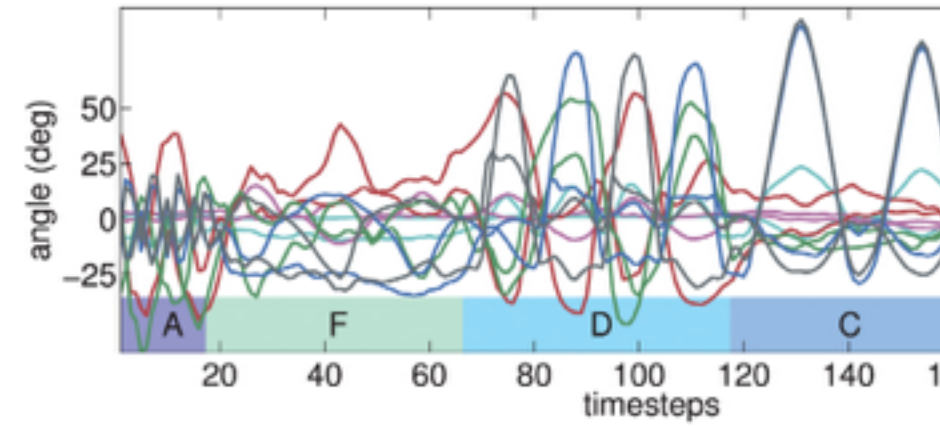
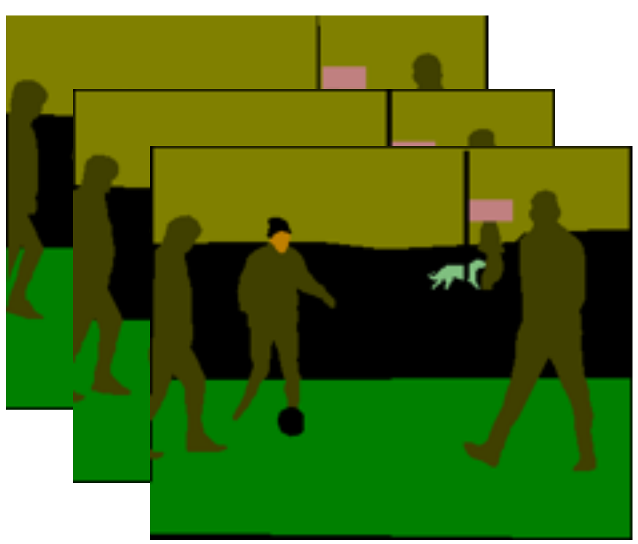
Bayesian Nonparametric Discovery of Layers and Parts from Scenes and Objects

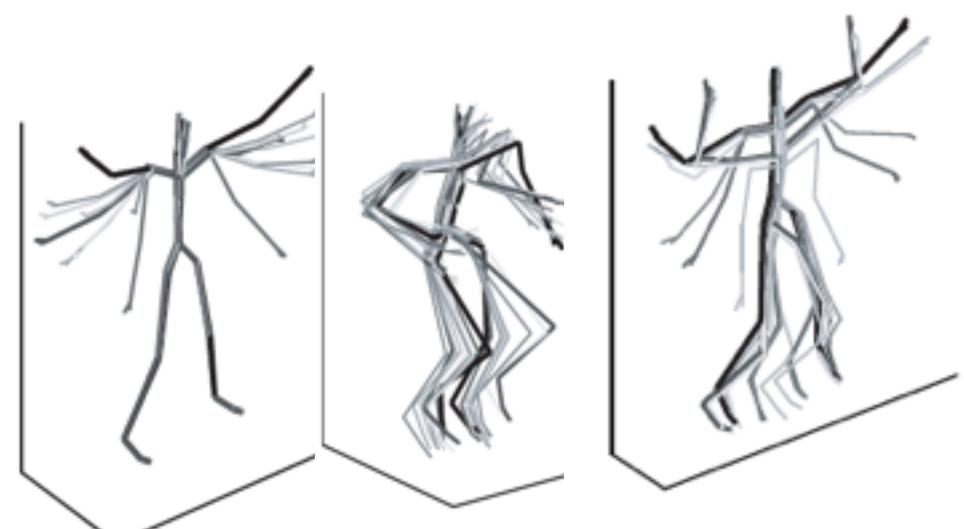
Soumya Ghosh

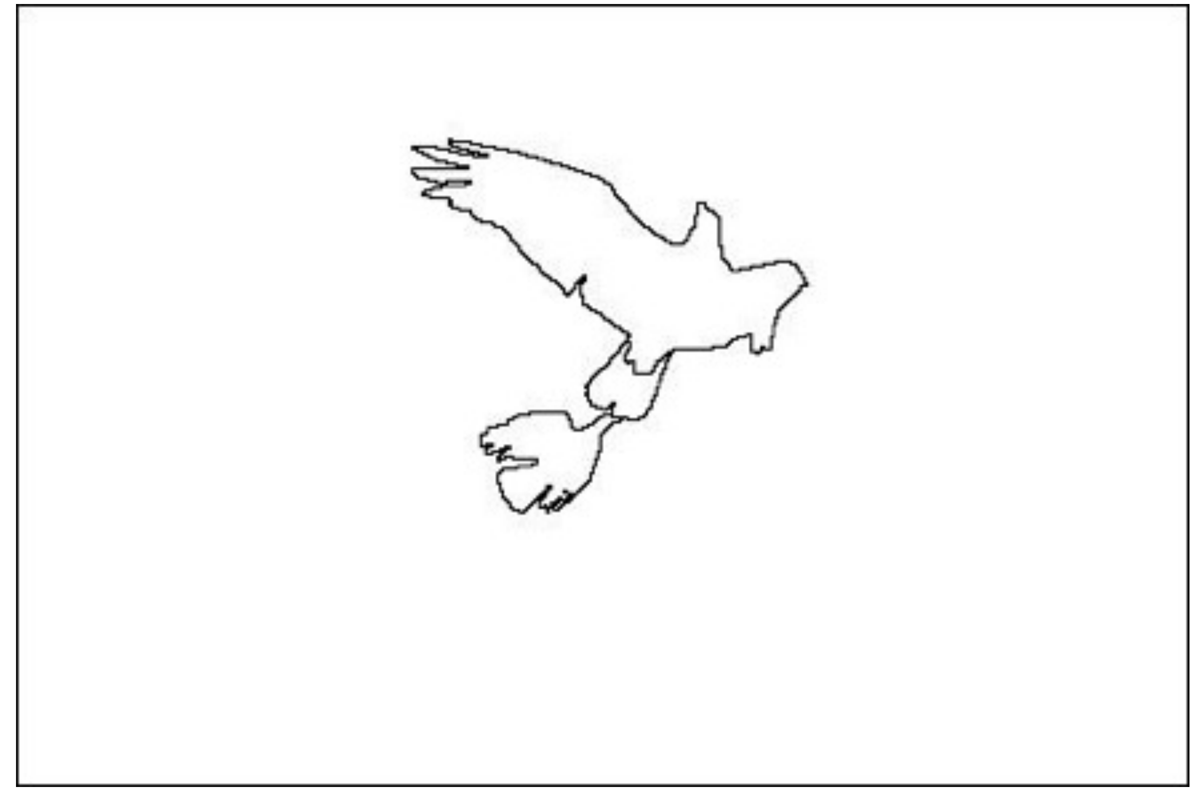
Advisor: Erik Sudderth

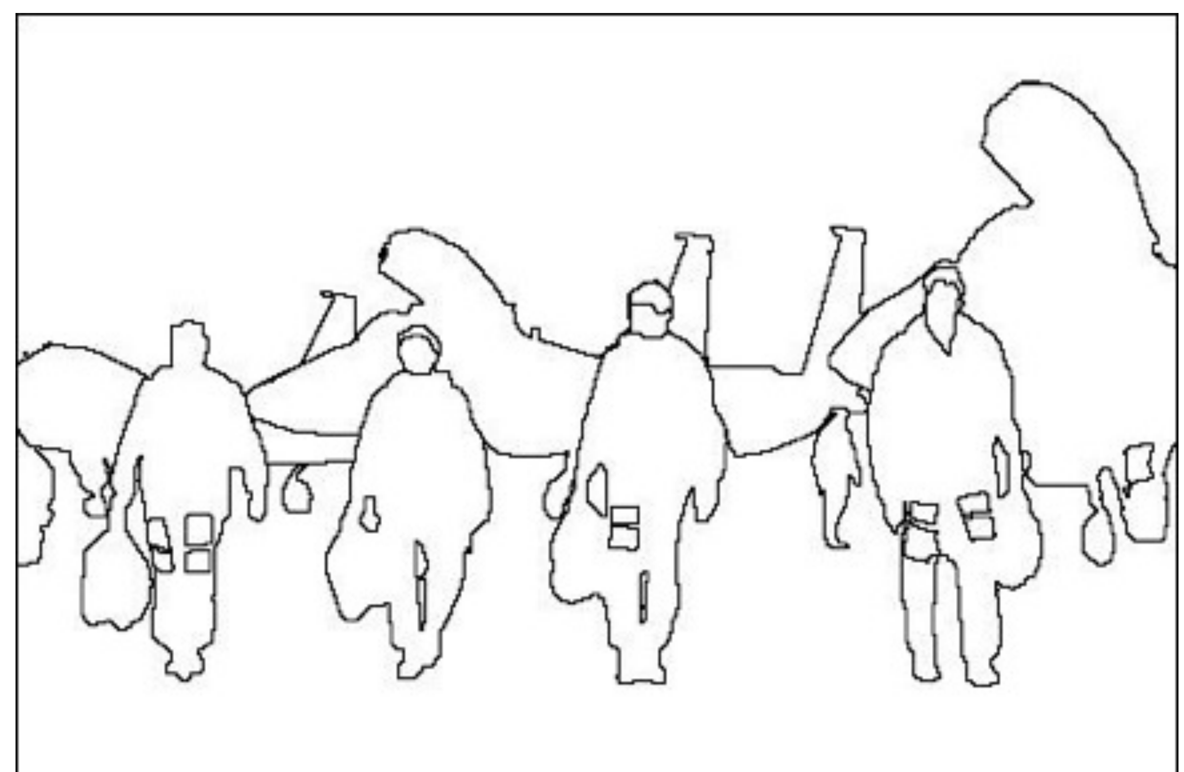
Committee: Michael Black and James Hays

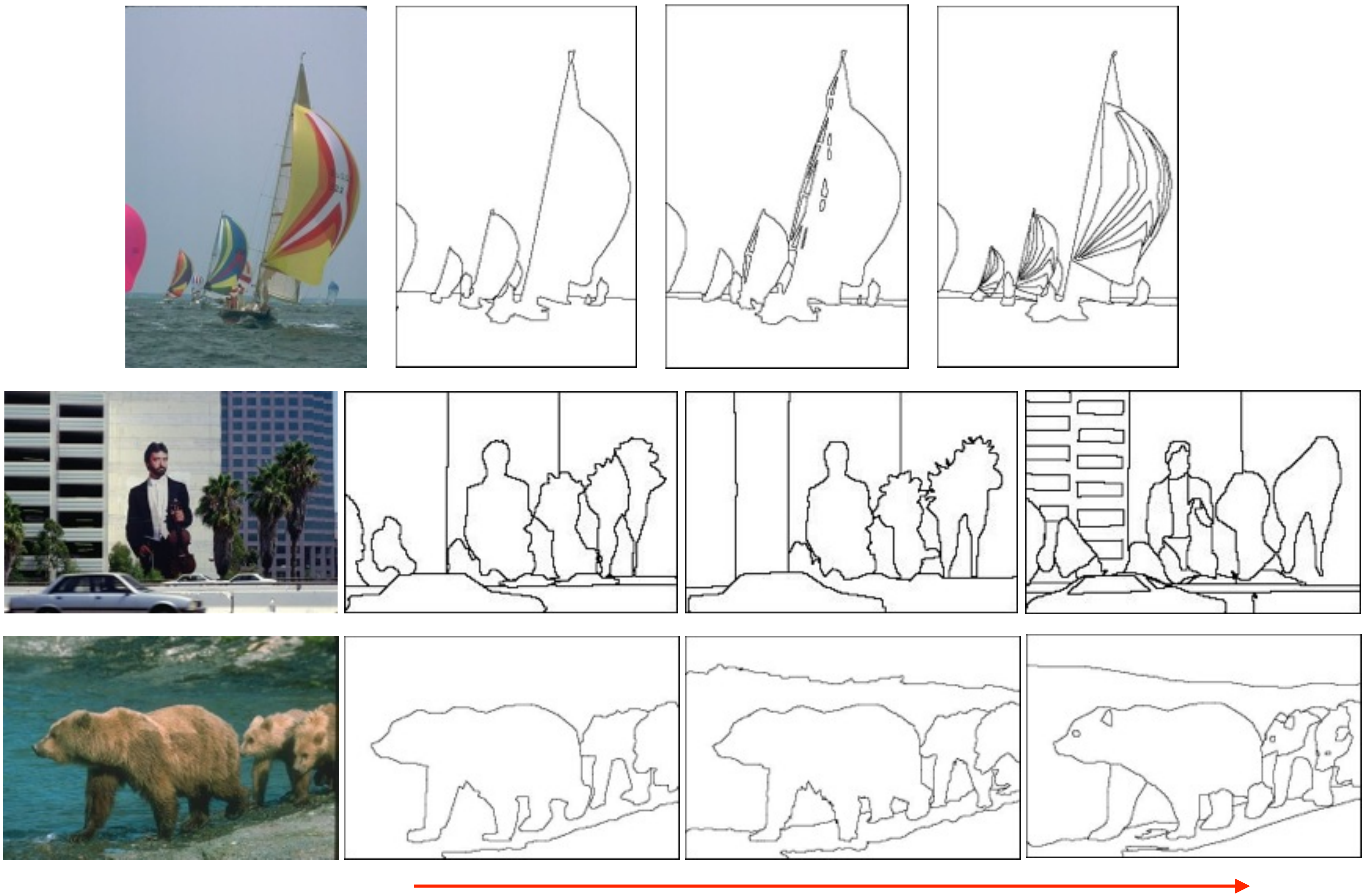

angle (deg)
50
25
0
−25
A
F
D
C
20
40
60
80
100
120
140
16
timesteps







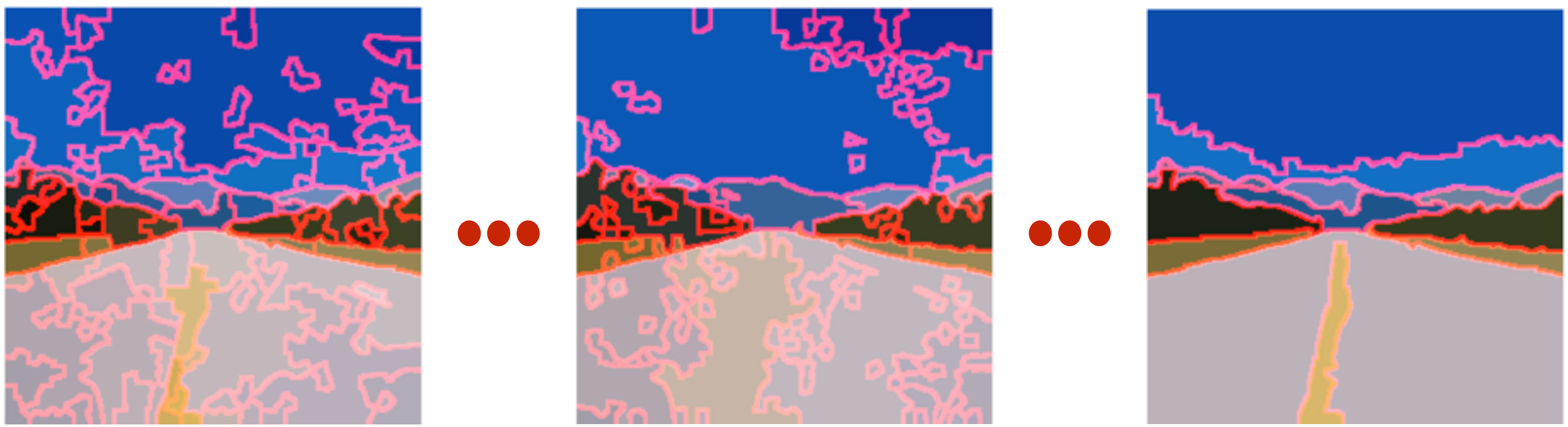


Human Segmentations

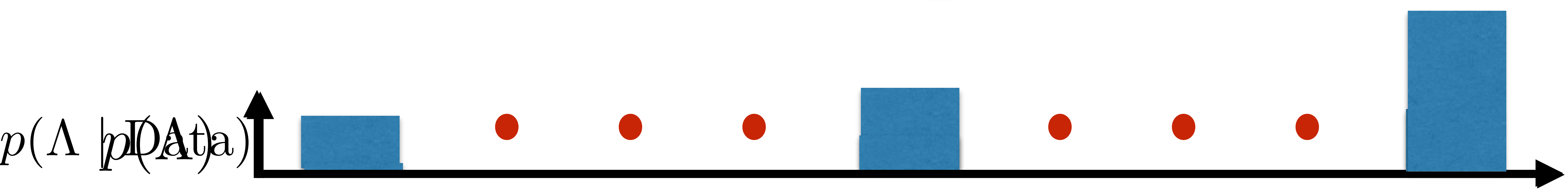
Model Desiderata

- Automatic model selection - adapt to variability in image/video/object complexity
- Manage uncertainty - retain a distribution over possible explanations
- Model spatial and temporal correlations
- Learn from human explanations

Adapting to complexity: Distributions over partitions

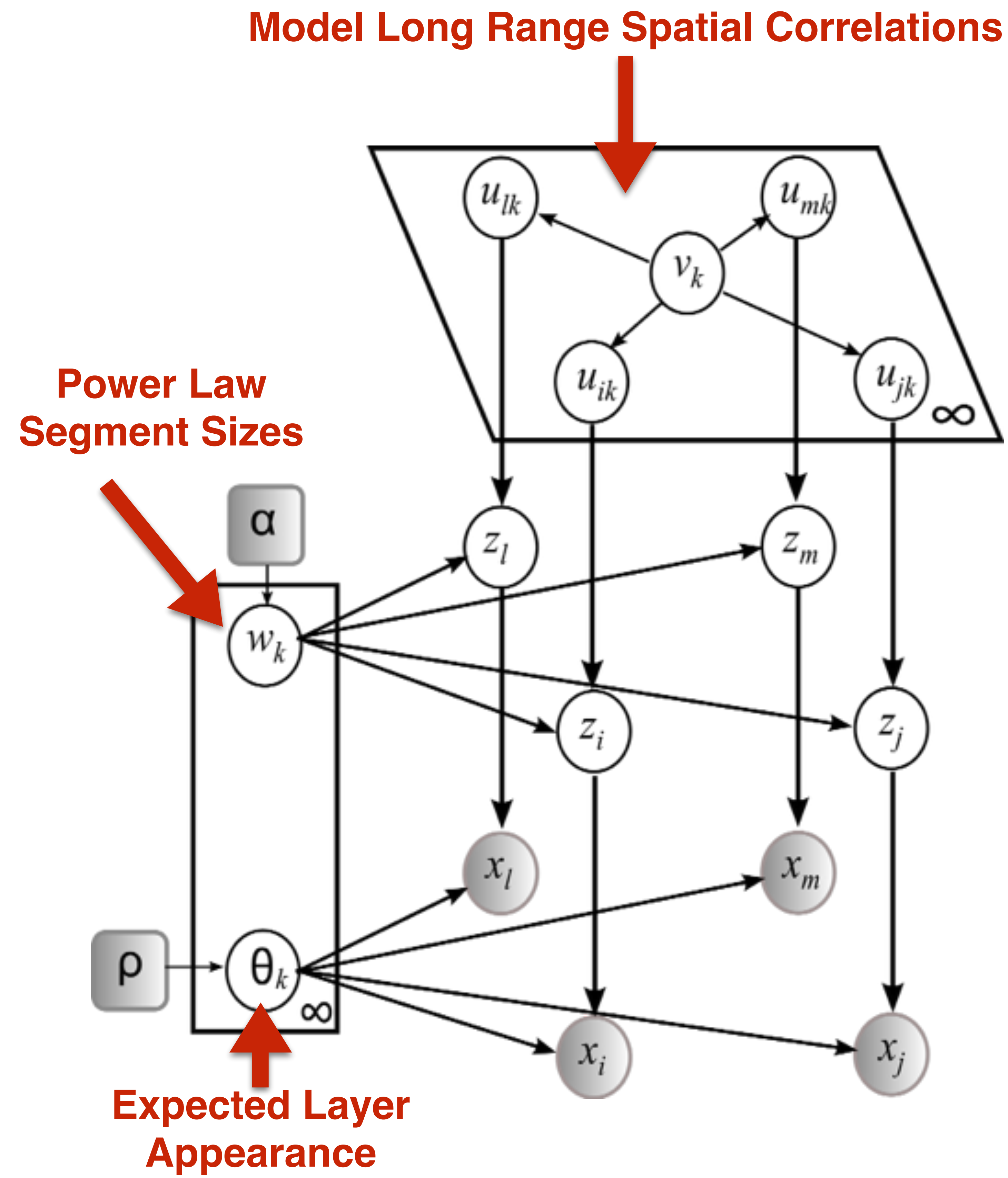


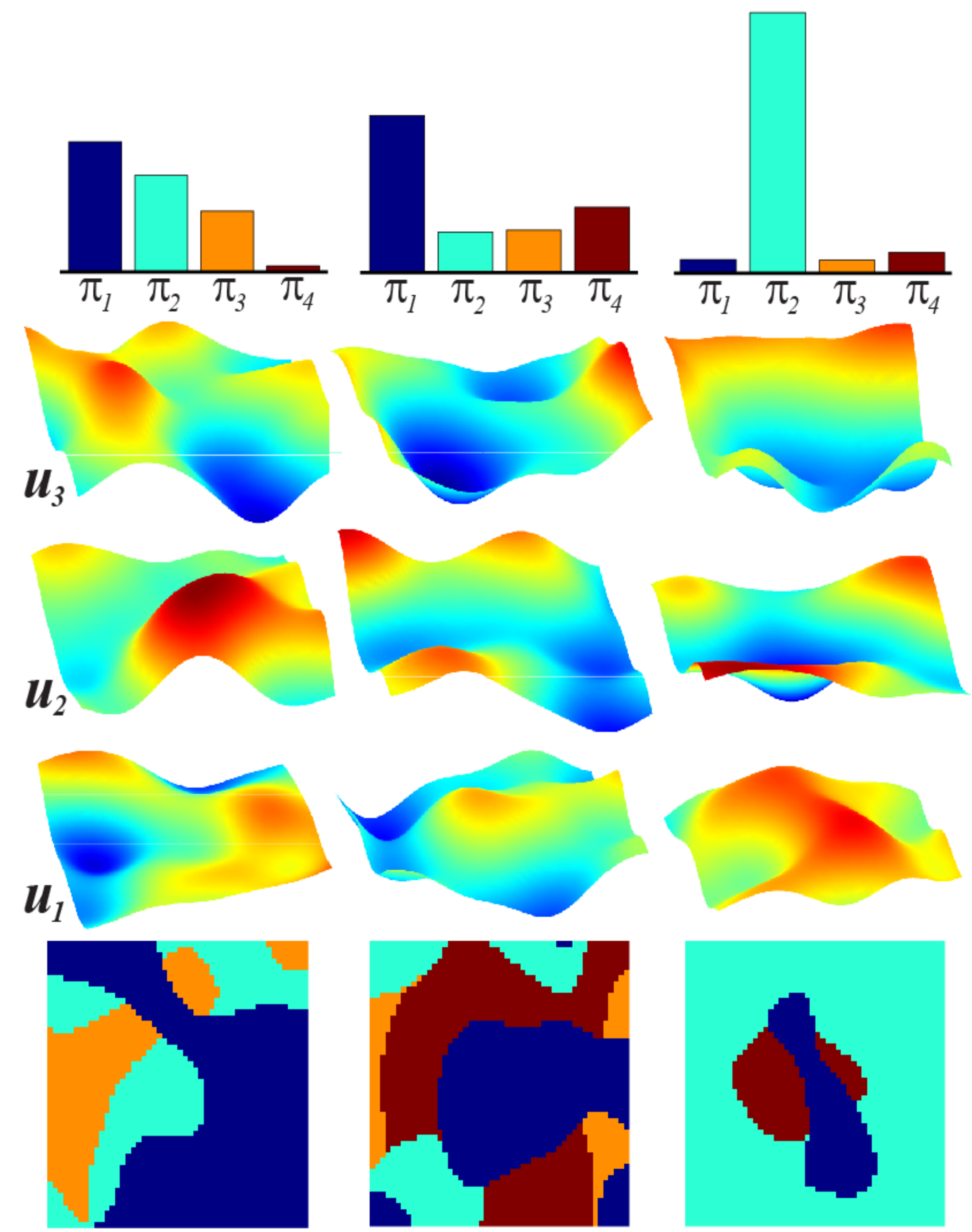
Λ_1 Λ_2 Λ_3



$$\Lambda^* \sim p(\Lambda \mid \text{Data})$$

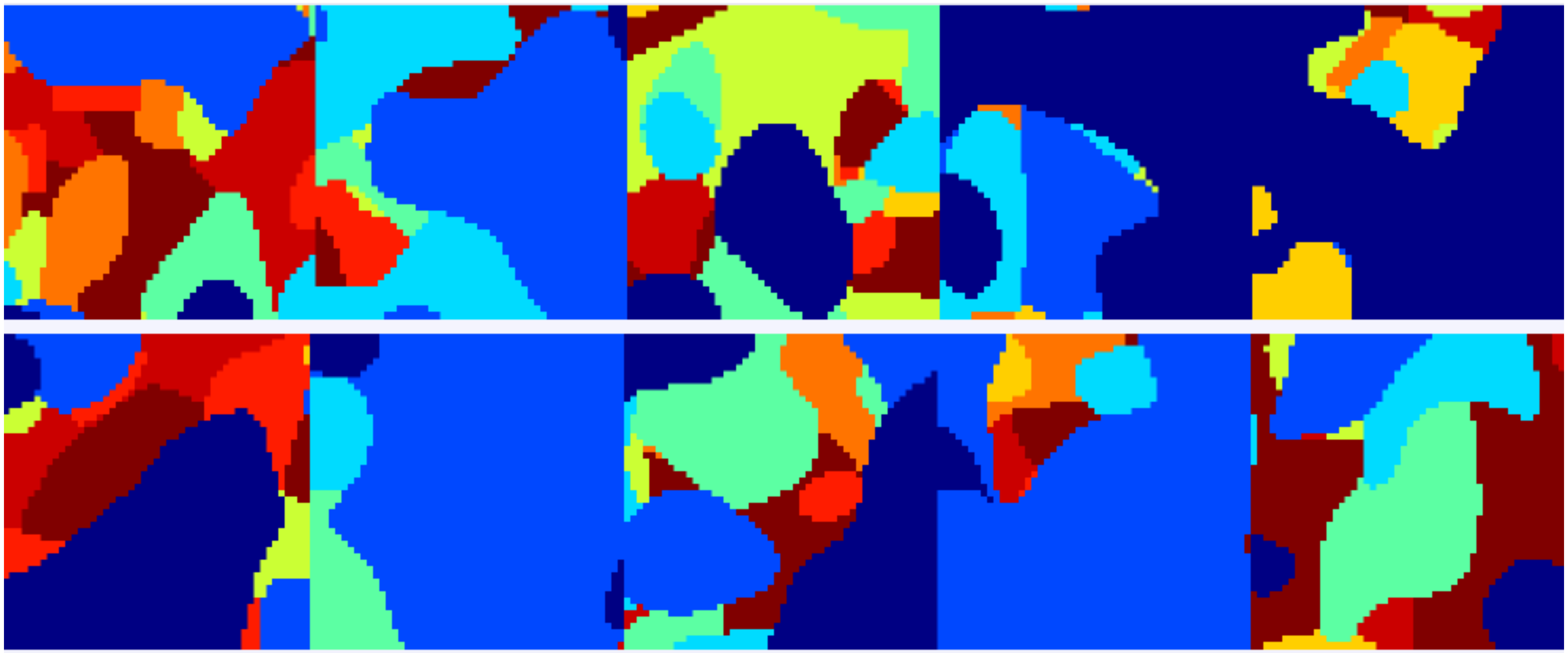
Spatially Coupled PY Processes



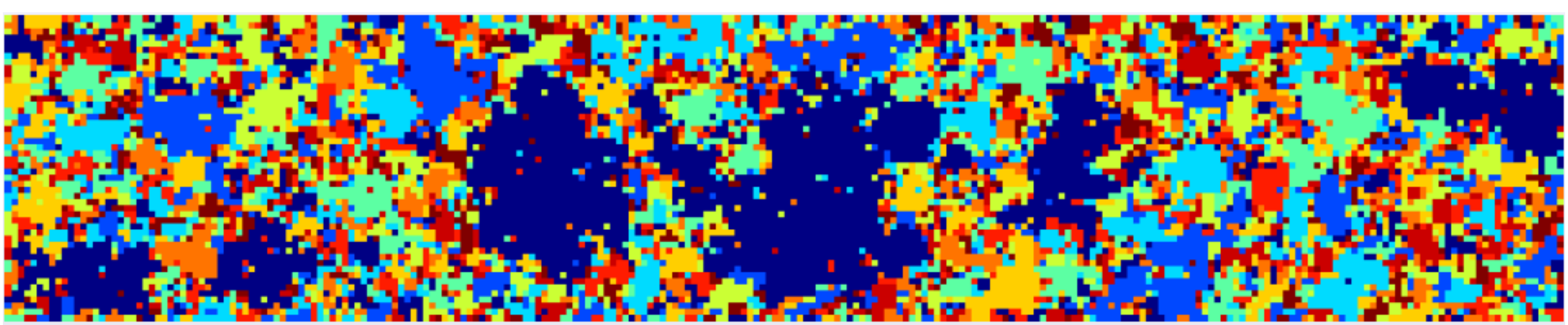


Ghosh & Sudderth, CVPR 2012
Sudderth & Jordan, NIPS 2008

Generative Samples



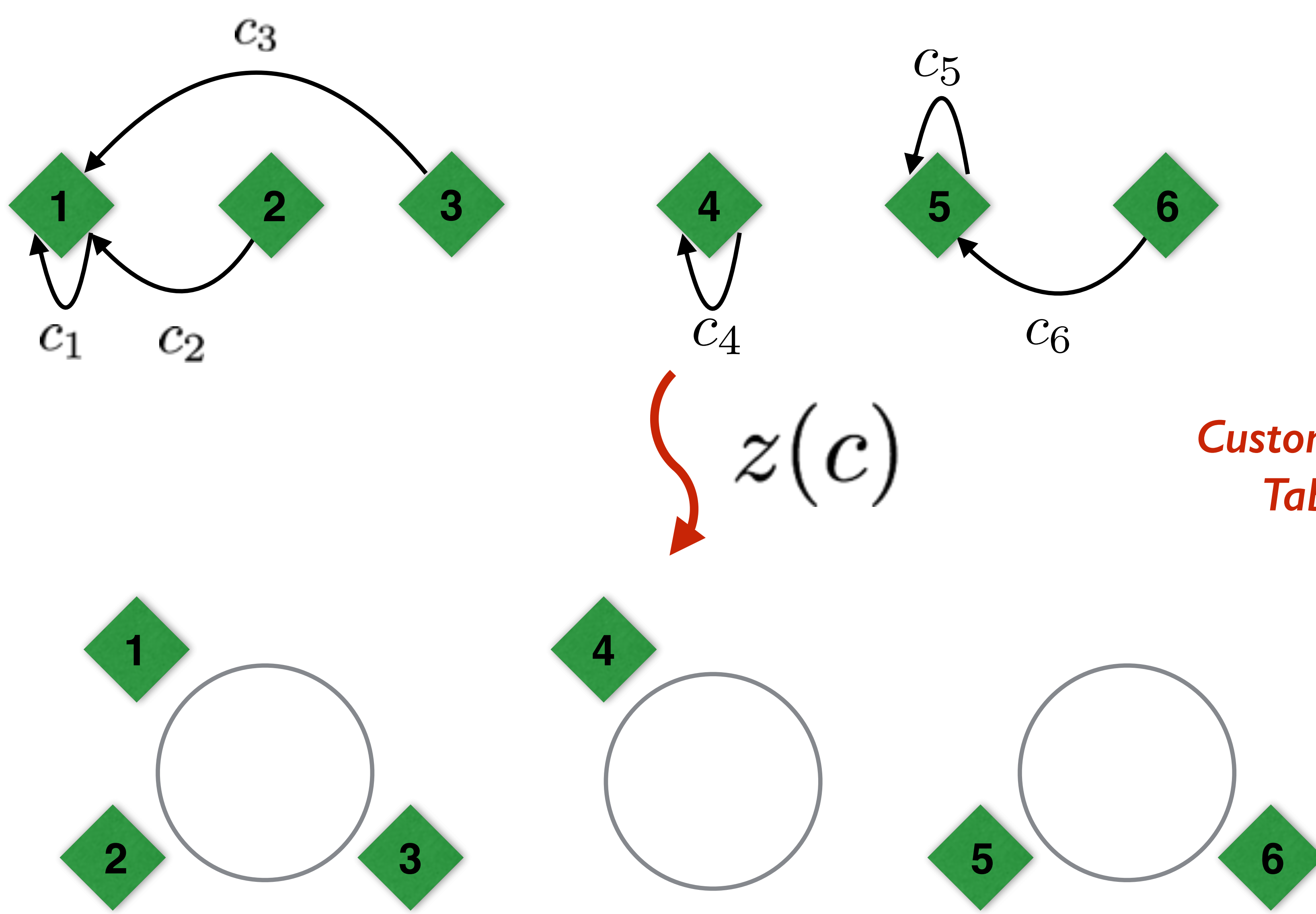
Samples from a Potts Markov Random Field (MRF) model:



