Machine Learning for Language Processing
ACS 2015/16

Stephen Clark

L5: Topic Modelling and LDA

UNIVERSITY OF CAMBRIDGE

Probabilistic Topic Modelling

- We want to find *themes* (or *topics*) in documents
  - useful for e.g. search or browsing
- We don't want to do supervised topic classification
  - rather not fix topics in advance nor do manual annotation
- Need an approach which automatically teases out the topics
- This is essentially a *clustering* problem - can think of both words and documents as being clustered

Key Assumptions behind LDA

- Documents exhibit multiple topics (but typically not many)
- LDA is a probabilistic model with a corresponding *generative process*
  - each document is assumed to be generated by this (simple) process
- A *topic* is a distribution over a fixed vocabulary
  - these topics are assumed to be generated first, before the documents
- Only the number of topics is specified in advance

Example Topics

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

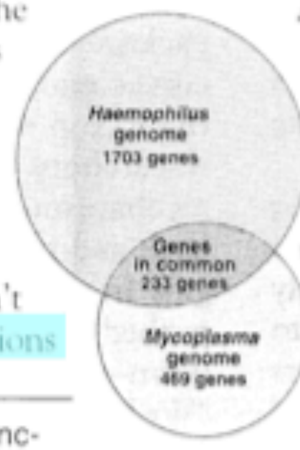
Taken from Blei 2012 ICML tutorial

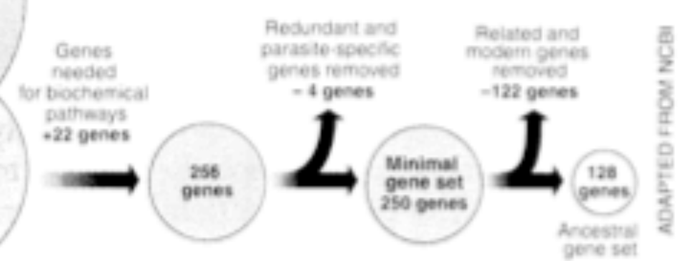
Documents and Topics

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an





Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

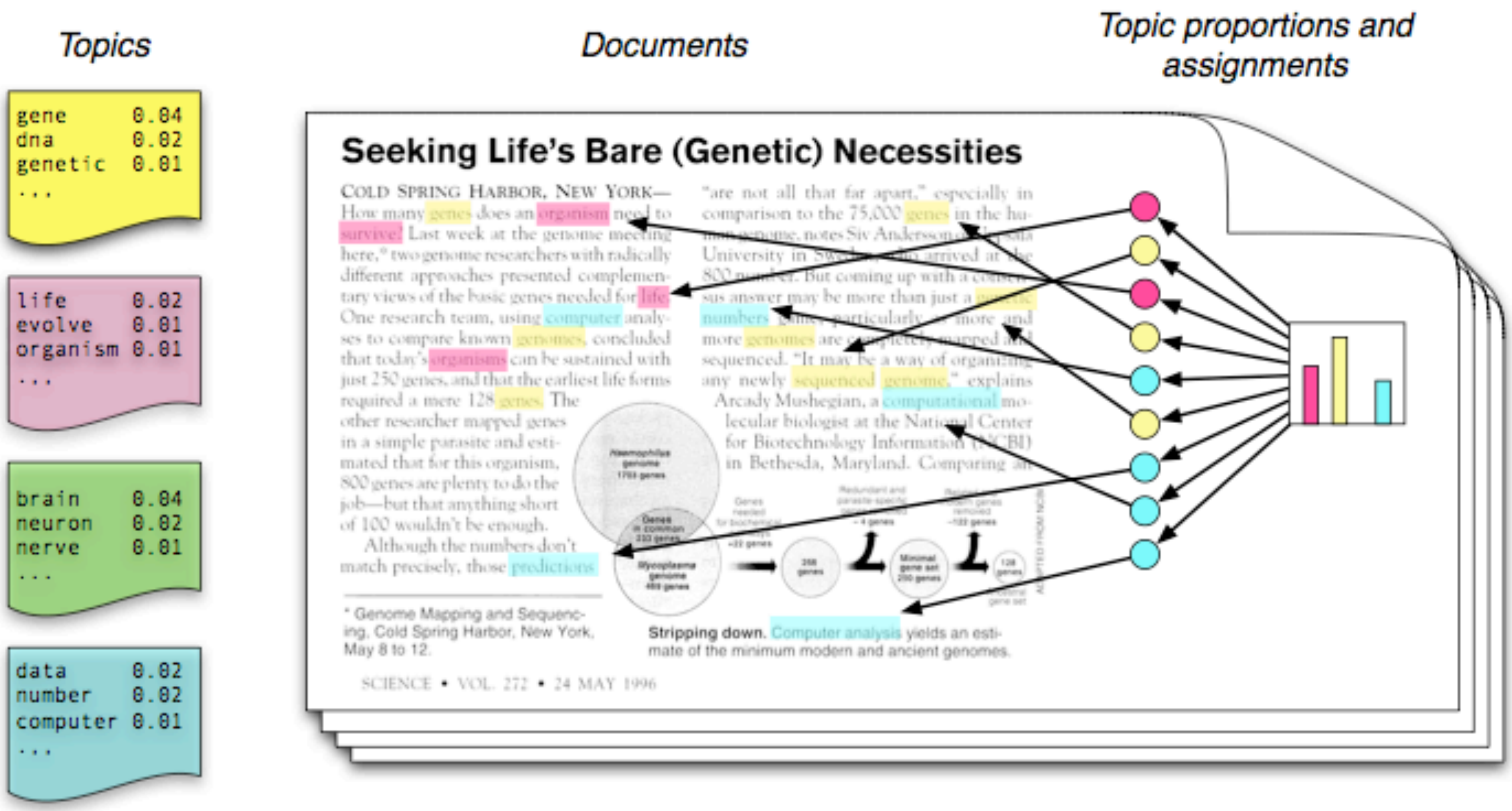
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Taken from Blei 2012 ICML tutorial

