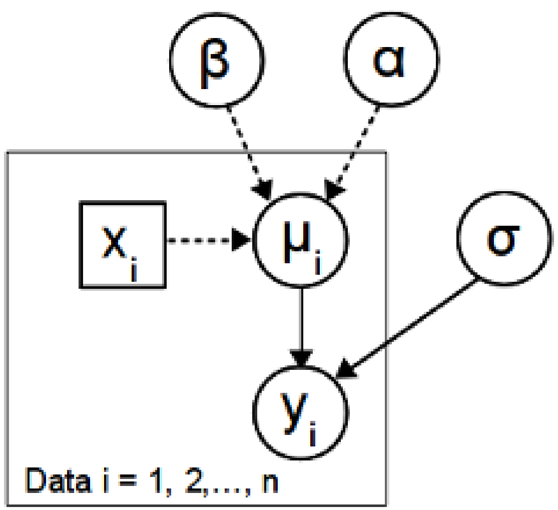
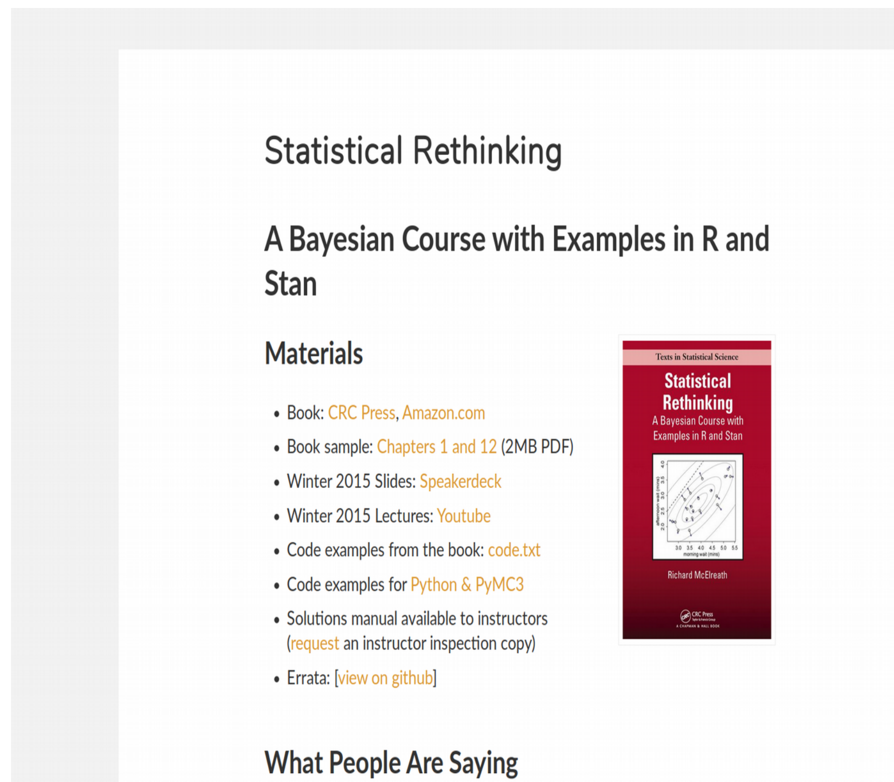
Introduction to probabilistic programming:

Leo Lahti & Ville Laitinen, VIB/KU Leuven & University of Turku, Finland

Lecture 1: Introduction to Bayesian thinking

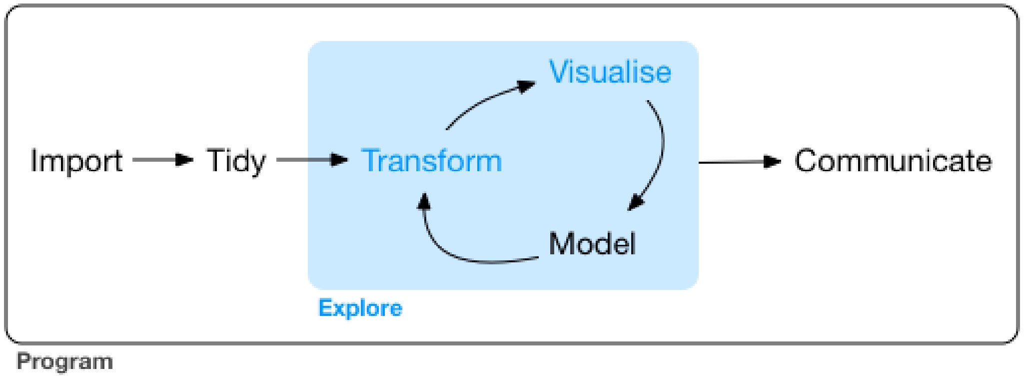




Today's contents & aims:
probabilistic programming

- **Theory**: intuition building on probabilistic models
- **Methods**: context within the overall data science workflow
- **Tools**: Hands-on experience in implementing and interpreting probabilistic models in R
- **Application**: Hands-on exercises
- **More**: Where to learn more

After this lecture, you will have a better understanding on how to use probabilistic programming as an alternative to classical models



Schedule

9:00 - 10:30 Introduction to Bayesian thinking and tools

10:30 - 11:30 Exercise session 1

11:30 - 12:00 Discussion & Conclusions

12:00 - 13:00 Lunch break

13:00 - 13:30 Extending probabilistic models

13:30 - 14:30 Exercise session 2

13:30 - 14:30 Discussion & Conclusions

15:00 - 15:15 Coffee break

15:15 - 15:45 Further remarks

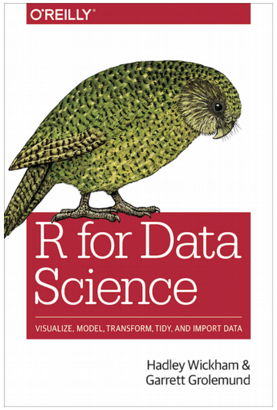
15:45 - 16:30 Exercise session 3

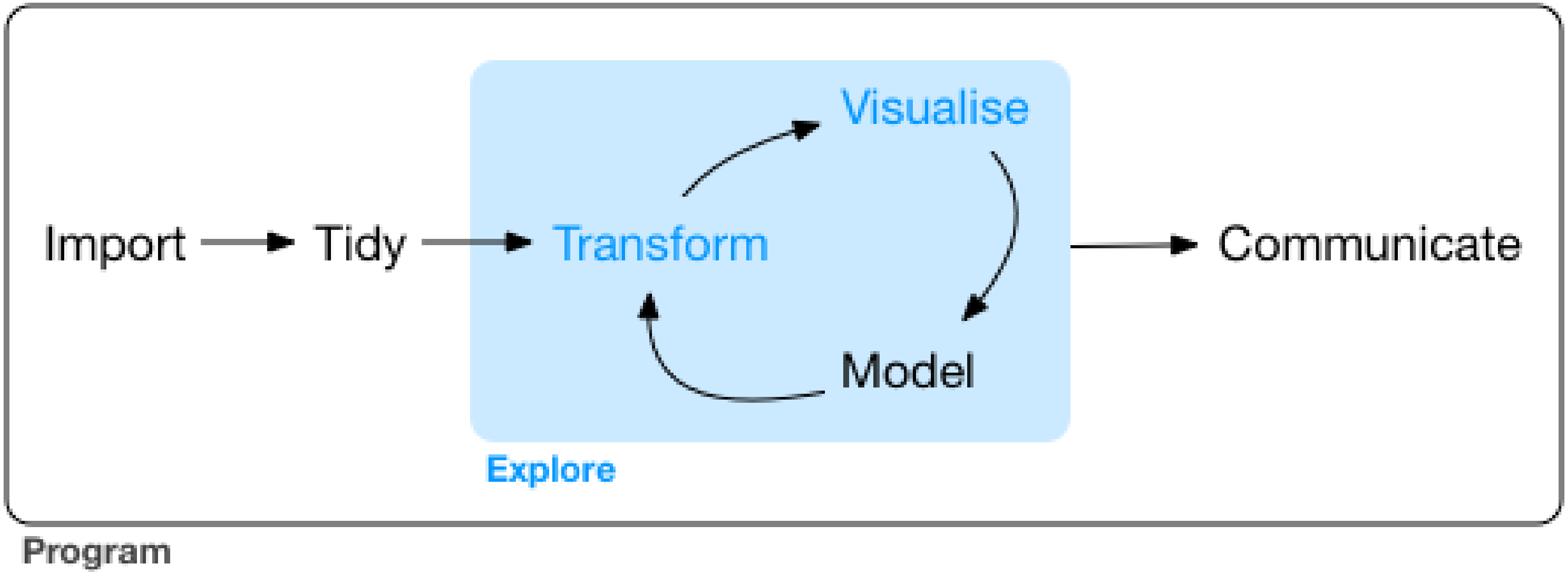
16:30 - 17:00 Discussion & Conclusions

Recap: data science workflow

- Statistical techniques deal with uncertainty in data analysis and modeling
- Tidy & trustworthy data is a prerequisite for successful analysis !
- Automation and standardization facilitate transparency, reproducibility, and model building

http://r4ds.had.co.nz





Bayesian (probabilistic) approach

x Data θ Parameters

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \sim P(x|\theta)P(\theta)$$

posterior likelihood prior

Frequentist (classical): maximize likelihood of the data, given parameters

Bayesian: evaluate posterior distribution of the parameters, given data and priors

Linear regression as a probabilistic model

$$\boxed{y \sim N(\alpha + \beta x, \sigma)}$$

Express the assumptions in a mathematical/graphical form

Express all parameters & variables as random variables

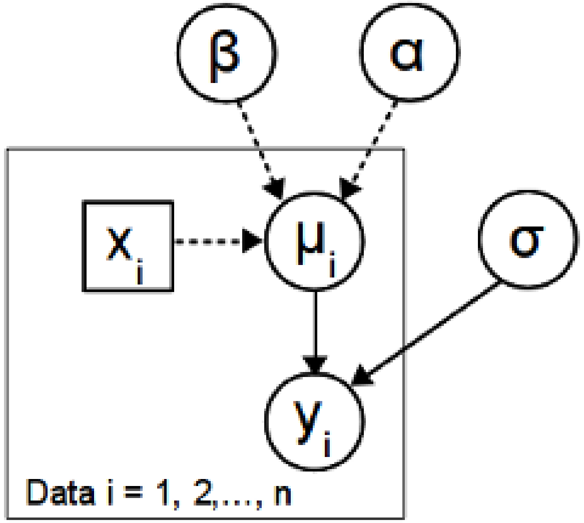
Indicate dependencies between variables

Linear regression as a probabilistic model

$$\boxed{y \sim N(\alpha + \beta x, \sigma)}$$

Express the assumptions in a mathematical/graphical form

$$y \sim N(\mu, \sigma \ \ )$$
$$\mu = \alpha + \beta x$$



$y_i \sim \text{Normal}(\mu_i,\sigma)$
$\mu_i = \alpha + \beta x_i$

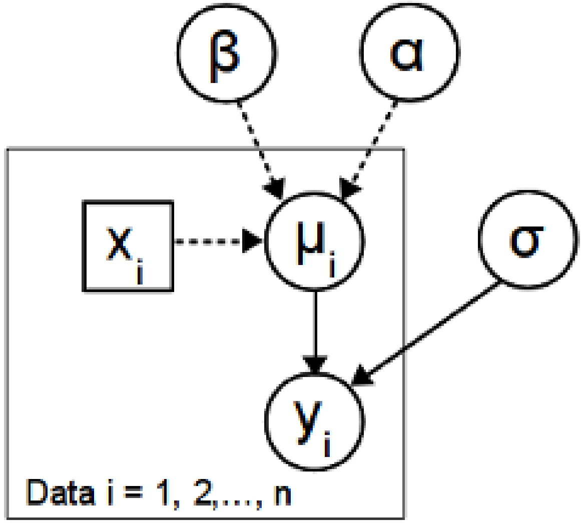
Indicate dependencies between variables

Linear regression as a probabilistic model

$$\boxed{y \sim N(\alpha + \beta x, \sigma)}$$

Express all parameters & variables as random variables

$$y \sim N(\mu, \sigma\ )$$
$$\mu = \alpha + \beta x$$
$$\alpha, \beta \sim N(0, 10^2)$$
$$\sigma \sim Cauchy(0, 5)$$



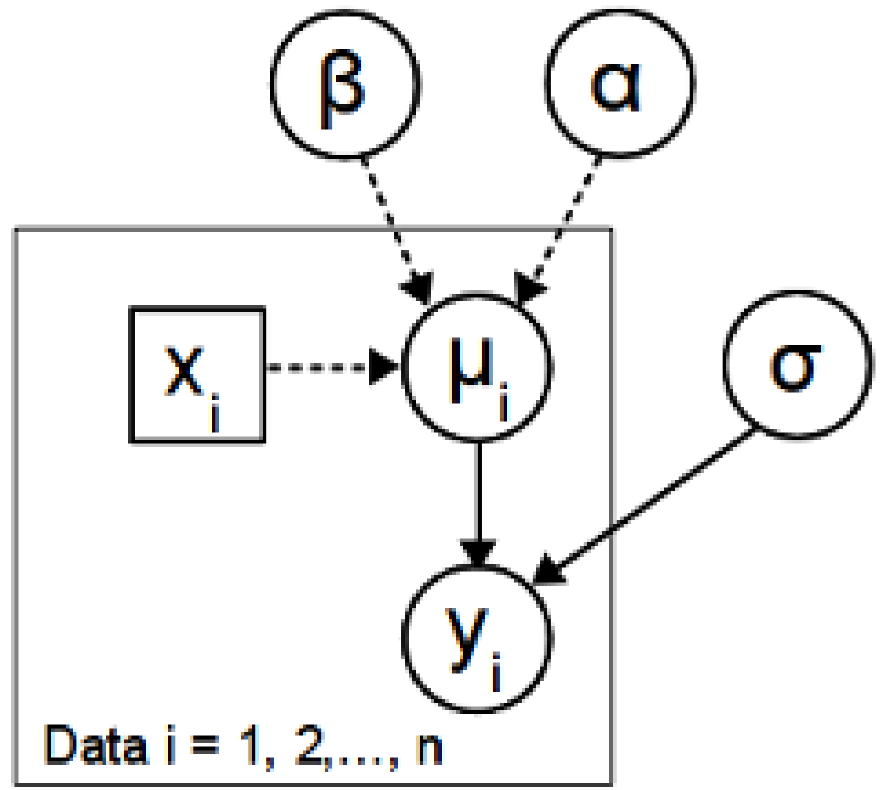
$y_i \sim \text{Normal}(\mu_i, \sigma)$
$\mu_i = \alpha + \beta x_i$

Linear regression as a probabilistic model

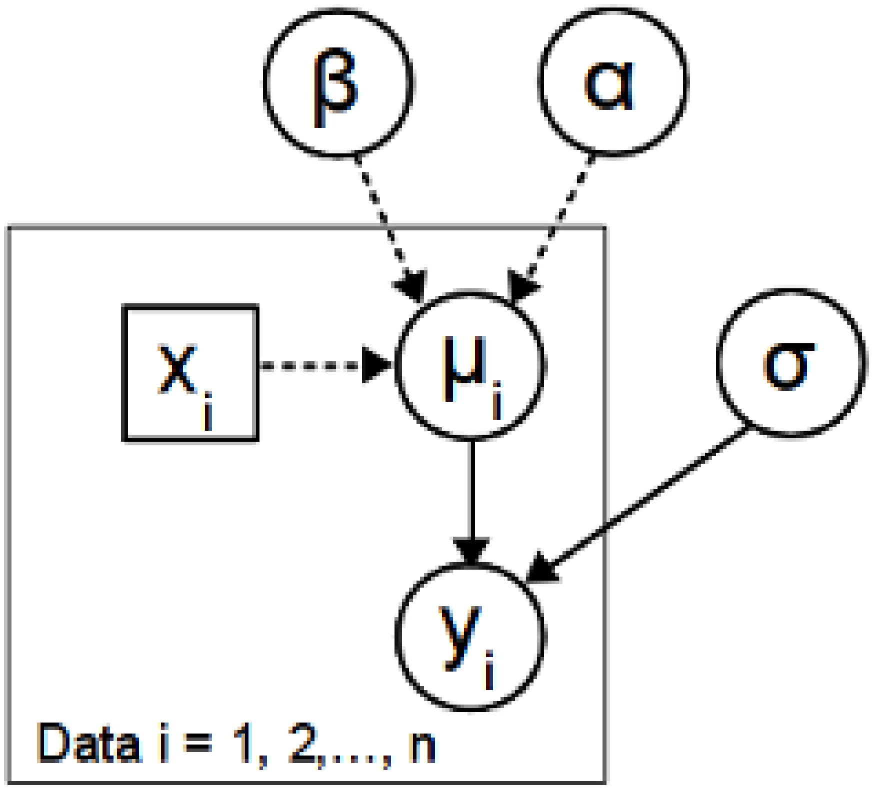
$$\boxed{y \sim N(\alpha + \beta x, \sigma)}$$

