A Supertag-Context Model for Weakly-Supervised CCG Parser Learning

Dan Garrette	U. Washington
Chris Dyer	CMU
Jason Baldridge	UT-Austin
Noah A. Smith	CMU

Contributions

1. A **new generative model** for learning CCG parsers from *weak supervision*

2. A way to select Bayesian **priors** that capture properties of CCG

3. A Bayesian **inference procedure** to learn the parameters of our model

Type-Level Supervision

- Unannotated text
- Incomplete tag dictionary: word $\mapsto$ {tags}

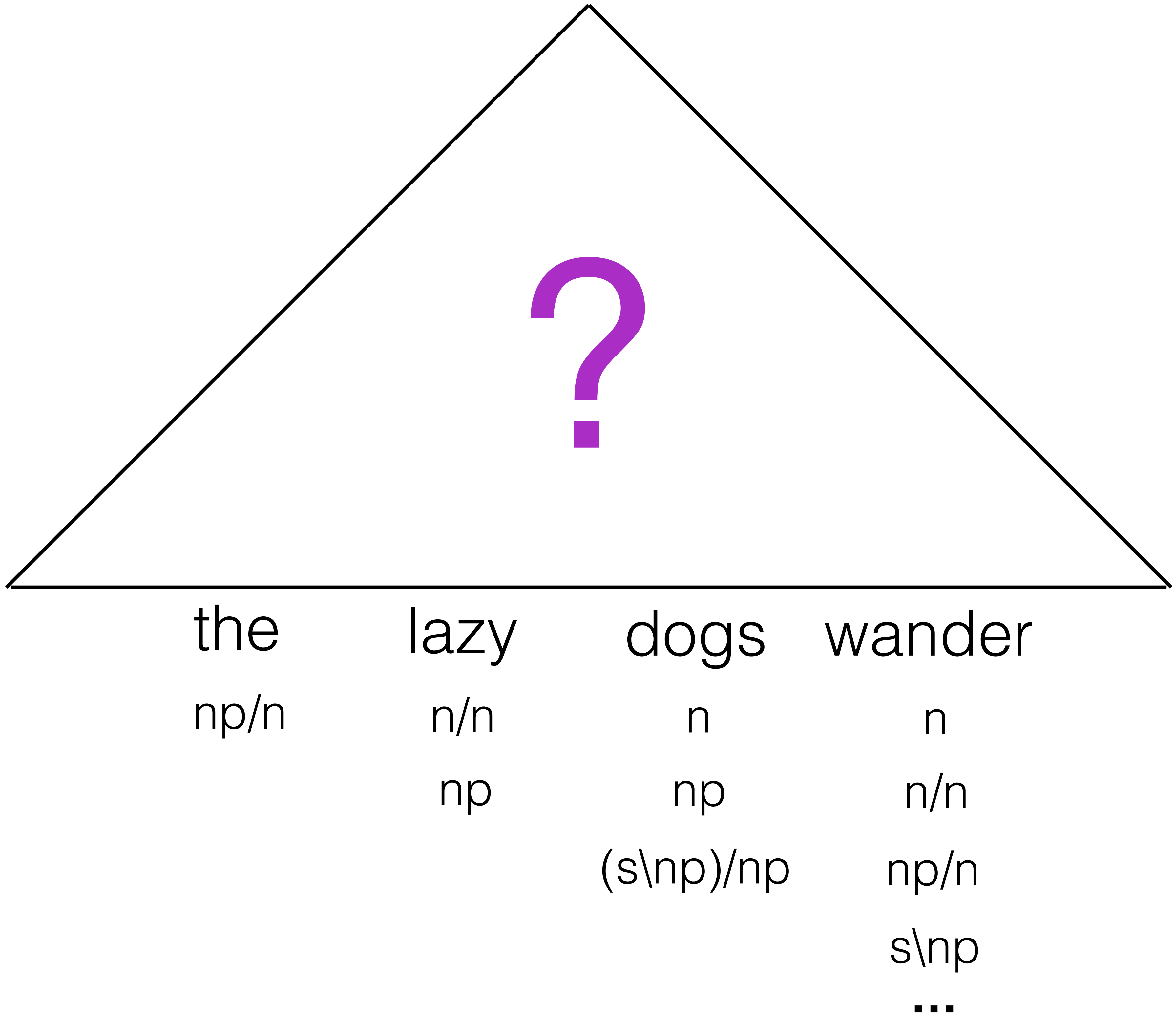
Type-Level Supervision

the	lazy	dogs	wander
np/n	n/n	n	
	np	np	
		(s\np)/np	

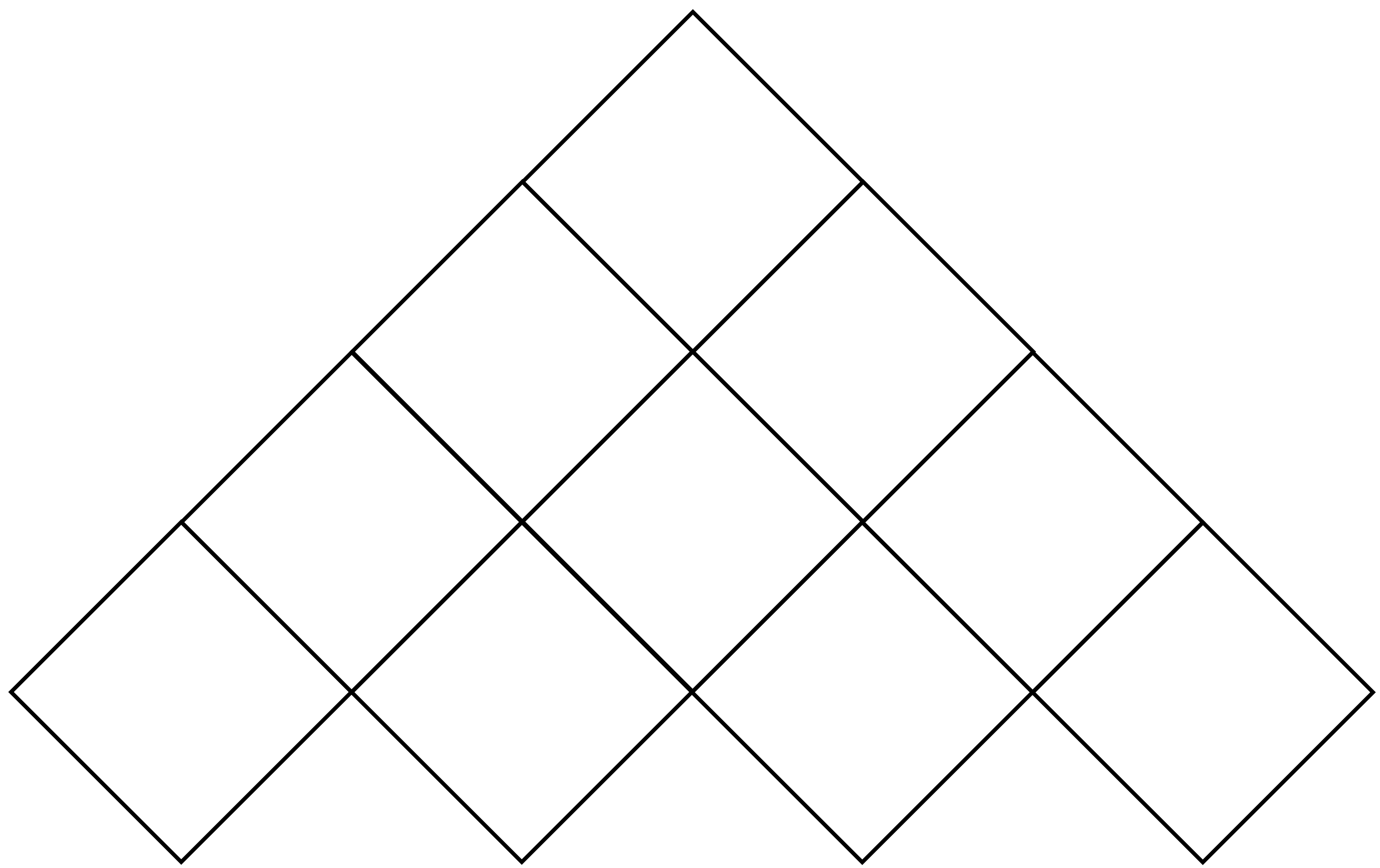
Type-Level Supervision

the	lazy	dogs	wander
np/n	n/n	n	n
	np	np	n/n
		(s\np)/np	np/n
			s\np
			...

