Class 20: Clustering

MGSC 310

Prof. Jonathan Hersh

Class 20 Announcements

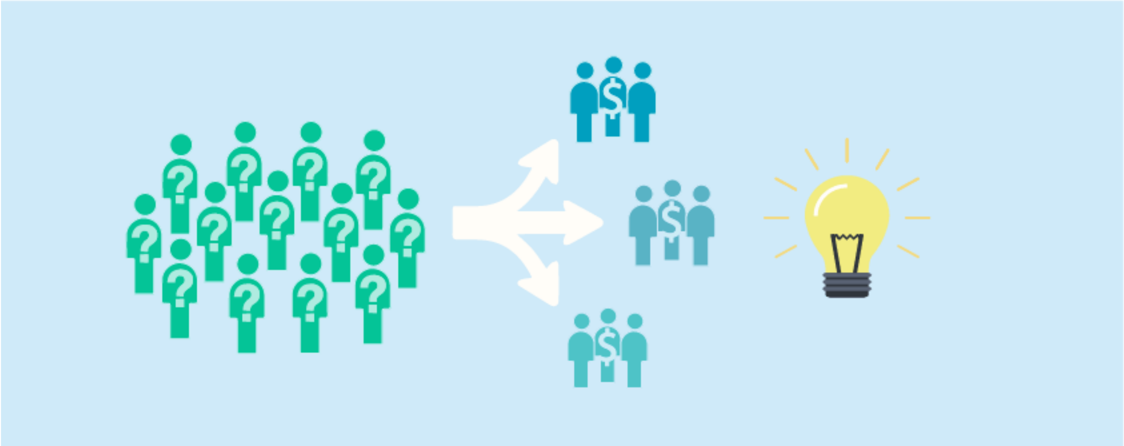
1. No Quiz This Week
2. Pset 5 due Tuesday, November 17 (posted)
   - Problem set 6 canceled ☹
3. Final Project
   - Feedback sent via Canvas. Most look great!
   - Upload model by ~~Nov 19~~ Dec 1
   - Signup for timeslot for final project presentation:
     - https://tinyurl.com/310FinalProject230
     - https://tinyurl.com/310FinalProject400

Class 20: Outline

1. K-Mean Clustering
2. Hierarchical Clustering
3. Clustering in R
4. Lab (Time Permitting)

What is unsupervised learning?

- All of the machine learning we've encountered so far has been supervised learning such as regression
- This lecture will describe **unsupervised learning**
- In unsupervised learning, we observe x_1, x_2, x_p, features but we don't observe any Ys



Goals of unsupervised learning

- **Since we don't observe $Y's$, we can't predict anything**

- The goal is more subtle here: can we discover interesting patterns in the data? Can we discover useful subgroups?

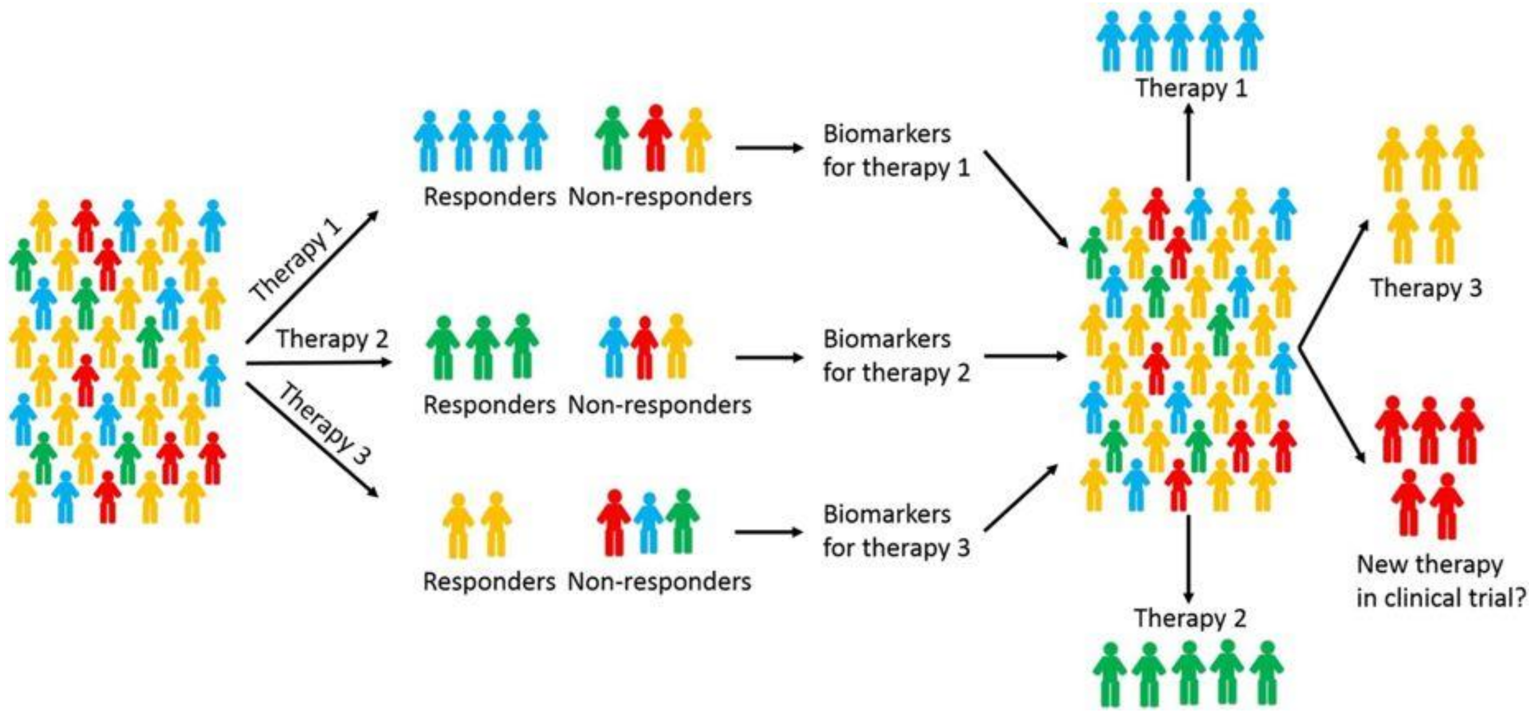
Challenge of unsupervised learning

- **Because we have no "truth", the end result needs interpretation**

- We often have to bring our own contextual understanding to a fitted unsupervised model

- Some examples of unsupervised learning….

Personalized Cancer Treatment by Genetic Characteristics



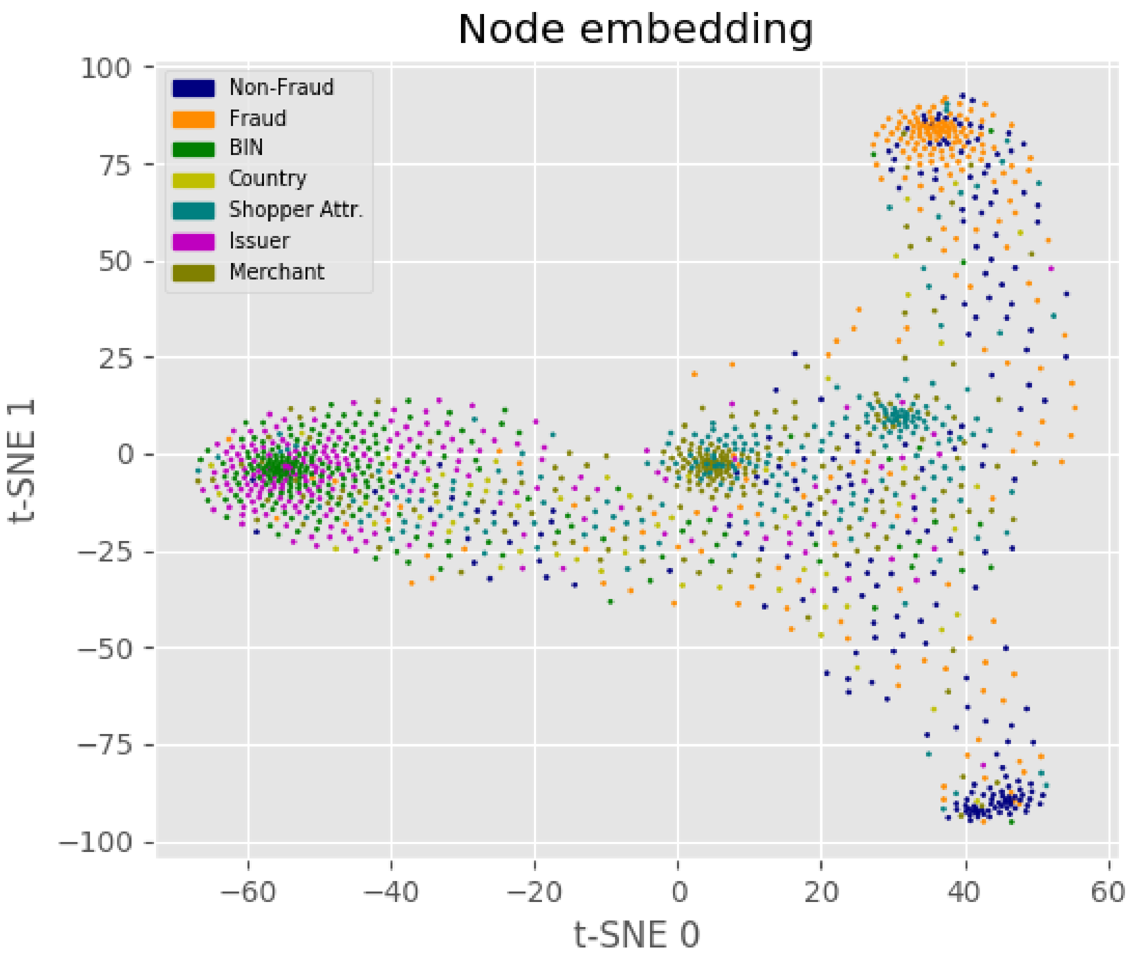
- Xs: patient genetic expressions
  - Very high dimensional! 1000000s of genes for each patient
- Method: group patients by genes
- Goal: better cancer therapy targeting

Group Website Visitors By Site Behavior

Events Greater than Average	All Clusters 144,198 Users Avg # Events	Cluster 2 45,530 Users Avg # Events	Cluster 1 48,245 Users Avg # Events	Cluster 3 29,739 Users Avg # Events	Cluster 4 20,684 Users Avg # Events
1 Search Song or Video	5.28	9.14 +1.0 σ	4.04 -0.3 σ	0.99 -1.1 σ	5.85 +0.2 σ
2 Select Song or Video	5.31	9.19 +1.0 σ	3.92 -0.4 σ	0.97 -1.1 σ	6.28 +0.3 σ
3 Share Song or Video	1.53	3.37 +1.0 σ	1.11 -0.2 σ	0.17 -0.7 σ	0.44 -0.6 σ
4 Concert Landing Screen	1.15	2.53 +0.9 σ	0.80 -0.2 σ	0.20 -0.7 σ	0.31 -0.6 σ
5 Purchase Ticket	0.93	2.1 +0.9 σ	0.61 -0.3 σ	0.13 -0.6 σ	0.25 -0.5 σ
6 Download Song or Video	2.06	3.49 +0.8 σ	2.34 +0.2 σ	0.67 -0.8 σ	0.26 -1.0 σ
7 Add Content to Cart	1.66	2.89 +0.8 σ	1.89 +0.1 σ	0.45 -0.8 σ	0.17 -1.0 σ
8 Purchase Song or Video	1.34	2.4 +0.8 σ	1.5 +0.1 σ	0.30 -0.8 σ	0.13 -0.9 σ
9 Play Song or Video	3.99	6.08	2.51	0.71	7.53

