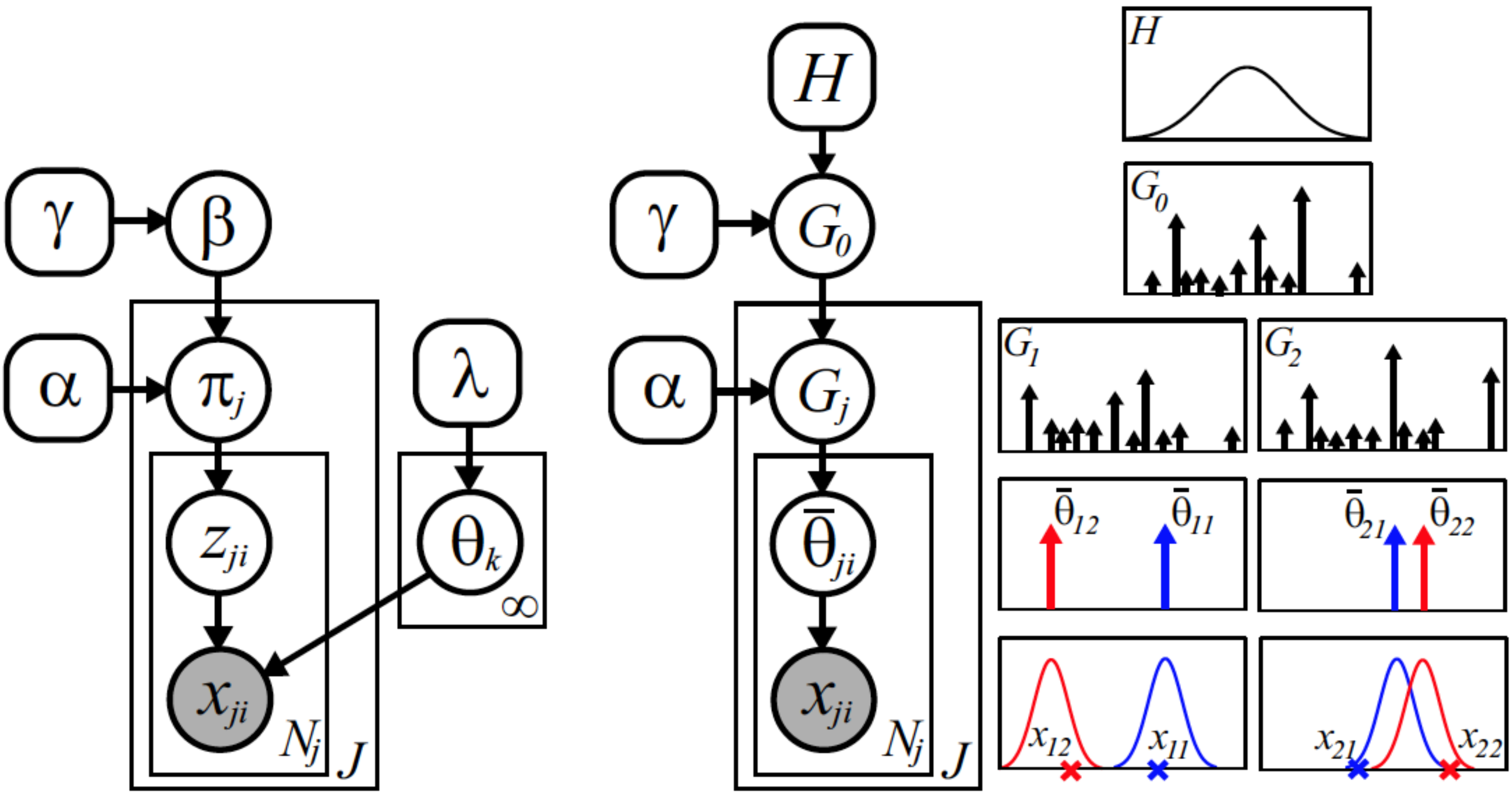
Applied Bayesian Nonparametrics

Special Topics in Machine Learning
Brown University CSCI 2950-P, Fall 2011

October 6: Hierarchical, Nested, and Transformed Dirichlet Processes

Hierarchical Dirichlet Process



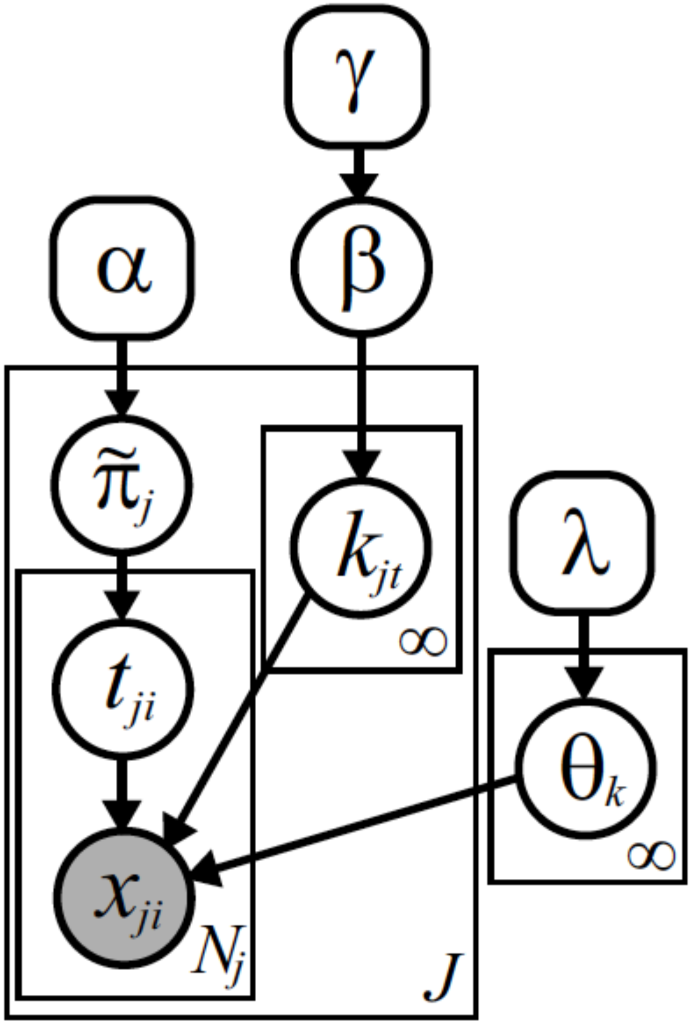
Hierarchical Dirichlet Process

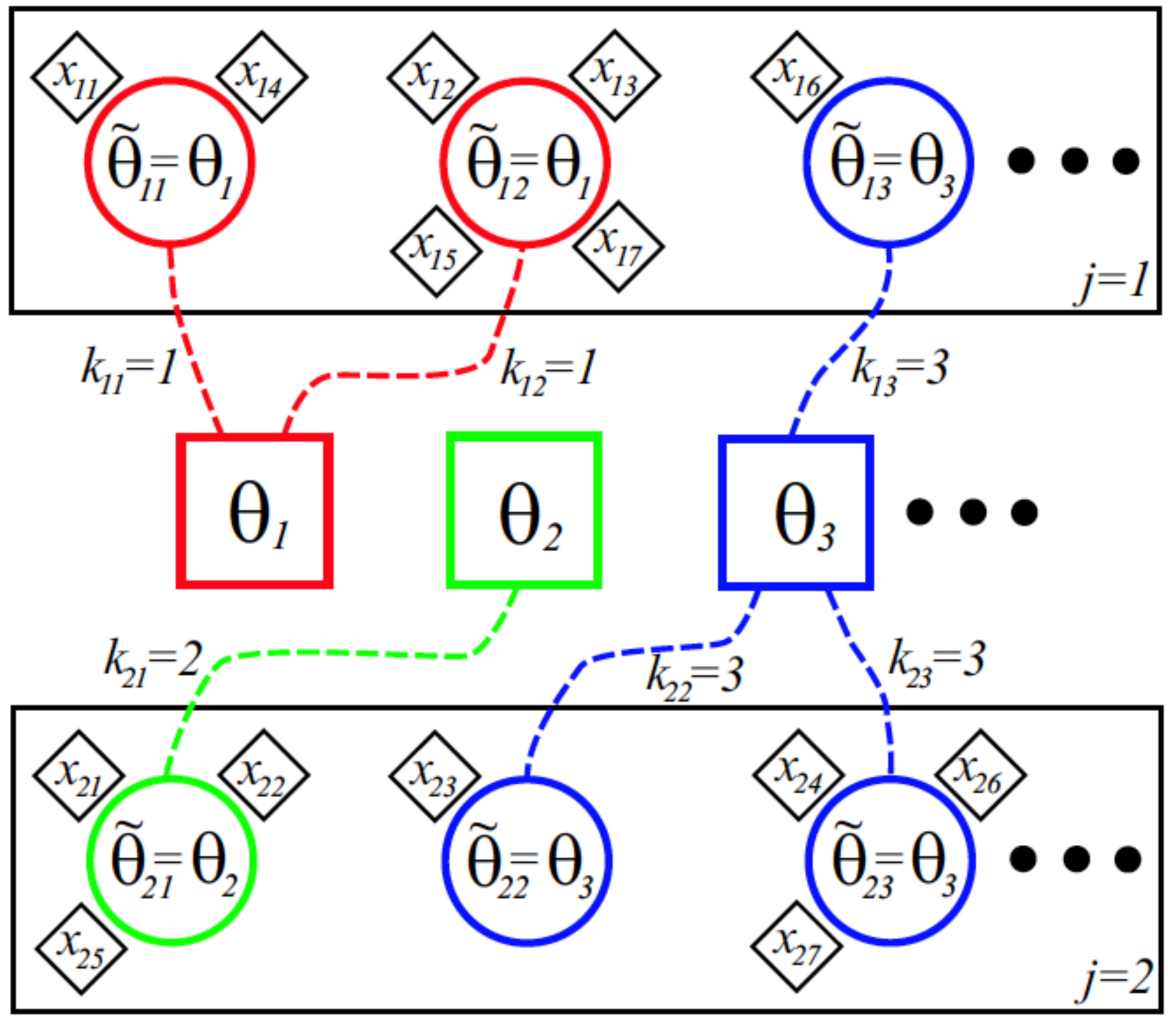
$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k) \qquad \begin{aligned} \boldsymbol{\beta} &\sim \mathrm{GEM}(\gamma) \\ \theta_k &\sim H(\lambda) \qquad k = 1, 2, \ldots \end{aligned}$$

$$G_j(\theta) = \sum_{t=1}^{\infty} \widetilde{\pi}_{jt} \delta(\theta, \widetilde{\theta}_{jt}) \qquad \begin{aligned} \widetilde{\boldsymbol{\pi}}_j &\sim \mathrm{GEM}(\alpha) \\ \widetilde{\theta}_{jt} &\sim G_0 \qquad t = 1, 2, \ldots \end{aligned}$$

$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k) \qquad \pi_{jk} = \sum_{t \mid k_{jt} = k} \widetilde{\pi}_{jt}$$

Chinese Restaurant Franchise

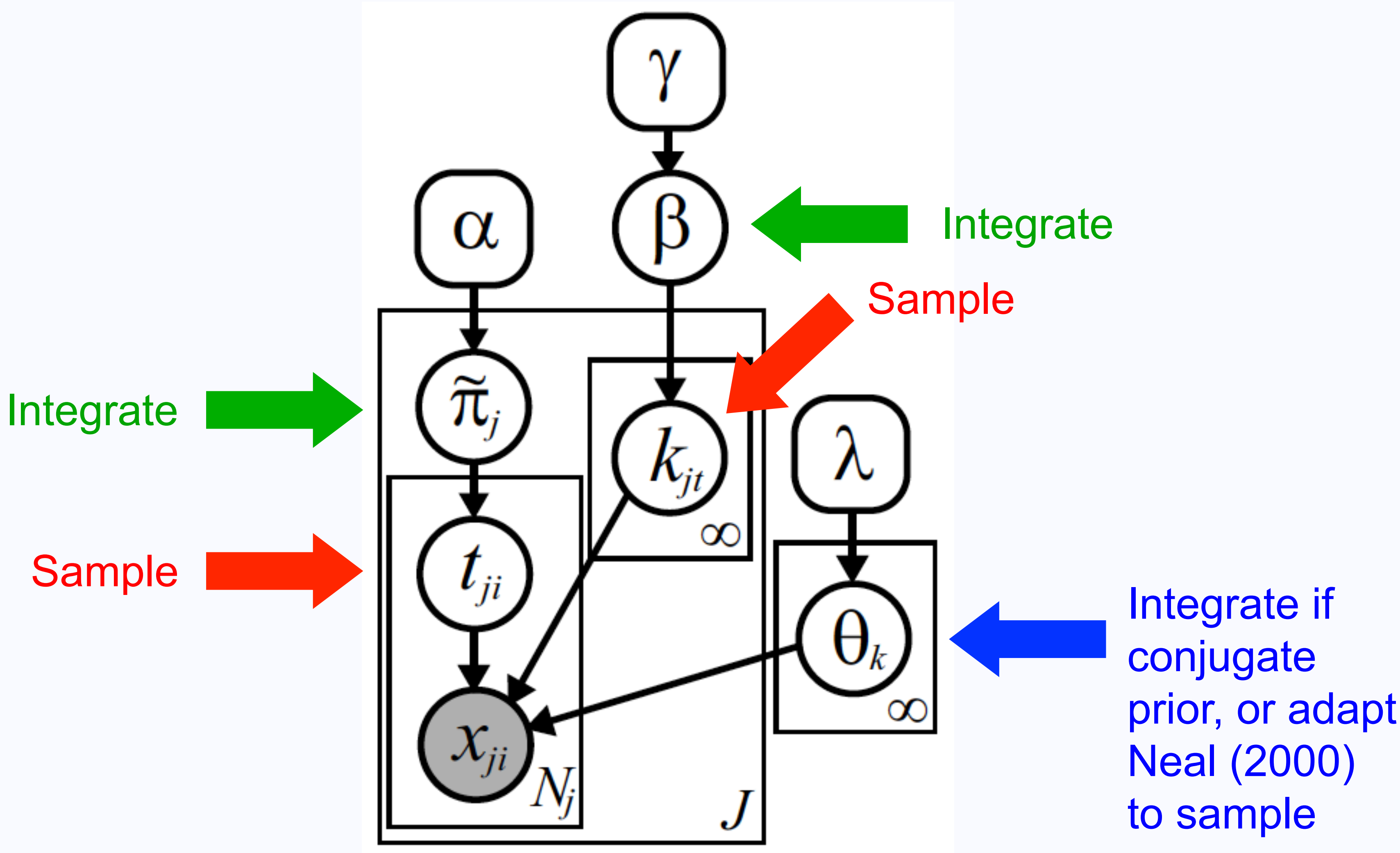




$$p(t_{ji} \mid t_{j1}, \ldots, t_{ji-1}, \alpha) \propto \sum_t N_{jt}\delta(t_{ji}, t) + \alpha\delta(t_{ji}, \bar{t})$$

$$p(k_{jt} \mid \mathbf{k}_1, \ldots, \mathbf{k}_{j-1}, k_{j1}, \ldots, k_{jt-1}, \gamma) \propto \sum_k M_k\delta(k_{jt}, k) + \gamma\delta(k_{jt}, \bar{k})$$

HDP CRF Gibbs Sampler



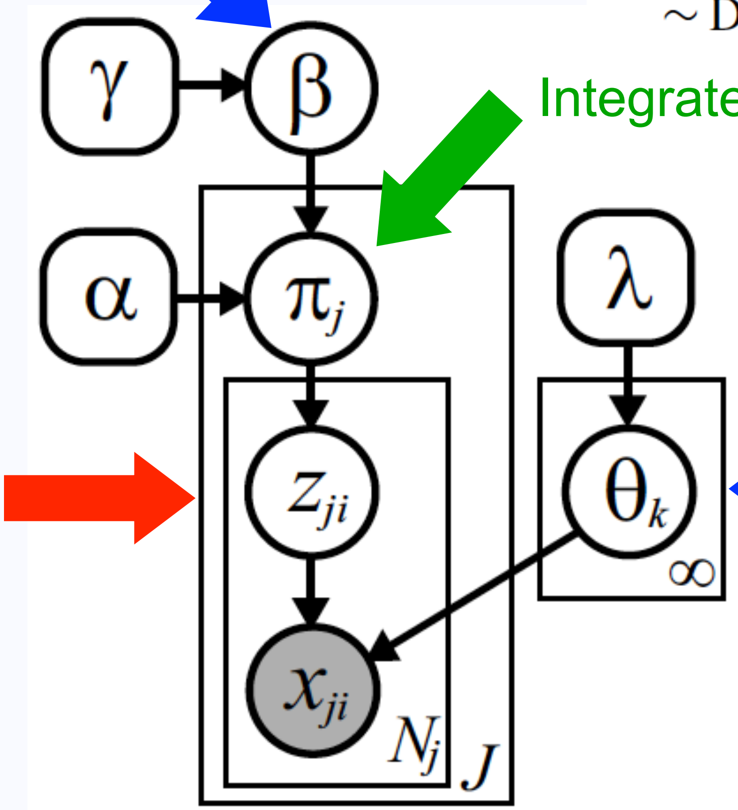
No finite truncation required…

HDP Direct Assignment Sampler

Sample using
auxiliary variable
trick involving
CRF tables

$$\left(\sum_{k\in K_1}\pi_{jk},\ldots,\sum_{k\in K_r}\pi_{jk}\right) \sim \mathrm{Dir}\left(\alpha_0\sum_{k\in K_1}\beta_k,\ldots,\alpha_0\sum_{k\in K_r}\beta_k\right)$$

