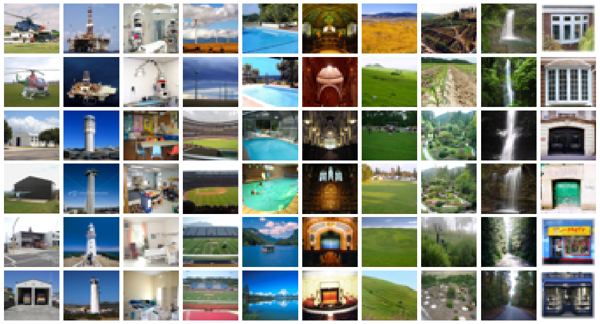
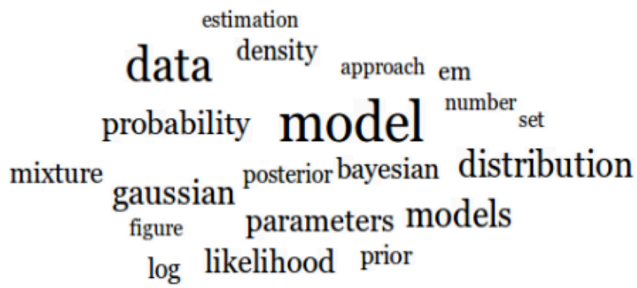
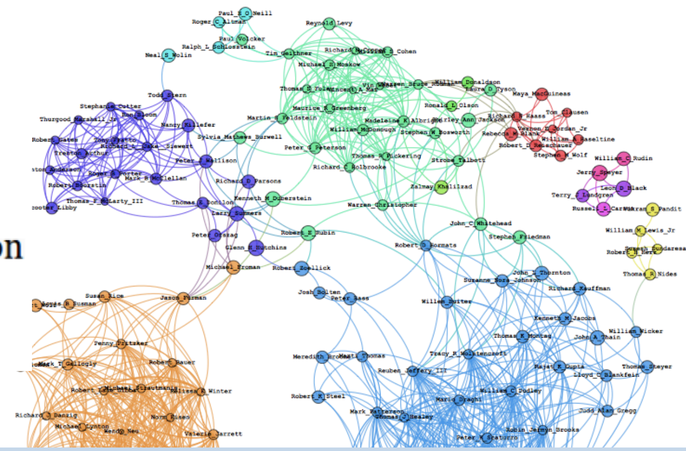






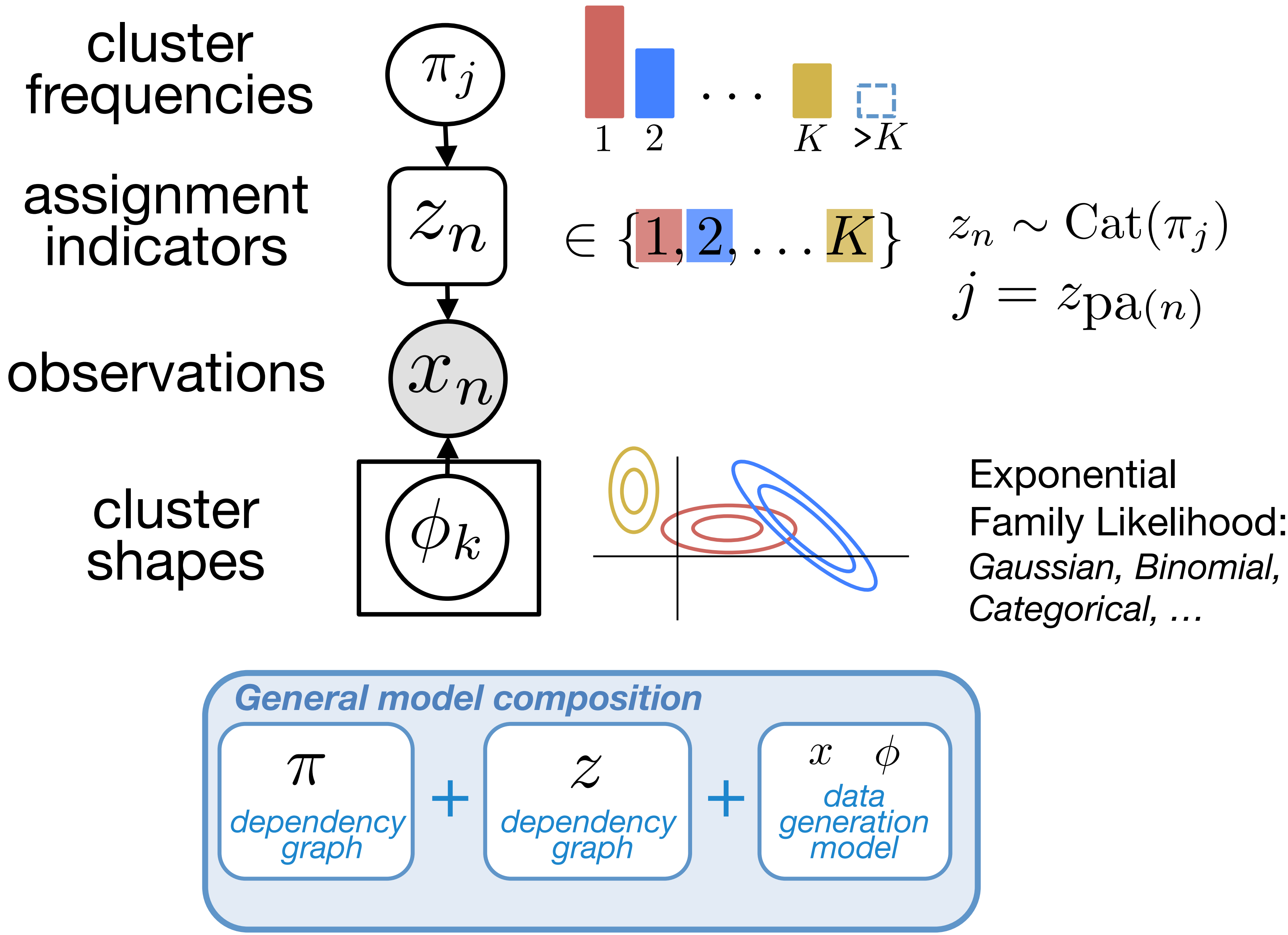
Reliable Variational Learning for Hierarchical Dirichlet Processes

Erik Sudderth

Brown University Computer Science

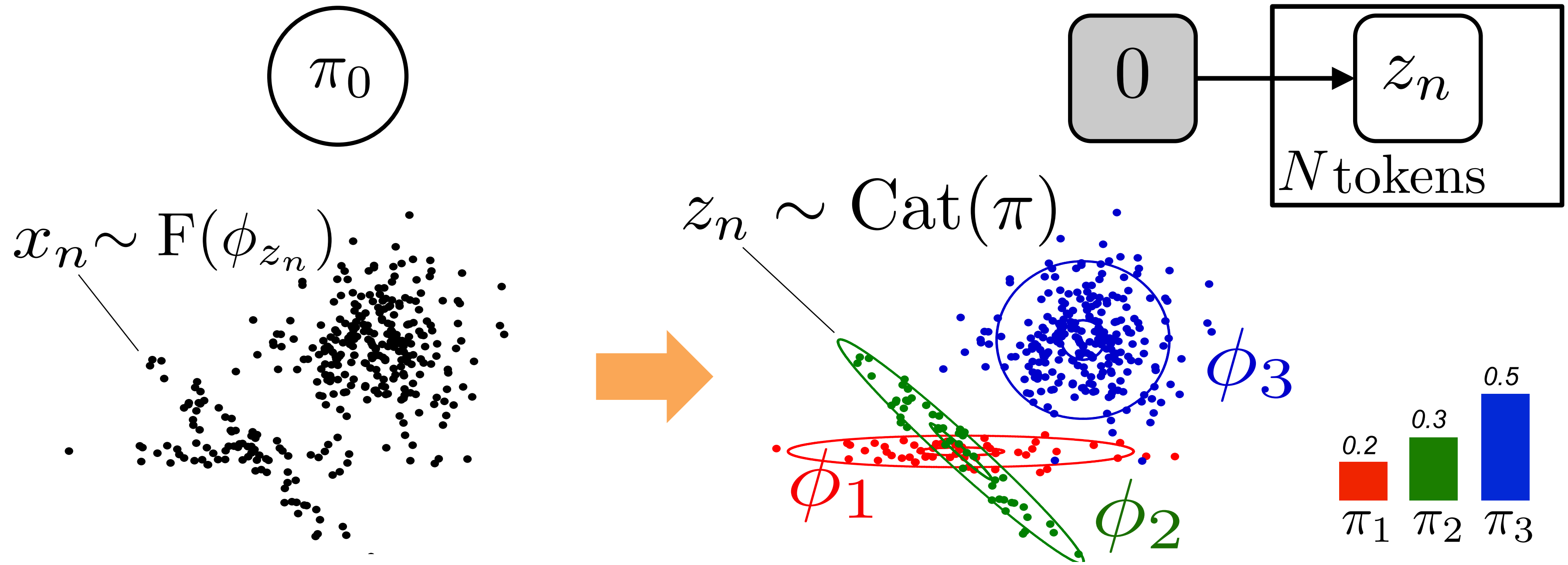
Joint work with
Michael Hughes & Dae Il Kim

Bayesian Nonparametric Clustering



BNP Mixture Models

Cluster Frequency Graph *Cluster Assignment Graph*



Stick-breaking prior on cluster frequencies:

$\pi_1 = v_1$

0 1

$\pi_2 = v_2(1 - v_1)$

$\pi_3 = v_3(1 - v_2)(1 - v_1)$

$$\pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell) \quad \vdots$$

Dirichlet Process:

$v_k \sim \mathrm{Beta}(1, \alpha)$

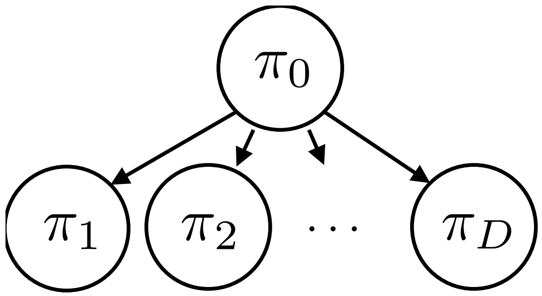
Pitman-Yor Process:

$v_k \sim \mathrm{Beta}(1 - \sigma, \alpha + k\sigma)$

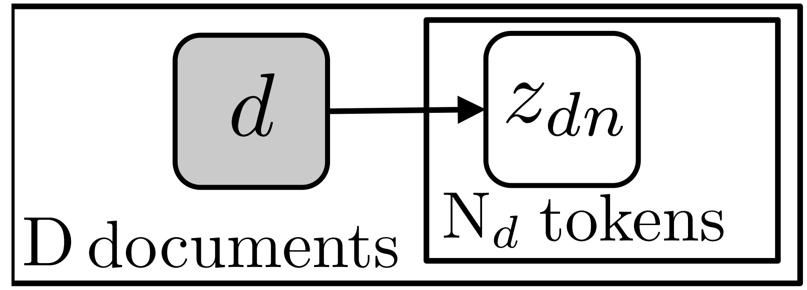
Also finite Dirichlet, …

BNP Admixture (Topic) Models

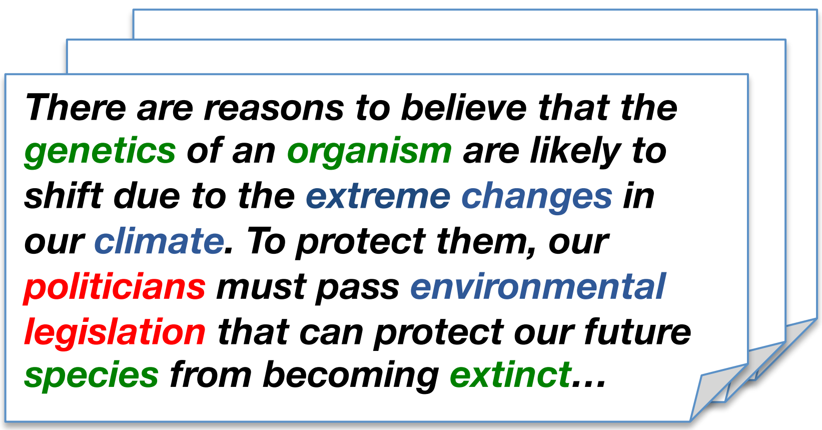
Cluster Frequency Graph

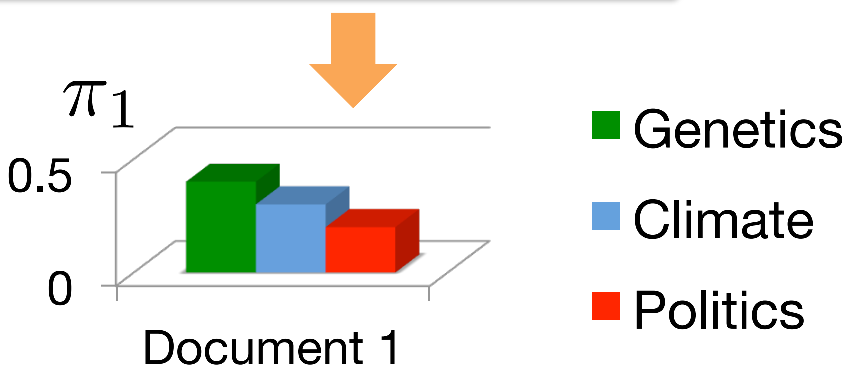


Cluster Assignment Graph



$$z_{dn} \sim \mathrm{Cat}(\pi_d)$$





Hierarchical DP *(Teh et al., 2006) prior on group-specific cluster frequencies, or doc-specific topic frequencies:*

