Data Visualization II

Principles of Bivariate Graphics

Last week

- Visualization in three phases: three “how” questions to ask

  1. ***How*** humans ***perceive*** graphical information

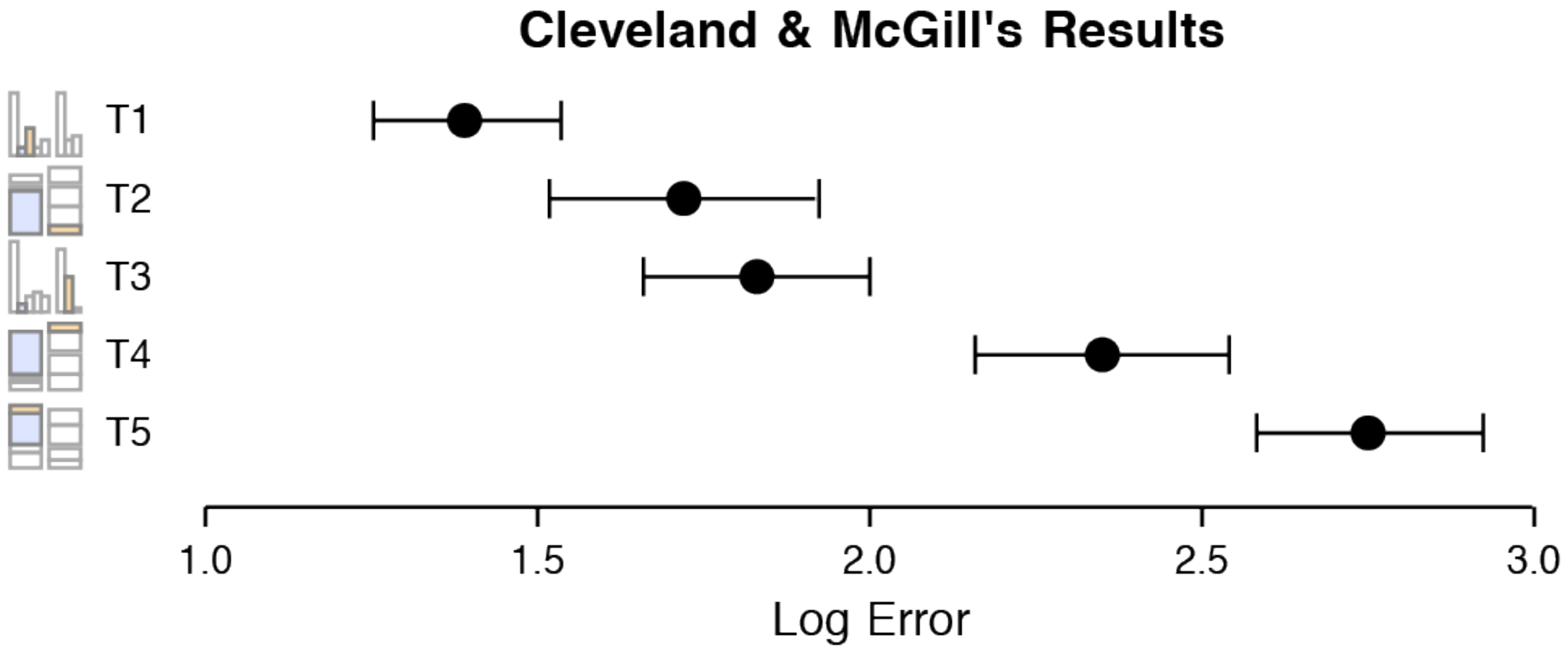
  2. ***How*** human perception means you should ***design*** your graphics

  3. ***How***, practically, you ***draw*** your graphics with software

Human Perceptual Efficiencies and Inefficiencies

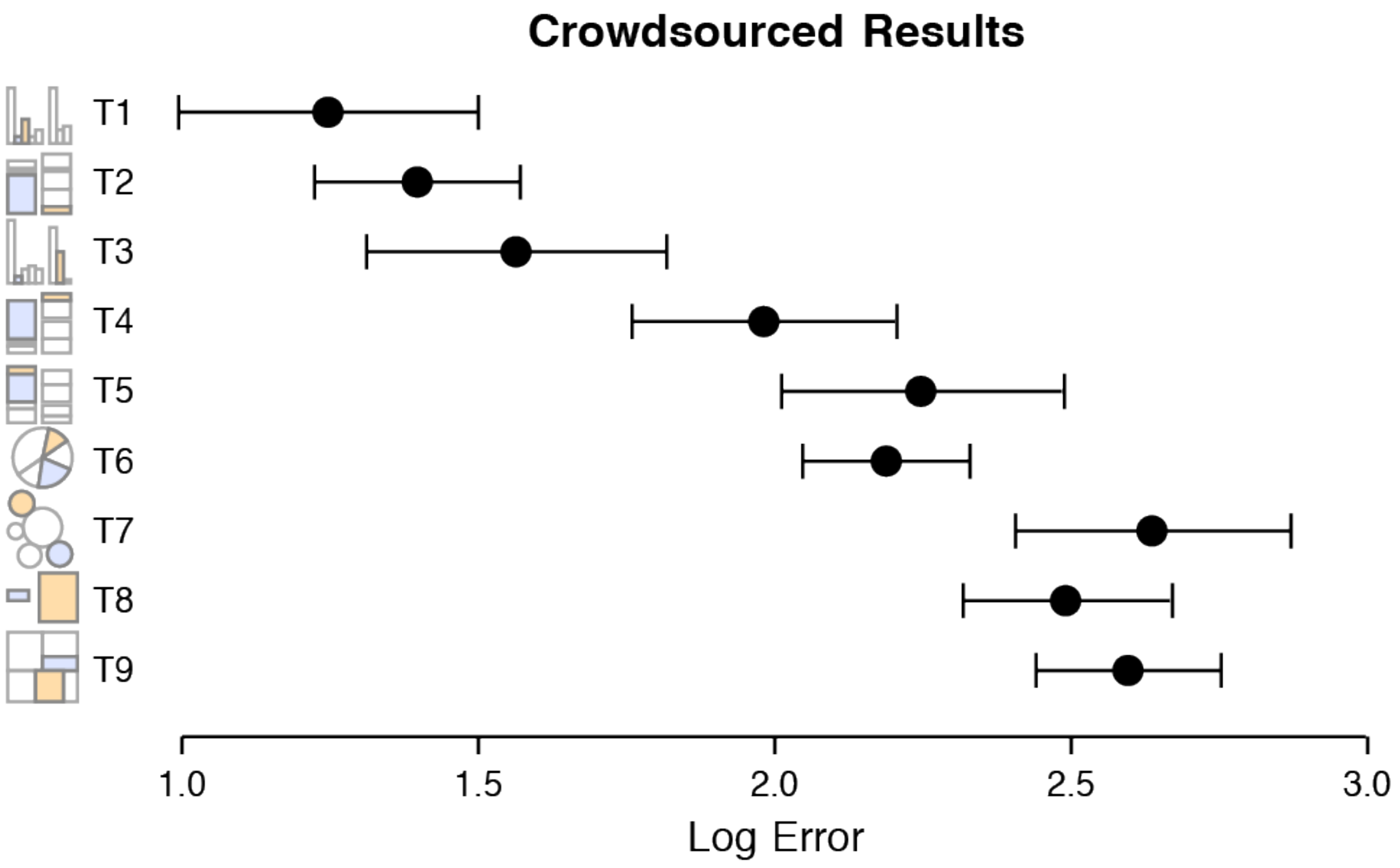
Ease and efficacy of perceptual tasks

- Another representation (Healy, 2019)



Replicated

- Heer and Bostock (2010)
  - Note the continued strength, especially, of position on a scale



Drawing attention (Healy, 2019)

- While color and shape is *not* always effective for conveying scale
  - They can be useful for drawing attention to points, but mind cognitive capacities

What makes a graphic good and bad?

“Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design … [It] consists of complex ideas communicated with clarity, precision, and efficiency. … [It] is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space … [It] is nearly always multivariate … And graphical excellence requires telling the truth about the data.”

– Edward Tufte (1983) ***p. 51***

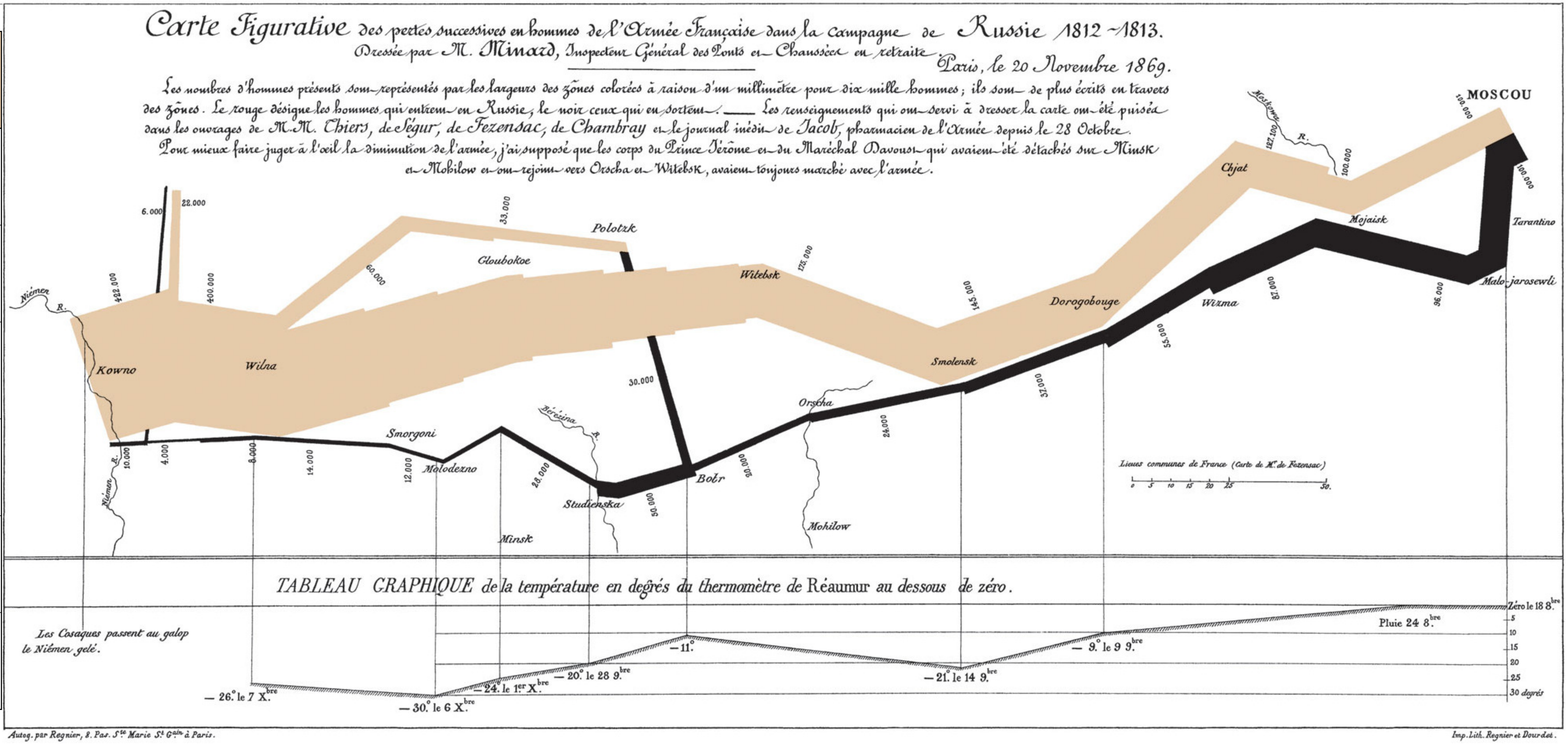
as quoted in Healy (2019)

Clarifying

- Healy (2019). Three varieties of badness:
  - **Aesthetic** - a graph that is tasteless or inconsistent
  - **Substantive** - there is a problem with the data being presented
  - **Perceptual** - a graph can be “confusing or misleading because of how people perceive and process what they are looking at”
- **Chartjunk**
- **Data-Ink Ratio**

Minard’s Napoleon

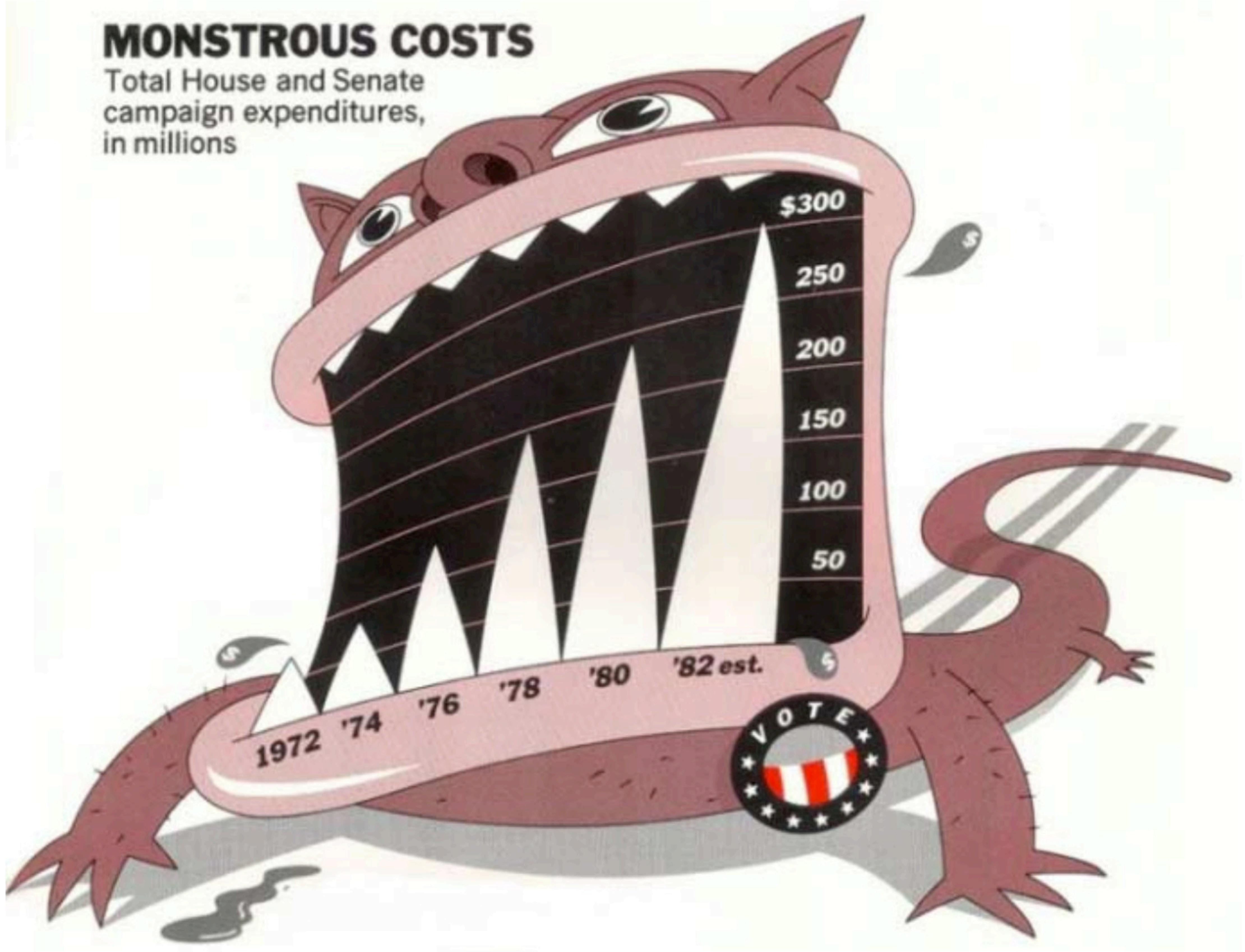
Information	Aesthetic
Size of Napoleon’s Grande Armée	Width of path
Longitude of the army’s position	x-axis
Latitude of the army’s position	y-axis
Direction of the army’s movement	Color of path
Date of points along retreat path	Text below plot
Temperature during the army’s retreat	Line below plot



https://www.andrewheiss.com/blog/2017/08/10/exploring-minards-1812-plot-with-ggplot2/

Complicating the story

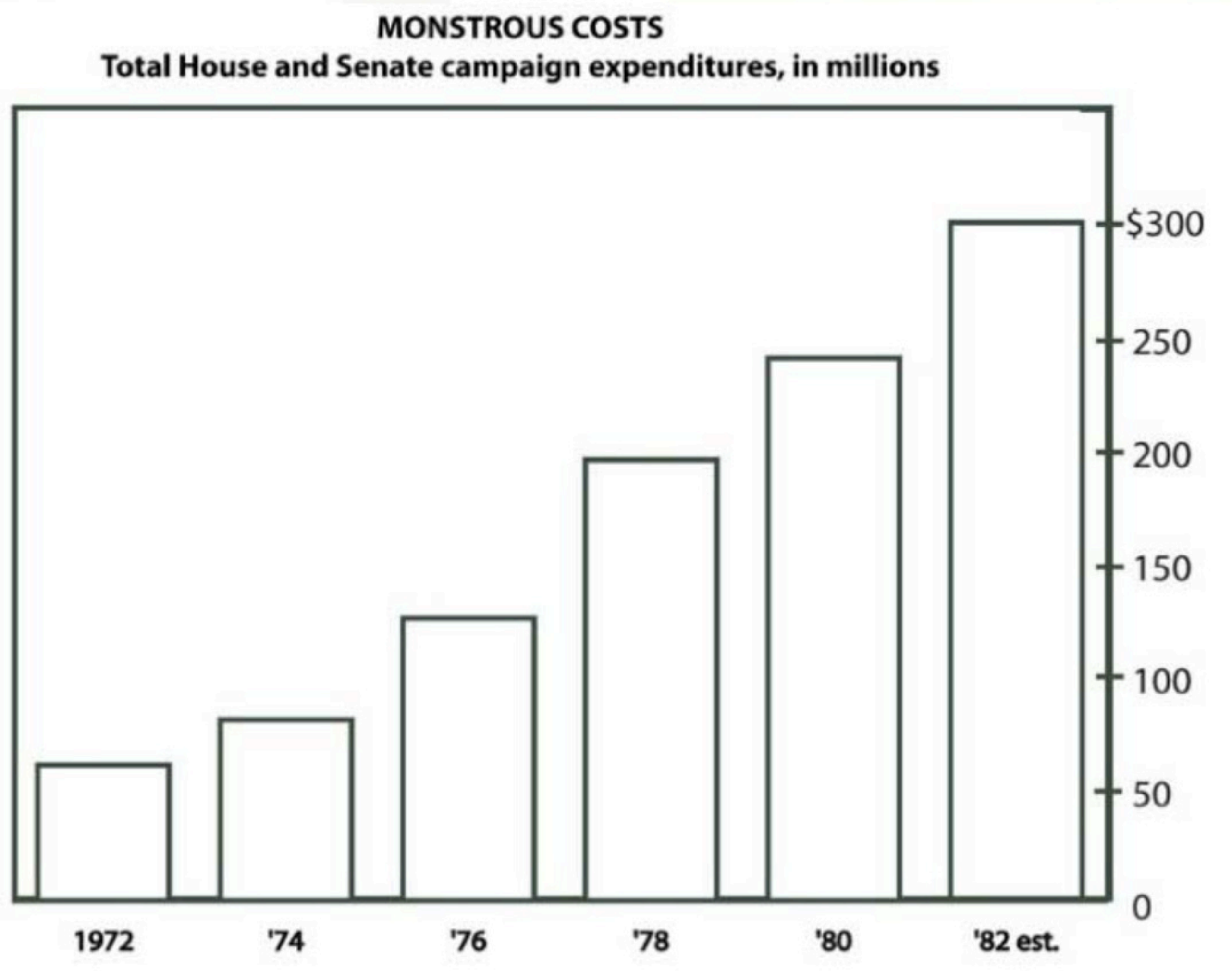
- A figure that breaks all the rules



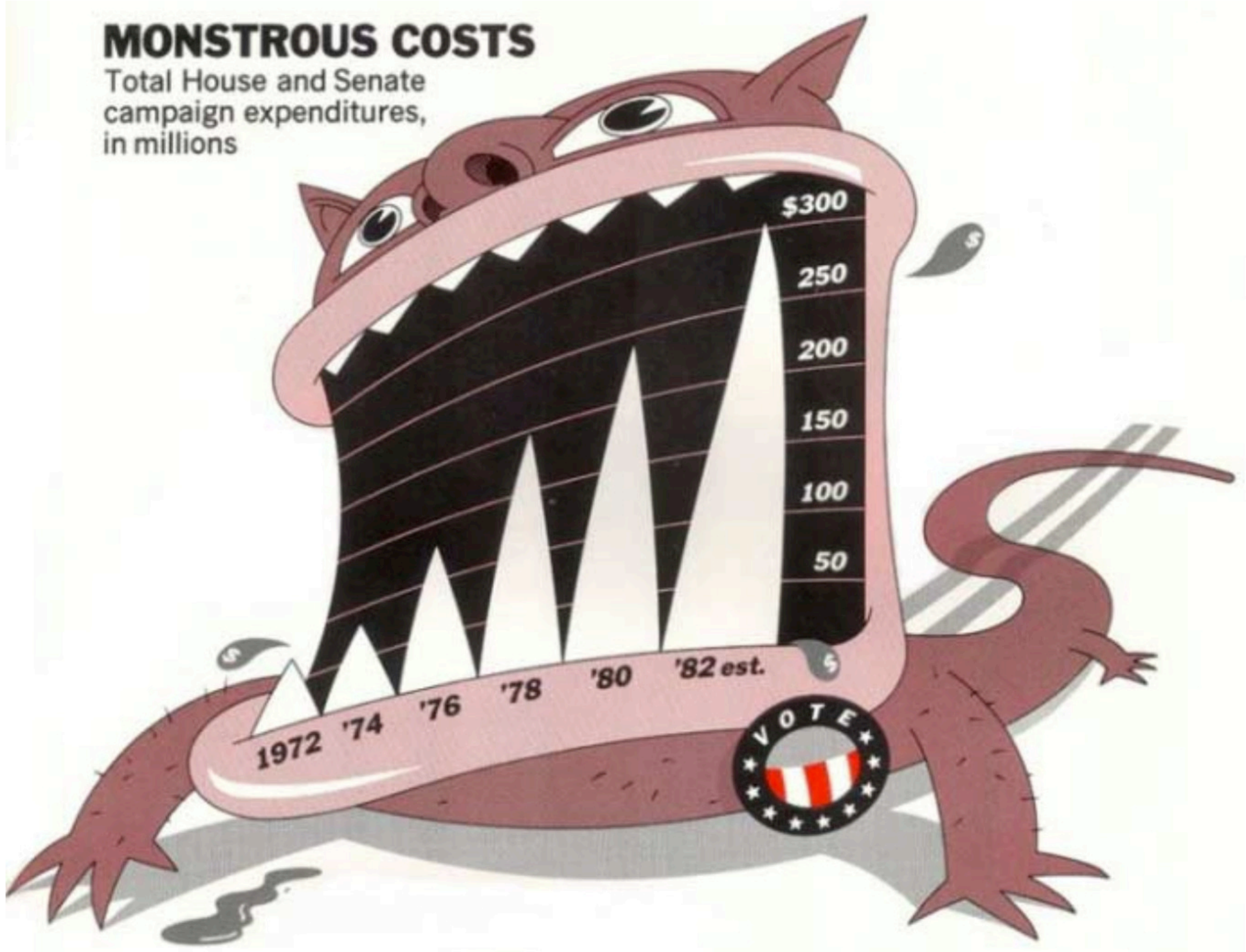
(Bateman (2010); image: Nigel Holmes)

