CLUSTERING ANALYSIS

DATA 3300 – Carly Fox, PhD

OBJECTIVES

Define Cluster Analysis, data type requirements, and business applications

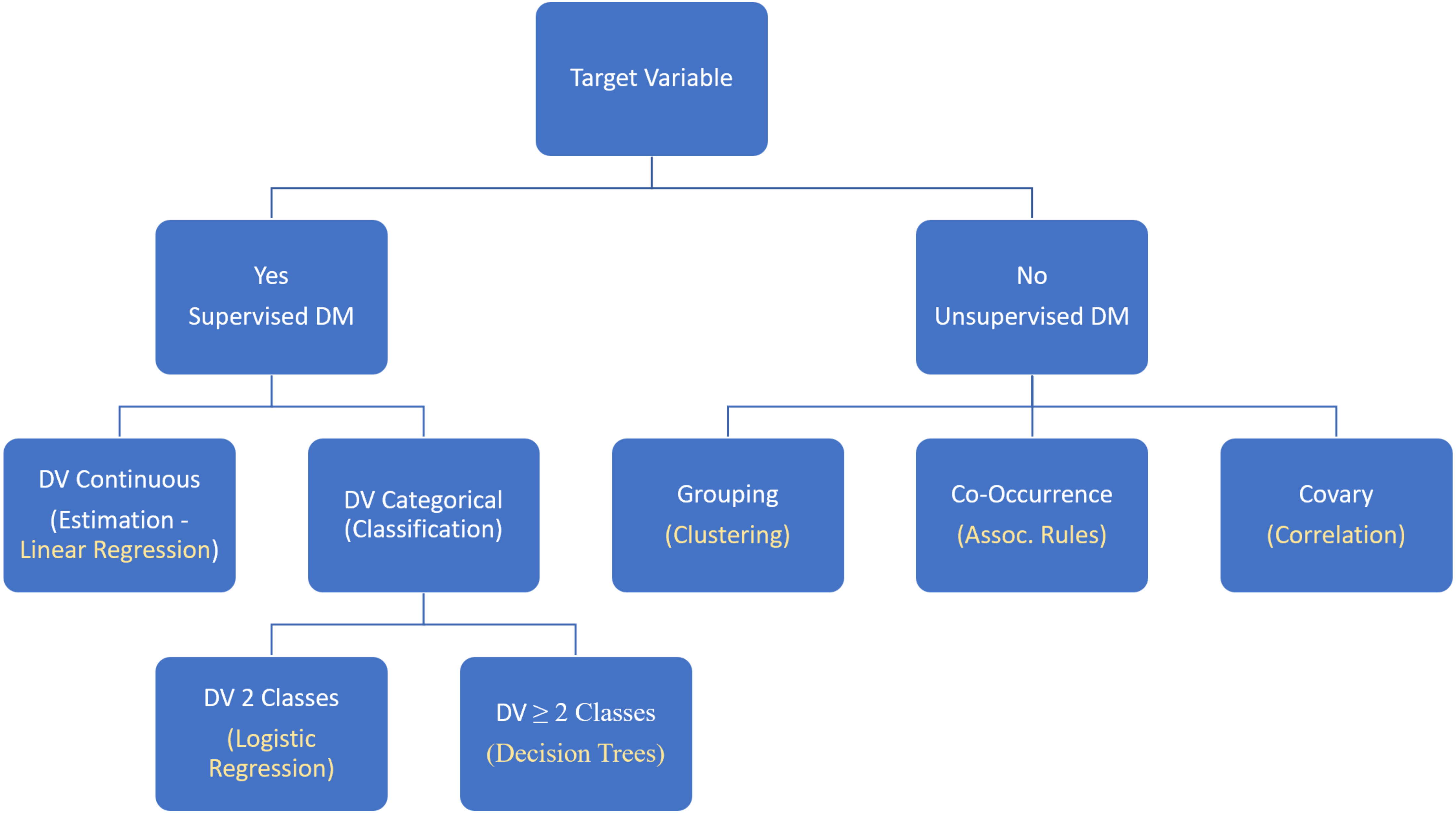
Understand what *k*-means clustering is and what *k* stands for

Calculate similarity/dissimilarity between observations (Euclidean Distance)

Describe the basic process of the *k*-means algorithm, including centroids

Understand the potential effect of outliers

Examine the limitations of Cluster Analysis



Target Variable

Yes

Supervised DM

No

Unsupervised DM

DV Continuous

(Estimation - Linear Regression)

DV Categorical (Classification)

Grouping

(Clustering)

Co-Occurrence

(Assoc. Rules)

Covary

(Correlation)

DV 2 Classes

(Logistic Regression)

DV ≥ 2 Classes

(Decision Trees)

CLUSTERING ANALYSIS OVERVIEW

Type of Analysis: Unsupervised; looking for natural relationships, not trying to predict a target variable

Type of Data: Quantitative (interval/ratio) and or qualitative (ordinal/nominal) – with additional preprocessing – may be used

Type of Business Qs: Do cases (e.g., customers, employees, etc.) tend to cluster into natural groups that we can use for an actionable purpose?

- Do certain groups of customers tend to display similar purchasing patterns?

- Are there certain clients who have a higher risk profile than others?

APPLICATIONS OF CLUSTERING ANALYSIS

Market/Customer Segmentation: Grouping people according to their similarity across several dimensions (attributes) related to a product under consideration

Sales Segmentation: Clustering types of customers by which products purchased

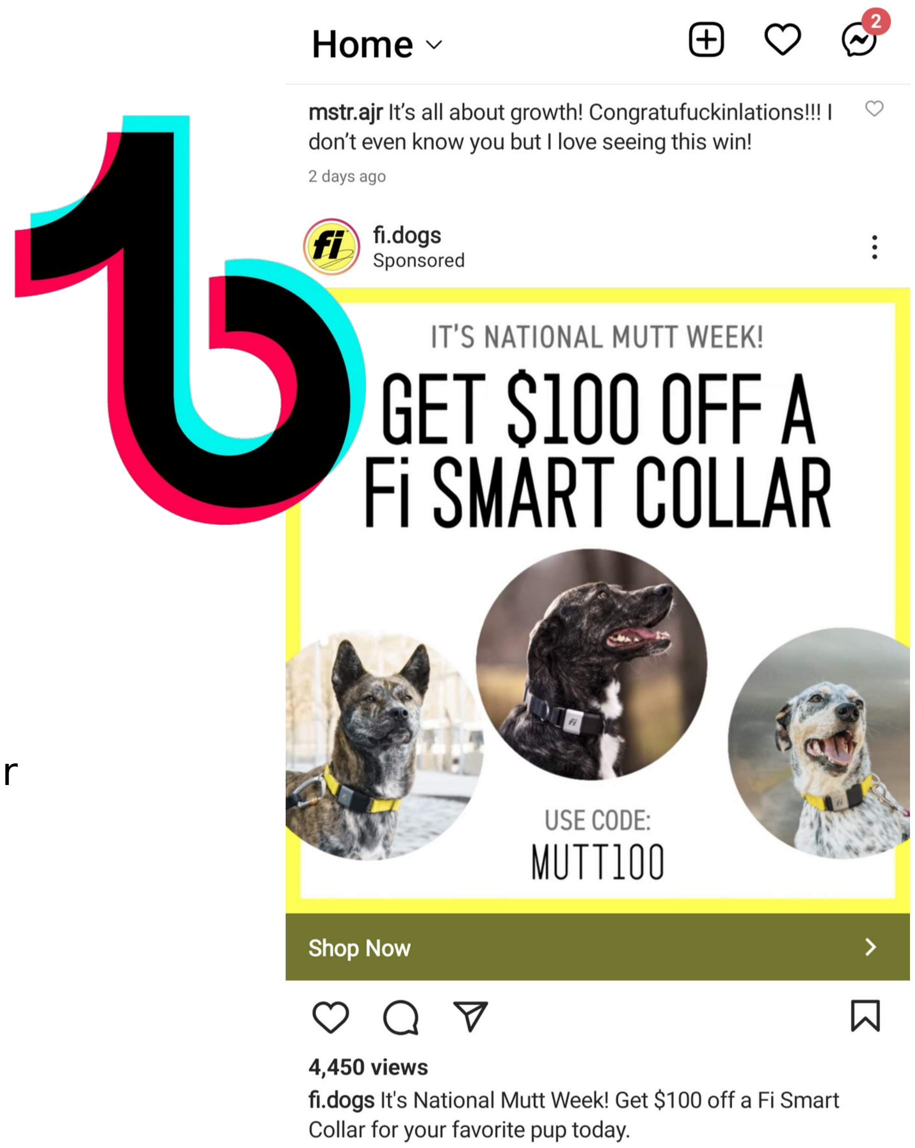
Credit Risk: Clustering types of customers based on their credit history

Operations: Promoting based on a person’s performance or segmenting high performers

Insurance: IDing groups of motor insurance policy holders with a high average claim cost

City-Planning: IDing groups of houses according to their house type, value, and geographical location

Geographical: IDing areas of similar land use



mstr.ajr It’s all about growth! Congratufuckinlations!!! I don’t even know you but I love seeing this win!

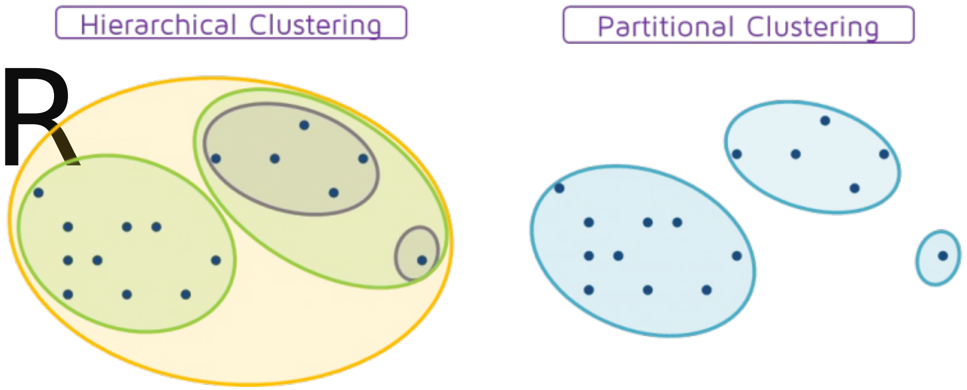
2 days ago

fi.dogs
Sponsored

4,450 views

fi.dogs It's National Mutt Week! Get $100 off a Fi Smart Collar for your favorite pup today.