$$\pi_0 \sim \mathrm{Stick}(\gamma)$$

$$\pi_d \sim \mathrm{DP}(\alpha\pi_0)$$

- *Mean cluster frequencies:*

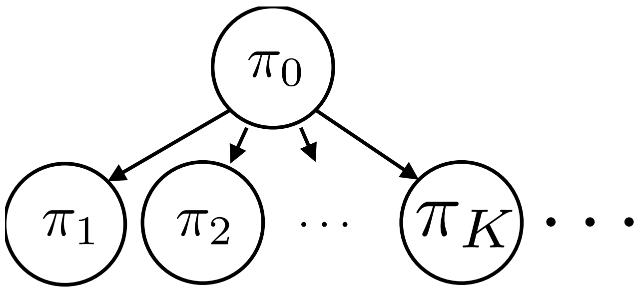
$$\mathbb{E}[\pi_d] = \pi_0$$

- *Sparse topic usage for*

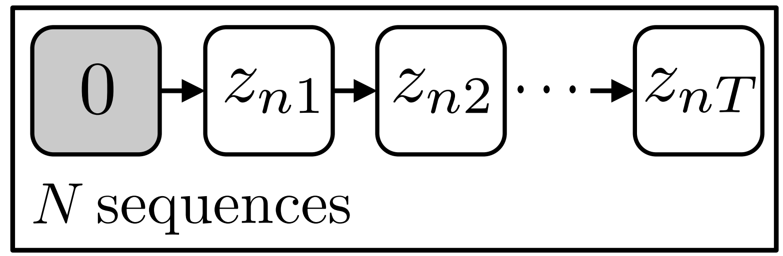
$$\alpha < 1$$

BNP Hidden Markov Models

Cluster Frequency Graph

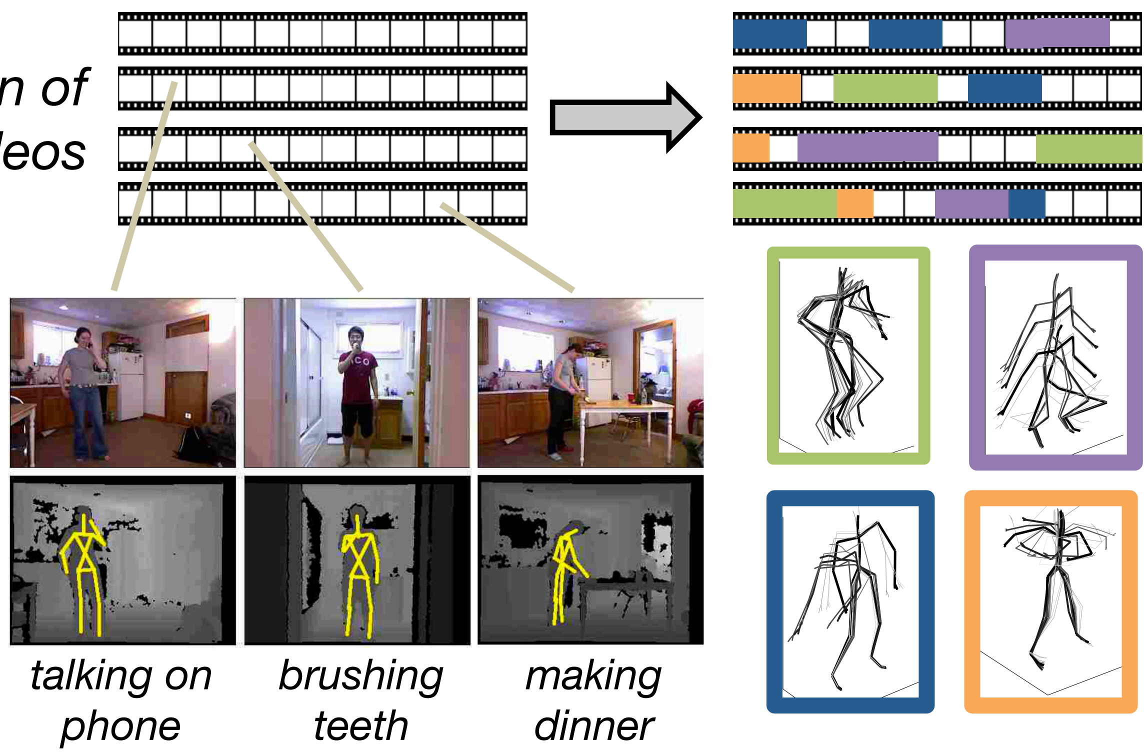


Cluster Assignment Graph



$$z_{nt} \sim \mathrm{Cat}(\pi_{z_{n,t-1}})$$

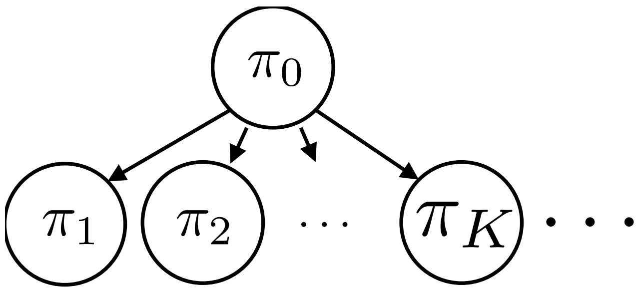
Collection of RGB-D videos



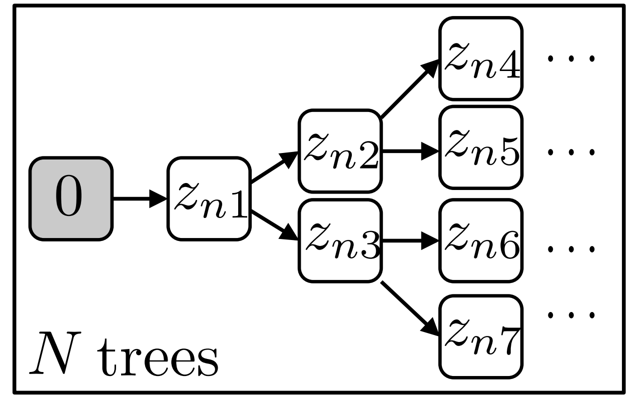
Segmentation of common actions

BNP Hidden Markov Trees

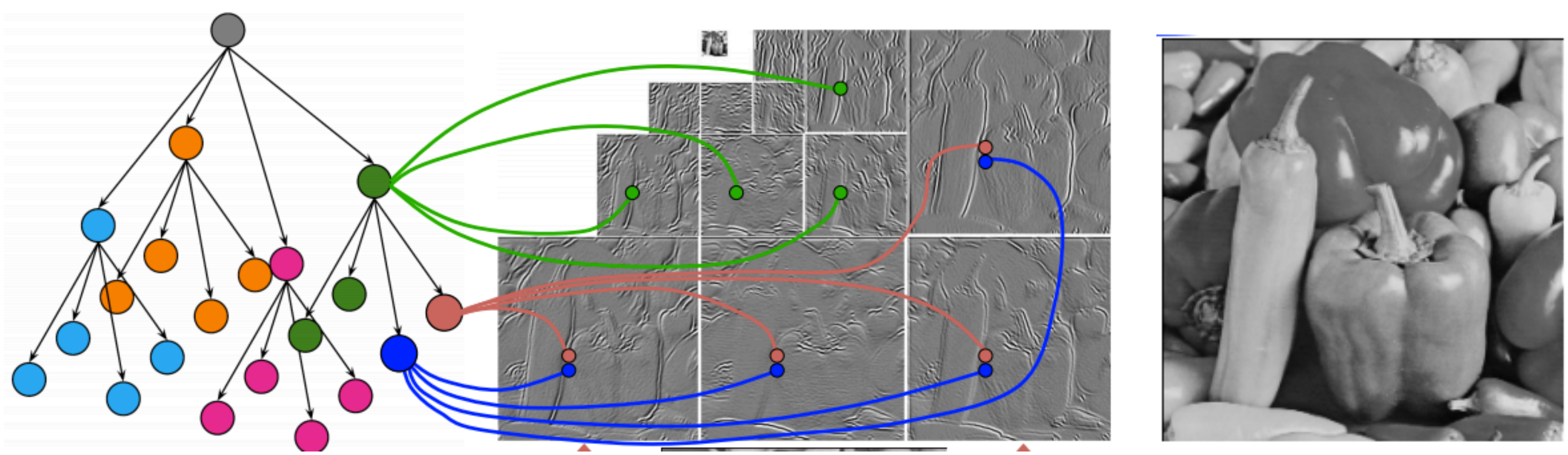
Cluster Frequency Graph



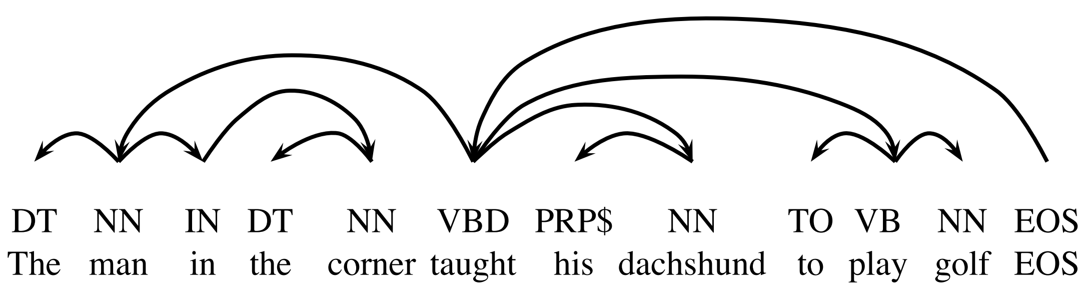
Cluster Assignment Graph



Natural Image Statistics
(Kivinen et al., ICCV 2007)

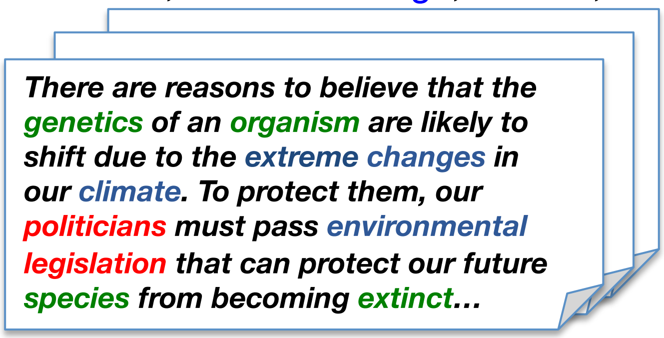


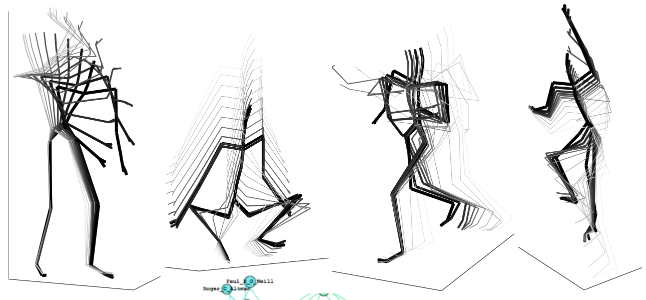
Natural Language Dependence
(Finkel et al., ACL 2007)

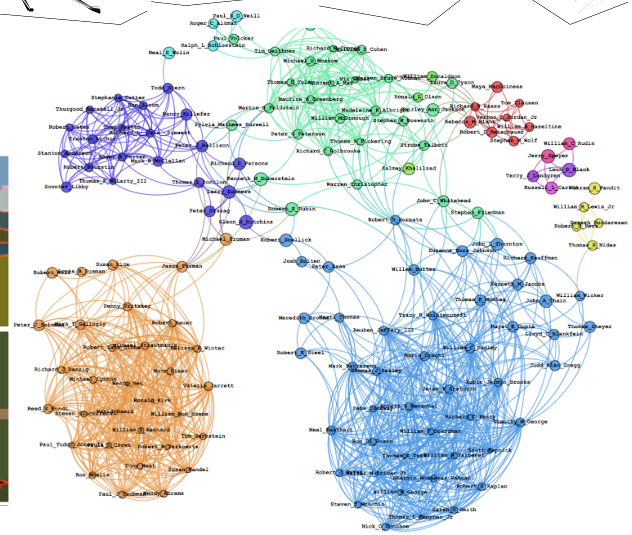


Learning Structured BNP Models

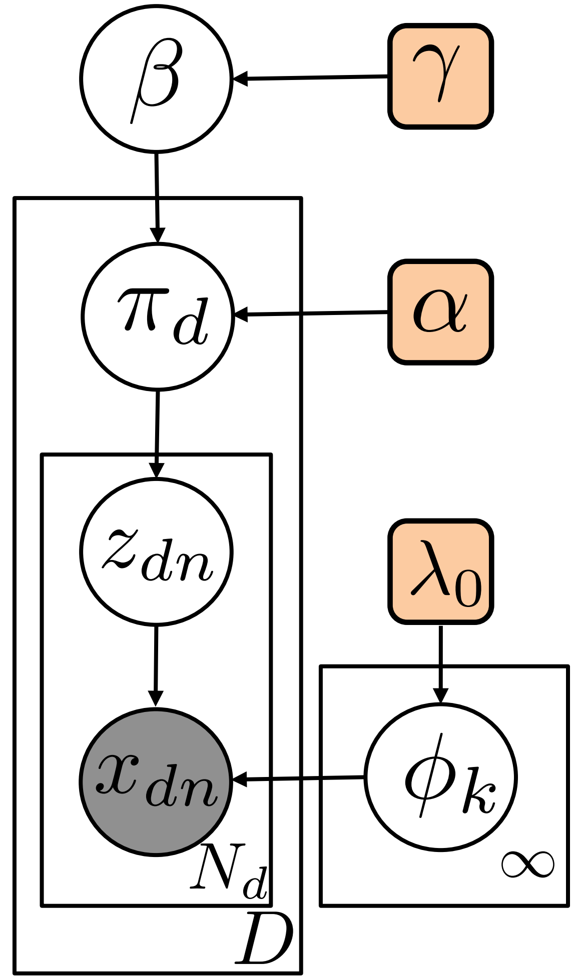
Genetics, Climate Change, Politics, …





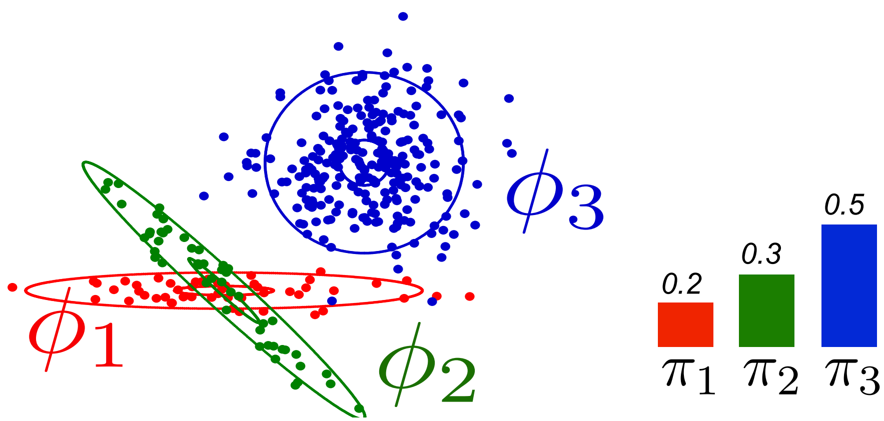


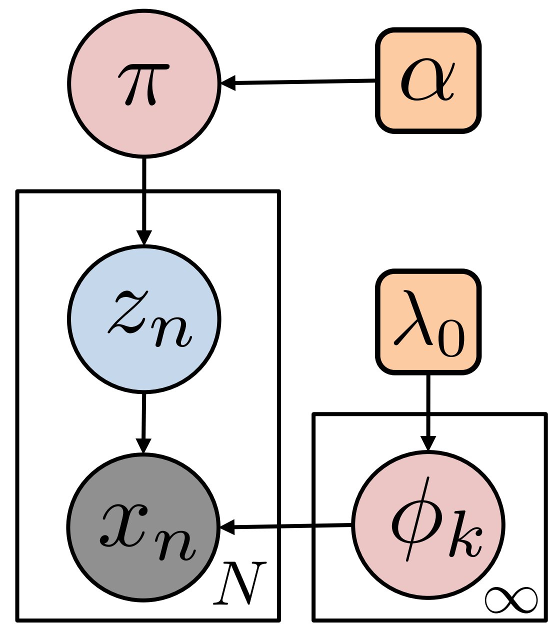
- ***Nonparametric:*** Data-driven discovery of model structure: *topics, behaviors, objects, communities…*
- ***Reliable:*** Structure driven by data and modeling assumptions, not heuristic algorithm initializations
- ***Parsimonious:*** Want a single model structure with good predictive power, not full posterior uncertainty



Hierarchical Dirichlet Process
(Teh et al., JASA 2006)

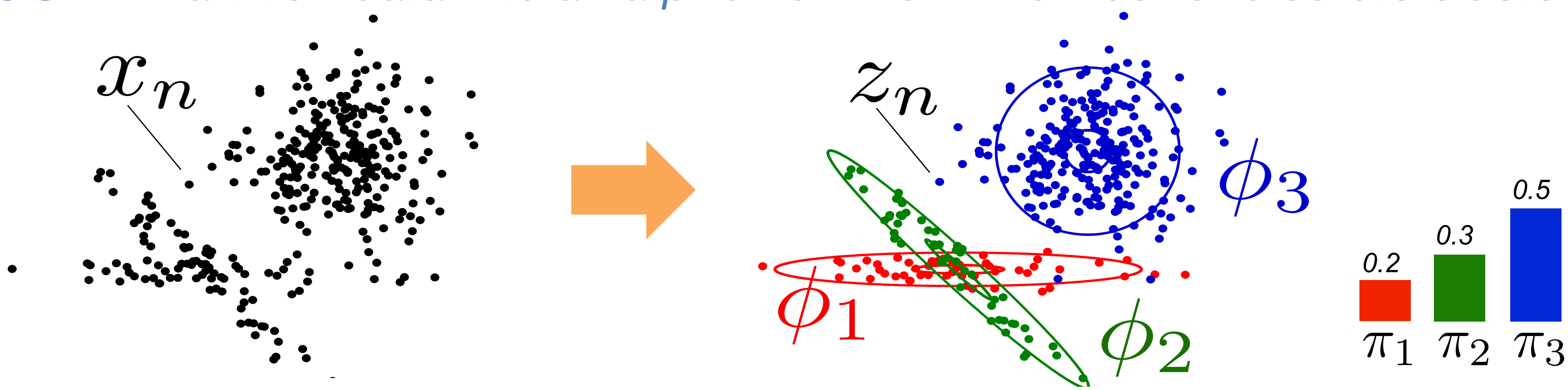
Variational Inference for Dirichlet Process Mixtures





Dirichlet Process Mixtures

GOAL: *Partition data into an a priori unknown number of discrete clusters.*



$$\phi_k \sim H(\lambda_0) \qquad \pi \sim \mathrm{Stick}(\alpha)$$

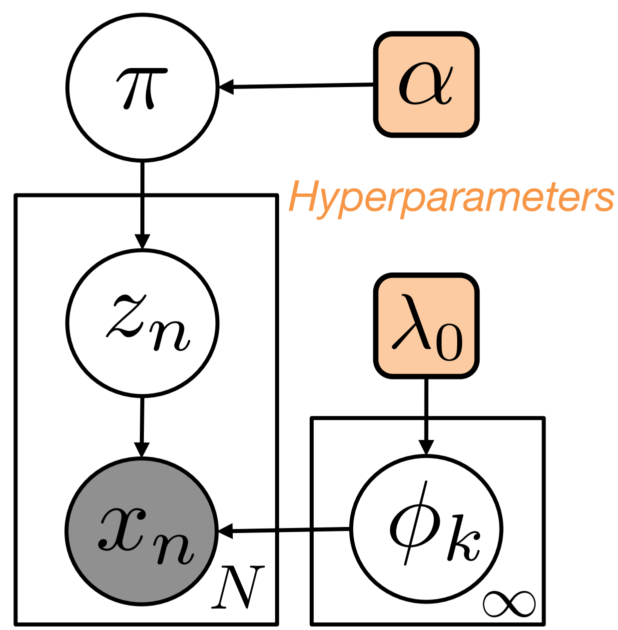
Each observation n = 1, 2, …, N:

- Cluster assignment: $z_n \sim \mathrm{Cat}(\pi)$
- Observed value: $x_n \sim \mathrm{F}(\phi_{z_n})$

Exponential family with conjugate prior:

$$f(x_n \mid \phi_k) = \exp\big(\phi_k^T t(x_n) - a(\phi_k)\big)$$

$t(x_n) \in \mathbb{R}^D$ are sufficient statistics



Variational Bounds

*Bayesian Learning: Maximize the **marginal likelihood** of our observed data*

➢ For any *variational distribution* $q(z, v, \phi)$:

$$\log p(x \mid \alpha, \lambda_0) = \log \sum_z \iint p(x, z, v, \phi \mid \alpha, \lambda_0) \, dv d\phi$$

Jensen's Inequality

$$\geq \mathbb{E}_q[\log p(x, z, v, \phi \mid \alpha, \lambda_0)] - \mathbb{E}_q[\log q(z, v, \phi)] = \mathcal{L}(q)$$

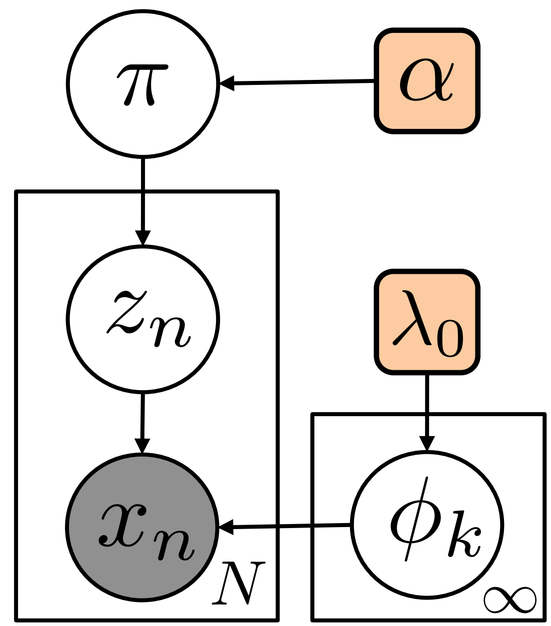
Expected log-likelihood (negative of "average energy")

Variational entropy

$$\pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell)$$

➢ Maximizing this bound recovers true posterior:

$$\mathcal{L}(q) = \log p(x \mid \alpha, \lambda_0) \\ -\mathrm{KL}(q(z, v, \phi) \,||\, p(z, v, \phi \mid x, \alpha, \lambda_0))$$



➢ The simplest *mean field* variational methods create tractable algorithms via *assumed independence*:

$$q(z, v, \phi) = q(z) q(v, \phi)$$

Approximating Infinite Models

$$q(z_n = k) = r_{nk}$$

Beta Distribution

Exponential Family from Conjugate Prior

$$q(z, v, \phi) = q(z)q(v, \phi) = \left[\prod_{n=1}^{N} q(z_n)\right] \cdot \left[\prod_{k=1}^{\infty} q(v_k)q(\phi_k)\right]$$

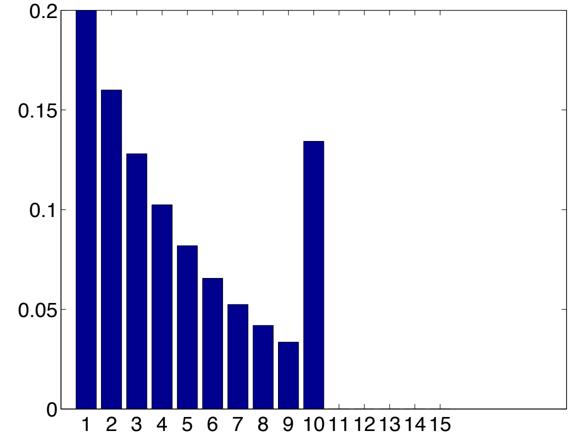
Categorical distribution with unbounded support, and infinitely many potential clusters!

