Choices in statistical graphics: My stories

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

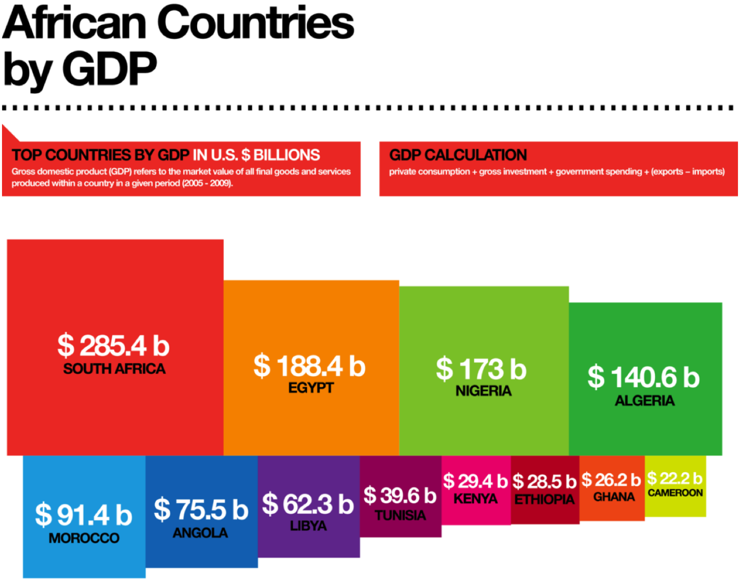
New York Data Visualization Meetup
14 Jan 2013

My earlier talk on tradeoffs in statistical graphics

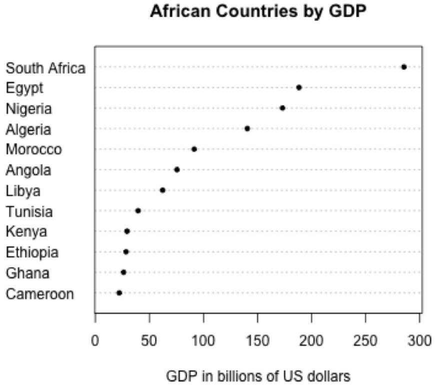
- Originally: Infoviz vs. stat graphics
  - The best information visualizations are grabby, visually striking
  - The best statistical graphics reveal patterns and discrepancies
  - Different goals, different looks
- Lots of negative reactions
  - (Some) infofiz people felt we were trivializing their work
  - (Some) statisticians felt we gave infofiz too much respect
- Our new theme: tradeoffs in statistical graphics

We did not come to mock . . .

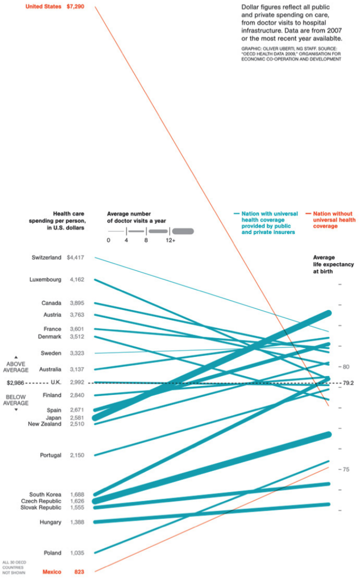
Instead, compare a bare-bones infographic . . .



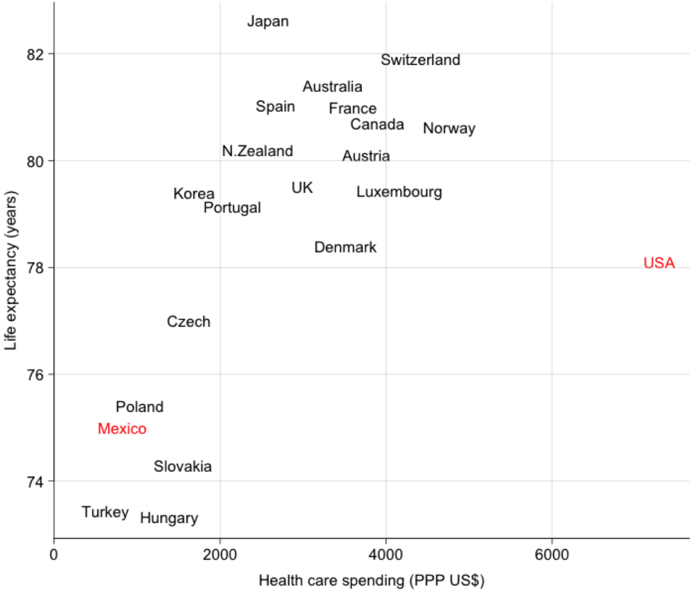
To a corresponding statistical graphic . . .



Another example . . .

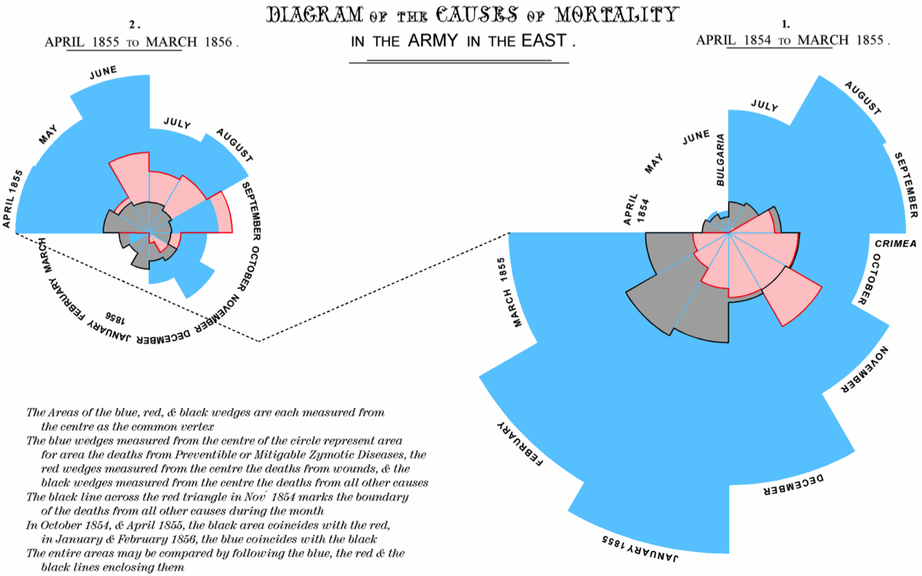


The statistician's version . . .

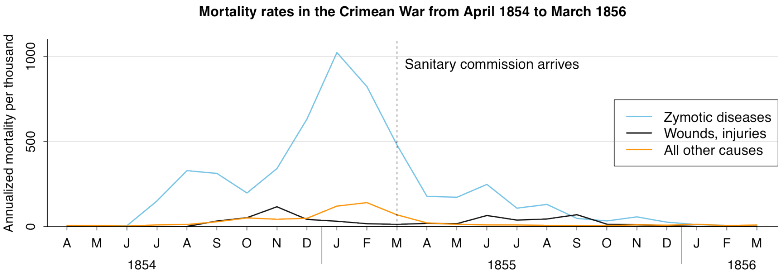


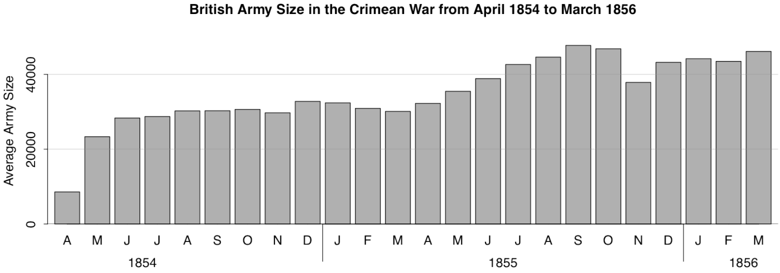
A legendary early infographic . . .

http://www.Florence-Nightingale-Avenging-Angel.co.uk/Coxcomb.htm



How we would display it . . .





For those of you reading this talk off the web

- I'm not saying that the boring plots (constructed by Antony Unwin and myself using R) are better than Florence Nightingale's beautiful images!
- Rather, I'm saying that Nightingale's graphic and ours serve different purposes:
  - She dramatizes the problem with a unique and visually-appealing image that draws the casual viewer in deeper
  - We display the data to reveal patterns, for viewers who are already interested in the problem
- In any case, this is not my main point today. We'll spend most of our time discussing the choices involved in graphs that I've made over the years.
- Now, back to our regularly scheduled presentation ...

