A Bayesian Workflow for Data Analysis

Paul Bürkner

The Bayes Theorem

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

Rethinking the Bayes Theorem

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta) = p(y, \theta)$$

Why use Bayesian Statistics?

Advantages and Disadvantages of Bayesian Statistics

Advantages:

- Natural approach to expressing uncertainty
- Ability to incorporate prior information
- Increased modeling flexibility
- Full posterior distribution of parameters
- Natural propagation of uncertainty

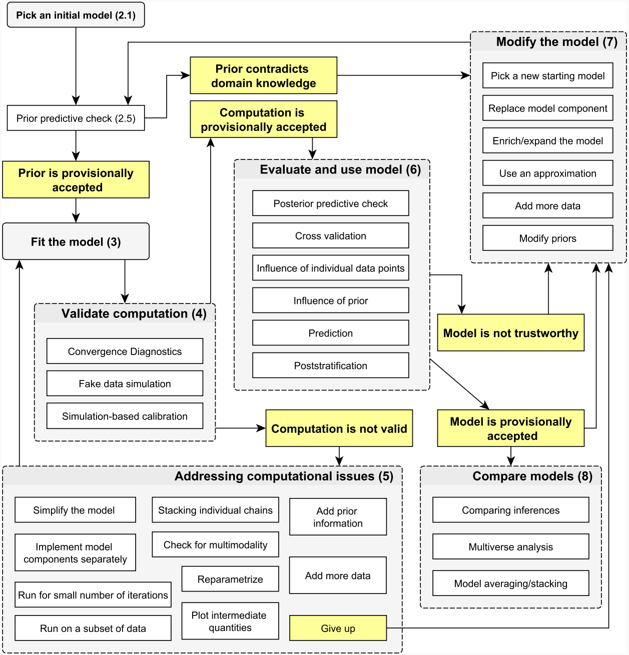
Disadvantages:

- Slow Speed of model estimation

Topic for Today

(Aspects of) a Bayesian workflow for data analysis

Gelman A., Vehtari A., Simpson D., Margossian, C., Carpenter, B. and Yao, Y., Kennedy, L., Gabry, J., **Bürkner P. C.**, & Modrák M. (2020). Bayesian Workflow. *https://arxiv.org/abs/2011.01808*


Pick an initial model (2.1)
Prior predictive check (2.5)
Prior contradicts domain knowledge
Computation is provisionally accepted
Prior is provisionally accepted
Fit the model (3)
Modify the model (7)
Pick a new starting model
Replace model component
Enrich/expand the model
Use an approximation
Add more data
Modify priors
Evaluate and use model (6)
Posterior predictive check
Cross validation
Influence of individual data points
Influence of prior
Prediction
Poststratification
Model is not trustworthy
Validate computation (4)
Convergence Diagnostics
Fake data simulation
Simulation-based calibration
Computation is not valid
Model is provisionally accepted
Addressing computational issues (5)
Simplify the model
Stacking individual chains
Add prior information
Implement model components separately
Check for multimodality
Reparametrize
Add more data
Run for small number of iterations
Run on a subset of data
Plot intermediate quantities
Give up
Compare models (8)
Comparing inferences
Multiverse analysis
Model averaging/stacking

