michael.landis@wustl.edu
landislab.github.io
@landismj

Phylogenetic inference and graphical models

Molecular Evolution Workshop

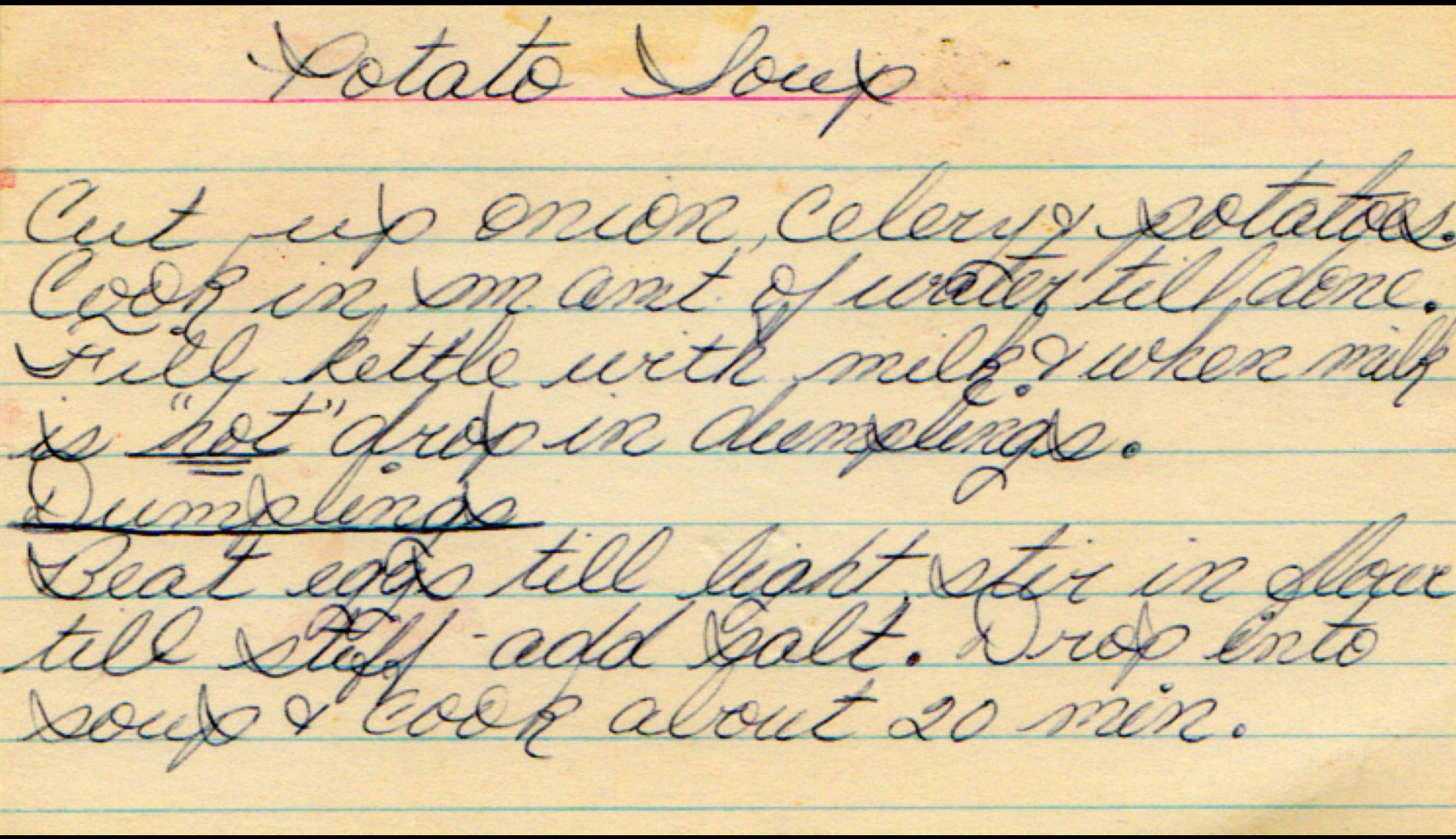
Woods Hole, MA

Michael Landis

Wash U in St. Louis

Sunday, August 4

What's in a recipe?



Potato Soup

Cut up onion, celery & potatoes. Cook in small amt of water till done. Fill kettle with milk & when milk is "hot" drop in dumplings.

Dumplings

Beat eggs till light, stir in flour till stiff - add salt. Drop into soup & cook about 20 min.

ingredients

Potato Soup

Cut up onion, celery & potatoes. Cook in amt of water till done. Fill kettle with milk & when milk is "hot" drop in dumplings.

Dumplings

Beat eggs till light, stir in flour till stiff - add salt. Drop into soup & cook about 20 min.

steps ingredients

Potato Soup

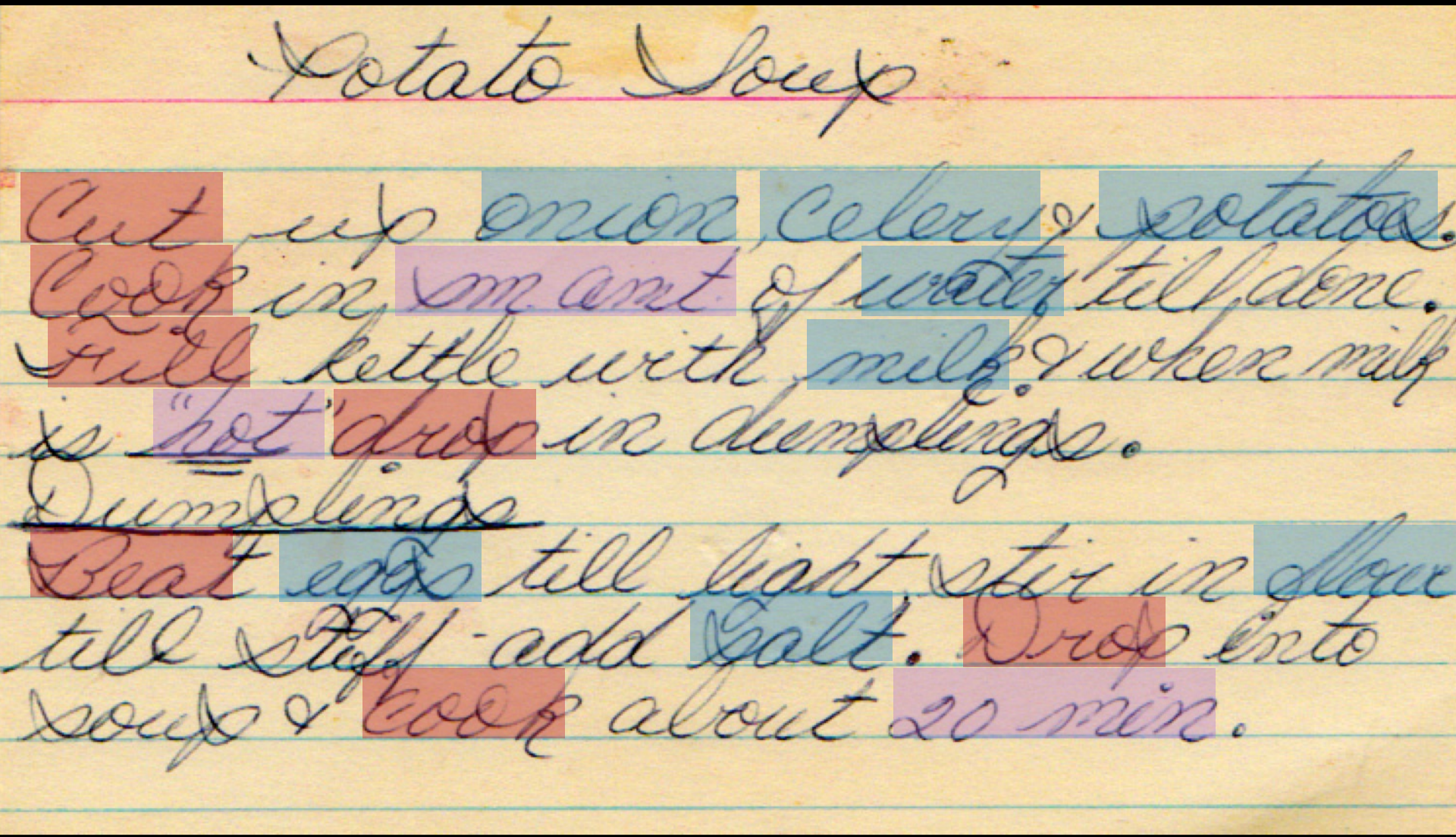
Cut up onion, celery & potatoes. Cook in amt of water till done. Fill kettle with milk & when milk is "hot" drop in dumplings.

Dumplings

Beat eggs till light, stir in flour till stiff - add salt. Drop into soup & cook about 20 min.

steps

ingredients


Potato Soup

Cut up onion, celery & potatoes. Cook in amt of water till done. Fill kettle with milk & when milk is "hot" drop in dumplings.

Dumplings

Beat eggs till light, stir in flour till stiff - add salt. Drop into soup & cook about 20 min.

quantities??

Estimating ingredient quantities is an *inference* problem

Recipe		Model
Ingredients	⟷	Parameters
Steps	⟷	Dependencies
Quantities?	⟷	Parameter values?
Meal	⟷	Data

steps??

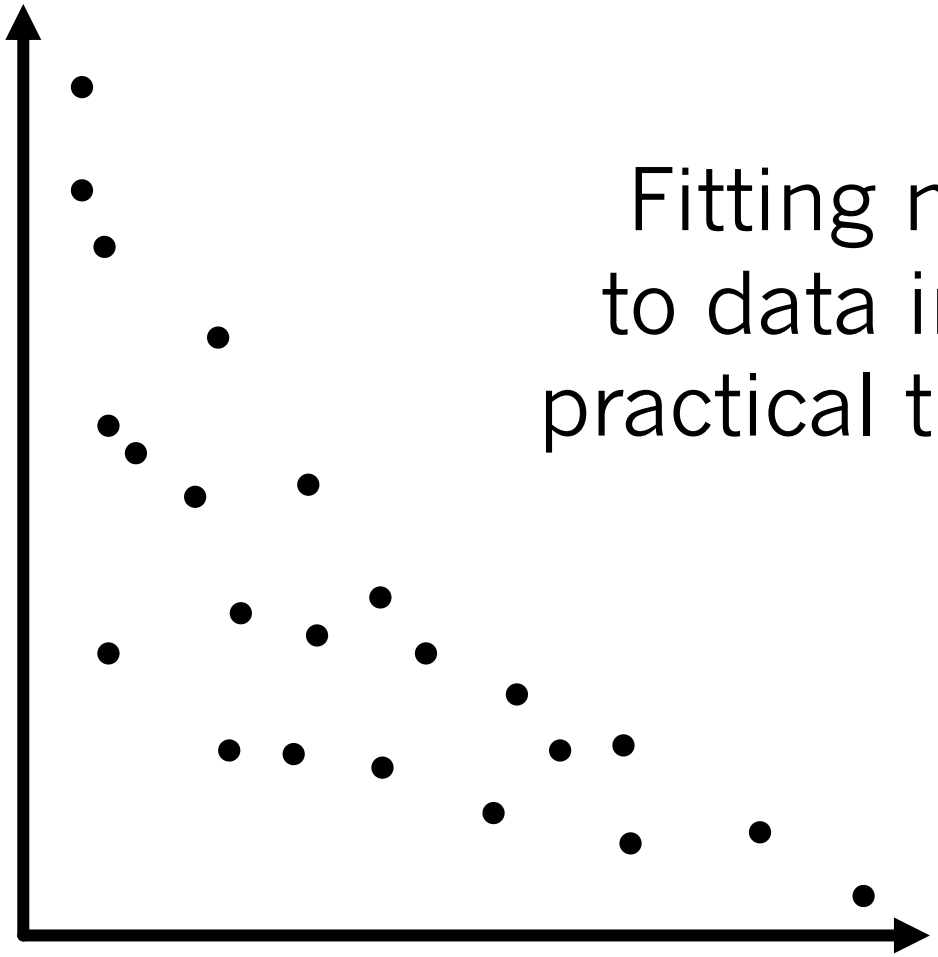
ingredients??

quantities??

Recovering the recipe is a *modeling* problem

Recipe		Model
Ingredients?	⟷	Parameters?
Steps?	⟷	Dependencies?
Quantities?	⟷	Parameter values?
Meal	⟷	Data

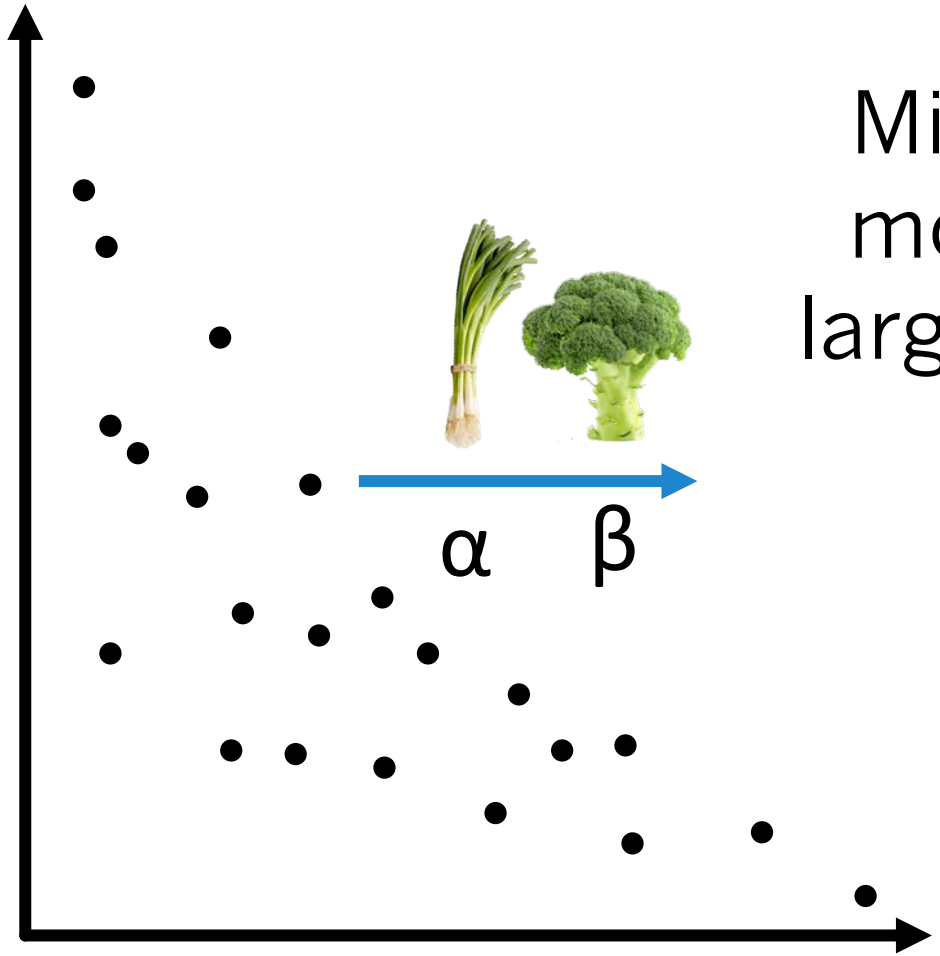
Practicality
time, scale,
cost, ease



Fitting models
to data involves
practical trade-offs

Purity
realism,
correctness,
accuracy,
elegance

Practicality
time, scale,
cost, ease



Minor changes to models can bring large improvements

Purity
realism,
correctness,
accuracy,
elegance

Learned quite a bit already

Basics of probability

Frequentist/Bayesian inference

Substitution processes

Site-rate models

Partition models

Tree space

Pruning algorithm

Model selection

...

About to learn much more

Basics of probability
Frequentist/Bayesian inference
Substitution processes
Site-rate models
Partition models
Tree space
Pruning algorithm
Model selection
Birth-death processes
Selection models
Protein models
Coalescent processes
Phylogenetic network models
Species tree estimation
Species delimitation
...

Practical model literacy

memorize features
for each specific model

– or –

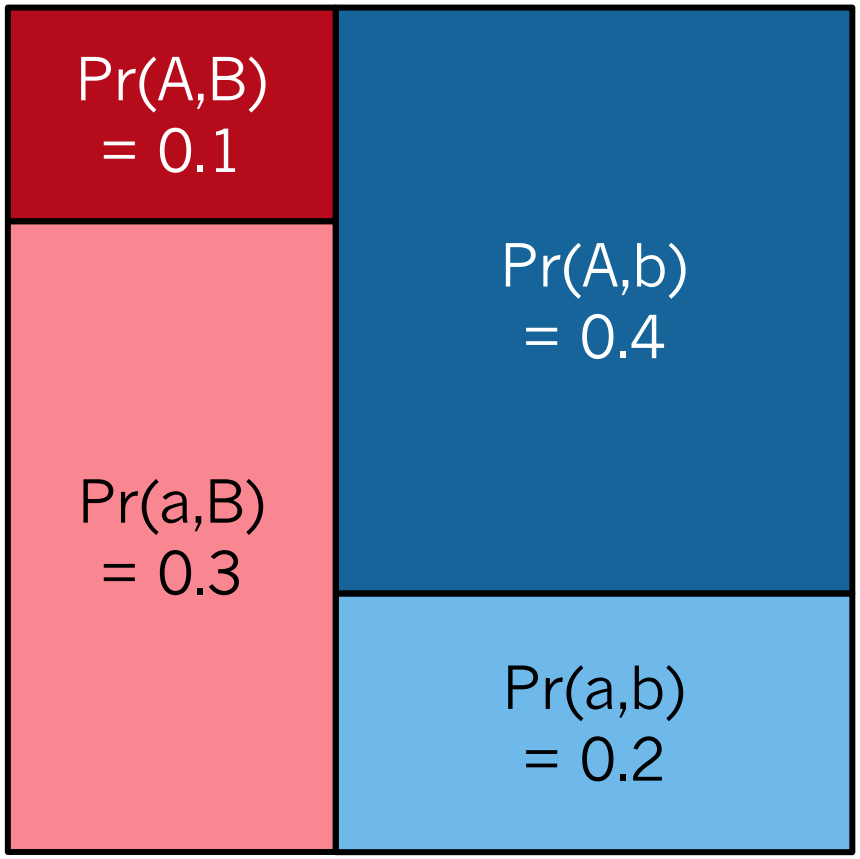
learn to identify, classify, modify
model variants

Outline

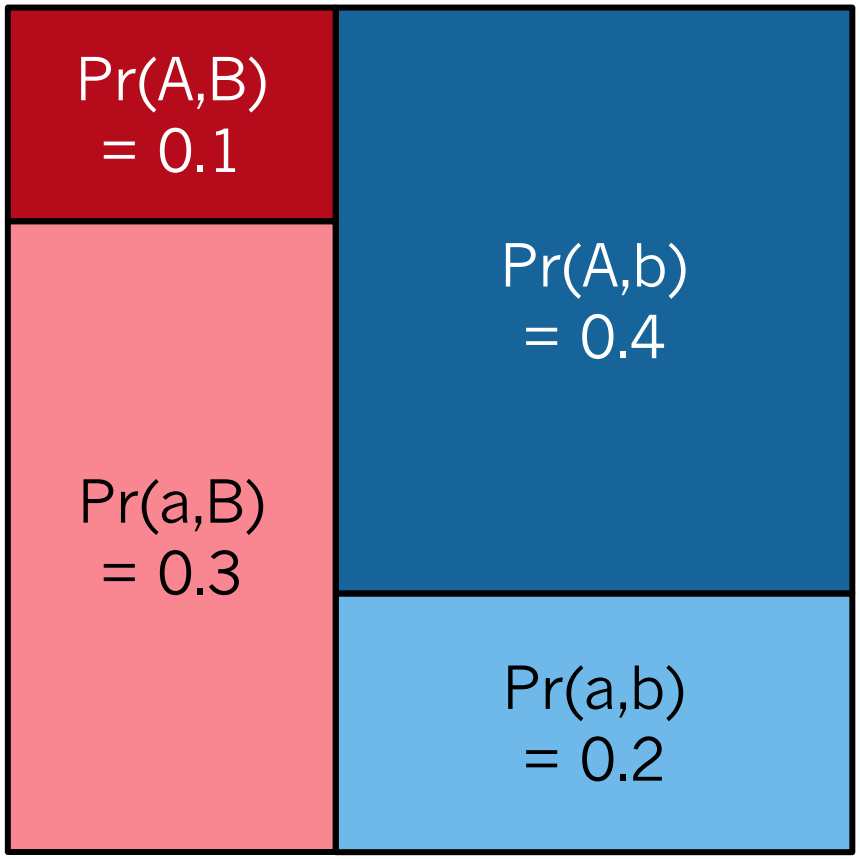
1. Introduction & Motivation
2. What's in a model?
3. Probabilistic Graphical Models (PGMs)
4. PGMs in Phylogenetics
5. Navigating model space

(break)

A quick probability refresher



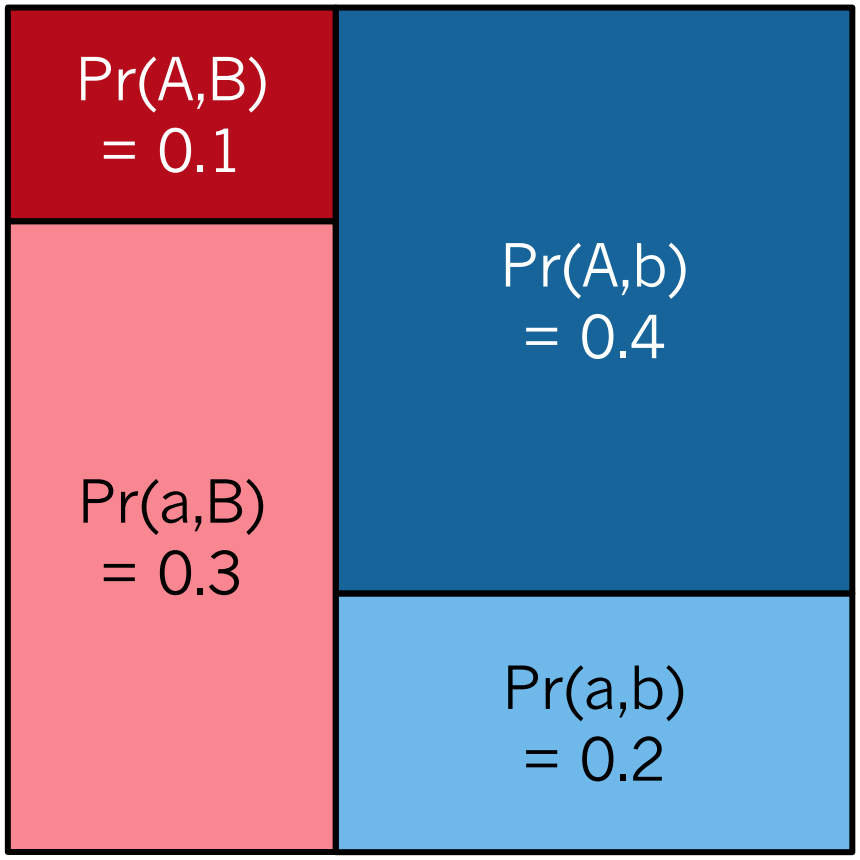
A quick probability refresher



A : You love to eat lobster
a : You hate to eat lobster

B : You are a lobster
b : You are *not* a lobster

Probability rules

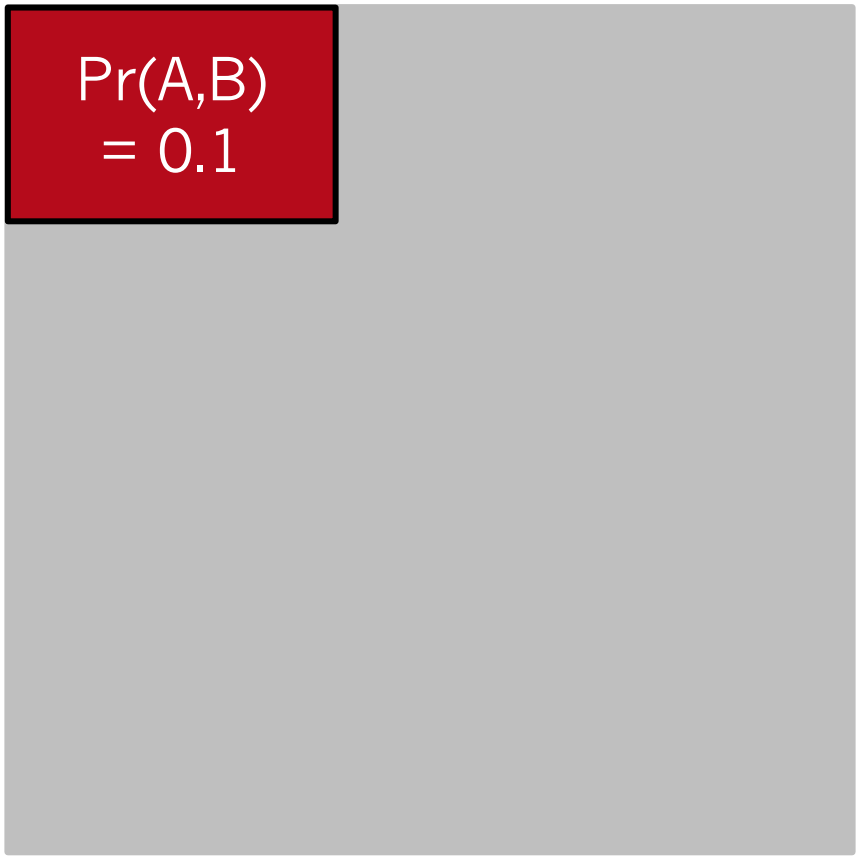


Each outcome probability is from 0 to 1

All outcome probabilities sum to 1

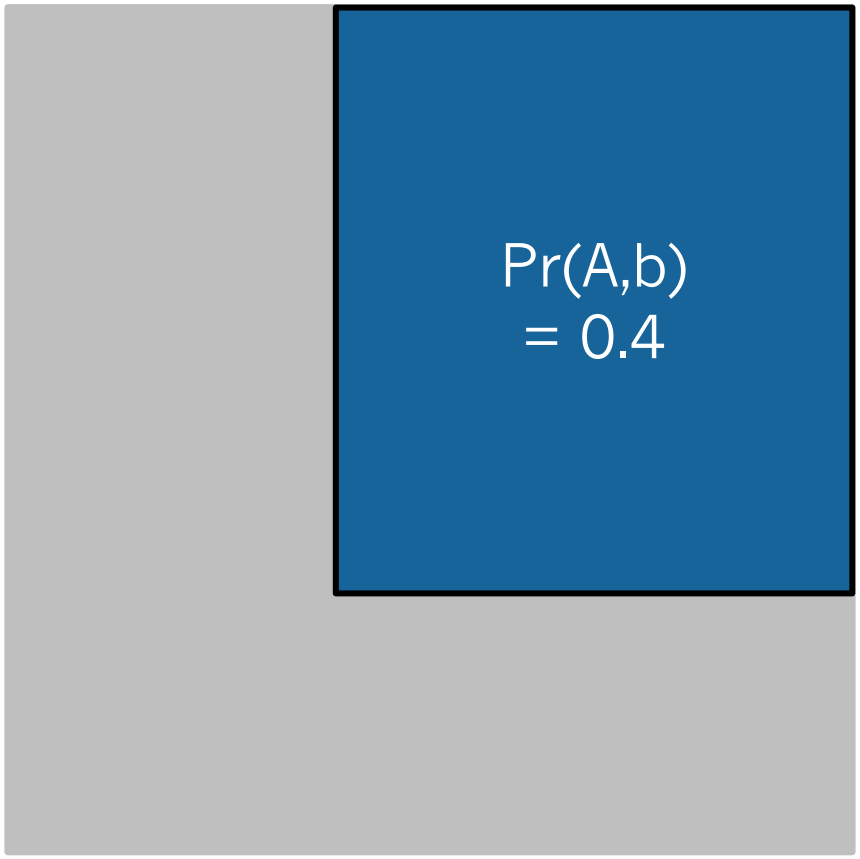
Every outcome is assigned a single probability

Joint probability of A and B



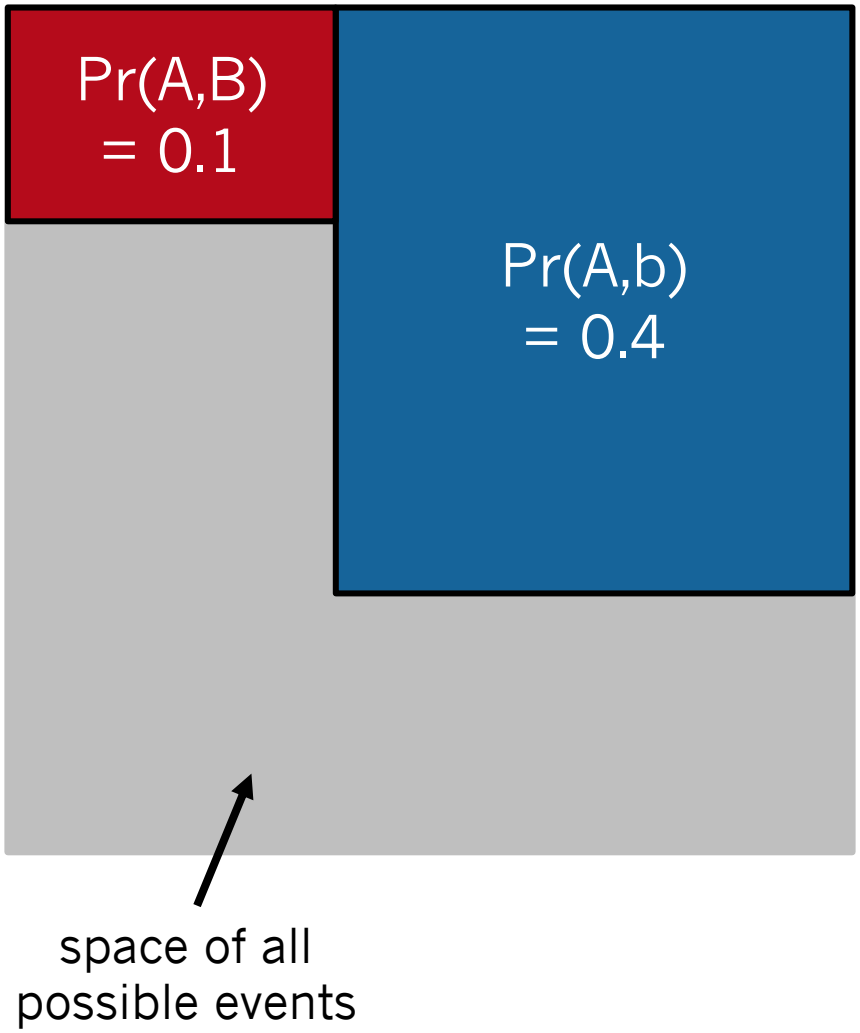
Pr(A,B) = 0.1

Joint probability of A and b



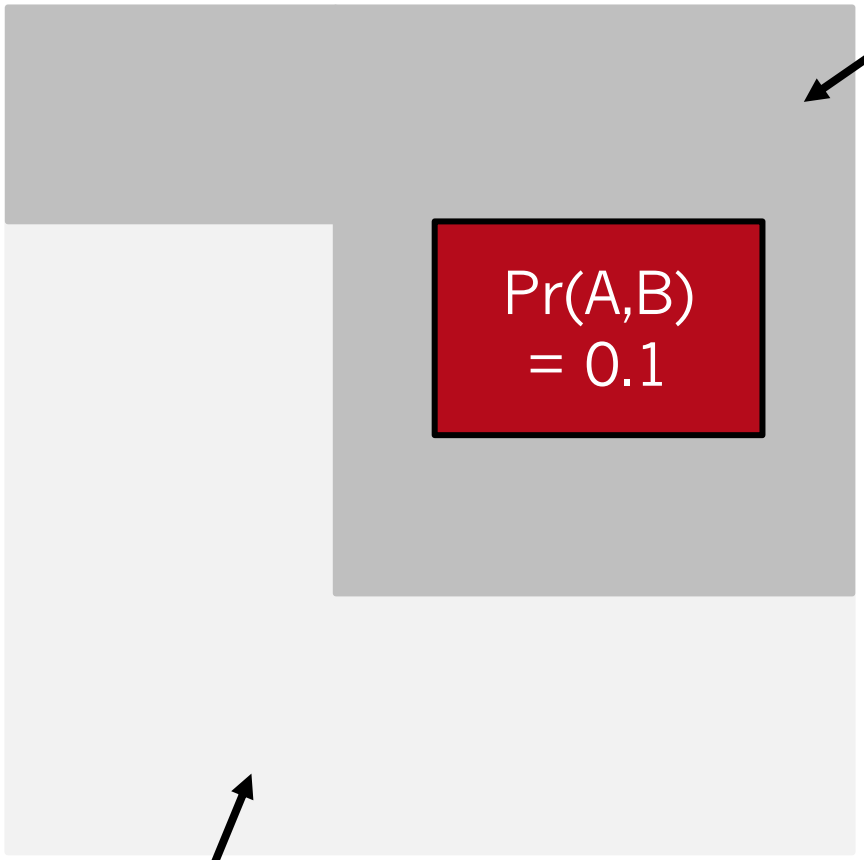
Pr(A,b) = 0.4

Marginal probability of A



$$\Pr(A) = \Pr(A,B) + \Pr(A,b)$$
$$= 0.1 + 0.4$$
$$= 0.5$$

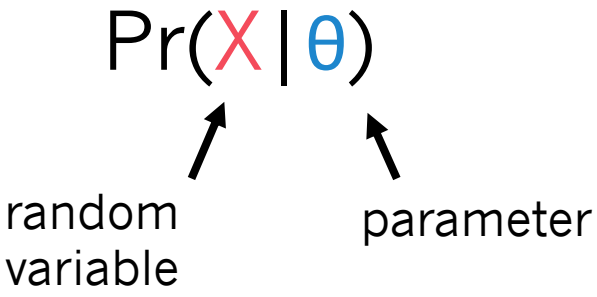
Conditional probability of B given A



$$\Pr(B|A) = \Pr(A,B) / \Pr(A)$$
$$= 0.1 / (0.1 + 0.4)$$
$$= 0.2$$

Probability distributions

"probability of X given θ"



Probability distributions

"probability of X given θ"

$$\Pr(X \mid \theta)$$

random variable

parameter

"X has an Exponential distribution"

$$X \sim \text{Exponential}(\theta)$$

random variable

parameter

Joint distributions as models

Joint probability

$$\Pr(X, Y, \theta, \mu)$$

Joint distributions as models

Joint probability

$$\Pr(X, Y, \theta, \mu)$$

Joint probability, factored

$$\Pr(X|\theta)\ \Pr(Y|\theta,\mu)\ \Pr(\theta|\mu)\ \Pr(\mu)$$

Joint distributions as models

Joint probability

$$\Pr(X, Y, \theta, \mu)$$

Joint probability, factored

$$\Pr(X|\theta)\ \Pr(Y|\theta,\mu)\ \Pr(\theta|\mu)\ \Pr(\mu)$$

Joint distribution

$$\mu \sim \text{Uniform}(0,10)$$
$$\theta \sim \text{Exponential}(1)$$
$$X \sim \text{Exponential}(\theta)$$
$$Y \sim \text{Gamma}(\theta,\mu)$$

Model definitions

Conceptual

Probabilistic

Graphical

Computational

A conceptual model

How might we estimate mutation rate with *K* loci from parent-offspring in yeast?

```
Locus 1     *
Parent     TTTATGT
Offspring  TCTATGT
```

```
Locus 3      *  *
Parent     AGCTATGGG
Offspring  ACCGATGGG
```

```
Locus 4
Parent     TGAA
Offspring  TGAA
```

```
Locus 2        *
Parent     CTCCG
Offspring  CTCAA
```

```
Locus K           * **
Parent     GGTCTATTCCC
Offspring  GGTCTAATGGC
```

A conceptual model

How might we estimate mutation rate with K loci from parent-offspring in yeast?

Number of mutations should be related to total number of sites and mutation rate

A conceptual model

How might we estimate mutation rate with K loci from parent-offspring in yeast?

Number of mutations should be related to total number of sites and mutation rate

The rate should also be similar to estimates for other eukaryotes

Model definitions

Conceptual

Probabilistic

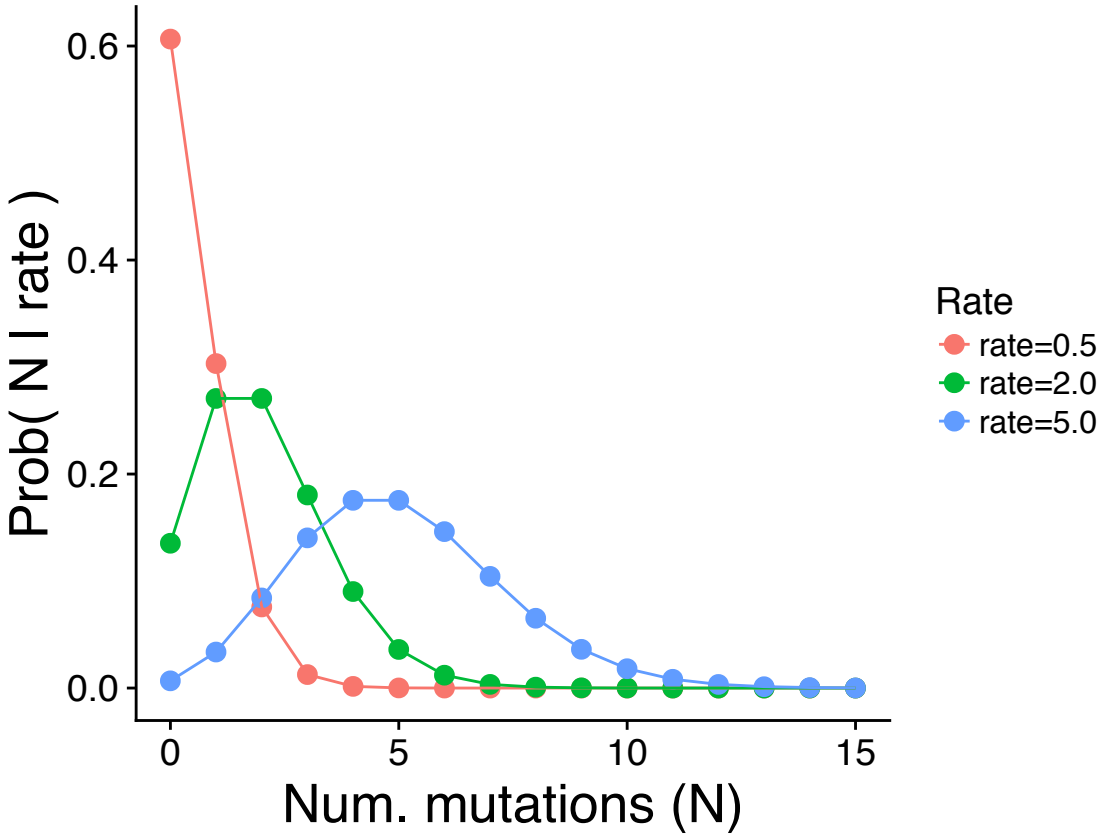
Graphical

Computational

A probabilistic model

Model per-locus mutation counts with Poisson

$N \sim \text{Poisson}(\lambda)$



A probabilistic model

Model per-locus mutation counts with Poisson

$N \sim \text{Poisson}(\lambda)$

Per-locus mutation rate is the per-site mutation rate times the number of sites in the locus

$\lambda = r \times L$

A probabilistic model

Model per-locus mutation counts with Poisson

$N \sim \text{Poisson}(\lambda)$

Per-locus mutation rate is the per-site mutation rate times the number of sites in the locus

$\lambda = r \times L$

Estimate *r* using an empirical prior (eukaryotes)

$r \sim \text{Exponential}(1/10^{-8})$

A probabilistic model

Model per-locus mutation counts with Poisson

$N \sim \text{Poisson}(\lambda)$

Per-locus mutation rate is the per-site mutation rate times the number of sites in the locus

$\lambda = r \times L$

Estimate *r* using an empirical prior (eukaryotes)

$r \sim \text{Exponential}(1/10^{-8})$

Substitute N_k, λ_k, and L_k to generalize for *K* loci

A probabilistic model

Model probability

$$\Pr(N, r \mid L, \mu) = \Pr(r \mid \mu) \prod_k \Pr(N_k \mid r, L_k)$$

A probabilistic model

Model probability

$$\Pr(N, r \mid L, \mu) = \Pr(r \mid \mu) \prod_k \Pr(N_k \mid r, L_k)$$

Model distribution

$$\begin{aligned} r &\sim \text{Exponential}(1/\mu) \\ \lambda_k &= r \times L_k \\ N_k &\sim \text{Poisson}(\lambda_k) \end{aligned}$$

Model definitions

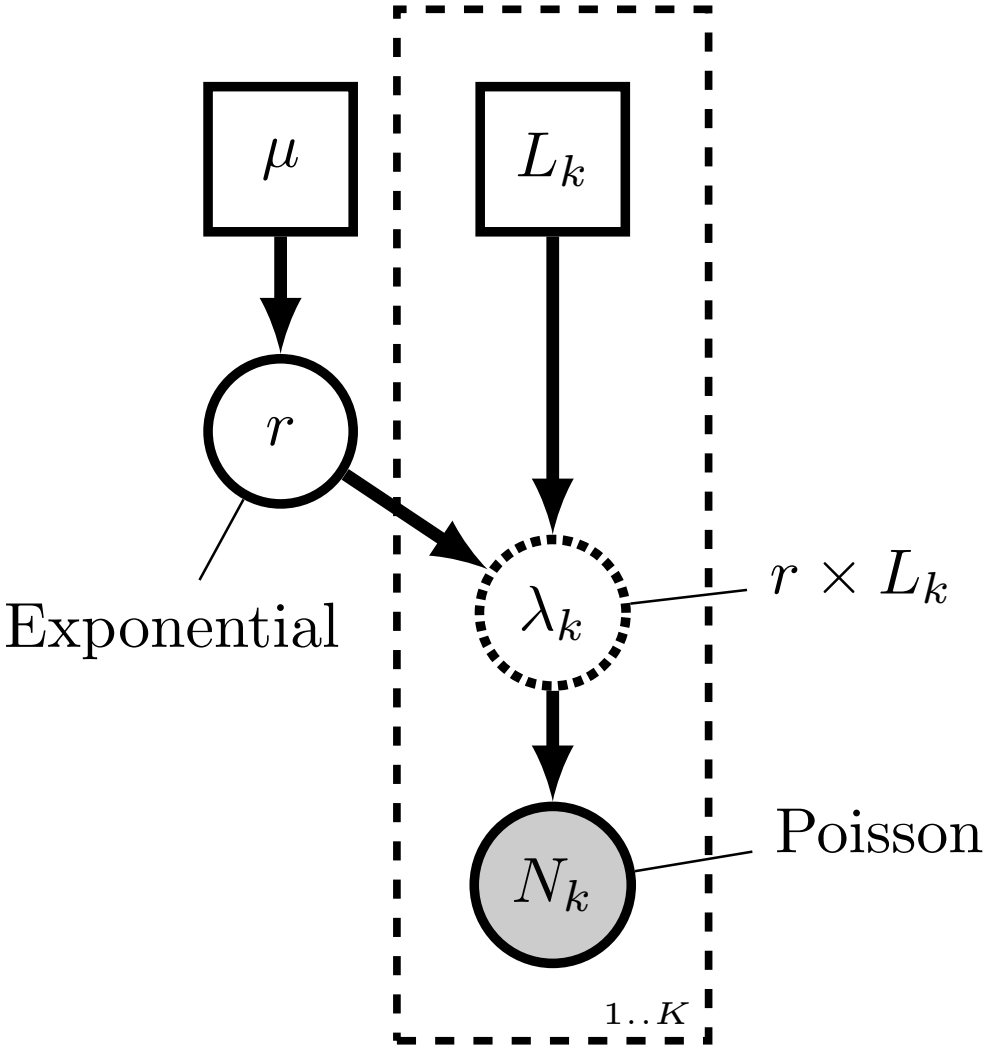
Conceptual

Probabilistic

Graphical

Computational

A probabilistic graphical model



probabilistic model
a set of random variables and dependencies

+

graph
a set of nodes and edges

=

probabilistic graphical model (PGM)
a probabilistic model that uses a graph
to encode its variables and their dependencies

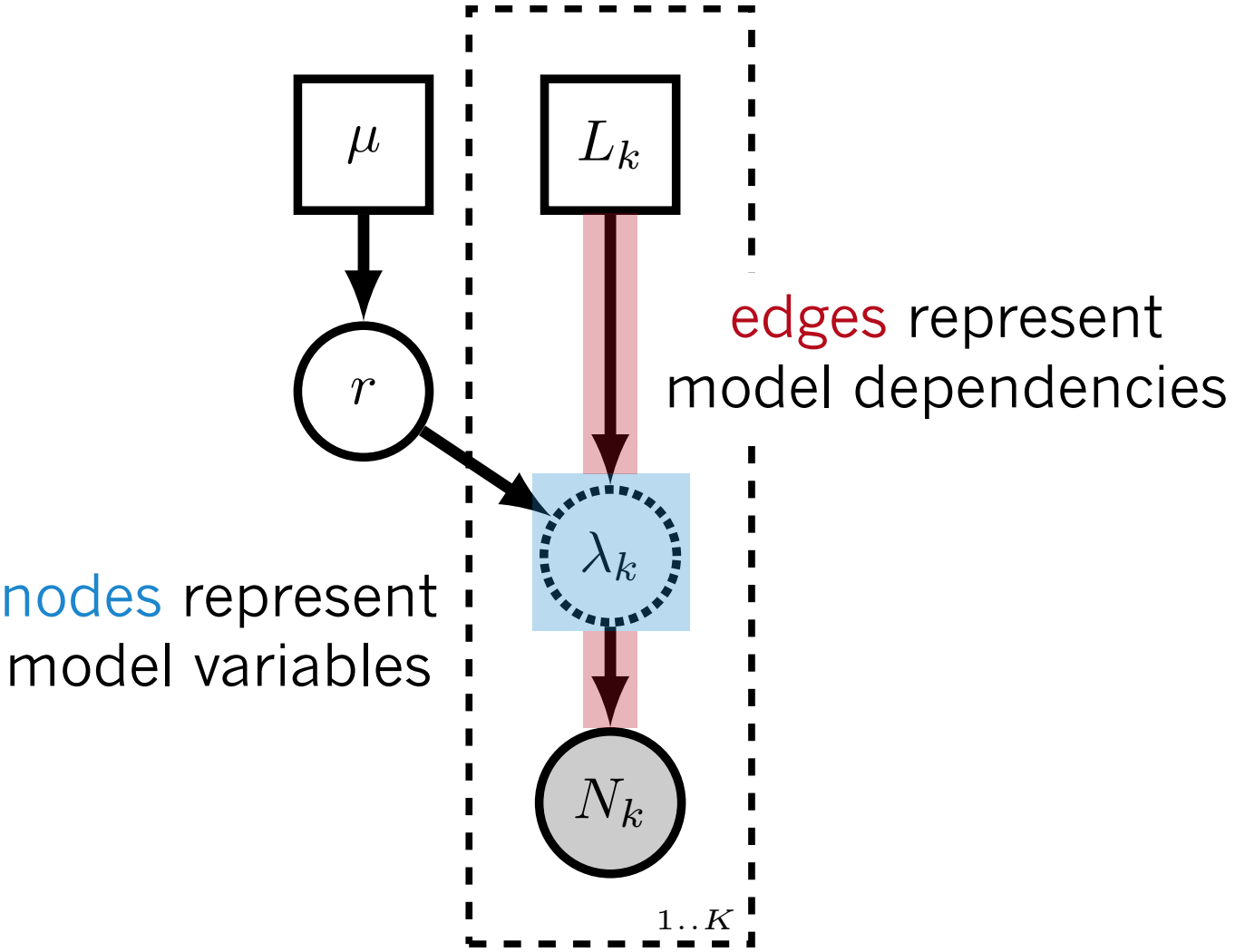
probabilistic graphical model (PGM)

a probabilistic model that uses a graph to encode its variables and their dependencies

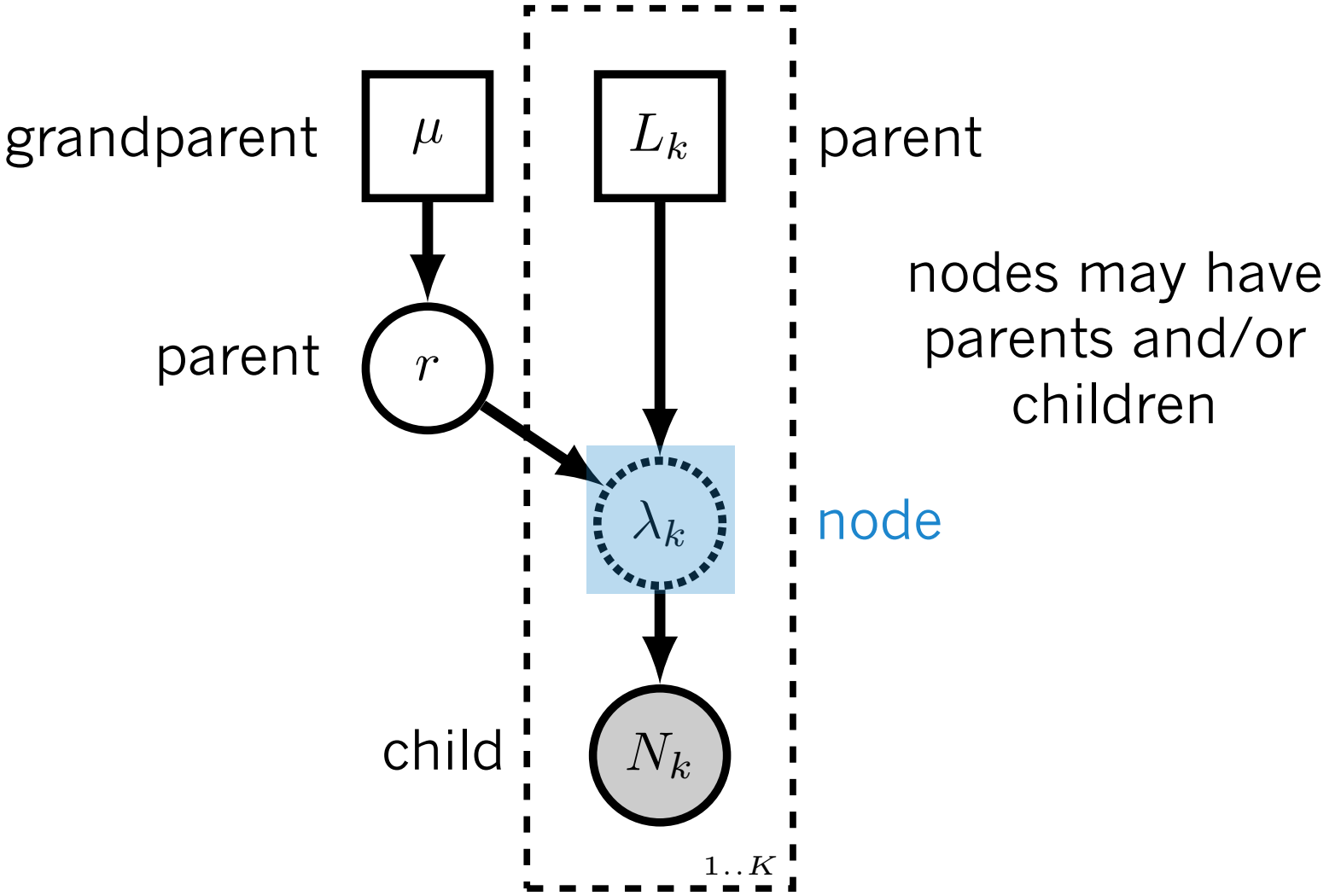
Uses

- Algorithm design
- Model design
- Blueprinting ideas
- Communication
- Teaching

