# Bayesian Statistics and Hierarchical Bayesian Modeling for Psychological Science

## Lecture 04


Lei Zhang

Social, Cognitive and Affective Neuroscience Unit (SCAN-Unit)

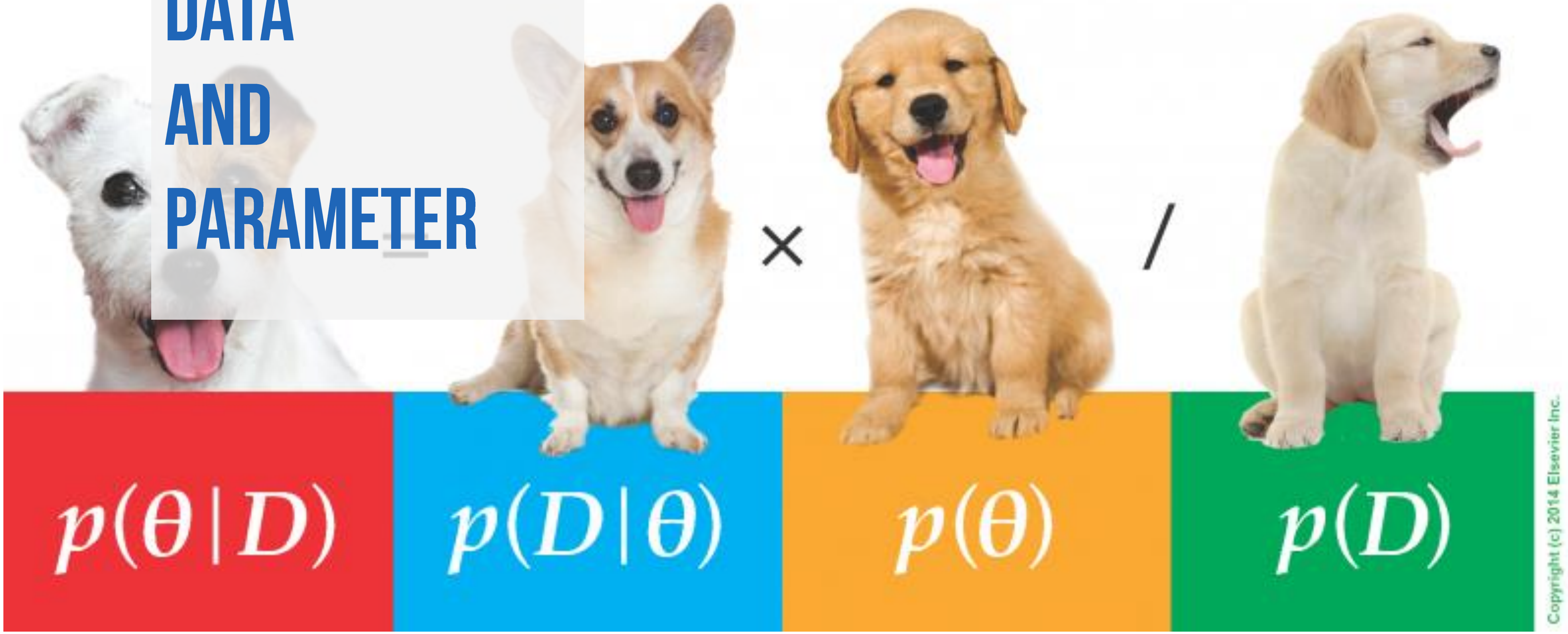
Department of Basic Psychological Research and Research Methods

lei.zhang@univie.ac.at
lei-zhang.net
@lei_zhang_lz


https://github.com/lei-zhang/BayesCog_Wien


universität wien
Fakultät für Psychologie

# LINKING DATA AND PARAMETER



# Linking Data and Parameter

$\theta$ $\qquad$ $D$

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

# Linking Data and Parameter

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

# Linking Data and Parameter



**Likelihood**
How plausible is the data given our parameter is true?

**Prior**
How plausible is our parameter before observing the data?

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

**Posterior**
How plausible is our parameter given the observed data?

**Evidence**
How plausible is the data under all possible parameters?

# What is $p(\text{Data}|\,\vartheta)$

- This is the "Model"
- Data is fixed, $\vartheta$ varies
- Not a probability distribution
  - the sum is not "one"

$$Pr(X=0\,|\,\theta)=Pr(T,T\,|\,\theta)=Pr(T\,|\,\theta)\times Pr(T\,|\,\theta)=(1-\theta)^2$$

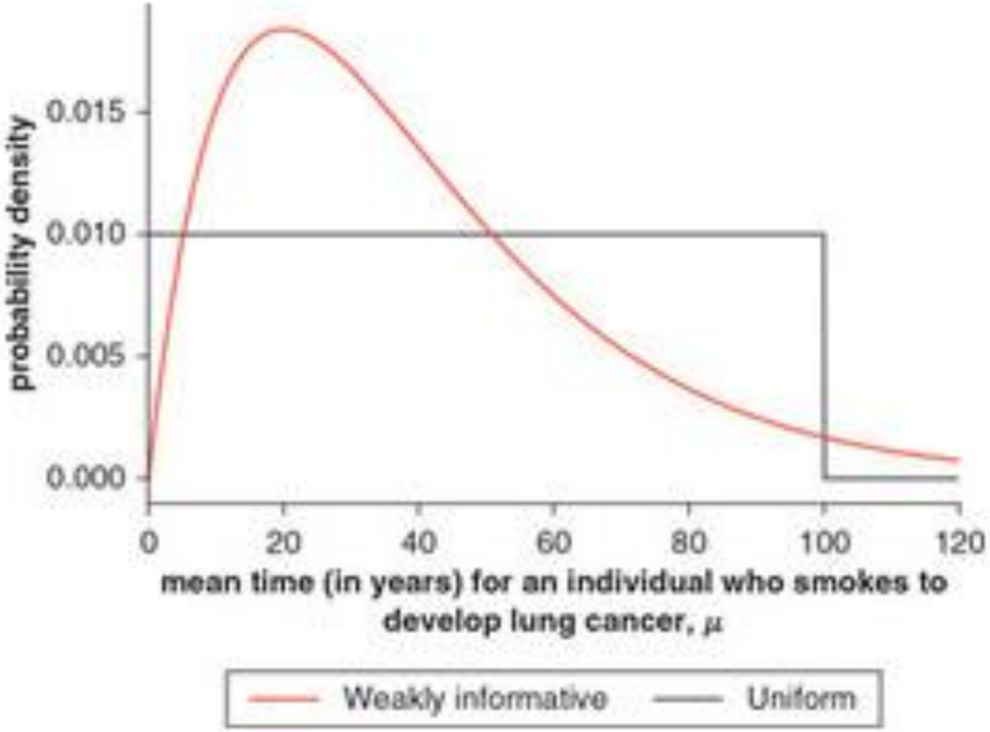
$$Pr(X=1\,|\,\theta)=Pr(H,T\,|\,\theta)+Pr(T,H\,|\,\theta)=2\times Pr(T\,|\,\theta)\times Pr(H\,|\,\theta)=2\theta(1-\theta)$$

$$Pr(X=2\,|\,\theta)=Pr(H,H\,|\,\theta)=Pr(H\,|\,\theta)\times Pr(H\,|\,\theta)=\theta^2.$$

| Probability of coin landing heads up, $\theta$ | Number of heads, $X$ 0 | 1 | 2 | Total |
|---|---|---|---|---|
| 0.0 | 1.00 | 0.00 | 0.00 | 1.00 |
| 0.2 | 0.64 | 0.32 | 0.04 | 1.00 |
| 0.4 | 0.36 | 0.48 | 0.16 | 1.00 |
| 0.6 | 0.16 | 0.48 | 0.36 | 1.00 |
| 0.8 | 0.04 | 0.32 | 0.64 | 1.00 |
| 1.0 | 0.00 | 0.00 | 1.00 | 1.00 |
| Total | 2.20 | 1.60 | 2.20 | |

