Bayesian Nonparametric Mixture, Admixture, and Language Models

Yee Whye Teh
University of Oxford

Nov 2015

Overview

- Bayesian nonparametrics and random probability measures
  - Mixture models and clustering
- Hierarchies of Dirichlet processes
  - Modelling document collections with topic models
  - Modelling genetic admixtures in human populations
- Hierarchies of Pitman-Yor processes
  - Language modelling with high-order Markov models and power law statistics
  - Non-Markov language models with the sequence memoizer

Bayesian Nonparametrics

- Data $x_1,\ldots,x_n$ assume iid from an underlying distribution μ:

$$x_i|\mu \overset{\text{iid}}{\sim} \mu$$

- Inference on μ nonparametrically, within a Bayesian framework:

$$\mu \sim \mathcal{P}$$

- "There are two desirable properties of a prior distribution for nonparametric problems:

  (I) The support of the prior distribution should be large—with respect to some suitable topology on the space of probability distributions on the sample space.

  (II) Posterior distributions given a sample of observations from the true probability distribution should be manageable analytically."

— Ferguson (1973)

[Hjort et al (eds) 2010]

Dirichlet Process

- Random probability measure

$$\mu \sim \mathrm{DP}(\alpha, H)$$

- For each partition $(A_1, \ldots A_m)$,

$$(\mu(A_1), \ldots, \mu(A_m)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_m))$$

- Cannot use Kolmogorov Consistency Theorem to construct the DP:
  - Space of probability measures not in the product σ-field on $[0,1]^B$.
  - Use a countable generator $\mathcal{F}$ for B and view $\mu \in [0,1]^{\mathcal{F}}$.
- Easier constructions:
  - Define an infinitely exchangeable sequence with directing random measure μ.
  - Define a gamma process and normalizing it.
  - Explicit construction using the stick-breaking process.

[Ferguson 1973, Blackwell-McQueen 1973, Sethuraman 1994, Pitman 2006]

Dirichlet Process

- Analytically tractable posterior distribution.

- Well-studied process:
  - ranked-ordered masses have Poisson-Dirichlet distribution.
  - Size-bias permuted masses have simple iid Beta structure.
  - Corresponding exchangeable random partition described by the Chinese restaurant process.

- Large support over space of probability measures in weak topology.
  - Variety of convergence (and non-convergence) results.
- Draws from DP are discrete w.p. 1.

Dirichlet Process Mixture Models

- Draws from DPs are discrete probability measures:

$$\mu = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$$

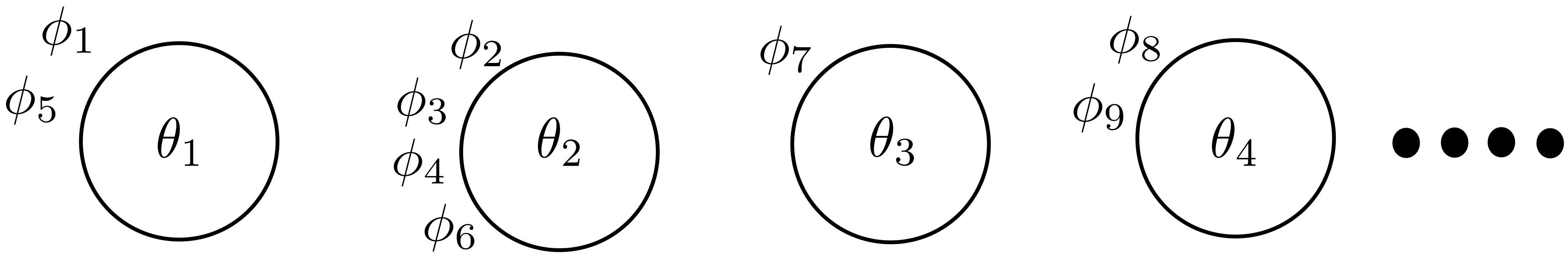
  where w_k, θ_k are random.
- Typically use within a hierarchical model,

$$\phi_i | \mu \overset{\text{iid}}{\sim} \mu \qquad\qquad x_i | \phi_i \sim F(\phi_i)$$

  leading to nonparametric mixture models.
- Discrete nature of μ induces repeated values among $\phi_{1:n}$.
  - Induces a partition Π of $[n] = \{1,\ldots,n\}$.
  - Leads to a clustering model with an unbounded/infinite number of clusters.
- Properties of model for cluster analysis depends on the properties of the induced random partition Π (a Chinese restaurant process (CRP)).
- Generalisations of DPs allow for more flexible prior specifications.

[Antoniak 1974, Lo 1984]

Chinese Restaurant Processes



$$p(\text{sit at table } k) = \frac{c_k}{\alpha + \sum_{j=1}^{K} c_j} \qquad p(\text{table serves dish } y) = H(y)$$

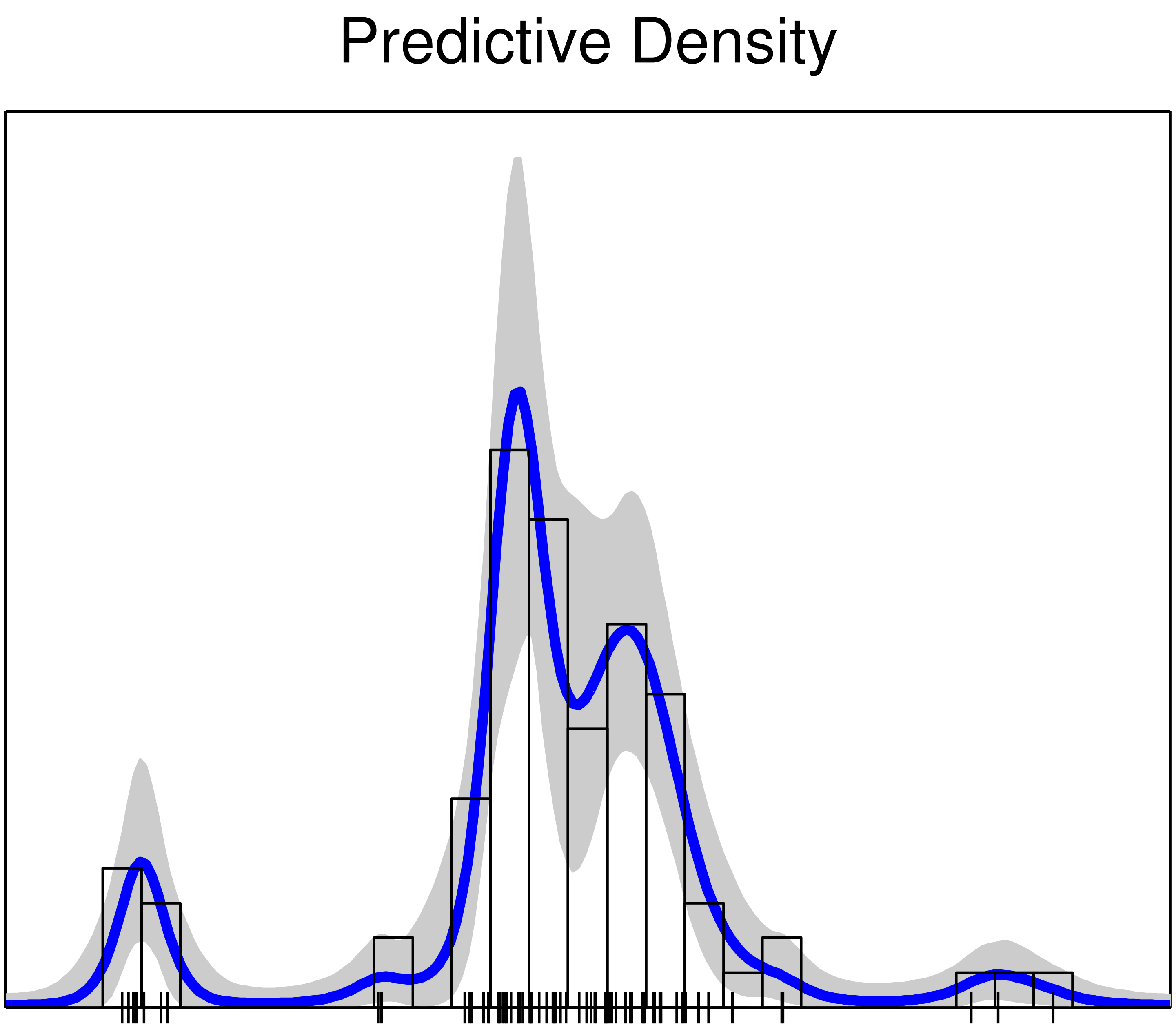
$$p(\text{sit at new table}) = \frac{\alpha}{\alpha + \sum_{j=1}^{K} c_j} \qquad i \text{ sits at table } j: \quad \phi_i = \theta_j$$

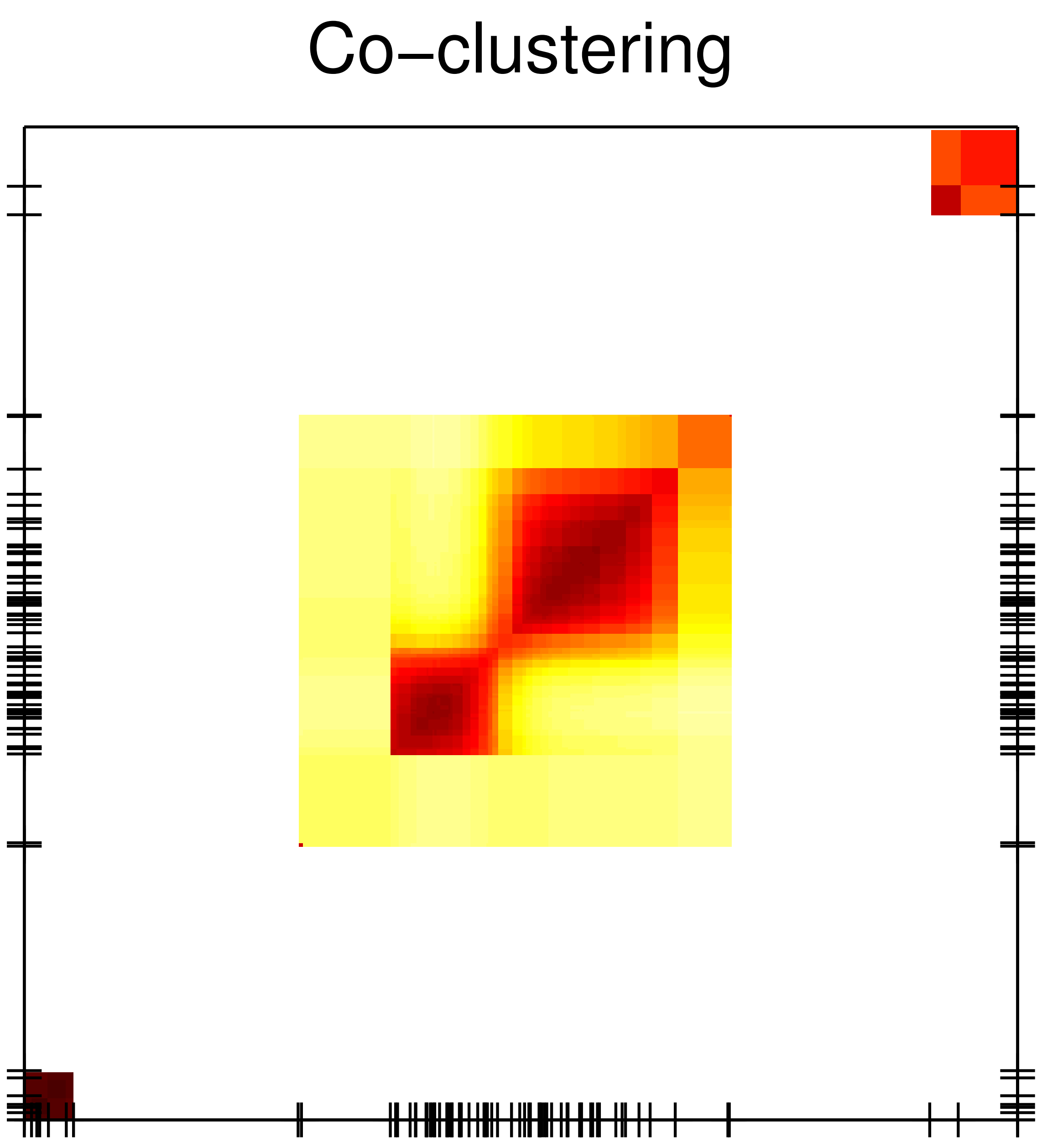
- Defines an exchangeable stochastic process over sequences $\phi_1, \phi_2, \ldots$
- The de Finetti measure [Kingman 1978] is the Dirichlet process,

$$\begin{aligned} \mu &\sim \mathrm{DP}(\alpha, H) \\ \phi_i | \mu &\sim \mu \qquad i = 1, 2, \ldots \end{aligned}$$

[Blackwell & McQueen 1973, Pitman 2006]

