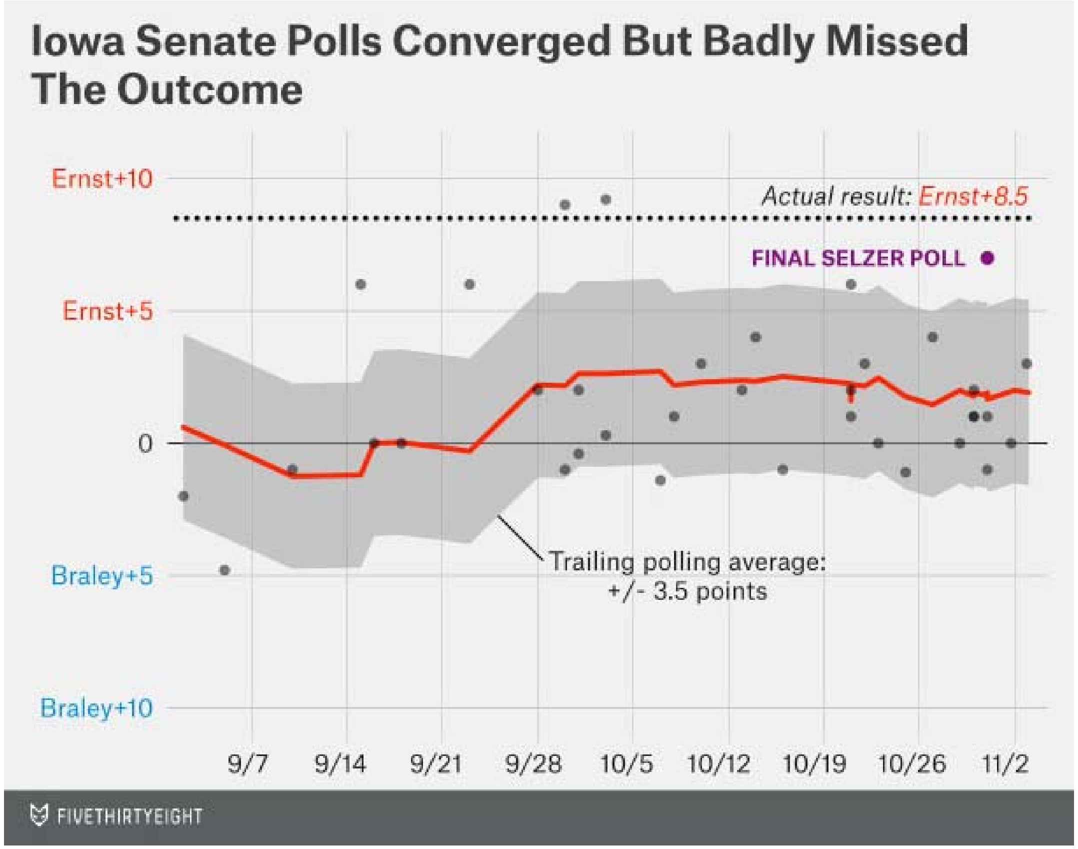
Thinking about Graphs

The Grammar of Graphics and Stata

Reconstructing a graph

From

http://fivethirtyeight.com/features/heres-proof-some-pollsters-are-putting-a-thumb-on-the-scale/