Lambert (2018)

# What is $p(\vartheta)$?




Lambert (2018)

# What is *p*(*Data*)?



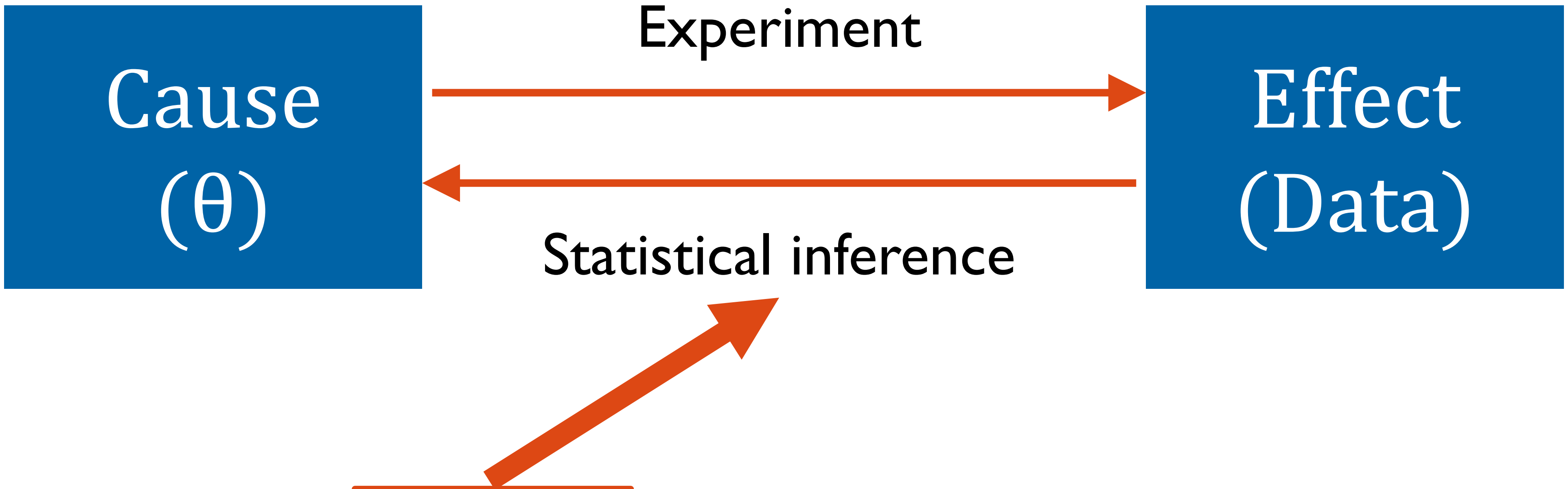
discrete parameters

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\boxed{\sum_{\theta^*} p(D \mid \theta^*)\, p(\theta^*)}}$$

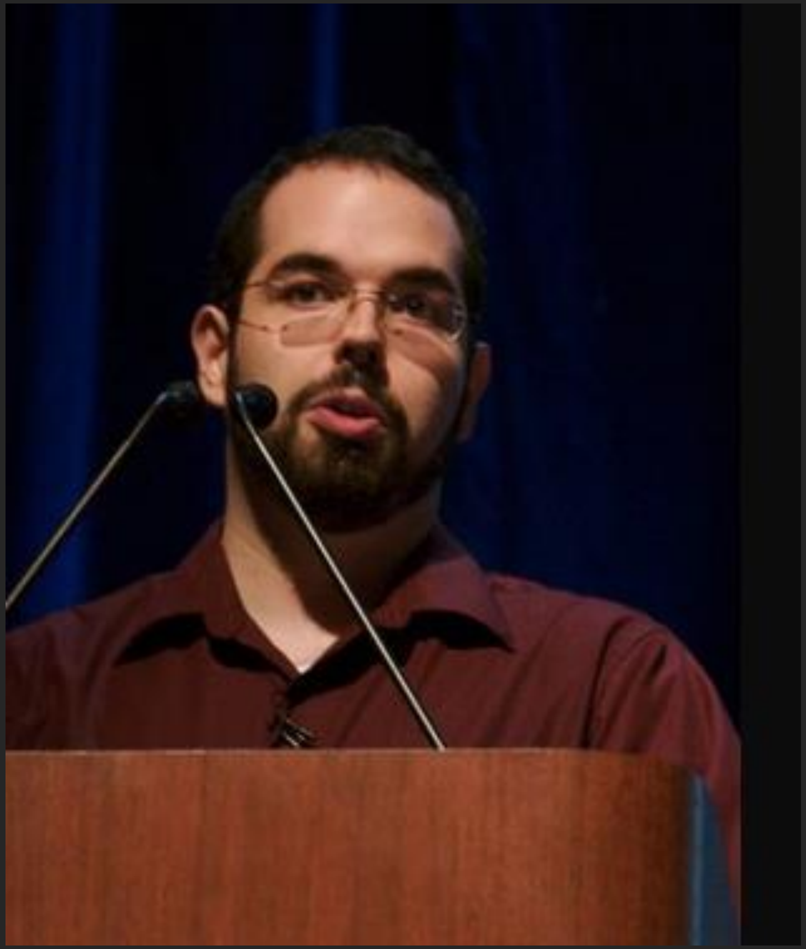
continuous parameters

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\boxed{\int p(D \mid \theta^*)\, p(\theta^*)\, d\theta^*}}$$

# Why the Bayes' theorem is important?




$$\boxed{p(\theta|D)} = \frac{p(D|\theta)p(\theta)}{p(D)}$$



*"Probability is orderly opinion and <u>inference</u> from data <u>is</u> nothing other than <u>the revision of</u> such <u>opinion</u> in the light of relevant new information."*

Eliezer S. Yudkowsky

# BINOMIAL MODEL

# Binomial Model



- You are curious how much of the surface is covered in water.
- You will toss the globe up in the air.
- You will record whether or not the surface under your right index finger is water (W) or land (L).
- You might observe: W L W W W L W L W
- → 6/9 = 0.666667?
- Is it right? If not, what to do next?

McElreath (2016)

# Steps of (Bayesian) Modeling?

**A data story**

Think about how the data might arise.
It can be *descriptive* or even *causal*.

