

# Dated Bayesian phylogenetic analysis to infer parameters of epidemiological interest

**Sam Hong**  
**Rega Institute of Medical Research**  
**KU Leuven - Belgium**  
**[samuel.hong@kuleuven.be](mailto:samuel.hong@kuleuven.be)**

# Some applied questions addressed by viral genomic epidemiology



*Tuesday* What virus is causing the outbreak?

*Wednesday* What species are involved in virus transmission?

*Thursday* When did the outbreak begin?

*Thursday* How many introductions have there been?

*Friday* Where did the outbreak begin?

*Friday* How fast is the virus evolving?

*Friday* How rapidly is the virus transmitting?

*Friday* What factors drive an outbreak?

Workshop Materials  
and Detailed Agenda:



Ask questions, participate, network and join the Workshop's WhatsApp groups ☺

# Some applied questions addressed by viral genomic epidemiology



*Tuesday* What virus is causing the outbreak?

*Wednesday* What species are involved in virus transmission?

*Thursday* When did the outbreak begin?

*Thursday* How many introductions have there been?

*Friday* Where did the outbreak begin?

*Friday* How fast is the virus evolving?

*Friday* How rapidly is the virus transmitting?

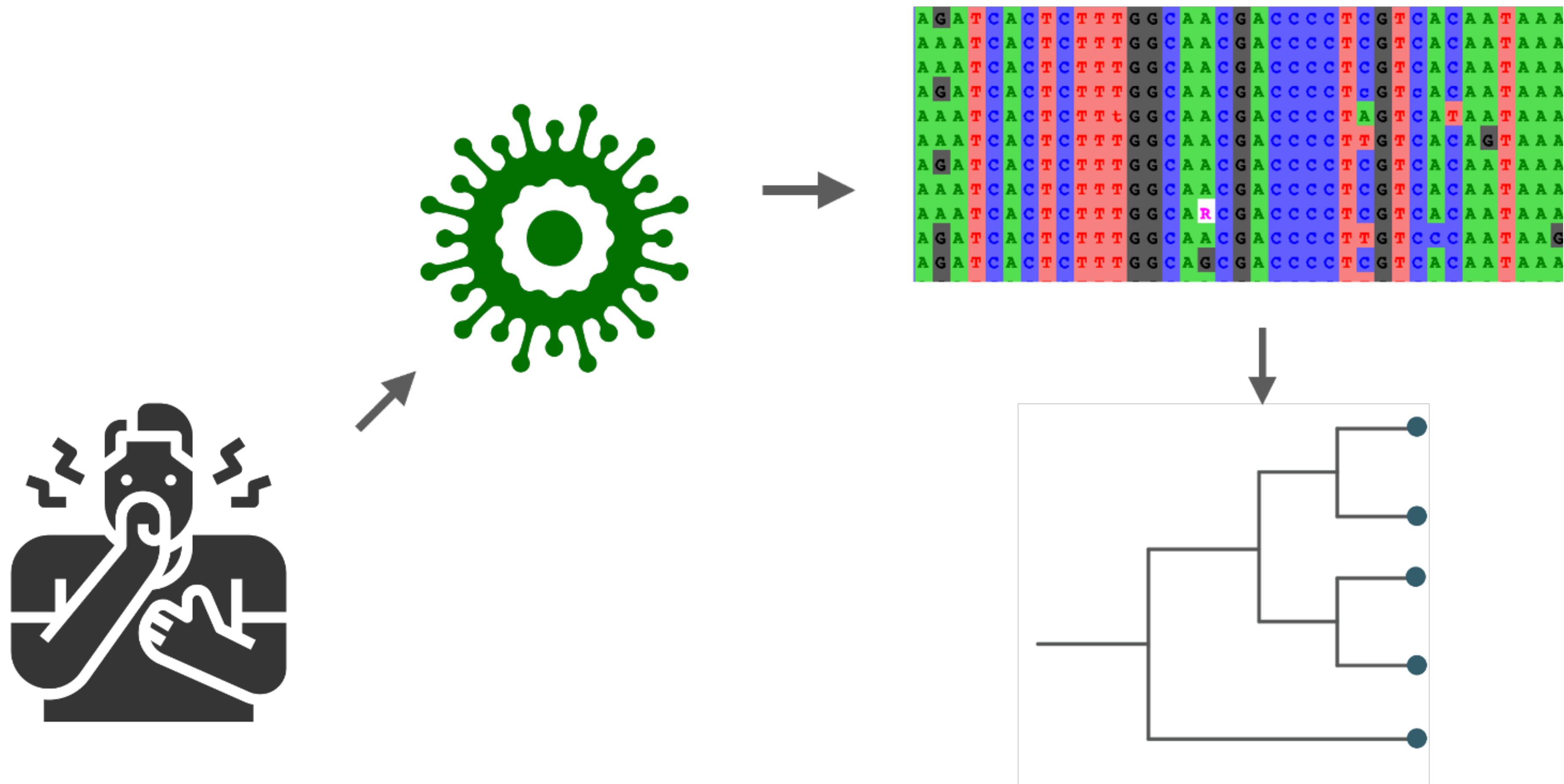
*Friday* What factors drive an outbreak?

Workshop Materials  
and Detailed Agenda:



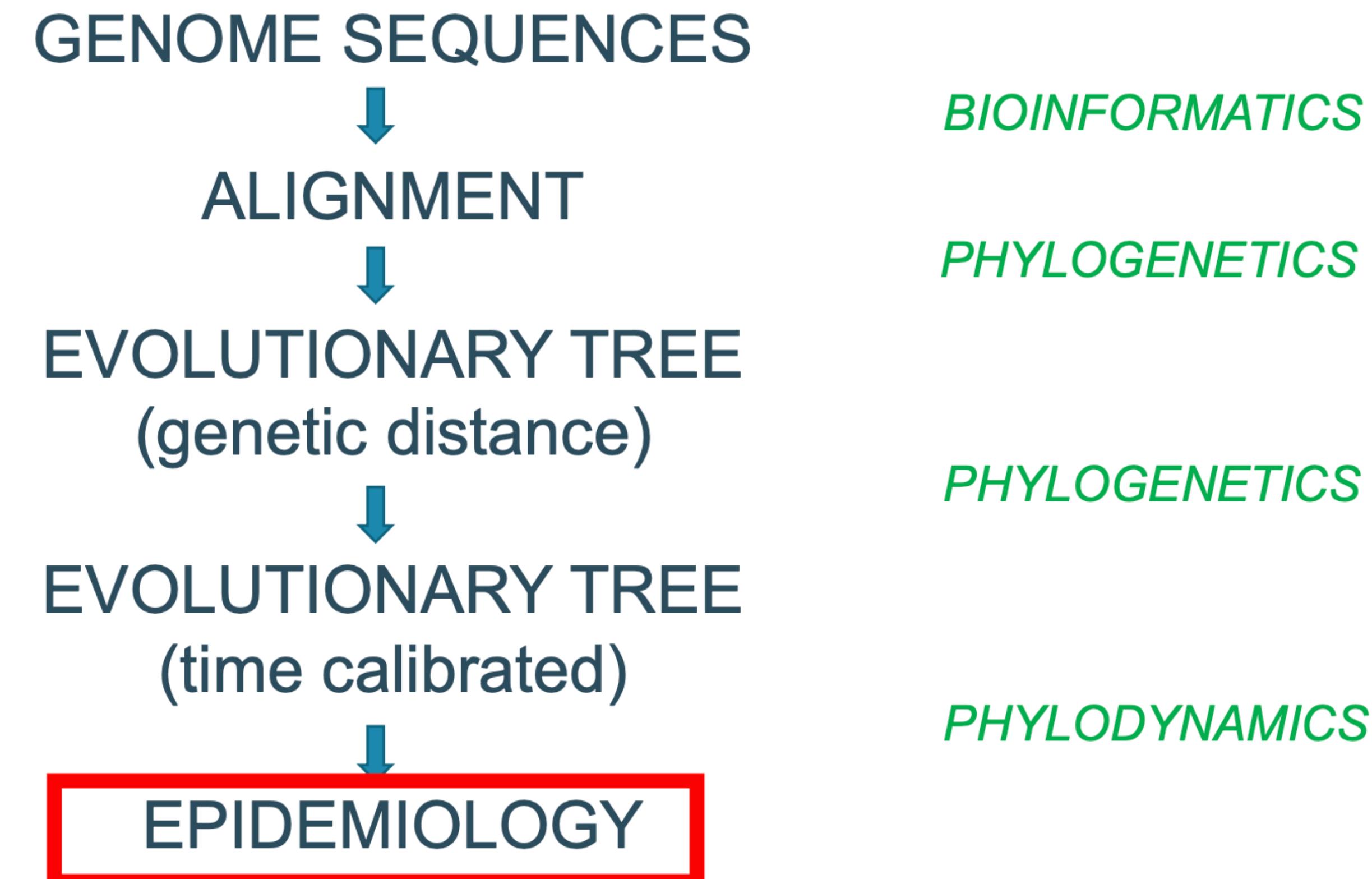
Ask questions, participate, network and join the Workshop's WhatsApp groups ☺

# Genomic Epidemiology



# From sequence to epidemiology

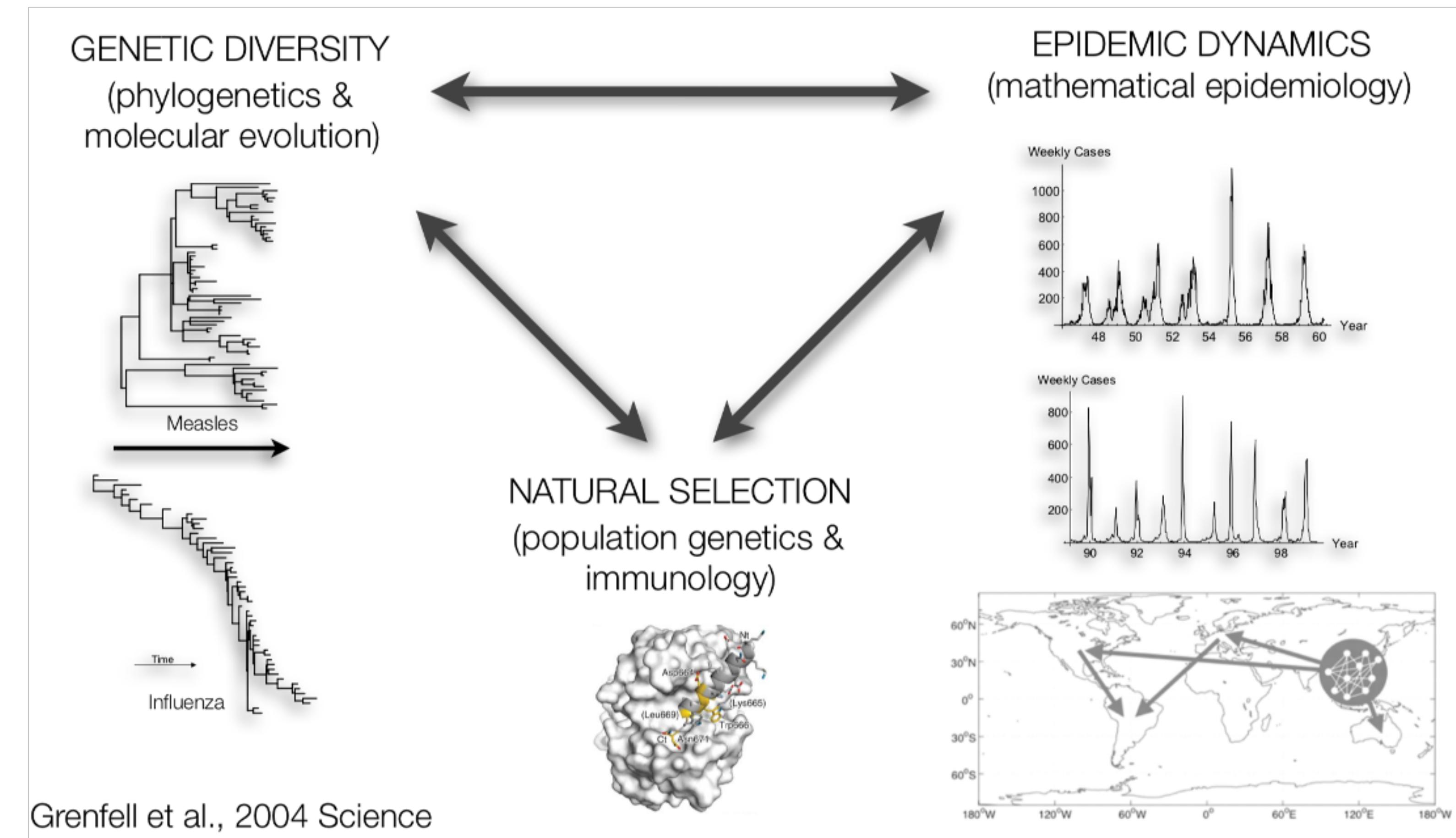
*Alignment Methods*  
*Sequence Evolution Models*  
*Phylogenetic Reconstruction*  
*Molecular Clock Models*  
*Phylodynamic Models*



Adapted from Philippe Lemey

# Phyldynamics

*In many infectious pathogens, evolutionary and epidemiological dynamics occur within the same time scale.*

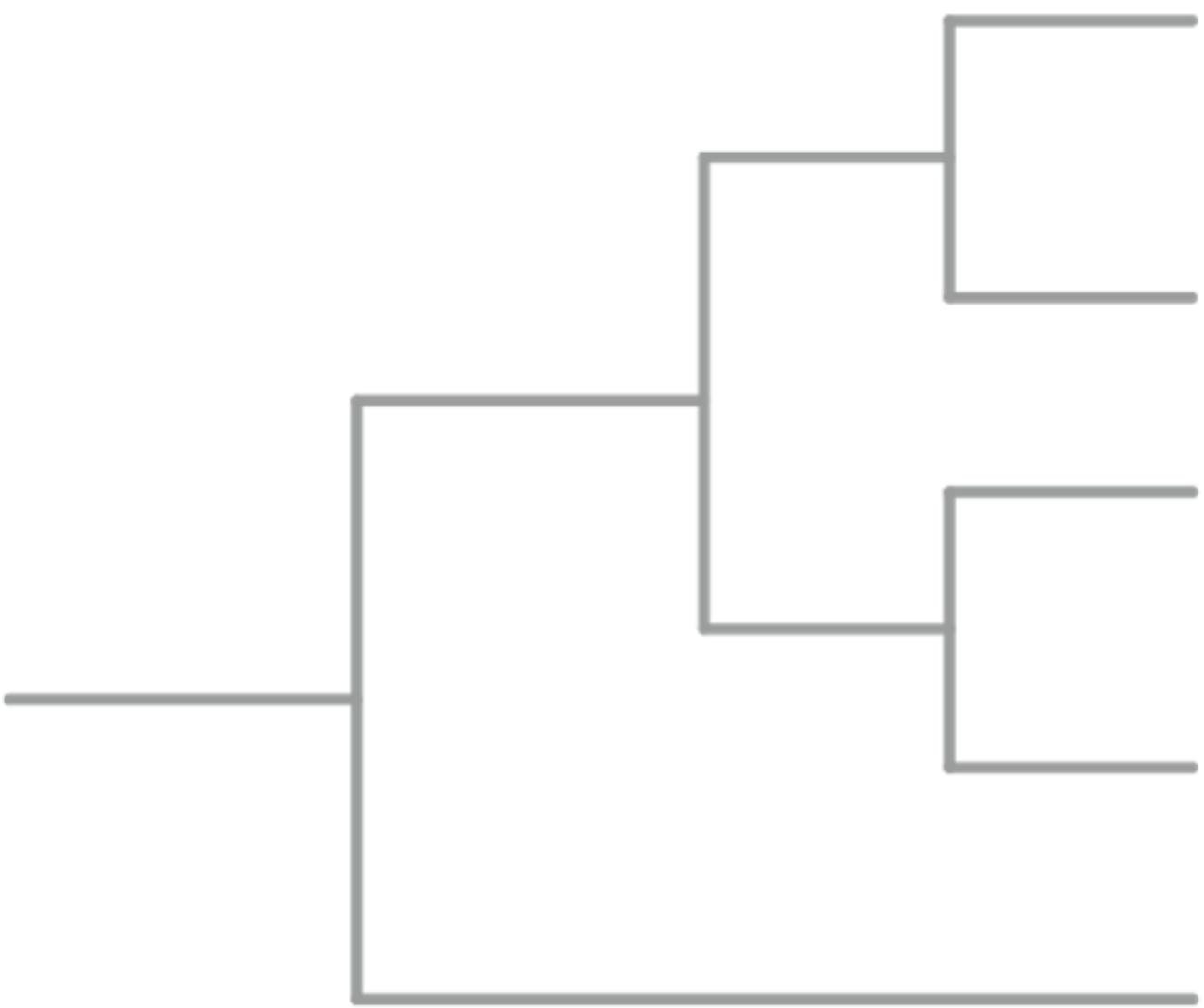


# Bayesian Phylogenetics

Maximum Likelihood

$$\operatorname{argmax}_{\theta} L(\theta)$$

Likelihood:  $L(\theta) \propto P(Data|\theta)$

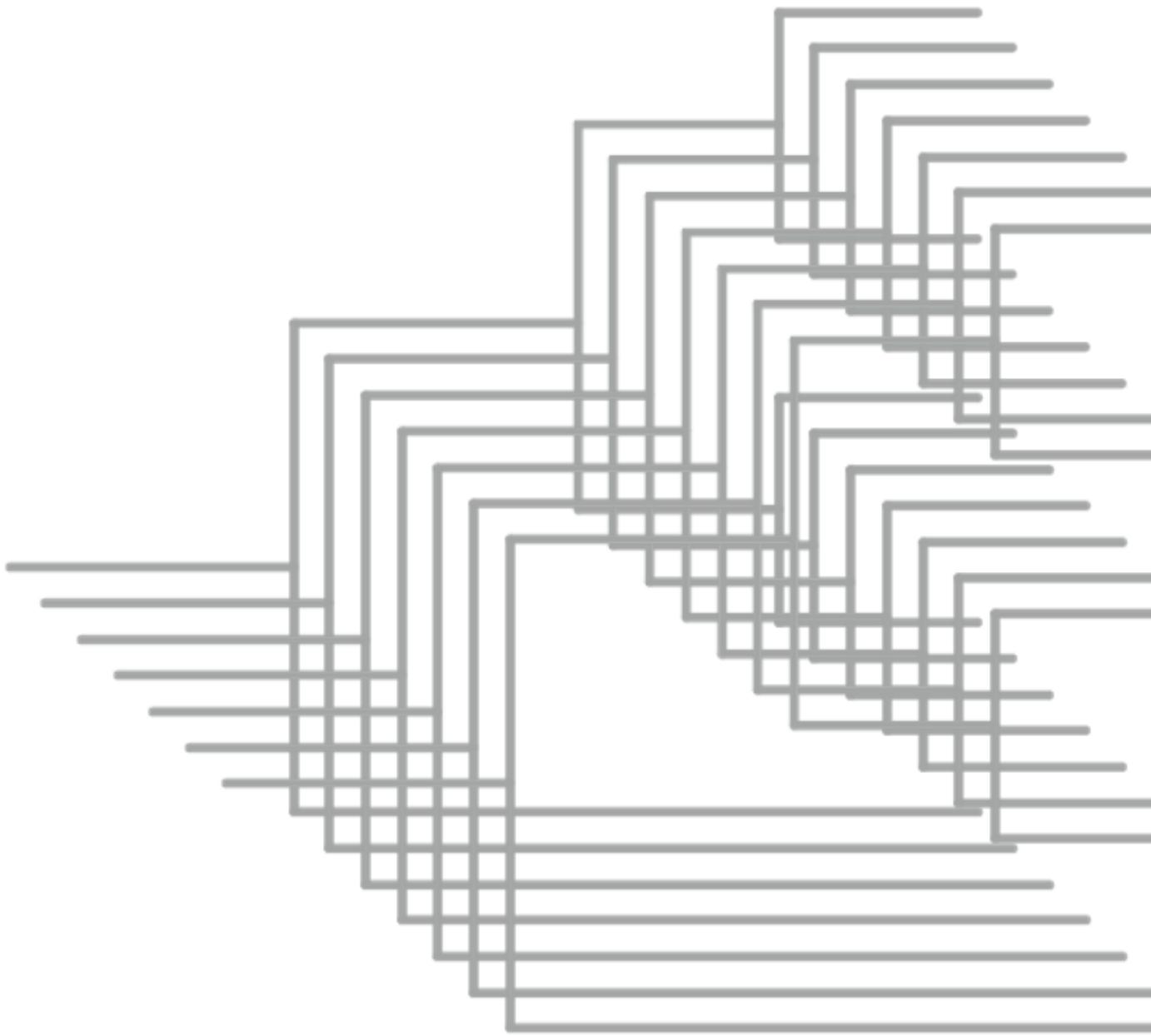


Posterior

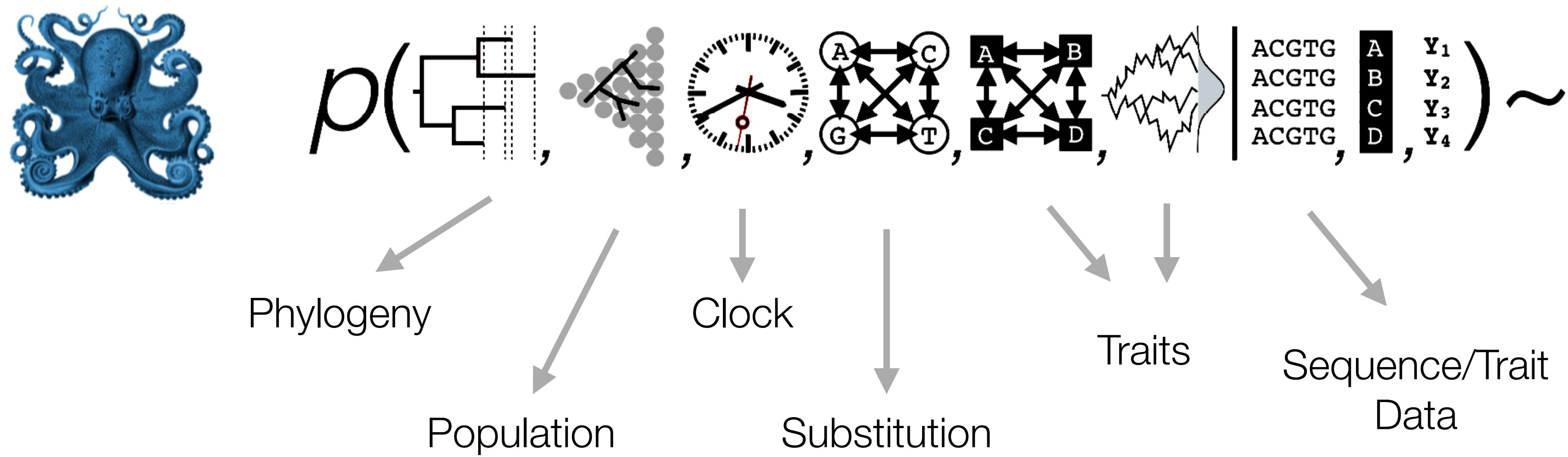
Bayesian

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

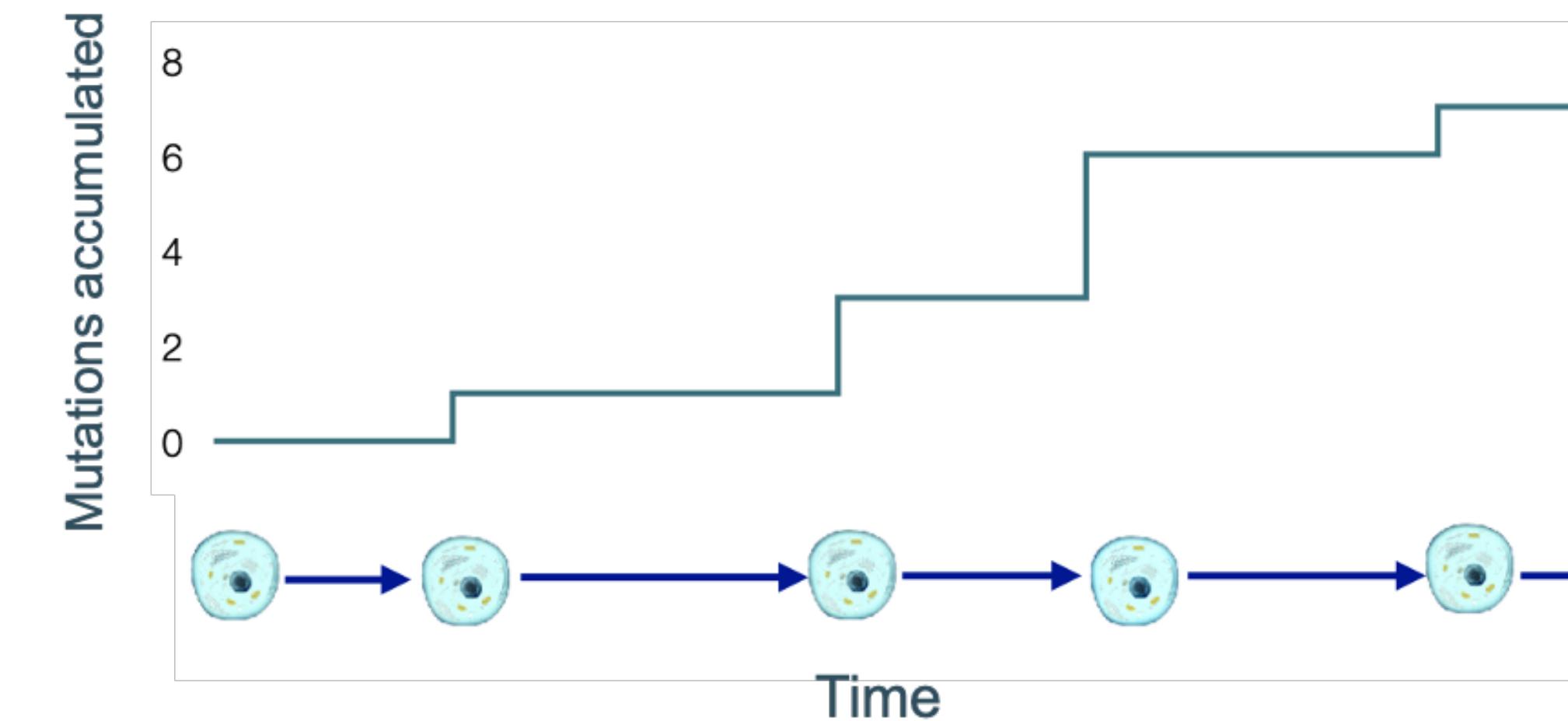
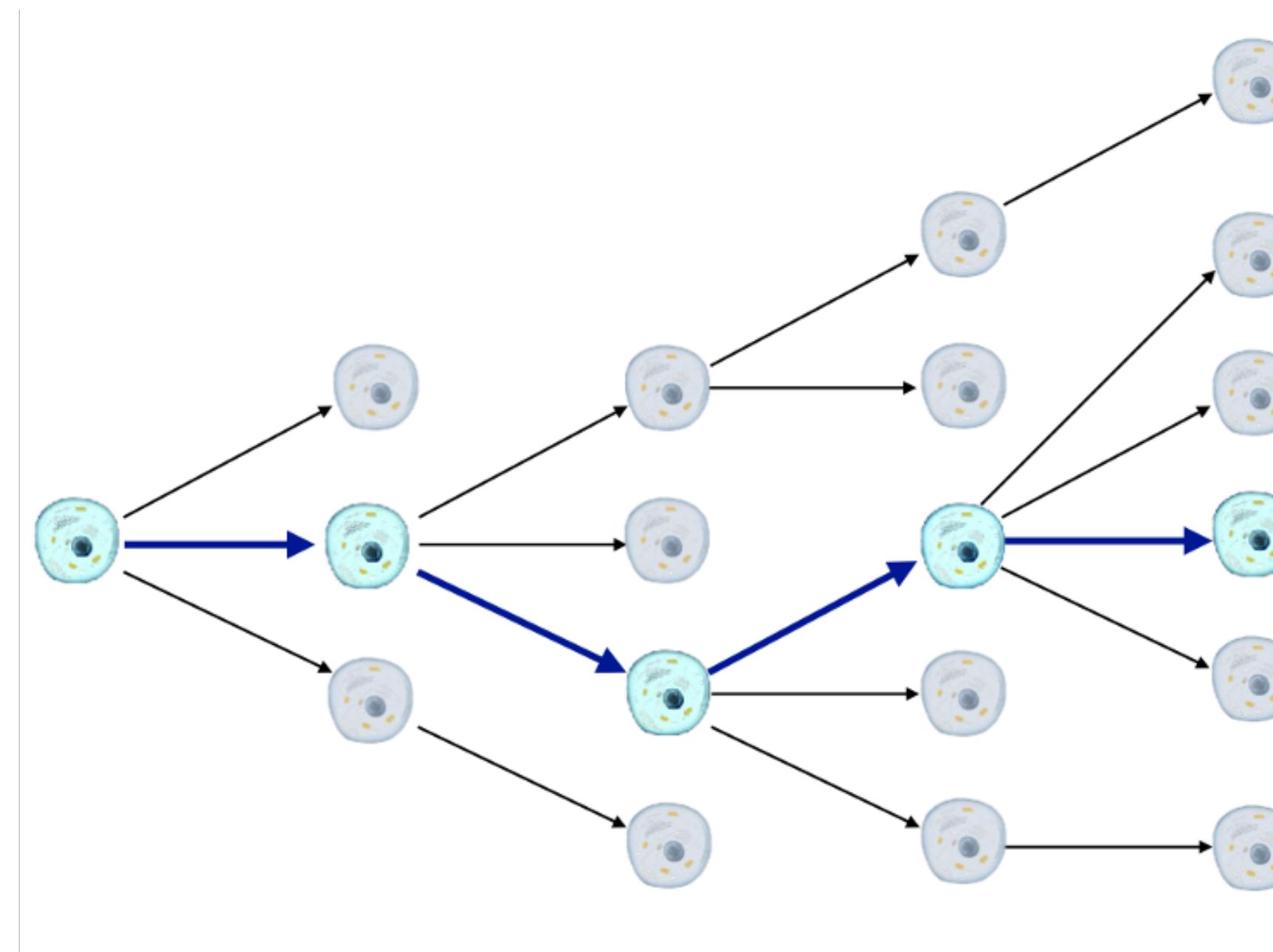
Prior



