Walmart

Cluster Analysis

Mallikharjuna MV & Ojaswini Chhabra
International Data & Analytics

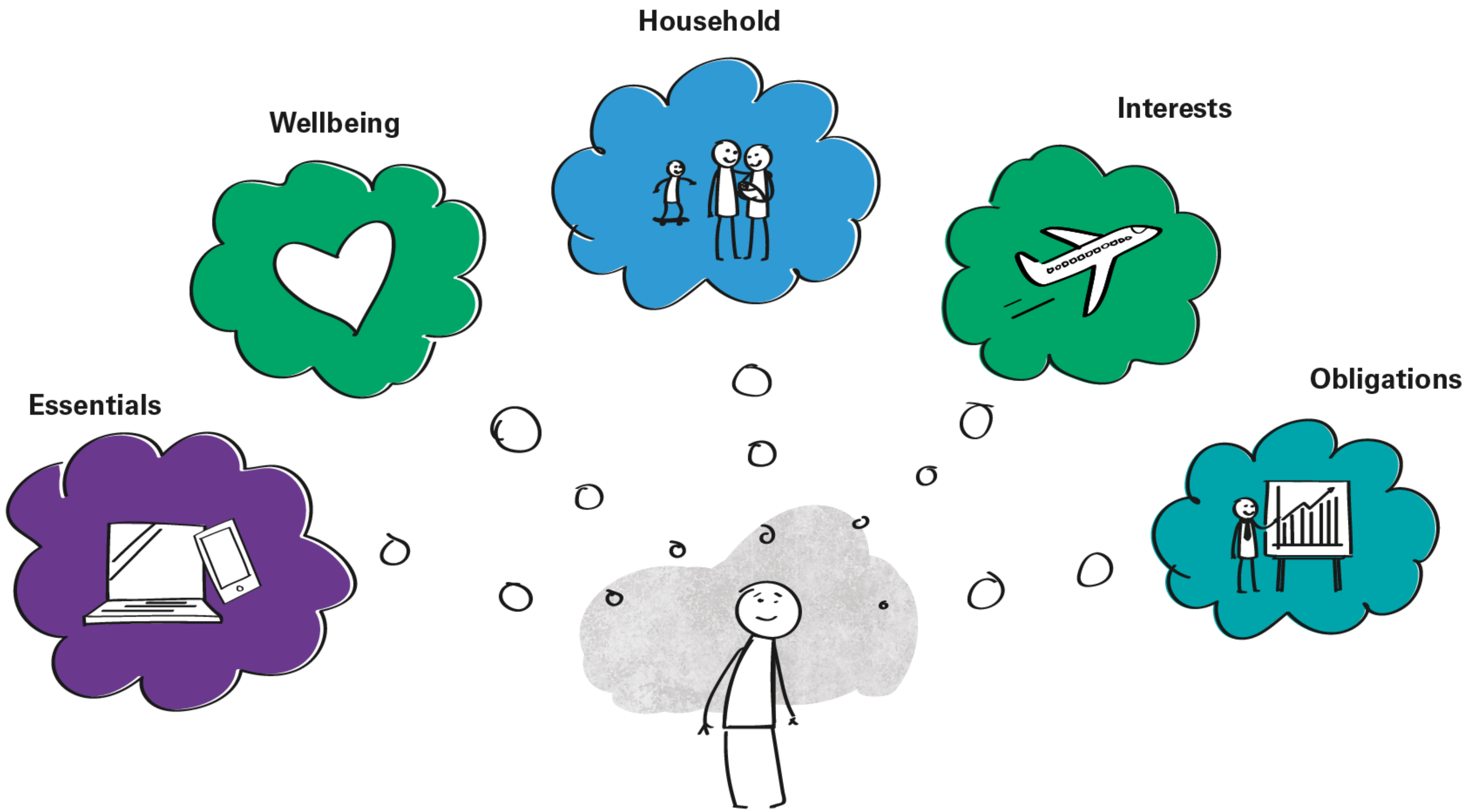
11-Sep-2020

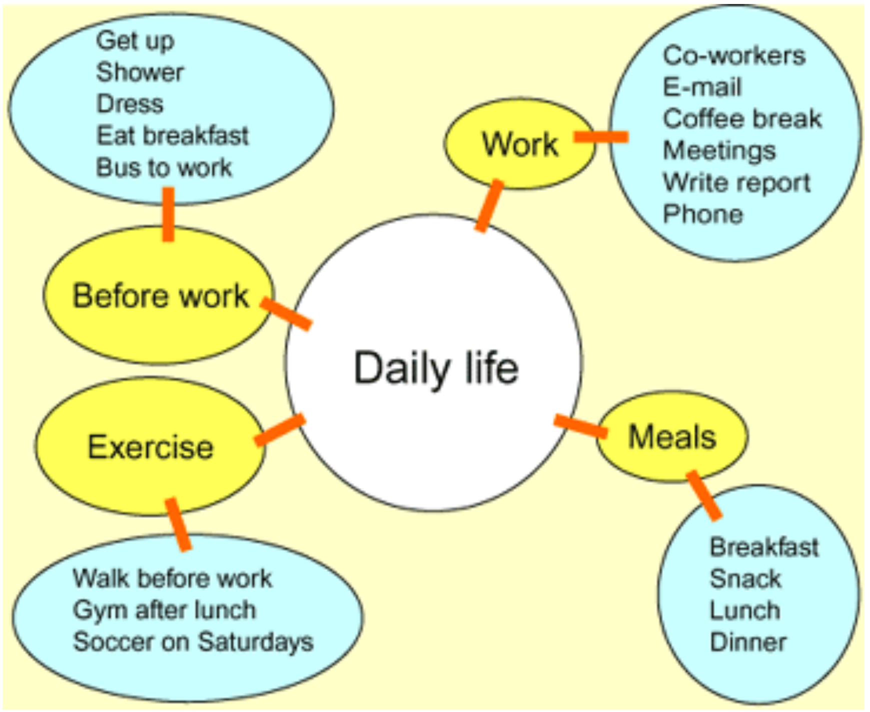
Agenda

- Clustering – An Overview
- What is Clustering?
- Common Distance Measures
- K-Means Clustering & DBSCAN Algorithms
- Walmart Use Case
- Q&A

CLUSTERING – AN OVERVIEW

Clustering in Daily Lives

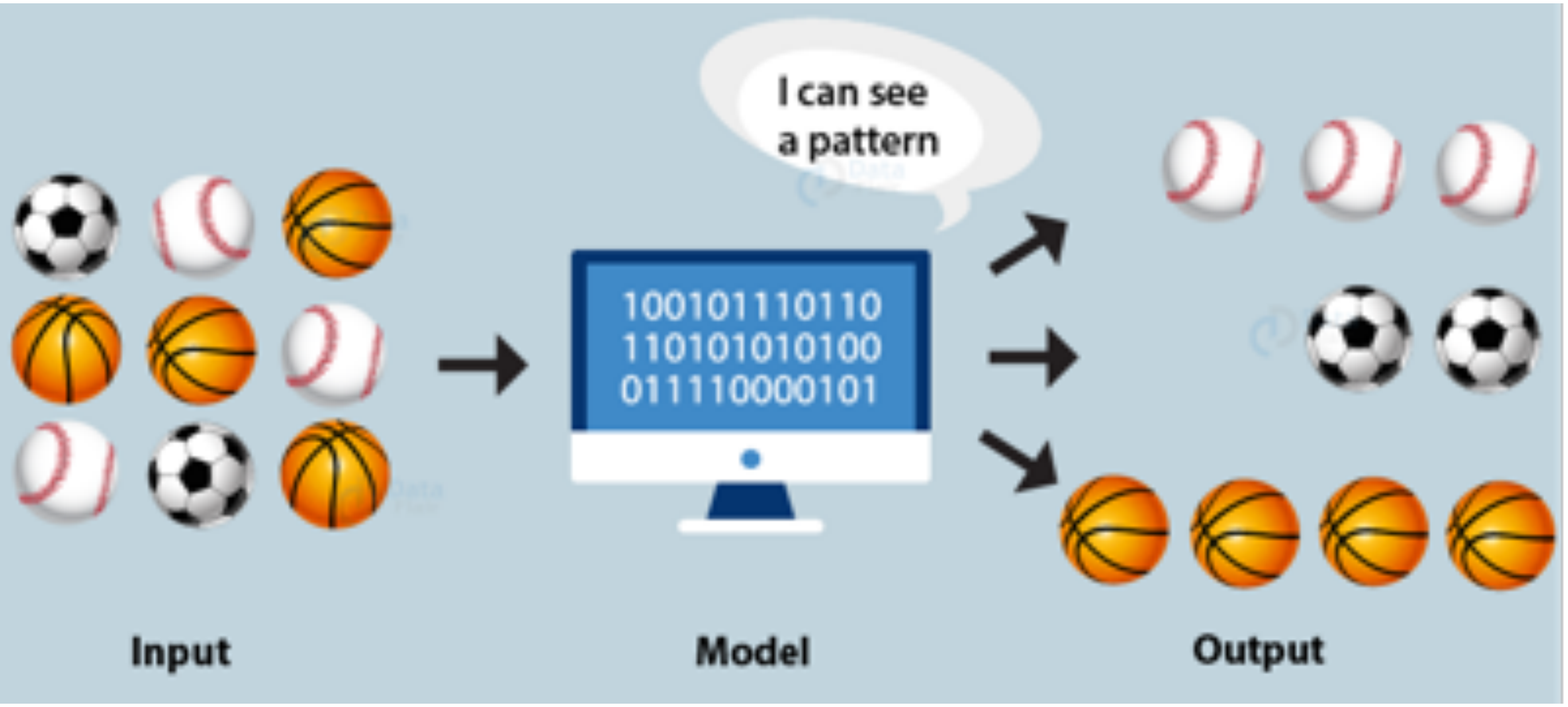




https://www.liveworkstudio.com/insights/how-people-cluster-experiences-in-their-life/