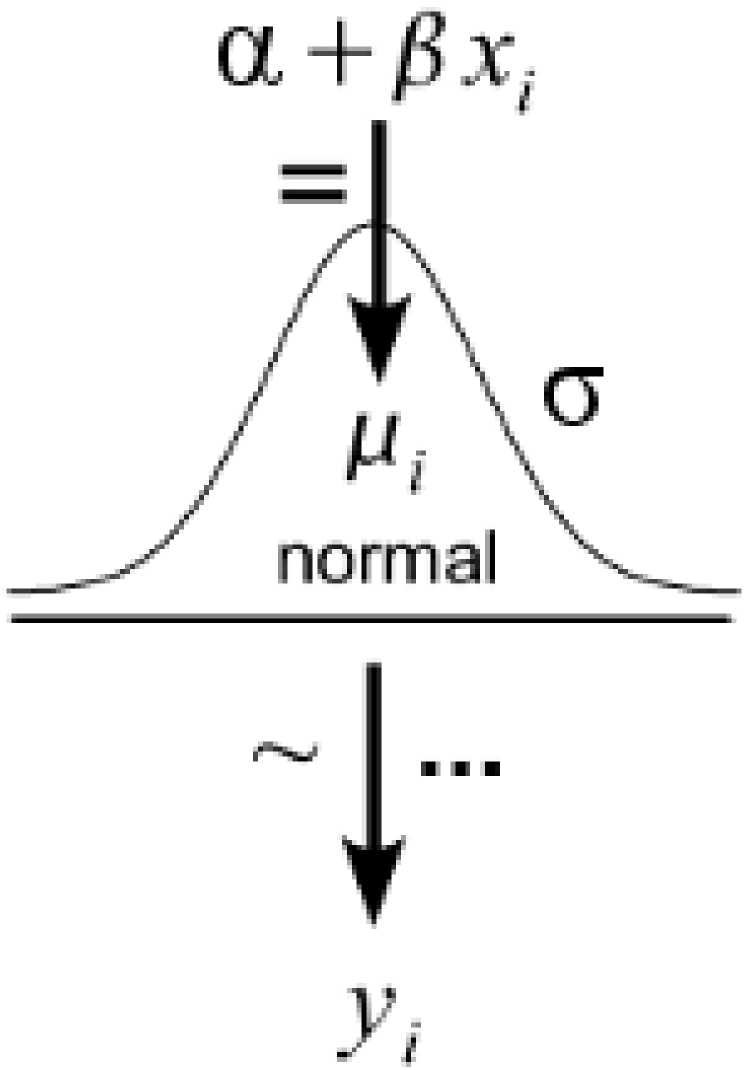
Bring in priors: intuitive generative model across all (observed and latent) variables

$$y \sim N(\mu, \sigma\ )$$
$$\mu = \alpha + \beta x$$
$$\alpha, \beta \sim N(0, 10^2)$$
$$\sigma \sim Cauchy(0, 5)$$

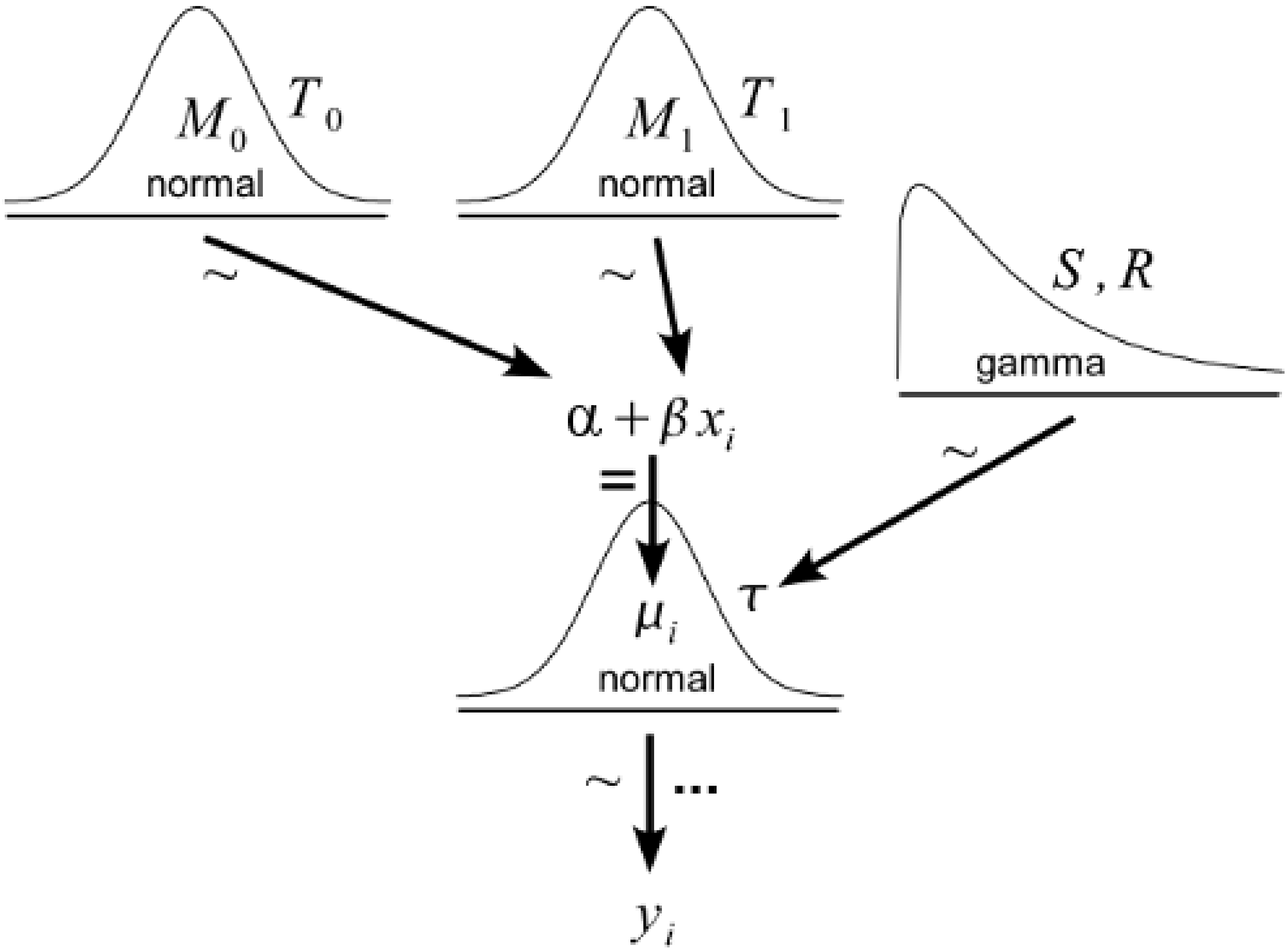


Alternative graphical representation (Kruschke model)

