Advanced Introduction to Machine Learning

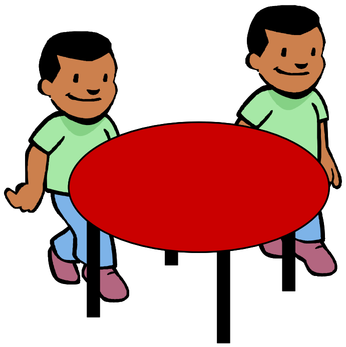
10715, Fall 2014

Nonparametric Bayesian Models

--Learning/Reasoning in Open Possible Worlds

Eric Xing

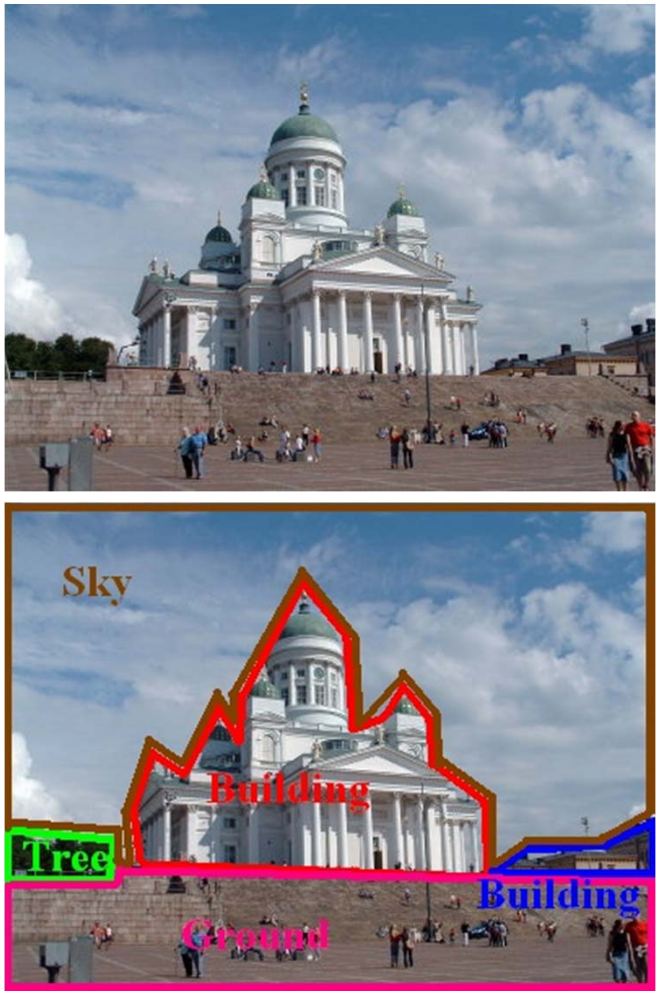
Lecture 19, November 12, 2014



Reading:

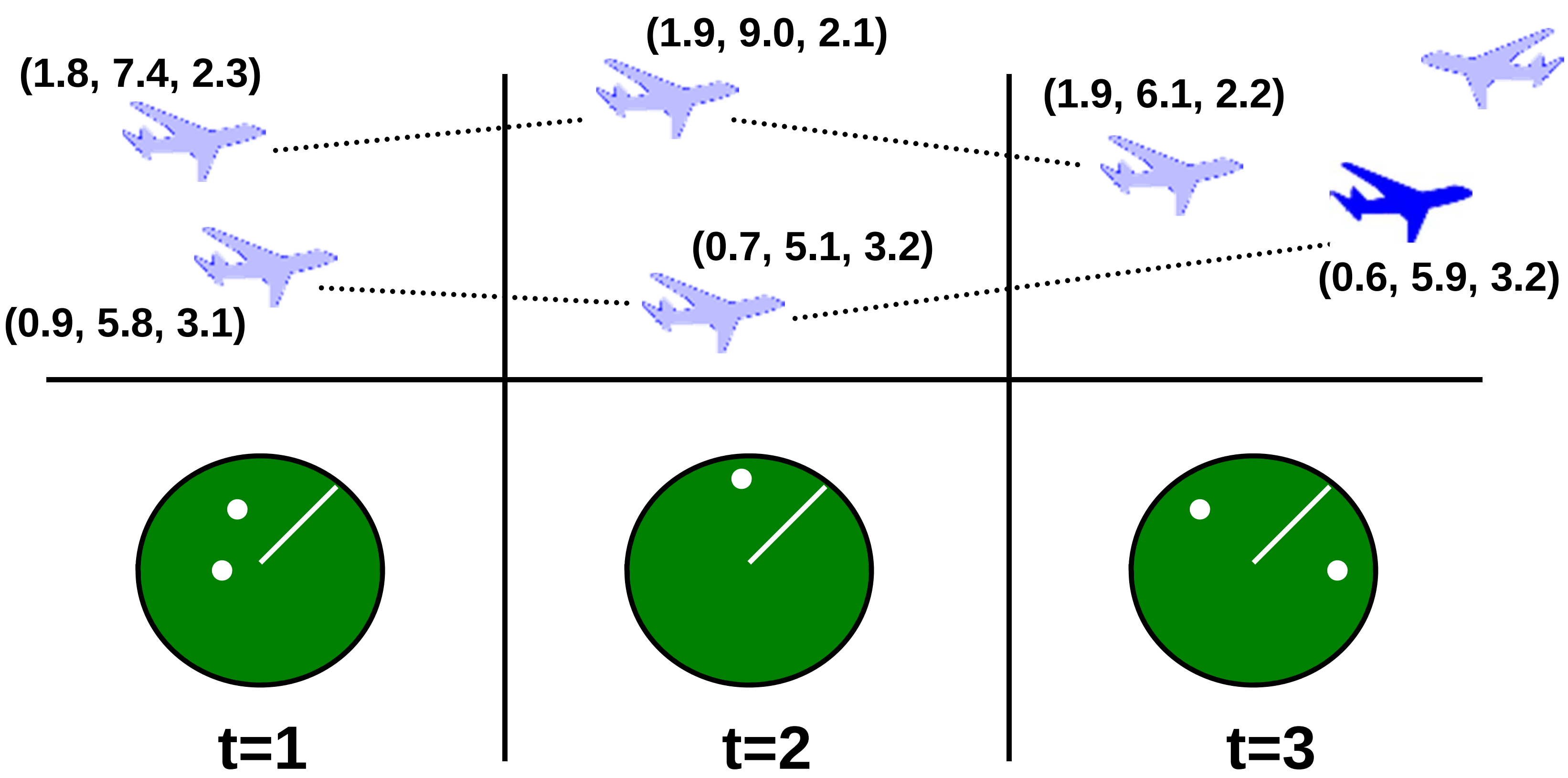
Clustering

Image Segmentation

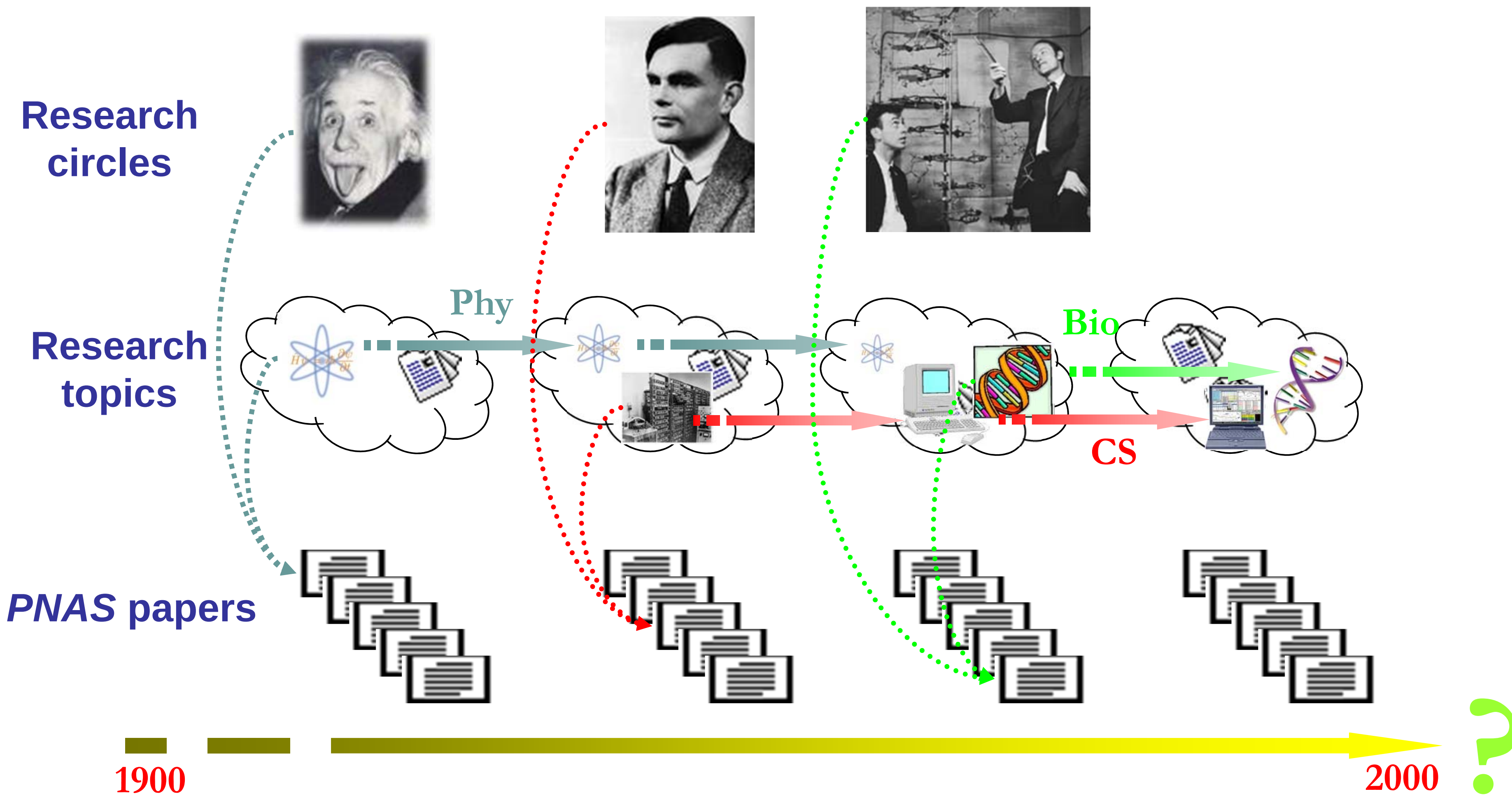


- How to segment images?
  - Manual segmentation (very expensive)
  - Algorithm segmentation
    - K-means
    - Statistical mixture models
    - Spectral clustering
- Problems with most existing algorithms
  - Ignore the spatial information
  - Perform the segmentation one image at a time
  - Need to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

