Statistical modeling with stochastic processes

Alexandre Bouchard-Côté

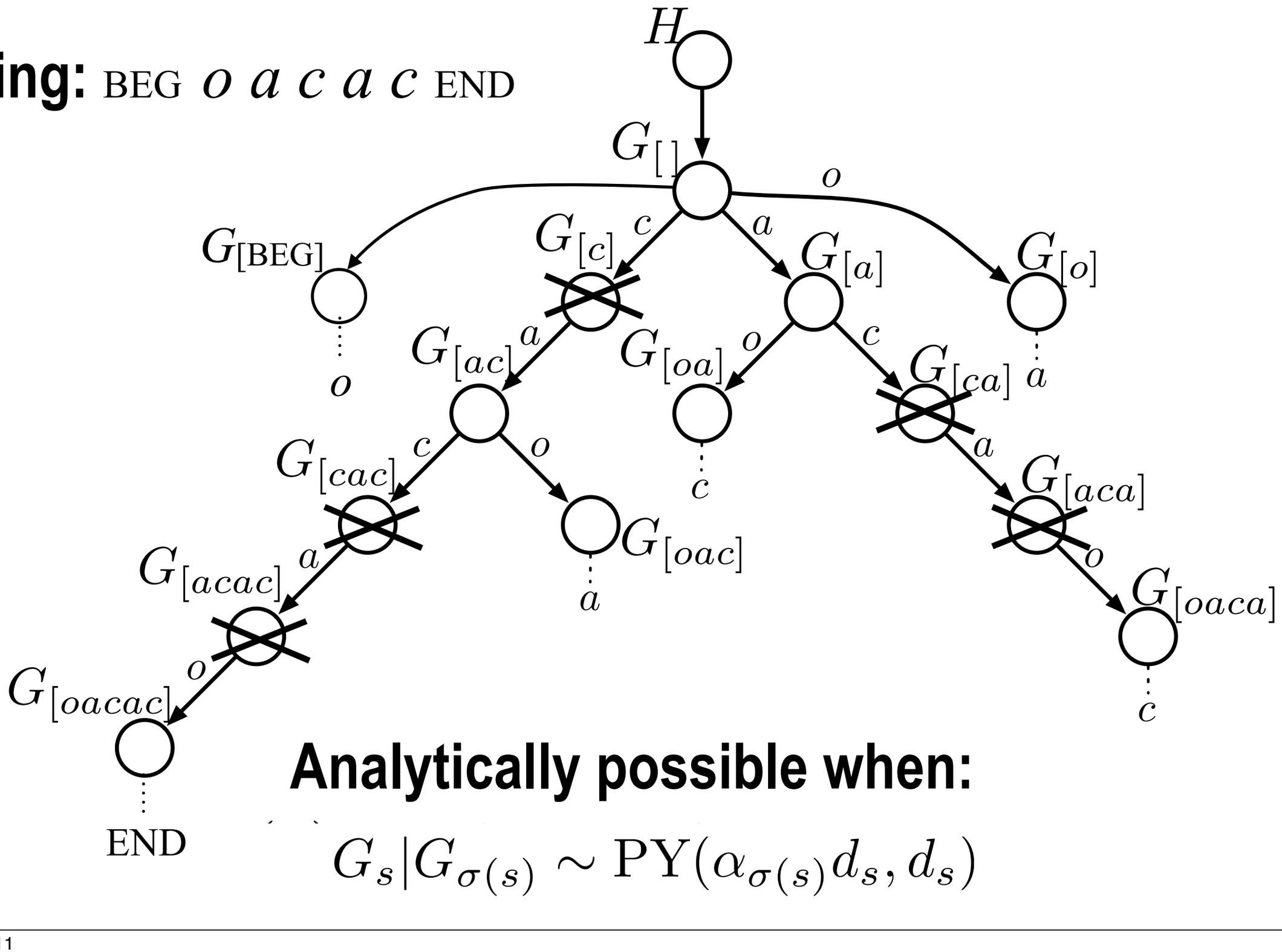
Lecture 12, Wednesday April 6

Program for today

- Dependent Dirichlet Processes
- CTMCs, trees and random hierarchies

Marginalization

Training: BEG $o\ a\ c\ a\ c$ END

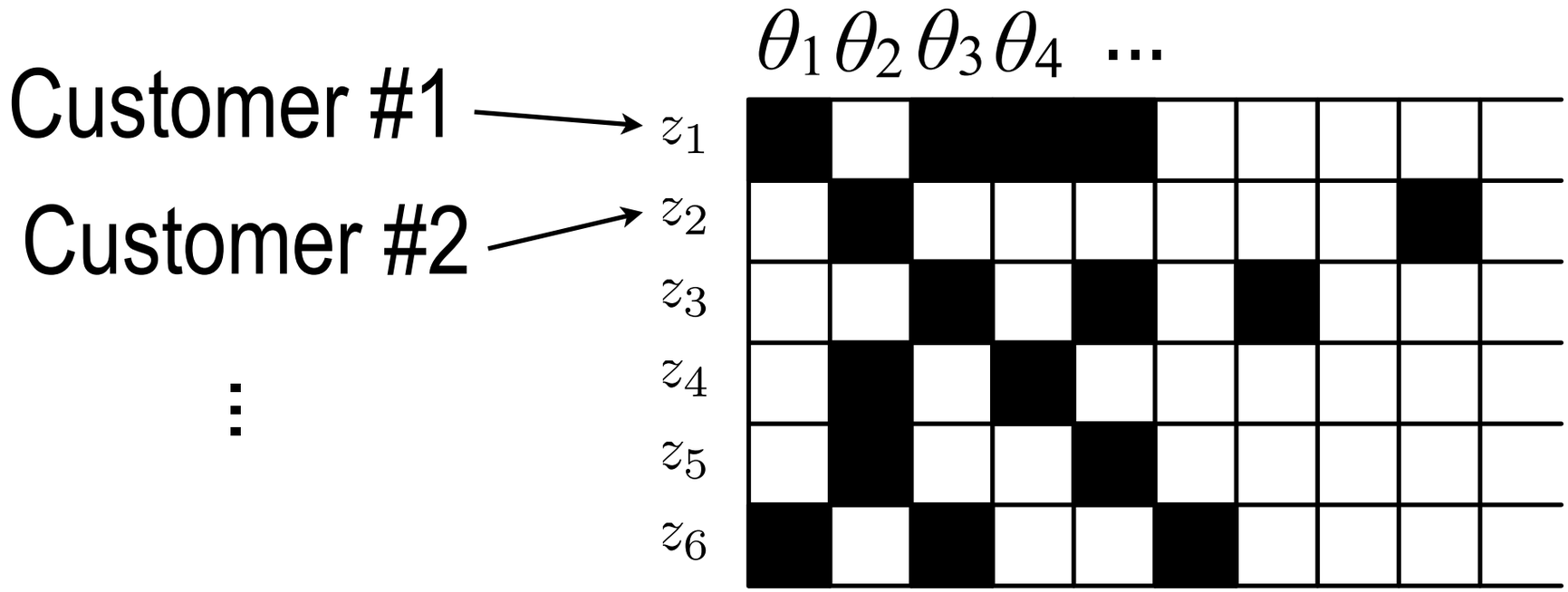


Analytically possible when:

$$G_s|G_{\sigma(s)} \sim \mathrm{PY}(\alpha_{\sigma(s)}d_s, d_s)$$

Predictive distribution: restaurant metaphor

Instead of a sit-down restaurant, think of a buffet with an infinite sequence of dishes θ_i sampled by customers



Poisson processes

Another random discrete measure, but unnormalized:

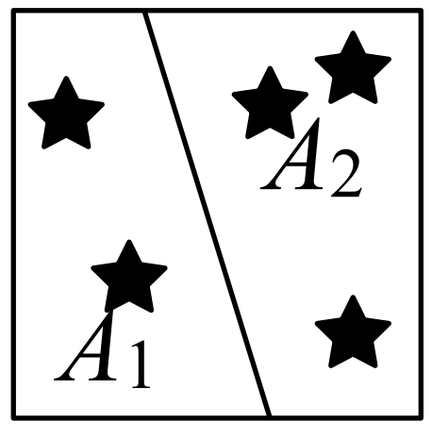
Let P_0 be a distribution on a sample space Ω (the base distribution) and $(A_1, ..., A_k)$ be a partition of Ω. We say

$$P \sim \mathrm{PP}(P_0)$$

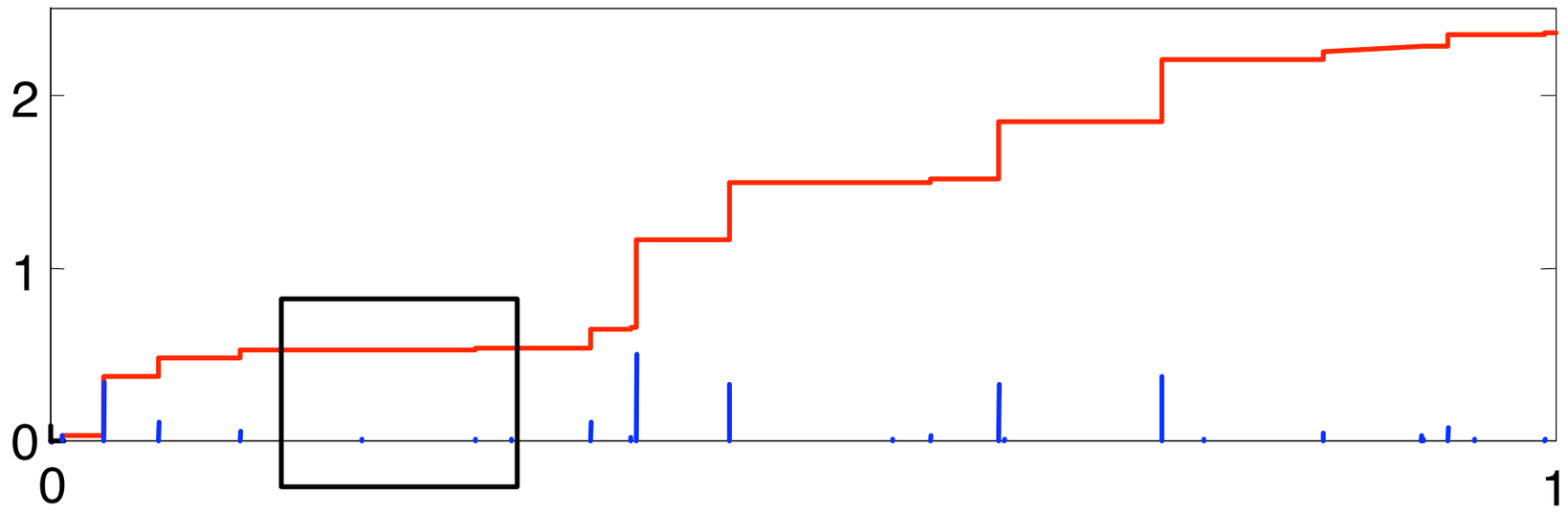
i.e., P is a Poisson Process, if

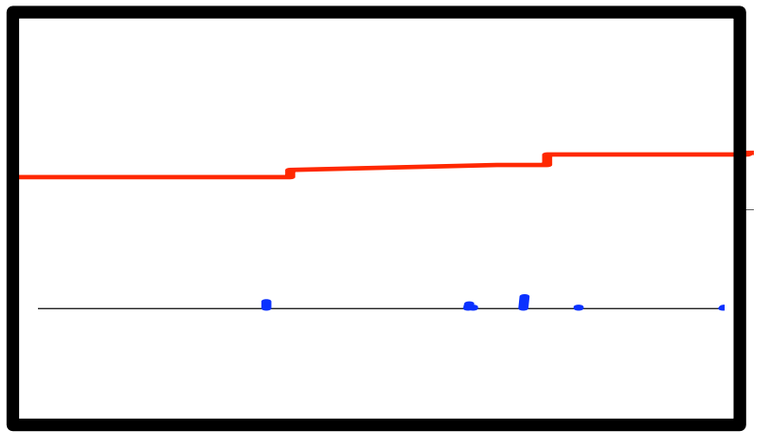
$$P(A_1) \overset{\text{ind.}}{\sim} \mathrm{Poi}(P_0(A_1))$$

for all partitions and all k.