Complicating the story

- A figure that breaks all the rules
  - And a cleaner version



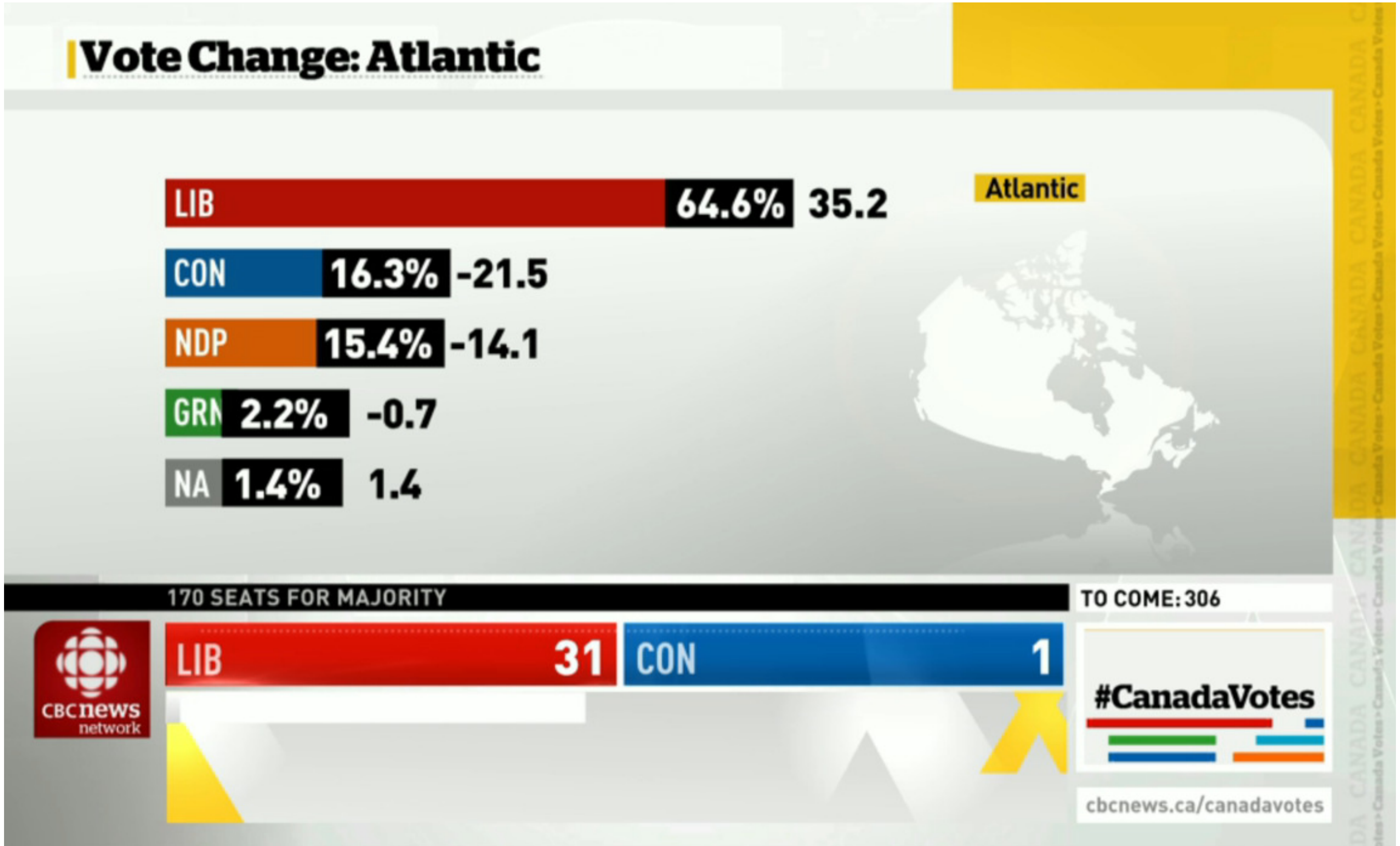
  - Yet the original is more memorable

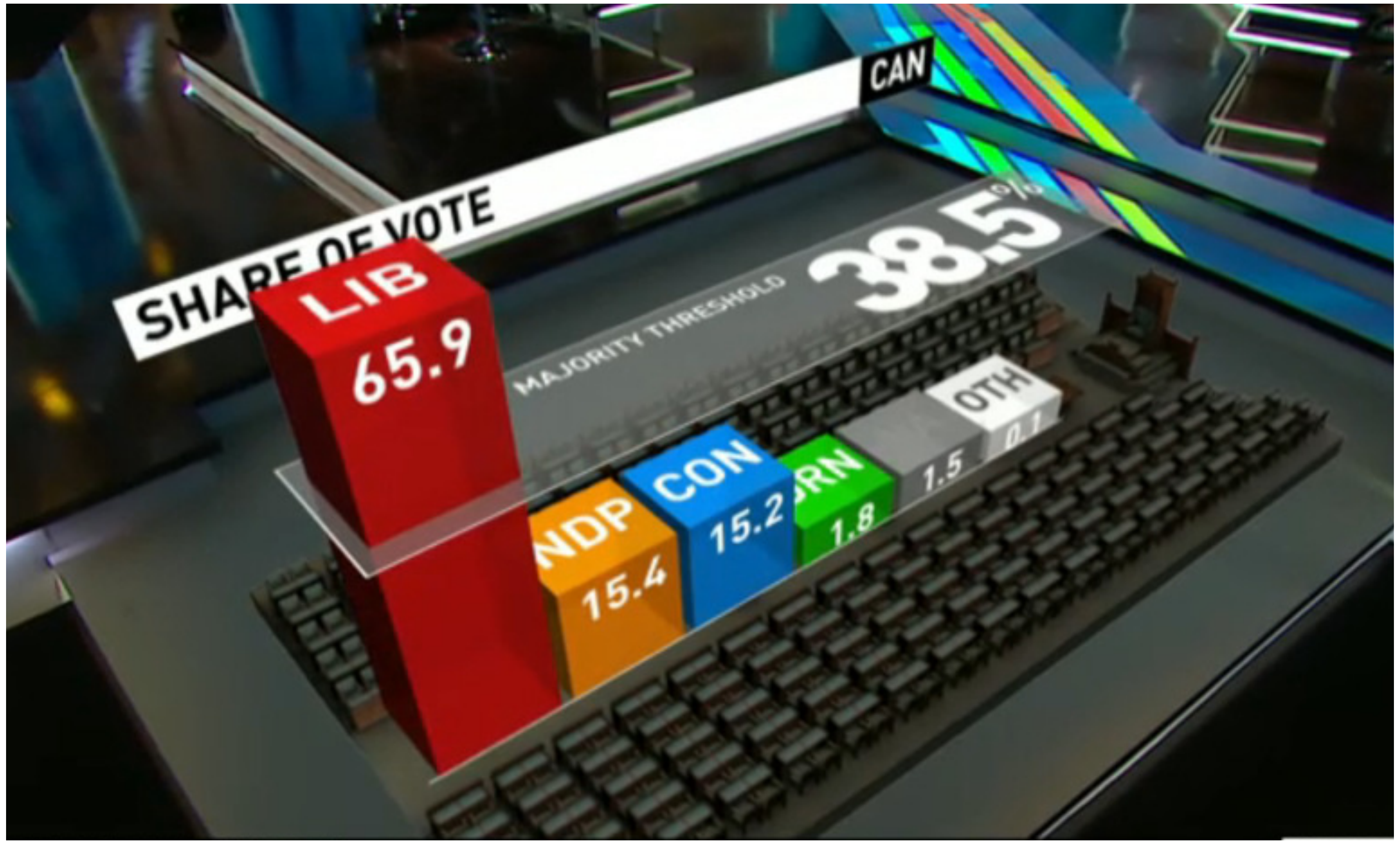


(Bateman (2010); image: Nigel Holmes)

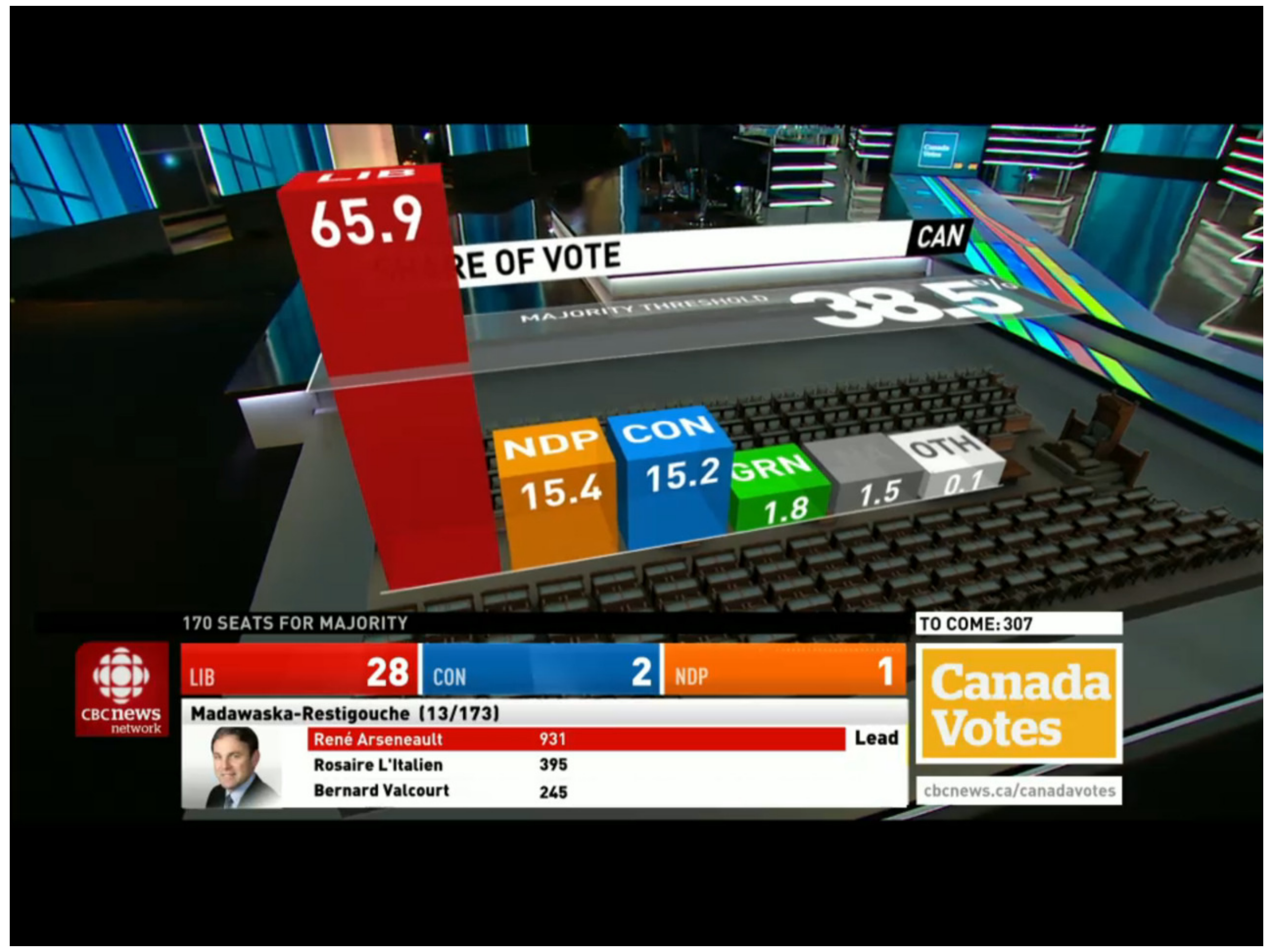
Bad Graphics

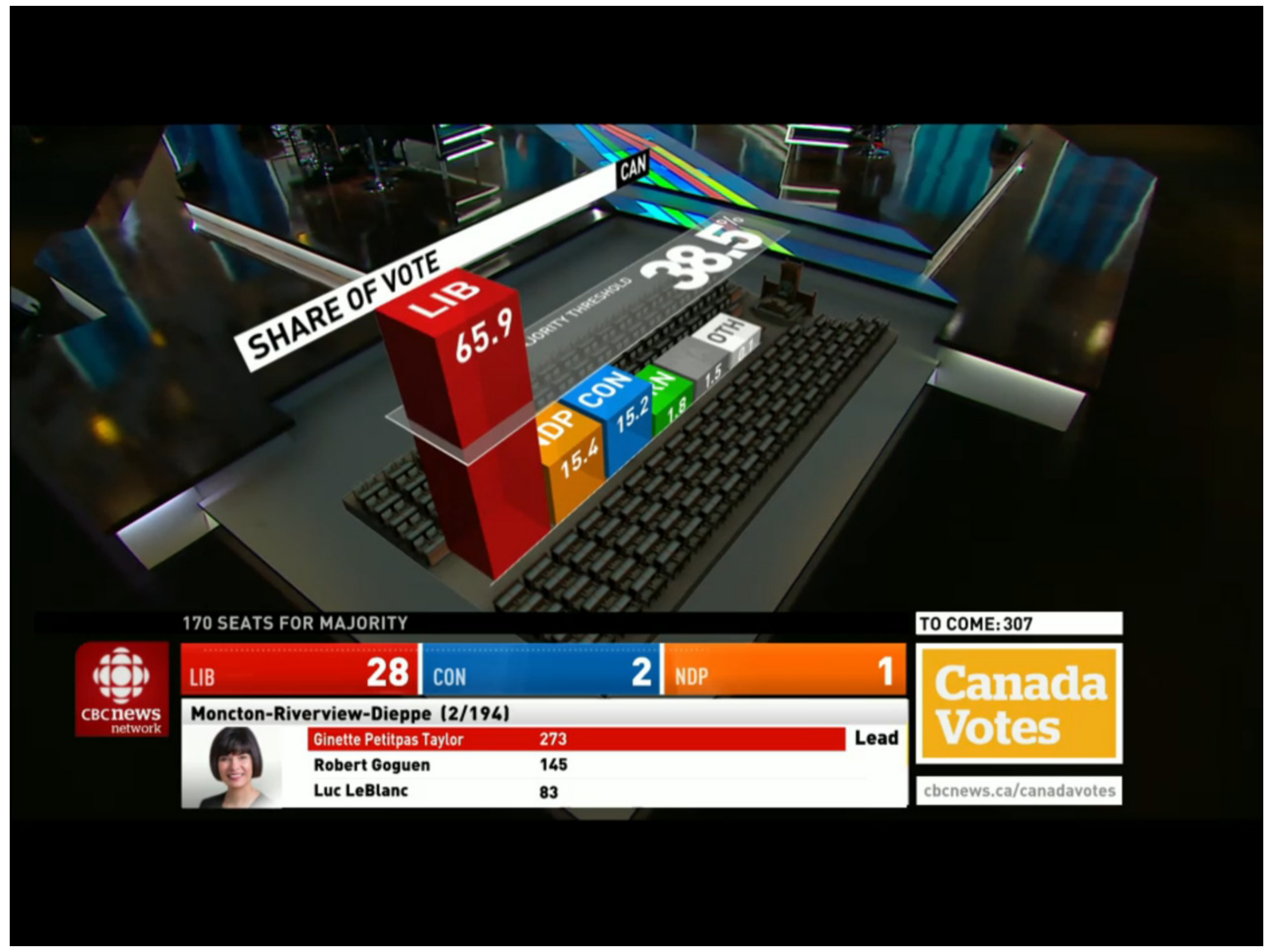
We also talked about some heinous graphics . . . how about some more?

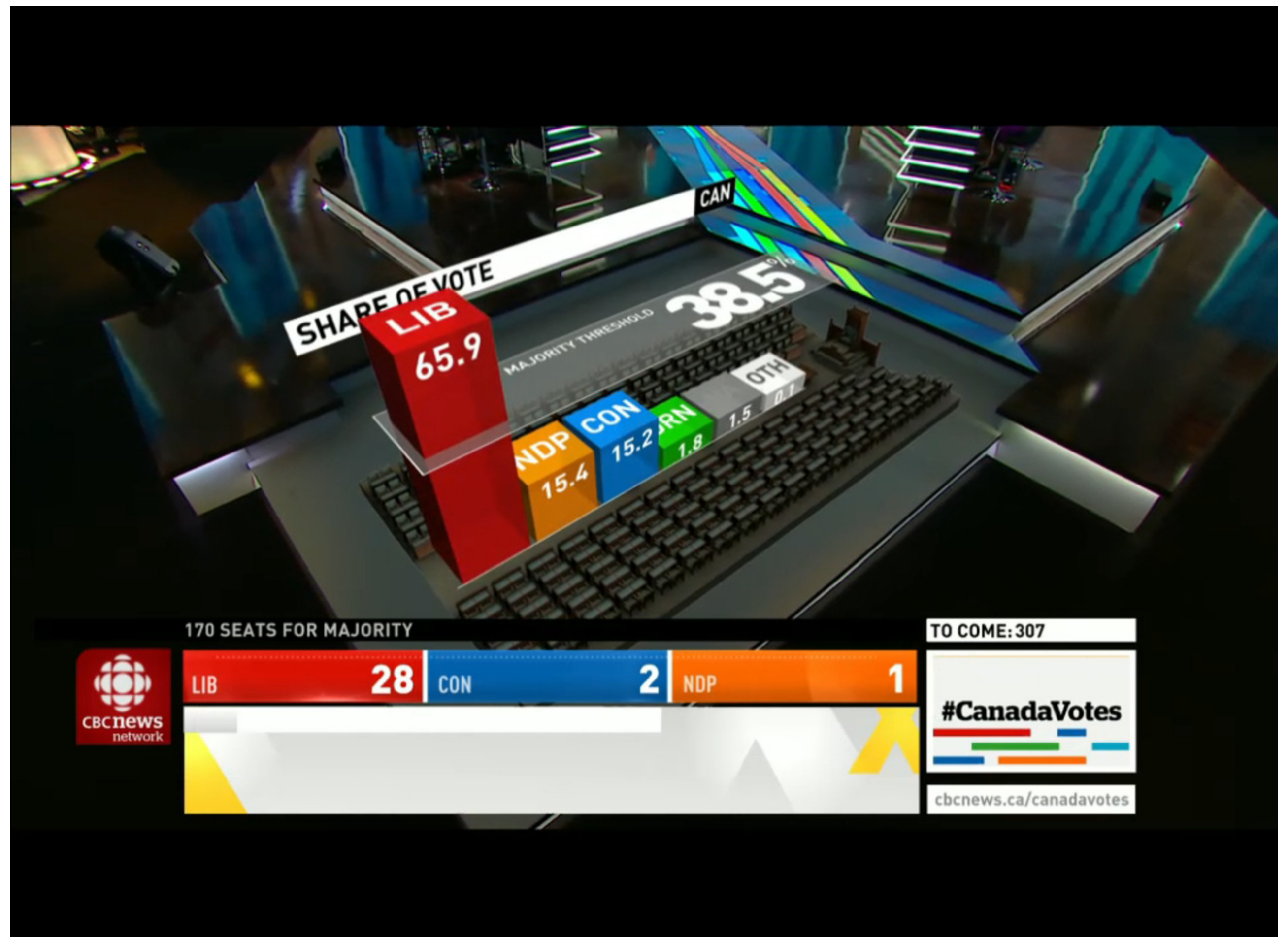


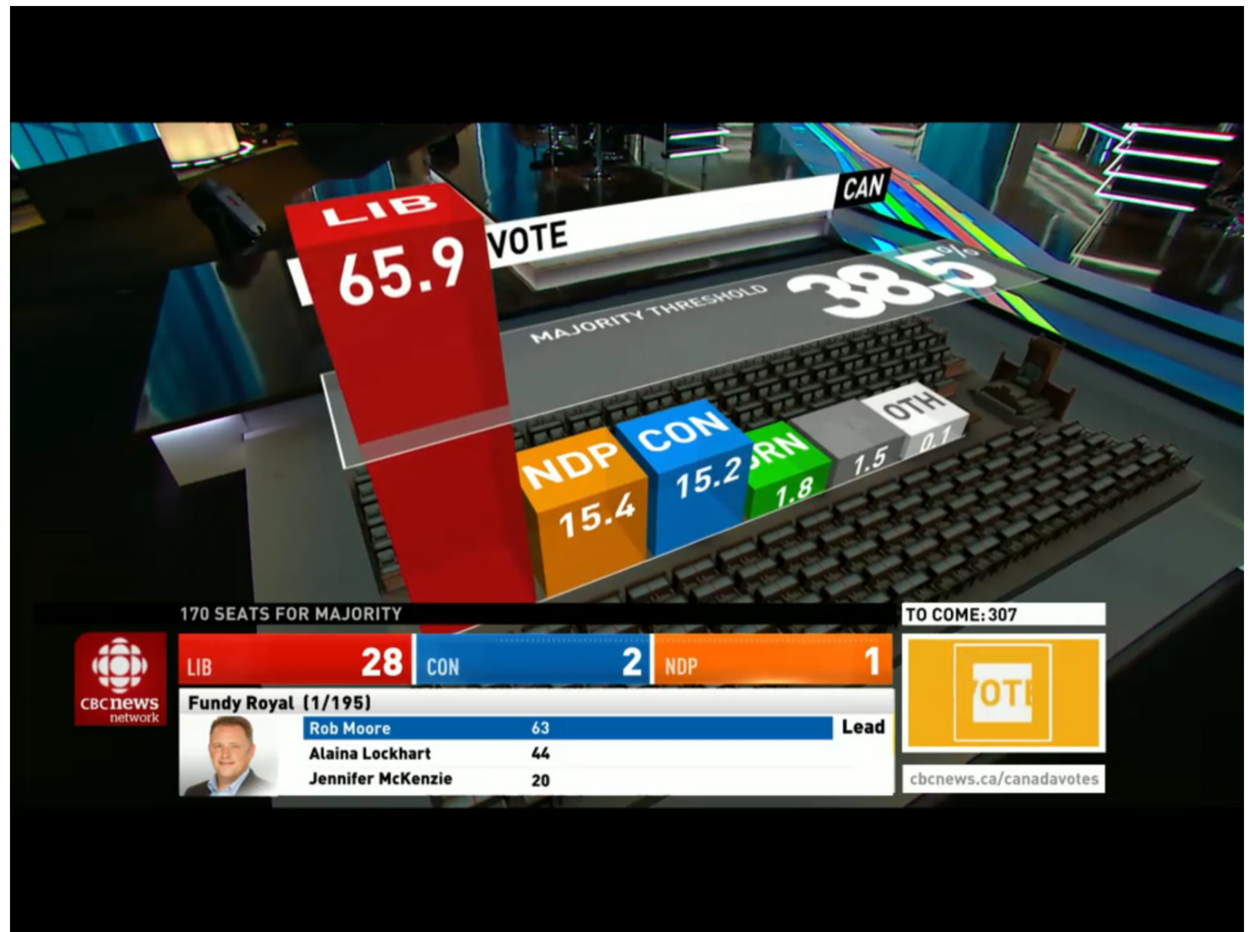

CAN
SHARE OF VOTE
LIB
65.9
MAJORITY THRESHOLD
38.5%
NDP
15.4
CON
15.2
1.8
1.5
OTH
0.1

this thing MOVED! and zoomed. and stuff.