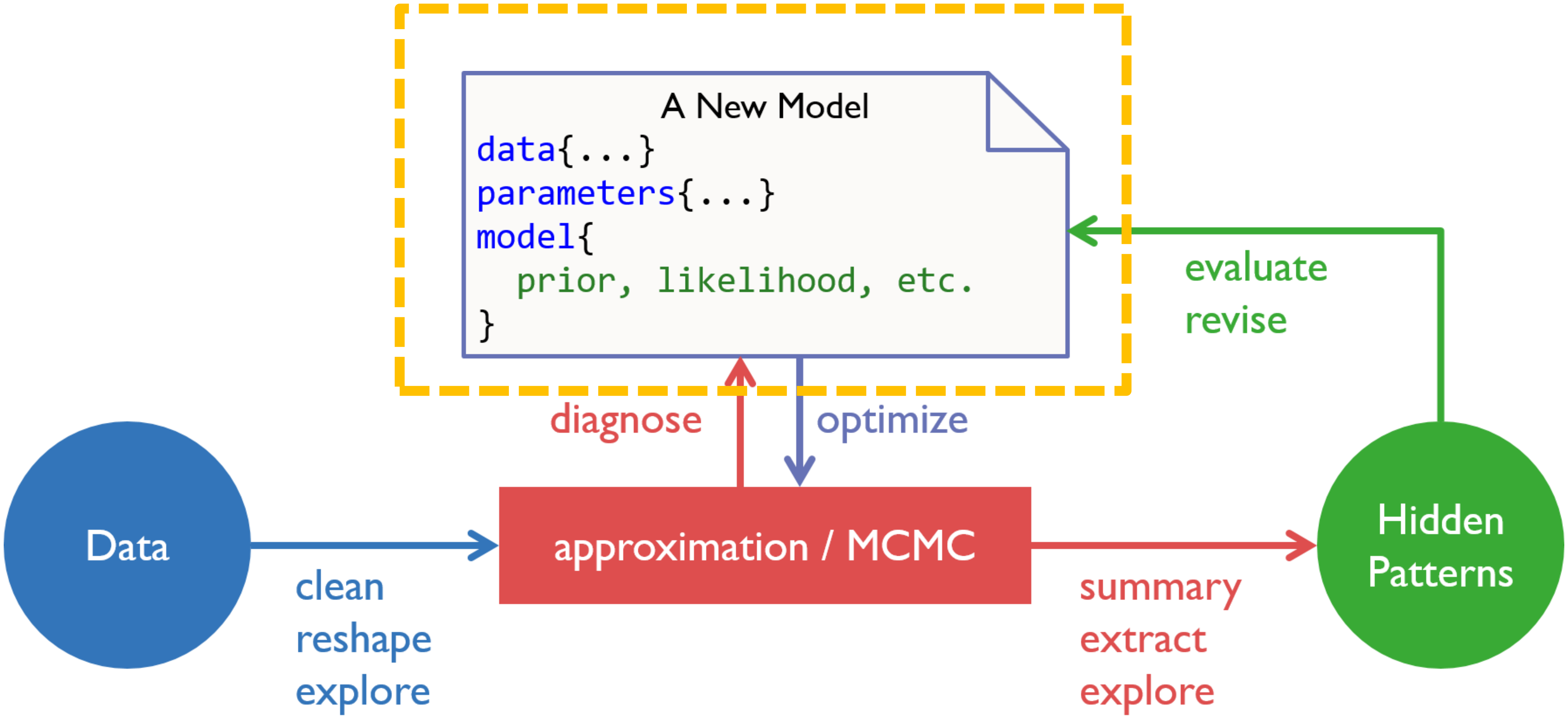
**Update**

Educate your model by feeding it with data.
Bayesian Update:
- update the prior, in light of data, to produce posterior
- the updated posterior then becomes the prior of next update

**Evaluate**

Compare model with reality.
Revise your model.

McElreath (2016)



A New Model
data{...}
parameters{...}
model{
prior, likelihood, etc.
}
evaluate
revise
diagnose
optimize
Data
approximation / MCMC
Hidden Patterns
clean
reshape
explore
summary
extract
explore

# A Data Story of the Globe

- The true proportion of water covering the globe is $p$.
- A single toss of the globe has a probability $p$ of producing a water (W) observation.
- It has a probability $(1 - p)$ of producing a land (L) observation.
- Each toss of the globe is independent of the others.

McElreath (2016)

# Components of a Model

think about the likelihood function (of Binomial):

$$p(\theta \mid D) = \frac{\boxed{p(D \mid \theta)}\, p(\theta)}{\int p(D \mid \theta^*)\, p(\theta^*)\, d\theta^*}$$

$$p(w \mid N, p) = \binom{N}{w} p^w (1-p)^{N-w}$$

$\left.\begin{array}{l} N\text{: total number of observations} \\ w\text{: number of water} \end{array}\right\}$ known (data)

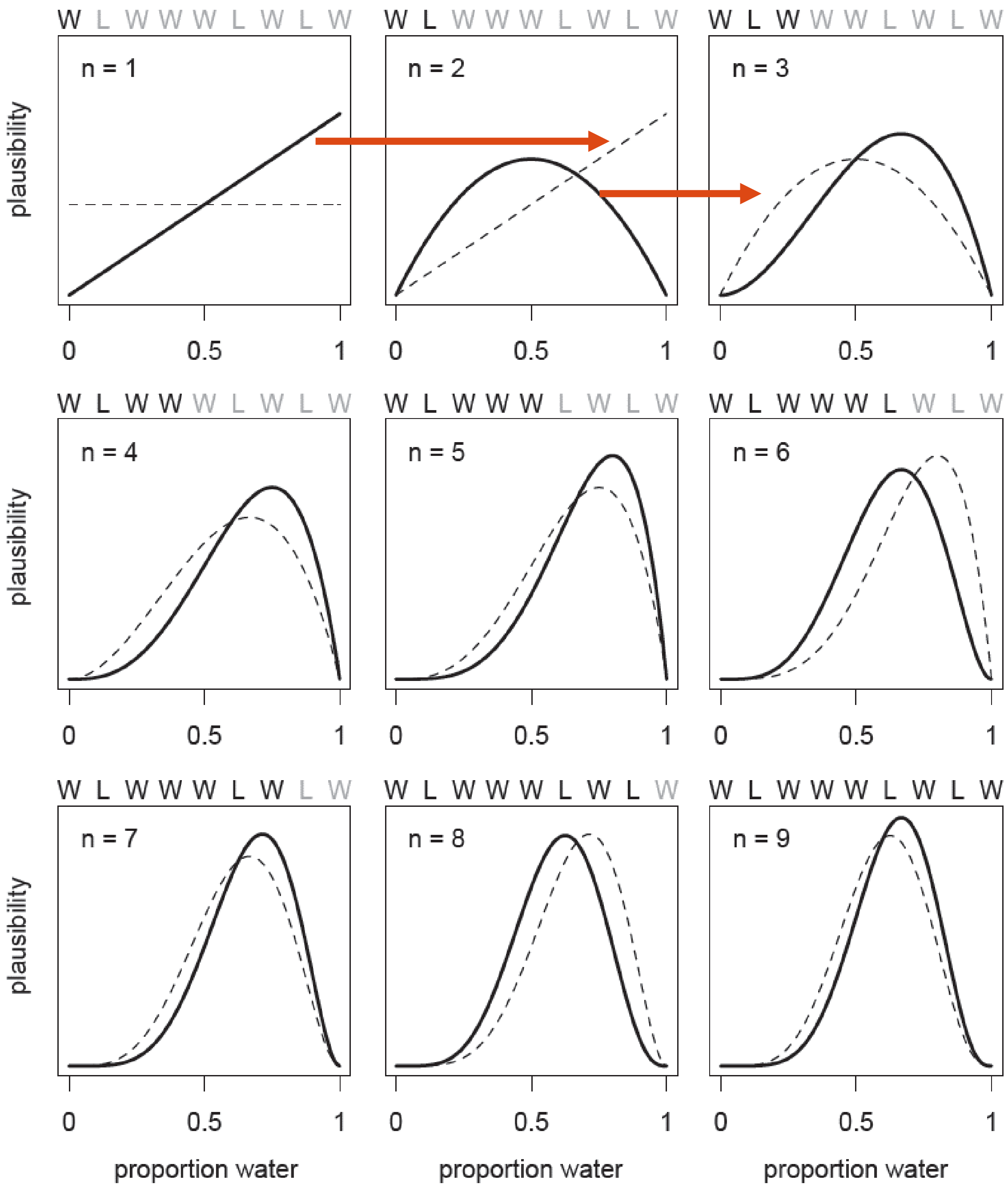
$p$: proportion of water — unknown (parameter)

# Update

cognitive model

statistics

computing




- order doesn't matter
- 2/3 is most likely
- others are not ruled out

McElreath (2016)

