Bayesian Statistics and Hierarchical Bayesian Modeling for Psychological Science

Lecture 06

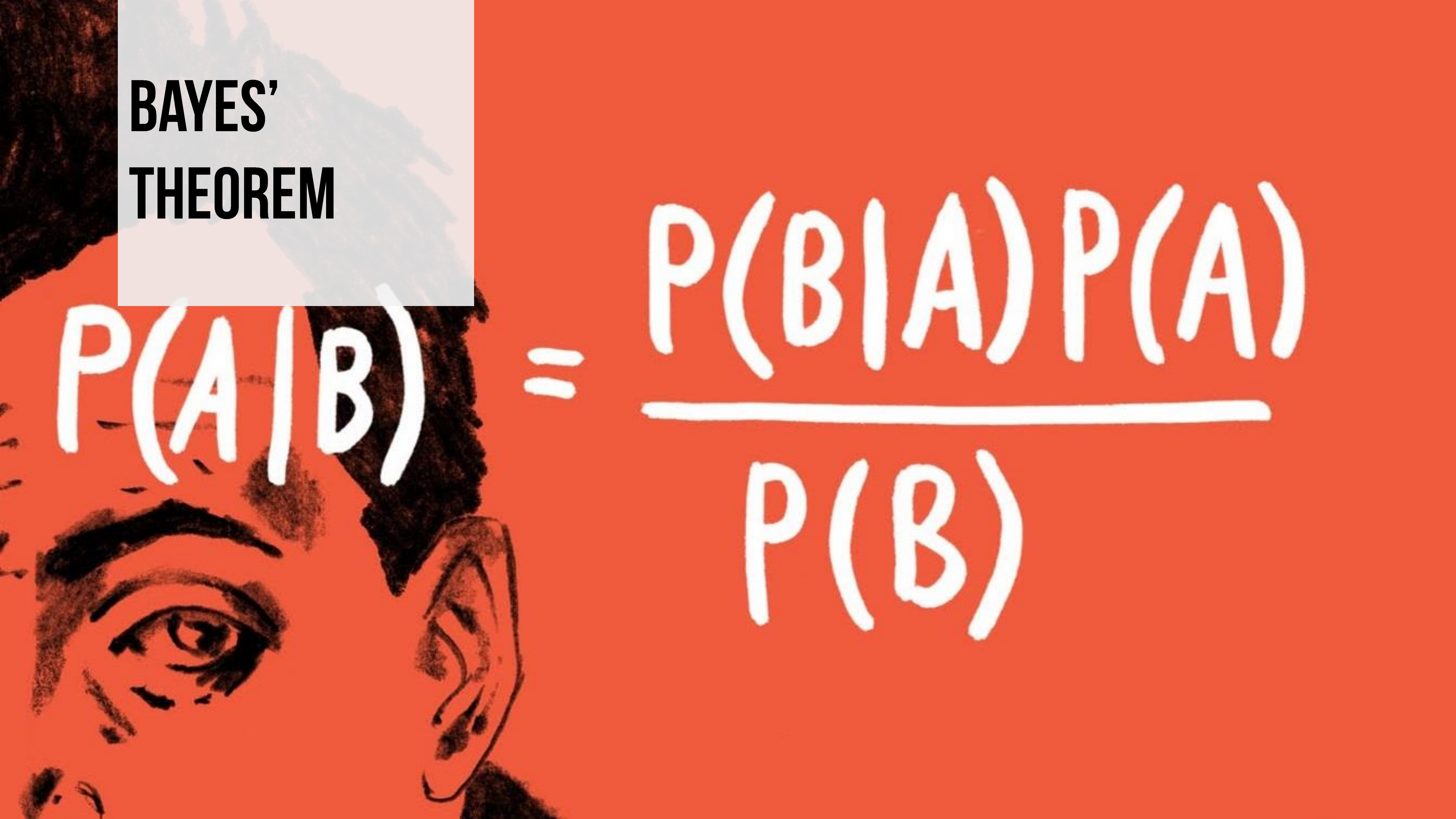
Lei Zhang
Social, Cognitive and Affective Neuroscience Unit (SCAN-Unit)
Department of Basic Psychological Research and Research Methods

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

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

universität wien
Fakultät für Psychologie

BAYES' THEOREM



Bayes' theorem

$$p(A,B) = p(B,A)$$

$$p(A,B) = p(A|B)\,p(B)$$

$$p(B,A) = p(B|A)\,p(A)$$

$$p(A|B)\,p(B) = p(B|A)\,p(A)$$

$$\boxed{p(A \mid B) = \frac{p(B \mid A)\,p(A)}{p(B)}}$$

Example

		disease X	
		0	1
symptoms Y	0	0.5	0.1
	1	0.1	0.3

Joint probability : $P(X = 0, Y = 1) = 0.1$

$$\sum_{x,y} P(X = x, Y = y) = 1$$

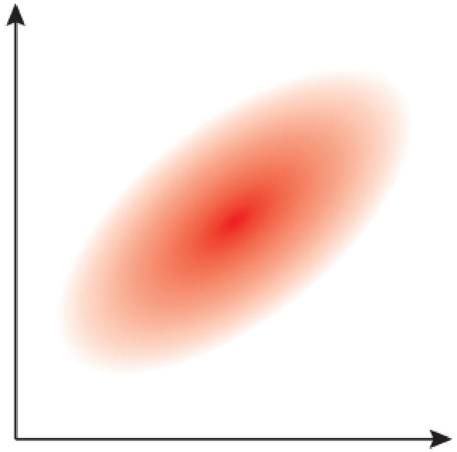
Marginal probability :

$$P(Y = 1) = 0.1 + 0.3 = 0.4$$

$$P(X = 0) = 0.1 + 0.5 = 0.6$$

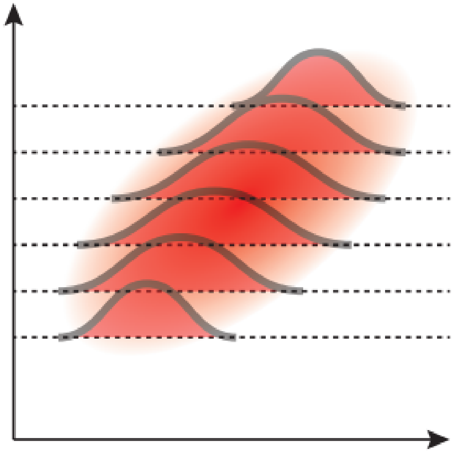
$$P(X = x) = \sum_{y} P(X = x, Y = y)$$

joint distribution



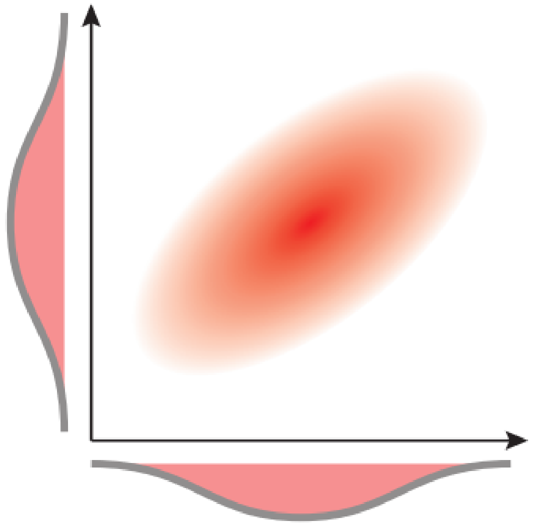
The "co-distribution" of x and y.

conditional distribution



The probability distribution of x, given that we know the value of y.

mariginal distribution



The density of x- (or y-) values, without knowing the other's value.

Credit: [Eric Ma]

Bayesian **warm-up**?

Second Example

Row	Column			Marginal
	$\ldots$	c	$\ldots$	
$\vdots$		$\vdots$		
r	$\cdots$	$p(r, c) = p(r \mid c)\, p(c)$	$\cdots$	$p(r) = \sum_{c^*} p(r \mid c^*)\, p(c^*)$
$\vdots$		$\vdots$		
Marginal		$p(c)$		

Kruschke (2015)

Second Example

	Hair color				
Eye color	Black	Brunette	Red	Blond	Marginal (Eye color)
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{\sum_{A^*} p(B \mid A^*)\, p(A^*)}$$

Kruschke (2015)

Exercise VI

Suppose that in the general population, the probability of having a rare disease is 1/1000. We denote the true presence or absence of the disease as the value of a parameter, ϑ, that can have the value ϑ = ☹ if disease is present in a person, or the value ϑ = ☺ if the disease is absent. The base rate of the disease is therefore denoted $p(\vartheta = ☹) = 0.001$.