TYPES OF CLUSTER ANALYSIS



Partitional (Non-Hierarchical)

A division of objects (data instances) into non-overlapping subsets (clusters) such that each object belongs to exactly one cluster

Divide the dataset of size *N* objects into *M* clusters

K-Means Clustering ☾ most used non-hierarchical method in business analytics

What we’re focusing on

Hierarchical

A set of nested clusters organized as a hierarchical tree

Produces a set of nested clusters in which each pair of objects or clusters is progressing nested in a larger cluster until only one remains

EX: doctors, nested within hospitals, nested within states, nested within the US

CHAID tree most used in business analytics

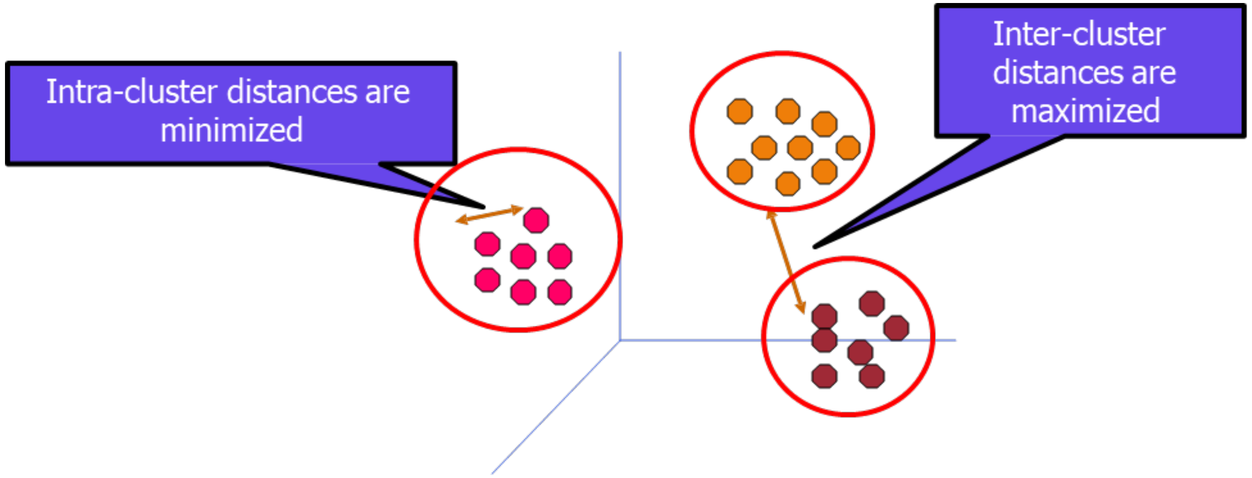
IMPORTANT TERMS

Cluster ☾ group of objects (cases, points, observations, members, customers, etc.) – *not attributes* – that are similar to each other with respect to variable(s) of interest

Clustering Analysis ☾ Data mining method used to find naturally occurring groups of cases/observations/objects in the sample that are:

- Homogenous within groups (i.e., high intra-class similarity)

- Heterogenous between each group (i.e., low inter-class similarity)



Intra-cluster distances are minimized

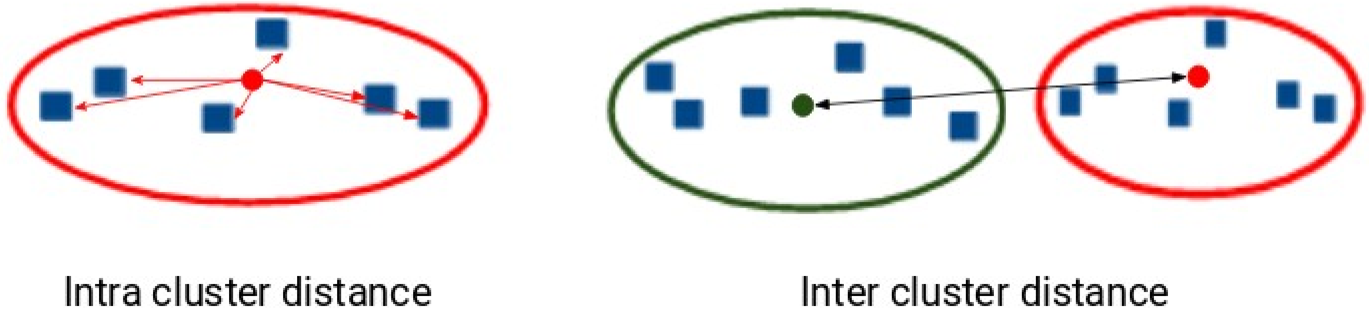
Inter-cluster distances are maximized

CONSIDERATIONS IN CLUSTER ANALYSIS

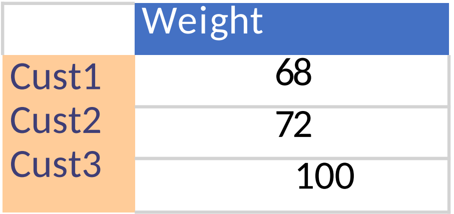
If the aim is to build clusters (e.g., divide a sample or population into groups of similar objects), how do we do that?

- What defines **similarity/dissimilarity**
- How do you define **distance** between clusters?

Remember: trying to max similarity (min distance) within clusters and max dissimilarity (max distance) between clusters

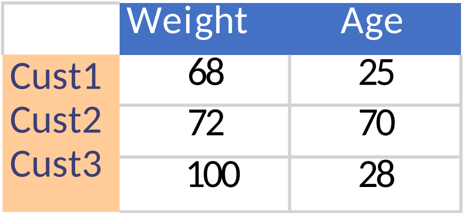


SIMILARITY & DISSIMILARITY



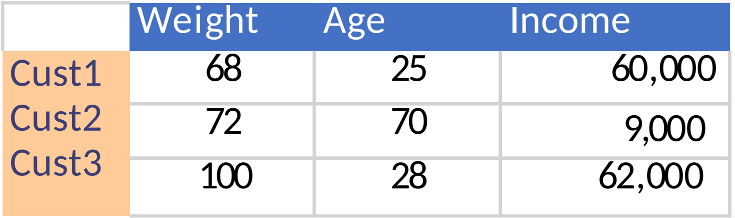
	Weight
Cust1	68
Cust2	72
Cust3	100

Which two customers are similar?



	Weight	Age
Cust1	68	25
Cust2	72	70
Cust3	100	28

Which two customers are similar now?



	Weight	Age	Income
Cust1	68	25	60,000
Cust2	72	70	9,000
Cust3	100	28	62,000

Which two customers are similar in this case?

QUANTIFYING SIMILARITY – MEASURES OF DISTANCE

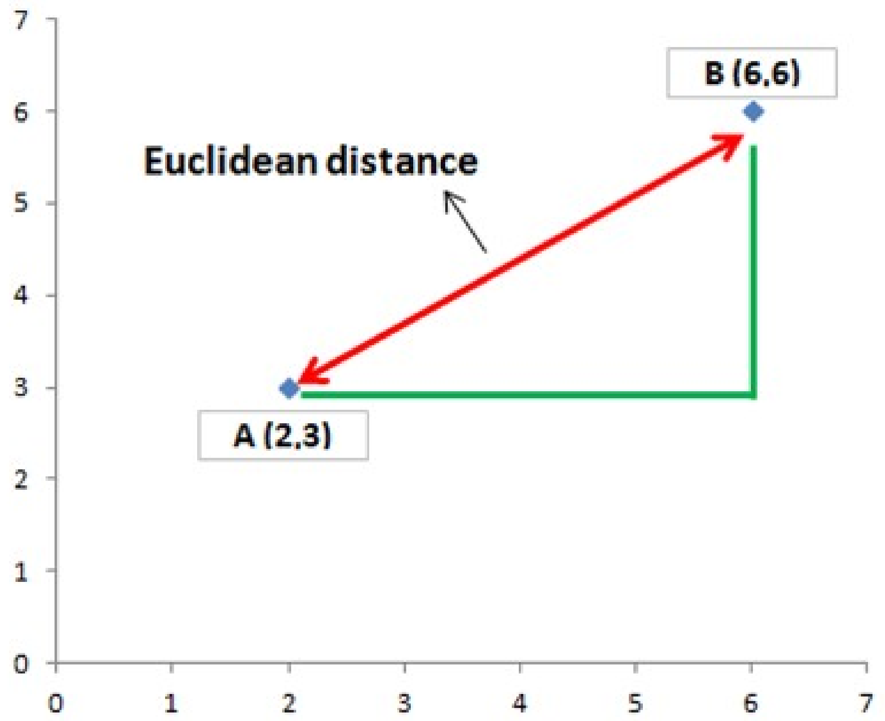
The **similarity of two observations** can be quantified by calculating their ***distance*** from one another

Straightforward when there’s on a single variable

Multiple variables require an ***aggregate distance measure*** like ***Euclidean Distance*** (e.g., remember the cartesian coordinate system?)

$$distance = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

FEELING NOSTALGIC ABOUT GEOMETRY?



$$\boldsymbol{Euclidean\ distance}\ (\boldsymbol{a},\boldsymbol{b}) = \sqrt{(\boldsymbol{a_1}-\boldsymbol{b_1})^2 + (\boldsymbol{a_2}-\boldsymbol{b_2})^2}$$

Other types of distance metrics

CALCULATING DISTANCE

	Weight
Cust1	68
Cust2	72
Cust3	100

- Cust1 vs Cust2 : (68-72)= **4**
- Cust2 vs Cust3 : (72-100) = **28**
- Cust3 vs Cust1 : (100-68) =**32**

