

# Multi-Modal Hierarchical Dirichlet Process Model

Predicting Image Annotation and Image-Object Label  
Correspondence

---

*Oksana Yakhnenko* and Vasant Honavar  
Iowa State University



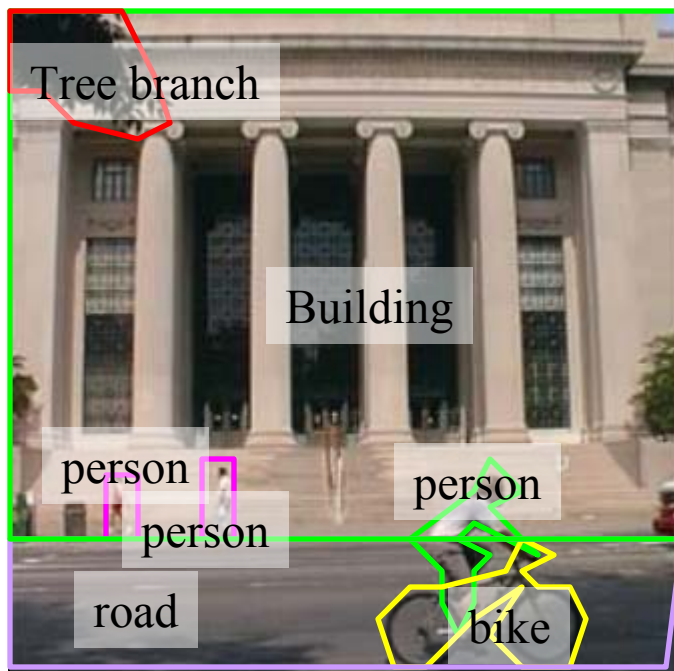
# Outline

---

- Image annotation and object recognition
- Modeling images and text using correlations
  - Probabilistic correlation: Multi-Modal Hierarchical Dirichlet Process
  - Linear correlation: Kernel Multiple Linear Regression (KMLR)
  - Automatic Kernel Selection for MLR
- Ongoing and future work
- Conclusion

# Image/object classification

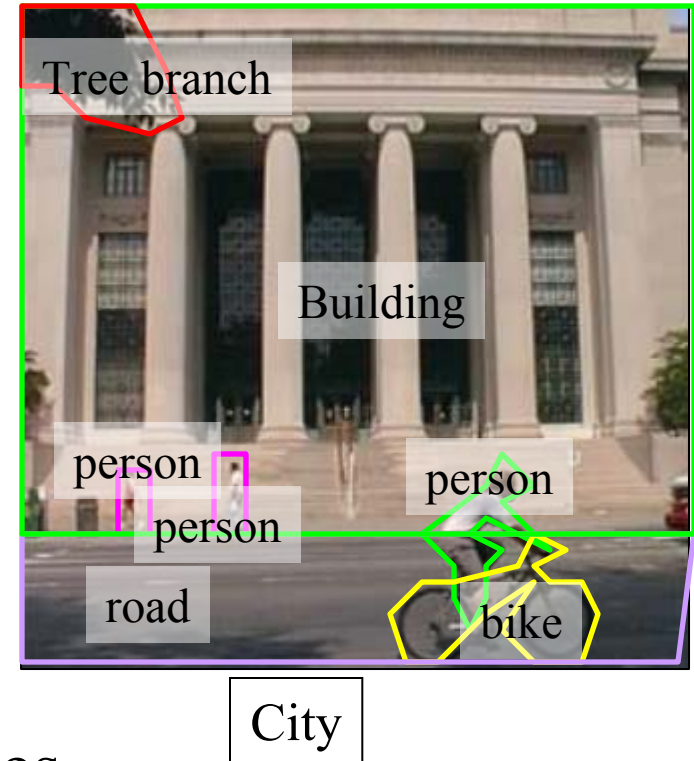
---



City

# Holistic scene understanding

- Interested to know:
  - What the image is about
  - What the objects in the image are
  - What the classes of the objects are
  - What the interactions and relations between the objects are



Reducing to classification task requires labeled data



---

Problem: labeled data is expensive!

Took me 10 minutes to label 1 image for animation in the previous slide... Can you imagine having to label more?

How can we take advantage of available data?