Integrate

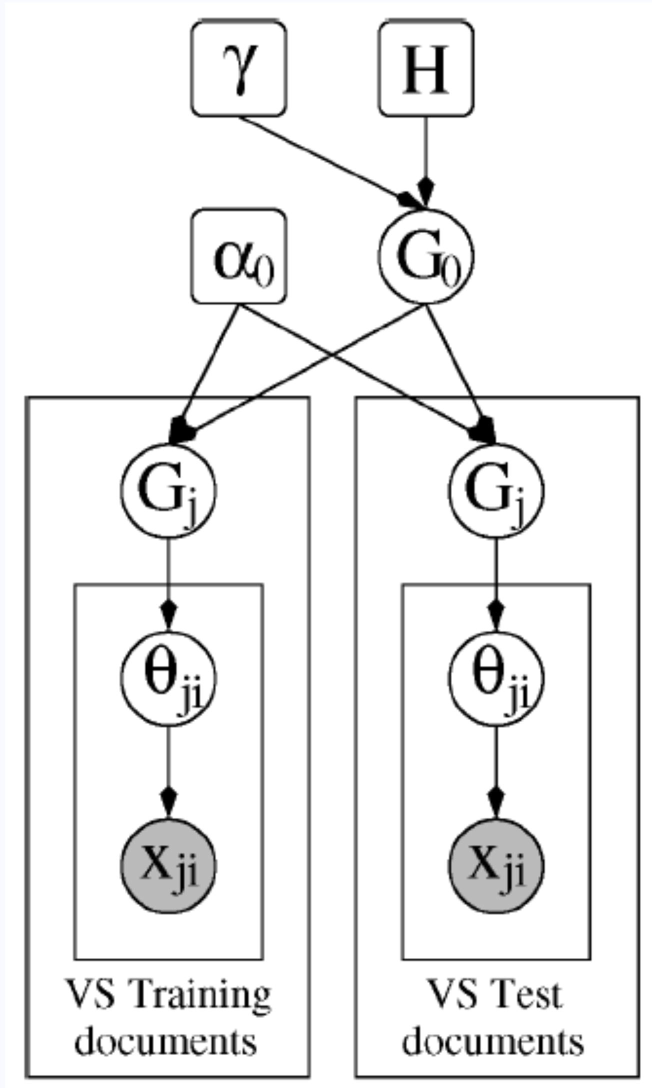


Sample via
weighted CRP

Integrate if
conjugate
prior, or adapt
Neal (2000)
to sample

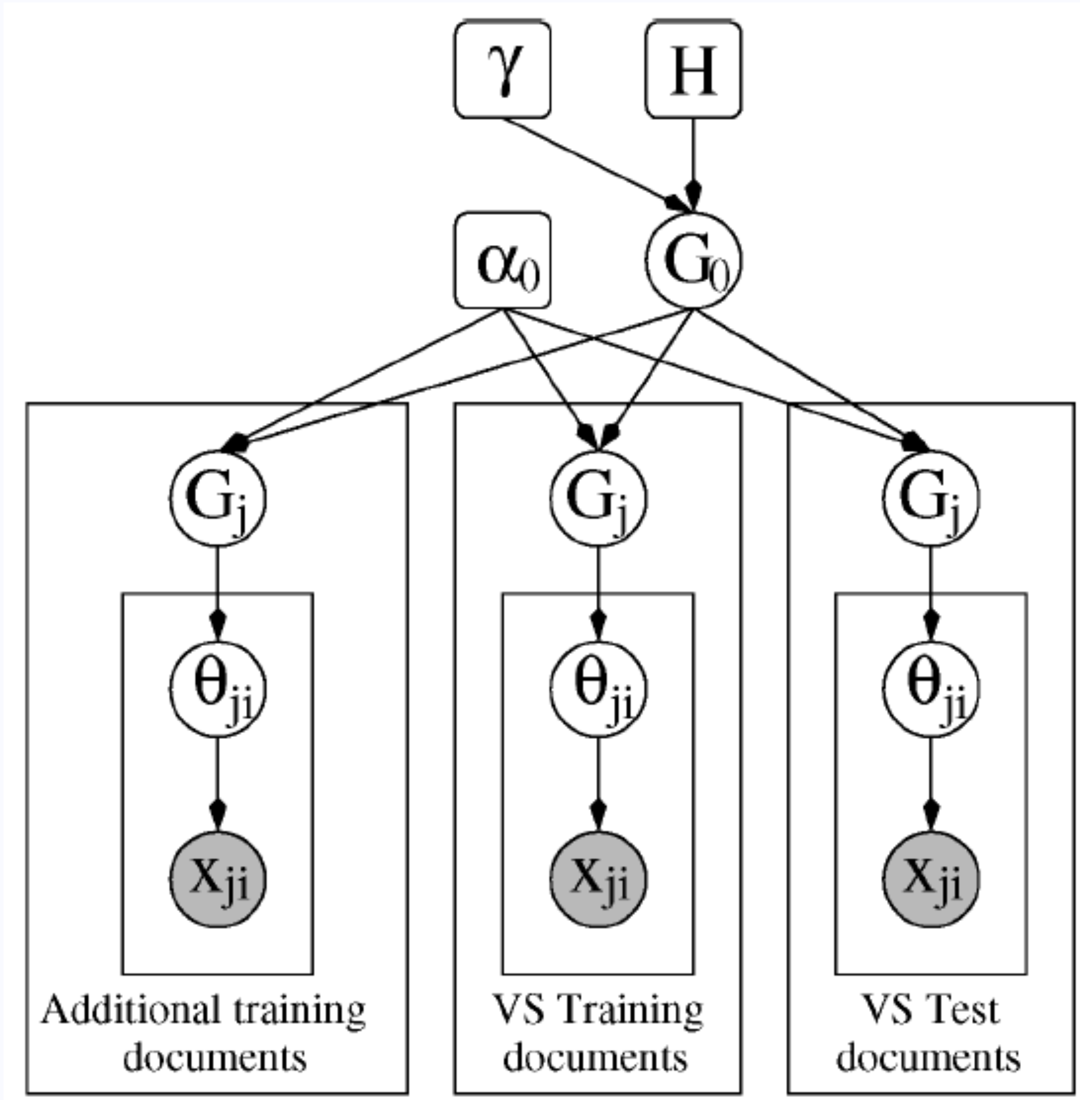
No finite truncation required…

Categorized Documents



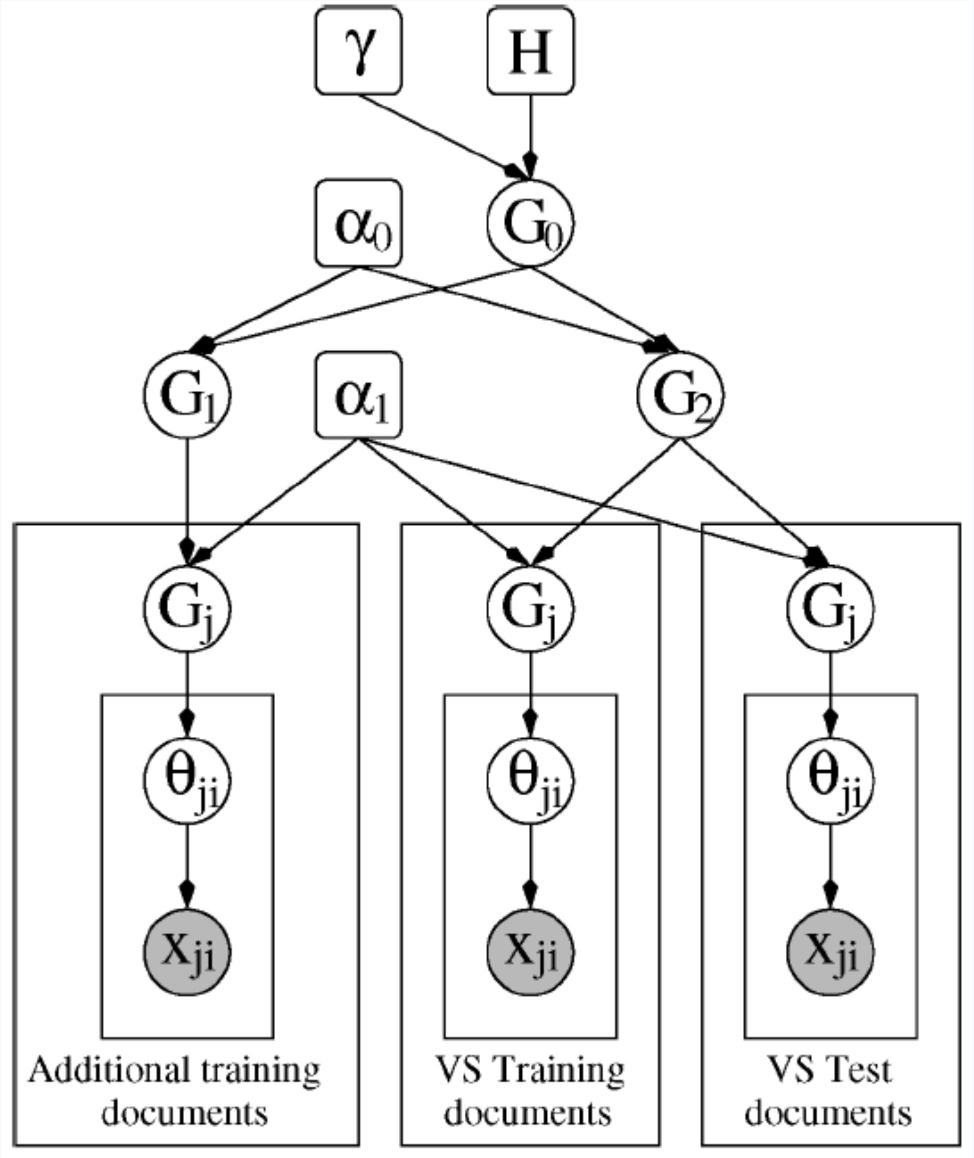
M1: Each category is treated independently

Categorized Documents



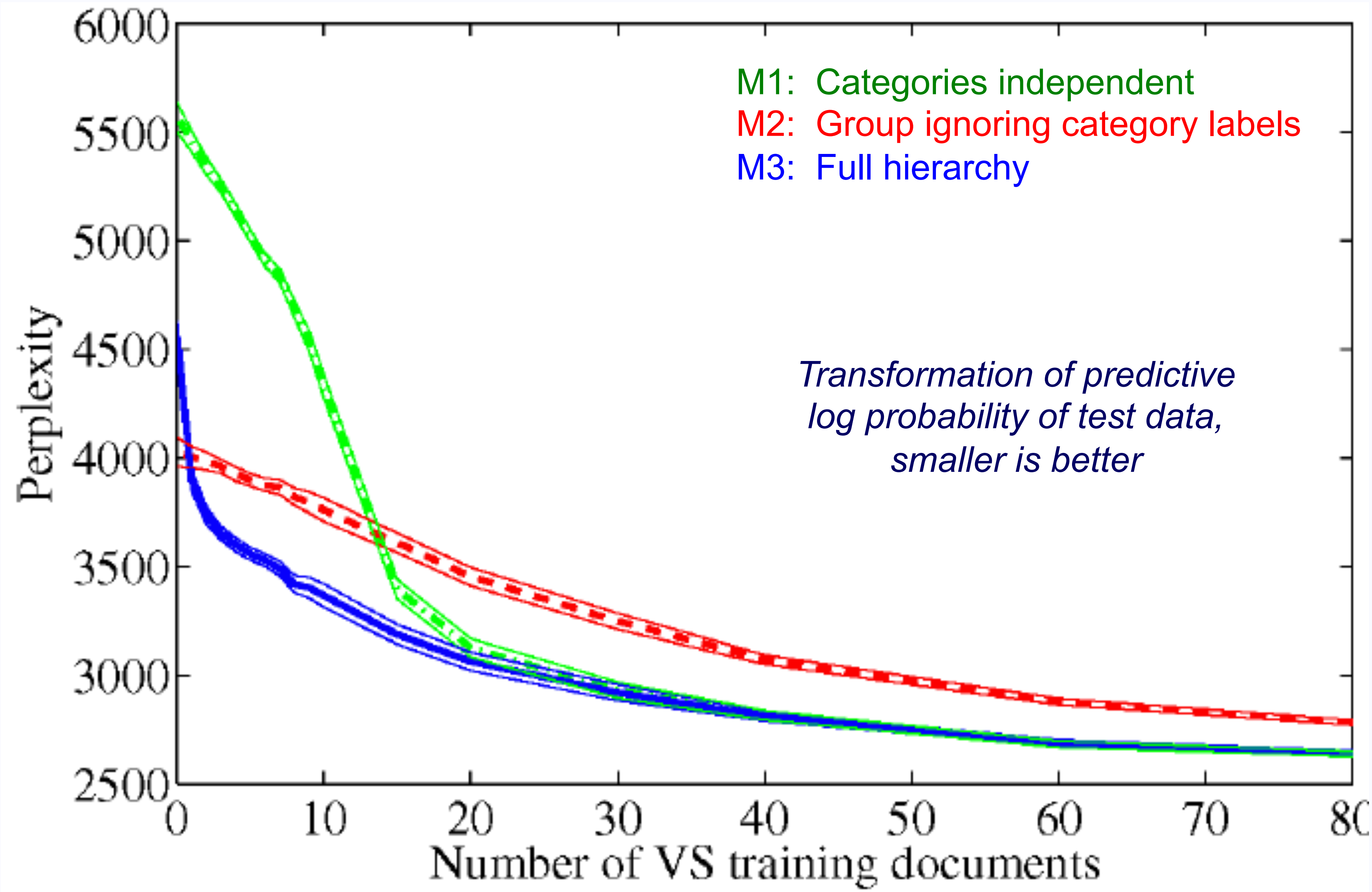
M2: Ignore category labels, treat as one large dataset

Categorized Documents



M3: Fully hierarchy, documents more similar within than between categories

Categorized Documents



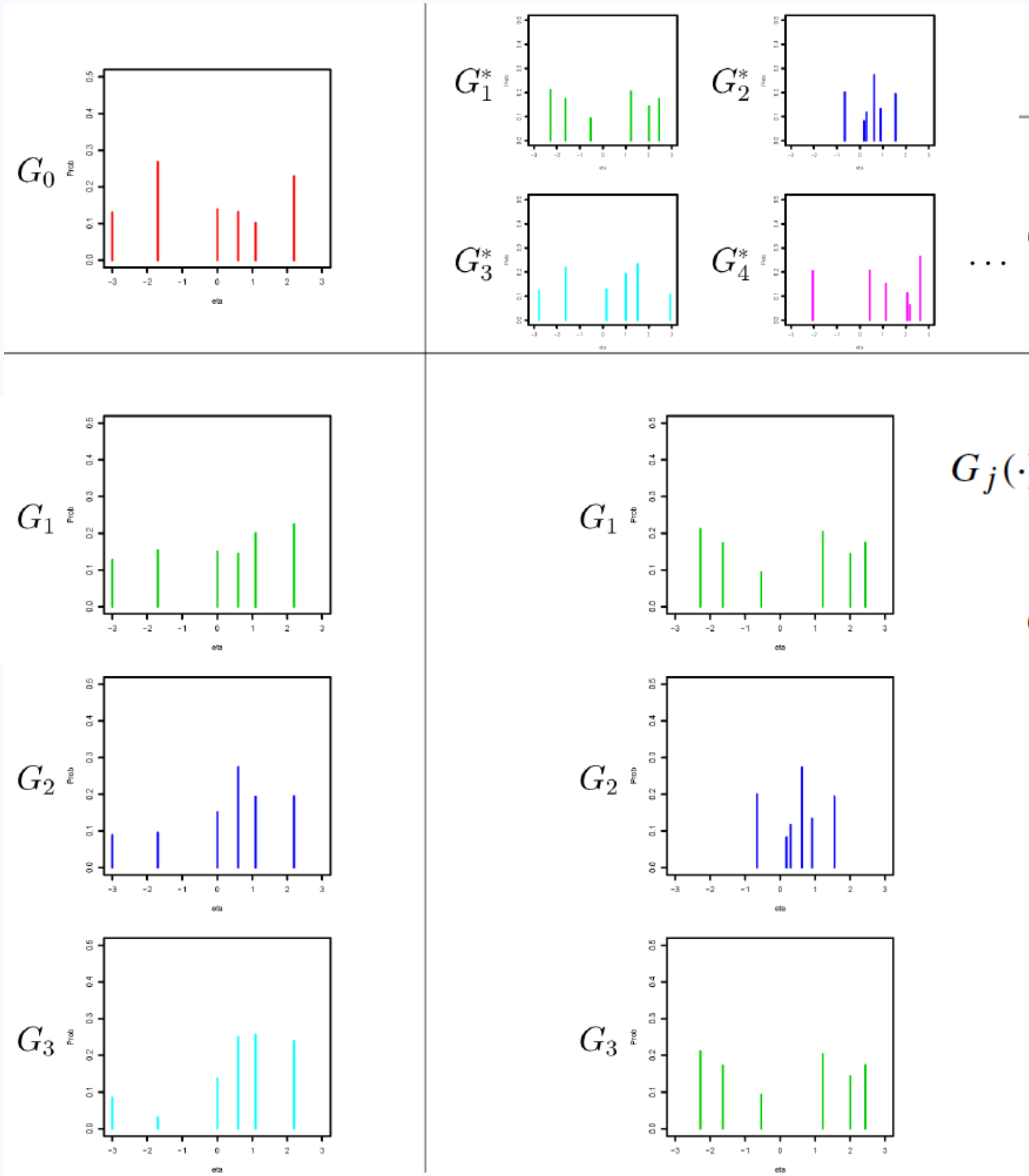
Hierarchical DP vs. Nested DP

HDP

$$G_j \sim \mathrm{DP}(\alpha G_0)$$
$$G_0 \sim \mathrm{DP}(\beta H)$$

$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k)$$

$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k)$$



NDP

$$G_j \sim Q$$
$$Q \sim \mathrm{DP}(\alpha \mathrm{DP}(\beta H))$$

$$G_j(\cdot) \sim Q \equiv \sum_{k=1}^{\infty} \pi_k^* \delta_{G_k^*(\cdot)}$$

$$G_k^*(\cdot) \equiv \sum_{l=1}^{\infty} w_{lk}^* \delta_{\theta_{lk}^*}(\cdot)$$

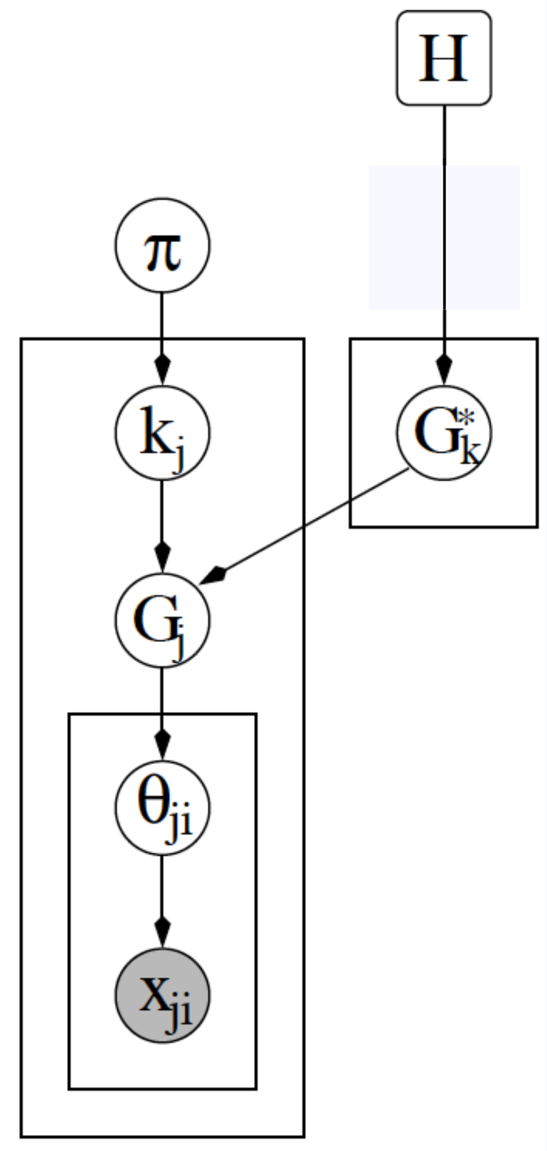
$$\boldsymbol{\theta}_{lk}^* \sim H$$

The NDP: Simpler than it seems

1. Partition your data groups according to a Chinese restaurant process with hyperparameter α
2. For each cluster in this partition, independently sample an "infinite" mixture model from a Dirichlet process prior with hyperparameter β
3. Treat these clusters as new "super-groups", generate the data i.i.d. from the corresponding DP mixture (independently of other clusters)

Gives a simple correlation structure:

$$\mathrm{cor}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) = \begin{cases} \dfrac{1}{(1+\beta)}, & j = j' \\ \dfrac{1}{(1+\alpha)(1+\beta)}, & j \neq j' \end{cases}$$



Graph by Teh, 2007

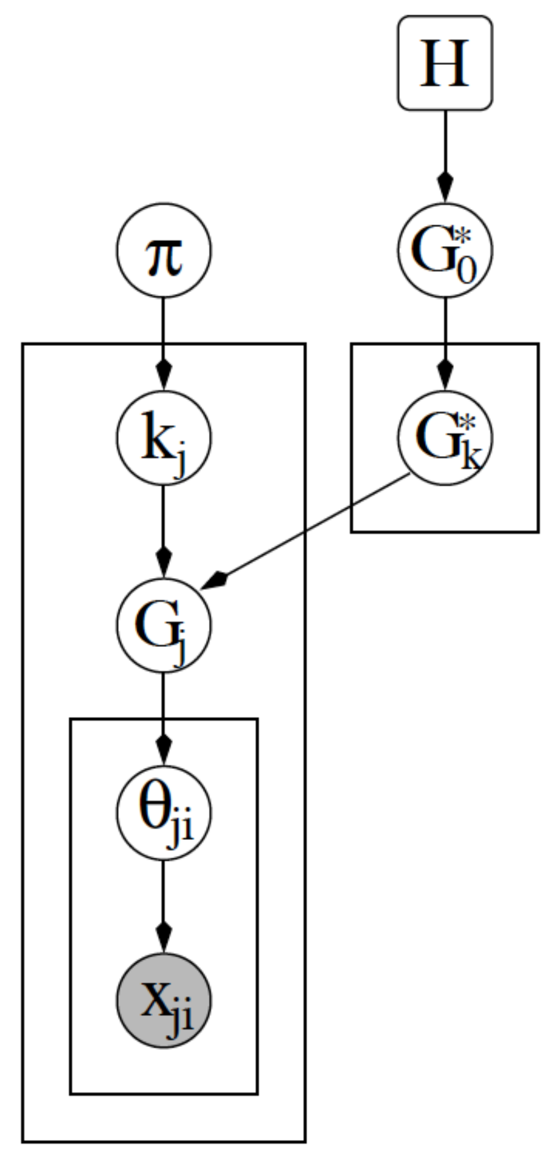
The NDP: Simpler than it seems

1. Partition your data groups according to a Chinese restaurant process with hyperparameter α
2. For each cluster in this partition, independently sample an "infinite" mixture model from a Dirichlet process prior with hyperparameter β
3. Treat these clusters as new "super-groups", generate the data i.i.d. from the corresponding DP mixture (independently of other clusters)

Gives a simple correlation structure:

$$\mathrm{cor}(\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_{i'j'}) = \begin{cases} \dfrac{1}{(1+\beta)}, & j = j' \\ \dfrac{1}{(1+\alpha)(1+\beta)}, & j \neq j' \end{cases}$$

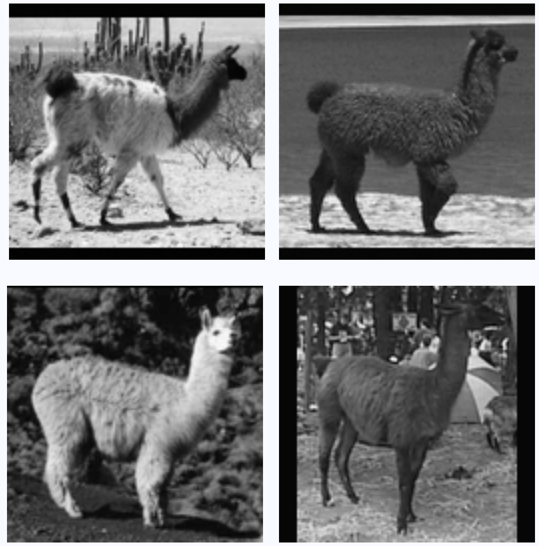
Hybrid of HDP and NDP allows sharing of parameters among the nested DP's clusters (not directly considered in Rodriguez JASA 2008, but mentioned in comments)



Graph by Teh, 2007

Visual Object Categorization

Bicycles



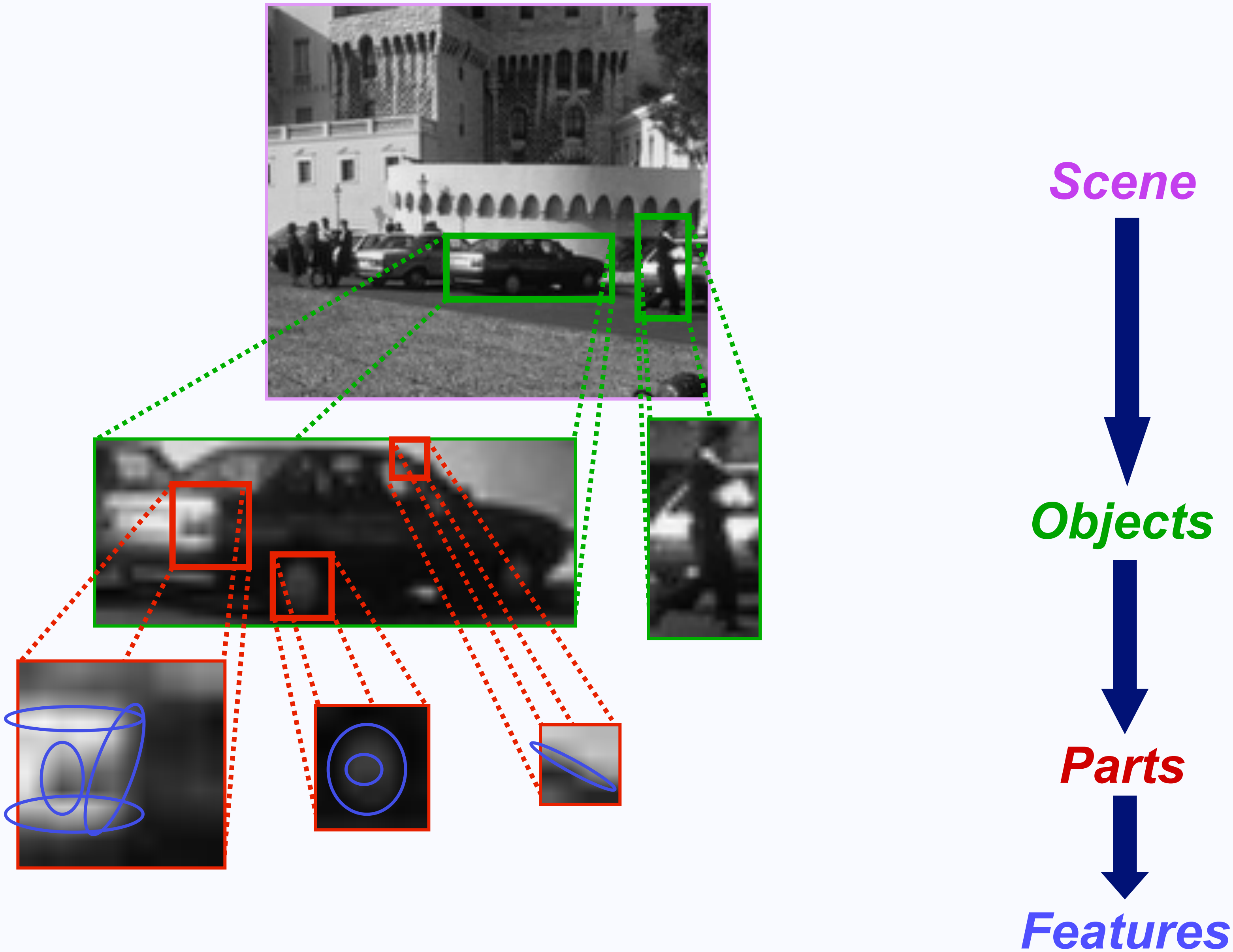
Llamas

Cannons

GOALS:

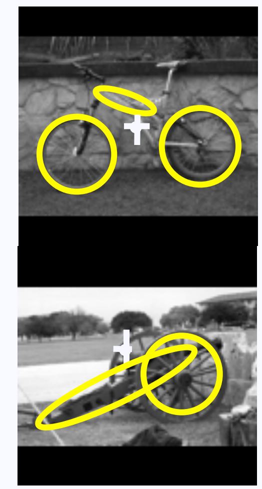
- Visually *recognize* and *localize* object categories
- Robustly *learn* appearance models from few examples
  - Use hierarchical models to *transfer* knowledge among categories
  - Nonparametric, *Dirichlet process* prior gives flexibility

Scenes, Objects, and Parts



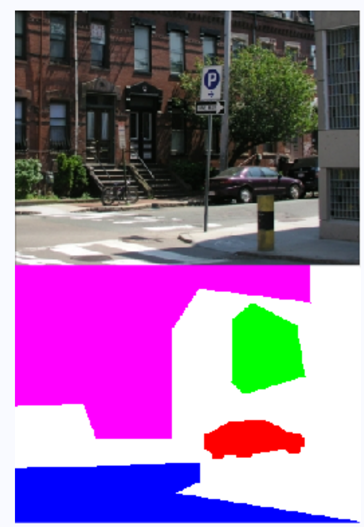
Outline

Object Recognition with Shared Parts



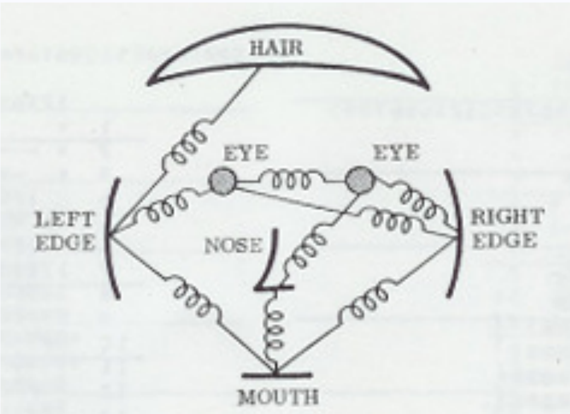
- Learning parts via Dirichlet processes
- Hierarchical DP model for 16 object categories

Multiple Object Scenes

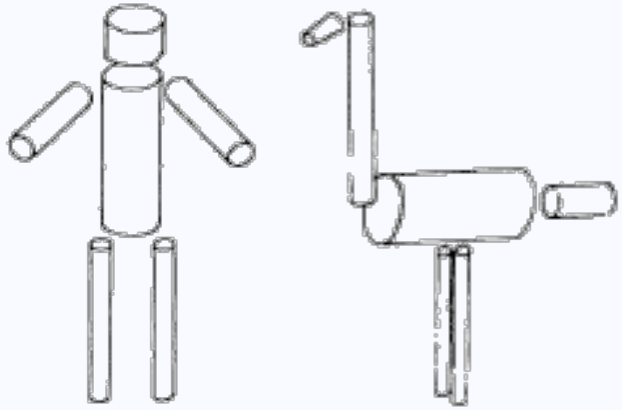


- Transformed Dirichlet processes
- Part-based models for visual scenes

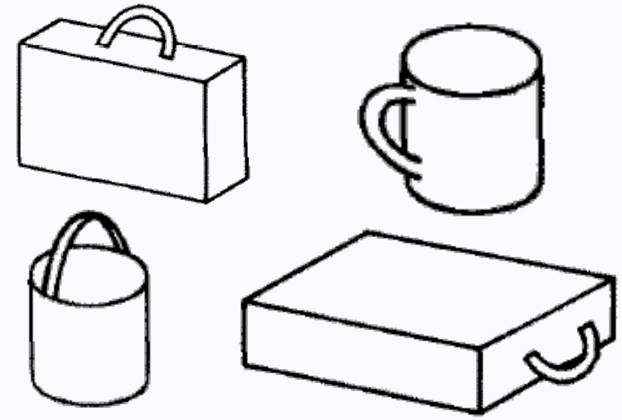
Part-Based Models for Objects



Pictorial Structures
Fischler & Elschlager, 1973

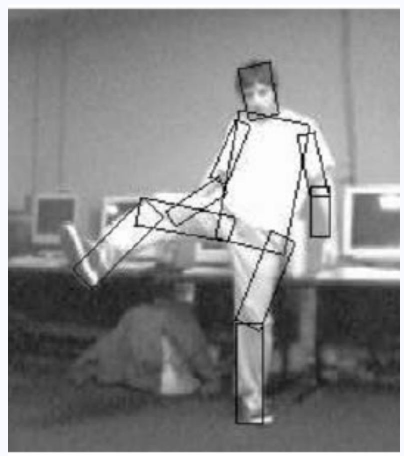


Generalized Cylinders
Marr & Nishihara, 1978

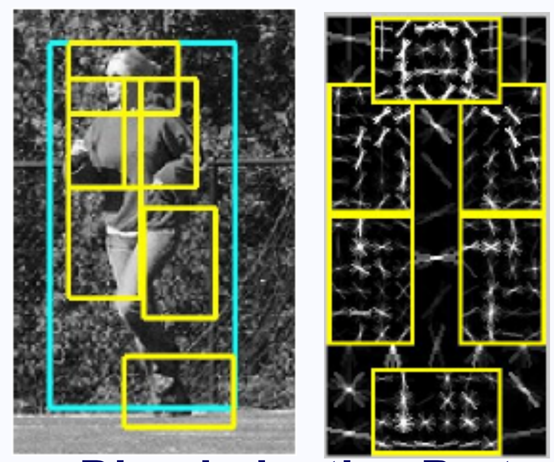


Recognition by Components
Biederman, 1987

Constellation Model
Perona, Weber, Welling, Fergus, Fei-Fei, 2000 to …

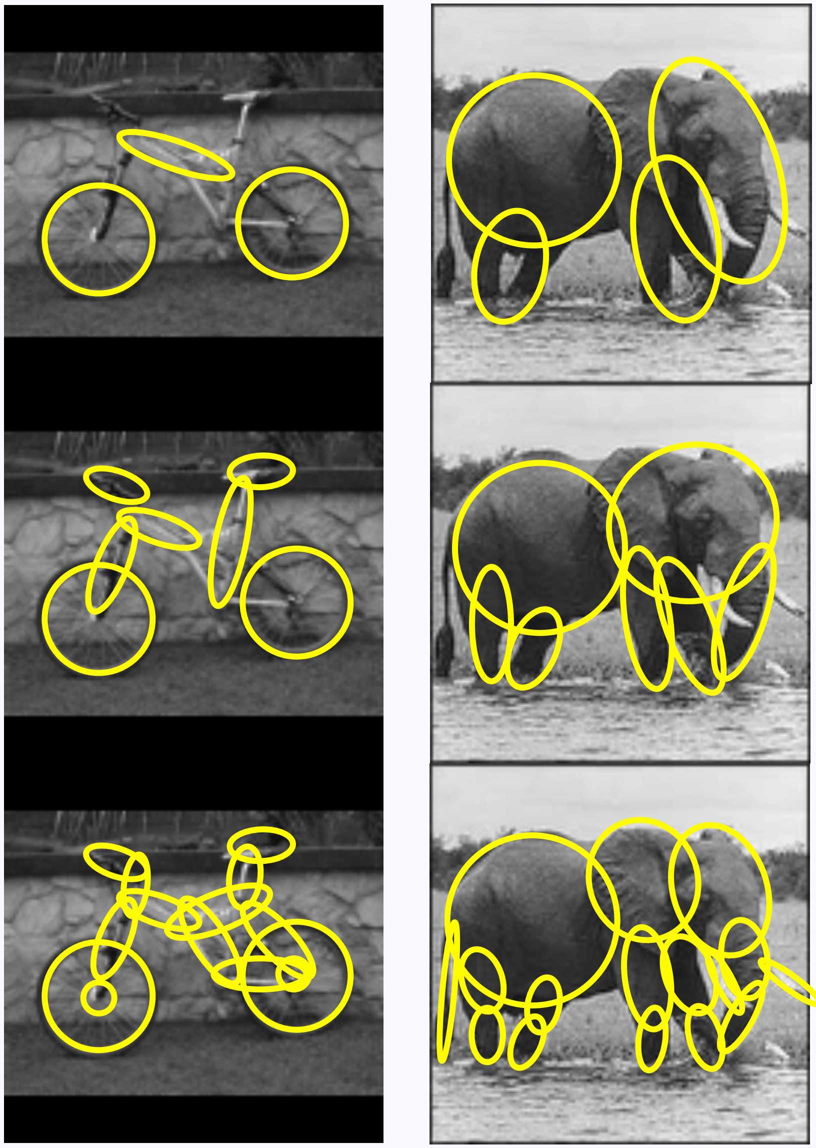


Efficient Matching
Felzenszwalb & Huttenlocher, 2005



Discriminative Parts
Felzenszwalb, McAllester, Ramanan, 2008 to …

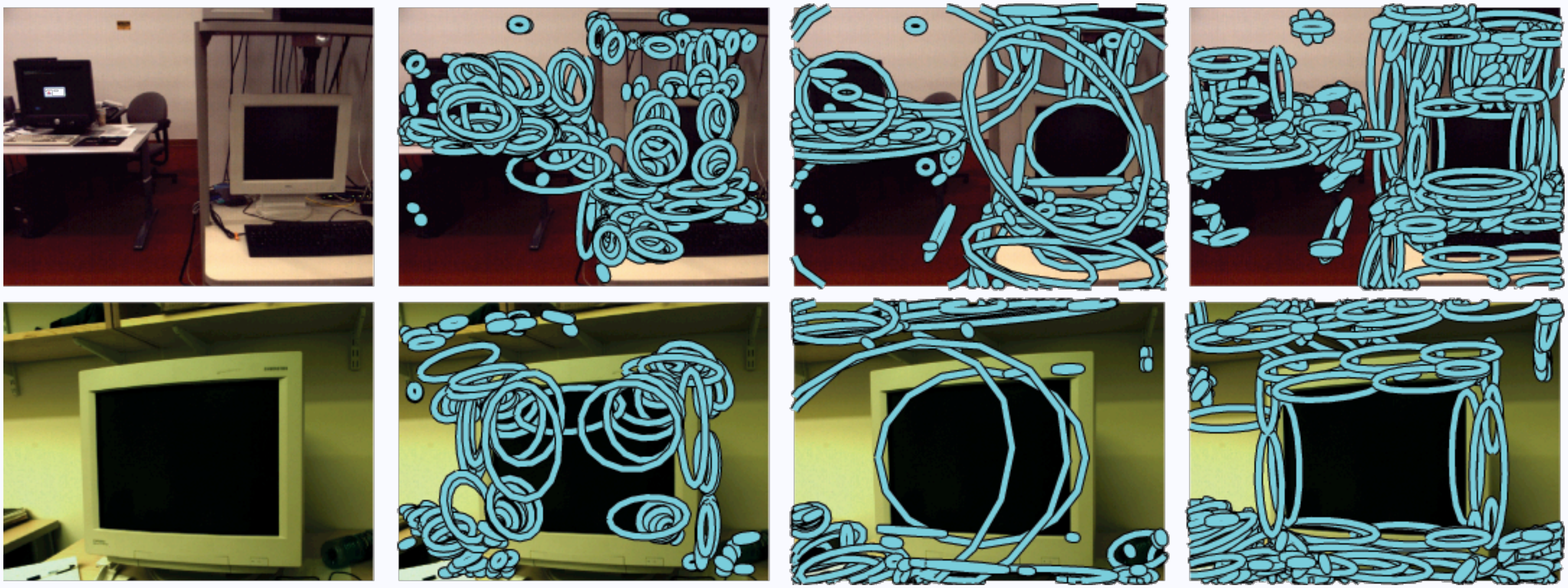
Counting Objects & Parts



How many parts?

How many objects?

From Images to Features



Affinely Adapted Harris Corners

Maximally Stable Extremal Regions

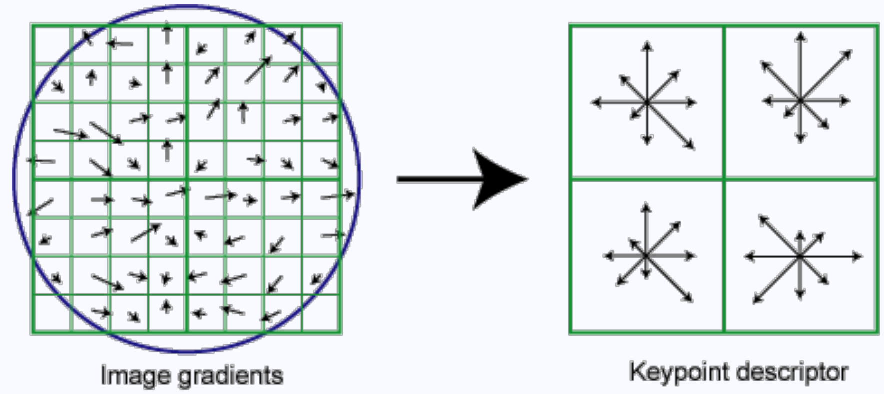
Linked Sequences of Canny Edges

- Some invariance to lighting & pose variations
- Dense, multiscale, over-segmentation of image

A Discrete Feature Vocabulary

SIFT Descriptors

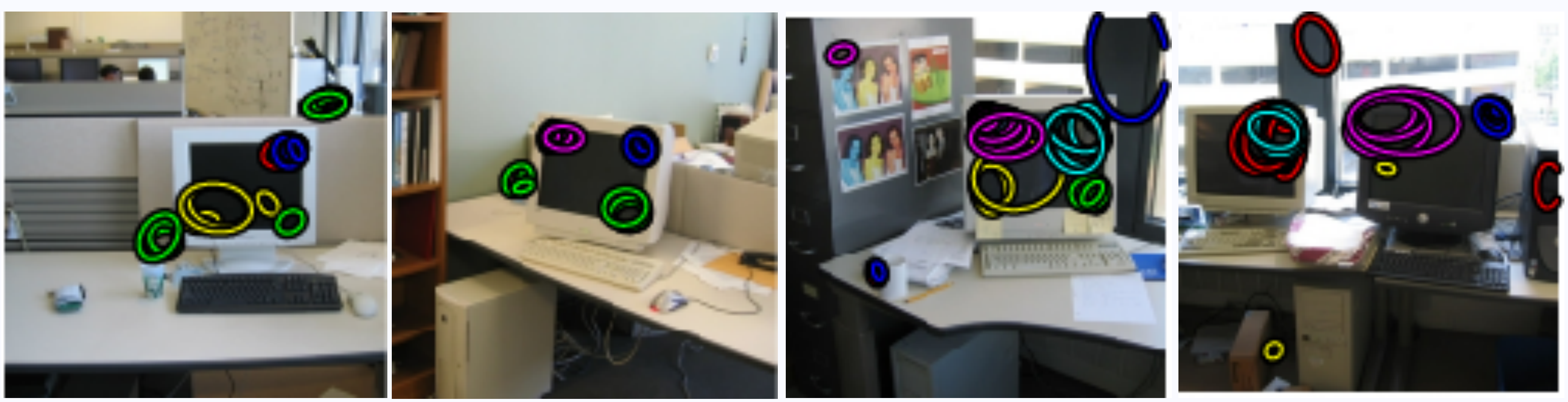
- Normalized histograms of orientation energy
- Compute ~1,000 word dictionary via K-means
- Map each feature to nearest *visual word*



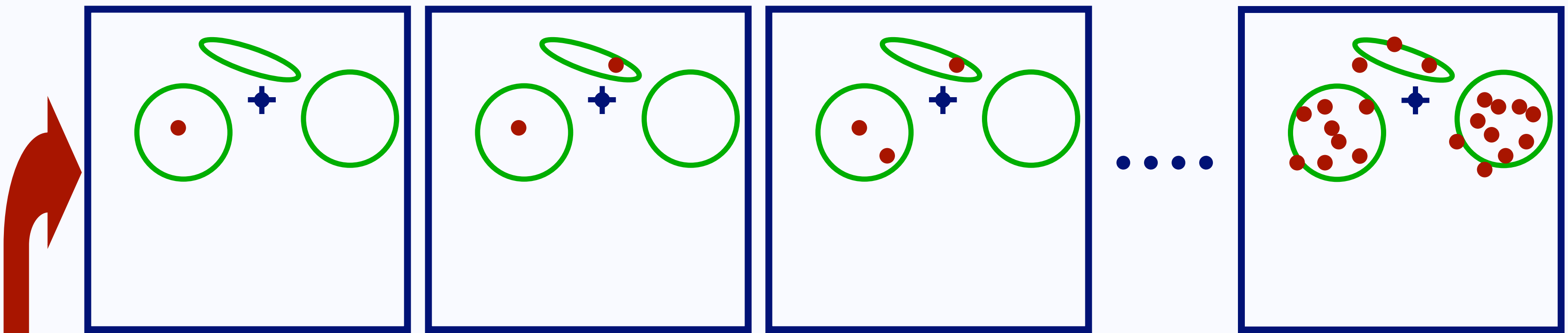
Lowe, IJCV 2004

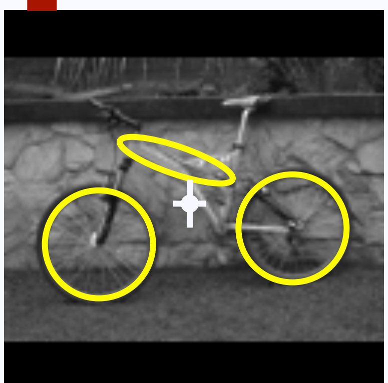
$w_{ji} \longrightarrow$ appearance of feature i in image j

$v_{ji} \longrightarrow$ 2D position of feature i in image j



Generative Model for Objects

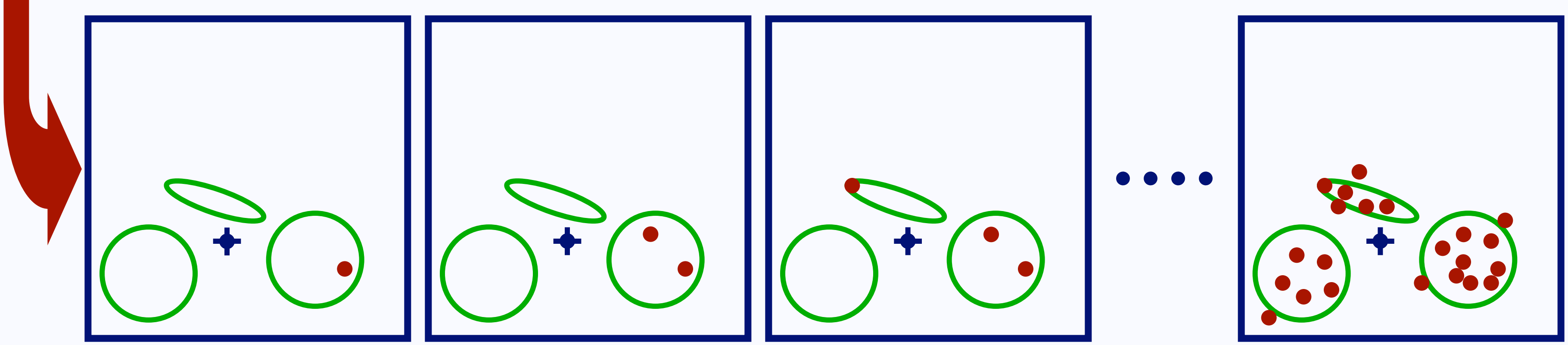




For each image: Sample a reference position

For each feature:

- ➢ Randomly choose one part
- ➢ Sample from that part's feature distribution



Objects as Mixture Models

- For a fixed reference position, our generative model is equivalent to a finite mixture model:

$$p(\underbrace{w_{ji}}_{\text{Feature appearance}}, \underbrace{v_{ji}}_{\text{Feature position}} | \rho_j) = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{Pr(part)}} \underbrace{\eta_k(w_{ji})}_{\text{Pr(appearance $\mid$ part)}} \underbrace{\mathcal{N}(v_{ji}; \mu_k + \rho_j, \Lambda_k)}_{\text{Pr(position $\mid$ part)}}$$

- How many parts should we choose?
  - ➢ Too few reduces model accuracy
  - ➢ Too many causes overfitting & poor generalization

Objects as Distributions

$$p(w_{ji}, v_{ji} | \rho_j) = \sum_{k=1}^{\infty} \pi_k \underbrace{\eta_k(w_{ji})}_{\text{Pr(appearance | part)}} \underbrace{\mathcal{N}(v_{ji}; \mu_k + \rho_j, \Lambda_k)}_{\text{Pr(position | part)}}$$

Feature appearance

Feature position

DP prior

- Parts are defined by *parameters*, which encode distributions on visual features:

$$\theta_k = \{\eta_k \, , \mu_k \, , \Lambda_k\}$$

- Objects are defined by *distributions* on the infinitely many potential part parameters:

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad\qquad \boldsymbol{\pi} \sim \text{Stick}(\alpha)$$

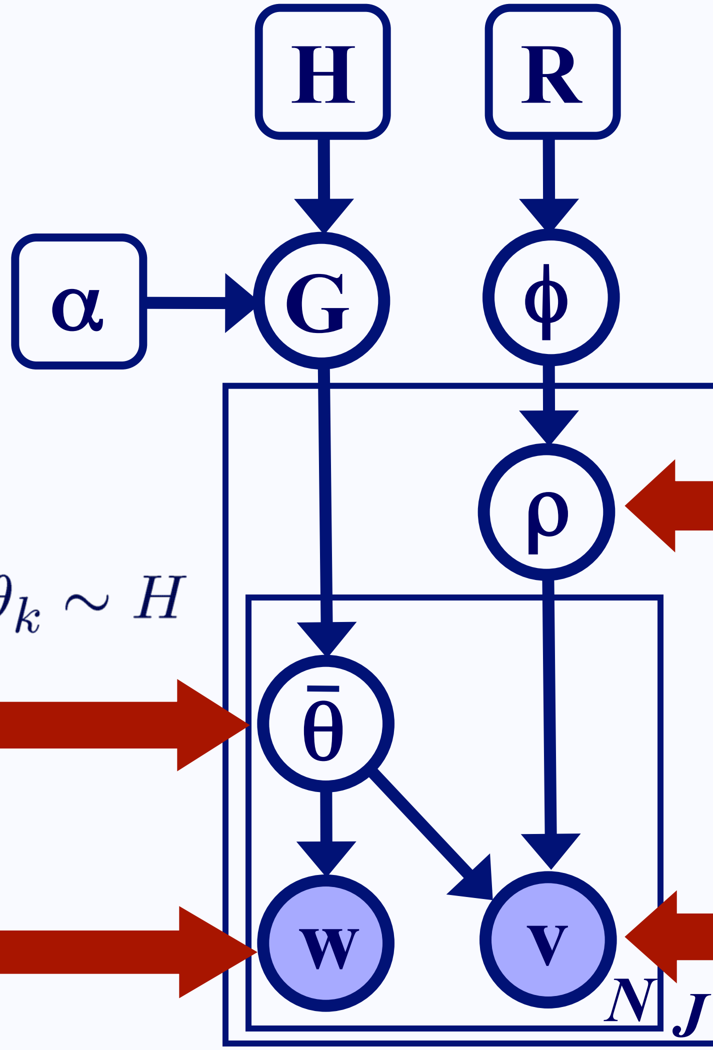
Dirichlet Process Object Model

Part-based object model sampled from DP prior:

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

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

$$\boldsymbol{\pi} \sim \mathrm{Stick}(\alpha) \qquad \theta_k \sim H$$



For each of N features, sample *part parameters:*

$$\bar{\theta}_{ji} \sim G(\theta)$$

Sample multinomial *feature appearance:*

$$w_{ji} \sim \bar{\eta}_{ji}(w)$$

For each of J images, sample a *reference position:*

$$\rho_j \sim \mathcal{N}(\rho; \phi)$$

Sample Gaussian *feature position:*

$$v_{ji} \sim \mathcal{N}(v; \bar{\mu}_{ji} + \rho_j, \bar{\Lambda}_{ji})$$

$$\bar{\theta}_{ji} = \{\bar{\eta}_{ji}, \bar{\mu}_{ji}, \bar{\Lambda}_{ji}\}$$

Learning DPs: Gibbs Sampling

Part Scale &
Visual Sparsity
(insensitive)

Image Scale
(insensitive)

Sample

Integrate

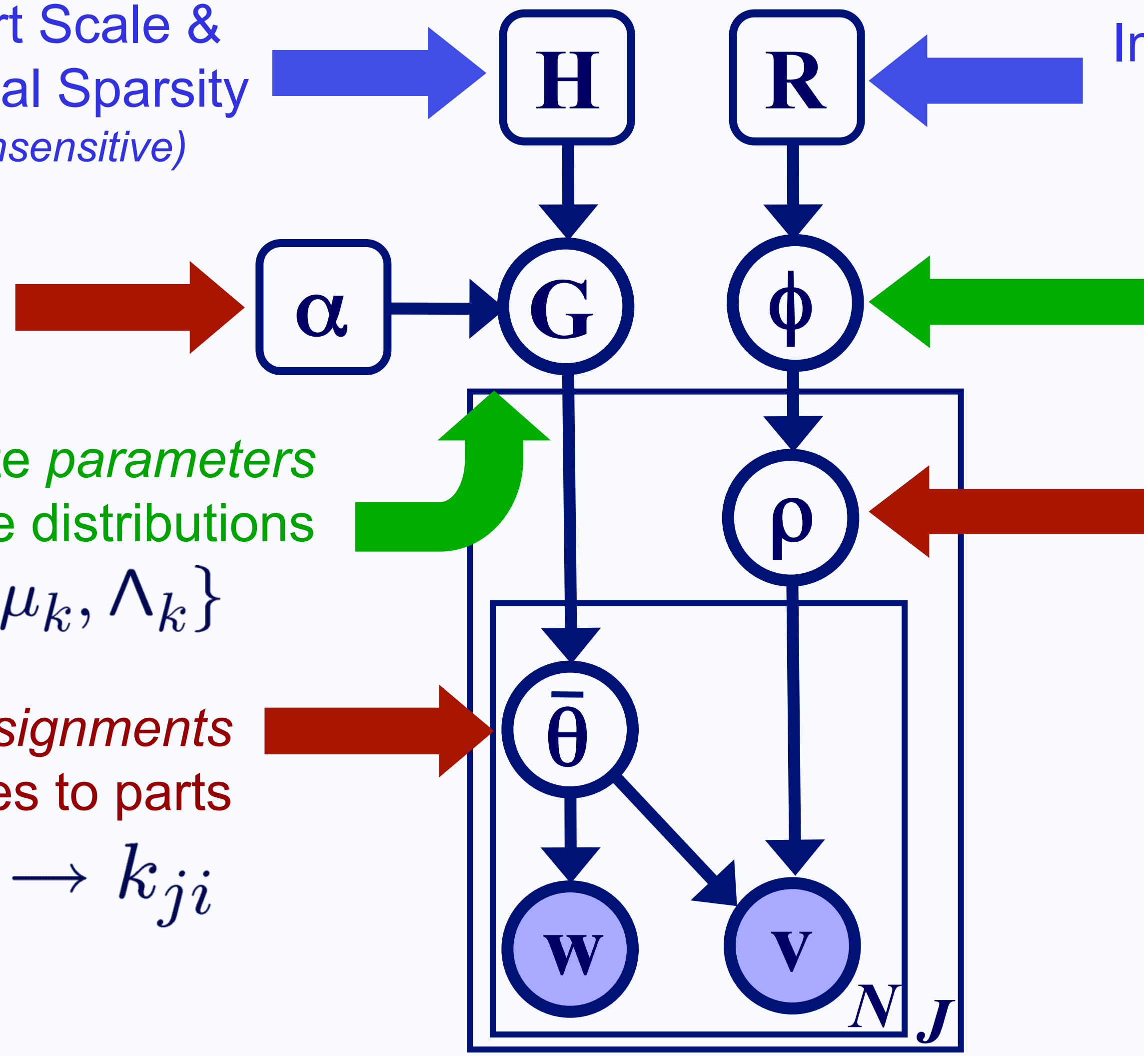
Integrate *parameters*
defining feature distributions

$$\theta_k = \{\eta_k, \mu_k, \Lambda_k\}$$

Sample

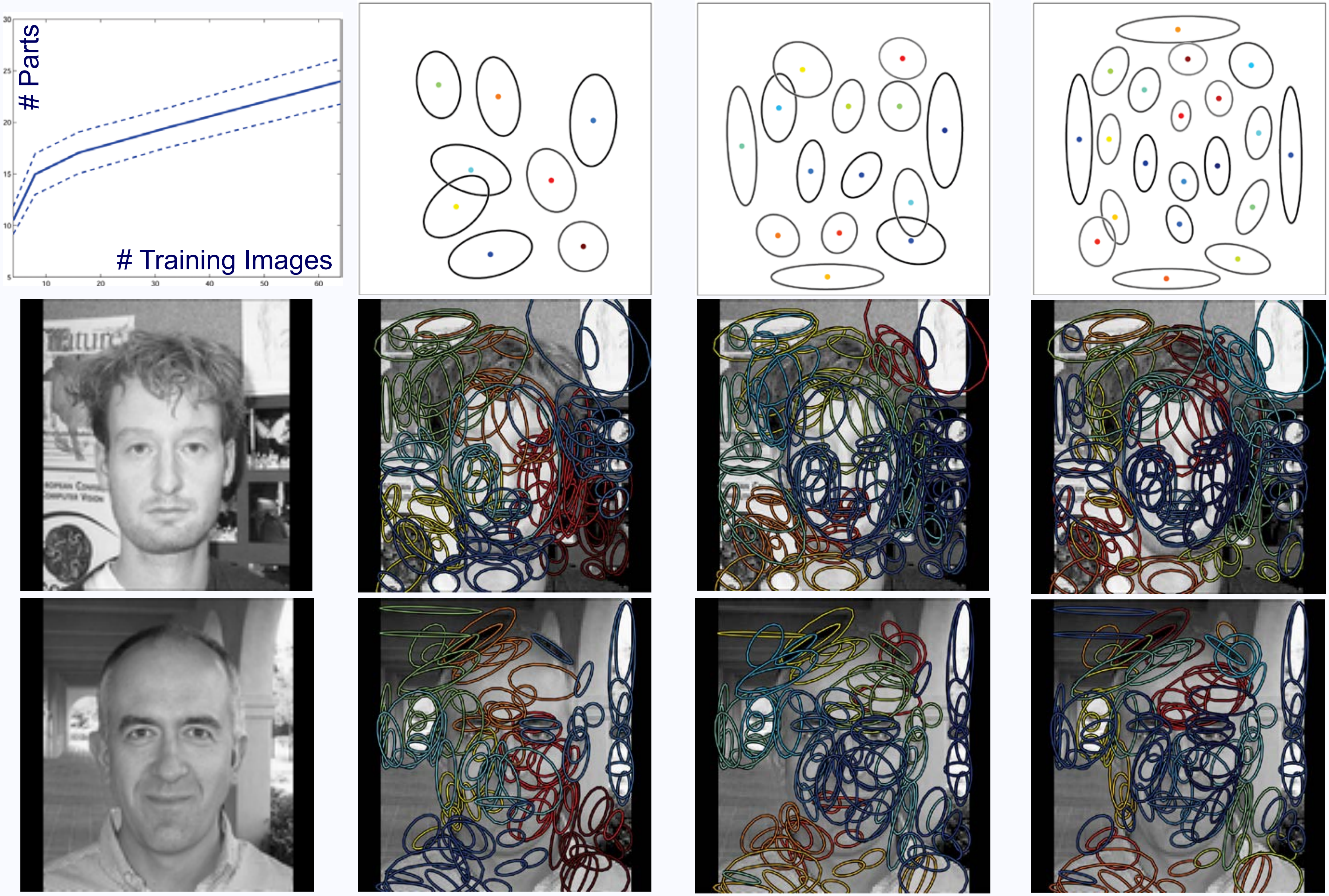
Sample *assignments*
clustering features to parts

$$(w_{ji}, v_{ji}) \rightarrow k_{ji}$$



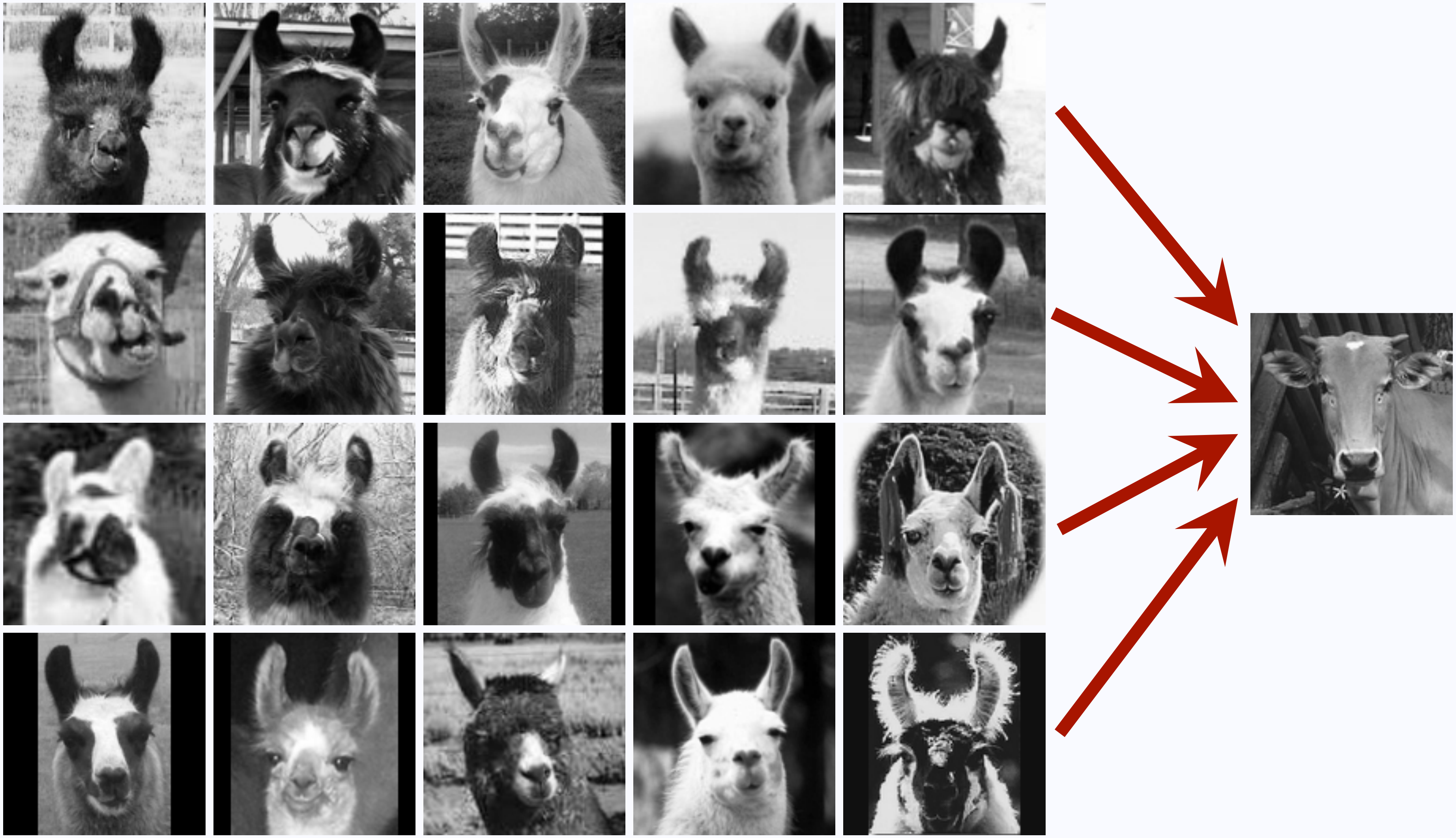
Dirichlet processes have many desirable analytic properties, which lead to efficient *Rao-Blackwellized* learning algorithms

Decomposing Faces into Parts



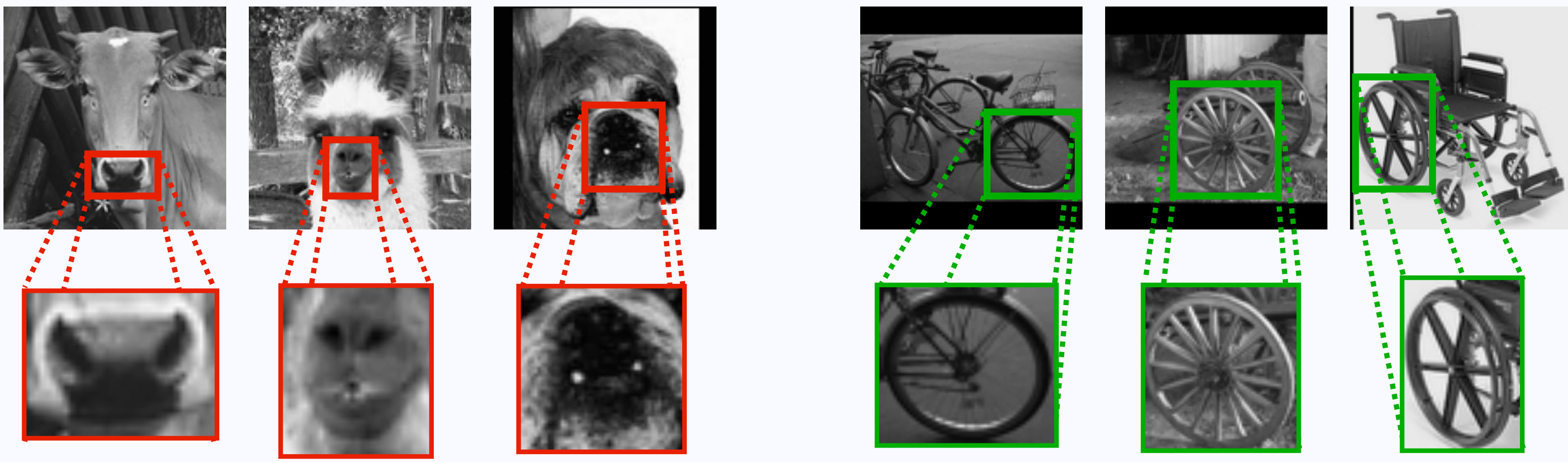
4 Images *16 Images* *64 Images*

Generalizing Across Categories



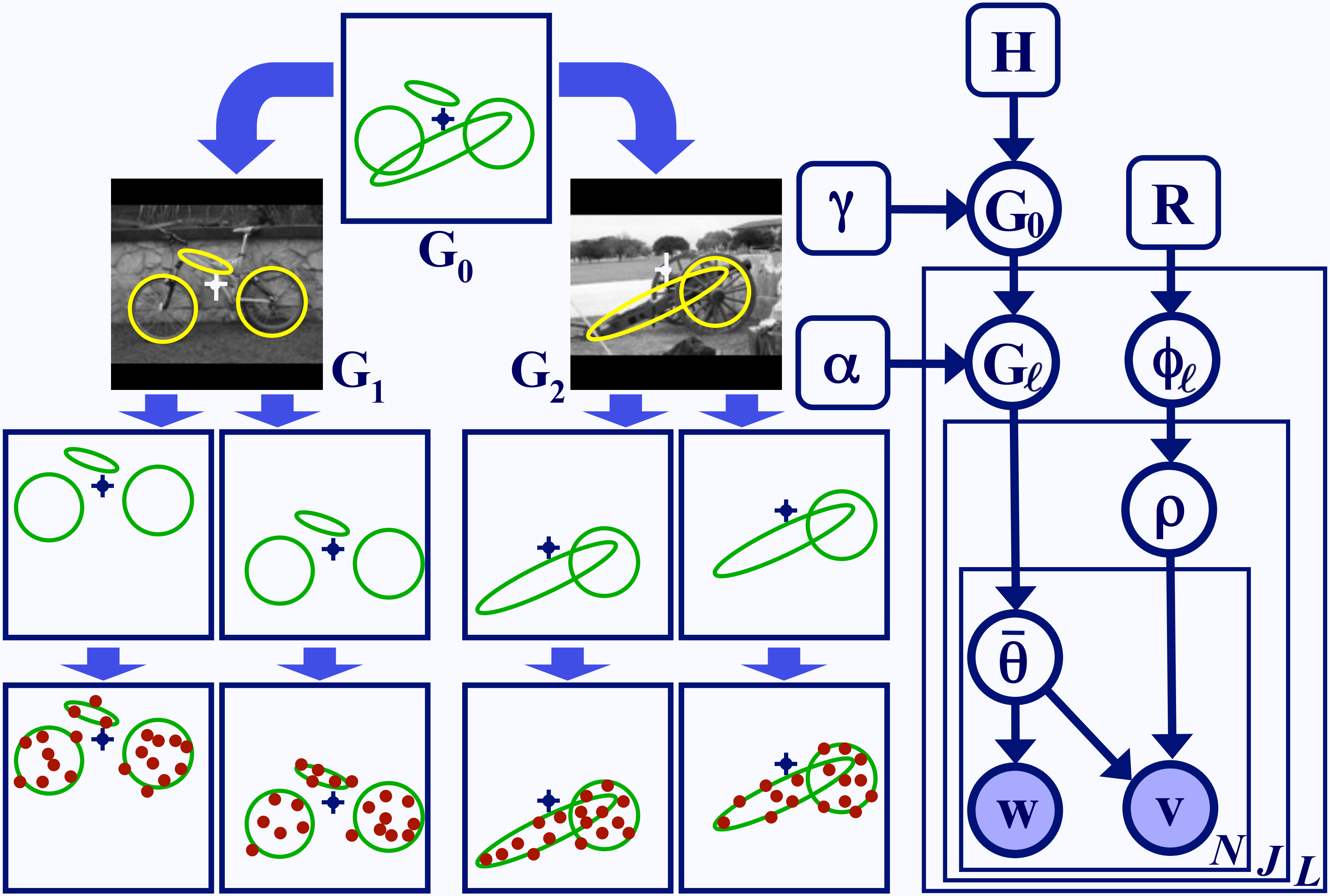
Can we transfer knowledge from one object category to another?

Learning Shared Parts

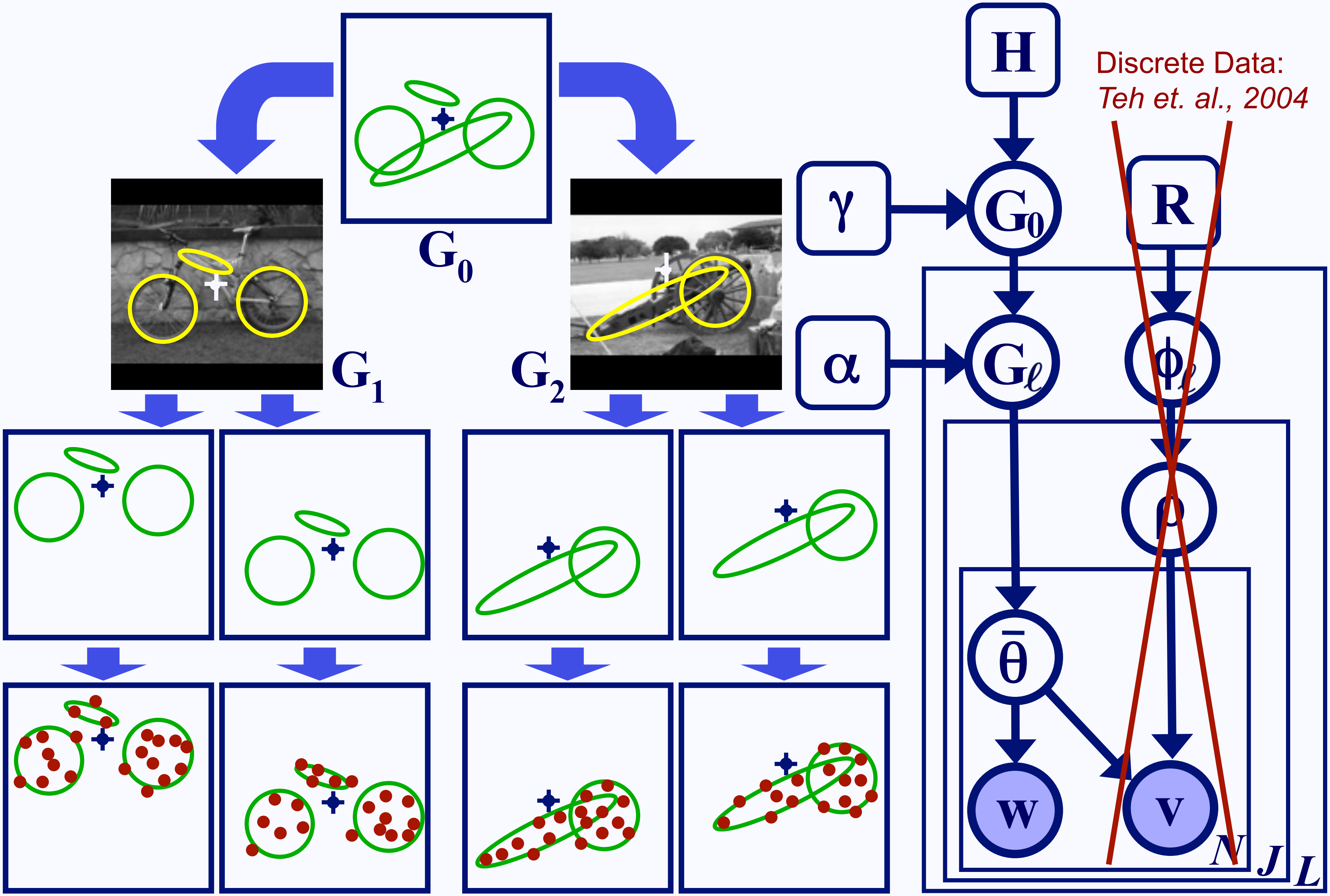


- Objects are often locally similar in appearance
- Discover *parts* shared across categories
  - ➢ How many total parts should we share?
  - ➢ How many parts should each category use?