UNIVERSITY OF CAMBRIDGE

Documents and Topics



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Taken from Blei 2012 ICML tutorial

The Generative Process

To generate a document:

1. Randomly choose a distribution over topics

2. For each word in the document

   a. randomly choose a topic from the distribution over topics
   b. randomly choose a word from the corresponding topic (distribution over the vocabulary)

- Note that we need a distribution over a distribution (for step 1)

- Note that words are generated independently of other words (unigram bag-of-words model)

The (Formal) Generative Process

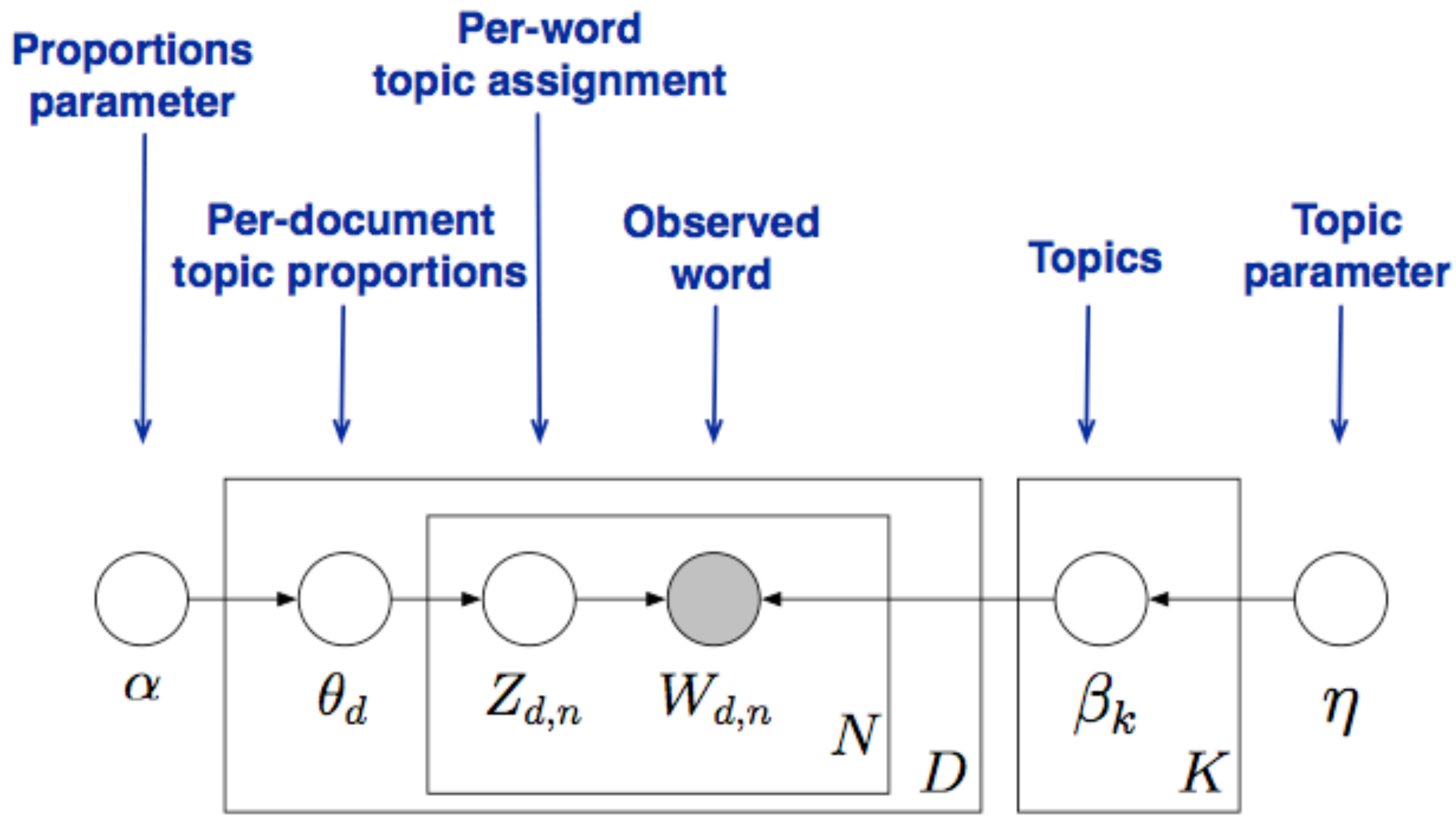
- Some notation:
  - $\beta_{1:K}$ are the topics where each β_k is a distribution over the vocabulary
  - θ_d are the topic proportions for document d
  - $\theta_{d,k}$ is the topic proportion for topic k in document d
  - z_d are the topic assignments for document d
  - $z_{d,n}$ is the topic assignment for word n in document d
  - w_d are the observed words for document d

- The joint distribution (of the hidden and observed variables):

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})$$

UNIVERSITY OF CAMBRIDGE

LDA as a Graphical Model



$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Taken from Blei 2012 ICML tutorial

