FROM LDA TO VISION VIA POPULATION STRUCTURE

Sharmanska V., Lampert C.H. - IST Austria
{viktoriia.sharmanska, chl}@ist.ac.at

ICVSS 2011
Sicily ~ 11-16 July
International Computer Vision Summer School

Abstract

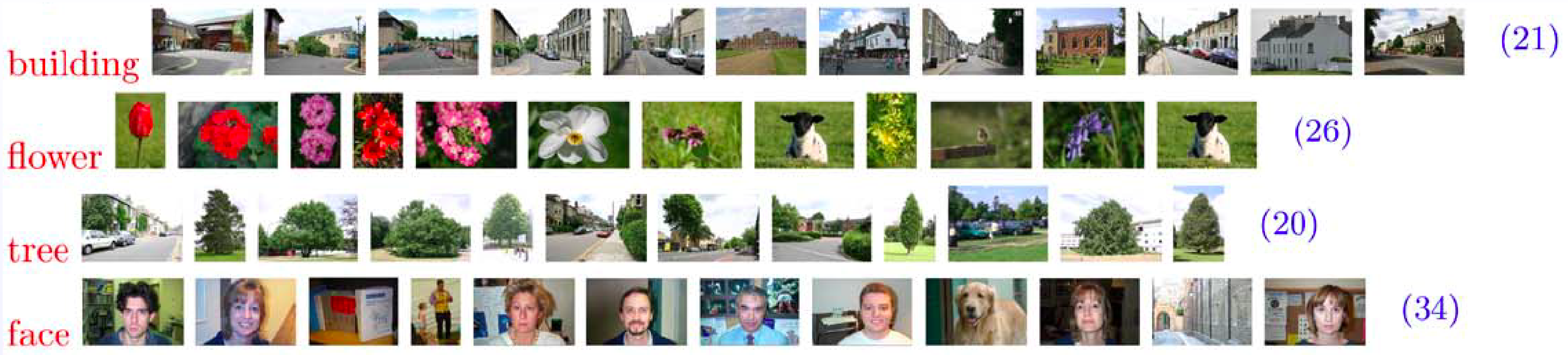
STRUCTURE is a model-based clustering method, which infers population structure and assigns individuals to populations; the model considers each individual as a mixture of a few source populations. Latent Dirichlet allocation (LDA) is a generative approach for topic modeling tasks in text processing; the model finds the thematic structure of a collection of documents. We show that these two models describe the same generative process.

Goal

LDA:
Infer (latent) topic structure of text documents [1]