General theme

- All graphs are comparisons
- All of statistics are comparisons

Specific recommendations

- Multiple plots per page (small multiples)
- Don't clutter each plot
- Line plots are great—they facilitate more comparisons

Don't clutter each plot: example

From *Graph Design for the Eye and Mind* by Stephen Kosslyn:

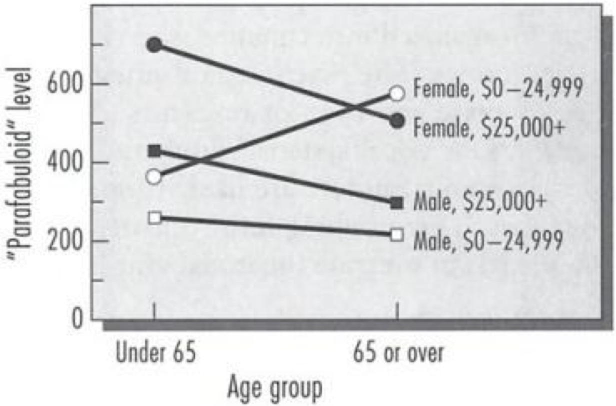
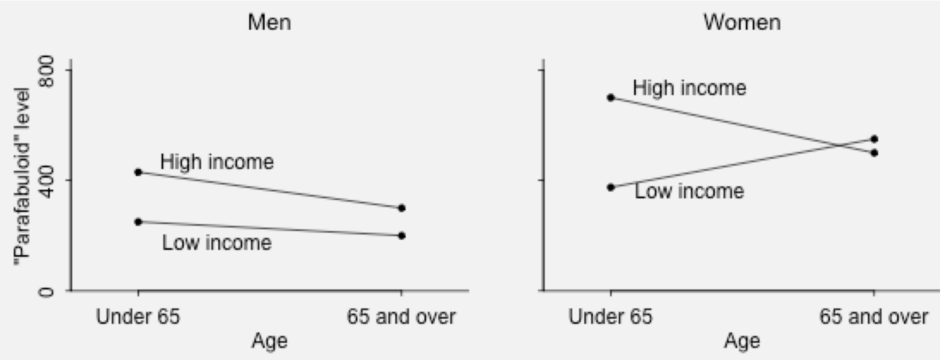
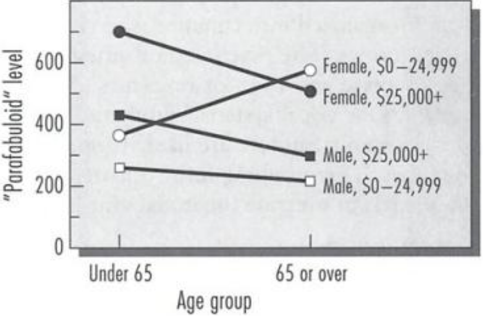
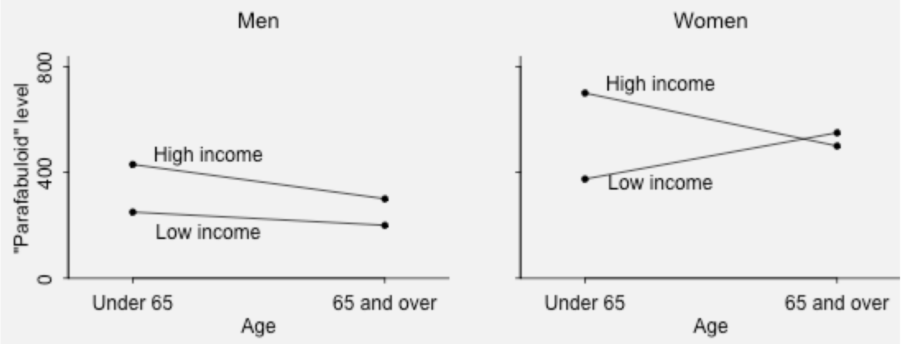


Figure 1.6. The contrasting slope of one line makes the odd group easy to spot; no such visual cue can be given in a table.

Redo using small multiples!

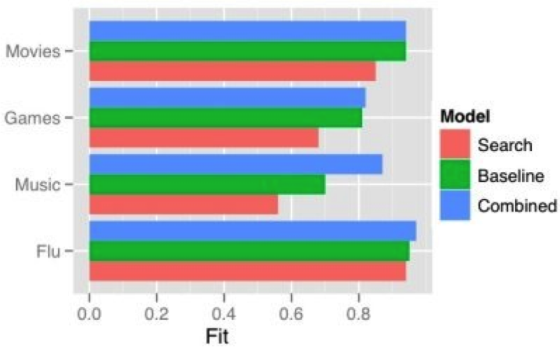



"Parafabuloid" level
600
400
200
0
Female, $0–24,999
Female, $25,000+
Male, $25,000+
Male, $0–24,999
Under 65
65 or over
Age group

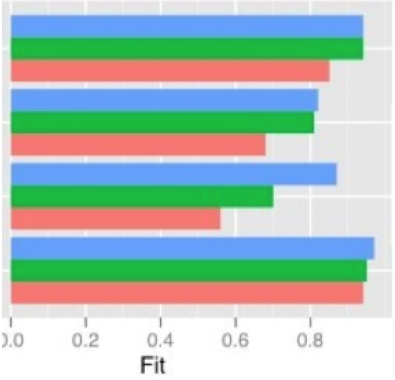

Men
Women
"Parafabuloid" level
800
400
0
High income
Low income
High income
Low income
Under 65
65 and over
Age
Under 65
65 and over
Age

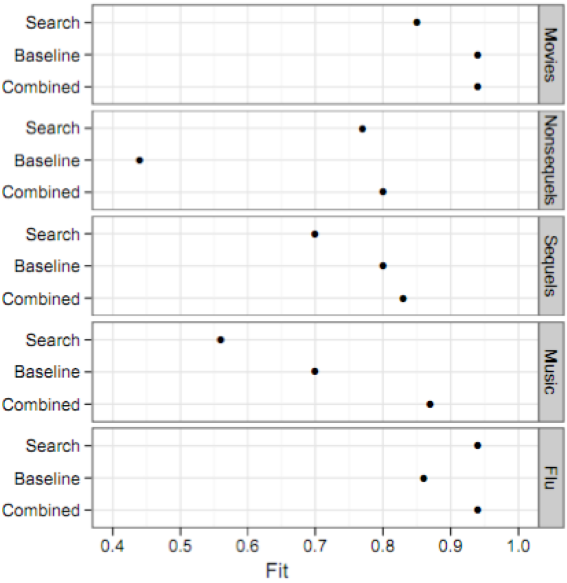
Line plots: Cleveland's principle

- Always ask: What is the comparison?
- Example: an analysis from market research

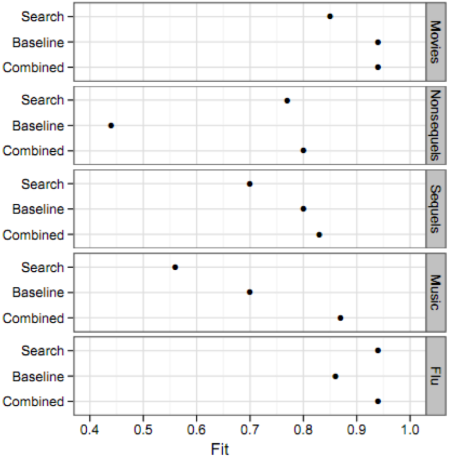


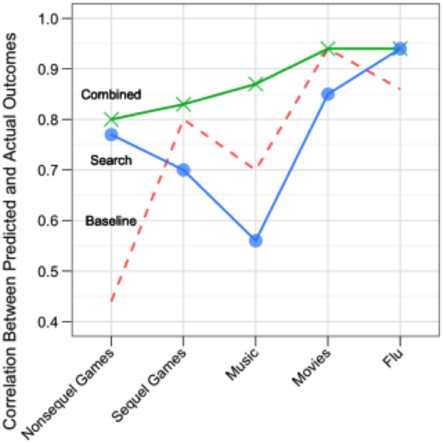
Improvement?





