



Cluster Analysis and Its Applications

Part 1

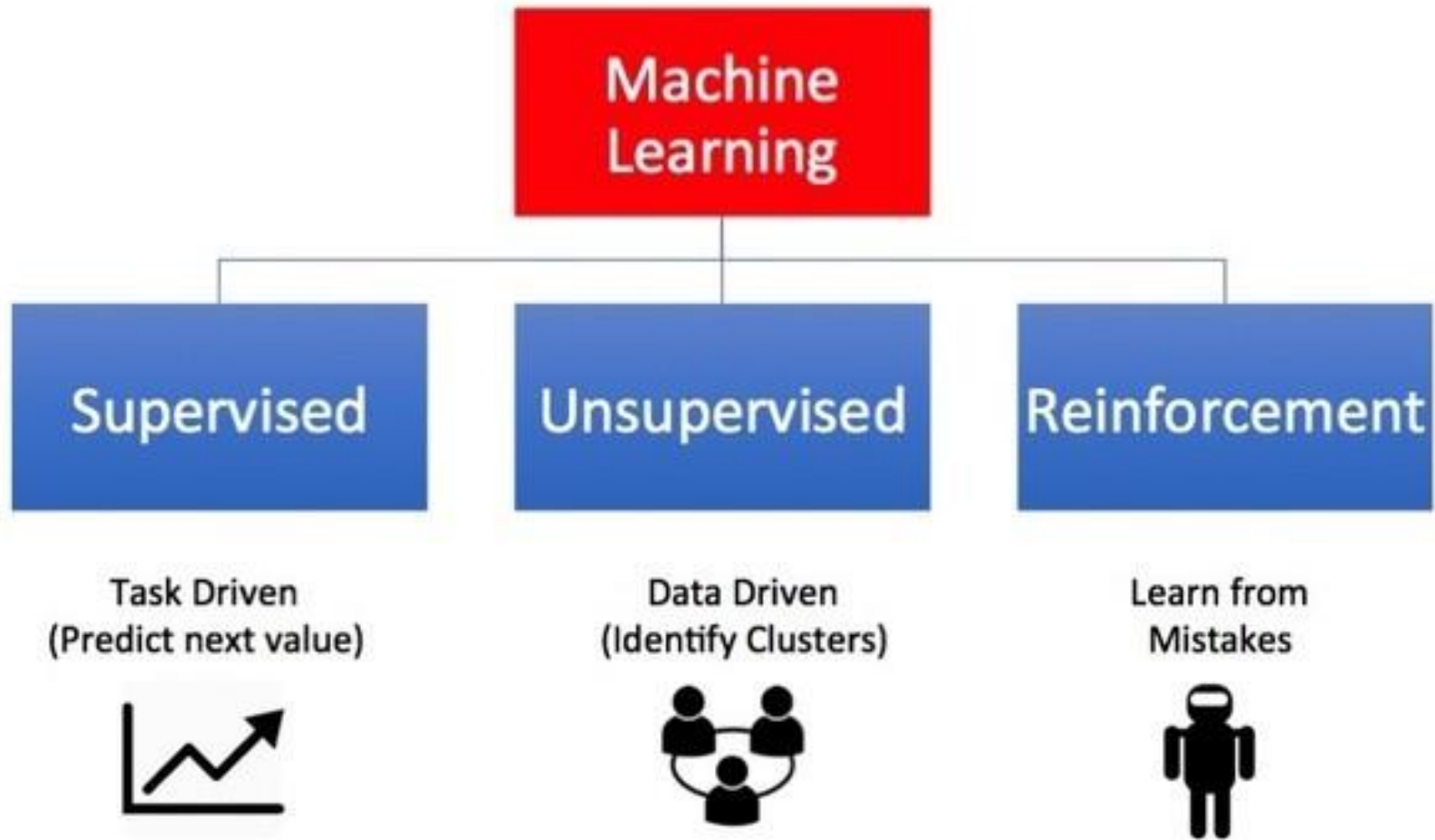
Albert C. Yang, M.D., Ph.D.

Institutes of Brain Science/Digital Medicine Center
National Yang-Ming University