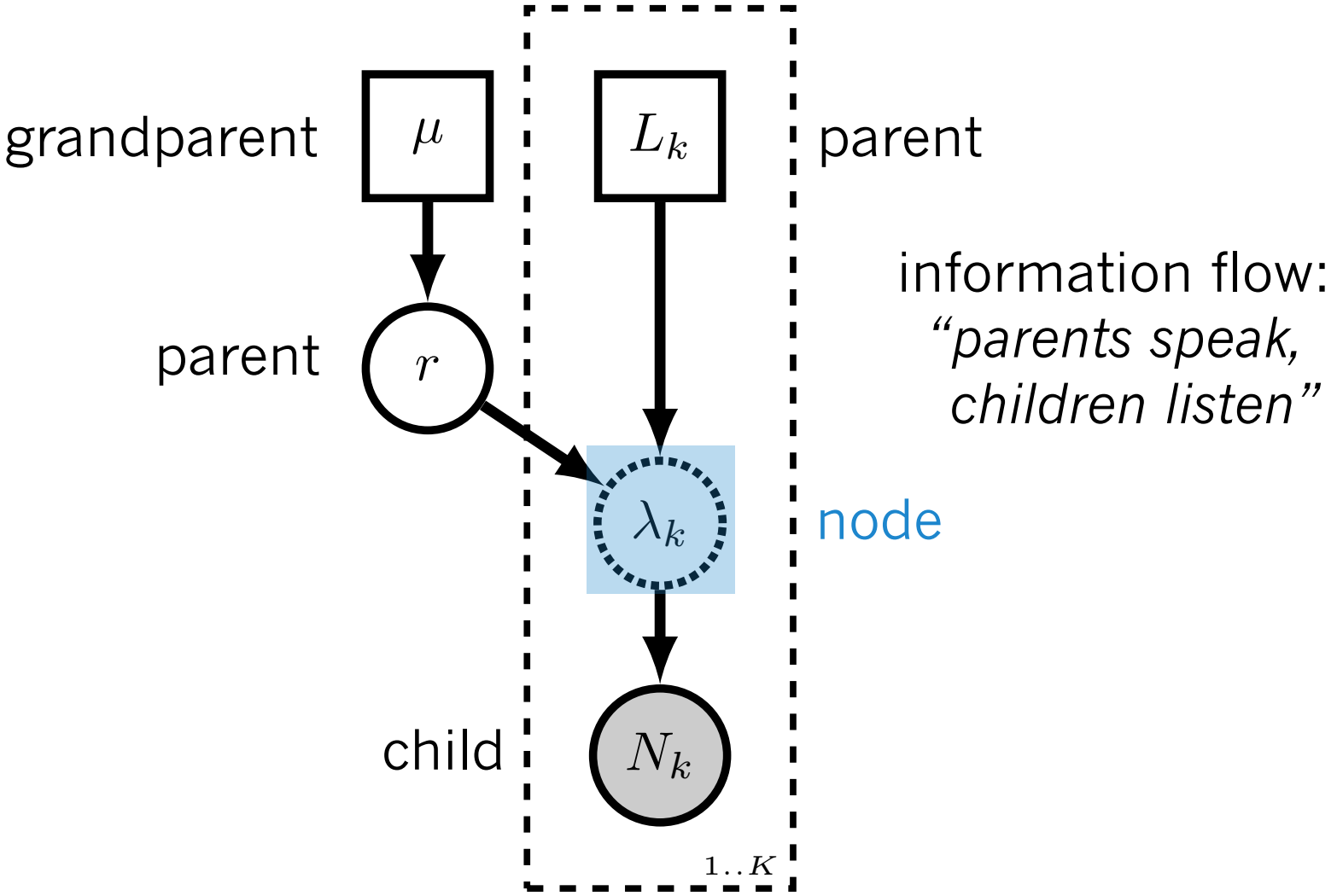
PGM anatomy


μ
L_k
r
λ_k
N_k
$1..K$
edges represent
model dependencies
nodes represent
model variables

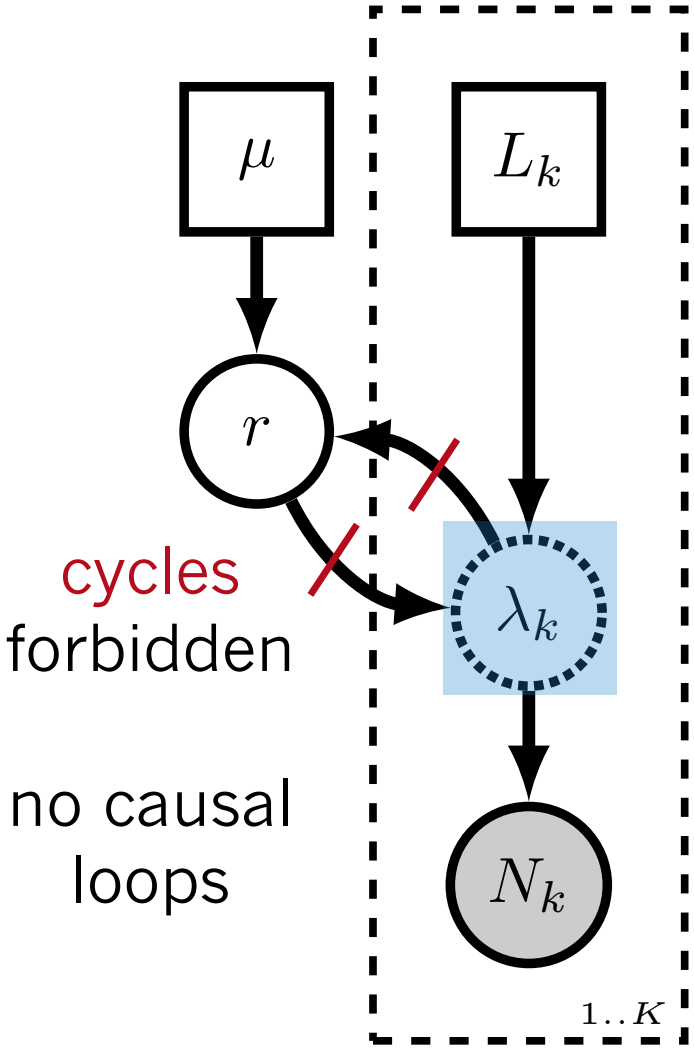
PGM anatomy


grandparent
μ
L_k
parent
parent
r
nodes may have parents and/or children
λ_k
node
child
N_k
$1..K$

PGM anatomy


grandparent
μ
L_k
parent
parent
r
information flow:
"parents speak, children listen"
λ_k
node
child
N_k
$1..K$

PGM anatomy



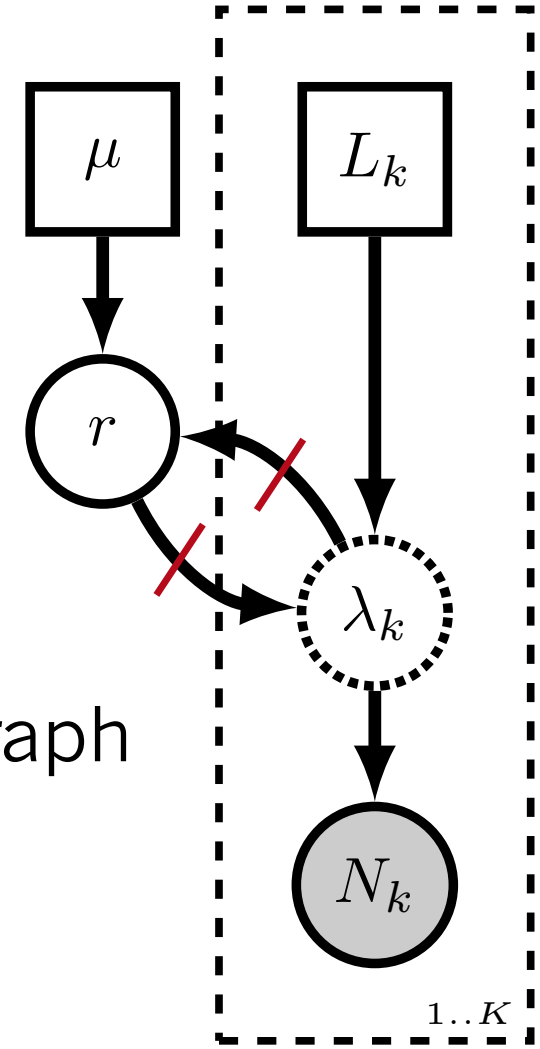
cycles forbidden

no causal loops

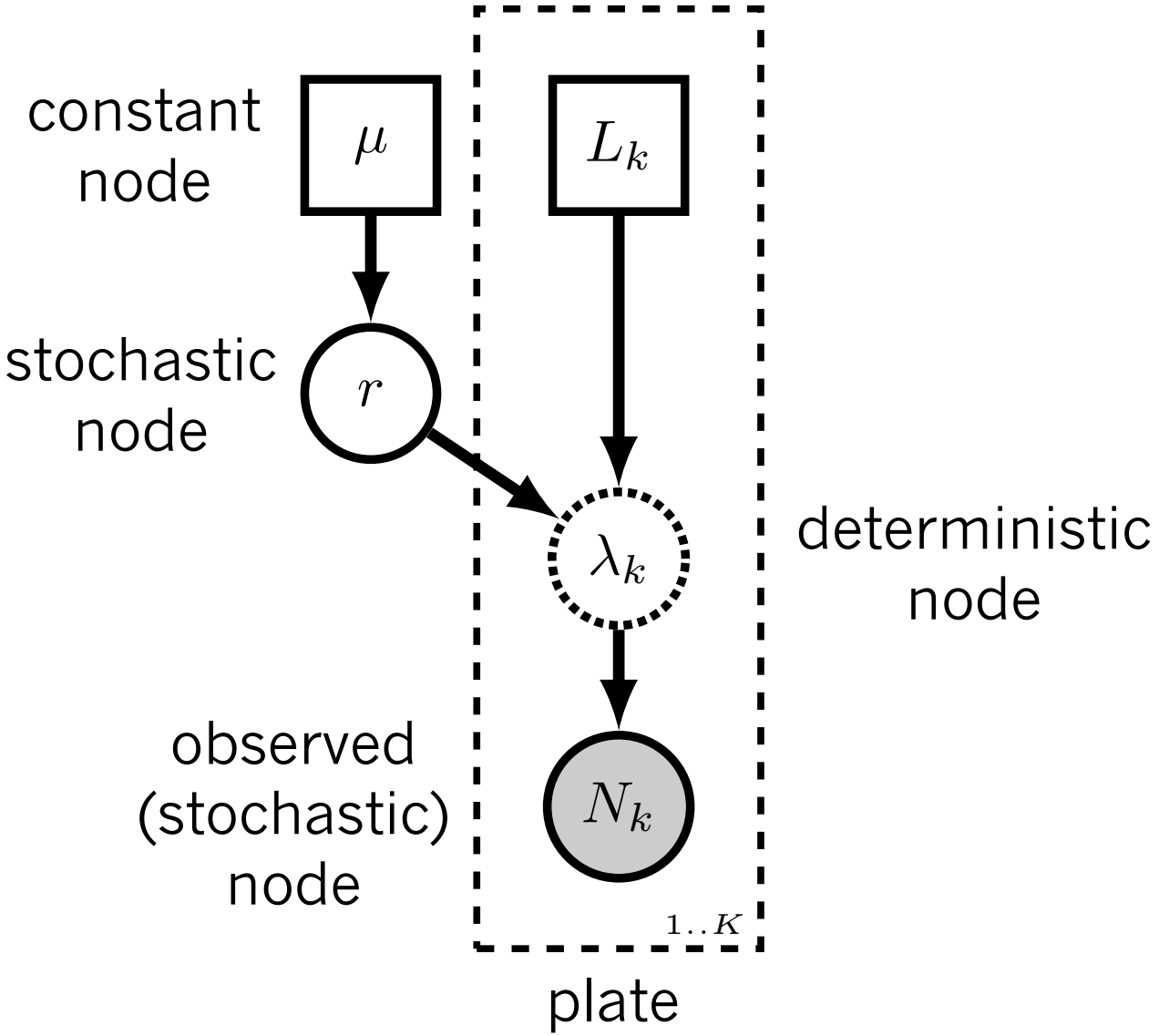
node and grandchild and grandparent (!?)

PGM anatomy

Directed acyclic graph (DAG)



PGM anatomy

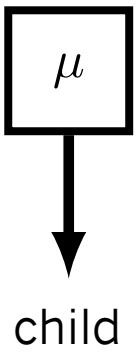

constant node
μ
L_k
stochastic node
r
λ_k
deterministic node
observed (stochastic) node
N_k
$1..K$
plate

Constant node

"equals"
fixed value
asserted or known

the rate for a prior

$\mu = 1$



Constant node

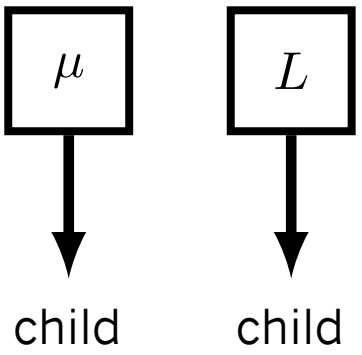
"equals"
fixed value
asserted or known

the rate for a prior

$\mu = 1$

the length of a locus

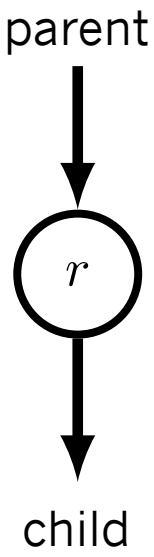
$L = 1030$



Stochastic node

"distributed by"
dynamic value
learned

per-site mutation rate
$r \sim \text{Exponential}(\mu)$

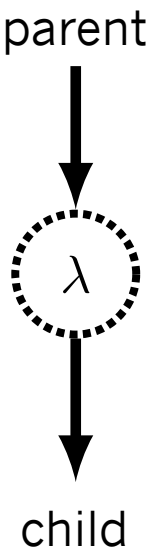


Deterministic node

"determined by"
dynamic value
learned

per-locus mutation rate

$\lambda := L \times r$



Observed (stochastic) node

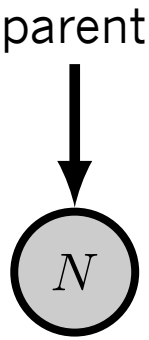
"distributed by"

observed value

fixed (at observation)

de novo mutations (1 locus)

$N \sim \text{Poisson}(\lambda)$



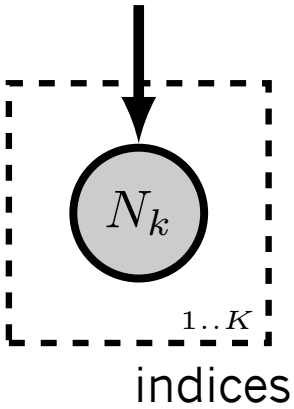
Plates

repeats model structure

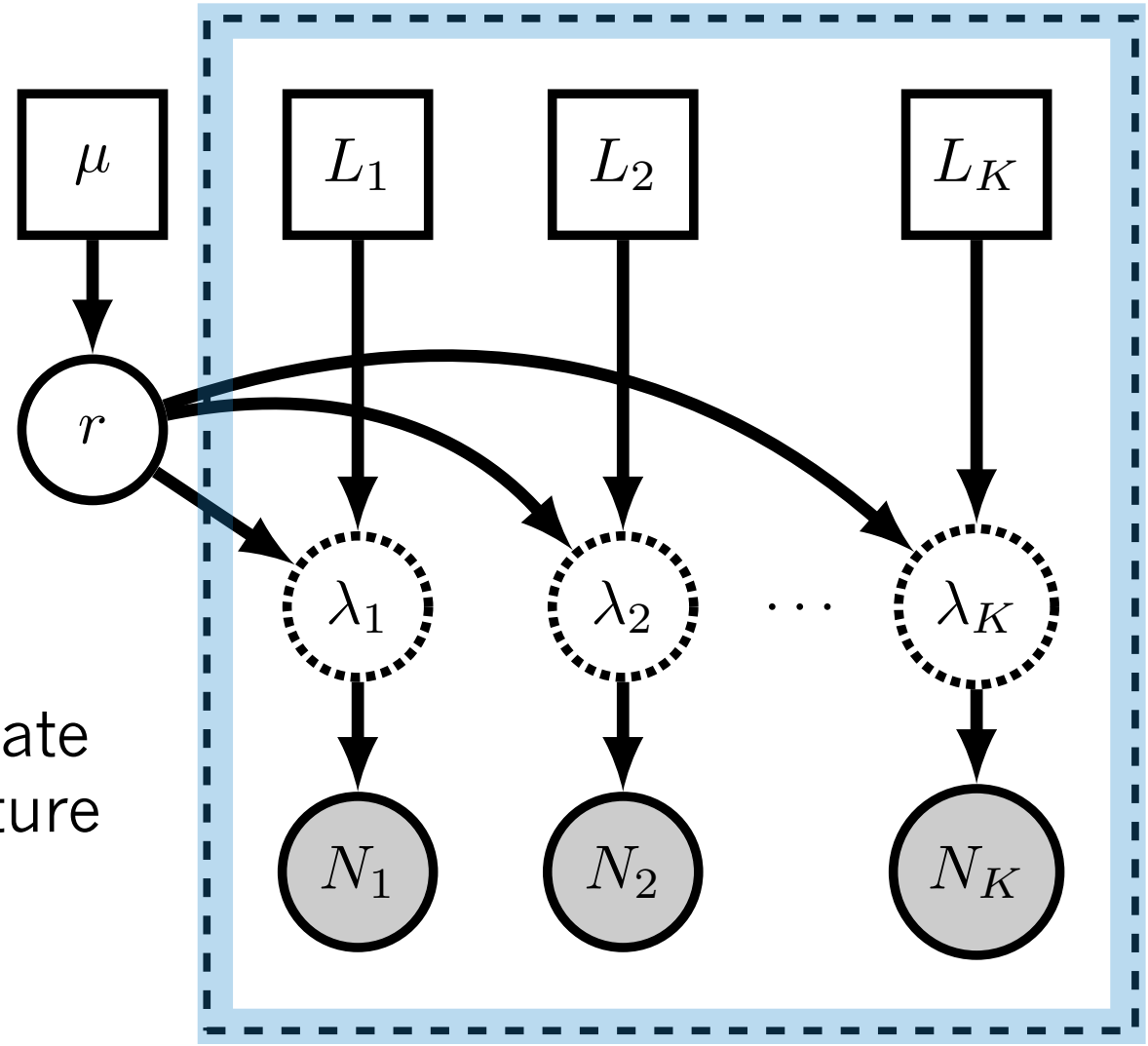
simplifies graph

de novo mutations (K loci)

$N_k \sim \text{Poisson}(\lambda_k)$

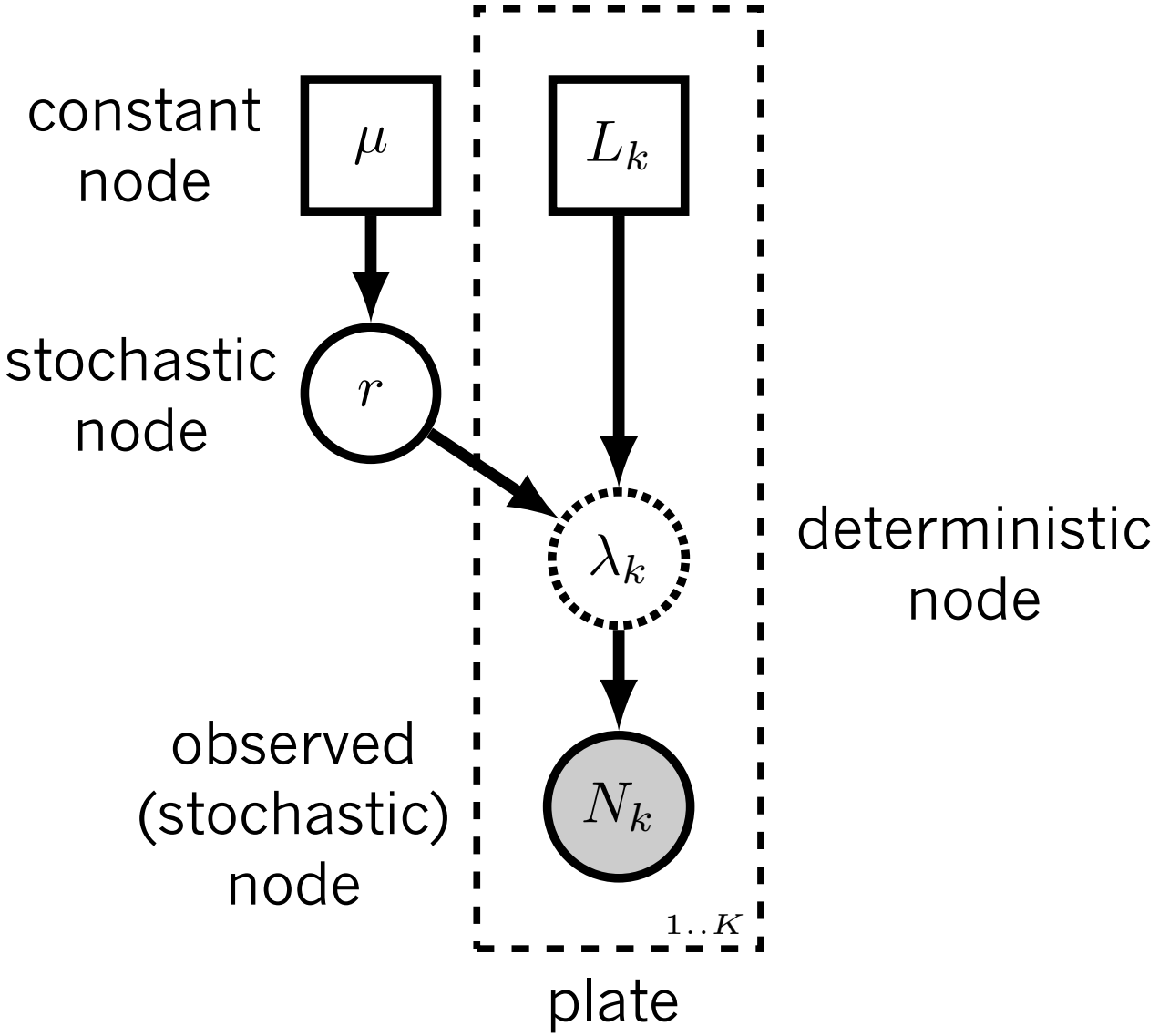


PGM anatomy

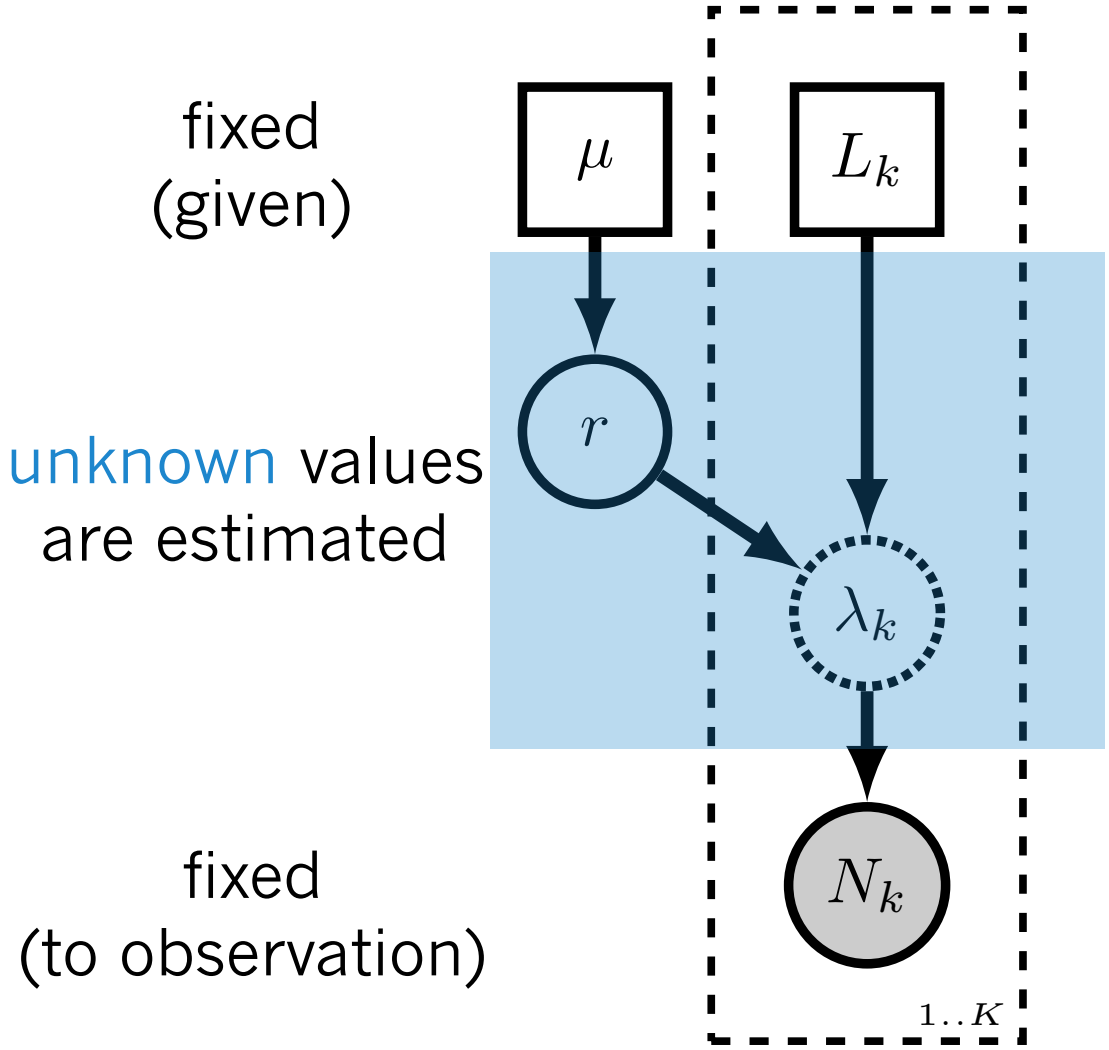


plates replicate model structure

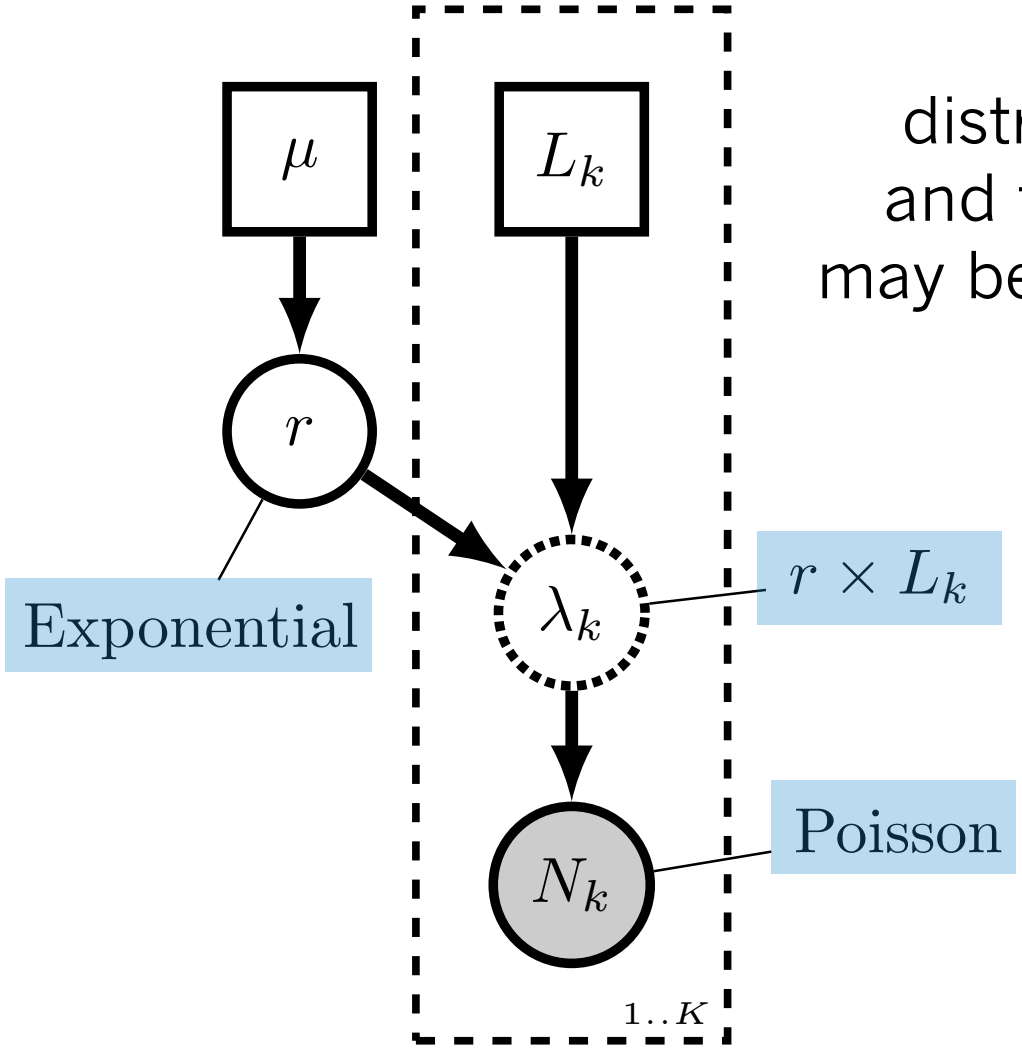
PGM anatomy


constant node
μ
L_k
stochastic node
r
λ_k
deterministic node
observed (stochastic) node
N_k
$1..K$
plate

PGM anatomy

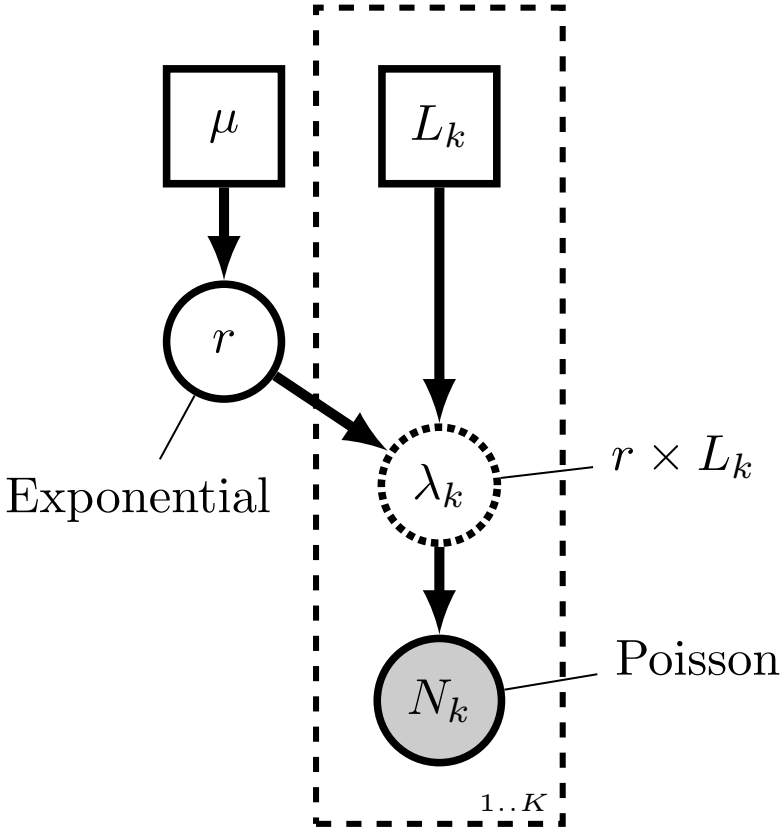

fixed
(given)
unknown values
are estimated
fixed
(to observation)
μ
L_k
r
λ_k
N_k
$1..K$

PGM anatomy



distributions
and functions
may be annotated

Example of computing model probability using PGM

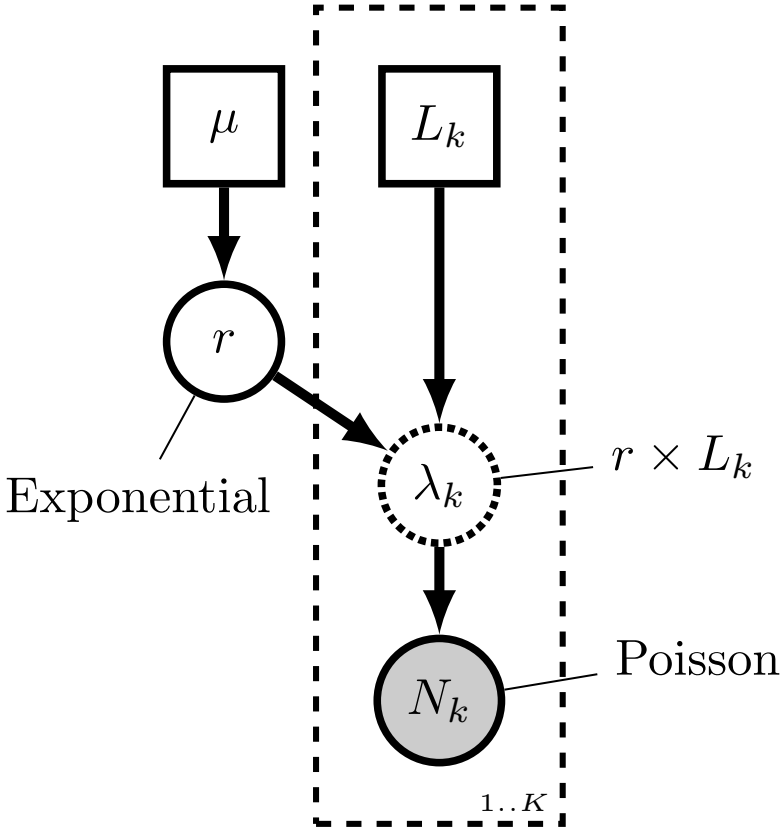


One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

r = ???

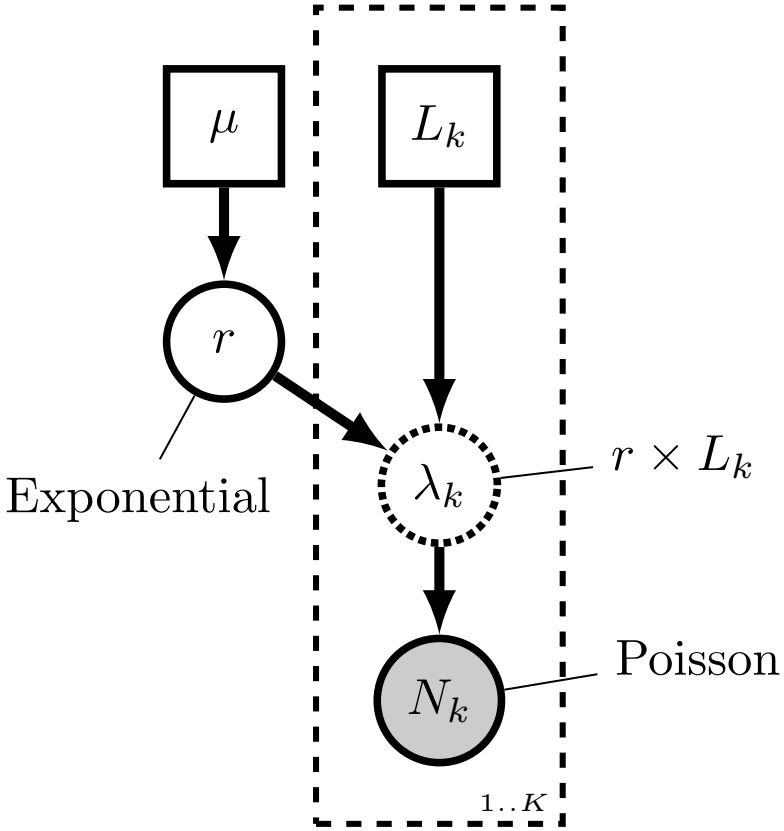


One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

r = 3.0×10^{-7}



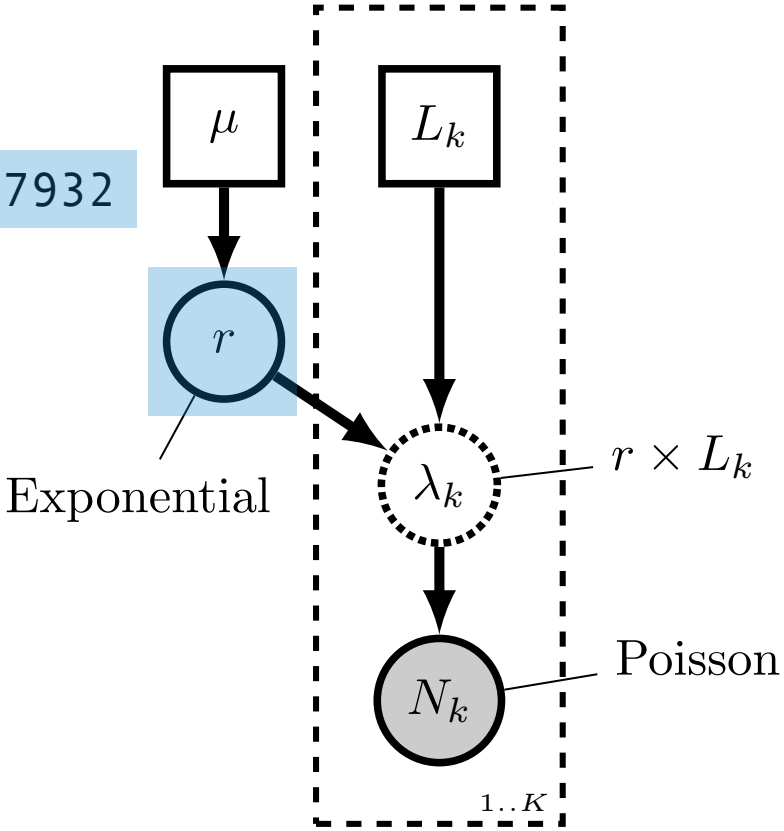
One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

r = 3.0×10^{-7}

$\ln \Pr(r=3.0*10^{-7} \mid \mu=1/10^{-8}) =$ -11.57932



One locus (K=1)

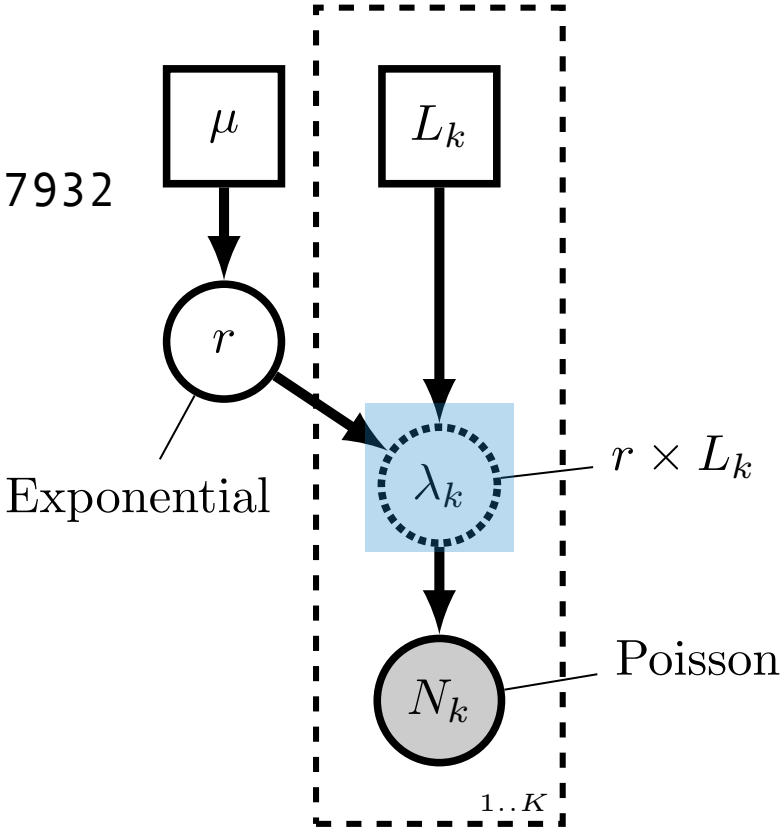
N_1 = 3 mutations

L_1 = 1030bp

r = 3.0×10^{-7}

$\ln \Pr(r=3.0*10^{-7} \mid \mu=1/10^{-8}) = -11.57932$

$\lambda_1 = 3.0*10^{-7} * 1030 = 0.0003090$



One locus (K=1)

N_1 = 3 mutations

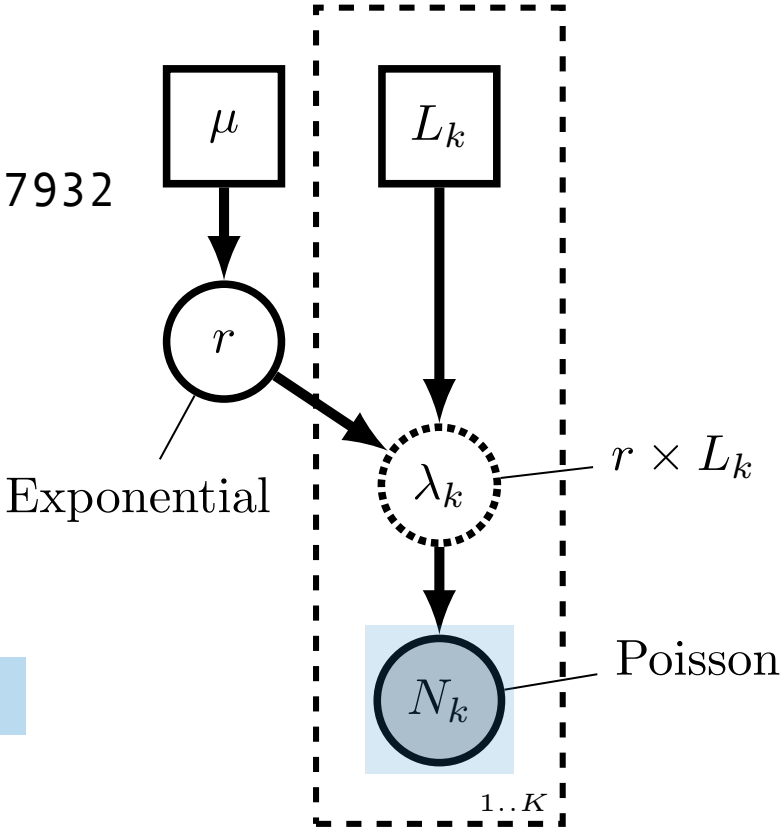
L_1 = 1030bp

r = 3.0×10^{-7}

$\ln \Pr(r=3.0*10^{-7} \mid \mu=1/10^{-8}) = -11.57932$

$\lambda_1 = 3.0*10^{-7} * 1030 = 0.0003090$

$\ln \Pr(N_1=3 \mid \lambda_1=0.0003090) =$ -26.0386



One locus (K=1)

N_1 = 3 mutations

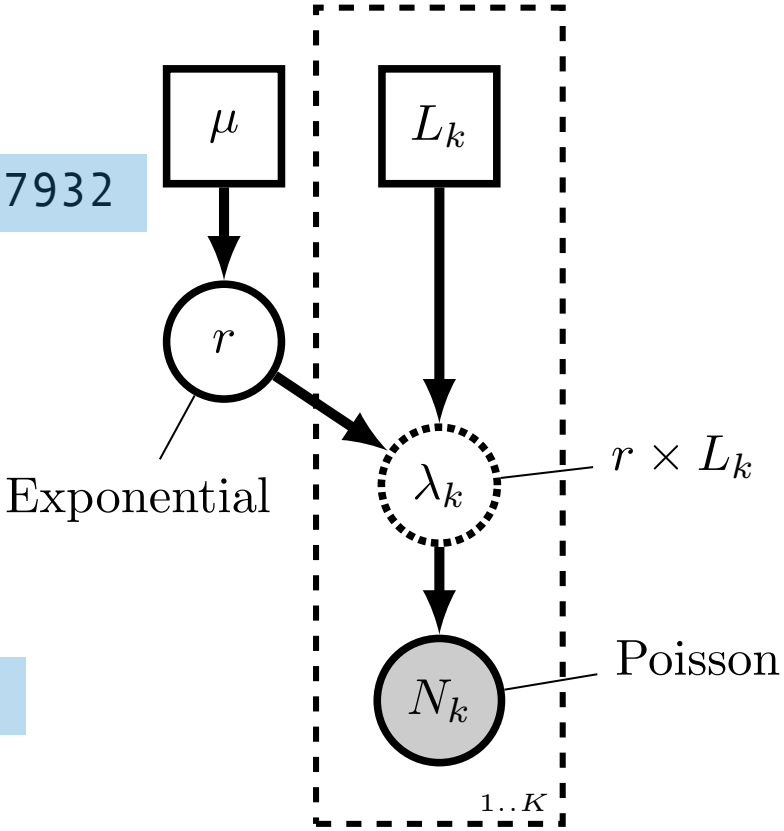
L_1 = 1030bp

r = 3.0×10^{-7}

$\ln \Pr(r=3.0*10^{-7} \mid \mu=1/10^{-8}) = -11.57932$

$\lambda_1 = 3.0*10^{-7} * 1030 = 0.0003090$

$\ln \Pr(N_1=3 \mid \lambda_1=0.0003090) = -26.0386$



$\ln \Pr(N_1=3, r=3.0*10^{-7} \mid \lambda_1=0.0003090, \mu=1/10^{-8}) = -37.6179$

model log-probability

One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

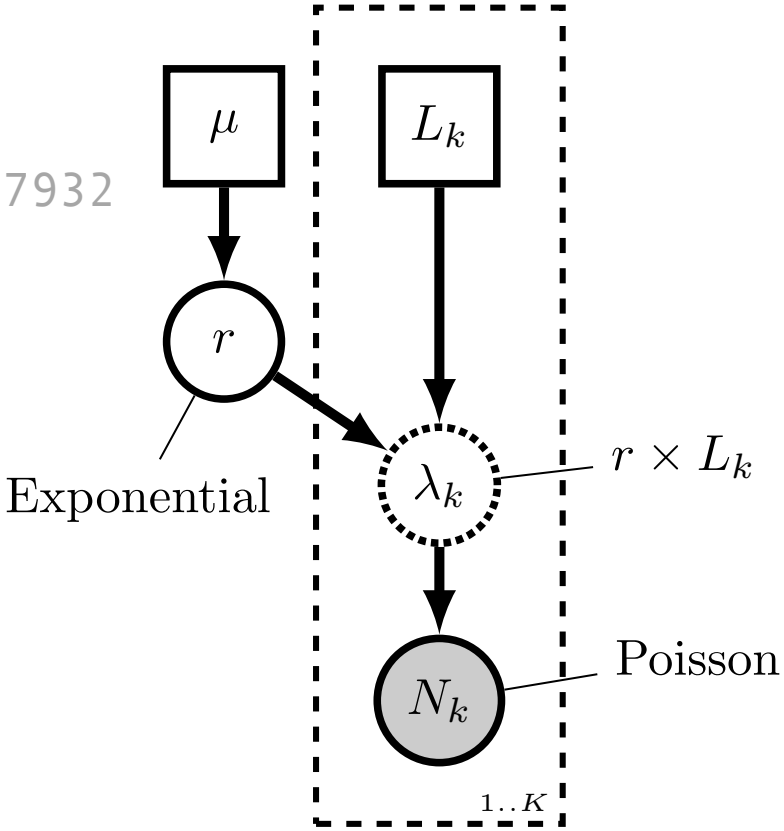
r = 2.7×10^{-7}

$\ln \Pr(r=3.0*10^{-7} \mid \mu=1/10^{-8}) = -11.57932$

$\lambda_1 = 3.0*10^{-7} * 1030 = 0.0003090$

$\ln \Pr(N_1=3 \mid \lambda_1=0.0003090) = -26.0386$

$\ln \Pr(N_1=3, r=3.0*10^{-7} \mid \lambda_1=0.0003090, \mu=1/10^{-8}) = -37.6179$



One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

r = 2.7×10^{-7}

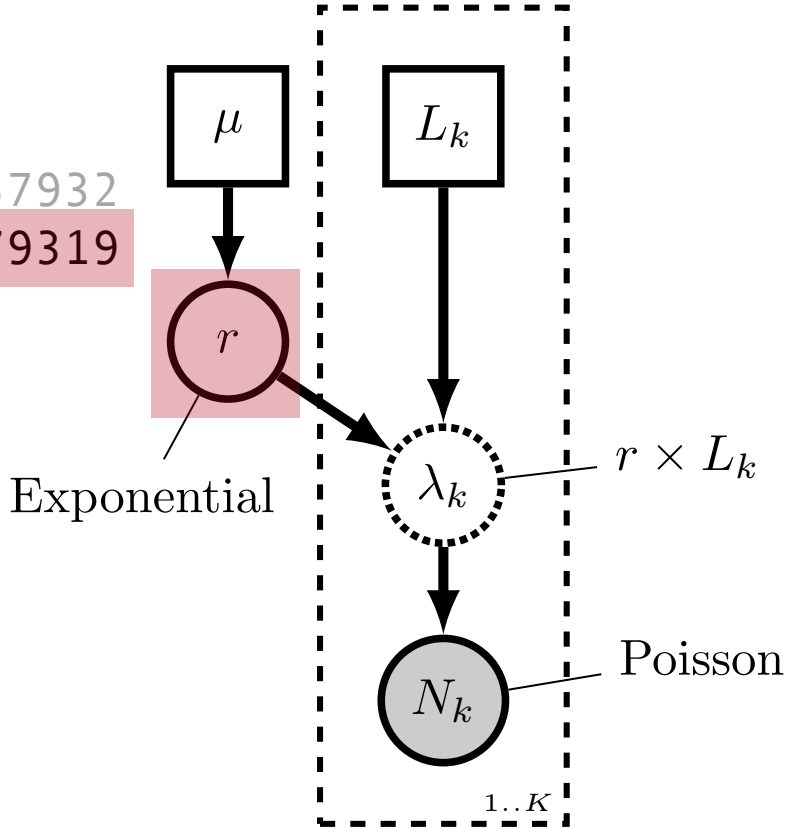
$\ln \Pr(r=3.0*10^{-7} \mid \mu=1/10^{-8}) = -11.57932$