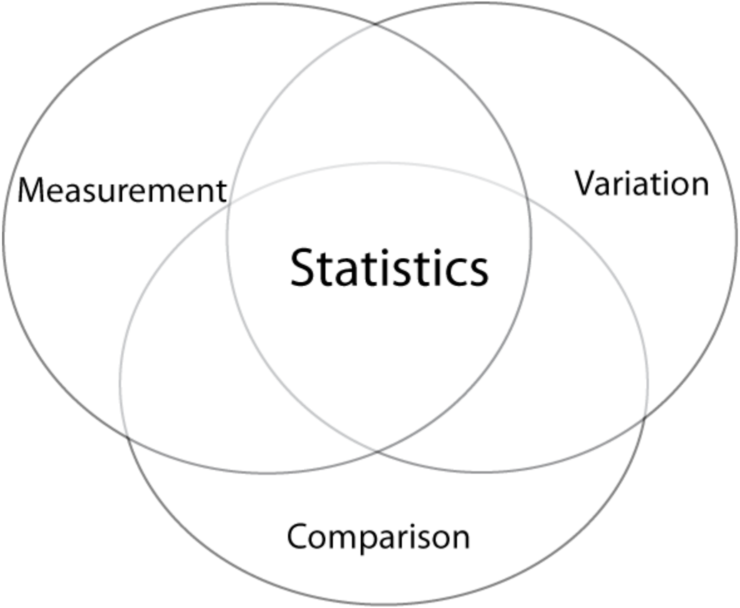
Line plot is better





Consider the comparisons you can make!

Statistics is . . .



Today's talk

- (Some of) my examples from (nearly) 30 years of applied resarch
- Choices involved in making the graphs
- What works, what doesn't, and why
- You must participate!

1984: "The effects of solar flares on single event upset rates"

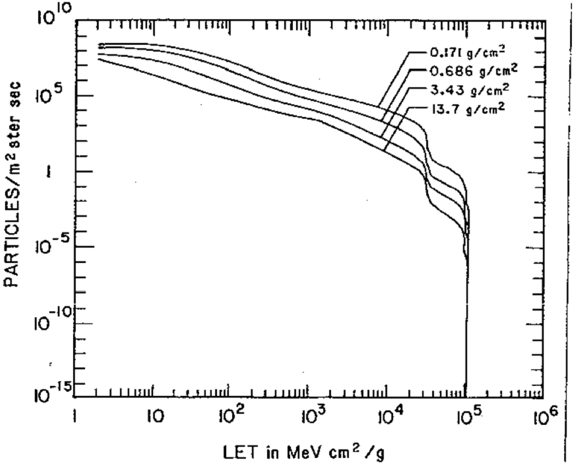
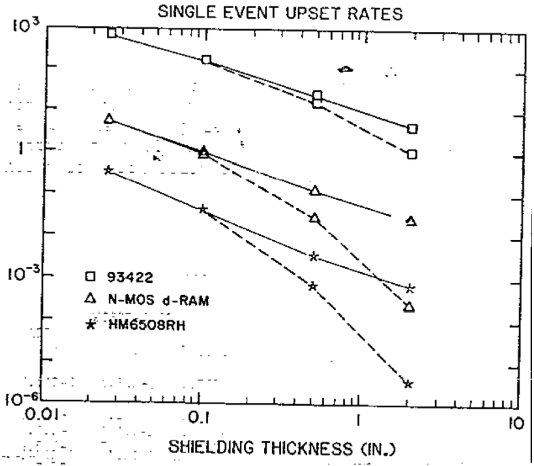


Figure 4: The integral LET spectra for the composite worst-case solar flare particle event outside the magnetosphere. These spectra are behind aluminum shielding of the indicated thicknesses. These thicknesses correspond to 0.025, 0.1, 0.5 and 2.0

1984: "The effects of solar flares on single event upset rates"



1986: "Reduced subboundary misalignment in SOI films scanned at low velocities"

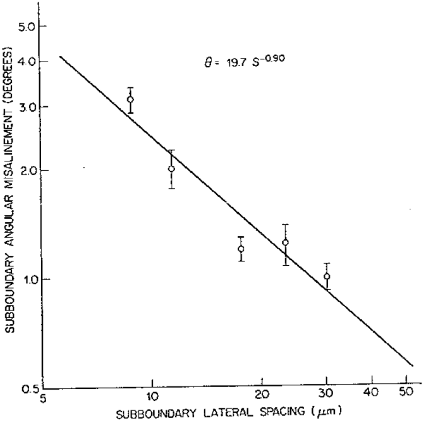
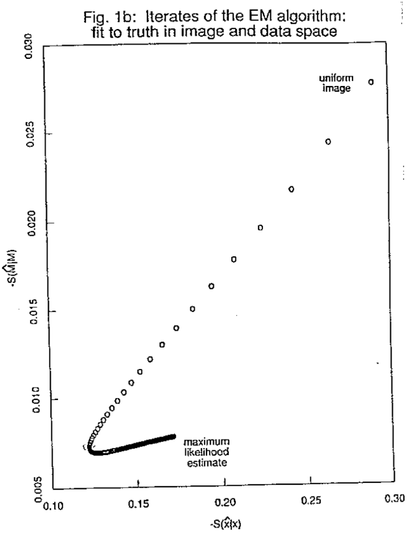


Fig. 9 Measured average crystallographic angular misalignment $\bar{\theta}$ for a number of subboundaries as a function of the average lateral spacing $\bar{s}$ of those subboundaries as obtained from the experiment of Fig. 8.

1989: "Constrained maximum entropy methods in an image reconstruction problem"



1990: "Estimating the electoral consequences of legislative redistricting"

Table 1. Votes Received by Democrats and Republicans in Ohio Legislative House Districts,

1972						1974			
District	Democrat	Republican	District	Democrat	Republican	District	Democrat	Republican	D
1	18,250	22,798	51	22,488	16,951	1	20,490	15,107	
2	25,679	17,130	52	24,336	14,083	2	18,669	11,969	
3	0	33,954	53	25,932	8,997	3	12,778	20,272	
4	23,684	10,212	54	22,780	15,229	4	15,765	9,813	
5	21,723	16,130	55	20,198	9,583	5	11,711	9,708	
6	28,309	0	56	21,603	10,678	6	20,584	5,763	
7	20,334	12,675	57	16,533	17,114	7	20,193	9,778	
8	16,622	3,656	58	13,587	22,105	8	11,153	2,261	
9	11,946	10,396	59	14,877	20,234	9	9,566	0	
10	12,383	5,316	60	14,556	13,940	10	8,277	1,890	
11	20,091	18,539	61	16,507	17,825	11	22,398	5,221	
12	18,337	20,561	62	23,668	13,428	12	9,865	19,599	
13	16,688	1,970	63	13,868	18,402	13	10,687	966	
14	22,865	11,218	64	13,984	22,593	14	11,478	8,087	
15	21,401	0	65	11,710	29,134	15	15,905	1,936	
16	27,783	12,701	66	15,500	30,156	16	21,909	10,403	
17	24,511	15,716	67	20,409	17,931	17	22,327	11,274	
18	28,805	14,454	68	21,489	15,574	18	22,416	8,138	
19	17,687	23,463	69	16,592	21,816	19	12,431	10,832	

1990: "Estimating the electoral consequences of legislative redistricting"

0.5	11234
	5667789999
0.6	0000111122223333344444
	55667778888888899999999
0.7	0000001111122222233333333444444
	55555556667778889
0.8	00011111122333
	66799
0.9	11344

Figure 1. Stem-and-Leaf Plot of the Proportion of the Vote Received by a Party in a Contested District Election, Immediately Preceding an Election in Which That Party Was Unopposed in That District.

1990: "Estimating the electoral consequences of legislative redistricting"

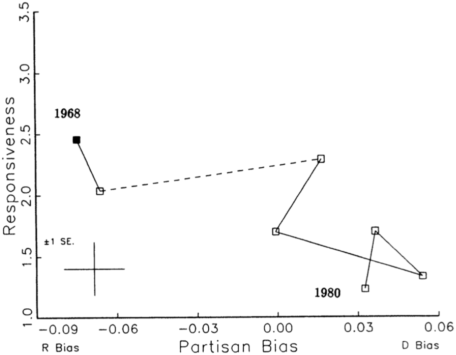


Figure 7. Ohio House, 1968–1980.

