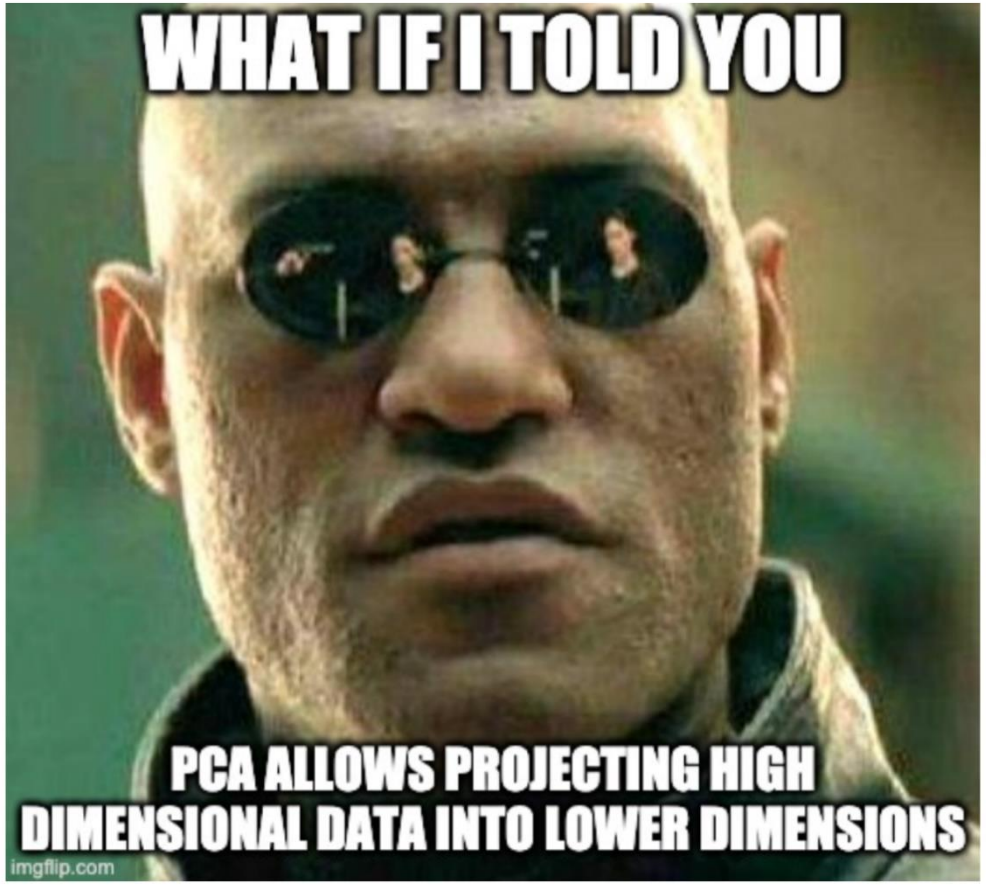
PCA, clustering and conclusions



Overview

Principal components analysis (PCA) continued

Clustering

Brief mention/pointers to additional topics

- Ethics
- String manipulation
- Interactive "shiny" apps

Conclusions

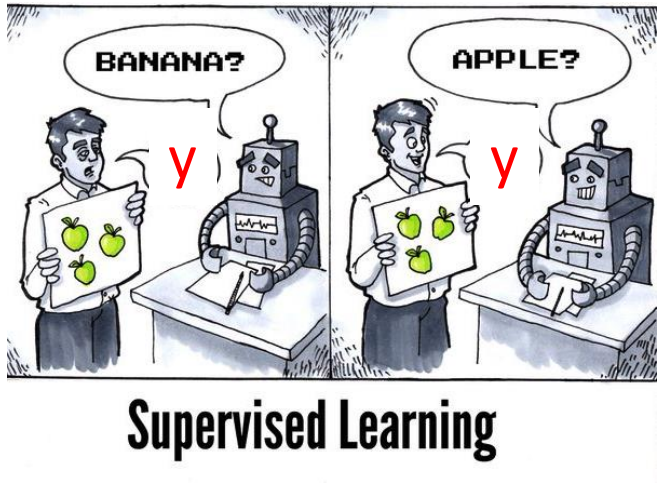
Principal Component Analysis

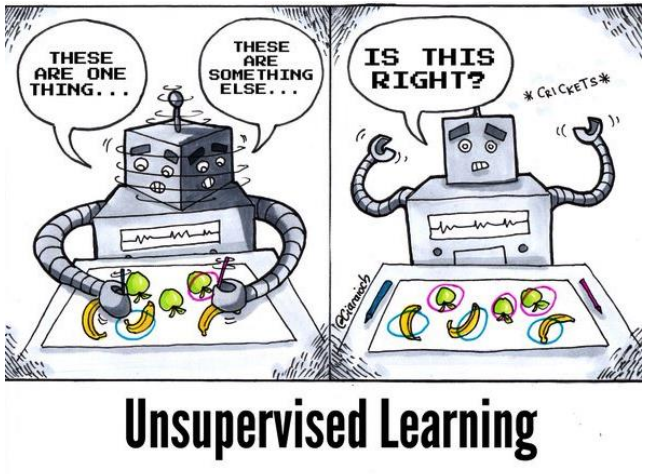
Unsupervised learning

We will discuss two types of unsupervised learning:

1. **Dimensionality reduction** where we try to find a smaller set of features that captures most of the variability in the data
   - Principal component analysis (PCA)

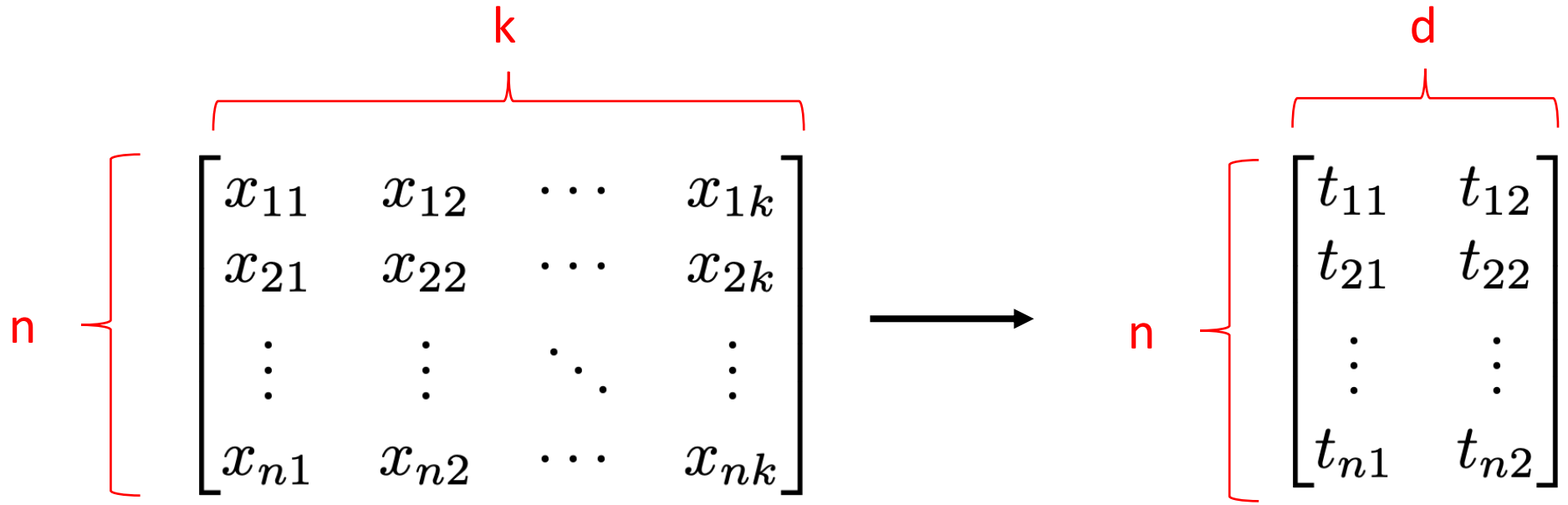
2. **Clustering** where we try to group similar data points together



