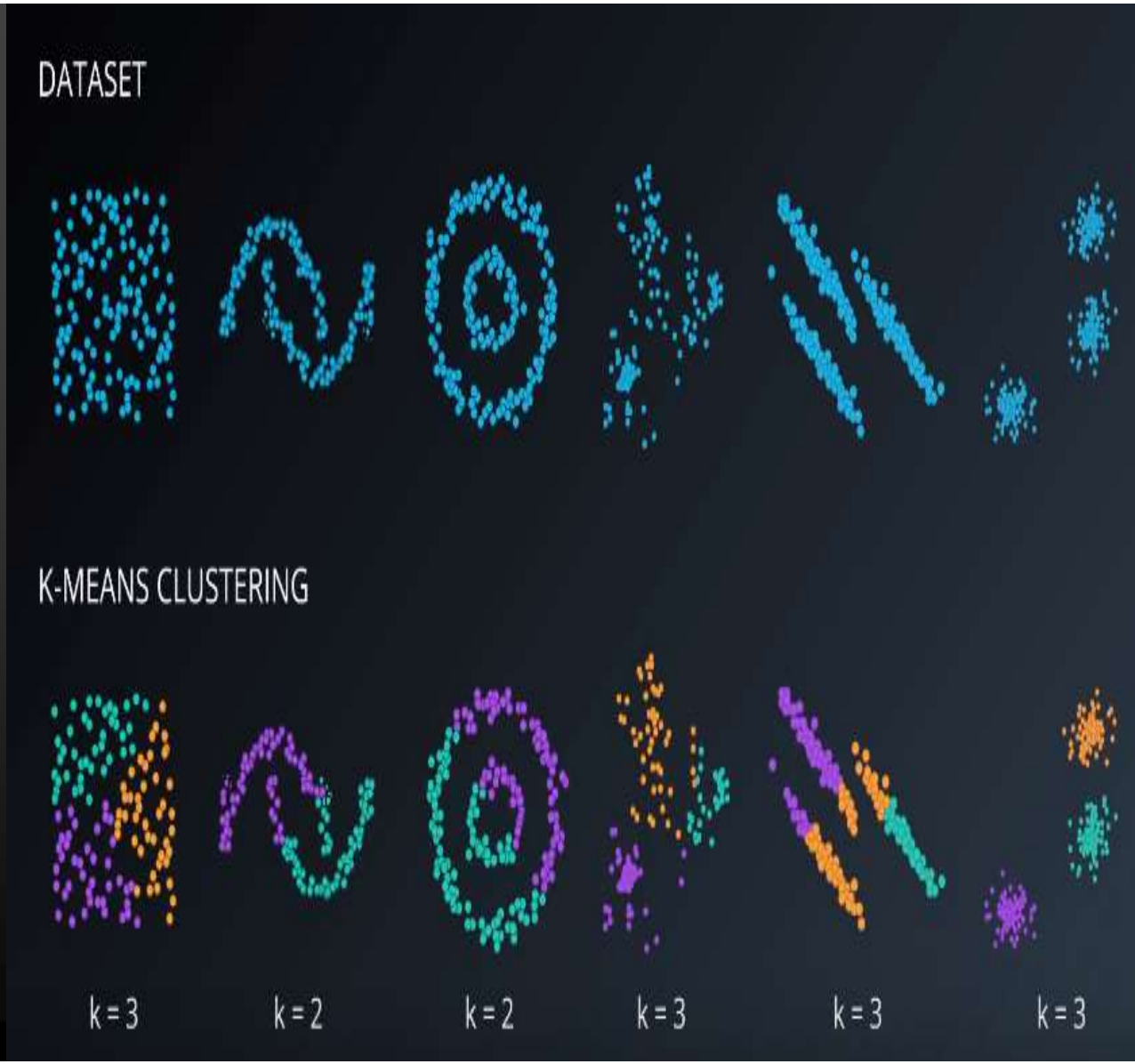
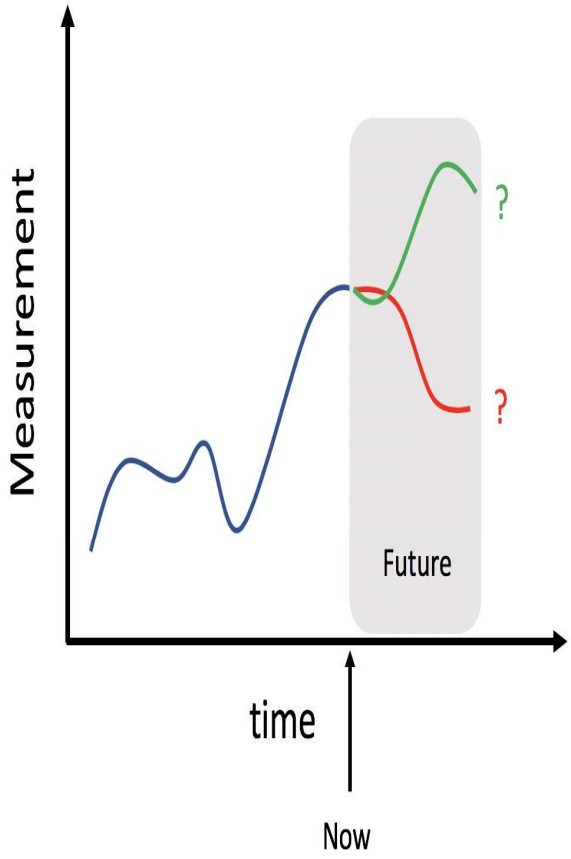
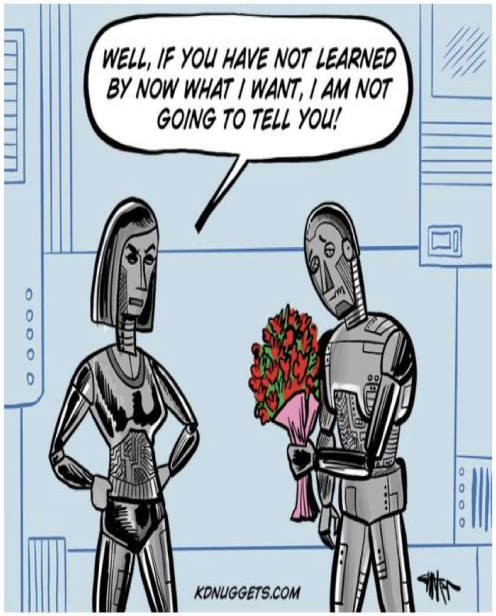
Study of Internal Indices on Time Series Data

Shailesh Kumar Jha



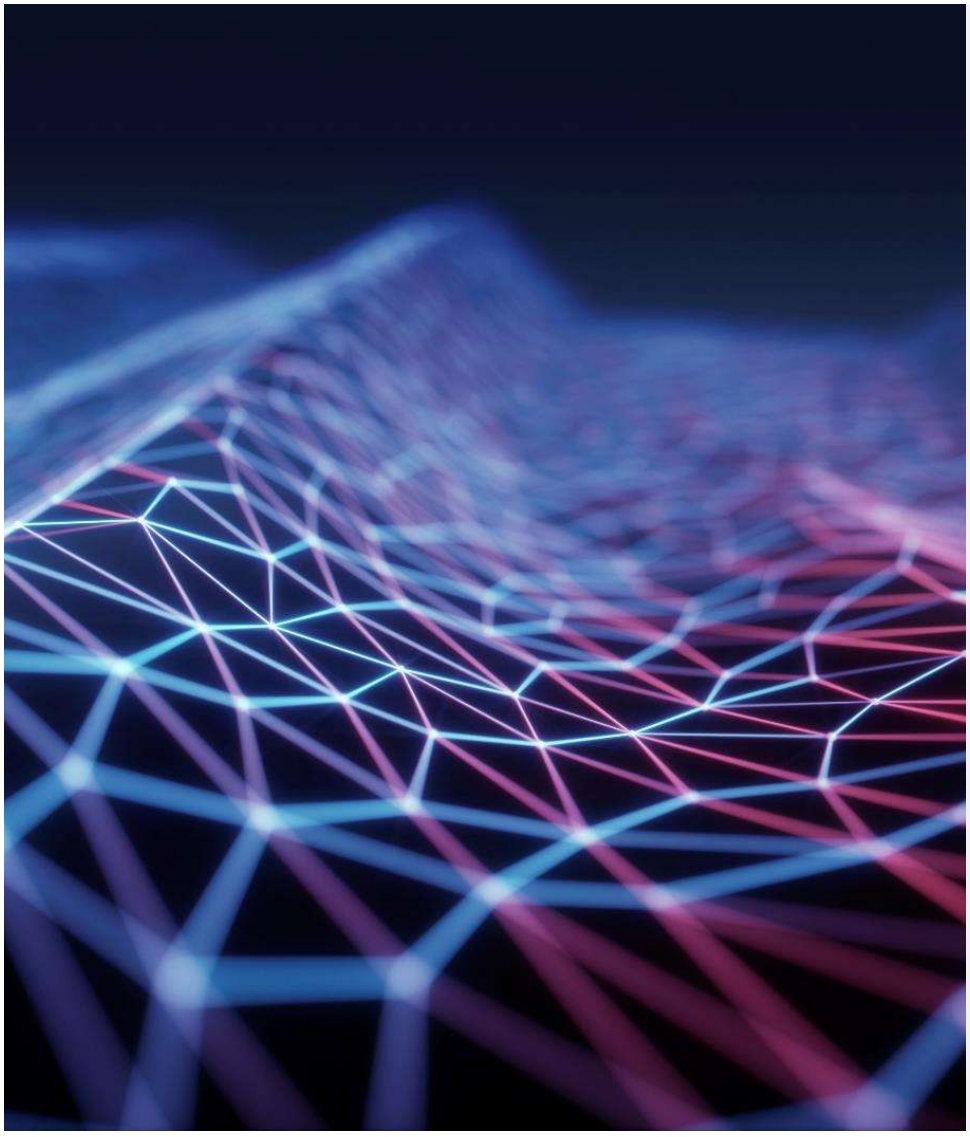
Machine Learning Time Series Problem





What is Time Series Data

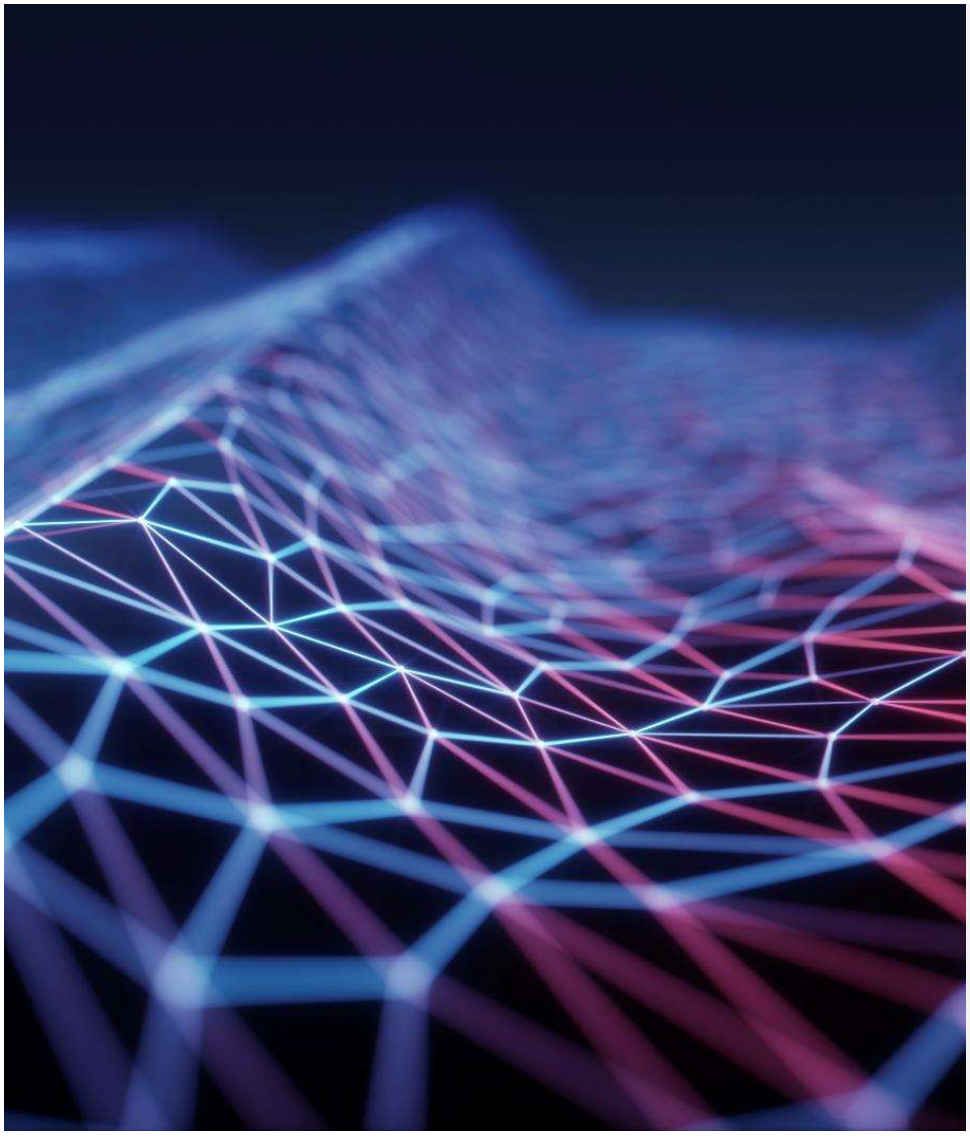
- ✓ Data over a time period
- ✓ Can be used to analyze historical data
- ✓ Forecasting of future trends
- ✓ Applied to various domains



External Indices

Measure of clustering quality

- An external index is a measure of agreement between two partitions where the first partition is the a priori known clustering structure, and the second results from the clustering. (Dudoit et al., 2002).
- We evaluate the results of a clustering algorithm based on a known cluster structure of a data set (or cluster labels)
- We make use of information about the ground truth (e.g. true class labels, if available).



Internal Indices

Measure of clustering quality

- Internal indices are used to measure the goodness of a clustering structure without external information (Tseng et al., 2005).
- Evaluates results using quantity and features inherent in the dataset.
- No information about the ground truth available.

Clustering Methods under consideration

Unsupervised Learning

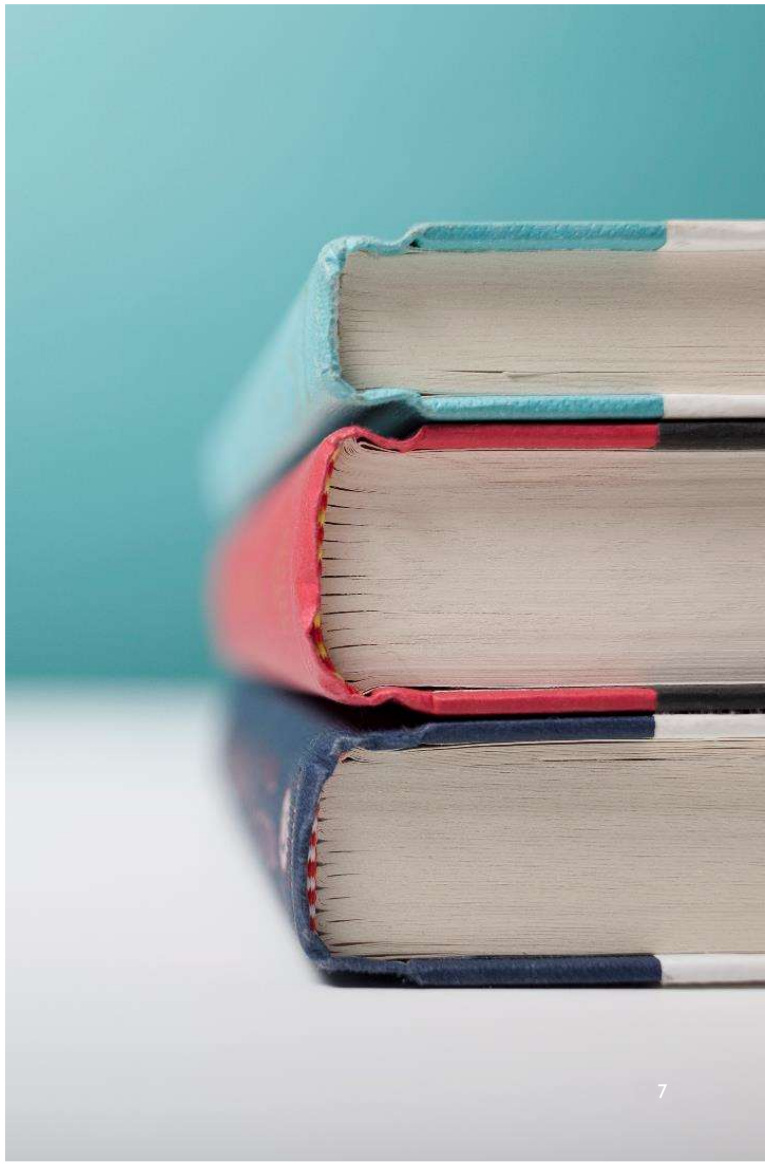
- Kmeans Clustering
- Hierarchical Single Linkage
- Hierarchical Complete Linkage
- Hierarchical Average Linkage
- DDC(Distance Density Clustering)

Purpose of the study

- To study if we can effectively use DTW instead of Euclidean distance for internal indices.
- To explore if we have a clear winner between the clustering methods that we are studying.

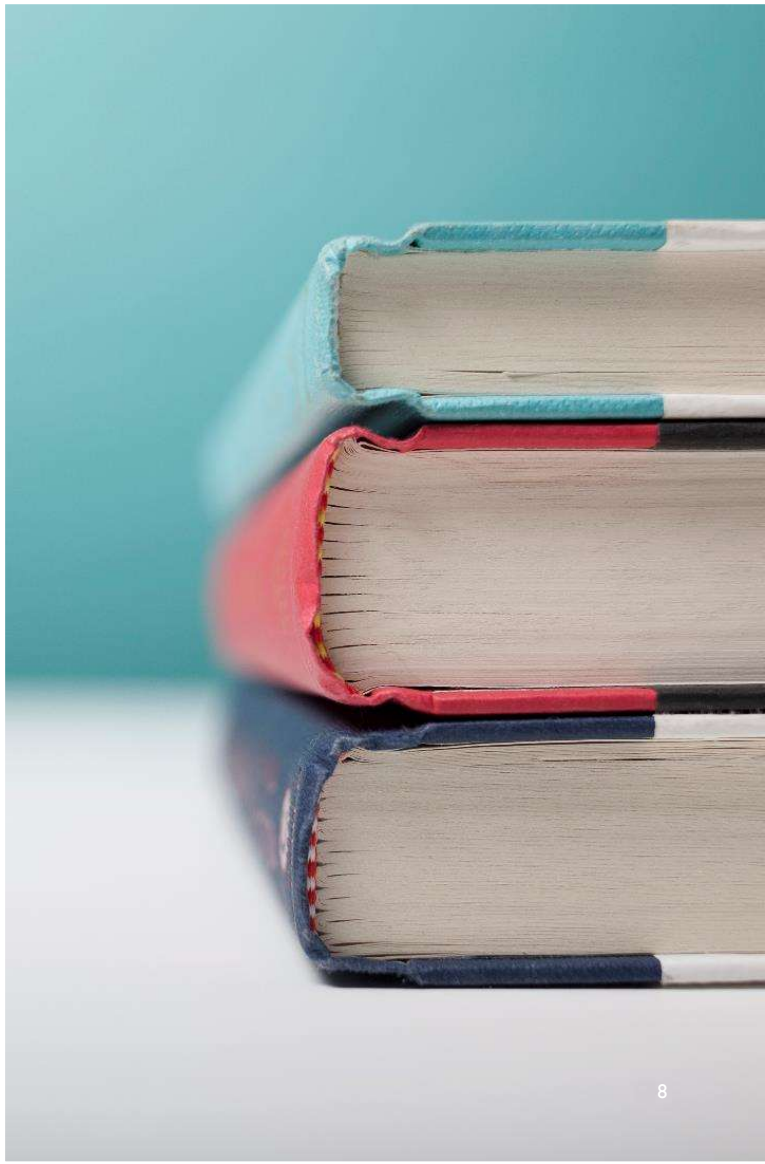
Datasets

- UCR Repository
- 7 internal indices algorithms used
- 5 program for each internal index algorithm
- 10 rounds for Kmeans averaging
- 10 datasets used



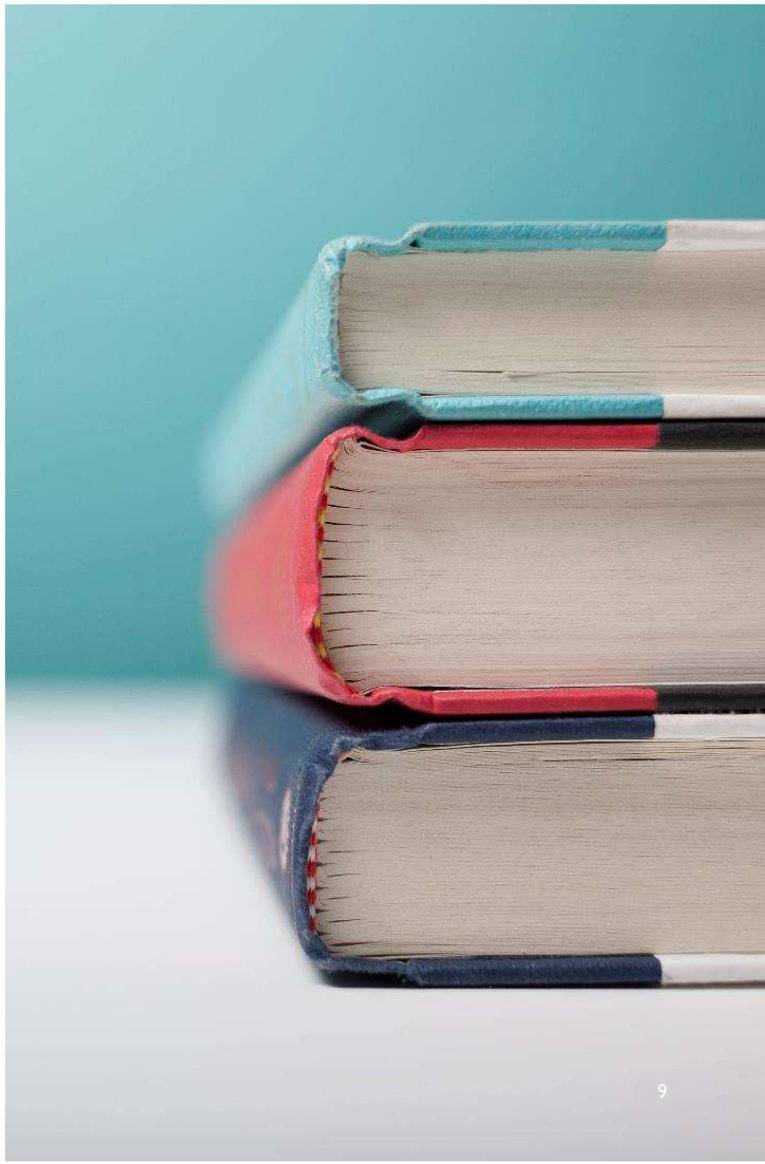
Datasets

- **ArrowHead:** The arrowhead data consists of outlines of the images of arrowheads. The shapes of the projectile points are converted into a time series using the angle-based method.
- **BeetleFly:** It is a database of binary images developed for testing MPEG-7 shape descriptors. It is used for testing contour/image and skeleton-based descriptors
- **BirdChicken:** It is a database of binary images developed for testing MPEG-7 shape descriptors. It is used for testing contour/image and skeleton-based descriptors
- **Coffee:** Food spectrographs are used in chemometrics to classify food types, a task that has obvious applications in food safety and quality assurance. The coffee data set is a two-class problem to distinguish between Robusta and Aribica coffee beans.
- **GunPoint:** This dataset involves one female actor and one male actor making a motion with their hand. The two classes are: Gun-Draw and Point