The Probabilistic Programming Language Stan

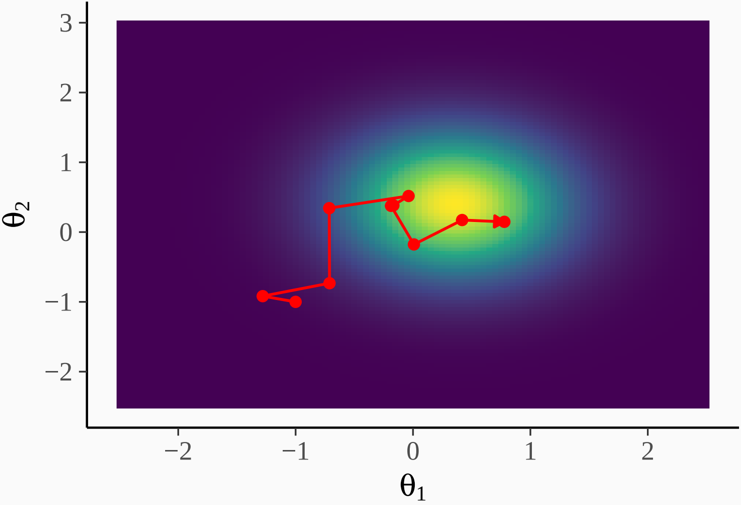
Stan

https://mc-stan.org/

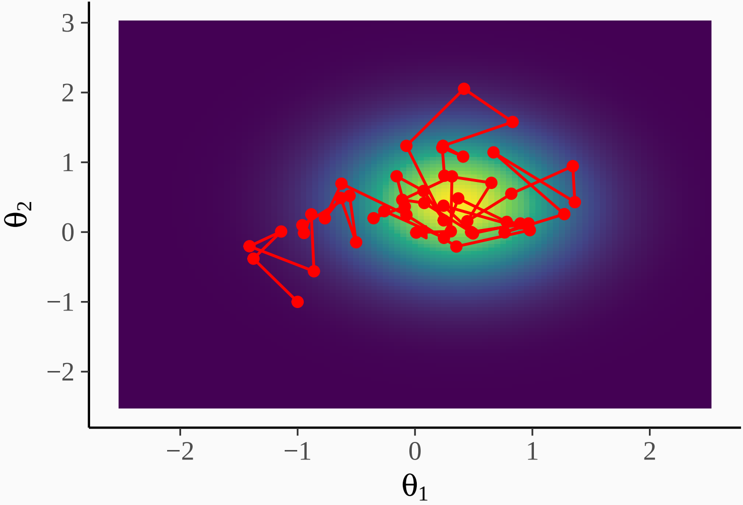
Stan syntax: Linear Regression

```
data {
  int<lower=1> N;  // total number of observations
  vector[N] y;  // response variable
  int<lower=1> K;  // number of regression coefficients
  matrix[N, K] X;  // predictor design matrix
}
parameters {
  vector[K] b;  // regression coefficients
  real<lower=0> sigma;  // residual standard deviation
}
model {
  vector[N] mu = X * b;  // predicted means
  b ~ normal(0, 10);  // prior
  sigma ~ exponential(1);  // prior
  y ~ normal(mu, sigma);  // likelihood
}
```

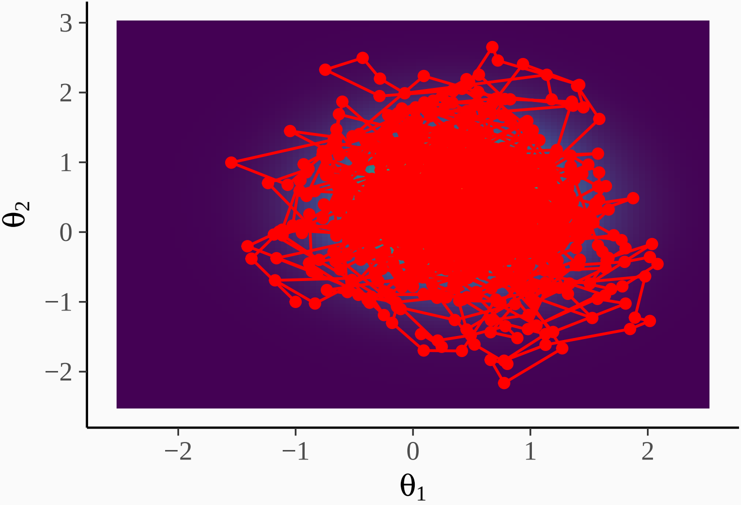
MCMC Sampling: A Single Chain (10 Iterations)



MCMC Sampling: A Single Chain (50 Iterations)



MCMC Sampling: A Single Chain (1000 Iterations)



All we care about are expectations

Expectation of some function f over the distribution $p(\theta \mid y)$:

$$\mathbb{E}_p(f) = \int f(\theta)\, p(\theta \mid y)\, \mathrm{d}\theta$$

Monto-Carlo Estimator

Having obtained exact random draws $\{\theta_s\}$ from $p(\theta \mid y)$:

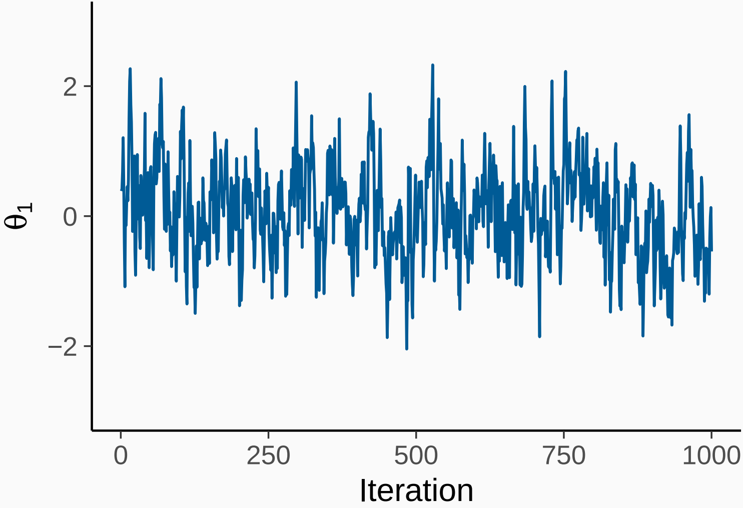
$$\frac{1}{S}\sum_{s=1}^{S} f(\theta_s) \sim \text{Normal}\left(\mathbb{E}_p(f), \sqrt{\frac{\text{Var}_p(f)}{\text{S}}}\right)$$

Markov-Chain Monto-Carlo Estimator

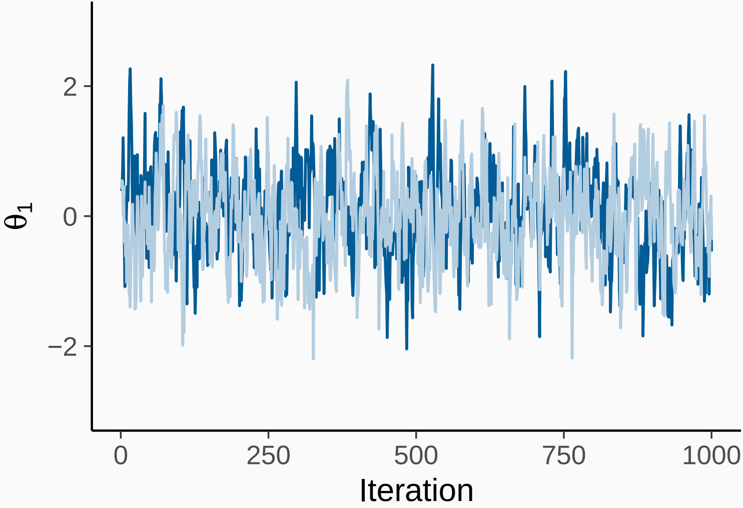
Assuming *geometric ergodicity* of a Markov Chain $\{\theta_s\}$:

$$\frac{1}{S}\sum_{s=1}^{S} f(\theta_s) \sim \text{Normal}\left(\mathbb{E}_p(f), \sqrt{\frac{\text{Var}_p(f)}{\text{ESS}}}\right)$$