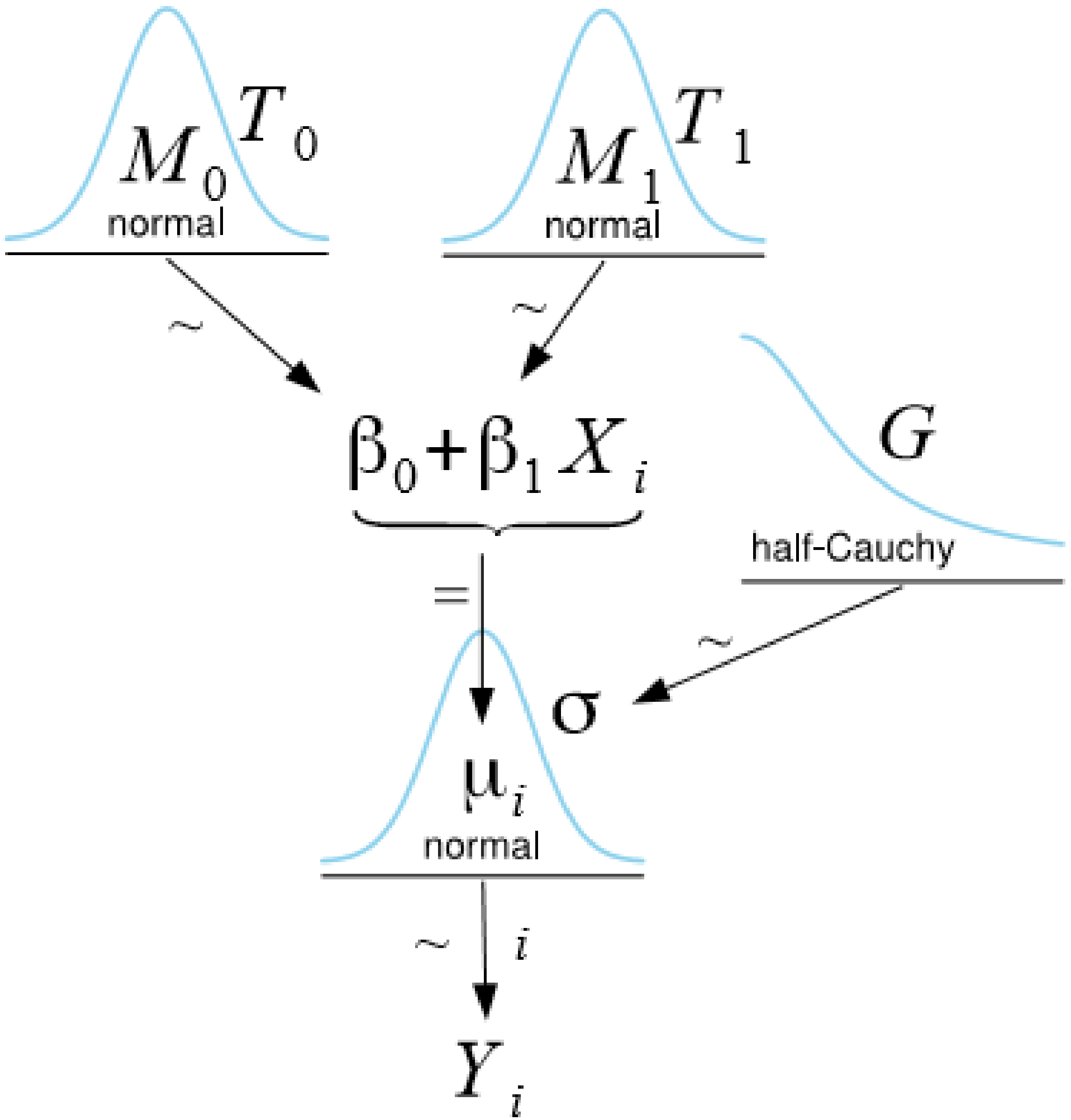




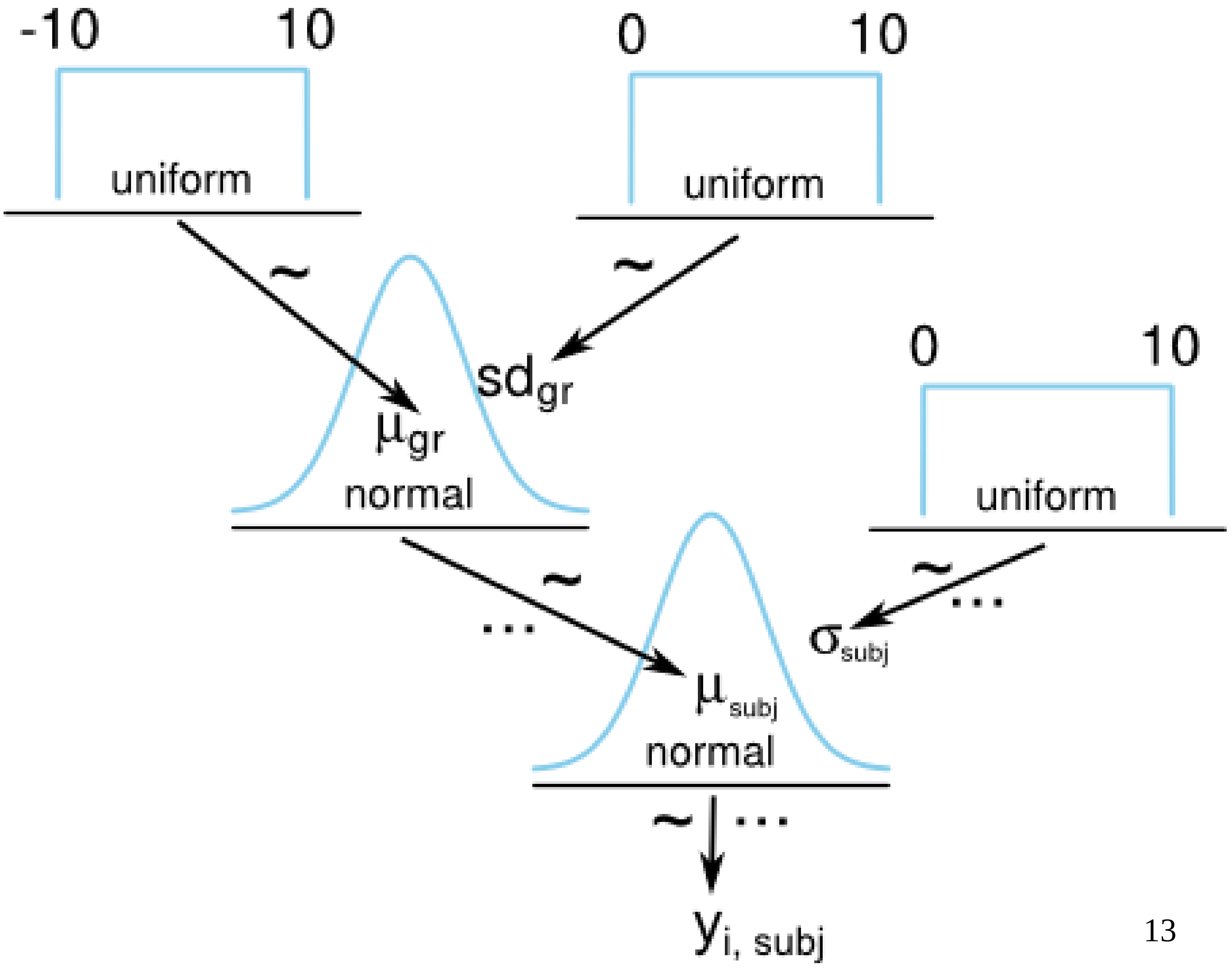
Role of priors



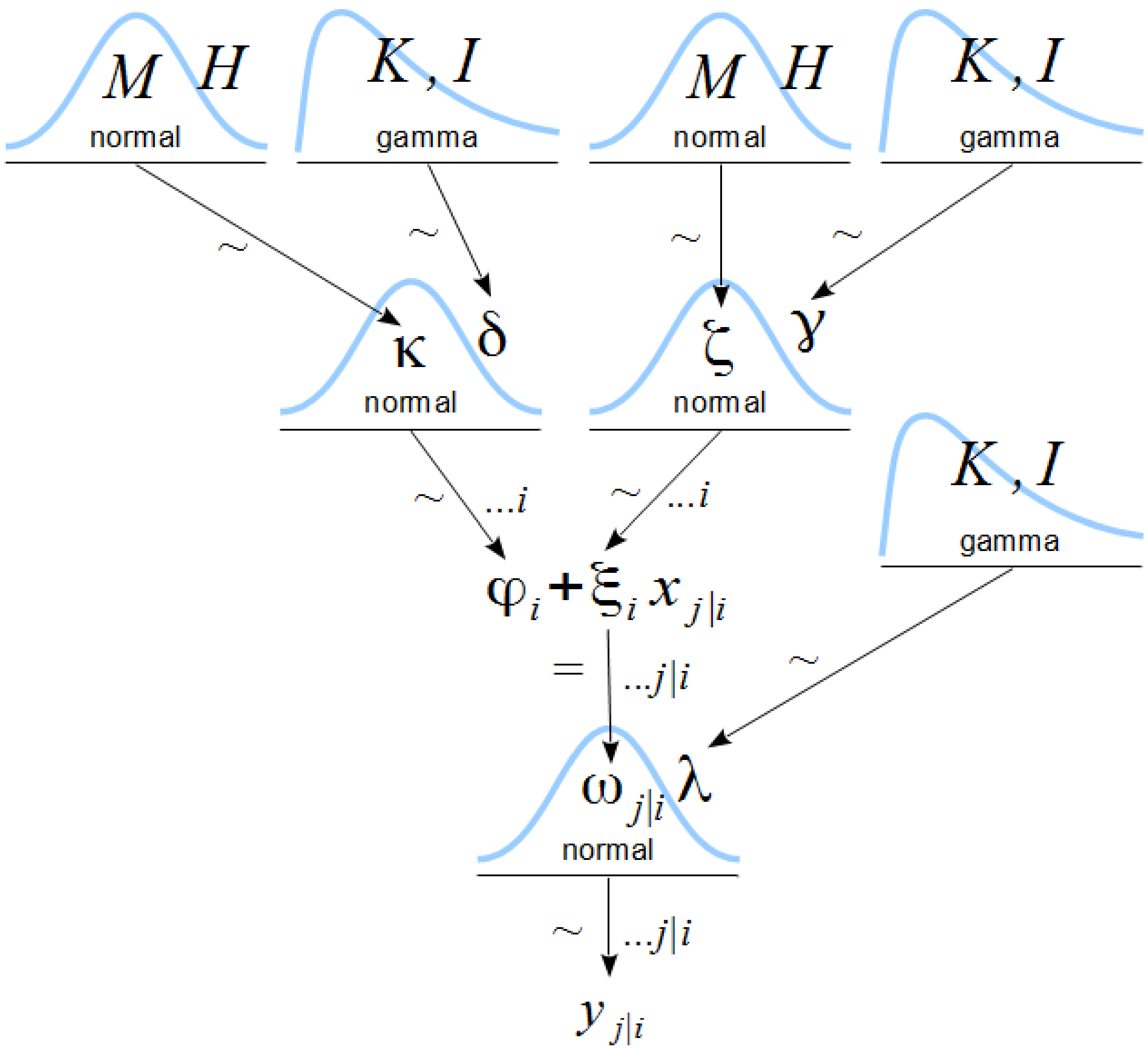
Role of priors



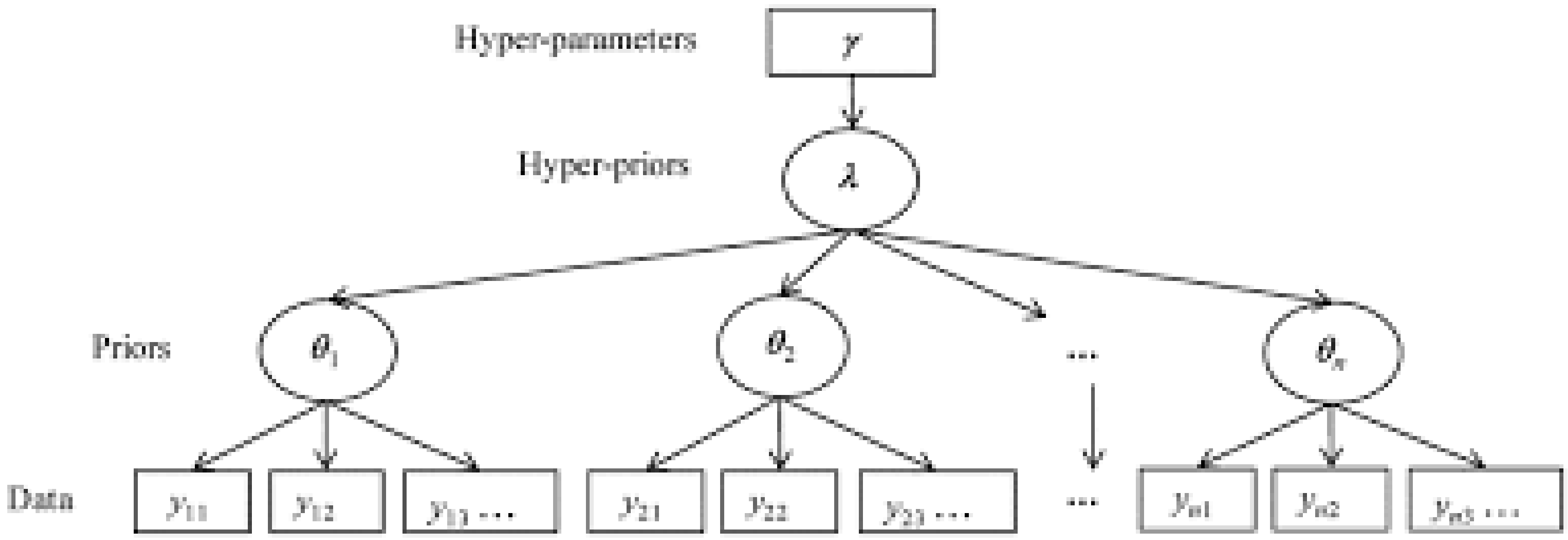
Role of priors



Role of hyperpriors



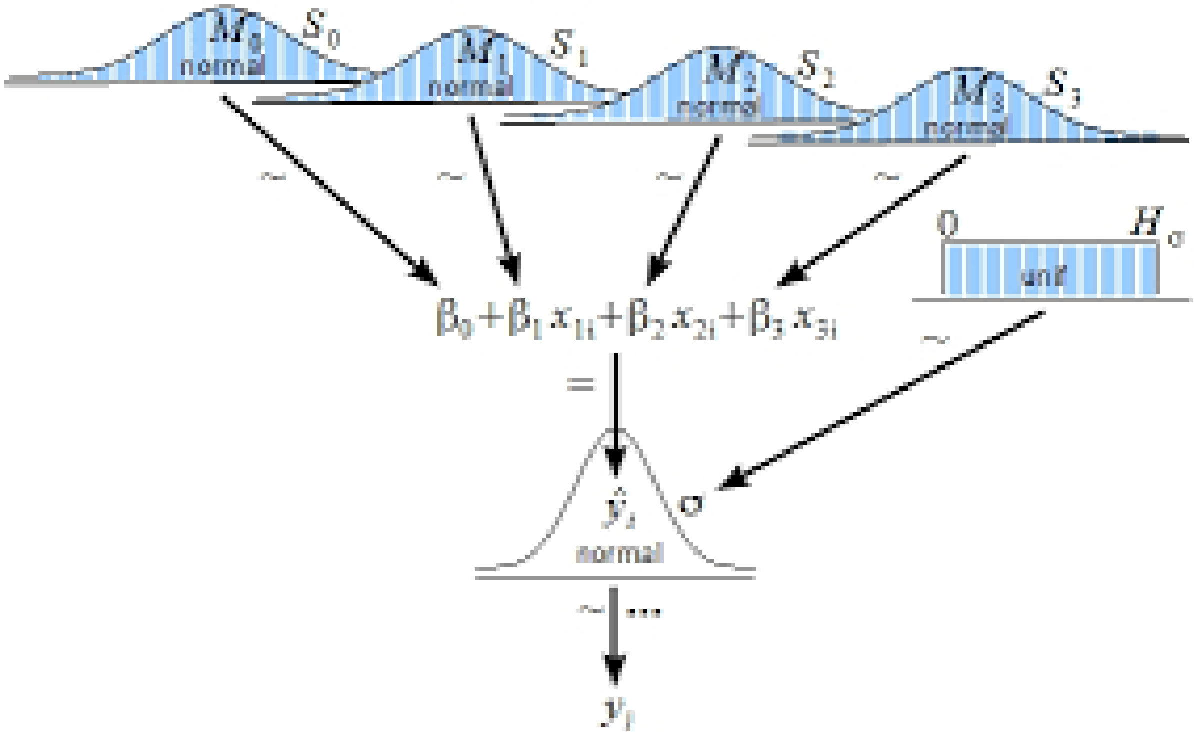
Levels of hierarchy



Full linear model

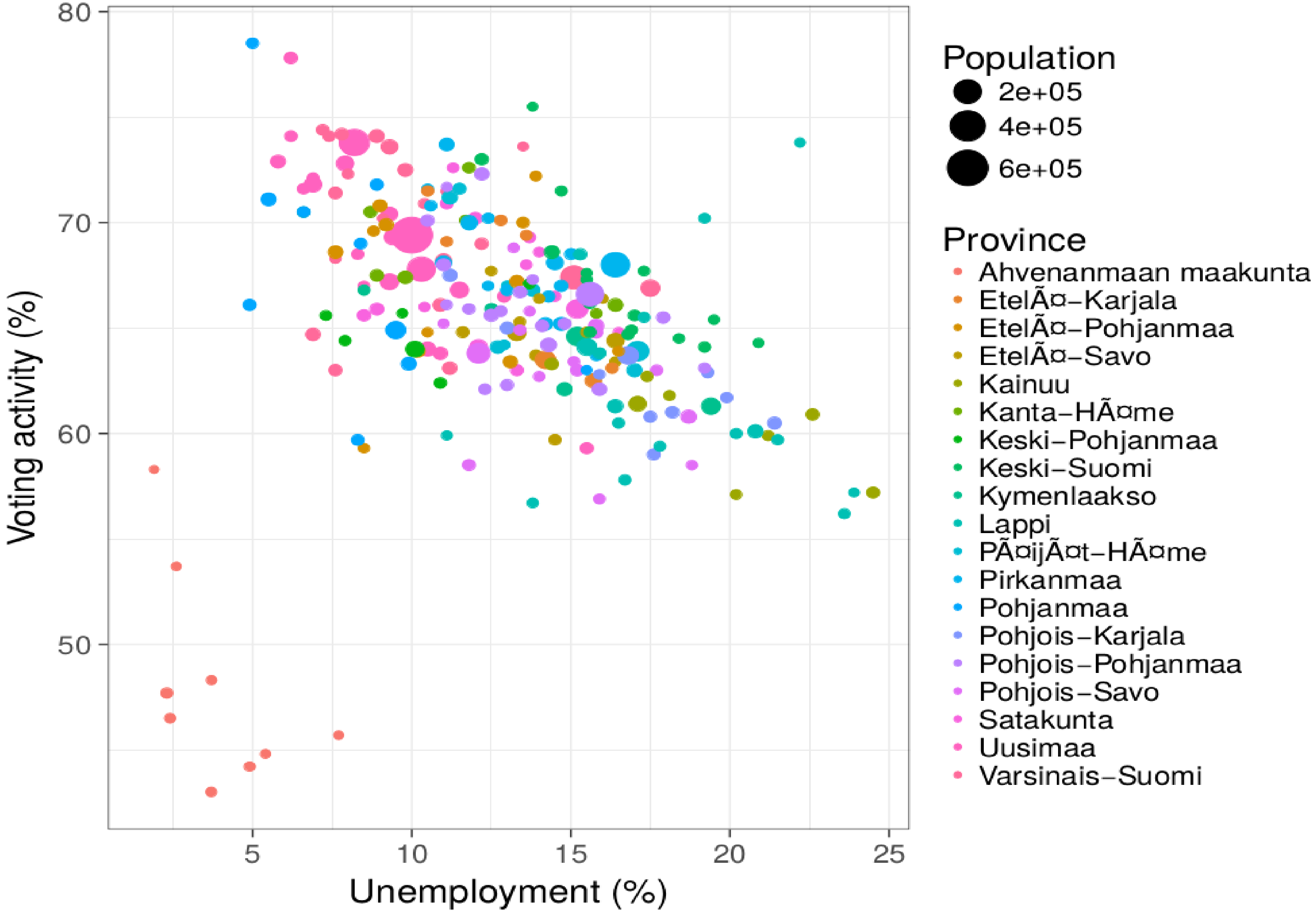
Priors can bring in prior information and support the analysis

Parameter pooling can be used to regularize the model



Example: unemployment and voting activity (Finland, 2015)

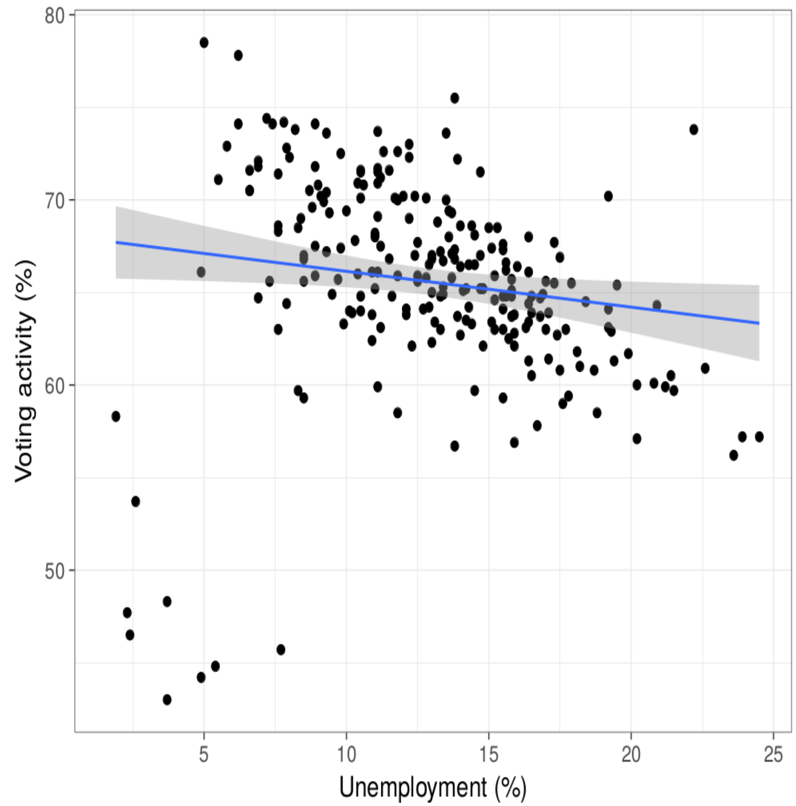
Data: Statistics Finland & Land Survey Finland (retrieved with pxweb & sorvi R packages)



Standard linear model: voting (y) ~ unemployment (x)

$$y \sim N(\alpha + \beta x, \sigma)$$

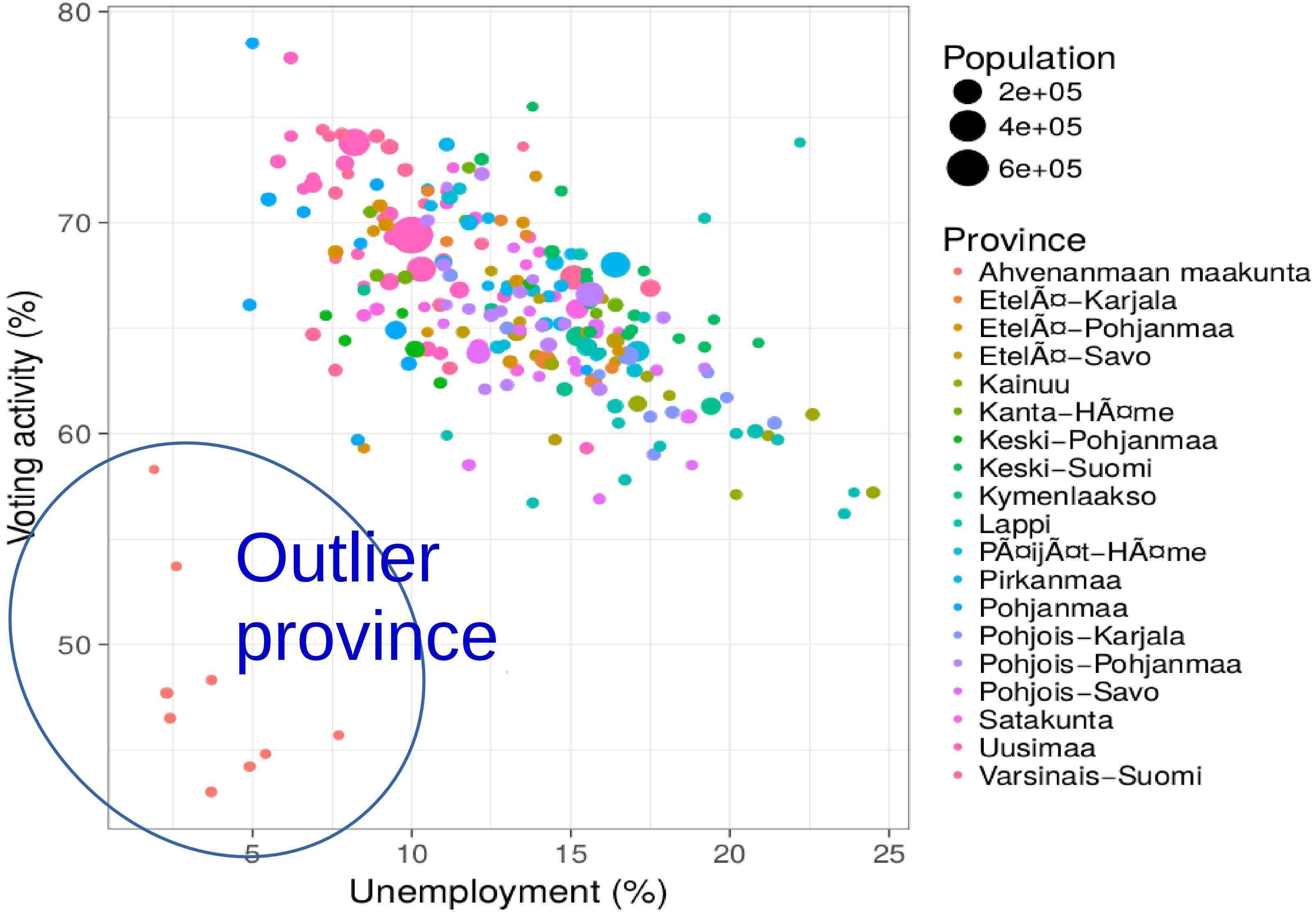
y = voting activity



X = unemployment

Unemployment and voting activity (Finland, 2015)

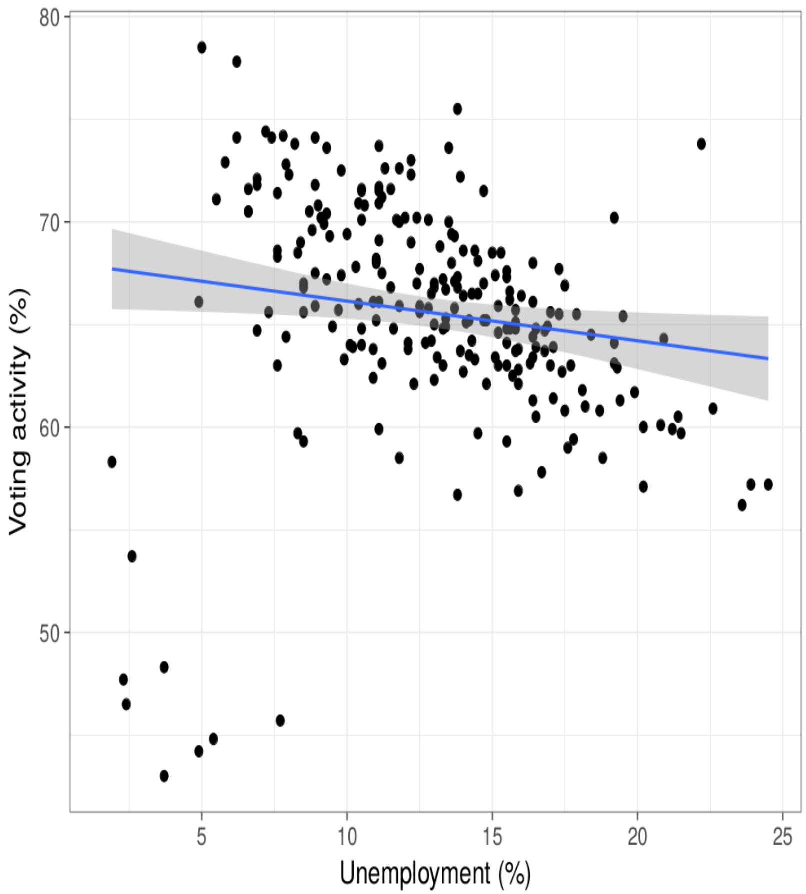
Data: Statistics Finland & Land Survey Finland (retrieved with pxweb & sorvi R packages)



