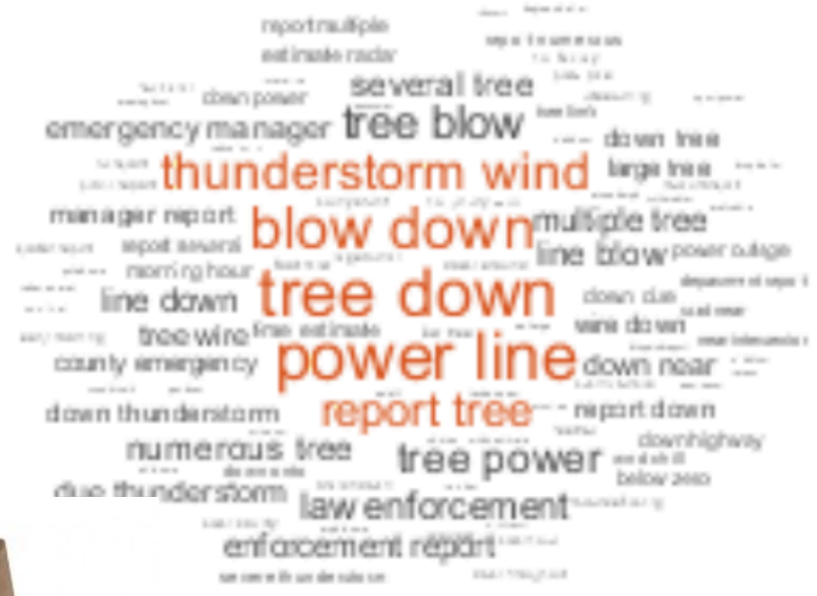
Latent Dirichlet Allocation

Authors: David M. Blei, Andrew Y. Ng, and Michael I. Jordan
Presenter: Leah Li


wind gust
gust knot
mph wind
gust mph
tree report
report down
estimate mph
power outage
damage report
measure wind
gust measure
international airport
thunderstorm produce


thunderstorm wind
blow down
tree down
power line
several tree
tree blow
emergency manager
report multiple
estimate radar
down power
line down
tree wire
county emergency
down thunderstorm
report tree
numerous tree
tree power
law enforcement
enforcement report
multiple tree
line blow
power outage
down near
report down
downhighway
below zero


WORDS


gust mph
wind gust
sustain wind
peak gust
wind mph
peak wind
flood stage
high wind
measure sustain
mph highwaydot sensor
measure peak
heavy rain


emergency management
flooding occur
rain fell
inch rain
heavy rain
drought condition
reduce visibility
management report
thunderstorm move
report trees
rain fall
damage report
county produce
minor flooding
trees down
quarter mile
half inches
mile less
two inch
inches rain