	Weight	Age
Cust1	68	25
Cust2	72	70
Cust3	100	28

- Cust1 vs Cust2 : sqrt($((68-72)^2 + (25-70)^2)$ = **44.9**
- Cust2 vs Cust3 : **50.54**
- Cust3 vs Cust1 : **32.14**

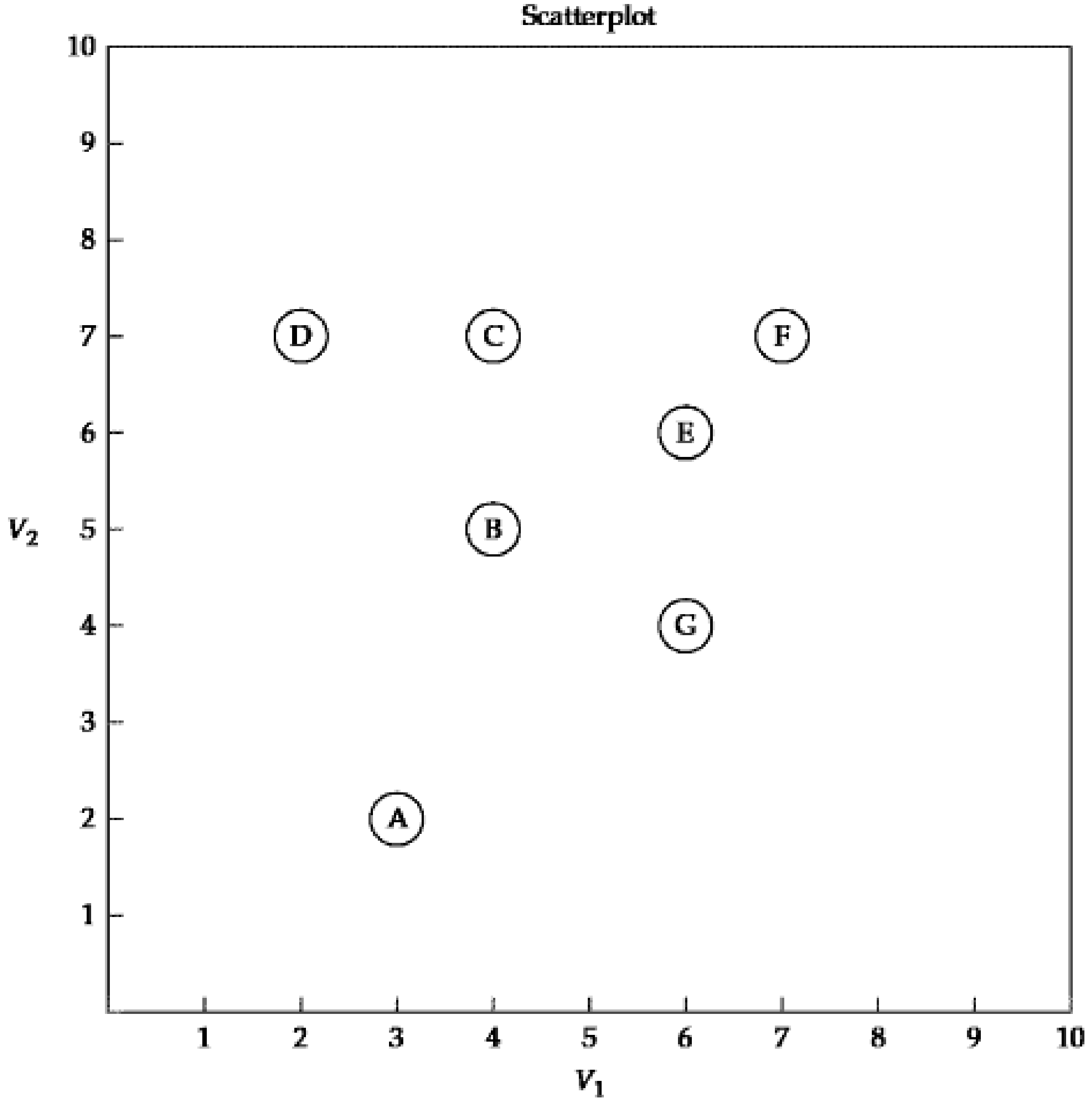
$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(x_{ki} - x_{kj}\right)^2}$$

EXAMPLE CASE

Simple Example: Suppose a marketing researcher wishes to determine market segments in a community based on patterns of loyalty to brands and stores. A small sample of seven respondents is selected as a pilot test of how cluster analysis is applied. Two measures of loyalty - V1(store loyalty) and V2(brand loyalty) - were measured for each respondent on a 0-10 scale.

Clustering Variable	**Respondents**						
	A	**B**	**C**	**D**	**E**	**F**	**G**
Store Loyalty	3	4	4	3	6	7	6
Brand Loyalty	2	4	7	7	6	7	4

CALCULATING DISTANCE



CALCULATING DISTANCE

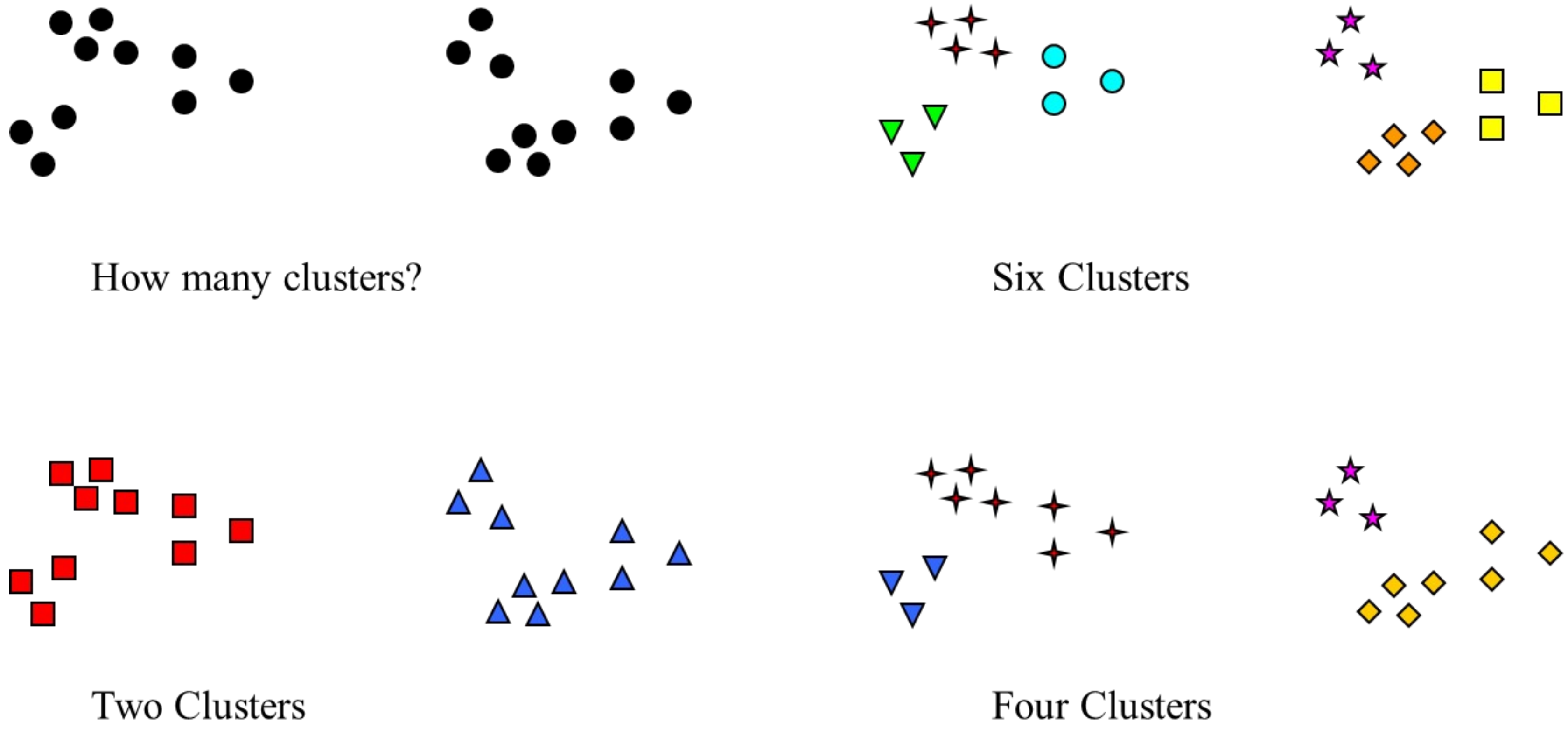
roximity Matric of Euclidean Distance Between Each Observation Set

Observations	Observations						
	A	**B**	**C**	**D**	**E**	**F**	**G**
A	-						
B	3.162	-					
C	5.099	2.000	-				
D	5.099	2.828	2.000	-			
E	5.000	2.236	2.236	4.123	-		
F	6.403	3.606	3.000	5.000	1.414	-	
G	3.606	2.236	3.606	5.000	2.000	3.162	-

$$d_{Euclidean}\,(A,B) = \sqrt{(V_{1(A)} - V_{1(B)}\,)^2 + \;(V_{2(A)} - V_{2(B)}\,)^2}$$

$$d_{Euclidean}\,(A,B) = \sqrt{(3-4\,)^2 + \;(2-5\,)^2} = 3.162$$

CLUSTERING IS **NOT** AN **EXACT SCIENCE**



How many clusters?