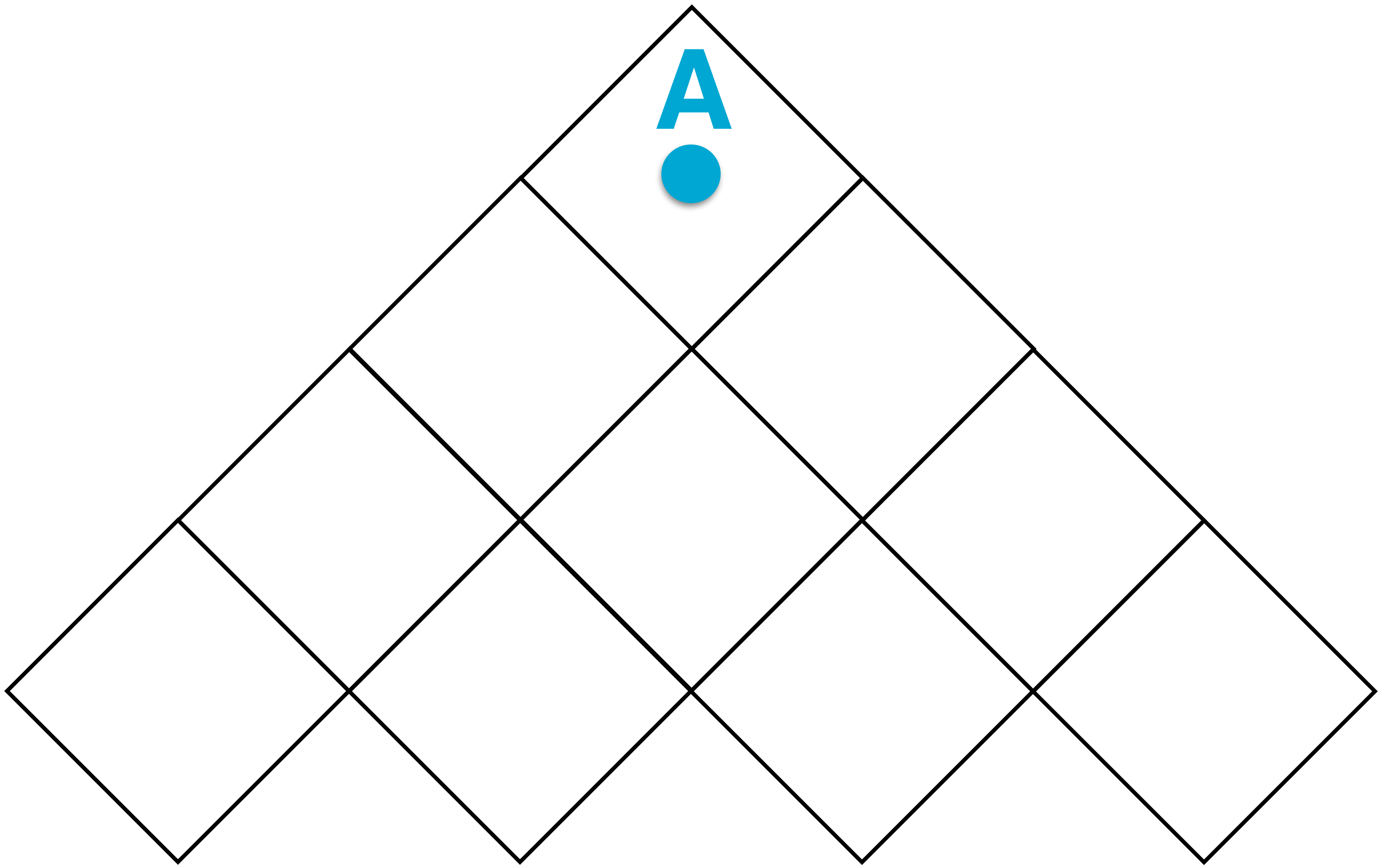
Type-Level Supervision



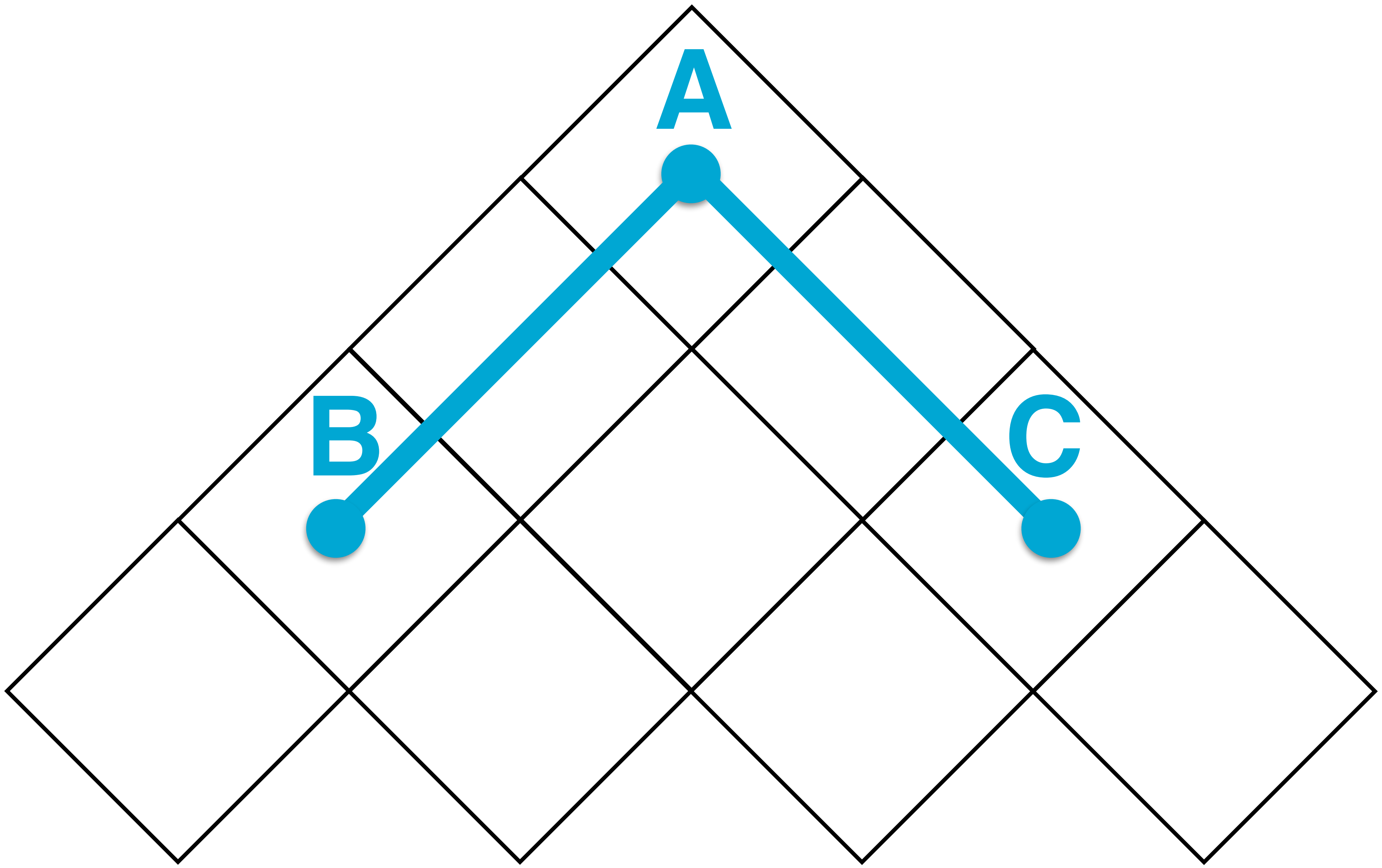
PCFG: Local Decisions



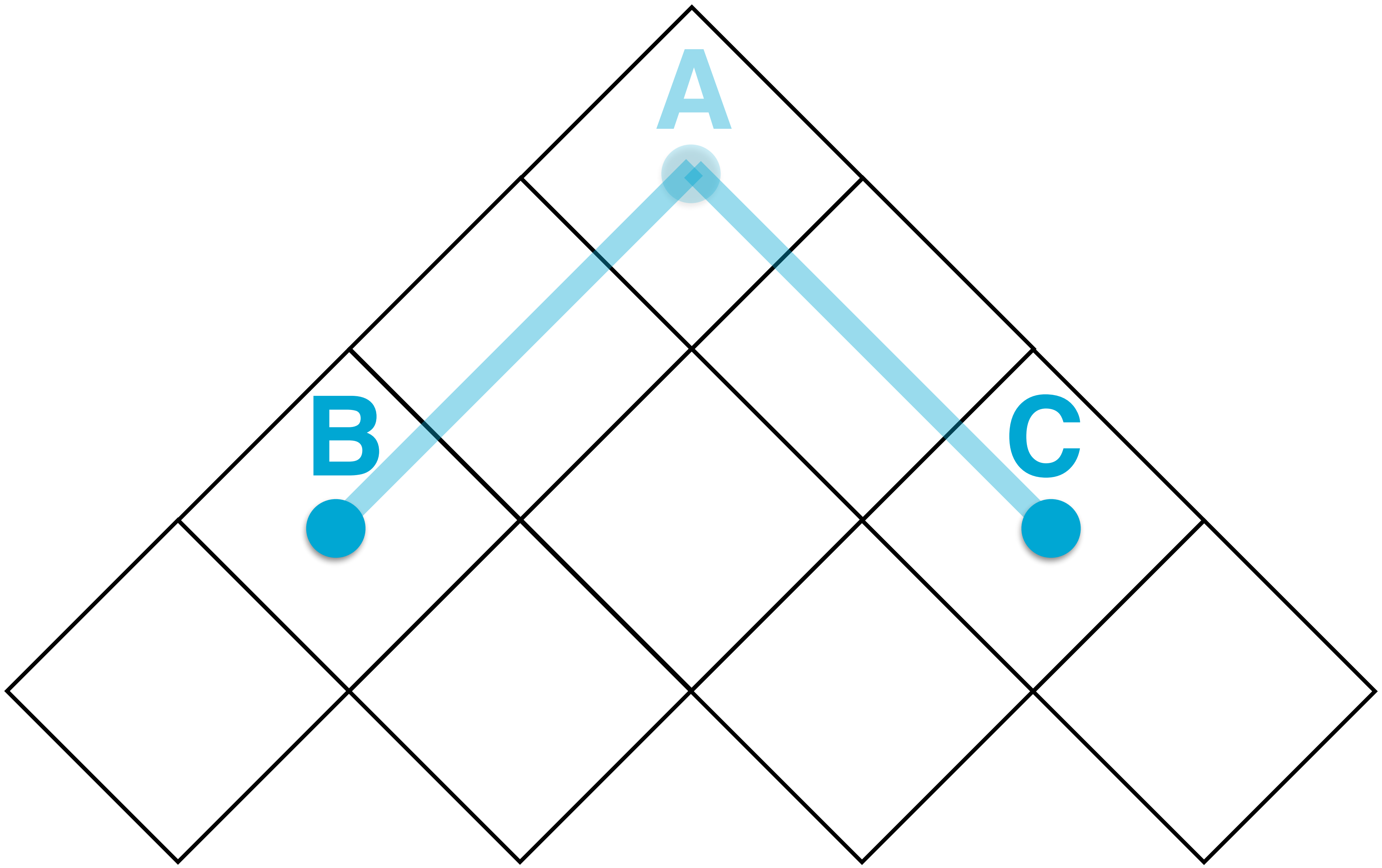
PCFG: Local Decisions



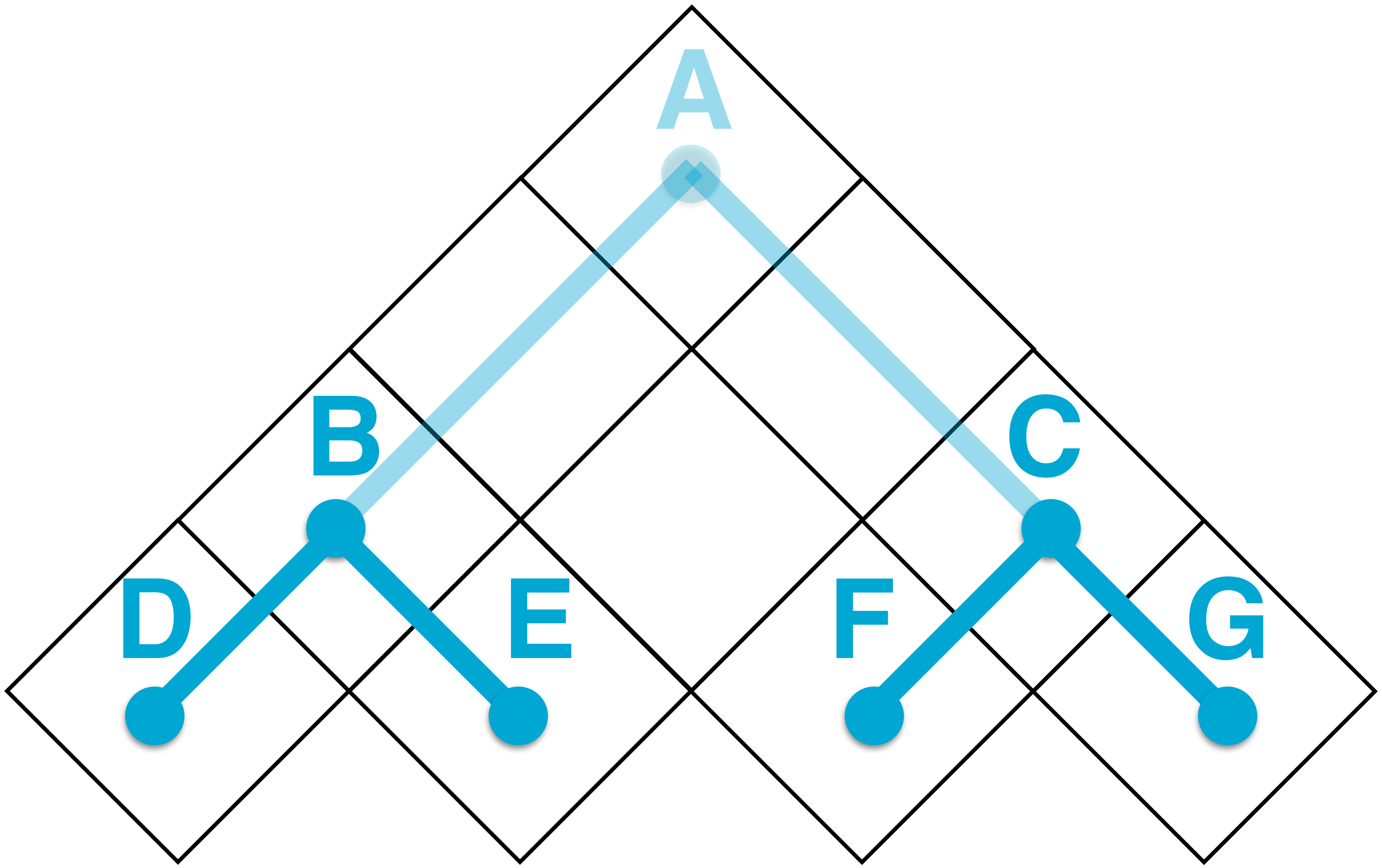
PCFG: Local Decisions



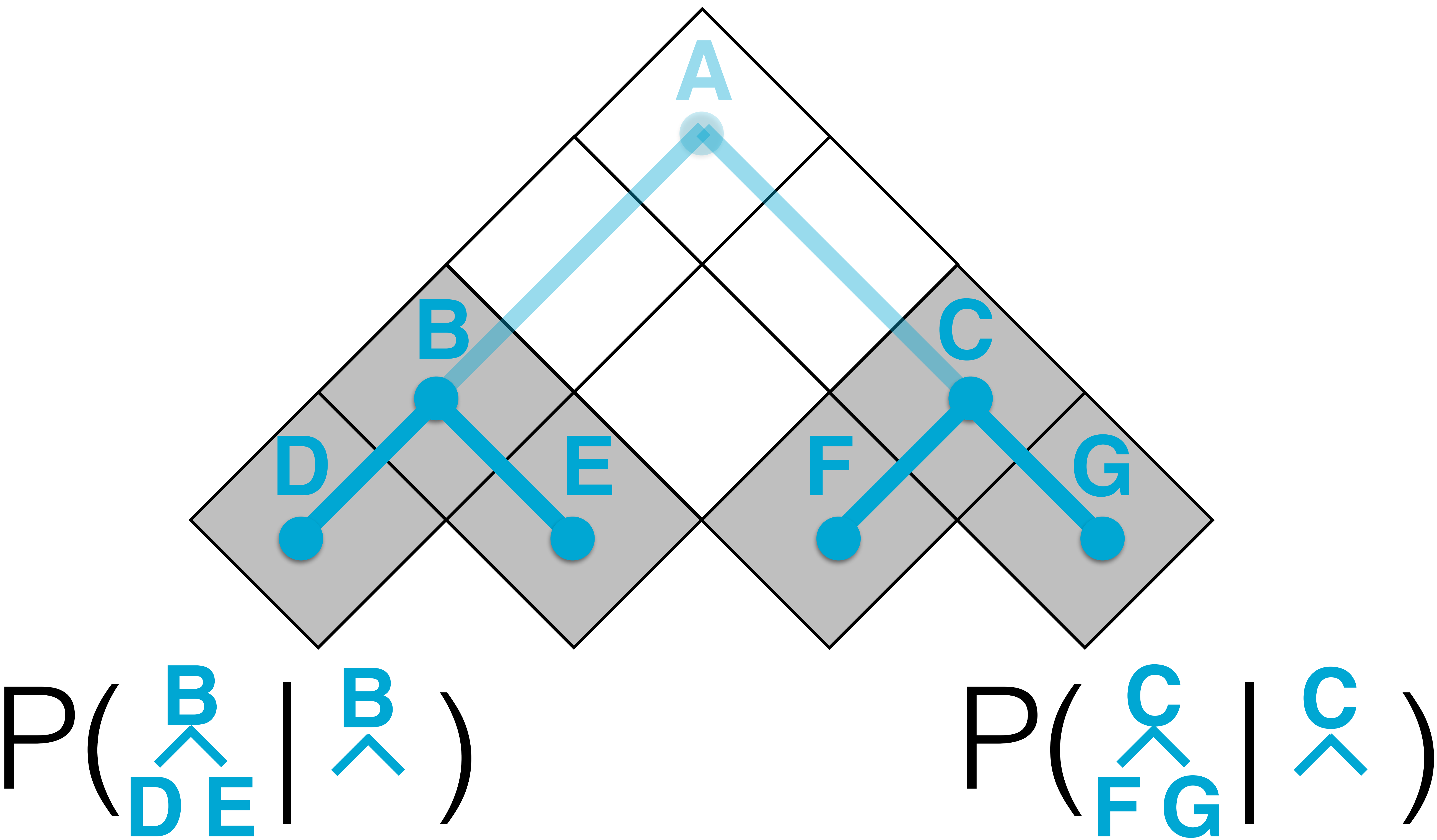
PCFG: Local Decisions



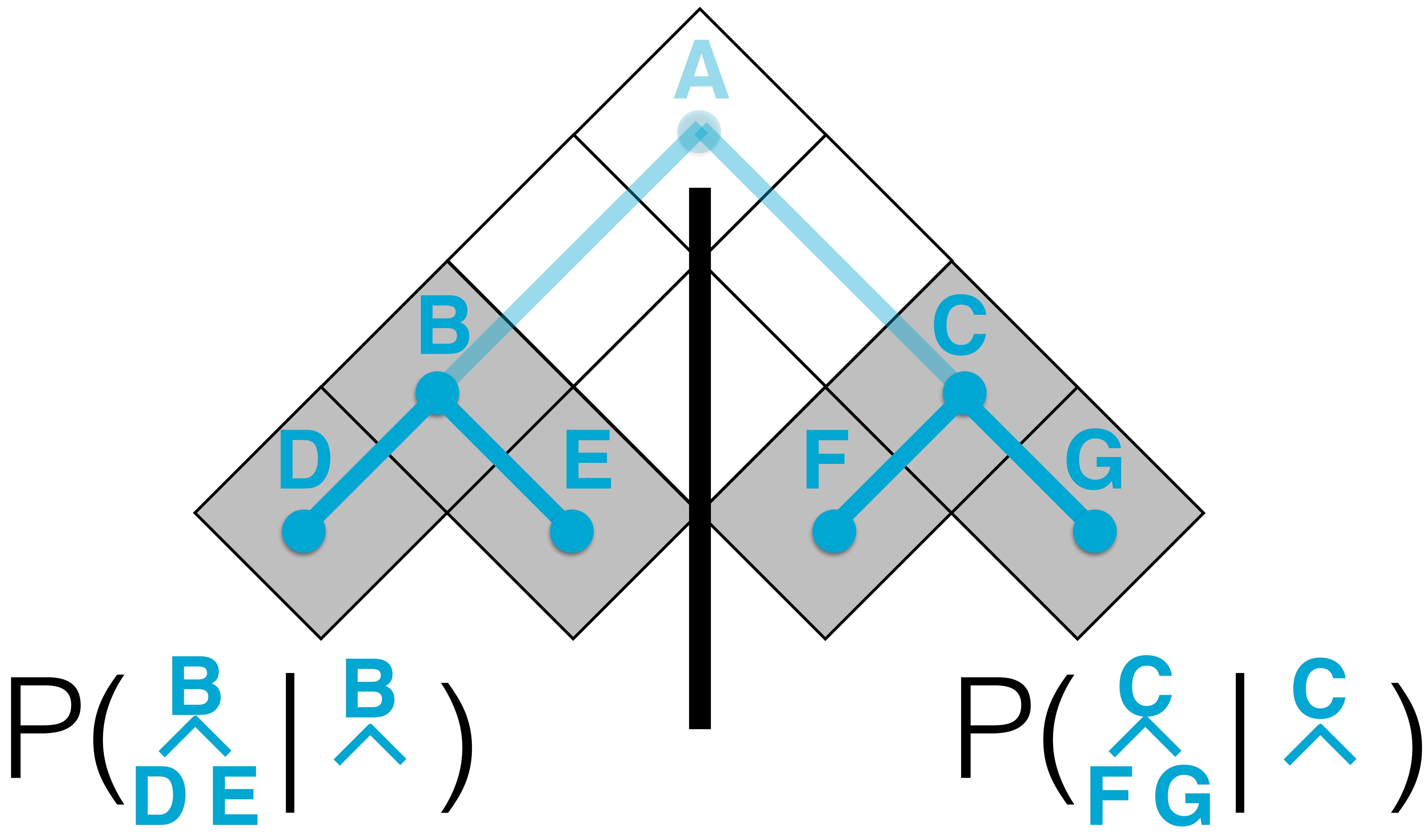
PCFG: Local Decisions



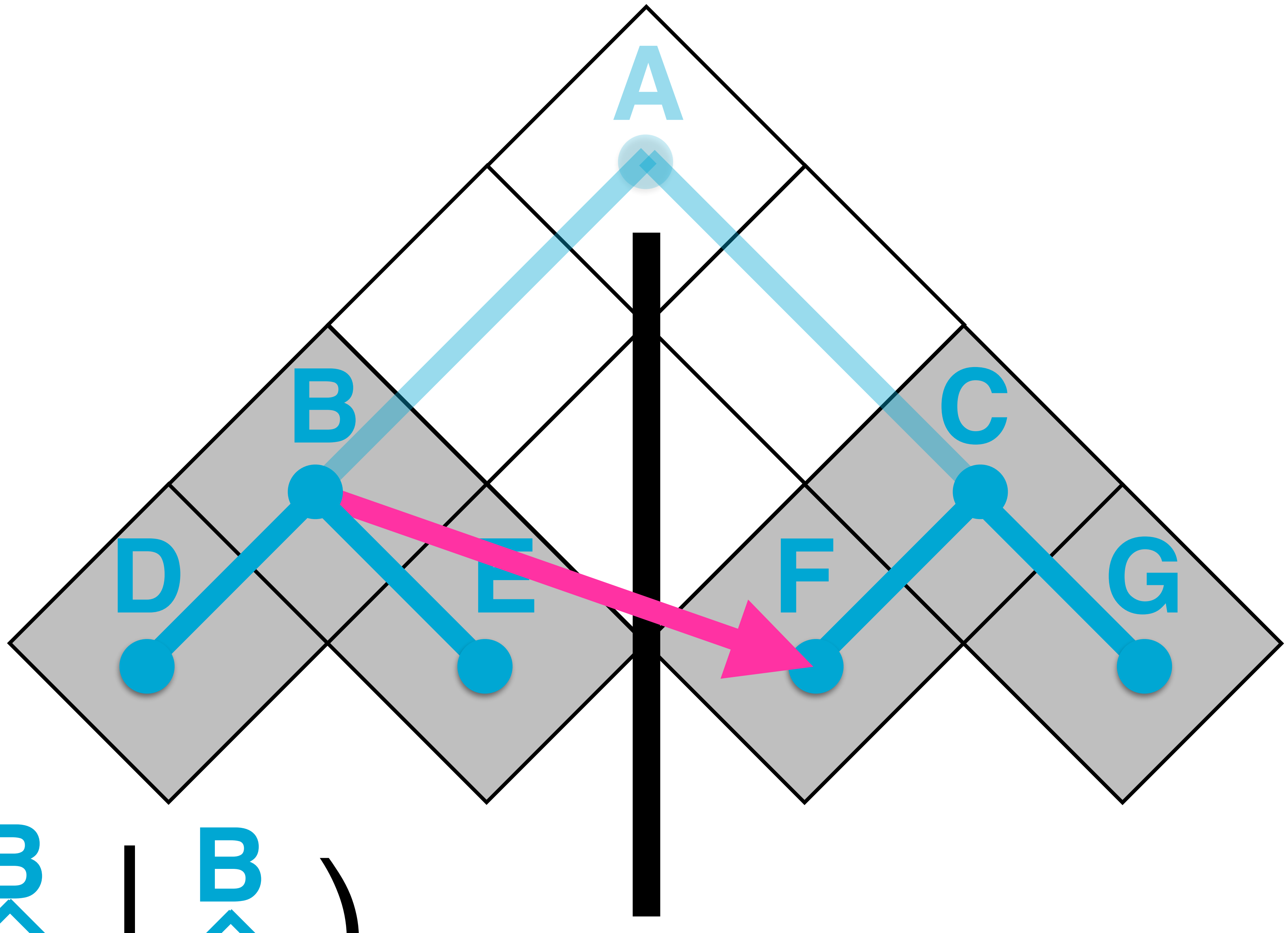
PCFG: Local Decisions



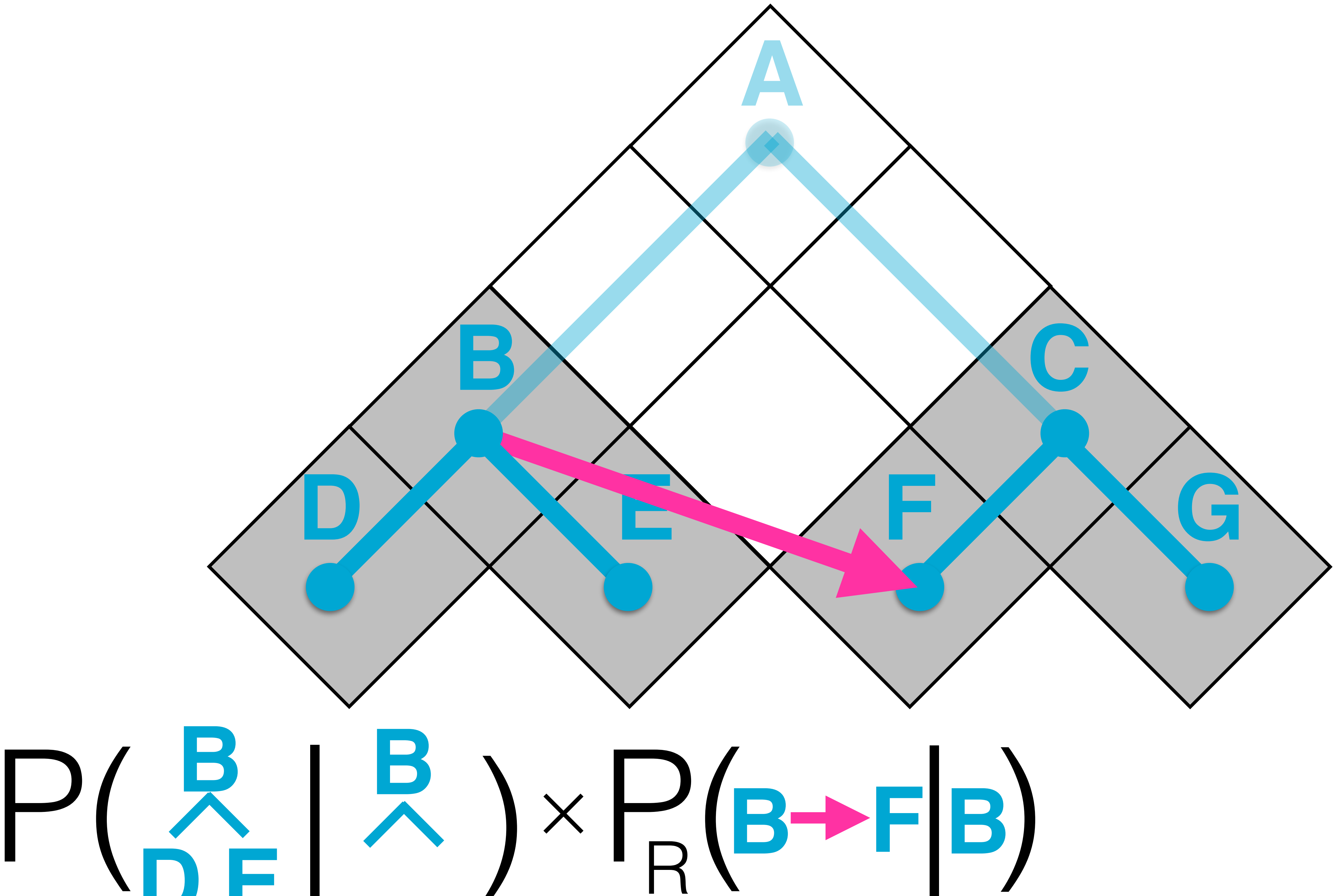
PCFG: Local Decisions



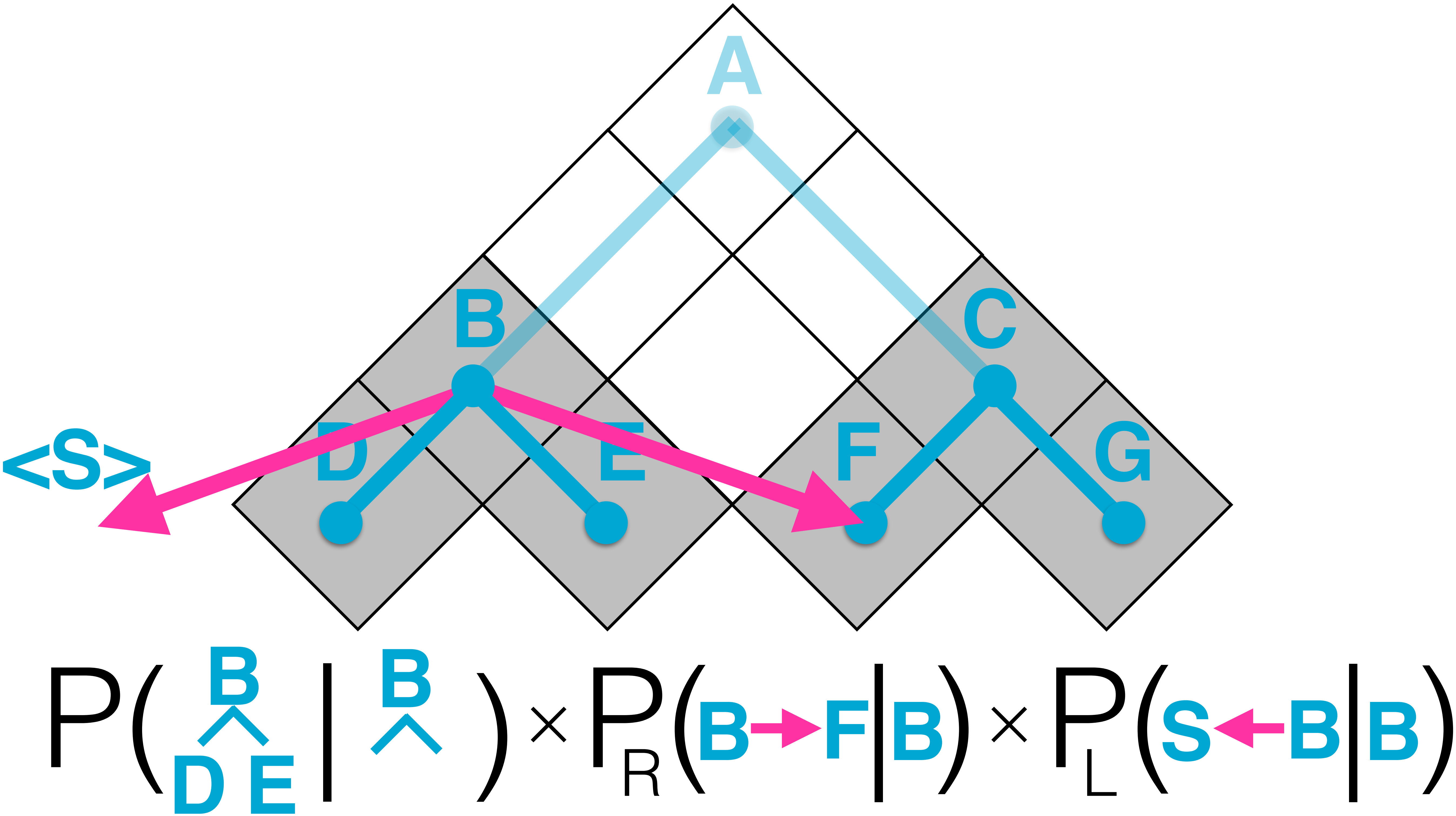