https://medium.com/data-science-group-iitr/clustering-described-63e62833099e

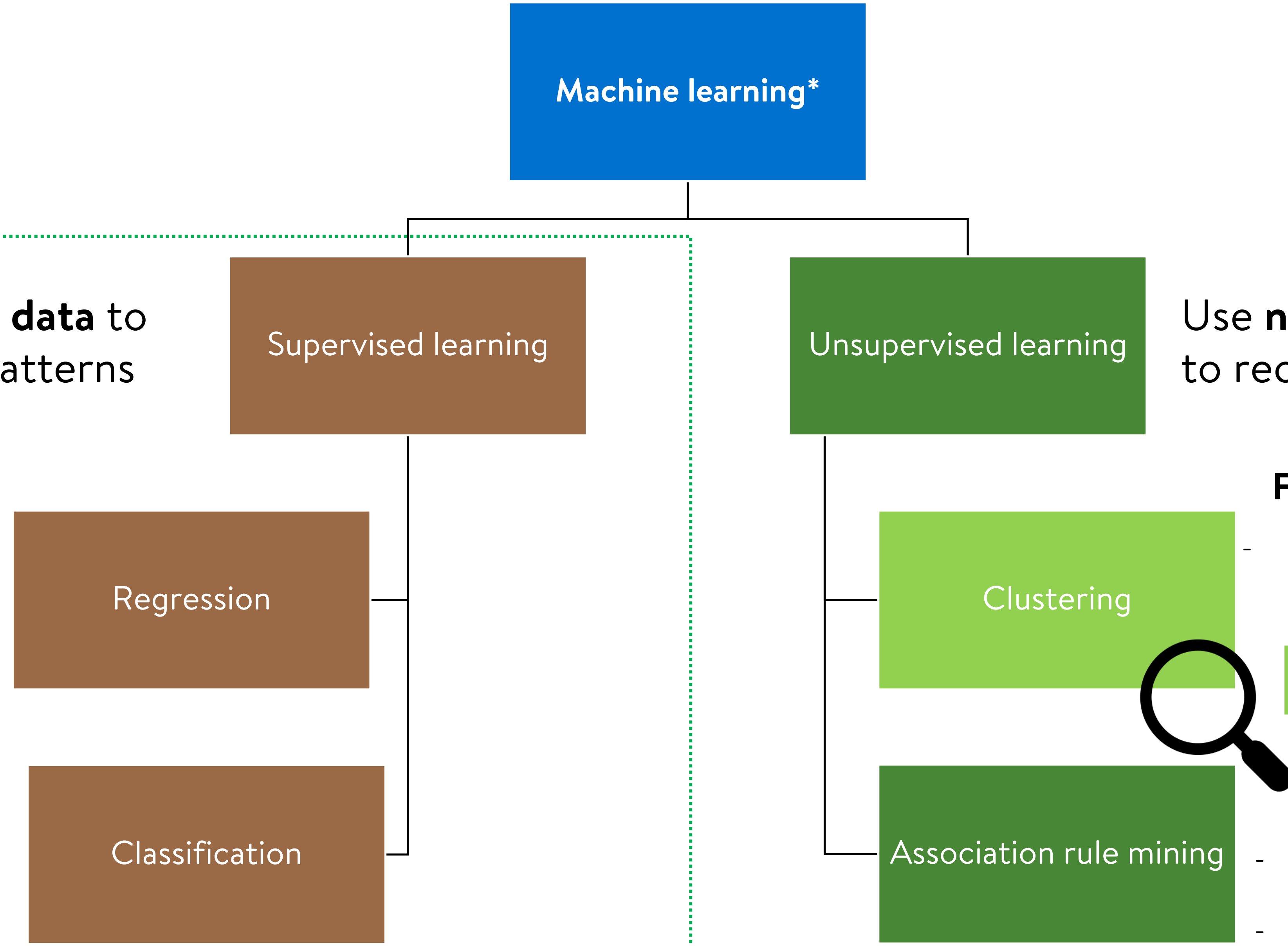
Clustering by Machine Learning



https://data-flair.training/blogs/clustering-in-machine-learning/

A brief recount of machine learning

Broadly, the science of detecting patterns in data



Use **labeled data** to recognize patterns

Use **non-labeled** data to recognize patterns

Predicting a continuous variable

- *How much $'s will the customer spend in Walmart*
- *How much $'s will the customer spend in Walmart grocery*

Predicting a class

-Will a customer buy/not buy from us in the next 3 months ?
-Will a customer buy/not buy on Online grocery in the next 6 months ?

Finding segments in data

- *What are the different Walmart customer segments*

Focal area

Finding association between units of analysis

- *If a customer buys beer, what is he likely to buy as well in the same basket?*
- *If a customer buys a lawnmower & a rake n his current transaction – what is he likely to buy in his next transaction*

What is Clustering?

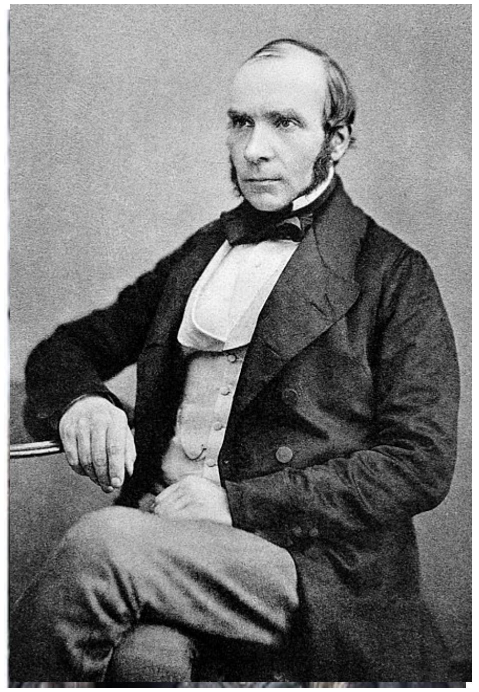
What is Clustering?

Clustering is a process to discover hidden structures of the data, possibly as a prelude to more focused analysis or decision processes:

1. A way to decompose a data set into subsets with each subset representing a group with similar characteristics.
2. When we cluster observations, we seek to partition them into distinct groups such that objects in the same group are more similar to each other in some sense than to objects of different groups.
3. The groups are known as clusters, and each cluster gets a distinct label called cluster id

Historic application of clustering

In the middle of the 19th century a Cholera claimed lives across Asia, Africa, and Europe.



John Snow a London Physician, marked the places where the deaths happened on a map and used that data to cluster. An interesting pattern emerged— the cases were clustered around certain intersections where there were polluted wells.

