



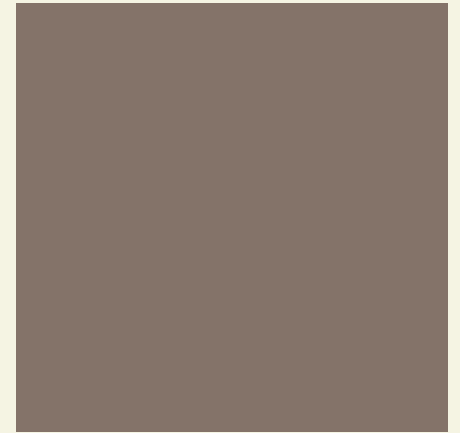
# ICCPP-STATISTICS

- Cluster Analysis

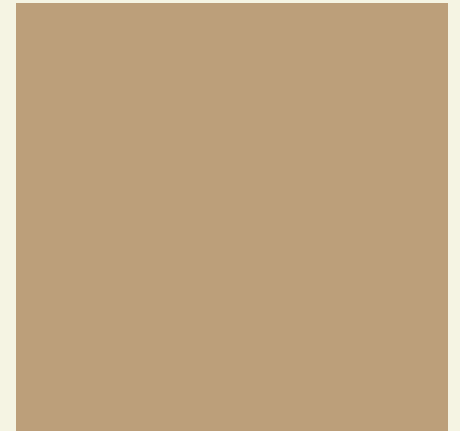
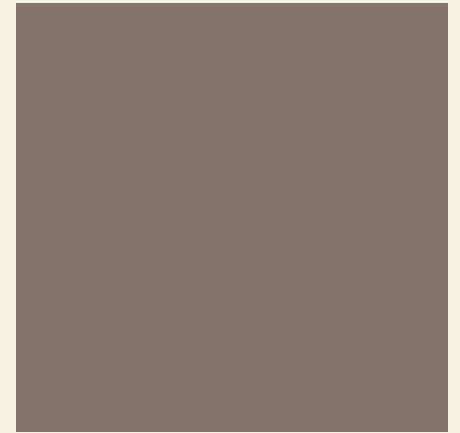
## Vishal Lohchab

*Scientific Assistant of  
Prof. Dr. Hans-Werner Gessmann  
Director ICCPP International*



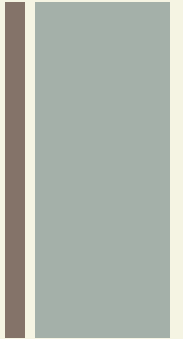


Joseph Zubin  
(1900-1990)  
Cluster Analysis



Robert Tryon  
(1901-1967)  
Cluster Analysis

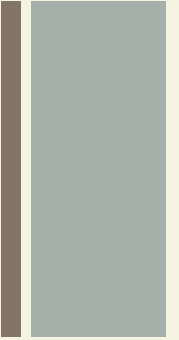
# + Definition



- Cluster analysis is a statistical classification technique in which a set of objects or points with similar characteristics are grouped together in clusters.
- The aim of cluster analysis is to organize observed data into meaningful structures in order to gain further insight from them.

# + Use of Cluster Analysis

- It is used to classify different objects into groups in such a way that the similarity between two objects is maximal if they belong to the same group and minimal if they do not belong to the same group.



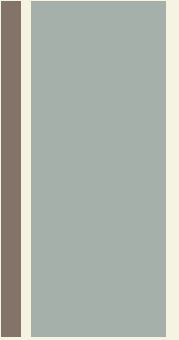
# + Use of Cluster Analysis

- Unlike many other statistical methods, cluster analysis is typically used when there is no assumption made about the likely relationships within the data.
- It provides information about where associations and patterns in data exist, but not what those might be or what they mean.

+

# Clusters can be based on factors like

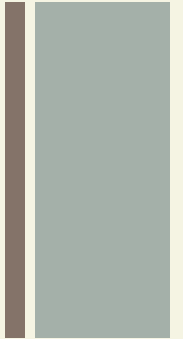
- Distance-based Clustering: Items are sorted based on their proximity (or distance).
- For example, cancer cases might be clustered together if they are in the same geographic location.



+

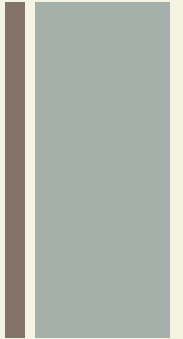
Clusters can be based on factors like

- Conceptual Clustering: Items are grouped by factors that items have in common.
- For example, cancer clusters could be grouped by “people who work in manufacturing.”





# + K Means Clustering



- Clustering is just a way to group a set of data into smaller sets.