Object Recognition and Tracking

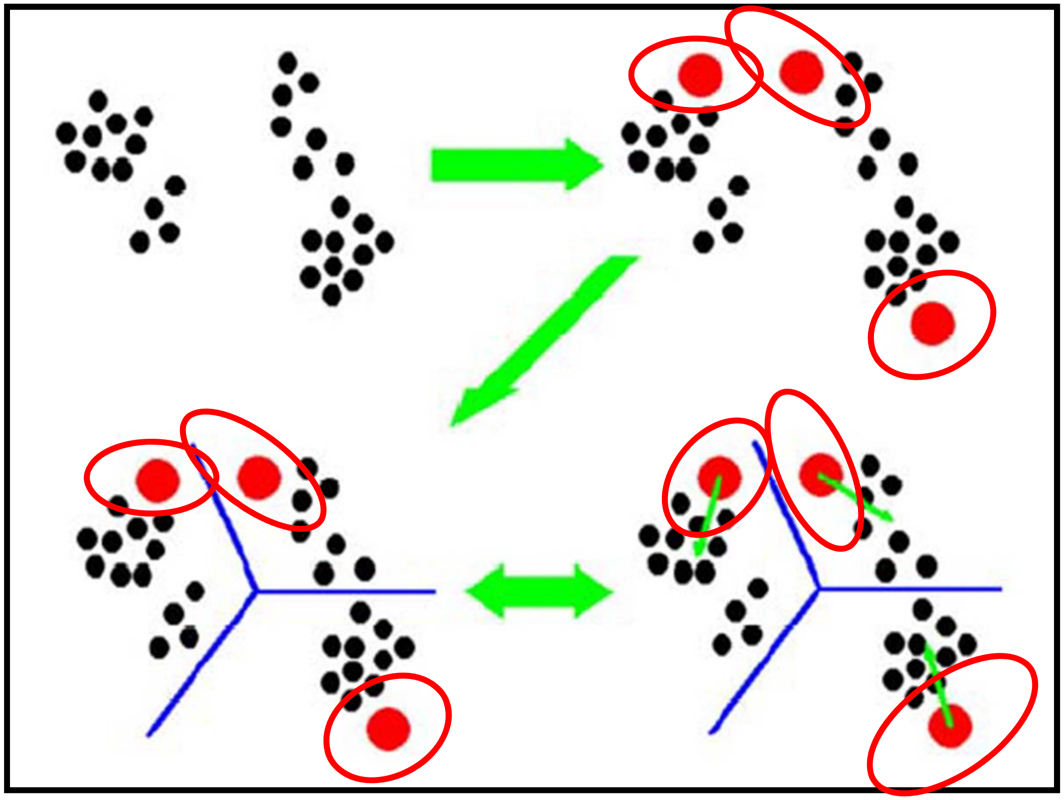


The Evolution of Science



A Classical Approach

- Clustering as Mixture Modeling



- Then "model selection"

Model Selection vs. Posterior Inference

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

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

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

- Posterior inference:

  we want to handle uncertainty of model complexity explicitly

$$p(M \mid D) \propto p(D \mid M) p(M)$$

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

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

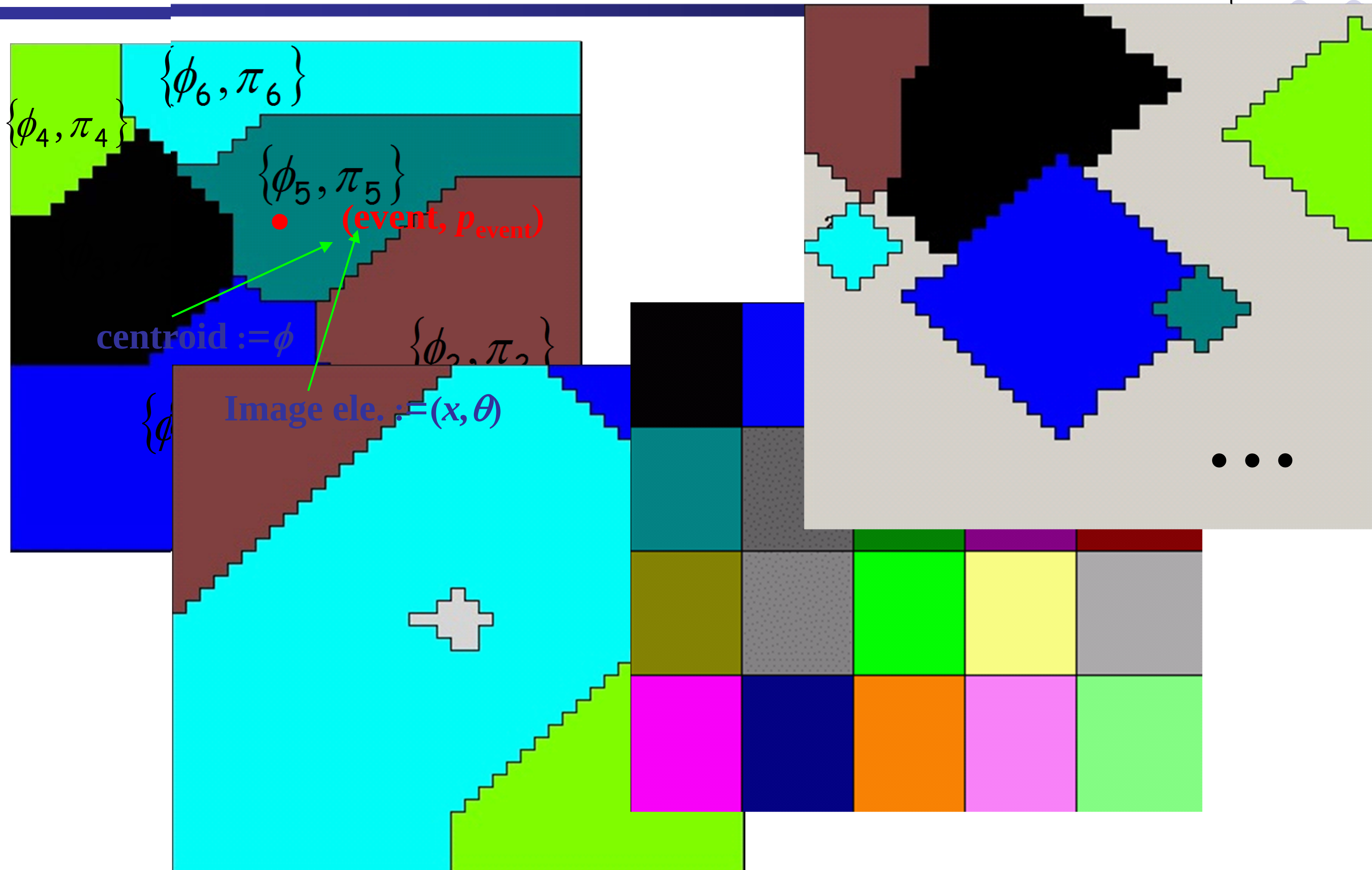
Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Application: joint multiple population clusteri
- Dynamic Dirichlet Process
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary