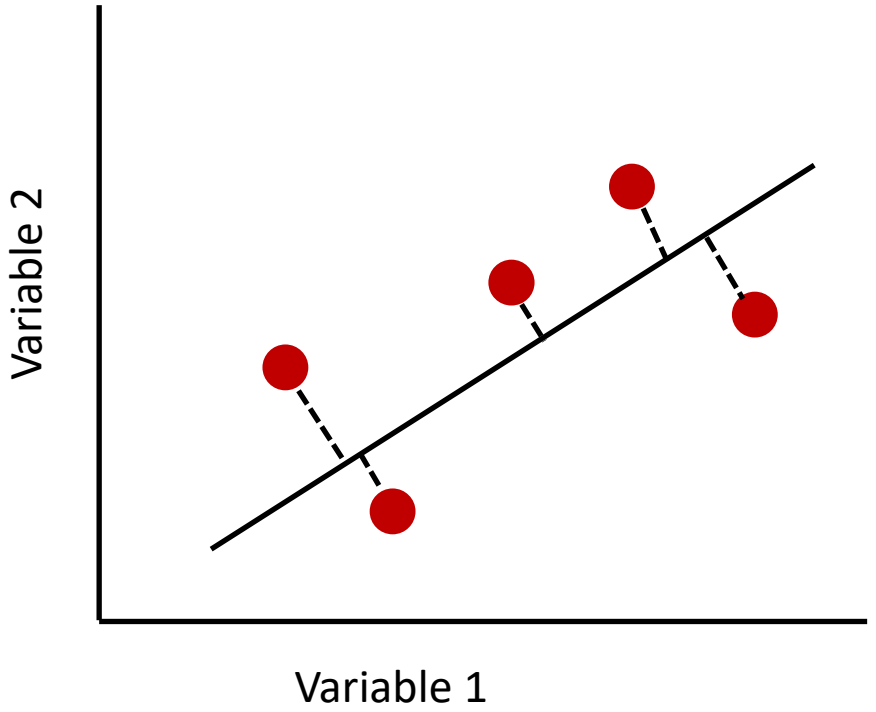


Principal Component Analysis

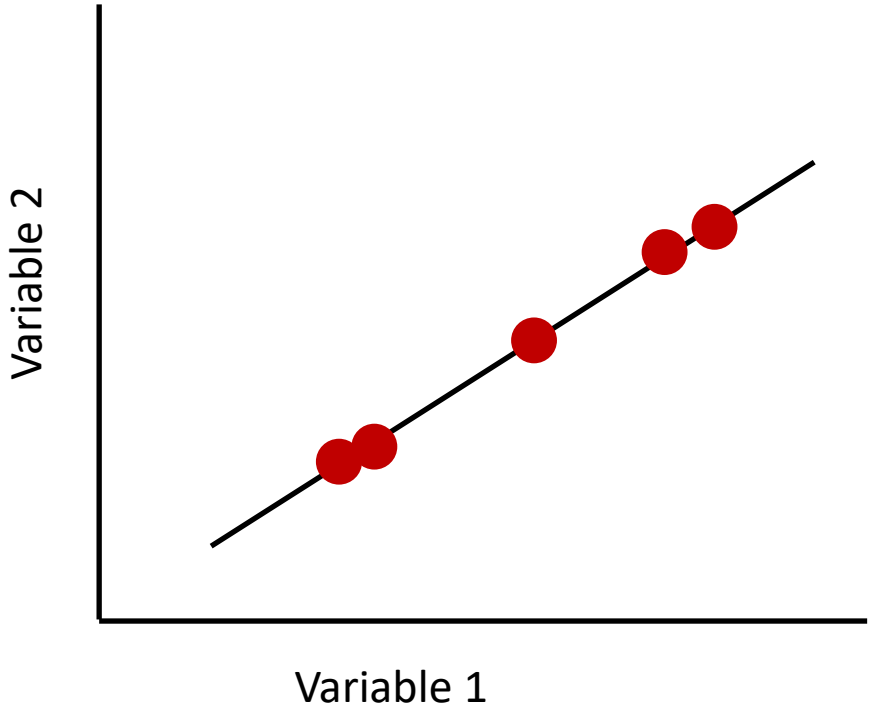
Principal Component Analysis is a dimensionality method that tries to capture most of the variability in the original data



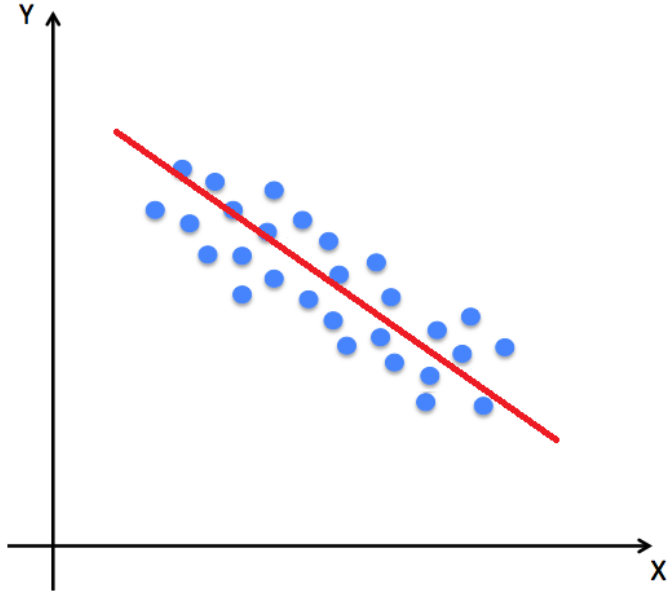
Principal Component Analysis



Principal Component Analysis



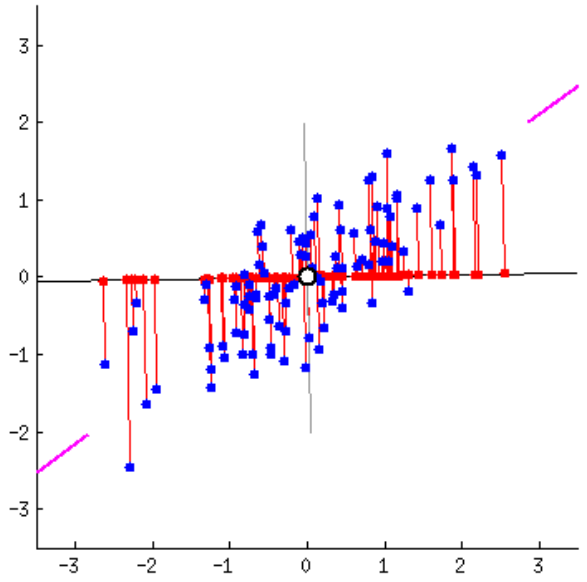
Principal Component Analysis



The loadings for the first principal component are found by finding the projection vector $A_1 = (\alpha_{11}, \alpha_{21}, \alpha_{k1})$ such that the variance of the t_i is maximized

Find the α's that maximize:

$$\frac{1}{n-1}\sum_{i=1}^{n} t_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\alpha_{11}z_{i1} + \alpha_{21}z_{i2} + \ldots + \alpha_{k1}z_{ik})^2$$



Subject to the constraint:

$$\sum_{j=1}^{k} \alpha_{j1}^2 = 1$$

The Second Principal Component

The second principal component scores t_{i2} is the linear combination of the $z_1, z_2, \ldots, z_k$ that has maximal variance and is **uncorrelated** with the first principal component scores t_{i1}

- $t_{i2} = \alpha_{12}z_1 + \alpha_{22}z_2 + \ldots + \alpha_{k2}z_k$
- $cor(T_1, T_2) = 0$

This is equivalent of having A_1 be orthogonal to A_2

- $A_1^T A_2 = 0$

$$\sum_{j=1}^{k} \alpha_{j1} \cdot \alpha_{j2} = 0$$

First principal component

$$\begin{bmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{n1} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{bmatrix} \begin{bmatrix} \alpha_{11} \\ \alpha_{21} \\ \vdots \\ \alpha_{k1} \end{bmatrix}$$