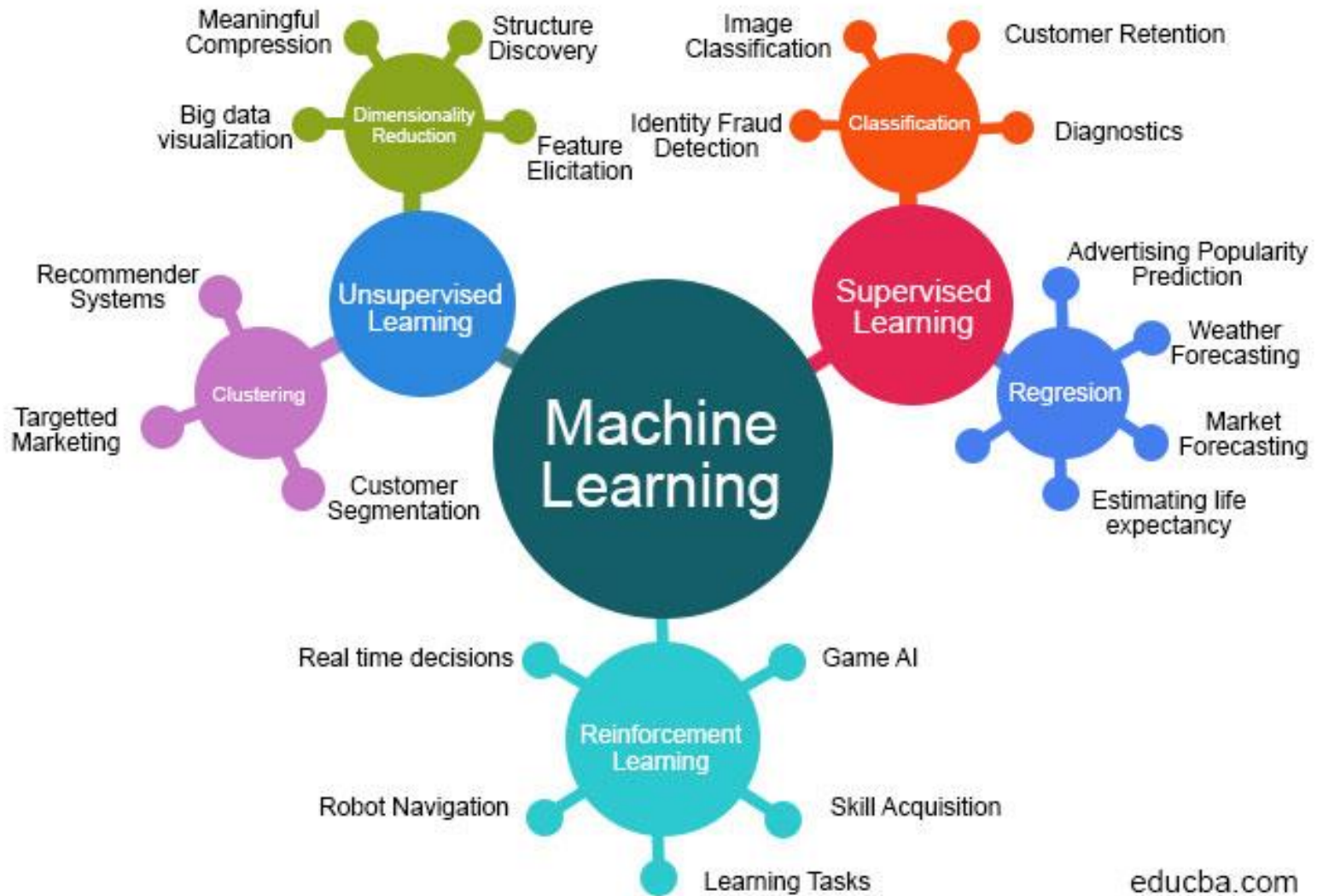
May 7, 2020

accyang@gmail.com

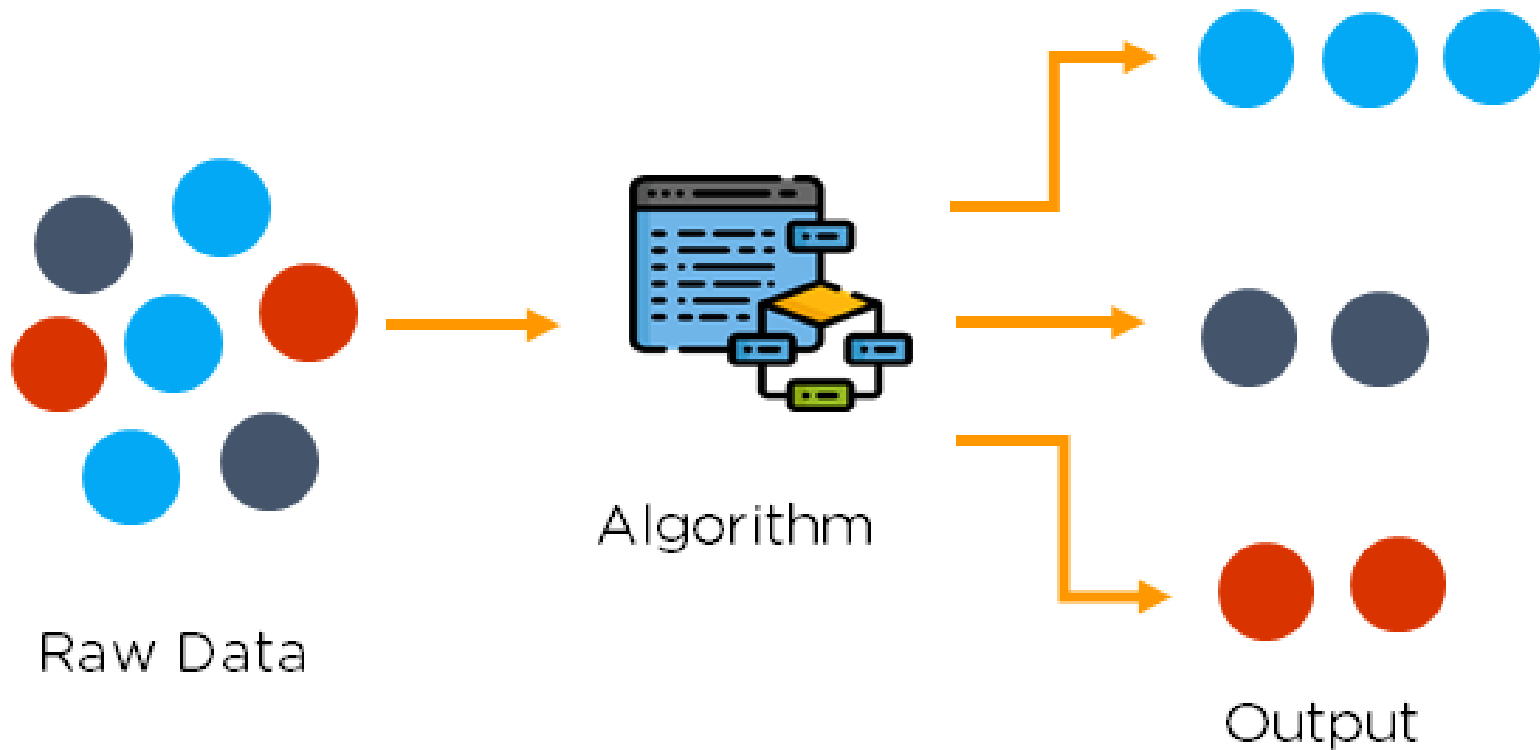
Types of Machine Learning



Machine Learning Algorithms



Cluster Analysis



<https://www.quora.com/What-is-the-difference-between-k-means-and-hierarchical-clustering>

Measuring Similarity

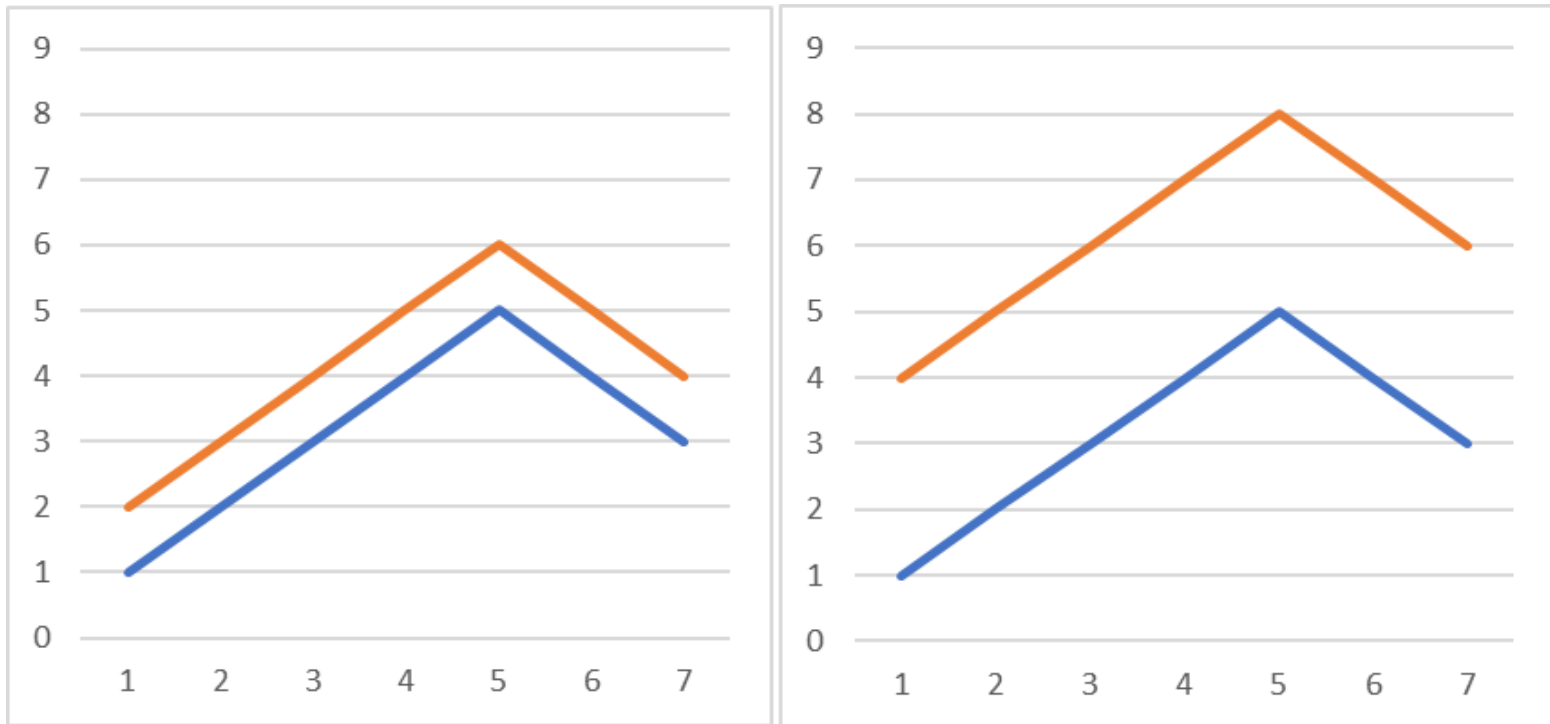
- A set of rules that serve as criteria for grouping or separating items
- Correlational measures
 - E.g. Pearson's correlation
- Distance Measures
 - Higher values indicate greater dissimilarity