$\ln \Pr(r=2.7*10^{-7} \mid \mu=1/10^{-8}) = -8.579319$

$\lambda_1 = 3.0*10^{-7} * 1030 = 0.0003090$

$\ln \Pr(N_1=3 \mid \lambda_1=0.0003090) = -26.0386$

$\ln \Pr(N_1=3, r=3.0*10^{-7} \mid \lambda_1=0.0003090, \mu=1/10^{-8}) = -37.6179$



One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

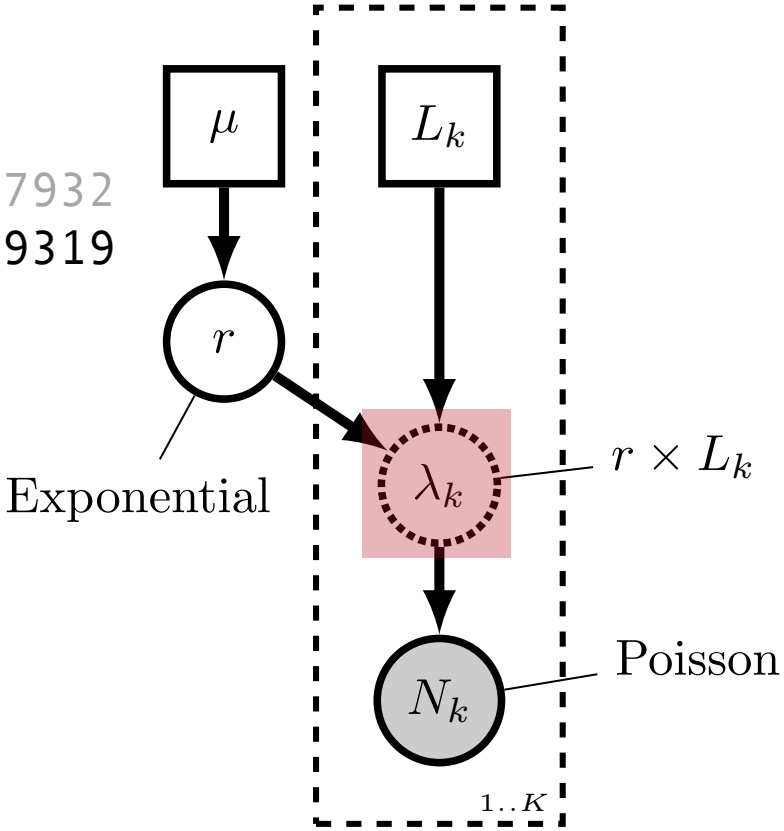
r = 2.7 x 10^{-7}

$\ln \Pr(r=3.0*10^{-7} \mid \mu=1/10^{-8}) = -11.57932$

$\ln \Pr(r=2.7*10^{-7} \mid \mu=1/10^{-8}) = -8.579319$

$\lambda_1 = 3.0*10^{-7} * 1030 = 0.0003090$

$\lambda_1 = 2.7*10^{-7} * 1030 = 0.0002781$



$\ln \Pr(N_1=3 \mid \lambda_1=0.0003090) = -26.0386$

$\ln \Pr(N_1=3, r=3.0*10^{-7} \mid \lambda_1=0.0003090, \mu=1/10^{-8}) = -37.6179$

One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

r = 2.7 x 10^{-7}

ln Pr(r=3.0*10^{-7} | μ=1/10^{-8}) = -11.57932

ln Pr(r=2.7*10^{-7} | μ=1/10^{-8}) = -8.579319

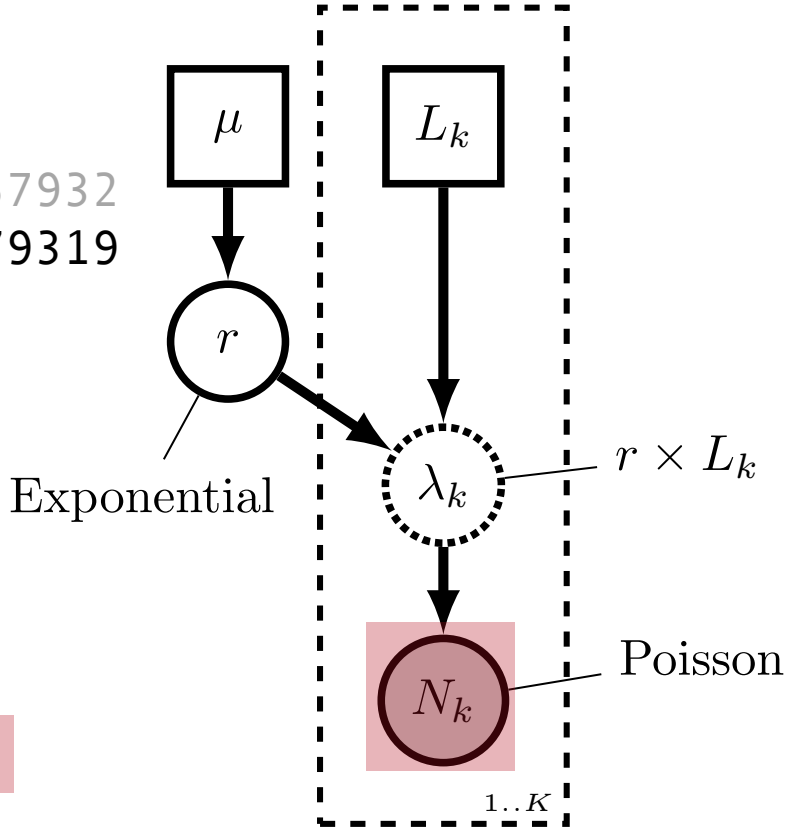
λ_1 = 3.0*10^{-7} * 1030 = 0.0003090

λ_1 = 2.7*10^{-7} * 1030 = 0.0002781

ln Pr(N_1=3 | λ_1=0.0003090) = -26.0386

ln Pr(N_1=3 | λ_1=0.0002781) = -26.3546

ln Pr(N_1=3, r=3.0*10^{-7} | λ_1=0.0003090, μ=1/10^{-8}) = -37.6179

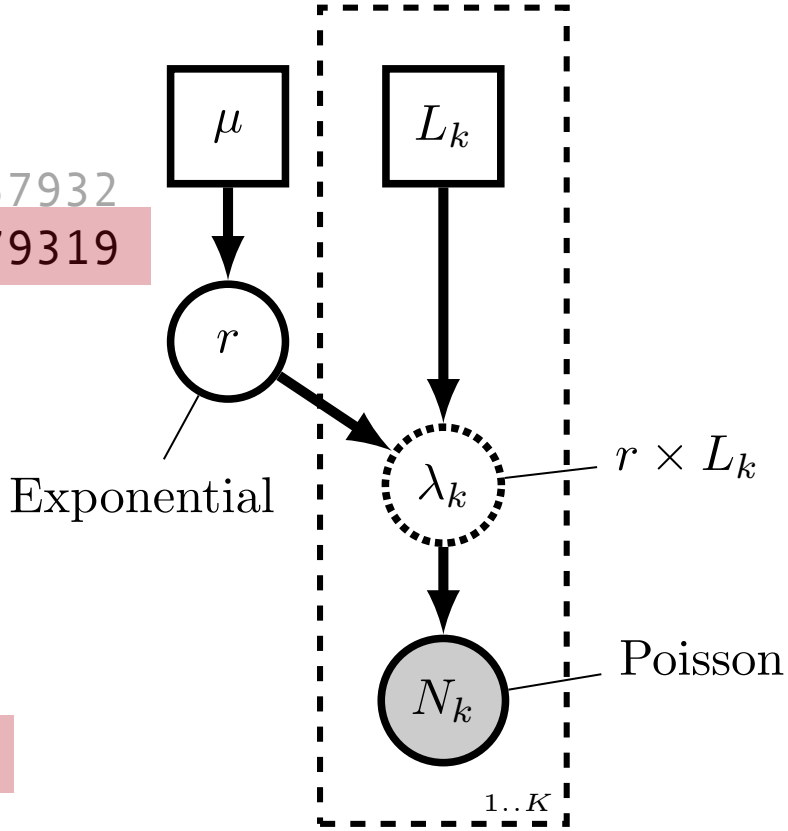


One locus (K=1)

N_1 = 3 mutations

L_1 = 1030bp

r = 2.7 x 10^{-7}



ln Pr(r=3.0*10^{-7} | μ=1/10^{-8}) = -11.57932

ln Pr(r=2.7*10^{-7} | μ=1/10^{-8}) = -8.579319

λ_1 = 3.0*10^{-7} * 1030 = 0.0003090

λ_1 = 2.7*10^{-7} * 1030 = 0.0002781

ln Pr(N_1=3 | λ_1=0.0003090) = -26.0386

ln Pr(N_1=3 | λ_1=0.0002781) = -26.3546

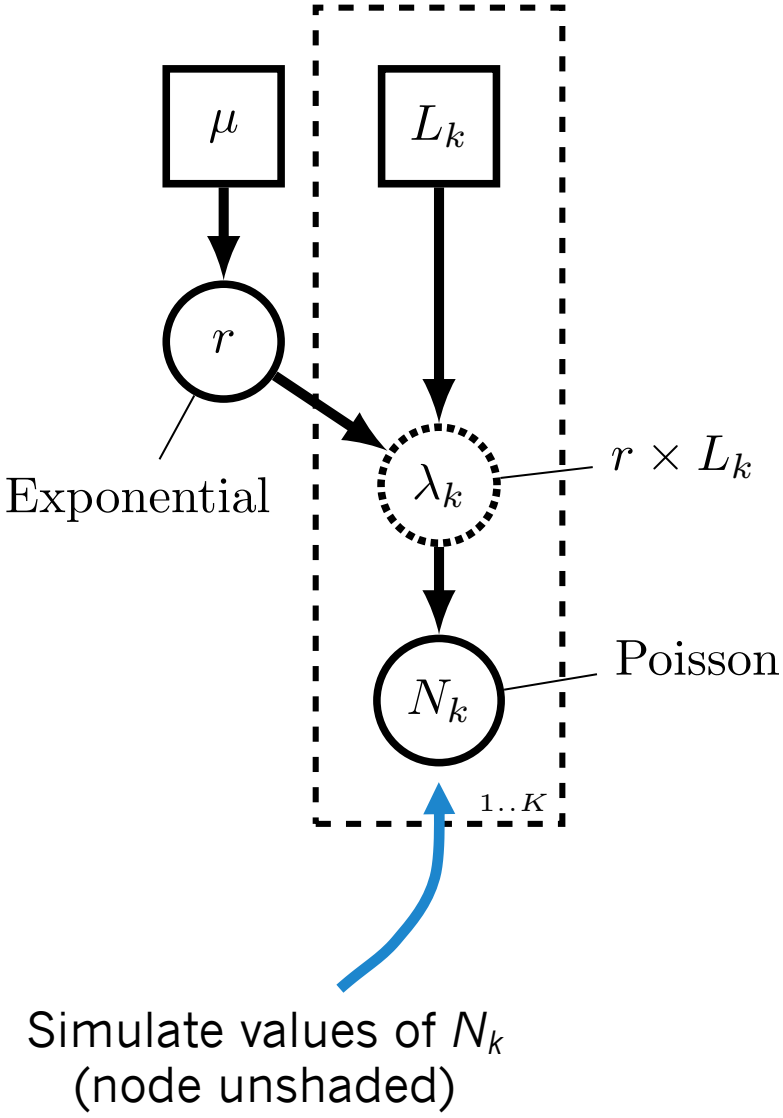
ln Pr(N_1=3, r=3.0*10^{-7} | λ_1=0.0003090, μ=1/10^{-8}) = -37.6179

ln Pr(N_1=3, r=2.7*10^{-7} | λ_1=0.0002781, μ=1/10^{-8}) = -34.9340

new model log-probability

Generative model

use same PGM both to compute model likelihood and to simulate data

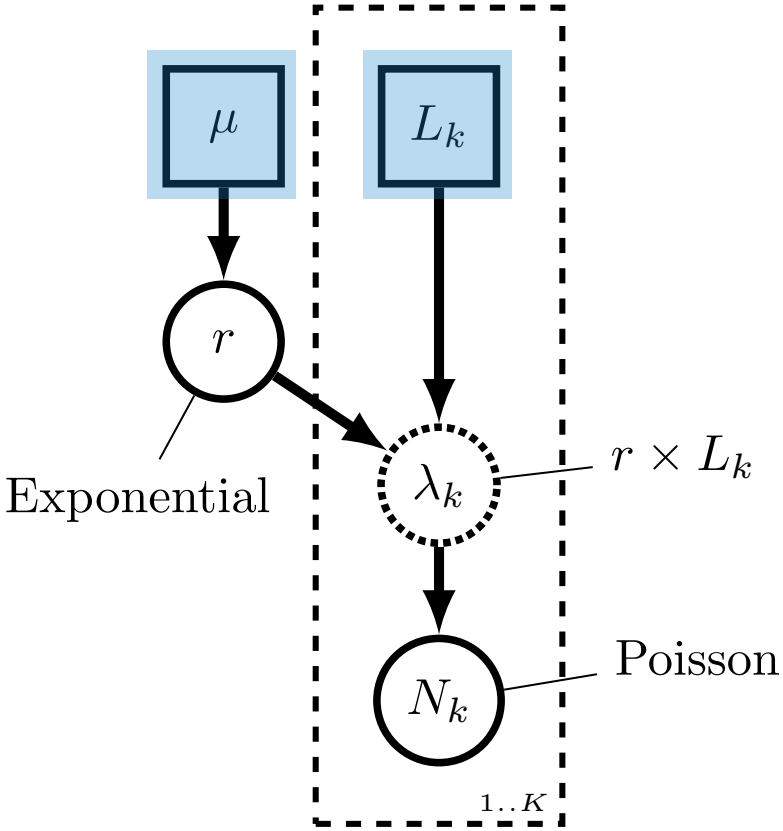


Simulation using DAG structure

use parent values to draw child values

Set initial conditions (for locus 1)

- $\mu = 1/10^{-8}$
- $L_1 = 1030$ bp



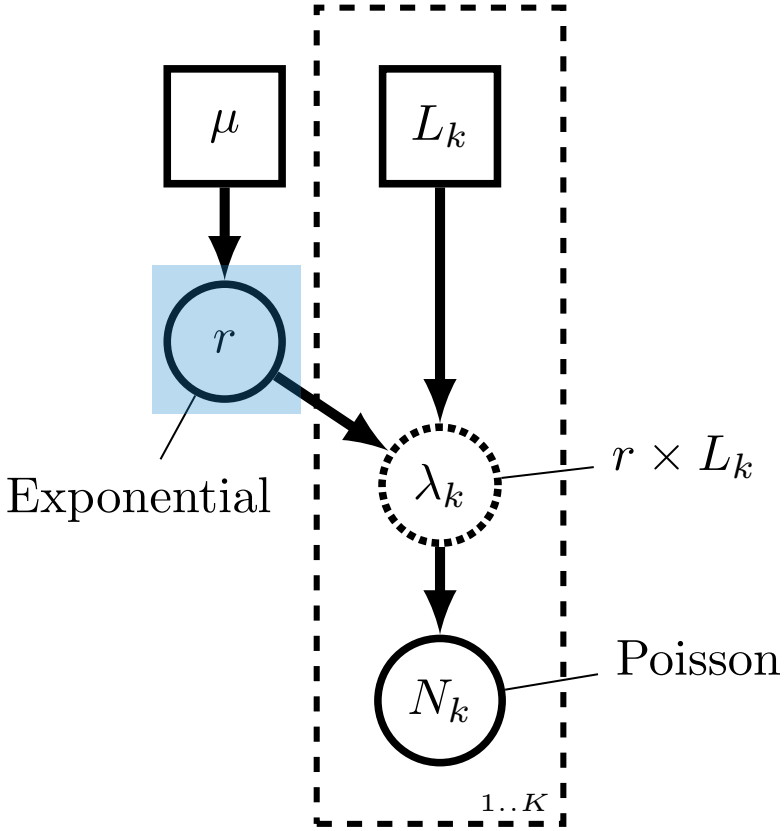
Simulation using DAG structure
use parent values to draw child values

Set initial conditions (for locus 1)

μ=1/10^{-8}

L_1=1030 bp

Sample r = 2.5*10^{-7} from Exp(1/10^{-8})



Simulation using DAG structure

use parent values to draw child values

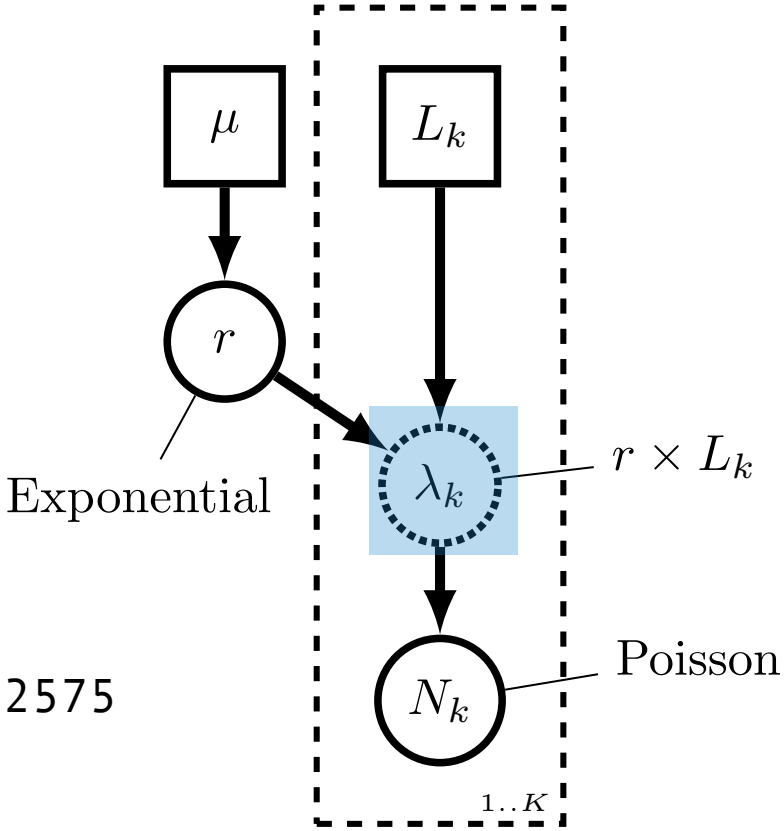
Set initial conditions (for locus 1)

$\mu = 1/10^{-8}$

$L_1 = 1030$ bp

Sample $r = 2.5*10^{-7}$ from $\text{Exp}(1/10^{-8})$

Determine $\lambda_1 = 1030 * 2.5*10^{-7} = 0.0002575$



Simulation using DAG structure

use parent values to draw child values

Set initial conditions (for locus 1)

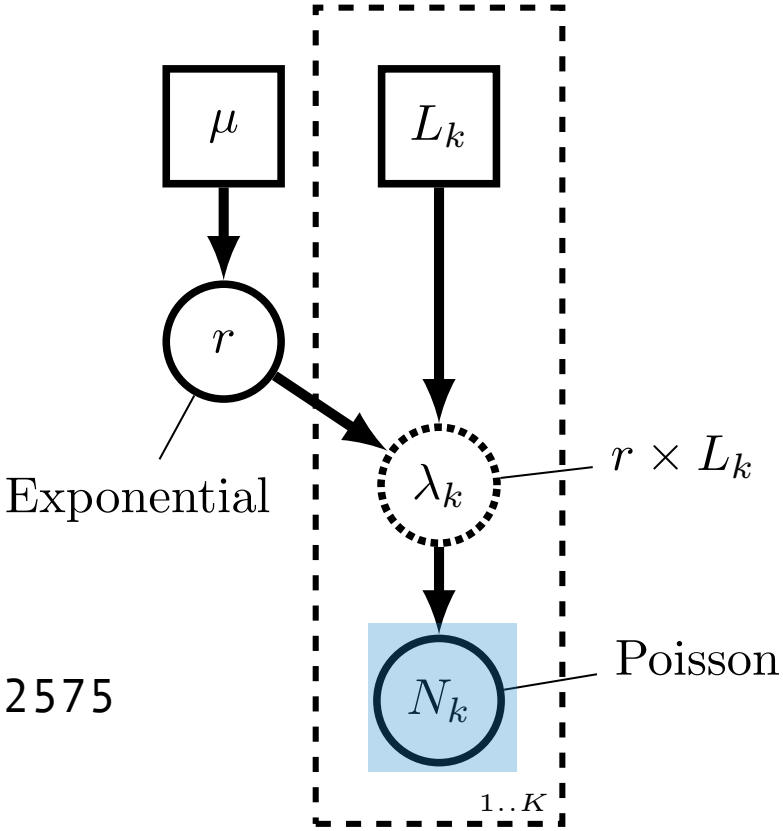
μ=1/10^{-8}

L_1=1030 bp

Sample r = 2.5*10^{-7} from Exp(1/10^{-8})

Determine λ_1 = 1030 * 2.5*10^{-7} = 0.0002575

Sample N_1 = 0 from Poisson(0.0002575)



Simulation using DAG structure
use parent values to draw child values

Set initial conditions (for locus 2)

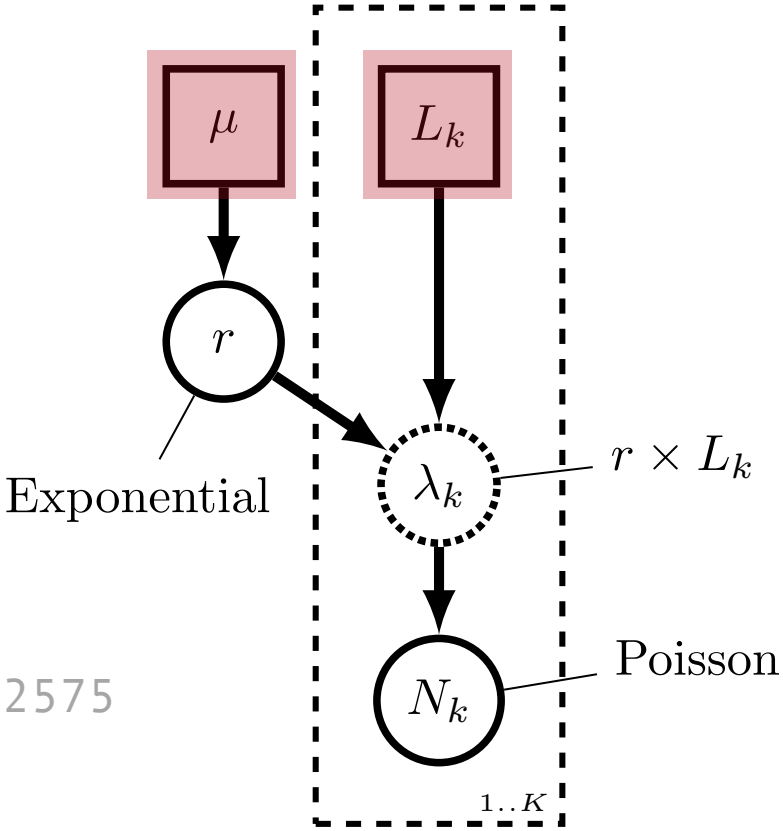
$\mu=1/10^{-8}$ $\mu=1/10^{-8}$

$L_1=1030$ bp $L_2=3.2*10^9$ bp

Sample $r = 2.5*10^{-7}$ from Exp($1/10^{-8}$)

Determine $\lambda_1 = 1030 * 2.5*10^{-7} = 0.0002575$

Sample $N_1 = 0$ from Poisson(0.0002575)



Simulation using DAG structure
use parent values to draw child values

Set initial conditions (for locus 2)

$\mu = 1/10^{-8}$ $\mu = 1/10^{-8}$

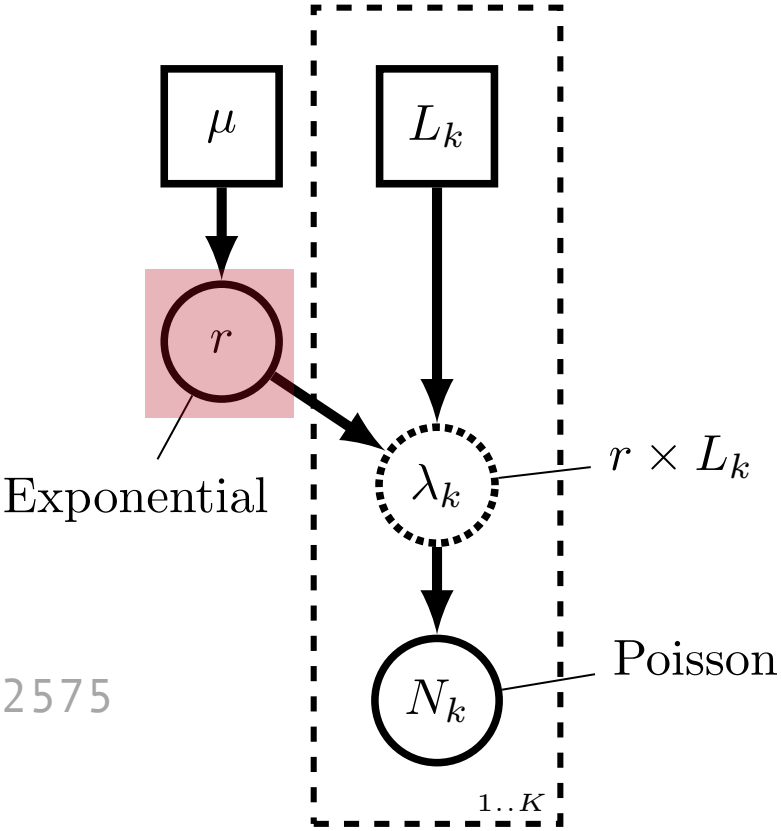
$L_1 = 1030$ bp $L_2 = 3.2 * 10^9$ bp

Sample $r = 2.5 * 10^{-7}$ from $Exp(1/10^{-8})$

Sample $r = 1.2 * 10^{-8}$ from $Exp(1/10^{-8})$

Determine $\lambda_1 = 1030 * 2.5 * 10^{-7} = 0.0002575$

Sample $N_1 = 0$ from $Poisson(0.0002575)$

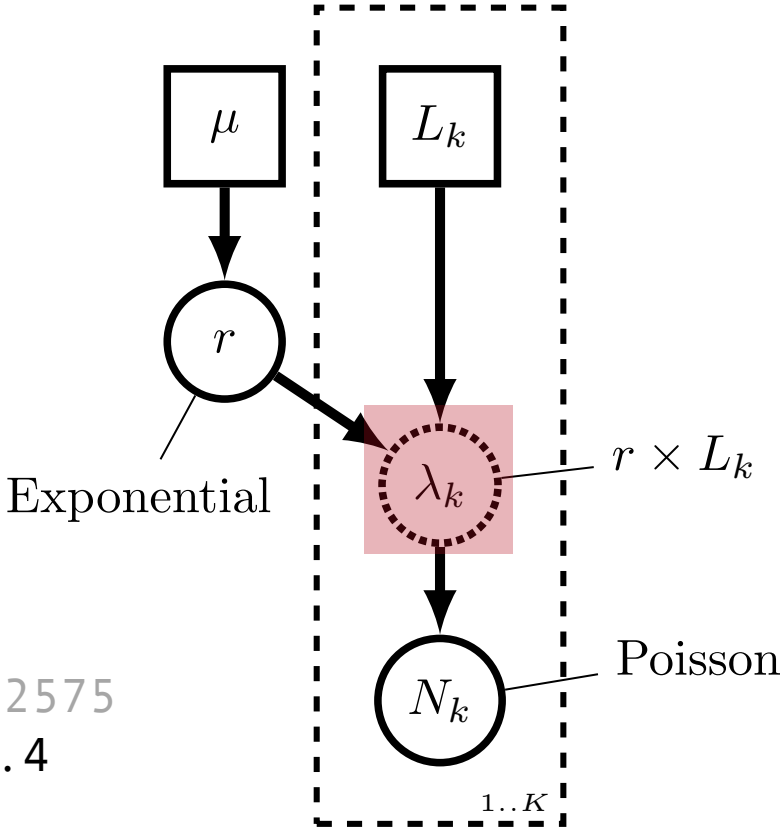


Simulation using DAG structure
use parent values to draw child values

Set initial conditions (for locus 2)

$\mu = 1/10^{-8}$ $\mu = 1/10^{-8}$

$L_1 = 1030$ bp $L_2 = 3.2*10^9$ bp



Sample $r = 2.5*10^{-7}$ from Exp($1/10^{-8}$)

Sample $r = 1.2*10^{-8}$ from Exp($1/10^{-8}$)

Determine $\lambda_1 = 1030 * 2.5*10^{-7} = 0.0002575$

Determine $\lambda_2 = 3.2*10^9 * 1.2*10^{-8} = 38.4$

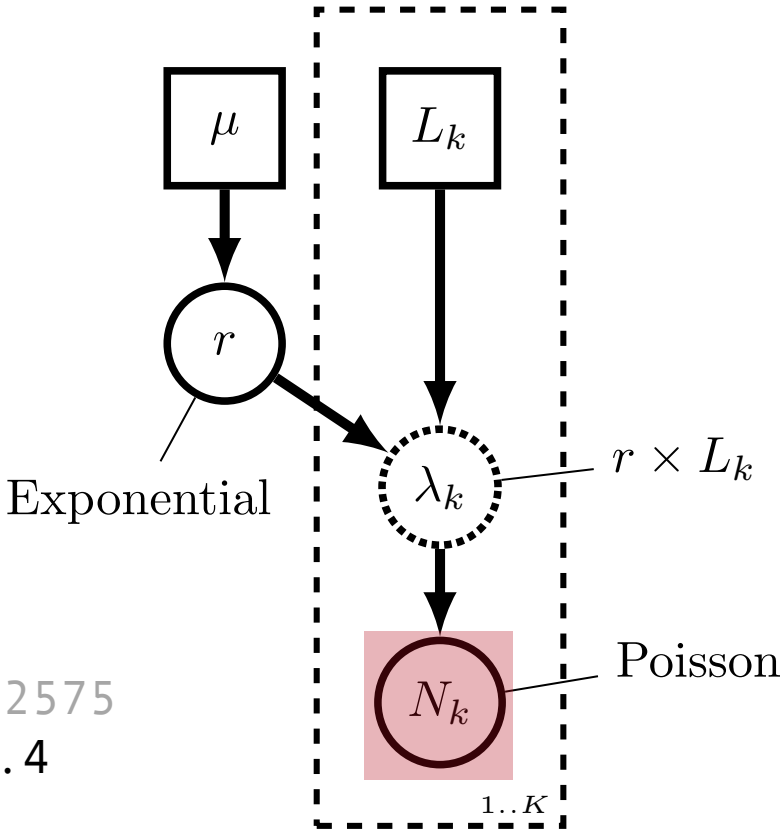
Sample $N_1 = 0$ from Poisson(0.0002575)

Simulation using DAG structure
use parent values to draw child values

Set initial conditions (for locus 2)

$\mu=1/10^{-8}$ $\mu=1/10^{-8}$

$L_1=1030$ bp $L_2=3.2*10^9$ bp



Sample $r = 2.5*10^{-7}$ from Exp($1/10^{-8}$)

Sample $r = 1.2*10^{-8}$ from Exp($1/10^{-8}$)

Determine $\lambda_1 = 1030 * 2.5*10^{-7} = 0.0002575$

Determine $\lambda_2 = 3.2*10^9 * 1.2*10^{-8} = 38.4$

Sample $N_1 = 0$ from Poisson(0.0002575)

Sample $N_2 = 44$ from Poisson(38.4)

Simulation experiment

Biological data is messy, limited

Simulated data is perfect, abundant

Simulation experiment

Biological data is messy, limited

Simulated data is perfect, abundant

A model that performs poorly with simulated data generally performs worse with biological data!

Simulation experiment

Biological data is messy, limited

Simulated data is perfect, abundant

A model that performs poorly with simulated data generally performs worse with biological data!

Simulation experiment

1. Analyze simulated data under controlled settings
   vary dataset size, parameters, models, etc.
2. Translate simulation results to guide biological analyses

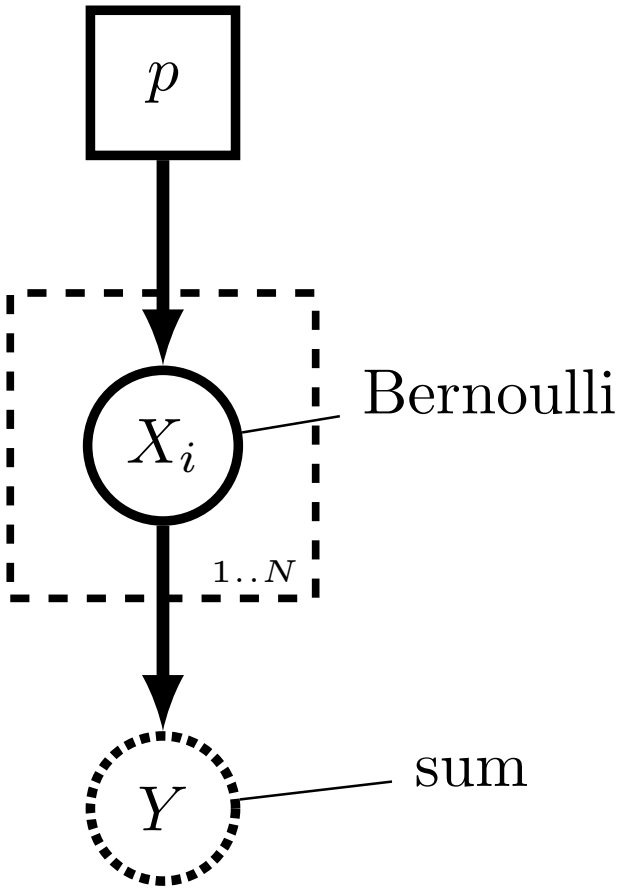
(more later)

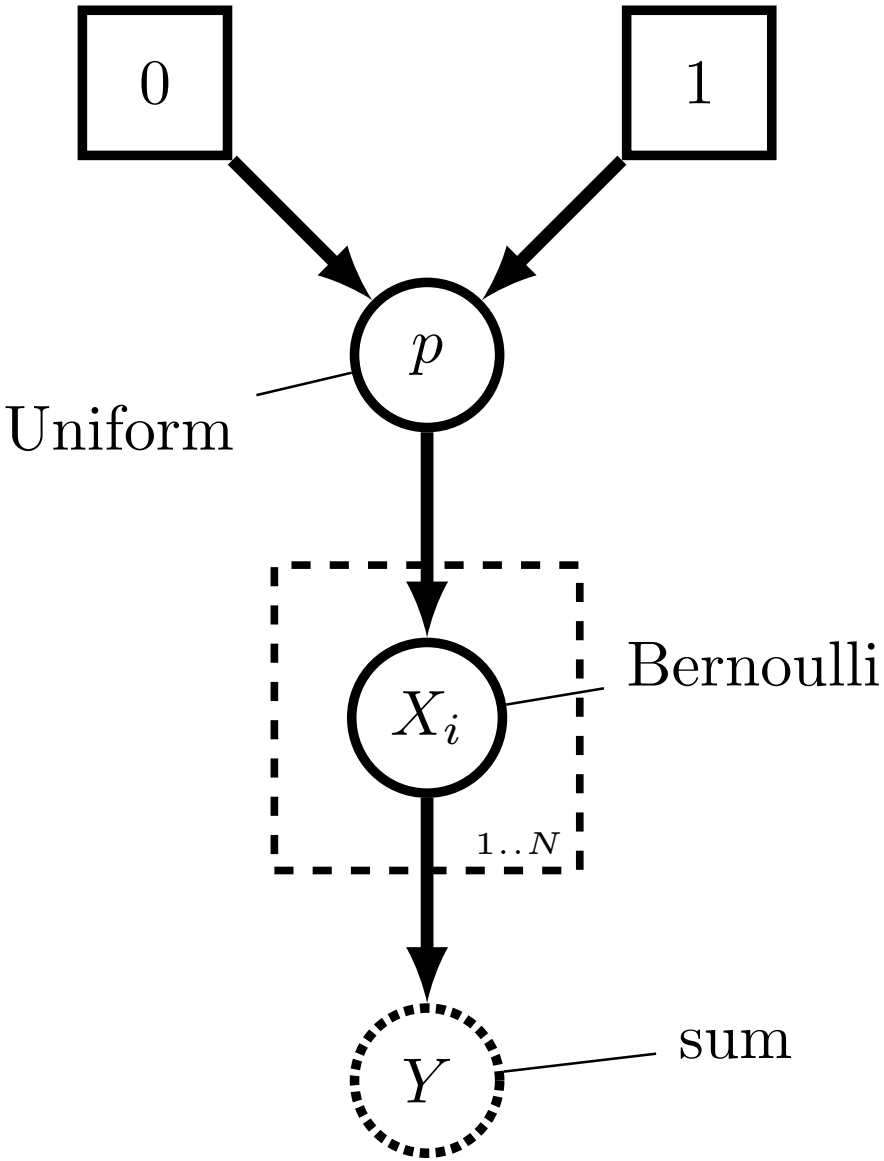
Draw a graphical model for

$$p = 0.5$$
$$X_i \sim \text{Bernoulli}(p)$$
$$Y = X_1 + X_2 + \ldots + X_N$$

Extra challenge

Treat p as stochastic node that takes any value from 0 to 1 with equal probability


p
X_i
Bernoulli
$1..N$
Y
sum

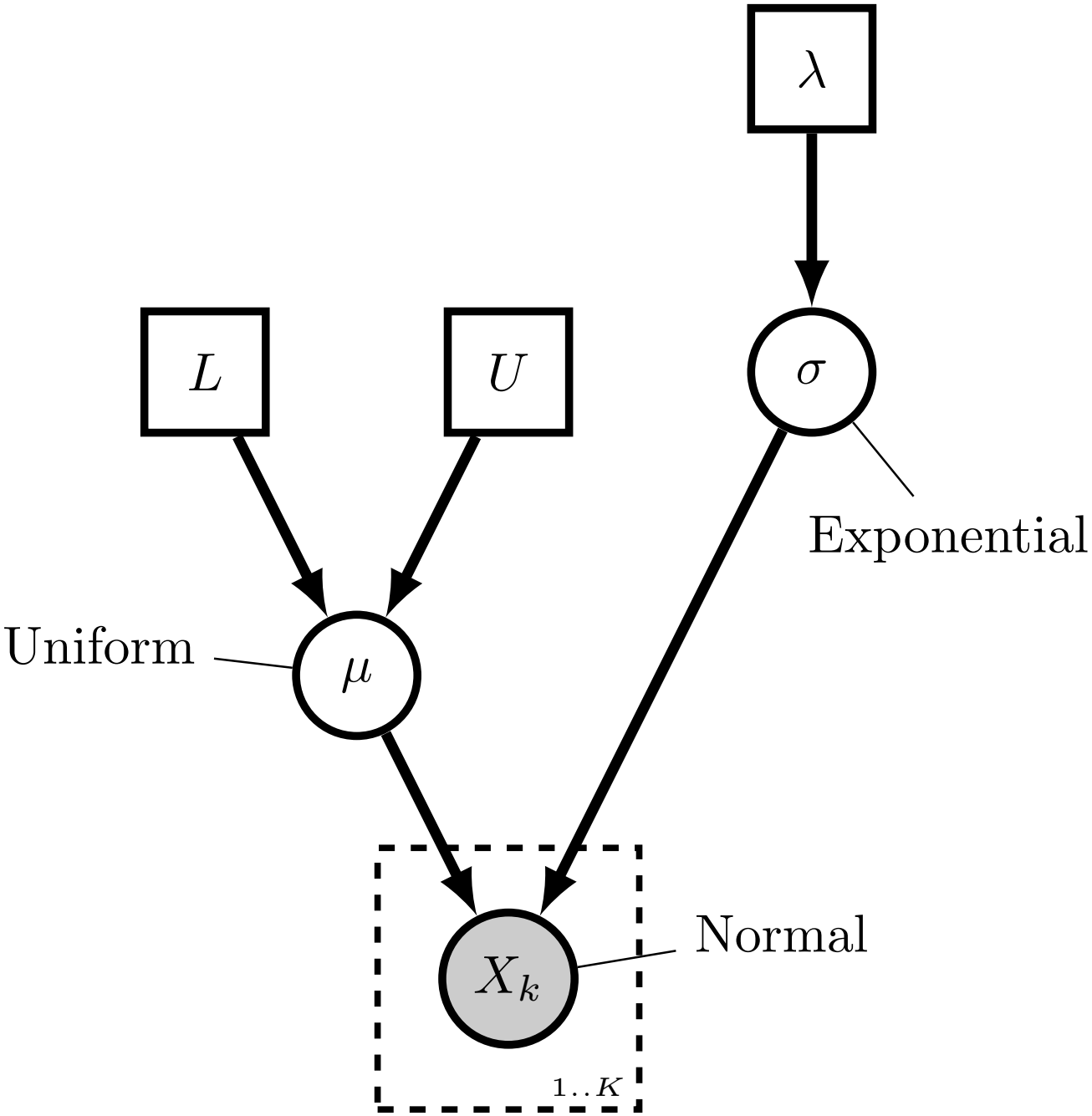

0
1
p
Uniform
X_i
Bernoulli
$1..N$
Y
sum

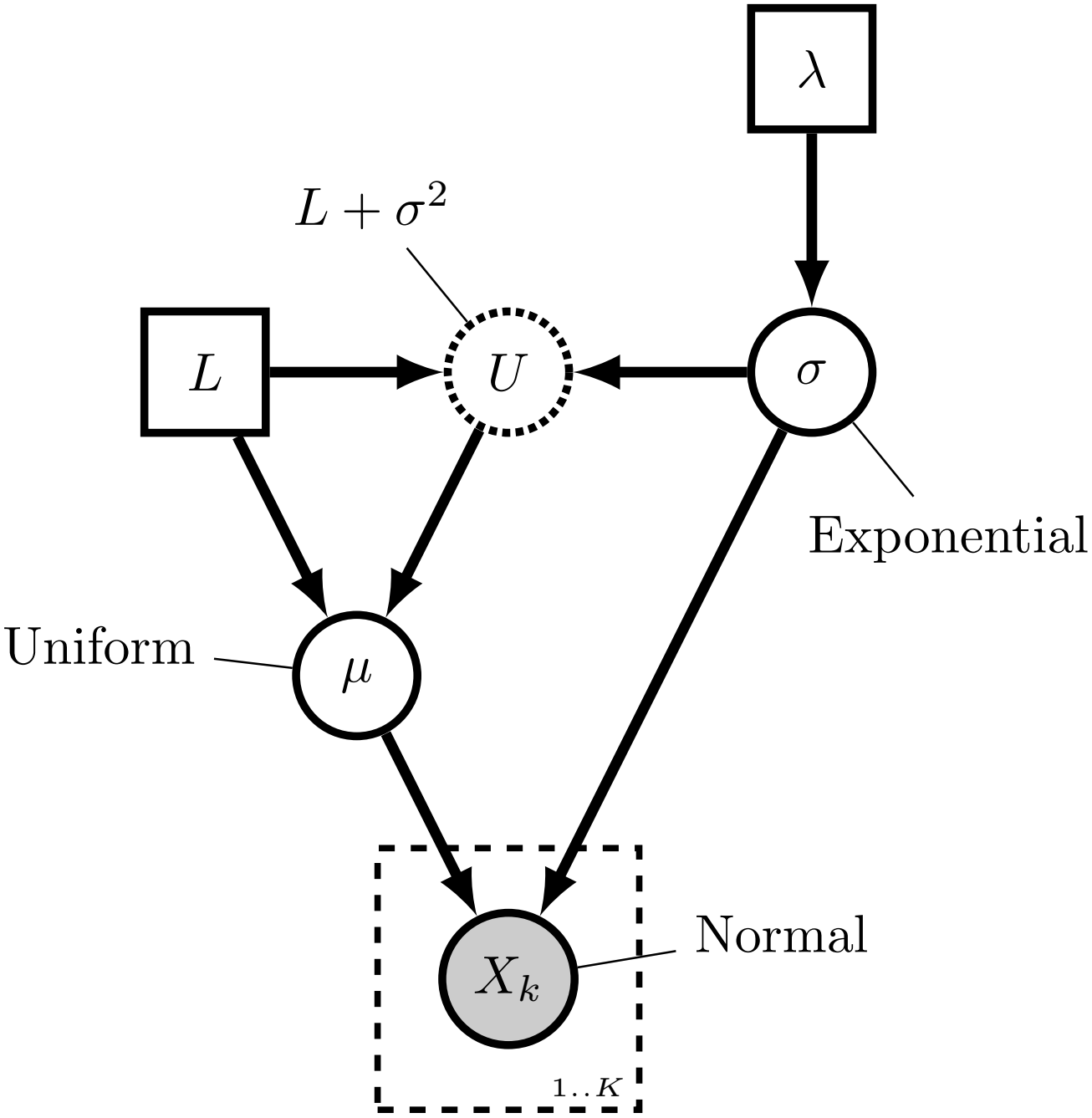
Draw a graphical model for

$\lambda = 1$
$L = 3$
$U = 10$
$\sigma \sim \text{Exponential}(\lambda)$
$\mu \sim \text{Uniform}(L, U)$
$X_k \sim \text{Normal}(\mu, \sigma)$ – *observed data*

Extra challenge

Modify U so its value determined by $L+\sigma^2$


λ
L
U
σ
Exponential
Uniform
μ
X_k
Normal
$1..K$


λ
$L + \sigma^2$
L
U
σ
Exponential
Uniform
μ
X_k
Normal
$1..K$

Model definitions

Conceptual

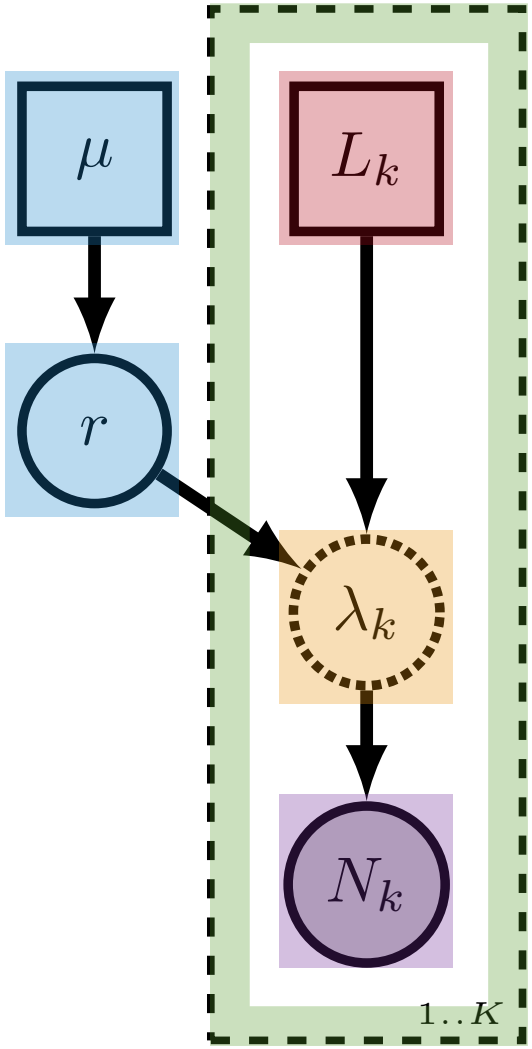
Probabilistic

Graphical

Computational

Computational

```
# input
data = [   0,   3,    3 ]
L    = [ 310, 780, 1030 ]

# per-site mutation rate
mu = 1e-8
r_site ~ dnExponential(1/mu)

# mutation counts per locus
K = data.size()
for (k in 1:K) {
    r_loci[k] := r_site * L[k]
    n_muts[k] ~ dnPoisson(r_loci[k])
    n_muts[k].clamp( data[k] )
}
```



RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language

Sebastian Höhna ✉, Michael J. Landis, Tracy A. Heath, Bastien Boussau, Nicolas Lartillot, Brian R. Moore, John P. Huelsenbeck, Fredrik Ronquist Author Notes

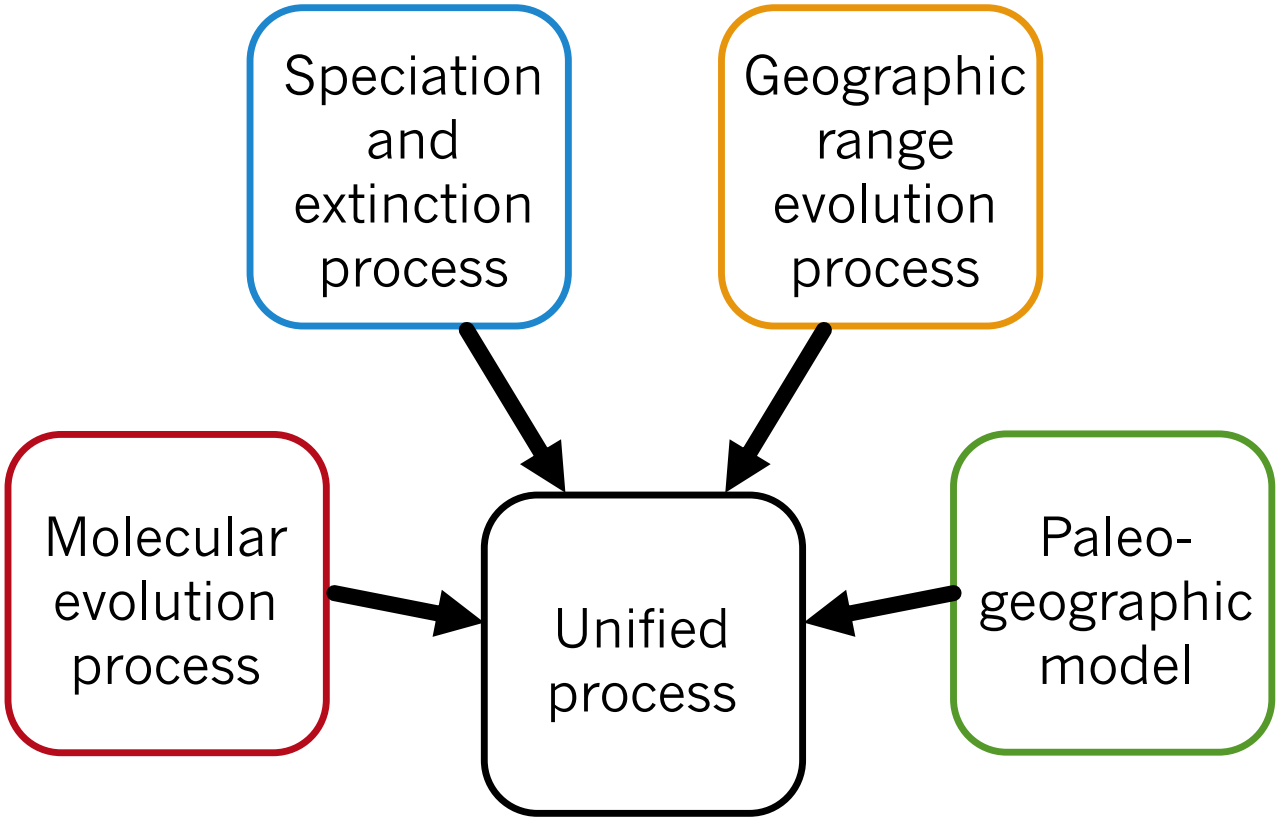
Systematic Biology, Volume 65, Issue 4, 1 July 2016, Pages 726–736,
https://doi.org/10.1093/sysbio/syw021
Published: 28 May 2016 **Article history** ▾

RevBayes

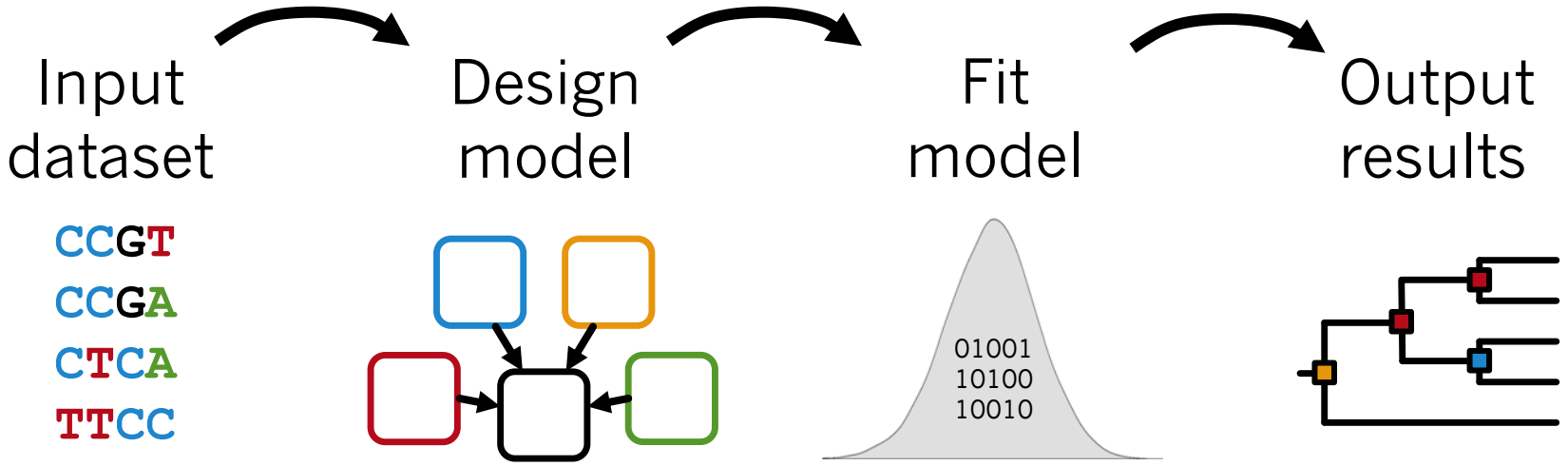
1. Fast
2. Flexible
3. Easy-to-use

http://revbayes.com

Model design with RevBayes



Phylogenetic inference with RevBayes

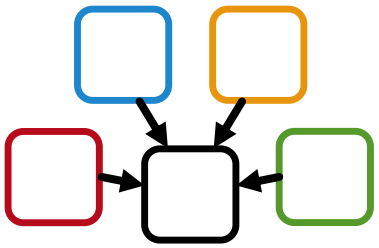


Phylogenetic inference with RevBayes

Input dataset

CCGT
CCGA
CTCA
TTCC

Design model



Probabilistic graphical models

Fit model

01001
10100
10010

Output results

Familiar phylogenetic components

Taxon data (what evolved)
Tree model
Rate matrix
Clock model
Site-rate variation
Fossilization model
Biogeographic model
Continuous trait model
...