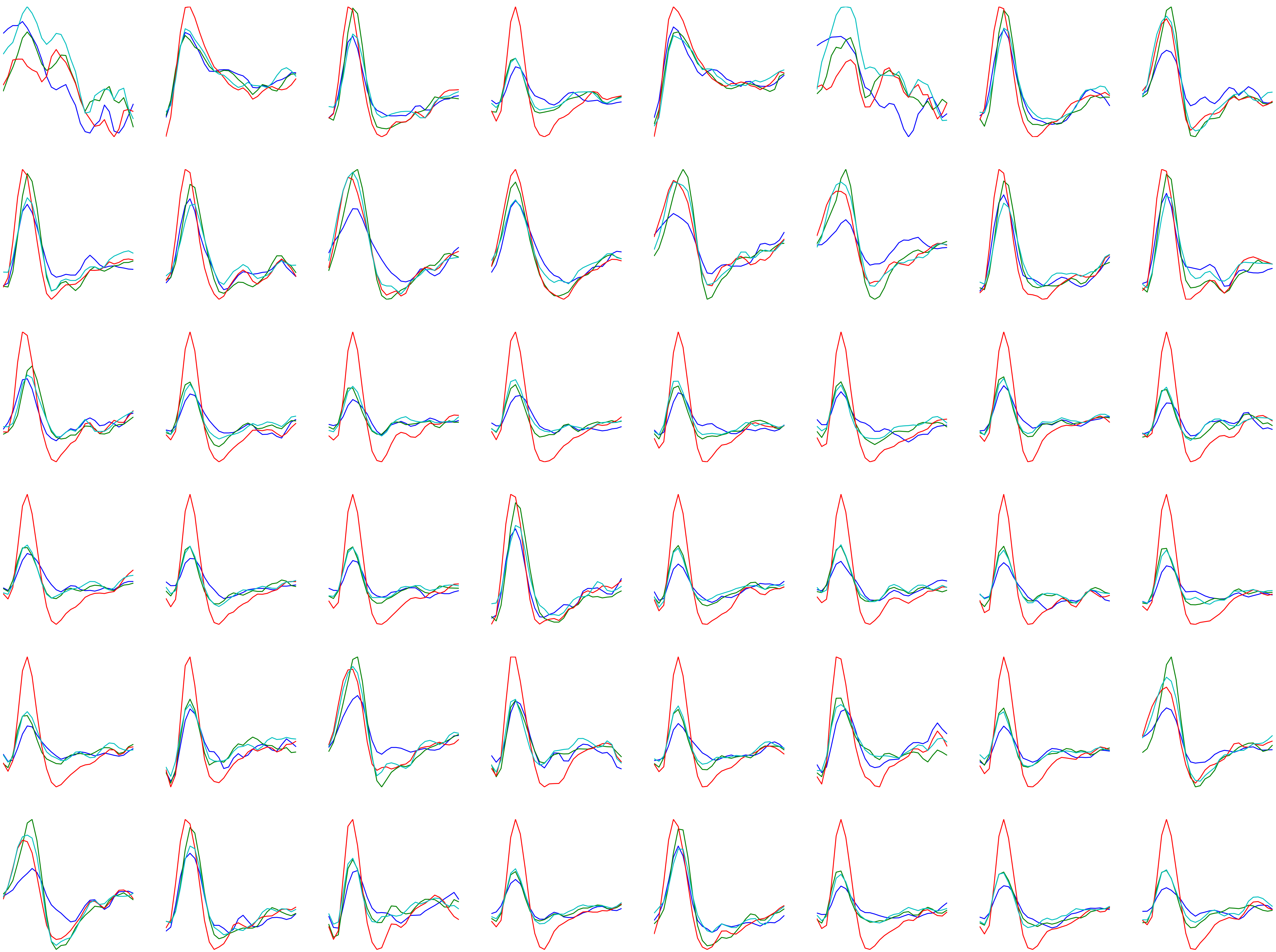
Density Estimation and Clustering



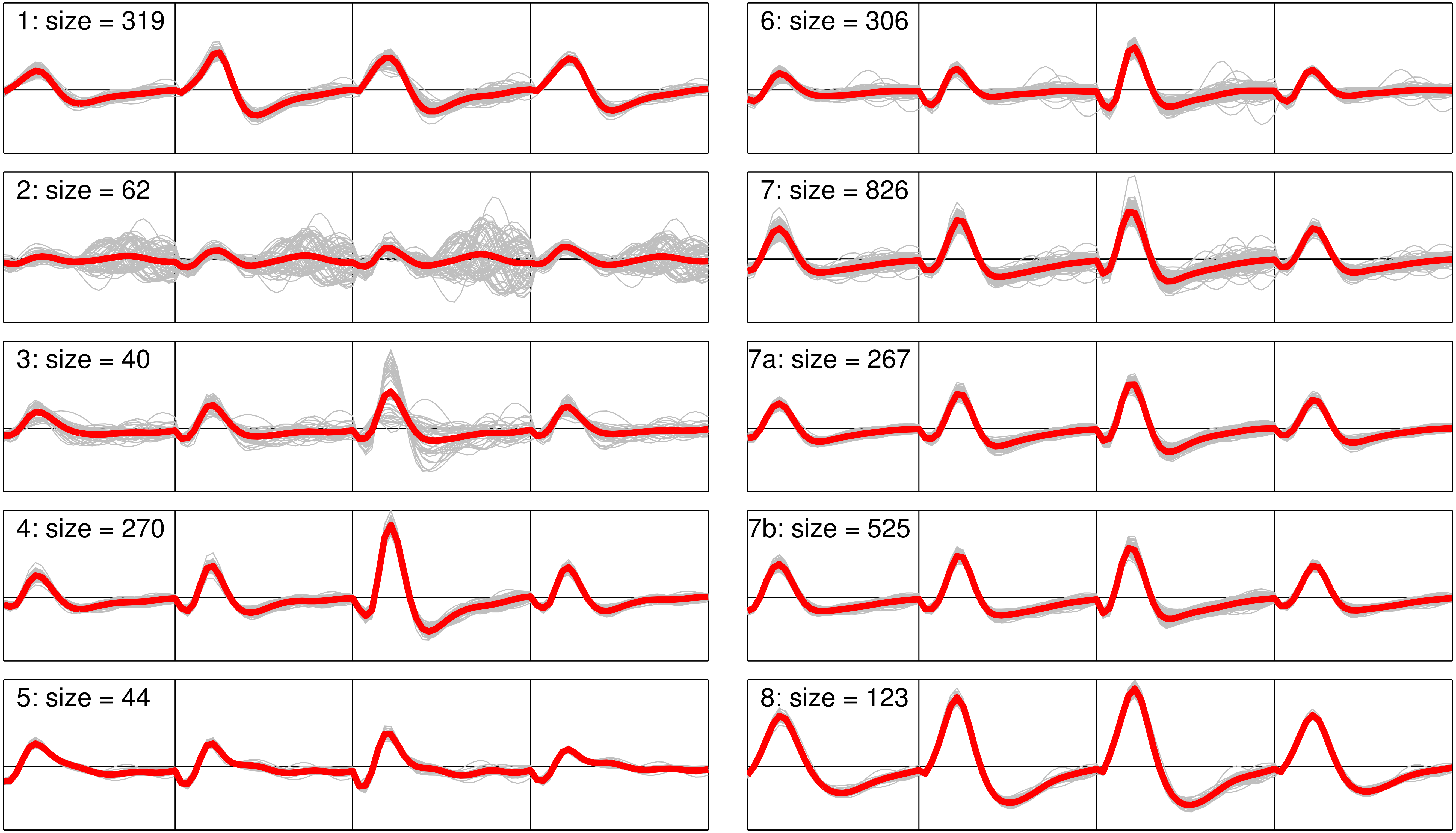


[Favaro & Teh 2013]

Spike Sorting

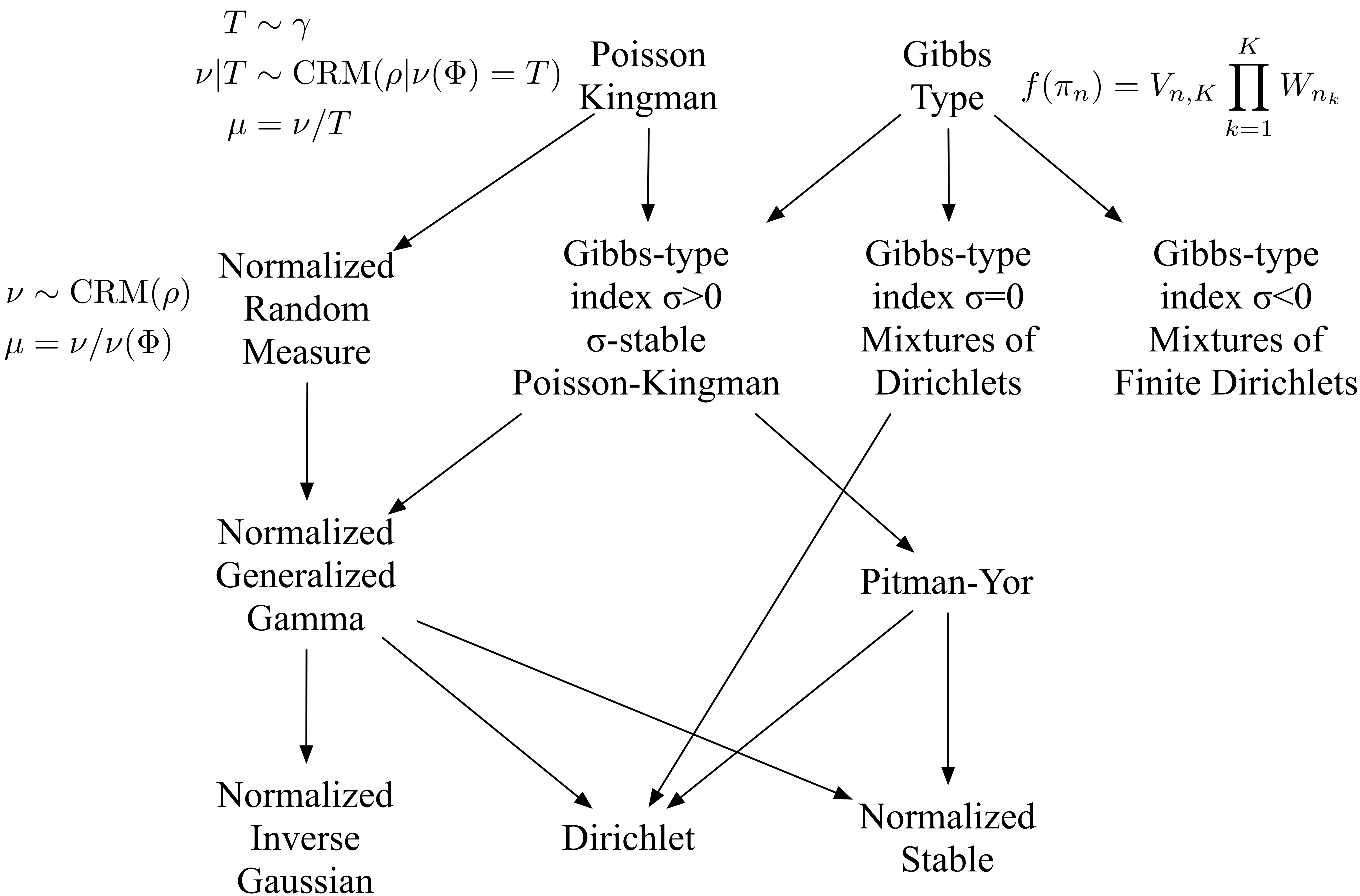


Spike Sorting



[Favaro & Teh 2013]

Families of Random Probability Measures



Gibbs Type Partitions

- An exchangeable random partition Π is of Gibbs type if
$$p(\Pi_n = \pi_n) = V_{n,K} \prod_{k=1}^{K} W_{n_k}$$
where π has K clusters with sizes $n_1, \ldots, n_K$.
- Exchangeability and Gibbs form implies that wlog:
$$W_m = (1-\sigma)(2-\sigma)\cdots(m-1-\sigma)$$
where $-\infty \leq \sigma \leq 1$.
- The number of clusters K grows with n, with asymptotic distribution
$$\frac{K_n}{f(n)} \to S_\sigma$$
for some random variable S_σ, where $f(n) = 1, \log n, n^\sigma$ for $\sigma < 0, = 0, > 0$.
- Choice of S_σ and σ arbitrary and part of prior specification.
  - $\sigma < 0$: Bayesian finite mixture model
  - $\sigma = 0$: DP mixture model with hyper prior on α
  - $\sigma > 0$: σ-stable Poisson-Kingman process mixture model

[Gnedin & Pitman 2006, De Blasi et al 2015, Lomeli et al 2015]

Other Uses of Random Probability Measures

- Species sampling [Lijoi, Pruenster, Favaro, Mena]
- Nonparametric regression [MacEachern, Dunson, Griffin etc]
- Flexibly modelling heterogeneity in data

- More general random measures:
  - Survival analysis [Hjort 1990]
  - Feature models [Griffiths & Ghahramani 2011, Broderick et al 2012]

- Building more complex models via different motifs:
  - hierarchical Bayes
  - measure-valued stochastic processes
  - spatial and temporal processes
  - relational models

[Hjort et al (eds) 2010]

Overview

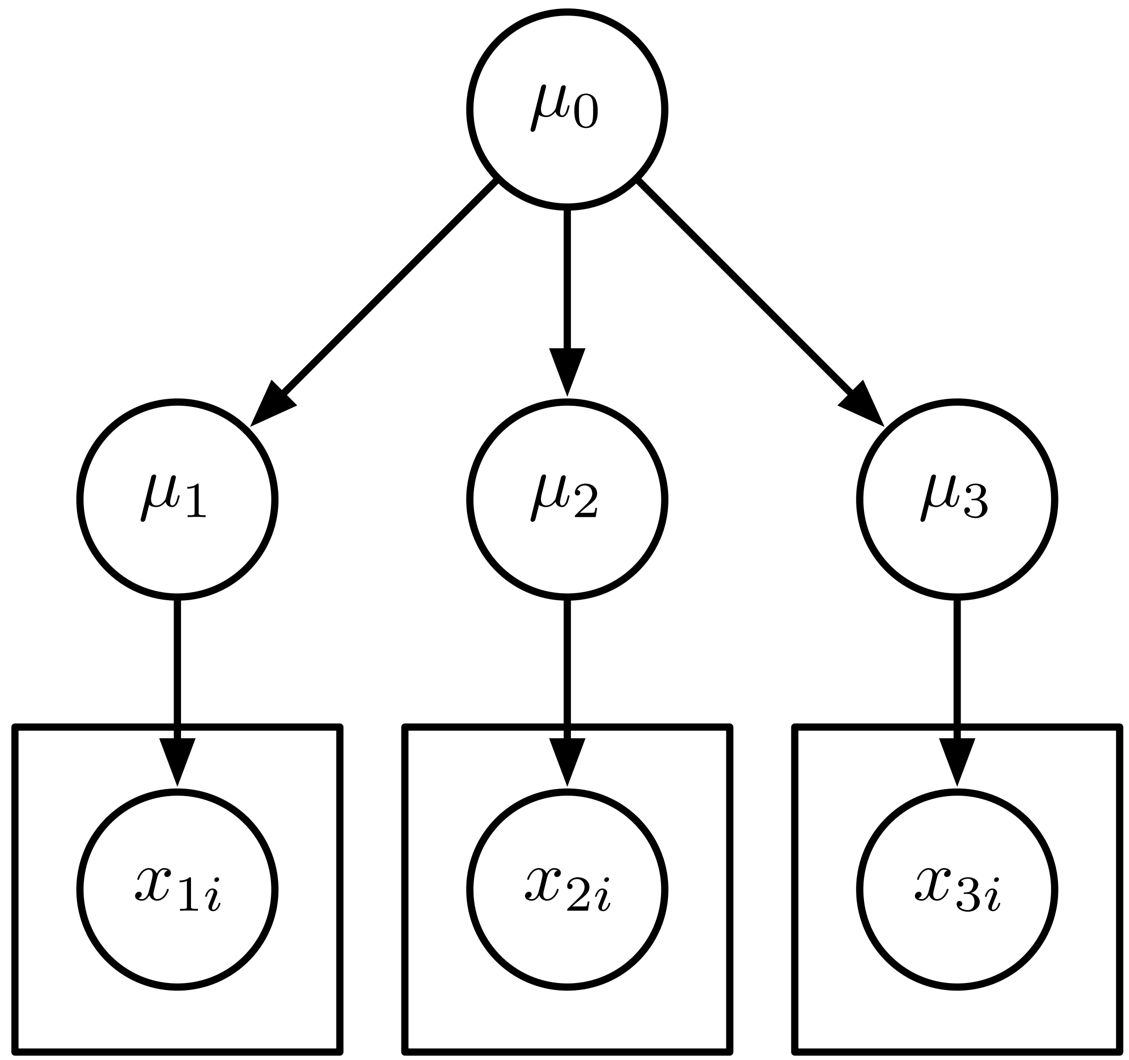
- Bayesian nonparametrics and random probability measures
  - Mixture models and clustering

- Hierarchies of Dirichlet processes
  - Modelling document collections with topic models
  - Modelling genetic admixtures in human populations

- Hierarchies of Pitman-Yor processes
  - Language modelling with high-order Markov models and power law statistics
  - Non-Markov language models with the sequence memoizer

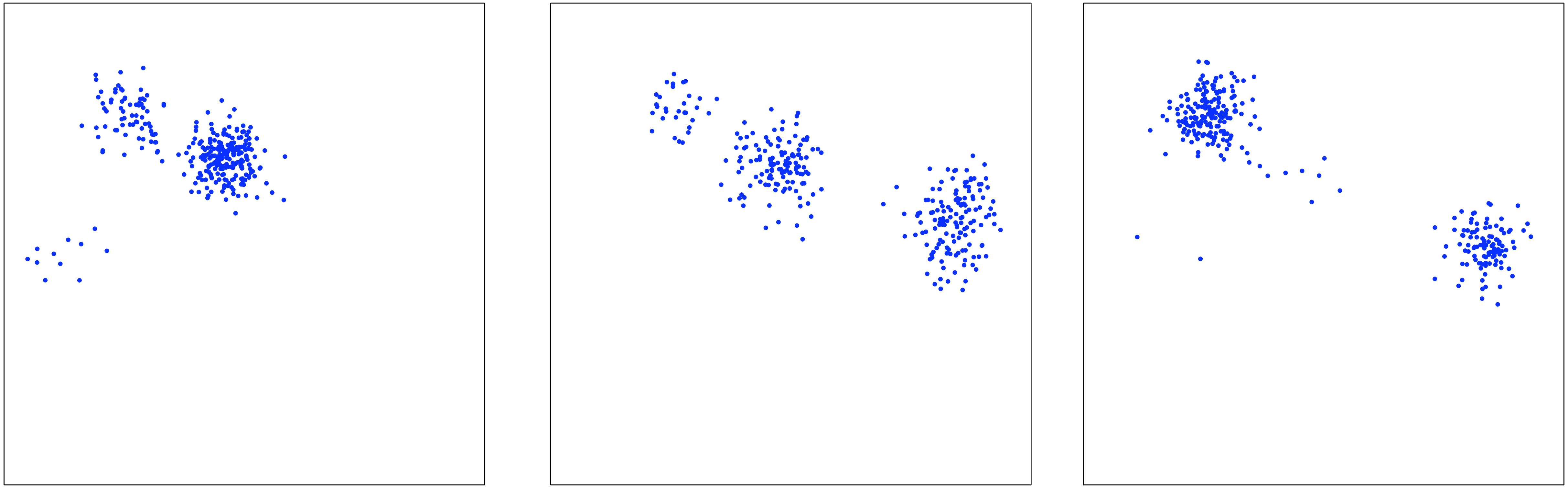
Hierarchical Bayesian Models

- Hierarchical modelling an important overarching theme in modern statistics [Gelman et al, 1995, James & Stein 1961].