Six Clusters

Two Clusters

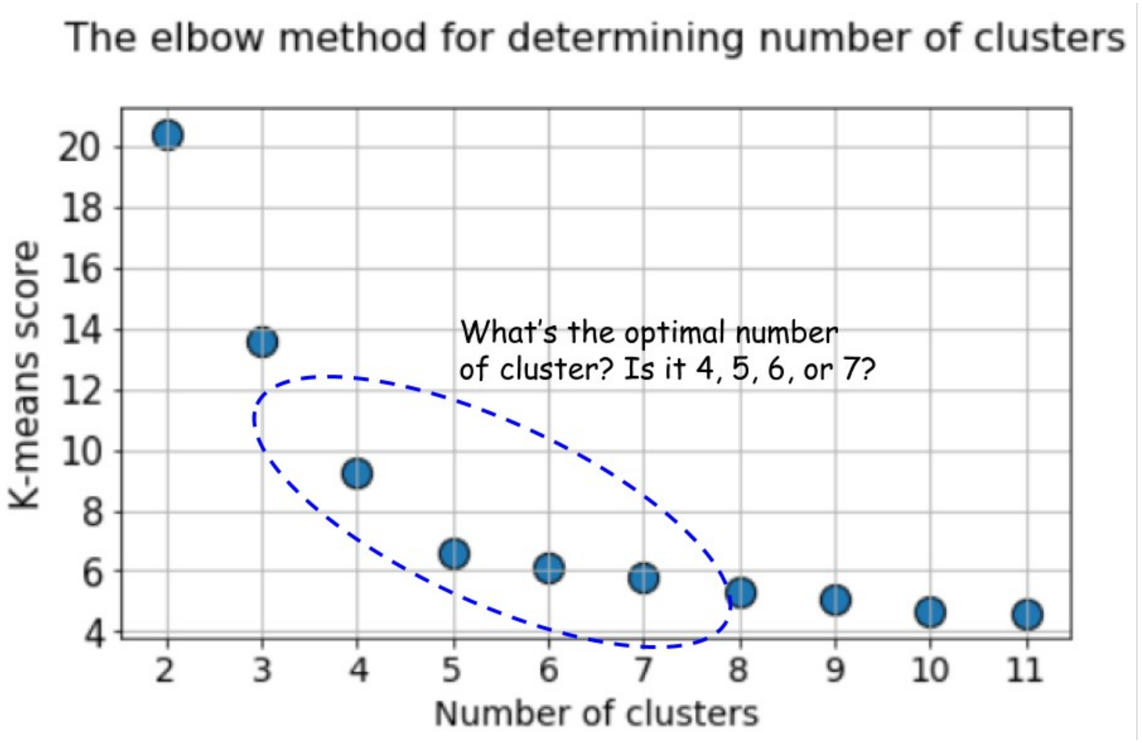
Four Clusters

HOW DO YOU DETERMINE CLUSTER SIZE ()?

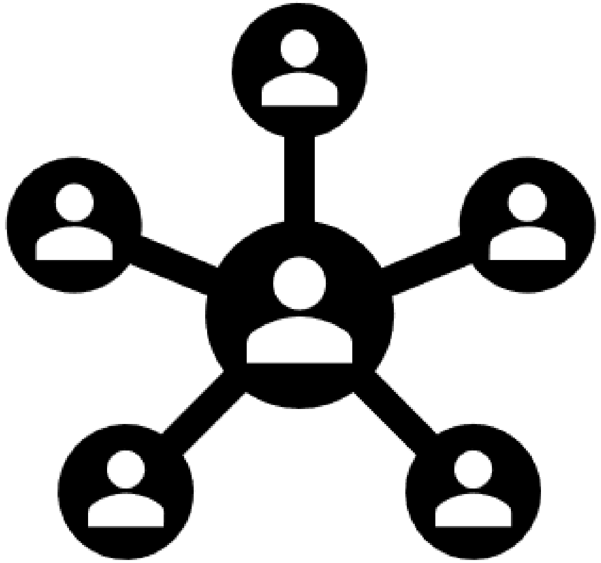
Tractability: The client identifies the number of clusters, typically based on the ease with which something can be managed

Post-hoc evaluation: Practice of performing several analyses using different values for *k*, then reviewing the results to determine which *k* value is most suitable

Elbow rule: Identify the value of *k* after which increasing the value of *k* adds very little to the clusters in terms of further decreasing the within-cluster distance or increasing the between-cluster distance



CLUSTERING ANALYSIS P2



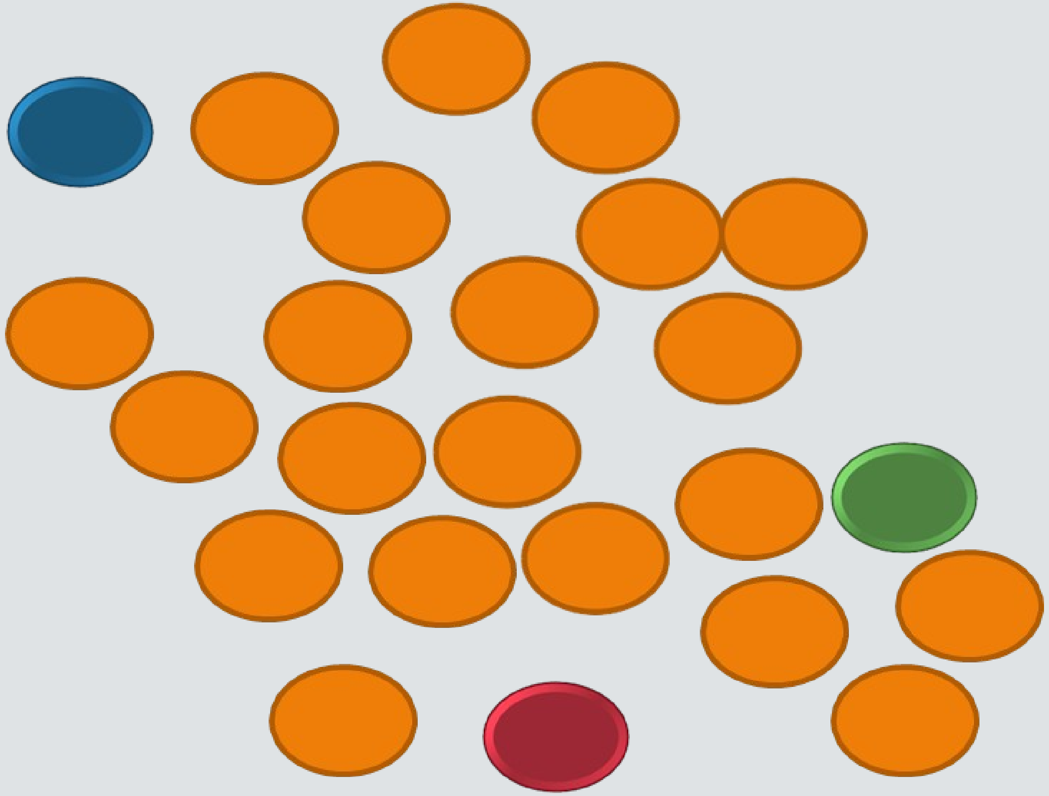
- MEANS CLUSTERING – ALGORITHM STEPS

1. Number of clusters is set by the analyst
2. An initial set of “seeds” (aggregation centroids) is provided
    - Starts with the first *k* elements
    - Other seeds (randomly selected or explicitly defined)
3. Given a certain fixed threshold value, all units are assigned to the nearest cluster’s seed (centroid)
4. New seeds are computed
5. Repeat steps 3-4 until no reclassification of observations is necessary