# Tagged images

---



Road, car, stoplight,  
sidewalk, person,  
people, sky, clouds,  
buildings, clock-tower

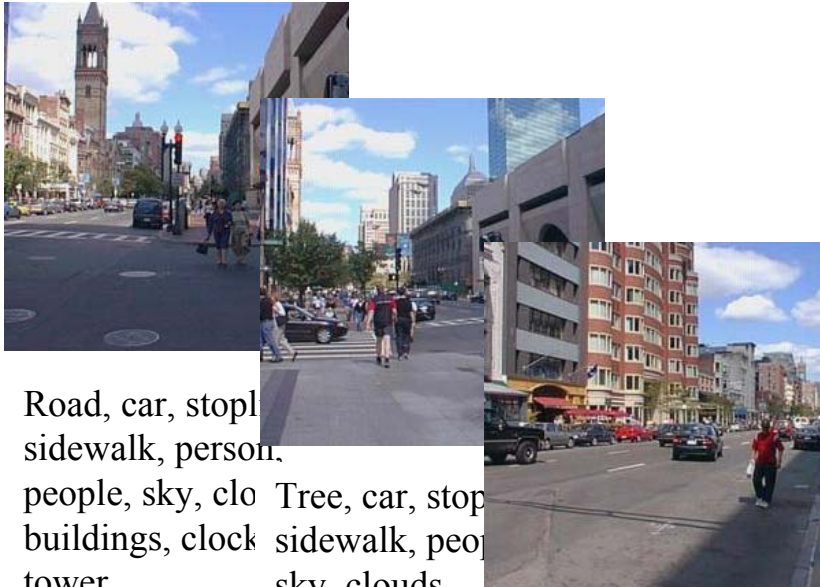


Tree, car, stoplight,  
sidewalk, people, sky,  
clouds, building,  
skyscraper



Cars, road, person,  
cloudy sky, buildings

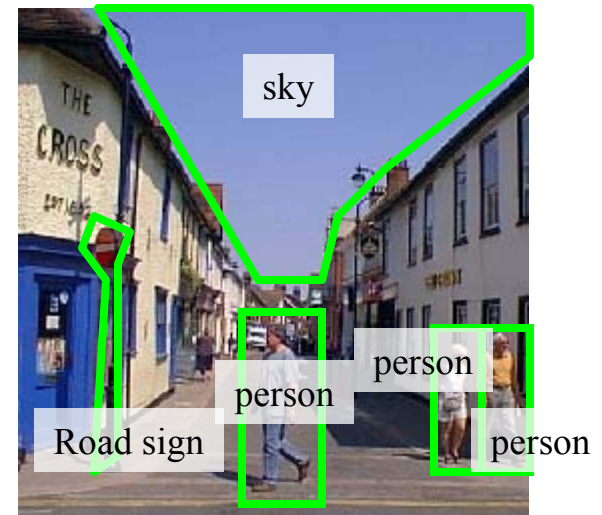
# Object recognition in “weakly-labeled” images



Road, car, stopl  
sidewalk, person,  
people, sky, clo  
buildings, clock  
tower

Tree, car, stop  
sidewalk, peo  
sky, clouds,  
building, skys

Cars, road person,  
cloudy sky,  
buildings





# Problems:

---

- Image annotation
  
- Object recognition using image annotations



# Outline

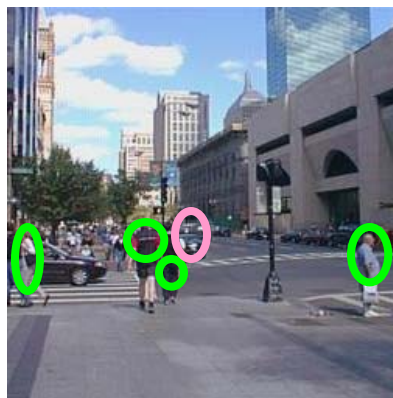
---

- Image annotation and object recognition
- **Modeling images and text using correlations**
  - Probabilistic correlation: Multi-Modal Hierarchical Dirichlet Process
  - Linear correlation: Kernel Multiple Linear Regression (KMLR)
  - Automatic Kernel Selection for MLR
- Ongoing and future work
- Conclusion

# Learning by correlating



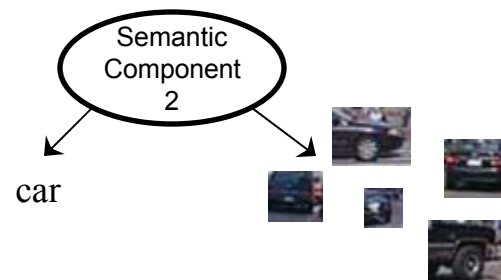
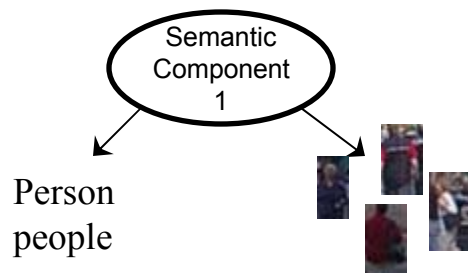
Road, car, stoplight,  
sidewalk, person,  
people, sky, clouds,  
buildings, clock-  
tower