Talk Outline

- Distance dependent partitions
  - Parts from articulated 3D objects
- Hierarchical distance dependent partitions
  - Activity discovery from MoCap data
- Learning distance dependent models
  - Image and video segmentation

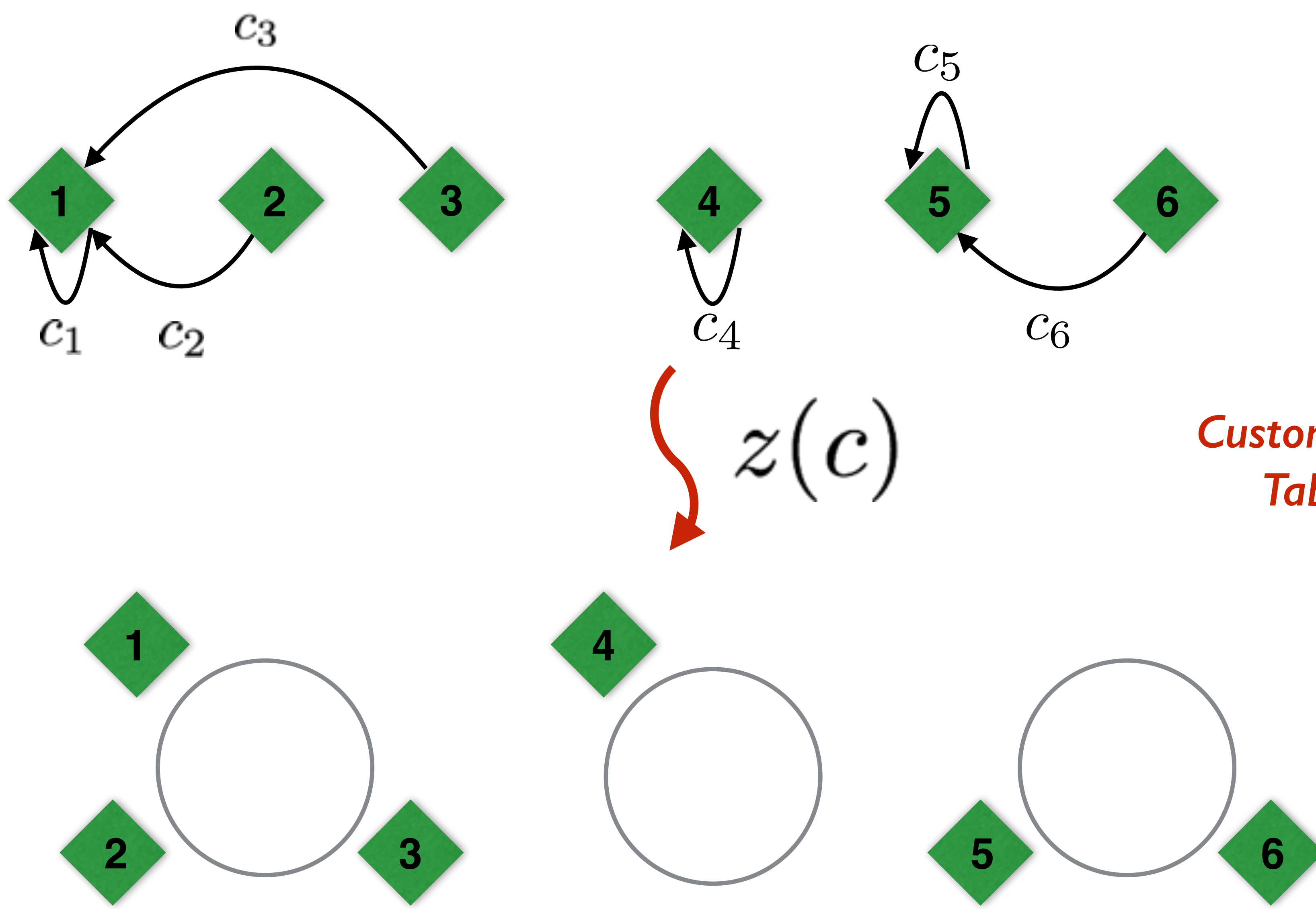
A distribution over partitions: Chinese Restaurant Process



Customers = Data Instances
Tables = Components

Probability of a customer joining a table $\propto \begin{cases} n_k & \text{if } k \text{ is an existing table} \\ \alpha & k \text{ is a new table} \end{cases}$

Distance dependent Chinese Restaurant Process (ddCRP)



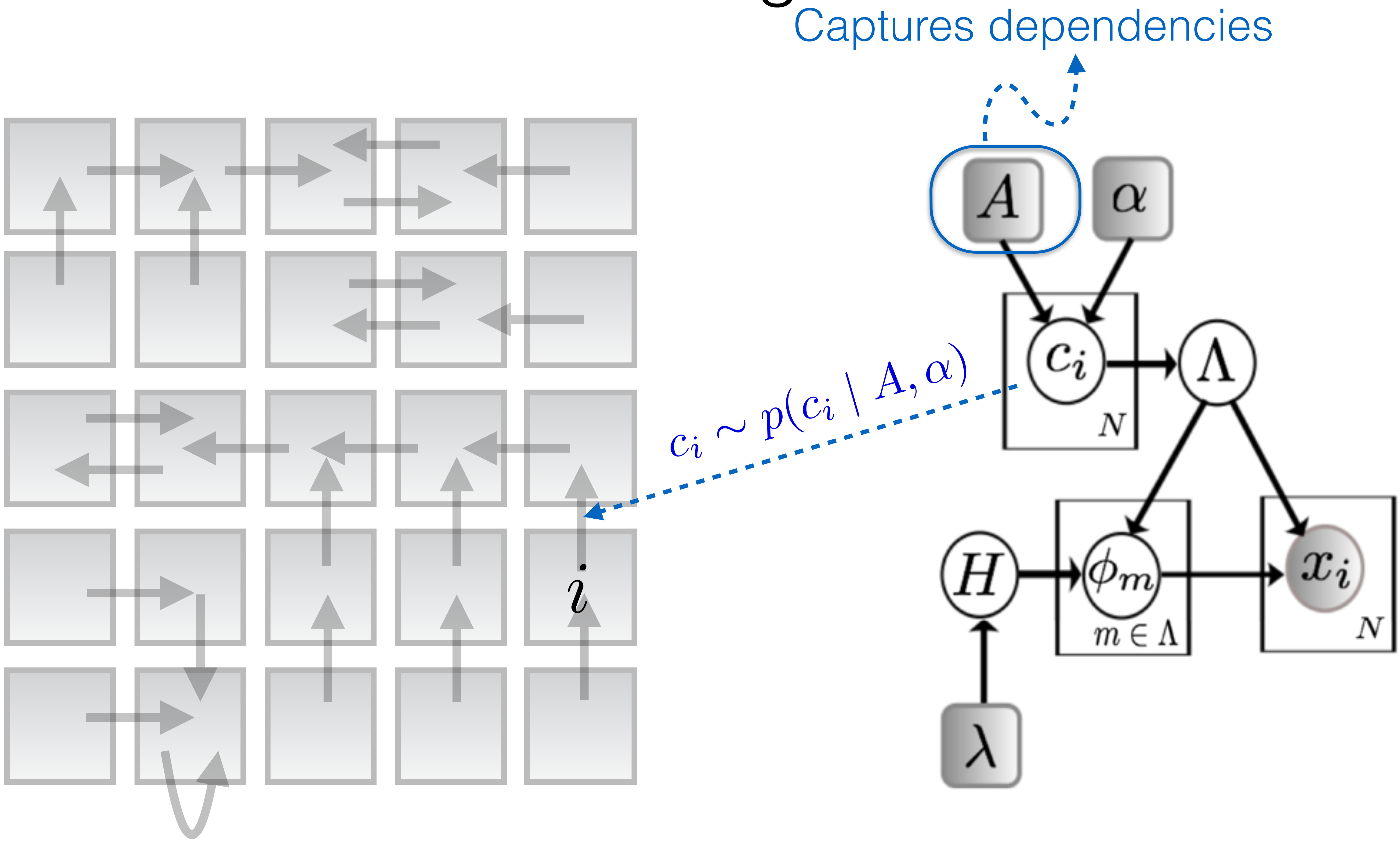
Customers = Data Instances
Tables = Components

$$p(c_n = m \mid A, \alpha) \propto \begin{cases} A_{mn} & \text{if m} \neq \text{n}, \\ \alpha & \text{if m} = \text{n}. \end{cases}$$

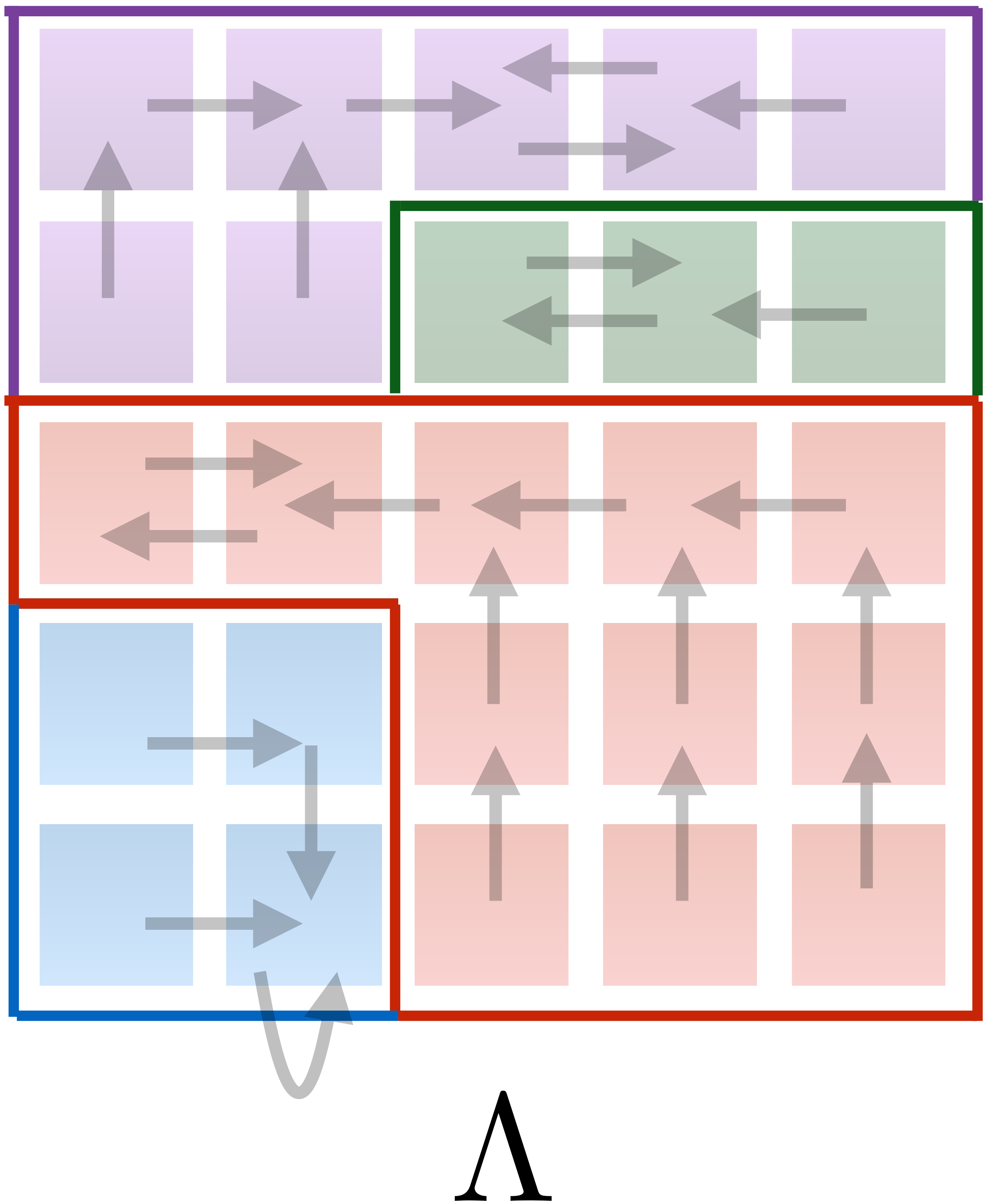
$$p(z(c) \mid A, \alpha) = \prod_{n=1}^{N} p(c_n \mid A, \alpha)$$

Blei & Frazier, JMLR 2011

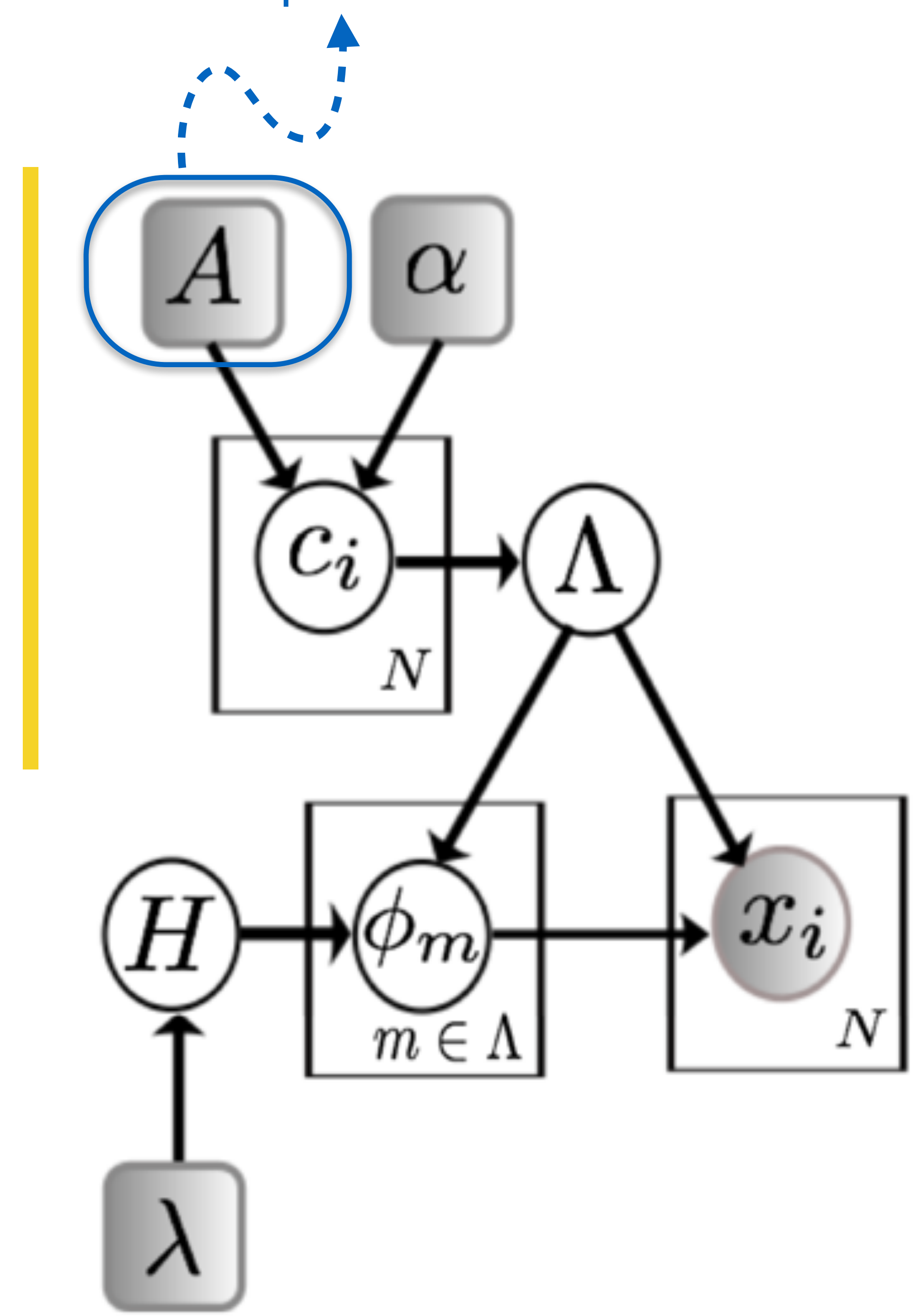
Models for heterogeneous data



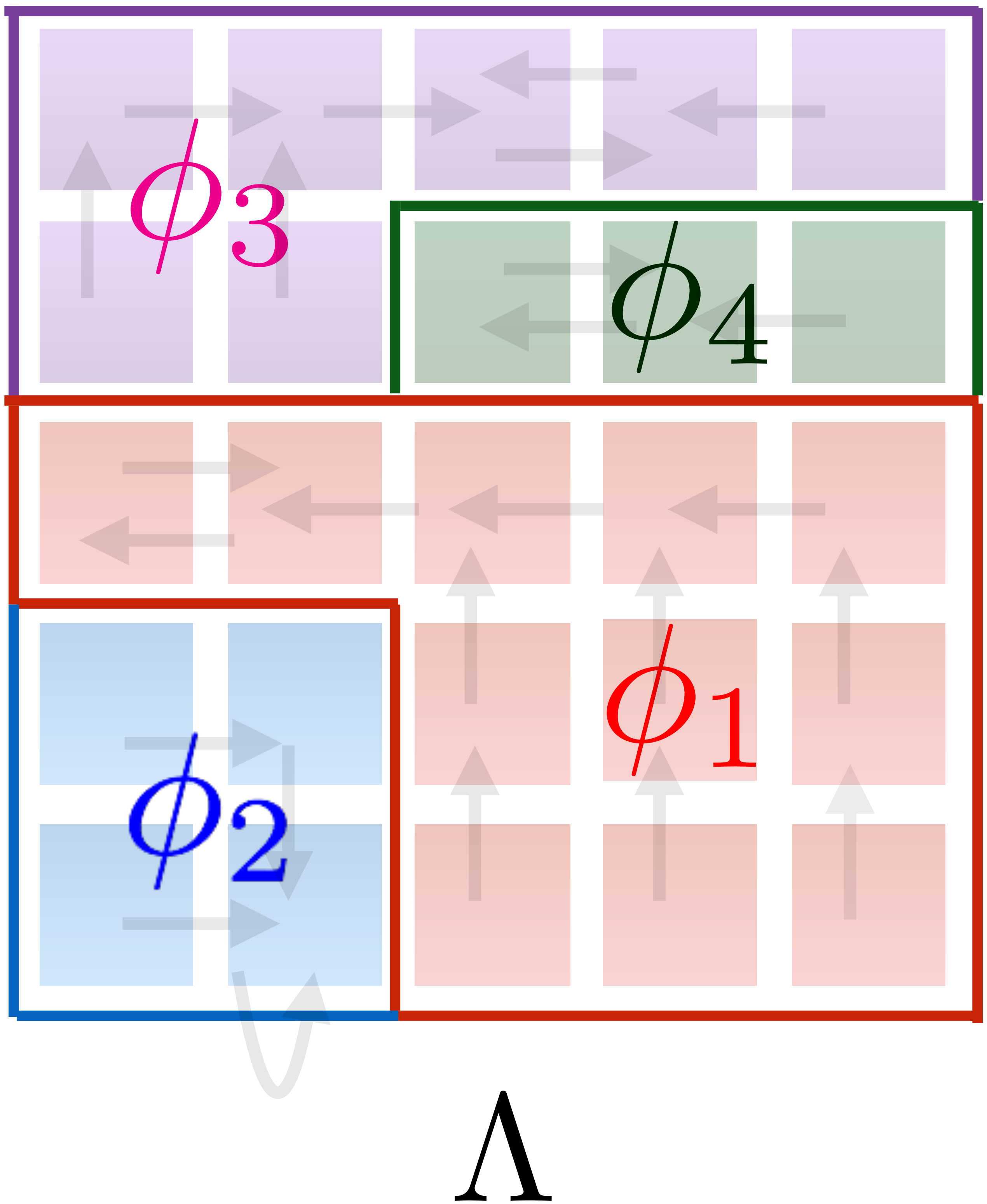
Models for heterogeneous data

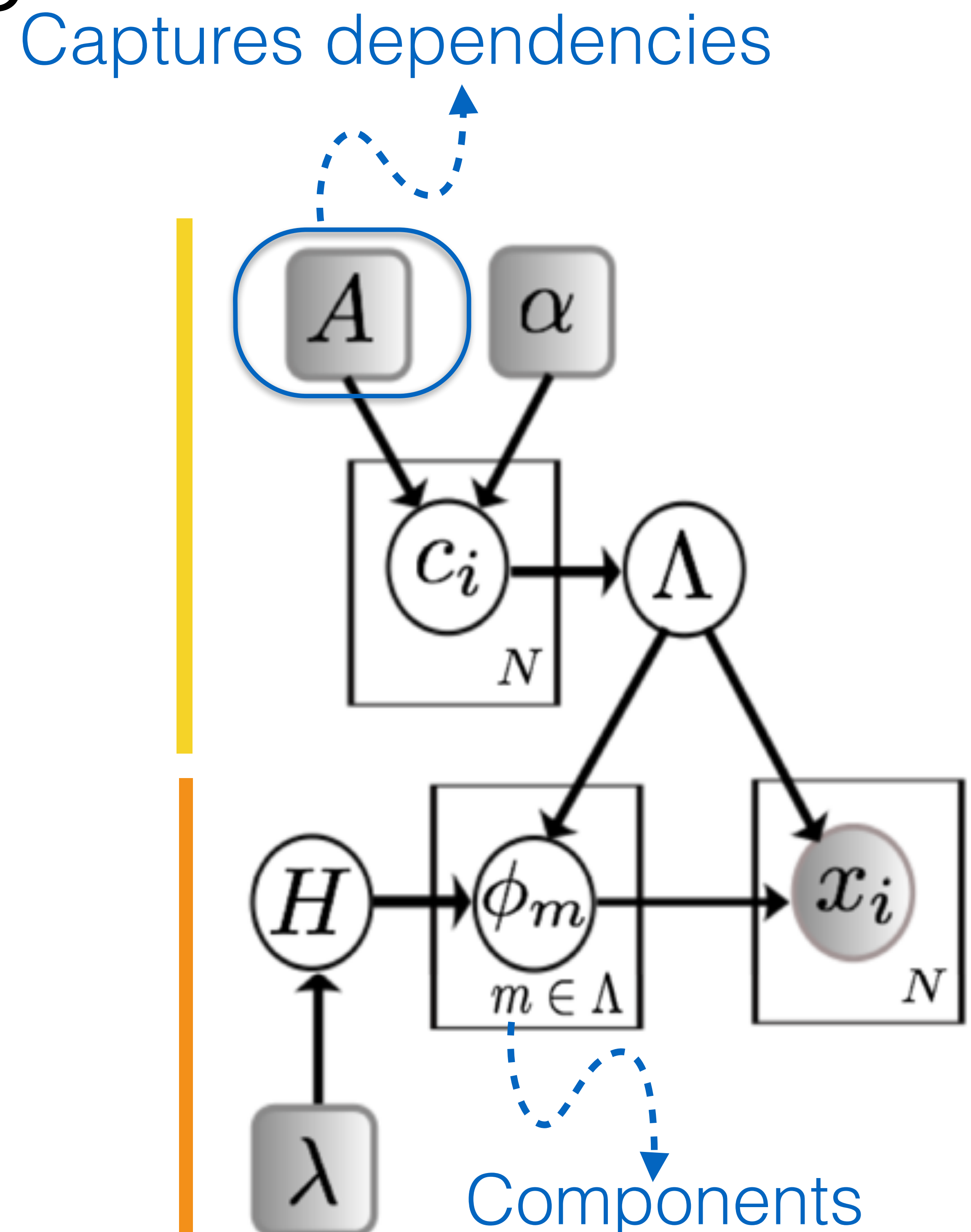


Captures dependencies



Models for heterogeneous data



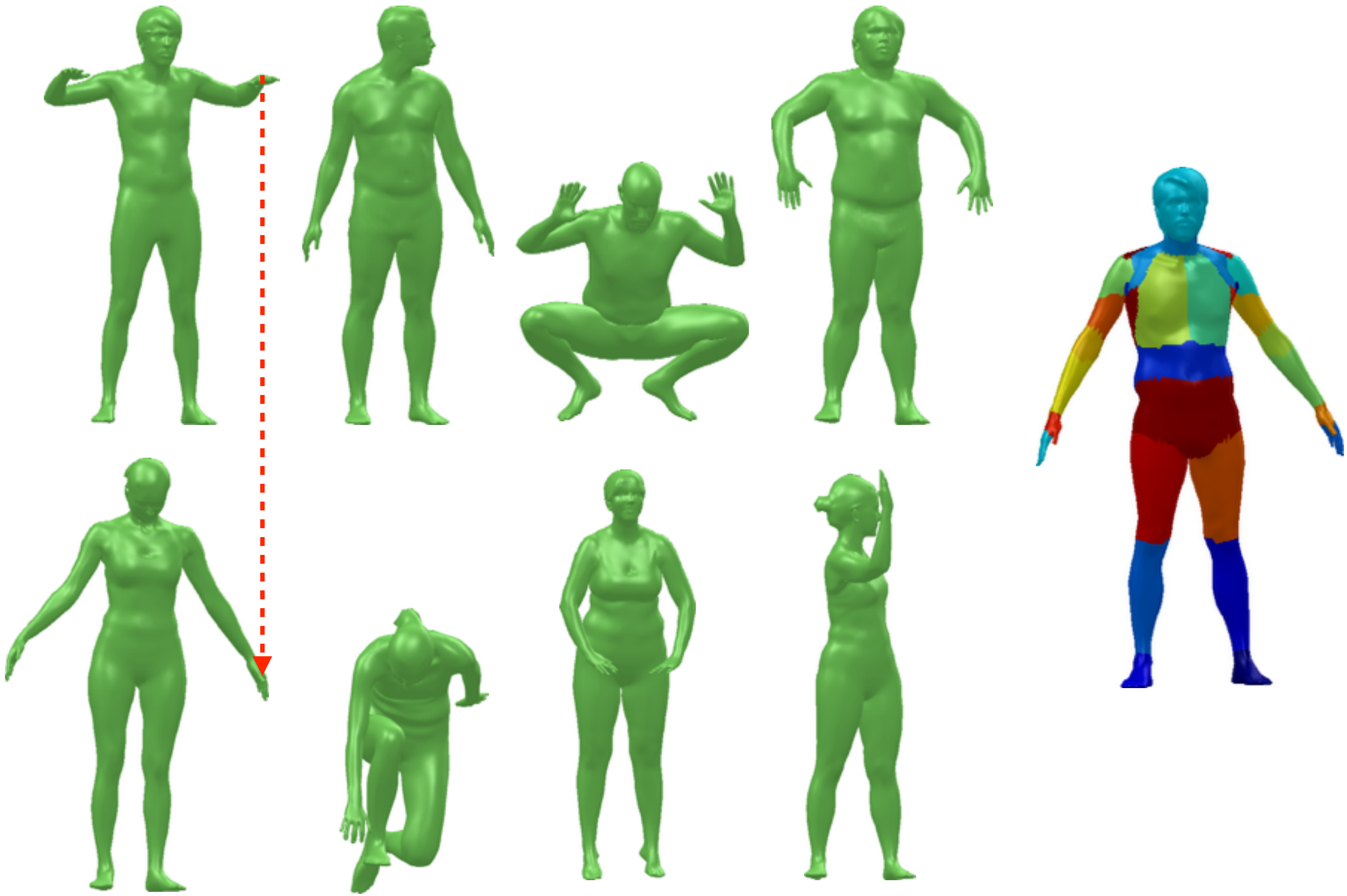


$$\phi_m \sim H(\lambda), \forall m \in \Lambda$$
$$x_i \sim \phi_m, \forall i \in m$$

Talk Outline

- Distance dependent partitions
  - Parts from articulated 3D objects
- Hierarchical distance dependent partitions
  - Activity discovery from MoCap data
- Learning distance dependent models
  - Image and video segmentation

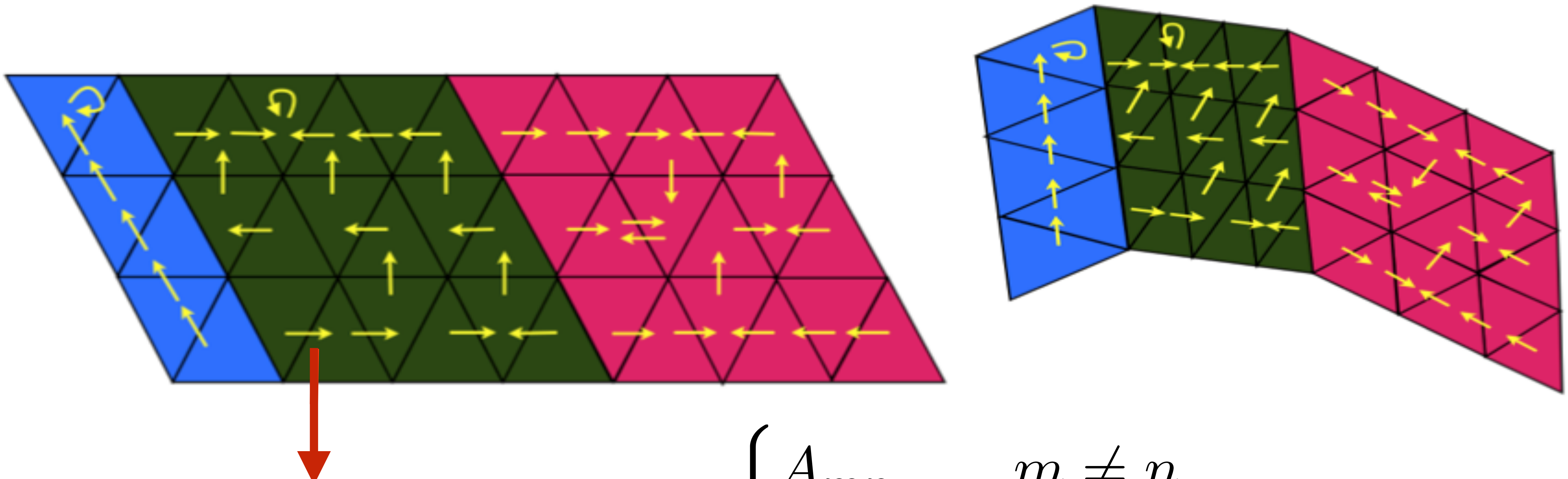
Parts from Deformations



Discovering Parts from Deformations: Big Picture

- **Cluster:** Mesh faces.
- **Prior:** over the space of plausible mesh partitions.
- **Likelihood:** Given segmentation into parts, model how multiple bodies deform across many poses.
- **Posterior:** Explored through MCMC.