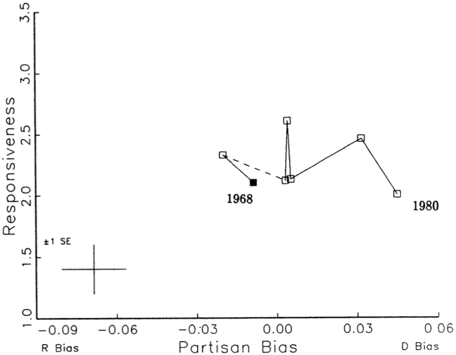
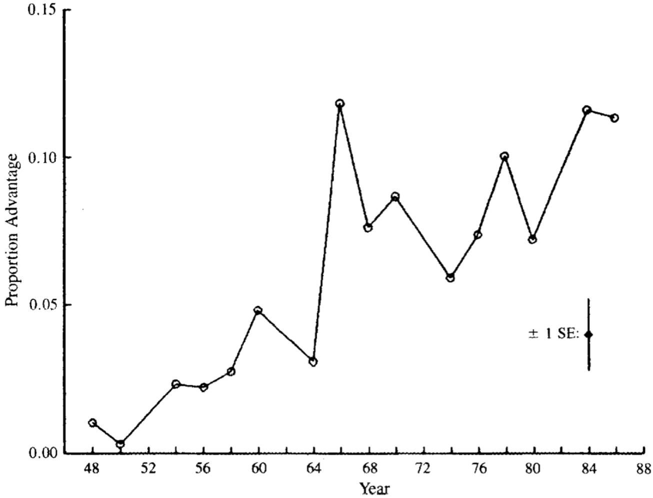


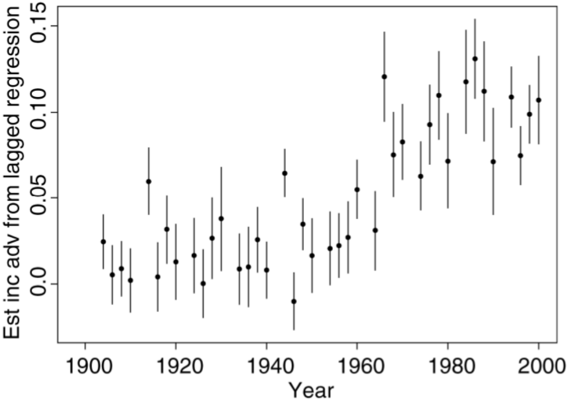
Figure 9. Wisconsin House, 1968–1980.

1991: "Systemic consequences of incumbency advantage in U.S. House elections"

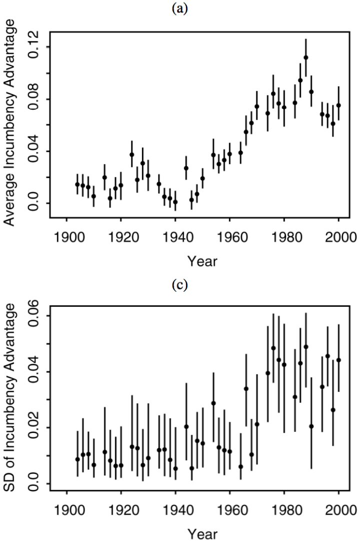
Figure 3. Estimates of Incumbency Advantage



2008: "Estimating incumbency advantage and its variation, as an example of a before/after study"



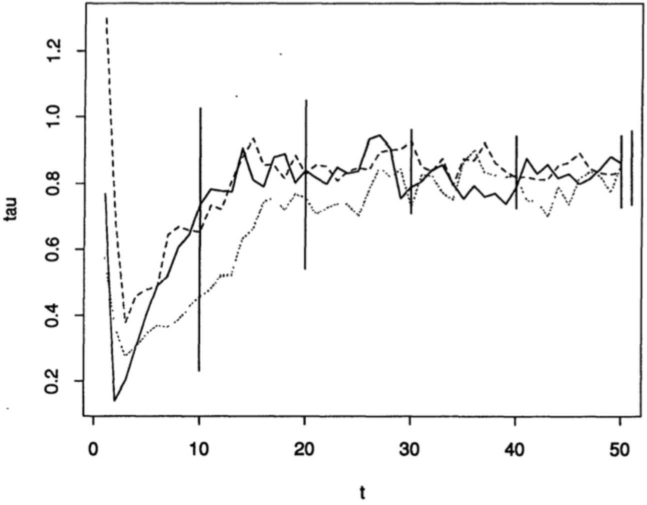
2008: "Estimating incumbency advantage and its variation, as an example of a before/after study"



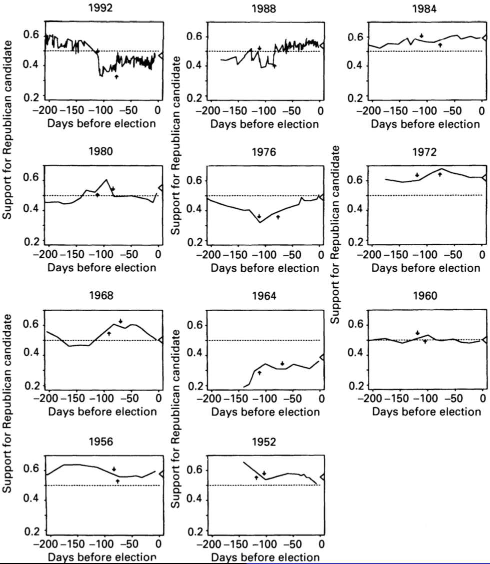
1992: "Inference from iterative simulation using multiple sequences"

	Normal-theory posterior interval			Potential scale reduction		Simulated quantiles				
	2.5%	$\hat{\mu}$	97.5%	Est.	97.5%	2.5%	25%	Median	75%	97.5%
α_1	5.66	5.73	5.80	1.00	1.00	5.66	5.71	5.73	5.76	5.80
α_2	5.82	5.89	5.95	1.00	1.00	5.82	5.86	5.89	5.91	5.95
α_3	5.64	5.71	5.78	1.00	1.01	5.65	5.69	5.71	5.73	5.78
α_4	5.64	5.71	5.77	1.00	1.02	5.64	5.68	5.71	5.73	5.77
α_5	5.51	5.58	5.65	1.00	1.01	5.51	5.56	5.58	5.60	5.65
α_6	5.73	5.80	5.86	1.00	1.00	5.73	5.77	5.80	5.82	5.86
α_7	5.79	5.86	5.92	1.00	1.00	5.79	5.83	5.86	5.88	5.92
α_8	5.52	5.59	5.66	1.00	1.00	5.52	5.56	5.59	5.61	5.65
α_9	5.48	5.55	5.62	1.00	1.00	5.49	5.53	5.55	5.57	5.62
α_{10}	5.71	5.77	5.84	1.00	1.01	5.71	5.75	5.77	5.80	5.84
α_{11}	5.65	5.72	5.78	1.00	1.01	5.65	5.69	5.72	5.74	5.78
α_{12}	5.66	5.73	5.80	1.00	1.00	5.66	5.71	5.73	5.75	5.80
α_{13}	5.97	6.03	6.10	1.00	1.00	5.96	6.01	6.03	6.05	6.10
α_{14}	5.93	6.01	6.09	1.00	1.01	5.93	5.98	6.01	6.04	6.09
α_{15}	6.08	6.19	6.29	1.03	1.07	6.08	6.15	6.19	6.22	6.29
α_{16}	6.11	6.19	6.27	1.01	1.03	6.10	6.16	6.19	6.22	6.26
α_{17}	6.00	6.07	6.14	1.01	1.02	5.99	6.04	6.07	6.09	6.14
σ_α	0.09	0.14	0.21	1.00	1.00	0.10	0.12	0.14	0.16	0.21
β	0.17	0.32	0.47	1.01	1.02	0.17	0.27	0.32	0.37	0.48
λ	0.07	0.12	0.19	1.02	1.04	0.07	0.10	0.12	0.14	0.18
τ	0.74	0.85	0.96	1.02	1.05	0.74	0.81	0.85	0.88	0.96
σ_{obs}	0.18	0.19	0.20	1.01	1.02	0.18	0.18	0.19	0.19	0.20
$\sigma_\alpha/\sigma_{obs}$	0.50	0.74	1.10	1.00	1.00	0.51	0.64	0.73	0.85	1.11
$-2\log(\text{density})$	727.81	747.33	766.86	1.01	1.01	729.98	739.92	746.88	753.84	768.35

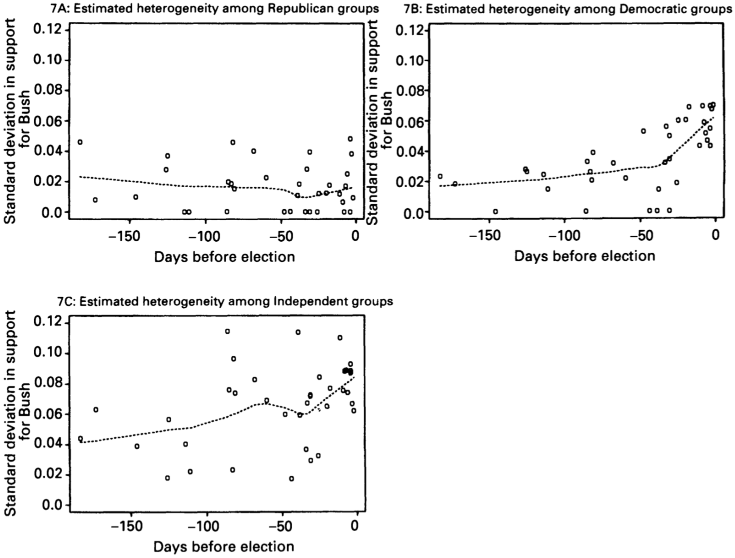
1992: "Inference from iterative simulation using multiple sequences"