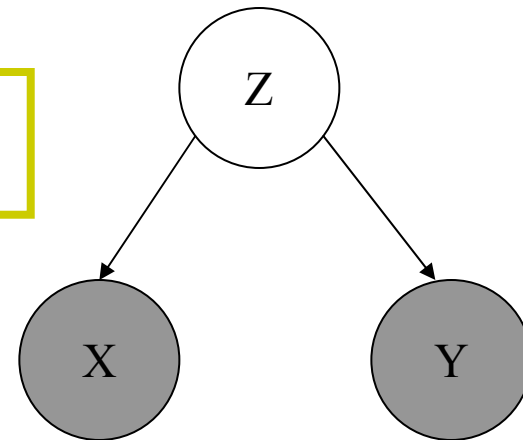
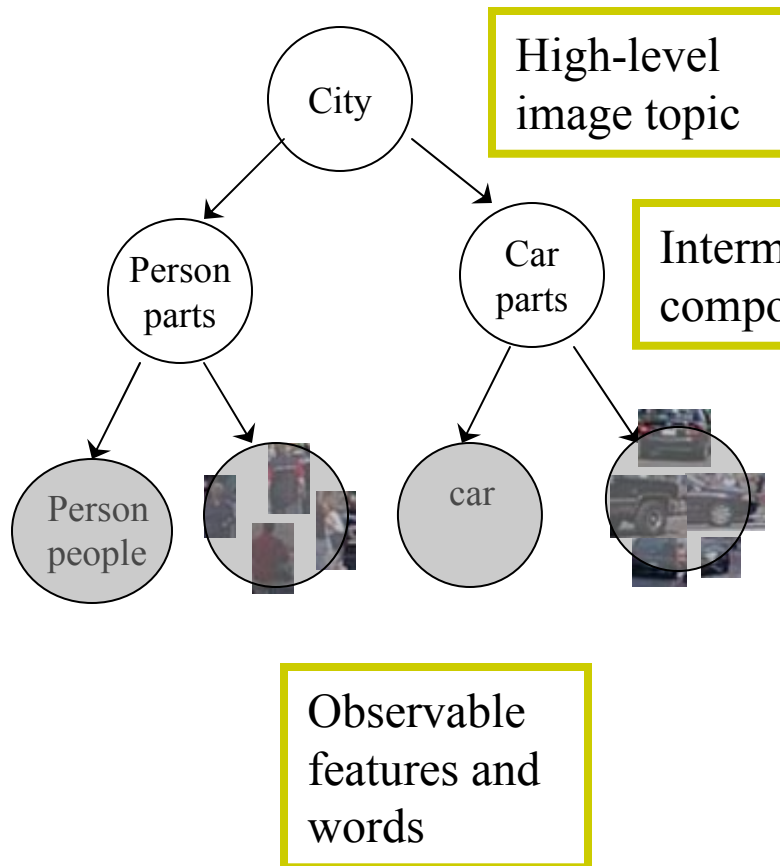
Tree, car, stoplight,  
sidewalk, people,  
sky, clouds,  
building, skyscraper



