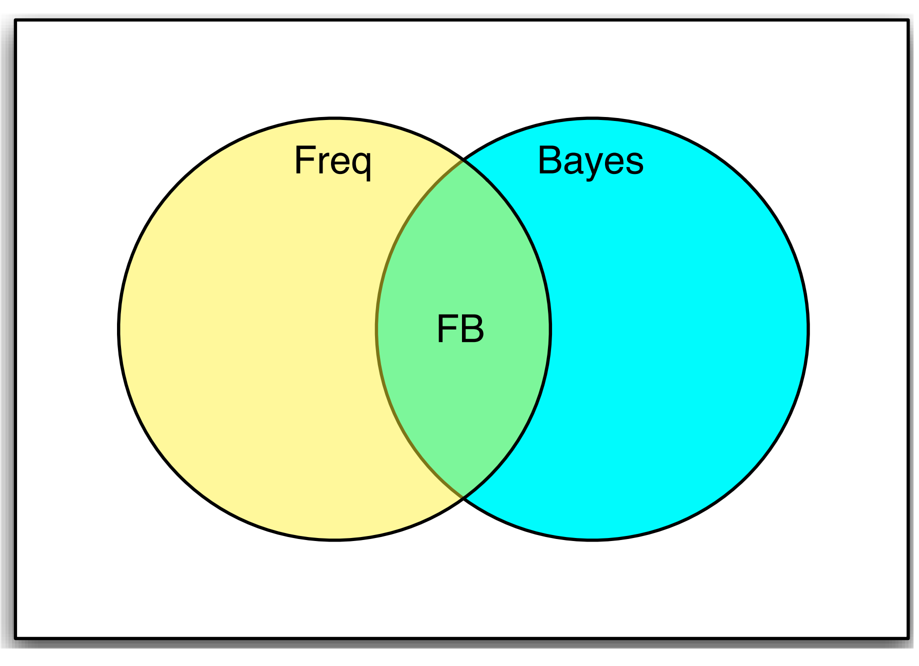
CS109/Stat121/AC209/E-109
Data Science
Bayesian Methods

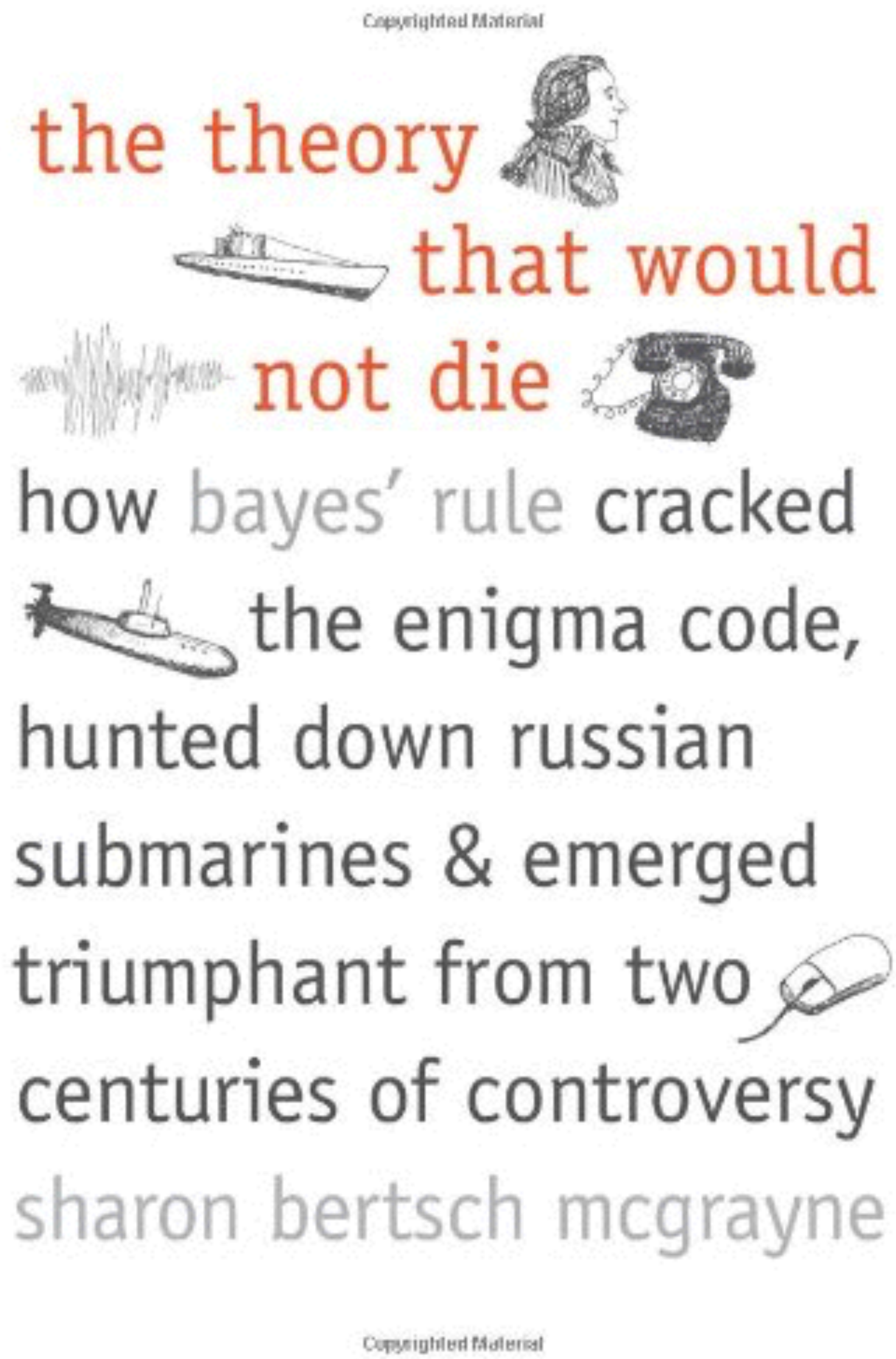
Hanspeter Pfister, Joe Blitzstein, Verena Kaynig



