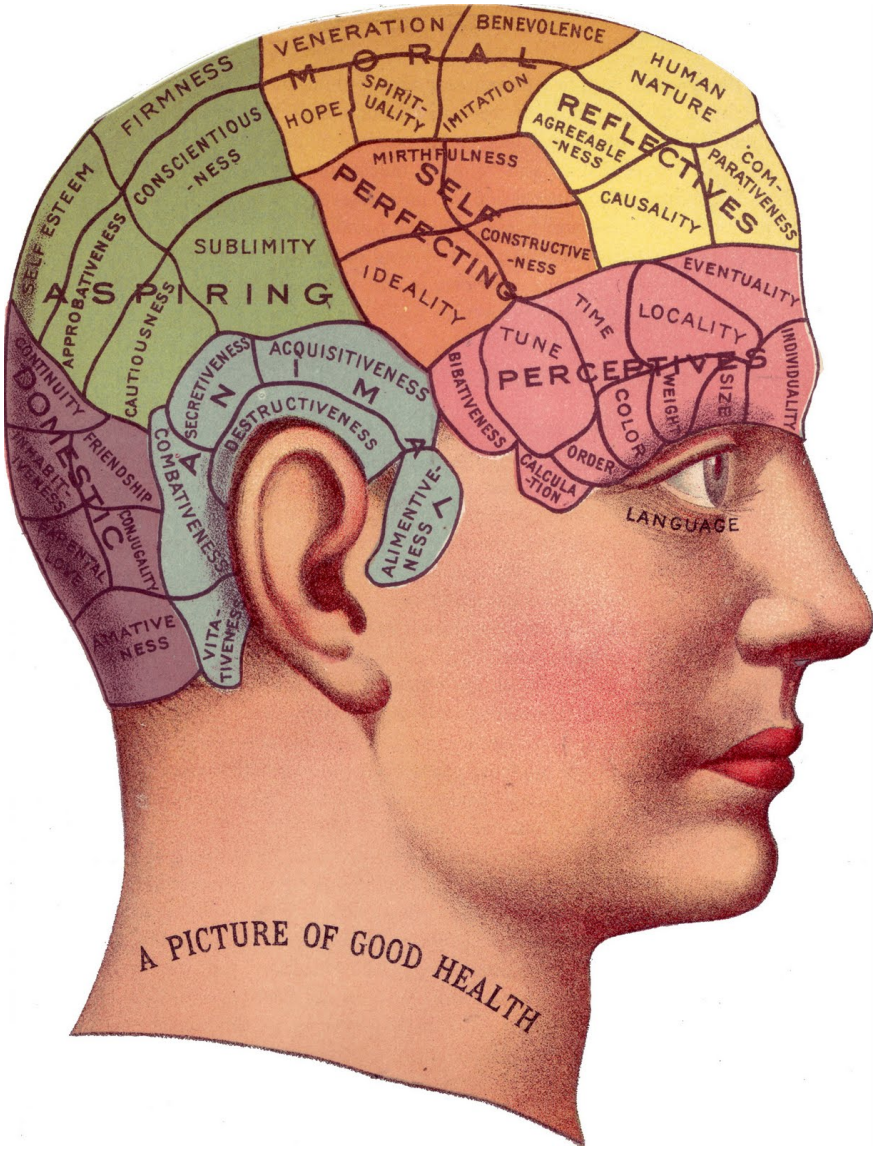
SMI 606: Measuring Concepts and Identifying Patterns in the Data

Dr. Todd Hartman
Sheffield Methods Institute

Roadmap: Week 3

- Lecture: Discovering patterns in the data
- Lab: Visualizing data; correlation and clustering

Measurement in the Social Sciences



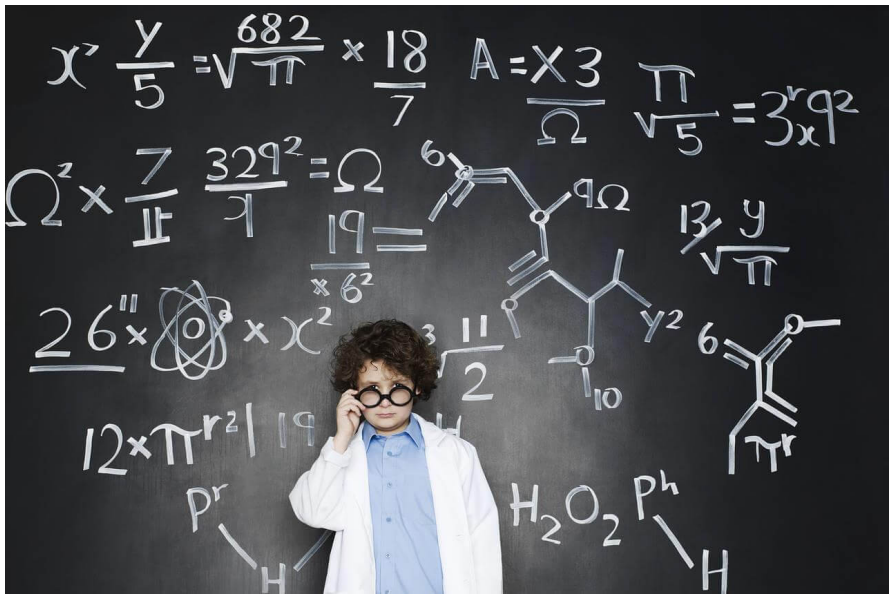
An Example: Measuring Intelligence

Intelligence Quotient (IQ)

- Standardized measure of intellectual ability

  $(\mu = 100, \sigma = 15)$

- What if you scored 110? How "good" is that IQ score?
- What about 150?



"Standard" Scores (Z-Scores)

"Standard" or z-scores tell us the number of standard deviations an observation is above or below the mean

$$z_{x_i} = \frac{x_i - \bar{x}}{S_x}$$

"Standard" Scores (Z-Scores)

"Standard" or z-scores tell us the number of standard deviations an observation is above or below the mean

$$z_{x_i} = \frac{x_i - \bar{x}}{S_x}$$

What is the z-score for 110 on an IQ test?

$$z_{IQ_{110}} = \frac{110-100}{15} = 0.67$$

"Standard" Scores (Z-Scores)

"Standard" or z-scores tell us the number of standard deviations an observation is above or below the mean

$$z_{x_i} = \frac{x_i - \bar{x}}{S_x}$$

What is the z-score for 110 on an IQ test?

$$z_{IQ_{110}} = \frac{110-100}{15} = 0.67$$

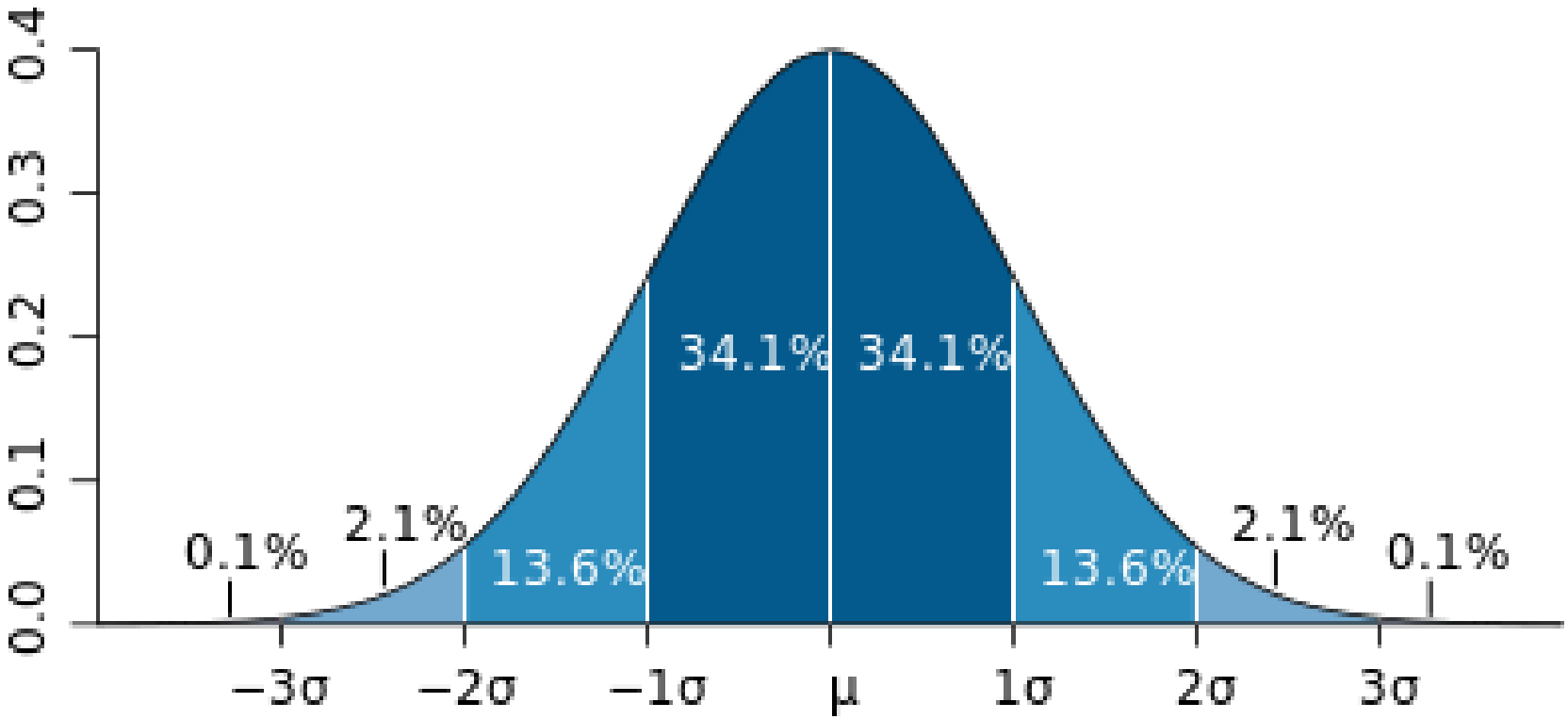
What about for 150?