Car, road, person,  
cloudy sky,  
buildings

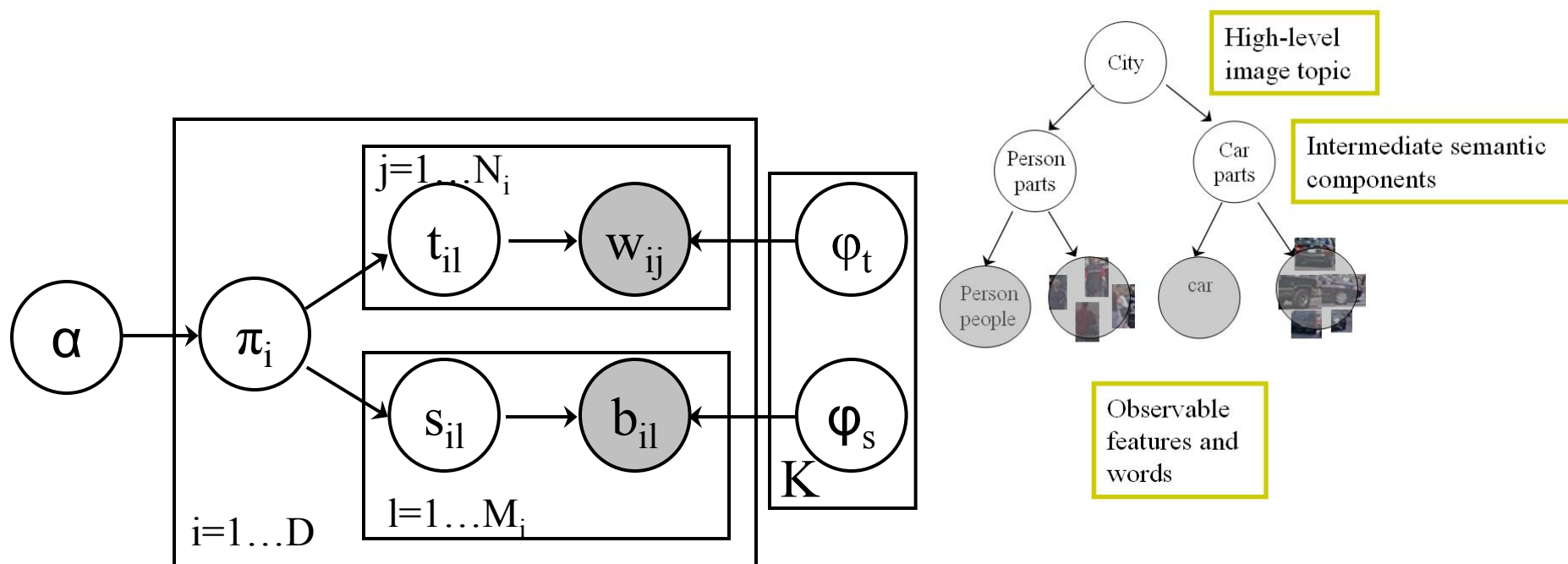


# Probabilistic correlation for images and text



Probabilistic correlation for 2 views

# Multi-modal Latent Dirichlet Allocation



Need to know the number of mixture components in advance

# How many mixture components?

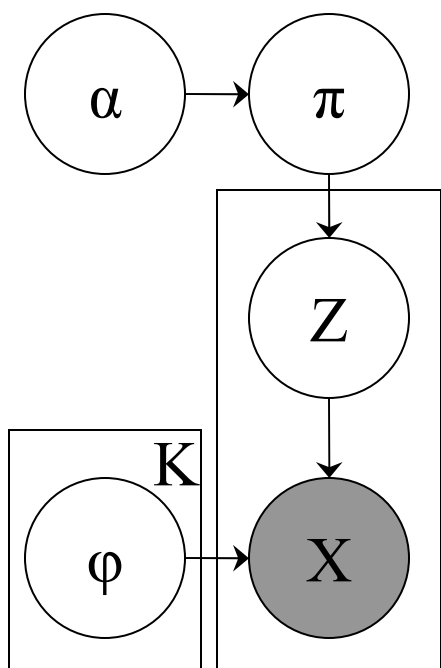
---

- The number of mixture components is fixed by the model
  - Overfitting
    - Maximum likelihood estimation can keep increasing likelihood of the data as number of mixture components increases
  - Model Selection
    - Generally, a separate model is trained for a given number
    - The number resulting in the best-performing model is chosen

# Solution: Dirichlet Process

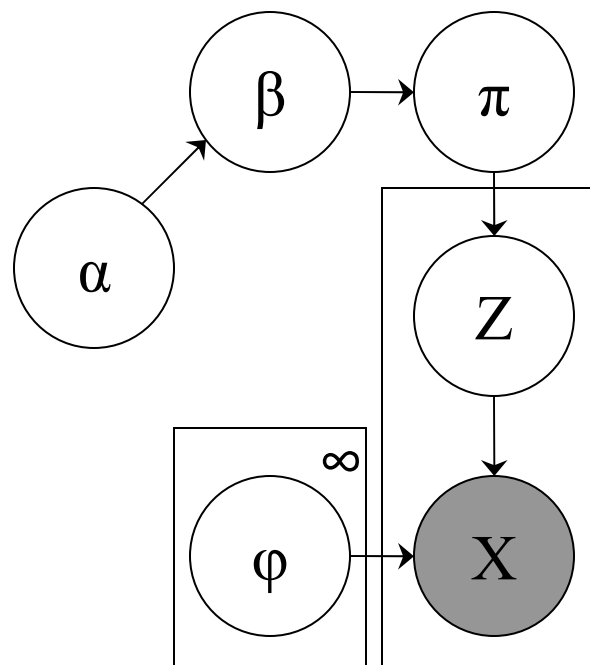
- Stochastic process that allows countably infinite number of mixture components

Finite mixture model



Fixed prior distribution  
 $\pi \sim \text{Dirichlet}(\alpha)$

Dirichlet Process



Stochastic process to generate prior  
 $\pi \sim \text{DP}(\beta)$

# DP using Stick-breaking construction

---

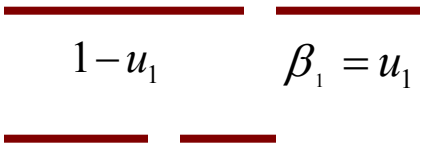
Goal: generative process for  $\beta_1 \beta_2 \dots$  so that  $\sum_{i=1}^{\infty} \beta_i = 1$

Assume a stick of unit length 

Generate proportions  $u_i \sim \text{Beta}(1, \alpha)$

Break the stick:

$$\beta_i = u_i \prod_{j=1}^{i-1} (1 - u_j)$$


$$\begin{array}{cc} \overline{\hspace{2cm}} & \overline{\hspace{1cm}} \\ 1 - u_1 & \beta_1 = u_1 \\ \overline{\hspace{2cm}} & \overline{\hspace{1cm}} \\ (1 - u_1)(1 - u_2) & \beta_2 = u_2 (1 - u_1) \end{array}$$