Familiar phylogenetic components

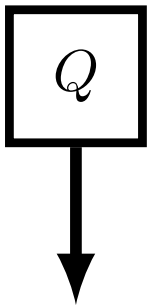
Taxon data (what evolved)
Tree model
Rate matrix
Clock model
Site-rate variation
Fossilization model
Biogeographic model
Continuous trait model
...

$$Q_{\mathrm{JC}} = \begin{pmatrix} - & 1/3 & 1/3 & 1/3 \\ 1/3 & - & 1/3 & 1/3 \\ 1/3 & 1/3 & - & 1/3 \\ 1/3 & 1/3 & 1/3 & - \end{pmatrix}$$

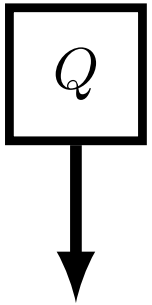
Substitution model

Jukes-Cantor / JC

$$Q_{\mathrm{JC}} = \begin{pmatrix} - & 1/3 & 1/3 & 1/3 \\ 1/3 & - & 1/3 & 1/3 \\ 1/3 & 1/3 & - & 1/3 \\ 1/3 & 1/3 & 1/3 & - \end{pmatrix}$$



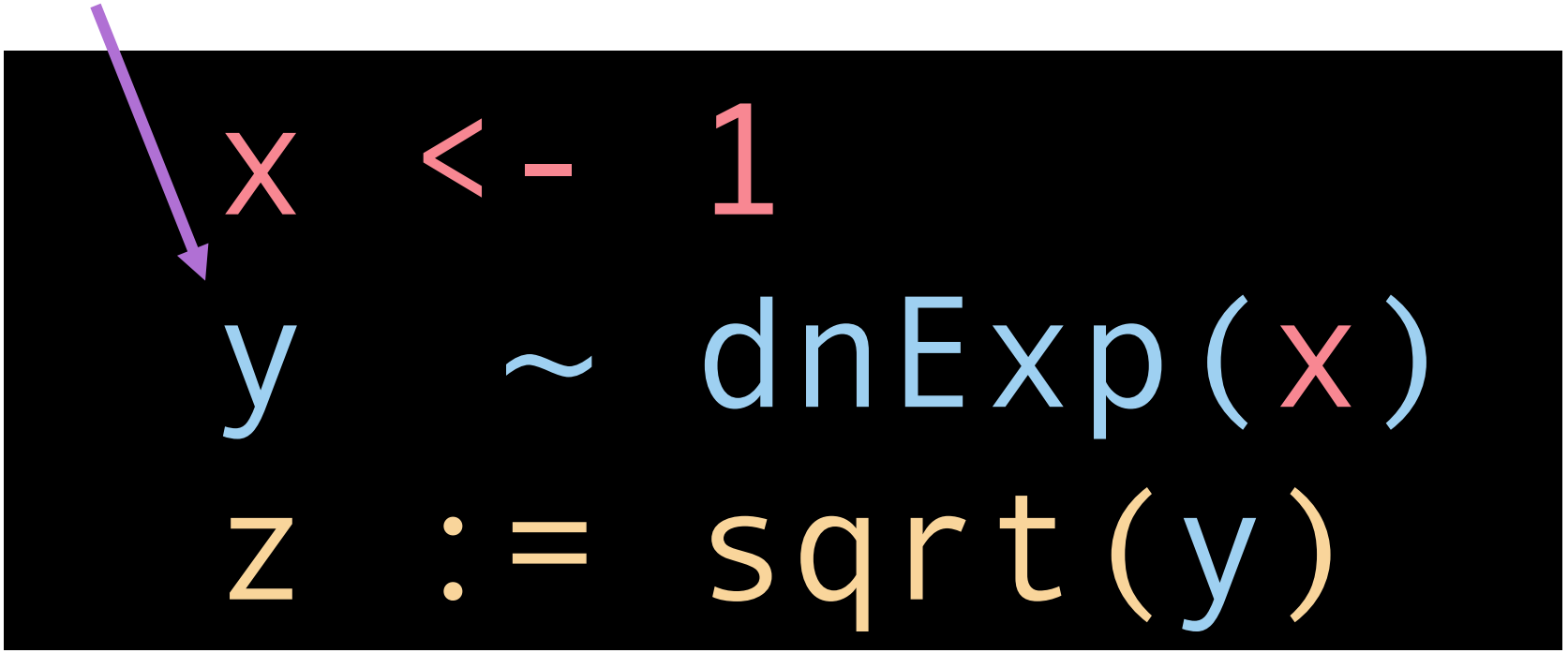
$$Q_{\mathrm{JC}} = \begin{pmatrix} - & 1/3 & 1/3 & 1/3 \\ 1/3 & - & 1/3 & 1/3 \\ 1/3 & 1/3 & - & 1/3 \\ 1/3 & 1/3 & 1/3 & - \end{pmatrix}$$

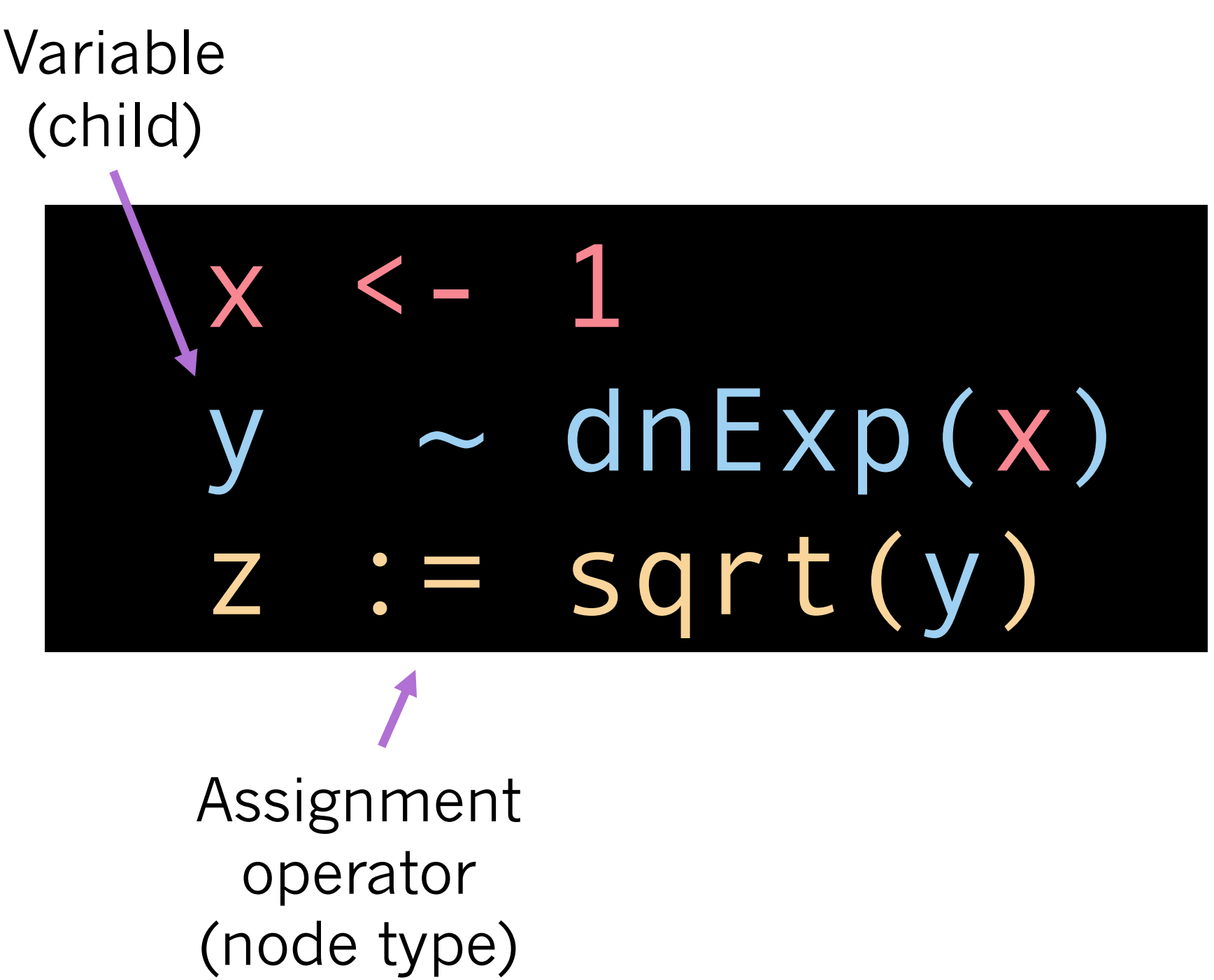


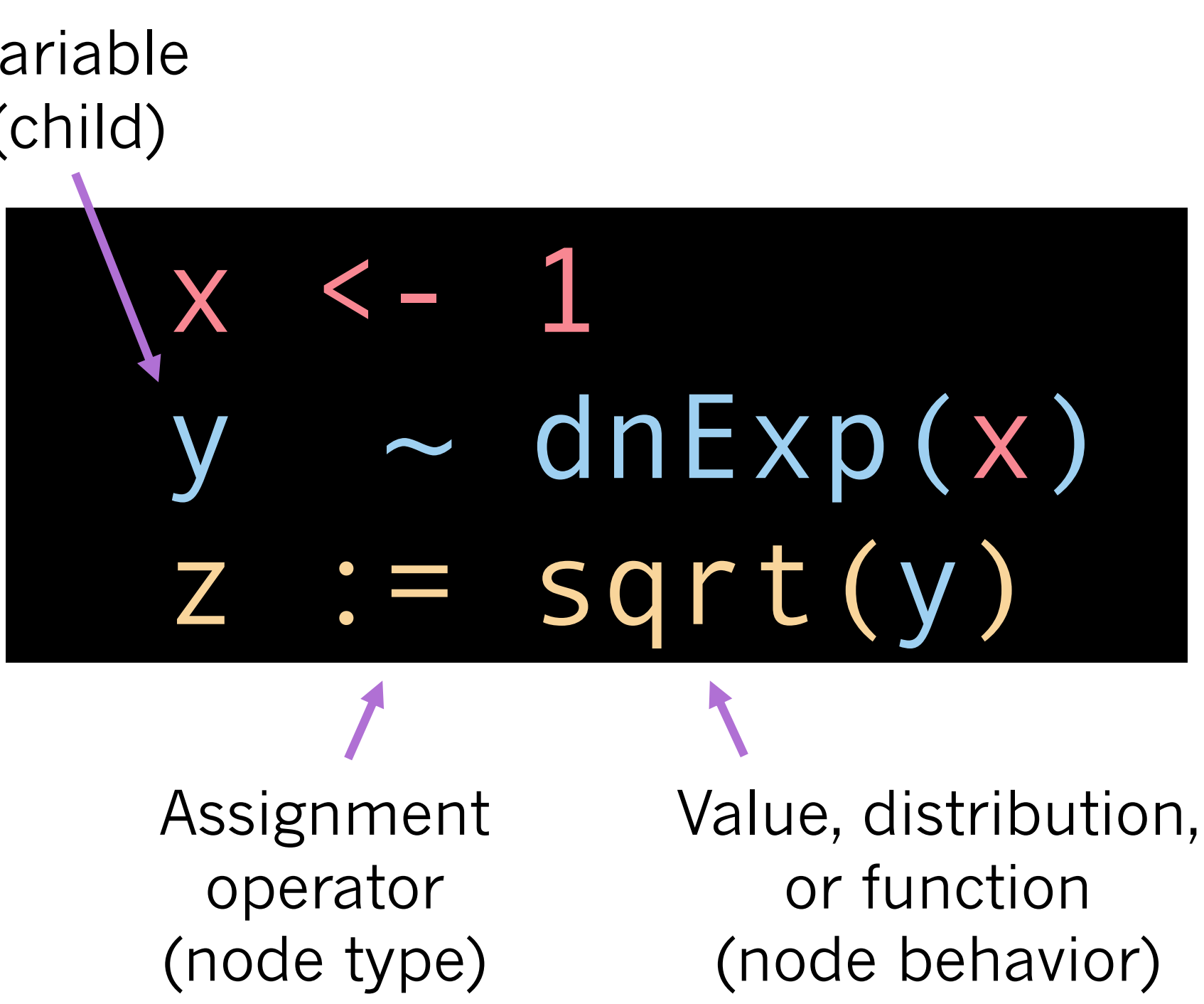

```
# Jukes-Cantor model
Q <- fnJC(4)
```

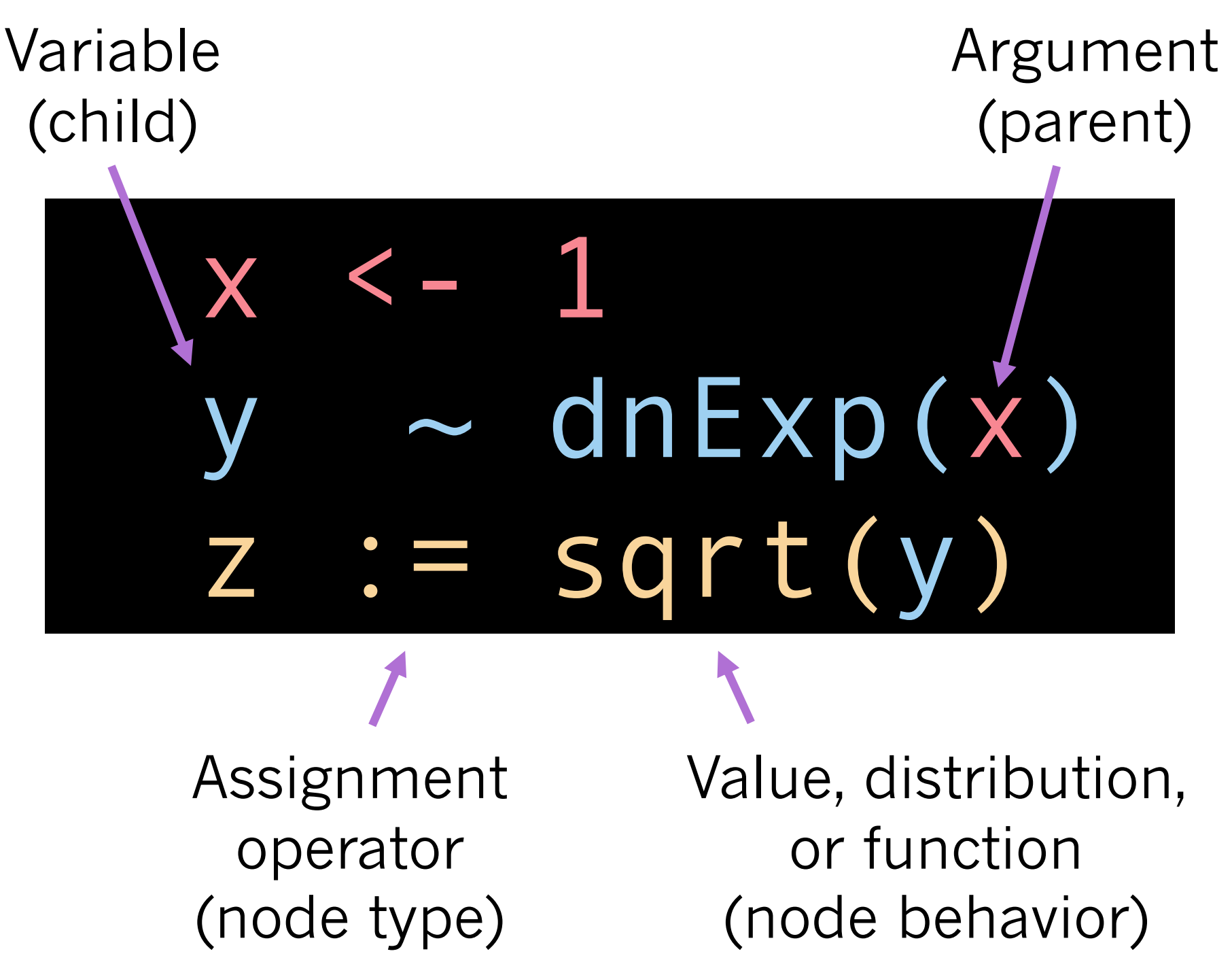
```
x <- 1
y  ~ dnExp(x)
z := sqrt(y)
```

Variable
(child)

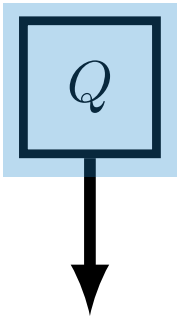



Variable
(child)
x <- 1
y ~ dnExp(x)
z := sqrt(y)
Assignment
operator
(node type)


Variable
(child)
x <- 1
y ~ dnExp(x)
z := sqrt(y)
Assignment
operator
(node type)
Value, distribution,
or function
(node behavior)


Variable
(child)
Argument
(parent)
x <- 1
y ~ dnExp(x)
z := sqrt(y)
Assignment
operator
(node type)
Value, distribution,
or function
(node behavior)

$$Q_{\mathrm{JC}} = \begin{pmatrix} - & 1/3 & 1/3 & 1/3 \\ 1/3 & - & 1/3 & 1/3 \\ 1/3 & 1/3 & - & 1/3 \\ 1/3 & 1/3 & 1/3 & - \end{pmatrix}$$



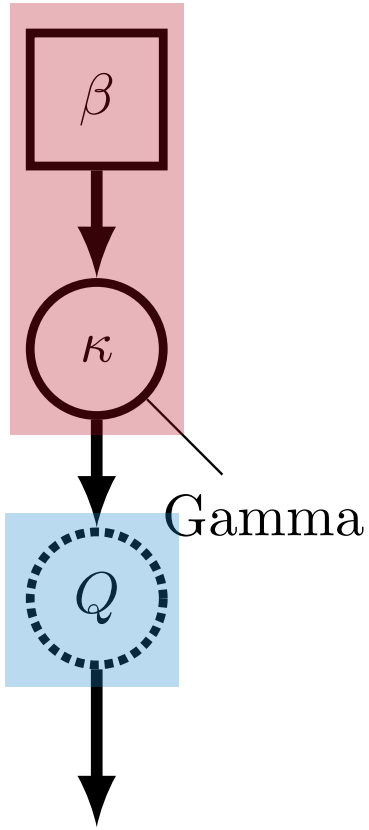

```
# Jukes-Cantor model
Q <- fnJC(4)
```

Substitution model

Kimura-2P / K2P

$$Q_{\text{K80}} = \begin{pmatrix} - & 1 & \kappa & 1 \\ 1 & - & 1 & \kappa \\ \kappa & 1 & - & 1 \\ 1 & \kappa & 1 & - \end{pmatrix}$$

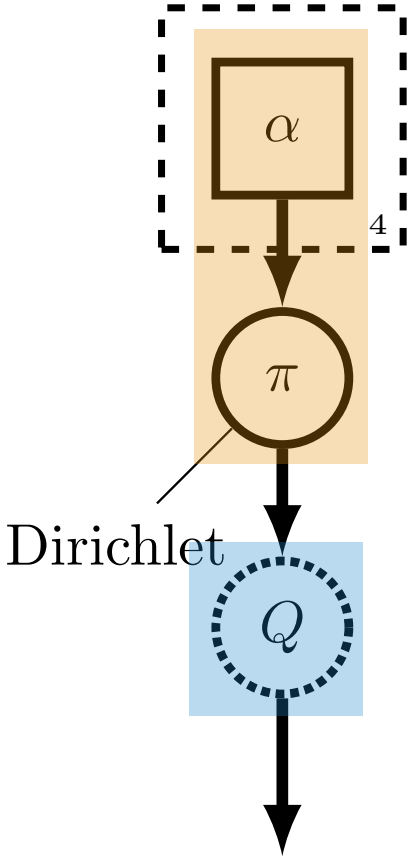
```
# Ti/Tv rate ratio, mean 1
beta <- 1
kappa ~ dnGamma(beta, beta)

# Kimura 2-parameter model
Q := fnK80(kappa)
```



Substitution model

Felsenstein-81 / F81

$$Q_{\text{F81}} = \begin{pmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{pmatrix}$$



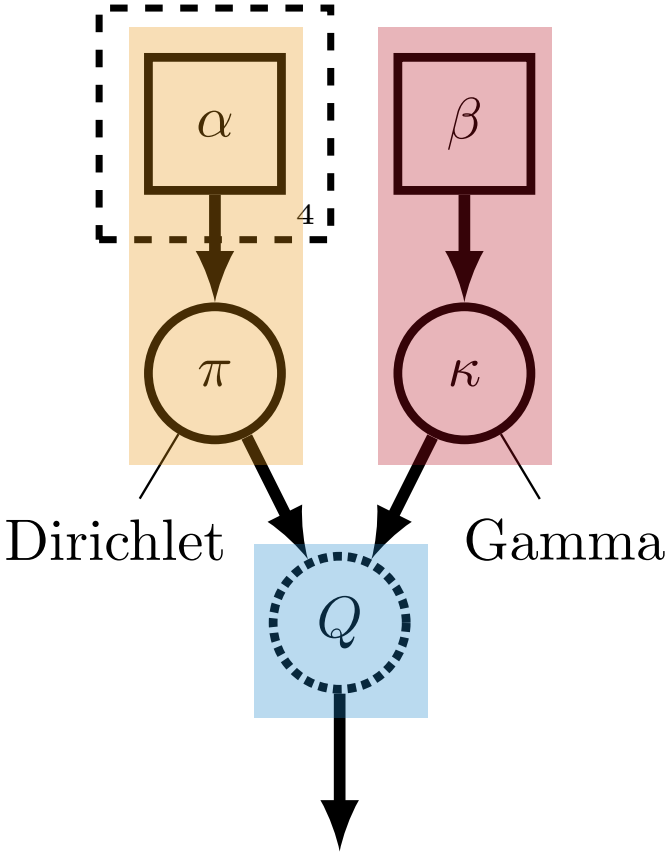

```
# base frequencies, “flat”
alpha <- [1, 1, 1, 1]
pi ~ dnDirichlet(alpha)

# Felsenstein-81 model
Q := fnF81(pi)
```

Substitution model

Hasegawa-Kishino-Yano / HKY

$$Q_{\text{HKY}} = \begin{pmatrix} - & \pi_C & \kappa\ \pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\ \pi_T \\ \kappa\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\ \pi_C & \pi_G & - \end{pmatrix}$$

```
# base frequencies, “flat”
alpha <- [1, 1, 1, 1]
pi ~ dnDirichlet(alpha)

# Ti/Tv rate ratio, mean 1
beta <- 1
kappa ~ dnGamma(beta, beta)

# HKY model
Q := fnHKY(pi, kappa)
```

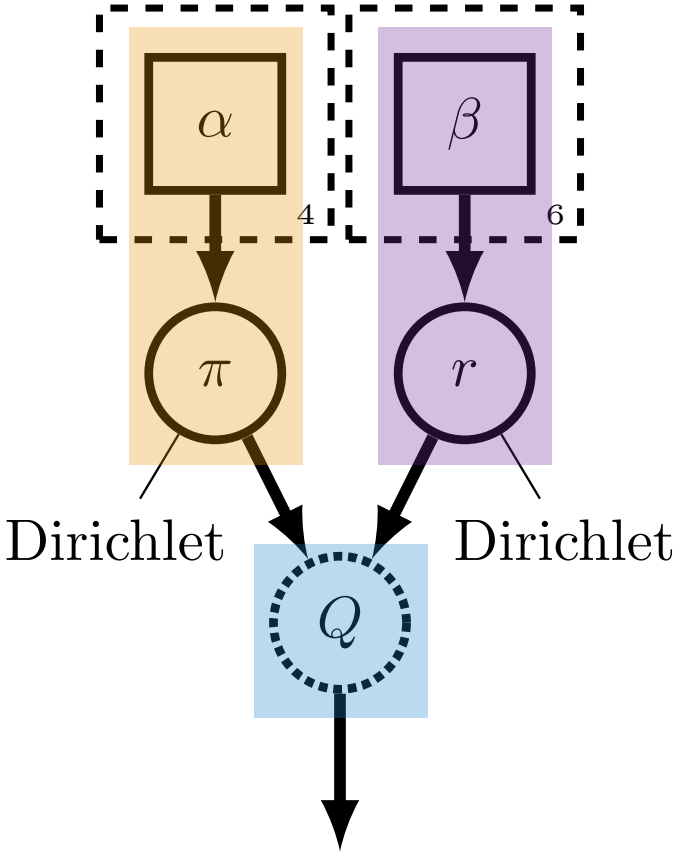


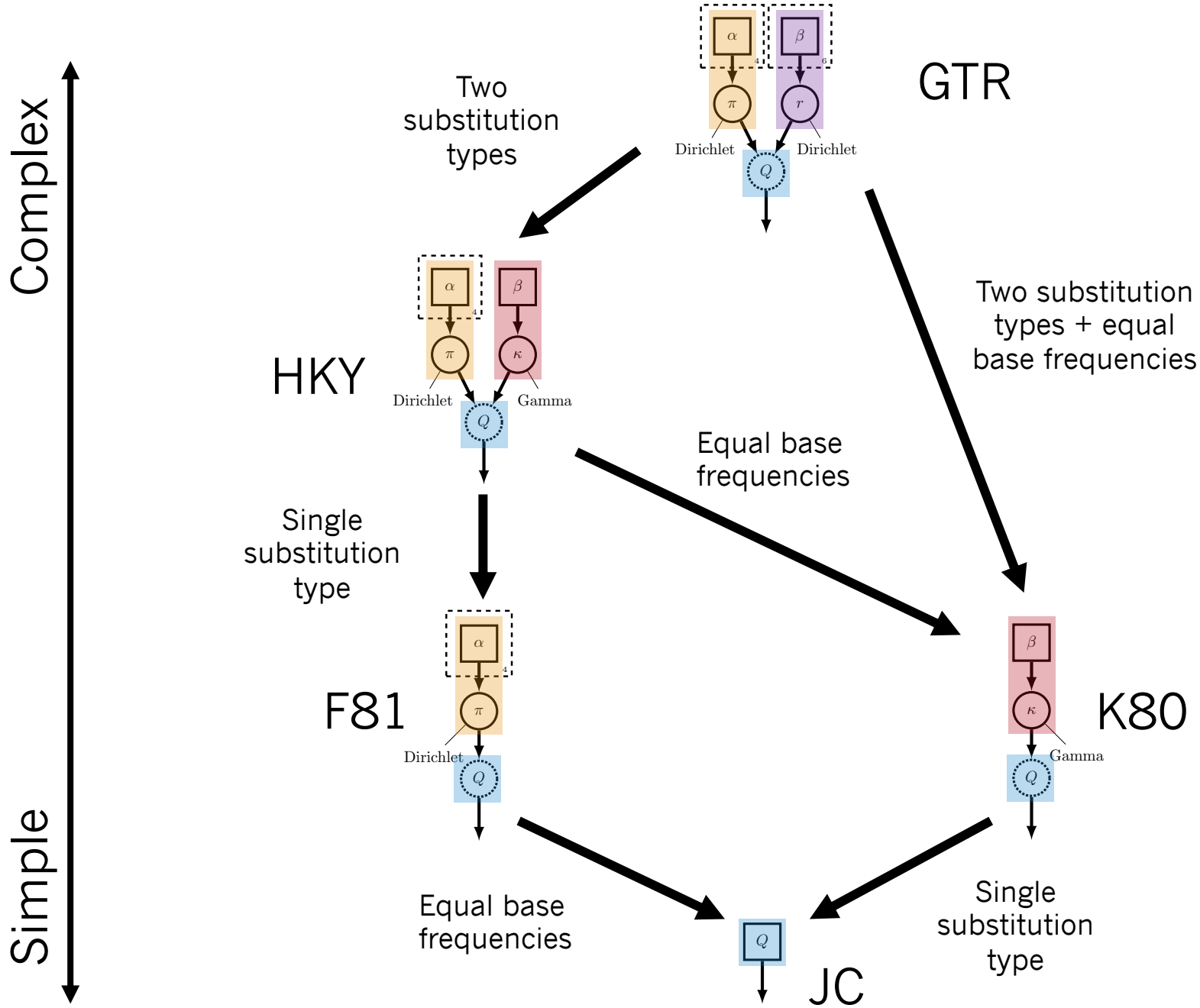
Substitution model

Generalized-Time-Reversible / GTR

$$Q_{\text{GTR}} = \begin{pmatrix} - & r_{AC}\,\pi_C & r_{AG}\,\pi_G & r_{AT}\,\pi_T \\ r_{AC}\,\pi_A & - & r_{CG}\,\pi_G & r_{CT}\,\pi_T \\ r_{AG}\,\pi_A & r_{CG}\,\pi_C & - & r_{GT}\,\pi_T \\ r_{AT}\,\pi_A & r_{CT}\,\pi_C & r_{GT}\,\pi_G & - \end{pmatrix}$$

```
# base frequencies, “flat”
alpha <- [1, 1, 1, 1]
pi ~ dnDirichlet(alpha)

# exch. rates, “flat”
beta <- [1, 1, 1, 1, 1, 1]
er ~ dnDirichlet(beta)

# GTR model
Q := fnGTR(pi, er)
```



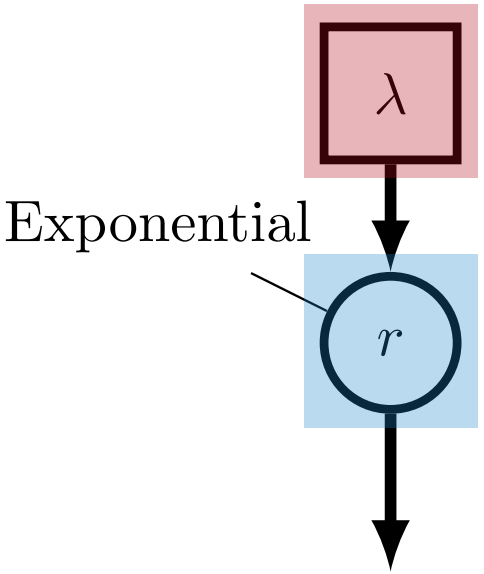


from Dave Swofford's talk

Clock models

Strict clock

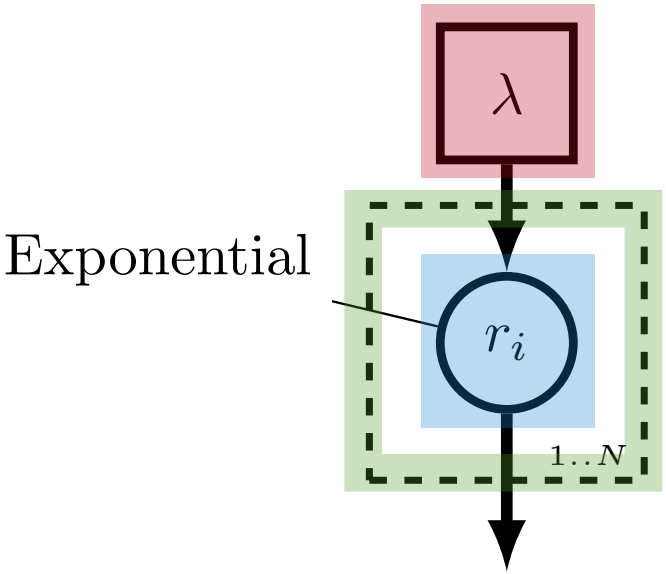
```
# strict clock rate mean (^-1)
lambda <- 1.0 / 0.1

# strict exponential clock
clock ~ dnExp(lambda)
```



Clock models

Centered relaxed clock
iid exponentials

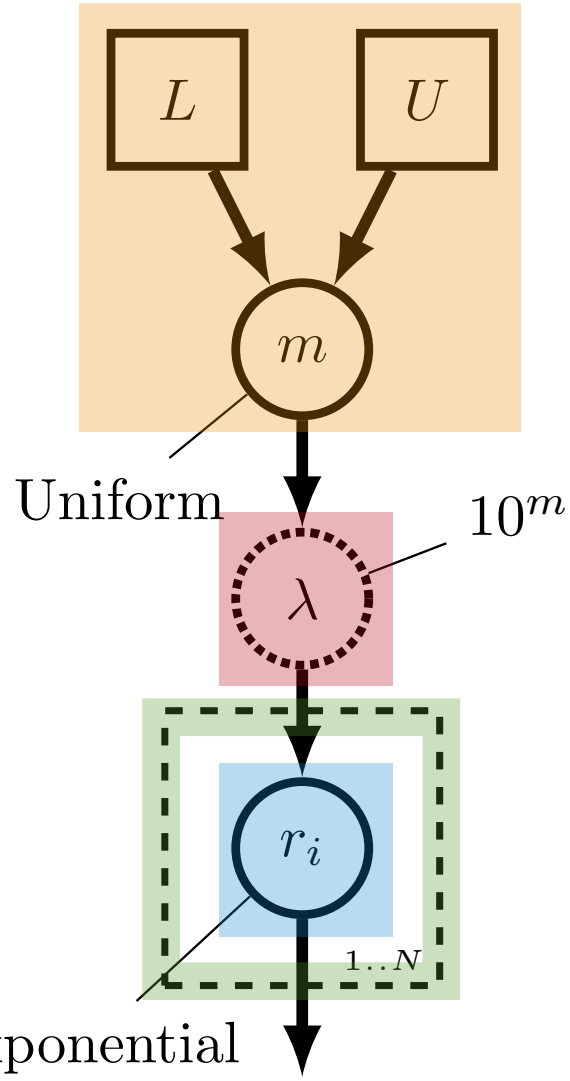
```
# relaxed clock rate mean (^-1)
lambda <- 1.0 / 0.1

# uncorrelated exponential clock
for (i in 1:n_branches) {
    clock[i] ~ dnExp(lambda)
}
```



Clock models

Uncentered relaxed clock iid exponentials

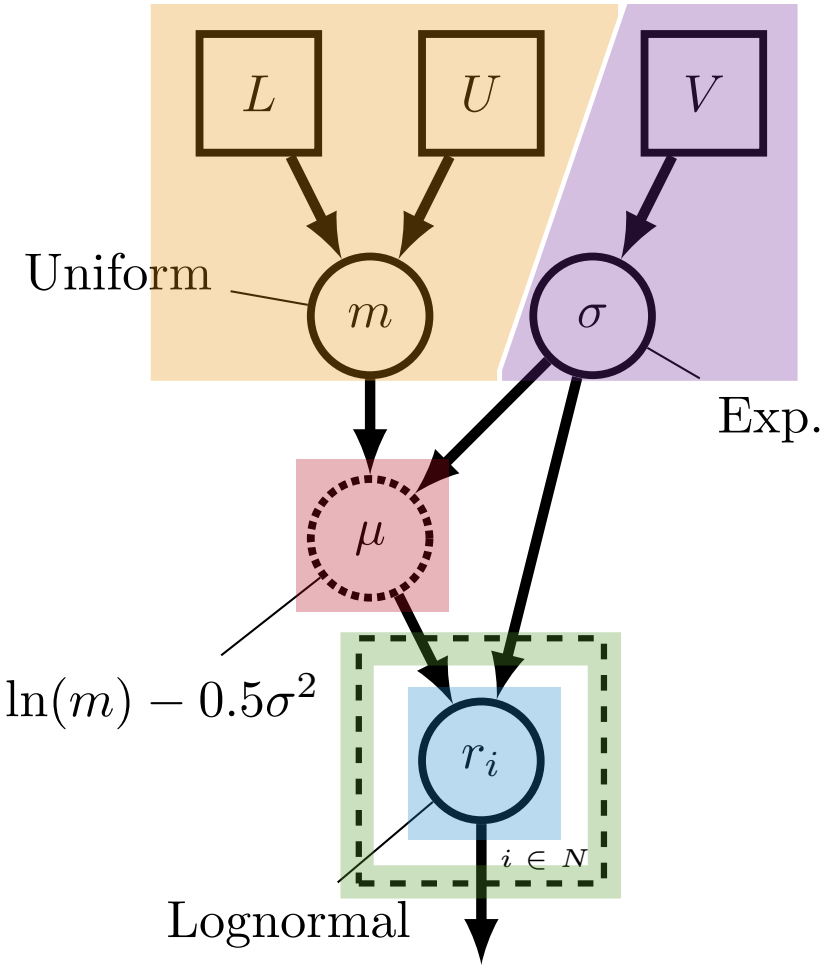
```
# order-of-magnitude of clock
lower <- -6
upper <- +1
m ~ dnUniform(lower, upper)

# relaxed clock rate mean (^-1)
lambda := 10^-m

# uncorrelated exponential clock
for (i in 1:n_branches) {
    clock[i] ~ dnExp(lambda)
}
```

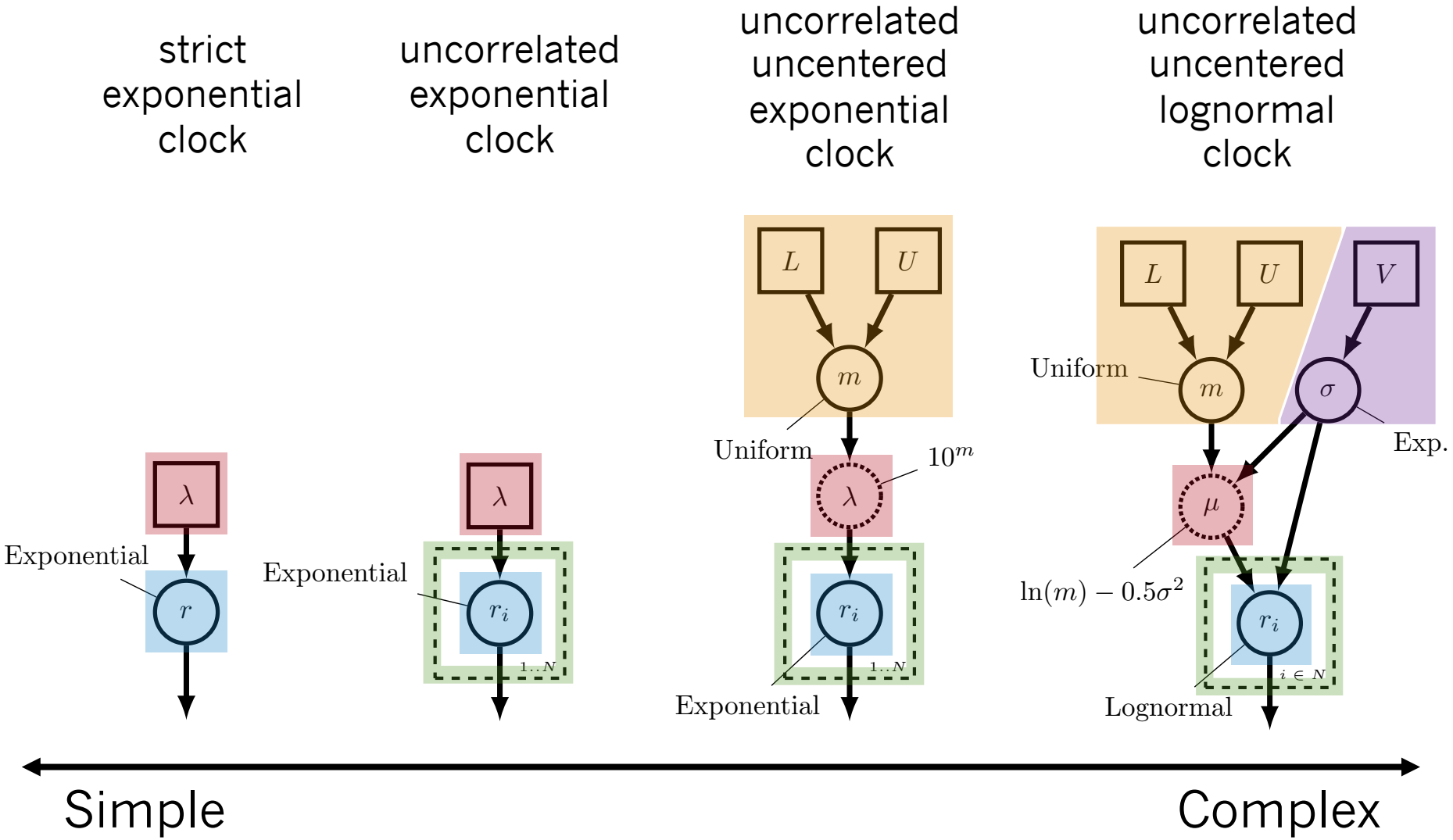


Clock models

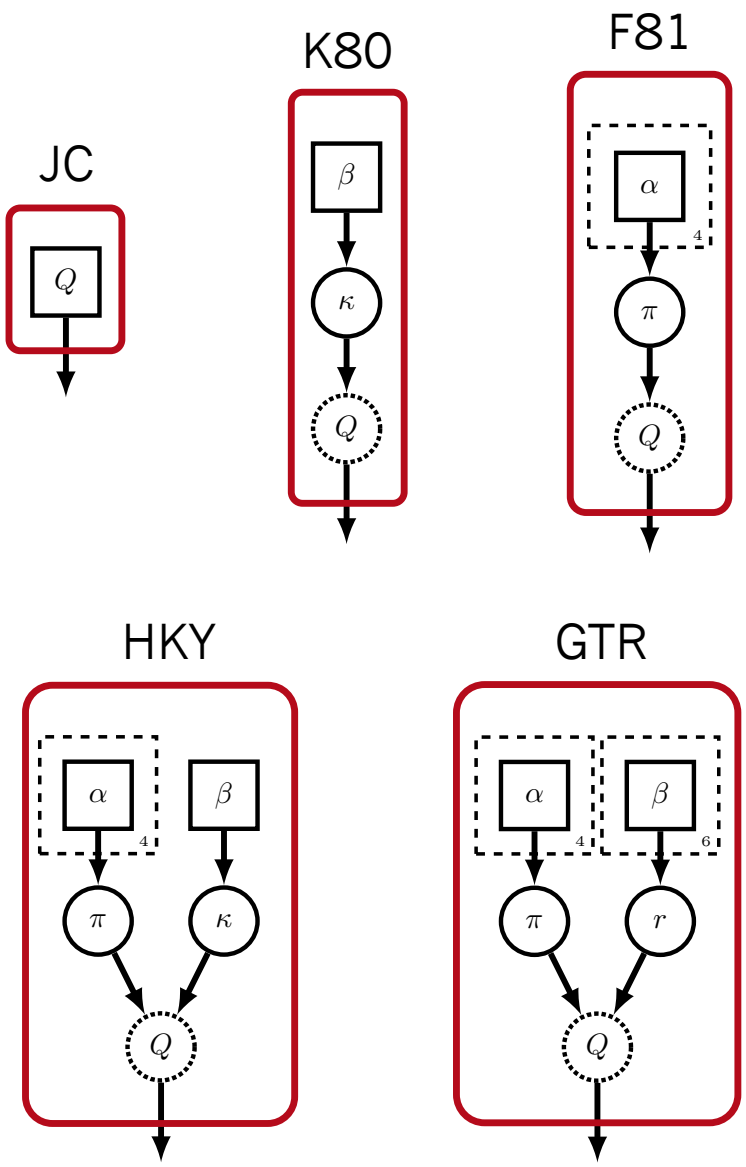
Uncentered relaxed clock iid lognormals

```
# order-of-magnitude of clock
lower <- -6
upper <- +1
m ~ dnUniform(lower, upper)

# relaxed clock rate variance
V <- 1
sigma ~ dnExp(V)

# relaxed clock rate mean
mu := m - 0.5*sigma^2

# uncorrelated lognormal clock
for (i in 1:n_branches) {
    r[i] ~ dnLognormal(mu, sigma)
}
```

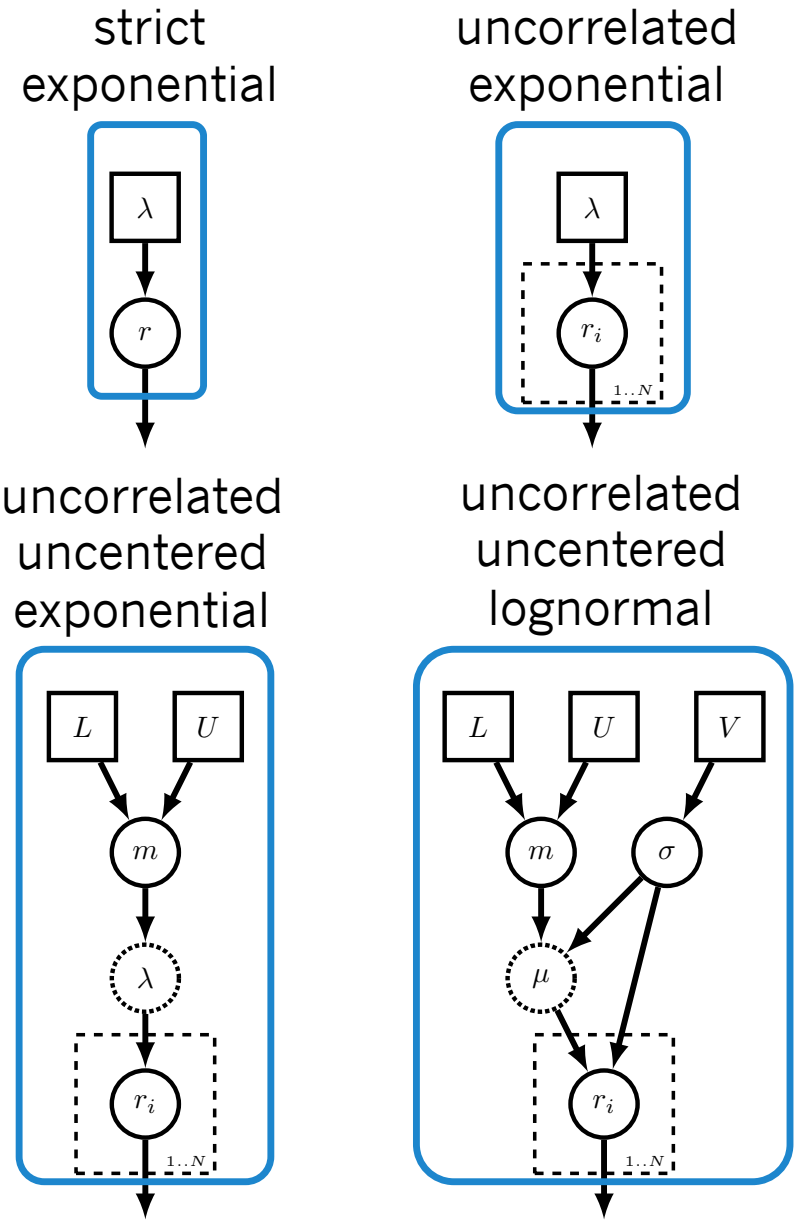



strict exponential clock
uncorrelated exponential clock
uncorrelated uncentered exponential clock
uncorrelated uncentered lognormal clock
λ
Exponential
r
λ
Exponential
r_i
1..N
L
U
m
Uniform
10^m
λ
r_i
1..N
Exponential
L
U
V
Uniform
m
σ
Exp.
μ
$\ln(m) - 0.5\sigma^2$
r_i
$i \in N$
Lognormal
Simple
Complex