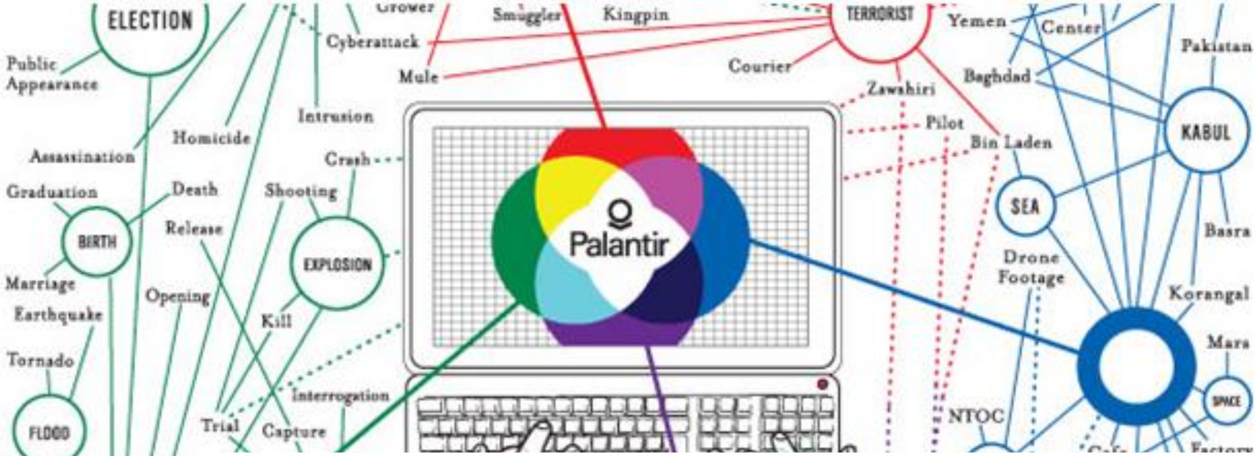
- Xs: visitor website behavior (search, purchase, length of time, add item to cart)
- Method: group visitors by site behavior
- Goal: better understanding of different types of visitor behavior

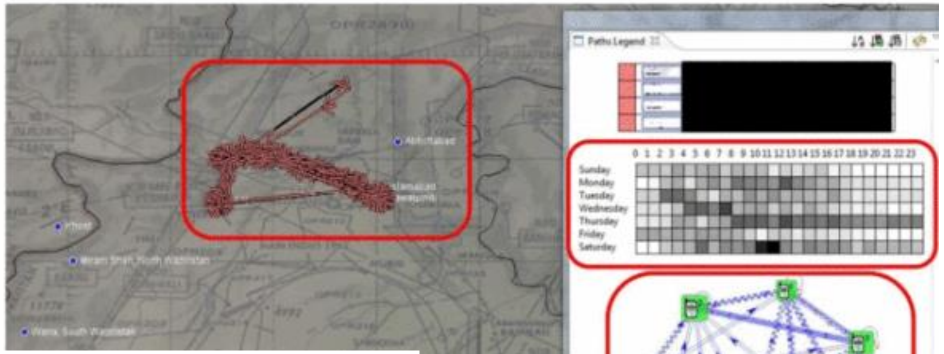
Group Credit Card Behavior Into Possible Fraudulent Behavior



- Xs: credit card purchase history, type of purchase, amount, frequency
- Method: group credit card behavior into fraud and non-fraud groups
- Goal: early warning indicator of financial fraud

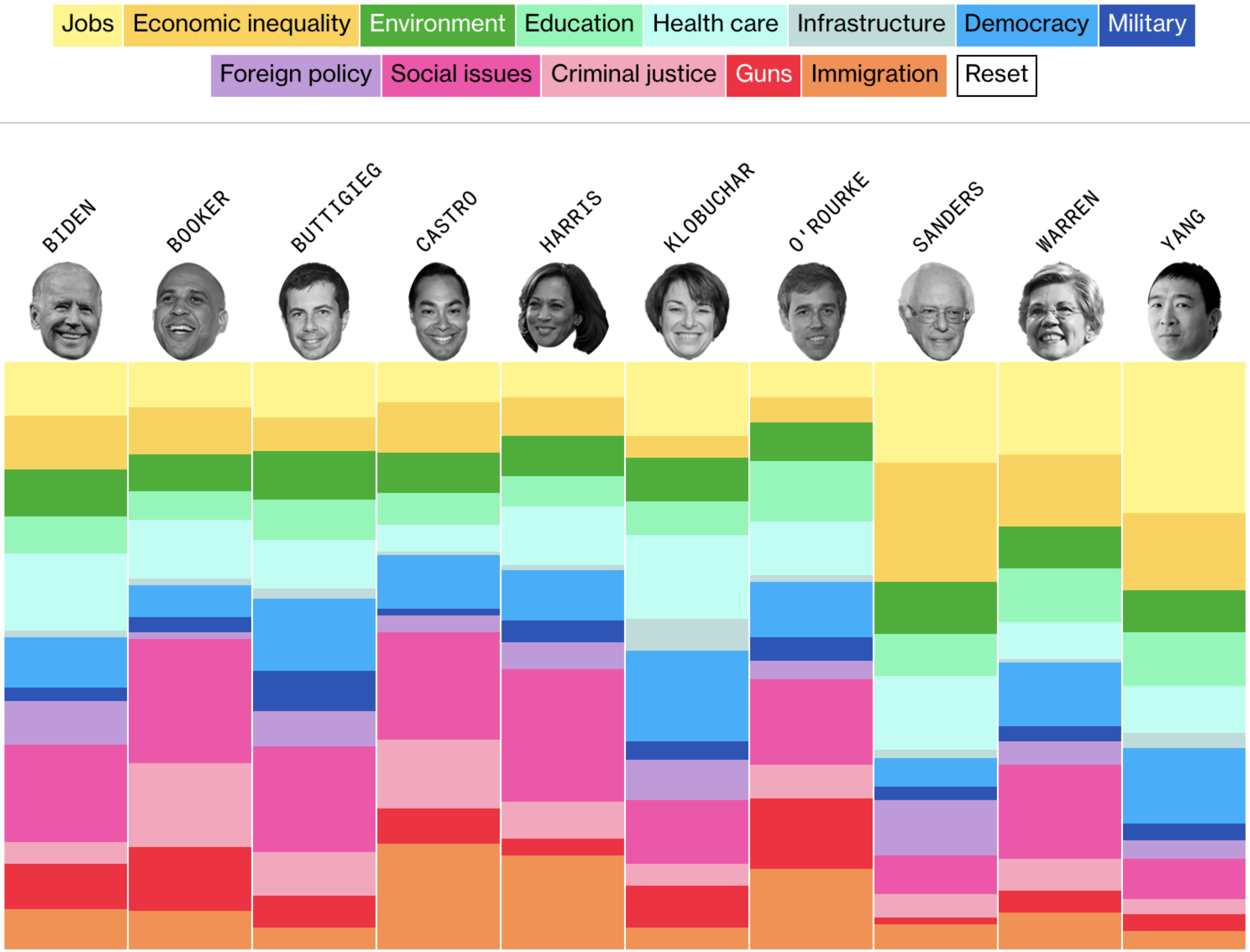
Did Unsupervised Learning Catch Bin Laden?





In-Q-Tel has invested in more than 100 startups, but only one has been associated with the 2011 killing of Osama bin Laden. For years, tech and security insiders have theorized that Palantir's analytics helped the government pick up bin Laden's scent. It's a link Palantir has never confirmed or denied. "I can't comment on our specific national security successes," Karp says when asked by *Fortune*. "Maybe a different way of answering is that not everybody likes our affiliation with national security, but we're very proud of it ... That also involves finding terrorists and sometimes taking them out."

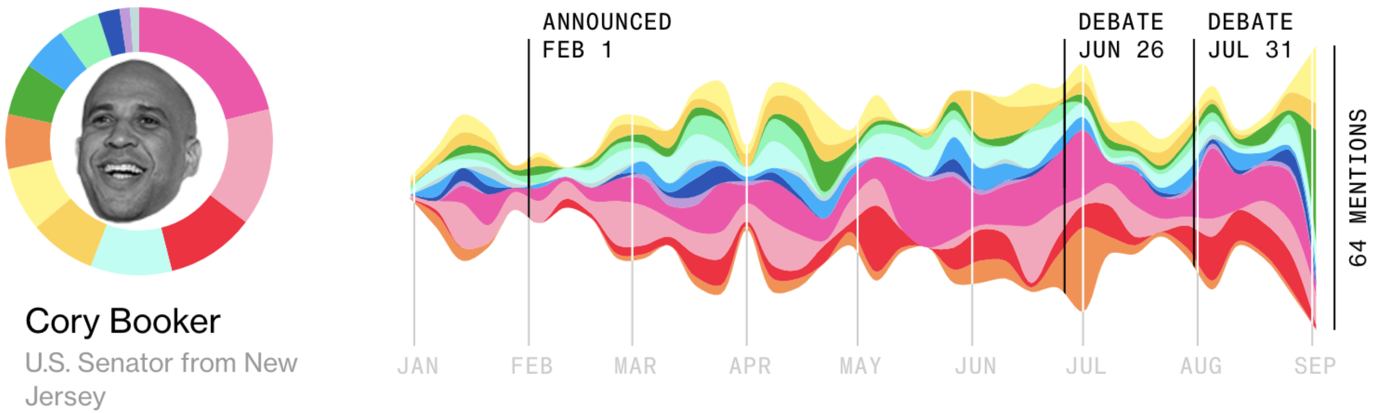
Characterizing Democratic Candidates' Tweets by Topics



- Xs: text of all 44,000 candidates tweets
- Method: group tweets by topic – jobs, inequality, health care, etc
- Goal: model of tweet "topic"

https://www.bloomberg.com/graphics/2020-democratic-presidential-candidate-policies/

Characterizing Democratic Candidates' Tweets by Topics



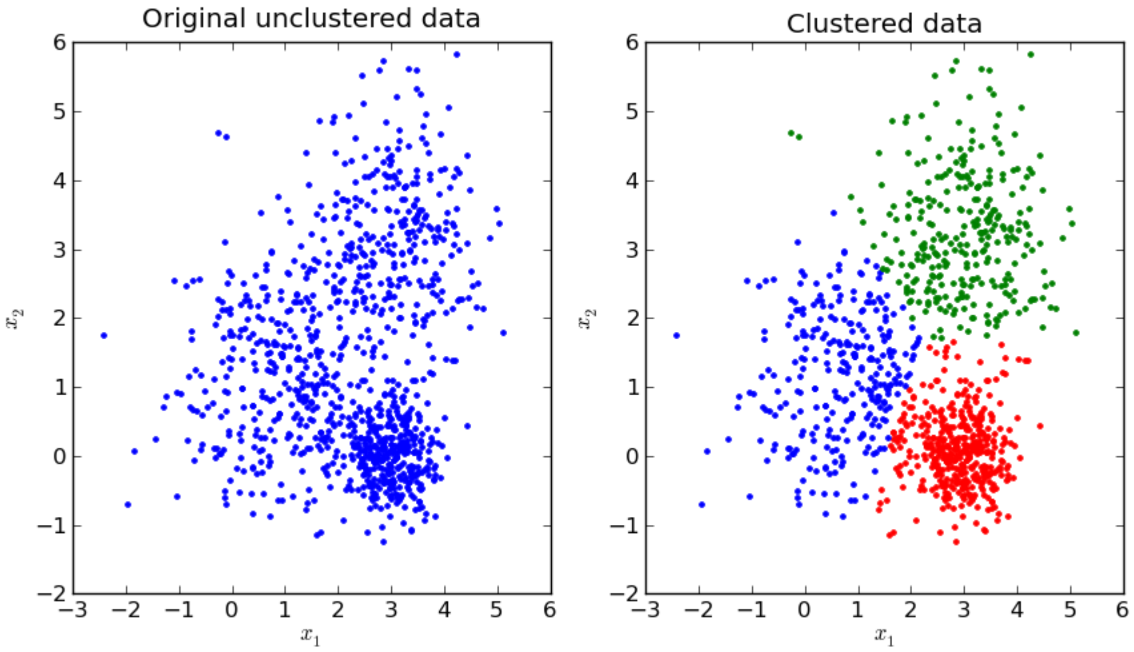
- Xs: text of all 44,000 candidates tweets
- Method: group tweets by topic – jobs, inequality, health care, etc
- Goal: model of tweet "topic"

https://www.bloomberg.com/graphics/2020-democratic-presidential-candidate-policies/

K-means clustering as a "game"