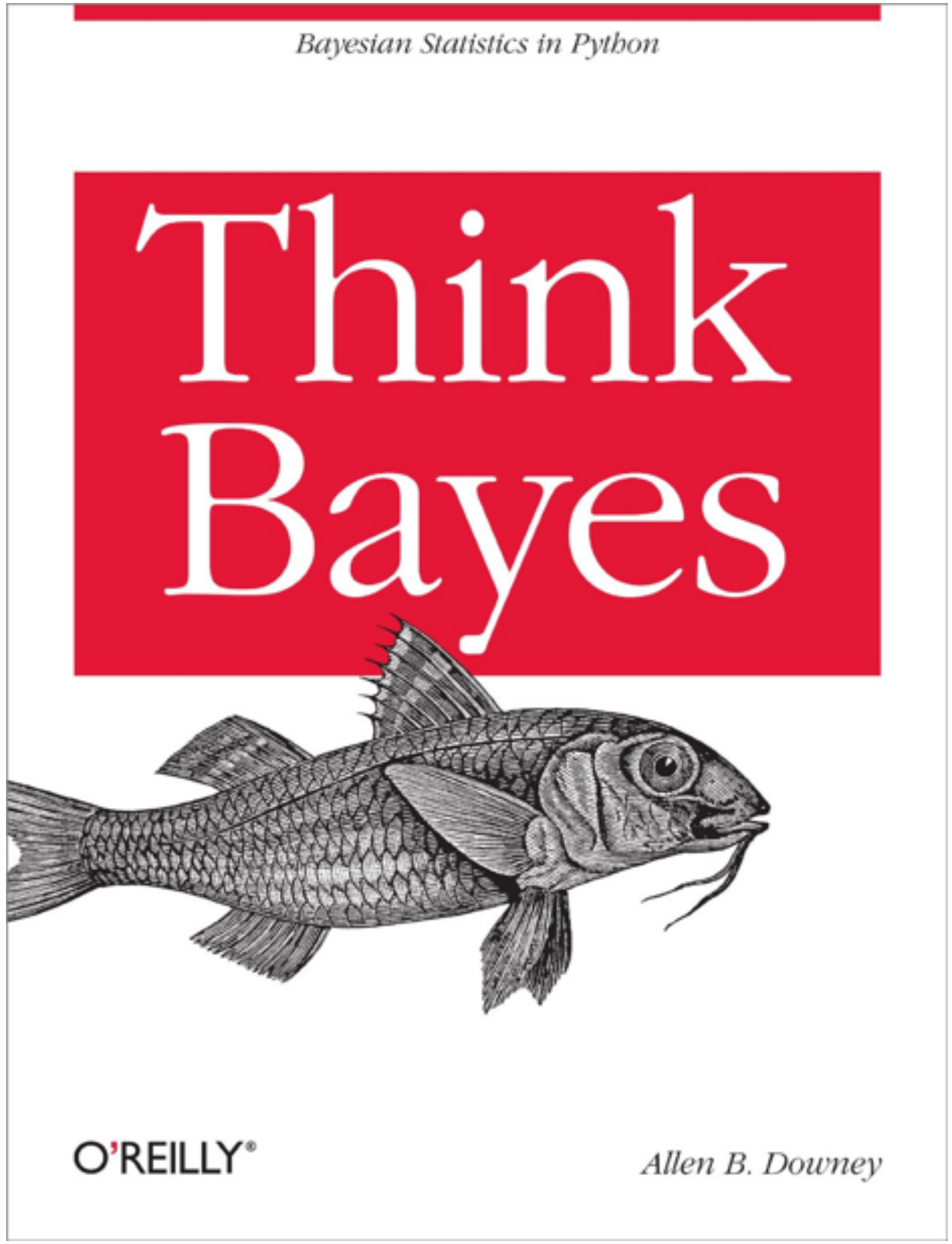
This Week

- Form project team if you haven't already. Try your best to have your team formed by next Tuesday, but in any case it is required to fill in the Google form by Tuesday Nov 3, 11:59 pm: http://goo.gl/forms/CzVRIuCZk6
- HW4 is due Thursday Nov 5, 11:59 pm.