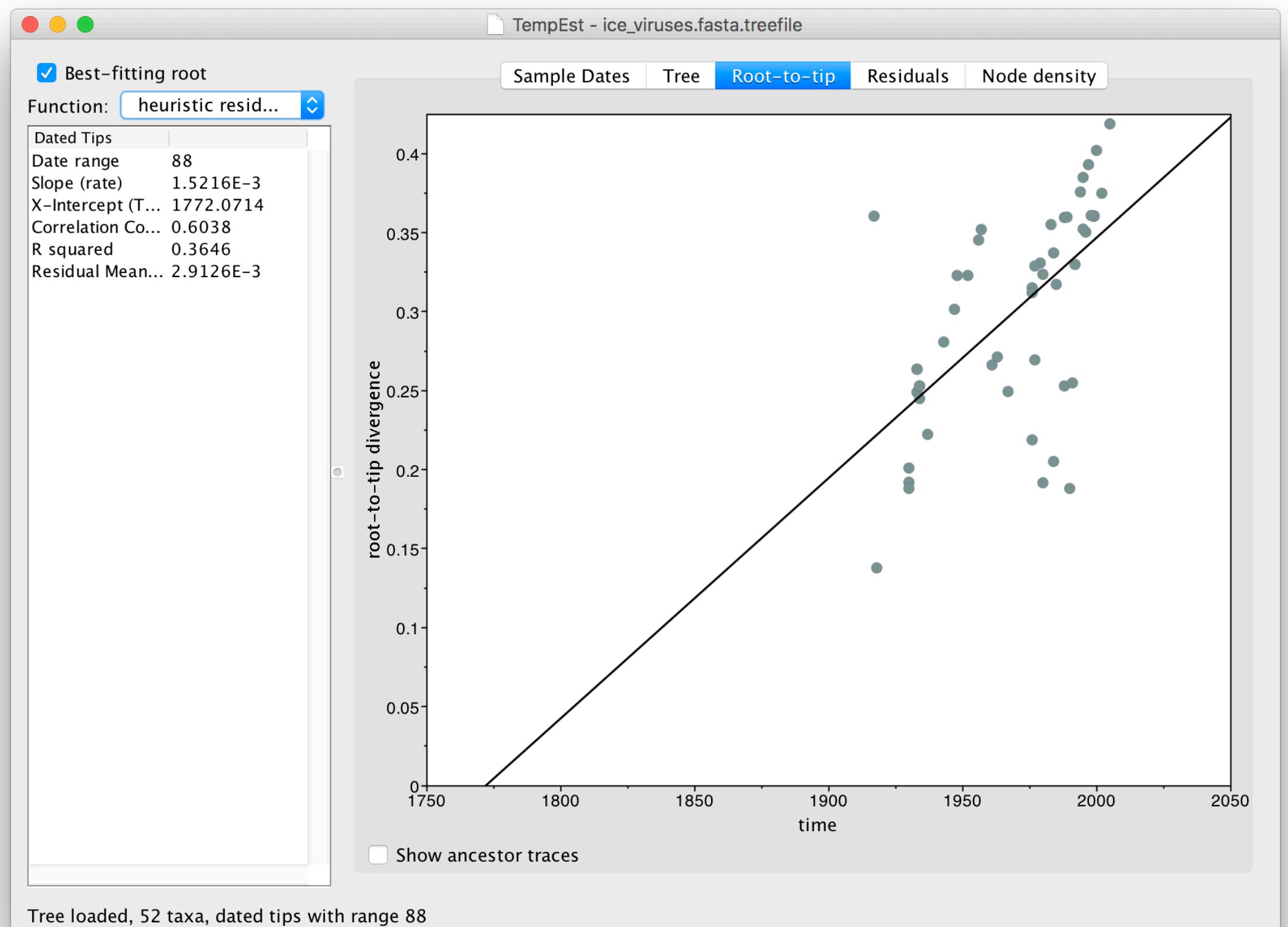
# Bayesian phyloodynamics with BEAST



# Learning from trees: molecular clocks

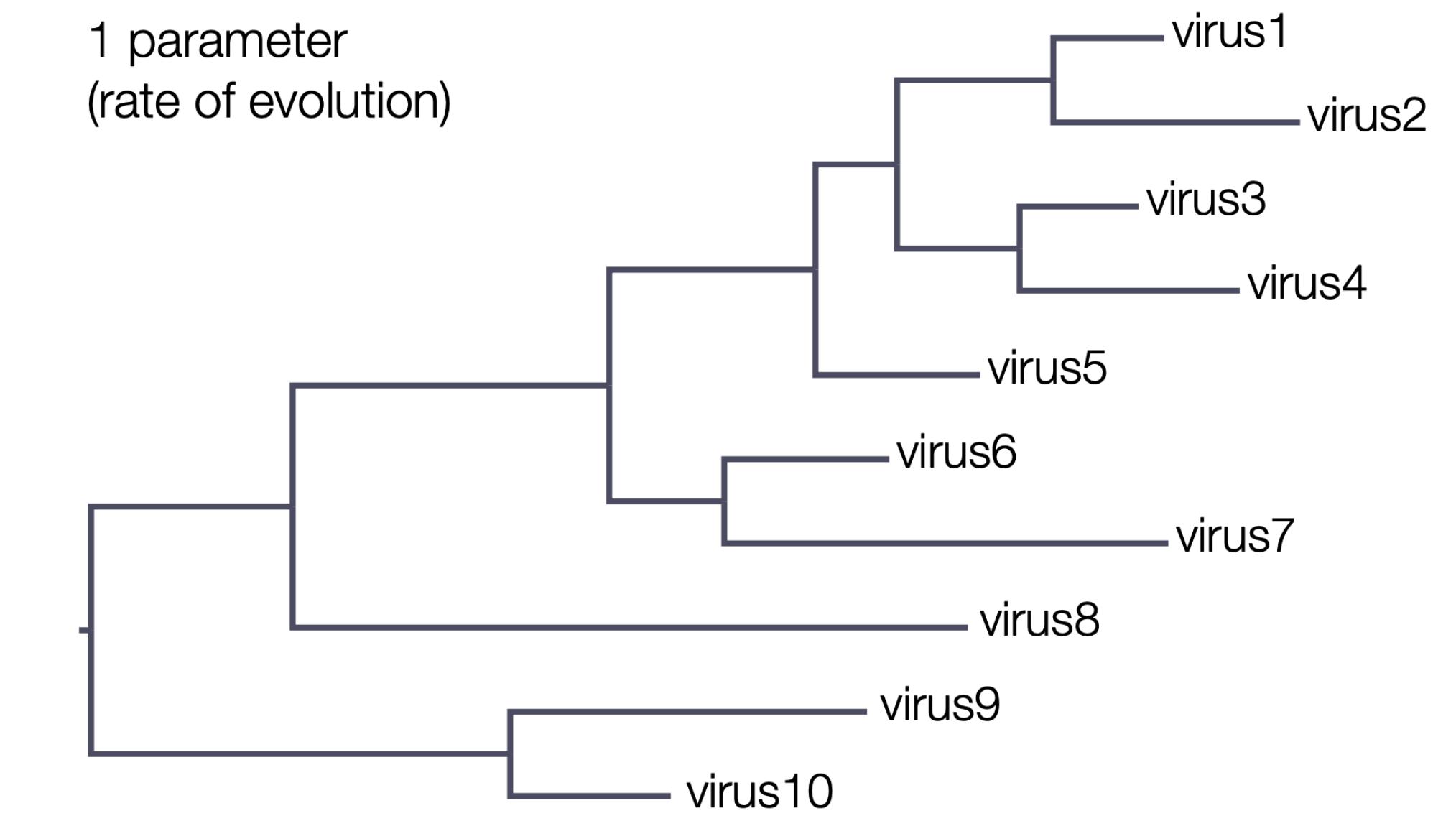


# Clock models



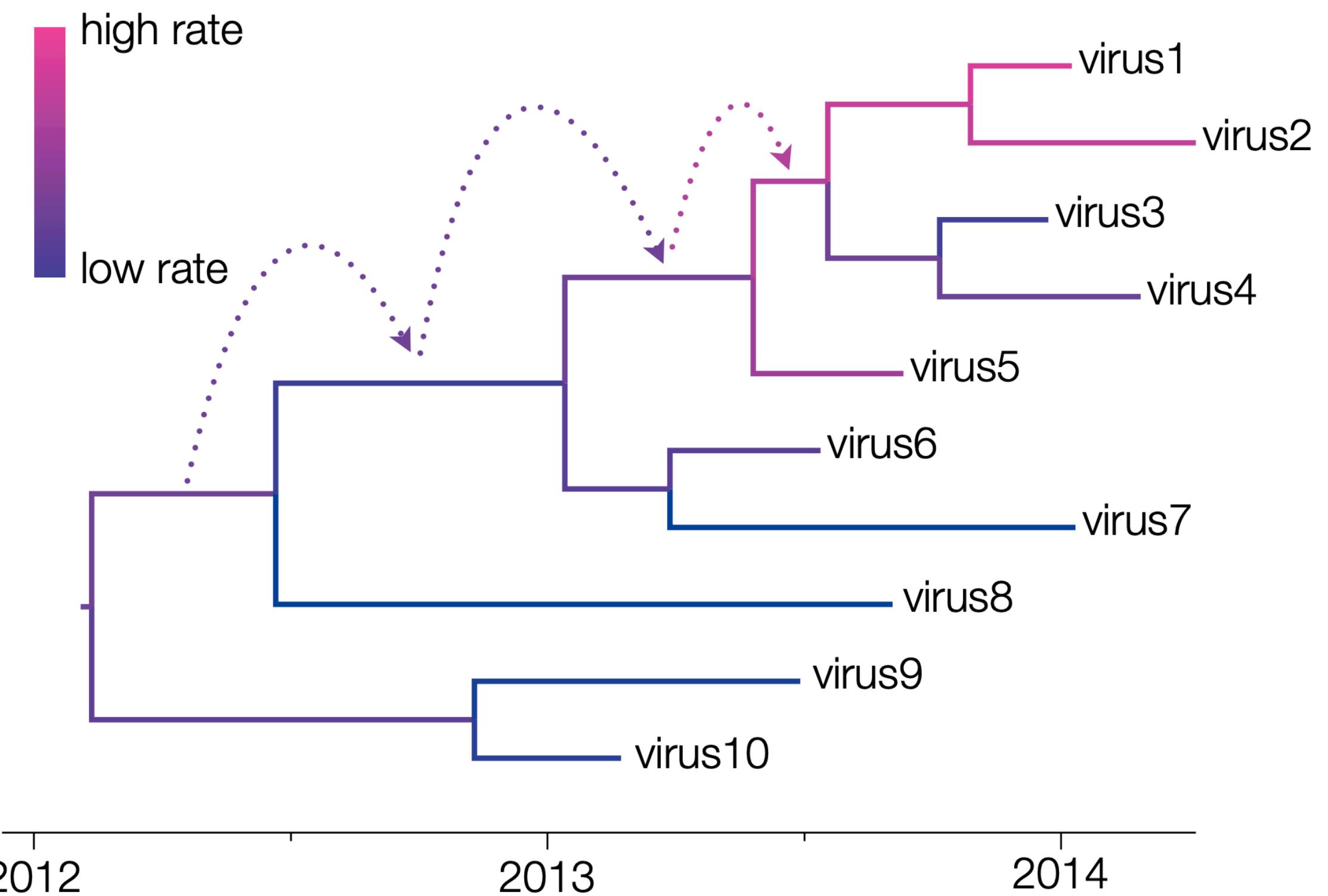
'strict' molecular clock

1 parameter  
(rate of evolution)