Topics

Documents

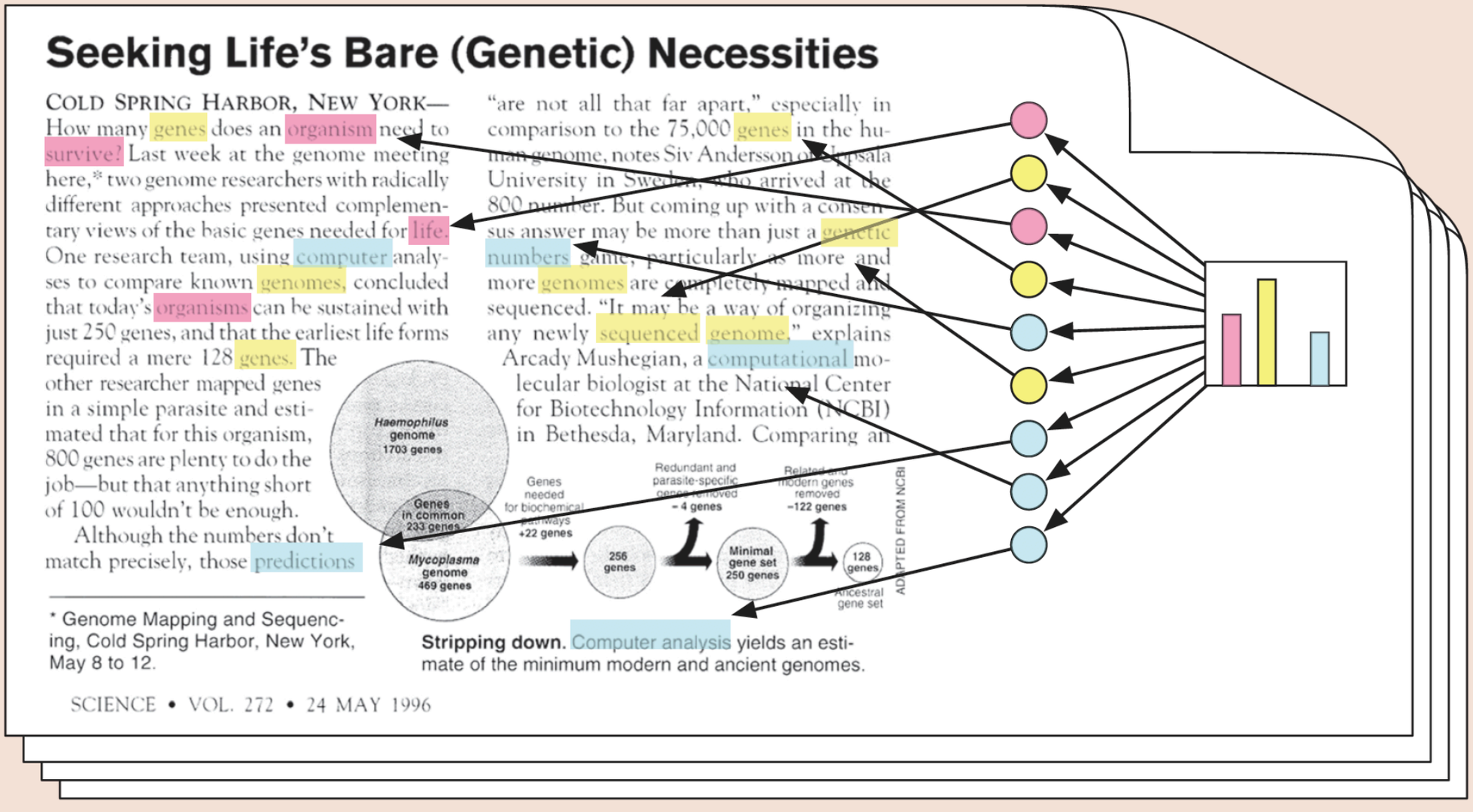
Topic proportions and assignments

gene	0.04
dna	0.02
genetic	0.01
. . .	

life	0.02
evolve	0.01
organism	0.01
. . .	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. . .	



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.

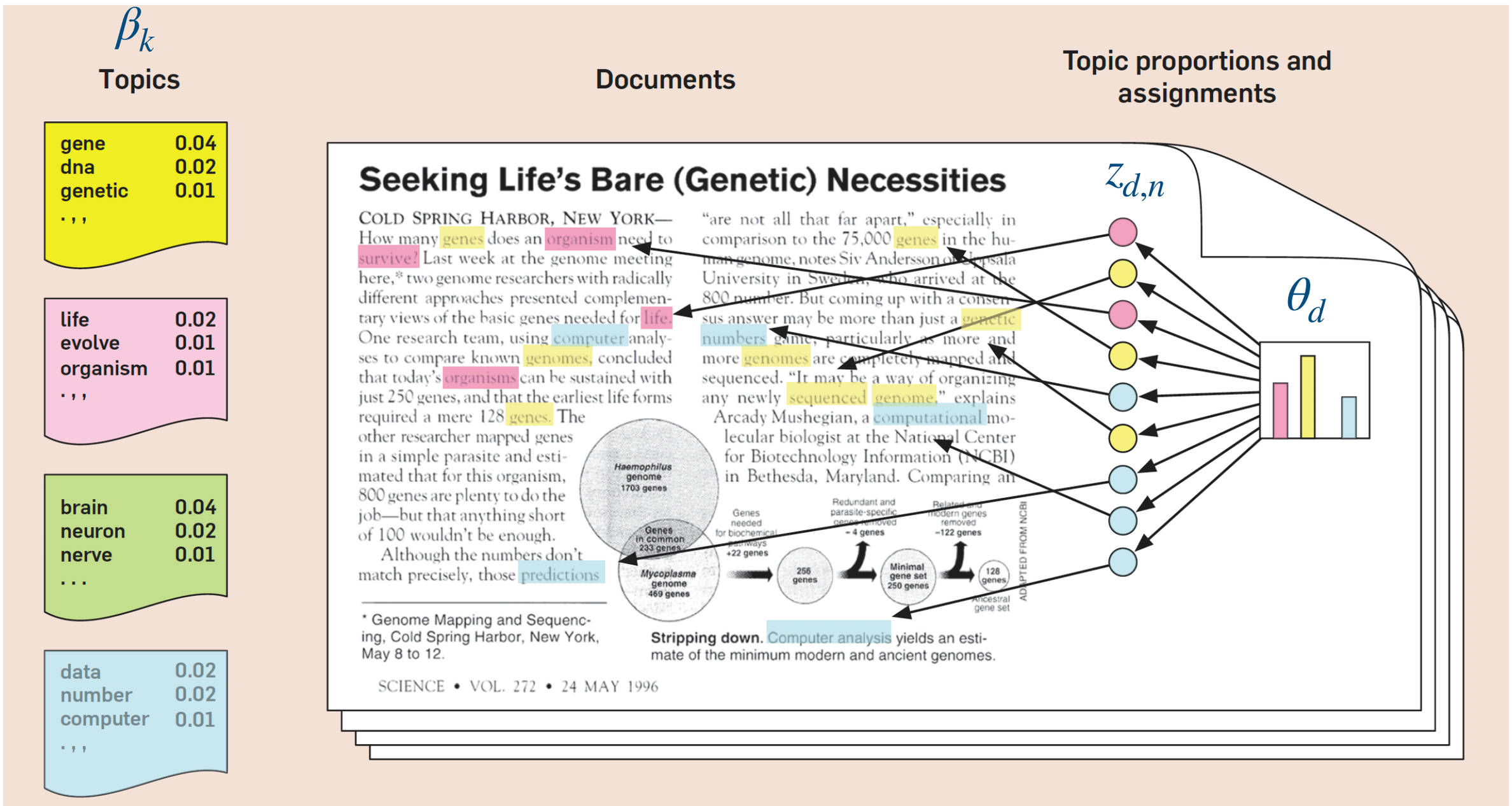
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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