A New Generative Model



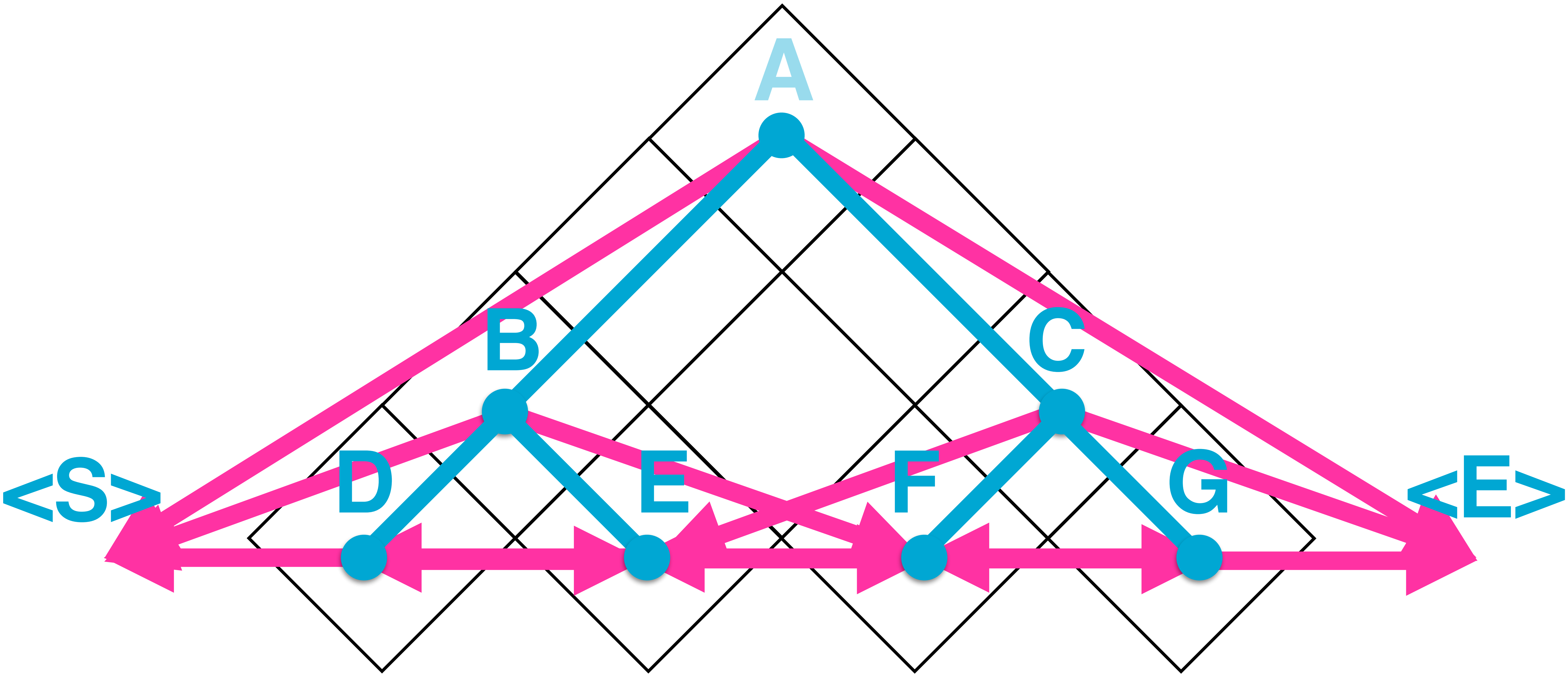
A New Generative Model



A New Generative Model



A New Generative Model

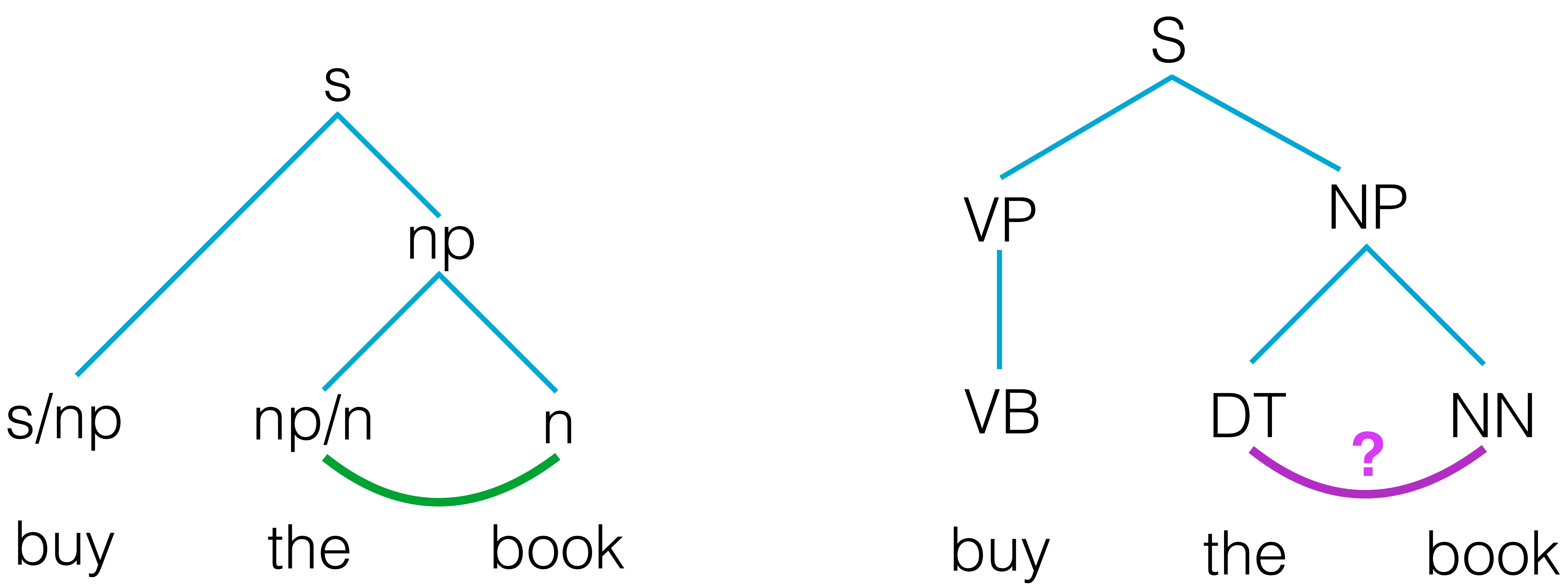


(This makes inference tricky… we'll come back to that)

Why CCG?

- The grammar formalism *itself* can be used to guide learning
    - Given any two categories, we always know whether they are combinable.
- We can extract *a priori* context preferences, before we even look at the data
    - Adjacent categories *tend* to be combinable.

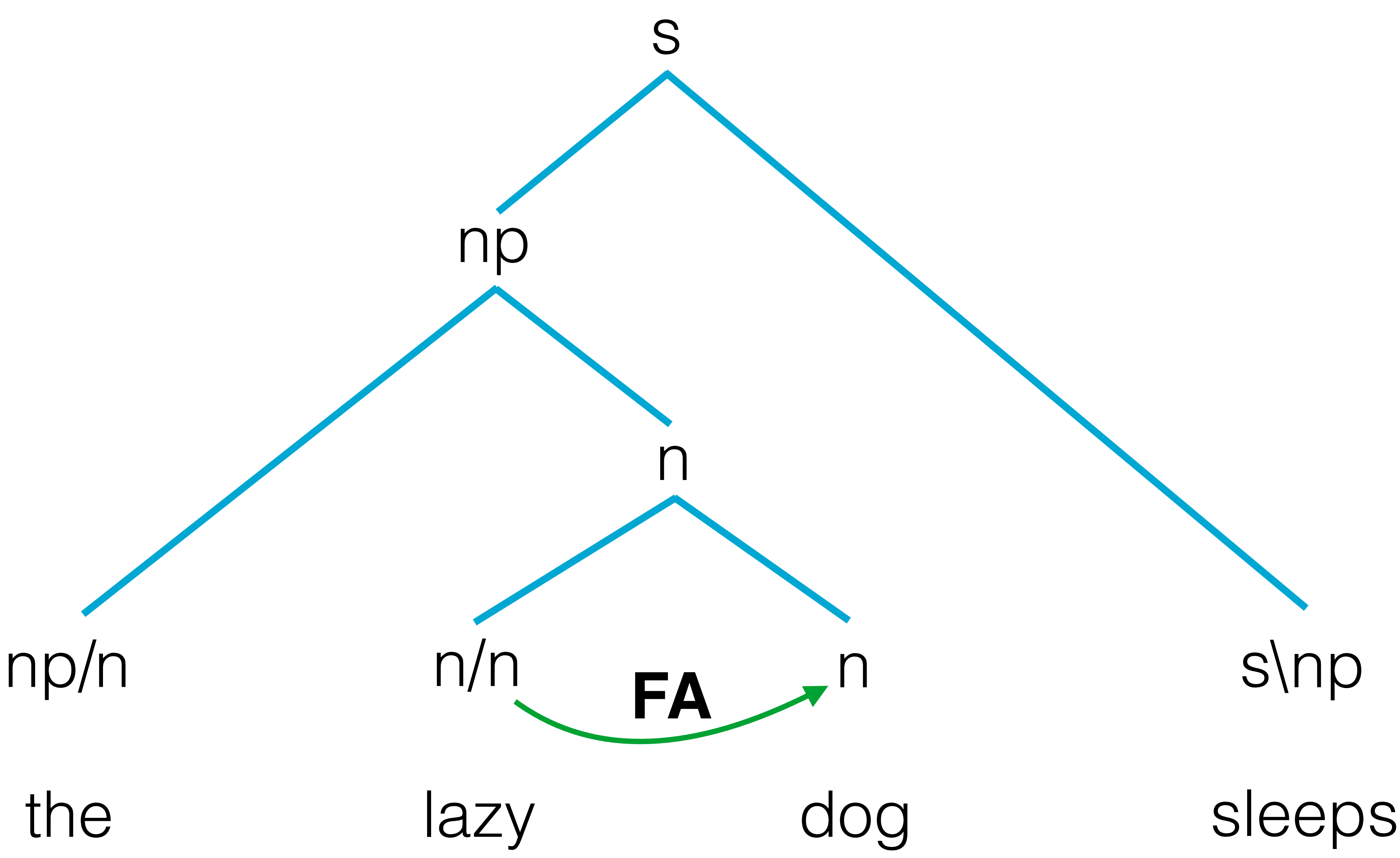
Why CCG?