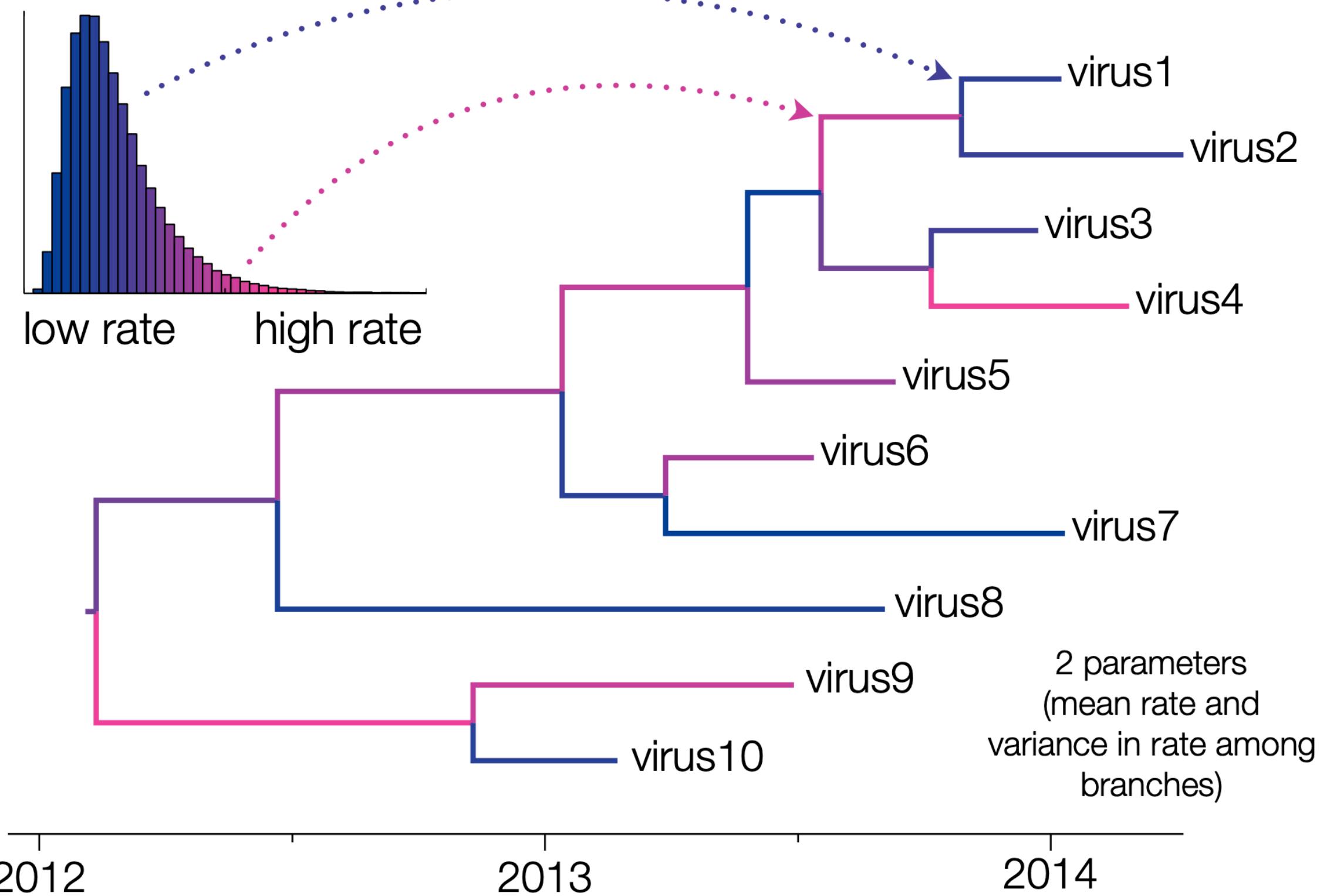


# Relaxed Clocks

autocorrelated relaxed clock



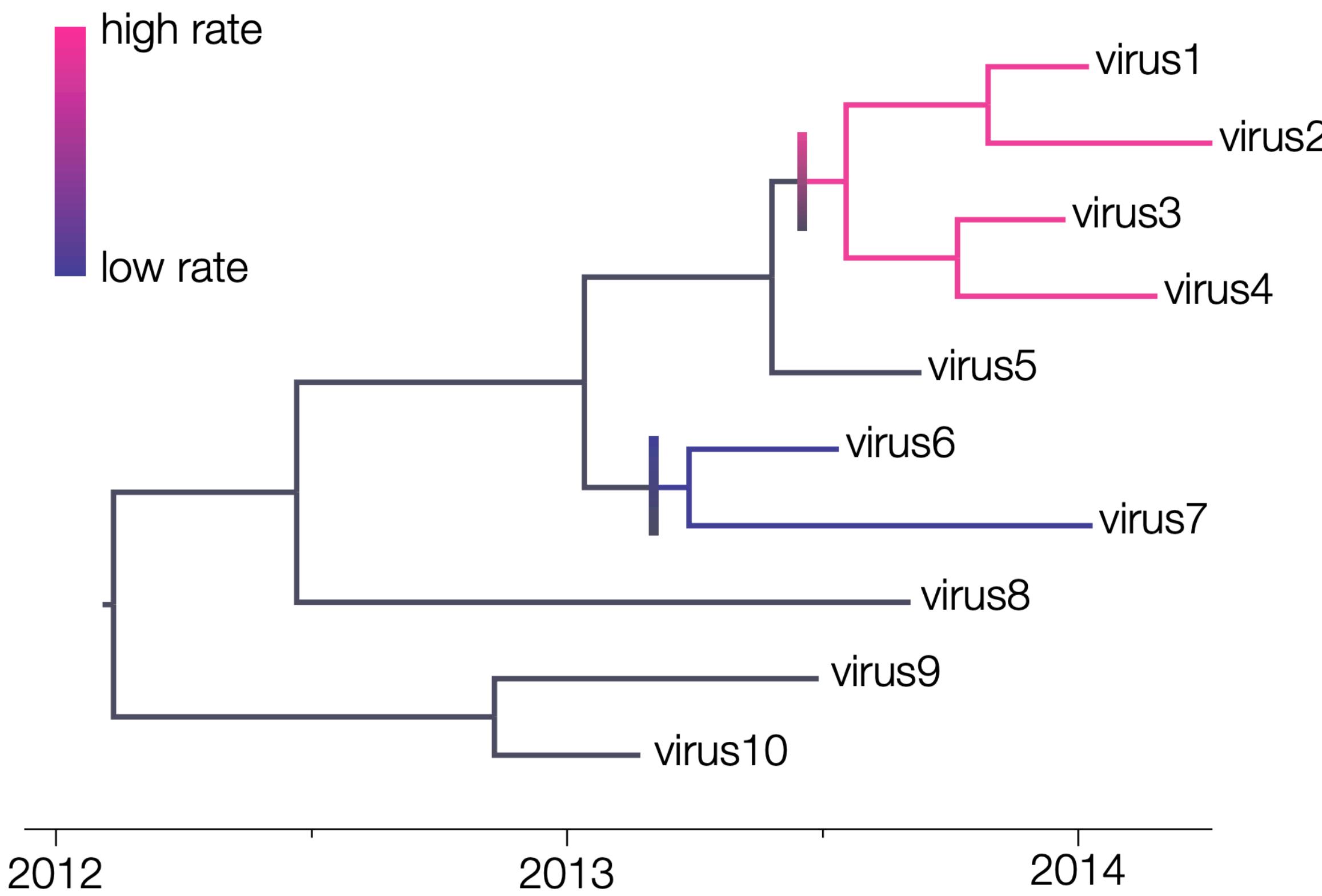
lognormal uncorrelated relaxed clock



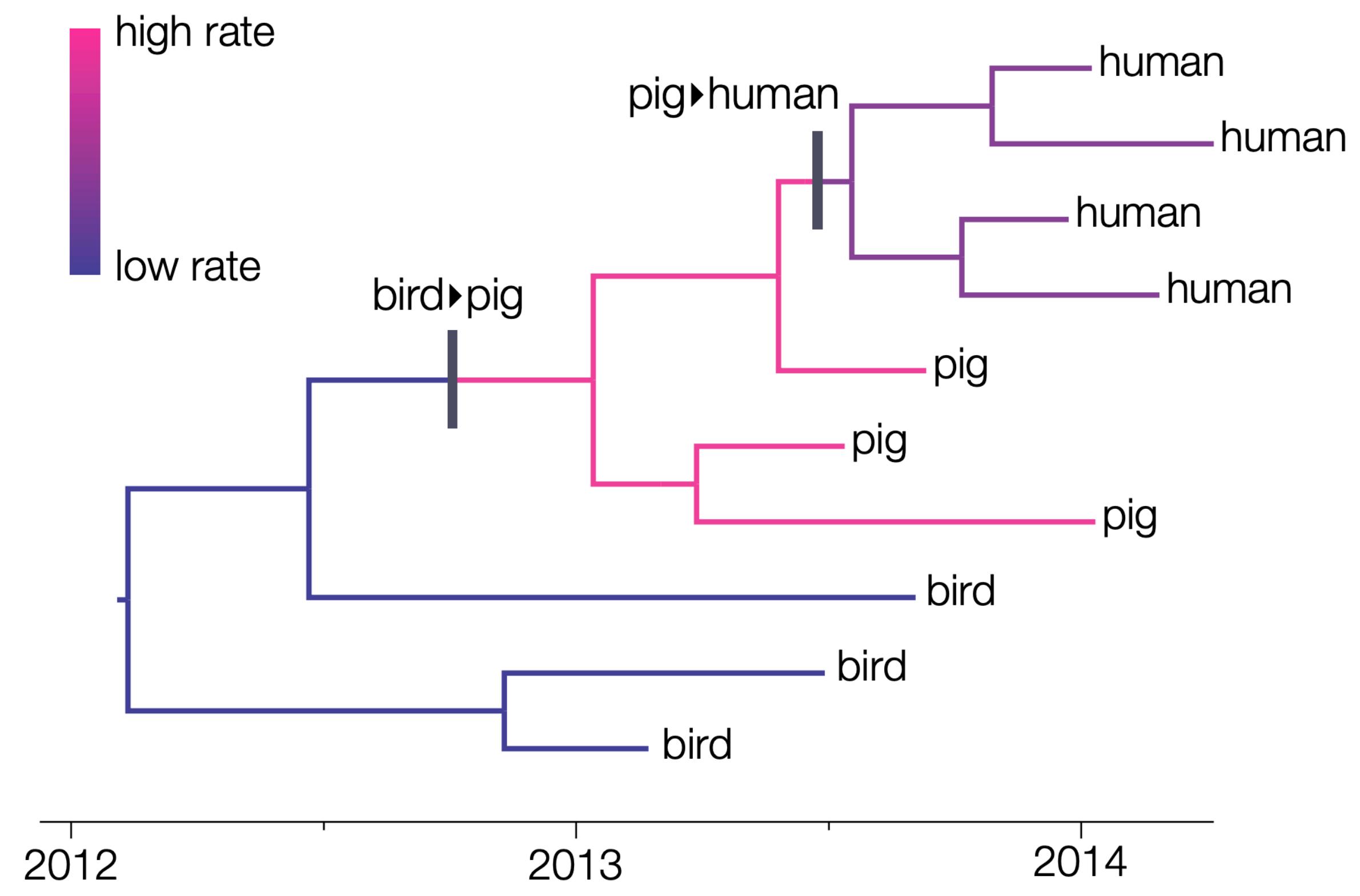
Adapted from Andrew Rambaut

# Local clocks

'local' molecular clock



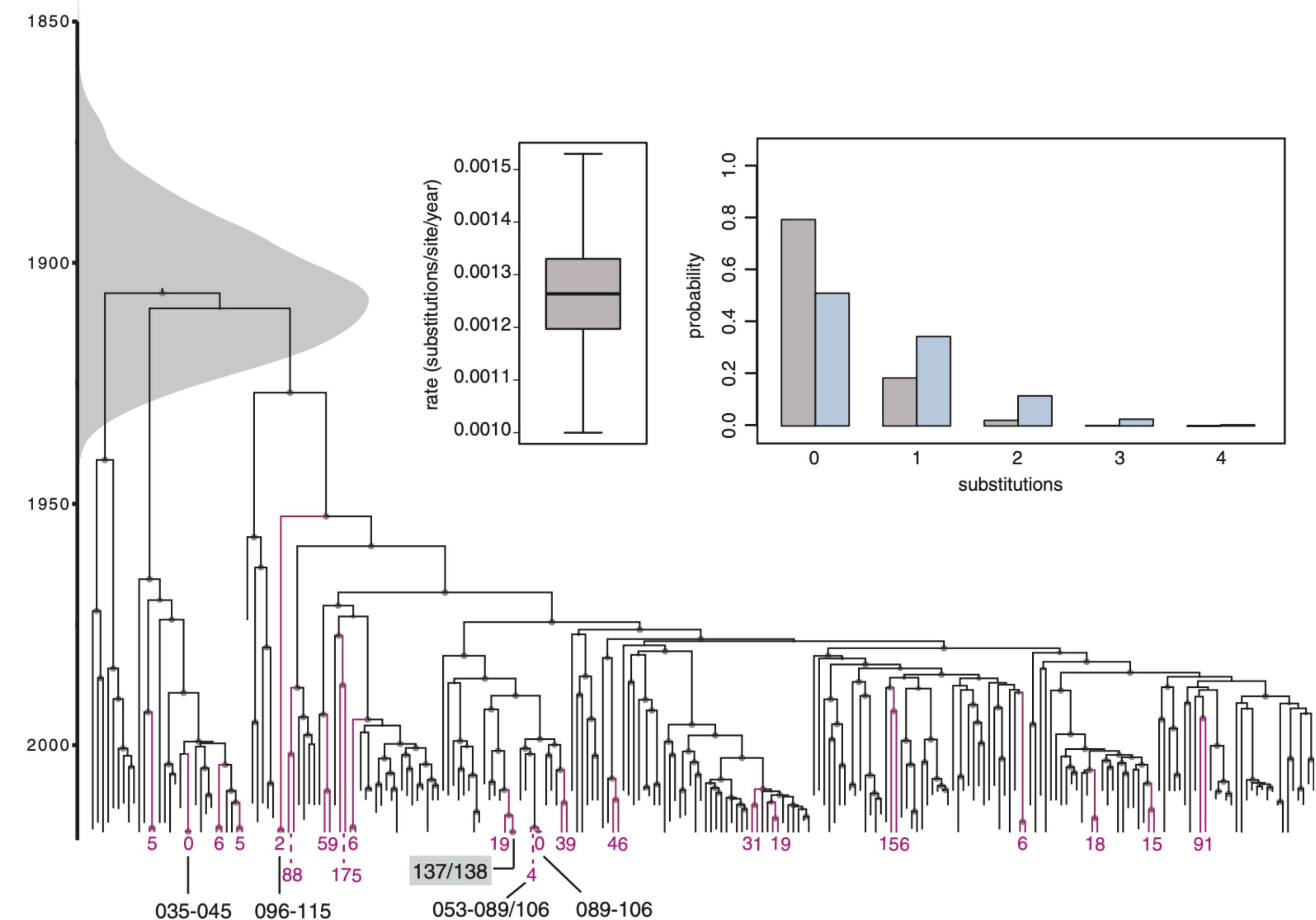
host specific local clock



Adapted from Andrew Rambaut

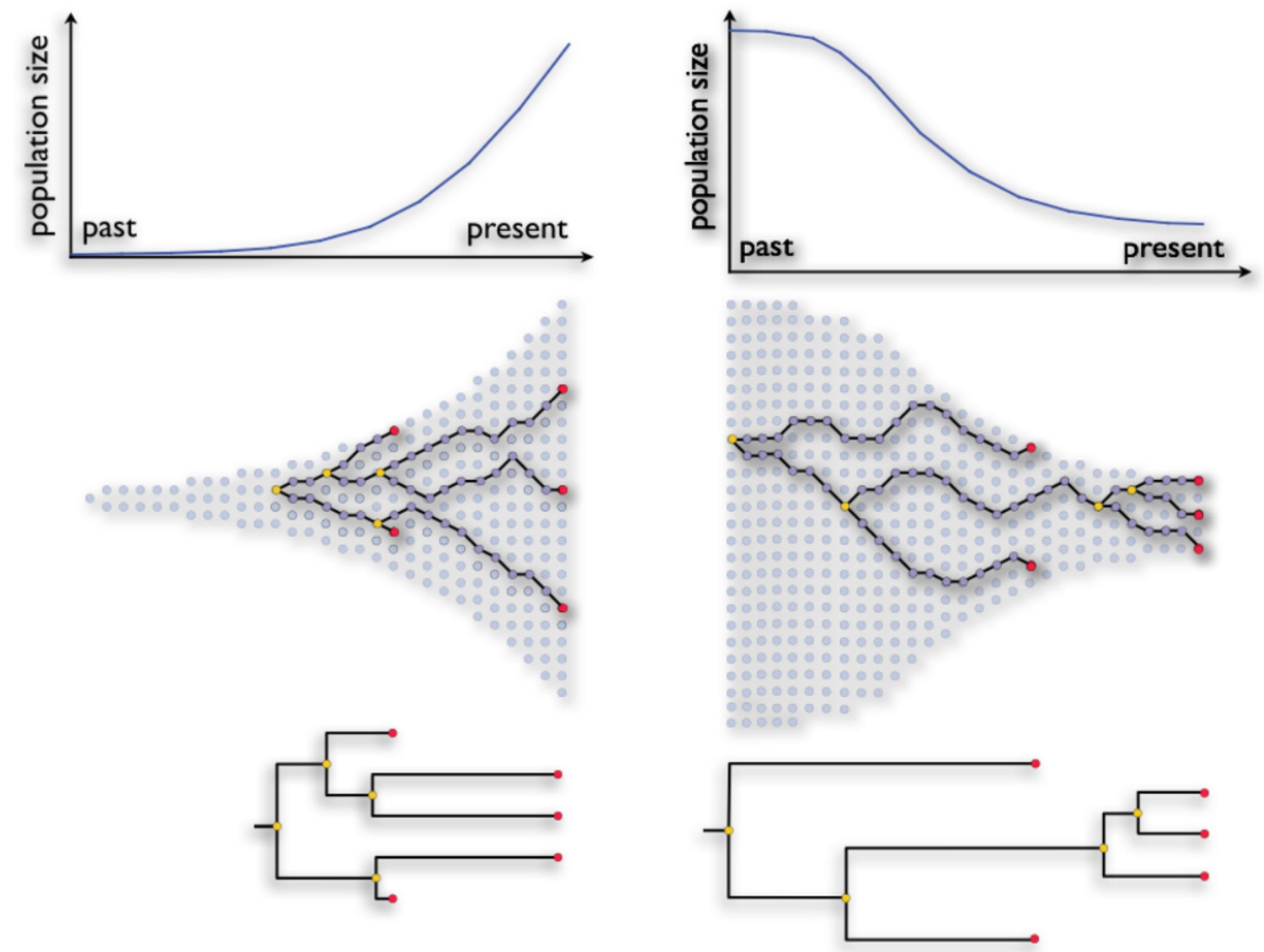
# Time calibration applications

- Transforming branch lengths into units of time allow us to answer epidemiological questions such as :
  - When did the outbreak begin? (TMRCA)
  - How fast is the virus evolving? (Clock rate)
  - What is the mode of transmission?



# Learning from trees: Coalescent

- Coalescent theory: Statistical model of the probability of two samples to merge (coalesce) after a given time
- We can learn about changes in population size from the shape of a tree:
  - The larger the population size, the longer it will take for two lineages to coalesce
  - Application: Calculate the exponential growth rate of an epidemic
    - How rapidly is the virus spreading? (Exponential growth)



# Coalescent: Exponential Growth

$$N_e(t) = N_0 e^{rt}$$

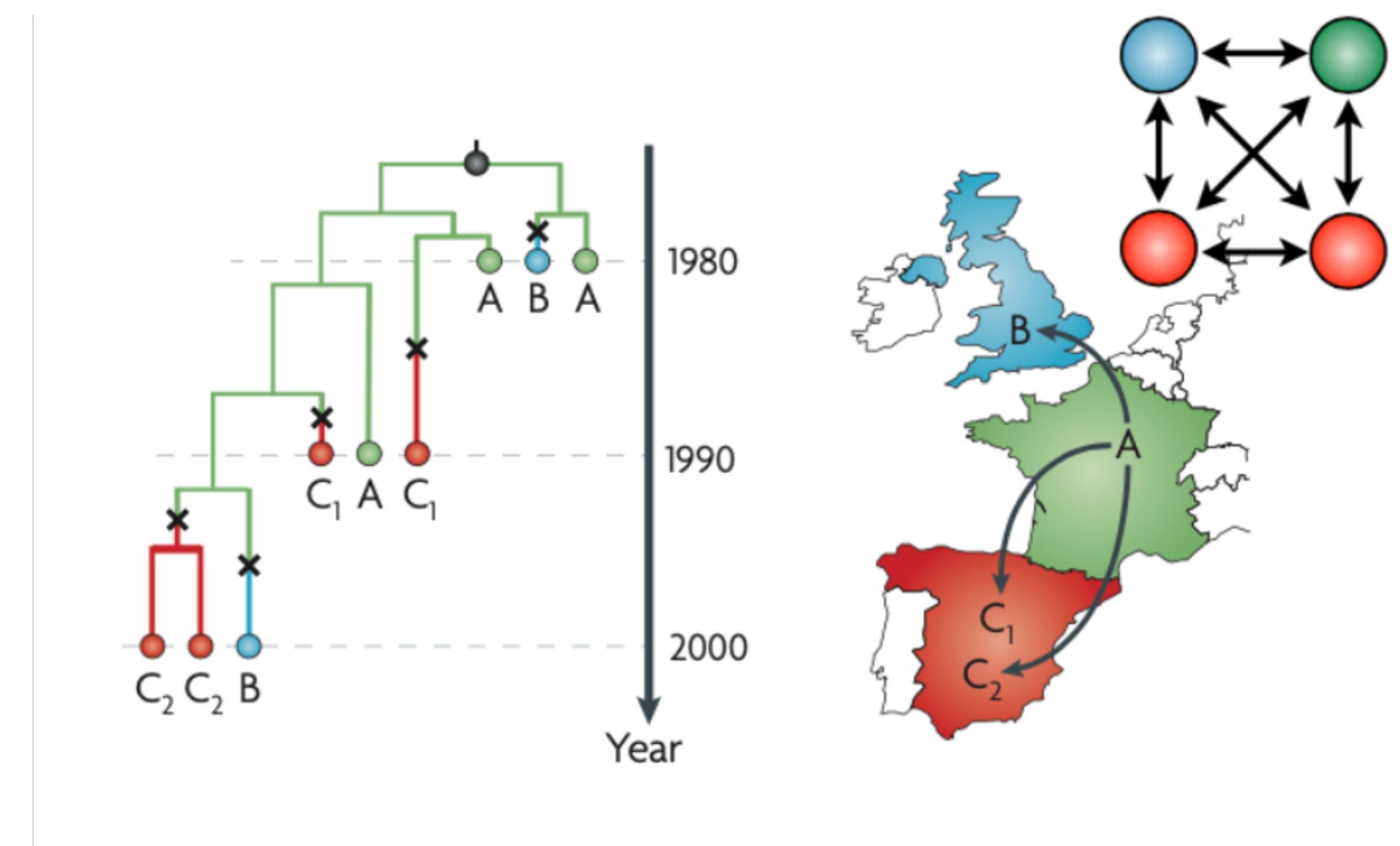
Effective population size      Initial population size      Growth rate

Doubling time:  $N_e(t_{double}) = 2N_0 = N_0 e^{rt_{double}}$  →  $t_{double} = \frac{\log(2)}{\text{Growth rate}}$

Solve for t

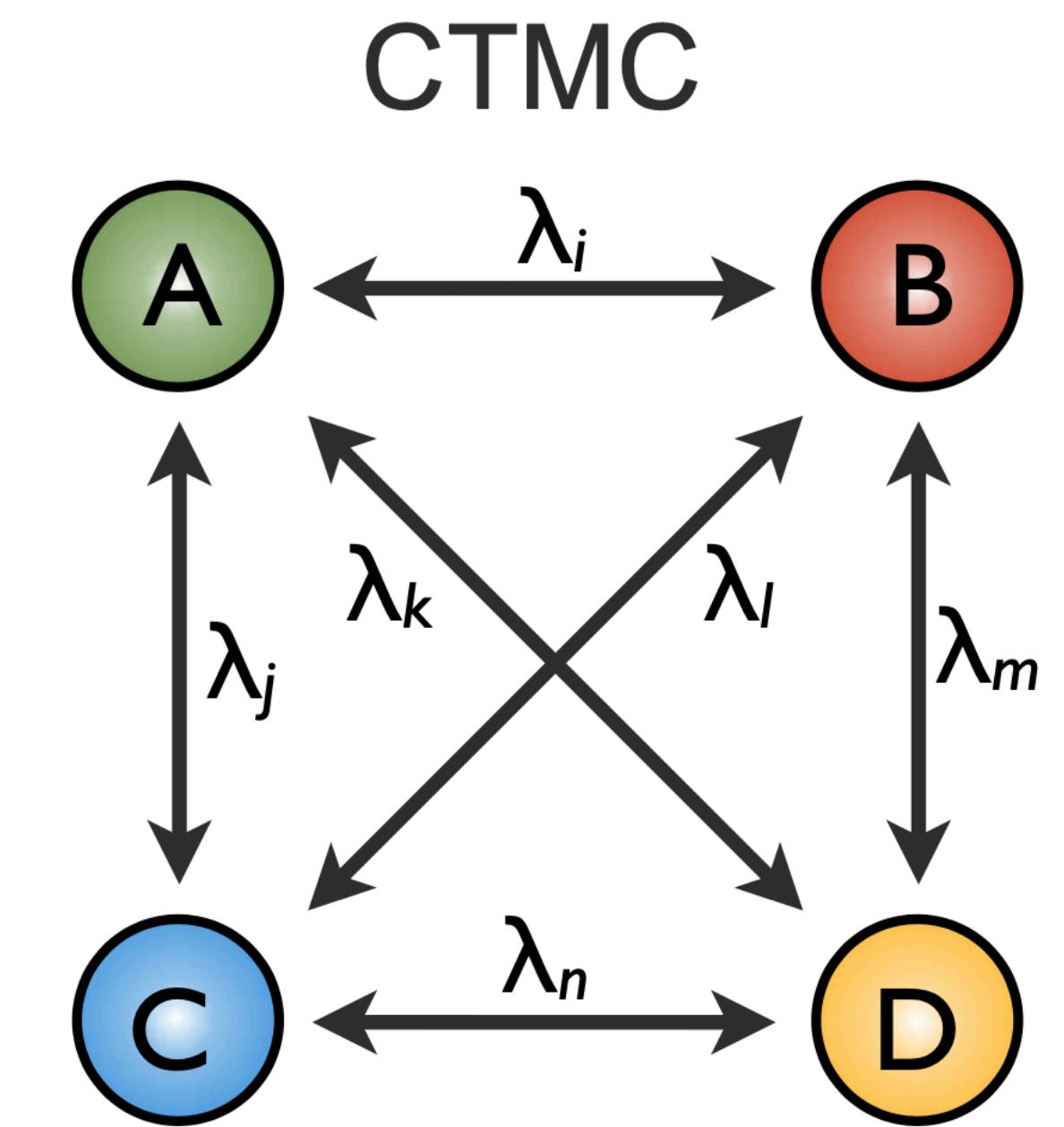
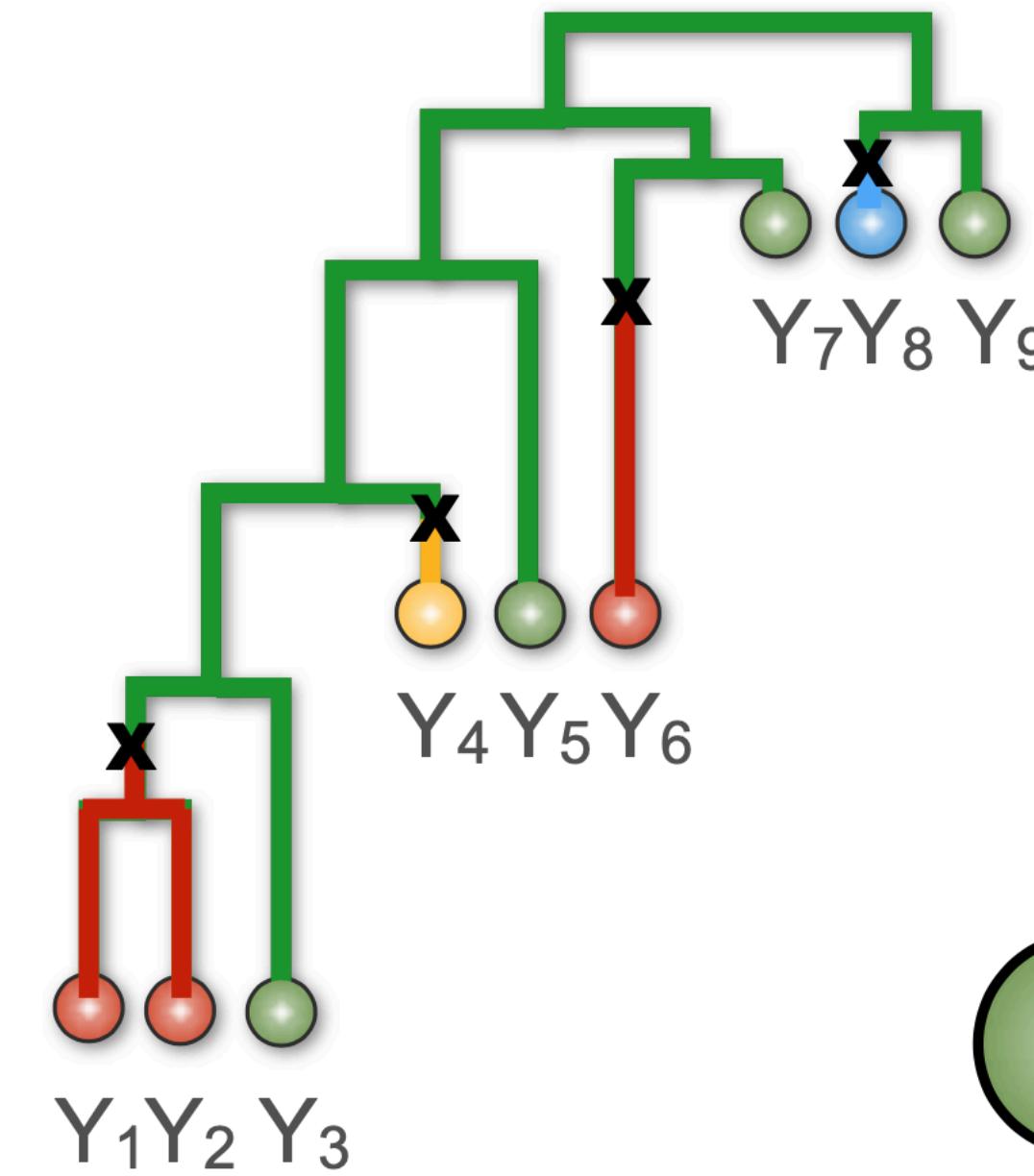
# Learning from trees: Phylogeography

- Discrete trait analysis: model discrete locations as a trait coloring the tree
- Reconstruct probability of jumping from one location to another and expected number of jumps
- Answer questions such as:
  - How many introductions?
  - Where did the outbreak begin?
  - What factors drive spread?

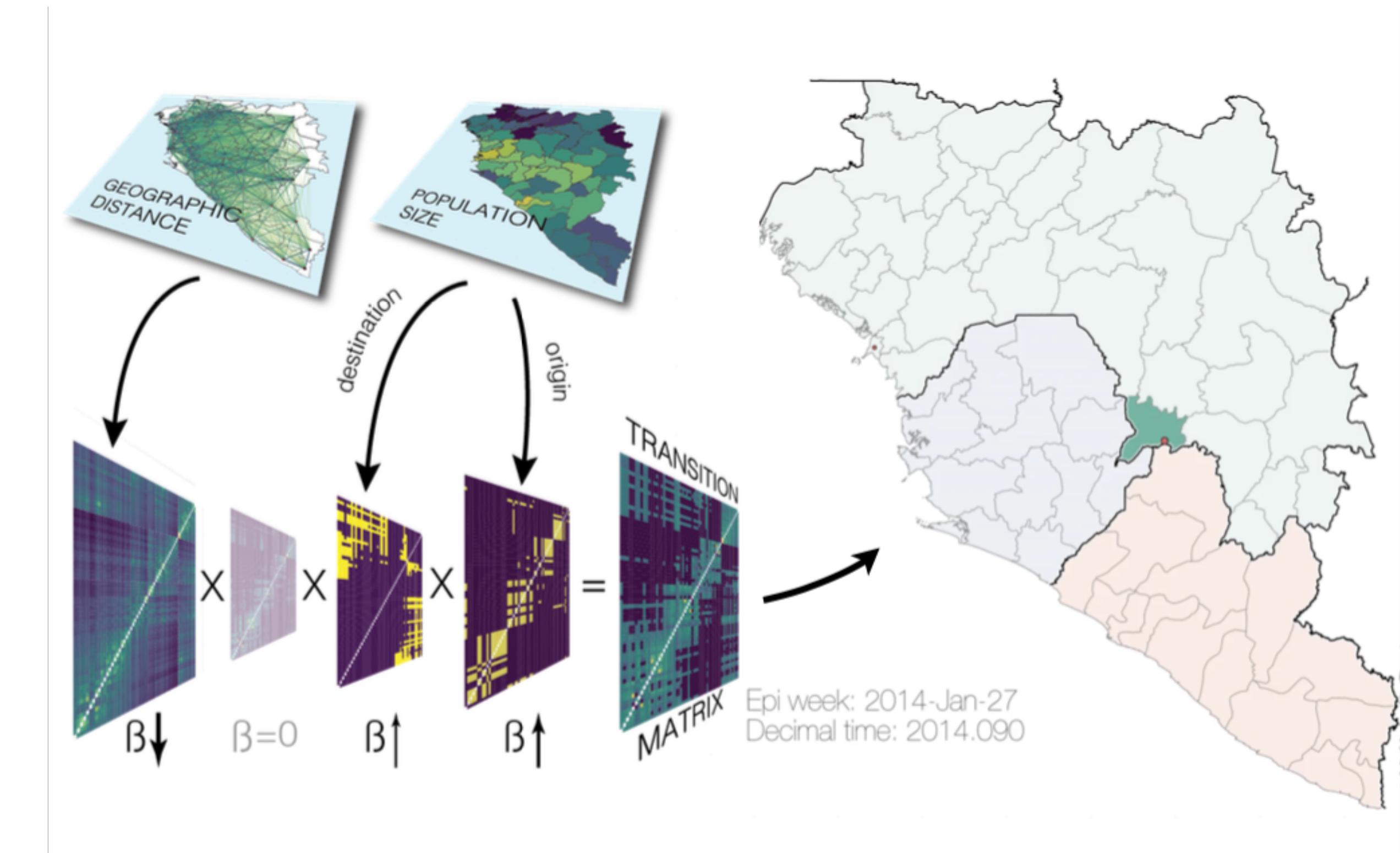
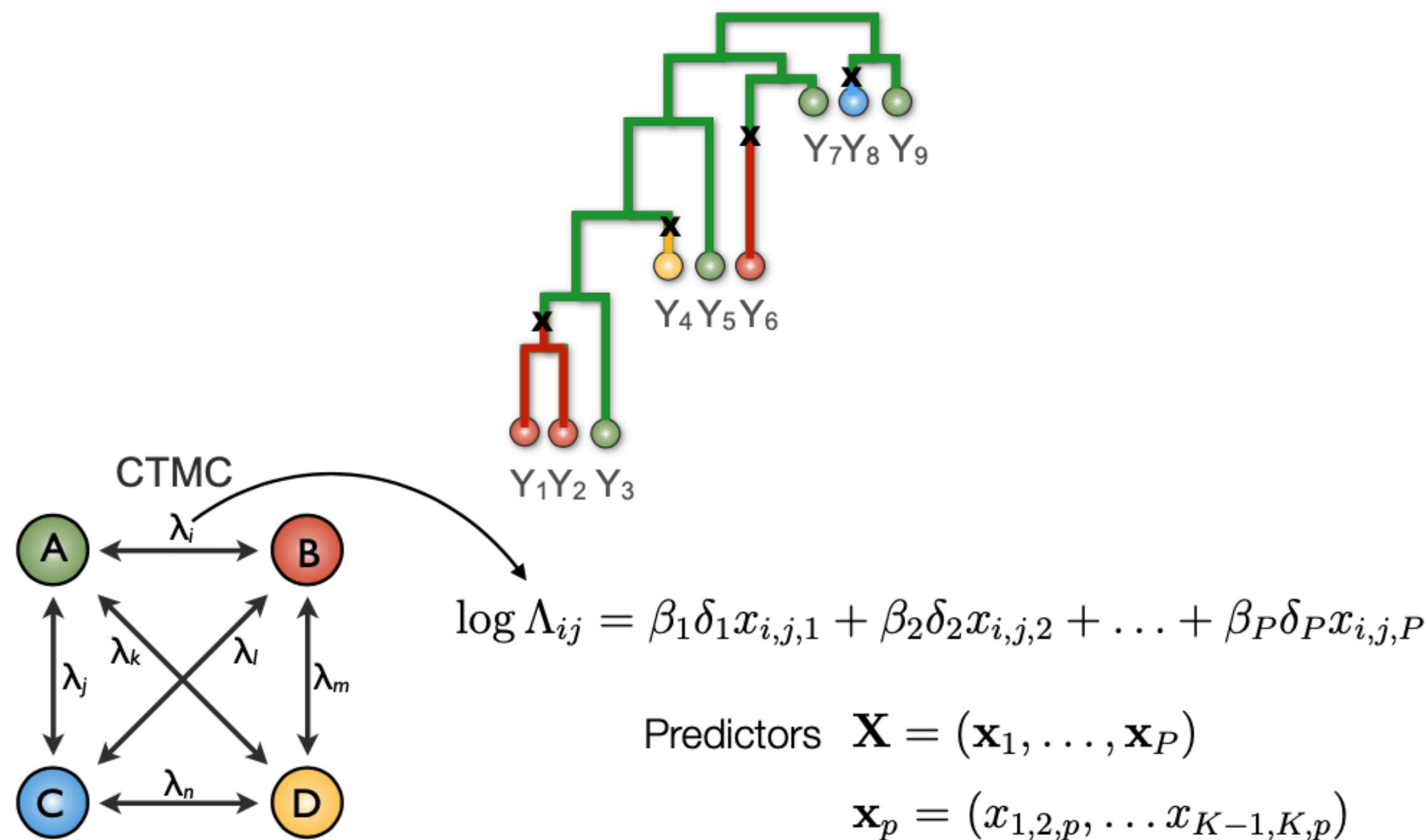


# Discrete phylogeography

- Model spatial spread as a random process characterized by a rate matrix that describes the probability of transitioning to a location given the current location
  - Analogous to nucleotide substitution models
- Transition rates can be symmetric or asymmetric
- Use rates to reconstruct the location of internal nodes
  - How many introductions?
  - What is the origin?