Suppose(1): a test for the disease that has a 99% hit rate: $p(T = + | \vartheta = ☹) = 0.99$

Suppose(2): the test has a false alarm rate of 5%: $p(T = + | \vartheta = ☺) = 0.05$

Q: Suppose we sample a person at random from the population, administer the test, and it comes up positive. What is the posterior probability that the person has the disease?

Exercise VI

Q: What is the posterior probability that the person has the disease?

→ $p(\vartheta = ☹ \mid T = +)$

Kruschke (2015)

Exercise VI

	Disease		
Test result	**$\theta = \ddot{\frown}$ (present)**	**$\theta = \ddot{\smile}$ (absent)**	**Marginal (test result)**
$T = +$	$p(+\mid\ddot{\frown})\,p(\ddot{\frown}) = 0.99 \cdot 0.001$	$p(+\mid\ddot{\smile})\,p(\ddot{\smile}) = 0.05 \cdot (1 - 0.001)$	$\sum_\theta p(+\mid\theta)\,p(\theta)$
$T = -$	$p(-\mid\ddot{\frown})\,p(\ddot{\frown}) = (1 - 0.99) \cdot 0.001$	$p(-\mid\ddot{\smile})\,p(\ddot{\smile}) = (1 - 0.05) \cdot (1 - 0.001)$	$\sum_\theta p(-\mid\theta)\,p(\theta)$
Marginal (disease)	$p(\ddot{\frown}) = 0.001$	$p(\ddot{\smile}) = 1 - 0.001$	1.0

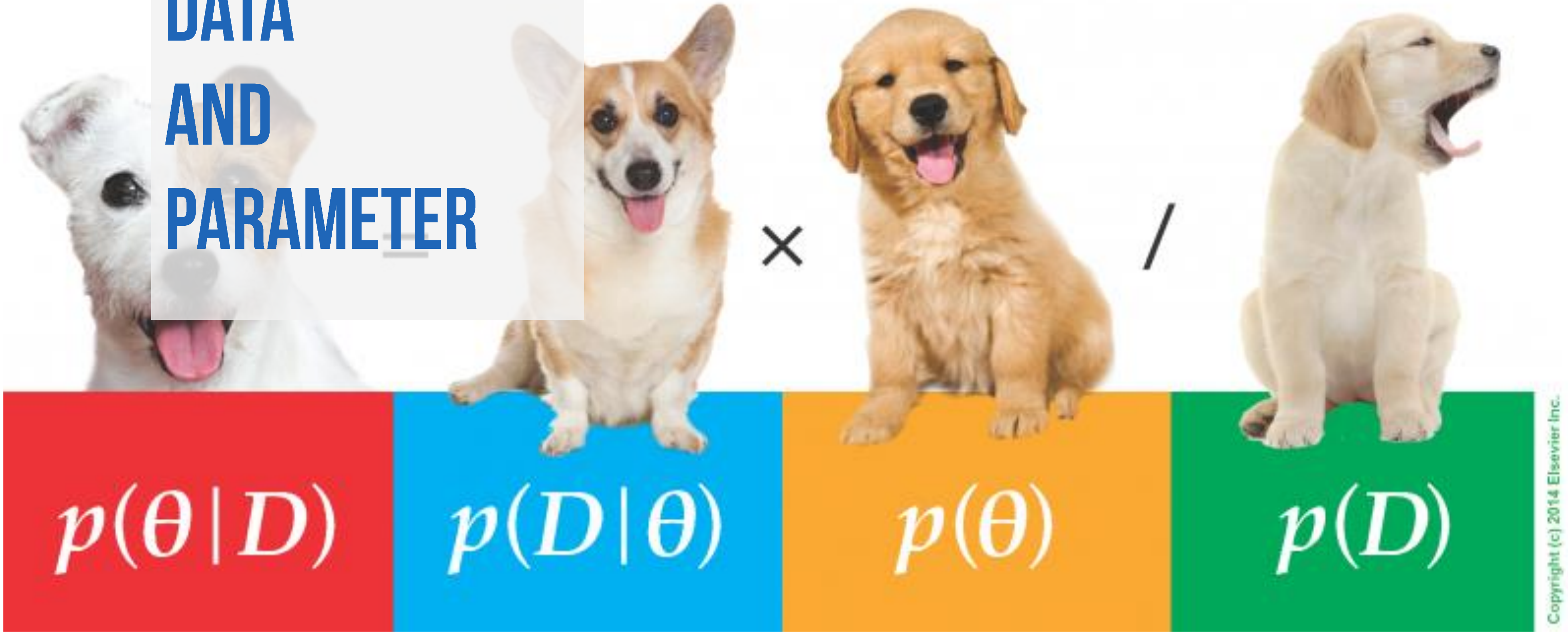
$$p(\theta = \ddot{\frown} \mid T = +) = \frac{p(T = + \mid \theta = \ddot{\frown})\,p(\theta = \ddot{\frown})}{\sum_\theta p(T = + \mid \theta)\,p(\theta)}$$

$$= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.05 \cdot (1 - 0.001)}$$

$$= 0.019$$

Kruschke (2015)

LINKING DATA AND PARAMETER



Linking Data and Parameter

θ $\qquad\qquad$ 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, ϑ varies
- Not a probability distribution
  - the sum is not "one"

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

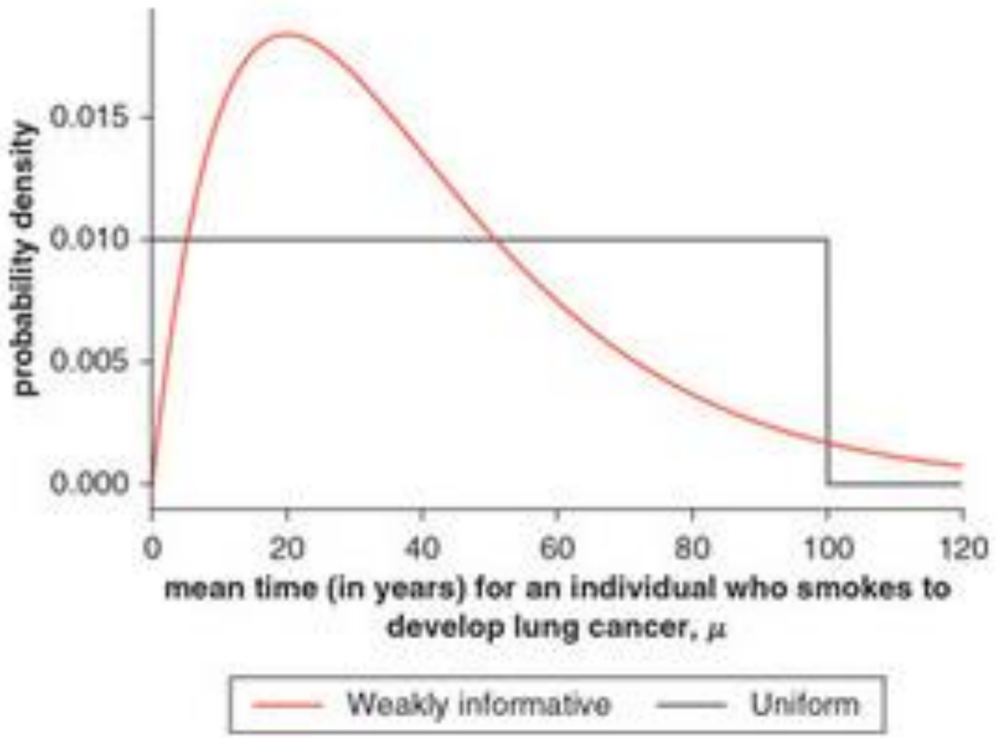
$$Pr(X=1\,|\,\theta)=Pr(H,T\,|\,\theta)+Pr(T,H\,|\,\theta)=2\times Pr(T\,|\,\theta)\times Pr(H\,|\,\theta)=2\theta(1-\theta)$$

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

Probability of coin landing heads up, θ	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*)?