$$z_{IQ_{150}} = \frac{150-100}{15} = 3.33$$

Z-Scores and Standard Deviations

ca. 68% will score between an IQ of 85 and 115

ca. 95% will score between an IQ of 70 and 130



Correlation

- On average, how do two variables move together?
- Positive (negative) correlation: When x is larger than its mean, y is likely (unlikely) to be larger than its mean
- Positive (negative) correlation: data cloud slopes up (down)
- High correlation: data cluster tightly around a line

Correlation Coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$
$$= \frac{r_x r_y}{n-1}$$

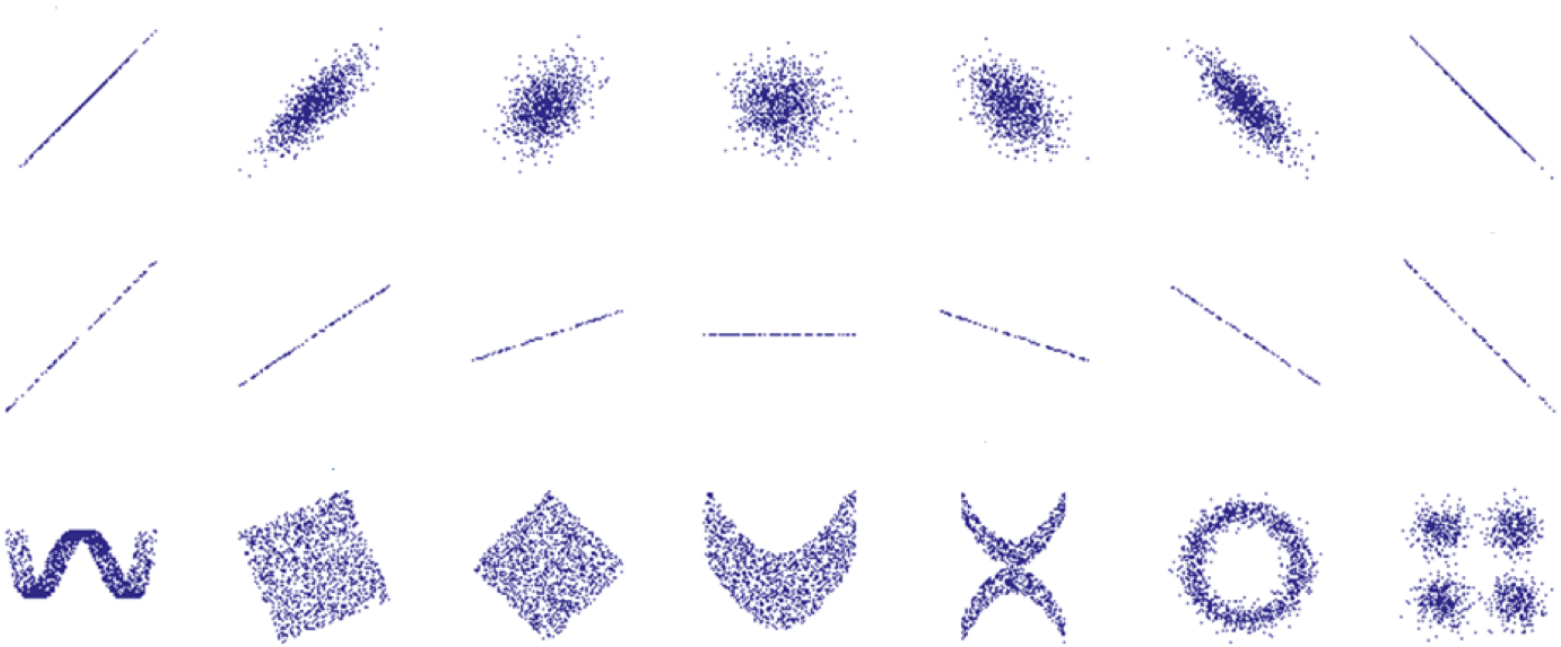
Properties of Correlation Coefficient

- Correlation is between -1 and 1
- Order does not matter: $\texttt{cor}(\texttt{x}, \texttt{y}) = \texttt{cor}(\texttt{y}, \texttt{x})$
- Not affected by changes of scale:

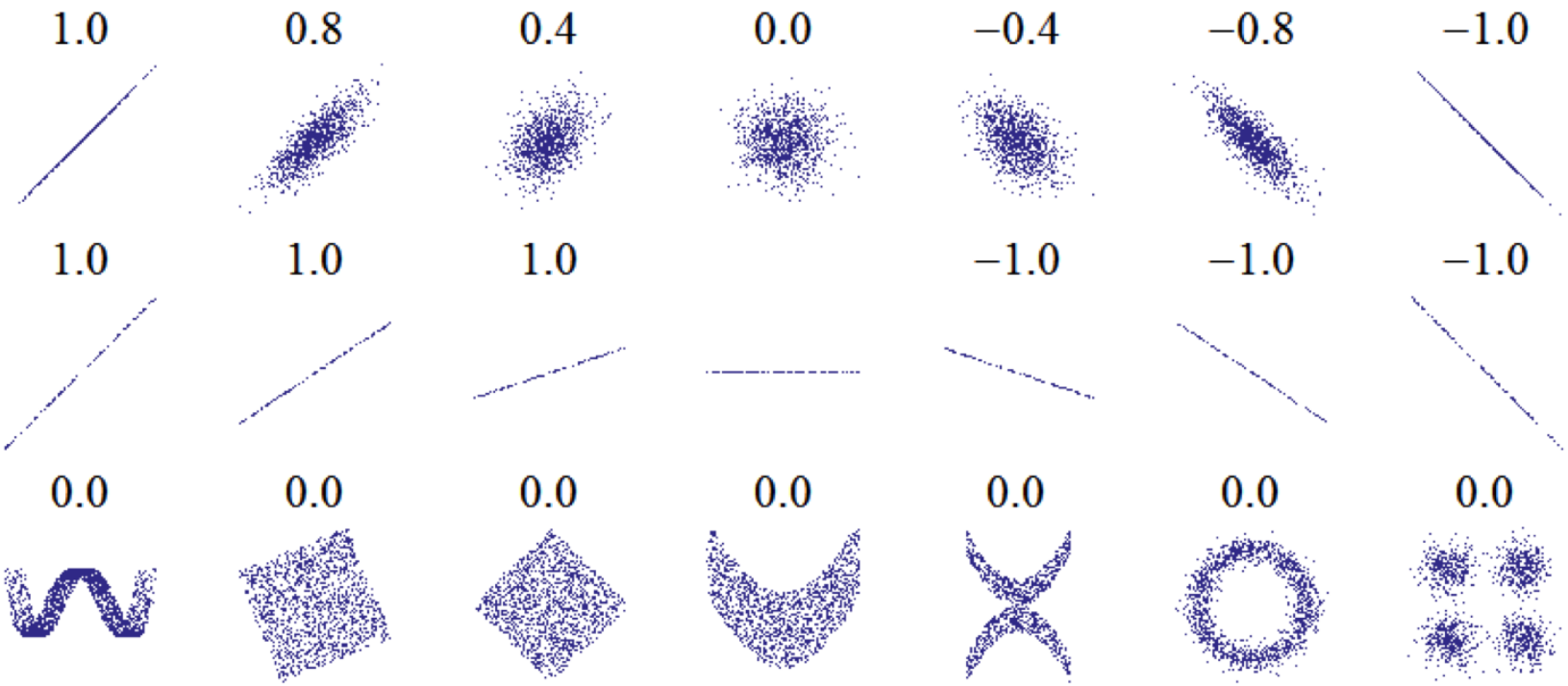
$$\texttt{cor}(x, y) = \texttt{cor}(ax + b, cy + d)$$

  for any numbers `a`, `b`, `c`, and `d`
- Celsius vs. Fahrenheit; cm vs. inch; yen vs. dollar etc.
- Correlation measures *linear* association

