*- "This is how you randomize a beam search!"*

Wouter Kool, Herke van Hoof, Max Welling

*- "You will never have duplicate samples again!"*

# STOCHASTIC BEAMS AND WHERE TO FIND THEM

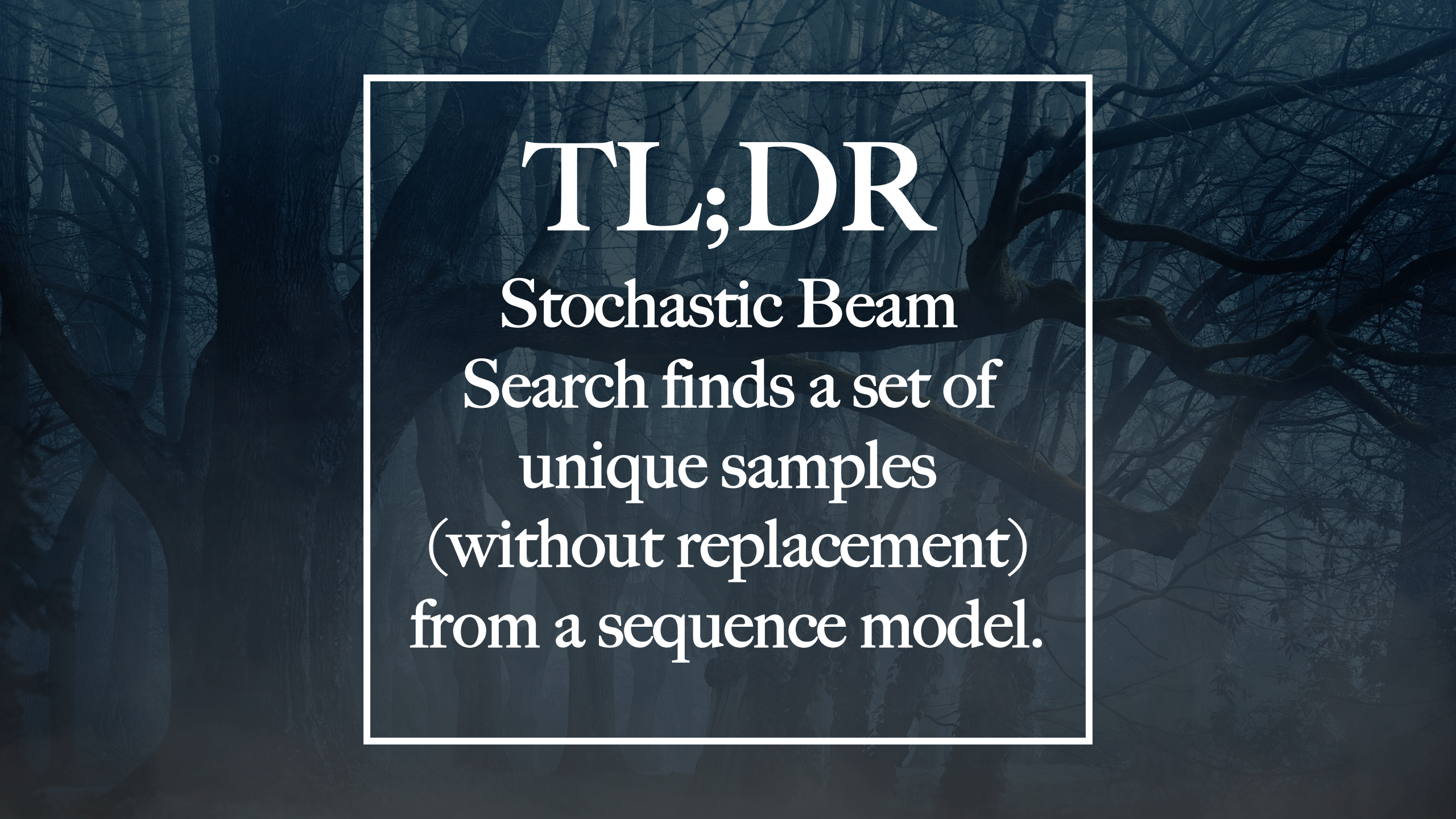
## The Gumbel-Top-$k$ Trick for Sampling Sequences Without Replacement