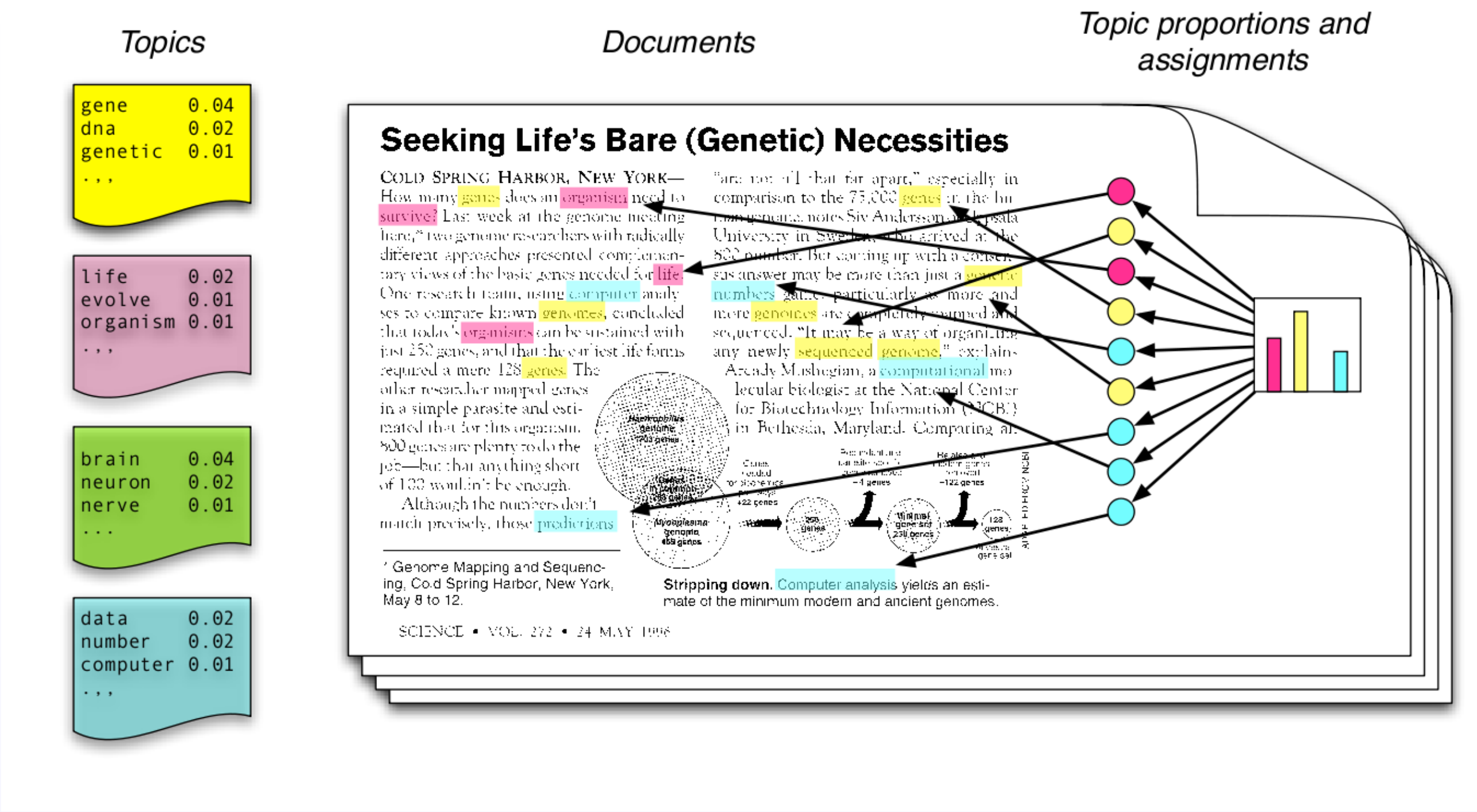
Population STRUCTURE:
Infer populations and understand how individuals originated from them based on their genotypes [2]