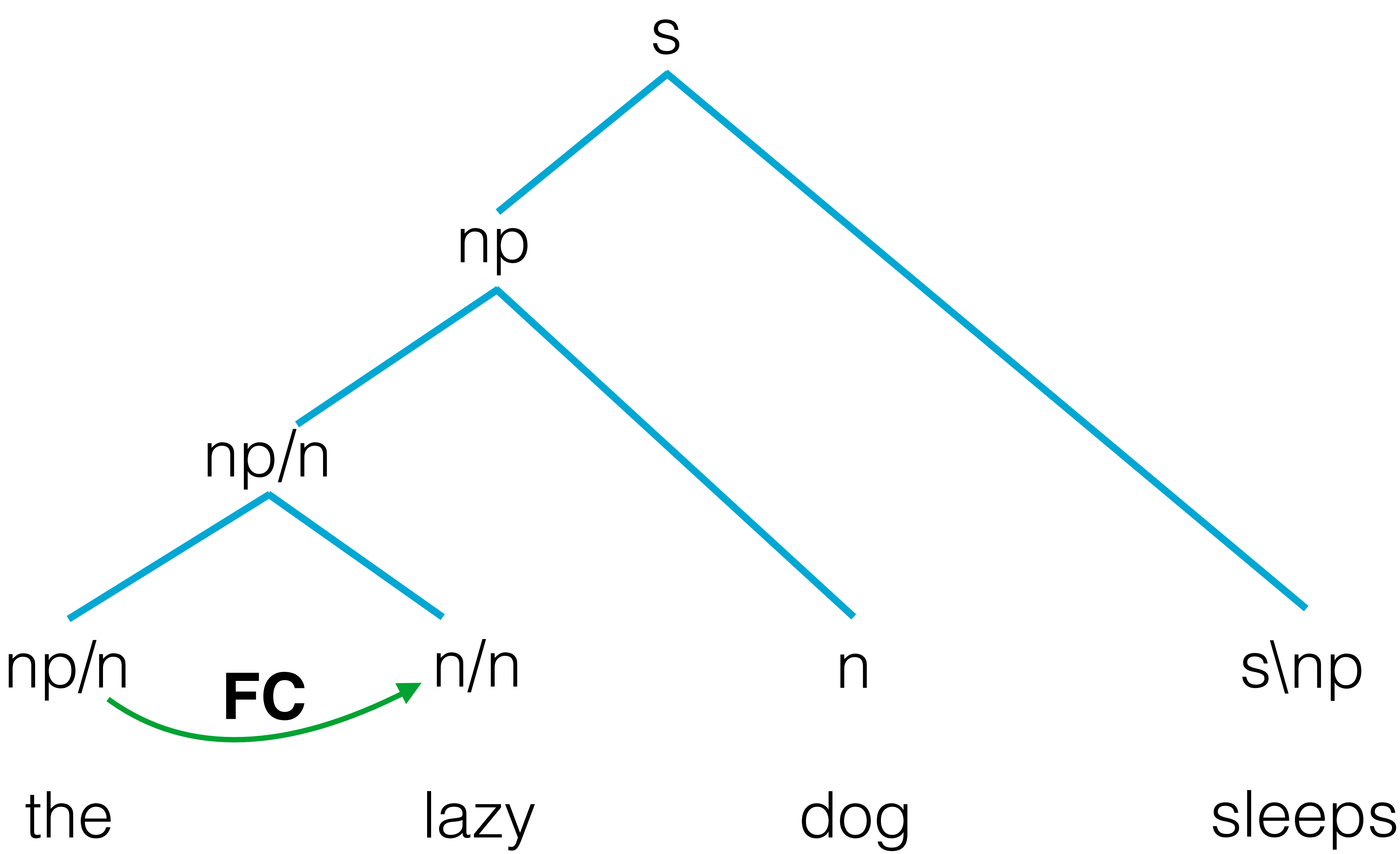
universal, intrinsic grammar properties

all relationships must be learned

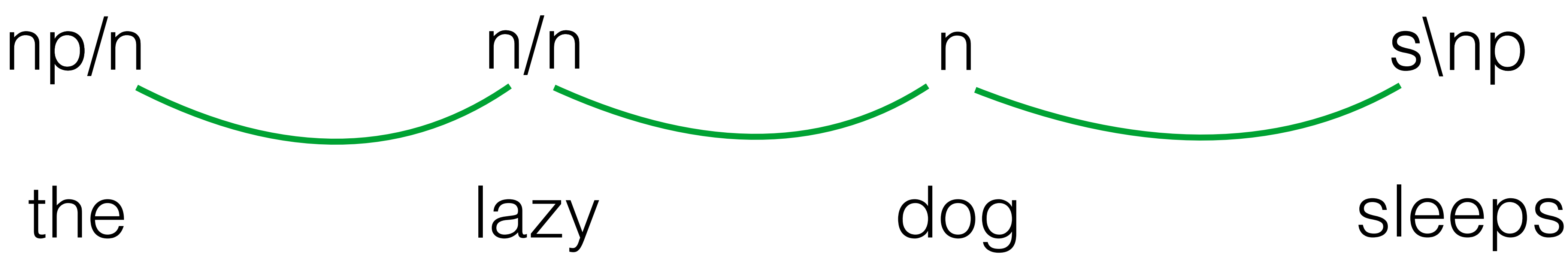
CCG Parsing



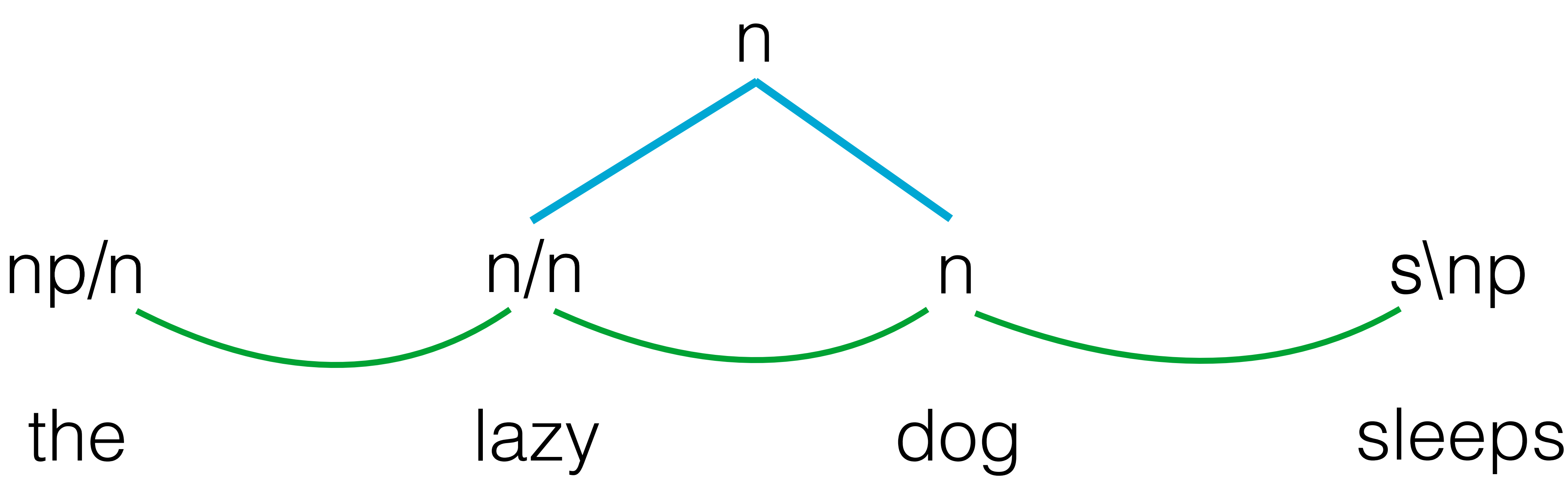
CCG Parsing



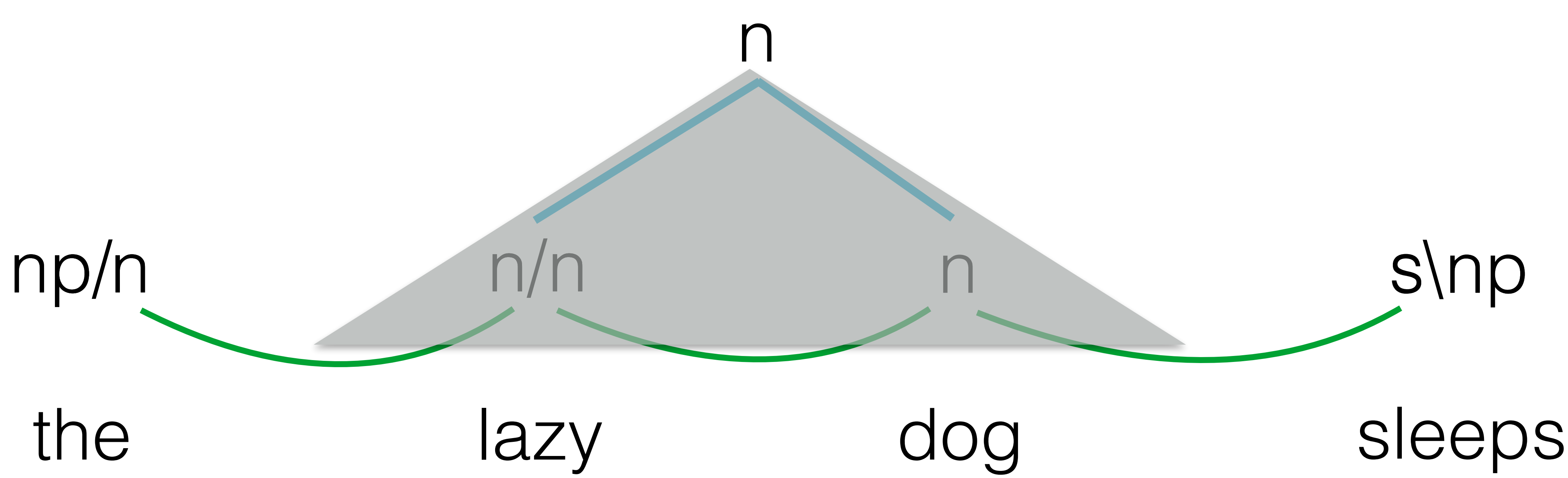
Supertag Context



Supertag Context



Supertag Context

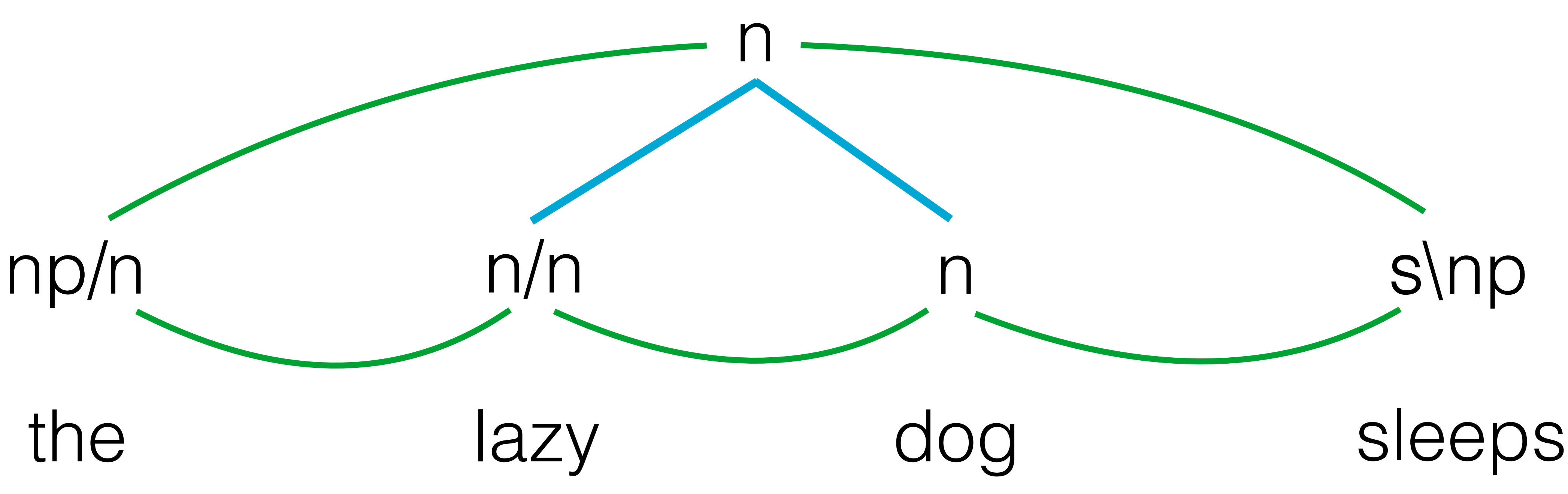


Supertag Context

np/n n s\np

the lazy dog sleeps

Supertag Context



Constituent Context

- Klein & Manning showed the value of modeling context with the Constituent Context Model (CCM)

the lazy dog sleeps

[Klein & Manning 2002]

Constituent Context

DT (JJ NN) VBZ

[Klein & Manning 2002]

Constituent Context

“substitutability”

DT (JJ NN) VBZ

lazy dog

[Klein & Manning 2002]

Constituent Context

"substitutability"

DT (NN) VBZ

dog

[Klein & Manning 2002]

Constituent Context

"substitutability"

DT (JJ JJ NN) VBZ

big lazy dog

[Klein & Manning 2002]

Constituent Context

"substitutability"

DT (~Noun) VBZ

[Klein & Manning 2002]

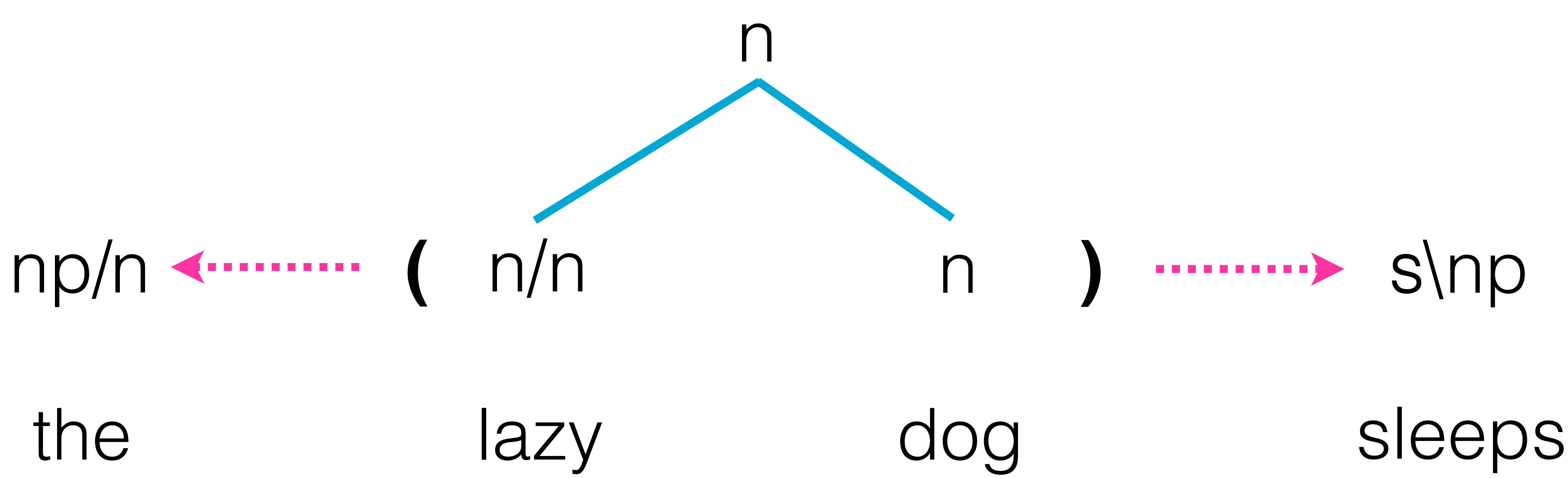
Constituent Context

"substitutability"

DT ( ) VBZ

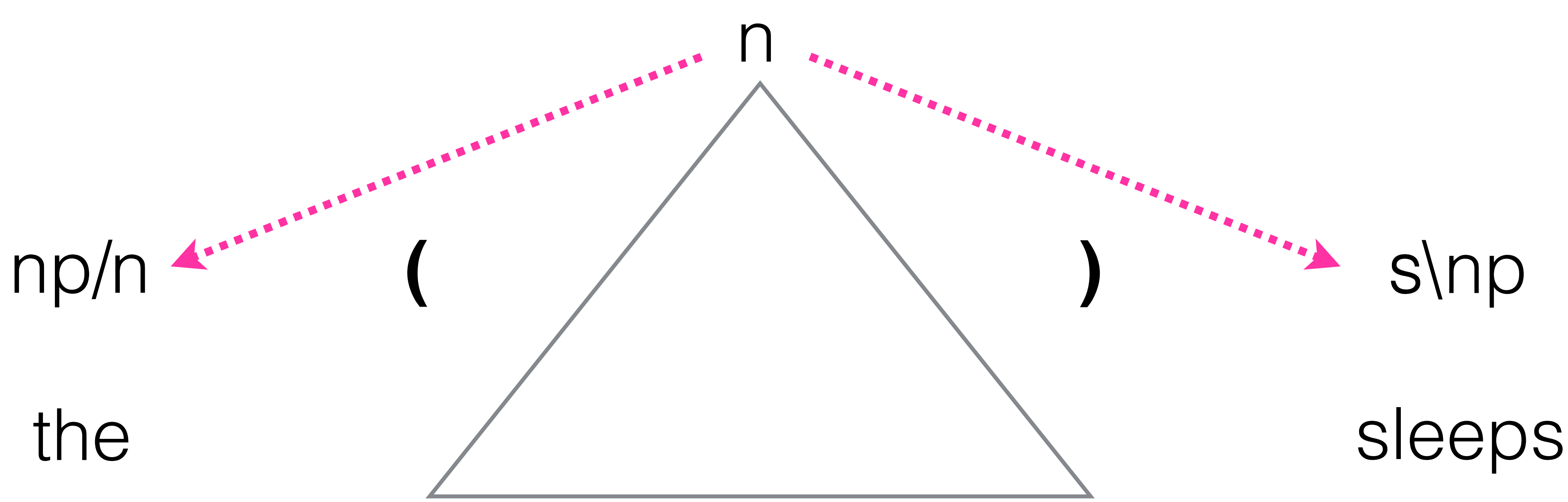
[Klein & Manning 2002]

Supertag Context



Supertag Context

- We know the constituent label
- We know if it’s a fitting context, even before looking at the data



This Paper

1. A **new generative model** for learning CCG parsers from *weak supervision*

2. A way to select Bayesian **priors** that capture properties of CCG

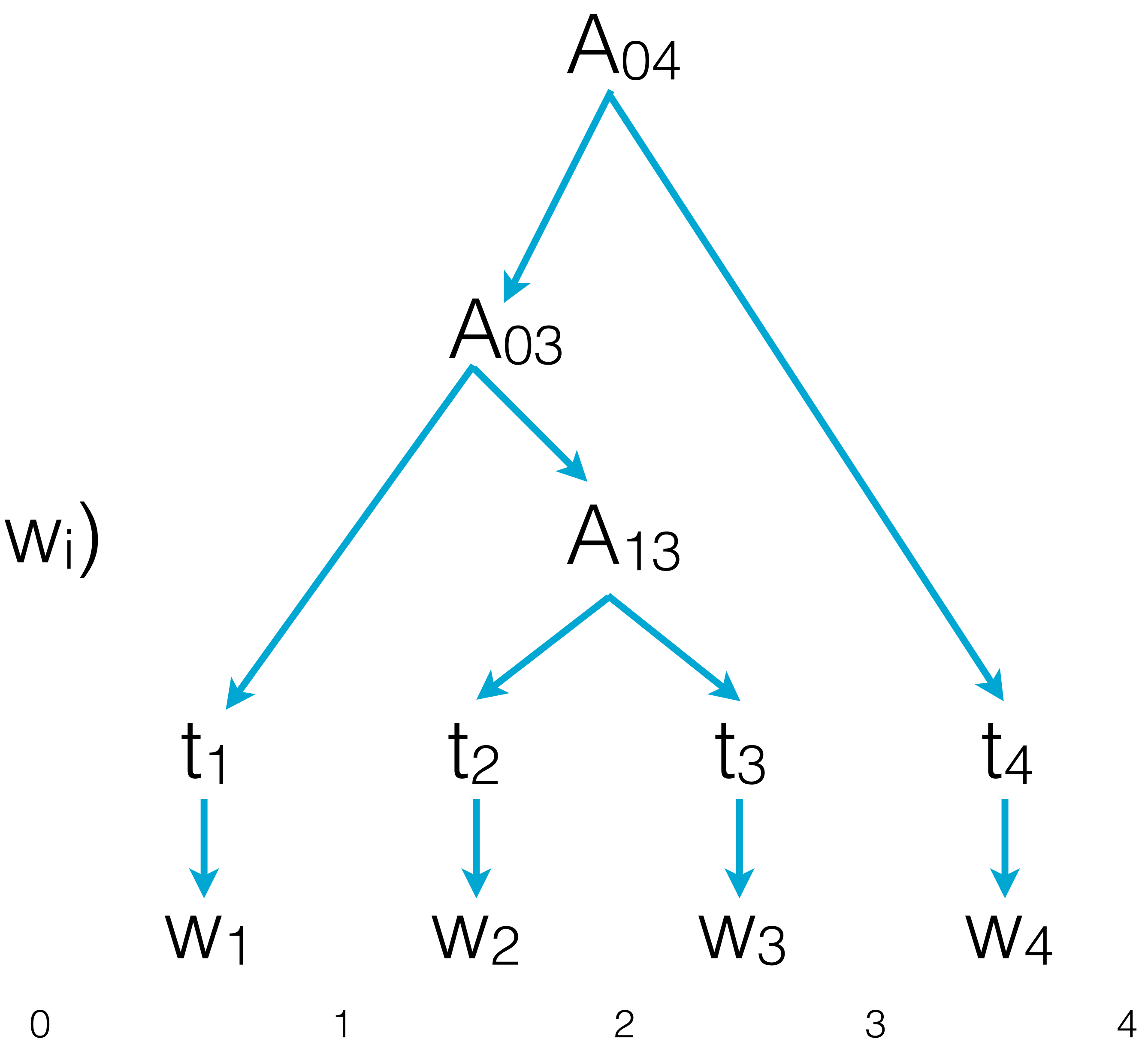
3. A Bayesian **inference procedure** to learn the parameters of our model

Supertag-Context Parsing

Standard PCFG

$P(A_{root})$

$P(A \rightarrow A_{left}\ A_{right}\ \text{OR}\ w_i)$



Supertag-Context Parsing

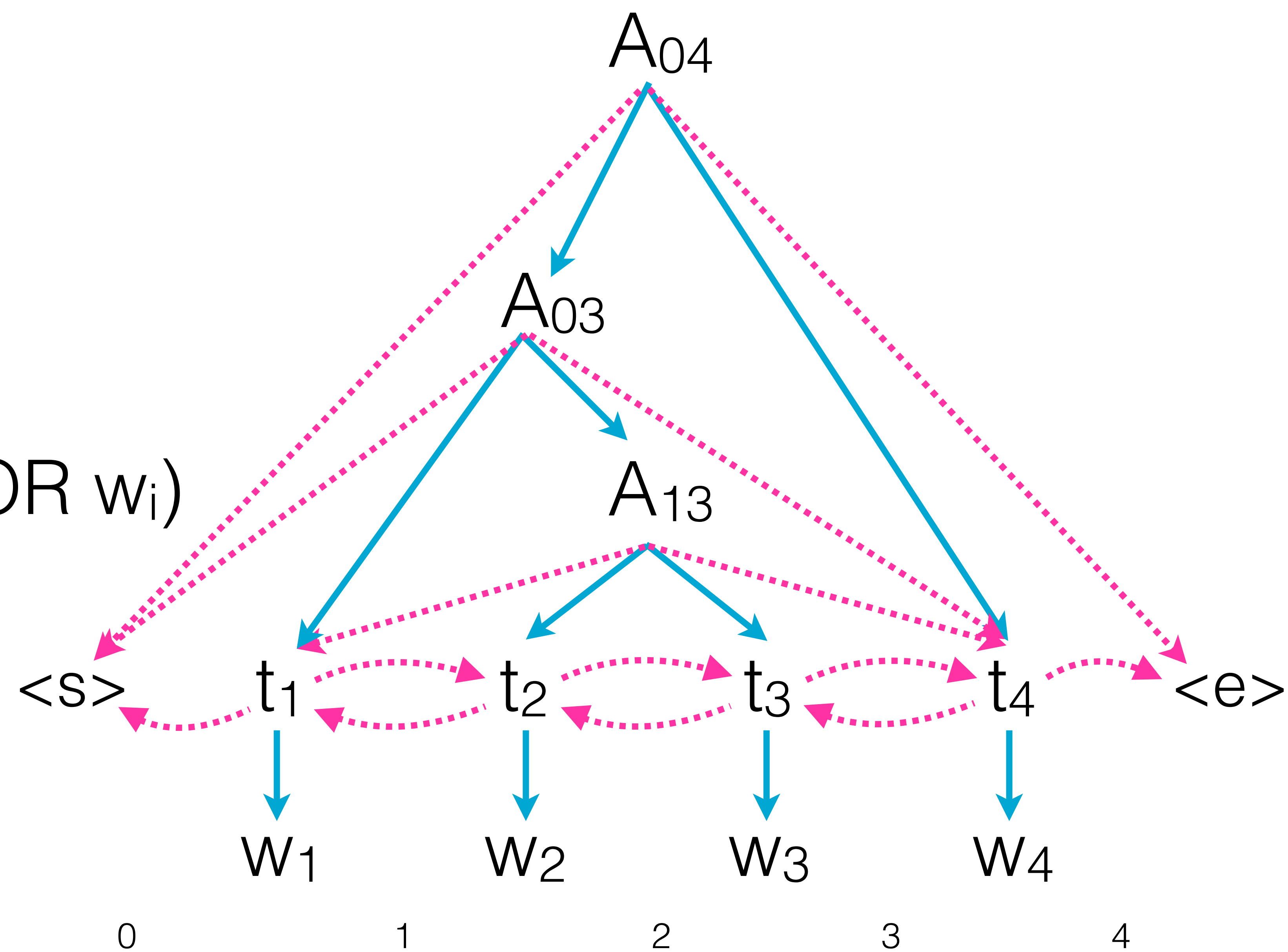
With Context

$P(A_{root})$

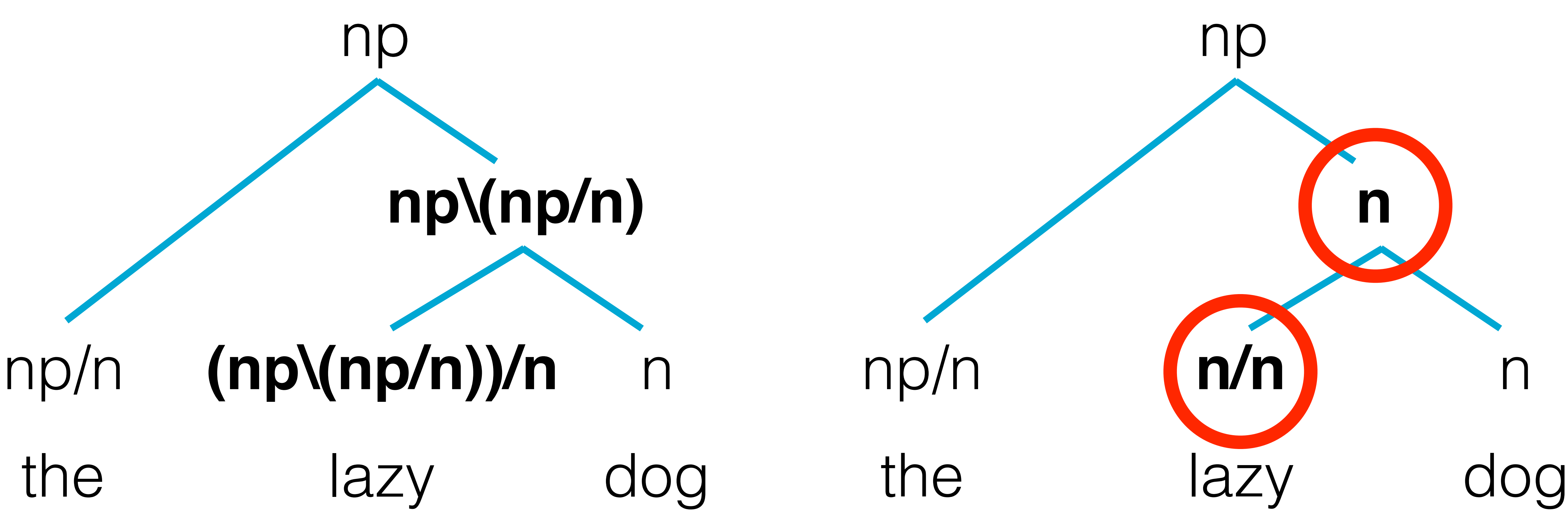
$P(A \rightarrow A_{left}\ A_{right}\ \text{OR}\ w_i)$