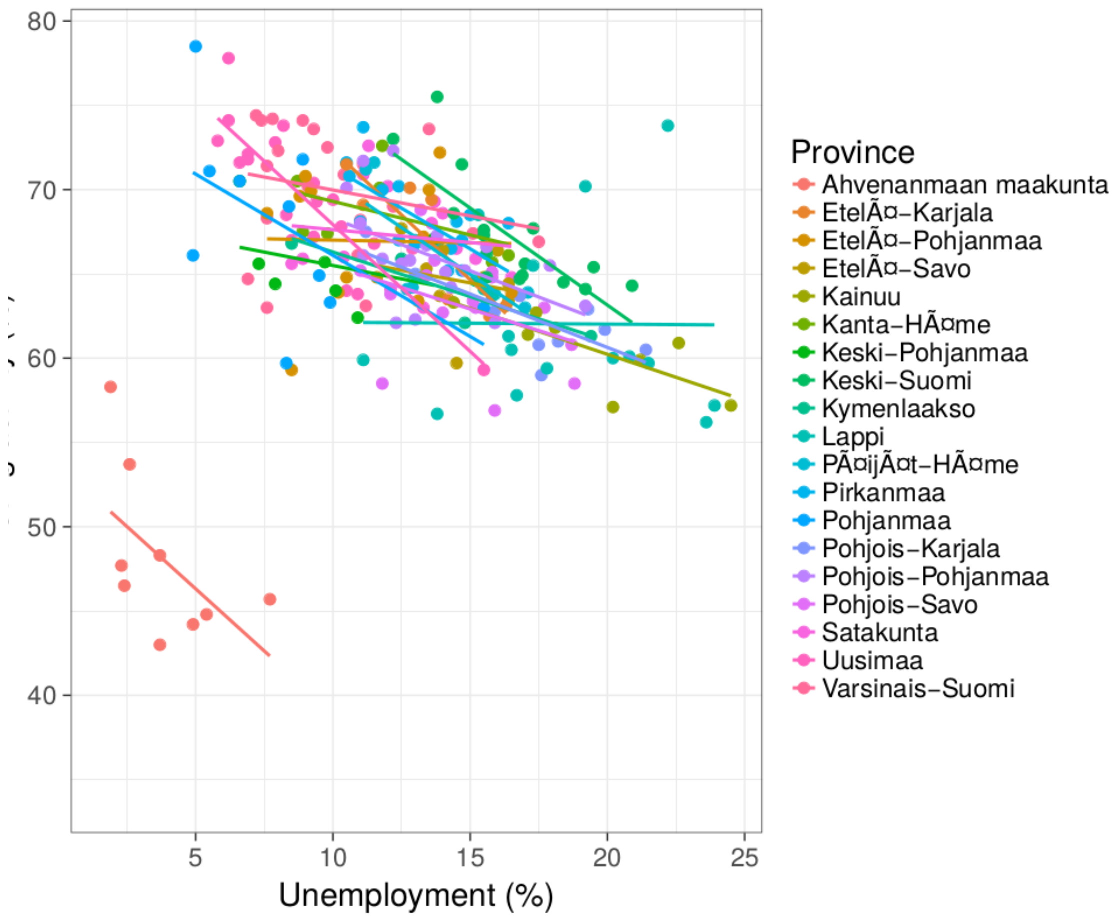
Linear models with confounders

Data: Statistics Finland & Land Survey Finland (retrieved with pxweb & sorvi R packages)

All municipalities together



Per province

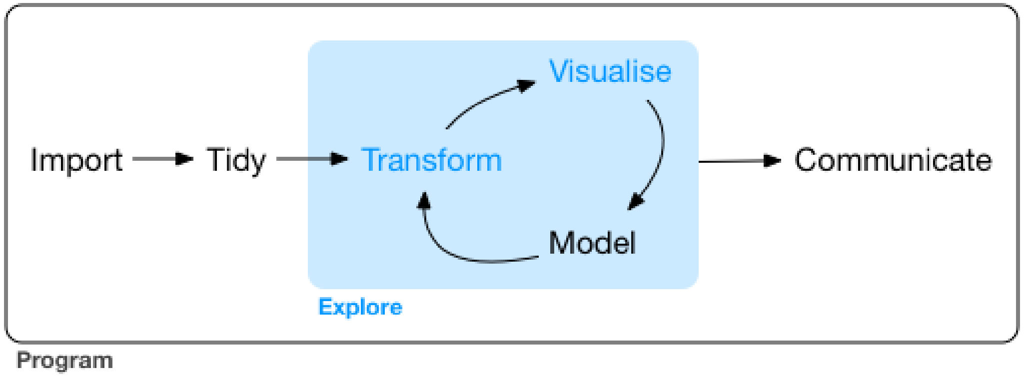


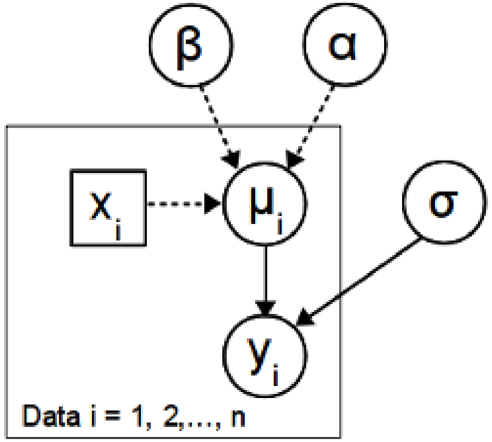
What is Stan ?

mc-stan.org

Programming language for probabilistic analysis

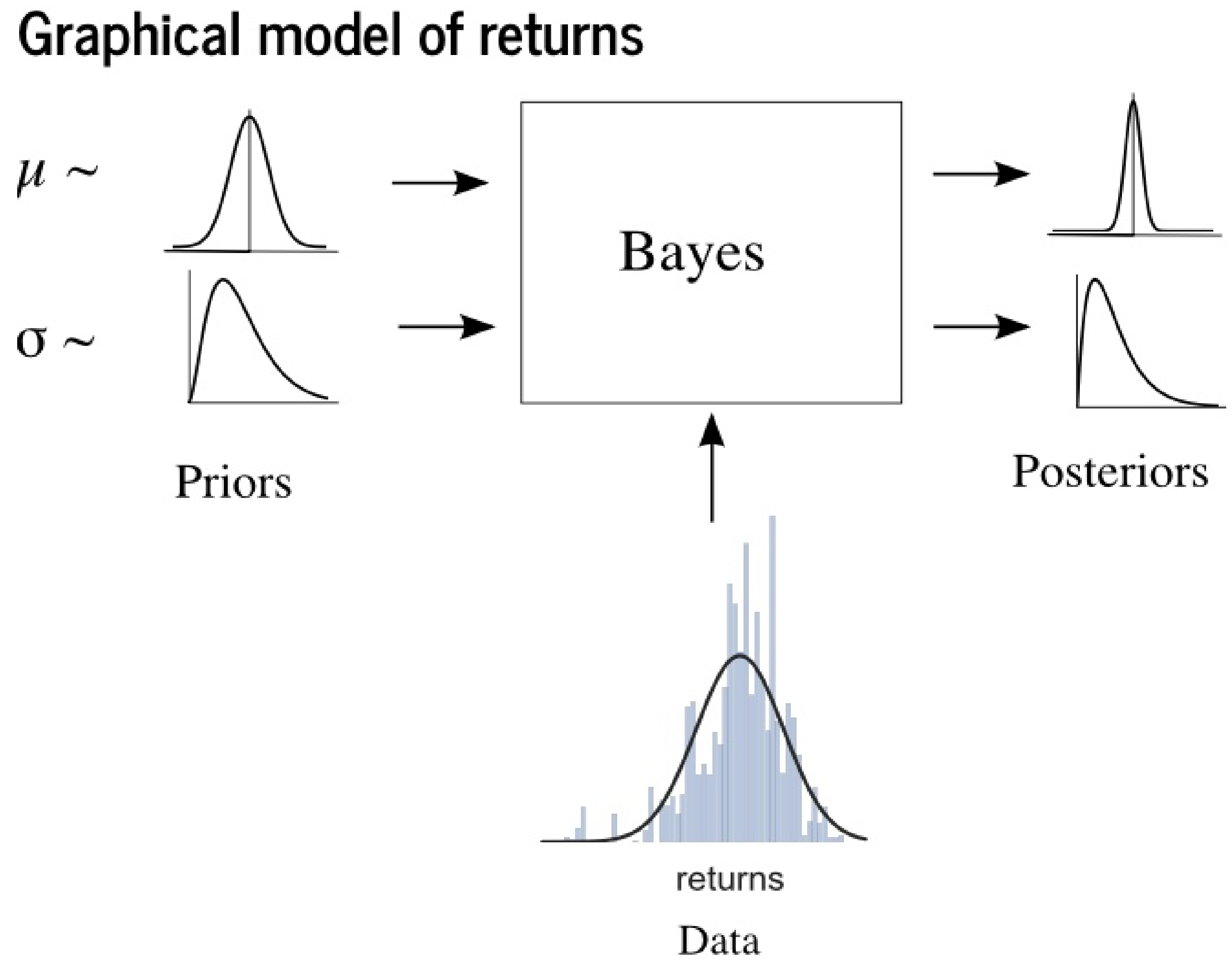
- Intuitive mathematical syntax to describe data and models
- Abstraction: technical details in the background. Separates modeling from inference & application from technicalities
- Powerful open interfaces in R/Python/Julia/etc
- Interpretation tools (shinystan, bayesplot)
- Support: comprehensive tutorials, active user communities





Modeling philosophy

- Define priors for the model parameters
- Update the beliefs based on observed data
- Analyze the posterior distributions (not point estimates!)



TRADITIONAL MACHINE LEARNING

Model and Fitting Algorithm are Conflated and Black Box

```
e.g. fit = nnet(x, y,
size = 2, decay = 5e-4,
maxit = 200)
```

POINT ESTIMATE PREDICTIONS

Stan

Model is Exposed in the Stan Program

```
model {
  y ~ normal(alpha +
       beta * x, sigma);
}
```

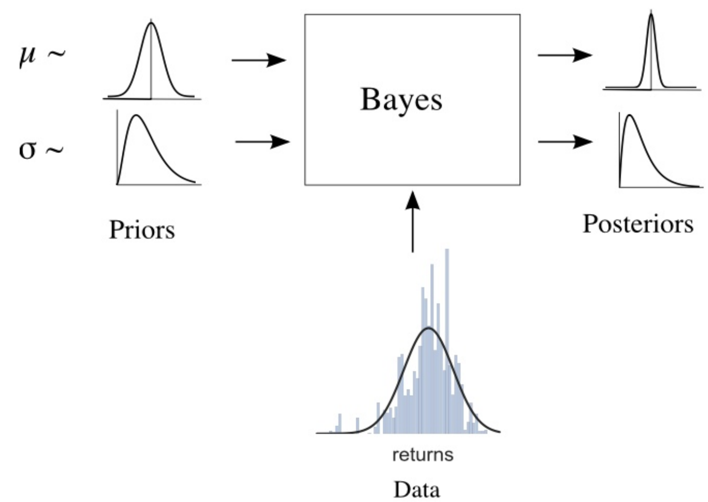
General Purpose Estimation Algorithms: HMC with NUTS, ADVI

PARAMETER DISTRIBUTIONS

PREDICTIVE DISTRIBUTION

DECISION THEORY

Graphical model of returns



Shortcuts for standard models: rstanarm

```
fit <- stan_glm(voting ~ unemployment,
                data = df,
                family = gaussian)
```

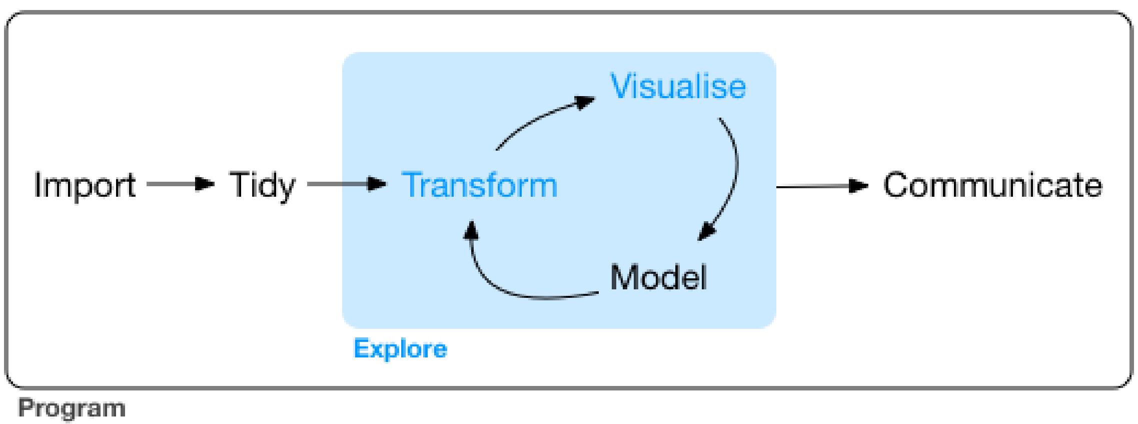
- Precompiled common regression models with default priors
- Write regular code. Add stan_ to the front, and add a prior (e.g. glm -> stan_glm)
- Detailed tutorials available!

Exercise session 1
https://microbiome.github.io/microbiome/rstanarm.html

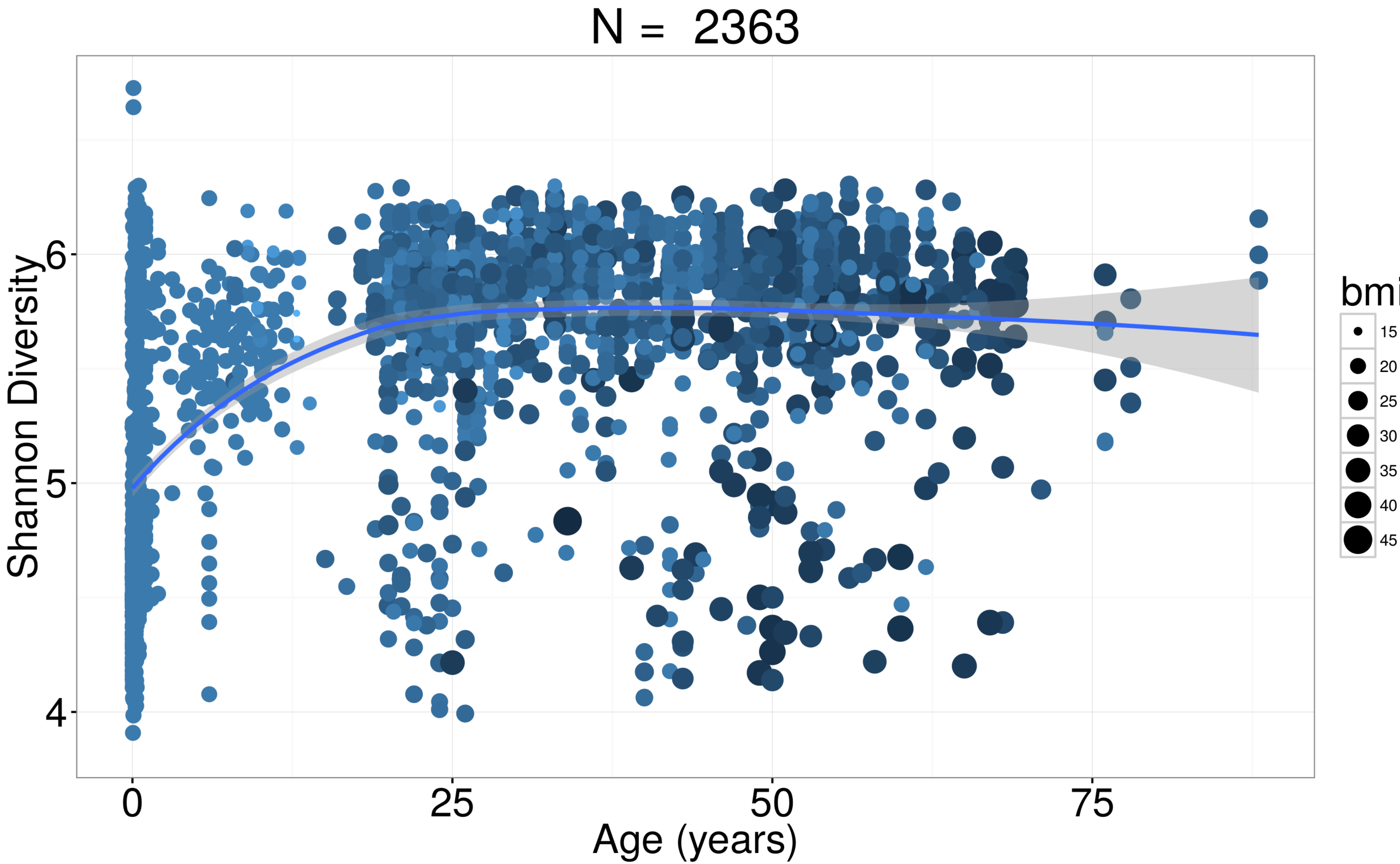
The exercises demonstrate how to:

- organize the data in stan-compatible format
- implement and run simple probabilistic models
- diagnose,visualize, interpret and revise