# Factors that drive spread?



Dudas et al. 2017

Adapted from Philippe Lemey

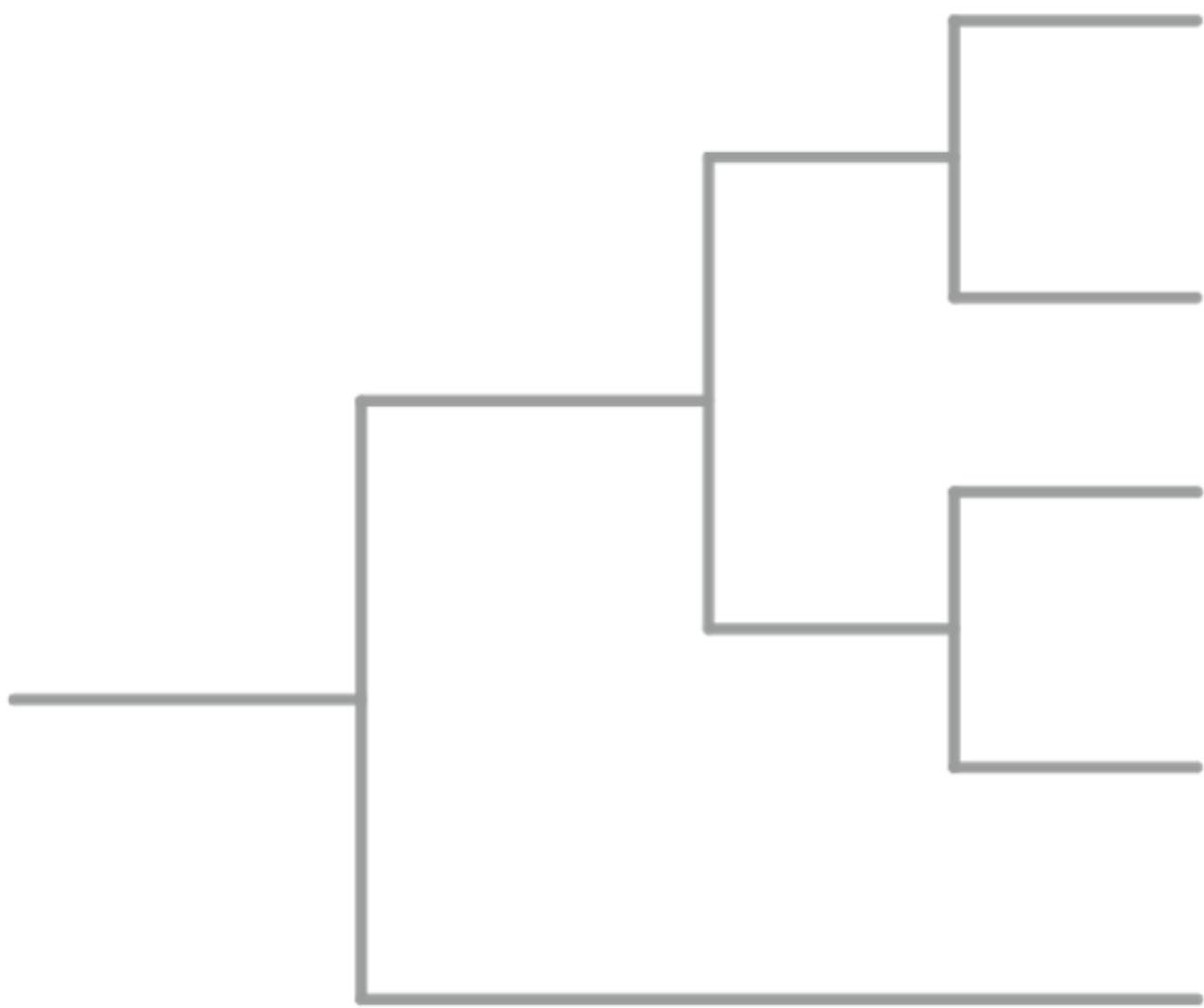
**How does Bayesian  
Phylogenetic Inference work?**

# Bayesian Phylogenetics

Maximum Likelihood

$$\operatorname{argmax}_{\theta} L(\theta)$$

Likelihood:  $L(\theta) \propto P(Data|\theta)$

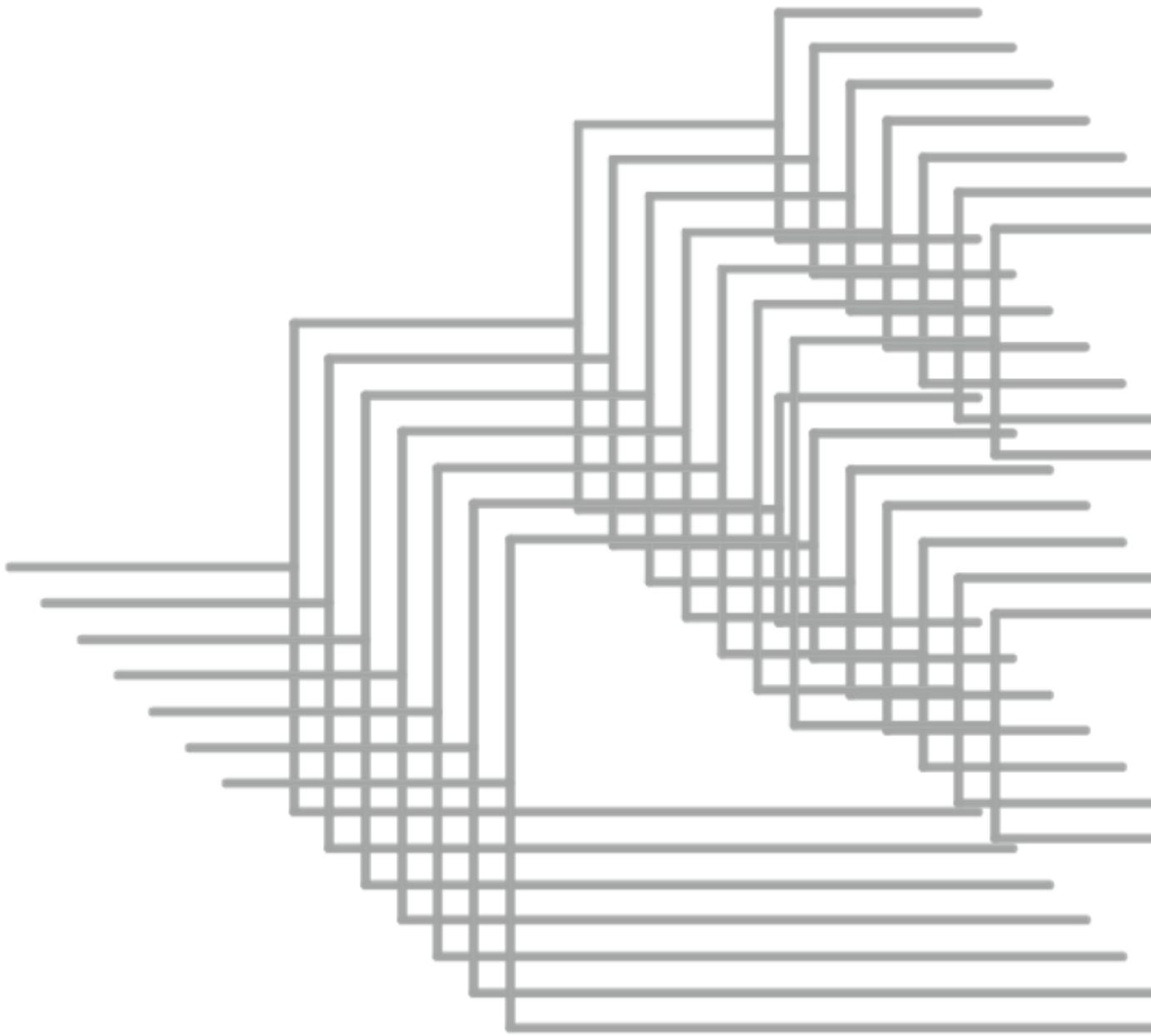


Posterior

Bayesian

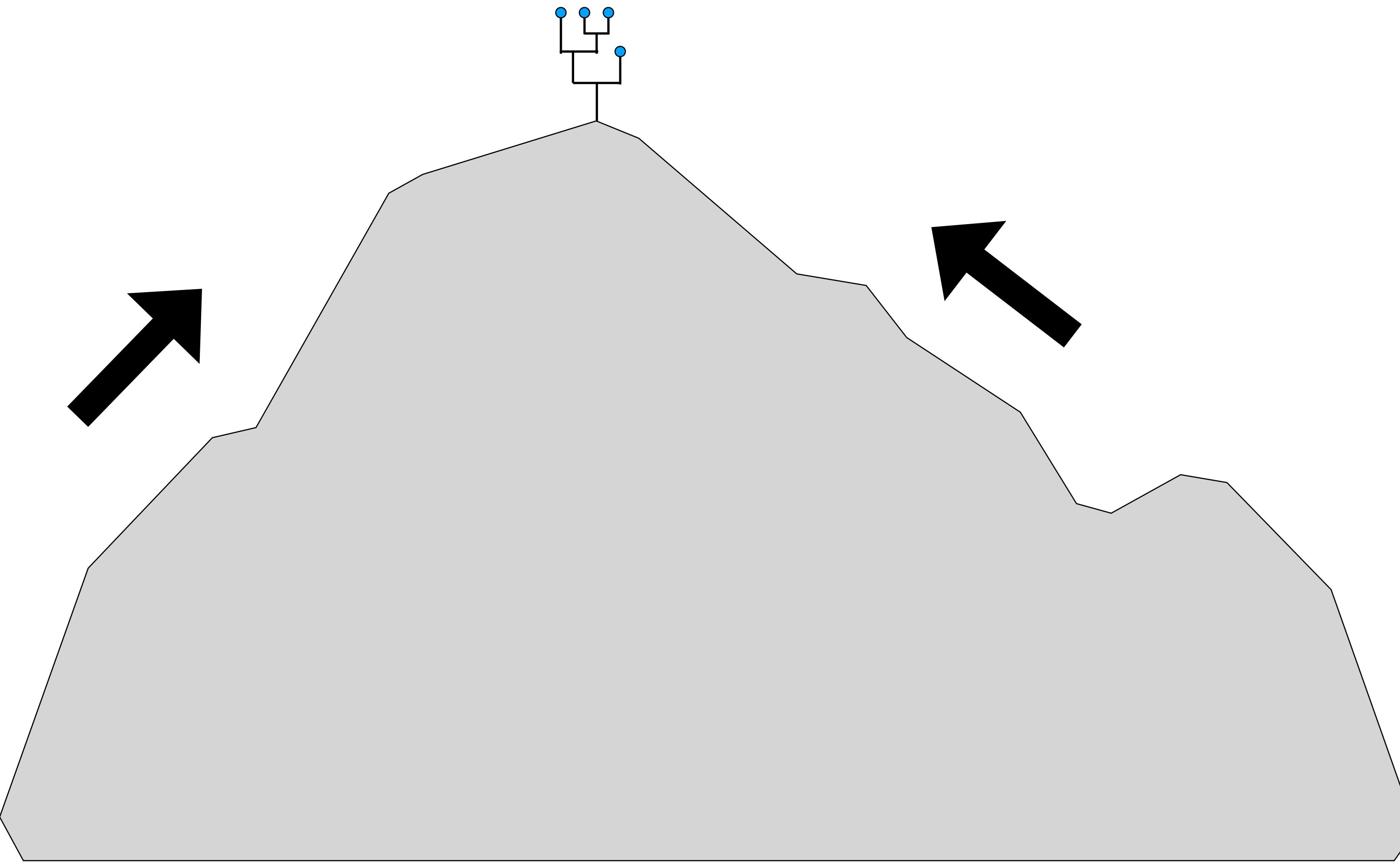
$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)}$$