Datasets

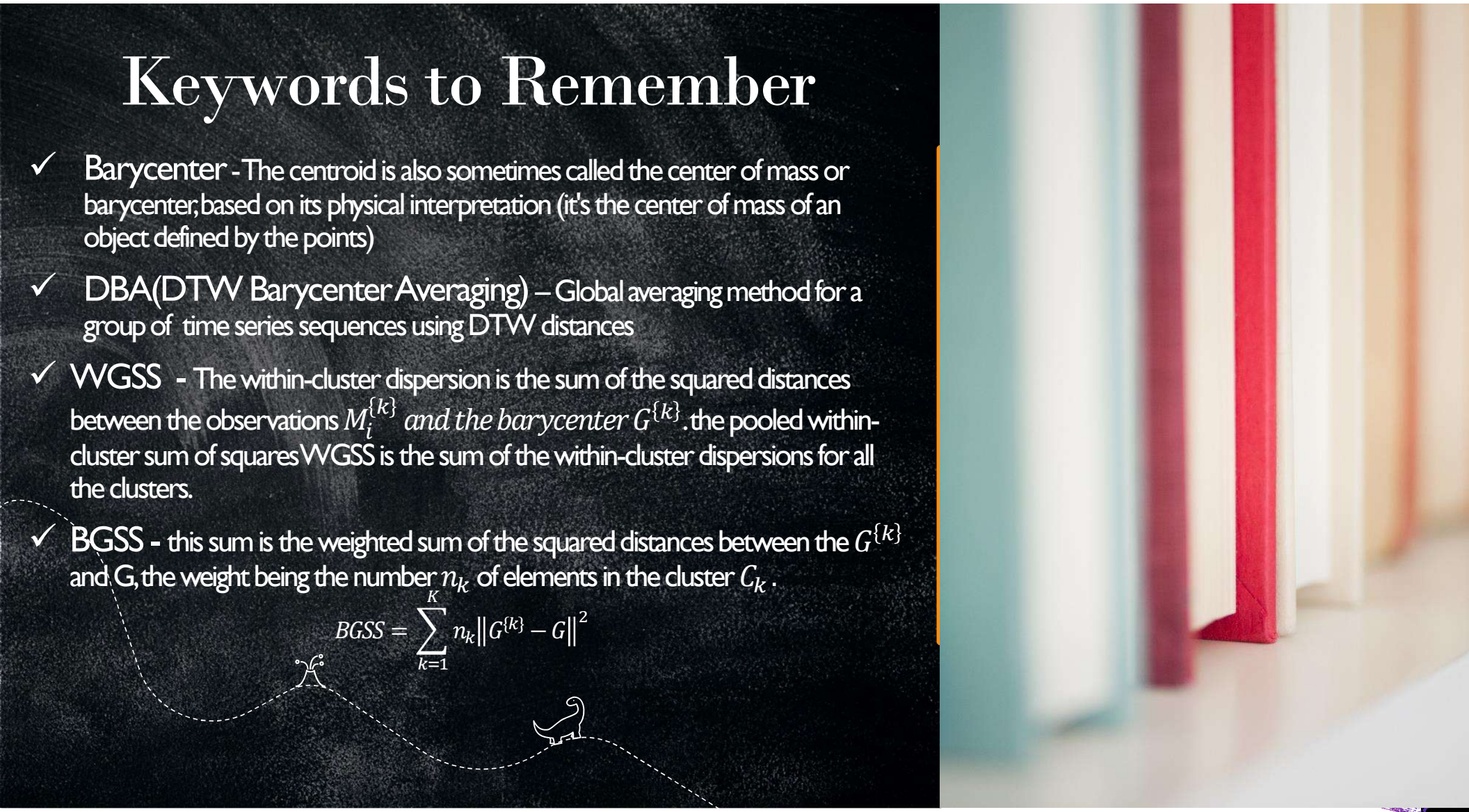
- **Ham:** The ham data includes measurements from 19 Spanish and 18 French dry-cured hams.
- **Meat:** Food spectrographs are used in chemometrics to classify food types, a task that has obvious applications in food safety and quality assurance. The classes are chicken, pork and turkey.
- **SonyAIBORobotSurface:** The robot has roll/pitch/yaw accelerometers. The task is to detect the surface being walked on which is cement or carpet.
- **ToeSegmentation2:** Motions in the database containing the keyword walk are classified by their motion descriptions into two categories. The first is the normal walk, with only walk in the motion descriptions. The other is the abnormal walk, with the motion descriptions containing hobble walk, walk wounded leg, walk on toes bent forward, hurt leg walk, drag bad leg walk, or hurt stomach walk
- **Wine:** This dataset involves wine classification by spectrograph



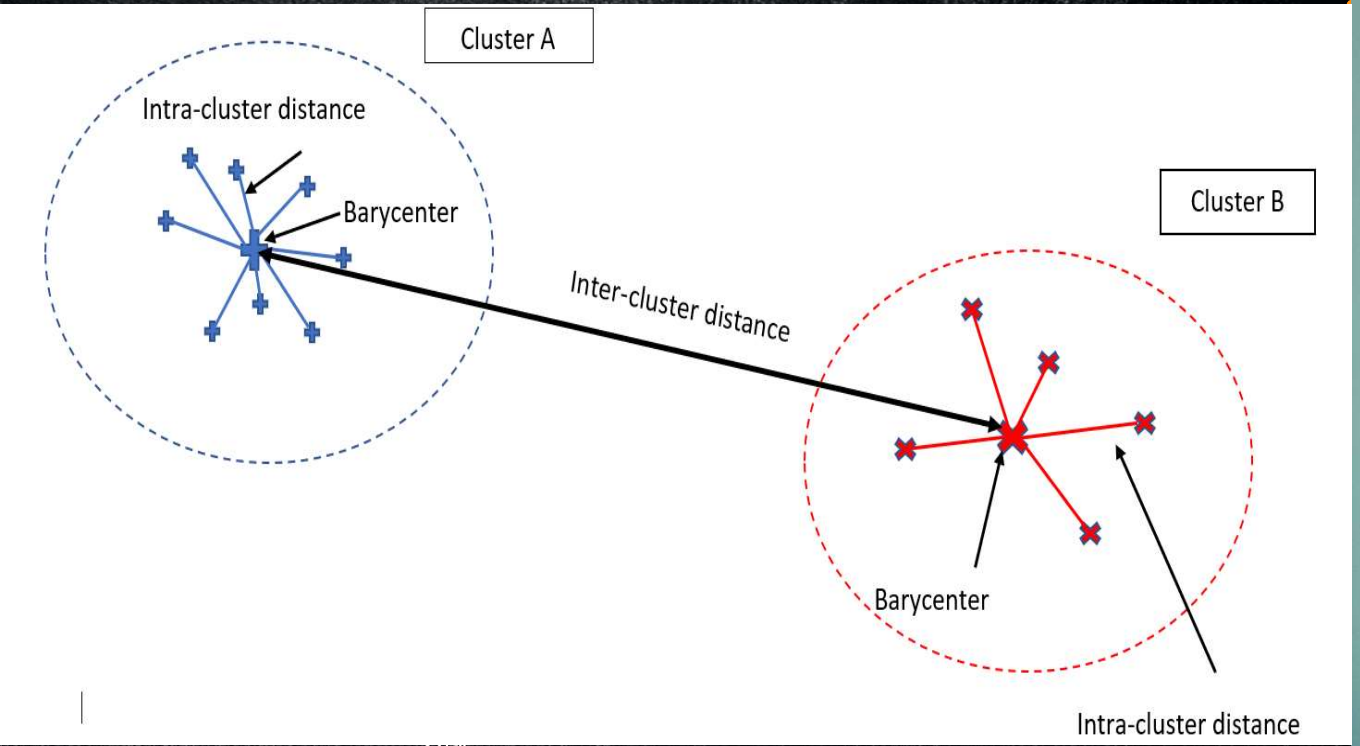
Keywords to Remember

- ✓ **Barycenter** - The centroid is also sometimes called the center of mass or barycenter, based on its physical interpretation (it's the center of mass of an object defined by the points)
- ✓ **DBA(DTW Barycenter Averaging)** – Global averaging method for a group of time series sequences using DTW distances
- ✓ **WGSS** - The within-cluster dispersion is the sum of the squared distances between the observations $M_i^{\{k\}}$ *and the barycenter* $G^{\{k\}}$. the pooled within-cluster sum of squares WGSS is the sum of the within-cluster dispersions for all the clusters.
- ✓ **BGSS** - this sum is the weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C_k.

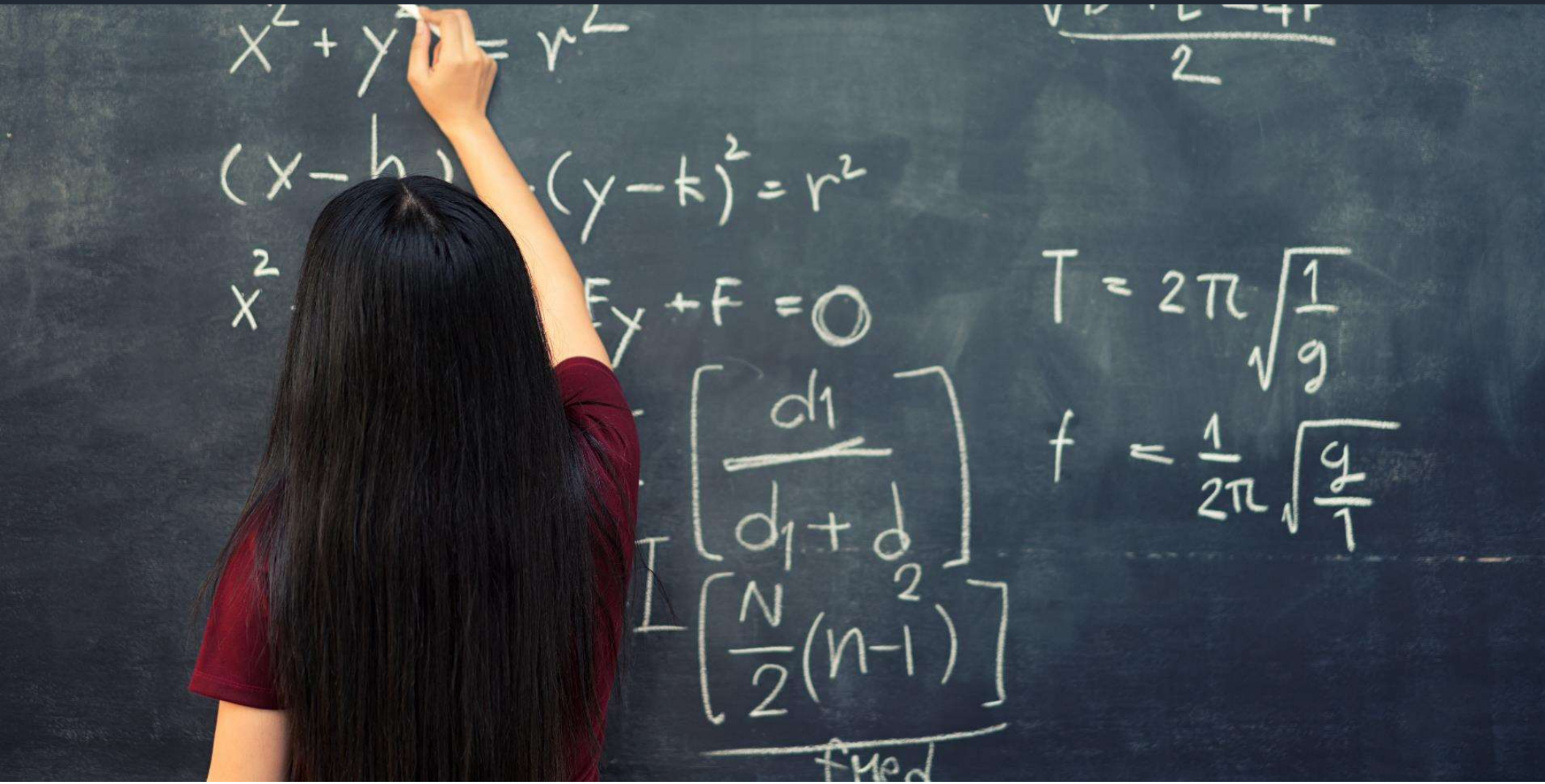
$$BGSS = \sum_{k=1}^{K} n_k \left\| G^{\{k\}} - G \right\|^2$$

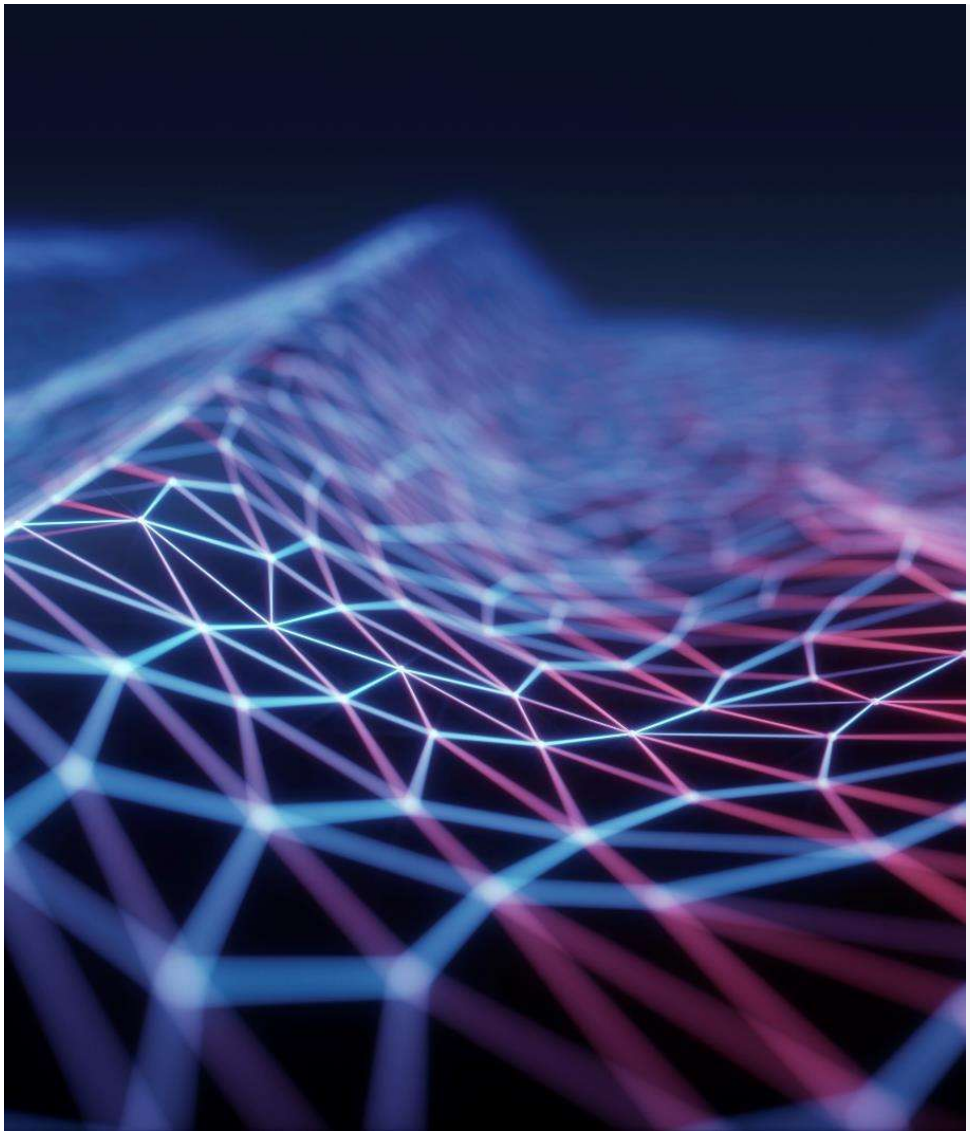


Distances under consideration



Algorithms and Experimental Results





Ball Hall Index

- The mean dispersion of a cluster is the mean of the squared distances of the points of the cluster with respect to their barycenter.
- The Ball-Hall index is the mean, through all the clusters, of their mean dispersion
- The one with the maximum difference is the best method

Mathematically speaking

Ball Hall Index is

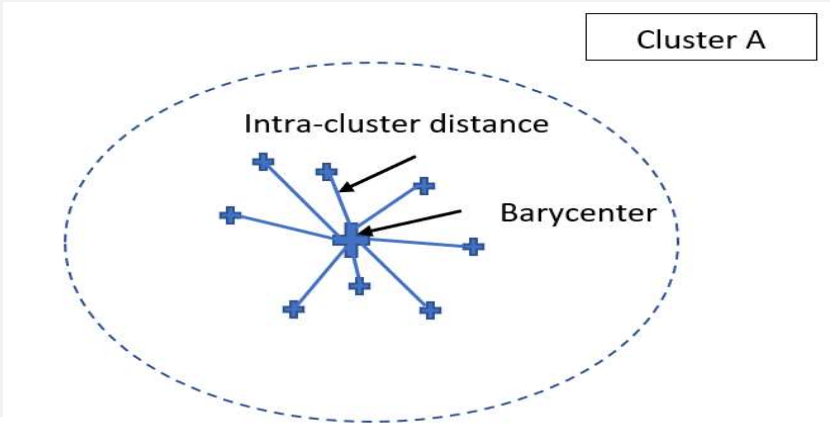
$$C = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i \in I_k}\left\|M_i^{\{k\}} - G^{\{k\}}\right\|^2$$