- Tell the computer how many groups (k) you think the data should be split into

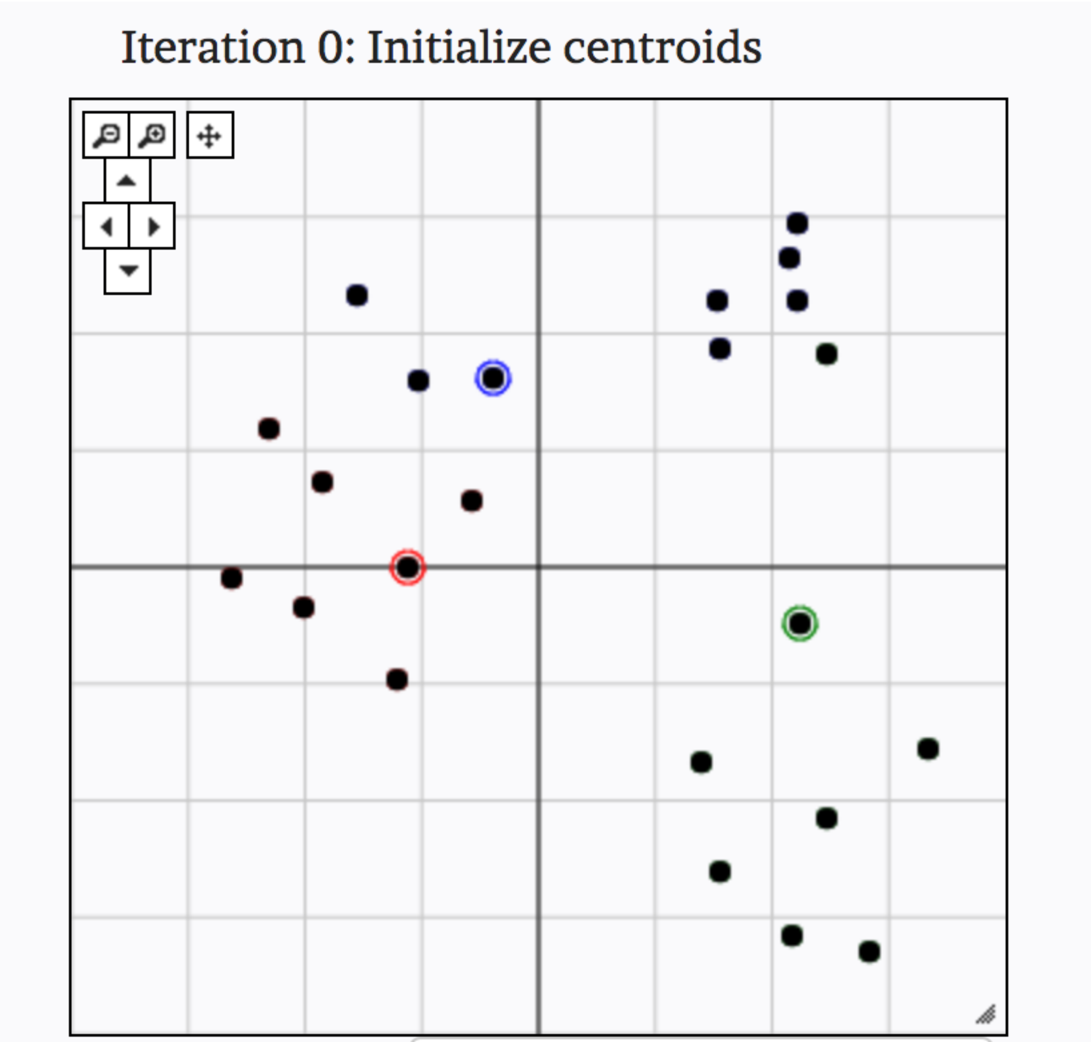
- The computer splits the objects into k groups such that the groups are most similar



K-means clustering algorithm

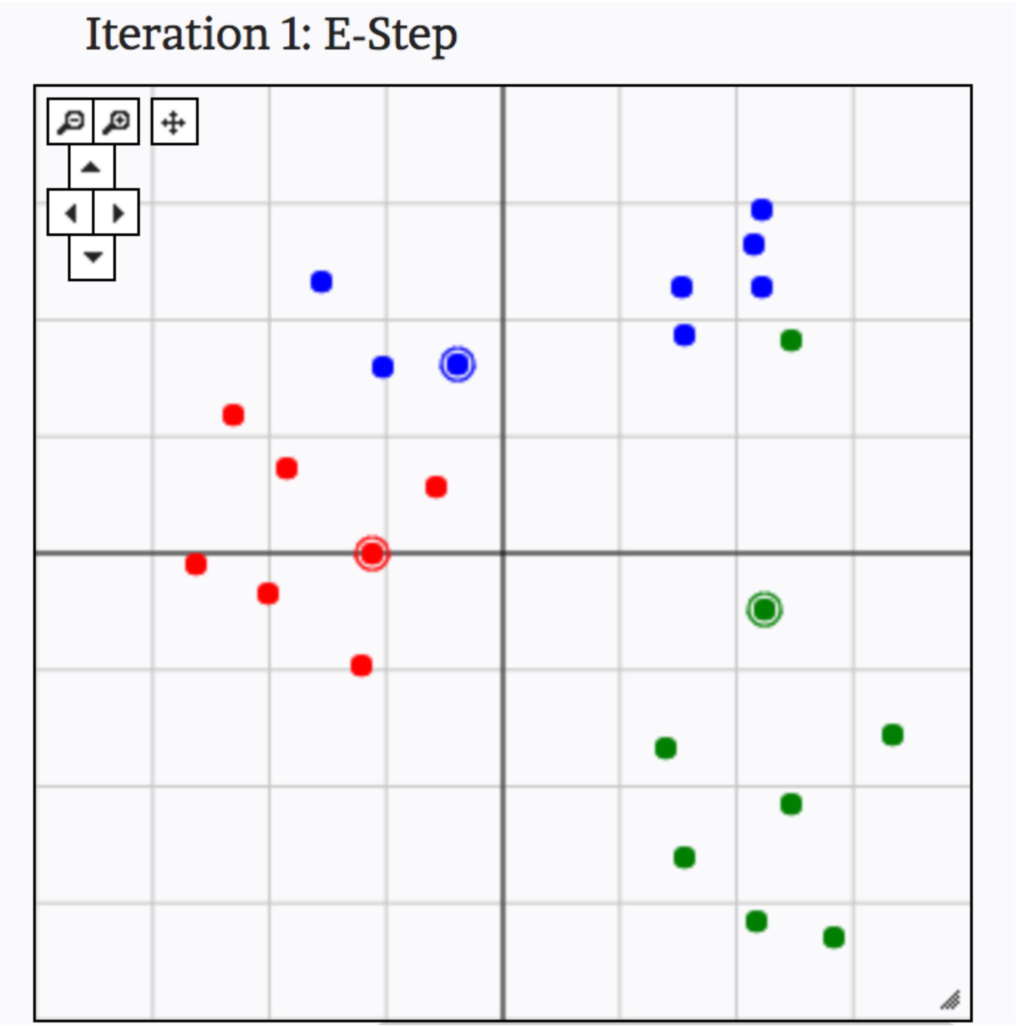
1. Decide how many clusters we want. Call this K
2. Randomly assign a number, $1, \ldots, K$, to each of the observations. (Initial cluster assignment)
3. Iterate until clusters stop changing:
   - **Expectation-step:** For each of the K clusters, compute the cluster centroid (center point, i.e. means for the k-th cluster)
   - **Maximization-step:** Assign each observation to the cluster whose centroid is closest (in Euclidean distance)

