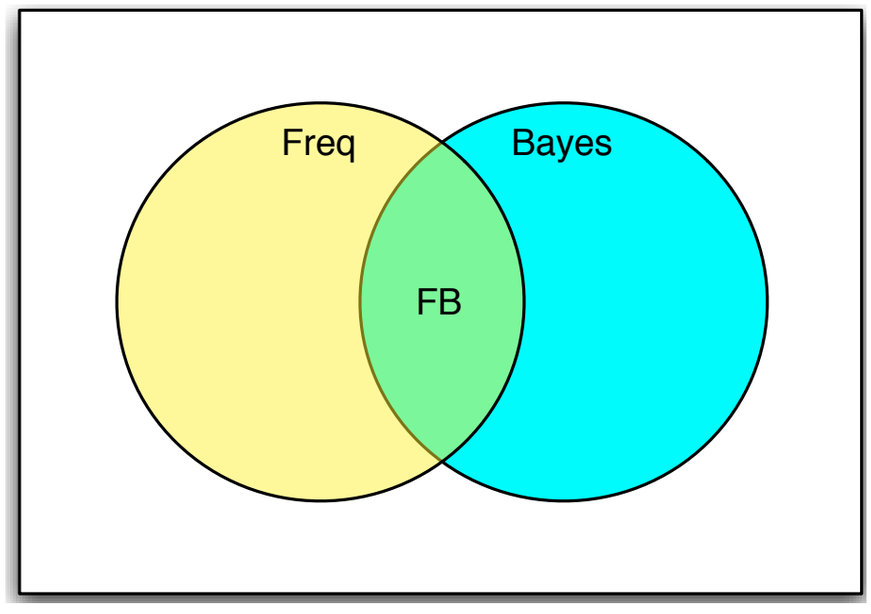
CS109/Stat121/AC209/E-109
Data Science
Bayesian Methods

Hanspeter Pfister & Joe Blitzstein
pfister@seas.harvard.edu / blitzstein@stat.harvard.edu



This Week

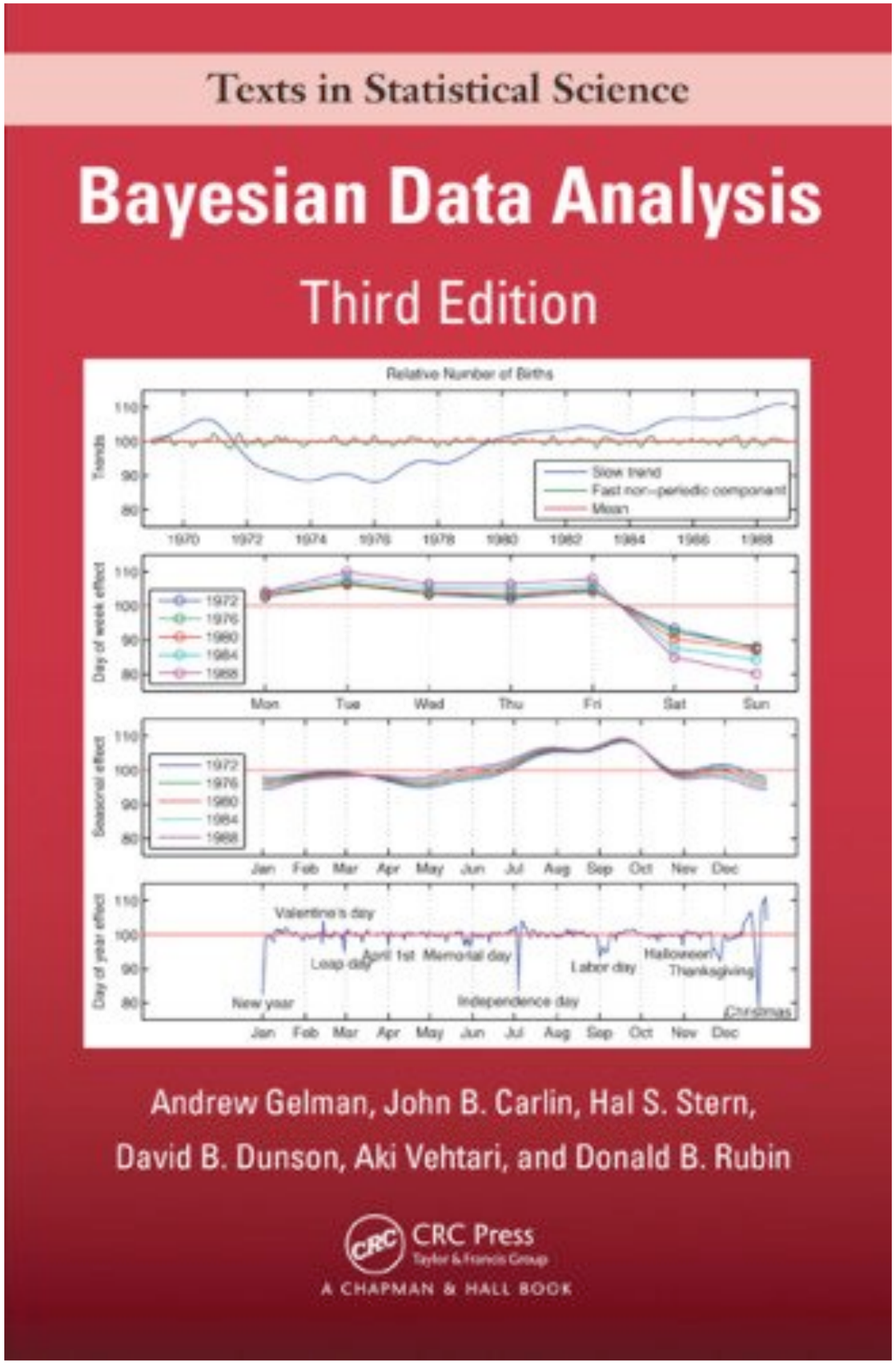
- HW3 due next Thursday (Oct 17) at 11:59 pm – start now!
- Friday lab **10-11:30 am** in MD G115

Nate Silver is Hiring:
The Rise of Data Journalism?

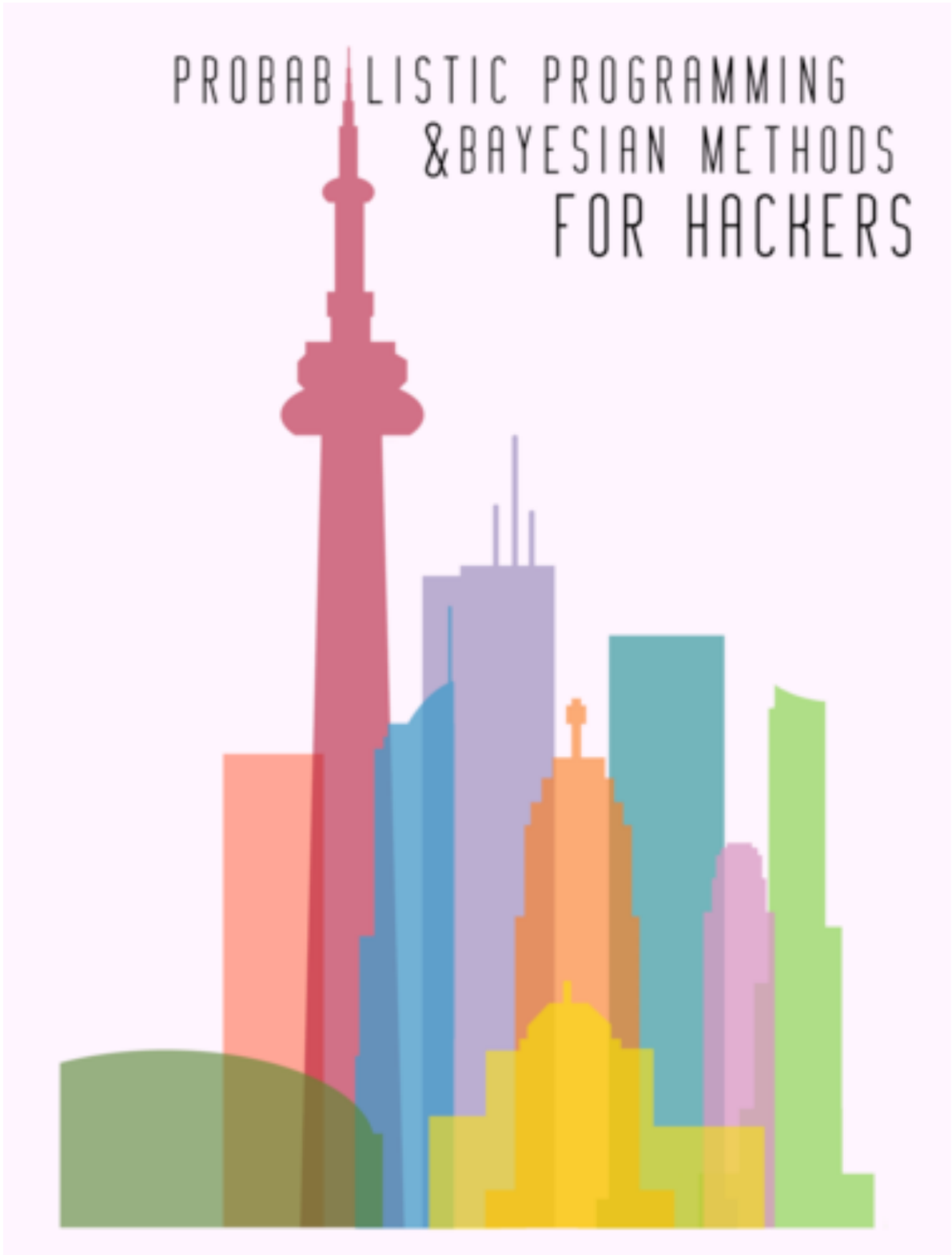
http://www.fivethirtyeight.com/2013/09/seeking-lead-writers-in-sports-politics.html

"FiveThirtyEight is conducting a search for lead writers in three of our most important content verticals: sports, politics and economics. ...
These are high-profile, full-time positions for people with an outstanding combination of writing and statistical skills. ...
We are intrigued by candidates who can combine traditional reporting with critical, empirical analysis ...
Programming skills, database skills, and familiarity with statistical software packages are clear positives. ...
So is the demonstrated ability to produce high-quality charts, graphics, and interactive features."