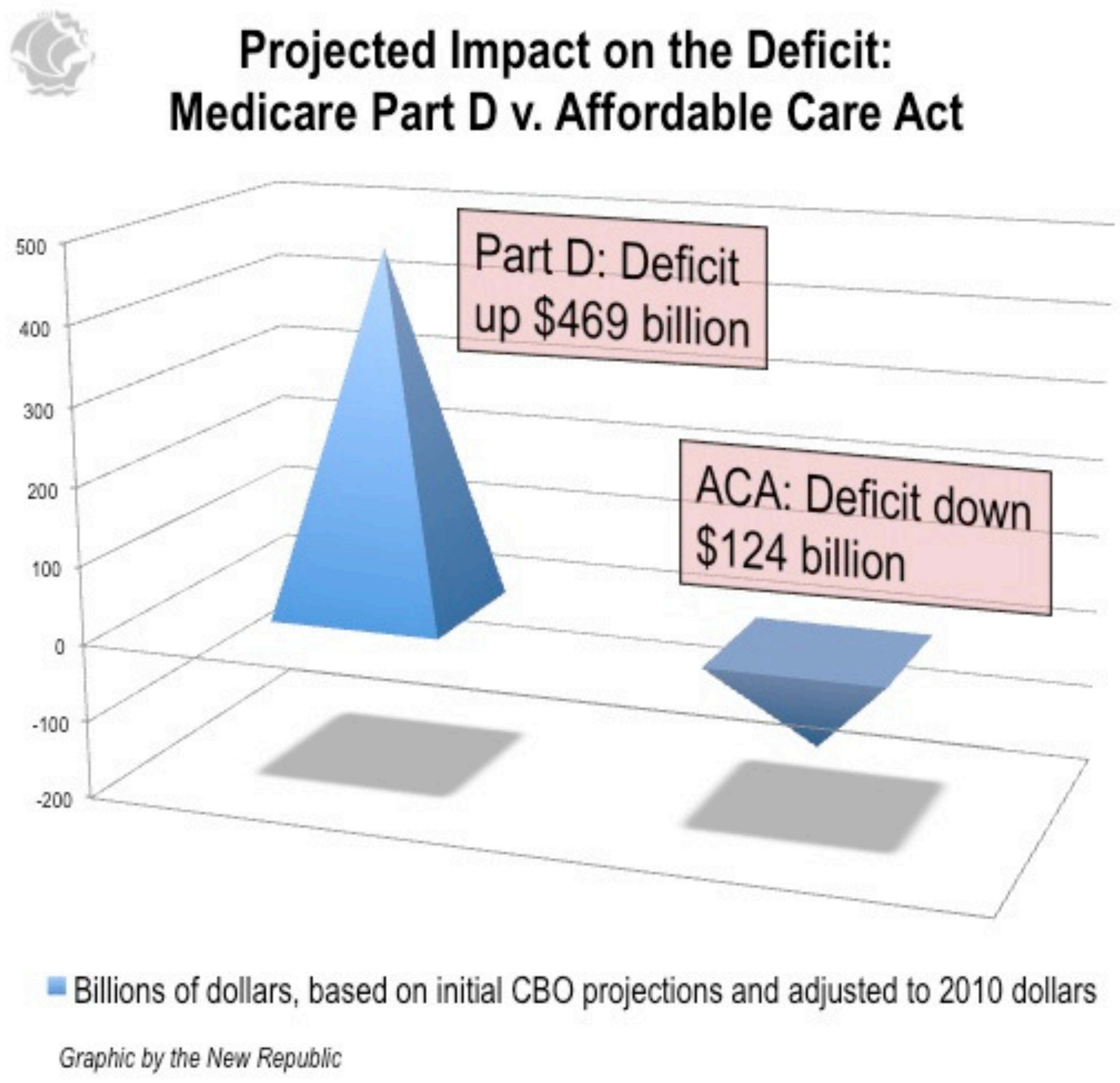




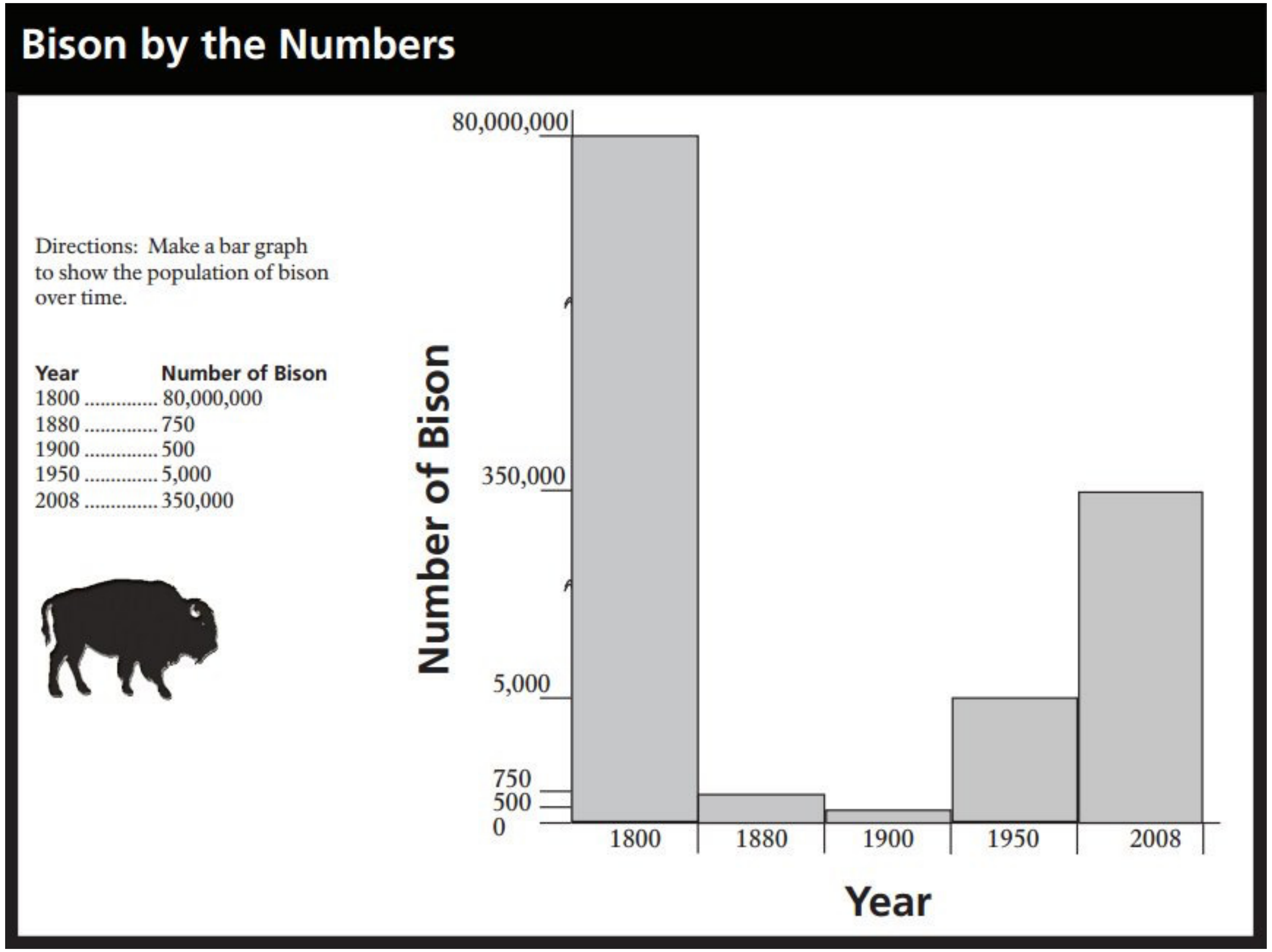




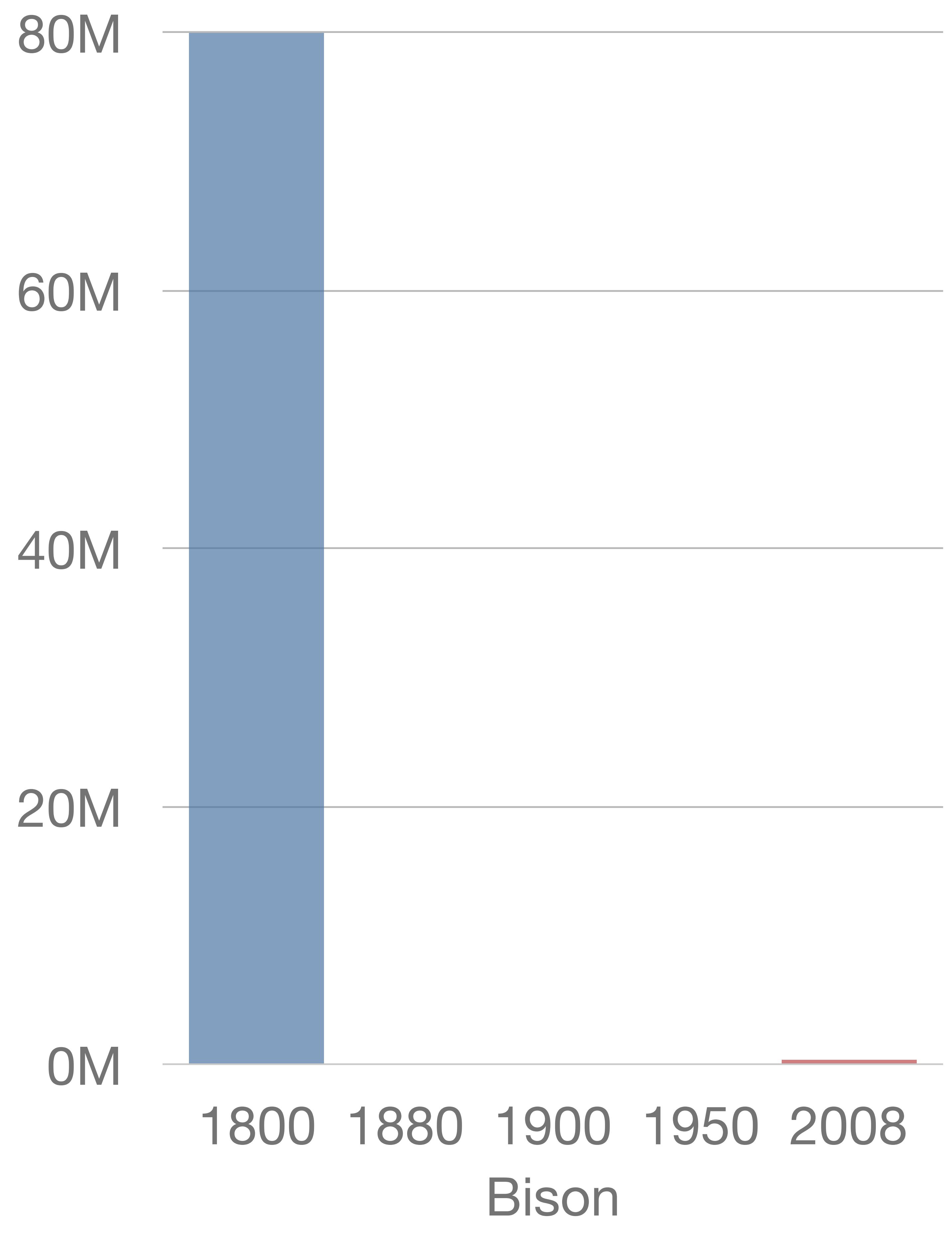
More bad graphics



More bad graphics

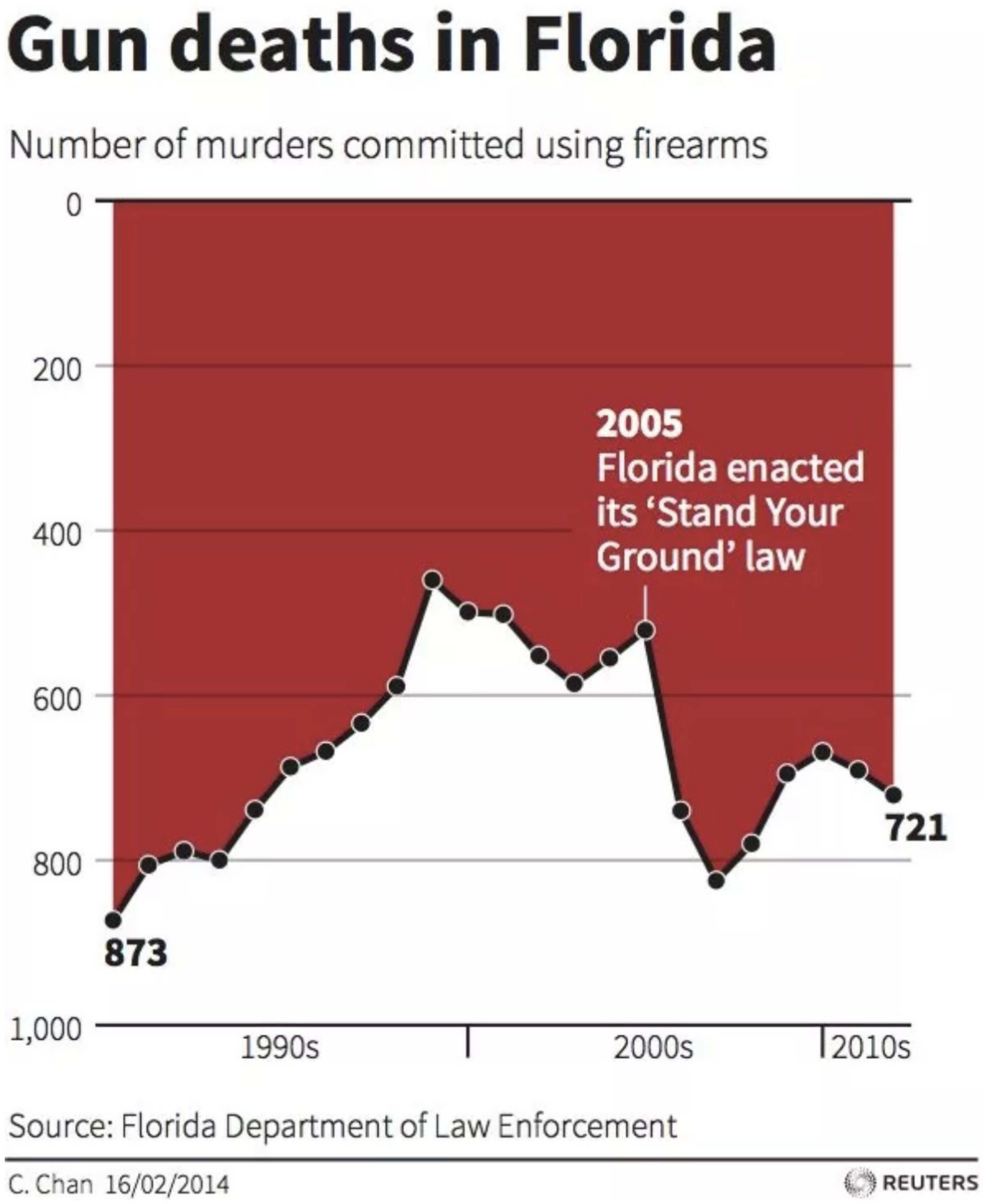


A *less* terrible bison graph

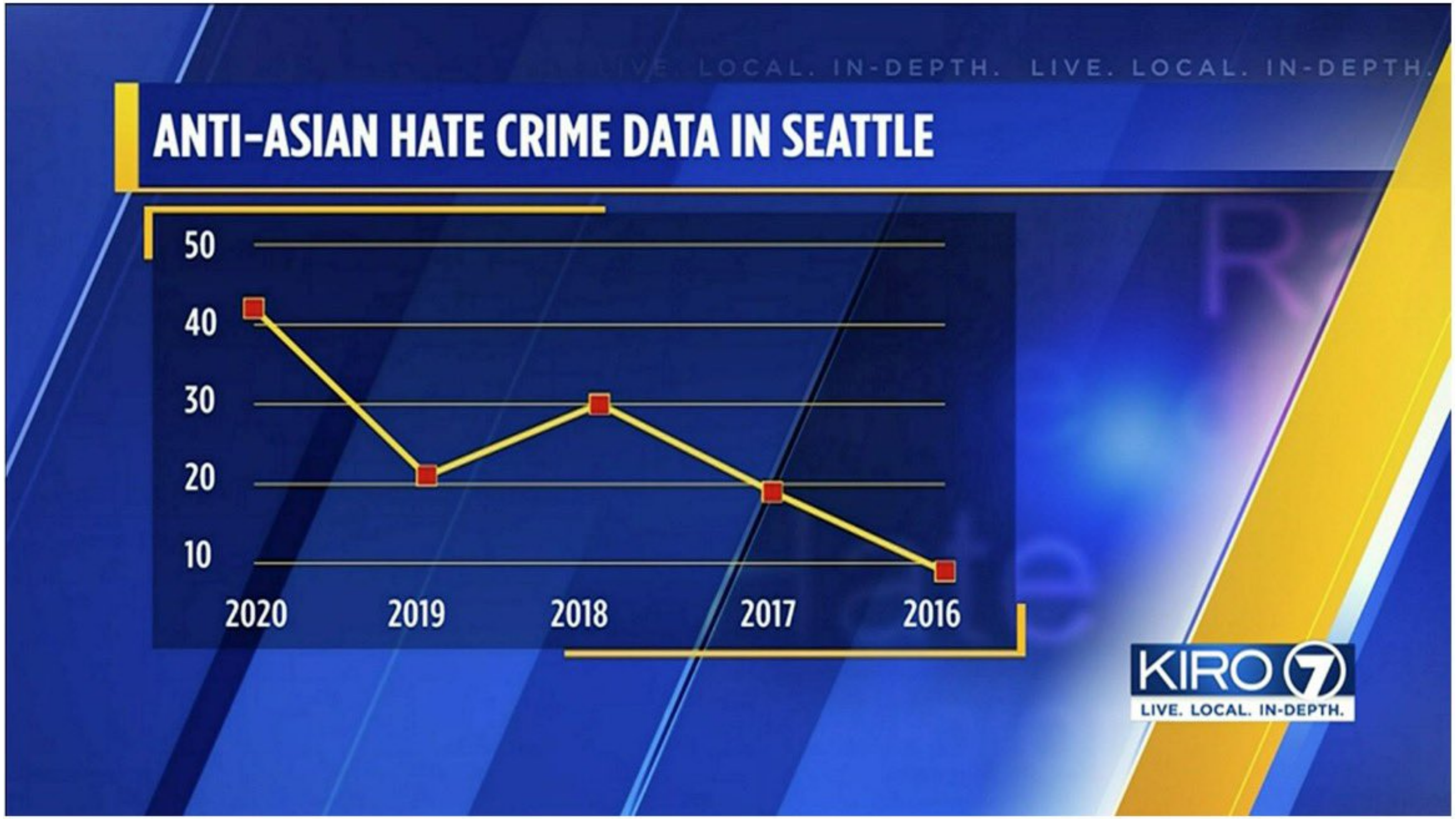


Humans are quite bad at understanding how large big numbers really are

More bad graphics



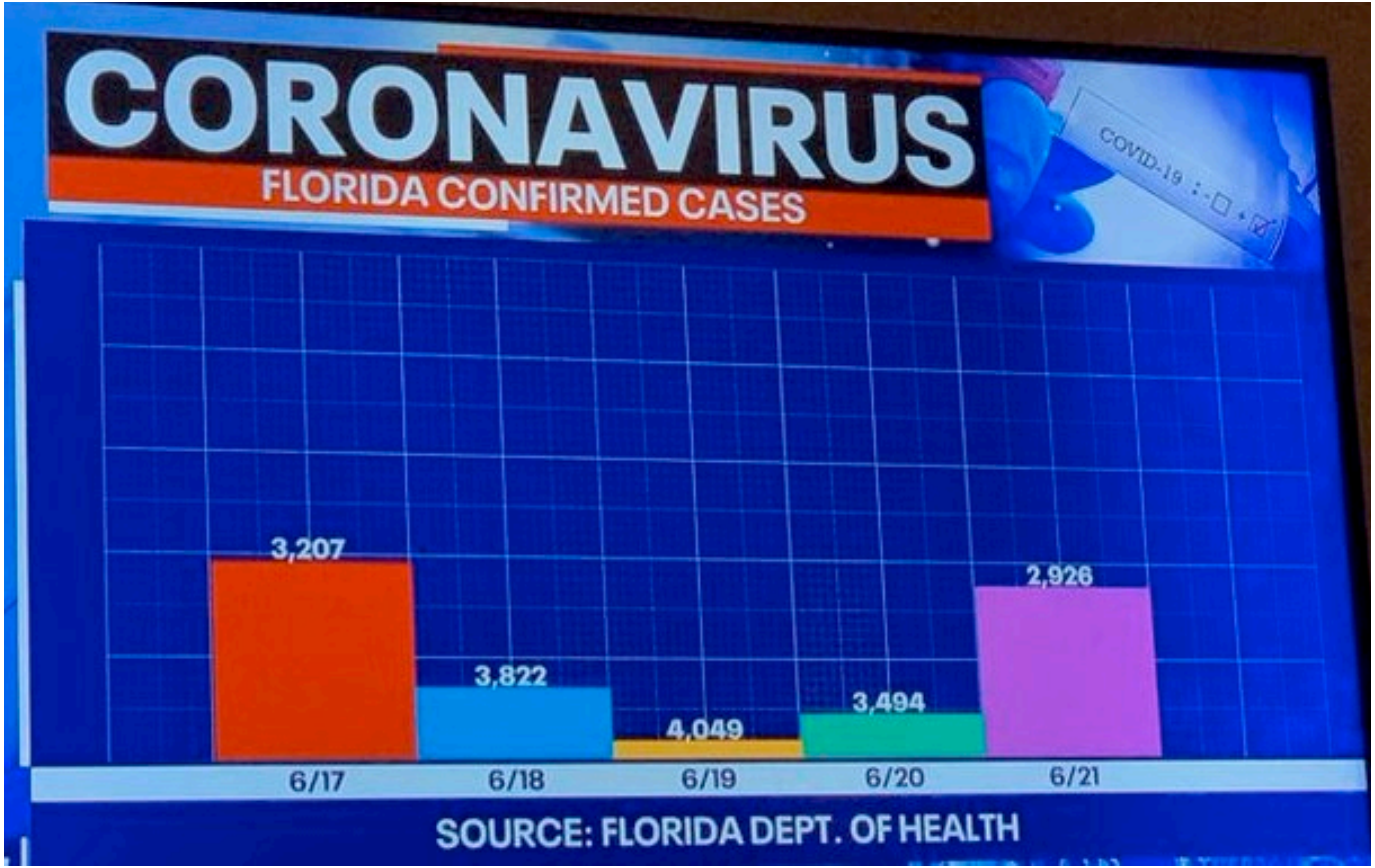
More bad graphics

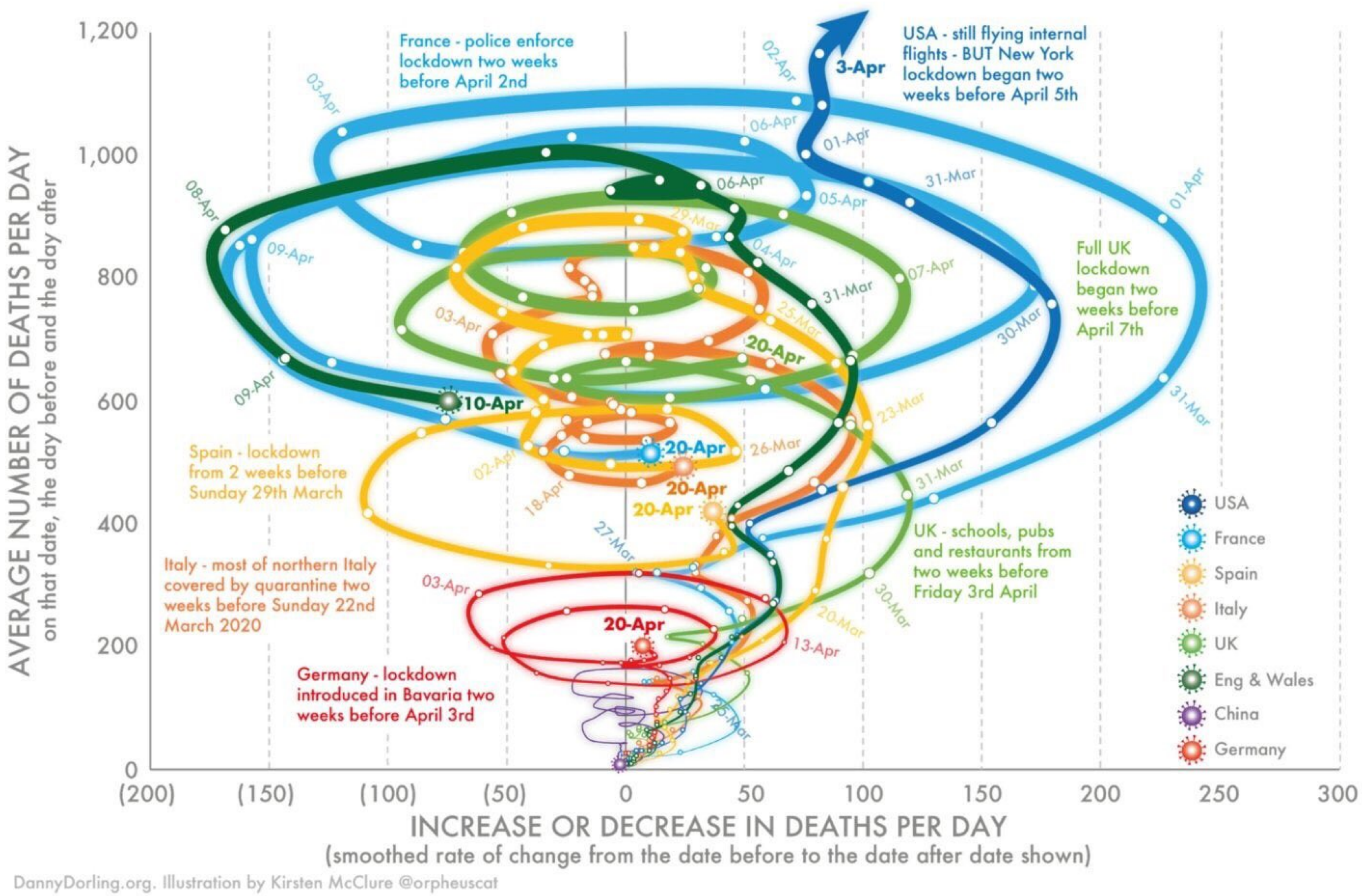


VIDEO: Anti-Asian hate crimes continue in Seattle and across U.S.

COVID yielded new terrible opportunities for terrible graphics

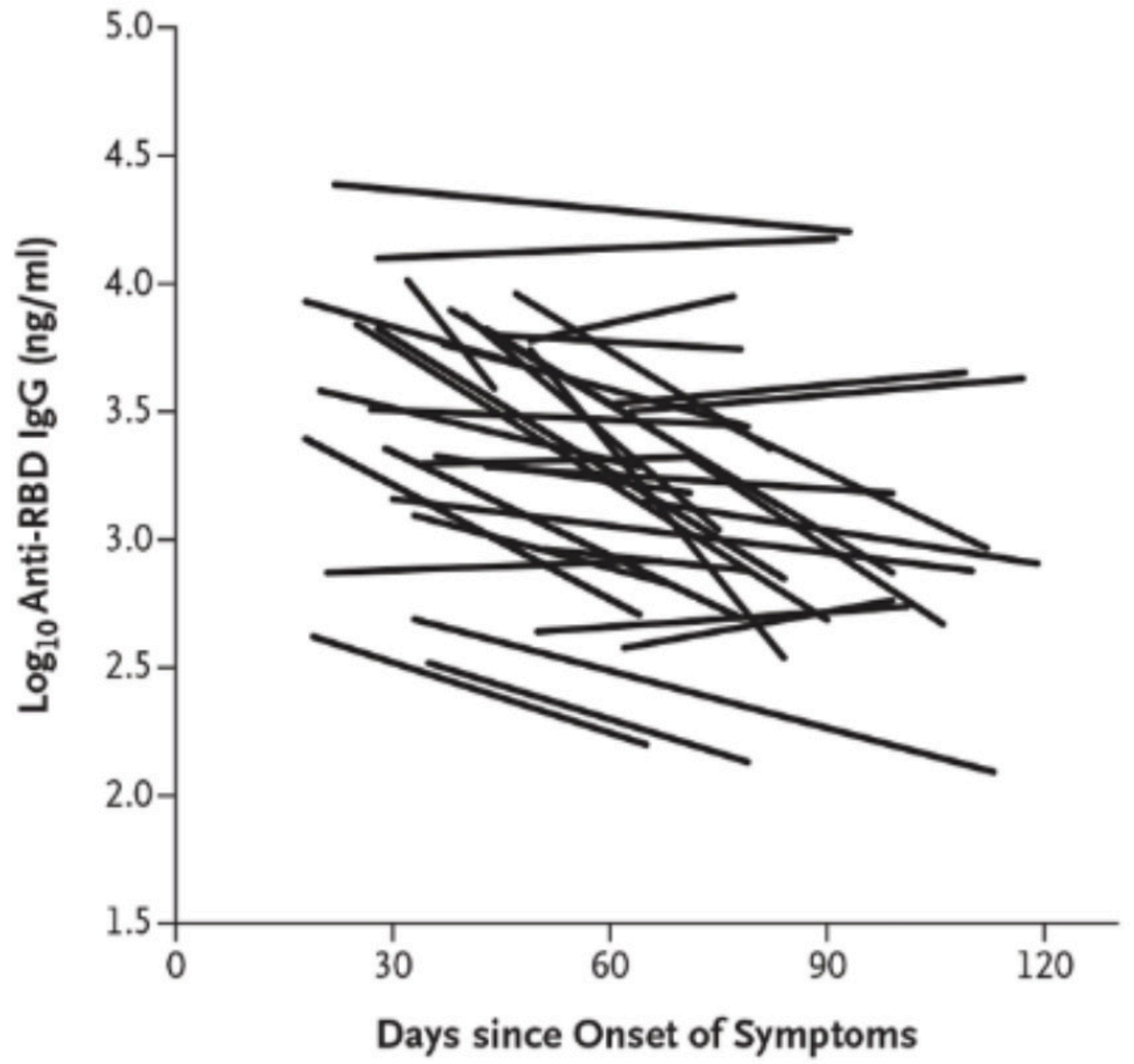
The perfect storm of opportunity - new data and emerging new data tools


CORONAVIRUS
FLORIDA CONFIRMED CASES
COVID-19
3,207
3,822
4,049
3,494
2,926
6/17
6/18
6/19
6/20
6/21
SOURCE: FLORIDA DEPT. OF HEALTH

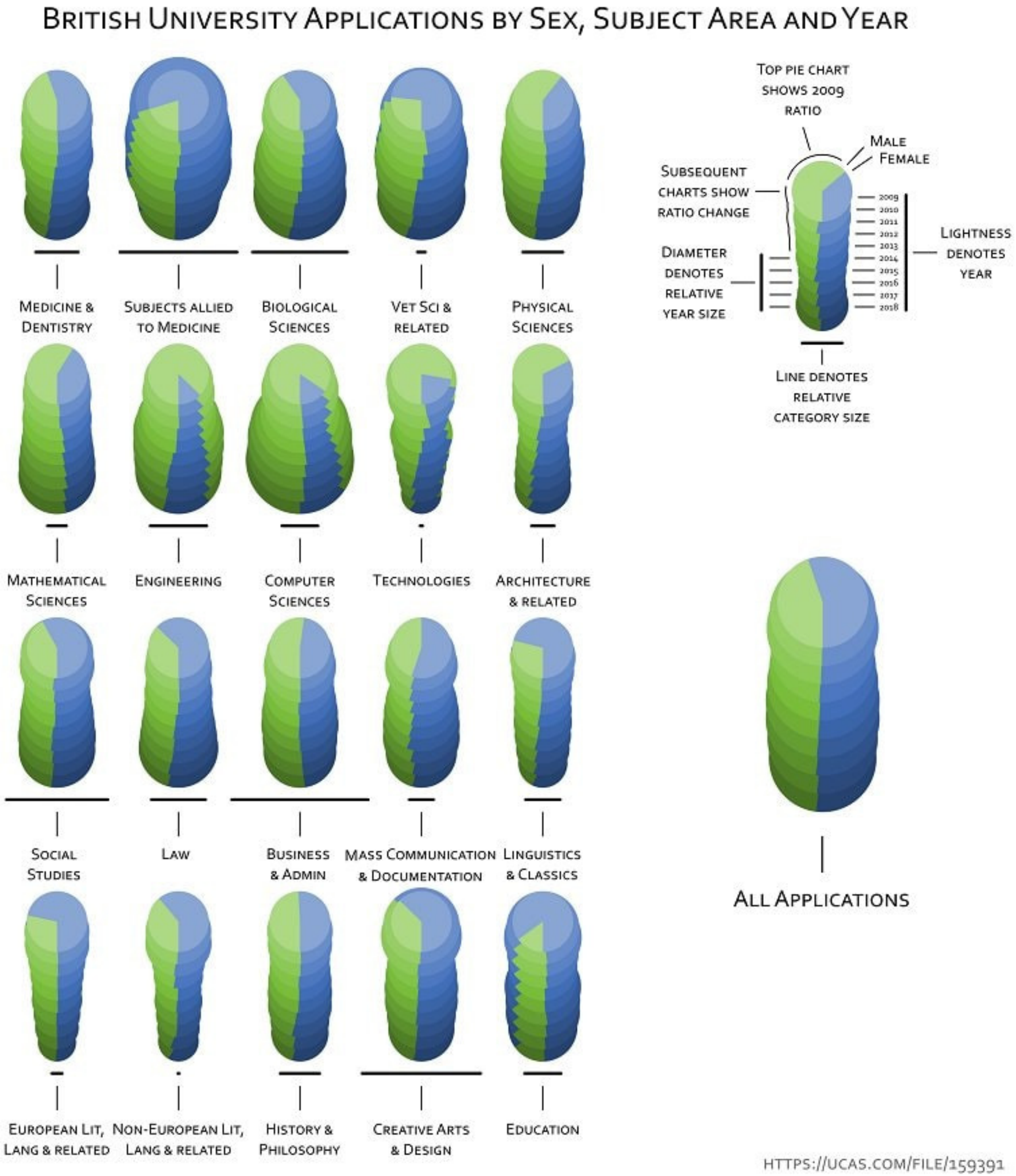

France - police enforce lockdown two weeks before April 2nd
USA - still flying internal flights - BUT New York lockdown began two weeks before April 5th
3-Apr
Full UK lockdown began two weeks before April 7th
Spain - lockdown from 2 weeks before Sunday 29th March
UK - schools, pubs and restaurants from two weeks before Friday 3rd April
Italy - most of northern Italy covered by quarantine two weeks before Sunday 22nd March 2020
Germany - lockdown introduced in Bavaria two weeks before April 3rd
10-Apr
20-Apr
USA
France
Spain
Italy
UK
Eng & Wales
China
Germany
AVERAGE NUMBER OF DEATHS PER DAY
on that date, the day before and the day after
INCREASE OR DECREASE IN DEATHS PER DAY
(smoothed rate of change from the date before to the date after date shown)
DannyDorling.org. Illustration by Kirsten McClure @orpheuscat

