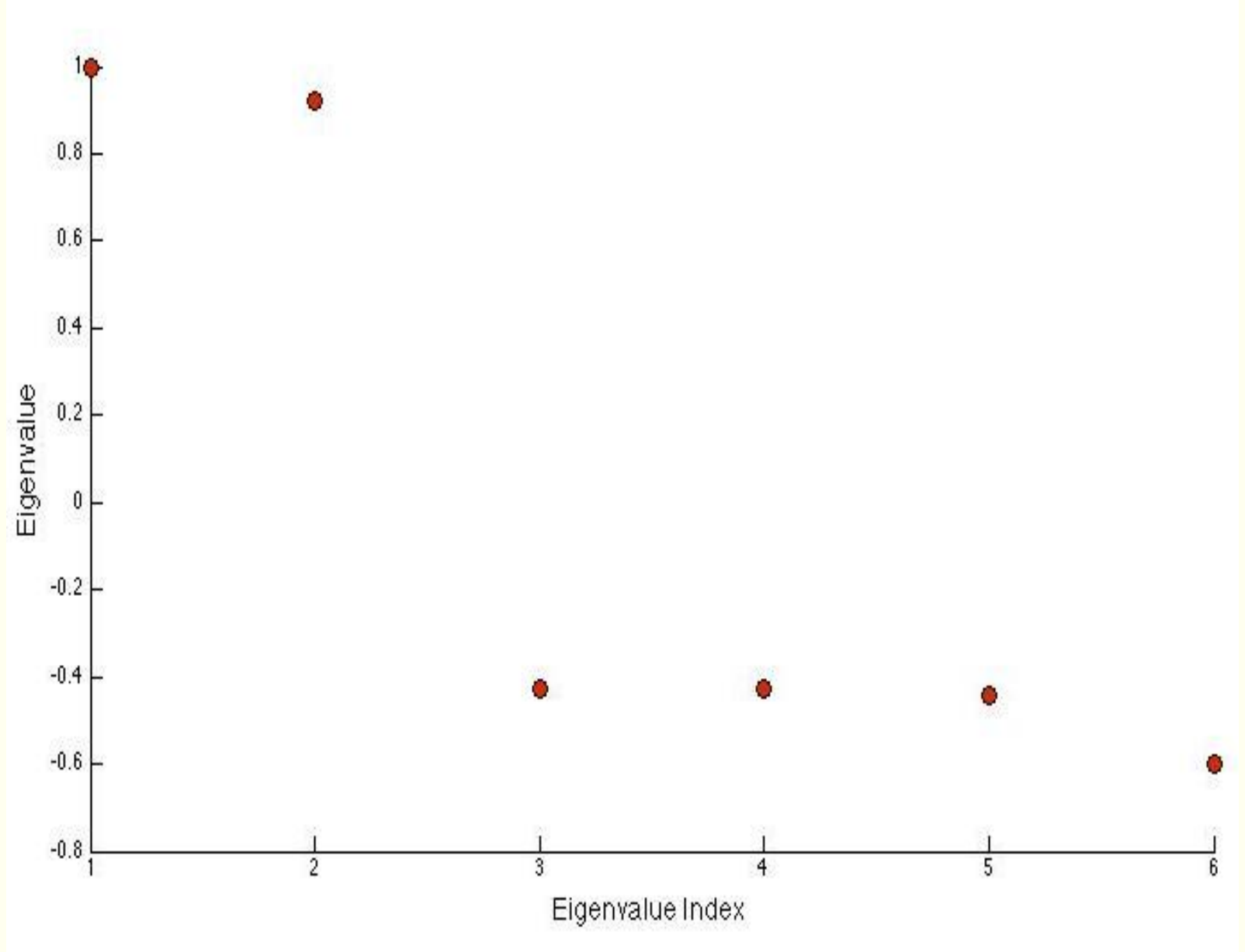
- Each row and column represent a point
- Each matrix entry is the number of times its corresponding row and column are clustered together

Iterated Consensus Clustering

- Treats consensus matrix as a new set of data
- Clusters consensus matrix based on similarities in previous groupings
- Terms in consensus matrix below a certain threshold are dropped
- Iterates this process until convergence