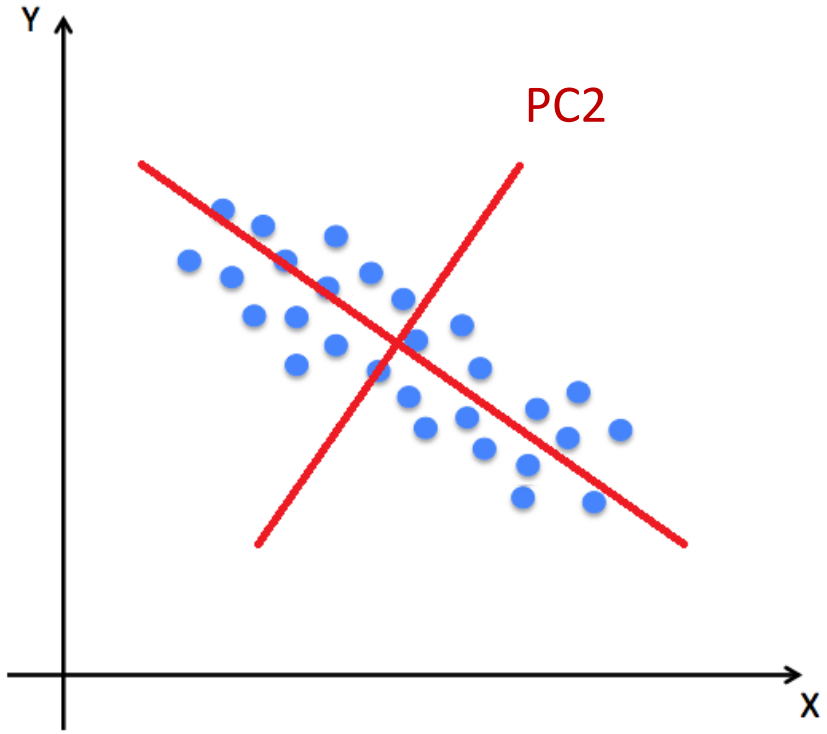
Second principal component

$$\begin{bmatrix} t_{12} \\ t_{22} \\ \vdots \\ t_{n2} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{bmatrix} \begin{bmatrix} \alpha_{12} \\ \alpha_{22} \\ \vdots \\ \alpha_{k2} \end{bmatrix}$$

Geometric interpretation of the second PC

Find the direction that maximizes the variance of t_i's

- Data projected on to the principal component is most spread out that is perpendicular (orthogonal) to the other PCs



First and second principal components

$$\begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ \vdots & \vdots \\ t_{n1} & t_{n2} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nk} \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \\ \vdots & \vdots \\ \alpha_{k1} & \alpha_{k2} \end{bmatrix}$$

Higher Principal Components

We continue this process until we find all the principal component scores, T_1, T_2, … T_d

- The principal component scores are unique up to a sign flip $T_i = -T_i$
  - To find the principal components what is really done is an eigenvalue decomposition of the covariance matrix.

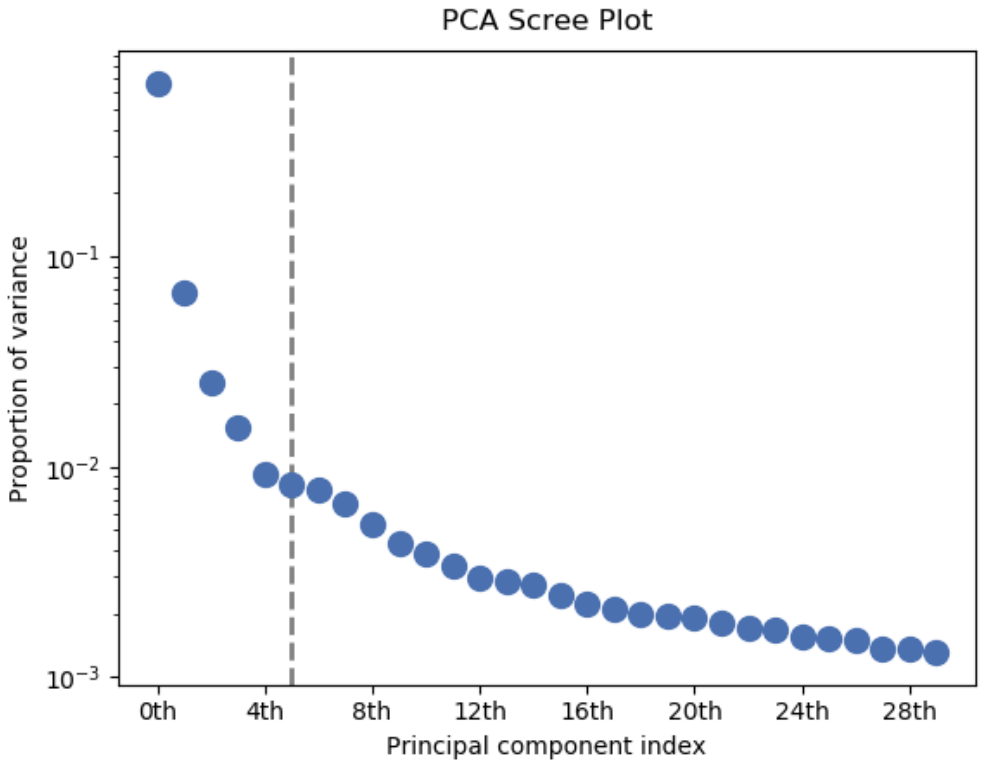
All principal components

$$\begin{bmatrix} t_{11} & t_{12} & \dots & t_{1d} \\ t_{21} & t_{22} & \dots & t_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \dots & t_{nd} \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1k} \\ z_{21} & z_{22} & \dots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{nk} \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \dots & \alpha_{1d} \\ \alpha_{21} & \alpha_{22} & \dots & \alpha_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{k1} & \alpha_{k2} & \dots & \alpha_{kd} \end{bmatrix}$$

Deciding how many PCs to use

A **scree plot** shows the PVE as a function of PC number

- The number of PCs chosen is often selected by looking for the “elbow” in this plot
  - i.e., point where PVE stop dramatically dropping and levels off



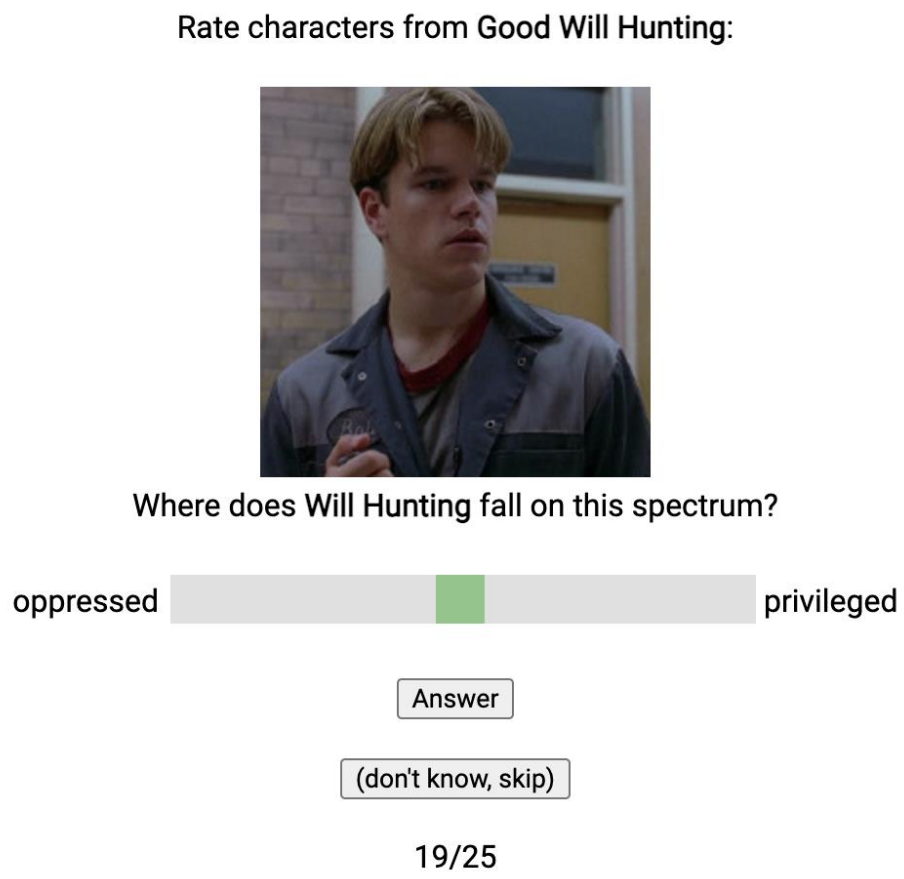
PCA example: personality traits of fictional characters

The Open-Source Psychometrics Project conducted a survey where they got ratings of 235+ personality traits from 800 fictional characters.