From PP to Gamma Process to DP





Campbell's theorem

Assume P_0 is a probability measure, f is bounded, and $P \sim \mathrm{PP}(P_0)$.

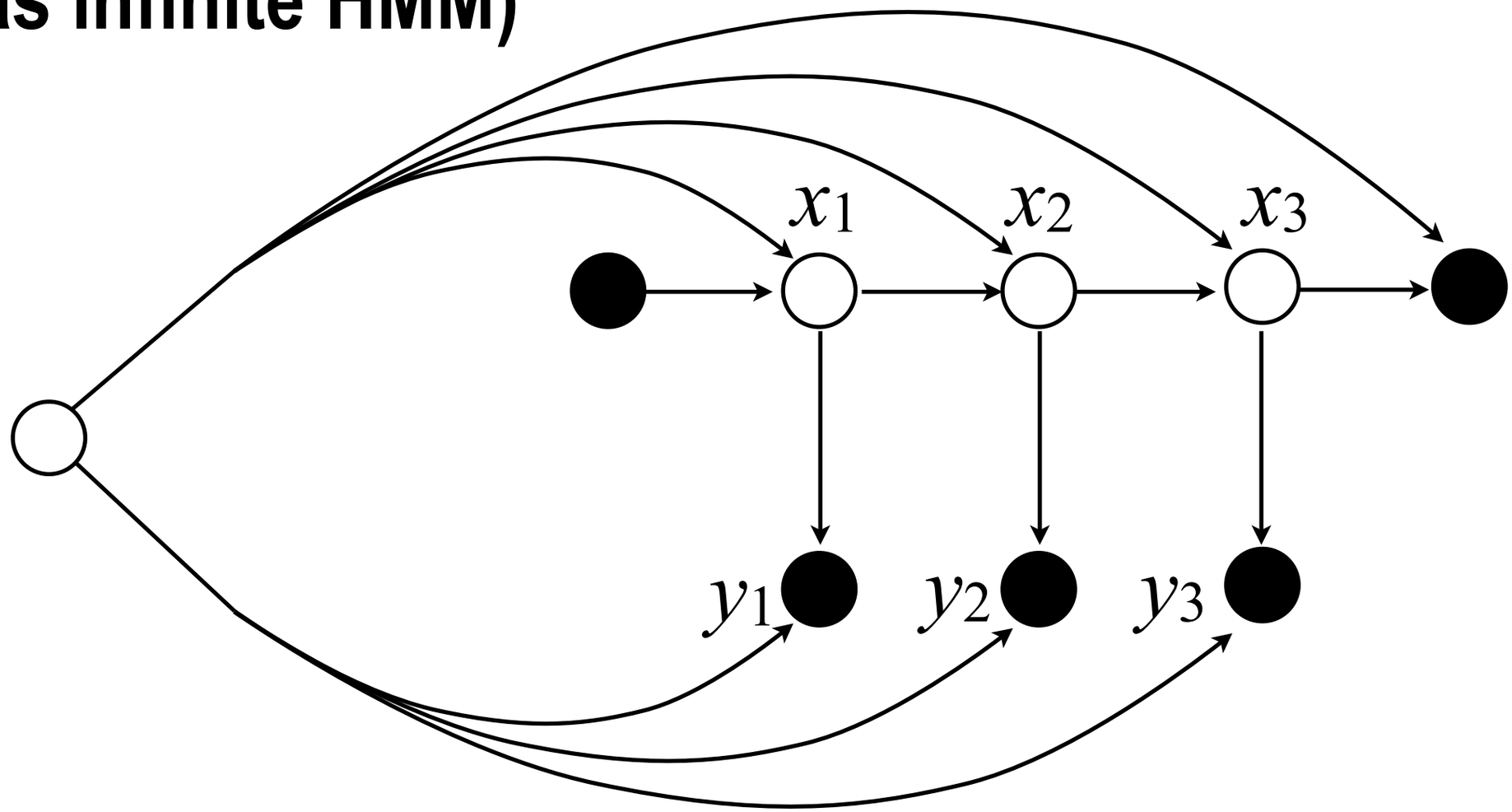
Let also: $$\Sigma = \sum_{X \in P} f(X)$$

Then: $$\mathbb{E}\left[e^{it\Sigma}\right] = \exp\left\{\int_\Omega (e^{itf(x)} - 1) P_0(\,\mathrm{d}x)\right\}$$

Dependent Dirichlet Processes

Desired model

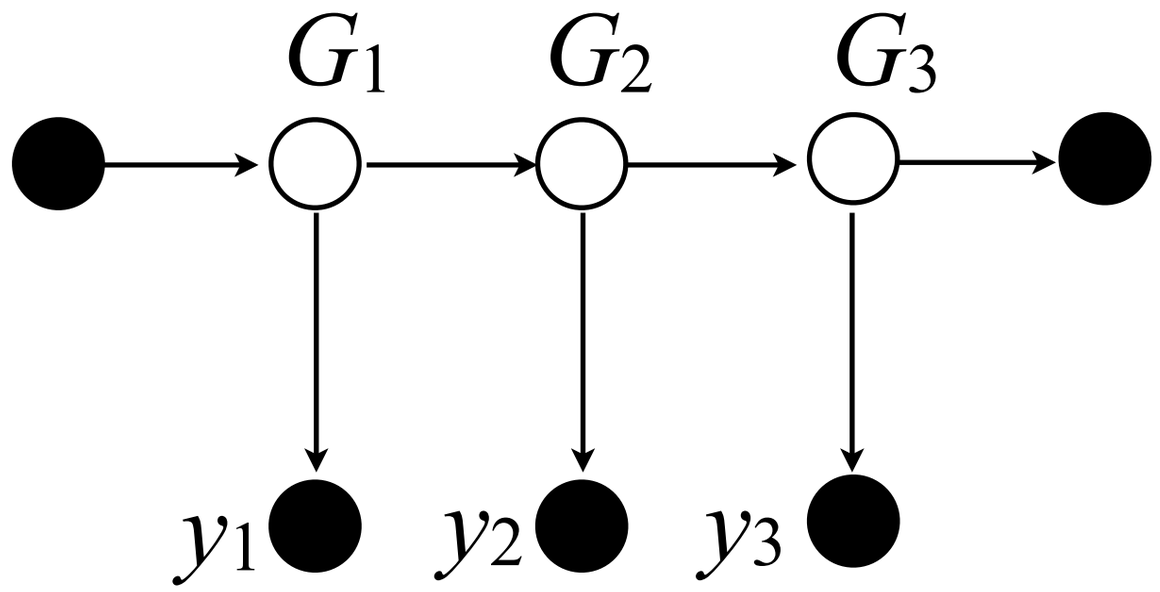
Sequential (as infinite HMM)