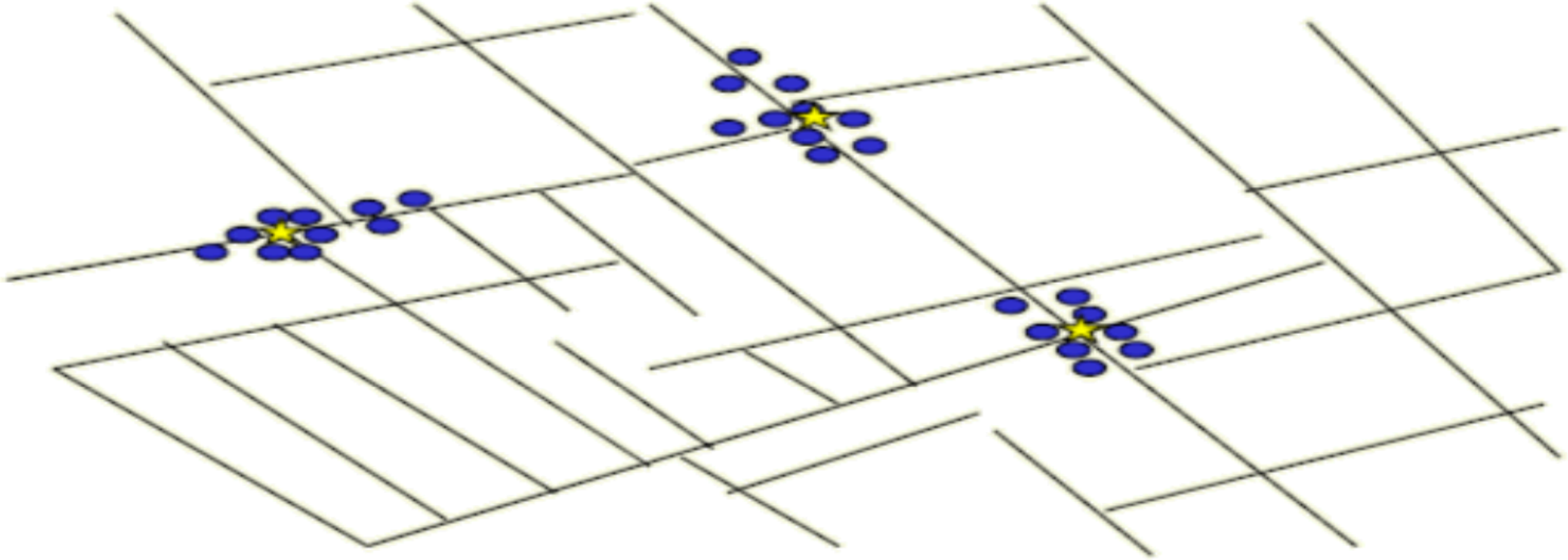
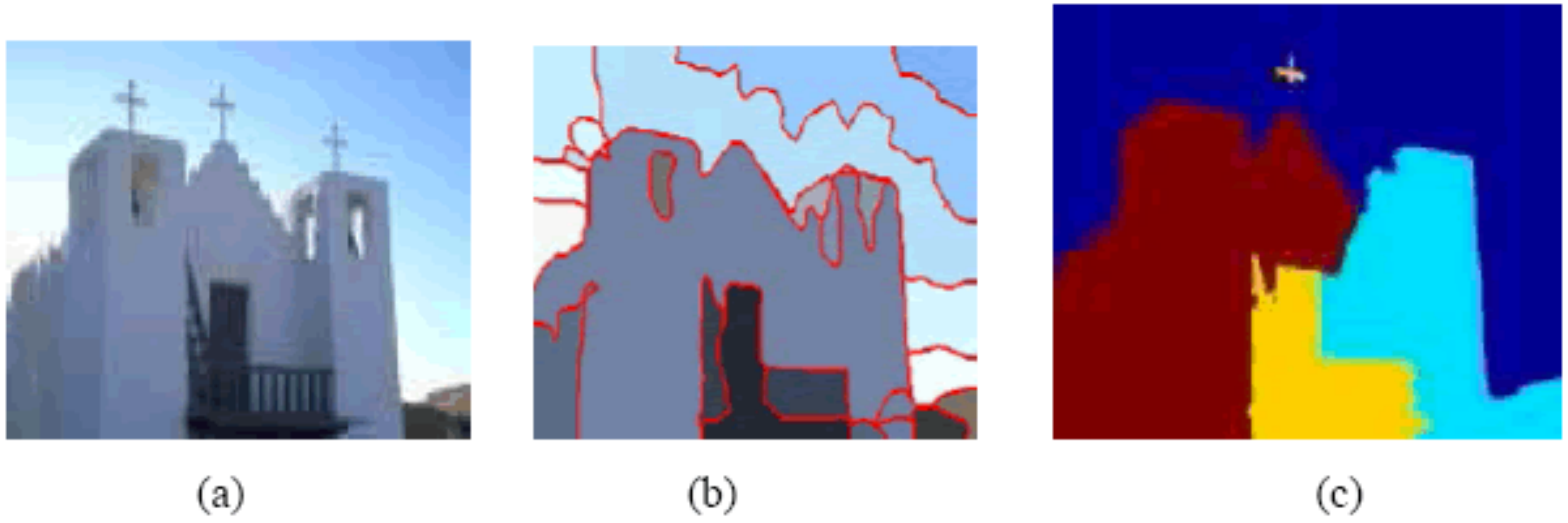


Image Segmentation



(a)

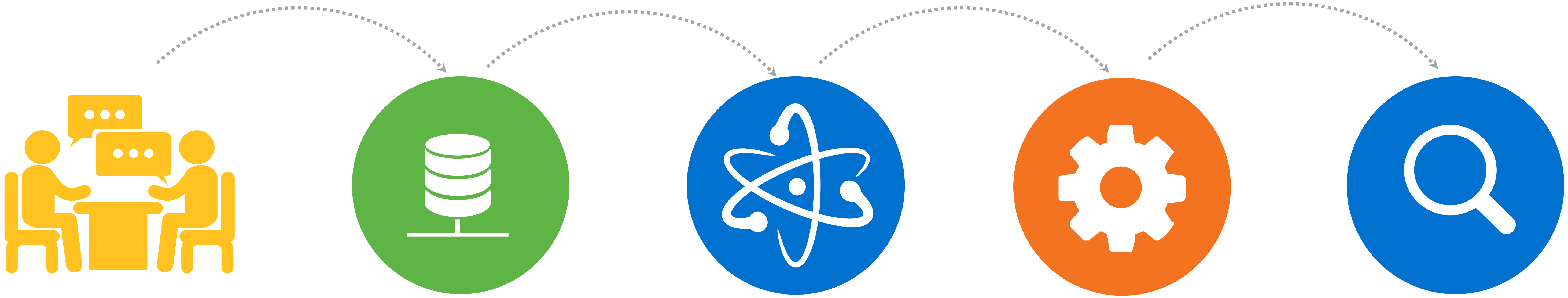
(b)

(c)

Figure 1: (a) is the original image; (b) and (c) are the segmentation results.

https://www.hilarispublisher.com/open-access/image-segmentation-by-using-linear-spectral-clustering-2167-0919-1000143.pdf

What do we need for Clustering



Objective

Data Acquisition & Processing

Proximity/ Distance Measure

Clustering Technique

Criterion function to Evaluate

Distance Measures

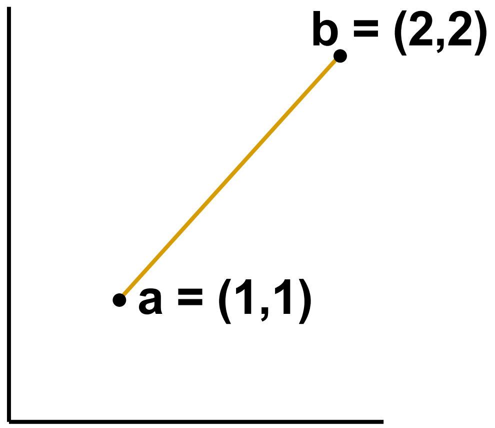
Distance Measure for Continuous Variables

Irrespective of the clustering algorithm, we need a way of defining distance measure. This is central to all the goals of cluster analysis.

Distance $\left(\mathrm{d}(\mathrm{a},\mathrm{b})\right)$ will be small if records a & b are similar