ddCRP Prior over Mesh Partitions

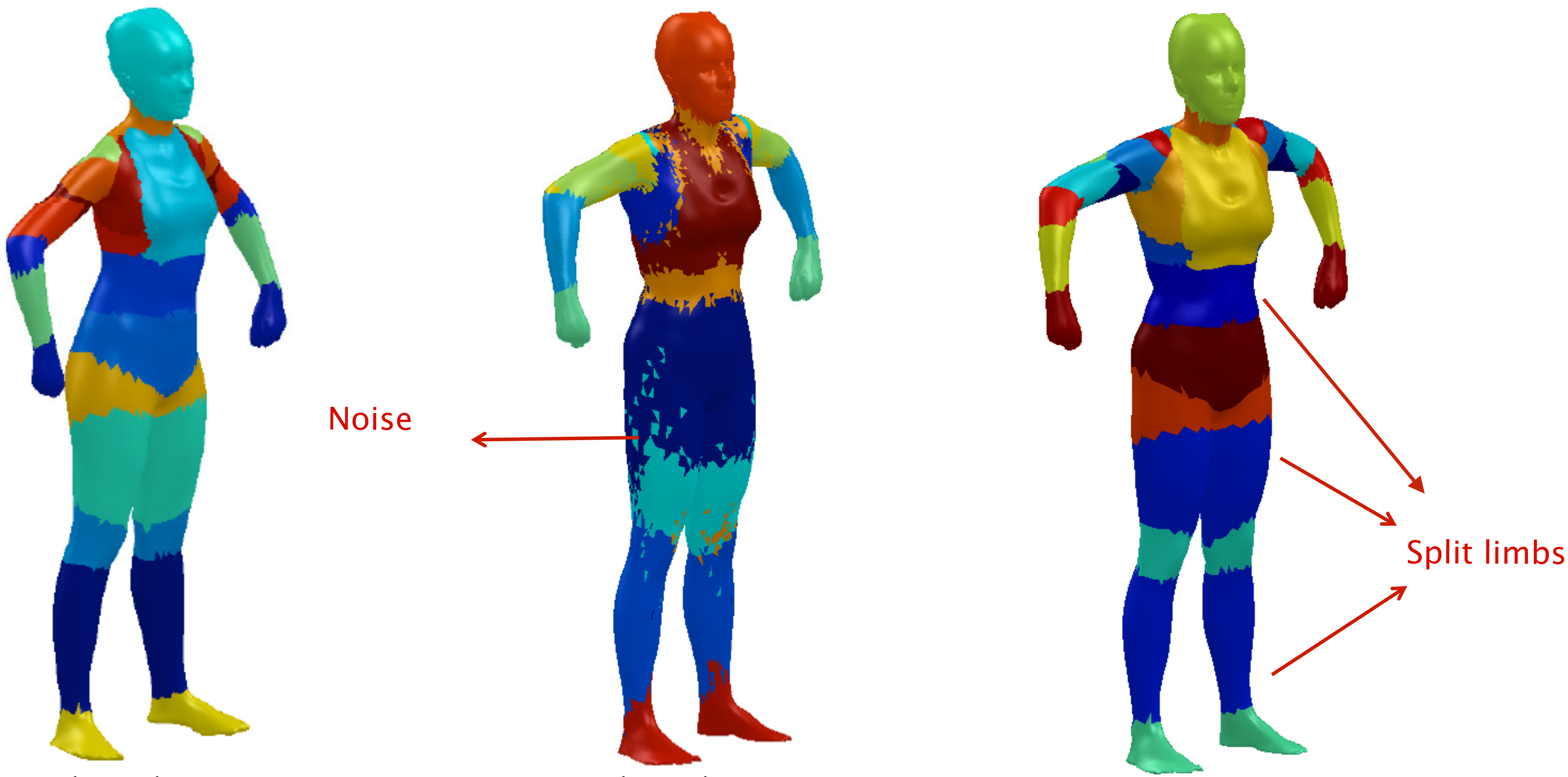


$$p(c_m = n \mid A, \alpha) \propto \begin{cases} A_{mn} & m \neq n, \\ \alpha & m = n. \end{cases}$$

- Mesh faces are only allowed to link to neighboring faces

$$A_{mn} = \mathbf{1}[d_{mn} \leq 1]$$

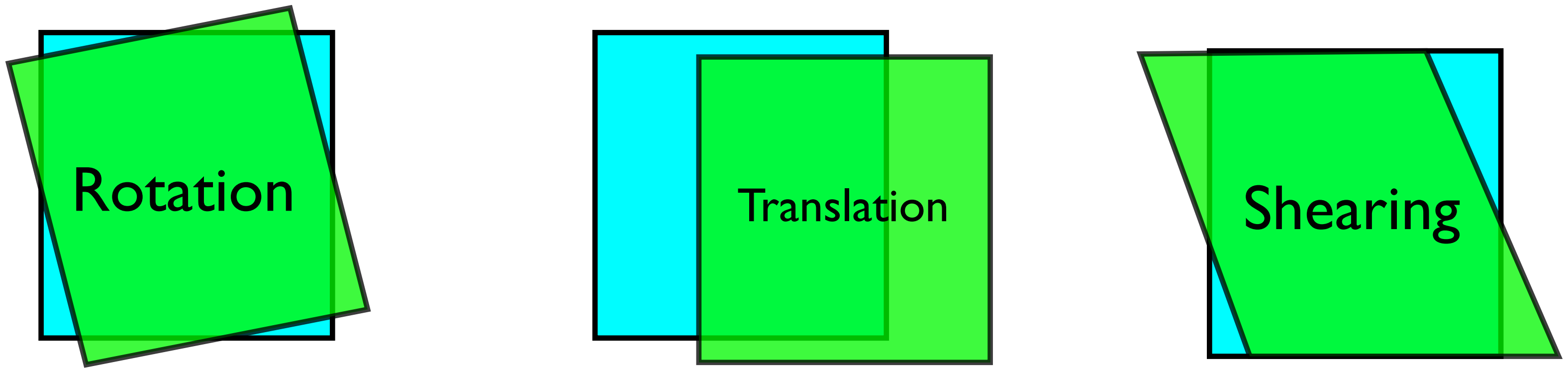
Prior over plausible partitions



$p(Z_1) > 0$ | $p(Z_2) = 0$ | $p(Z_3) = 0$

Desirable | ***Avoid: Noisy Parts*** | ***Avoid: Disconnected Parts***

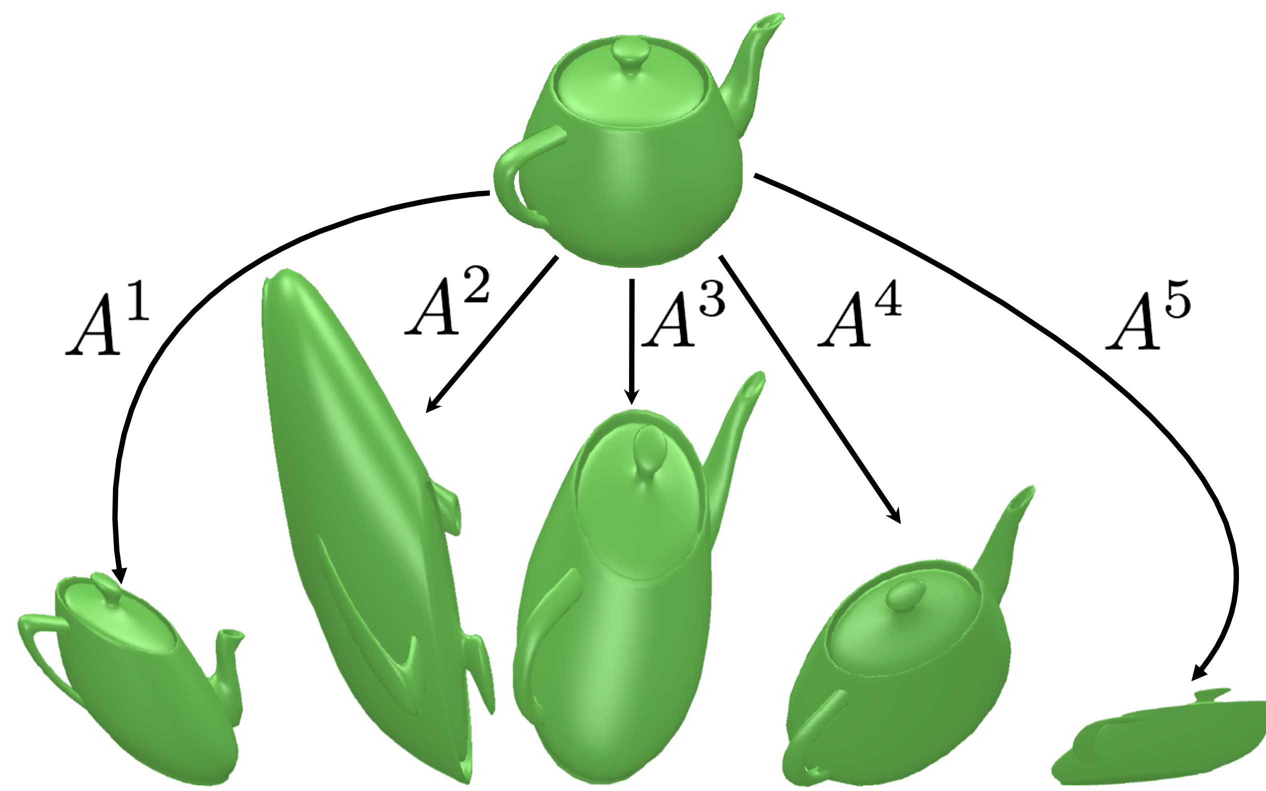
Modeling Part Deformations



Matrix Normal Inverse Wishart:

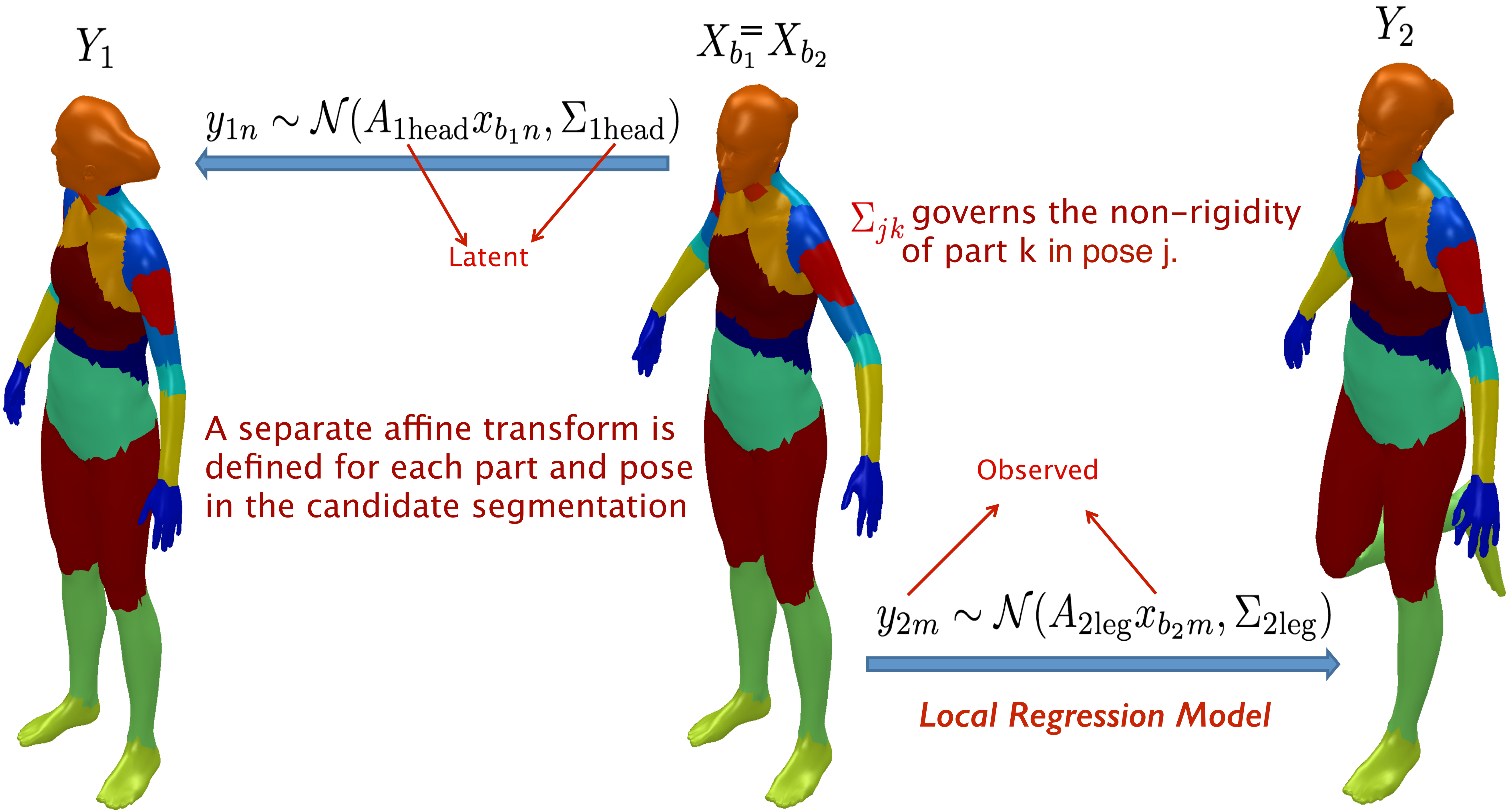
$$\Sigma \sim \mathcal{IW}(n_0, S_0)$$
$$A \mid \Sigma \sim \mathcal{MN}(M, \Sigma, K)$$

where $A \in \mathbb{R}^{3\times 4}$ is an affine transformation.



$$A^1 \ldots A^5 \sim \mathcal{MNIW}(.)$$

Generative Affine Likelihoods



Marginal Affine Likelihoods

For each part and pose combination analytically marginalize over *all possible* affine transformations

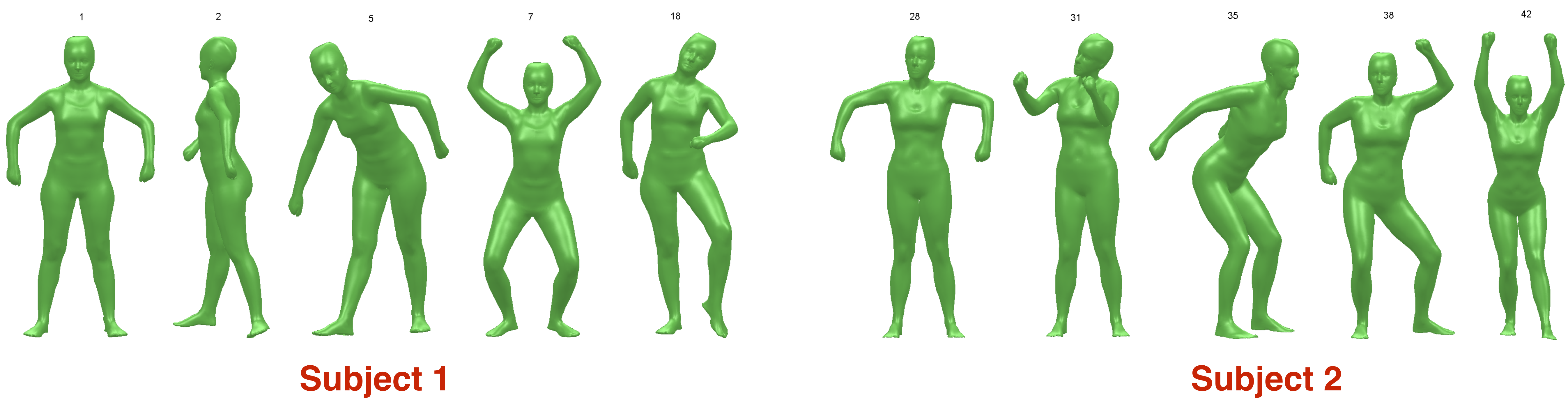
$$\boxed{p(Y_{jk} \mid X_{jk}) = \int \underbrace{p(Y_{jk}, A_{jk}, \Sigma_{jk} \mid X_{jk}) dA_{jk}, d\Sigma_{jk}}_{\textbf{Marginal Likelihood}}}$$

Bayesian Model Selection:

- Improper merges have low marginal likelihoods
- Improper splits are "suspicious coincidences" and end up with lower marginal likelihoods

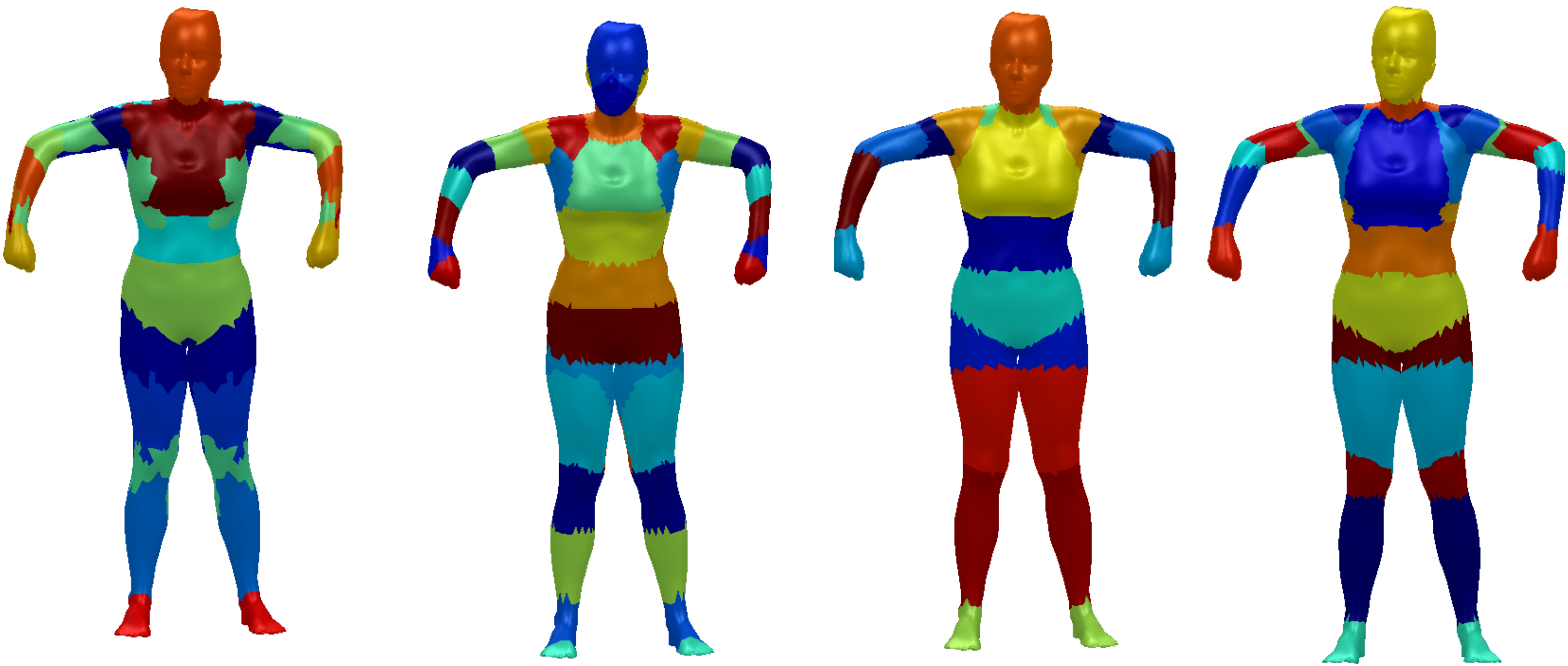
Human Bodies in Motion

- 56 Aligned scans from two human subjects
- Wide variability in poses, limited variability in body shapes



Ghosh et al., NIPS 2012

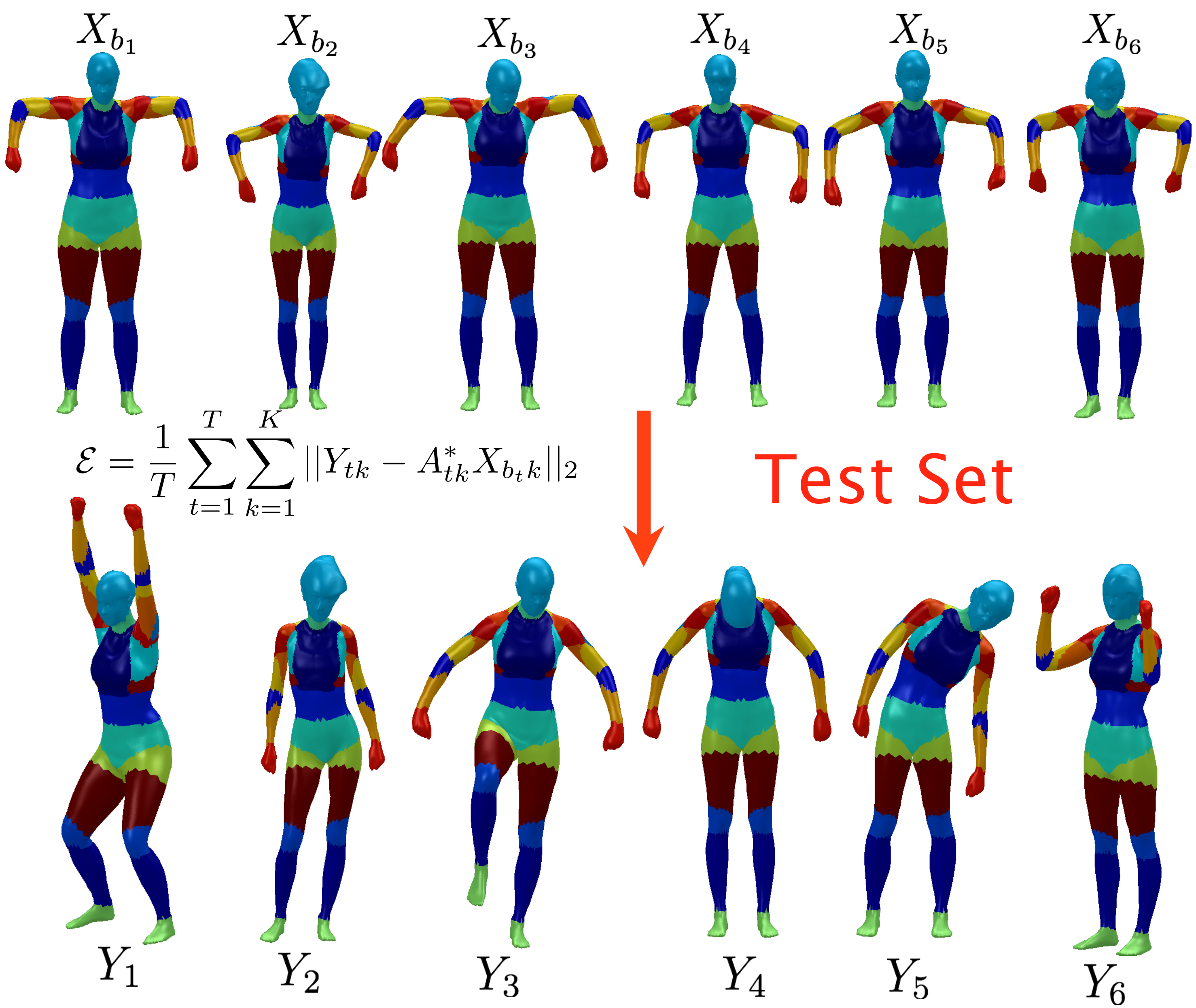
Visual Comparisons



Agglomerative | **Spectral Clustering** *Liu & Zhang, 2004* | **CRP** | **ddCRP**

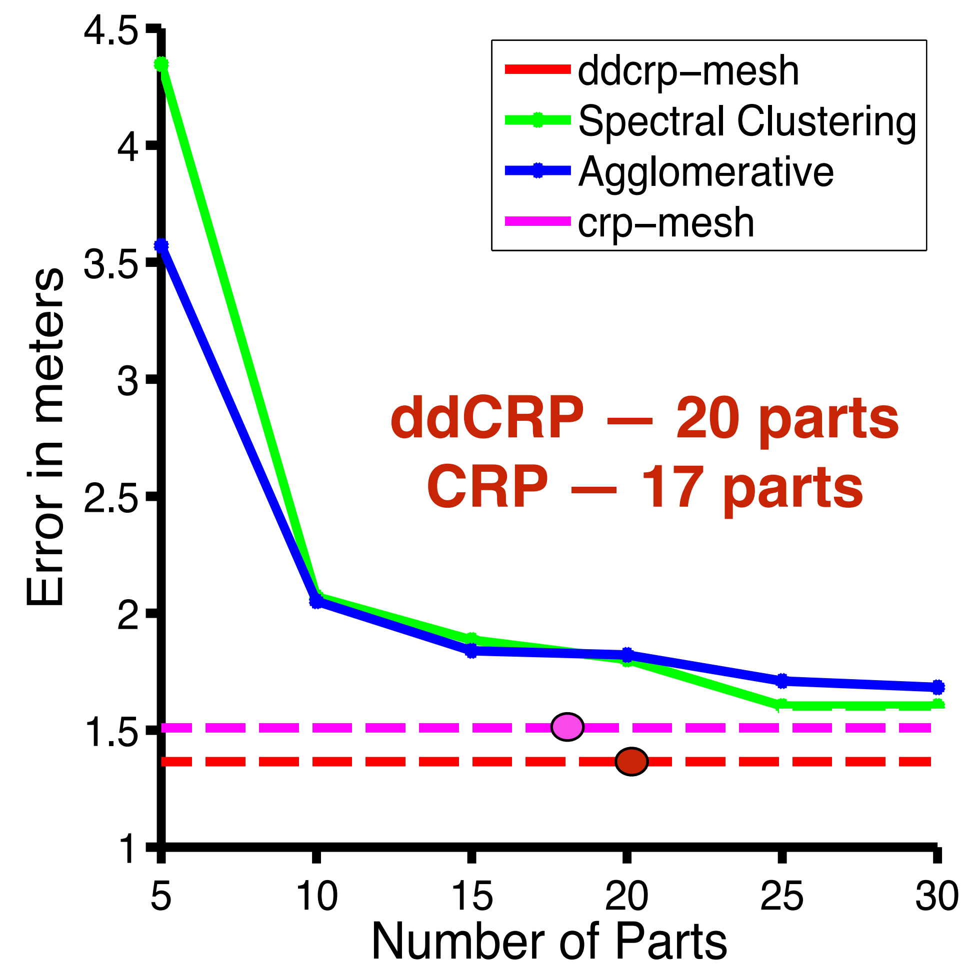
Ghosh et al., NIPS 2012

Quantitative Evaluation



$$\mathcal{E} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}||Y_{tk} - A^*_{tk}X_{b_tk}||_2$$

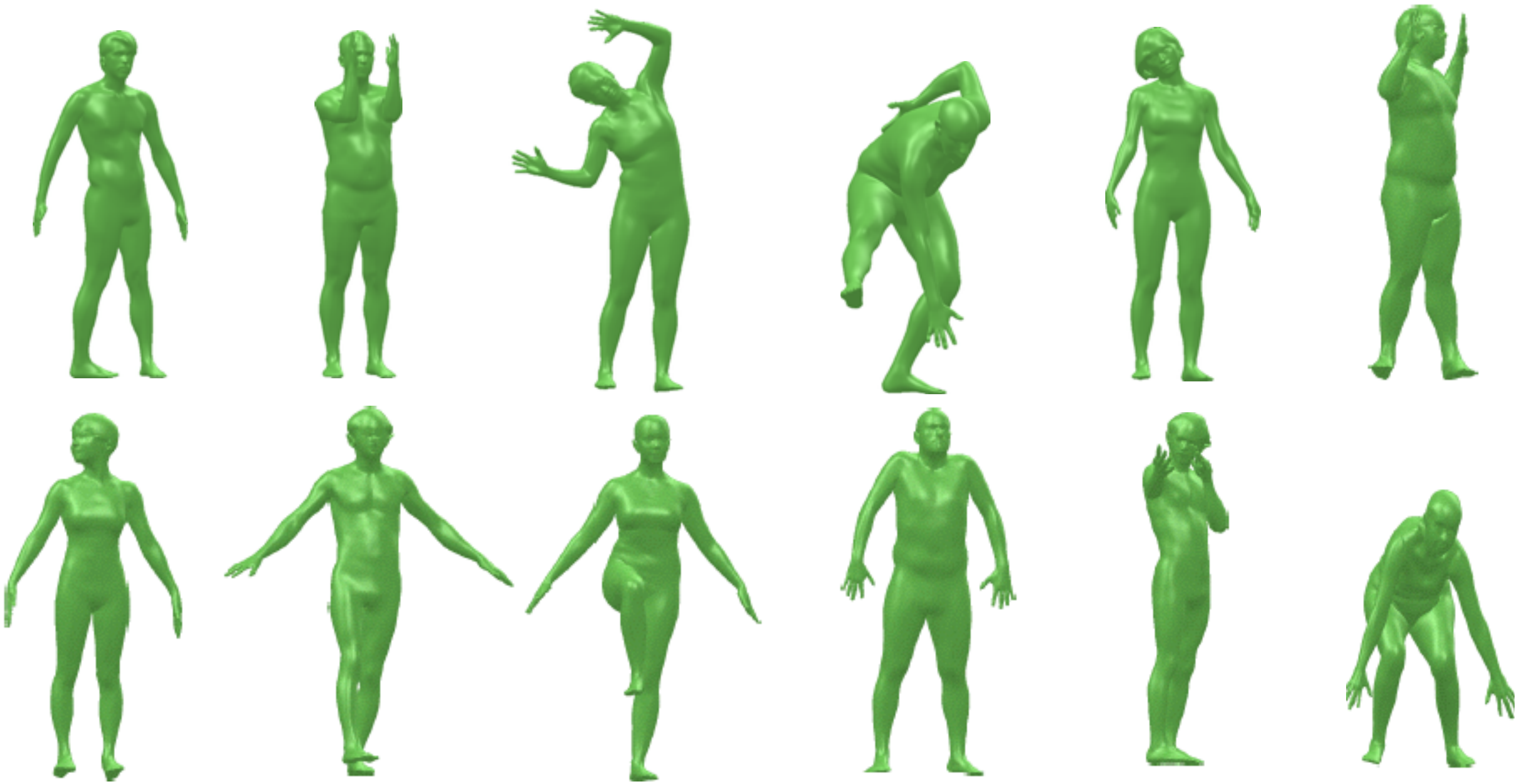
12 Test meshes from six subjects.