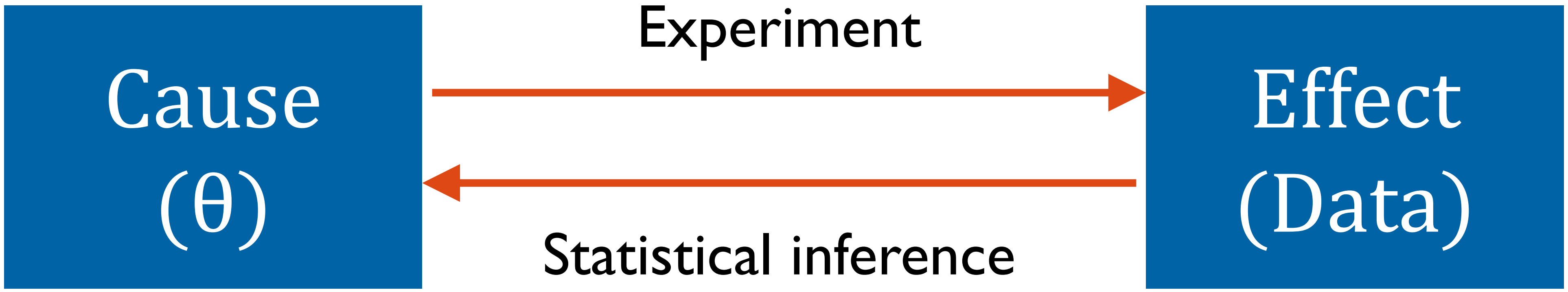
discrete parameters

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

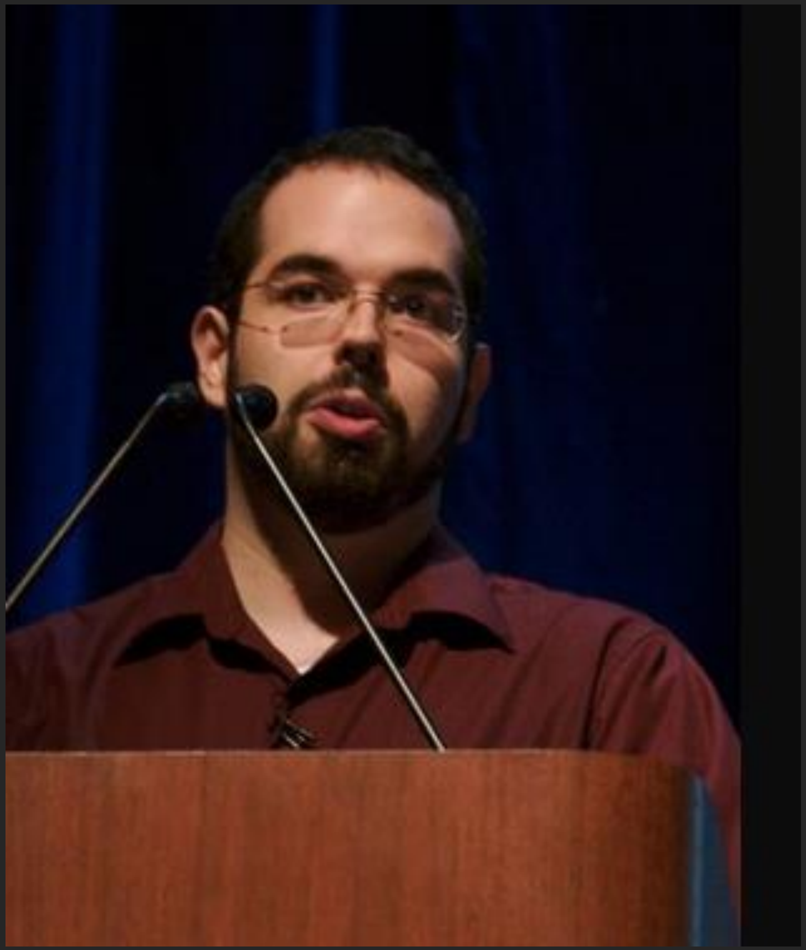
continuous parameters

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

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

Bayesian Statistics and Hierarchical Bayesian Modeling for Psychological Science

Lecture 07

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

BINOMIAL MODEL

Binomial Model

cognitive model
statistics
computing

- 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?

Steps of (Bayesian) Modeling?

cognitive model
statistics
computing

A data story

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

Update

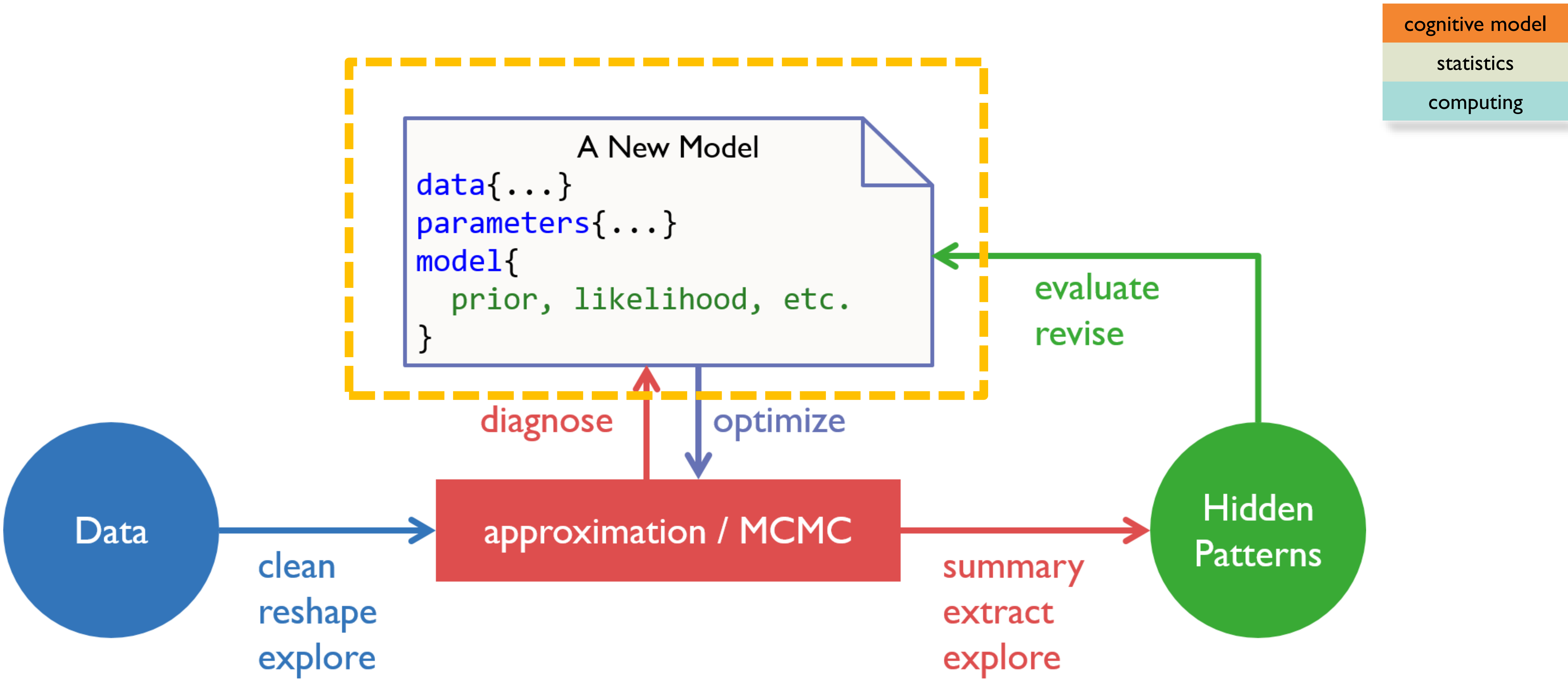
Educate your model by feeding it with data.
Bayesian Update:

- 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)


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

A Data Story of the Globe

cognitive model
statistics
computing

- The true proportion of water covering the globe is ϑ.
- A single toss of the globe has a probability p of producing a water (W) observation.
- It has a probability $(1 - \vartheta)$ 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):