Rate matrices across sites



Clock rates along branches



PGM modules

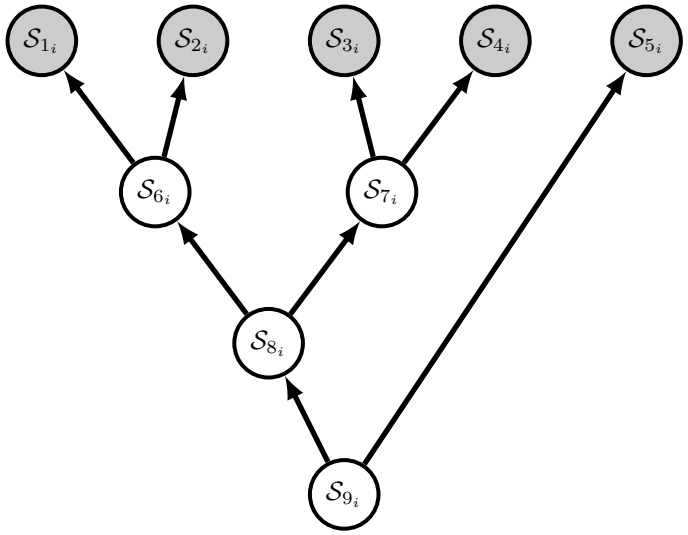
Definition

a module is a PGM that subgraph that performs a distinct, integrated function

Modules

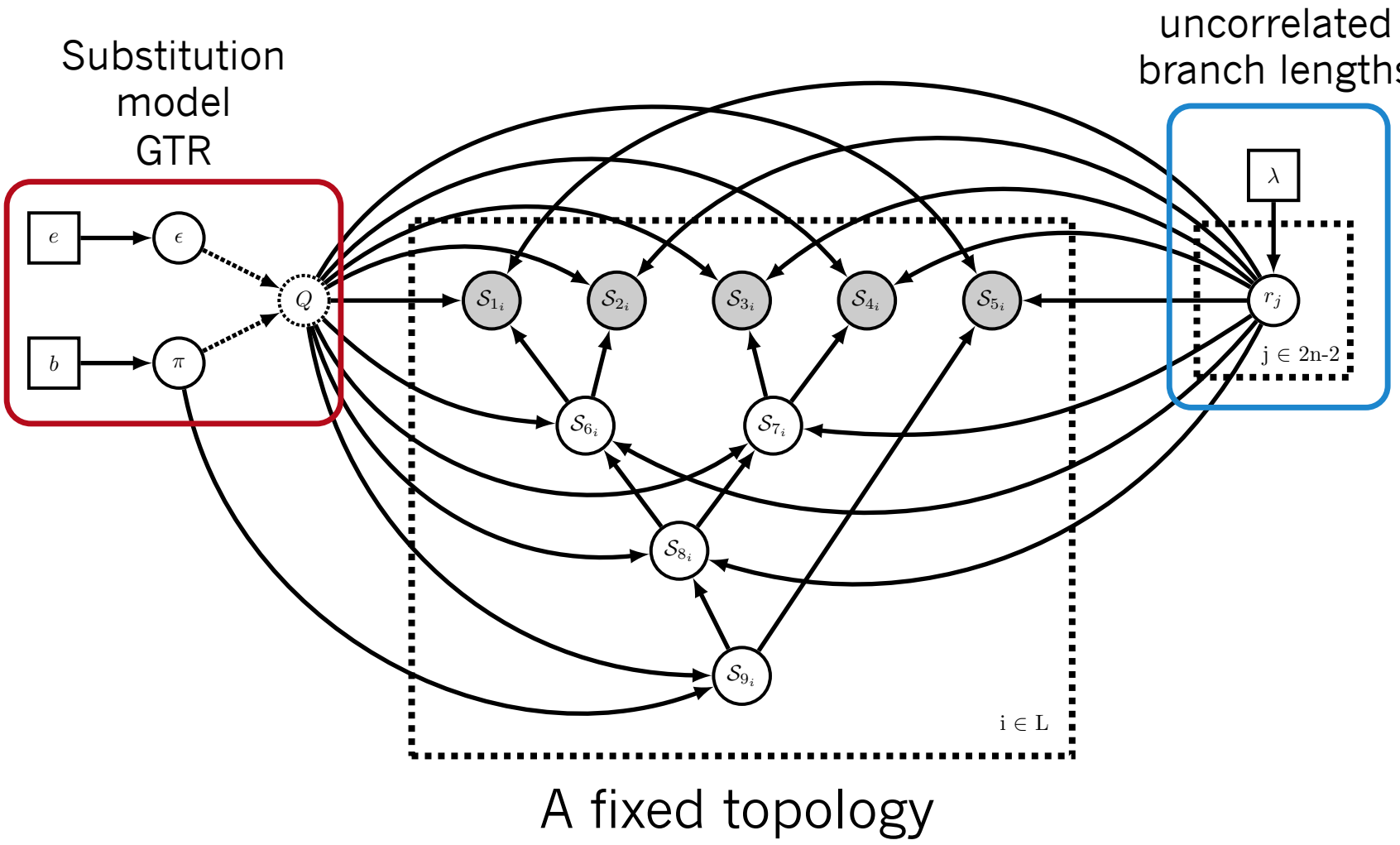
- Reveal “higher level” relationships
- Provide templated functionality
- Mask details unless needed

Phylogenetic substitution model

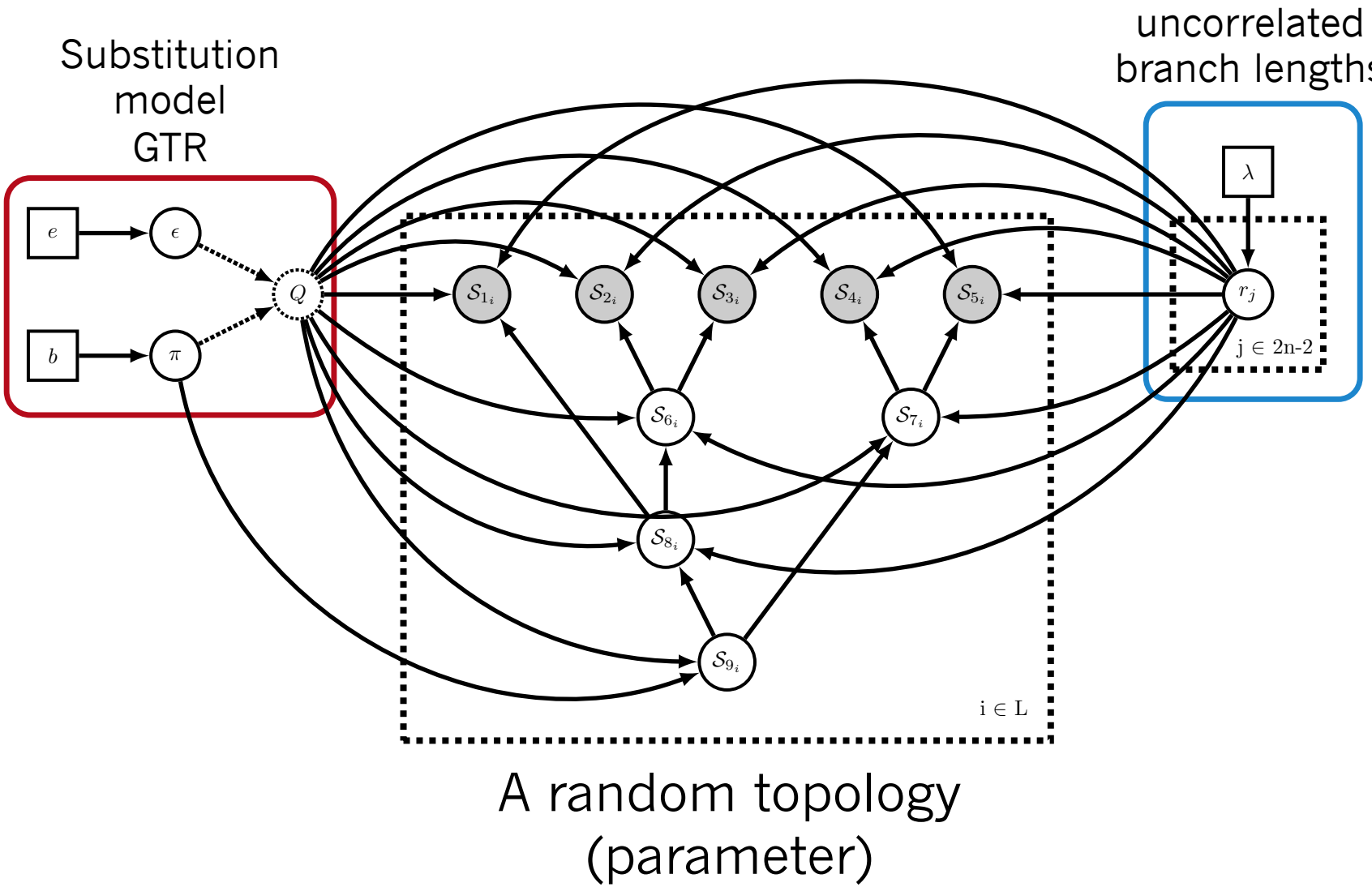


A fixed topology

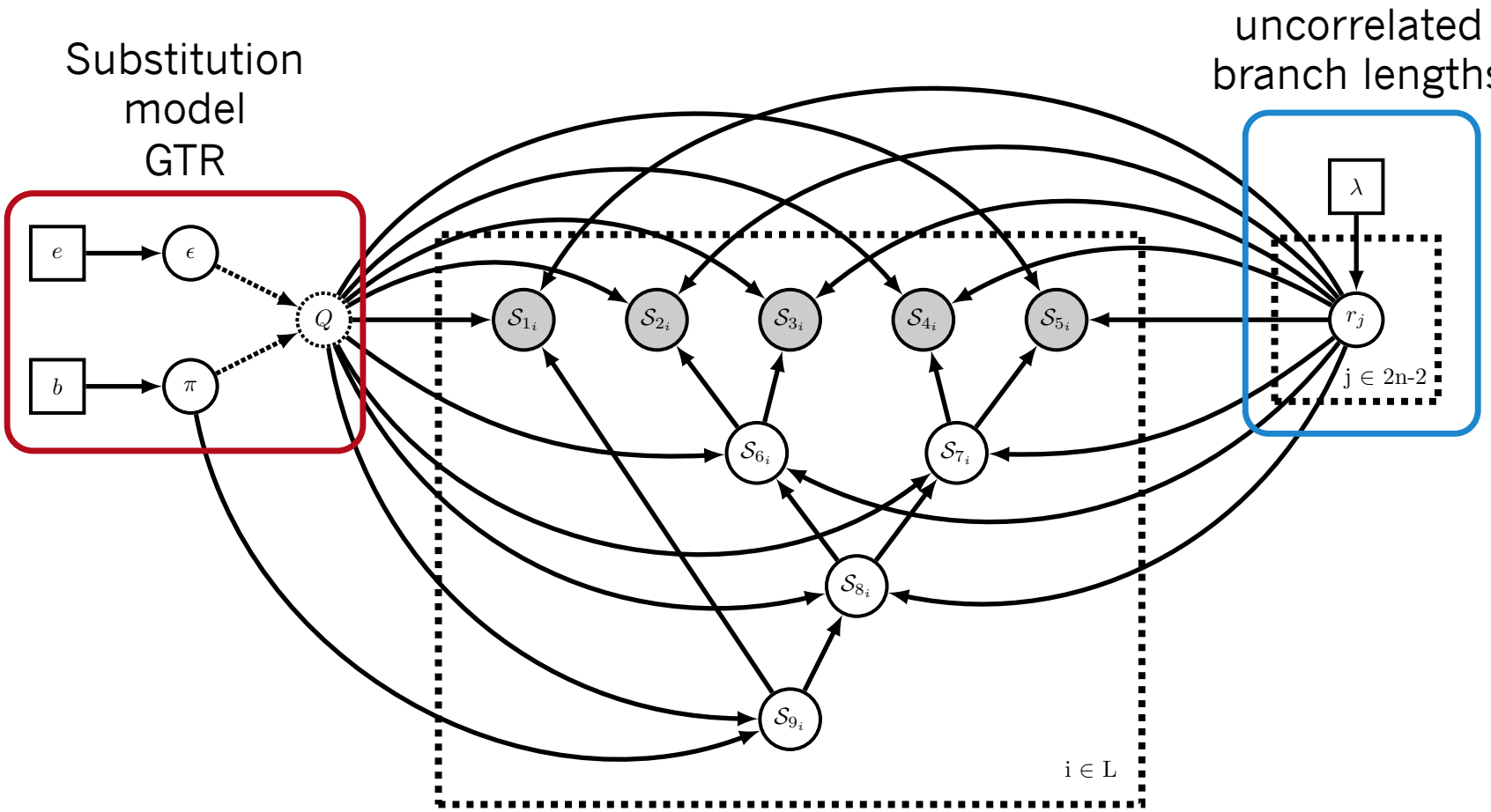
Phylogenetic substitution model



Phylogenetic substitution model

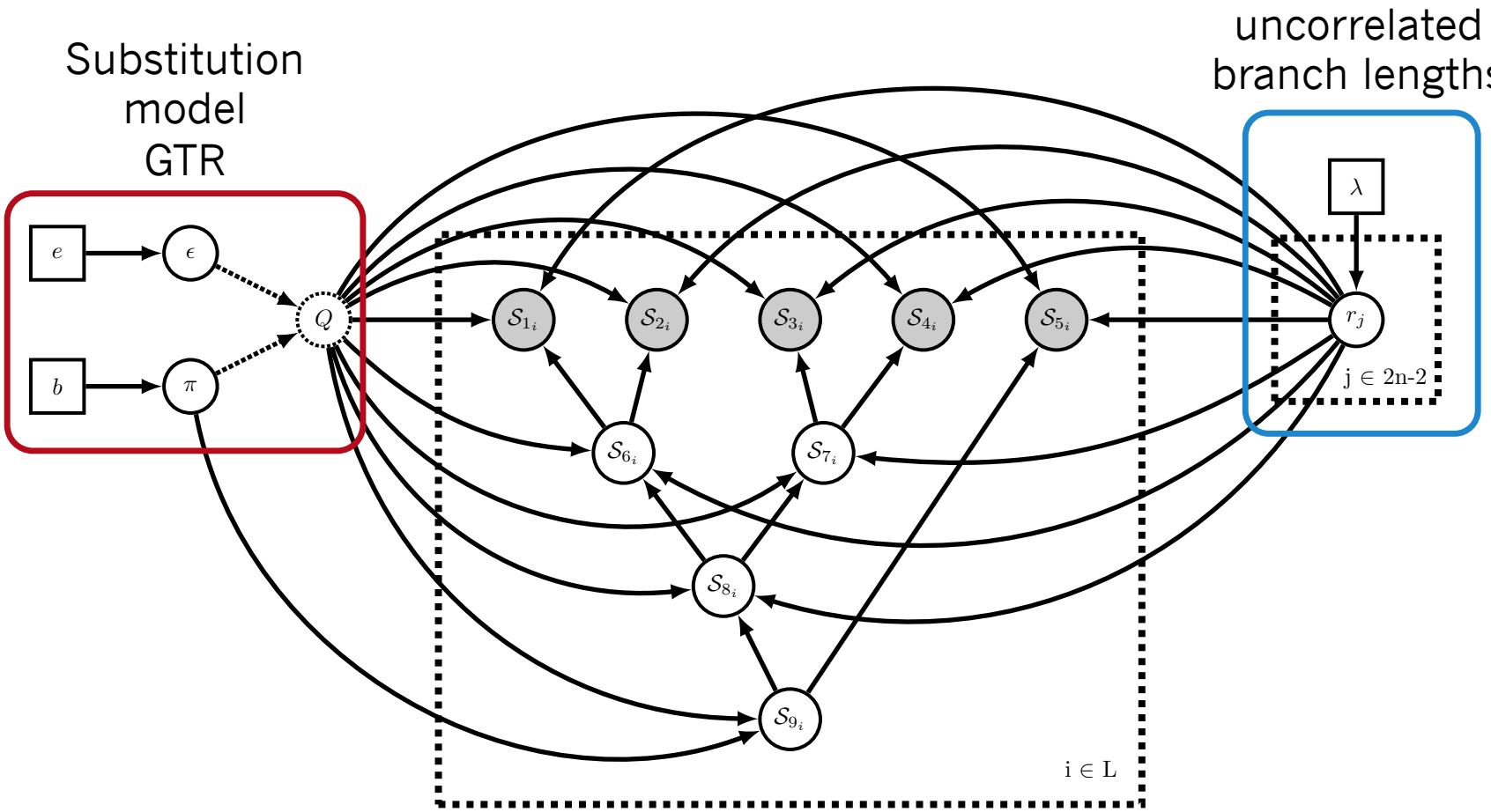


Phylogenetic substitution model



A random topology
(parameter)

Phylogenetic substitution model



Is a tree a parameter?

Is a tree a graph?

Is a tree a model?

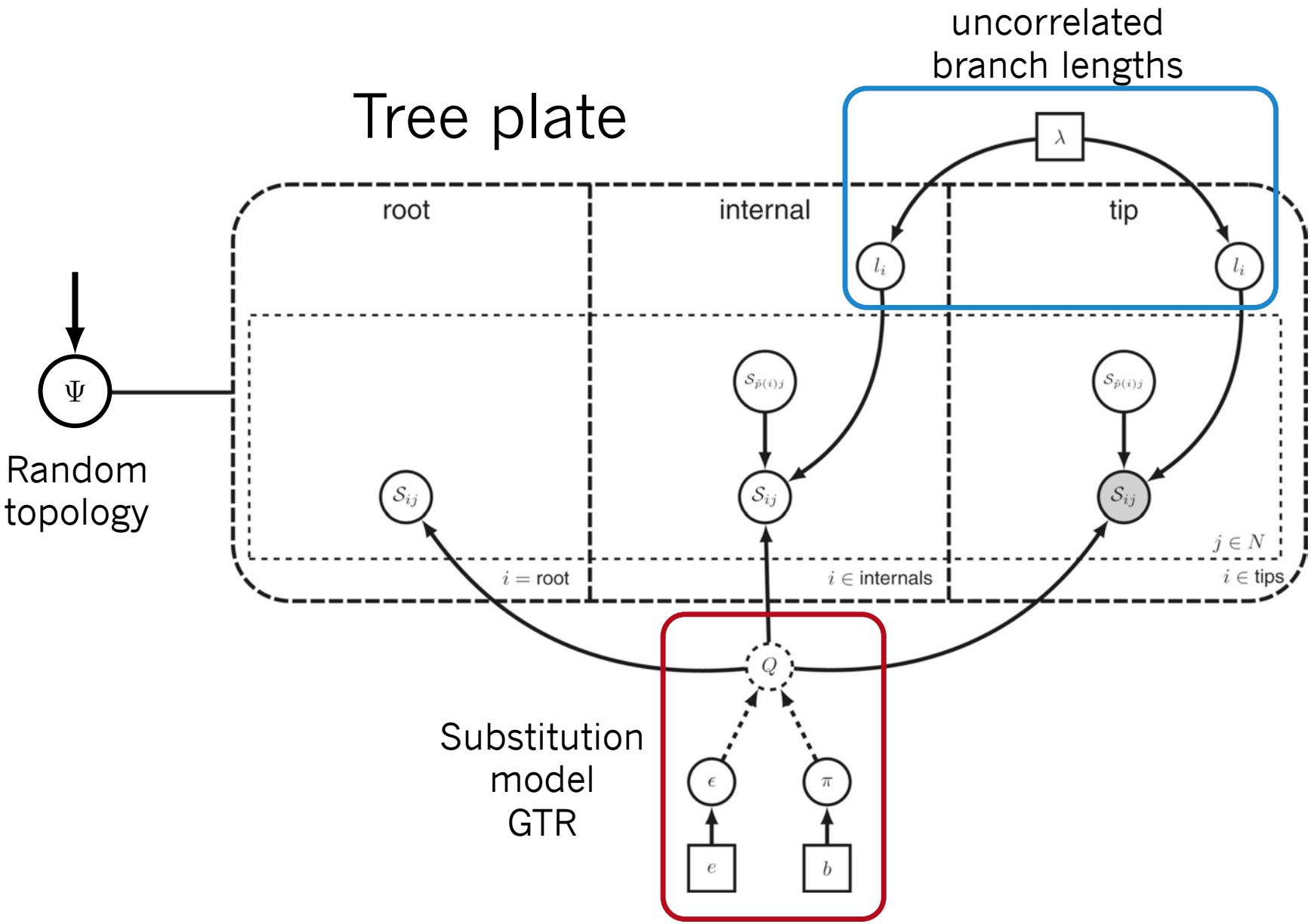
The Holy Treenity

Is a tree a parameter?
Yes

Is a tree a graph?
Yes

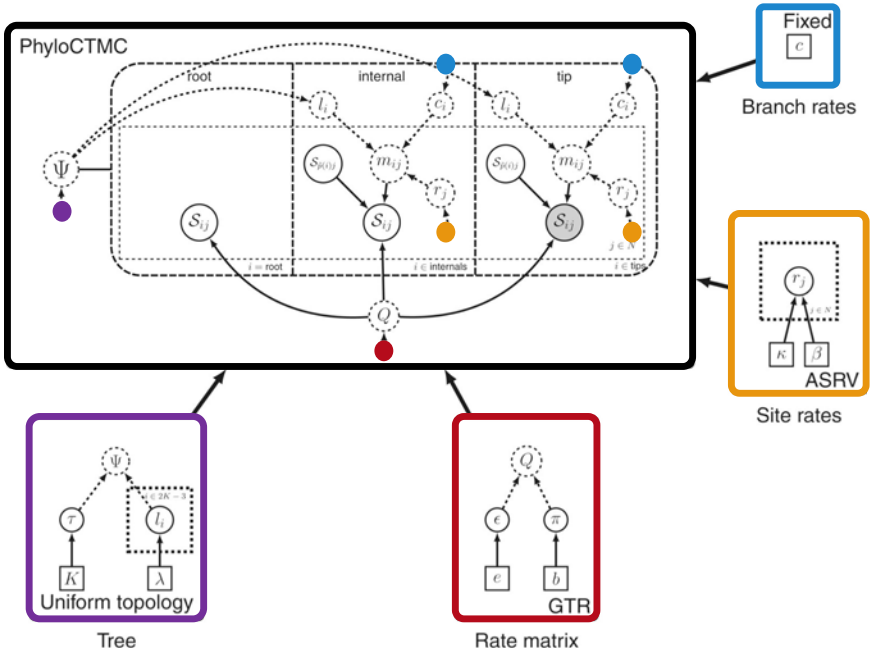
Is a tree a model?
Yes

Phylogenetic substitution model

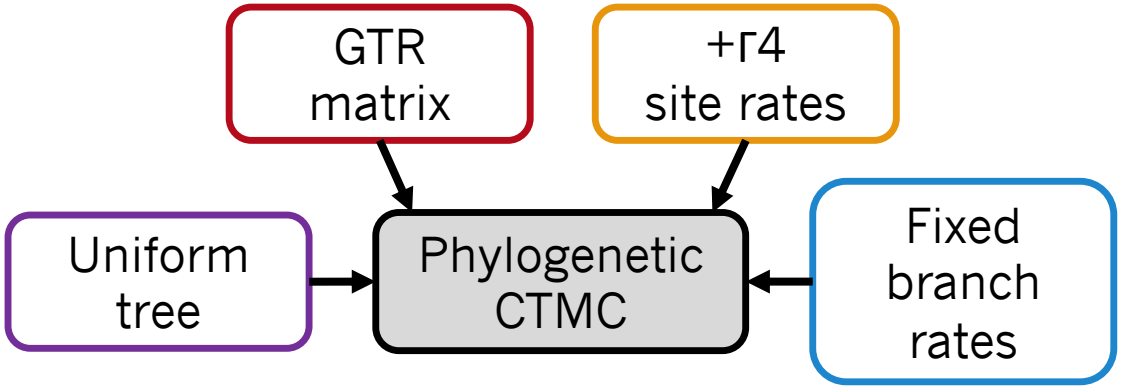


Phylogenetic substitution model

Expanded view

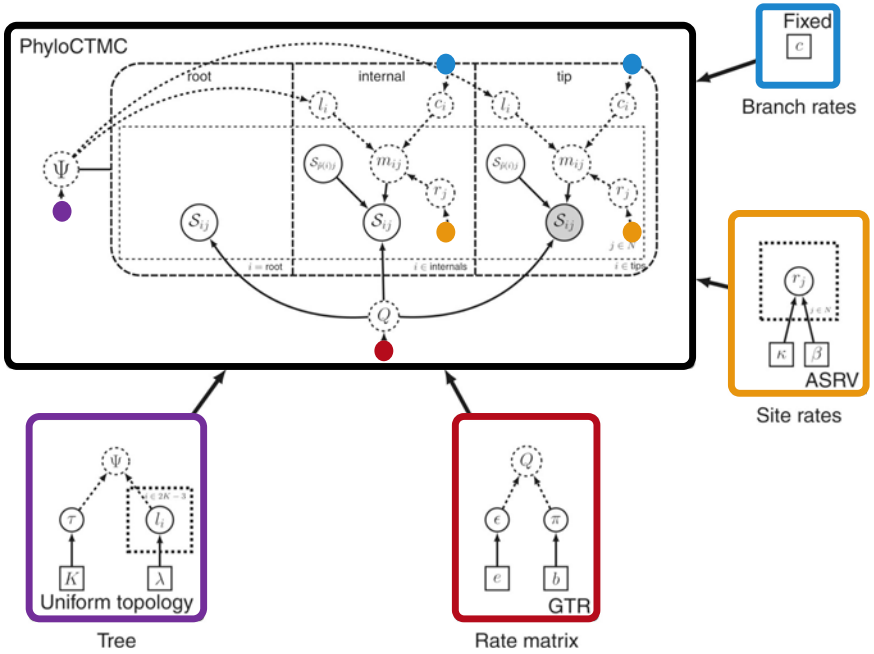


Modular view

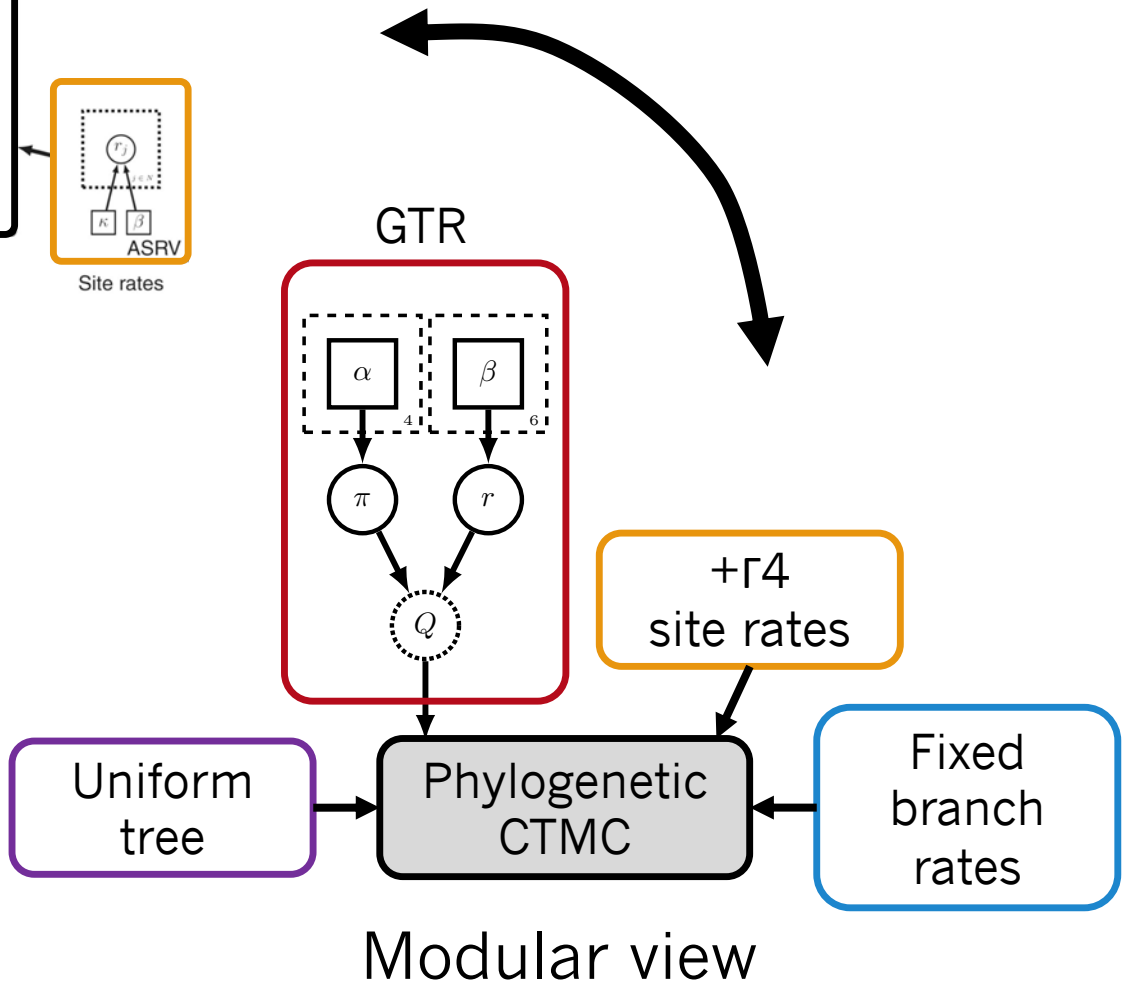


Phylogenetic substitution model

Expanded view

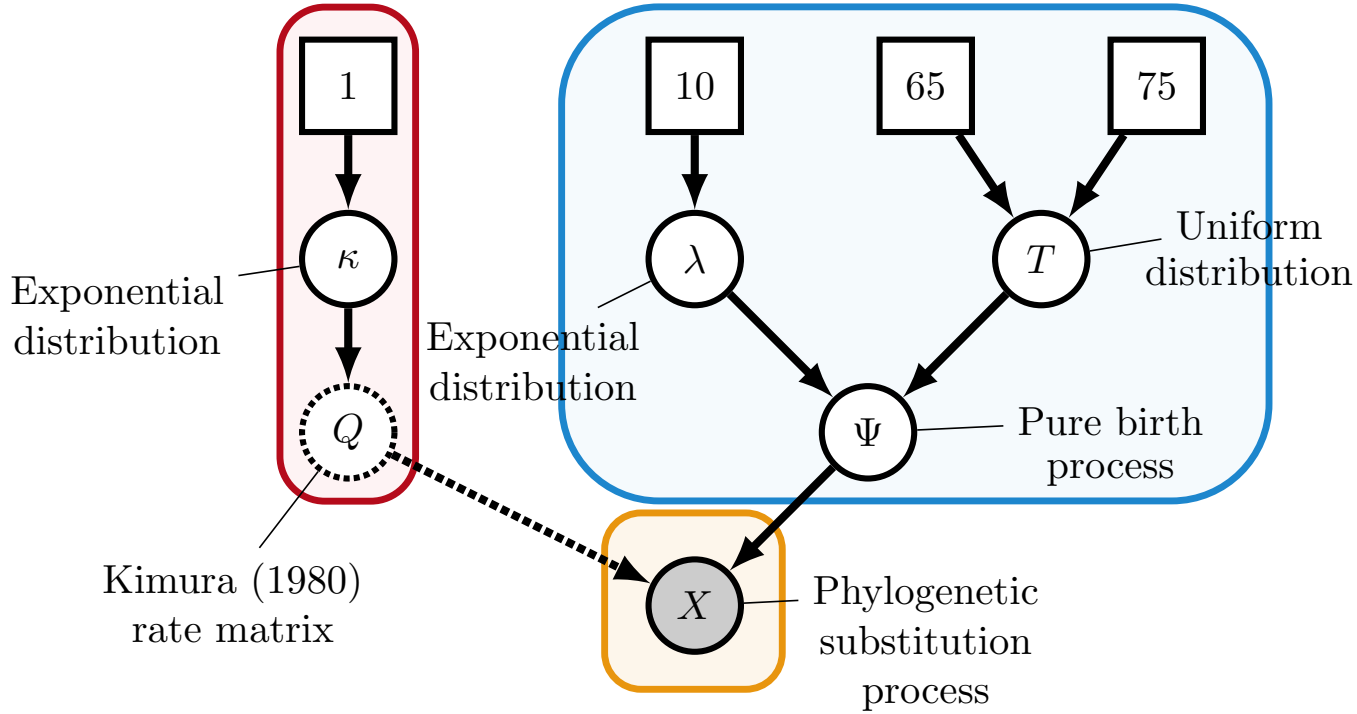


Modular view



```
# pure birth prior and hyperpriors
birth ~ dnExponential(10)
tree_age ~ dnUnform(65, 75)
tree ~ dnBDP(birth, tree_age, taxa)
# rate matrix
kappa ~ dnExponential(1)
Q := fnK80(kappa)
# phylogenetic substitution process
seq ~ dnPhyloCTMC(Q, tree, “DNA”)
seq.clamp(data)
```

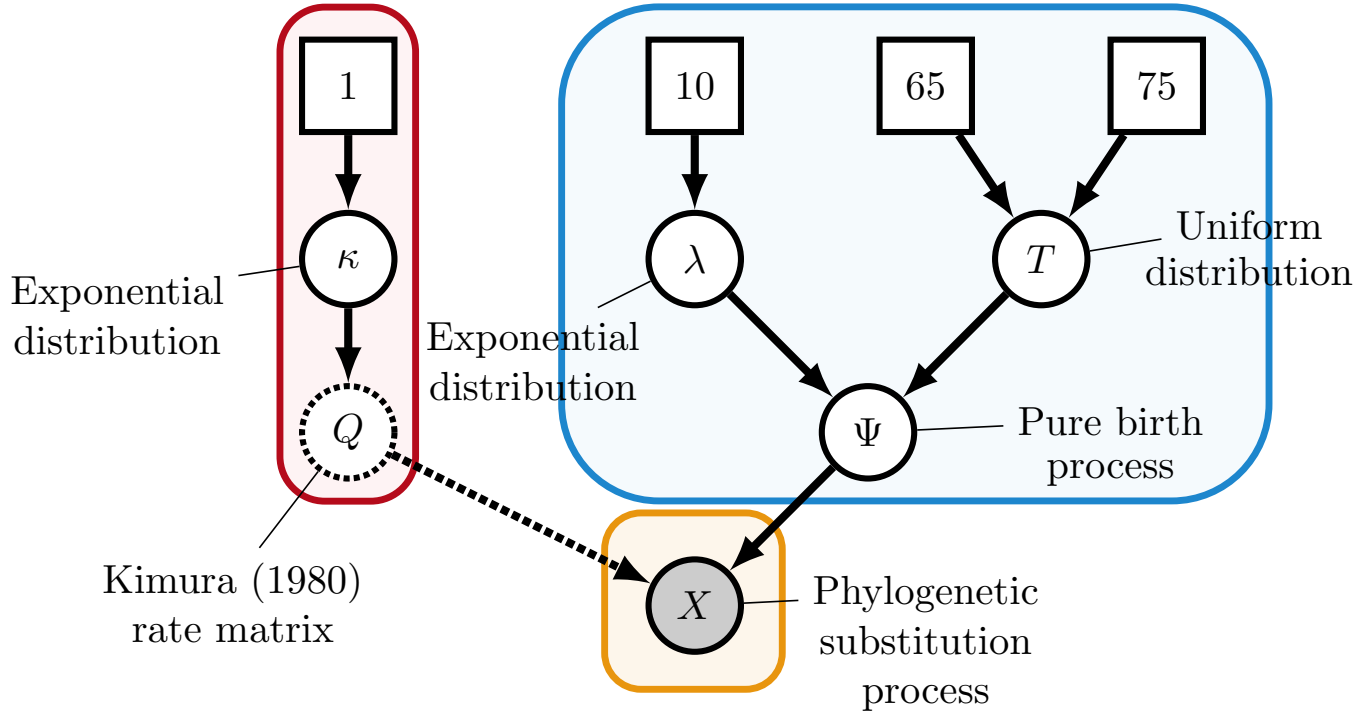
RevBayes script



Phylogenetic graphical model (Yule + K80)

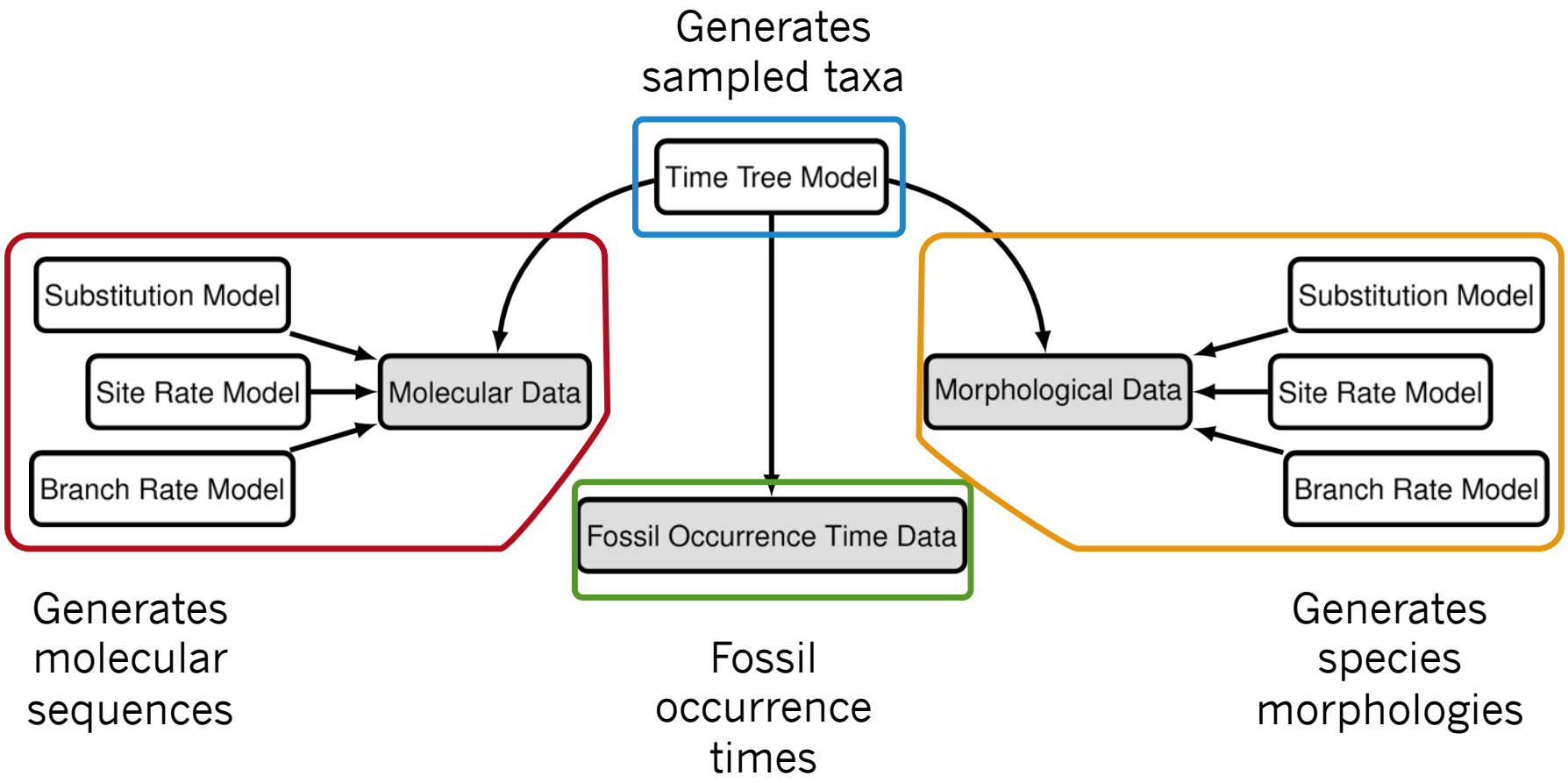
```
# pure birth prior and hyperpriors
source("module_tree_yule.Rev")


# rate matrix
source("module_Q_K80.Rev")

# phylogenetic substitution process
source("module_CTMC.Rev")
```

RevBayes script



Phylogenetic graphical model (Yule + K80)

Modular View of Combined-Evidence Analysis under a Fossilized Birth-Death Process



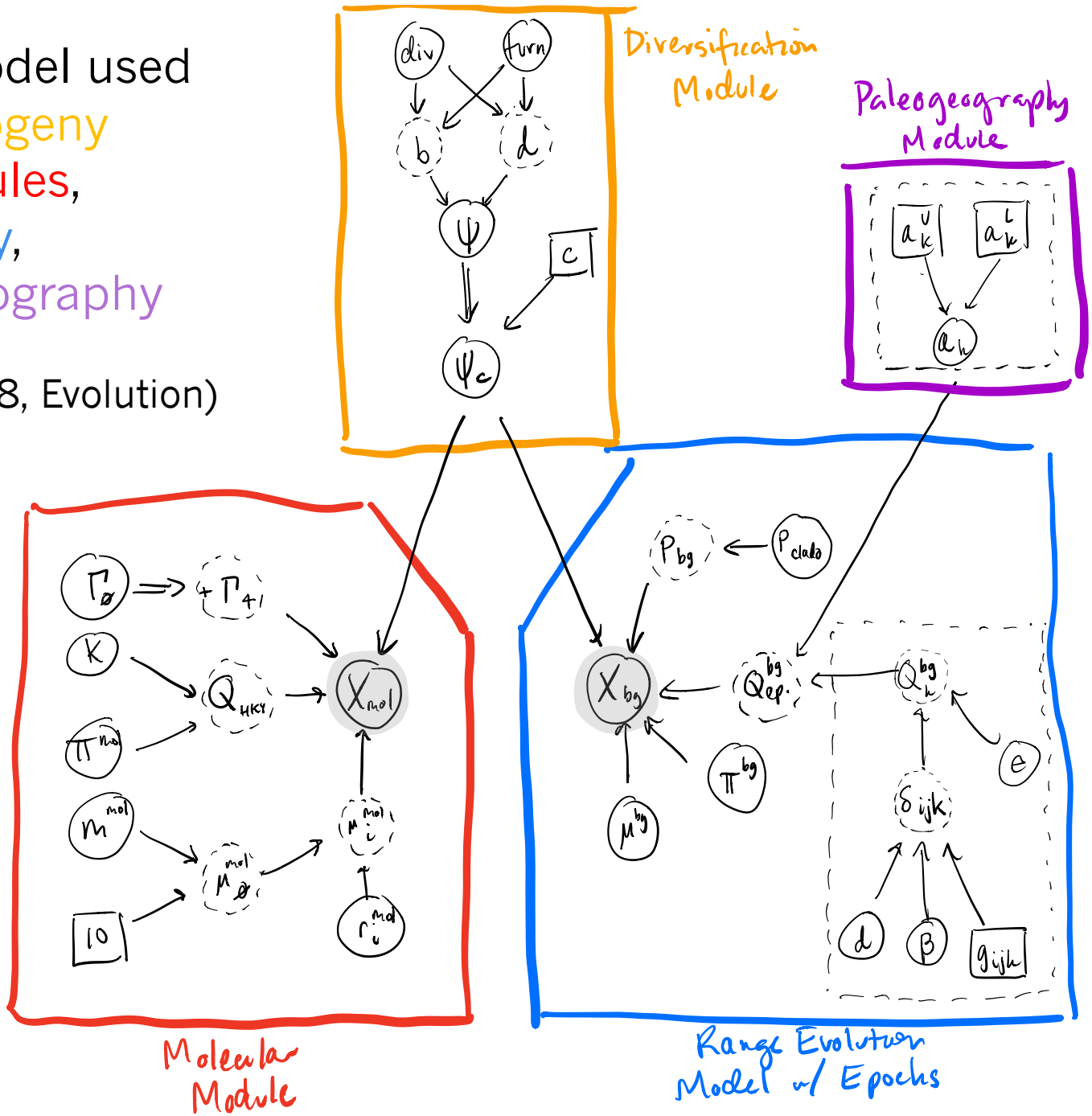
(Tracy will teach this evening)

https://revbayes.github.io/tutorials/fbd/

Sketch of model used to date phylogeny using molecules, biogeography, and paleogeography

Landis et al. (2018, Evolution)

Hawaiian silverswords



All possible
processes

Generates biological data

●

All possible processes

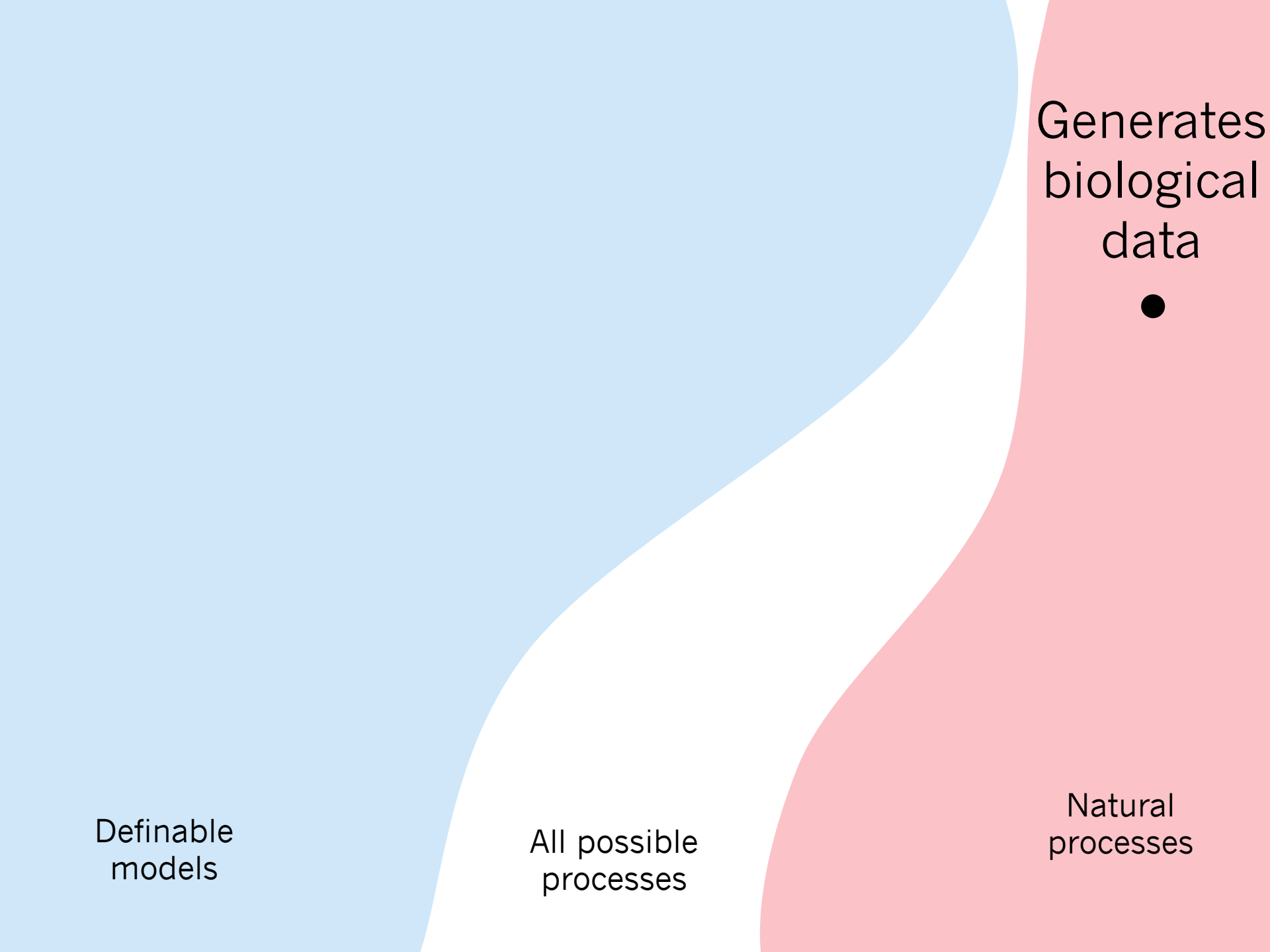
Natural processes

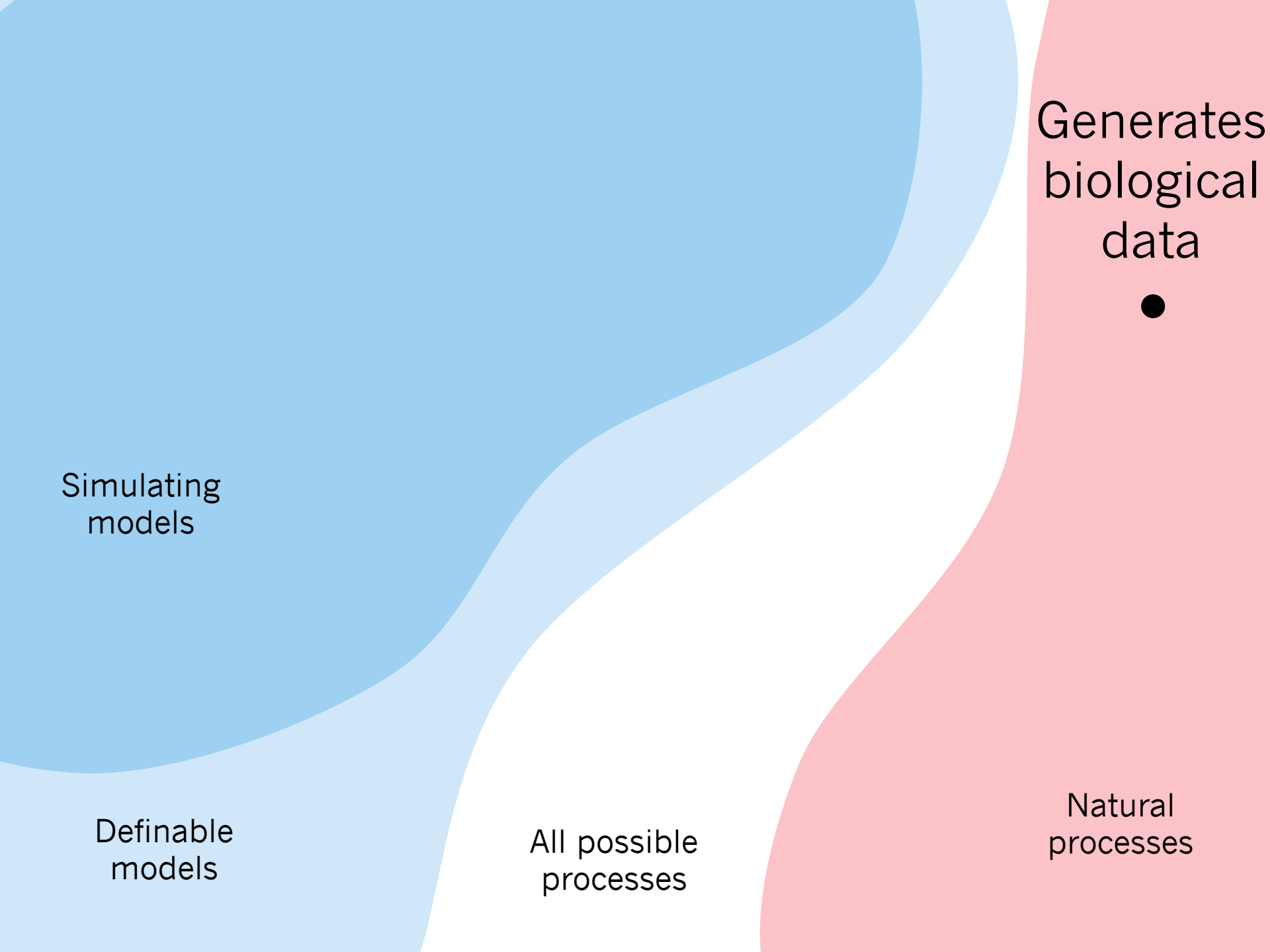
Studying evolution

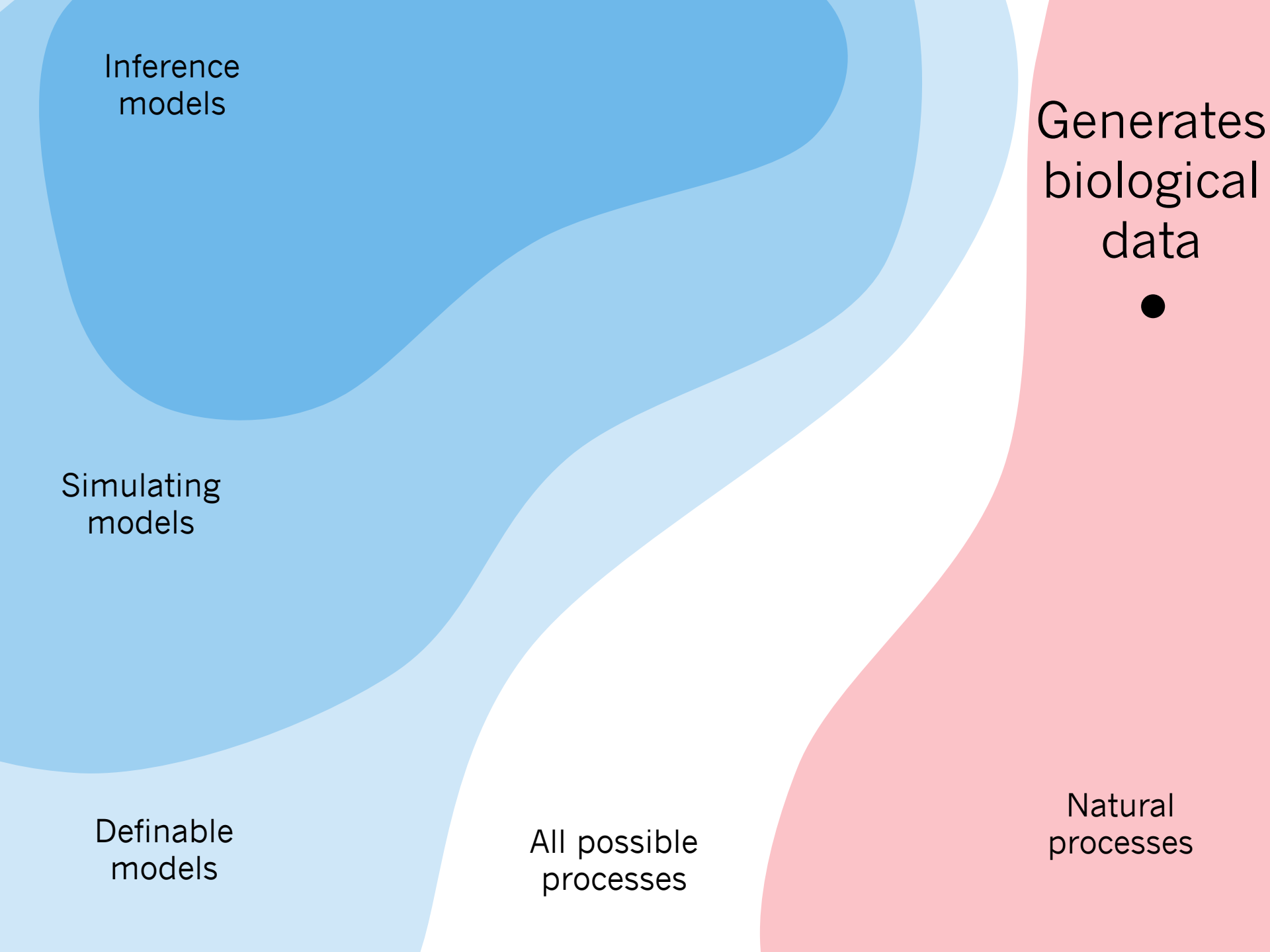
Common descent
    phylogenetic inference