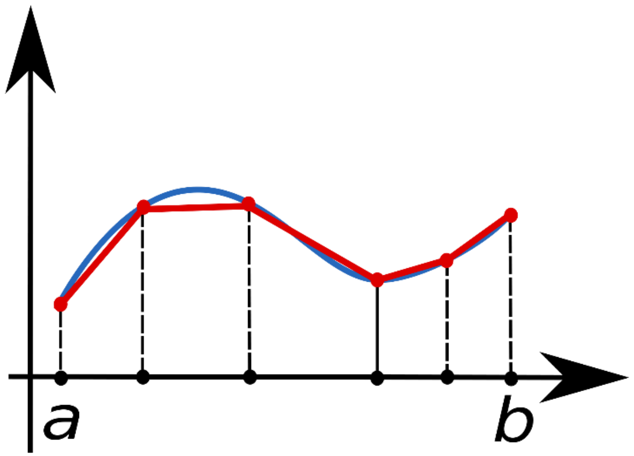
# Solve it by Grid Approximation

discrete parameters

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\boxed{\sum_{\theta^*} p(D \mid \theta^*)\, p(\theta^*)}}$$

continuous parameters

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\boxed{\int p(D \mid \theta^*)\, p(\theta^*)\, d\theta^*}}$$



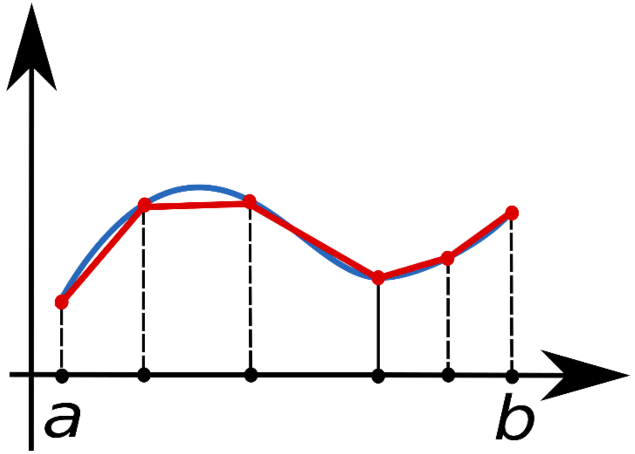
# Binomial Model – Grid Approximation



```r
p_start <- 0; p_end <- 1; n_grid <- 20
w <- 6; N <- 9

# define grid
p_grid <- seq( from = p_start ,
             to = p_end , length.out = n_grid )

# define prior
prior <- rep(1 , n_grid)

# compute likelihood at each value in grid
likelihood <- dbinom(w , size = N , prob = p_grid )

# compute product of likelihood and prior
unstd.posterior <- likelihood * prior

# standardize the posterior, so it sums to 1
posterior <- unstd.posterior / sum(unstd.posterior)
```

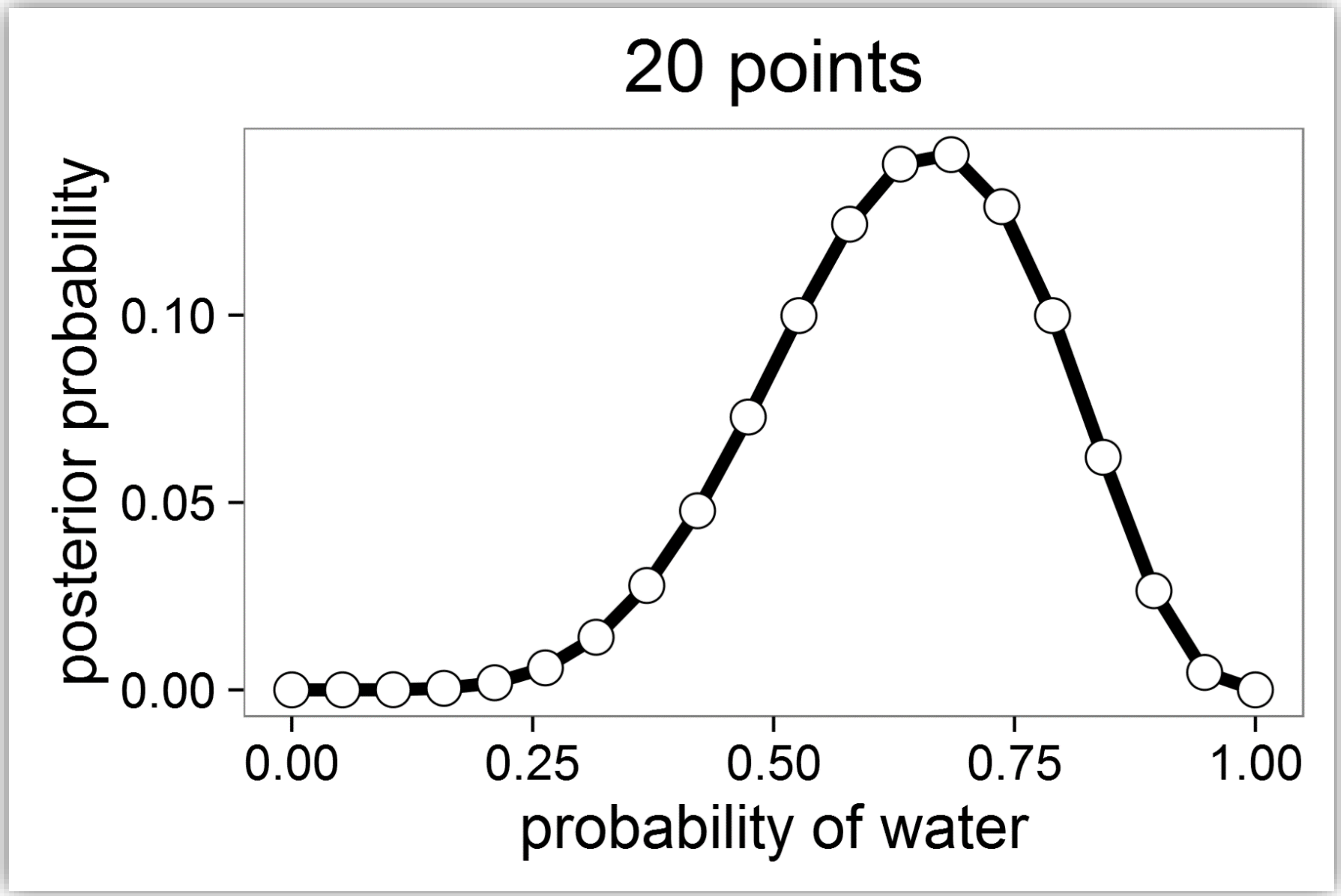
$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\int p(D \mid \theta^*)\, p(\theta^*)\, d\theta^*}$$

