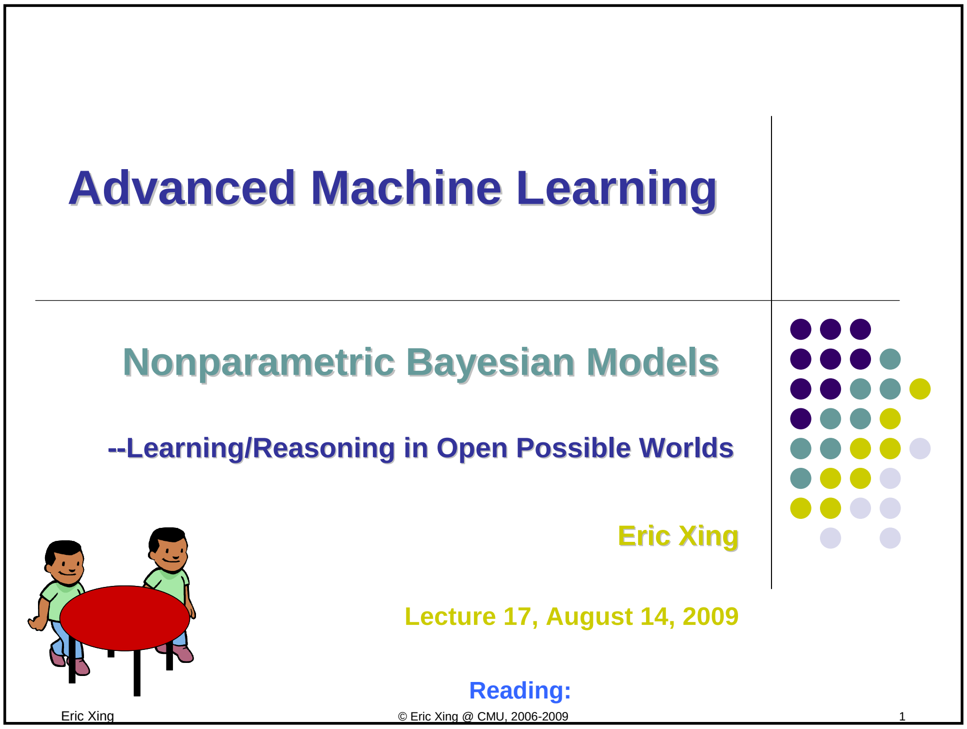


Advanced Machine Learning

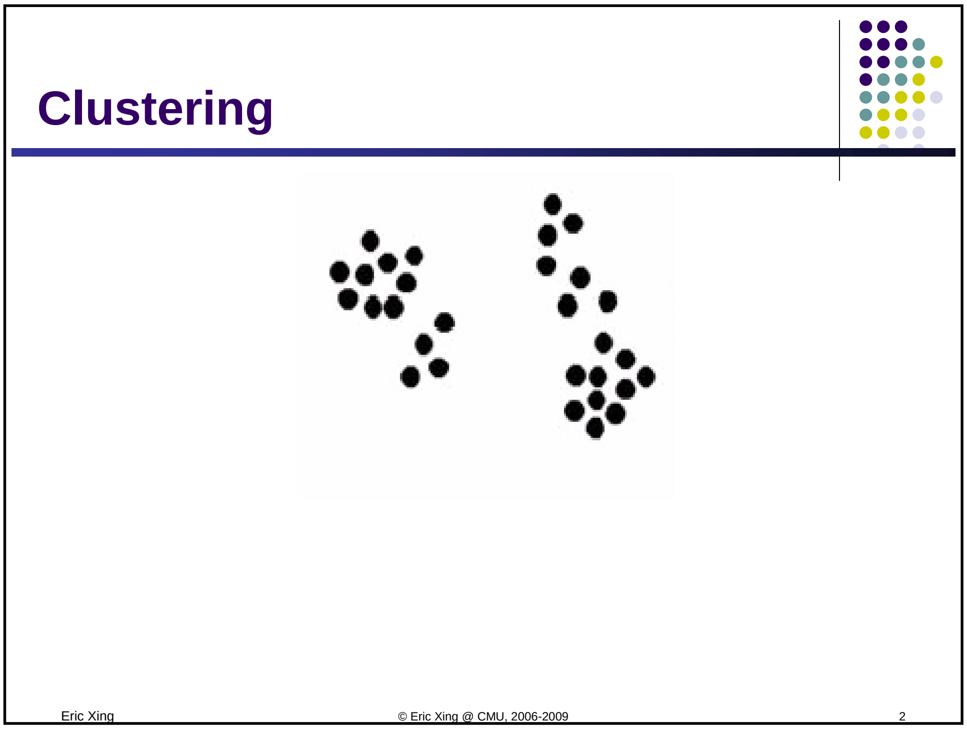
Nonparametric Bayesian Models

--Learning/Reasoning in Open Possible Worlds

Eric Xing

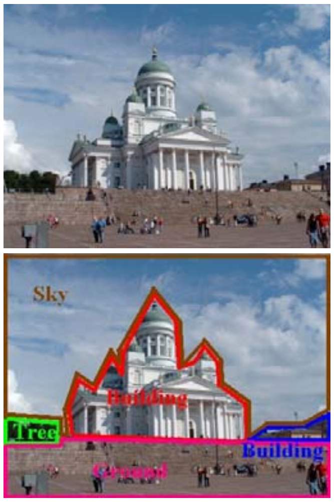
Lecture 17, August 14, 2009

Reading:



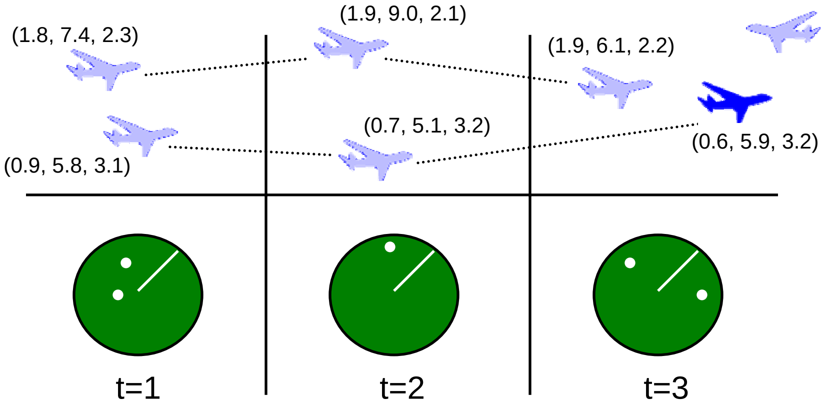
Clustering

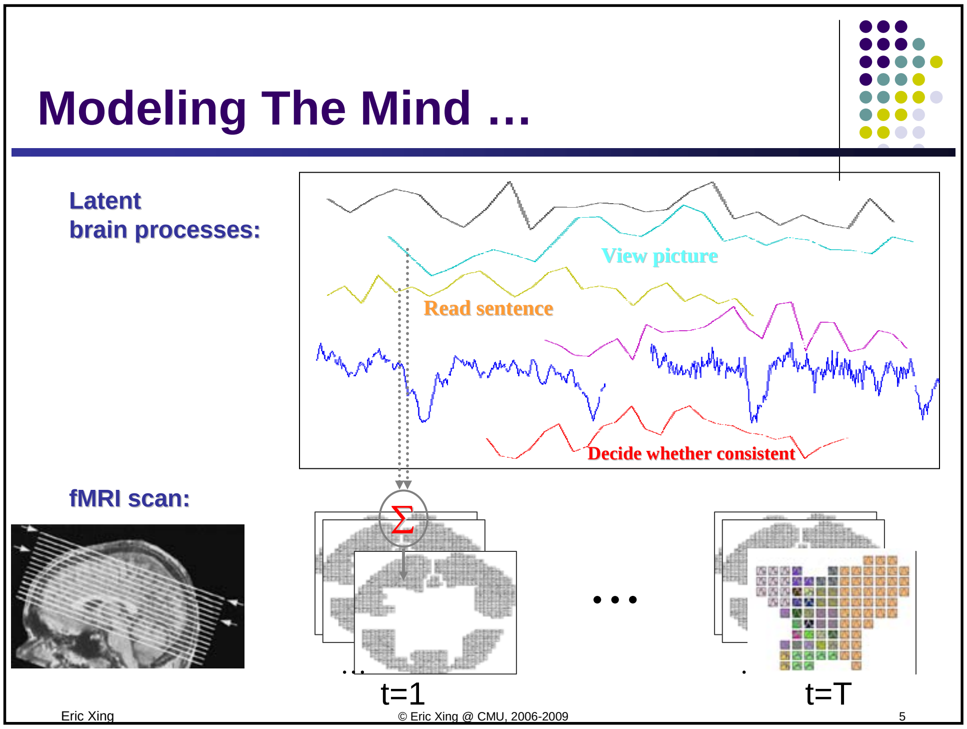
Image Segmentation



- How to segment images?
  - Manual segmentation (very expensive)
  - Algorithm segmentation
    - K-means
    - Statistical mixture models
    - Spectral clustering
- Problems with most existing algorithms
  - Ignore the spatial information
  - Perform the segmentation one image at a time
  - Need to specify the number of segments *a priori*