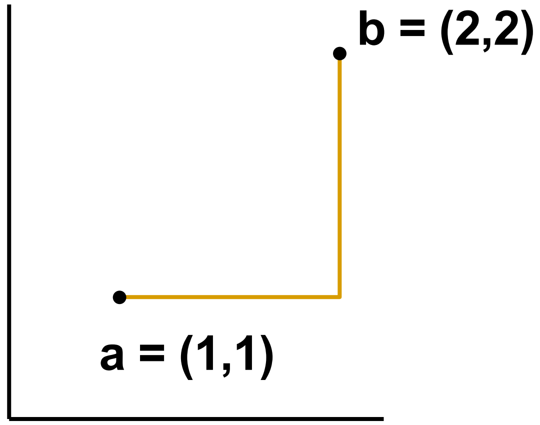
Euclidian Distance

$$d(a,b) = \sqrt{\sum_{k=1}^{N} (a_k - b_k)^2} = \sqrt{2}$$



Manhattan Distance

$$d(a,b) = \sum_{k=1}^{N} |a_k - b_k| = 2$$

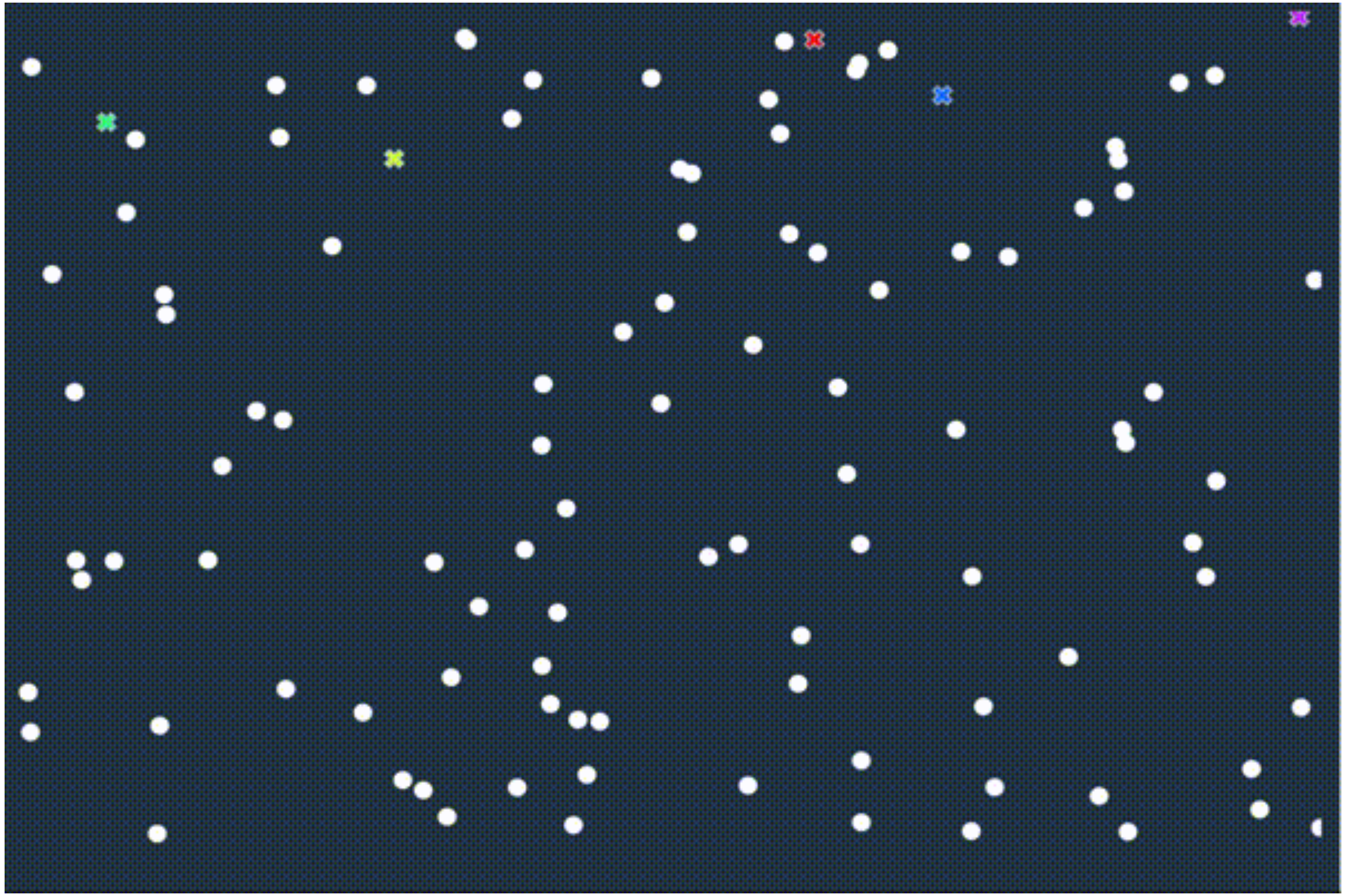


Distance Measure Properties

$\mathrm{d}(\mathrm{a},\mathrm{b}) = \mathrm{d}(\mathrm{b},\mathrm{a})$	Symmetric
$\mathrm{d}(\mathrm{a},\mathrm{a}) = 0$	Self − Similarity
$\mathrm{d}(\mathrm{a},\mathrm{b}) \geq 0$	Non − Negative
$\mathrm{d}(\mathrm{a},\mathrm{c}) + \mathrm{d}(\mathrm{b},\mathrm{c}) \geq \mathrm{d}(\mathrm{a},\mathrm{b})$	Triangle Inequality

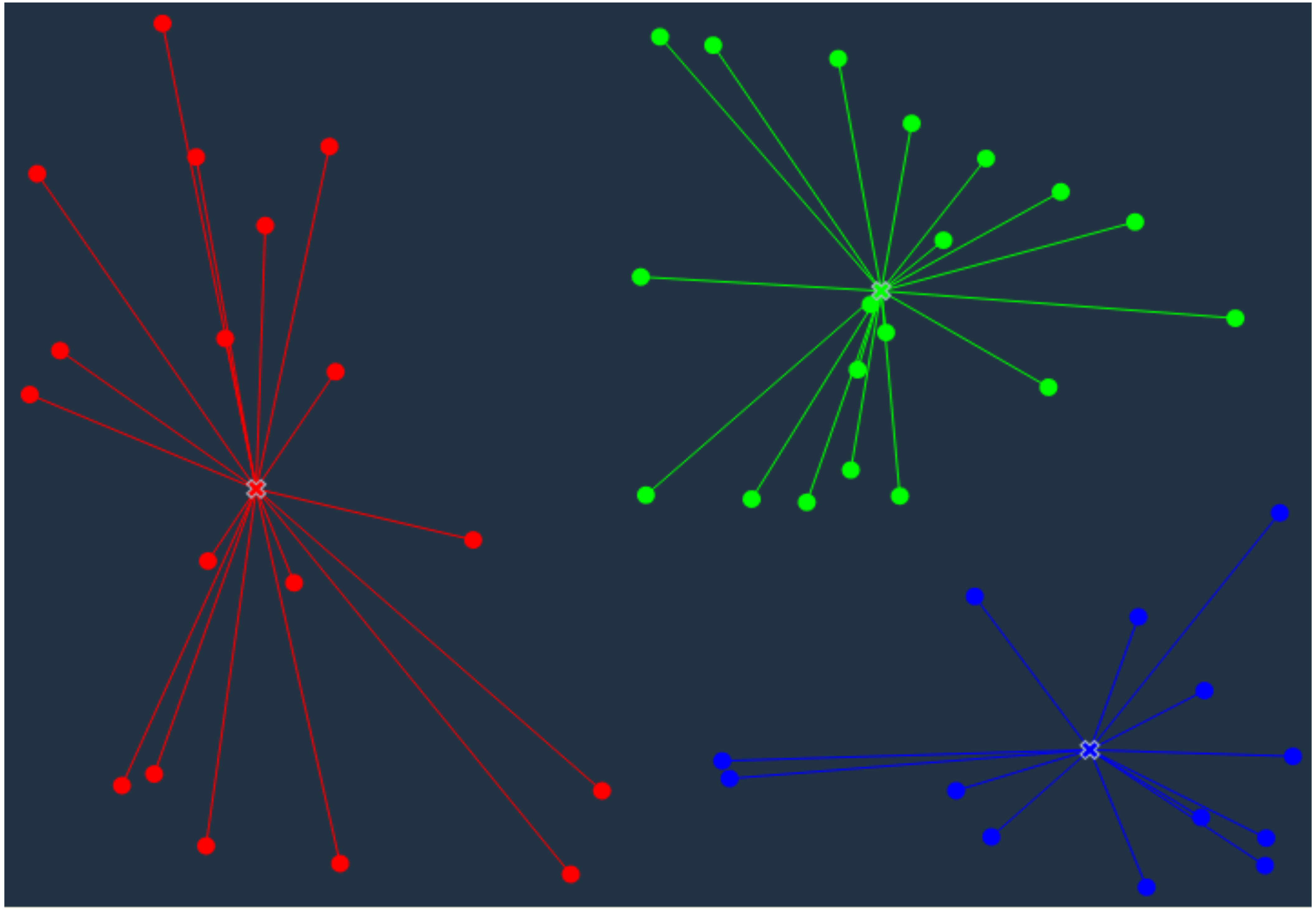
K-Means Clustering

K-Means Clustering Visual Flow



https://yourstory.com/2019/04/data-science-clustering-for-future

K-Means Clustering



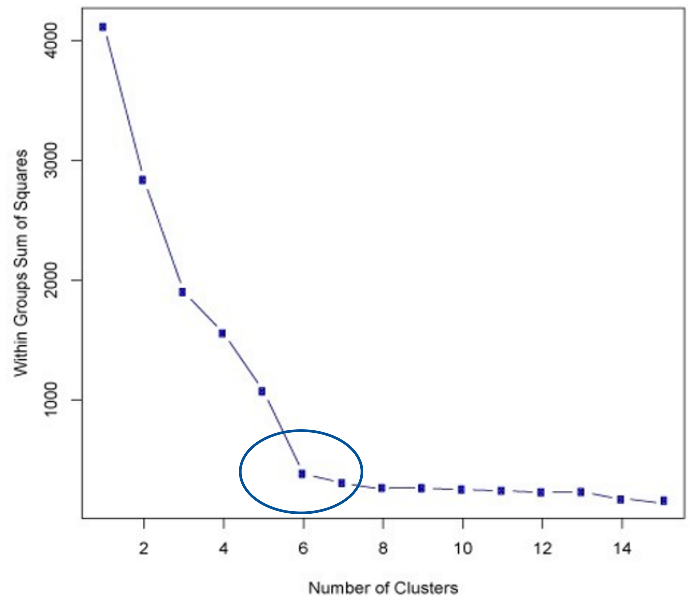
https://yourstory.com/2019/04/data-science-clustering-for-future

In the update phase the centroid of each cluster is calculated

Note : based on the distance of some boundary points from the new centroid, they get reallocated to different cluster

K-Means Clustering Steps

1. K-means input data: continuous variables
2. Standardize the variables to common scale. For example, $\frac{X-Average(X)}{StdDev(X)}$
3. K-means clustering partitions data into K disjoint sets or clusters where K is a pre-specified number. It can range from 1 to n where n is number of data points
4. Iterating stops when Loss Function $SSE = \sum_{i=1}^{K} \sum_{x\in C_i} d(x, cenriod\ of\ Cluster_i)^2$ converges
5. Experiment with different values of K
6. Select optimal K by plotting SSE plot



Advantages and Disadvantages

Advantages :