$$p(w \mid N, p) = \binom{N}{w} p^w (1-p)^{N-w}$$

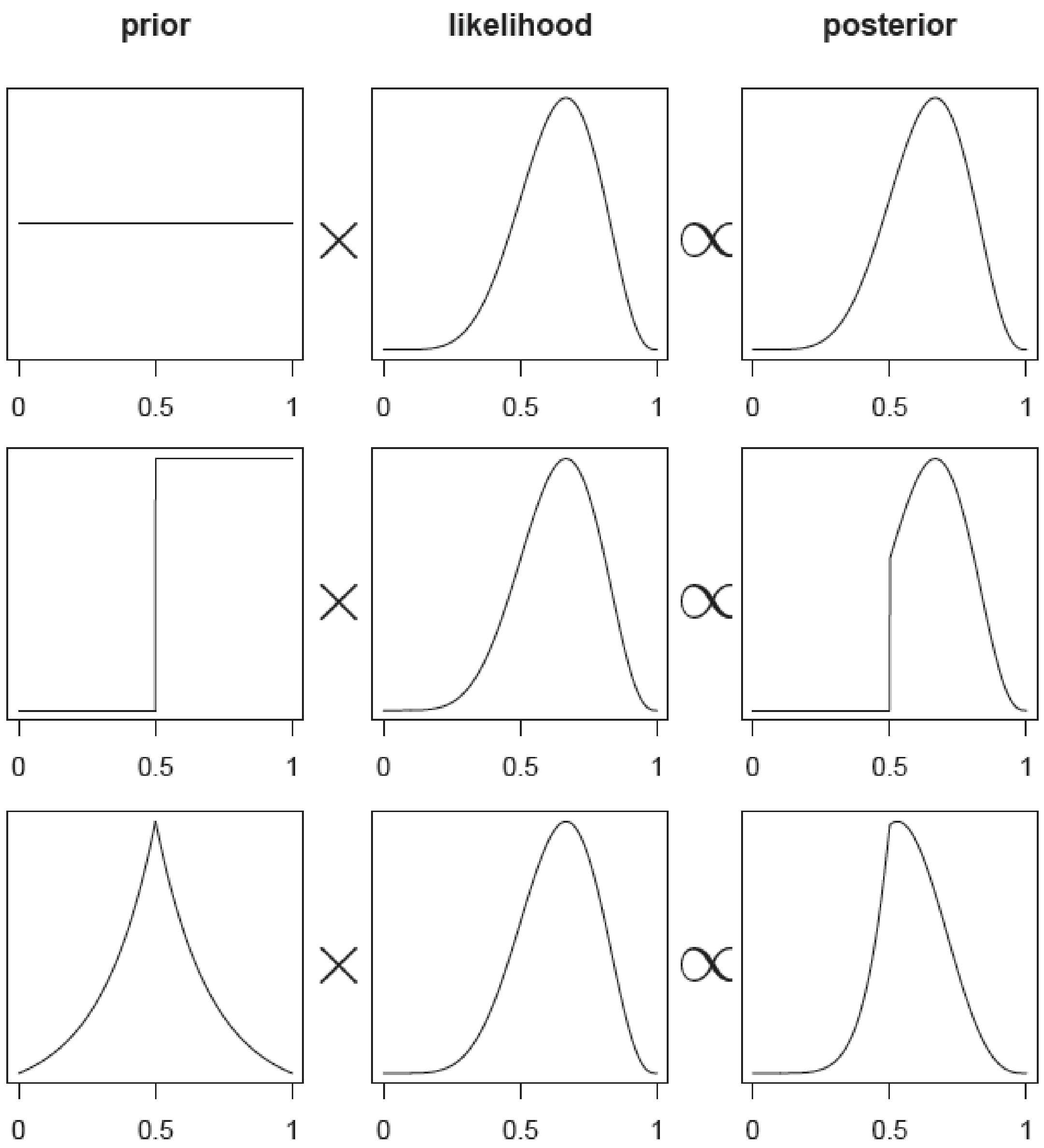


# Binomial Model – Grid Approximation

cognitive model
statistics
computing



# Impact of Prior




McElreath (2016)

# Exercise VII

```
…/BayesCog/02.binomial_globe/_scripts/binomial_globe_grid.R
```

TASK: run a grid approximation with `grid_size = 50`

# Components of a Model



grid approximation for
2 parameters?
5 parameters?
10 parameters?

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\boxed{\int p(D \mid \theta^*)\, p(\theta^*)\, d\theta^*}}$$

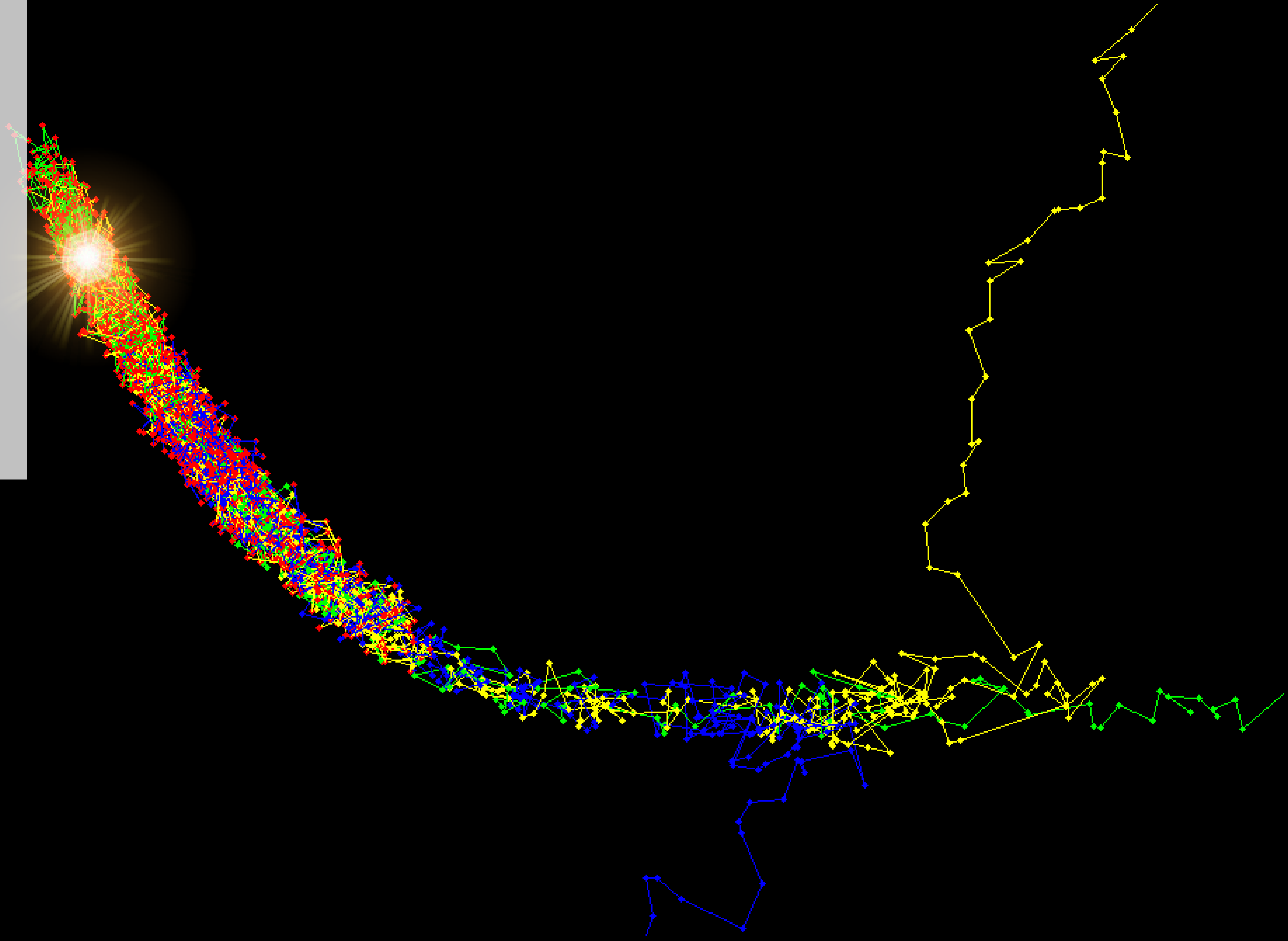
$$p(data) = \int_{\text{All}\,\theta_1} \int_{\text{All}\,\theta_2} p(data, \theta_1, \theta_2)\, \mathrm{d}\theta_1 \mathrm{d}\theta_2$$