UNIVERSITY OF AMSTERDAM

ORTEC
OPTIMIZE YOUR WORLD

AMLAB
Amsterdam
Machine Learning Lab

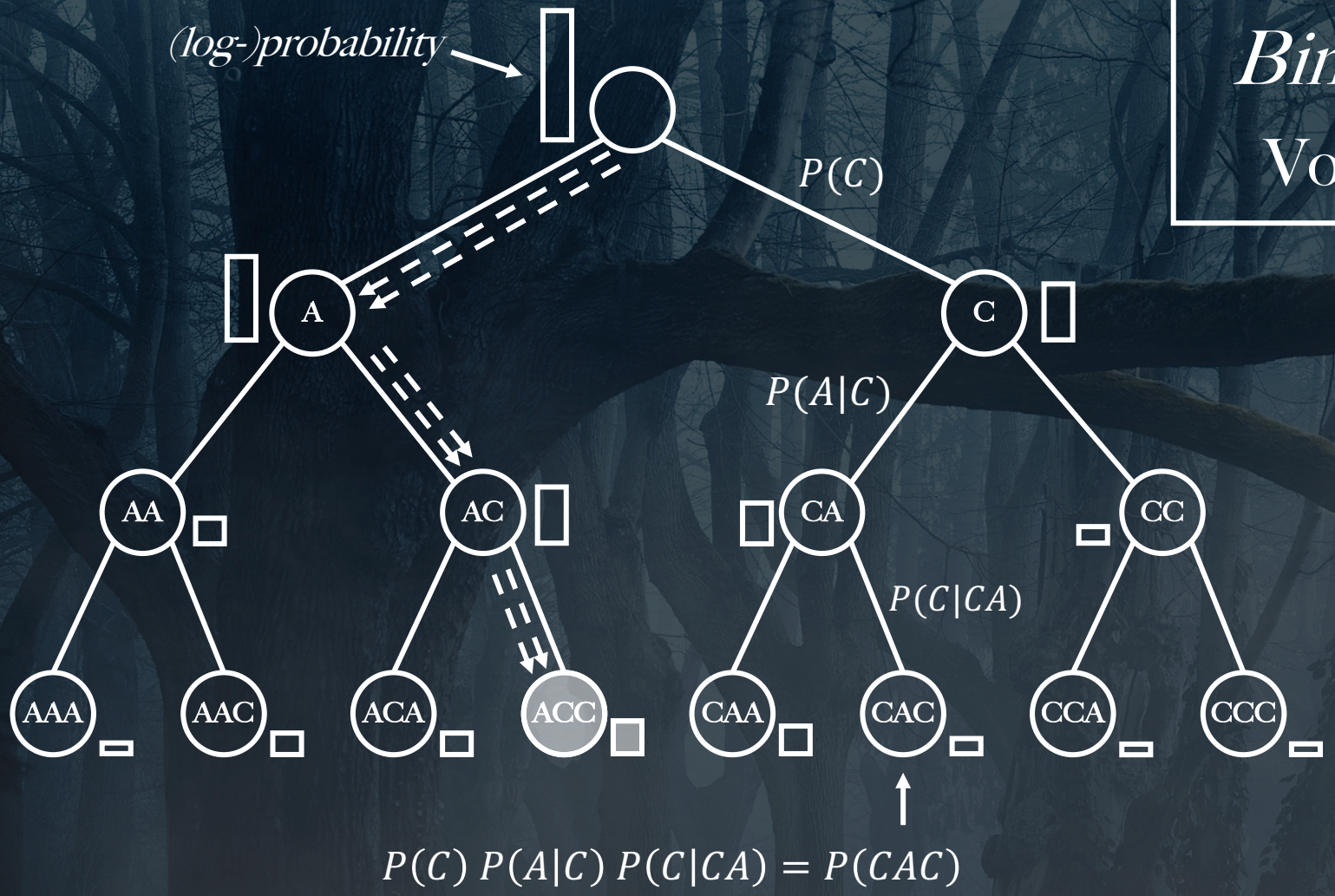
★★★★★
ICML 2019
BEST PAPER
HONORABLE MENTION


TL;DR
Stochastic Beam Search finds a set of unique samples (without replacement) from a sequence model.

# Example

*Binarese* language model

Vocabulary: {**A**bra, **C**adabra}




*What if we want a **sample** from our model?*

# The Gumbel-Max Trick

*"Prof. Gumbeldore"*

(Gumbel, 1945; Maddison et al., 2014)

$\phi_i = \log p_i$

log-probability

$G_i \sim \text{Gumbel}(0)$

Gumbel noise

$G_{\phi_i} \sim \text{Gumbel}(\phi_i)$

*perturbed log-probability*

# The Gumbel-Max Trick

*"Prof. Gumbeldore"*

(Gumbel, 1945; Maddison et al., 2014)