But this time: a forgetful model, where

$$\lim_{t \to \infty} \mathbb{P}(z_{t+s} = i | z_s = i) = 0$$

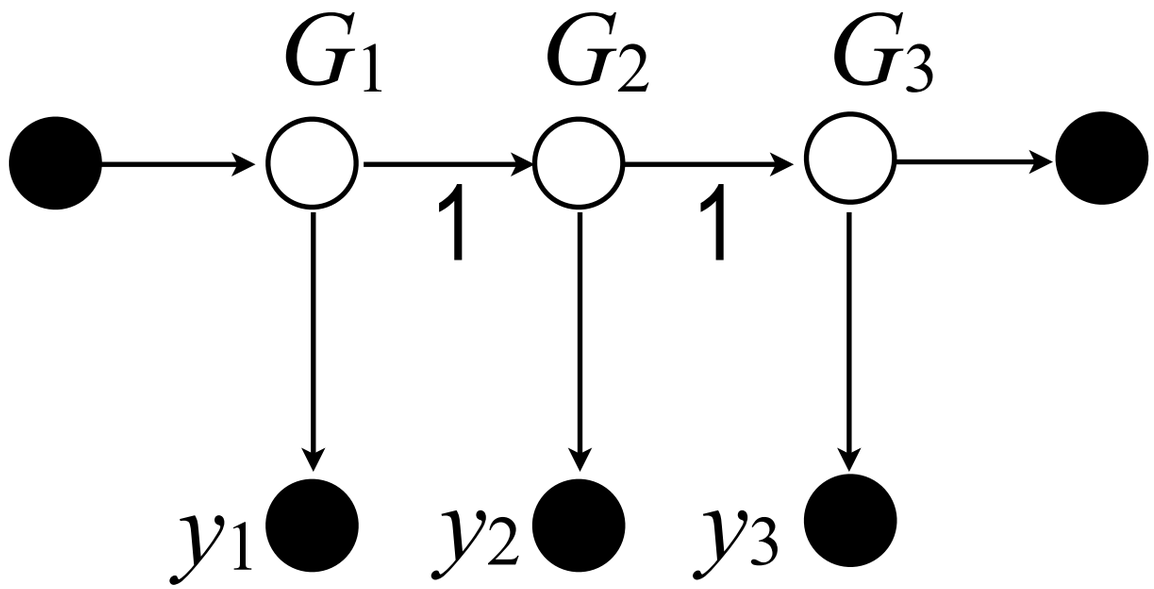
Inspiration: AR models

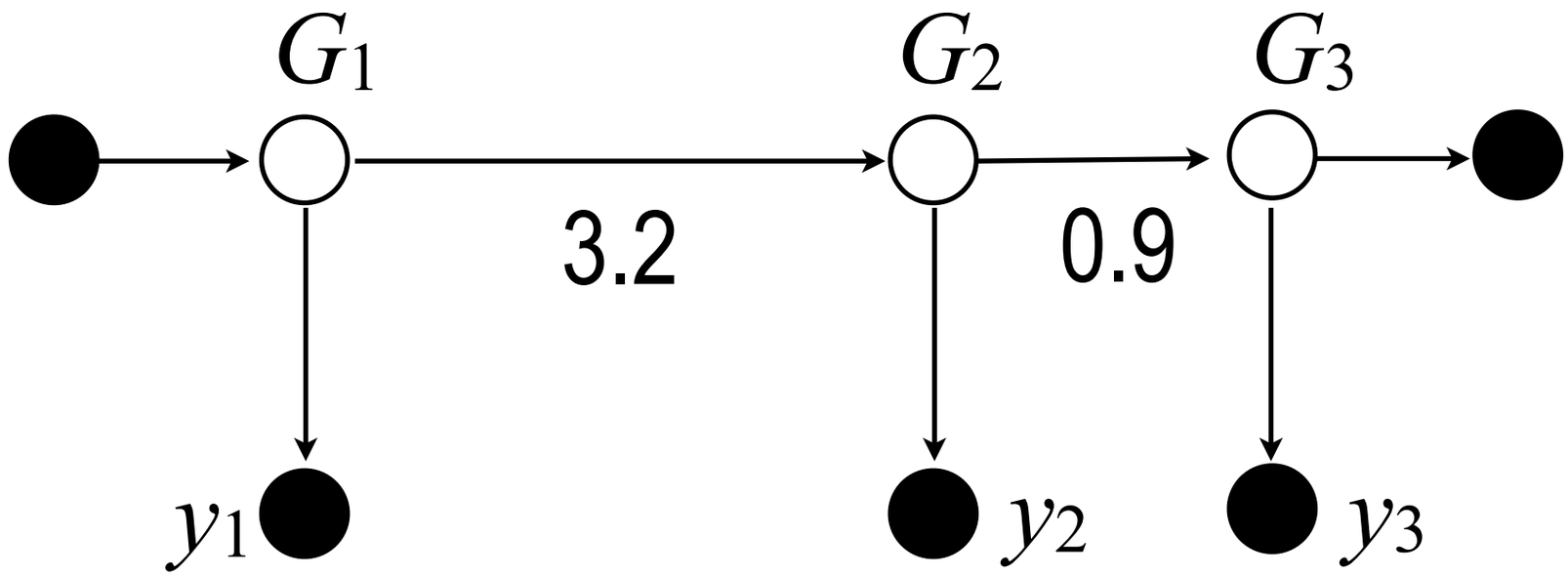


$$G_1 \sim \mathrm{DP}(\alpha_0, H)$$

$$G_{t+1} = w_t G_t + (1 - w_t)\epsilon_t$$

Time continuous version?

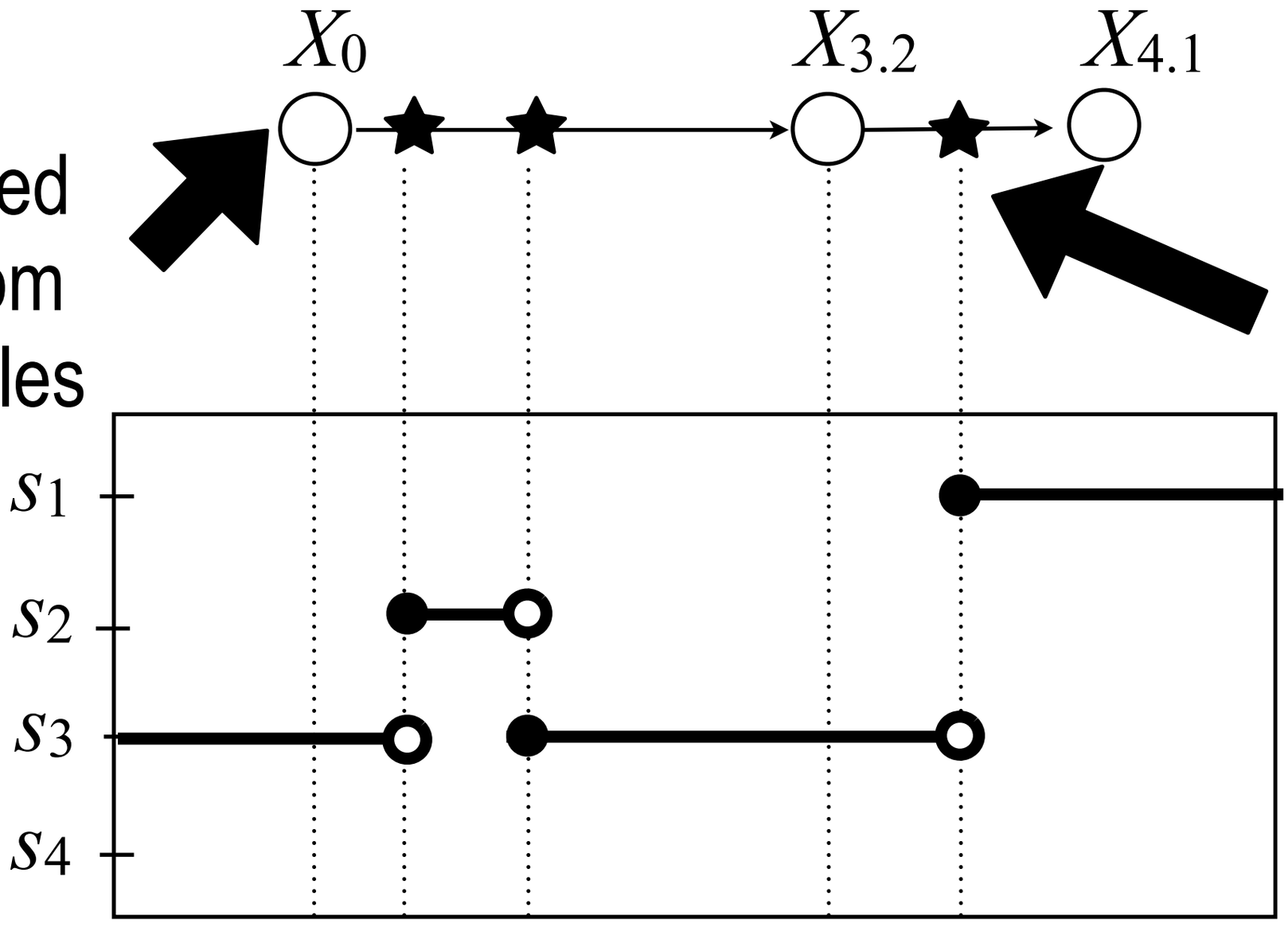




First: continuous time, finite state space models (CTMCs)