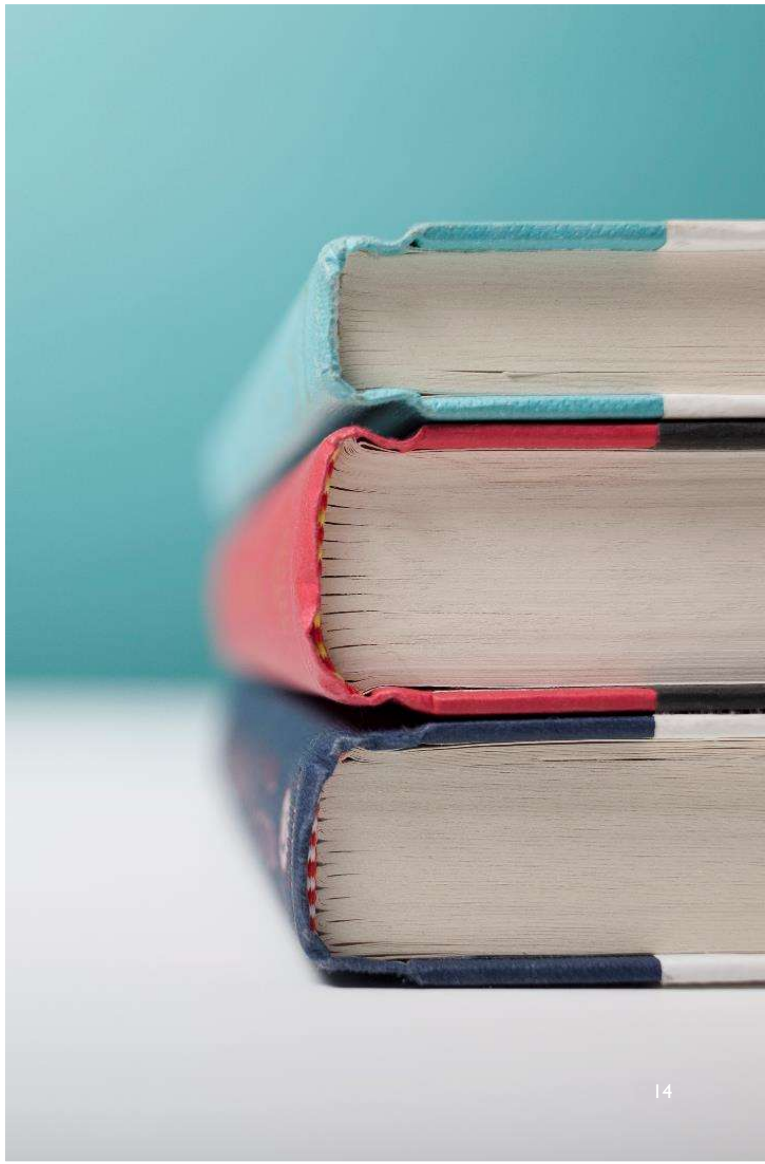
Where K is the number of clusters

n_k number of observations in cluster k

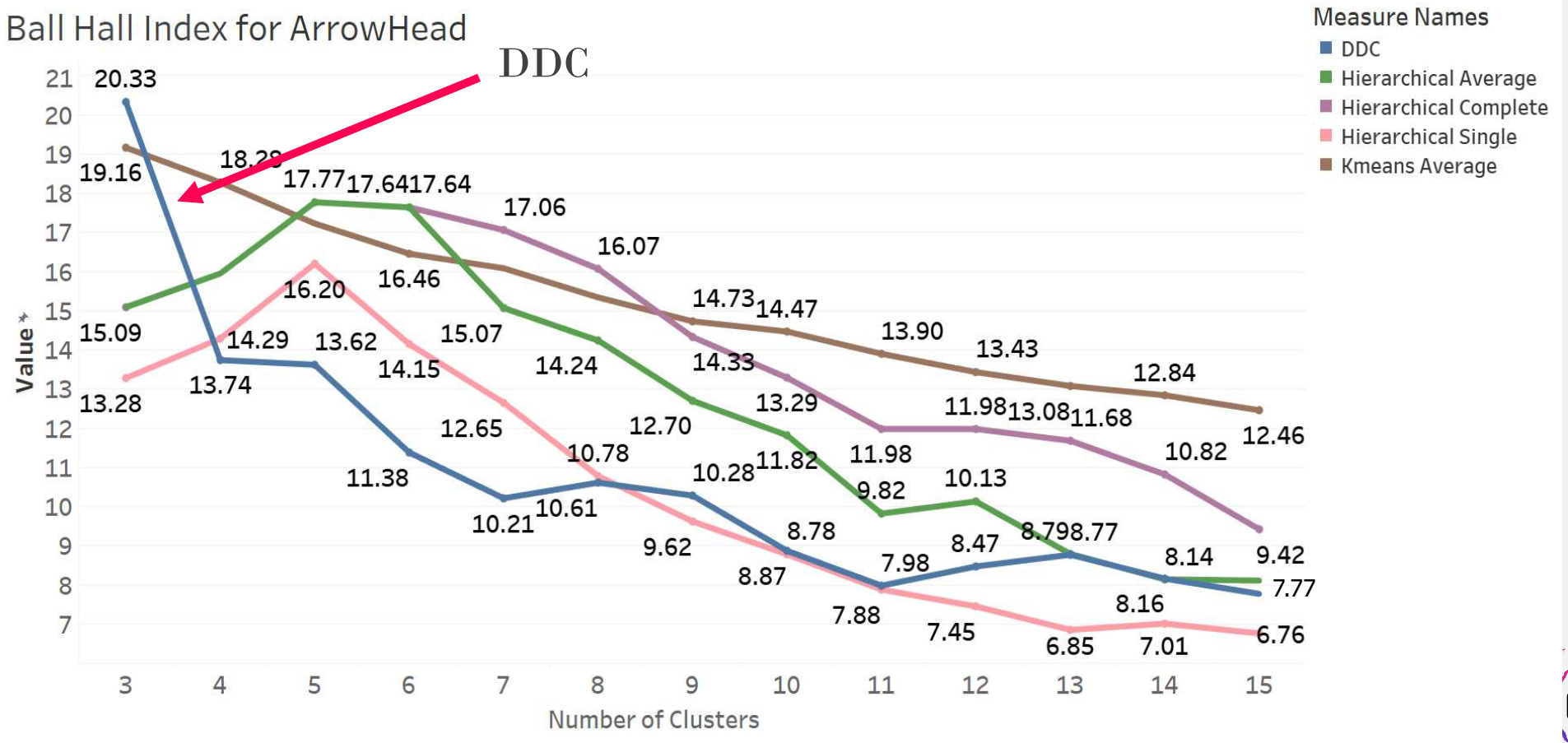
M is a point in the cluster and,

G is the barycenter of the cluster

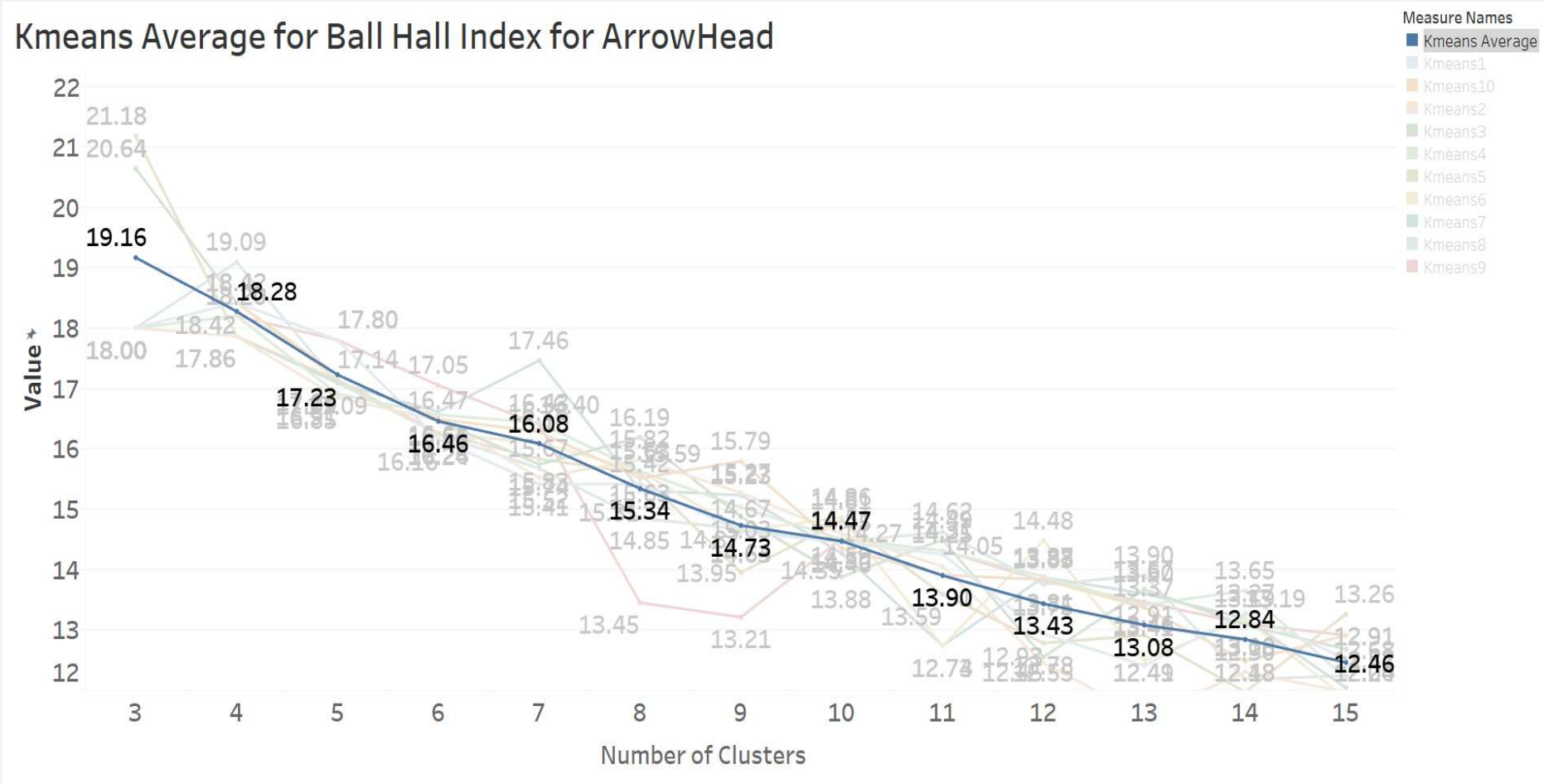




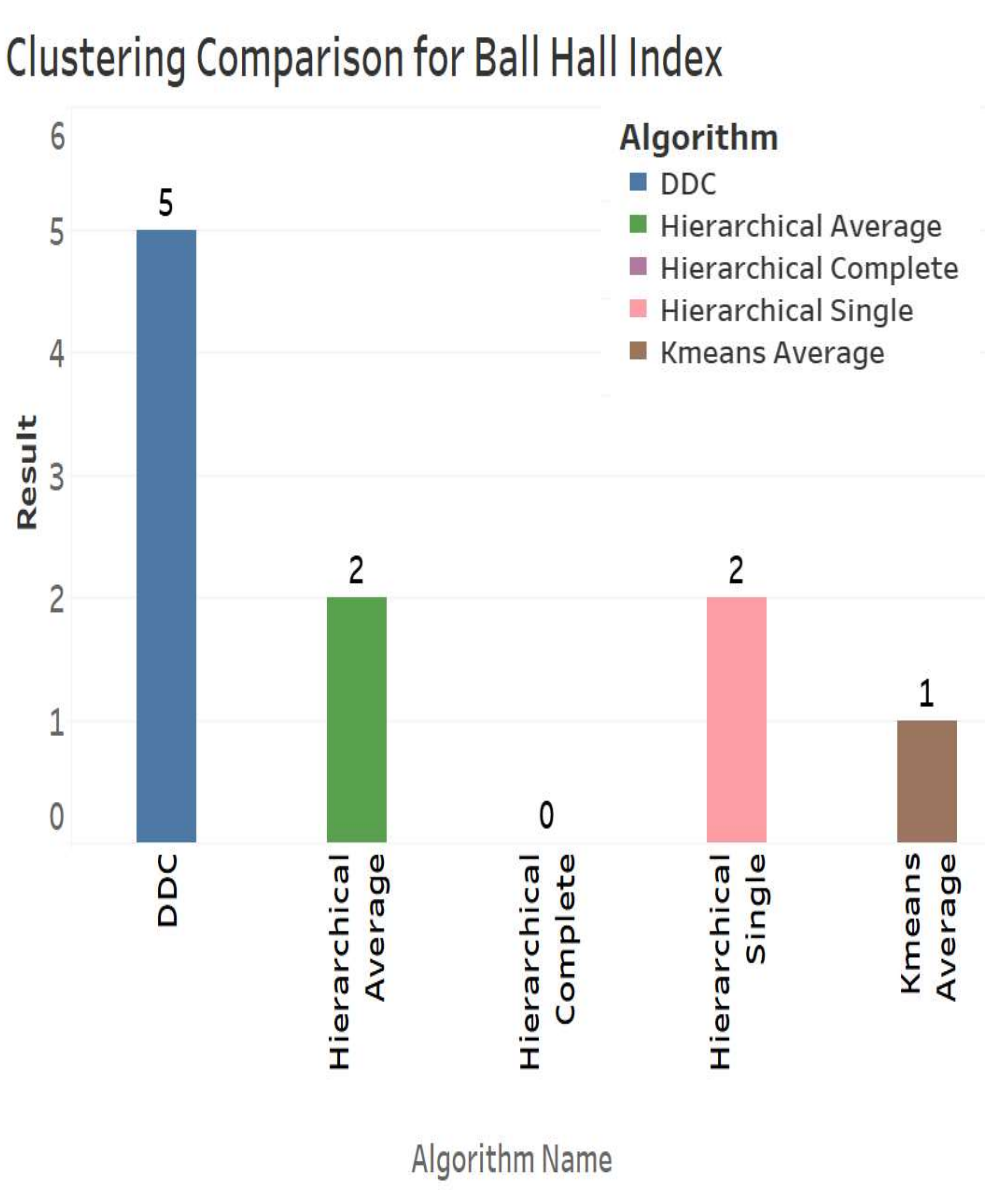
Experimental Result

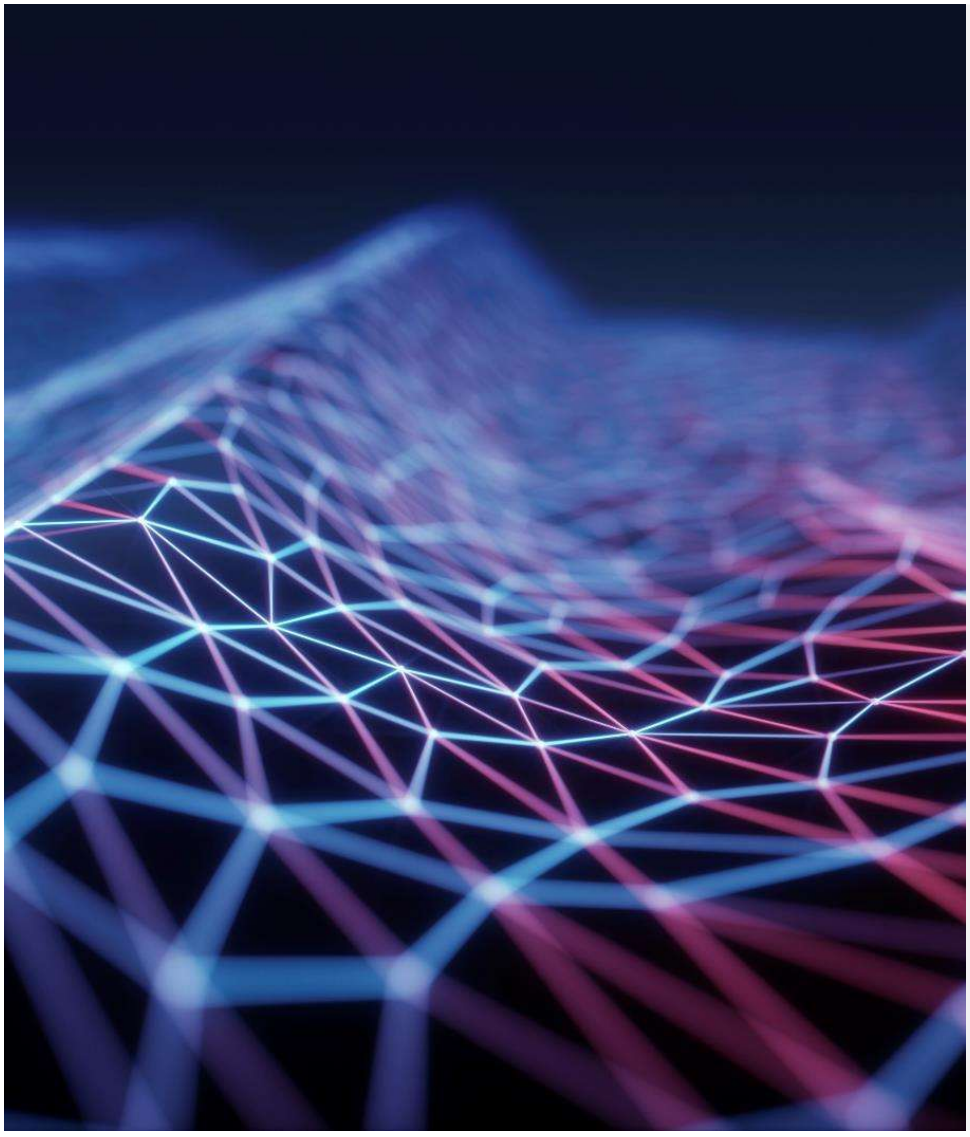


Kmeans Averaging



Clustering Algorithm Comparison





Banfeld-Raftery Index

- This index is the weighted sum of the logarithms of the traces of the variance-covariance matrix of each cluster.
- The quantity Tr(W $G^{\{k\}}$)/n_k can be interpreted as the mean of the squared distances between the points in cluster C_k and their barycenter $G^{\{k\}}$
- The one with the minimum value is the best clustering here.

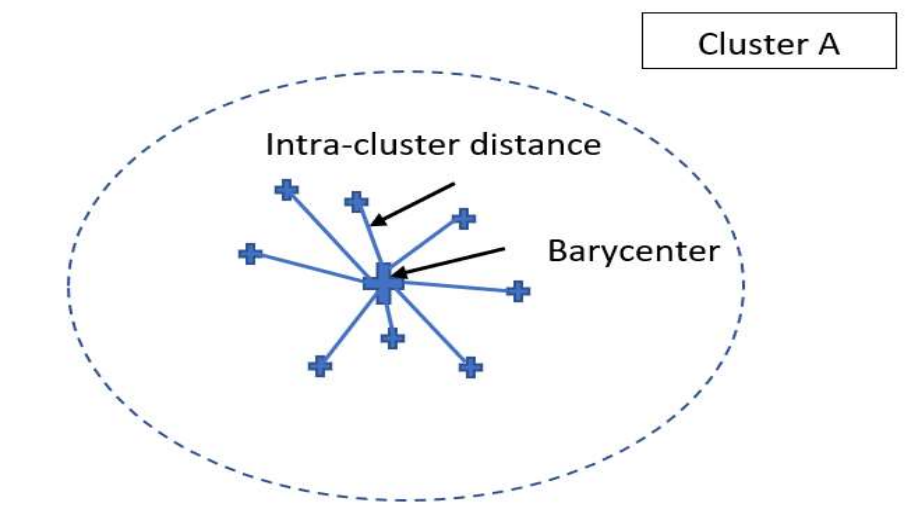
Mathematically speaking

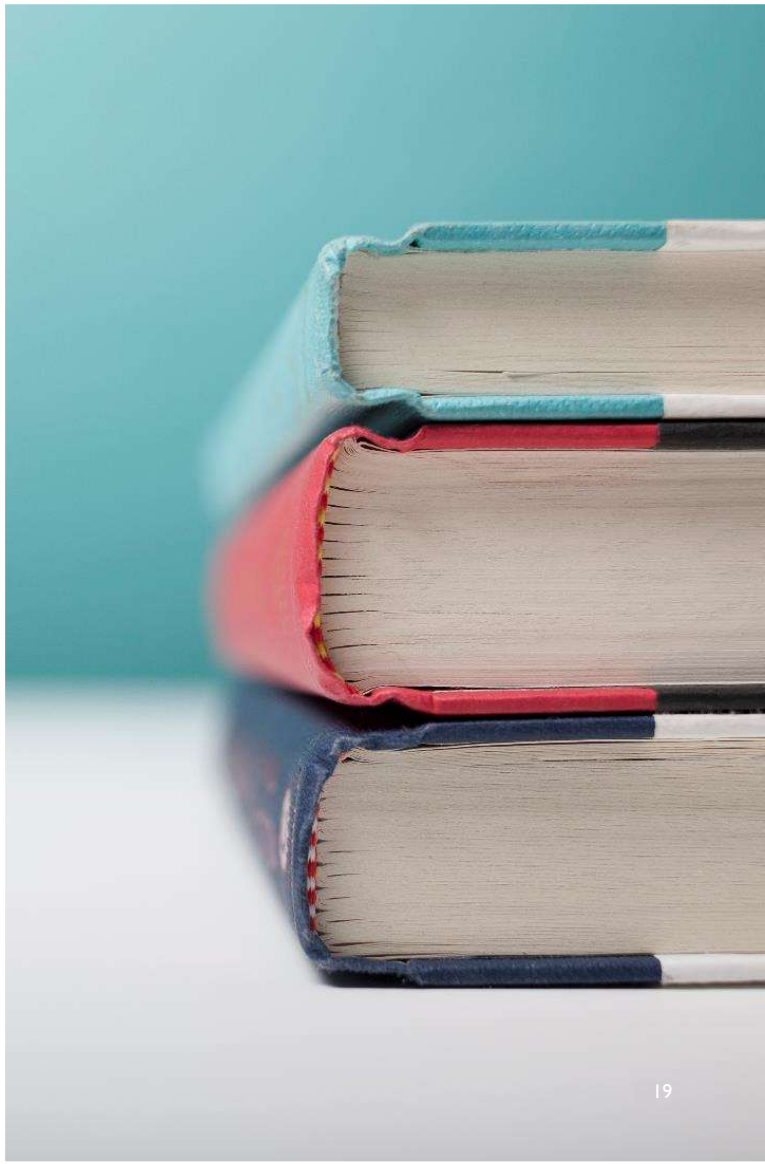
Banfeld-Raftery index is

$$C = \sum_{k=1}^{K} n_k \log\left(\frac{Tr\left(WG^{\{k\}}\right)}{n_k}\right)$$