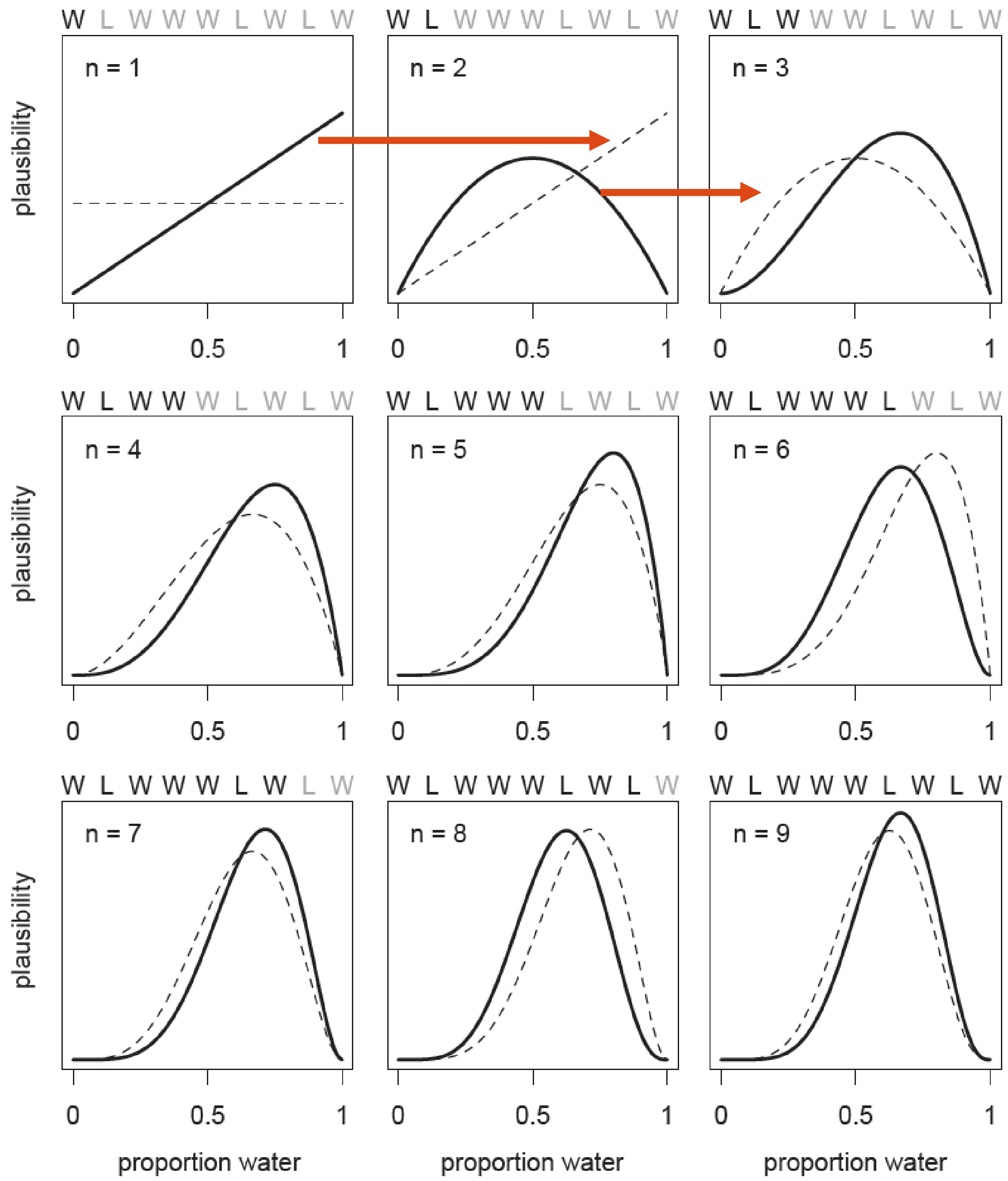
$$p(\theta \mid D) = \frac{\boxed{p(D \mid \theta)}\, p(\theta)}{\int p(D \mid \theta^*)\, p(\theta^*)\, d\theta^*}$$

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

N: total number of observations
w: number of water } known (data)

ϑ : proportion of water   unknown (parameter)

Update



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

McElreath (2016)

Solve it by Grid Approximation

cognitive model

statistics

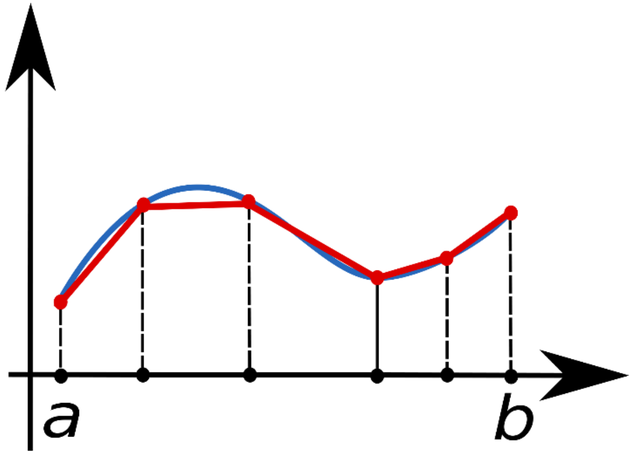
computing

discrete parameters

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

continuous parameters

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

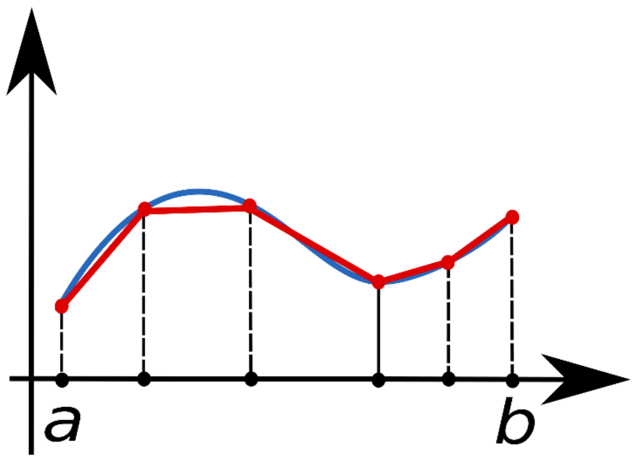


Binomial Model – Grid Approximation

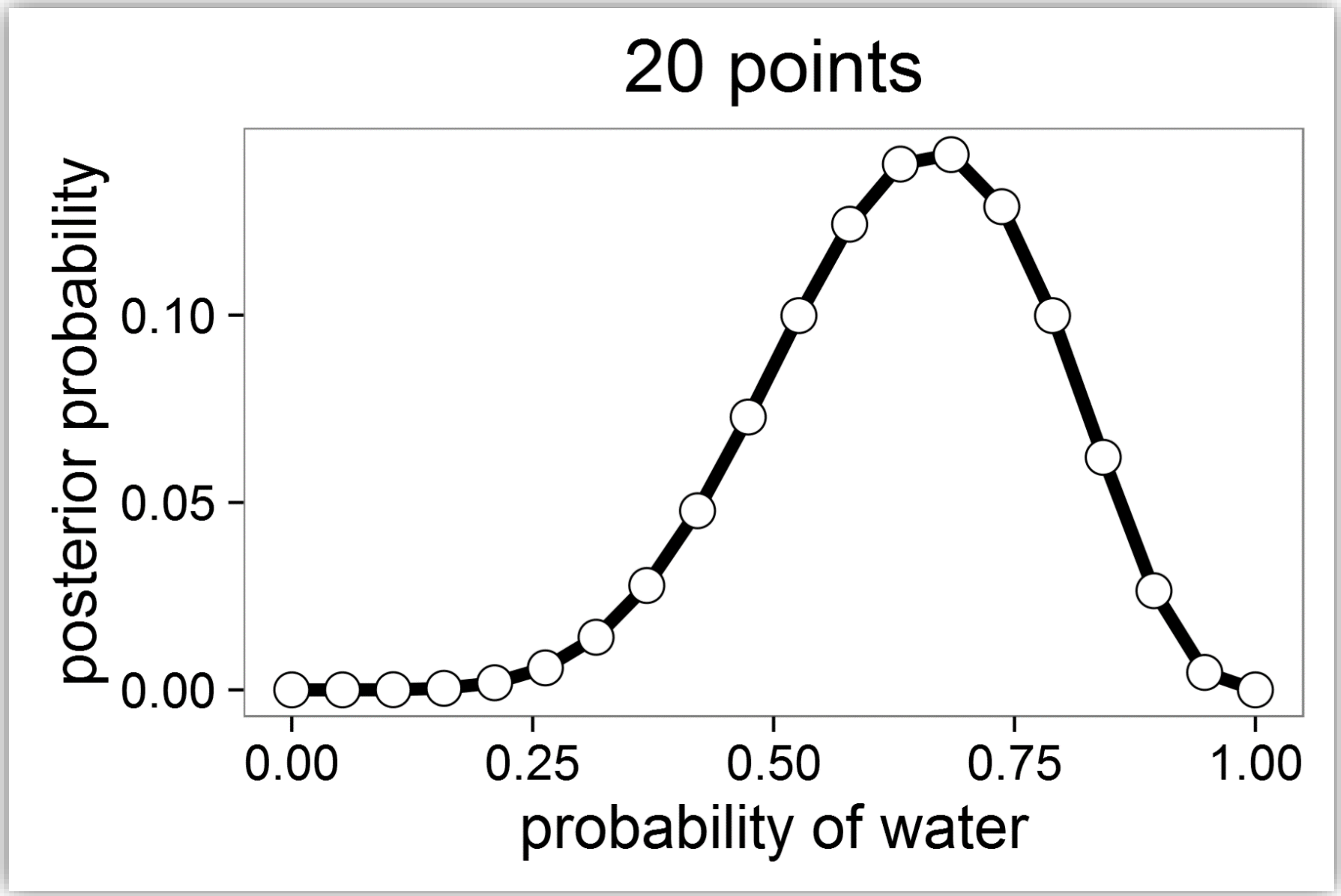

```
theta_start <- 0; theta_end <- 1; n_grid <- 20
w <- 6; N <- 9

# define grid
theta_grid <- seq(from = theta_start, to = theta_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 = theta_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)
```