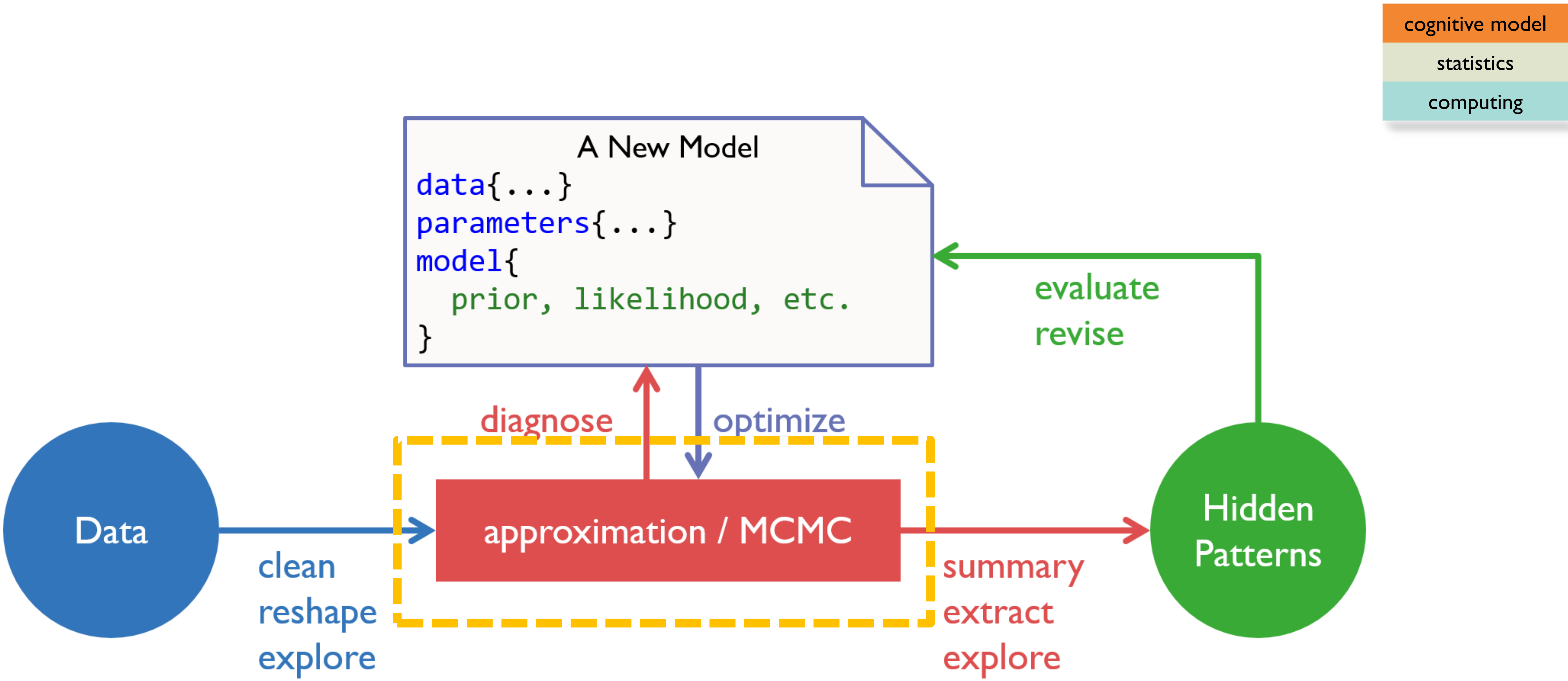
$$p(data) = \int_{\mu_1} \int_{\sigma_1} \cdots \int_{\mu_{100}} \int_{\sigma_{100}} \underbrace{p(data \mid \mu_1, \sigma_1, \ldots, \mu_{100}, \sigma_{100})}_{\text{likelihood}} \times \underbrace{p(\mu_1, \sigma_1, \ldots, \mu_{100}, \sigma_{100})}_{\text{prior}}$$

$$\mathrm{d}\mu_1 \mathrm{d}\sigma_1 \ldots \mathrm{d}\mu_{100} \mathrm{d}\sigma_{100},$$

$$\boxed{p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)}$$

# MARKOV CHAIN MONTE CARLO




cognitive model
statistics
computing
A New Model
data{...}
parameters{...}
model{
prior, likelihood, etc.
}
evaluate
revise
diagnose
optimize
Data
clean
reshape
explore
approximation / MCMC
summary
extract
explore
Hidden Patterns

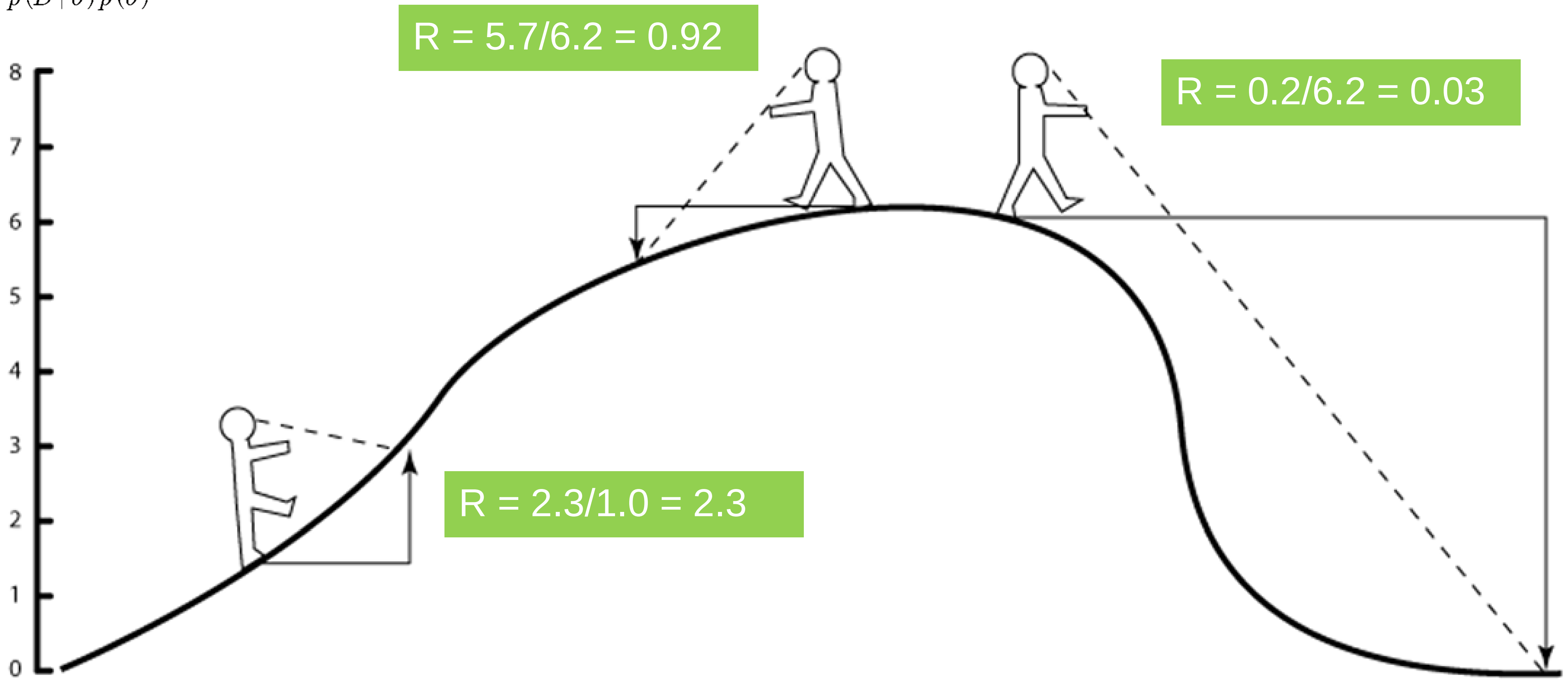
# Solving the Problem by Approximation

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$$

| Deterministic Approximation | Stochastic Approximation |
| --- | --- |
| → Variational Bayes | → Sampling Methods |