Questions toward reconstruction

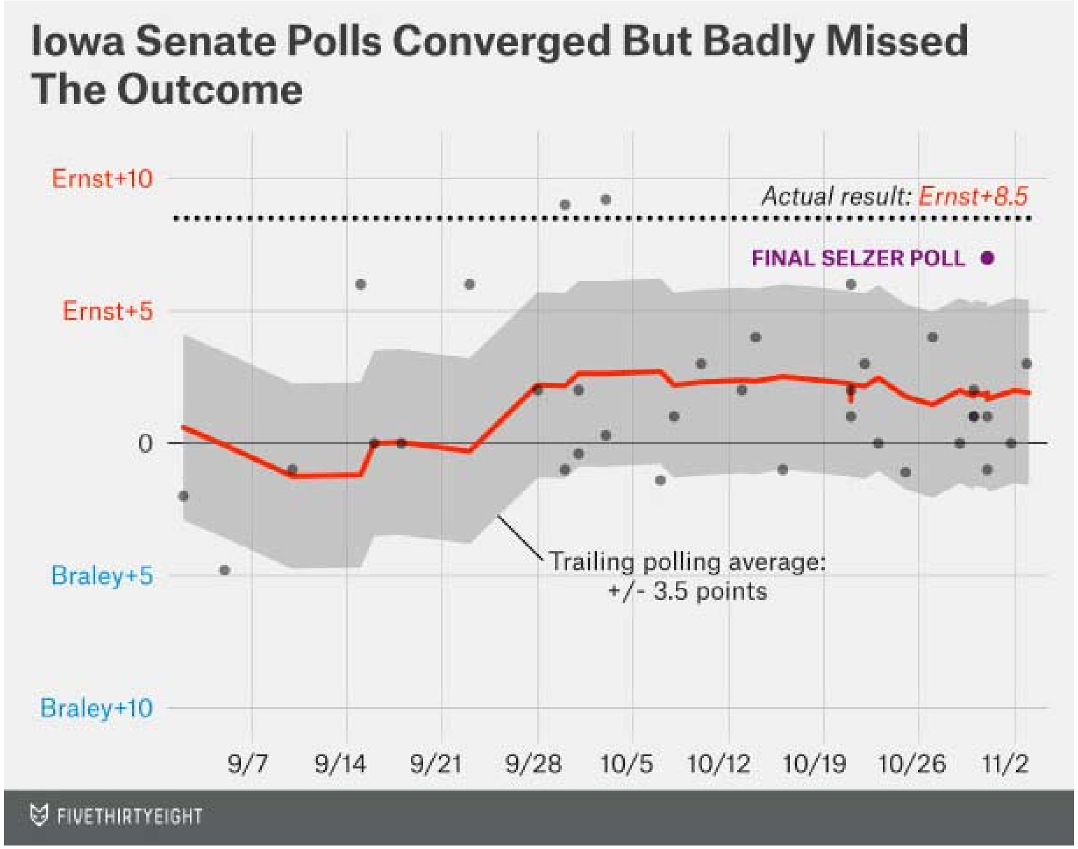
- What are the graphical elements?
- How are they related to data?
- How are they arranged on the screen/paper?
- How are they decorated?

Graphical elements

Points

Line(s)

Area


Iowa Senate Polls Converged But Badly Missed The Outcome

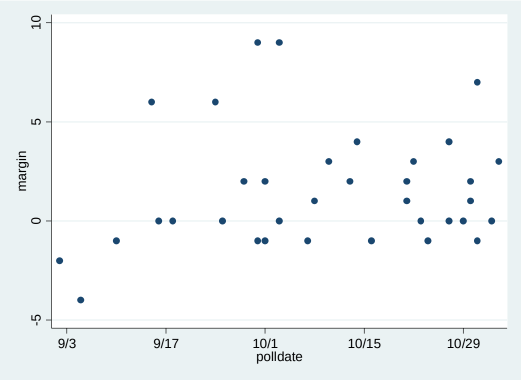
Relation to data

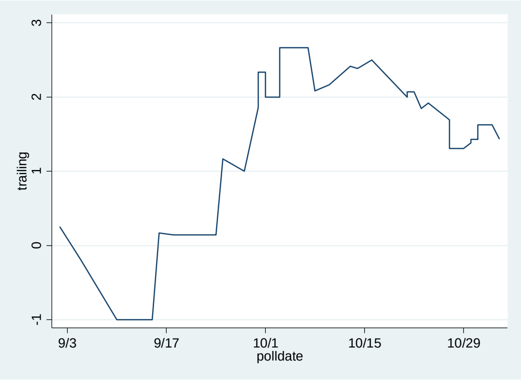
- Points: polling margins versus dates, essentially a scatter plot
- Lines:
    - Grid lines, some emphasized
    - Trailing margin is polling averages versus dates, connected (a.k.a. a line plot)
- Area: a fixed range around the trailing margin