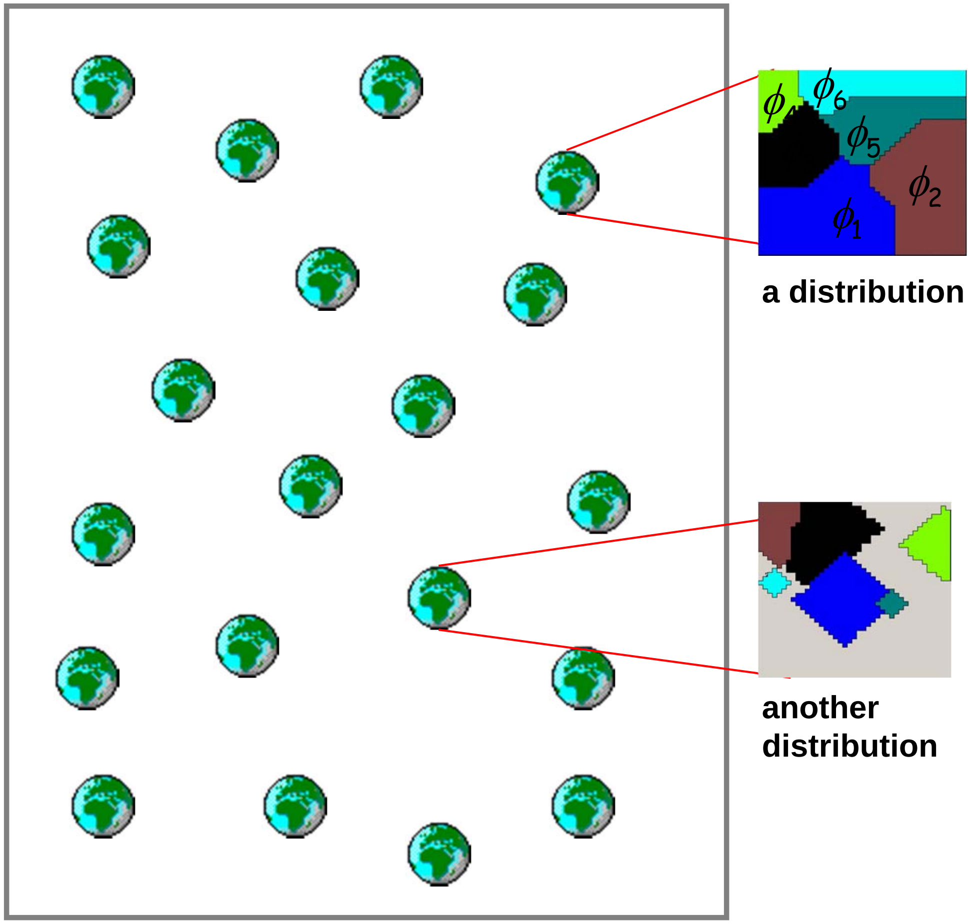
Clustering

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

Random Partition of Probability Space



Dirichlet Process



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

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

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

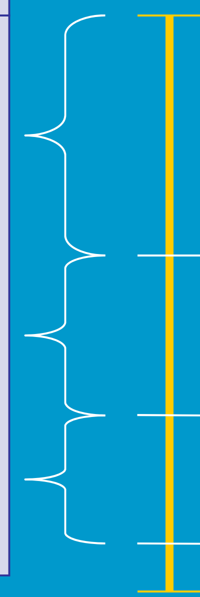
$$\sum_{k=1}^{\infty} \pi_k = 1$$

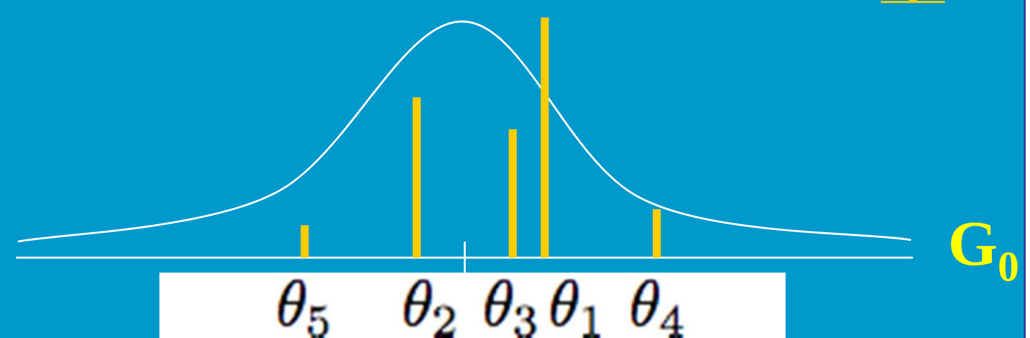
$$\pi_k = \beta_k \prod_{j=1}^{k-1} (1 - \beta_k)$$

Mass

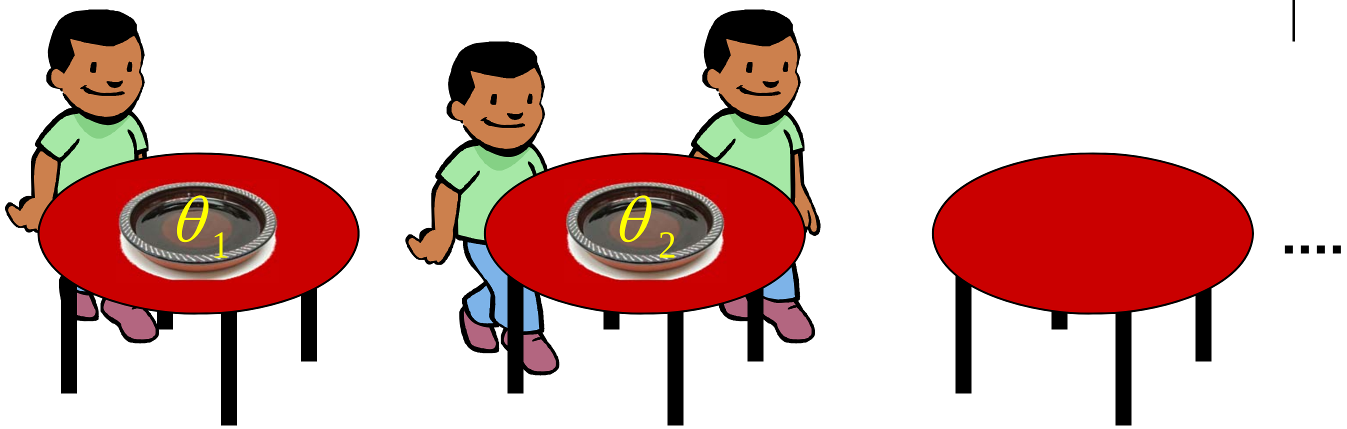
$$\beta_k \sim \mathrm{Beta}(1, \alpha)$$

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





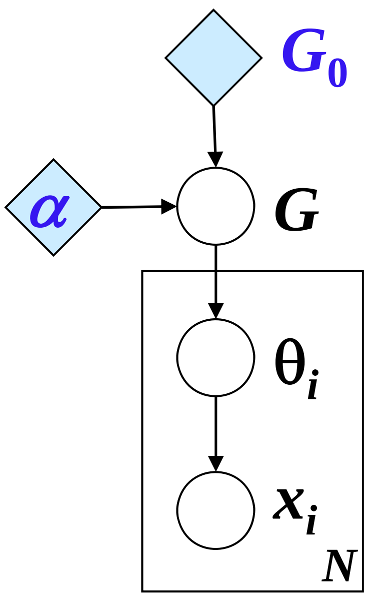
Chinese Restaurant Process



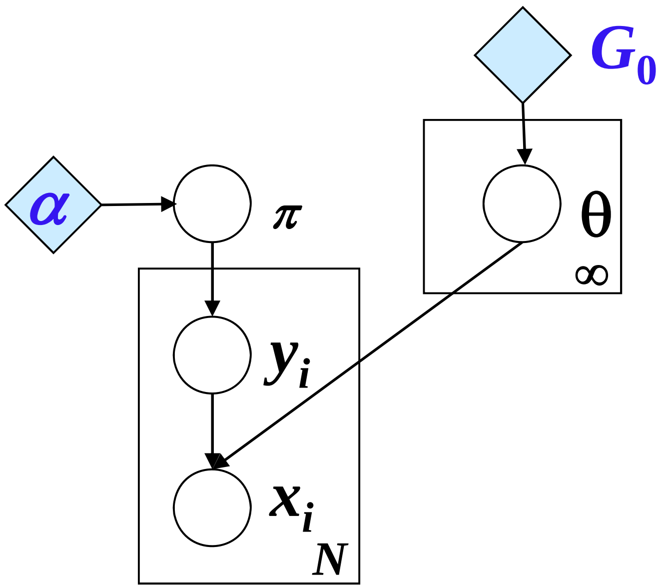
$$
P(c_i = k \mid \mathbf{c}_{-i}) = \begin{array}{cccc}
1 & 0 & & 0 \\
\frac{1}{1+\alpha} & \frac{\alpha}{1+\alpha} & & 0 \\
\frac{1}{2+\alpha} & \frac{1}{2+\alpha} & & \frac{\alpha}{2+\alpha} \\
\frac{1}{3+\alpha} & \frac{2}{3+\alpha} & & \frac{\alpha}{3+\alpha} \\
\frac{m_1}{i+\alpha-1} & \frac{m_2}{i+\alpha-1}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

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

Graphical Model Representations of DP

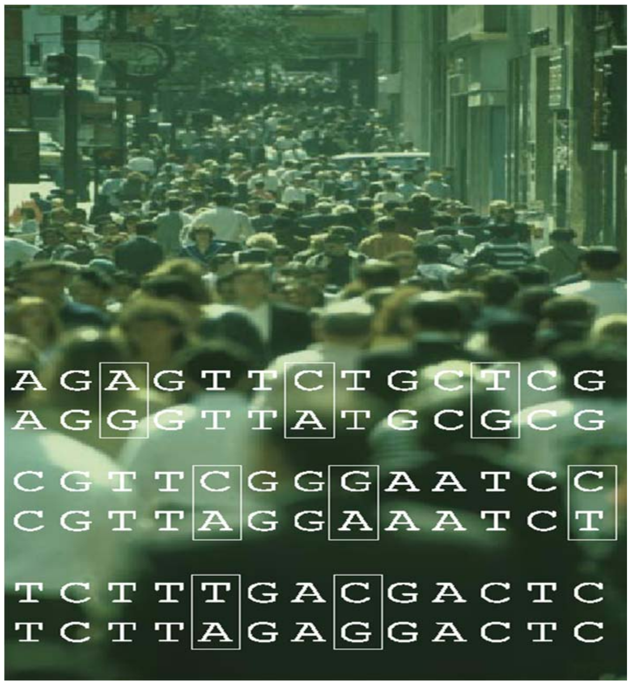


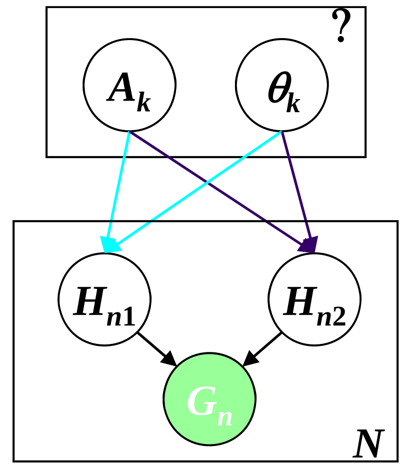
The CRP construction



The Stick-breaking construction

Ancestral Inference



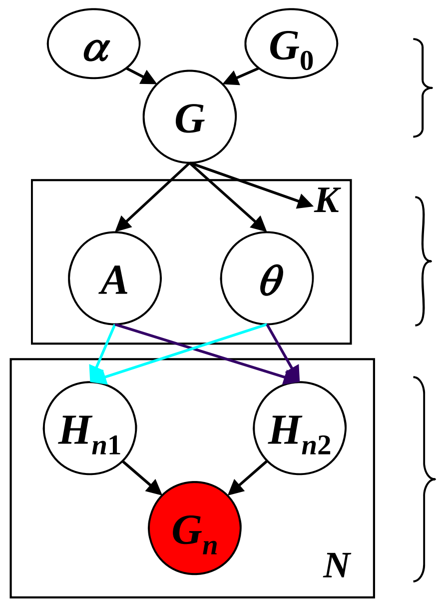


Essentially a clustering problem, but …

- 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

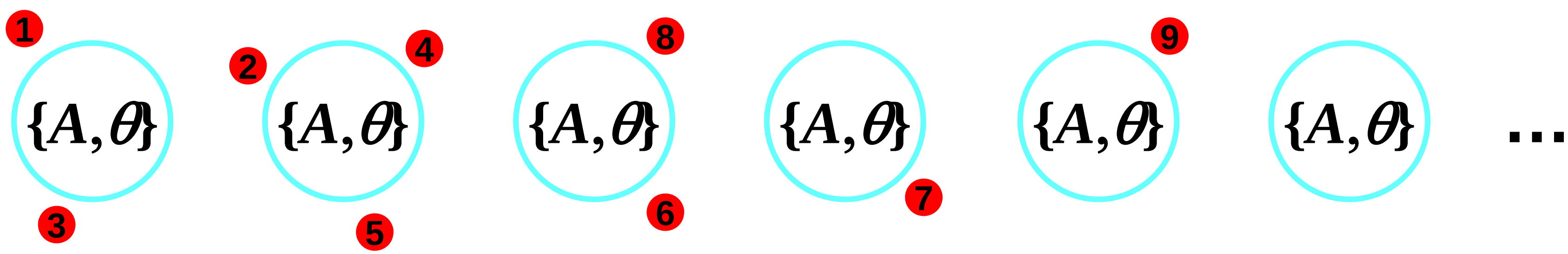
infinite mixture components (for population *haplotypes*)

