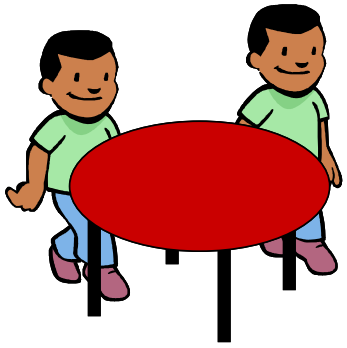
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Dirichlet Process and Hiarchical DP

-- case studies in genetics and text analysis

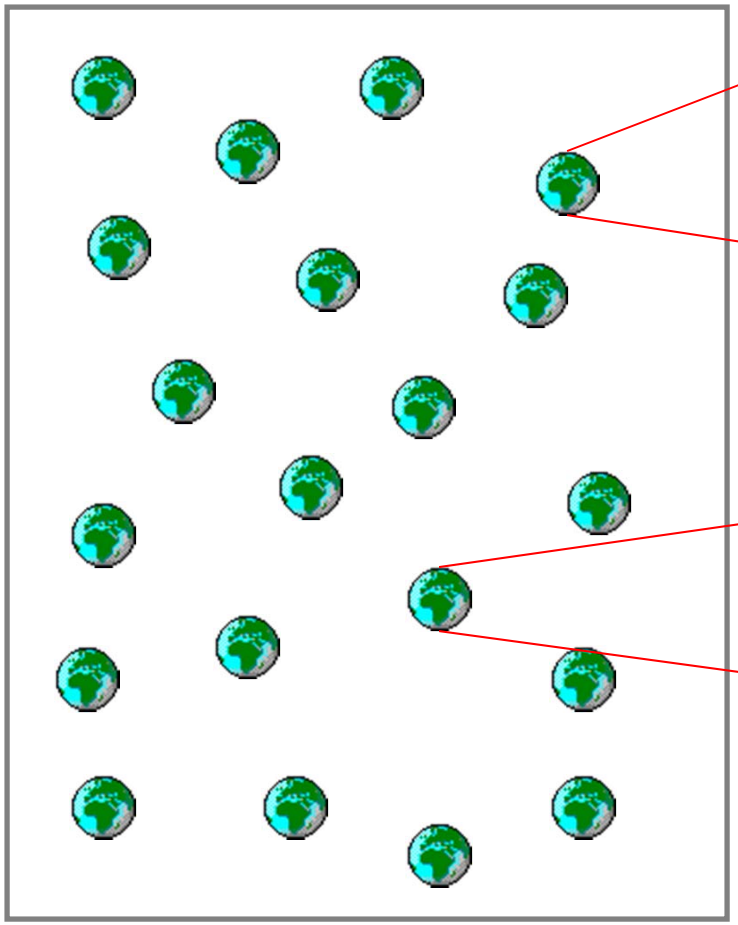


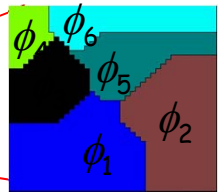
Eric Xing

Lecture 20, March 31, 2014

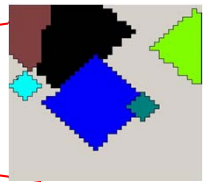
Reading:

Dirichlet Process





a distribution



another distribution

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

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

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

Thus a Dirichlet Process G defines a distribution of distribution

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

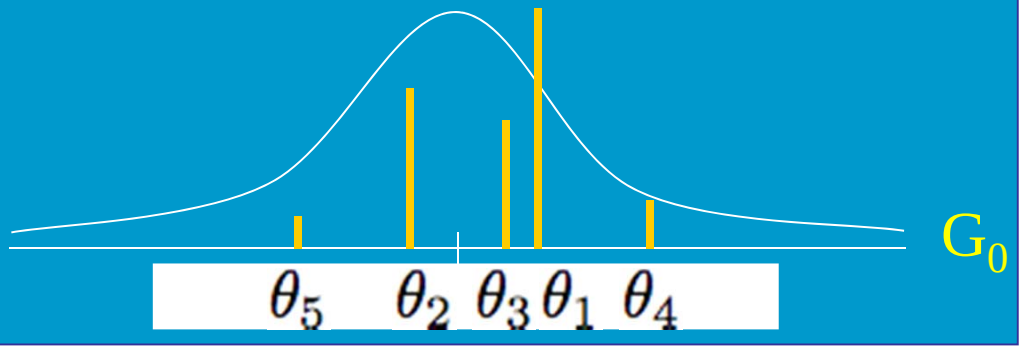
$$\sum_{k=1}^{\infty} \pi_k = 1$$

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

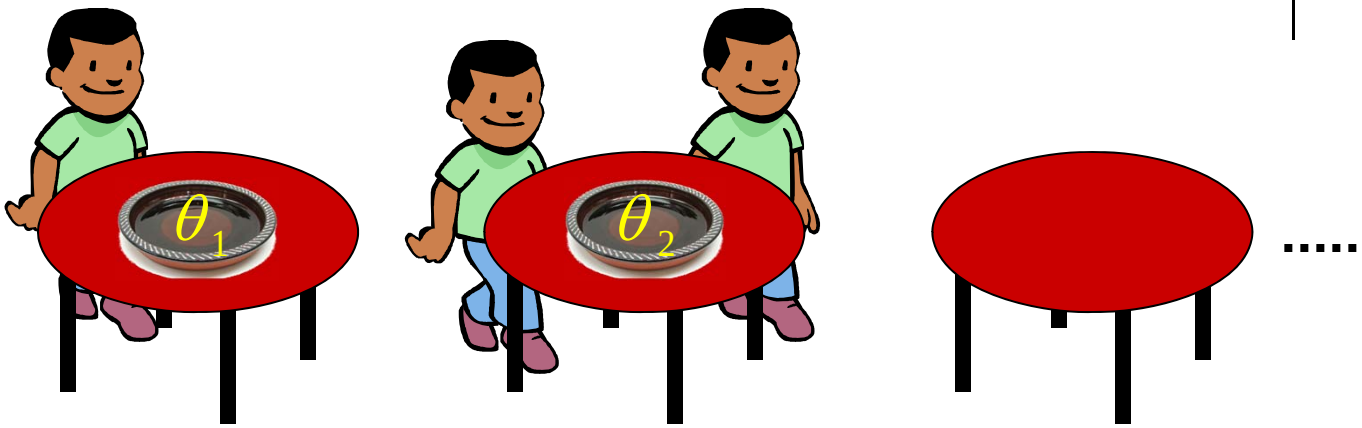
Mass

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

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



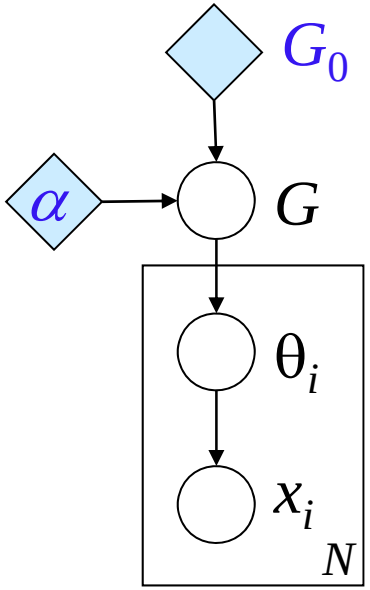
Chinese Restaurant Process



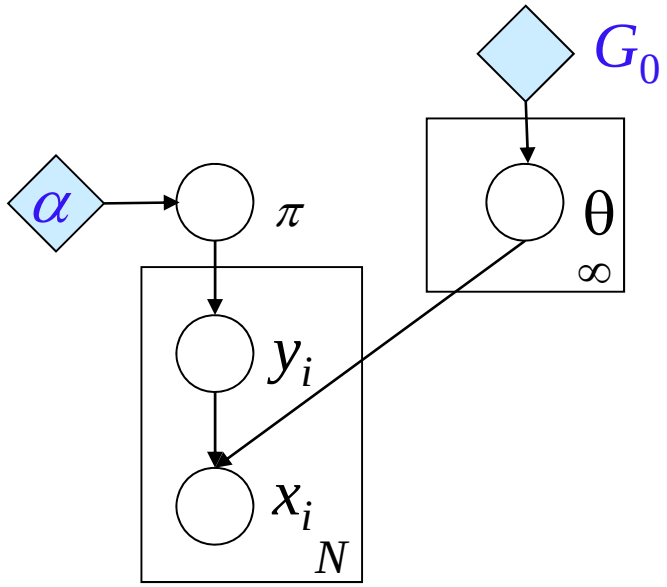
$$P(c_i = k \mid \mathbf{c}_{-i}) = \begin{matrix} 1 \\ \frac{1}{1+\alpha} \\ \frac{1}{2+\alpha} \\ \frac{1}{3+\alpha} \\ \frac{m_1}{i+\alpha-1} \end{matrix} \quad \begin{matrix} 0 \\ \frac{\alpha}{1+\alpha} \\ \frac{1}{2+\alpha} \\ \frac{2}{3+\alpha} \\ \frac{m_2}{i+\alpha-1} \end{matrix} \quad \begin{matrix}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

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

Graphical Model Representations of DP mixture

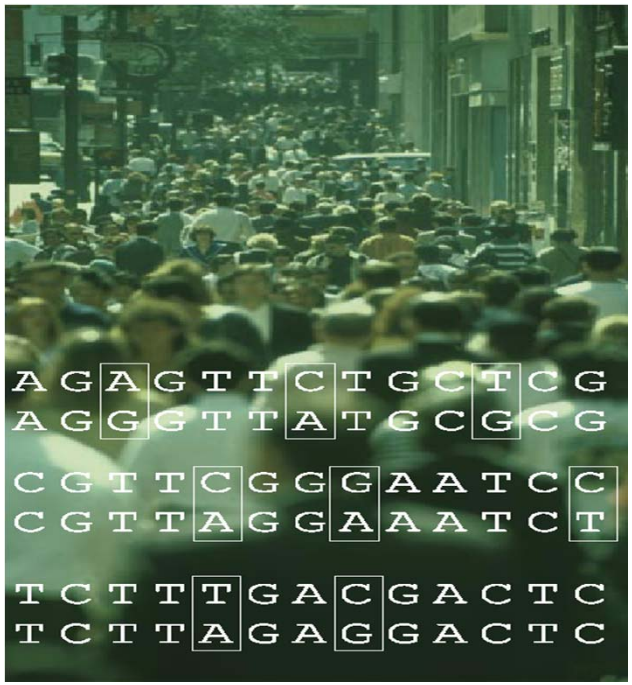


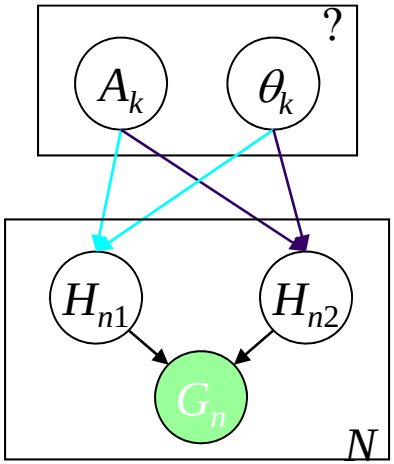
The CRP construction



The Stick-breaking construction

Case I: Ancestral Inference

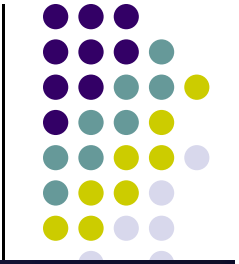




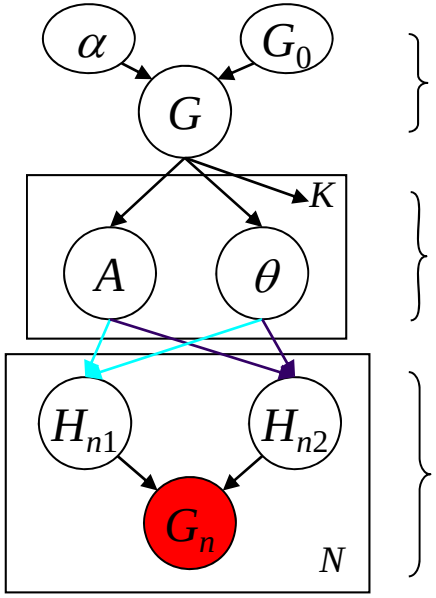
Essentially a clustering problem, but …

- 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



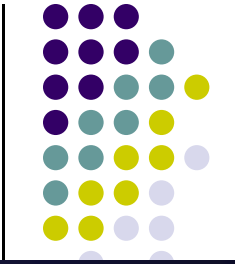
DP

infinite mixture components
(for population *haplotypes*)

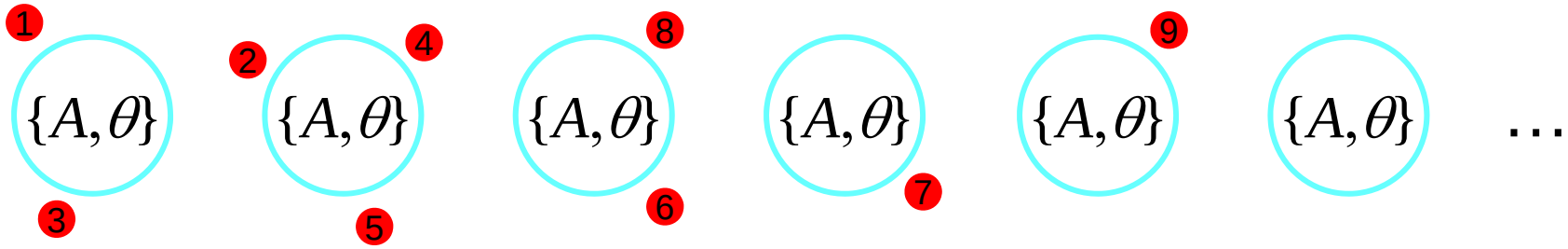
Likelihood model
(for individual
haplotypes and *genotypes*)

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

The DP Mixture of Ancestral Haplotypes

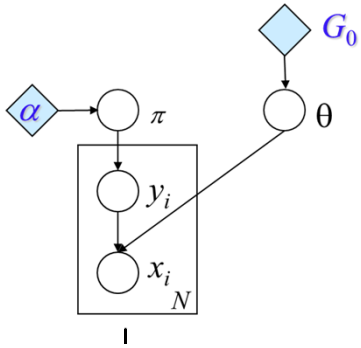


- The customers around a table in CRP form a cluster
  - associate a mixture component (*i.e.*, a population haplotype) with a table
  - sample $\{a, \theta\}$ at each table from a base measure G_0 to obtain the population haplotype and nucleotide substitution frequency for that component



