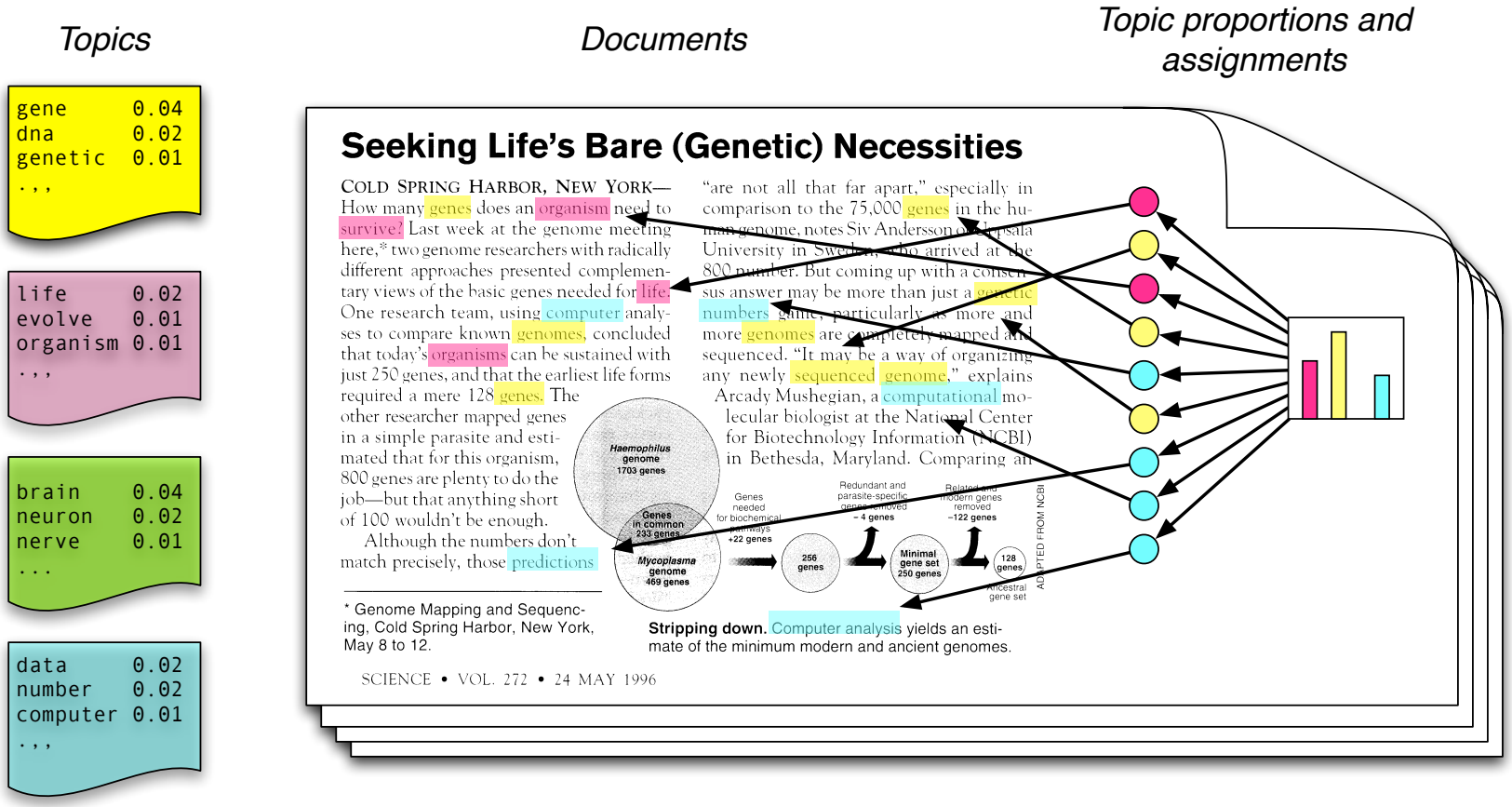
Topic Models for Texts and Images in Representation Space