Object Recognition and Tracking

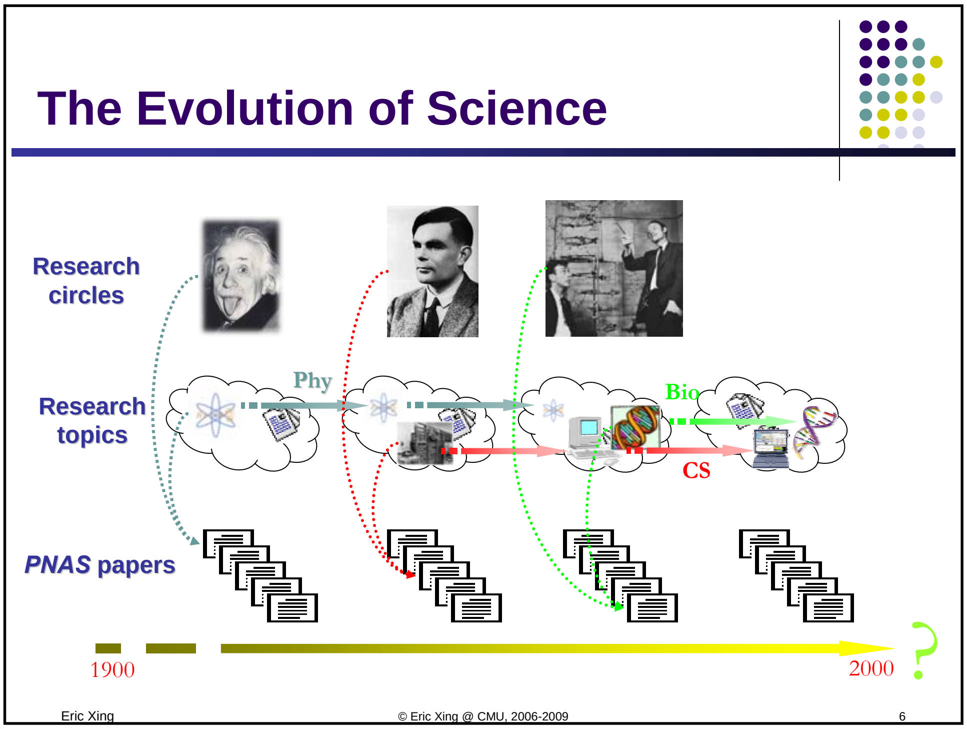




Modeling The Mind …

Latent brain processes:

fMRI scan:

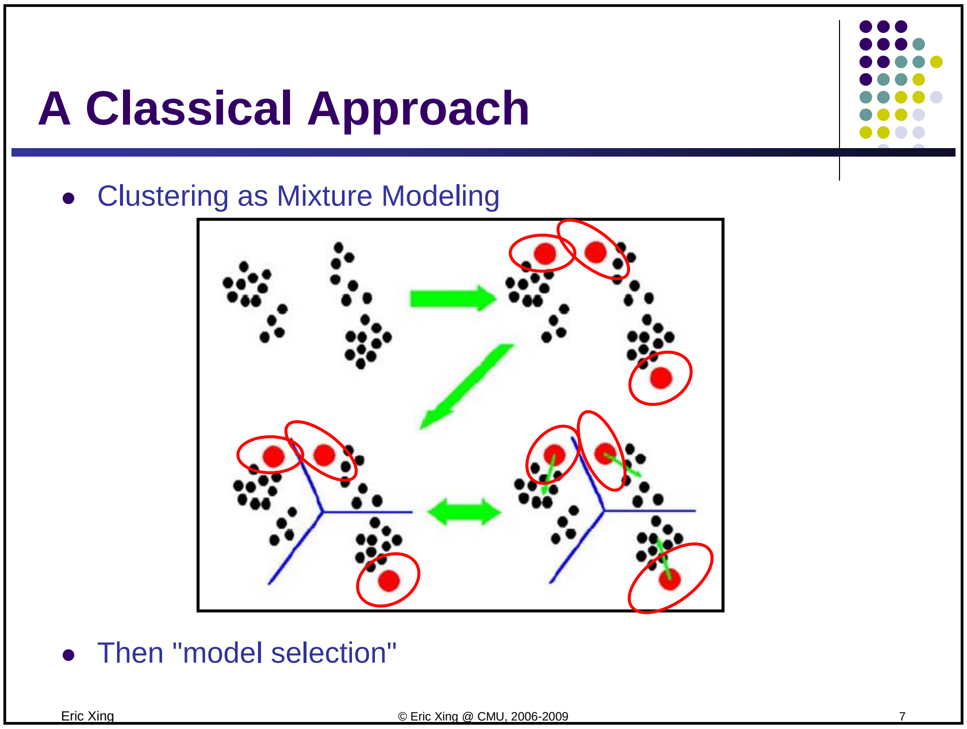


The Evolution of Science

Research circles

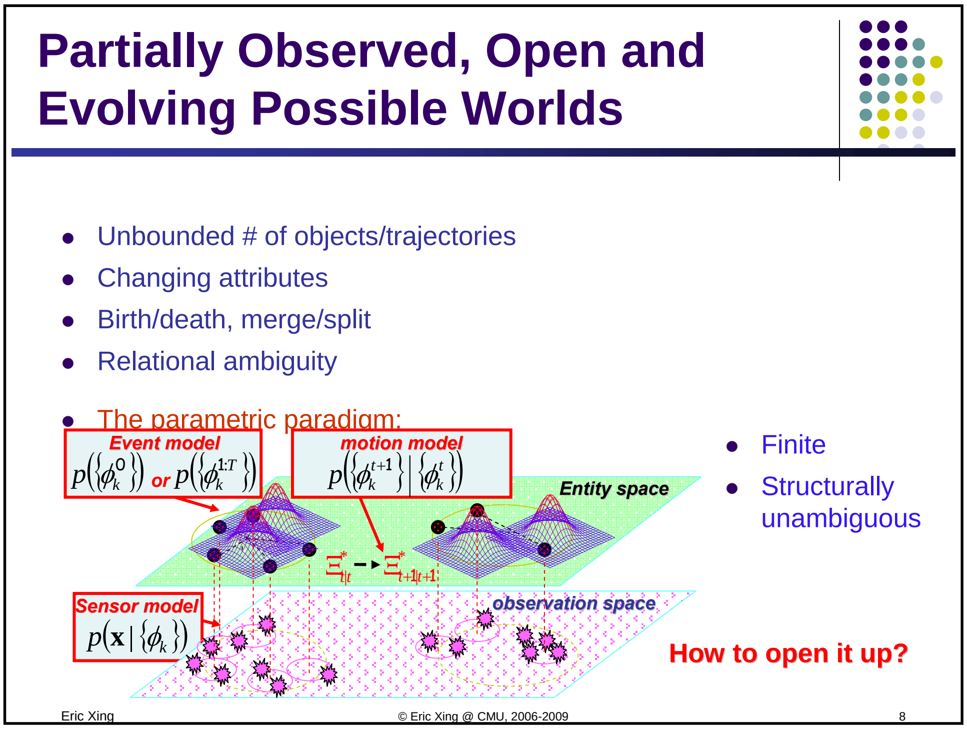
Research topics

PNAS papers



A Classical Approach

- Clustering as Mixture Modeling
- Then "model selection"