Cluster vs. Factor Analysis

- Cluster Analysis
 - Based on distance/proximity measures

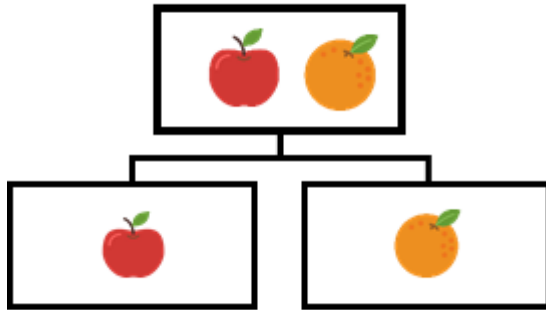
- Factor Analysis
 - Based on correlation (patterns)

Correlation vs. Distance

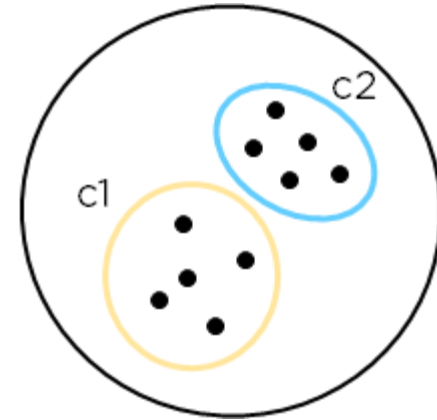


Both graph have the same correlation coefficient $r = 1$, but has different distance

Hierarchical vs K-Means Clustering

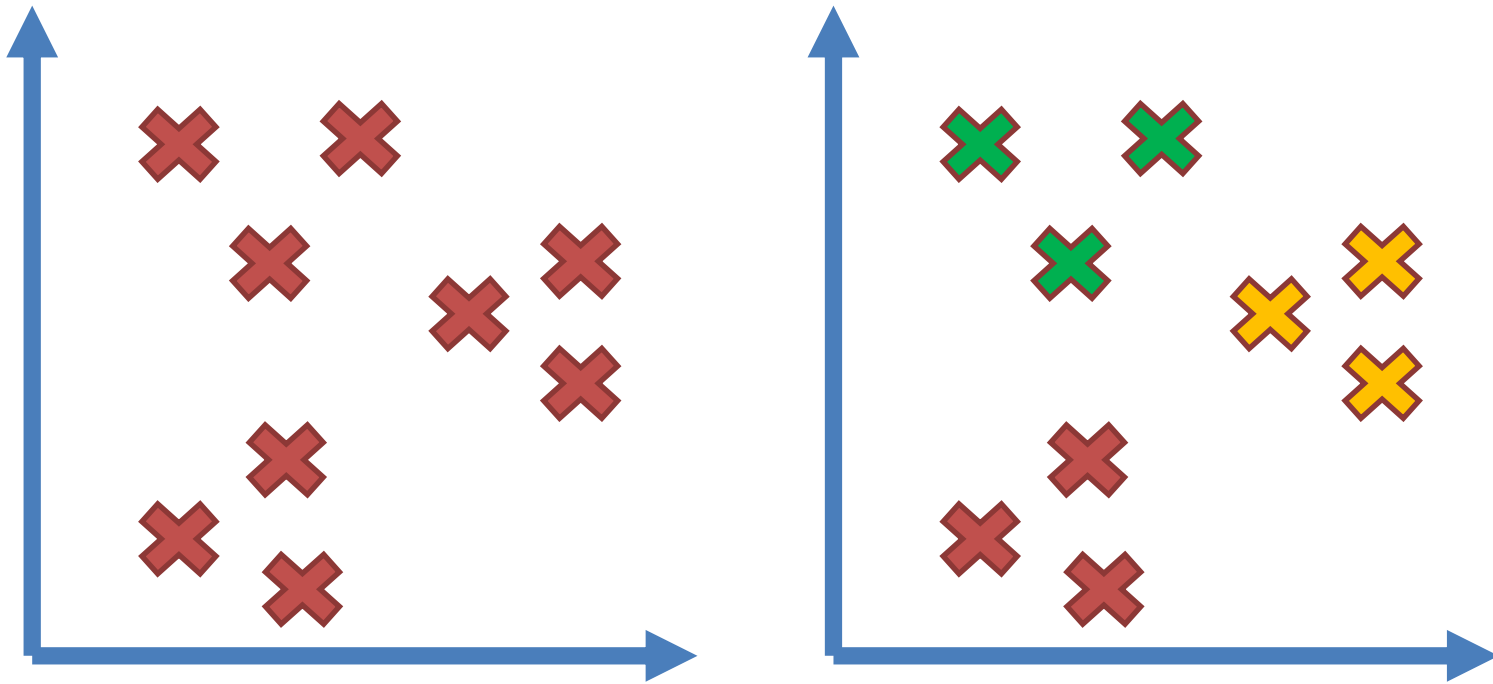


Hierarchical Clustering



K-Means Clustering

Cluster Analysis



Clustering

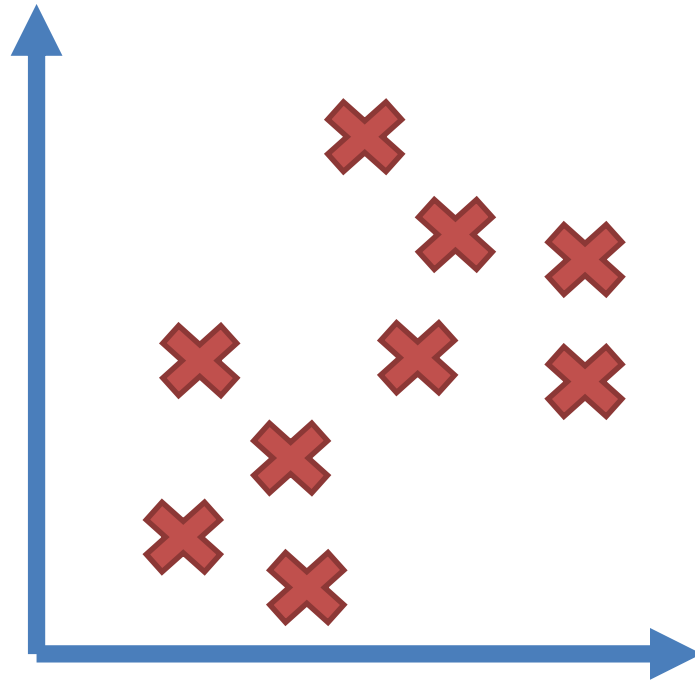


K-Means Cluster Analysis

- Step 1: Choose the number K of clusters
- Step 2: Select random K points as centroid
- Step 3: Assign each data point to the closest centroid
- Step 4: Compute and place the new centroid of each cluster
- Step 5: Reassign each data point to the new closet centroid.
 - If any reassignment occurred, go to step 4, otherwise the analysis is done.

