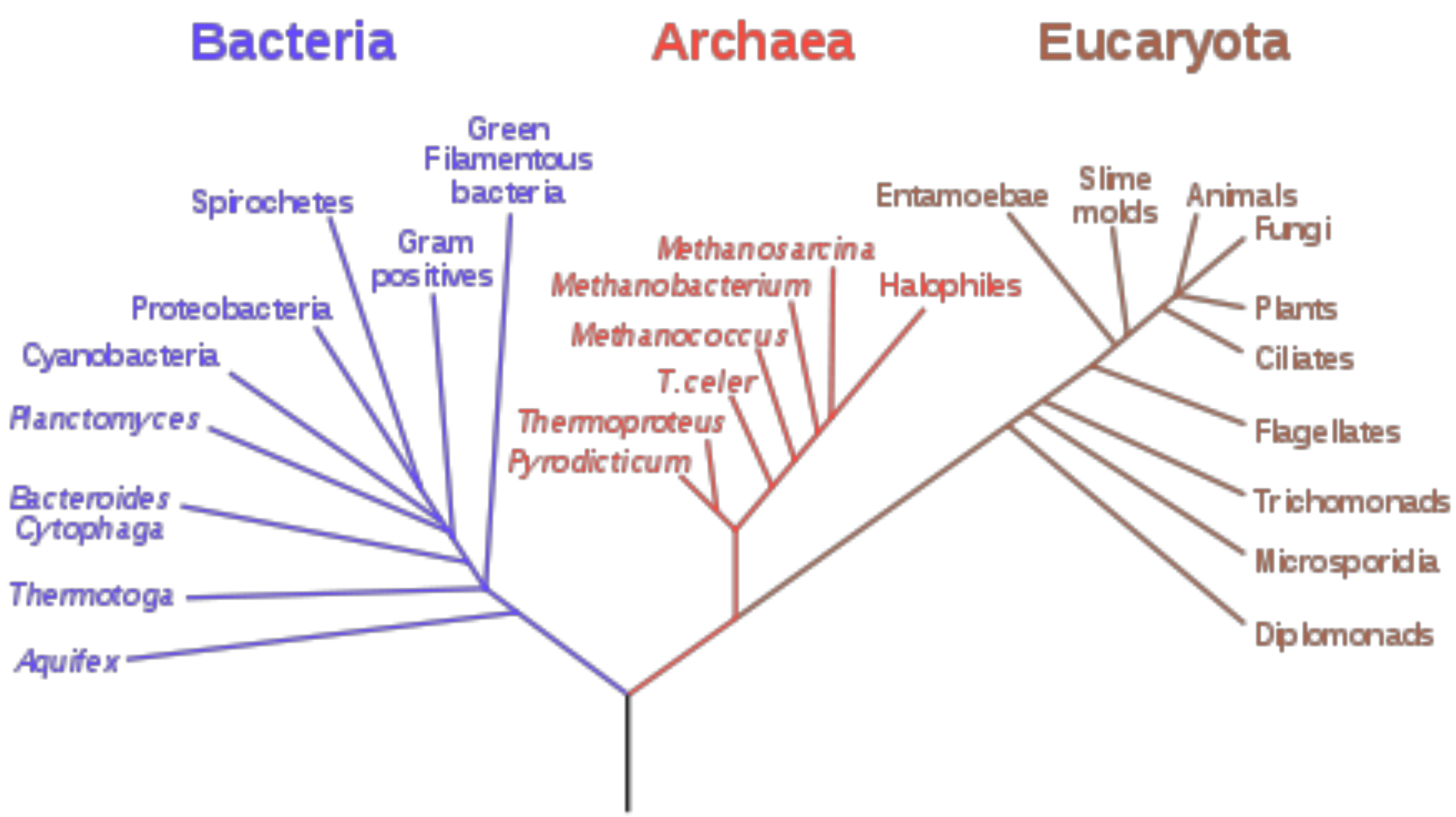
Flexible Priors for Deep Hierarchies

Jacob Steinhardt

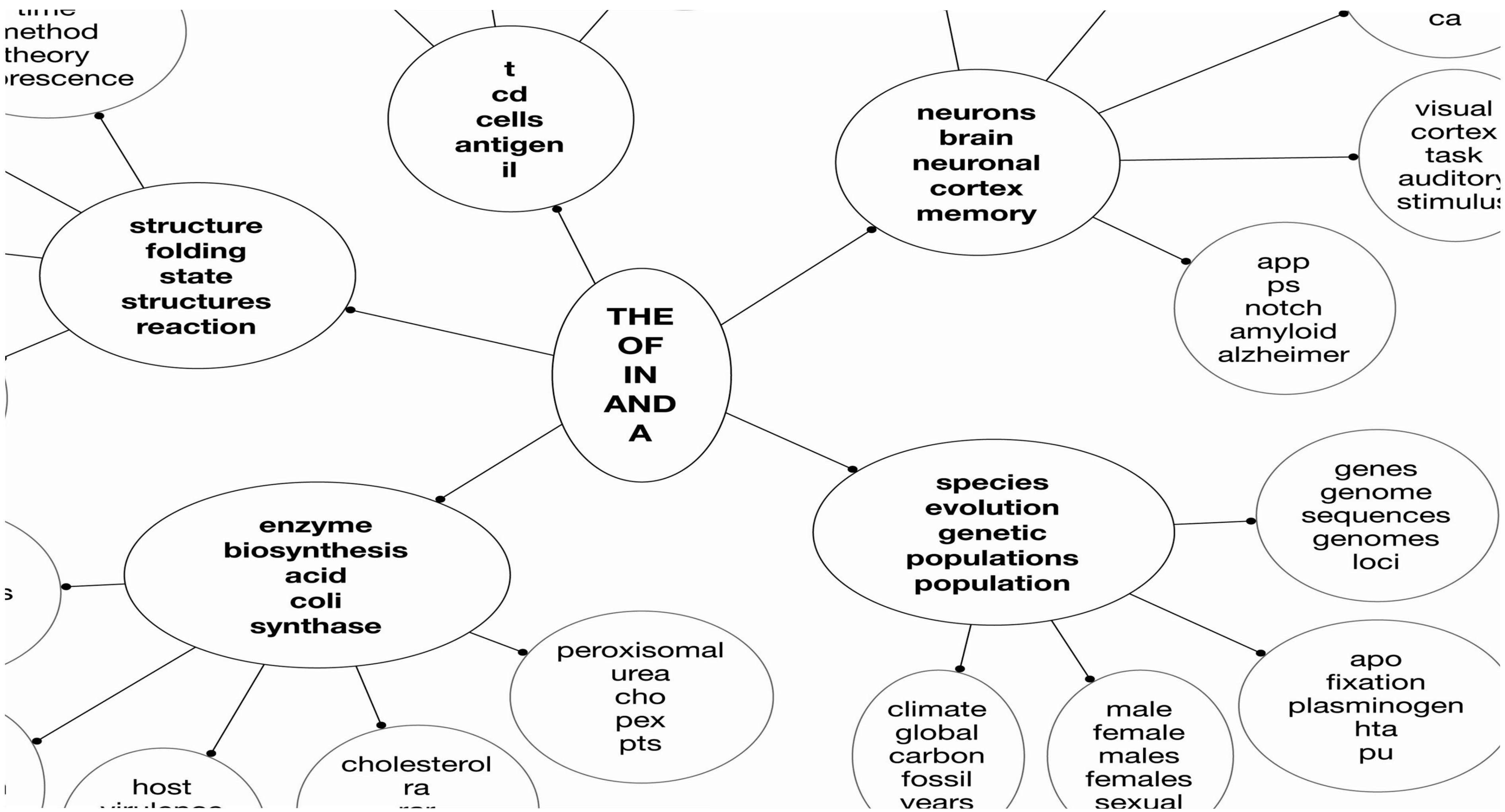
Hierarchical Modeling

- many data are well-modeled by an underlying tree



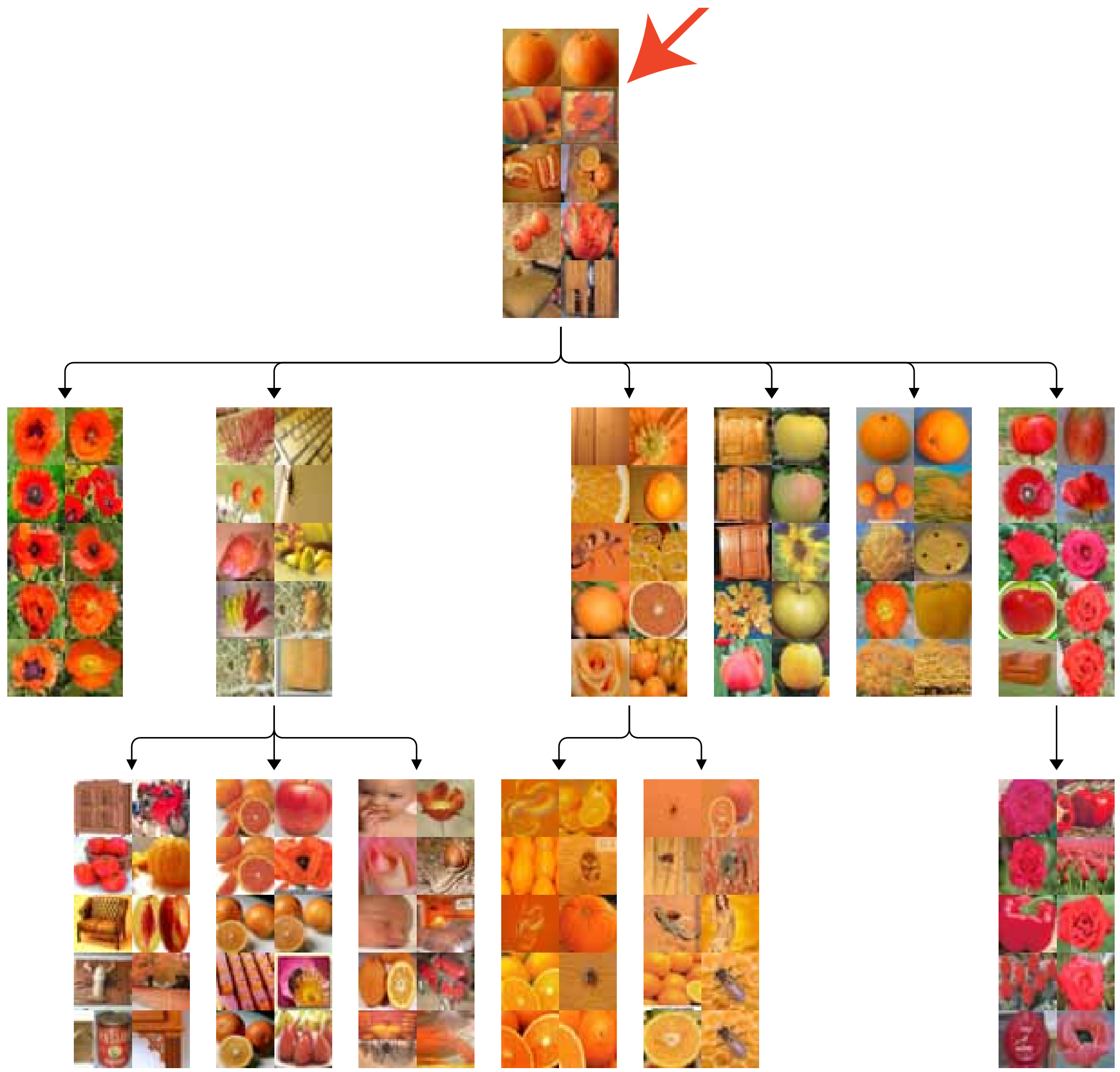
Hierarchical Modeling

- many data are well-modeled by an underlying tree