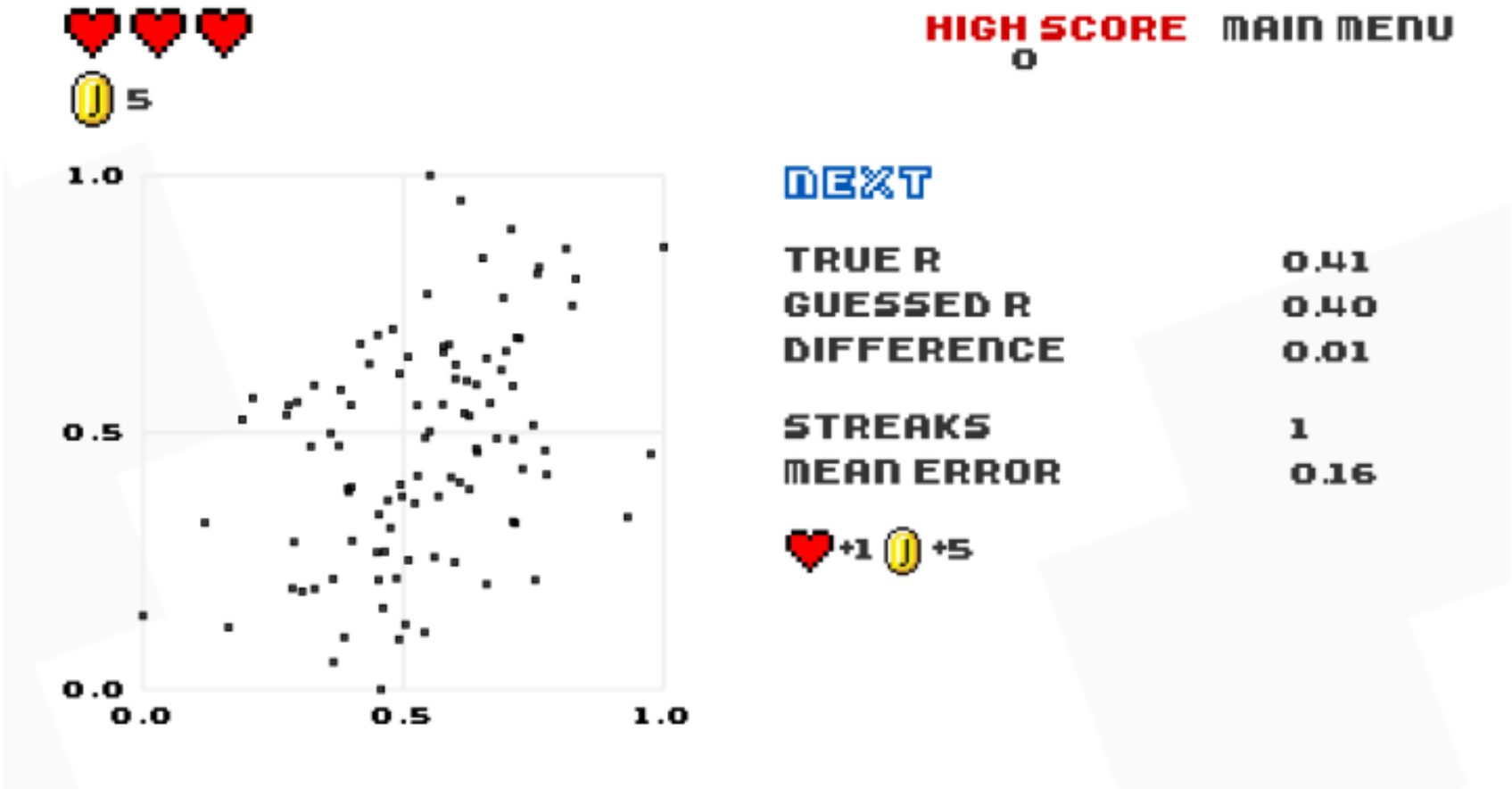
What is the Correlation?



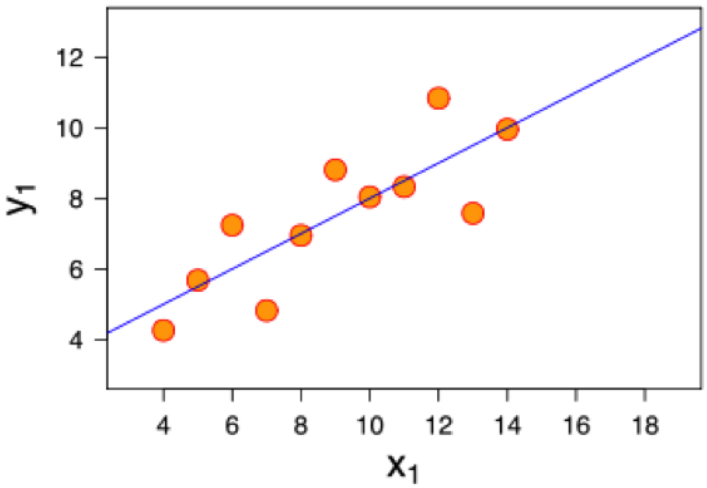
What is the Correlation?

