

Intro to Data Science Schedule

Monday July 12

9-9:30: Course Overview

9:30-10:30: Coding & Python Basics

10:30-10:45: Break

10:45-12:15: Fundamentals of Data Manipulation

12:15-1:15: Lunch

1:15-3:15: Data Visualization

3:15-3:30: Break

3:30-4:30: Ethical Issues in Data Science

4:30-5: Questions & Project Planning

5-5:30: Artificial Intelligence (AI) Institute

Tuesday July 13

9-9:30: Clustering Using K-Means

10:30-10:45: Break

10:45-12:15: Regression Methods

12:15-1:15: Lunch

1:15-4:00: Group Work on Course Project

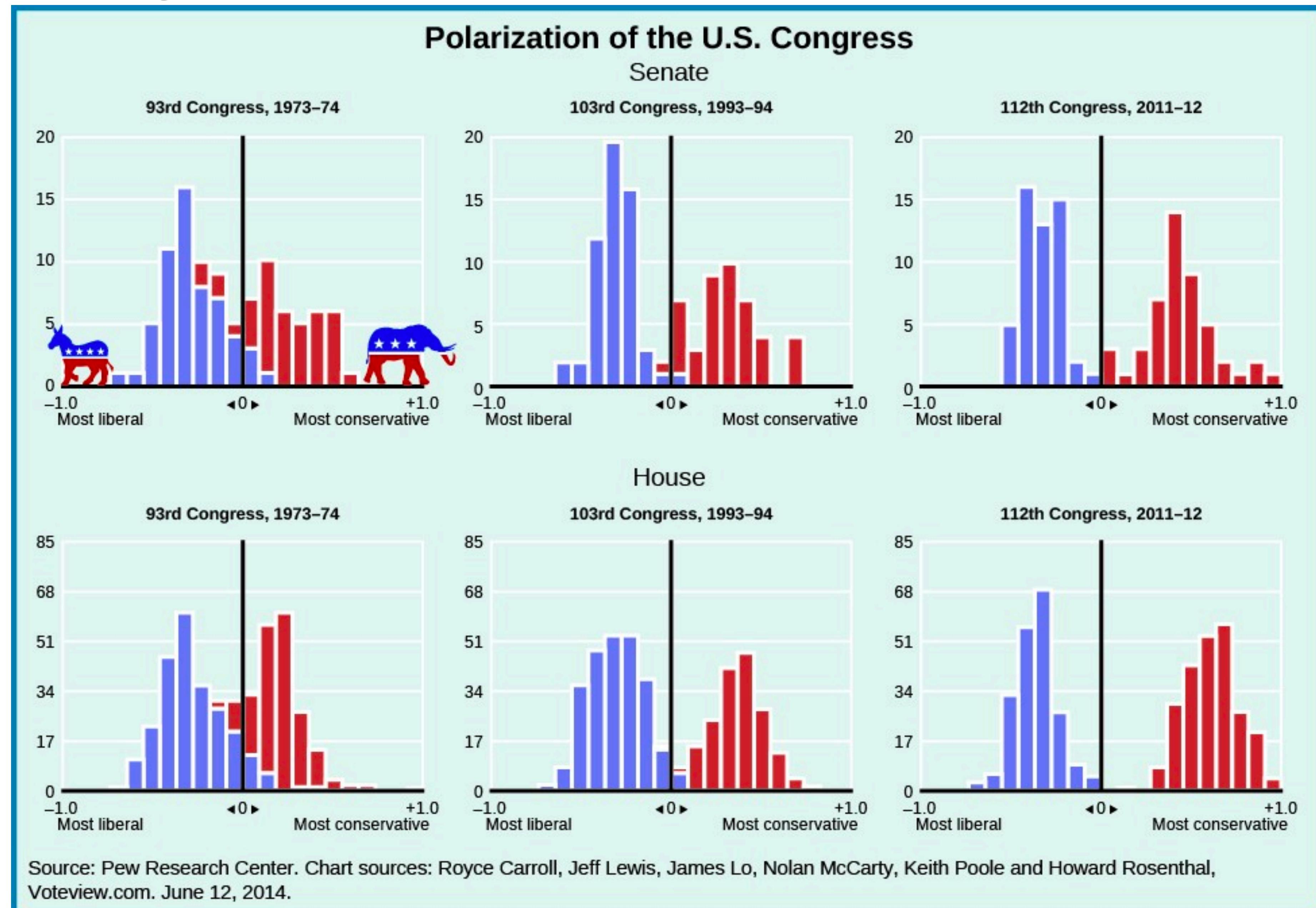
4:00-4:30: Project Presentations and Discussion