Let's use PCA to assess:

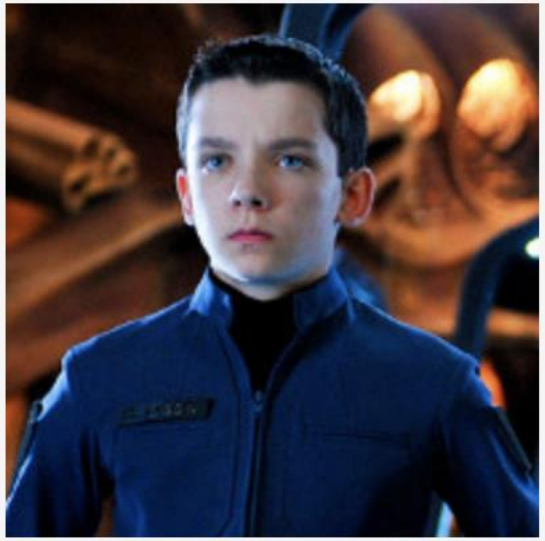
- How to personality traits commonly covary
- Which fictional characters are most similar

If you want to find out which fictional character you are most similar you can take their "Which Character" personality quiz



Let’s try the PCA in R…

The best match between the self assessment you provided and the profile of a fictional character as rated by other people who have taken this survey is the character **Ender Wiggin** (Ender's Game).



84% match

Your traits versus their traits are graphed below (click on points for labels).

Clustering

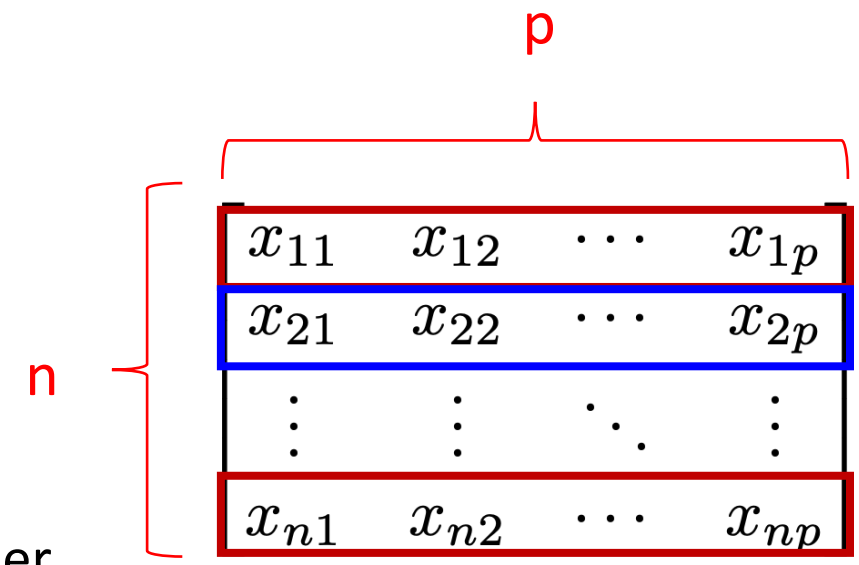
How many clusters?

Six Clusters

Two Clusters

Four Clusters

Clustering



Clustering divides n data points x_i's into subgroups

- Data points in the same group are similar/homogeneous
- Data points in different groups are different from each other

Examples:

- Examining gene expression levels to group cancer types together
- Examining consumer purchasing behavior to perform market segmentation

Clustering can be:

- **Flat**: no structure beyond dividing points into groups
- **Hierarchical**: Population is divided into smaller and smaller groups (tree like structure)

K-means clustering

K-means clustering partitions the data into ***K*** distinct, non-overlapping clusters

- i.e., each data point x_i belongs to exactly one cluster C_k

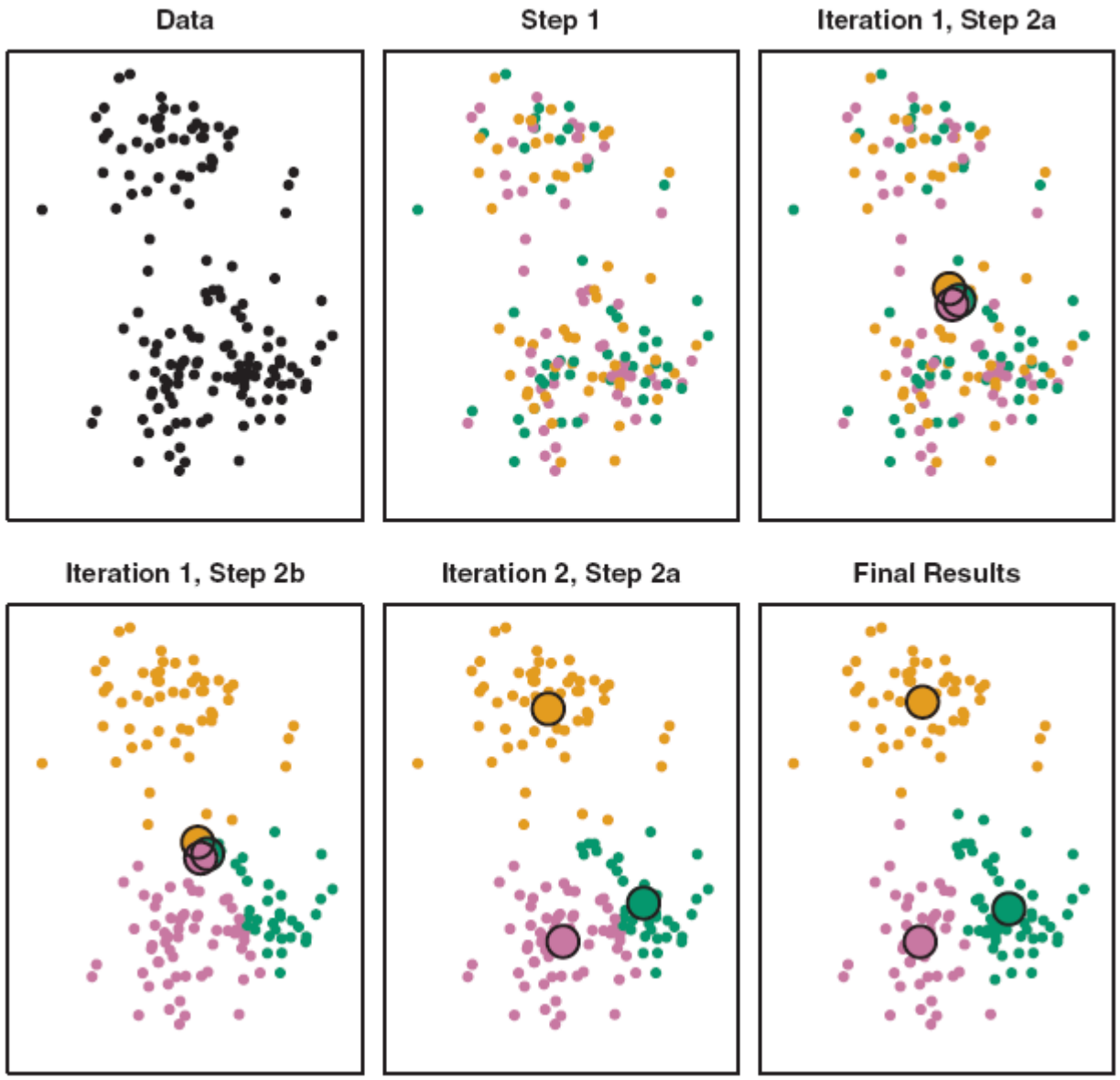
The number of clusters, ***K***, needs to be specified prior to running the algorithm

The goal is to minimize the within-cluster variation for some measure $W(C_k)$

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

K-means clustering

1. Randomly assign points to clusters C_k
2. Calculate cluster centers as means of points in each cluster
3. Assign points to the closest cluster center
4. Recalculate cluster center as the mean of points in each cluster
5. Repeat steps 3 and 4 until convergence



K-means clustering

Because only a local minimum is found, different random initializations will lead to different solutions

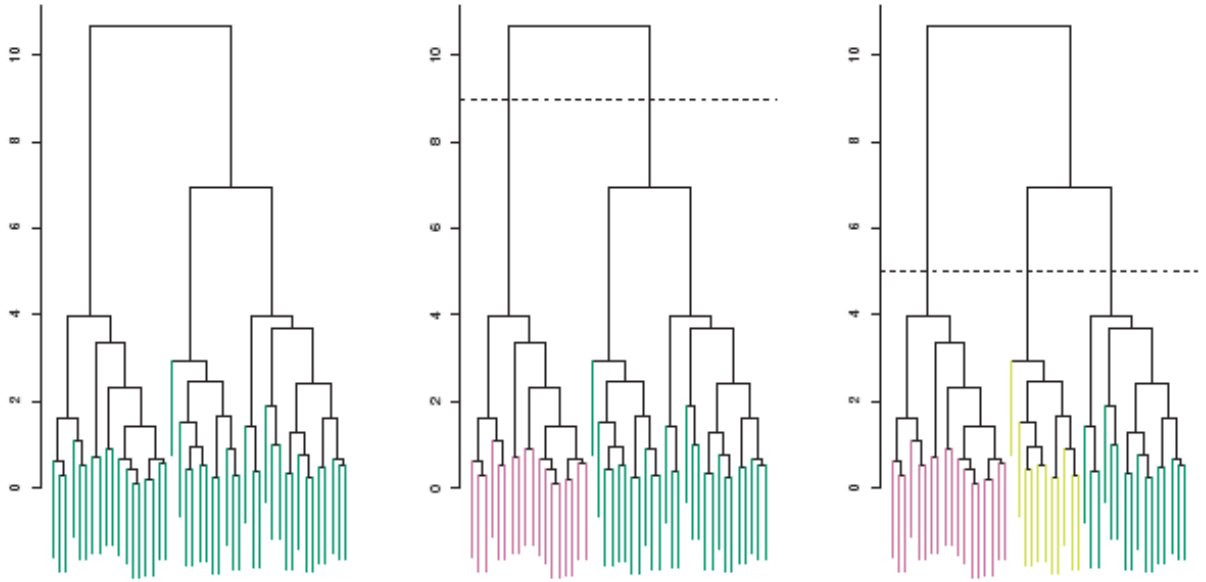
- One should run the algorithm multiple times to get better solutions

Hierarchical clustering

In **hierarchical clustering** we create a dendrogram which is a tree-based representation of successively larger clusters.

We can cut the dendrogram at any point to create as many clusters as desired

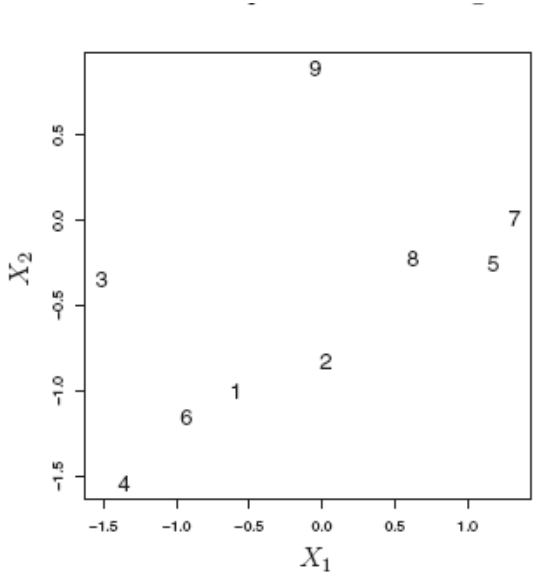
- i.e., don't need to specify the number of clusters, K, beforehand



Hierarchical clustering

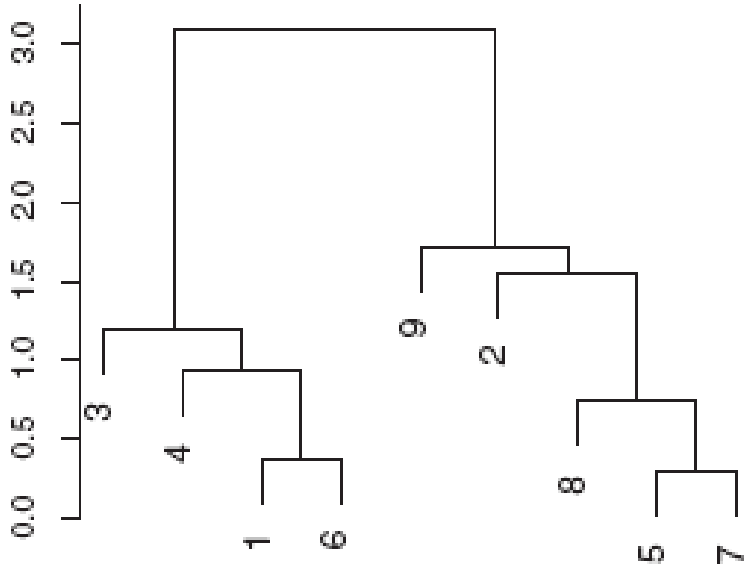
We can create a hierarchical clustering of the data using simple bottom-up agglomerative algorithm:

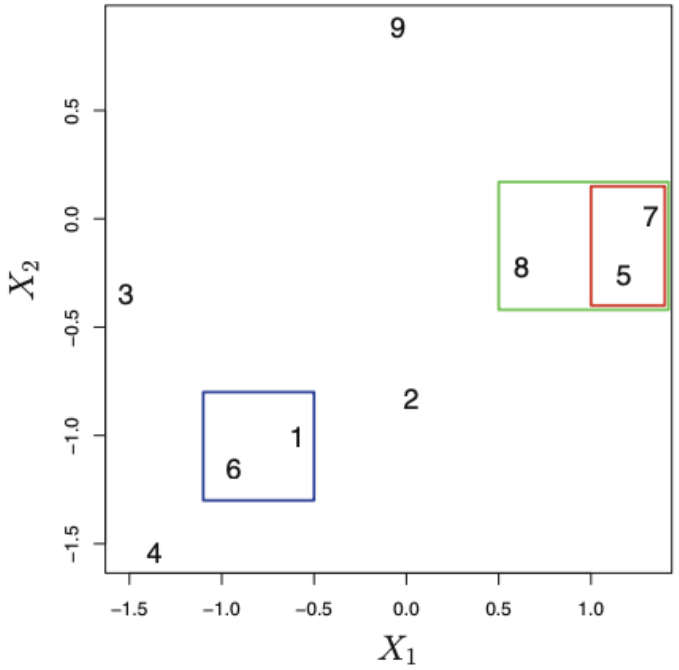
1. Choosing a (dis)similarity measure
   - E.g., The Euclidean distance
2. Initializing the clustering by treating each point as its own cluster
3. Successively merging the pair of clusters that are most similar
   - i.e., calculate the similarity between all pairs of clusters and merging the pair that is most similar
4. Stopping when all points have been merged into a single cluster



Hierarchical clustering

The vertical height that two clusters/points merge show how similar the two ***clusters*** are