  - 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

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

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

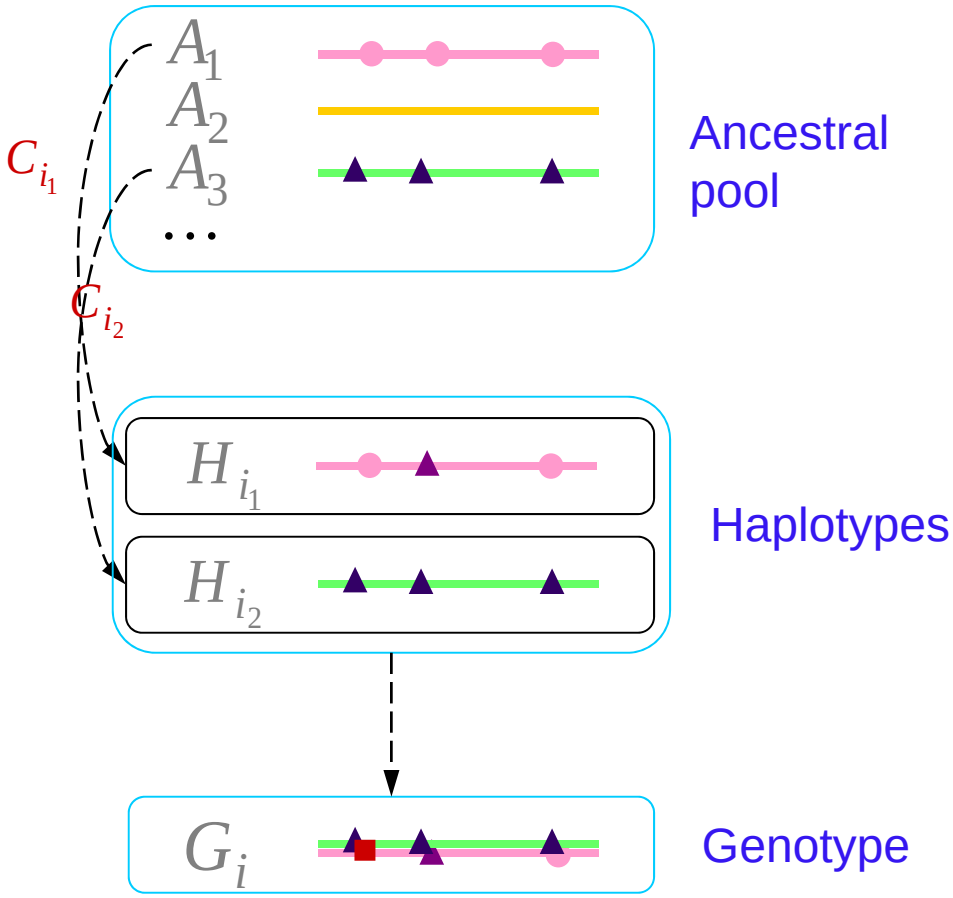
$$\rightarrow h_t = a_t \quad with \ prob. \ \theta$$

- Noisy observation model

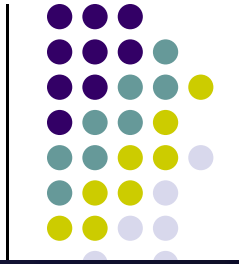
$$H_{i_1}, H_{i_2} \rightarrow G_i$$

$$P_G(g \mid h_1, h_2):$$

$$g_t = h_{1,t} \oplus h_{2,t} \quad with \ prob. \ \lambda$$



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}

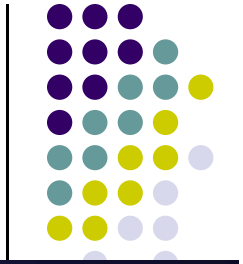
$$p(c_{i_e} = k \mid \mathbf{c}_{[-i_e]}, \mathbf{h}, \mathbf{a}) \propto p(c_{i_e} = k \mid \mathbf{c}_{[-i_e]})\, p(h_{i_e} \mid a_k, \mathbf{h}_{[-i_e]}, \mathbf{c})$$

Posterior Prior x Likelihood

CRP

- 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

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

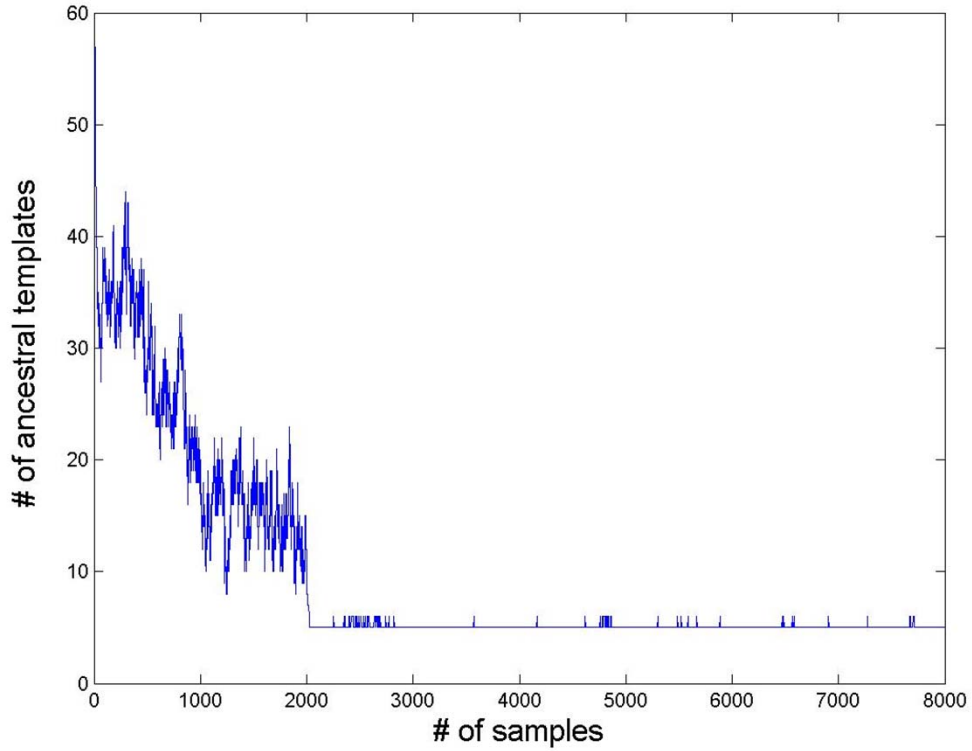
2. Sample a_k from

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

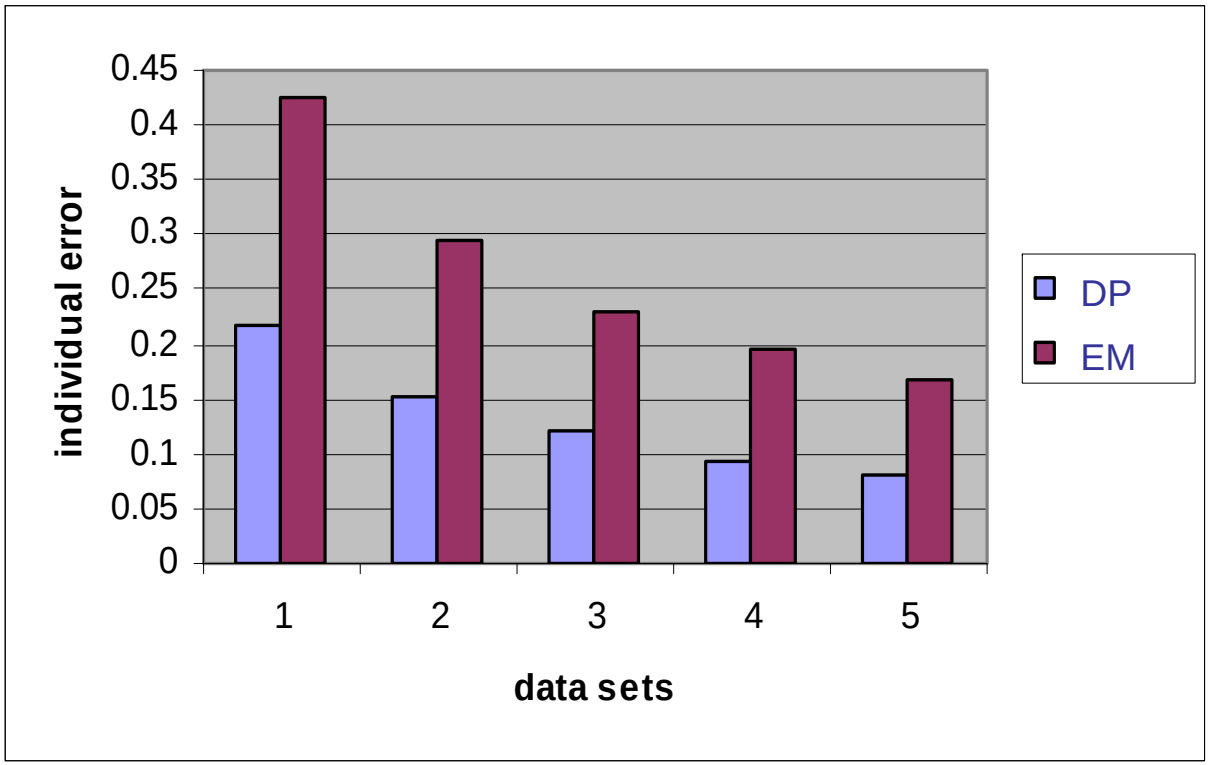
3. Sample $h_{ie}^{(j)}$ from $p(h_{i_e,t}^{(j)} | \mathbf{h}_{[-i_e,t]}^{(j)}, \mathbf{c}, \mathbf{a}, \mathbf{g})$

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

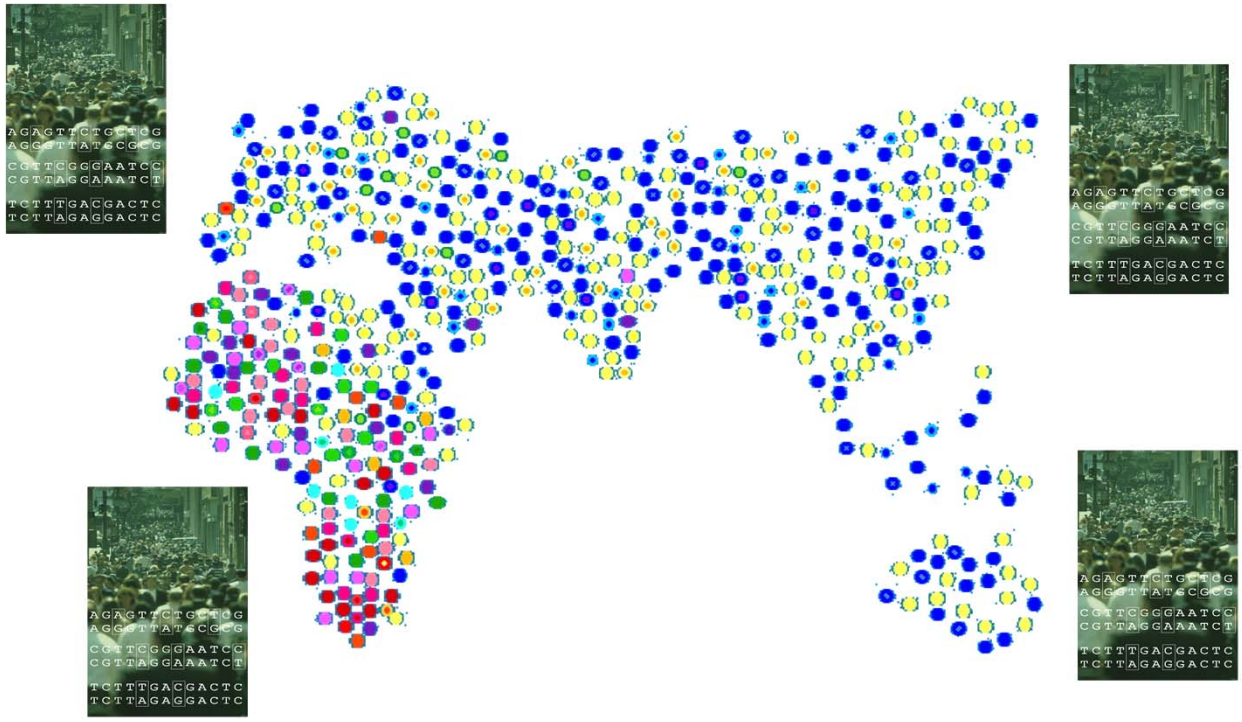
Convergence of Ancestral Inference



DP vs. Finite Mixture via EM

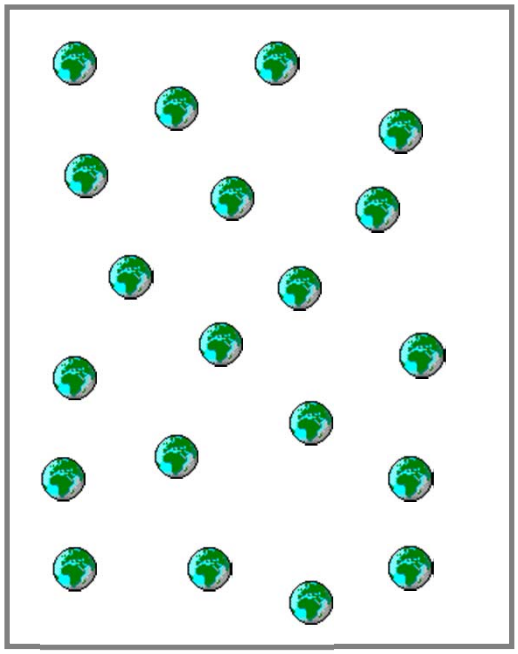


Multi-population Genetic Demography



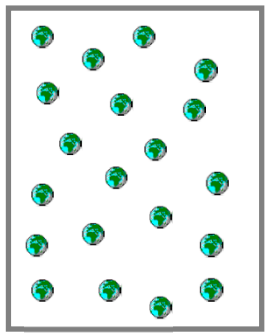
- Pool everything together and solve 1 hap problem?
  - --- ignore population structures
- Solve 4 hap problems separately?
  - --- data fragmentation
- Co-clustering … solve 4 *coupled* hap problems jointly

Solving Multiple Clustering Problems



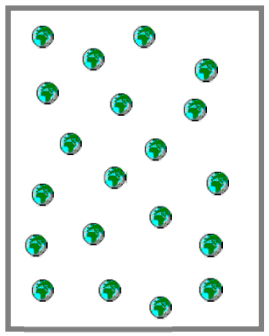
$G(\ ) \sim DP(\alpha G_0)$

1 3 2 4 5 6



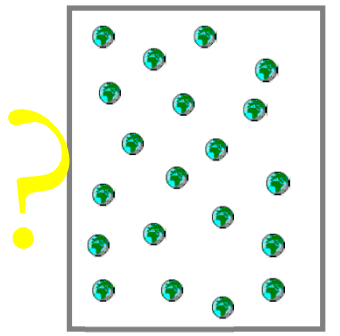
$G(\ ) \sim DP(\alpha G_0)$

Nature articles



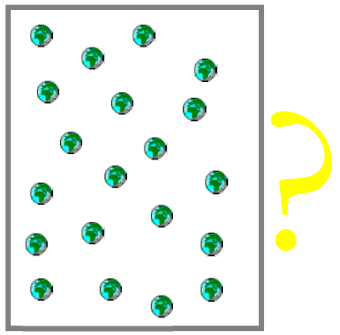
$G(\ ) \sim DP(\alpha G_0)$

PNAS articles



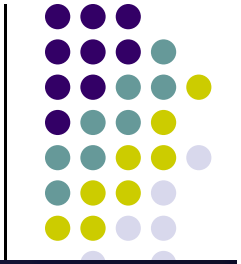
$G(\ ) \sim DP(\alpha G_0)$

Science articles

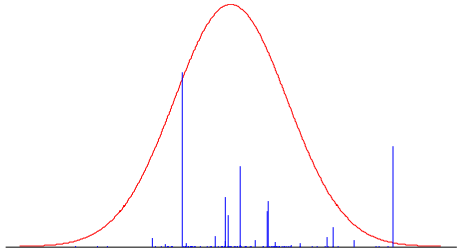


$G(\ ) \sim DP(\alpha G_0)$

How? And what is the challenge?

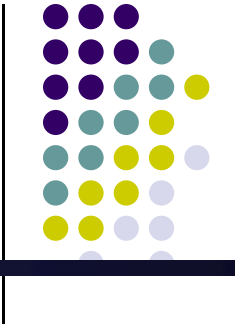


- How to share mixture components?