Partially Observed, Open and Evolving Possible Worlds

- Unbounded # of objects/trajectories
- Changing attributes
- Birth/death, merge/split
- Relational ambiguity
- The parametric paradigm:

Event model $p\left(\left\{\phi_k^0\right\}\right)$ or $p\left(\left\{\phi_k^{1:T}\right\}\right)$

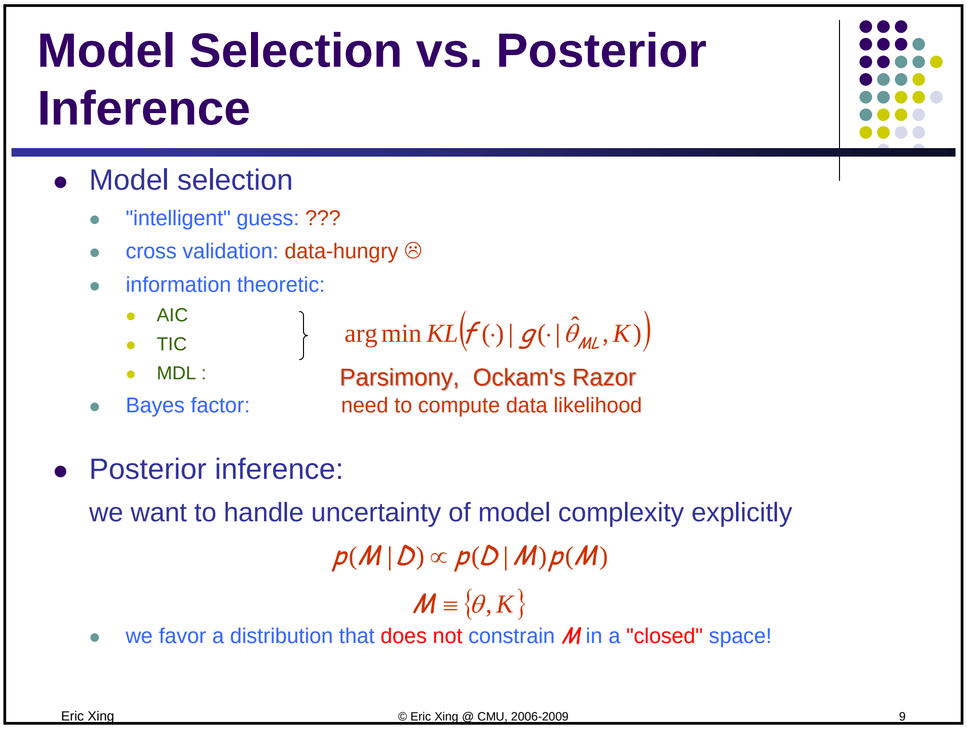
motion model $p\left(\left\{\phi_k^{t+1}\right\} \middle| \left\{\phi_k^t\right\}\right)$

Sensor model $p\left(\mathbf{x} \middle| \left\{\phi_k\right\}\right)$

$\Xi_{t|t}^* \dashrightarrow \Xi_{t+1|t+1}^*$

- Finite
- Structurally unambiguous

How to open it up?



Model Selection vs. Posterior Inference

- Model selection
  - "intelligent" guess: ???
  - cross validation: data-hungry ☹
  - information theoretic:
    - AIC
    - TIC
    - MDL :

    $$\left.\right\} \quad \arg\min KL\left(f(\cdot) \,|\, g(\cdot \,|\, \hat{\theta}_{ML}, K)\right)$$

    Parsimony, Ockam's Razor
  - Bayes factor: need to compute data likelihood

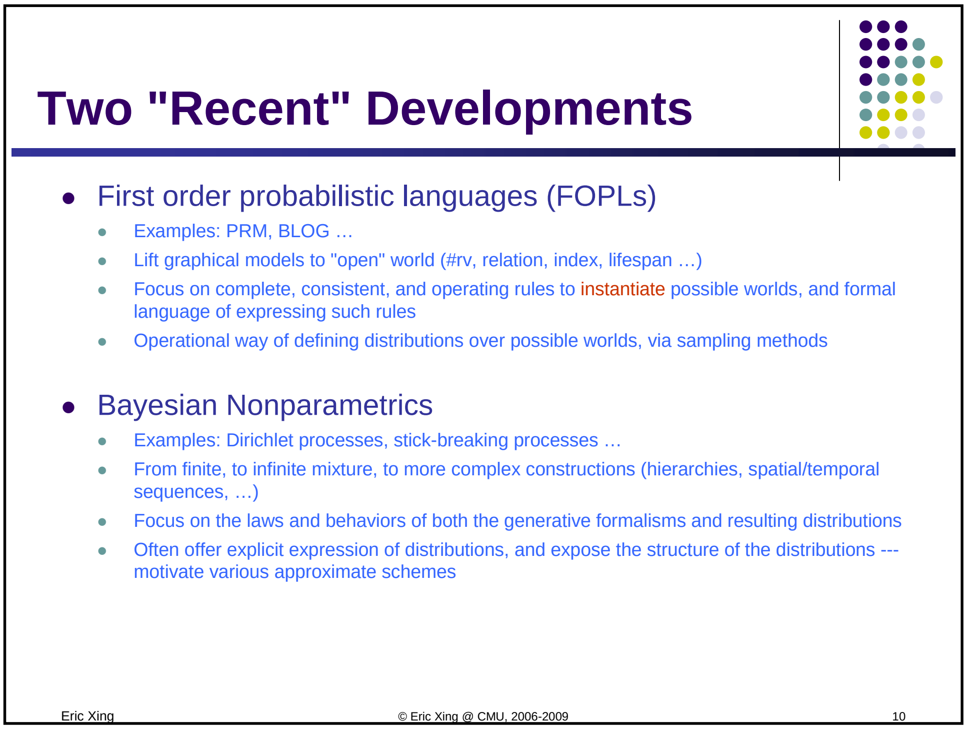
- Posterior inference:

  we want to handle uncertainty of model complexity explicitly

  $$p(M \,|\, D) \propto p(D \,|\, M)\, p(M)$$

  $$M \equiv \{\theta, K\}$$

  - we favor a distribution that does not constrain M in a "closed" space!

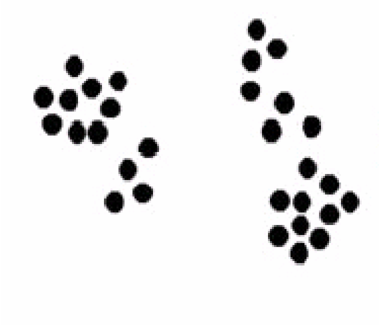


Two "Recent" Developments

- First order probabilistic languages (FOPLs)
  - Examples: PRM, BLOG …
  - Lift graphical models to "open" world (#rv, relation, index, lifespan …)
  - Focus on complete, consistent, and operating rules to instantiate possible worlds, and formal language of expressing such rules
  - Operational way of defining distributions over possible worlds, via sampling methods

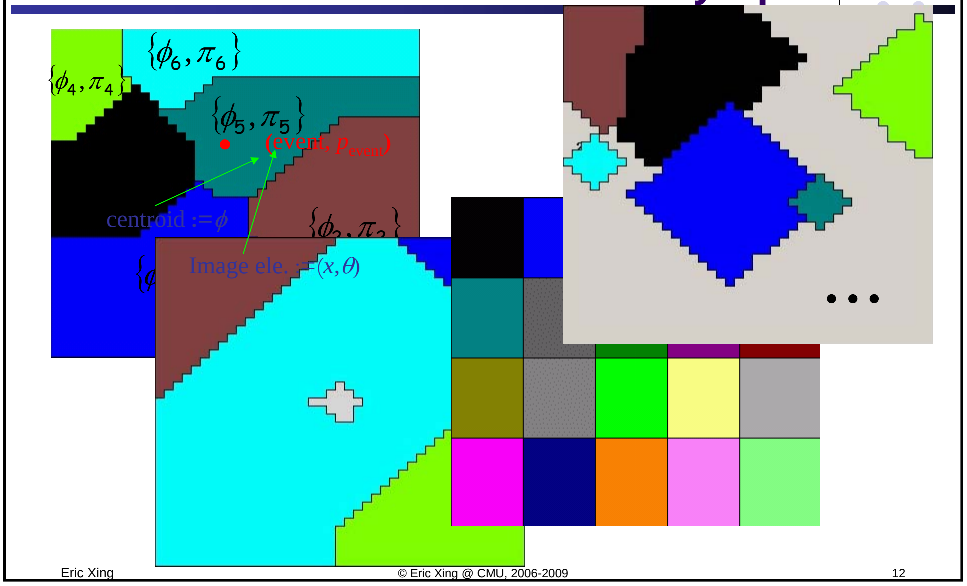
- Bayesian Nonparametrics
  - Examples: Dirichlet processes, stick-breaking processes …
  - From finite, to infinite mixture, to more complex constructions (hierarchies, spatial/temporal sequences, …)
  - Focus on the laws and behaviors of both the generative formalisms and resulting distributions
  - Often offer explicit expression of distributions, and expose the structure of the distributions --- motivate various approximate schemes