4:30-5:00: Final Thoughts and Questions

Project Logistics

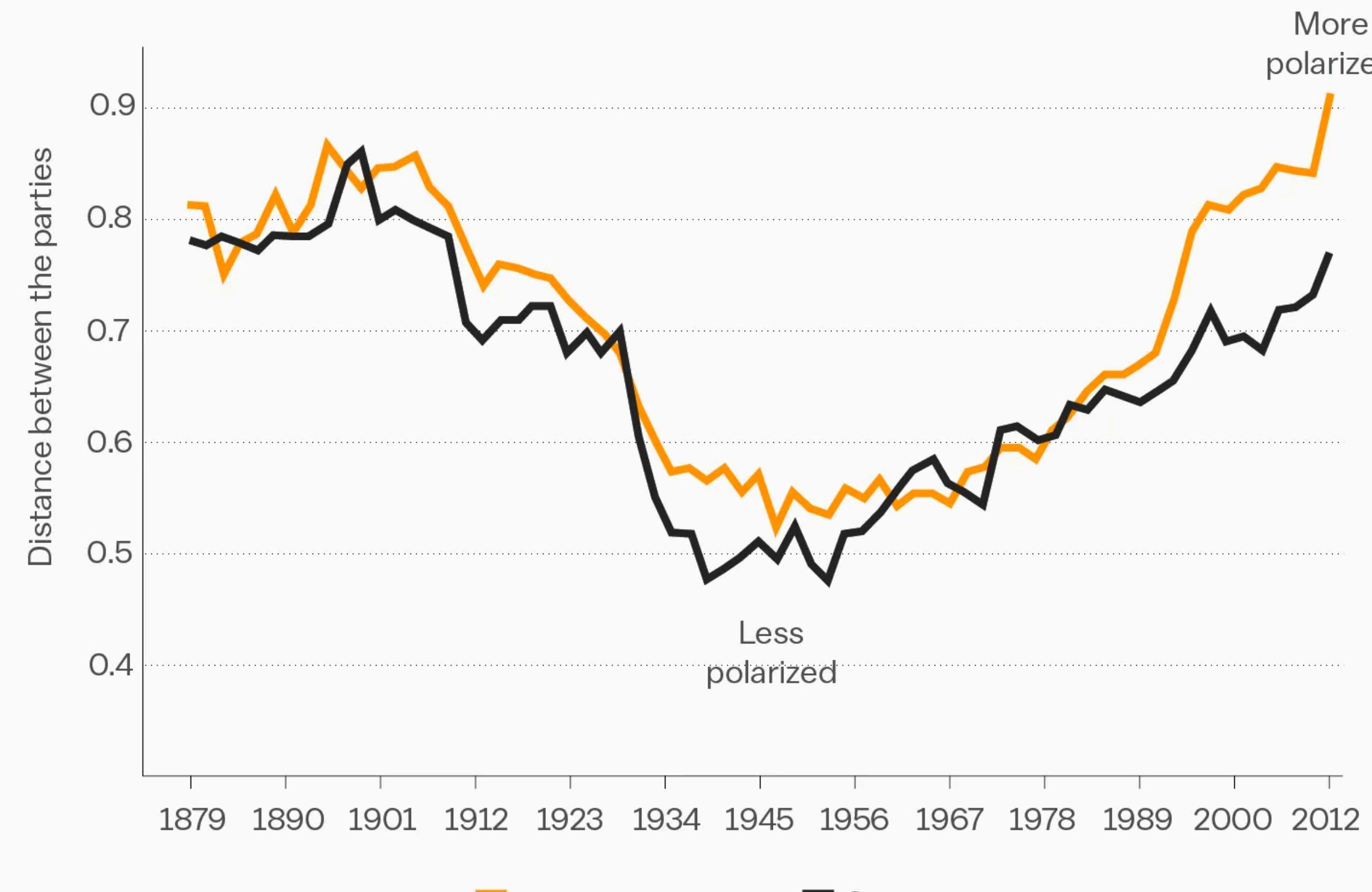
- **goals:**
 - get more independent experience working on data analysis and results interpretation in python
 - explore details of skills we cover that are of interest to you
 - have some freedom to pose and answer a specific research question about the project you choose
- work in pairs: in-person work together; virtual work together — will have a time block to choose shortly
- two prompts to choose from: suggested project work will be closely outlined, but you are free to jump ahead to more open ended parts
- expectation: get through at least most of the prompted portion of the project and be prepared to give a 1-2 minute report back on how it went, what you learned