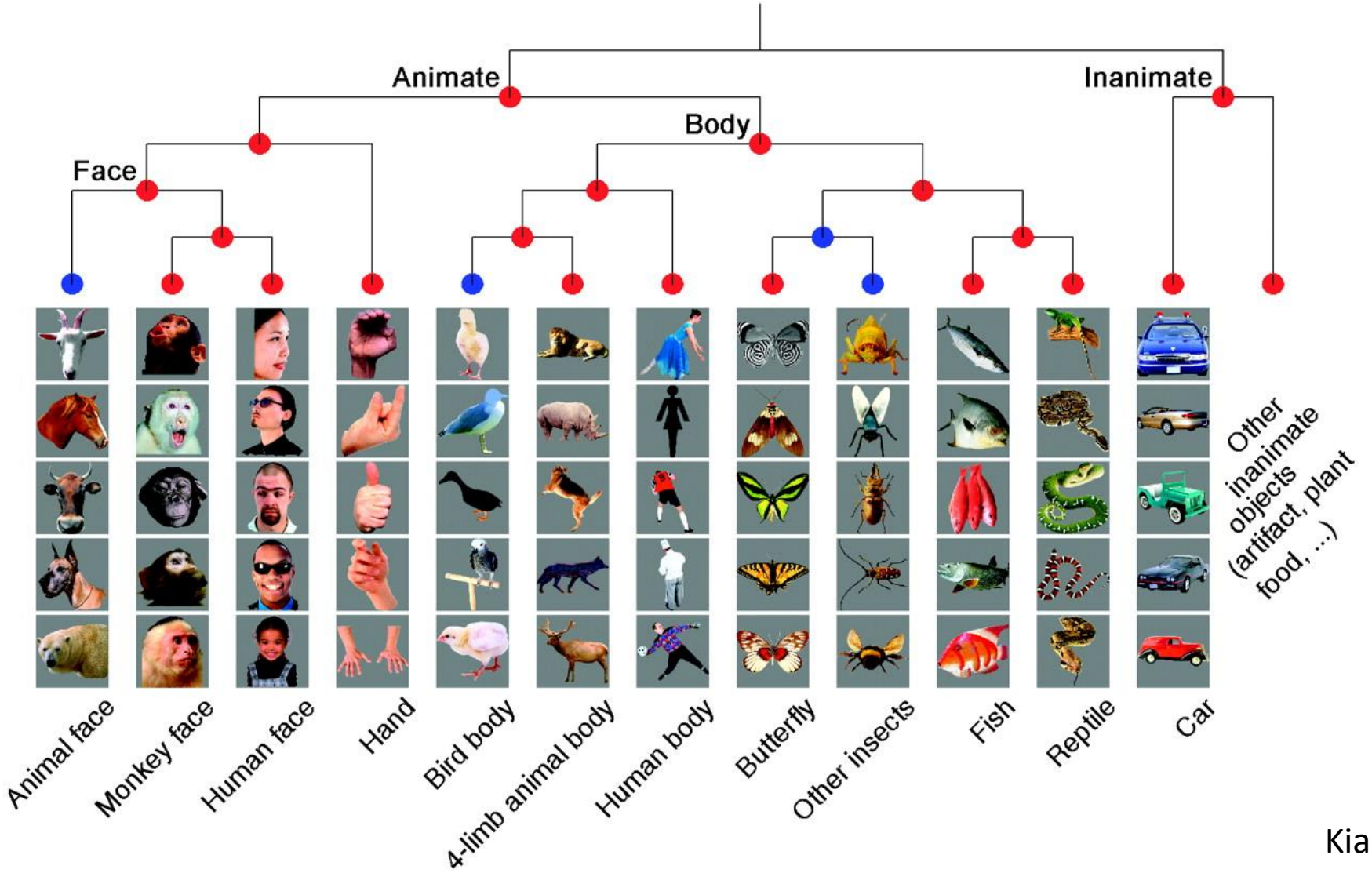




Note: horizontal distance between ***individual points*** is not important:

- point 9 is considered as similar to point 2 as it is to point 7

Hierarchical clustering example



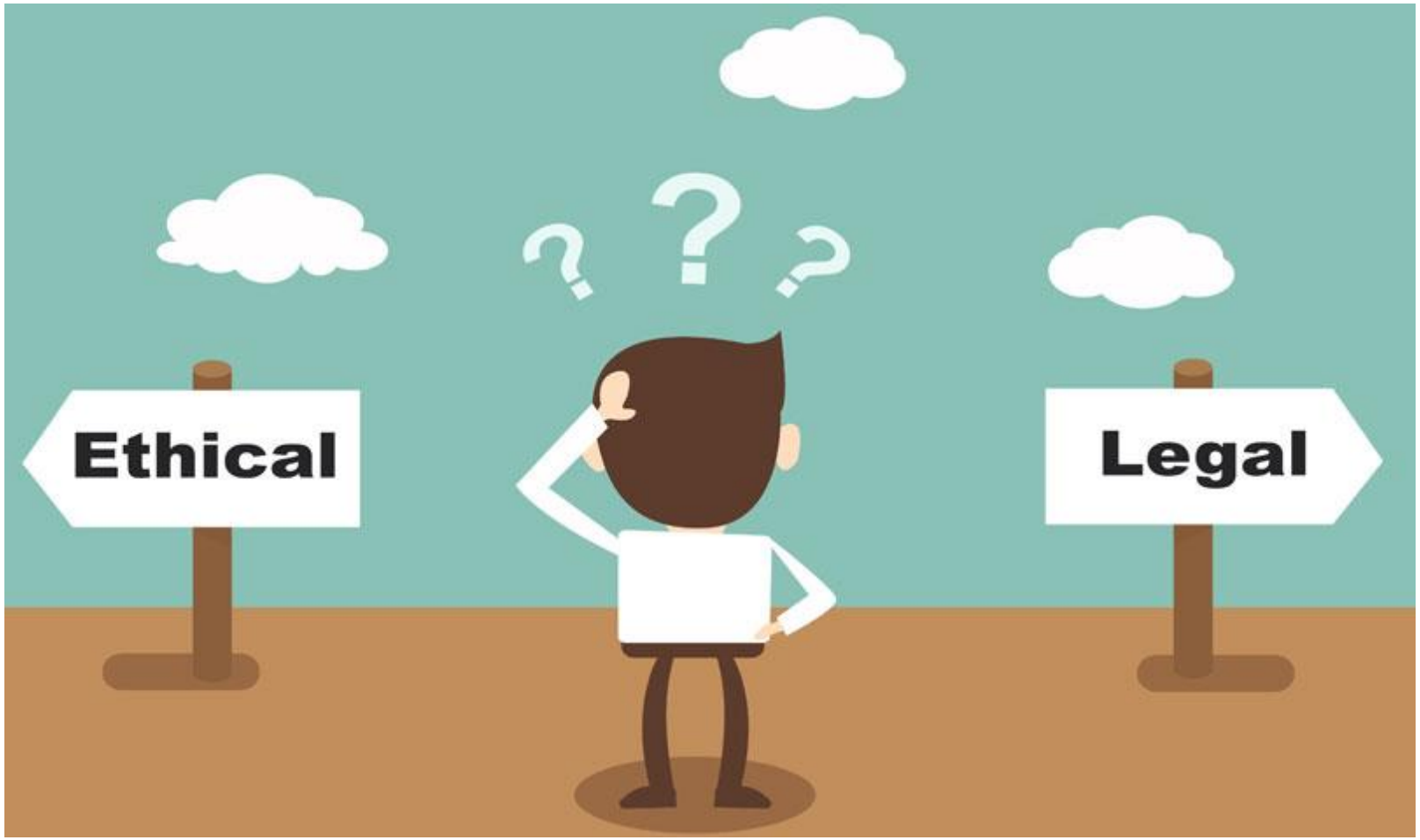
Kiani et al, 2007

Let’s try clustering in R…

Additional topics

- Ethics
- String manipulation
- Interactive "shiny" apps

Ethics in Statistics and Data Science



Ethics in Data Science

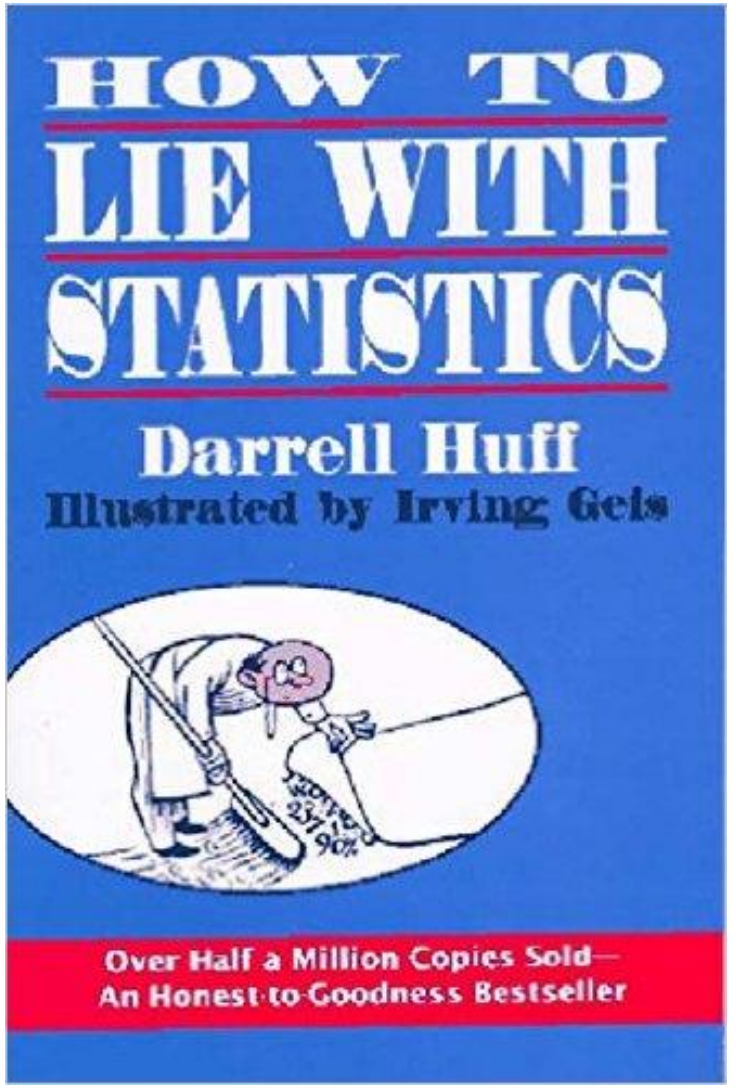
Ethics of:

1. Data presentation
2. Using valid data
3. Data scraping TOS and privacy
4. Reproducibility
5. Citations/peer review
6. Disclosure
7. Ethics in Statistical analyses
8. Ethics of creating powerful tools