The Theory That Would Not Die



Think Bayes

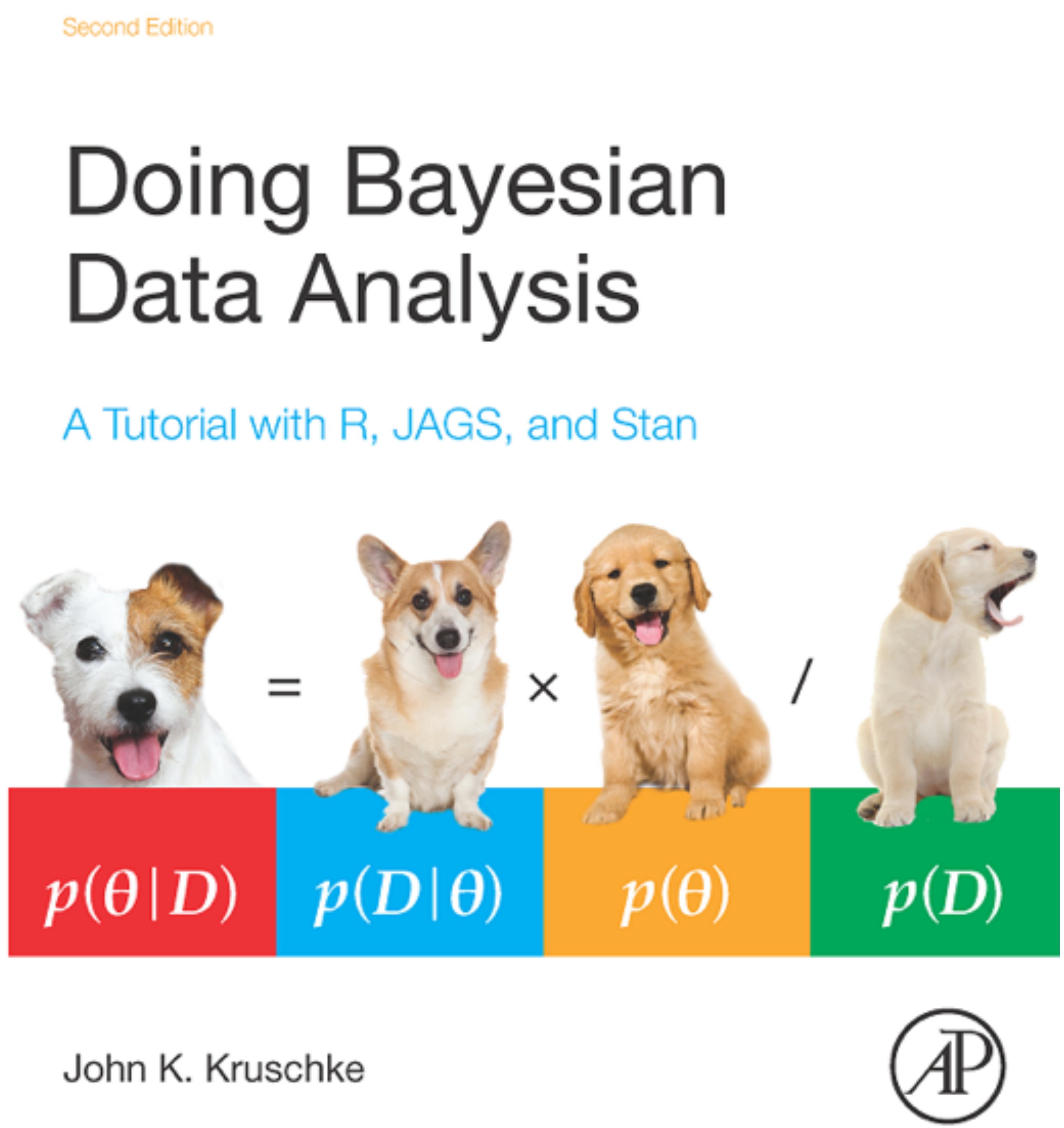


http://greenteapress.com/thinkbayes/

Probabilistic Programming and Bayesian Methods for Hackers

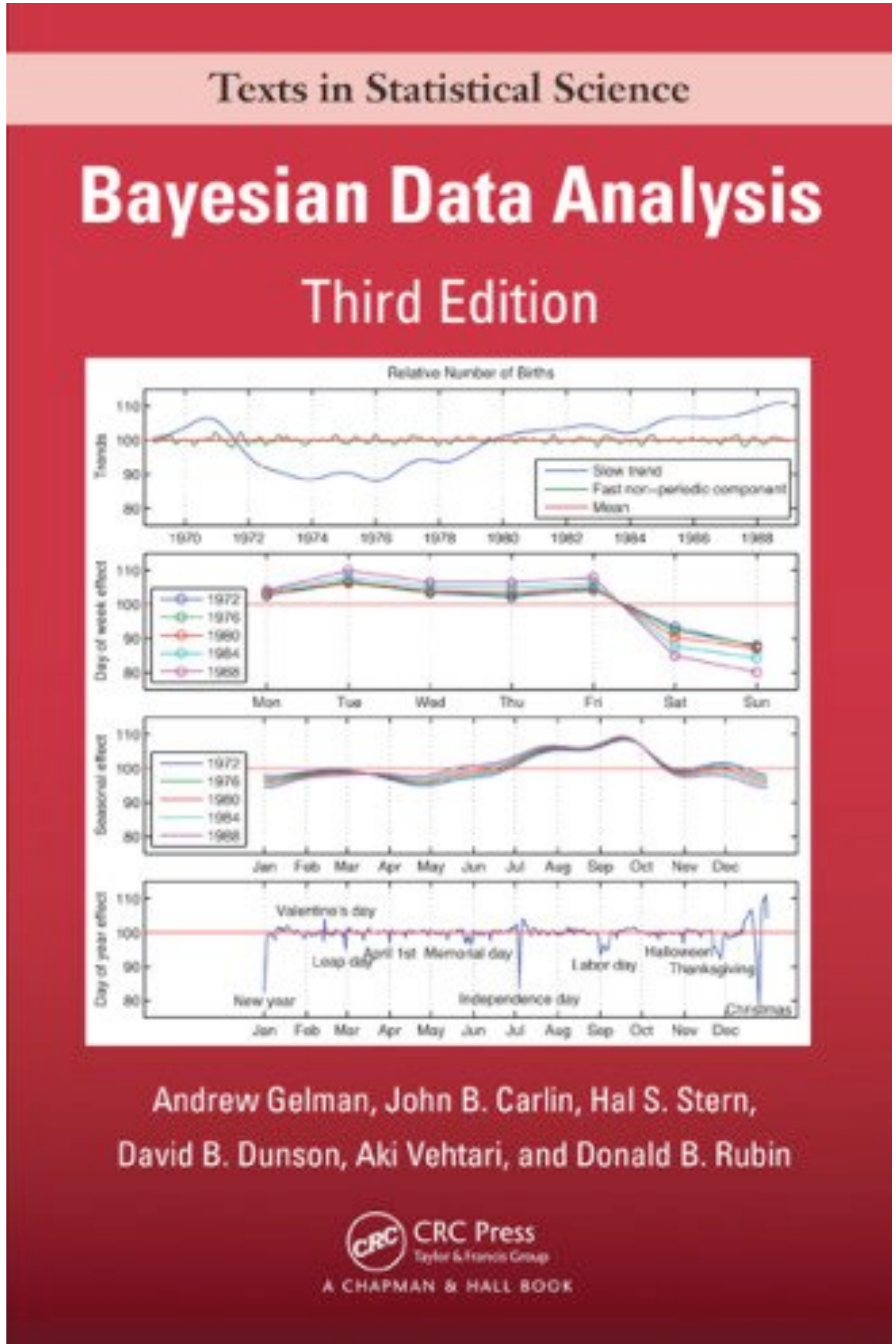
http://nbviewer.ipython.org/urls/raw.github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers/master/Prologue/Prologue.ipynb