Bayesian Data Analysis



Probabilistic Programming and Bayesian Methods for Hackers



http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Prologue/Prologue.ipynb

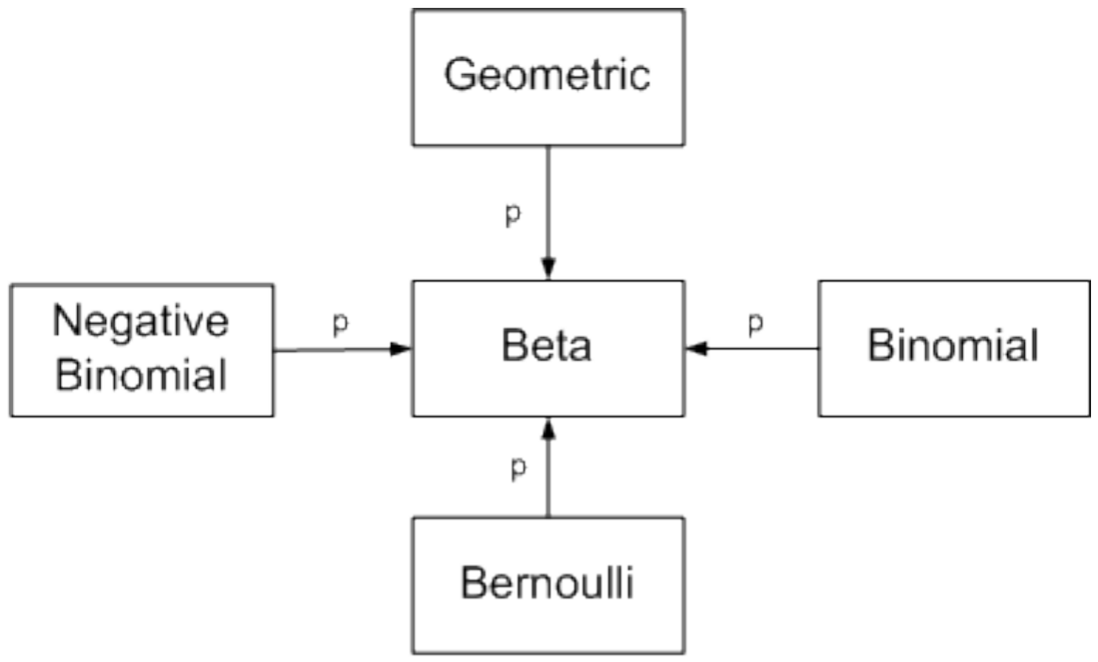
Full Probability Modeling

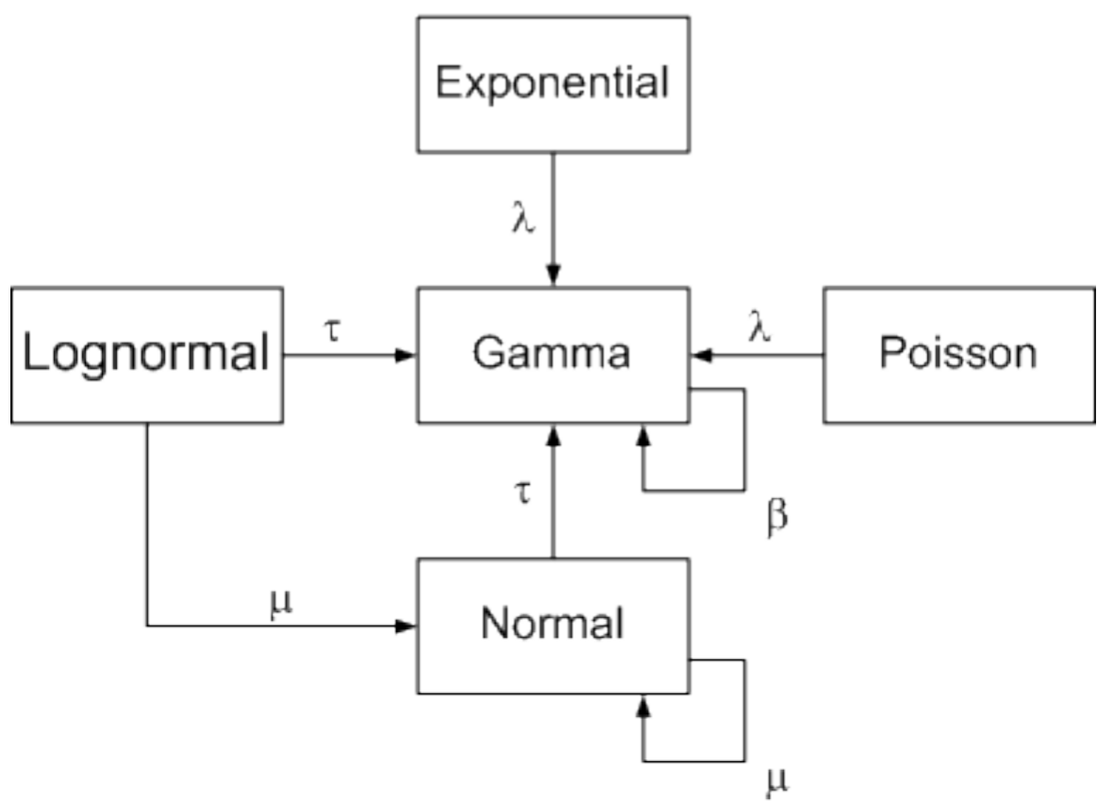
"The process of Bayesian data analysis can be idealized by dividing it into the following three steps:

1. Setting up a full probability model – a joint probability distribution for all observable and unobservable quantities in a problem...
2. Conditioning on observed data: calculating and interpreting the appropriate posterior distribution – the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.
3. Evaluating the fit of the model and the implications of the resulting posterior distribution..."

-- Gelman et al, Bayesian Data Analysis

Conjugate Priors





http://www.johndcook.com/conjugate_prior_diagram.html

Ranking Reddit Comments:

Example from Probabilistic Programming and Bayesian Methods for Hackers

[–] **Cocky_All_Day** 1957 points 2 days ago (**2355**|**401**)

If I ever found myself being attacked by a bear, what advice would you give me?

permalink source report give gold save **reply** hide child comments

> [–] **allenahansen** [S] 4848 points 2 days ago (**9861**|**5007**)
>
> If it's a Grizzly Bear, play dead. If you're in California, it's a Black Bear. Fight back with everything you've got because it's trying to kill you. If it's a Polar Bear, you're fucked.
>
> permalink source parent report save give gold **reply**
>
> > [–] **ExBoop** 4052 points 2 days ago (**6960**|**2910**)
> >
> > If it's brown, stay down. If it's black, attack. Now confirmed to be true.
> >
> > permalink source parent report save give gold **reply**
> >
> > > [–] **IrkenInvaderGir** 4439 points 2 days ago (**8395**|**3948**)
> > >
> > > If it's white, good night.
> > >
> > > permalink source parent report save give gold **reply**

http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter4_TheGreatestTheoremNeverTold/LawOfLargeNumbers.ipynb

Ranking Reddit Comments: A Simple Model

$$\text{number of upvotes} \sim \text{Bin}(n, p)$$

$$\text{conjugate prior: } p \sim \text{Beta}(a, b), \text{ pdf} \propto p^{a-1}(1-p)^{b-1}$$