Top-Down Model Truncation

Blei & Jordan, 2006; Ishwaran & James, 2001

$$q(z_n) = \mathrm{Cat}(z_n \mid r_{n1}, r_{n2}, \ldots, r_{nK})$$

$$q(v, \phi) = \left[\prod_{k=1}^{K} q(\phi_k)\right] \cdot \left[\prod_{k=1}^{K-1} q(v_k)\right], \quad v_K = \prod_{k=1}^{K-1}(1 - v_k).$$



$$\alpha = 4, K = 10$$

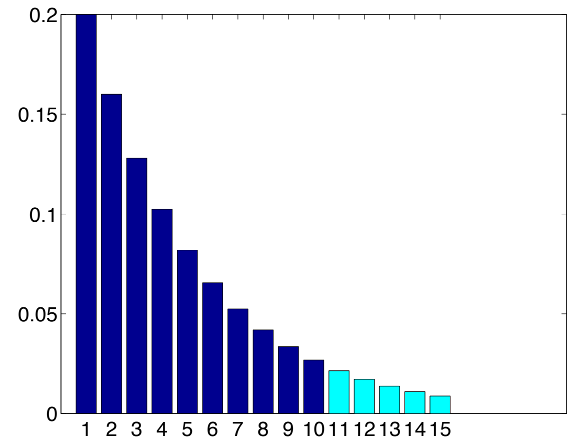
Bottom-Up Assignment Truncation

Bryant & Sudderth, 2012; Teh, Kurihara, & Welling, 2008

$$q(z_n) = \mathrm{Cat}(z_n \mid r_{n1}, r_{n2}, \ldots, r_{nK}, 0, 0, 0, \ldots)$$

$$q(v, \phi) = \prod_{k=1}^{\infty} q(v_k)q(\phi_k)$$

For any k>K, optimal variational distributions equal prior & need not be explicitly represented



Batch Variational Updates

A Bayesian nonparametric analog of Expectation-Maximization (EM)

$$q(z, v, \phi) = \left[\prod_{n=1}^{N} q(z_n \mid r_n)\right] \cdot \left[\prod_{k=1}^{\infty} \mathrm{Beta}(v_k \mid \alpha_{k1}, \alpha_{k0}) h(\phi_k \mid \lambda_k)\right]$$

$$q(z_n) = \mathrm{Cat}(z_n \mid r_{n1}, r_{n2}, \ldots, r_{nK}, 0, 0, 0, \ldots)$$ *for some K>0*

Update Assignments (The Expectation Step):

For all N data,

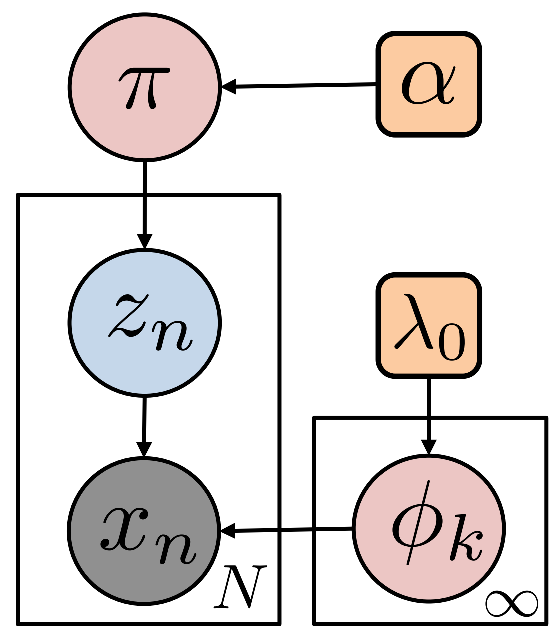
$$r_{nk} \propto \exp(\mathbb{E}_q[\log \pi_k(v)] + \mathbb{E}_q[\log p(x_n \mid \phi_k)]) \qquad \text{for } k \leq K$$

Update Cluster Parameters (The Other Expectation Step):

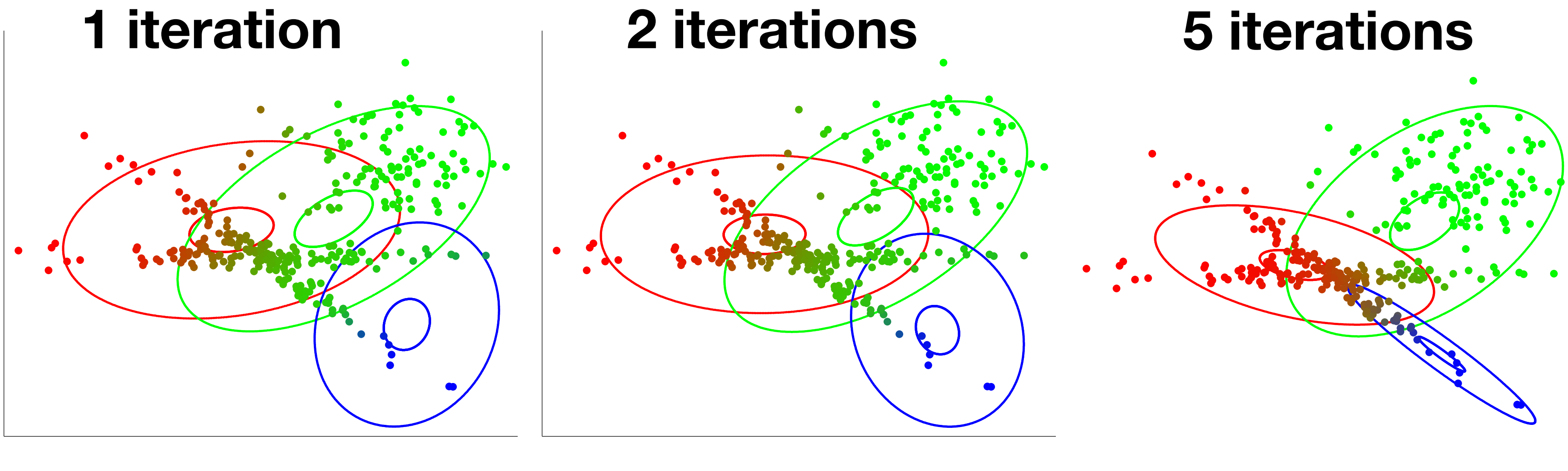
$$s_k^0 \leftarrow \textstyle\sum_{n=1}^{N} r_{nk} t(x_n)$$

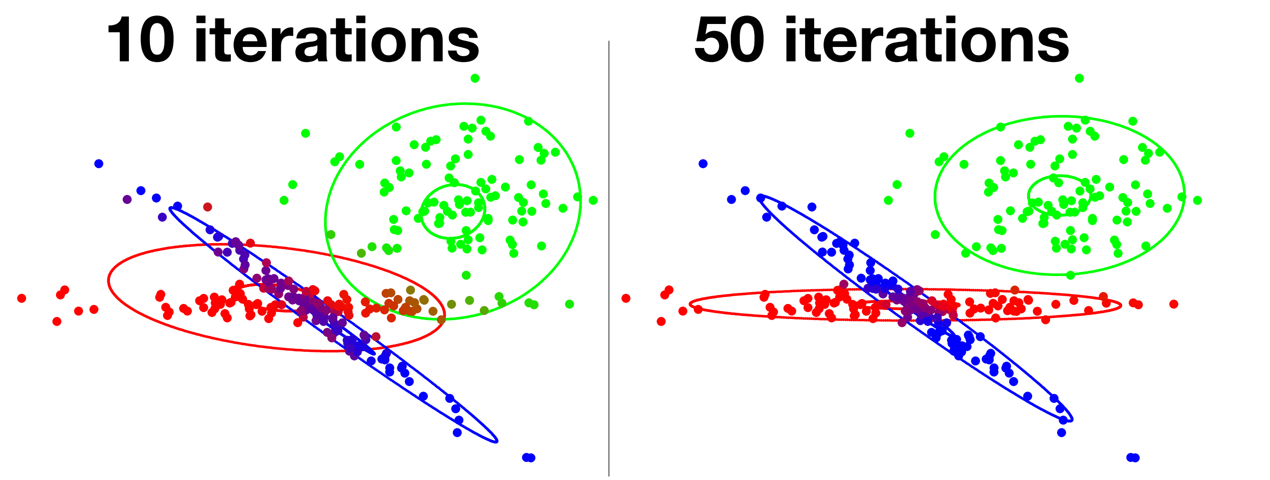
$$\lambda_k \leftarrow \lambda_0 + s_k^0$$

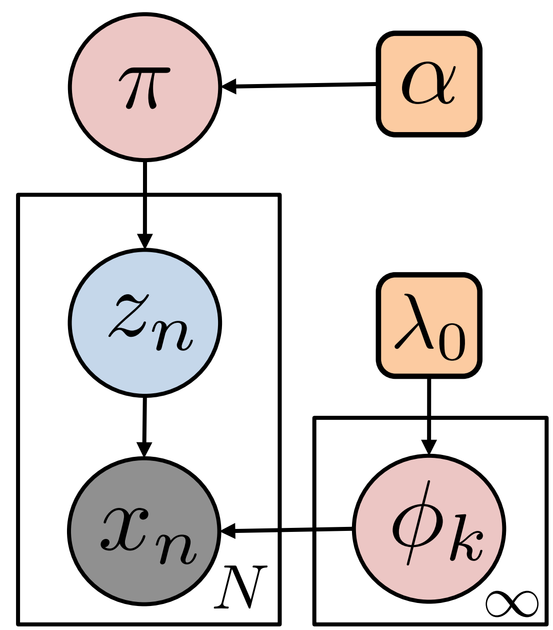
Expected counts and sufficient statistics are only non-zero for first K clusters.



Variational EM: Convergence





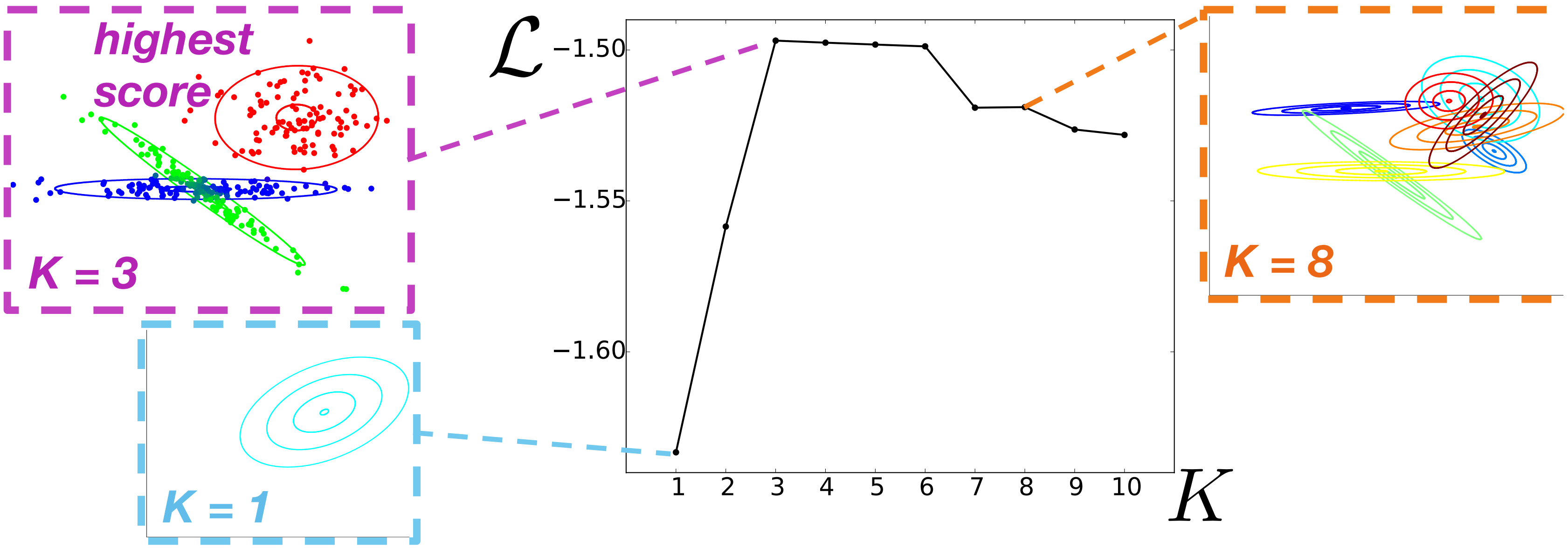


+ Likelihood bound monotonically increases to mode
– Each iteration must examine all data (SLOW)

Bayesian Model Selection

Maximizing marginal likelihood enables Bayesian model selection

$$\log p(x) \geq \mathbb{E}_q[\log p(x, z, v, \phi \mid \alpha, \lambda_0)] - \mathbb{E}_q[\log q(z, v, \phi)] = \mathcal{L}(q)$$

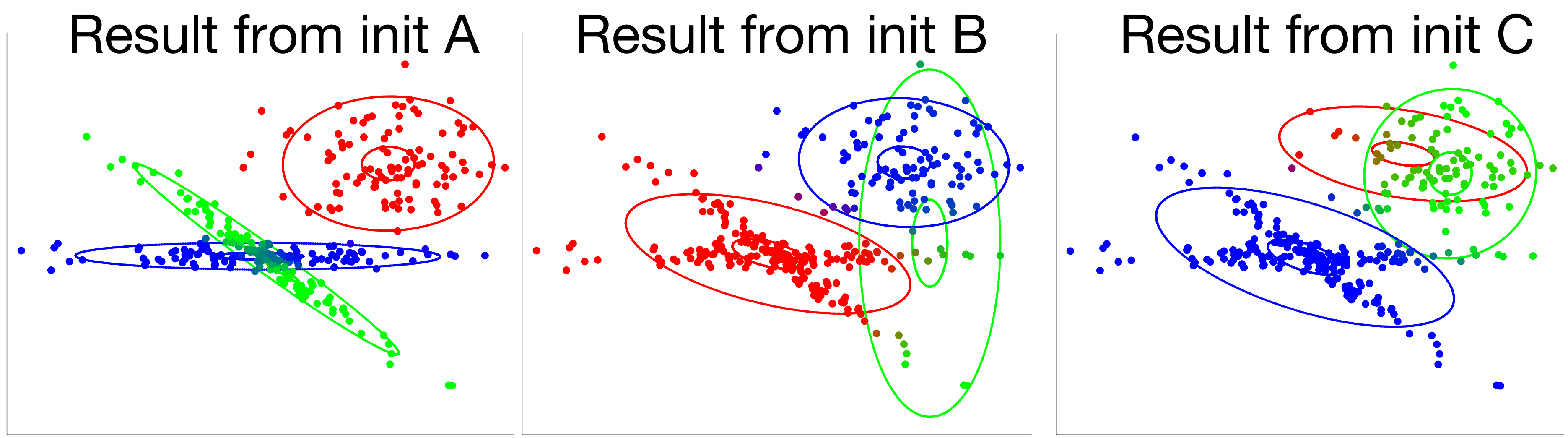


\+ Allows Bayesian comparison of hypotheses with varying complexity *K*.
For BNP models, MAP estimation will cause severe overfitting!

– Truncation level *K* is fixed, must fit many different models (EXPENSIVE)

Variational EM: Local Optima

Final clusters can be (highly) sensitive to initialization!