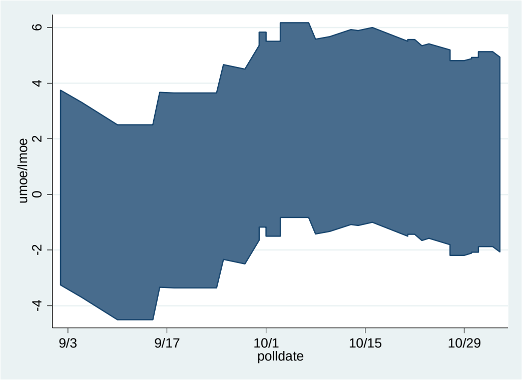
- Given the points, the lines and area can be calculated

Arrangement

- Think in layers, points on top of lines on top of area



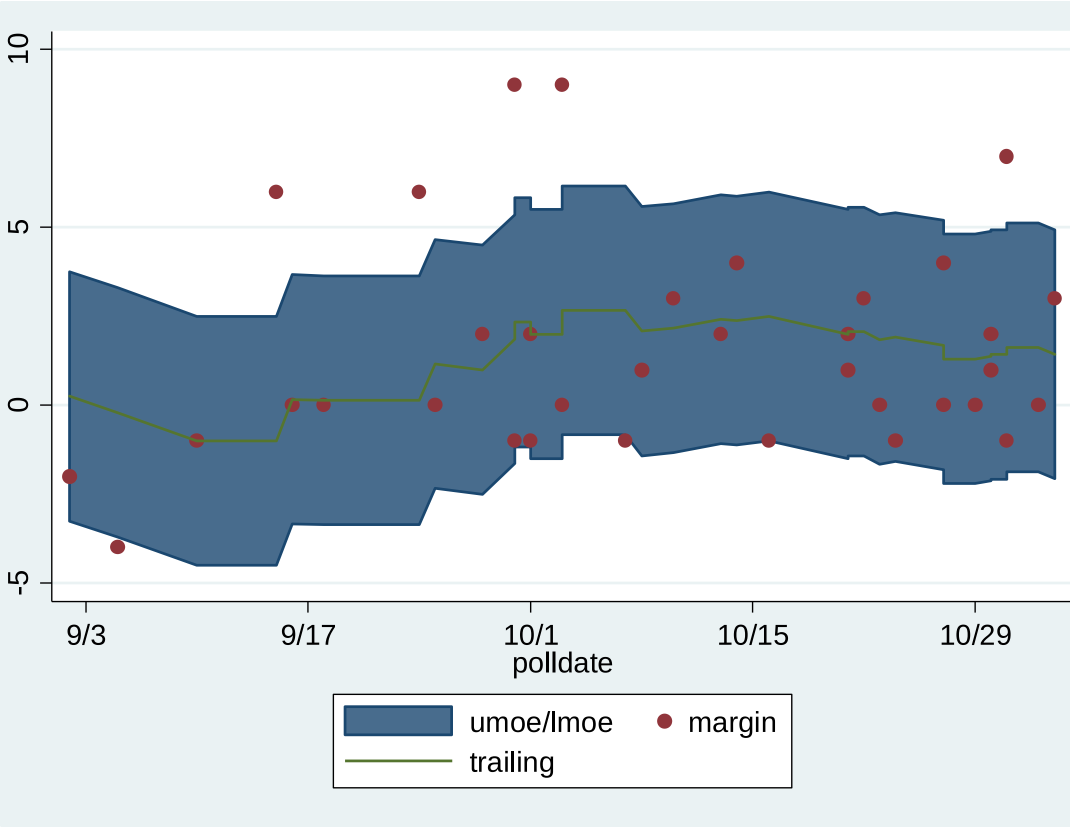




Layered together

Notice the scales now match.

The scales/coordinates are critical to how the elements are aligned on the page, and with each other.