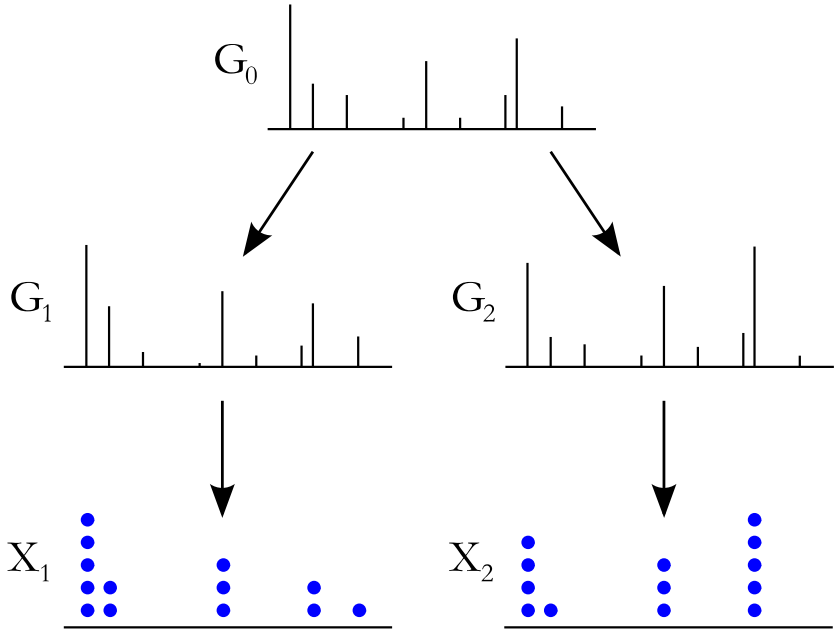
  - Because the base measure is *continuous*, we have zero probability of picking the same component twice.
  - If we want to pick the same topic twice, we need to use a *discrete* base measure.
    - For example, if we chose the base measure to be $H = \sum_{k=1}^{K} \alpha_k \delta_{\beta_k}$
  - We want there to be an infinite number of topics, so we want an *infinite, discrete* base measure.
  - We want the location of the topics to be random, so we want an *infinite, discrete, random* base measure.

Hierarchical Dirichlet Process (Teh et al, 2006)



- Solution: Sample the base measure from a Dirichlet process!

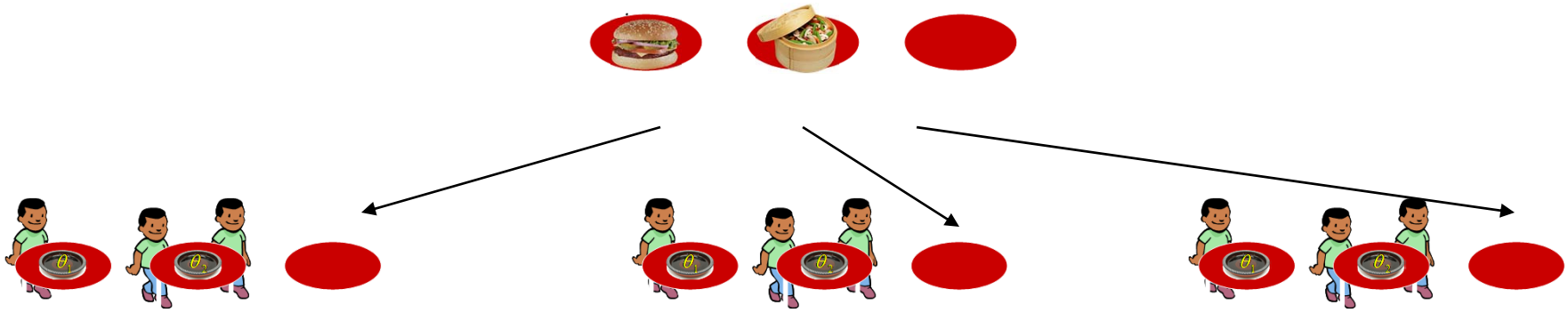
$$G_0 \sim \mathrm{DP}(\gamma, H)$$
$$G_m \sim \mathrm{DP}(\alpha, G_0)$$



Hierarchical Dirichlet Process

[Teh et al., 2005, Xing et al. 2005]

- **Two level CRP scheme: The Chinese restaurant franchise**
  - At the *i*-th step in *j*-th "group",



– *Choose* θ_k *with prob.* $\dfrac{m_{jk}}{\sum_k m_{jk} + \alpha_0}$

– *Go to the upper level DP*

with prob. $\dfrac{\alpha_0}{\sum_k m_{jk} + \alpha_0}$

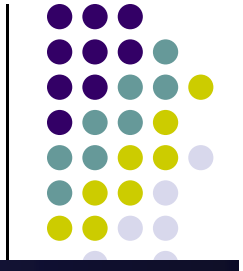
Oracle

Choose θ_k *with prob.* $\dfrac{n_k}{\sum n_k + \gamma}$

Draw a new sample

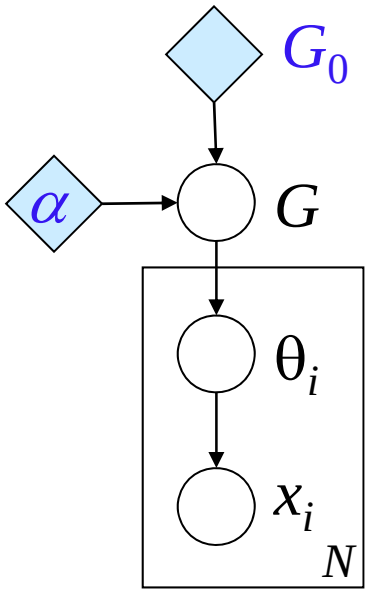
with prob. $\dfrac{\gamma}{\sum n_k + \gamma}$

Chinese restaurant franchise

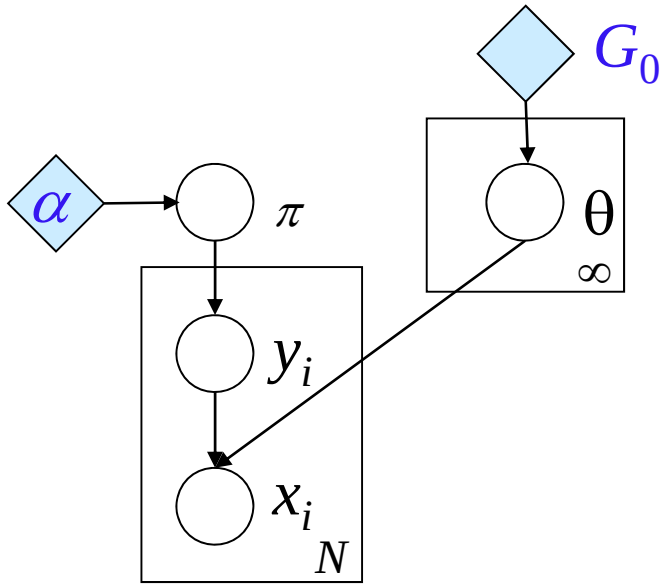


- Imagine a *franchise* of restaurants, serving an infinitely large, global menu.
- Each table in each restaurant orders a single dish.
- Let n_{rt} be the number of customers in restaurant r sitting at table t.
- Let m_{rd} be the number of tables in restaurant r serving dish d.
- Let $m_{.d}$ be the number of tables, across *all* restaurants, serving dish d.

Recall: Graphical Model Representations of DP

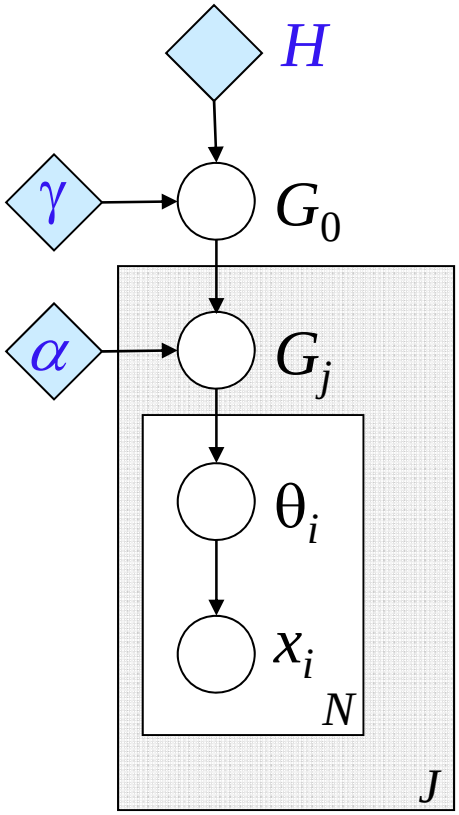


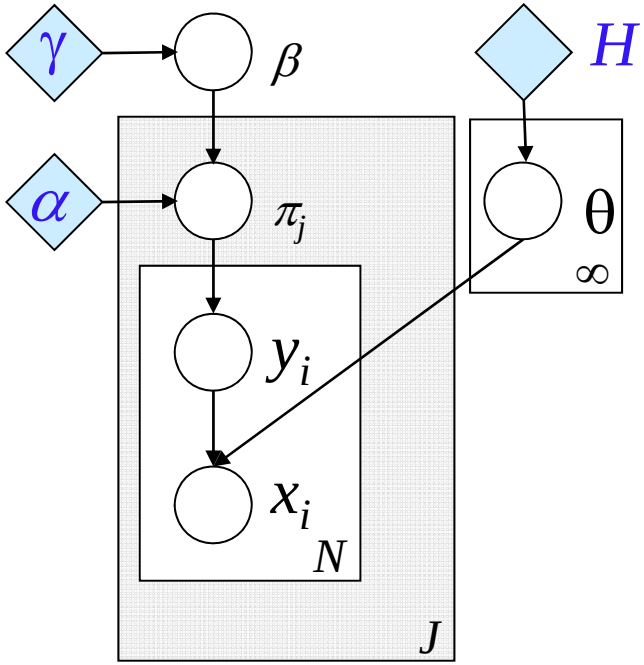
The Pólya urn construction



The Stick-breaking construction

Hierarchical DP Mixture





$\theta_k \sim H$

Stick(α,β):

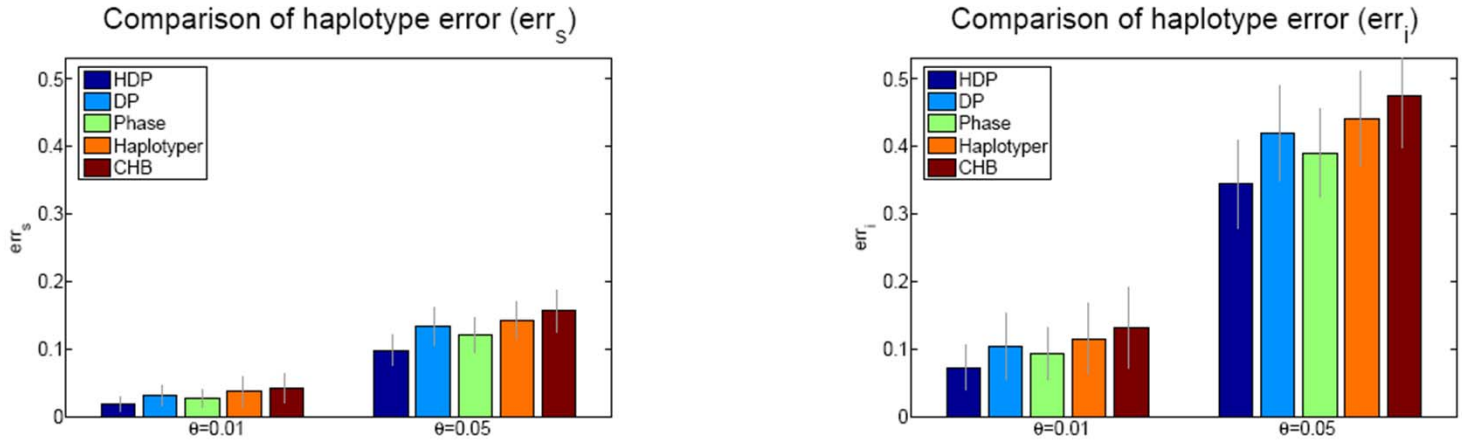
$$\pi'_{jk} \sim \text{Beta}\left(\alpha\beta_k, \alpha\left(1-\sum_{l=1}^{k}\beta_l\right)\right), \quad \pi_{jk} = \pi'_{jk}\prod_{l=1}^{k-1}\left(1-\pi'_{jl}\right).$$

$$\beta = \text{Stick}(\gamma), G_0 = \sum_{k=1}^{\infty}\beta_k\delta(\theta_k)$$

$$\pi_j = \text{Stick}(\alpha,\beta), G_j = \sum_{k=1}^{\infty}\pi_k\delta(\theta_k)$$

Results - Simulated Data

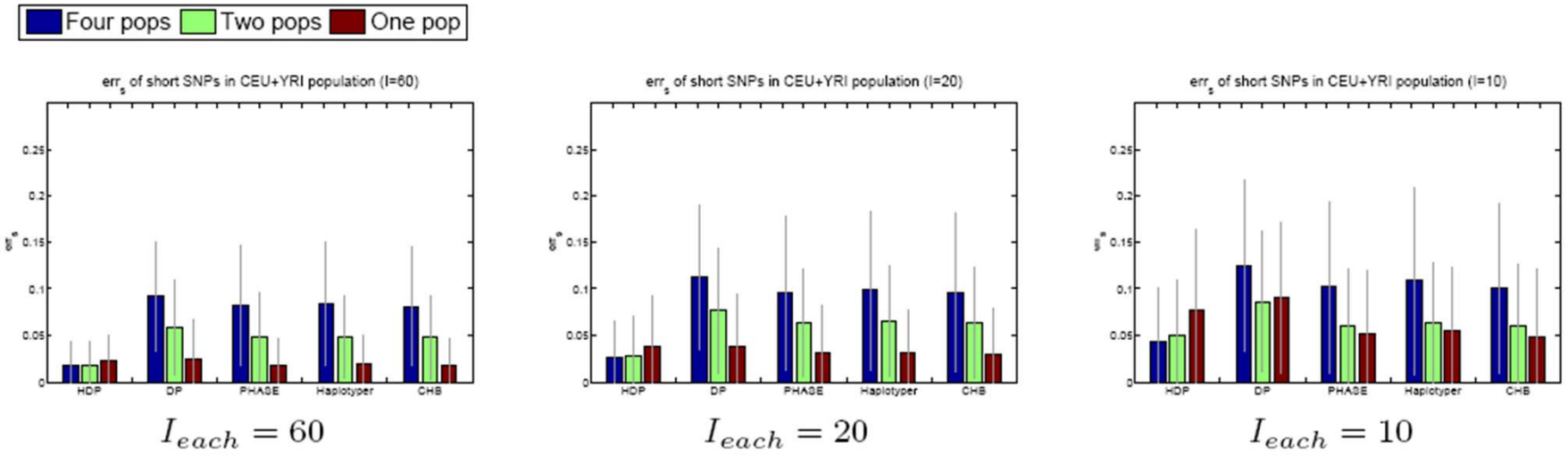
- 5 populations with 20 individuals each (two kinds of mutation rates)
- 5 populations share parts of their ancestral haplotypes
- the sequence length = 10