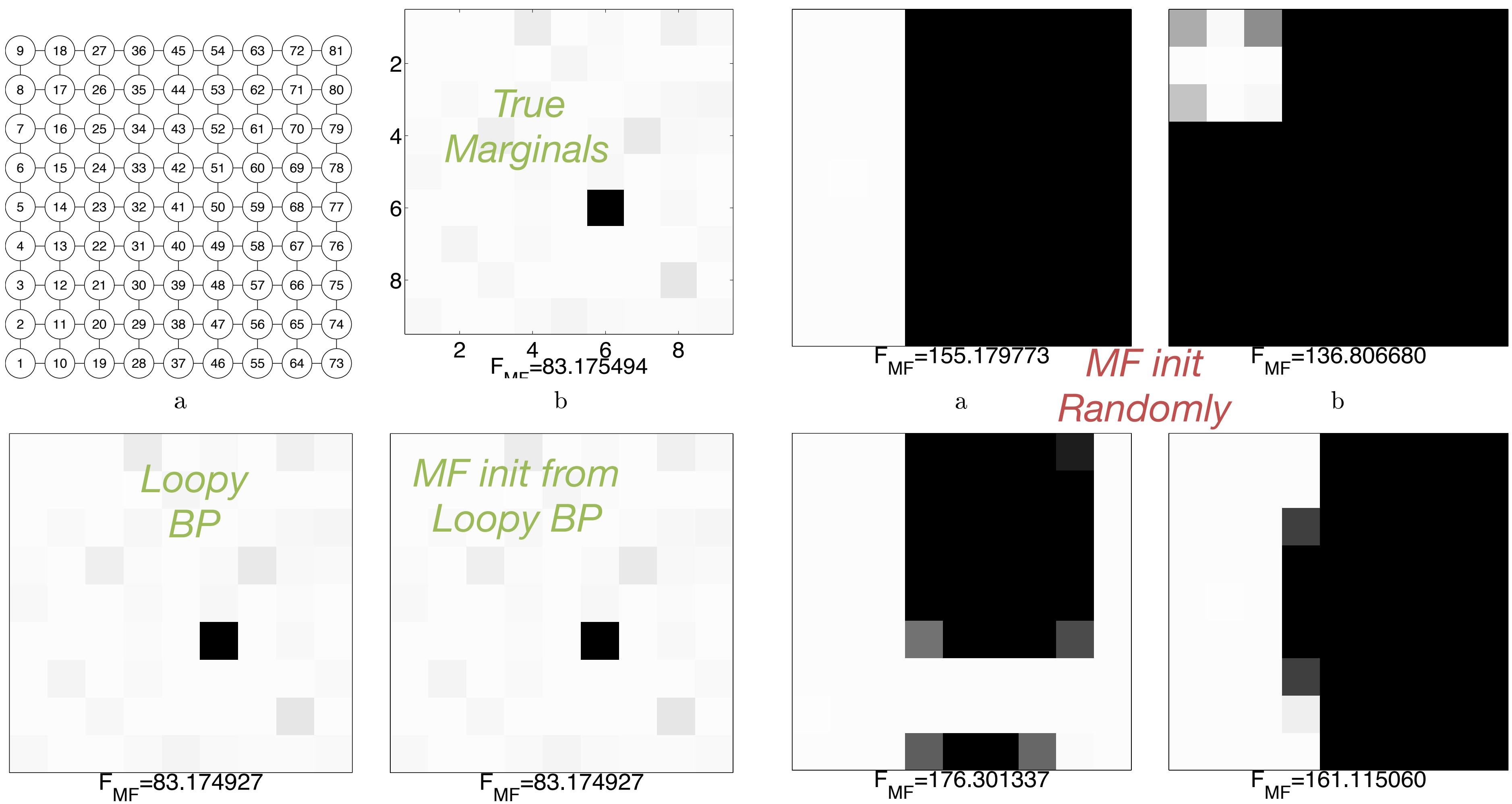
Heuristics commonly used in practice:

- Run from many different random initializations
- Use application intuition to engineer reasonable initializations
- Repeat for each complexity hypotheses (number of clusters *K*)

Requires expertise, not-big datasets,
and often compromises in model sophistication.

Mean Field versus Loopy BP

Toroidal 9x9 Grid with Attractive Binary Potentials (Weiss 2001)

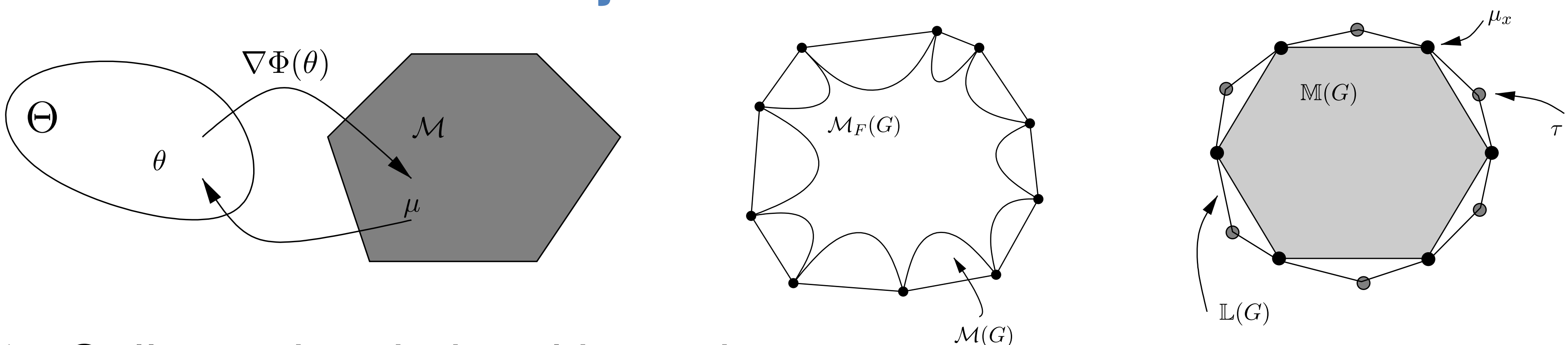


Optimize mean field via coordinate ascent on node marginals.

Objective versus Algorithm

Variational Inference Objectives:

Wainwright & Jordan, 2008



- Collapsed variational bounds
- Bethe and Kikuchi variational expansions
- Loop series expansions and cycle polytopes
- Fractional, reweighted, and convexified variational methods
- …

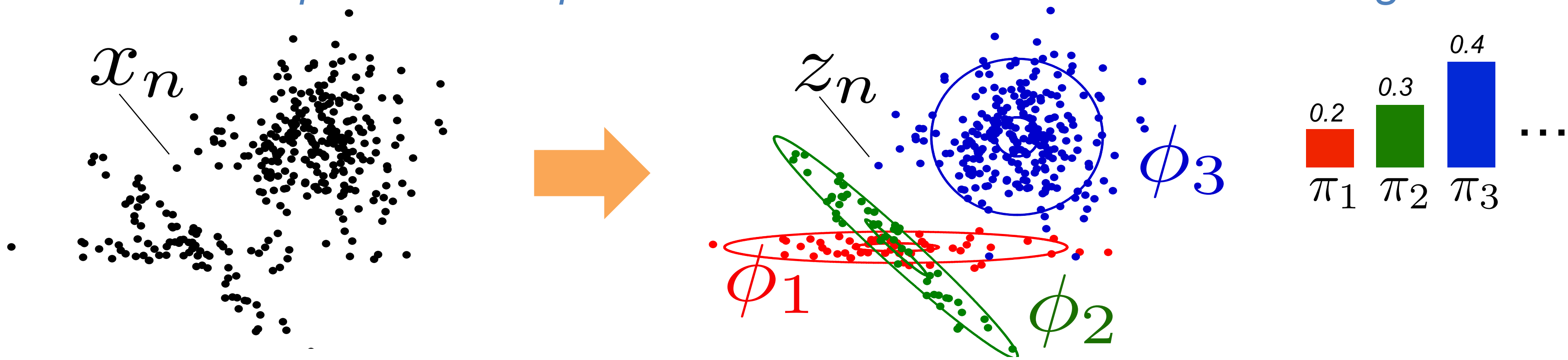
Variational Inference Algorithms:

- **Coordinate ascent:** Pick one free parameter, fix others, take step towards improving objective
- For non-convex objectives, we need improved algorithms!

Why not MCMC?
It’s asymptotically exact…

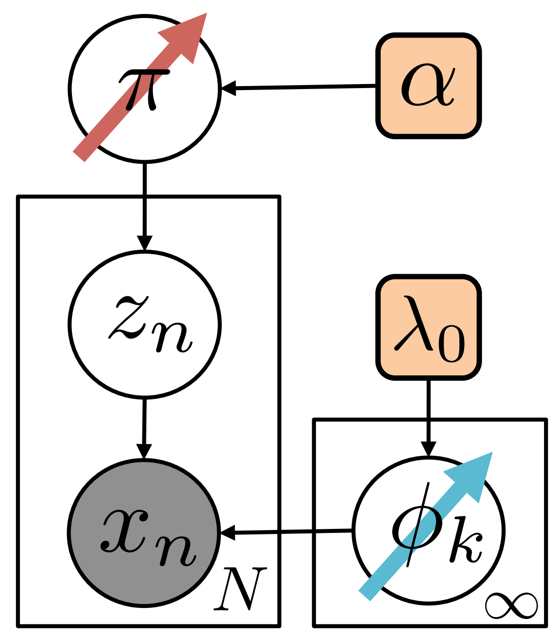
MCMC for DP Mixtures

Can we sample from the posterior distribution over data clusterings?



Given any fixed partition z:

- Marginalize cluster frequencies via *species sampling prediction rule* (Chinese restaurant process)
- Via *conjugacy* of base measure to exponential family likelihood, marginalize cluster shape parameters

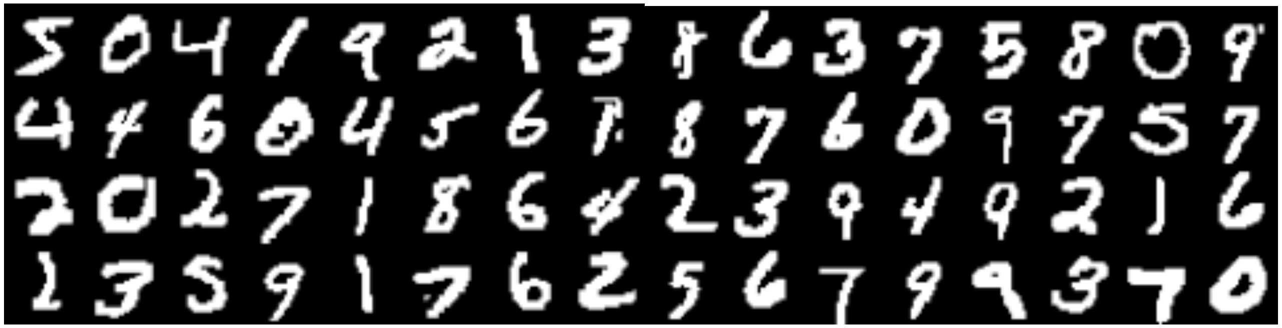


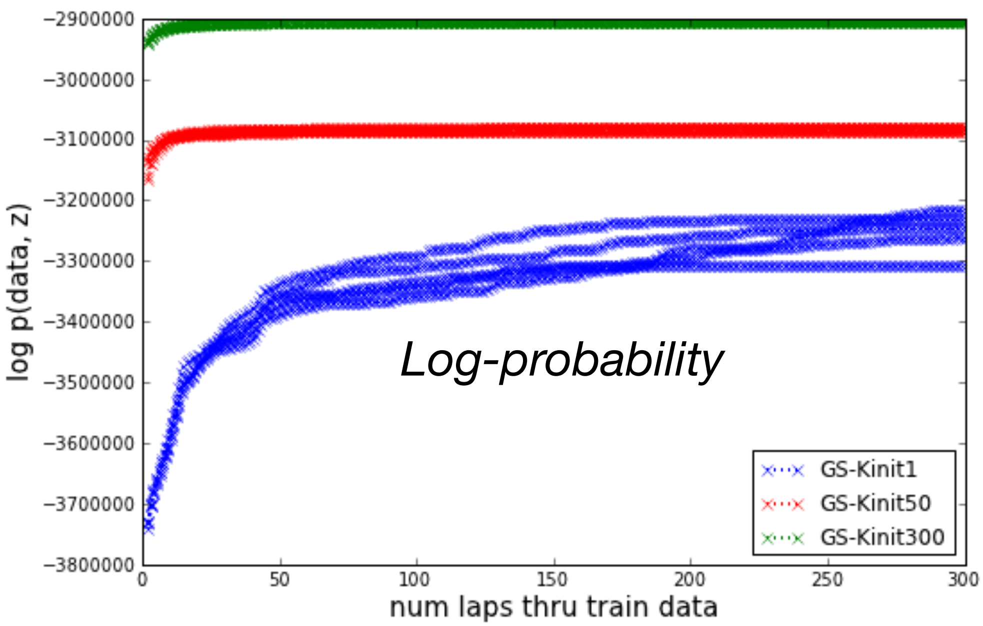
Gibbs Sampler: *(Neal 1992, MacEachern 1994)*

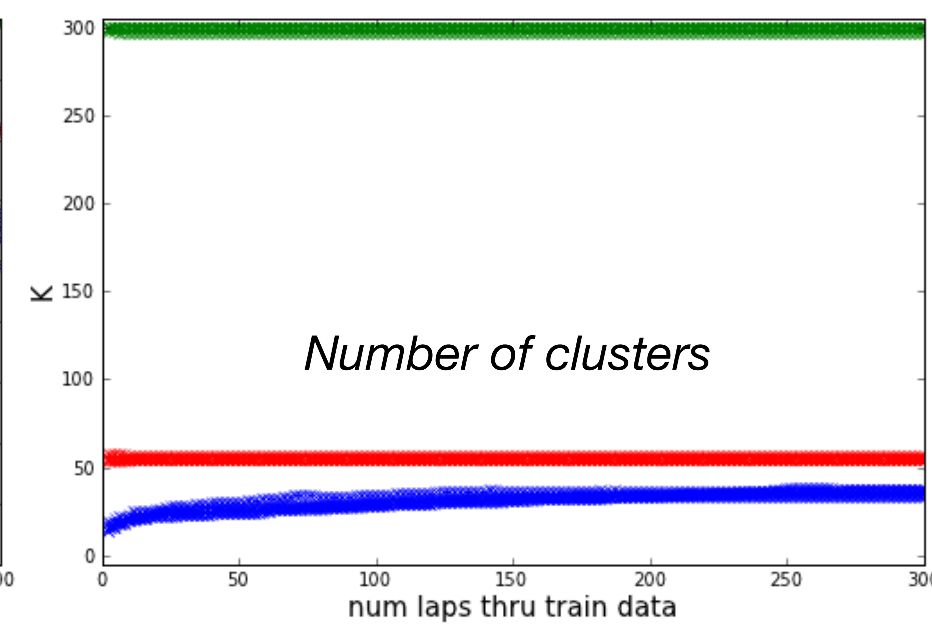
Iteratively resample cluster assignment for one observation, fixing all others.

Mixing for DP Mixture Samplers

MNIST: *60,000 digits projected to 50 dimensions via PCA.*



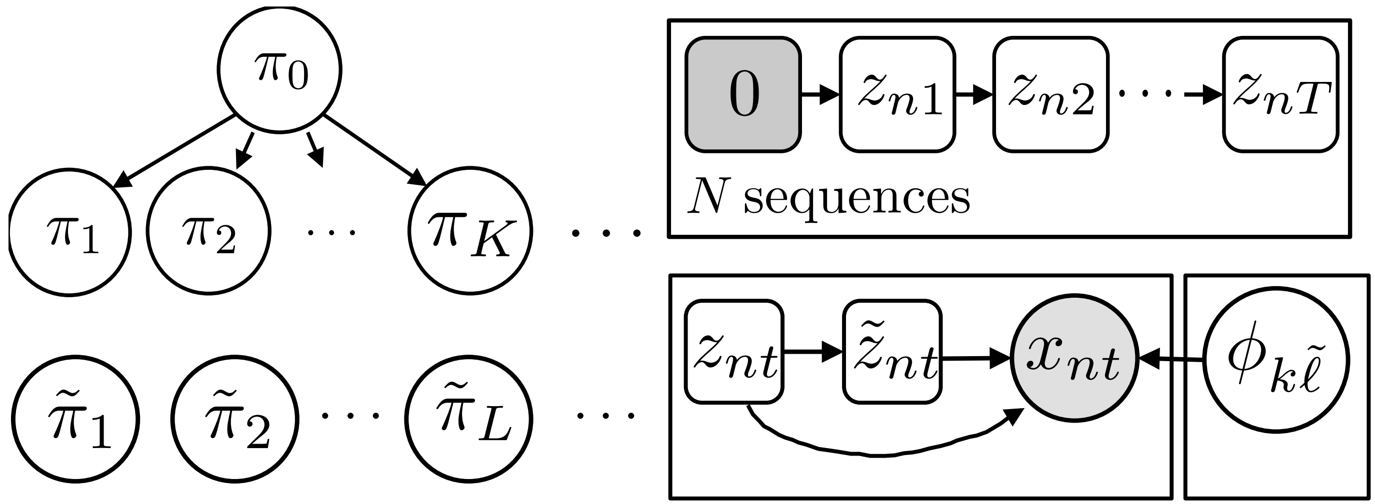




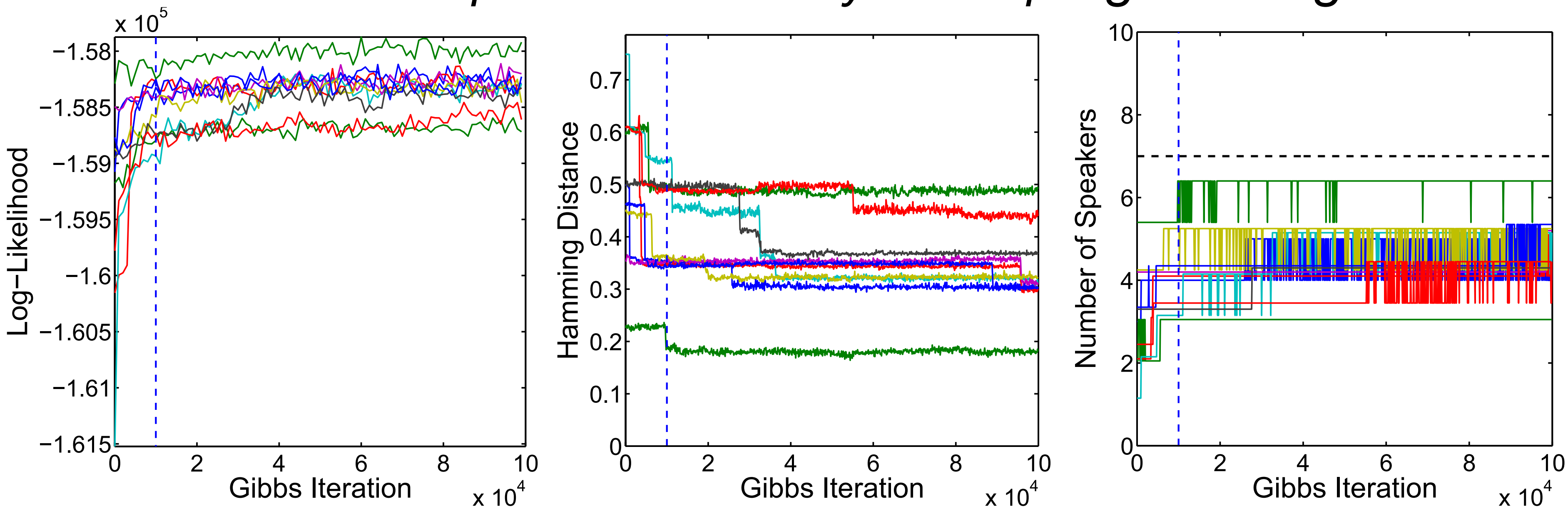
- Five random initializations from *K=1*, *K=50*, *K=300* clusters
- Need good initialization for good results. Can we do better?