Measure error in predicted motion for the candidate segmentations

Ghosh et al., NIPS 2012

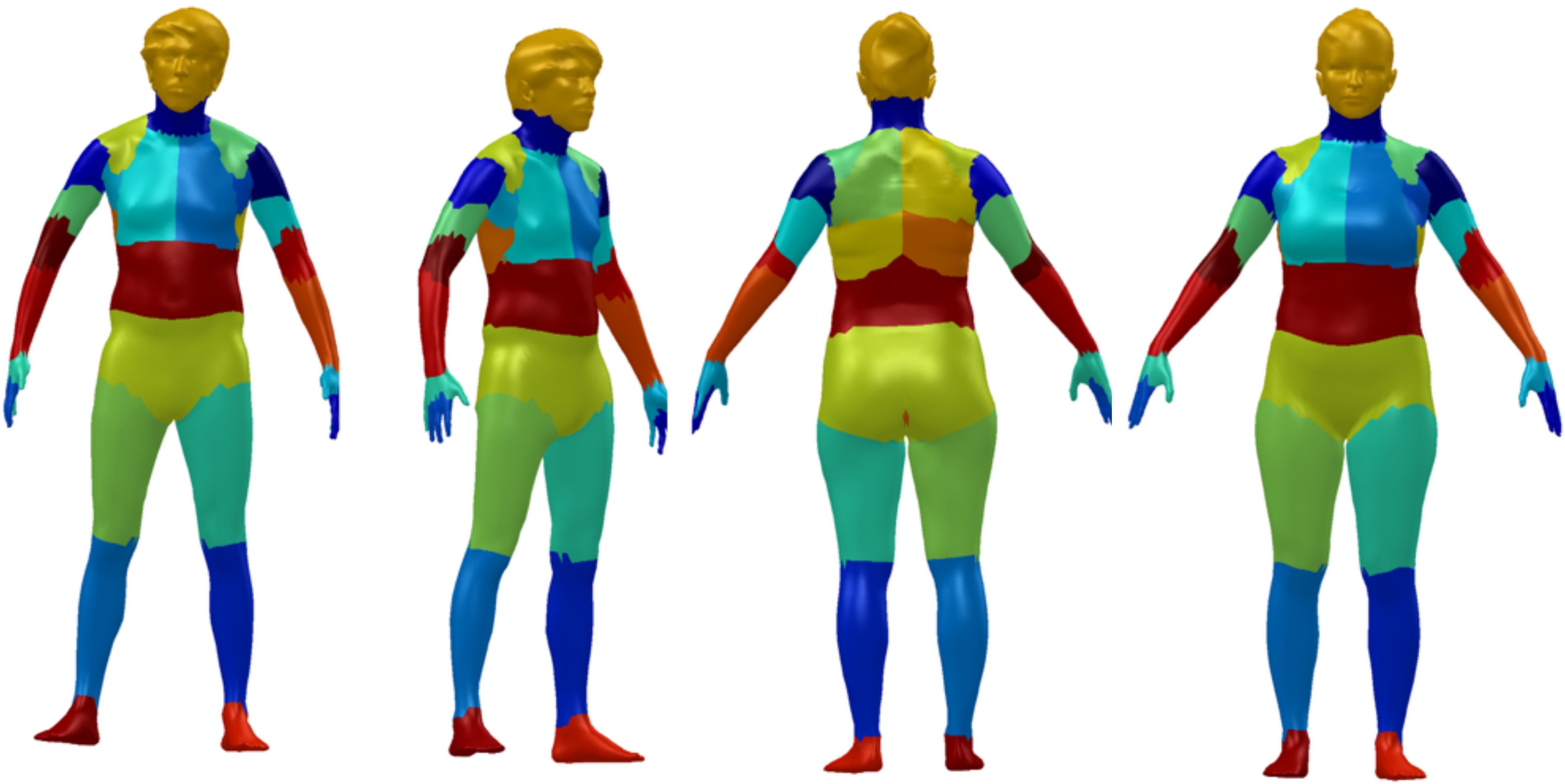
Large Scale Studies



1732 meshes, 78 subjects, ~22,000 mesh faces

Wide variability in both body shapes and poses.

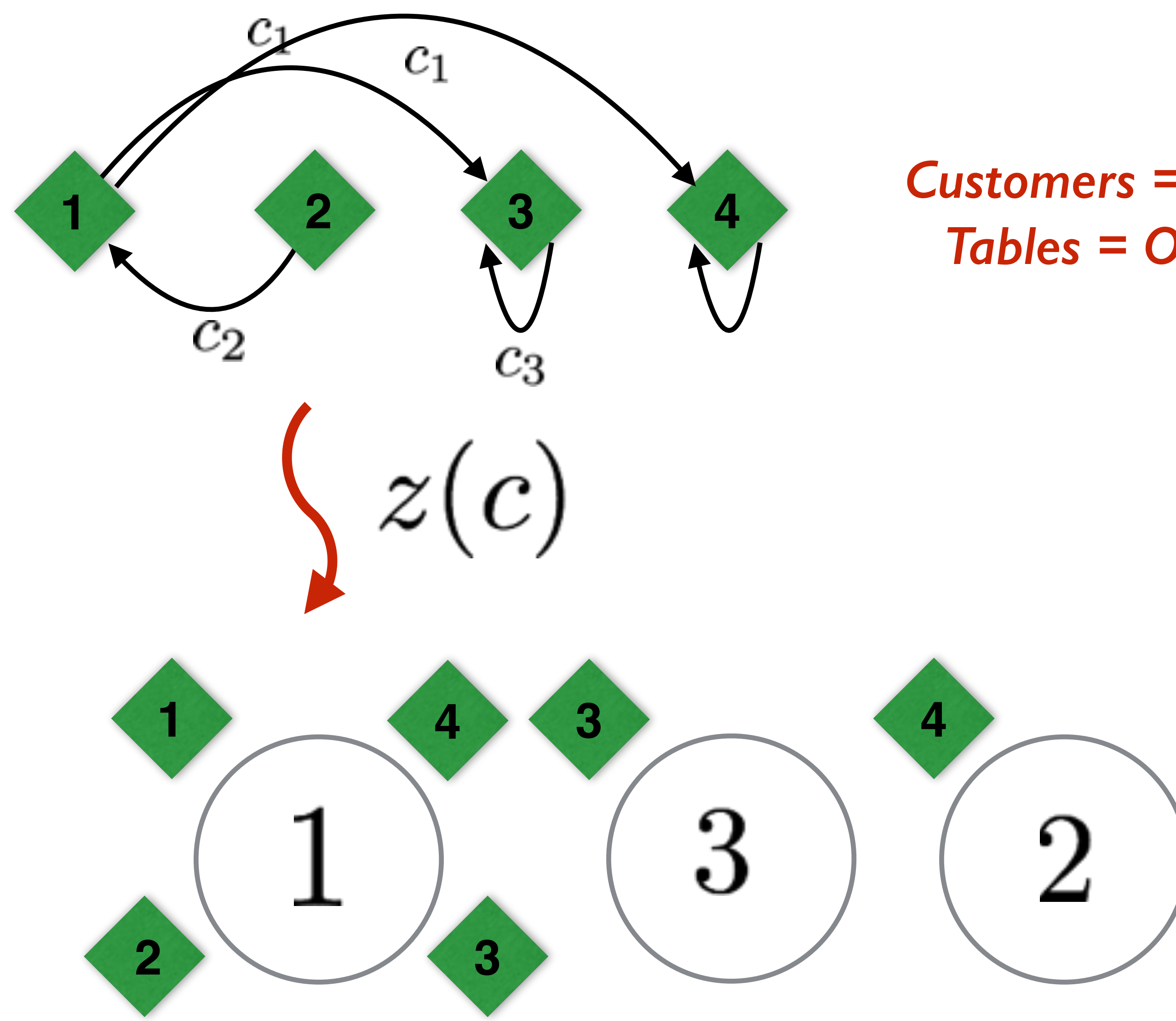
Segmented Bodies



Computer generated meshes

Inference through Gibbs Sampling

Collapsed Sampler:
Only need to sample links,
other random variables are
analytically marginalized out.



Customers = Mesh Faces
Tables = Object Parts

Table structure

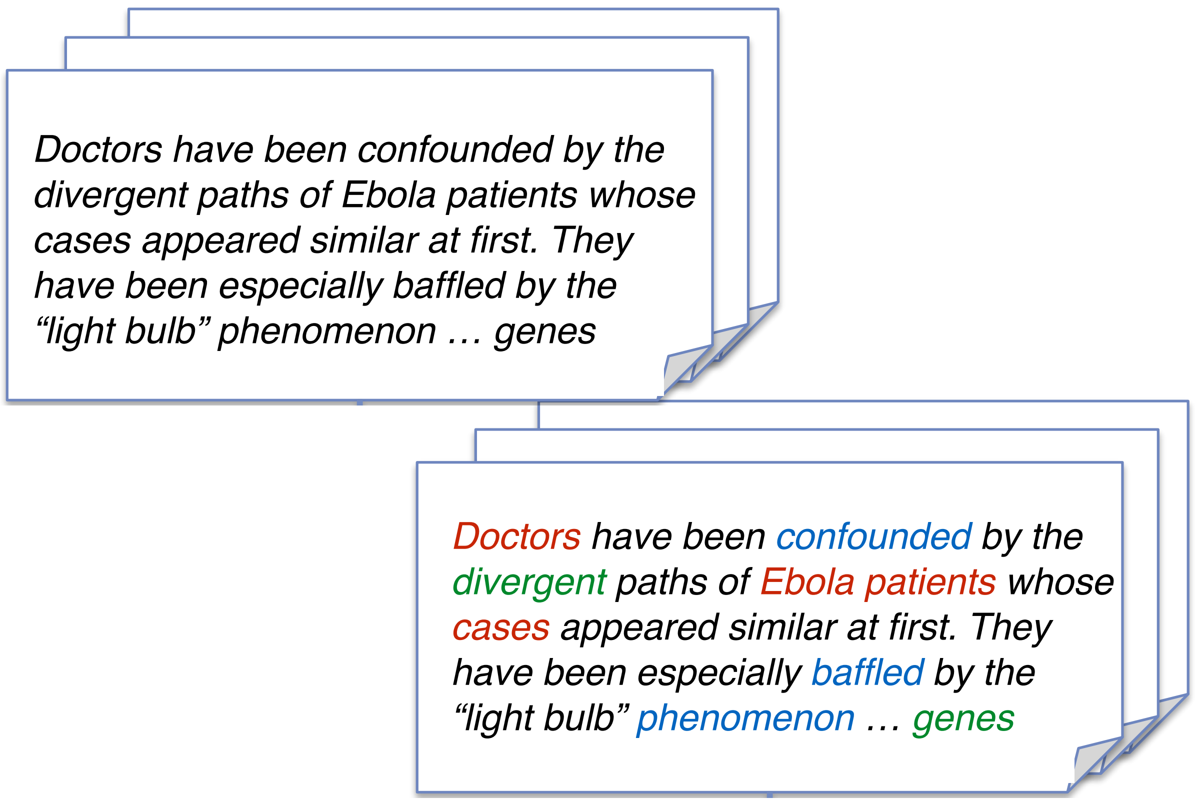
Segmentation

Local changes in the link structure lead to large changes in the partition structure

Talk Outline

- Distance dependent partitions
    - Parts from articulated 3D objects
- Hierarchical distance dependent partitions
    - Activity discovery from MoCap data
- Learning distance dependent models
    - Image and video segmentation

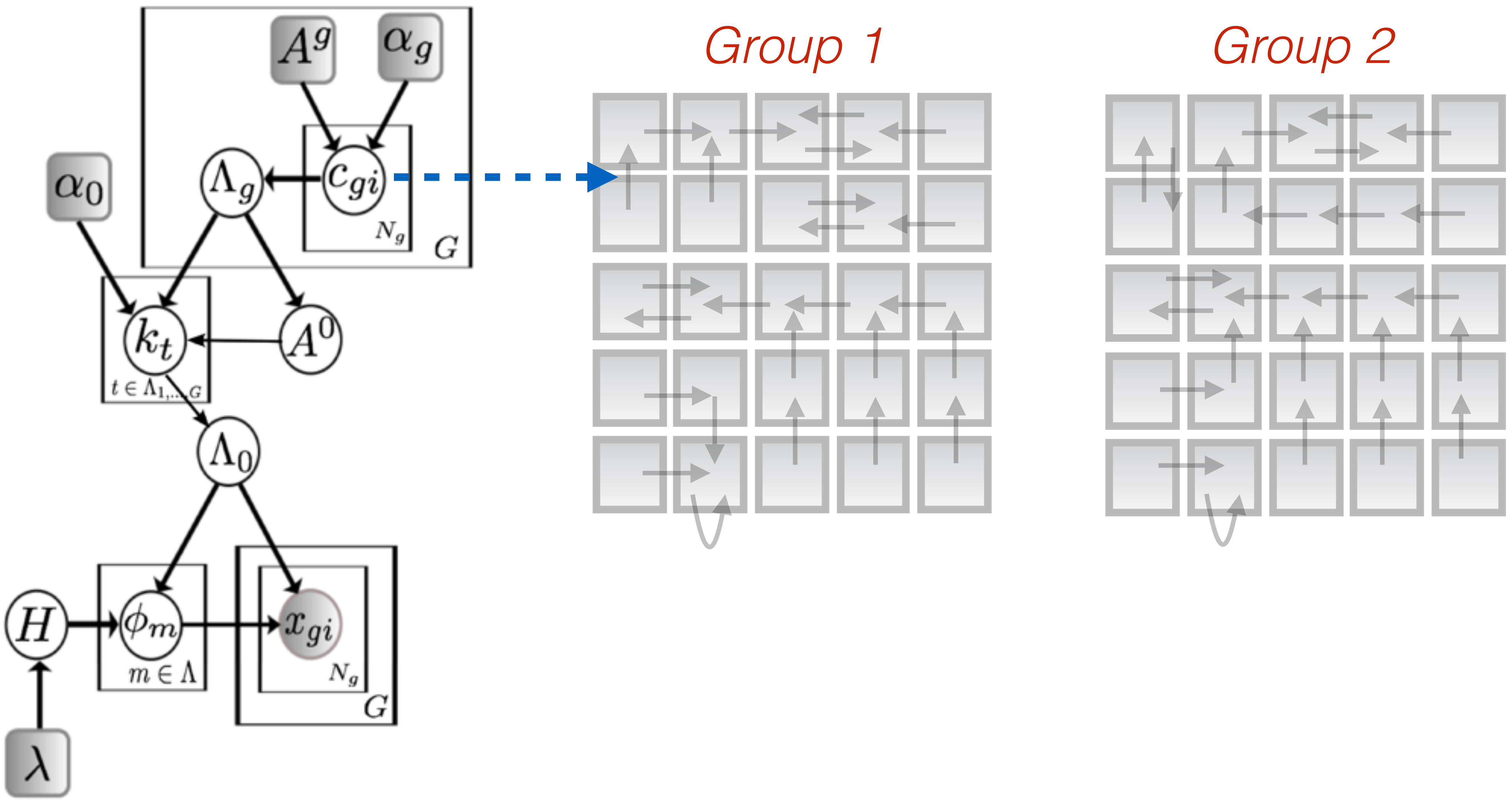
Hierarchical Distance Dependent Partitions



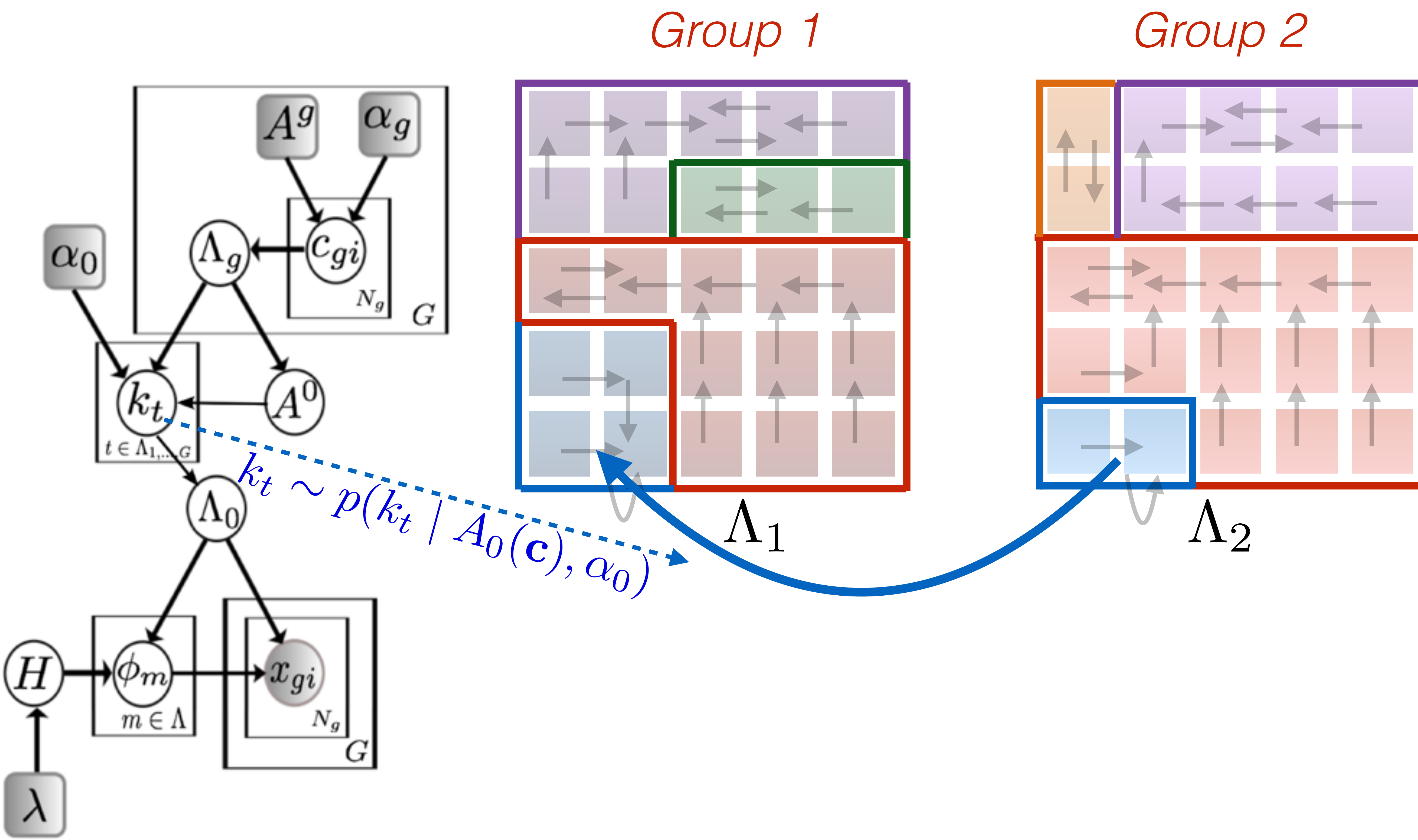
Model affinities between both data points and latent clusters.

Ghosh et al., UAI 2014

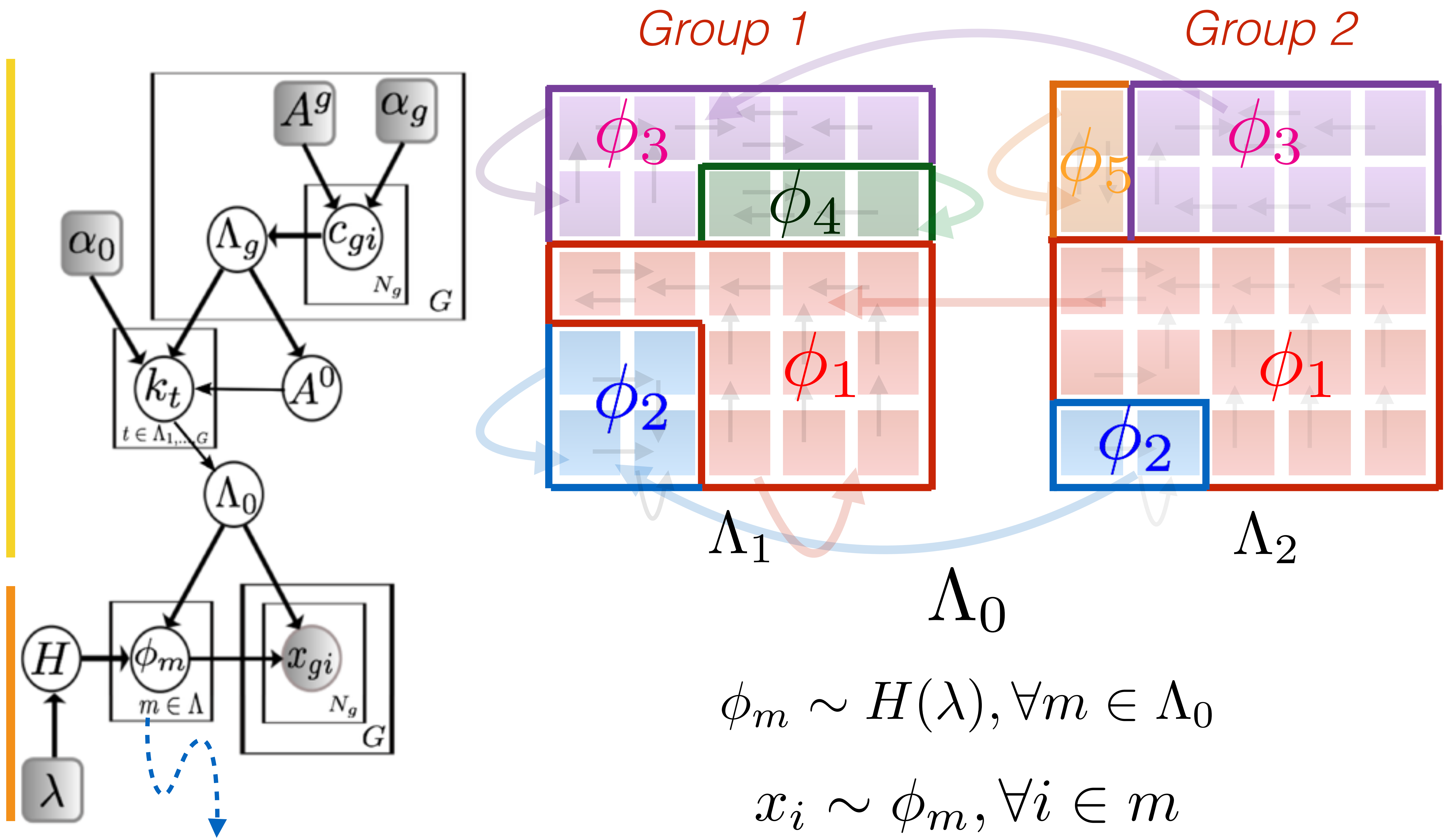
Hierarchical ddCRP



Hierarchical ddCRP



Hierarchical ddCRP

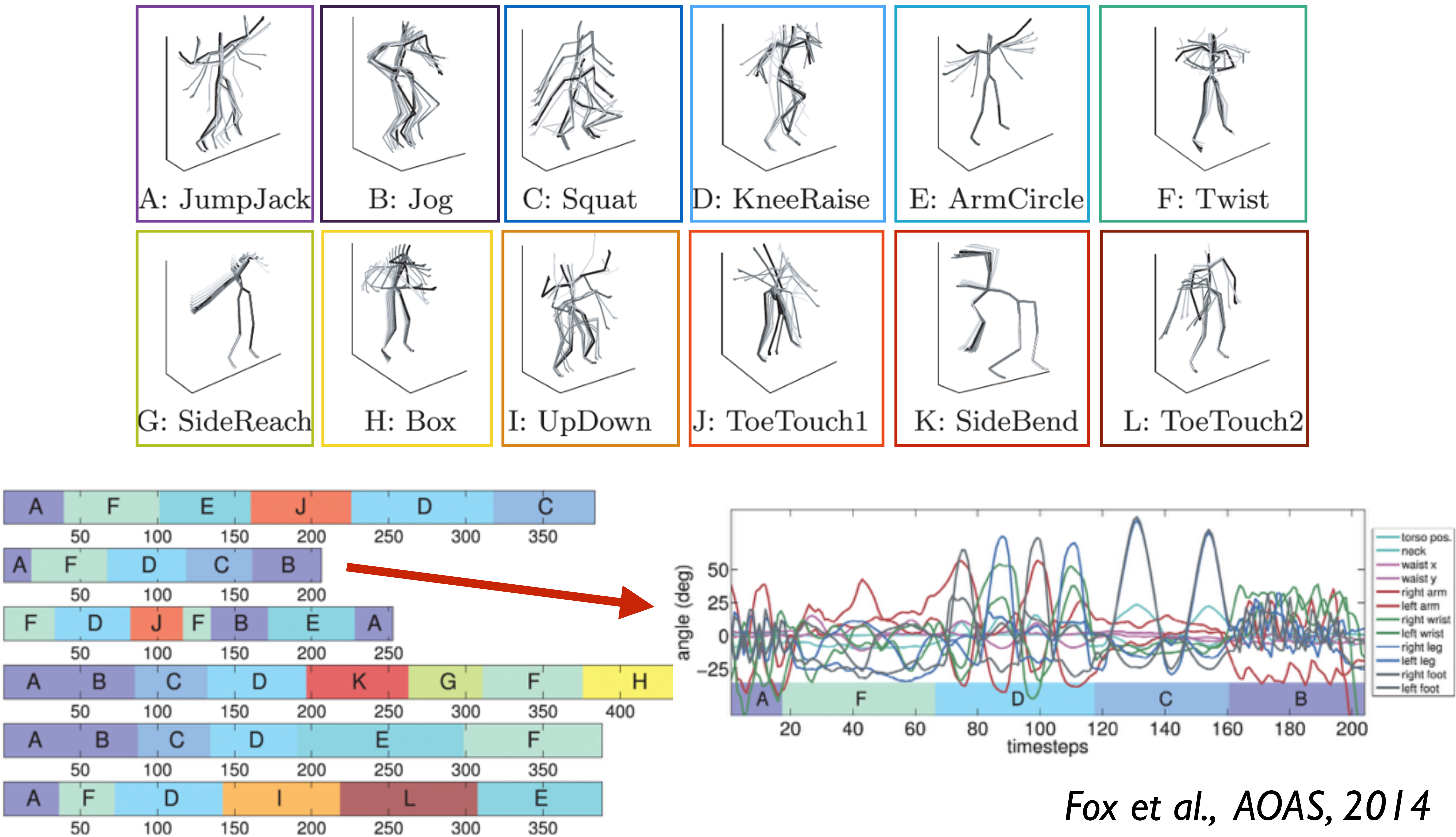


$$\phi_m \sim H(\lambda), \forall m \in \Lambda_0$$

$$x_i \sim \phi_m, \forall i \in m$$

Components shared across groups

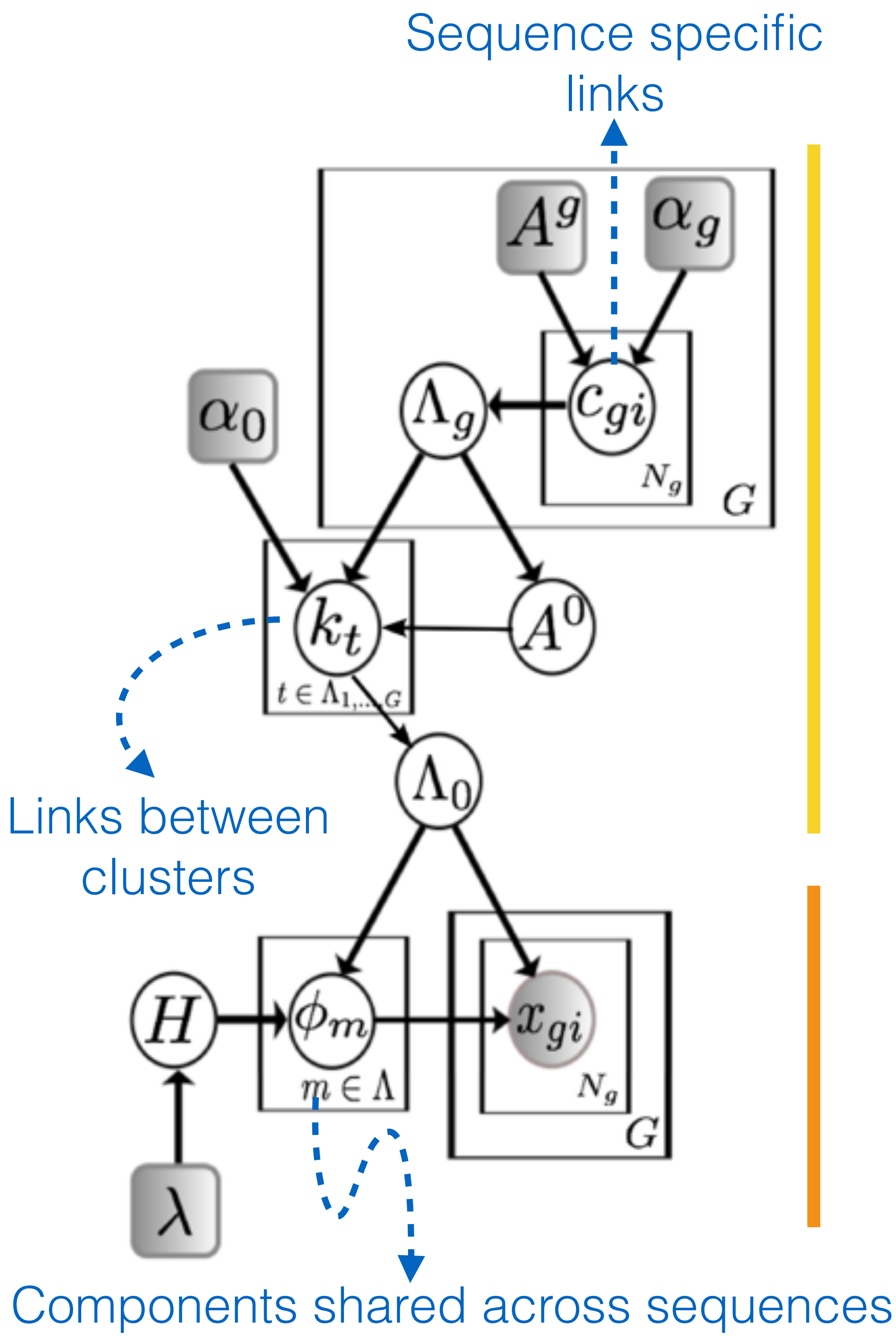
Activity Recognition



Fox et al., AOAS, 2014

mocap.cs.cmu.edu

Hierarchical Auto Regressive Mixtures



- Sequence specific ddCRP models:

$$p(c_{gi} = gj \mid \alpha_g, A^g) \propto \begin{cases} \exp(-\frac{(i-j)}{N_g^\gamma}) & i > j, \\ 0 & i < j, \\ 1 & i = j. \end{cases}$$

- Global CRP across sequences:

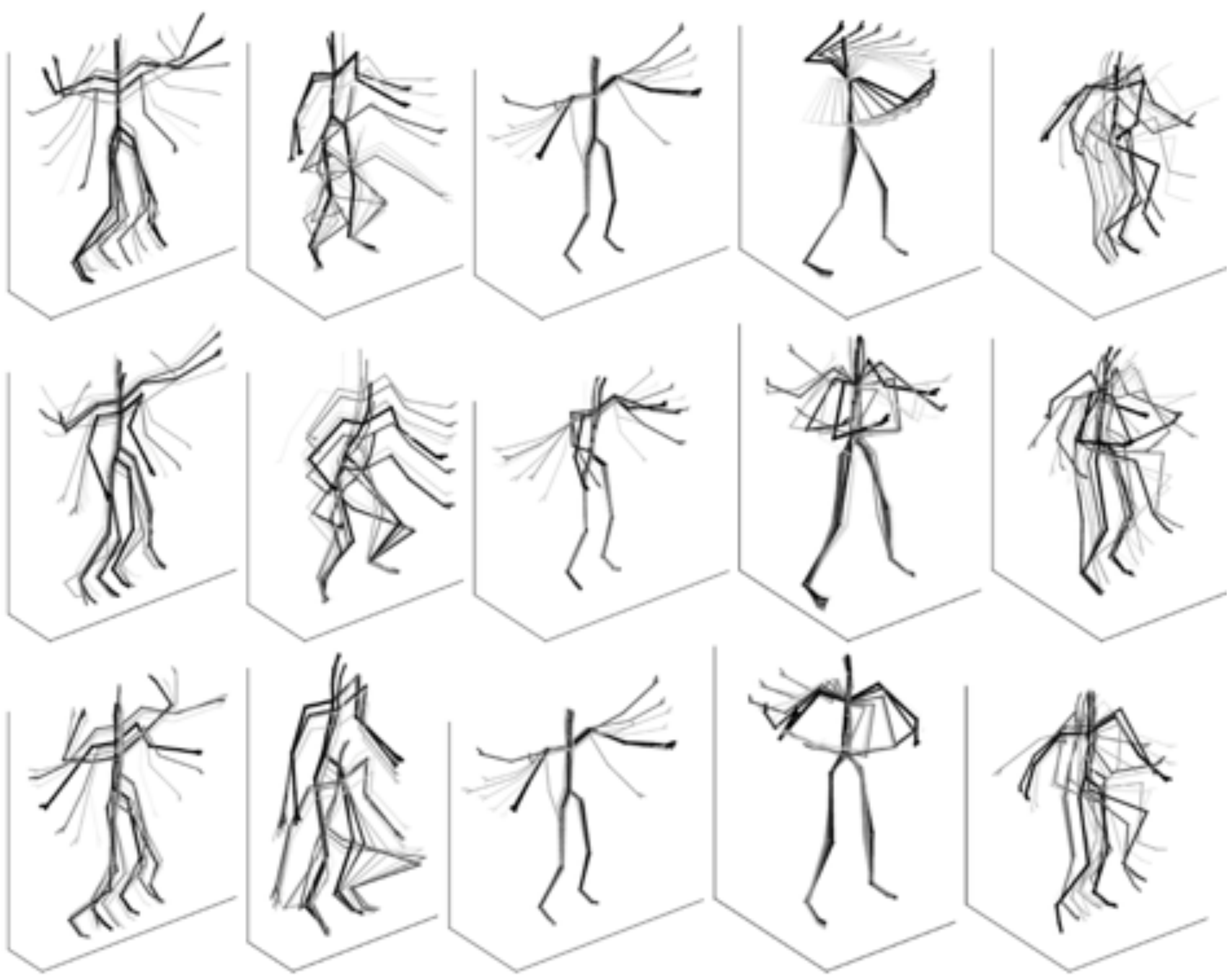
$$\Lambda_0 \sim \mathrm{CRP}(T(\mathbf{c}), \alpha_0)$$