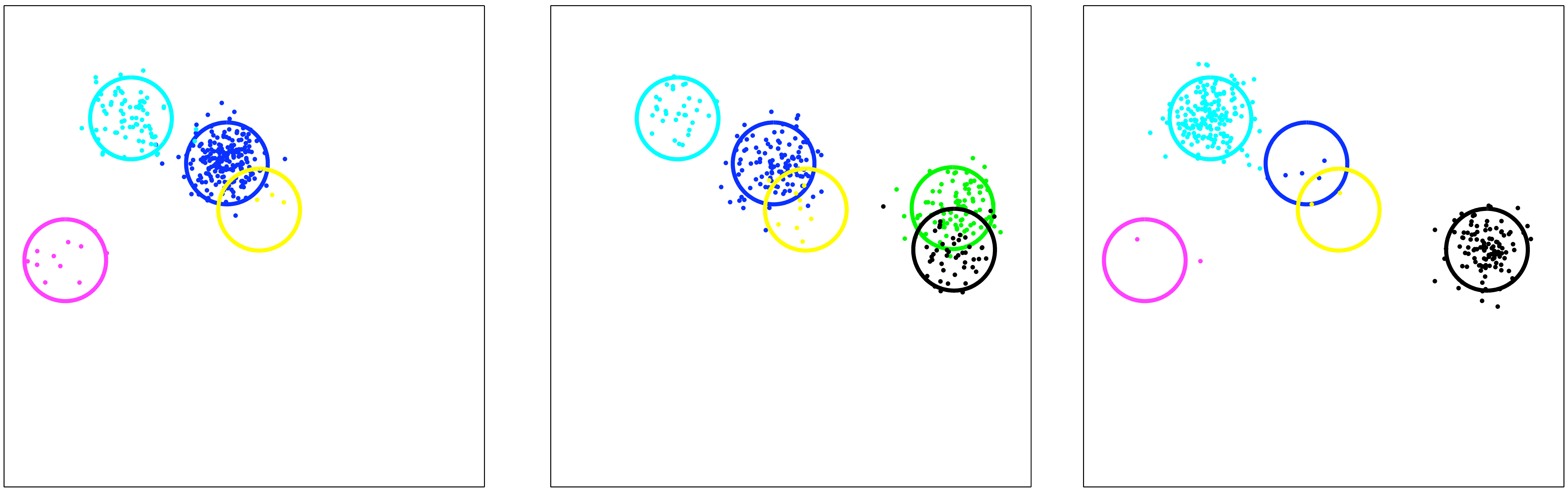
- In machine learning, have been used for multitask learning, transfer learning, learning-to-learn and domain adaptation.

Clustering of Related Groups of Data



- Multiple groups of data.
- Wish to cluster each group, using DP mixture models.
- Clusters are shared across multiple groups.

Clustering of Related Groups of Data



- Multiple groups of data.
- Wish to cluster each group, using DP mixture models.
- Clusters are shared across multiple groups.

Document Topic Modeling

- Model each document as a **bag of words** coming from an underlying set of topics [Hofmann 2001, Blei et al 2003].

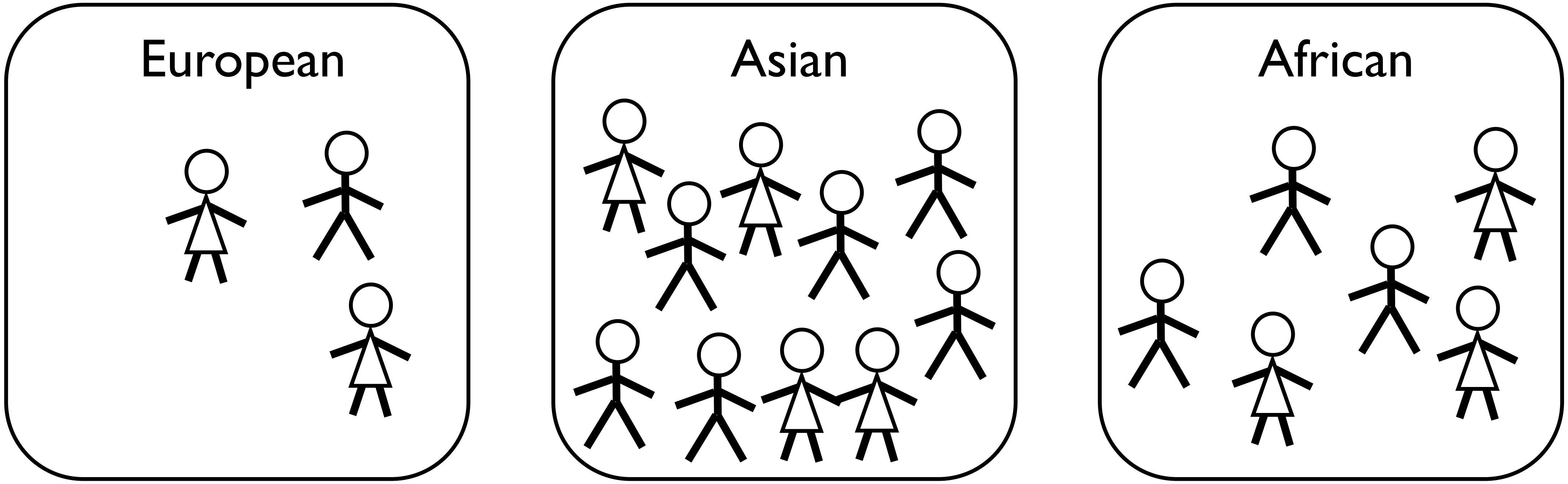
CARSON, Calif., April 3 - Nissan Motor Corp said it is raising the suggested retail price for its cars and trucks sold in the United States by 1.9 pct, or an average 212 dollars per vehicle, effective April 6....
DETROIT, April 3 - Sales of U.S.-built new cars surged during the last 10 days of March to the second highest levels of 1987. Sales of imports, meanwhile, fell for the first time in years, succumbing to price hikes by foreign carmakers.....

Auto industry
Market economy
US geography
Plain old English

- Summarize documents.
- Document/query comparisons.

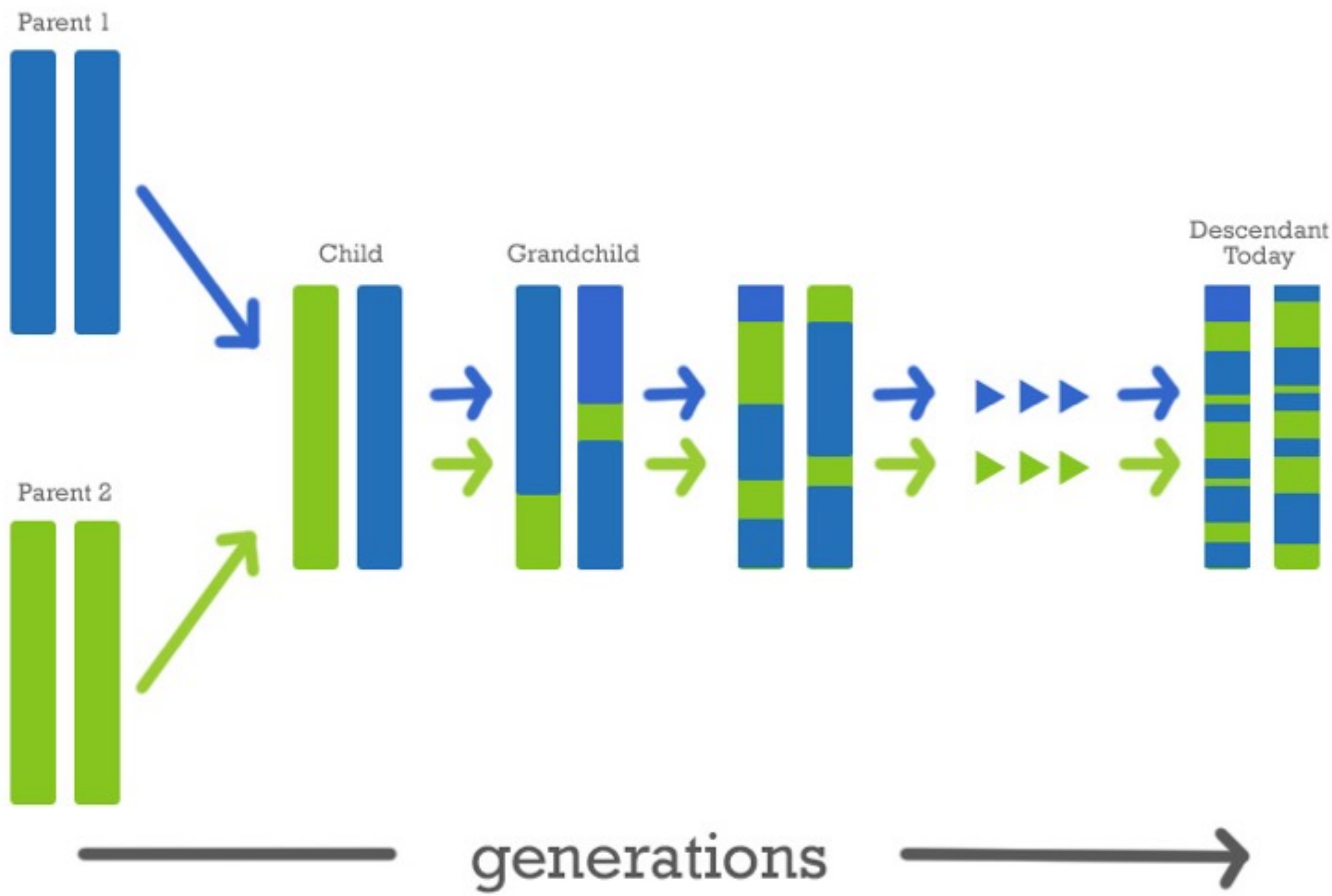
- Topics are shared across documents.
- Don't know #topics beforehand.

Multi-Population Genetics

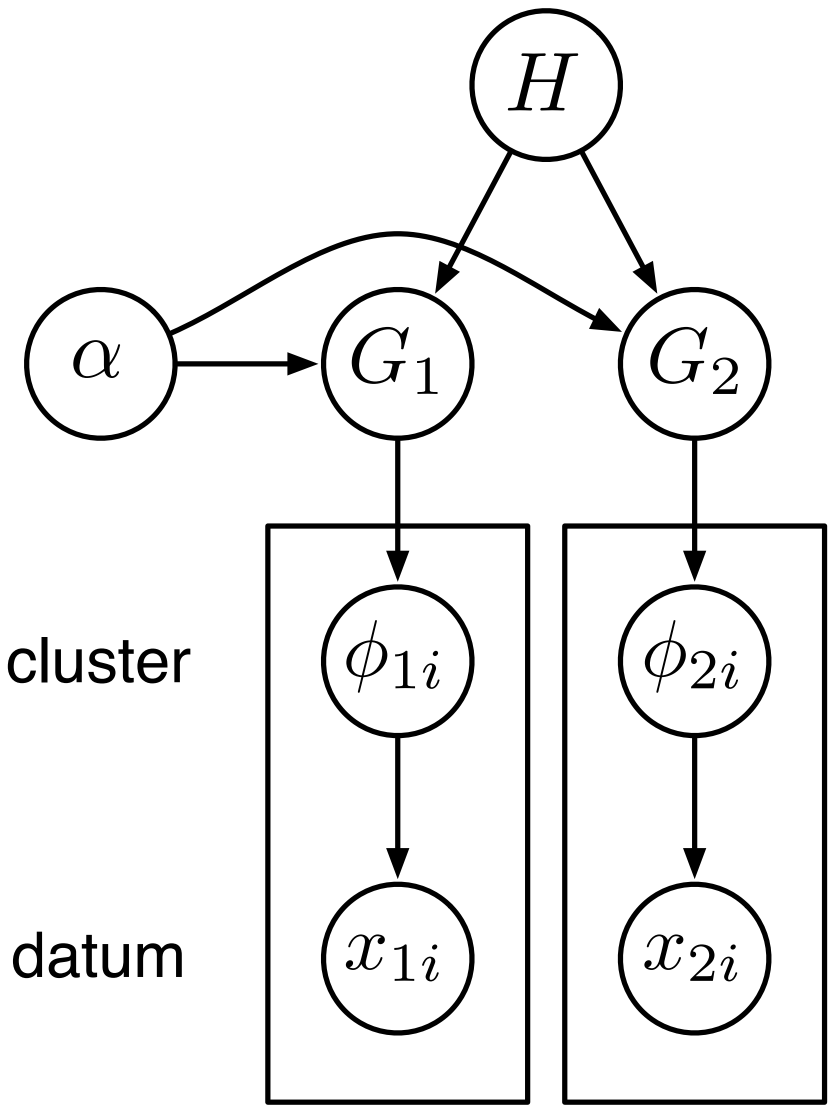


- Individuals can be clustered into a number of genotypes, with each population having a different proportion of genotypes [Xing et al 2006].
- Sharing genotypes among individuals in a population, and across different populations.
- Indeterminate number of genotypes.