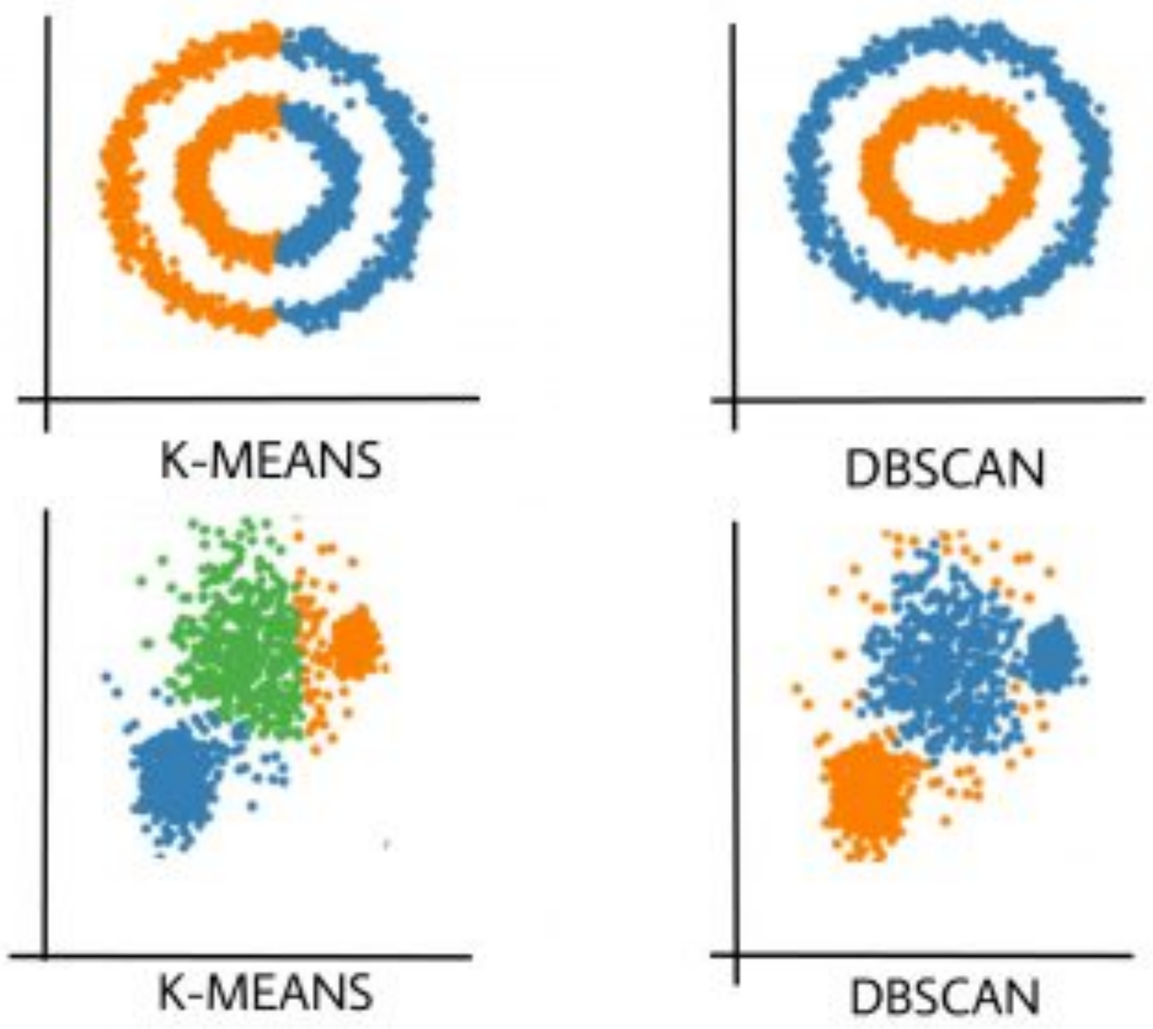
- Simple: Easy to understand and to implement
- Efficient: Time complexity - O(tkn), where
    - n is the number of data points
    - k is the number of clusters
    - t is the number of iterations.

Disadvantages:

- User needs to specify the number of clusters (k)
- Only applicable if the mean is defined
- Sensitive to outliers
    - Outliers are the data points which are far away from the other points or errors in the data

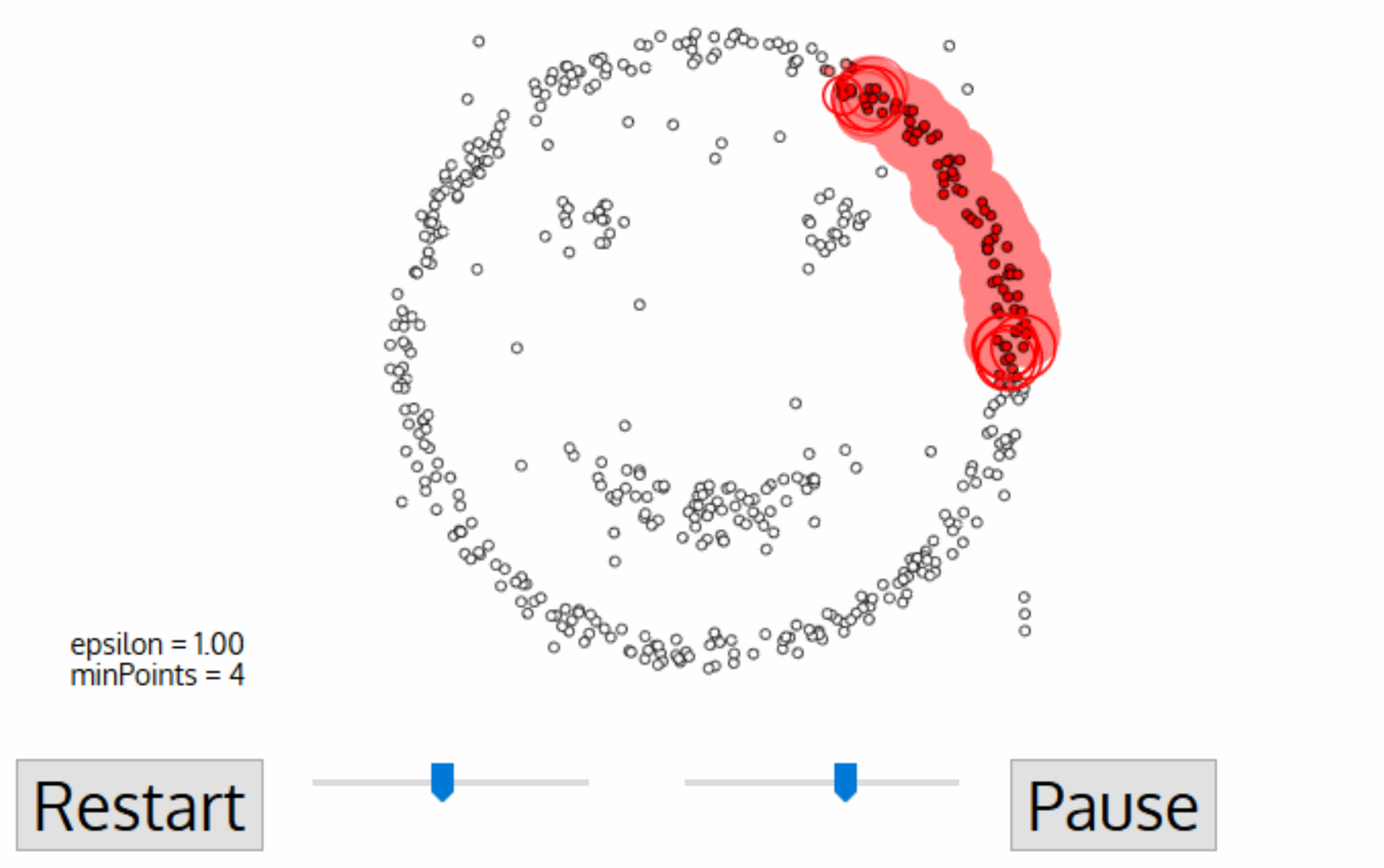
Where K-Means Fails

K-Means Clustering fails in cases of non-round shaped clusters or different density clusters!