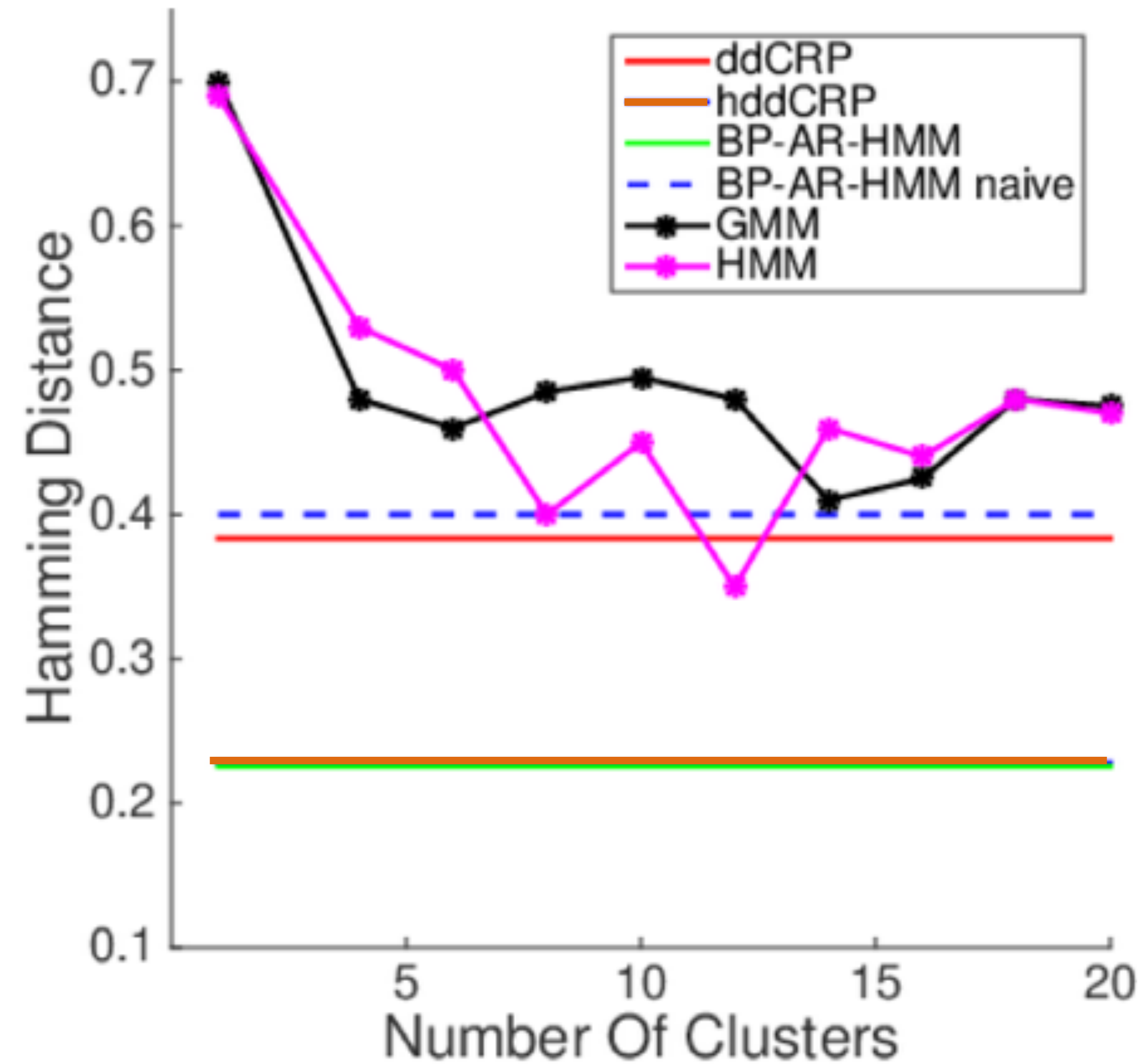
- Autoregressive likelihoods:

$$x_{gt} = B_m x_{gt-1} + \epsilon_m$$

$$B_m, \epsilon_m \sim H(\lambda)$$

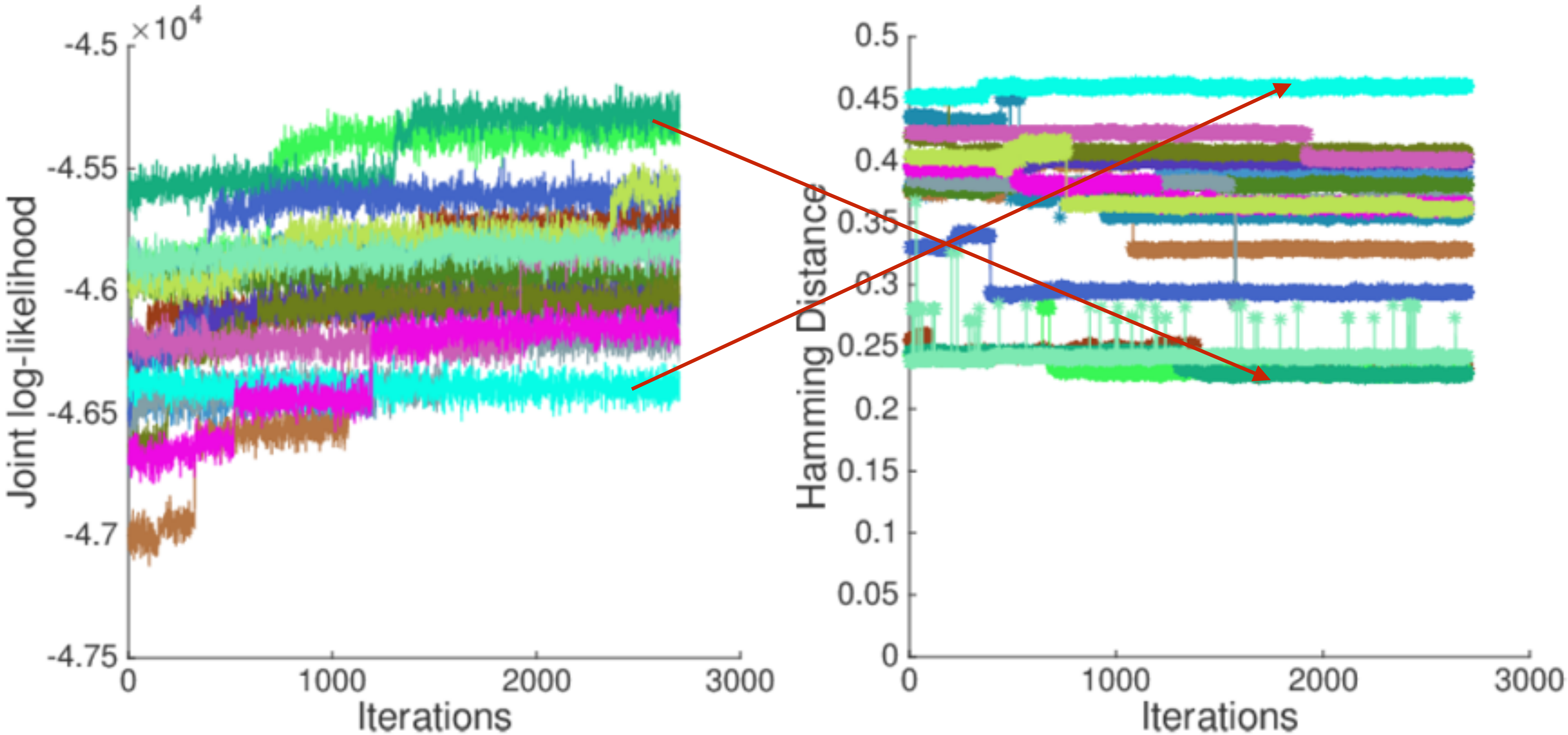
Discovered Activities





Examples of activities discovered by hddCRP

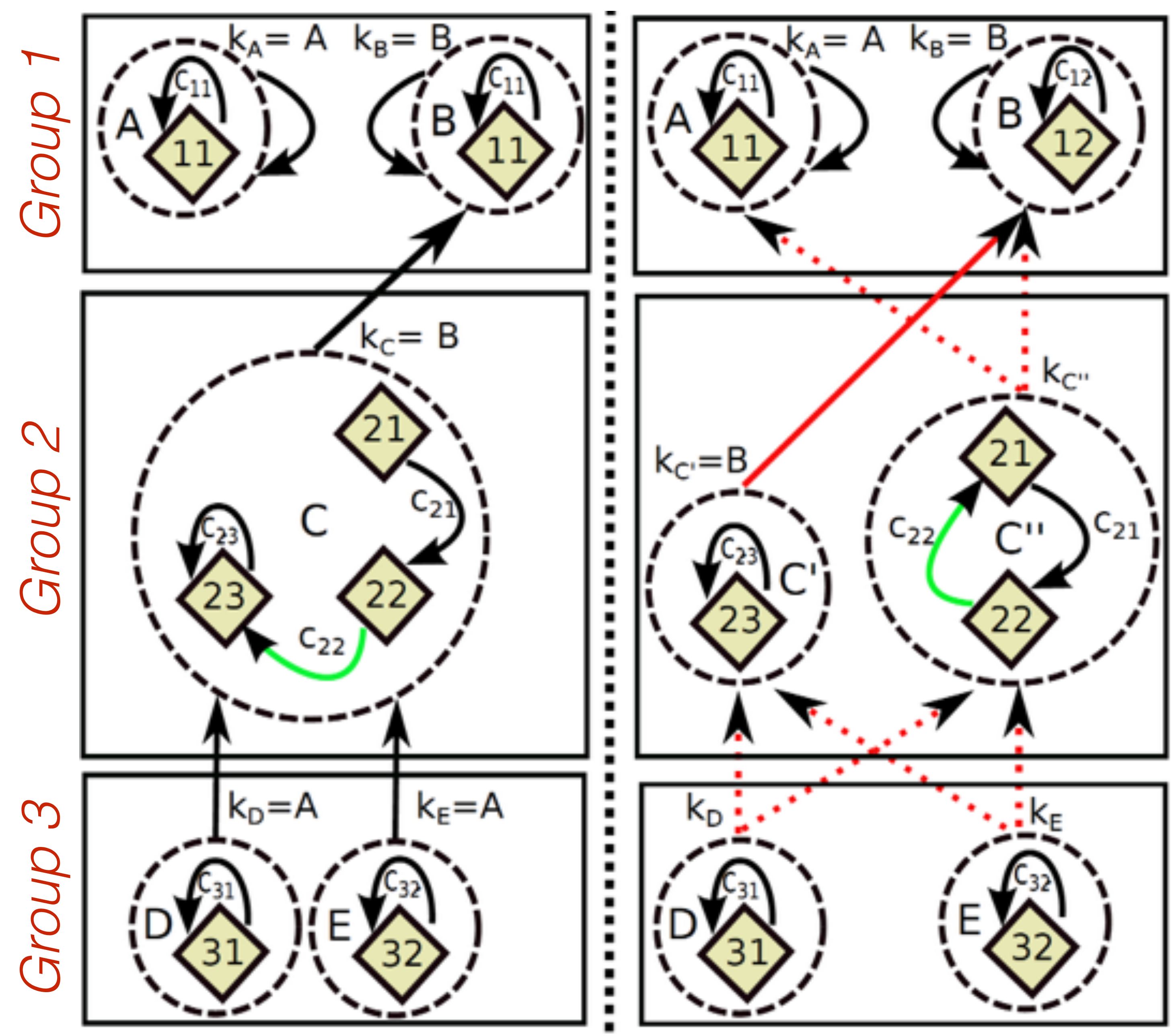
External Model Validation



15 independent MCMC chains

Inference

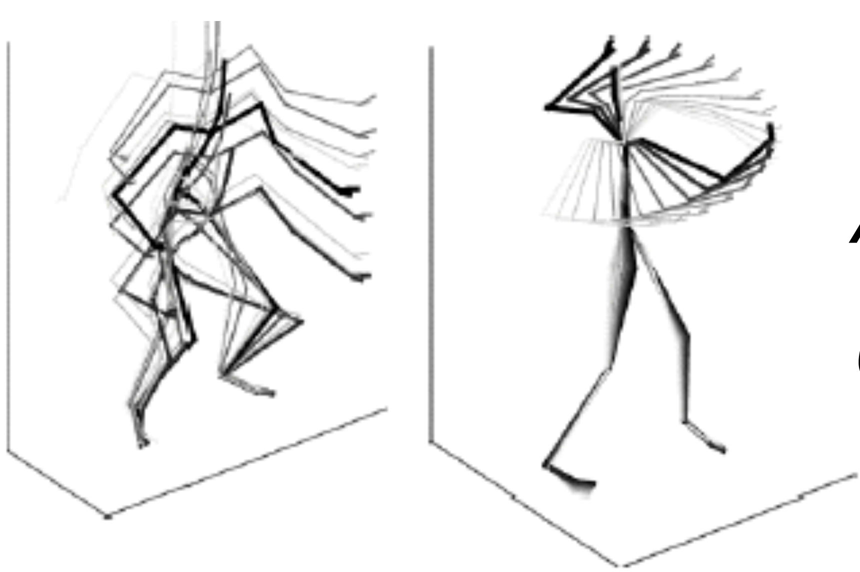
- More involved.
- No simple Gibbs sampler, need to resort to Metropolis Hastings.
- Nonetheless “efficient” MH samplers can be crafted.



Ghosh et al., UAI, 2014

Summary

Articulated object segmentation through ddCRP mixtures



Activity discovery via hierarchical distance dependent models

Talk Outline

- Distance dependent partitions
    - Parts from articulated 3D objects
- Hierarchical distance dependent partitions
    - Activity discovery from MoCap data
- Learning distance dependent models
    - Image and video segmentation

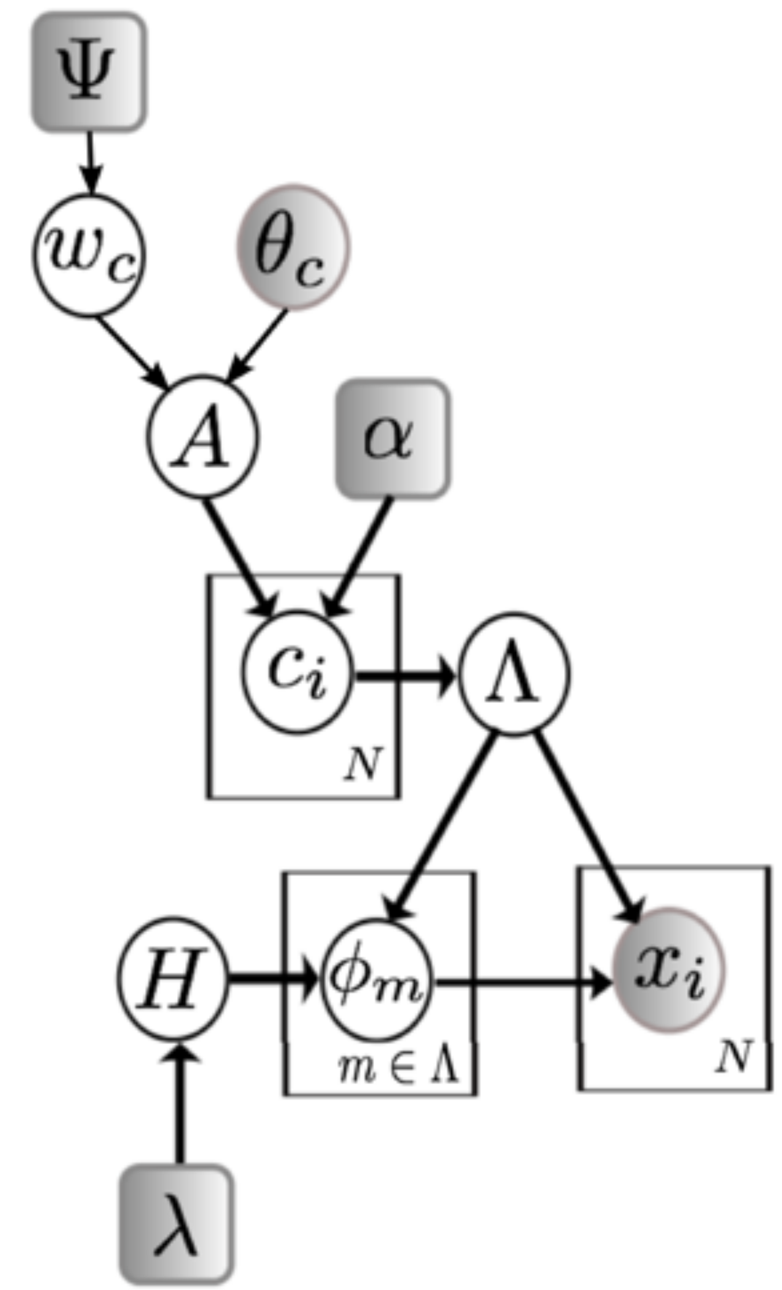
Feature Augmented Models

$$p(c_i = j \mid A) \propto A_{ij}$$

$$A_{ij} = f(w_c^T \theta_{ij}^c)$$

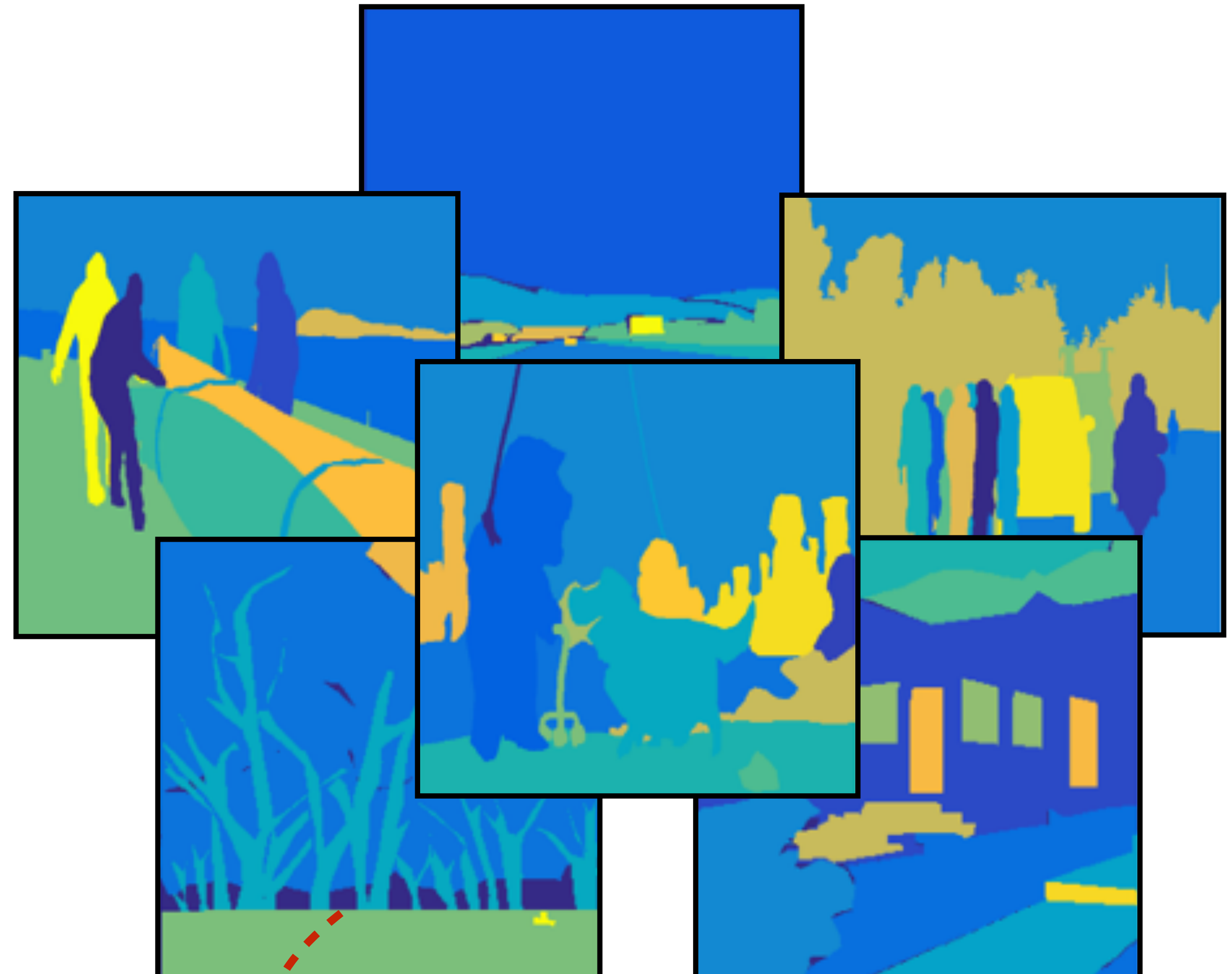
Features encoding similarity

Latent variables governing contribution of features



Learning From Partitions

- Moderate sized databases of partitions available for image and video collections.
- *Uncertainty* in labeled partitions
- Partitions are observed, but *links are not.*



$$y_i = \mathbb{N}^{N_k \times 1} = [1,1,2,4,\ldots,3,3]$$

$$Y = \{y_1 \ldots y_D\}$$

Approximate Bayesian Computation

- Noisy partitions - human interpretations vary
    - Appropriate noise model? Unclear, *ABC instead*
- Likelihood free inference:
    - Match "*interesting*" model statistic with observed data statistic

Marin et al., Stat Comput, 2012

Auxiliary Training Model

$$p(\mathbf{c}, w_c, Y) \propto p(w_c) \prod_{d=1}^{D} p(c_d \mid w_c) \mathbf{1}(z(\mathbf{c}_d), y_d)$$

$$\mathbf{1}(y_a, y_b) = \begin{cases} 1 \text{ if } \Delta(y_a, y_b) < \epsilon, \\ 0 \text{ otherwise.} \end{cases}$$

Probability restricted to partitions close to training data.

Marin et al., Stat Comput, 2012

Loss Aware Model

- Notion of closeness captured through a task specific loss function:

$$\Delta(y_a, y_b) = 1 - \mathrm{RI}(y_a, y_b)$$

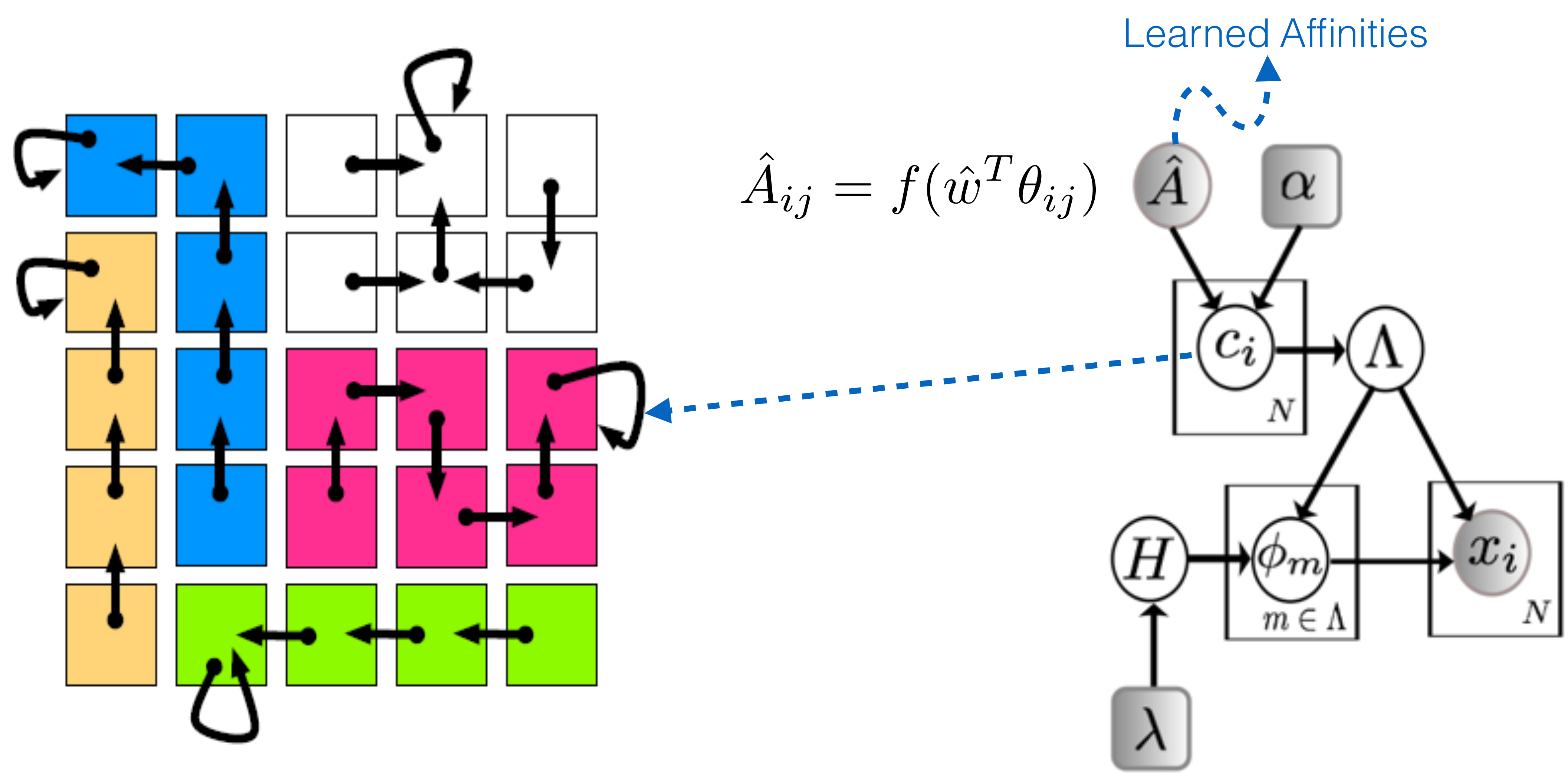
- *Marginalize* over the exponentially large space of *latent links* using MCMC

- Efficient ABC variant for sampling from the *auxiliary training model*

Talk Outline

- Distance dependent partitions
  - Parts from articulated 3D objects
- Hierarchical distance dependent partitions
  - Activity discovery from MoCap data
- Learning distance dependent models
  - Image and video segmentation

Image Segmentation



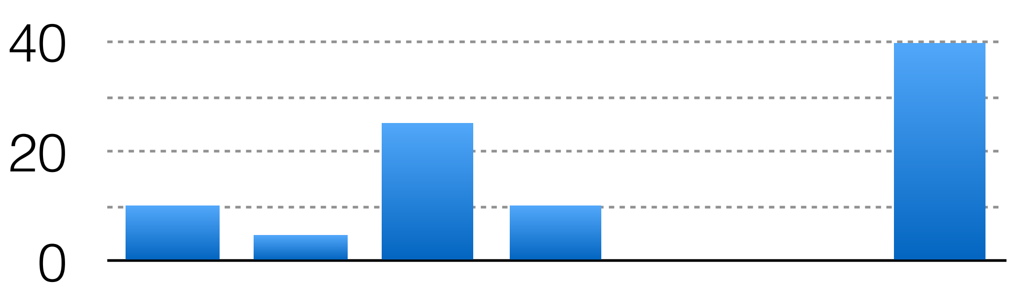
Generative features:

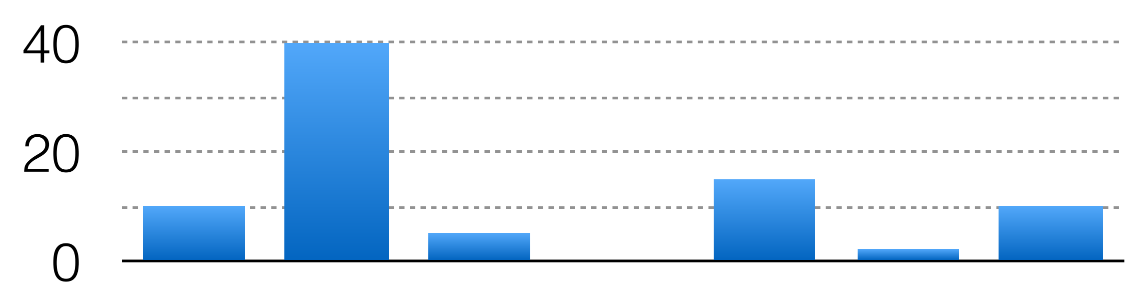
$$\theta_{ij} = \{\mathrm{row}_i - \mathrm{row}_j,\ \mathrm{col}_i - \mathrm{col}_j\}$$

Conditional features:

$$\theta_{ij} = \{\mathrm{row}_i - \mathrm{row}_j,\ \mathrm{col}_i - \mathrm{col}_j, \mathrm{edge}_{ij}\}$$

Image Representation





$$x_i^{color} \sim \mathrm{Mult}(\phi_m^{color}) \qquad x_i^{texture} \sim \mathrm{Mult}(\phi_m^{texture})$$