Genetic Admixtures

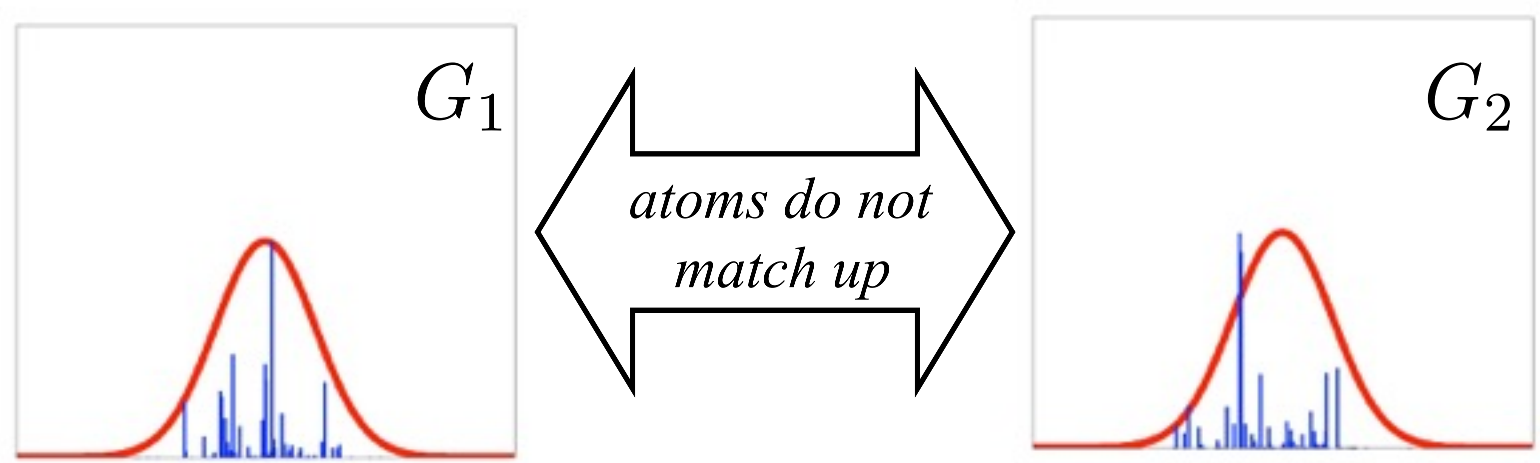


Dirichlet Process Mixture for Grouped Data?



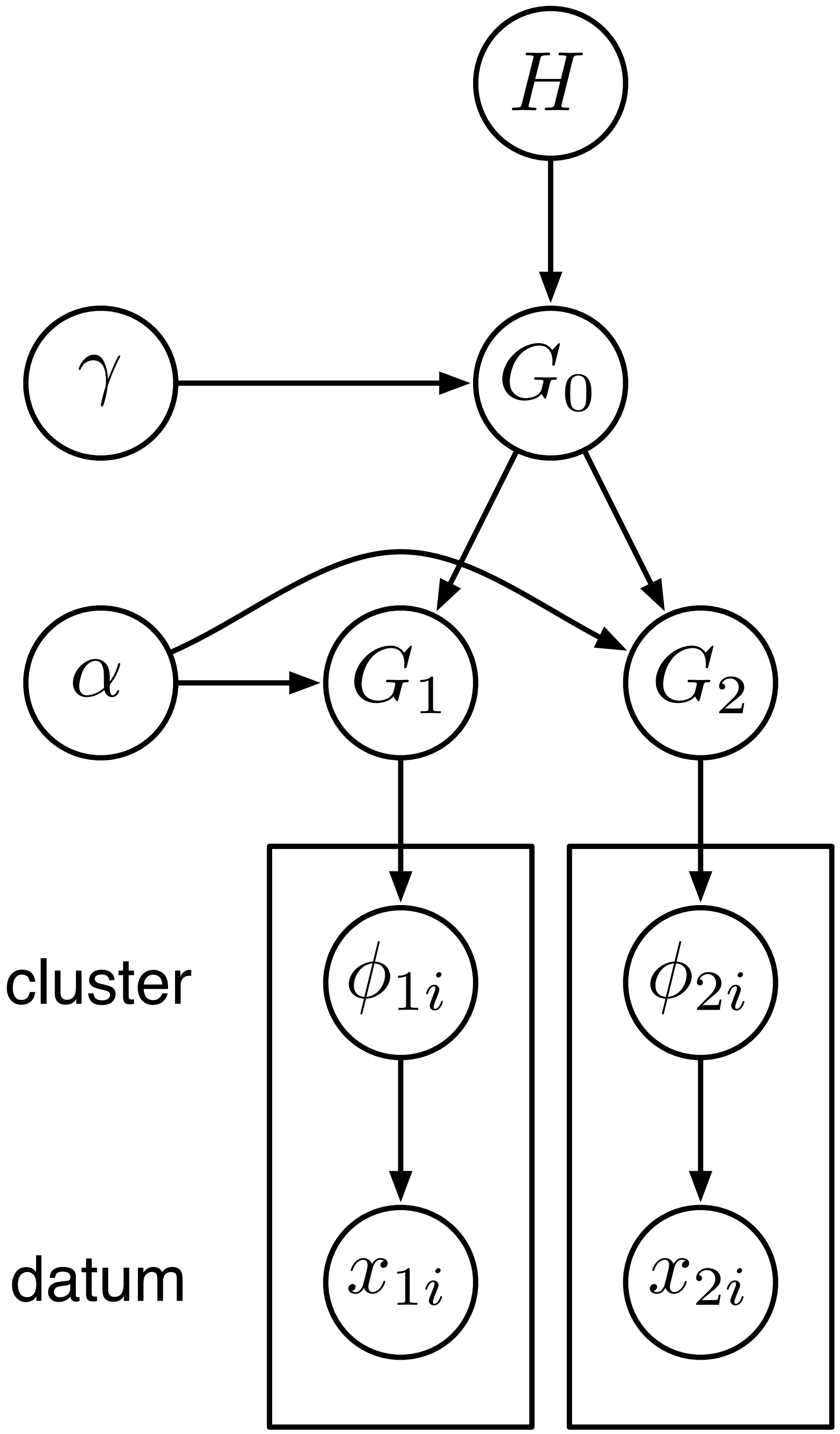
- Introduce dependencies between groups by making parameters random?

- If H is smooth, then clusters will not be shared between groups.



- But if the base distribution were discrete….

Hierarchical Dirichlet Process Mixture Models



- Making base distribution discrete forces groups to share clusters.

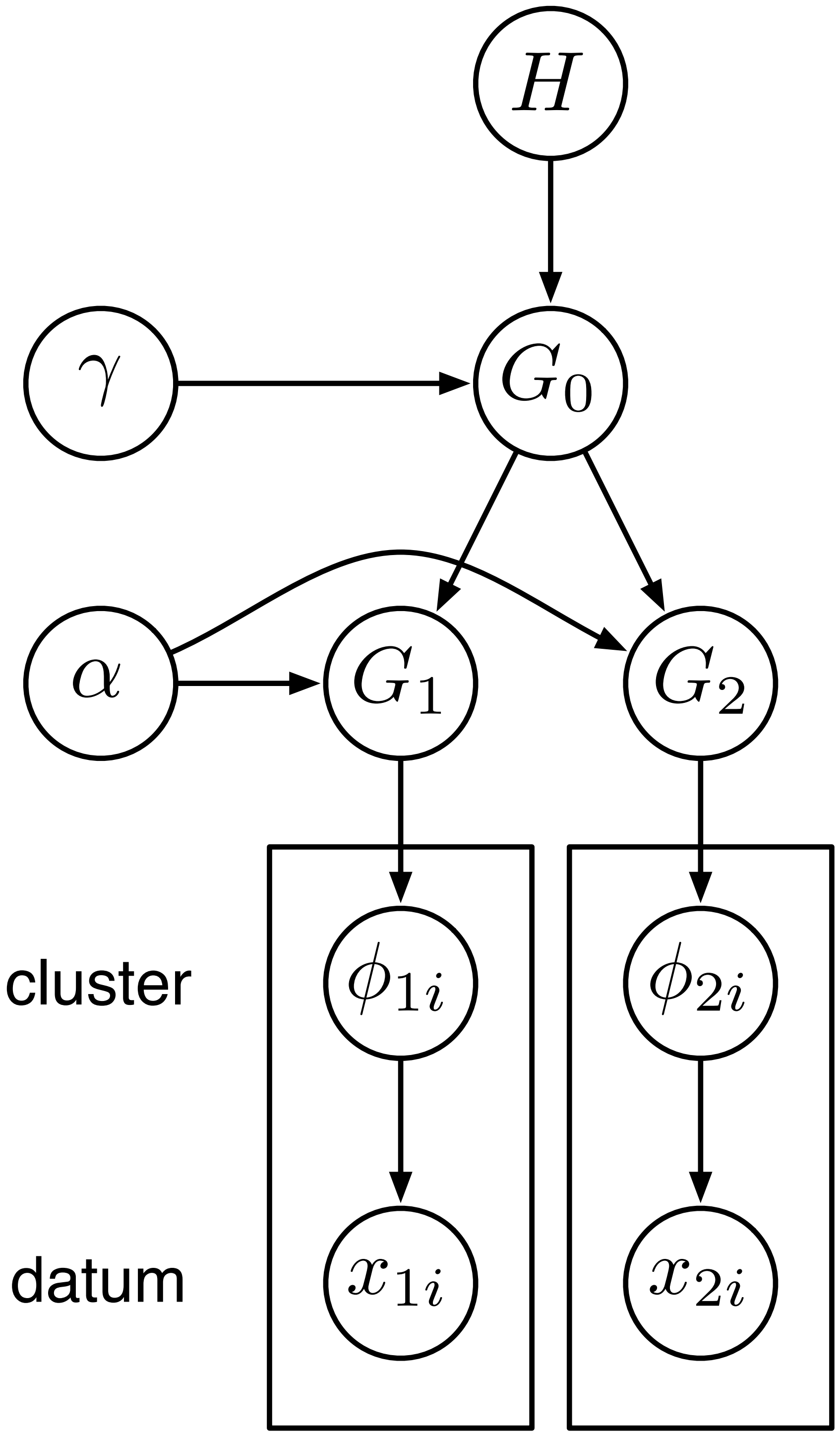
- Hierarchical Dirichlet process:

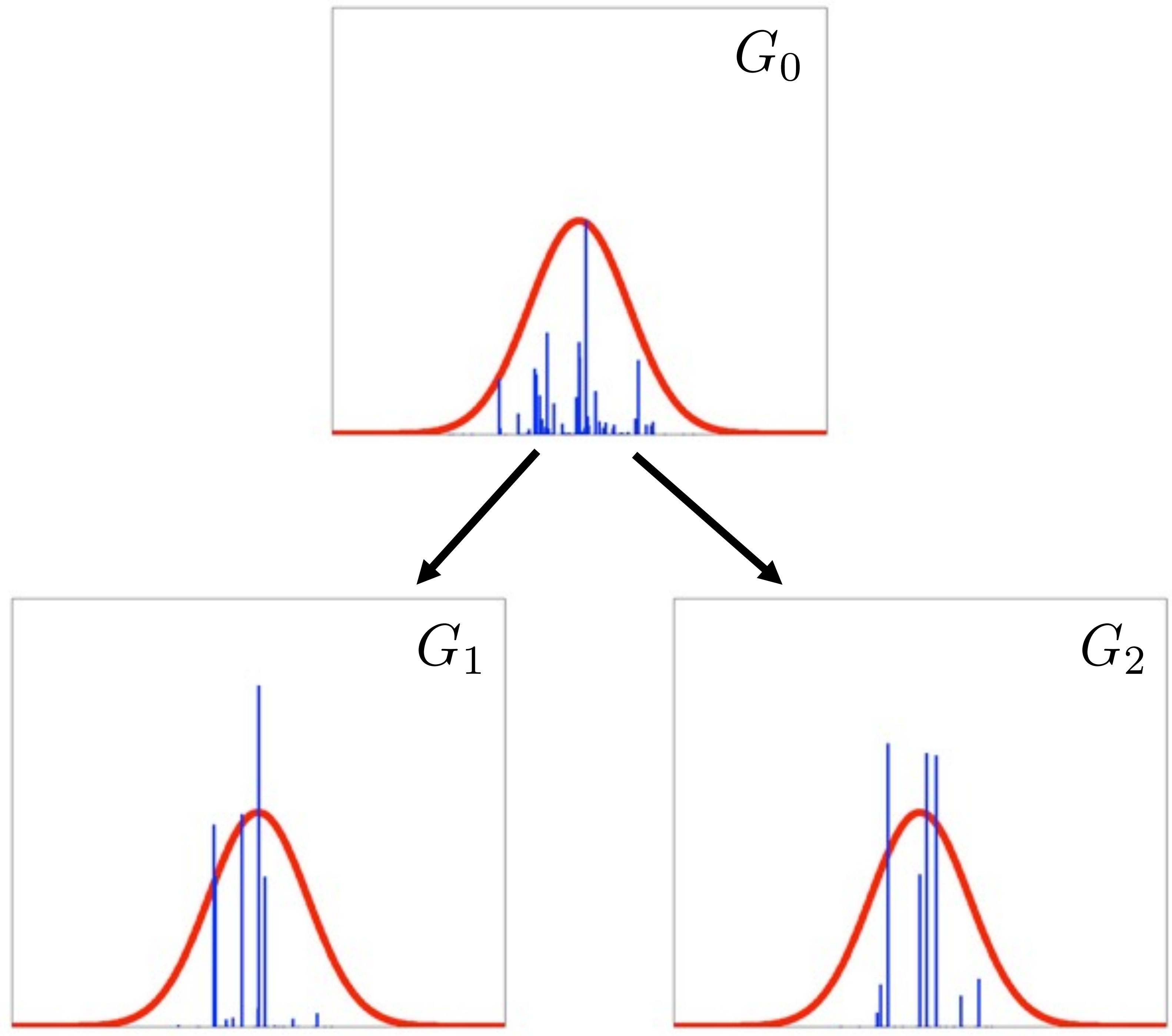
$$\begin{aligned} G_0 &\sim \mathrm{DP}(\gamma, H) \\ G_1 | G_0 &\sim \mathrm{DP}(\alpha, G_0) \\ G_2 | G_0 &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

- Extension to deeper hierarchies is straightforward.