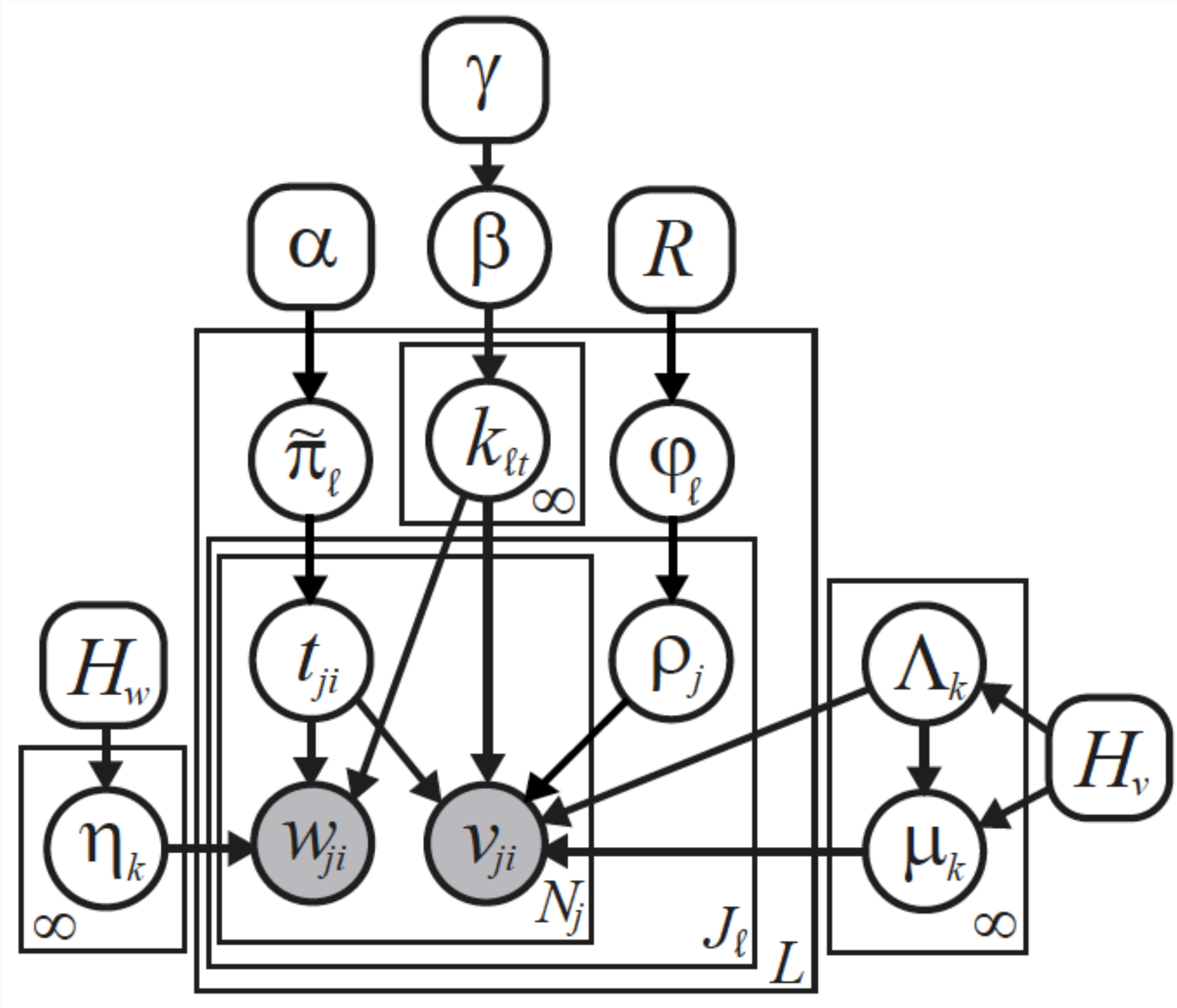
Hierarchical DP Object Model



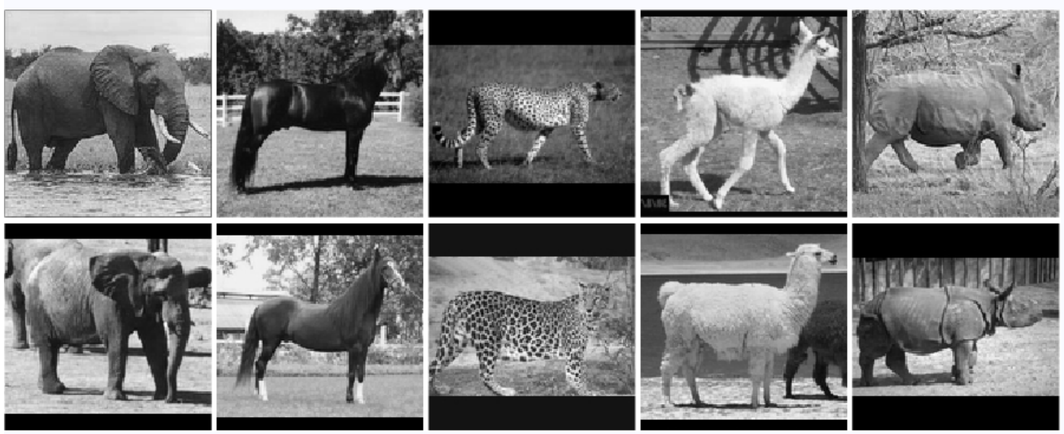
Hierarchical DP Object Model

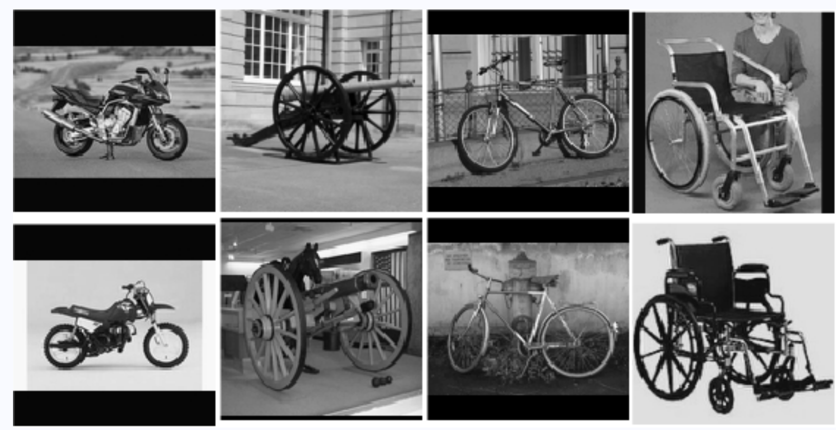


Chinese Restaurant Franchise



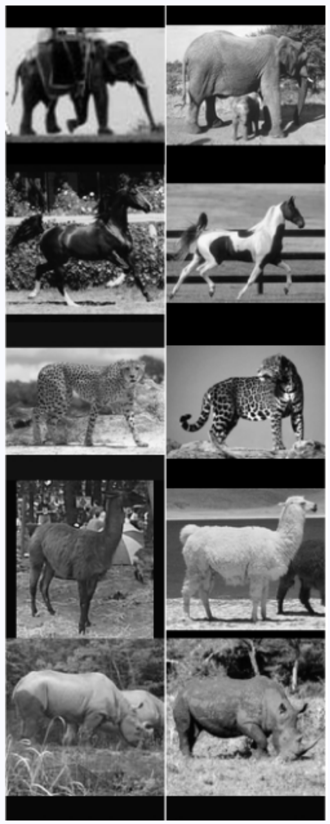
Sharing Parts: 16 Categories



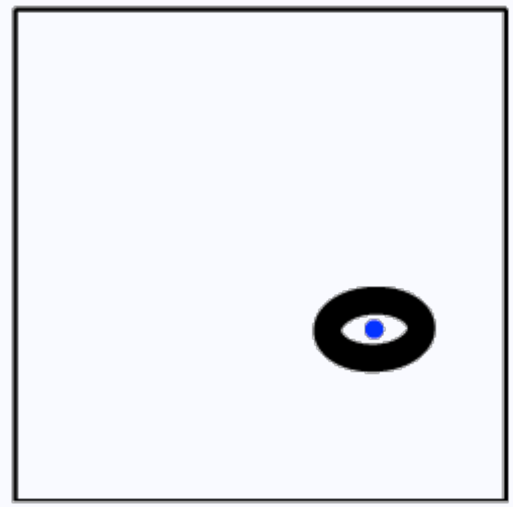


- Caltech 101 Dataset (Li & Perona)
- Horses (Borenstein & Ullman)
- Cat & dog faces (Vidal-Naquet & Ullman)
- Bikes from Graz-02 (Opelt & Pinz)
- Google…

Visualization of Shared Parts

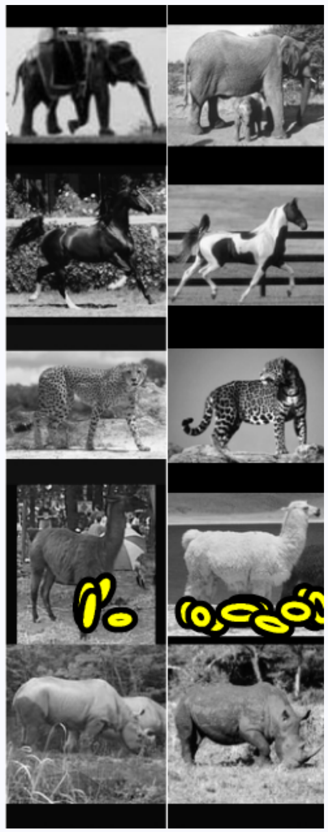


Pr(appearance | part)

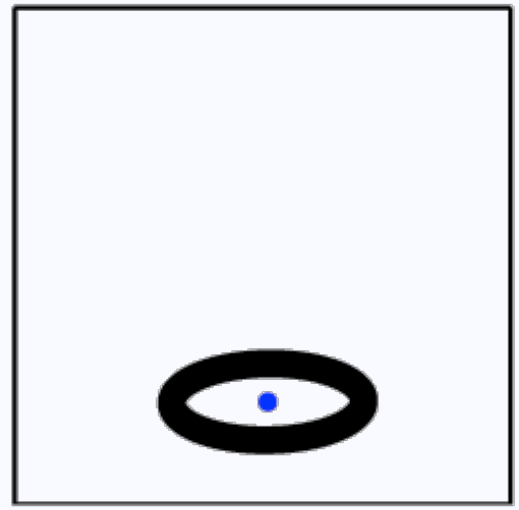


Pr(position | part)

Visualization of Shared Parts

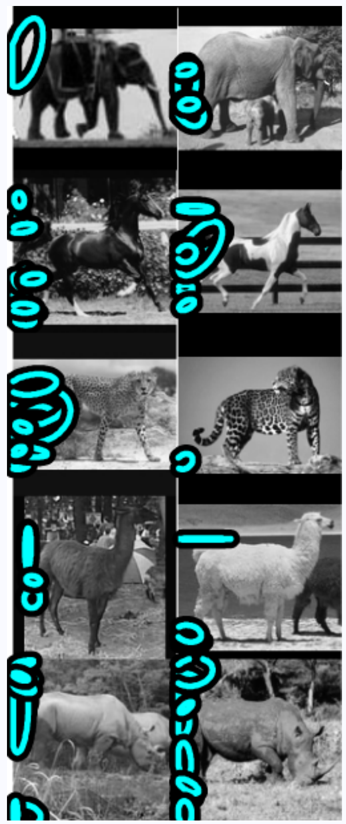


Pr(appearance | part)

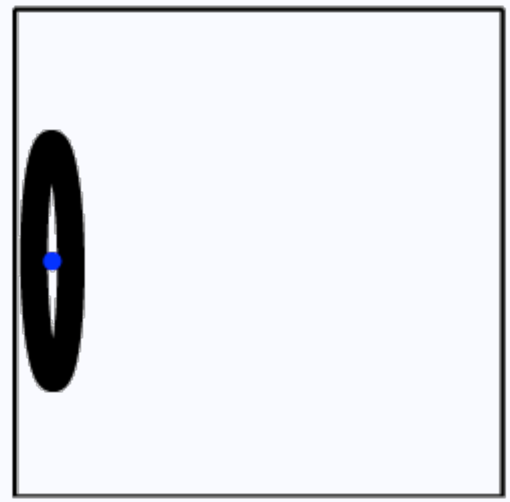


Pr(position | part)

Visualization of Shared Parts

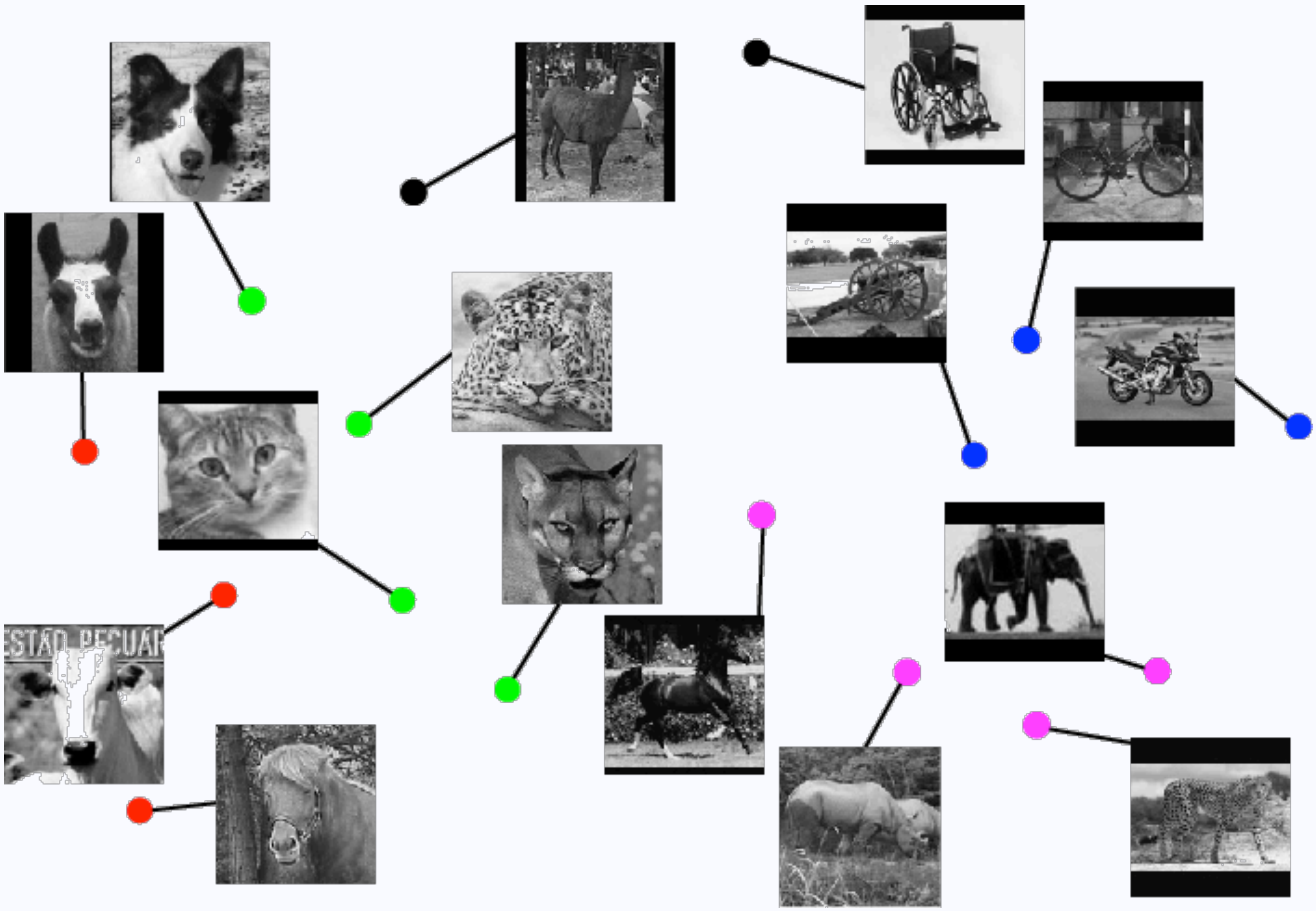


Pr(appearance | part)



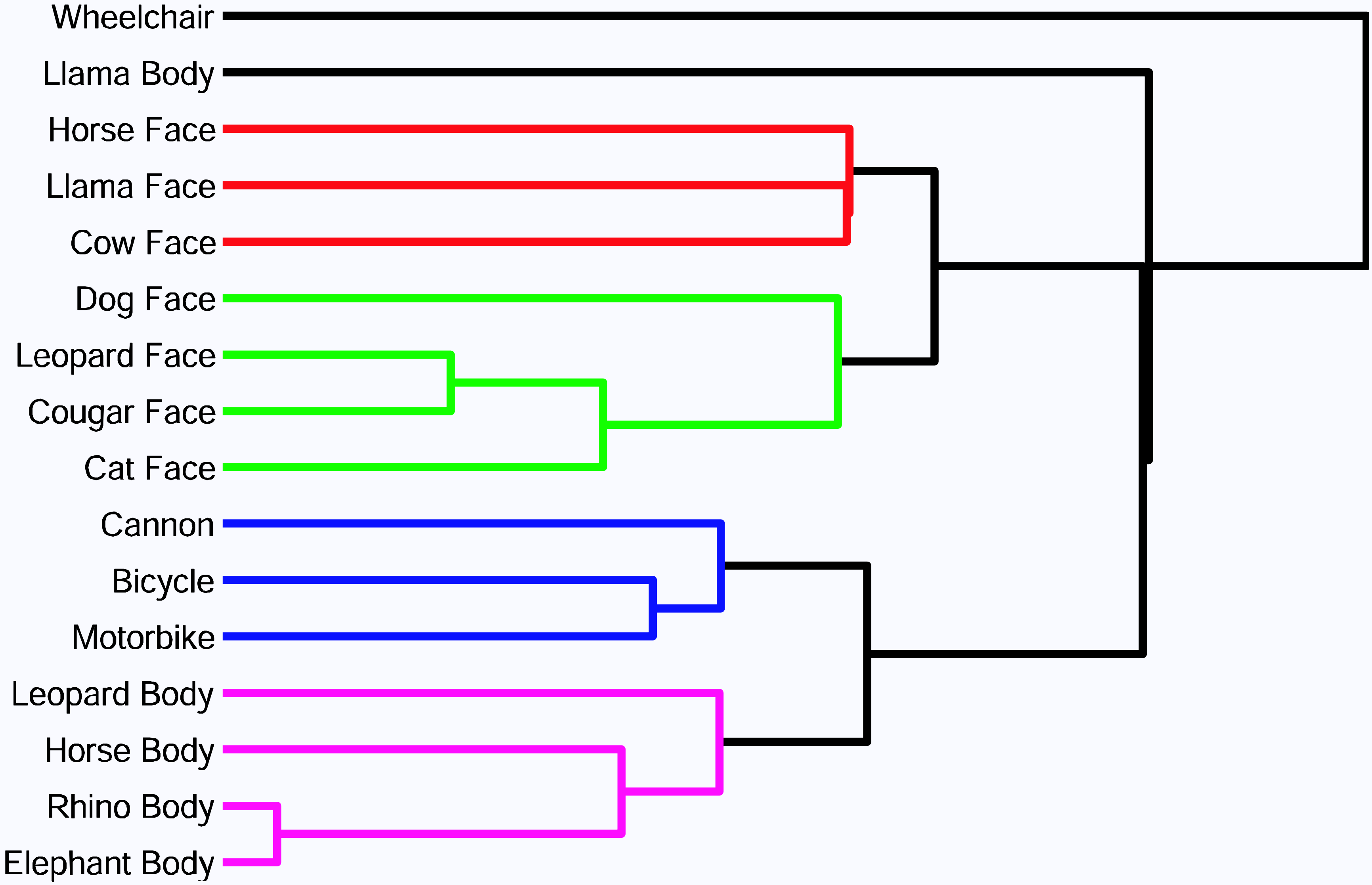
Pr(position | part)

Visualization of Part Densities



MDS Embedding of Pr(part | object)

Visualization of Part Densities

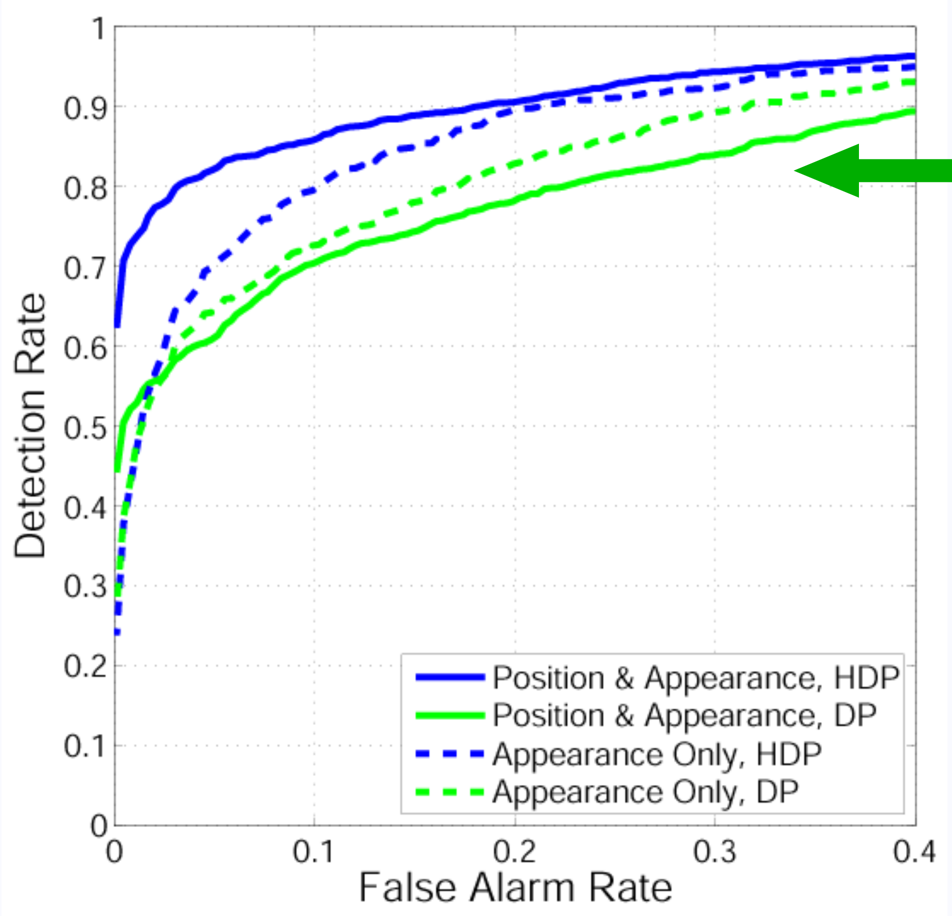


Hierarchical Clustering of Pr(part | object)

