GR5065 HW4 Answer Key

1 Wages

```
library(dplyr)
if (!any(list.files() == "ACS.zip")) {
  ROOT <- "https://www2.census.gov/programs-surveys/acs/data/pums/2021/1-Year/"
  FILE <- "csv_pct.zip" # Connecticut
  download.file(paste0(ROOT, FILE), destfile = "ACS.zip")
  unzip("ACS.zip")
}
ACS <- readr::read_csv(list.files(pattern = "csv$"), show_col_types = FALSE) %>%
  select(RT:WAOB) %>%
  filter(WAGP > 0) %>%
  mutate(censored = WAGP == max(WAGP))
```

1.1 Generative Model

```
ACS <- mutate(ACS,
              SCHL = as.integer(SCHL),
              SEX = SEX - 1L,
              AGEP = AGEP / 10)
X <- model.matrix(log(WAGP) ~ SCHL + SEX + AGEP + I(AGEP ^ 2),
                  data = ACS)
X <- X[ , -1] # drop intercept
X <- sweep(X, MARGIN = 2, STATS = colMeans(X), FUN = `-`)
```

This generative model for potentially top-coded wages has almost the same structure as a survival model with censoring that we discussed in Week11.

$$
\begin{aligned}
\forall n : y_n &\equiv \begin{cases} y_n^* & \text{if } y_n^* < c \\ c & \text{if } y_n^* \geq c \end{cases} \\
\forall n : y_n^* &\equiv e^{\eta_n + \epsilon_n} \\
\forall n : \epsilon_n &\sim \mathcal{N}(0, \sigma) \\
\sigma &\sim \mathcal{E}(r) \\
\forall n : \eta_n &\equiv \gamma + \sum_{k=1}^{K} \beta_k (x_{nk} - \overline{x}_k) \\
\gamma &\sim \mathcal{N}(m_0, s_0) \\
\forall k : \beta_k &\sim \mathcal{N}(m_k, s_k)
\end{aligned}
$$

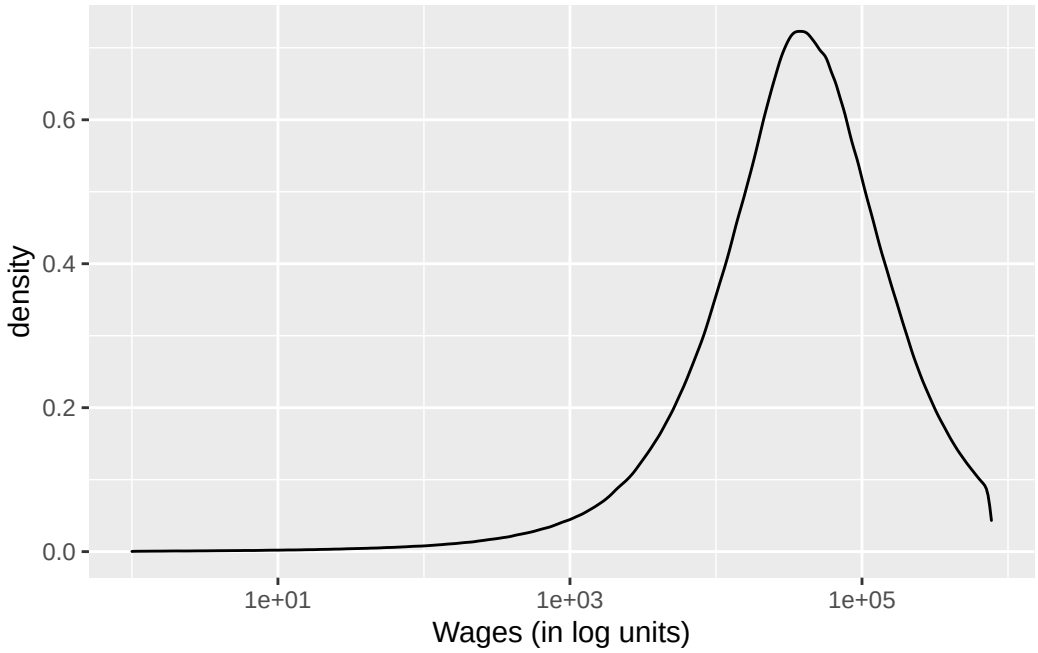
To draw outcomes, we can do

```
m_0 <- log(40000) # about 10.6
s_0 <- 0.5
m <- c( 0.1, # SCHL
       -0.2, # SEX (female)
        0.1, # AGEP
        0 )  # AGEP^2
s <- c(rep(0.2, 3), 0.05)
r <- 1

c <- max(ACS$WAGP)
R <- 1000

draws <- tibble(
  gamma  = rnorm(R, m_0, s_0),
  beta_1 = rnorm(R, m[1], s[1]),
  beta_2 = rnorm(R, m[2], s[2]),
  beta_3 = rnorm(R, m[3], s[3]),
  beta_4 = rnorm(R, m[4], s[4]),
  sigma  = rexp(R, r)
) %>%
  rowwise %>%
  summarize(eta = gamma + beta_1 * X[ , 1] + beta_2 * X[ , 2] +
              beta_3 * X[ , 3] + beta_4 * X[ , 4],
            epsilon = rnorm(nrow(X), 0, sigma),
            y_star = exp(eta + epsilon),
            y = pmin(y_star, c)) %>%
  ungroup
```

```
library(ggplot2)
ggplot(draws) +
  geom_density(aes(y)) +
  scale_x_log10(limits = c(1, c - 1)) +
  labs(x = "Wages (in log units)")
```