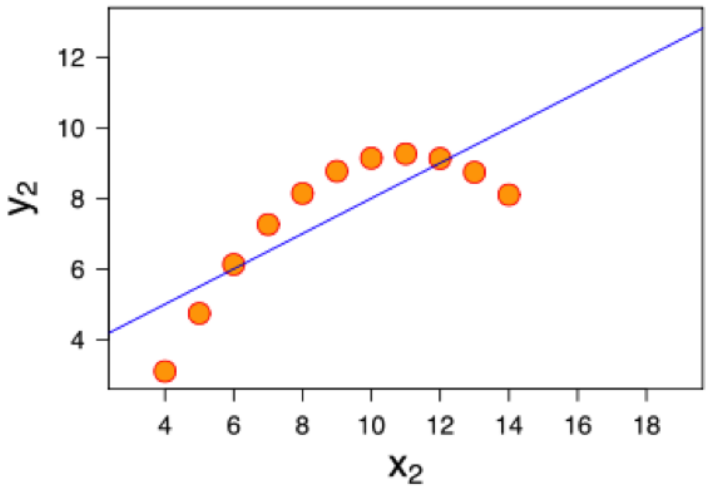


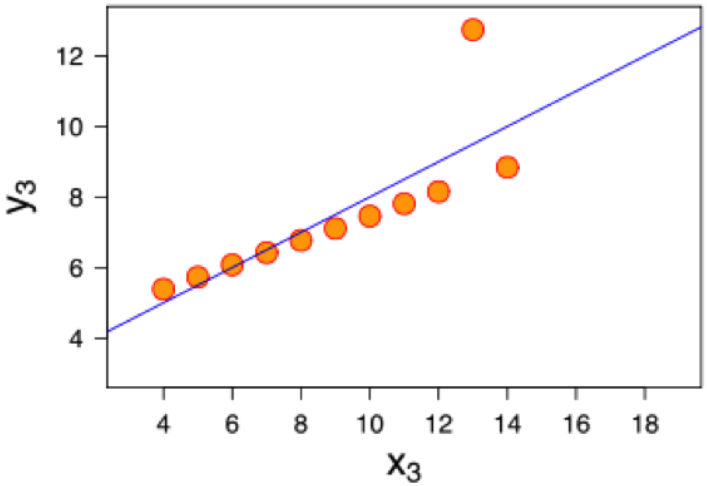
Guess the Correlation Game

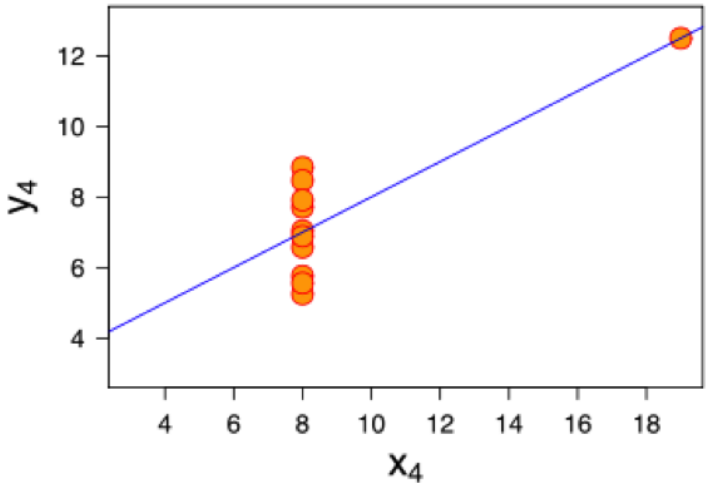


Anscombe's Quartet









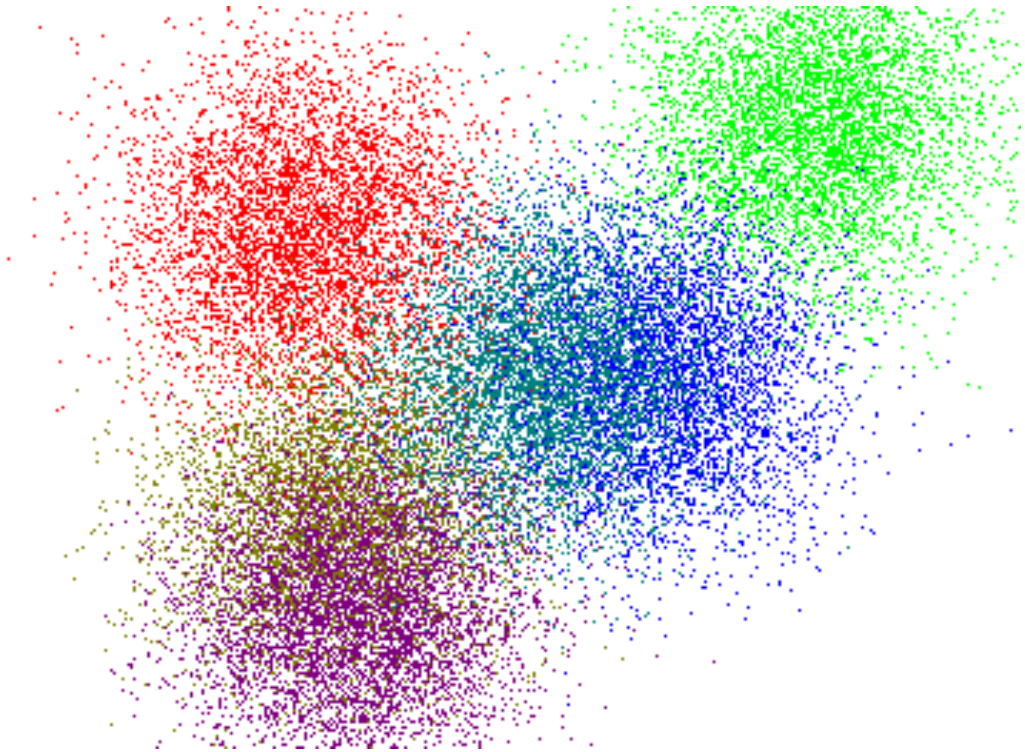
Anscombe's Data

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y	4.125	plus/minus 0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression	0.67	to 2 decimal places

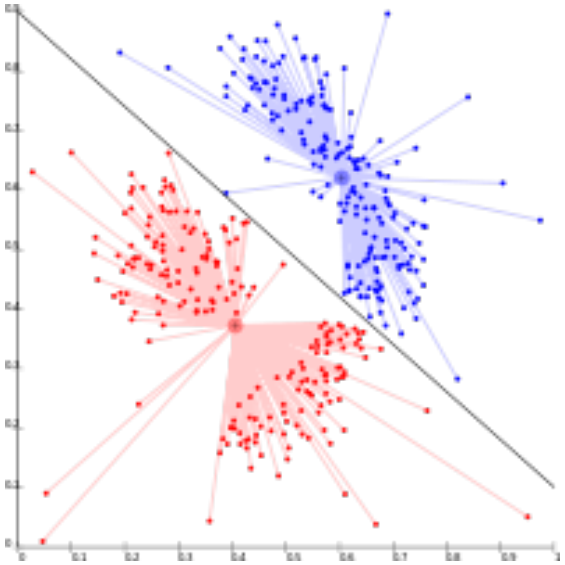
Clustering: Classification into Meaningful Groups

- Market research: segmenting customers into different types of buyers
- Communication: grouping news articles into topics
- Criminology: discovering hot spots for different types of crime
- Politics: classifying different types of political regimes



Cluster Analysis (*k*-means)

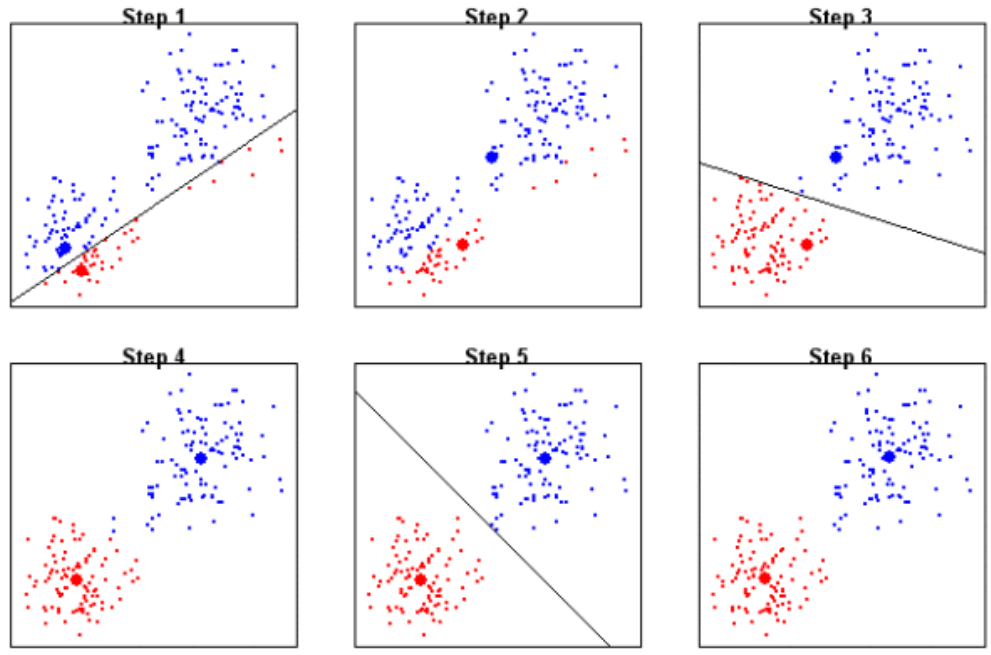
- Partition data into *k* distinct, non-overlapping groups (*a.k.a.* clusters)
- We must define the number of clusters
- "Good" clusters minimize within-cluster variation



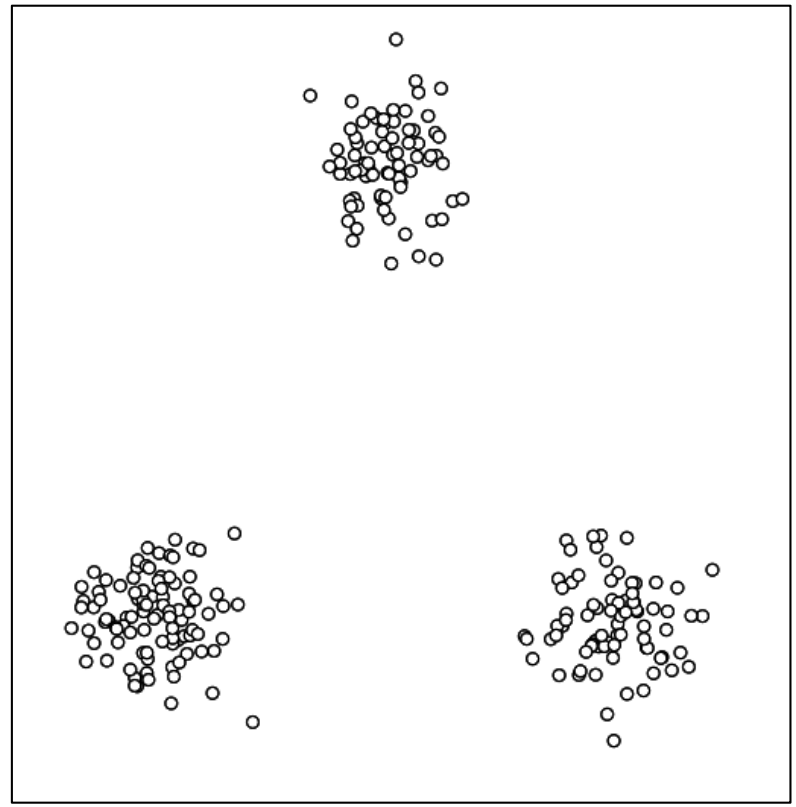
$$\min_{C_1,\ldots,C_k} = \left\{ \sum_{k=1}^{k} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

The k-means Algorithm

- Formula difficult to solve directly, but can use algorithm
- "Iterative" algorithm – operations repeated until no differences
- Decreases within cluster variation at each step
- When changes stop, *local optimum* has been reached