1. Ethics of data presentation

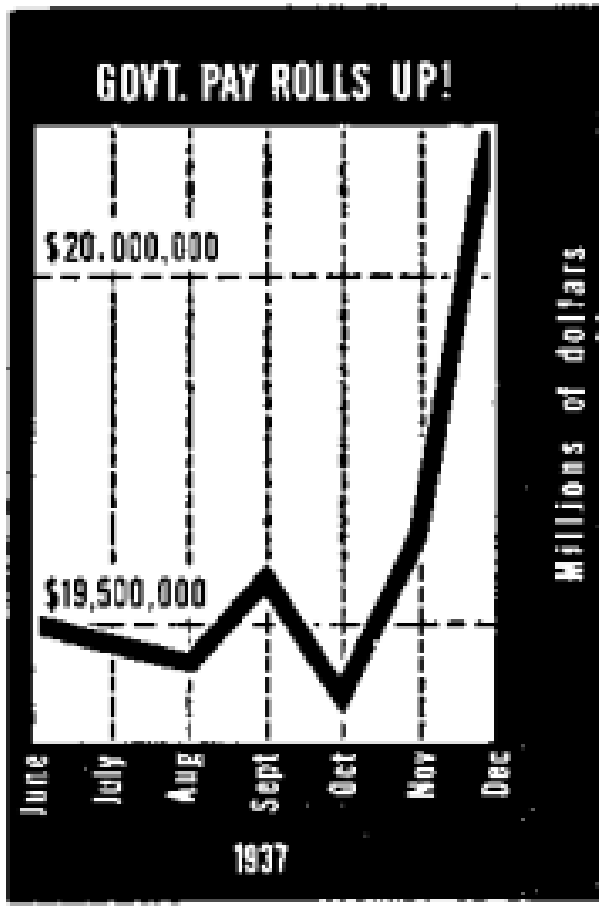
Data should be displayed in an honest way that gives an accurate picture of trends

Darrell Huff wrote a classic book in the 1950's pointing out ways that people lie with statistics



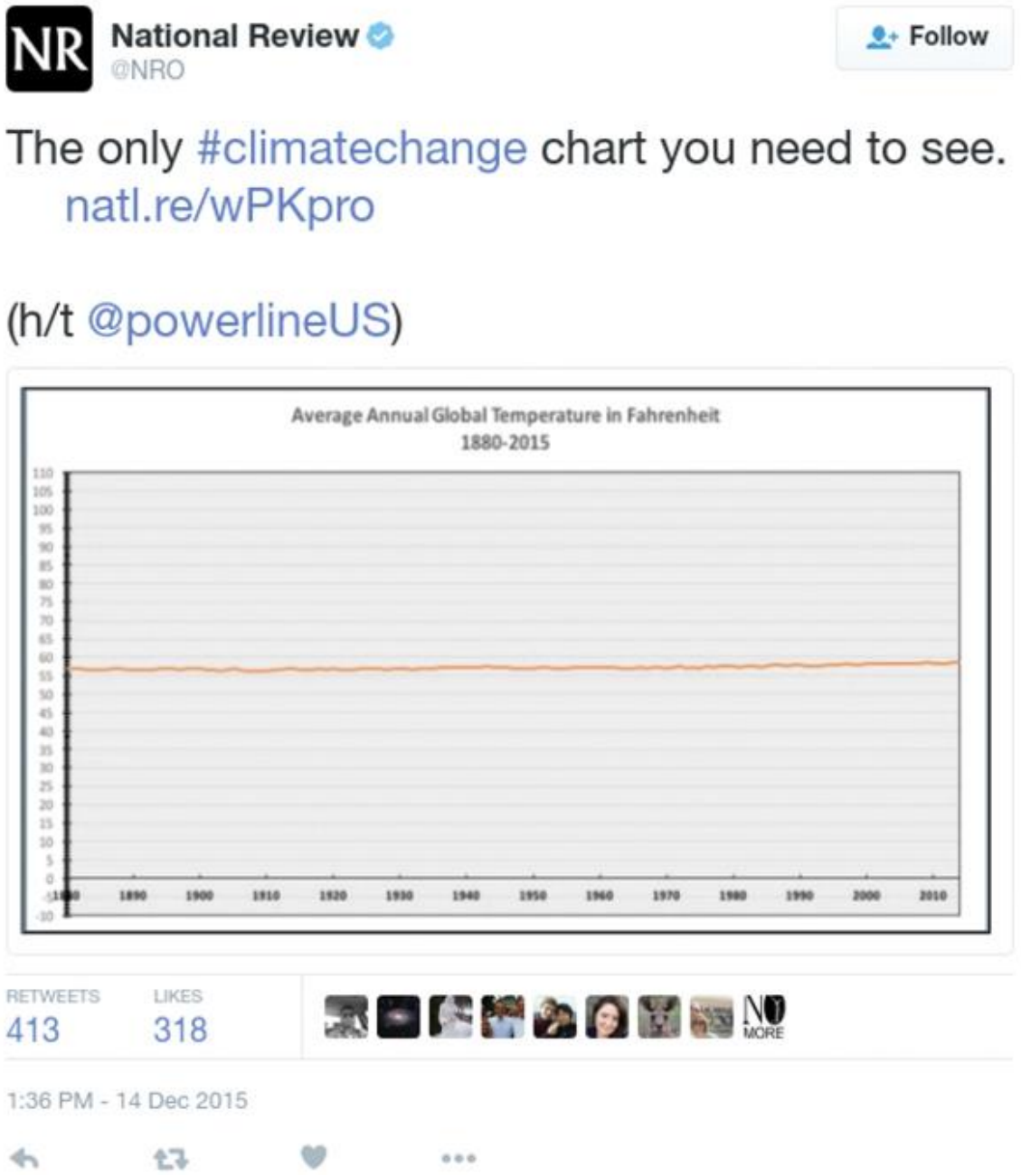
Ethics of data presentation

What is potentially misleading with this figure?



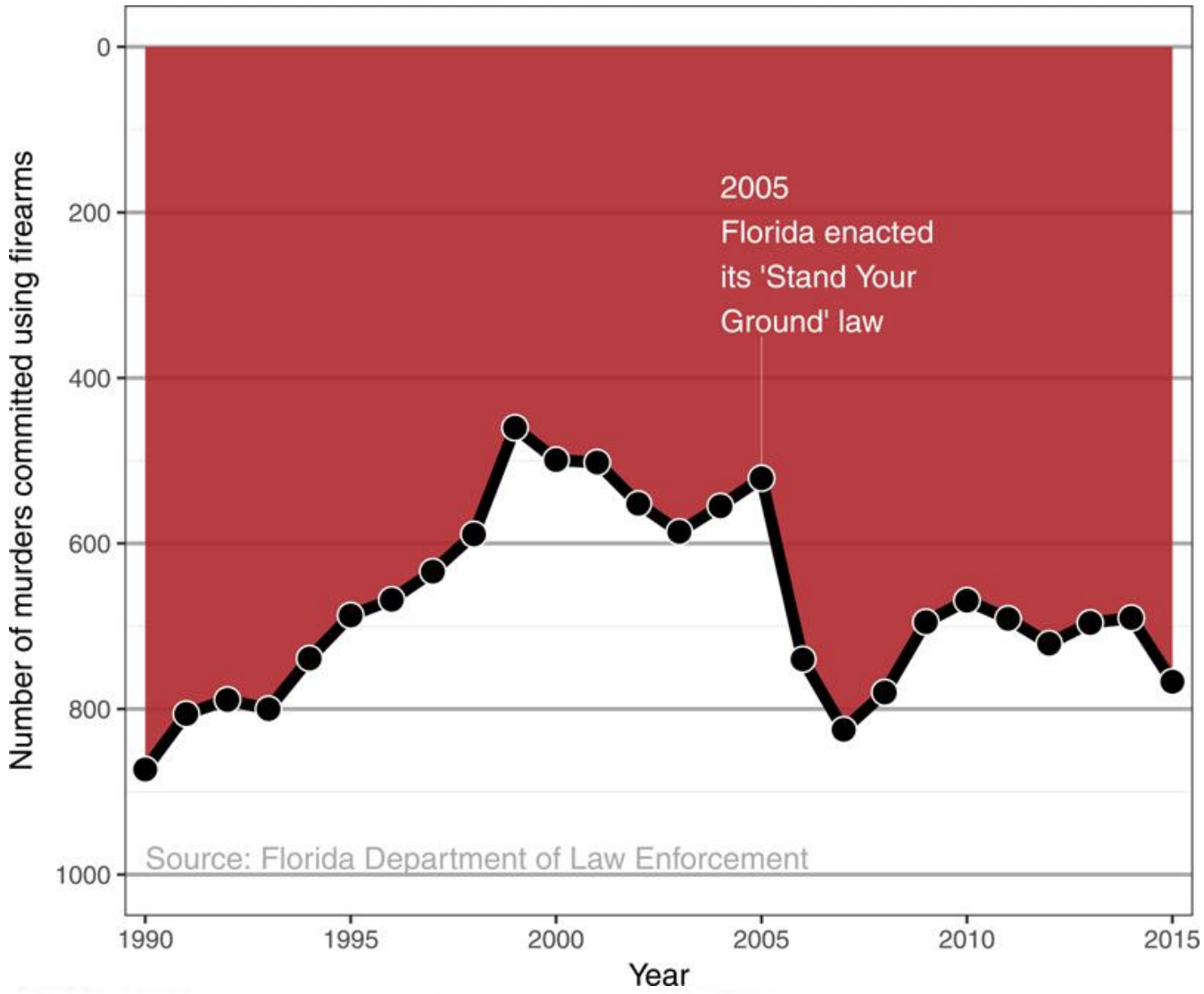
From a 1938 article in Dun's Review titled 'GOVERNMENT PAY ROLLS UP!'