Prior

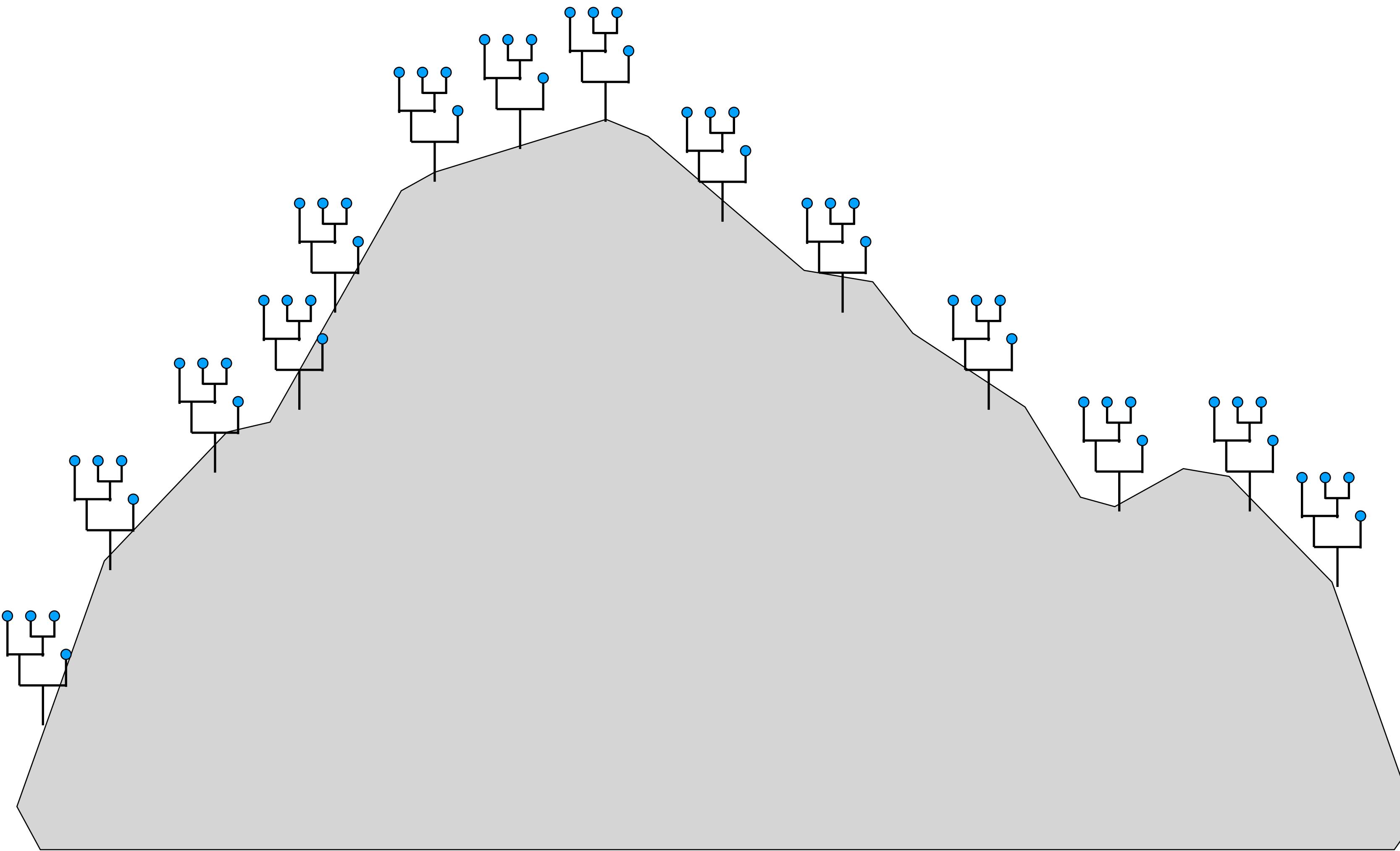


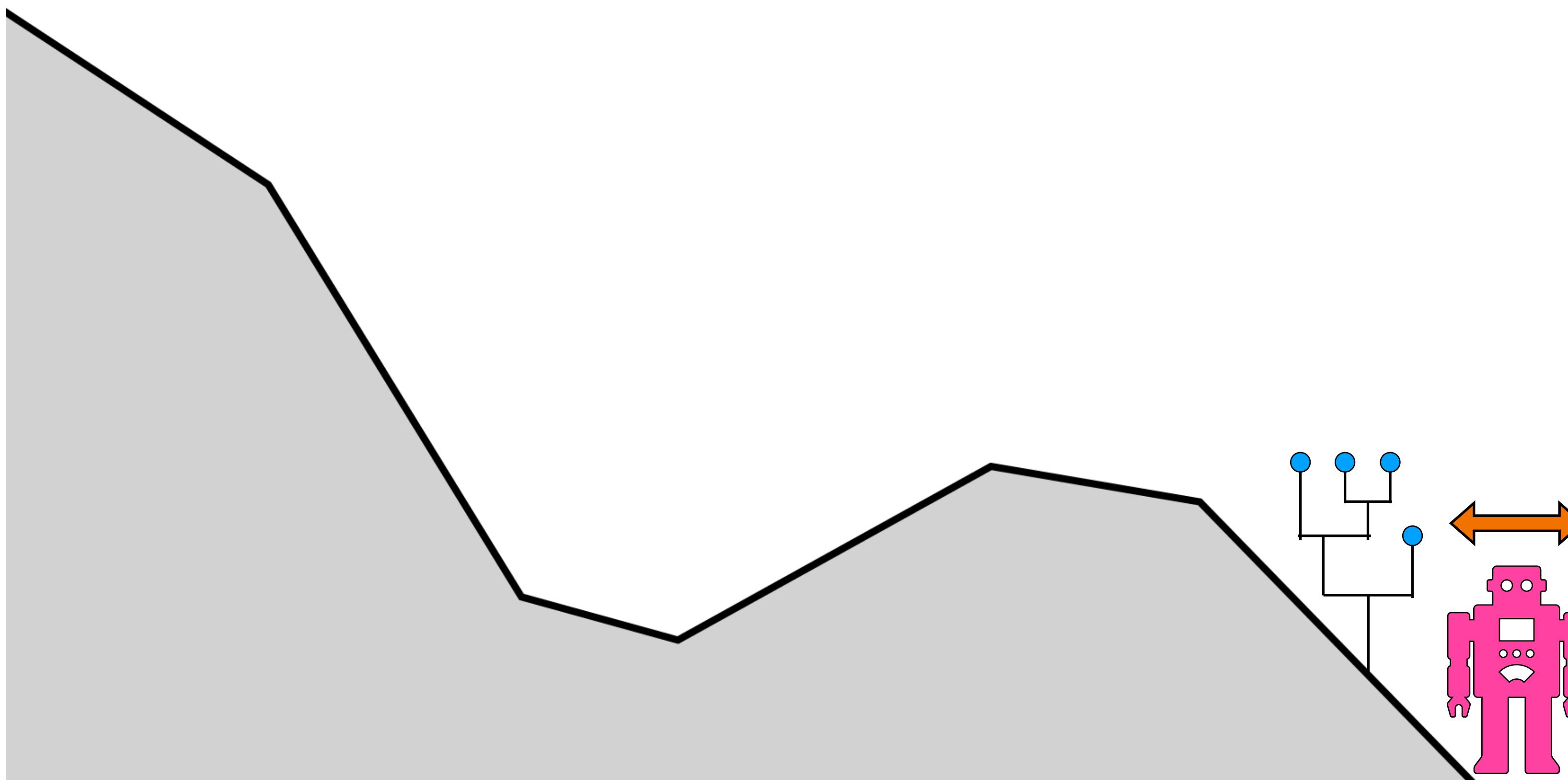


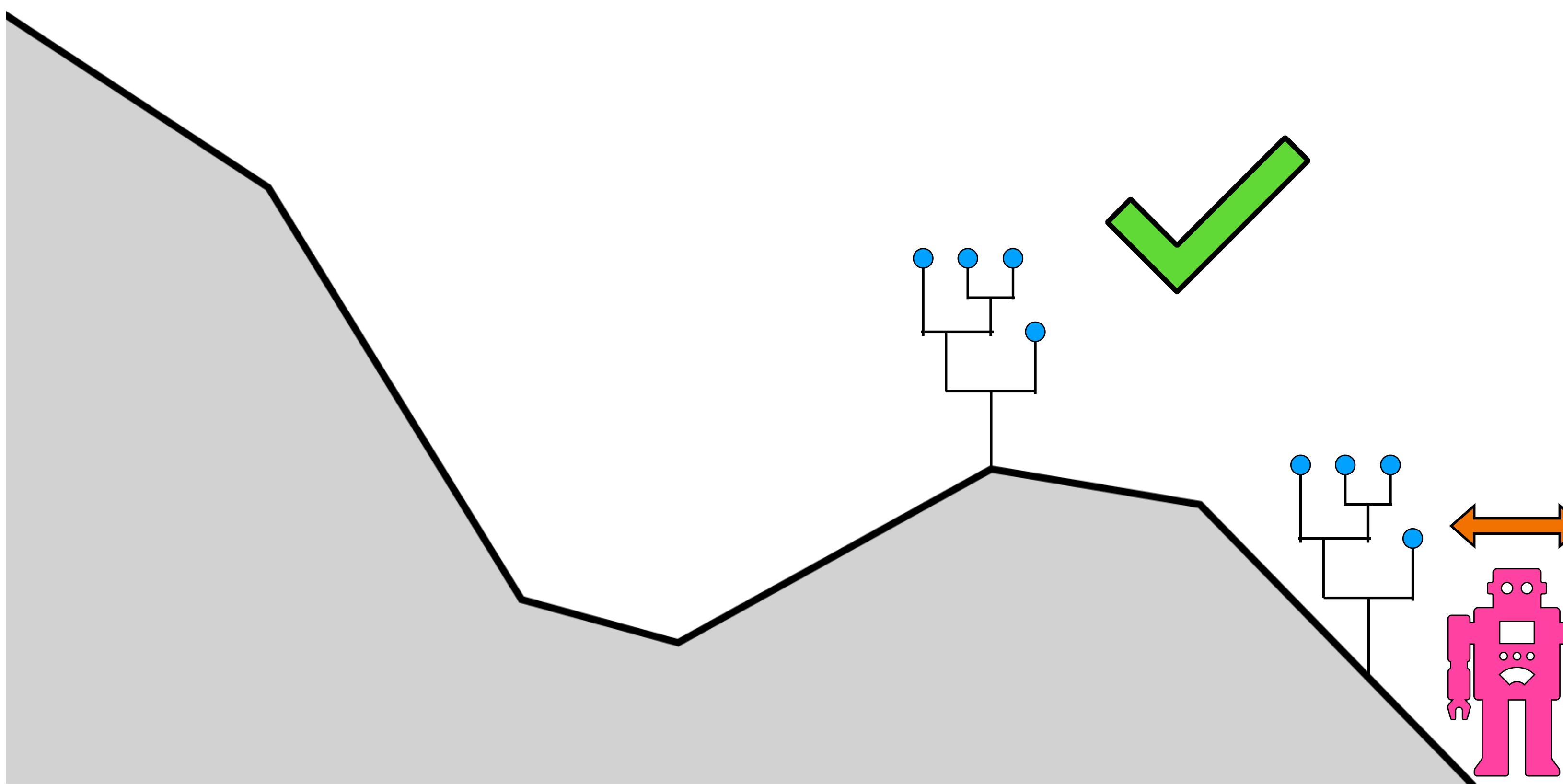
# Maximum Likelihood

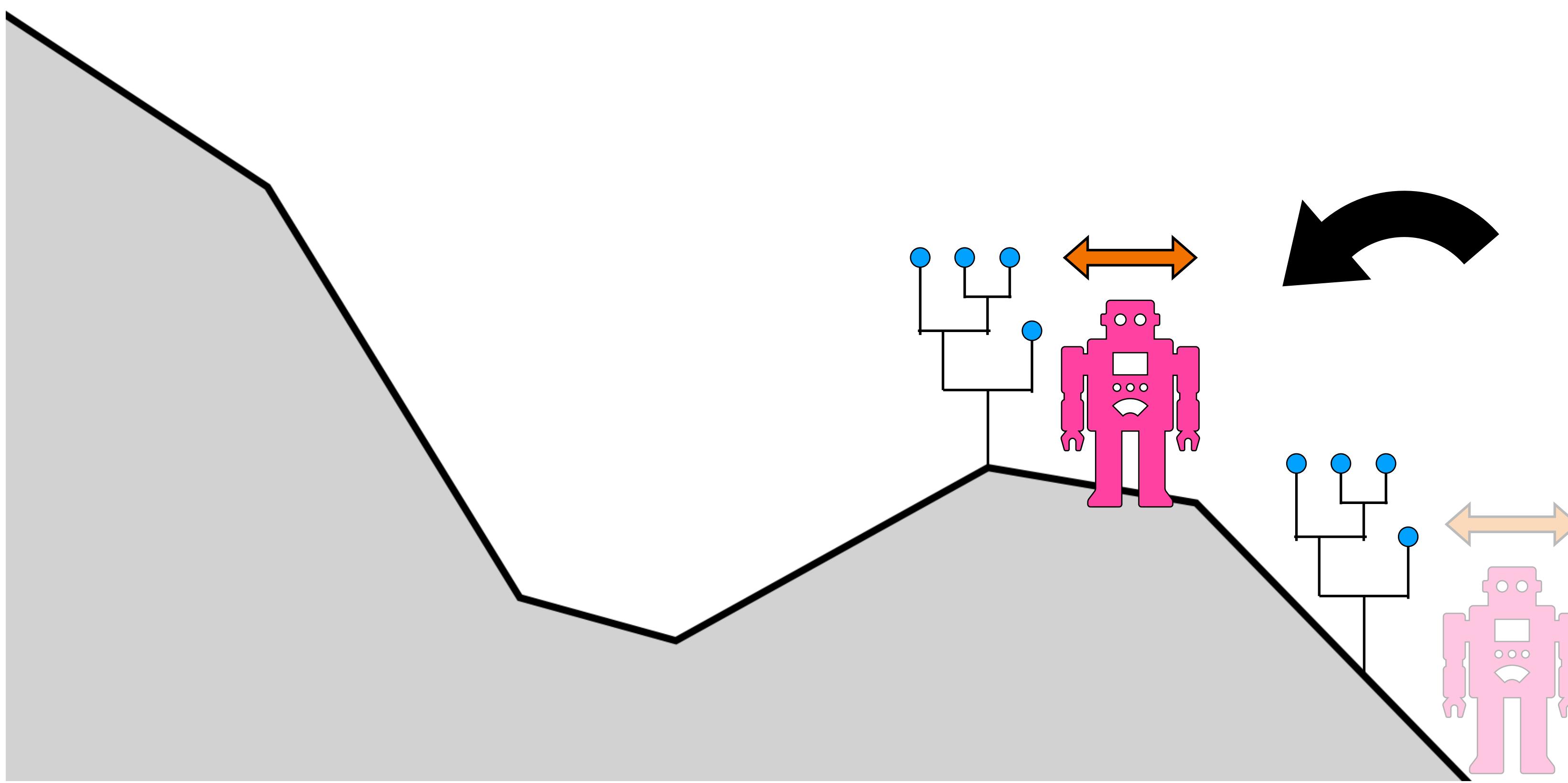


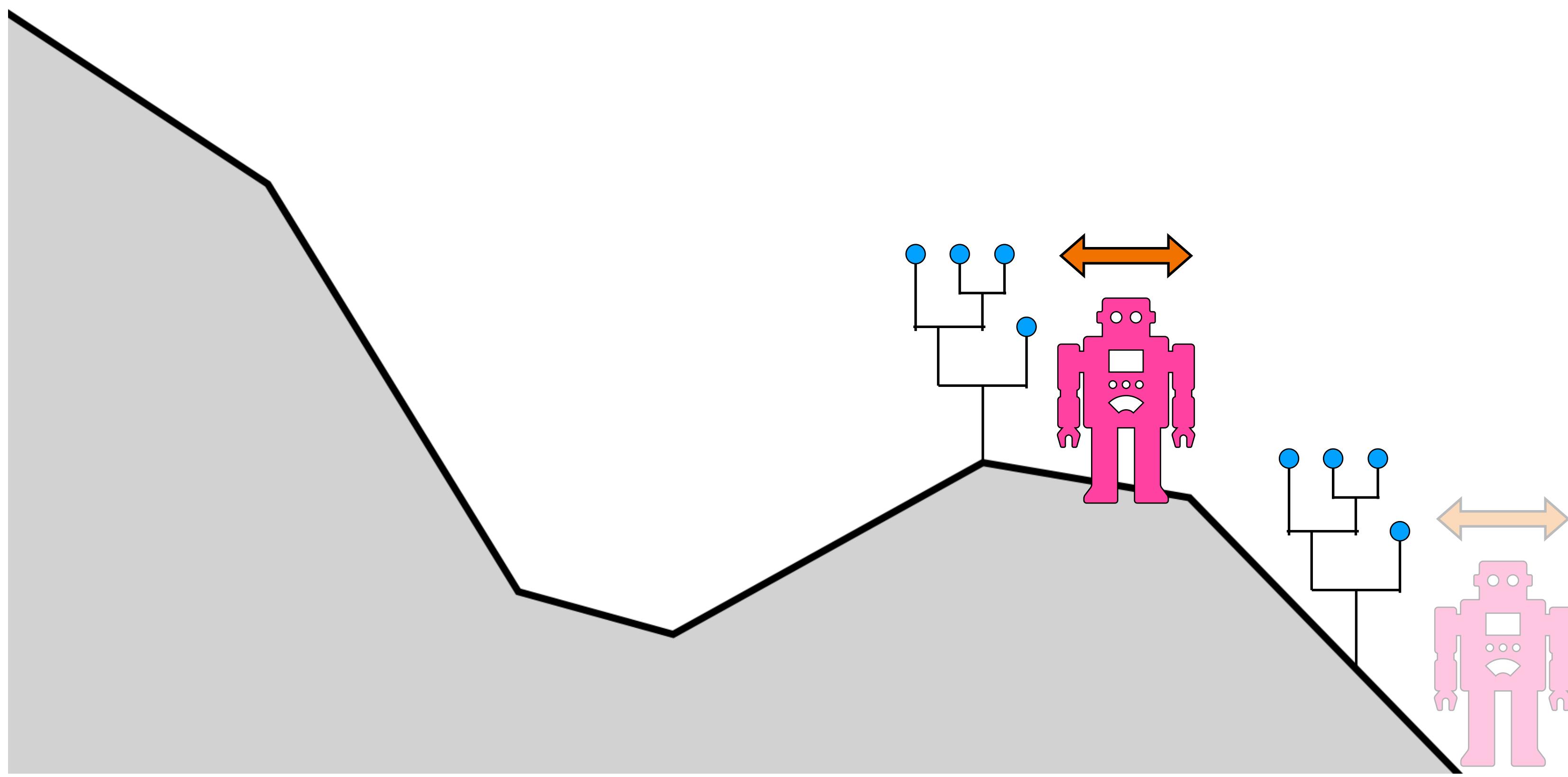
# Bayesian

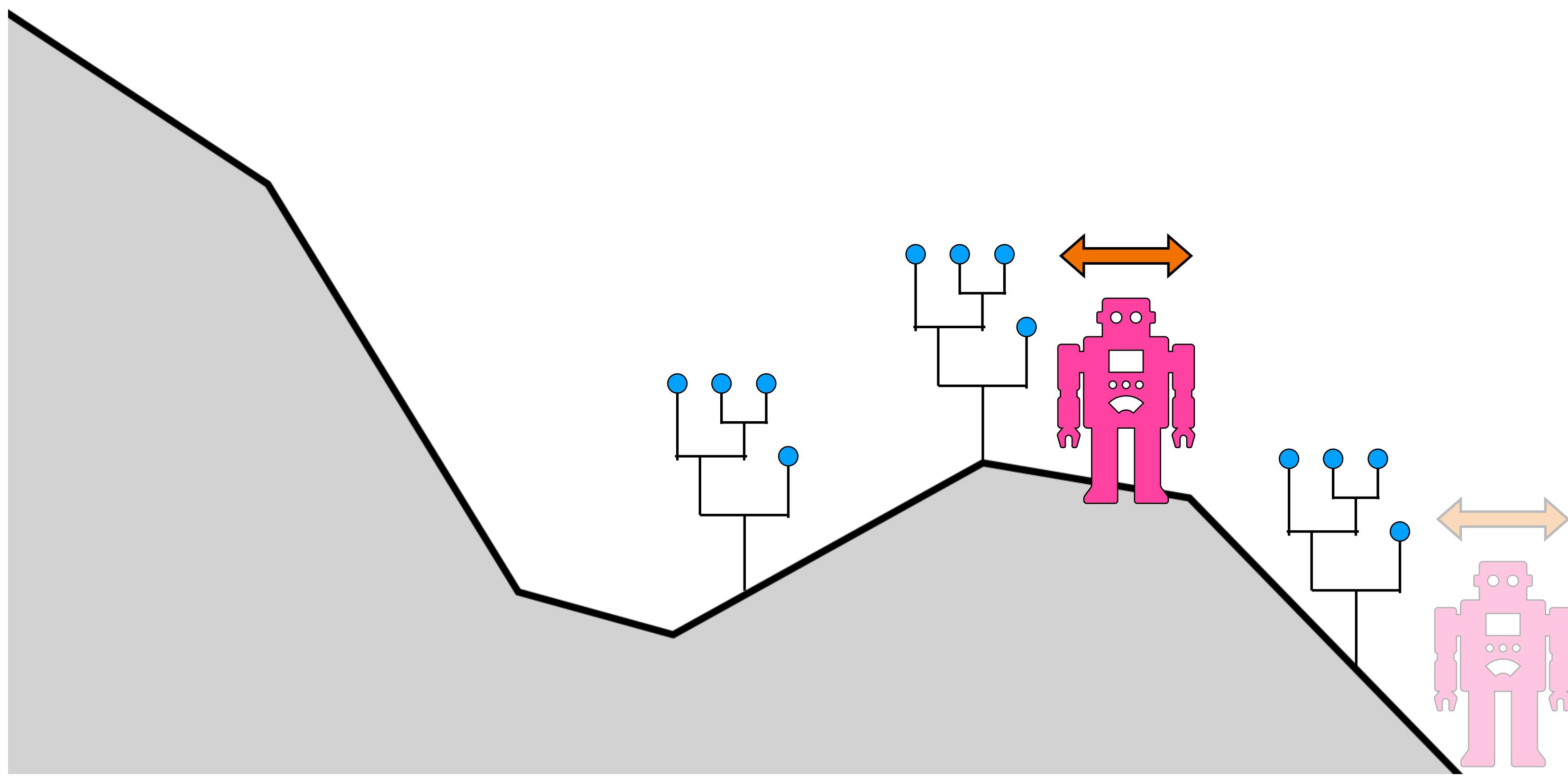


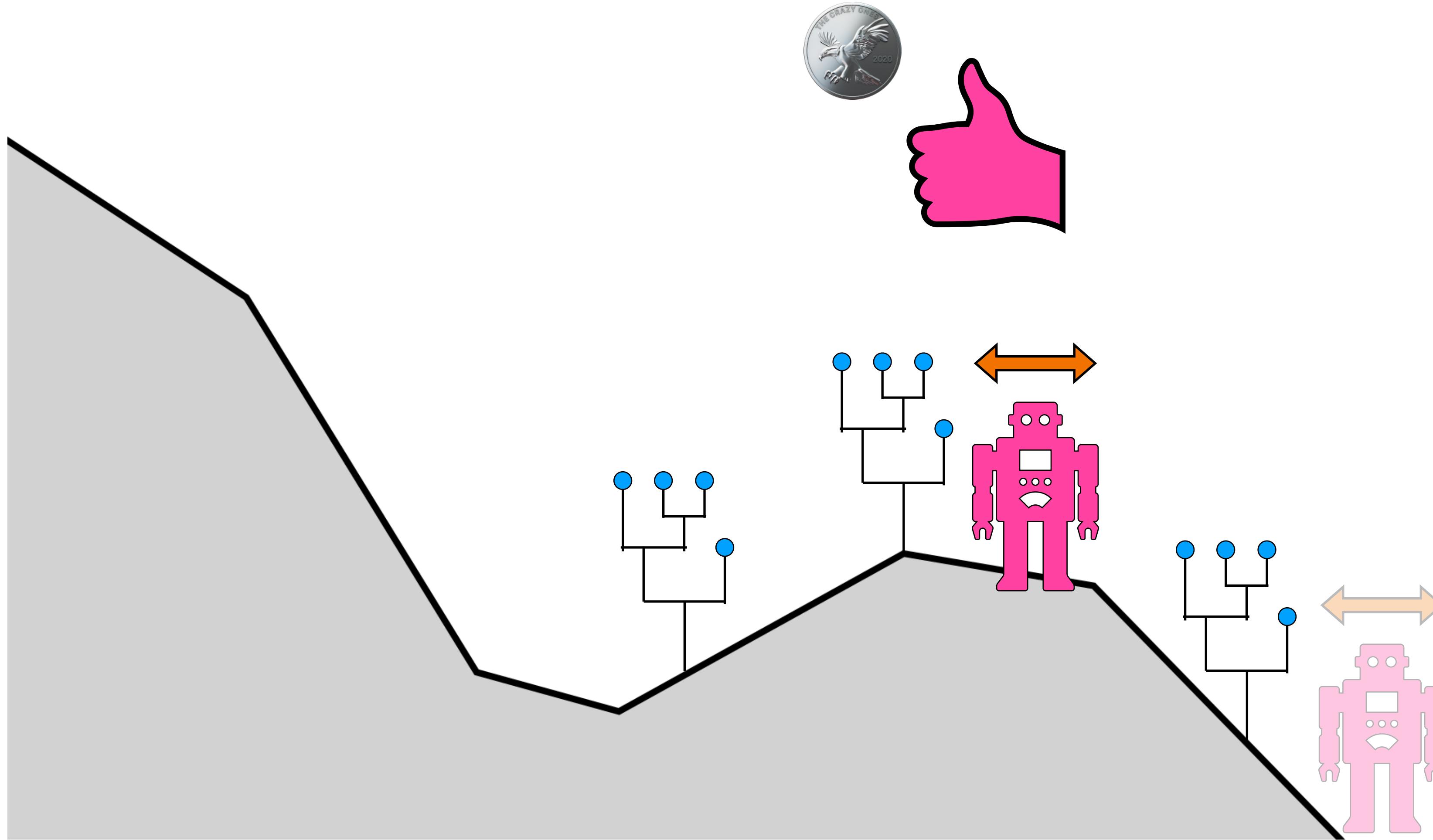


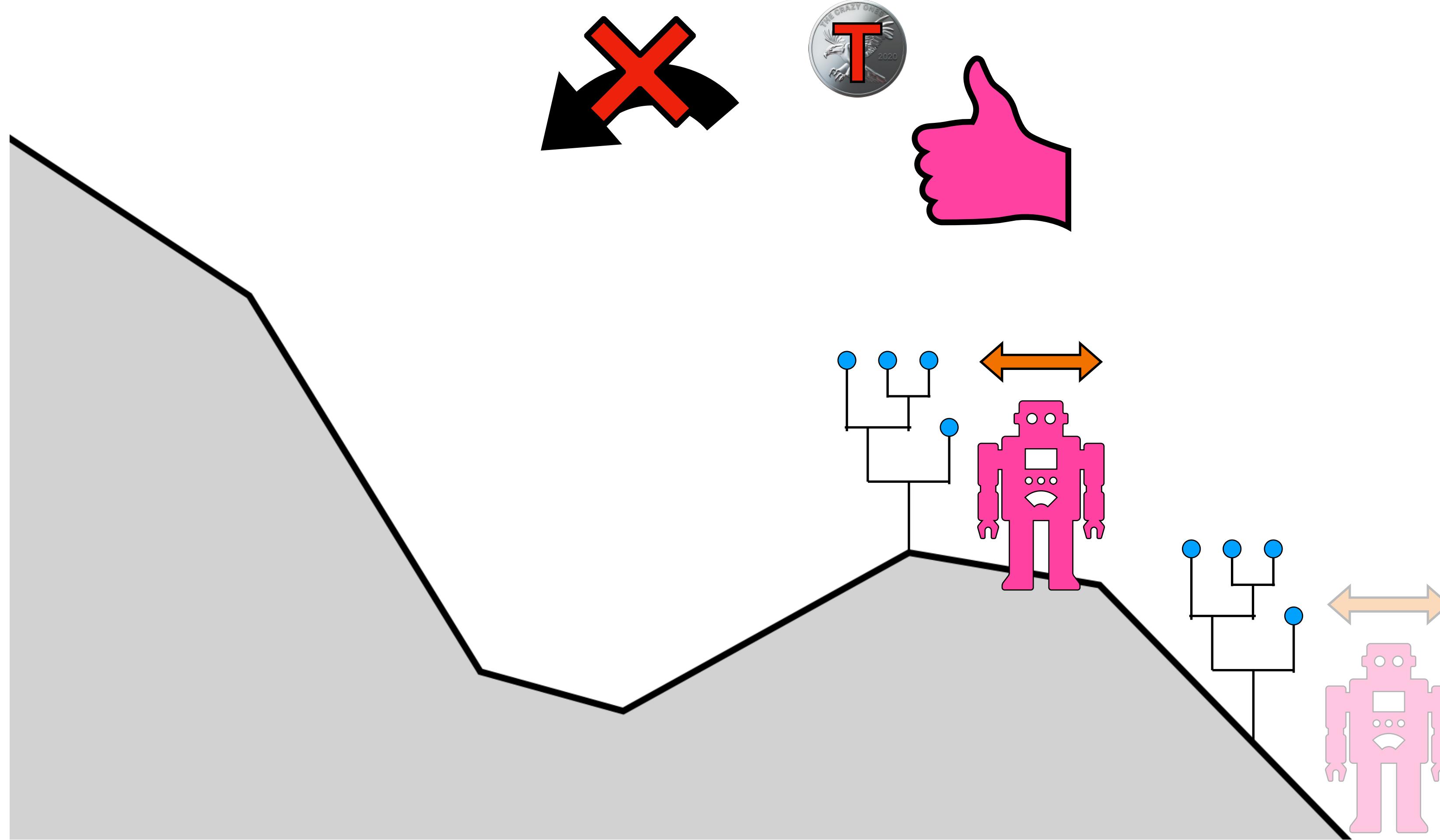


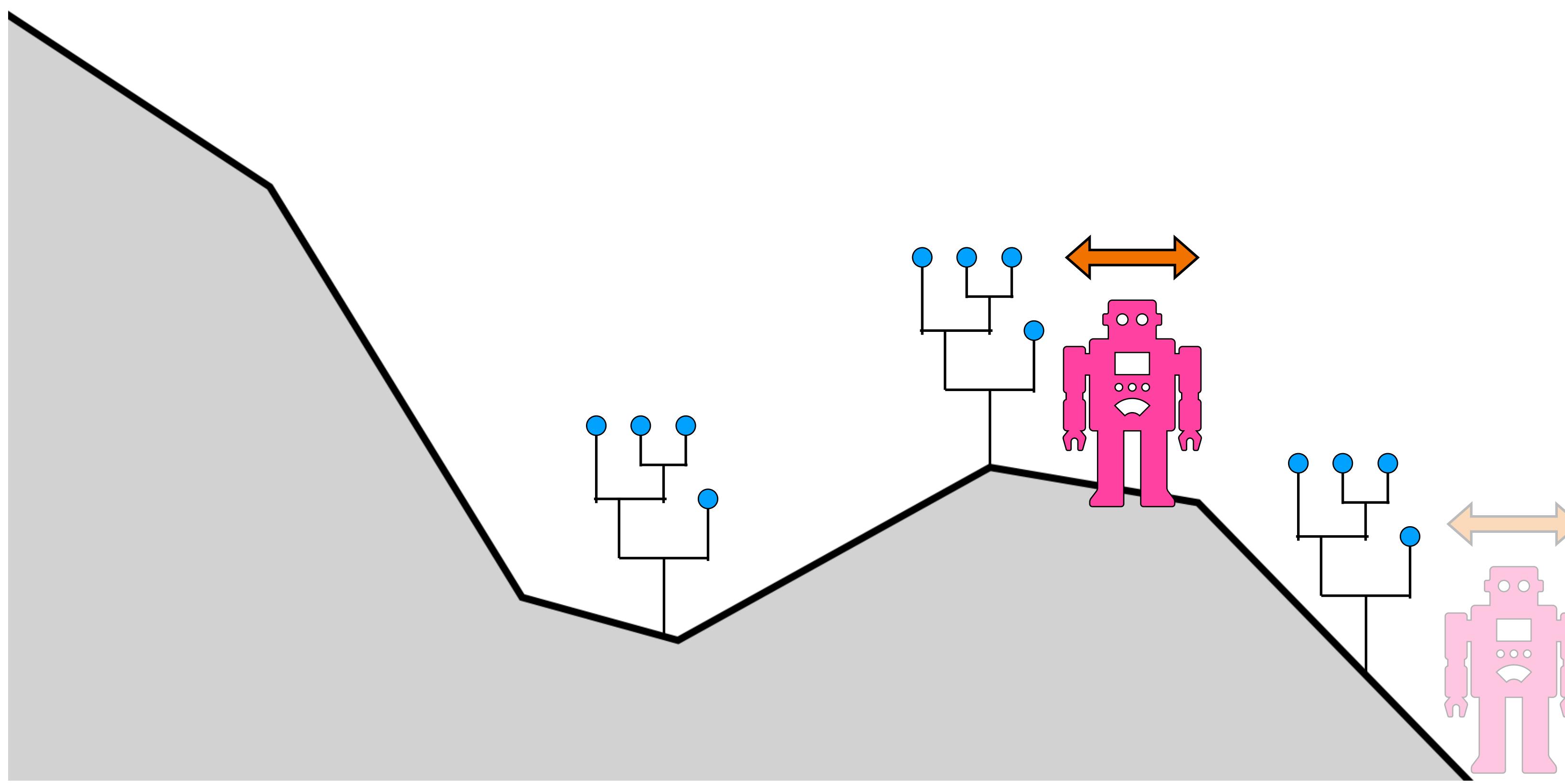


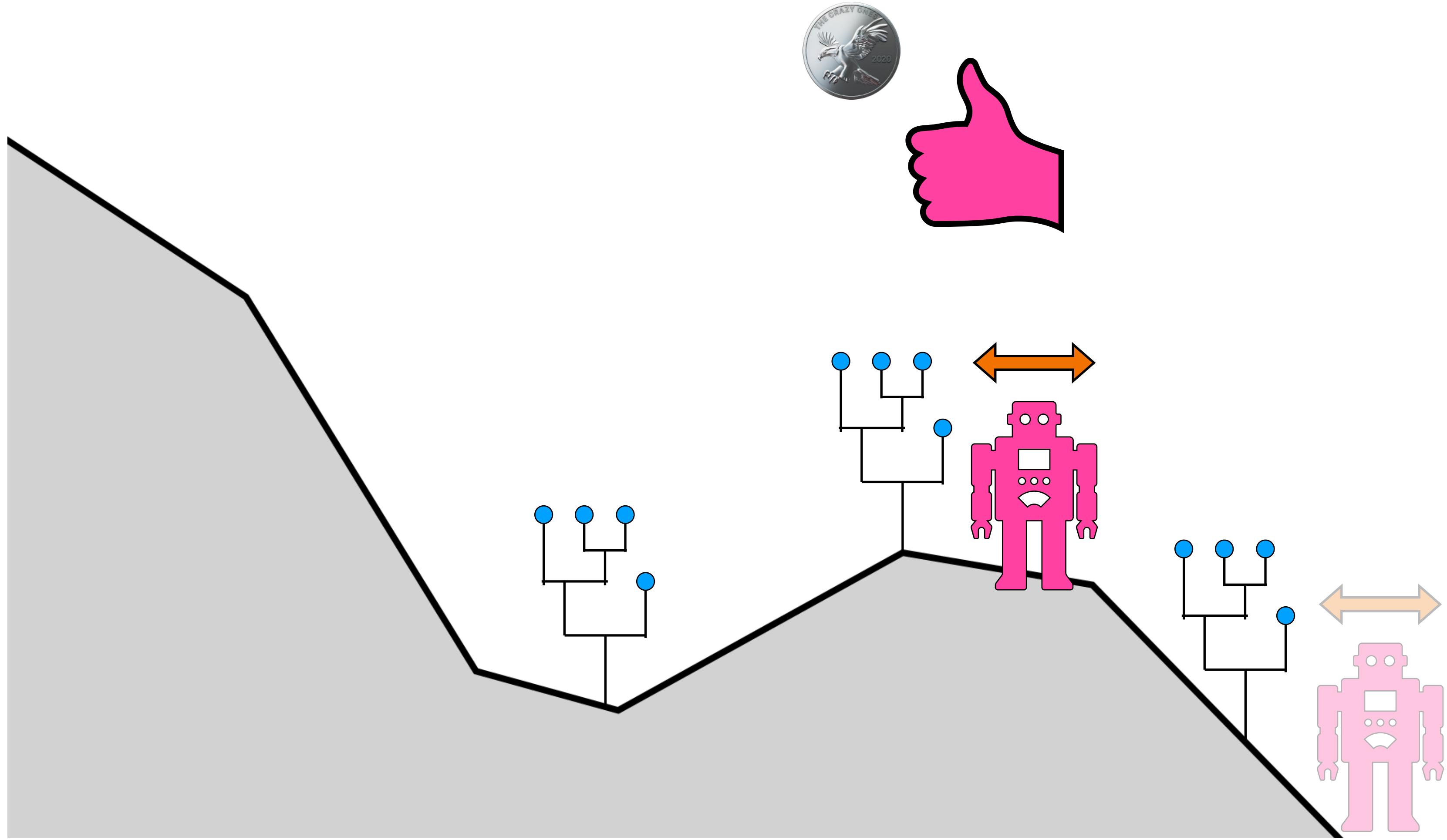


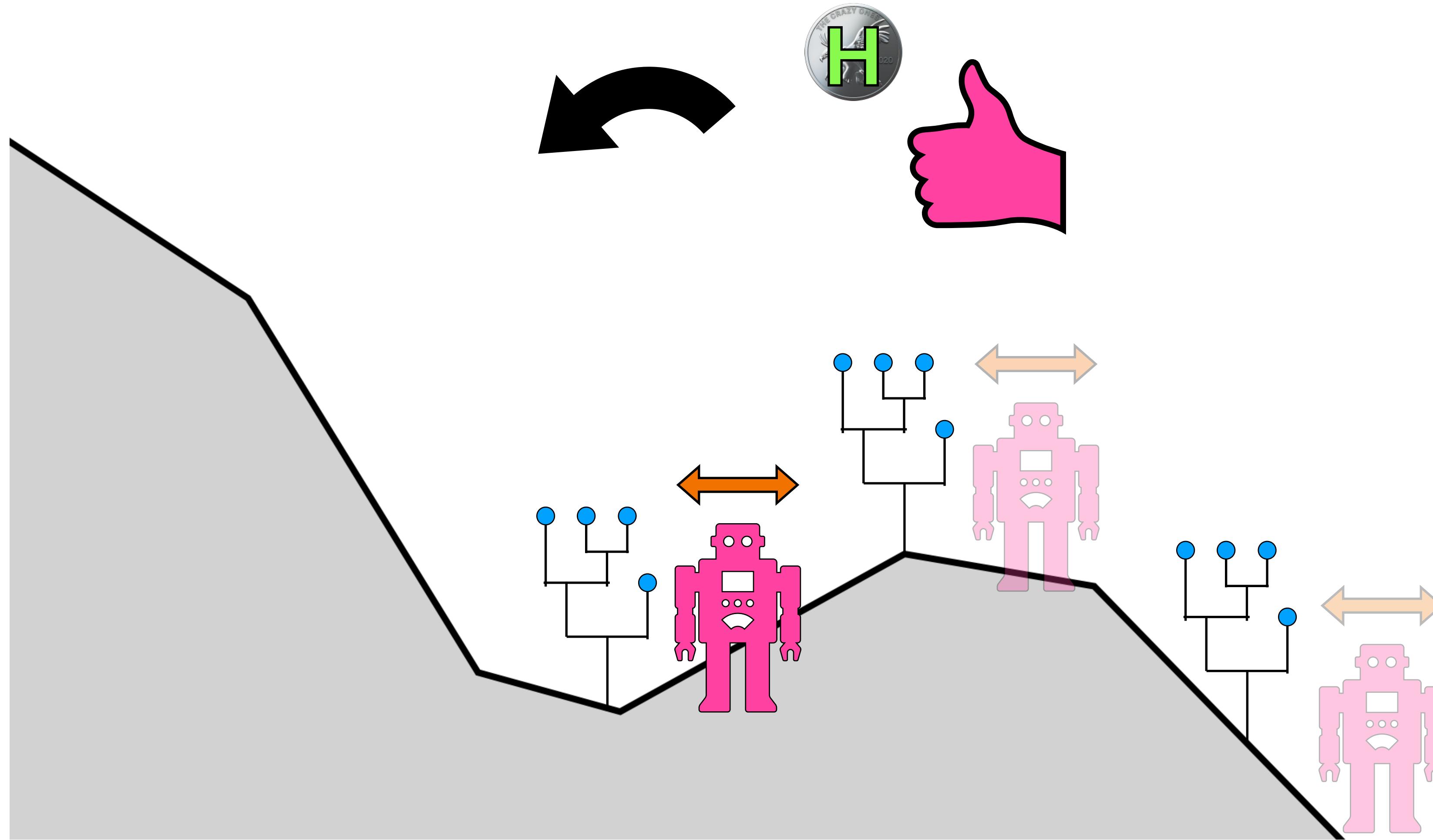


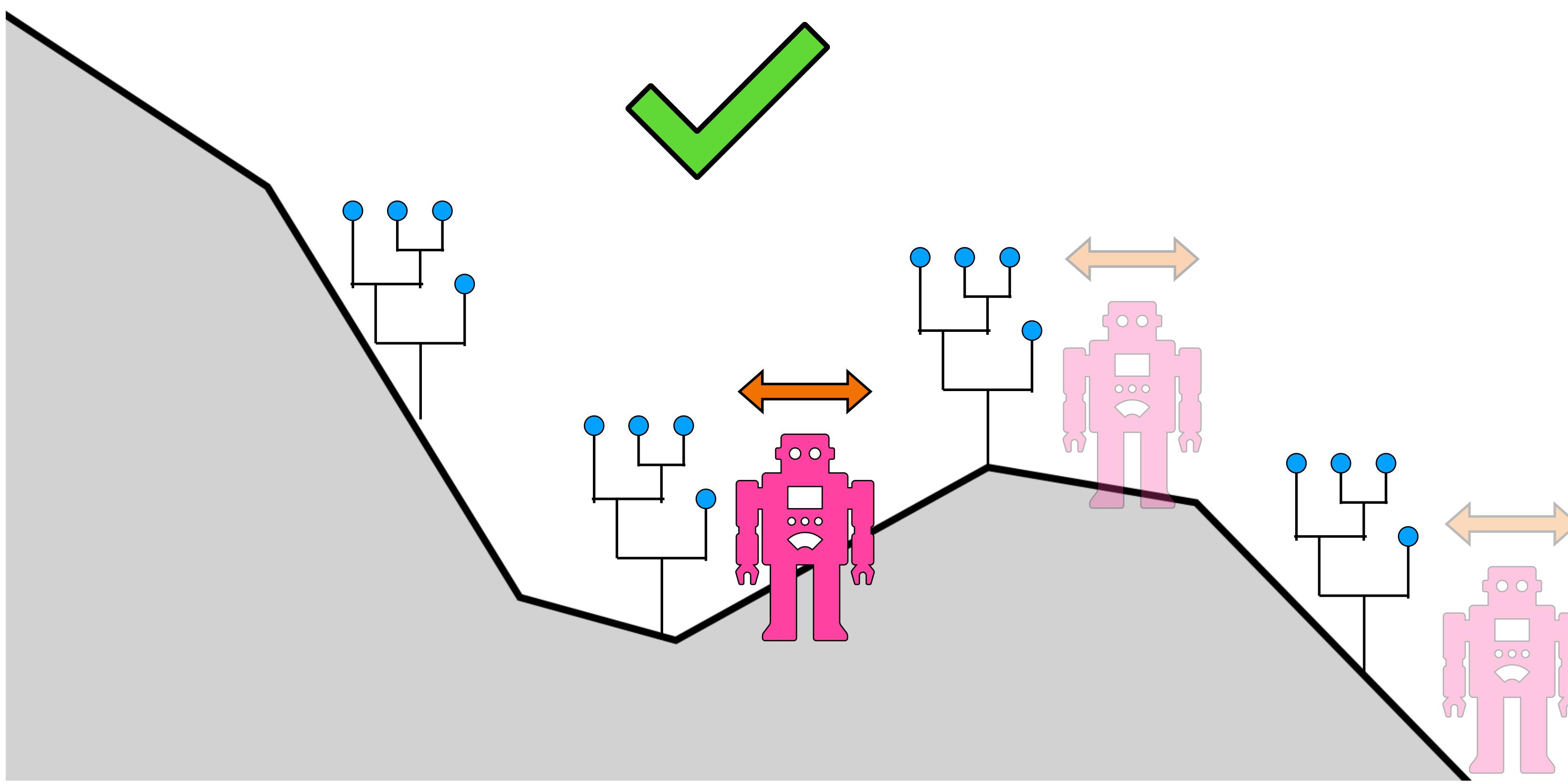


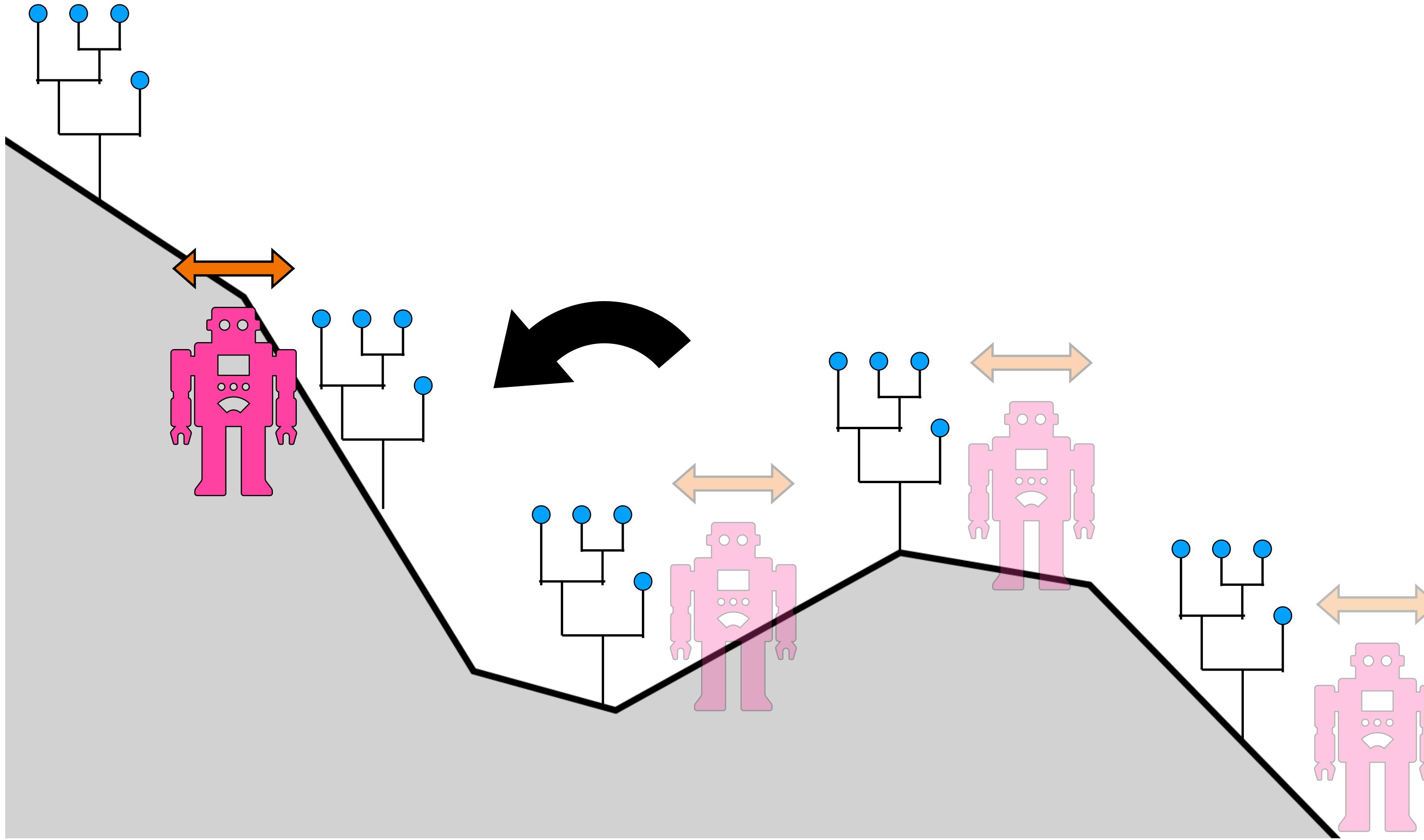


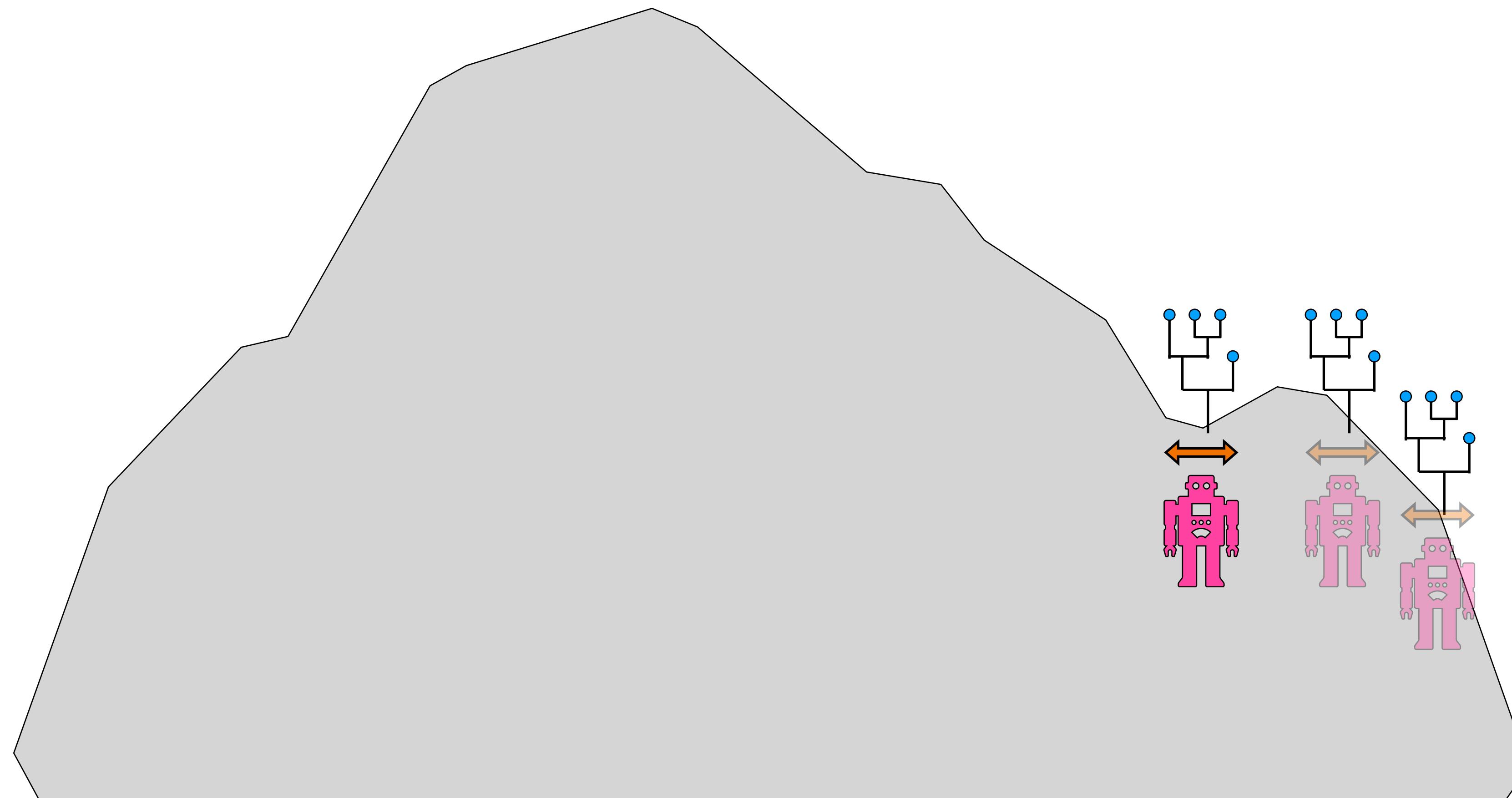


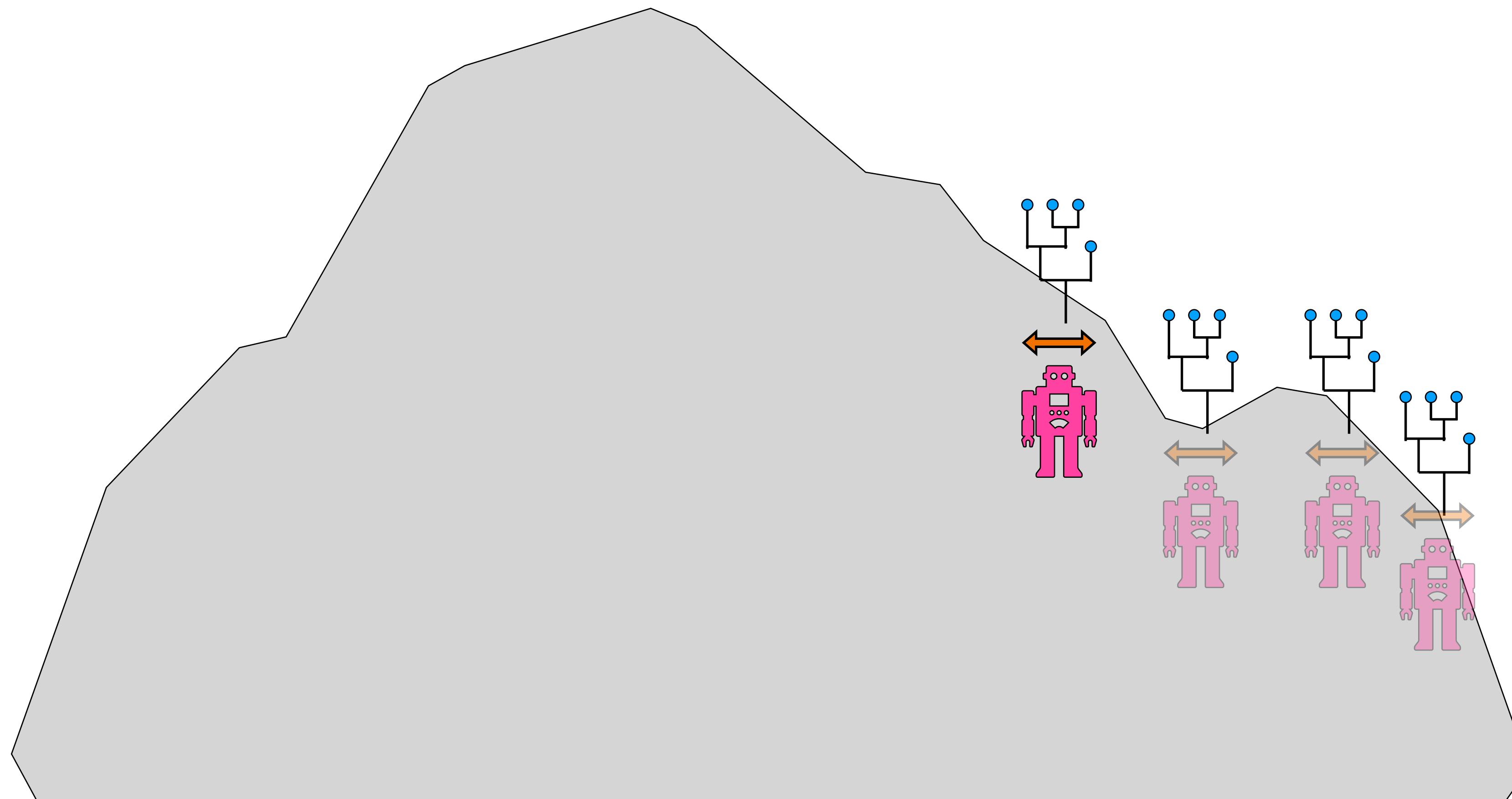


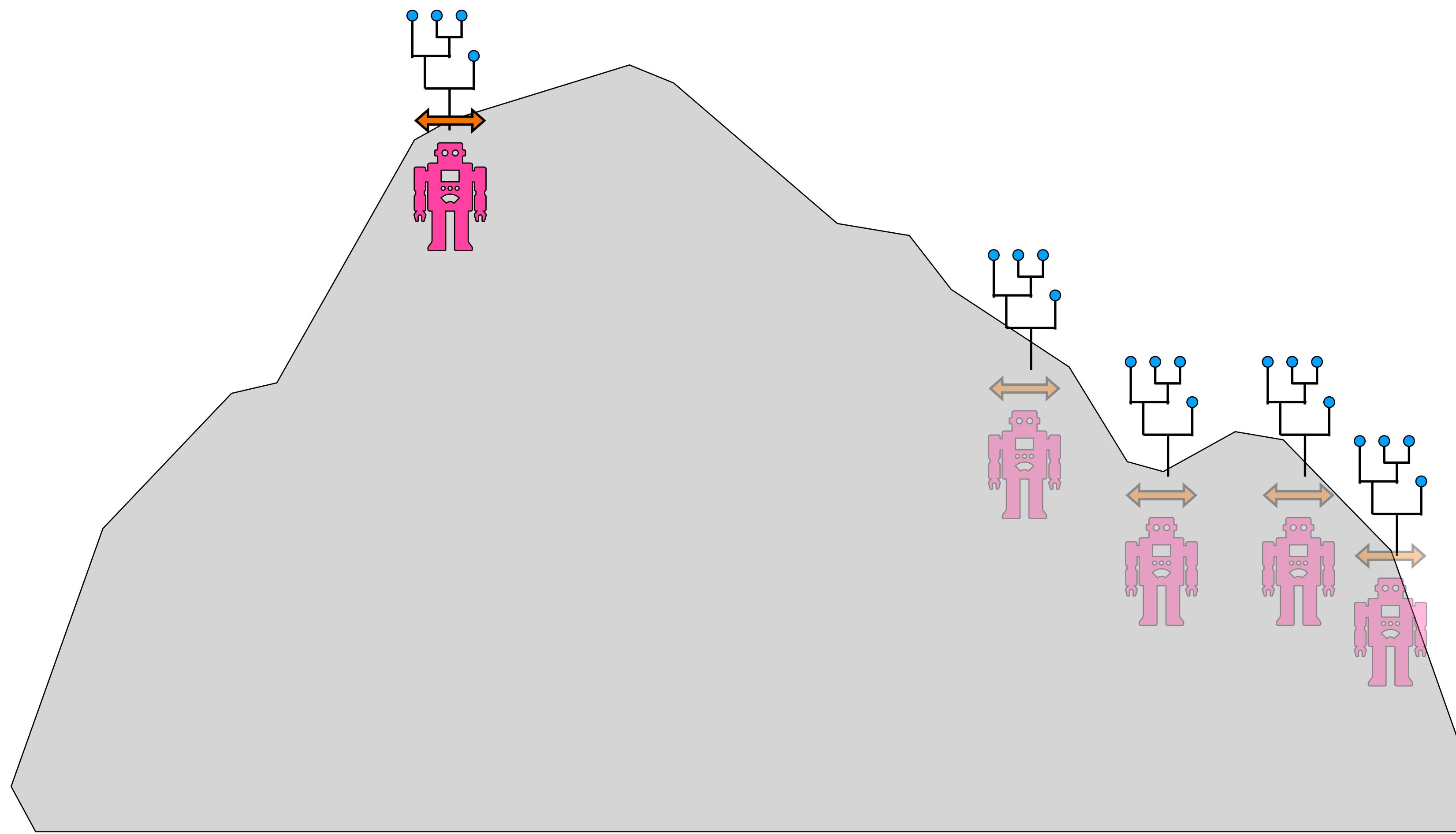


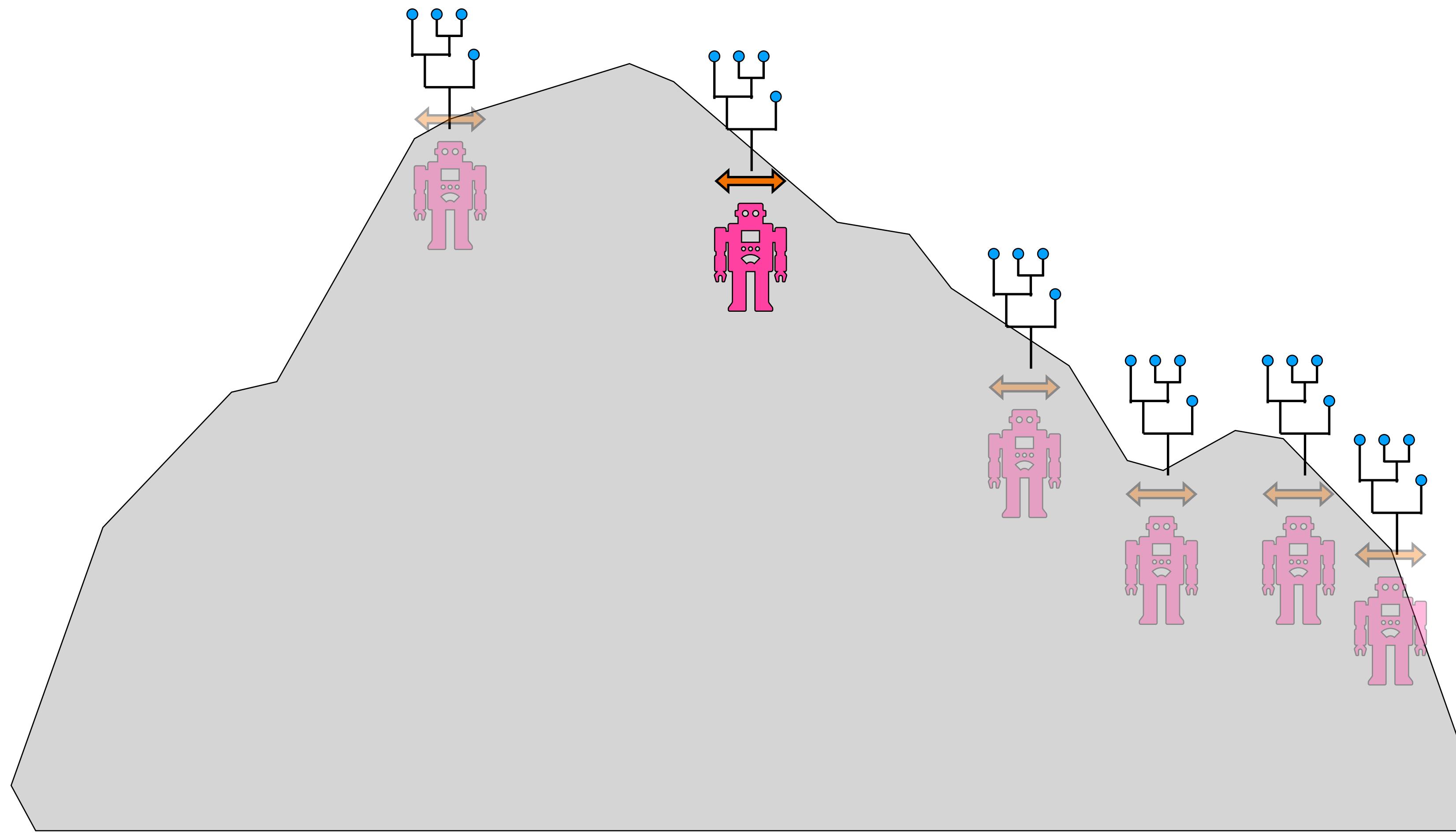


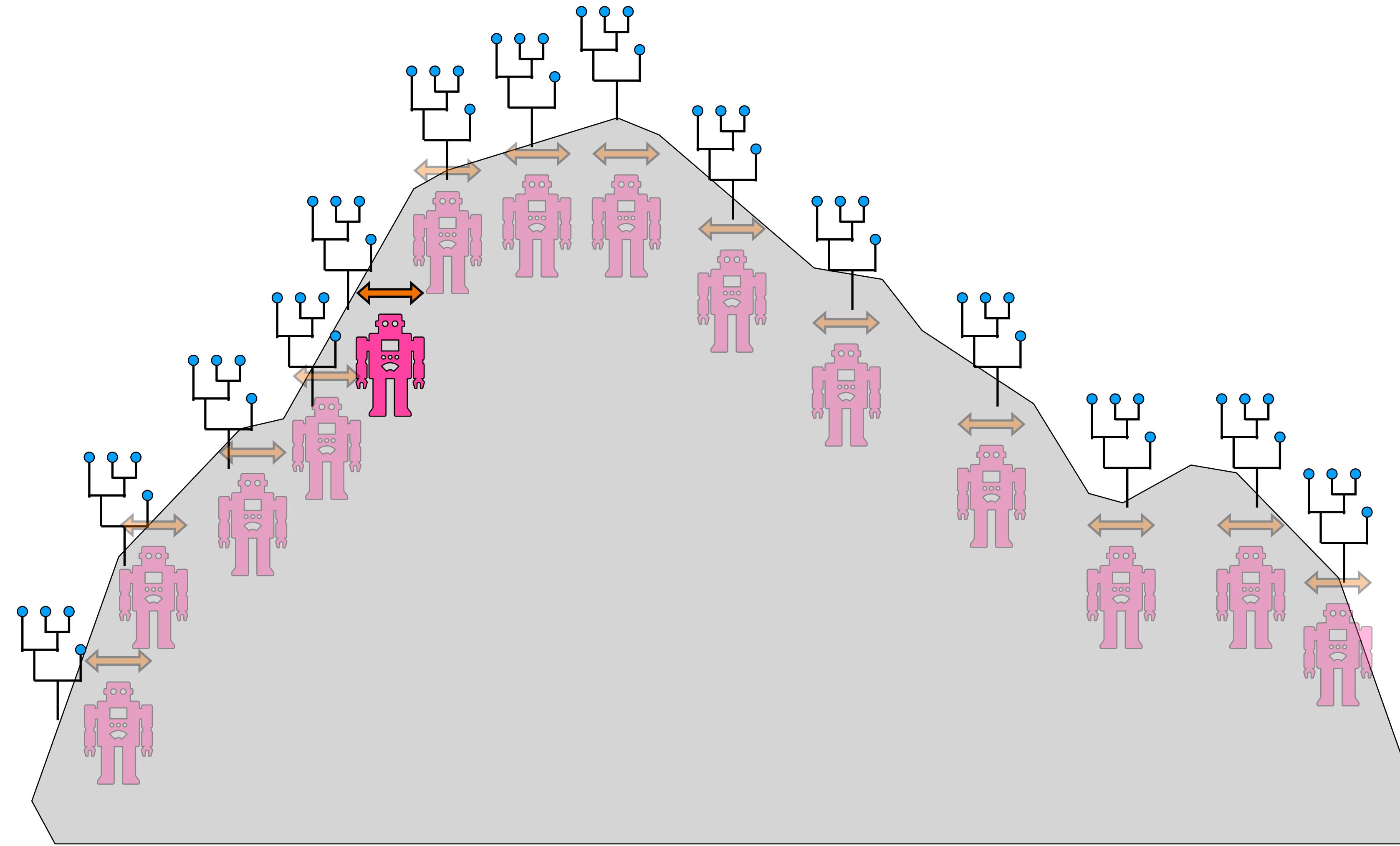






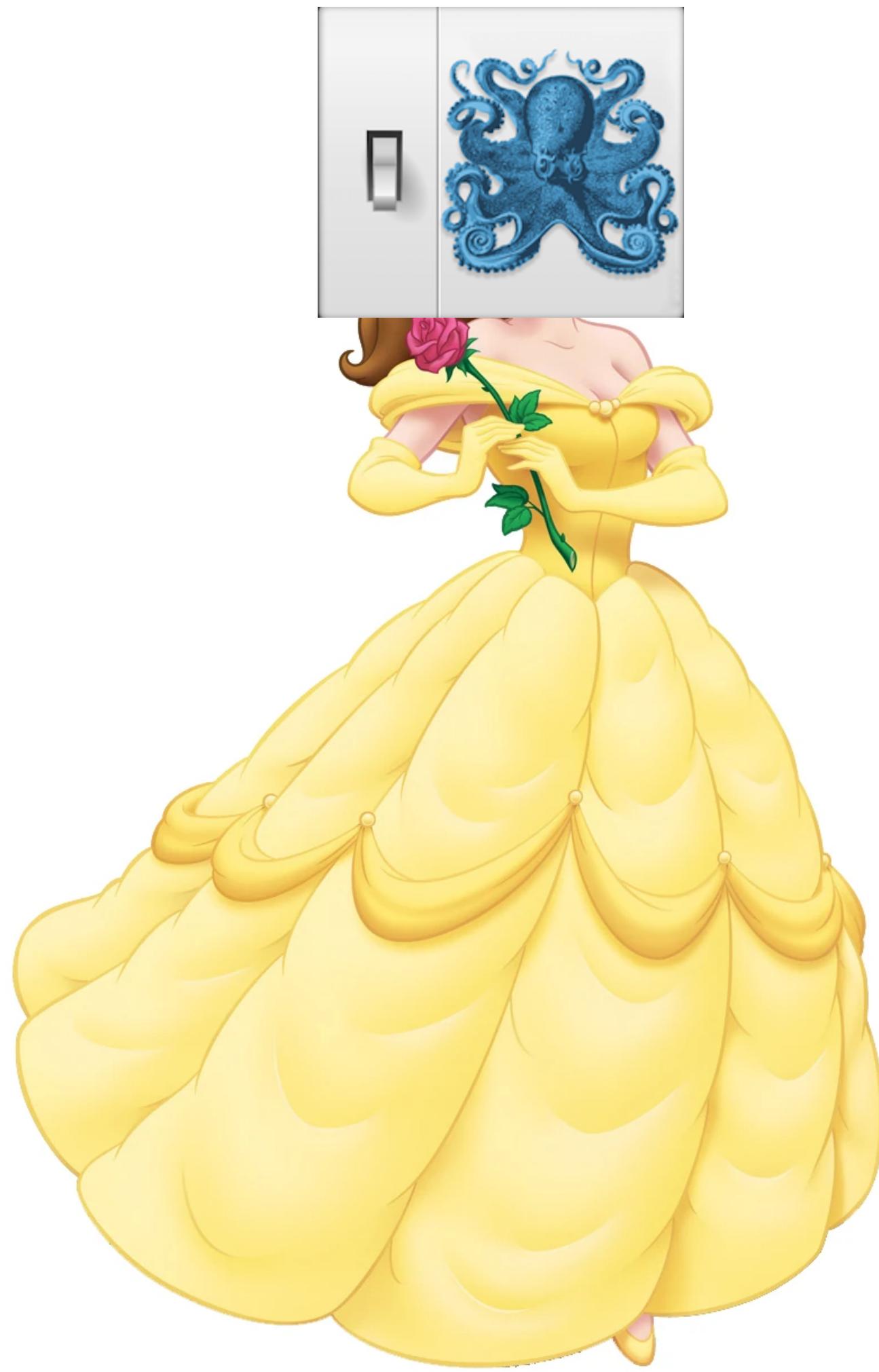






**How do I infer trees with  
BEAST?**

# BEAUi & the BEAST



Model specification (BEAUi)

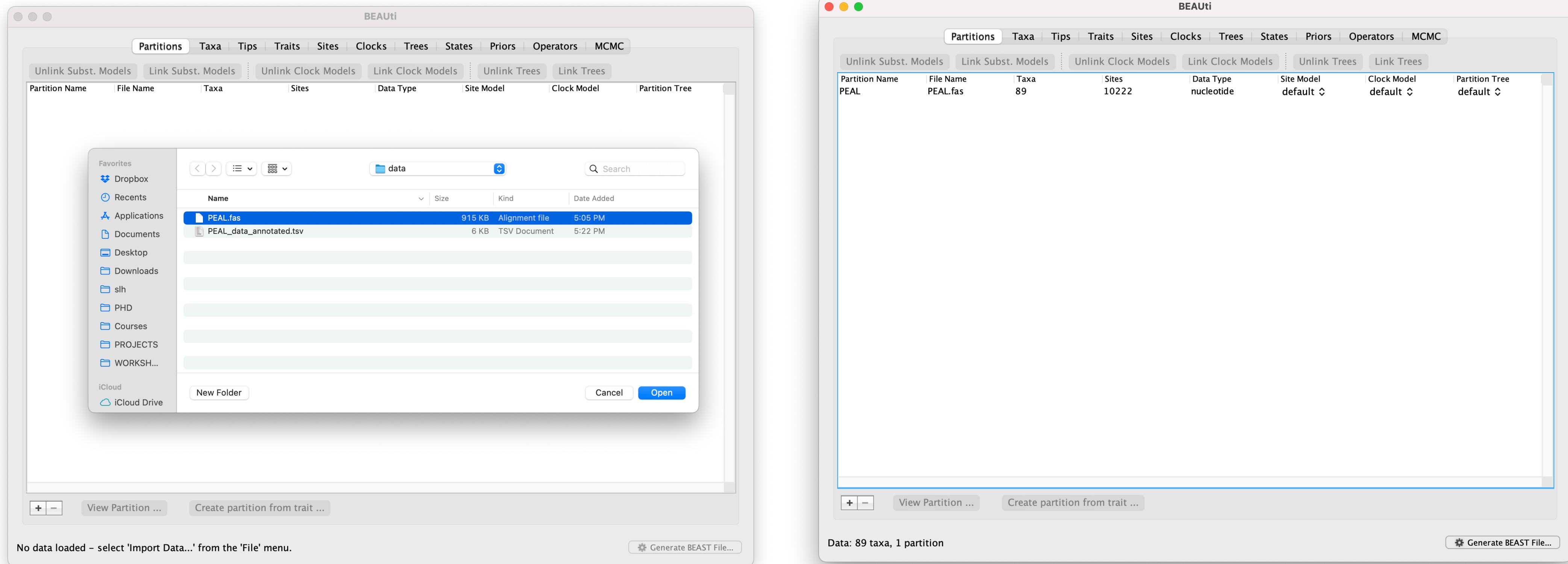


MCMC sampler (BEAST)

# Setting up the analysis in BEAUti

- Go to <> and download the data required for the analysis
- The data consists of:
  - A fasta file with 89 sequences of YFV from the 2017-18 outbreak
  - A TSV file with metadata for the 89 sequences
- Open BEAUti by typing *beauti* in the terminal

# Loading sequence data



Click on File → Import data and load the fasta file (alternatively you can drag it to the window)

# Loading location metadata

The image displays three screenshots of the BEAUti software interface, illustrating the steps to load location metadata:

- Top Left Screenshot:** Shows the "Import Traits..." dialog. A file browser window is open, showing two files: "PEAL.fas" (915 KB, Alignment file) and "PEAL\_data\_annotated.tsv" (6 KB, TSV Document). The "PEAL\_data\_annotated.tsv" file is selected.
- Middle Left Screenshot:** Shows the main BEAUti window with the "Traits" tab selected. A trait named "binary\_loc" is listed with its type set to "discrete". Below the trait list is a large table of data rows, each containing a taxon name and a value (either PEAL or Other).
- Bottom Right Screenshot:** Shows the "Partitions" tab selected. A table lists partitions: "PEAL" (File Name: PEAL.fas, Taxa: 89, Sites: 10222, Data Type: nucleotide, Site Model: default, Clock Model: default) and "binary\_loc" (File Name: PEAL\_data\_annotated.tsv, Taxa: 89, Sites: 1, Data Type: discrete, Site Model: binary\_loc, Clock Model: binary\_loc). Buttons for "View Partition ..." and "Create partition from trait..." are visible at the bottom.