Quantifying Political Polarization



Party polarization, 1879 - 2012

The ideological gap between the Democratic and Republican caucuses



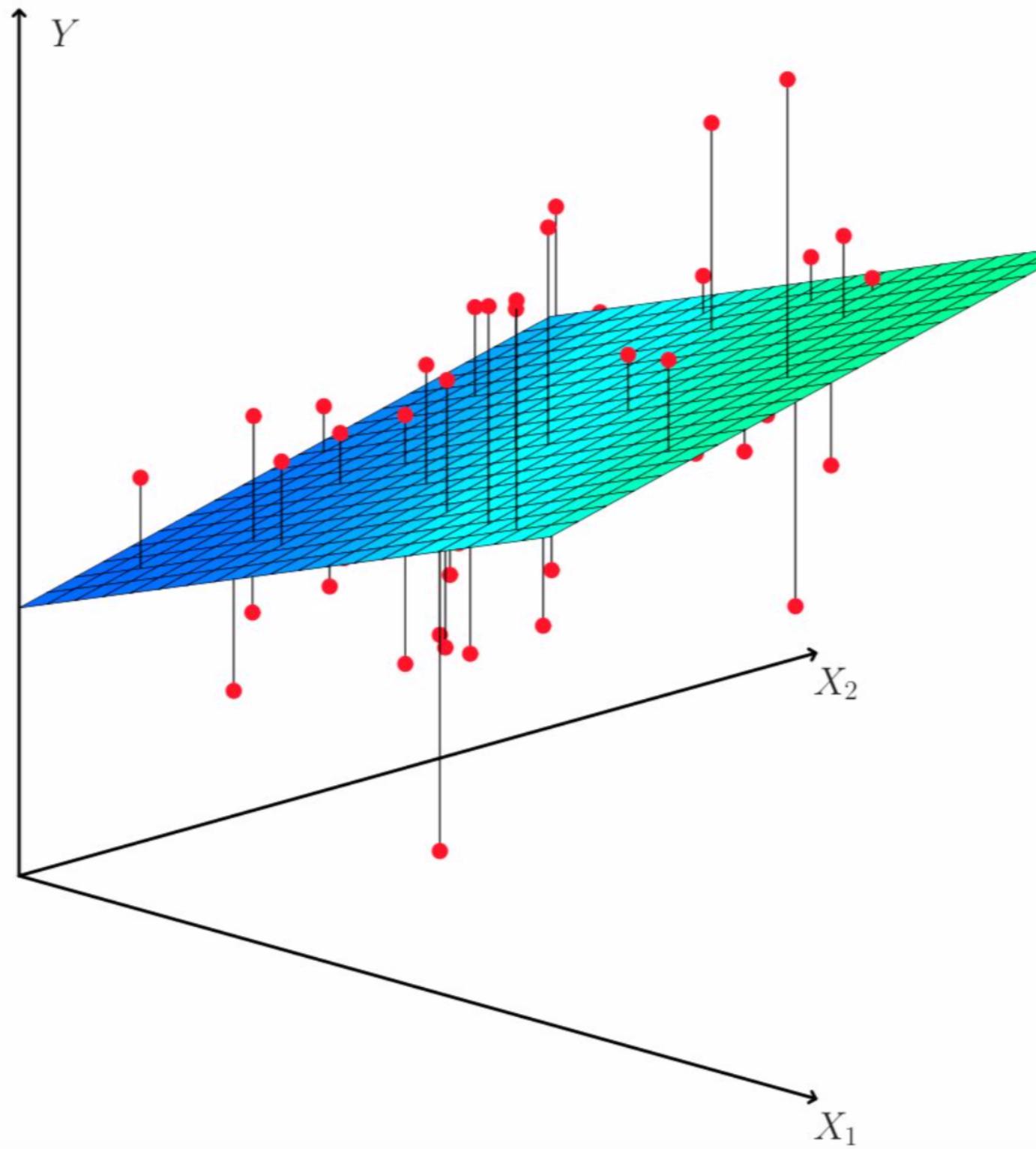
SOURCE: Voteview.com and DW-NOMINATE scores

Vox

- use clustering to test hypothesis that the US Senate (and the House) has become more polarized in terms of voting blocks across time
- could test other countries; could analyze with or without knowledge of party

- 1) Load data set on Congressional voting from a 2 year time increment
- 2) Clean data to obtain wide form with yea/nay/abstain indicated for each senator and corresponding bill vote
- 3) Compute and plot yeas/nays of each senator in a term; look to identify distinguishable clusters
- 4) Perform principal components analysis on the vote distribution of all senators and plot first two components to again see if data clusters
- 5) Perform clustering and compare to visualized distribution of yeas/nays as well as principal components
- 6) Compute a measure of how distinct the clusters are
- 7) Make comparisons of cluster separation score across time