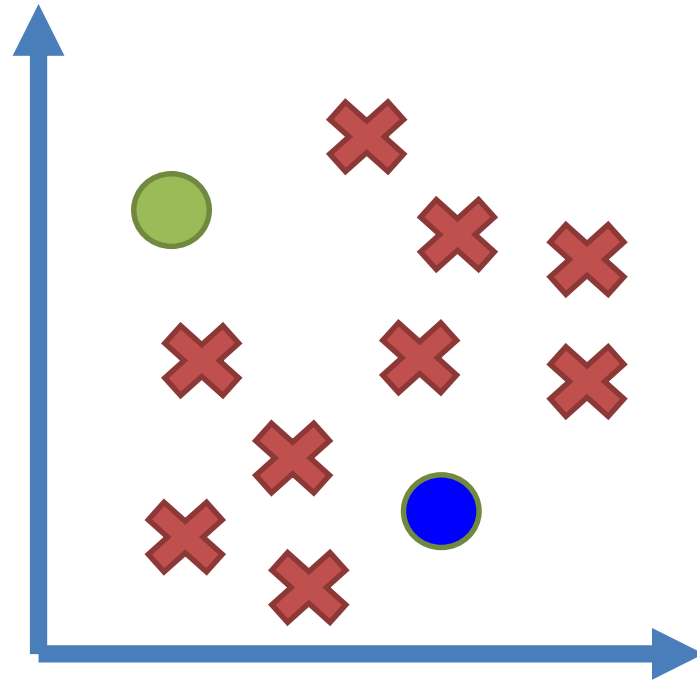
K-Means Cluster Analysis

- Step 1: Choose the number K of clusters ($K = 2$)



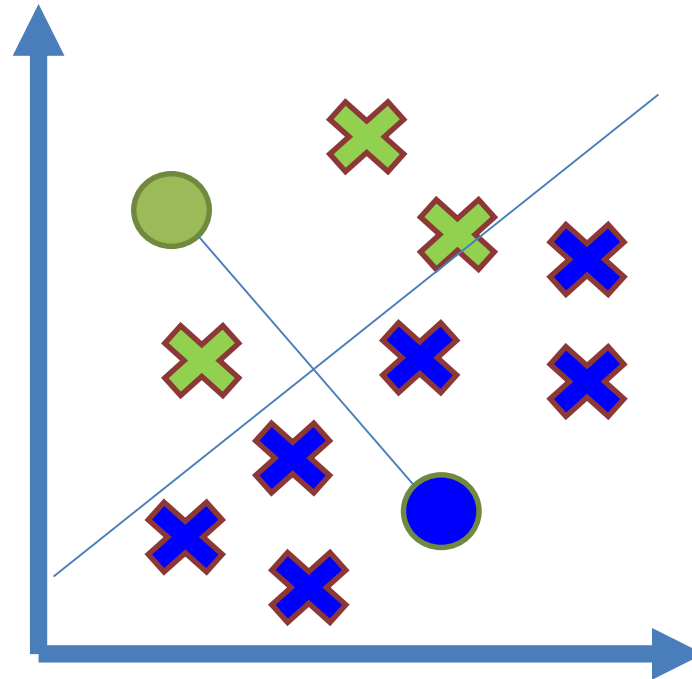
K-Means Cluster Analysis

- Step 2: Select random K points as centroid



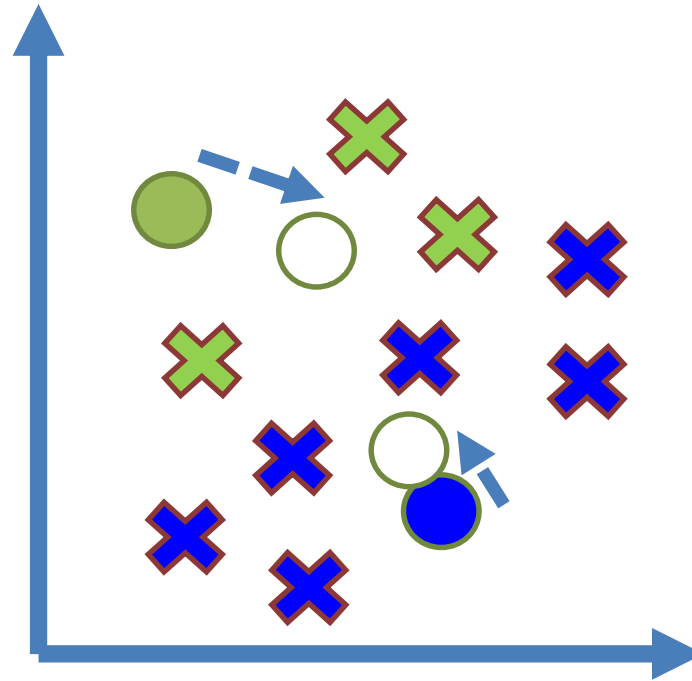
K-Means Cluster Analysis

- Step 3: Assign each data point to the closest centroid



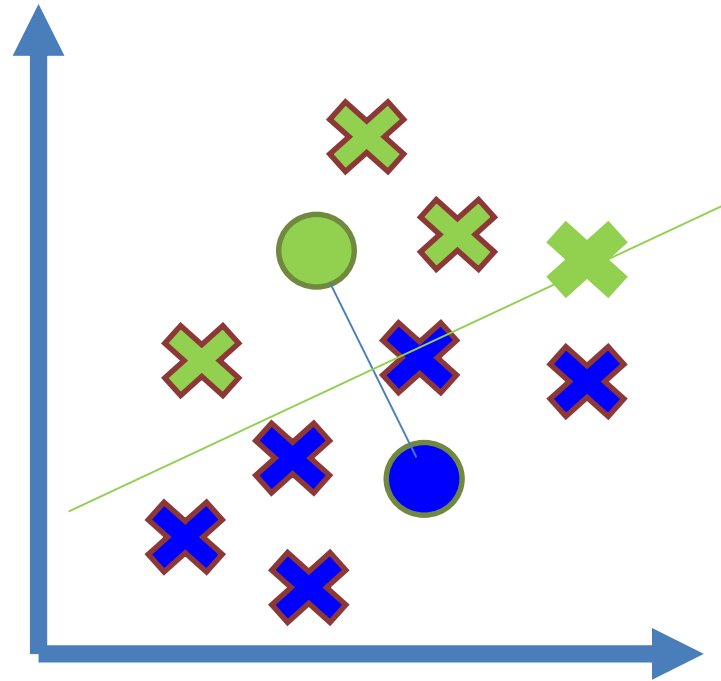
K-Means Cluster Analysis

- Step 4: Compute and place the new centroid of each cluster



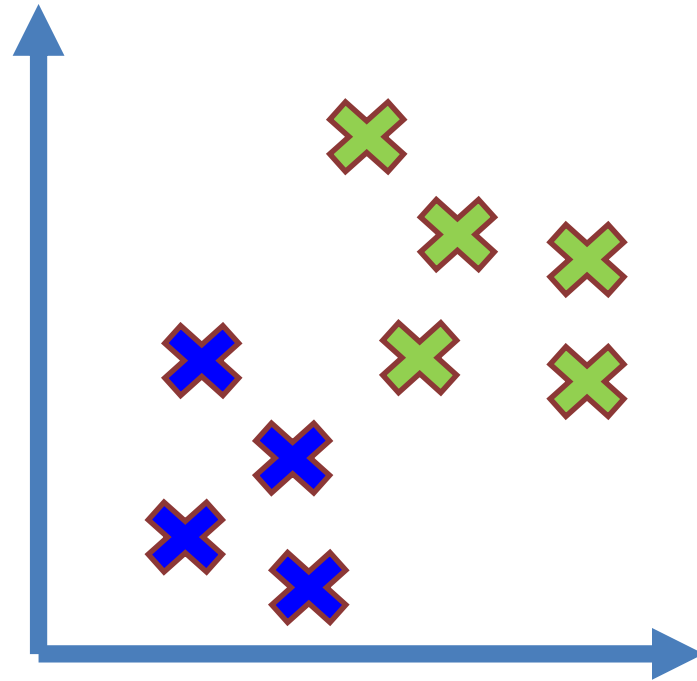
K-Means Cluster Analysis

- Step 5: Reassign each data point to the new closet centroid.