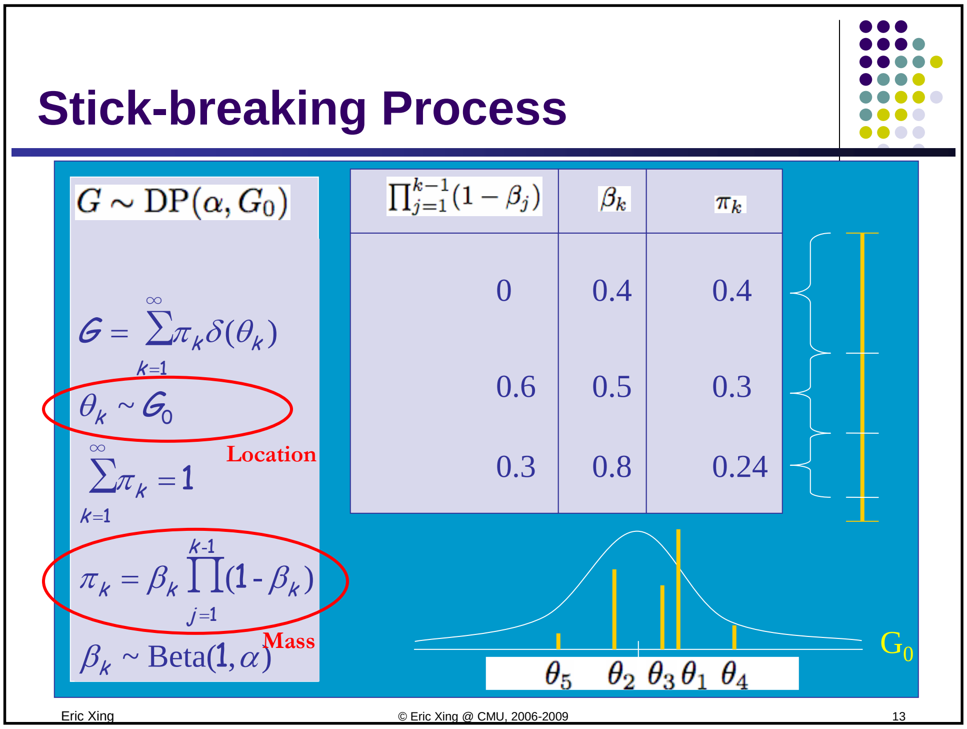
Clustering



- How to label them ?
- How many clusters ???

Random Partition of Probability Space





Stick-breaking Process

$G \sim \mathrm{DP}(\alpha, G_0)$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k)$$

$\theta_k \sim G_0$

Location

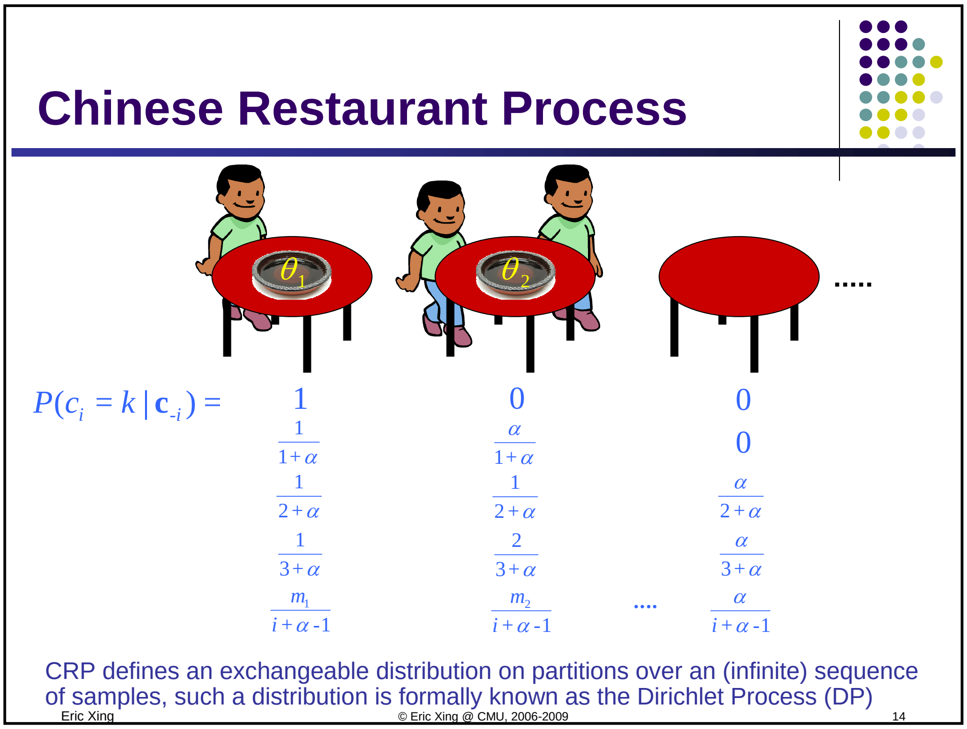
$$\sum_{k=1}^{\infty} \pi_k = 1$$

$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_k)$$

Mass

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

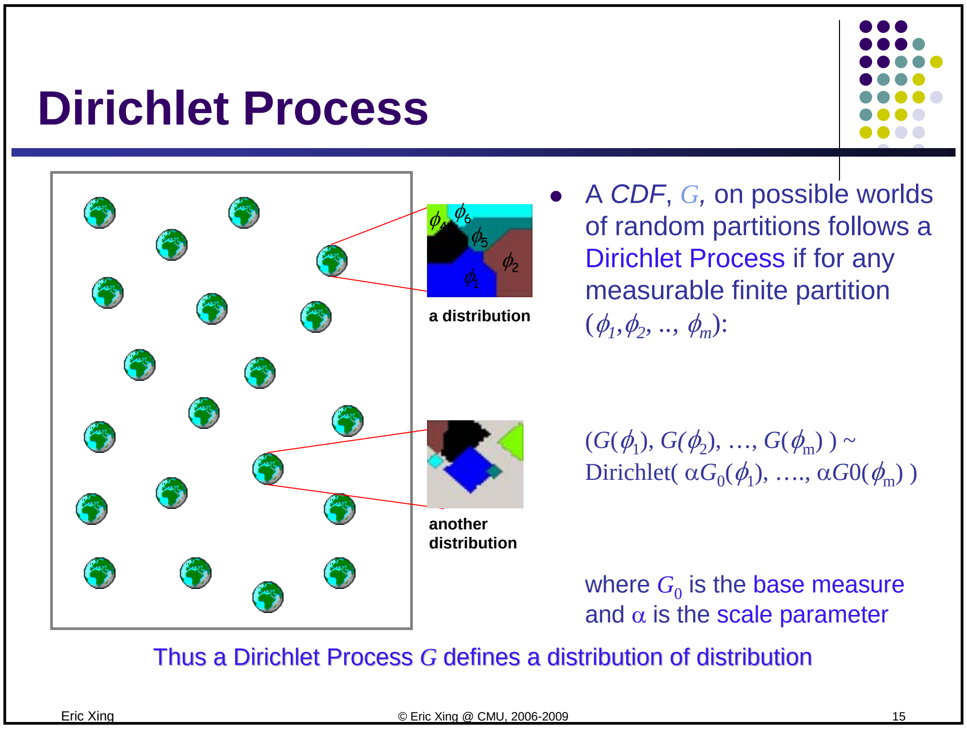
$\prod_{j=1}^{k-1}(1-\beta_j)$	β_k	π_k
0	0.4	0.4
0.6	0.5	0.3
0.3	0.8	0.24