# An MCMC Robot

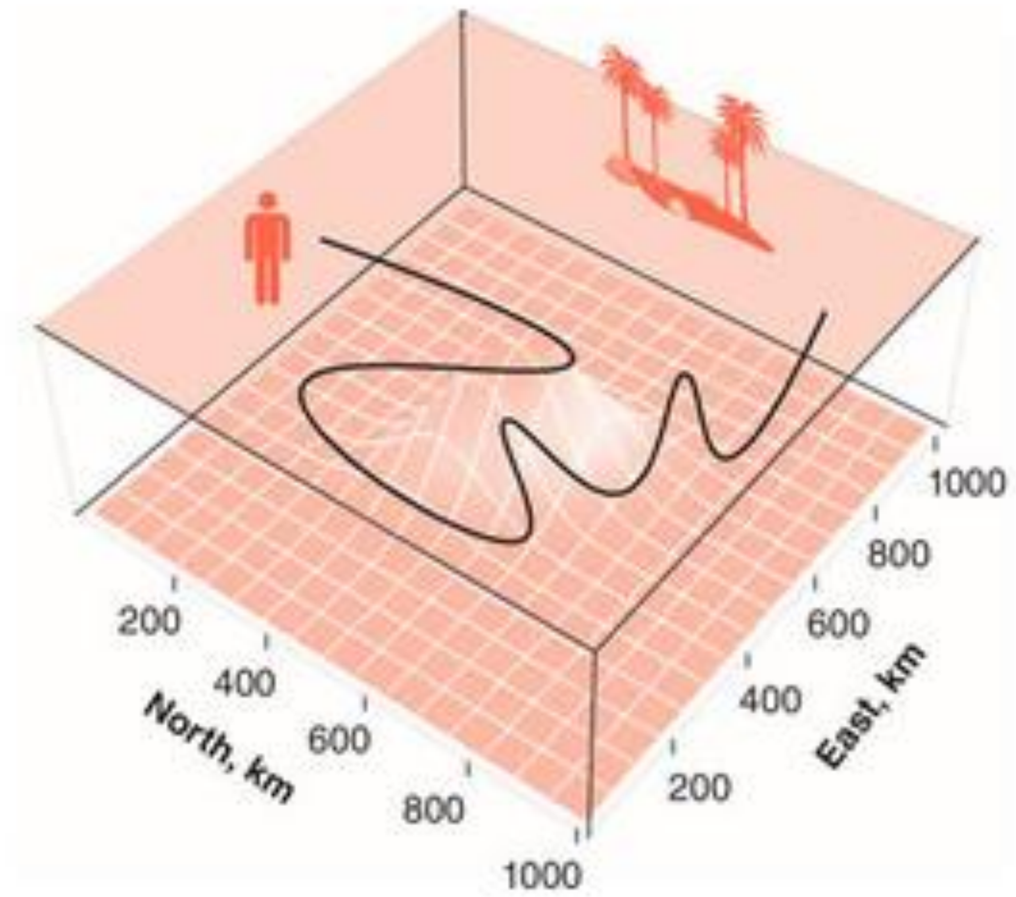
$p(\theta \mid D) \propto p(D \mid \theta) p(\theta)$




http://marple.eeb.uconn.edu/mcmcrobot/

# An MCMC Robert in 3D

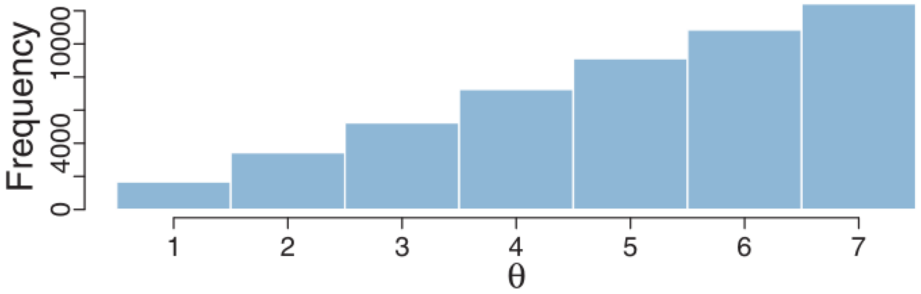





Lambert (2018)