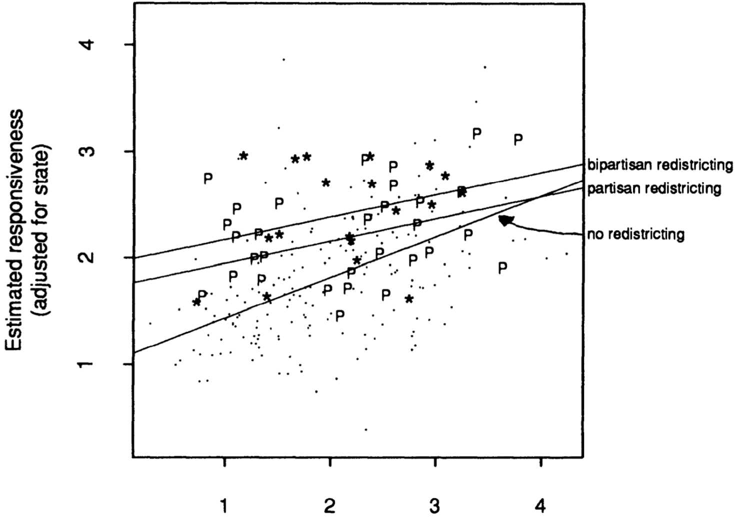
1993: "Why are American Presidential election campaign polls so variable when votes are so predictable?"



1993: "Why are American Presidential election campaign polls so variable when votes are so predictable?"



1994: "Enhancing democracy through legislative redistricting"



1995: "Pre-election survey methodology: details from nine polling organizations, 1988 and 1992"

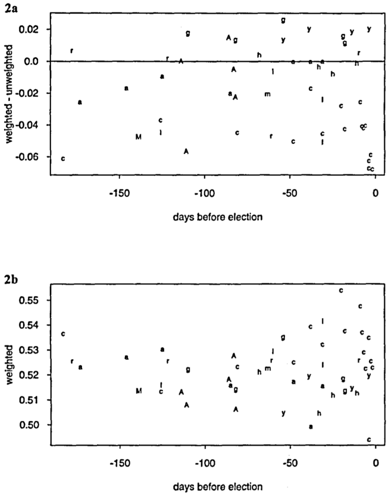
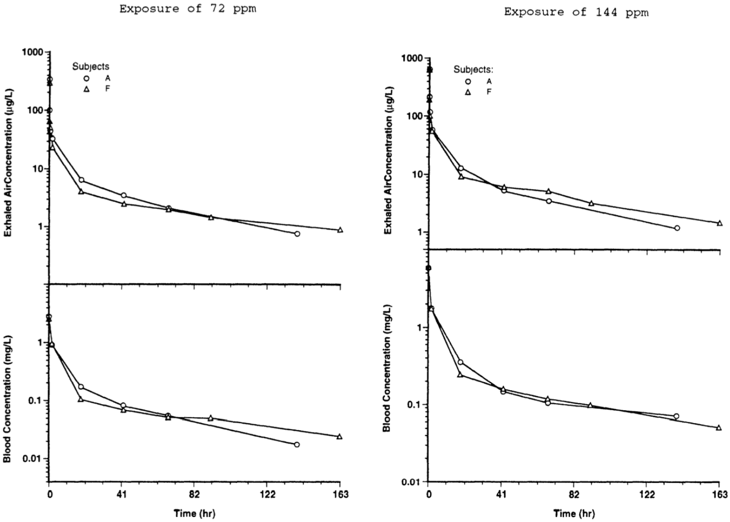


Figure 2. Fig. 2*a*, Effect of weighting on proportion of women. Fig. 2*b*, Proportion of women over time. *a* = ABC/*Washington Post*/ Chilton; *c* = CBS; *g* = Gallup; *h* = Harris; *l* = *Los Angeles Times;* *m* = Media General/AP; *r* = Roper; *y* = Yankelovich. Capital let-

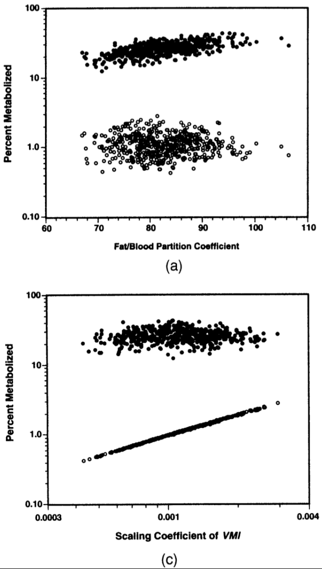
1996: "Physiological pharmacokinetic analysis using population modeling and informative prior distributions"



1996: “Physiological pharmacokinetic analysis using population modeling and informative prior distributions”

Parameter	Population prior	Posterior distributions for individuals						Population posterior
		A	*B*	*C*	*D*	*E*	*F*	
Ventilation/perfusion ratio (VPR)	1.6(×÷ 1.3) ×÷ 1.3	1.16 ×÷ 1.15	1.26 ×÷ 1.15	1.19 ×÷ 1.14	1.33 ×÷ 1.15	1.22 ×÷ 1.15	.961 ×÷ 1.15	1.19 ×÷ 1.13
Blood flow, well-perfused tissues (Fwp)	.47(×÷ 1.17) ×÷ 1.17	.653 ×÷ 1.06	.658 ×÷ 1.07	.647 ×÷ 1.07	.660 ×÷ 1.06	.626 ×÷ 1.08	.606 ×÷ 1.08	.637 ×÷ 1.06
Blood flow, poorly perfused tissues (Fpp)	.20(×÷ 1.22) ×÷ 1.22	.121 ×÷ 1.12	.123 ×÷ 1.13	.127 ×÷ 1.13	.123 ×÷ 1.12	.132 ×÷ 1.13	.134 ×÷ 1.13	.129 ×÷ 1.11
Blood flow, fat (Ff)	.07(×÷ 1.27) ×÷ 1.27	.048 ×÷ 1.13	.0442 ×÷ 1.13	.0462 ×÷ 1.14	.0437 ×÷ 1.13	.0507 ×÷ 1.14	.0582 ×÷ 1.14	.0488 ×÷ 1.12
Blood flow, liver (Fl)	.25(×÷ 1.15) ×÷ 1.15	.173 ×÷ 1.15	.170 ×÷ 1.16	.175 ×÷ 1.15	.168 ×÷ 1.15	.185 ×÷ 1.16	.195 ×÷ 1.15	.179 ×÷ 1.11
Volume, well-perfused tissues (Vwp)	.27(×÷ 1.36) ×÷ 1.36	.189 ×÷ 1.14	.201 ×÷ 1.15	.202 ×÷ 1.15	.201 ×÷ 1.15	.183 ×÷ 1.15	.188 ×÷ 1.14	.196 ×÷ 1.09
Volume, poorly perfused tissues (Vpp)	.55(×÷ 1.17) ×÷ 1.17	.649 ×÷ 1.04	.636 ×÷ 1.05	.636 ×÷ 1.05	.636 ×÷ 1.05	.655 ×÷ 1.04	.65 ×÷ 1.04	.641 ×÷ 1.03
Volume, liver (Vl)	.033(×÷ 1.1) ×÷ 1.1	.032 ×÷ 1.1	.033 ×÷ 1.1	.033 ×÷ 1.1	.033 ×÷ 1.1	.033 ×÷ 1.1	.032 ×÷ 1.1	.033 ×÷ 1.04
Partition coeff, blood/air (Pba)	12(×÷ 1.5) ×÷ 1.3	15.1 ×÷ 1.04	16.4 ×÷ 1.03	15.3 ×÷ 1.04	15.6 ×÷ 1.04	18.7 ×÷ 1.04	15.8 ×÷ 1.04	16.0 ×÷ 1.11
Partition coeff, well-perfused (Pwp)	4.8(×÷ 1.5) ×÷ 1.3	1.83 ×÷ 1.15	1.98 ×÷ 1.16	1.95 ×÷ 1.16	2.00 ×÷ 1.16	1.83 ×÷ 1.15	1.83 ×÷ 1.14	1.92 ×÷ 1.12
Partition coeff, poorly perfused (Ppp)	1.6(×÷ 1.5) ×÷ 1.3	2.94 ×÷ 1.08	2.59 ×÷ 1.09	2.51 ×÷ 1.09	2.76 ×÷ 1.08	4.06 ×÷ 1.09	2.96 ×÷ 1.09	2.90 ×÷ 1.15
Partition coeff, fat (Pf)	125(×÷ 1.5) ×÷ 1.3	82.3 ×÷ 1.08	69.1 ×÷ 1.08	73.9 ×÷ 1.08	49.1 ×÷ 1.08	171 ×÷ 1.09	85.4 ×÷ 1.07	84.1 ×÷ 1.28
Partition coeff, liver (Pl)	4.8(×÷ 1.5) ×÷ 1.3	2.93 ×÷ 1.32	3.07 ×÷ 1.33	3.21 ×÷ 1.32	3.09 ×÷ 1.33	3.16 ×÷ 1.33	2.94 ×÷ 1.32	3.08 ×÷ 1.12
Max metabolic rate in liver (VMl)	.042(×÷ 10) ×÷ 2	.0011 ×÷ 1.41	.00139 ×÷ 1.37	.00214 ×÷ 1.30	.00199 ×÷ 1.34	.00415 ×÷ 1.30	.00165 ×÷ 1.38	.00191 ×÷ 1.45
K_m in liver (KMl)	16(×÷ 10) ×÷ 1.5	.801 ×÷ 1.63	.754 ×÷ 1.61	.660 ×÷ 1.59	.742 ×÷ 1.57	.650 ×÷ 1.59	.771 ×÷ 1.60	.729 ×÷ 1.20