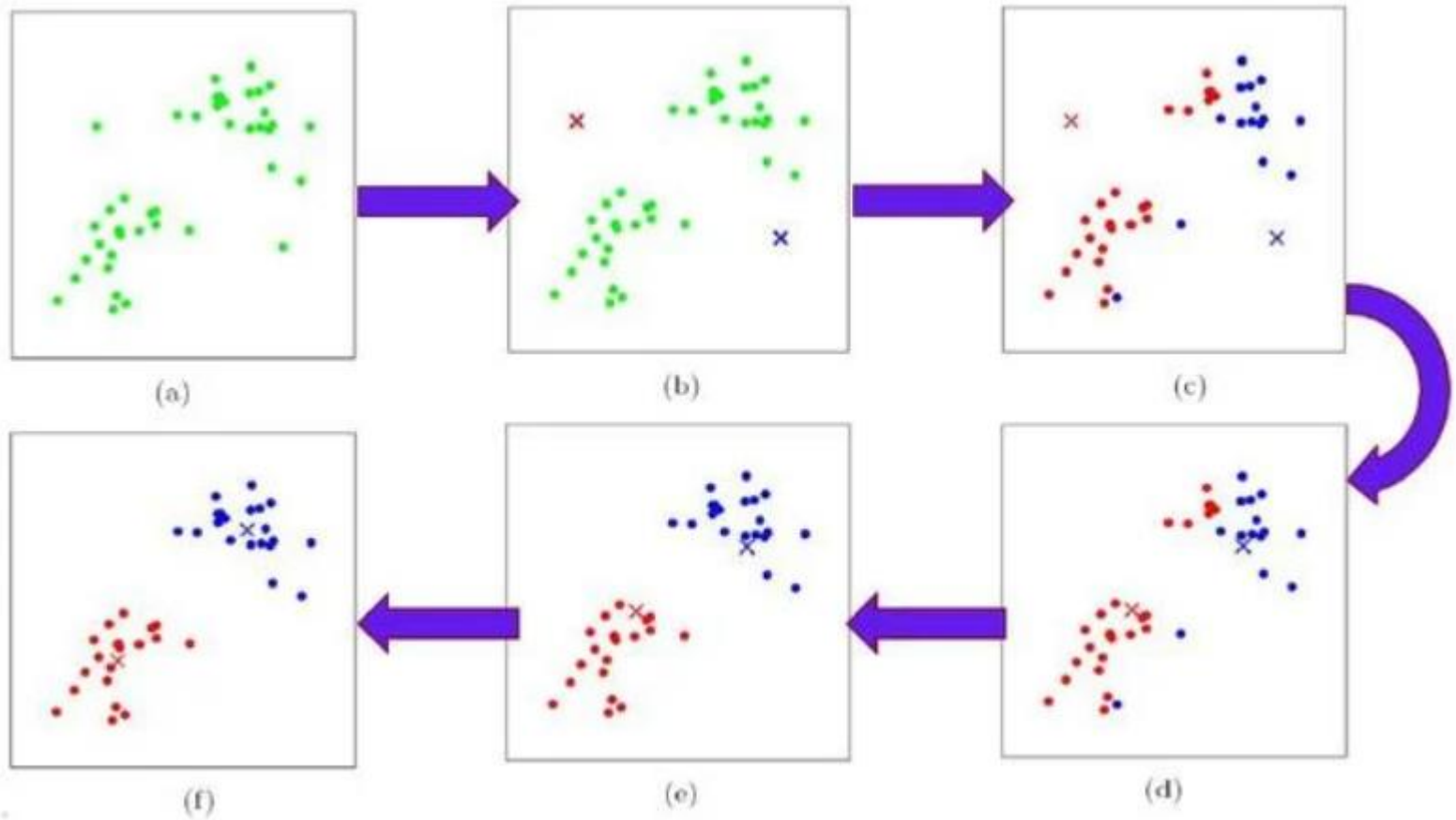


K-Means Cluster Analysis

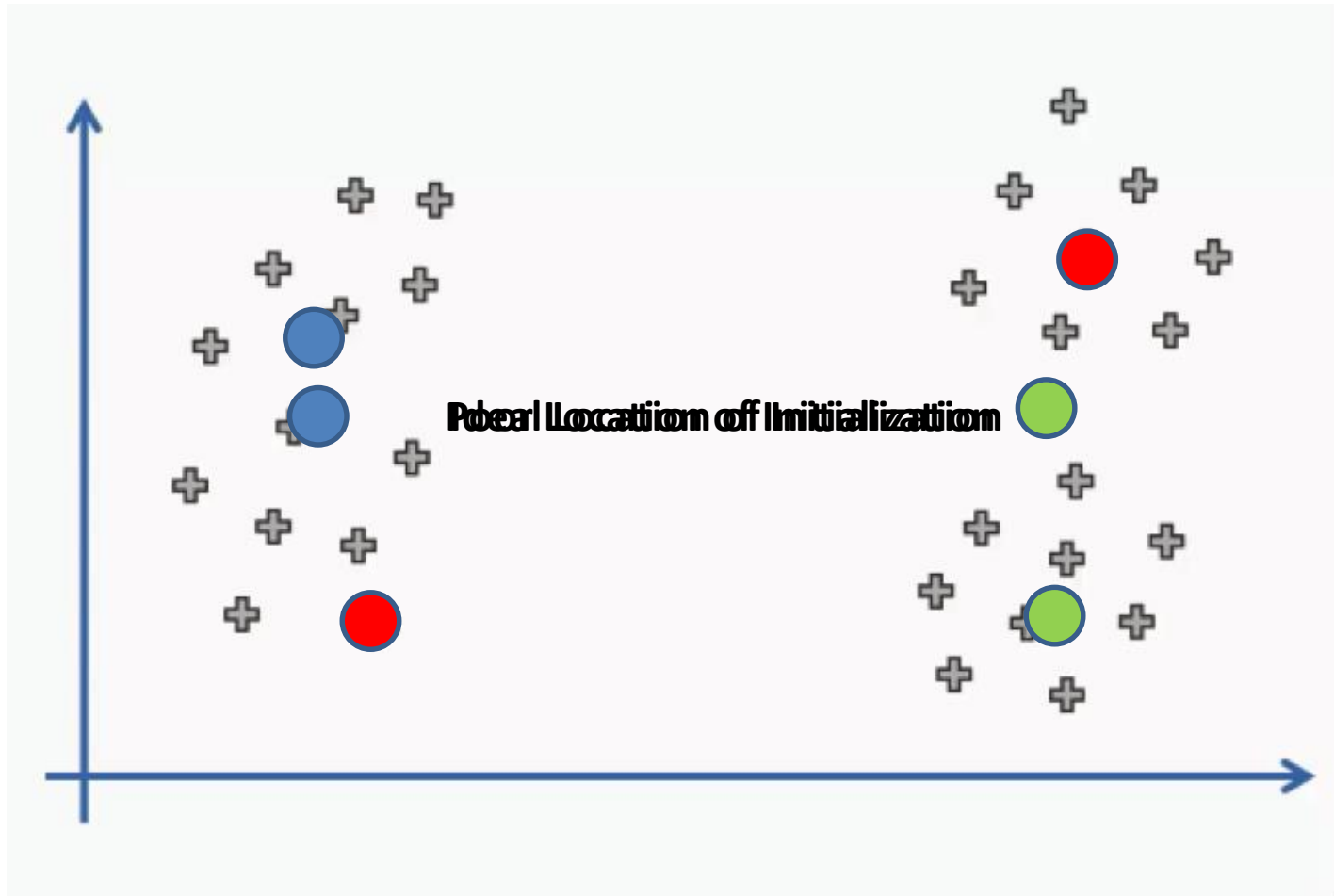
- Final Clusters



K-Means Cluster Analysis: Recap



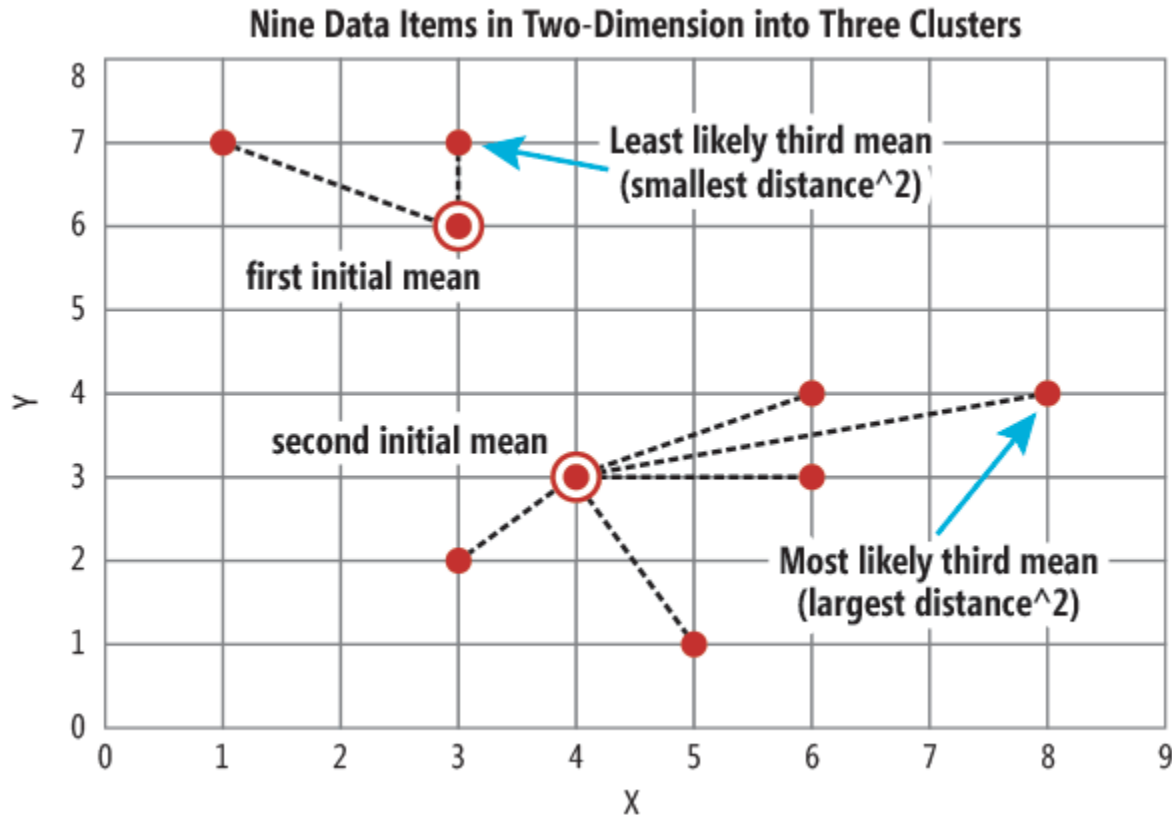
K-Means: Random Initialization Trap



<https://www.superdatascience.com/blogs/self-organizing-maps-soms-extra-k-means-clustering-part-2/>

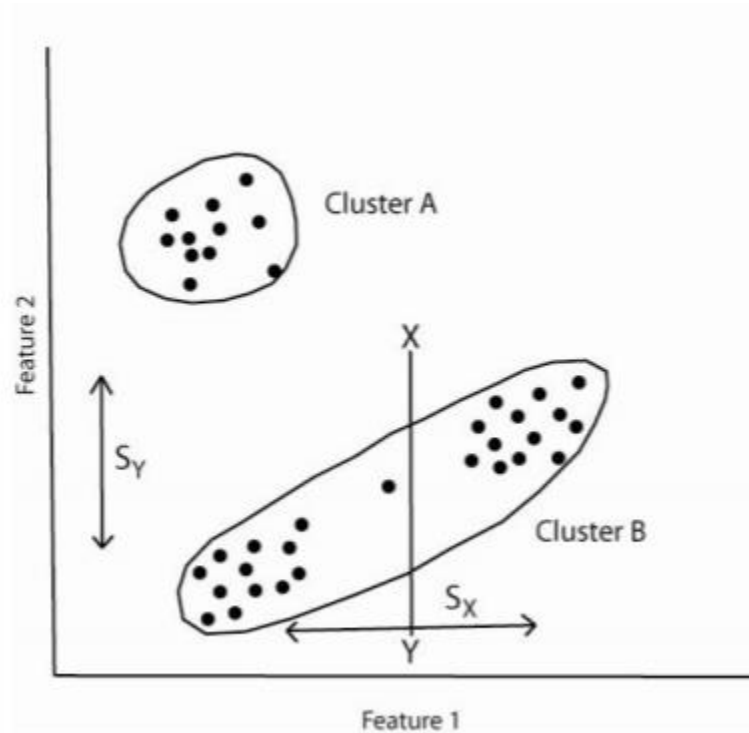
Solutions to Random Initialization Trap

- K-Means++



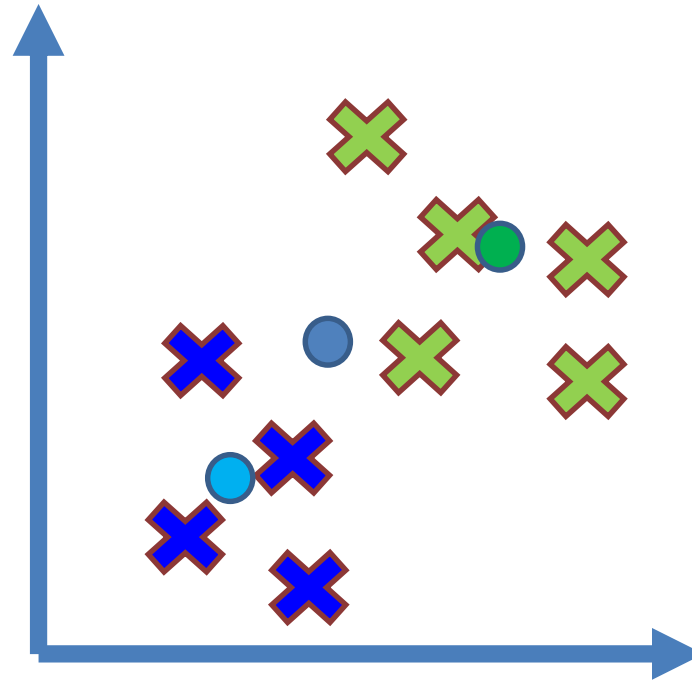
Solutions to Random Initialization Trap

- Iterative Self-Organizing Data Analysis Technique (ISODATA)