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

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

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

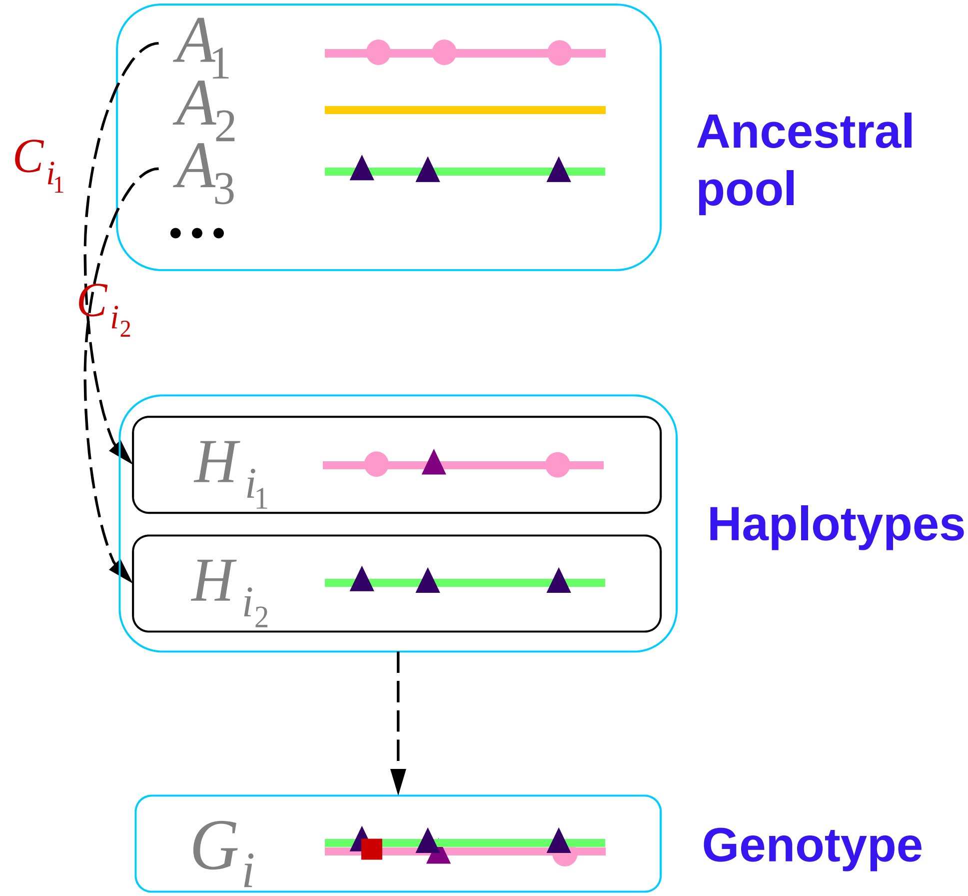
$$\rightarrow h_t = a_t \quad with\ prob.\ \theta$$

- Noisy observation model

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

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

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



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

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

2. Sample a_k from

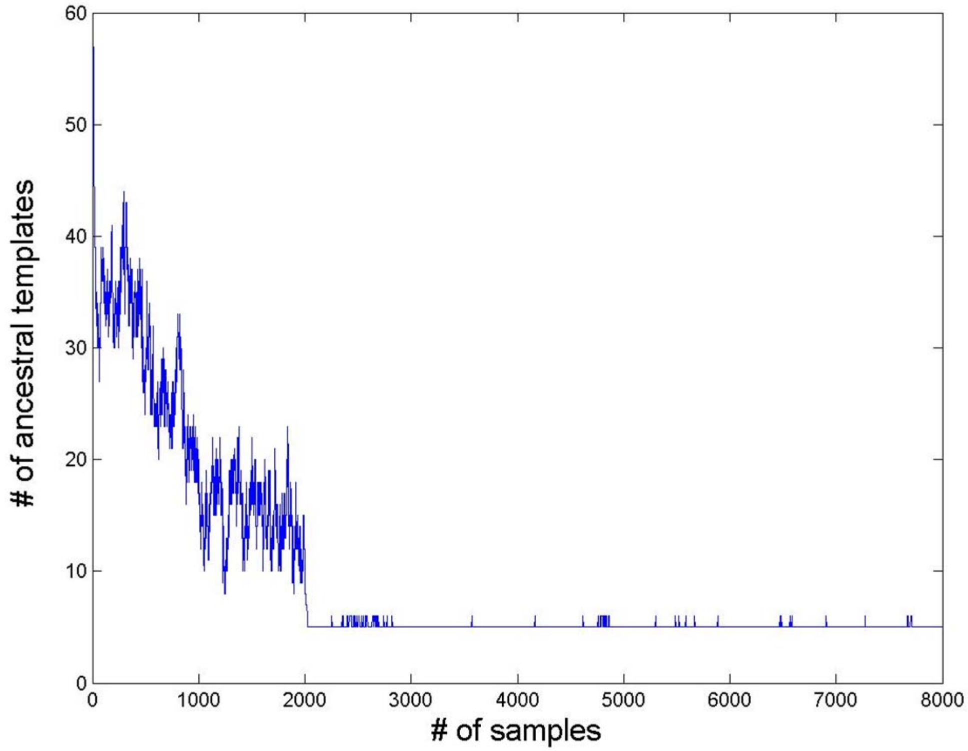
$$p(a_{k,t} | \mathbf{c}, \mathbf{h}) \propto \prod_{j, i_e | c_{i_e,t}^{(j)} = k} p(h_{i_e,t}^{(j)} | a_{k,t}, l_{k,t}^{(j)})$$
$$= \frac{\Gamma(\alpha_h + l_{k,t})\Gamma(\beta_h + l'_{k,t})}{\Gamma(\alpha_h + \beta_h + m_k)(|B| - 1)^{l'_{k,t}}} R(\alpha_h, \beta_h)$$

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

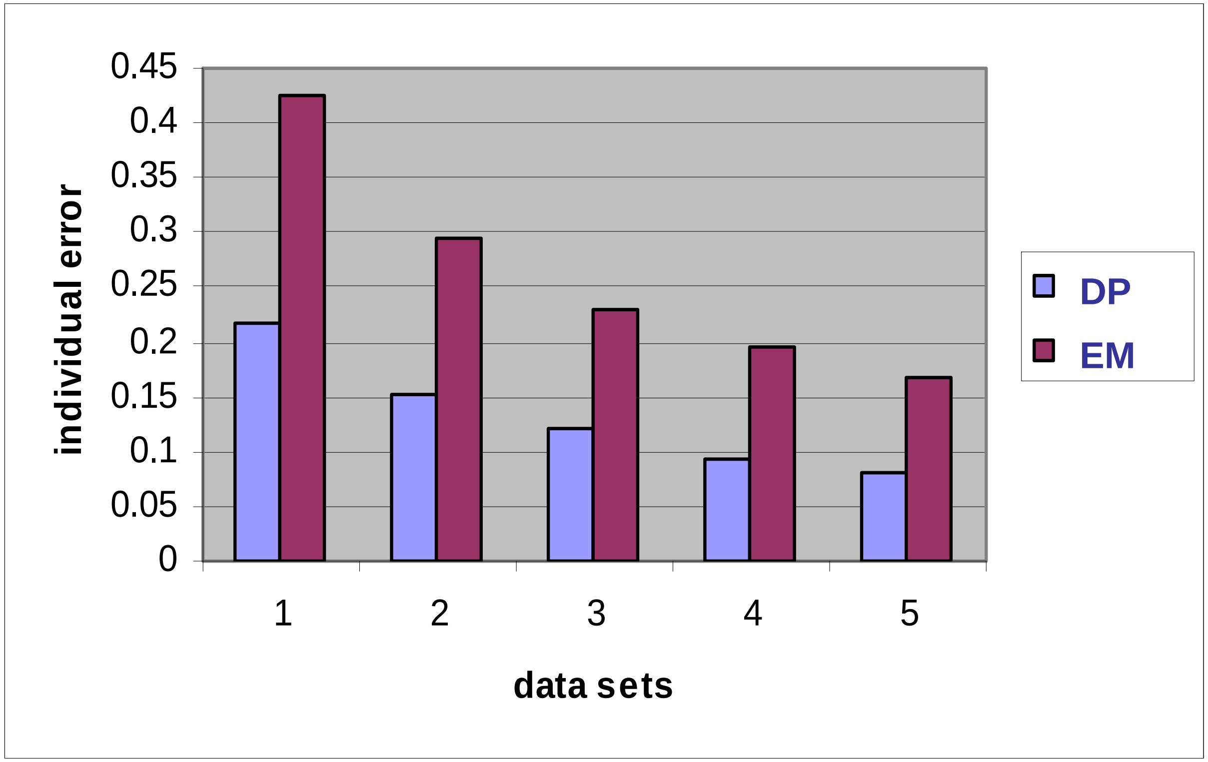
$$p(h_{i_e,t}^{(j)} | \mathbf{h}_{[-i_e,t]}^{(j)}, \mathbf{c}, \mathbf{a}, \mathbf{g})$$