Multiple Linear Regression



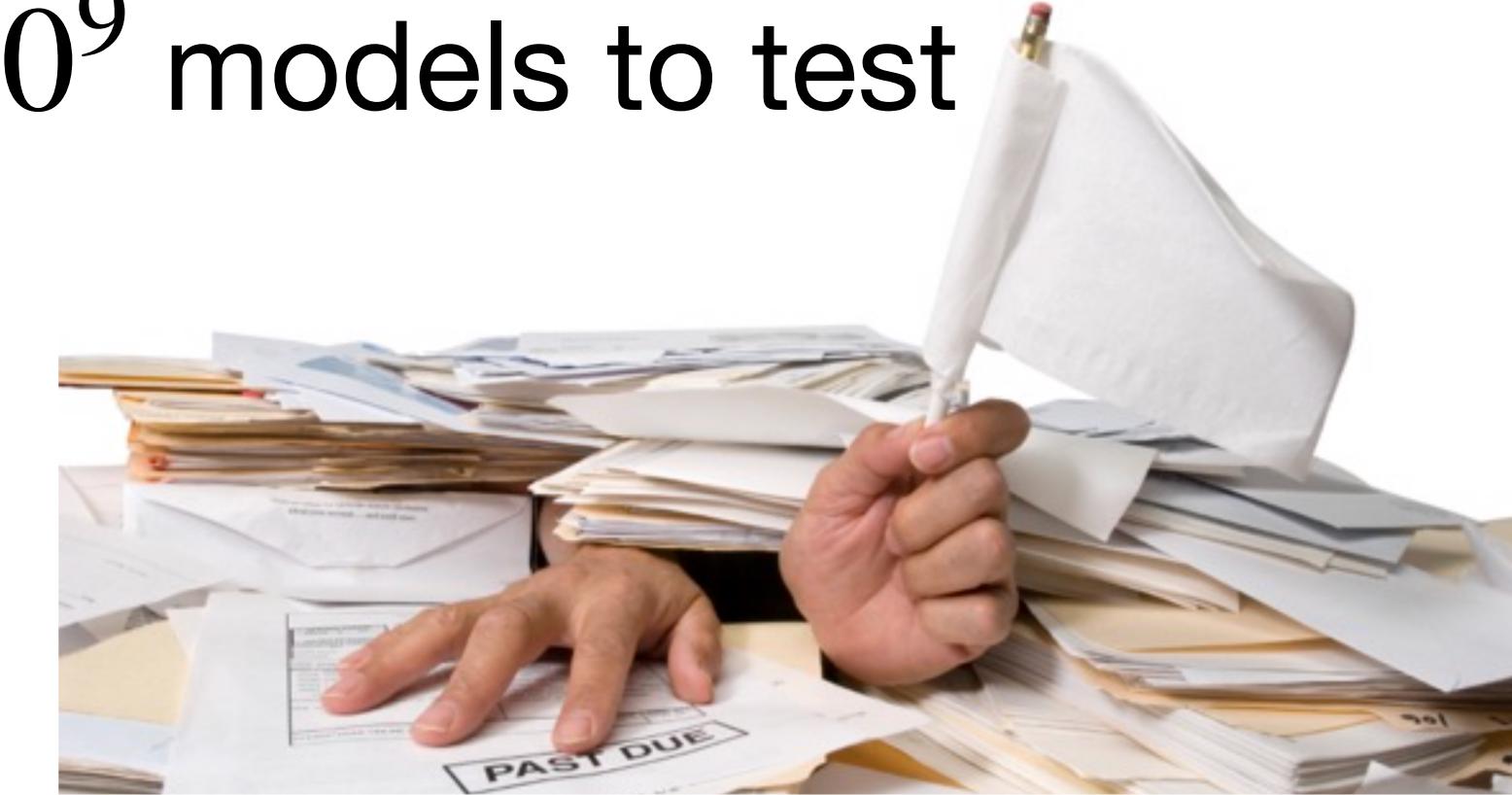
- Is at least one feature useful in predicting the response?
- Do all of the features help to explain the response?
- Or can we reduce to a few?
- How well does the model fit the data?
- How well does a subset of features do?

Model Selection: Which features to keep?

- Try all p features, choose the best combo.
- There are 2^p possible models, so for $p = 30$ that's $2^{30} \approx 10^9$ models to test



selection: A greedy algorithm for selecting features



- 1) Fit a null model with an intercept but no slope
- 2) Fit p individual SLR models – 1 for each feature. Add to the null model the one that improves the null model best (e.g., decreases SSE most, increases F-statistic most)
- 3) Fit $p - 1$ MLR – 1 for each of the remaining features along with the feature you isolated from step 2. Add one that improves the model performance most.

Backward selection: greedy algorithm for removing features

- 1) Fit model with all available features
- 2) Remove the feature with the largest p -value (i.e. the significant feature)
- 3) Repeat until some stopping criterion is reached. (e.g., some threshold SSE, or some fixed number of features)



Pick project groups

Pairs: In-person learners should pair together and virtual should pair together

Tomorrow after lunch, we will create breakout rooms, one for each project group

We'll take 5 minutes now for folks to pick groups, but can help people find partners if needed

You can wait and choose your topic tomorrow after the morning sessions