$$\max_i G_{\phi_i} \sim \text{Gumbel}\left(\log \sum_i \exp \phi_i\right)$$

max and argmax are *independent*

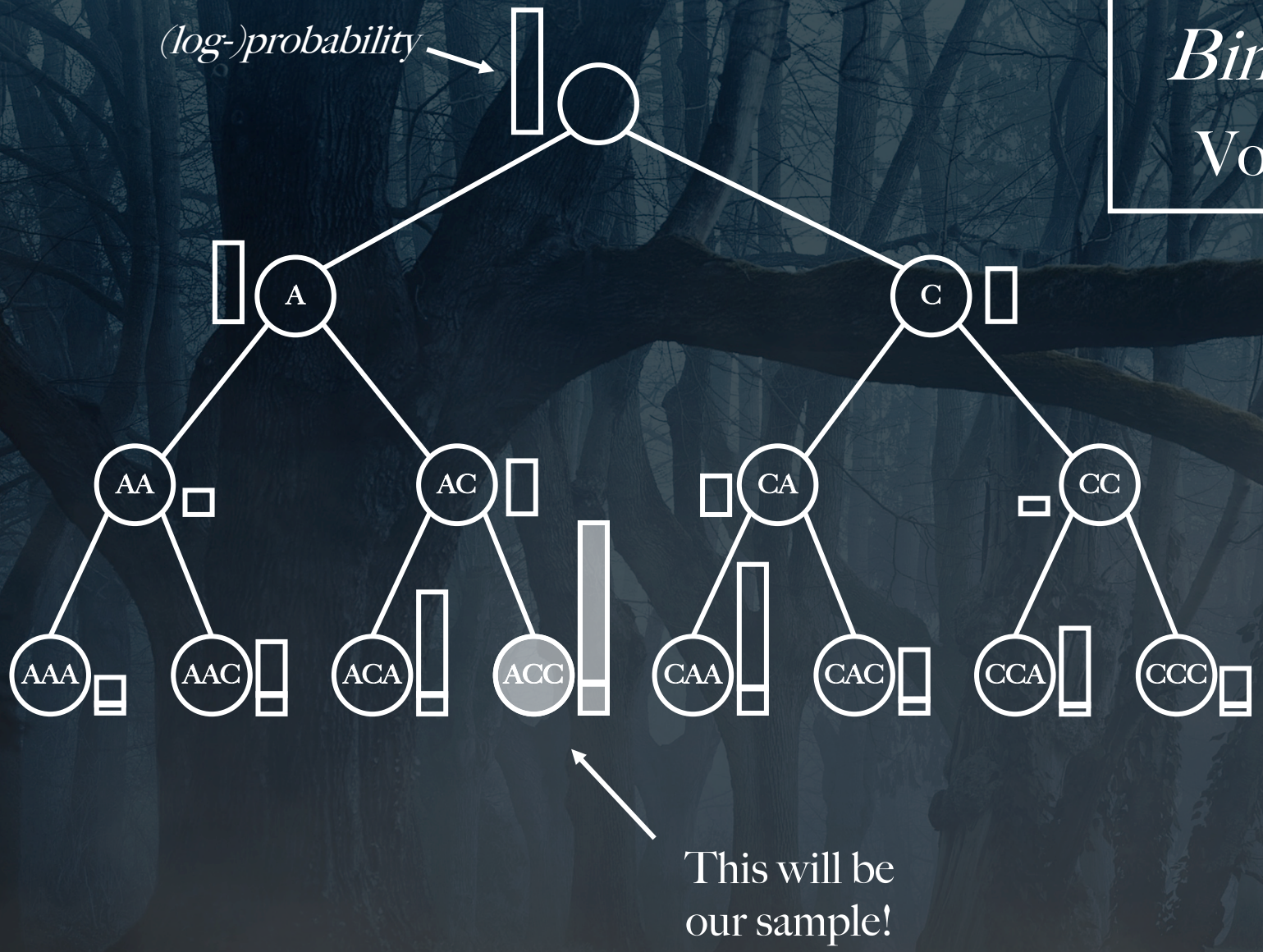
$$I = \operatorname{argmax}_i G_{\phi_i} \sim \text{Categorical}(p_i)$$

$$P(I = i) = p_i$$

# Example

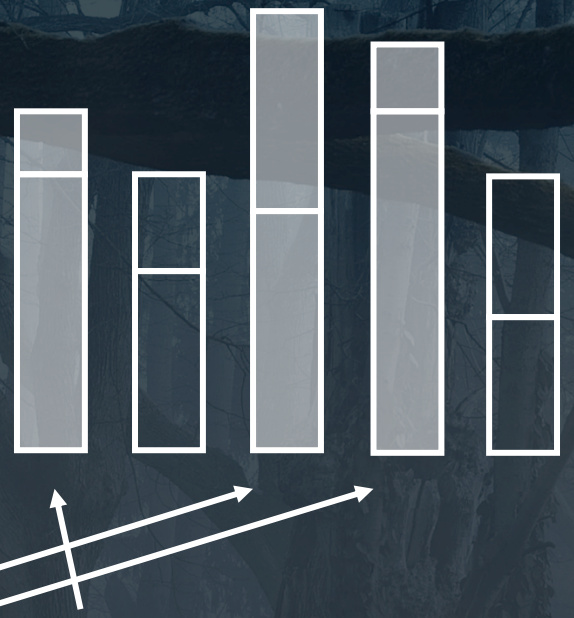
*Binarese* language model

Vocabulary: {**A**bra, **C**adabra}




*What if we want a **sample** from our model?*

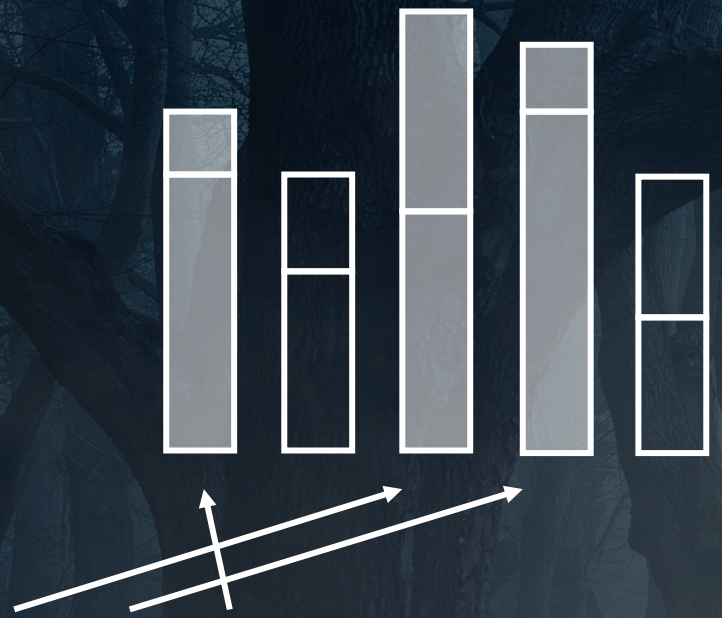
*What happens if, instead of 1 (one), we take the k largest elements (top k)?*