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

Convergence of Ancestral Inference



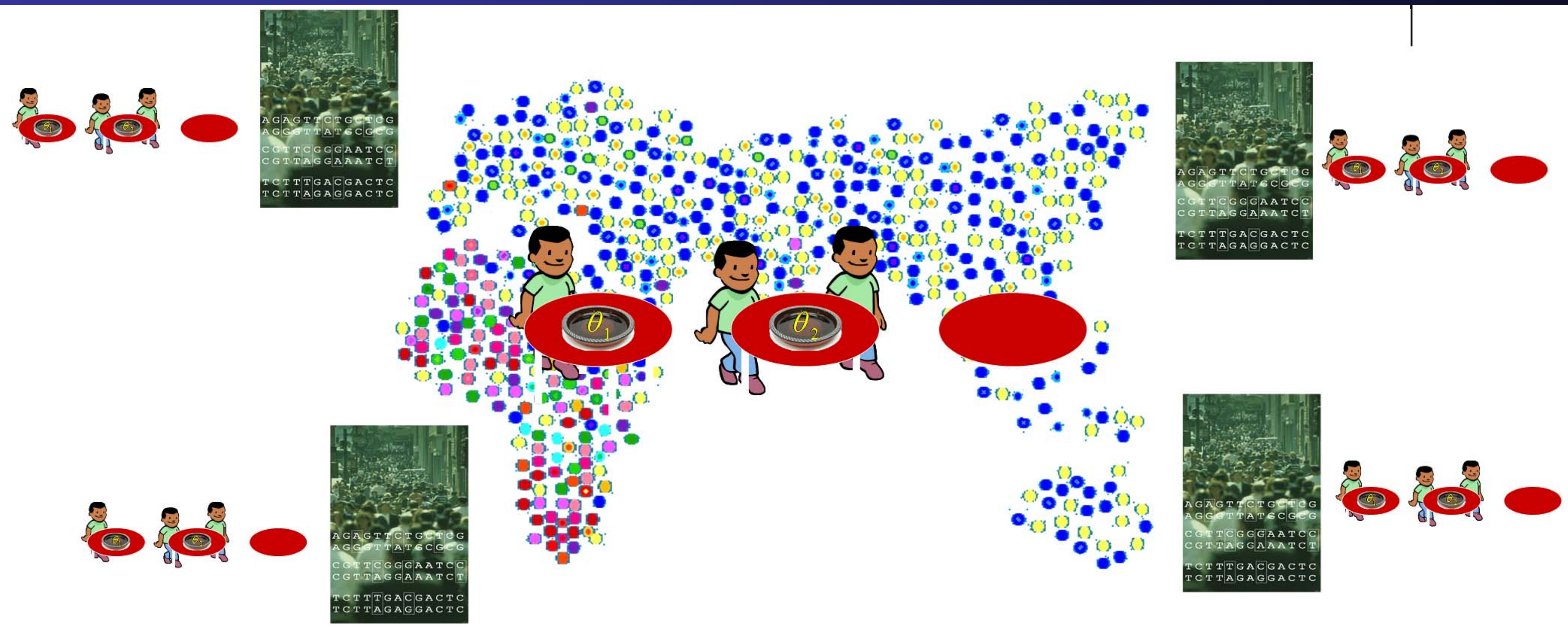
DP vs. Finite Mixture via EM



Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Application: joint multiple population clustering
- Dynamic Dirichlet Process
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary

Multi-population Genetic Demography

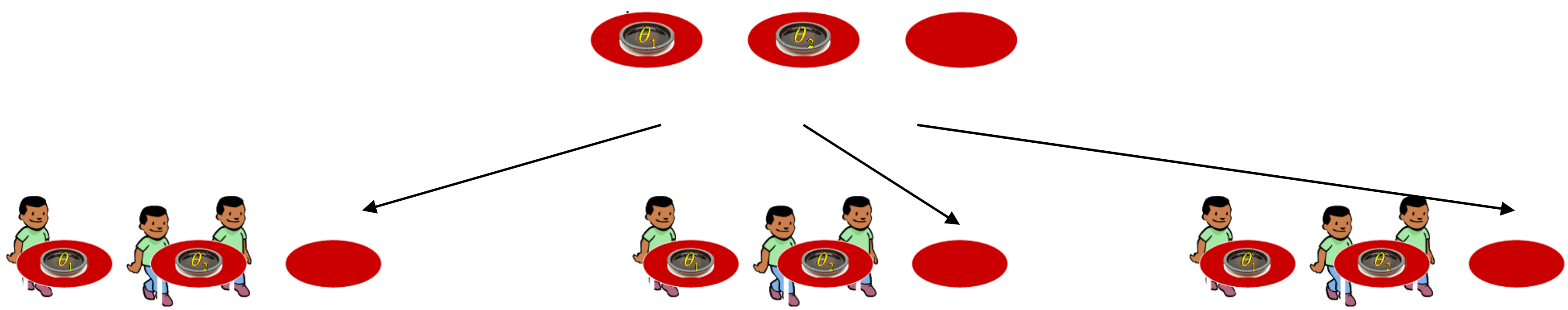


- Pool everything together and solve 1 hap problem?
  - --- ignore population structures
- Solve 4 hap problems separately?
  - --- data fragmentation
- Co-clustering … solve 4 *coupled* hap problems jointly

Hierarchical Dirichlet Process

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

- **Two level Pólya urn scheme**
  - At the *i*-th step in *j*-th "group",



$$-\,Choose\ \theta_k\ with\ prob.\ \frac{m_{jk}}{\sum_k m_{jk} + \alpha_0}$$

$$-\,Go\ to\ the\ upper\ level\ DP$$

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

Oracle

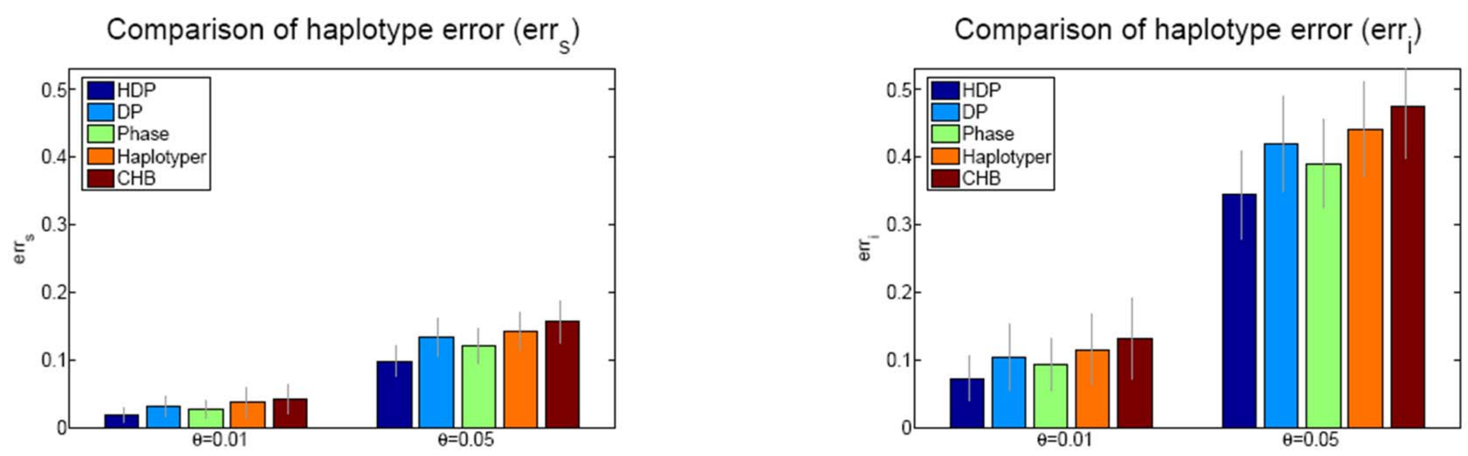
$$Choose\ \theta_k\ with\ prob.\ \frac{n_k}{\sum n_k + \gamma}$$

$$Draw\ a\ new\ sample$$

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

Results - Simulated Data

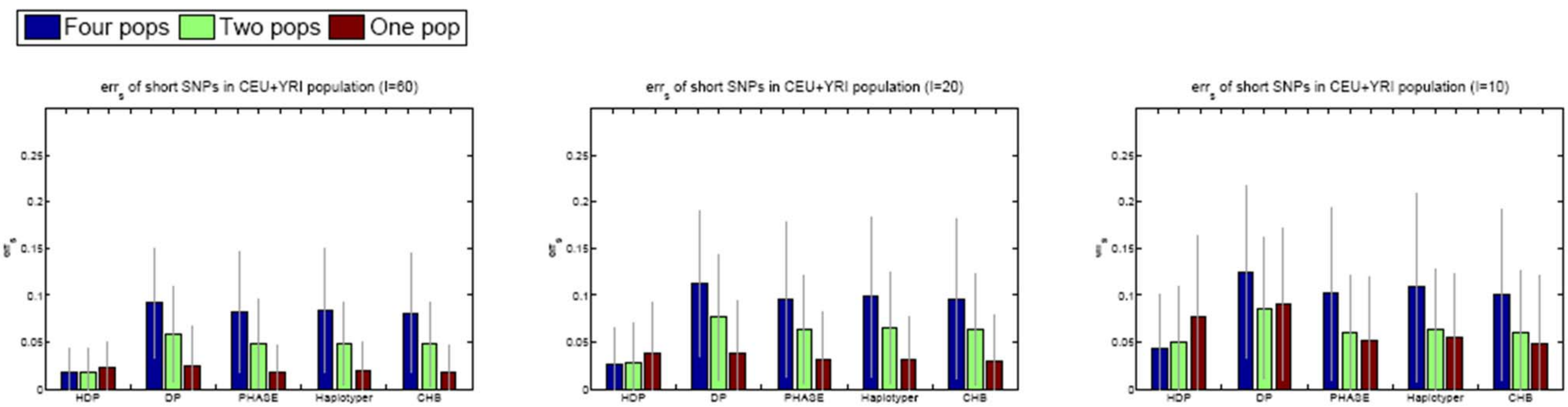
- 5 populations with 20 individuals each (two kinds of mutation rates)
- 5 populations share parts of their ancestral haplotypes
- the sequence length = 10