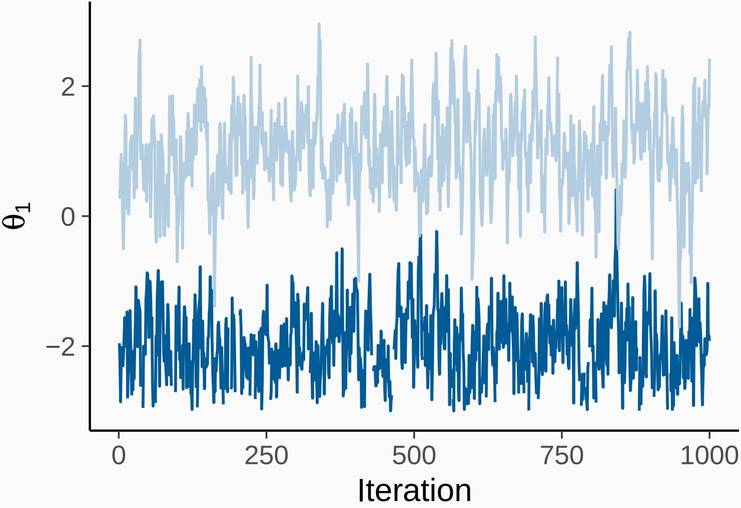
Trace Plots: Visualizing a Single Chain



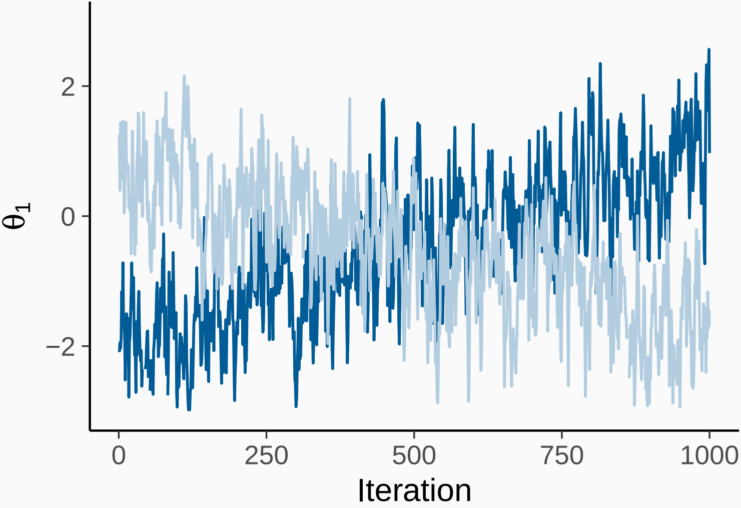
Trace Plots: Visualizing Multiple Chains



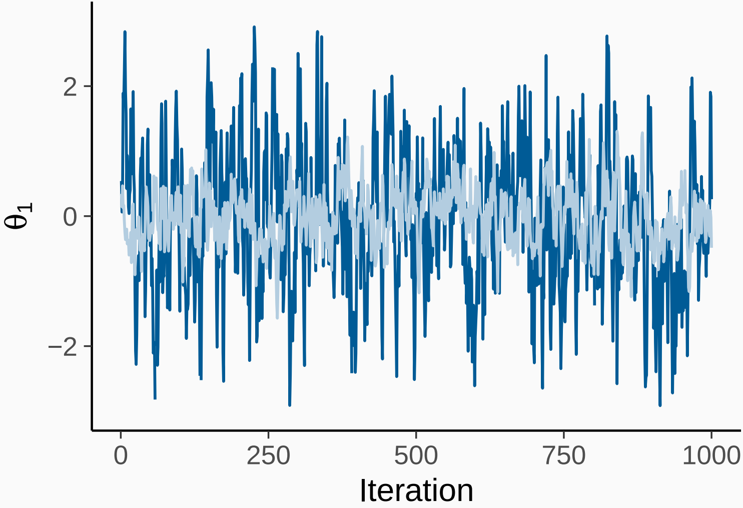
Chains with Different Locations



Non-Stationary Chains



Chains with Different Variances



Traditional MCMC Diagnostics

Between Chain Variance:

$$B = \frac{N}{M-1} \sum_{m=1}^{M} (\overline{\theta}^{(.m)} - \overline{\theta}^{(..)})^2$$

Within Chain Variance:

$$W = \frac{1}{M(N-1)} \sum_{m=1}^{M} \sum_{n=1}^{N} (\theta^{(nm)} - \overline{\theta}^{(.m)})^2$$

Potential Scale Reduction Factor:

$$\widehat{R} = \sqrt{\frac{\frac{N-1}{N} W + \frac{1}{N} B}{W}}$$

Effective Sample Size:

$$\mathrm{ESS} = \frac{N\,M}{\hat{\tau}}$$

Problems with the Traditional MCMC Diagnostics

(1) We do not detect differences of chains with infinite means

(2) We do not detect non-convergence in the tails of the distribution

(3) We cannot properly localize convergence problems

Solutions provided in:

Vehtari A., Gelman A., Simpson D., Carpenter B., & **Bürkner P. C.** (2020). Rank-normalization, folding, and localization: An improved Rhat for assessing convergence of MCMC. *Bayesian Analysis*. 1–28. doi:10.1214/20-BA1221