$P(A \rightarrow t_{left})$

$P(A \rightarrow t_{right})$



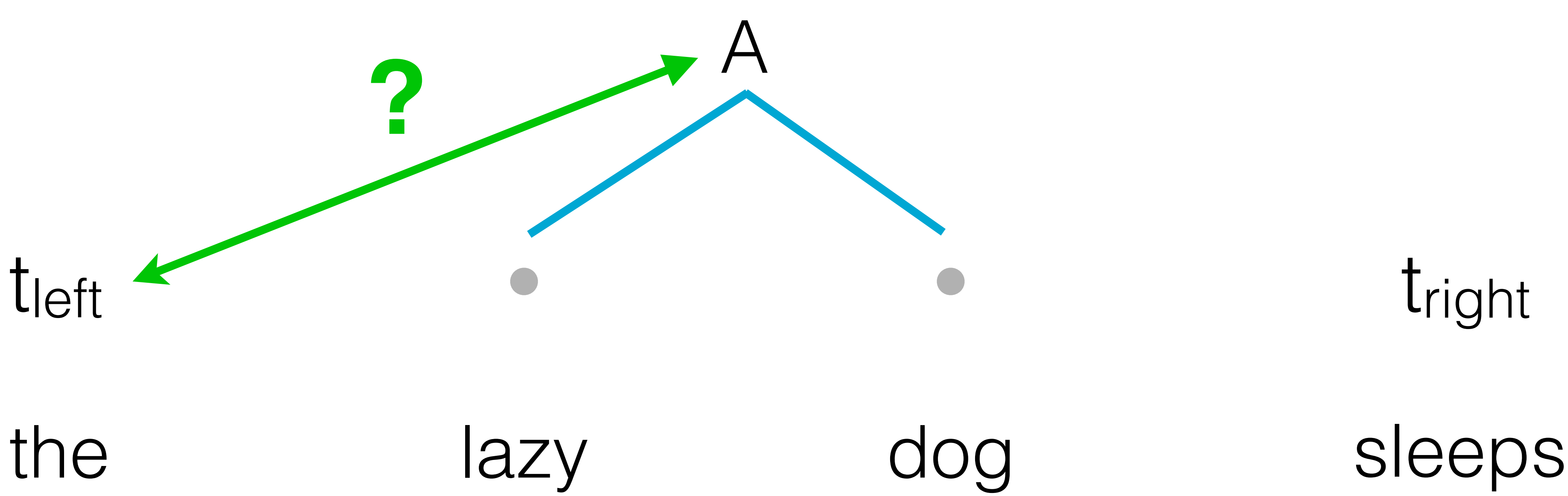
Prior on Categories



[Garrette, Dyer, Baldridge, and Smith, 2015]

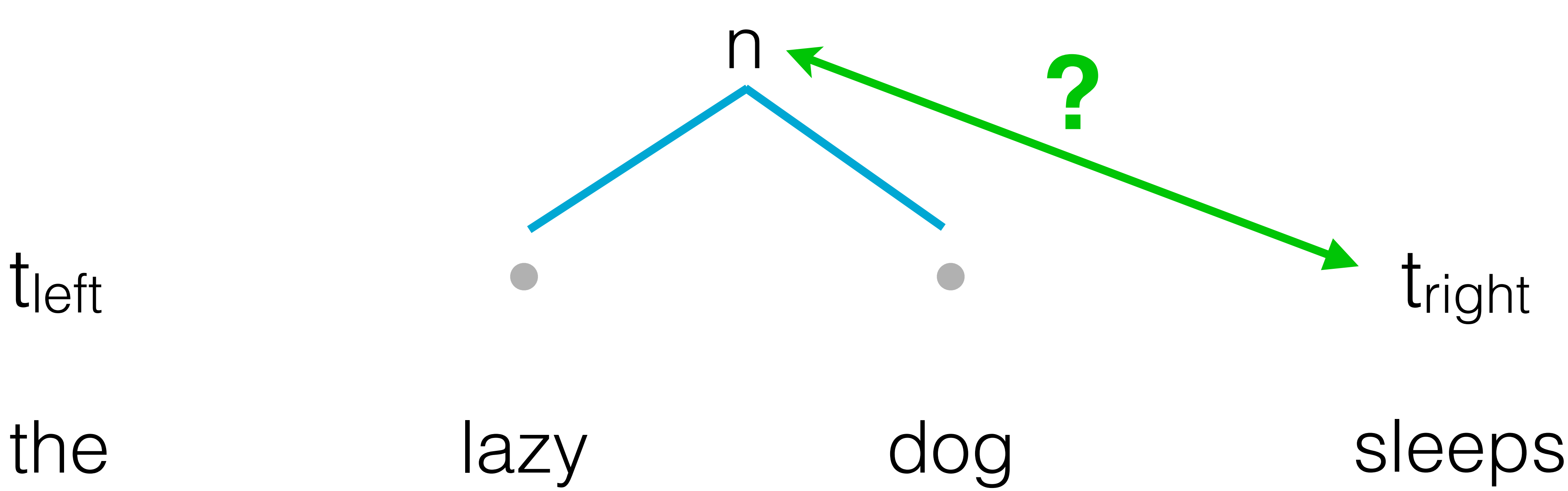
Supertag-Context Prior

$$P_{\text{L-prior}}(t_{\text{left}} \mid A) \propto \begin{cases} 10^5 & \text{if } t_{\text{left}} \text{ can combine with } A \\ 1 & \text{otherwise} \end{cases}$$



Supertag-Context Prior

$$P_{R\text{-prior}}(t_{right} \mid A) \propto \begin{cases} 10^5 & \text{if } A \text{ can combine with } t_{right} \\ 1 & \text{otherwise} \end{cases}$$



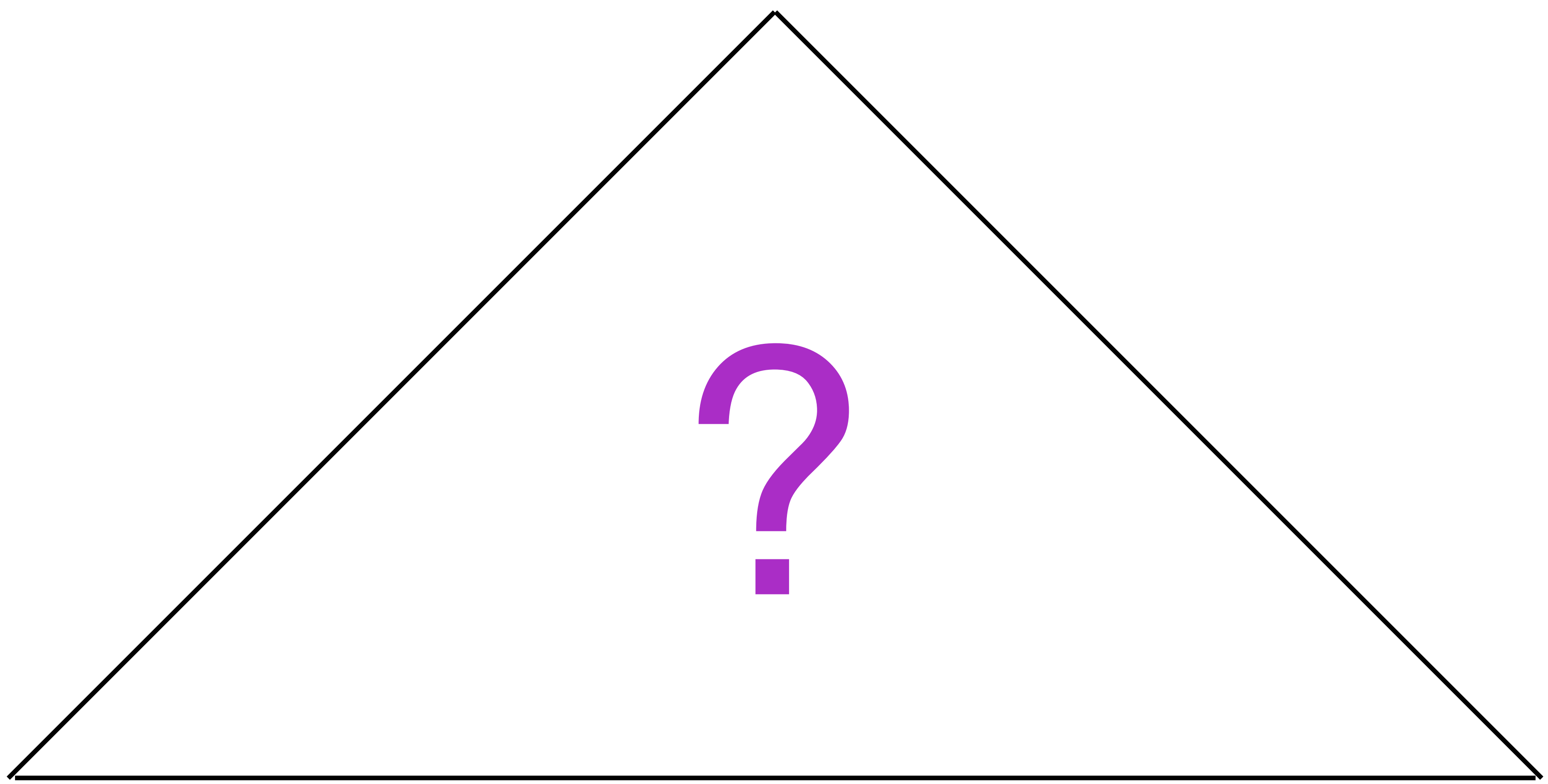
This Paper

1. A **new generative model** for learning CCG parsers from *weak supervision*

2. A way to select Bayesian **priors** that capture properties of CCG

3. A Bayesian **inference procedure** to learn the parameters of our model

Type-Level Supervision



the	lazy	dogs	wander
np/n	n/n	n	
	np	np	
		(s\np)/np	

Type-Supervised Learning

unlabeled corpus

tag dictionary

universal properties of the CCG formalism

Posterior Inference

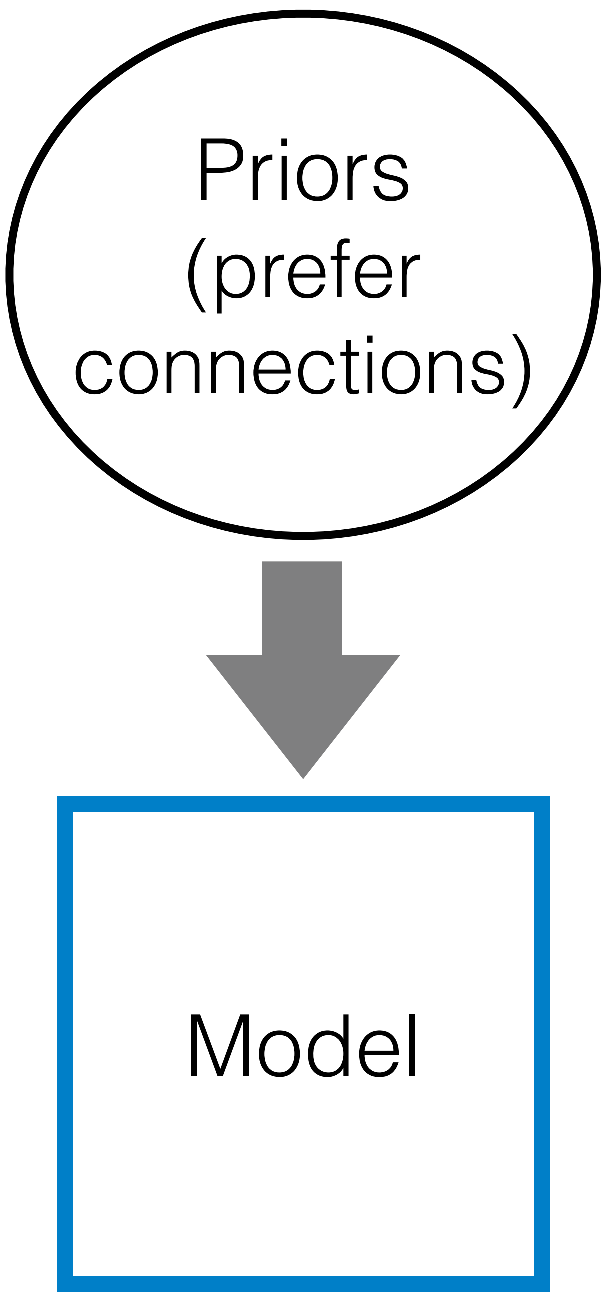
- A Bayesian inference procedure will make use of our linguistically-informed priors
- But we can't do sampling like a PCFG
    - Can't compute the inside chart, even with dynamic programming.

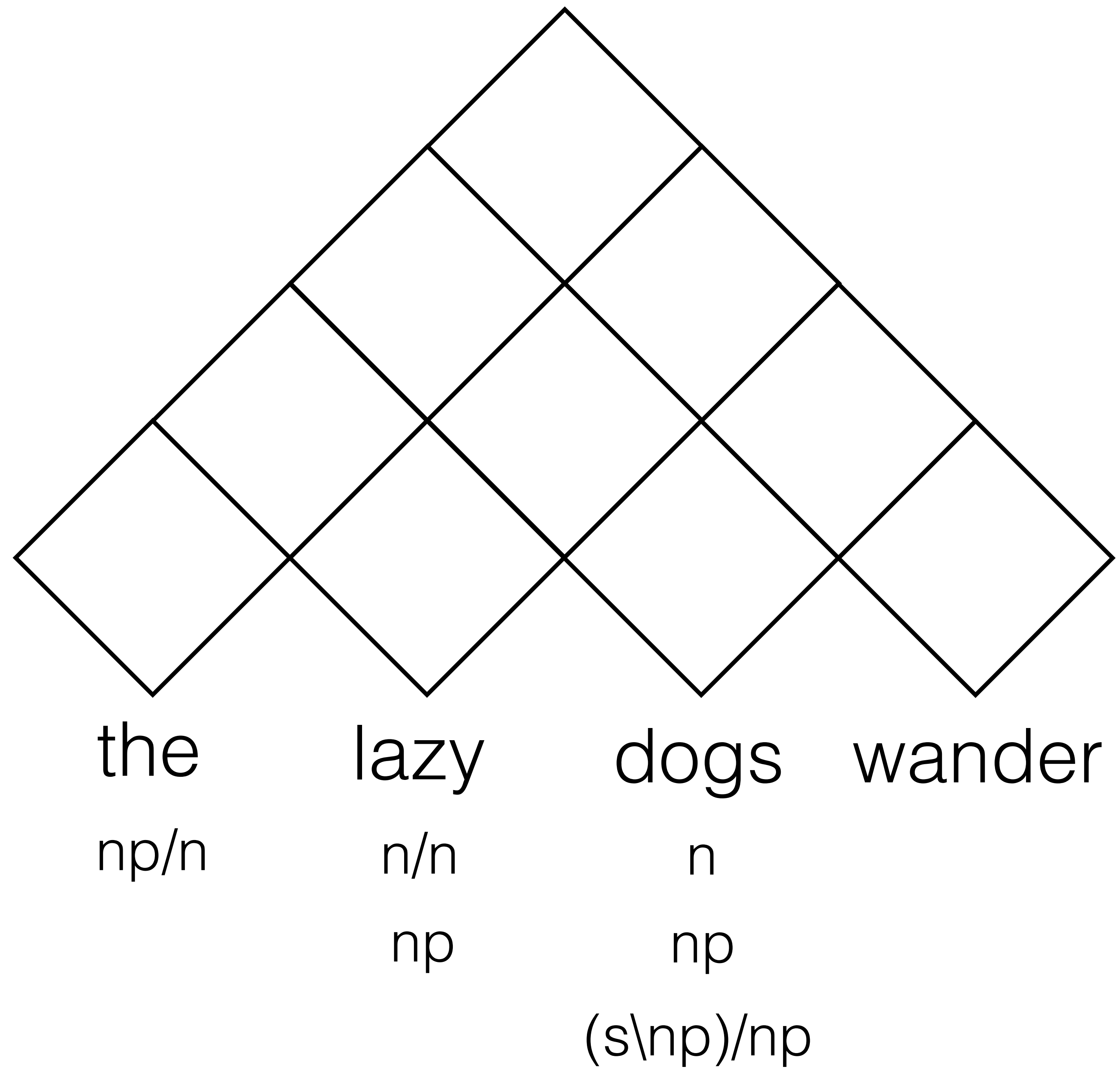
Sampling via Metropolis-Hastings

Idea:

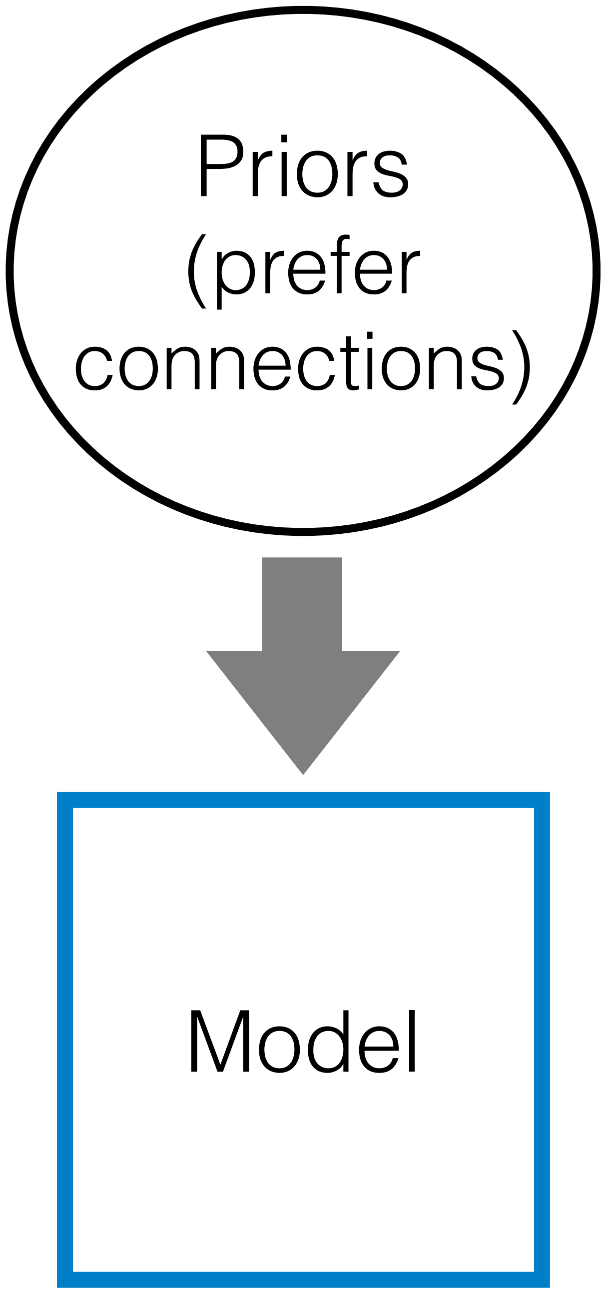
- Sample tree from an efficient **proposal** distribution
  - (PCFG parameters) (Johnson et al. 2007)
- Accept according to the **full** distribution
  - (Context parameters)