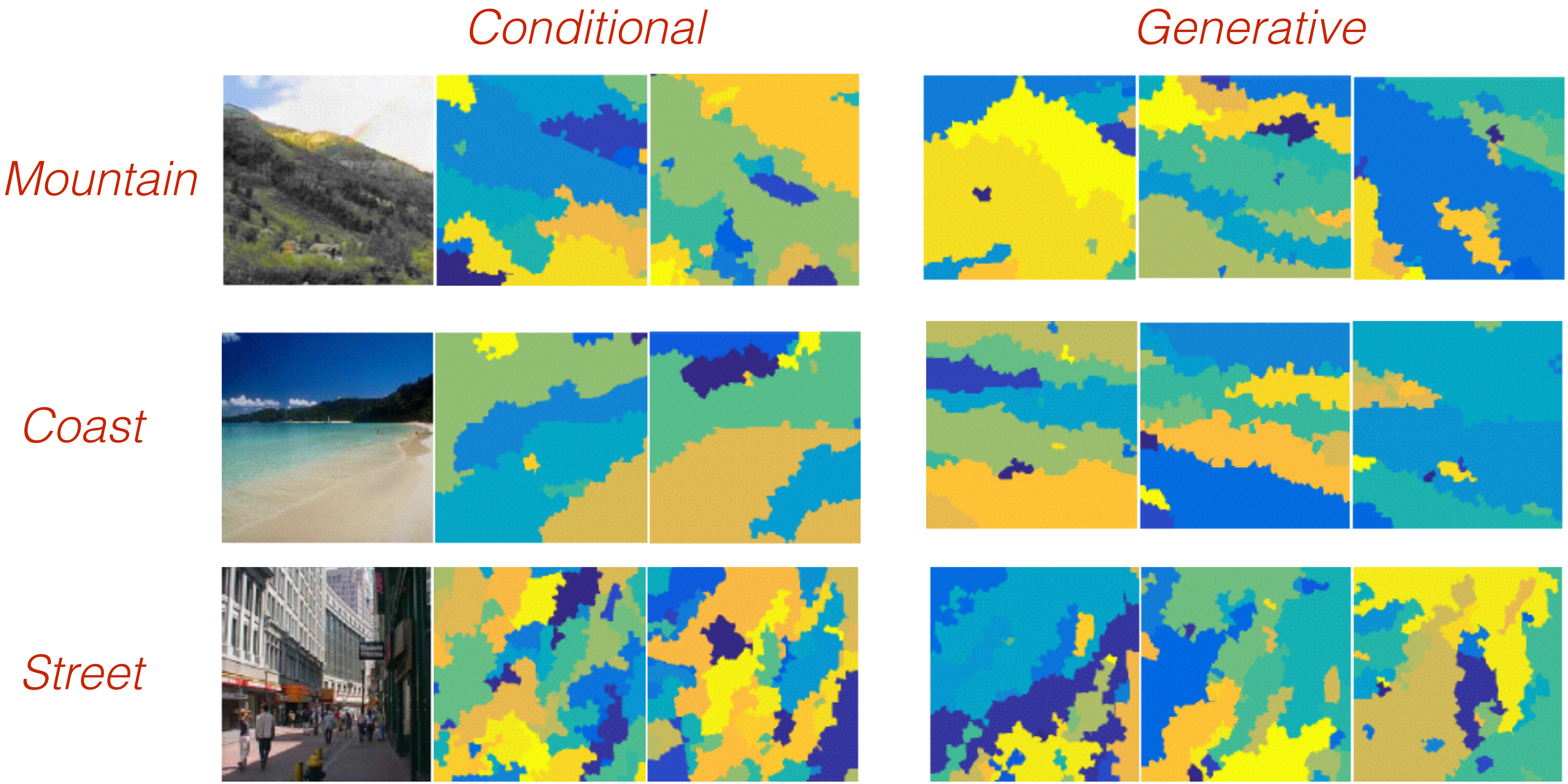
Each super-pixel is described through histograms (~120 bin) of color and texture

Eight Natural Scene Category Dataset (LabelMe)

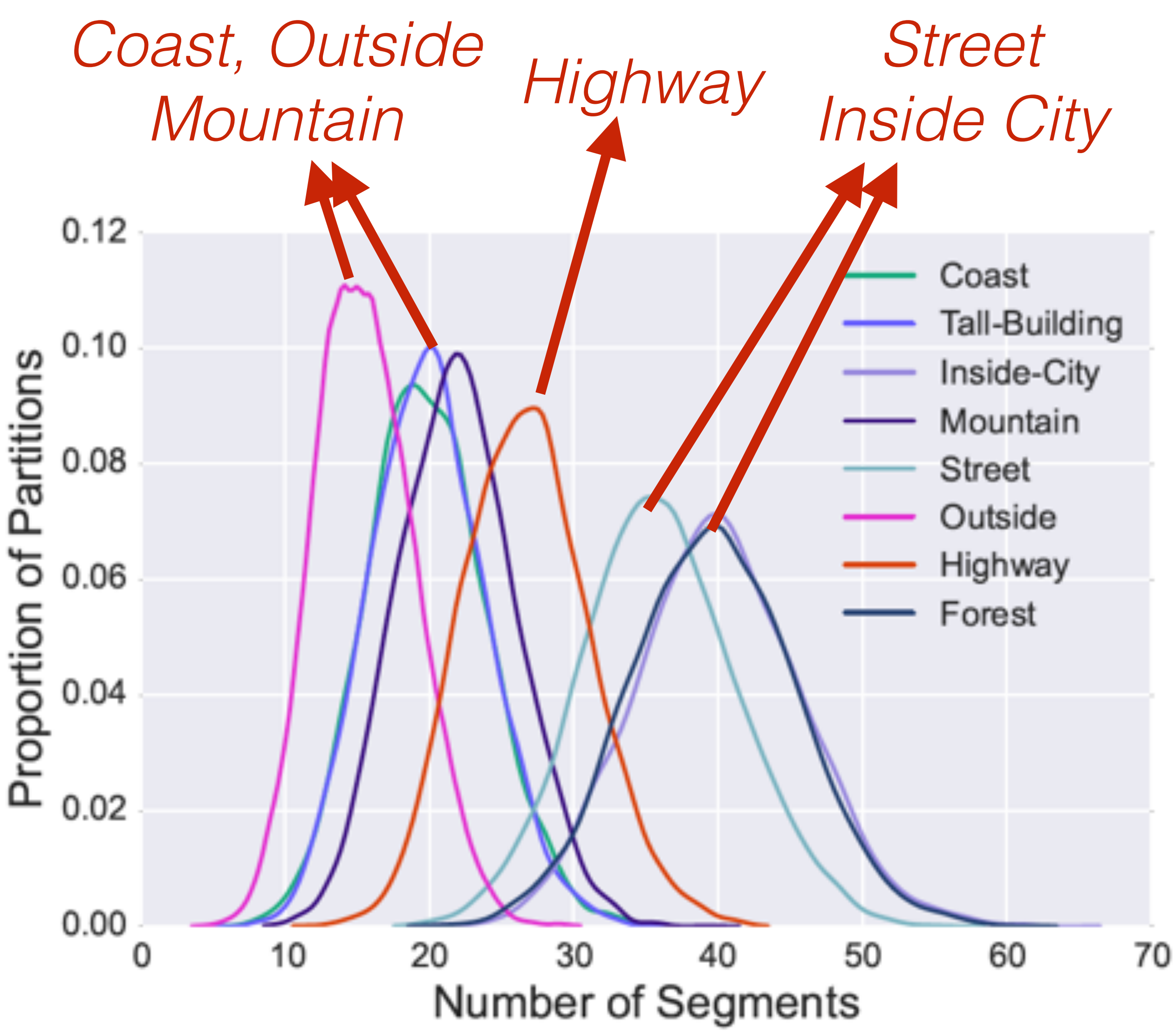
400 train and 800 test images

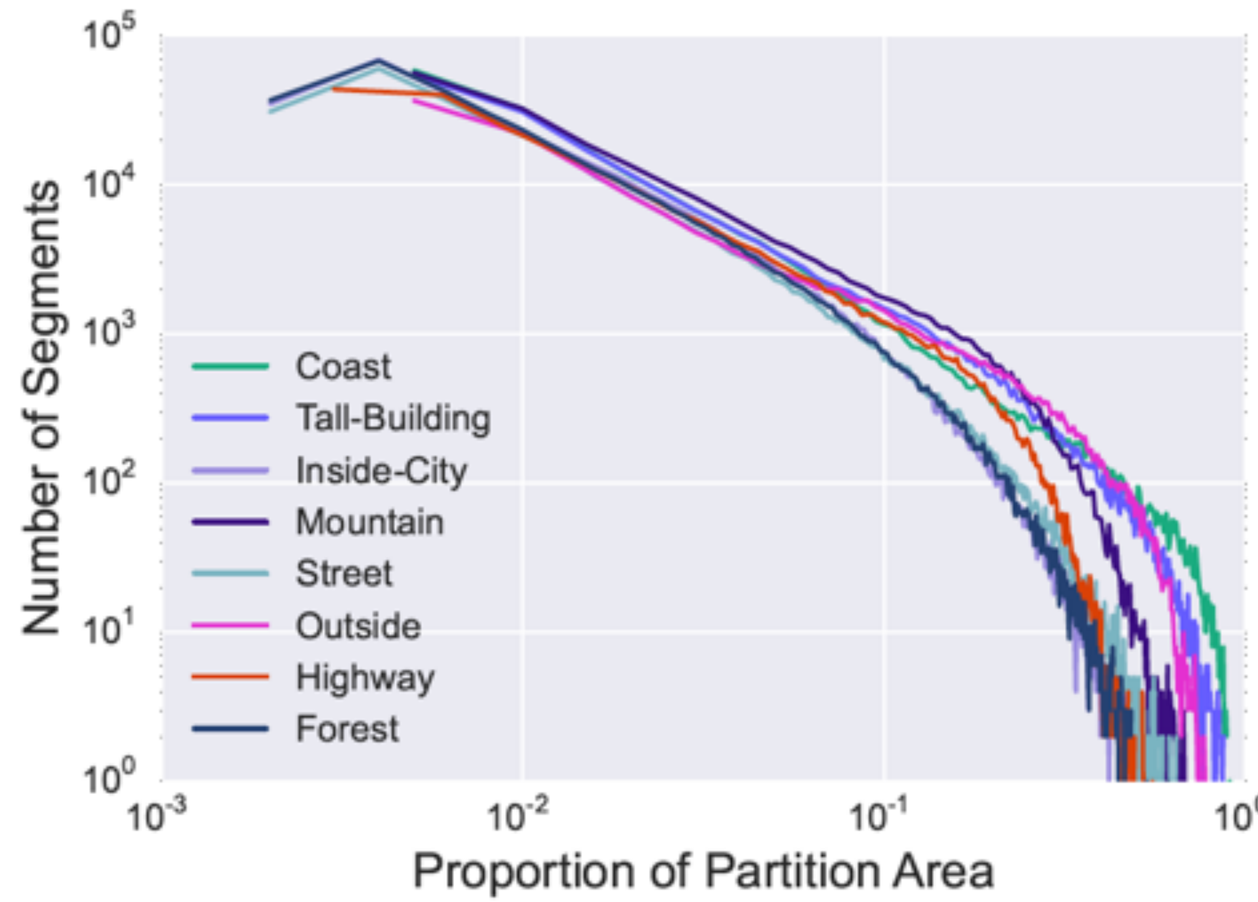
Oliva and Torralba, 2001

Samples from learned models



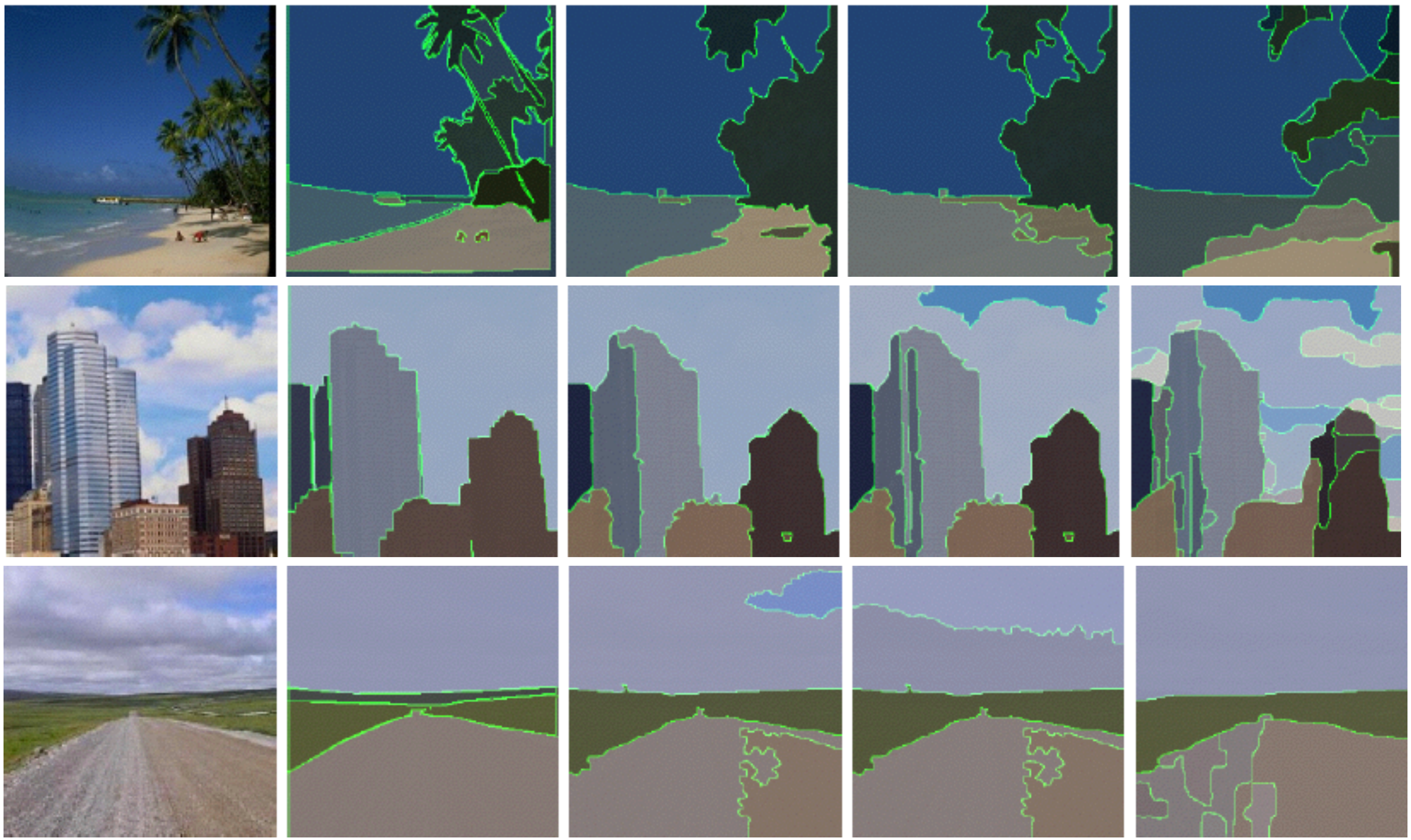
Monte Carlo Statistics





Statistics from 10,000 partitions sampled from generative affinities

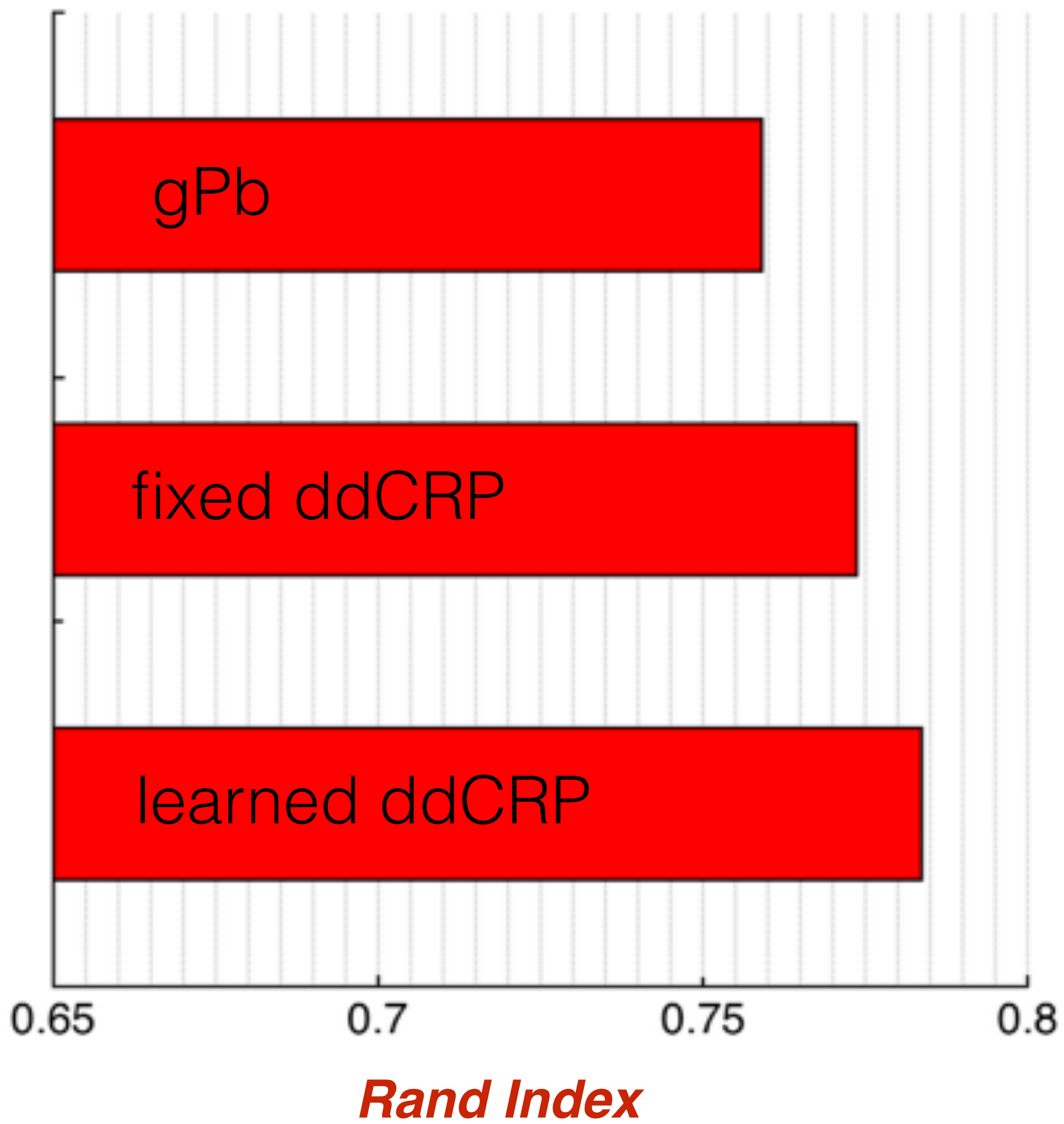
Qualitative Results



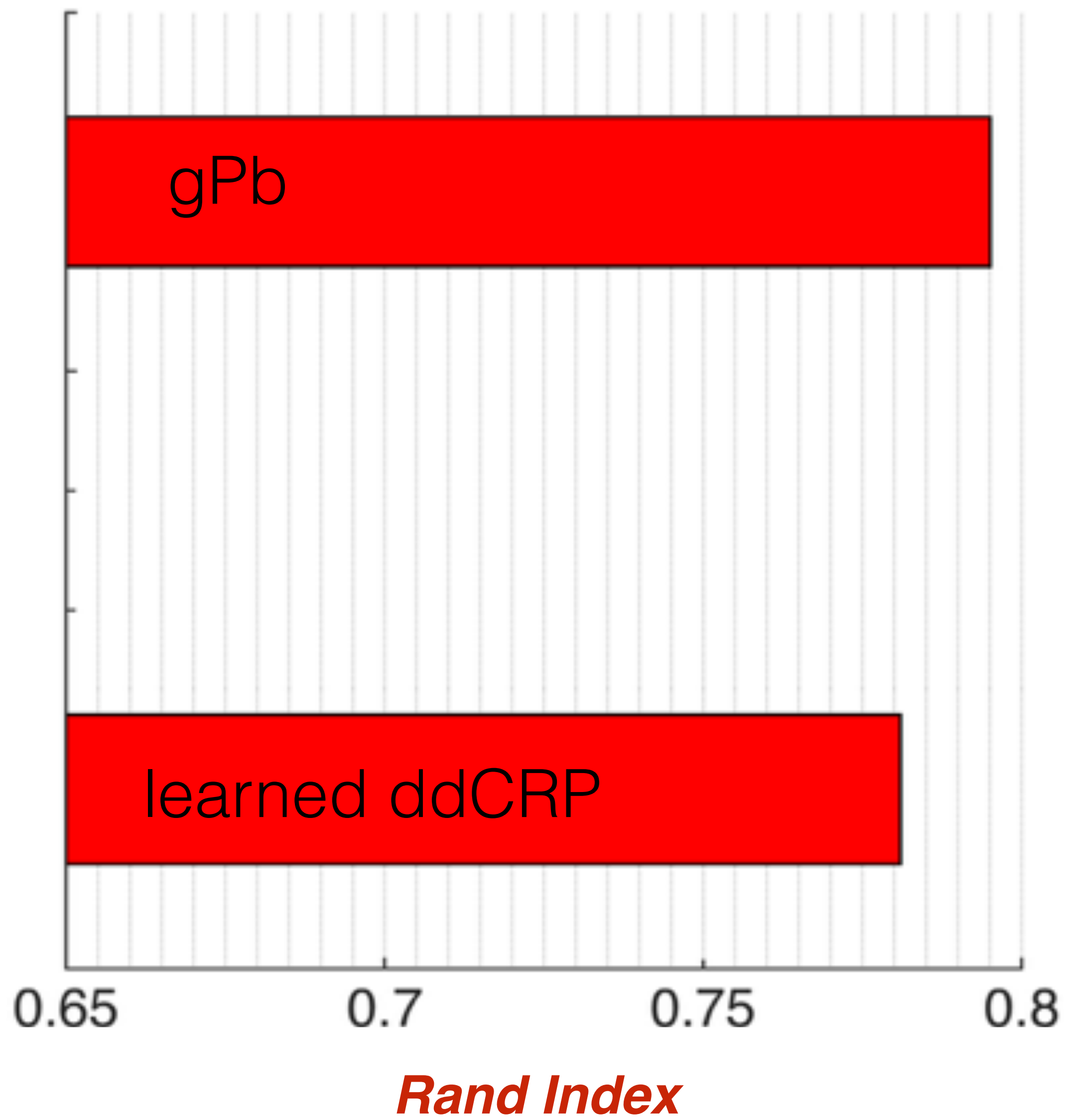
GT learned ddCRP fixed ddCRP gPb

Quantitative results

LabelMe



BSDS300

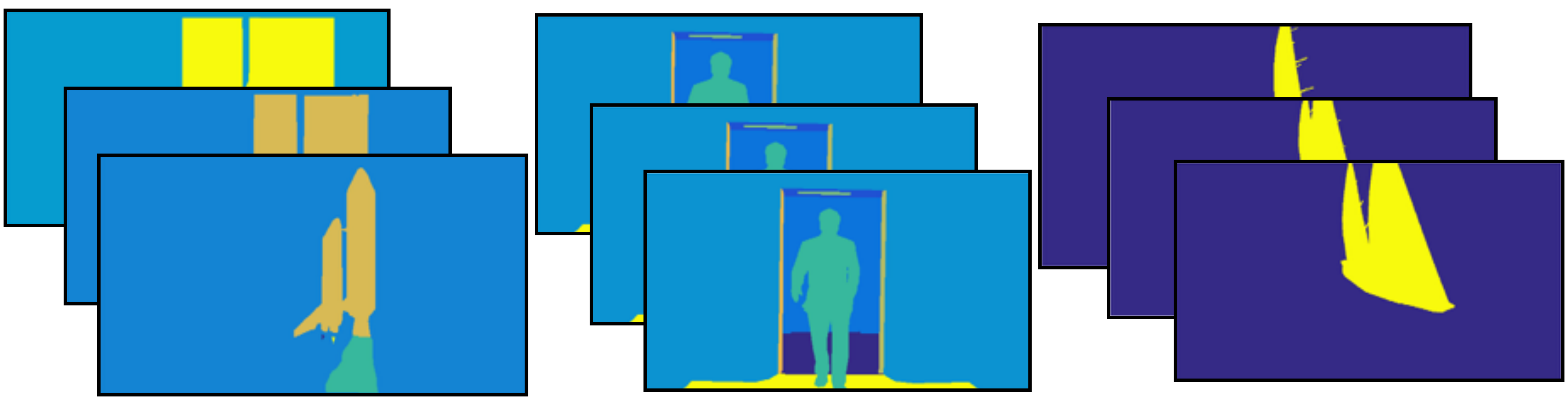


Learning in hierarchical models

- Auxiliary model for training now needs to account for links between clusters

$$p(\mathbf{c}, \mathbf{k}, w, Y) \propto p(w) \prod_{d=1}^{D} p(c_d \mid w_c) p(k_d \mid c_d, w_k) \mathbf{1}(z(\mathbf{c}_d, k_d), y_d)$$

$$w = \{w_c, w_k\}$$



VSB 100 - 40 training videos

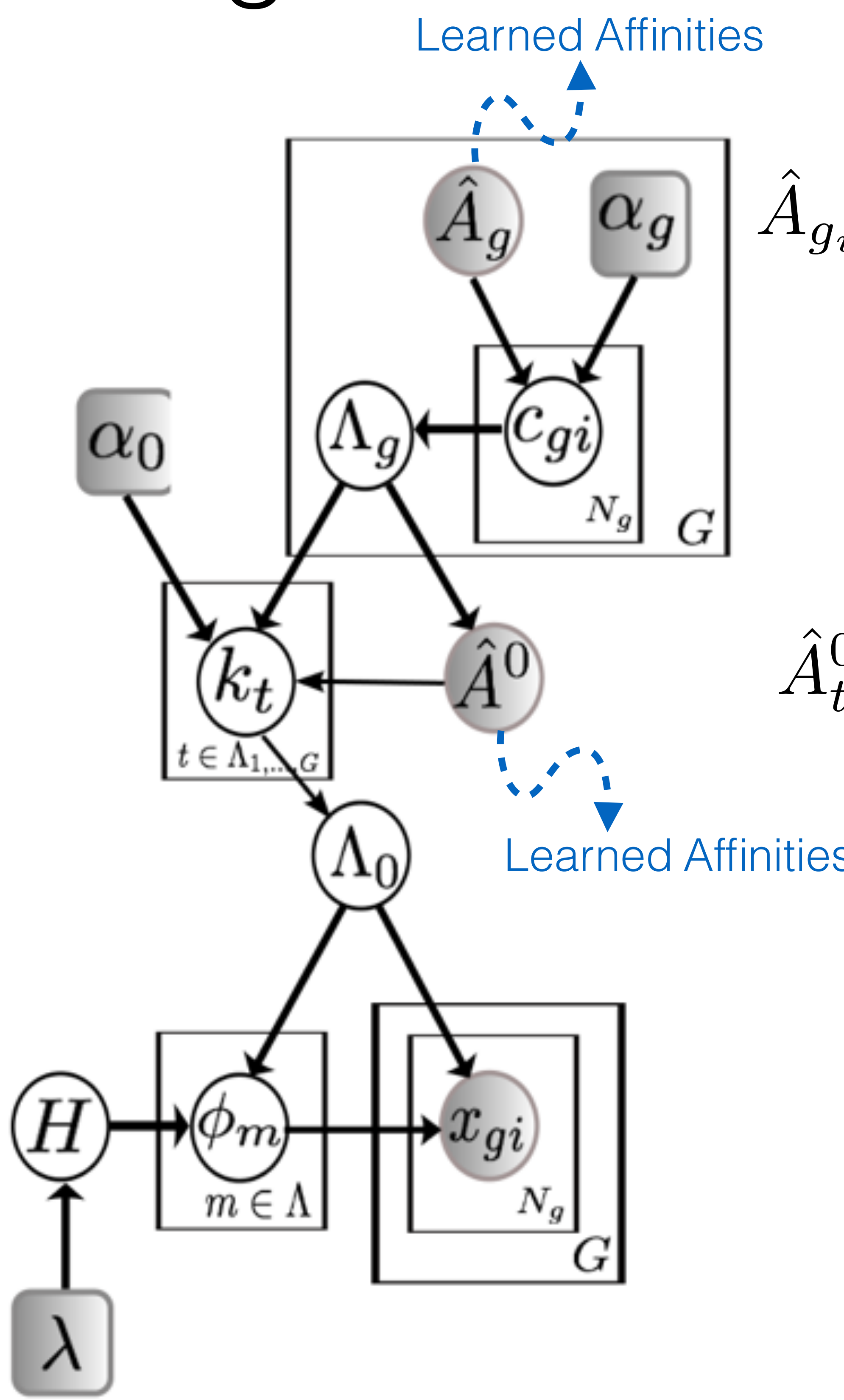
Video Segmentation

Features between superpixels:

$$\theta_{ij} = \{\mathrm{row}_i - \mathrm{row}_j,\\ \mathrm{col}_i - \mathrm{col}_j,\\ \mathrm{edge}_{ij}\}$$