$k = 3$

$$I_1, \ldots, I_k = \underset{i}{\arg\operatorname{top}\, k}\; G_{\phi_i}$$

# The 'Gumbel-Top-$k$' Trick



$$I_1, \dots, I_k = \arg\operatorname*{top}_i k \, G_{\phi_i}$$

$$P(I_1 = i_1, \dots, I_k = i_k)$$
$$= p_{i_1} \cdot \frac{p_{i_2}}{1 - p_{i_1}} \cdot \dots \cdot \frac{p_{i_k}}{1 - \sum_{\ell=1}^{k-1} p_{i_\ell}}$$
$$= \prod_{j=1}^{k} \frac{p_{i_j}}{1 - \sum_{\ell=1}^{j-1} p_{i_\ell}}$$

Also known as Plackett-Luce

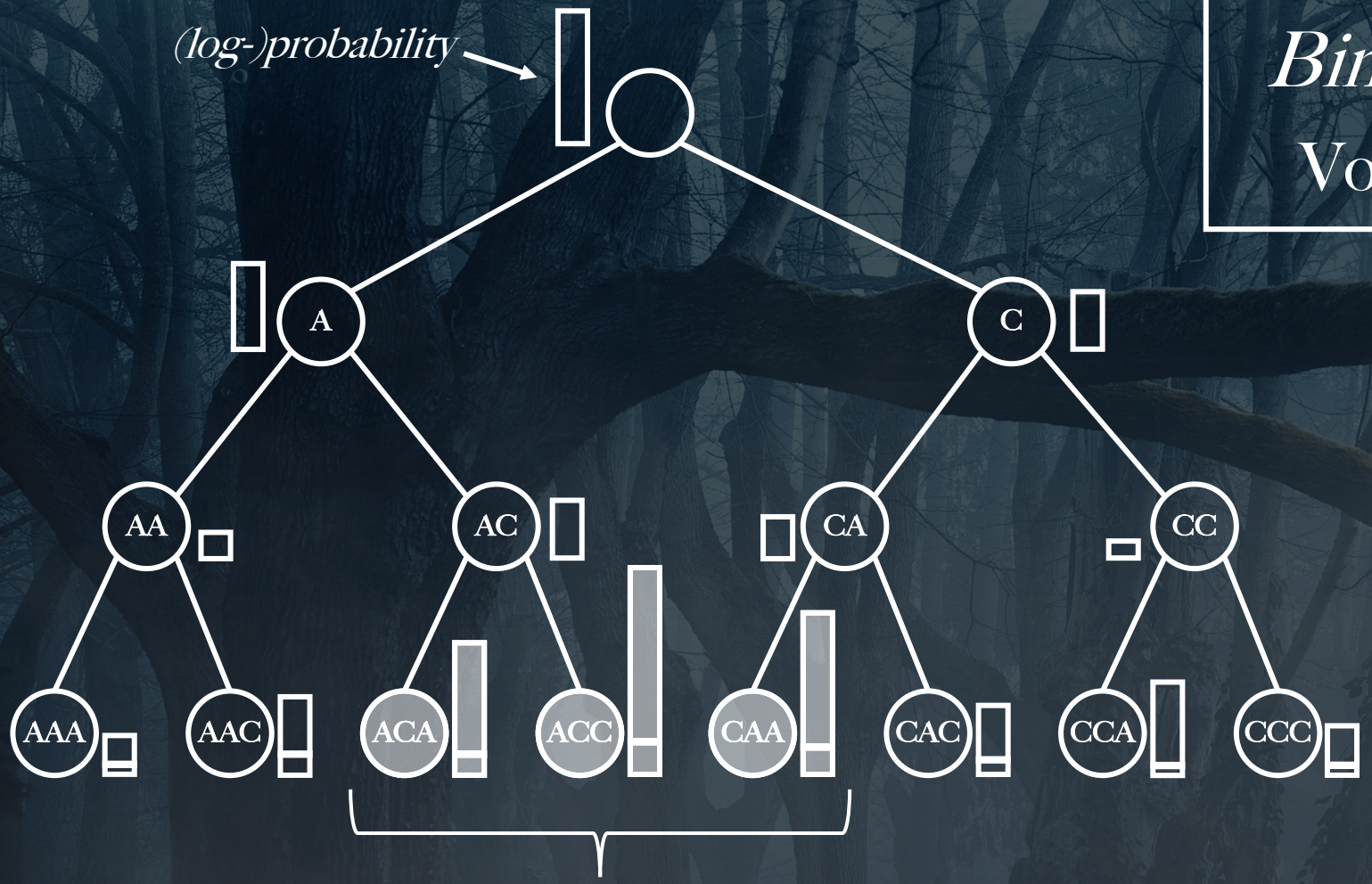
*This is equivalent to repeated sampling without replacement!*