Kui Tang and Sameer Lal

Columbia University

18 February 2015

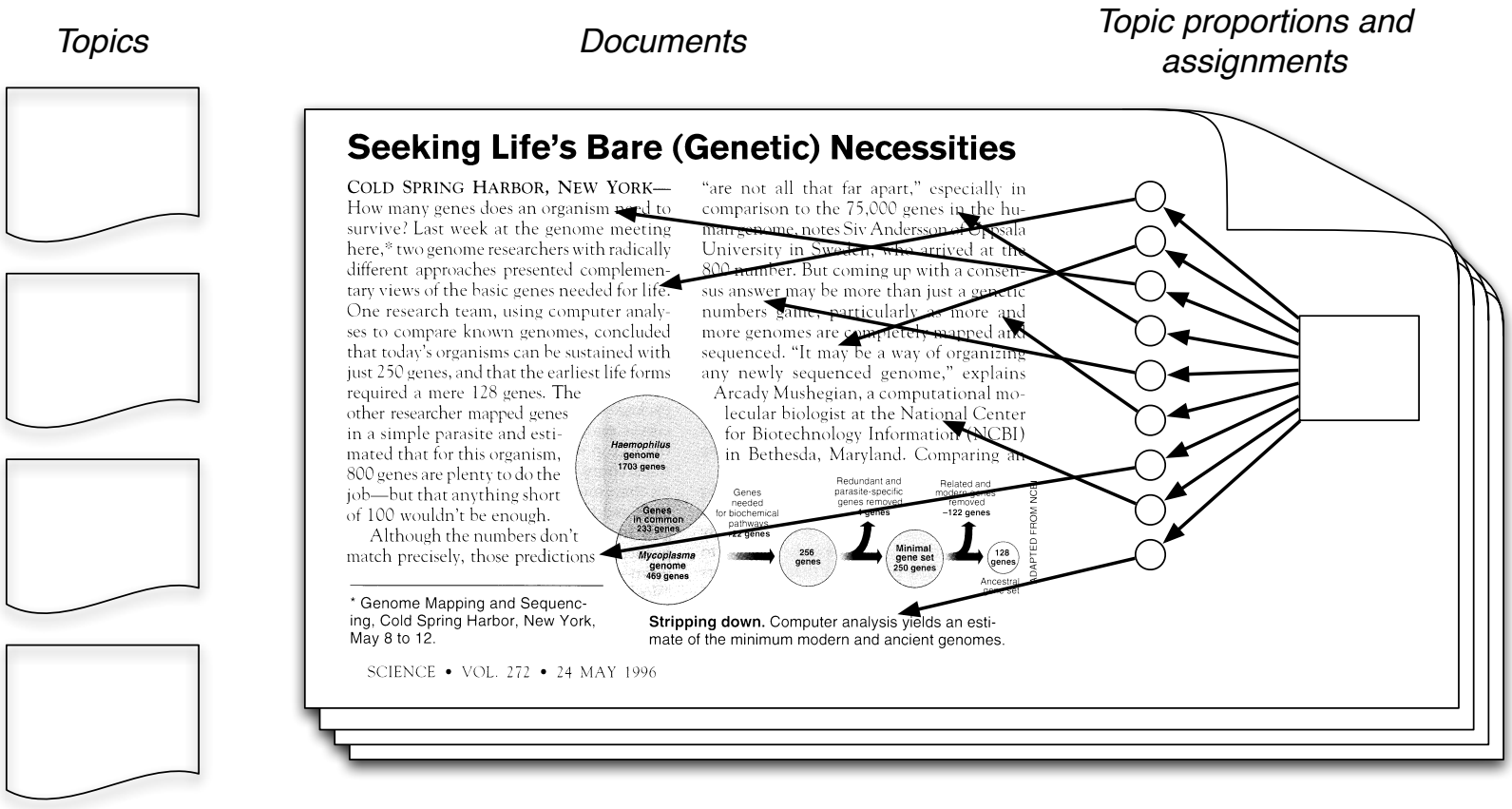
Topic Models



- LDA assumes that there are K topics shared by the collection.
- Each document exhibits the topics with different proportions.
- Each word is drawn from one topic.
- We discover the structure that best explain a corpus.

Slide stolen from D. Blei.