This looks reasonable, although it took several tries to avoid having it be too spread out and right-skewed. It is still somewhat right-skewed but the median is a little over $40,000. A small percentage of values are top-coded (although the percentage of top-coded values in the actual data is even smaller). It certainly does not look normal, but the model does not claim that the outcome is normal irrespective of the parameters (and predictors). The generative process clearly assumes that the errors (in log-units) are normal with mean zero and standard deviation σ. When we marginalize over σ, the errors have heavier tails, which contributes to the skew in predicted wages (in dollars).

1.2 Posterior Distribution

```
library(brms)
options(mc.cores = parallel::detectCores())
get_prior(log(WAGP) | cens(censored) ~
            SCHL + SEX + AGEP + I(AGEP ^ 2),
```

```
            data = ACS,
            family = gaussian)
```

```
                  prior     class    coef group resp dpar nlpar bound
                 (flat)         b
                 (flat)         b    AGEP
                 (flat)         b IAGEPE2
                 (flat)         b    SCHL
                 (flat)         b     SEX
student_t(3, 10.8, 2.5) Intercept
   student_t(3, 0, 2.5)     sigma
      source
     default
(vectorized)
(vectorized)
(vectorized)
(vectorized)
     default
     default
```

```
my_prior <-
  prior(normal(10.6, 0.5), class = "Intercept") +
  prior(normal(0.1, 0.2),  class = "b", coef = "AGEP") +
  prior(normal(0.0, 0.05), class = "b", coef = "IAGEPE2") +
  prior(normal(0.1, 0.2),  class = "b", coef = "SCHL") +
  prior(normal(-0.2, 0.2), class = "b", coef = "SEX") +
  prior(exponential(1), class = "sigma")
```

```
post <- brm(log(WAGP) | cens(censored) ~
              SCHL + SEX + AGEP + I(AGEP ^ 2),
            data = ACS,
            family = gaussian,
            prior = my_prior)
```

Under our posterior distribution, females are expected to have lower wages than otherwise identical males, and we are essentially certain of this.

```
hypothesis(post, "SEX < 0")
```

```
Hypothesis Tests for class b:
  Hypothesis Estimate Est.Error CI.Lower CI.Upper Evid.Ratio Post.Prob Star
1  (SEX) < 0    -0.39      0.02    -0.42    -0.37        Inf         1    *
---
'CI': 90%-CI for one-sided and 95%-CI for two-sided hypotheses.
'*': For one-sided hypotheses, the posterior probability exceeds 95%;
for two-sided hypotheses, the value tested against lies outside the 95%-CI.
Posterior probabilities of point hypotheses assume equal prior probabilities.
```