K-means clustering in action



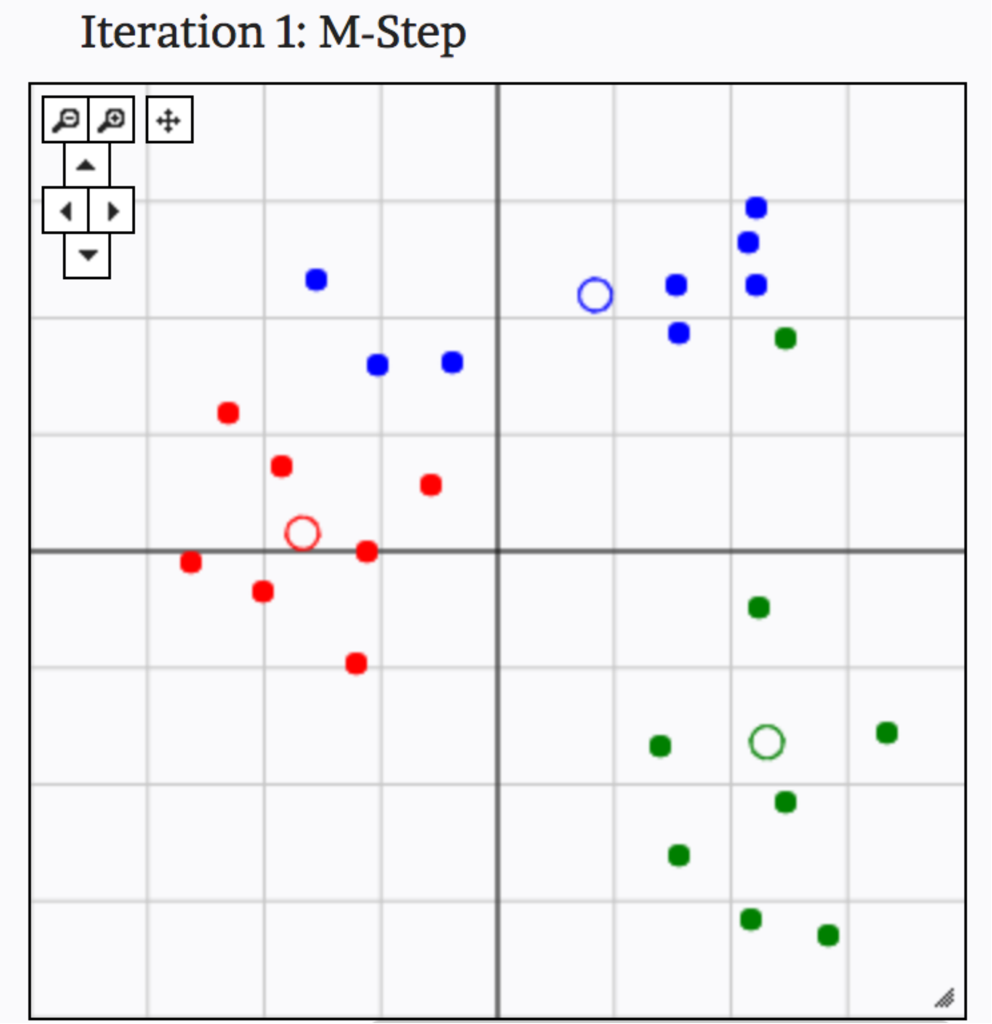
- From: http://util.io/k-means

K-means clustering in action



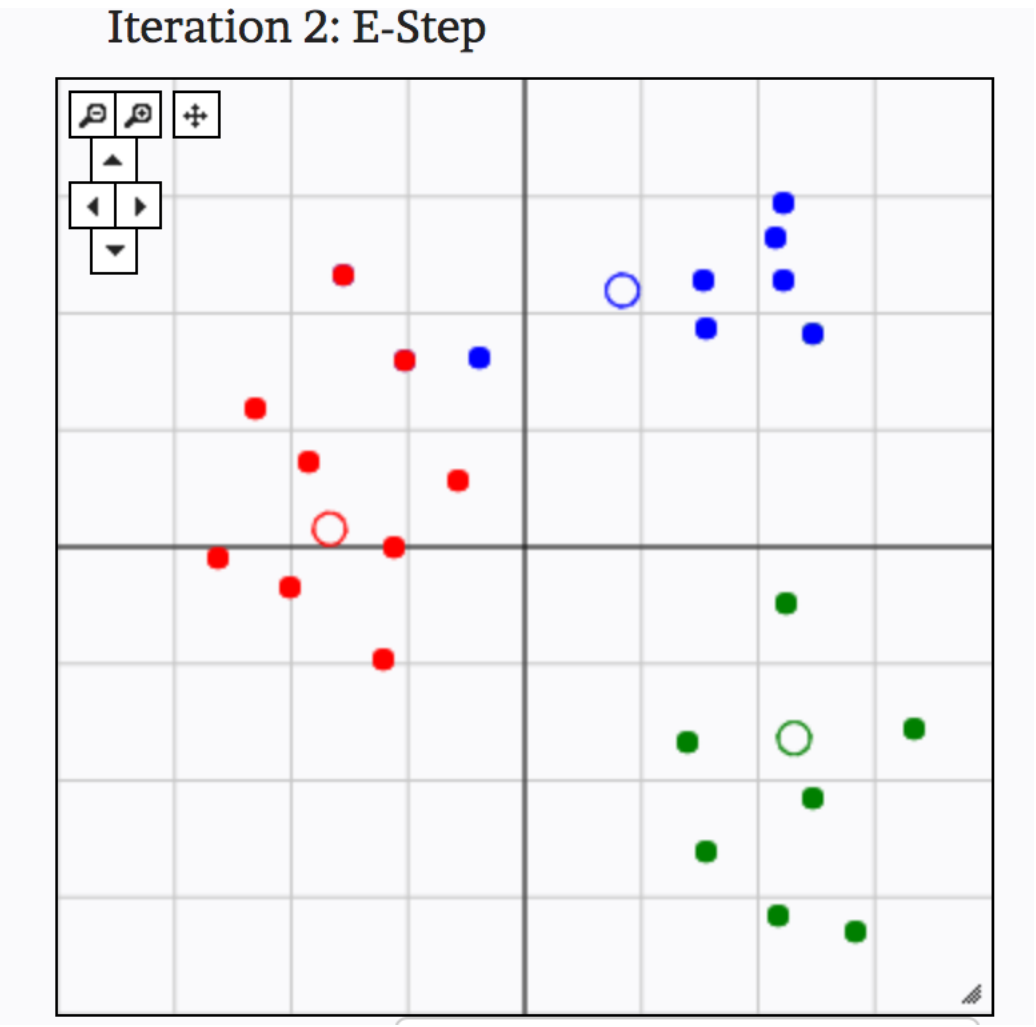
- From: http://util.io/k-means

K-means clustering in action



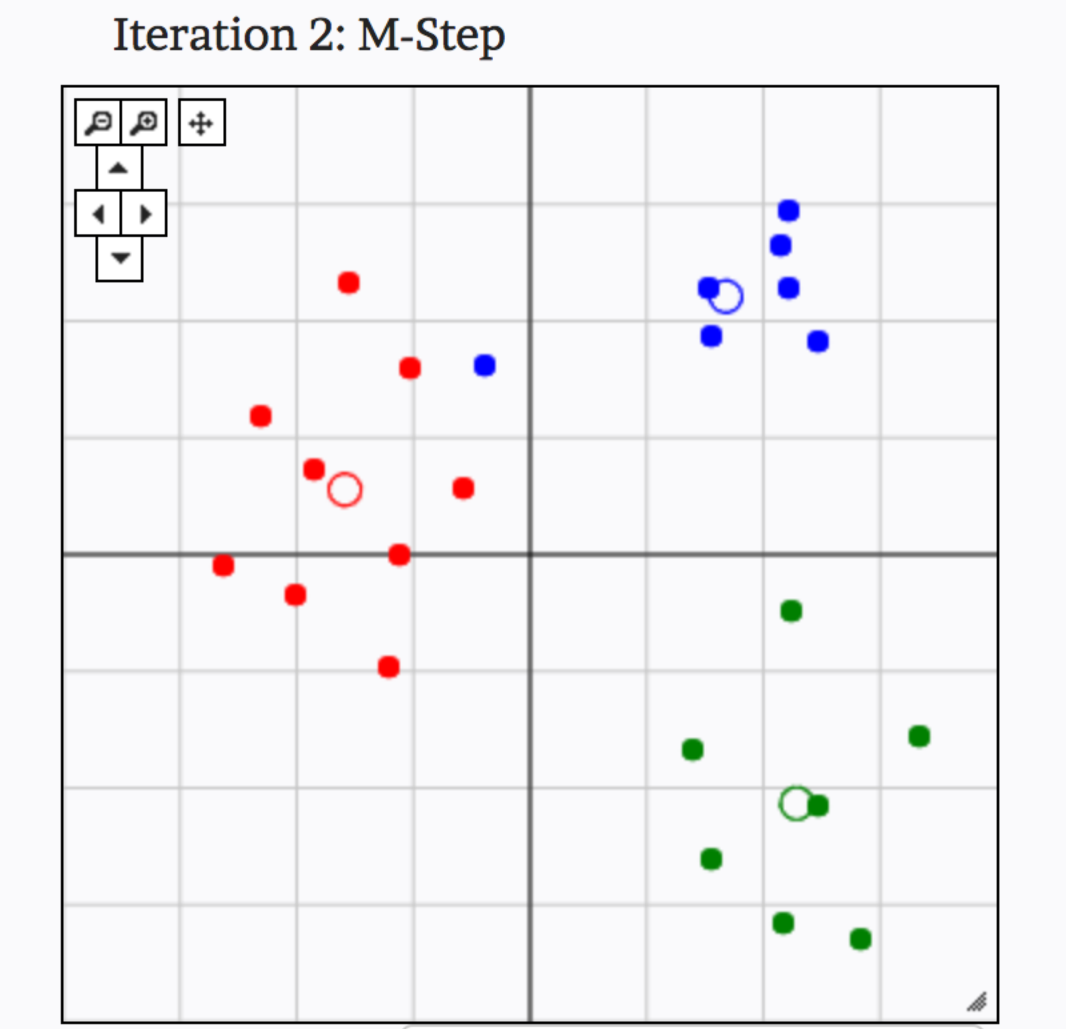
- From: http://util.io/k-means

K-means clustering in action



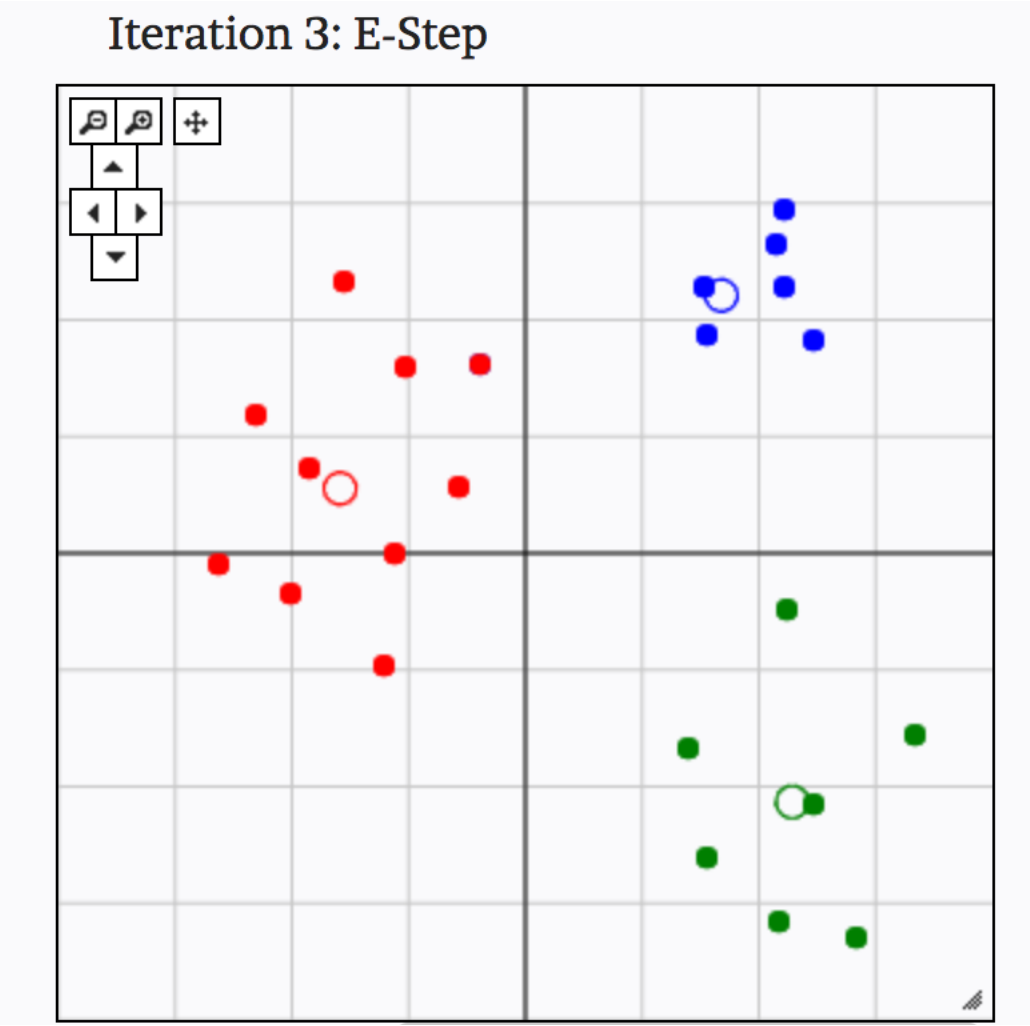
- From: http://util.io/k-means

K-means clustering in action



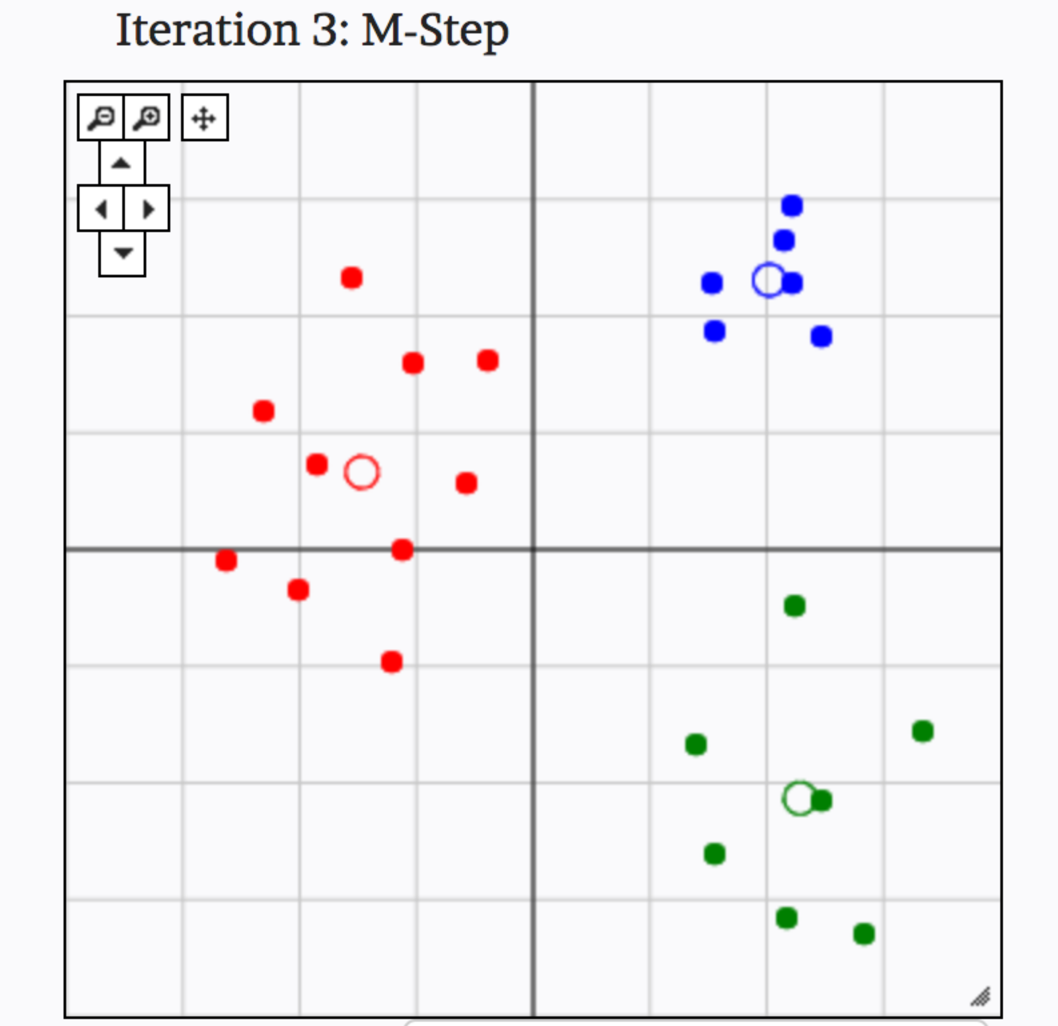
- From: http://util.io/k-means

K-means clustering in action



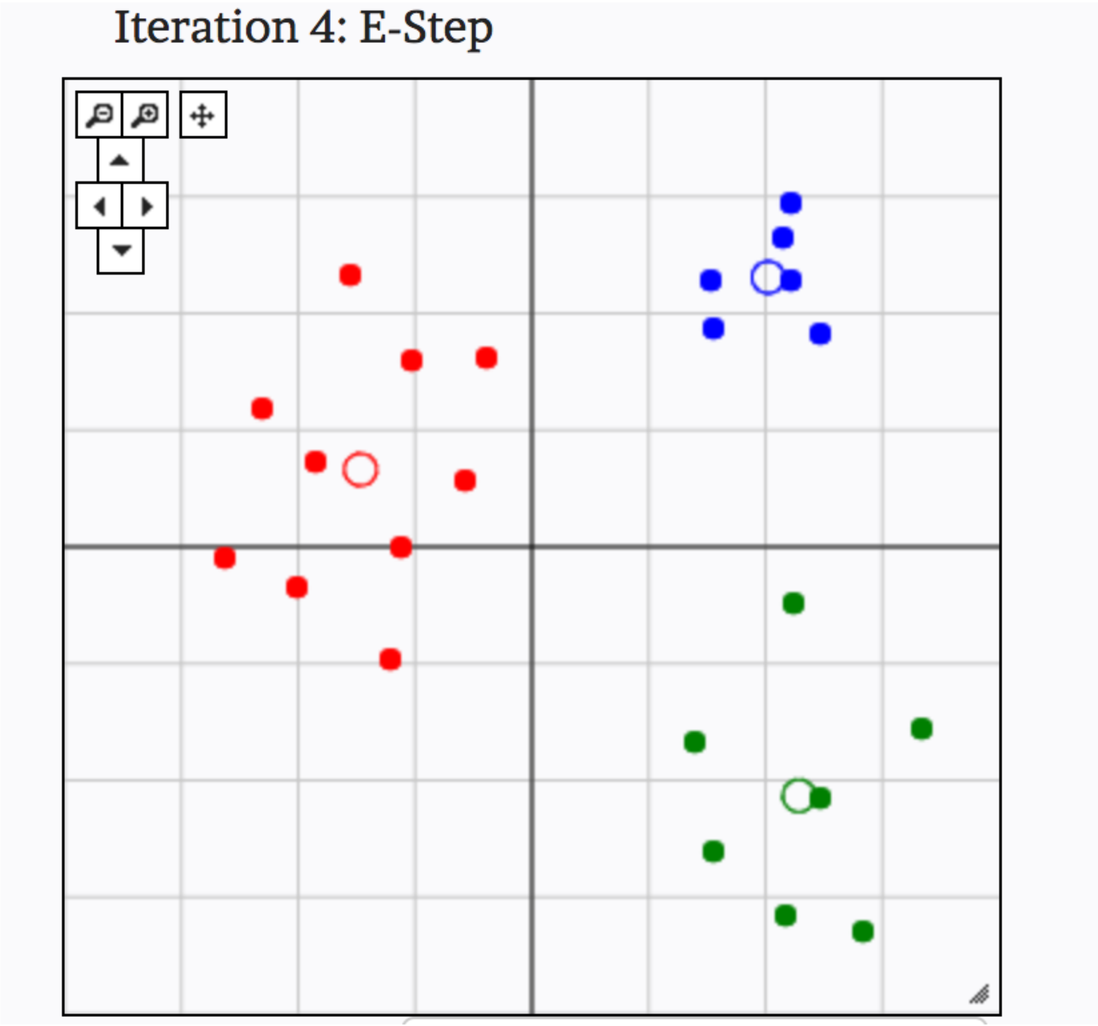
- From: http://util.io/k-means

K-means clustering in action



- From: http://util.io/k-means

K-means clustering in action



- From: http://util.io/k-means

Hierarchical clustering algorithm

1. Begin with n observations and calculate all of the pairwise dissimilarities. Treat each observation as its own cluster
2. For $i = n, n - 1, . . . , 2$:
   - Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the clusters that are most similar. Fuse these two clusters.
   - Compute the new pairwise inter-cluster dissimilarities among the $i - 1$ remaining clusters

Example Data:

Rowname	X1	X2
A	1	1
B	2	3
C	1	1.5
D	3	4
E	4	4.5

- Lets use hierarchical clustering on this simple data.
- Goal: which rows are most similar based on X1 and X2?

Example: Pairwise Dissimilarity Matrix (Manhattan Norm)

Dissimilarity Matrix

	A	B	C	D	D
A	0				
B					
C					
D					
E					

Dissimilarity for A versus A is:

$$\sum_{k=1}^{p} \left| x_k^A - x_k^A \right|$$

$$\left(\left| x_1^A - x_1^A \right| + \left| x_2^A - x_2^A \right| \right)$$

$$(|1 - 1| + |1 - 1|) = 0$$

Dataset

Rowname	X1	X2
A	1	1
B	2	3
C	1	1.5
D	3	4
E	4	4.5

Example: Pairwise Dissimilarity Matrix (Manhattan Norm)

	A	B	C	D	D
A	0				
B	3.0				
C					
D					
E					

Dissimilarity for A versus B is:

$$\sum_{k=1}^{p} \left|x_k^A - x_k^B\right|$$

$$\left(\left|x_1^A - x_1^B\right| + \left|x_2^A - x_2^B\right|\right)$$

$$(|1 - 2| + |1 - 3|) = 3$$

Rowname	X1	X2
A	1	1
B	2	3
C	1	1.5
D	3	4
E	4	4.5

Example: Pairwise Dissimilarity Matrix (Manhattan Norm)

	A	B	C	D	D
A	0				
B	3.0				
C	0.5				
D					
E					

Dissimilarity for A versus C is:

$$\sum_{k=1}^{p} |x_k^A - x_k^C|$$

$$(|x_1^A - x_1^C| + |x_2^A - x_2^C|)$$

$$(|1 - 1| + |1 - 1.5|) = 0.5$$

Rowname	X1	X2
A	1	1
B	2	3
C	1	1.5
D	3	4
E	4	4.5

Pairwise Dissimilarity Matrix

	A	B	C	D	D
A	$\boxed{0}$				
B	3.0	0			
C	$\boxed{0.5}$	2.5	0		
D	5.0	2.0	4.5	0	
E	6.5	3.5	6.0	1.5	0

- A and C have the lowest pairwise dissimilarity
- Therefore we create a new group that includes A-C and continue on with pairwise dissimilarity

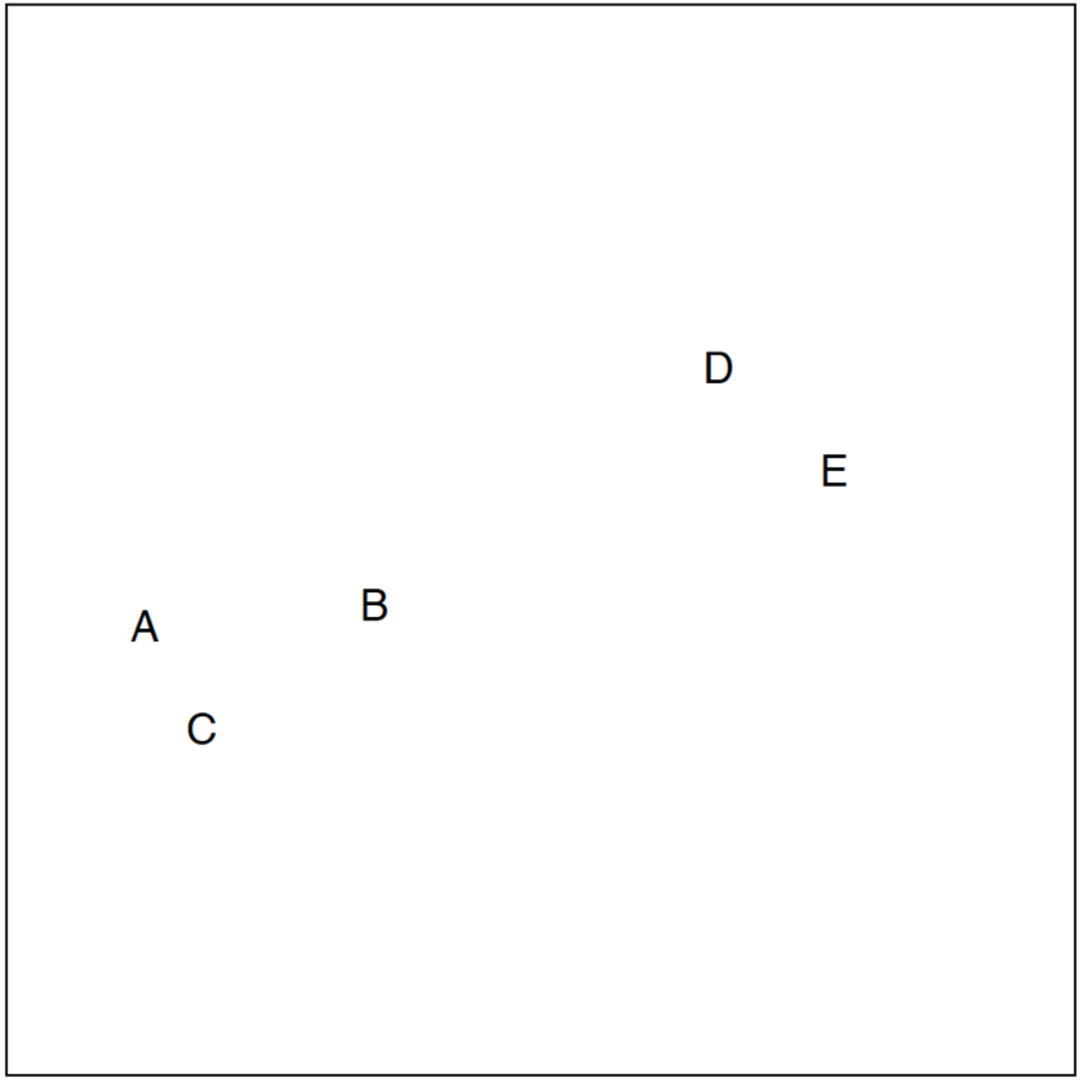
Pairwise Dissimilarity Matrix

	A	B	C	D	D
A	$\boxed{0}$				
B	3.0	0			
C	$\boxed{0.5}$	2.5	0		
D	5.0	2.0	4.5	0	
E	6.5	3.5	6.0	1.5	0

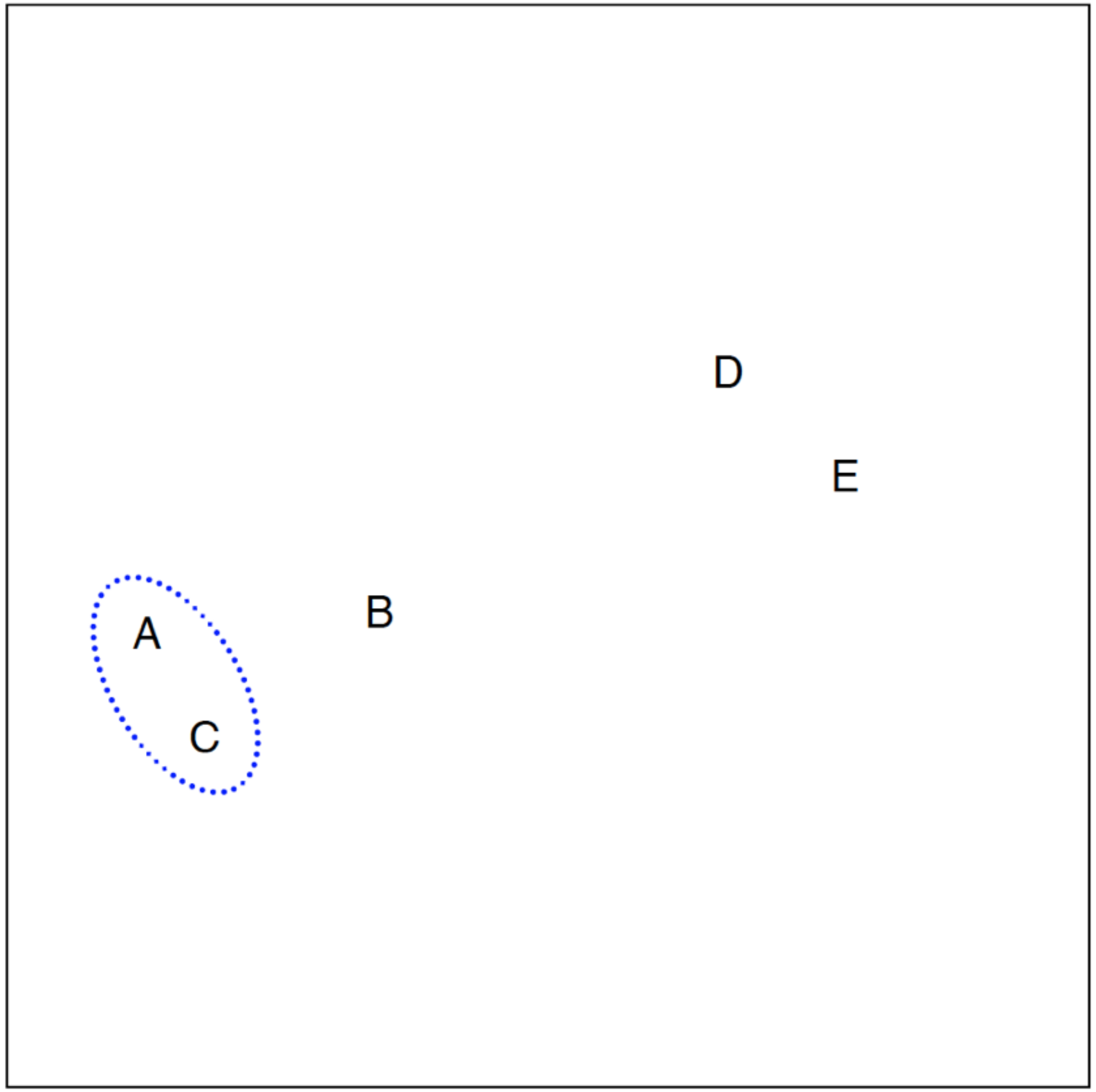
- Therefore we create a new group that includes A-C (averaging variables) and continue on with pairwise dissimilarity