k-Means Clustering

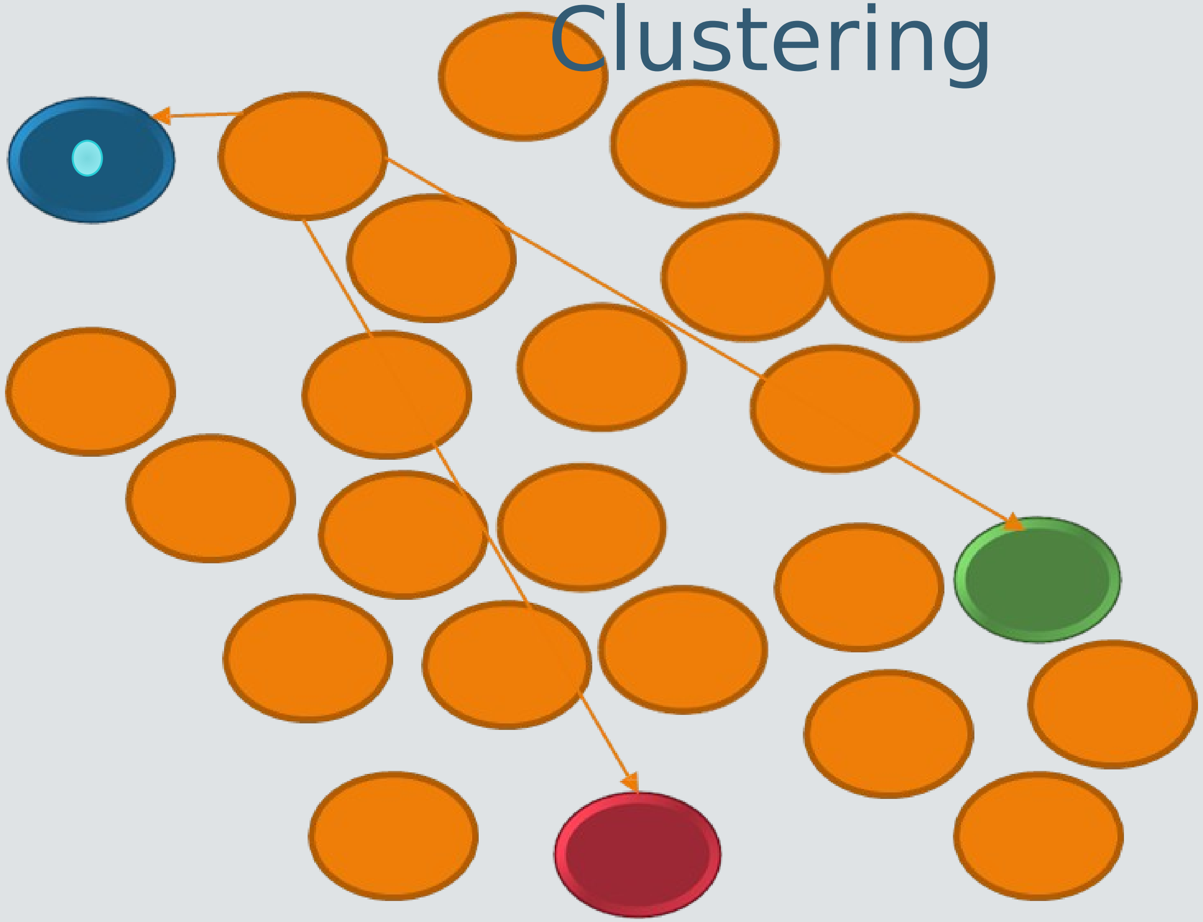
Overall sample

k-Means Clustering



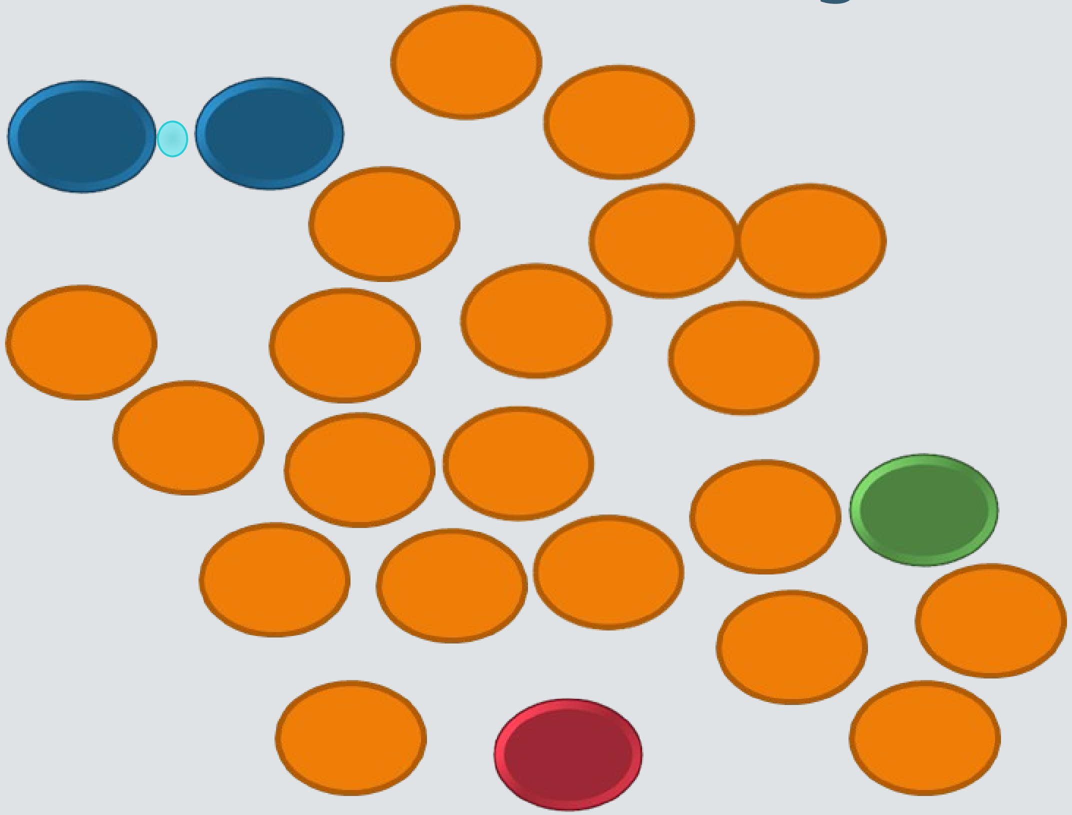
Fix the Number of Cluster seeds or centroids

k-Means Clustering



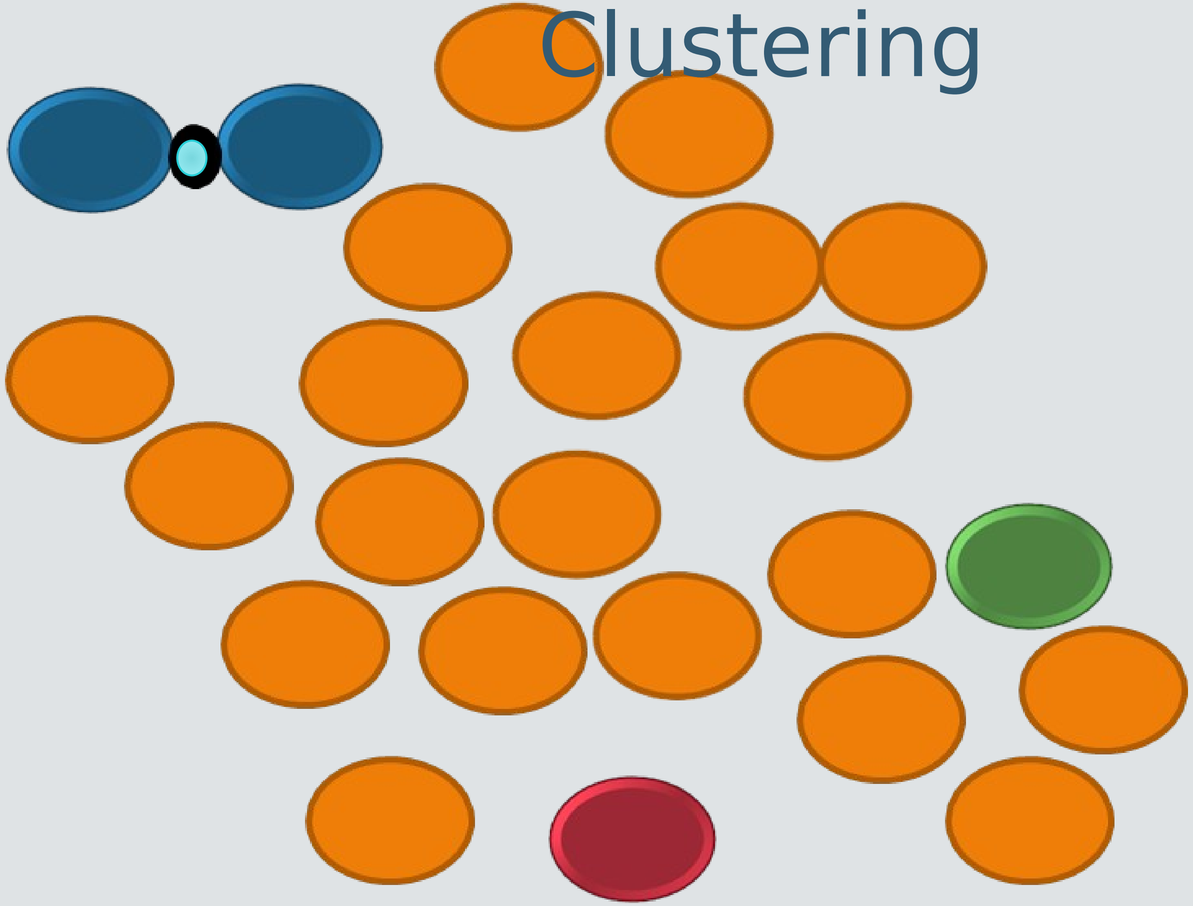
Calculate the distance of each case from all centroids

k-Means Clustering



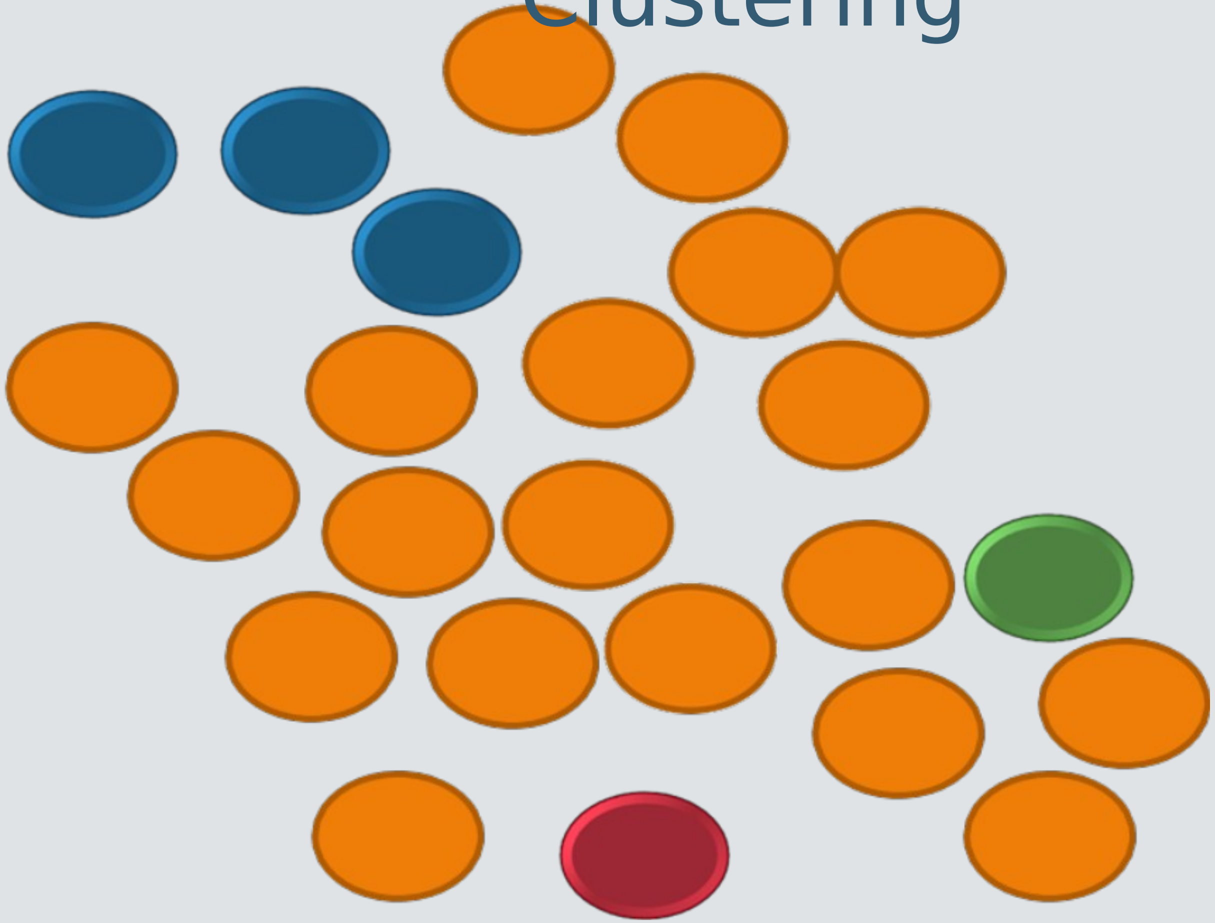
Assign each case to nearest cluster centroid