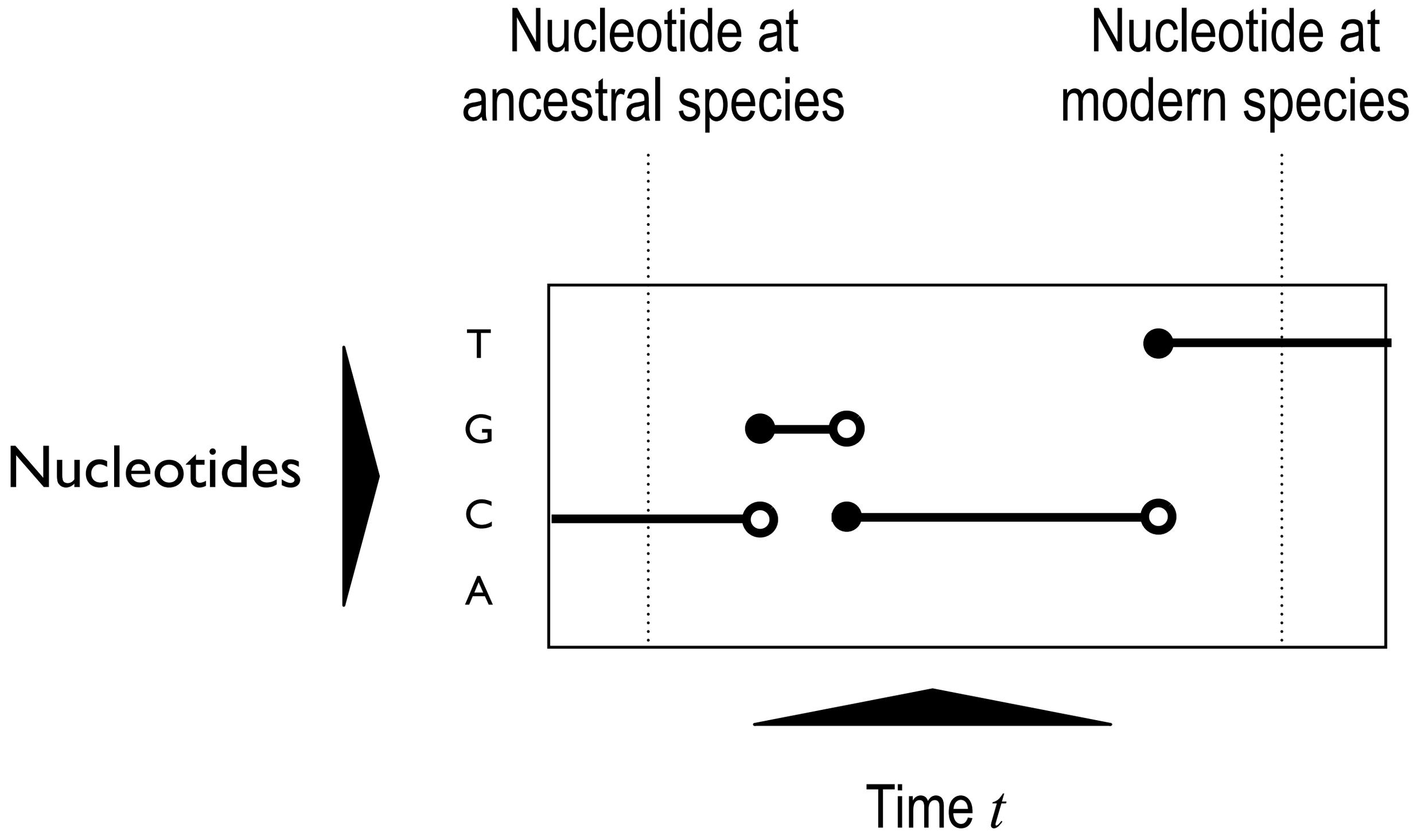
CTMC

Queried random variables

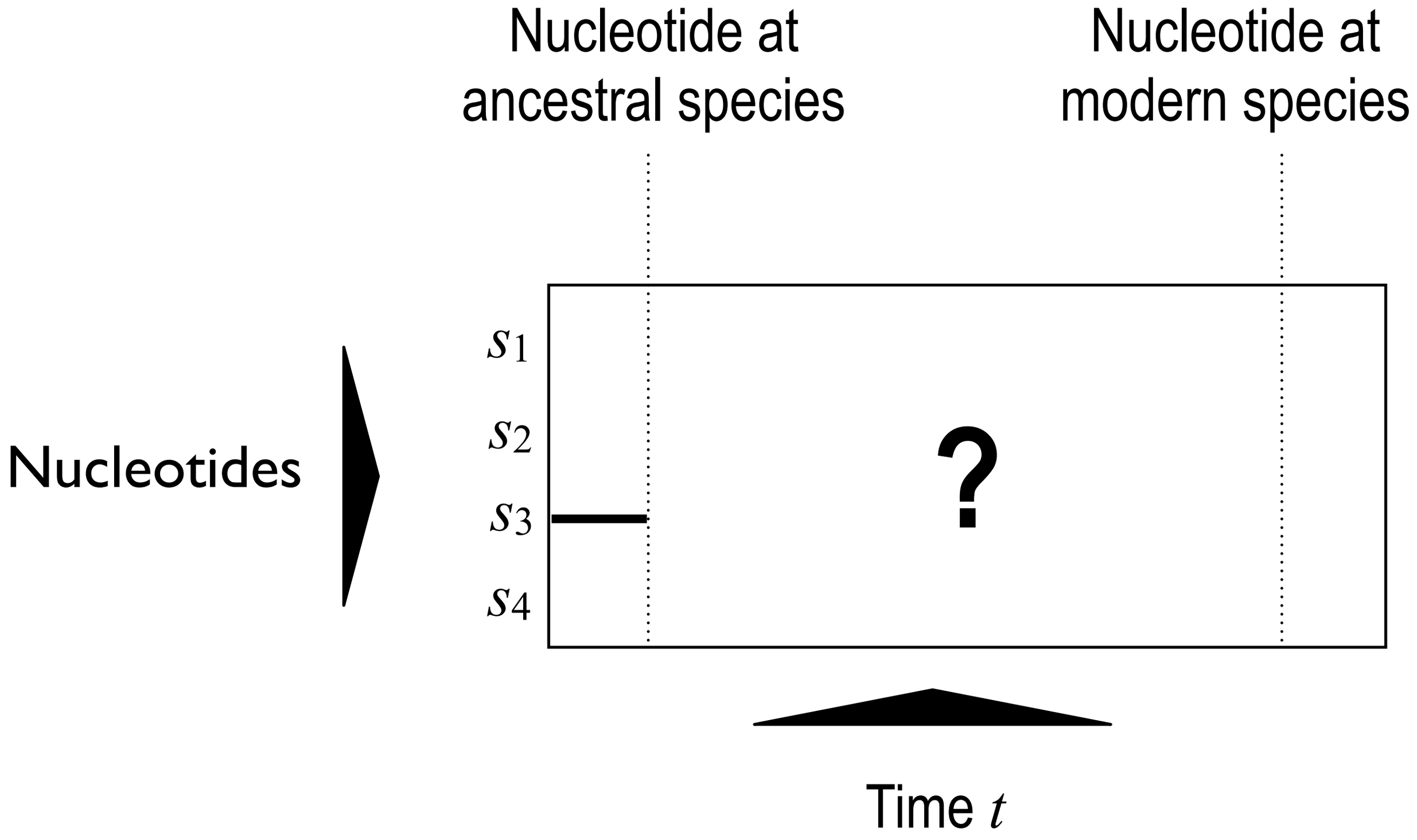


Changes can occur at points of the continuum between query nodes (want to marginalize over this)