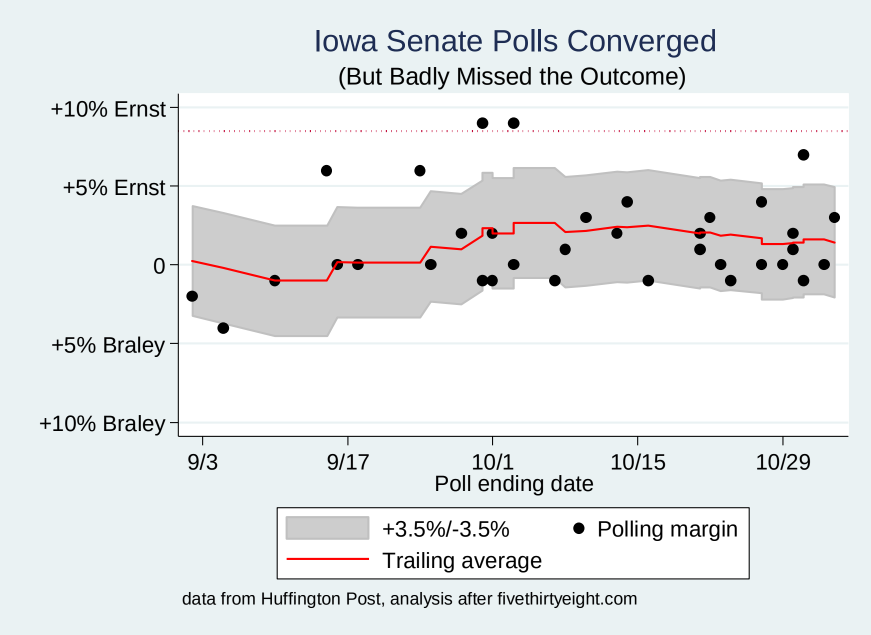


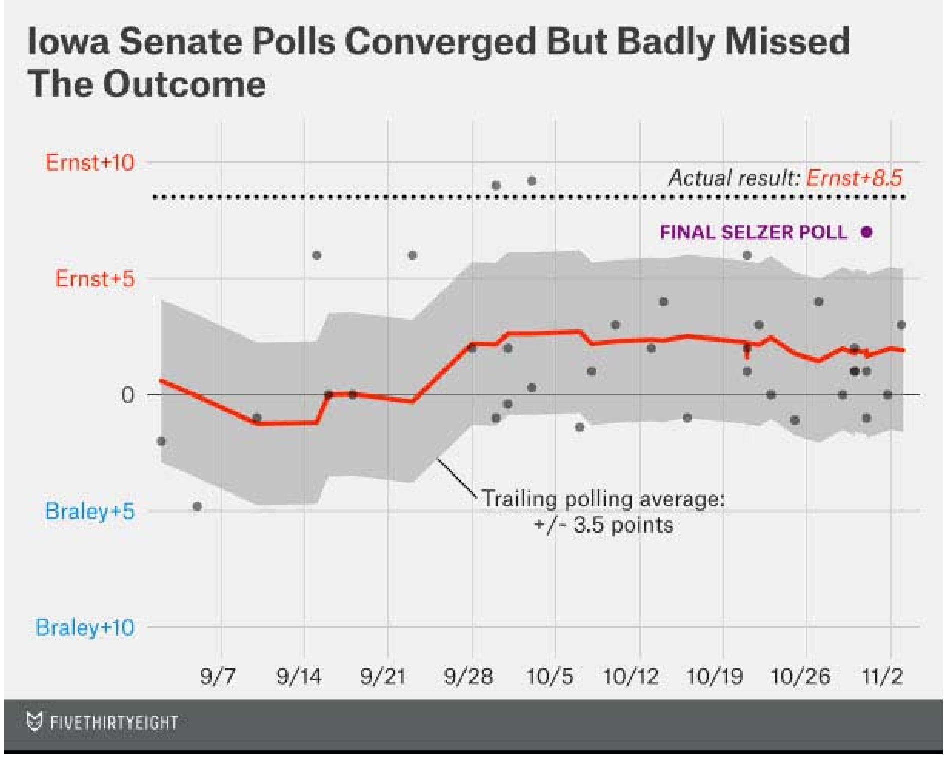
Decoration/Aesthetics

- Titles and footnotes
- Color, weight, etc. of graphical elements
- Axis and legend text
- Grid or guidelines

- Etc. – there tend to be a large number of options at this point

Reconstructed





Programming

- The final program starts with the end result in mind
  - Get the data, convert data types and layout (long vs. wide) as necessary
  - Calculate data values needed
  - Specify the graphics