DBSCAN

DBSCAN – Visual



epsilon = 1.00
minPoints = 4

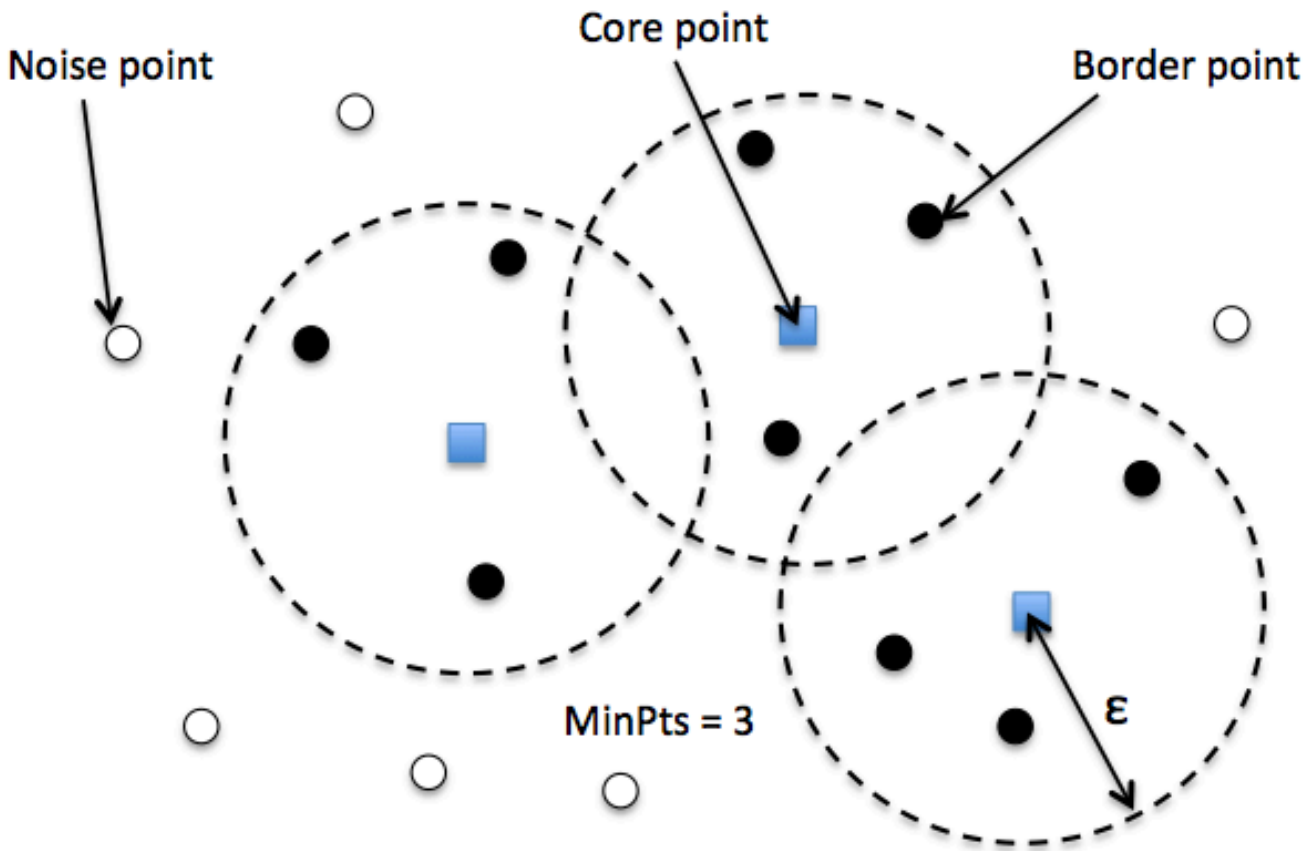
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

DBSCAN - Overview

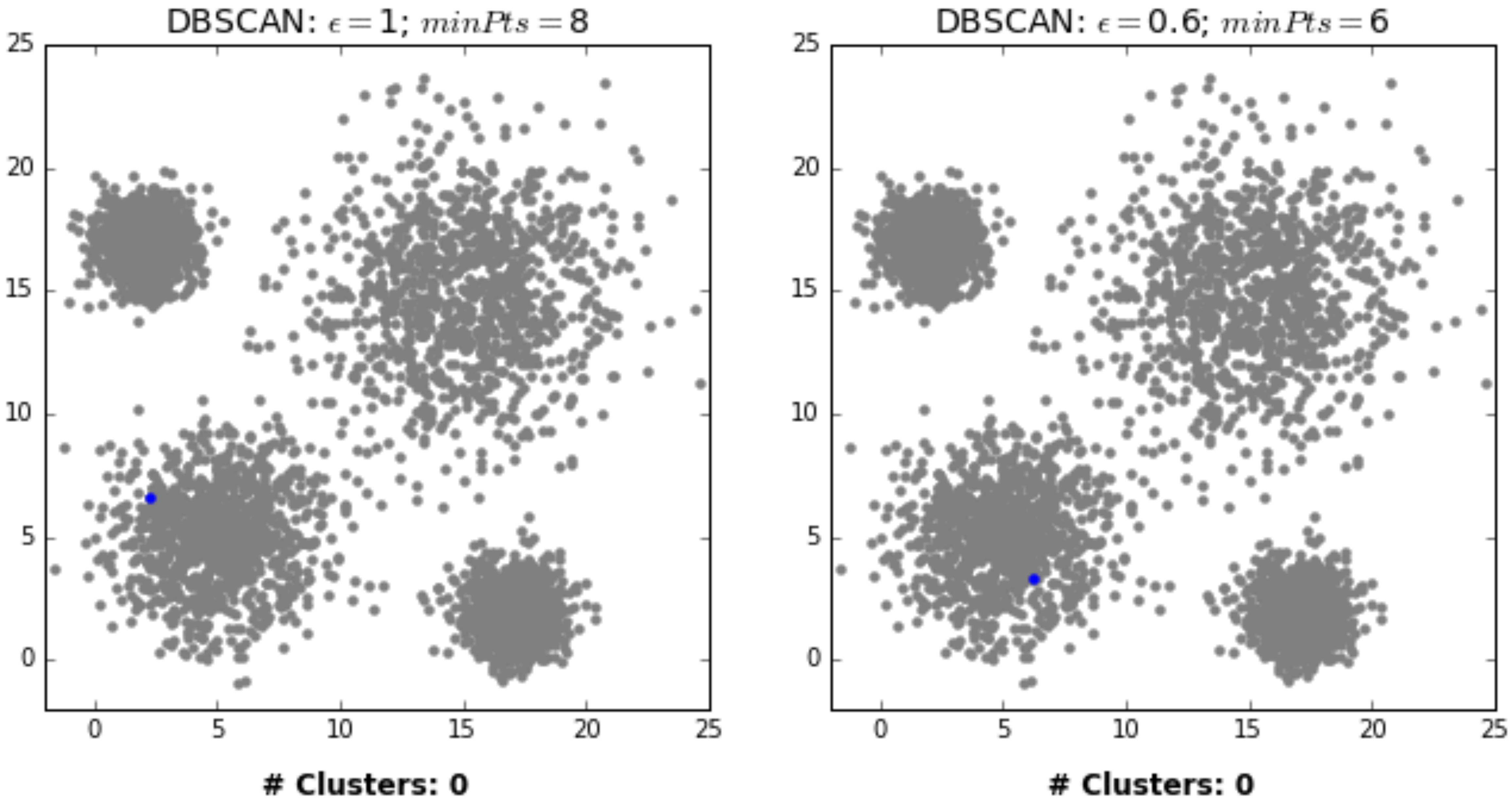
1. DBSCAN means **D**ensity **B**ased **S**patial **C**lustering of **A**pplication with **N**oise
2. Input Parameters:
   - Epsilon (ϵ): The size of epsilon neighborhood
   - MinPts: Minimum Points in the neighborhood
3. Density at point p: number of points within a circle of radius ϵ

DBSCAN – Core, Border & Noise Points

1. A point is a core point if the density at that point has more than a specified number of points (MinPts)
2. A border point density has fewer than MinPts, but is in the neighborhood of a core point
3. A noise point is any point that is not a core point or a border point.



DBSCAN – Visual



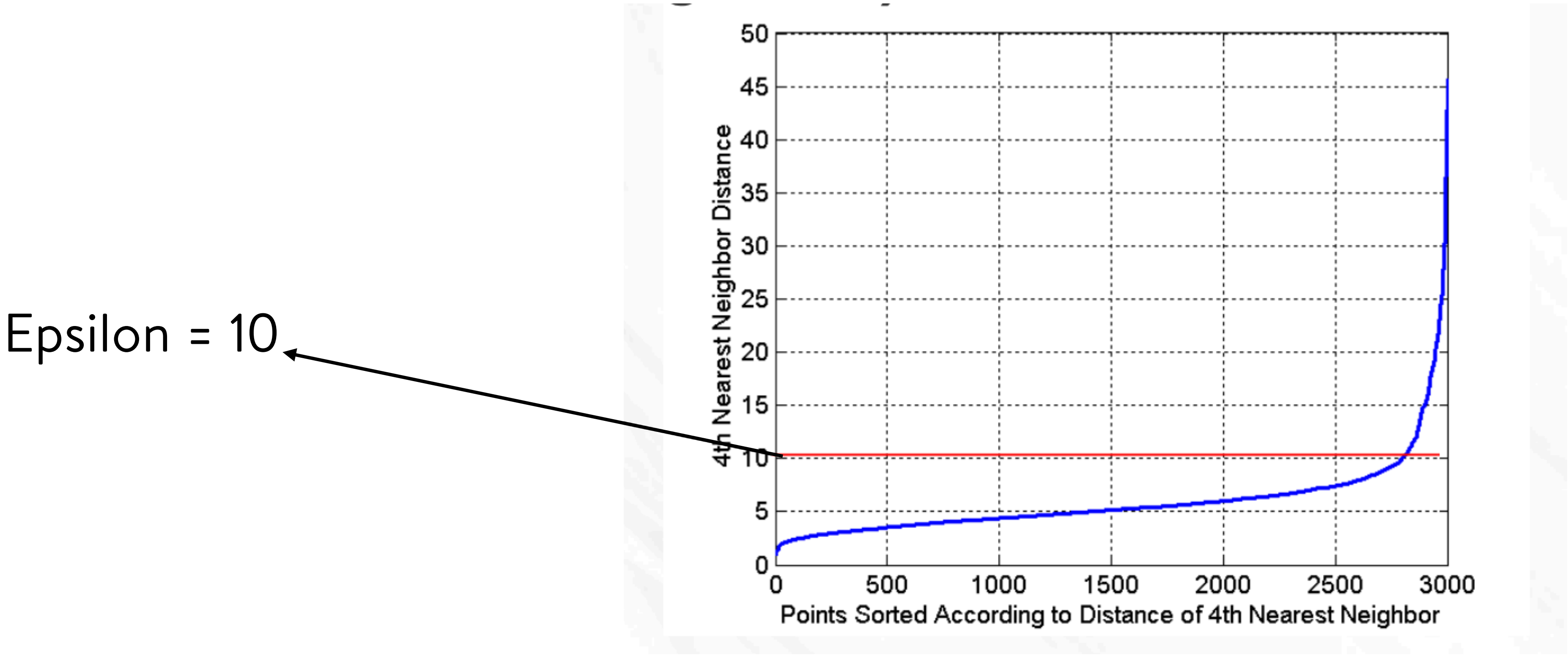
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

DBSCAN – Steps

1. Define suitable distance measure
2. Select suitable MinPts & Epsilon
3. Let ClusterCount=0. For every point p:
   1. If p it is not a core point, assign a null label to it [e.g., zero]
   2. If p is a core point, a new cluster is formed [with label ClusterCount:= ClusterCount+1]
   3. Then find all points density-reachable from p and classify them in the cluster.
   4. Repeat this process until all the points have been visited.