Haplotype error

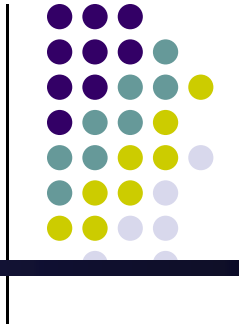
Results - International HapMap DB

- Different sample sizes, and different # of sub-populations



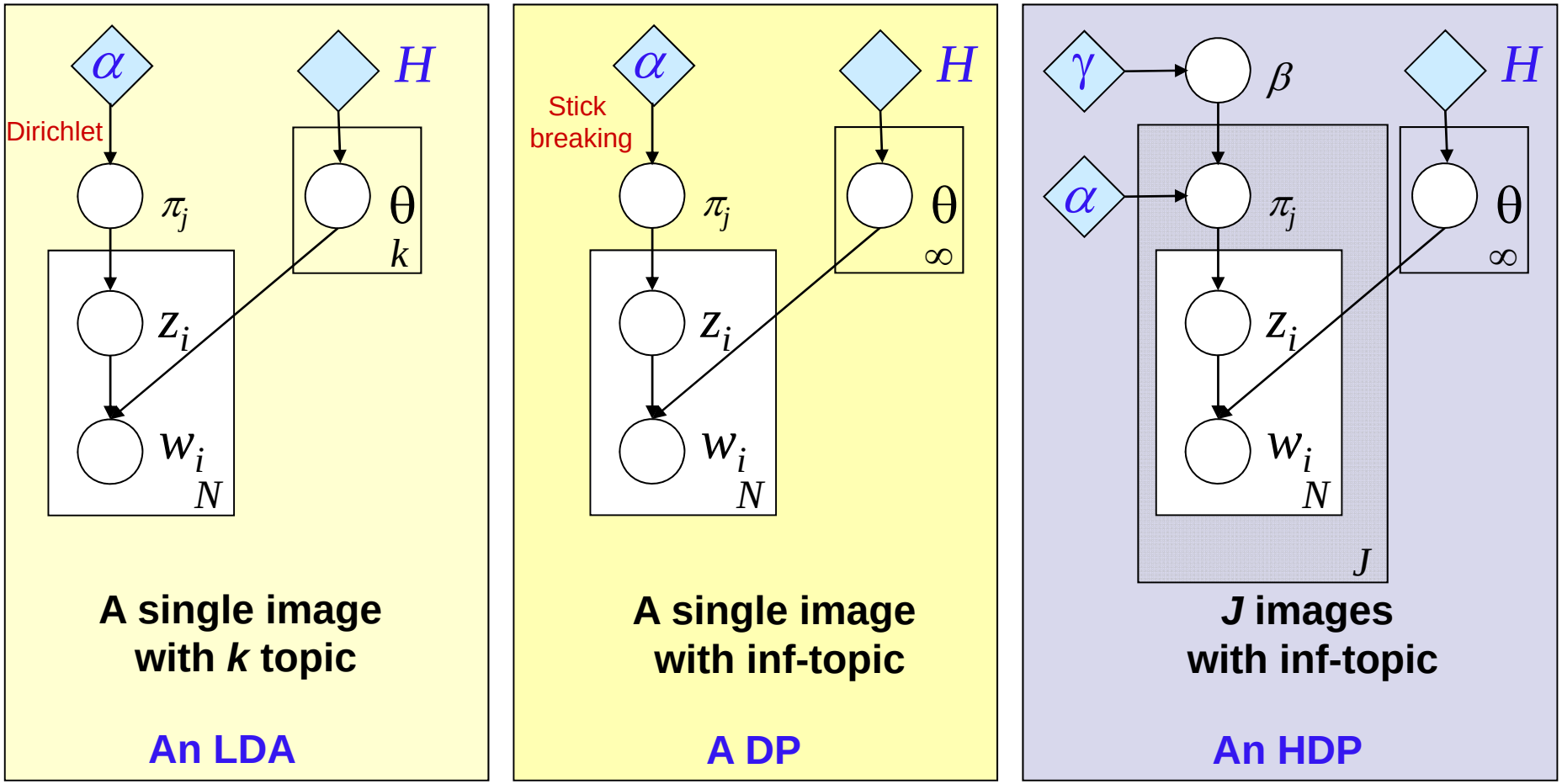
$I_{each} = 60$ $I_{each} = 20$ $I_{each} = 10$

Constructing a topic model with infinitely many topics

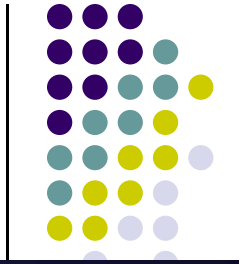


- LDA: Each distribution is associated with a distribution over *K* topics.
- Problem: How to choose the number of topics?
- Solution:
  - Infinitely many topics!
  - Replace the Dirichlet distribution over topics with a Dirichlet process!
- Problem: We want to make sure the topics are *shared* between documents

Infinite Topic Models



An infinite topic model



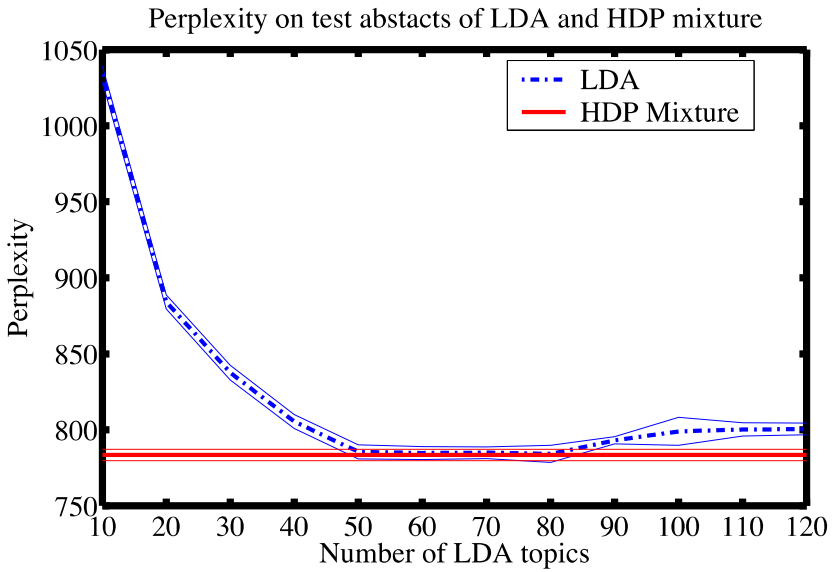
- Restaurants = documents; dishes = topics.

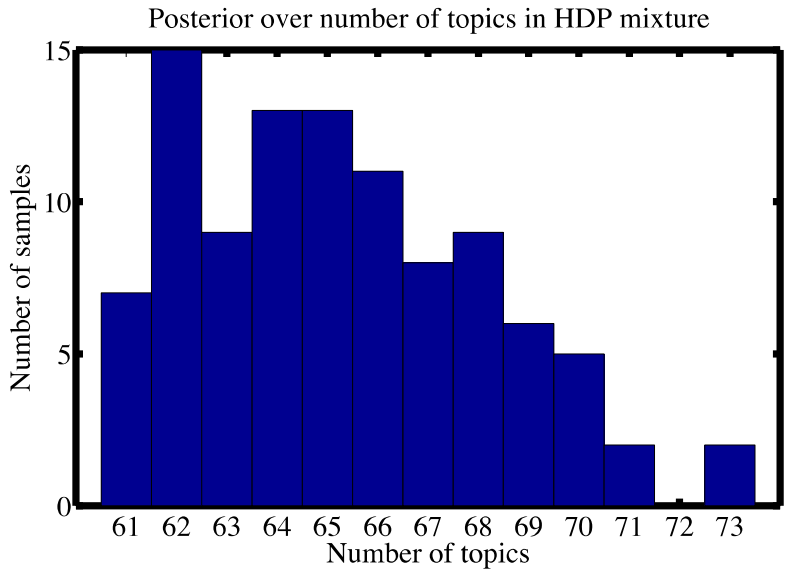
- Let *H* be a V-dimensional Dirichlet distribution, so a sample from *H* is a distribution over a vocabulary of *V* words.
- Sample a global distribution over topics,

$$G_0 := \sum_{k=1}^{\infty} \pi_k \delta_{\beta_k} \sim \mathrm{DP}(\alpha, H)$$

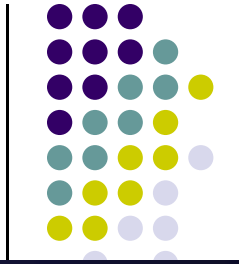
- For each document *m=1,…,M*
  - Sample a distribution over topics, $G_m \sim \mathrm{DP}(\gamma, G_0)$.
  - For each word *n=1,…,N_m*
    - Sample a topic $\phi_{mn} \sim \mathrm{Discrete}(G_0)$.
    - Sample a word $w_{mk} \sim \mathrm{Discrete}(\phi_{mn})$.

The “right” number of topics





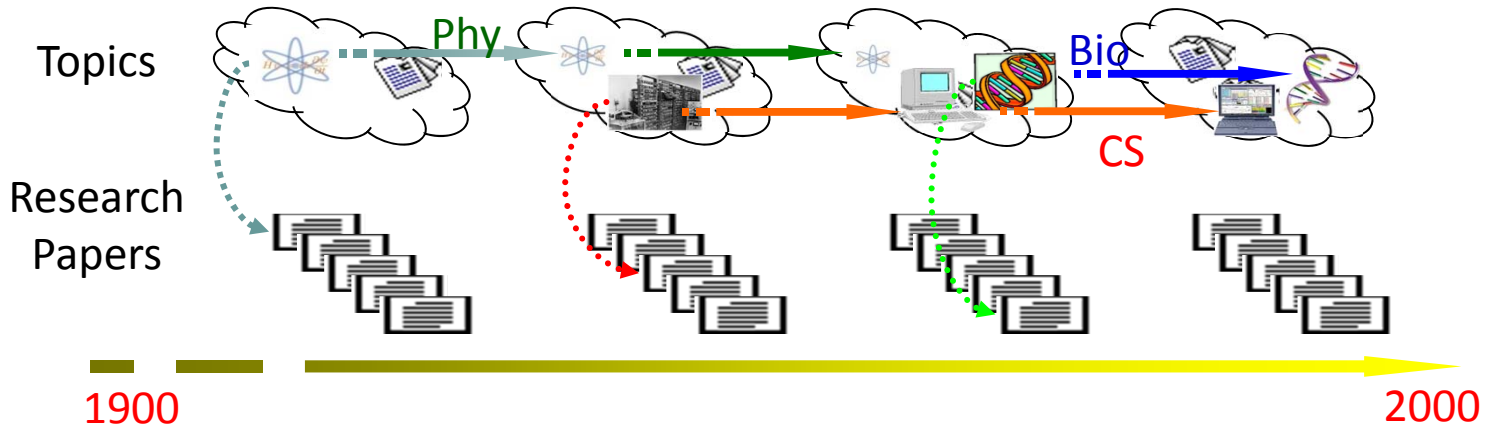
Dynamic Dirichlet Process



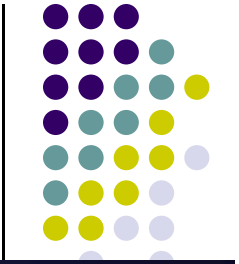
- Two Main Ideas:
  - Infinite HMM: a hidden Markov DP (see appendix)
  - Dependent DP/HDP: directly evolving a DP/HDP

Evolutionary Clustering

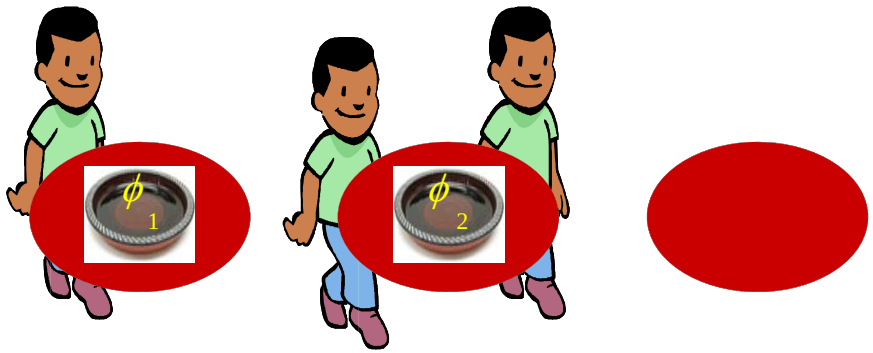
- Adapts the number of mixture components over time
  - Mixture components can die out
  - New mixture components are born at any time
  - Retained mixture components parameters evolve according to a Markovian dynamics



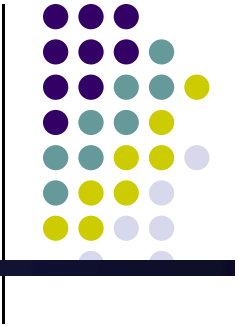
The Chinese Restaurant Process



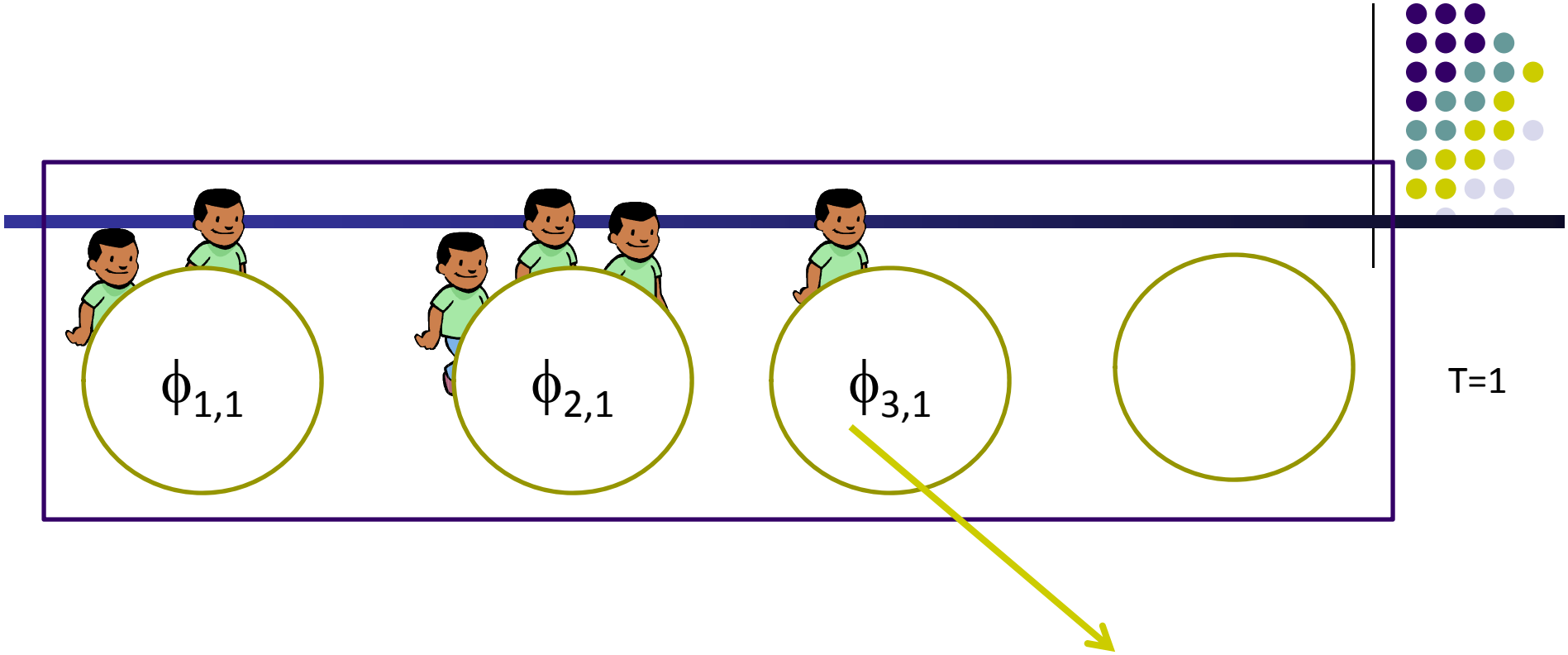
- Customers correspond to data points
- Tables correspond to clusters/mixture components
- Dishes correspond to parameter of the mixtures



Temporal DPM [Ahmed and Xing 2008]



- The Recurrent Chinese Restaurant Process
  - The restaurant operates in epochs
  - The restaurant is closed at the end of each epoch
  - The state of the restaurant at time epoch *t* depends on that at time epoch *t-1*
    - Can be extended to higher-order dependencies.