NEJM
@NEJM

Correspondence: Rapid Decay of Anti–SARS-CoV-2 Antibodies in Persons with Mild Covid-19 #COVID19 #SARSCoV2

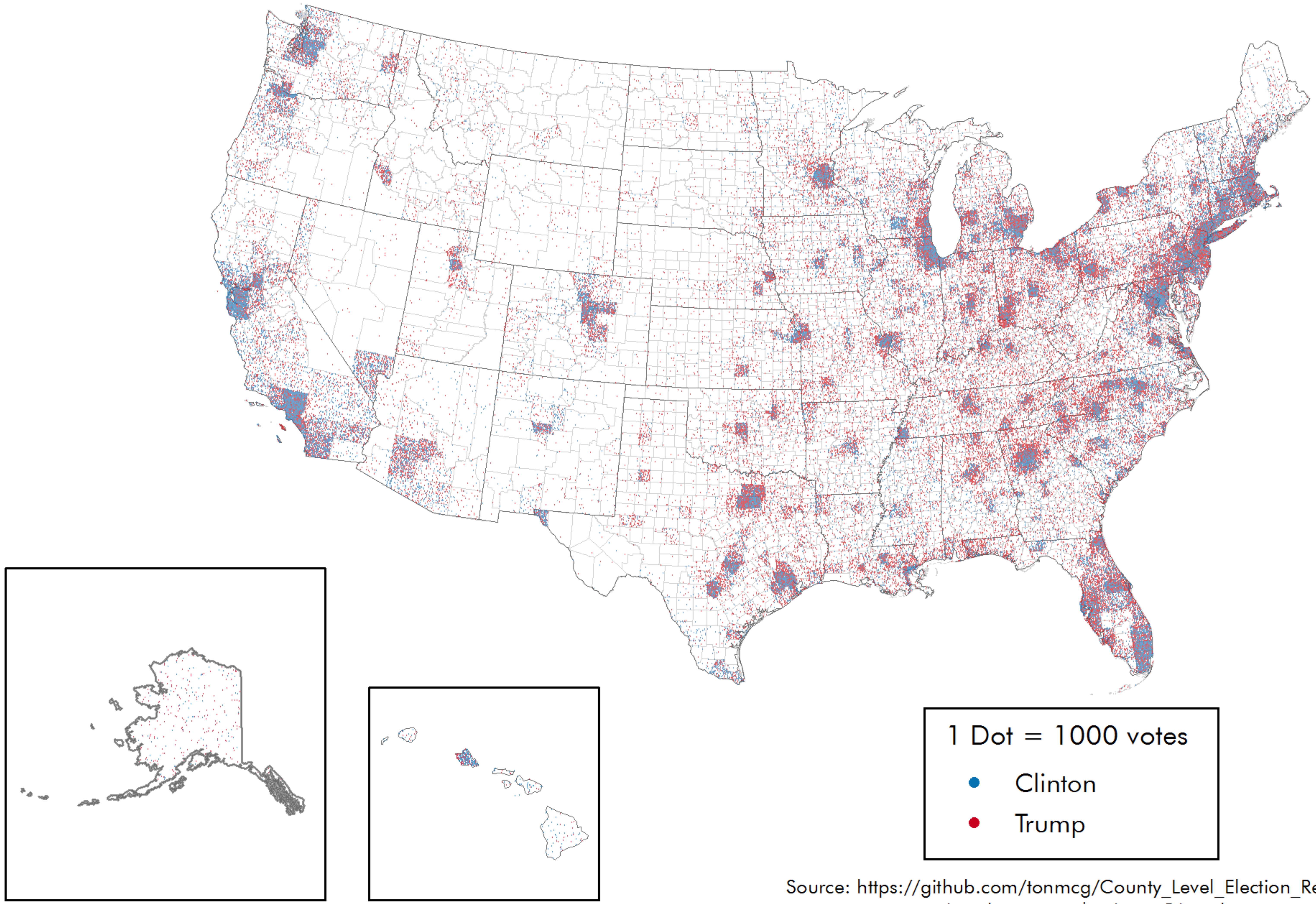


Aggressively Bad



How about some good graphics?

Not As Red As You Think:
2016 Presidential Election Results By County Vote Tally

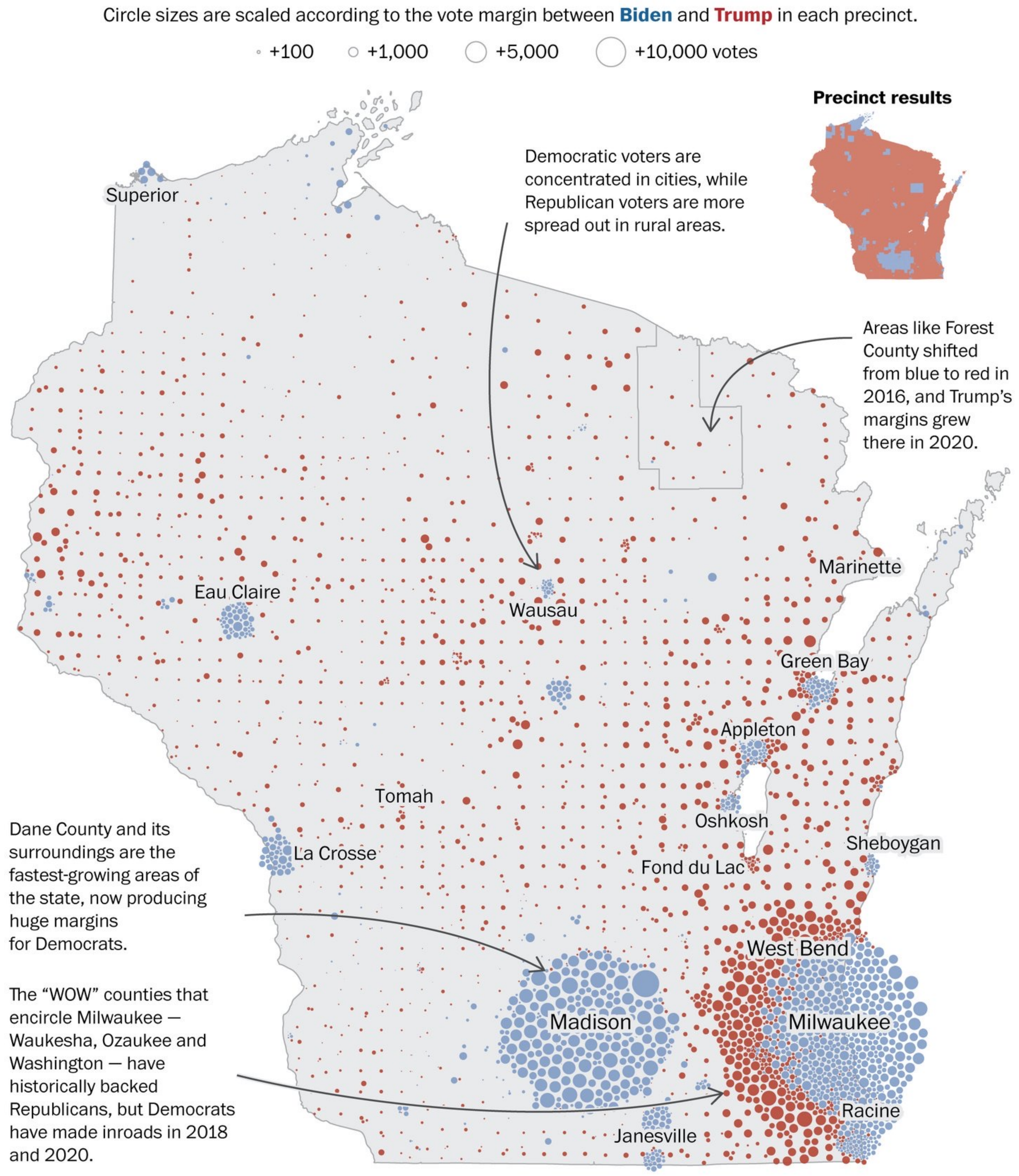


Source: https://github.com/tonmcg/County_Level_Election_Results_12-16
igotcharts.com | twitter: @igotcharts

More *good* graphics

How Wisconsin voted in the 2020 presidential election

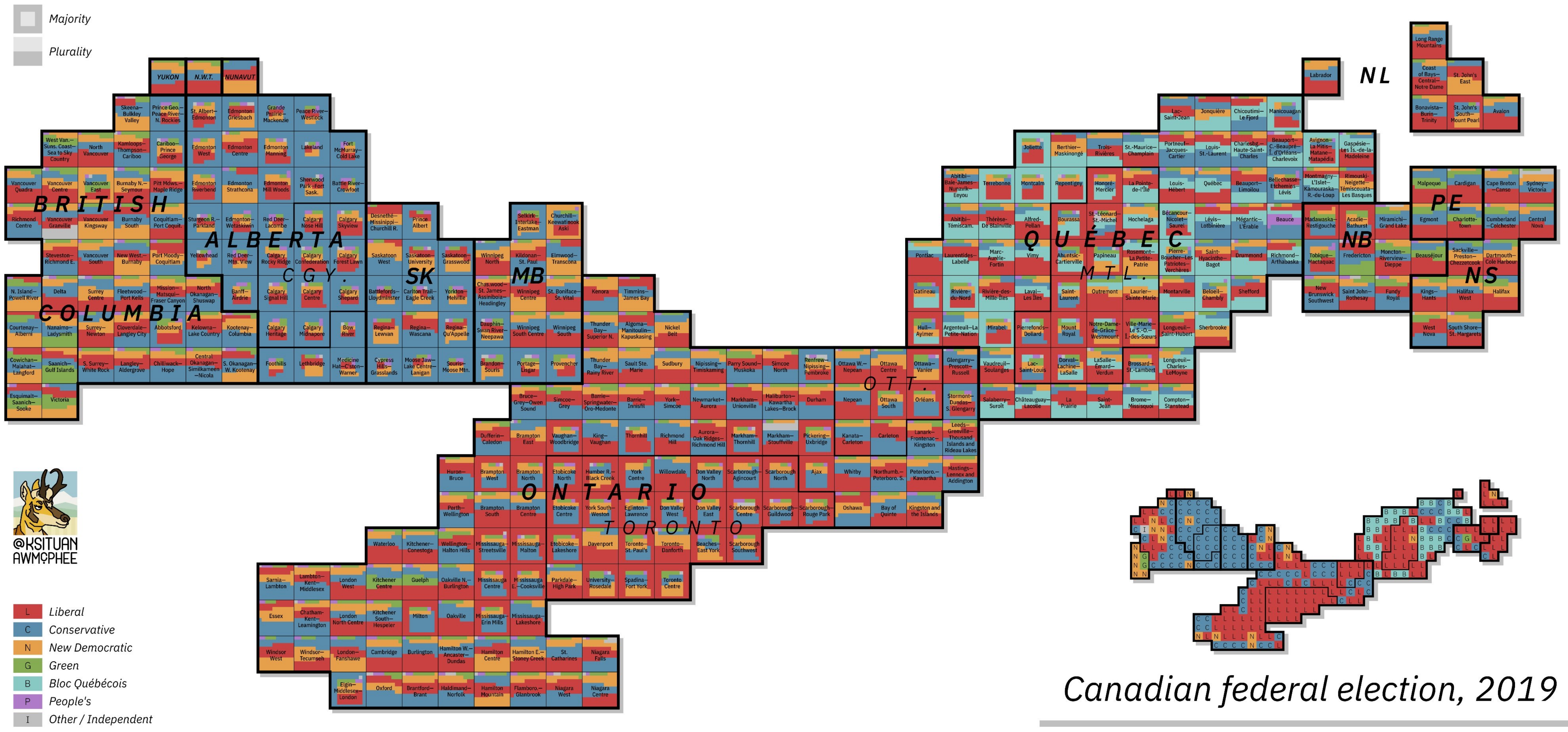
Narrow margins have defined recent presidential elections in Wisconsin. Biden won by just 20,682 votes out of 3.3 million cast. Four years earlier, Trump had bested Hillary Clinton by only 22,748 of almost 3 million votes cast.



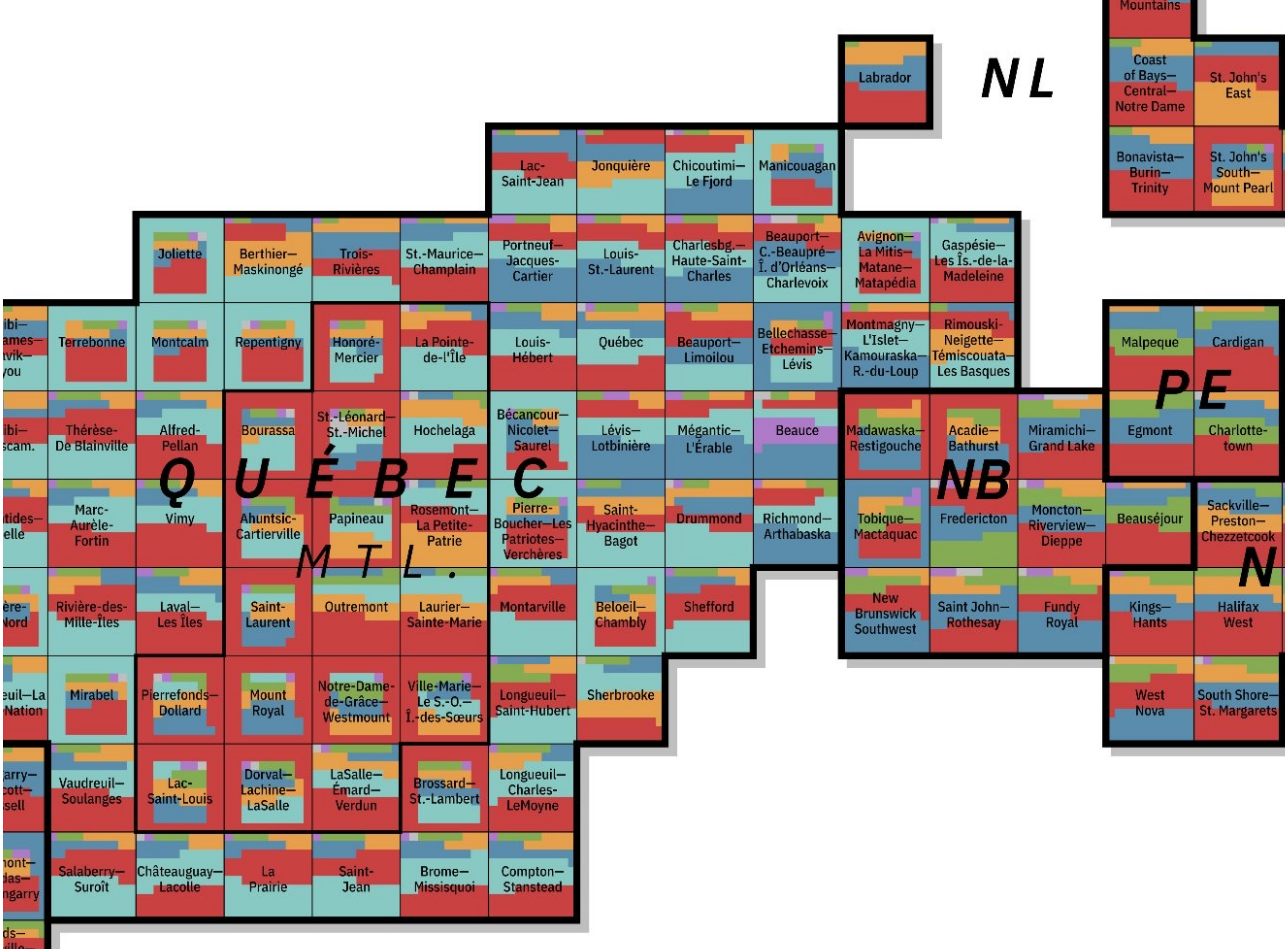
Note: Circles are clustered in areas with many precincts, so locations are approximate.

Source: Decision Desk HQ

ADRIÁN BLANCO AND SHELLY TAN/THE WASHINGTON POST



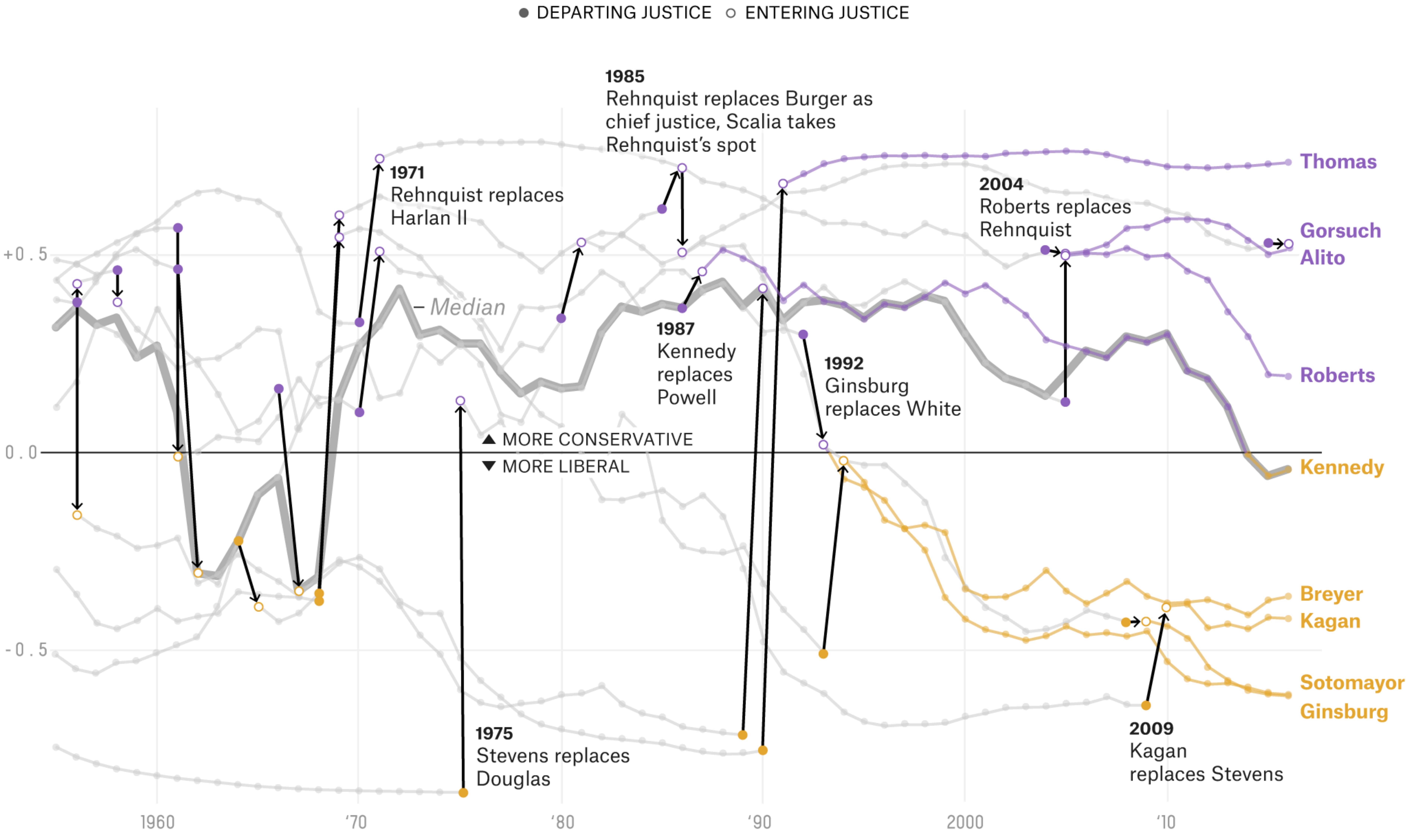
Canadian federal election, 2019



NL
Mountains
Labrador
Coast of Bays—Central—Notre Dame
St. John's East
Bonavista—Burin—Trinity
St. John's South—Mount Pearl
Lac-Saint-Jean
Jonquière
Chicoutimi—Le Fjord
Manicouagan
Joliette
Berthier—Maskinongé
Trois-Rivières
St.-Maurice—Champlain
Portneuf—Jacques-Cartier
Louis-St.-Laurent
Charlesbg.—Haute-Saint-Charles
Beauport—C.-Beaupré—Î. d'Orléans—Charlevoix
Avignon—La Mitis—Matane—Matapédia
Gaspésie—Les Îs.-de-la-Madeleine
Terrebonne
Montcalm
Repentigny
Honoré-Mercier
La Pointe-de-l'Île
Louis-Hébert
Québec
Beauport—Limoilou
Bellechasse—Etchemins—Lévis
Montmagny—L'Islet—Kamouraska—R.-du-Loup
Rimouski-Neigette—Témiscouata—Les Basques
Malpeque
Cardigan
PE
Thérèse-De Blainville
Alfred-Pellan
Bourassa
St.-Léonard—St.-Michel
Hochelaga
Bécancour—Nicolet—Saurel
Lévis—Lotbinière
Mégantic—L'Érable
Beauce
Madawaska—Restigouche
Acadie—Bathurst
Miramichi—Grand Lake
Egmont
Charlotte-town
QUÉBEC
NB
Marc-Aurèle-Fortin
Vimy
Ahuntsic-Cartierville
Papineau
Rosemont—La Petite-Patrie
Pierre-Boucher—Les Patriotes—Verchères
Saint-Hyacinthe—Bagot
Drummond
Richmond—Arthabaska
Tobique—Mactaquac
Fredericton
Moncton—Riverview—Dieppe
Beauséjour
Sackville—Preston—Chezzetcook
MTL.
N
Rivière-des-Mille-Îles
Laval—Les Îles
Saint-Laurent
Outremont
Laurier—Sainte-Marie
Montarville
Beloeil—Chambly
Shefford
New Brunswick Southwest
Saint John—Rothesay
Fundy Royal
Kings—Hants
Halifax West
Mirabel
Pierrefonds—Dollard
Mount Royal
Notre-Dame-de-Grâce—Westmount
Ville-Marie—Le S.-O.—Î.-des-Sœurs
Longueuil—Saint-Hubert
Sherbrooke
West Nova
South Shore—St. Margarets
Vaudreuil—Soulanges
Lac-Saint-Louis
Dorval—Lachine—LaSalle
LaSalle—Émard—Verdun
Brossard—St.-Lambert
Longueuil—Charles-LeMoyne
Salaberry—Suroît
Châteauguay—Lacolle
La Prairie
Saint-Jean
Brome—Missisquoi
Compton—Stanstead

Supreme Court departures and replacements

Supreme Court justices from 1955-2016 by their ideological leanings, based on their Judicial Common Space scores



Years refer to Supreme Court terms, which run from October to September.

FiveThirtyEight

SOURCES: LEE EPSTEIN, CHAD WESTERLAND, THE JOURNAL OF LAW, ECONOMICS, AND ORGANIZATION

Bivariate Graphics

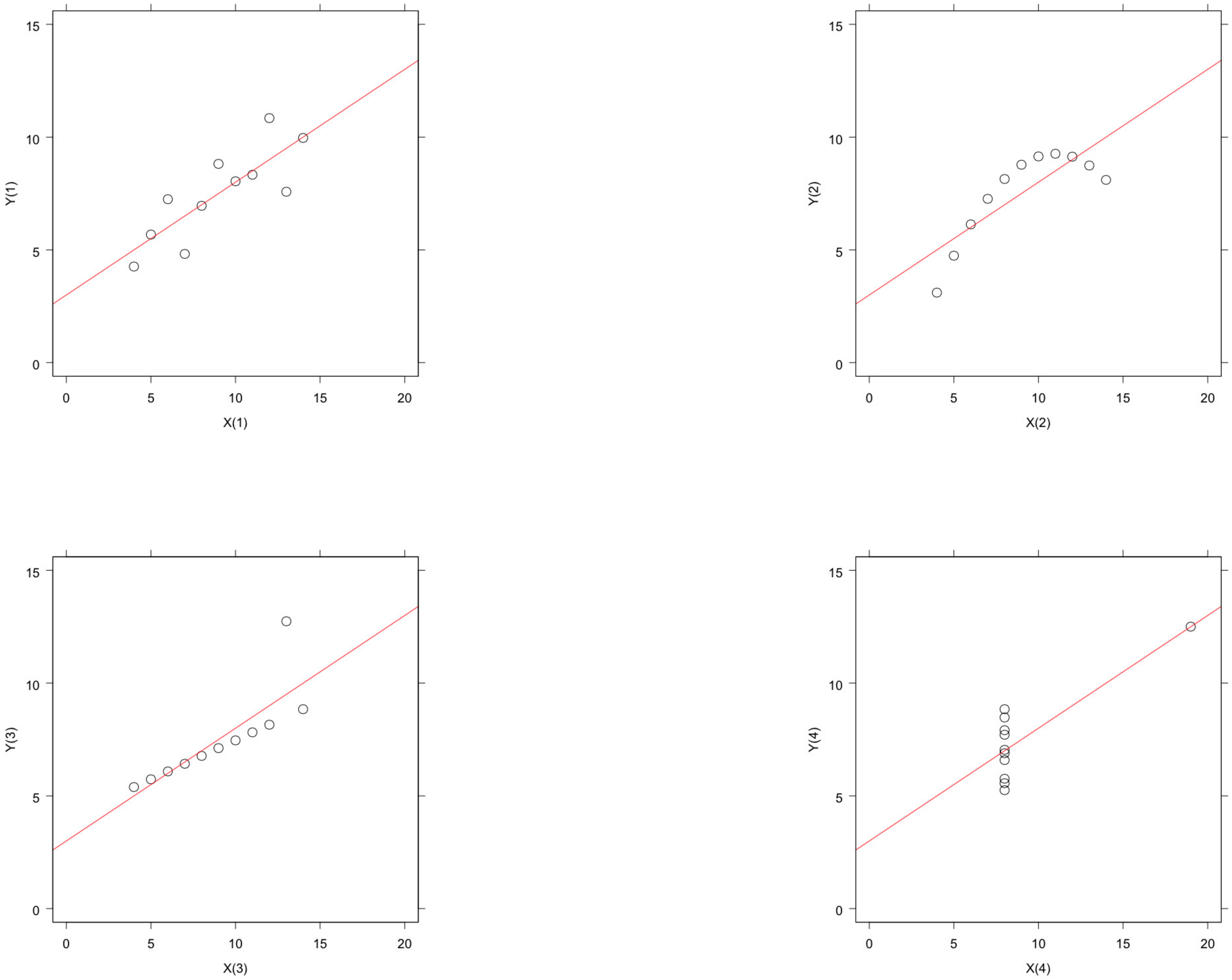
- Last time: looked at univariate graphics, and a number of different ways to examine univariate graphics
  - Now we're back!
- First . . . why should we care about bivariate graphics?
  - To convince you . . .