MCMC for HDP-HMM Diarization

***GOAL:** Recover unknown set of people, and when each one spoke, from audio data*

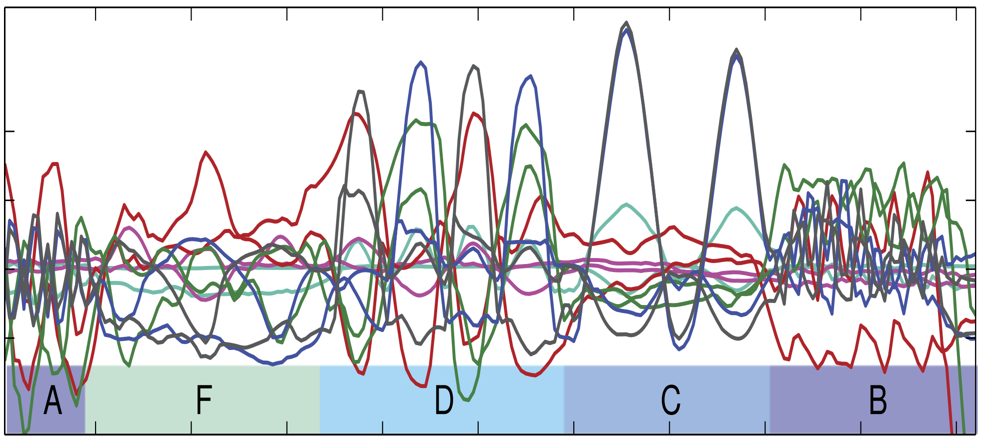


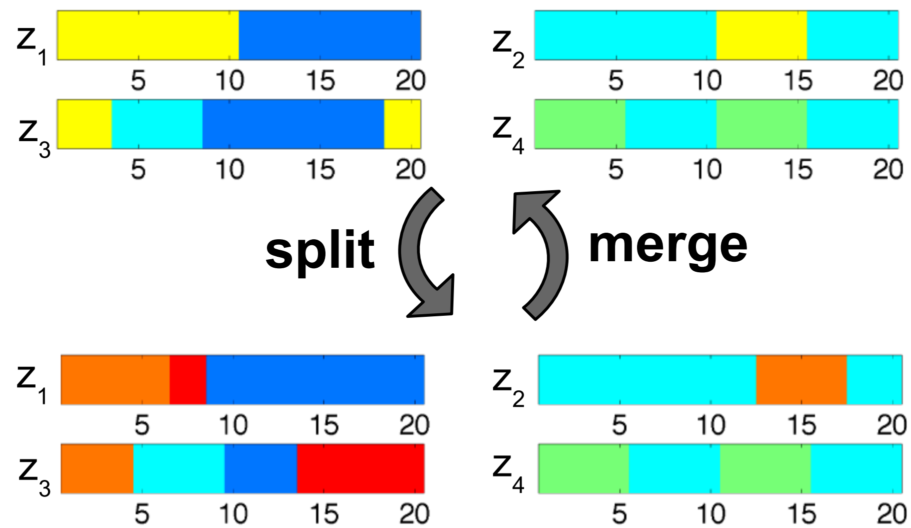
Blocked Gibbs sampler based on dynamic programming:



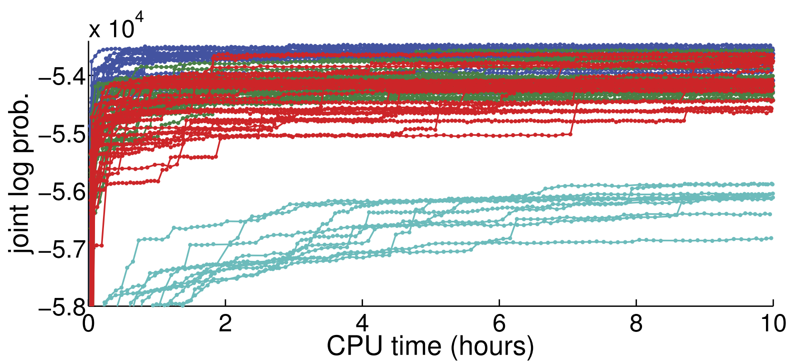
Fox, Sudderth, Jordan, & Willsky, AOAS 2011

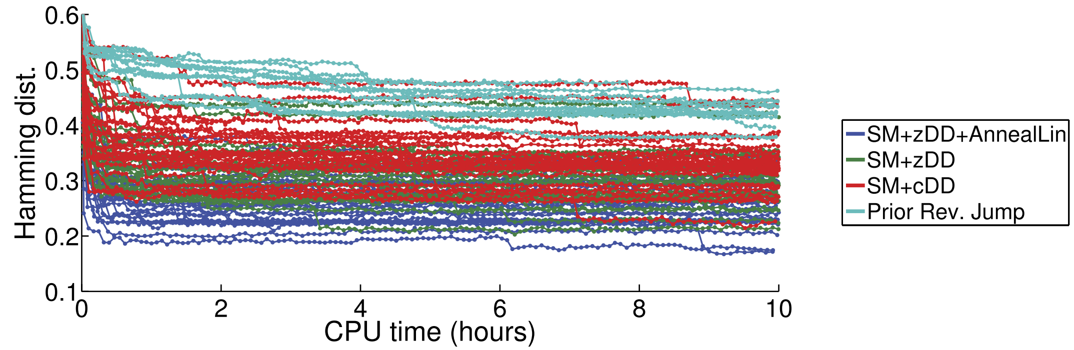
Reversible Jump MCMC?





Sequentially allocated split-merge RJ-MCMC for BP-HMM:





Correct MCMC proposals versus **annealed acceptance ratio.**

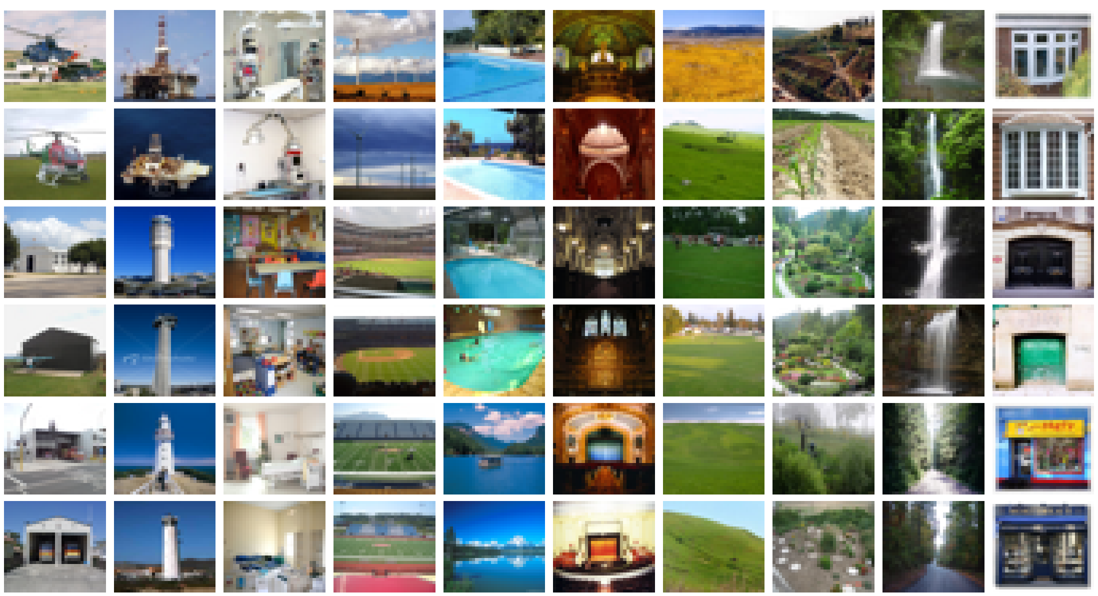
Combinatorial factors overwhelming for big datasets!

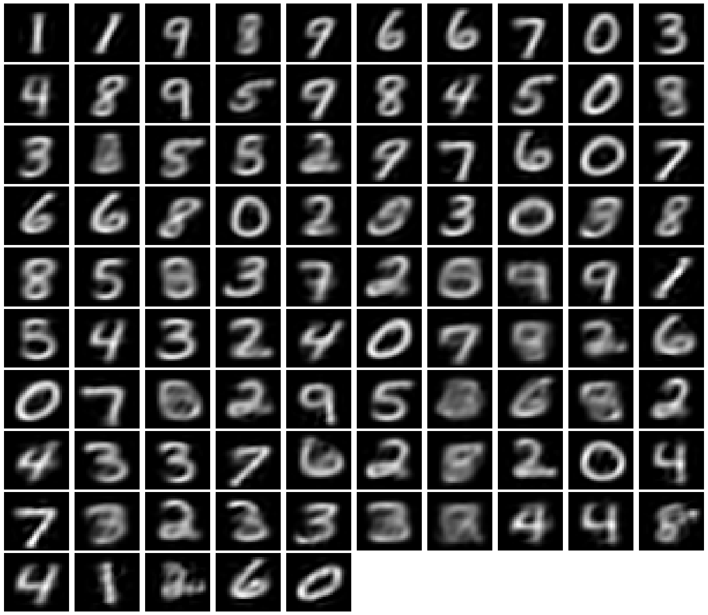
Fox, Hughes, Sudderth, & Jordan, AOAS 2014

Memoized Variational Inference for Dirichlet Process Mixture Models

Michael Hughes & E. Sudderth

2013 Conference on Neural Information Processing Systems





Stochastic Variational Inference

Hoffman, Blei, Paisley, & Wang, JMLR 2013

Stochastically partition large dataset into *B* smaller *batches*:

Update: *For each batch b*

$$r(\mathcal{B}_b) \leftarrow \mathrm{Estep}(x(\mathcal{B}_b), \alpha, \lambda)$$

For cluster k = 1, 2, … K:

$$s_k^b \leftarrow \sum_{n \in \mathcal{B}_b} r_{nk} t(x_n)$$

$$\lambda_k^b \leftarrow \lambda_0 + \frac{N}{|\mathcal{B}_b|} s_k^b$$

$$\lambda_k \leftarrow \rho_t \lambda_k^b + (1-\rho_t)\lambda_k$$

batch stats give noisy estimate of (natural) gradient

Apply similar updates to stick weights.

Data

$x(\mathcal{B}_1)$

$x(\mathcal{B}_2)$

⋮

$x(\mathcal{B}_b)$

⋮

$x(\mathcal{B}_B)$

Learning Rate

$$\rho_t \triangleq (\rho_0 + t)^{-\kappa}$$

Robbins-Monro convergence condition:

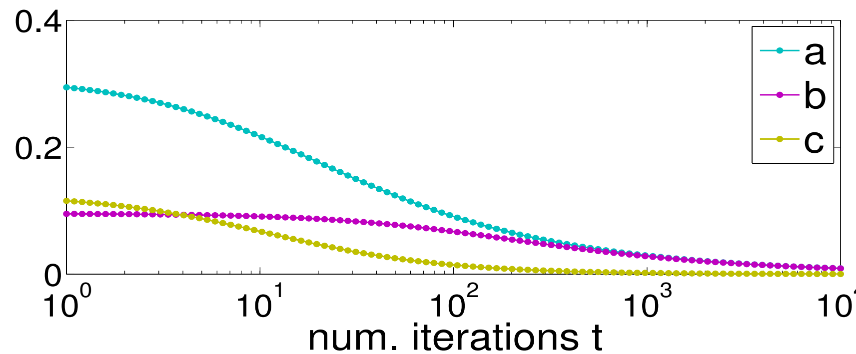
$$\sum_t \rho_t \rightarrow \infty \qquad \kappa \in (.5, 1]$$

$$\sum_t \rho_t^2 < \infty$$

Properties of stochastic inference:

+ Per-iteration cost is low
+ Initial progress is rapid

- Objective is highly non-convex, so convergence guarantee is weak
- Sensitivity to batch size & learning rate



Memoized Variational Inference

Hughes & Sudderth, NIPS 2013; Neal & Hinton 1999

Memoization: Storage (caching) of results of previous computations

Update: *For each batch b*

$$r(\mathcal{B}_b) \leftarrow \mathrm{Estep}(x(\mathcal{B}_b), \alpha, \lambda)$$

For cluster k = 1, 2, … K:

$$s_k^0 \leftarrow s_k^0 - s_k^b$$
$$s_k^b \leftarrow \sum_{n \in \mathcal{B}_b} r_{nk} t(x_n)$$
$$s_k^0 \leftarrow s_k^0 + s_k^b$$
$$\lambda_k \leftarrow \lambda_0 + s_k^0$$

Apply similar updates to stick weights.

batch stats allow exact estimation from partial E-steps

Data

$$x(\mathcal{B}_1),\ x(\mathcal{B}_2),\ \vdots,\ x(\mathcal{B}_b),\ \vdots,\ x(\mathcal{B}_B)$$

Batch Summaries

$$\begin{array}{cccc} s_1^1 & s_2^1 & \cdots & s_K^1 \\ s_1^2 & s_2^2 & \cdots & s_K^2 \\ \vdots & \vdots & & \vdots \\ s_1^B & s_2^B & \cdots & s_K^B \end{array}$$

Global Summary

$$s_1^0 \quad s_2^0 \quad \cdots \quad s_K^0$$

$$s_k^0 = s_k^1 + s_k^2 + \ldots s_k^B$$

Properties of memoized inference:

- \+ Per-iteration cost is low
- \+ Initial progress is rapid
- \+ Insensitive to batch size, no learning rate
- – Requires storage proportional to number of batches (NOT number of observations)

Memoized Variational Inference

Hughes & Sudderth, NIPS 2013; Neal & Hinton 1999

Memoization: Storage (caching) of results of previous computations

Update: *For each batch b*

$r(\mathcal{B}_b) \leftarrow \mathrm{Estep}(x(\mathcal{B}_b), \alpha, \lambda)$

For cluster k = 1, 2, … K:

$s_k^0 \leftarrow s_k^0 - s_k^b$

$s_k^b \leftarrow \sum_{n \in \mathcal{B}_b} r_{nk} t(x_n)$

$s_k^0 \leftarrow s_k^0 + s_k^b$

$\lambda_k \leftarrow \lambda_0 + s_k^0$

batch stats allow exact estimation from partial E-steps

Apply similar updates to stick weights.

Data

$x(\mathcal{B}_1)$

$x(\mathcal{B}_2)$

$\vdots$

$x(\mathcal{B}_b)$

$\vdots$

$x(\mathcal{B}_B)$

Batch Summaries

$$\begin{array}{cccc} s_1^1 & s_2^1 & \cdots & s_K^1 \\ s_1^2 & s_2^2 & \cdots & s_K^2 \\ \vdots & \vdots & & \vdots \\ s_1^B & s_2^B & \cdots & s_K^B \end{array}$$

Global Summary

$$\begin{array}{cccc} s_1^0 & s_2^0 & \cdots & s_K^0 \end{array}$$

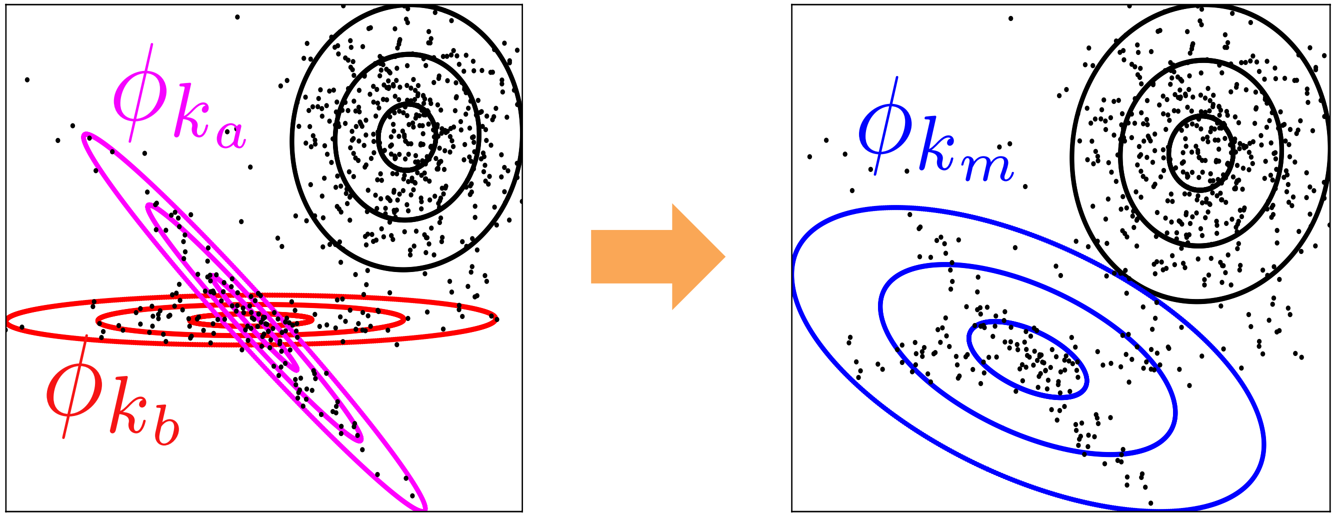
$$s_k^0 = s_k^1 + s_k^2 + \ldots s_k^B$$

An Inspiration:

A Stochastic Gradient Method with an Exponential Convergence Rate for Strongly-Convex Optimization with Finite Training Sets. N. Le Roux, M. Schmidt, F. Bach, NIPS 2012.

Memoized Cluster Merges

Merge two clusters into one for parsimony, accuracy, efficiency.