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

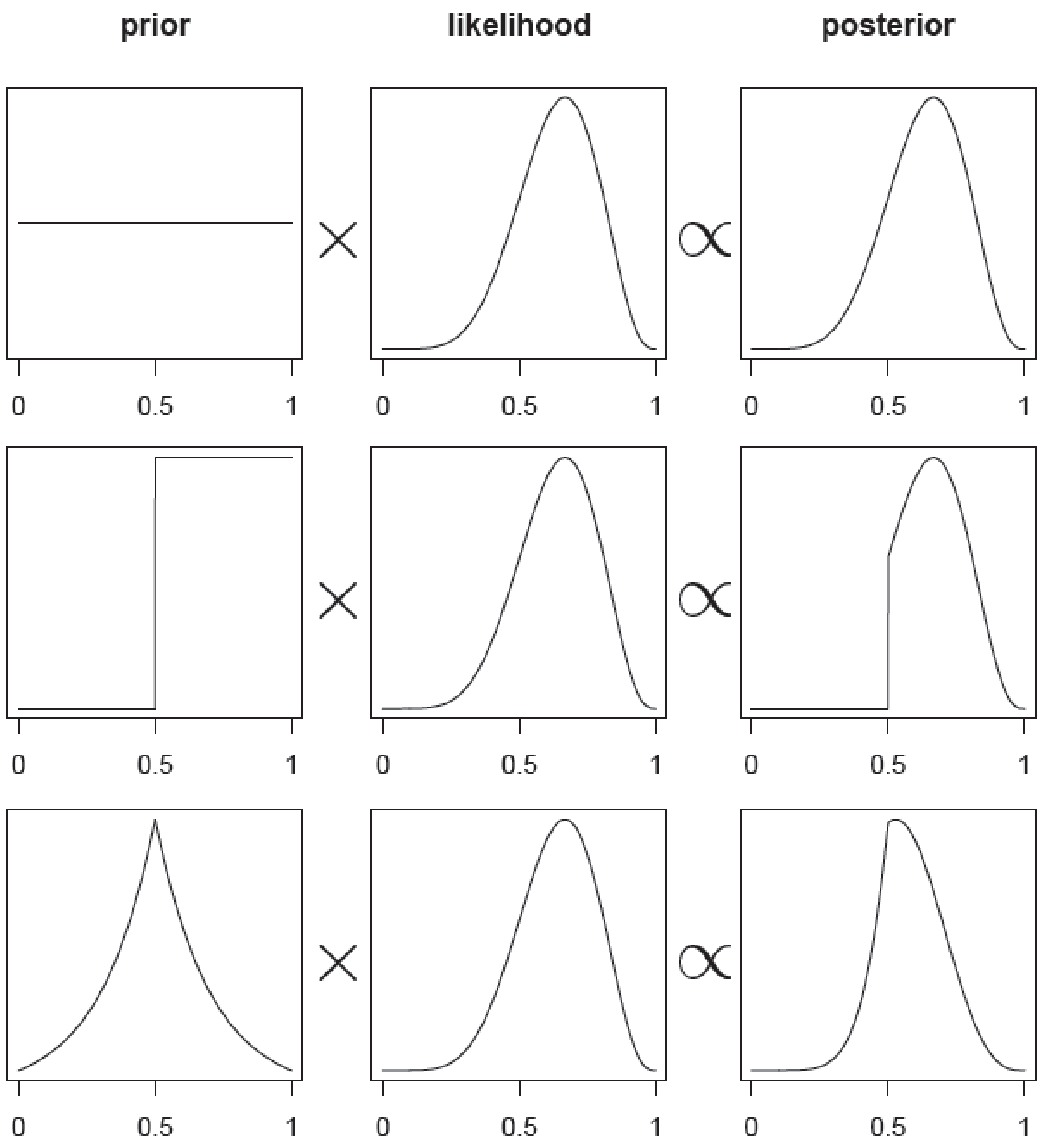
$$p(w \mid N, \theta) = \binom{N}{w} \theta^w (1-\theta)^{N-w}$$



Binomial Model – Grid Approximation



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?

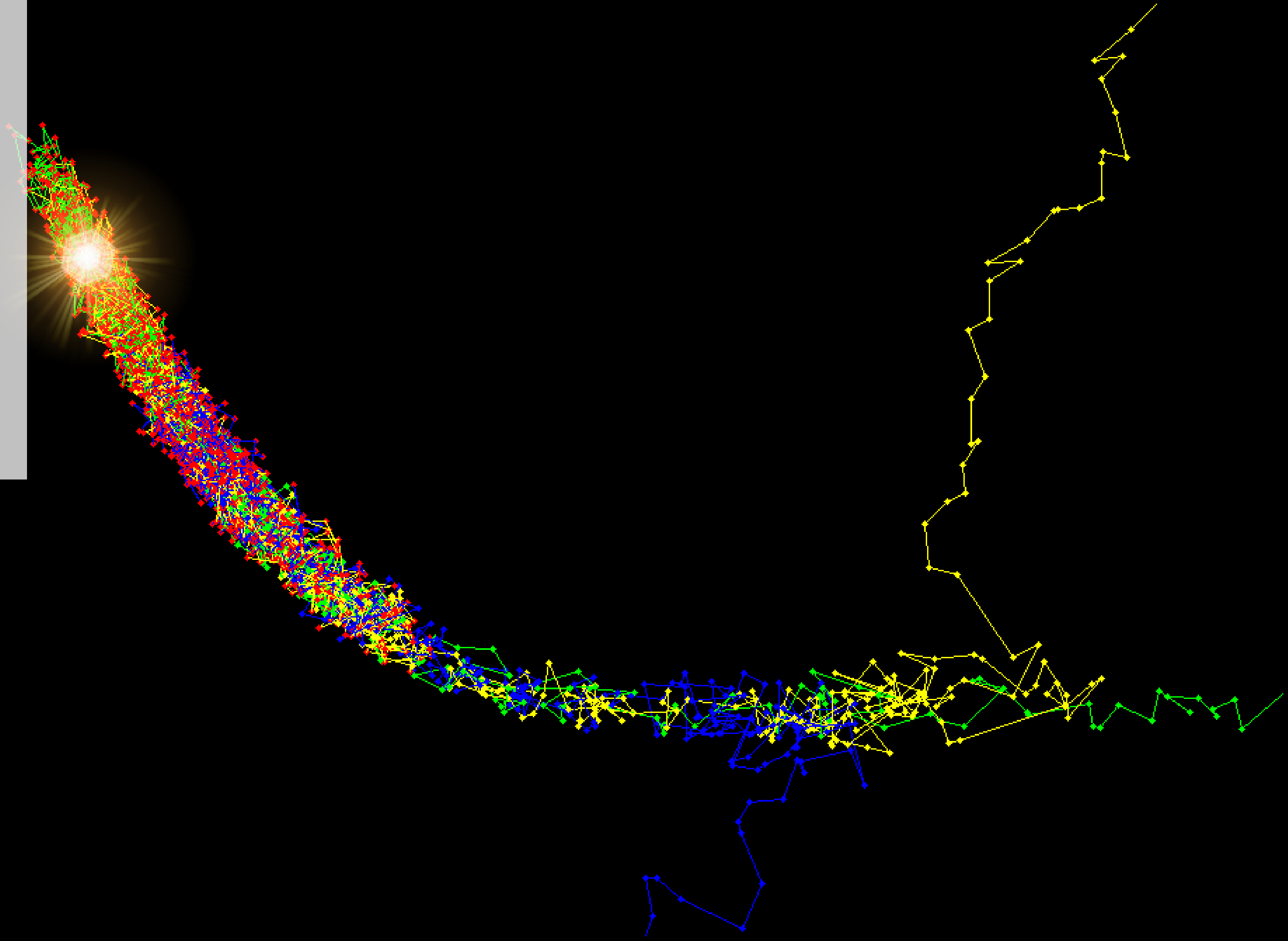
$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{\boxed{\int p(D \mid \theta^*)\, p(\theta^*)\, d\theta^*}}$$