Equivalent OLS Estimates

- $Y_i = 3.00 + .5X + e_i$
- $R^2 = .67$
- Standard Errors: (1.12) and (0.12)

the critical statistics for each regression are ***identical***

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



Y(1)
X(1)
Y(2)
X(2)
Y(3)
X(3)
Y(4)
X(4)

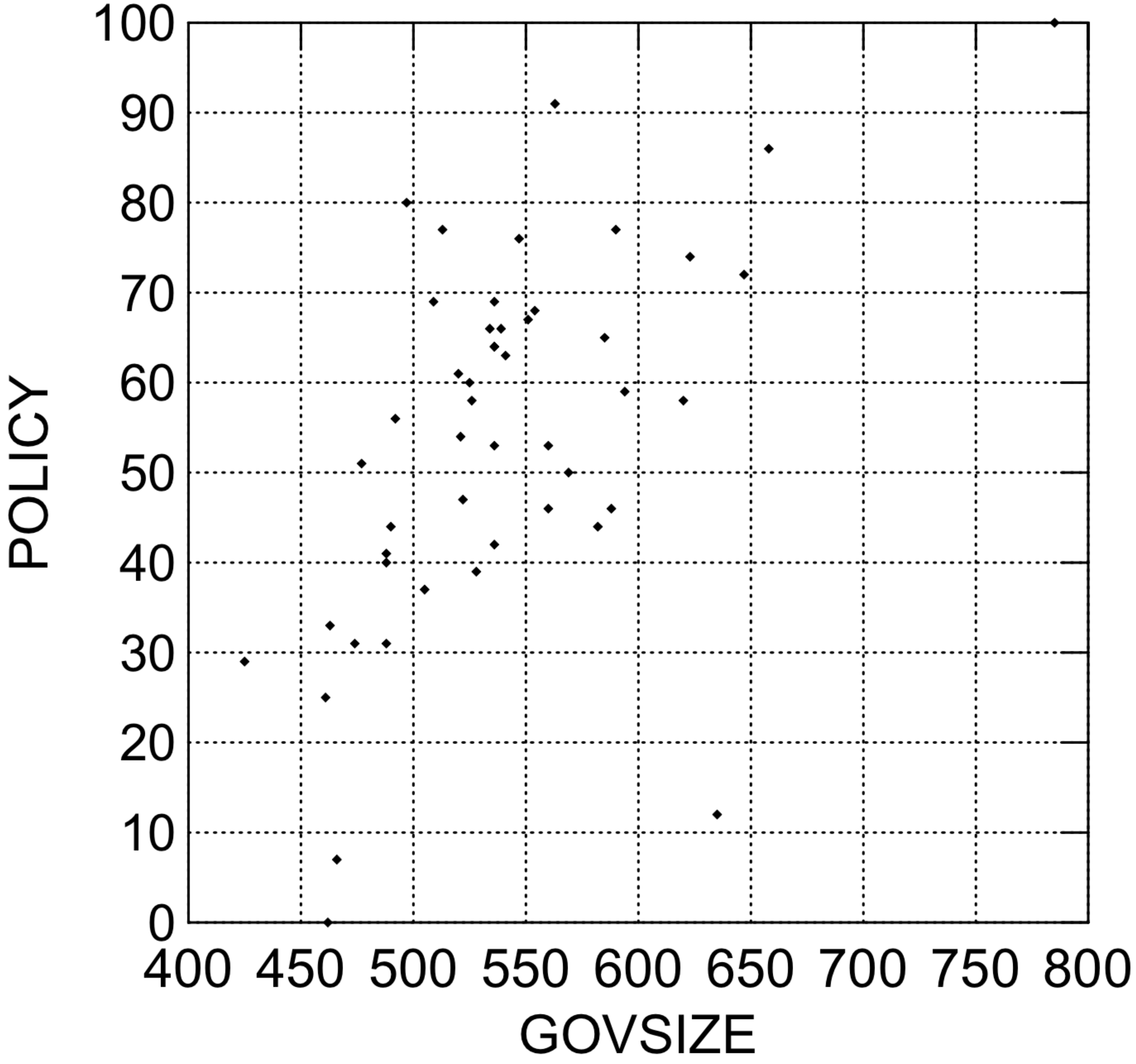
Aw, shucks. Silly OLS.

- These datasets, despite what the OLS regressions tell us, are not remotely identical.
    - *You should be starting to get suspicious of unexamined OLS estimates …*
- Graphics can help us diagnose many significant issues with OLS stemming from nonlinearity, non spherical error terms, and outliers
- And bivariate graphics can even help us with truly nonlinear data!

So, how are we going to do this?

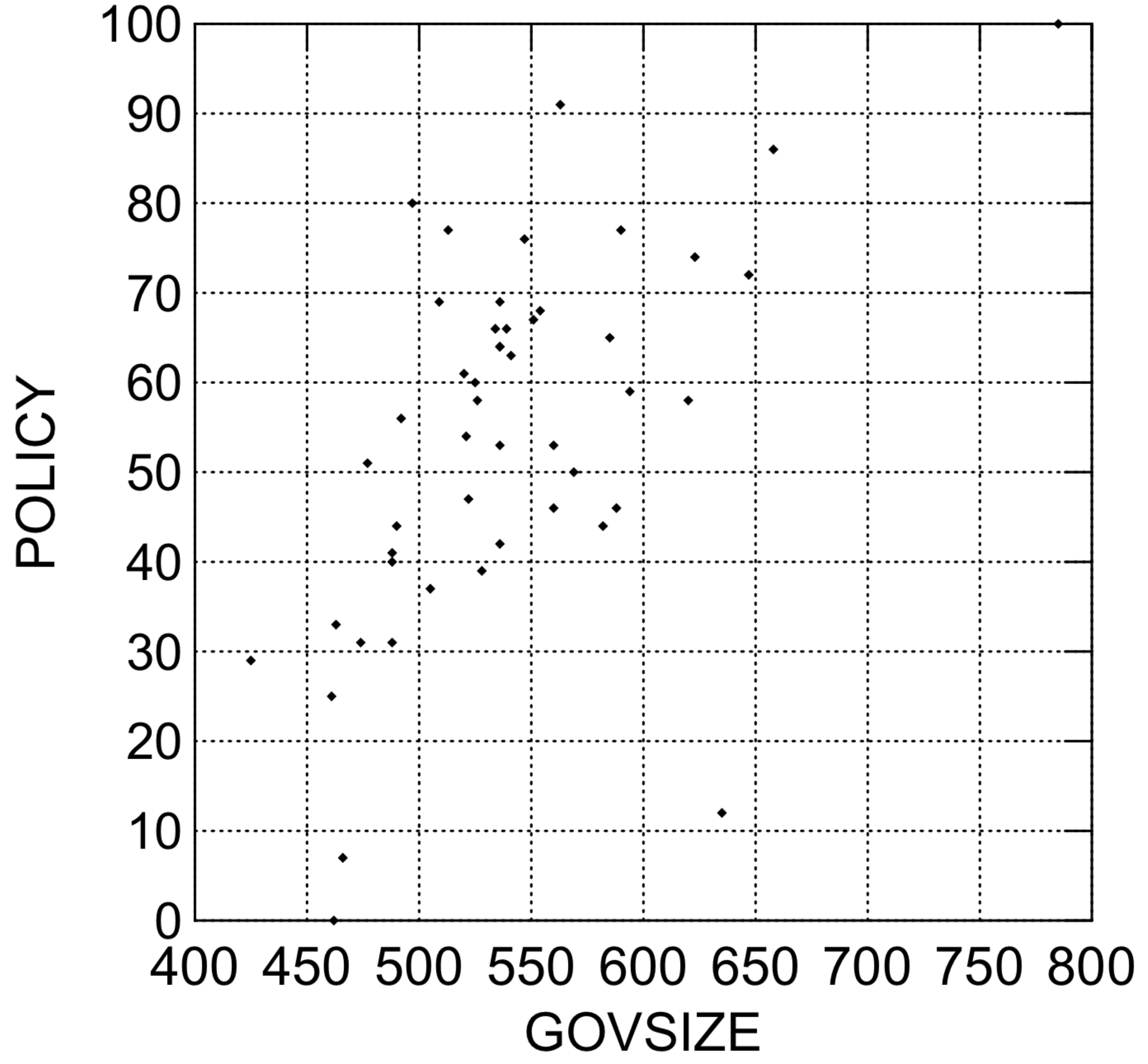
- Guidelines for bivariate graphs follow same conventions as univariate graphics.
  - Keep graphs clean and simple, and focus on your data.
  - Follow cognitive guidelines and rules as before; eliminate clutter, think about the **data-ink ratio**.
- Like univariate graphs, bivariate graphs have a dominant citizen - the **scatterplot**. Like histograms, scatterplots can be drawn well . . . or poorly.

An ugly scatterplot



Aggressively difficult to see properly

Why is it so ugly?

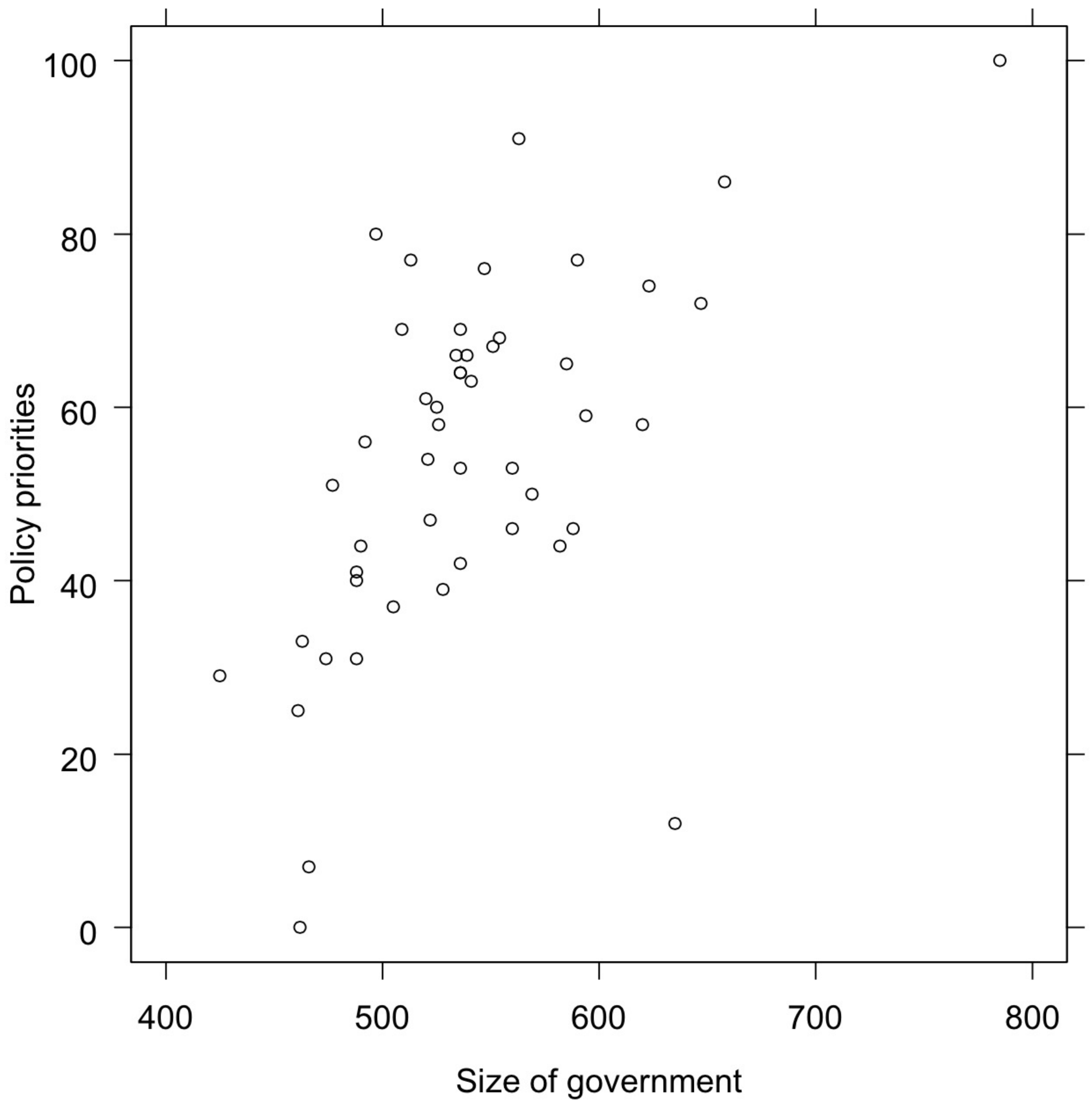


- Doesn’t follow the rules!
  - Why all those gridlines?
  - Why so many so many ticks?
  - Origin labels jammed together
  - LABELS ANGRY! YELLING AT YOU! SHORTENED TOO! FOR NO GOOD REASON!
  - Where are those data points again? Can you squint and see them?
  - What is this even a graph of? Curling policy?
- ***Data-ink ratio is poor***

How do I make not-ugly graphs?

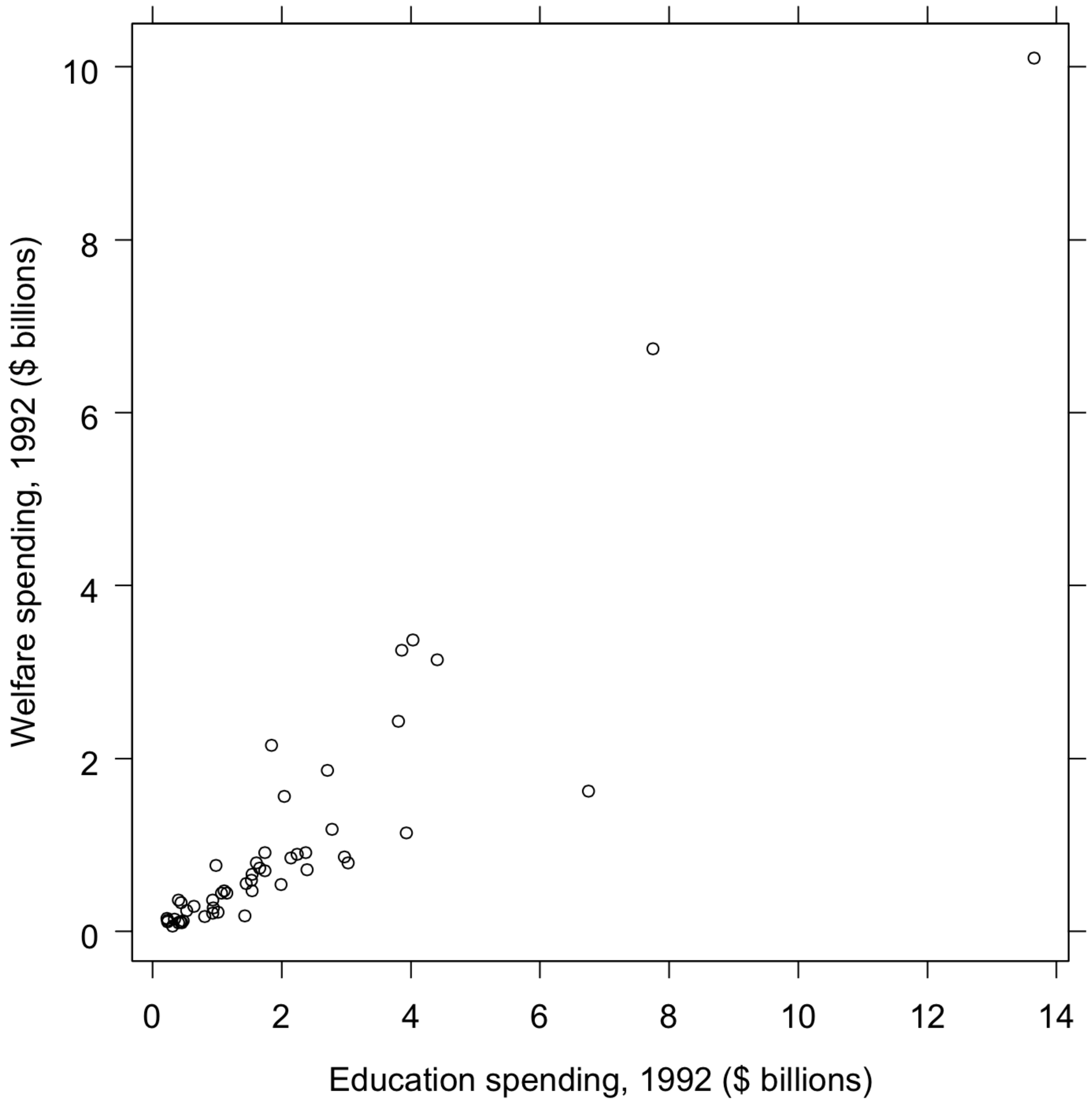
- Use axes to enclose the plot region
- Clear axis labels
- Rectangular grid lines within the plot region are unnecessary
- Plotting symbols should be visually prominent (data!)
- Tick marks point outward, not inward over the plot region
- Few tick marks on each axis
- Data rectangle should be slightly smaller than the scale rectangle
- When (and if) necessary, transform data values so plotted points fill up the data rectangle more fully
- Labels should be written in plain language

A not as ugly scatterplot

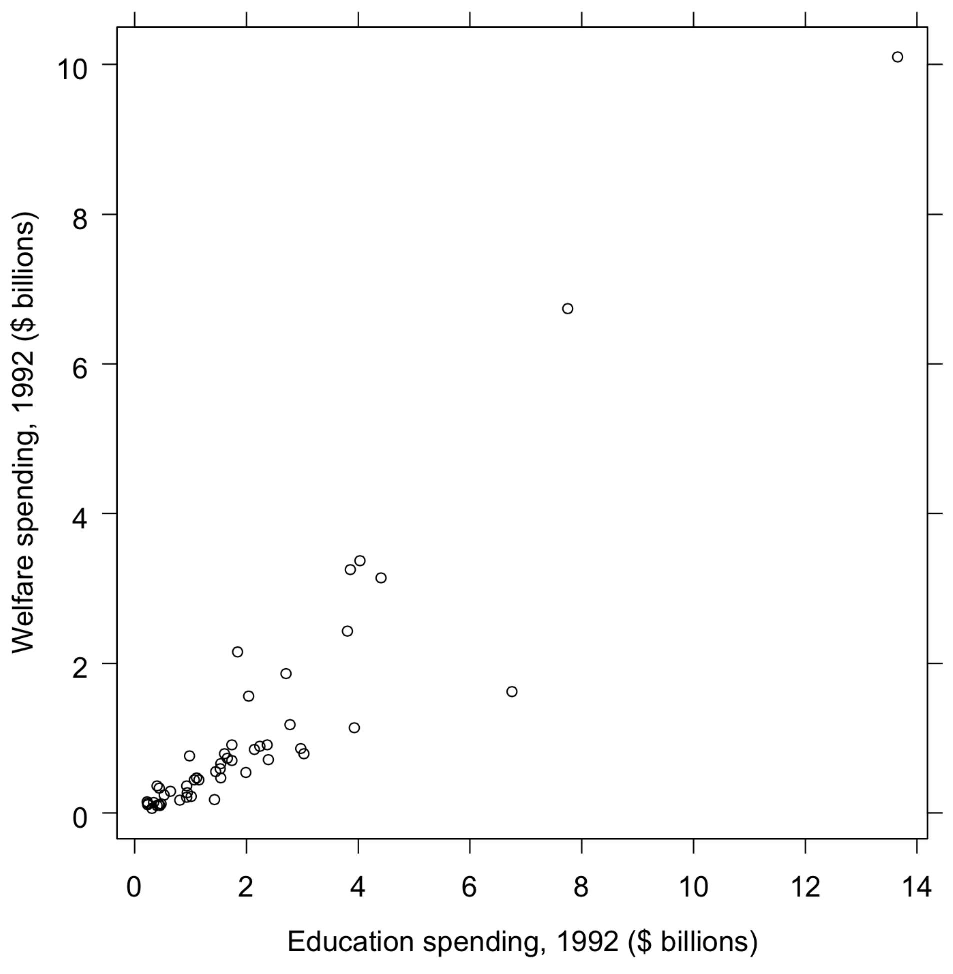


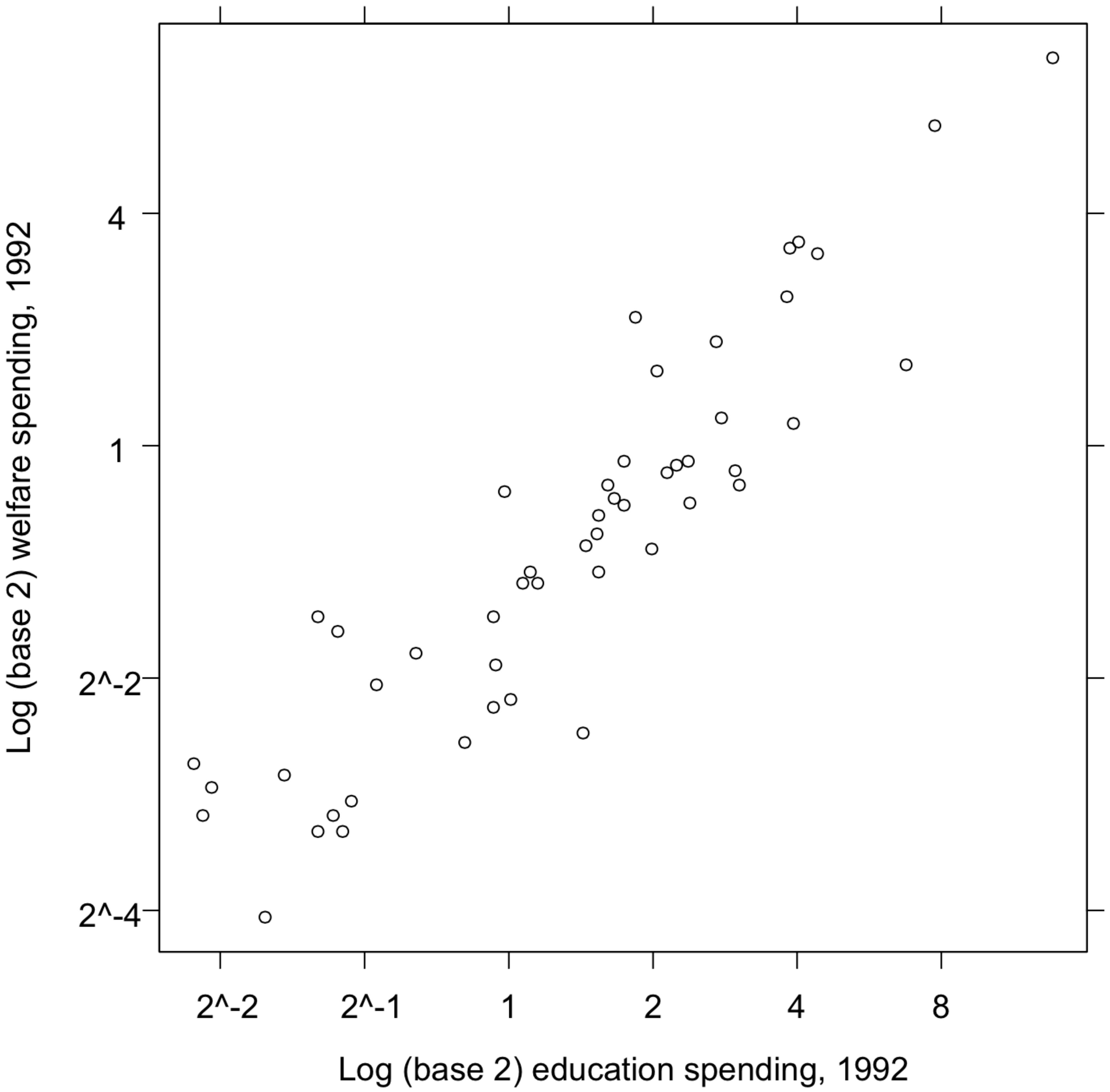
Ahhh . . .