Chinese Restaurant Process

$$P(c_i = k \mid \mathbf{c}_{-i}) = \begin{array}{ccc} 1 & 0 & 0 \\ \frac{1}{1+\alpha} & \frac{\alpha}{1+\alpha} & 0 \\ \frac{1}{2+\alpha} & \frac{1}{2+\alpha} & \frac{\alpha}{2+\alpha} \\ \frac{1}{3+\alpha} & \frac{2}{3+\alpha} & \frac{\alpha}{3+\alpha} \\ \frac{m_1}{i+\alpha-1} & \frac{m_2}{i+\alpha-1} \quad .... & \frac{\alpha}{i+\alpha-1} \end{array}$$

CRP defines an exchangeable distribution on partitions over an (infinite) sequence of samples, such a distribution is formally known as the Dirichlet Process (DP)



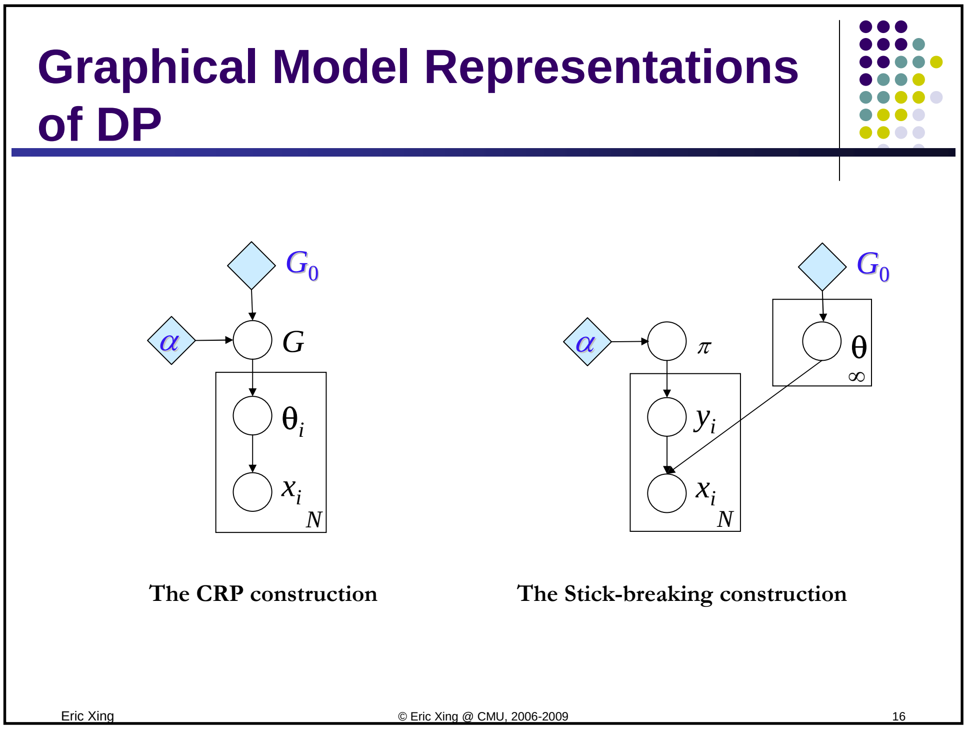
Dirichlet Process

- A *CDF*, G, on possible worlds of random partitions follows a Dirichlet Process if for any measurable finite partition $(\phi_1, \phi_2, .., \phi_m)$:

$(G(\phi_1), G(\phi_2), \ldots, G(\phi_m)) \sim$ Dirichlet($\alpha G_0(\phi_1), \ldots., \alpha G0(\phi_m)$)

where G_0 is the base measure and α is the scale parameter

Thus a Dirichlet Process G defines a distribution of distribution

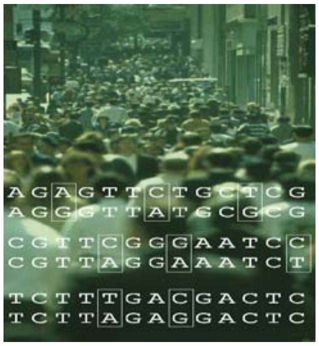


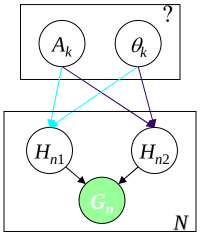
Graphical Model Representations of DP

The CRP construction

The Stick-breaking construction

Ancestral Inference


A G A G T T C T G C T C G
A G G G T T A T G C G C G
C G T T C G G G A A T C C
C G T T A G G A A A T C T
T C T T T G A C G A C T C
T C T T A G A G G A C T C

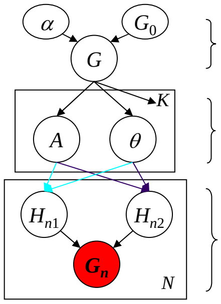


Essentially a clustering problem, but …

- Better recovery of the ancestors leads to better haplotyping results (because of more accurate grouping of common haplotypes)
- True haplotypes are obtainable with high cost, but they can validate model more subjectively (as opposed to examining saliency of clustering)
- Many other biological/scientific utilities

Example: DP-haplotyper [Xing et al, 2004]

- Clustering human populations



DP

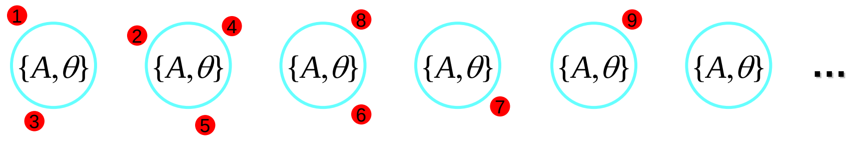
infinite mixture components (for population *haplotypes*)

Likelihood model (for individual haplotypes and *genotypes*)

- Inference: Markov Chain Monte Carlo (MCMC)
  - Gibbs sampling
  - Metropolis Hasting

The DP Mixture of Ancestral Haplotypes

- The customers around a table in CRP form a cluster
  - associate a mixture component (*i.e.*, a population haplotype) with a table
  - sample $\{a, \theta\}$ at each table from a base measure G_0 to obtain the population haplotype and nucleotide substitution frequency for that component



  - With $p(h|\{A, \theta\})$ and $p(g|h_1,h_2)$, the CRP yields a posterior distribution on the number of population haplotypes (and on the haplotype configurations and the nucleotide substitution frequencies)

Inheritance and Observation Models

- Single-locus mutation model

$$A_{C_{i_e}} \rightarrow H_{i_e}$$

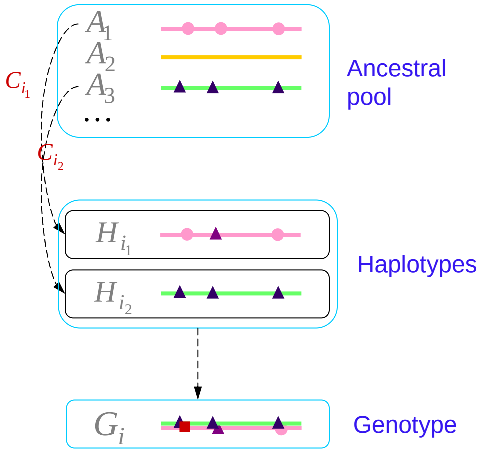
$$P_H(h_t \mid a_t, \theta) = \begin{cases} \theta & \text{for } h_t = a_t \\ \dfrac{1-\theta}{|B|-1} & \text{for } h_t \neq a_t \end{cases}$$