We will go through examples from the rstanarm tutorial

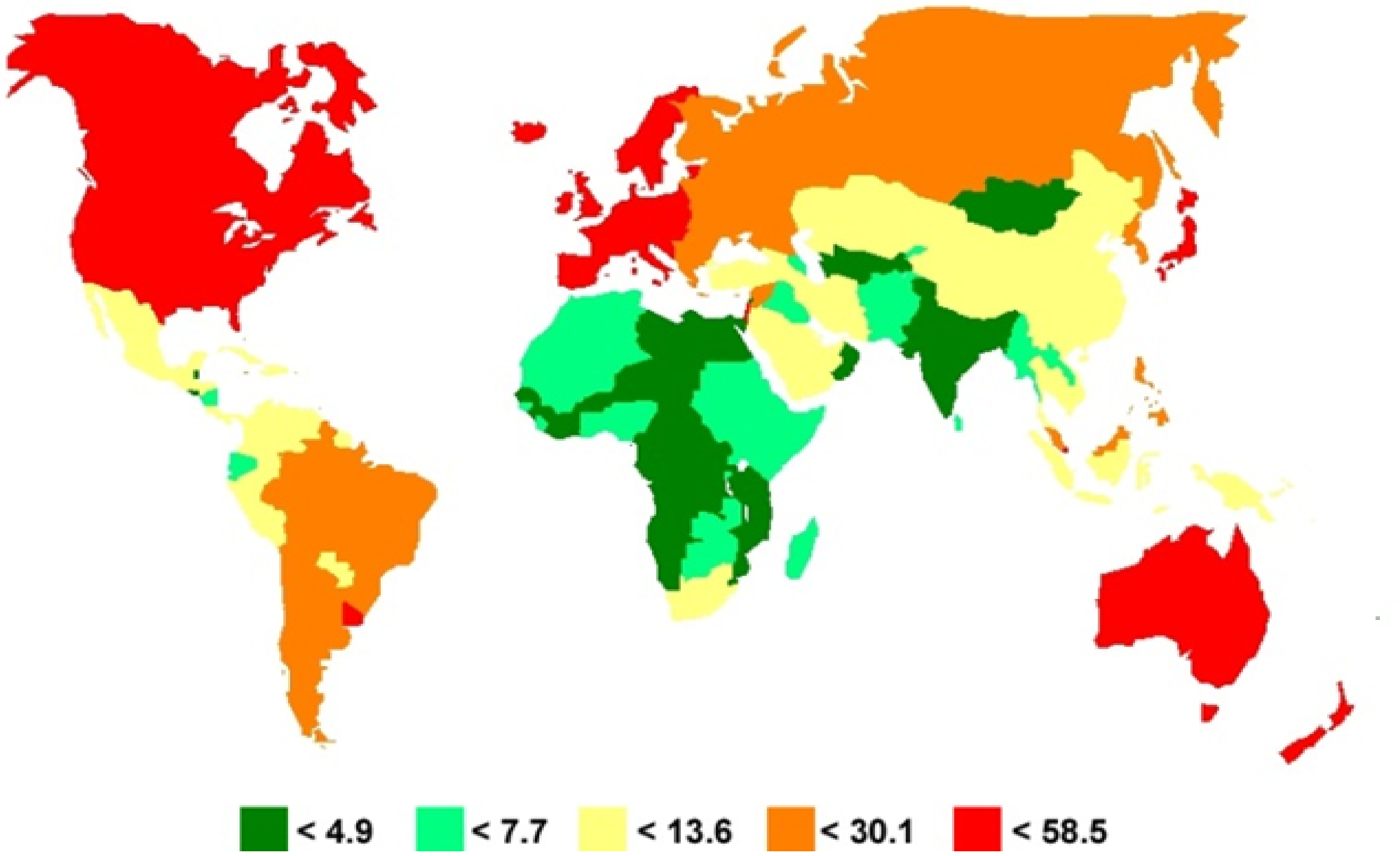


Microbiome diversity and age: healthy & normal obese subjects



Data: HITChip Atlas;
healthy fecal RBB samples

Colon cancer prevalence



Colon cancer rates per country

theguardian
Winner of the Pulitzer prize 2014

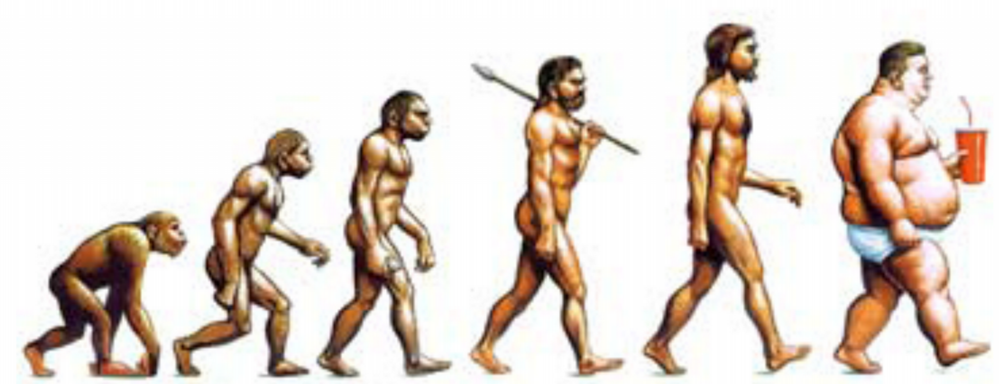
UK election world sport football opinion culture business lifestyle fashion environment tech travel all

home › science

Medical research

Bowel cancer risk may be reduced by rural African diet, study finds

Tests on subjects who swapped a fatty, meat-heavy diet for foods rich in beans and vegetables found a drop in biological markers for cancer in just two weeks



Diet swap – Two weeks ?

nature COMMUNICATIONS

Home | About the journal | Authors and referees | Browse archive | Search

nature.com ▸ journal home ▸ current month ▸ abstract

ARTICLE PREVIEW
view full access options ▸

NATURE COMMUNICATIONS | ARTICLE

Fat, fibre and cancer risk in African Americans and rural Africans

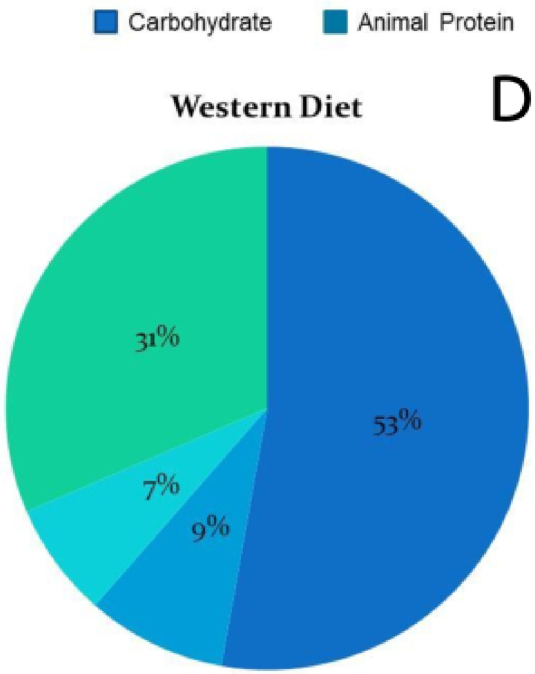
Stephen J. D. O'Keefe, Jia V. Li, Leo Lahti, Junhai Ou, Franck Carbonero, Khaled Mohammed, Joram M. Posma, James Kinross, Elaine Wahl, Elizabeth Ruder, Kishore Vipperla, Vasudevan Naidoo, Lungile Mtshali, Sebastian Tims, Philippe G. B. Puylaert, James DeLany, Alyssa Krasinskas, Ann C. Benefiel, Hatem O. Kaseb, Keith Newton + *et al.*

Affiliations | Contributions | Corresponding author

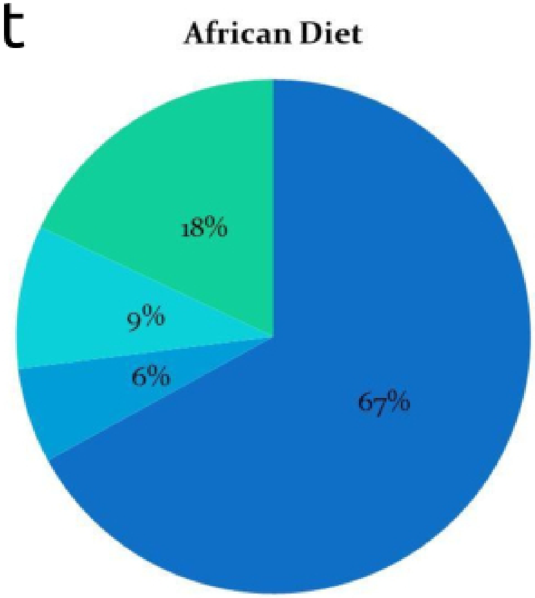
Nature Communications **6**, Article number: 6342 | doi:10.1038/ncomms7342
Received 23 May 2014 | Accepted 20 January 2015 | Published 28 April 2015

Diet

West



19 g fiber/ 2500 kcal



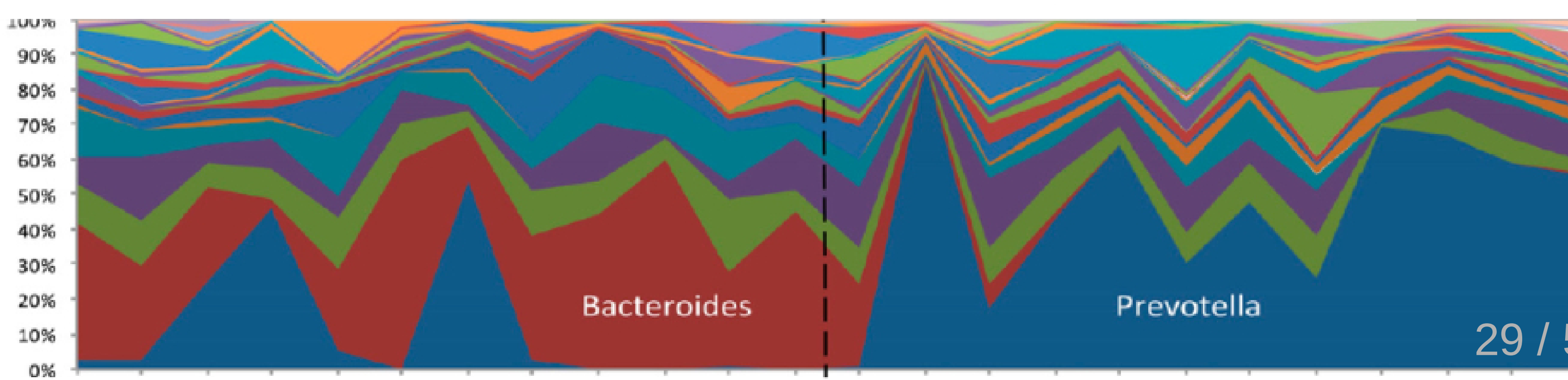
51 g fiber/ 2500 kcal

Africa

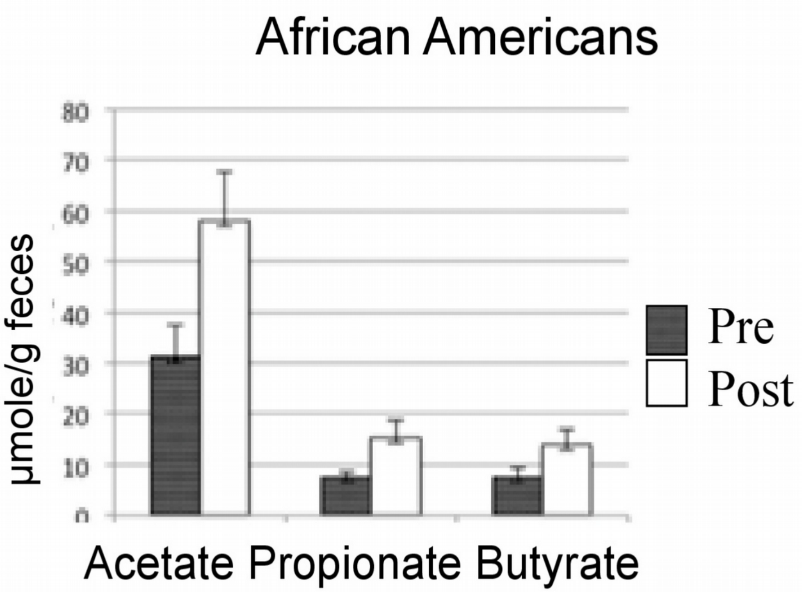
Microbiota composition

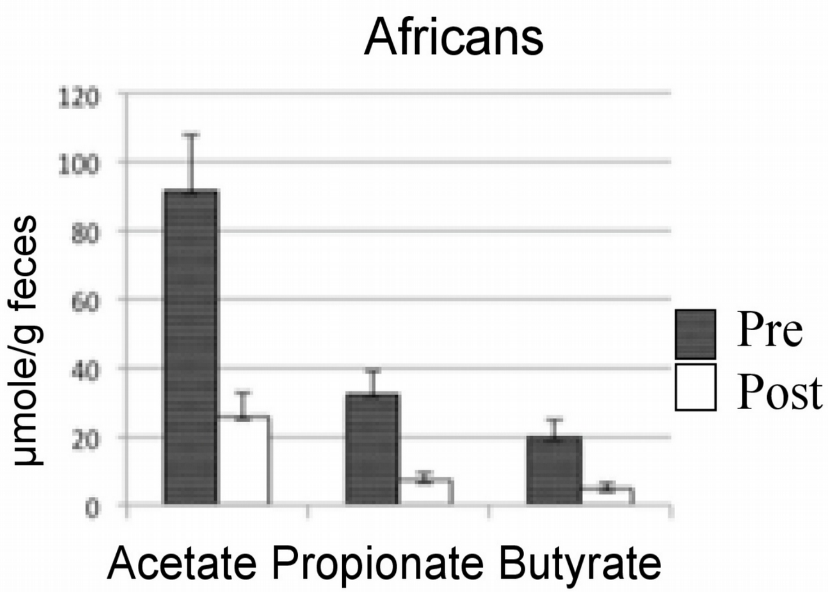
Colon cancer rates:
-Africans:<10:100,000
-African Americans: >65:100,000

Ou et al. Am J Clin Nutr. Jul 2013; 98(1): 111–120



Impact diet exchange on SCFA



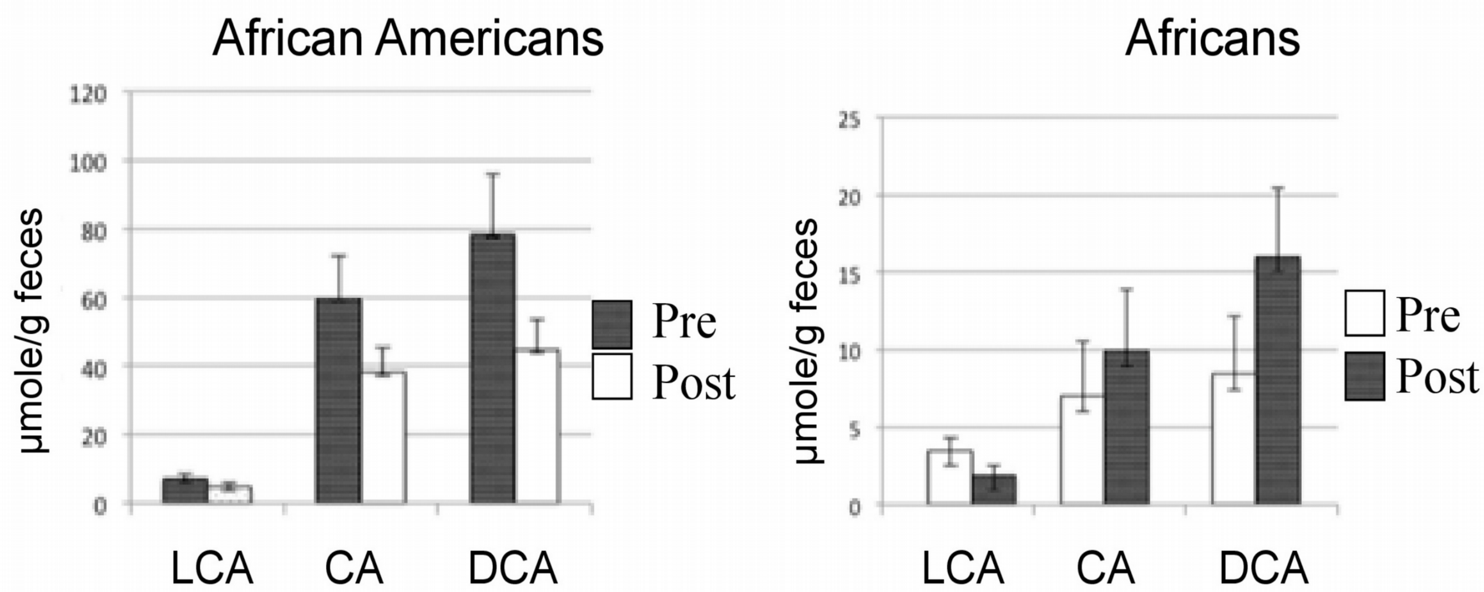


Reciprocally increased SCFA production with African diet: reported anti-inflammatory and anti-carcinogenic properties !

See also
Louis et al. Nat. Rev. Microbiol Sept 2014

O'Keefe et al. Nat. Comm. 6:6342, 2015

Impact diet exchange on bile acids



LCA: Lithocholic Acid, CA: Cholic Acid, DCA: Deoxycholic Acid

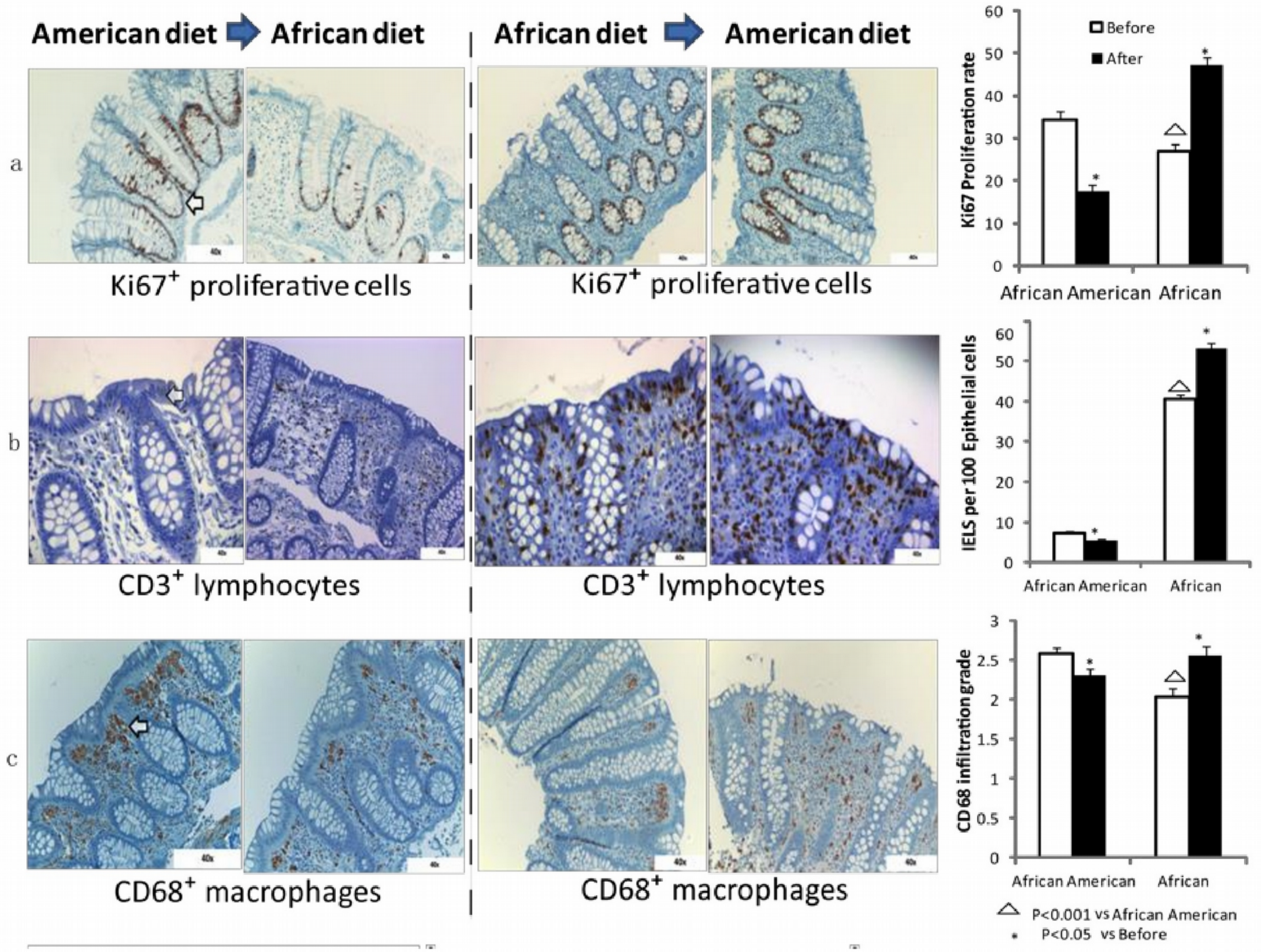
Primary and secondary bile acids reciprocally increased with American diet: reported pro-inflammatory and carcinogenic properties !