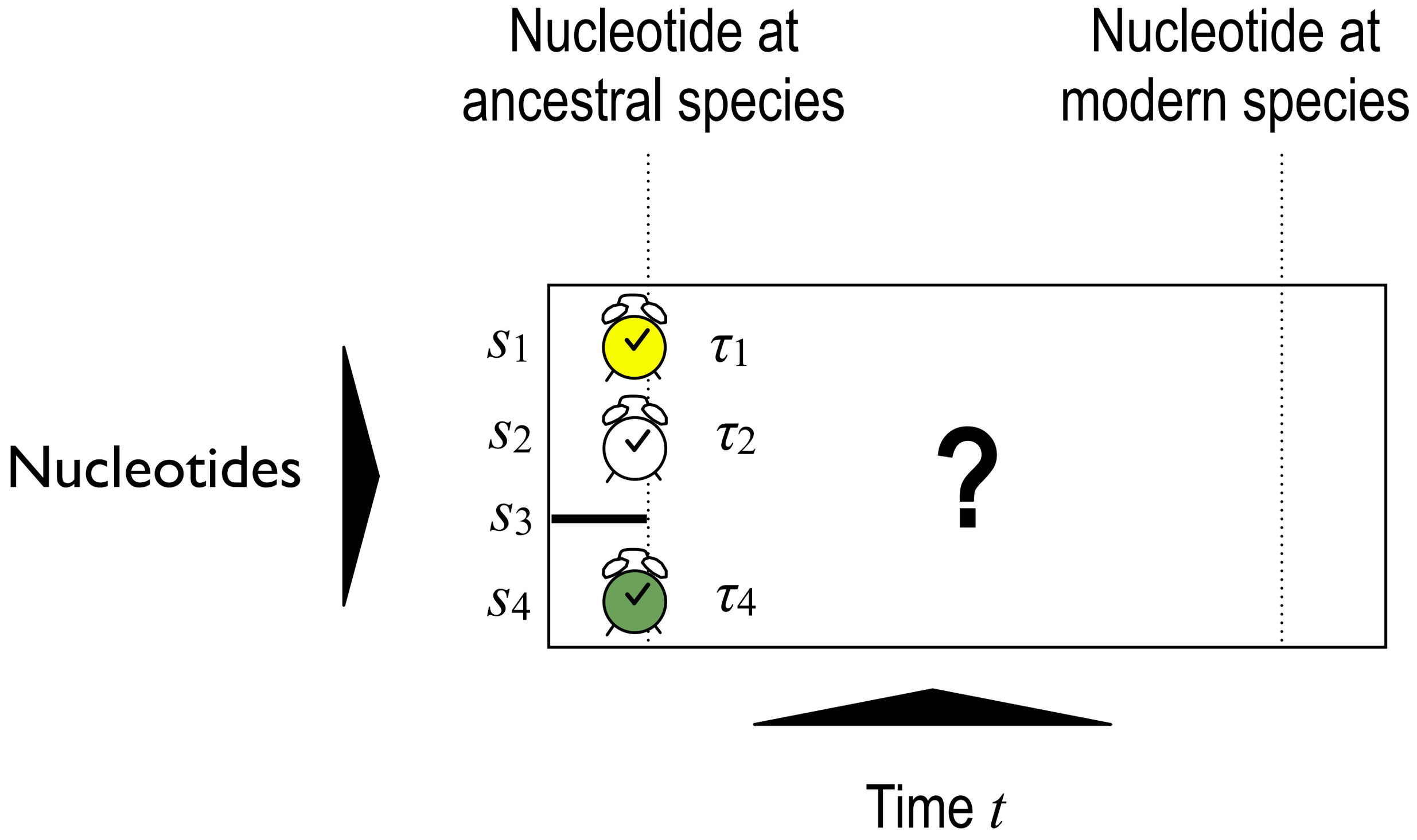
Example: Models for DNA evolution



Generating random CTMC paths

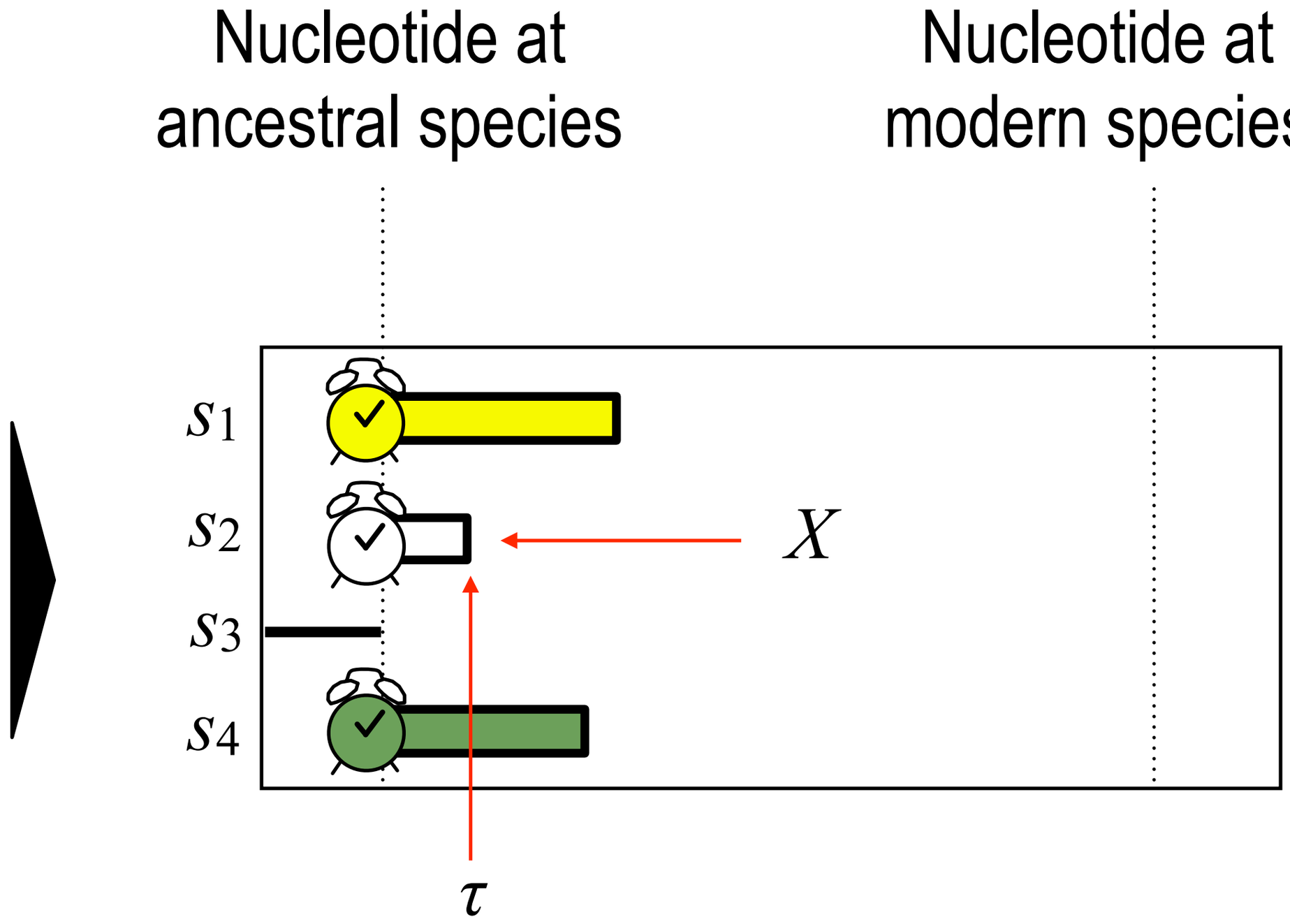


Generating random CTMC paths



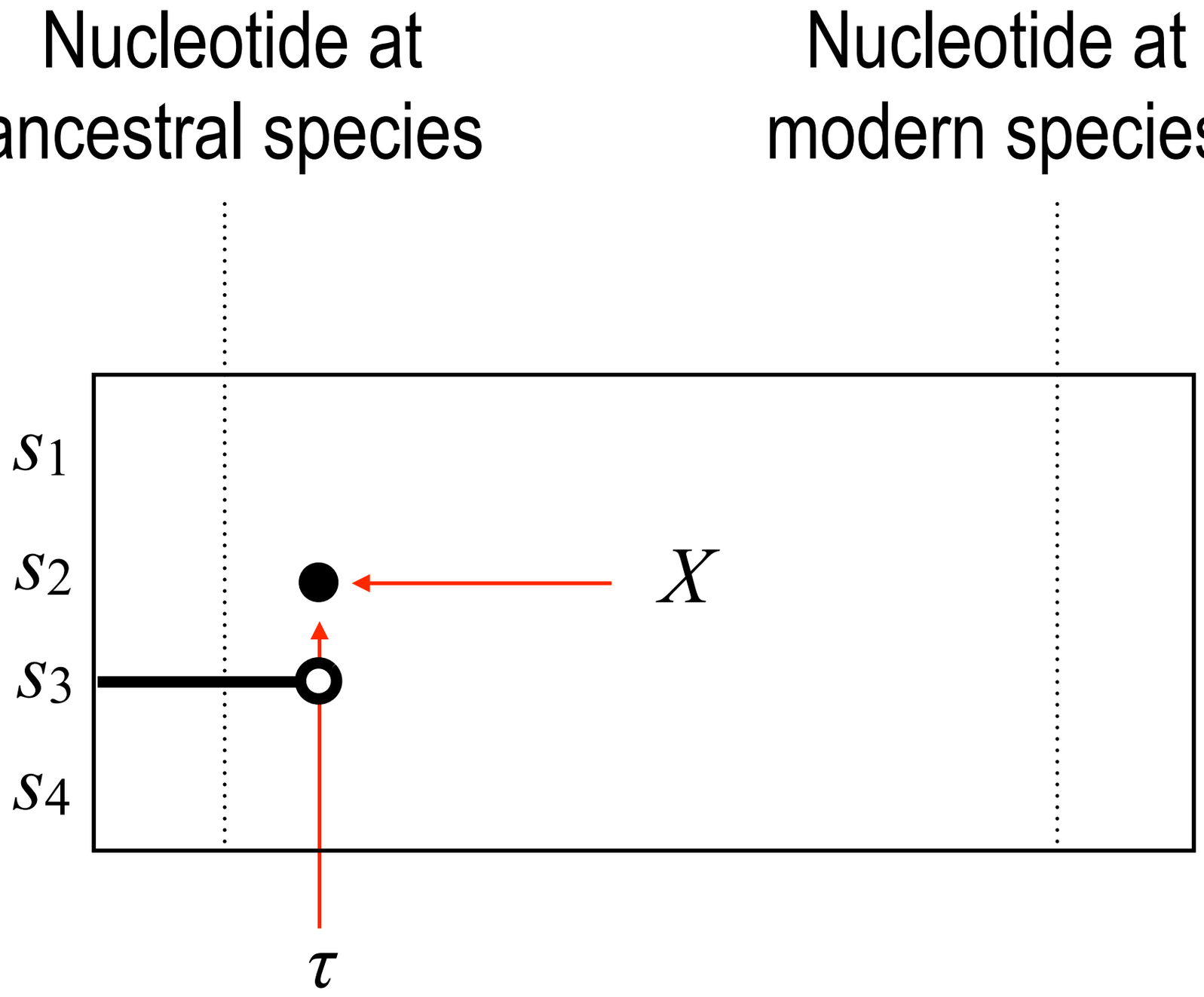
Generating random CTMC paths

Nucleotides



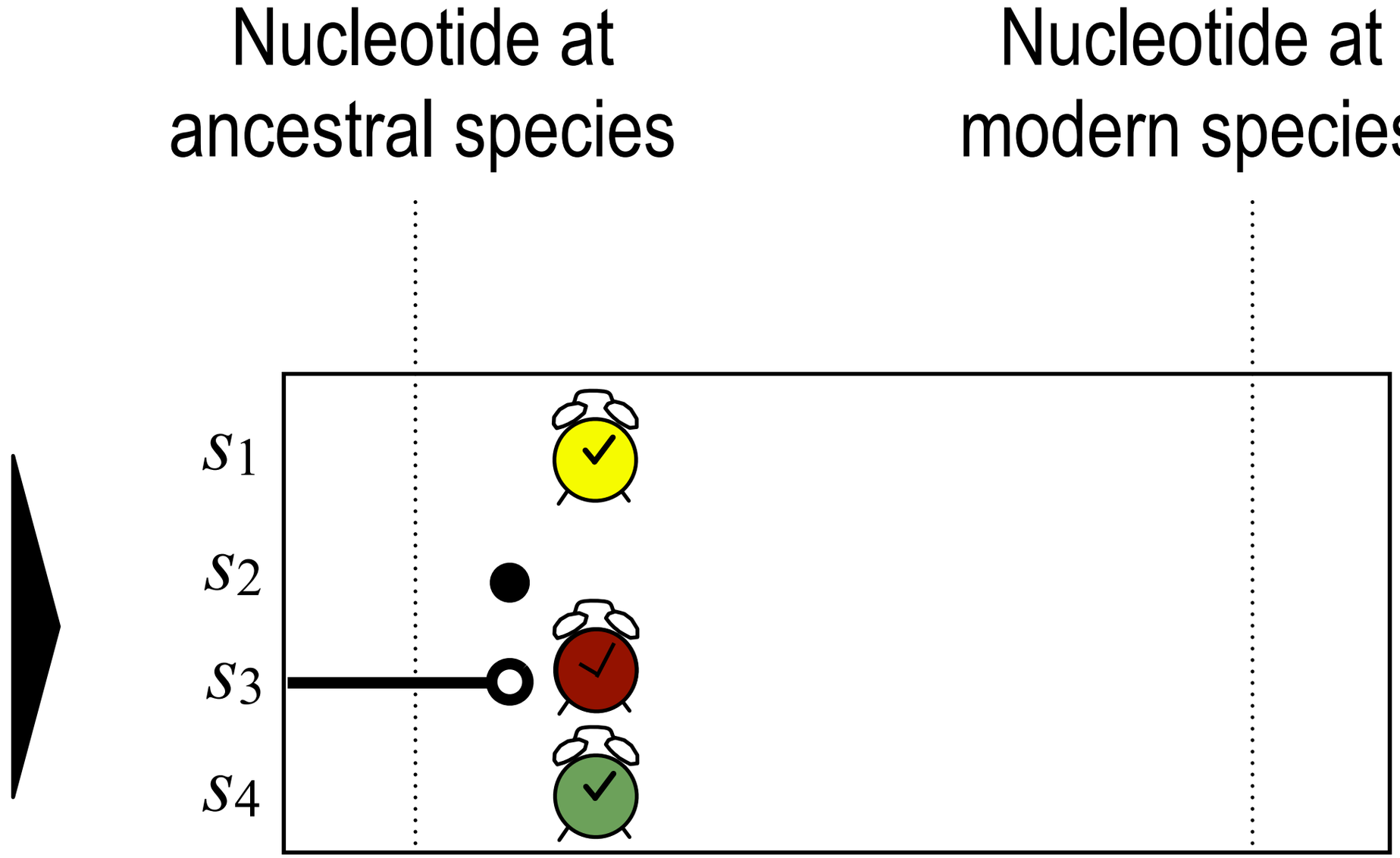
Generating random CTMC paths

Nucleotides



Generating random CTMC paths

Nucleotides

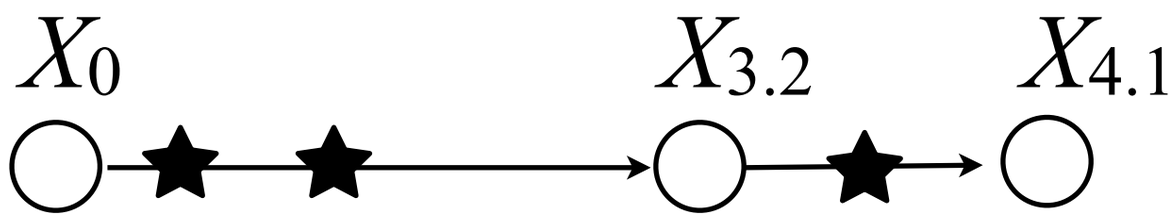


CTMC: special case

Suppose: $q_{ij} = c$ for all $i \neq j$ and some constant $c > 0$

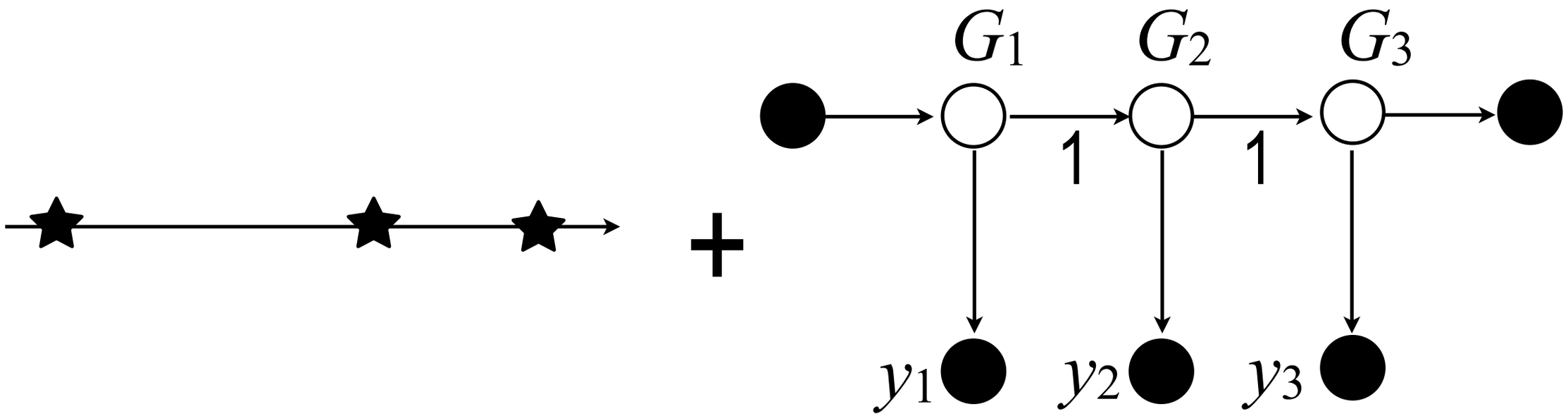
Then: the points of mutation are distributed according to a Poisson process:

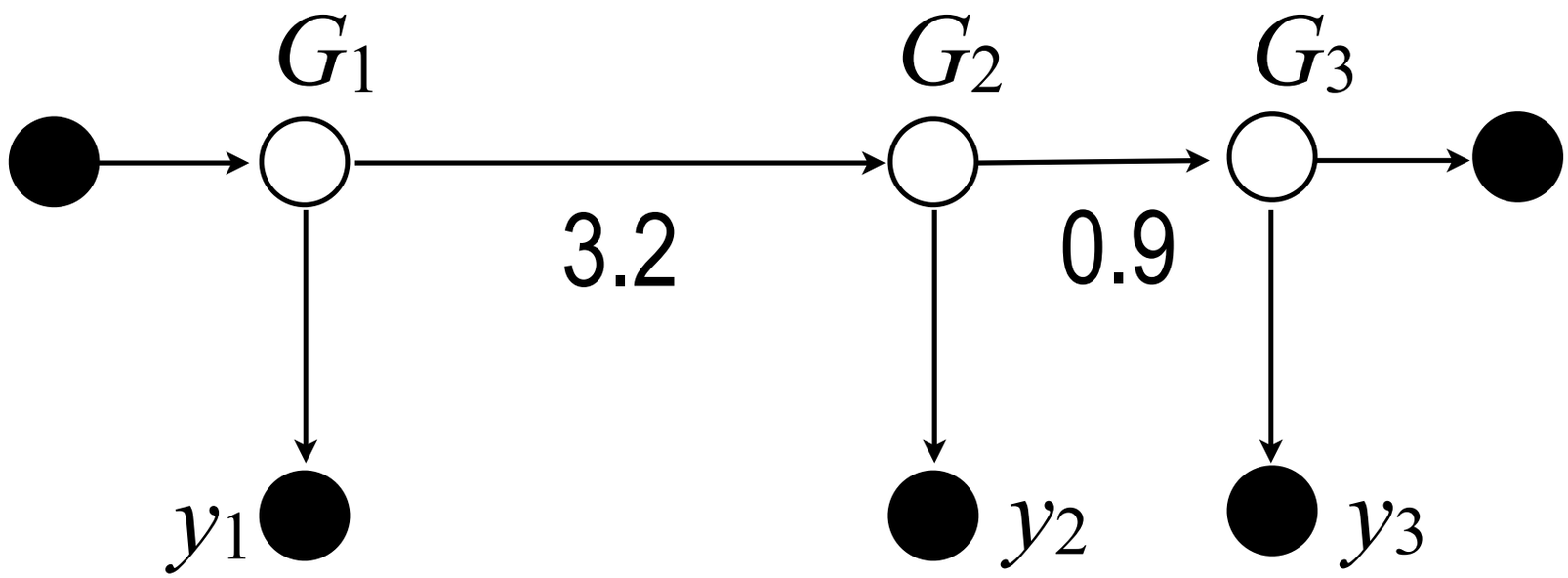
$$\{\bigstar\} \sim \text{PP}(c, \text{Uni}(0, 4.1))$$



Application to DDP: make the location of insertion of sticks distributed according to a Poisson process

Time continuous version

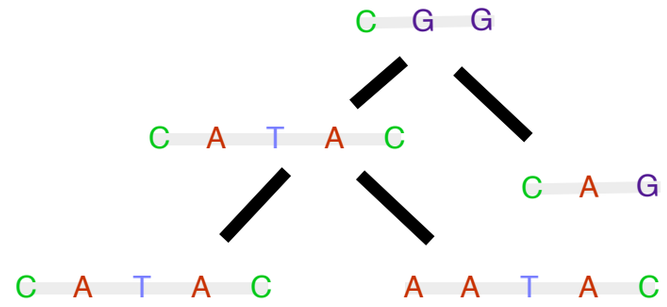




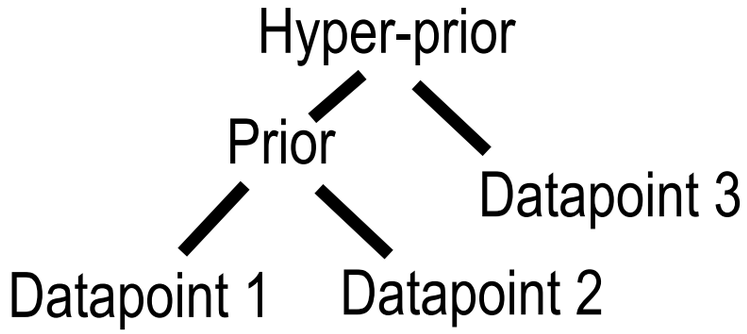
Quick overview of phylogenetics

Applications

Evolution/ language change



Hierarchical models with random hierarchies



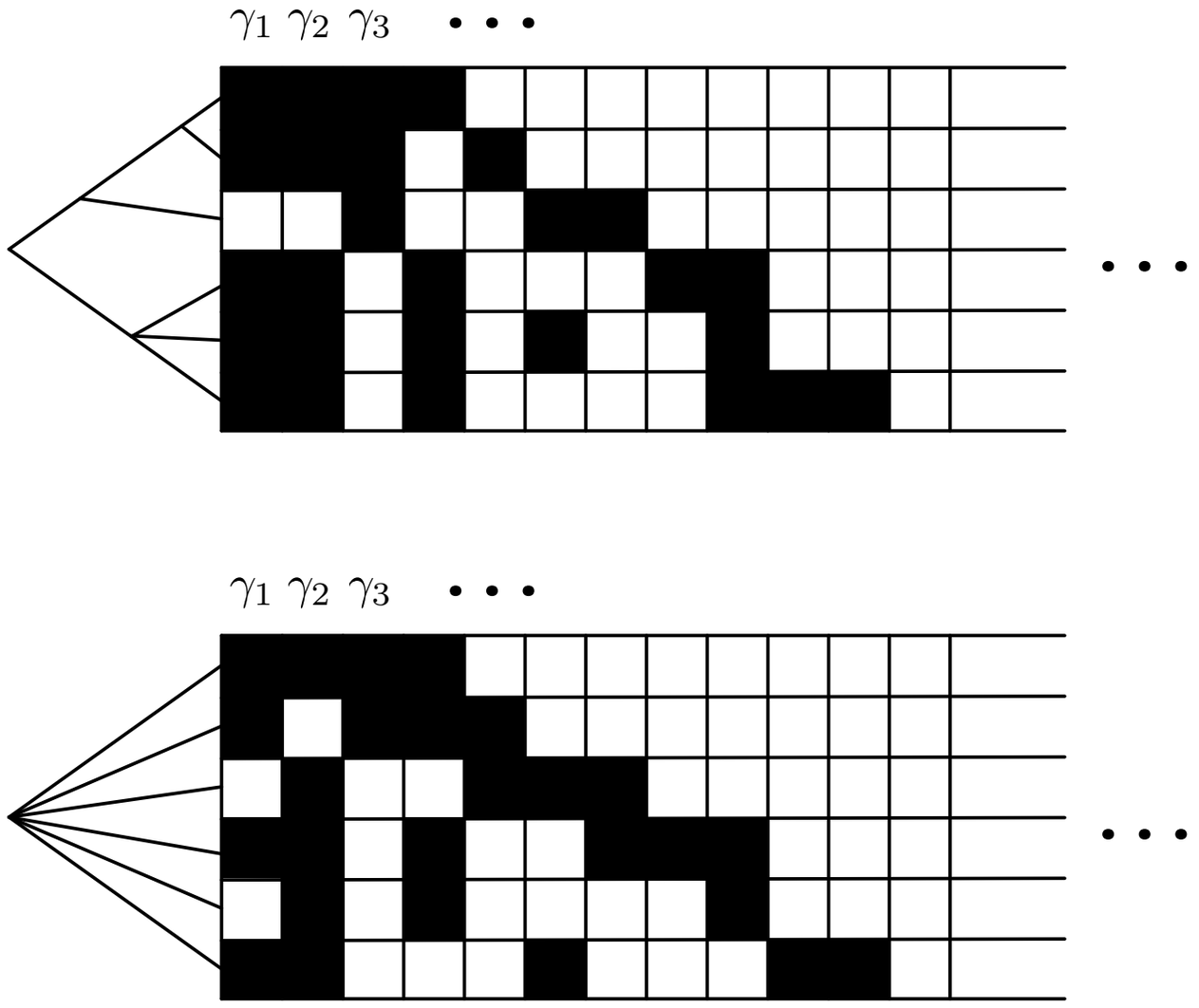
Random hierarchies: example

Suppose we want a model similar to IBP, but with hierarchical grouping on the customers

We need:

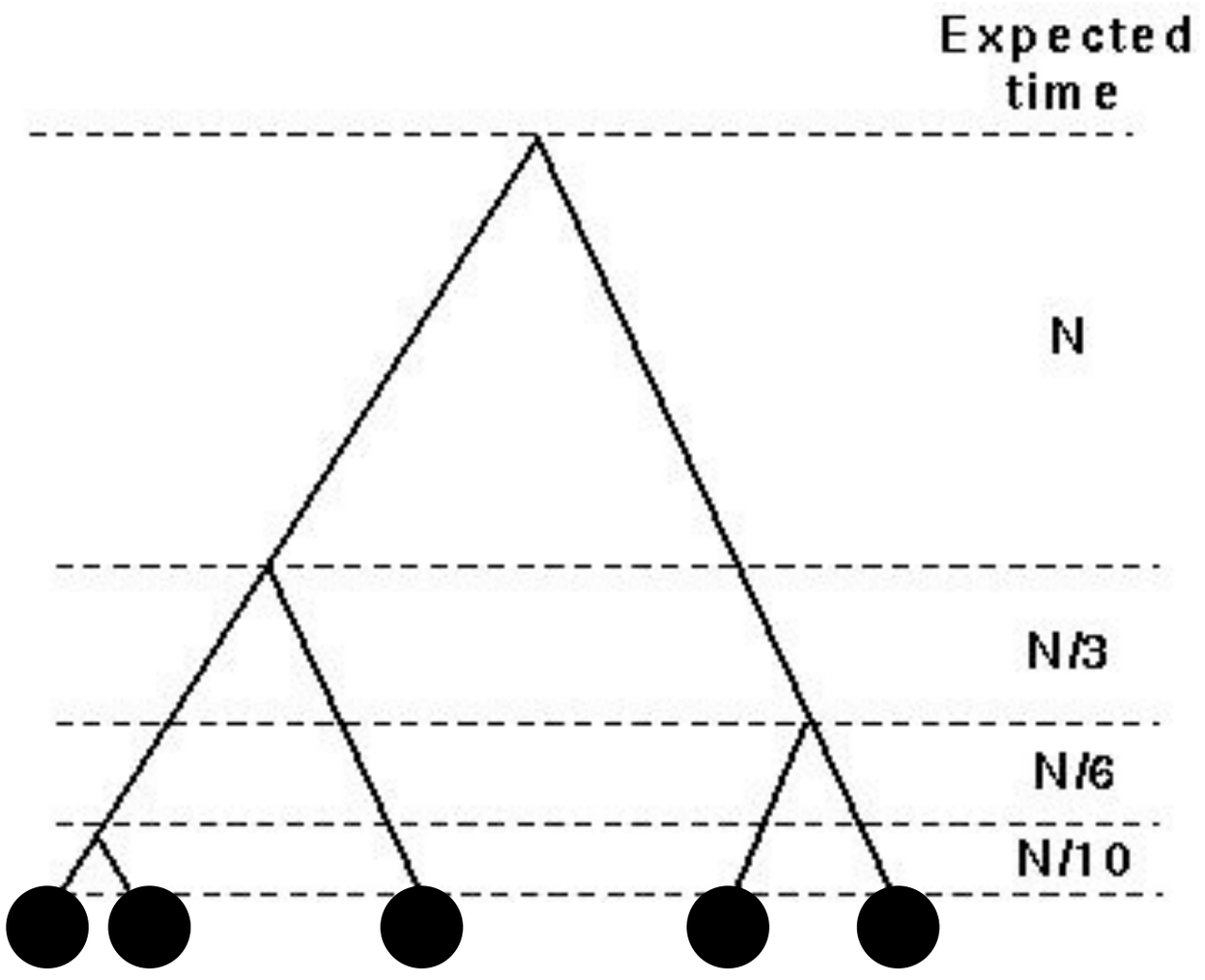
- A prior over trees (hierarchy structures)