Tempo & mode
    macroevolutionary models

Rewinding deep time
    ancestral state estimation


Generates biological data
Definable models
All possible processes
Natural processes


Generates biological data
Simulating models
Definable models
All possible processes
Natural processes


Inference models
Simulating models
Definable models
All possible processes
Natural processes
Generates biological data

Inference models

Simulating models

Definable models

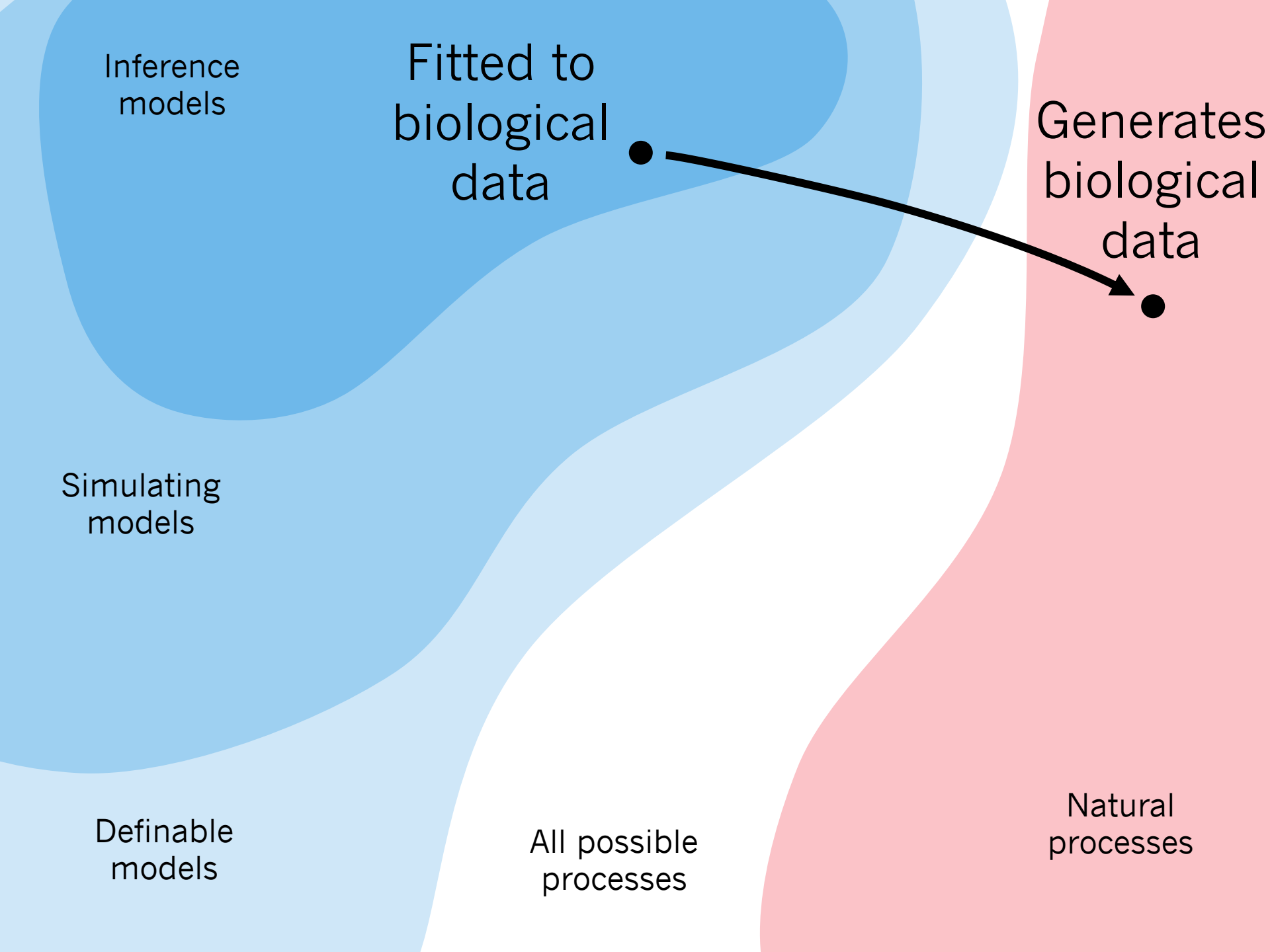
All possible processes

Generates biological data

●

~~No exact model~~

Natural processes


Inference models
Fitted to biological data
Generates biological data
Simulating models
Definable models
All possible processes
Natural processes

Inference models

Fitted to biological data

Generates biological data

Simulating models

How useful is our wrong model?

Definable models

All possible processes

Box

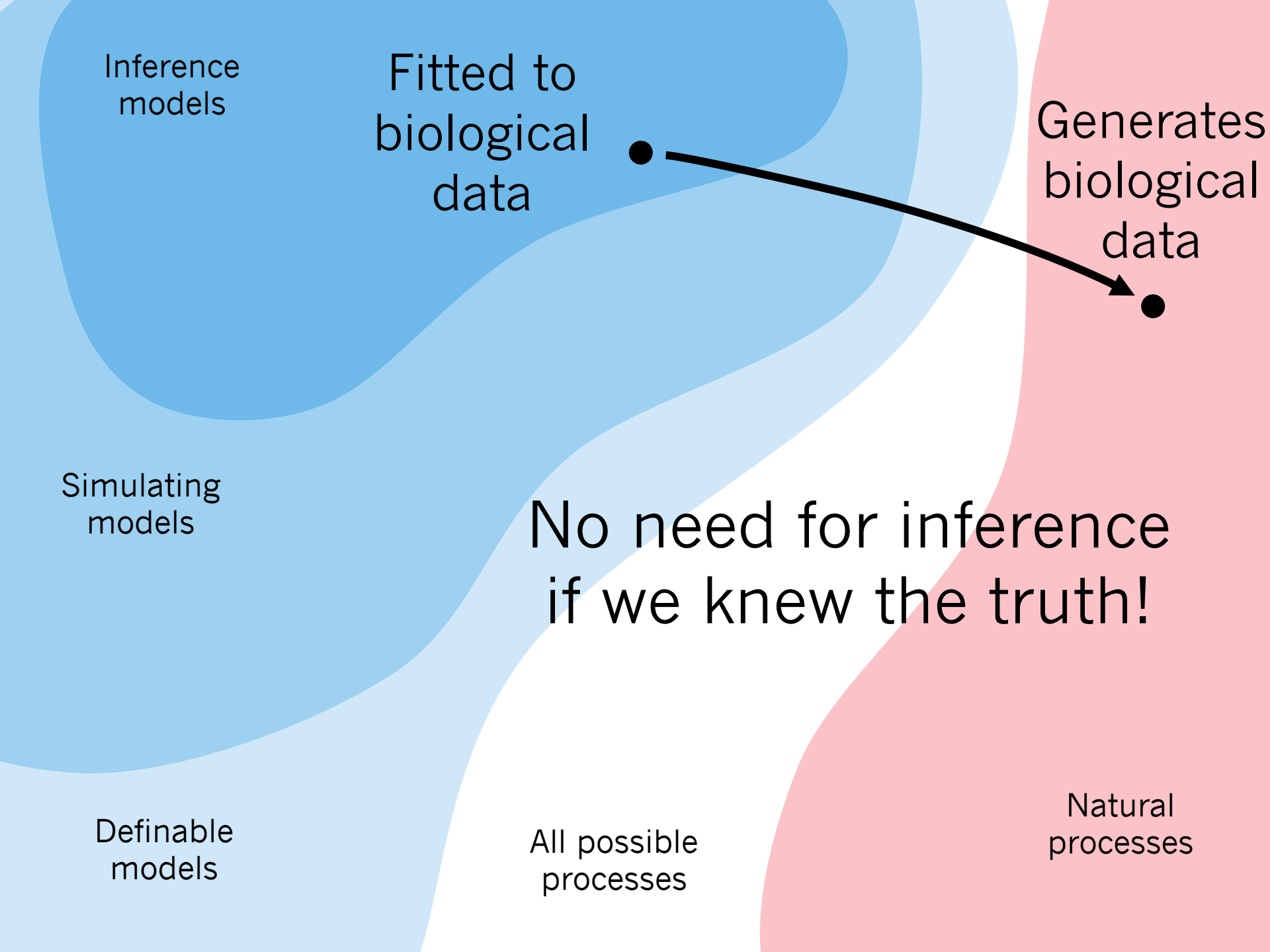
Natural processes

Assessing models

accuracy	how close are our estimates to the true parameter values?
precision	how are we certain of our parameter estimates?
power	what size datasets are needed to estimate the truth?
sensitivity	are parameter estimates responsive to prior/data changes?
performance	does the model require excessive time or resources?
selection	do we reliably select the true model among competing models?
adequacy	does the model generate data resembling the true data?

Assessing models

accuracy	how close are our estimates to the true parameter values?
precision	how are we certain of our parameter estimates?
power	what size datasets are needed to estimate the truth?
sensitivity	are parameter estimates responsive to prior/data changes?
performance	does the model require excessive time or resources?
selection	do we reliably select the true model among competing models?
→ adequacy	does the model generate data resembling the true data?


Inference models
Fitted to biological data
Generates biological data
Simulating models
No need for inference if we knew the truth!
Definable models
All possible processes
Natural processes

Simulating the truth

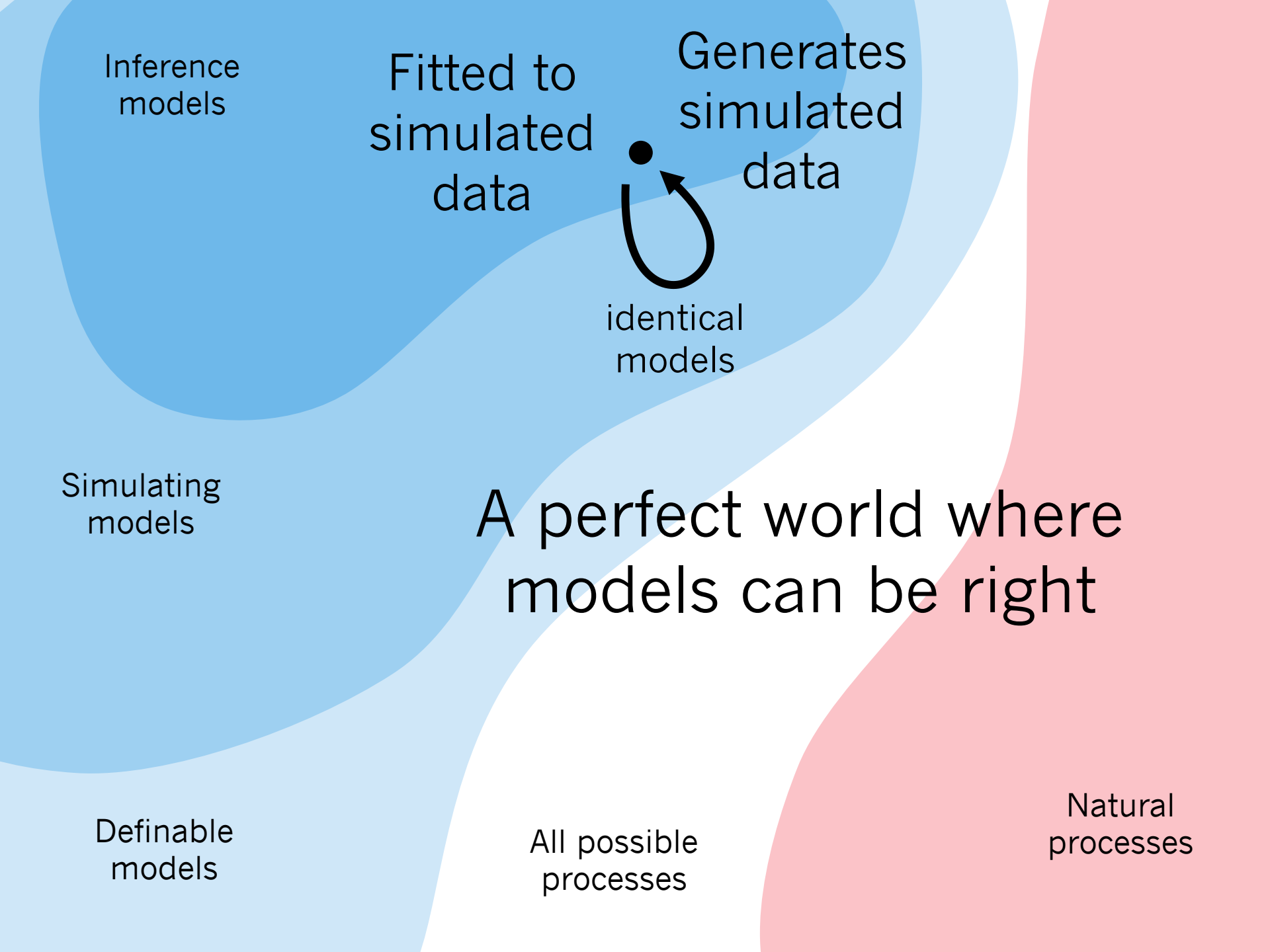
Poor model performance under known conditions will be equally bad (or worse) under unknown conditions

Simulating the truth

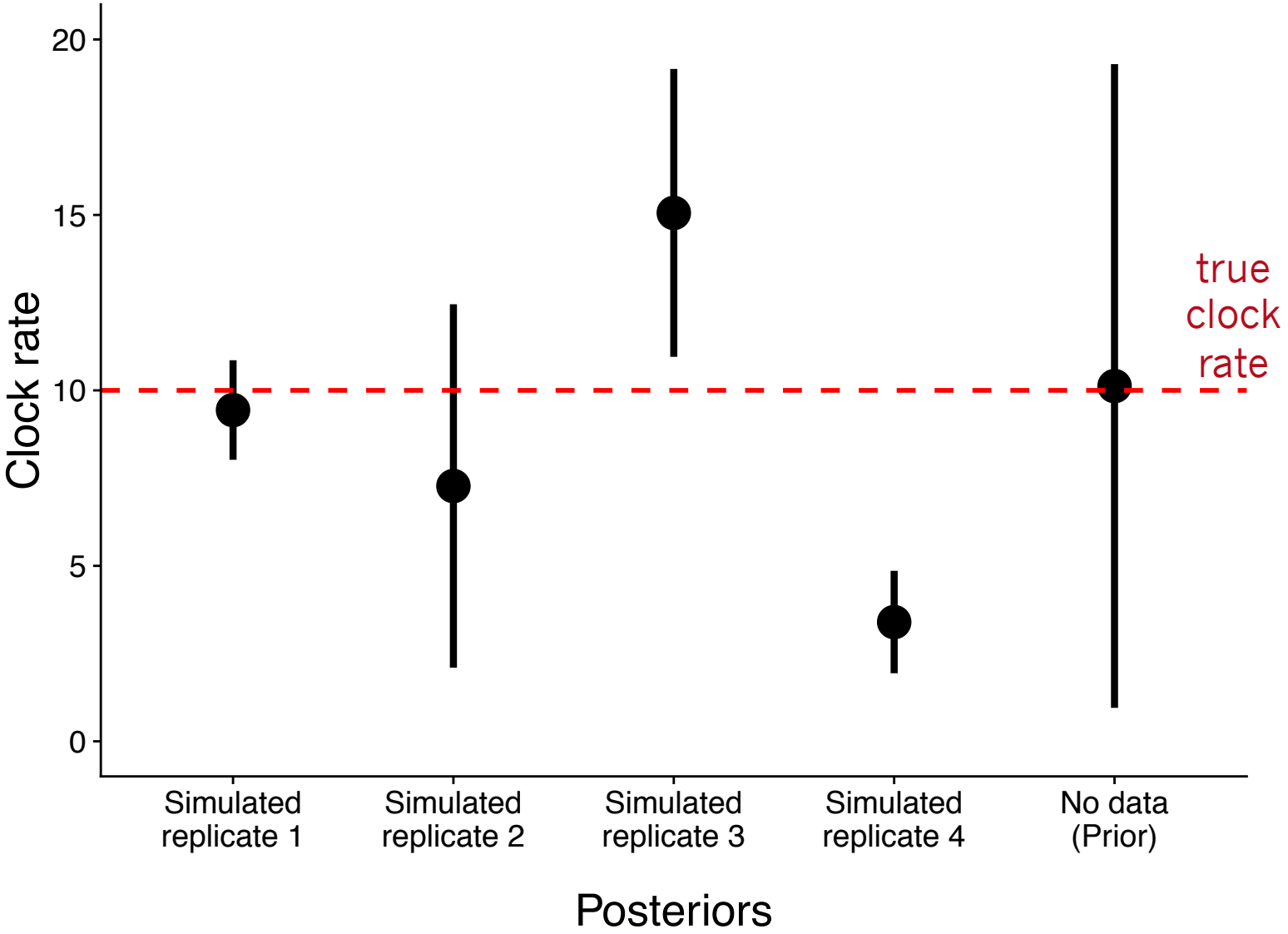
Poor model performance under known conditions will be equally bad (or worse) under unknown conditions

Simulations are crucial!

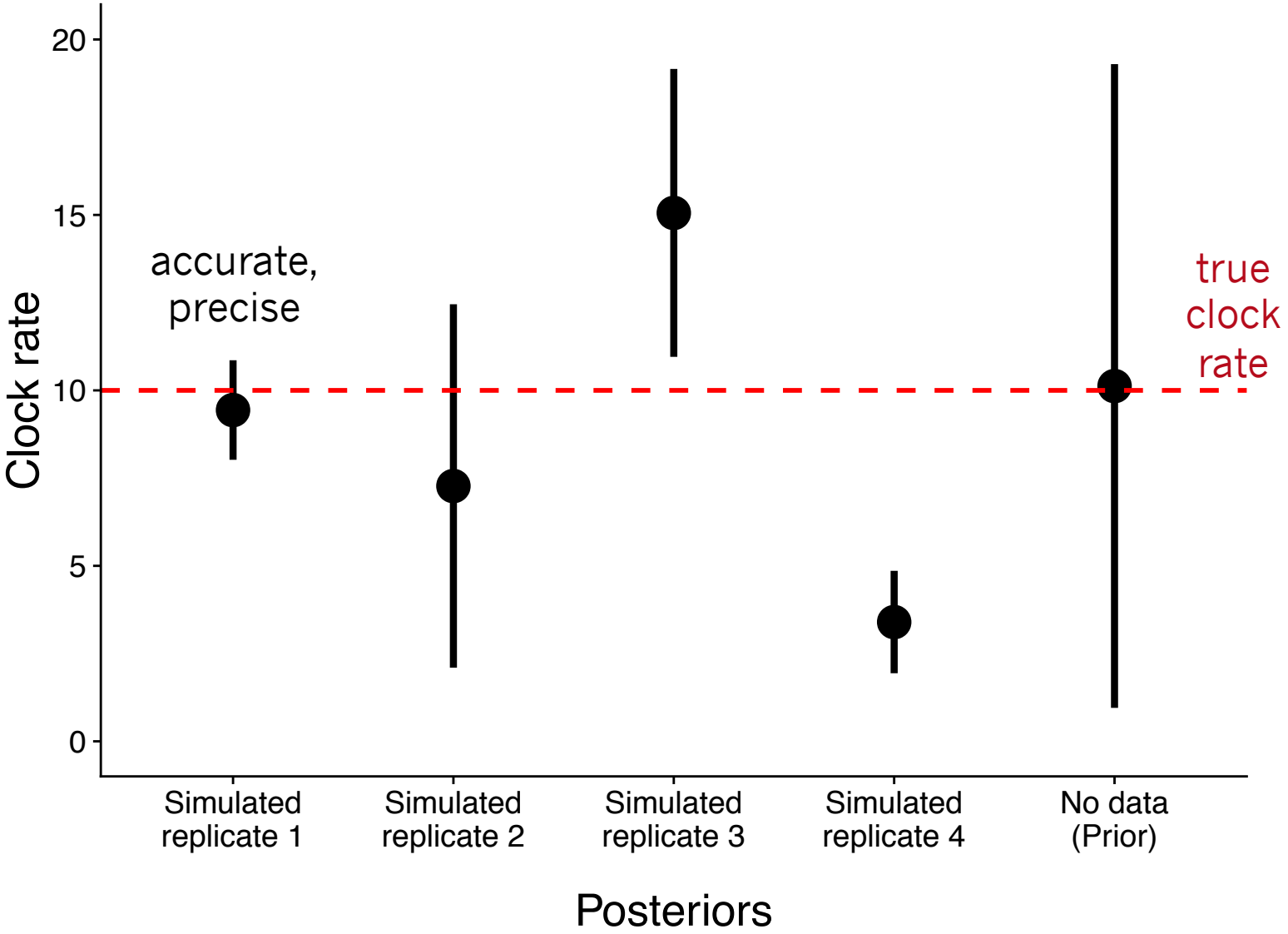
1. Test model/method behavior
2. Reveal a model's personality
3. Provide context for empirical results


Inference models
Fitted to simulated data
Generates simulated data
identical models
Simulating models
A perfect world where models can be right
Definable models
All possible processes
Natural processes