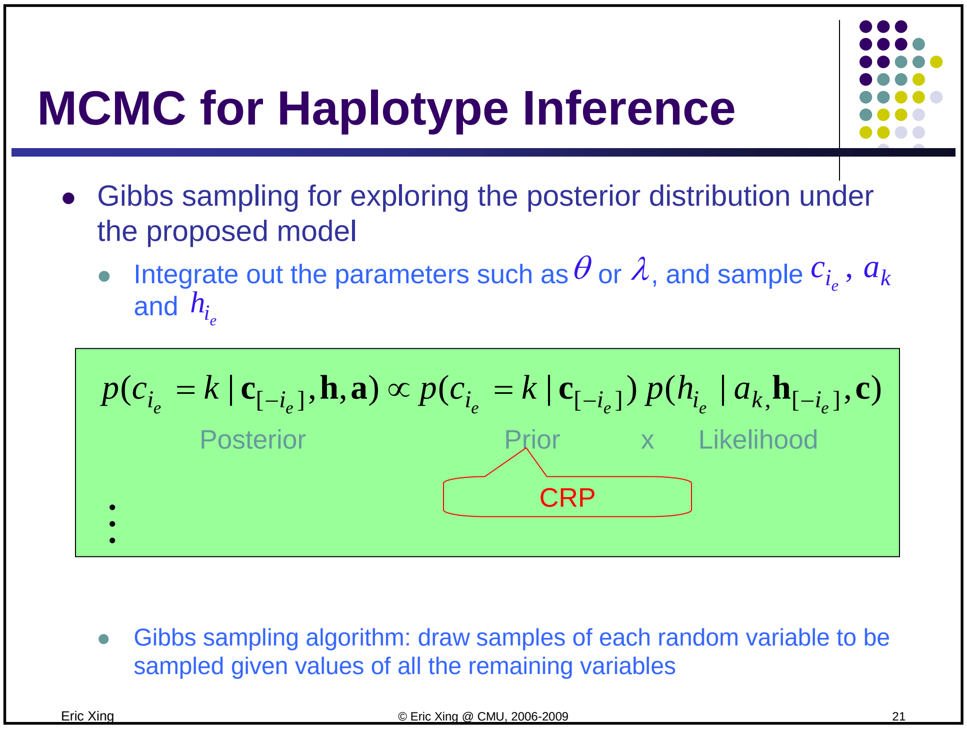
$$\rightarrow h_t = a_t \quad with \quad prob.\ \theta$$

- Noisy observation model

$$H_{i_1}, H_{i_2} \rightarrow G_i$$

$$P_G(g \mid h_1, h_2):$$
$$g_t = h_{1,t} \oplus h_{2,t} \quad with\ \ prob.\ \lambda$$





MCMC for Haplotype Inference

- Gibbs sampling for exploring the posterior distribution under the proposed model
  - Integrate out the parameters such as θ or λ, and sample c_{i_e} , a_k and h_{i_e}

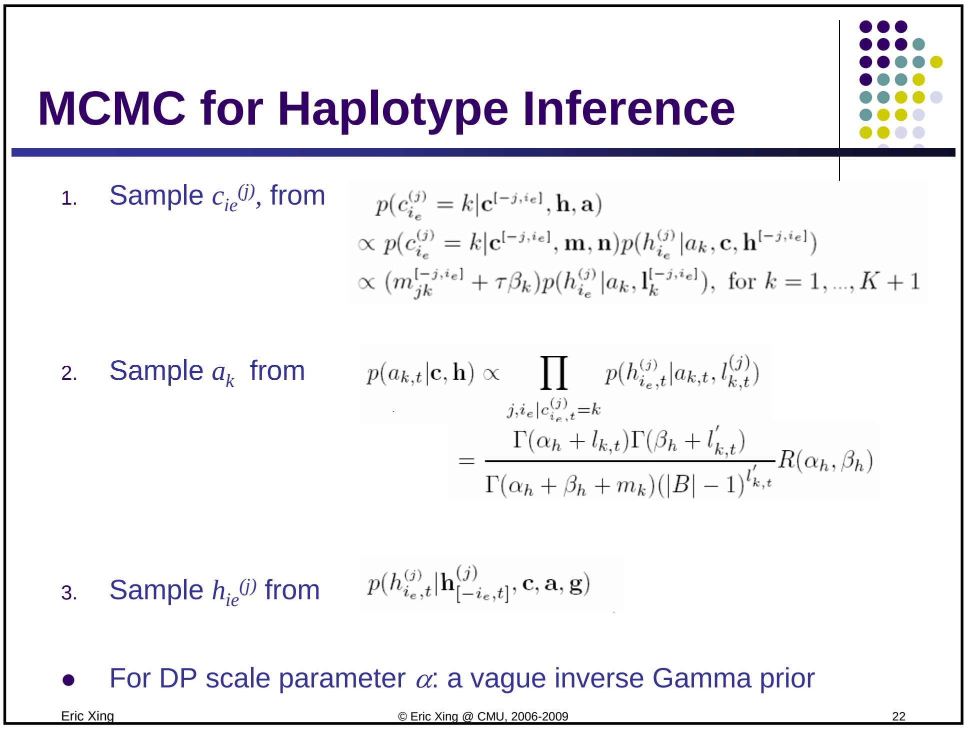
$$p(c_{i_e}=k \mid \mathbf{c}_{[-i_e]},\mathbf{h},\mathbf{a}) \propto p(c_{i_e}=k \mid \mathbf{c}_{[-i_e]})\; p(h_{i_e} \mid a_{k,}\mathbf{h}_{[-i_e]},\mathbf{c})$$

Posterior Prior x Likelihood

CRP

$\vdots$

  - Gibbs sampling algorithm: draw samples of each random variable to be sampled given values of all the remaining variables



MCMC for Haplotype Inference

1. Sample $c_{ie}^{(j)}$, from

$$p(c_{i_e}^{(j)}=k|\mathbf{c}^{[-j,i_e]},\mathbf{h},\mathbf{a})$$
$$\propto p(c_{i_e}^{(j)}=k|\mathbf{c}^{[-j,i_e]},\mathbf{m},\mathbf{n})p(h_{i_e}^{(j)}|a_k,\mathbf{c},\mathbf{h}^{[-j,i_e]})$$
$$\propto (m_{jk}^{[-j,i_e]}+\tau\beta_k)p(h_{i_e}^{(j)}|a_k,\mathbf{l}_k^{[-j,i_e]}),\ \text{for } k=1,...,K+1$$

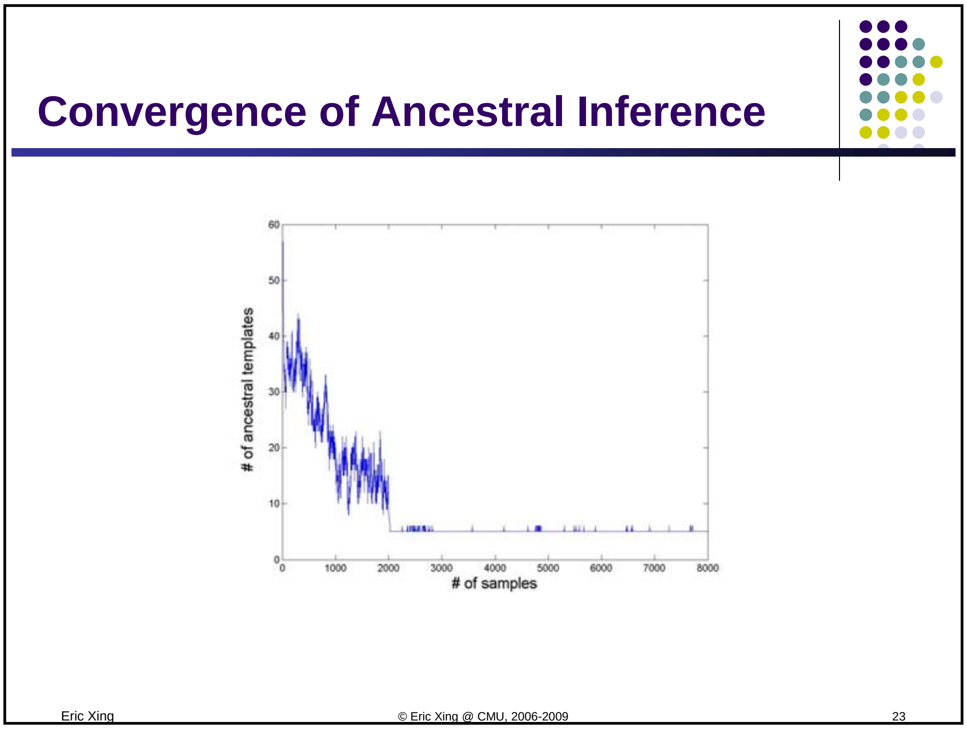
2. Sample a_k from

$$p(a_{k,t}|\mathbf{c},\mathbf{h}) \propto \prod_{j,i_e|c_{i_e,t}^{(j)}=k} p(h_{i_e,t}^{(j)}|a_{k,t},l_{k,t}^{(j)})$$
$$=\frac{\Gamma(\alpha_h+l_{k,t})\Gamma(\beta_h+l'_{k,t})}{\Gamma(\alpha_h+\beta_h+m_k)(|B|-1)^{l'_{k,t}}}R(\alpha_h,\beta_h)$$