1996: "Physiological pharmacokinetic analysis using population modeling and informative prior distributions"



1997: "Poststratification into many categories using hierarchical logistic regression"

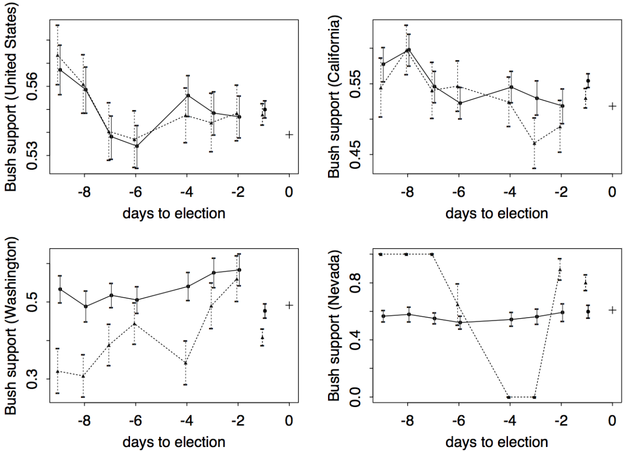


Figure 3: Estimated Bush support estimated separately from seven individual polls taken shortly before the election: for (a) the entire U.S. (excluding Alaska, Hawaii, and the District of Columbia), (b) a large state (California), (c) a medium-sized state (Washington), and (d) a small state (Nevada). Each plot shows the raking estimates as a dotted line and the estimates from hierarchical model

1998: "Estimating the probability of events that have never occurred: When is your vote decisive?"

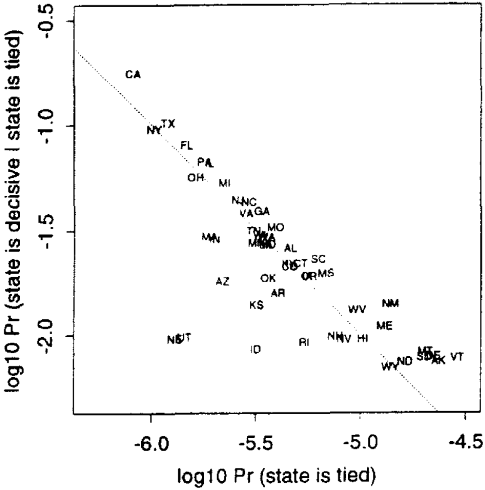
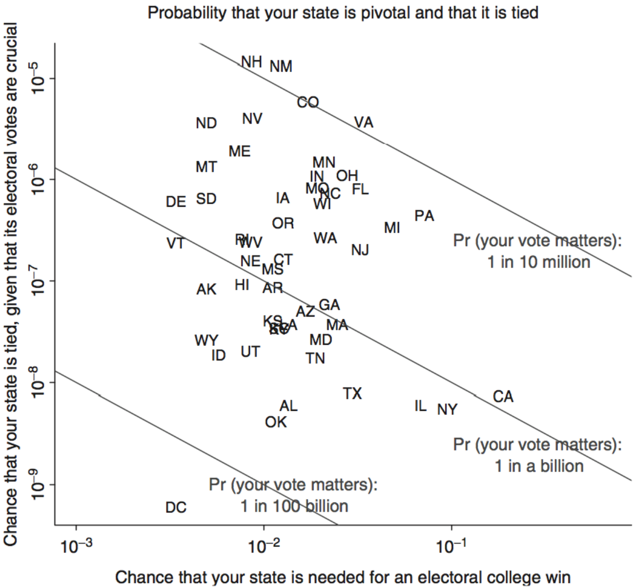


Figure 3. Probability That a State Is Decisive Given Tied Versus the Probability That the State Is Tied for 1992 Plotted on a Log Scale. ···,

2009: "The probability your vote will make a difference"



1999: “All maps of parameter estimates are misleading”

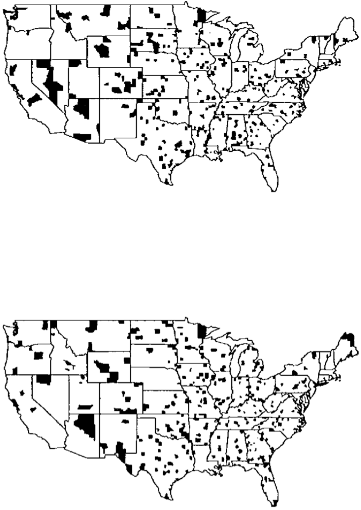


Fig. 6. Four multiple imputations. For each map, the shaded counties are those in which the imputed rates, θ_j, drawn from their posterior distribution, are in the top 10 per cent of U.S. counties, for that imputation. Compare these maps to the map of the highest true county parameters in Figure 4. These maps have no systematic artefacts due to variation in the

2000: "Type S error rates for classical and Bayesian single and multiple comparison procedures"

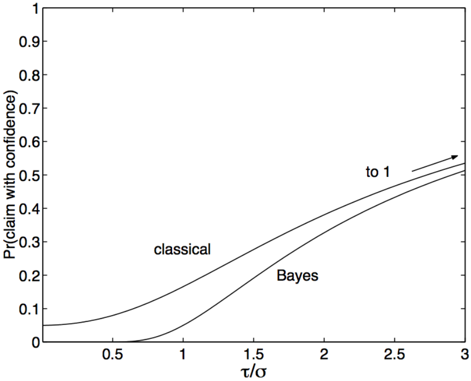
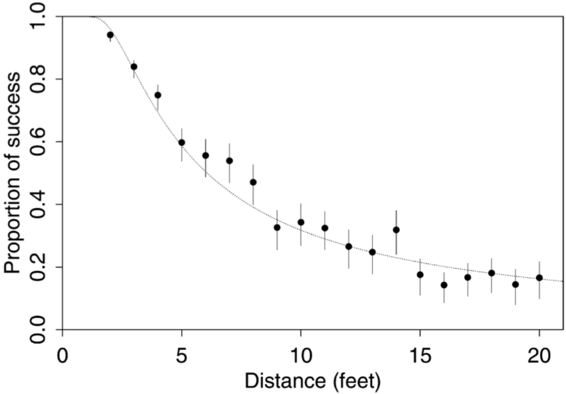
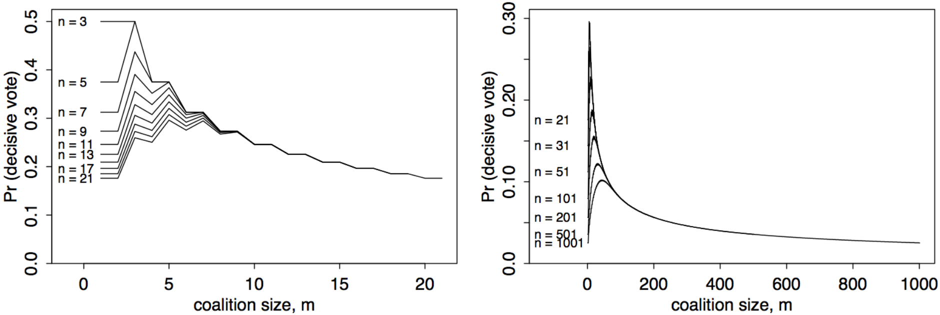


Figure 2: Probability of making a claim with confidence for classical and Bayesian comparisons: long-run frequencies are shown as a function of the variance ratio τ/σ.