k-Means Clustering



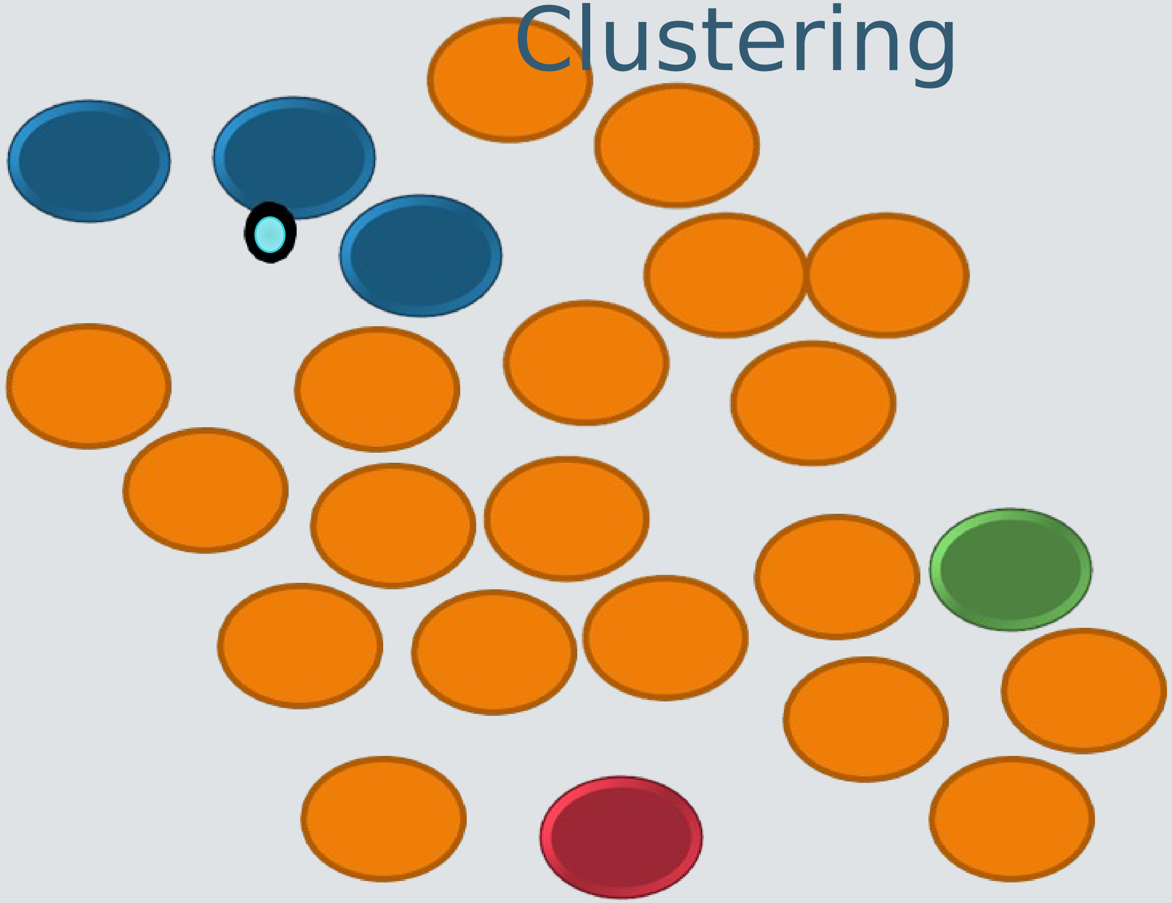
Recalculate the cluster centroids

k-Means Clustering

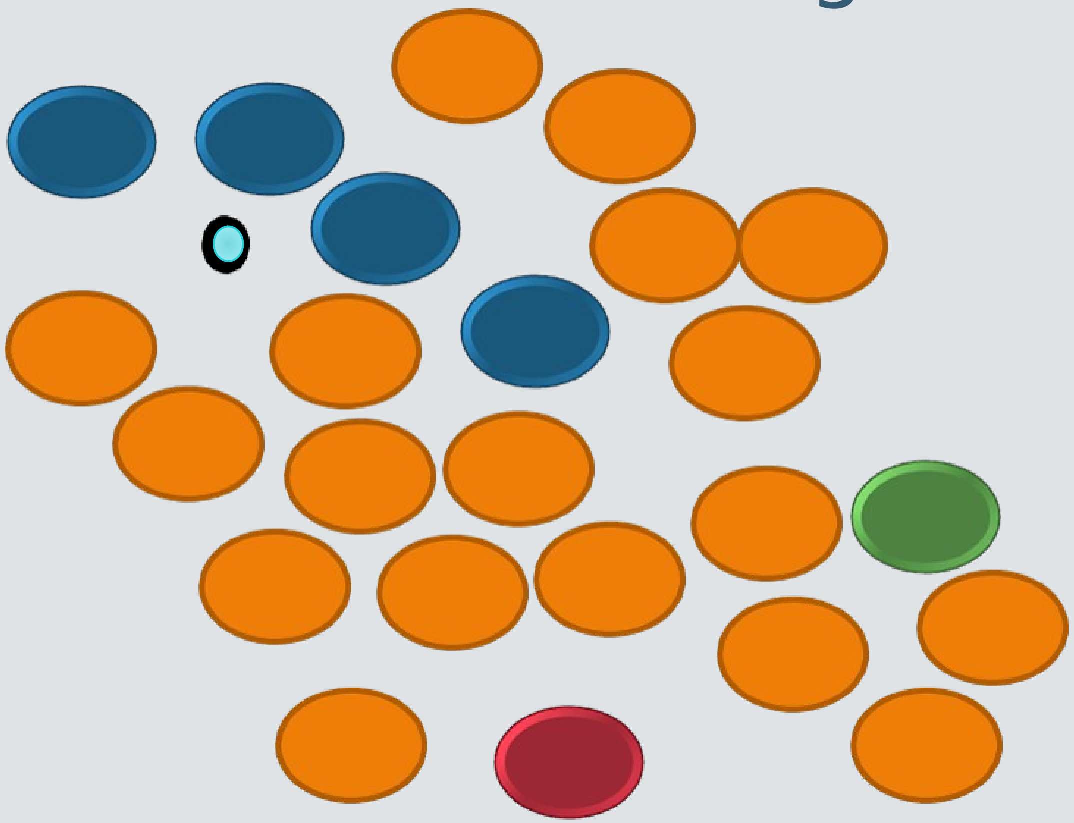


Repeat...