Transforming variables



Transforming variables



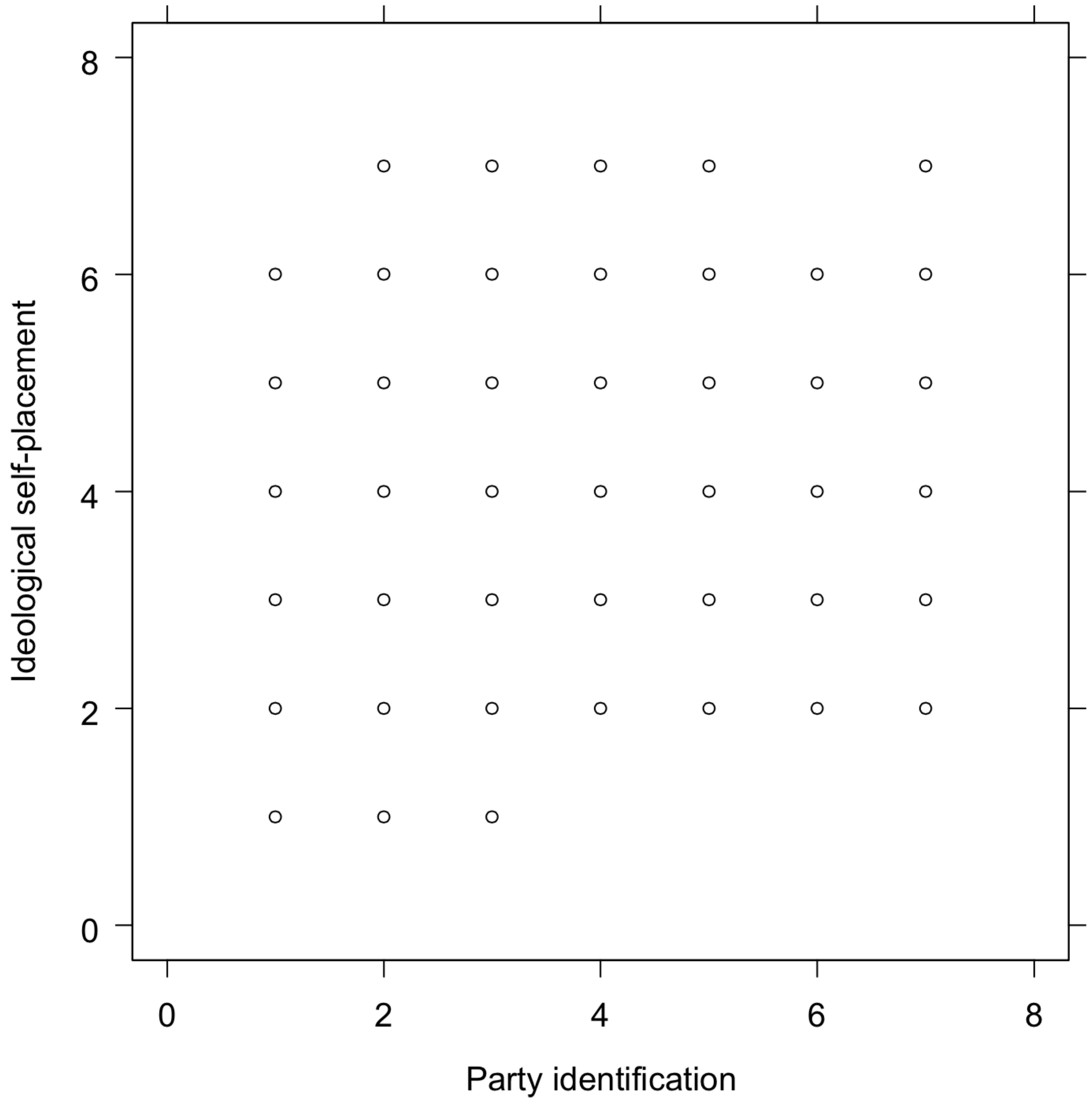


Enhancements

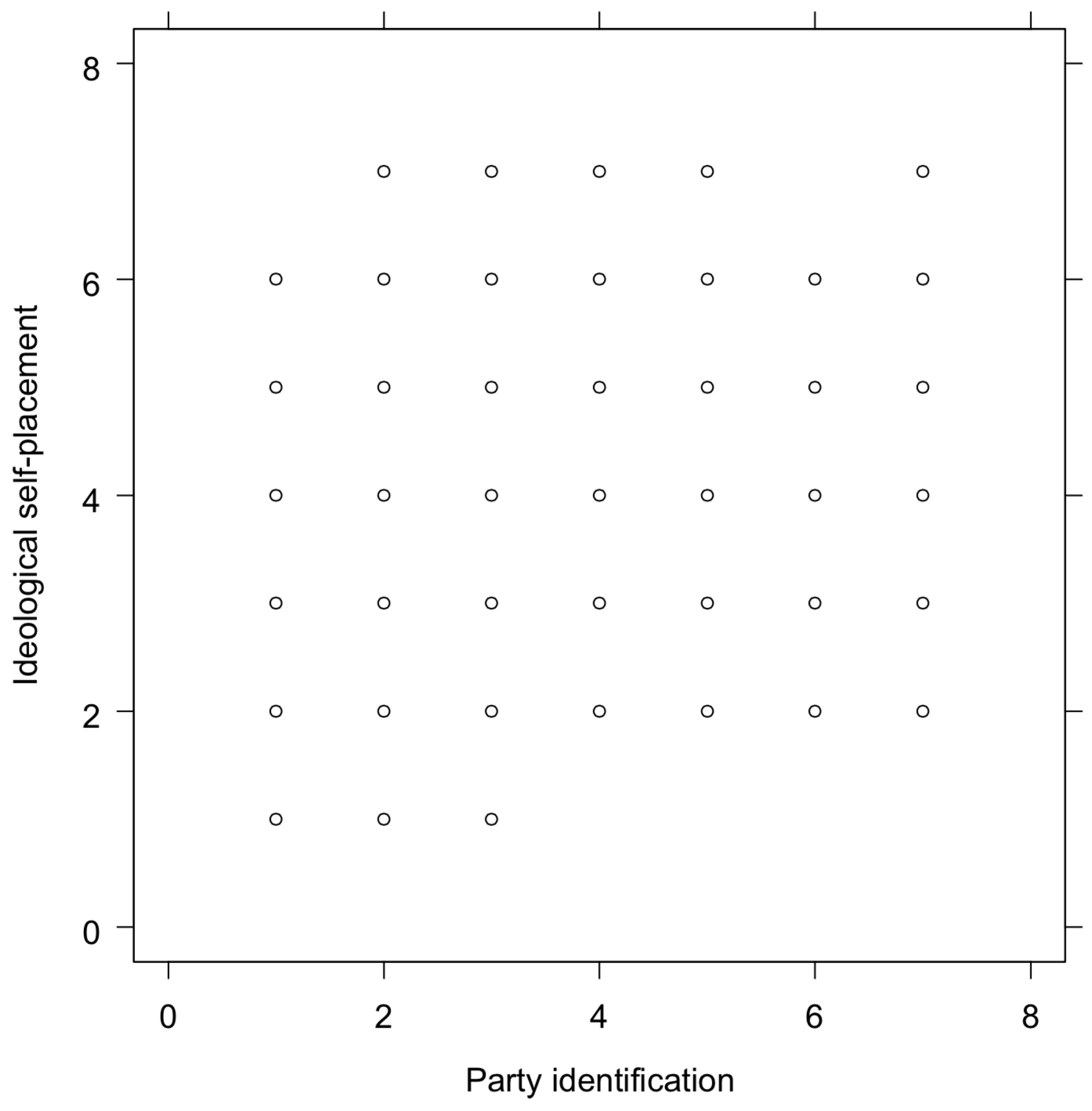
- How else can we make better looking, more informative scatterplots?
  - Jitter data to prevent overplotting and deal appropriately with repeated data points
  - Marginal displays of univariate information (box plots!)
  - Label data points (but use sparingly - data-ink ratio)

Jittering

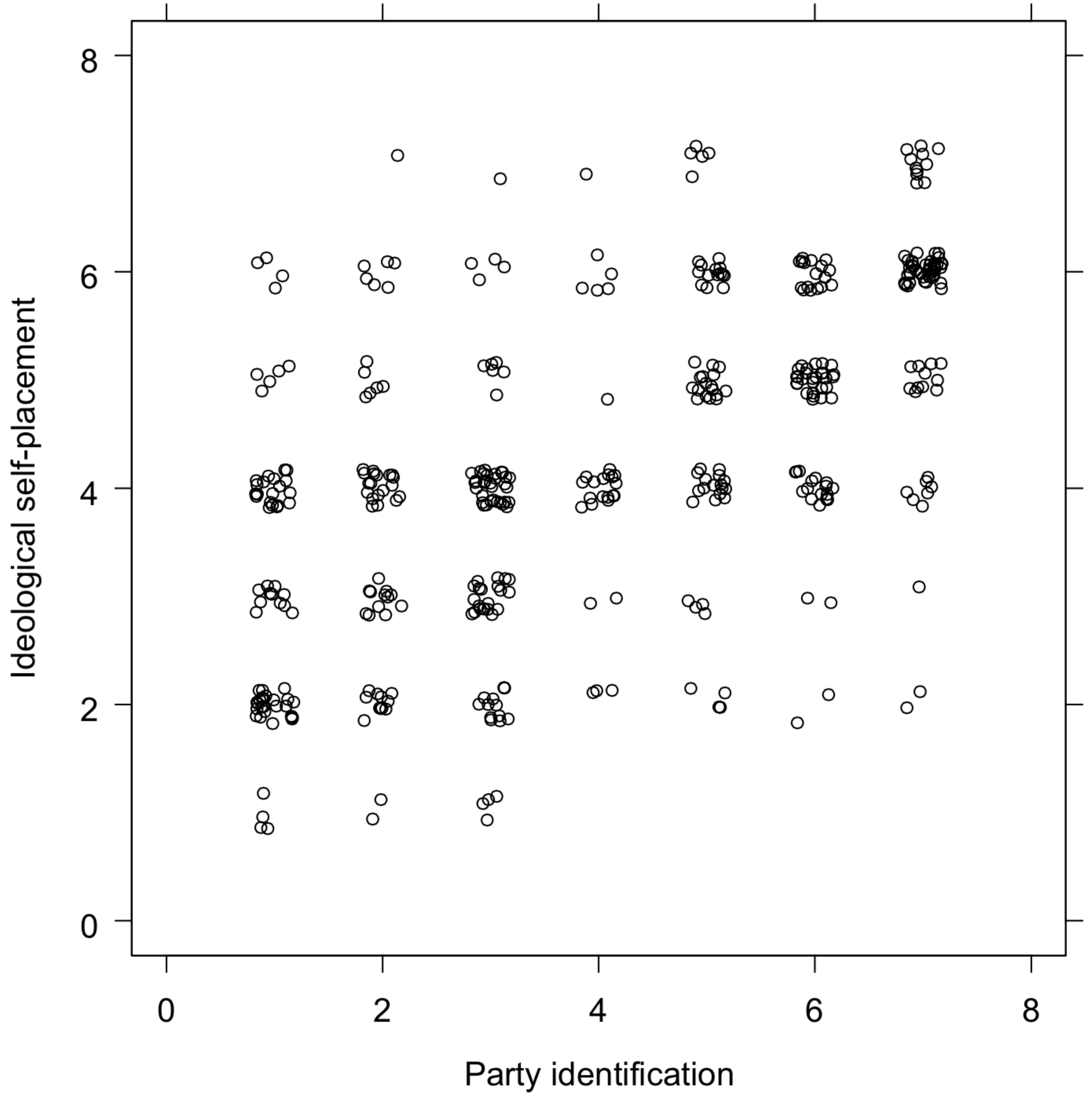
- What's going on in this figure?



Jittering

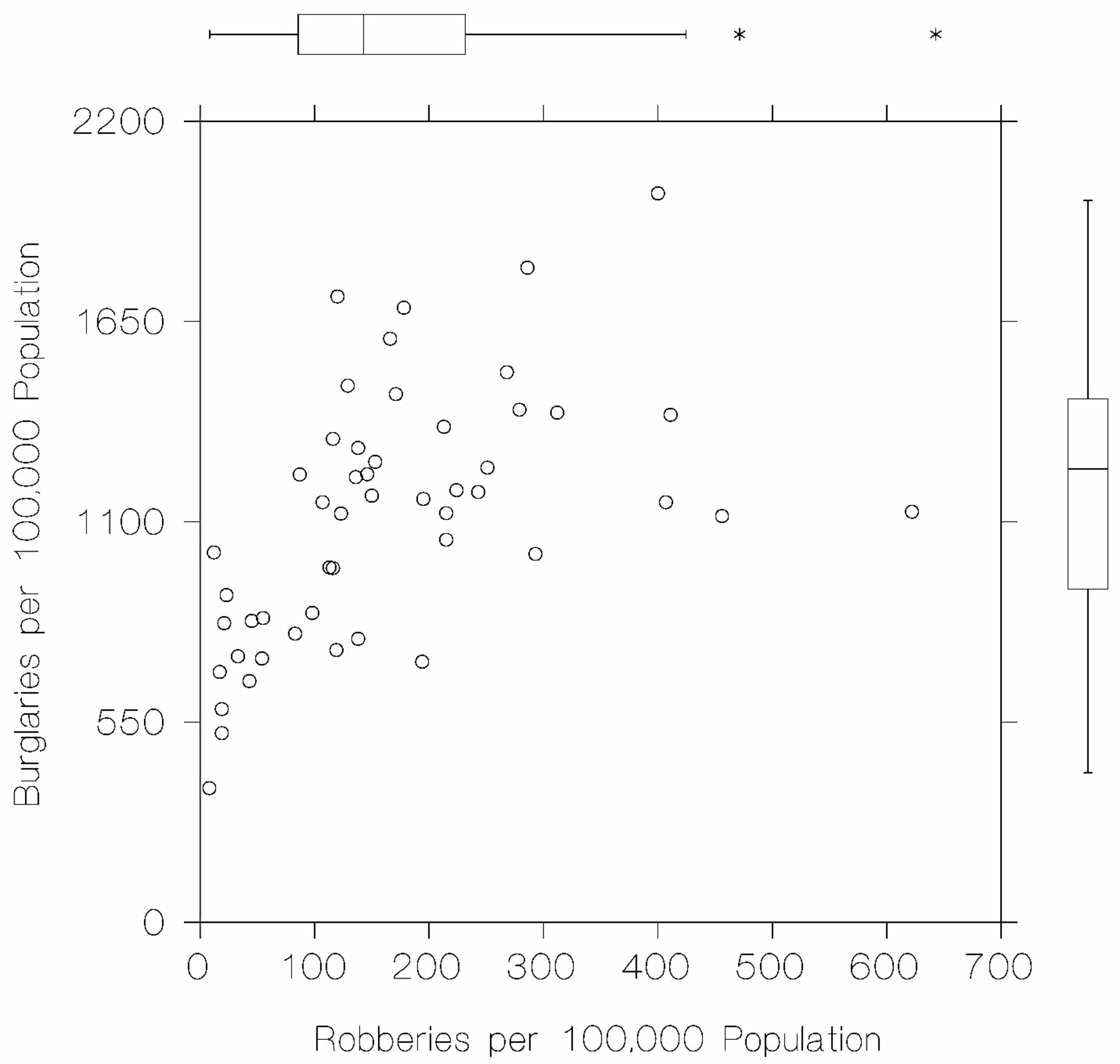


Ideological self-placement
Party identification
0
2
4
6
8

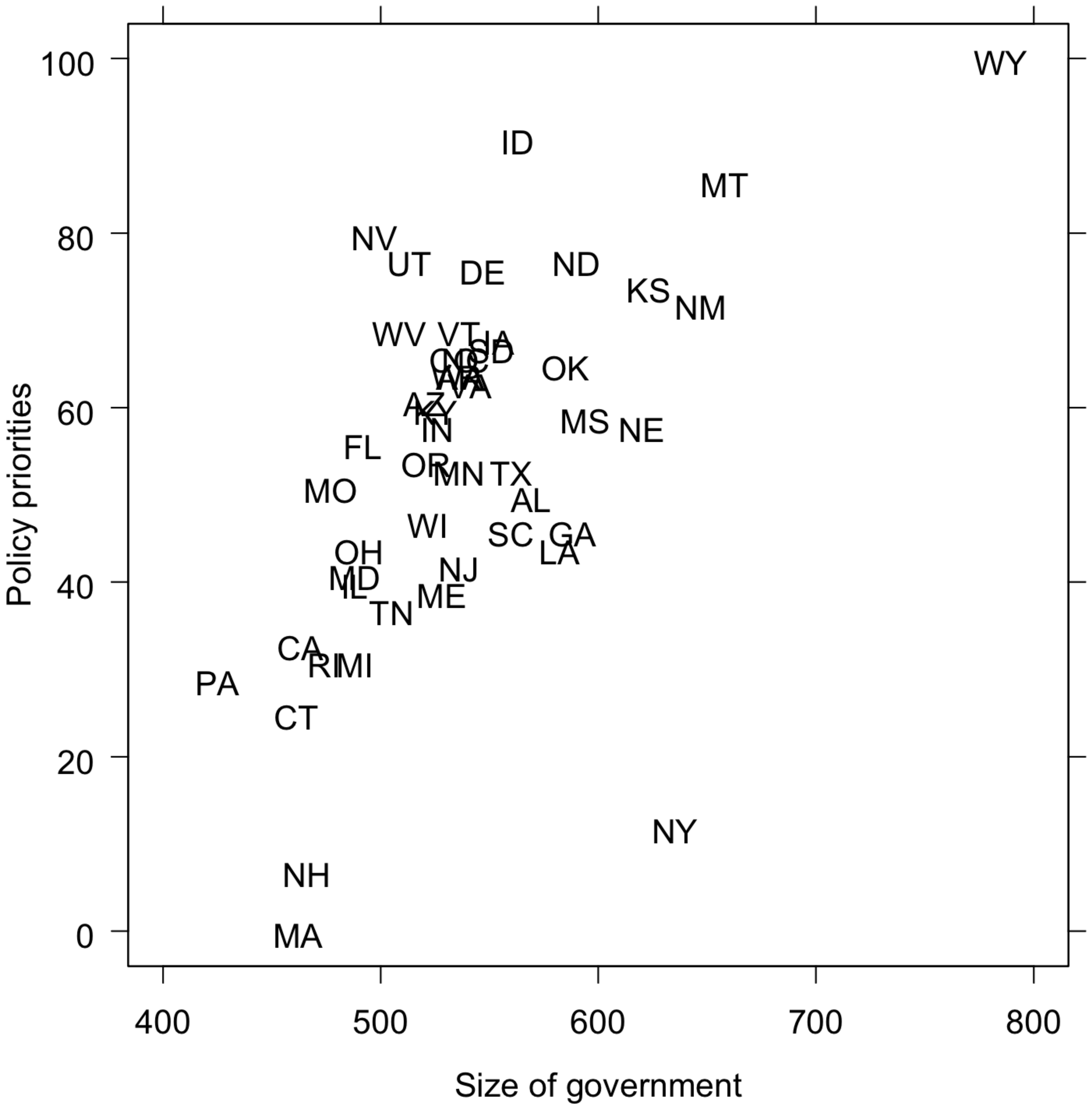


Ideological self-placement
Party identification
0
2
4
6
8

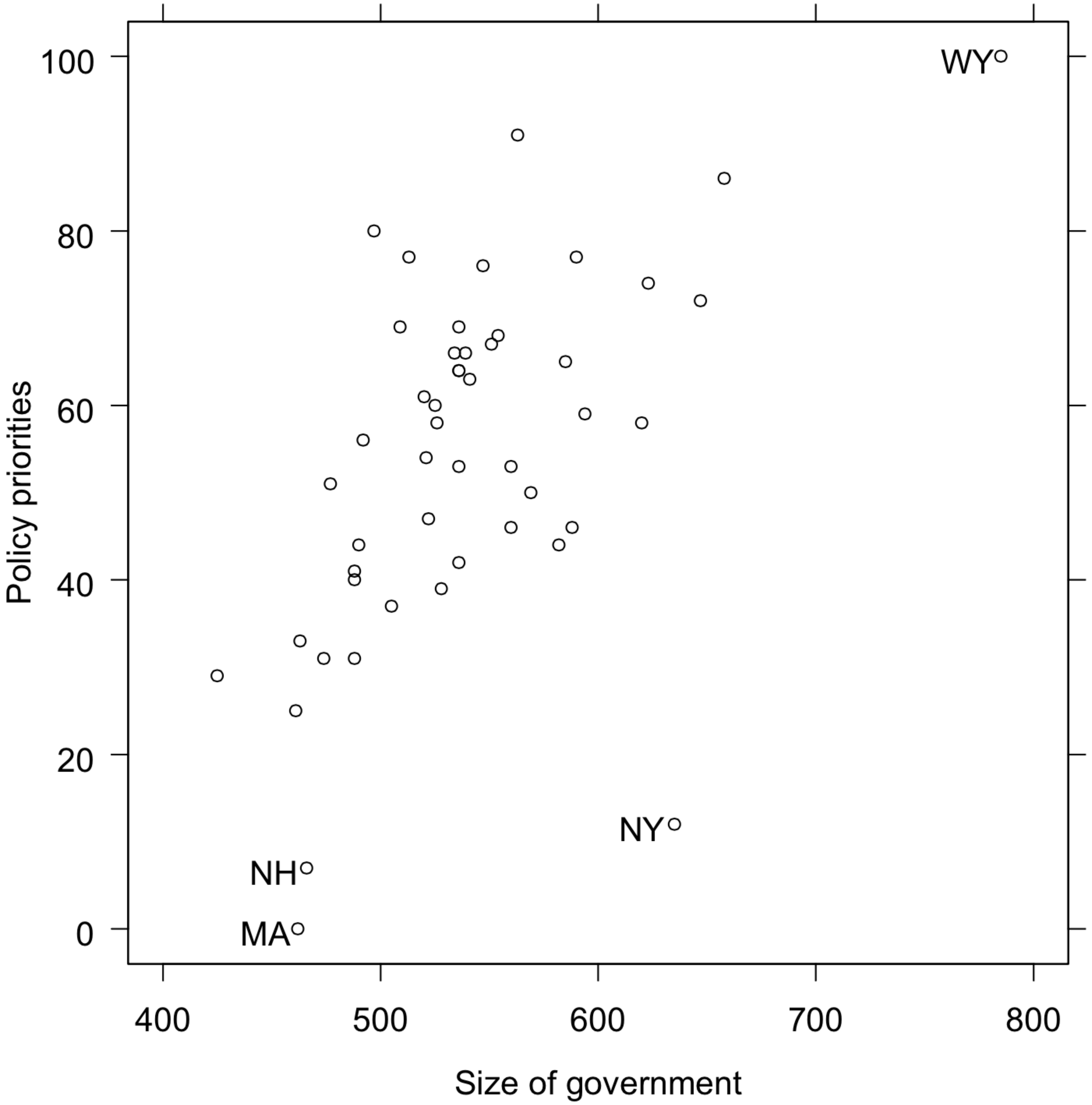
Univariate axes for clarity



A not super great use of point labels



A better use of point labels



What else can we do?

- Primary point to visual displays of bivariate data is to better assess *functional dependence* - a fancy phrase for the kind of relationship existing between variables

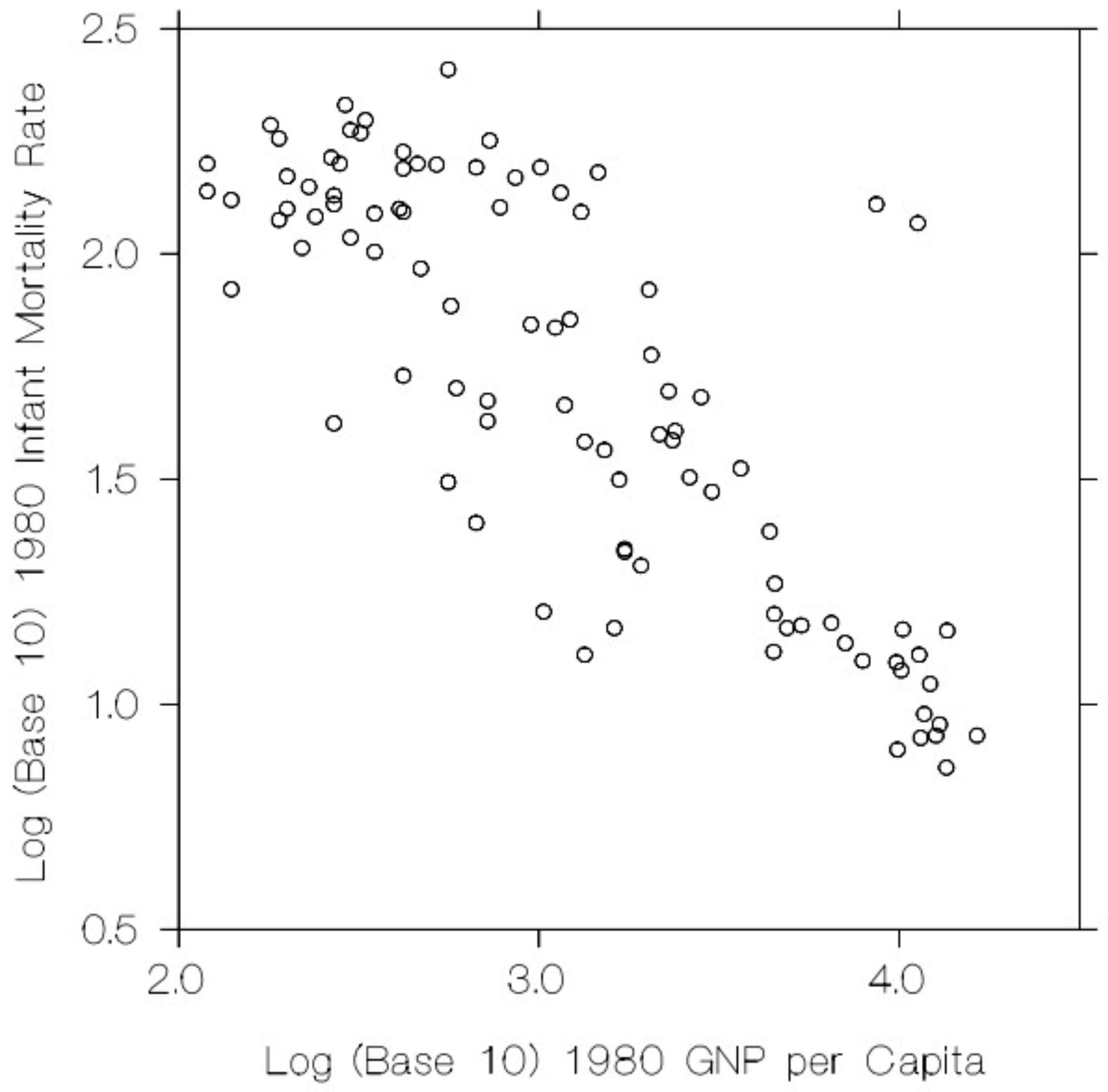
- *Slice* a scatterplot with box plots (or something else) to give a better visual flow and abstract from individual data points

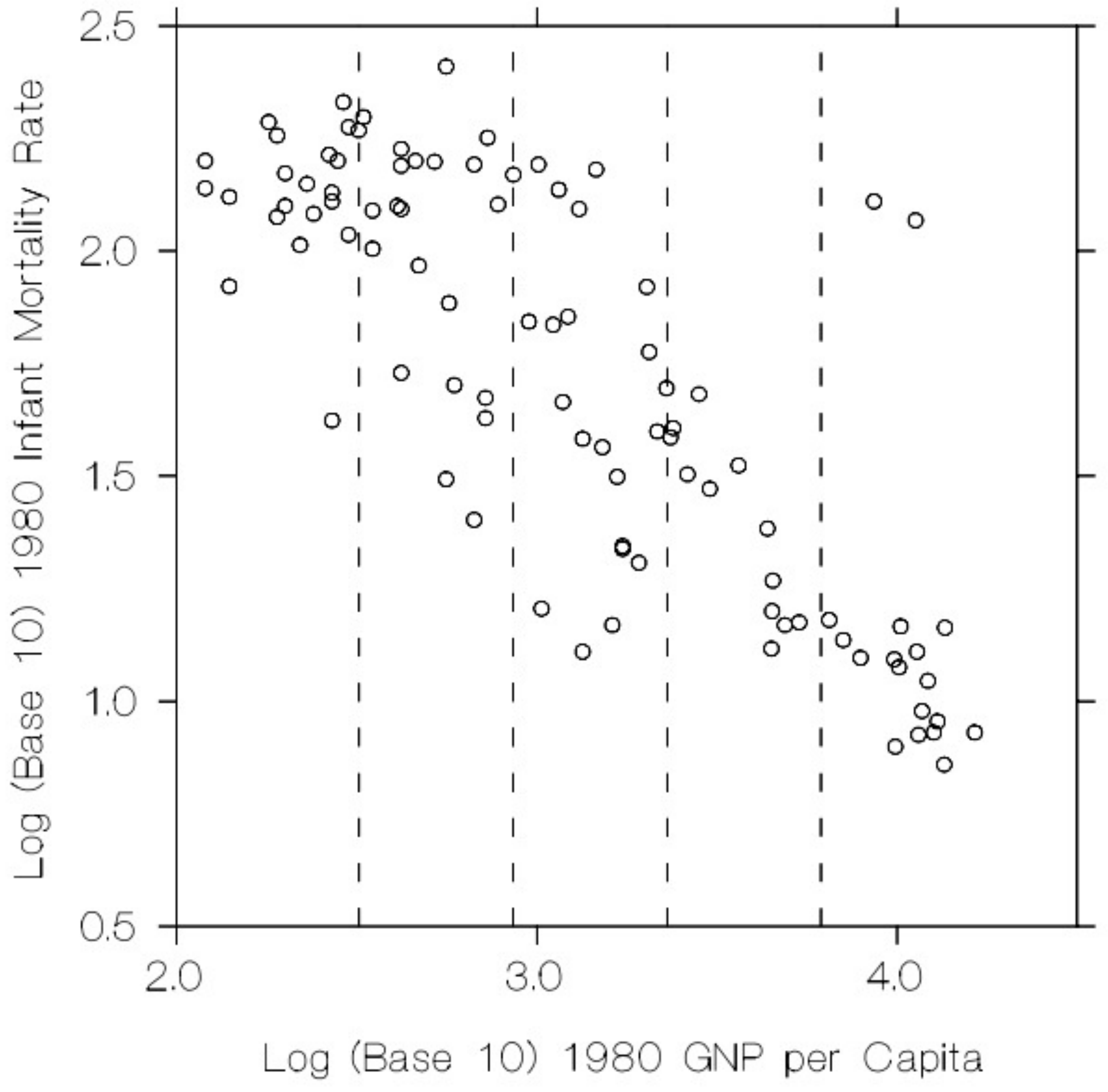
    - where you divide a plot region on the scatterplot into vertical “slices”, enabling nice visual displays