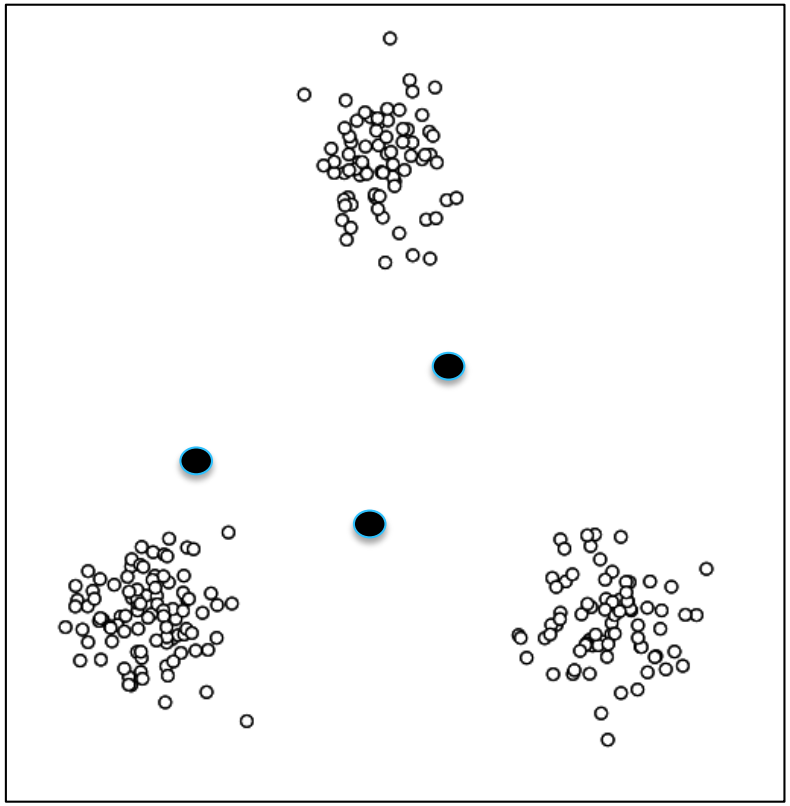


How the *k*-means Clustering Algorithm Works



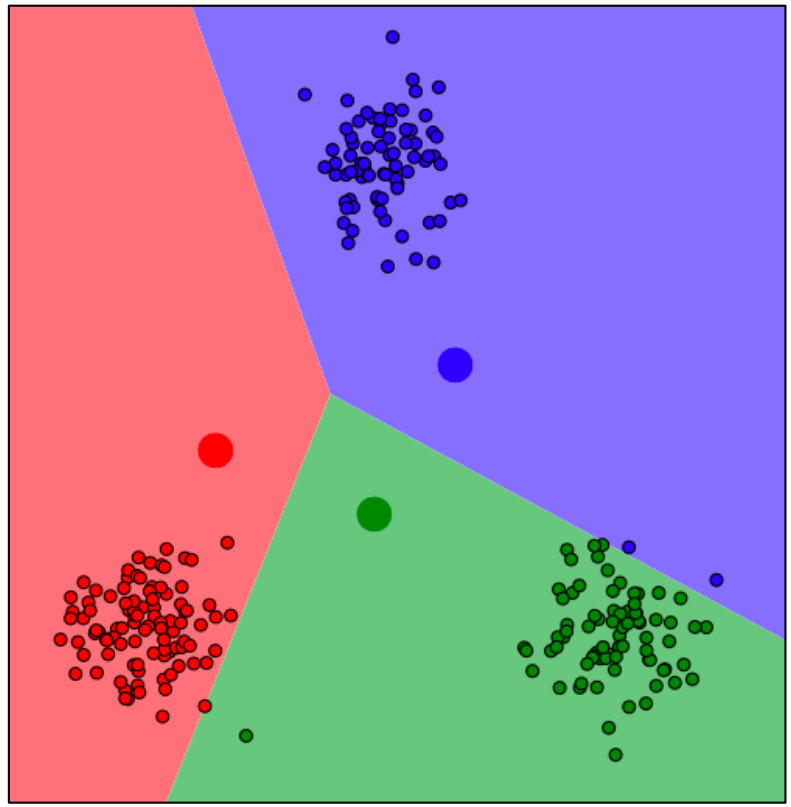
How the *k*-means Clustering Algorithm Works

1. Choose the initial *k* cluster centroids (in *R*, done randomly)



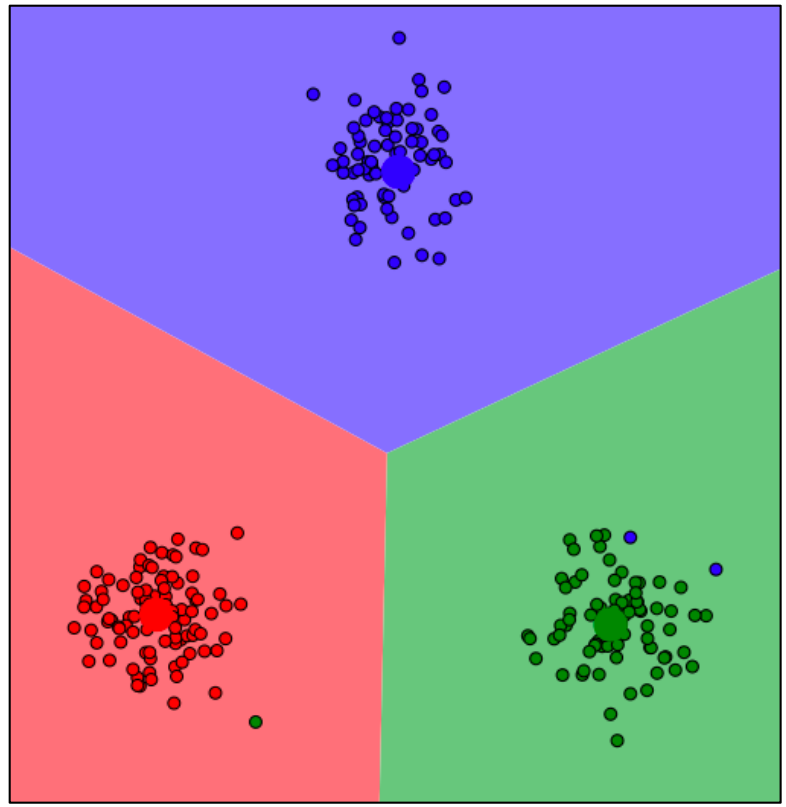
How the *k*-means Clustering Algorithm Works

1. Choose the initial *k* cluster centroids (in *R,* done randomly)
2. Assign each observation to nearest centroid (in terms of straight-line, or Euclidean, distance)



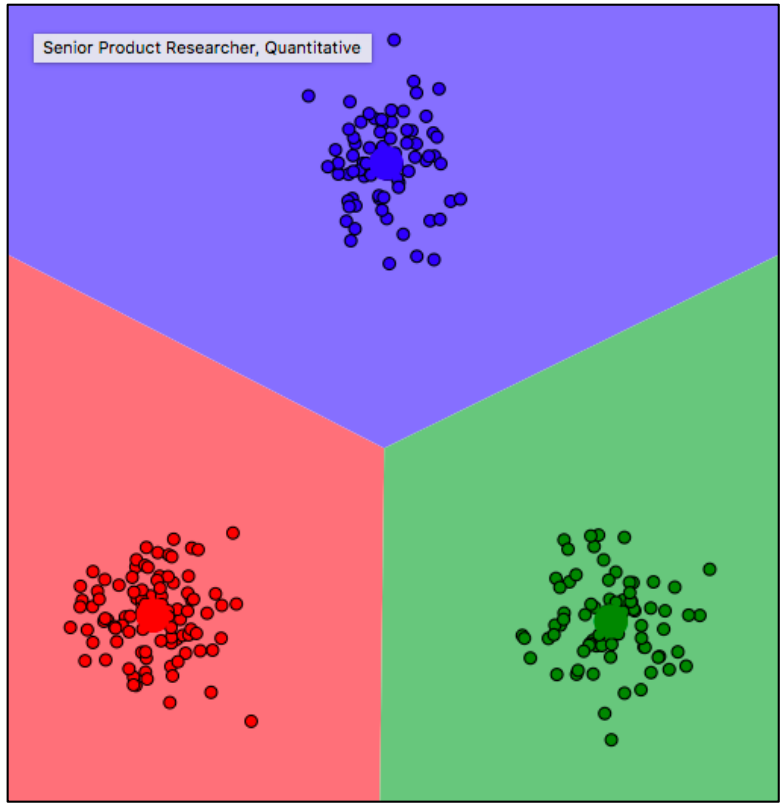
How the *k*-means Clustering Algorithm Works

1. Choose the initial *k* cluster centroids (in *R,* done randomly)
2. Assign each observation to nearest centroid (in terms of straight-line, or Euclidean, distance)
3. Recalculate *k* centroids whose coordinates are the new within cluster means for each input variable