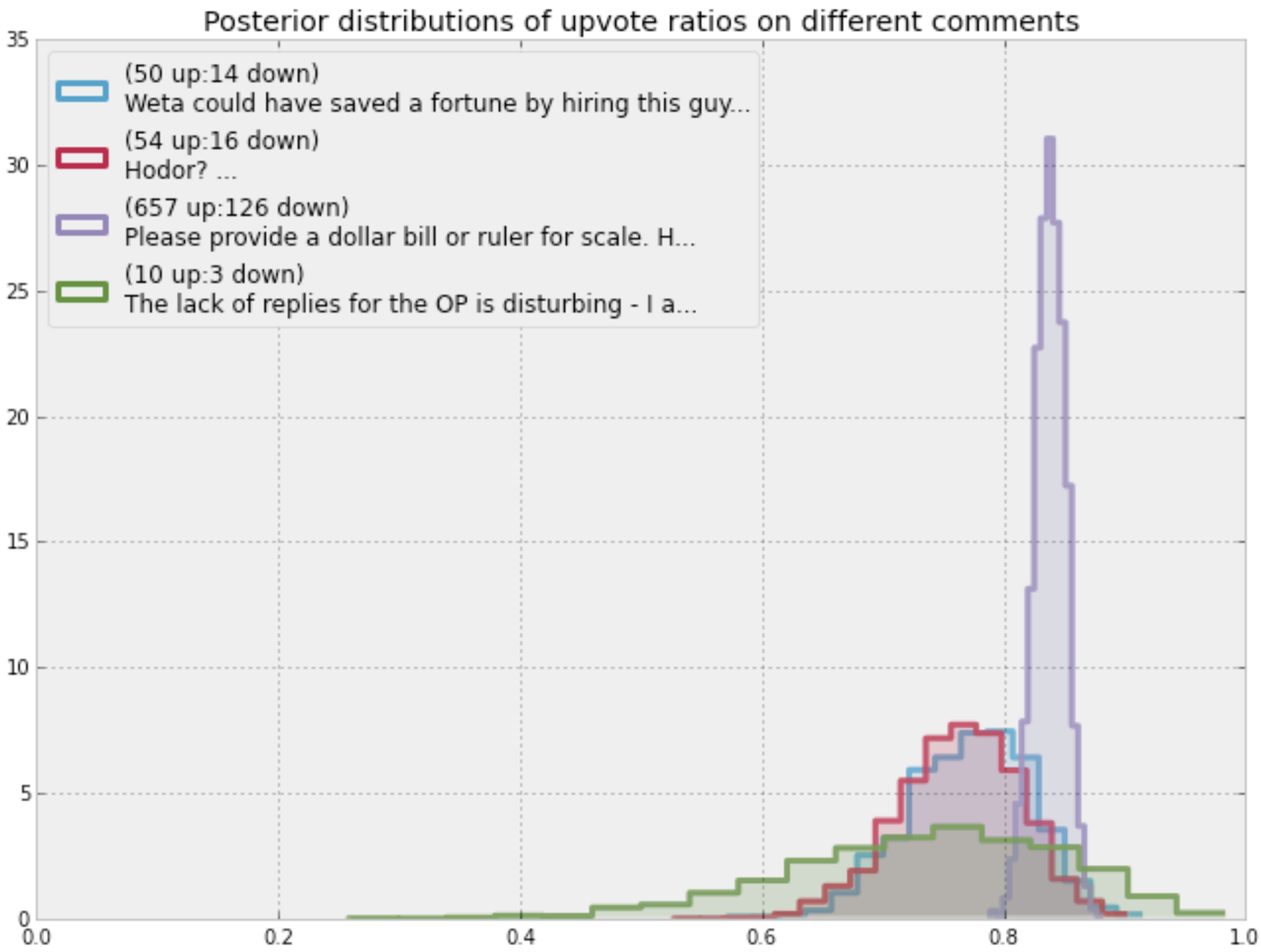
$$\text{posterior: } p|\text{data} \sim \text{Beta}(a + \#\text{upvotes}, b + \#\text{downvotes})$$

Ranking Reddit Comments

Why not just add "pseudocounts" and then use proportion? Why bother with Bayes?

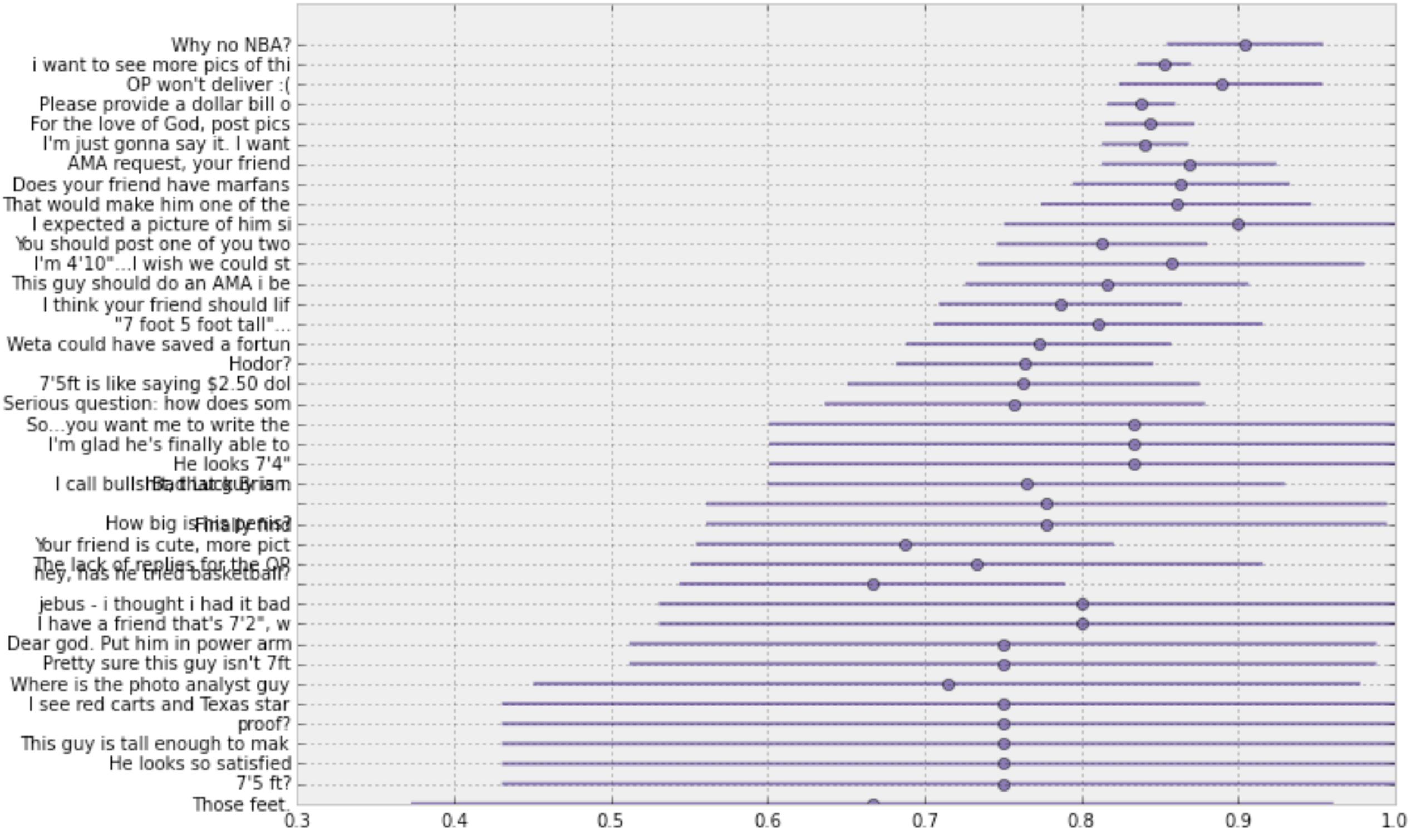
For example, the Agresti-Coull method adds 2 successes and 2 failures.

Posterior Distributions for Reddit Comments



http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter4_TheGreatestTheoremNeverTold/LawOfLargeNumbers.ipynb