Detection Task

versus

Detection Results



6 Training Images per Category

(ROC Curves)

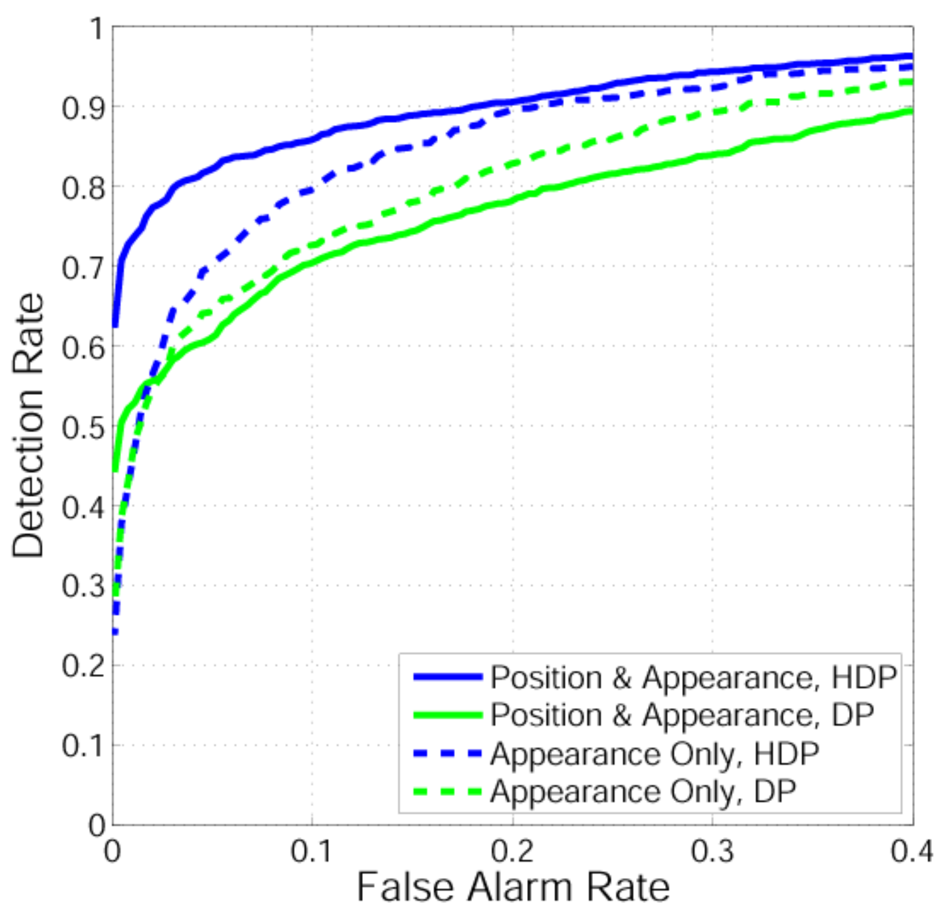
Shared Parts

more accurate than

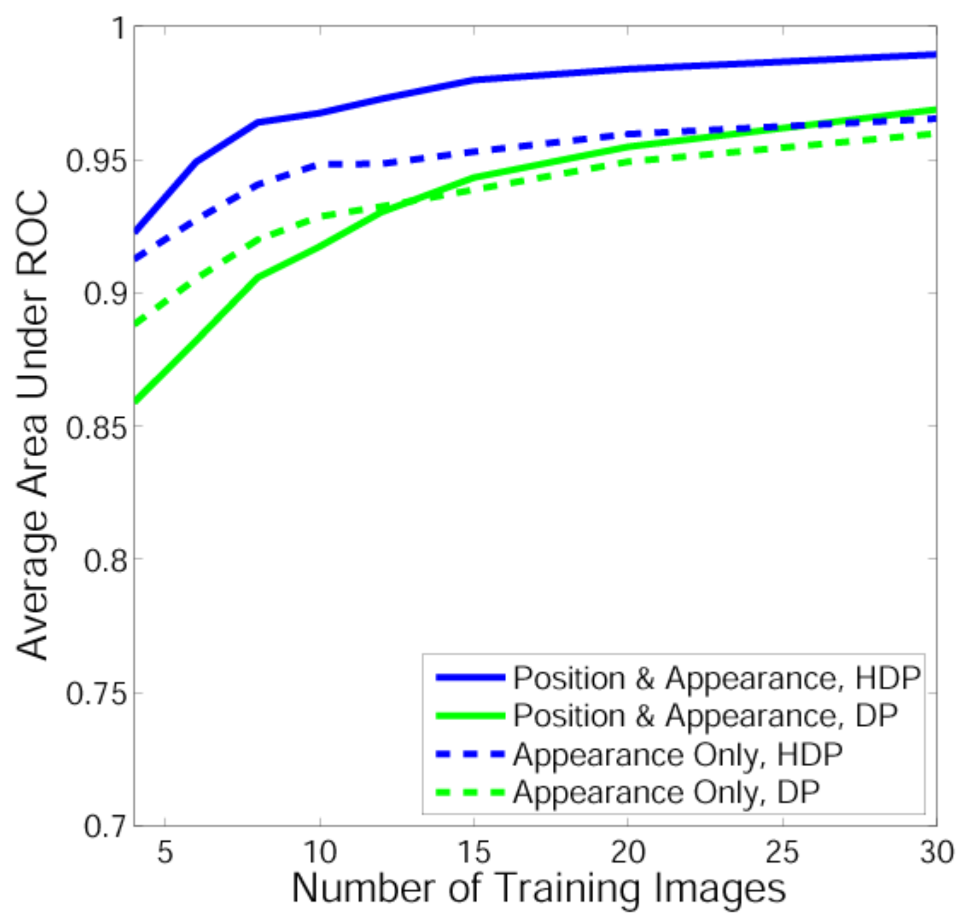
Unshared Parts

Modeling feature positions ***improves shared*** detection, but ***hurts unshared*** detection

Detection Results

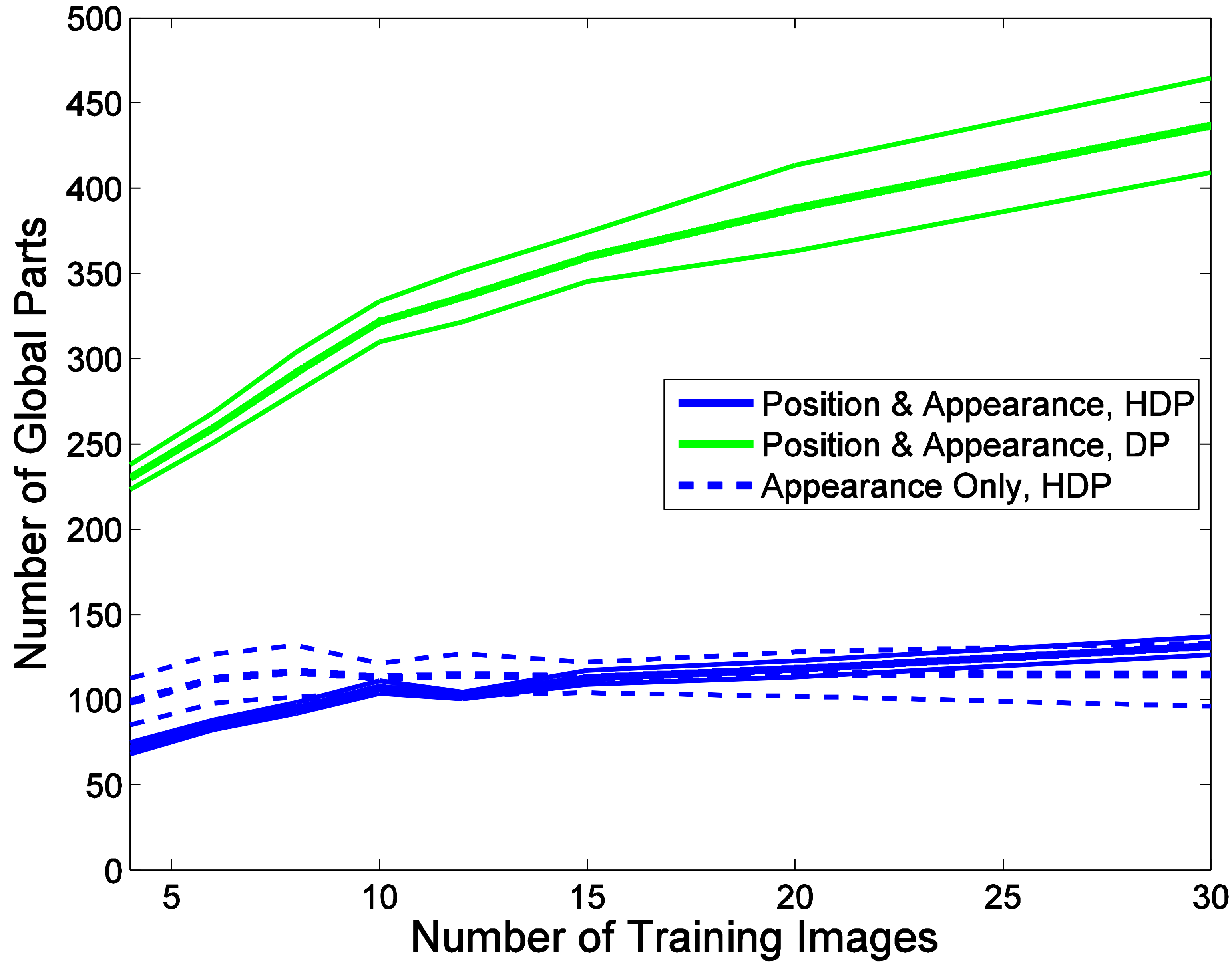


6 Training Images per Category
(ROC Curves)



Detection vs. Training Set Size
(Area Under ROC)

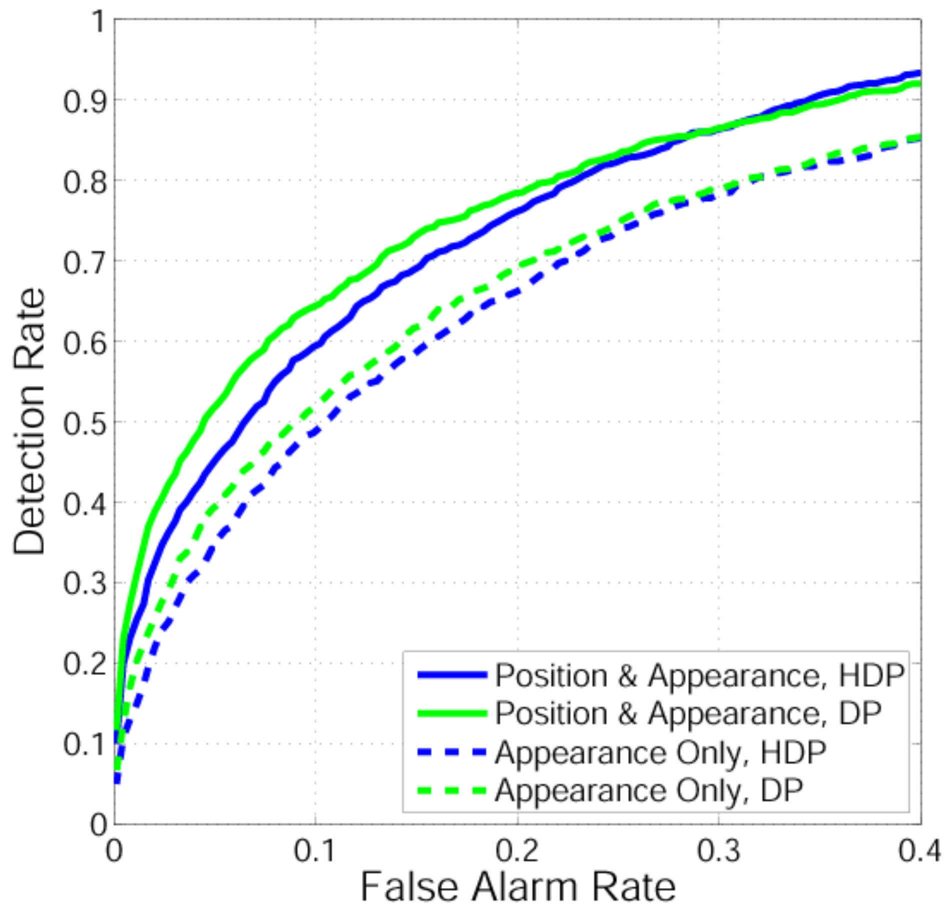
Sharing Simplifies Models



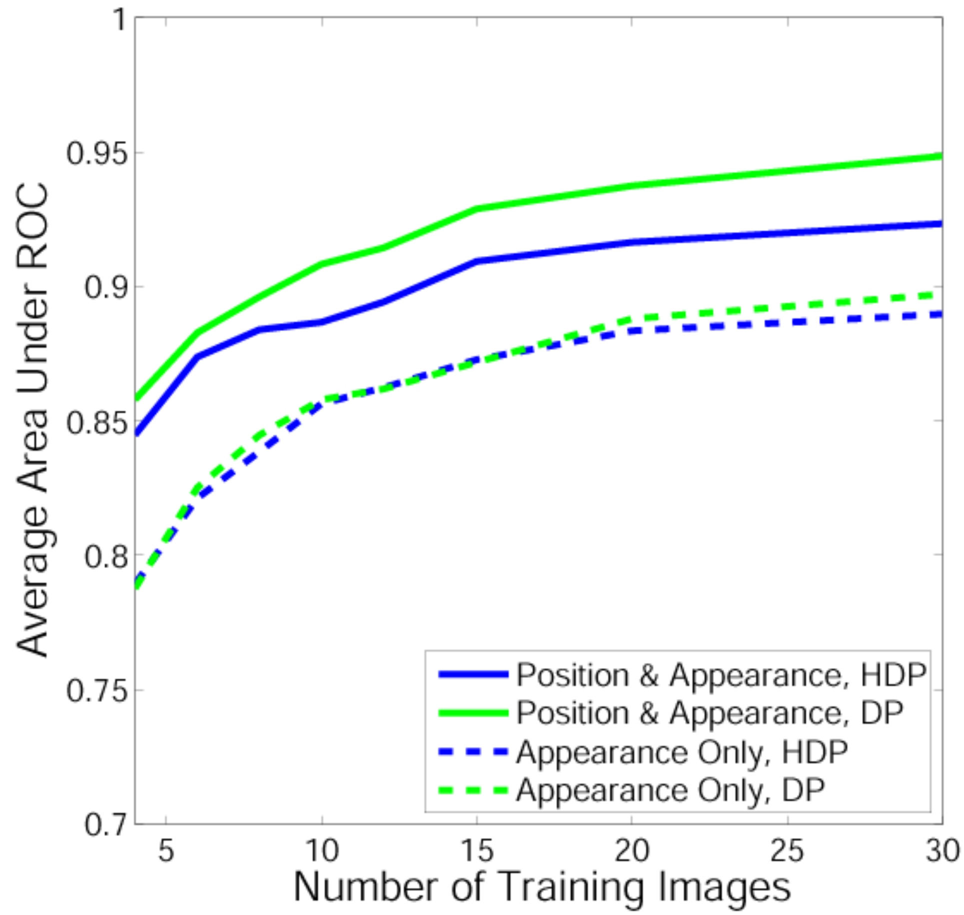
Recognition Task

versus

Recognition Results



6 Training Images per Category
(ROC Curves)

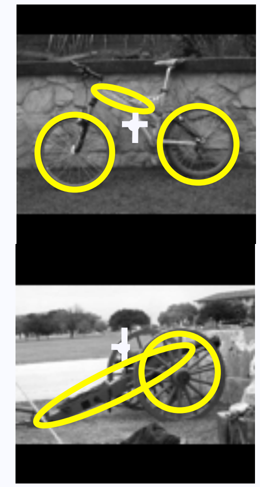


Detection vs. Training Set Size
(Area Under ROC)

Outline

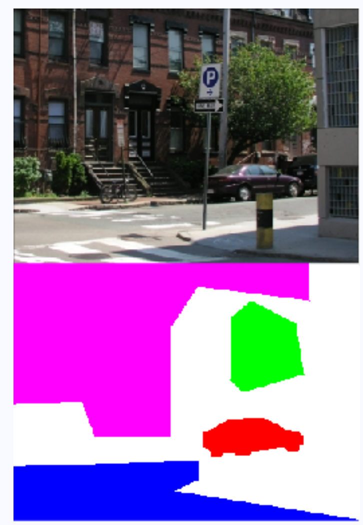
Object Recognition with Shared Parts

- Learning parts via Dirichlet processes
- Hierarchical DP model for 16 object categories