# DP using Stick-breaking construction

---

Goal: generative process for  $\beta_1 \beta_2 \dots$  so that  $\sum_{i=1}^{\infty} \beta_i = 1$

Assume a stick of unit length 

Generate proportions  $u_i \sim \text{Beta}(1, \alpha)$


**Notation:**  $\beta \sim \text{GEM}(\alpha)$

Break the stick:

**Use truncated Dirichlet process:**

$\beta_k = 0$  for  $k > K$

$$\beta_i = u_i \prod_{j=1}^{i-1} (1 - u_j)$$



$$(1 - u_1)(1 - u_2) \quad \beta_2 = u_2 (1 - u_1)$$

# Variational Inference

---

Given unknown distribution  $p$  approximate it with  $q$

$$q^*(\theta, z) = \arg \min_{q \in \mathcal{Q}} KL(q(\theta, z) \| p(\theta, z | x))$$

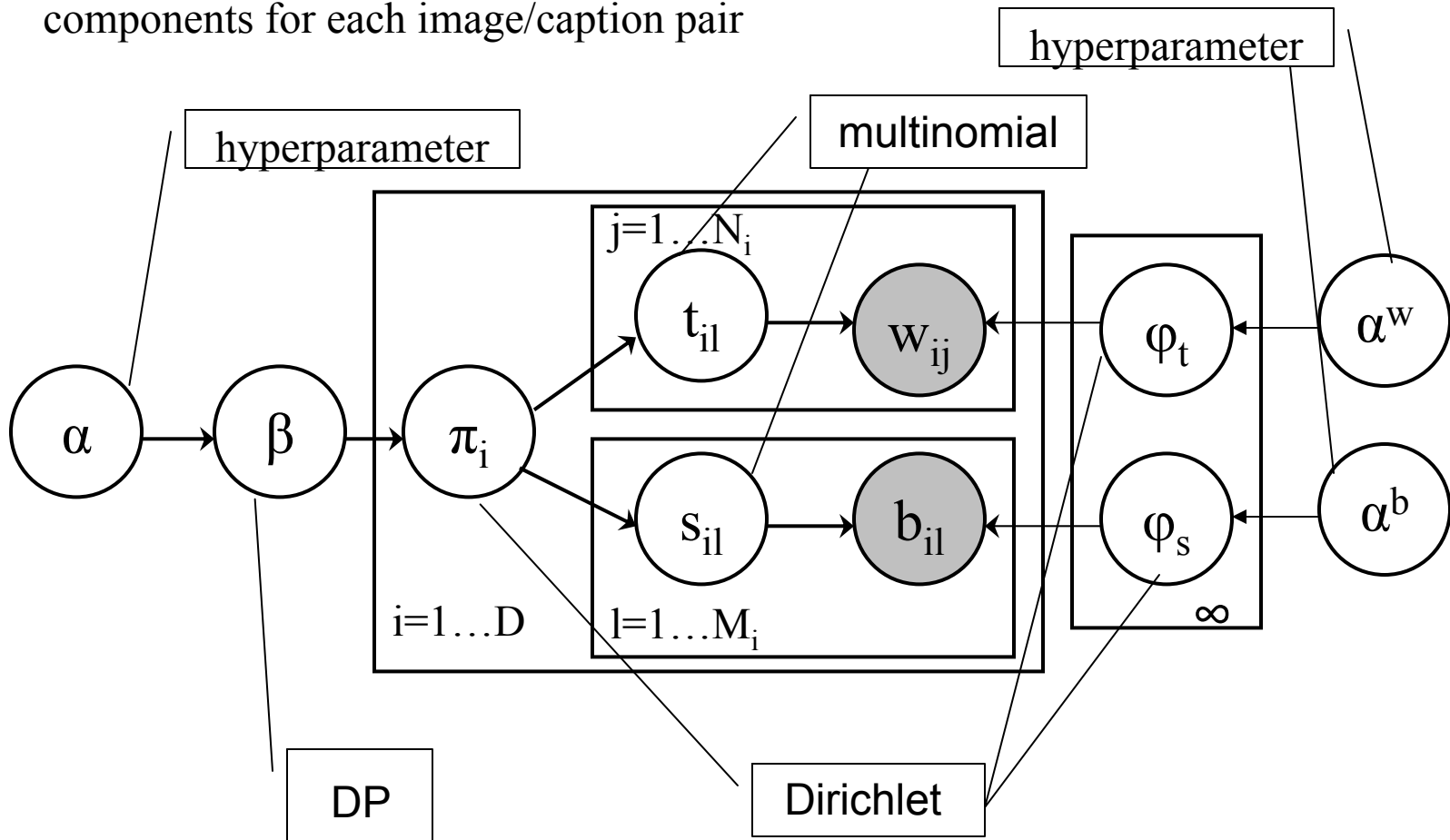
Assume fully factorized distribution:  $q(\theta) = \prod_{i=1}^n q_i(\theta_i)$