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

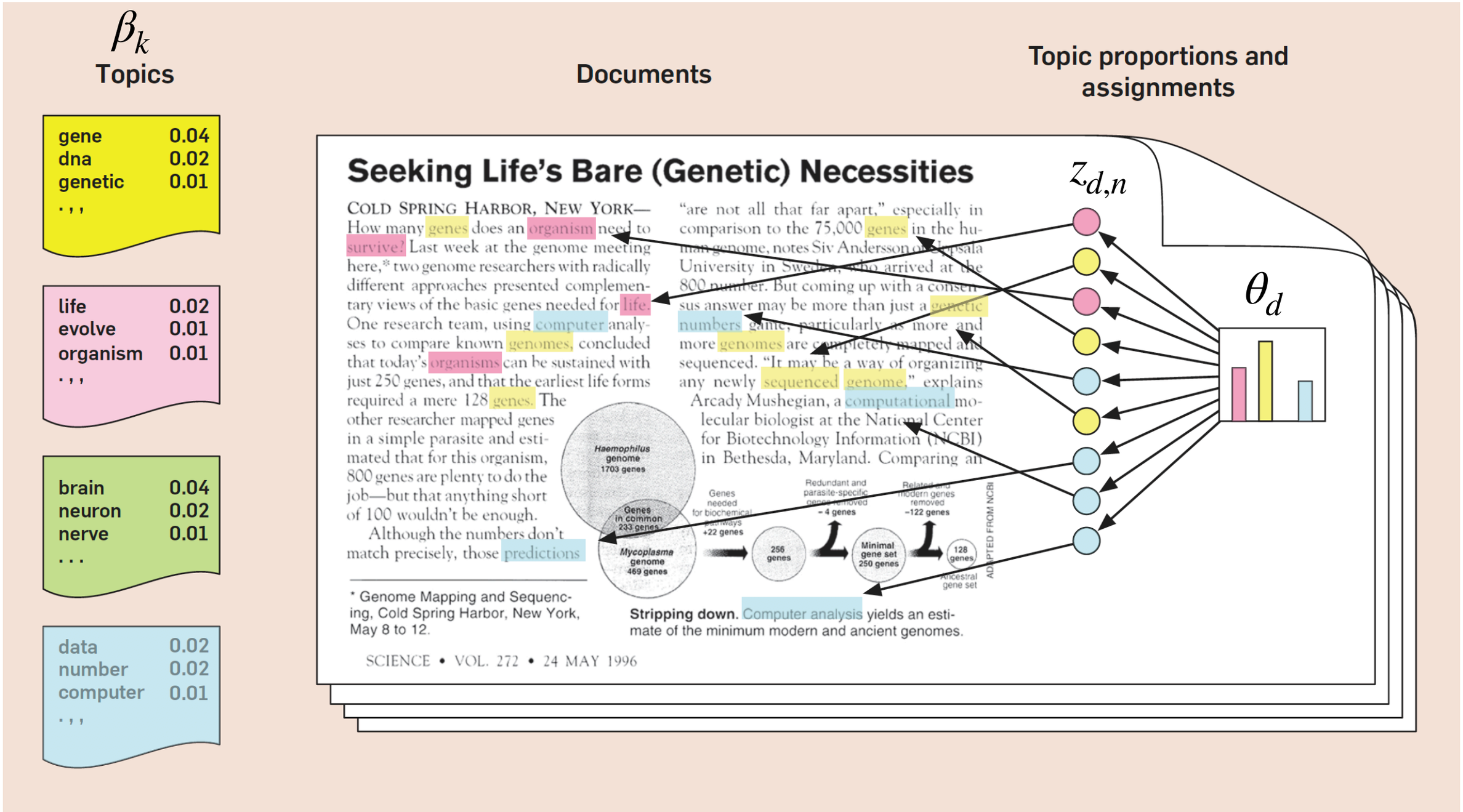
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

- Our goal is to infer or estimate the hidden variables, i.e. computing their distribution conditioned on the documents.

$$P(topics, proportions, assignments \mid documents) = P(\beta, z, \theta \mid w)$$



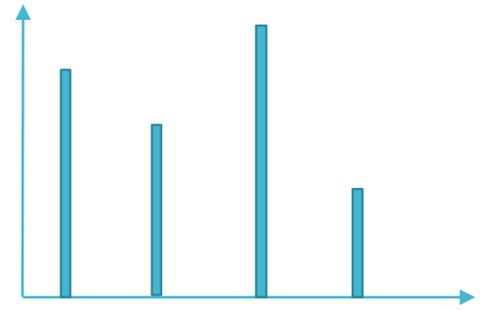
- Use posterior expectations ($E(\beta \mid w)$ for the corpus, $E(\theta \mid w)$ for each document) to perform the tasks: information retrieval, document similarity, exploration.