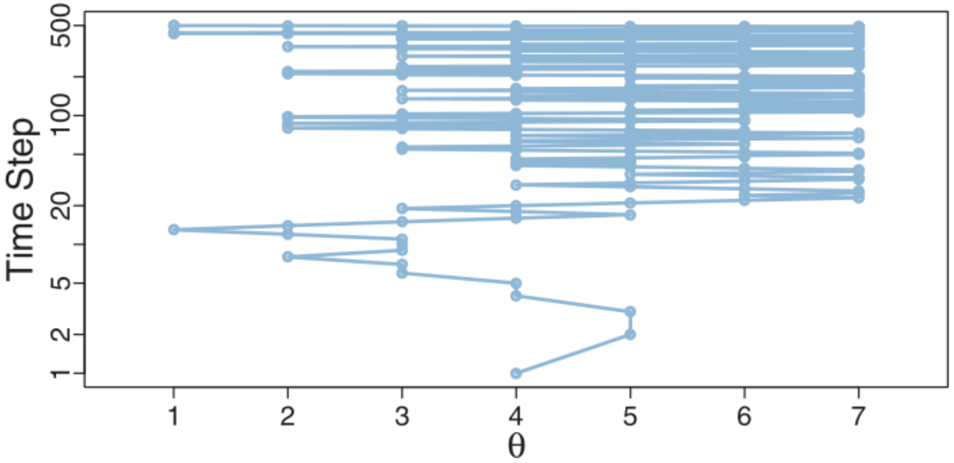
# Sampling Example





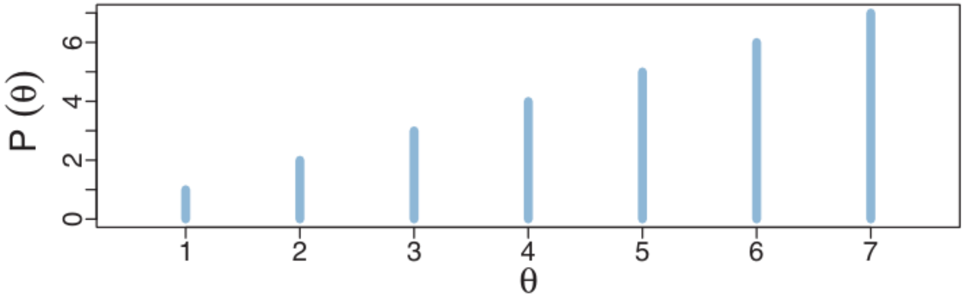

MCMC summary

⇑




MCMC trace

⇑

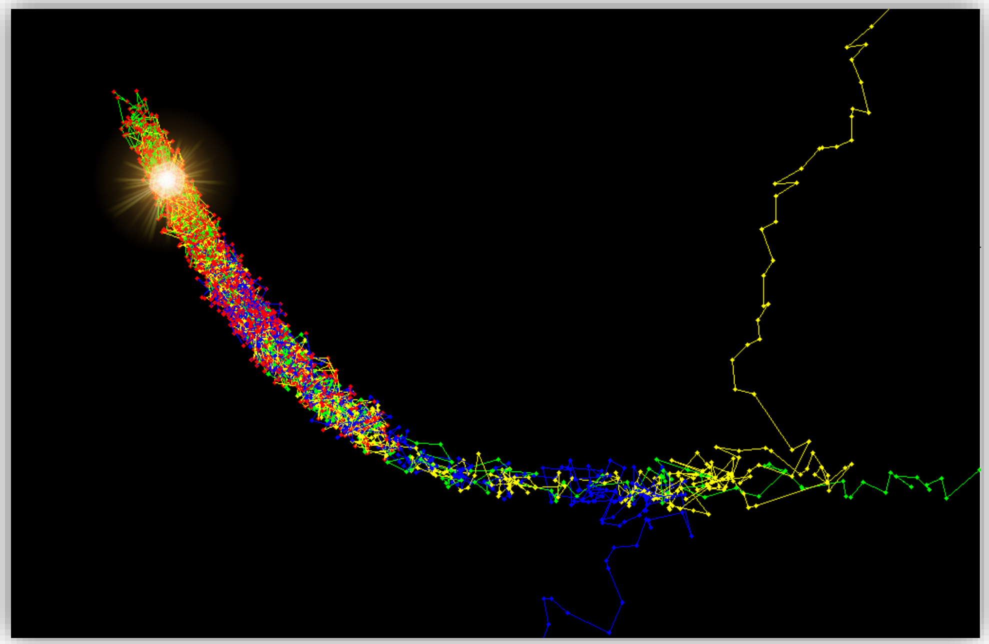


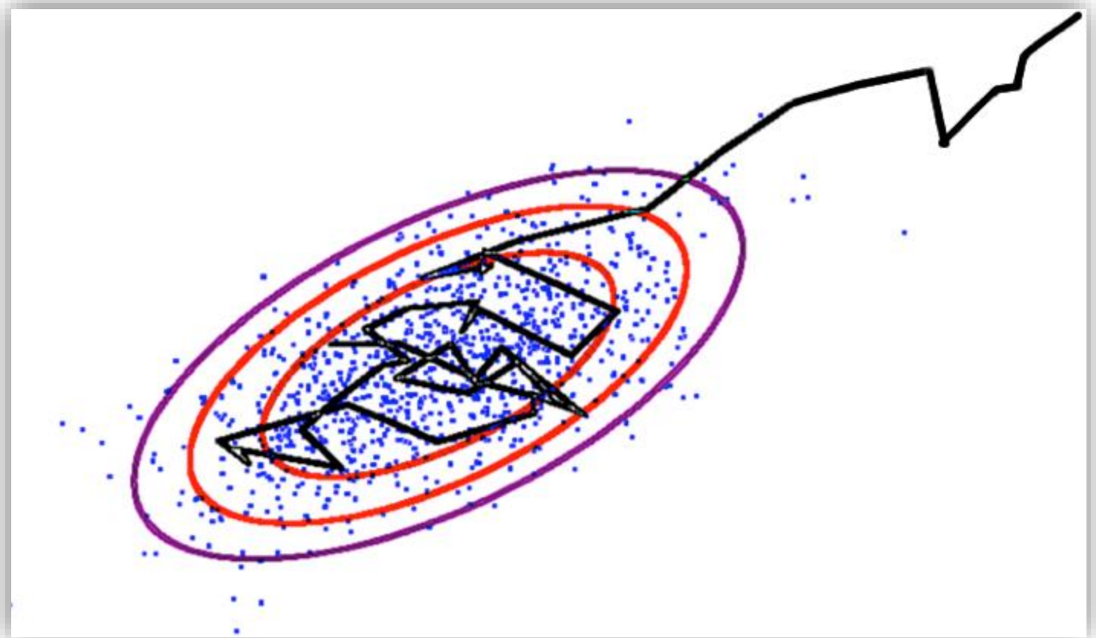

True distribution

Kruschke (2015)

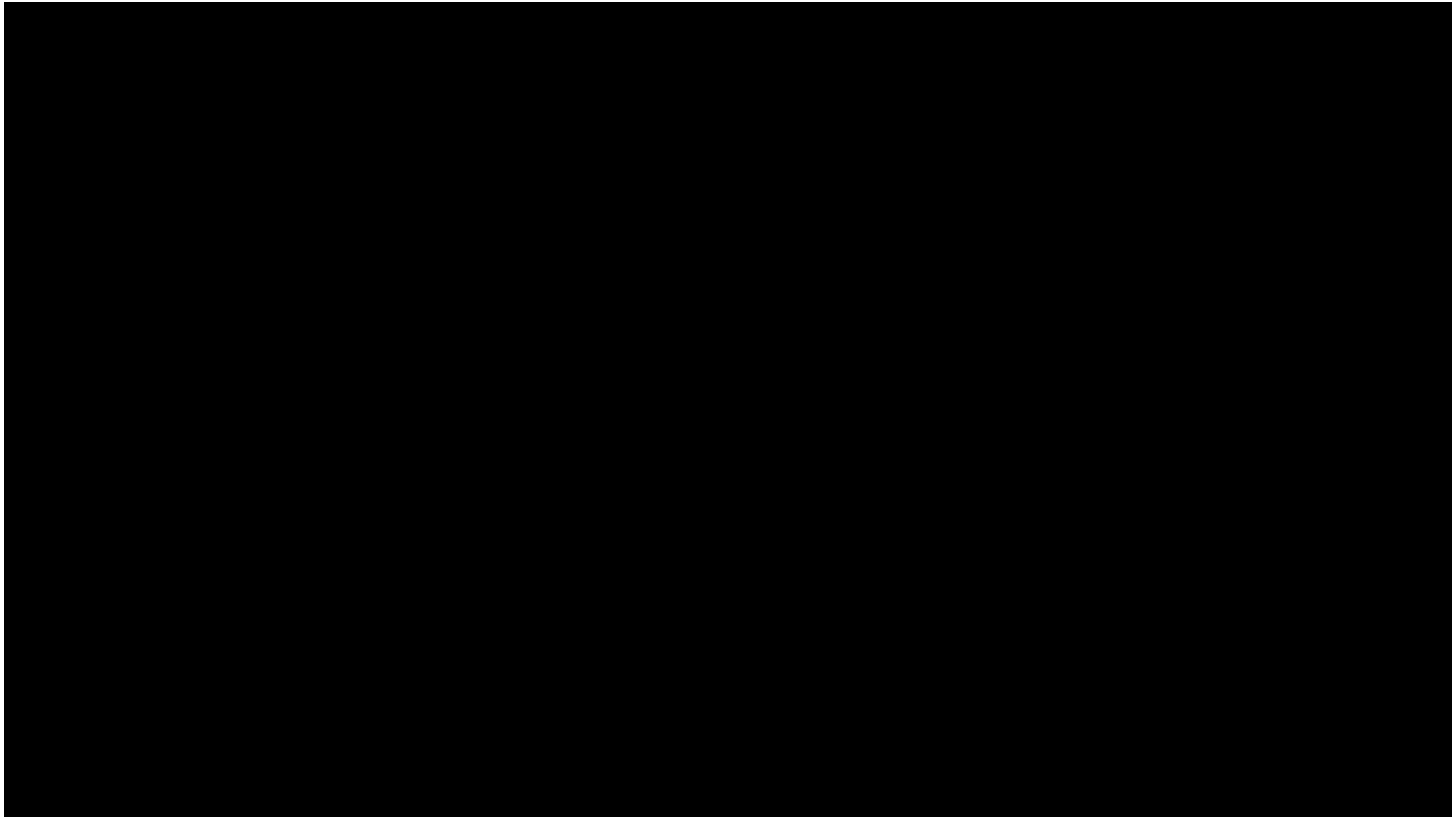
# Visual Example







# Let's watch a video!



Credit: Thomas Wiecki

# MCMC Sampling Algorithms

- Rejection sampling
- Importance sampling
- Metropolis algorithm
- Gibbs sampling → JAGS
- HMC sampling*

Stan!

*: also PyMC3