(Vieira, 2014)

# Example

*Binarese* language model

Vocabulary: {**A**bra, **C**adabra}



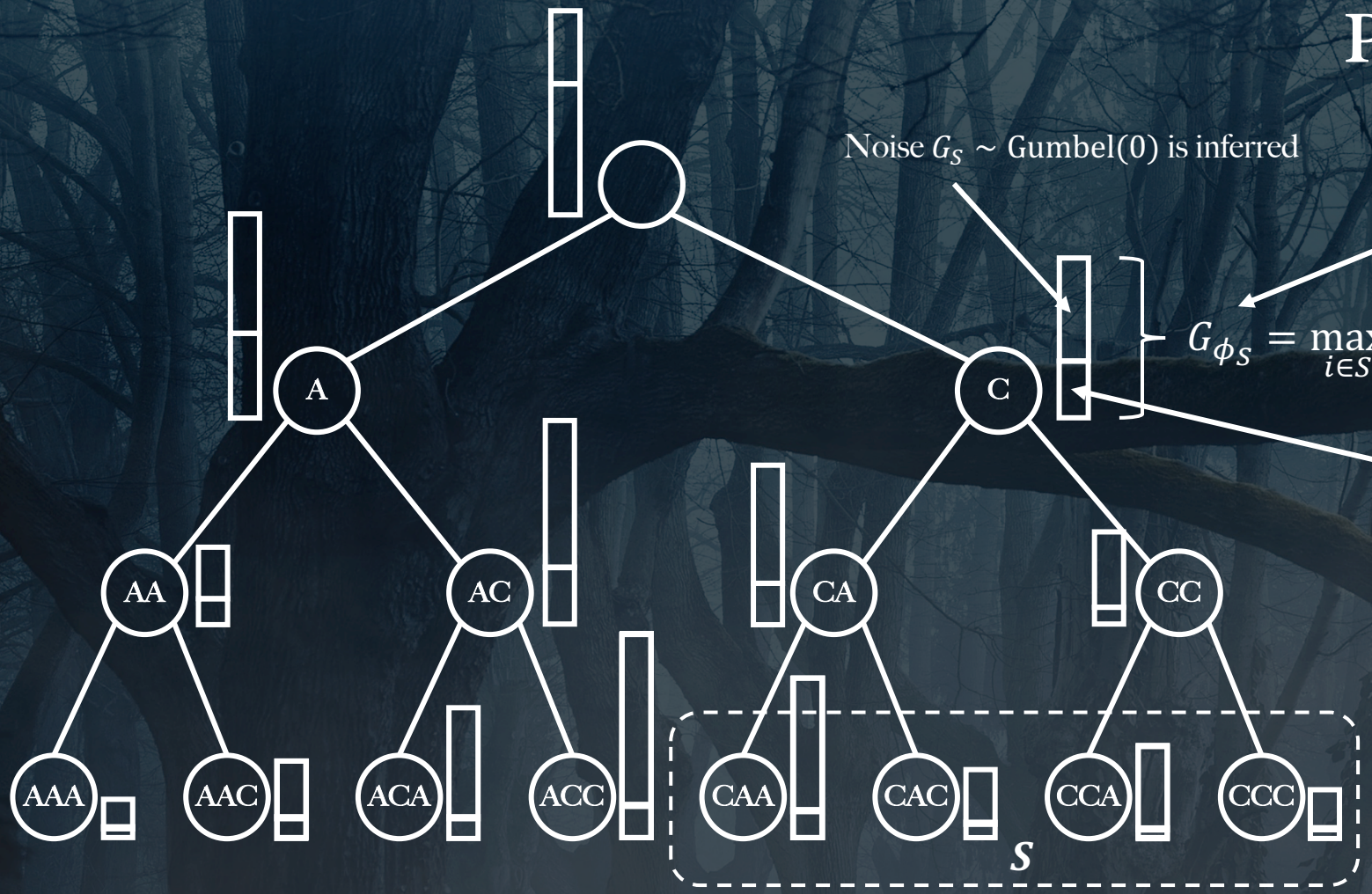

This will be our set of samples!

*We can get a set of unique samples from our model!*

# PROBLEM

In general, constructing the full tree is not possible...

... but we don't have to!