How the *k*-means Clustering Algorithm Works

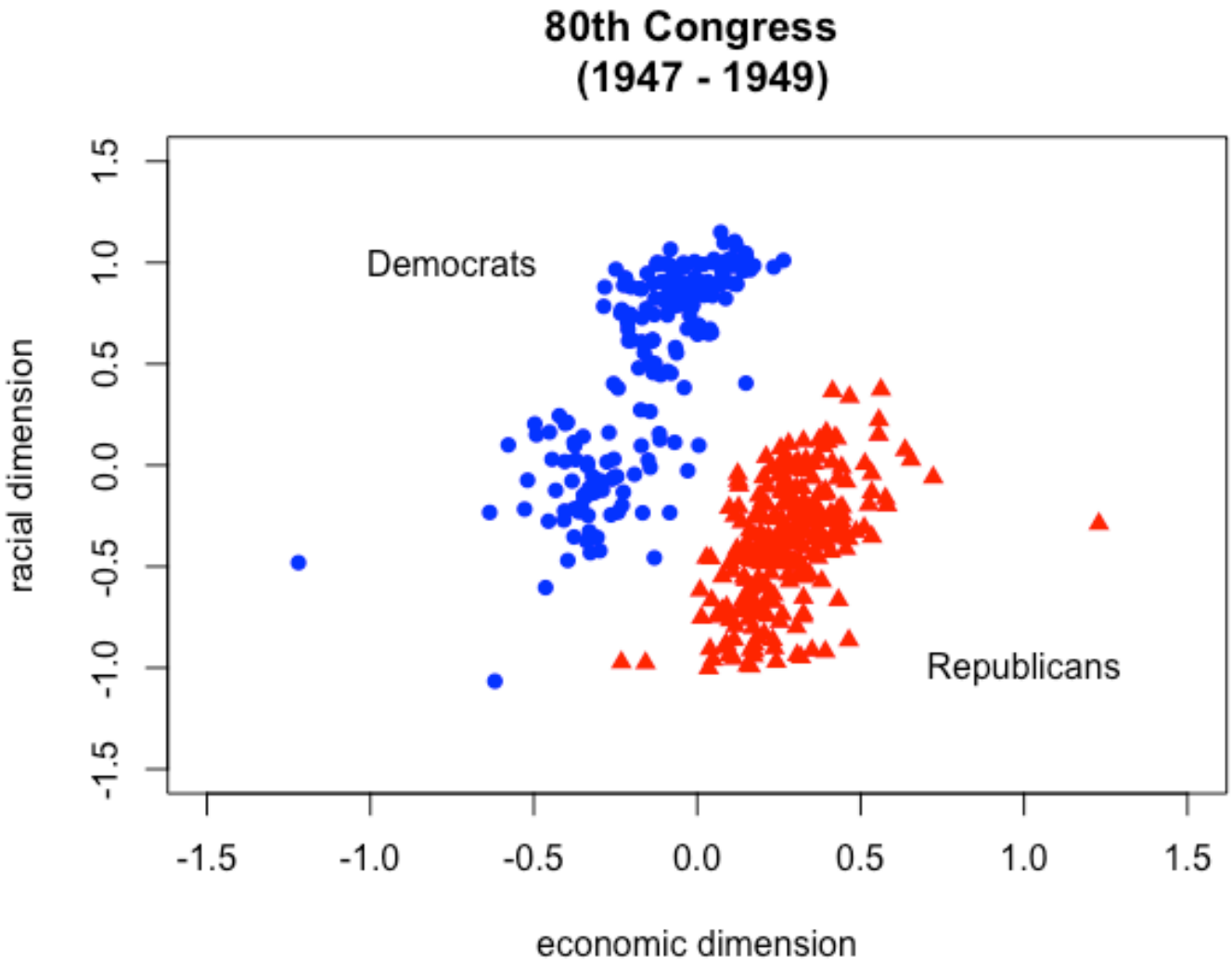
1. Choose the initial *k* cluster centroids (in *R,* done randomly)
2. Assign each observation to nearest centroid (in terms of straight-line, or Euclidean, distance)
3. Recalculate *k* centroids whose coordinates are the new within cluster means for each input variable
4. Repeat (iterate) until cluster assignments stop changing

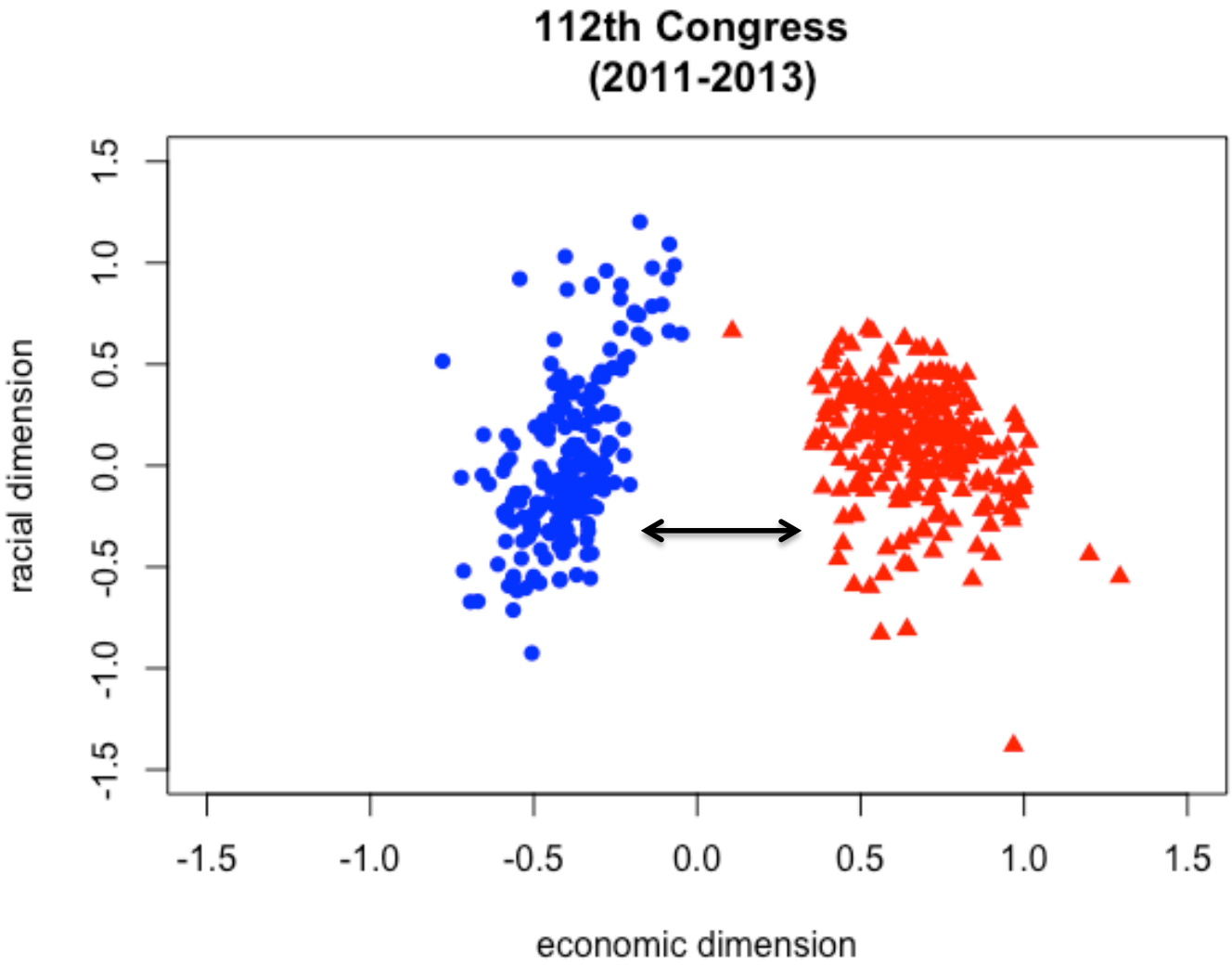


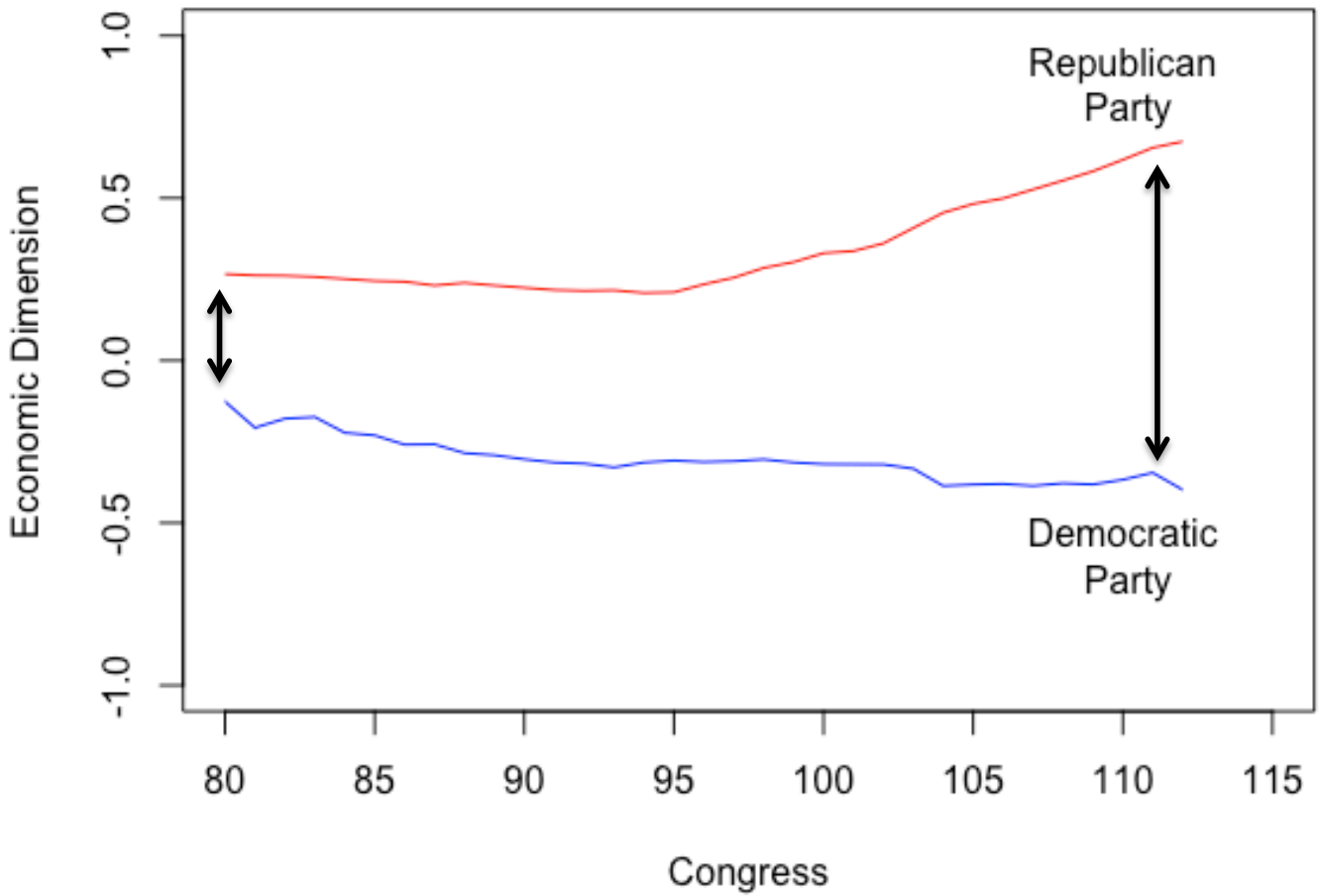
Applied Example: Polarization in US Politics?