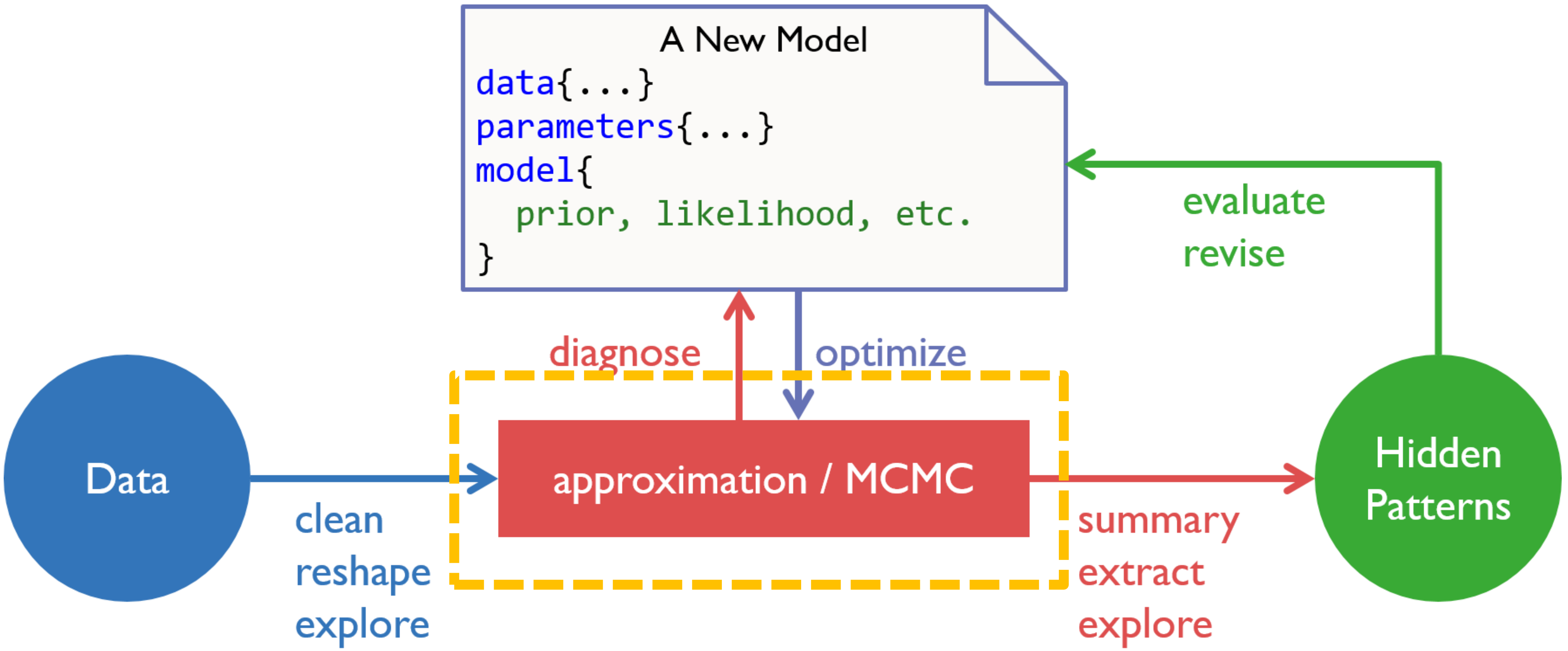
$$p(data) = \int_{\mathrm{All}\,\theta_1} \int_{\mathrm{All}\,\theta_2} p(data, \theta_1, \theta_2)\, \mathrm{d}\theta_1 \mathrm{d}\theta_2$$

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

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

MARKOV CHAIN MONTE CARLO




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

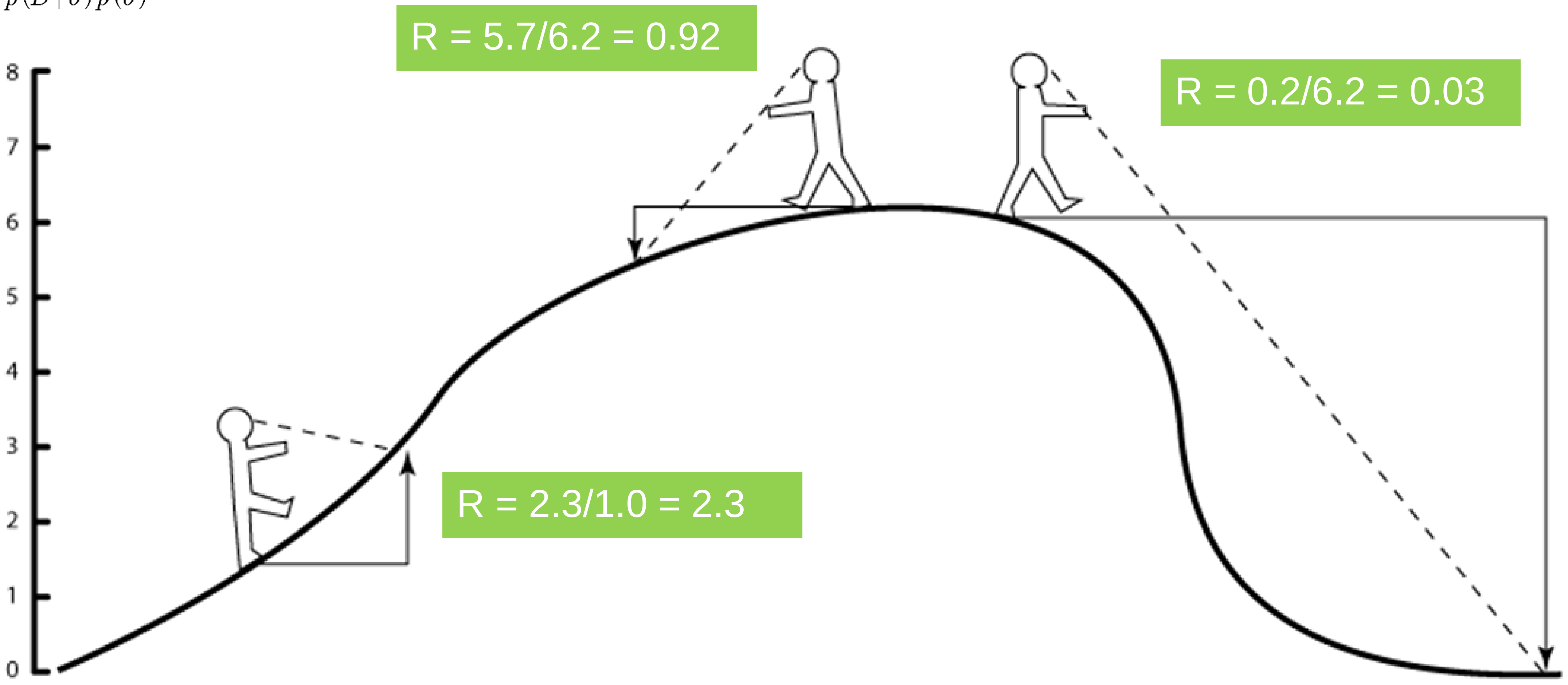
Solving the Problem by Approximation

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

Deterministic Approximation	Stochastic Approximation
→ Variational Bayes	→ Sampling Methods

An MCMC Robot

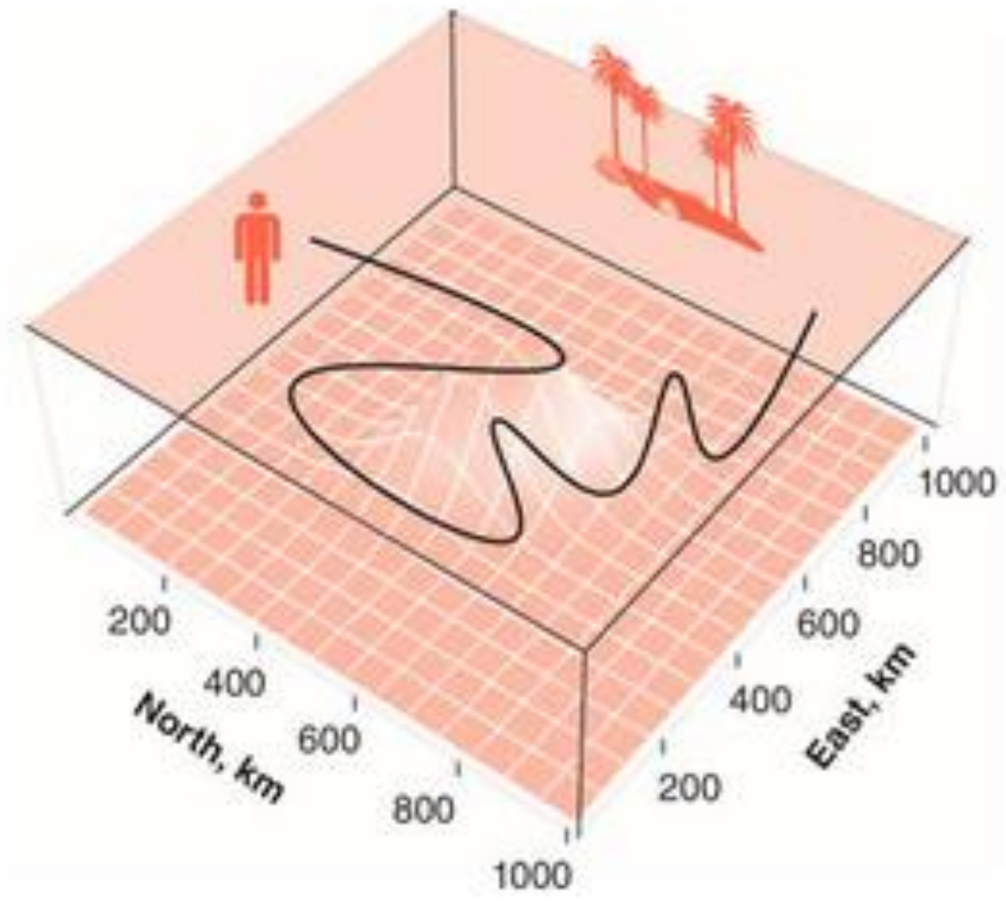
$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$$