Doing Bayesian Data Analysis

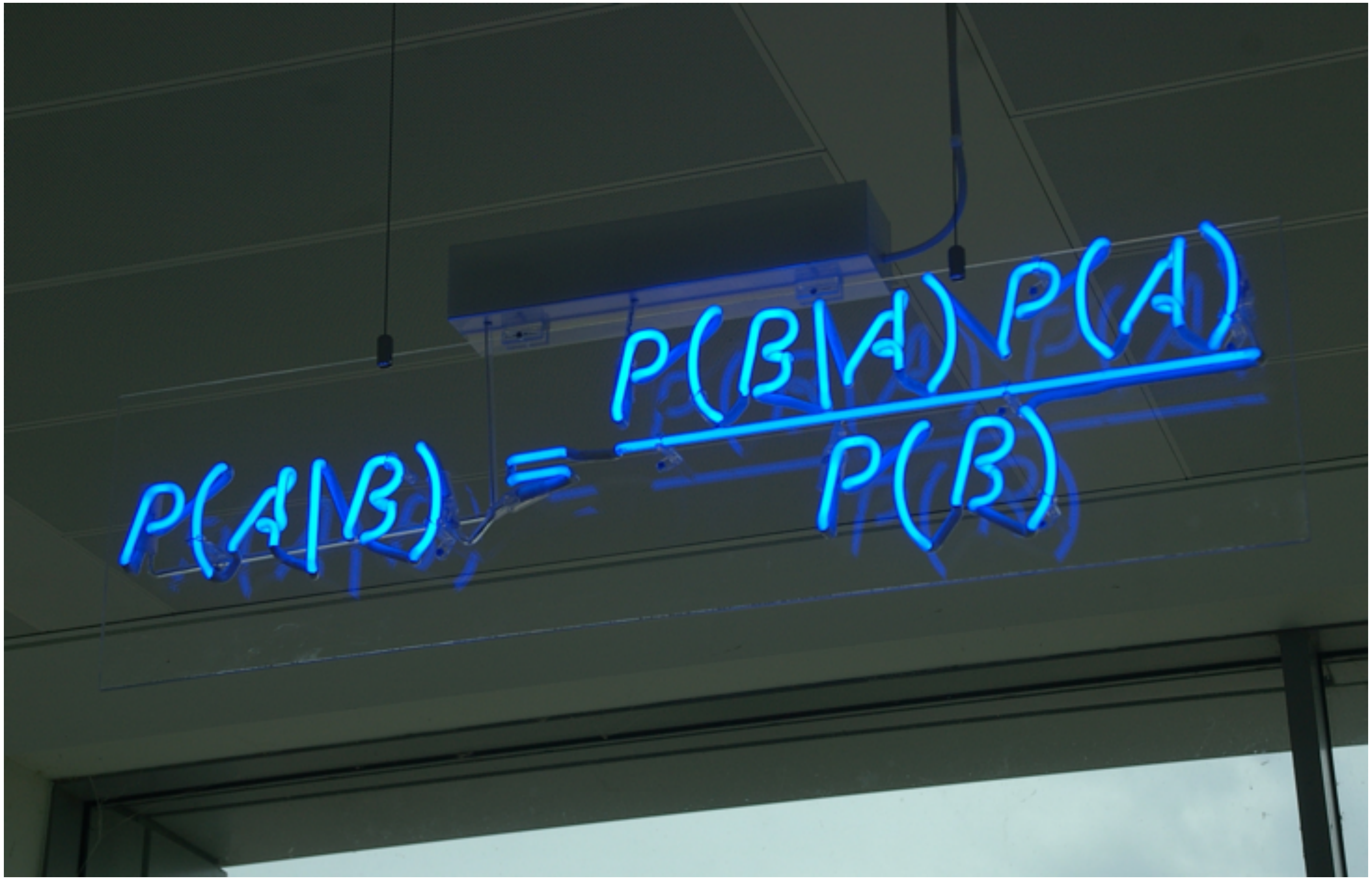


https://sites.google.com/site/doingbayesiandataanalysis/

Bayesian Data Analysis



Bayes' rule



Bayes' rule

prior probability for A

posterior probability for A

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' rule, likelihood version

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Treating the data y as fixed,

$$p(\theta|y) \propto L(\theta)p(\theta)$$

Bayes' rule says the posterior density is proportional to the likelihood function times the prior density.

Discriminative vs. Generative Classifiers

What to model and what not to model?

discriminative: just model $p(y|x)$

generative: give a full probability model
$$p(x,y)=p(x)p(y|x)=p(y)p(x|y)$$

Generative Models

$$P(Y=1|X=x) = \frac{f(x|Y=1)P(Y=1)}{f(x|Y=1)P(Y=1) + f(x|Y=0)P(Y=0)}$$

(by Bayes' Rule)

Then can model the densities f(x|Y=1), f(x|Y=0).

Naive Bayes Spam Filter

Consider 10 words that occur frequently in spam, and let W_j be the event that the jth word appears in the email.

A certain email uses the 1st and 10th words but not the rest. What's the probability that it is spam?

$$P(\text{spam}|W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10}) = \frac{P(W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10}|\text{spam})P(\text{spam})}{P(W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10})}$$

Expand denominator with law of total probability

$$P(W) = P(W|\text{spam})P(\text{spam}) + P(W|\text{not spam})P(\text{not spam})$$

Naive Bayes Spam Filter

$$P(\text{spam}|W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10}) = \frac{P(W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10}|\text{spam})P(\text{spam})}{P(W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10})}$$

Naive Bayes assumption: *conditional independence* given spam, and also *conditional independence* given not spam.

$$P(W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10}|\text{spam}) = P(W_1|\text{spam})P(W_2^c|\text{spam}) \ldots P(W_{10}|\text{spam})$$

$$P(W_1, W_2^c, W_3^c, \ldots, W_9^c, W_{10}|\text{not spam}) = P(W_1|\text{not spam})P(W_2^c|\text{not spam}) \ldots P(W_{10}|\text{not spam})$$

Huge assumption but huge simplification in statistical and computational complexity.