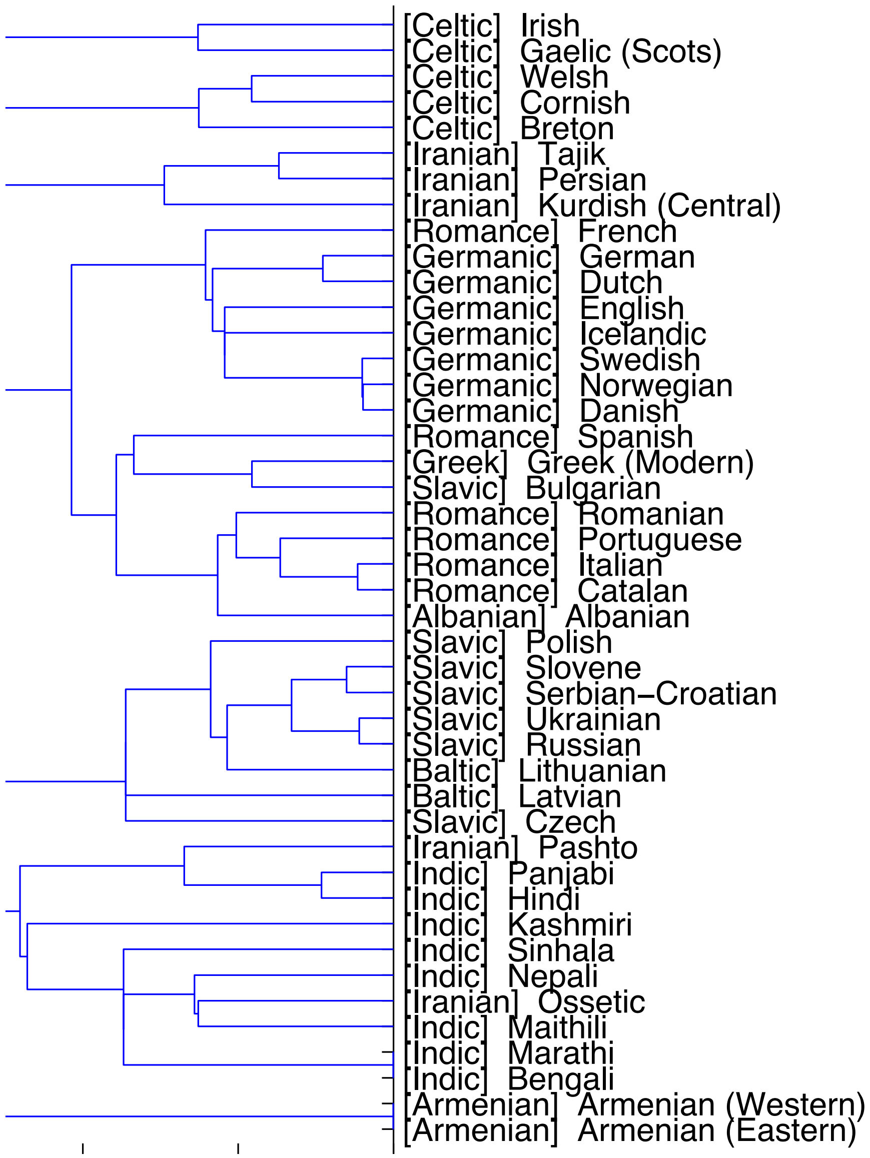
Hierarchical Modeling

- many data are well-modeled by an underlying tree



Hierarchical Modeling

- many data are well-modeled by an underlying tree



Hierarchical Modeling

Hierarchical Modeling

- advantages of hierarchical modeling:

Hierarchical Modeling

- advantages of hierarchical modeling:
  - captures both broad and specific trends

Hierarchical Modeling

- advantages of hierarchical modeling:
  - captures both broad and specific trends
  - facilitates transfer learning

Hierarchical Modeling

- advantages of hierarchical modeling:
  - captures both broad and specific trends
  - facilitates transfer learning
- issues:

Hierarchical Modeling

- advantages of hierarchical modeling:
  - captures both broad and specific trends
  - facilitates transfer learning
- issues:
  - the underlying tree may not be known

Hierarchical Modeling

- advantages of hierarchical modeling:
  - captures both broad and specific trends
  - facilitates transfer learning
- issues:
  - the underlying tree may not be known
  - predictions in deep hierarchies can be strongly influenced by the prior

Learning the Tree

Learning the Tree

- major approaches for choosing a tree:

Learning the Tree

- major approaches for choosing a tree:
  - agglomerative clustering

Learning the Tree

- major approaches for choosing a tree:
  - agglomerative clustering
  - Bayesian methods (place prior over trees)

Learning the Tree

- major approaches for choosing a tree:
  - agglomerative clustering
  - Bayesian methods (place prior over trees)
    - stochastic branching processes

Learning the Tree

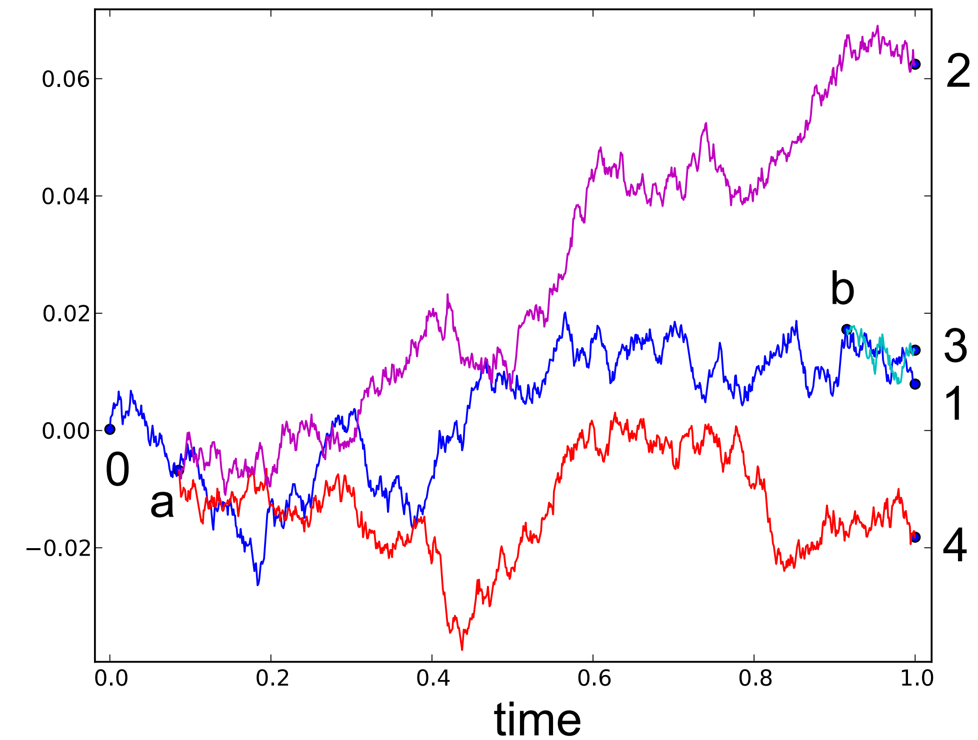
- major approaches for choosing a tree:
  - agglomerative clustering
  - Bayesian methods (place prior over trees)
    - stochastic branching processes
    - nested random partitions

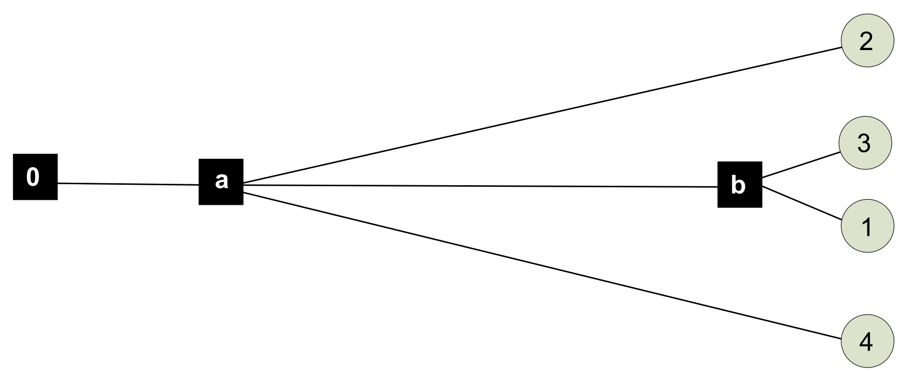
Agglomerative Clustering

- start with each datum in its own subtree
- iteratively merge subtrees based on a similarity metric
- issues:
  - can't add new data
  - can't form hierarchies over latent parameters
  - difficult to incorporate structured domain knowledge

Stochastic Branching Processes

- fully Bayesian model
- data starts at top and branches based on an arrival process (Dirichlet diffusion trees)
- can also start at bottom and merge (Kingman coalescents)





Stochastic Branching Processes

- many nice properties
  - infinitely exchangeable
  - complexity of tree grows with the data
- latent parameters must undergo a continuous-time diffusion process
  - unclear how to construct such a process for models over discrete data