Haplotype error

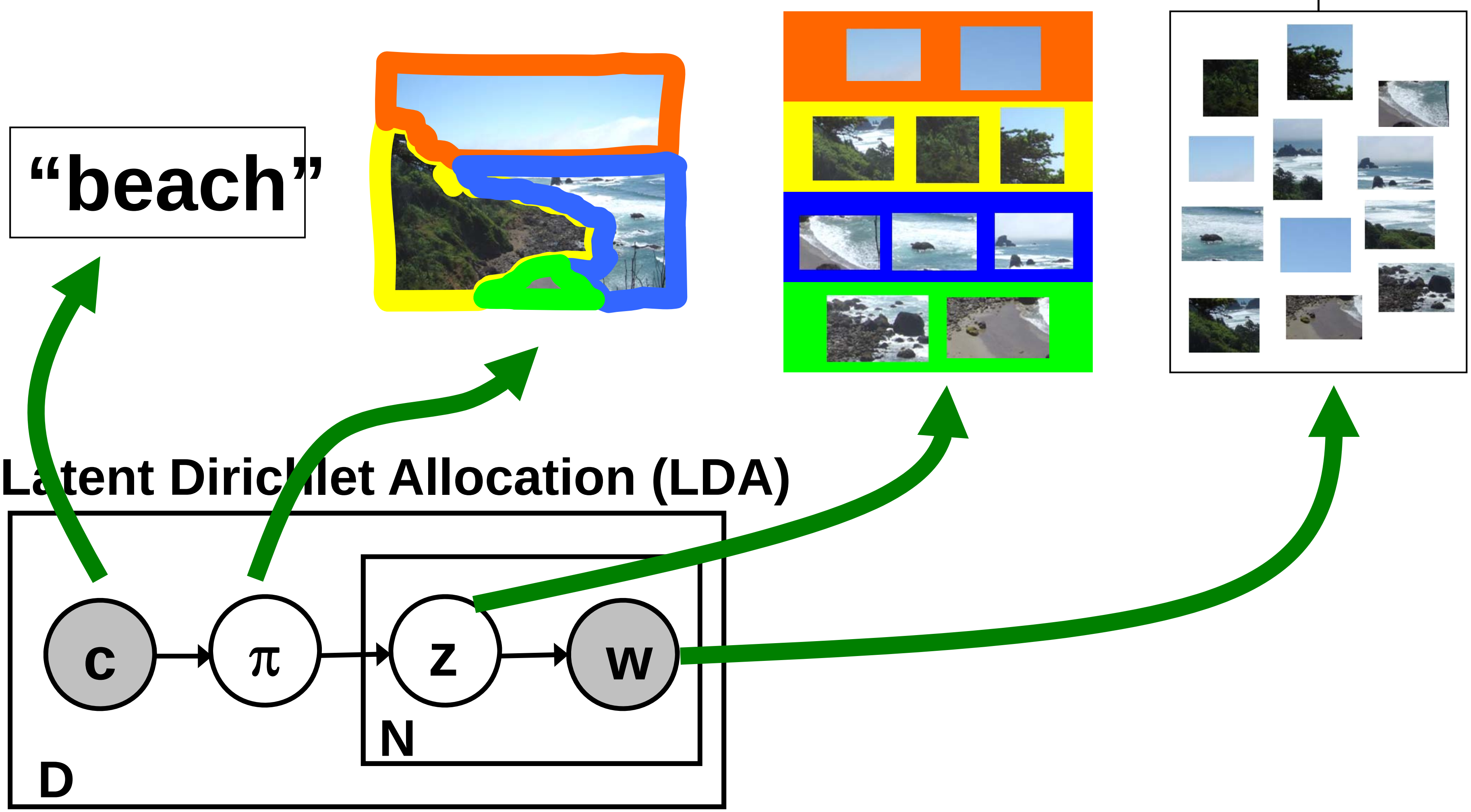
Results - International HapMap DB

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

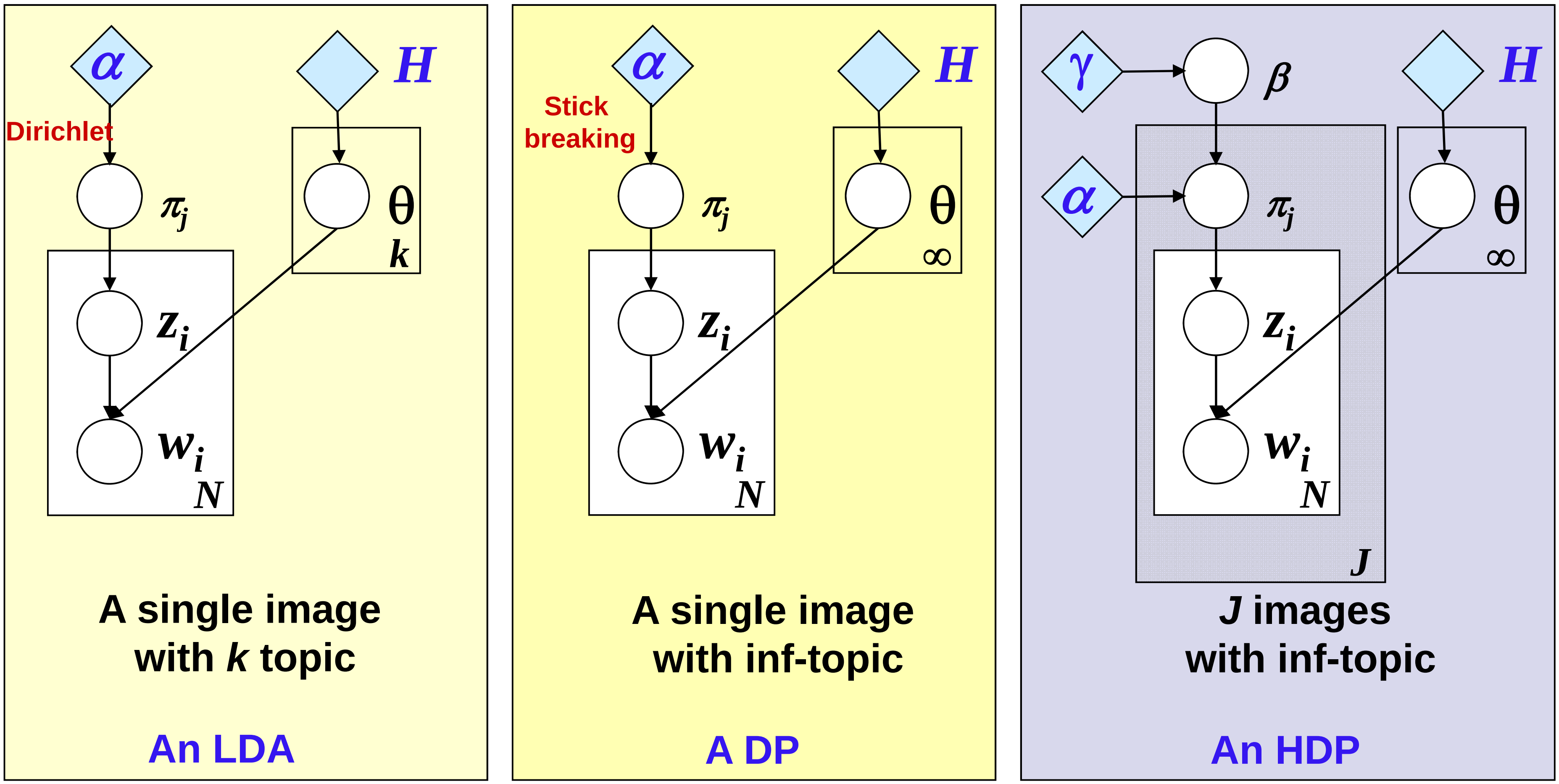


$I_{each} = 60$ $I_{each} = 20$ $I_{each} = 10$

Topic Models for Images



Infinite Topic Model for Image

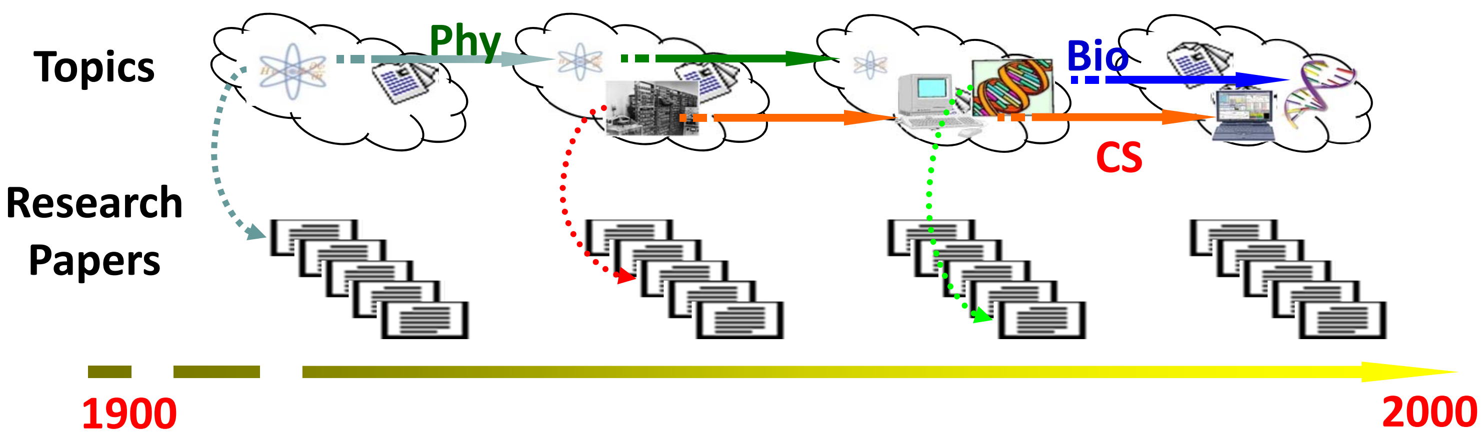


Outline

- Motivation and challenge
- Dirichlet Process and Infinite Mixture
  - Formulation
  - Approximate Inference algorithm
  - Example: population clustering
- Hierarchical Dirichlet Process and Multi-Task Clustering
  - Formulation
  - Application: joint multiple population clustering
- Dynamic Dirichlet Process
  - Temporal DPM
  - Application: evolutionary clustering of documents
- Summary