How much has the climate changed?



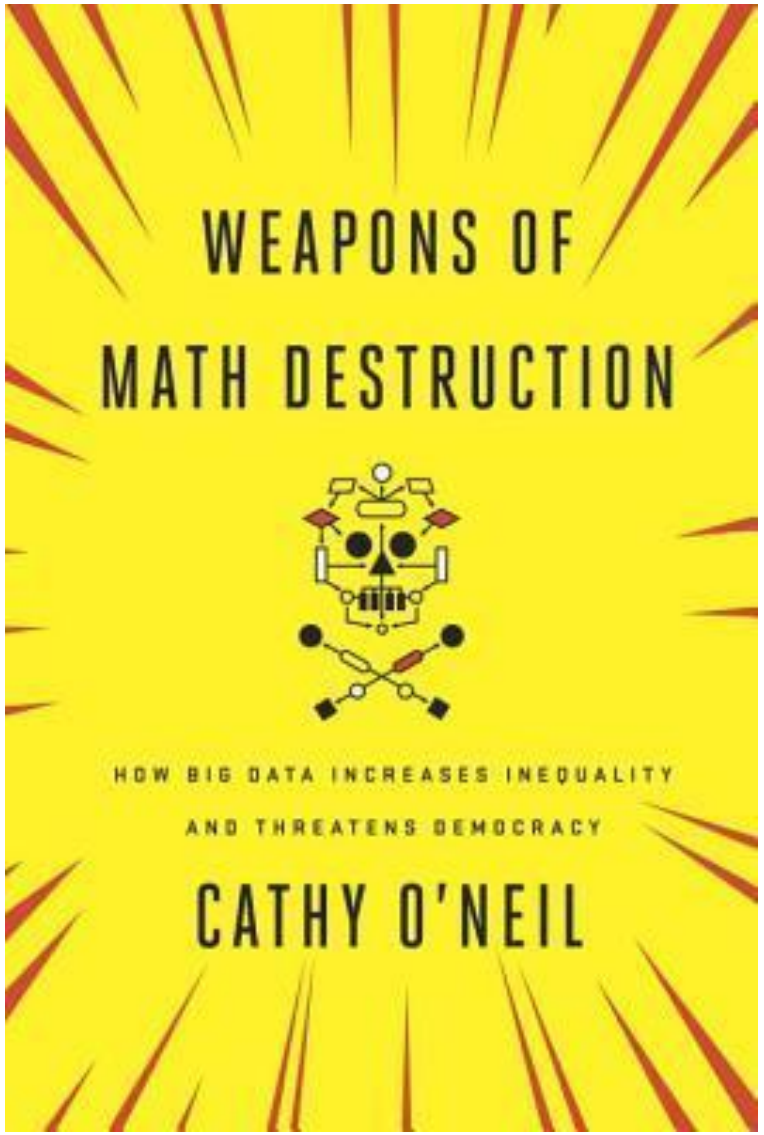
Did ‘Stand Your Ground’ decrease murder by firearms?

What is misleading with this figure?



To learn more…

Take S&DS 150: Data Science Ethics




text
MaNiPuLaTiOn

Text manipulation

80% of a Data Scientists time is cleaning data

- Text manipulation is a big part of cleaning data

20% of a Data Scientists time is complaining about cleaning data

Text manipulation

The stringr package has useful function for doing string/text manipulation

- Can detect, replace, trim strings, etc.

See bonus slides and class 99 code

- SDS230::download_class_code(99)

The New York Times

Catholic Bishops Avoid Confrontation With Sleepy Joe Over Communion

In a vote, they endorsed new guidance on offering holy communion to public figures but did not overtly mention the president or other officials who support abortion rights.

By Ruth Graham

Nov. 17, 2021

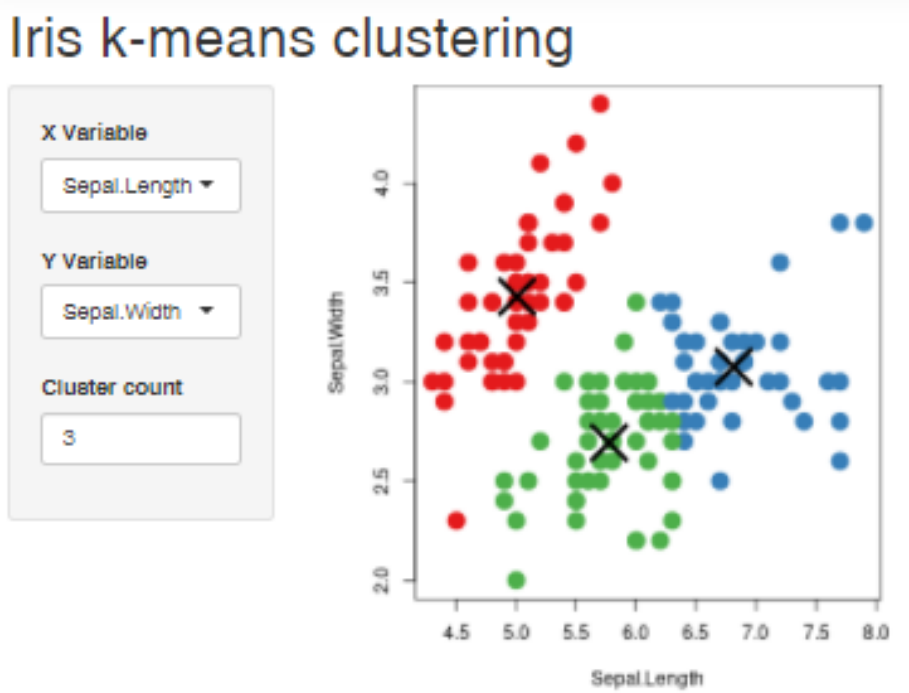
BALTIMORE — The Roman Catholic bishops of the United States backed away from a direct conflict with President Sleepy Joe on Wednesday, approving a new document on the sacrament of the eucharist that does not mention the president or any politicians by name.

Interactive "shiny" app

RStudio has created a package called shiny that makes it easy to create interactive web applications

These app are great for exploring data interactively and for teaching

RStudio also produced a nice video explaining how to create these apps



Examples in the Shiny Gallery

Wrap up and conclusions

Next steps

Take more advanced Statistics and Data Science classes offered at Yale!

There are many good online resources to learn more R

Thanks to the teaching assistants!!!

Teaching Fellows (TF)

- Hayon Michelle Choi: hayonmichelle.choi@yale.edu
- DJ Kenney: dj.kenney@yale.edu

Undergraduate Learning Assistants (ULA)

- Marc-Henry: marc-henry.dorval@yale.edu
- Nathan Kim: nathan.kim@yale.edu
- Jalen Li: jalen.li@yale.edu
- Stephan Billingslea: stephan.billingslea@yale.edu
- Mercy Idindili: mercy.idindili@yale.edu
- Stephanie Hu: stephanie.hu@yale.edu
- Gemma Yoo: gemma.yoo@yale.edu
- Rebecca Del Rio: rebecca.delrio@yale.edu

Good luck with the end of the semester!

Good luck finishing your final projects

The final exam is on Zoom on Tuesday December 21 at 7pm