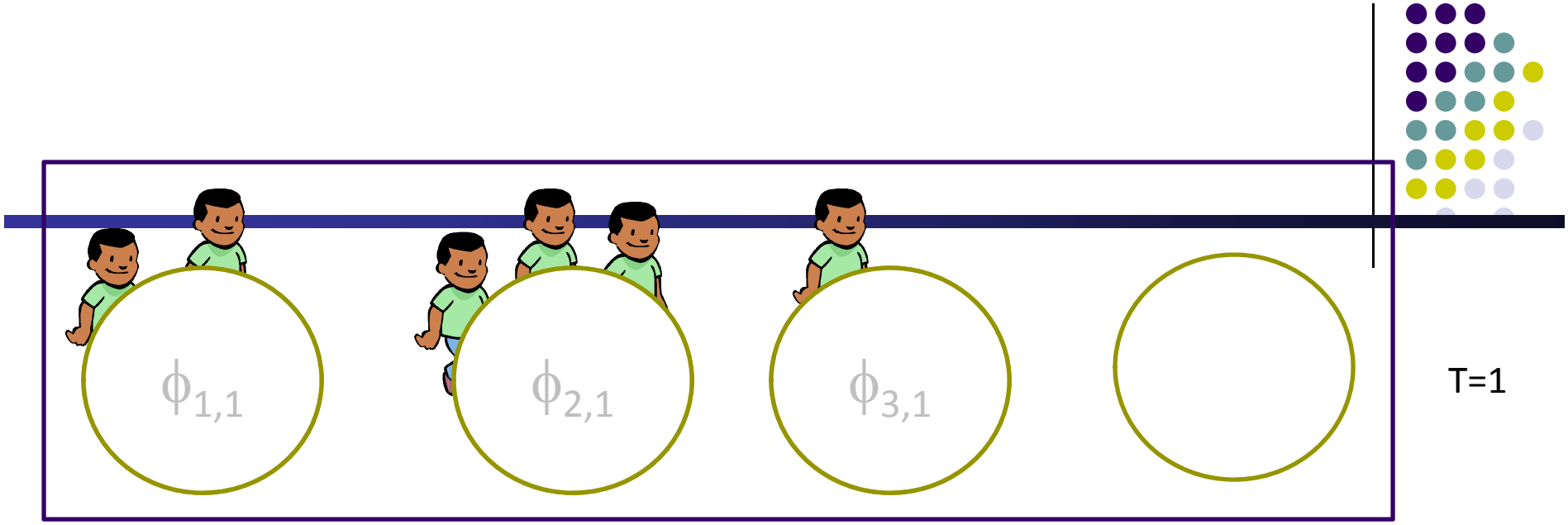


Dish eaten at table 3 at time epoch 1
OR the parameters of cluster 3 at time epoch 1

Generative Process

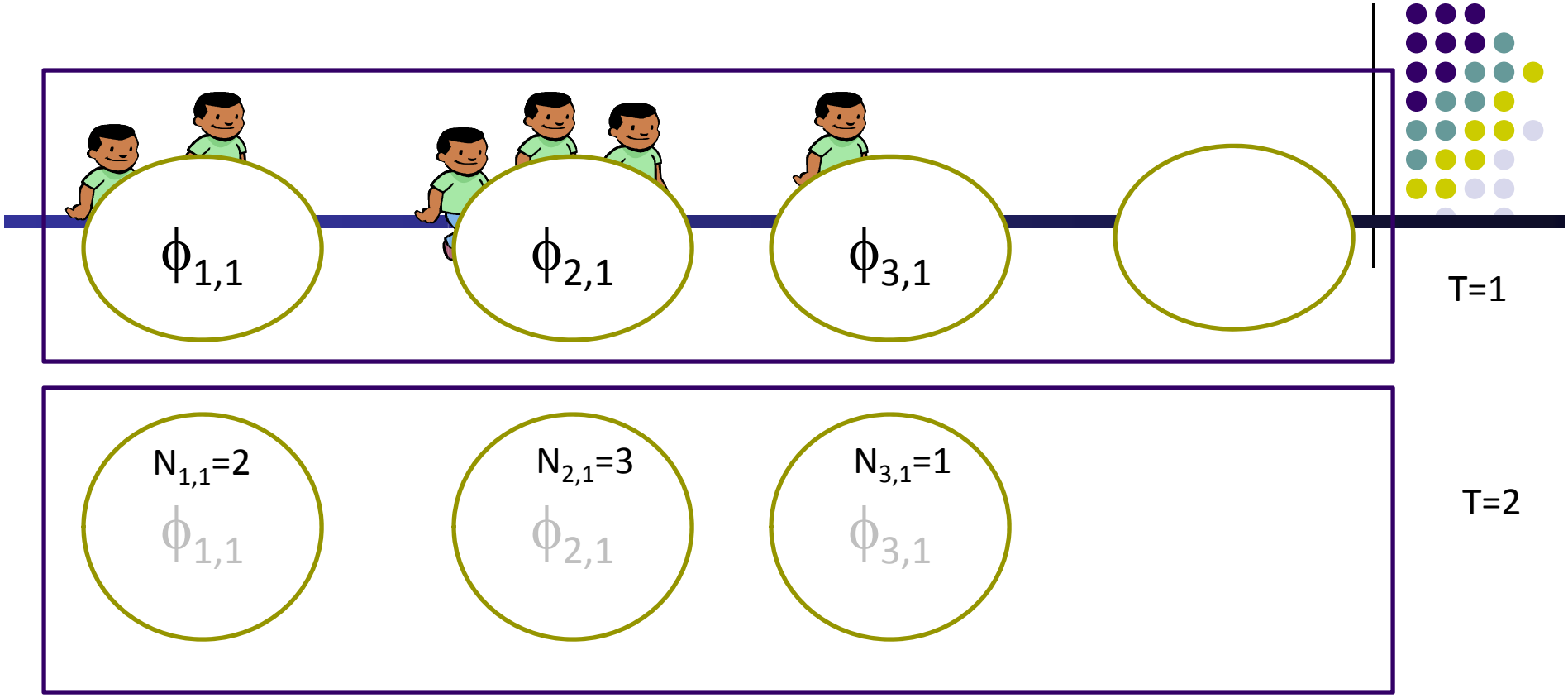
-Customers at time T=1 are seated as before:

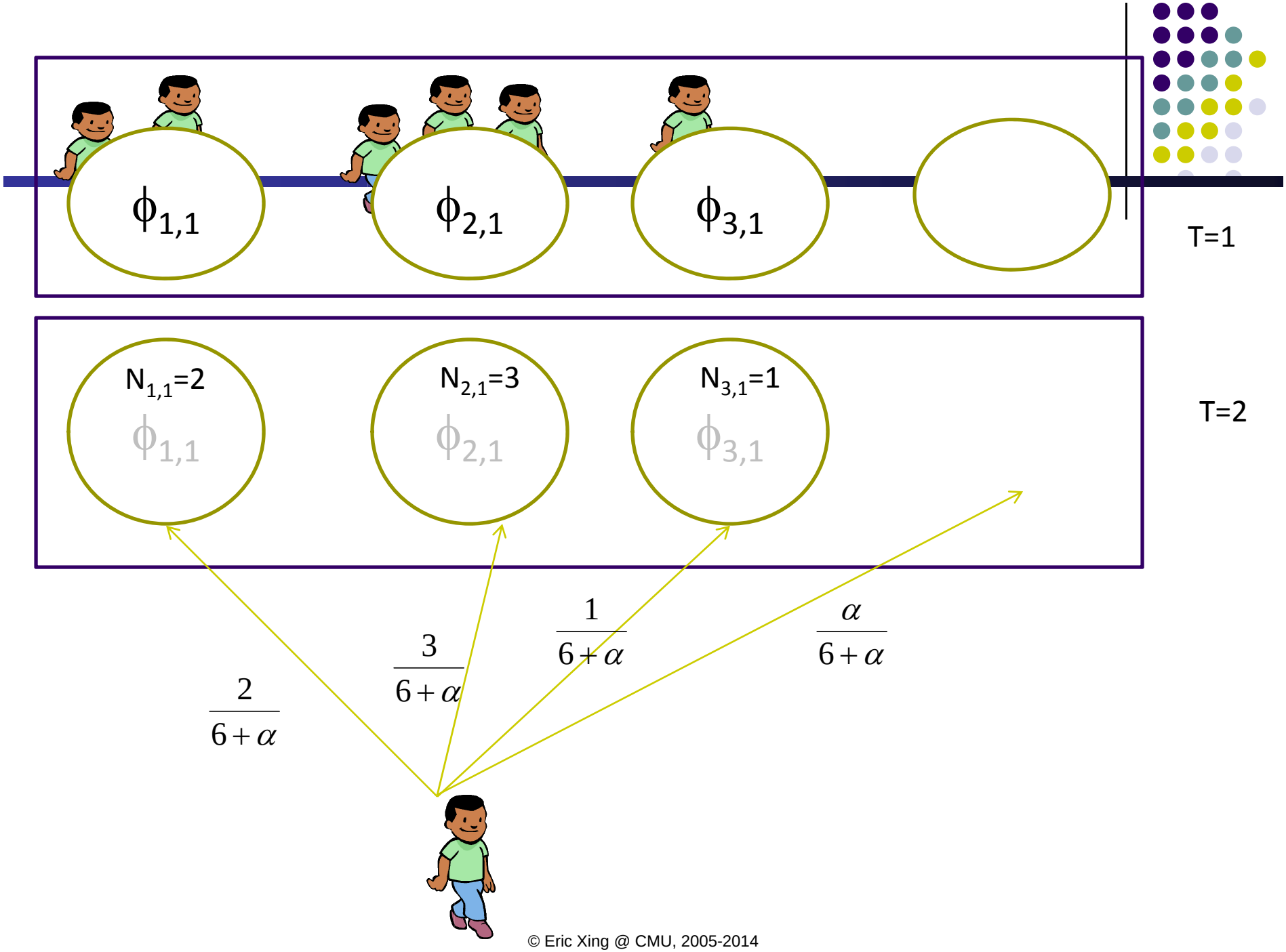
- Choose table $j \propto N_{j,1}$ and Sample $x_i \sim f(\phi_{j,1})$
- Choose a new table $K+1 \propto \alpha$
  - Sample $\phi_{K+1,1} \sim G_0$ and Sample $x_i \sim f(\phi_{K+1,1})$