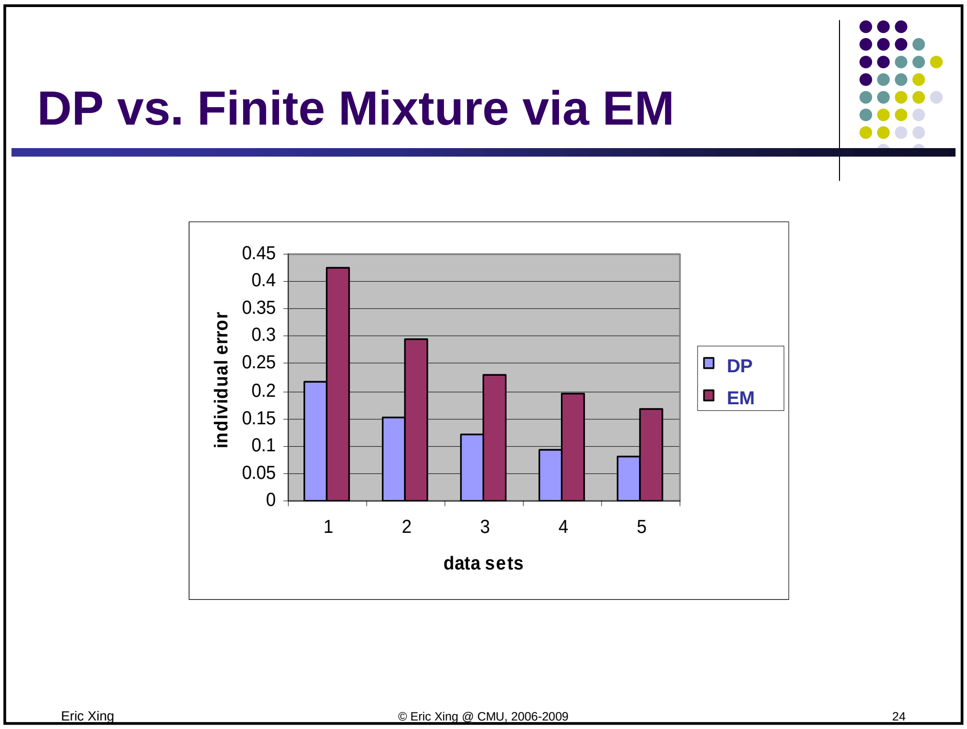
3. Sample $h_{ie}^{(j)}$ from

$$p(h_{i_e,t}^{(j)}|\mathbf{h}_{[-i_e,t]}^{(j)},\mathbf{c},\mathbf{a},\mathbf{g})$$

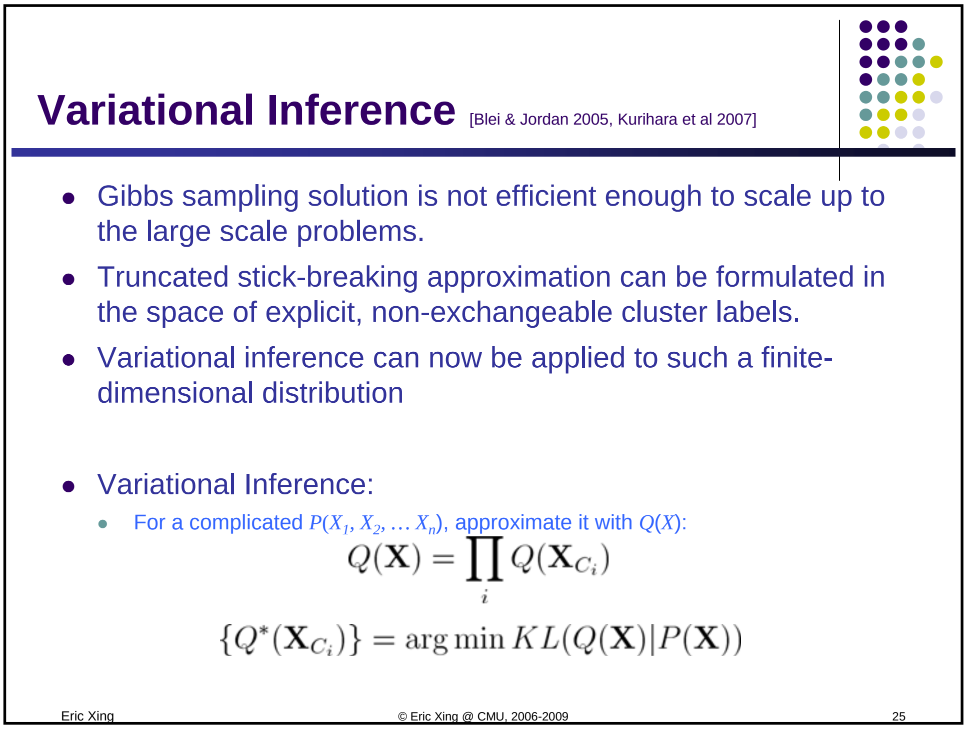
- For DP scale parameter α: a vague inverse Gamma prior



Convergence of Ancestral Inference



DP vs. Finite Mixture via EM



Variational Inference [Blei & Jordan 2005, Kurihara et al 2007]

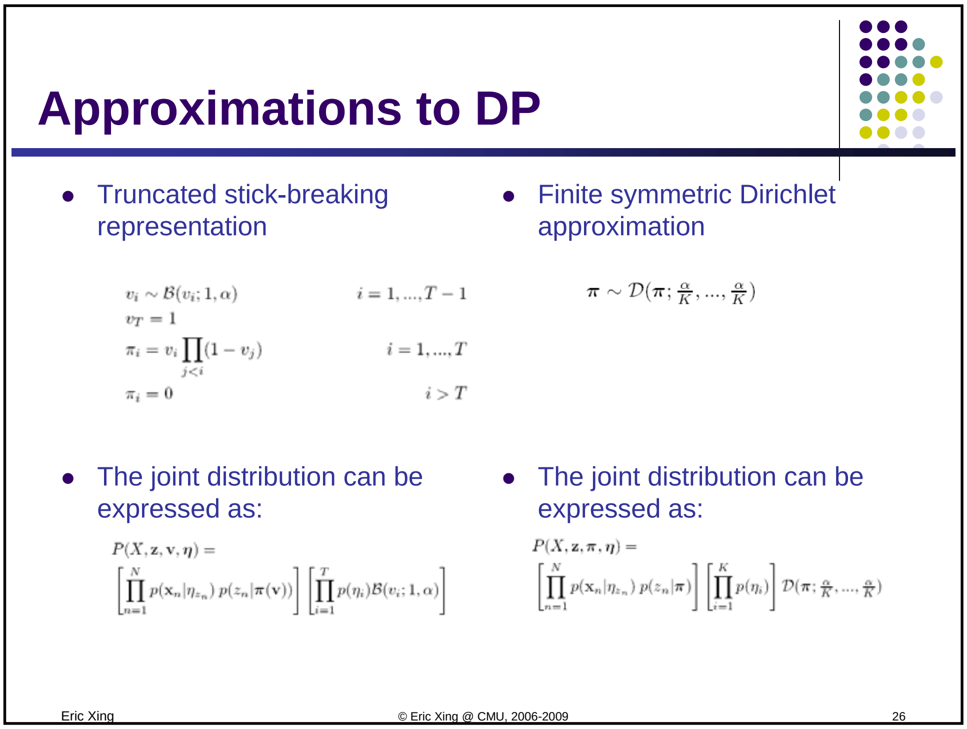
- Gibbs sampling solution is not efficient enough to scale up to the large scale problems.
- Truncated stick-breaking approximation can be formulated in the space of explicit, non-exchangeable cluster labels.
- Variational inference can now be applied to such a finite-dimensional distribution