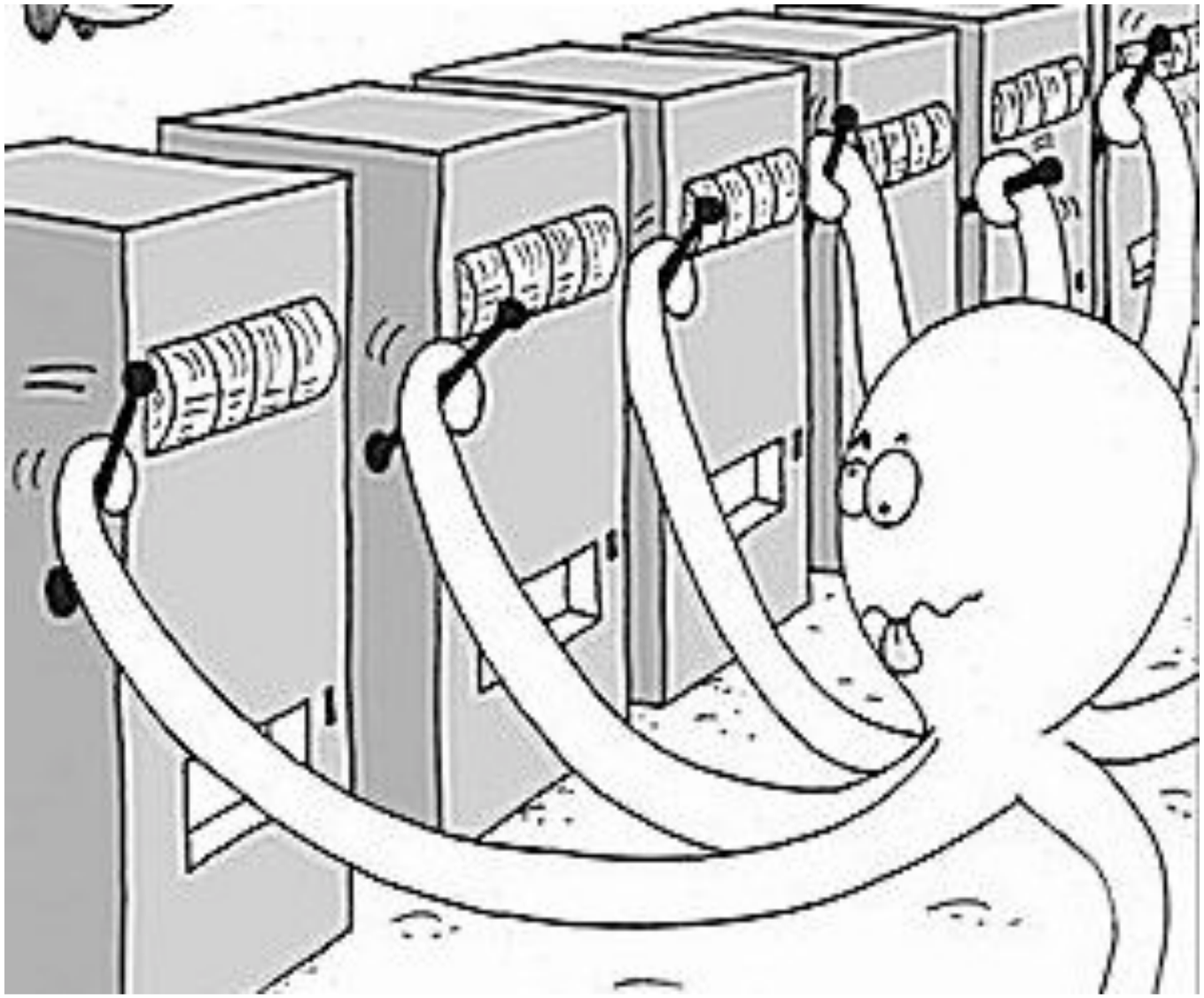
Ranking Reddit Comments by Posterior Quantiles



http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter4_TheGreatestTheoremNeverTold/LawOfLargeNumbers.ipynb

Bayesian Bandits

Example from Probabilistic Programming and Bayesian Methods for Hackers



http://research.microsoft.com/en-us/projects/bandits/

N slot machines, each with its own unknown probability of giving a prize. Exploration-exploitation tradeoff.

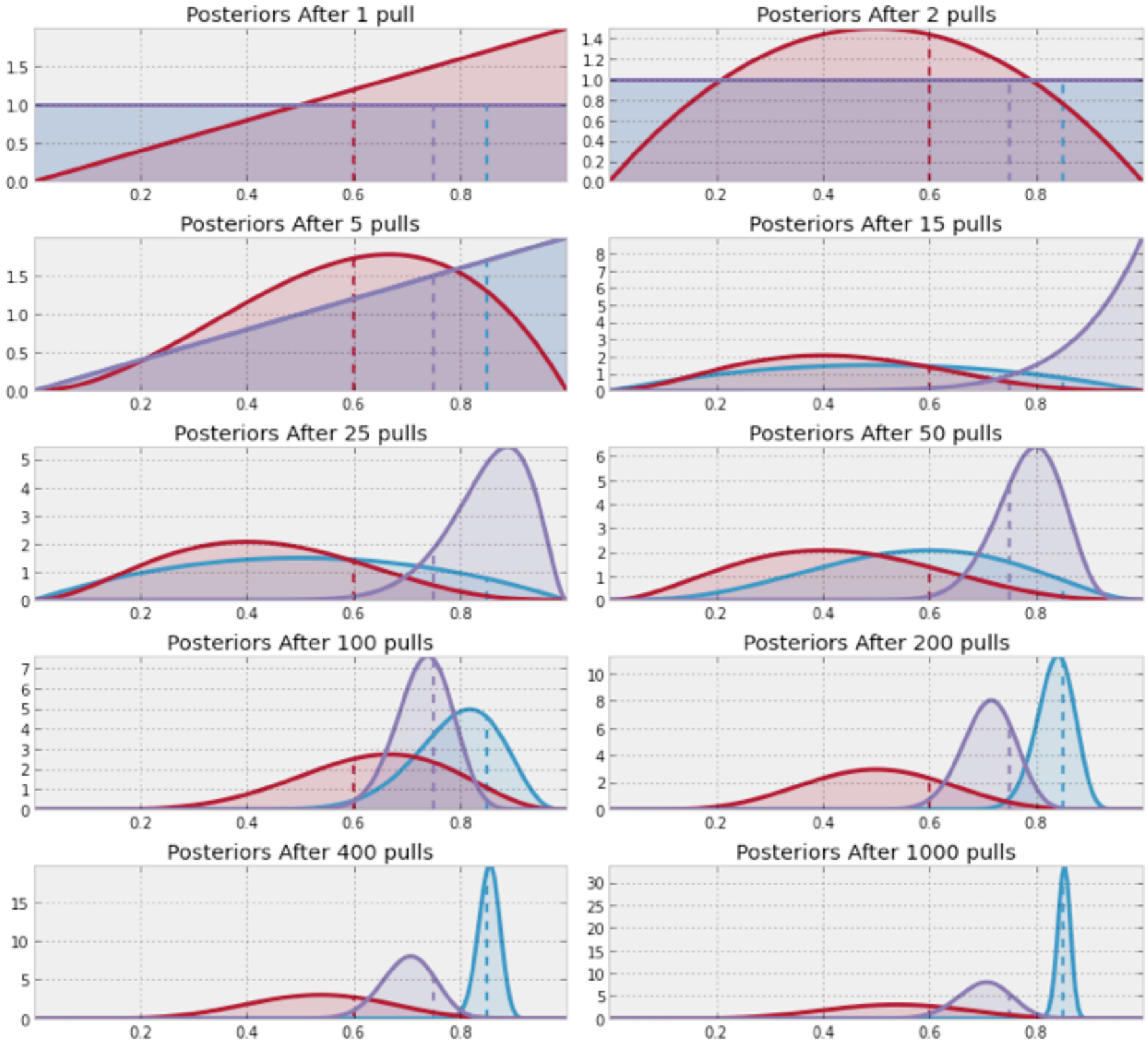
Bayesian Bandits

Example from Probabilistic Programming and Bayesian Methods for Hackers

A fast, simple Bayesian algorithm:

1. sample from the prior of each bandit
2. select the bandit with the largest sampled value
3. update the prior for that bandit (the posterior becomes the new prior)
4. repeat.

Bayesian Bandits



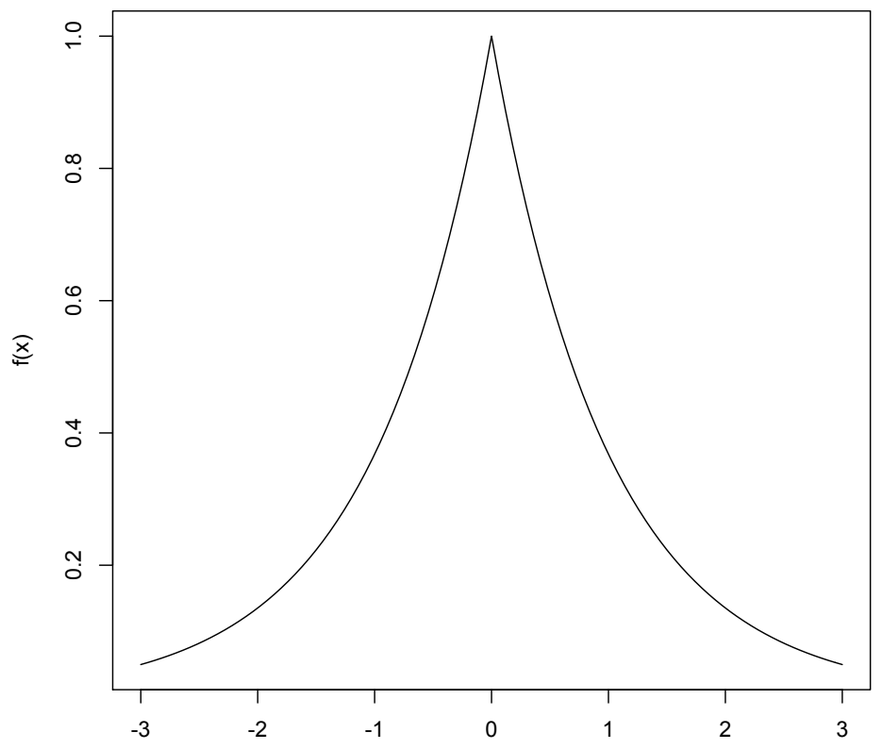
http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Chapter6_Priorities/Priors.ipynb