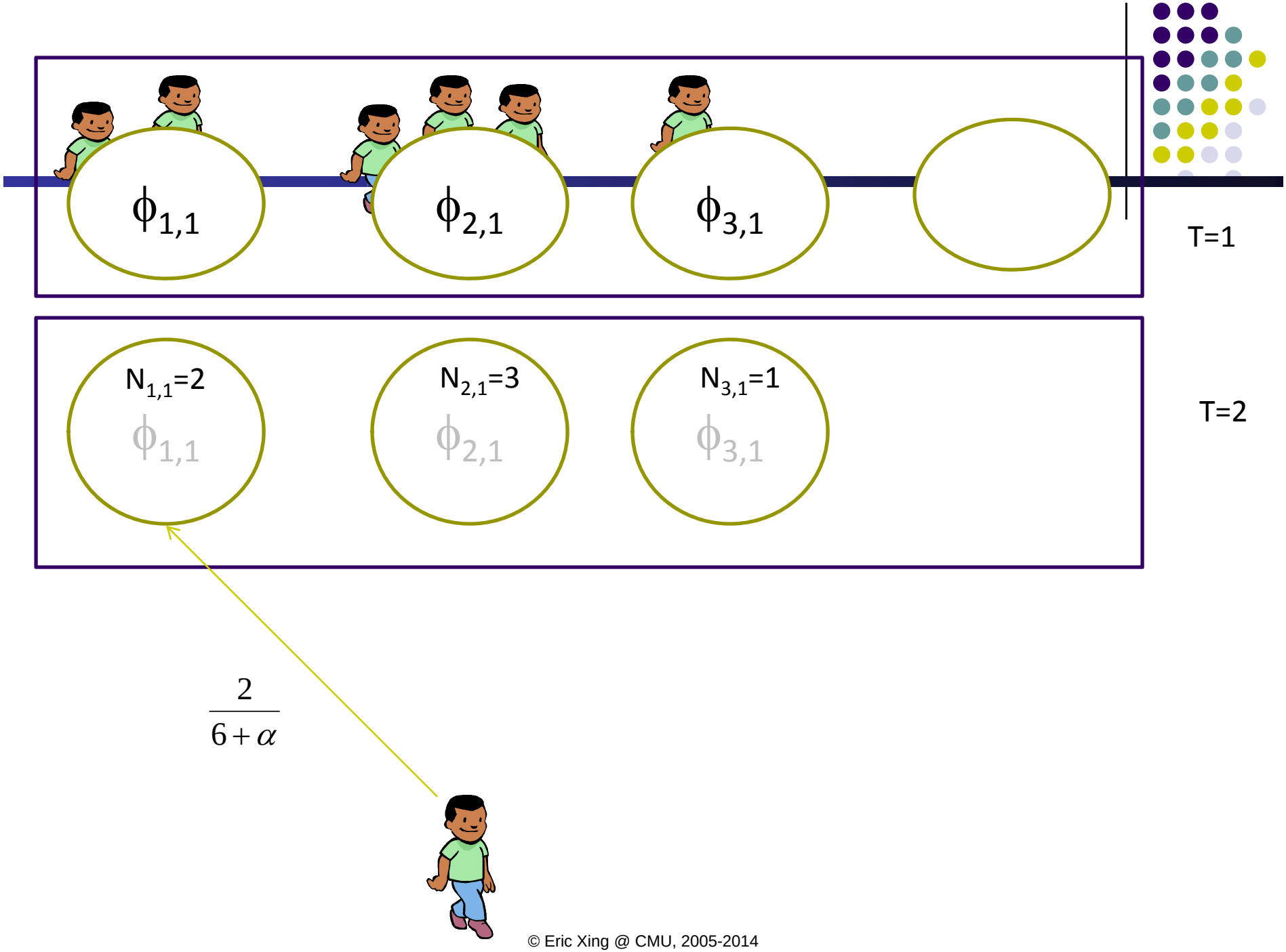


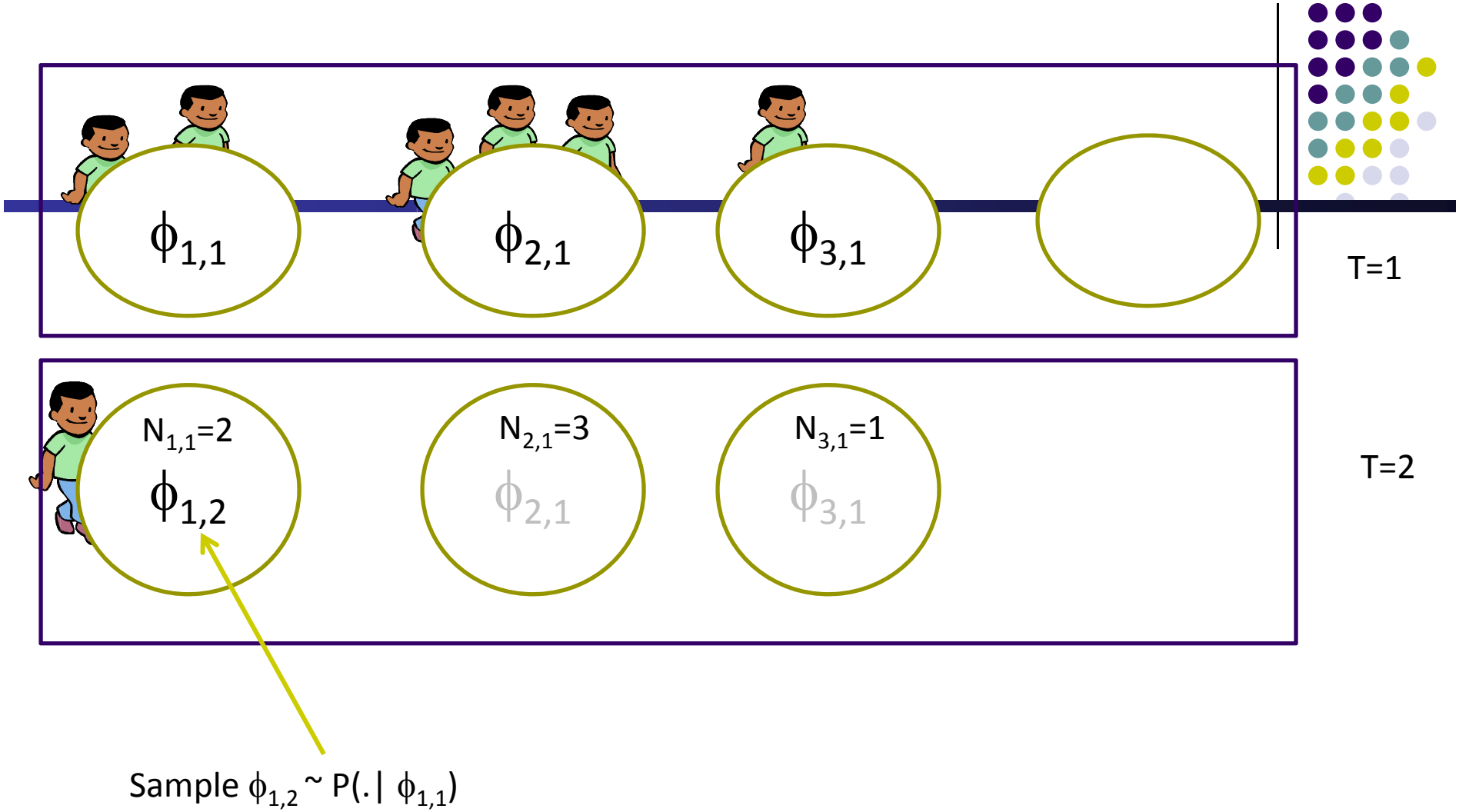
$N_{1,1}=2$ $N_{2,1}=3$ $N_{3,1}=1$

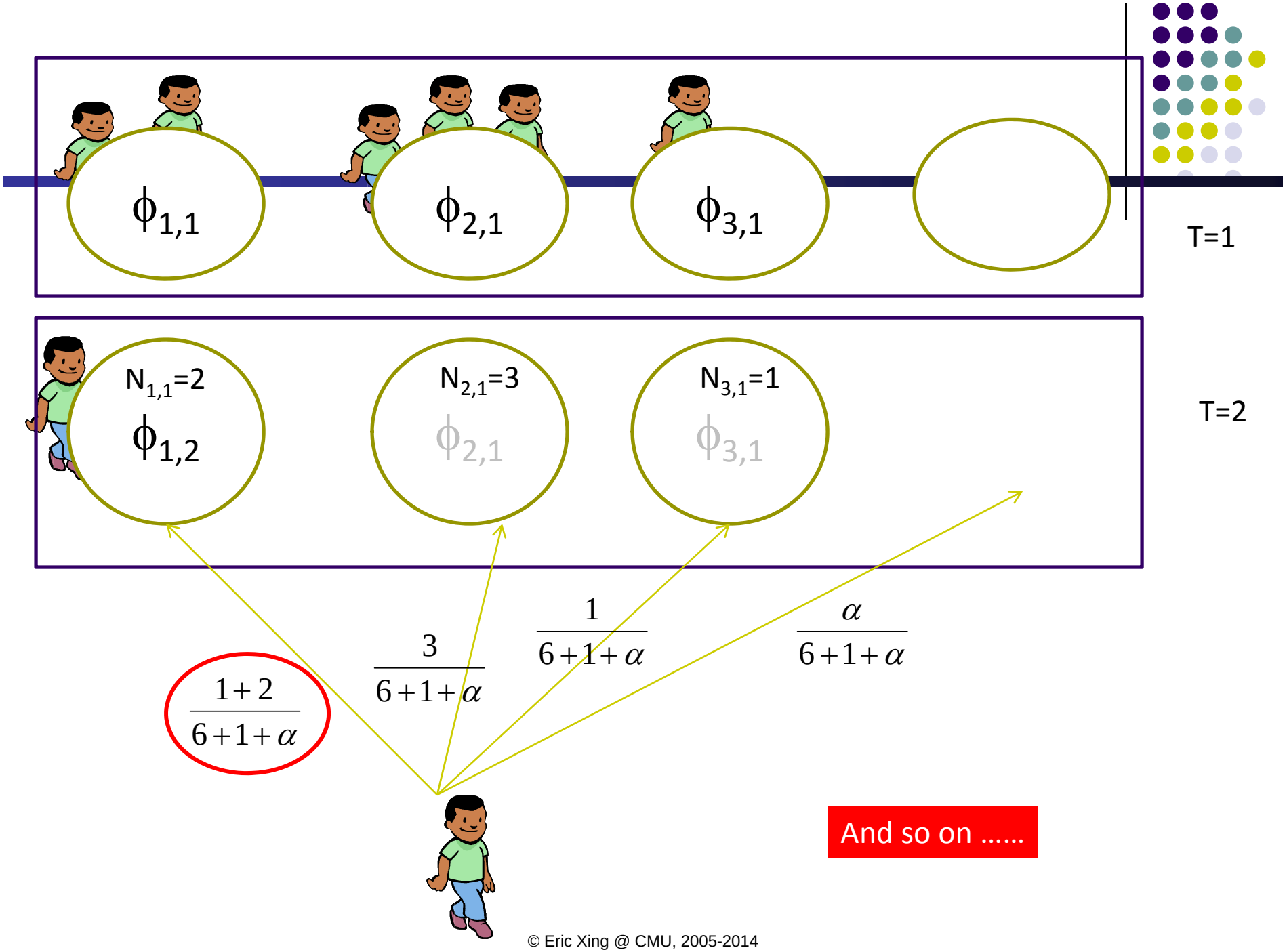
T=2

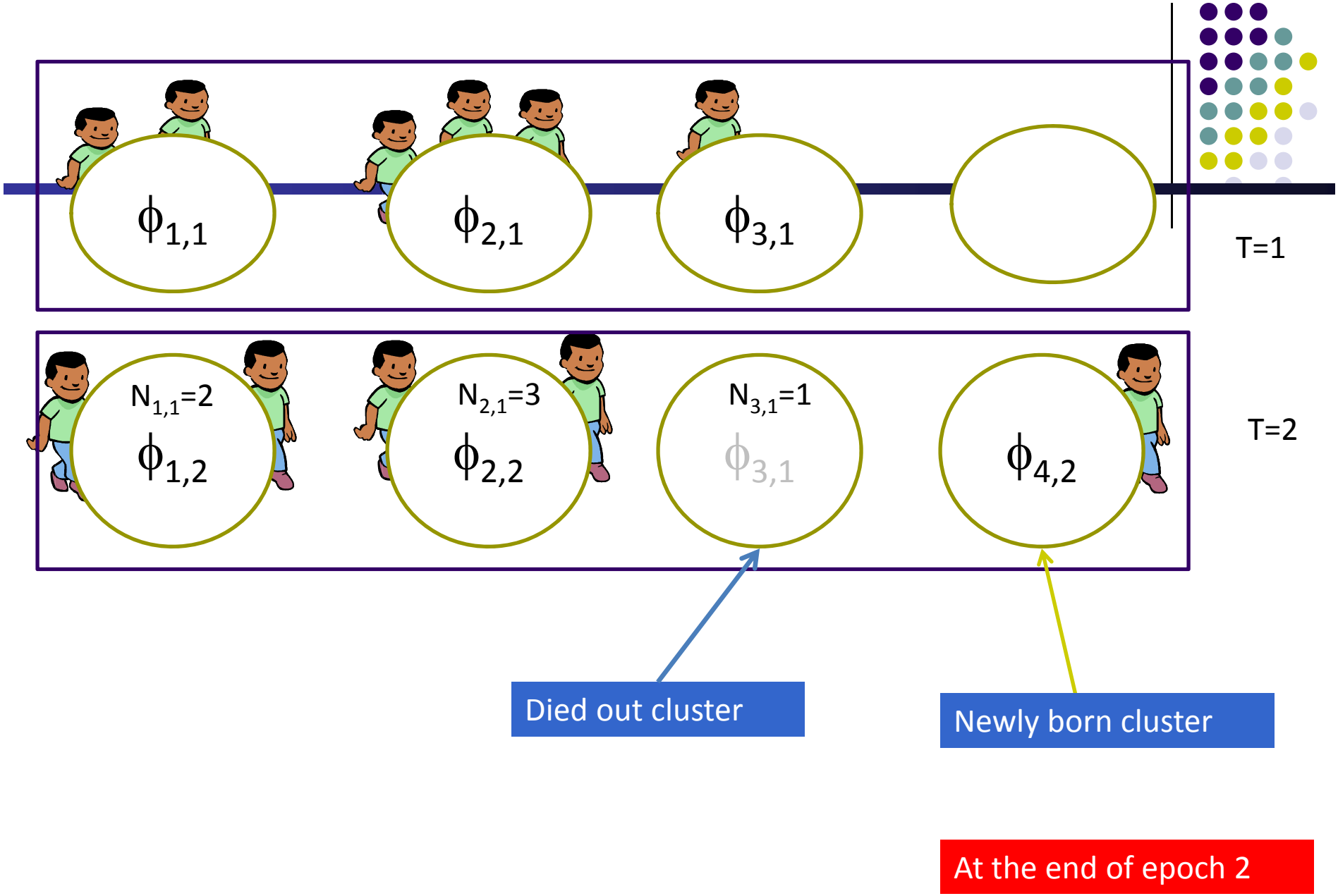


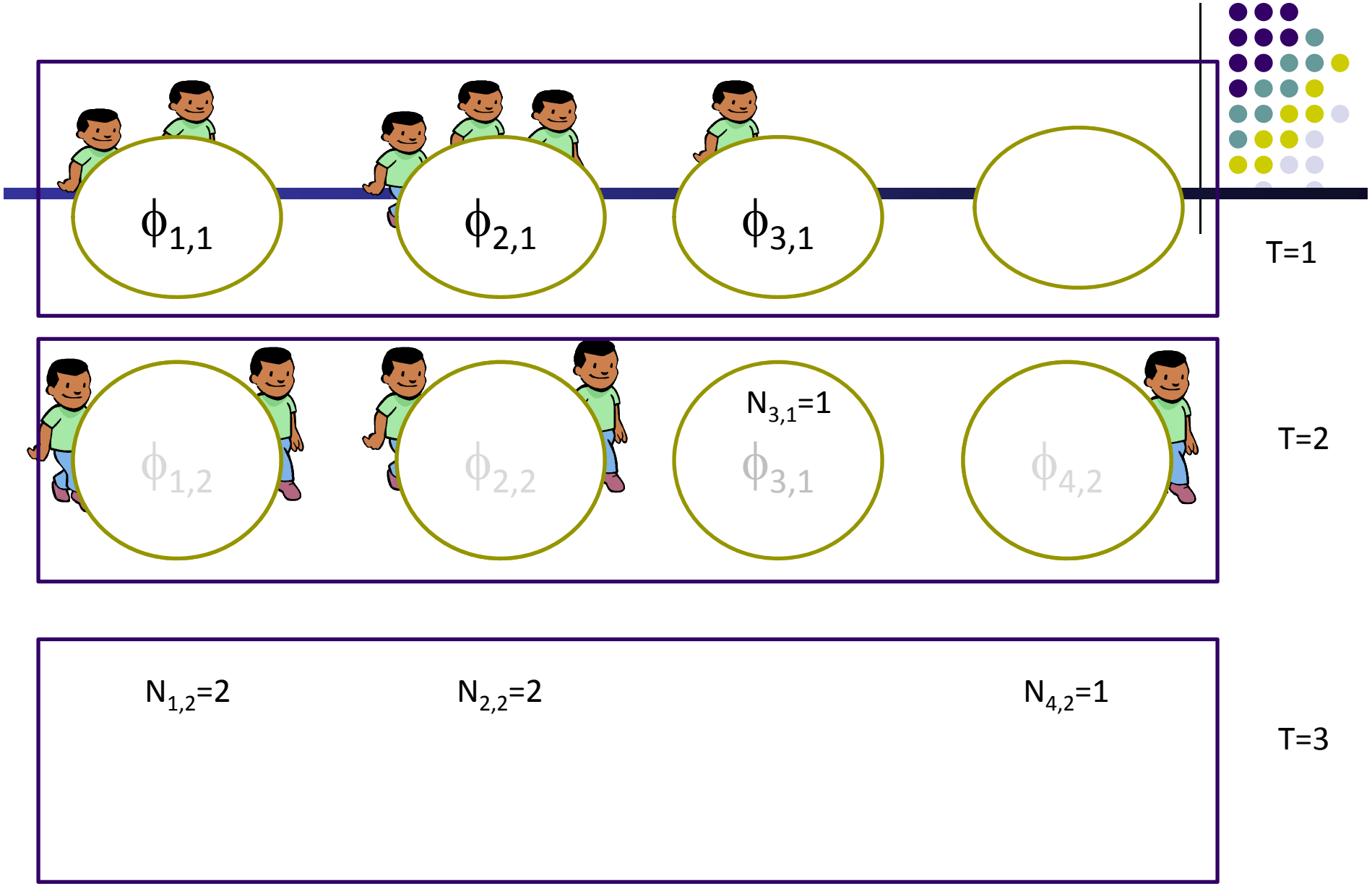




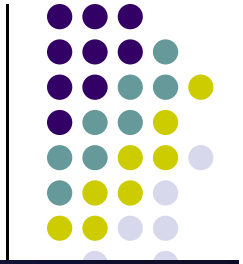








Temporal DPM



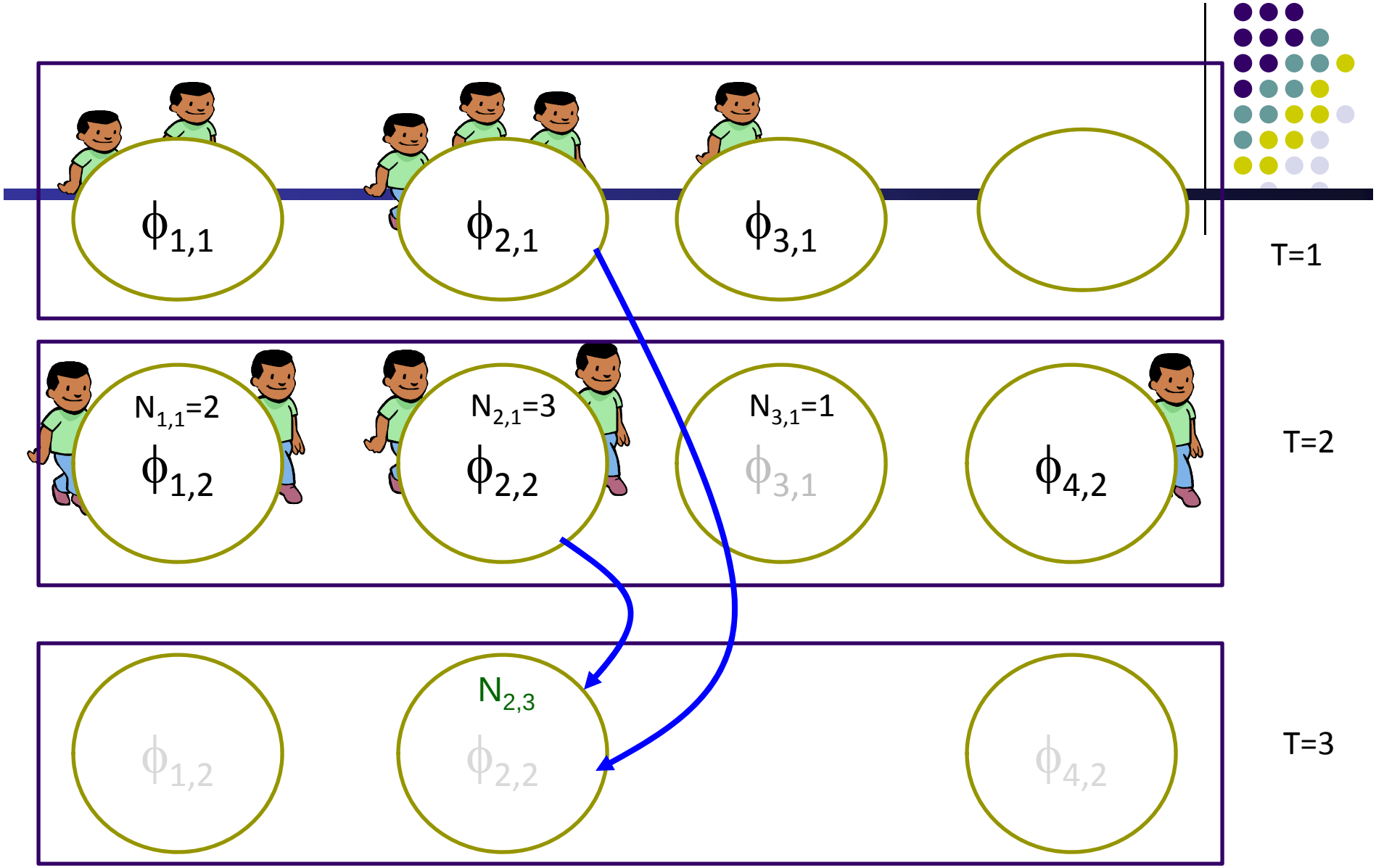
- Can be extended to model higher-order dependencies
- Can decay dependencies over time
  - Pseudo-counts for table *k* at time *t is*