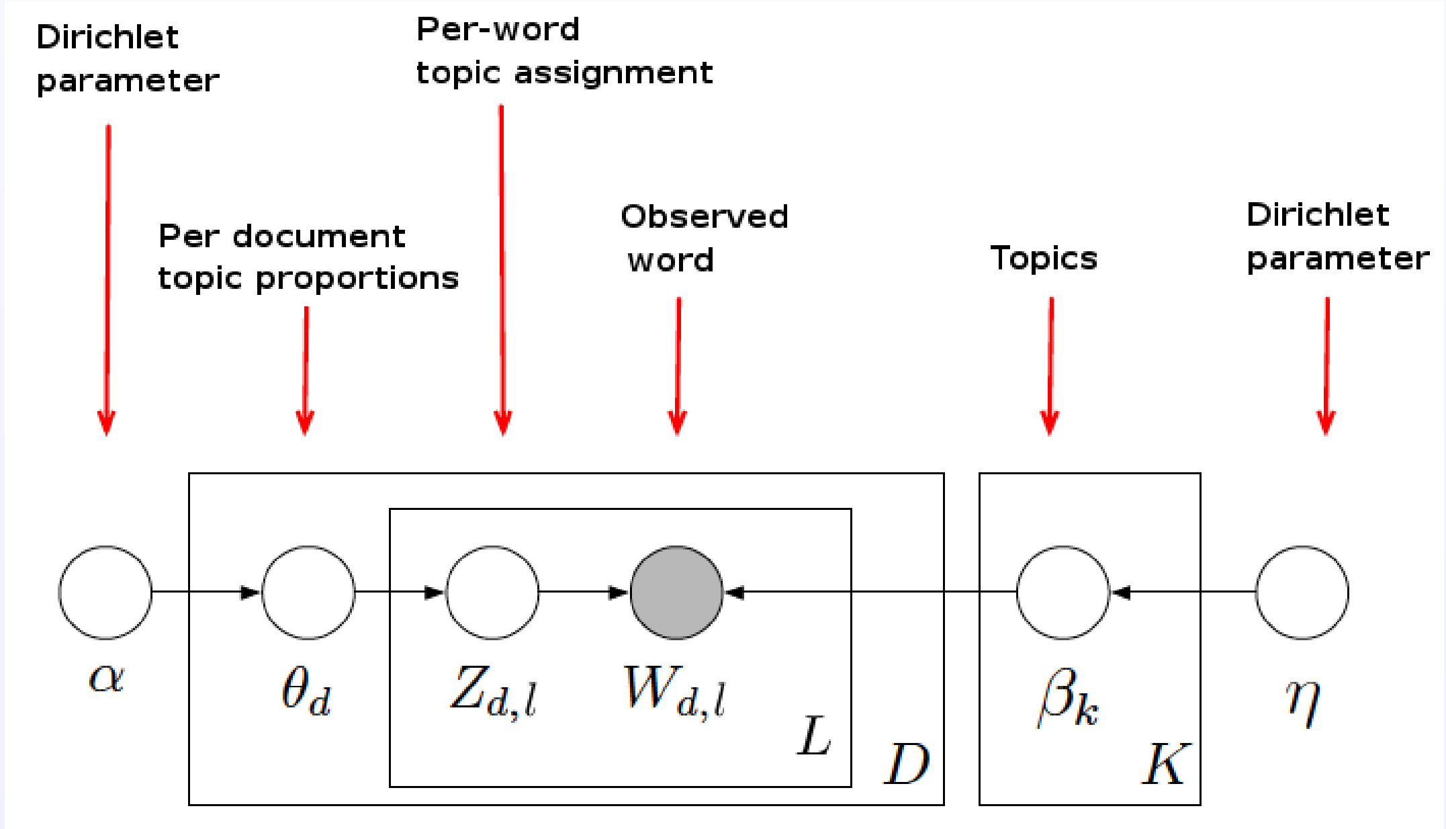
Computer Vision: Infer the object categories that images are composed of [3]