Random Partitions

- stick-breaking process: a way to partition the unit interval into countably many masses $\pi_1, \pi_2, \ldots$
- draw β_k from Beta(1,γ)
- let $\pi_k = \beta_k \times (1-\beta_1) \ldots (1-\beta_{k-1})$
- the distribution over the π_k is called a *Dirichlet process*

Random Partitions

- suppose $\{\pi_k\}_{k=1,...,\infty}$ are drawn from a Dirichlet process
- for n=1,..,N, let X_n ~ Multinomial($\{\pi_k\}$)
- induces distribution over partitions of {1,...,N}
- given partition of {1,...,N}, add X_{N+1} to a part of size s with probability $s/(N+\gamma)$ and to a new part with probability $\gamma/(N+\gamma)$
  - *Chinese restaurant process*

Nested Random Partitions

- a tree is equivalent to a collection of nested partitions
- nested tree <=> nested random partitions
- partition at each node given by Chinese restaurant process
- issue: when to stop recursing?

Martingale Property

- martingale property:

$$E[f(\theta_{child}) \mid \theta_{parent}] = f(\theta_{parent})$$

- implies $E[f(\theta_v) \mid \theta_u] = f(\theta_u)$ for any ancestor u of v
- says that learning about a child does not change beliefs in expectation

Doob’s Theorem

Doob's Theorem

- Let $\theta_1, \theta_2, \ldots$ be a sequence of random variables such that $E[f(\theta_{n+1}) \mid \theta_n] = f(\theta_n)$ and $\sup_n E[|\theta_n|] < \infty$.

Doob's Theorem

- Let $\theta_1, \theta_2, \ldots$ be a sequence of random variables such that $E[f(\theta_{n+1}) \mid \theta_n] = f(\theta_n)$ and $\sup_n E[|\theta_n|] < \infty$.
- Then $\lim_{n\to\infty} f(\theta_n)$ exists with probability 1.

Doob's Theorem

- Let $\theta_1, \theta_2, \ldots$ be a sequence of random variables such that $E[f(\theta_{n+1}) \mid \theta_n] = f(\theta_n)$ and $\sup_n E[|\theta_n|] < \infty$.
- Then $\lim_{n\to\infty} f(\theta_n)$ exists with probability 1.
- Intuition: each new random variable reveals more information about $f(\theta)$ until it is completely determined.