[Teh et al 2006]

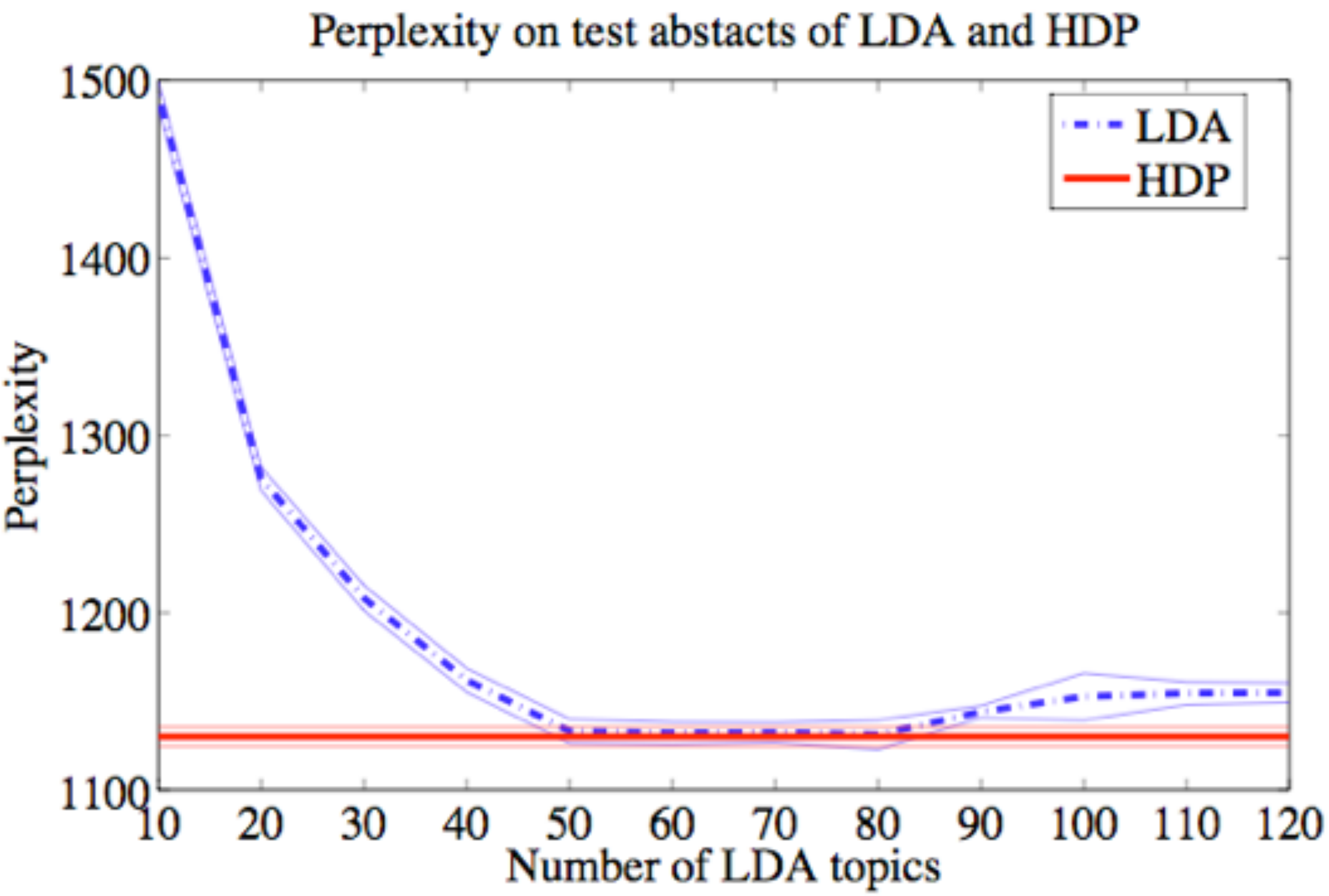
Hierarchical Dirichlet Process Mixture Models

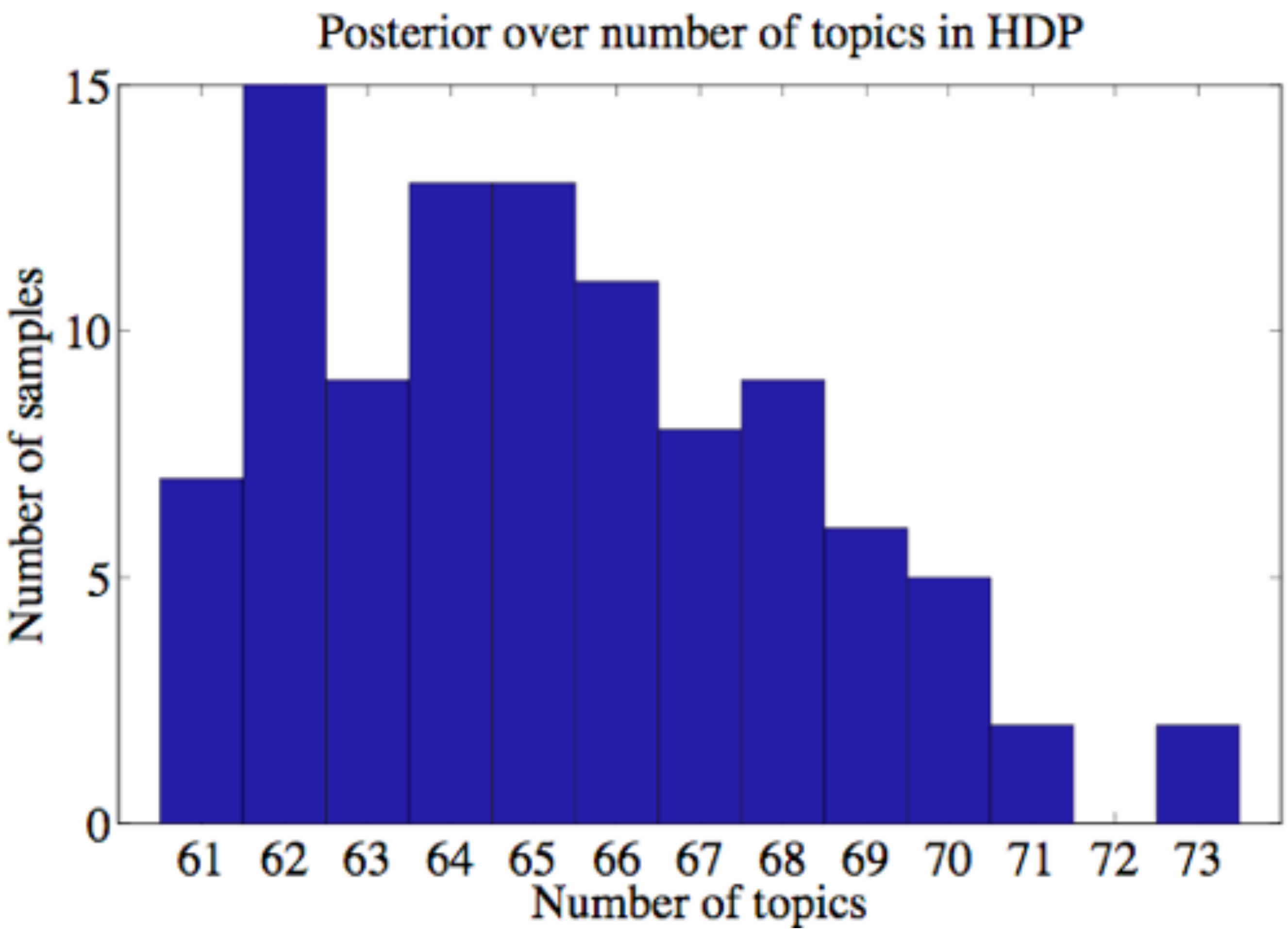




Document Topic Modeling

- Comparison of HDP and latent Dirichlet allocation (LDA).
- LDA is a parametric model, for which model selection is needed.
- HDP bypasses this step in the analysis.



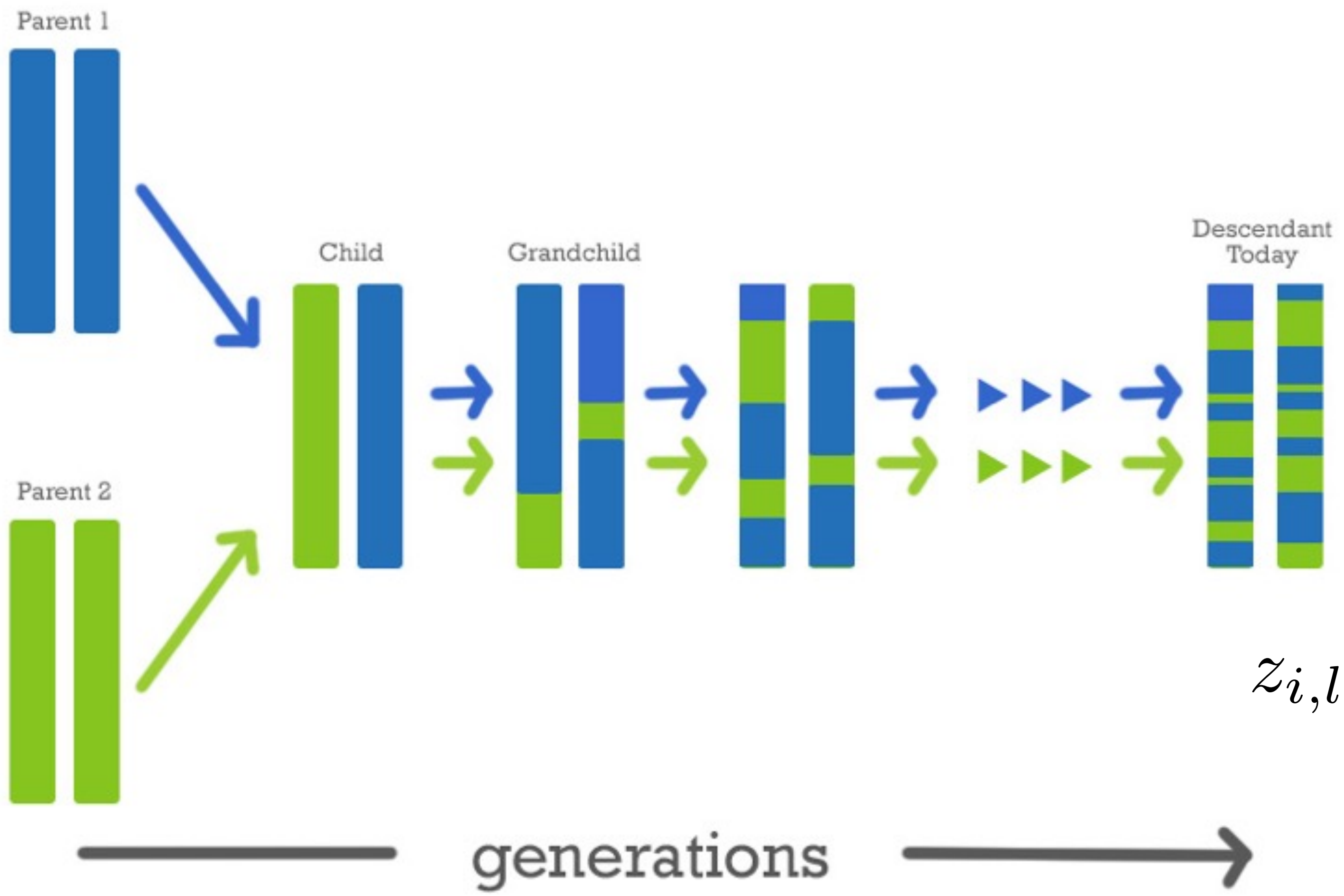


Shared Topics

- Used a 3-level HDP to model shared topics in a collection of machine learning conference papers.
- Shown are the two largest topics shared between Visual Sciences section and four other sections.
- Topics are summarized by the 10 most frequent words in it.

Cognitive Science		**Neuroscience**		**Algorithms & Architecture**		**Signal Processing**	
task	examples	cells	visual	algorithms	distance	visual	signals
representation	concept	cell	cells	test	tangent	images	separation
pattern	similarity	activity	cortical	approach	image	video	signal
processing	Bayesian	response	orientation	methods	images	language	sources
trained	hypotheses	neuron	receptive	based	transformation	image	source
representations	generalization	visual	contrast	point	transformations	pixel	matrix
three	numbers	patterns	spatial	problems	pattern	acoustic	blind
process	positive	pattern	cortex	form	vectors	delta	mixing
unit	classes	single	stimulus	large	convolution	lowpass	gradient
patterns	hypothesis	fig	tuning	paper	simard	flow	eq

Genetic Admixtures



$$G_0 \sim \mathrm{DP}(\gamma, H)$$
$$G_i | G_0 \sim \mathrm{DP}(\alpha, G_0)$$

$$s_{i,l+1} \sim \mathrm{Bernoulli}(e^{-rd_l})$$
$$z_{i,l+1} | s_{i,l+1}, z_{il} \sim \begin{cases} \delta_{z_{il}} & \text{if } s_{i,l+1} = 1, \\ G_i & \text{if } s_{i,l+1} = 0. \end{cases}$$
$$x_{il} | z_{il} = \theta_k \sim \mathrm{Discrete}(\theta_{kl})$$

[de Iorio et al 2015]

Overview

- Bayesian nonparametrics and random probability measures
  - Mixture models and clustering
- Hierarchies of Dirichlet processes
  - Modelling document collections with topic models
  - Modelling genetic admixtures in human populations
- Hierarchies of Pitman-Yor processes
  - Language modelling with high-order Markov models and power law statistics
  - Non-Markov language models with the sequence memoizer

Sequence Models for Language and Text

- Probabilistic models for sequences of words and characters, e.g.

south, parks, road

s, o, u, t, h, _, p, a, r, k, s, _, r, o, a, d

- Uses:
  - Natural language processing: speech recognition, OCR, machine translation.
  - Compression.
  - Cognitive models of language acquisition.
  - Sequence data arises in many other domains.

Markov Models for Language and Text

- Probabilistic models for sequences of words and characters.

P(south parks road) =
P(south)*
P(parks | south)*
P(road | south parks)

Andrey Markov

- Usually makes a Markov assumption:

P(south parks road) ~
P(south)*
P(parks | south)*
P(road | parks)

George E. P. Box

- Order of Markov model typically ranges from ~3 to > 10.

High Dimensional Estimation

- Consider a high order Markov models:

$$P(\text{sentence}) = \prod_i P(\text{word}_i | \text{word}_{i-N+1} \dots \text{word}_{i-1})$$

- Large vocabulary size means naïvely estimating parameters of this model from data counts is problematic for N>2.

$$P^{\text{ML}}(\text{word}_i | \text{word}_{i-N+1} \dots \text{word}_{i-1}) = \frac{C(\text{word}_{i-N+1} \dots \text{word}_i)}{C(\text{word}_{i-N+1} \dots \text{word}_{i-1})}$$

- Naïve regularization fail as well: most parameters have *no* associated data.
  - Smoothing.
  - Hierarchical Bayesian models.