LASSO and Sparsity

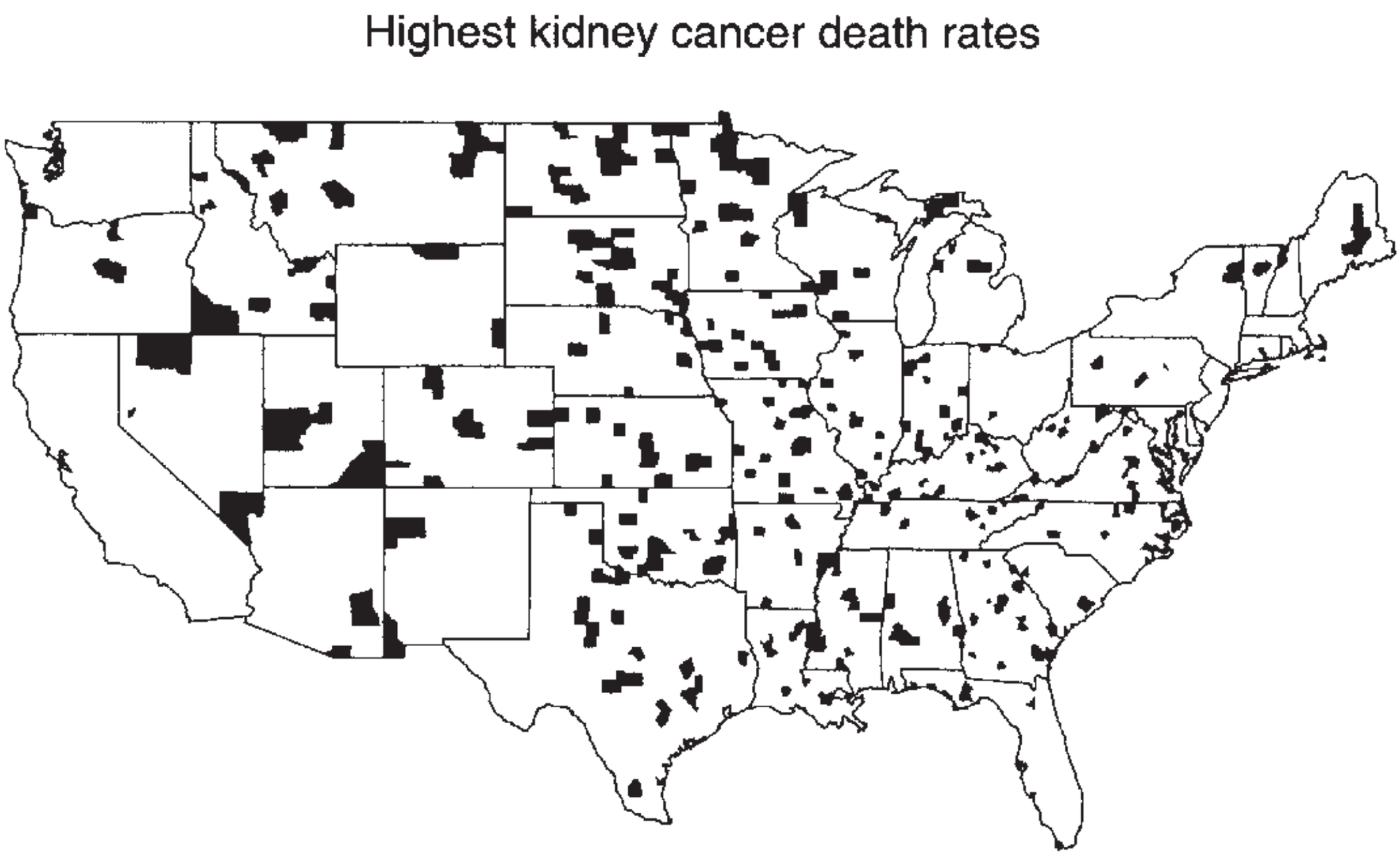
In a linear regression model, in place of minimizing the sum of squared residuals, LASSO says to minimize

$$\mathrm{SSR}(\boldsymbol{\beta} : \lambda) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,.$$

Bayesian interpretation: posterior mode, with independent Laplace priors on the parameters.



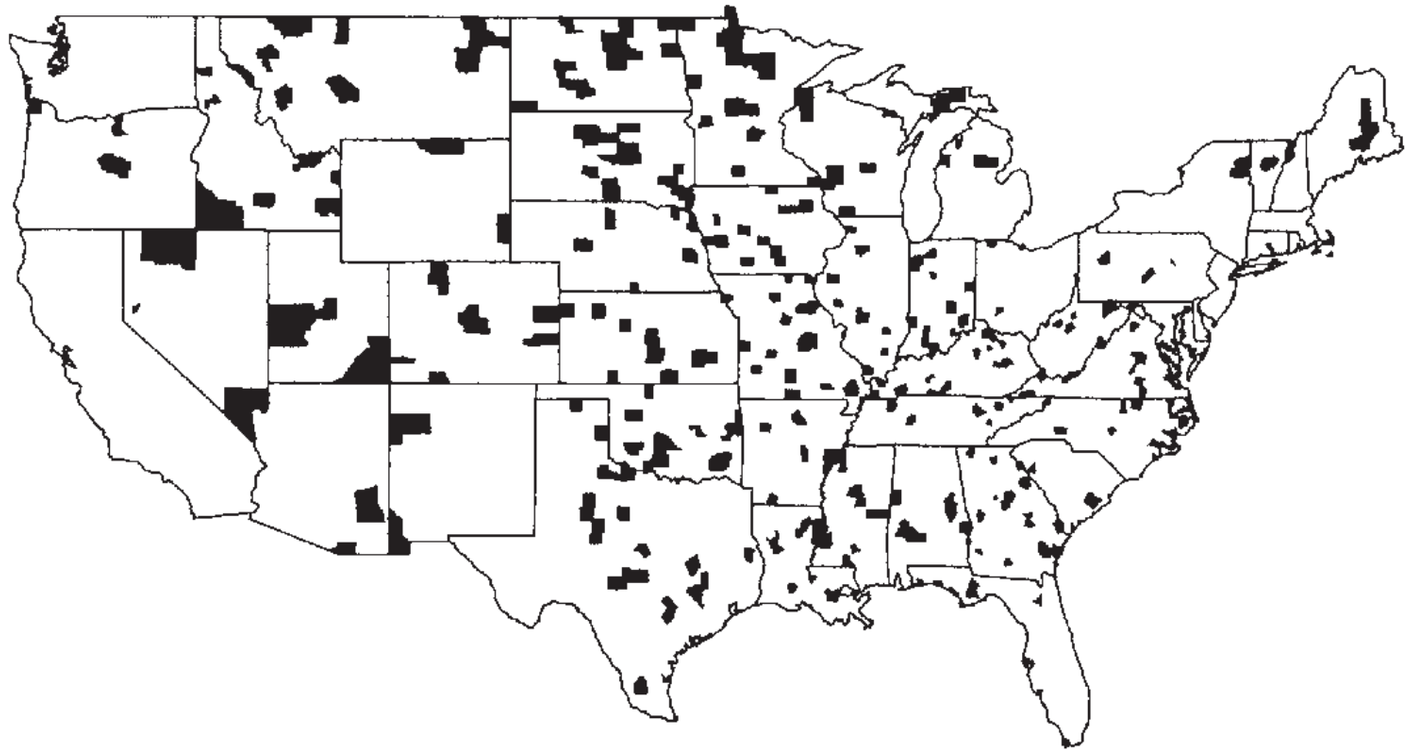
Kidney Cancer Example from Bayesian Data Analysis (Gelman et al)



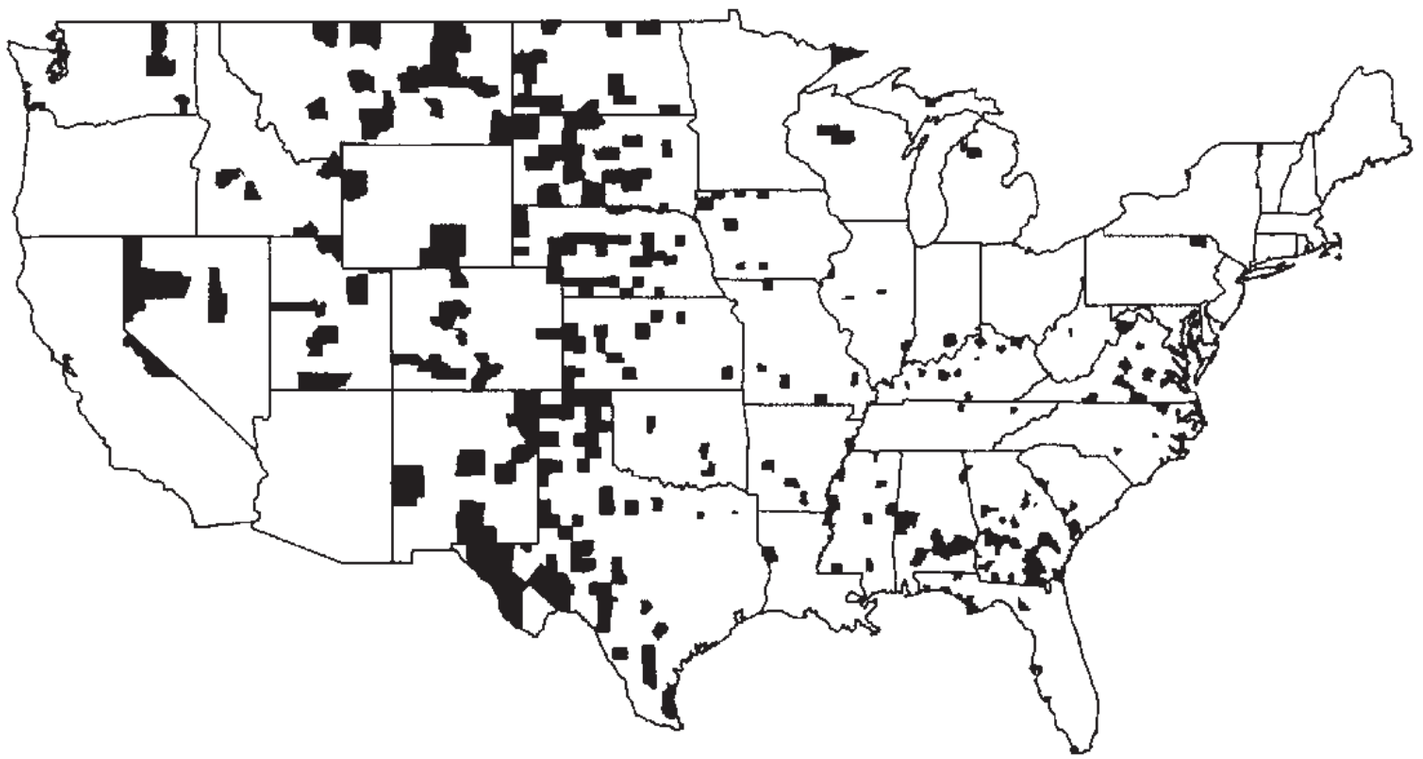
U.S. counties with the highest 10% of kidney cancer death rates (age-adjusted)