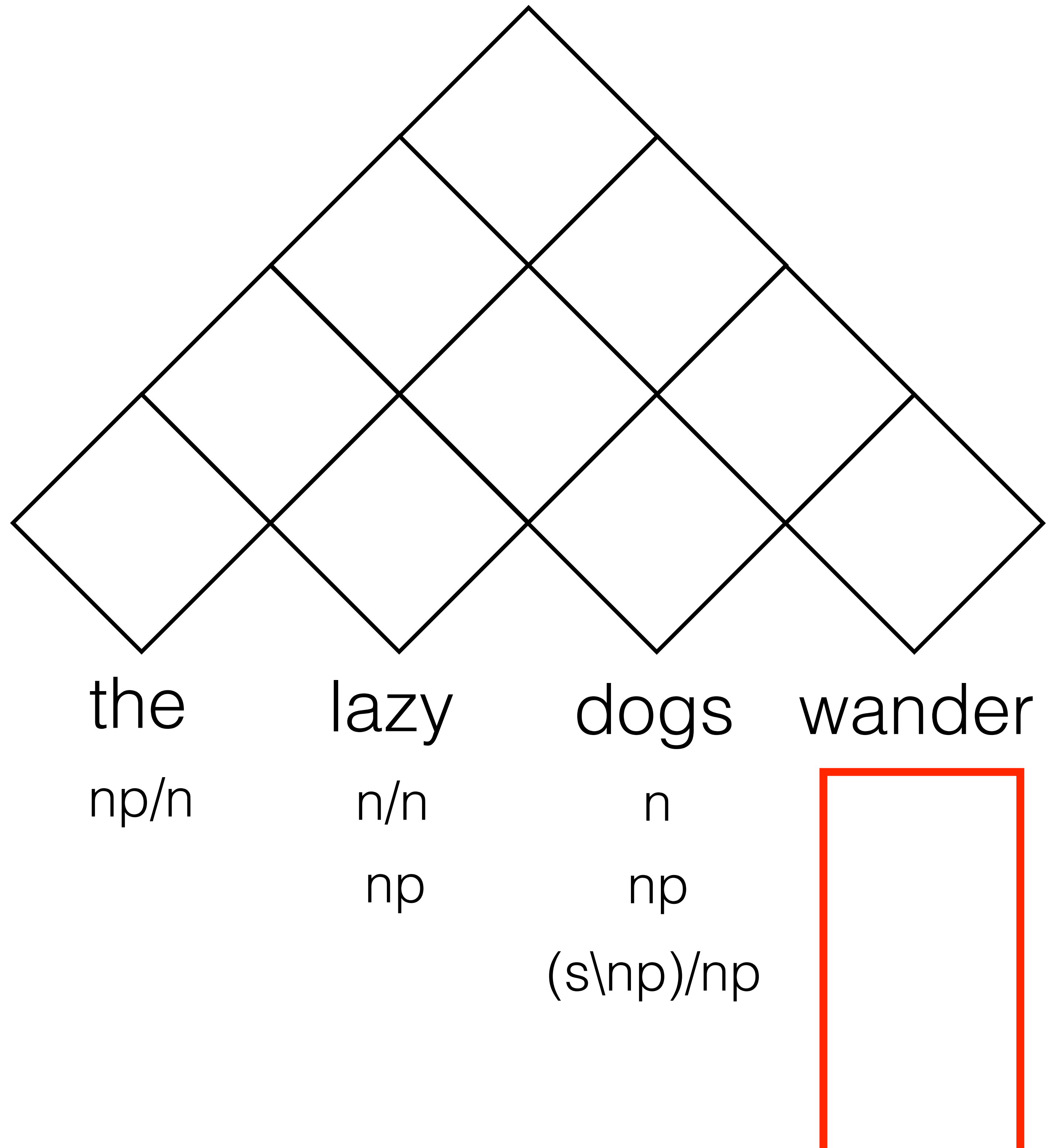
Posterior Inference



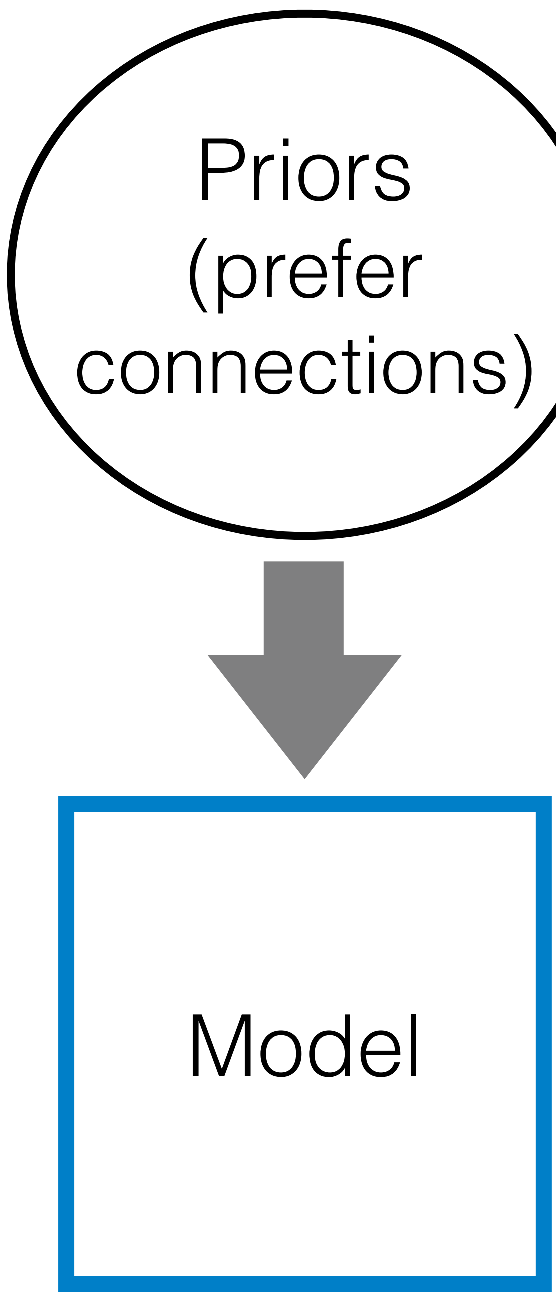


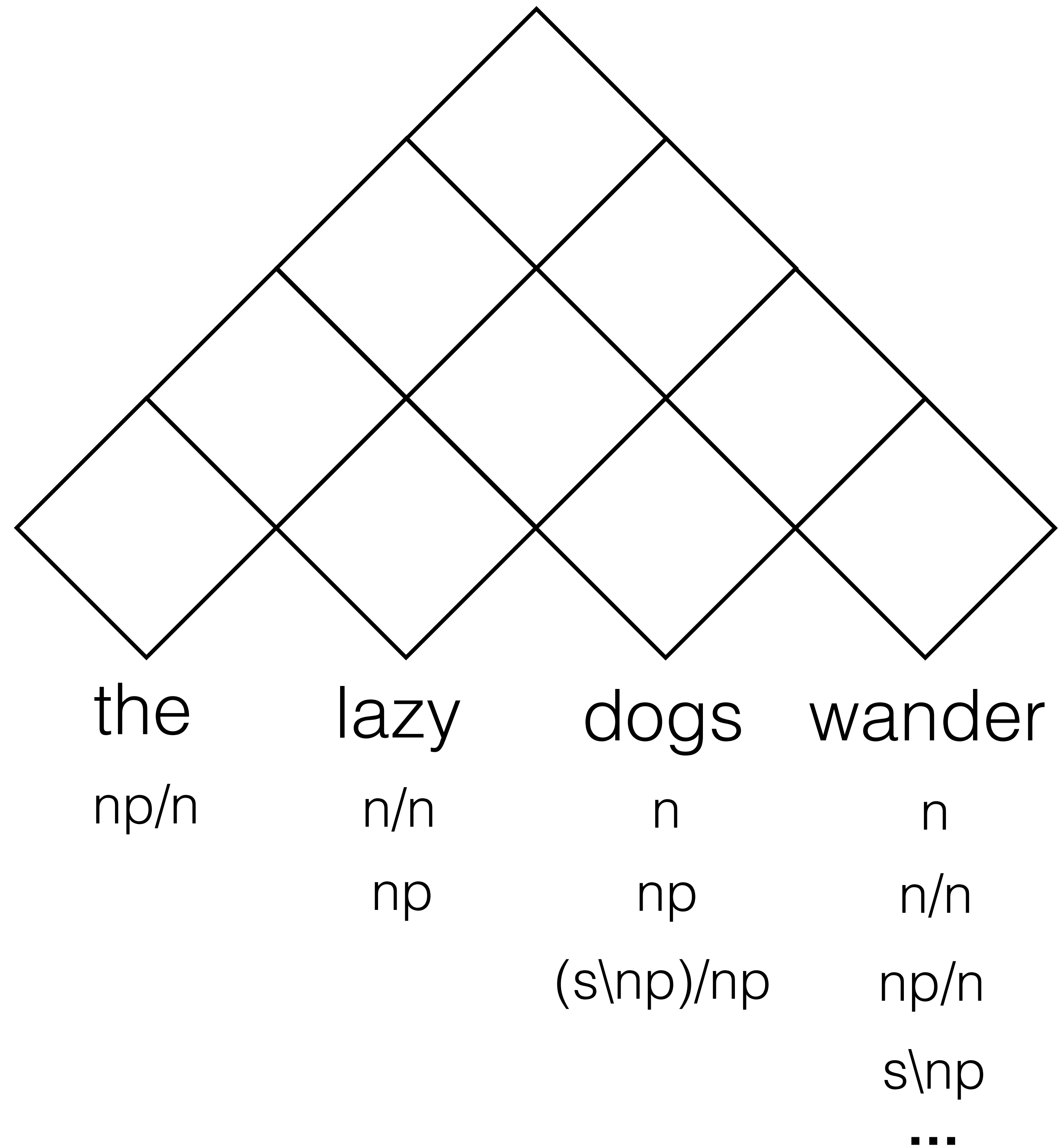
Posterior Inference



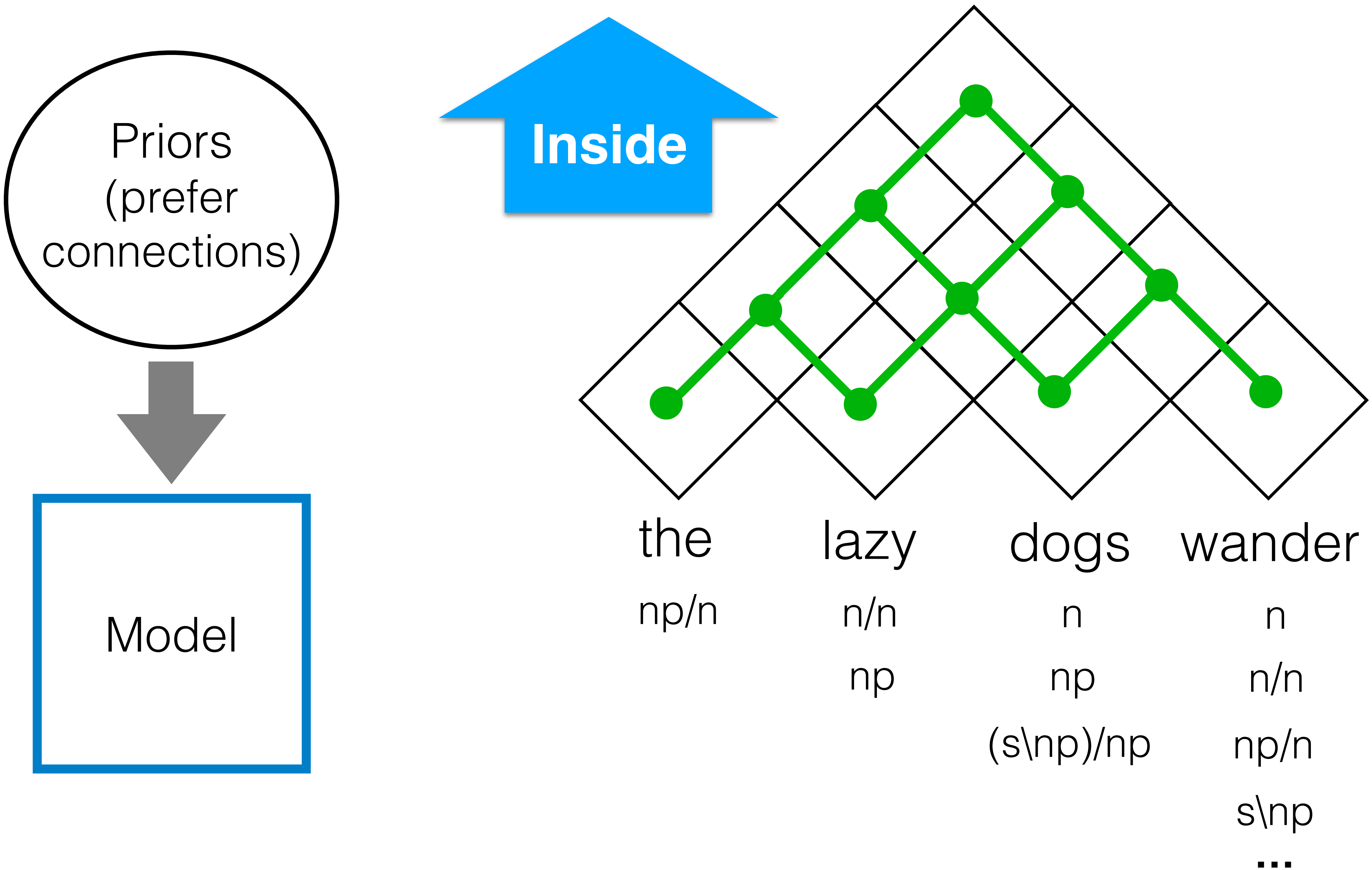


Posterior Inference

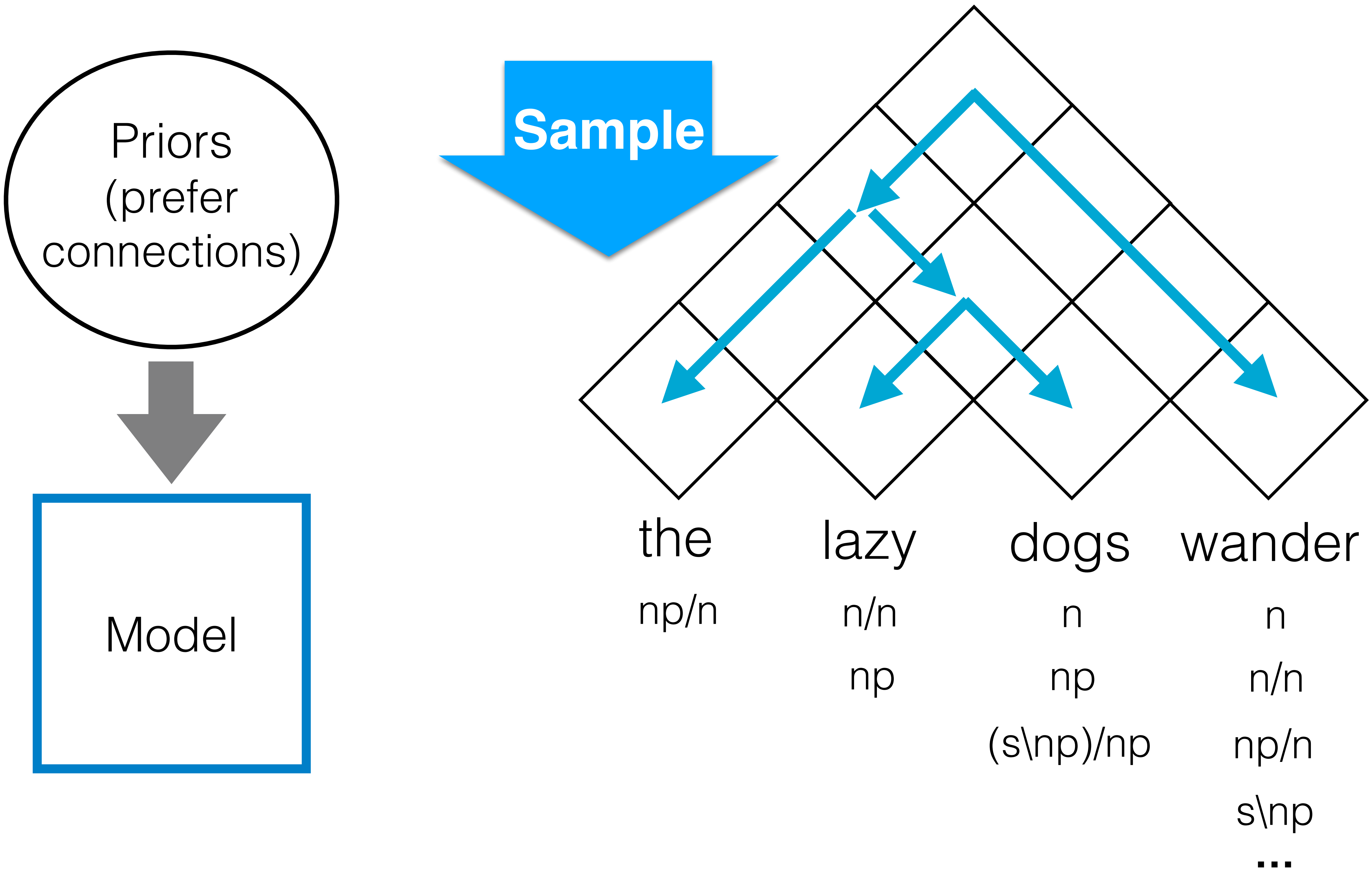




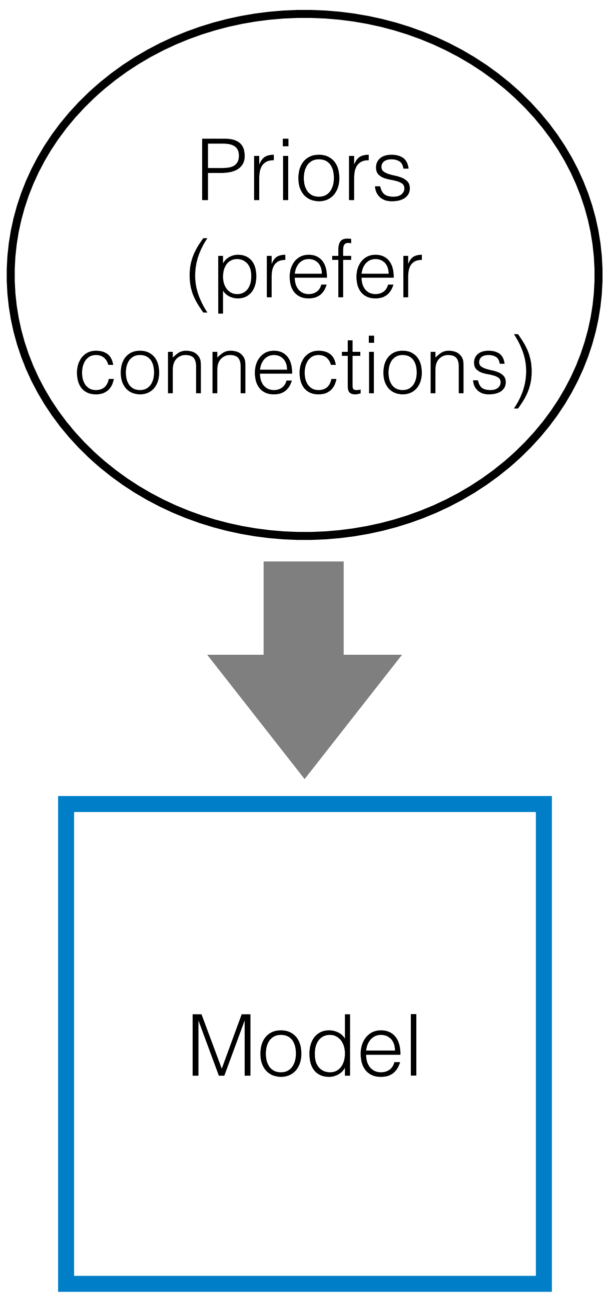
Posterior Inference

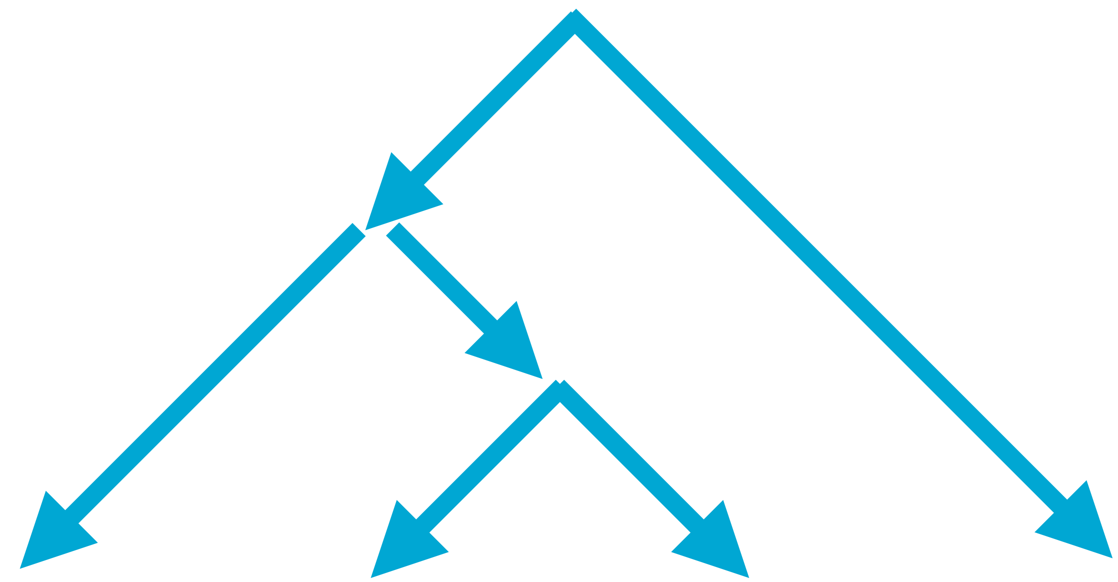


Posterior Inference

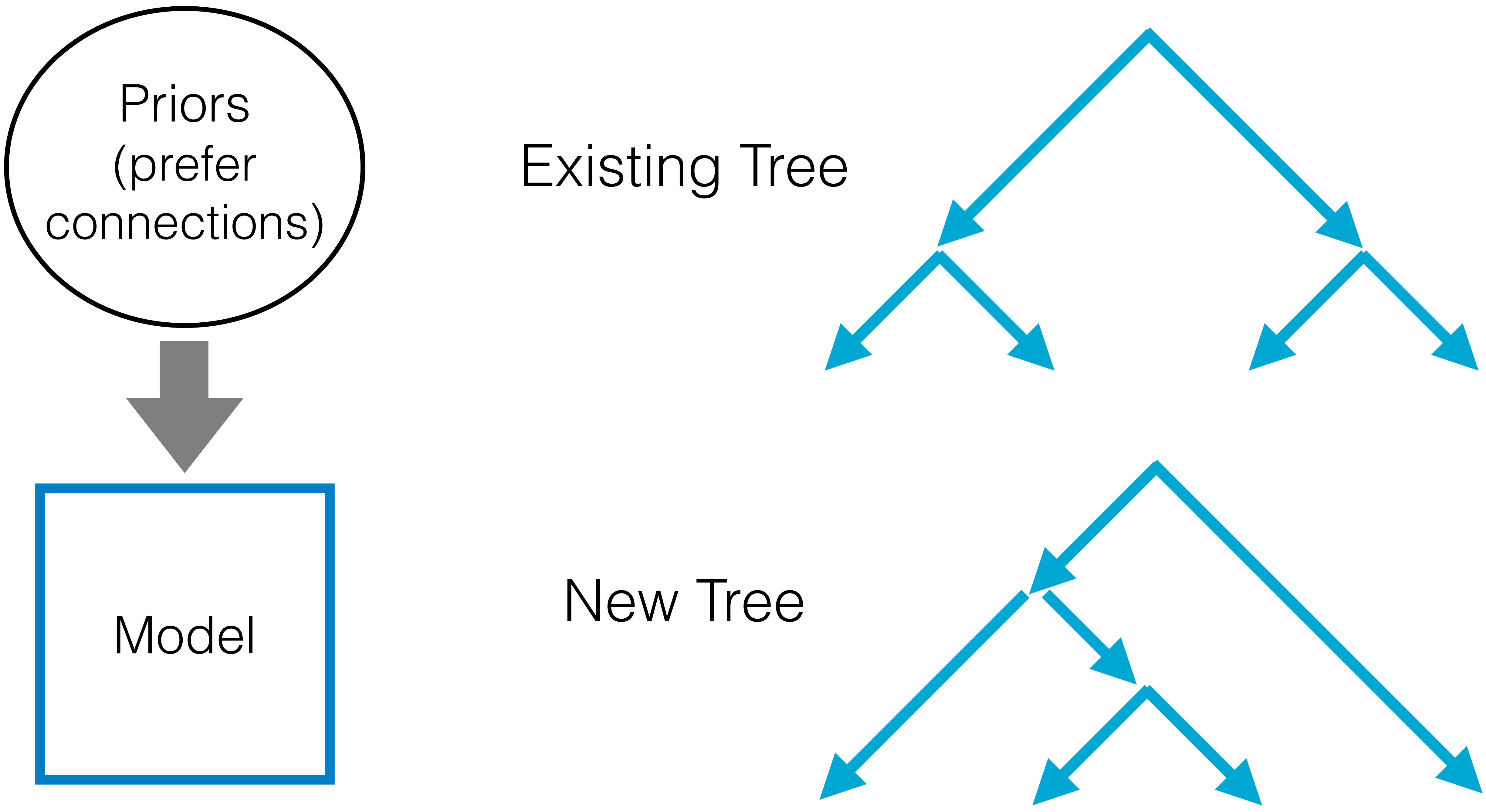


Metropolis-Hastings

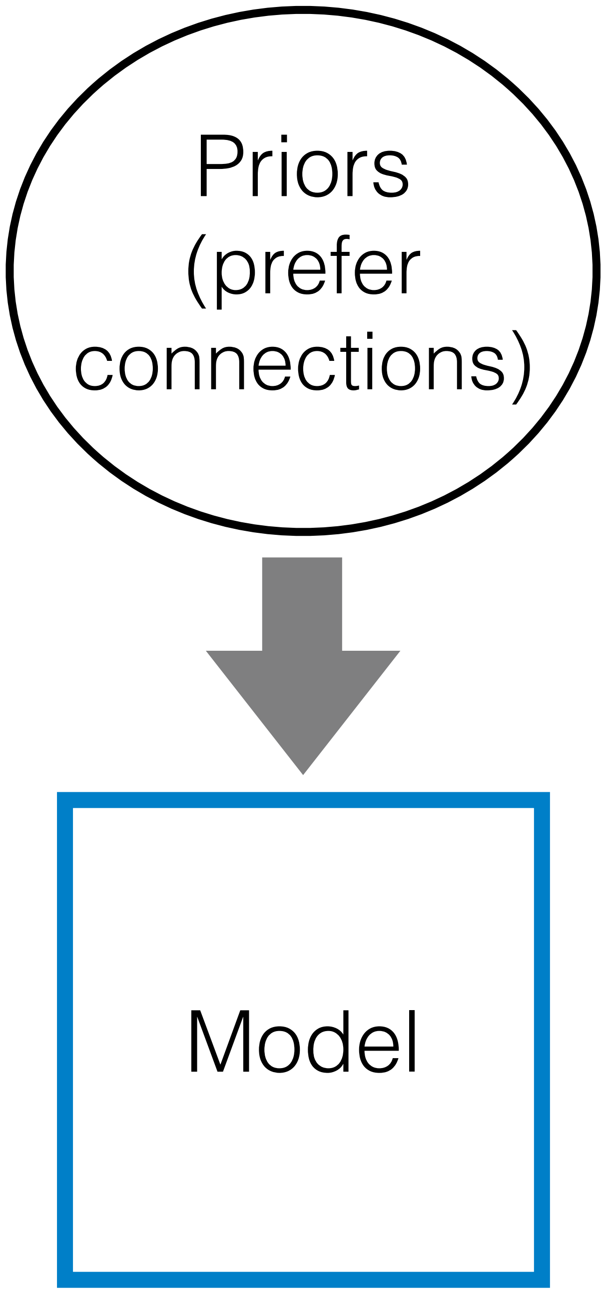




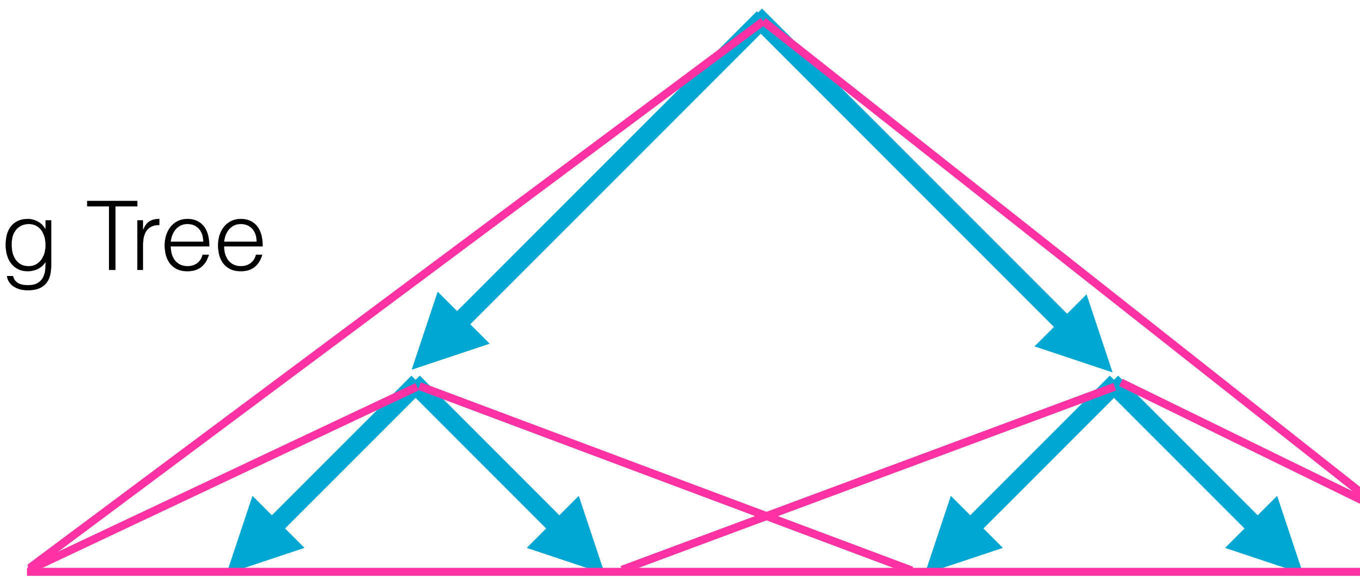
Metropolis-Hastings



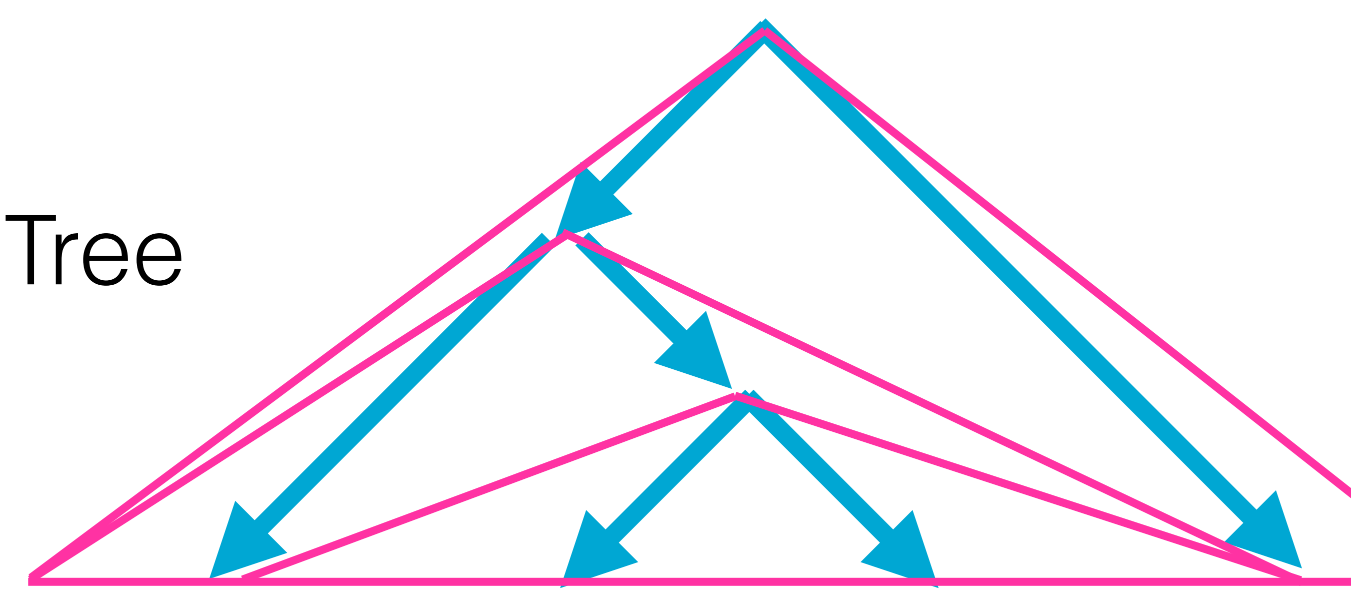
Metropolis-Hastings



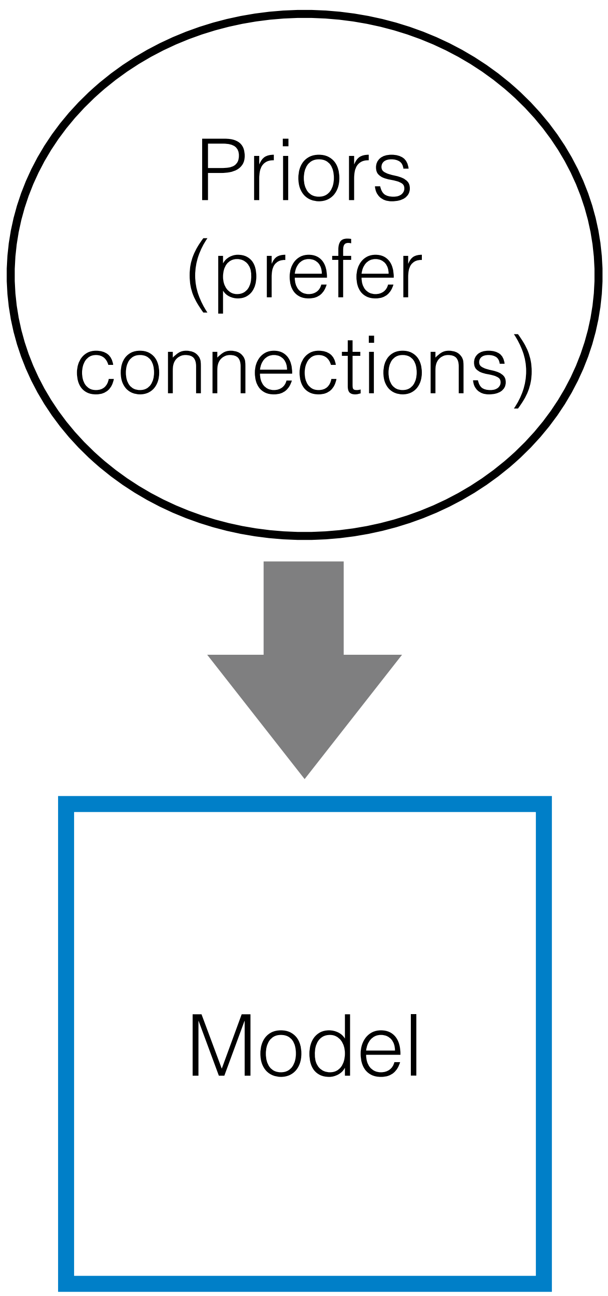
Existing Tree



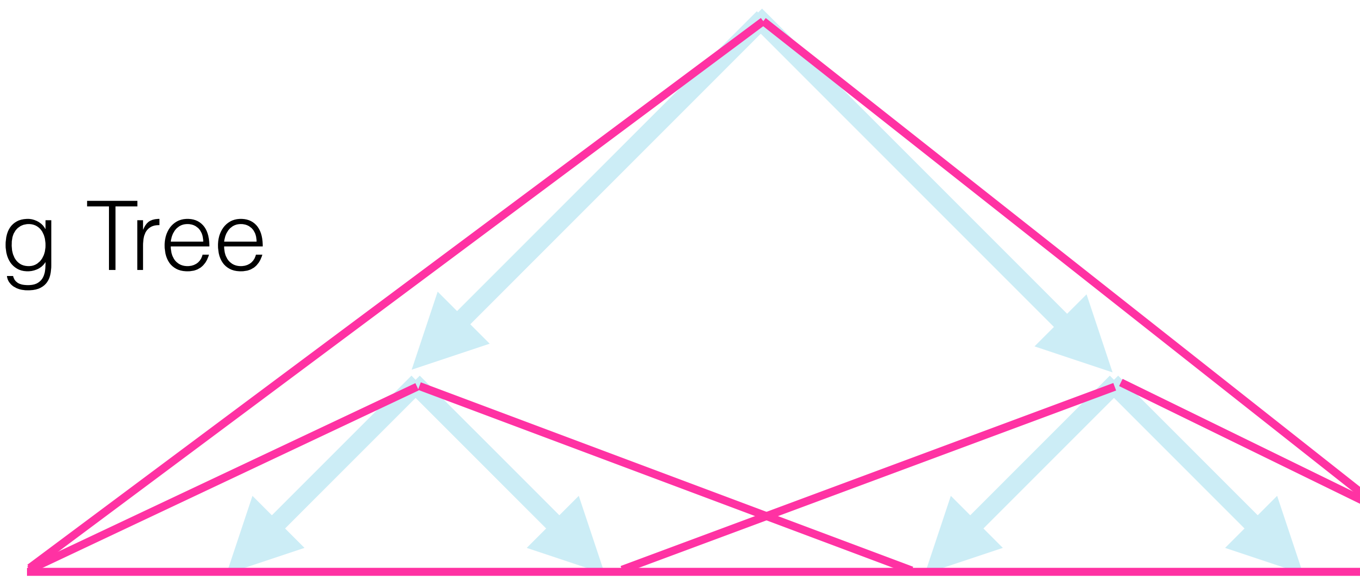
New Tree



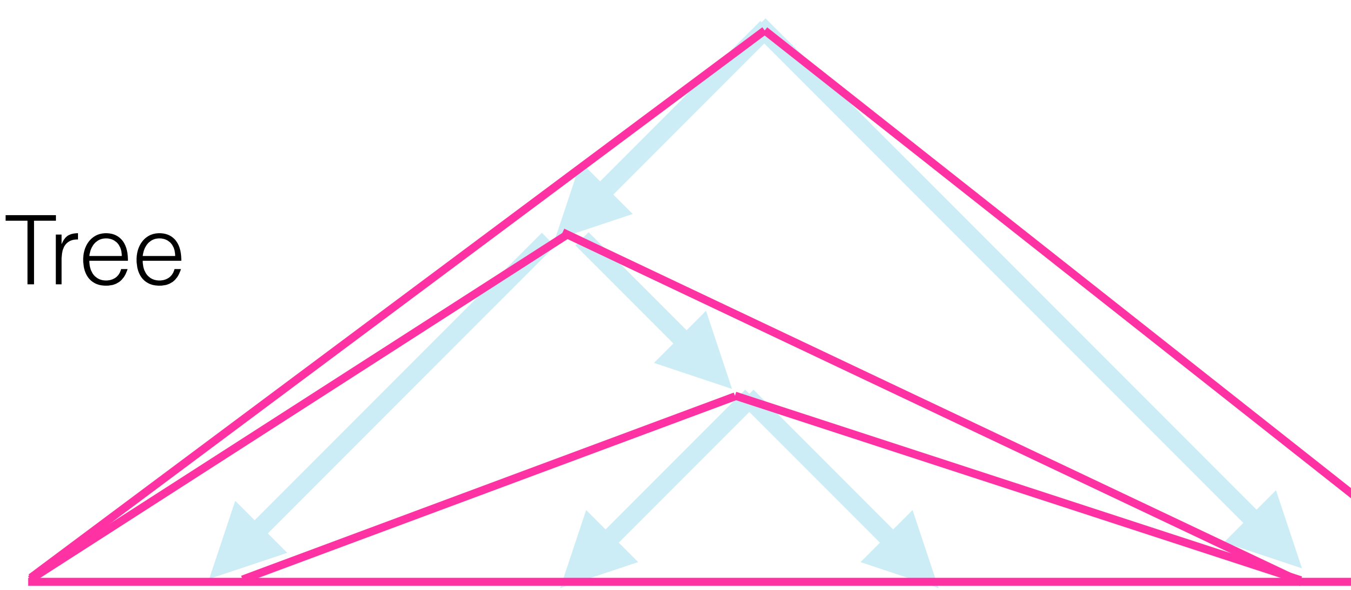
Metropolis-Hastings