Smoothing in Language Models

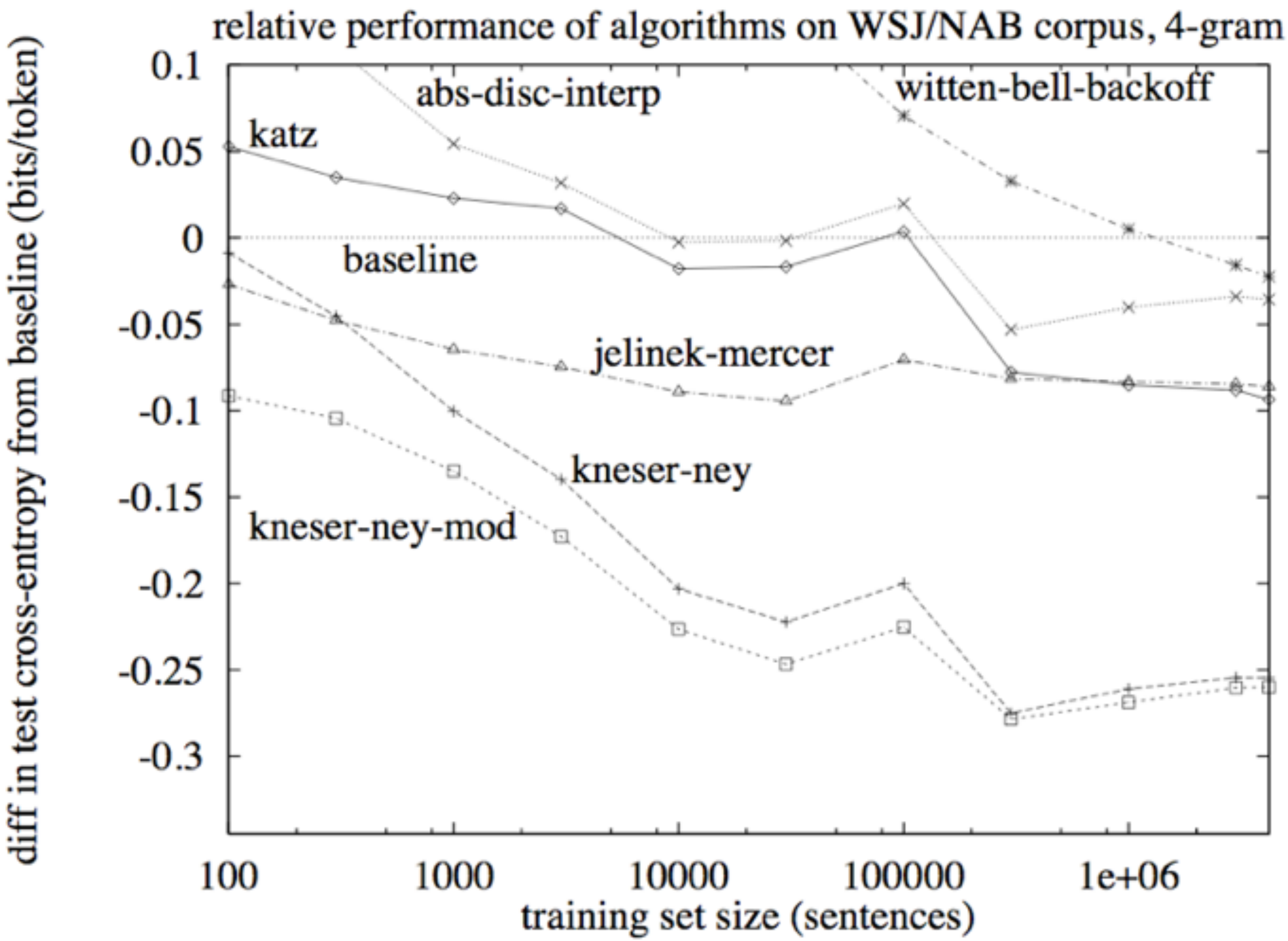
- Smoothing is a way of dealing with data sparsity by combining large and small models together.

$$P^{\text{smooth}}(\text{word}_i|\text{word}_{i-N+1}^{i-1}) = \sum_{n=1}^{N} \lambda(n) Q_n(\text{word}_i|\text{word}_{i-n+1}^{i-1})$$

$$\begin{aligned} & P^{\text{smooth}}(\text{road}|\text{south parks}) \\ = \quad & \lambda(3)Q_3(\text{road}|\text{south parks}) + \\ & \lambda(2)Q_2(\text{road}|\text{parks}) + \\ & \lambda(1)Q_1(\text{road}|\emptyset) \end{aligned}$$

- Combines expressive power of large models with better estimation of small models (cf bias-variance trade-off and hierarchical modelling).

Smoothing in Language Models

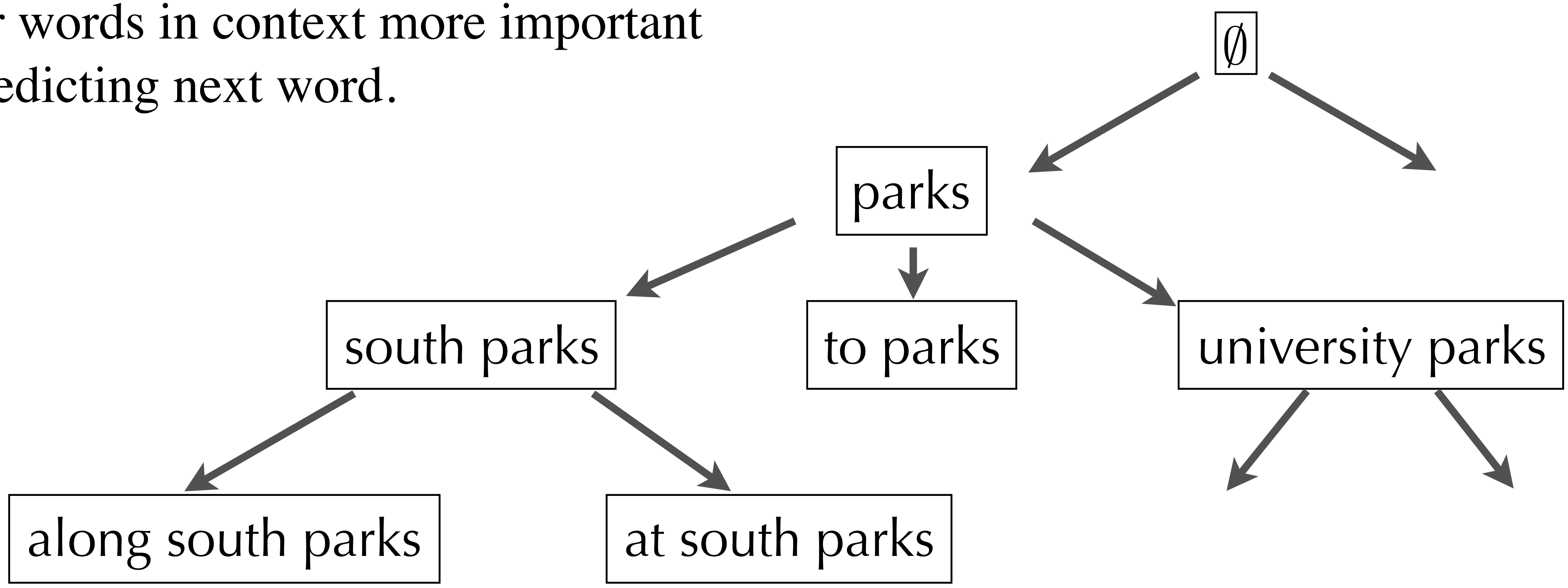


[Chen and Goodman 1998]

Context Tree

- *Context* of conditional probabilities naturally organized using a tree.
- Smoothing makes conditional probabilities of neighbouring contexts more similar.
- Later words in context more important in predicting next word.

$$
\begin{aligned}
&P^{\text{smooth}}(\text{road}|\text{south parks}) \\
=&\lambda(3)Q_3(\text{road}|\text{south parks})+ \\
&\lambda(2)Q_2(\text{road}|\text{parks})+ \\
&\lambda(1)Q_1(\text{road}|\emptyset)
\end{aligned}
$$



Hierarchical Bayesian Models on Context Tree

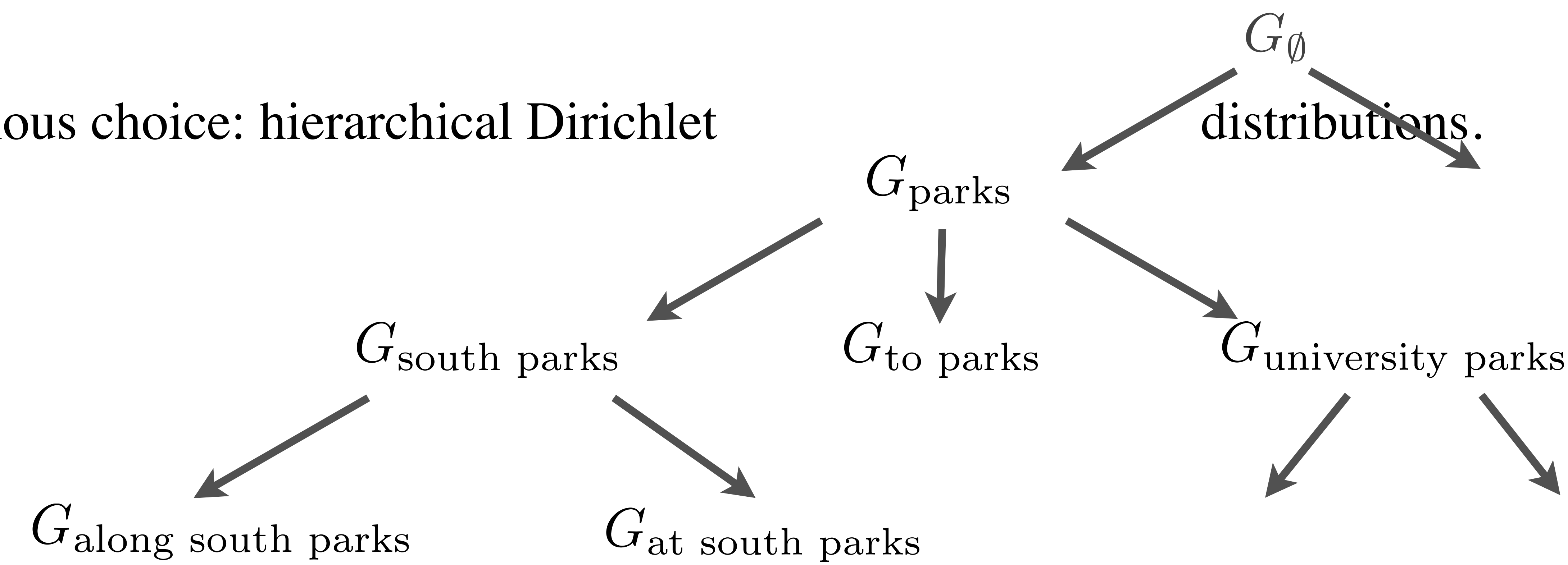
- Parametrize the conditional probabilities of Markov model:

$$P(\text{word}_i = w | \text{word}_{i-N+1}^{i-1} = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in \text{vocabulary}}$$

- G_u is a probability vector associated with context u.

- Obvious choice: hierarchical Dirichlet distributions.



[MacKay and Peto 1994]

Hierarchical Dirichlet Language Models

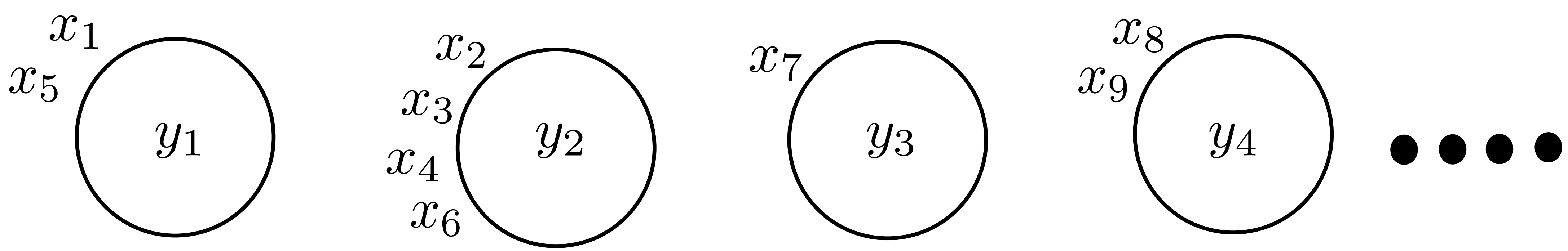
- What is $P(G_u|G_{\mathrm{pa}(u)})$? [MacKay and Peto 1994] proposed using the standard Dirichlet distribution over probability vectors.

T	N-1	IKN	MKN	HDLM
2×10^6	2	148.8	144.1	191.2
4×10^6	2	137.1	132.7	172.7
6×10^6	2	130.6	126.7	162.3
8×10^6	2	125.9	122.3	154.7
10×10^6	2	122.0	118.6	148.7
12×10^6	2	119.0	115.8	144.0
14×10^6	2	116.7	113.6	140.5
14×10^6	1	169.9	169.2	180.6
14×10^6	3	106.1	102.4	136.6

- We will use Pitman-Yor processes instead [Pitman and Yor 1997], [Ishwaran and James 2001].

Exchangeable Random Partition

- Easiest to understand them using Chinese restaurant processes.



$$p(\text{sit at table } k) = \frac{c_k - d}{\theta + \sum_{j=1}^{K} c_j} \qquad p(\text{table serves dish } y) = H(y)$$

$$p(\text{sit at new table}) = \frac{\theta + dK}{\theta + \sum_{j=1}^{K} c_j} \qquad i \text{ sits at table } c\text{:} \quad x_i = y_c$$

- Defines an exchangeable stochastic process over sequences $x_1, x_2, \ldots$
- The de Finetti measure [Kingman 1978] is the Pitman-Yor process,

$$\begin{aligned} G &\sim \mathrm{PY}(\theta, d, H) \\ x_i &\sim G \qquad i = 1, 2, \ldots \end{aligned}$$

- [Pitman & Yor 1997]

Power Law Properties of Pitman-Yor Processes

- Chinese restaurant process:

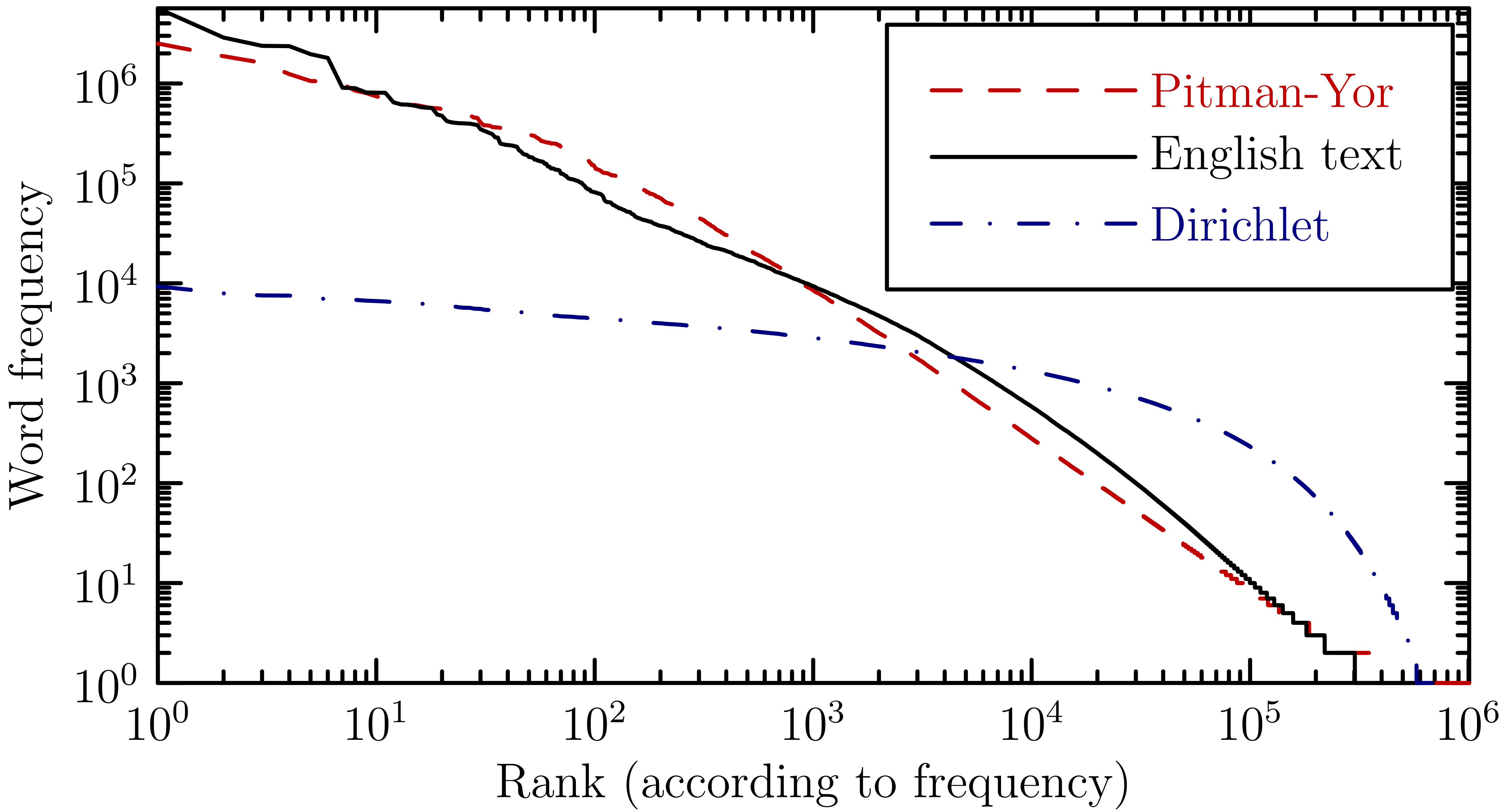
$$
\begin{aligned}
p(\text{sit at table } k) &\propto c_k - d \\
p(\text{sit at new table}) &\propto \theta + dK
\end{aligned}
$$

- Pitman-Yor processes produce distributions over words given by a power-law distribution with index $1 + d$.
  - Customers = word instances, tables = dictionary look-up;
  - Small number of common word types;
  - Large number of rare word types.