Latent Variable Models

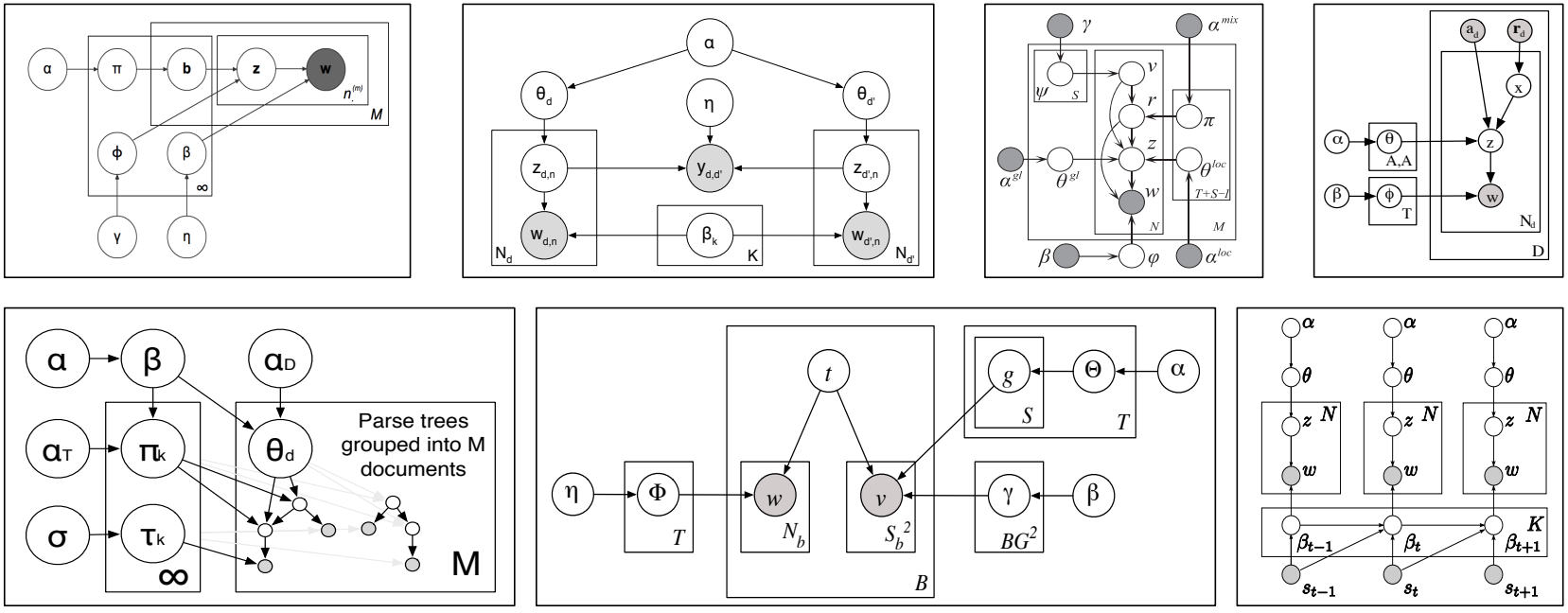


- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Slide stolen from D. Blei.

Bayesian Networks



Slide stolen from D. Blei.

- Shaded variables are *observed*, other variables are *hidden*.
- A model is our hypothesis for how data are generated.
- We *condition* on observations to update our hypothesis.

Multimodal Documents

A milking machine in action

Typical cattle yard in Northern Iowa, USA

farms also grow their own feed, typically including corn, alfalfa, and hay. This is fed directly to the cows, or stored as silage for use during the winter season. Additional dietary supplements are added to the feed to improve milk production. [10]

4.2 Poultry farms

Poultry farms are devoted to raising chickens (egg layers or broilers), turkeys, ducks, and other fowl, generally for