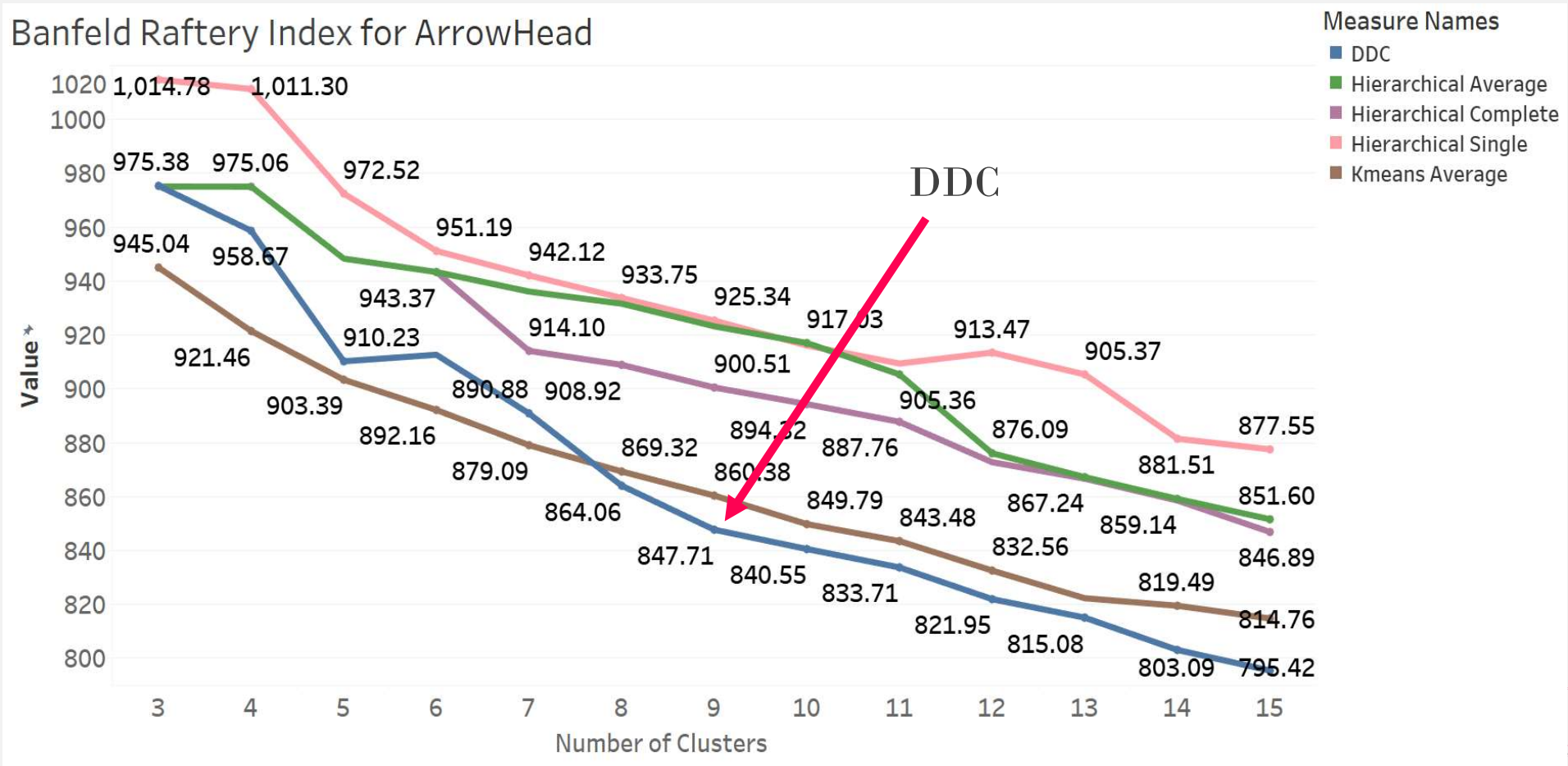
Where K is the number of clusters

$WG^{\{k\}}$ is the Within Group Scatter Matrix

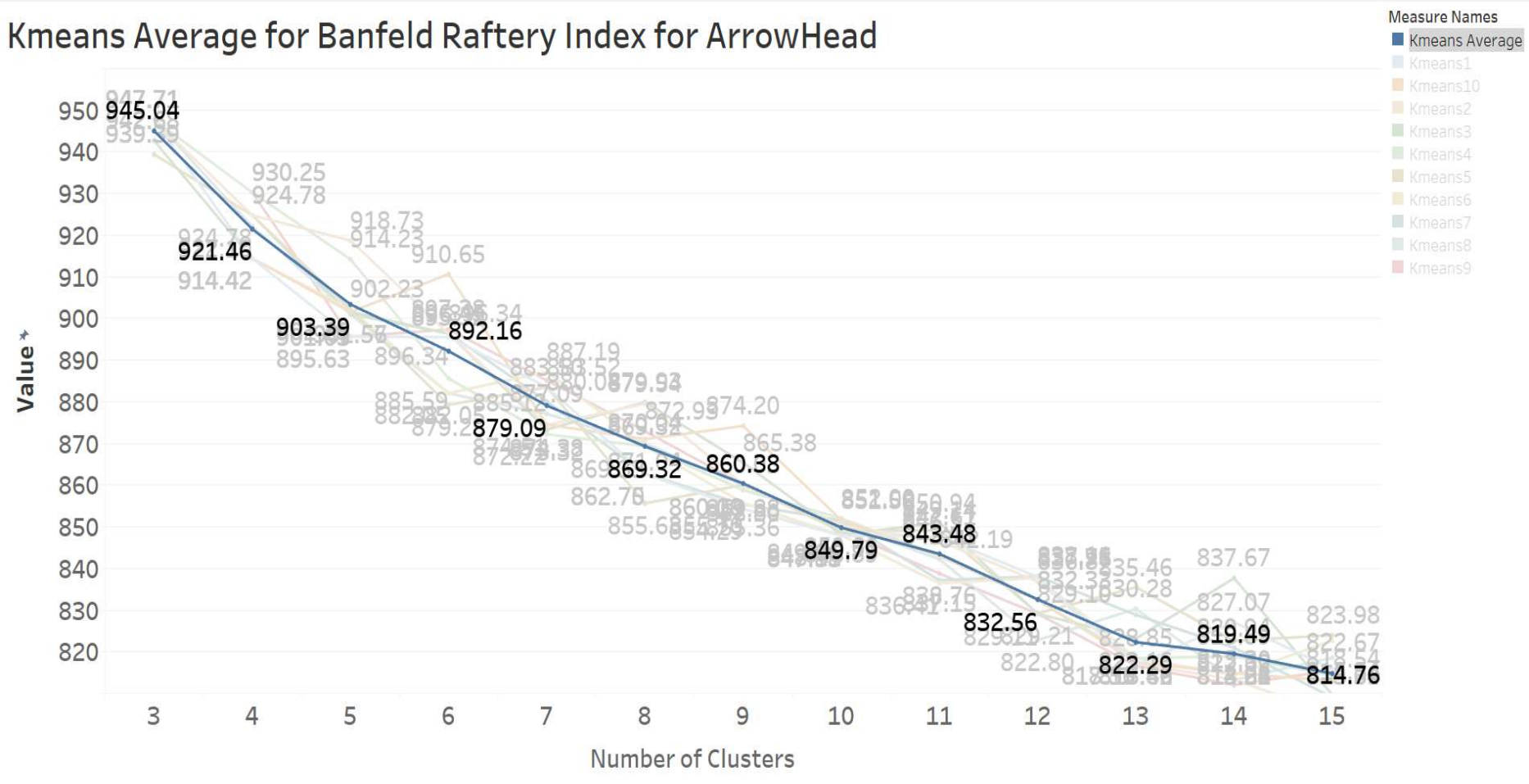




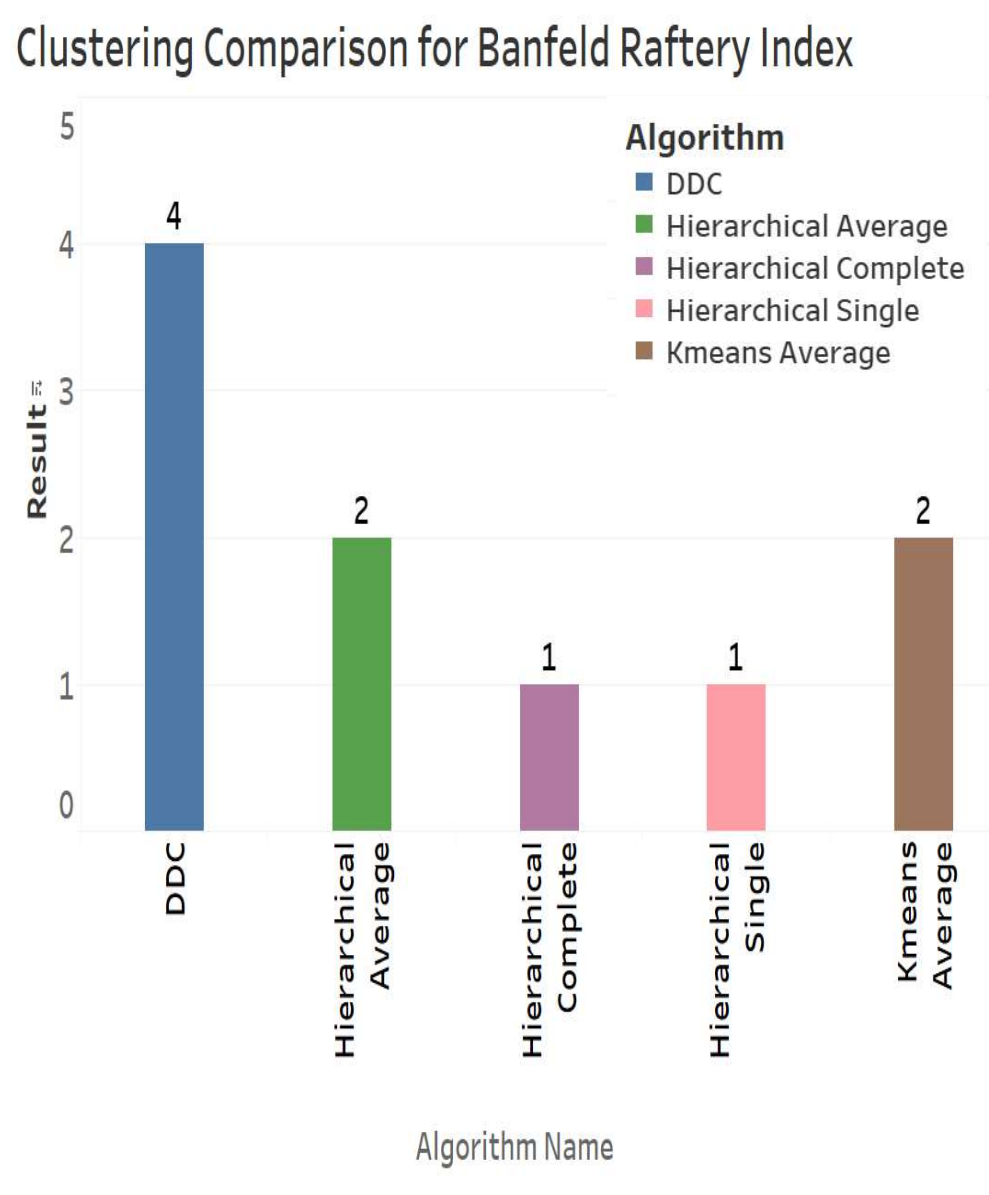
Experimental Result

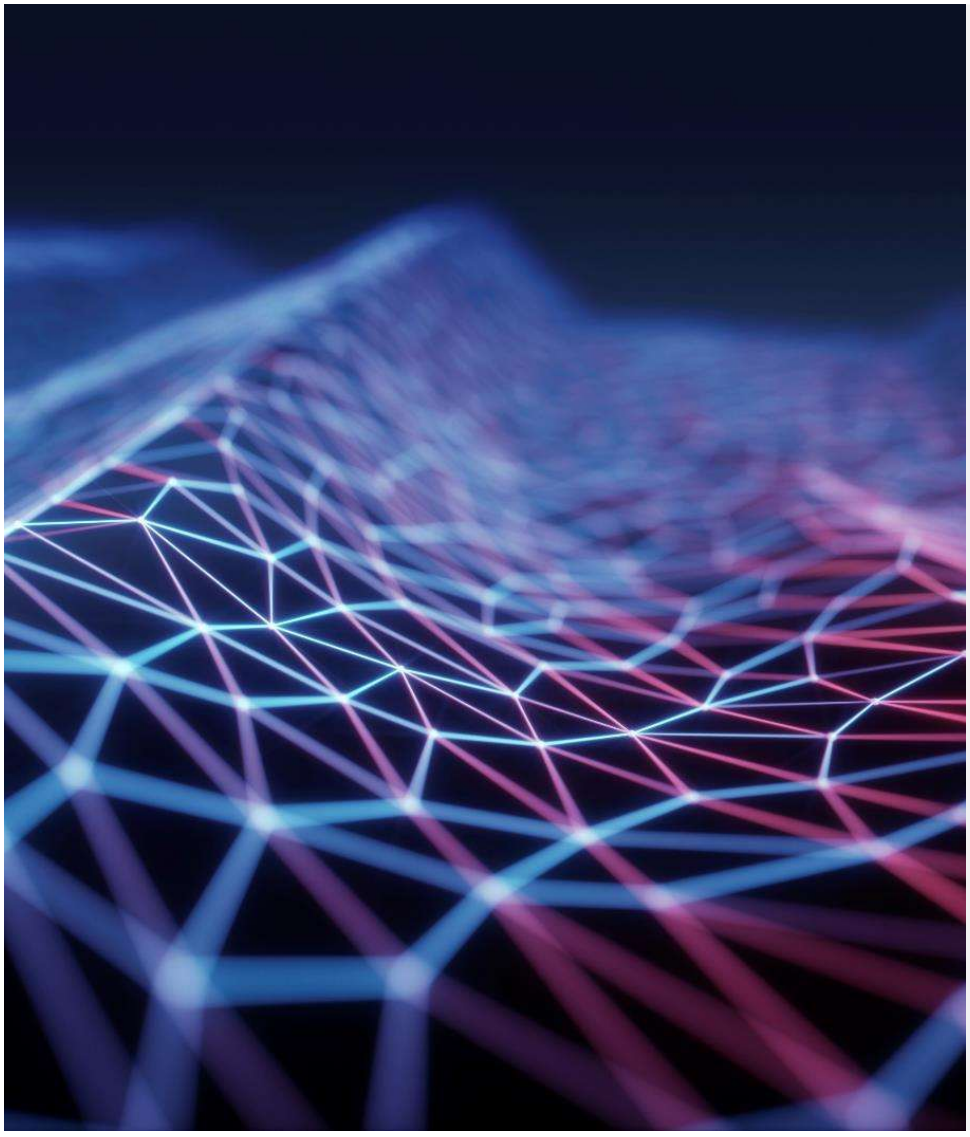


Kmeans Averaging



Clustering Algorithm Comparison





Calinski-Harabasz Index

- This index is given as the ratio of the weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C_k and the sum of the squared distances between the observations $M^{\{k\}}$ i and the barycenter $G^{\{k\}}$ of the cluster.

- This is divided by the ratio of number of cluster subtracted from total number of clusters divided by one less than total number of clusters.

- The bigger the index value, the better is the clustering.

Mathematically speaking

Calinski-Harabasz Index is,

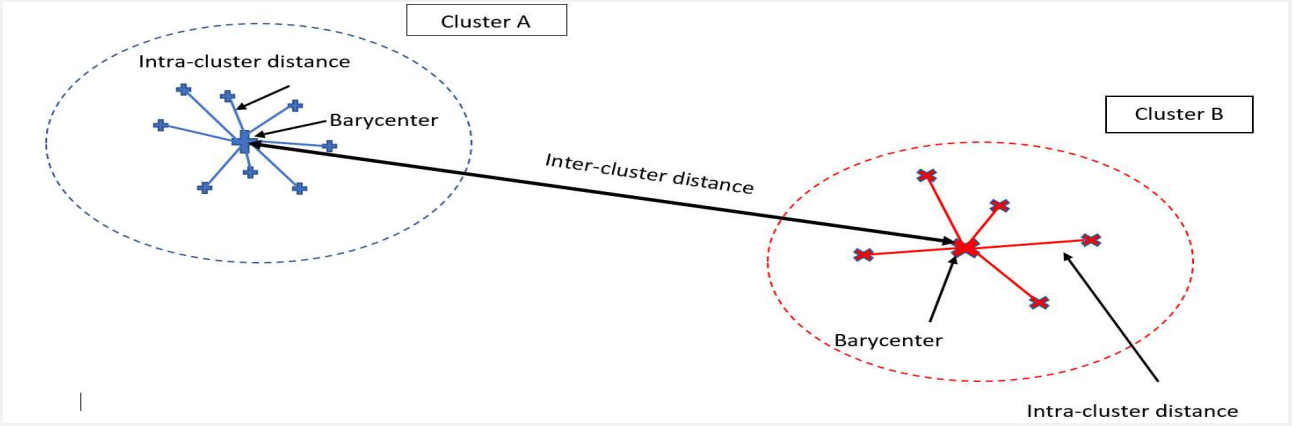
$$C = \frac{BGSS}{WGSS}\frac{N-K}{K-1}$$

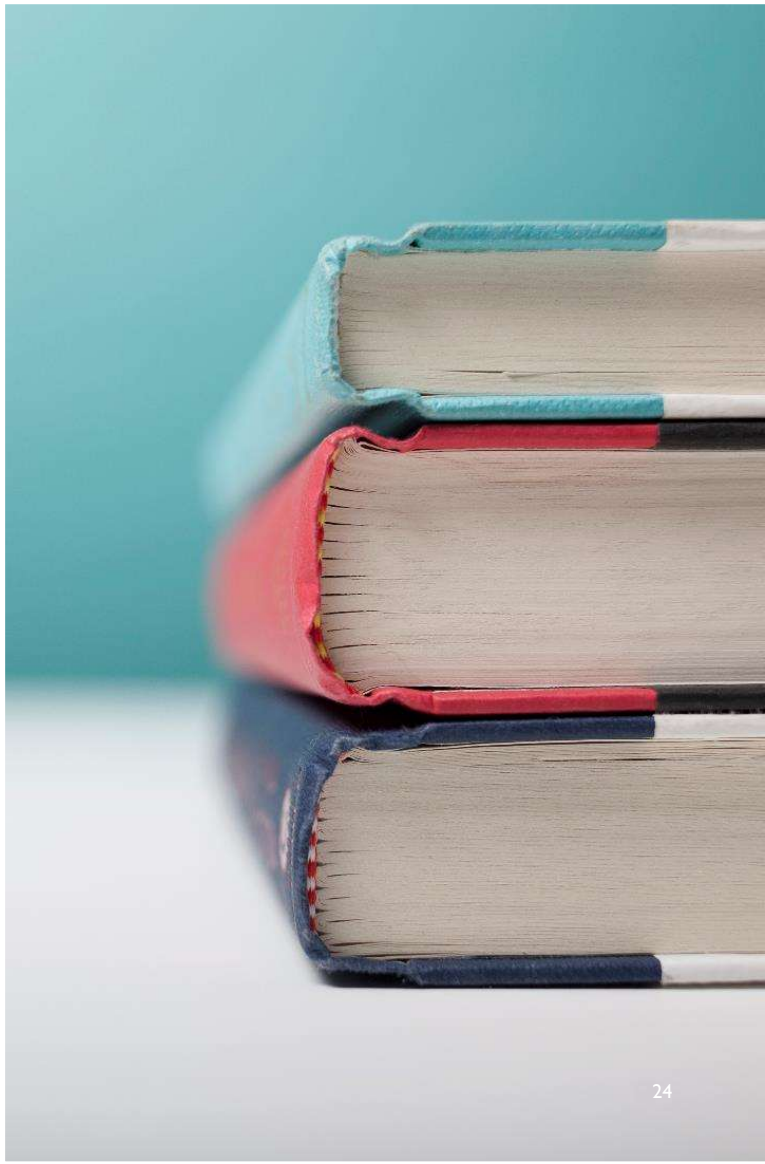
Where K is the number of clusters

N is the total number of observations

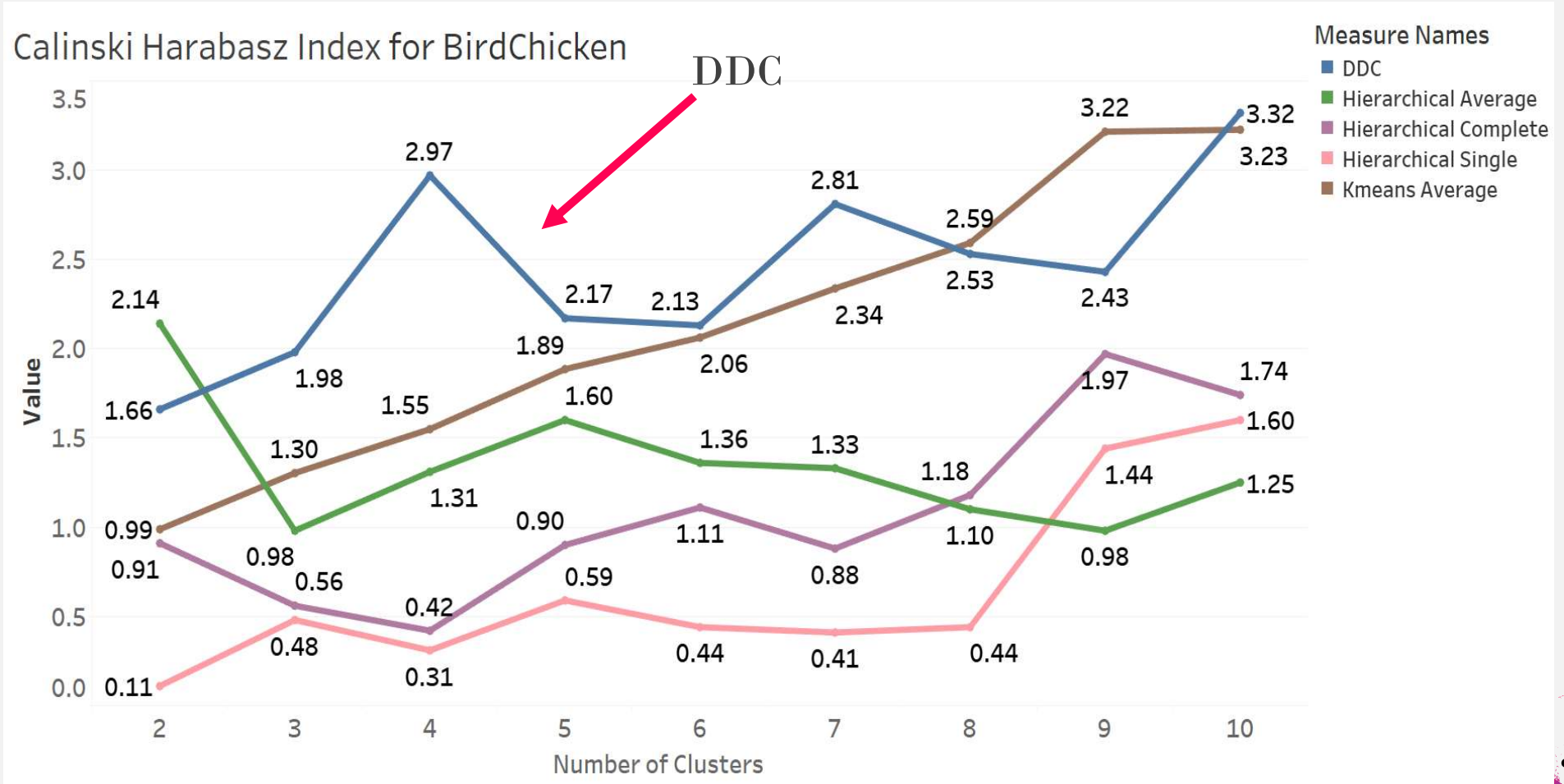
$BGSS$ is weighted sum of the squared distances between the $G^{\{k\}}$ and G, the weight being the number n_k of elements in the cluster C_k

WGSS is the sum of the squared distances between the observations $M^{\{k\}}$ i and the barycenter $G^{\{k\}}$ of the cluster