See also
Louis et al. Nat. Rev. Microbiol Sept 2014

O'Keefe et al. Nat. Comm. 6:6342, 2015

Diet swap (2 weeks) and colon cancer risk

Reciprocal impact on epithelial biomarkers for colon cancer risk seen already after two weeks !



O'Keefe et al. Nat. Comm. 6:6342, 2015

Measurements w.r.t diet swap:

- before (ED1, HE1 and HE2),
- during (DI1 & DI2)
- 2 weeks after (ED2)

Groups

- African American (AA)
- Native African (NA)

Covariates

- Overweight (bmi_group)
- Gender (sex)
- Nationality (AA / NA)

nature COMMUNICATIONS

Home | About the journal | Authors and referees | Browse archive | Search

nature.com ▸ journal home ▸ current month ▸ abstract

ARTICLE PREVIEW

view full access options ▸

NATURE COMMUNICATIONS | ARTICLE

Fat, fibre and cancer risk in African Americans and rural Africans

Stephen J. D. O'Keefe, Jia V. Li, Leo Lahti, Junhai Ou, Franck Carbonero, Khaled Mohammed, Joram M. Posma, James Kinross, Elaine Wahl, Elizabeth Ruder, Kishore Vipperla, Vasudevan Naidoo, Lungile Mtshali, Sebastian Tims, Philippe G. B. Puylaert, James DeLany, Alyssa Krasinskas, Ann C. Benefiel, Hatem O. Kaseb, Keith Newton + *et al.*

Affiliations | Contributions | Corresponding author

Nature Communications **6**, Article number: 6342 | doi:10.1038/ncomms7342
Received 23 May 2014 | Accepted 20 January 2015 | Published 28 April 2015

Stan vs traditional ML

- Bring in prior information
- Define models explicitly
- Point estimates and posteriors for model parameters available
- Express and fit complex models with millions of parameters; hierarchical and non-standard models
- Communicate uncertainty
- Fast and simple definition of complex models
- Full & Approximate Bayes; optimization (HMC, NUTS, L-BFGS)

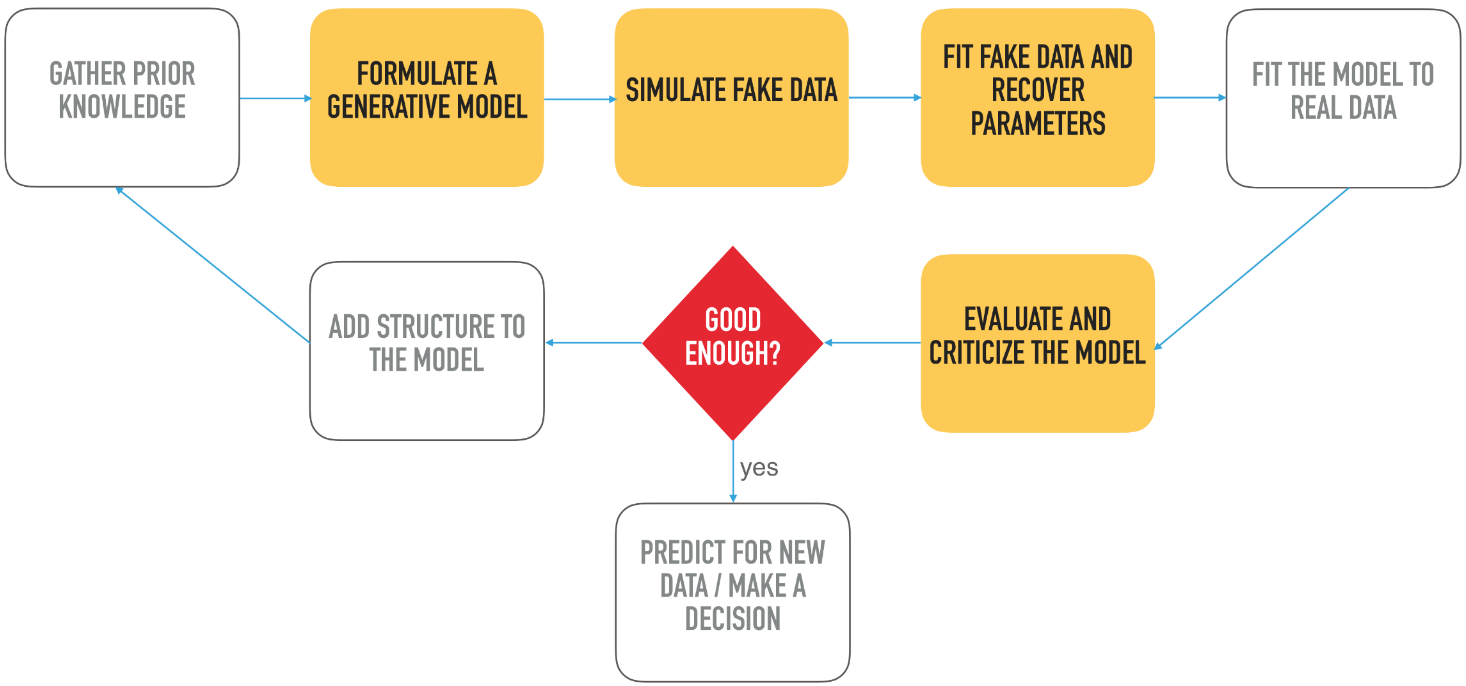
Increasingly used in science & analytics: climate models, clinical drug trials, genomics and cancer biology, population dynamics, psycholinguistics, social networks, finance and econometrics, sports, publishing, recommender systems..

Iterative Bayesian workflow

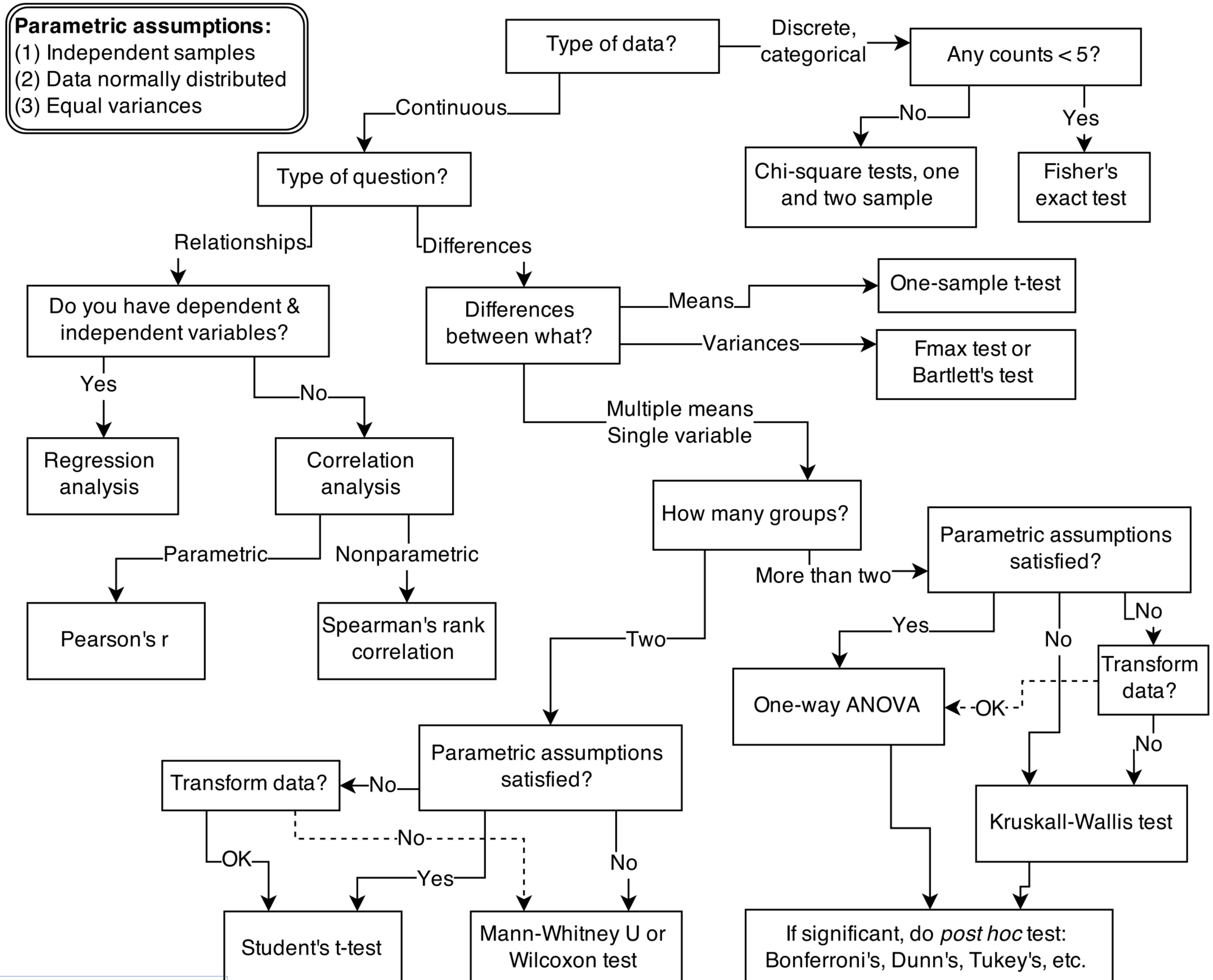
Construct the model: express the joint probability distribution of all variables (e.g. as a graph)

Incorporate data (observations)

Perform inference: learn parameters of the latent variables



Expanding conventional statistical models becomes cumbersome when complexity grows

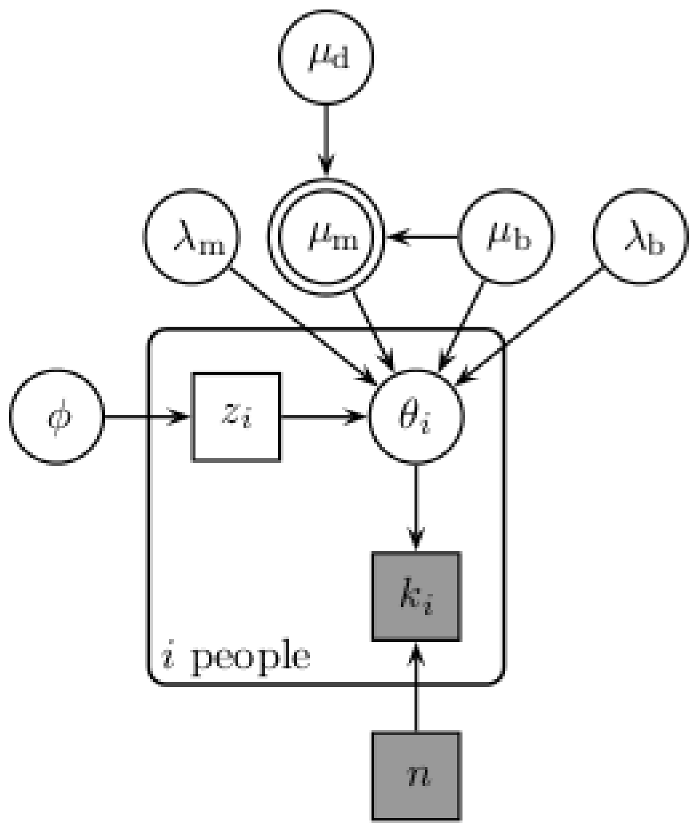


How about:
- Generalized linear models
- Hierarchical models
- Missing data
- Measurement errors
- Alternative explanations

McElreth (2015)

Expanding probabilistic models

More complex models also possible:
flexible modeling framework

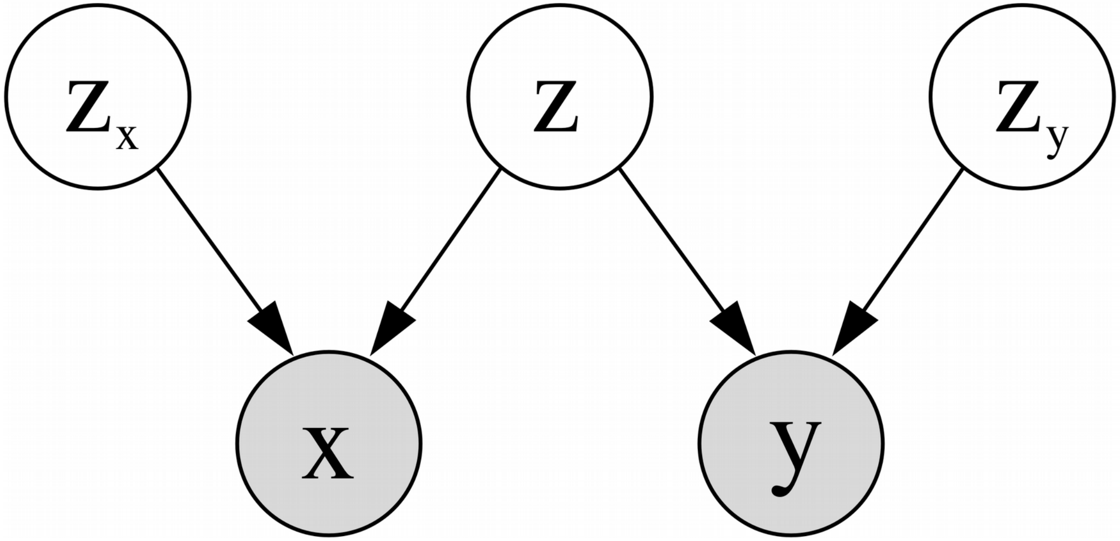


$$
\begin{aligned}
\mu_{\mathrm{b}} &\sim \mathrm{Beta}(1,1) \\
\mu_{\mathrm{d}} &\sim \mathrm{Gaussian}(0,0.5)_{\mathcal{I}(0,\infty)} \\
\lambda_{\mathrm{b}} &\sim \mathrm{Uniform}(40,800) \\
\lambda_{\mathrm{m}} &\sim \mathrm{Uniform}(4,100) \\
z_i &\sim \mathrm{Bernoulli}(\phi) \\
\theta_i &\sim \begin{cases} \mathrm{Beta}(\mu_{\mathrm{b}}\lambda_{\mathrm{b}},(1-\mu_{\mathrm{b}})\,\lambda_{\mathrm{b}}) & \text{if } z_i = 0 \\ \mathrm{Beta}(\mu_{\mathrm{m}}\lambda_{\mathrm{m}},(1-\mu_{\mathrm{m}})\,\lambda_{\mathrm{m}}) & \text{if } z_i = 1 \end{cases} \\
k_i &\sim \mathrm{Binomial}(\theta_i,n) \\
\mathrm{logit}\mu_{\mathrm{m}} &\leftarrow \mathrm{logit}\mu_{\mathrm{b}} - \mu_{\mathrm{d}} \\
\phi &\sim \mathrm{Beta}(5,5)
\end{aligned}
$$

Graphical model for inferring membership of two latent groups, consisting of malingerers and *bona fide* participants.