$p(\theta \mid w)$ Per document topic distributions

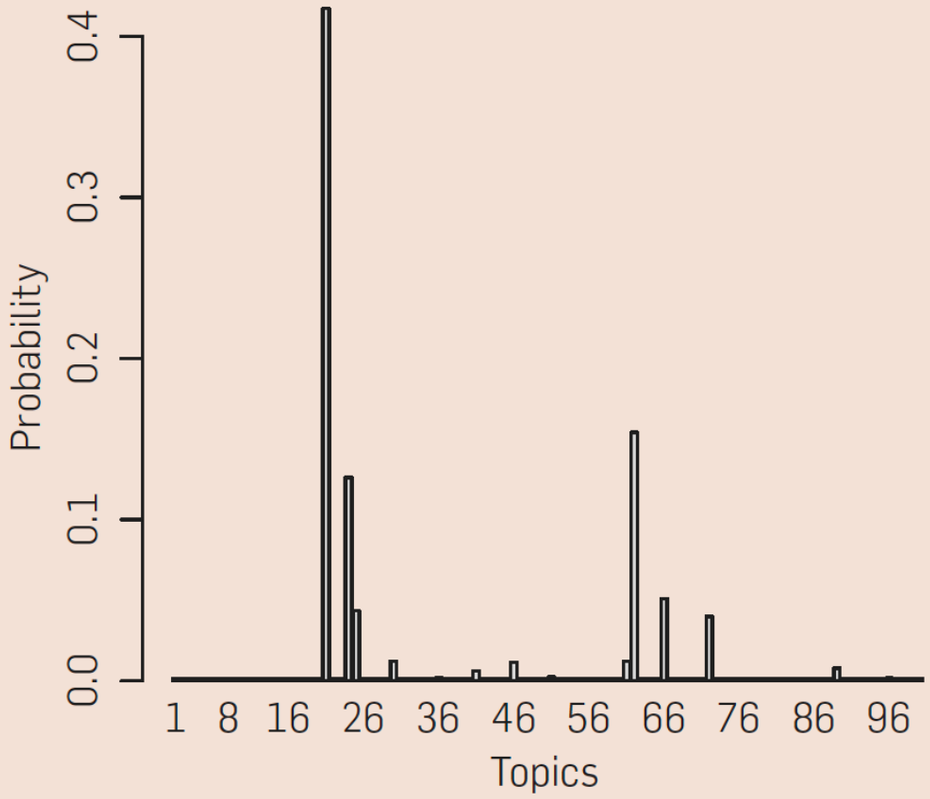
$p(\beta \mid w)$ Topic word distributions

θ_d:



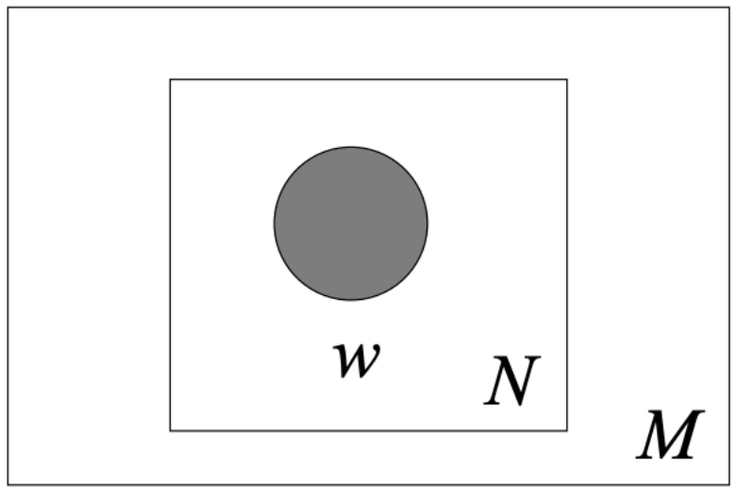
β:

	Word probabilities for each topic		
Topics			



“Genetics”	**“Evolution”**	**“Disease”**	**“Computers”**
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

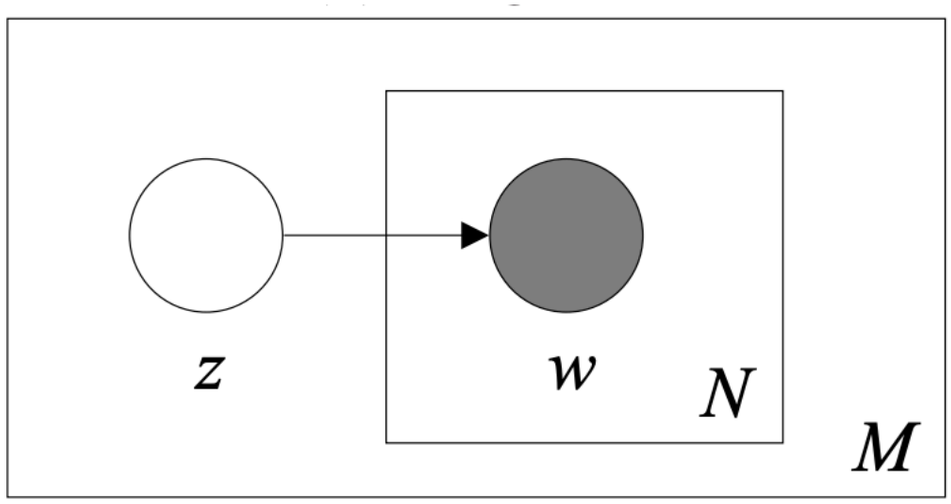
Unigram Model



- the words of every document are drawn independently from a single multinomial distribution