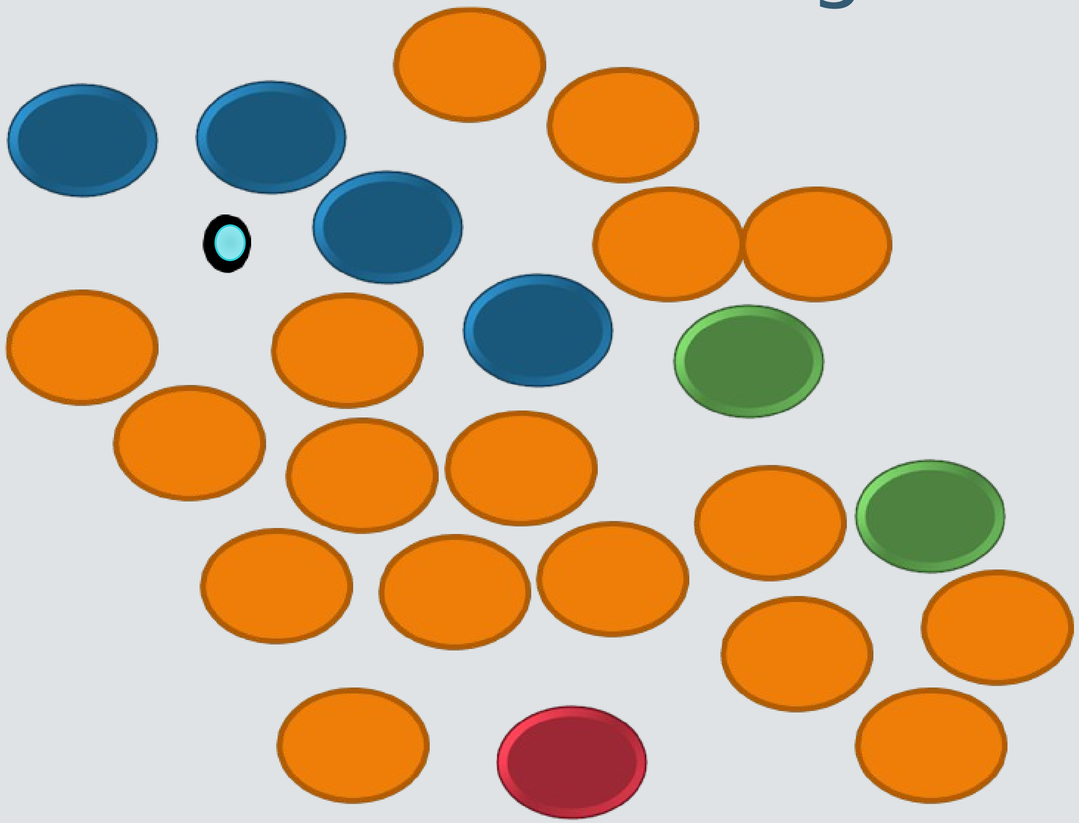
k-Means Clustering



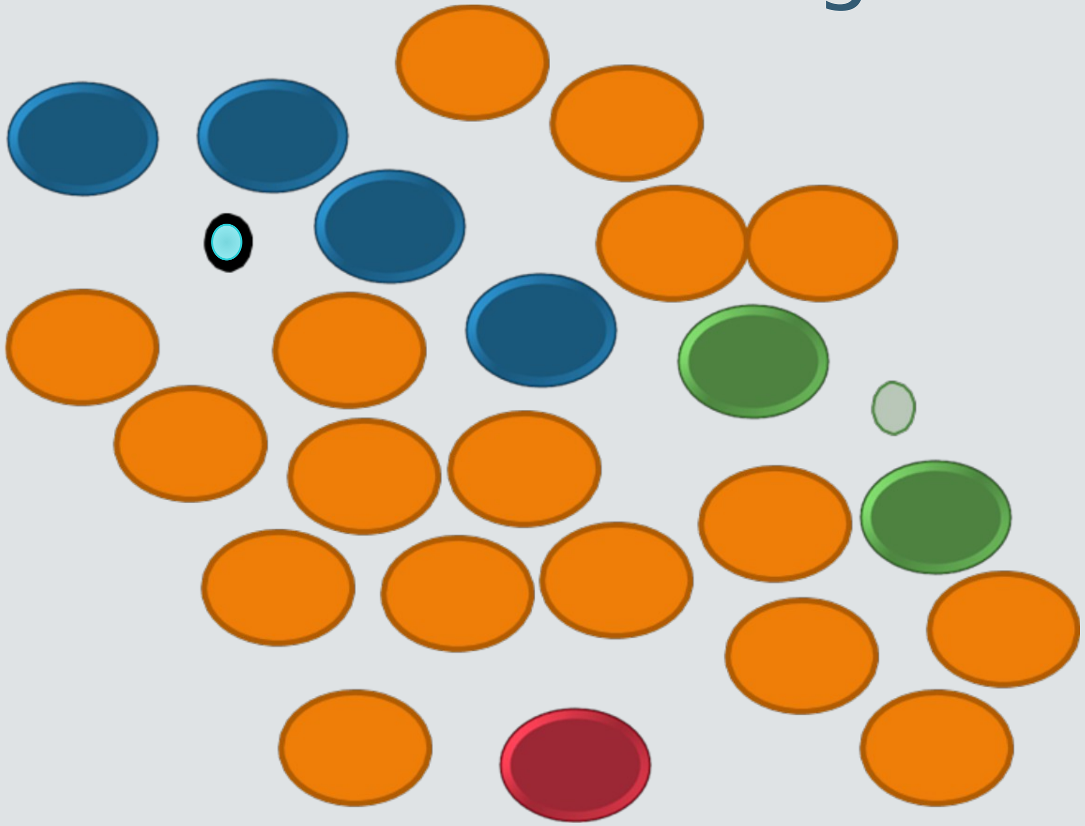
k-Means Clustering



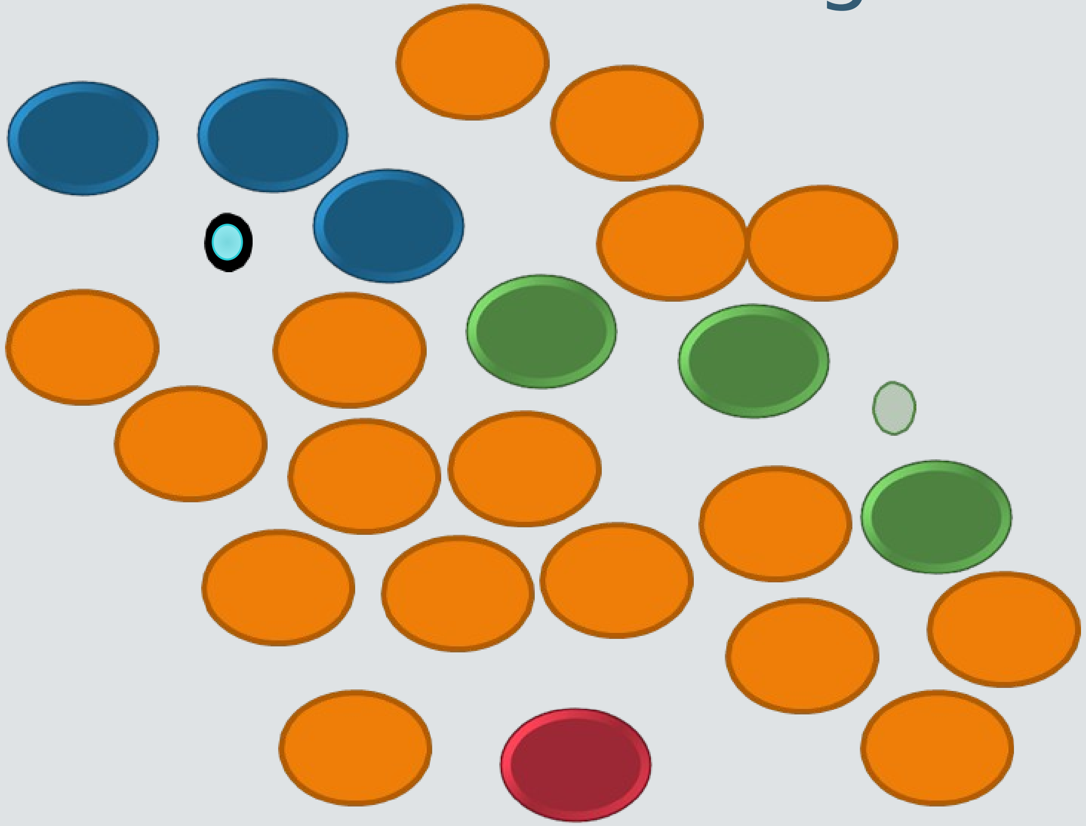
k-Means Clustering



k-Means Clustering

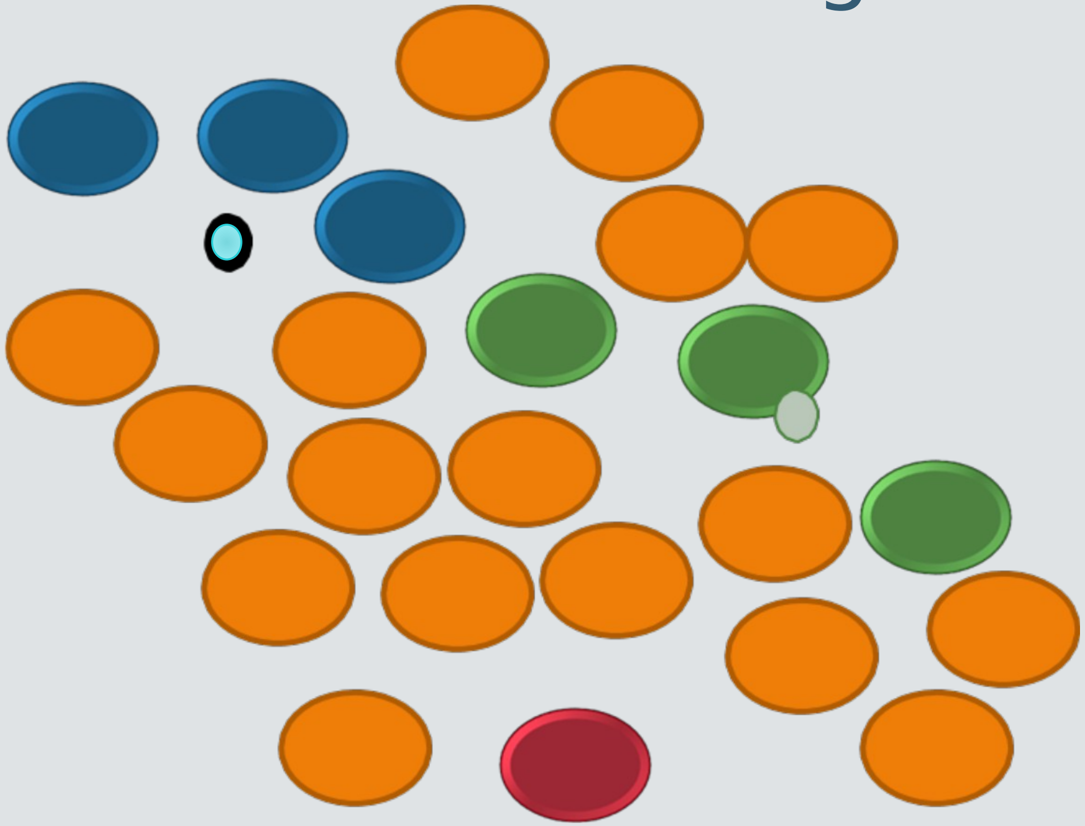


k-Means Clustering



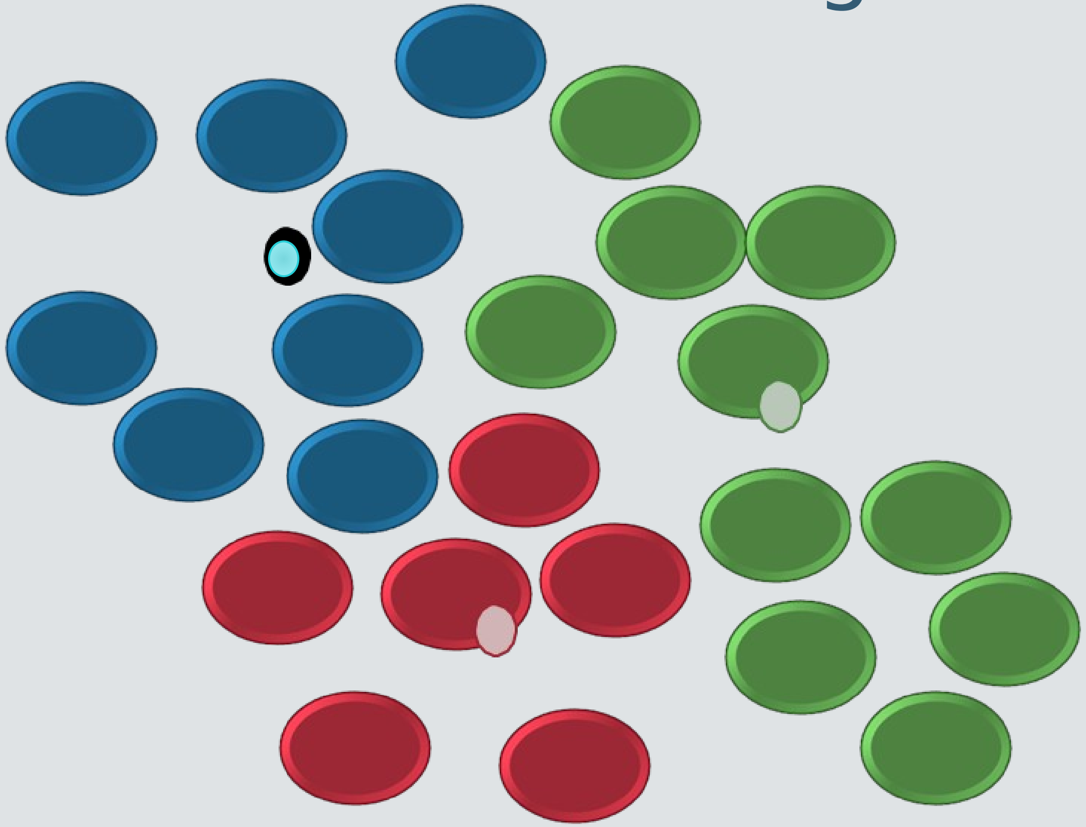
Reassign cases as needed after changing cluster centroids

k-Means Clustering



Re-compute cluster centroids as new cases are assigned

k-Means Clustering



Continue until there is no significant change between each iteration

THE BANE OF OUR (ANALYST) EXISTENCE: OUTLIERS 😠

Outliers can severely distort the representativeness of the results of cluster analysis

They **<u>should be removed</u> IF** the outliers represent:

Aberrant observations not representative of the population

Observations of small or insignificant segments within the population and of no interest to the analysis objectives

They **<u>should be NOT be removed</u> IF:**

There is undersampling/poor representation of relevant groups in the population

The sample should be augmented to ensure representation of these groups

How to detect outliers?

Their appearance in cluster solutions as single-member or small clusters

Using the 68-95-99 rule

Seeing if a value is two or more standard deviations from the mean

ISSUES OF SCALE

Some analyses (e.g., k-means) work better when vars have similar ranges/are on similar scales to prevent overweighting certain variables

This can be accomplished via:

Normalization ☾ Every var is fit into the same range (e.g., between 0-1) or on the same scale

Standardization ☾ Every value is calculated in terms of SD from the mean**;** typically, though a *Z-score conversion, sets the mean to 0 with an SD of 1*

$$z = \frac{x - \mu}{\sigma} \qquad \begin{array}{l} \mu = \text{Mean} \\ \sigma = \text{Standard Deviation} \end{array}$$