Multiple Object Scenes

- Transformed Dirichlet processes
- Part-based models for visual scenes

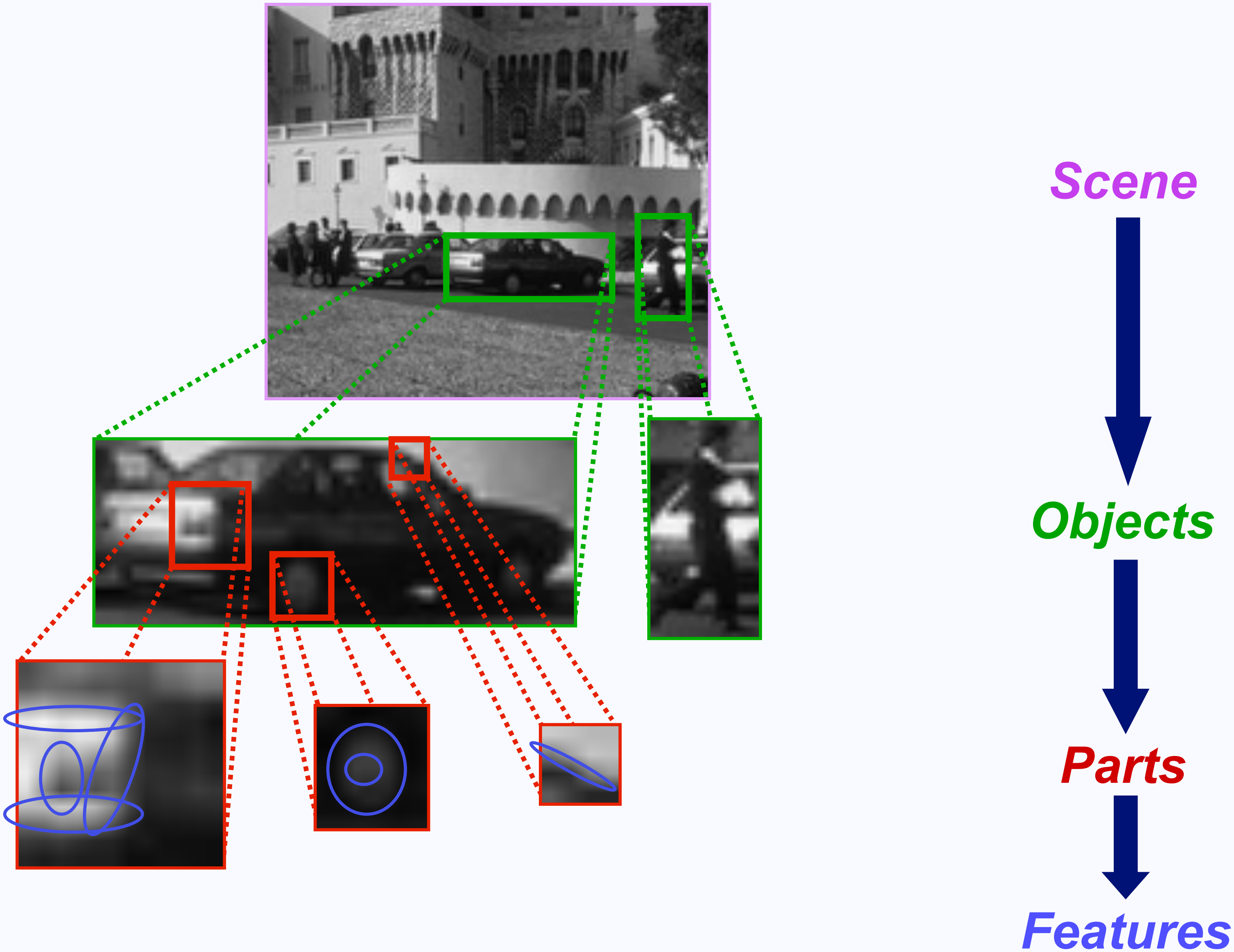


Detecting Objects in Scenes

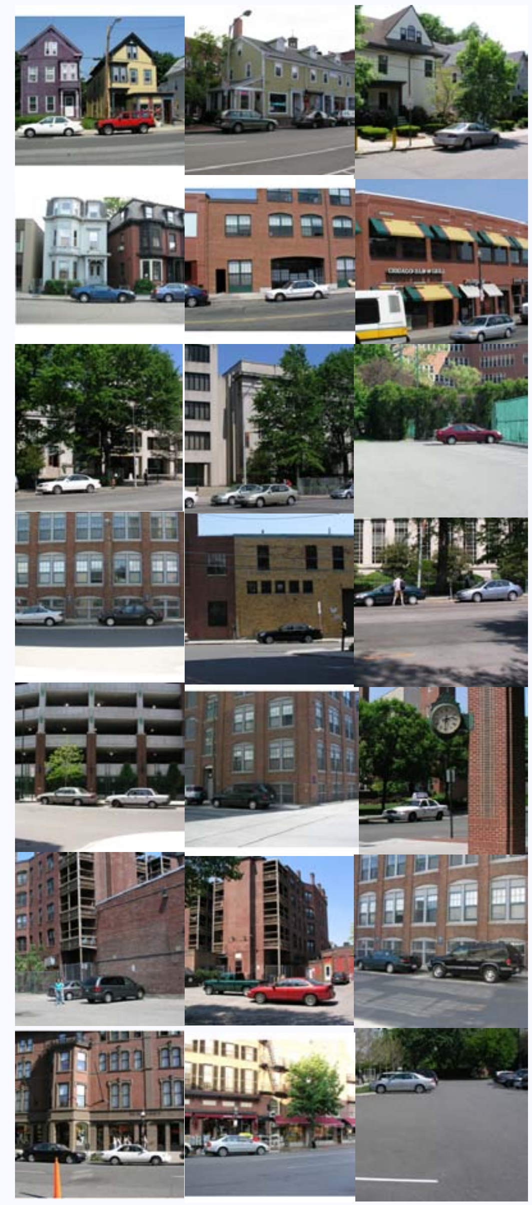
Sliding Window Approach

Greedy Feature Extraction Approach

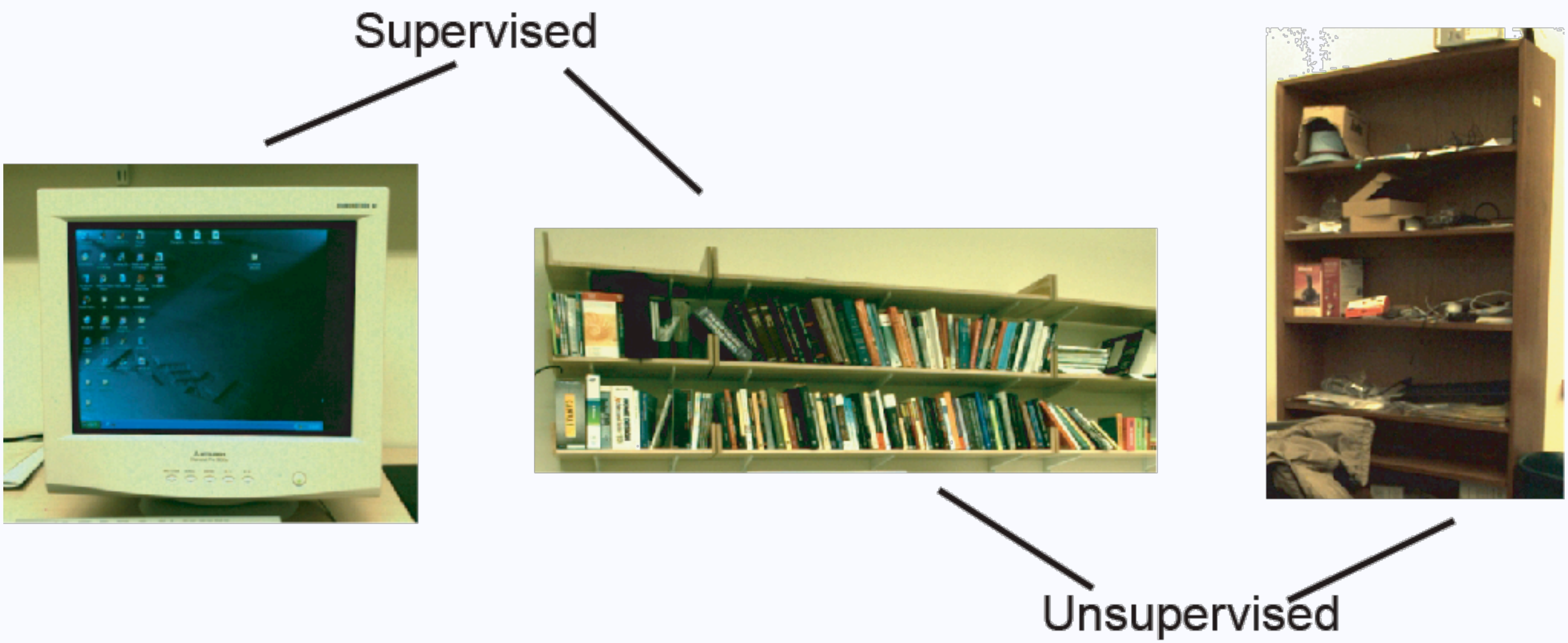
Scenes, Objects, and Parts



Semi-supervised Learning

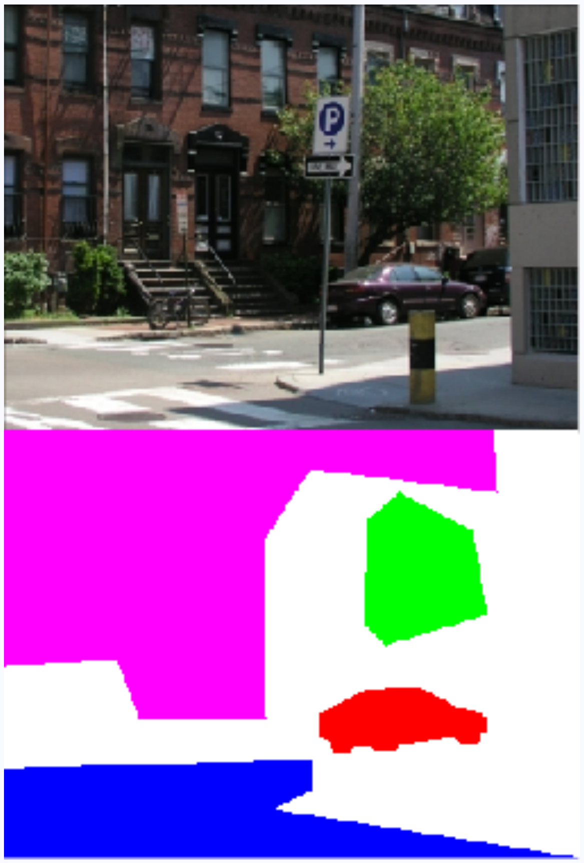


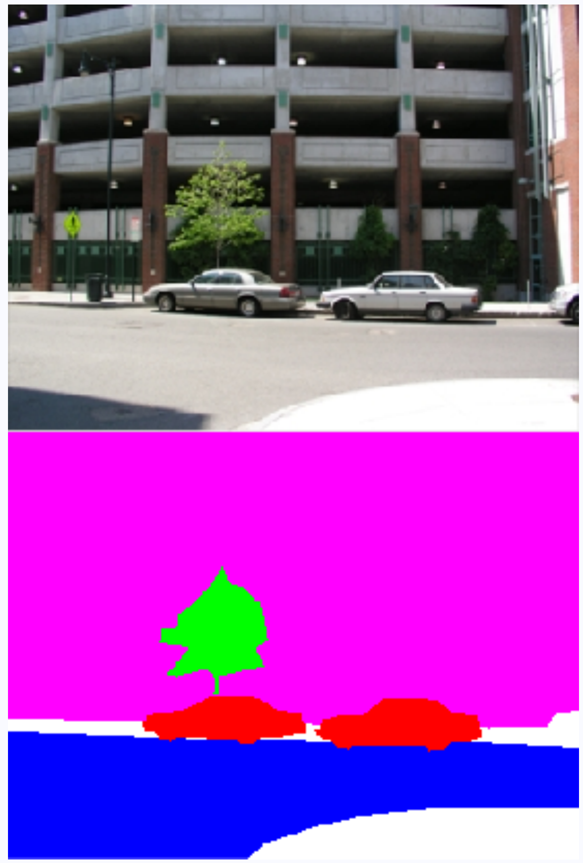
Object vs. Visual Categories



- Assume training data contains object category labels
- Discover underlying visual categories automatically

Multiple Object Scenes





- How many cars are there?
- Where are those cars in the scene?

Standard dependent Dirichlet process models (Gelfand et. al., 2005) inappropriate

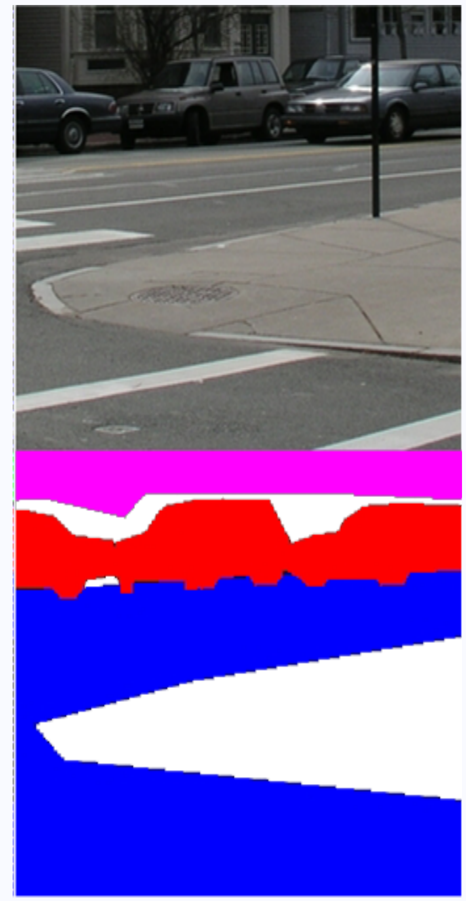
Spatial Transformations

- Let global DP clusters model objects in a *canonical* coordinate frame
- Generate images via a random *set of transformations:*

$$\tau((\mu, \Lambda); \rho) = (\mu + \rho, \Lambda)$$

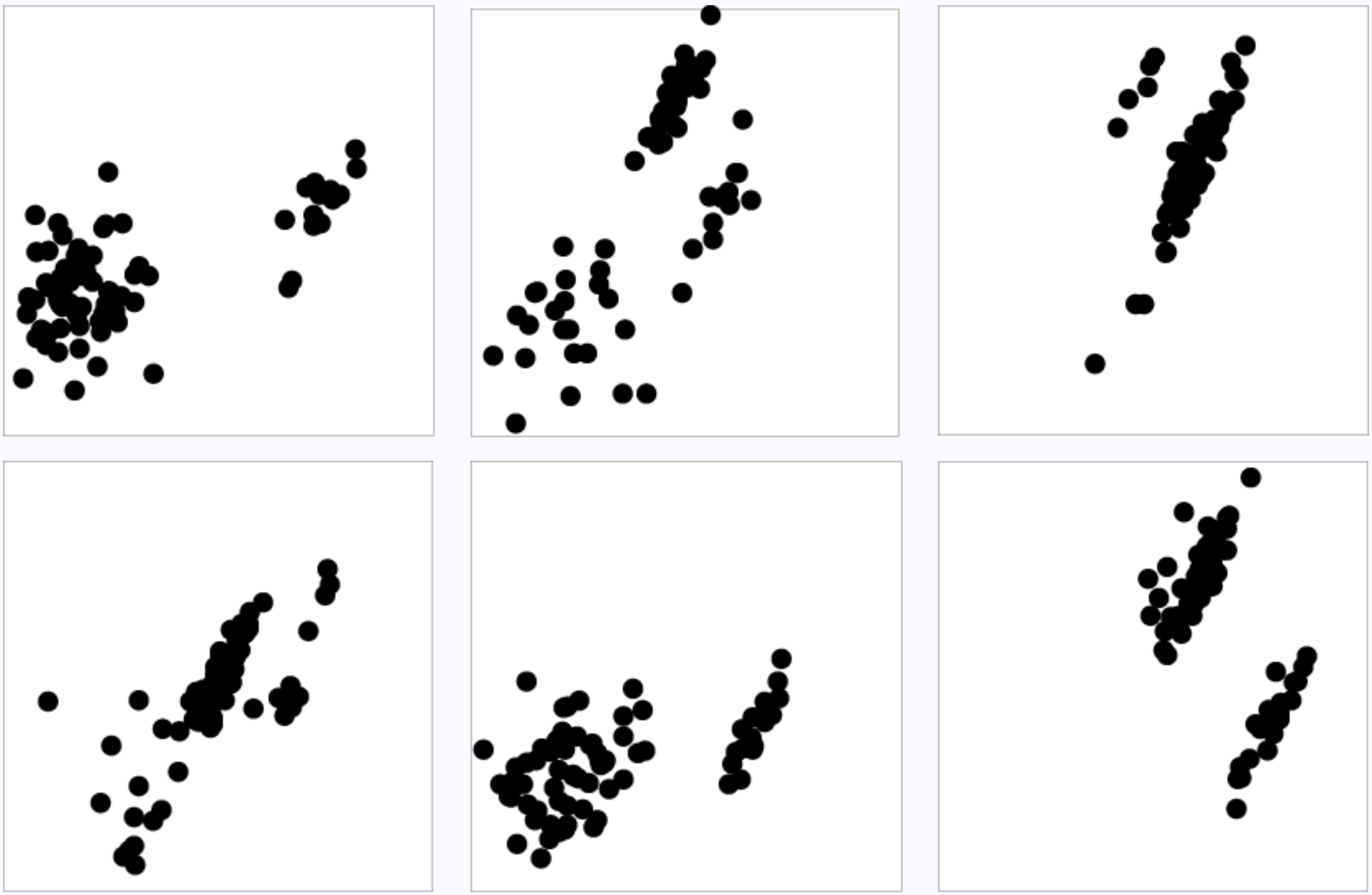
Parameterized family of transformations

Shift cluster from canonical coordinate frame to object location in a given image

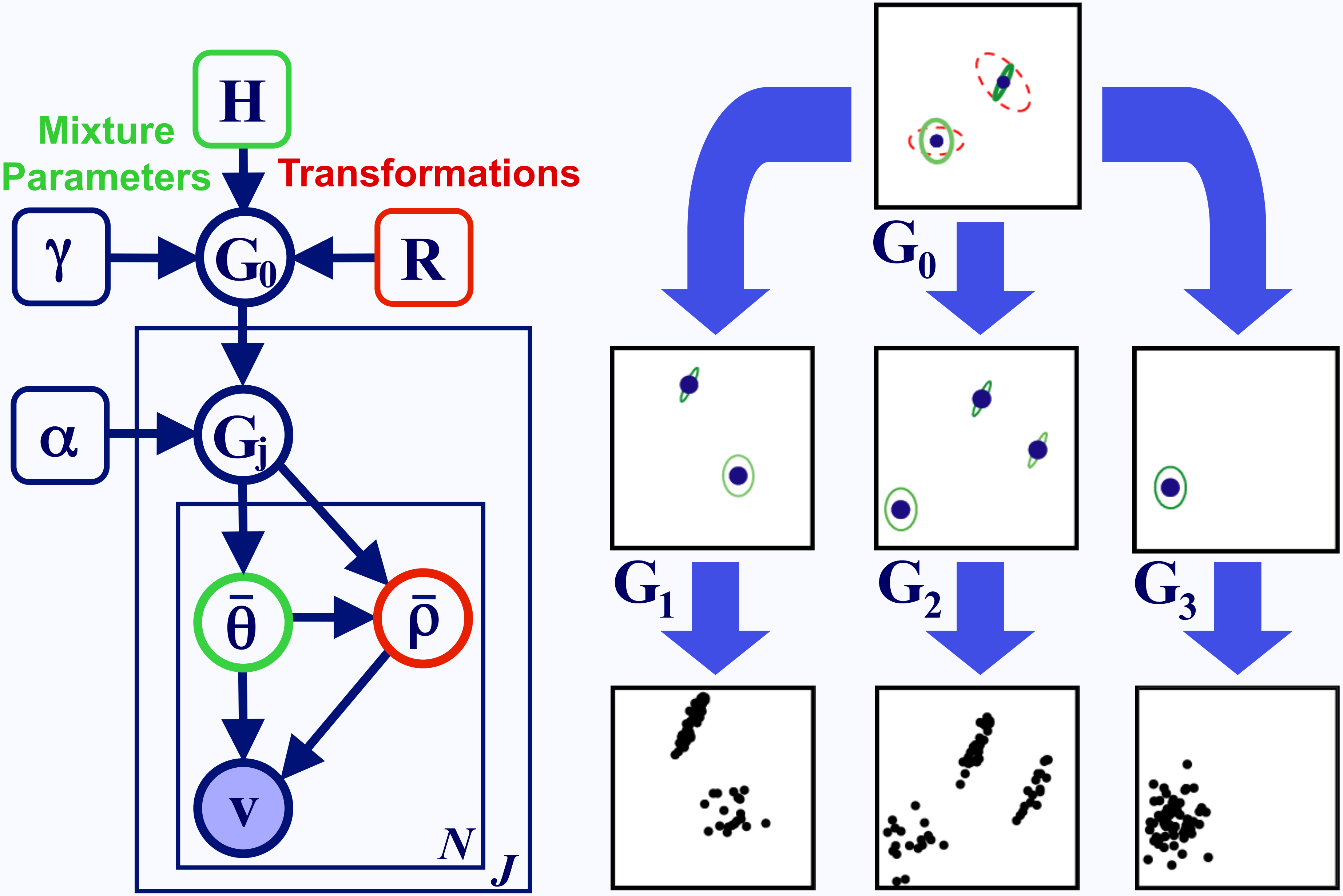


Layered Motion Models *(Wang & Adelson, Jojic & Frey)*
Nonparametric Transformation Densities *(Learned-Miller & Viola)*

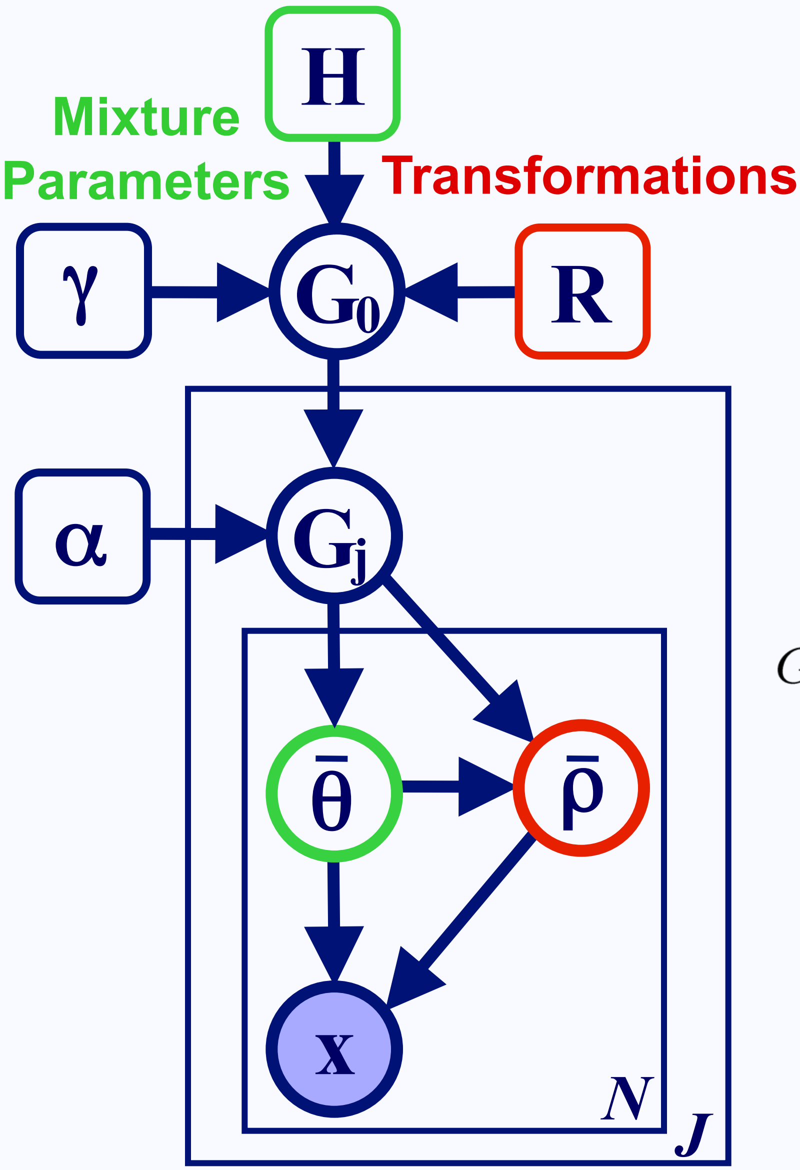
A Toy World: Bars & Blobs



Transformed Dirichlet Process



Transformed Dirichlet Process



Global mixture over parameters & transformations (translations):

$$G_0(\theta, \rho) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k) q\left(\rho \mid \phi_k\right)$$

$$\boldsymbol{\beta} \sim \mathrm{Stick}(\gamma) \qquad \theta_k \sim H \qquad \phi_k \sim R$$

Images generated from a set of transformed global densities:

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

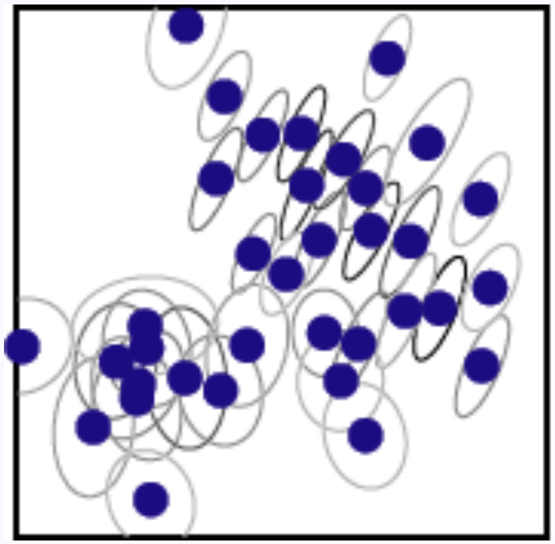
$$G_j(\theta, \rho) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k) \underbrace{\left[\sum_{\ell=1}^{\infty} \omega_{jk\ell} \delta(\rho, \rho_{jk\ell}) \right]}_{\boldsymbol{\omega}_{jk} \sim \mathrm{Stick}(\alpha \beta_k)}$$

Sample each feature independently:

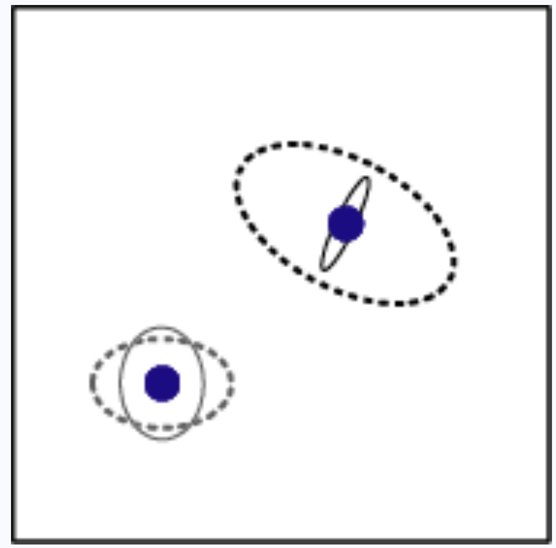
$$(\bar{\theta}_{ji}, \bar{\rho}_{ji}) \sim G_j(\theta, \rho)$$

$$x_{ji} \sim f\left(x \mid \tau(\bar{\theta}_{ji}; \bar{\rho}_{ji})\right)$$

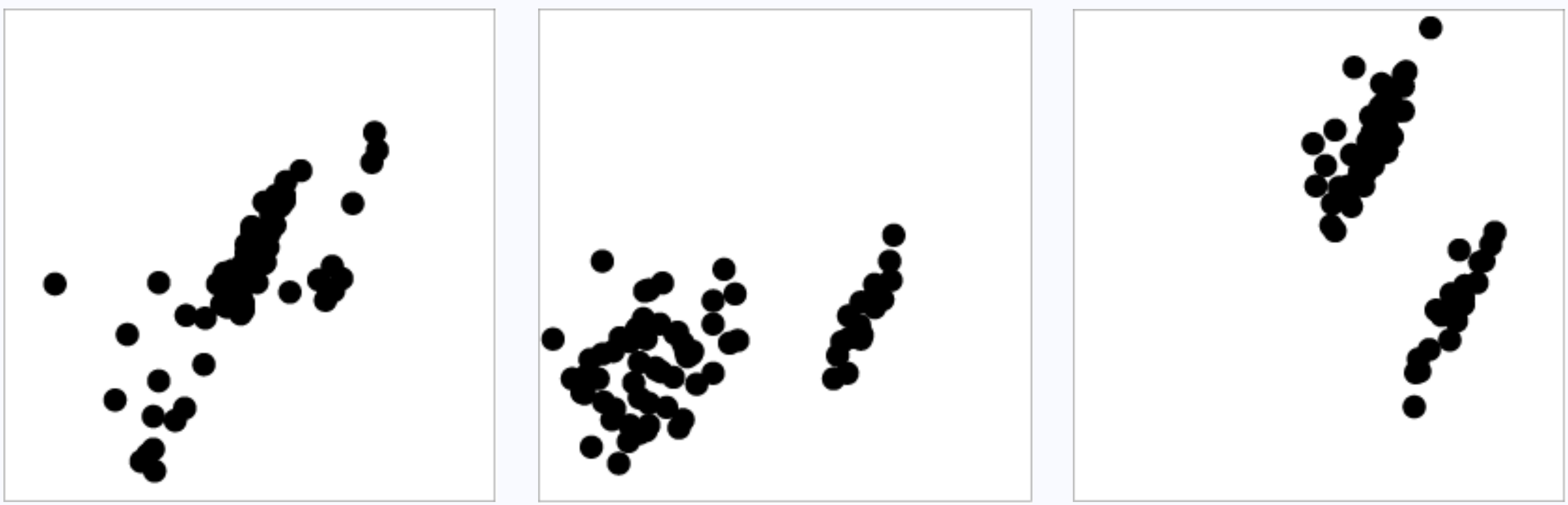
Importance of Transformations



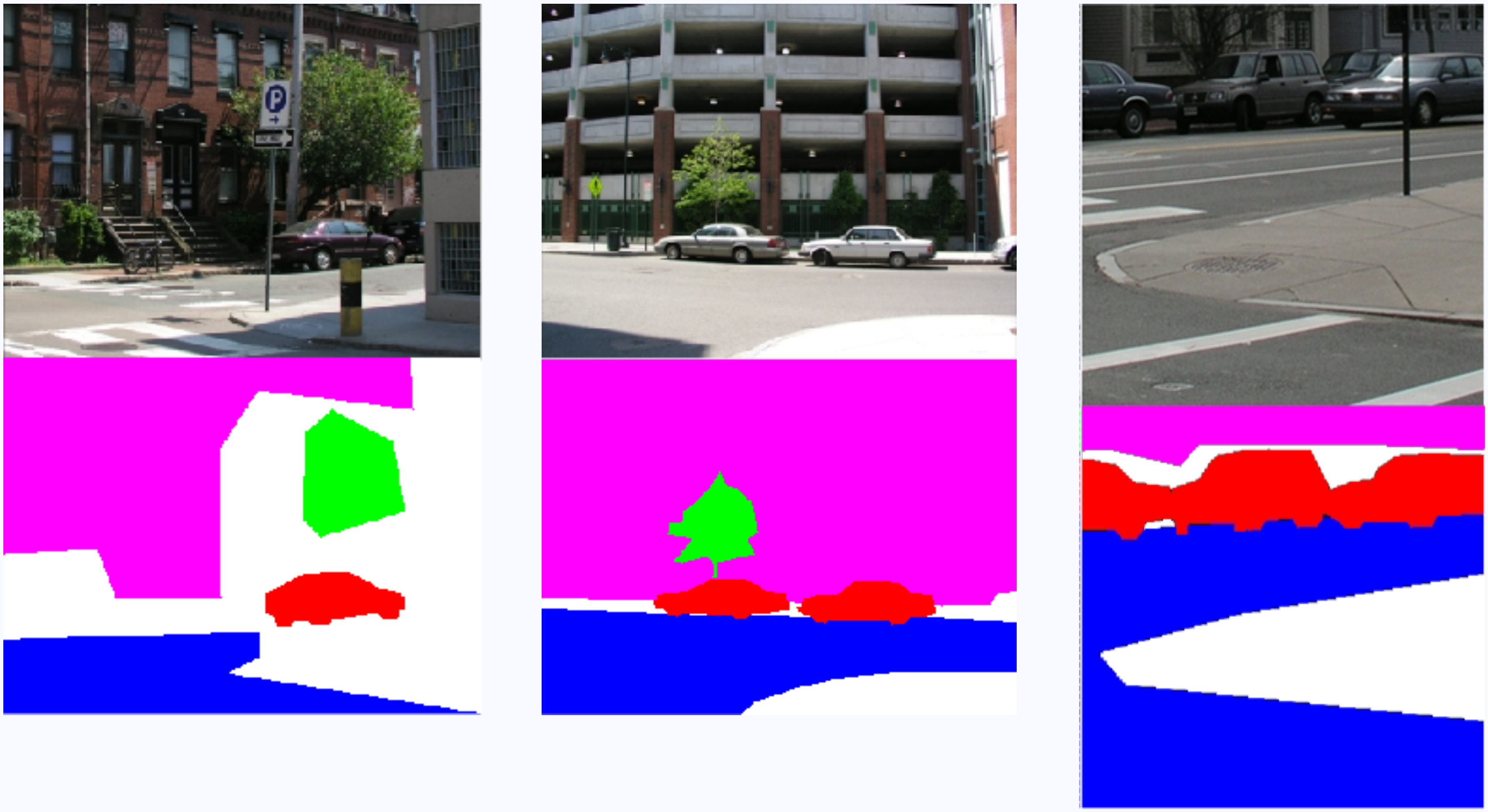
HDP



TDP

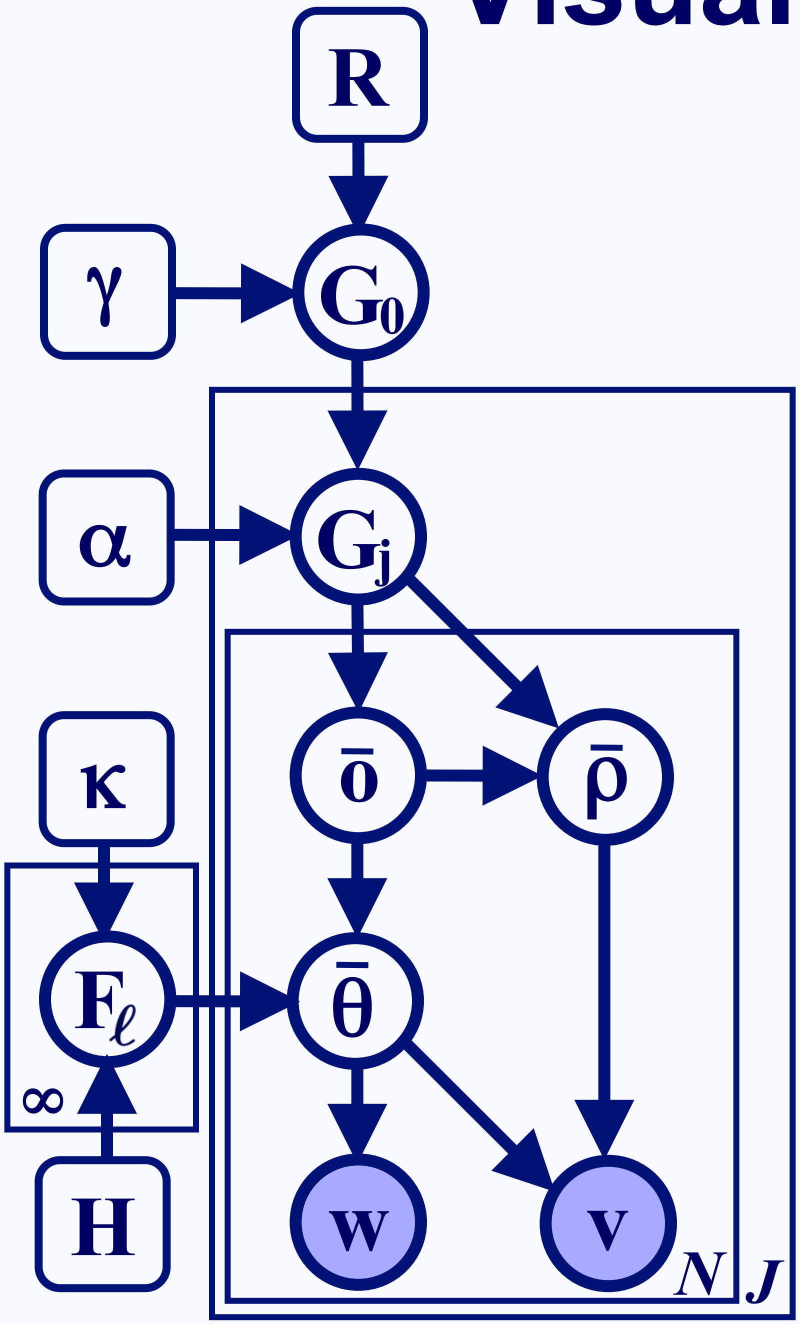


Counting & Locating Objects

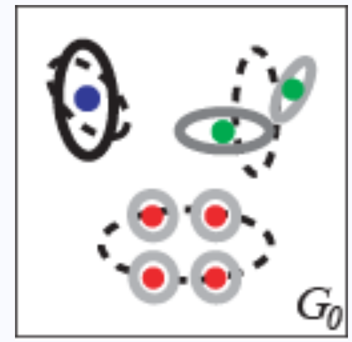


- How many cars are there? *Dirichlet Processes*
- Where are those cars in the scene? *Transformations*

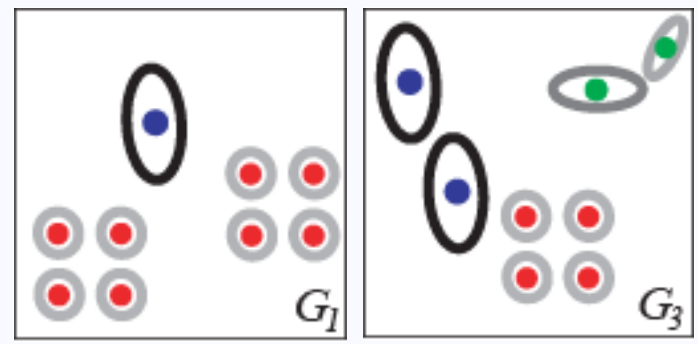
Visual Scene TDP



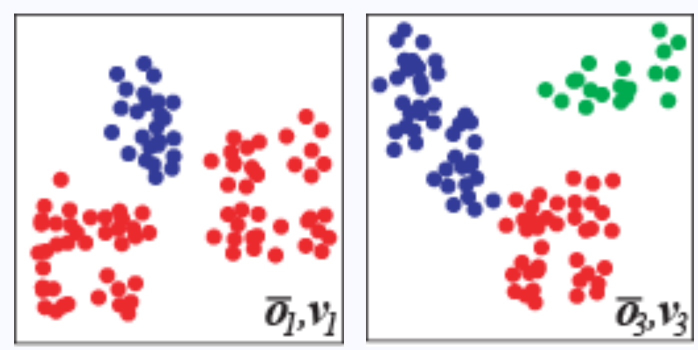
Global Density
Object category
Part size & shape
Transformation prior



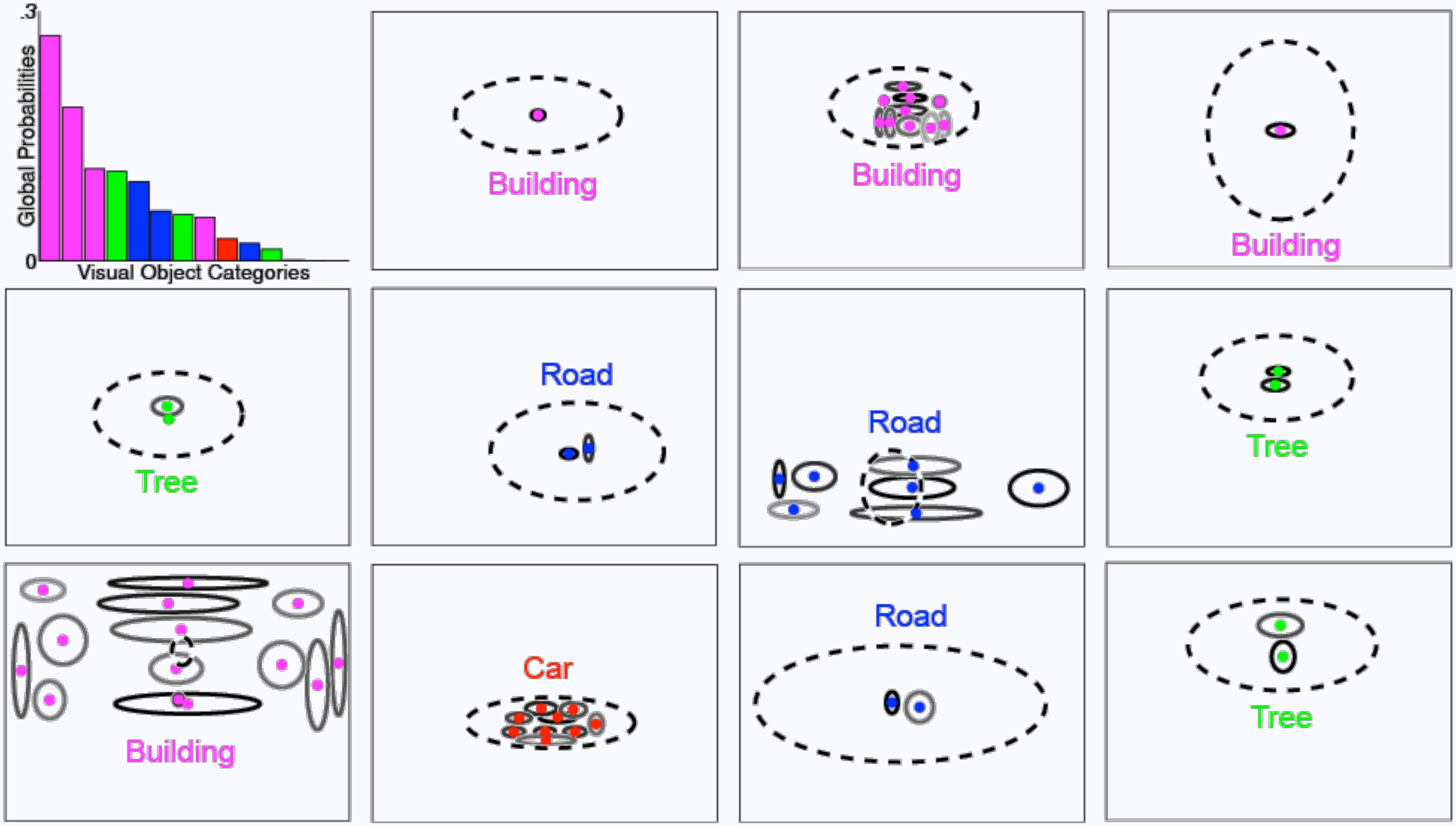
Transformed Densities
Object category
Part size & shape
Instance locations



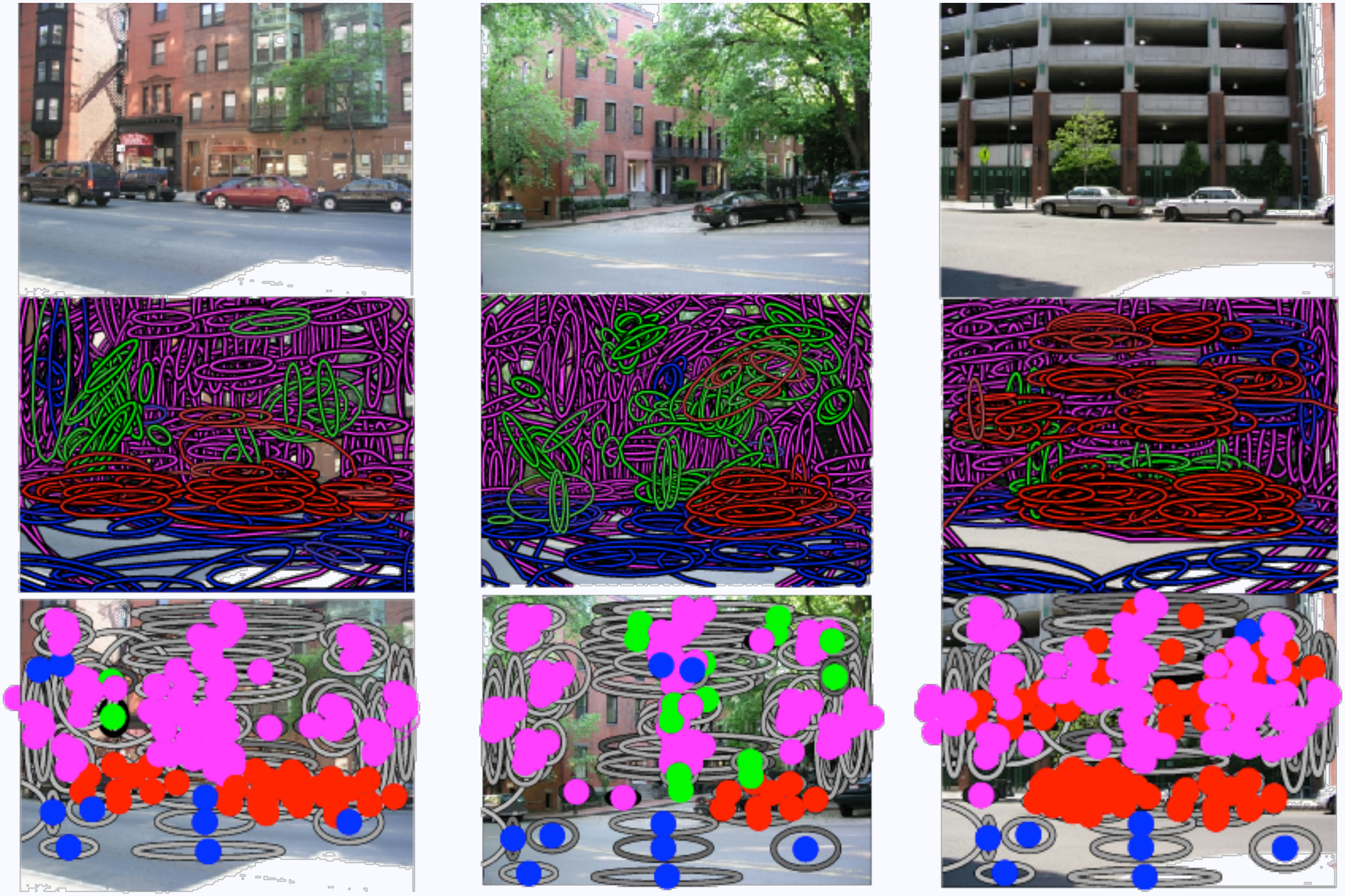
2D Image Features
Appearance
Location



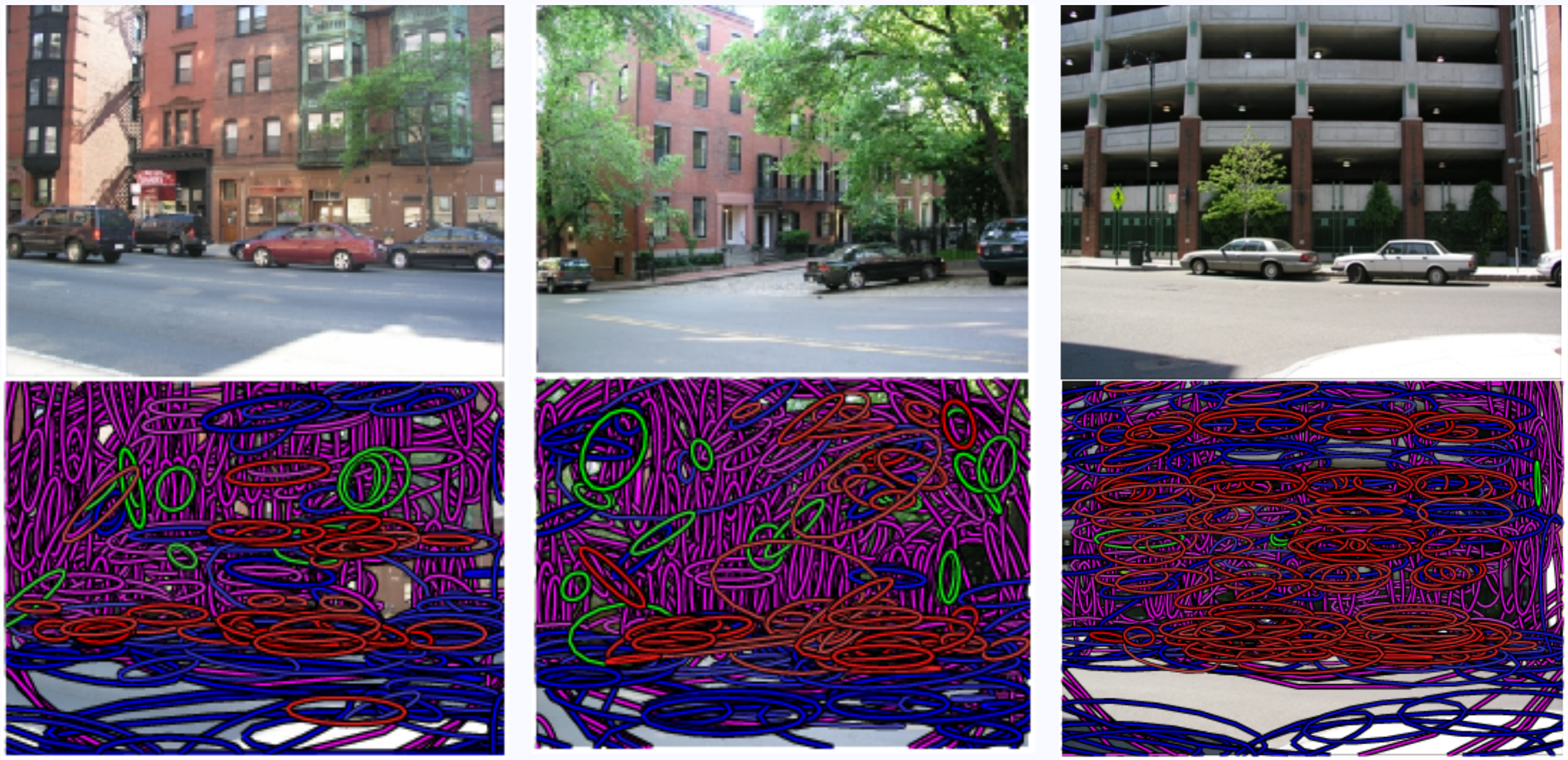
Street Scene Visual Categories



Street Scene Segmentations



Appearance Only



- “Bag of features” model, ignores feature positions
- Inferior segmentations, cannot count objects

Segmentation Performance