Naive Bayes

Naive conditional independence assumption:

$$f_j(x_1, \ldots, x_d) = f_{j1}(x_1) f_{j2}(x_2) \ldots f_{jd}(x_d)$$

Often unrealistic, but still may be *useful* esp. since it leads to a drastic reduction in the number of parameters to estimate.

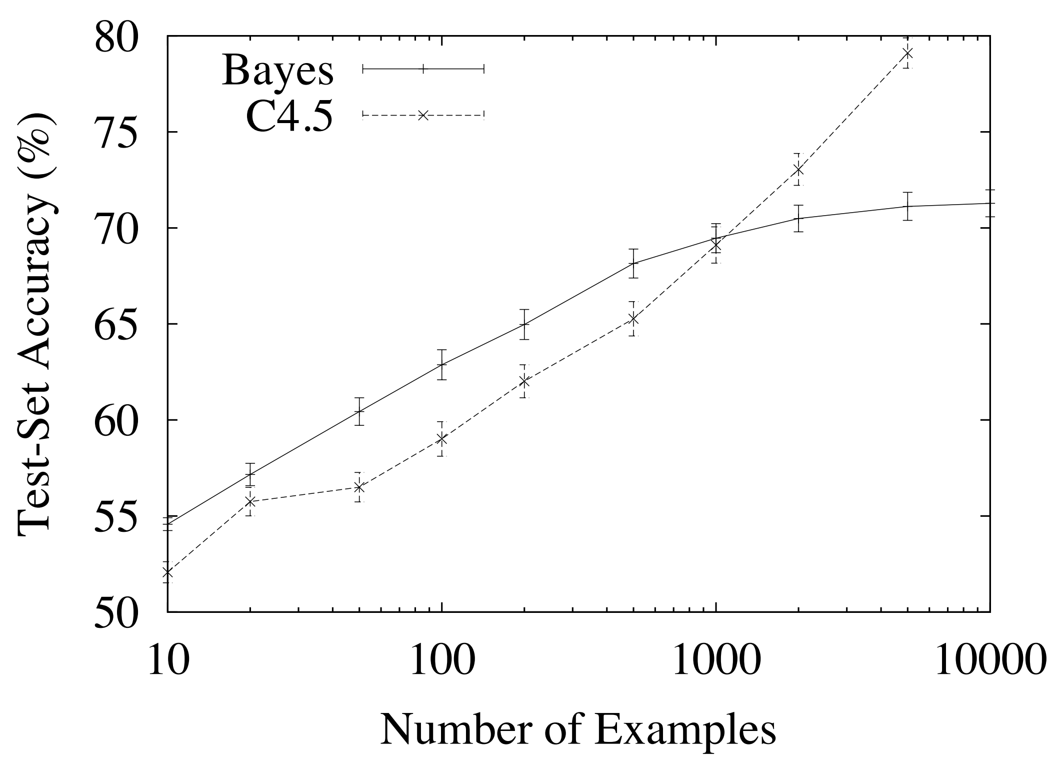


Figure 2: Naive Bayes can outperform a state-of-the-art rule learner (C4.5rules) even when the true classifier is a set of rules.