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

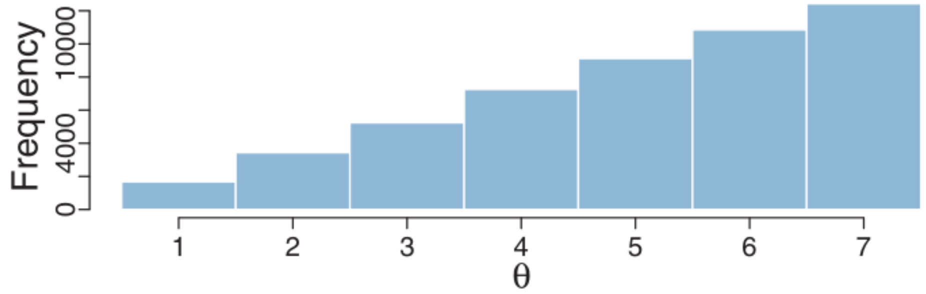
An MCMC Robert in 3D

cognitive model
statistics
computing

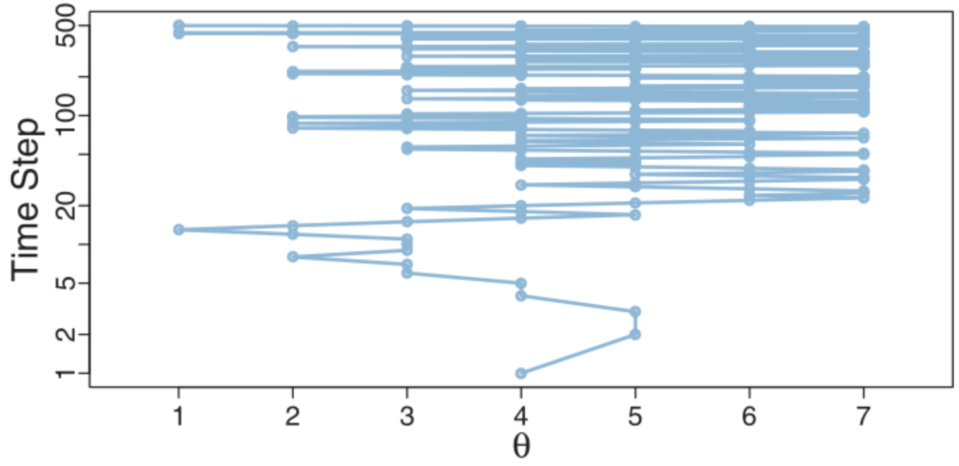


Lambert (2018)

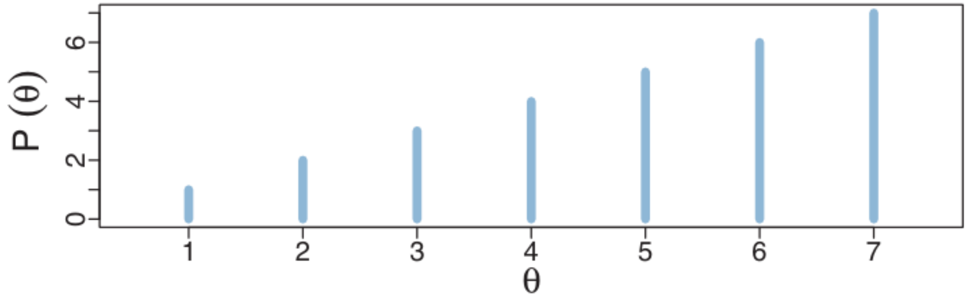
Sampling Example: Discrete



MCMC summary



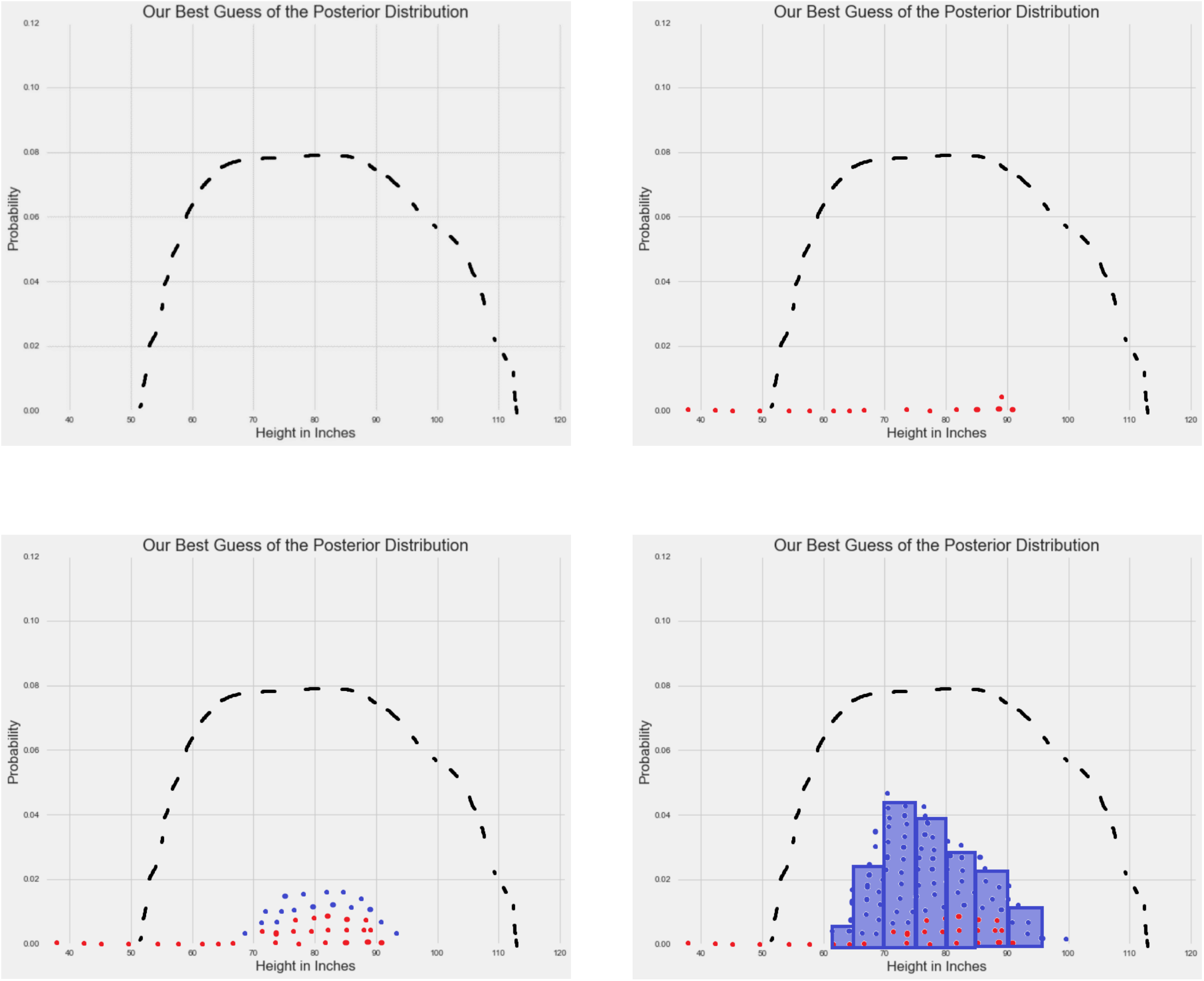
MCMC trace



True distribution

Kruschke (2015)