Click on *Traits* → *Import Traits...* and load the metadata tsv. The click on *Create partition from trait...*

# Setting up sequence dates

The screenshot shows the BEAUti software interface with the 'Tips' tab selected. In the main window, the 'Use tip dates' checkbox is checked. Below it, the 'Parse Dates' button is highlighted. A modal dialog titled 'Parse date values for all taxa' is open, showing a list of sequence names and their corresponding dates. The dialog includes options for defining dates by order, prefix, regular expression, number, or calendar date, and a date format field set to 'yyyy-MM-dd'. At the bottom, 'Tip date sampling' is set to 'Off'.

**BEAUti**

Partitions Taxa **Tips** Traits Sites Clocks Trees States Priors Operators MCMC

Use tip dates

Parse Dates Import Dates Set Dates Clear Dates Set Uncertainty

Dates as: Years Since some time in the past Specify origin date: unable to parse date

**Parse date values for all taxa**

The date is given by a numerical field in the taxon label that is:

- Defined just by its order
- Defined by a prefix and its order
- Defined by regular expression (REGEX)
- Parse as a number
- Parse as a calendar date
- Parse calendar dates with variable precision

Order: last

Prefix: |

Add the following value to each: 1900

...unless less than: 16

...in which case add: 2000

Date format: yyyy-MM-dd

Tip date sampling: Off

Data: 89 taxa, 2 partitions

**BEAUti**

Partitions Taxa **Tips** Traits Sites Clocks Trees States Priors Operators MCMC

Use tip dates

Parse Dates Import Dates Set Dates Clear Dates Set Uncertainty

Dates as: Years Since some time in the past Specify origin date: unable to parse date

Name	Date	Uncertainty	Height
New_mosquitoes pool119 Brazil SP SaoPaulo_PEAL 2018-01-05	2018.0109589041097	0.0	0.05205479452047257
New_mosquitoes pool120 Brazil SP SaoPaulo_PEAL 2018-01-05	2018.0109589041097	0.0	0.05205479452047257
New_mosquitoes pool177 Brazil SP SaoPaulo_PEAL 2018-01-11	2018.027397260274	0.0	0.035616438356100844
New_mosquitoes pool182 Brazil SP SaoPaulo_PEAL 2018-01-11	2018.027397260274	0.0	0.035616438356100844
New_mosquitoes pool192 Brazil SP SaoPaulo_PEAL 2018-01-11	2018.027397260274	0.0	0.035616438356100844
New_mosquitoes pool24 Brazil SP SaoPaulo_PEAL 2017-12-26	2017.9835616438356	0.0	0.07945205479450124
New_mosquitoes pool66 Brazil SP SaoPaulo_PEAL 2018-01-03	2018.0054794520547	0.0	0.05753424657541473
New_mosquitoes pool81 Brazil SP SaoPaulo_PEAL 2018-01-04	2018.0082191780823	0.0	0.05479452054782996
New_mosquitoes pool95 Brazil SP SaoPaulo_PEAL 2018-01-04	2018.0082191780823	0.0	0.05479452054782996
New_NHP 2754 Brazil SP SaoPaulo_P...	2017.9835616438356	0.0	0.07945205479450124
New_NHP 2743 Brazil SP SaoPaulo_P...	2017.972602739726	0.0	0.09041095890415818
New_NHP 74213 Brazil SP SaoPaulo_P...	2017.9534246575342	0.0	0.1095890410958873
New_NHP 74300 Brazil SP SaoPaulo_P...	2017.9616438356165	0.0	0.10136986301358775
New_NHP 74301 Brazil SP SaoPaulo_P...	2017.964383561644	0.0	0.09863013698623035
New_NHP 74327 Brazil SP SaoPaulo_P...	2017.9671232876713	0.0	0.09589041095887296
New_NHP 73553 Brazil SP SaoPaulo_P...	2017.8986301369864	0.0	0.16438356164371726
New_NHP 73635 Brazil SP Mairipora...	2017.9068493150685	0.0	0.15616438356164508
New_NHP 73816 Brazil SP SaoPaulo_P...	2017.9260273972602	0.0	0.13698630136991596
New_NHP 73820 Brazil SP SaoPaulo_P...	2017.9260273972602	0.0	0.13698630136991596
New_NHP 73821 Brazil SP SaoPaulo_P...	2017.9260273972602	0.0	0.13698630136991596
New_NHP 73941 Brazil SP SaoPaulo_P...	2017.9342465753425	0.0	0.12876712328761641
New_NHP 73943 Brazil SP SaoPaulo_P...	2017.9342465753425	0.0	0.12876712328761641
New_NHP 73946 Brazil SP Mairipora...	2017.9342465753425	0.0	0.12876712328761641
New_NHP 74342 Brazil SP SaoPaulo_P...	2017.986301369863	0.0	0.0767123287614384
New_NHP 74718 Brazil SP SaoPaulo_P...	2018.0054794520547	0.0	0.05753424657541473
New_NHP 74747 Brazil SP SaoPaulo_P...	2018.0082191780823	0.0	0.05479452054782996
New_NHP 74329 Brazil SP SaoPaulo_P...	2017.9671232876713	0.0	0.09589041095887296
New_NHP 74347 Brazil SP SaoPaulo_P...	2017.9698630136986	0.0	0.09315068493151557
New_NHP 74348 Brazil SP SaoPaulo_P...	2017.9698630136986	0.0	0.09315068493151557
New_NHP 74405 Brazil SP SaoPaulo_P...	2017.972602739726	0.0	0.09011005800115919

Tip date sampling: Off Apply to taxon set: All taxa

Data: 89 taxa, 2 partitions

Generate BEAST File...

Click on *Tips* → *Parse Dates* and set up the dates by parsing them from the sequence names

# Setting up substitution models

The image displays two side-by-side screenshots of the BEAUti software interface, version 1.8.1, used for phylogenetic analysis.

**Left Screenshot (Nucleotide Substitution Model):**

- Substitution Model:** default  
binary\_loc
- Substitution Model:** HKY
- Base frequencies:** Estimated
- Site Heterogeneity Model:** Gamma
- Number of Gamma Categories:** 4
- Partition into codon positions:** Off
- Link/Unlink parameters:**
  - Unlink substitution rate parameters across codon positions
  - Unlink rate heterogeneity model across codon positions
  - Unlink base frequencies across codon positions
- Buttons:** Use Yang96 model, Use SRD06 model

**Bottom Buttons:** Clone Settings..., Generate BEAST File...

**Right Screenshot (Discrete Traits Substitution Model):**

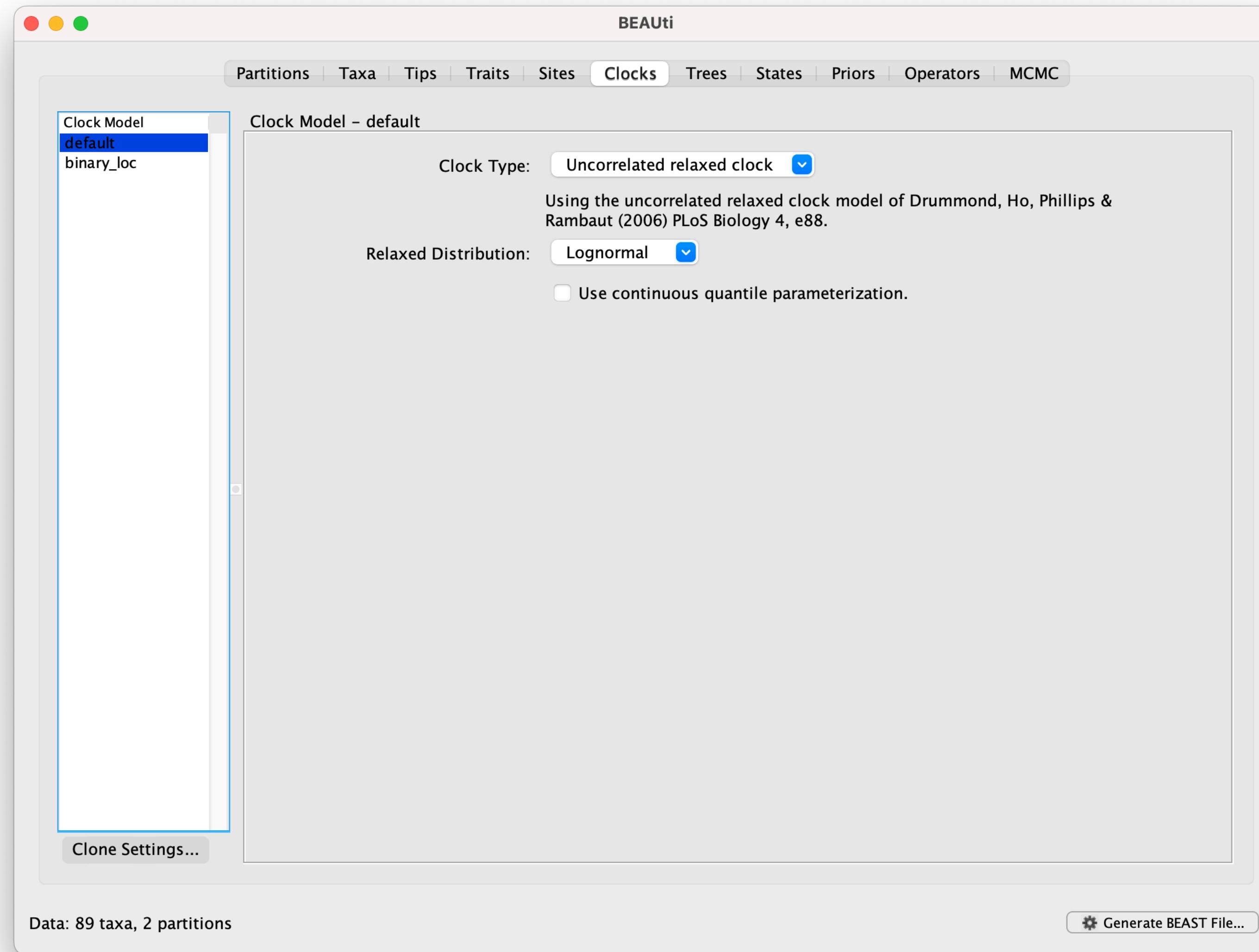
- Substitution Model:** default  
binary\_loc
- Discrete Trait Substitution Model:** Asymmetric substitution model
- Checkboxes:** Infer social network with BSSVS, Setup GLM

**Bottom Buttons:** Clone Settings..., Generate BEAST File...

**Common Interface Elements:** The top navigation bar includes tabs for Partitions, Taxa, Tips, Traits, Sites, Clocks, Trees, States, Priors, Operators, and MCMC. The status bar at the bottom of each window indicates "Data: 89 taxa, 2 partitions".

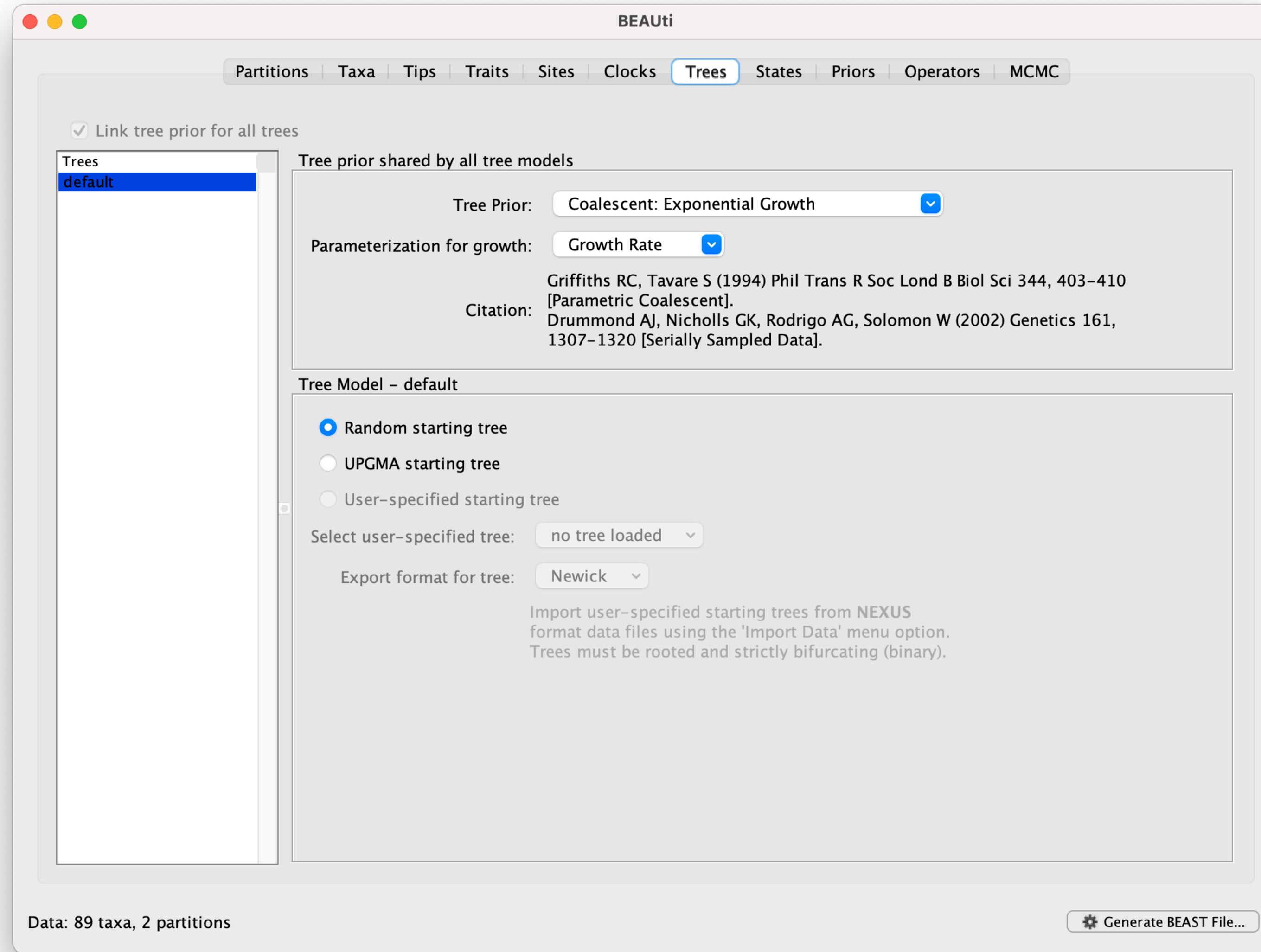
Click on the **Sites** tab and set up the nucleotide substitution model as *HKY with Gamma heterogeneity and 4 categories*. Set up an *Asymmetric substitution model* for the location trait

# Setting up the clock model



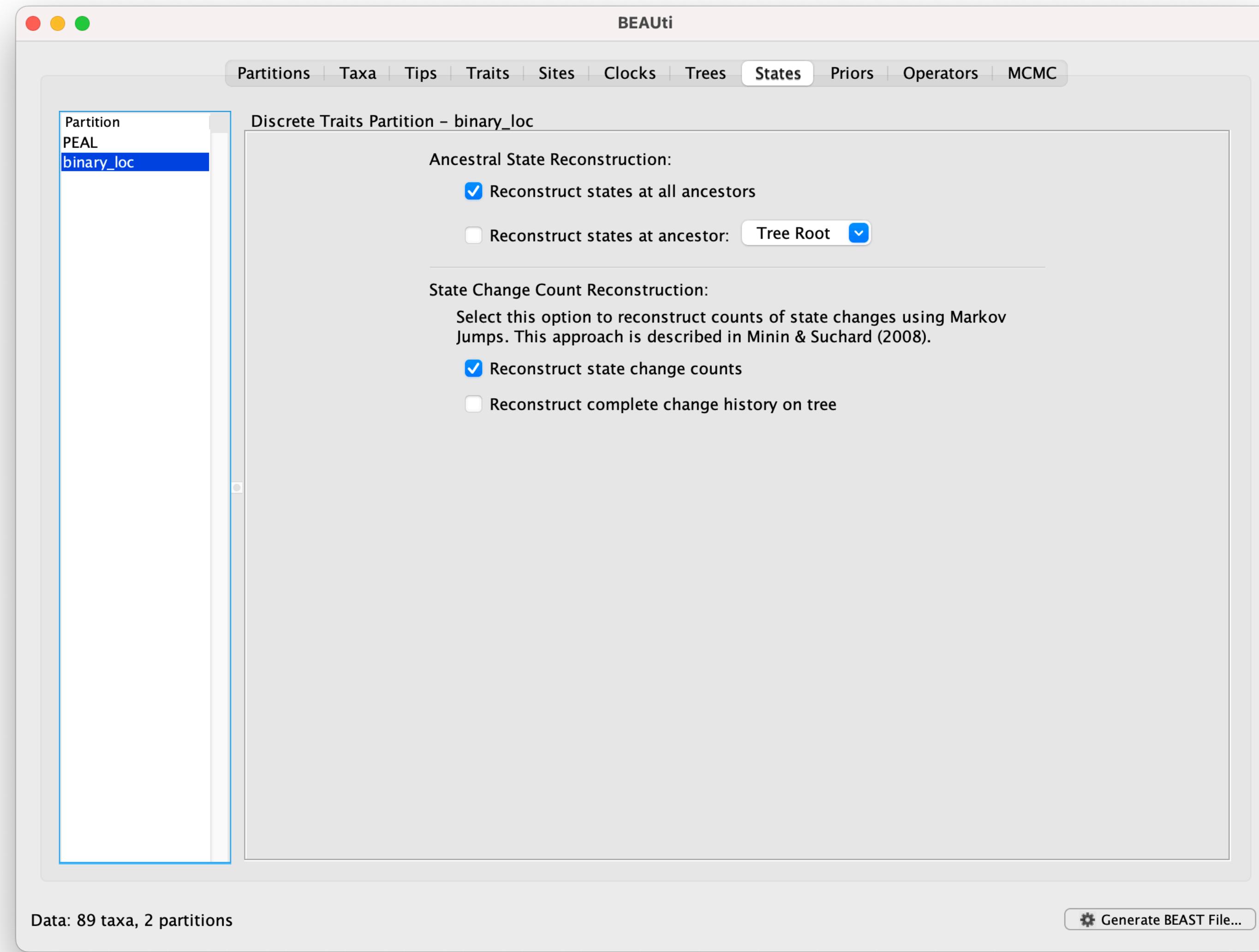
Click on the *Clocks* tab and set up the clock type as *Uncorrelated relaxed clock* with a *Lognormal* distribution. Leave the default *Strict clock* for the location partition.

# Setting up the coalescent model



Click on the *Trees* tab and set up the tree prior as *Coalescent: Exponential Growth* with a *Growth Rate* parameterization.

# Setting up the ancestral state reconstruction



Click on the *States* tab and on the discrete trait partition, click on *Reconstruct state change counts*

# Setting up the priors

The screenshot shows the BEAUti software interface with the title bar "BEAUti". Below the title bar is a menu bar with tabs: Partitions, Taxa, Tips, Traits, Sites, Clocks, Trees, States, **Priors**, Operators, and MCMC. The "Priors" tab is currently selected. A checkbox labeled "Use classic priors/operators" is unchecked. Below this is a table listing parameters and their current prior settings:

Parameter	Prior	Bound	Description
kappa	* LogNormal [1, 1.25], initial=2	[0, ∞]	HKY transition-transversion parameter
frequencies	* Dirichlet [1,1]	[0, ∞]	base frequencies
default.ulcl.mean	* Approx. Reference Prior, init...	[0, ∞]	uncorrelated lognormal relaxed clock mean
default.ulcl.stdev	* Exponential [0.333333], init...	[0, ∞]	uncorrelated lognormal relaxed clock stdev
binary_loc.clock.rate	* Approx. Reference Prior, init...	[0, ∞]	substitution rate
treeModel.rootHeight	* Using Tree Prior in [0.29315...	[0.293151..., ∞)	root height of the tree
exponential.popSize	LogNormal [1, 5], initial=1	[0, ∞]	coalescent population size parameter
exponential.growthRate	Laplace [0, 100], initial=0	(-∞, ∞)	coalescent growth rate parameter
binary_loc.frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	discrete state frequencies
binary_loc.rates	* Gamma [1, 1], initial=1	[0, ∞]	discrete trait instantaneous transition rates
binary_loc.root.frequencies	* Uniform [0, 1], initial=0.25	[0, 1]	discrete state root frequencies

At the bottom of the window, there are three buttons: "Link parameters together", "Link parameters into a hierarchical model", and "Unlink parameters". A note below these buttons states: "\* Marked parameters currently have a default prior distribution. You should check that these are appropriate." At the very bottom left is the text "Data: 89 taxa, 2 partitions" and at the bottom right is a button labeled "Generate BEAST File...".

Click on the *Priors* tab and leave the default priors for all parameters except for

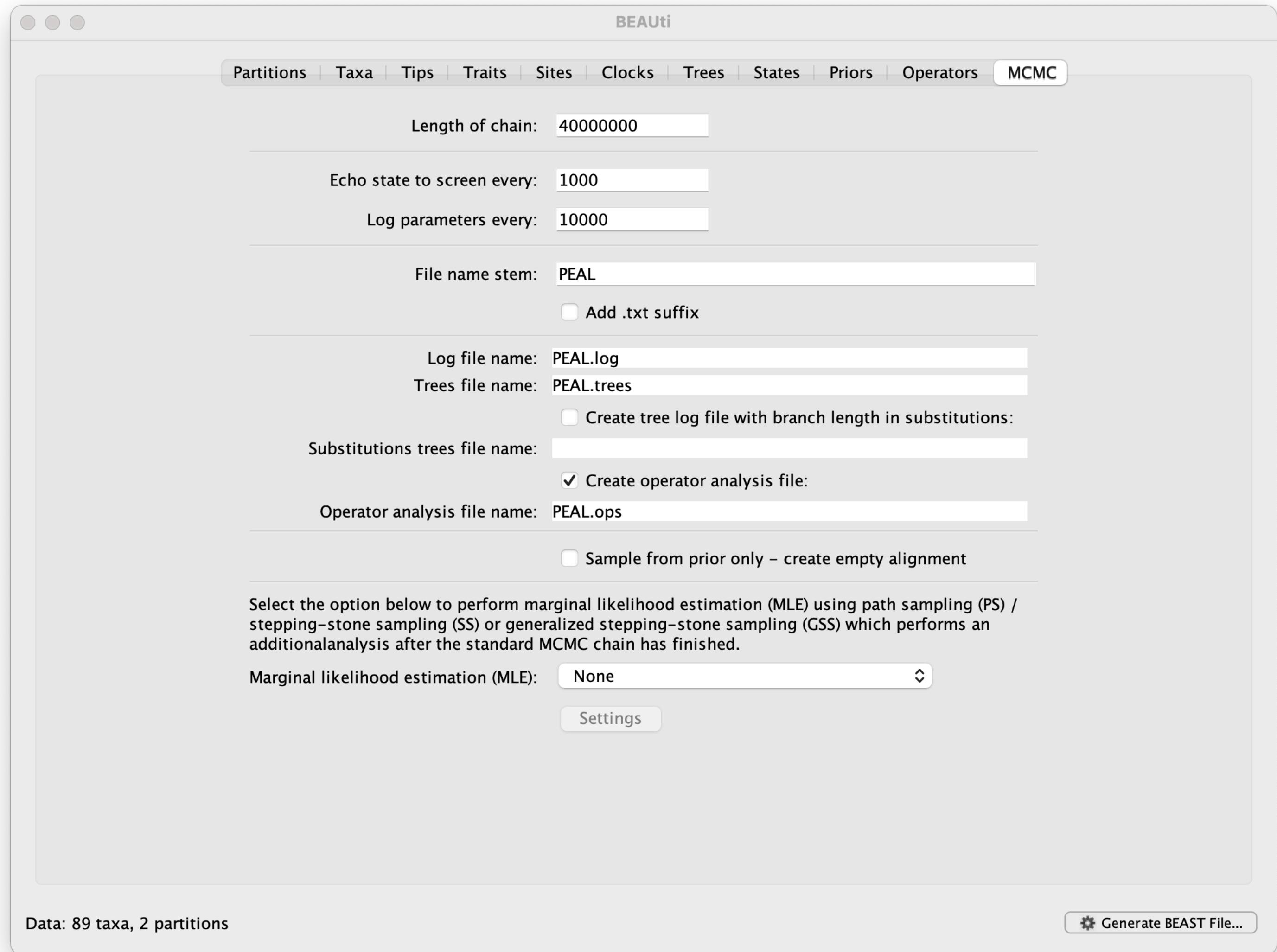
- *exponential.popSize*:  
LogNormal [1,5]
- *Exponential.growthRate*:  
Laplace [0,100]

# Setting up the operators

The screenshot shows the BEAUti software interface with the title bar "BEAUti". Below the title bar is a menu bar with tabs: Partitions, Taxa, Tips, Traits, Sites, Clocks, Trees, States, Priors, Operators (which is currently selected), and MCMC. A sub-menu for "Operator mix" is open, showing "classic operator mix" as the selected option. Below this, there is a table titled "In use | Operates on" with columns: In use, Operates on, Type, Tuning, Weight, and Description. The table lists various operators, many of which are checked (indicated by blue checkmarks). The operators include: Multiple (adaptiveMultivariate, weight 30.0), kappa (scale, weight 1.0), frequencies (deltaExchange, weight 1.0), default.ulcl.mean (scale, weight 3.0), default.ulcl.stdev (scale, weight 3.0), default.UCLD mean and heights (upDown, weight 3.0), default.branchRates.categories (swap, weight 10.0), default.branchRates.categories (integerUniform, weight 10.0), binary\_loc.clock.rate (scale, weight 3.0), binary\_loc.Substitution rate an... (upDown, weight 3.0), Tree (subtreeSlide, weight 30.0), Tree (narrowExchange, weight 30.0), Tree (wideExchange, weight 3.0), Tree (wilsonBalding, weight 3.0), treeModel.rootHeight (scale, weight 3.0), Internal node heights (uniform, weight 30.0), Tree (subtreeLeap, weight 89.0), Tree (subtreeJump, weight 8.9), exponential.popSize (scale, weight 3.0), exponential.growthRate (randomWalk, weight 3.0), binary\_loc.rates (scaleIndependently, weight 15.0), and binary\_loc.root.frequencies (deltaExchange, weight 1.0). At the bottom left of the window, it says "Data: 89 taxa, 2 partitions". At the bottom right, there is a button labeled "Generate BEAST File...".