Domingos, http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

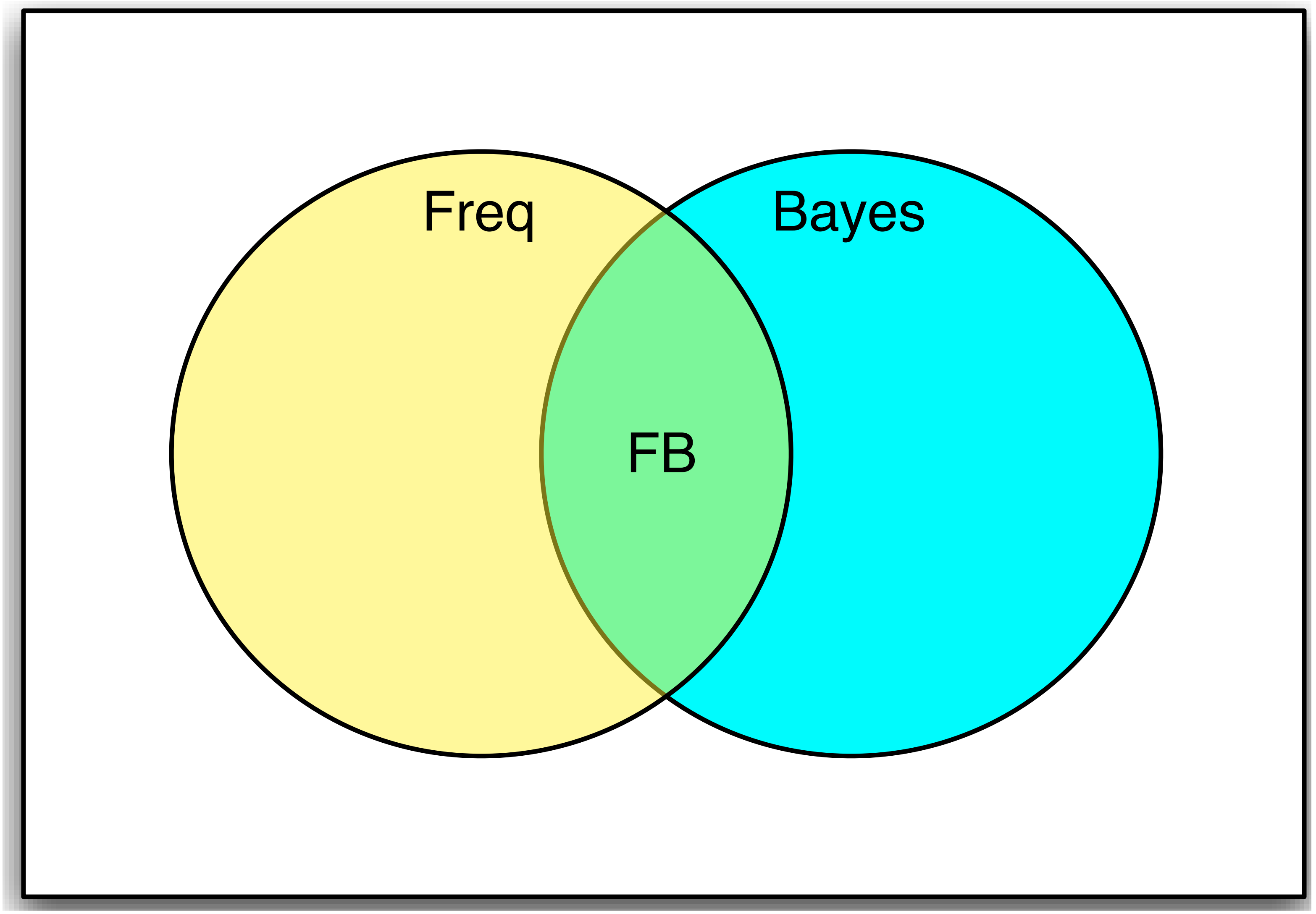
Full Probability Modeling

"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

Bayes-Frequency Reconciliation



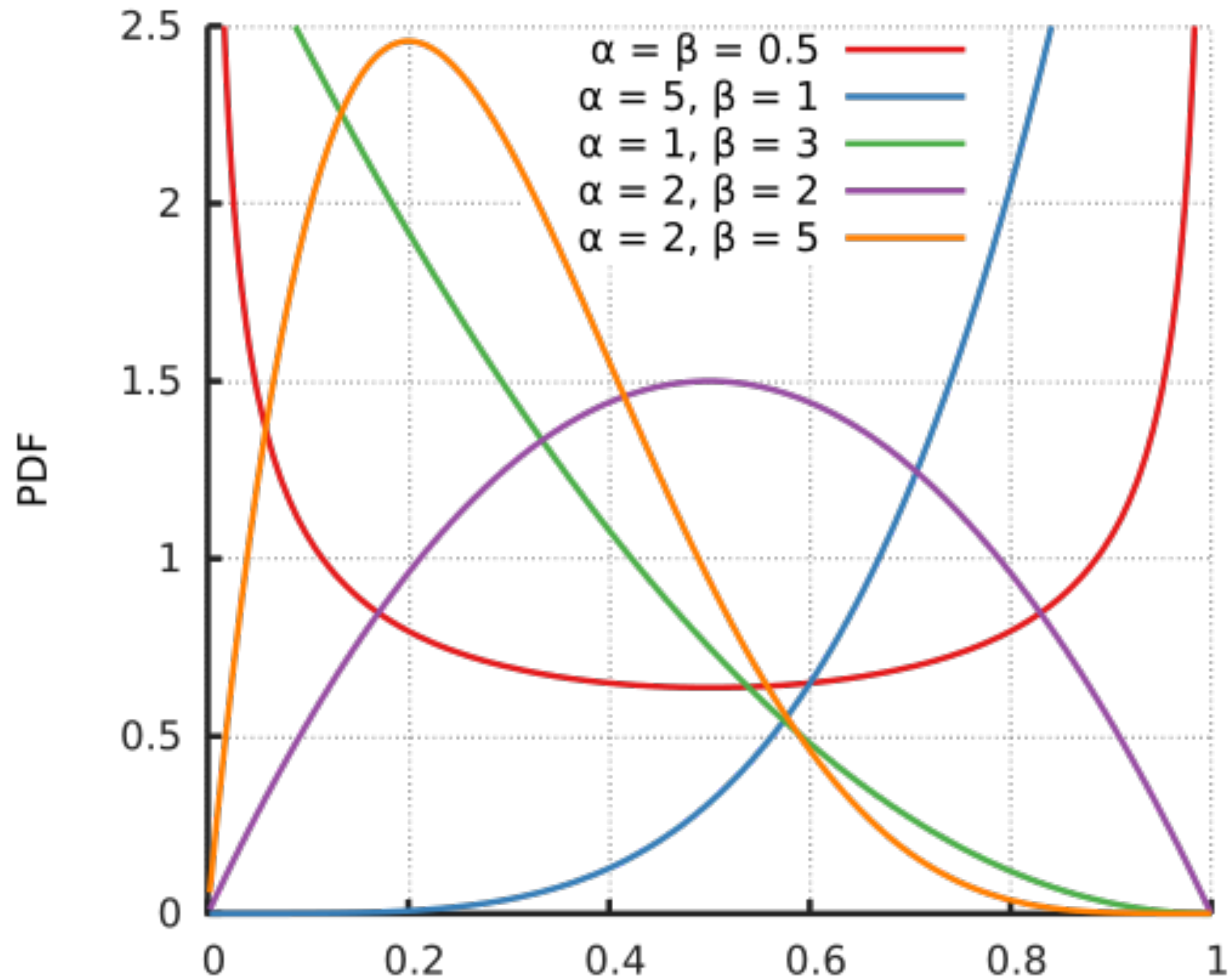
Think like a Bayesian, check like a frequentist.