- New cluster takes over all responsibility for data assigned to old clusters:

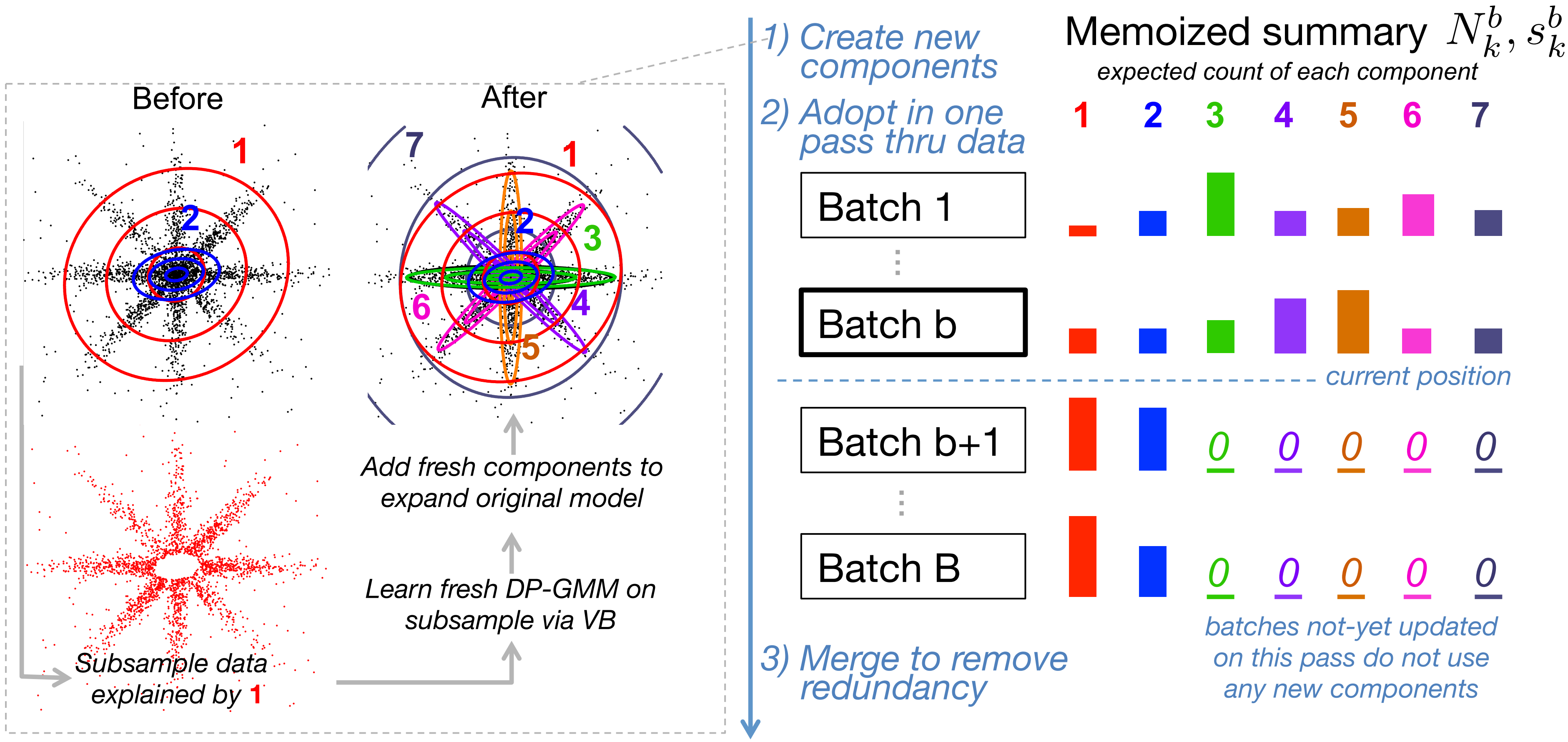
$$r_{nk_m} \leftarrow r_{nk_a} + r_{nk_b} \quad \Rightarrow \quad s^0_{k_m} \leftarrow s^0_{k_a} + s^0_{k_b}$$

- No batch processing required, efficiently evaluate via *memoized* statistics
- Accept or reject via *exact* full-dataset likelihood bound: $\mathcal{L}(q_{\text{merge}}) > \mathcal{L}(q)?$

Requires memoized entropy sums for candidate pairs of clusters; efficient implementation limits overhead.

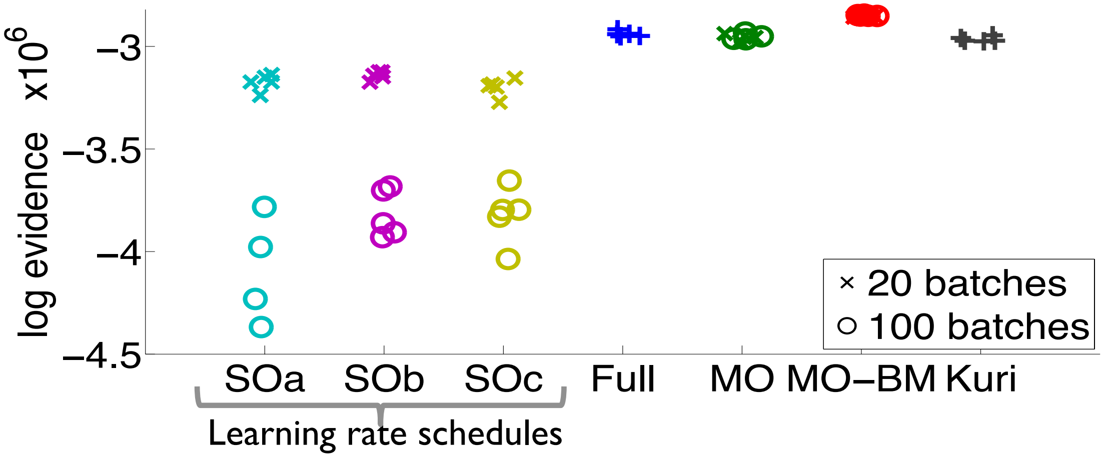
Memoized Cluster Births

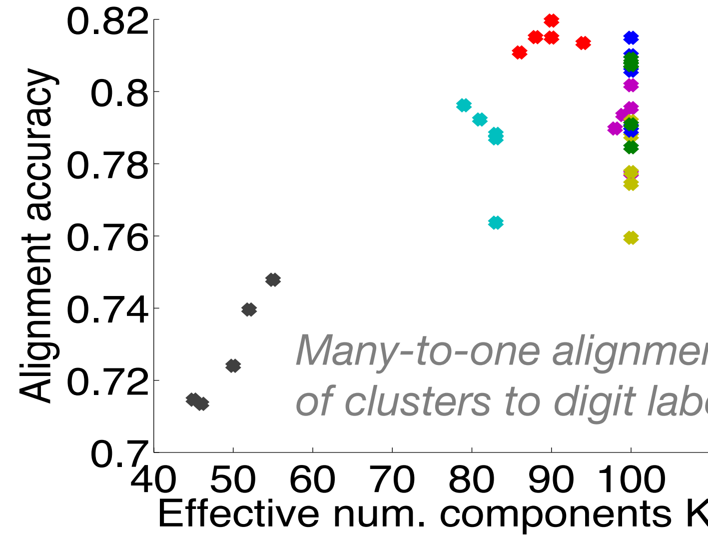
GOAL: Effective & efficiently verifiable cluster creation for general likelihoods.



Clustering Handwritten Digits

MNIST: *60,000 digits projected to 50 dimensions via PCA.*

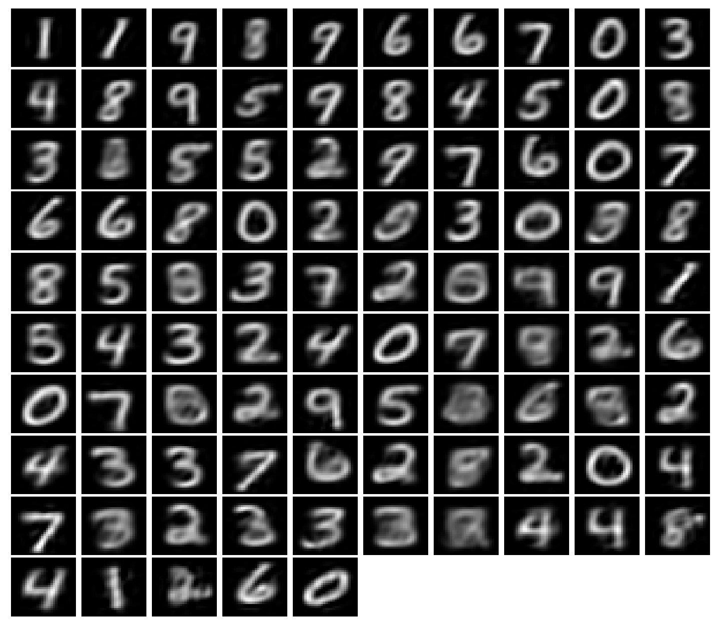




Batch, memoized, & memoized birth-merge

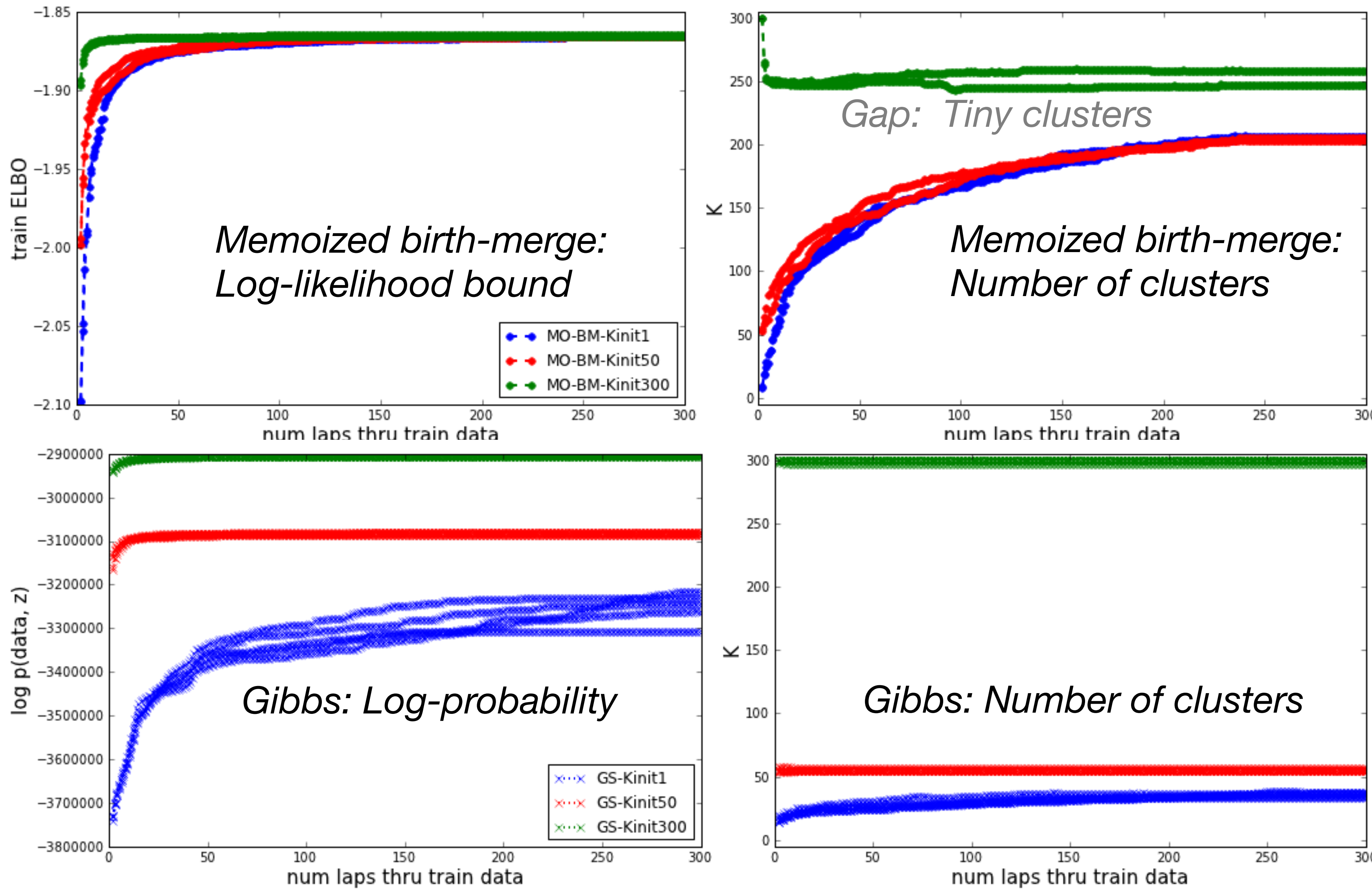
Stochastic variational: Rate a, Rate b, Rate c

Kurihara: Accelerated variational, NIPS 2006



Memoized birth-merge from K=1 has highest accuracy while using fewer clusters.

MNIST: Variational versus Gibbs

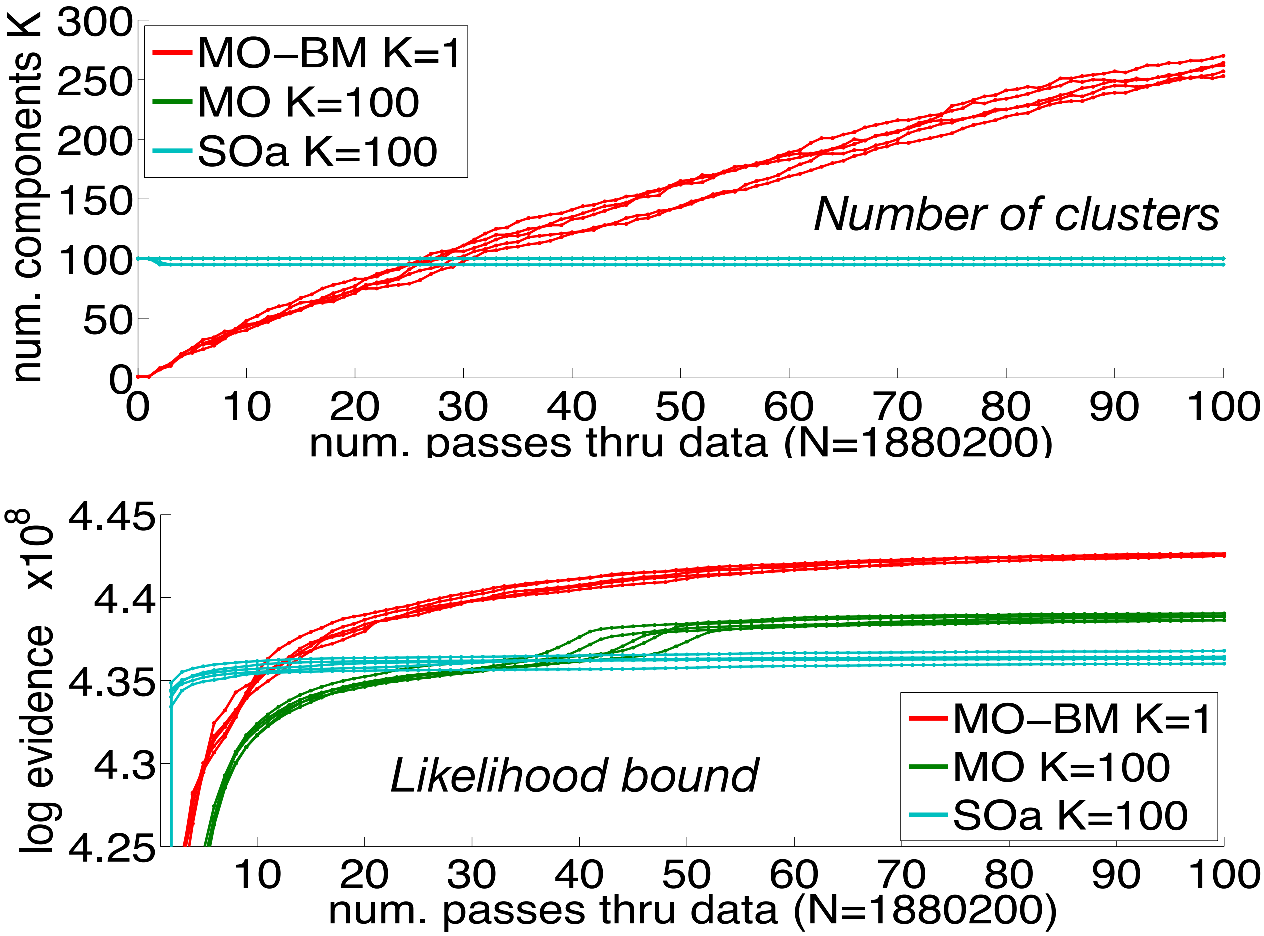


- Five random initializations from *K=1*, *K=50*, *K=300* clusters
- Diagonal-covariance Gaussians (change from previous slides)

Clustering Image Patches

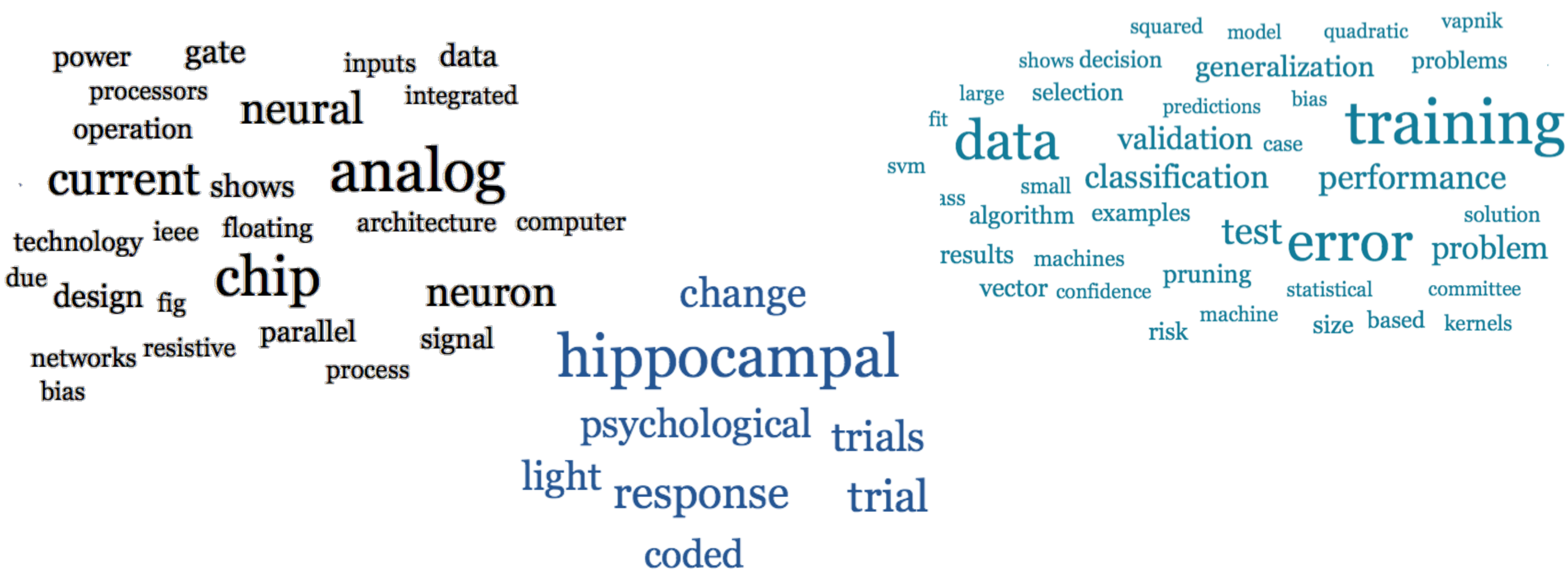
8x8 Image Patches (BSDS): *N=1.88 million*

- Memoized birth-merge allows growth in model complexity
- Effective performance as density model for image denoising



Memoized Variational Inference for Hierarchial DP Topic Models

Michael Hughes, Dae Il Kim, & E. Sudderth



Hierarchical DP Topic Model

Generalization of Latent Dirichlet Allocation (LDA, Blei 2003) by Teh et al. JMLR 2006.
Dependent Dirichlet process (DDP, MacEachern 1999) with group-specific weights.