Sampling Example: Continuous

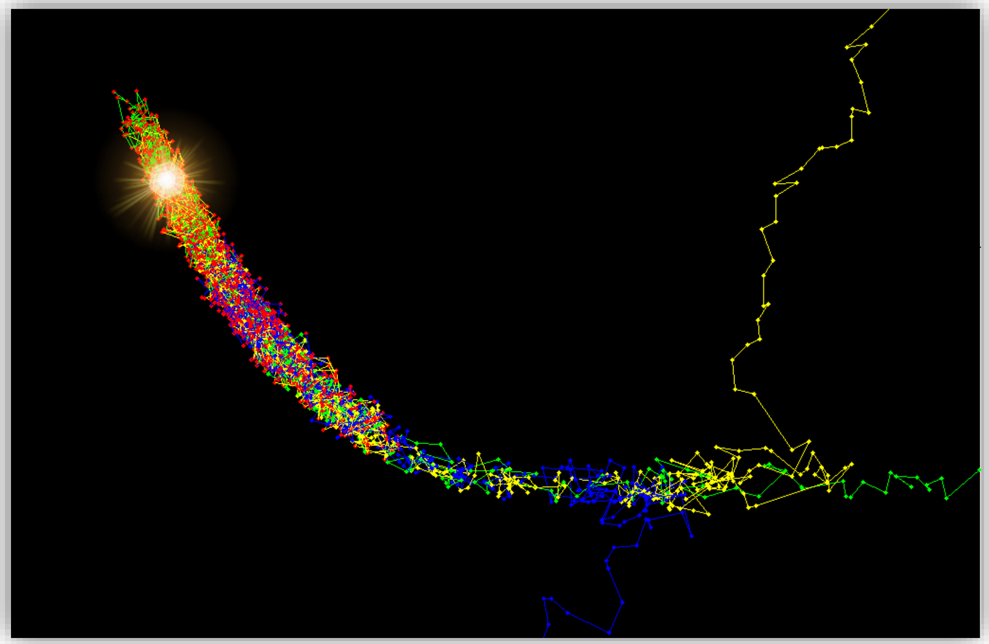


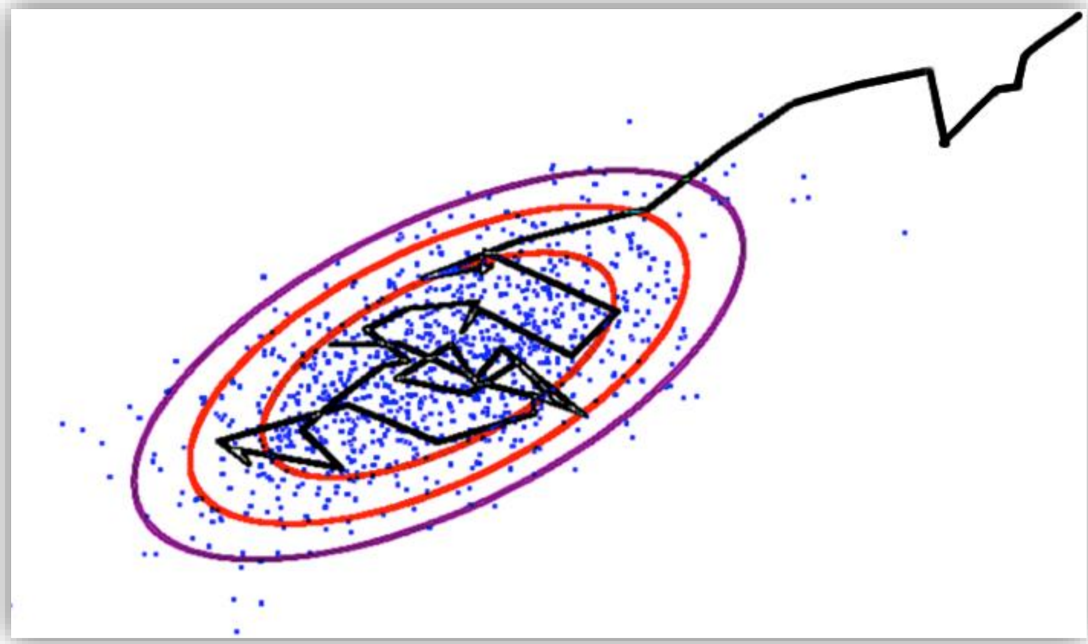
cognitive model
statistics
computing

[A Zero-Math Introduction to Markov Chain Monte Carlo Methods]

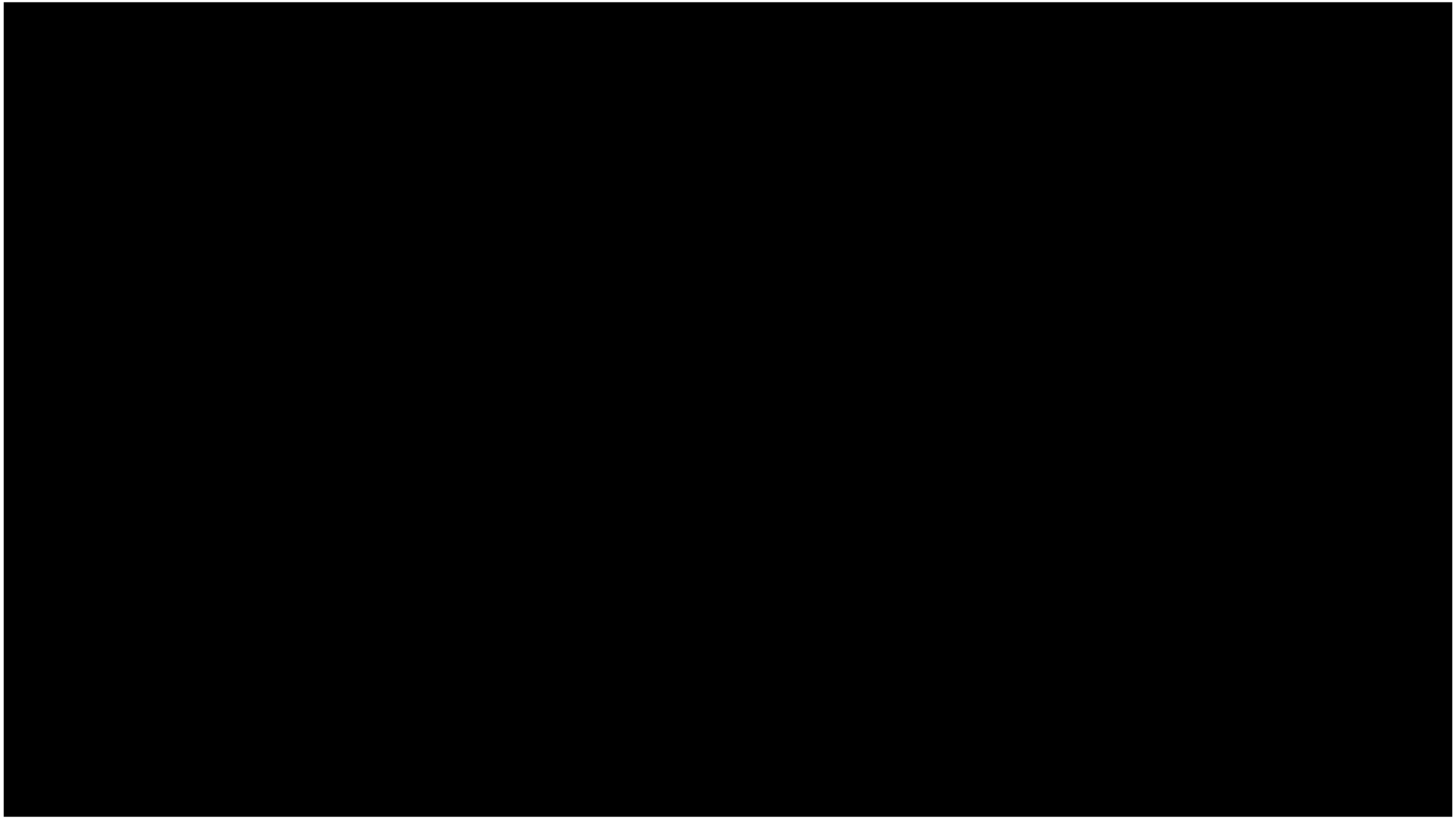
Visual Example

cognitive model
statistics
computing





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 PyMC3