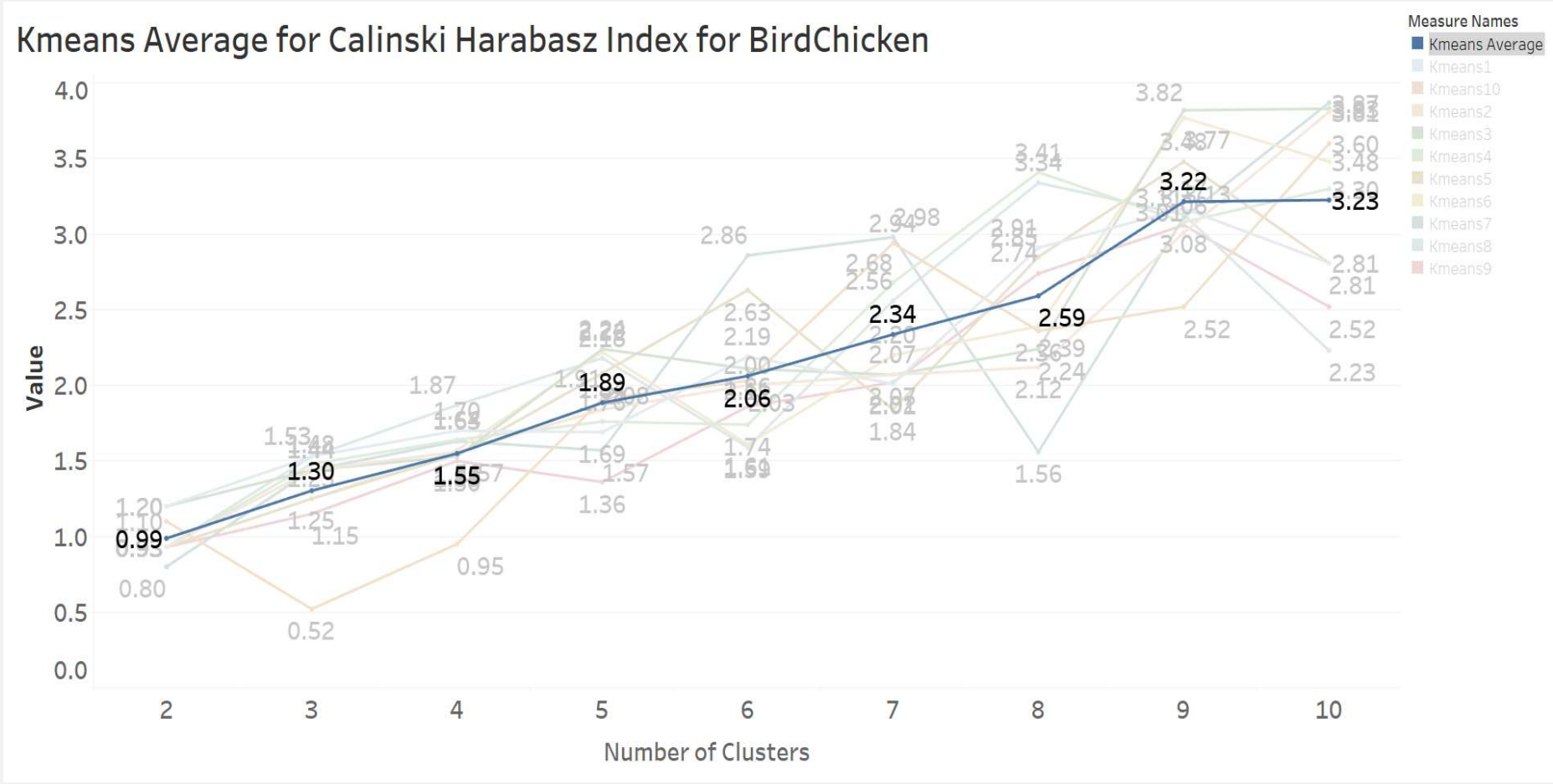




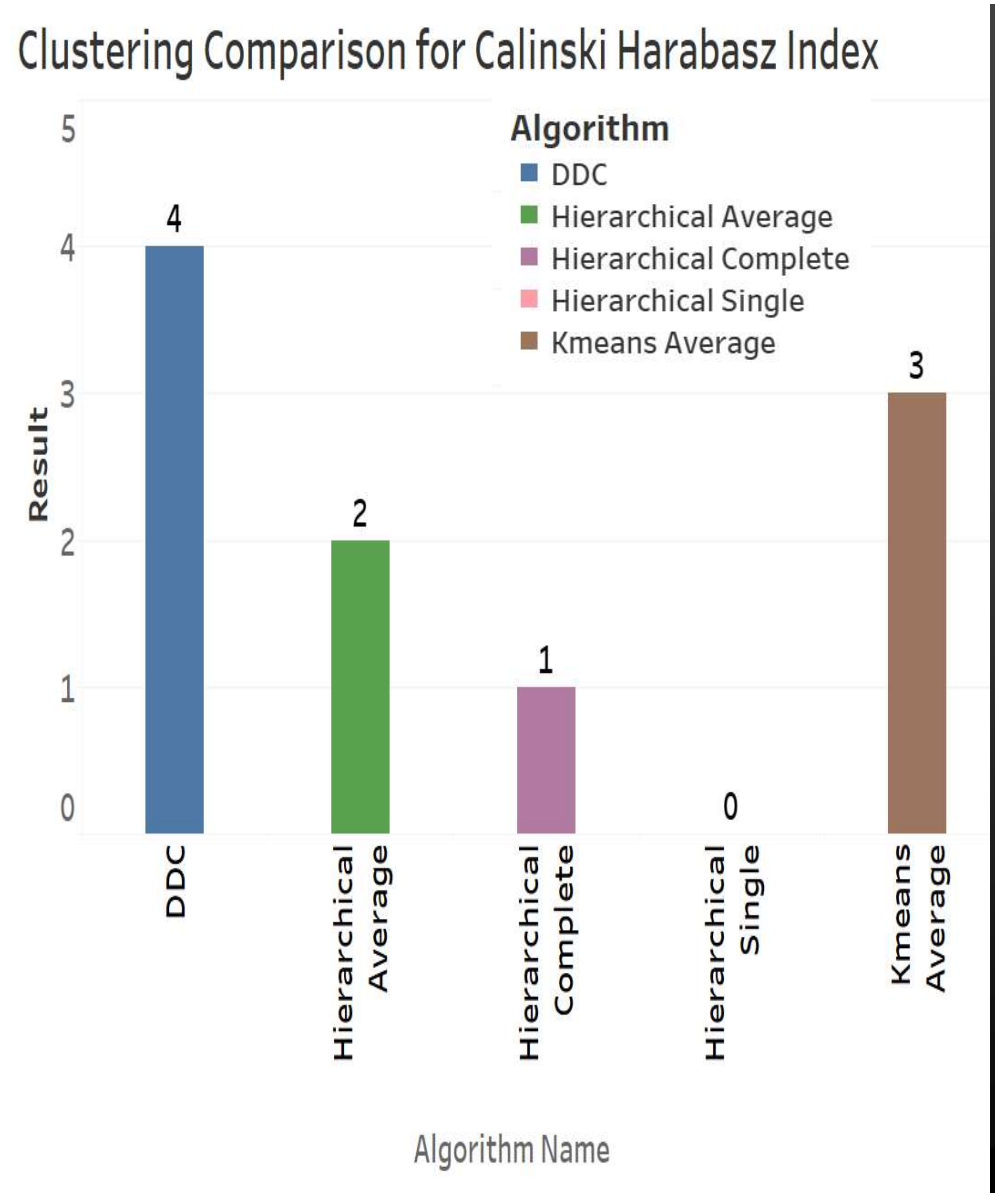
Experimental Result

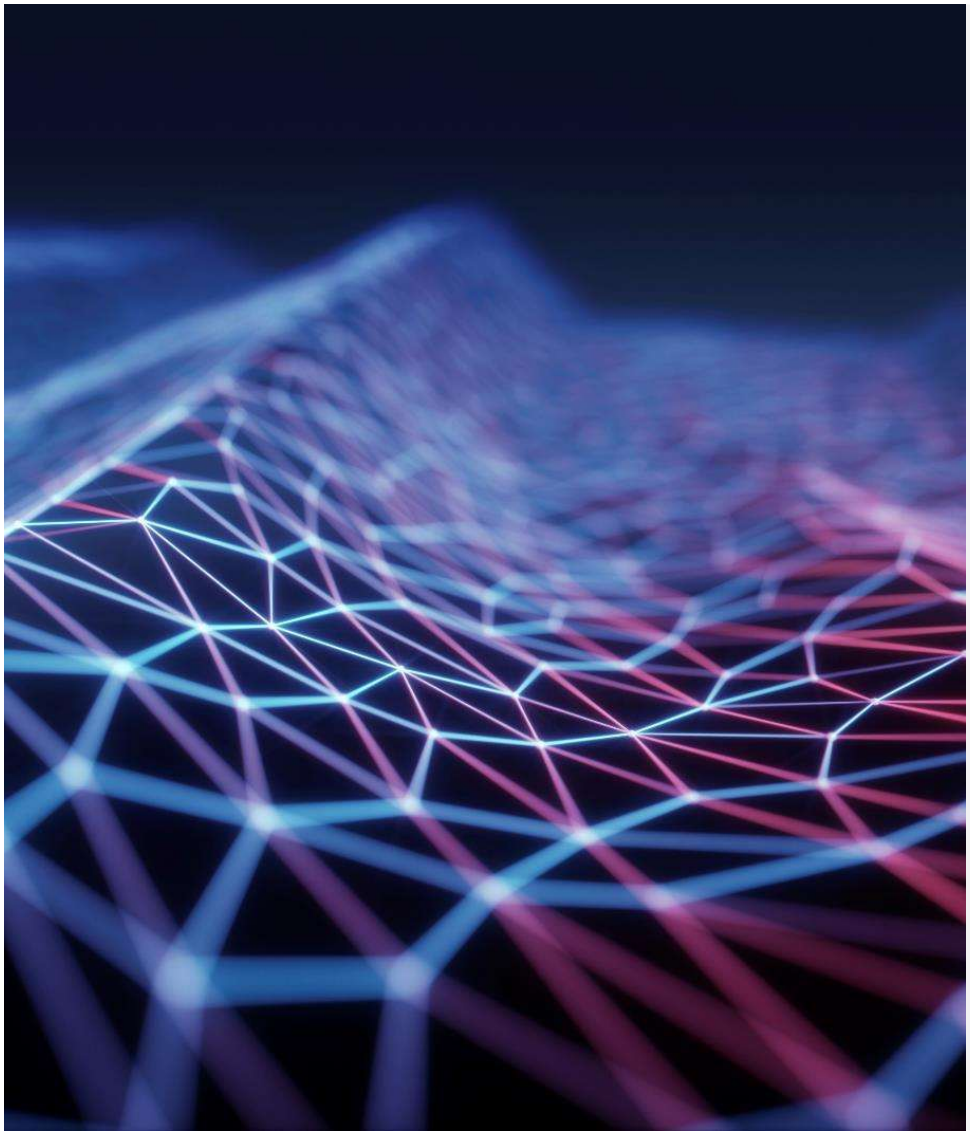


Kmeans Averaging



Clustering Algorithm Comparison





Davies-Bouldin Index

- Summation of mean value of points in cluster C_k to its barycenter $G^{\{k\}}$ and $C_{k'}$ to its barycenter $G^{\{k\prime\}}$ divided by distance between the barycenters
- The Mean value of the maximum of the above value among all the clusters is the Davies-Bouldin Index.
- The lower the value, the better clustering it is

Mathematically speaking

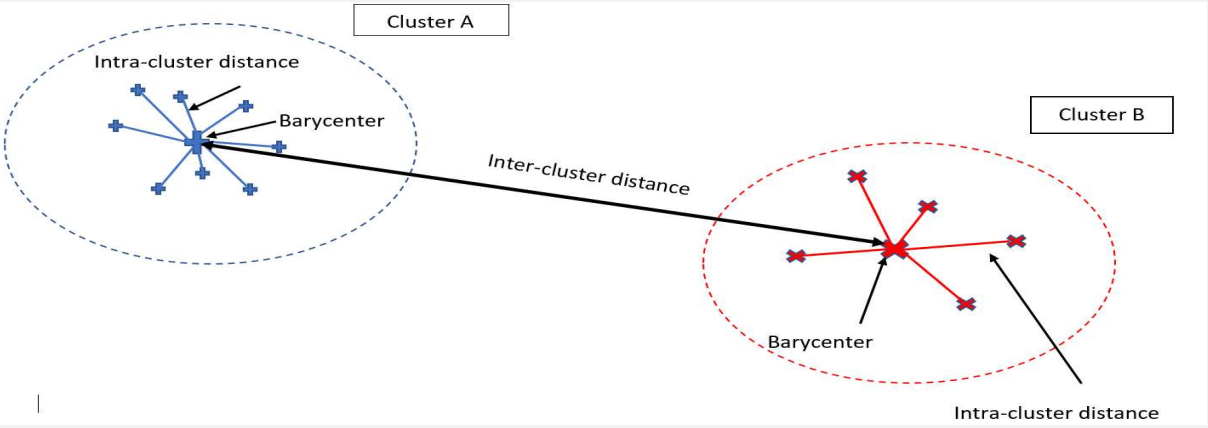
Davies-Bouldin Index is,

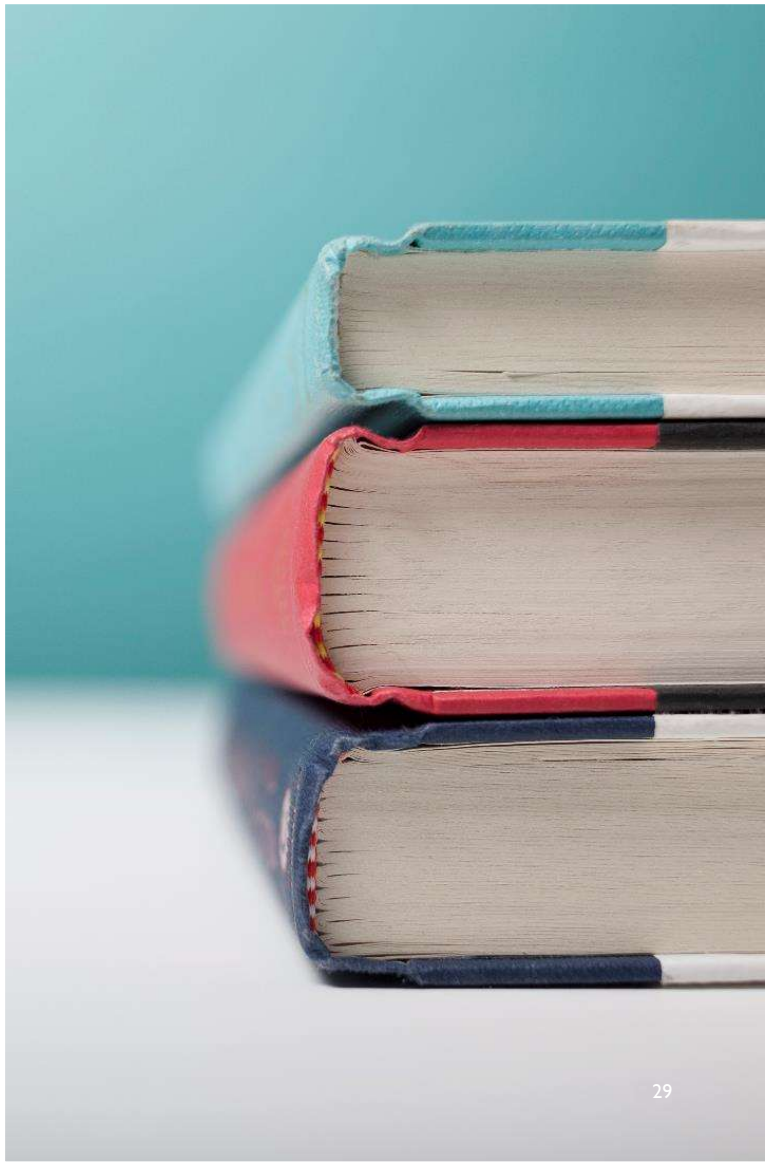
$$C = \frac{1}{K}\sum_{k=1}^{K} M_k = \frac{1}{K}\sum_{k=1}^{K} \max_{k' \neq k}\left(\frac{\delta_k + \delta_{k'}}{\Delta k k'}\right)$$

Where K is the number of clusters

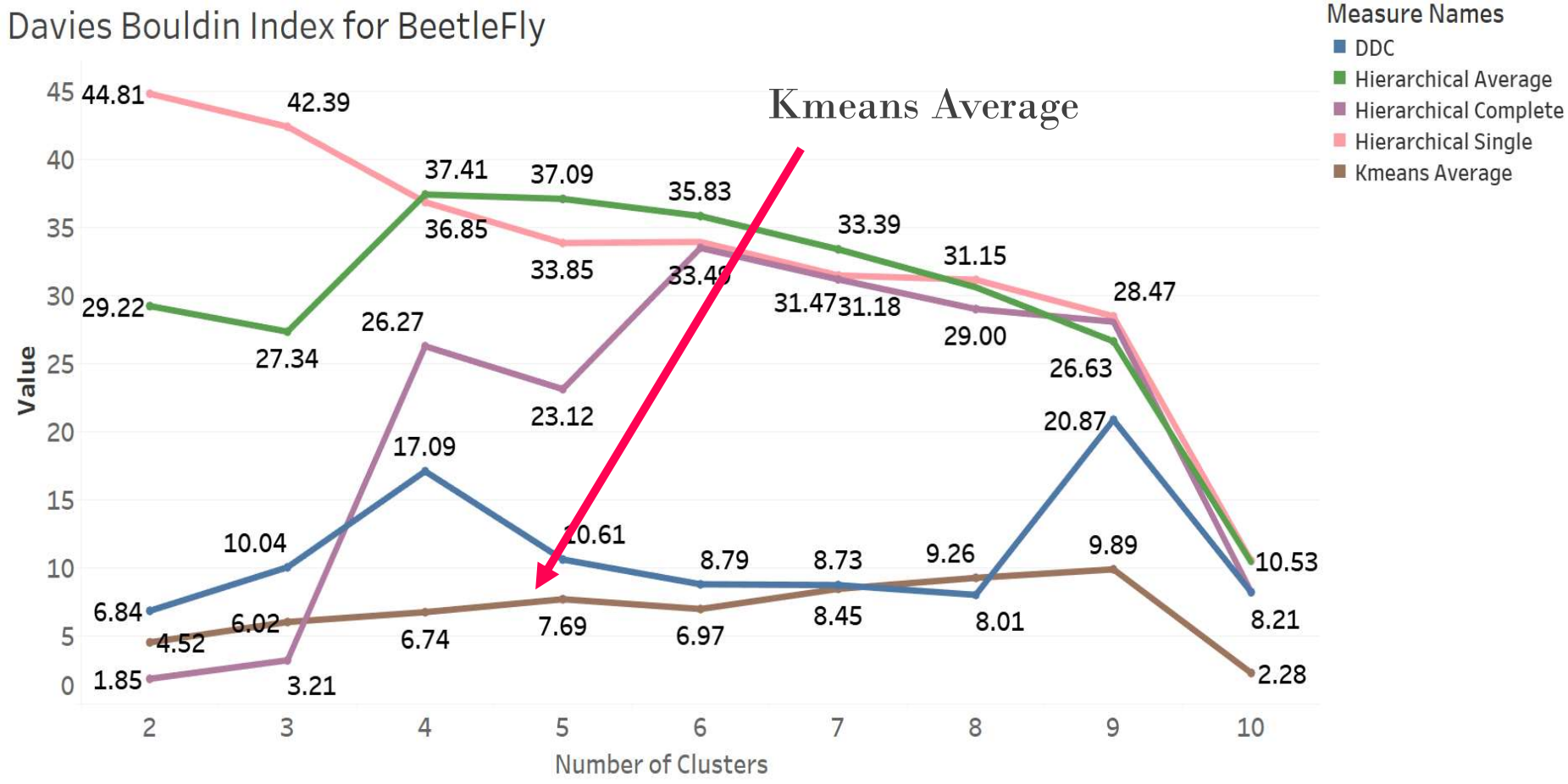
$\delta_k + \delta_{k'}$ is the sum of the mean distance of the points belonging to cluster C_k with their barycenter $G^{\{k\}}$

$\Delta k k'$ is the distance between the barycenters $G^{\{k\}}$ and $G^{\{k'\}}$ of clusters C_k and $C_{k'}$