- Variational Inference:
  - For a complicated $P(X_1, X_2, \dots X_n)$, approximate it with $Q(X)$:

$$Q(\mathbf{X}) = \prod_i Q(\mathbf{X}_{C_i})$$

$$\{Q^*(\mathbf{X}_{C_i})\} = \arg\min KL(Q(\mathbf{X})|P(\mathbf{X}))$$

Eric Xing © Eric Xing @ CMU, 2006-2009 25



Approximations to DP

- Truncated stick-breaking representation

$$\begin{aligned} &v_i \sim \mathcal{B}(v_i; 1, \alpha) && i = 1, \dots, T-1 \\ &v_T = 1 && \\ &\pi_i = v_i \prod_{j<i}(1 - v_j) && i = 1, \dots, T \\ &\pi_i = 0 && i > T \end{aligned}$$

- The joint distribution can be expressed as:

$$P(X, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi}(\mathbf{v}))\right] \left[\prod_{i=1}^{T} p(\eta_i)\mathcal{B}(v_i; 1, \alpha)\right]$$

- Finite symmetric Dirichlet approximation

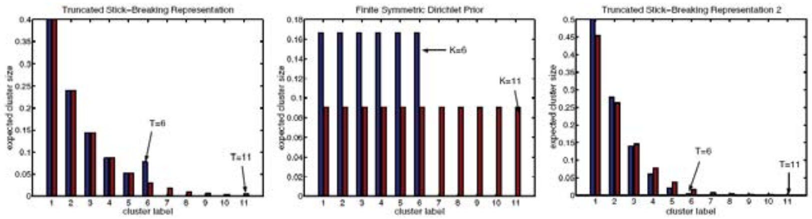
$$\pi \sim \mathcal{D}(\pi; \tfrac{\alpha}{K}, \dots, \tfrac{\alpha}{K})$$

- The joint distribution can be expressed as:

$$P(X, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi})\right] \left[\prod_{i=1}^{K} p(\eta_i)\right] \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, \dots, \tfrac{\alpha}{K})$$

Eric Xing © Eric Xing @ CMU, 2006-2009 26

TDP vs. TSB



- TDP is size biased
- cluster labels is NOT interchangeable under TDP but is interchangeable under TSB

Marginalization

$$P(X, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\, p(z_n|\boldsymbol{\pi})\right] \left[\prod_{i=1}^{K} p(\eta_i)\right] \mathcal{D}(\boldsymbol{\pi}; \tfrac{\alpha}{K}, \ldots, \tfrac{\alpha}{K})$$

- In variational Bayesian approximation, we assume a factorized form for the posterior distribution.
- However it is not a good assumption since changes in ***π*** will have a considerable impact on ***z***.

If we can integrate out π , the joint distribution is given by

$$P(X, \mathbf{z}, \boldsymbol{\eta}) = \left[\prod_{n=1}^{N} p(\mathbf{x}_n|\eta_{z_n})\right] p(\mathbf{z}) \left[\prod_{i=1}^{\infty} p(\eta_i)\right]$$

For the TSB representation:

$$p_{\mathrm{TSB}}(\mathbf{z}) = \prod_{i<T} \frac{\Gamma(1+N_i)\Gamma(\alpha+N_{>i})}{\Gamma(1+\alpha+N_{\geq i})} \ \alpha$$

For the FSD representation:

$$p_{\mathrm{FSD}}(\mathbf{z}) = \frac{\Gamma(\alpha)\prod_{k=1}^{K}\Gamma(N_k+\frac{\alpha}{K})}{\Gamma(N+\alpha)\Gamma(\frac{\alpha}{K})^K}$$

VB inference

- We can then apply the VB inference on the four approximations

$$\{Q^*(\mathbf{X}_{C_i})\} = \arg\min KL(Q(\mathbf{X})|P(\mathbf{X}))$$

The approximated posterior distribution for TSB and FSD are

$$Q_{\mathtt{TSB}}(\mathbf{z}, \boldsymbol{\eta}, \mathbf{v}) = \left[\prod_{n}^{N} q(z_n)\right]\left[\prod_{i=1}^{T} q(\eta_i) q(v_i)\right] \qquad Q_{\mathtt{FSD}}(\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\pi}) = \left[\prod_{n}^{N} q(z_n)\right]\left[\prod_{k=1}^{K} q(\eta_k)\right] q(\boldsymbol{\pi})$$

Depending on marginalization or not, ***v*** and π may be integrated out.

Experimental results

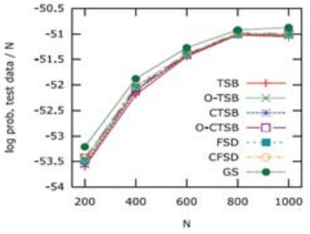


Figure 2: Average log probability per data-point for test data as a function of N.

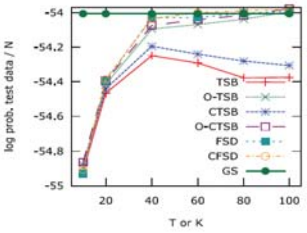
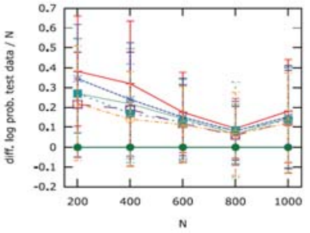
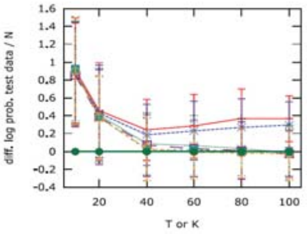
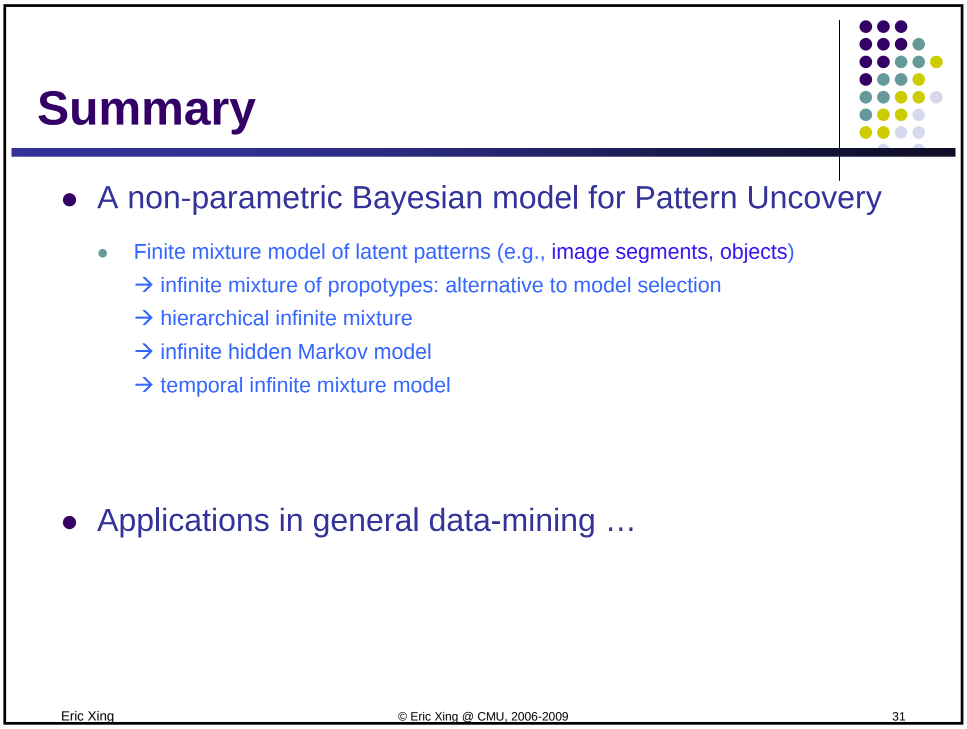


Figure 4: Average log probability per data-point for test data as a function of T (for TSB methods) or K (for FSD methods).







Summary

- A non-parametric Bayesian model for Pattern Uncovery
  - Finite mixture model of latent patterns (e.g., image segments, objects)
    → infinite mixture of propotypes: alternative to model selection
    → hierarchical infinite mixture
    → infinite hidden Markov model
    → temporal infinite mixture model

- Applications in general data-mining …