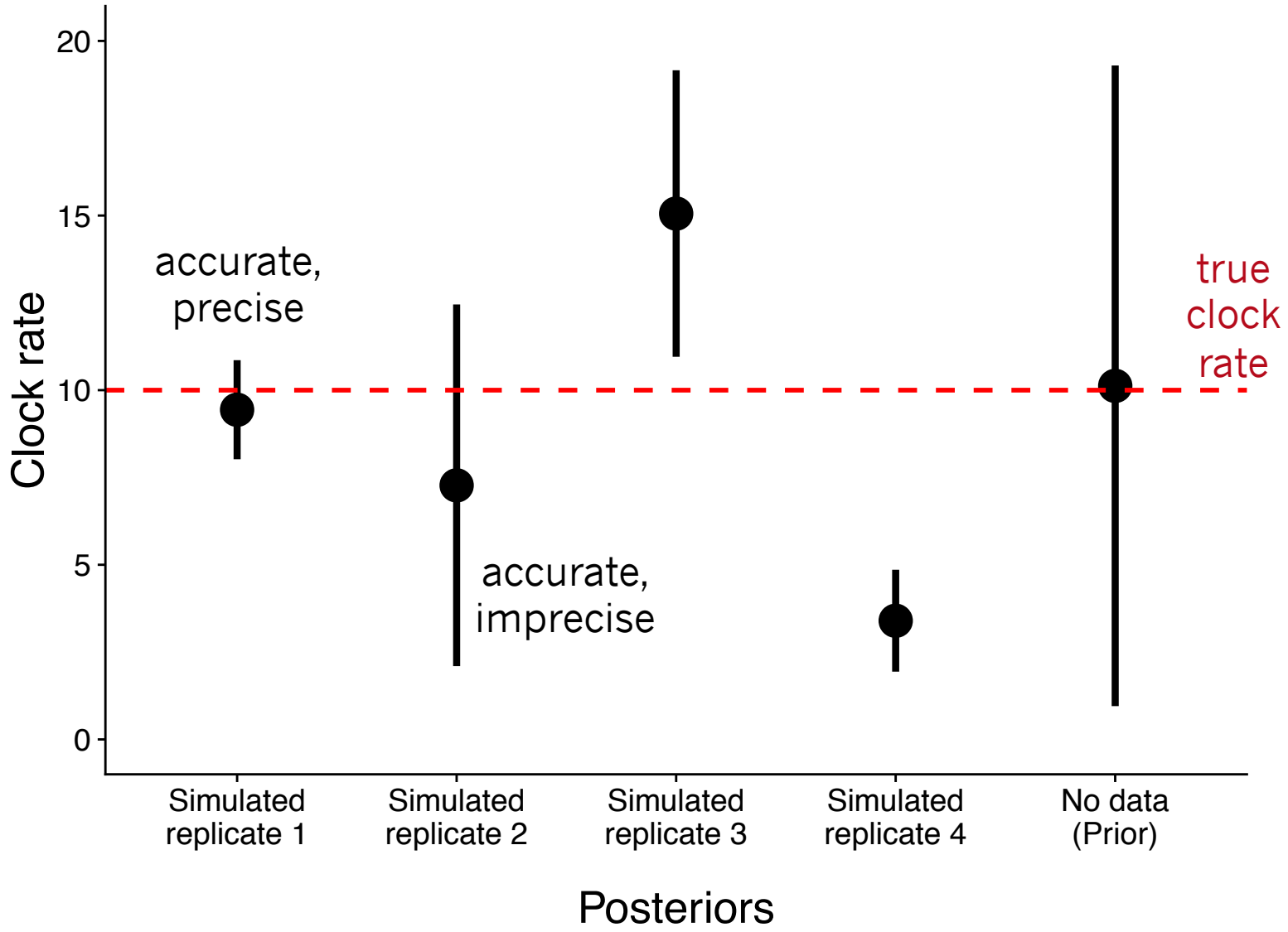
Accuracy and precision



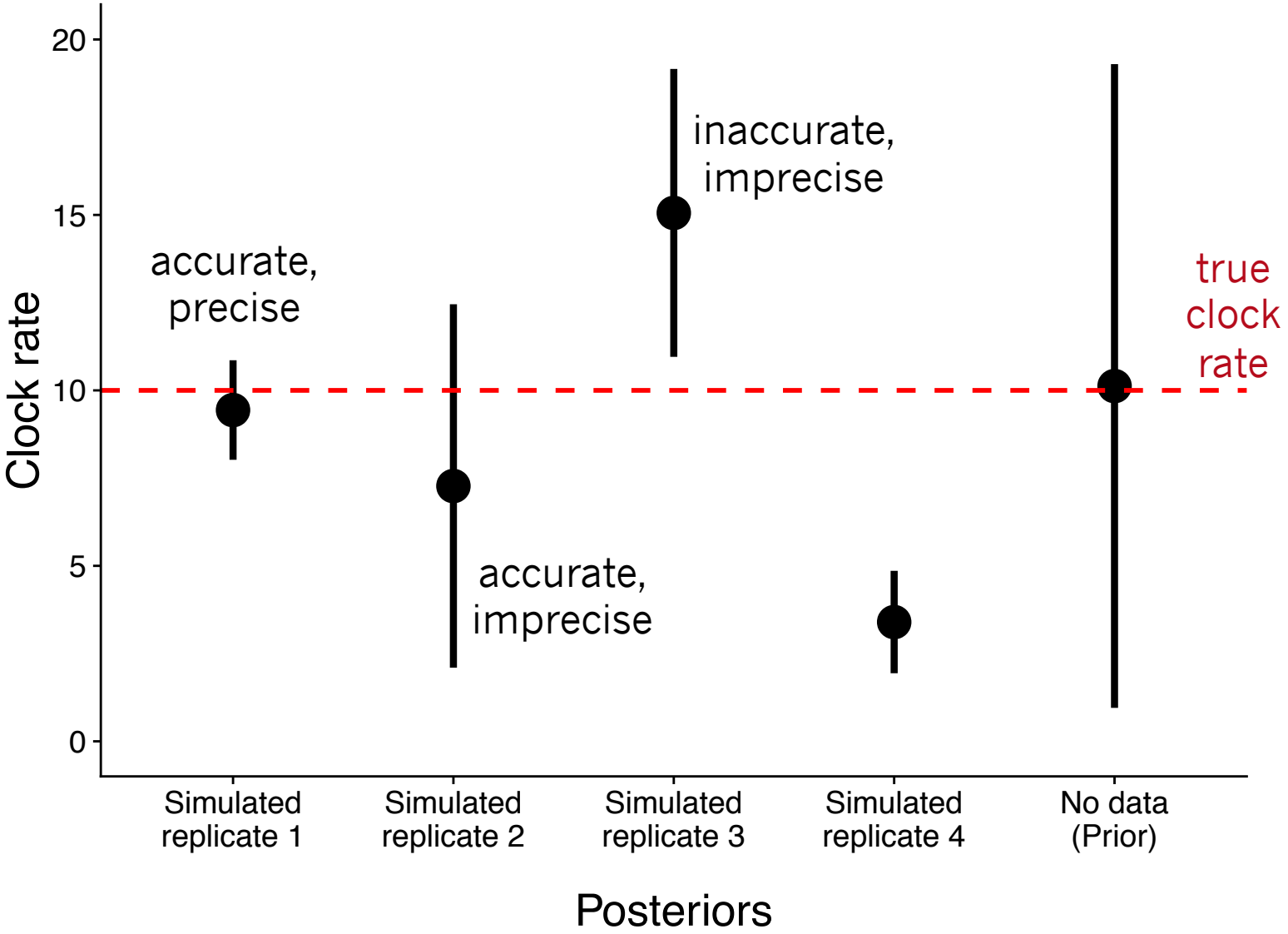
Accuracy and precision



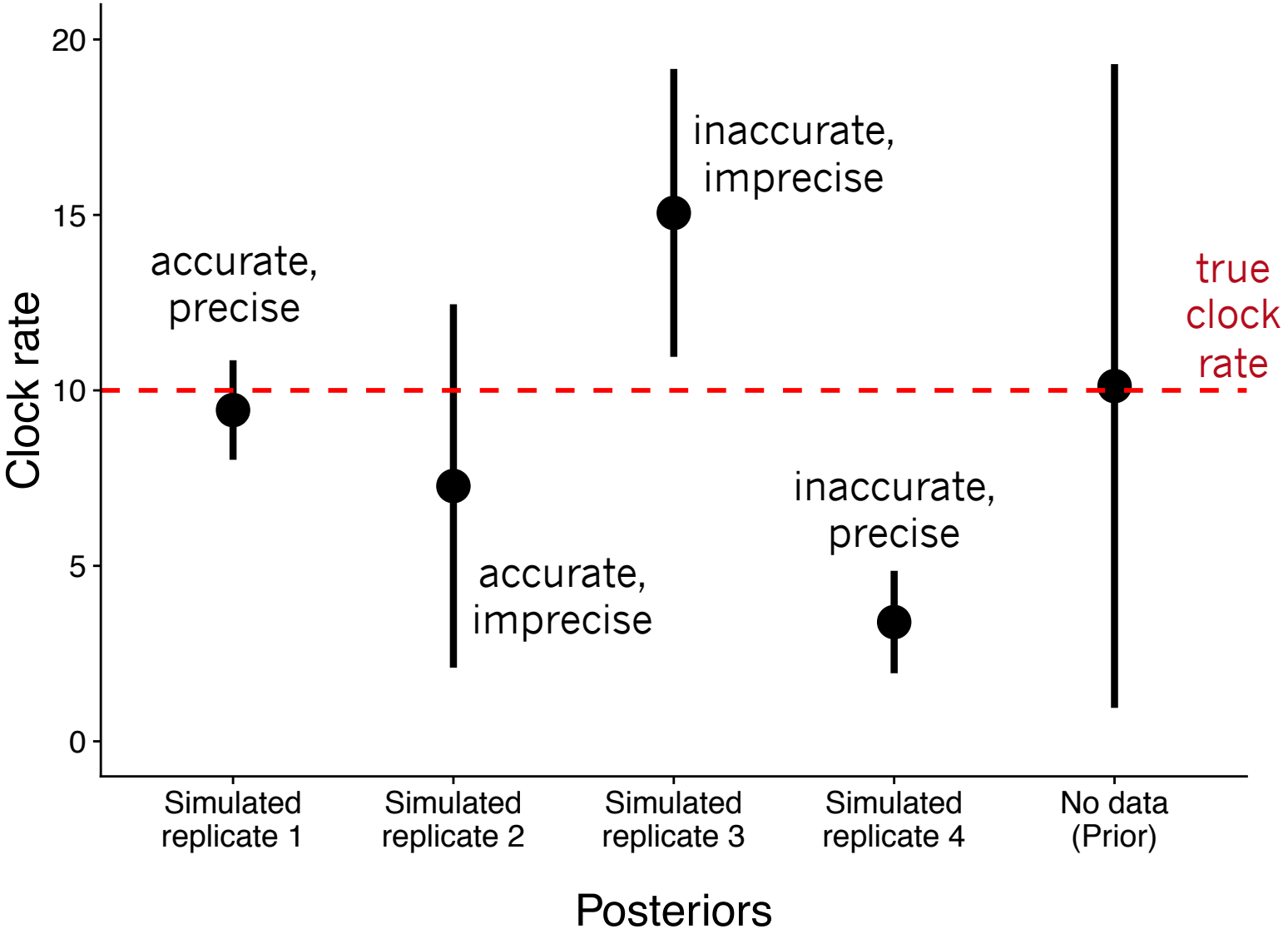
Accuracy and precision



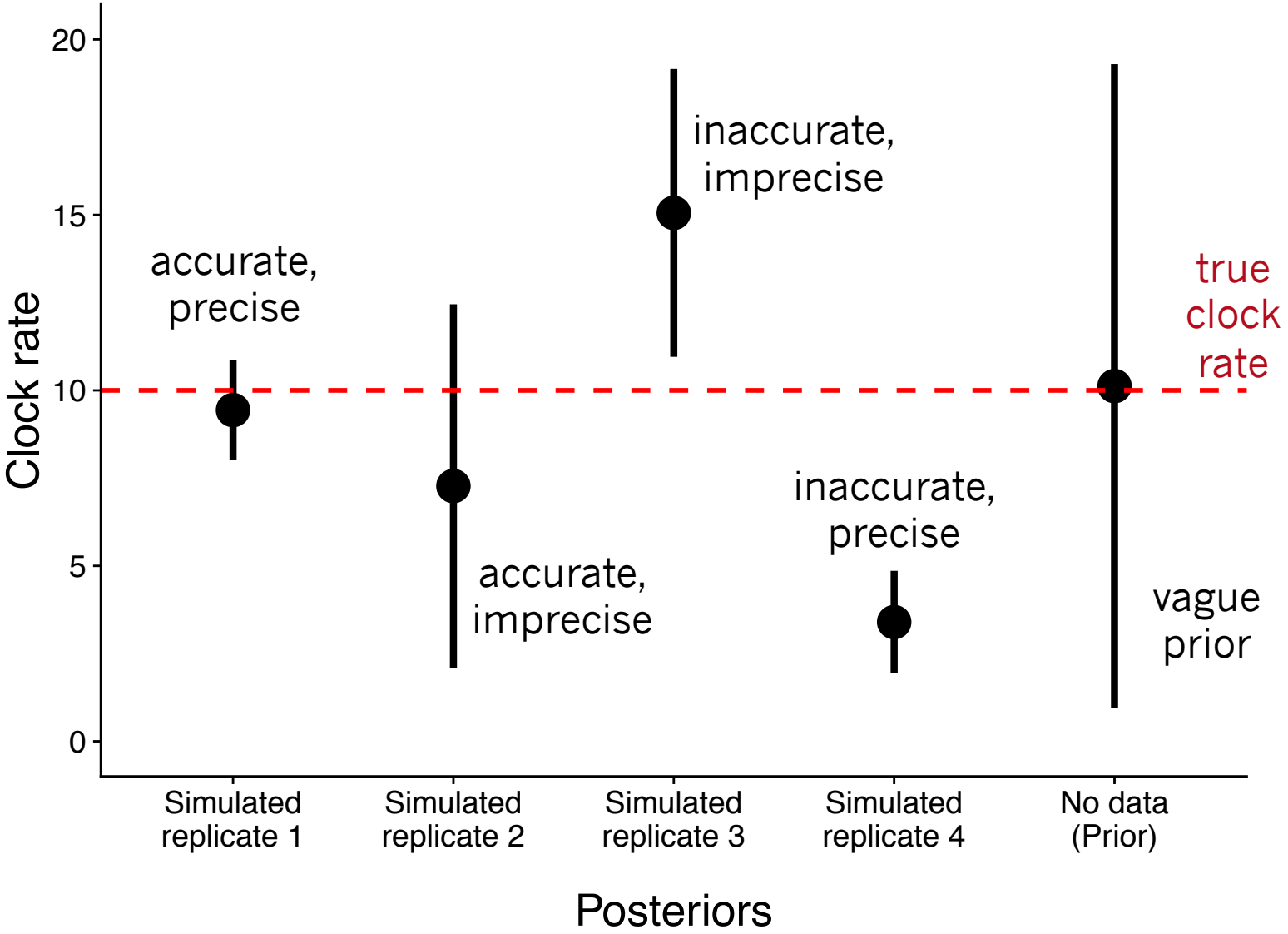
Accuracy and precision



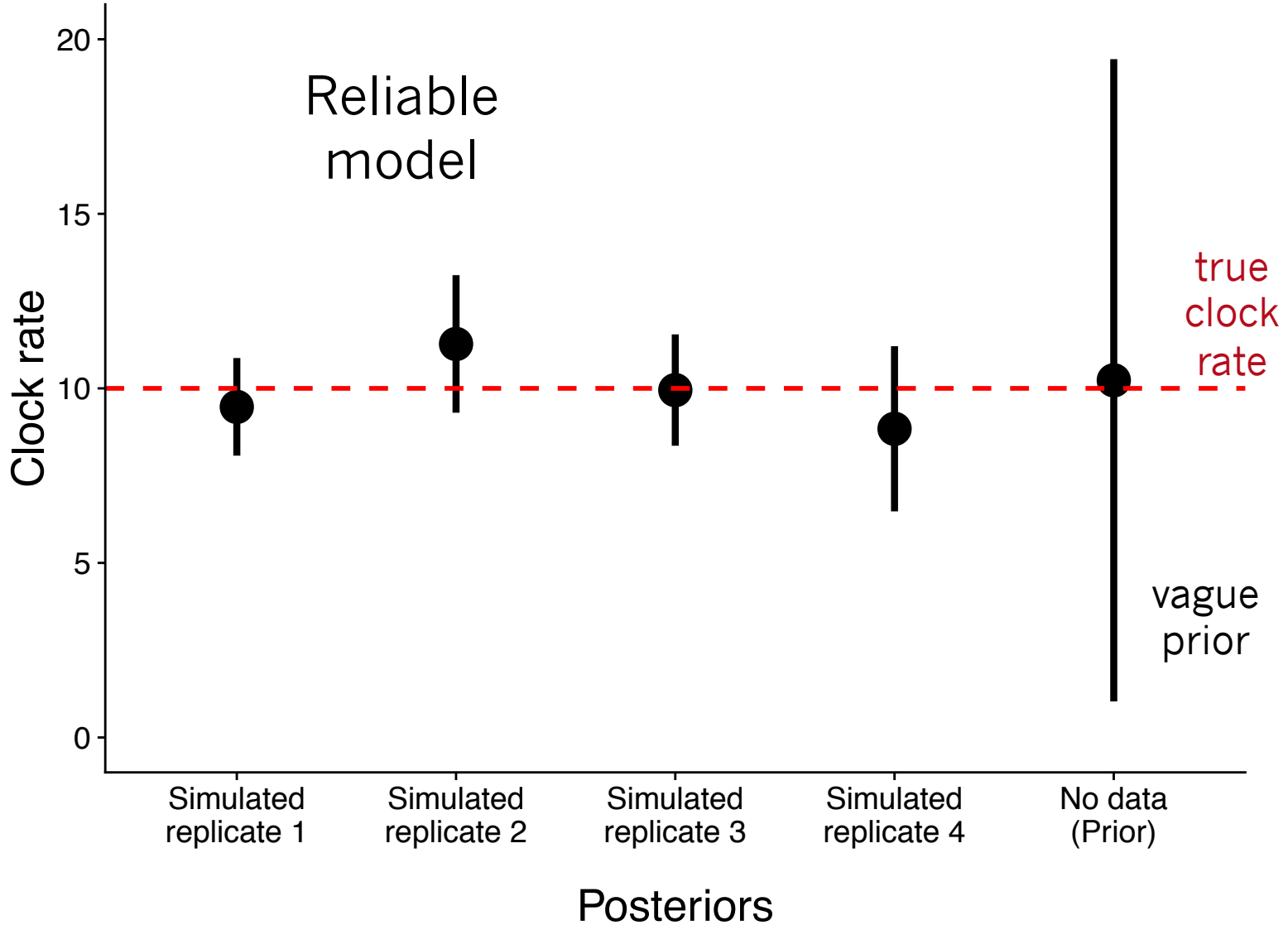
Accuracy and precision



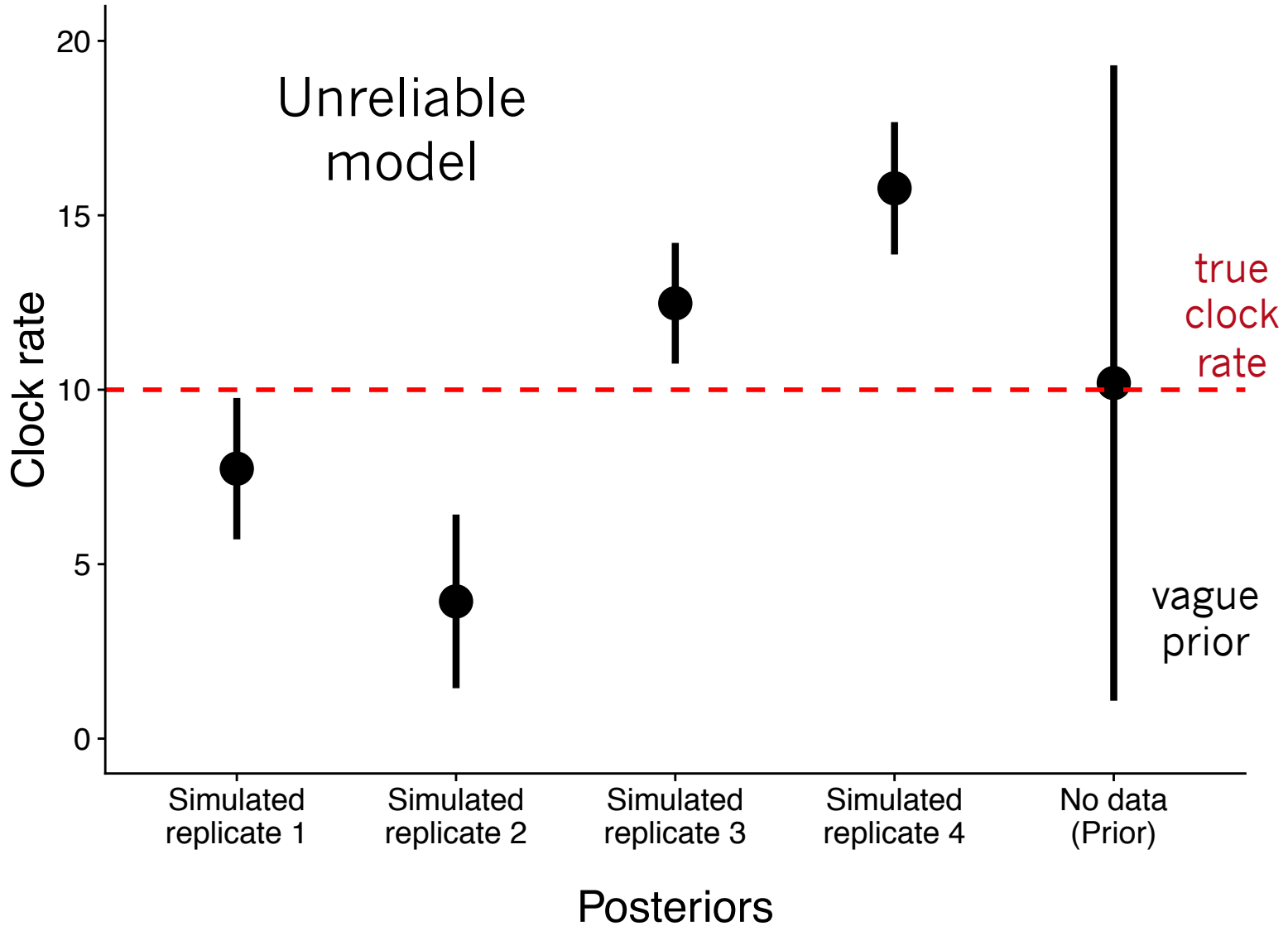
Accuracy and precision



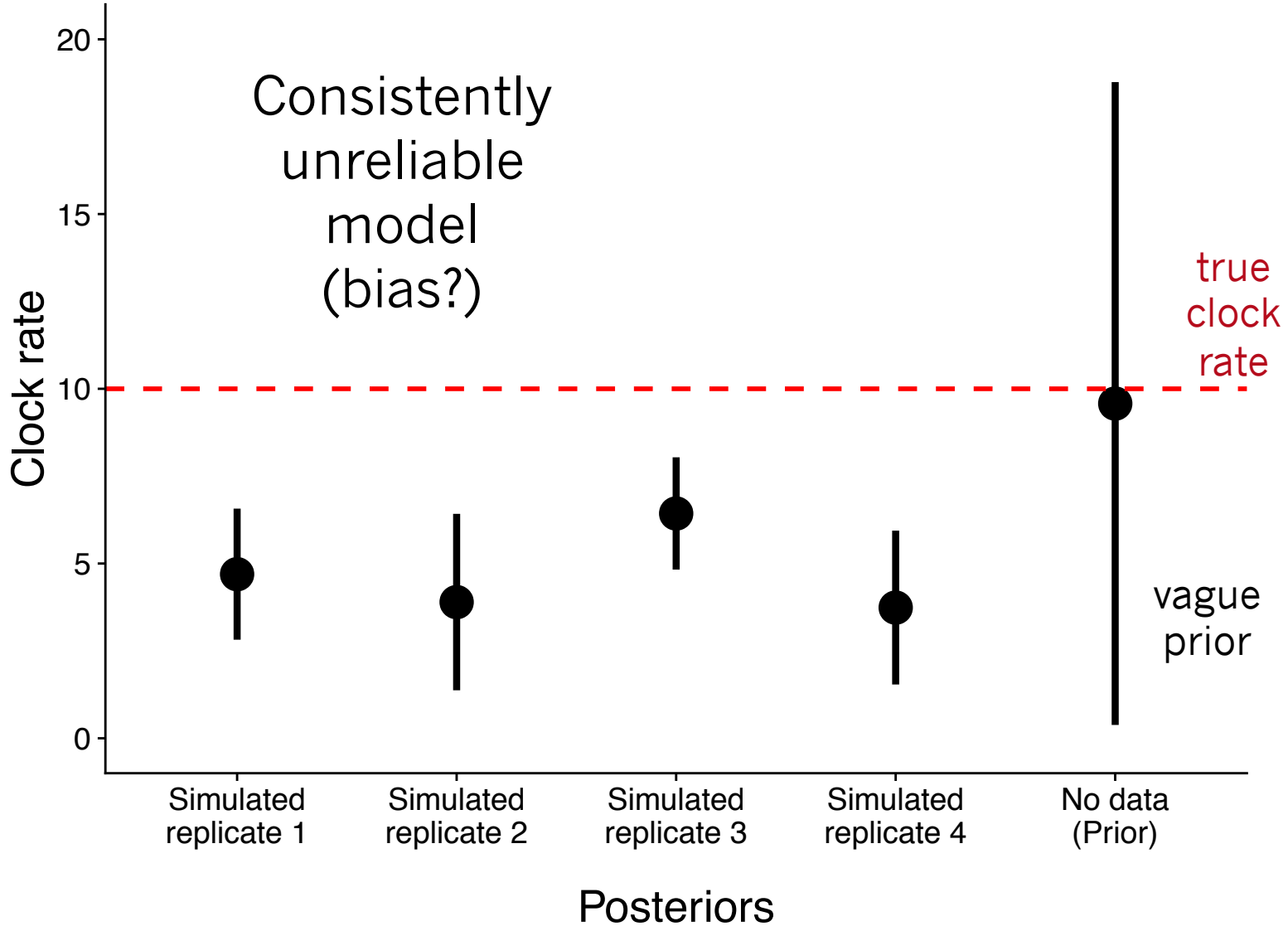
Accuracy and precision



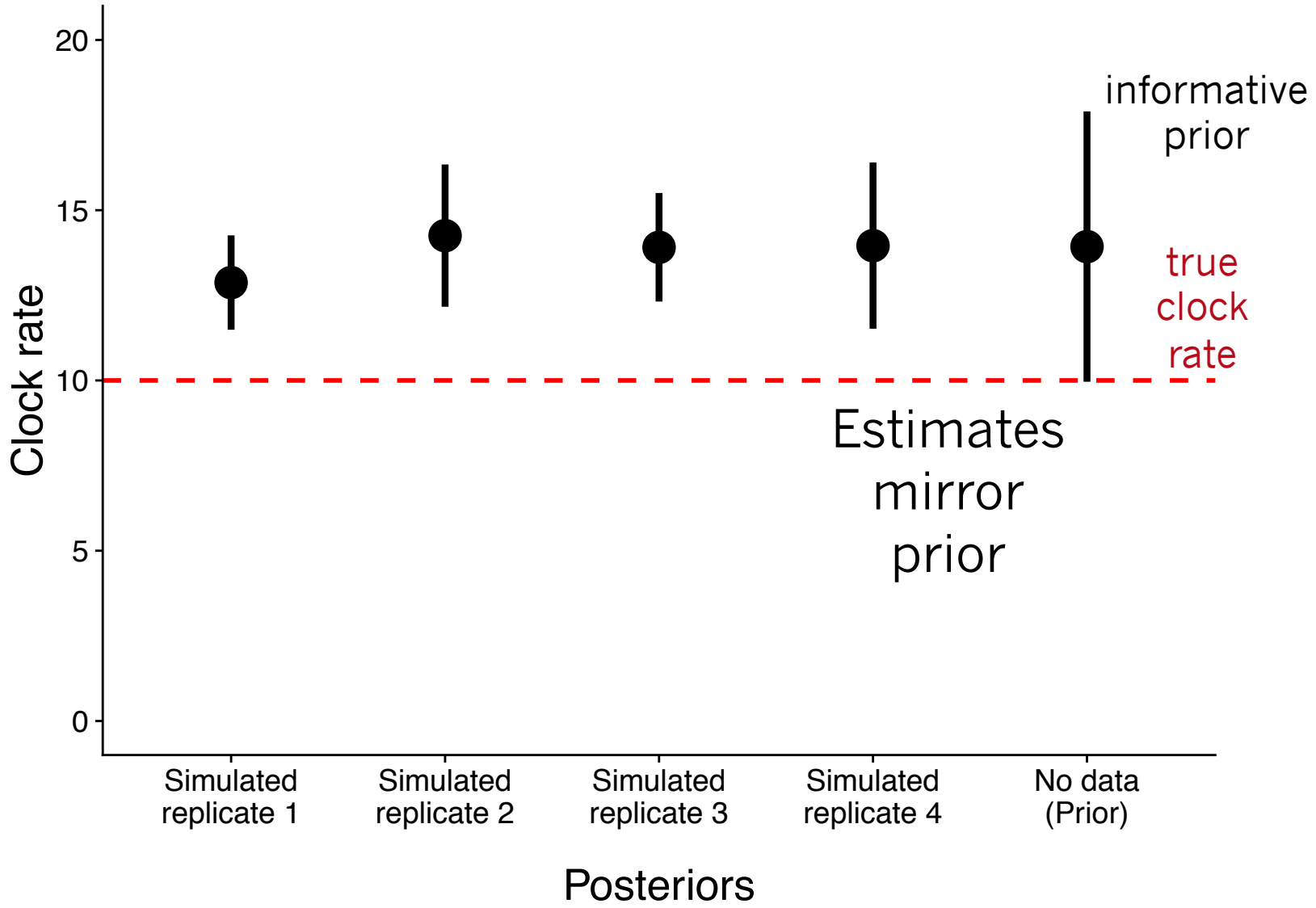
Accuracy and precision



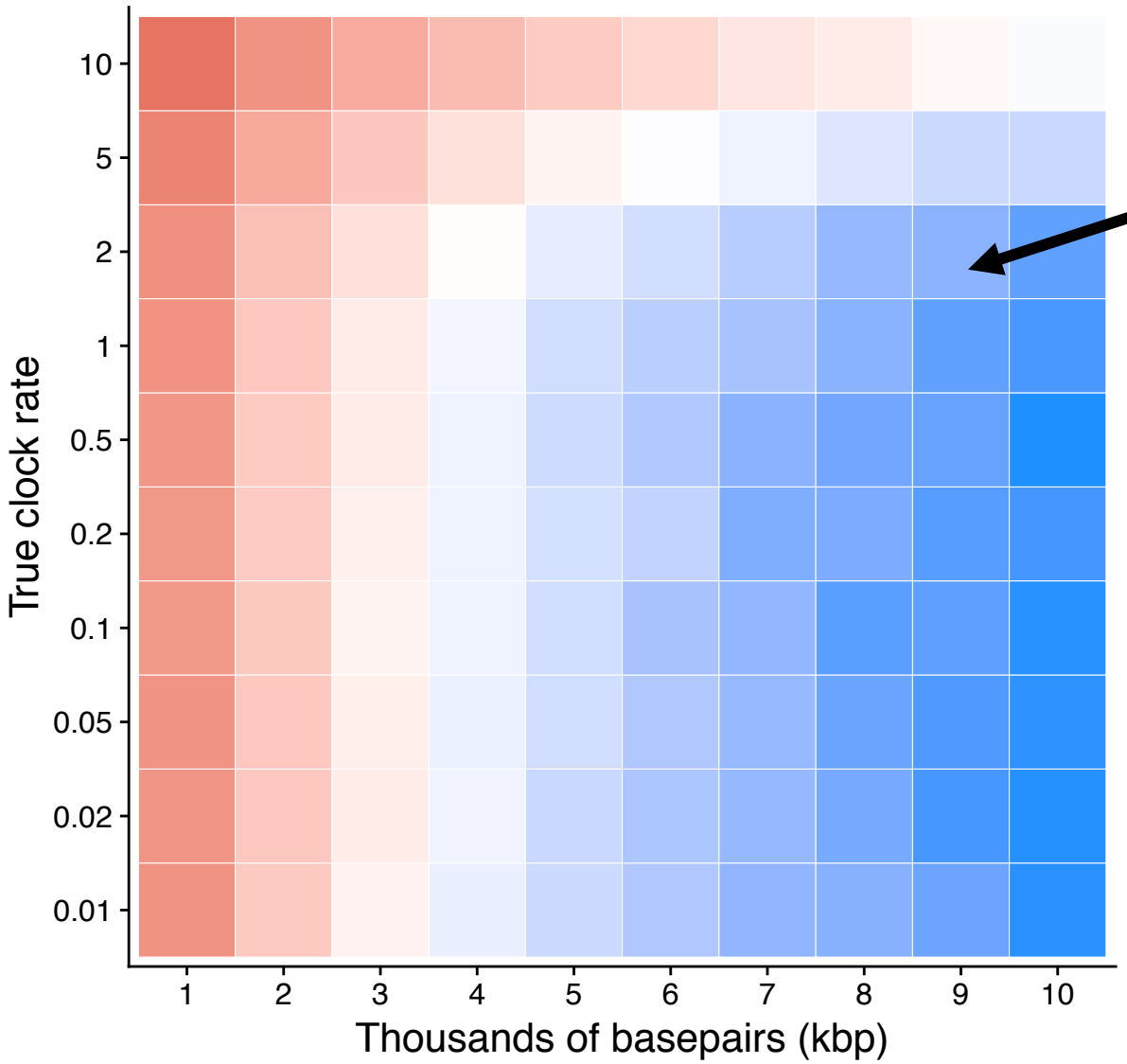
Accuracy and precision

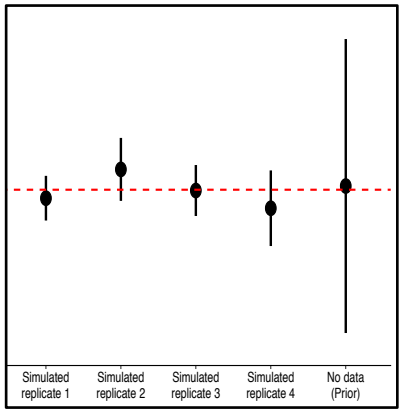


Prior sensitivity



Accuracy and power





Accuracy measurement

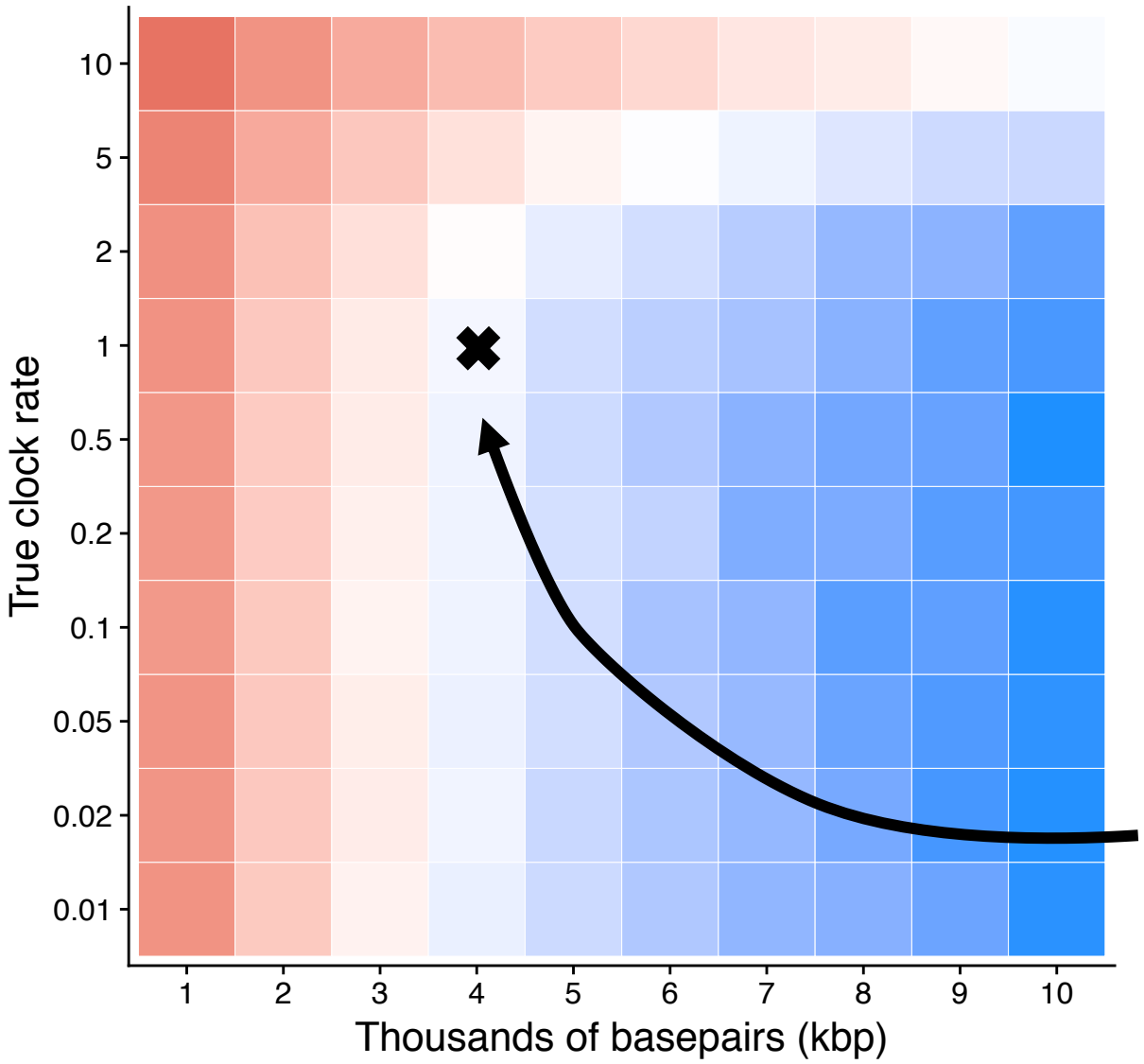
1 good

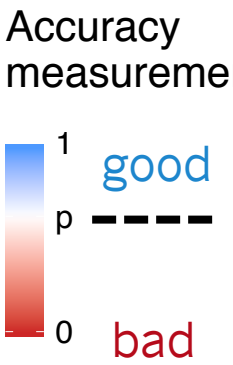
p

0 bad

Each cell reports frequency of pass/fail for an accuracy test for 100 tests

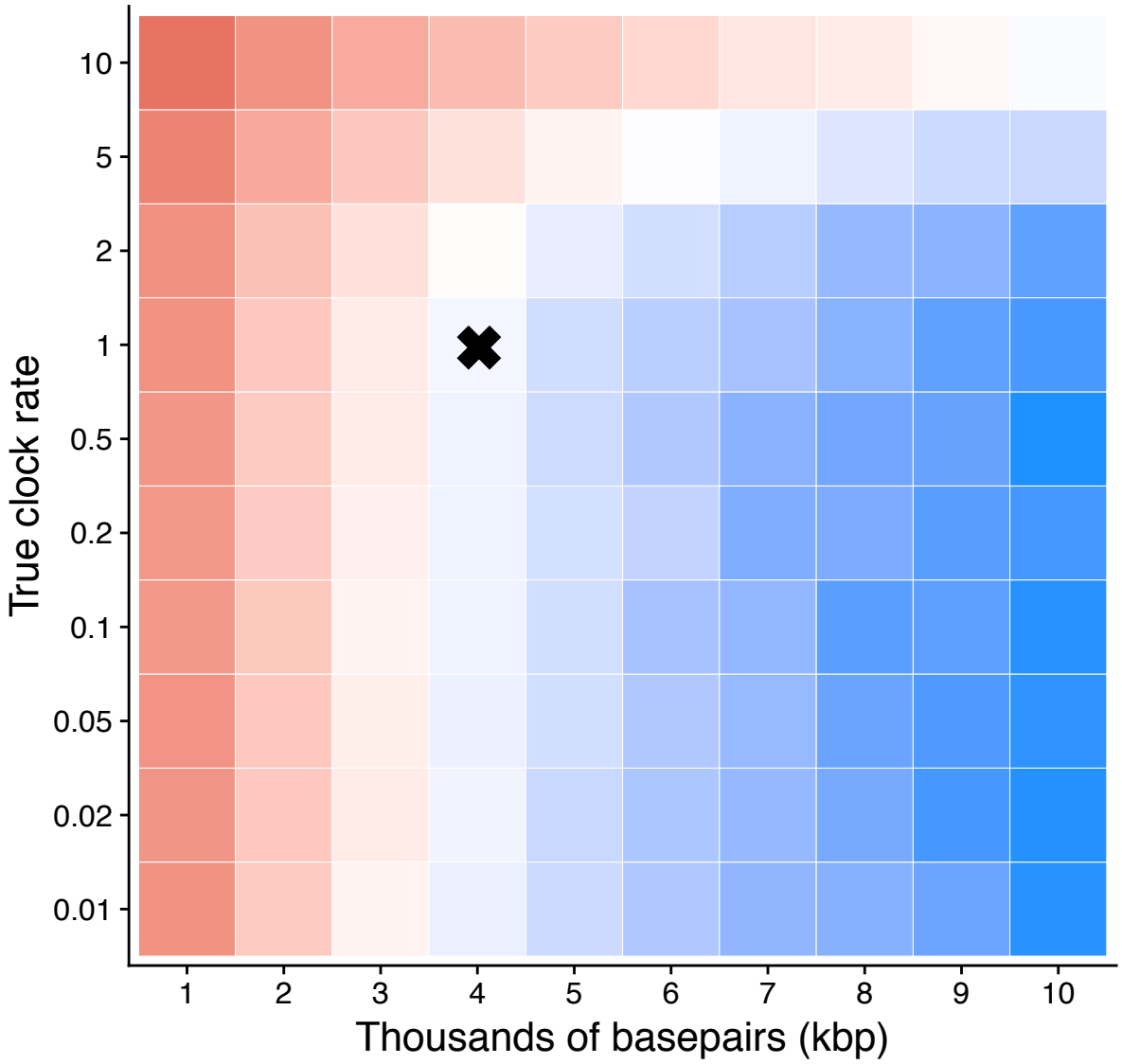
Accuracy and power





Could we estimate a rate equal to 1 using 4kbp of biological data?

Accuracy and power



Only 4 kbp needed for desired *accuracy* for rate = 1

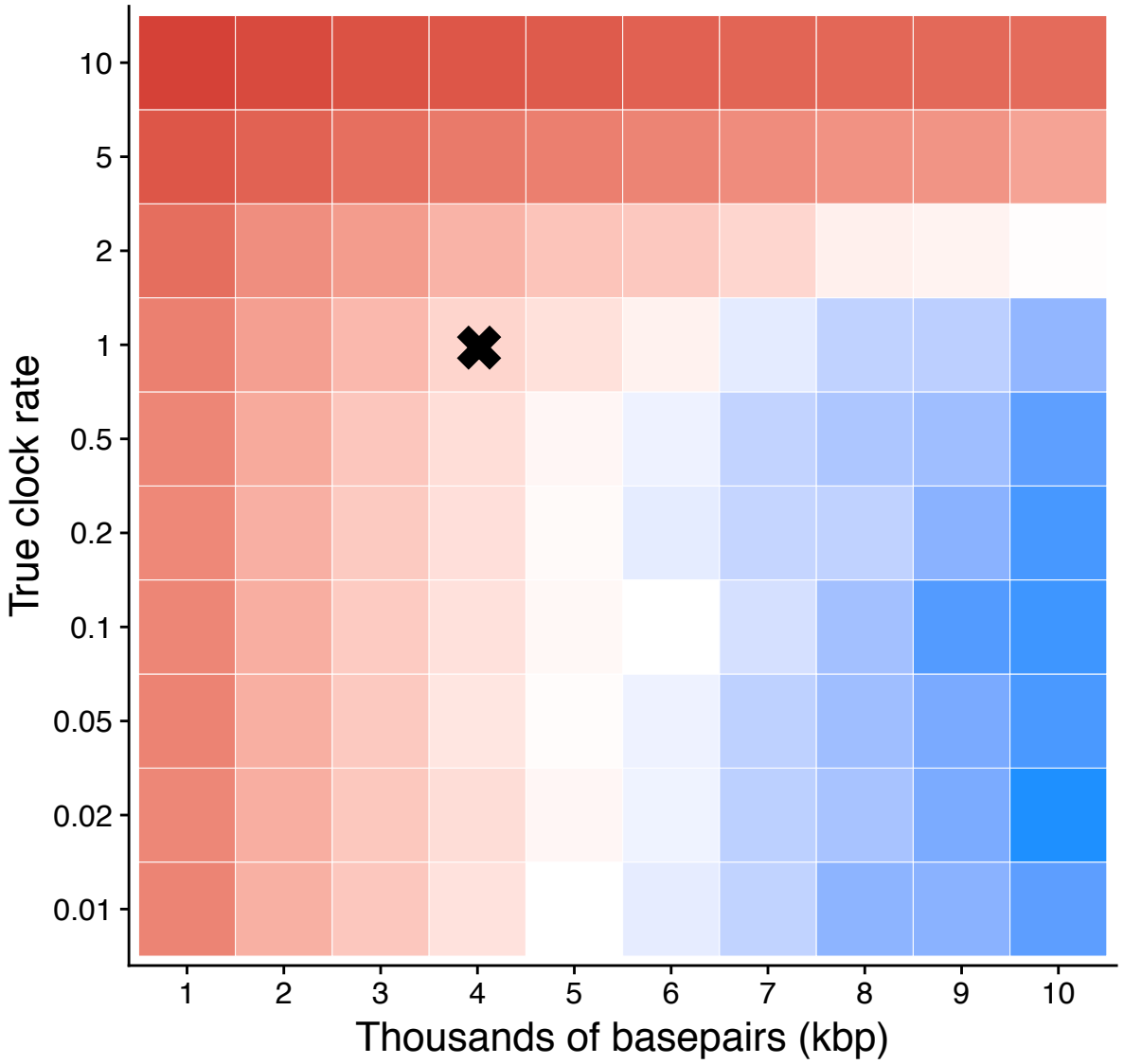
Accuracy measurement

1 good

p

0 bad

Precision and power



Only 4 kbp needed for desired *accuracy* for rate = 1

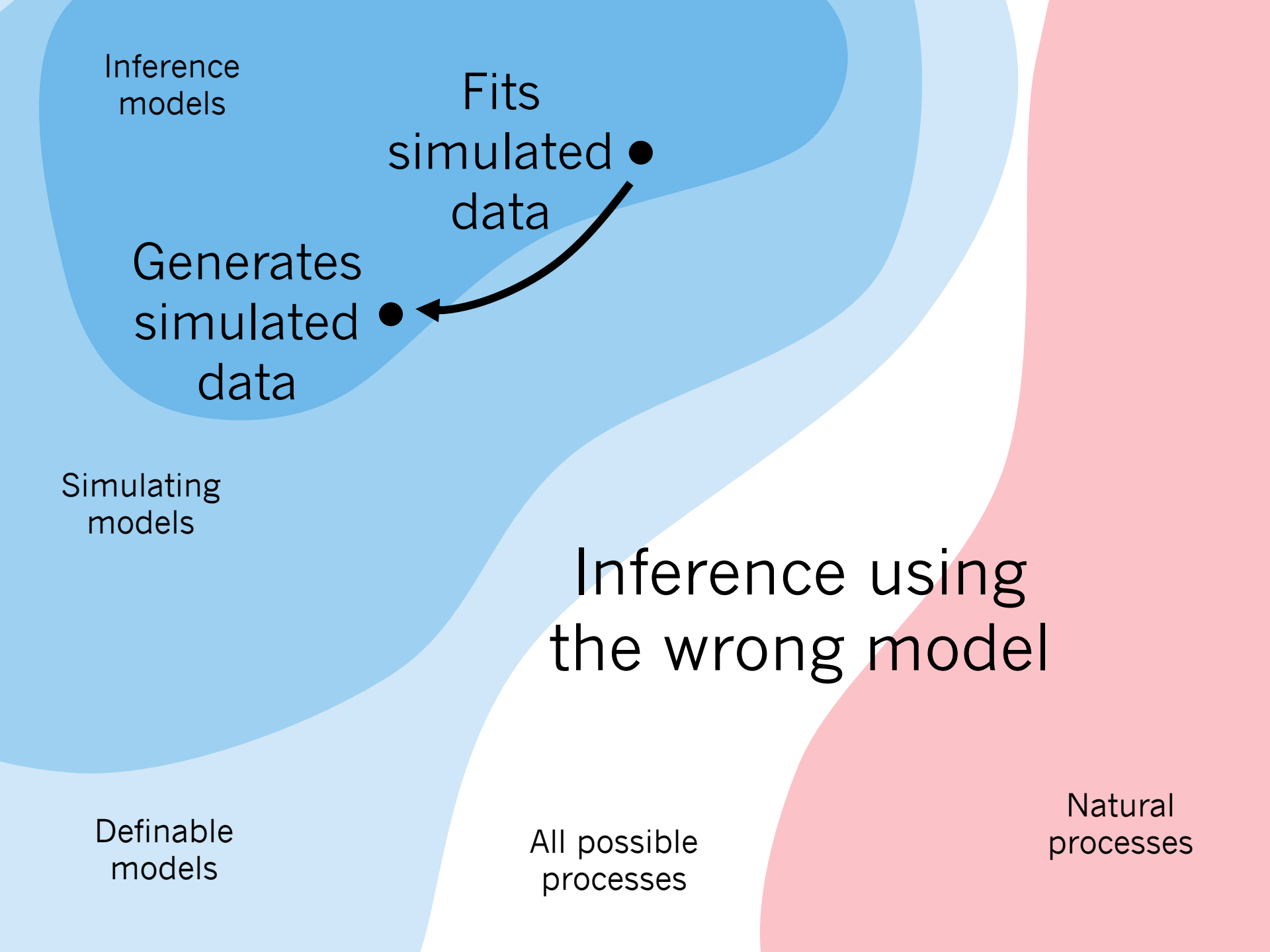
Precision measurement

1 good

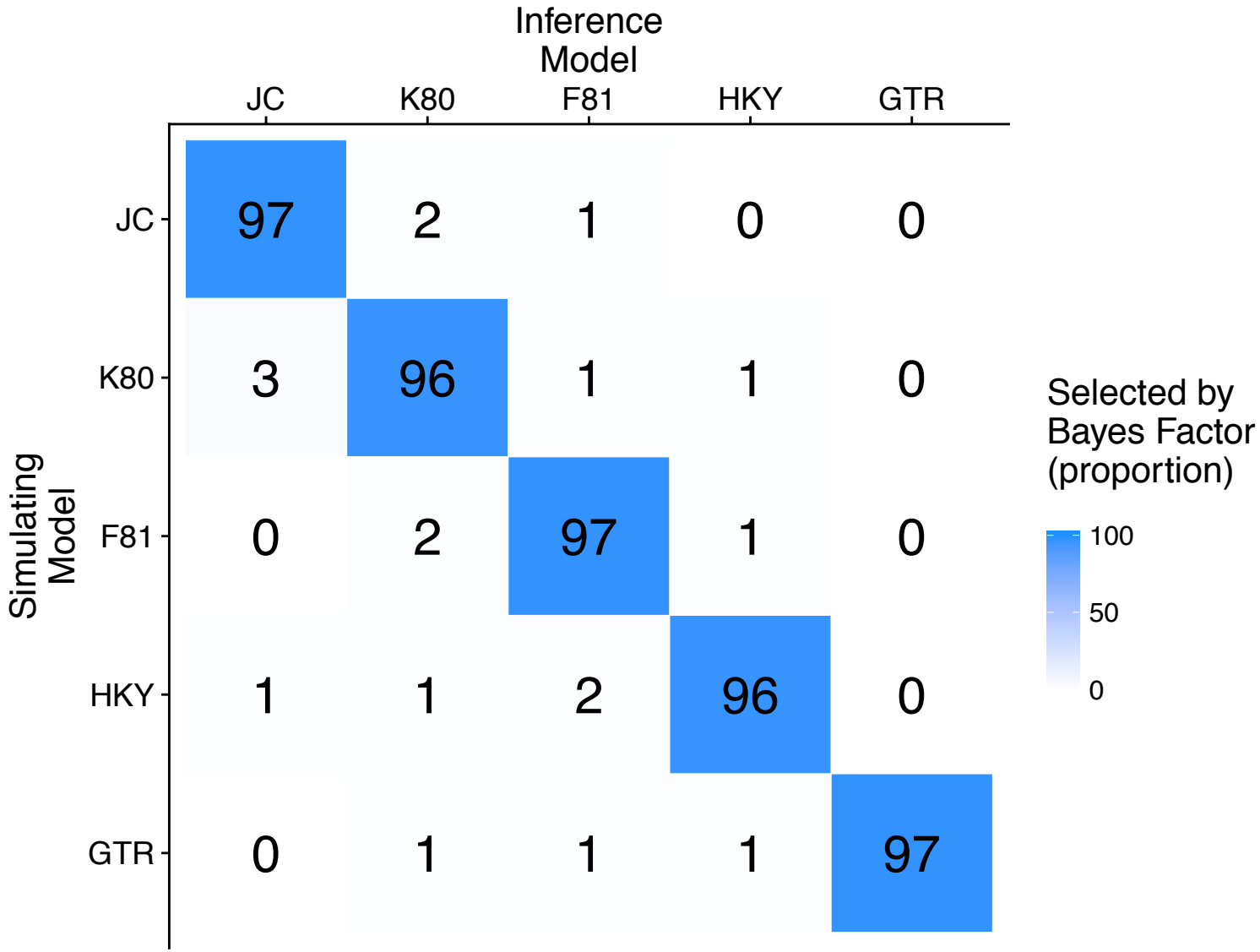
p

0 bad

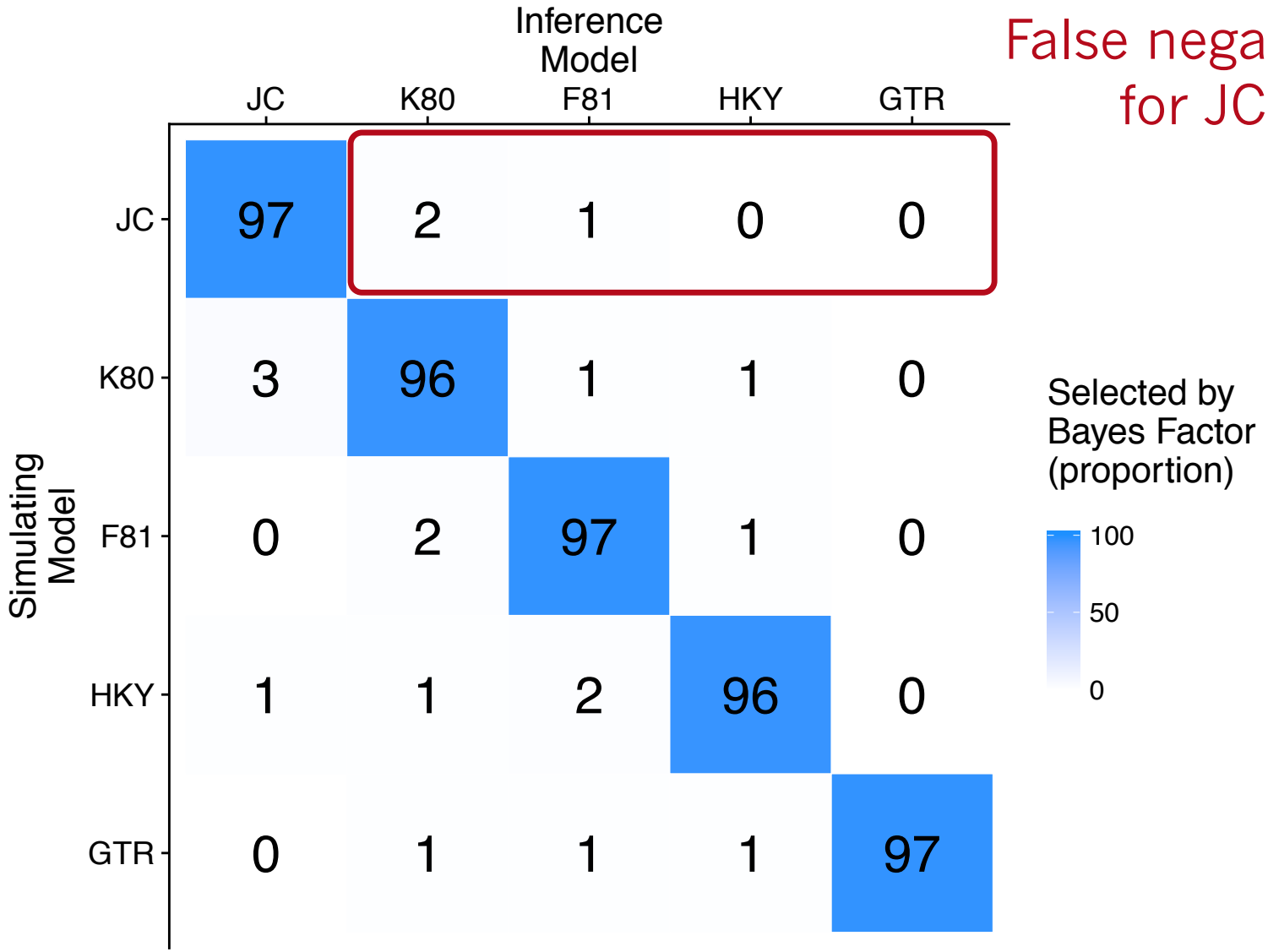
...but 4 kbp generally lacks sufficient *precision* for rate = 1


Inference models
Fits simulated data
Generates simulated data
Simulating models
Inference using the wrong model
Definable models
All possible processes
Natural processes

Model selection



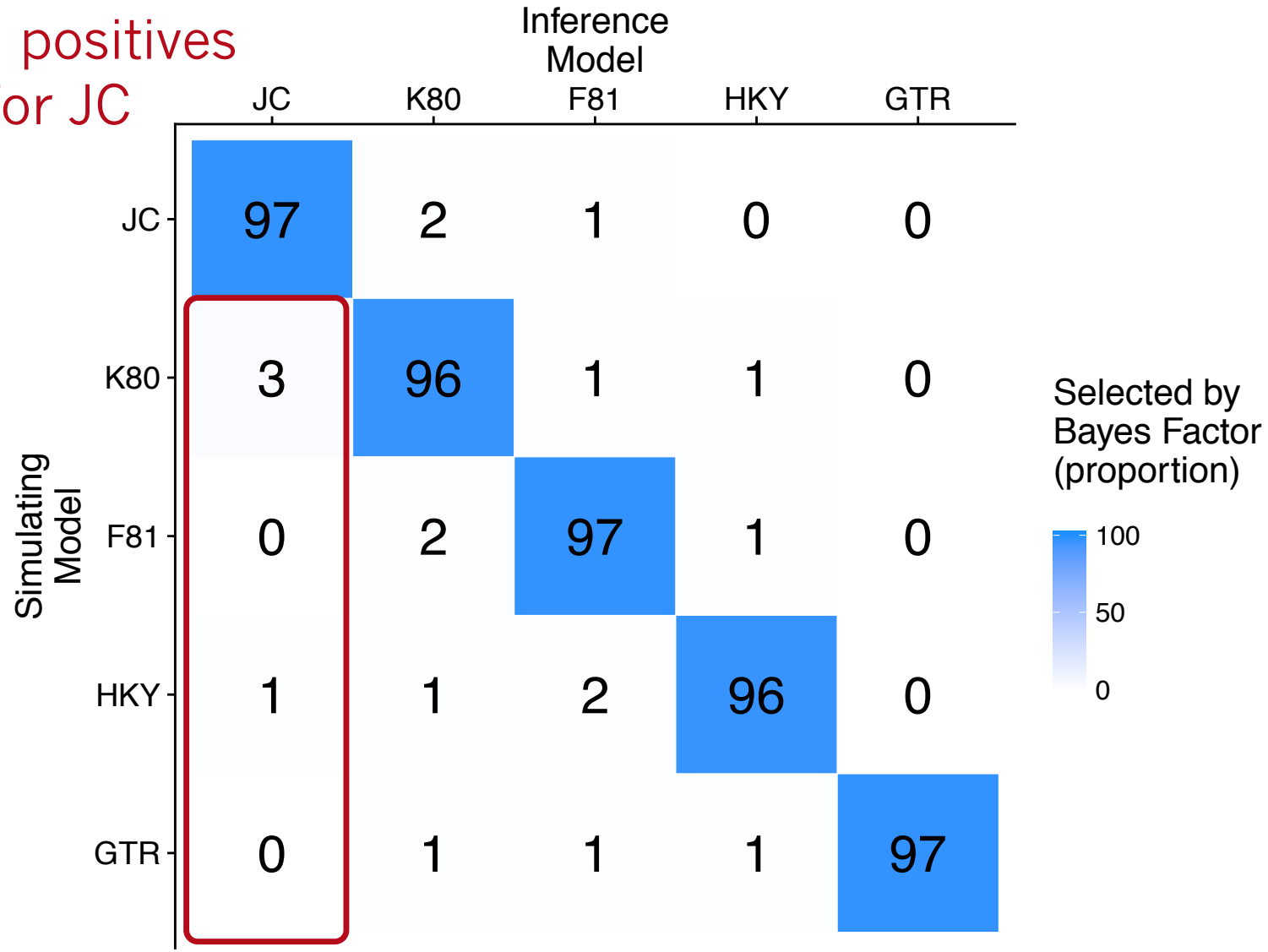
Model selection



False negatives for JC

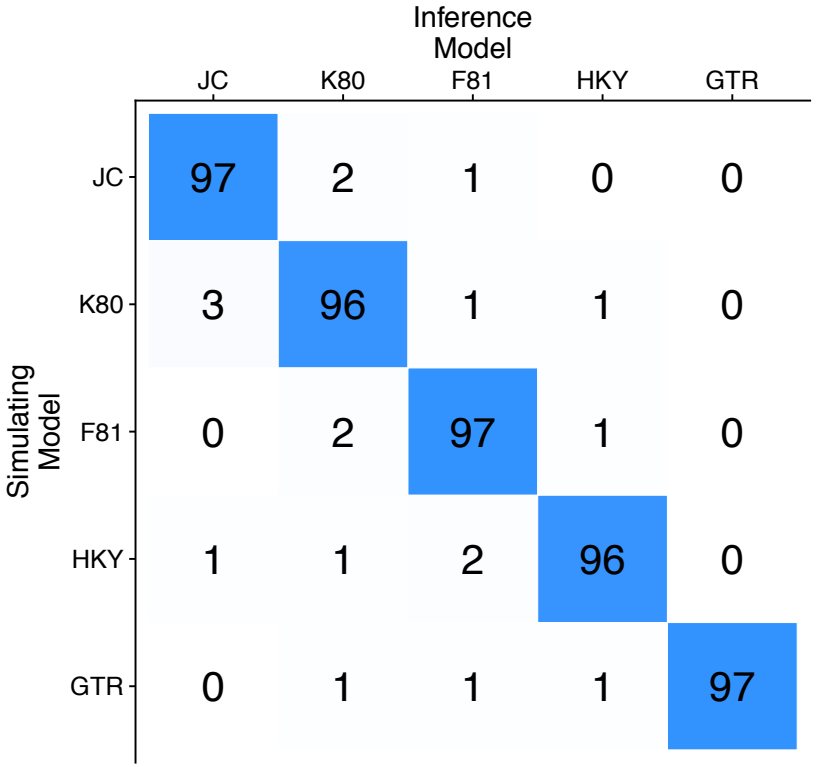
Model selection

False positives for JC

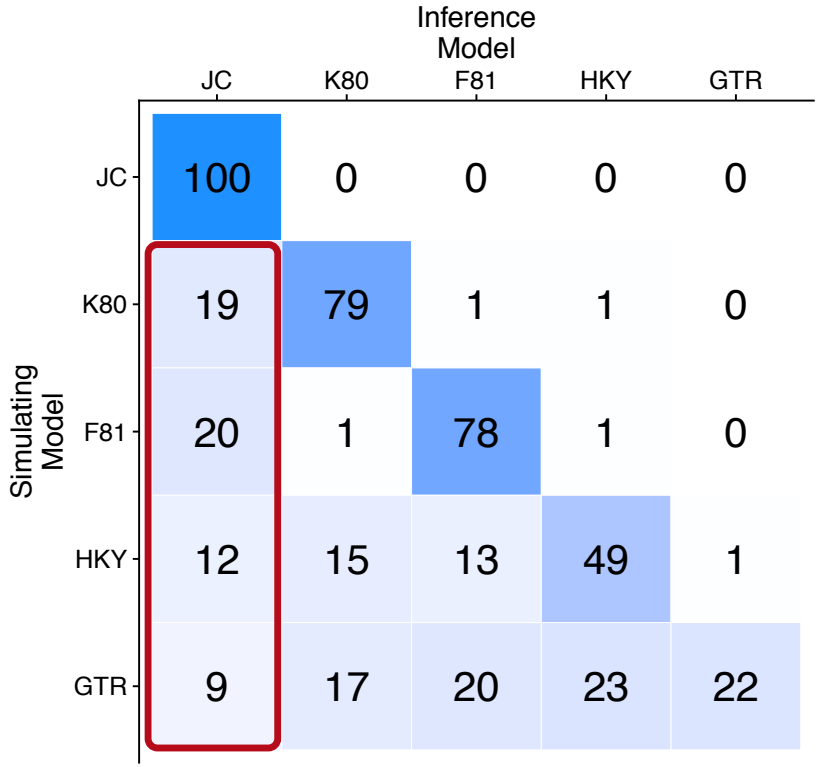


Model selection

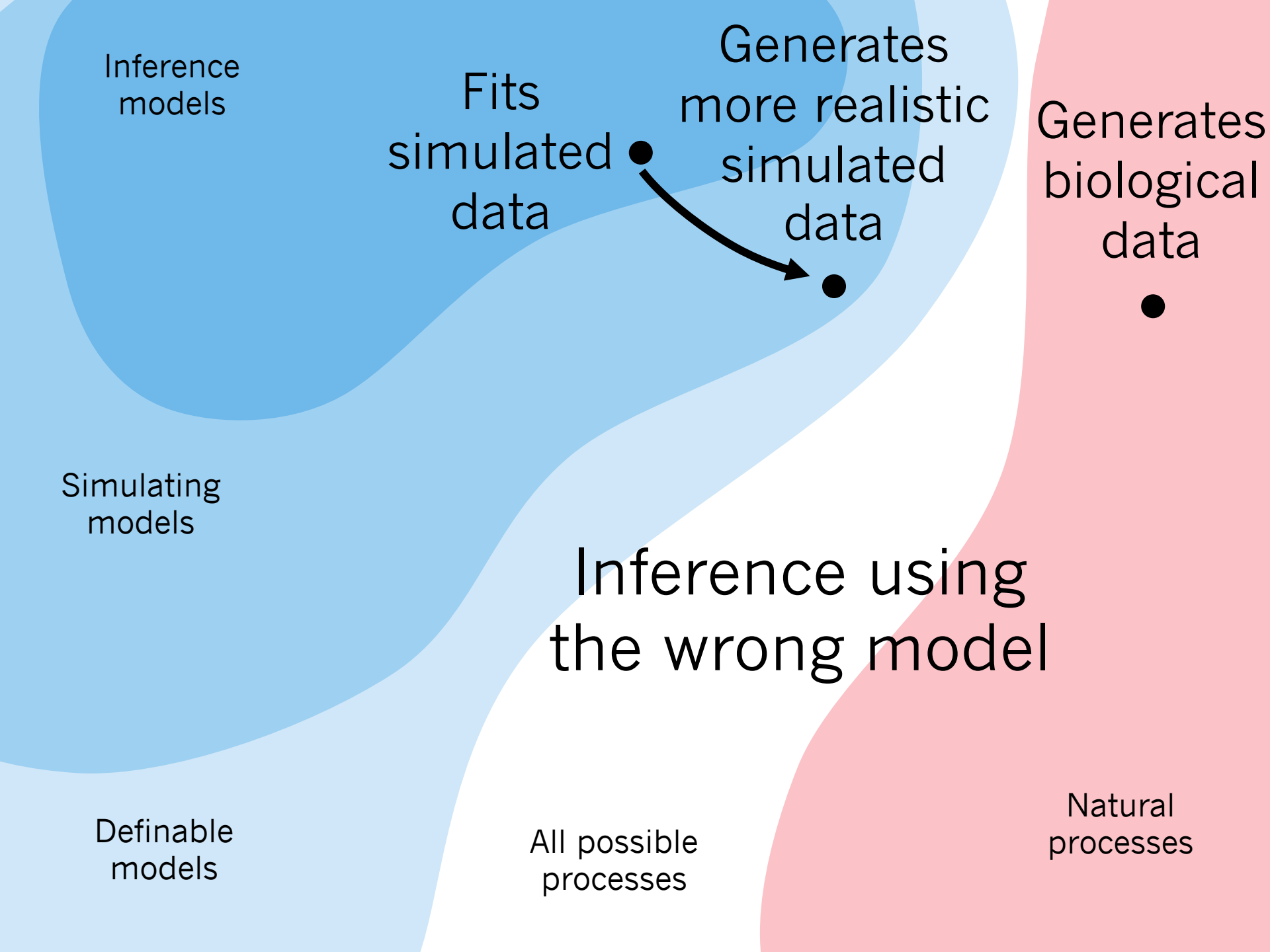
10000 sites,
distinct parameters



100 sites,
JC-like parameters



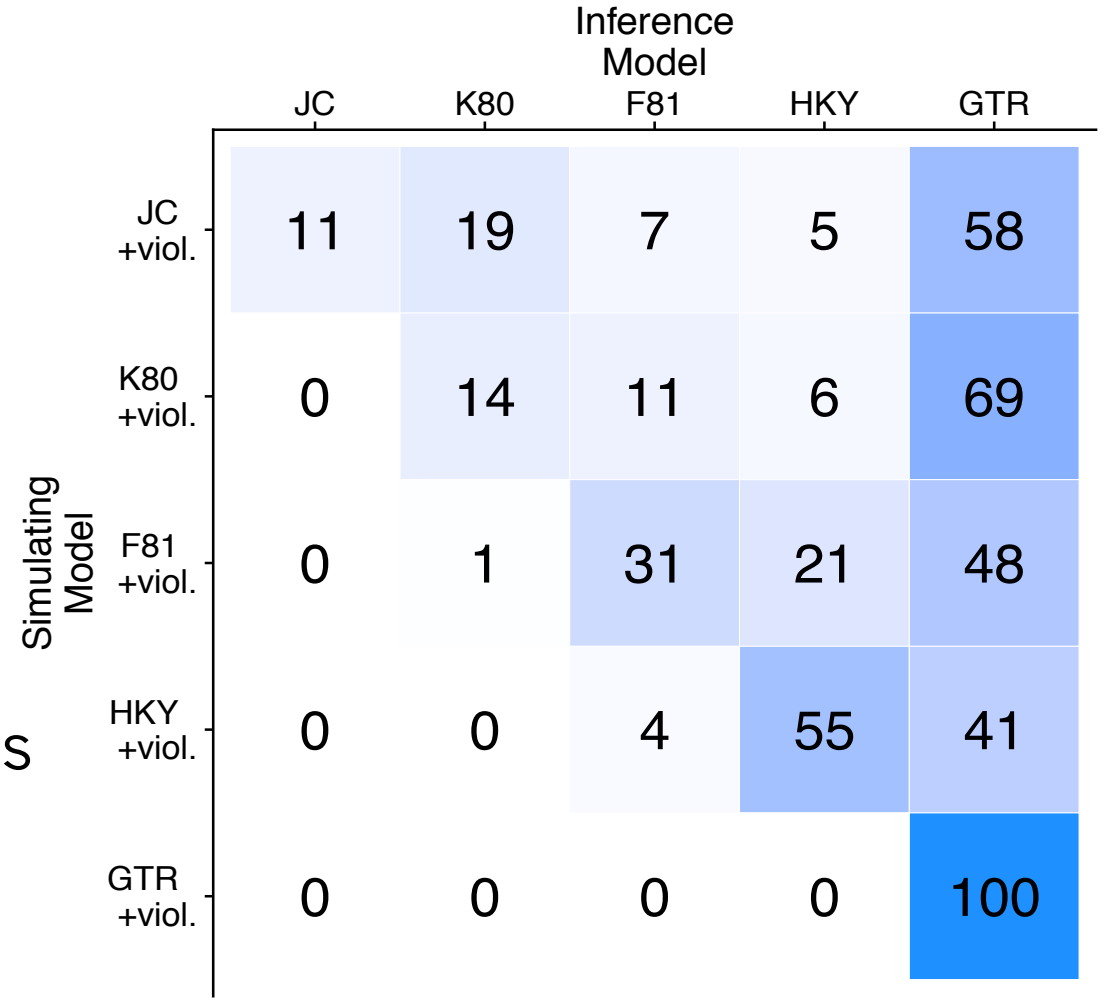
need more data
to detect faint signal!