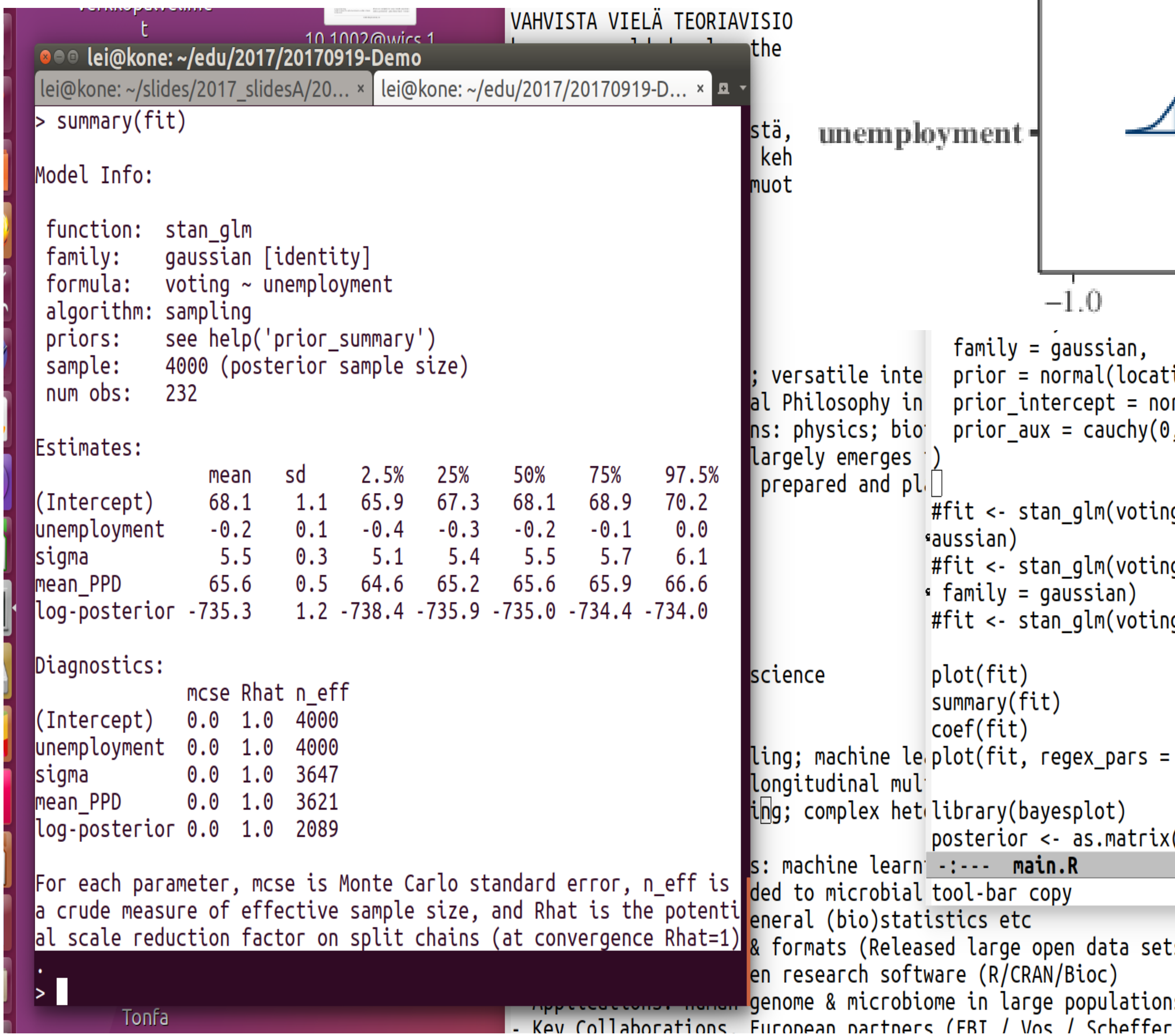
A practical example: probabilistic canonical correlation analysis

$$\begin{cases} X = W_x \mathbf{z} + \varepsilon_x \\ Y = W_y \mathbf{z} + \varepsilon_y \end{cases}$$



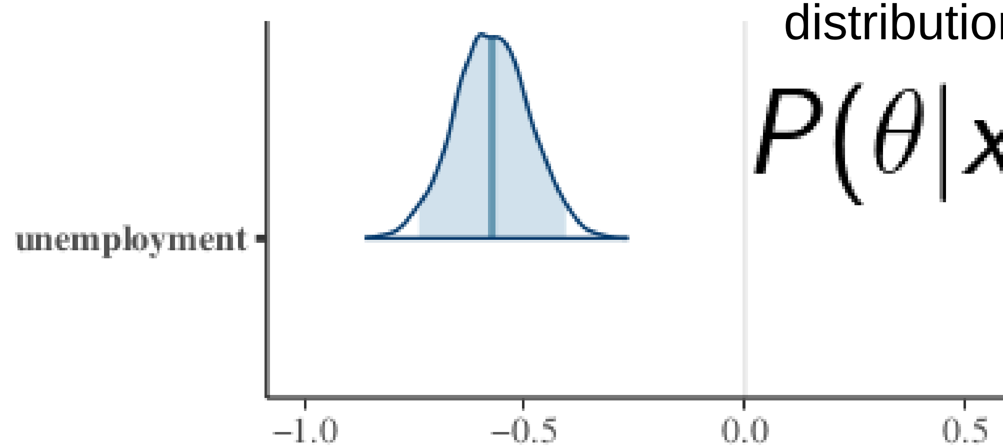
Summarizing model output

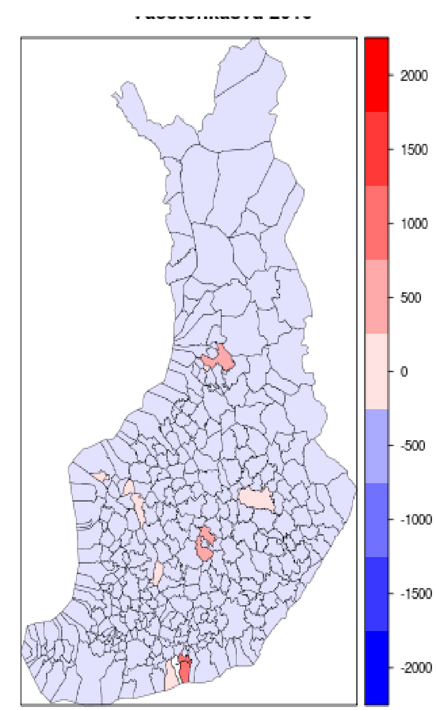
```
summary(fit)
```



Posterior distribution

$$P(\theta|x)$$

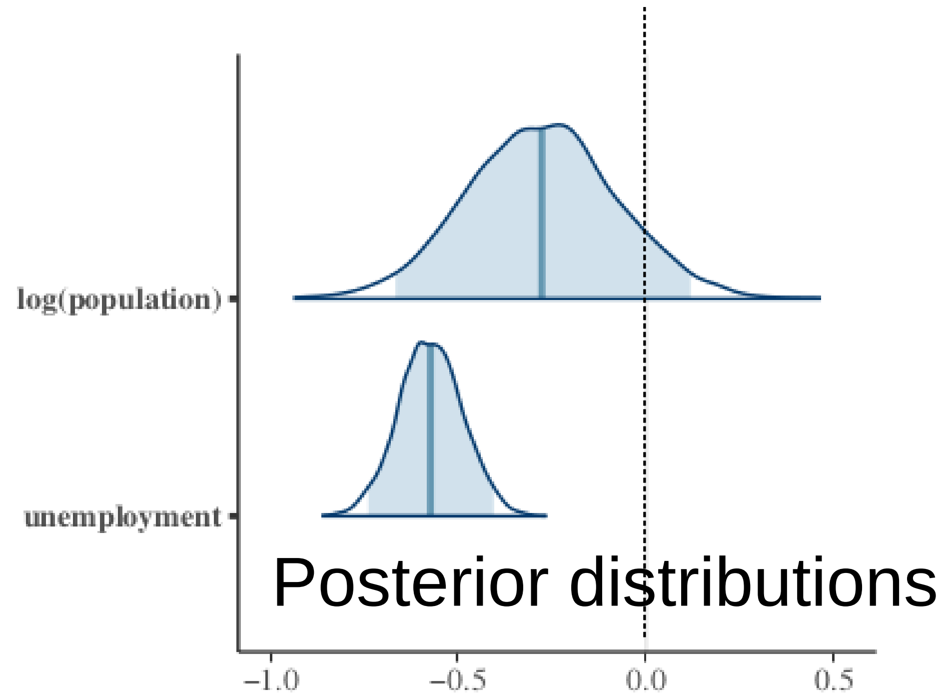


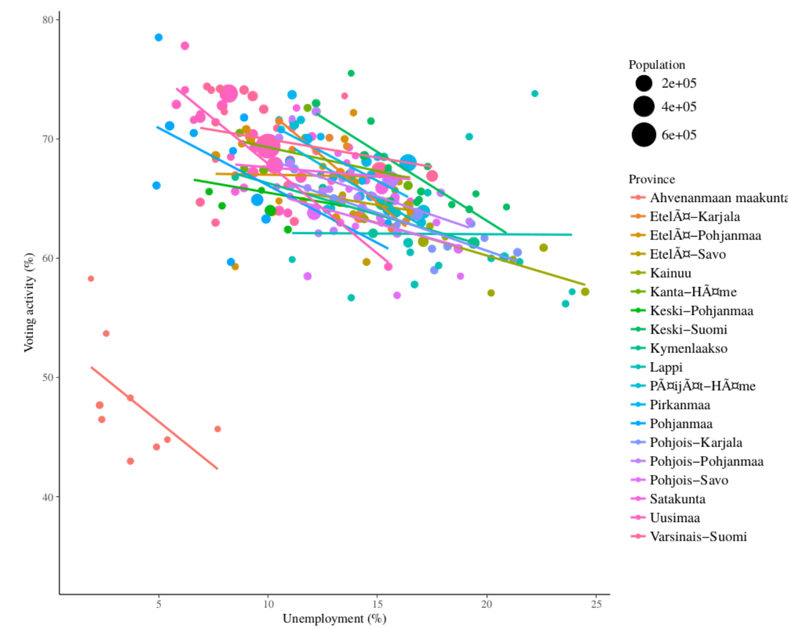


Add population size, provinces & explicit priors

```
fit <- stan_glm(
  // model & data
  voting ~ unemployment +
           province +
           log(population),
  data = df,
  family = gaussian,

  // priors
  prior = normal(location = 0,
                       scale = 10),
  prior_intercept = normal(0, 10),
  prior_aux = cauchy(0, 5)
)
```

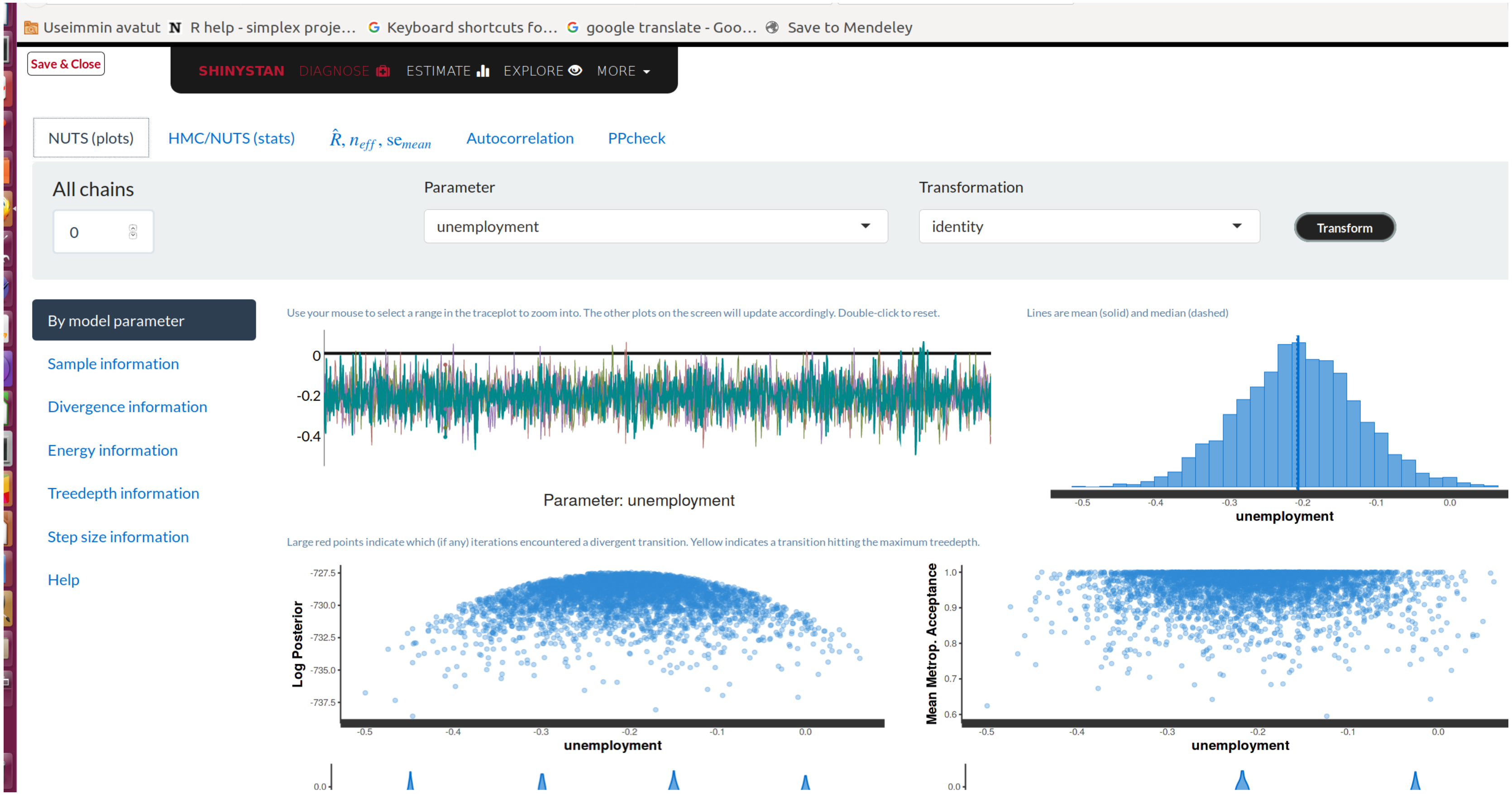




Interactive model diagnostics with ShinyStan

```
launch_shinystan(stan_model)
```

- This thing is pretty awesome.
- [Demo outside of slides.]

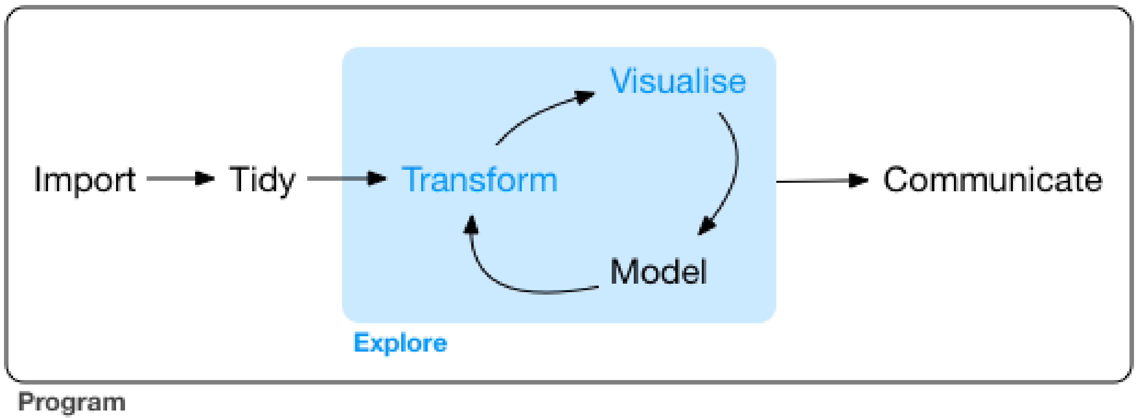


Exercise session 2
https://microbiome.github.io/microbiome/rstanarm.html

The exercises demonstrate how to:

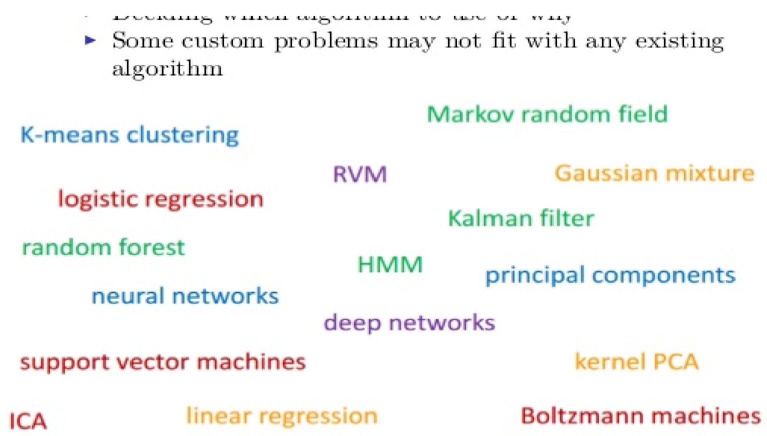
- extend the models
- diagnose,visualize, interpret and revise

We will go through examples from the rstanarm tutorial



Challenges in adopting machine learning

- Overwhelmingly many ML methods
- Difficulty in choosing and justifying the algorithms
- Custom problems may not have any existing algorithms



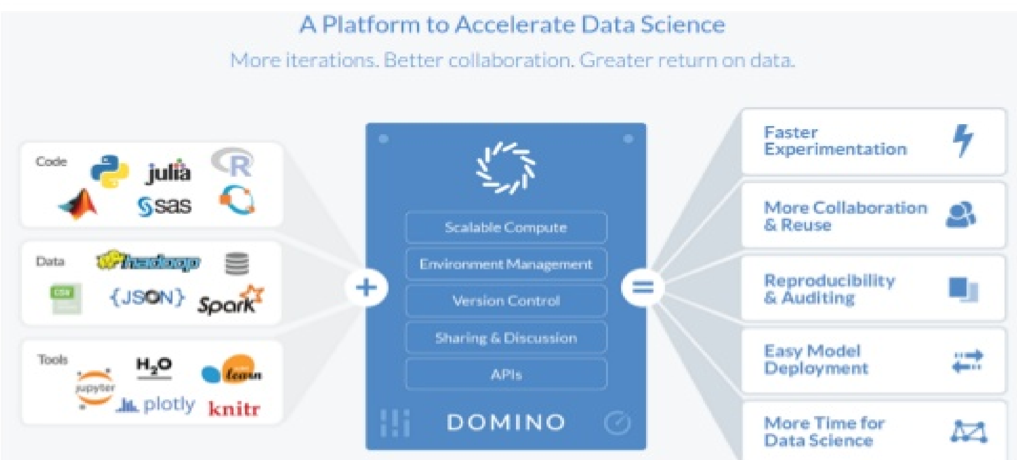


Fig: Daniel Emaasit

Solution: Model-based machine learning

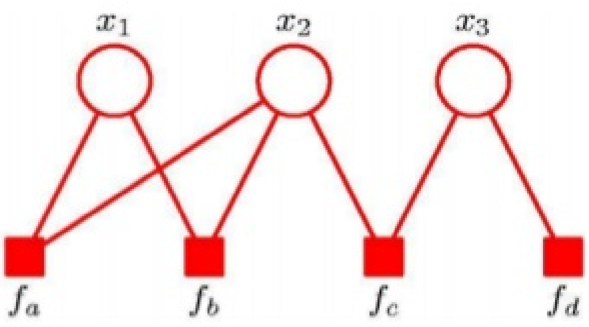
Single framework which has expressive power across a wide range of models: adoption of Bayesian ML and probabilistic graphical models.