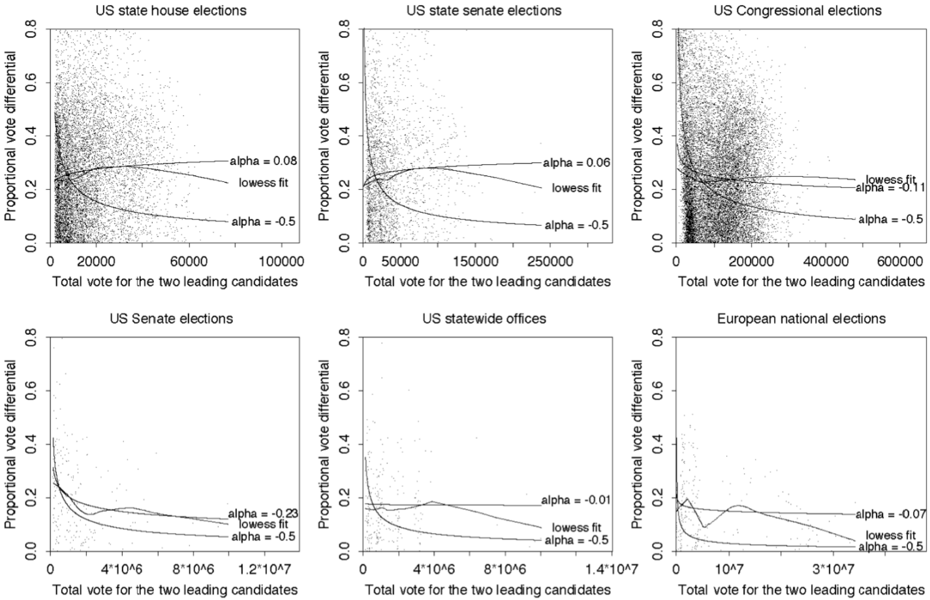
2002: "A probability model for golf putting"



2003: "Forming voting blocs and coalitions as a prisoner's dilemma: a possible theoretical explanation for political instability"



2004: "Standard voting power indexes don't work"



2005: “Multiple imputation for model checking: completed-data plots with missing and latent data”

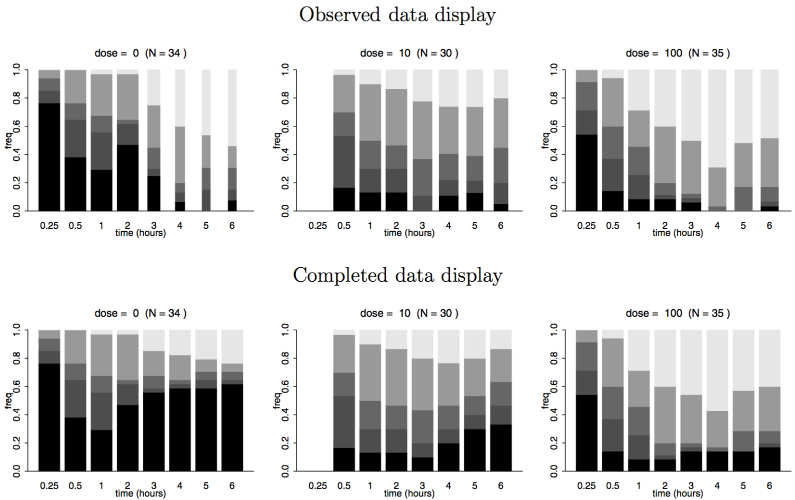
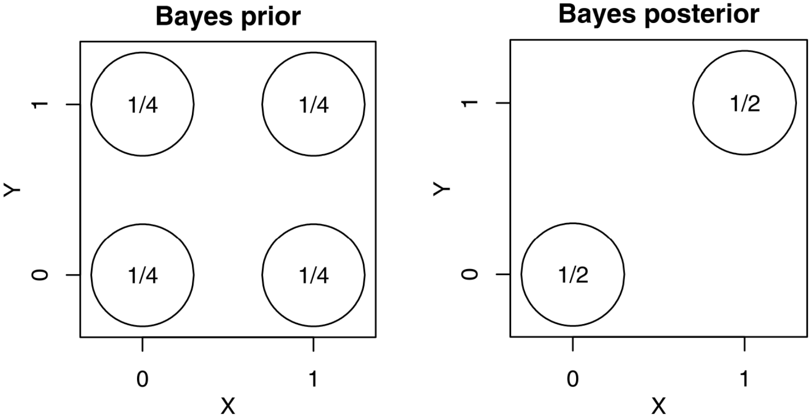
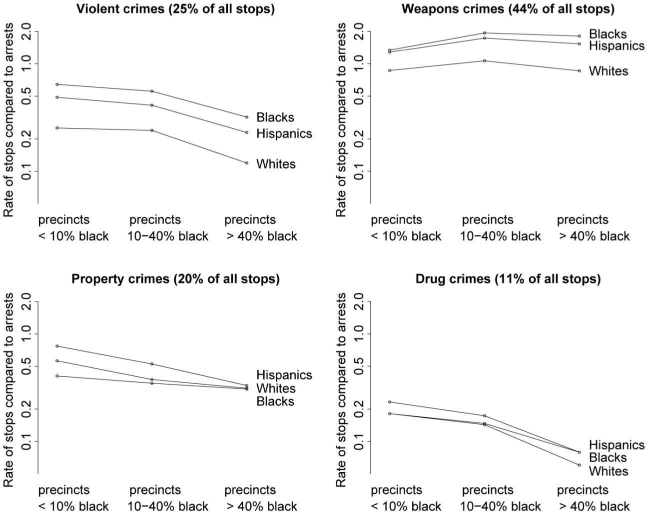


Figure 4. Summary of pain-relief responses over time under different doses from the clinical trial with nonignorable dropout discussed in Section 3.2. In each summary bar, the shadings from bottom to top indicate “no pain relief” and intermediate levels up to “complete pain relief.” The graphs in the top row include only the persons who have not dropped out (with the

2006: “The boxer, the wrestler, and the coin flip”



2007: "An analysis of the NYPD's stop-and-frisk policy in the context of claims of racial bias"



2009: "Beautiful political data"

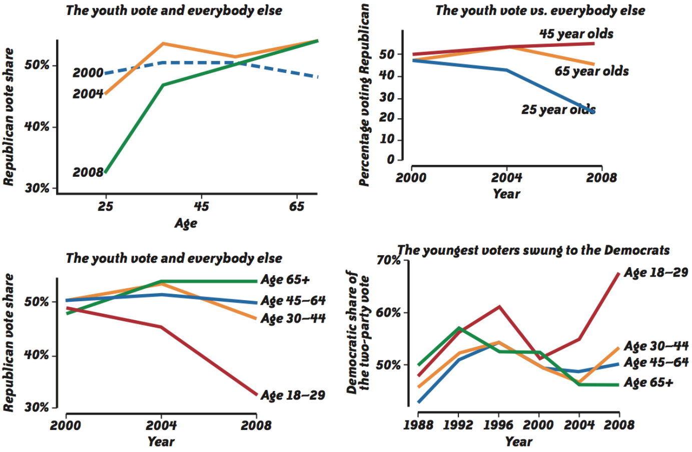


FIGURE 19-3. Some graphs showing recent patterns of voting by age. The top-left graph shows my first attempt, created on election night based on immediate exit poll data. The top-right graph was created by Hober Short, a student who saw my graph on the Web and made his own, displaying time on the x-axis. The lower-left graph is my cleaned-up version of Short's graph, labeling all four age categories directly on the lines of the graph. All these graphs show the dramatic difference between 2008 and the two previous elections. Finally, the lower-right graph extends the data back to 1988, showing that Bill Clinton in 1996

2010: “Public opinion on health care reform”