Solution to the minimization problem:  $q(\theta_i) \propto \exp(E_{q_{-i}} \log p(\theta_i | \theta_{-i}))$

Generalized EM:

# MoM-HDP: full model

Use HDP instead of DP to allow variable distribution of mixture components for each image/caption pair

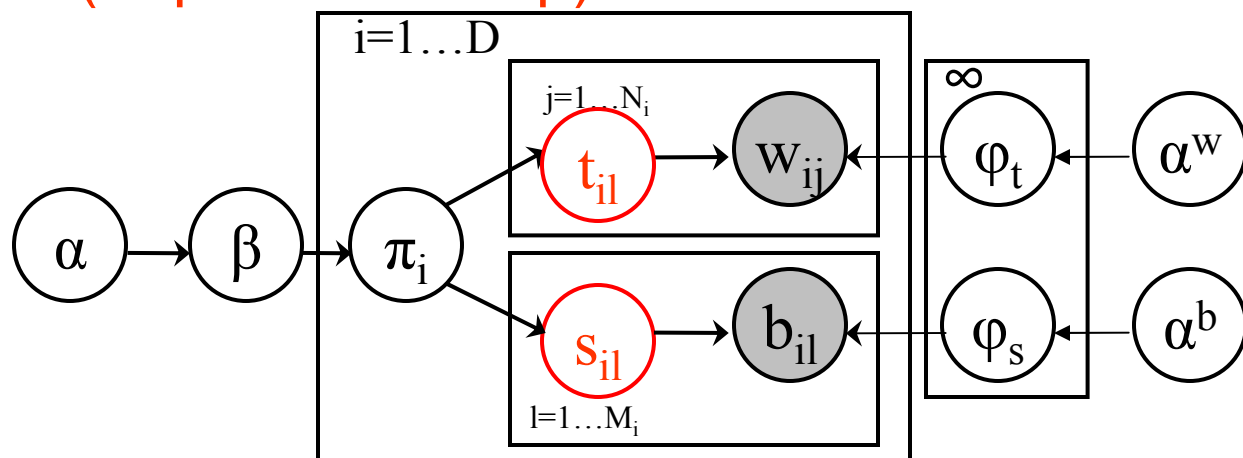


# Variational inference for MoM-HDP: generalized EM

## Multinomial update (Expectation step)

$$q(t_j) = \sum_{j=1}^{N_i} q(t_i, w_j)$$

$$q(t_i, w_j) = \exp(E_q \log \pi(i)) \exp(E_q \log \phi_{t_i}(w_j))$$



~~Use expected counts  
for general EM~~

Use DiGamma of expected counts  
for variational EM

Generate new expected counts  
using  $q(t)$ ,  $q(s)$

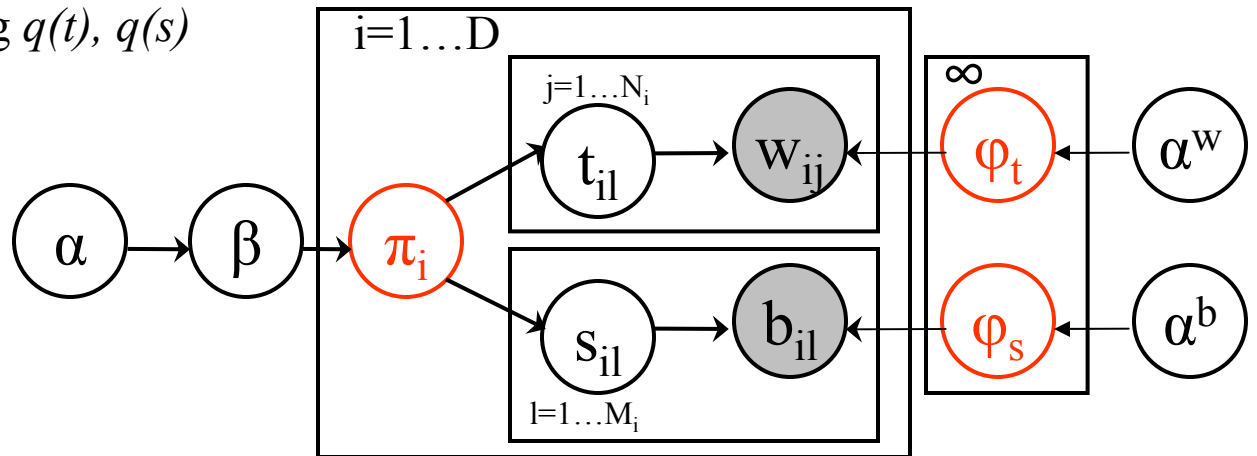
$$\exp(E_q \log \pi(i)) = \frac{\cancel{C(t_i)}}{\sum_k \cancel{C(t_k)}} \frac{\exp \Psi(\beta' + C(t_i))}{\exp \Psi(\sum (\beta' + C(t_k)))}$$

$$\exp(E_q \log \phi_{t_i}(w_j)) = \frac{\cancel{C_t(w_j)}}{\sum_k \cancel{C_t(w_k)}} \frac{\exp \Psi(\alpha_w' + C_{t_i}(w_j))}{\exp \Psi(\sum \alpha_w' + C_{t_i}(w_k))}$$