Eigengap Method

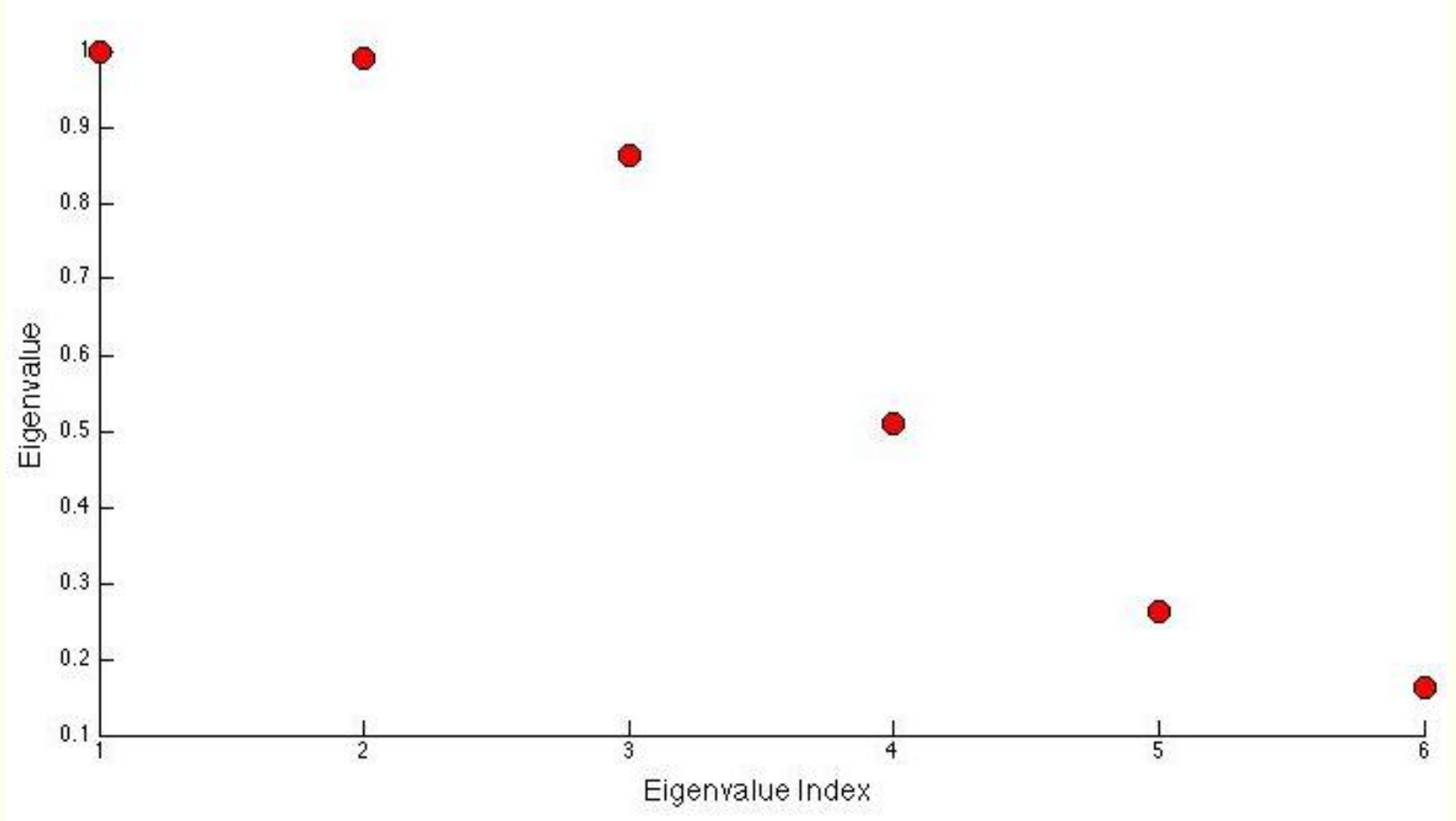


- Eigengap : the largest difference between consecutive eigenvalues
- Create a special "P Matrix" using the consensus matrix
- Sort the P Matrix eigenvalues
- The index of the eigenvalue before the eigengap is an approximation for k