On the *Operators* tab,  
leave all default values.

# Setting up the MCMC



Click on the *MCMC* tab and set the *Length of chain* to 40 million and *Log parameters every* 10,000 states. Generate an XML file for BEAST using the *Generate BEAST File* button on the bottom right.

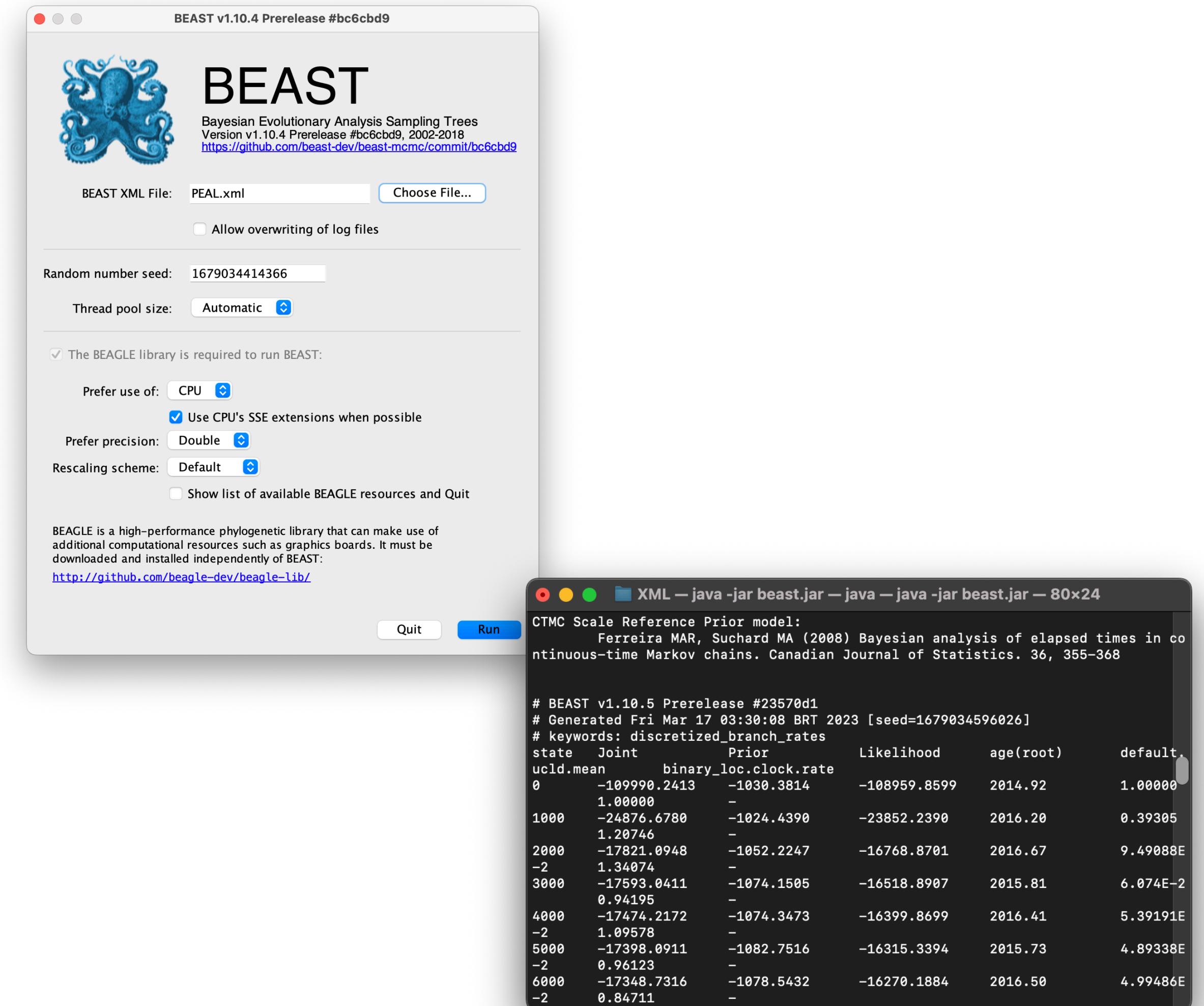
# Setting up the Markov Jump counts

```
<!-- START Ancestral state |  
reconstruction -->  
<parameter id="binary_loc.count" value=" 0.0 1.0 1.0 0.0"/>  
<parameter id="binary_loc.OthertoPEAL" value=" 0.0 1.0 0.0 0.0"/>  
<parameter id="binary_loc.PEALtoOther" value=" 0.0 0.0 1.0 0.0"/>  
  
<!-- END Ancestral state  
reconstruction -->
```

By default, *Reconstruct state change counts* in BEAUti will only create a parameter that logs all jumps from any two locations. To keep track of jumps into and out of Horto separately, open the generated XML file in a text editor and add the *OthertoPEAL* and *PEALtoOther* parameters below the parameter that logs all counts, in the Ancestral state reconstruction section of the *markovJumpsTreeLikelihood* block. The zeros and ones in the value field represent the transitions being counted represented with a flattened matrix with rows and columns for the different locations alphabetically ordered.

# Running the analysis in BEAST

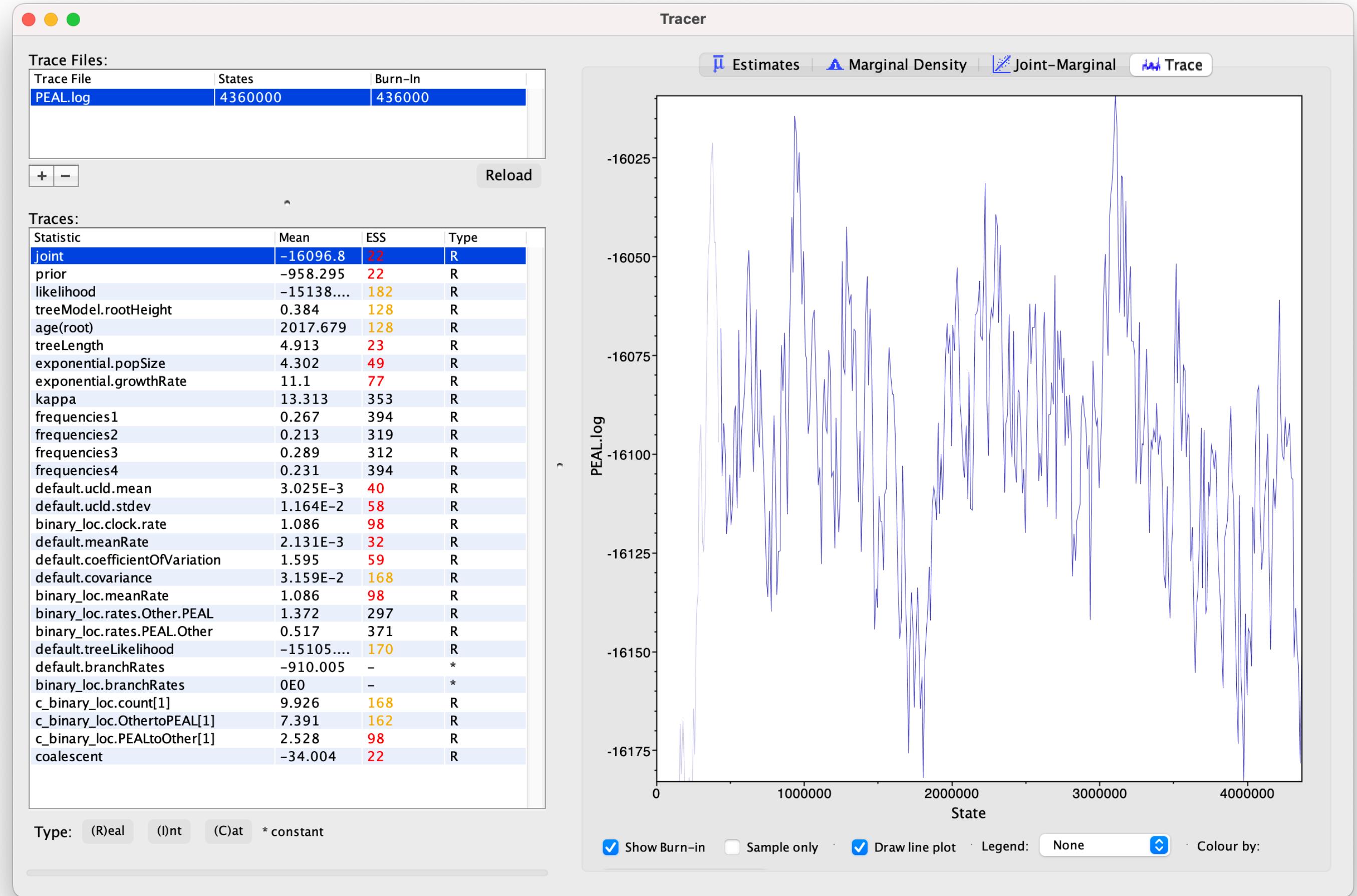
- Open BEAST by typing *beast* in the terminal
- Once BEAST is open, select the XML file for the analysis using the *Choose file* button.
- Click the *Run* button to start the analysis. Your terminal window should now be outputting the parameter values being explored by the MCMC sampler



# BEAST output files

- You should now see 3 output files being generated by BEAST:
  - A log file with all parameters being sampled in the analysis
  - A log file with the phylogeography rates being sampled
  - A trees file with all trees being sampled in the MCMC run
- We can use the Tracer to explore and visualize the output in the log files.

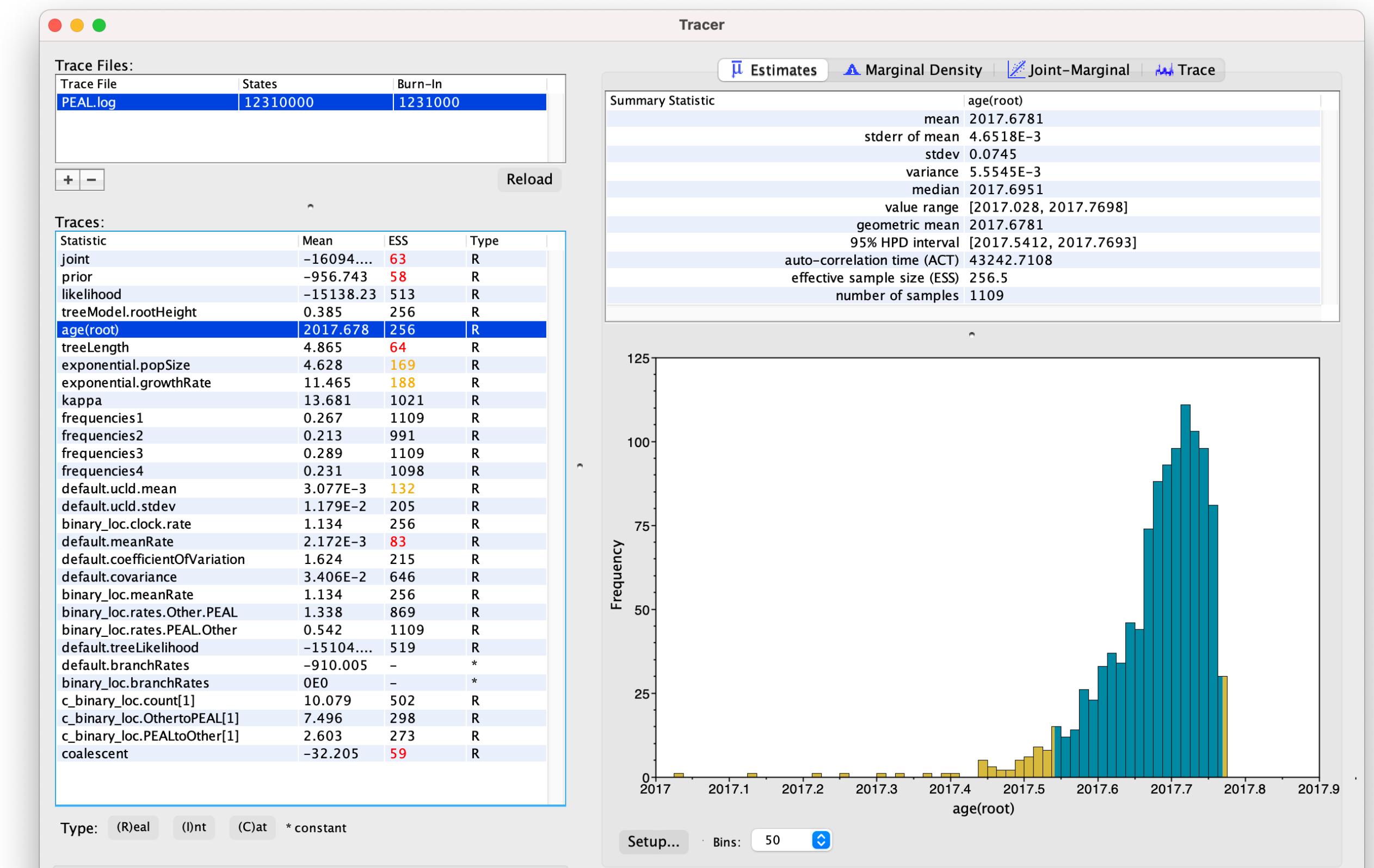
# Tracer



- Open Tracer and load the log file by dragging and dropping it into the *Trace Files* box
- Once loaded, you will see all parameters being logged on the *Traces* box on the left, and different visualizations on the tabs on the right.
- As a rule of thumb we want to keep running the analysis until all parameters in the log file have ESS values above 200.

# Exploring the Tracer results

- The values shown in Tracer represent the estimates of the parameters in our model being obtained by sampling across all trees in our posterior distribution.
- Thus, the uncertainty in our estimates is given for free without having to do any additional analysis.
- We can check the *95% Highest Posterior Density (HPD)* range for our parameters in the *Estimates* tab on the right. We call this the *credible interval* of our parameter, and it states that there is a 95% probability that the parameter value falls within this range.



# Back to the epidemiology

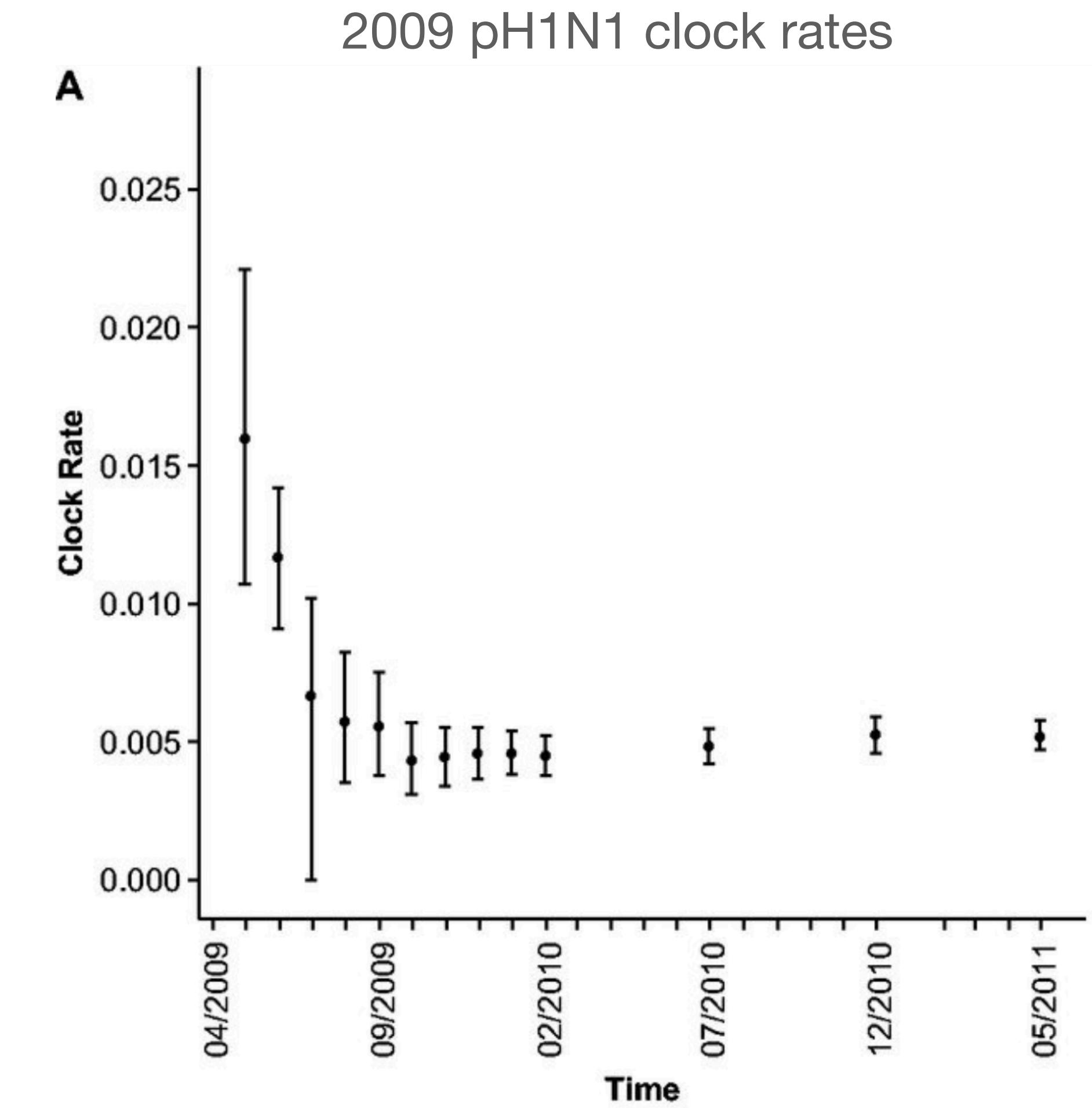
- When did the outbreak begin?
- How fast is the virus evolving on average?
- How many mutations would you expect after 1 week, given a genome size of 10kb?
- How many introductions to PEAL?
- How rapidly is the virus population growing?
- What is the doubling time of the virus?

# Back to the epidemiology

- When did the outbreak begin?
  - 2017.7 [2017.5 - 2017.8]
- How fast is the virus evolving on average?
  - $2.2\text{E-}3$  [ $1.3\text{E-}3$  -  $3.2\text{E-}3$ ]
- How many mutations would you expect after 1 week, given a genome size of 10kb?
  - 0.42 mutations [ $0.24$  -  $0.61$ ]
- How many introductions to PEAL?
  - 7.45 [5 - 11]
- How rapidly is the virus population growing?
  - Growth rate 10.9 [2.3 - 20.2]
- What is the doubling time of the virus?
  - Doubling time 23 days [12 - 109]

# A note on the evolutionary rate

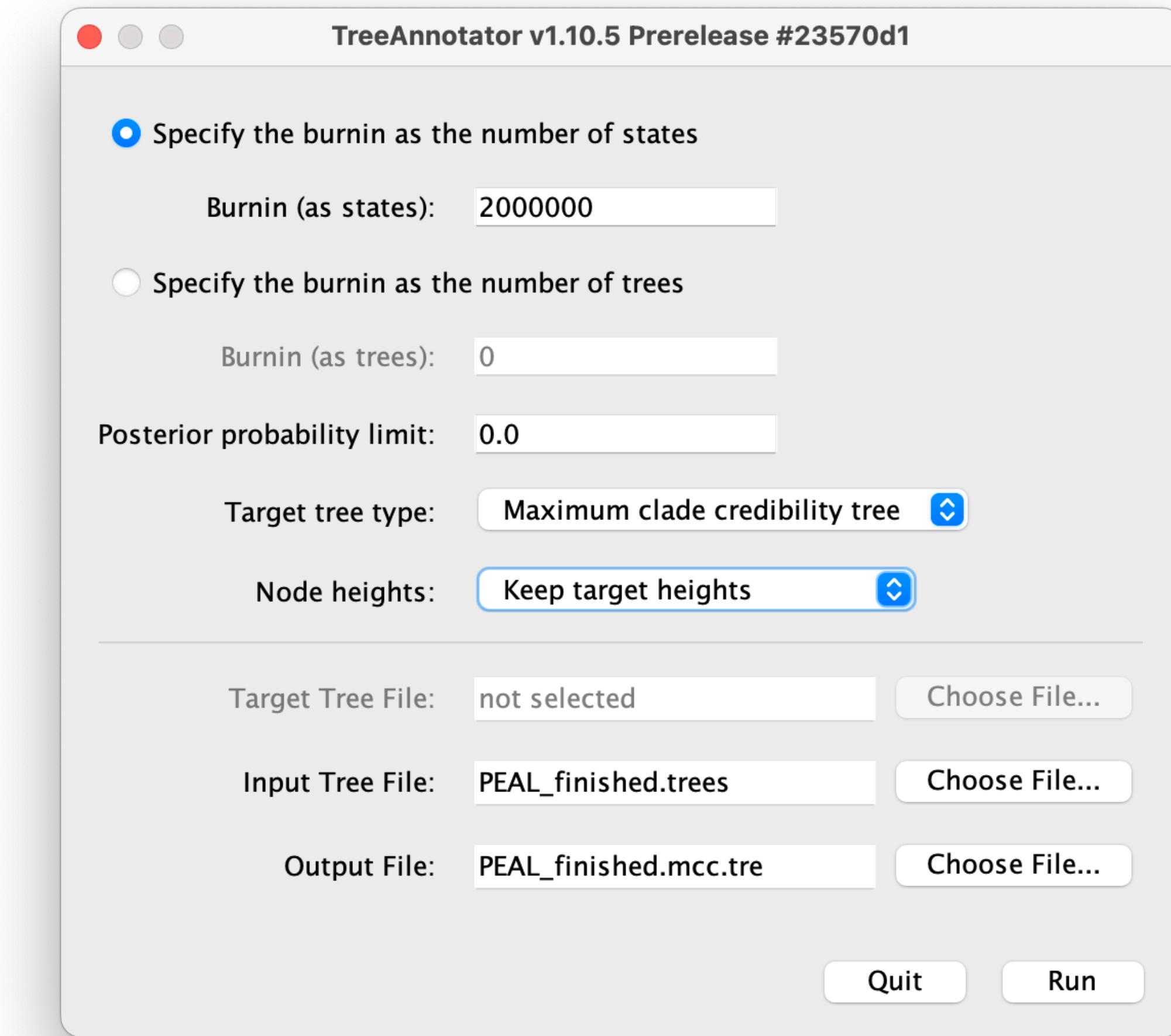
- Notice that the mean rate we obtain from our analysis,  $2.2\text{E-}3$  [ $1.2\text{E-}3$  to  $3.2\text{E-}3$ ] is much faster than the rate of  $\sim 9\text{E-}4$  reported in previous studies.
- An important thing to consider is that all of our sequences come from a short and densely sampled period of time at the beginning of the outbreak. The short amount of time of rapid growth in the virus population means that many mutations we observe might actually be transient polymorphisms, deleterious mutations that are still present in the population because purifying selection hasn't had enough time to act.
- This phenomenon of inflated evolutionary rates at the beginning of an outbreak can also be seen in other pathogens such as with the case of the 2009 pandemic H1N1 influenza virus.



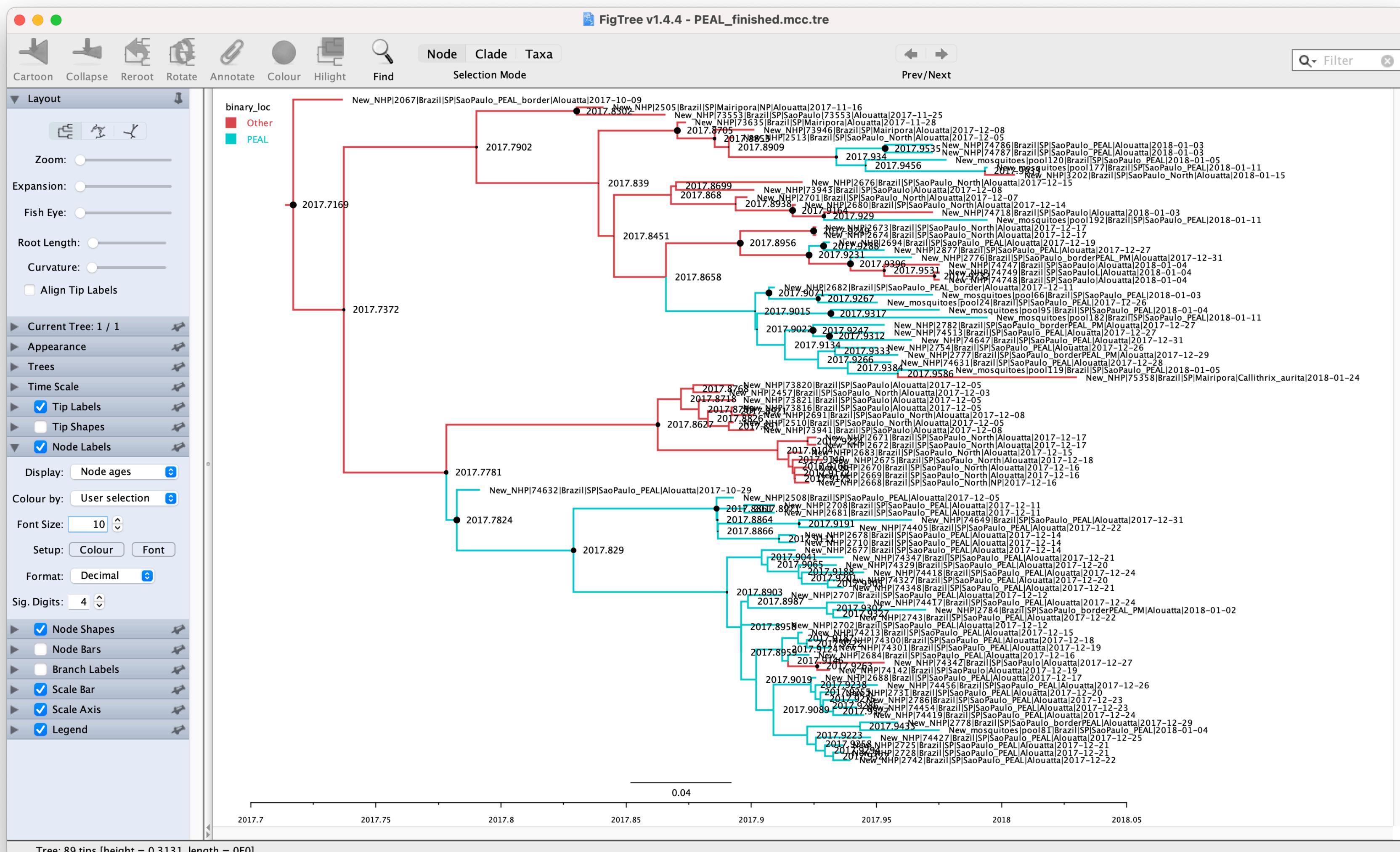
Meyer et al. 2009

# Summarizing posterior trees

- We can summarize posterior trees using TreeAnnotator to obtain a Maximum Clade Credibility (MCC) tree
- For each tree in the posterior, calculate the fraction of times each clade appears in the collection of trees
- The MCC tree is the tree that maximizes the product of these fractions
- Posterior support values are annotated in the internal nodes denoting the probability of each clade



# Visualizing the MCC tree



Finally, we can use FigTree to visualize the MCC tree with the annotations