# Variational inference for MoM-HDP: generalized EM

## Dirichlet update (Maximization step)

Get expected counts using  $q(t)$ ,  $q(s)$



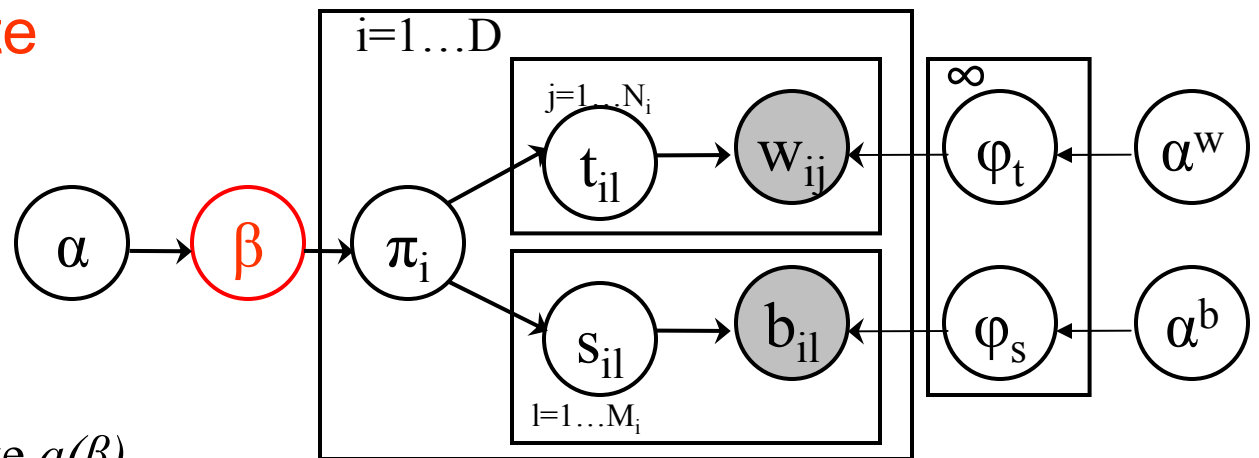
$$q(\pi_i) = \text{Dirichlet}(\alpha_\pi \beta + C(t_i) + C(s_i))$$

$$q(\phi_w^{t_i}) = \text{Dirichlet}(\alpha_w + C_{t_i}(w))$$

$$q(\phi_b^{s_i}) = \text{Dirichlet}(\alpha_b + C_{s_i}(b))$$

# Variational inference for MoM-HDP: generalized EM

## High-level update



Goal: maximize  $q(\beta)$

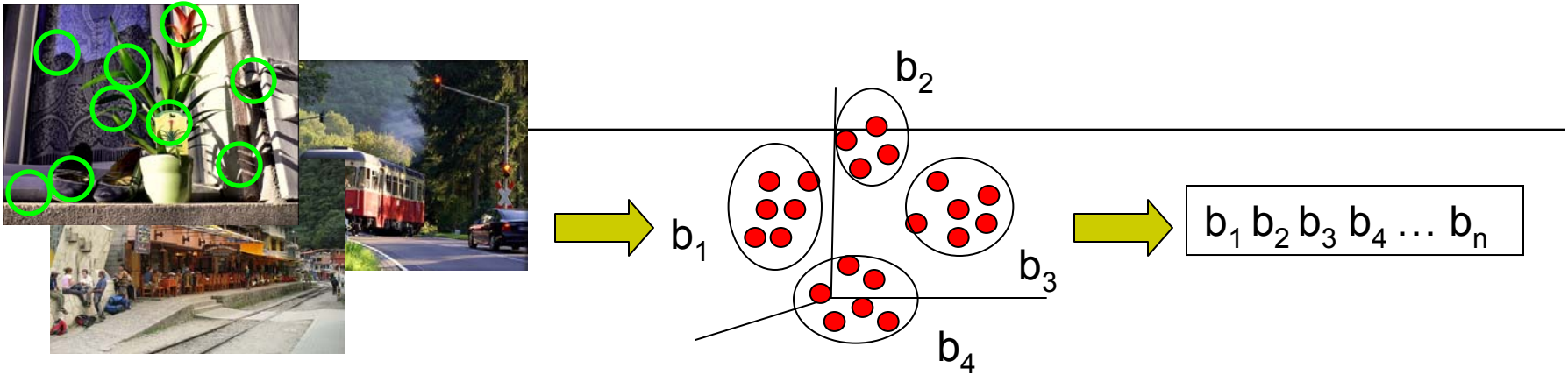
$$q(\beta) = E_q GEM(\beta) + E_q DP(\pi | \alpha_\pi \beta)$$