Noise $G_S \sim \text{Gumbel}(0)$ is inferred

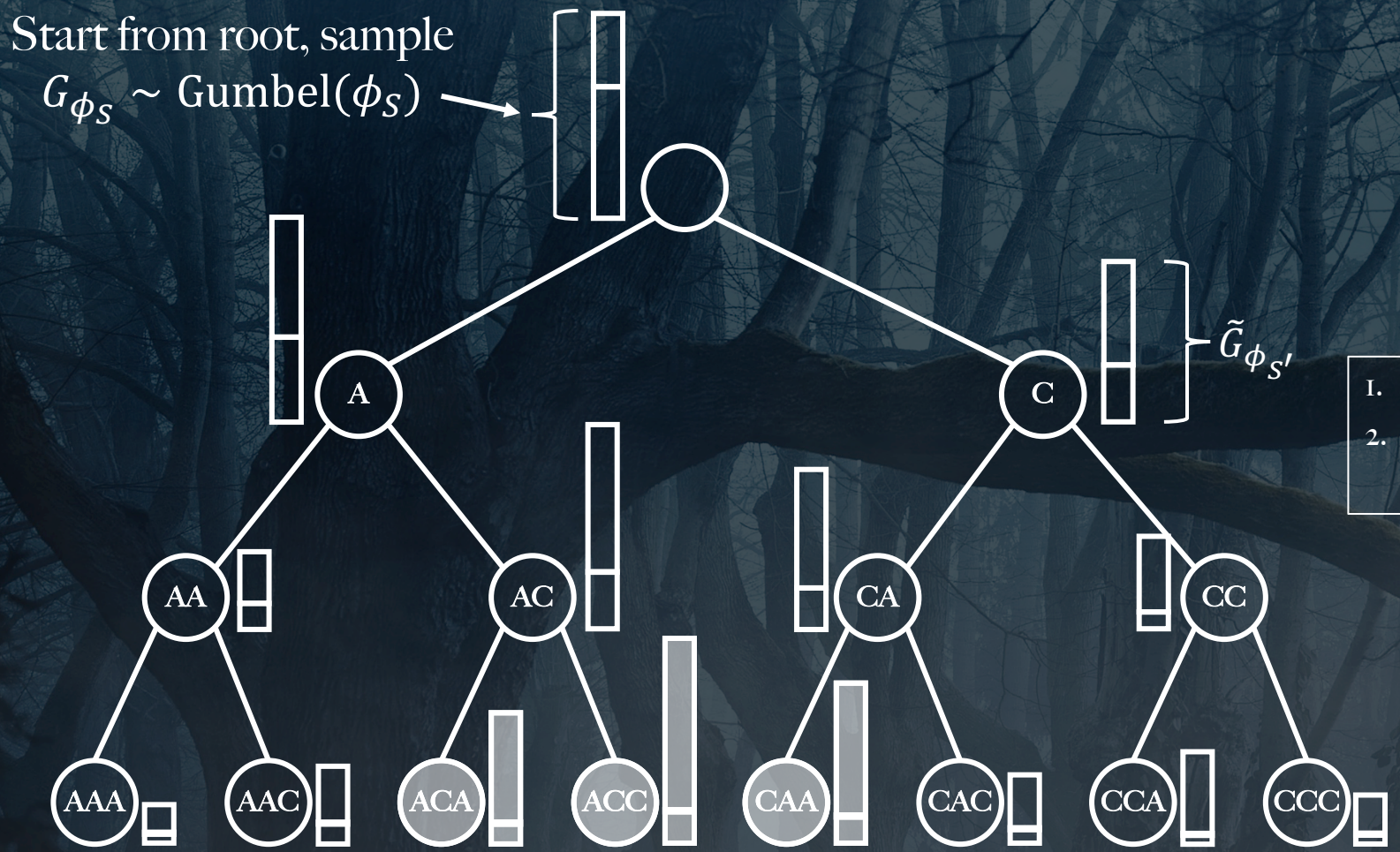
## Perturbed log-probability of partial sequence ("C")

$$G_{\phi_S} = \max_{i \in S} G_{\phi_i} \sim \text{Gumbel}\left(\log \sum_{i \in S} \exp \phi_i\right)$$

$\phi_S$ = log-probability of "C"

We can sample $G_{\phi_S} \sim \text{Gumbel}(\phi_S)$ **directly**

Look at maximum of perturbed log-probabilities in subtree




Sample children $G_{\phi_{S'}}$ *conditionally on* $\max_{S' \in \text{Children}(S)} G_{\phi_{S'}} = G_{\phi_S}$