$$p(\mathbf{w}) = \prod_{n=1}^{N} p(w_n).$$

Mixture of Unigrams Model

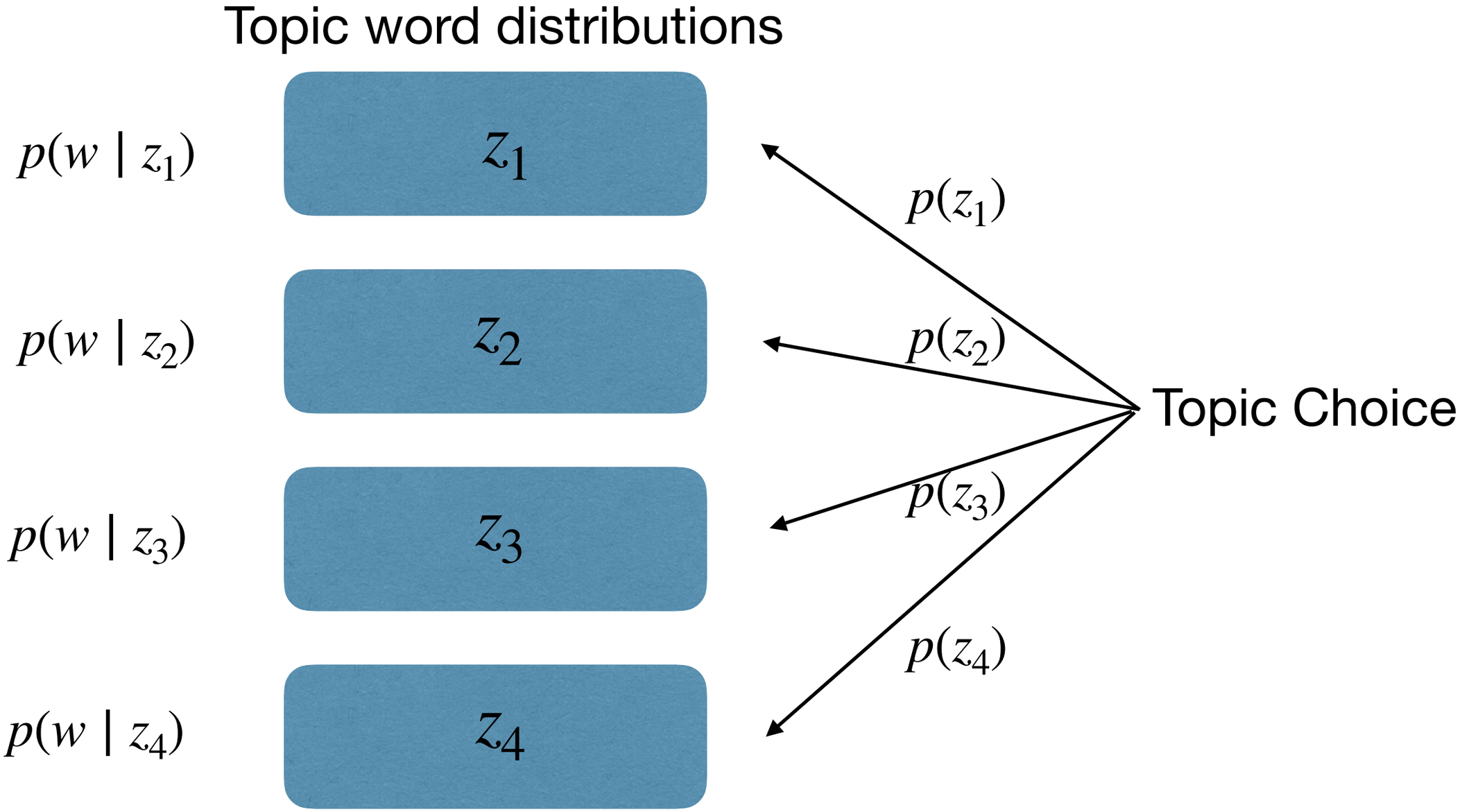


- Argument the unigram model with a discrete random topic variables z and obtain a mixture of unigrams model.