Simulation-Based Calibration

Idea based on the following identity:

$$p(\theta) = \int p(\theta \mid \tilde{y})\; p(\tilde{y} \mid \tilde{\theta})\; p(\tilde{\theta})\; d\tilde{y} d\tilde{\theta}$$

Repeat the following steps multiple times:

(1) Sample $\tilde{\theta} \sim p(\theta)$

(2) Sample $\tilde{y} \sim p(y \,|\, \tilde{\theta})$

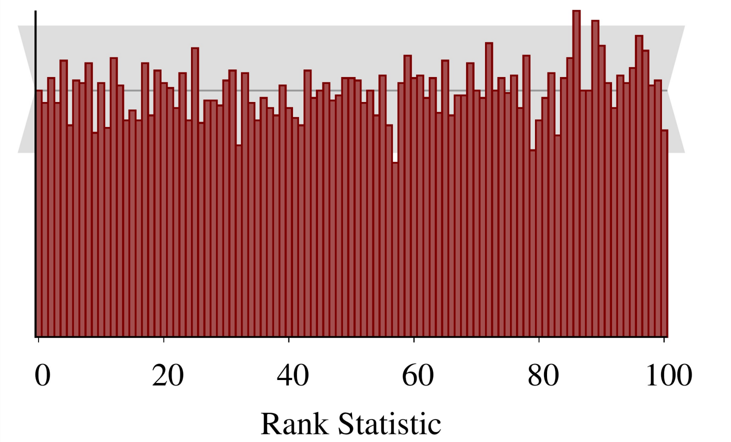
(3) Sample $\{\theta_1, \ldots, \theta_L\} \sim p(\theta|\tilde{y})$

(4) Compute $\mathrm{rank}(f(\tilde{\theta}), \{f(\theta_1), \ldots, f(\theta_L)\})$

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., & Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *https://arxiv.org/abs/1804.06788*

Simulation-Based Calibration: Illustration

Example for a well calibrated posterior:



Cross-Validation

Steps in cross-validation:

(1) Split the data into two Subsets: training data and test data
(2) Fit the model on the training data
(3) Evaluate the predictions on the test data
(4) Repeat (1) to (3) with multiple data splits
(5) Summarize the results of all splits

Types of cross-validation (selection):

- Leave-one-out cross-validation (LOO-CV)
- K-fold cross-validation (K-fold-CV)
- Leave-group-out cross-validation (LGO-CV)
- Leave-future-out cross-validation (LFO-CV)

Measures of Predictive Accuracy / Utility

Example measures for a single data split:

$$\begin{aligned}\text{ELPD} = \log\ p(y|y_{\text{tr}}) &= \log \int p(y|\theta)\ p(\theta|y_{\text{tr}})\ d\theta \\ &\approx \log\ \frac{1}{S}\sum_{s=1}^{S} p(y|\theta^{(s)})\end{aligned}$$

$$\text{RMSE} = \sqrt{\int (y-\hat{y})^2\ p(\hat{y}|y_{\text{tr}})\ d\hat{y}} \approx \sqrt{\frac{1}{S}\sum_{s=1}^{S}(y-\hat{y}^{(s)})^2}$$

Leave-One-Out Cross-Validation

Leave out a single observation y_i and predict by all other observations y_{-i} using the ELPD:

$$\text{ELPD} = \sum_{i=1}^{N} \log\ p(y_i | y_{-i})$$

(other measures are possible as well)

Important properties of LOO-CV:

- All possible N splits can be evaluated
- Can be approximated using the full model

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.

Importance Sampling

Approximate expectations over a target distribution $f(\theta)$ using an approximating proposal distribution $g(\theta)$:

$$\mathbb{E}_f[h(\theta)] = \int h(\theta) f(\theta)\, d\theta = \frac{\int h(\theta) f(\theta)\, d\theta}{\int f(\theta)\, d\theta} = \frac{\int h(\theta) r(\theta) g(\theta)\, d\theta}{\int r(\theta) g(\theta)\, d\theta}$$

Raw importance ratios:

$$r(\theta) = \frac{f(\theta)}{g(\theta)}$$

Approximation via $\theta^{(s)} \sim g(\theta)$:

$$\mathbb{E}_f[h(\theta)] \approx \frac{\sum_{s=1}^{S} h(\theta^{(s)}) r(\theta^{(s)})}{\sum_{s=1}^{S} r(\theta^{(s)})}$$

Case Study: Roaches

Research question: Does a treatment reduce the number of roaches?