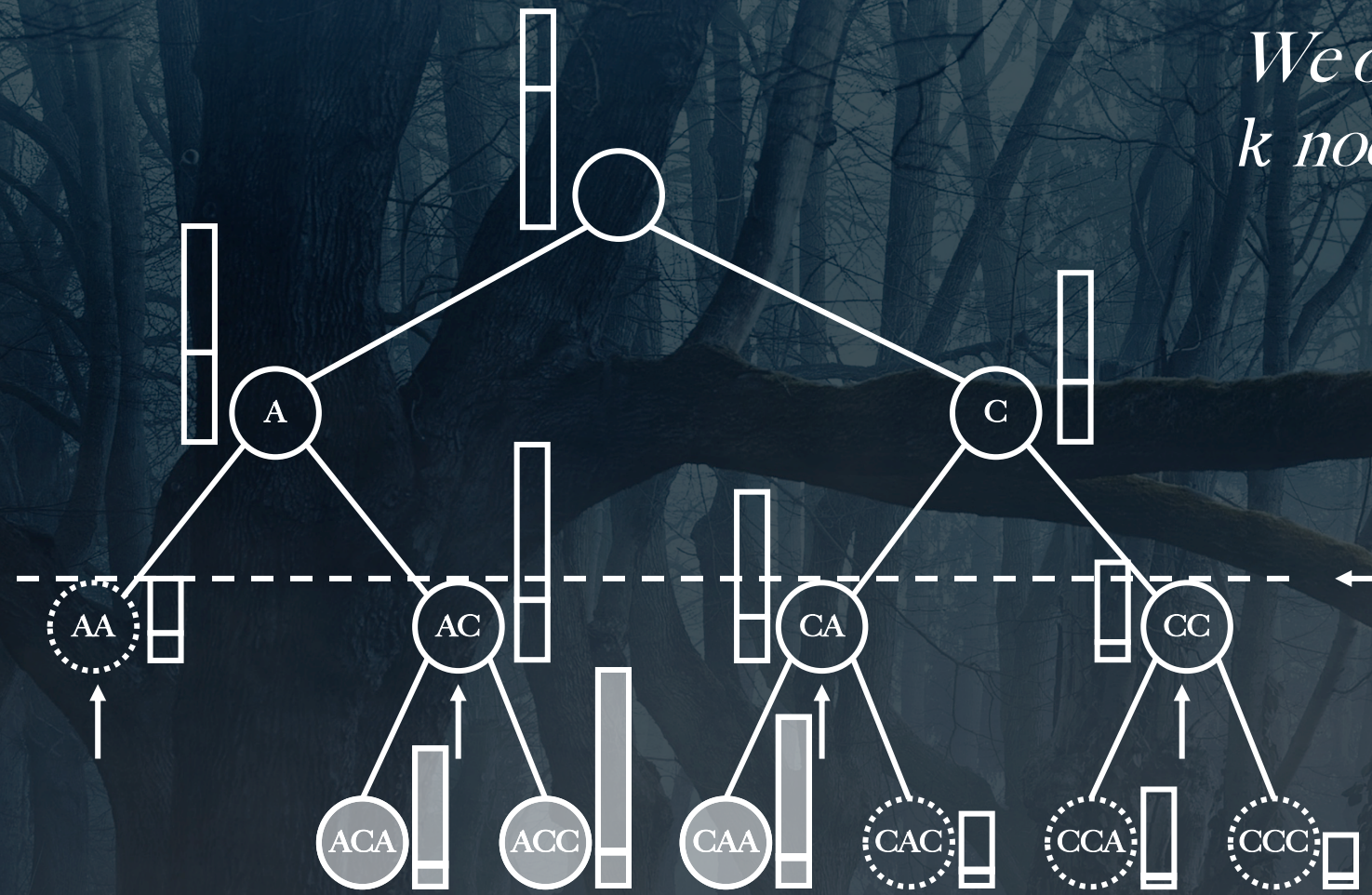
1. sample $G_{\phi_{S'}}$ independently, compute $Z = \max_{S'} G_{\phi_{S'}}$
2. 'shift' Gumbels in (negative) exponential space:
$\tilde{G}_{\phi_{S'}} = -\log\left(\exp(-G_{\phi_S}) - \exp(-Z) + \exp\left(-G_{\phi_{S'}}\right)\right)$

... the result is **equivalent** to sampling $G_{\phi_i}$ for leaves directly!

# Top-down sampling

(Maddison et al., 2014)

*We only need to expand the top k nodes at each level in the tree*




Threshold

Each top *k* node generates (at least) one leaf (maximum) above threshold

At least *k* leafs will be above threshold

Other nodes only generate leafs below threshold

No need to expand

# The Key Insight


# Stochastic Beam Search

*We only need to expand the top k nodes at each level in the tree*

↑

This is a *beam search*

Top *k* according to perturbed log-probability

= ← Gumbel-Top-*k* trick

Sampling (without replacement)

# Stochastic Beam Search

- A beam search that *samples* the nodes to expand

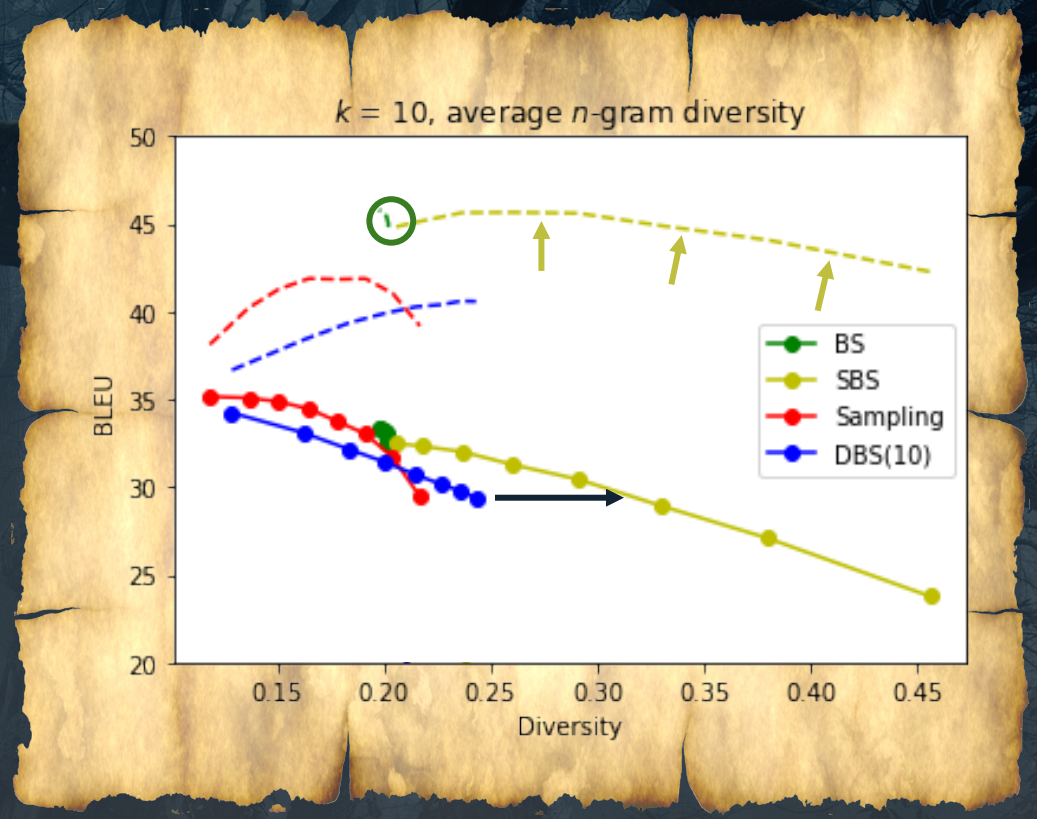
Important!