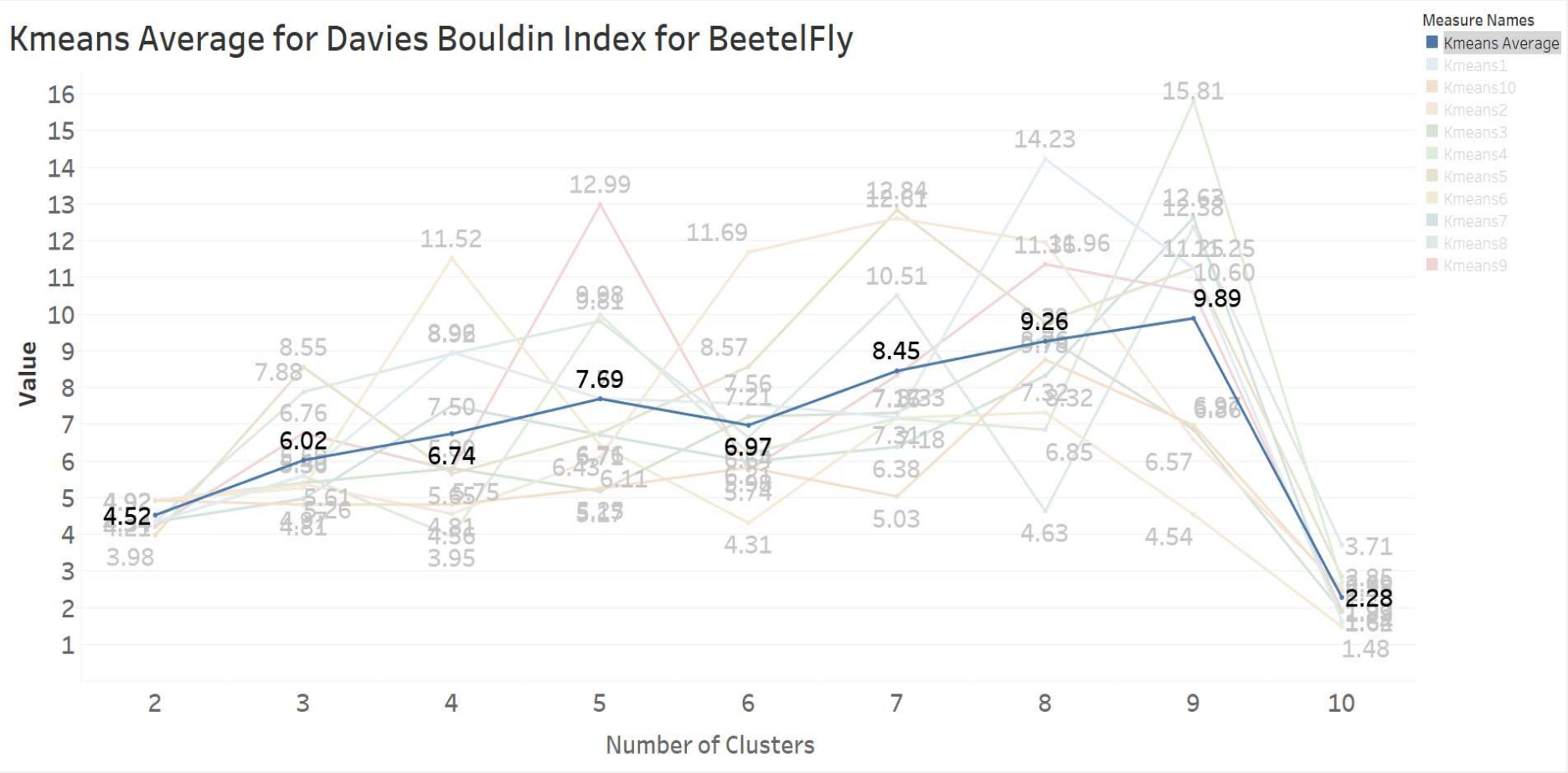




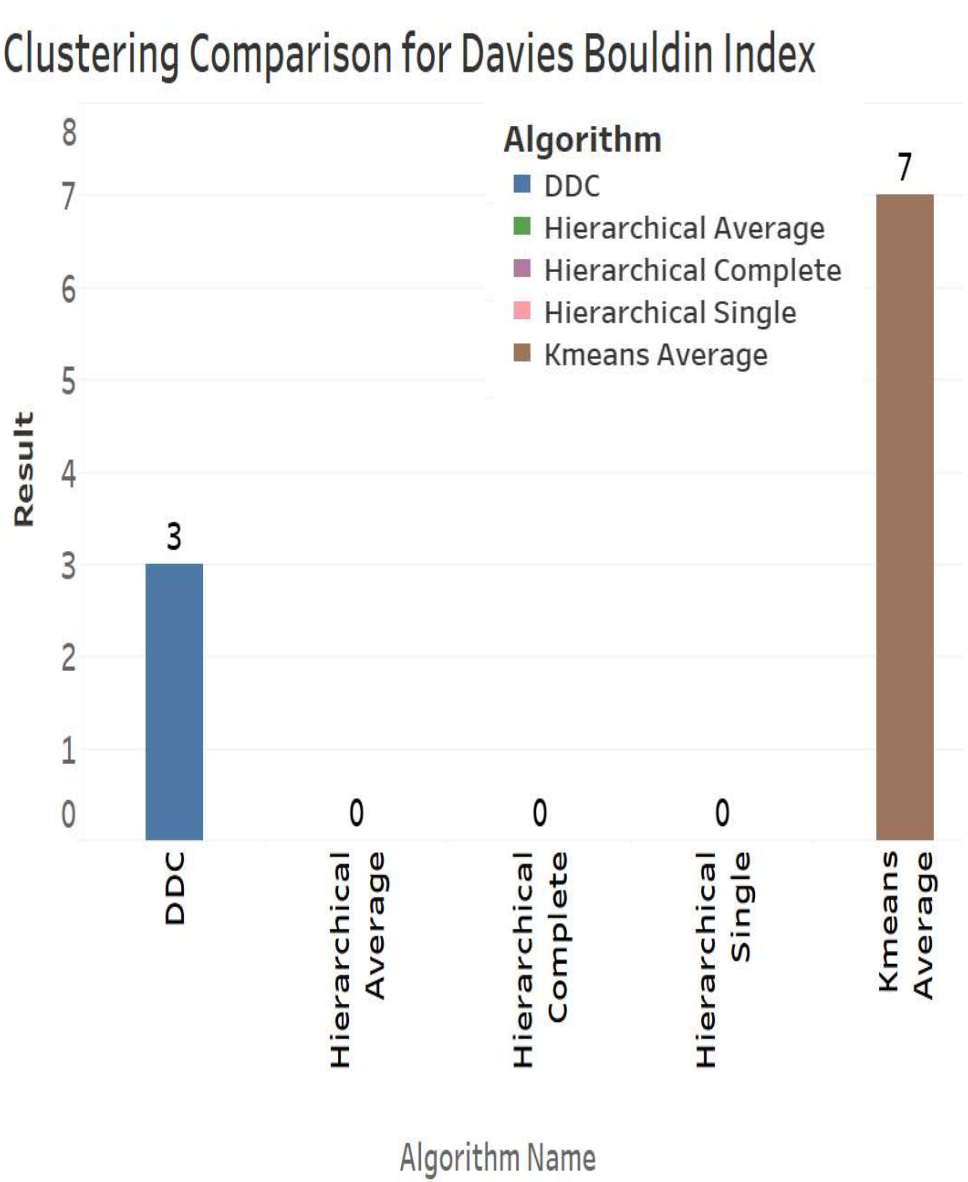
Experimental Result

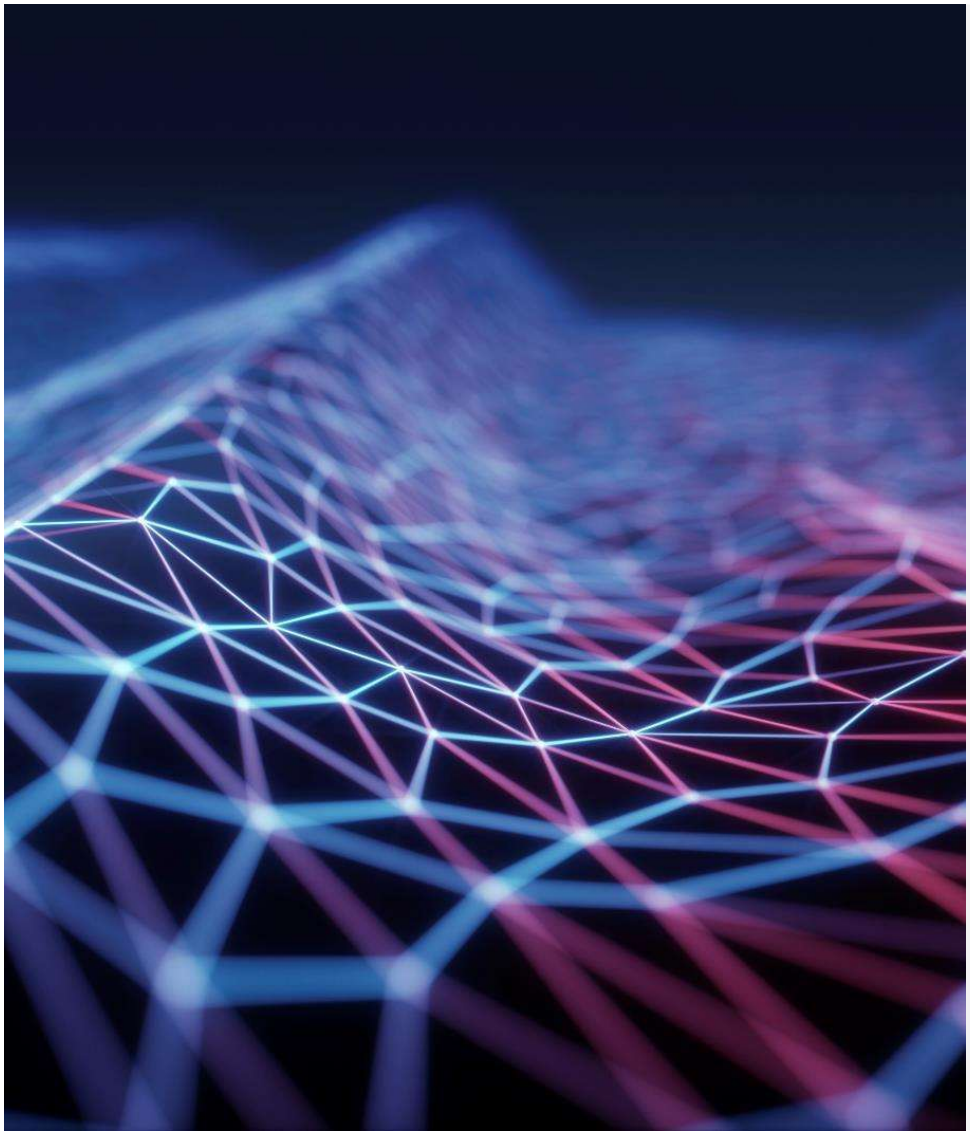


Kmeans Averaging



Clustering Algorithm Comparison





Log SS Ratio Index

- This index is logarithm of the ratio between BGSS and WGSS
- The one with the minimum difference has the best clustering.

Mathematically speaking

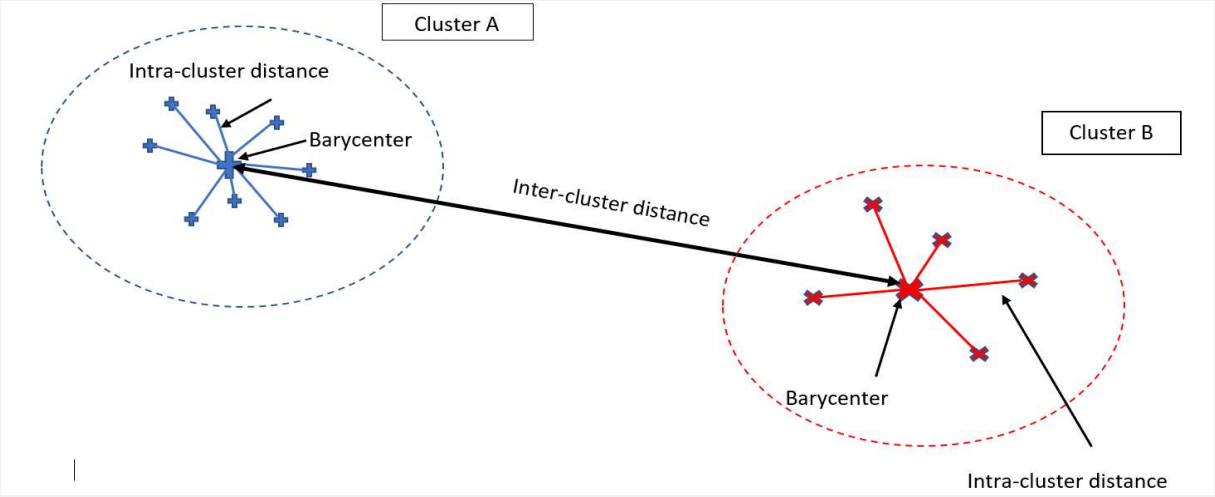
Log SS Ratio Index is,

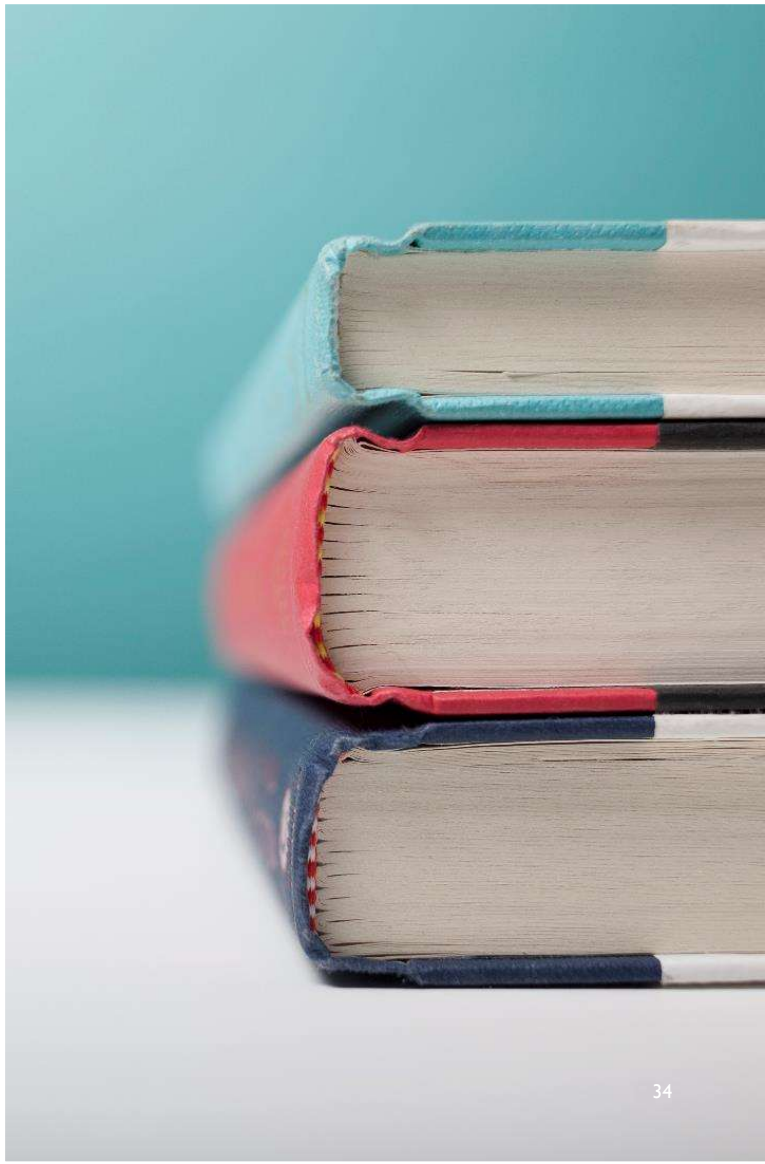
$$C = \sum_{k=1}^{K} \log(\frac{BGSS}{WGSS})$$

Where K is the number of clusters

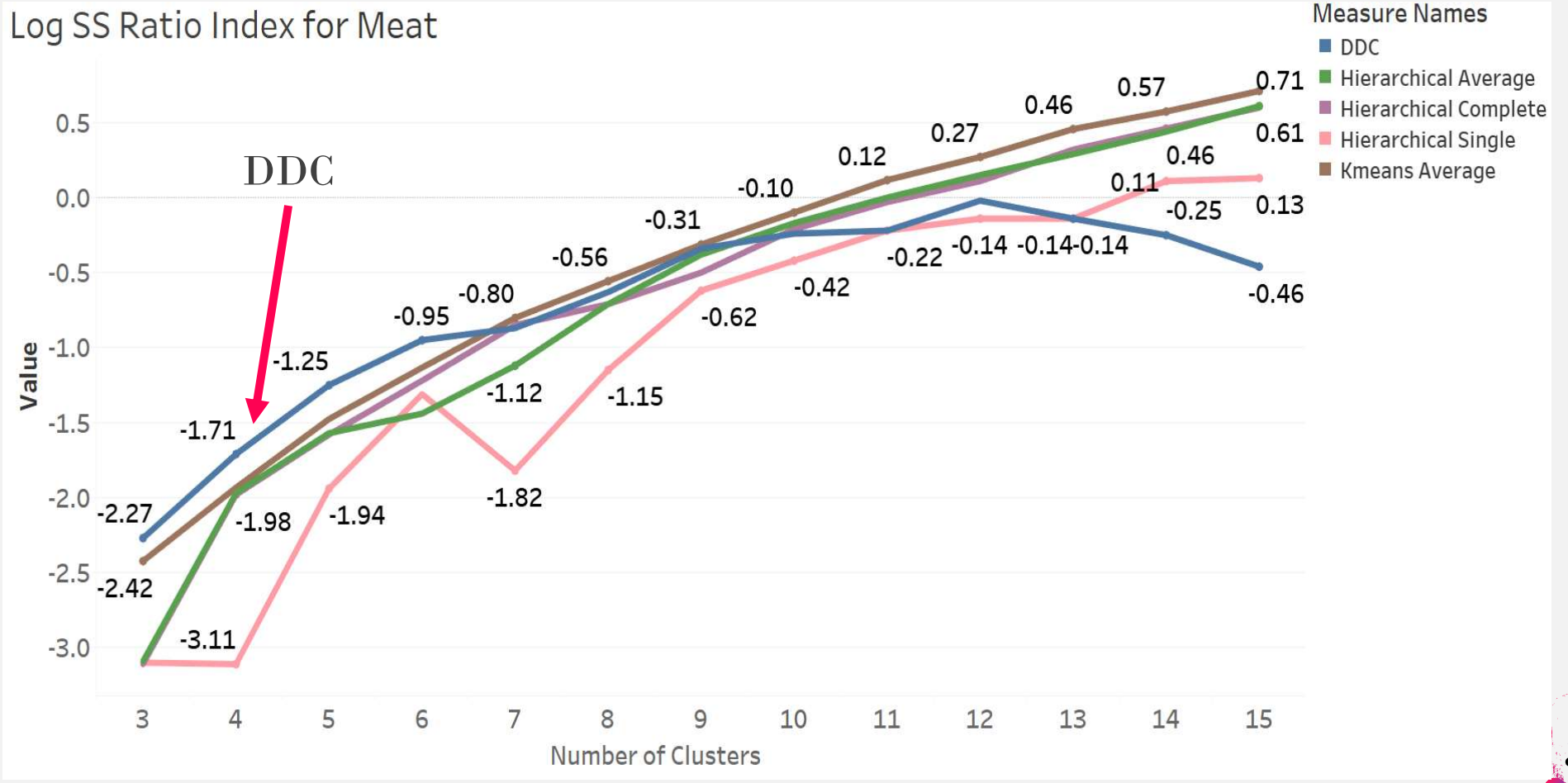
$WGSS$ is the Within Group Sum of Squares