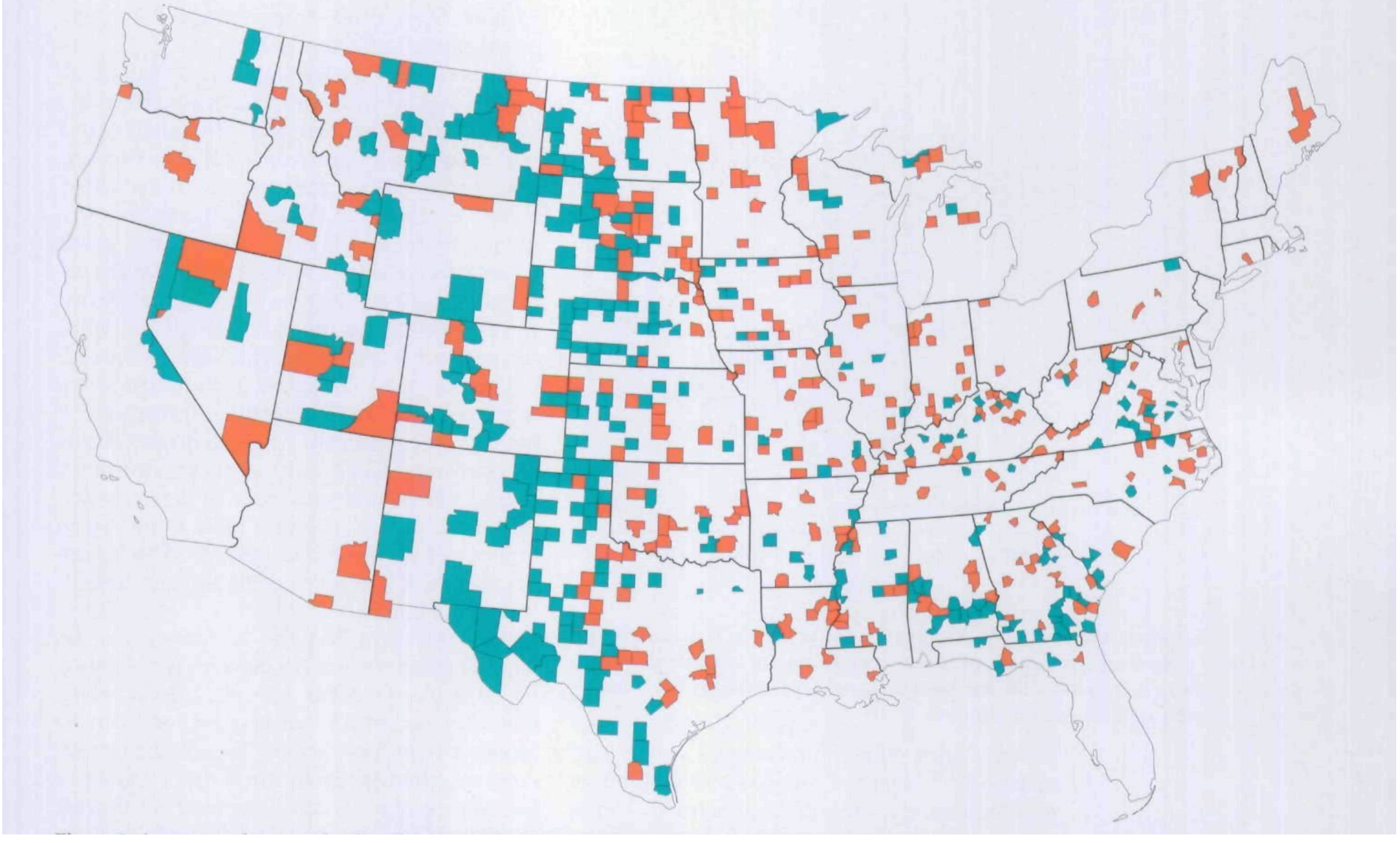
Kidney Cancer Example from Bayesian Data Analysis (Gelman et al)

Highest kidney cancer death rates



Lowest kidney cancer death rates

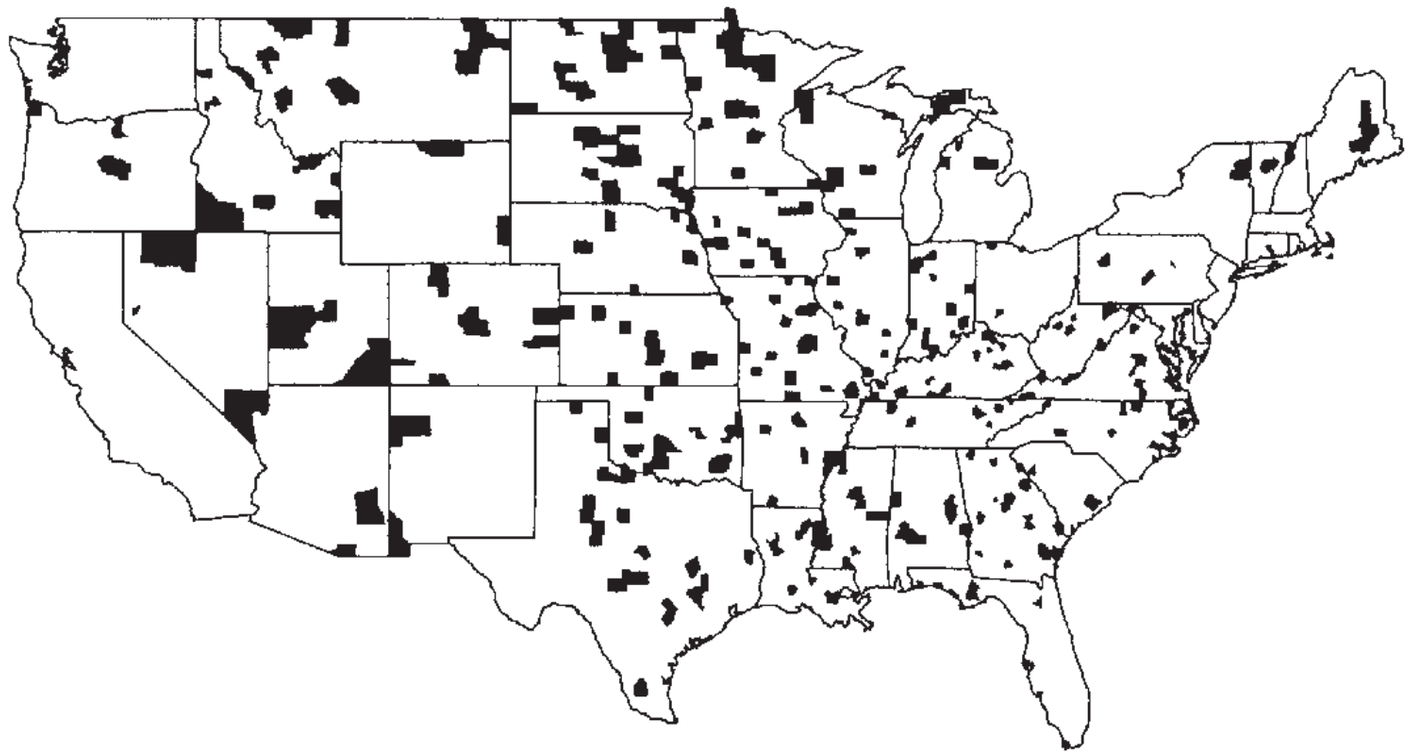




Wainer, "The Most Dangerous Equation" (American Scientist, 2007)

teal: lowest 10%
orange: highest 10%

Kidney Cancer Example from Bayesian Data Analysis (Gelman et al)



simple model:

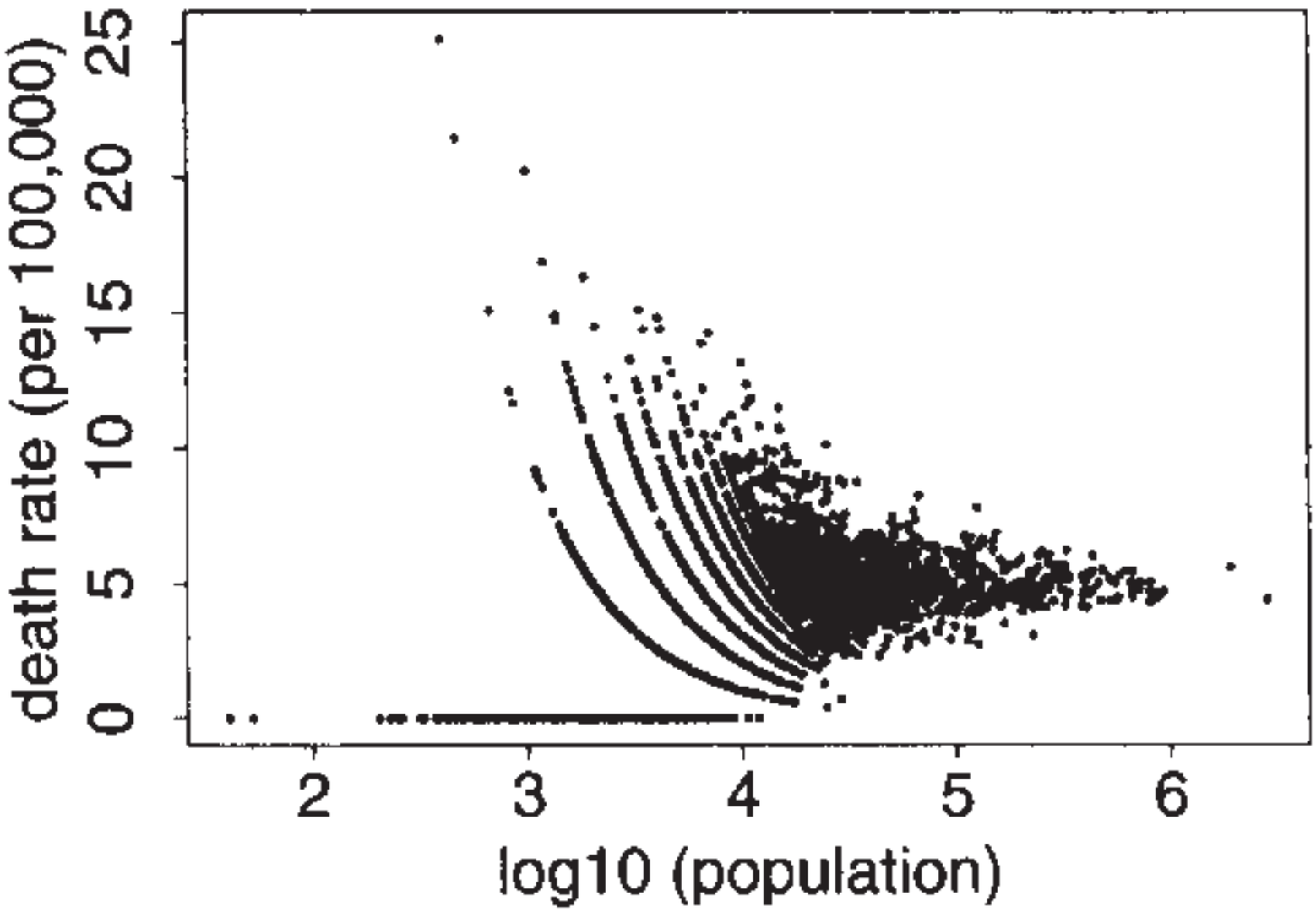
$$y_j \sim \text{Pois}(10 n_j \theta_j)$$

$$\theta_j \sim \text{Gamma}(\alpha, \lambda)$$

$$E(\theta_j | y_j) = w \frac{y_j}{10 n_j} + (1 - w) E(\theta_j)$$

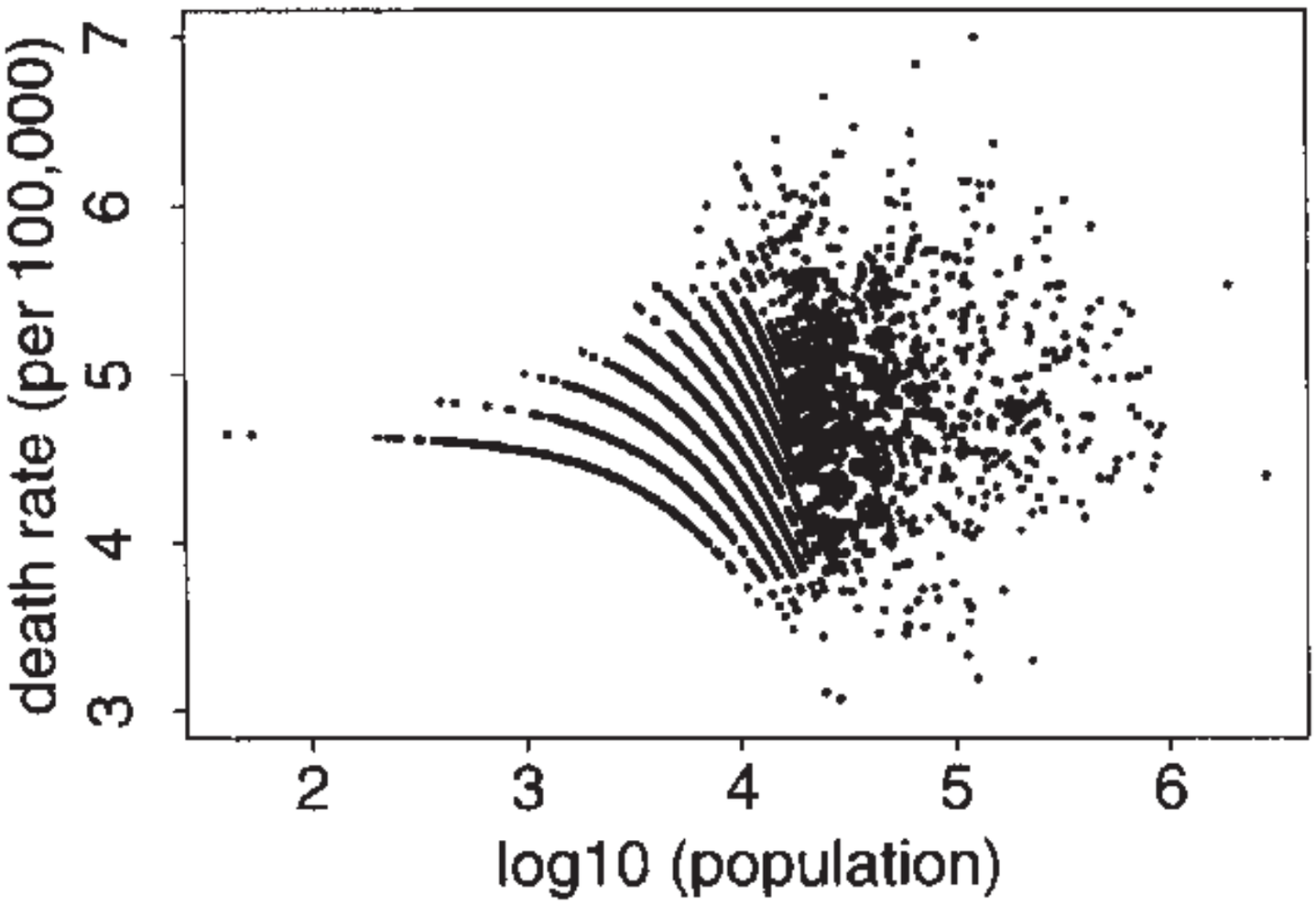
weighted combination of the data and the prior mean

Kidney Cancer Example from Bayesian Data Analysis (Gelman et al)