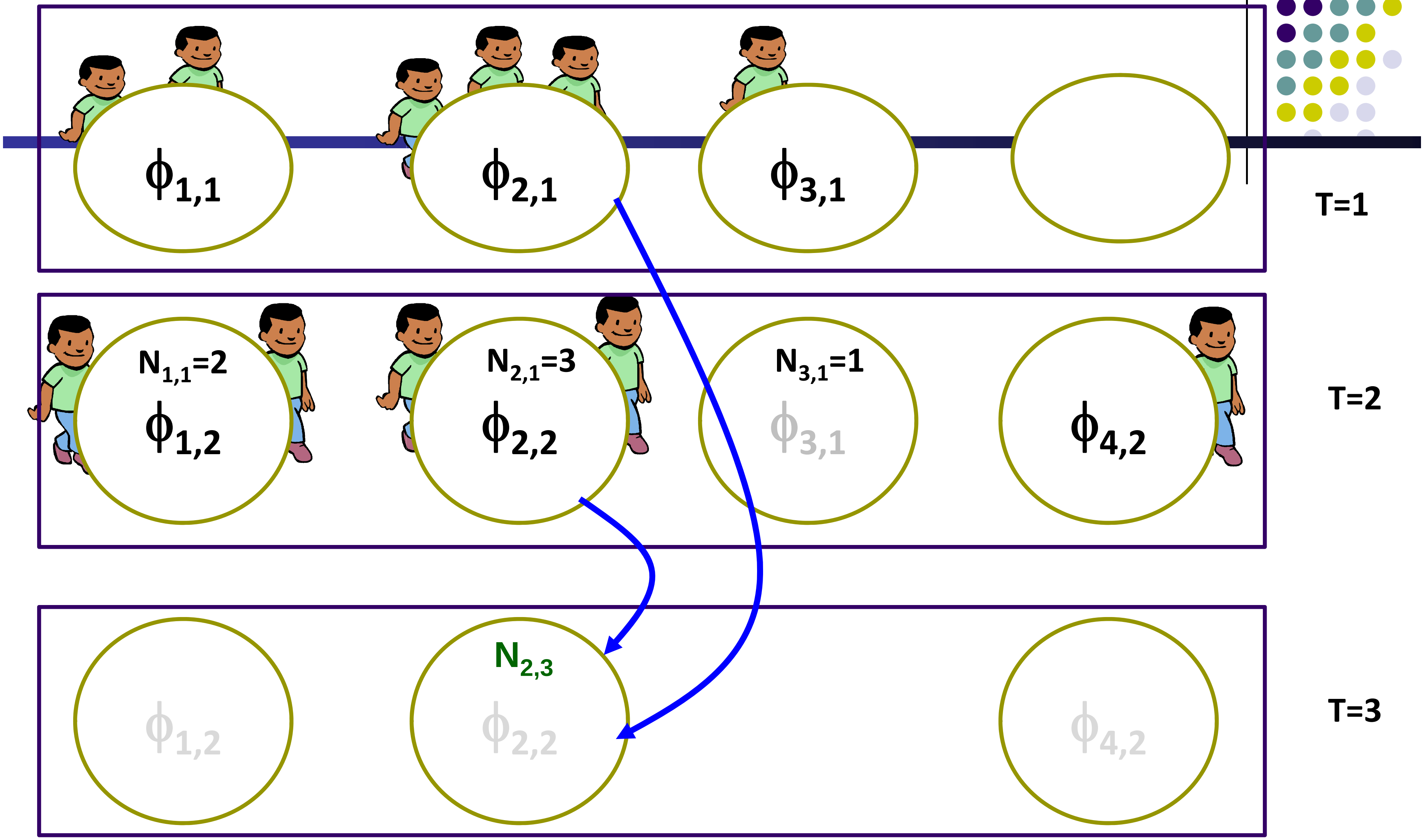
Evolutionary Clustering

- 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



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.