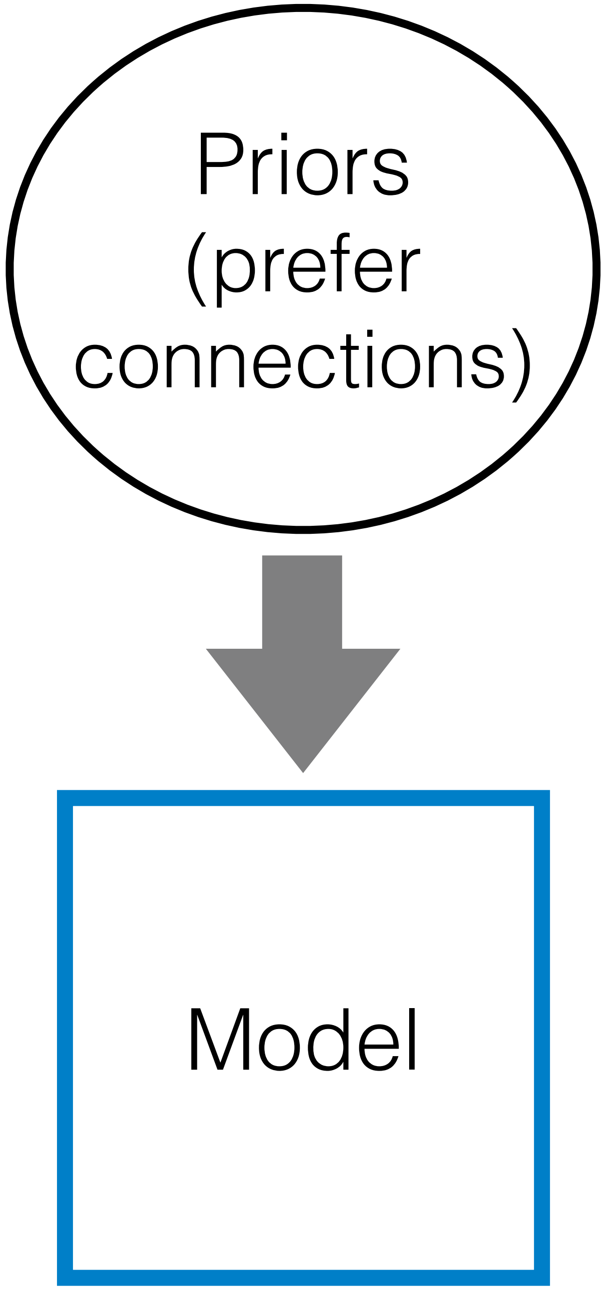
Existing Tree

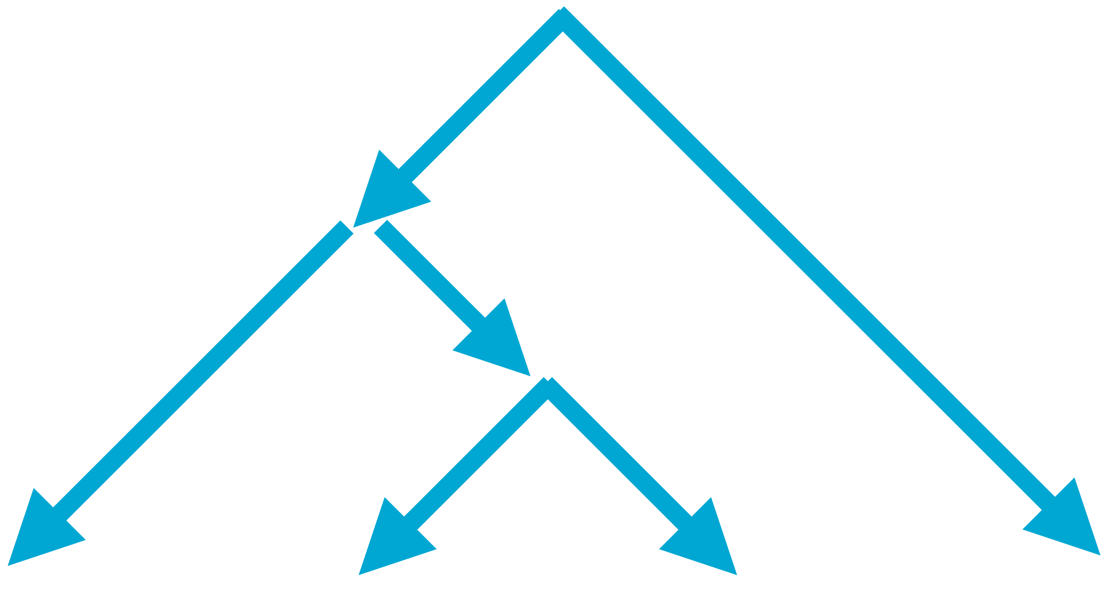


New Tree

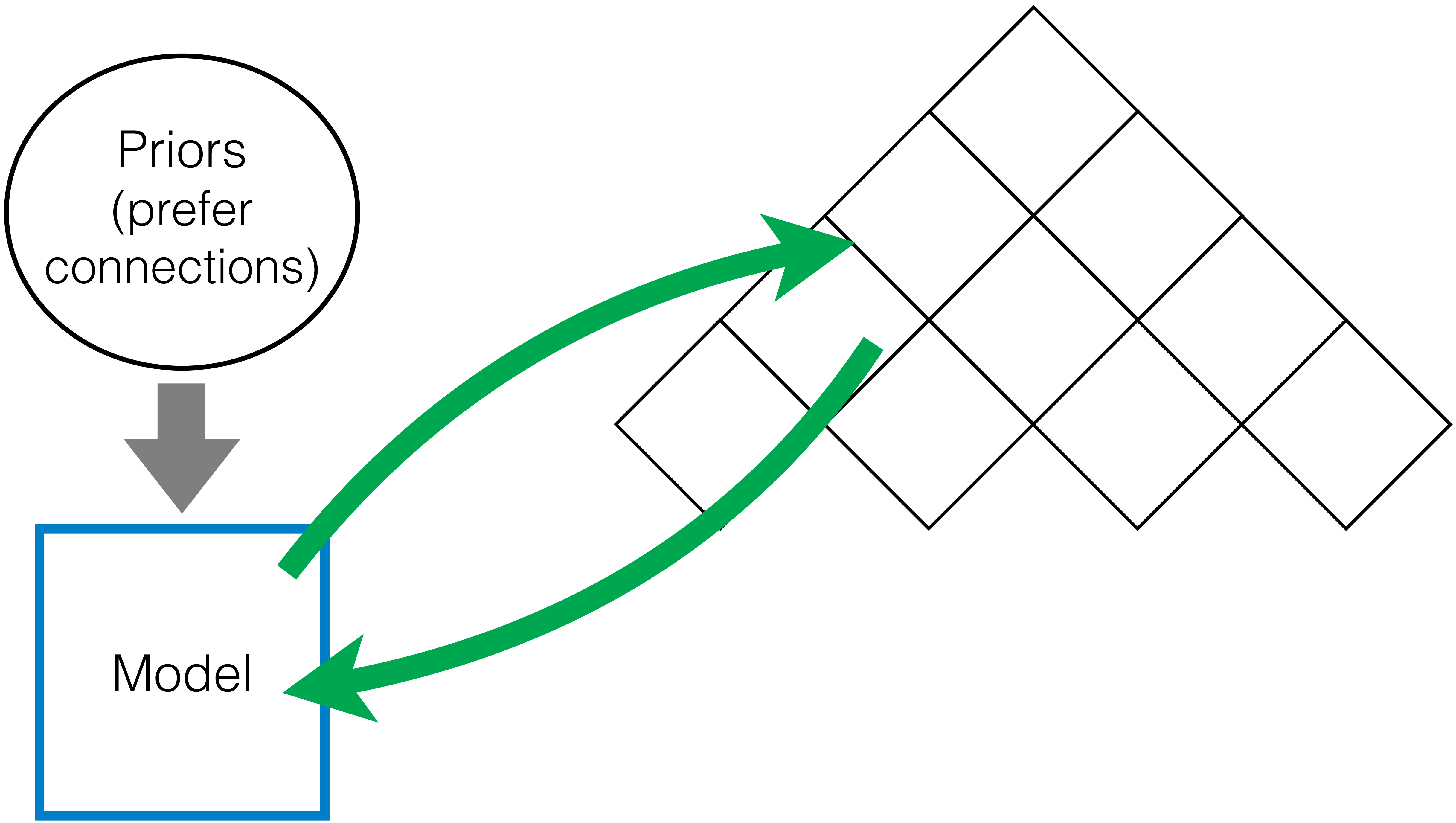


Metropolis-Hastings





Posterior Inference



Metropolis-Hastings

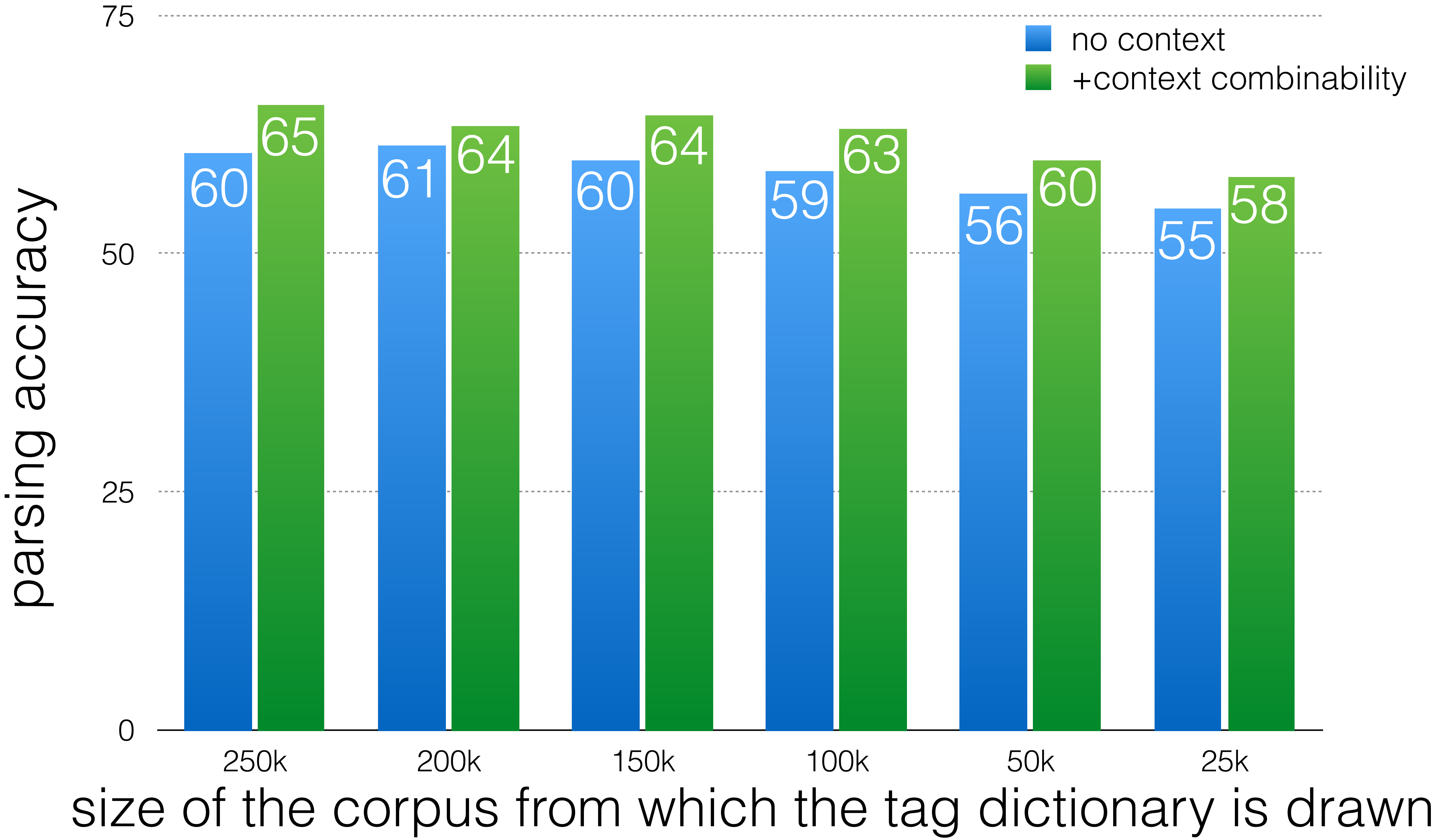
- Sample tree based only on the pcfg parameters
- Accept based only on the context
- New worse than old => less likely to accept

Experimental Results

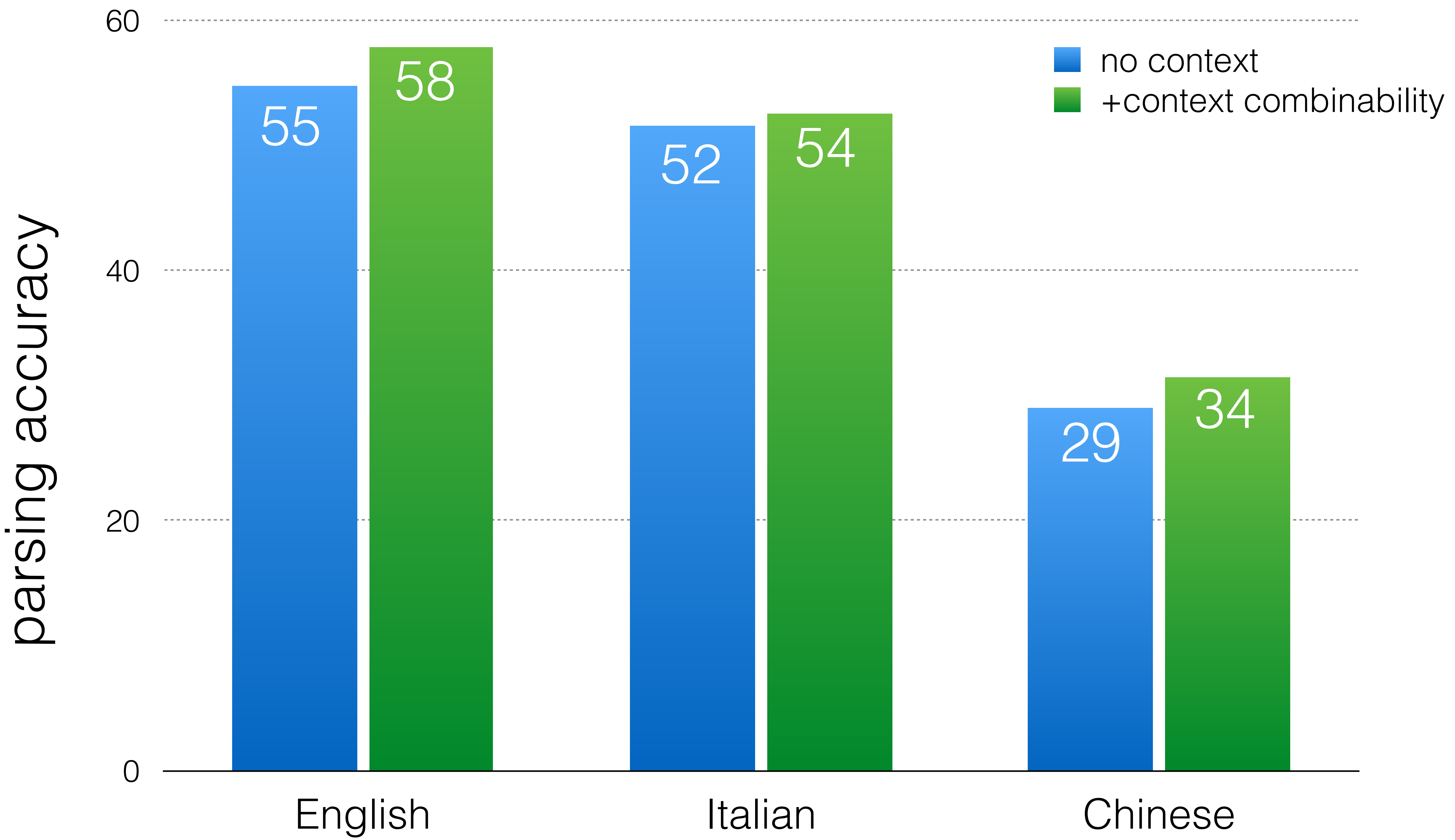
Experimental Question

- When supervision is incomplete, does modeling context, and biasing toward combining contexts, help learn better parsing models?

English Results



Experimental Results



25k token TD corpus

Conclusion

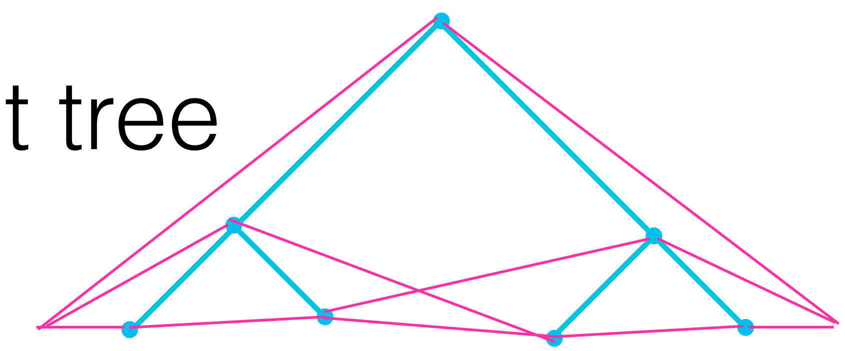
Under weak supervision, we can use
universal grammatical knowledge about **context**
to find trees with a **better global structure**.