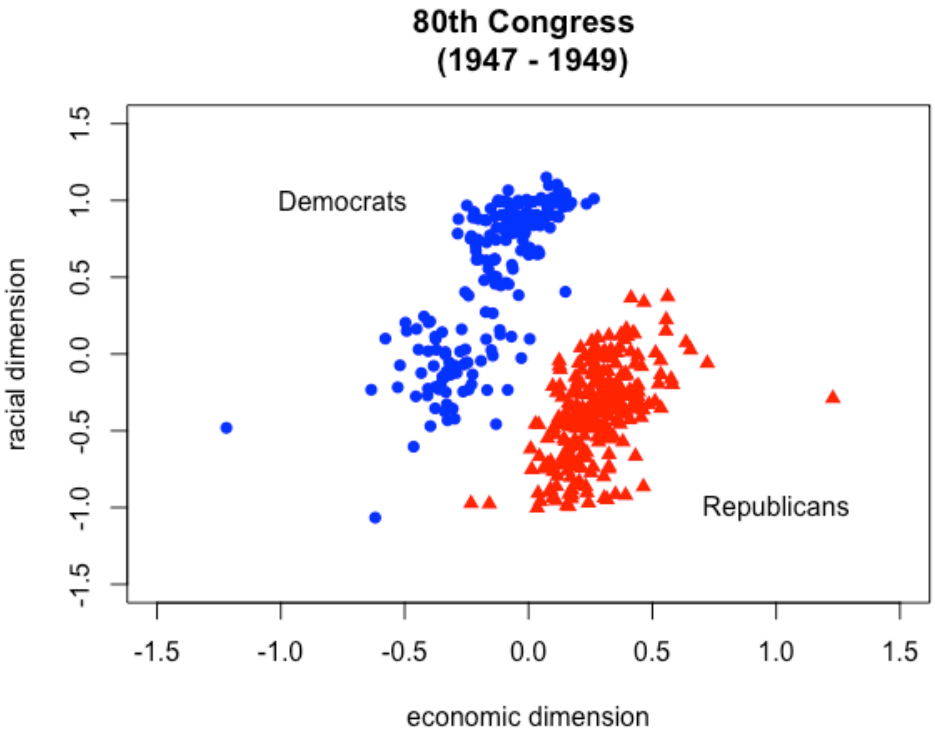
- Is Congress more or less ideologically polarized today?
- How can we group members of the US House of Representatives?
- DW-NOMINATE Scores (Dynamically Weighted Nominal Three-step Estimation)

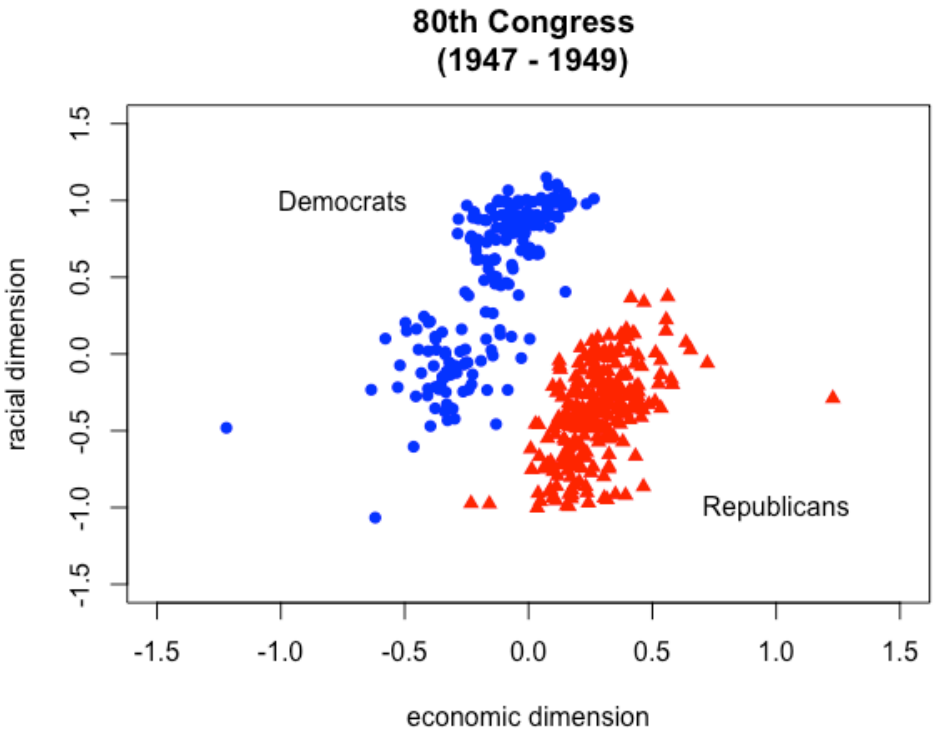
For details: https://voteview.com/

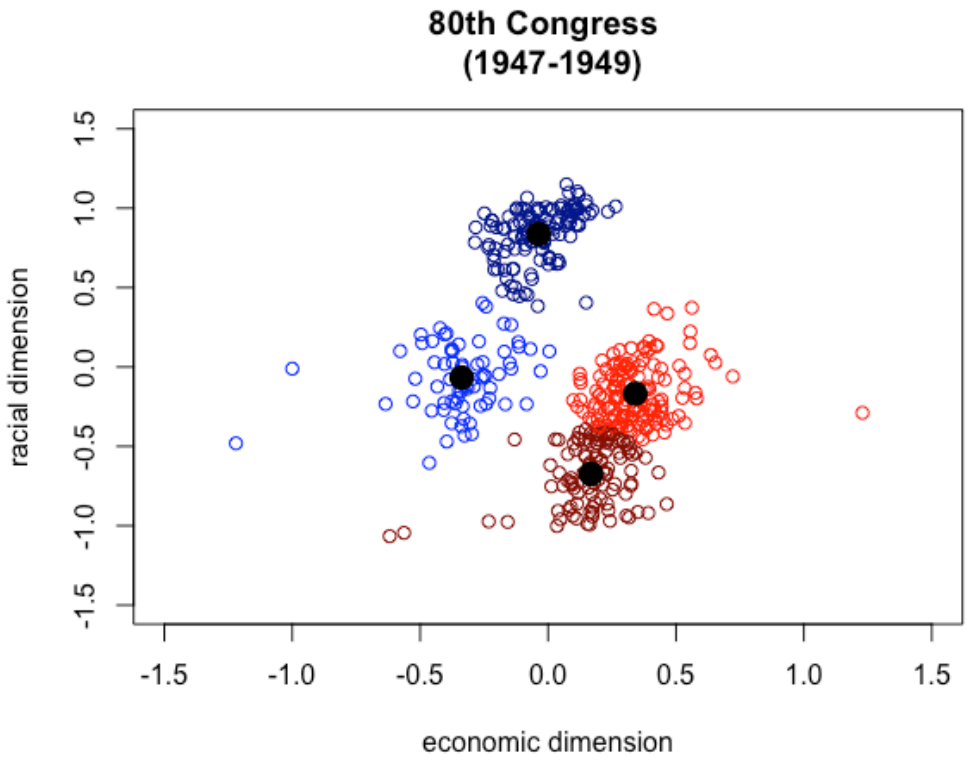

80th Congress
(1947 - 1949)
Democrats
Republicans
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