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

TDPM Generative Power

DPM

W=T

$\lambda = \infty$

TDPM

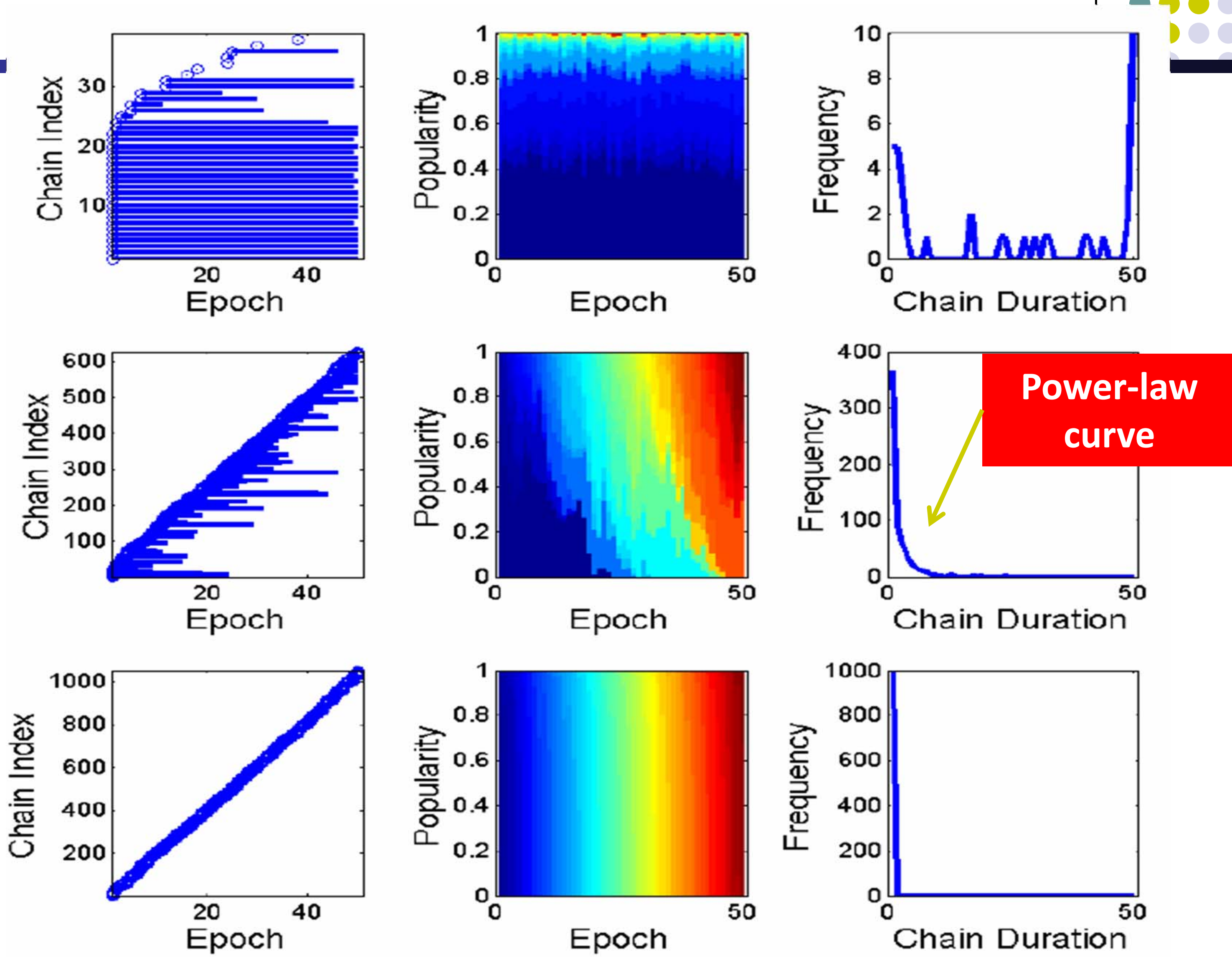
W=4

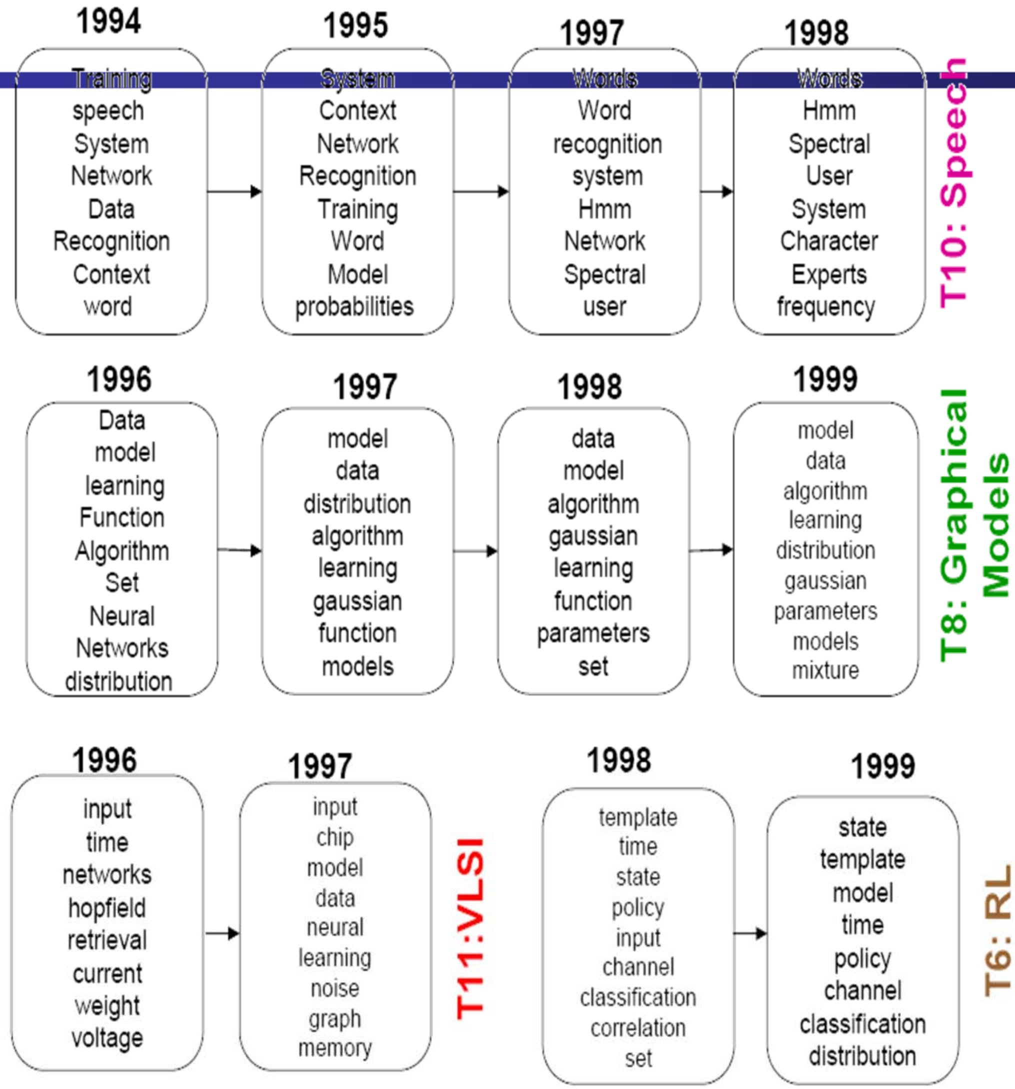
$\lambda = .4$

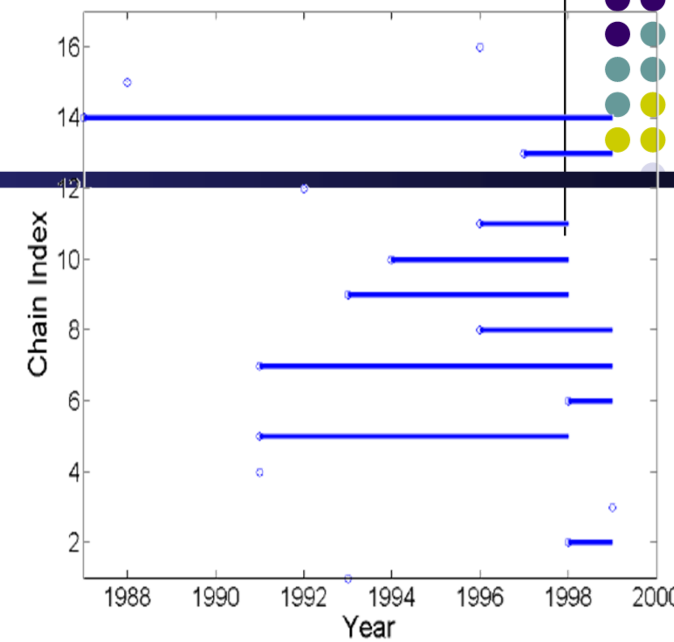
Independent DPMs

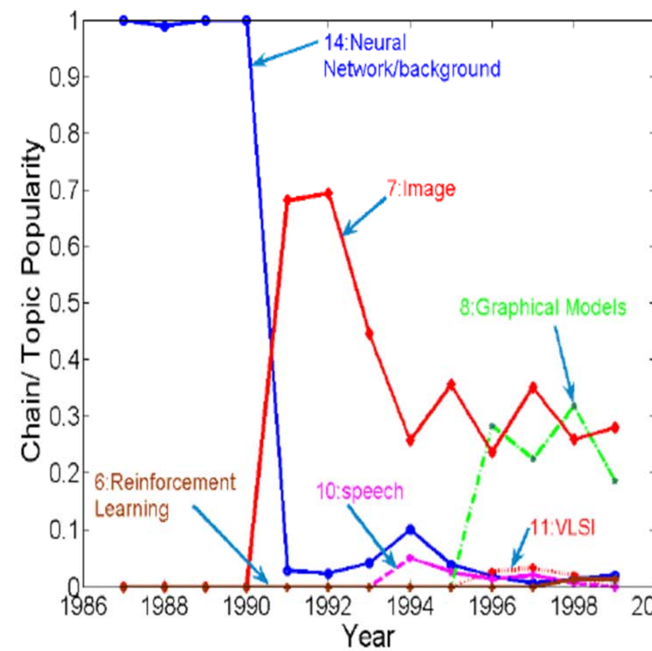
W= 0

$\lambda = ?$ (any)

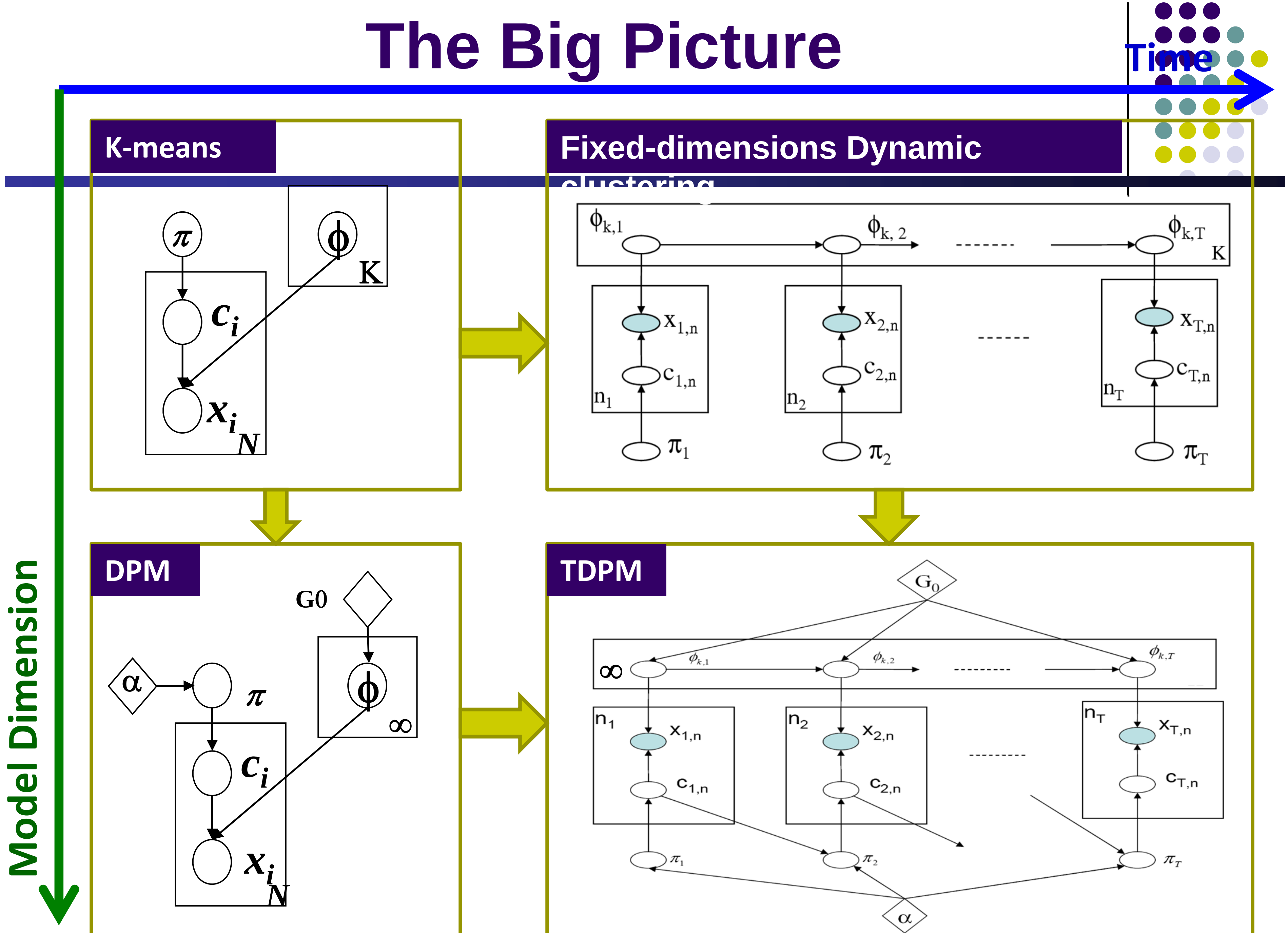



1994
Training speech System Network Data Recognition Context word
1995
System Context Network Recognition Training Word Model probabilities
1997
Words Word recognition system Hmm Network Spectral user
1998
Words Hmm Spectral User System Character Experts frequency
T10: Speech
1996
Data model learning Function Algorithm Set Neural Networks distribution
1997
model data distribution algorithm learning gaussian function models
1998
data model algorithm gaussian learning function parameters set
1999
model data algorithm learning distribution gaussian parameters models mixture
T8: Graphical Models
1996
input time networks hopfield retrieval current weight voltage
1997
input chip model data neural learning noise graph memory
T11:VLSI
1998
template time state policy input channel classification correlation set
1999
state template model time policy channel classification distribution
T6: RL


Chain Index
Year
16
14
12
10
8
6
4
2
1988
1990
1992
1994
1996
1998
2000


Chain/ Topic Popularity
Year
14:Neural Network/background
7:Image
8:Graphical Models
6:Reinforcement Learning
10:speech
11:VLSI
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1986
1988
1990
1992
1994
1996
1998
2000

The Big Picture

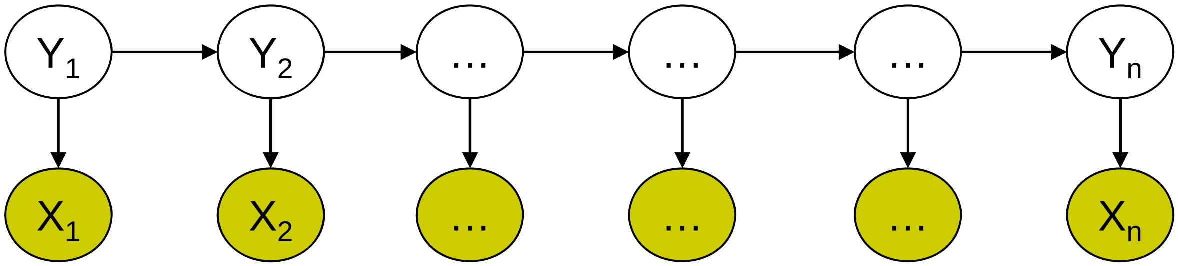


Summary

- 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 …

Shortcomings of Hidden Markov Model



- HMM models capture dependences between each state and only its corresponding observation
  - NLP example: In a sentence segmentation task, each segmental state may depend not just on a single word (and the adjacent segmental stages), but also on the (non-local) features of the whole line such as line length, indentation, amount of white space, etc.
- Mismatch between learning objective function and prediction objective function
  - HMM learns a joint distribution of states and observations $P(\mathbf{Y}, \mathbf{X})$, but in a prediction task, we need the conditional probability $P(\mathbf{Y}|\mathbf{X})$