$BGSS$ is the Between Group Sum of Squares

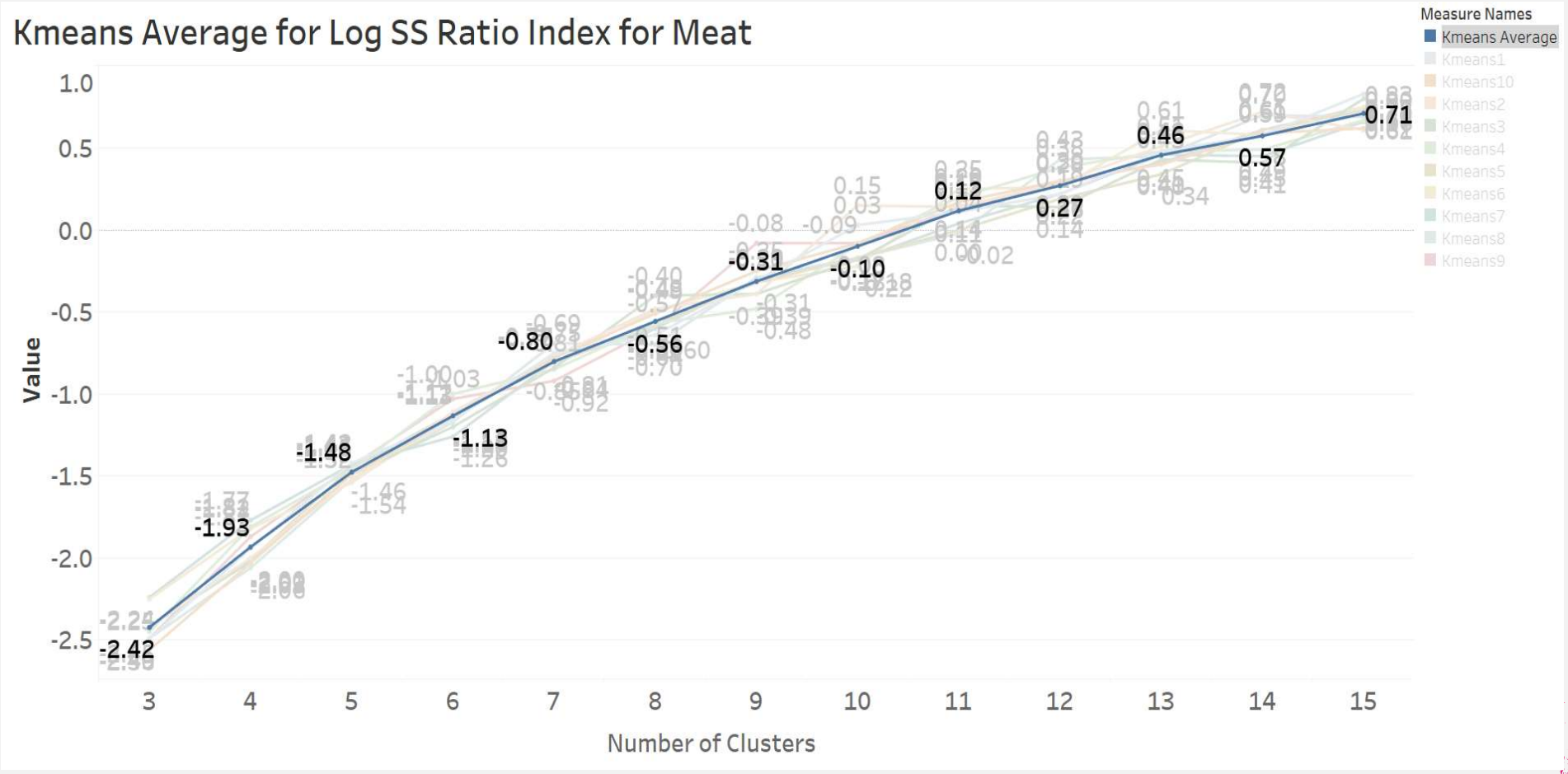




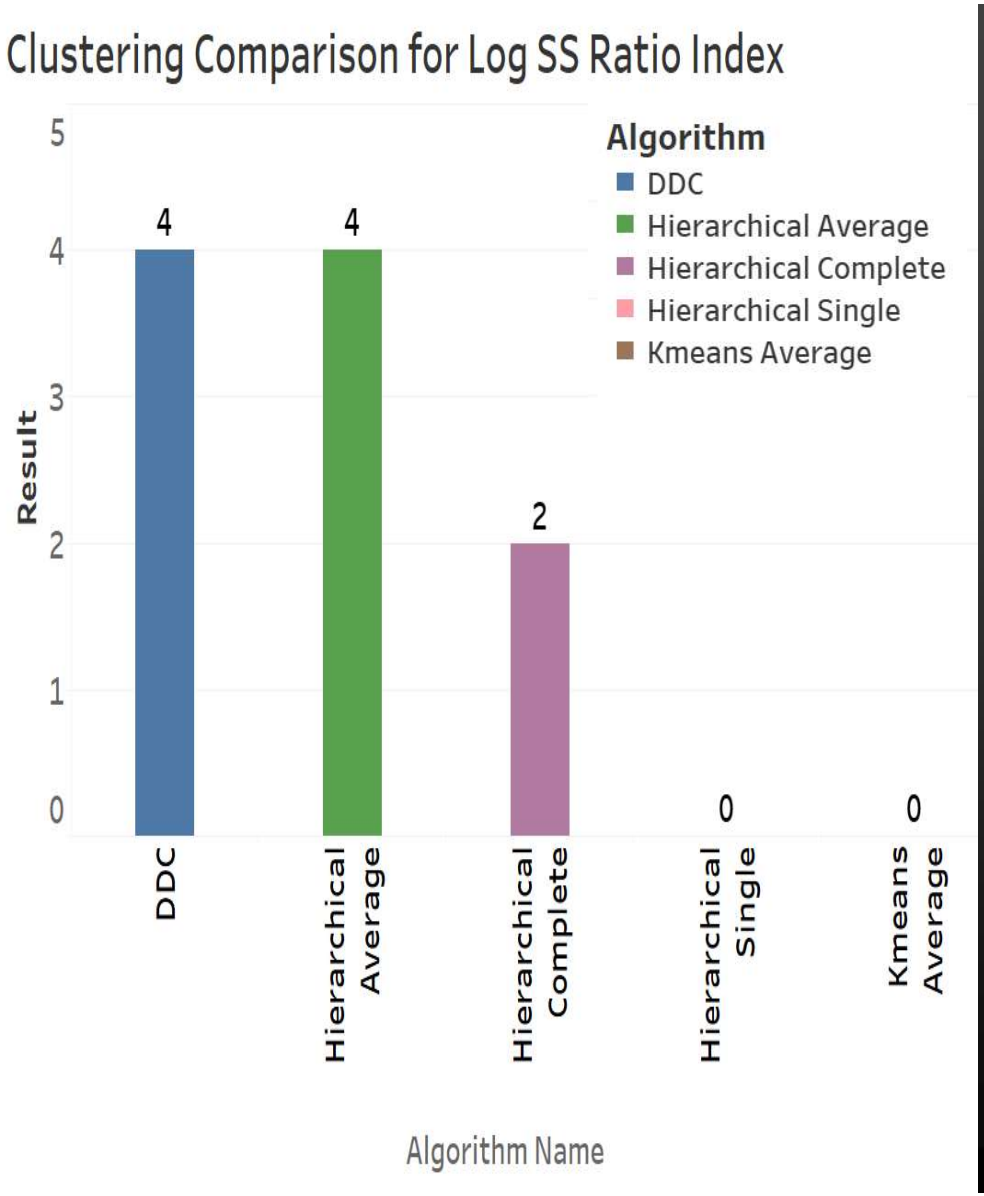
Experimental Result

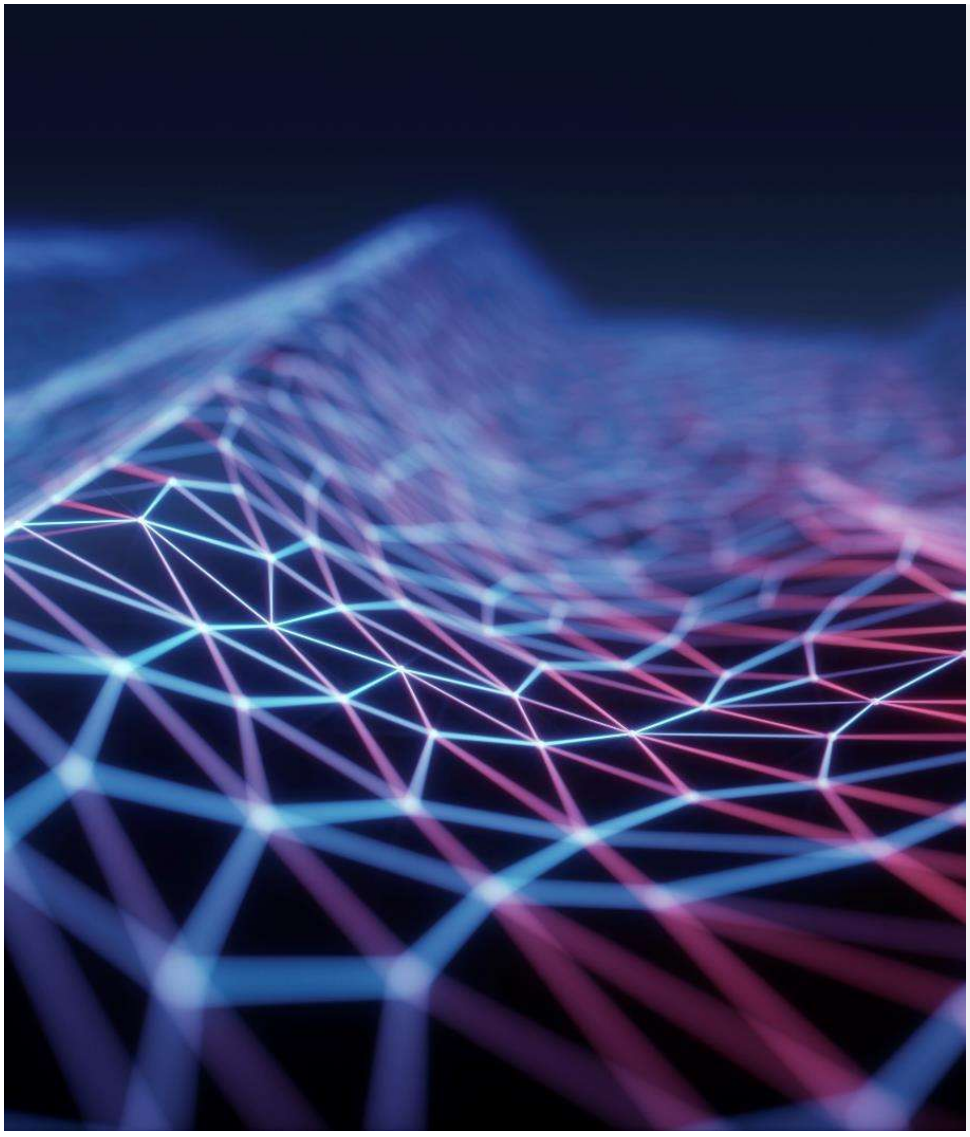


Kmeans Averaging



Clustering Algorithm Comparison





PBM Index

- This index is the square of the largest distance between two barycenters multiplied by the ratio of the sum of the distances of all the points to the barycenter G of the entire data set and the sum of the distances of the points of each cluster to their barycenter divided by the number of clusters.
- The bigger the value, better is the cluster validation.

Mathematically speaking

PBM Index is,

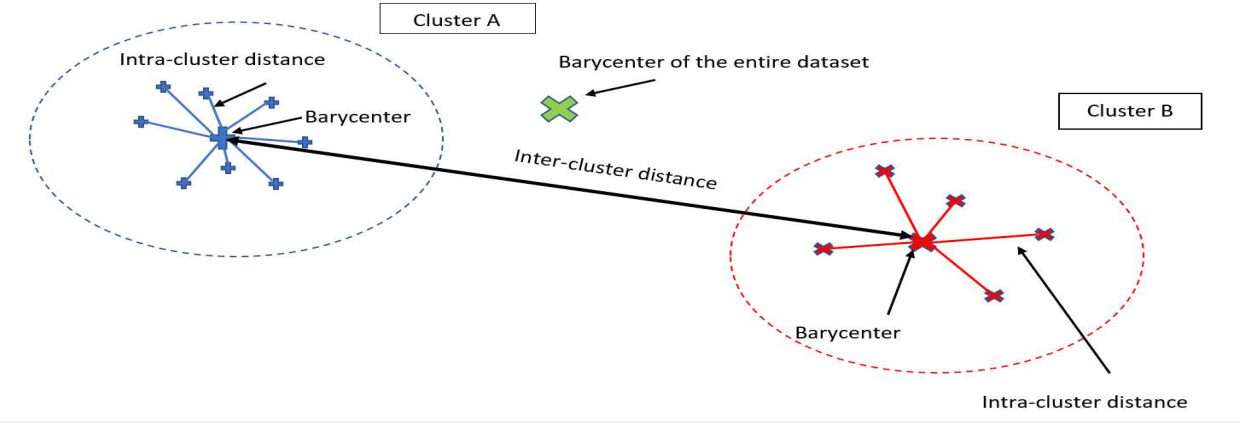
$$C = \frac{1}{K} \left(\frac{E_T}{E_W} X D_B\right)^2$$

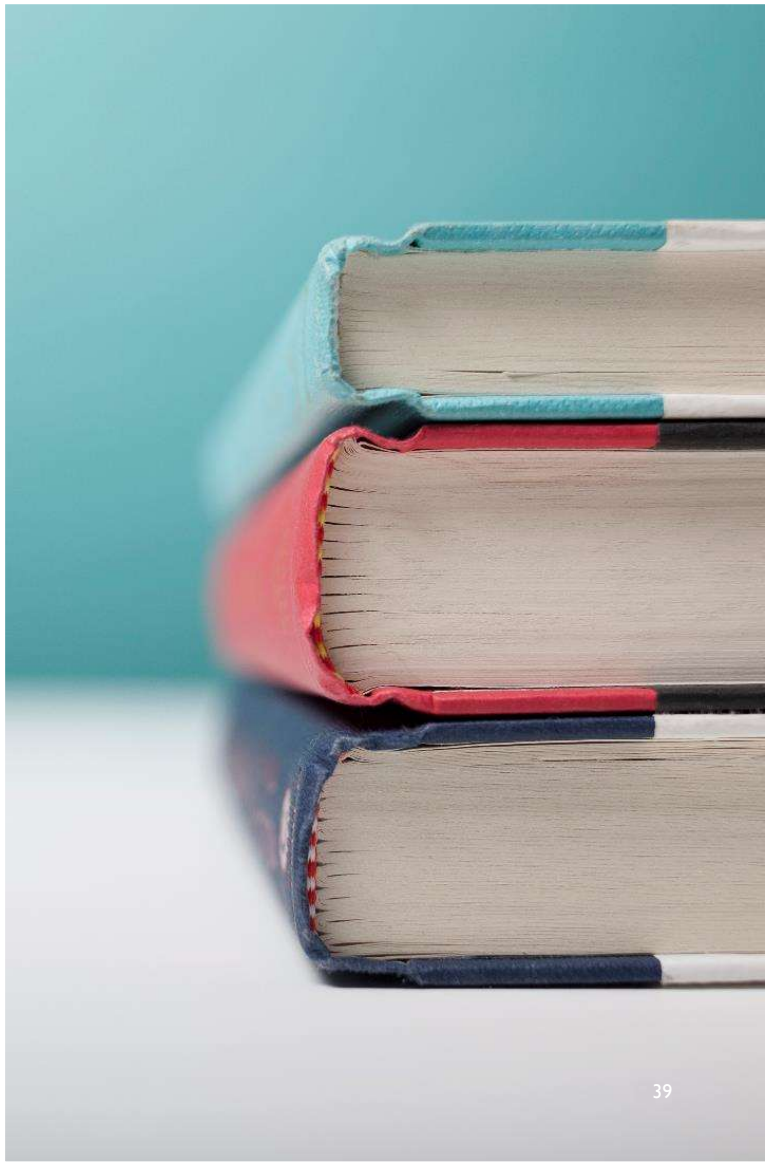
Where K is the number of clusters

E_T is the sum of the distances of all the points to the barycenter G of the entire data set.

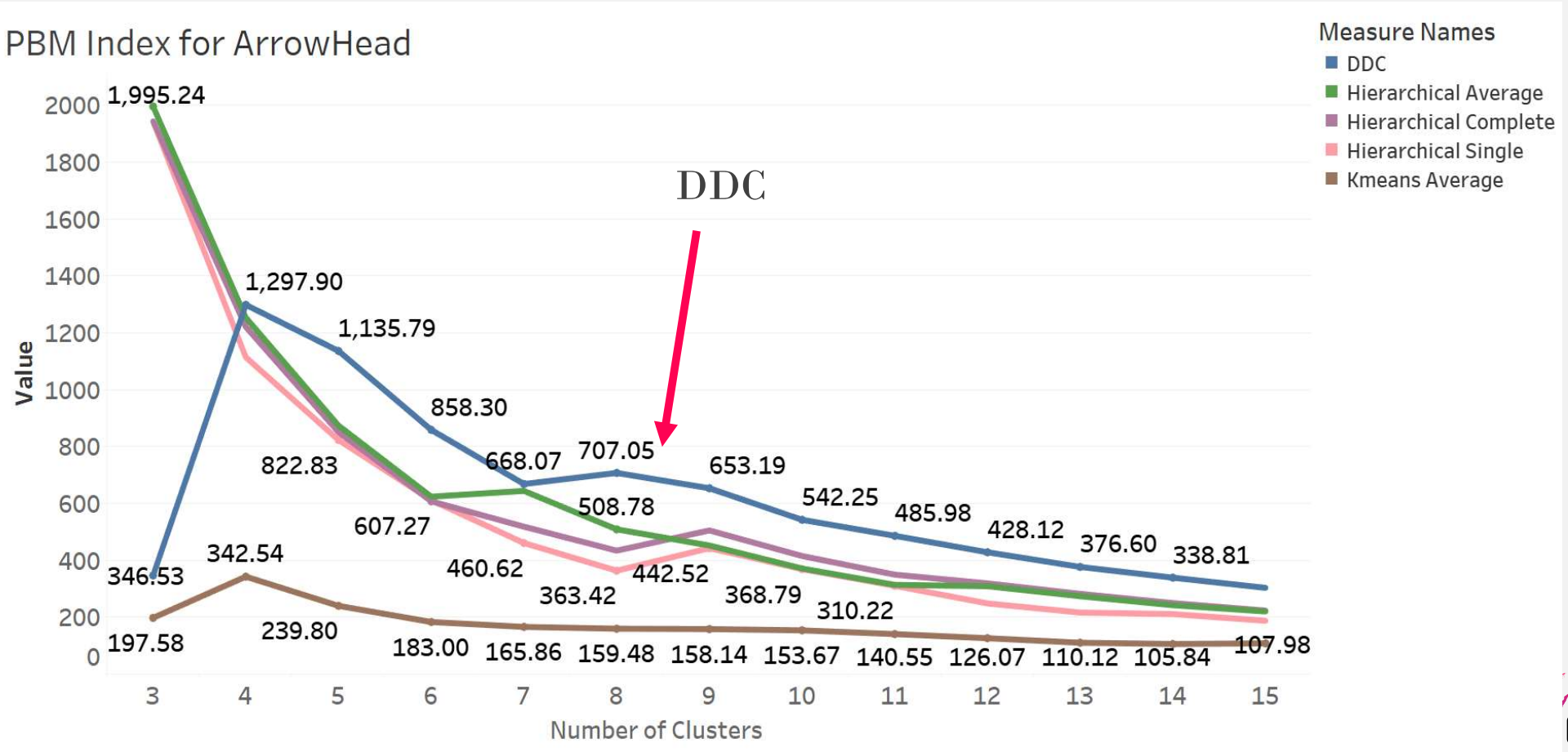
E_W is the sum of the distances of the points of each cluster to their barycenter.