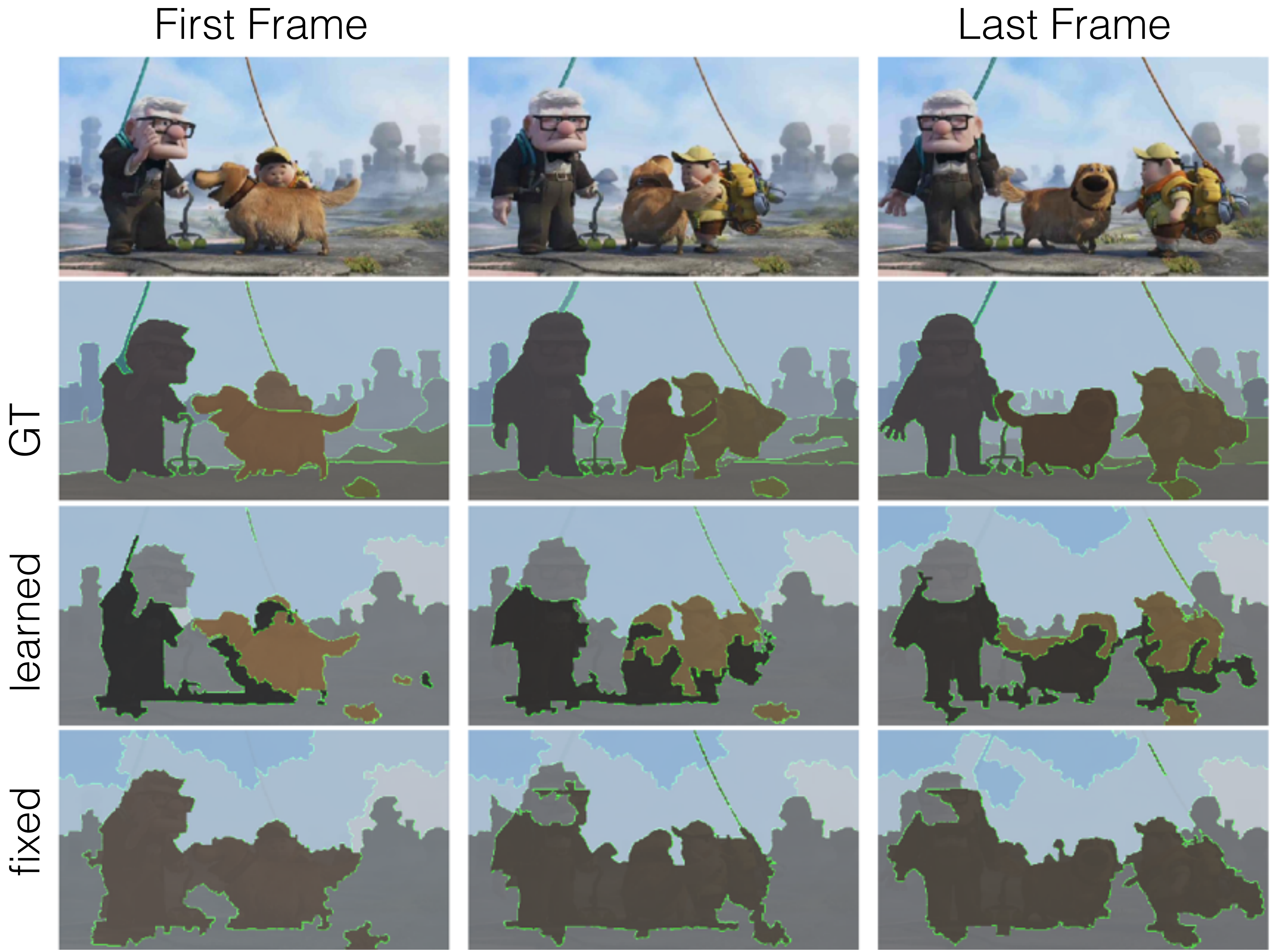
Features encoding similarity between segments:

$$\theta^k_{ts} = \{\psi(size_{ts},\\ shape_{ts},\\ locations_{ts})\}$$



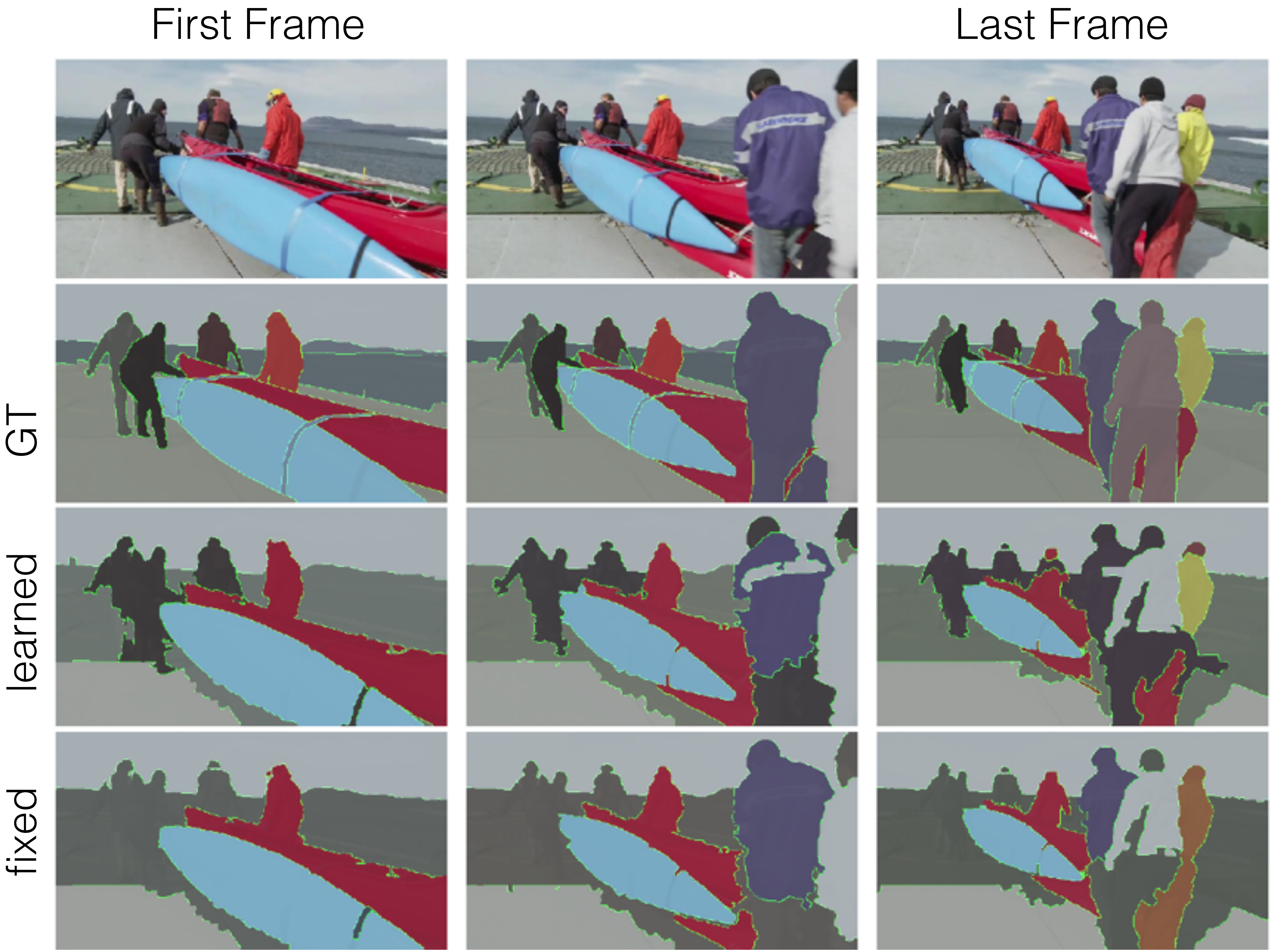
$$\hat{A}_{g_{ij}} = f(\hat{w}_c^T \theta_{ij})$$

$$\hat{A}^0_{ts} = f(\hat{w}_k^T \theta^k_{ts})$$

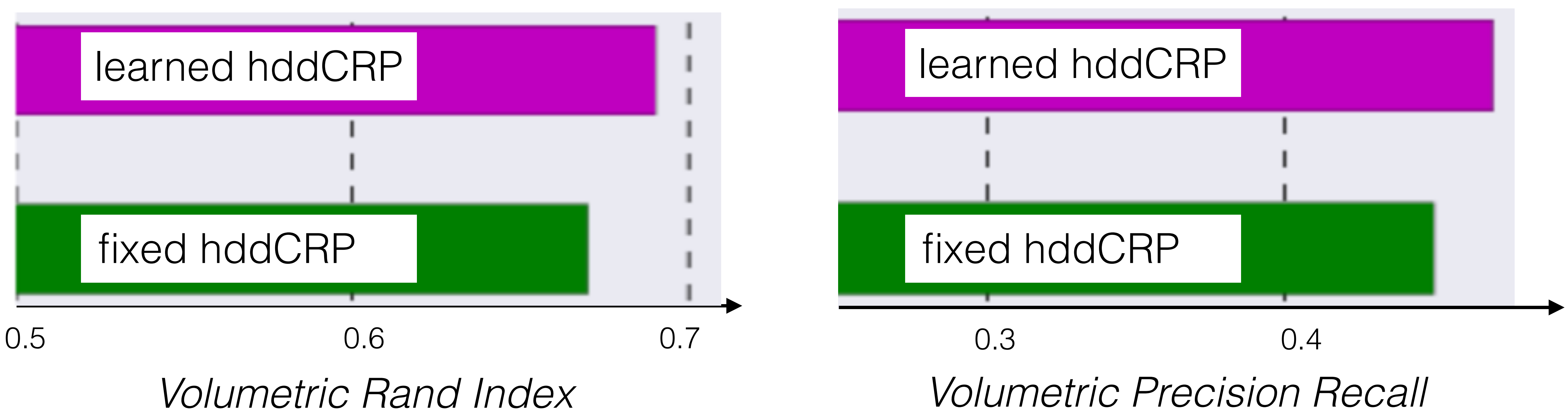


First Frame
Last Frame
GT
learned
fixed

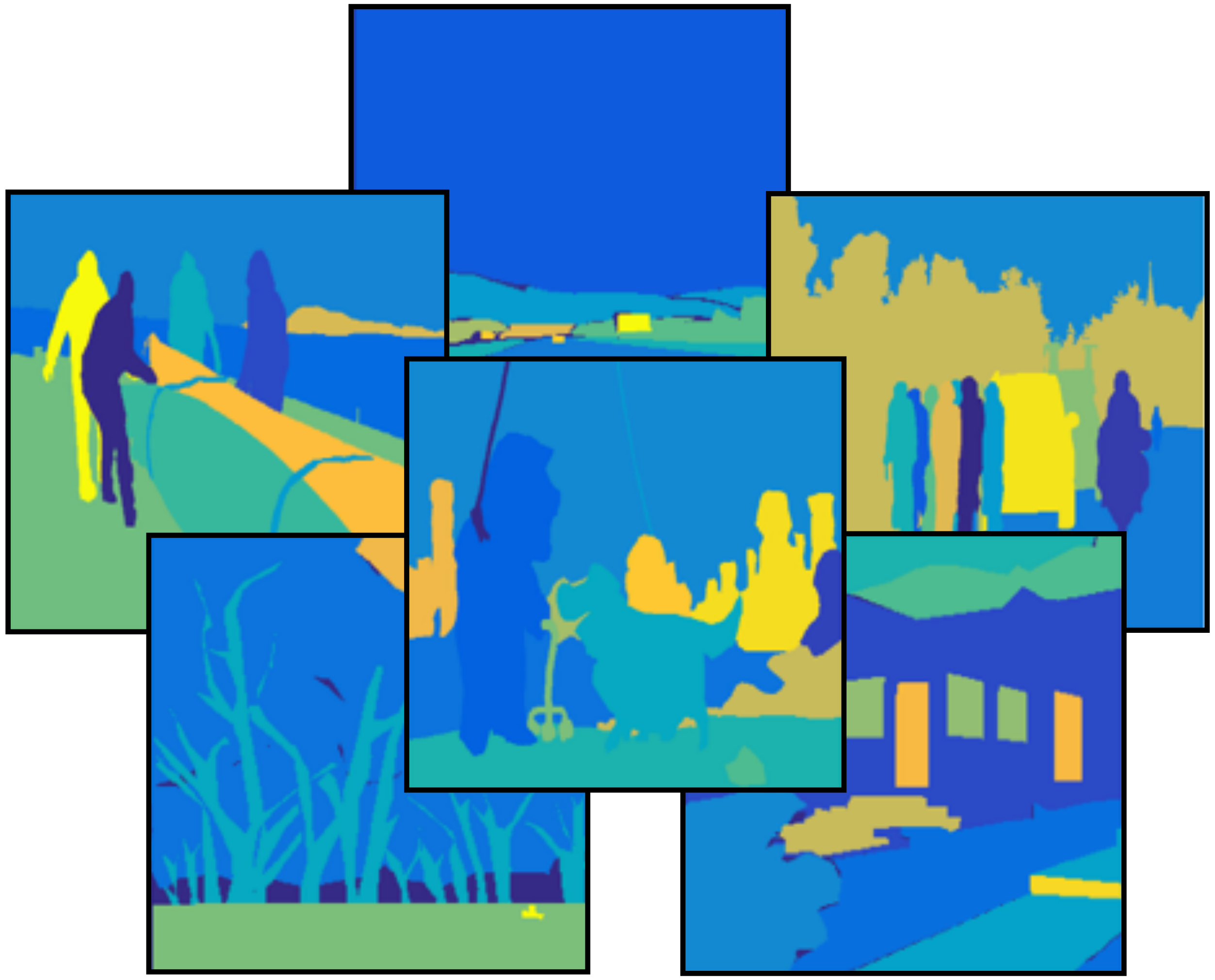

First Frame
Last Frame
GT
learned
fixed

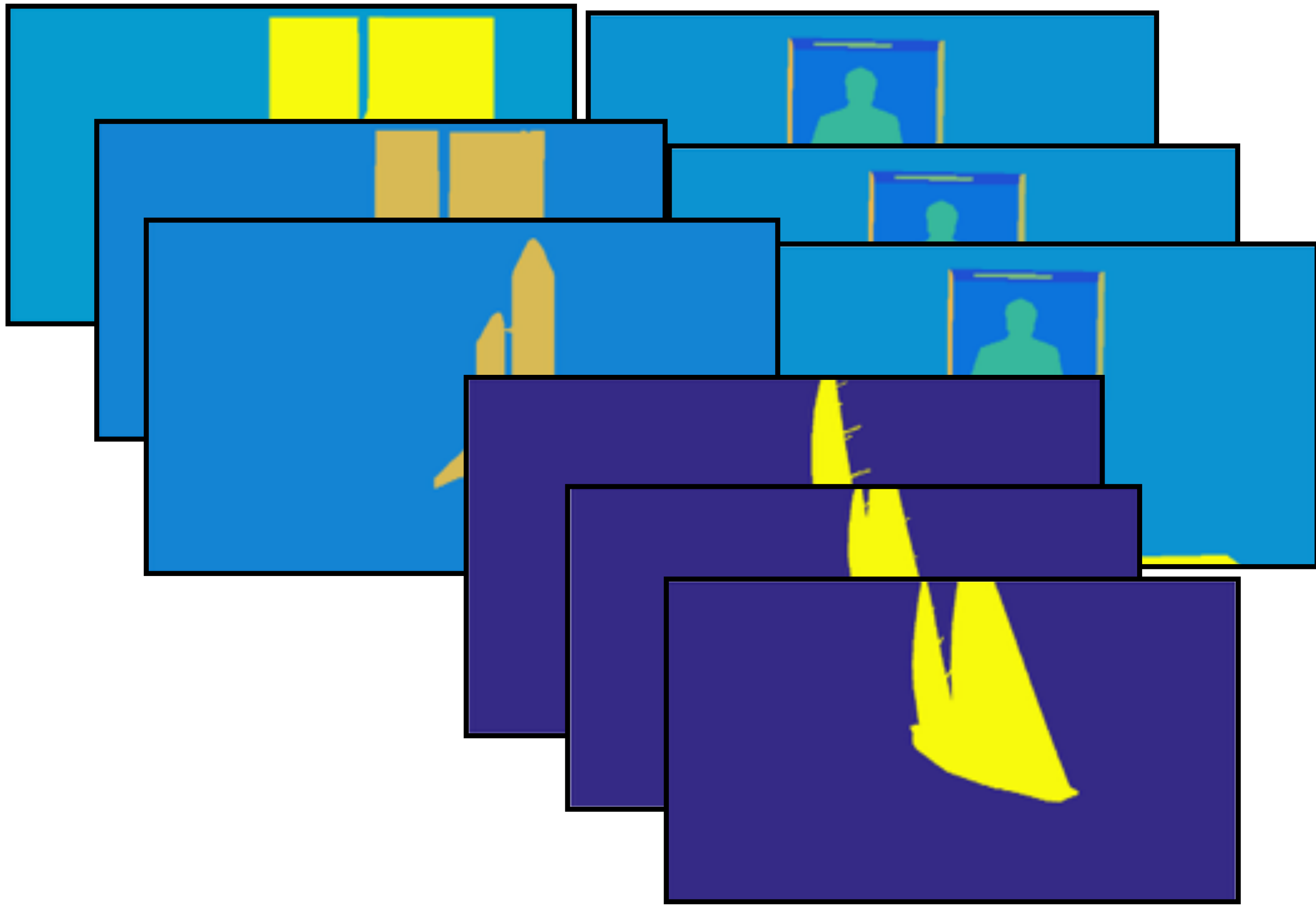

First Frame
Last Frame
GT
learned
fixed

Learning benefits hddCRP



Summary





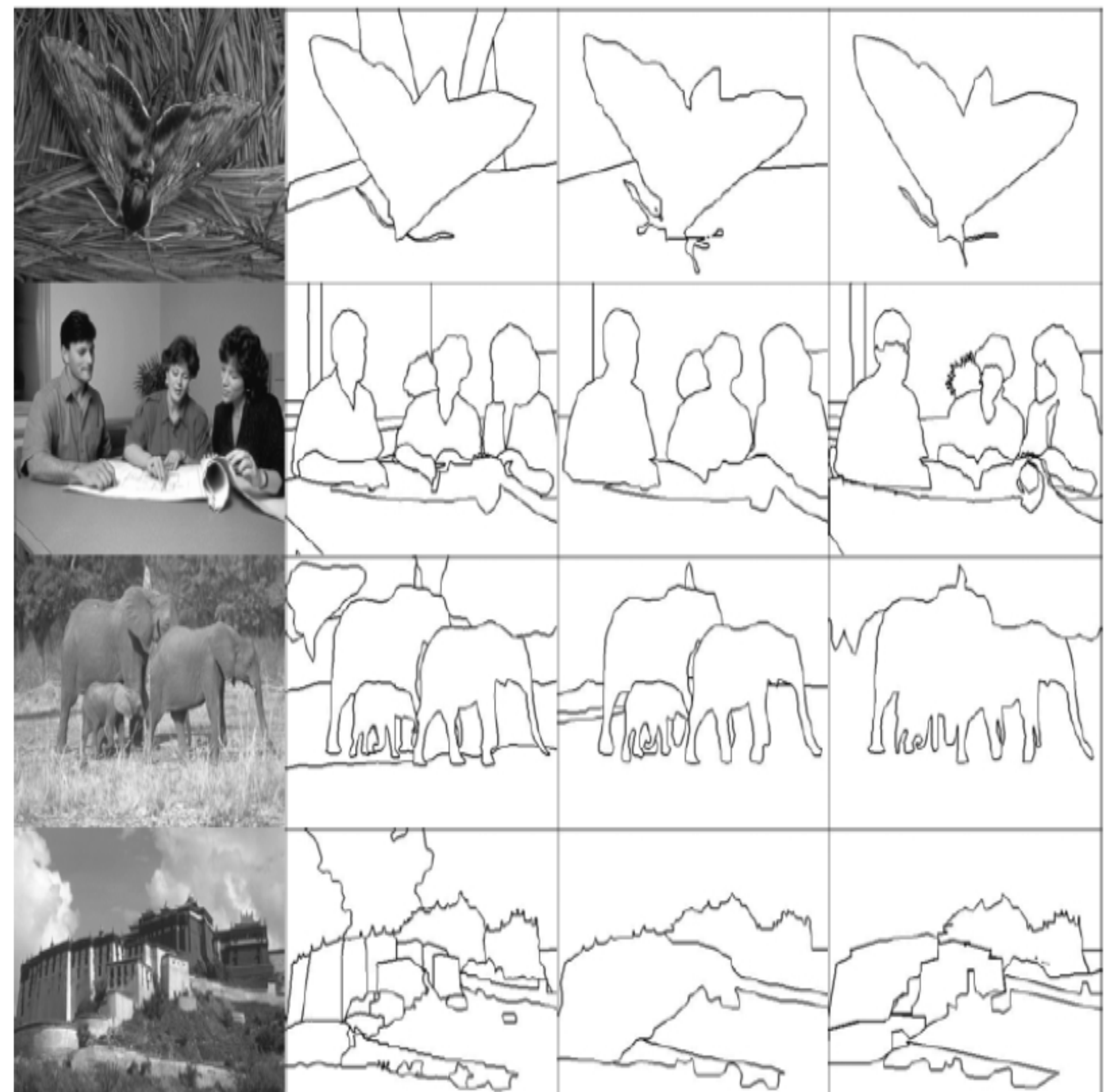
hddCRP and ddCRP affinities can be effectively learned from labeled partitions

Thank You

Questions?

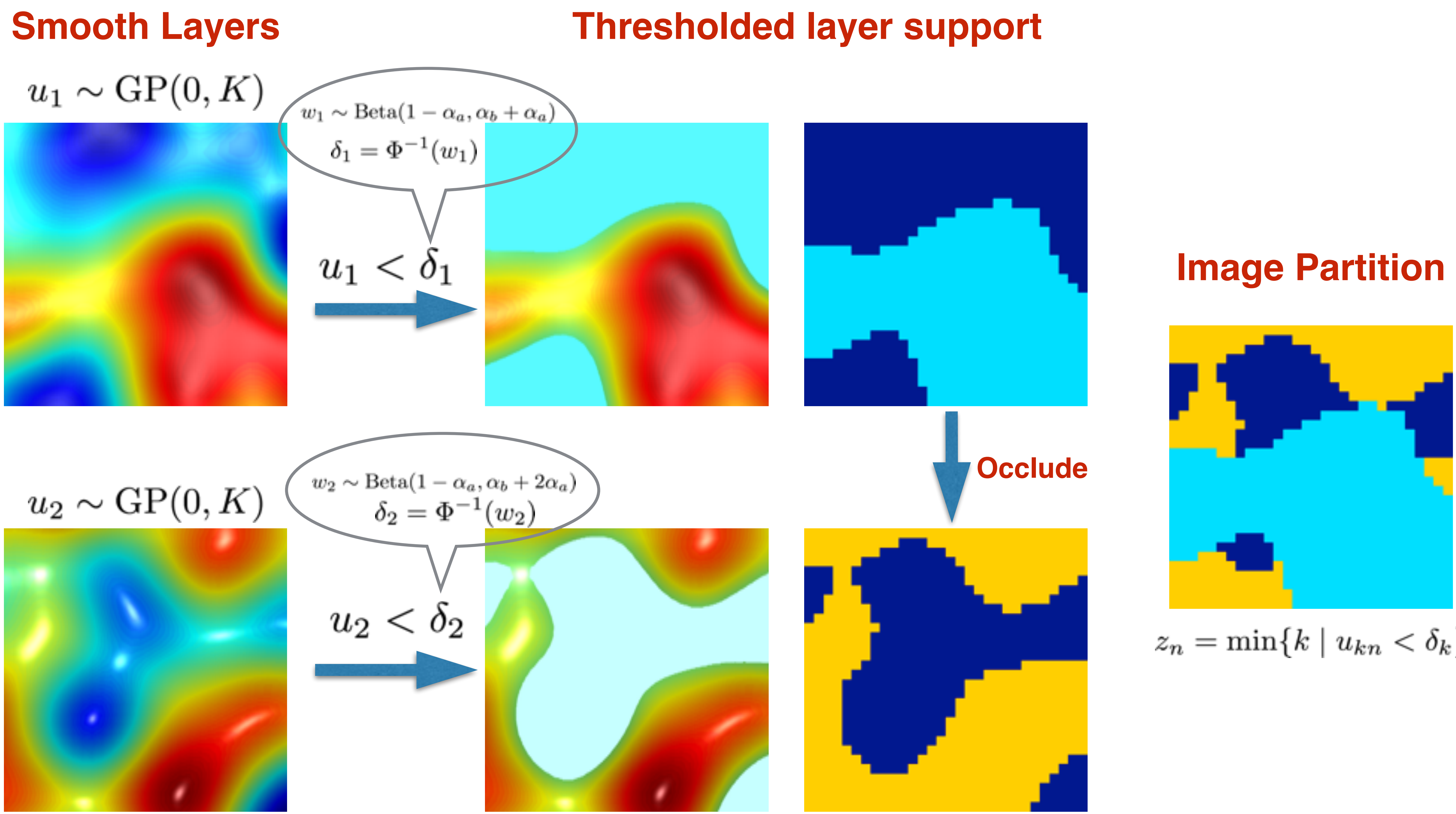
Statistics of Human Segments

- Human segment sizes follow power law behavior.



Sudderth & Jordan, NIPS 2008

Spatial Coupling through Layers



Sudderth & Jordan, 2008
Ghosh & Sudderth, 2012

Video Segmentation

- Features between superpixels — same as image segmentation.
- Features between segments — Shapes, sizes and positions.

$$\theta_{ts}^{k} = [\vartheta_{ts}, \varphi_{ts}, \frac{|\zeta_t - \zeta_s|}{S}]^T,$$

$$\vartheta_{ts} = \mathbf{1}_{[t,s|t \in g, s \in g]} [\frac{r_t - r_s}{R}, \frac{y_t - y_s}{Y}]^T,$$

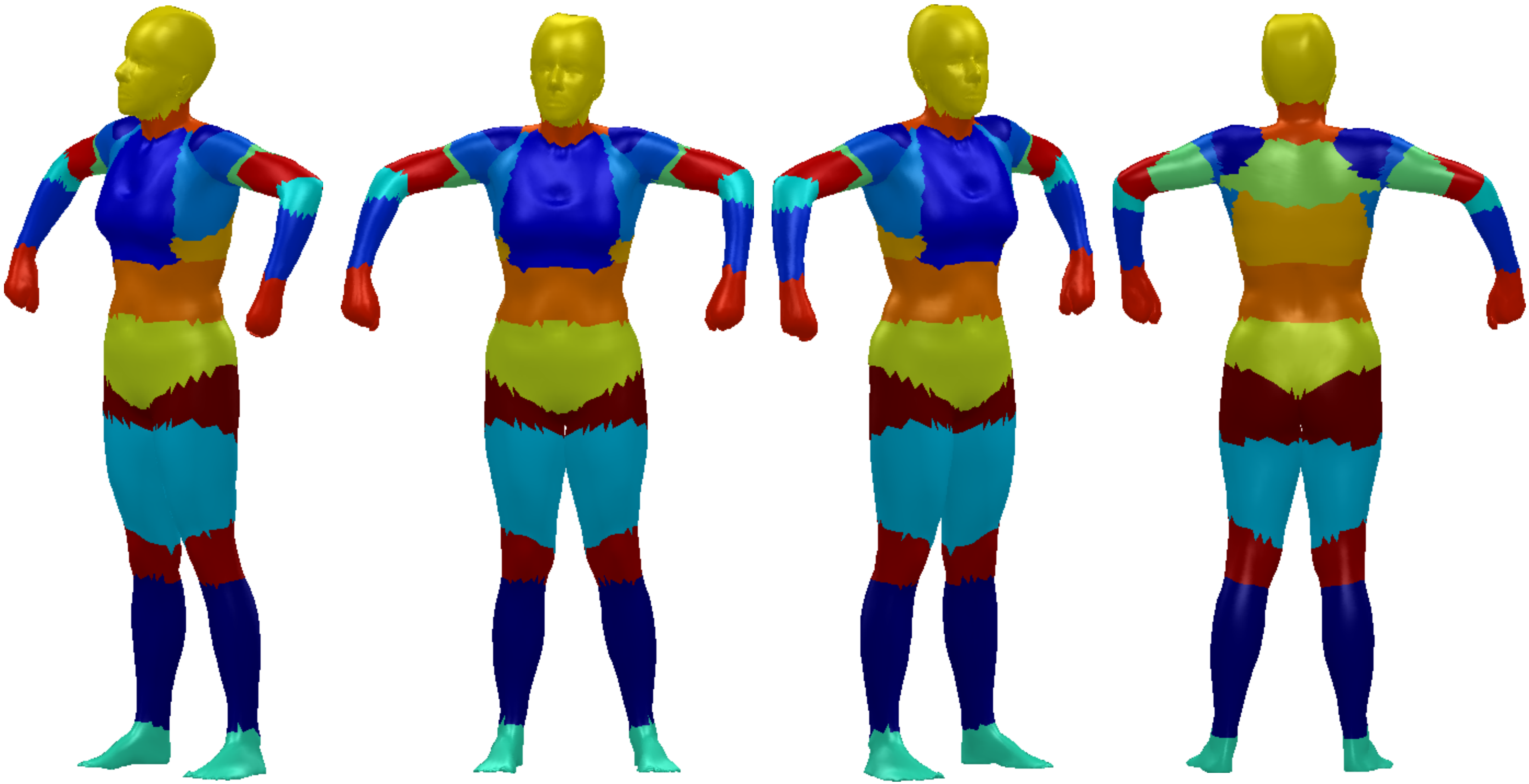
$$\varphi_{ts} = \mathbf{1}_{[t,s|t \in g+1, s \in g]} [\frac{|r_t - r_s|}{R}, \frac{|y_t - y_s|}{Y}, 1 - \frac{t \cap s}{t \cup s}]^T$$

MoCap Likelihoods

$$
\begin{aligned}
\Sigma_{z_{gi}} \mid n_0, S_0 &\sim \mathrm{IW}(n_0, S_0), \\
B_{z_{gi}} \mid M, \Sigma_{z_{gi}}, L &\sim \mathcal{MN}(M, \Sigma_{z_{gi}}, L), \\
\epsilon_{z_{gi}} &\sim \mathcal{N}(0, \Sigma_{z_{gi}}),
\end{aligned}
$$

Moderate robustness to alignment errors

Inferred Segmentation



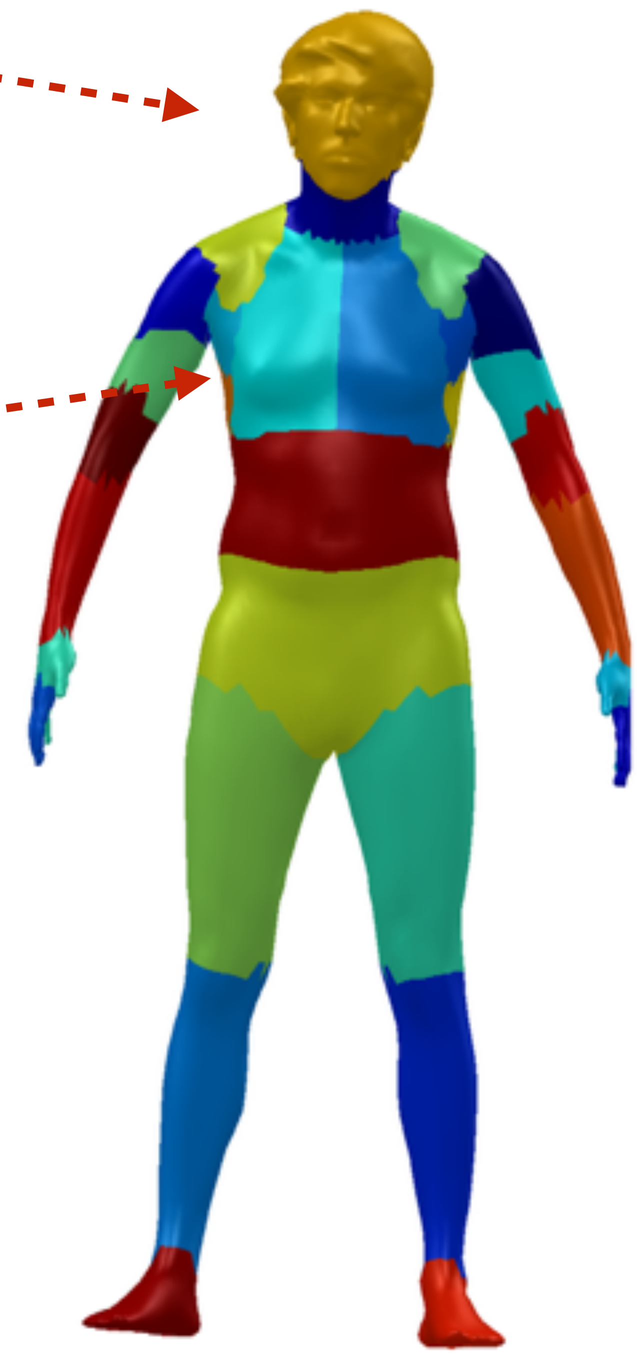
Segmentation with 20 Parts

Ghosh et al., NIPS 2012

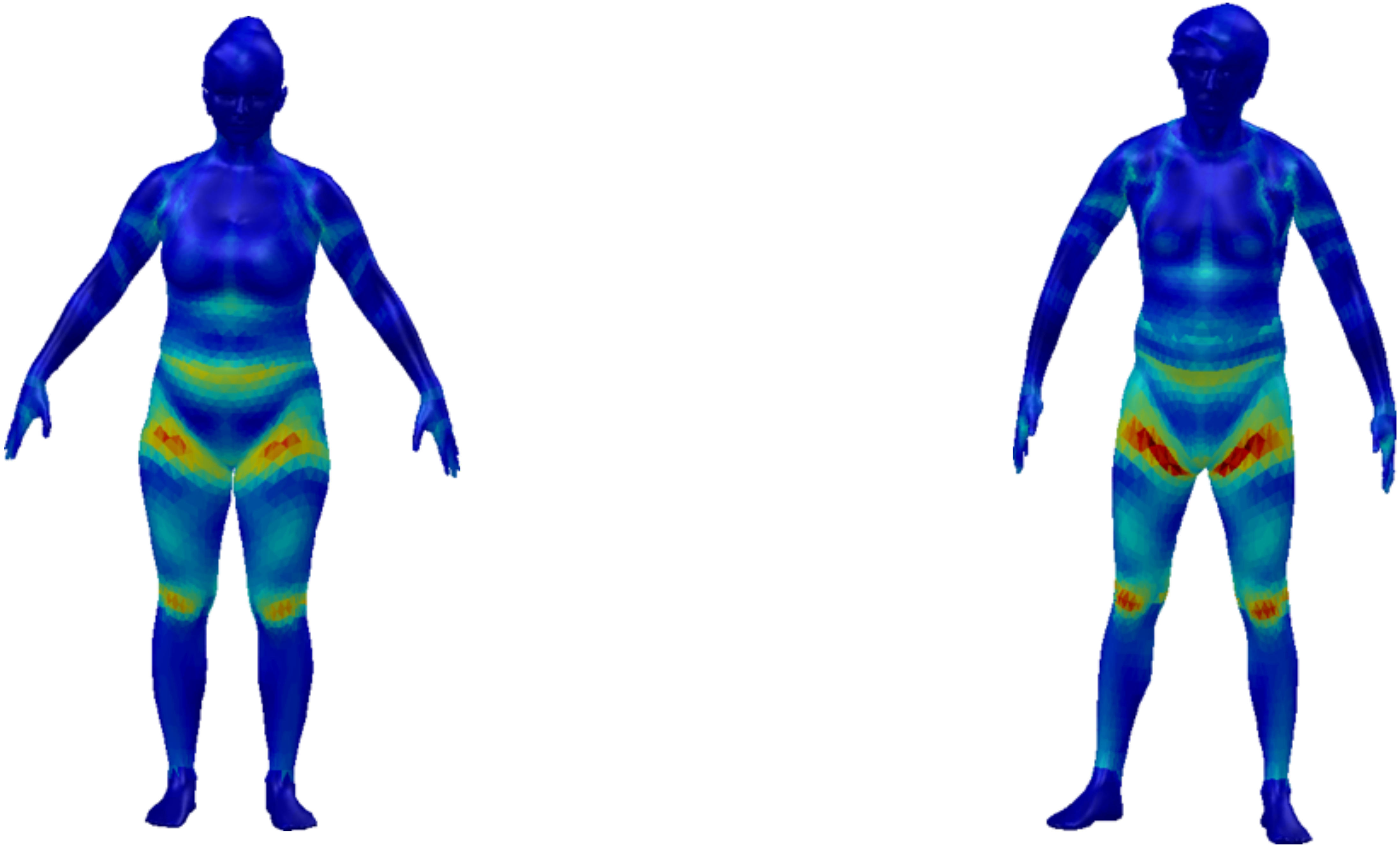
Axial Symmetry

$$\frac{p(Y_{left}^{head} \cup Y_{right}^{head} \mid X_{left}^{head} \cup X_{right}^{head})}{p(Y_{left}^{head} \mid X_{left}^{head})p(Y_{right}^{head} \mid X_{right}^{head})} > 1$$

$$\frac{p(Y_{left}^{chest} \cup Y_{right}^{head} \mid X_{left}^{chest} \cup X_{right}^{chest})}{p(Y_{left}^{chest} \mid X_{left}^{chest})p(Y_{chest}^{chest} \mid X_{right}^{chest})} < 1$$