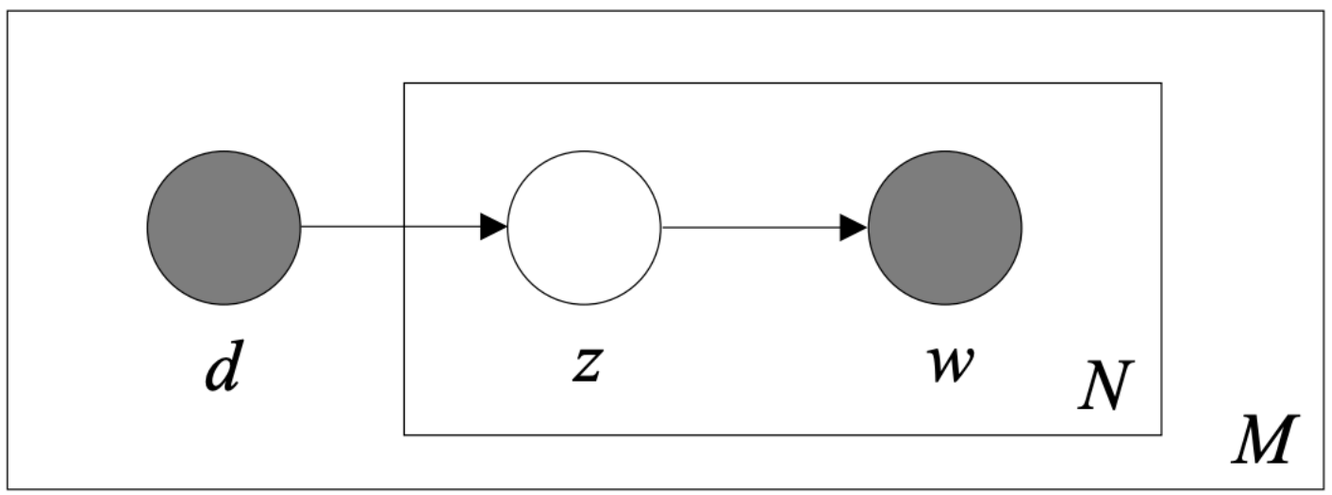
- Each document is generated by first choosing a topic z and then generating words independently from the conditional multinomial