    - Can use univariate graphical displays in each slice

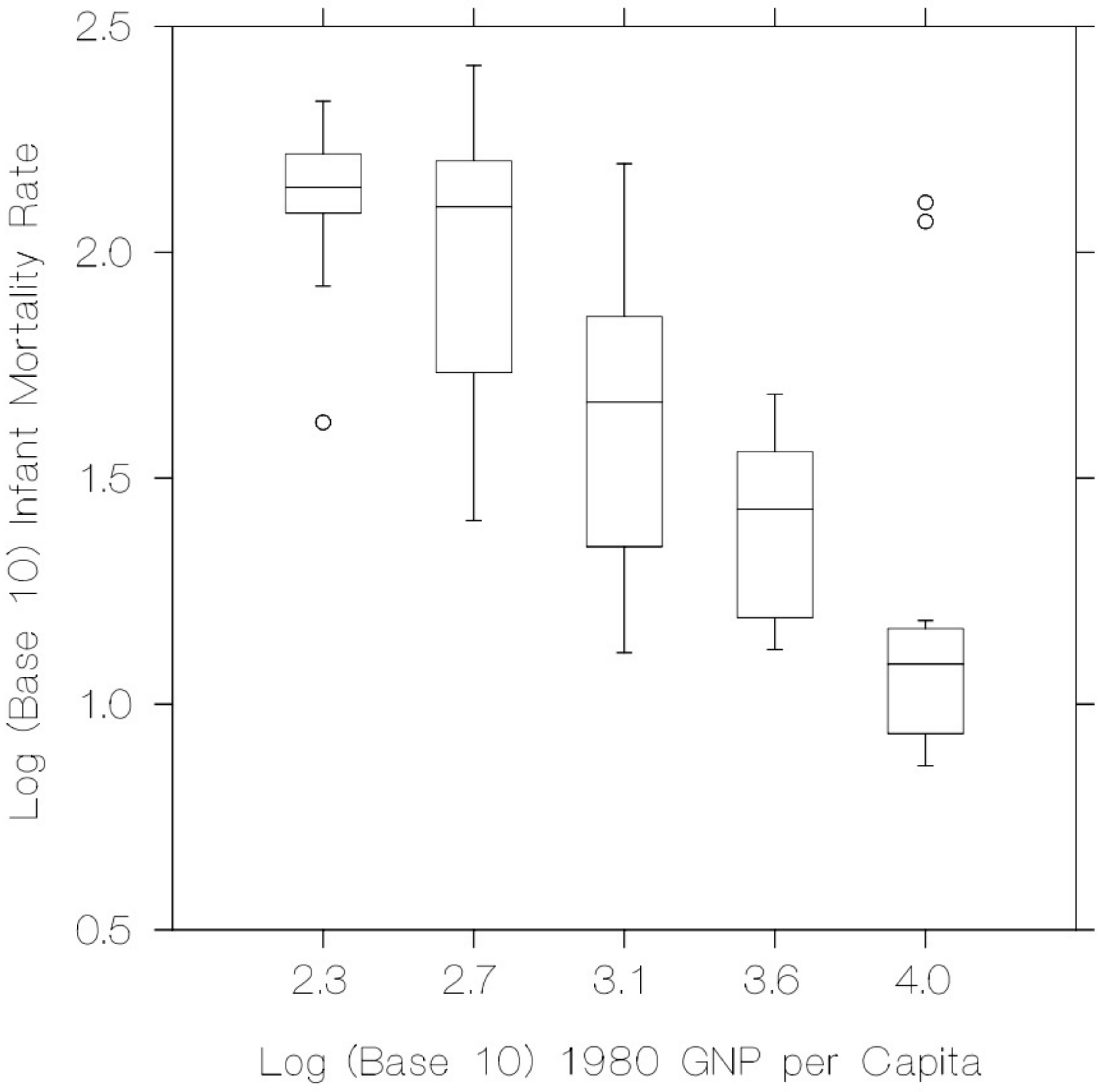
- Can allow for easier examination of relationships between variables

Slicing a scatterplot . . .

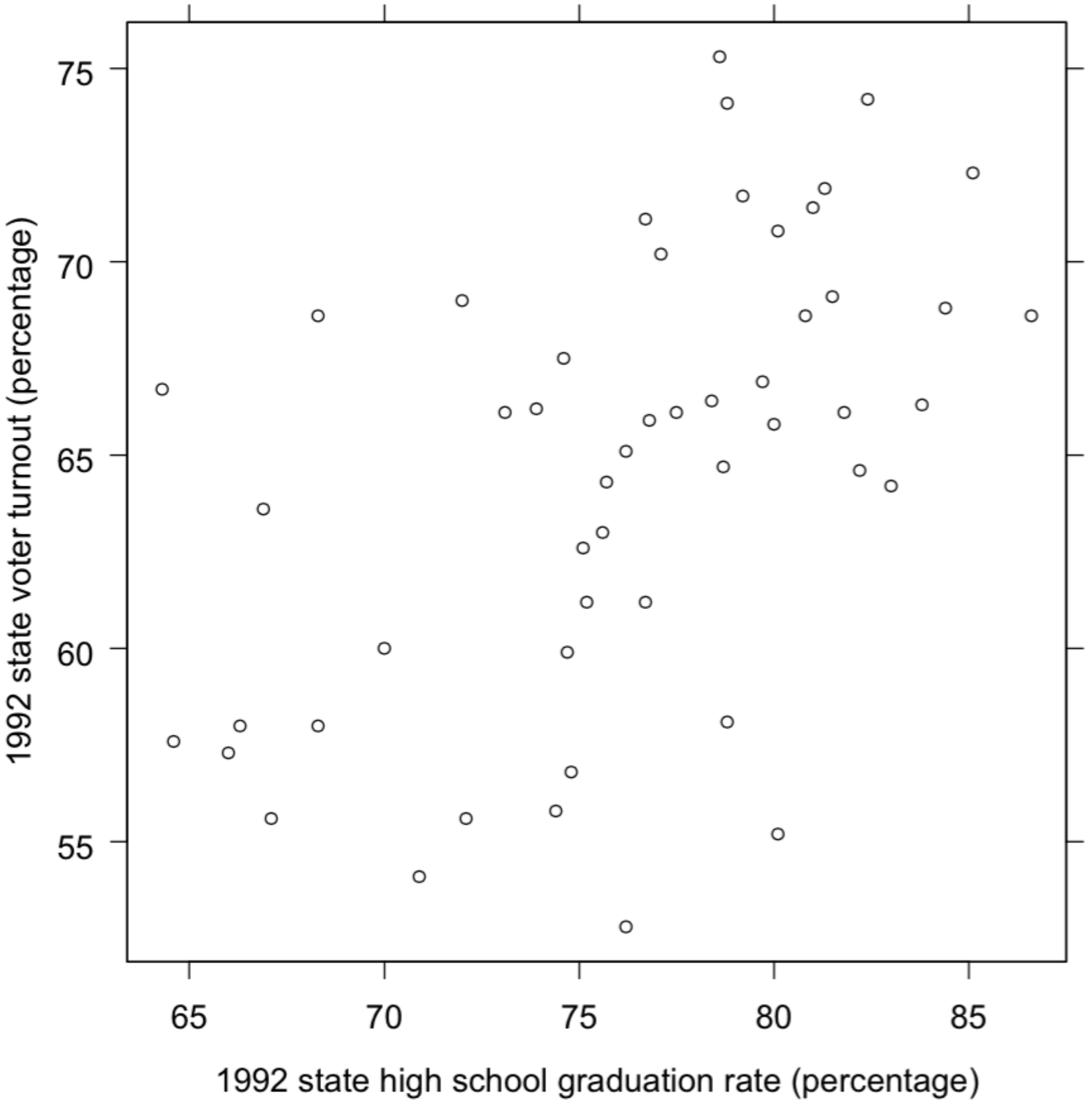




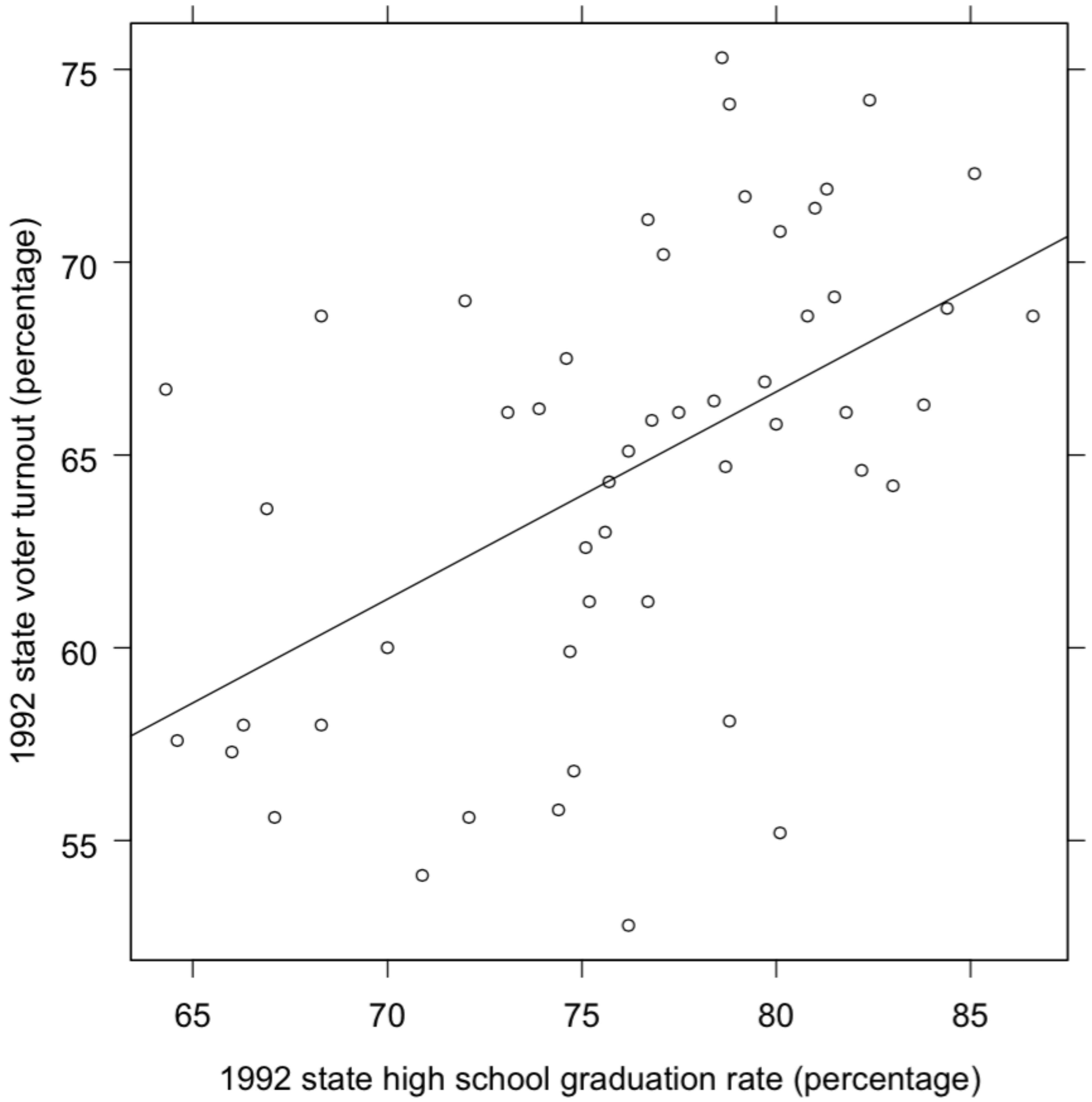
. . . and making box plots!



How can we clarify scatterplots with vague relationships?



OLS regression lines?



Doesn't, perhaps, capture the data well

OK, back to . . . (non)linearity

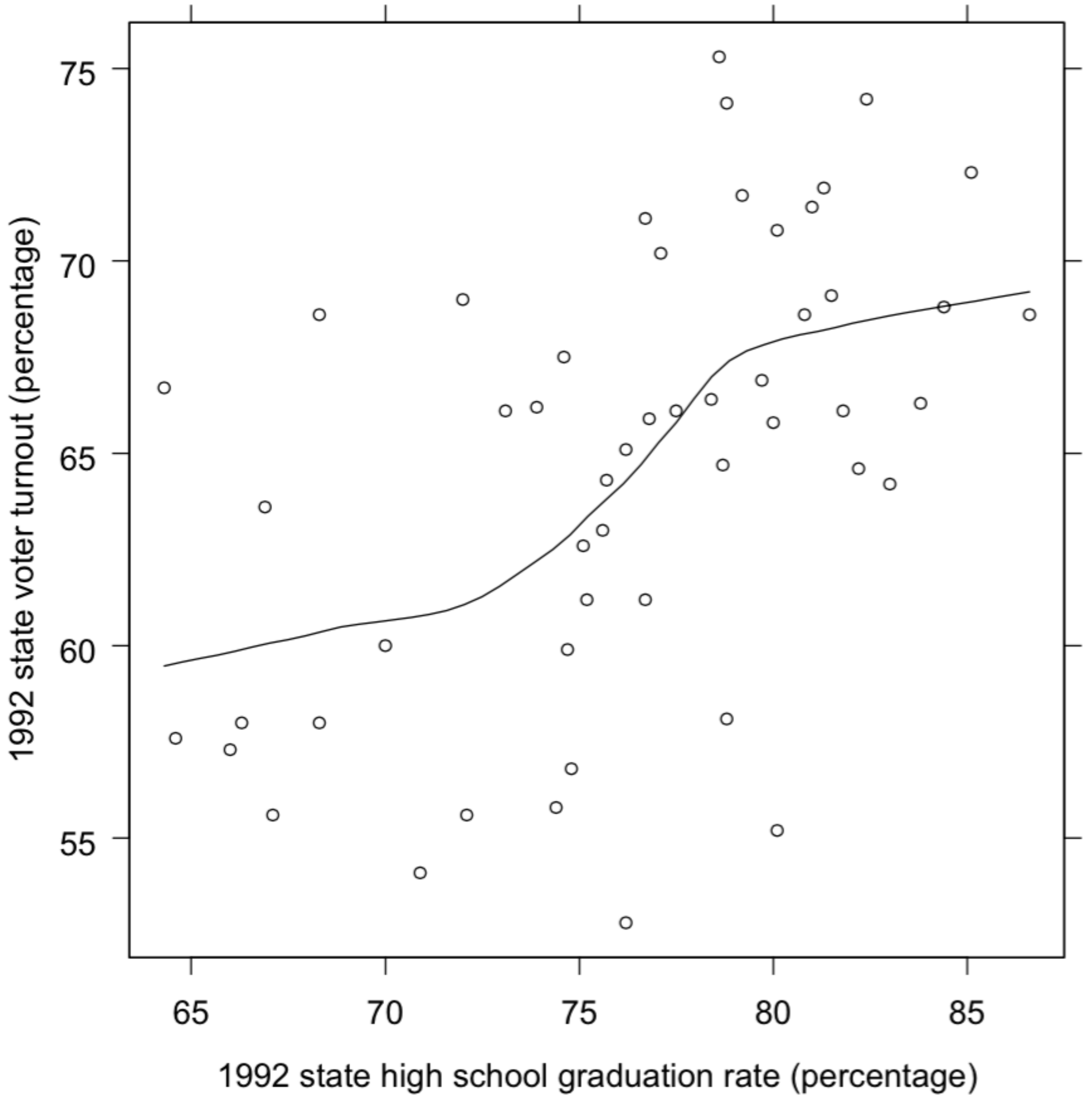
- To truly get at nonlinearity, it'd be nice to find simple ways to represent variable relationships besides lines
- Enter *scatterplot smoothing* - a very simple type of "nonparametric" regression
  - Specify no functional form, just average values across the axis
  - Generates a smooth curve that follows center of Y distribution across a range of X values

Loess smoothing

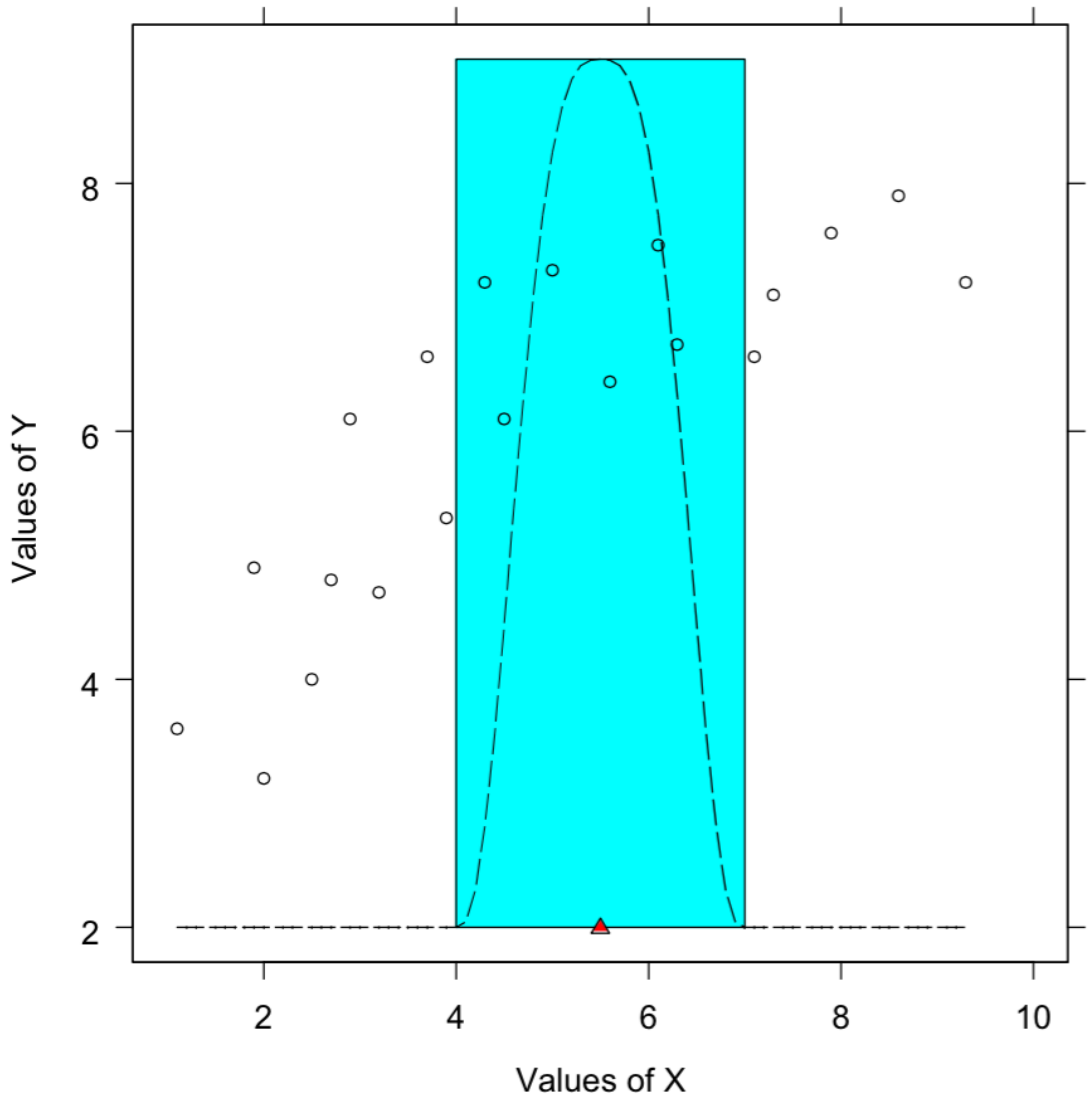
Loess = *Lo*cal Regr*ess*ion

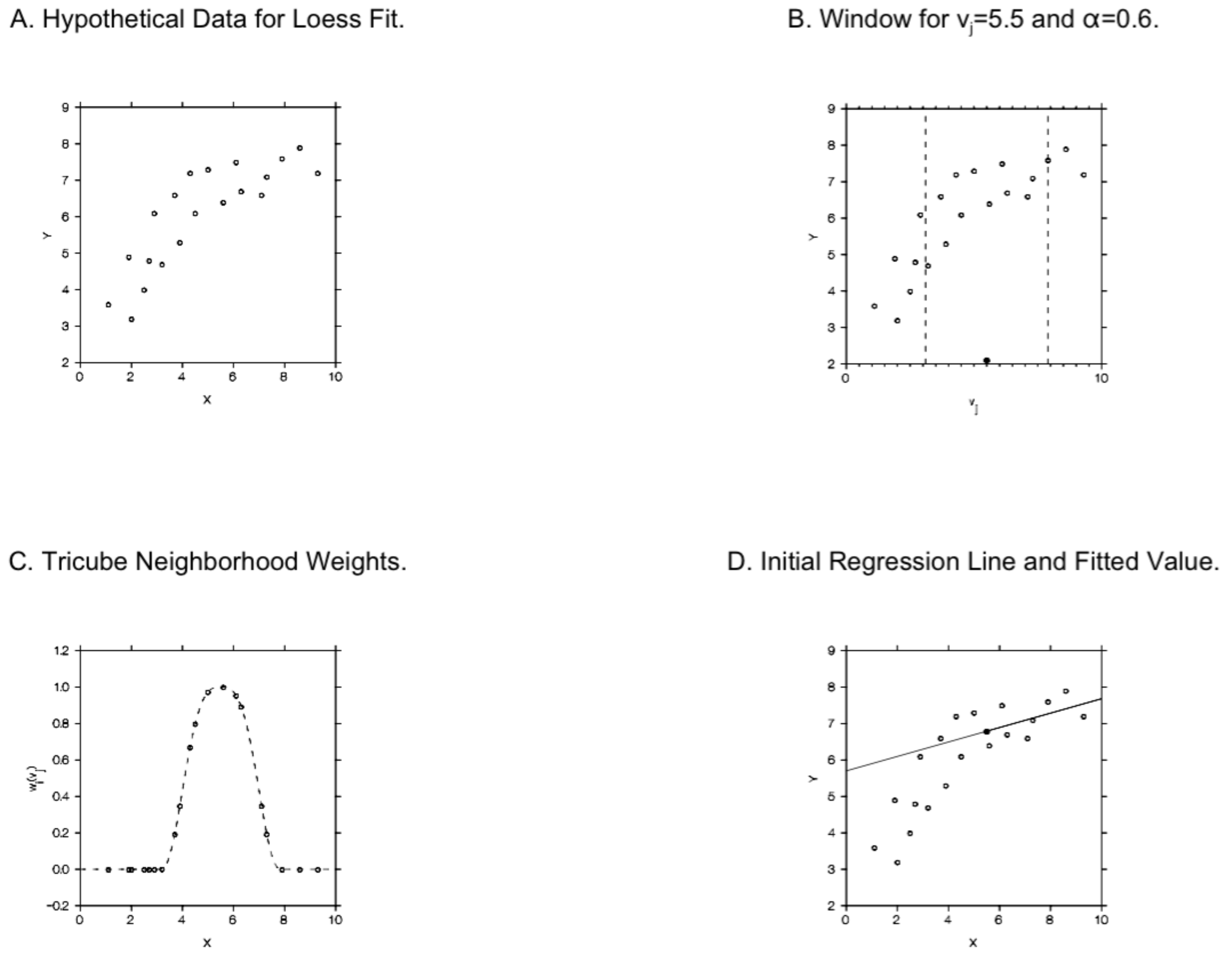
- Loess carries out a series of regressions within a “window” that “moves” across the range of X values
- Define equally space locations across X, perform a regression at each space, *weighting* according to proximity to the central point of the space
- Use regression estimates to predict value for current central Y point at X, and plot that
- As window moves across, plot all predicted value points and connect them with lines

So what does this look like?