Conjugate Priors: Beta-Binomial

$$X|p \sim \mathrm{Bin}(n, p)$$

$$p \sim \mathrm{Beta}(a, b)$$

$$f(p) \propto p^{a-1}(1-p)^{b-1}$$



https://en.wikipedia.org/wiki/Beta_distribution

Conjugate Priors: Beta-Binomial

$$X|p \sim \mathrm{Bin}(n, p)$$

$$p \sim \mathrm{Beta}(a, b)$$

Posterior is then $p|X = x \sim \mathrm{Beta}(a + x, b + n - x)$

Conjugate Priors: Normal-Normal

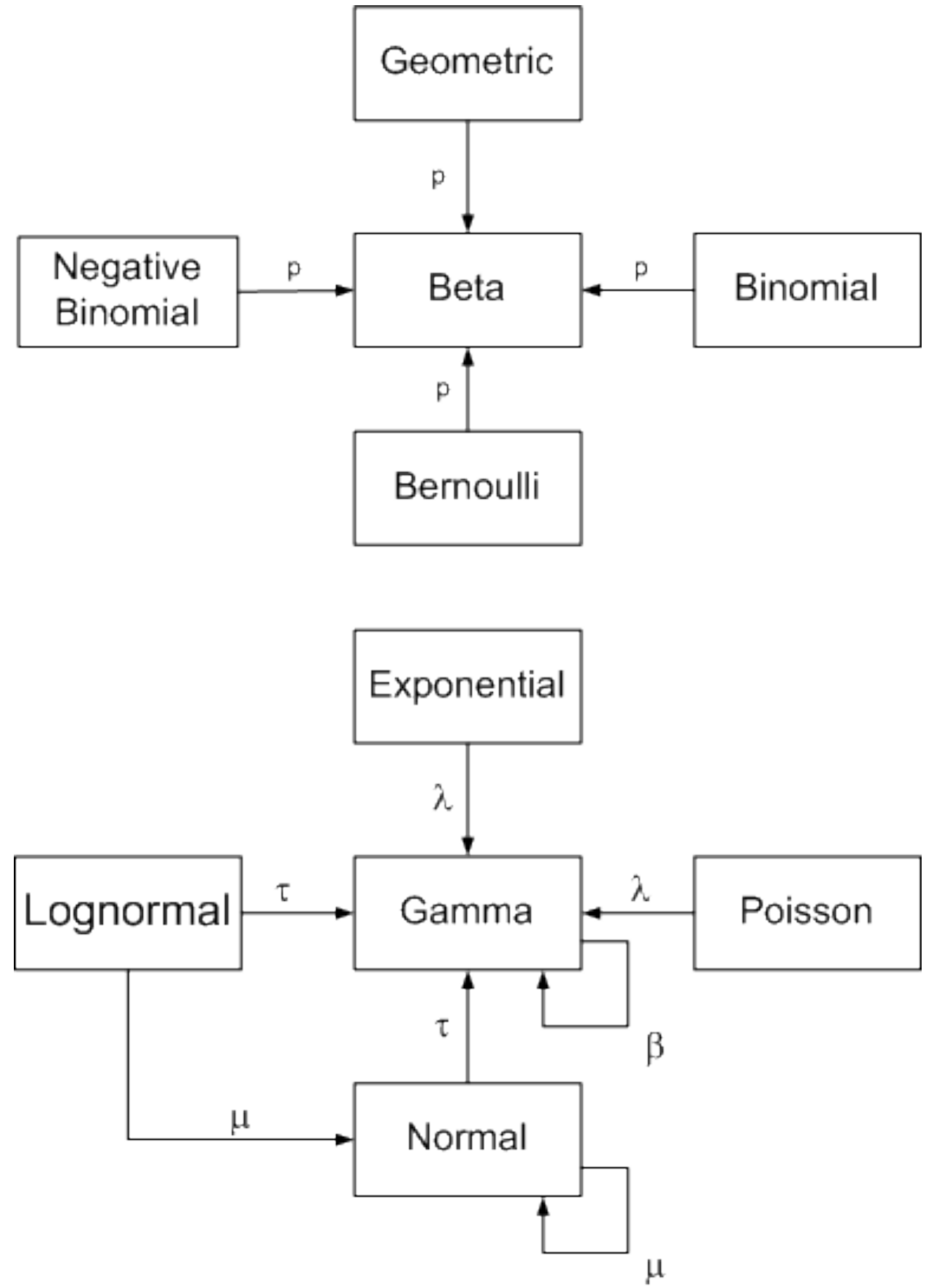
$$y|\mu \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mu \sim \mathcal{N}(\mu_0, \tau^2)$$

$$\text{Then } \mu|y \sim \mathcal{N}\left((1-B)y + B\mu_0, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