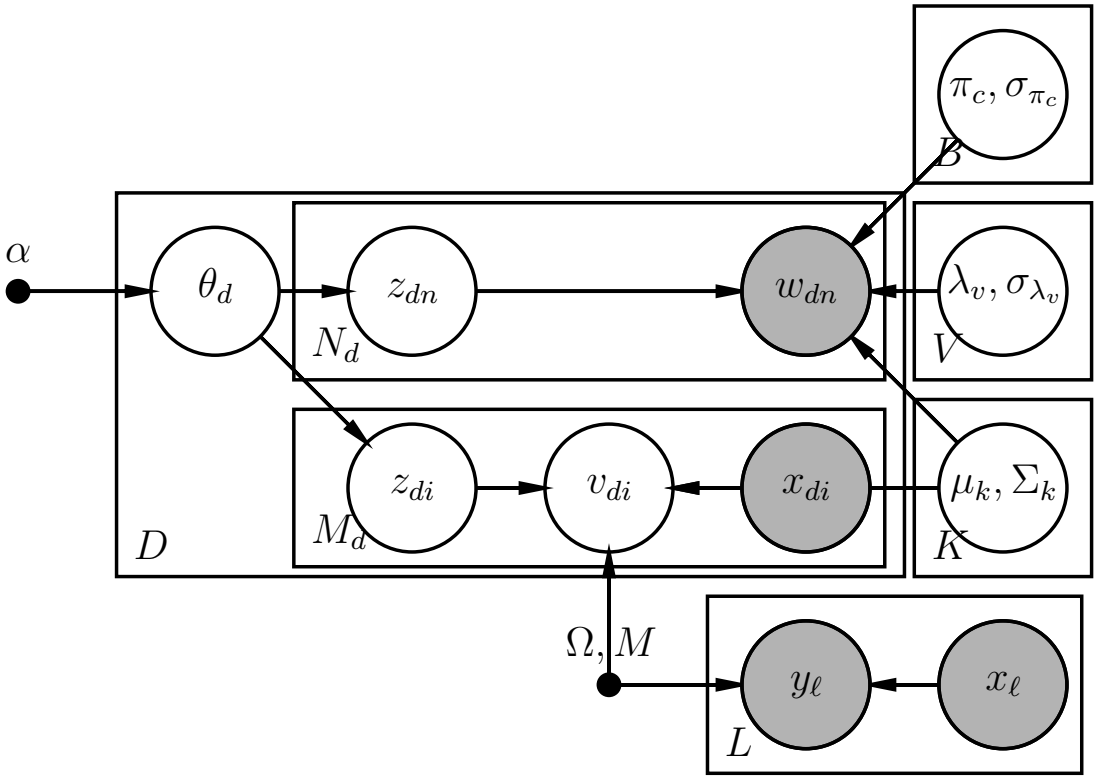
accompanied by the decoupling of political power from farm ownership.

5.1 Forms of ownership

In some societies (especially socialist and communist), collective farming is the norm, with either government ownership of the land or common ownership by a local group. Especially in societies without widespread industrialized farming, tenant farming and sharecropping are common; farmers either pay landowners for the right to use farmland or give up a portion of the crops.

- *We want to learn a topic model using text and images jointly.*
- Images and text complement each other.
- Captions aren't the whole story: cows in political contexts.

Gaussian Topic Models with CNNs



- Topics are (mixtures of) Gaussians.
- Words are latent vectors $\lambda_v \in \mathbb{R}^{D_W}$ using Bayesian word2vec.
- Images are latent vectors $v_{in} \in \mathbb{R}^{D_I}$ *conditioned* on raw images x_{di}. We have $v_{ni} \sim \mathcal{N}(M\text{CNN}_x(x_{ni}; \Omega), \Sigma)$ with Ω CNN parameters, M mapping to word vector space, and CNN_x feature representation output by CNN.

Variational Bayesian EM

To learn latent variable models, maximize the *marginal likelihood*

$$\max_{\theta} p(\mathbf{x}|\theta) = \int p(\mathbf{x}, \mathbf{z}|\theta) p(\mathbf{z}|\theta) d\mathbf{z}$$

This integral is intractable. Approximate instead with the *evidence lower bound (ELBO)*

$$\log p(\mathbf{x}|\theta) \geq E_{q(z|\phi)} \left[\log p(\mathbf{x}, \mathbf{z}|\theta) - \log q(\mathbf{z}|\theta)\right] =: \mathcal{L}(\theta, \phi)$$

where $q(\mathbf{z}|\phi)$ is a simple *variational distribution* which approximates the posterior $p(\mathbf{z}|\mathbf{x}, \theta)$.

Variational Bayesian EM:

- E Step: Update $\phi^{(t+1)} \leftarrow \arg\max_{\phi} \mathcal{L}(\theta^{(t)}, \phi)$
- M Step: Update $\theta^{(t+1)} \leftarrow \arg\max_{\theta} \mathcal{L}(\theta, \phi^{(t)})$