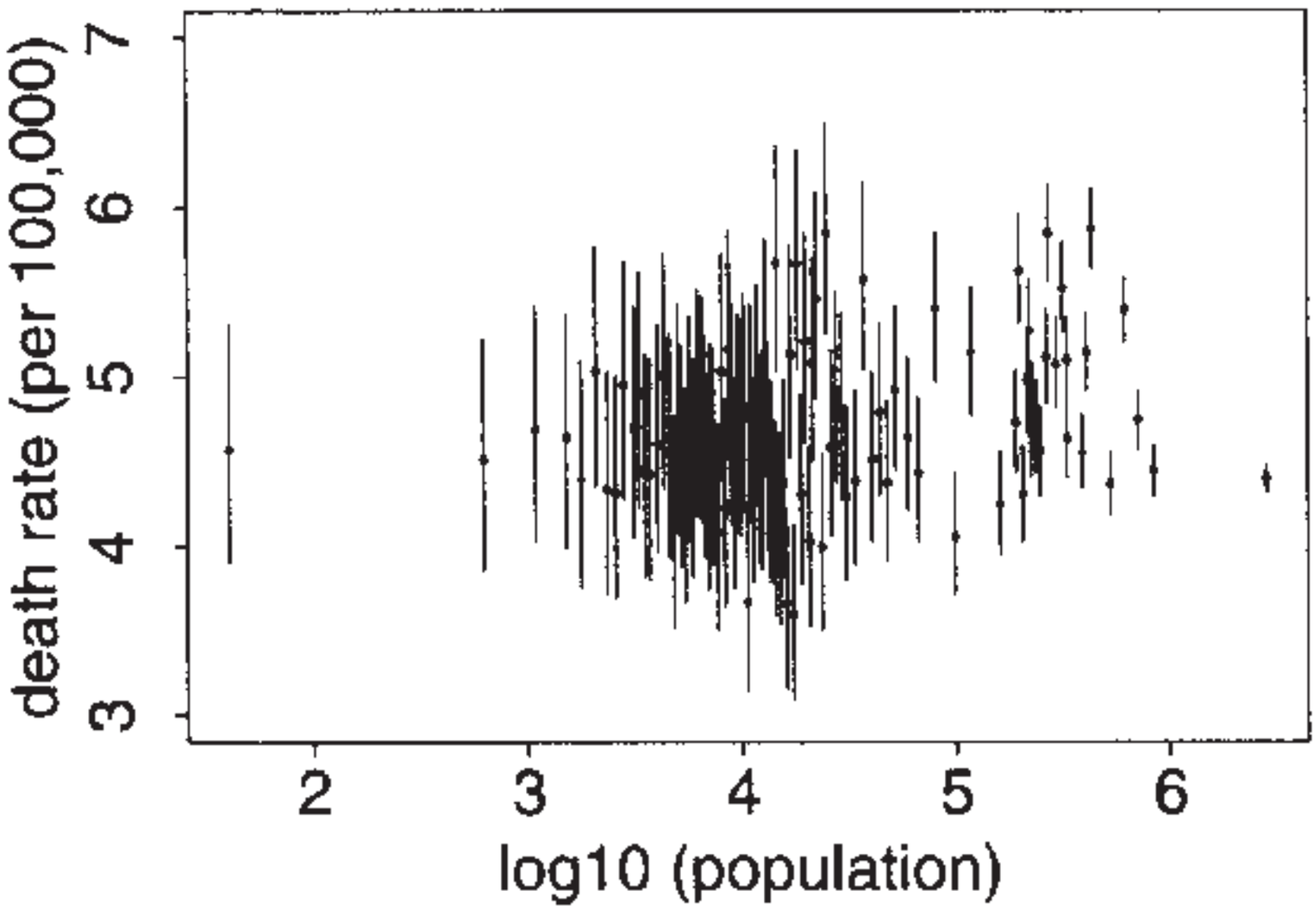
raw data: small counties account for almost all of the high and low death rates

Kidney Cancer Example from Bayesian Data Analysis (Gelman et al)



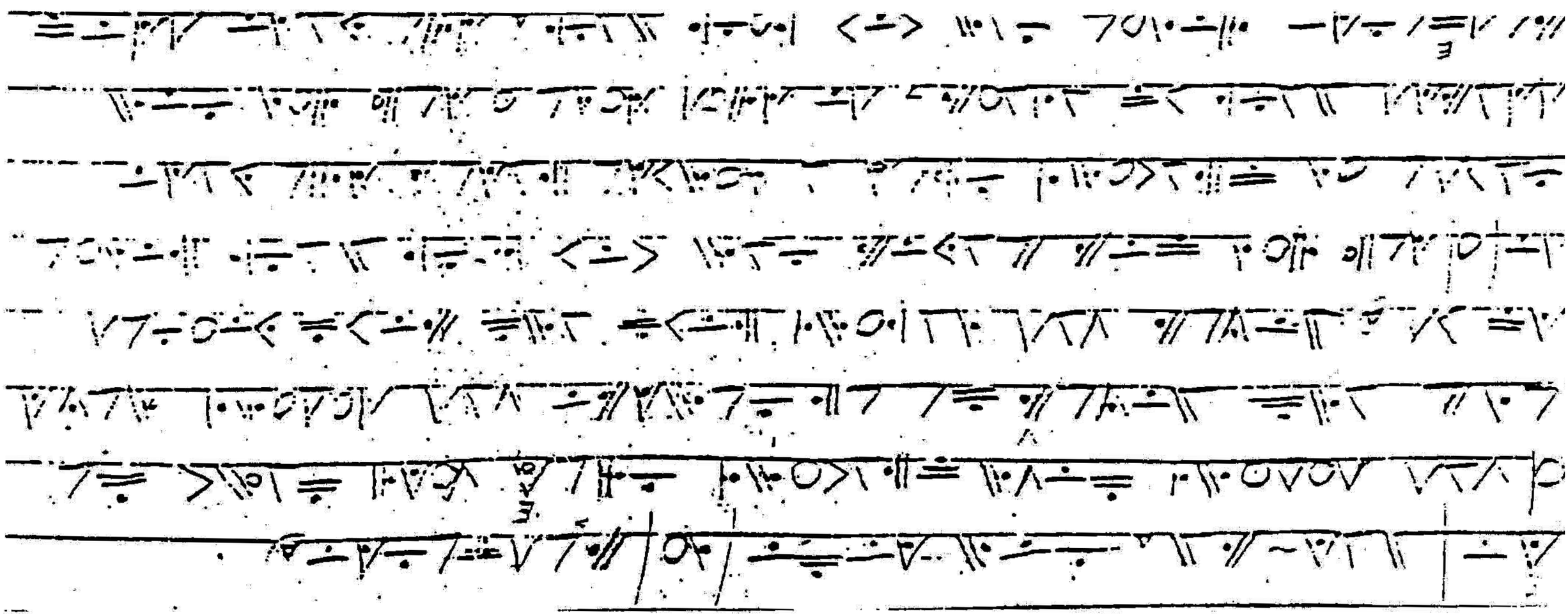
Bayes estimates: automatically accounts for regression toward the mean

Kidney Cancer Example from Bayesian Data Analysis (Gelman et al)



Bayesian posterior medians and 50% probability intervals

Markov Chain Monte Carlo (MCMC):
Diaconis-Coram Example



MCMCryptography

Get a transition matrix M(x,y) for English (the probability of going from letter x to letter y)

Define "plausibility"

$$\mathrm{Pl}(f) = \prod_i M\left(f(s_i), f(s_{i+1})\right),$$

"Try" to swap two random letters in the decoding, based on the ratio of plausibilities.

ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END

100 ER ENOHDLAE OHDLO UOZEOUNORU O UOZEO HD OITO HEOQSET IUROFHE HENO ITORUZAEN
200 ES ELOHRNDE OHRNO UOVEOULOSU O UOVEO HR OITO HEOQAET IUSOPHE HELO ITOSUVDEL
300 ES ELOHANDE OHANO UOVEOULOSU O UOVEO HA OITO HEOQRET IUSOFHE HELO ITOSUVDEL
400 ES ELOHINME OHINO UOVEOULOSU O UOVEO HI OATO HEOQRET AUSOWHE HELO ATOSUVMEL
500 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
600 ES ELOHINME OHINO UODEOULOSU O UODEO HI OATO HEOQRET AUSOWHE HELO ATOSUDMEL
900 ES ELOHANME OHANO UODEOULOSU O UODEO HA OITO HEOQRET IUSOWHE HELO ITOSUDMEL
1000 IS ILOHANMI OHANO RODIORLOSR O RODIO HA OETO HIOQUIT ERSOWHI HILO ETOSRDMIL
1100 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1200 ISTILOHANMITOHANOT ODIO LOS TOT ODIOTHATOEROTHIOQUIRTE SOWHITHILOTEROS DMIL
1300 ISTILOHARMITOHAROT ODIO LOS TOT ODIOTHATOENOTHIOQUINTE SOWHITHILOTENOS DMIL
1400 ISTILOHAMRITOHAMOT OFIO LOS TOT OFIOTHATOENOTHIOQUINTE SOWHITHILOTENOS FRIL
1600 ESTEL HAMRET HAM TO CE OL SOT TO CE THAT IN THE QUENTIOS WHETHEL TIN SOCREL
1700 ESTEL HAMRET HAM TO BE OL SOT TO BE THAT IN THE QUENTIOS WHETHEL TIN SOBREL
1800 ESTER HAMLET HAM TO BE OR SOT TO BE THAT IN THE QUENTIOS WHETHER TIN SOBLER
1900 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER
2000 ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER

to bat-rb. con todo mi respeto. i was sitting down playing chess with danny de emf and boxer de el centro was sitting next to us. boxer was making loud and loud voices so i tell him por favor can you kick back homie cause im playing chess a minute later the vato starts back up again so this time i tell him con respecto homie can you kick back. the vato stop for a minute and he starts up again so i tell him check this out shut the f**k up cause im tired of your voice and if you got a problem with it we can go to celda and handle it. i really felt disrespected thats why i told him. anyways after i tell him that the next thing I know that vato slashes me and leaves. dy the time i figure im hit i try to get away but the c.o. is walking in my direction and he gets me right dy a celda. so i go to the hole. when im in the hole my home boys hit doxer so now "b" is also in the hole. while im in the hole im getting schoold wrong and