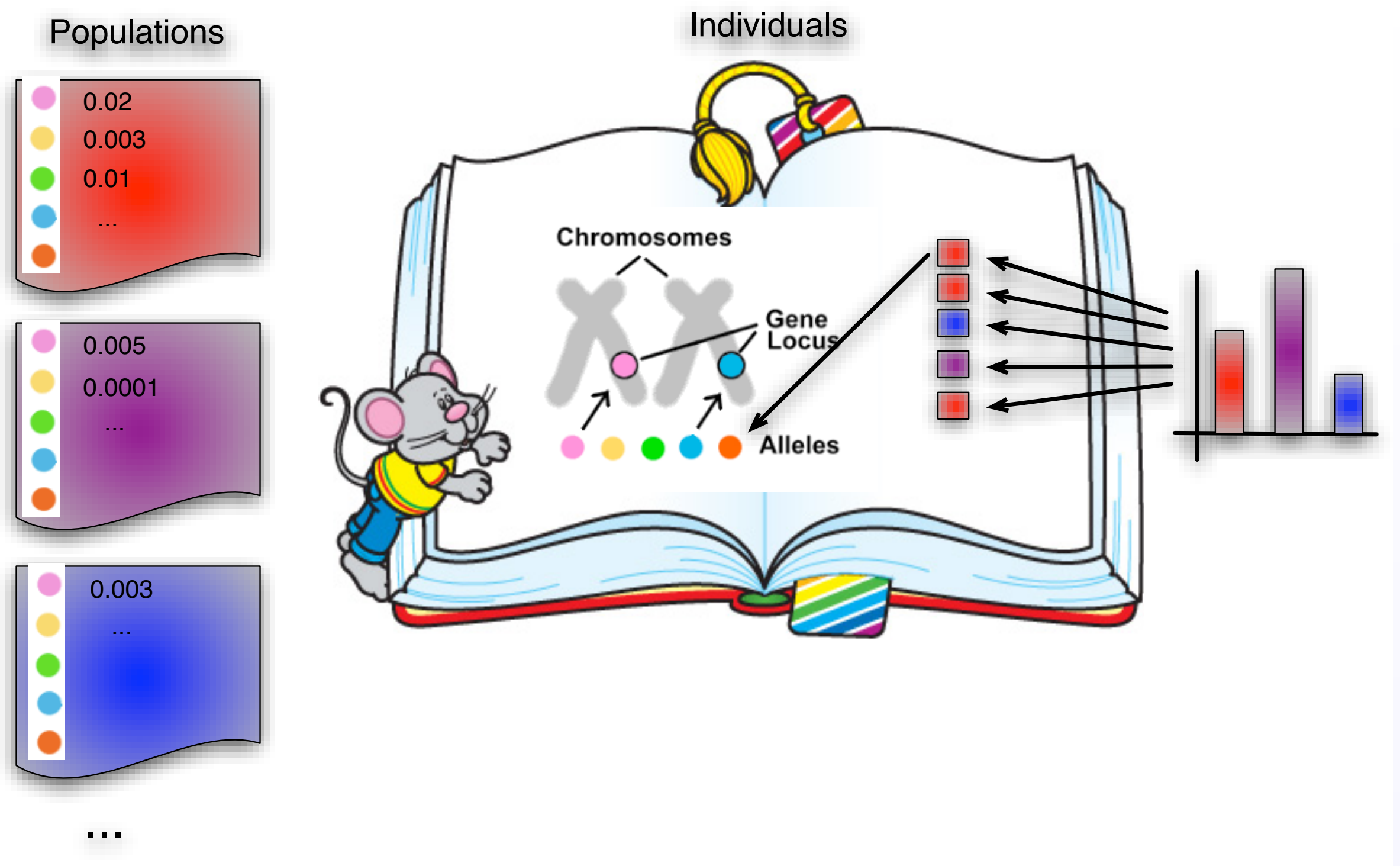
The observed data {documents, genotypes, images} is a mixture of latent components {topics, populations, objects}