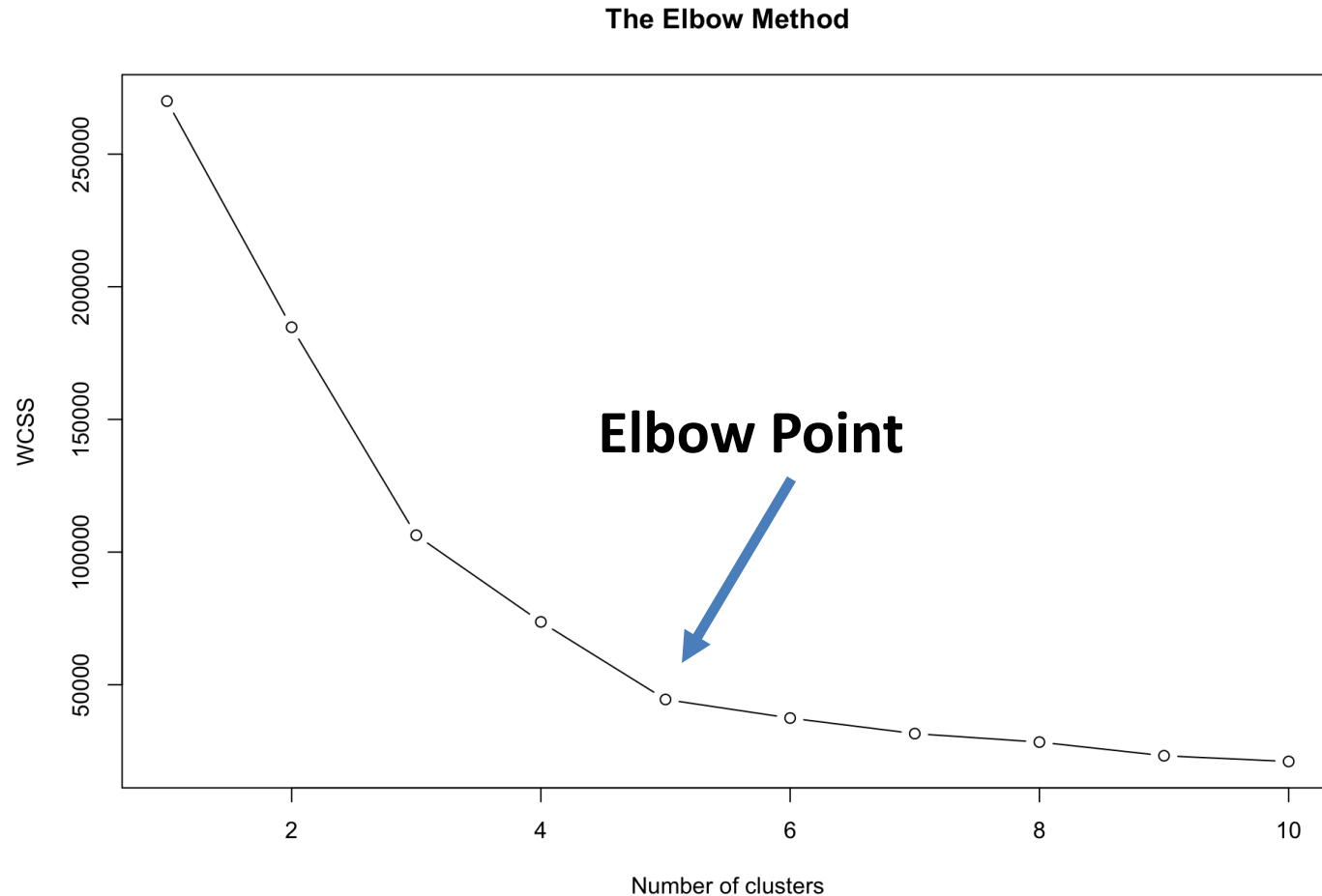


Determine the Number of Cluster (K)

Within Cluster Sum of Squares $\sum_{x_i \in c} (x_i - \bar{x})^2$



Determine the Number of Cluster (K)



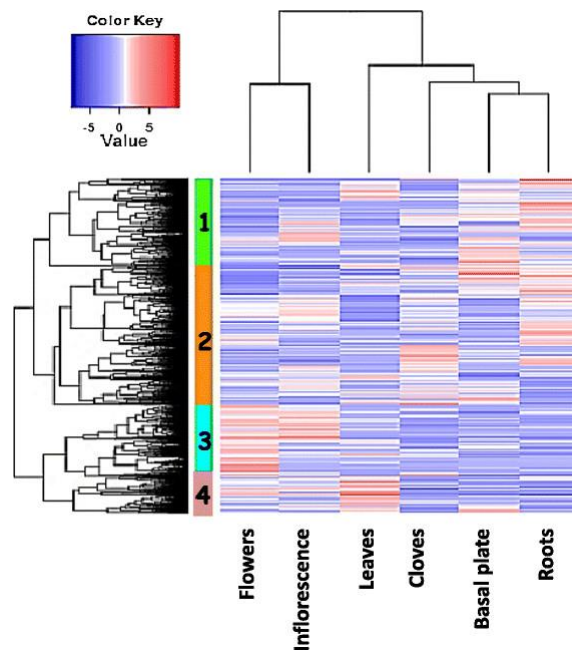
<https://rpubs.com/markloessi/499223>

Cluster Analysis is Used For...

- Data Reduction
 - A researcher may be faced with a large number of observations that can be meaningless unless classified into manageable groups.
- Hypothesis Generation
 - Cluster analysis is also useful when a researcher wishes to develop hypothesis concerning the nature of the data or to examine previously stated hypothesis.

Cluster Analysis is Used For...

- Relationship Identification
 - Define the structure of data by placing the most similar observations into groups.