Inference models
Fits simulated data
Generates more realistic simulated data
Generates biological data
Simulating models
Inference using the wrong model
Definable models
All possible processes
Natural processes

Model violations

Violation examples:

- site evolution depends on protein structure and function
- paralogy errors from bioinformatics pipeline

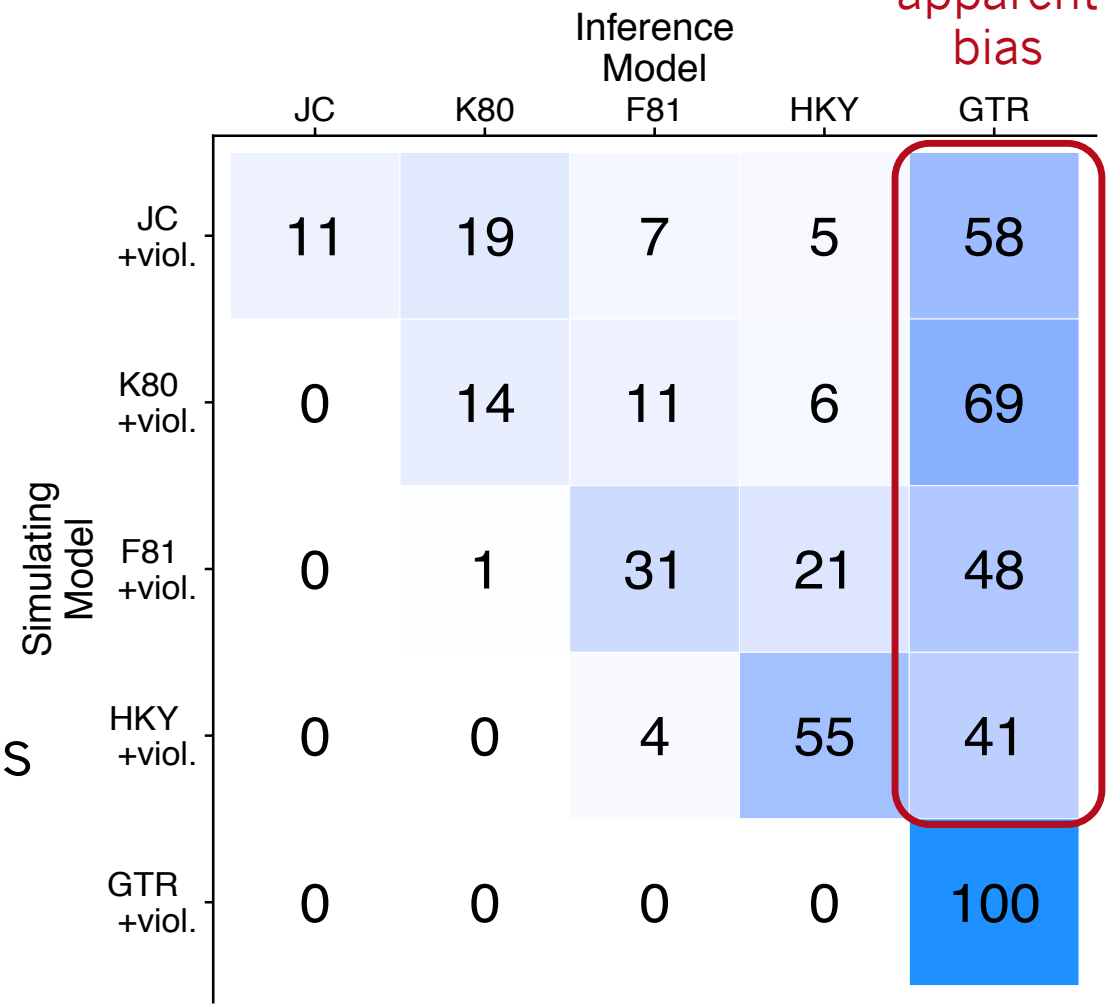


Model violations

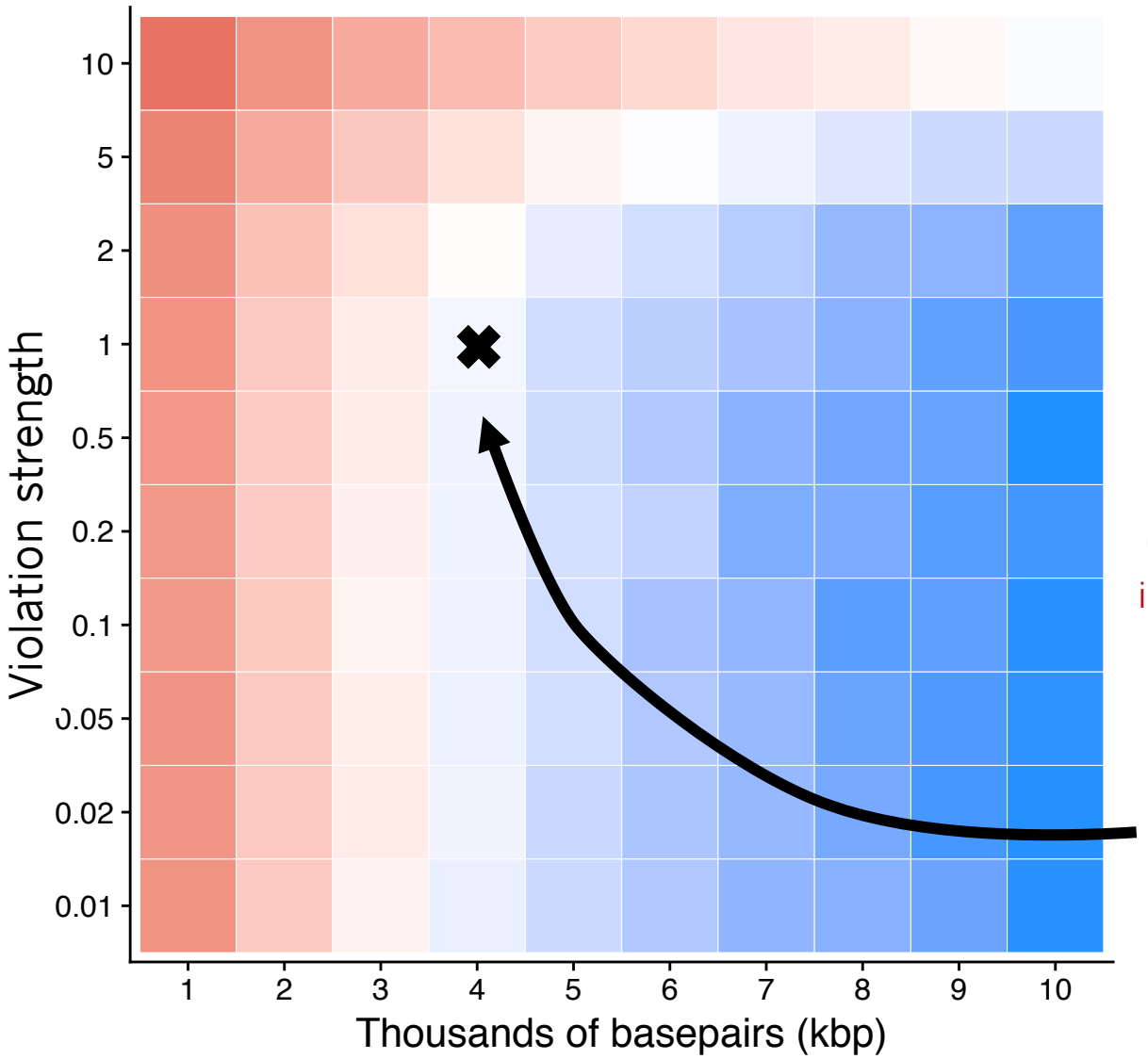
Violation examples:

site evolution depends on protein structure and function

paralogy errors from bioinformatics pipeline



Selection and power



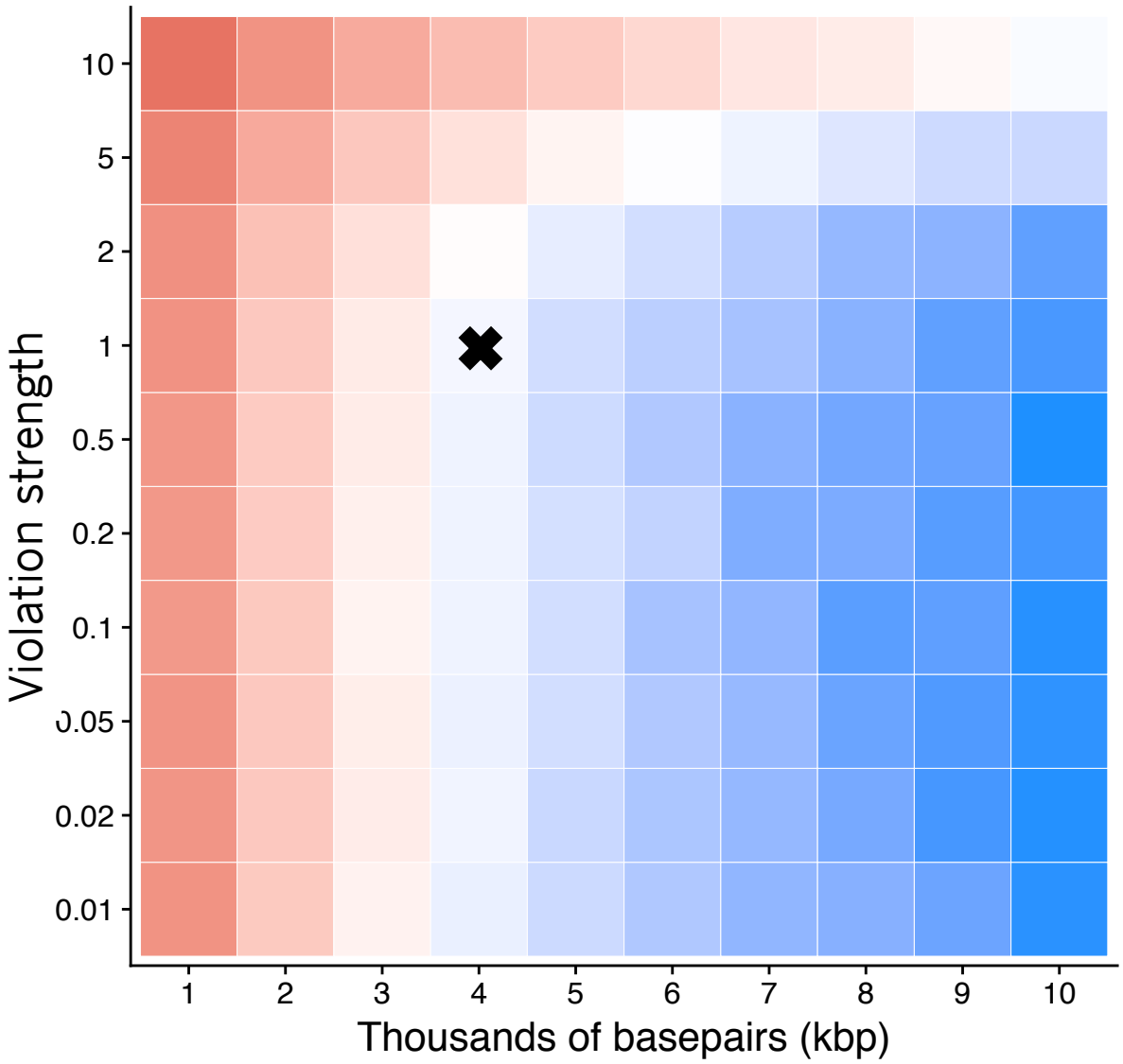
Under what conditions do we reliably select the true model?

correct model

incorrect model

False positive rate too large when violation strength >1

Selection and power

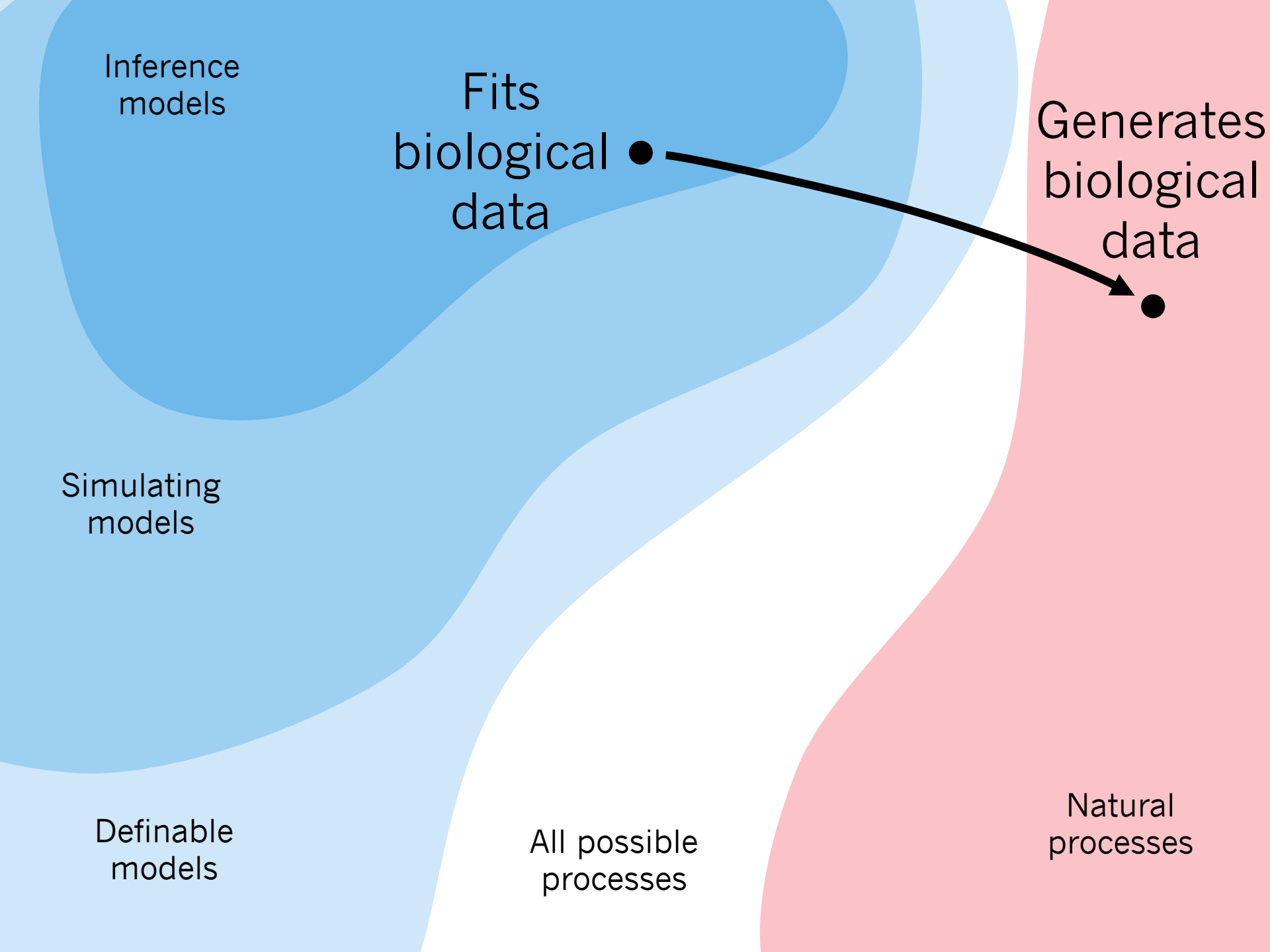


Under what conditions do we reliably select the true model?

correct model

incorrect model

		5	sim. model
	x	5	inf. model
	x	10	cond. 1
	x	10	cond. 2
	x	100	replicates
		250,000	analyses


Inference models
Fits biological data
Generates biological data
Simulating models
Definable models
All possible processes
Natural processes

Model adequacy with posterior prediction

Basic idea

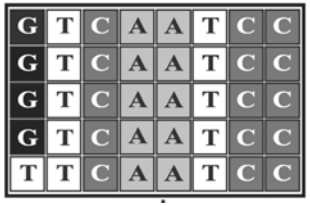
1. Ask model to simulate new datasets using the parameters that best explain your biological data

2. Do the simulated data resemble your biological data?

from Höhna et al (2017, Mol Biol Evol)

Model adequacy with posterior prediction

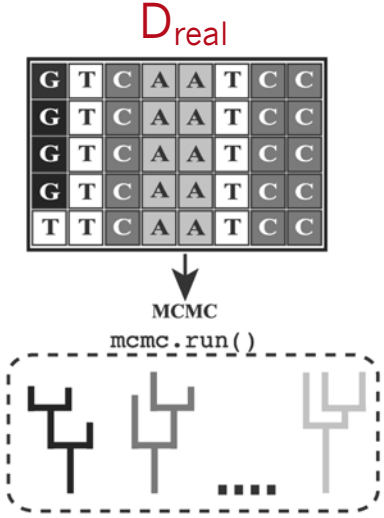
D_{real}



from Höhna et al (2017, Mol Biol Evol)

Model adequacy with posterior prediction

1. Estimate $Pr(\theta_{real} | D_{real})$

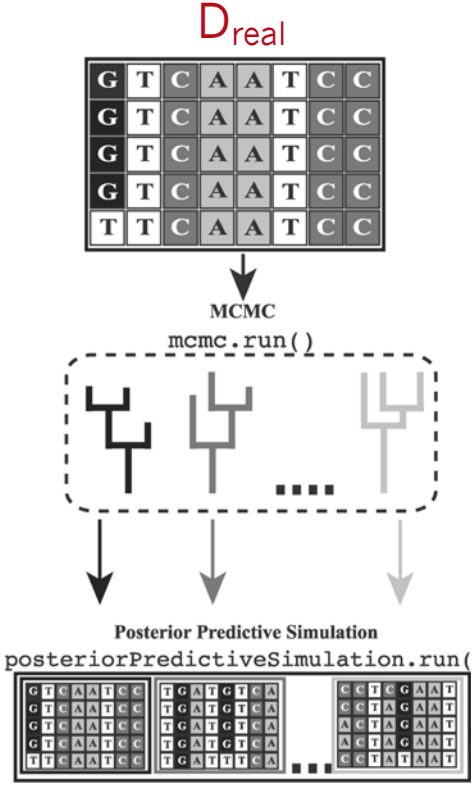


from Höhna et al (2017, Mol Biol Evol)

Model adequacy with posterior prediction

1. Estimate $Pr(\theta_{real} | D_{real})$

2. Simulate K datasets $D_{sim,k} \sim f(\theta_{real} | D_{real})$



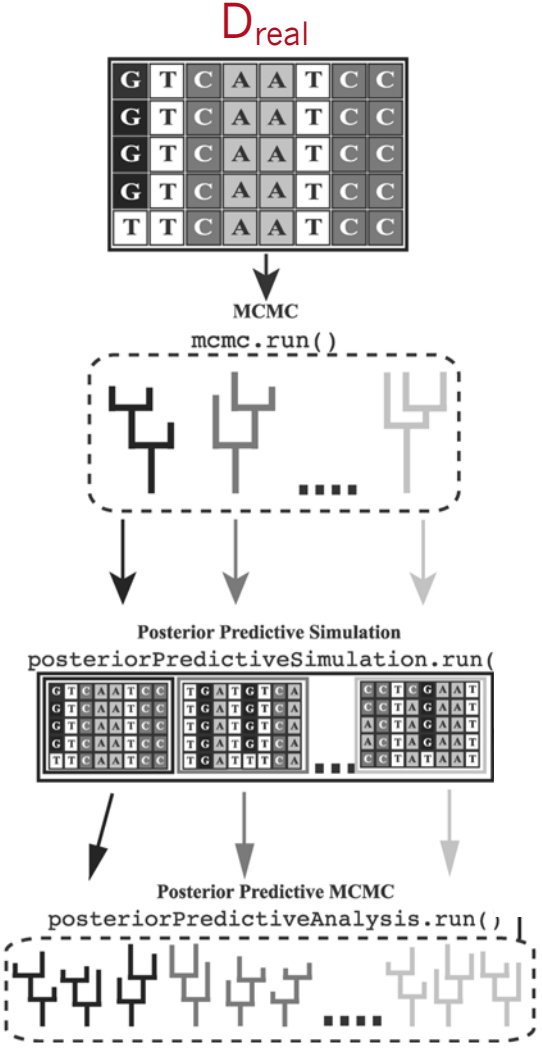
from Höhna et al (2017, Mol Biol Evol)

Model adequacy with posterior prediction

1. Estimate $Pr(\theta_{real} | D_{real})$

2. Simulate K datasets $D_{sim,k} \sim f(\theta_{real} | D_{real})$

3. Estimate K posteriors $Pr(\theta_{sim,k} | D_{sim,k})$



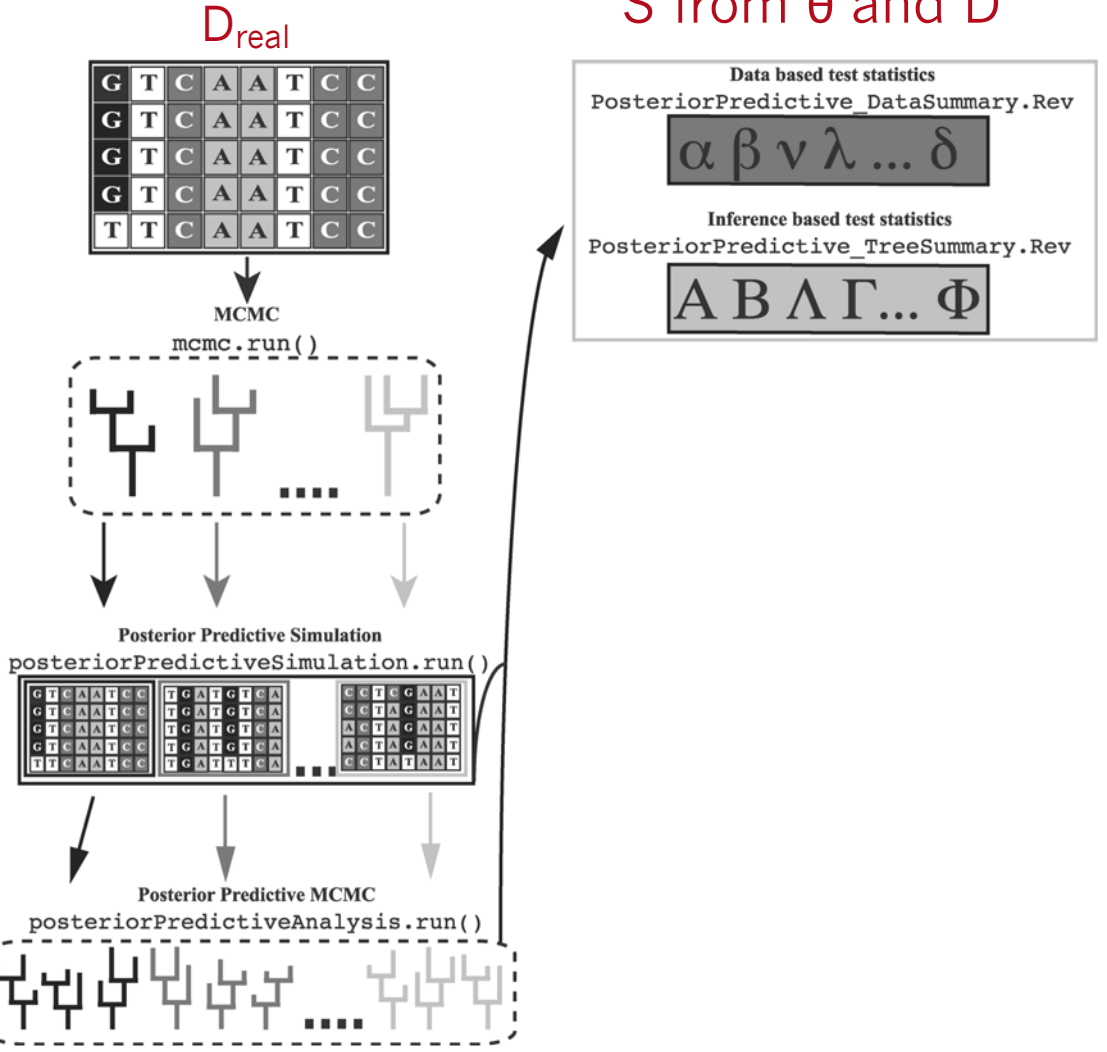
from Höhna et al (2017, Mol Biol Evol)

Model adequacy with posterior prediction

1. Estimate $Pr(\theta_{real} | D_{real})$

2. Simulate K datasets $D_{sim,k} \sim f(\theta_{real} | D_{real})$

3. Estimate K posteriors $Pr(\theta_{sim,k} | D_{sim,k})$



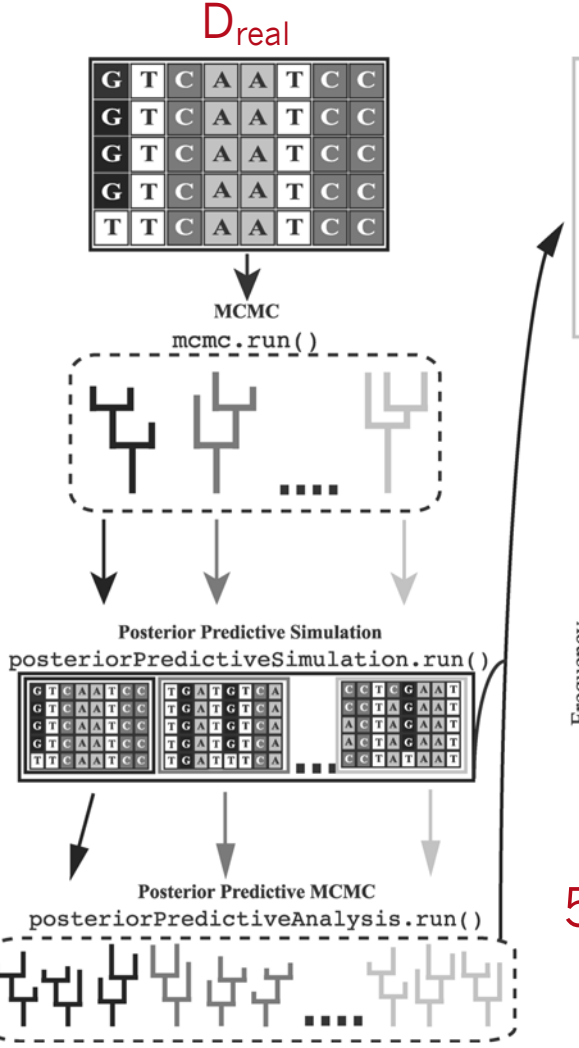
from Höhna et al (2017, Mol Biol Evol)

Model adequacy with posterior prediction

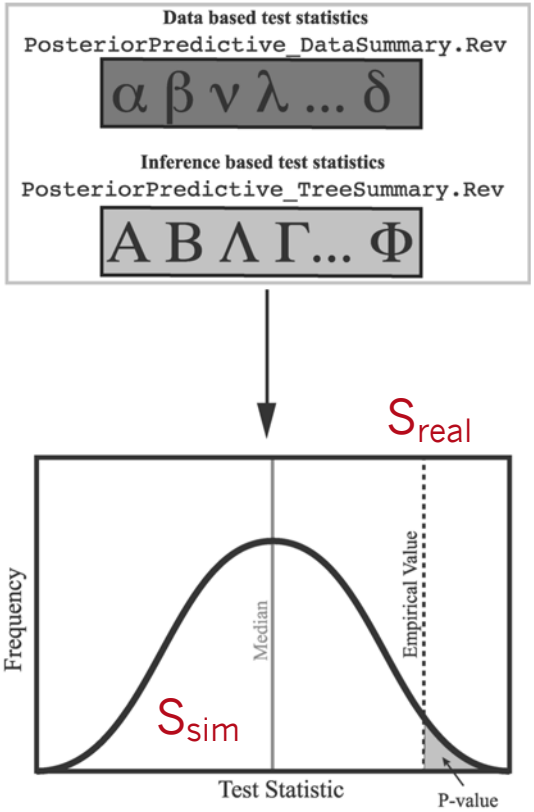
1. Estimate $Pr(\theta_{real} | D_{real})$

2. Simulate K datasets $D_{sim,k} \sim f(\theta_{real} | D_{real})$

3. Estimate K posteriors $Pr(\theta_{sim,k} | D_{sim,k})$



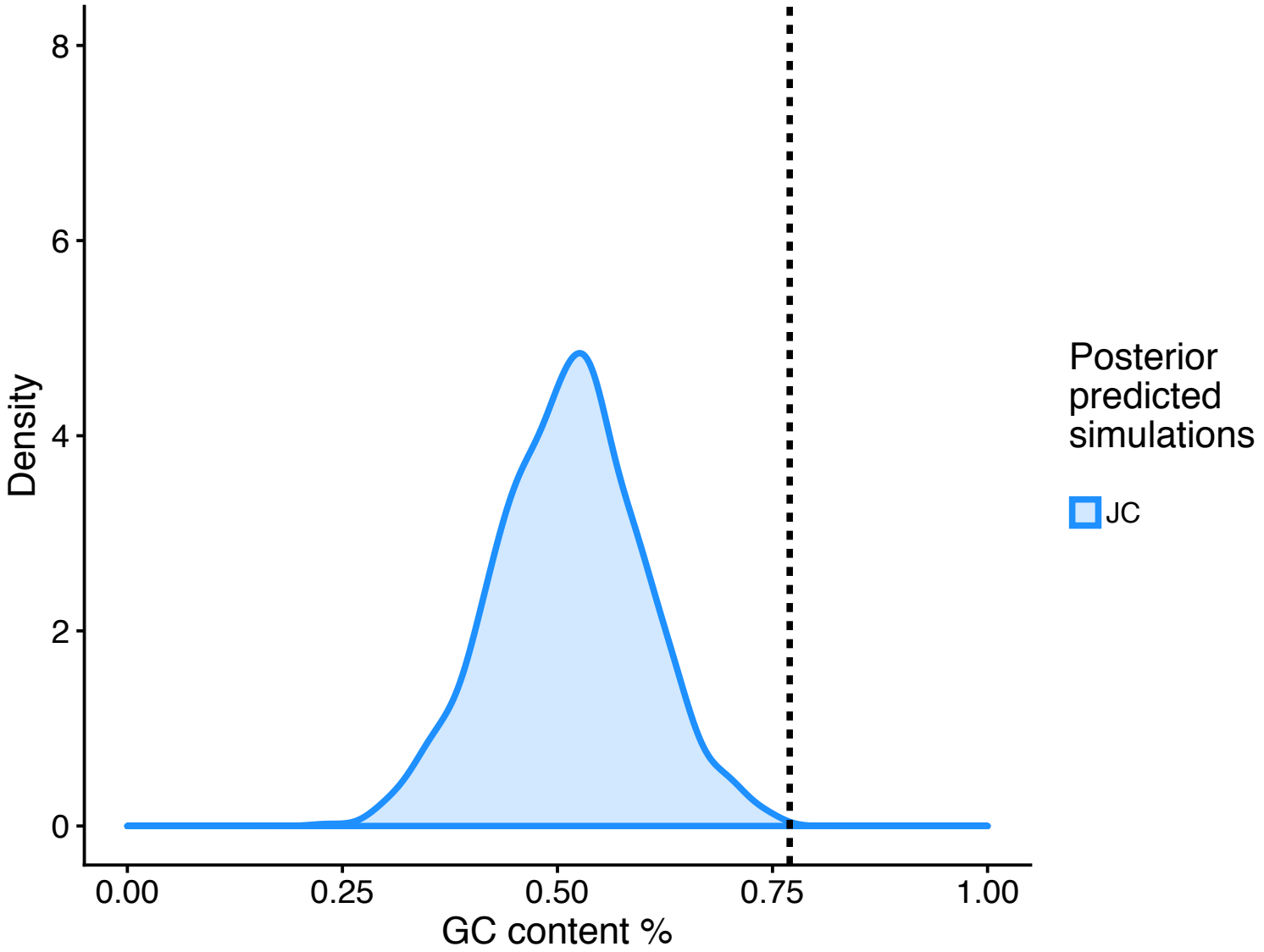
4. Collect test statistics S from θ and D



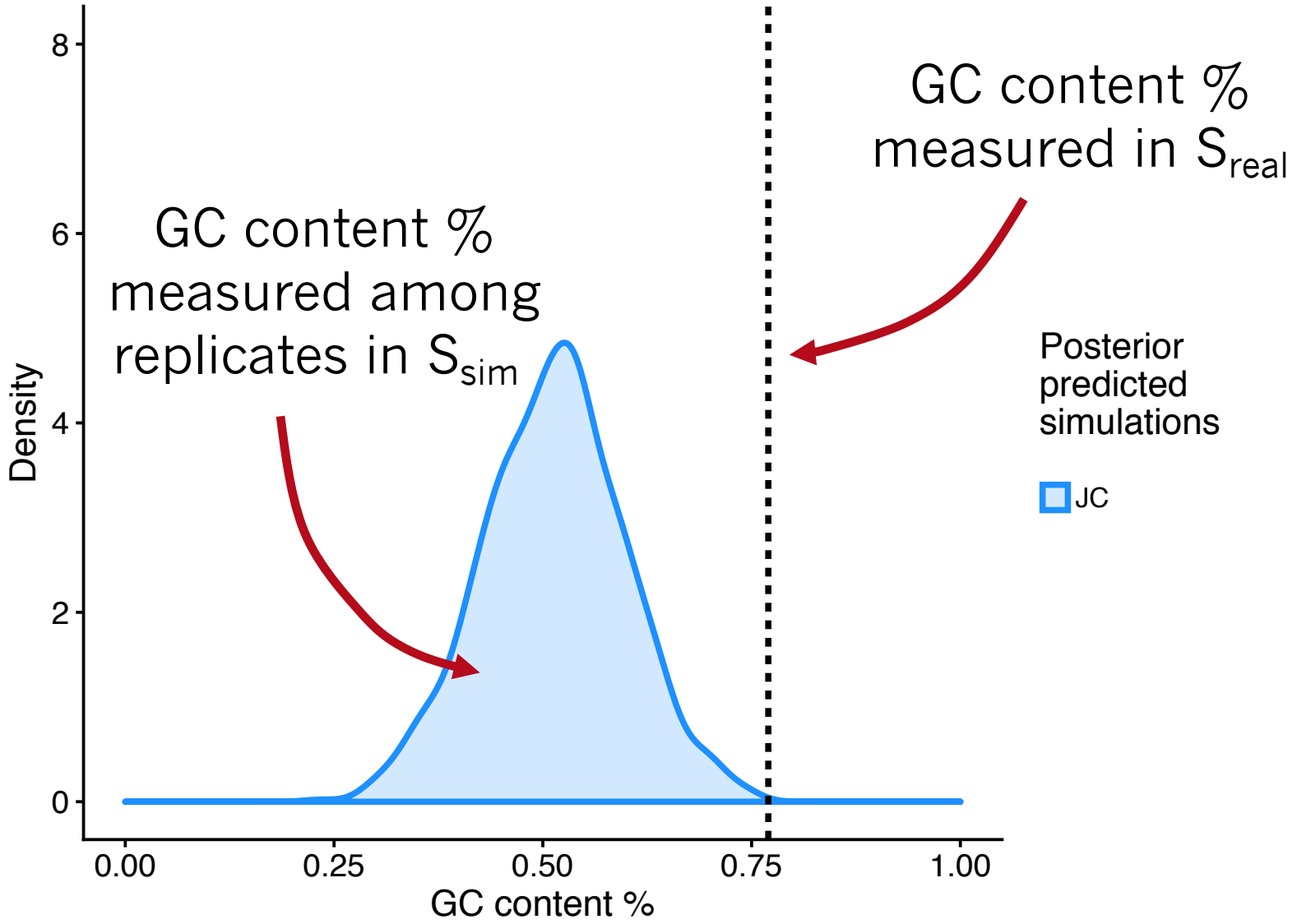
5. Does distribution of S_{sim} contain S_{real}?

from Höhna et al (2017, Mol Biol Evol)

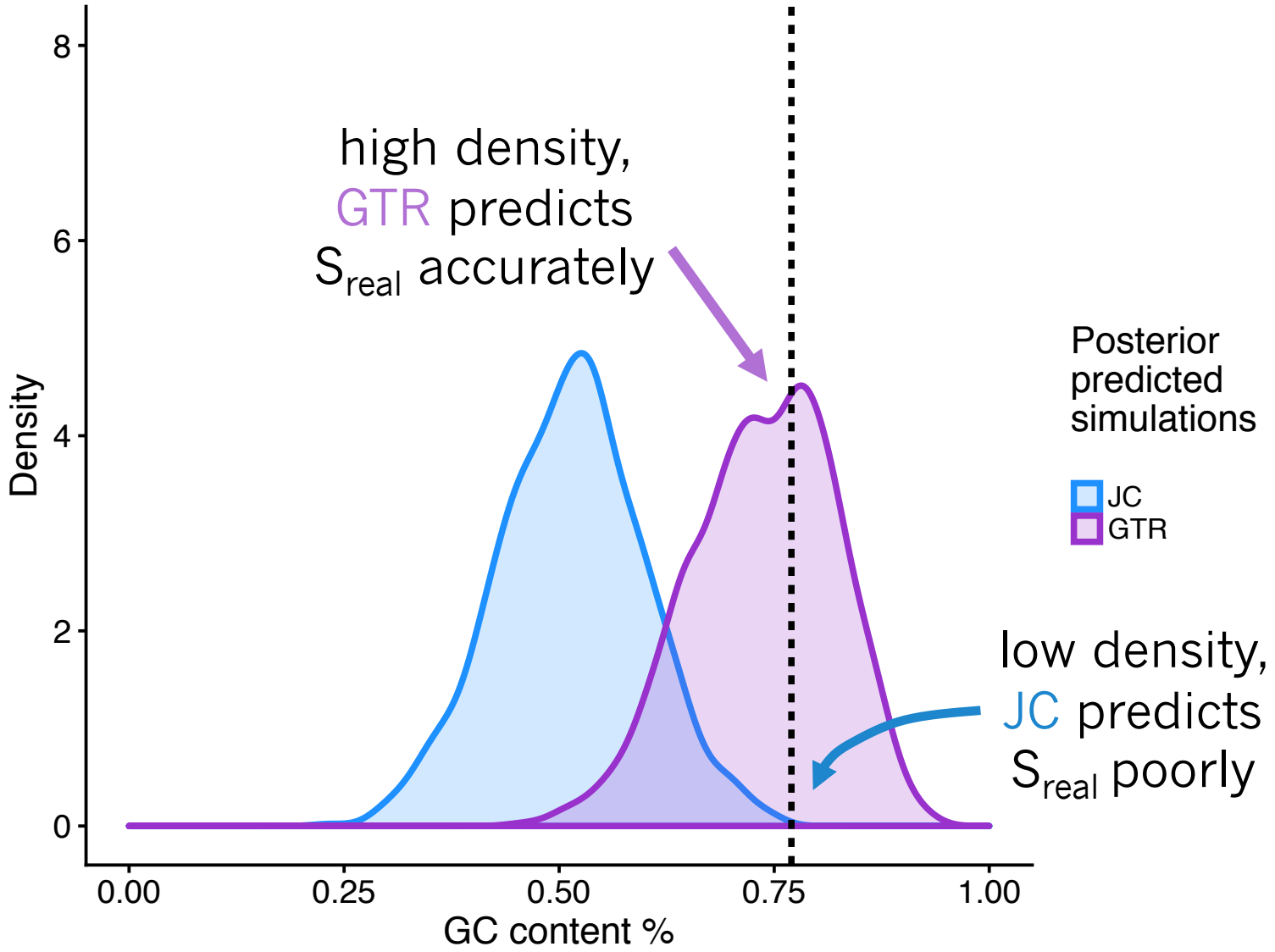
Model adequacy



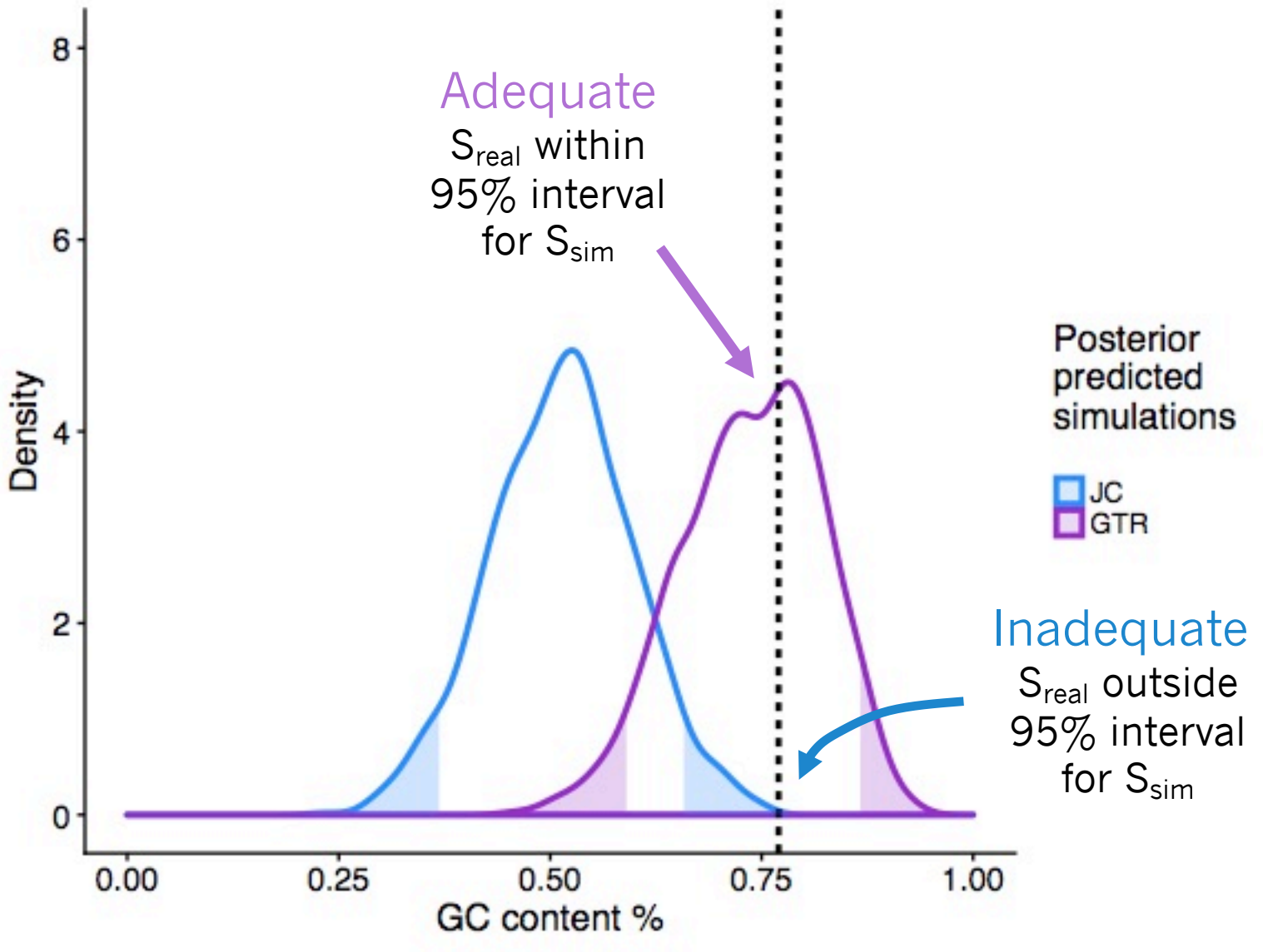
Model adequacy



Model adequacy



Model adequacy



Model smell

Do results make sense?

- Time runs backwards?
- Precambrian rabbits?

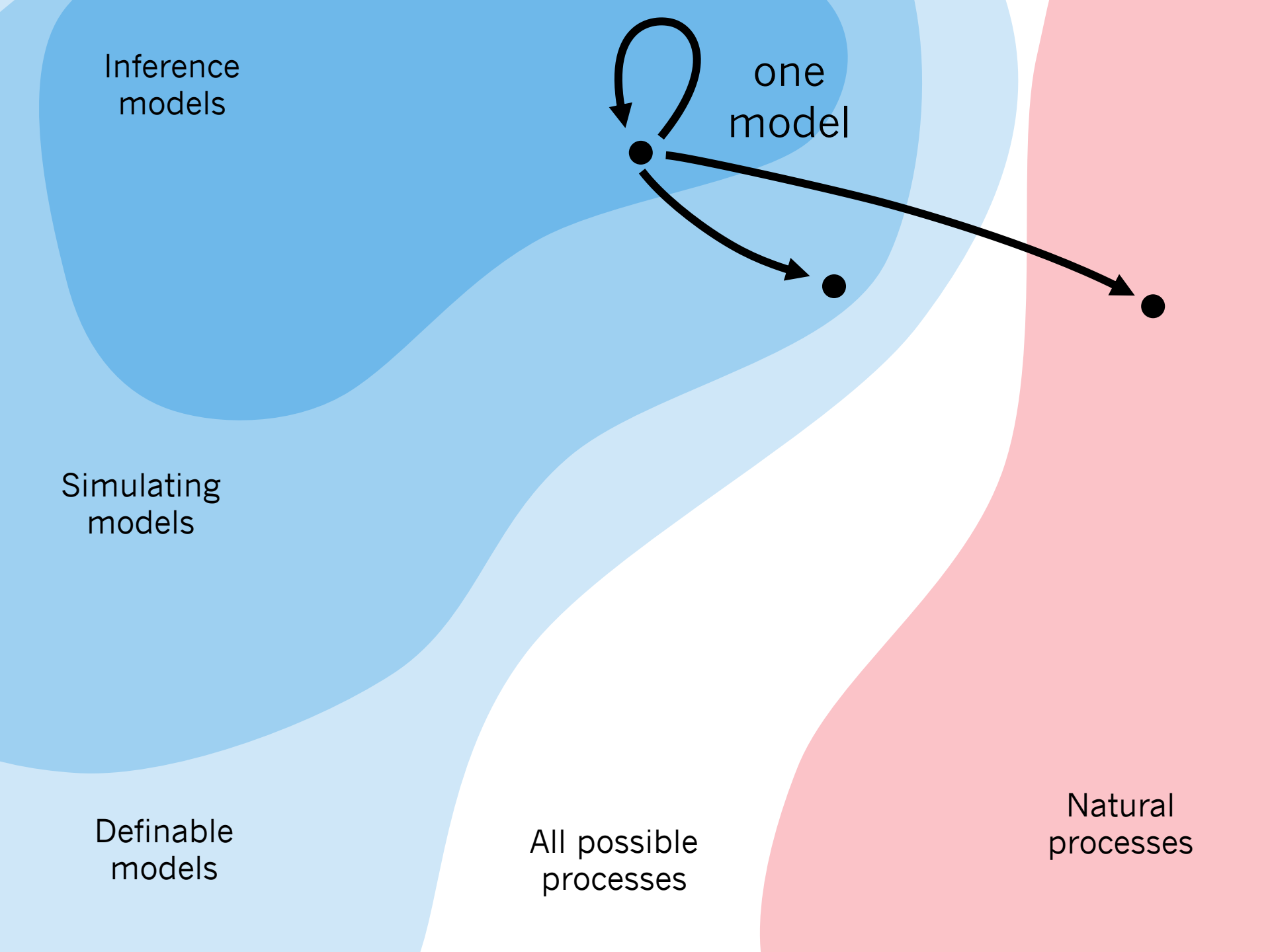
Model smell

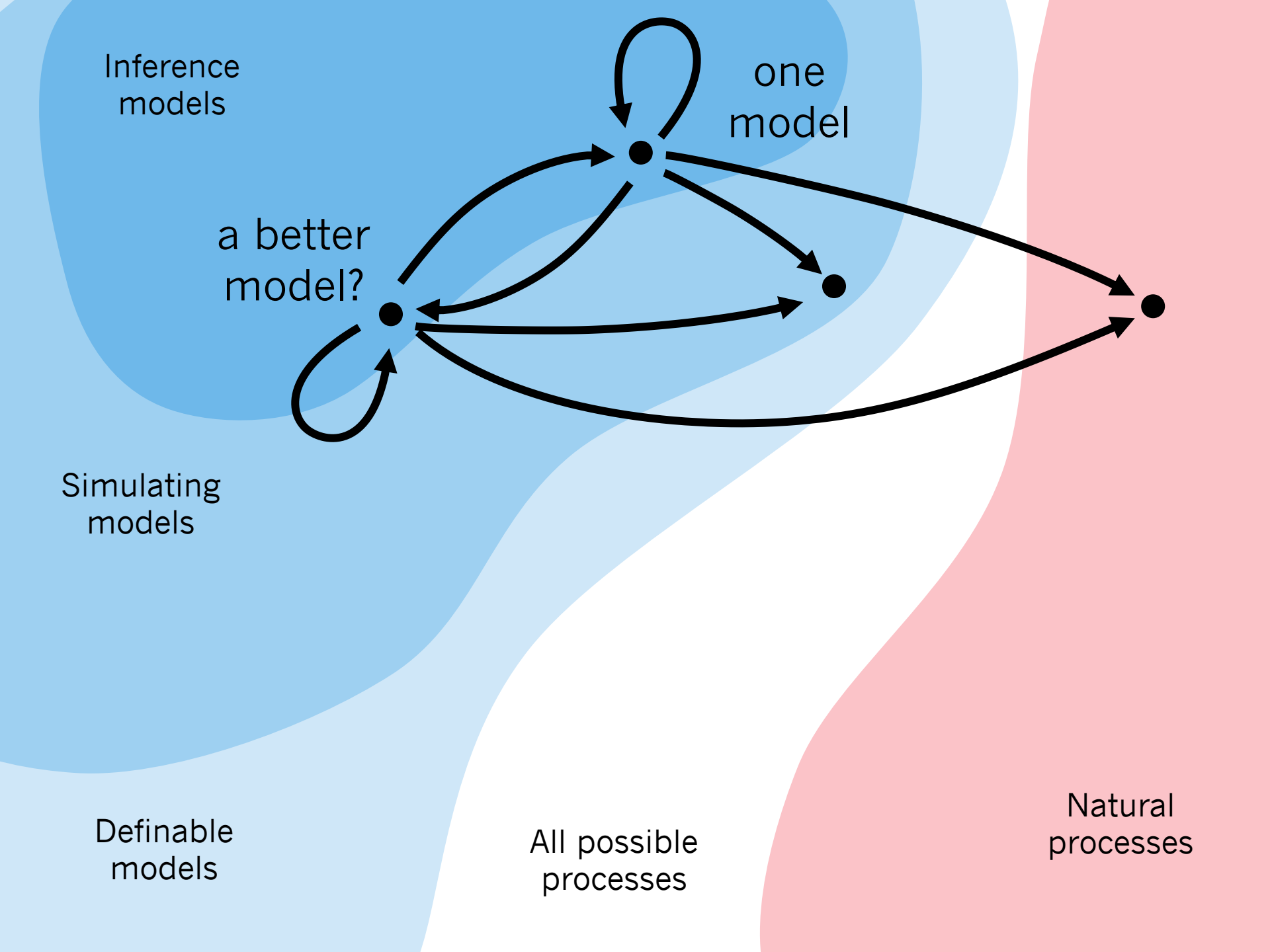
Do results make sense?

- Time runs backwards?
- Precambrian rabbits?

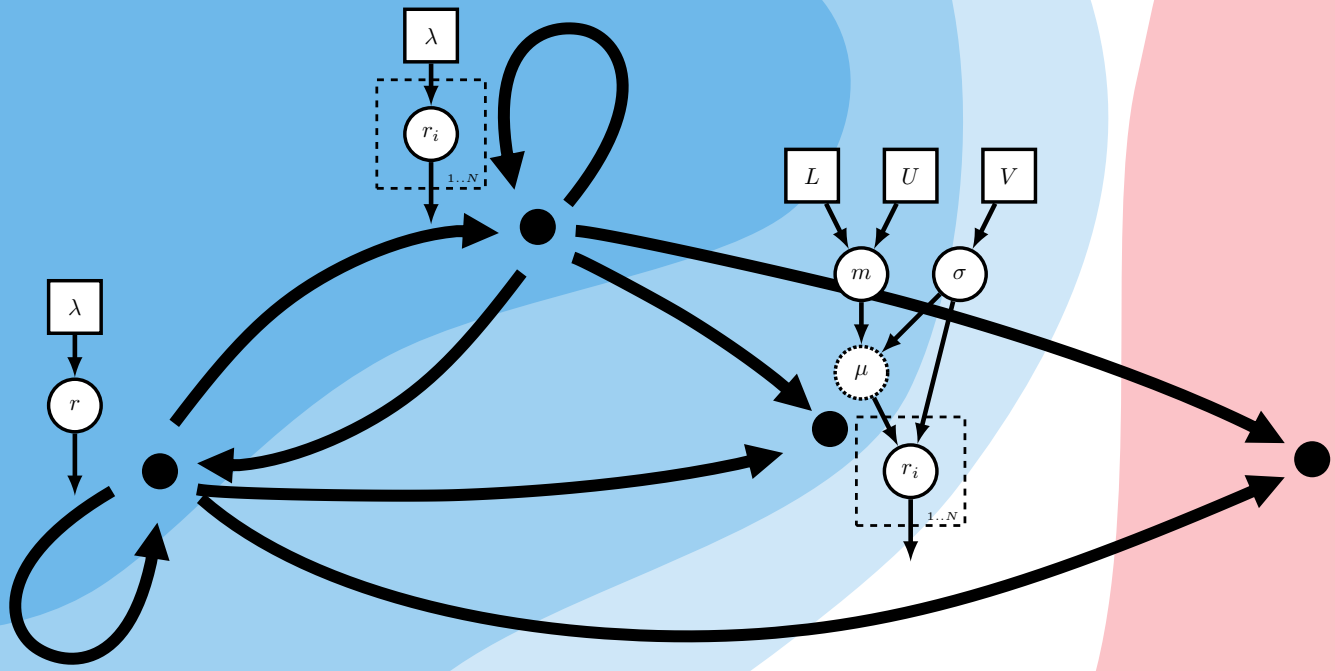
If not, slow down!

- Errors in data?
- Analysis misconfigured?
- Method glitch?


Inference models
one model
Simulating models
Definable models
All possible processes
Natural processes


Inference models
one model
a better model?
Simulating models
Definable models
All possible processes
Natural processes

Inference models



Simulating models

PGMs to navigate model space (more in lab)

Definable models

All possible processes

Natural processes