Rowname	X1	X2
A-C	1	1.25
B	2	3
D	3	4
E	4	4.5

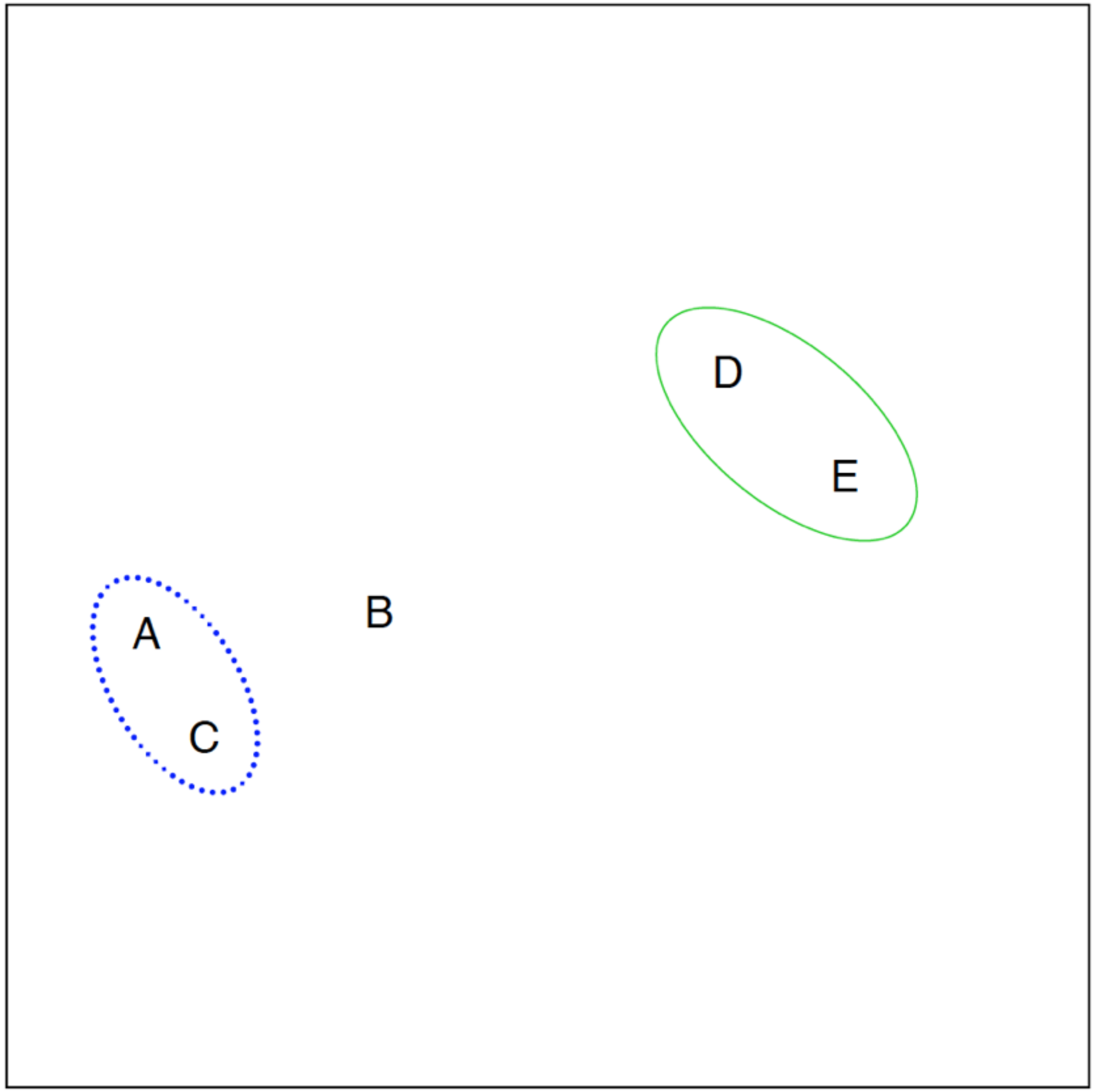
Hierarchical clustering in action



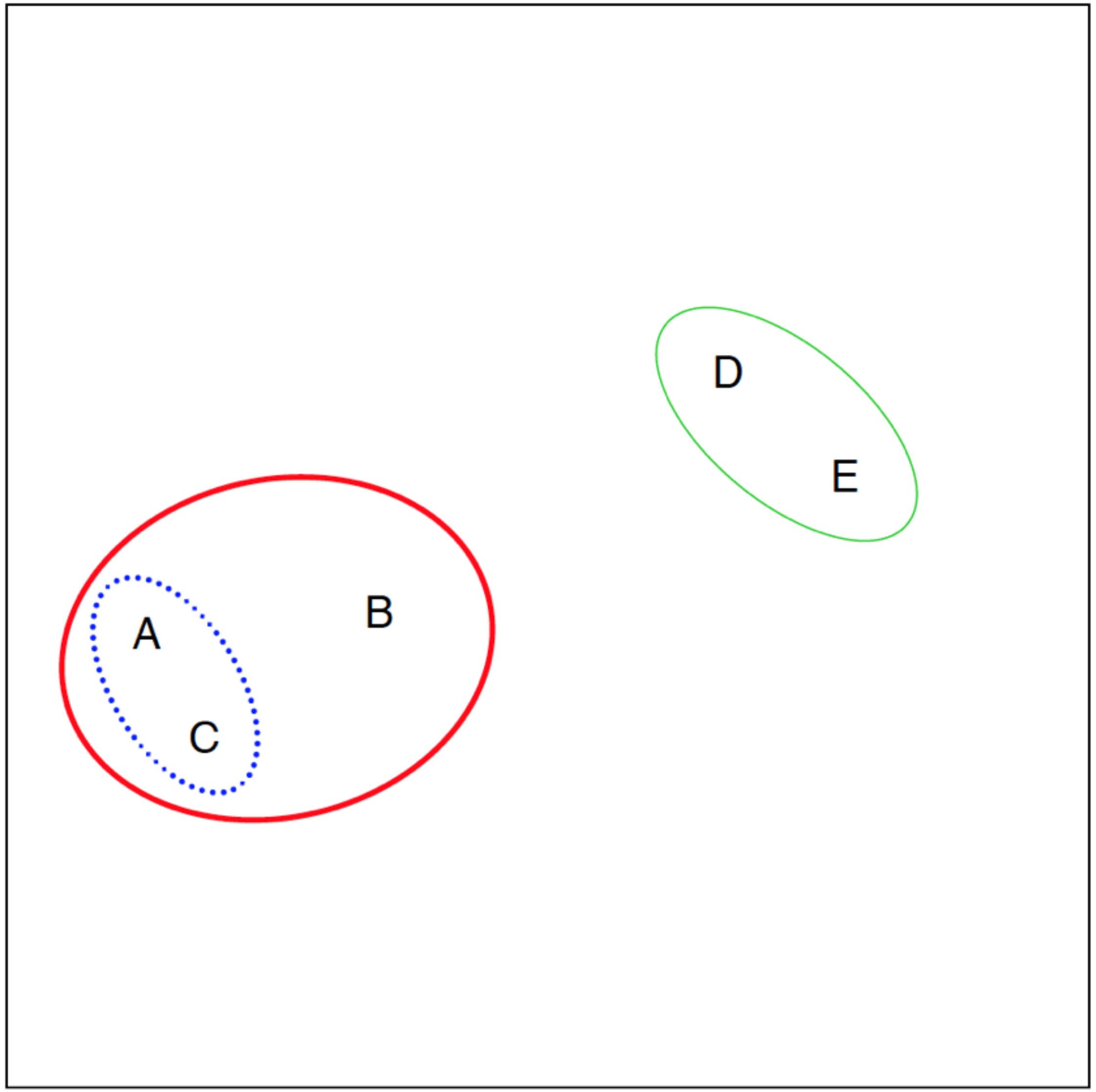
Hierarchical clustering in action



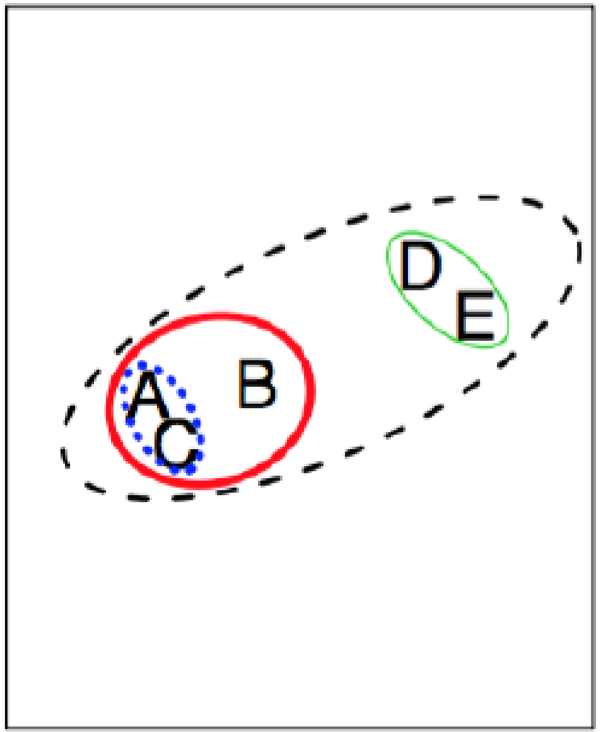
Hierarchical clustering in action

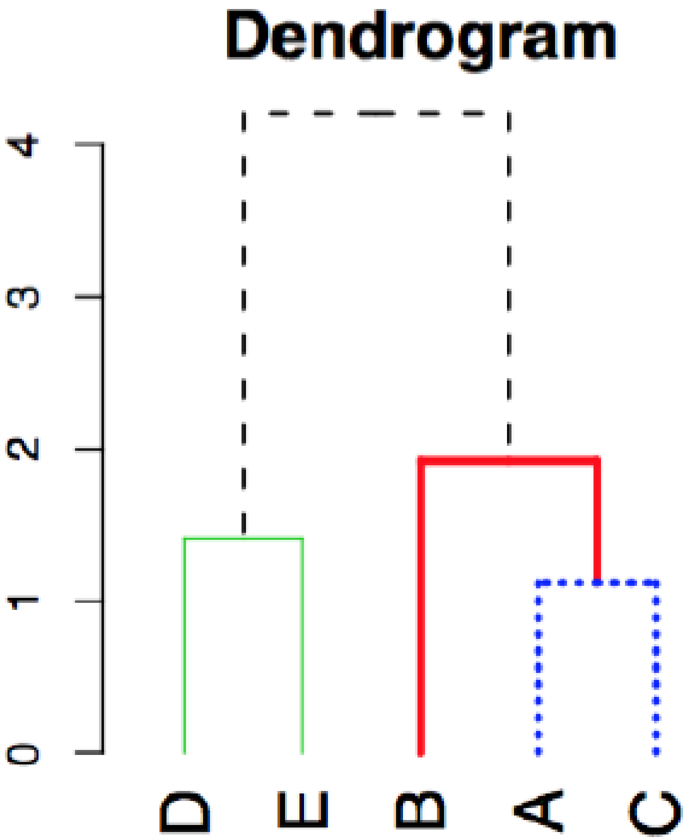


Hierarchical clustering in action



Connection to "dendrograms"





Customer Segmentation in R Using K-Means Clustering

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

Wholesale customers Data Set

Download: Data Folder, Data Set Description

Abstract: The data set refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories

Data Set Characteristics:	Multivariate	**Number of Instances:**	440	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	8	**Date Donated**	2014-03-31
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	304081

Source:

Margarida G. M. S. Cardoso, margarida.cardoso '@' iscte.pt, ISCTE-IUL, Lisbon, Portugal

Data Set Information:

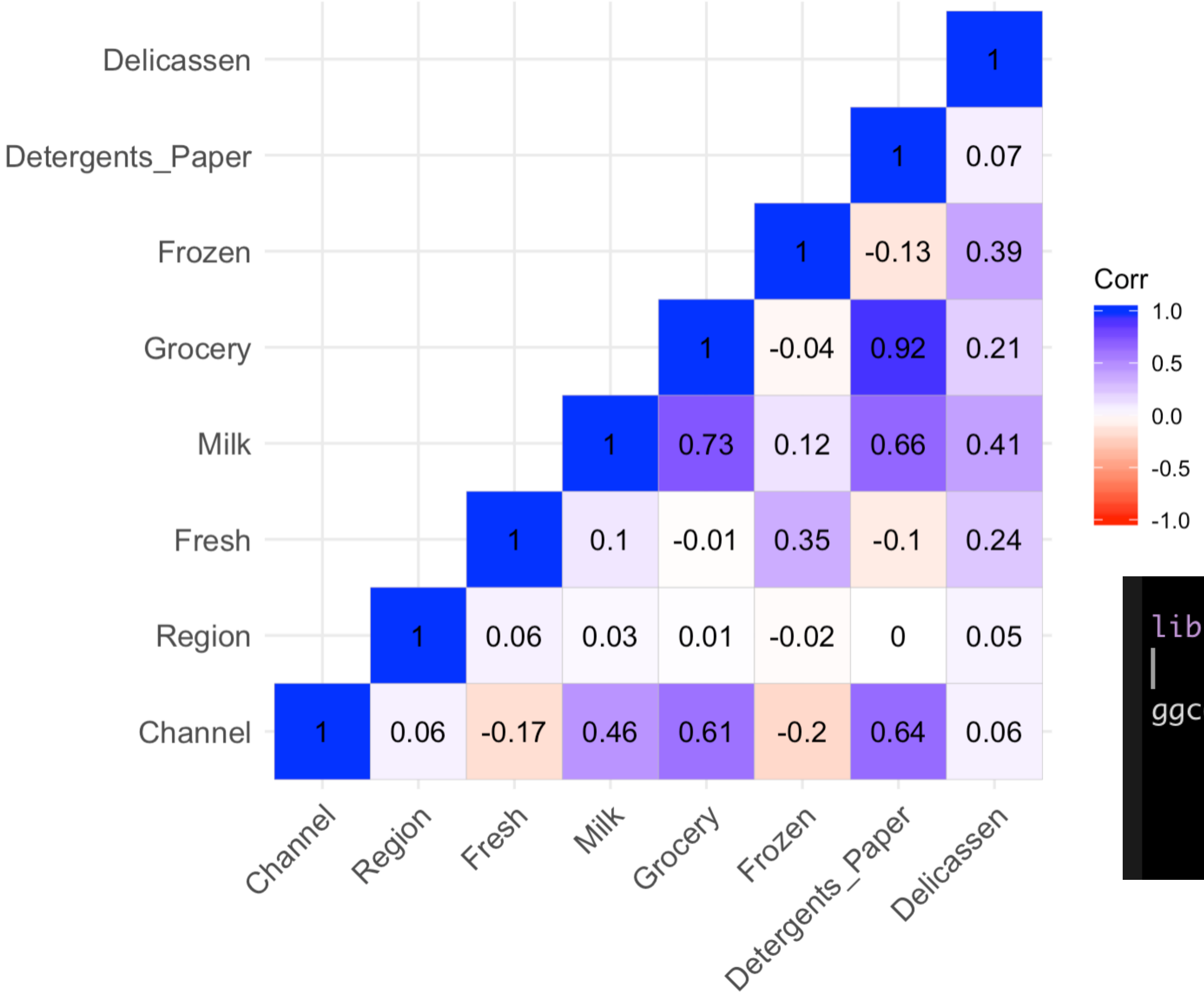
Provide all relevant information about your data set.

Customer Segmentation in R Using K-Means Clustering

```r
#-----------------------------------------
# Customer Segmentation
#-----------------------------------------

Customers_DF <-
  read_csv(here::here("datasets",
           "Wholesale_customers.csv"))

head(Customers_DF)

summary(Customers_DF)

View(Customers_DF)
```

```
> summary(Customers_DF)
    Channel          Region          Fresh             Milk
 Min.   :1.000   Min.   :1.000   Min.   :     3   Min.   :   55
 1st Qu.:1.000   1st Qu.:2.000   1st Qu.:  3128   1st Qu.: 1533
 Median :1.000   Median :3.000   Median :  8504   Median : 3627
 Mean   :1.323   Mean   :2.543   Mean   : 12000   Mean   : 5796
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.: 16934   3rd Qu.: 7190
 Max.   :2.000   Max.   :3.000   Max.   :112151   Max.   :73498
    Grocery          Frozen          Detergents_Paper    Delicassen
 Min.   :    3   Min.   :   25.0   Min.   :    3.0   Min.   :    3.0
 1st Qu.: 2153   1st Qu.:  742.2   1st Qu.:  256.8   1st Qu.:  408.2
 Median : 4756   Median : 1526.0   Median :  816.5   Median :  965.5
 Mean   : 7951   Mean   : 3071.9   Mean   : 2881.5   Mean   : 1524.9
 3rd Qu.:10656   3rd Qu.: 3554.2   3rd Qu.: 3922.0   3rd Qu.: 1820.2
 Max.   :92780   Max.   :60869.0   Max.   :40827.0   Max.   :47943.0
```

Correlation Matrix Across Variables



- Recall: if two variables have a correlation of 1, they are equivalent up to a constant

```
library('ggcorrplot')

ggcorrplot(round(cor(Customers_DF),2),
           type = "lower", insig = "blank",
           show.diag = TRUE, lab = TRUE,
           colors = c("red","white","blue"))
```

How to Pick k (Number of Clusters)?

1. Silhouette score:

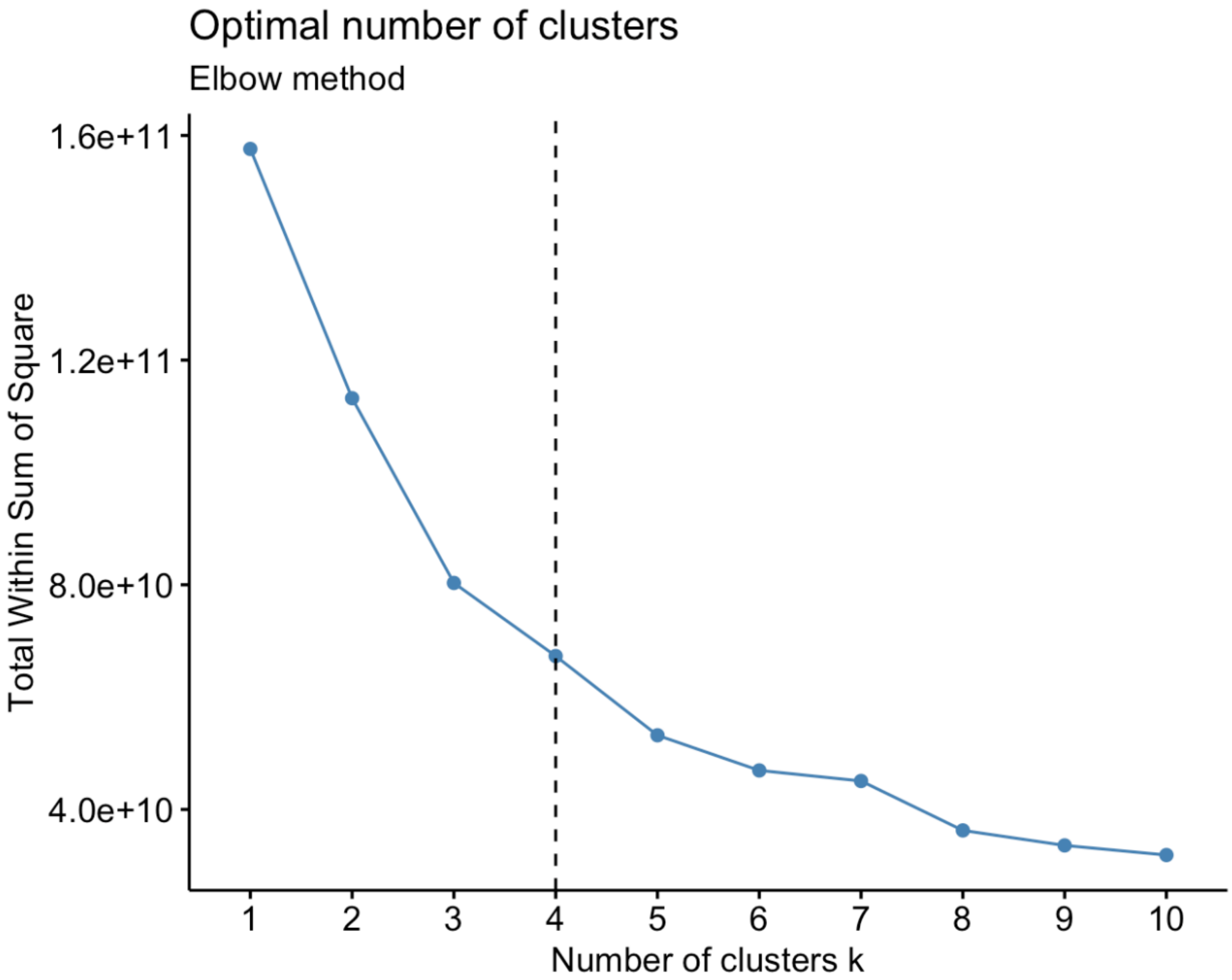
- Silhouette score: $s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$. Higher number
- a_i = measure of dissimilarity between i and other points in cluster. b_i = measure of dissimilarity between i all other points

2. **Elbow method**

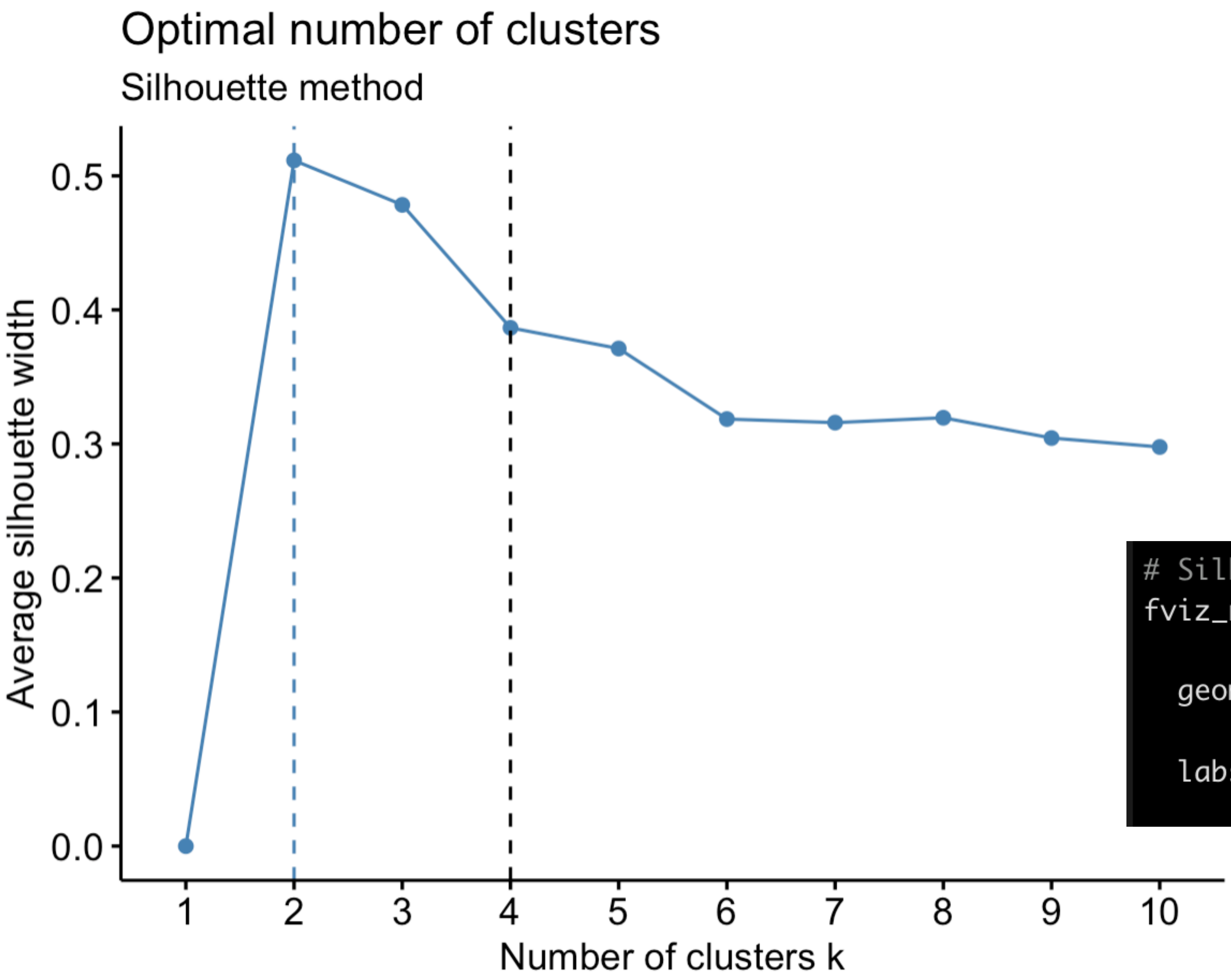
- Look for “kink” in within-cluster sum of square errors