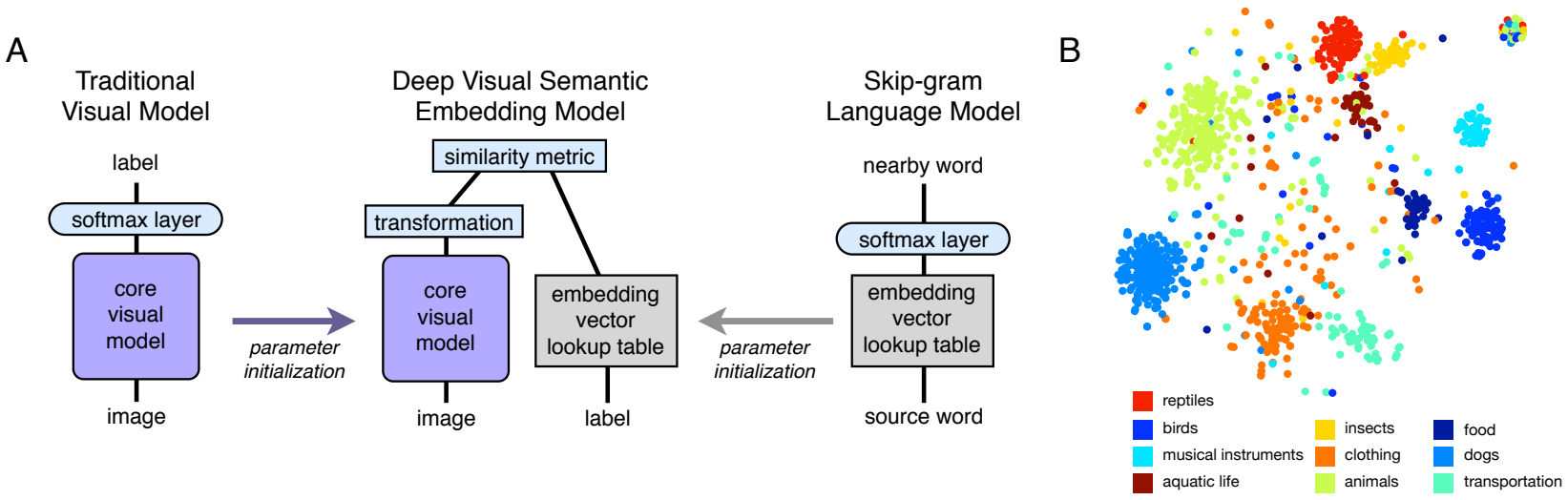
E step is variational Bayesian inference (Ranganath, Gerrish, and D. M. Blei, 2014; Wang and D. M. Blei, 2013). M step is learning (updating) a CNN with objective

$$\min_{\Omega} \sum_{\ell} L(y_\ell; \text{CNN}_y(x_\ell; \Omega)) + \frac{1}{2\sigma^2} \sum_{di} E_{q(v_{di}|\phi^{(t)})} \left[(v_{di} - \text{CNN}_x(x_{di}; \Omega))^2\right]$$

Why Do We Want to Do This?

- Constructing an unsupervised, non-discrimantive, model
- Difficult to measure performance (Wallach et al., 2009)
- Unspervised data can lead to better vector construction
- "...in general capture some distributional syntatic and semantic information." (Socher et al., 2013)
- Can this lead to a semantic understanding of multimodal data?

Related Methods



- Use Deep Boltzmann Machines (Srivastava and Salakhutdinov, 2014)
- Semi-supervised learning for effective generalization from smaller data sets (Kingma et al., 2014)
- Deep Visual-Semantic Embedding Model (DeViSE), use both unannotated text and trained image data for classification (Frome et al., 2013)

Implementation and Applications

- Image and text modalities: corpa with images, annotated images, images with captions
- Code EM algorithm with pretrained Caffe model and custom objective, compare with performing SGD on CNN and variational parameters jointly
- Explore other applications: modeling sub-topics, generalization to tangential classes, or image queries and search.
- Topic models also used for social networks and recommendation engines. Our method adds image features to those models.
- Compare with other algorithms

Thank You

- Questions?

References I

Frome, A. et al. (2013). "Devise: a deep visual-semantic embedding model". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. Burges et al. Curran Associates, Inc., pp. 2121–2129.

Kingma, D. P. et al. (2014). "Semi-supervised learning with deep generative models". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., pp. 3581–3589.

Ranganath, R., S. Gerrish, and D. M. Blei (Dec. 2014). "Black Box Variational Inference". In: *ArXiv e-prints*.

Socher, R. et al. (2013). "Reasoning with neural tensor networks for knowledge base completion". In: *Advances in Neural Information Processing Systems 26*. Ed. by C. Burges et al. Curran Associates, Inc., pp. 926–934.

References II

Srivastava, N. and R. Salakhutdinov (2014). "Multimodal learning with deep boltzmann machines". In: *Journal of Machine Learning Research* 15, pp. 2949–2980.

Wallach, H. M. et al. (2009). "Evaluation methods for topic models". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Quebec, Canada: ACM, pp. 1105–1112. DOI: 10.1145/1553374.1553515.

Wang, C. and D. M. Blei (2013). "Variational inference in nonconjugate models". In: *Journal of Machine Learning Research* 14.1, pp. 1005–1031.