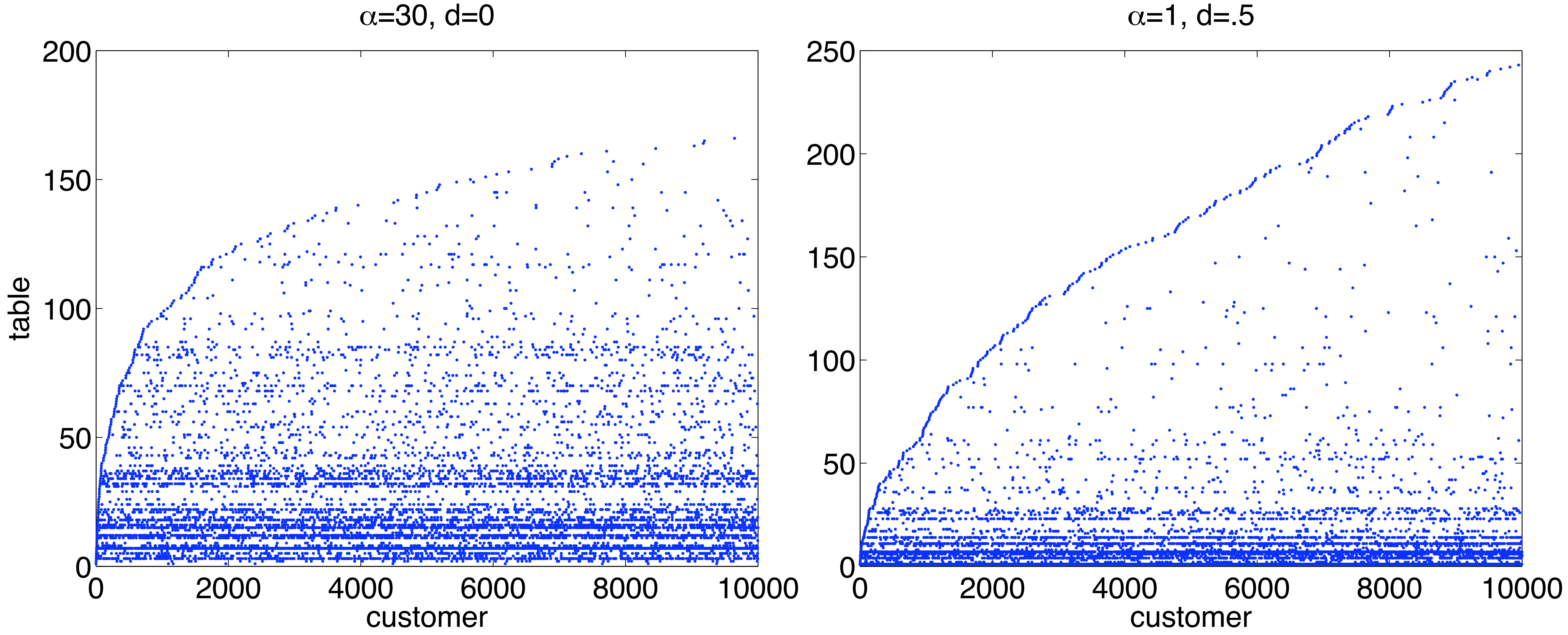
- This is more suitable for languages than Dirichlet distributions.

- [Goldwater, Griffiths and Johnson 2005] investigated the Pitman-Yor process from this perspective.

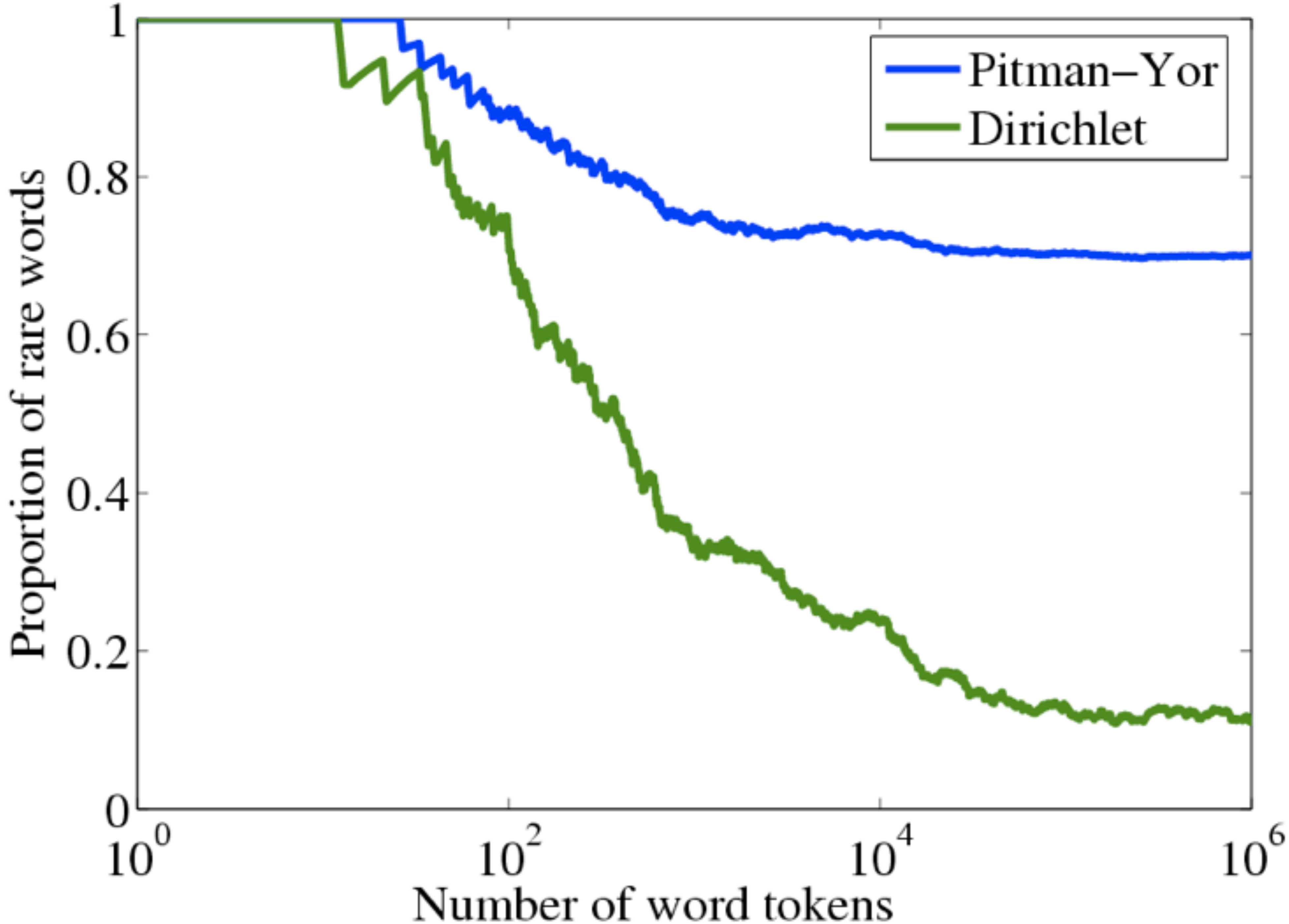
Pitman-Yor Processes



Power Law Properties of Pitman-Yor Processes



Power Law Properties of Pitman-Yor Processes



Hierarchical Pitman-Yor Language Models

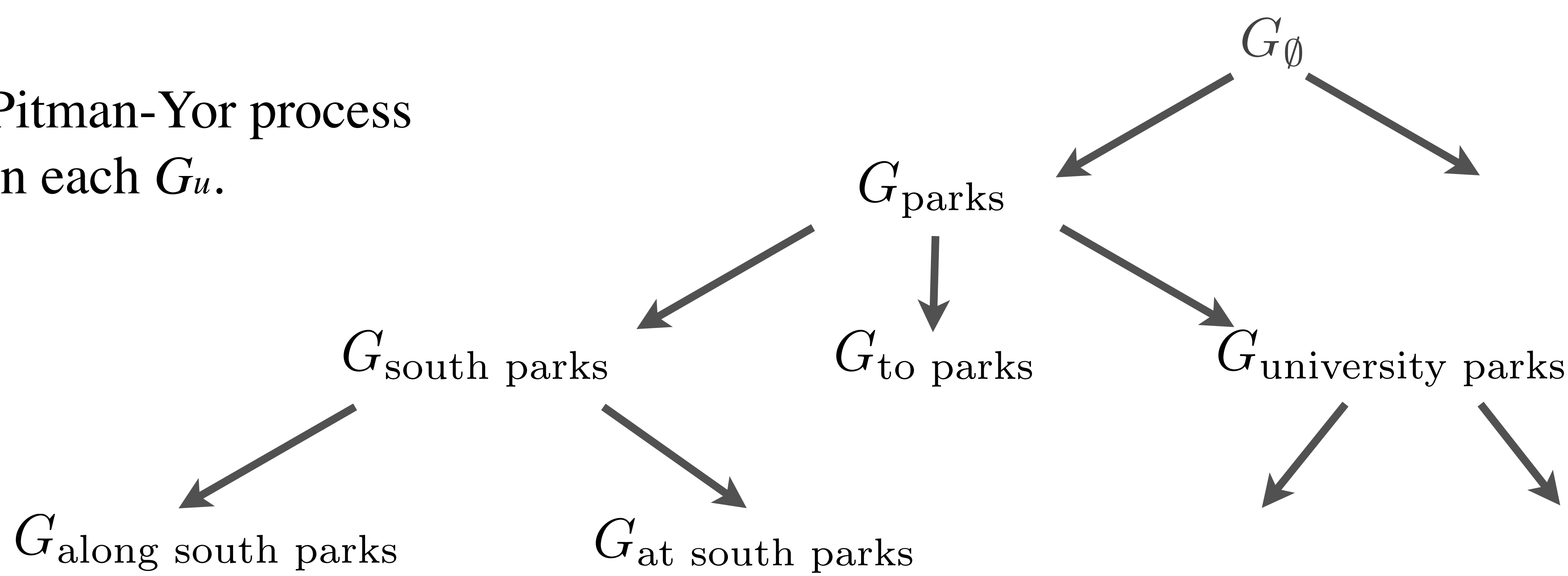
- Parametrize the conditional probabilities of Markov model:

$$P(\text{word}_i = w | \text{word}_{i-N+1}^{i-1} = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in \text{vocabulary}}$$

- G_u is a probability vector associated with context u.

- Place Pitman-Yor process prior on each G_u.



Hierarchical Pitman-Yor Language Models

- Significantly improved on the hierarchical Dirichlet language model.
- Results better Kneser-Ney smoothing, state-of-the-art language models.

T	N-1	IKN	MKN	HDLM	HPYLM
2×10^6	2	148.8	**144.1**	191.2	144.3
4×10^6	2	137.1	**132.7**	172.7	**132.7**
6×10^6	2	130.6	126.7	162.3	**126.4**
8×10^6	2	125.9	122.3	154.7	**121.9**
10×10^6	2	122.0	118.6	148.7	**118.2**
12×10^6	2	119.0	115.8	144.0	**115.4**
14×10^6	2	116.7	113.6	140.5	**113.2**
14×10^6	1	169.9	**169.2**	180.6	169.3
14×10^6	3	106.1	102.4	136.6	**101.9**

- Similarity of perplexities not a surprise---Kneser-Ney can be derived as a particular approximate inference method.

[Teh 2006]

Markov Models for Language and Text

- Usually makes a Markov assumption to simplify model:

$$P(\text{south parks road}) \sim P(\text{south}) * P(\text{parks} \mid \text{south}) * P(\text{road} \mid \text{south parks})$$

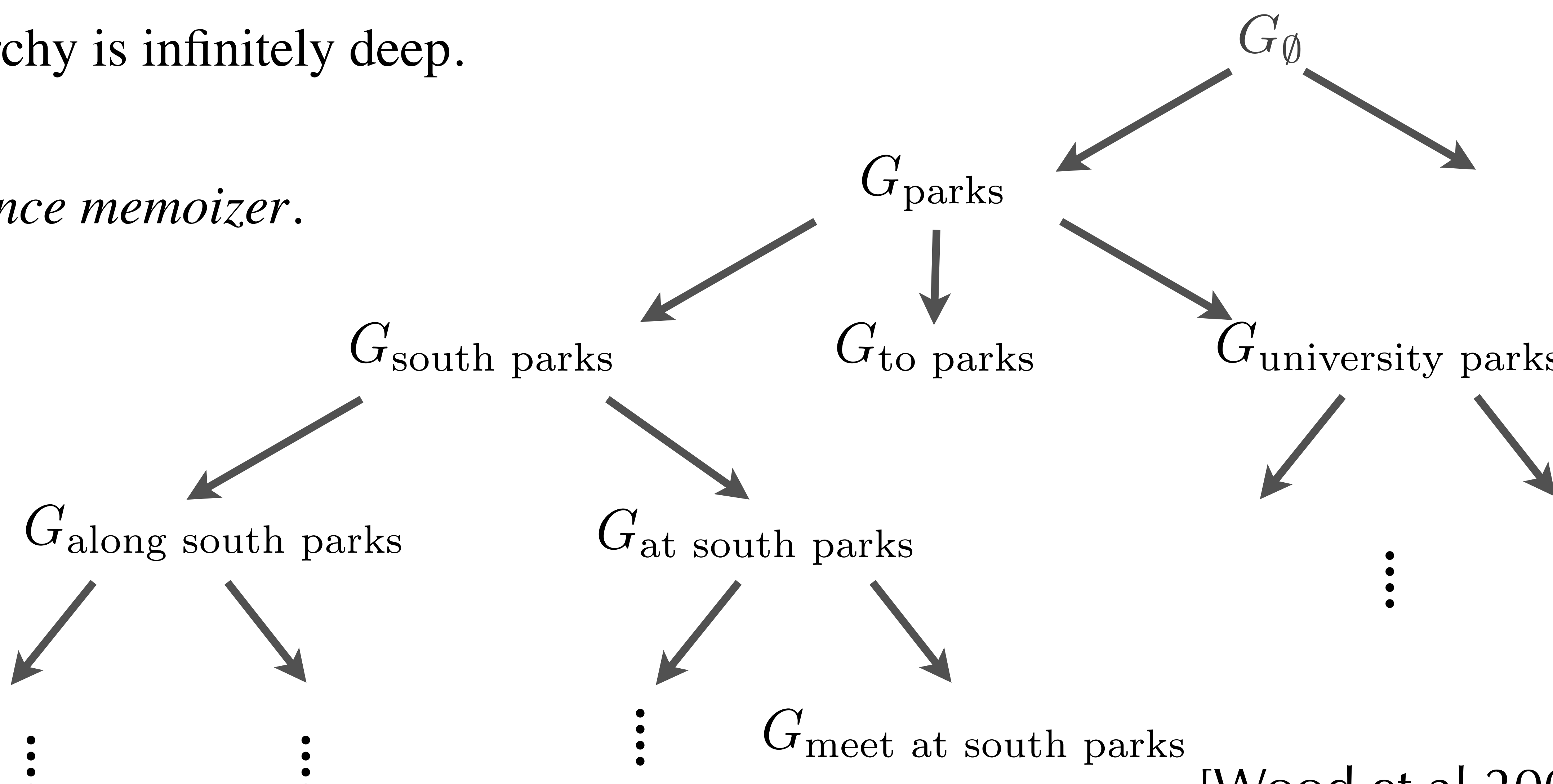
- Language models: usually Markov models of order 2-4 (3-5-grams).
- How do we determine the order of our Markov models?
- Is the Markov assumption a reasonable assumption?
  - Be nonparametric about Markov order...