- Global topic frequencies and parameters:

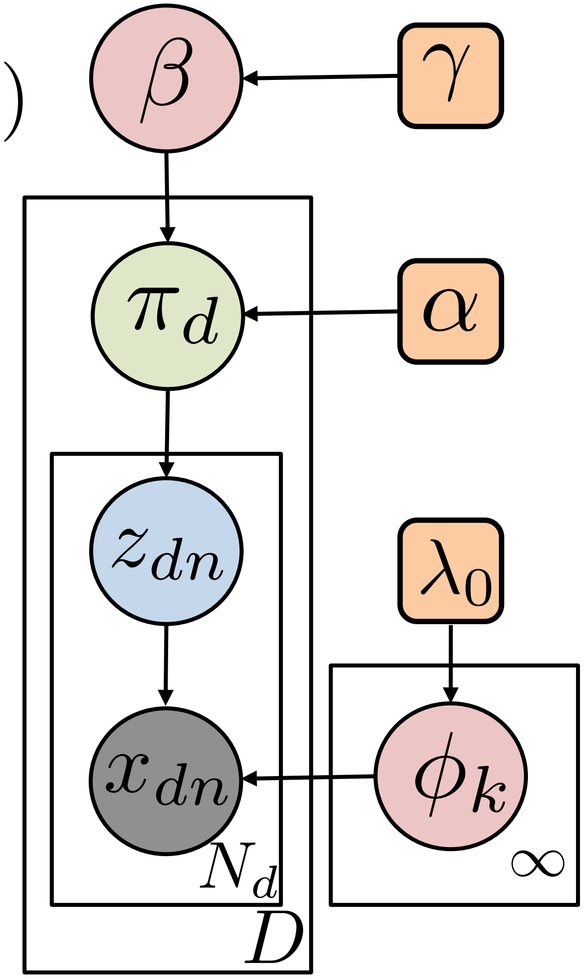
$$\beta_k = u_k \prod_{\ell=1}^{k-1} (1 - u_\ell) \qquad u_k \sim \mathrm{Beta}(1, \gamma)$$

$$\phi_k \sim \mathrm{Dirichlet}(\lambda_0)$$ *(sparse)*

- For each of *D* documents (groups):
  - Topic frequencies: $\pi_d \sim \mathrm{DP}(\alpha\beta)$

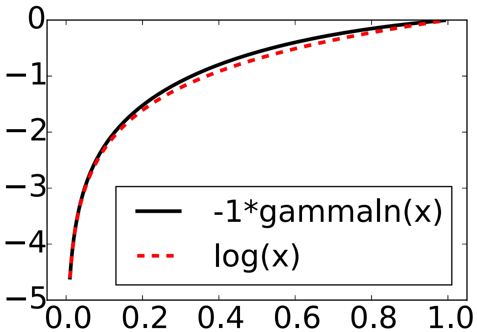
$$\mathbb{E}[\pi_{dk}] = \beta_k$$

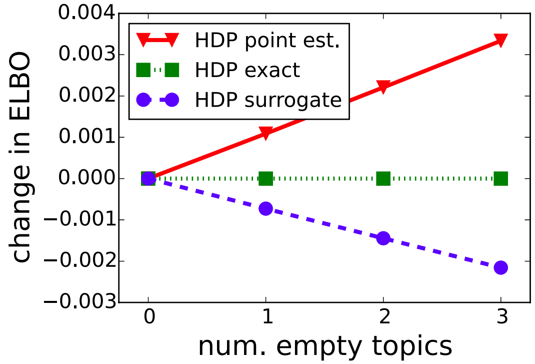
- For each of N_d words in document *d*:
  - Topic assignment: $z_{dn} \sim \mathrm{Cat}(\pi_d)$
  - Observed value: $x_{dn} \sim \mathrm{Cat}(\phi_{z_{dn}})$



Variational Learning of HDP Topics

Global Topic Frequencies: *Beta posteriors*

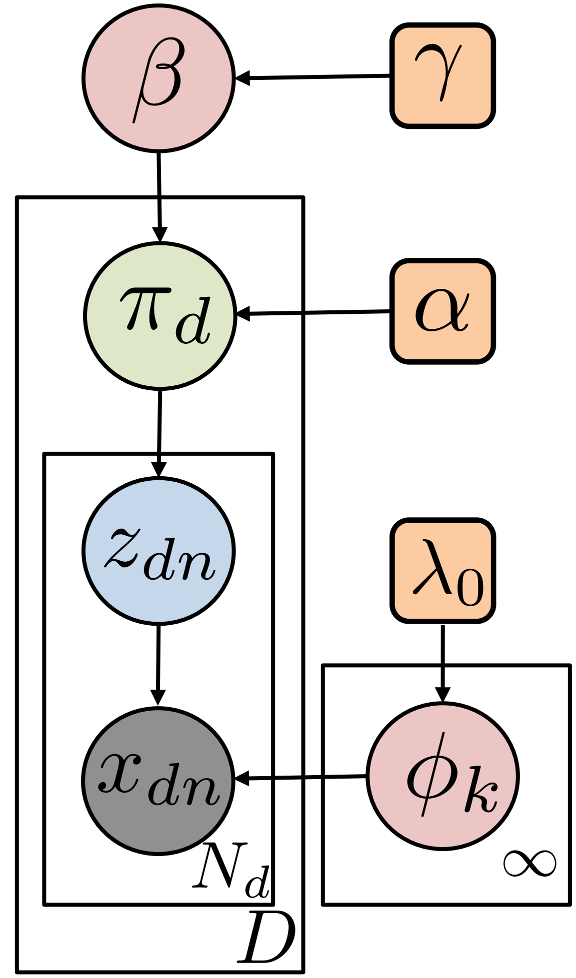




point estimate

numerical

logarithmic bound



Document-Topic Distributions:

- A DP induces a finite Dirichlet distribution on any finite partition. We consider *K+1* events:

$$(\pi_{d1}, \pi_{d2}, \ldots, \pi_{dK}, \pi_{d+}) \sim \mathrm{Dir}(\alpha, \beta)$$

$$\pi_{d+} = 1 - \sum_{k=1}^{K} \pi_{dk}$$

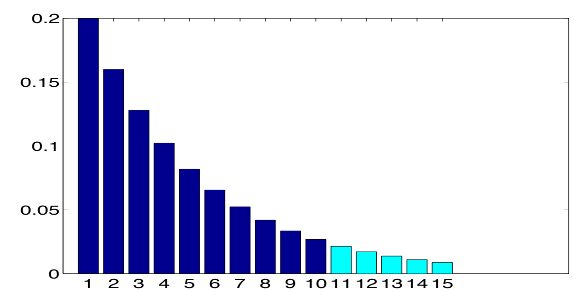
- Probabilities of *K* active topics, and infinite tail

Truncate Assignments:

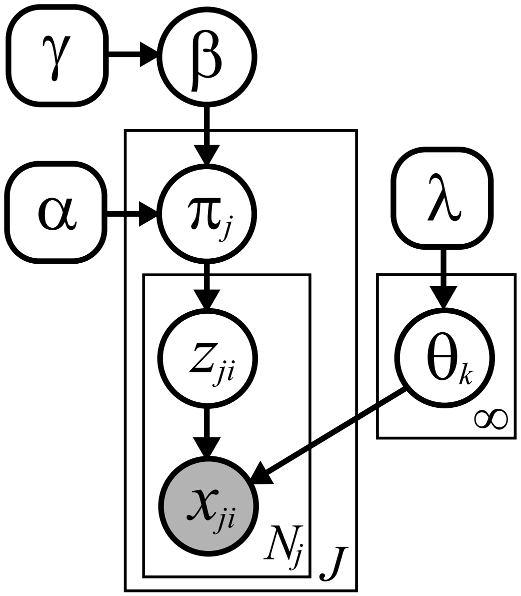
- For some current number of active topics *K*:

$$q(z_{dn}) = \mathrm{Cat}(z_{dn} \mid r_{dn1}, r_{dn2}, \ldots, r_{dnK}, 0, 0, \ldots)$$

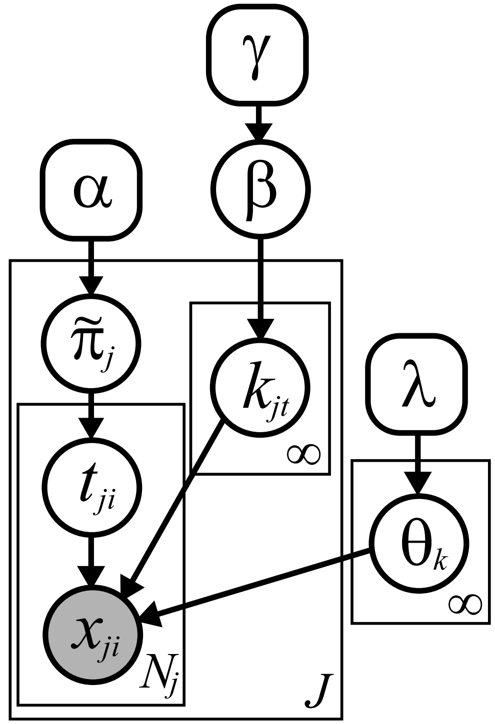
$$r_{dnk} \propto \exp(\mathbb{E}_q[\log \pi_{dk}(v_d)] + \mathbb{E}_q[\log p(x_{dn} \mid \phi_k)])$$



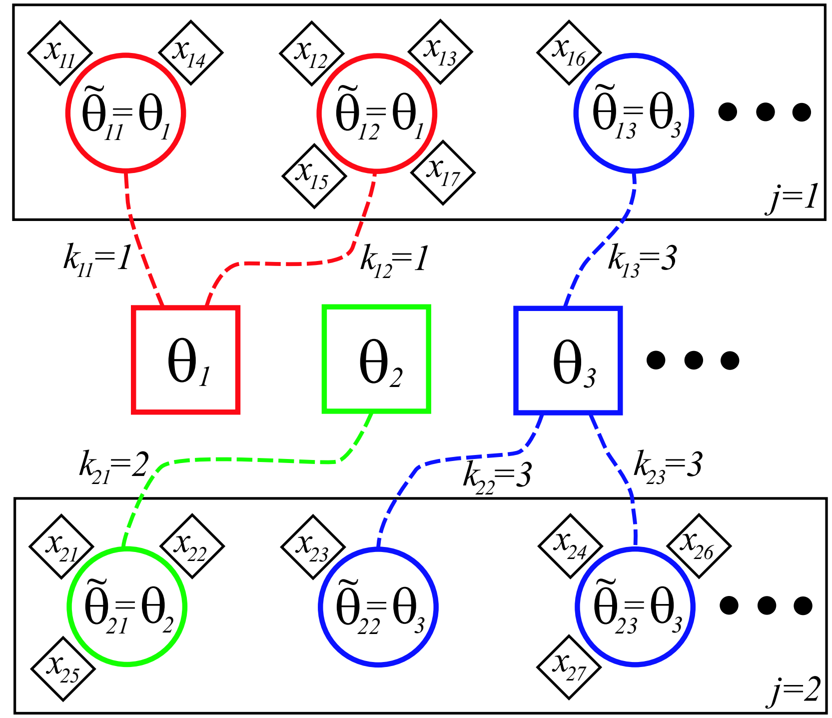
HDP Representations



HDP Direct Assignment



HDP Chinese Restaurant Franchise



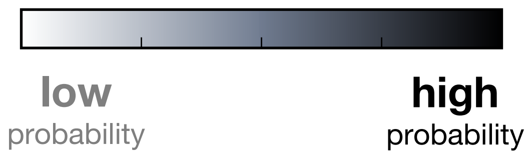
By introducing extra latent variables, the CRF:

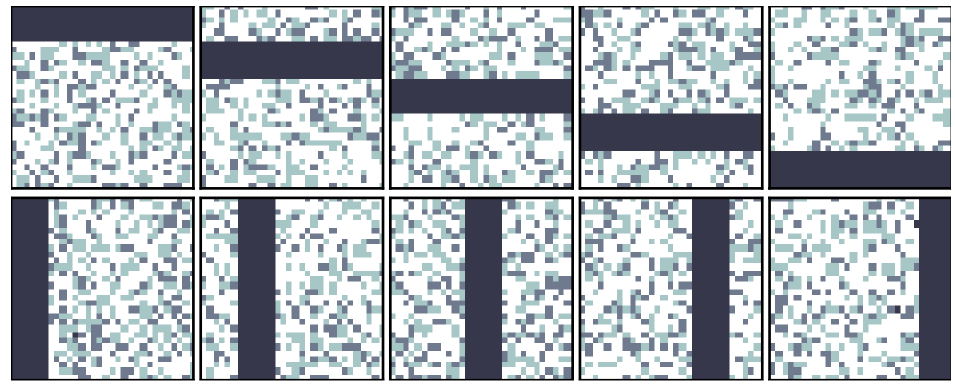
+ Makes all conditionals conjugate, closed-form inference
– Additional variables have very strong dependencies
– For both Gibbs and variational: slower, more local optima

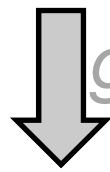
Toy Dataset: Bar Topics

10 Bar Topics:

900 vocabulary symbols arranged as 30x30 image, one pixel per word

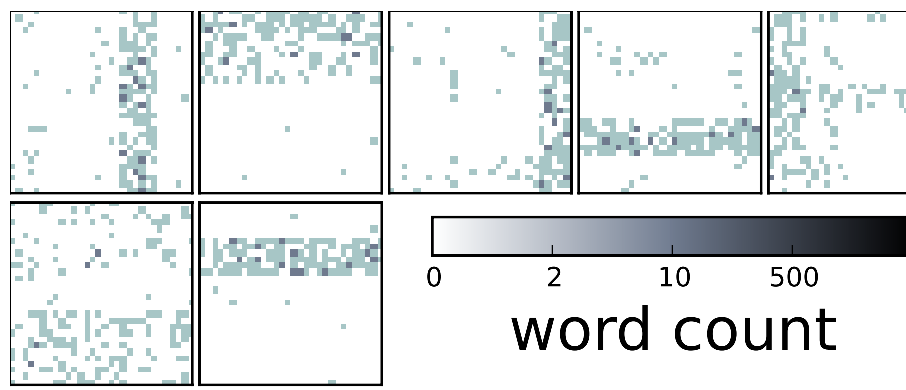




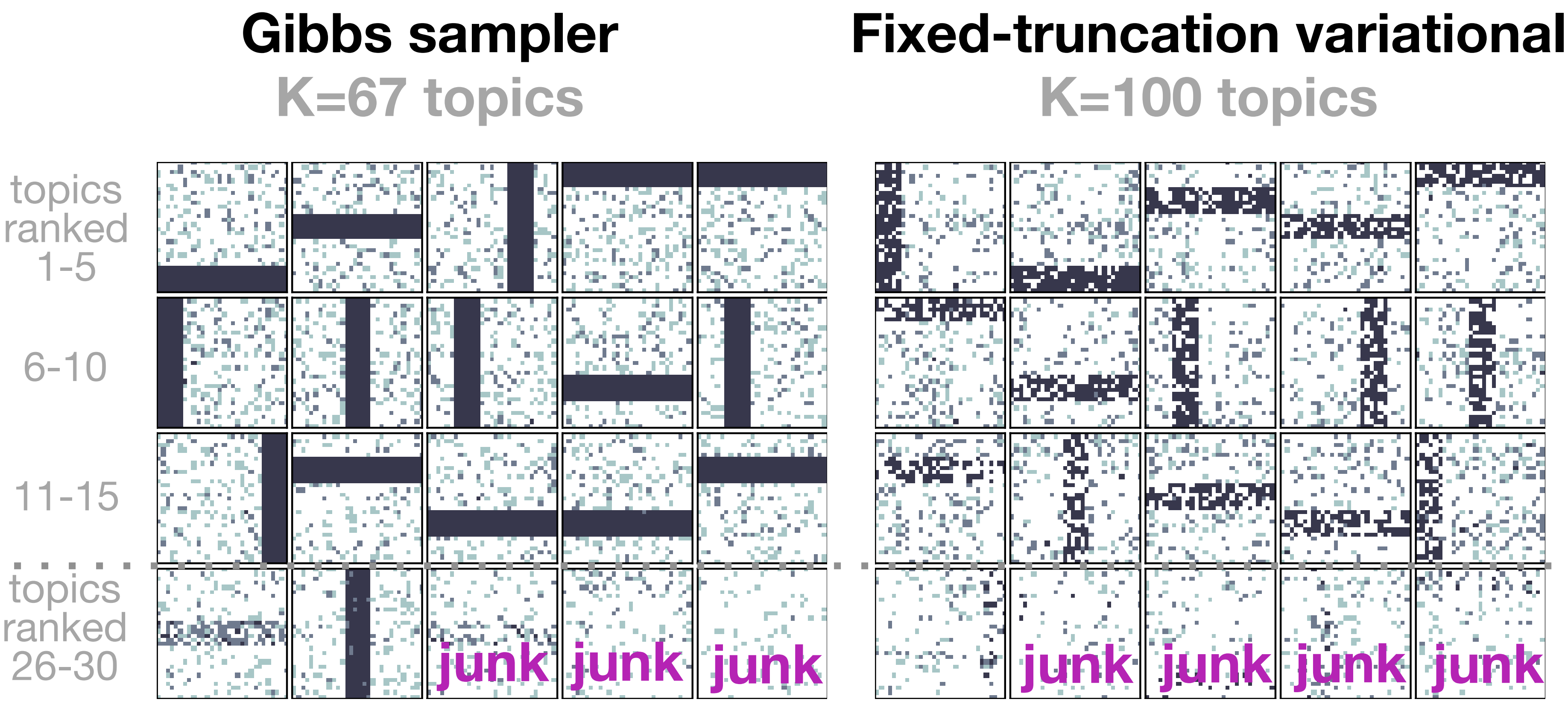


Example Docs:

Can we recover ***10 true topics*** *from 1000 observed documents?*



Toy Dataset: Bar Topics

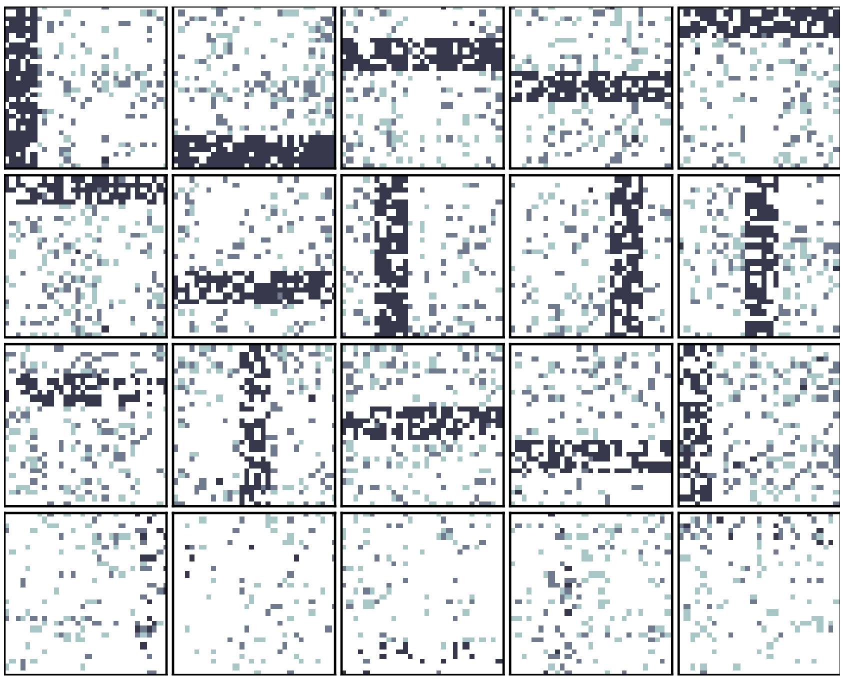


- Both methods produce far too many topics!
- Need **merge and delete moves** to find a compact set.

Toy Dataset: Bar Topics

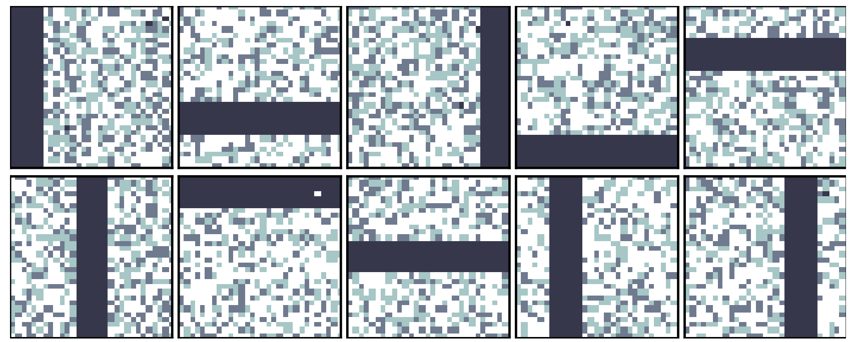
Memoized fixed-truncation

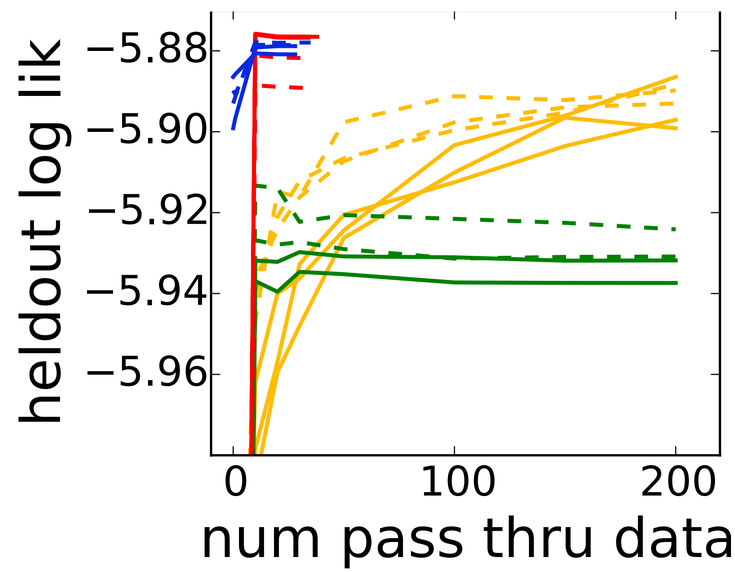
K=100 topics

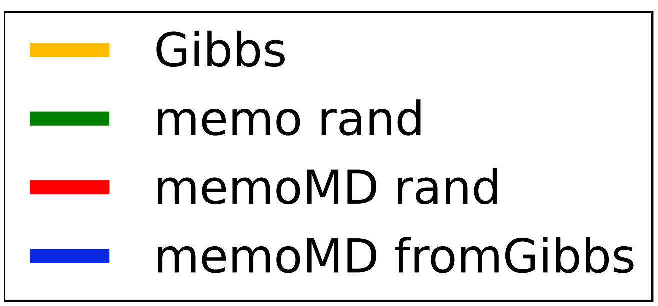


Memoized + merges, deletes

initial K=100 → final K=10







Refining HDP Topic Hypotheses

Accepted Merge

Correlation Score 0.54

1092.4	language	154.7	linguistic
364.4	latin	137.9	linguist
345.5	letter	122.5	language
332.4	dialect	122.4	speech
303.7	speak	103.1	linguistics
296.1	speaker	100.9	grammatical
290.7	sound	75.1	pronunciation
265.4	verb	71.7	suffix

Accepted Merge

Correlation Score 0.79

674.2	series	734.1	film
629.5	song	354.8	magazine
573.5	release	328.0	direct
519.8	star	313.2	production
489.1	television	296.1	actor
388.1	york	281.8	career
385.0	award	269.7	hollywood
371.4	friend	268.2	appeared

Accepted Delete

*Tokens from **deleted** topic **reassigned** to remaining topics, in document-specific fashion.*

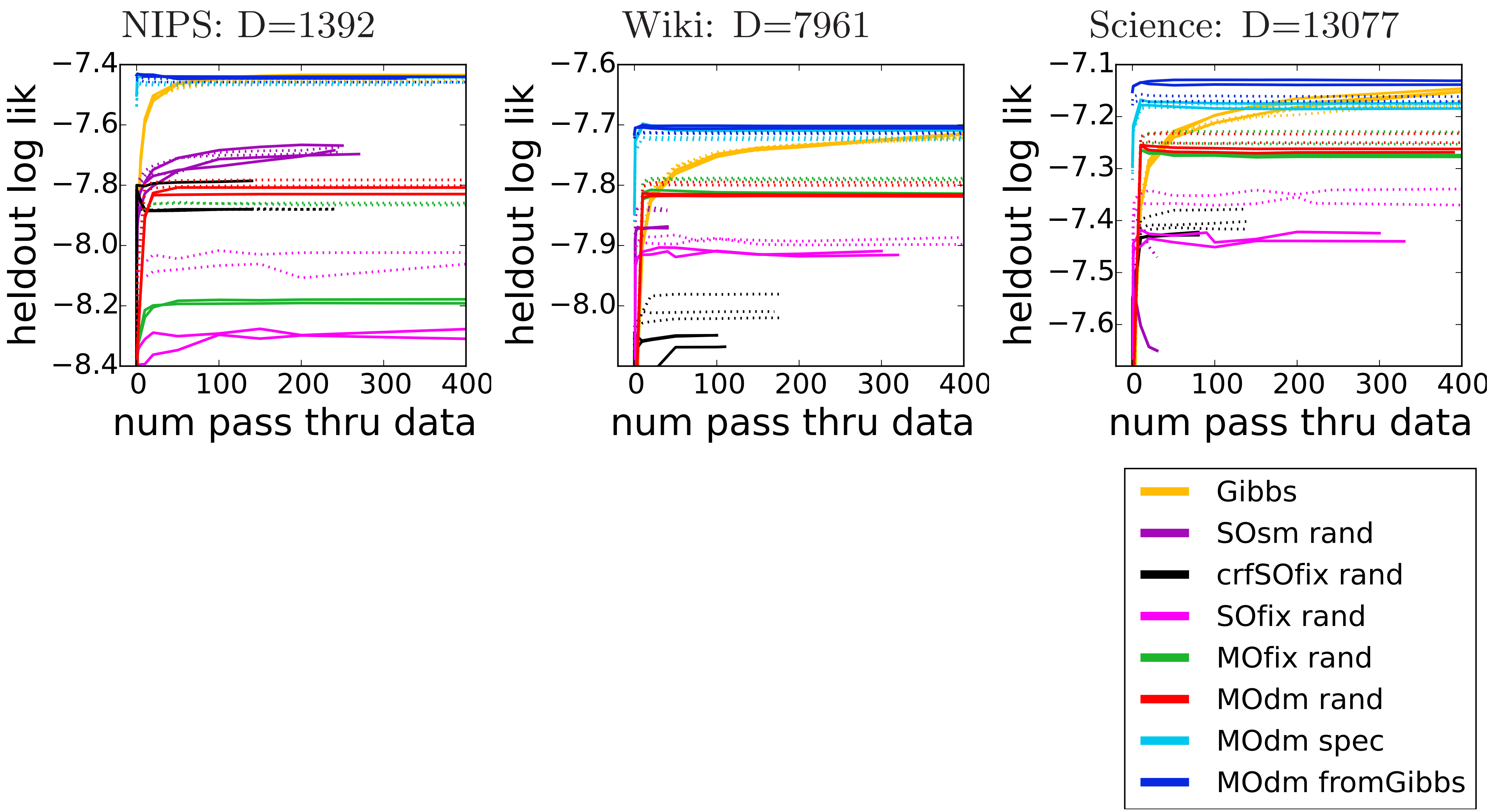
Size: 4611 tokens

100.4	engineering
84.9	science
64.5	computer
53.0	field
50.1	machine
49.8	mechanical
42.9	scientific
42.0	discipline
39.8	analysis
39.3	mathematics

	32682	**21165**	**32612**	**69562**	**58392**
	math	science	code	process	design
	function	theory	language	theory	engine
	theorem	scientific	computer	human	build
	define	mathematics	program	information	speed
	theory	scientist	programming	method	drive
	property	research	machine	approach	reduce
doc A	16.05	42.78	17.56	19.09	7.11
doc B	9.43	40.88	0	20.61	11.29
doc C	0	0	0	35.86	0
doc D	3.77	36.10	30.63	16.70	0

Net change in doc-topic count N_{dk} *after delete*

Analysis of Document Corpora

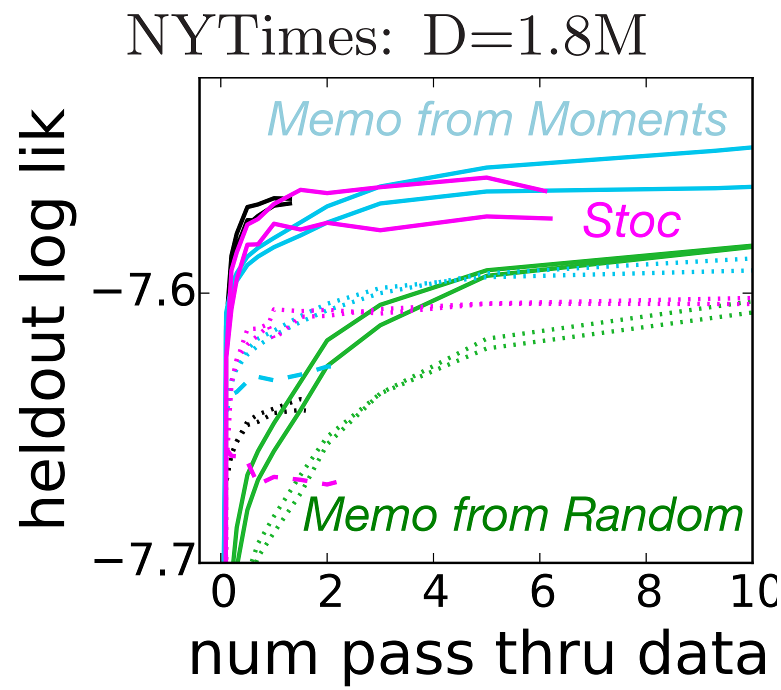


➢ On small-to-medium datasets, match or beat performance of MCMC with orders of magnitude less computation

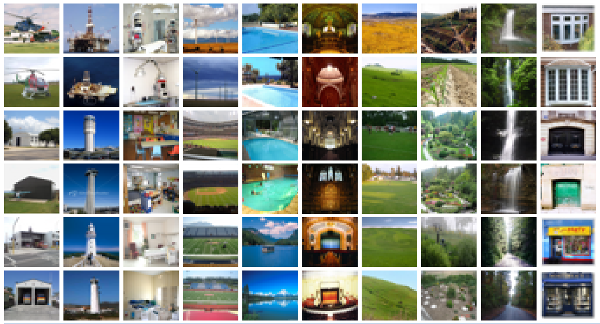
Analysis of Document Corpora

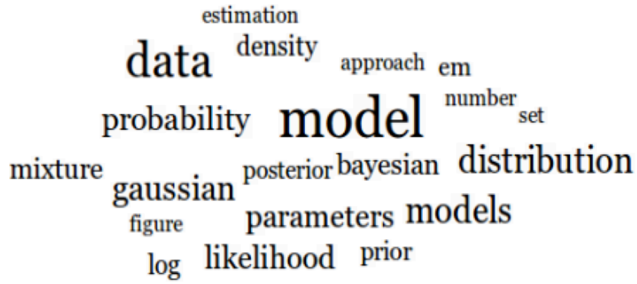
Spectral	+ Variational
0.009 ball	0.018 model
0.008 university	0.013 computer
0.007 says	0.012 models
0.006 science	0.011 problem
0.006 new	0.010 time
0.022 birds	0.019 birds
0.009 new	0.018 evolution
0.009 university	0.016 evolutionary
0.009 says	0.012 species
0.007 years	0.010 molecular
0.017 silicate	0.016 isotopic
0.010 metal	0.013 composition
0.010 high	0.012 ratios
0.009 melt	0.012 isotope
0.007 water	0.012 silicate

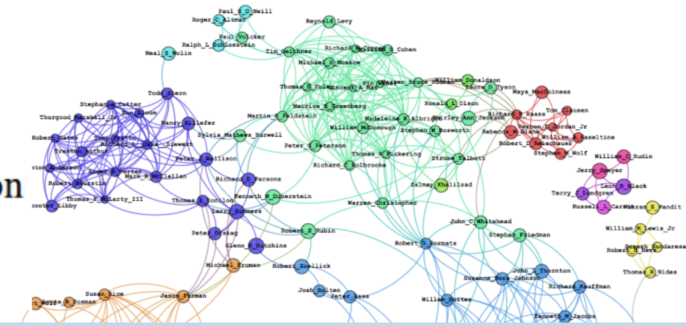
10 passes thru dataset



- Informative moment-based initialization useful (Arora et al. ICML13), but topics evolve in interesting ways.
- On large datasets, continual model improvement over many passes through data. Memoized & stochastic competitive.
- On small-to-medium datasets, match or beat performance of MCMC with orders of magnitude less computation

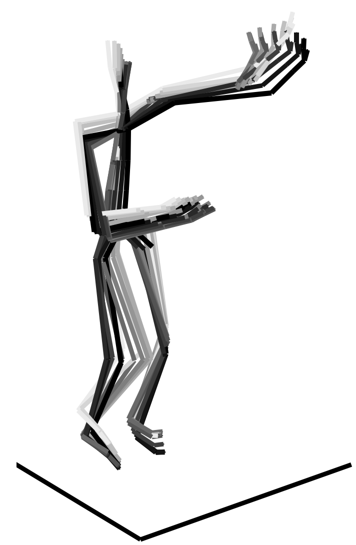






Reliable Variational Learning for Hierarchical Dirichlet Processes

- **Scalable:** Large-scale learning via stochastic or memoized updates
- **Reliable:** Birth-merge recovers structure informed by model & data, not inference algorithm limitations
- **Flexible:** Designed to be broadly applicable: space, time, networks, …



BNPy: Bayesian Nonparametric Learning in Python

Erik Sudderth @ Brown CS: http://cs.brown.edu/~sudderth/