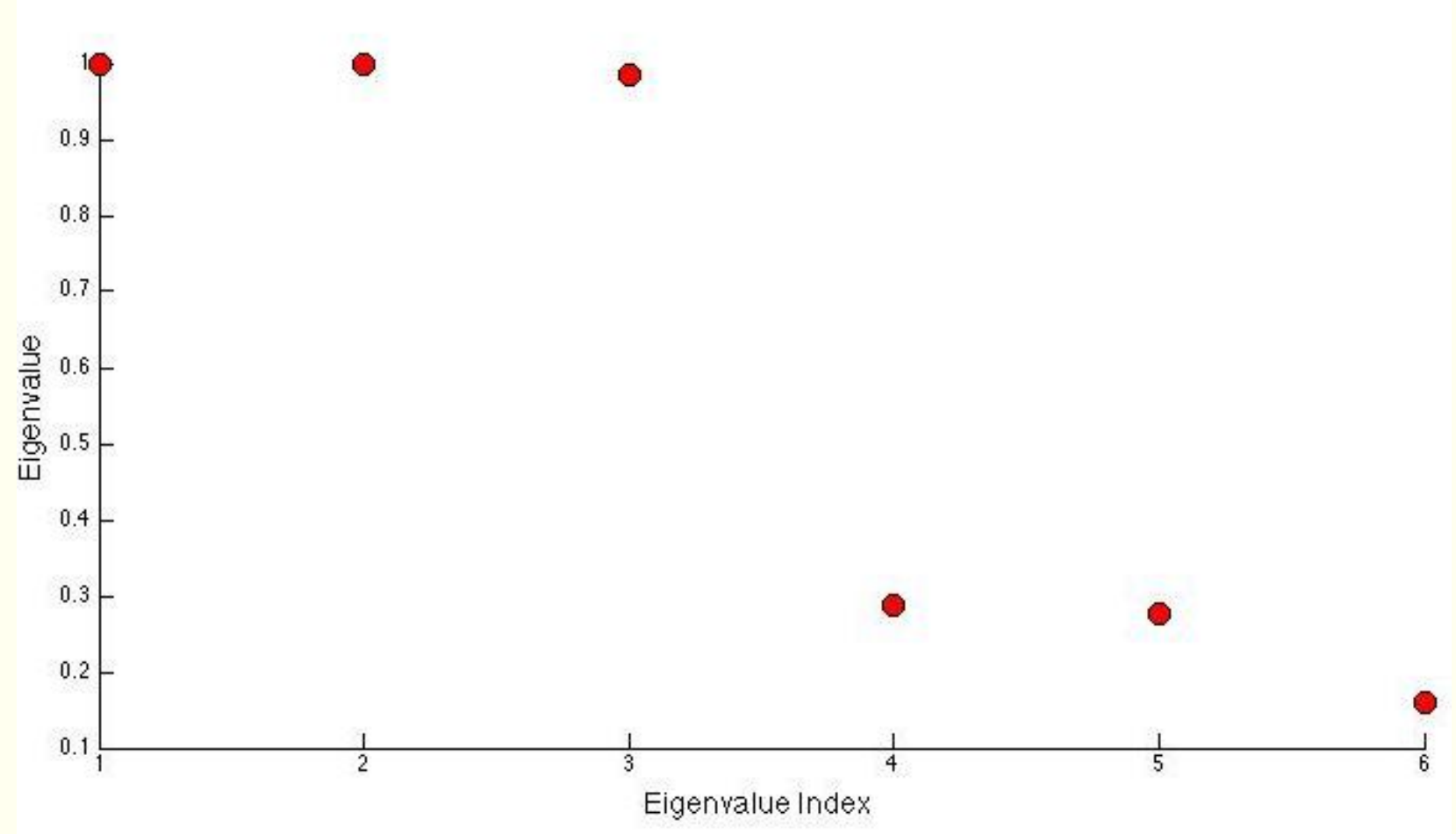
Results / Conclusion

Number of Clusters

Traditional Consensus Clustering



Iterated Consensus Clustering



- Iterated Consensus Clustering creates a larger eigengap, allowing for easier, and more confident, interpretation

Algorithm Consensus : Clustering Accuracy

Algorithm	1st Iteration	2nd Iteration	3rd Iteration
Alg 1	82%	89%	96%
Alg 2	79%	93%	96%
Alg 3	51%	88%	96%
Alg 4	89%	93%	96%
Alg 5	88%	95%	96%

- Errors are weeded out through iteration
- Most algorithms come to a consensus on the final clustering
- Final clustering improves upon many individual algorithms

Conclusion

- Iterated Consensus Clustering offers better results than traditional consensus clustering in:
  - Finding the number of clusters
  - Returning an appropriate clustering
- Calculations were done using the following techniques :
  Expectation Maximization, PDDP, k-Means, NMF, PCA, SVD

Acknowledgements

We would like to thank Dr. Carl Meyer and Shaina Race for their guidance through this project, as well as NSF and NSA for making this research possible.

NSA REU: H9823-10-1-0252
NSF REU: DMS-1063010
NSF REG: DMS-0943855