Non-Markov Models for Language and Text

- Model the conditional probabilities of each possible word occurring after each possible context (of unbounded length).

- Use hierarchical Pitman-Yor process prior to share information across all contexts.
- Hierarchy is infinitely deep.

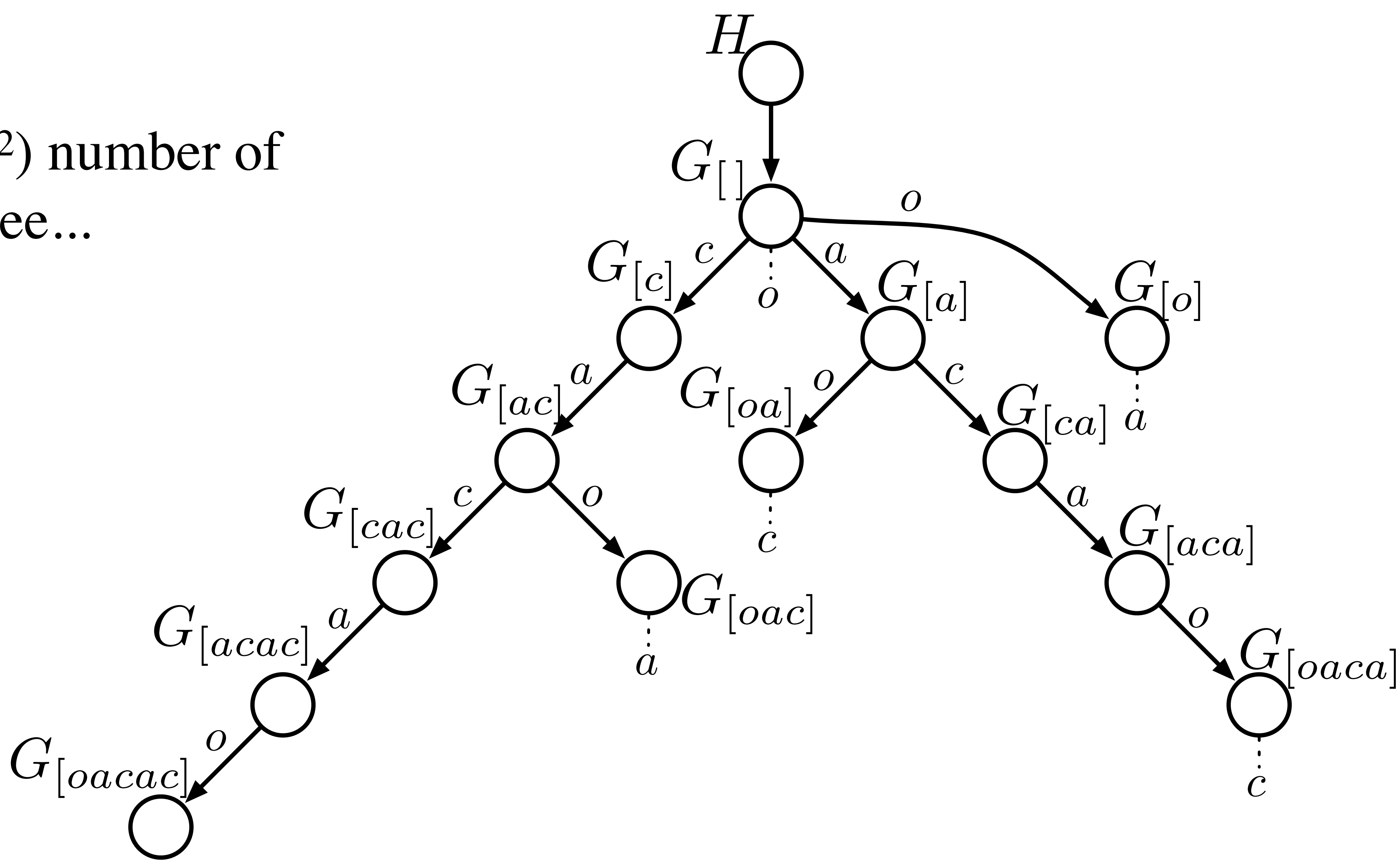
- *Sequence memoizer*.



[Wood et al 2009]

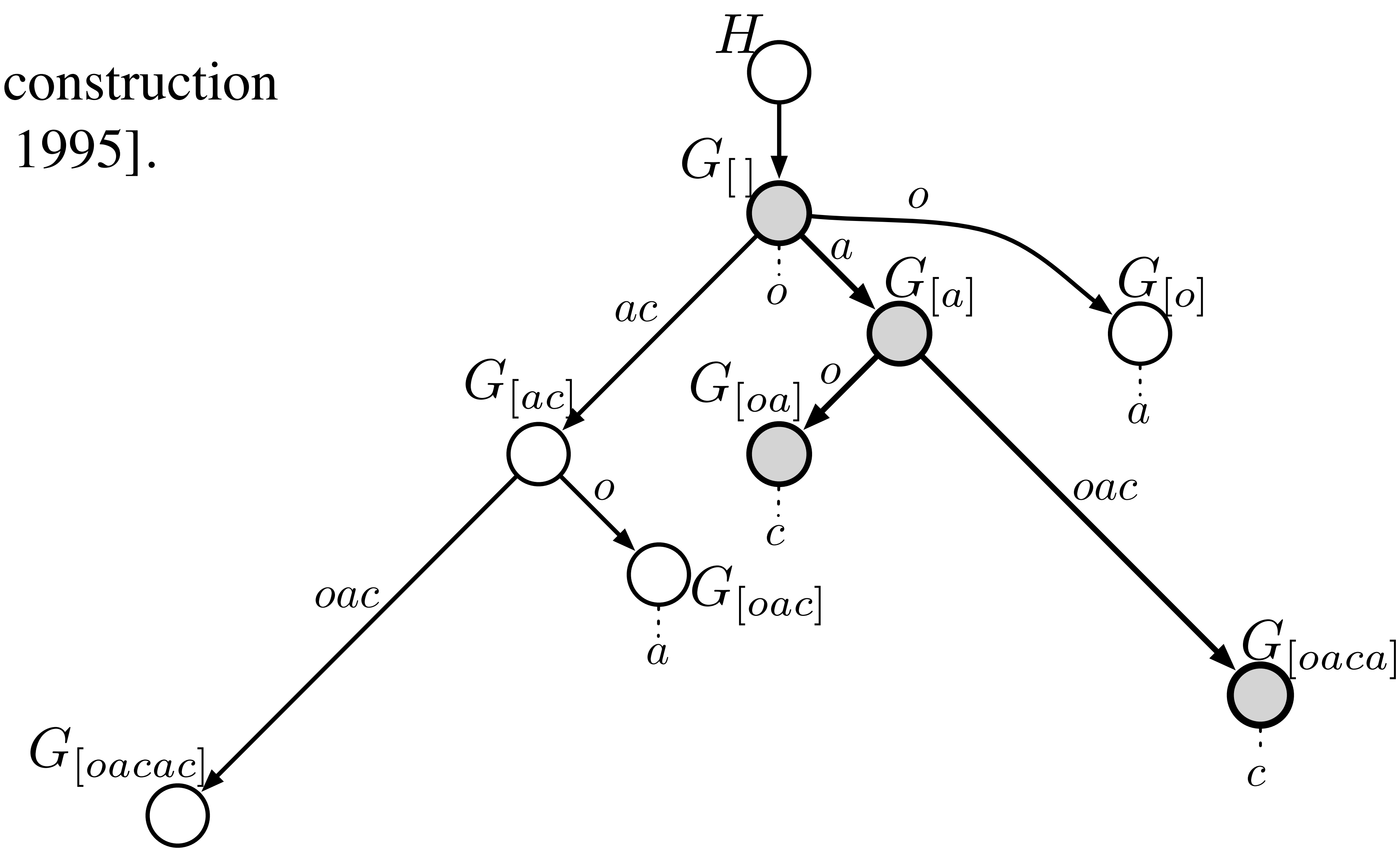
Model Size: Infinite -> $O(T^2)$

- The sequence memoizer model is very large (actually, infinite).
- Given a training sequence (e.g.: o,a,c,a,c), most of the model can be ignored (integrated out), leaving a finite number of nodes in context tree.
- But there are still $O(T^2)$ number of nodes in the context tree...



Model Size: Infinite -> O(T²) -> 2T

- Idea: integrate out non-branching, non-leaf nodes of the context tree.
- Resulting tree is related to a suffix tree data structure, and has at most *2T* nodes.
- There are linear time construction algorithms [Ukkonen 1995].



Closure under Marginalization

- In marginalizing out non-branching interior nodes, need to ensure that resulting conditional distributions are still tractable.

$$
\begin{tikzcd}
G_{[a]} \arrow[d, "\mathrm{PY}(\theta_2, d_2, G_{[a]})"] & & G_{[a]} \arrow[dd, "?"] \\
G_{[ca]} \arrow[d, "\mathrm{PY}(\theta_3, d_3, G_{[ca]})"] & & \\
G_{[aca]} & & G_{[aca]}
\end{tikzcd}
$$

- E.g.: If each conditional is Dirichlet, resulting conditional is not of known analytic form.

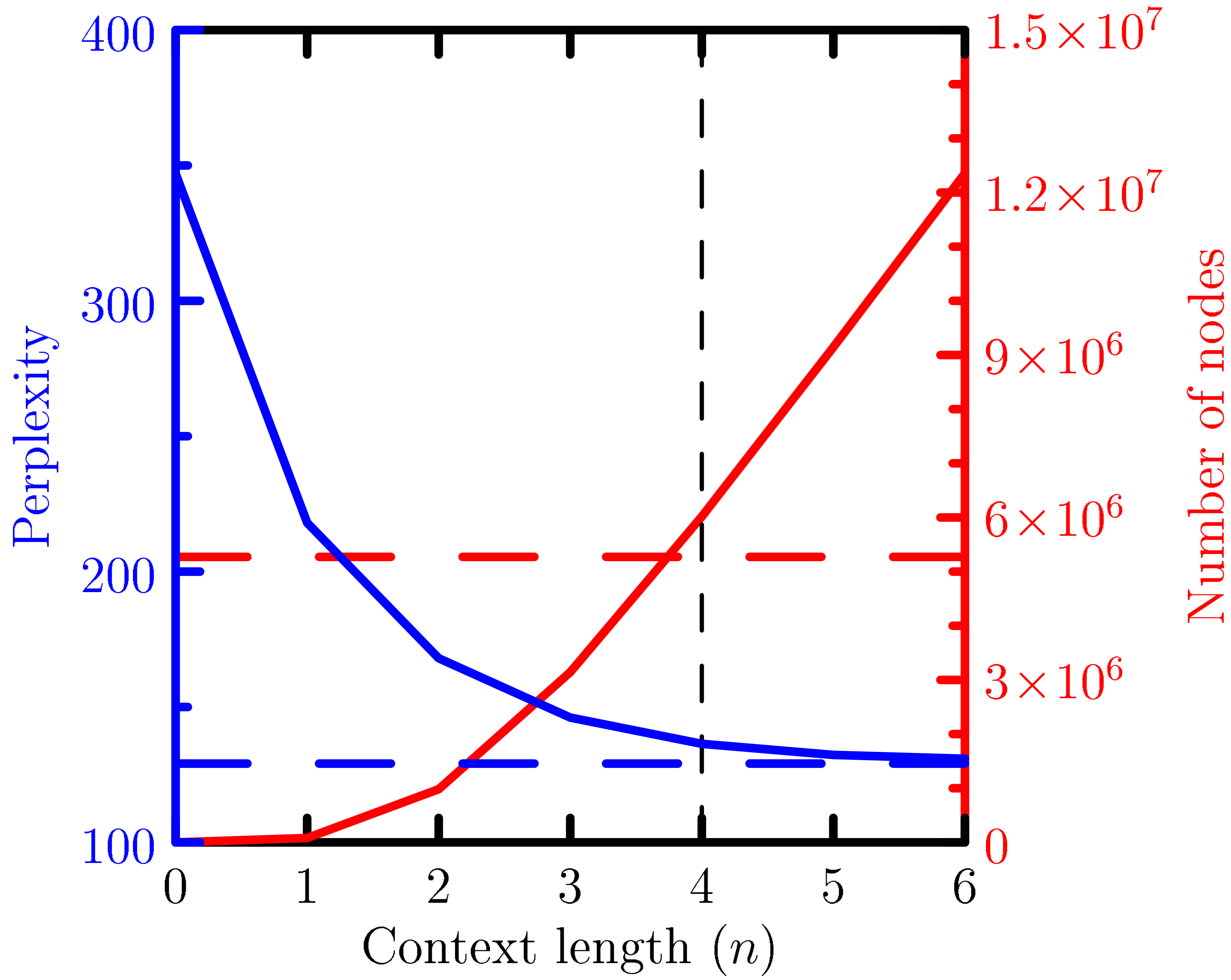
Closure under Marginalization

- In marginalizing out non-branching interior nodes, need to ensure that resulting conditional distributions are still tractable.

$$
\begin{tikzcd}
G_{[a]} \arrow[d, "\mathrm{PY}(\theta_2, d_2, G_{[a]})"] \\
G_{[ca]} \arrow[d, "\mathrm{PY}(\theta_2 d_3, d_3, G_{[ca]})"] \\
G_{[aca]}
\end{tikzcd}
\qquad
\begin{tikzcd}
G_{[a]} \arrow[dd, "\mathrm{PY}(\theta_2 d_3, d_2 d_3, G_{[a]})"] \\
\\
G_{[aca]}
\end{tikzcd}
$$

- Hierarchical construction is equivalent to coagulation, so the marginal process is Pitman-Yor distributed as well.

Comparison to Finite Order HPYLM



Compression Results

Model	Average bits/byte
gzip	2.61
bzip2	2.11
CTW	1.99
PPM	1.93
Sequence Memoizer	1.89

Calgary corpus
SM inference: particle filter
PPM: Prediction by Partial Matching
CTW: Context Tree Weigting
Online inference, entropic coding.

Summary

- Random probability measures are building blocks of many Bayesian nonparametric models.

- Motivated by problems in text and language processing, we discussed methods of constructing hierarchies of random measures.

- We used Pitman-Yor processes to capture the power law behaviour of language data.

- We used the equivalence between hierarchies and coagulations, and a duality between fragmentations and coagulations, to construct an efficient non-Markov language model.

UNIVERSITY OF OXFORD

Thank You!

Acknowledgements:
Matt Beal, Dave Blei, Mike Jordan,
Frank Wood, Jan Gasthaus,
Cedric Archambeau, Lancelot James.

Lee Kuan Yew Foundation
Gatsby Charitable Foundation
European Research Council
EPSRC

GATSBY

erc
European Research Council
Established by the European Commission

EPSRC
Engineering and Physical Sciences
Research Council