Explicit assumptions on plausible parameter ranges, distributions, and data generating processes

Fast, approximative inference algorithms

See Bishop (2013) & Winn (2015)

- Combine probability theory with graphs (e.g Factor Graph

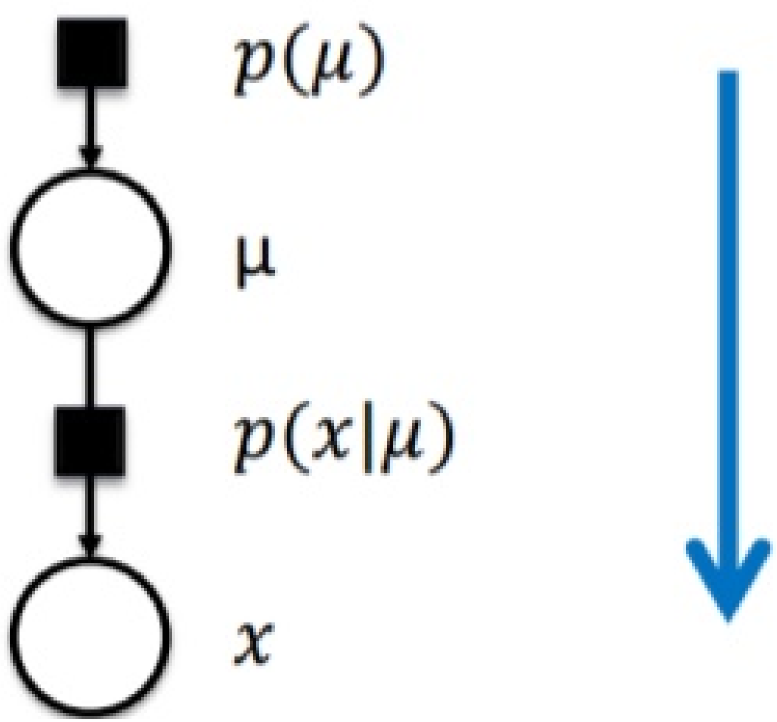


$$p(\mathbf{x}) = f_a(x_1, x_2) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_3)$$

$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s)$$

What is a Model in MBML?

- A Model:
  - is a set of assumptions, expressed in mathematical/graphical form
  - expresses all parameters, variables as random variables
  - shows the dependency between variables

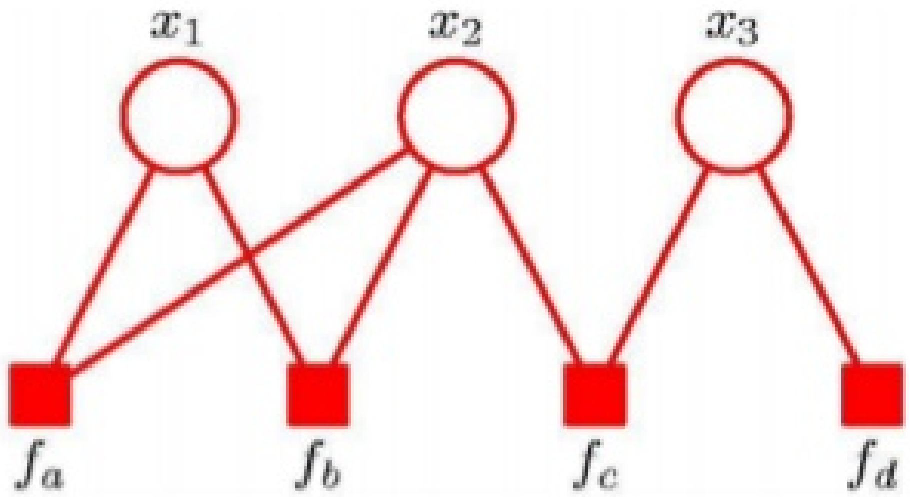


Key Ideas of MBML?

- MBML is built upon 3 key ideas
  - the use of Probabilistic Graphical Models (PGM)
  - the adoption of Bayesian ML
  - the application of fast, approximate inference algorithms

Key Idea 1: Probabilistic Graphical Models

- Combine probability theory with graphs (e.g Factor Graphs)



$$p(\mathbf{x}) = f_a(x_1, x_2) f_b(x_1, x_2) f_c(x_2, x_3) f_d(x_3)$$

$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s)$$

Key Idea 2: Bayesian Machine Learning

- Everything follows from two simple rules of probability theory

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta, m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D}, m)$	posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

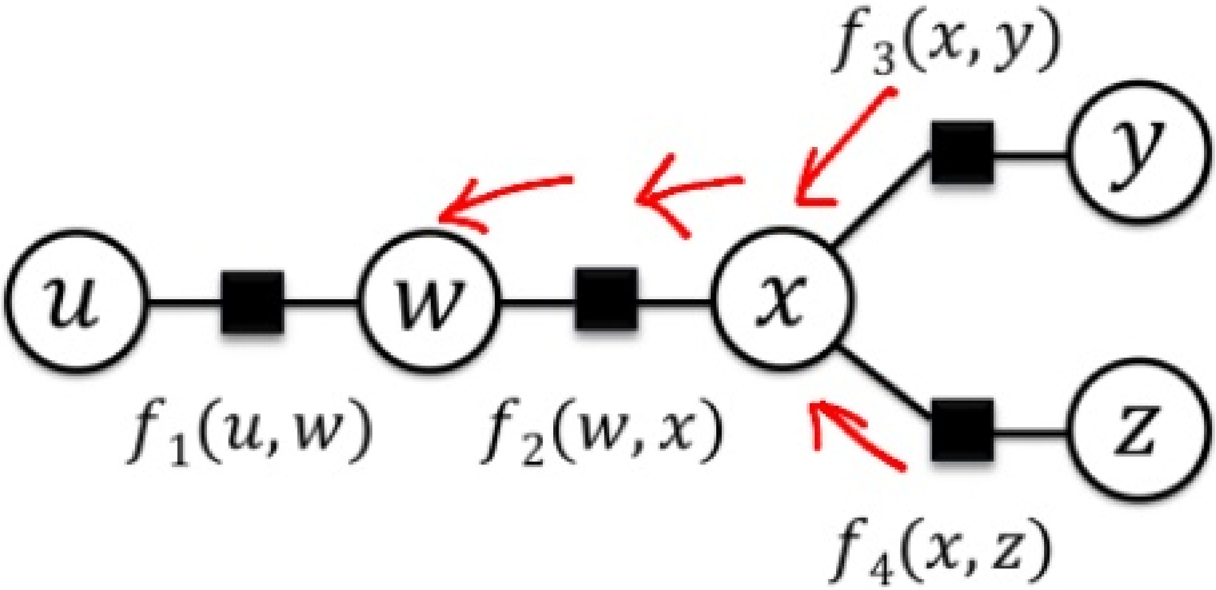
Model Comparison:

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m)\, d\theta$$

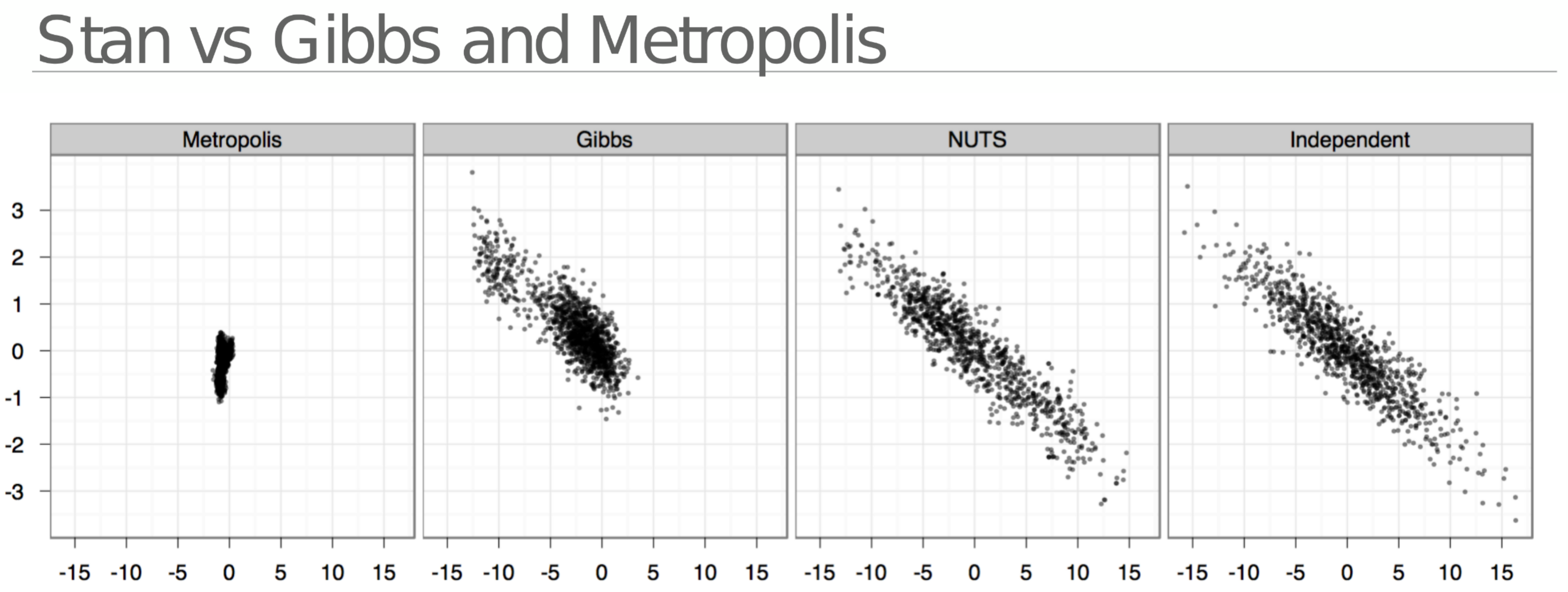
3. Fast, approximate inference algorithms

- Variational Bayes
- Expectation propagation
- Approximate Bayesian Learning (ABC)..?



Stan vs. Gibbs and Metropolis

- 2D projection of a highly correlated 250D distribution
- 1M samples from Metropolis / Gibss
- 1k samples from NUTS (in Stan!)



Bayesian approach

- Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty.
- Bayesians use probability as a language for describing uncertainty; anything which is (treated as) unknown has a probability distribution.
- The tools of probability give a coherent framework for updating beliefs with data

Probabilistic programming

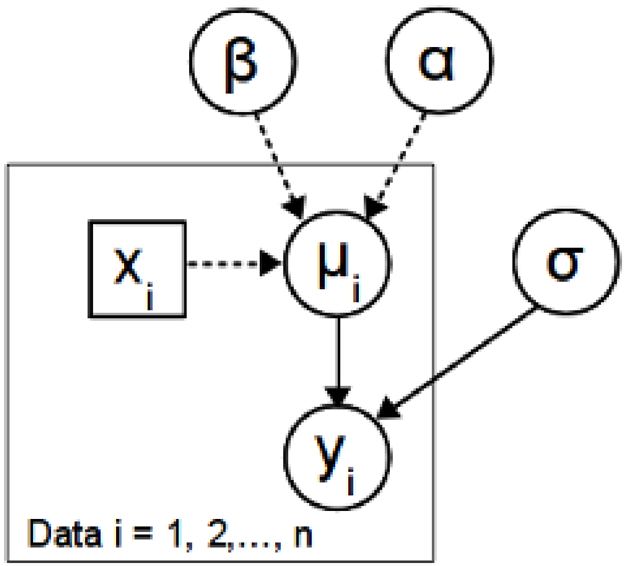
- Expands programming languages by adding support for random variables
- Software takes the model and then automatically generates inference routines (source code)
- Applicable to a wide variety of models
- Examples: Infer.net (C#, C++), Stan (R, Python, Julia, SPSS, SAS…), BUGS, Church, PyMC (Py)

Pros

- Separates models and inference. Implementing probabilistic models has taken a huge leap forward with probabilistic programming!
- Intuitive framework to fit complex non-standard models
- Rigorous way to handle uncertainties and prior information; helps with high dimensionality, small sample sizes
- Pooling of evidence as a compromise between collapsing all into a single model, or modeling all groups individually; avoid overfitting

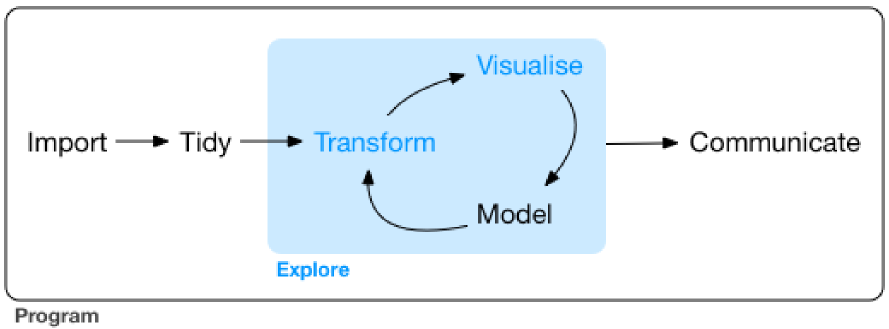
Cons

- Traditionally challenging to implement and fit
- Still more complex model specification
- Inference often slower than in traditional models
- Efficient use requires expertise
- Tools still developing



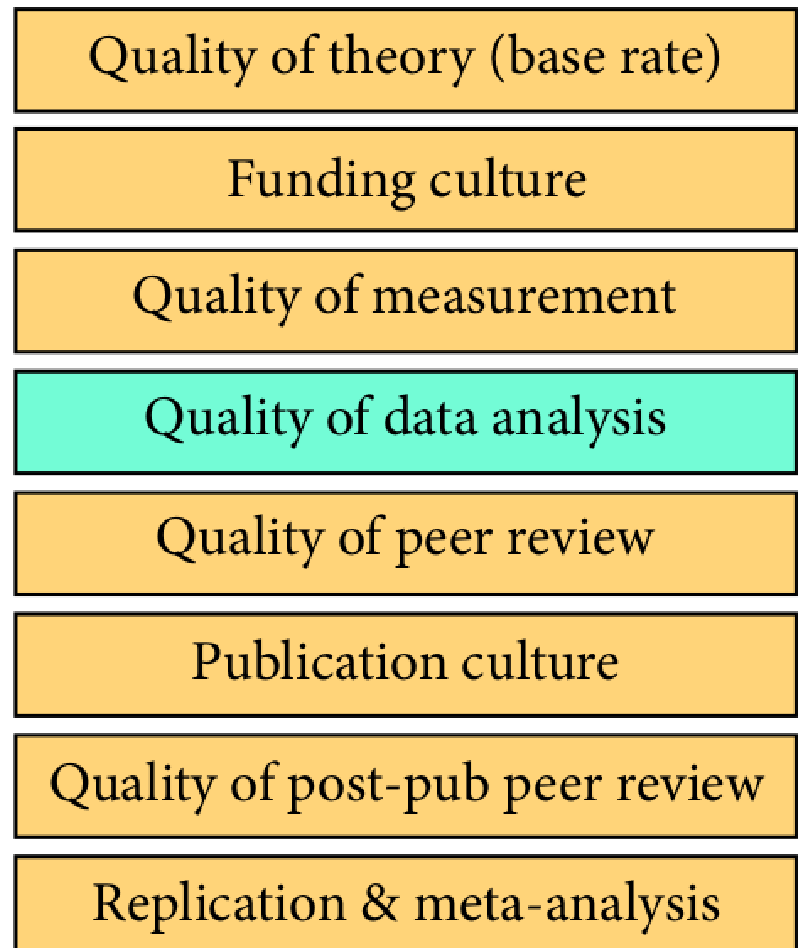
Tools still cannot replace expert knowledge

modeling choices, pitfalls, optimization, robustness, informative priors...



Further tools

- Visualization
- Simulations
- Predictive performance
- Model comparisons
- Etc.



McElreath (2015)

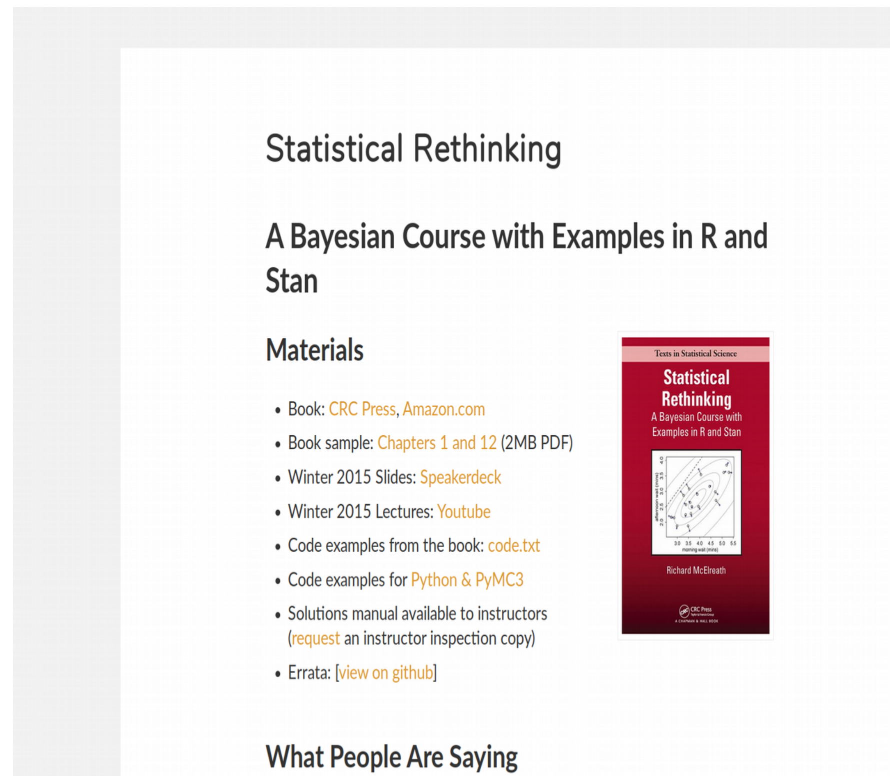
Conclusion

- Introduction to probabilistic modeling & tools
- Pros & cons in practical application

Suggested reading

check the online tutorial for slides & further resources

Contact:
http://www.iki.fi/Leo.Lahti
@antagomir
leo.lahti@iki.fi



References

- Stan manual: http://mc-stan.org/documentation/
- Google groups: http://mc-stan.org/community/
- R-package documentation: https://cran.r-project.org/web/packages/rstan/index.html
- Bishop (2013). Model-based machine learning. Philosophical transactions of the royal society A, 371:1-17.
- Winn, Bishop & Diethe (2015). Model-based machine learning. Microsoft Research Cambridge. www.mbmlbook.com

- Acknowledgement for multiple figures: Daniel Emaasit; I could not locate the source for all external figures; feel free to contact & will be added!

Feedback

Kindly fill in the feedback form

(linked also in the online material):

https://tinyurl.com/y7yfhhdh

Exercise session 3

https://microbiome.github.io/microbiome/rstanarm.html

The exercise:

- explore further aspects of rstanarm and rstan