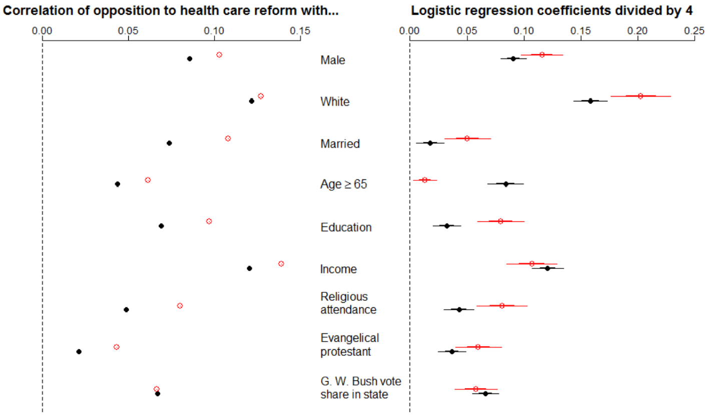
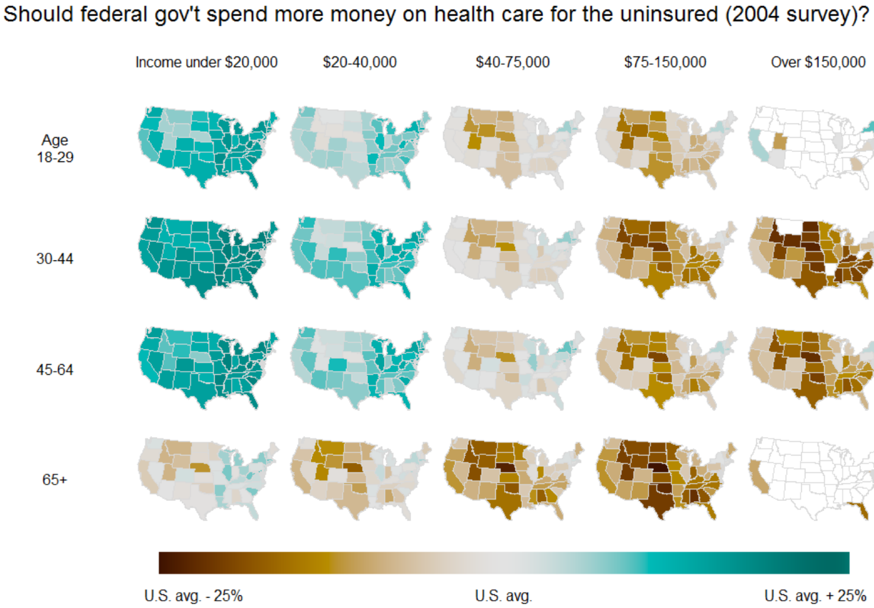
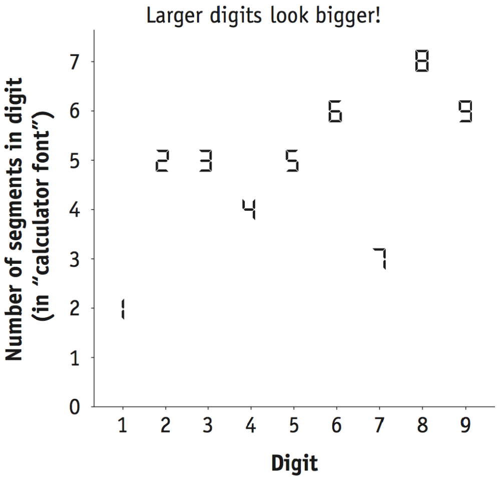


Figure 1. Correlations and logistic regression coefficients for predicting opposition to health care reform. The black closed circles are estimates for 2000 and the red open circles correspond to 2004. Logistic regression coefficients have been divided by 4 to correspond to approximate changes on the probability scale (e.g., Gelman and Hill, 2007), and the continuous inputs in the regression have been scaled by dividing by two standard deviations so that their coefficients are comparable to those of binary predictors

2010: "Public opinion on health care reform"



2011: "Tables as graphs: The Ramanujan principle"



2012: "Philosophy and the practice of Bayesian statistics"

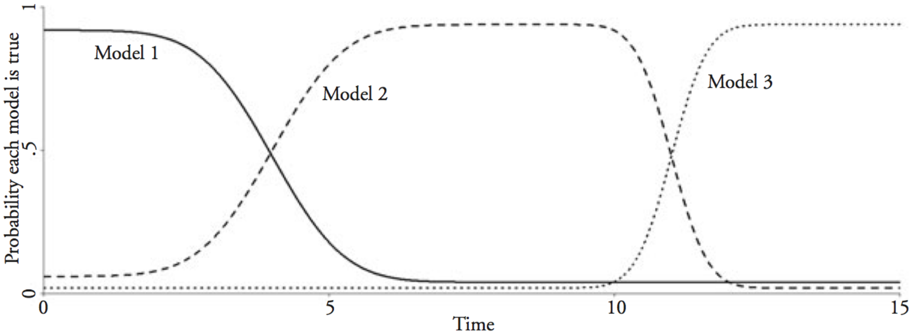
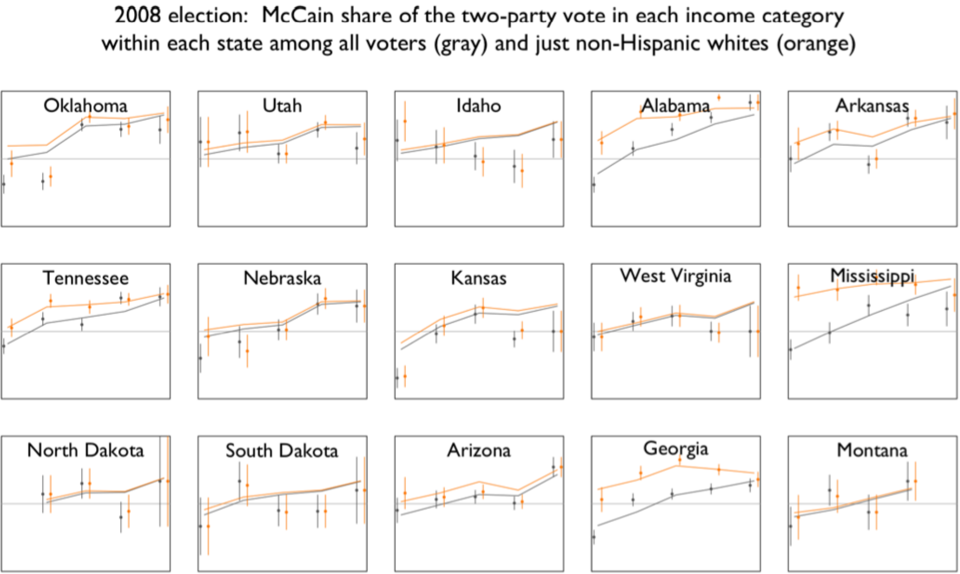


Figure 1. Hypothetical picture of idealized Bayesian inference under the conventional inductive philosophy. The posterior probability of different models changes over time with the expansion of the likelihood as more data are entered into the analysis. Depending on the context of the problem, the time scale on the *x*-axis might be hours, years, or decades, in any case long enough for information to be gathered and analysed that first knocks out hypothesis 1 in favour of hypothesis 2, which in turn is dethroned in favour of the current champion, model 3.

2013: "Election turnout and voting patterns"

2013: "Election turnout and voting patterns"



Notes

- Gradual improvements in technique . . . and understanding
- Often, what we're plotting is *not* "data"
- Research vs. publications: "Let me tell you about my first wife"

Take-home points

- Small multiples
- Line plots
- Try to make a display self-contained, then add words
- Graphs are comparisons

Some references

Andrew Gelman and Antony Unwin (2013). Infovis and statistical graphics: Different goals, different looks (with discussion by Stephen Few, Robert Kosara, Paul Murrell, and Hadley Wickham, and rejoinder by Gelman and Unwin). *Journal of Computational and Graphical Statistics*. [Our current views on tradeoffs in statistical graphics]

Andrew Gelman (2004). Exploratory data analysis for complex models (with discussion by Andreas Buja and rejoinder by Gelman). *Journal of Computational and Graphical Statistics* **13**, 755–787. [An expression of the idea that exploratory graphics are a form of model checking: the better the model, the more effective the graphics. Thus, statistical modeling and graphics are not competitors (as is often thought) but can work together.]

Andrew Gelman (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* **71**, 369–382. [A more formal exploration of the unity between statistical graphics and Bayesian modeling.]

Andrew Gelman, Cristian Pasarica, and Rahul Dodhia (2002). Let's practice what we preach: turning tables into graphs. *American Statistician* **56**, 121–130. [Proof of concept: we went through an issue of the *Journal of the American Statistical Association* and converted all the tables into graphs, in each case displaying all the information using less space.]

Andrew Gelman and Gary King (1993). Why are American Presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* **23**, 409–451. [We resolved in writing this paper to do all the analysis using graphs, no tables. It worked well: we told a story and backed it up with evidence.]