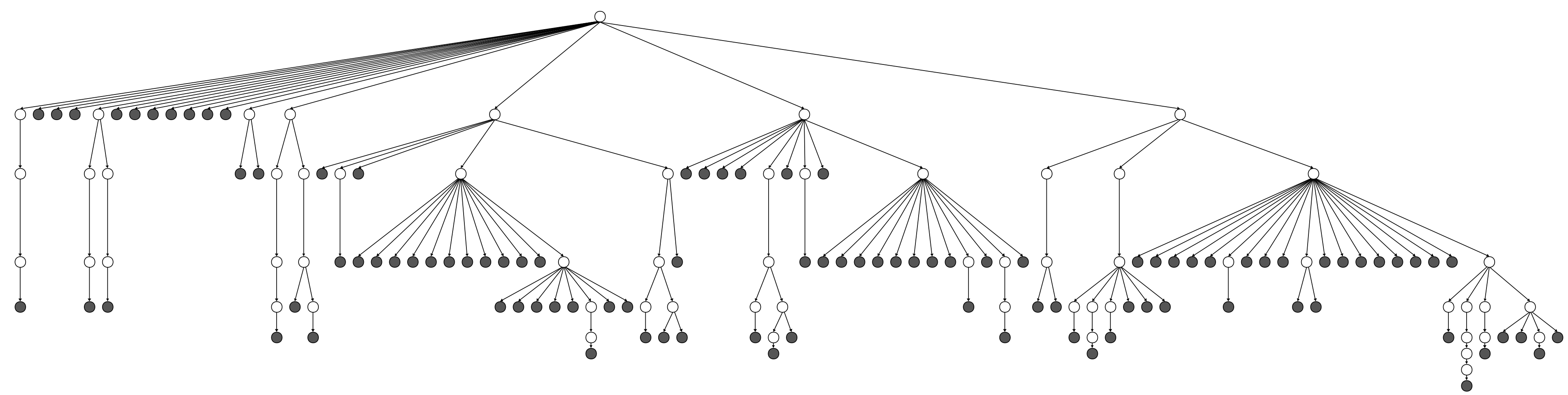
Doob's Theorem

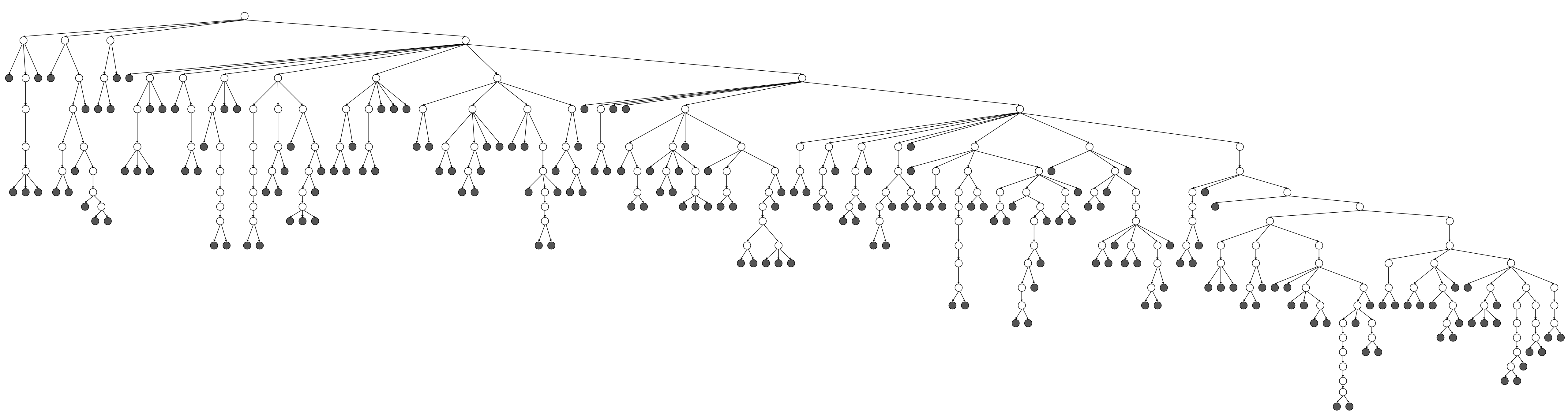
- Use Doob's theorem to build infinitely deep hierarchy
  - data associated with infinite paths $v_1, v_2, \ldots$ down the tree
  - each datum drawn from distribution parameterized by $\lim_n f(\theta_{v_n})$

Doob's Theorem

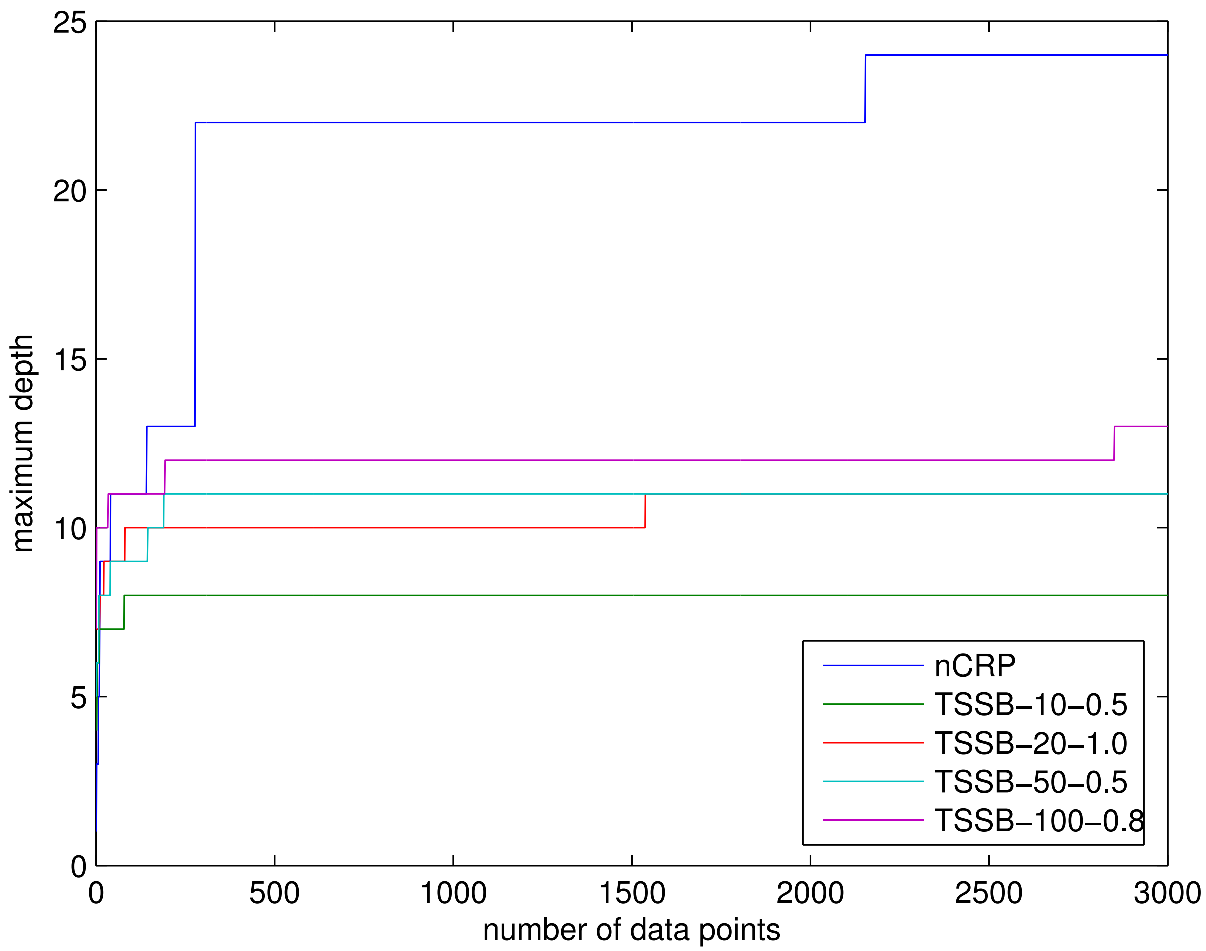
- all data have infinite depth
- can think of effective depth of a datum as first point where it is in a unique subtree
- effective depth is O(logN)