1.3 Pareto k Values

```
loo_post <- loo(post, save_psis = TRUE)
loo_post
```

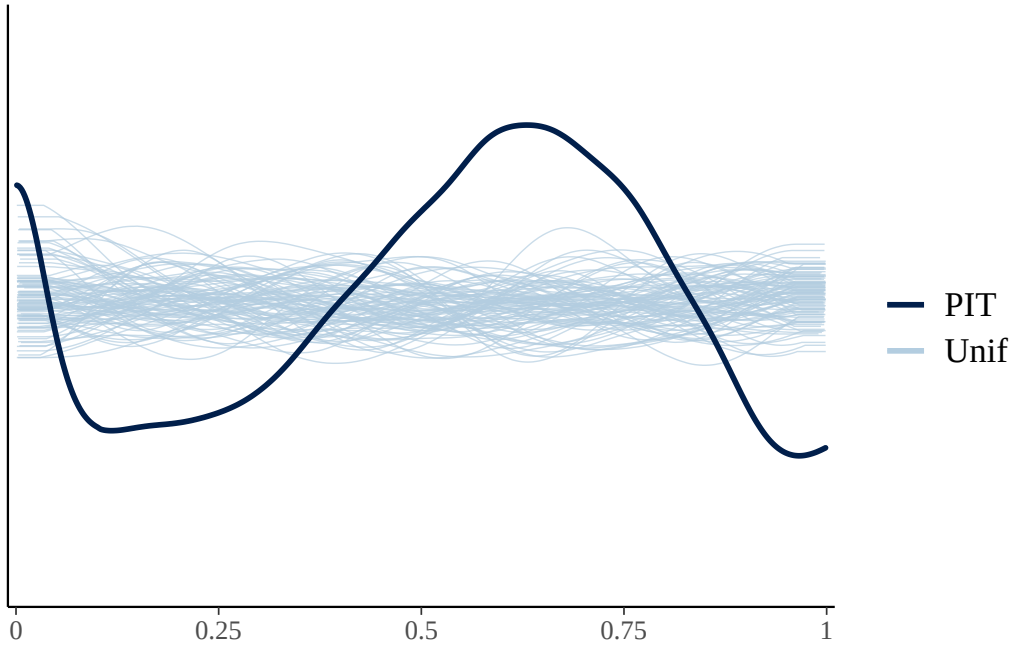
```
Computed from 4000 by 18483 log-likelihood matrix

         Estimate    SE
elpd_loo -28813.2 160.6
p_loo         9.2   0.3
looic     57626.4 321.2
------
Monte Carlo SE of elpd_loo is 0.1.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.
```

1.4 Posterior Predictive Checks

```
bayesplot::ppc_loo_pit_overlay(model.frame(post)[[1]],
                               yrep = posterior_predict(post),
                               lw = weights(loo_post$psis_object))
```