DBSCAN – Determining Epsilon

1. Choosing the value of the epsilon is not obvious.
2. There is a heuristic method we can use to at least get directional input. Calculate the distance from each point to its kth-nearest neighbor (eg. K=4) and then ordering the points in a plot based on this distance.
3. This plot tends to produce a plot containing a “knee” or “elbow”. The optimal value of epsilon is in or near that knee/elbow.



http://csc.csudh.edu/btang/seminar/slides/DBSCAN.pdf

Advantages and Disadvantages

Advantages :

- Does NOT require to specify the number of clusters
- Can find arbitrary shaped clusters
- Can identify outliers easily

Disadvantages:

- Sensitive to parameters epsilon and minpts
- If the data and scale are not well understood, choosing a meaningful distance threshold epsilon can be difficult.
- Does not work as well with clusters of varying density.

K-Means vs DBSCAN

K-Means Clustering	DBSCAN
User needs to specify the number of clusters (k)	Does NOT require to specify the number of clusters
Easy to understand and to implement	If the data and scale are not well understood, choosing a meaningful distance threshold epsilon can be difficult.
K-Means clustering fails in cases of non-round shaped clusters	Can find arbitrary shaped clusters
Sensitive to outliers	NOT sensitive to outliers

WALMART CASE STUDY:
Abbreviation and Spell Correction of Item Descriptions

Authors:
Ojaswini Chhabra
Mallikharjuna MV
Vivek Damodaran
Lance Levenson

Examples of item descriptions

- Item descriptions contains different representation of same word.
- For example, Backpack is mentioned as BKPK, BPACK…

COUNTRY_CODE	MDS_FAM_ID	FINELINE_DESC	SIGNING_DESC	ITEM1_DESC
US	109656710	PR LICENSE BACKPACKS	BARBIE BKPK 17"	PR BARBIE BKPK PRM
CA	60354268	MOD FASHION BKPKS	POLY BPACK GR 3PCE SET BUTTERFLY BLUE	GR 3PC BTTERFLY BKPK
US SAMS	54134840	DIVERTED HANDBAGS	HERSCHEL BKPK	HERSCHEL BKPK
US	87549707	APL RED BULK	RED DELICIOUS APPLES BULK	APL RED DEL 26
US	208315512	PR YOGURT SINGLE S	YOP ORG LF YGRT CHERRY STARBURST SS	PR YOP STRBURST CHRY
US	104135324	CHEESE SHREDDED	GREAT VALUE EXTRA SHARP CHED SHRED 16 O	GV SHR CHD 16Z
US	44762205	PREMIUM COFFEE	ALTO GRANDE COF	PR ALTO GRANDE COF
US	81226393	PINEAPPLE	DEL MONTE SLCE PNPPLE IN OWN JUICE 20 OZ	DM SLICE PINE
CA	57740594	ORGANIC FRUIT	ORGANIC GRAPES RED SDLS 454 GR	ORG GRAPE RED 454GR

Applications of Item Descriptions Cleaning

Walmart

Common Intl. Item Taxonomy

Corrected descriptions help in creating common item taxonomy across international markets.

Item Global Search

It helps the users in searching for relevant items in the SPInE tool.

Item Similarity

Text cleaning helps in item similarity model performance.

SPInE tool (wmlink/spine)

Common Descriptions Errors

Walmart

- Item descriptions are often described in similar (but not the same) manner.
- Eliminating duplicates & getting a unique set of words is a crucial pre-processing step to decrease the corpus size & increase prediction accuracy.
- There can be different kinds of correction types to remove redundancy:

Translation

Abbreviation Matching

Spell Correction

Singularization

CORRECTION TYPE	WORD	CORRECTION
TRANSLATION (Using Azure API & Spanish to English Dictionary)	MANZANA	APPLE
	TAMARINDO	TAMARIND
ABBREVIATION MATCHING	APL	APPLE
	WMELON	WATERMELON
SPELL CORRECTION	VAELNTINE	VALENTINE
	CARRIBBEAN	CARIBBEAN
SINGULARIZATION	APPLES	APPLE
	BACKPACKS	BACKPACK

Text Correction Foundations

ABBREVIATION MATCH RULE

- Matching a word to the closest word in the corpus that contain all the letters of the original word in the exact same order. This connects abbreviations to potential word matches.
- Example: BACKP matches with BACKPACK.

SPELL CORRECTION

- Getting a list of the closest dictionary suggestions for each word. This takes into account the number of deletions, insertions, or substitutions required to transform the original word to a potential match. Helps in matching the words that are misspellings, rather than abbreviations.
- Example: BACKPAK matches with BACKPACK.

SINGULARIZATION & TRANSLATION

- Plural words are converted to singular using NLP libraries as well as 's correction' rules.
- Words from other languages are translated to English using Azure API and Google Translate.
- Example: MOCHILA translates to BACKPACK (Spanish to English).

Text Correction Process

01

Pre-Processing

This involves converting to upper case, removing small, infrequent, and duplicate words, and creating a DTM.

02

Rule Based Correction

Abbreviation, Spell Correction, Translation, and Singularization rules are applied to match each word in the corpus with its closed similar word.

03

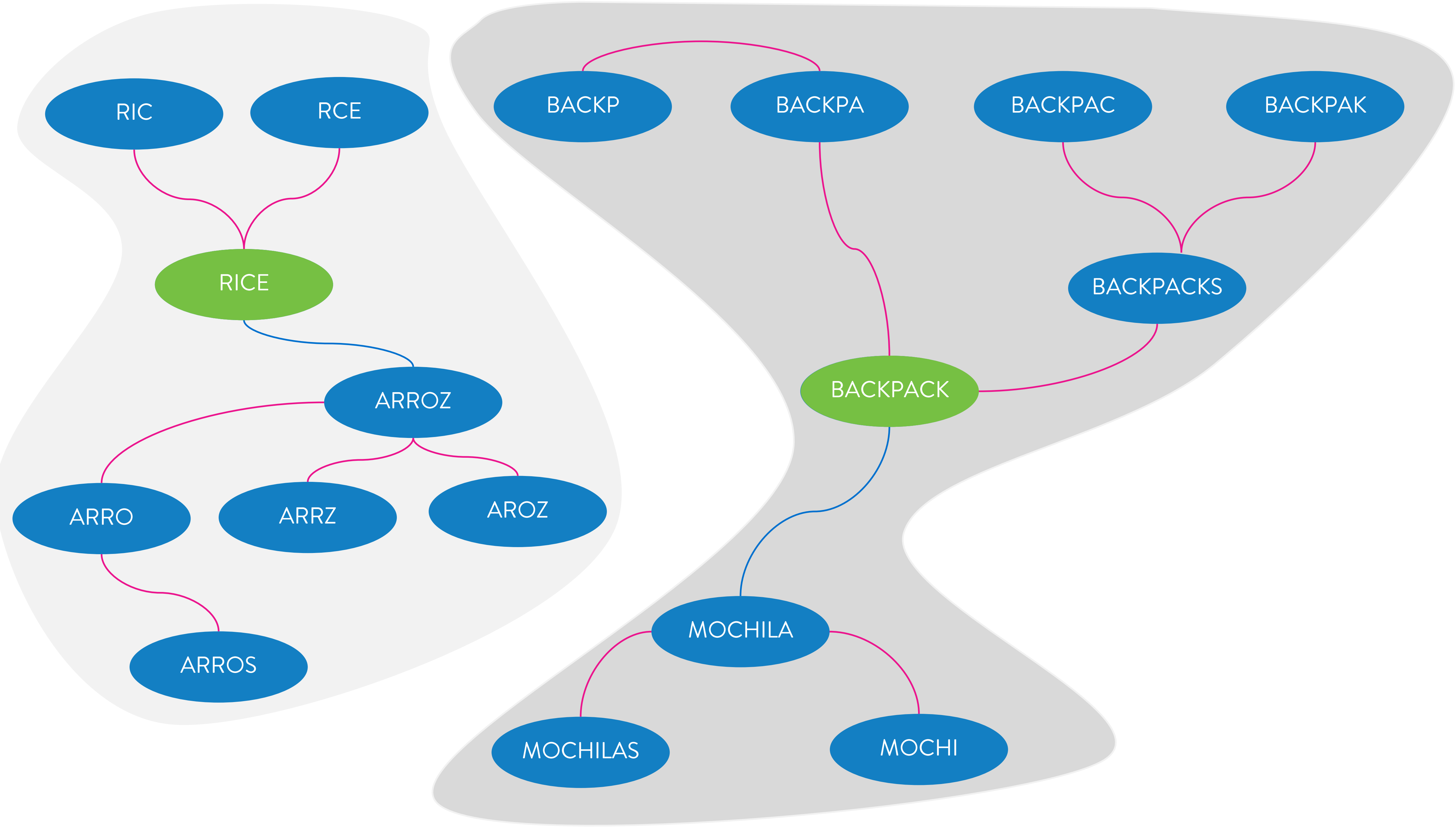
Word Clustering

DBSCAN is applied on the key-value pairs of matched words to obtain clusters of similar words. Once all the clusters are so obtained, every word in each of the resultant cluster is replaced by a single word in the cluster which is English, singular, and has the highest frequency.