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z).$$

$$p(\mathbf{w}) = \sum_{z} p(z) \prod_{n=1}^{N} p(w_n \mid z).$$



pLSI

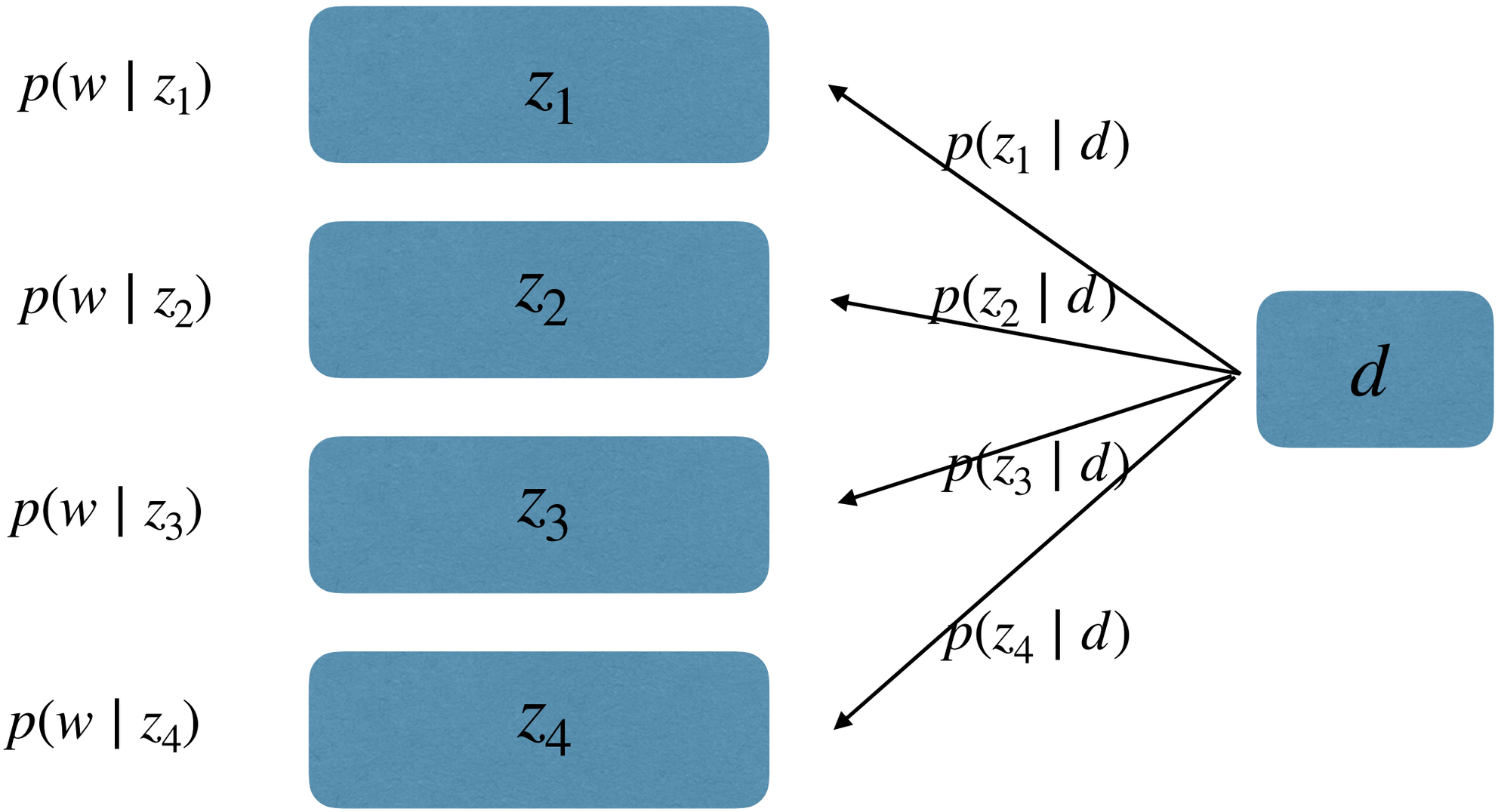


- The pLSI model attempts to relax the simplifying assumption made in the mixture of unigrams model that each document is generated from only one topic.