D_B is the largest distance between two barycenters

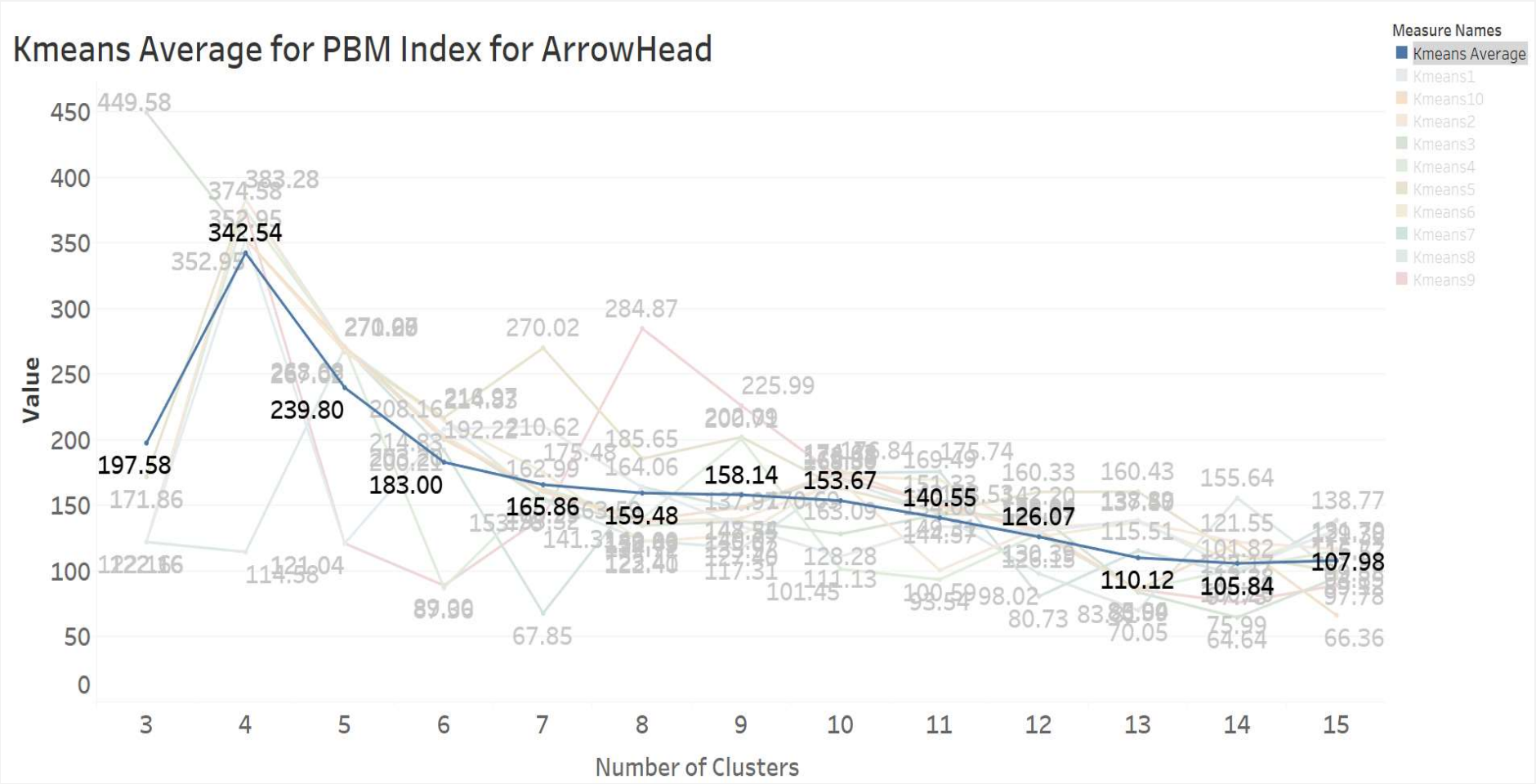




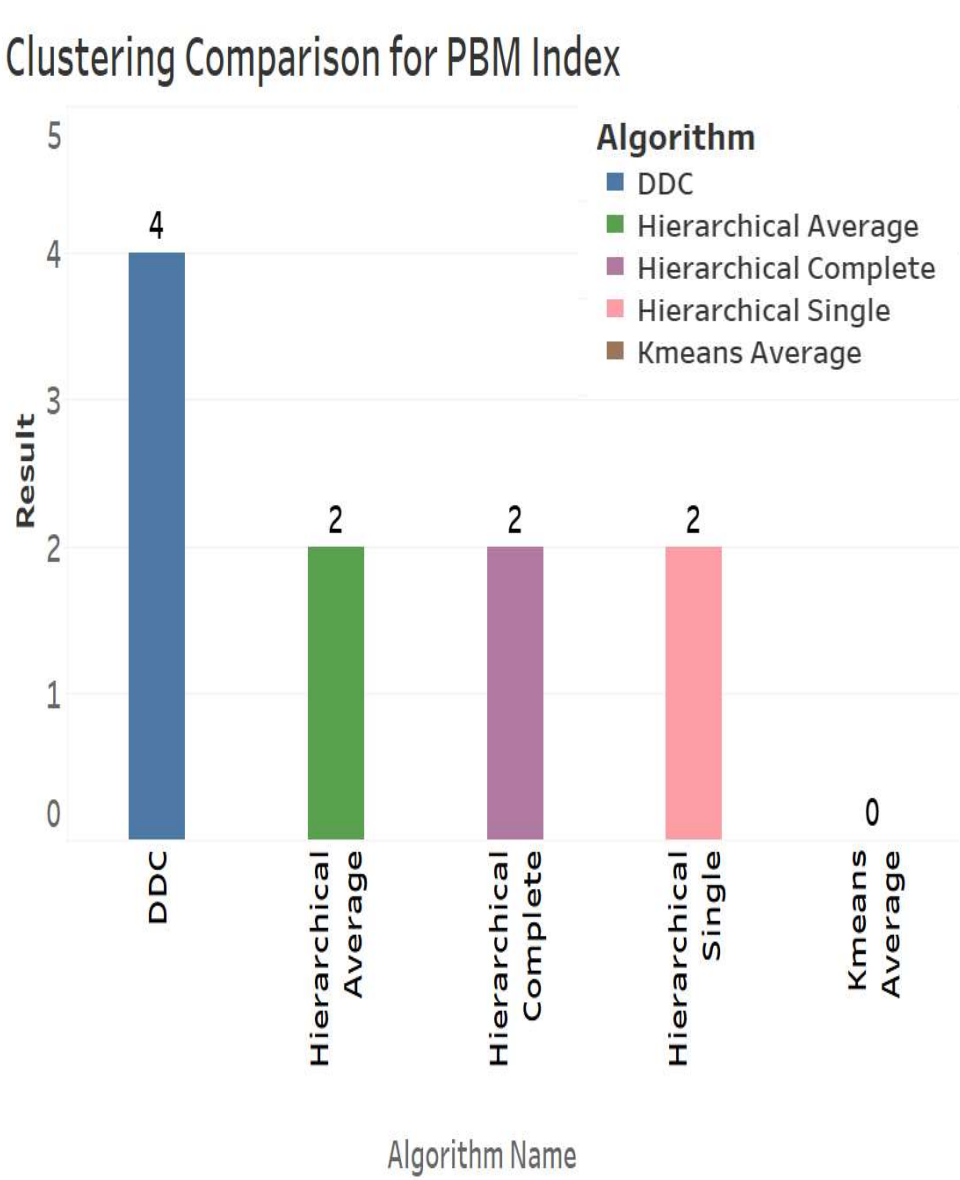
Experimental Result

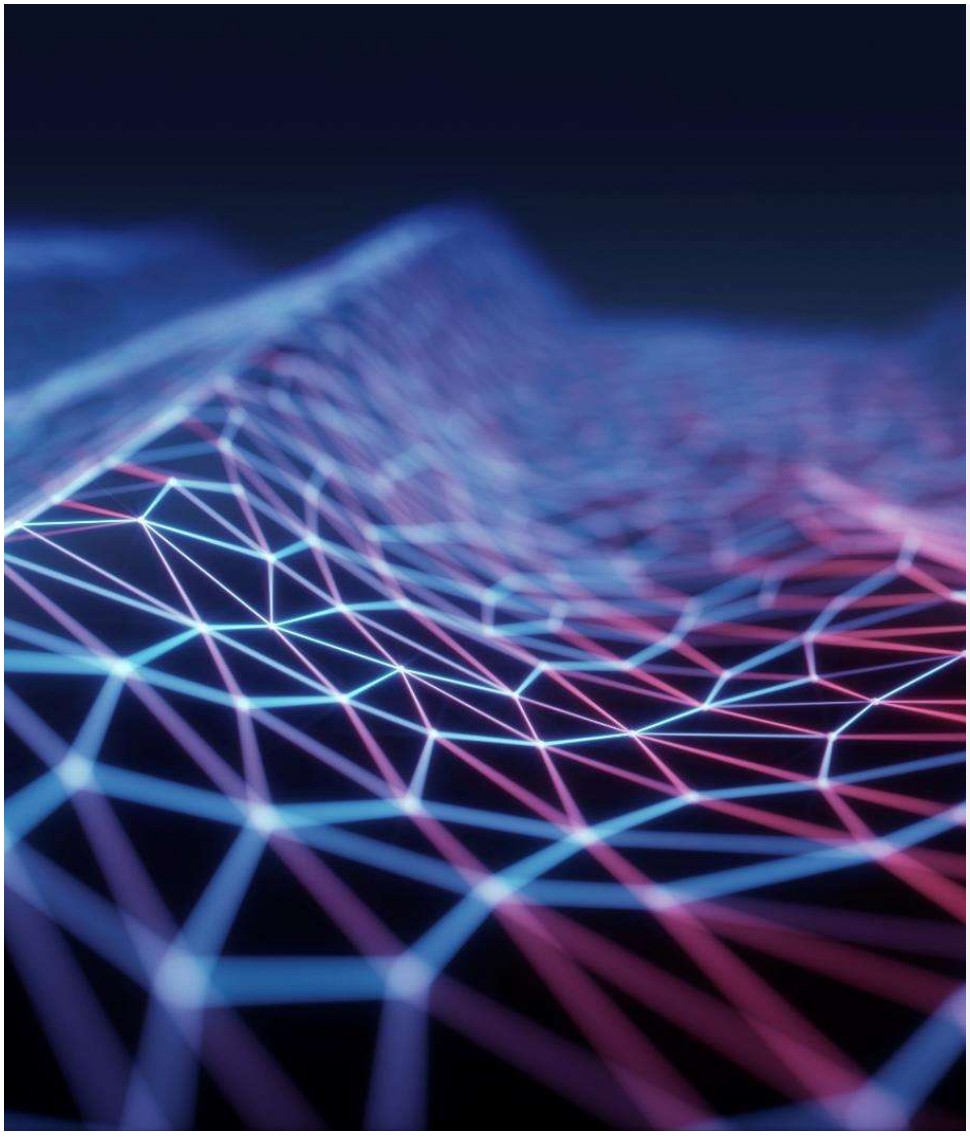


Kmeans Averaging



Clustering Algorithm Comparison





Ray-Turi Index

- the numerator is the mean of the squared distances of all the points with respect to the barycenter of the cluster they belong to.
- the denominator is the minimum of the squared distances $\Delta kk'$ between all the cluster barycenters.
- The lower the value, the better validation it is.

Mathematically speaking

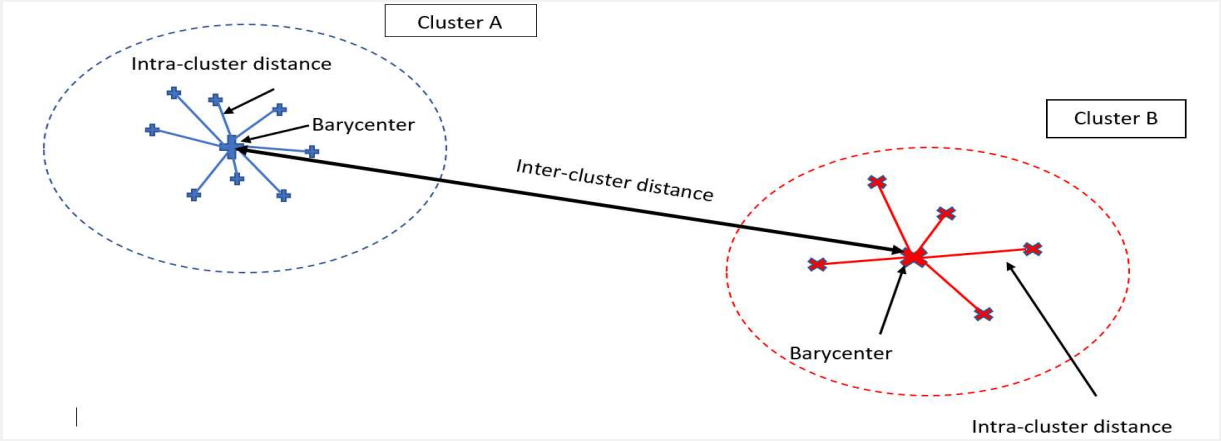
Ray Turi Index is,

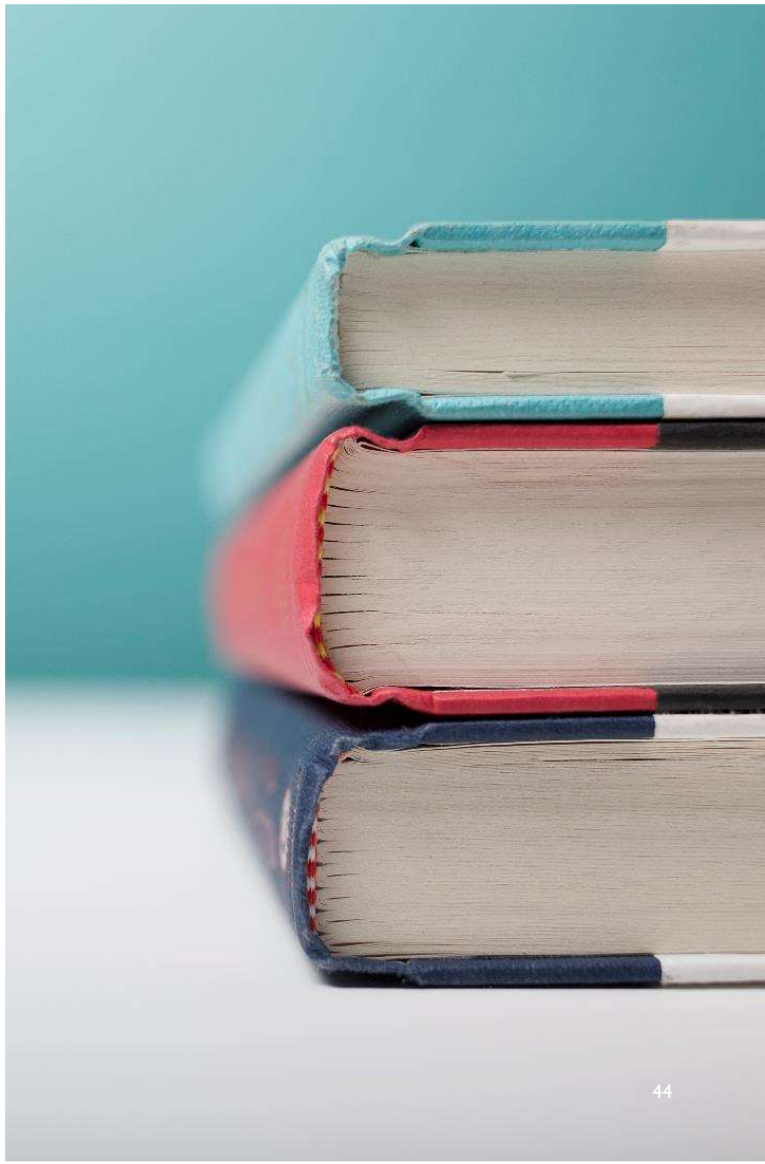
Mathematically speaking, the numerator is:

$$\frac{1}{N}\sum_{k=1}^{K} WGSS^{\{k\}} = \frac{1}{N} WGSS$$

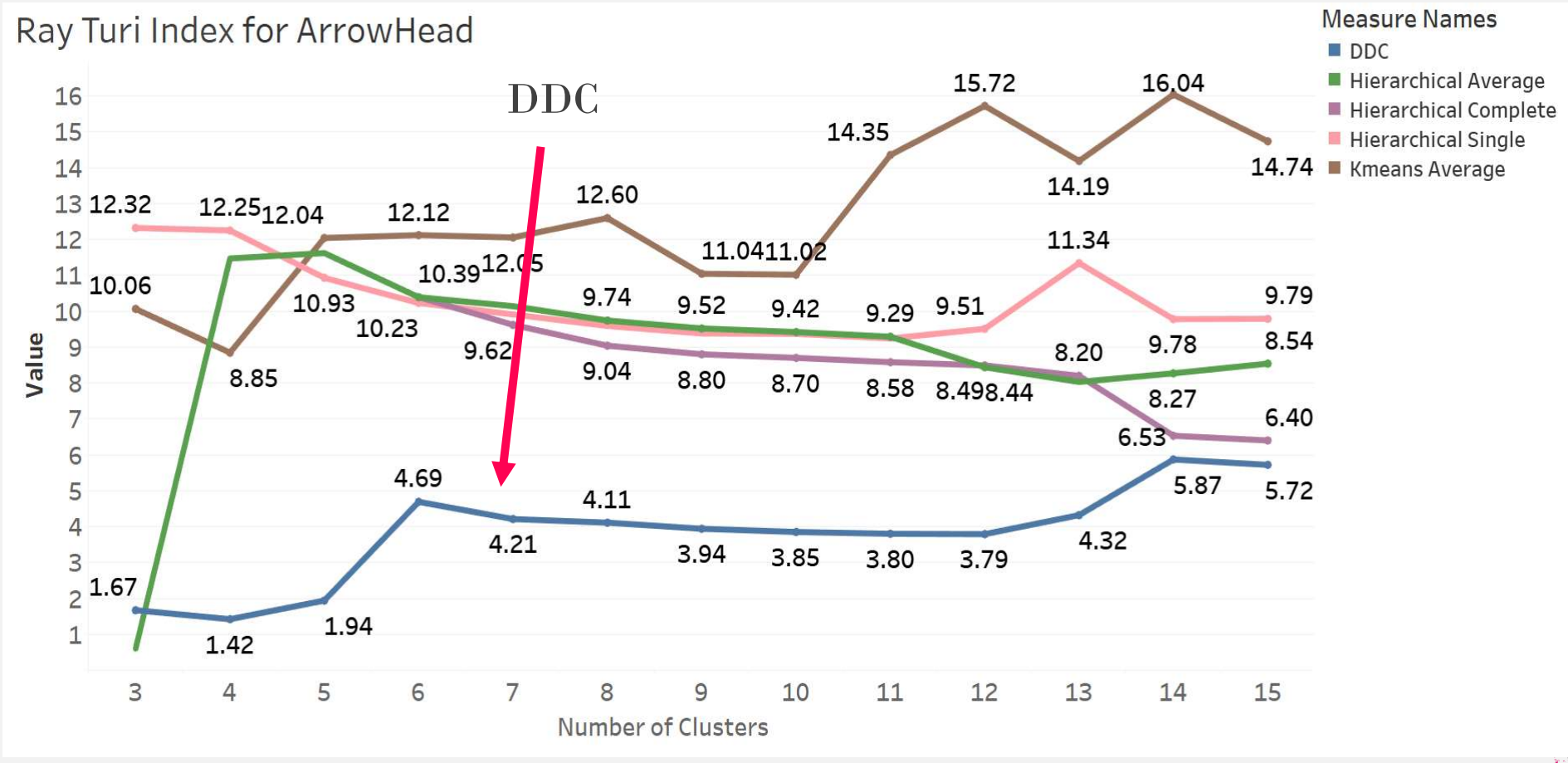
- And the denominator is, $\min_{k<k'} || G^{\{k\}} - G^{\{k'\}}||^2$

So, the Ray Turi Index can be written as $\frac{1}{N}\frac{WGSS}{\min_{k<k'} \Delta_{kk'}^2}$

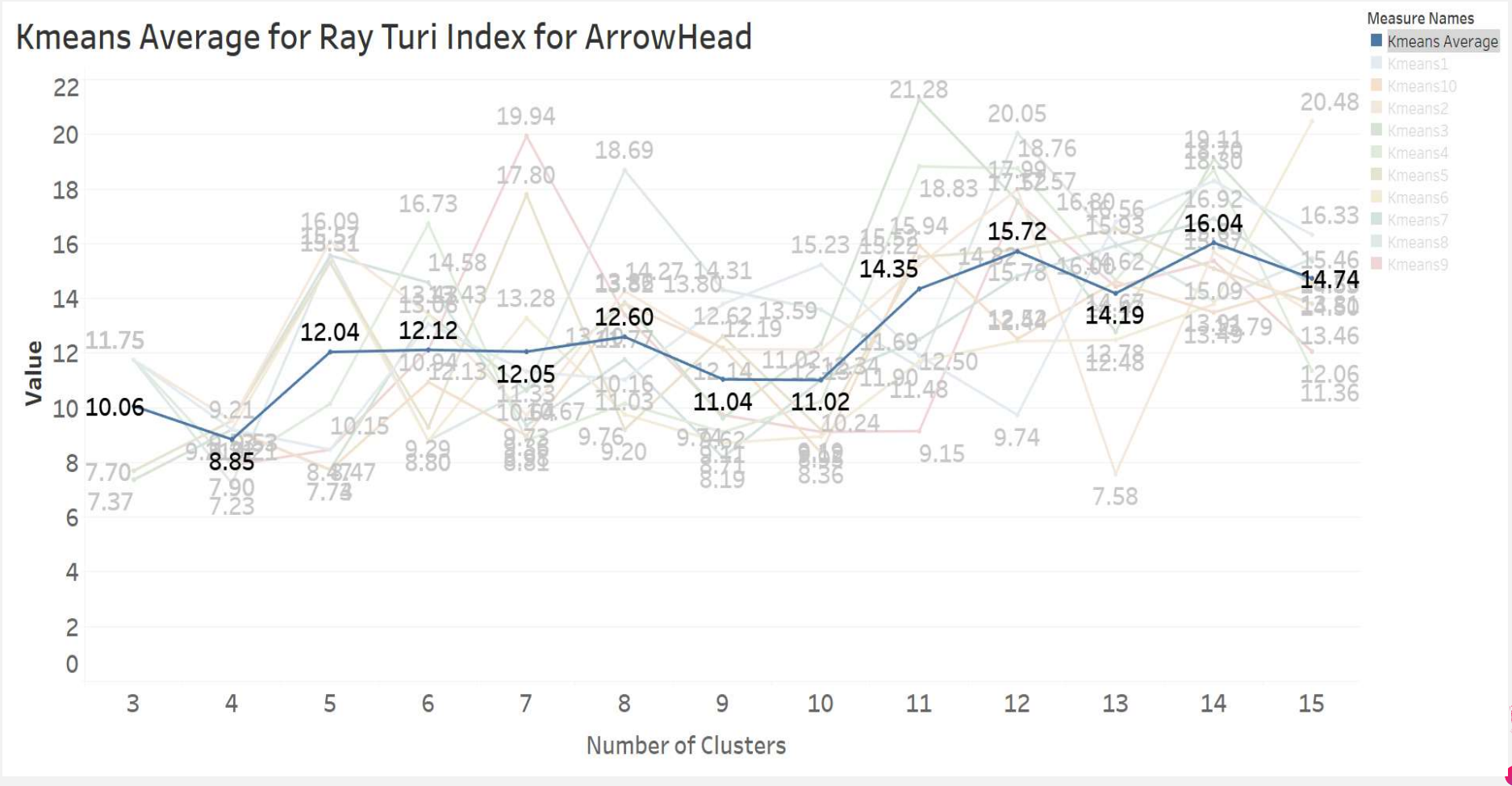




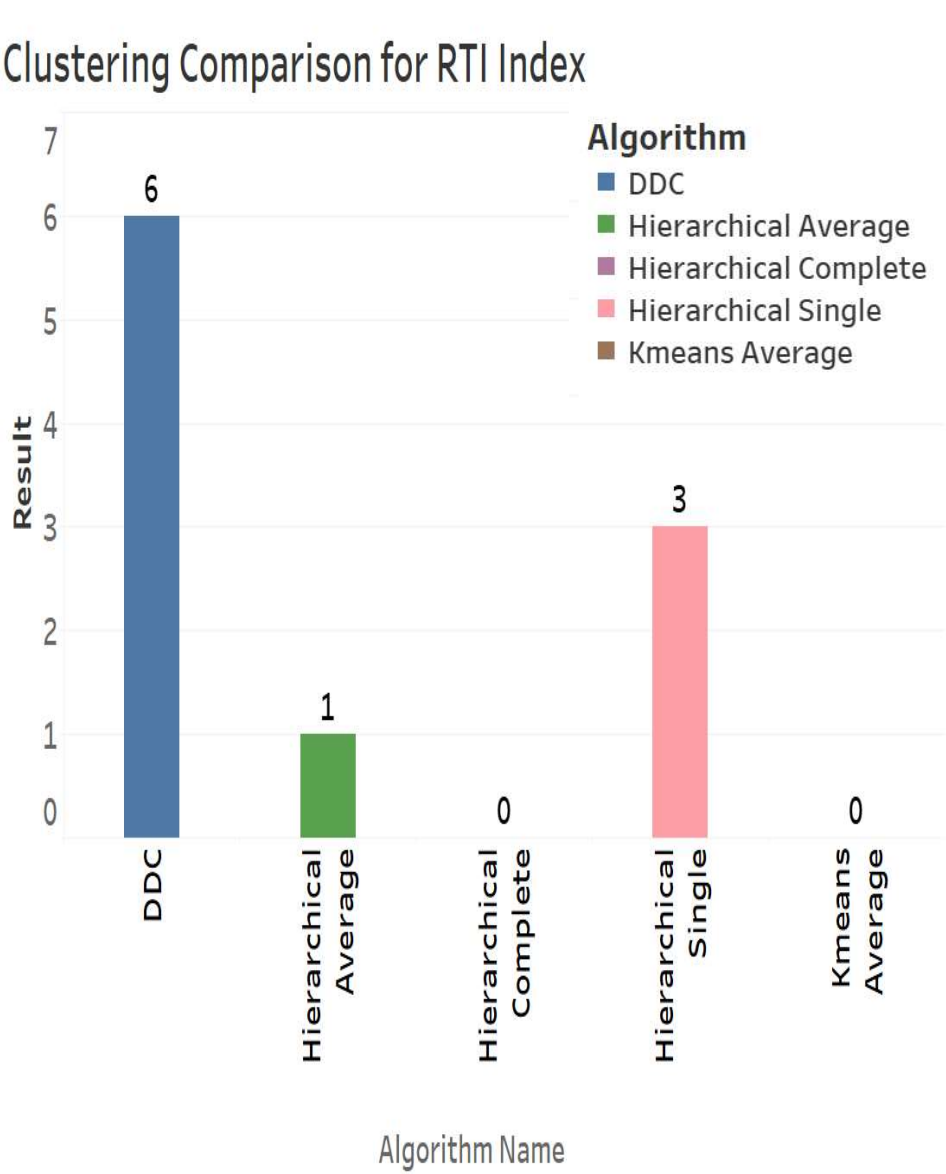
Experimental Result



Kmeans Averaging



Clustering Algorithm Comparison



Summary

Internal Indices	Intra-cluster distance	Inter-cluster distance	Ratio of inter and intra cluster distance	Lower Index value is better	Higher Index value is better	Lower value difference is better	Higher value difference is better	DDC	Hierarchical Single	Hierarchical Average	Hierarchical Complete	Kmeans Average
Ball Hall Index	X						X	X				
Banfeld-Raftery Index	X			X				X				
Calinski-Harabasz Index			X		X			X				
Davies-Bouldin Index			X	X								X
Log SS Ratio Index			X			X		X		X		
PBM Index	X				X			X				
Ray-Turi Index			X	X				X				

Conclusion

- ✓ Internal indices can be used for validating the goodness of clustering in time series data.
- ✓ DDC outperforms all other methods when we consider intra-cluster distances.
- ✓ When there is a ratio of BGSS and WGSS, all of them have almost equal performance in our datasets however, DDC has a slight edge over the others.
- ✓ On analysis, for Davies-Bouldin Index K-means random outperforms DDC because it has many single element clusters and due to which the mean distance increases hence the increase in the Index value which denotes a bad performance by DDC.
- ✓ A lower value of k for DDC clustering for Davies- Bouldin Index can augment its performance.

Future Work

- Experimentation on the more datasets.
- Document result on other internal indices like Dunn Index, Silhouette Index, and Xie Beni Index.
- Paper submission at ACM conference https://fods.acm.org/

Thank You