$$\sum_{w=1}^{W}\left(e^{\frac{-w}{\lambda}}N_{k,t-w}\right)$$

History size

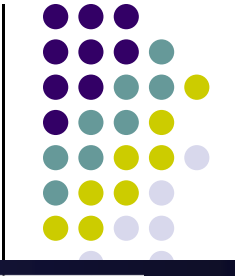
Decay factory

Number of customers sitting at table *K* at time epoch *t-w*



$$N_{2,3} = \sum_{w=1}^{W} \left(e^{\frac{-w}{\lambda}} N_{k,t-w} \right)$$

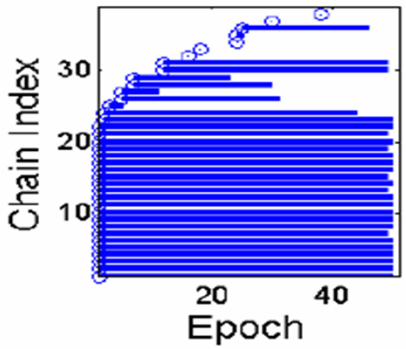
TDPM Generative Power

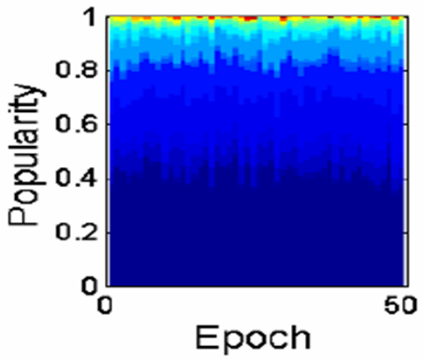


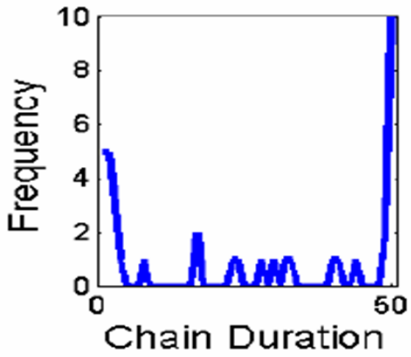
DPM

W=T

$\lambda = \infty$



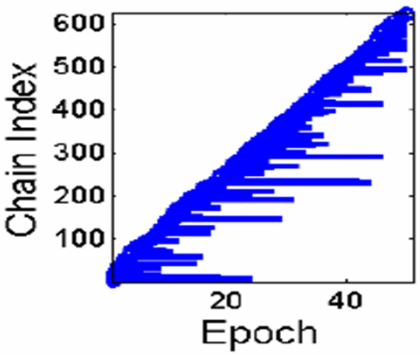


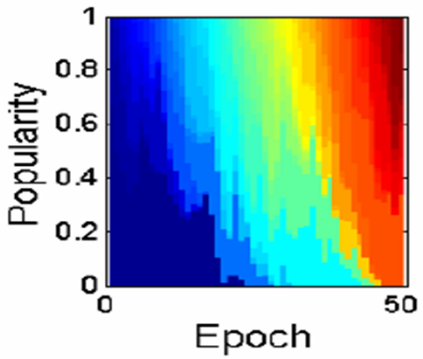


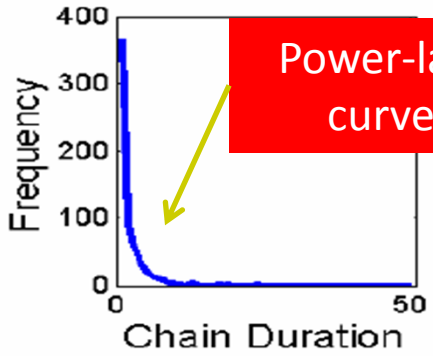
TDPM

W=4

$\lambda = .4$





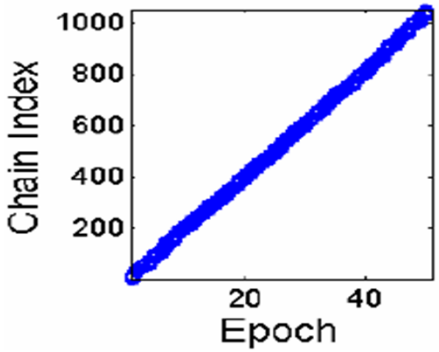


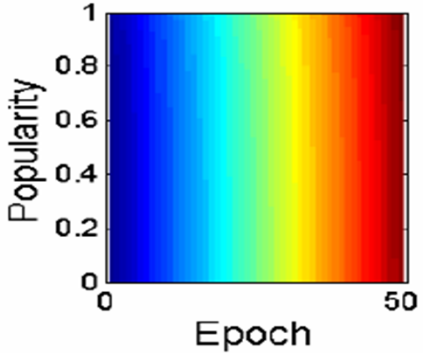
Power-law curve

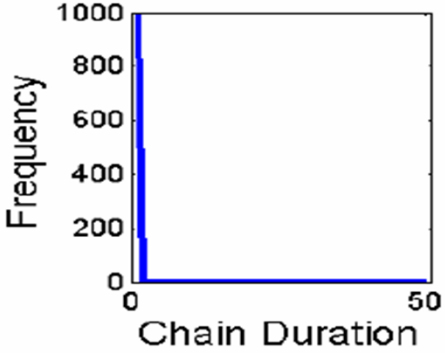
Independent DPMs

W= 0

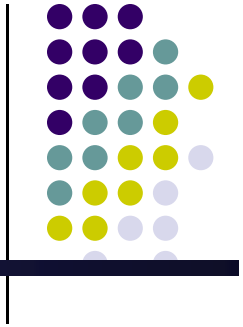
$\lambda = ?$ (any)



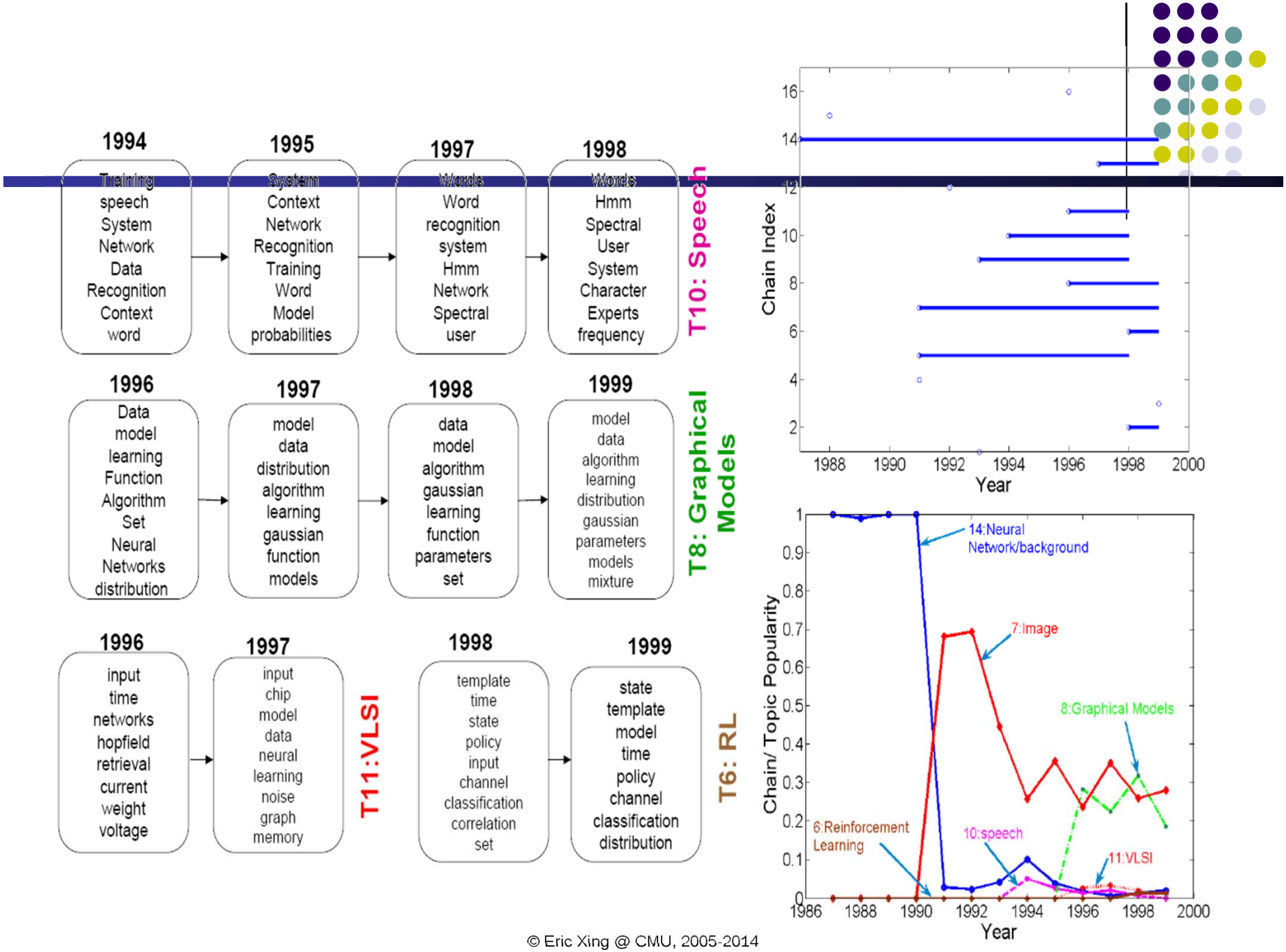




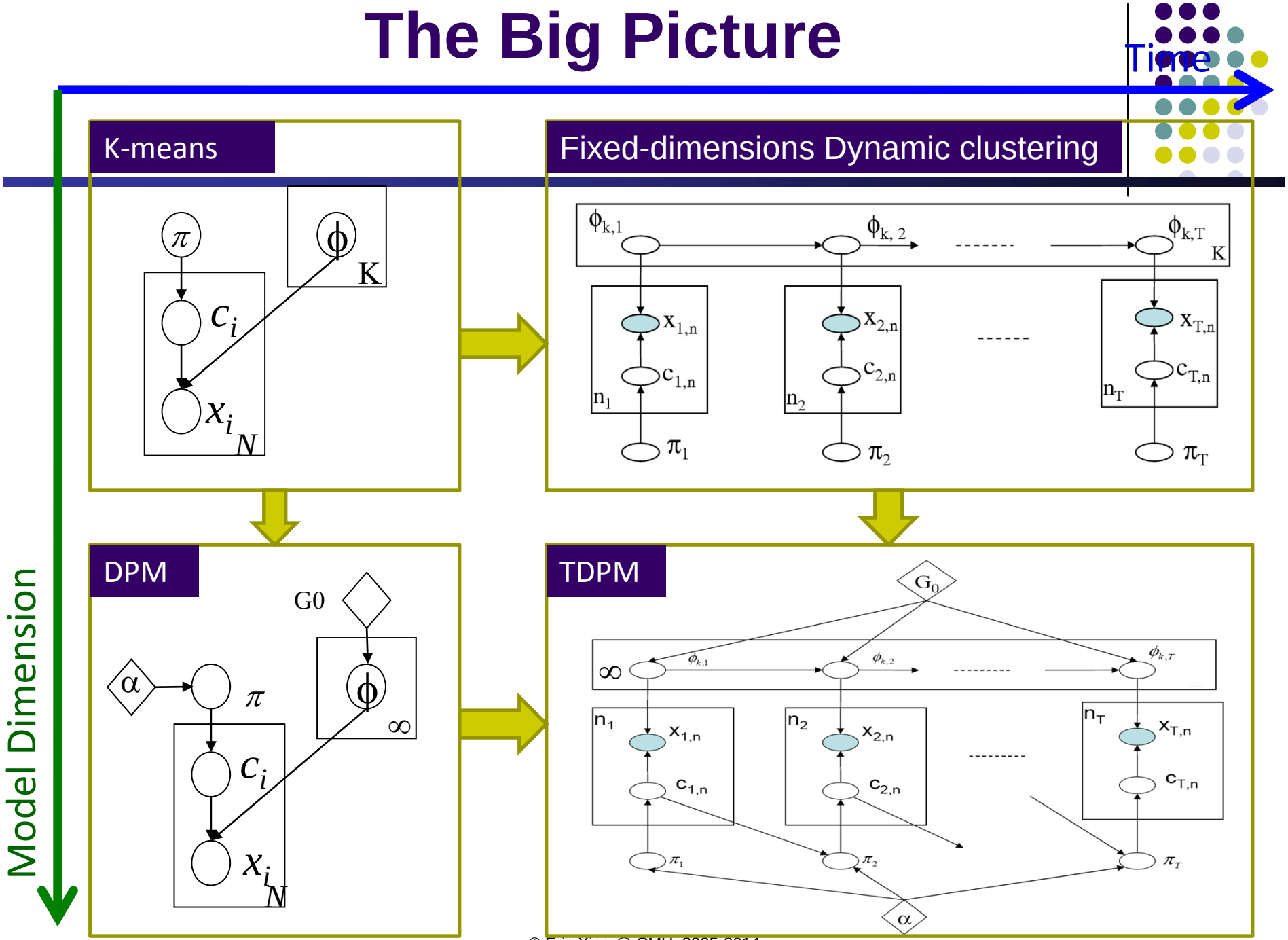
Results: NIPS 12



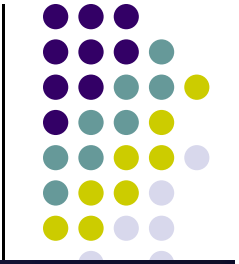
- Building a simple dynamic topic model
- Chain dynamics is as before
- Emission model for document $x_{k,t}$ is:
  - Project $\phi_{k,t}$ over the simplex
  - Sample $x_{k,t}|c_{t,i} \sim$ Multinomial$\left(.\middle|\text{Logisitic}(\phi_{k,t})\right)$
- Unlike LDA here a document belongs to one topic
- Use this model to analyze NIPS12 corpus
  - Proceeding of NIPS conference 1987-1999



The Big Picture



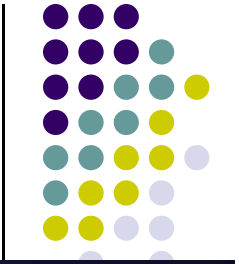
Summary



- 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
    → temporal infinite mixture model

- Applications in general data-mining …

Appendix:

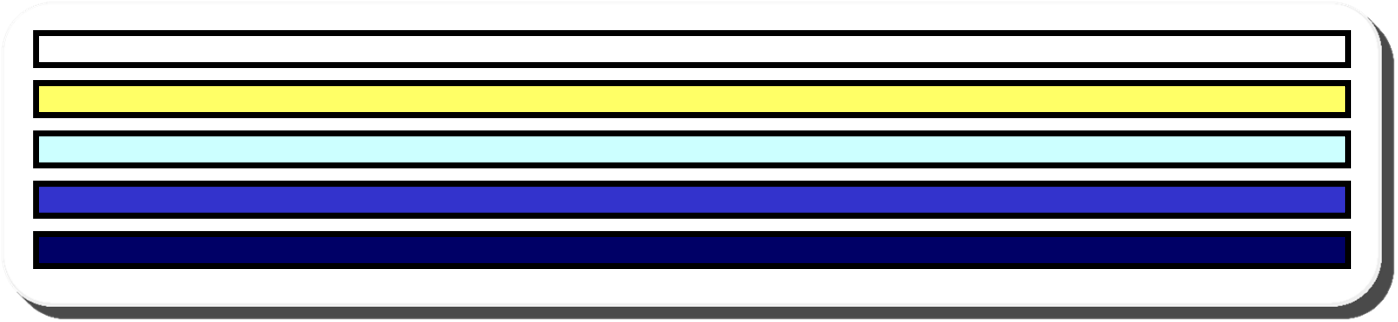


- What if we have an HMM with unknown # of states?
  - E.g., “recombination” over unknown number of chromosomes?