- The two ways you could group a set of data:

Quantitatively (using numbers)

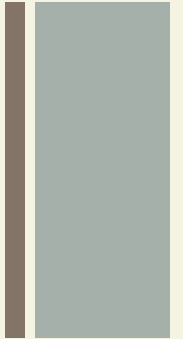
Qualitatively (using categories).

# + K-Means Clustering

- K-Means clustering is one of the simplest unsupervised learning algorithms that solves clustering problems using a quantitative method:

You pre-define a number of clusters and employ an algorithm name “simple” to sort your data.

# + K-Means Clustering



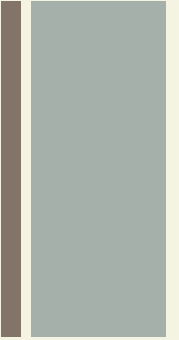
You have to use software for K-means clustering.  
Some programs that can perform clustering are:

- SPSS
- r
- MATLAB

+

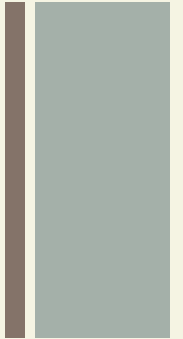
# General steps behind the K-means clustering algorithm

- Decide how many clusters ( $k$ ).
- Place  $k$  central points in different locations (usually far apart from each other).
- Take each data point and place it close to the appropriate central point. Repeat until all data points have been assigned.



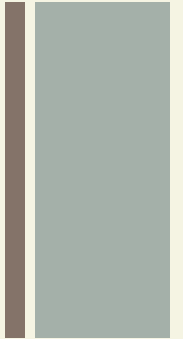
+

# General steps behind the K-means clustering algorithm



- Re-calculate  $k$  new central points as barycenters.
- Repeat the assigning of data points, this time to the new central point (the barycenter).
- Repeat 4 and 5 until the central points (barycenters) do not move any more.

# + K-Means Clustering



- K-Means clustering is to categorize  $n$  objects into  $k$  ( $k > 1$ ) pre-defined groups.
- The goal is to minimize the distance from each data point to the cluster.

# + K-Means Clustering

- In other words, to find:

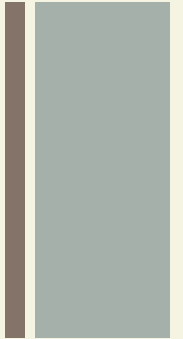
$$\arg \min_{\mathbf{S}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Where:

- $\mathbf{x}$  is a data point
- $k$  is the number of clusters
- $\boldsymbol{\mu}_i$  is the mean of the points in  $S_i$ .



# Cluster Analysis vs Discriminant Analysis

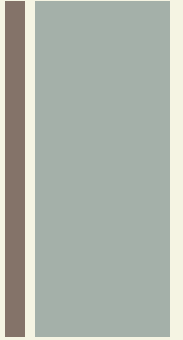


- Cluster analysis is very similar to discriminant analysis. Both methods involves separation into groups.
- However, cluster analysis is a way to identify the groups, while discriminant analysis requires you to know the groups before you begin analysis.





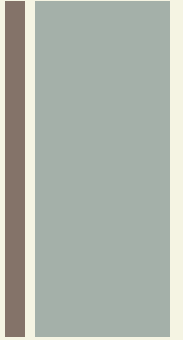
# Cluster Analysis vs Discriminant Analysis



- For example, let's say you had a group of psychiatric patients with abnormal behaviors.
- Cluster analysis could help you find distinct groups, like patients with a history of abuse, those with PTSD, or those experiencing hallucinations.



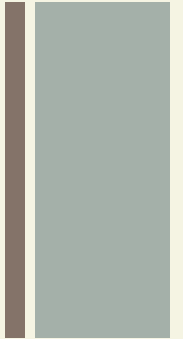
# Cluster Analysis vs Discriminant Analysis



- If you were to run discriminant analysis on the same group of people, you must know the patients diagnoses before you start placing them into groups.



# References



[https://en.wikipedia.org/wiki/Joseph\\_Zubin](https://en.wikipedia.org/wiki/Joseph_Zubin), date 27.11.21, 14:30 h MET

[https://en.wikipedia.org/wiki/Robert\\_Tryon](https://en.wikipedia.org/wiki/Robert_Tryon), date 27.11.21, 15:00 h MET

<https://www.techopedia.com/definition/30391/cluster-analysis>, date 27.11.21, 17:00 h MET

<https://www.qualtrics.com/au/experience-management/research/cluster-analysis/>, date 27.11.21, 15:30 h MET

Stephanie Glen. "Clustering and K Means: Definition & Cluster Analysis in Excel" From StatisticsHowTo.com: Elementary Statistics for the rest of us! <https://www.statisticshowto.com/clustering/>