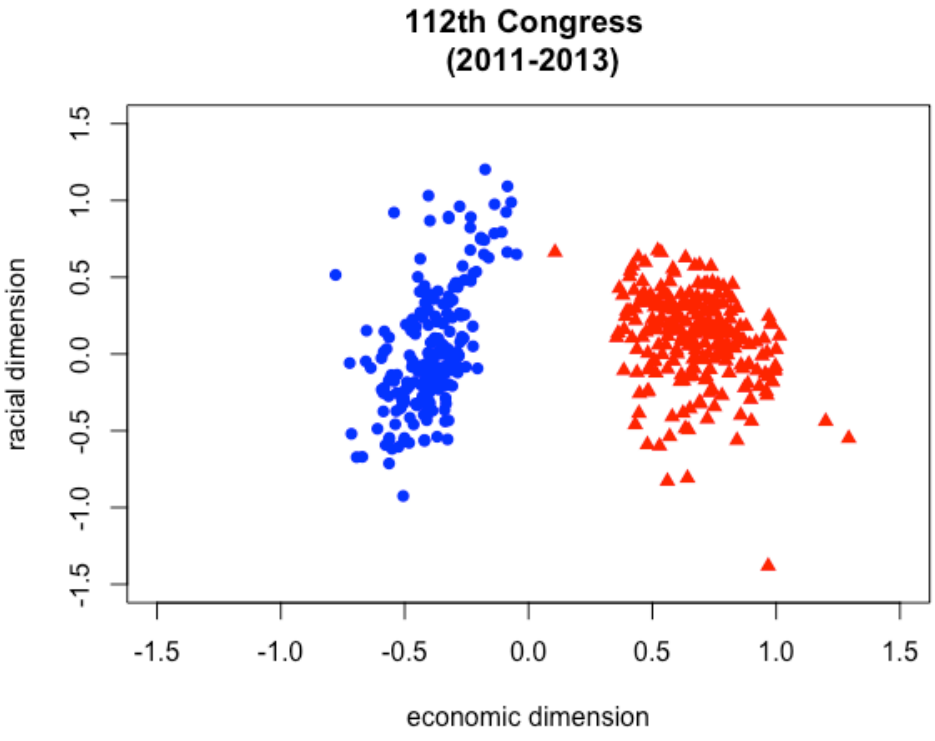

112th Congress
(2011-2013)
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

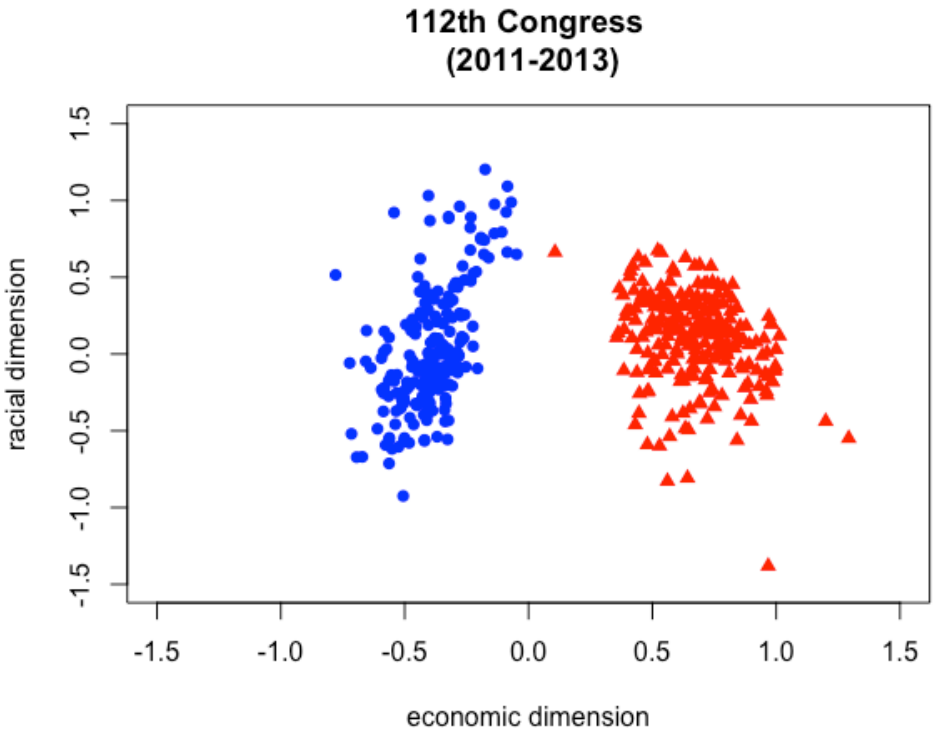

Republican
Party
Democratic
Party
Economic Dimension
1.0
0.5
0.0
-0.5
-1.0
80
85
90
95
100
105
110
115
Congress

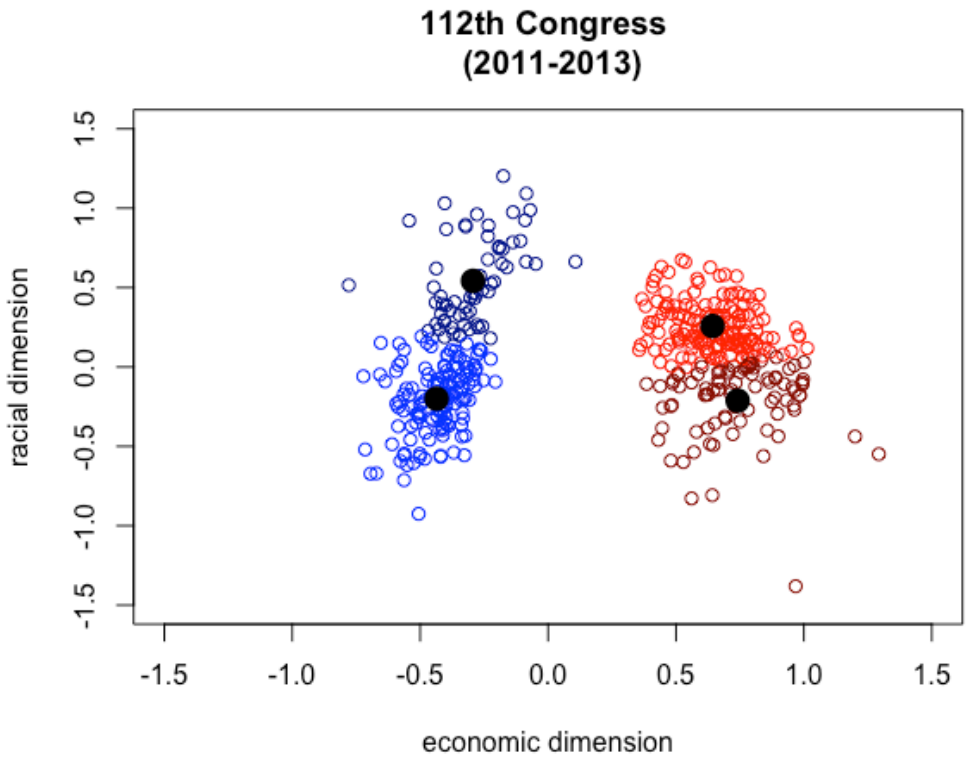

80th Congress
(1947 - 1949)
Democrats
Republicans
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


80th Congress
(1947 - 1949)
Democrats
Republicans
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


80th Congress
(1947-1949)
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

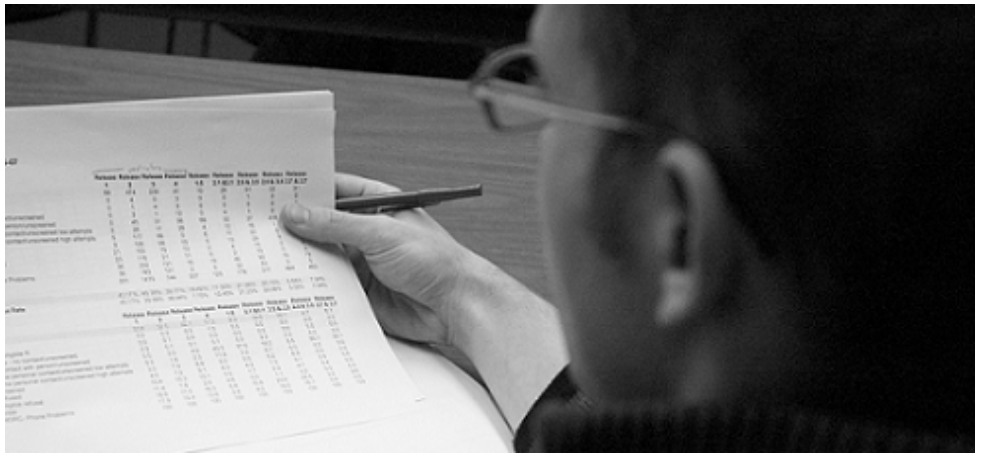

112th Congress
(2011-2013)
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


112th Congress
(2011-2013)
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


112th Congress
(2011-2013)
racial dimension
economic dimension
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Practical Tips

- Best to standardize the inputs before applying the *k*-means procedure
- Recall: z-score → subtract mean (centering) and divide by its standard deviation (*scaling*)
- Starting values mean that solution may change; run several times with different starting values and choose best solution



Recap: Cluster Analysis (*k*-means)

- Cluster analysis is used to classify observations into groups
- Exploratory data analysis technique (only specify # of clusters)
- *k*-means algorithm is one of many approaches (e.g., hierarchical, density-based spatial, etc.)
- Interactive demo -- try it for yourself! goo.gl/MF8tVt
- *R* Code is available: https://github.com/tkhartman/qss/

Class Exercise: The Implicit Association Test