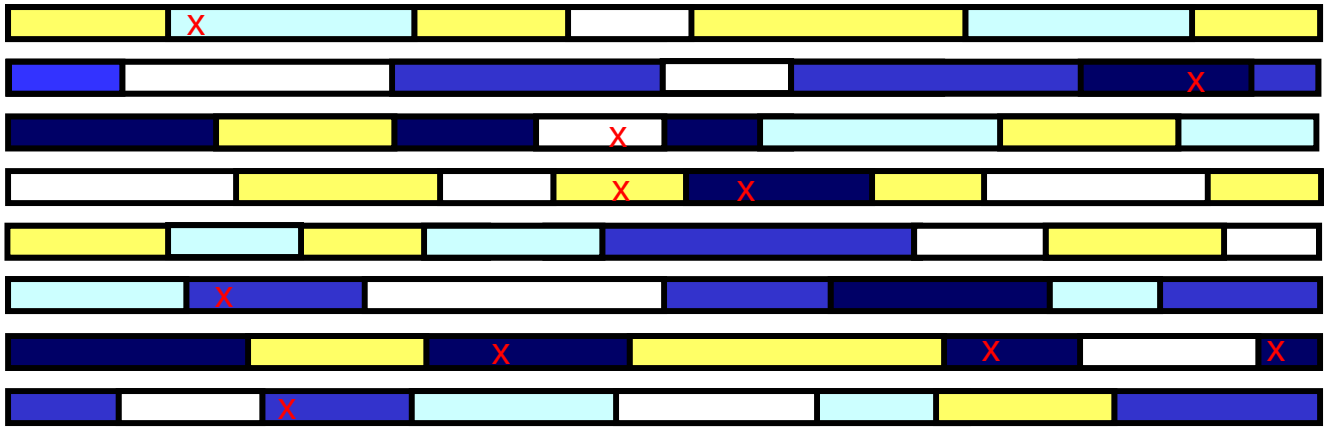
A common inheritance model to begin with …

Each individual haplotype is a mosaic of ancestral haplotypes

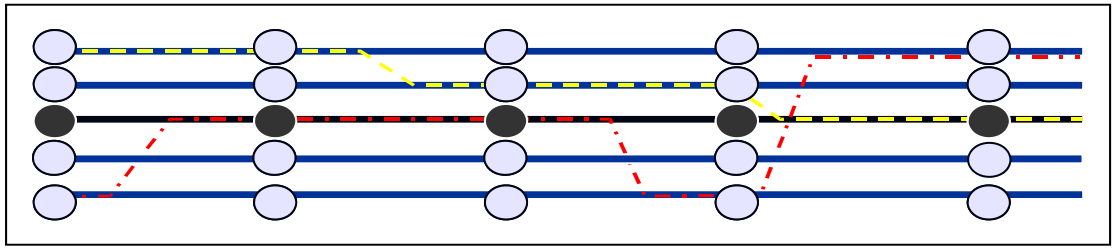
Ancestral chromosomes (K=5)

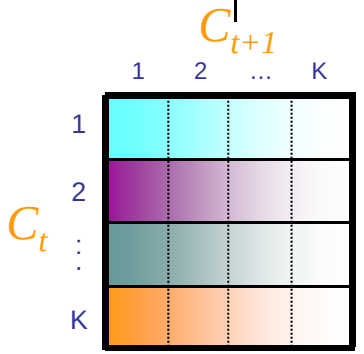


Individual chromosomes



The Hidden Markov Model





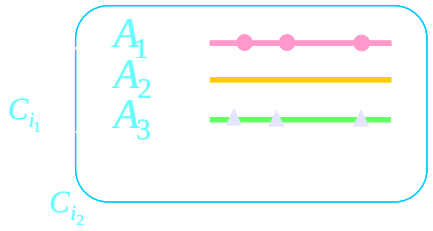
- Transition process: recombination

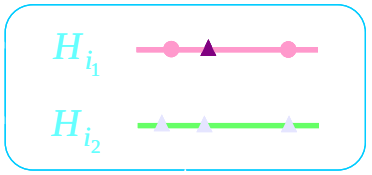
$$p(c_{i,t+1} = k' \mid c_{i,t} = k) = e^{-dr}\pi_{k,k'} + (1 - e^{-dr})\delta(k,k')$$

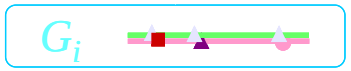
- Emission process: mutation

$$p(h_{i,t} \mid a_{k,t}, \theta_k) = \theta_k^{\,I(h_{i,t}=a_{k,t})}\left(\frac{1-\theta_k}{|B|-1}\right)^{I(h_{i,t}\neq a_{k,t})}$$

How many recombining ancestors?

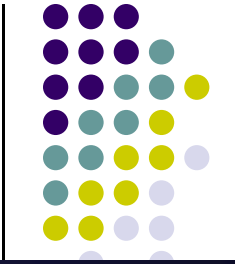




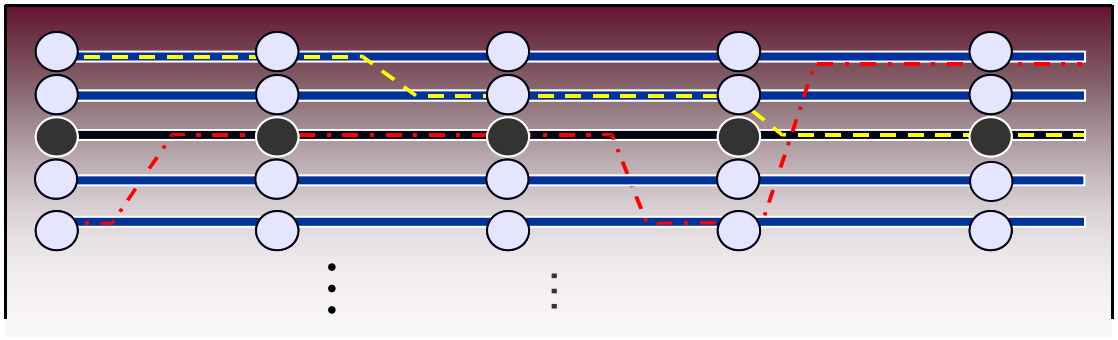


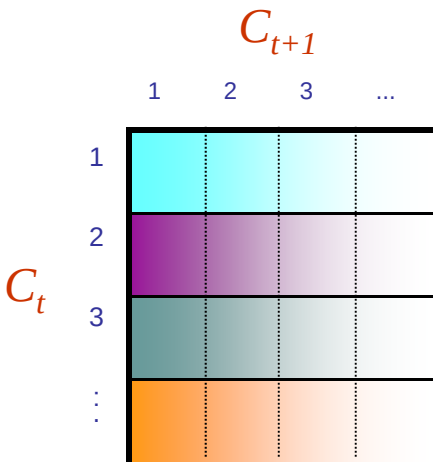
Hidden Markov Dirichlet Process

(Xing and Sohn, Bayesian Analysis, 2007)



- Hidden Markov Dirichlet process mixtures
  - Extension of HMM model to infinite ancestral space
    - Infinite dimensional transition matrix
    - Each row of the transition matrix is modeled with a DP



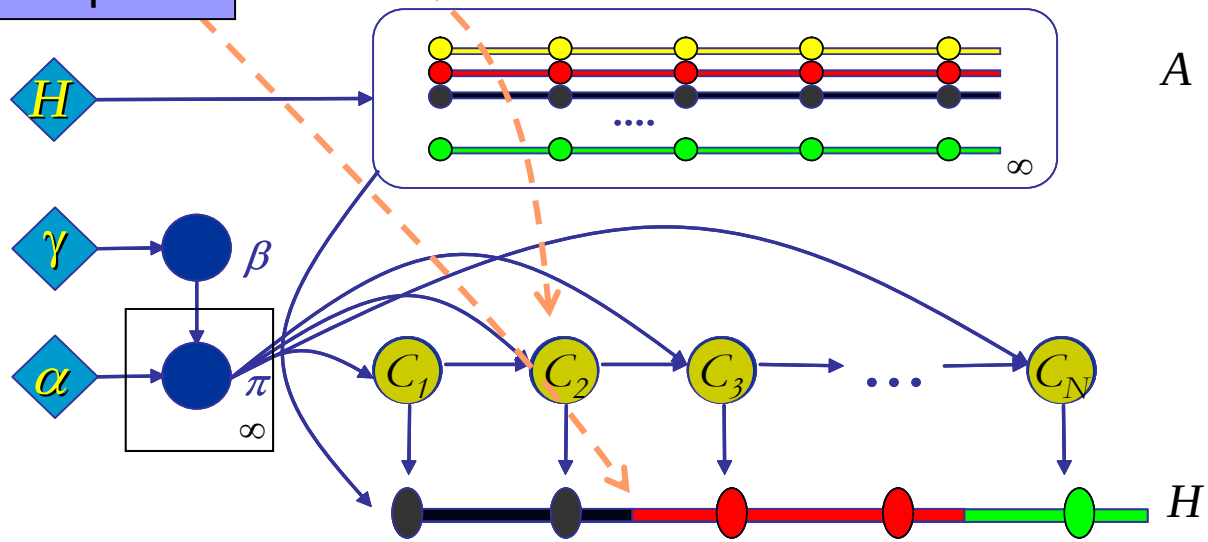


HMDP as a Graphical Model

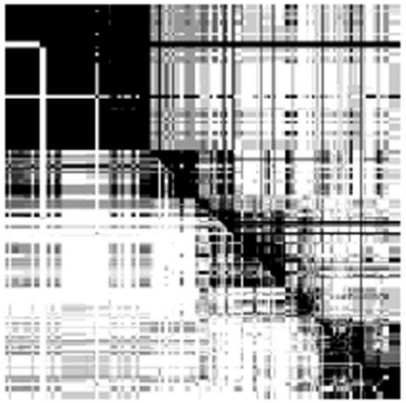
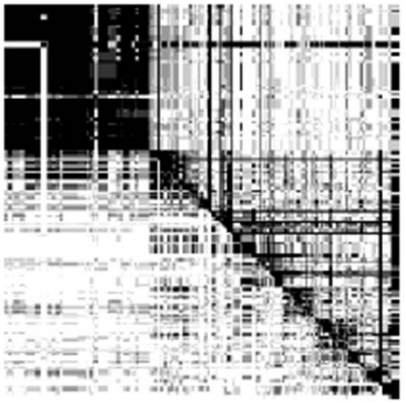
Ancestor allele reconstruction

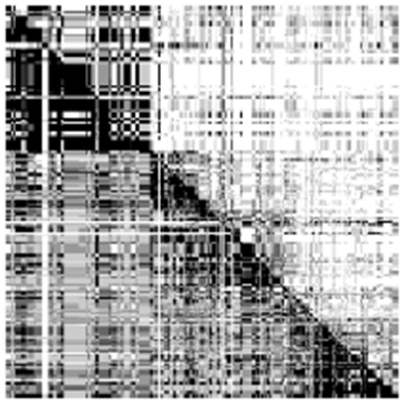
Inferring population structure

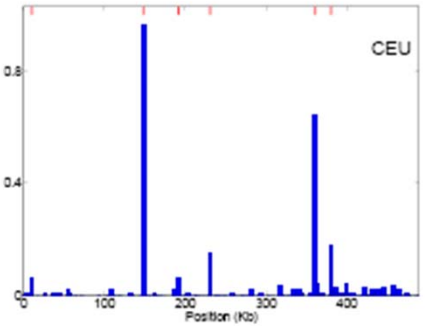
Inferring recombination hotspot

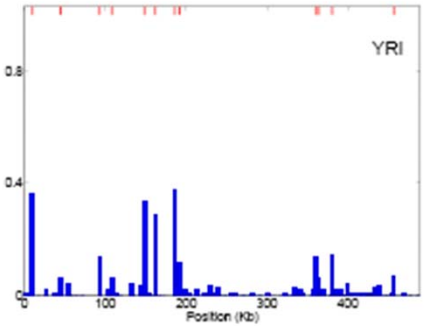


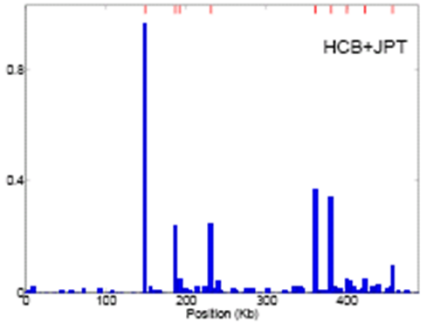
Recombination Analysis

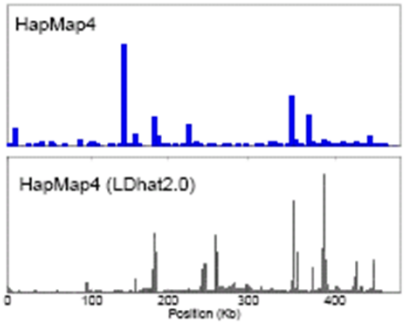





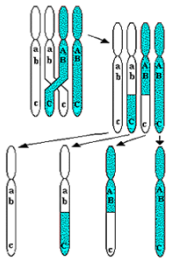








CEU YRI HCB+JPT HapMap4



		Stepcrum			*LDhat 2.0*			HMM ($K = 5$)		
	w_{tol}	0	± 1	± 2	0	± 1	± 2	0	± 1	± 2
ω=	FPR	0.16	0.11	0.07	0.19	0.09	0.06	0.18	0.12	0.11
3rd quartile	FNR	0.11	0	0	0.22	0.11	0.11	0.33	0.11	0.11
ω s.t.	FPR	0.16	0.11	0.07	0.22	0.11	0.07	0.18	0.12	0.11
FNR$\sim$FAR	FNR	0.11	0	0	0.22	0.12	0.11	0.33	0.11	0.11