where $B = \frac{\sigma^2}{\sigma^2+\tau^2}$ is the *shrinkage factor*

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

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

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

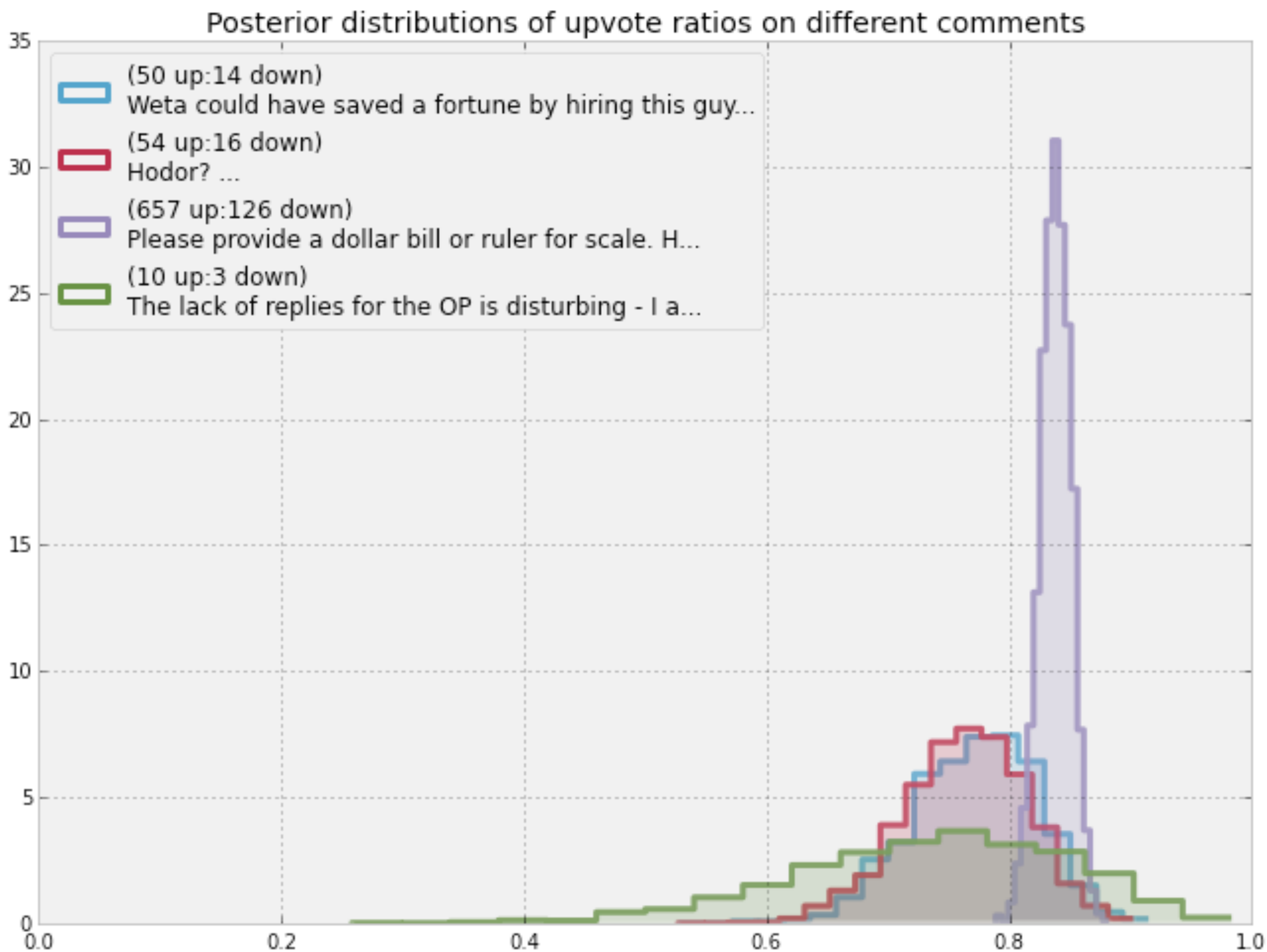
$$\text{posterior: } p|\text{data} \sim \mathrm{Beta}(a + \#\text{upvotes}, b + \#\text{downvotes})$$

Ranking Reddit Comments

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

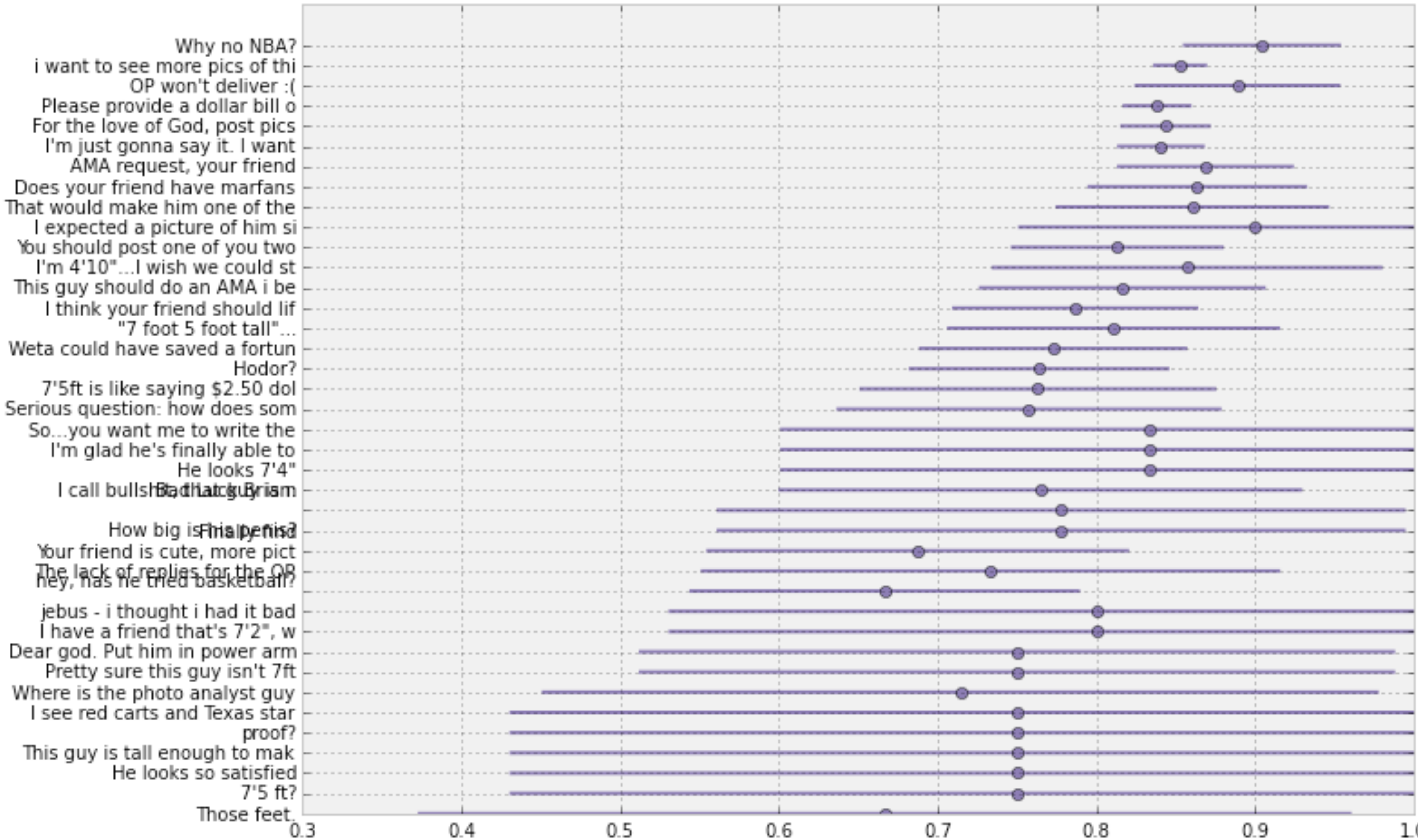
For example, the Agresti-Coull method adds 2 successes and 2 failures.

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