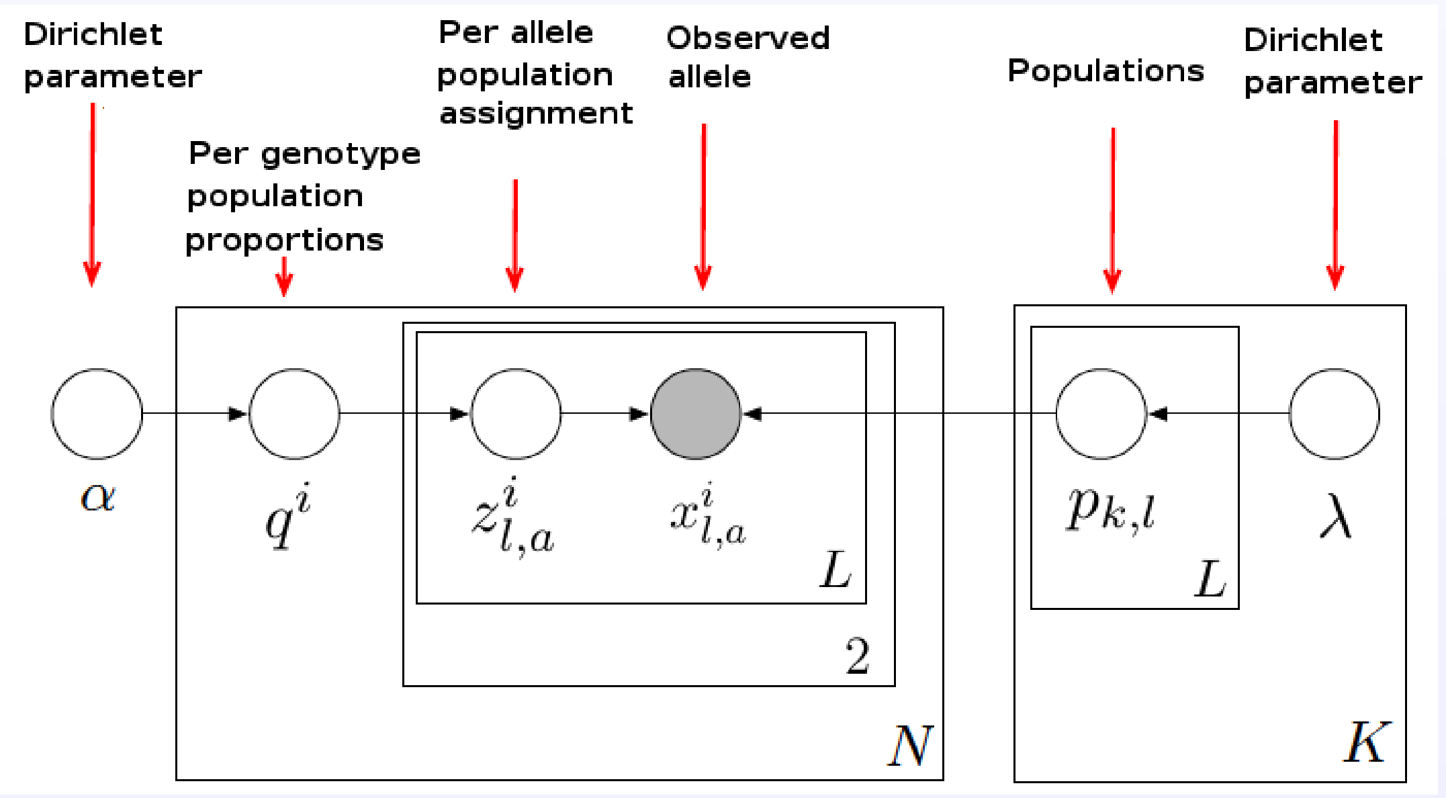
LDA





Population STRUCTURE





Equivalency

Documents
⇕
Genotypes

Words
⇕
Alleles

Topics
⇕
Populations

Generative Process: STRUCTURE View

(latent) P: populations

(latent) Q: admixture proportions in each genotype

(latent) Z: assigned populations for each allele

(observed) X: genotypes

The joint distribution of all variables (observed and latent):

$$p(Q,Z,X,P|\alpha,\lambda) = (\prod_{k=1}^{K}\prod_{l=1}^{L} p(P_{k,l}|\eta)) \left[p(Q|\alpha)(\prod_{l=1}^{L}\prod_{a=1}^{2} Q_{z^i_{l,a}} P_{z^i_{l,a},l,x^i_{l,a}}) \right]$$

$$q^i \sim Dir\ (\alpha)$$

$$p_{k,l} \sim Dir\ (\lambda),\ k = 1,\dots,K, l = 1,\dots,L,\ \lambda = (\lambda_1,\lambda_2,\dots,\lambda_{J_l})$$

$\lambda,\ \alpha$ are hyper-parameters

References

[1] Blei D.M., Ng A.Y., Jordan M.I., Latent Dirichlet Allocation, in *Journal of Machine Learning Research*, 2003

[2] Pritchard J.K., Stephens M. and Donnelly P., Inference of Population Structure Using Multilocus Genotype Data, in *Genetics*, 2000

[3] Tuytelaars T., Lampert C.H., Blaschko M.B., Buntine W., Unsupervised Object Discovery: A comparison, in *International Journal of Computer Vision*, 2010