Read more here about differences between Normalization & Standardization

INTERPRETATION OF CLUSTERS

Centroid Table

Variable	Mean	SD	Cluster_1	Cluster_2	Cluster_3	Cluster_4
Gender = F	0.48	0.50	1.00	0.54	0.37	0.00
Age_C	47.26	12.194	43.04	63.36	43.66	44.02
Dog_wt	67.49	26.107	57.67	72.82	90.73	57.31
Visits	8.45	5.934	8.09	13.80	4.99	7.96

hat can be interpreted about cluster 1? What about clusters 2,3,4?

e these differences substantial, or do they fail to show much variation?

the cluster centroids fit in with prior expectations?

VALIDATION & LIMITATIONS OF CLUSTERS

Important to **validate** your cluster analysis because its descriptive in nature and requires additional support

Limitations:

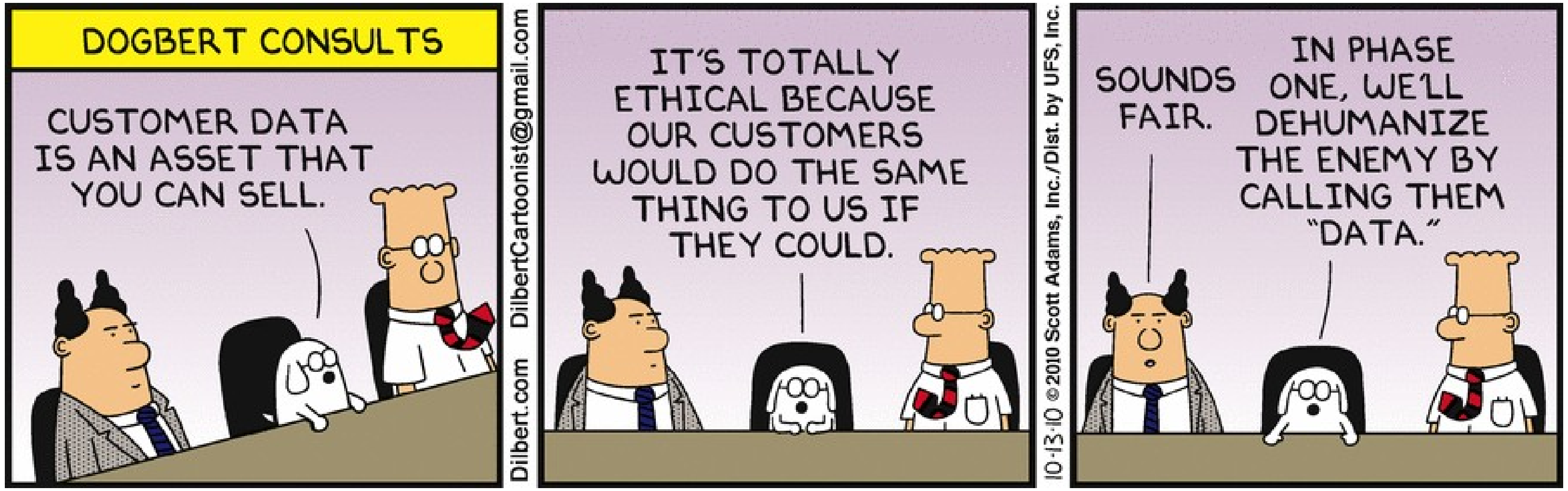
- No statistical basis upon which to draw inferences from your sample to the population

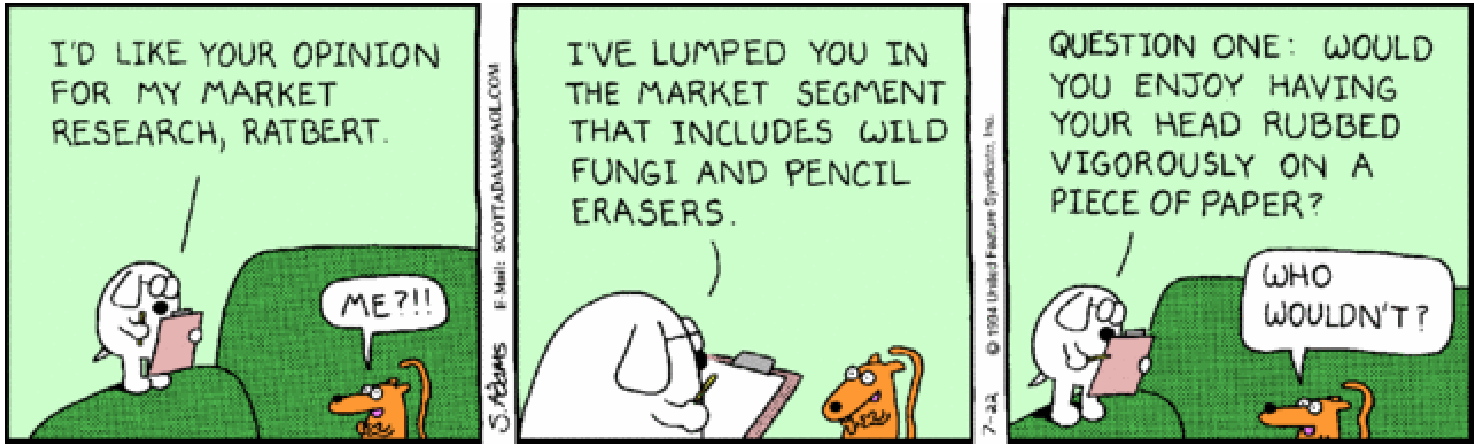
- Will always create *k* clusters, whether they exist or not

- Cluster solution is not generalizable to the data outside of the sample used to develop the cluster solution

- Humans like to ID patterns where they don’t necessarily exist, and beware of cognitive bias or bias in the data set

WEB COMICS (MAYBE) FOR THOUGHT


DOGBERT CONSULTS
CUSTOMER DATA IS AN ASSET THAT YOU CAN SELL.
Dilbert.com DilbertCartoonist@gmail.com
IT'S TOTALLY ETHICAL BECAUSE OUR CUSTOMERS WOULD DO THE SAME THING TO US IF THEY COULD.
10-13-10 © 2010 Scott Adams, Inc./Dist. by UFS, Inc.
SOUNDS FAIR.
IN PHASE ONE, WE'LL DEHUMANIZE THE ENEMY BY CALLING THEM "DATA."


I'D LIKE YOUR OPINION FOR MY MARKET RESEARCH, RATBERT.
ME?!!
S. Adams E-Mail: SCOTTADAMS@AOL.COM
I'VE LUMPED YOU IN THE MARKET SEGMENT THAT INCLUDES WILD FUNGI AND PENCIL ERASERS.
7-22 © 1994 United Feature Syndicate, Inc.
QUESTION ONE: WOULD YOU ENJOY HAVING YOUR HEAD RUBBED VIGOROUSLY ON A PIECE OF PAPER?
WHO WOULDN'T?

Moment of Chill