Deficiency

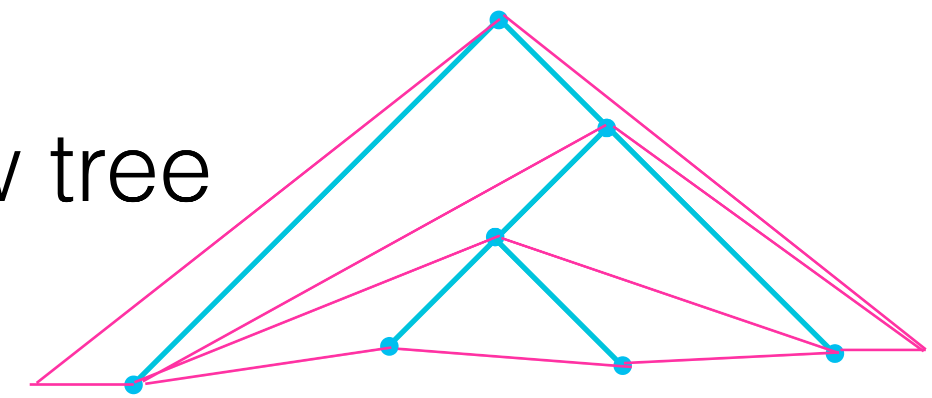
- Generative story has a “throw away” step if the context-generated nonterminals don’t match the tree.
- We sample only over the space of valid trees (condition on well-formed structures).
- This is a benefit of the Bayesian formulation.
- See Smith 2011.

Metropolis-Hastings

current tree



new tree



$$P_{\text{context}}(\mathbf{y}) = P_{\text{full}}(\mathbf{y}) \;/\; P_{\text{pcfg}}(\mathbf{y}) \qquad P_{\text{context}}(\mathbf{y'}) = P_{\text{full}}(\mathbf{y'}) \;/\; P_{\text{pcfg}}(\mathbf{y'})$$

$z \sim$ uniform(0,1)

$$\text{accept if } z < \frac{P_{\text{full}}(\mathbf{y'}) \,/\, P_{\text{pcfg}}(\mathbf{y'})}{P_{\text{full}}(\mathbf{y}) \;/\; P_{\text{pcfg}}(\mathbf{y})} = \frac{P_{\text{context}}(\mathbf{y'})}{P_{\text{context}}(\mathbf{y})}$$