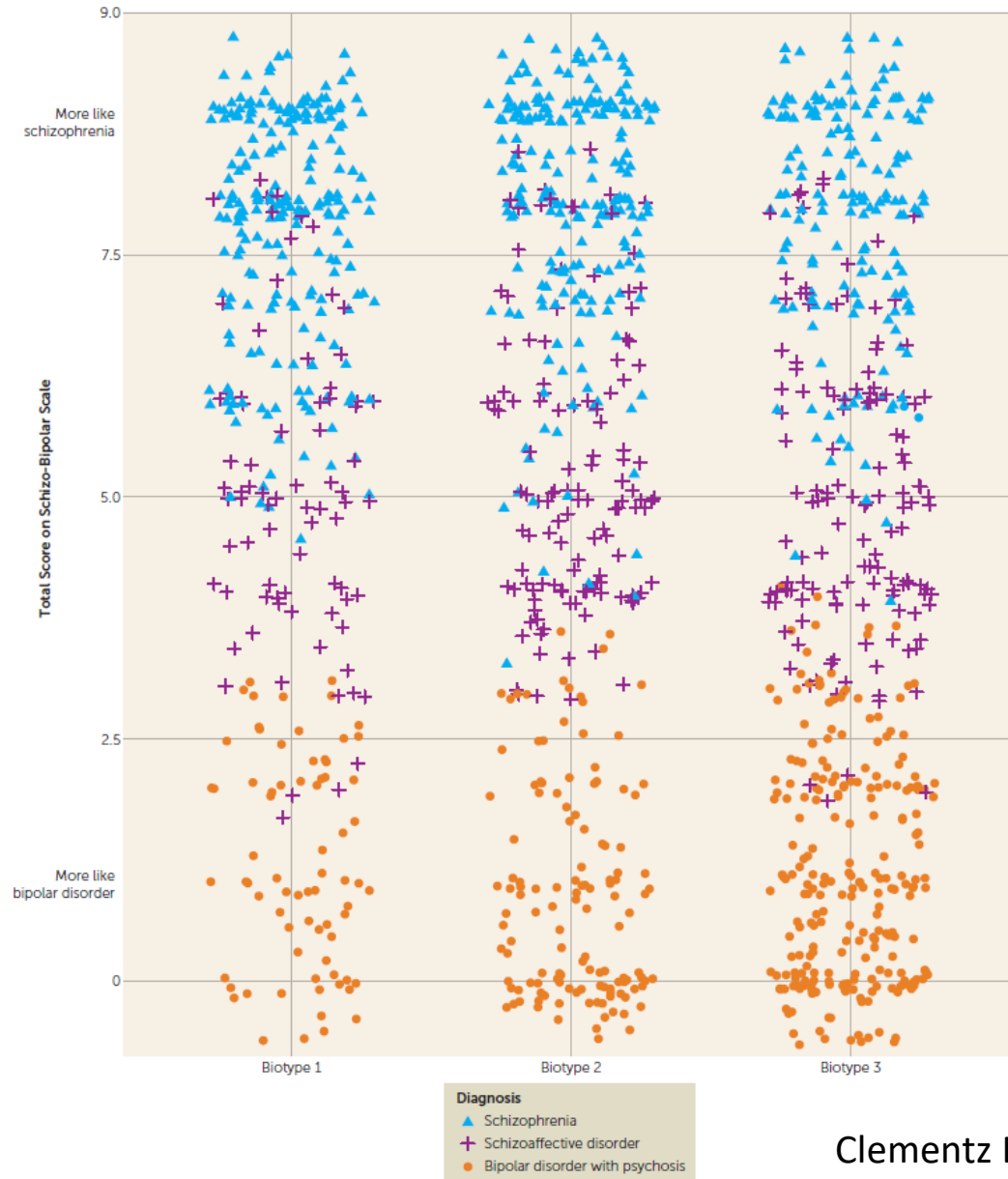


https://www.researchgate.net/figure/Hierarchical-cluster-analysis-of-gene-expression-patterns-in-six-vegetative-and_fig5_271222113

Limitations

- More descriptive than inferential
- Clusters are sometimes arbitrary
- Clusters are totally dependent upon the variables used

Identification of Distinct Psychosis Biotypes Using Brain-Based Biomarkers





GAP

Generalized Association Plots

GAP is a java-designed software for generalized association plots (Chen, 2002) and exploratory data analysis. It is programmed for the java runtime environment 1.5 (JRE version 1.5.0_04), which is available for most operating systems.

Last Updated: 2012/05/15

Authors:

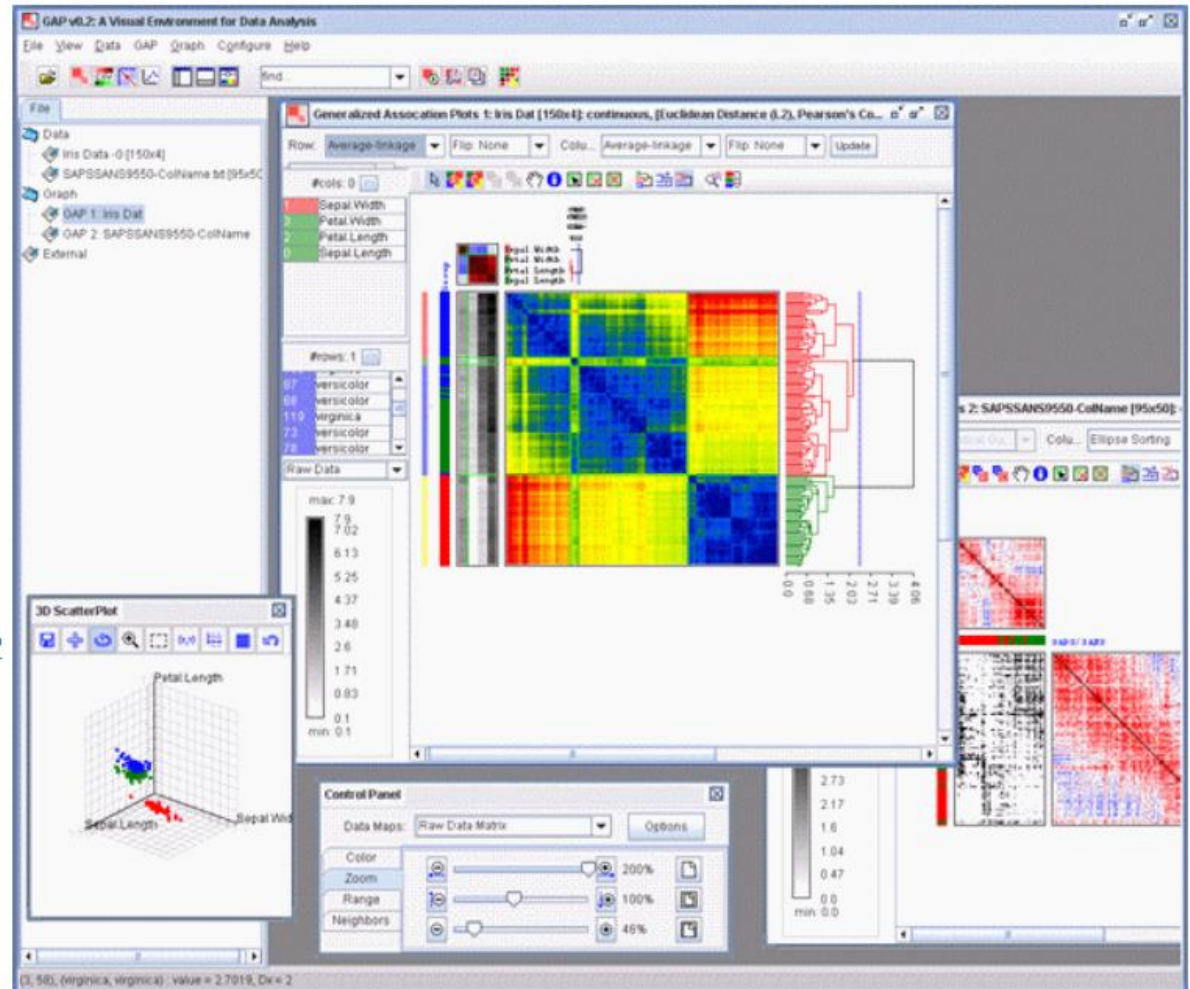
Dr. Han-Ming Wu
 Department of Mathematics,
 Tamkang University
 Tamsui, Taiwan, R.O.C.
 E-mail: hmwu AT mail.tku.edu.tw
 Homepage:
<http://www.hmwu.idv.tw>

Dr. Chun-houh Chen
 Institute of Statistical Science,
 Academia Sinica
 Taipei, Taiwan, R. O. C.
 E-mail: cchen AT stat.sinica.edu.tw
 Homepage:
<http://gap.stat.sinica.edu.tw>

Official Website of GAP Software:
<http://gap.stat.sinica.edu.tw/Software/GAP>

Mirror Website
<http://www.hmwu.idv.tw/GAPSoftware>

Current Version:
 v0.2.7, Build 20110331



The GAP Main Window [more screenshots...]