Since we use truncated HDP:

$$q(\beta) = E_q GEM(\beta) + E_q Dirichlet(\pi | \alpha_\pi \beta)$$

No closed form solution, but use gradient ascent with Quadratic Penalty to satisfy  $\sum_{i=1}^K \beta_i = 1$

# Representing Images: Bag of Visual Words



Training images

Testing images



Train, car

$b_i b_{i+1} \dots b_k$   
 $w_i w_{i+1} \dots w_k$



Person, building

$b_j b_{j+1} \dots b_l$   
 $w_j w_{j+1} \dots w_s$

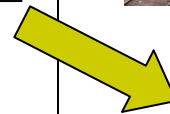
Model



$b_t b_{t+1} \dots$

?

$w$



# Data

---

## LabelMe:

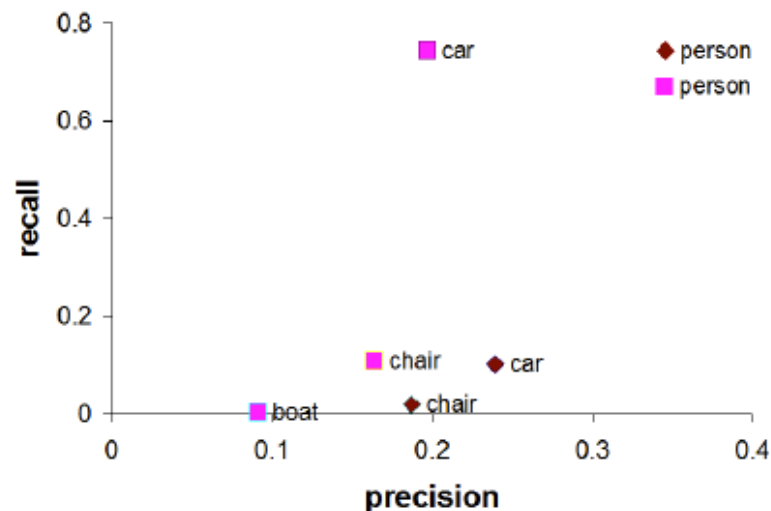
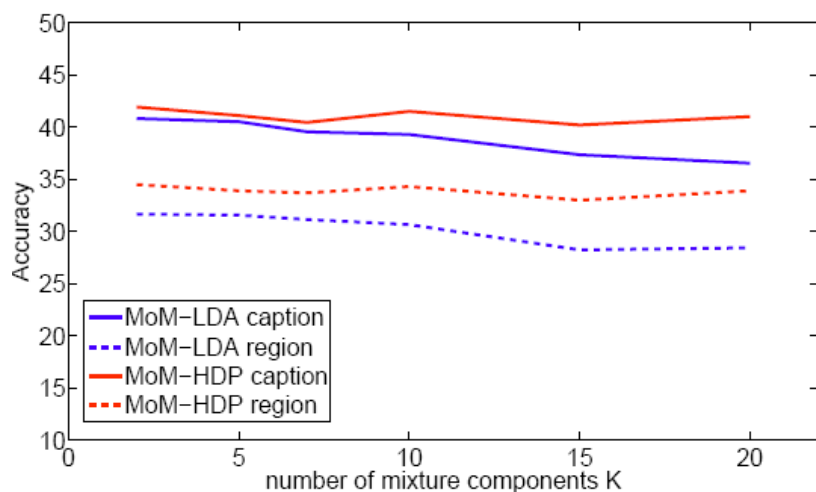
- Training set (7373),
- Test set (1513 images)
  - 14000 regions
- Between 4 and 19 objects
- On average each 10 words in caption
  - Lower cased and stemmed
  - 1700 distinct caption words
- Codebook of 1500 visual words

## VOC 2007:

- Training set: (2501)
- Test Set (4952)
  - ~15000 regions
- 20 possible caption words
- On average 4 words in caption
- Codebook of 1500 visual words

- Train the model on the image and caption information
- Test the model on the regions.

# Results for MoM-HDP (VOC 2009)



	per region	per caption
NB OneVsAll	30.56	38.03
LR OneVsAll	20.19	20.79
MoM-LDA	31.67	40.82
MoM-HDP	<b>34.5</b>	<b>41.92</b>
Chance prediction	5	5

# Results for MoM-HDP (LabelMe)

---

LabelME	image annot.	region annot.
MoM-LDA	15.56	10.5
MoM-HDP	34.84	<b>28.45</b>
NB OneVsAll	<b>38.2</b>	24.21



---

# Thank you

Email: [oksayakh@cs.iastate.edu](mailto:oksayakh@cs.iastate.edu)



# Selected References

---