Data set of 262 apartments with the following variables:

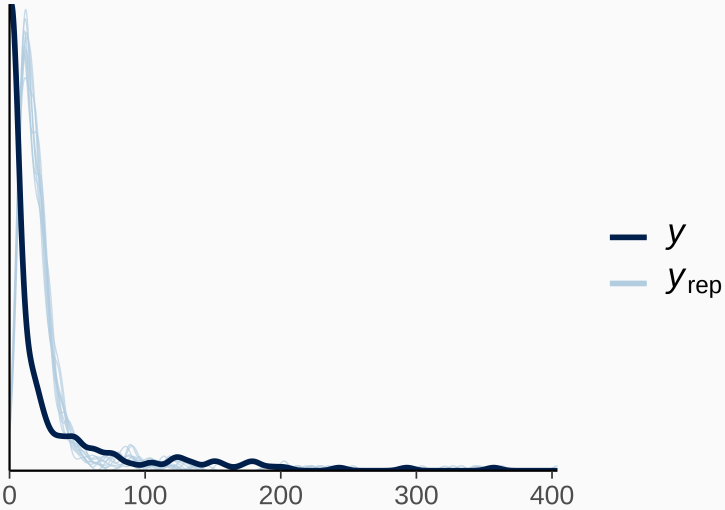
- `roach1`: Number of roaches counted before treatment within one hour (between 0 and 450)
- `y`: Number of roaches after treatment (between 0 and 357)
- `exposure2`: Time frame in which we counted `y` (between 0.2 and 4 hours)
- `treatment`: Dichotomous treatment indicator (0 or 1)

Choosing an initial model

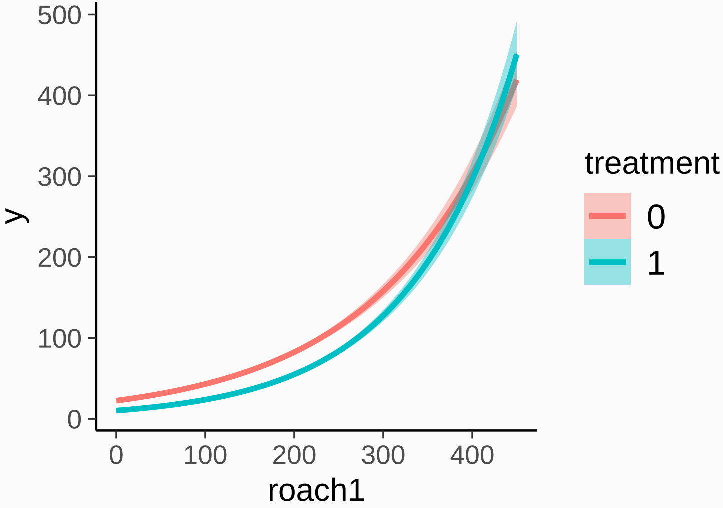
```r
model1 <- brm(
  y ~ treatment * roach1 + offset(log(exposure2)),
  family = poisson("log"),
  prior = prior(normal(0, 5), class = "b"),
  ...
)
```

Bürkner P. C. (2017). brms: An R Package for Bayesian Multilevel Models using Stan. *Journal of Statistical Software*. 80(1), 1-28. doi:10.18637/jss.v080.i01

Posterior Predictive Checking



Visualization of Predictions



Model Comparison

```r
model2 <- brm(
  y ~ treatment + roach1 + offset(log(exposure2)),
  family = poisson("log"),
  prior = prior(normal(0, 5), class = "b"),
  ...
)
```

```
##        elpd_diff se_diff
## model1   0.0       0.0
## model2 -20.6      91.9
```

Learn More

Learn more about me:

- Website: https://paul-buerkner.github.io/
- Publications: https://paul-buerkner.github.io/publications/
- Email: `paul.buerkner@gmail.com`
- Twitter: @paulbuerkner

Learn more about Stan:

- Website: http://mc-stan.org/
- Forums: http://discourse.mc-stan.org/