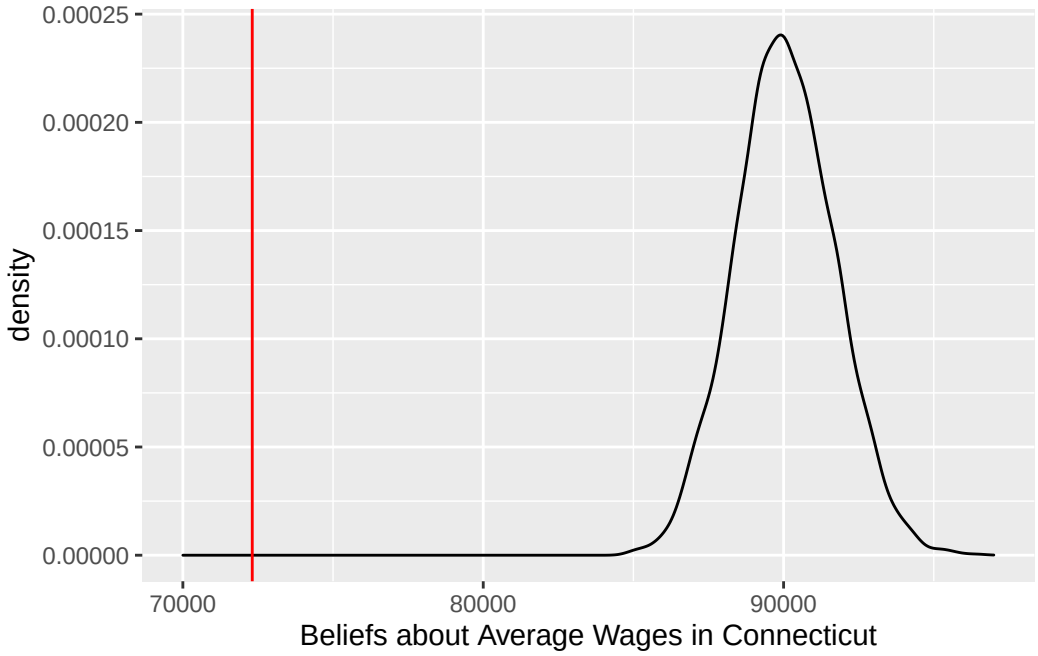


This fit is not great, as evidenced by the fact that the CDF of the Pareto smoothing weights is not distributed uniform. However, a lot of things look worse when you look at them rigorously. There are many Frequentist models of log-wages that are similar to the one above, and they only fit better in the sense that they only evaluate the fit at the parameter values that optimally fit the data. The LOOCV concepts utilized in the above posterior predictive check are based on the (estimated) Expected Log Predictive Density (ELPD) for N future observations *irrespective of the parameters*.

1.5 Posterior Predictive Distribution

```
avg_wages <- rowMeans(exp(posterior_predict(post)))
tibble(avg_wages = avg_wages) %>%
  ggplot() +
  geom_density(aes(avg_wages)) +
  geom_vline(aes(xintercept = mean(ACS$WAGP)), color = "red") +
  xlim(70000, 97000) +
  labs(x = "Beliefs about Average Wages in Connecticut")
```

```
Warning: Removed 1 rows containing non-finite values (stat_density).
```



Average predicted wages including those people whose observed wages are top-coded in the `ACS` is much larger (and more uncertain) that whatever estimate you might obtain from the censored data. This selection mechanism not only affects the average, but also the variance and any measure of wage inequality.

2 Culture

This question pertains to Leo Breiman's (2001a) article, along with the short responses by Cox, Efron, Hoadley, Parzen, and the rejoinder to those responses by Breiman (2001b).

Breiman (2001a) distinguishes between the "data modelling culture" and "algorithmic modeling culture". The data modeling culture consisted primarily of applied statisticians who start by assuming a stochastic model for the data, such as the errors in the outcome are normally distributed with mean zero and (perhaps unknown) standard deviation σ. The algorithmic modeling culture consisted primarily of engineers, computer scientists, researchers in industry, and other people without much connection to statistics departments at universities. Its goal is to develop algorithms to estimate the function that generated the data without assuming, for example, that it entails linearity or normally distributed errors with a constant σ.

Breiman (2001a) would consider Bayesians to be a subset of the data modelling culture because it asserts a generative model that typically makes functional form and parametric assumptions.

In fairness, Bayesian analysis was still not that prevalent in applied statistics as of 2000, so Breiman could perhaps be excused for not mentioning it beyond saying in passing that

> There is an old saying "If all a man has is a hammer, then every problem looks like a nail." The trouble for statisticians is that recently some of the problems have stopped looking like nails. I conjecture that the result of hitting this wall is that more complicated data models are appearing in current published applications. Bayesian methods combined with Markov Chain Monte Carlo are cropping up all over. This may signify that as data becomes more complex, the data models become more cumbersome and are losing the advantage of presenting a simple and clear picture of nature's mechanism.(204)

Bayesians would object to being lumped in with the Frequentists that Breiman (2001a) is implicitly referring to when he derides the data modelling culture. Efron (2001) notes that "A third front seems to have been opened in the long-running frequentist-Bayesian wars by the advocates of algorithmic prediction, who don't really believe in any inferential school." For example, Breiman (2001a) says

> Upon my return [to academia from consulting], I started reading the *Annals of Statistics*, the flagship journal of theoretical statistics, and was bemused. Everyarticle started with
>
> Assume that the data are generated bythe following model:
>
> followed by mathematics exploring inference, hypothesis testing and asymptotics.

However, one of the ways that Breiman (2001a) distinguishes the algorithmic modeling culture is by its emphasis on accuracy when predicting data in the testing set. Prediction comes very naturally in the Bayesian framework, as does evaluating a model based on its Expected Log Predictive Density (ELPD).