Only merge similarly moving parts across axis of symmetry

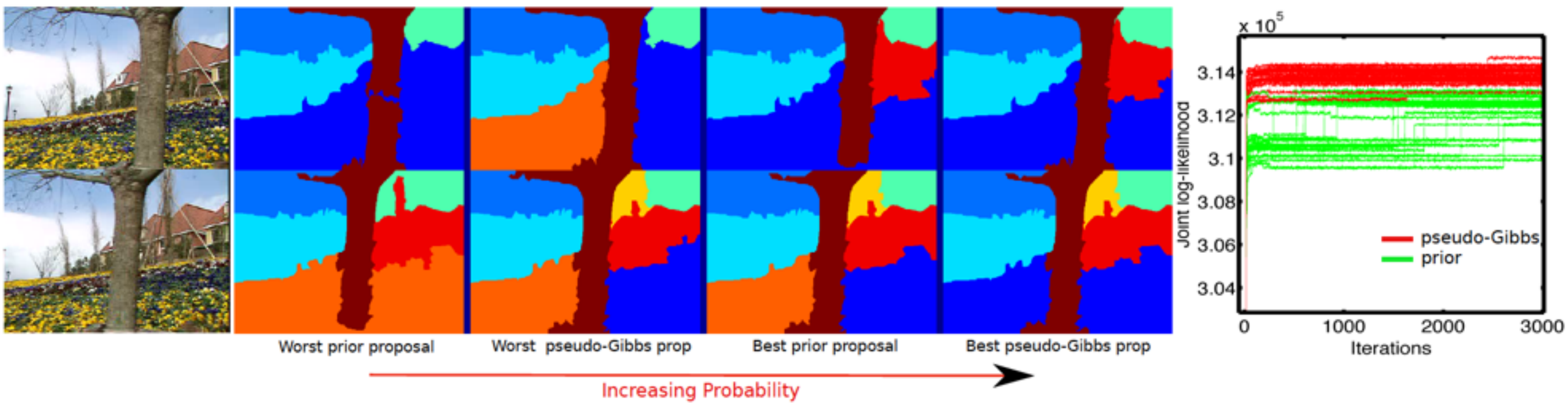


Measure of Rigidity

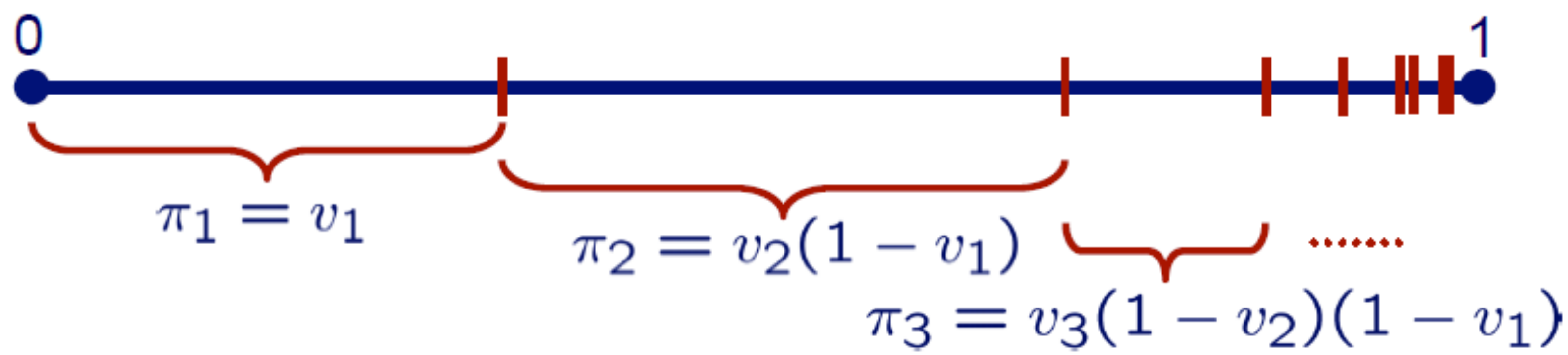
Inference

Algorithm 1: Hierarchical ddCRP sampler

For data instance $i \in \{1 \ldots N_G\}$ jointly propose data and affected cluster links $\{\mathbf{c}^*, \mathbf{k}^*\} \longleftarrow \text{ProposeLinks}(\mathbf{x}, \mathbf{k}, \mathbf{c}, \alpha_{1:G}, A^{1:G}, \alpha_0, A^0(\mathbf{c}))$.
Evaluate the proposal according to the Metropolis Hastings acceptance probability $a(\{\mathbf{c}^*, \mathbf{k}^*\}, \{\mathbf{c}, \mathbf{k}\})$. If the proposal is accepted, $\{\mathbf{c}^*, \mathbf{k}^*\}$ becomes the next state. If the proposal is rejected, the original configuration is retained.
For clusters $t \in T(\mathbf{c})$ resample cluster links via a Gibbs update: $k_t \sim p(k_t \mid \mathbf{k}_{-t}, \mathbf{c}, \mathbf{x}, \alpha_0, A^0(\mathbf{c}))$.


Worst prior proposal
Worst pseudo-Gibbs prop
Best prior proposal
Best pseudo-Gibbs prop
Increasing Probability
x 10^5
3.14
3.12
3.1
3.08
3.06
3.04
Joint log-likelihood
pseudo-Gibbs
prior
0
1000
2000
3000
Iterations

Stick Breaking to Layers



$$v_k = \mathbb{P}(z_i = k \mid z_i \neq k-1, \ldots, 1)$$

Sequential Binary Sampler:

$$b_{ki} \sim \text{Bernoulli}(v_k)$$

$$z_i = \min\{k \mid b_{ki} = 1\}$$

- *For each data instance i, go through the bins in order 1 through infinity.*
- *Toss a biased coin (with the probability of heads = v_k) for each bin .*
- *Pick the bin if the coin turns up heads*

MCMC Learning

- Marginalize over the exponentially large space of latent links
  - MCMC samples
- Explore the marginal posterior of the auxiliary training model:

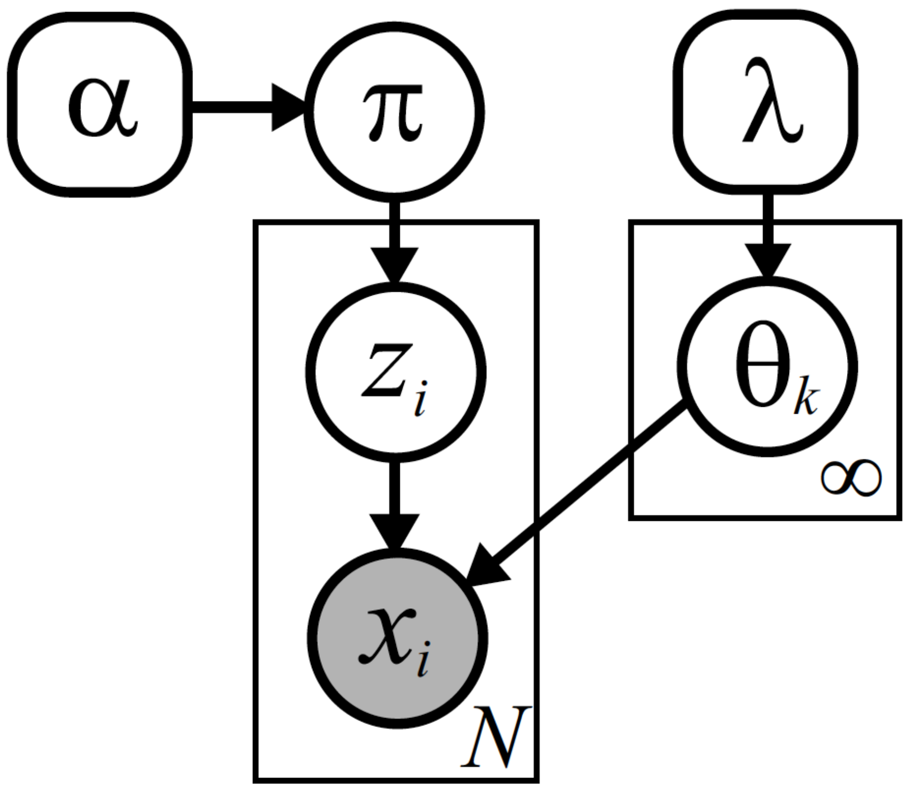
$$p(w_c \mid Y) = \sum_{\mathbf{c}} p(w_c, \mathbf{c} \mid Y) \approx \sum_{\mathbf{c^{(s')}}} p(w_c^{(s)}, \mathbf{c}^{(s')} \mid Y)$$

$$w_c^s, \mathbf{c}^s \sim p(w_c, \mathbf{c} \mid Y)$$

Random walk Proposal: $w_c^{t+1} \sim \mathcal{N}(w_c^{t+1} \mid w_c^t, \mathrm{scale} \times \mathbf{I})$

Gibbs Step: $c_{di} \mid \mathbf{c}_{-di}, w_c^*, Y \sim p(c_{di} \mid w_c^*)\delta(z(\mathbf{c_d}), y_d)$

Bayesian Nonparametric Priors



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

$$\sum_{k=1}^{\infty} \pi_k = 1 \qquad 0 \leq \pi_k \leq 1$$

$$\theta_k \sim \mathrm{H}(\lambda)$$

Pitman-Yor Process

Power Law Behavior

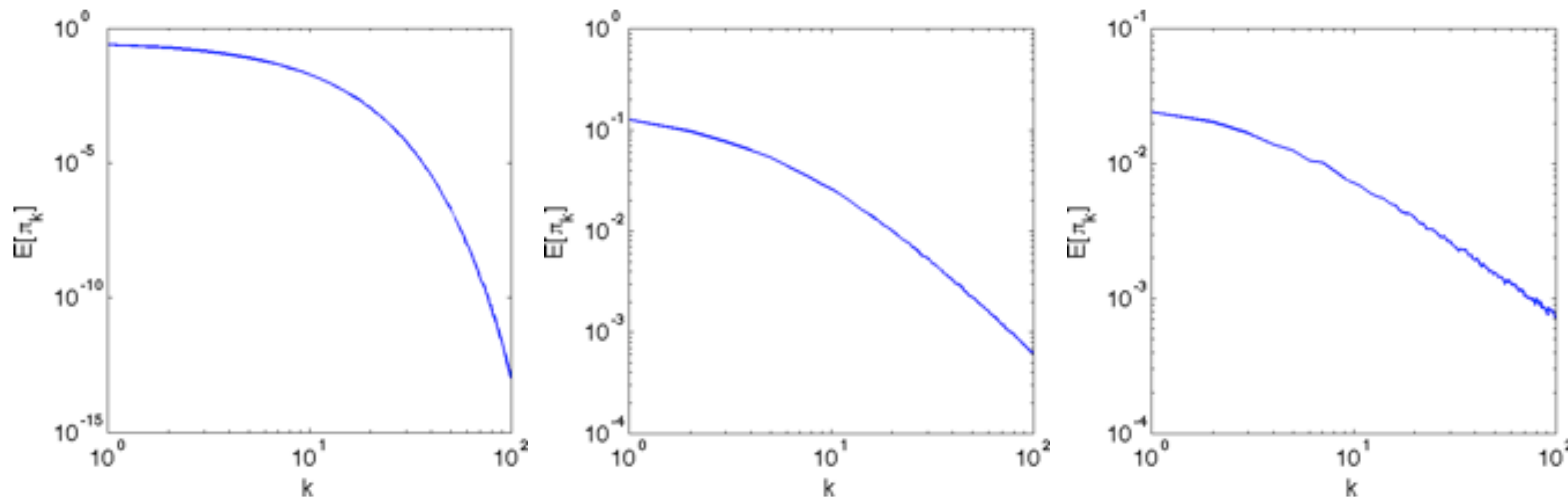
$$E[w_k] = \frac{1-\alpha_a}{(1+\alpha_b+(k-1)\alpha_a)}$$

$$\pi_k = w_k \prod_{l=1}^{k-1}(1-w_l)$$

$$w_k \sim \text{Beta}(a_k, b_k)$$

$$a_k = 1-\alpha_a$$

$$b_k = \alpha_b + k\alpha_a$$



$\alpha_a = 0, \alpha_b = 3$ $\alpha_a = 0.5, \alpha_b = 3$ $\alpha_a = 0.9, \alpha_b = 3$

Number of unique clusters in N observations: $O(\alpha_b N^{\alpha_a})$

Expected size of sorted component k: $O(k^{-\frac{1}{\alpha_a}})$

Hierarchical ddCRP

Sample local links:

$$p(c_{gi} = gj \mid \alpha_g, A^g) \propto \begin{cases} A^g_{ij} & i \neq j, \\ \alpha_g & i = j. \end{cases}$$

$$\Lambda_g = z(\mathbf{c}_g)$$

Sample global links:

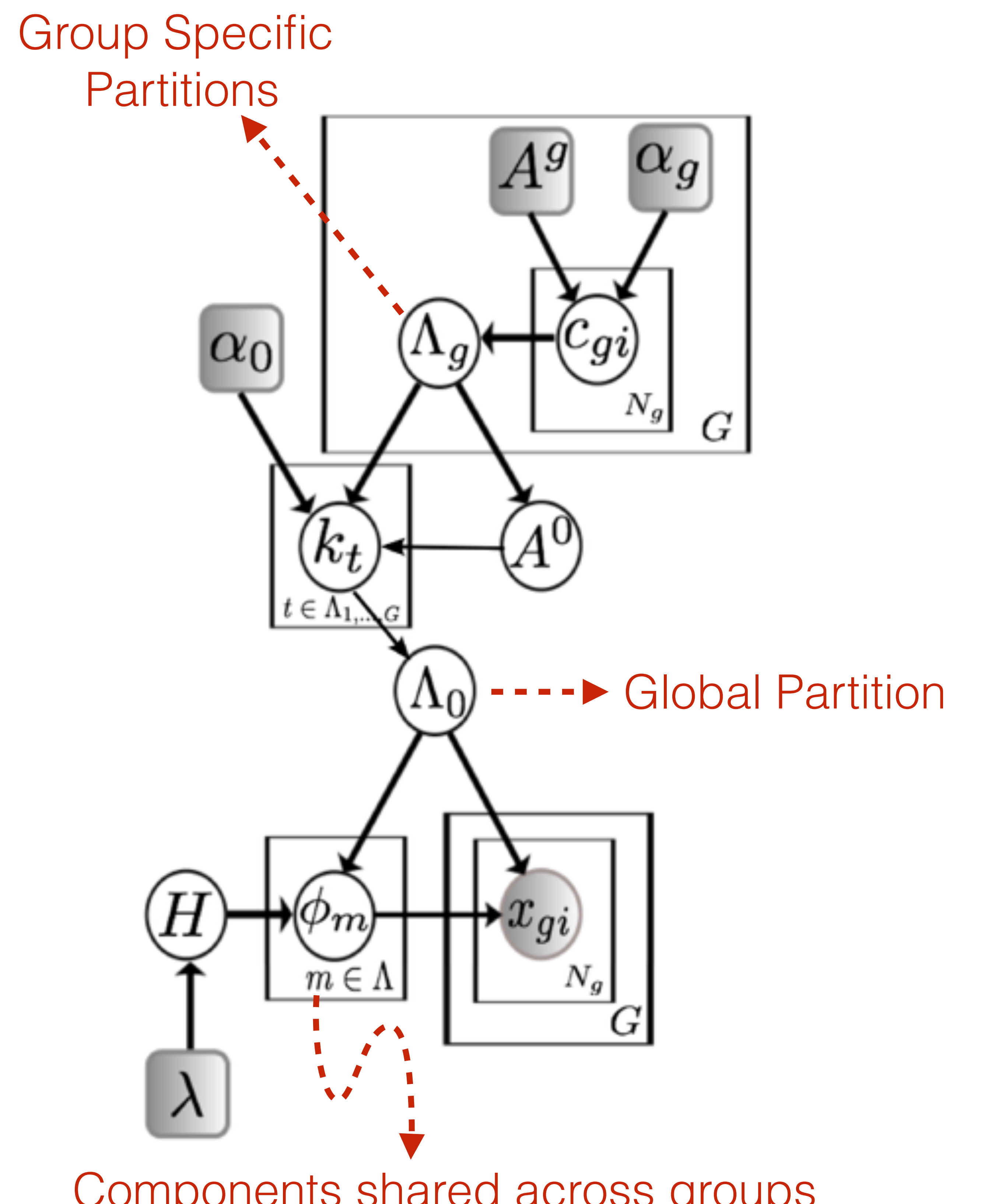
$$p(k_t = s \mid \alpha_0, A^0(\mathbf{c})) \propto \begin{cases} A^0_{ts}(\mathbf{c}) & t \neq s, \\ \alpha_0 & t = s. \end{cases}$$

$$\Lambda_0 = z(\mathbf{k})$$

Sample data generating parameters:

$$\phi_m \sim H(\lambda), \forall m \in \Lambda_0$$

$$x_i \sim \phi_m, \forall i \in m$$

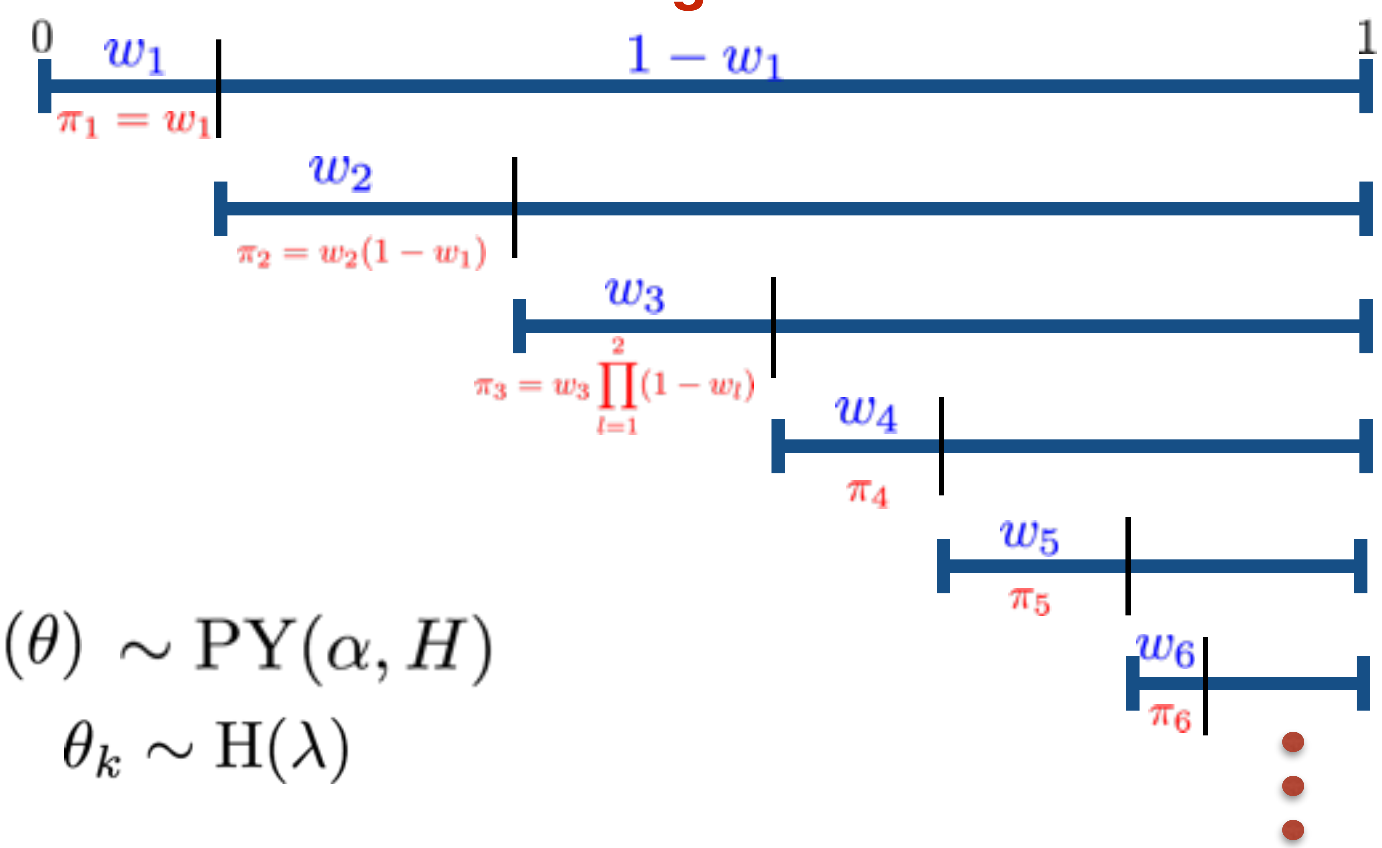


Pitman-Yor Process

- The Pitman-Yor process defines a distribution on infinite discrete measures, or partitions

$$\pi_k = w_k \prod_{l=1}^{k-1}(1 - w_l) \qquad w_k \sim \mathrm{Beta}(1 - \alpha_a, \alpha_b + k\alpha_a)$$

Stick Breaking Construction:



$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \sim \mathrm{PY}(\alpha, H)$$

$$\theta_k \sim \mathrm{H}(\lambda)$$

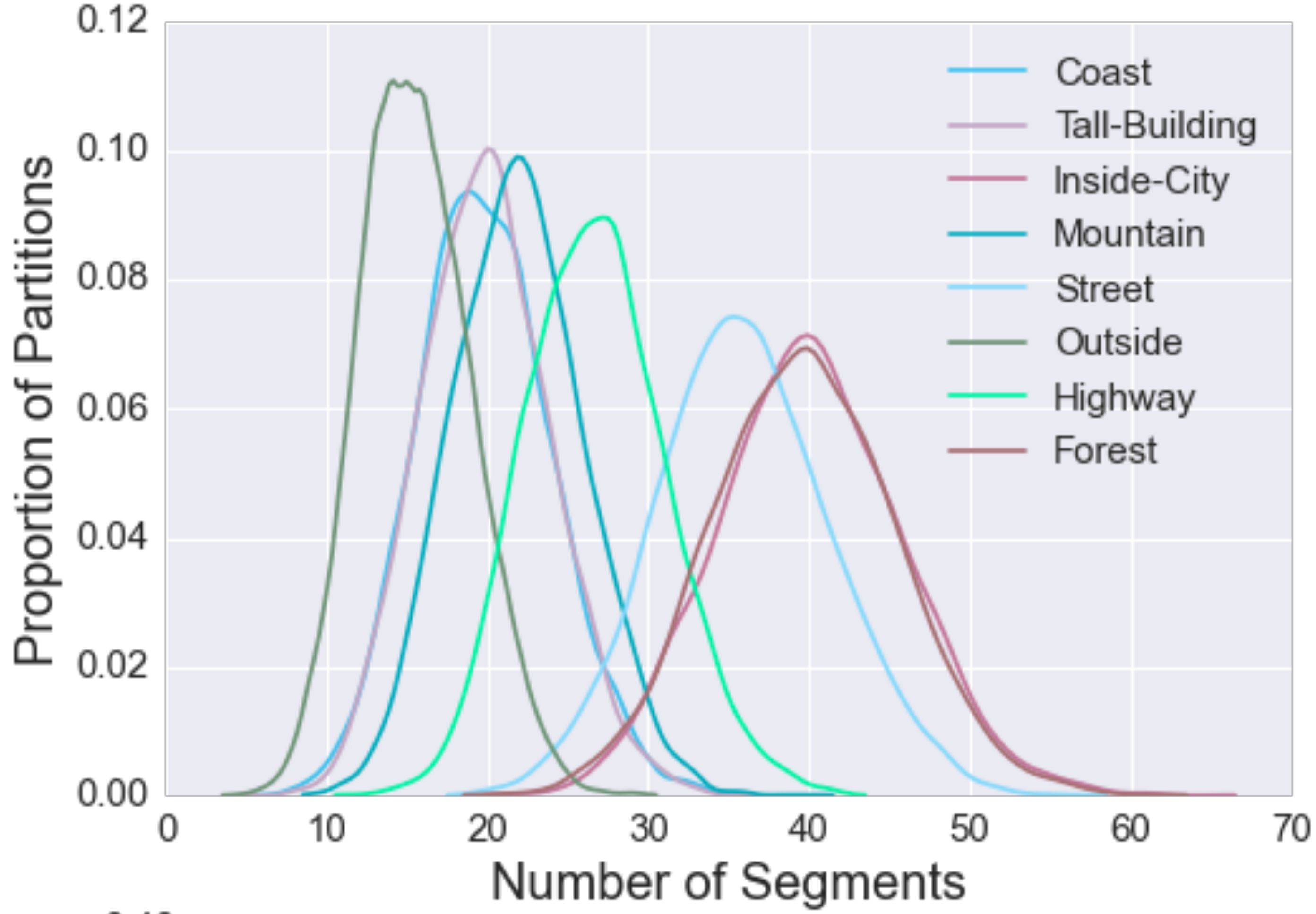
Sethuraman, 1994
Ishwaran and James, 2001

Video Segmentation

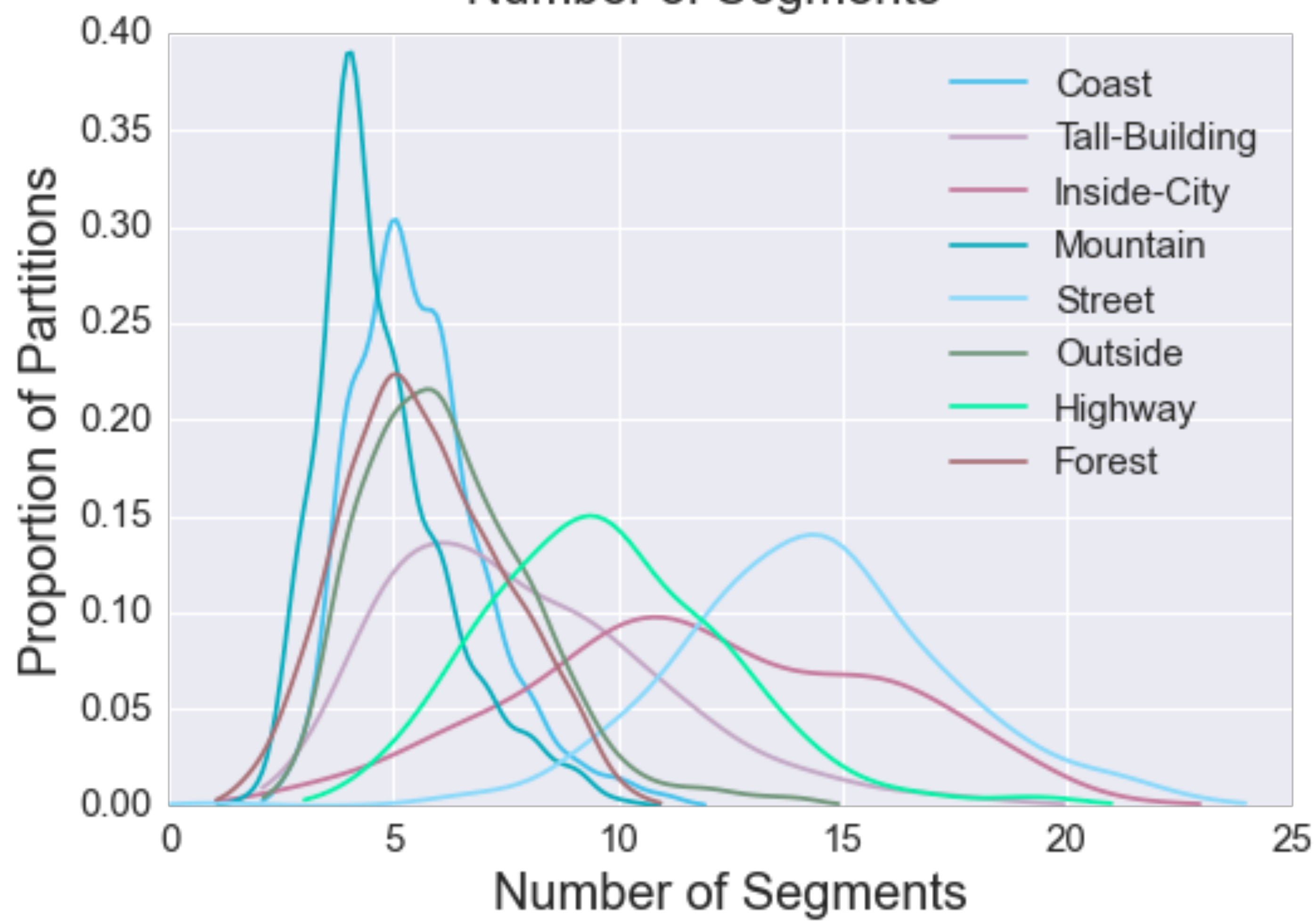
$$P = \frac{\sum_{i=1}^{M}\Big[\Big\{\sum_{s\in\mathbb{S}} \max_{g\in\mathbb{G}_i} |s\cap g|\Big\} - \max_{g\in\mathbb{G}_i} |g|\Big]}{M|\mathbb{S}| - \sum_{i=1}^{M} \max_{g\in\mathbb{G}_i} |g|}$$

$$R = \frac{\sum_{i=1}^{M} \sum_{g\in\mathbb{G}_i} \{\max_{s\in\mathbb{S}} |s\cap g| - 1\}}{\sum_{i=1}^{M} \{|\mathbb{G}_i| - \Gamma_{\mathbb{G}_i}\}}$$

VPR



0.12
0.10
0.08
0.06
0.04
0.02
0.00
Proportion of Partitions
0
10
20
30
40
50
60
70
Number of Segments
Coast
Tall-Building
Inside-City
Mountain
Street
Outside
Highway
Forest



0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Proportion of Partitions
0
5
10
15
20
25
Number of Segments
Coast
Tall-Building
Inside-City
Mountain
Street
Outside
Highway
Forest

Approximate Bayesian Computation

Algorithm 3 Likelihood-free MCMC sampler

Use Algorithm 2 to get a realisation $(\boldsymbol{\theta}^{(0)}, \mathbf{z}^{(0)})$ from the ABC target distribution $\pi_\varepsilon(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$

for $t = 1$ to N **do**

Generate $\boldsymbol{\theta}'$ from the Markov kernel $q(\cdot|\boldsymbol{\theta}^{(t-1)})$,

Generate $\mathbf{z}'$ from the likelihood $f(\cdot|\boldsymbol{\theta}')$,

Generate u from $\mathcal{U}_{[0,1]}$,

if $u \leq \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta}^{(t-1)})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(t-1)})}$ and $\rho\{\eta(\mathbf{z}'), \eta(\mathbf{y})\} \leq \varepsilon$ **then**

set $(\boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)}) = (\boldsymbol{\theta}', \mathbf{z}')$

else

$(\boldsymbol{\theta}^{(t)}, \mathbf{z}^{(t)}) = (\boldsymbol{\theta}^{(t-1)}, \mathbf{z}^{(t-1)})$,

end if

end for