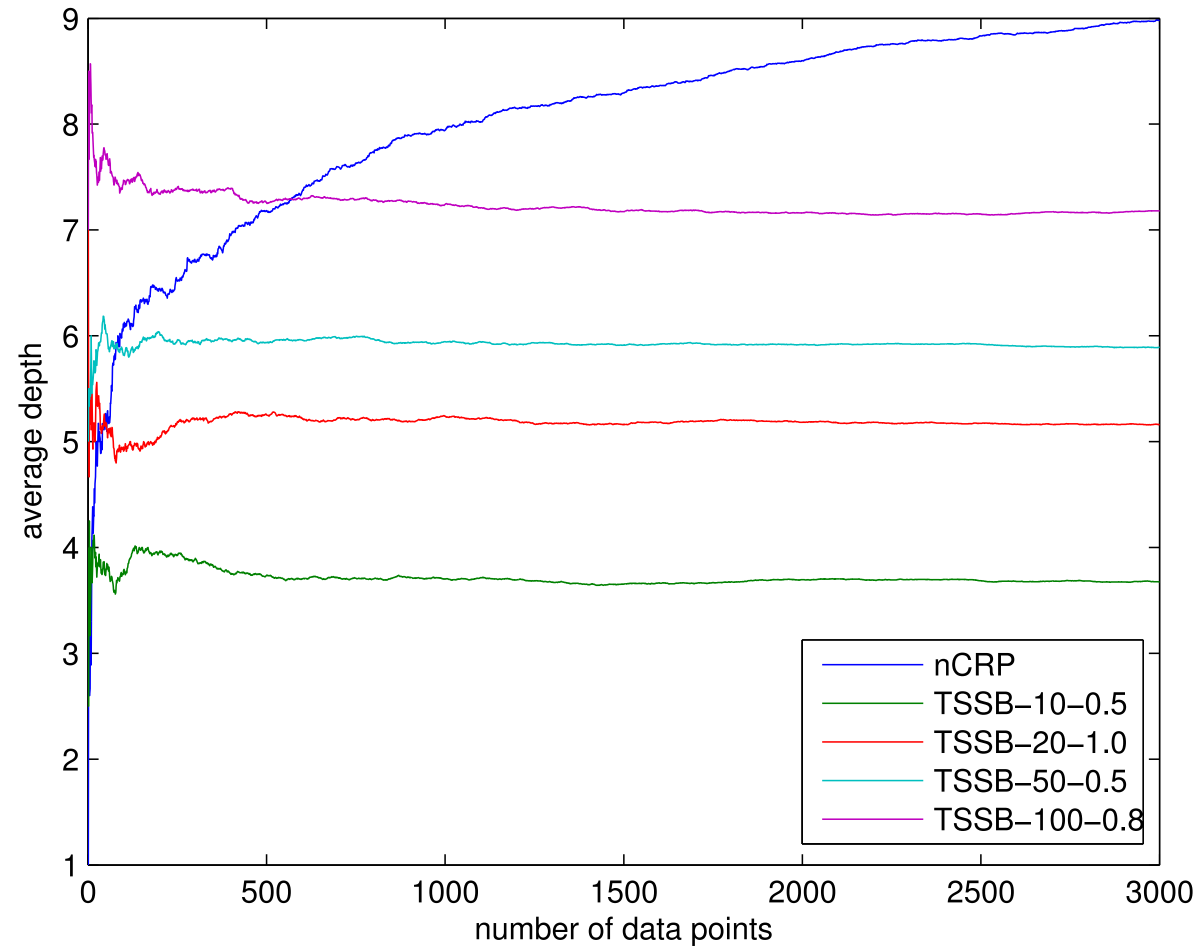
Letting the Complexity Grow with the Data





Letting the Complexity Grow with the Data





Hierarchical Beta Processes

Hierarchical Beta Processes

- θ_v lies in $[0,1]^D$

Hierarchical Beta Processes

- θ_v lies in $[0,1]^D$
- $\theta_{v,d} \mid \theta_{p(v),d} \sim \mathrm{Beta}(c\theta_{p(v),d}, c(1-\theta_{p(v),d}))$

Hierarchical Beta Processes

- θ_v lies in $[0,1]^D$
- $\theta_{v,d} \mid \theta_{p(v),d} \sim \mathrm{Beta}(c\theta_{p(v),d}, c(1-\theta_{p(v),d}))$
- martingale property for $f(\theta_v) = \theta_v$

Hierarchical Beta Processes

- θ_v lies in $[0,1]^D$
- $\theta_{v,d} \mid \theta_{p(v),d} \sim \text{Beta}(c\theta_{p(v),d}, c(1-\theta_{p(v),d}))$
- martingale property for $f(\theta_v) = \theta_v$
    - let θ denote the limit