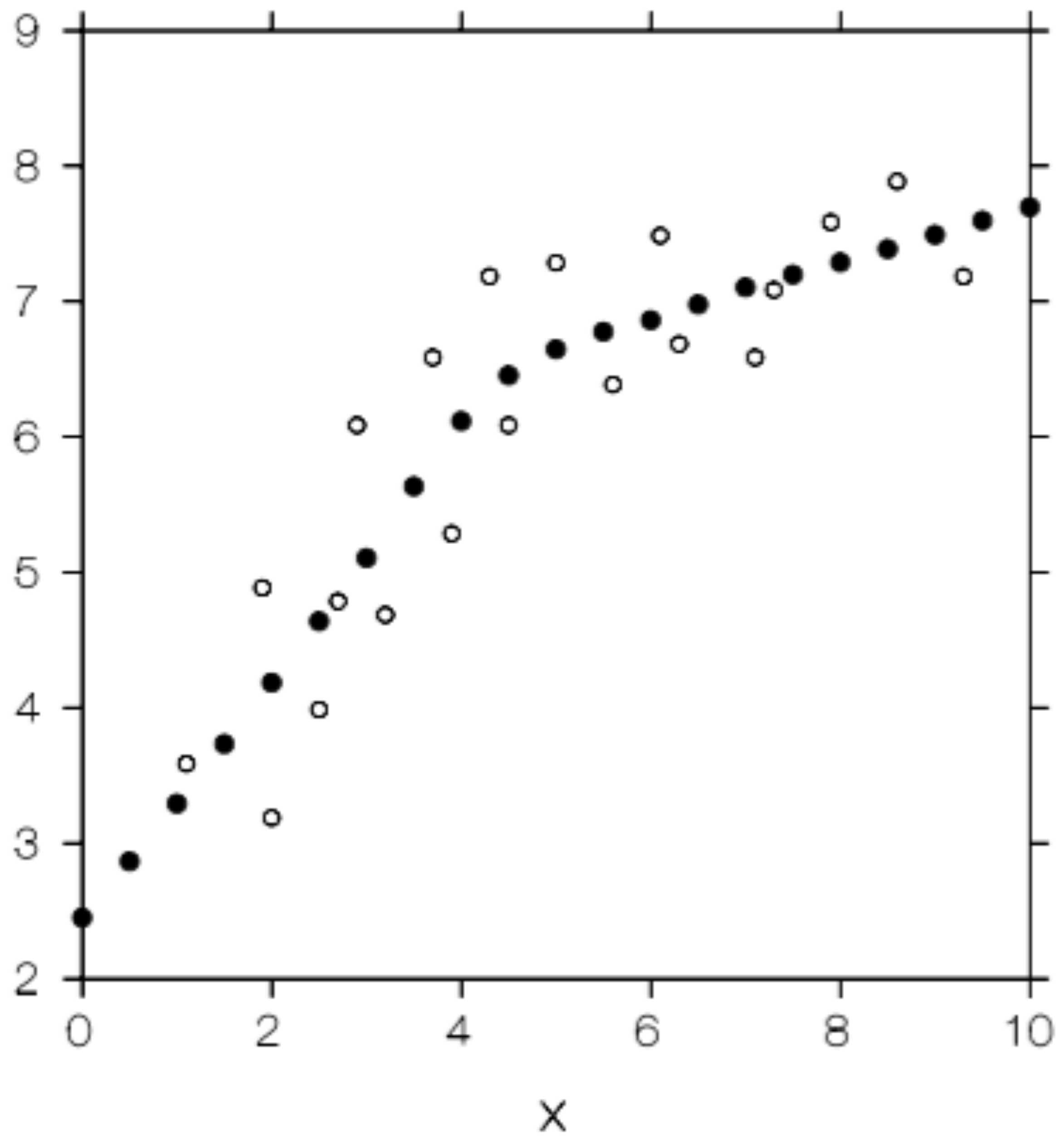
How does it work?
An illustration



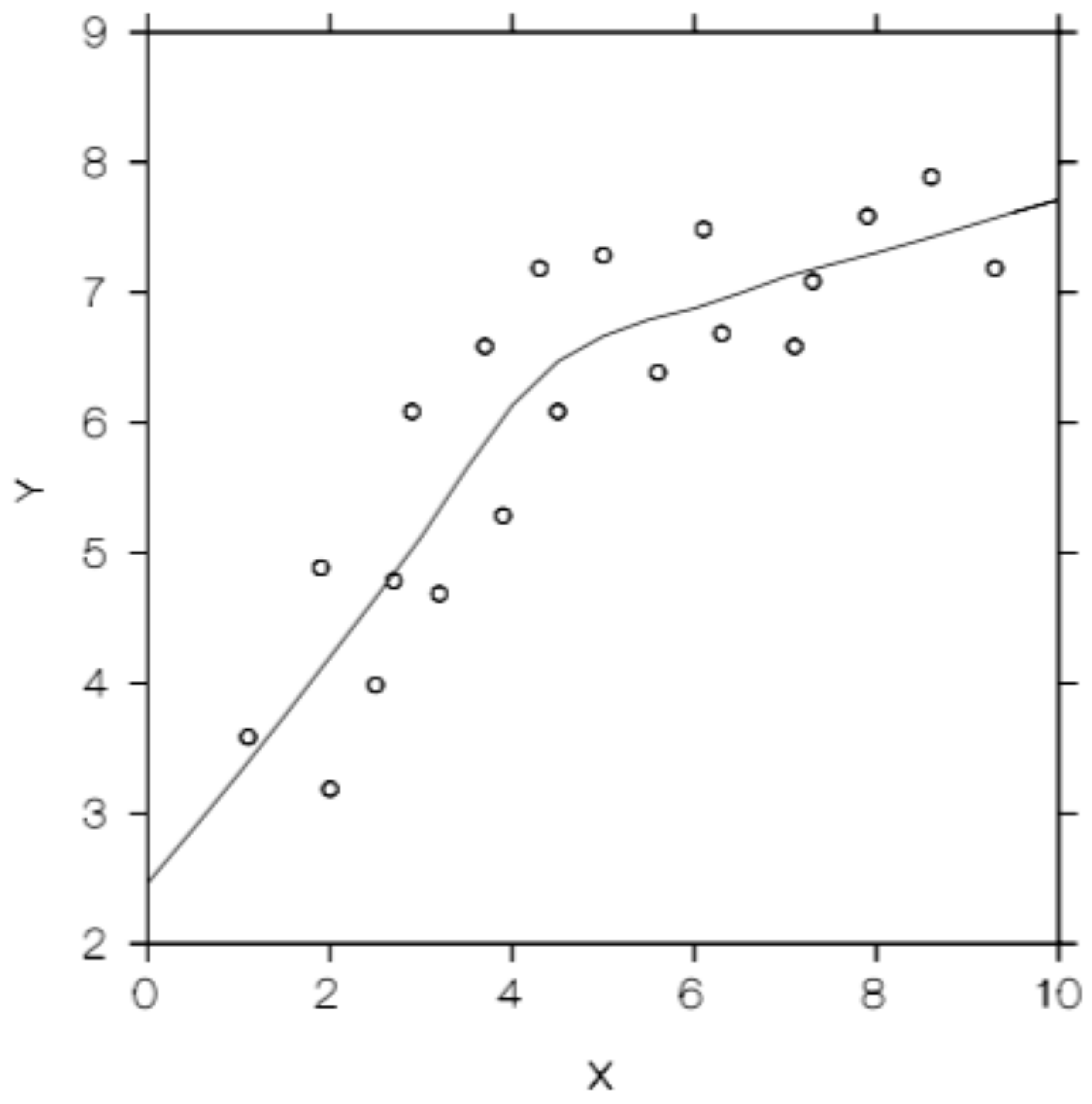

A. Hypothetical Data for Loess Fit.
B. Window for v_j=5.5 and α=0.6.
C. Tricube Neighborhood Weights.
D. Initial Regression Line and Fitted Value.
Y
X
v_j
$w_i(v_j)$

The final line

E. Complete Set of Fitted Values.



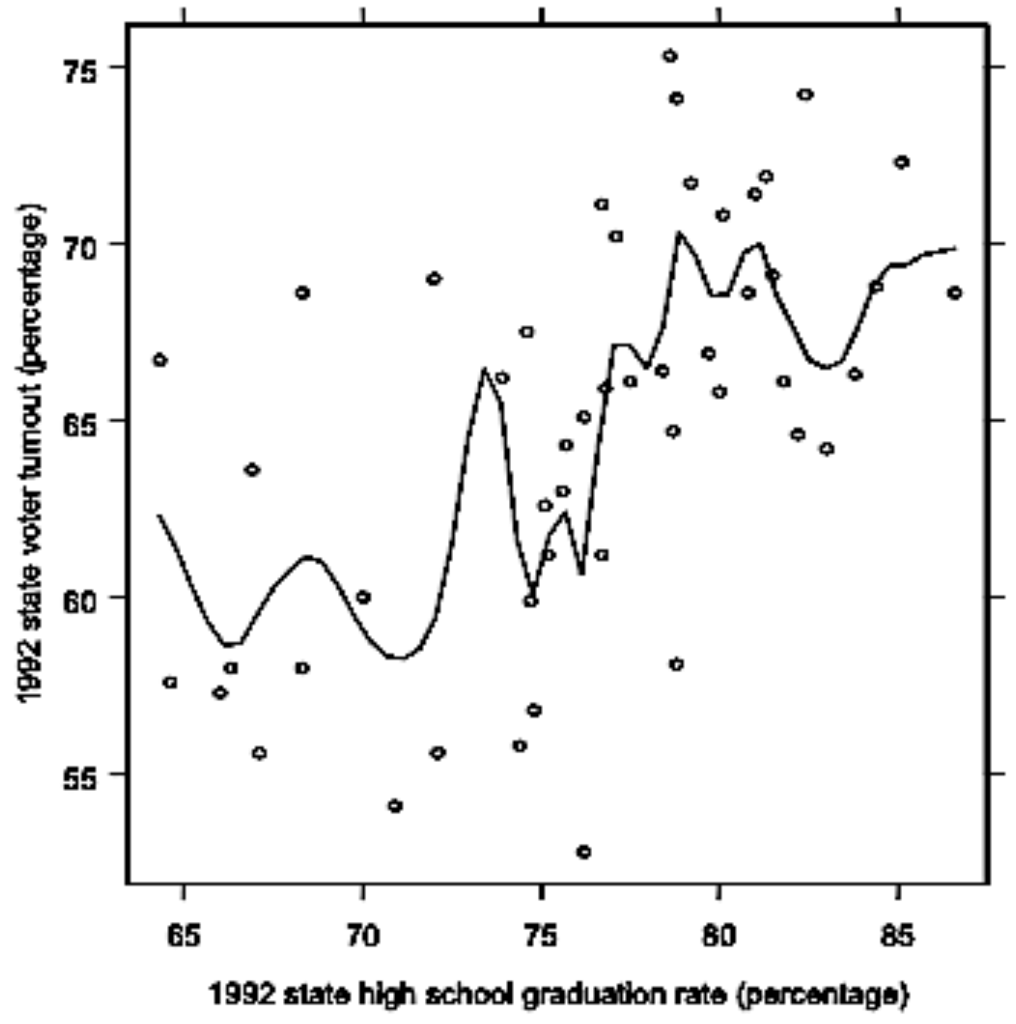
F. Loess Curve



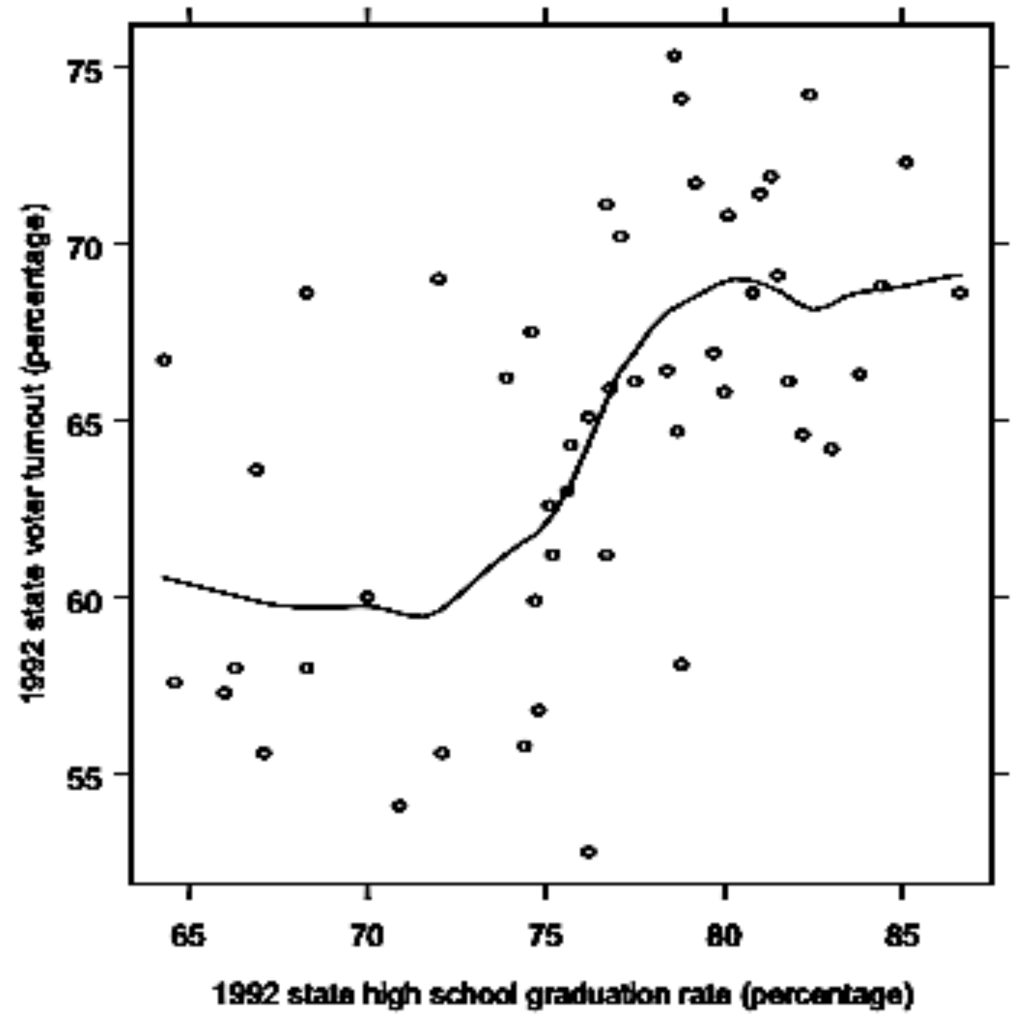
What are the limitations of loess?

- This is kind of similar to kernel density plots
- There's a similar parameter (a) identifying the "space" considered in each moving window
- And when that window changes . . .

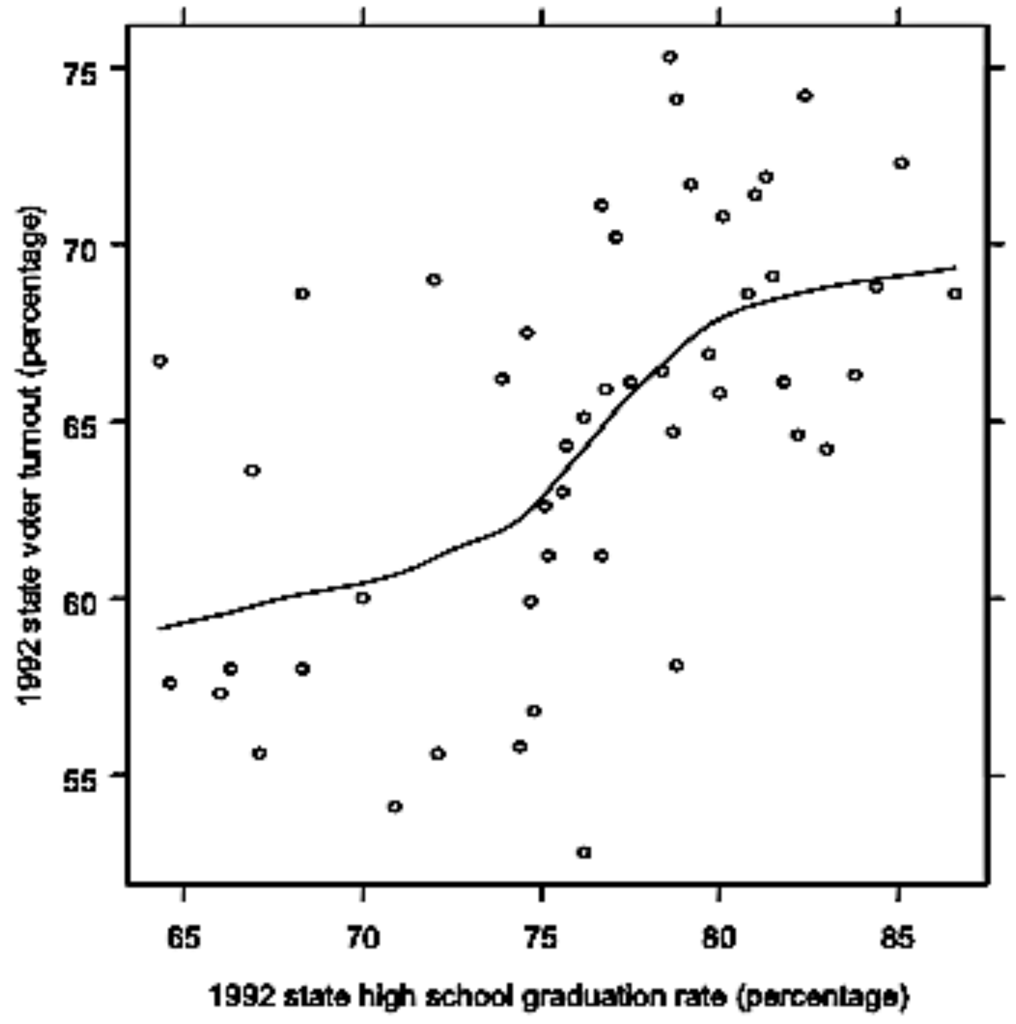
A. Loess Curve with α = 0.15



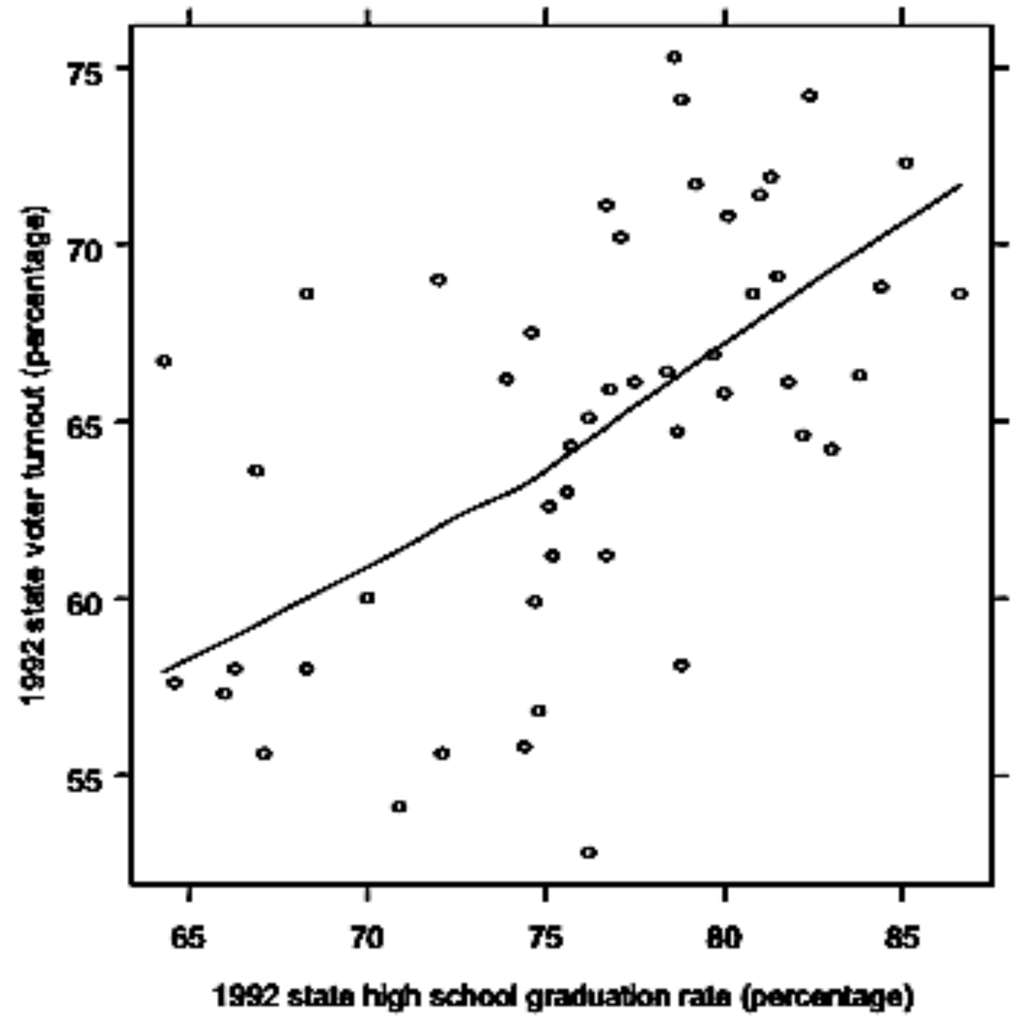
B. Loess Curve with α = 0.35



C. Loess Curve with α = 0.75

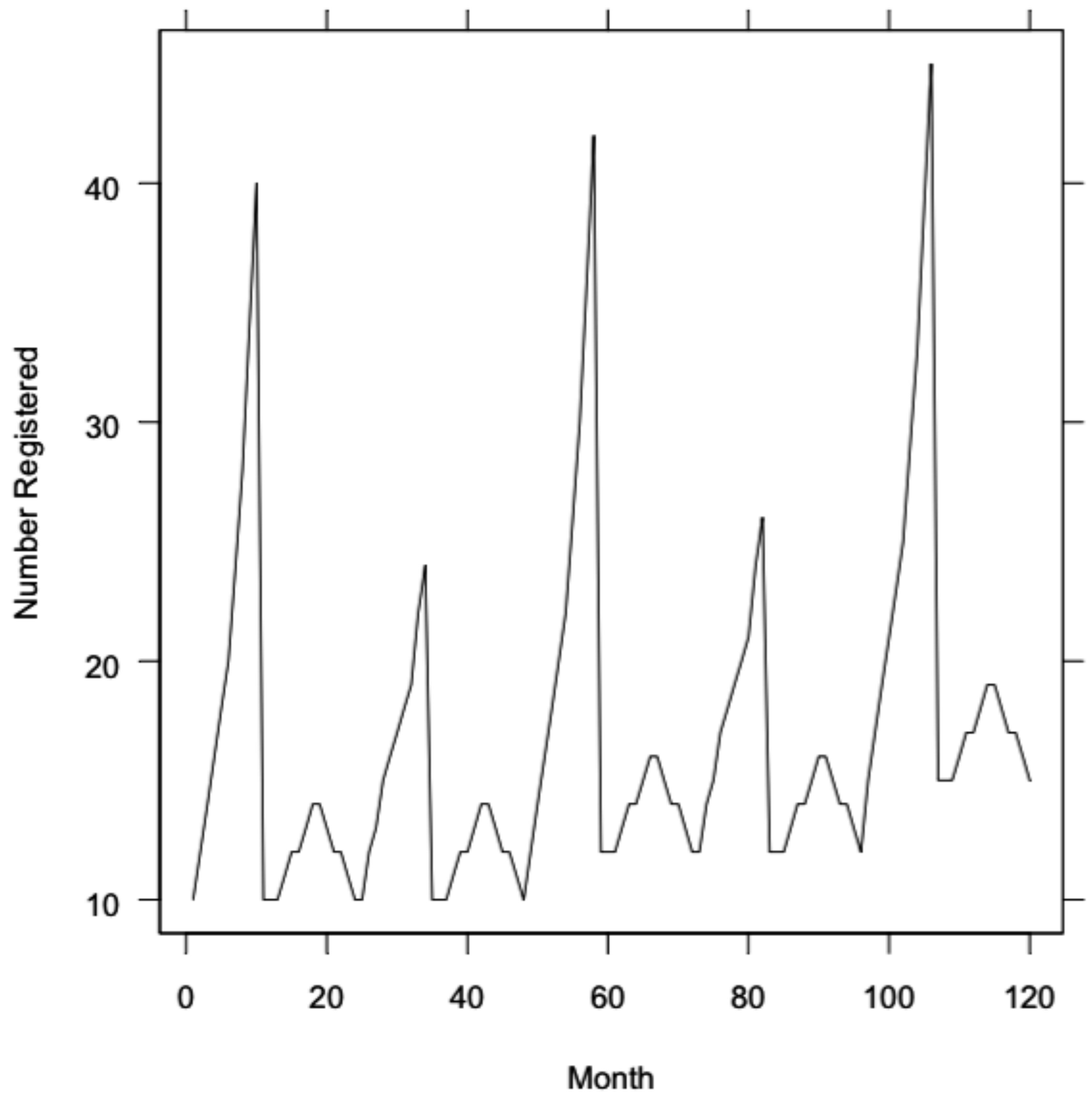


D. Loess Curve with α = 1.00

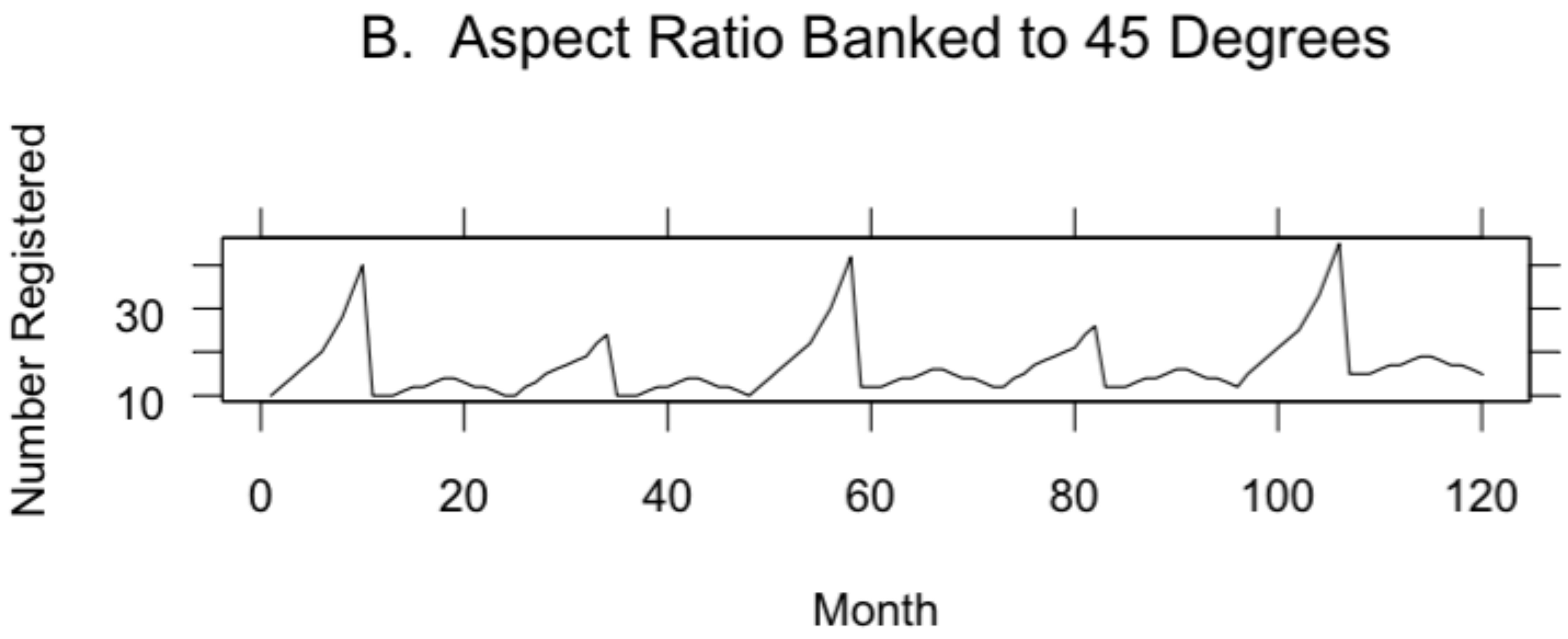


Finally, consider your scales

A. Aspect Ratio Set to 1.0



Tells a different story here, eh?



Same data . . . allows for better interpretation of “functional dependence”

Summing up

- There are not as many bivariate displays as univariate displays - the scatterplot covers a lot
- Same principles apply
- Consider using a loess smoothed line to examine linearity

Where else can I find about graphics?

- *The source of much of these lectures:*
  - “Statistical Graphics for Univariate and Bivariate Data”, Sage #117.
  - “Statistical Graphics for Visualizing Multivariate Data”, Sage #120.
- *Class text:* Kieran Healy’s book, “Data Visualization: A Practical Introduction” http://socviz.co and print
  - Nice overview of all three components (perception, principle, and practice) as well as developing graphical “taste”
  - Lots of kinds of graphics: univariate, bivariate, GIS, time, etc
- *The Godfather:* Tufte, assorted books, but especially “The Visual Display of Quantitative Information”
- Contemporary and technical: Yau, *Visualize This: The Flowing Data Guide*