- But... samples children *conditionally* on parent
- The result is a sample without replacement from the full sequence model
- Is a generalization of ancestral sampling ($k = 1$)

# Experiments

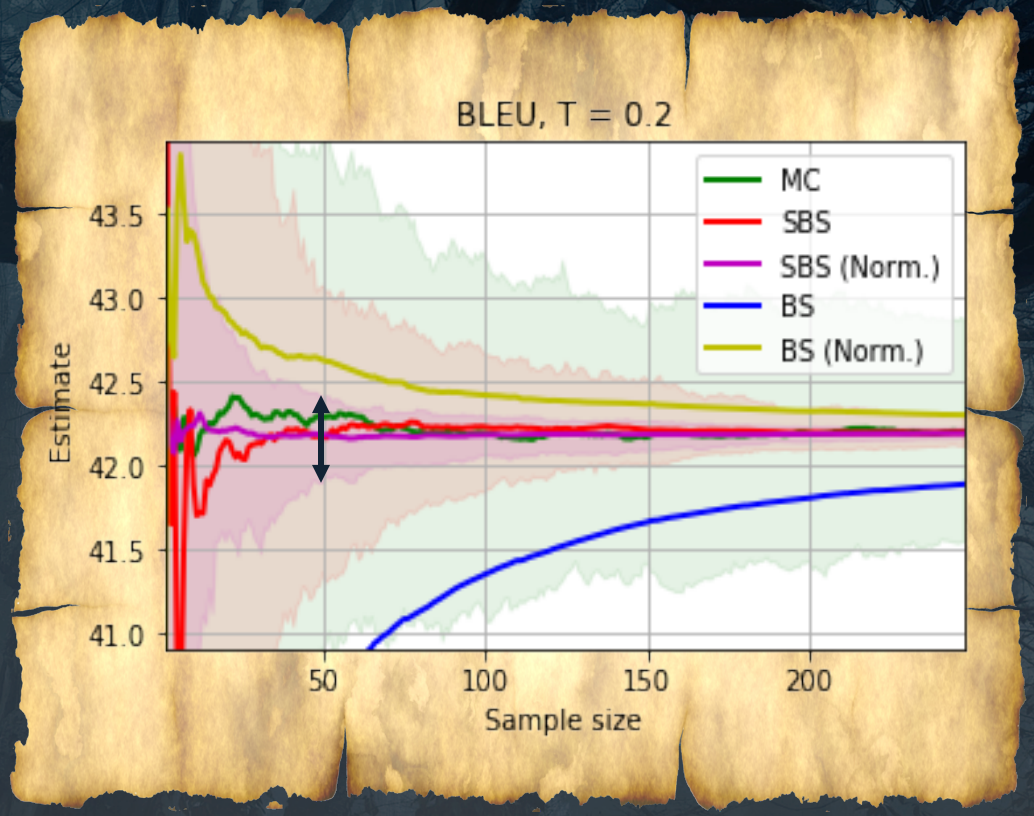
# Translation Diversity

- Generate $k$ translations
- Plot BLEU against diversity
- Vary softmax temperature
- Compare:
  - Beam Search
  - Stochastic Beam Search
  - Sampling
  - Diverse Beam Search (Vijayakumar et al., 2018)



# BLEU Score Estimation

- Estimate expected sentence-level BLEU
- Plot mean and 95% interval vs. num samples
- Compare:
  - Monte Carlo Sampling
  - Stochastic Beam Search with (normalized) Importance Weighted estimator
  - Beam Search with deterministic estimate



Wouter Kool, Herke van Hoof, Max Welling

# STOCHASTIC BEAMS AND WHERE TO FIND THE POSTER?


UNIVERSITY OF AMSTERDAM

ORTEC
OPTIMIZE YOUR WORLD

AMLAB
Amsterdam
Machine Learning Lab

#41