The Dirichlet Distribution

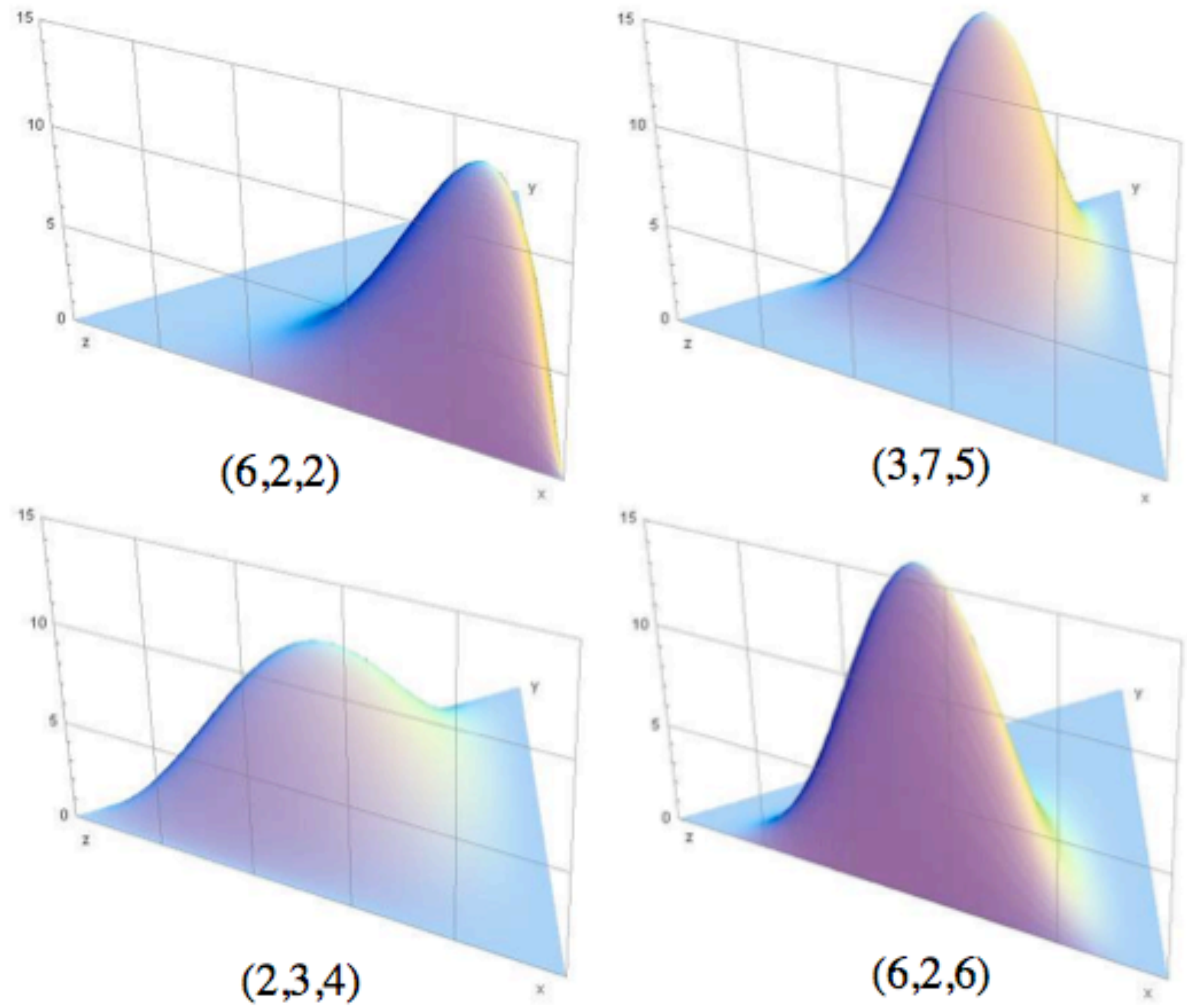
- **Dirichlet** (continuous) distribution with parameters $\boldsymbol{\alpha}$

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{d} \alpha_i)}{\prod_{i=1}^{d} \Gamma(\alpha_i)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}; \quad \text{for "observations"}: \sum_{i=1}^{d} x_i = 1, \quad x_i \geq 0$$

- $\Gamma()$ is the Gamma distribution

- Conjugate prior to the multinomial distribution
(form of posterior $p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M})$ is the same as the prior $p(\boldsymbol{\theta}|\mathcal{M})$)

The Dirichlet Distribution



- Parameters: $(\alpha_1, \alpha_2, \alpha_3)$

UNIVERSITY OF CAMBRIDGE

Parameter Estimation

- Main variables of interest:
  - β_k: distribution over vocabulary for topic k
  - $\theta_{d,k}$: topic proportion for topic k in document d
- Could try and get these directly, eg using EM (Hoffmann, 1999), but this approach not very successful
- One common technique is to estimate the posterior of the word-topic assignments, given the observed words, directly (whilst marginalizing out β and θ)
- Gibbs sampling is an example of a Markov Chain Monte Carlo (MCMC) technique

UNIVERSITY OF CAMBRIDGE

Estimates using Gibbs Sampling

- The Gibbs sampler produces the following estimate, where, following Steyvers and Griffiths:
  - z_i is the topic assigned to the ith token in the whole collection;
  - d_i is the document containing the ith token;
  - w_i is the word type of the ith token;
  - $\mathbf{z}_{-i}$ is the set of topic assignments of all other tokens;
  - $\cdot$ is any remaining information such as the α and η hyperparameters:

$$P(z_i = j | \mathbf{z}_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \eta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\eta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

where $\mathbf{C}^{WT}$ and $\mathbf{C}^{DT}$ are matrices of counts (word-topic and document-topic)

UNIVERSITY OF CAMBRIDGE

The Final Estimates

$$\beta_{ij} = \frac{C_{ij}^{WT} + \eta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\eta} \quad \theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

- Using the count matrices as before, where β_{ij} is the probability of word type i for topic j, and θ_{dj} is the proportion of topic j in document d

UNIVERSITY OF CAMBRIDGE