Different quantitative methods can be distinguished by how they address the four or five sources of uncertainty that we talked about on the first day of GR5065, which are:

1. Uncertainty about parameters in models
2. Uncertainty about which model is best
3. Uncertainty about what to do with model output
4. Uncertainty about whether the software works as intended
5. Uncertainty about whether the model holds with other data

Breiman and other supervised learners fixate on uncertainty about which model is best, particularly with regard to how well it predicts future data (as proxied by past data from the testing set). Bayesians use probability to describe all sources of uncertainty, but are unique in that they consider uncertainty about parameters in models to be fundamental and represent it with R draws from the prior or posterior distribution. Frequentist probability is conditional on the parameters, so there is no uncertainty about them. If Frequentists decide between

models, they start by associating one model with the null hypothesis and then either reject or fail to reject that null hypothesis. Thus, it is a yes-no decision rather than having anything to do with uncertainty. Frequentist software works fairly reliably if the data are generated according to its strong assumptions, and it is fairly well-established what should be done with the model output (i.e. testing null hypotheses) although those practices have been subject to long-standing criticisms.

In fact, there are several reasons why the Bayesian approach to prediction is preferable to the supervised learning approaches that Breiman (2001a) mentions. First, the ELPD is an expectation over future data and the parameters; in other words, it is a product of the assumed generative model but is irrespective of the parameters. Supervised learning obtains point estimates of the parameters using the training data only, and then predicts the outcomes in the testing set which still consists of observations from the past. Thus, the measures of predictive accuracy in supervised learning are only evaluated the optimal values of the parameters, which ignores the first source of uncertainty and is subject to idiosyncrasies about the training data. In addition, the conclusions as to which model predicts best are (or at least should be) limited to the particular testing set used, which may also have idiosyncrasies even if it was randomly sampled from the original data. Second, the Pareto Smoothed Importance Sampling (PSIS) weights used to estimate the ELPD can serve as a diagnostic for when the posterior distribution is especially sensitive to particular observations and can be used for posterior predictive checks to the data that were conditioned on.

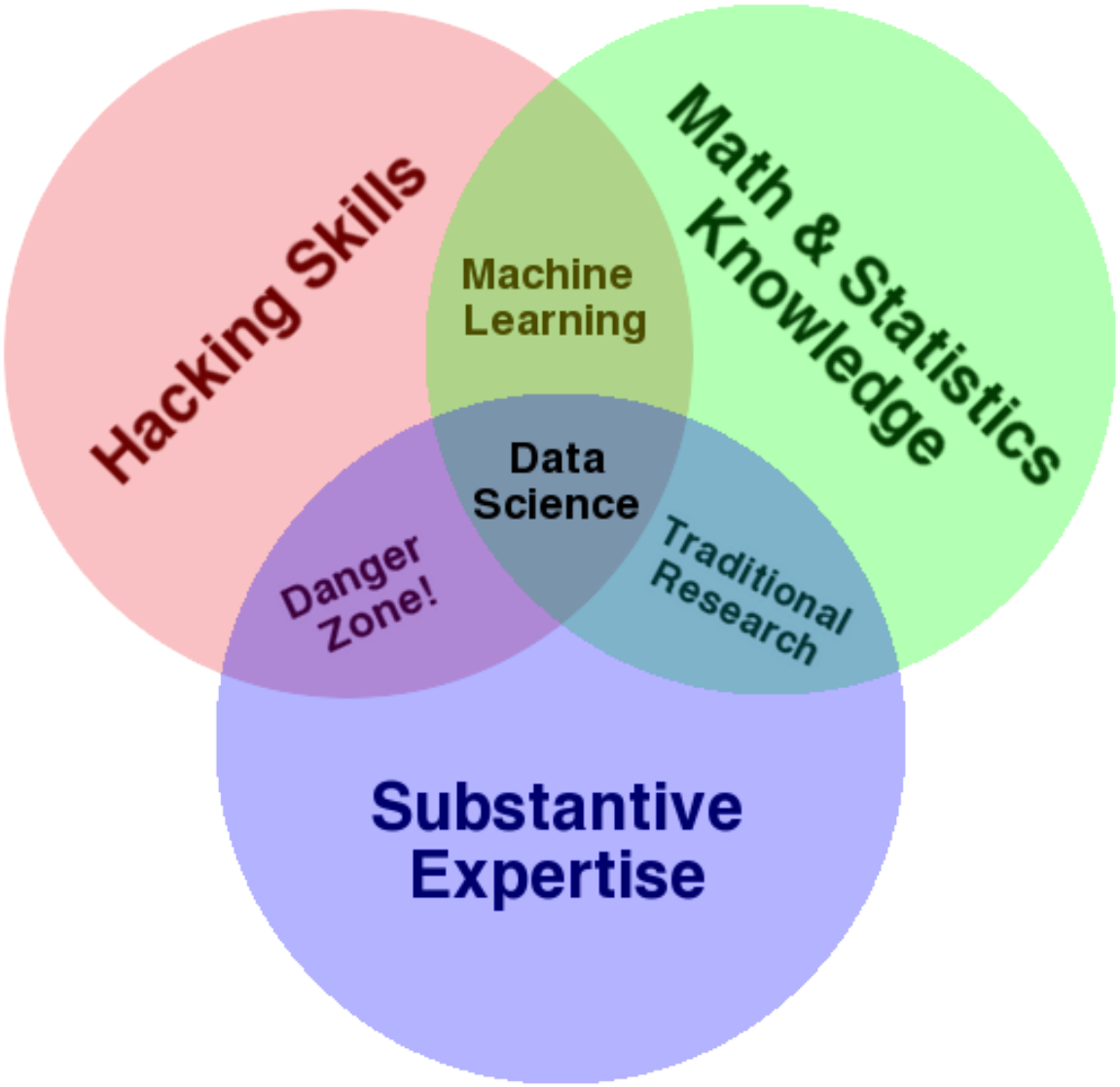
Breiman (2001a) emphasizes that there may be multiple models that yield very similar predictions, or by extension, many values of the parameters for a single model that yield very similar predictions. Thus, if your only goal is predictive accuracy, it does not matter much which parameter values are used because any that yield the same predictions are admissible. Cox (2001) takes issue with the fact that