- Given a tree, a likelihood model for the ‘evolution’ of feature indicators



Models over trees: example

Kingman's coalescent

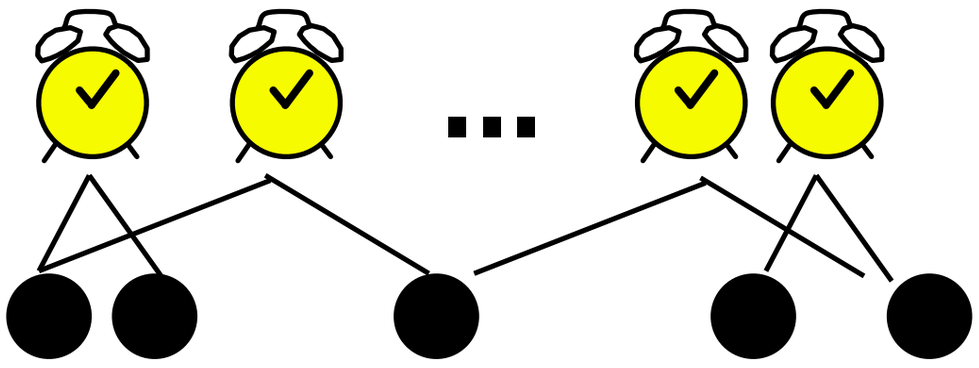


Models over trees: example

Kingman's coalescent

Models over trees: example

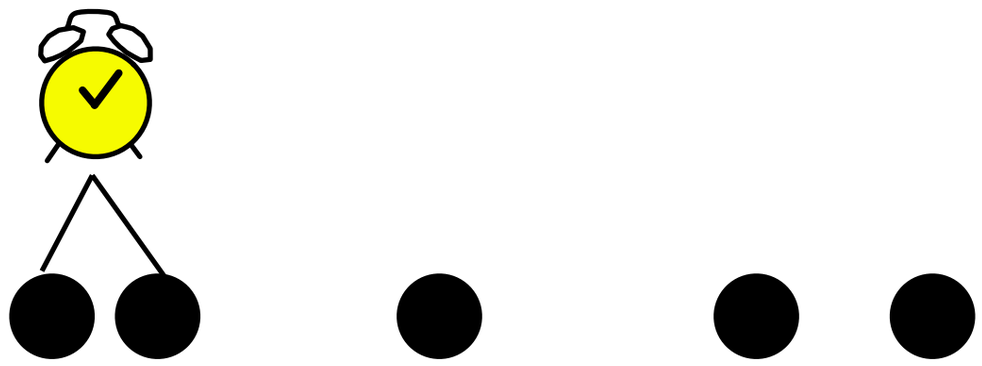
Kingman's coalescent



Simulate iid exp(1) clocks, one for each pair of points

Models over trees: example

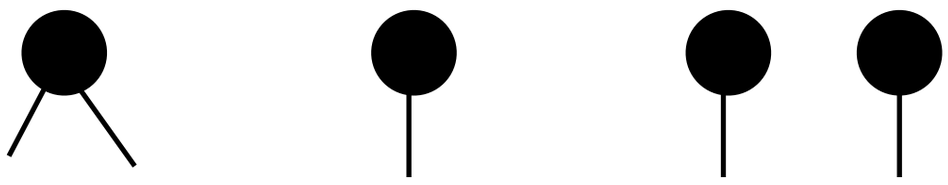
Kingman's coalescent



Again, the winner determines the first merge (coalescent) event

Models over trees: example

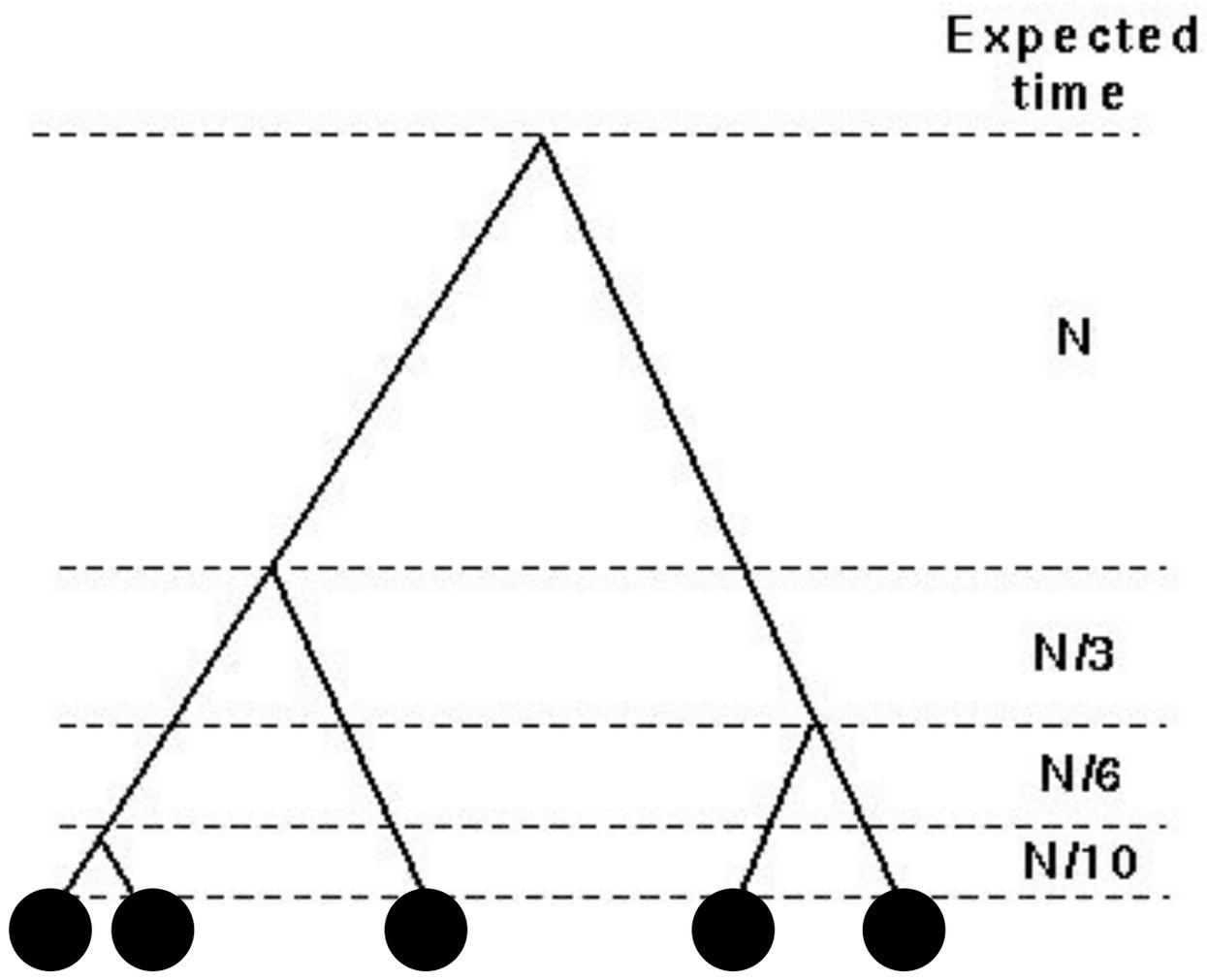
Kingman's coalescent



Start again, with
n-1 points now

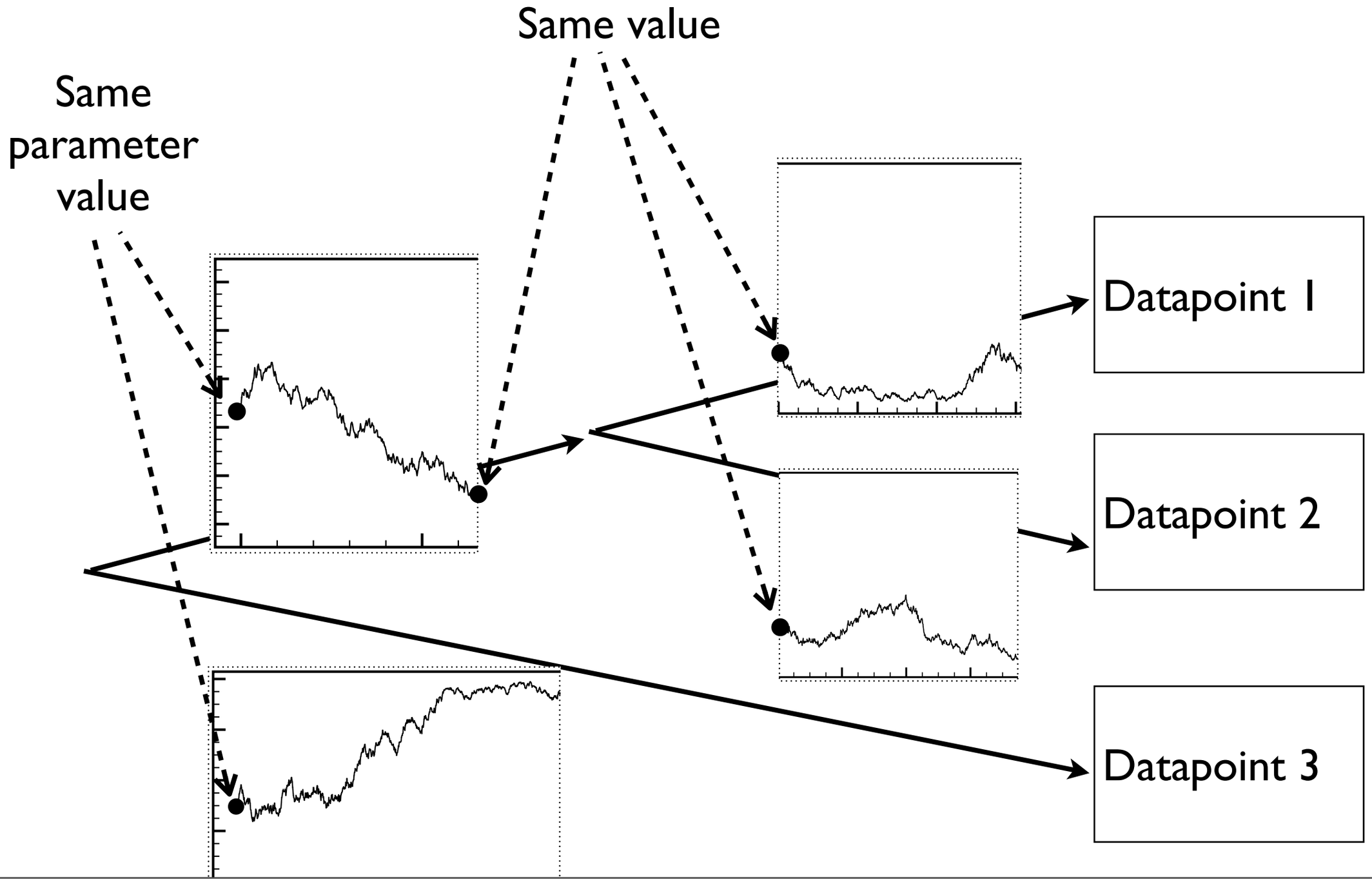
Models over trees: example

Until there is only
one node left



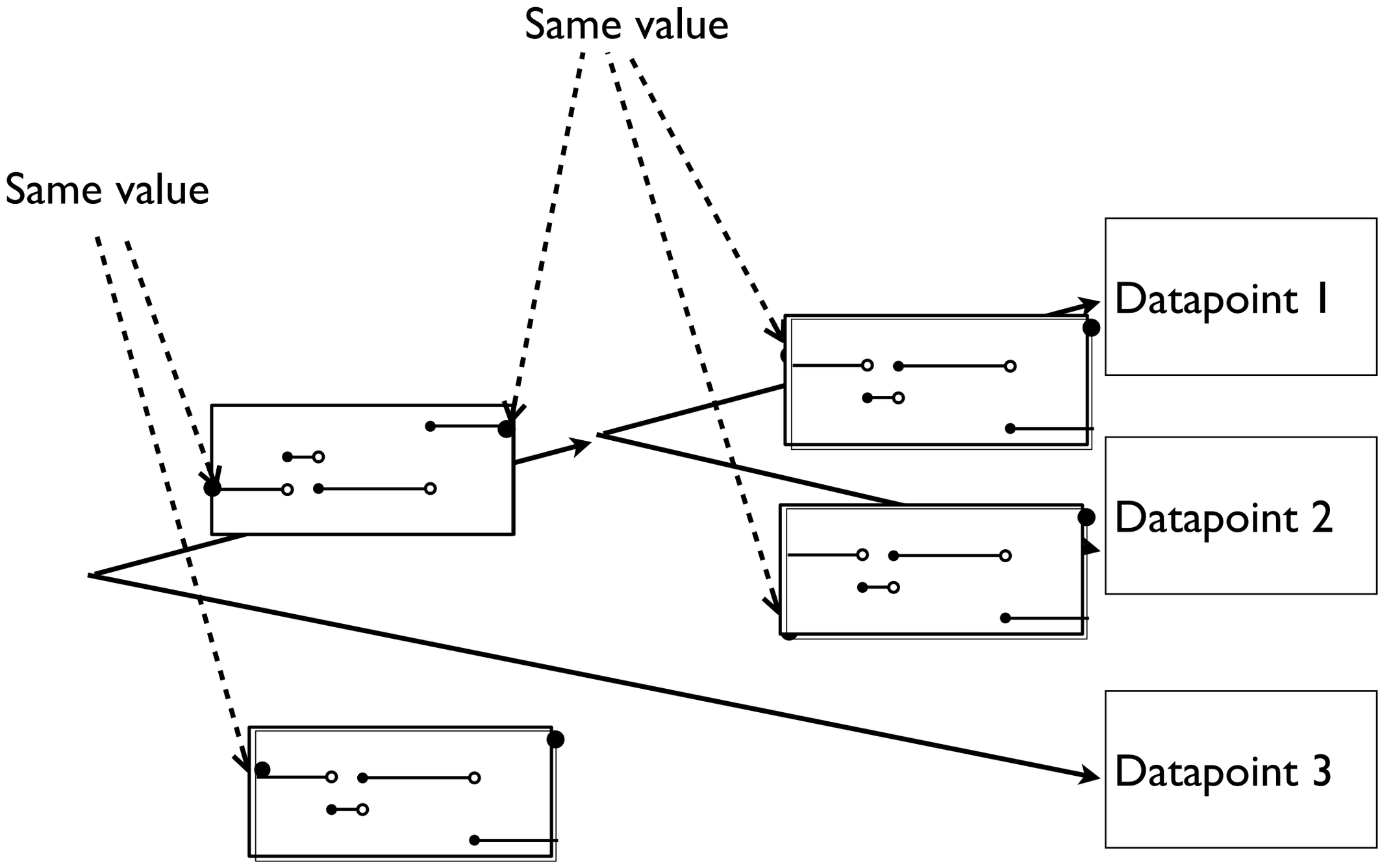
Likelihood model: examples

Continuous data: Brownian motion



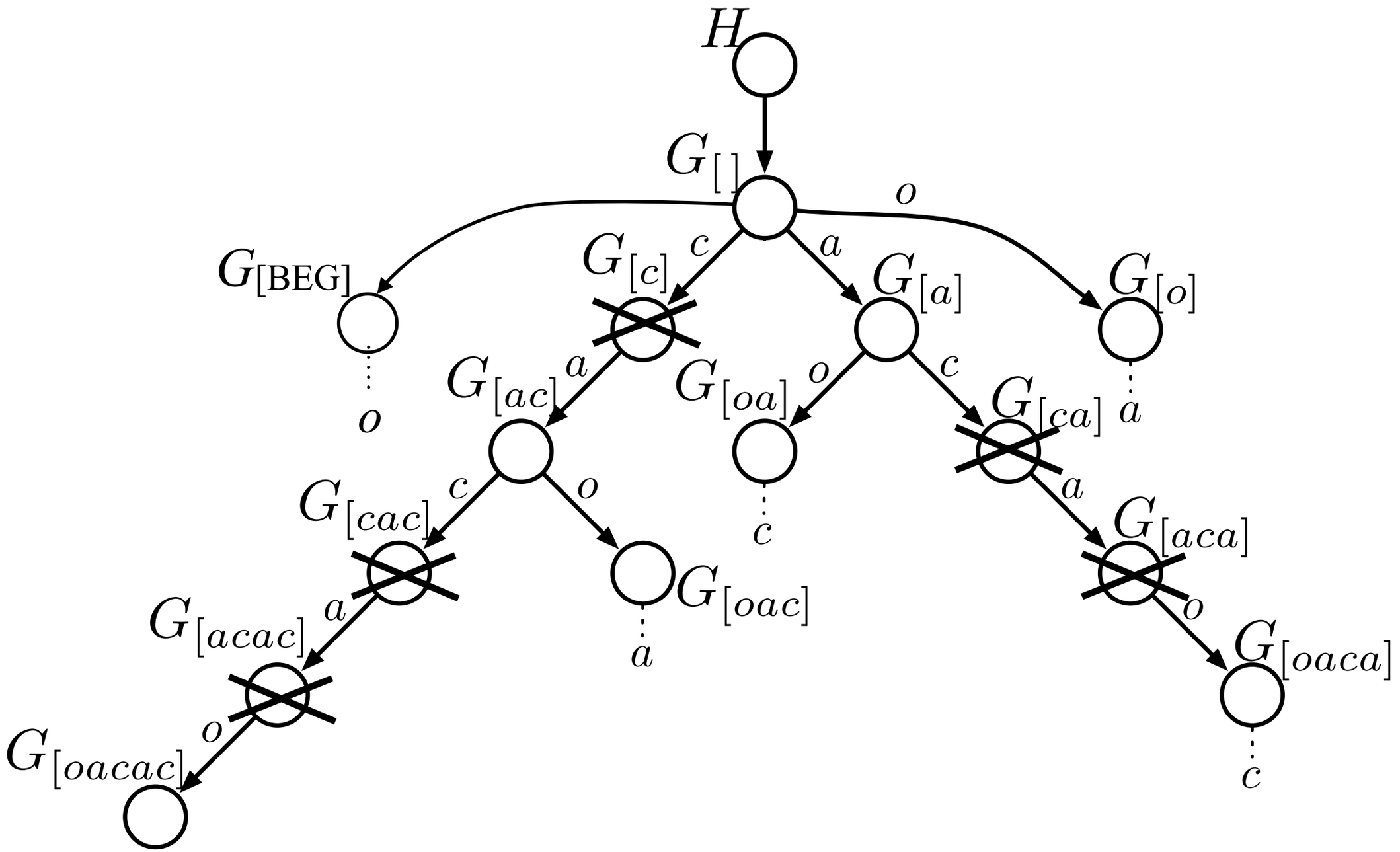
Likelihood model: examples

Discrete data: CTMC



Likelihood model: examples

Infinite dimensional: Coagulation Pitman-Yor?



Lots of other interesting topics!

Other models: nested Dirichlet process, kernel stick breaking process, nested CRP, infinite PCFG, string-valued CTMCs, Gaussian processes, Cox processes, diffusion processes

More general theories: Levy processes, completely random measure, tail-free processes (e.g. Polya tree), neutral to the right

Theoretical issues: consistency, Robins-Ritov paradox, mixing of MCMC samplers

Practical issues: fast, large scale inference; diagnosis