3. Gap Statistic Method

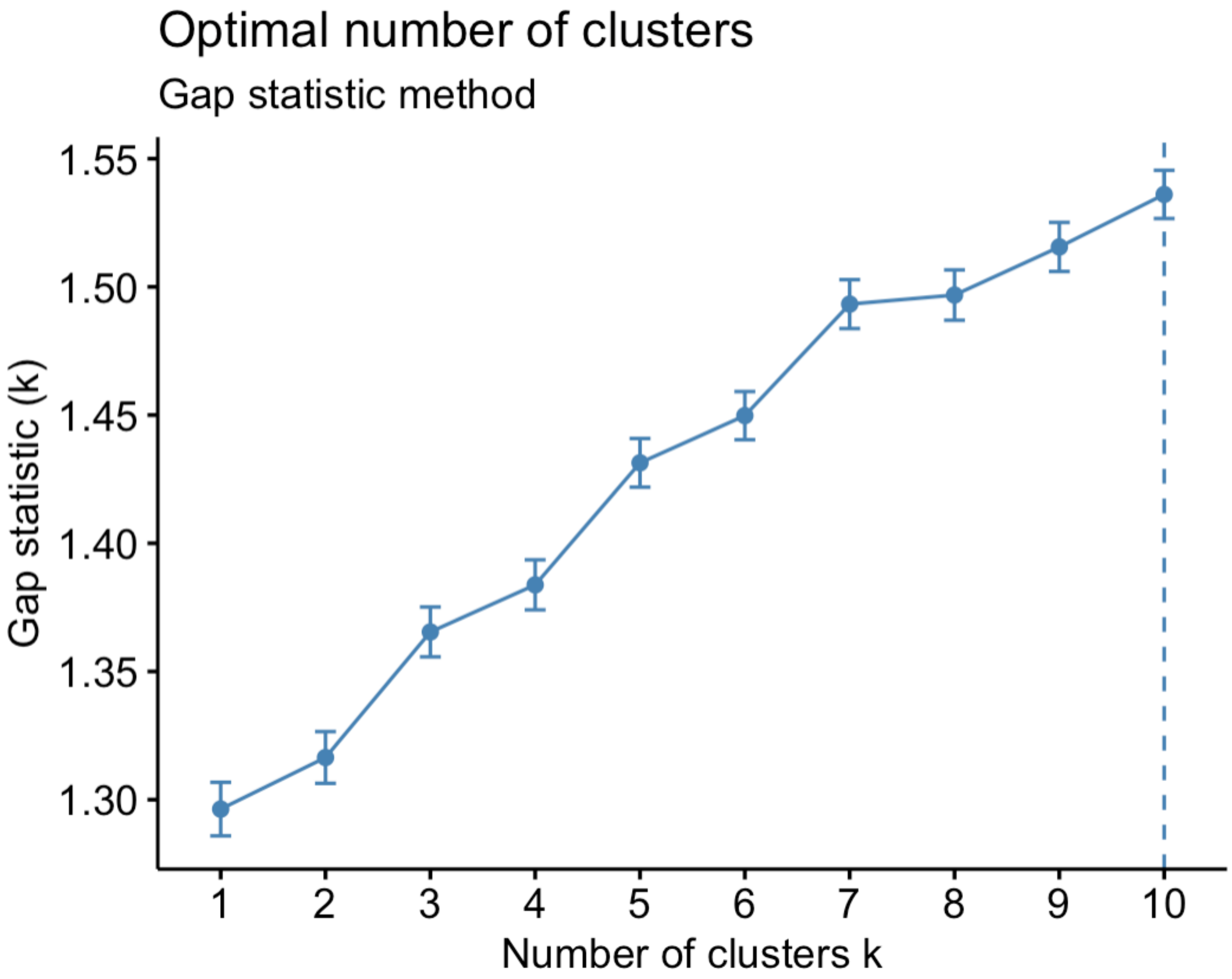
- Compares intra-cluster variation with “expected” under the null (no clustering)




```
# elbow method
fviz_nbclust(Customers_DF,
             kmeans,
             method = "wss") +
  geom_vline(xintercept = 4,
             linetype = 2)+
  labs(subtitle = "Elbow method")
```




```
# Silhouette method
fviz_nbclust(Customers_DF, kmeans,
             method = "silhouette") +
  geom_vline(xintercept = 4,
             linetype = 2)+
  labs(subtitle = "Silhouette method")
```




```
# Gap Statistic method
fviz_nbclust(Customers_DF,
             kmeans,
             nstart = 25,
             method = "gap_stat",
             nboot = 500) +
 labs(subtitle =
        "Gap statistic method")
```

Breaking Out The Big Guns: NbClust

- NbClust compares across 24 statistical methods for optimal cluster #s

```
install.packages('NbClust')
library('NbClust')
NbClust(Customers_DF,
        diss = NULL,
        distance = "euclidean",
        min.nc = 2,
        max.nc = 15,
        method = "kmeans")
```

```
> Nb_cl$Best.nc[1,]
         KL          CH   Hartigan         CCC      Scott    Marriot
          8           5          3           3          4          4
     TrCovW      TraceW   Friedman       Rubin     Cindex         DB
          3           3          3           8          9          8
 Silhouette        Duda   PseudoT2       Beale  Ratkowsky       Ball
          2           2          2           2          4          3
 PtBiserial        Frey    McClain        Dunn     Hubert    SDindex
          3           5          2          10          0          3
     Dindex        SDbw
          0          15
* 1 proposed 10 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

Estimate k-means Using k=3

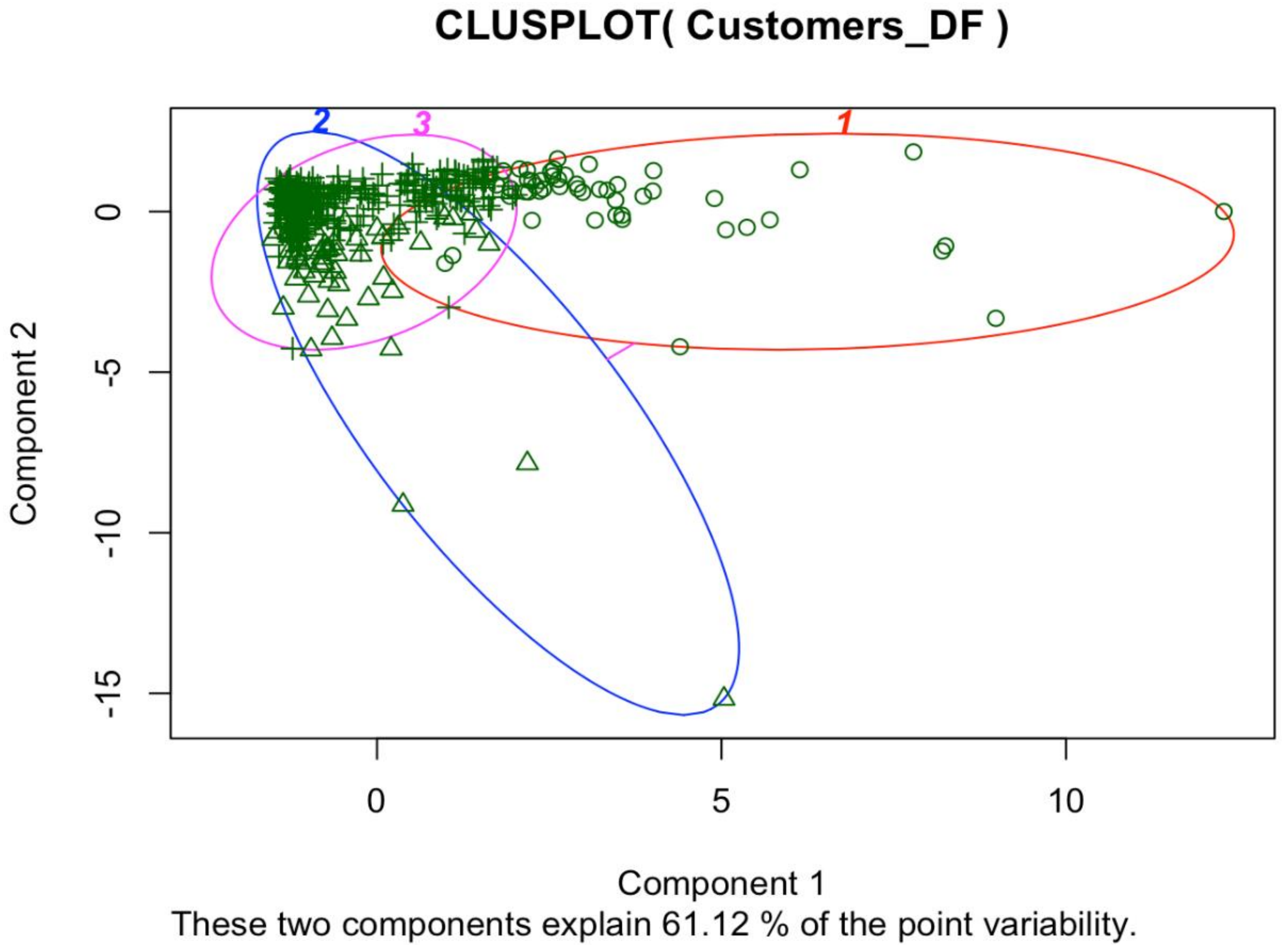
```r
# finally do kmeans
kmeans3 <- kmeans(Customers_DF,
                  centers = 3,
                  nstart = 25)

kmeans3$centers
```

- Centers of k-means show average X variable value for each cluster

```
> kmeans3$centers
   Channel   Region    Fresh      Milk   Grocery   Frozen Detergents_Paper Delicassen
1 1.960000 2.440000  8000.04 18511.420 27573.900 1996.680        12407.360   2252.020
2 1.133333 2.566667 35941.40  6044.450  6288.617 6713.967         1039.667   3049.467
3 1.260606 2.554545  8253.47  3824.603  5280.455 2572.661         1773.058   1137.497
```

- Three types of shoppers: group 1, 2, 3
- How do they differ?
- How would you characterize the three clusters?

```
clusplot(Customers_DF,
         kmeans3$cluster,
         color=TRUE,
         shade=FALSE,
         labels=5, lines=2)
```



Customer Clusters

Group	Fresh/Grocery Purchase?	Frozen/Deli Purchase?	Conceptual name?
1			
2			
3			

```
> kmeans3$centers
   Channel   Region    Fresh      Milk   Grocery   Frozen Detergents_Paper Delicassen
1 1.960000 2.440000  8000.04 18511.420 27573.900 1996.680        12407.360   2252.020
2 1.133333 2.566667 35941.40  6044.450  6288.617 6713.967         1039.667   3049.467
3 1.260606 2.554545  8253.47  3824.603  5280.455 2572.661         1773.058   1137.497
```

Lab (Time Permitting)

```r
#------------------------------------------------------------------------
# Lab Exercises
#------------------------------------------------------------------------
# 1. Load the customers data frame

Customers_DF <-
  read_csv(here::here("datasets",
                      "Wholesale_customers.csv"))

# 2. Use the `fviz_nbclust` function to calculate the optimal
#    number of clusters using the 'silhouette method'

# 3. How many clusters does the silhouette method suggest we should use?

# 4. Use the `fviz_nbclust` function to calculate the optimal
#    number of clusters using the 'total within sum of sqaure' method

# 5. How many clusters does the sum of square method suggest we should us

# 6. Cluster the data using the number of distinct clusters informed by
#    silhouette and sum of squares method. (Use the whole number average
#    between them if they disagree).
#    Use the kmeans function to calculate the clustering model and then
#    print out the centers and their average values.

# 7. Interpret each cluster. Give each one a name that is informative of
#    its component stores.

# 8. Suppose you run a business distributing groceries and supplies
#    and this dataset lists all of your customers and their purchases.
#    How might you use the cluster assignment you give to every store
#    in this dataset to better run your business?
```