DBSCAN to Cluster Similar Words

	RIC	RCE	RICE	ARROZ	ARRO	ARRZ	AROZ	ARROS	BACKP	BACKPA	BACKPAC	BACKPAK	BACKPACKS	BACKPACK	....
RIC	0		1												
RCE		0	1												
RICE	1	1	0	1											
ARROZ			1	0	1	1	1								
ARRO				1	0			1							
ARRZ				1		0									
AROZ				1			0								
ARROS					1			0							
BACKP									0	1					
BACKPA									1	0				1	
BACKPAC											0		1		
BACKPAK												0	1		
BACKPACKS											1	1	0	1	
BACKPACK										1			1	0	1
....														1	0

DBSCAN to Cluster Similar Words



Words in Corpus

Abbreviation, Singularization & Spell Correction Edges

Translation Edges

DBSCAN with MinPts = 2
Epsilon =1

Select the word within the graph cluster, which is English, singular & most frequent word in corpus

Description Correction Examples

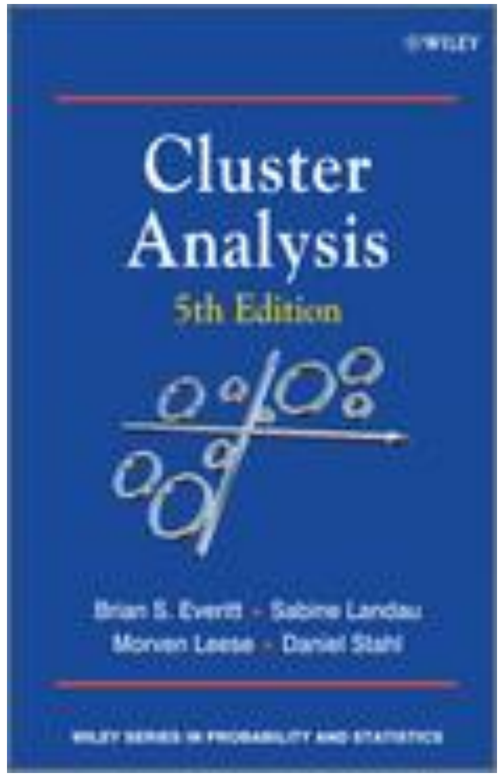
- ➢Using this approach, we have corrected 50,000+ words for 6 Million + items in 10 markets (US, US SAMS, CA, UK, MX, AR, Chile, Central America, China & Japan)
- ➢This solution is deployed to SPInE tool (wmlink/spine) from Feb-2019.

WORD_IN_DESC	CORRECTED_WORD
BACKP	BACKPACK
BACKPA	BACKPACK
BACKPAC	BACKPACK
BACKPACKS	BACKPACK
BACKPAK	BACKPACK
BACKPC	BACKPACK
BACKPK	BACKPACK
BCKPACK	BACKPACK
BCKPCK	BACKPACK
BKP	BACKPACK
BKPACK	BACKPACK
BKPK	BACKPACK
BKPKS	BACKPACK
BPACK	BACKPACK
BPACKS	BACKPACK
MOCHI	BACKPACK
MOCHILA	BACKPACK
MOCHILAS	BACKPACK

WORD_IN_DESC	CORRECTED_WORD
RIC	RICE
RCE	RICE
ARRZ	RICE
ARROZ	RICE
ARROS	RICE
ARRO	RICE
AROZ	RICE

References

O'REILLY
Hands-On Unsupervised Learning Using Python
How to Build Applied Machine Learning Solutions from Unlabeled Data
Ankur A. Patel


WILEY
Cluster Analysis
5th Edition
Brian S. Everitt · Sabine Landau
Morven Leese · Daniel Stahl
WILEY SERIES IN PROBABILITY AND STATISTICS

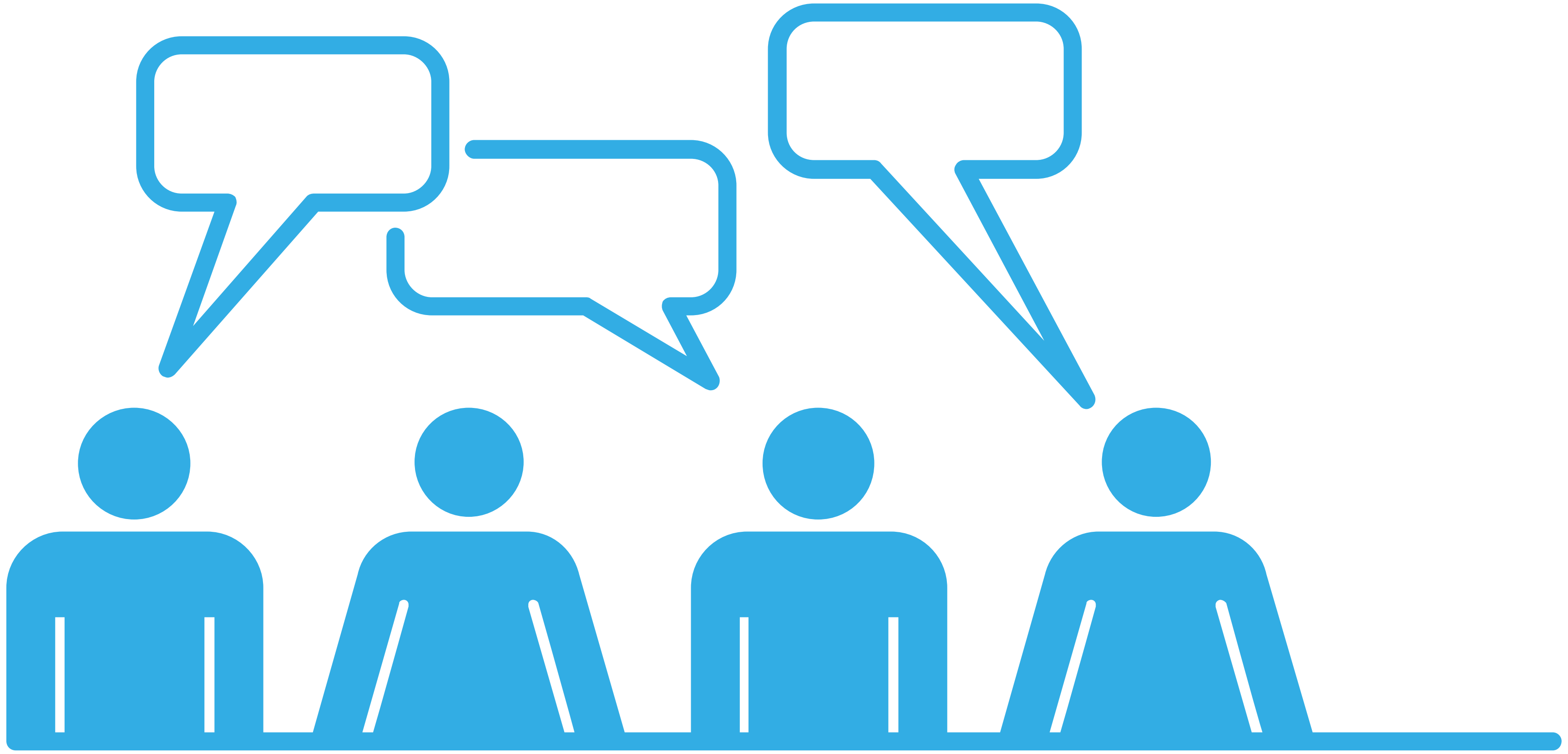
https://learning.oreilly.com/library/view/data-clustering/9781466558229/

https://learning.oreilly.com/library/view/hands-on-unsupervised-learning/9781492035633/

https://learning.oreilly.com/library/view/cluster-analysis-5th/9780470978443/

Text Correction example code: https://gecgithub01.walmart.com/m0m00zs/TechByte_Clustering

Q & A



Thank you !

Contact:
Mallikharjuna.MV@walmart.com
Ojaswini.Chhabra@walmart.com