Get, clean, convert the data

- `import delimited "Iowa HuffingtonPost.csv", clear`

- `generate LV = strpos(pop, "LV") > 0`
- `keep if LV    // Just use "likely voter" polls, not "registered voters"`

- `// Convert data from string to a form useful for graphing`
- `generate polldate = date(substr(date,strpos(date, "-")+2,.)+"/2014", "MDY")`
- `format polldate %tdnn/dd`
- `sort polldate  // sorting will make a nicer line graph, eventually`

- `rename margin spread`
- `generate margin = ernst - braley`

Calculate other needed data

```
• generate trailing21 = .
• forvalues i = 1/`=_N' {
•       local j = `i' - 1
•       generate win`i' = (polldate - polldate[`i']) >= -21 ///
•                       & (polldate - polldate[`i']) <= 0
•       egen trailing21`i' = total(margin) if win`i'==1
•       egen pool`i' = total(win`i')
•       generate trailmargin`i' = trailing21`i'/pool`i'
•       replace trailing21 = trailmargin`i' in `i' if pool`i' >= 3
•       drop win`i' trailing21`i' pool`i' trailmargin`i'
• }
• generate trailing = trailing21[_n-1]

• generate lmoe = trailing - 3.5
• generate umoe = trailing + 3.5
```

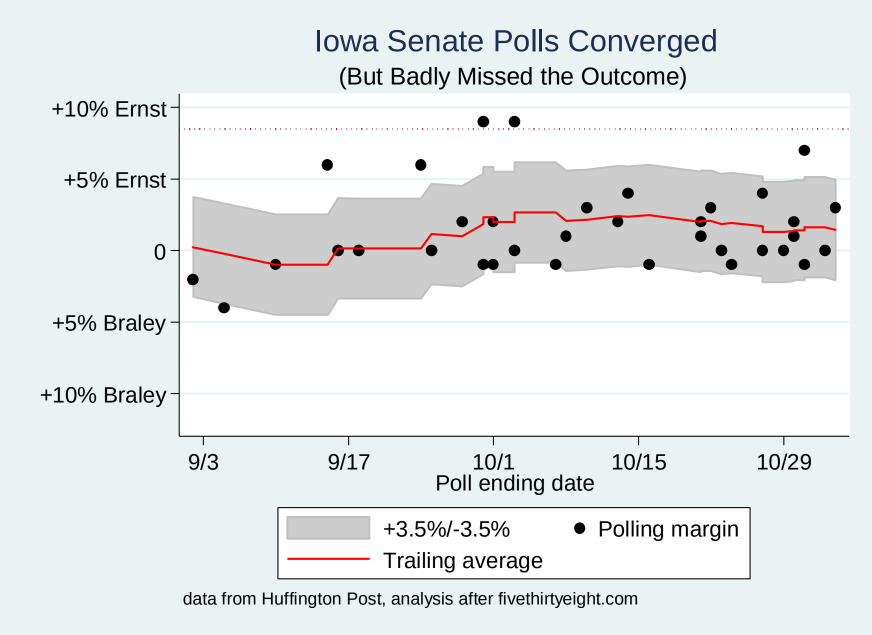
Basic graphical specification

```
• keep if polldate > td(1sep2014)

• twoway rarea umoe lmoe polldate || ///
•       scatter margin polldate || ///
•       line trailing polldate
```

With decoration

```
label variable polldate "Poll ending date"
label variable margin "Polling margin"
label variable trailing "Trailing average"
label variable lmoe "-3.5%"
label variable umoe "+3.5%"

twoway rarea umoe lmoe polldate, color(gs12) || ///
        scatter margin polldate, color(black) || ///
        line trailing polldate, color(red) ///
        yline(8.5, lpattern(dot)) yscale(range(-10(5)12)) ///
        ylabel(-10 "+10% Braley" -5 "+5% Braley" 0 "0" 5 "+5% Ernst" 10 "+10% Ernst", angle(0)) ///
        title("Iowa Senate Polls Converged") subtitle("(But Badly Missed the Outcome)") ///
        note("data from Huffington Post, analysis after fivethirtyeight.com")
```

After all the steps



data from Huffington Post, analysis after fivethirtyeight.com