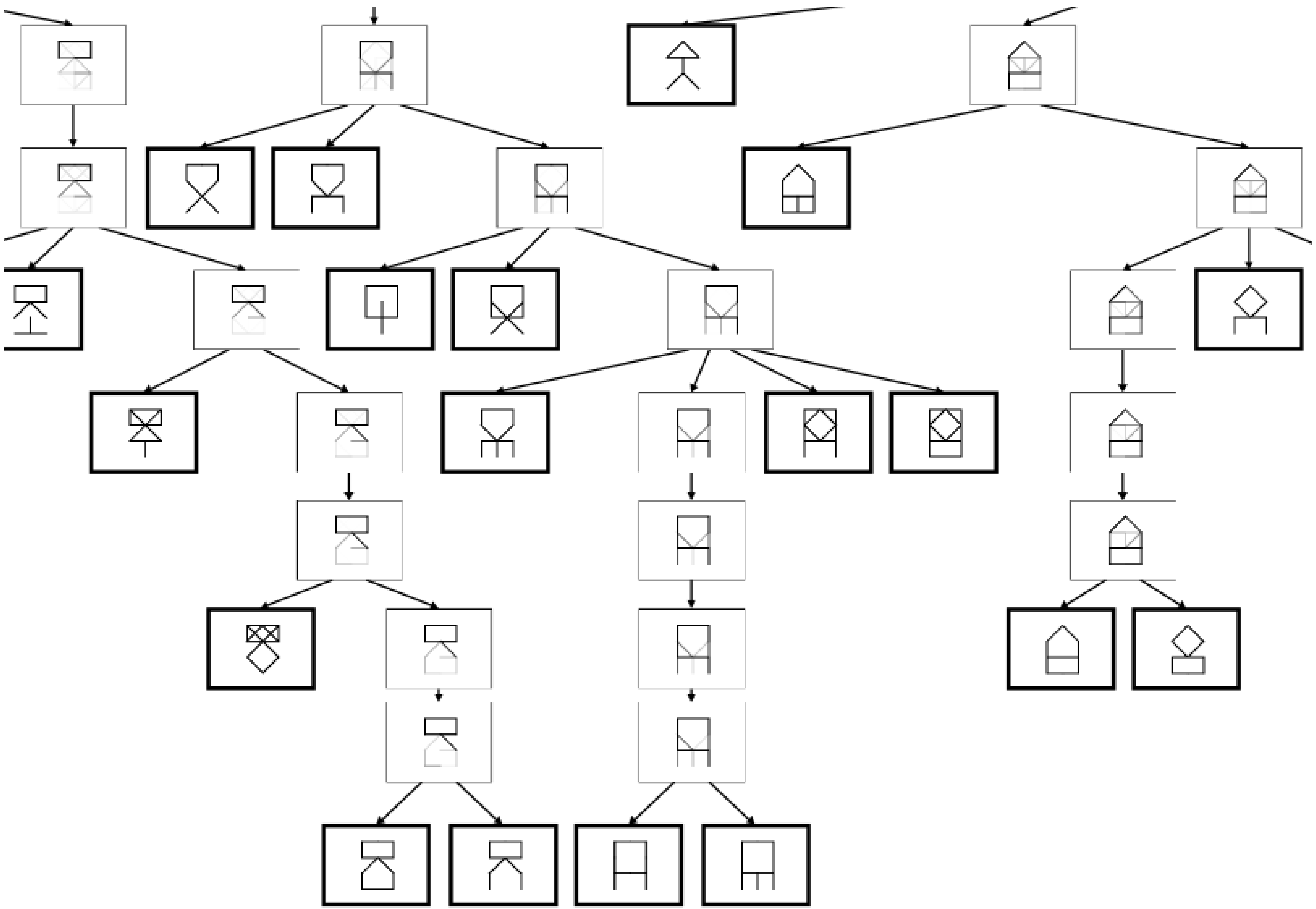
Hierarchical Beta Processes

- θ_v lies in $[0,1]^D$
- $\theta_{v,d} \mid \theta_{p(v),d} \sim \mathrm{Beta}(c\theta_{p(v),d}, c(1-\theta_{p(v),d}))$
- martingale property for $f(\theta_v) = \theta_v$
  - let θ denote the limit
- $X_d \mid \theta_d \sim \mathrm{Bernoulli}(\theta_d)$, where θ is the limit

Hierarchical Beta Processes

- θ_v lies in $[0,1]^D$
- $\theta_{v,d} \mid \theta_{p(v),d} \sim \mathrm{Beta}(c\theta_{p(v),d}, c(1-\theta_{p(v),d}))$
- martingale property for $f(\theta_v) = \theta_v$
  - let θ denote the limit
- $X_d \mid \theta_d \sim \mathrm{Bernoulli}(\theta_d)$, where θ is the limit
  - note that $X_d \mid \theta_{v,d} \sim \mathrm{Bernoulli}(\theta_{v,d})$ as well

Hierarchical Beta Processes



Priors for Deep Hierarchies

- for HBP, $\theta_{v,d}$ converges to 0 or 1
- rate of convergence: tower of exponentials

$$e^{e^{e^{e^{\cdots}}}}$$

- numerical issues + philosophically troubling

Priors for Deep Hierarchies

- inverse Wishart time-series
  - $\Sigma_{n+1} \mid \Sigma_n \sim \mathrm{InvW}(\Sigma_n)$
- converges to 0 with probability 1
- becomes singular to numerical precision
  - rate also given by tower of exponentials

Priors for Deep Hierarchies

- fundamental issues with iterated gamma distribution
  - $\theta_{n+1} \mid \theta_n \sim \Gamma(\theta_n)$
- instead, do $\theta_{n+1} \mid \theta_n \sim c\theta_n + d\phi_n$
  - $\phi_n \sim \Gamma(\theta_n)$

Priors for Deep Hierarchies