> Professor Breiman takes a rather defeatist attitude toward attempts to formulate underlying processes; is this not to reject the base of much scientific progress? The interesting illustrations given by Beveridge (1952), where hypothesized processes in various biological contexts led to important progress, even though the hypotheses turned out in the end to be quite false, illustrate the subtlety of the matter. Especially in the social sciences, representations of underlying process have to be viewed with particular caution, but this does not make them fruitless. (218)

Parzen (2001a) also gets at this point by distinguishing between "management" (with a goal of profit) and "science" (with a goal, perhaps never achieved, of seeking truth). Breiman (2001a, 2001b) says relatively little as to how the algorithmic modeling culture would contribute toward any aspect of science besides better predictions of future data. But I think Breiman would respond by saying the scientific conclusions drawn using by the data modeling culture are dubious because we do not know nature's generative model.

Breiman's views have become much more popular and much more mainstream in the last twenty years, even within statistics departments. Conway's blog post attempts to characterize

where the emerging field of data science fits in with the use of a Venn diagram:



Breiman (2001a) explicitly advocates for machine learning and against what Conway would call traditional research.

Data science includes many facets, including important ones that take place before (or as) the data are being collected and cleaned. But when it comes to the step of modeling the data, data science today tends toward the frontier with machine learning in Conway's Venn diagram. Presumably, Cox (2001), Parzen (2001), and others would question how much (algorithmic) modeling that is taught in data science programs and practiced at tech companies contributes to science.

Bayesians would have the same questions. If we were to equate or specialize Math & Statistics Knowledge to "probability", Substantive Expertise with the ability to specify "priors" on the parameters (as well as the other parts of a generative model), and Hacking Skills with "use of MCMC algorithms", then the intersection would characterize Bayesian analysis. Indeed, many of the techniques that were associated with machine learning in 2001, such as splines and neural networks, can be performed using Bayesian analysis today. Those advances in Bayesian technology have essentially been ignored in the Data Science Culture despite MCMC preceding the emergence of data science programs.

There are some legitimate concerns with how practical or necessary Bayesian inference is when conditioning on terrabytes of data, but most datasets in the social sciences are orders of magnitude smaller than that. Although MCMC makes Bayesian inference possible — in the sense that we do not have to work out non-elementary integrals symbolically — it does not render Bayesian inference easy for non-trivial models. Much of data science is about lowering barriers to using quantitative methods but Conway presumably would agree that being proficient in all aspects of his Venn diagram requires a lot of training, a lot of money to afford to take the necessary classes, etc. The perceived need to make data analysis easy and to make data analysts interchangeable has pushed Conway's Venn diagram northeast so that it is no longer as symmetric as Conway intended and leaves little room for Bayesian inference or decision theory.