- Given all parameters, we want to infer the distribution z a word is from

$$p(d, w_n) = p(d) \sum_z p(w_n|z) p(z|d).$$

$$p(d, w_n) = p(d) \sum_z p(w_n | z) p(z | d).$$

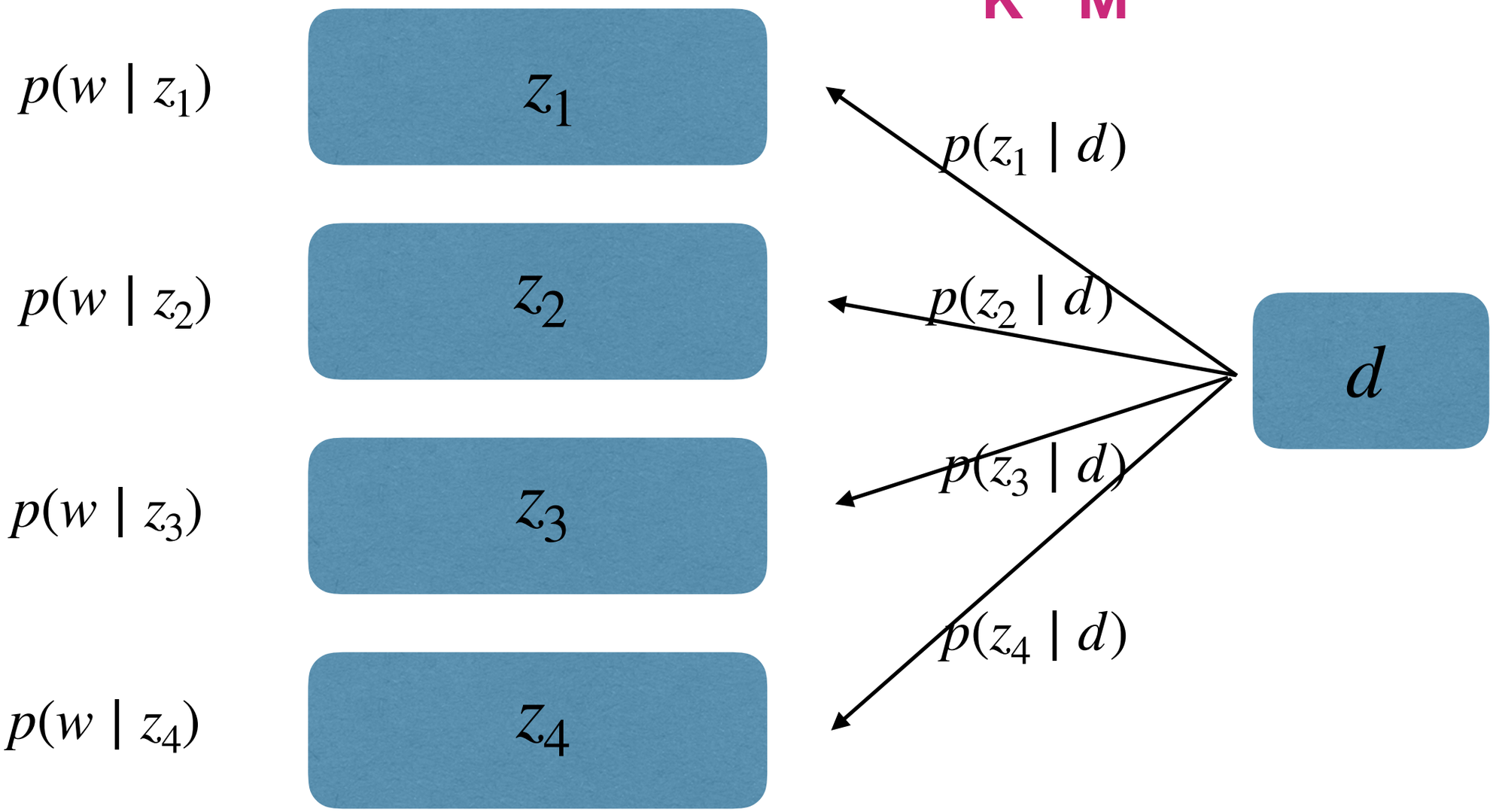


pLSI Deficiency

K * V

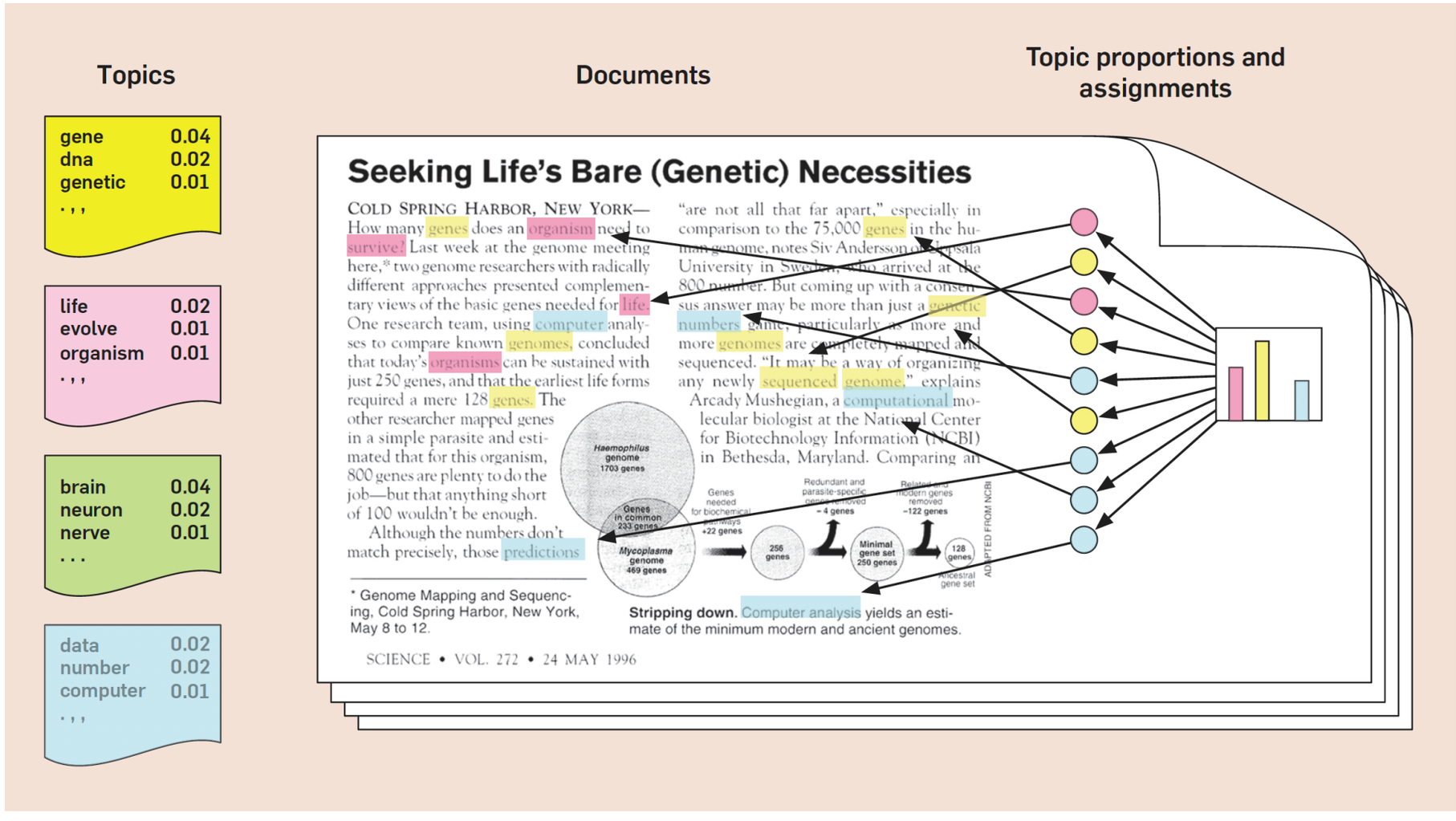
Topic word distributions

K * M

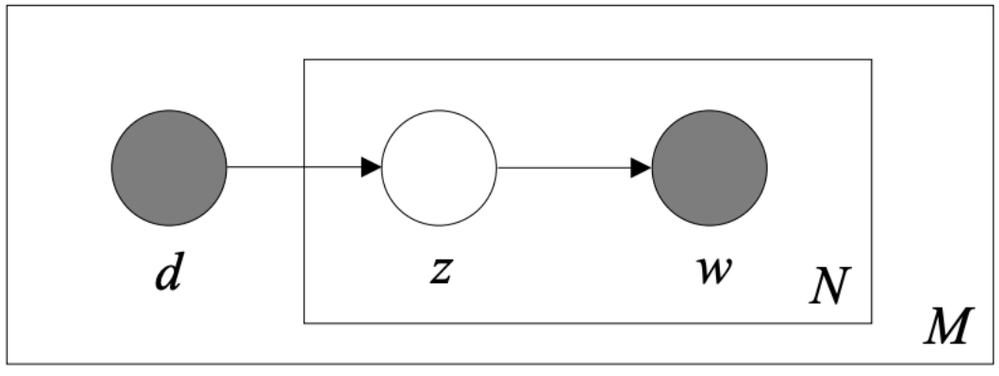


Exchangeability

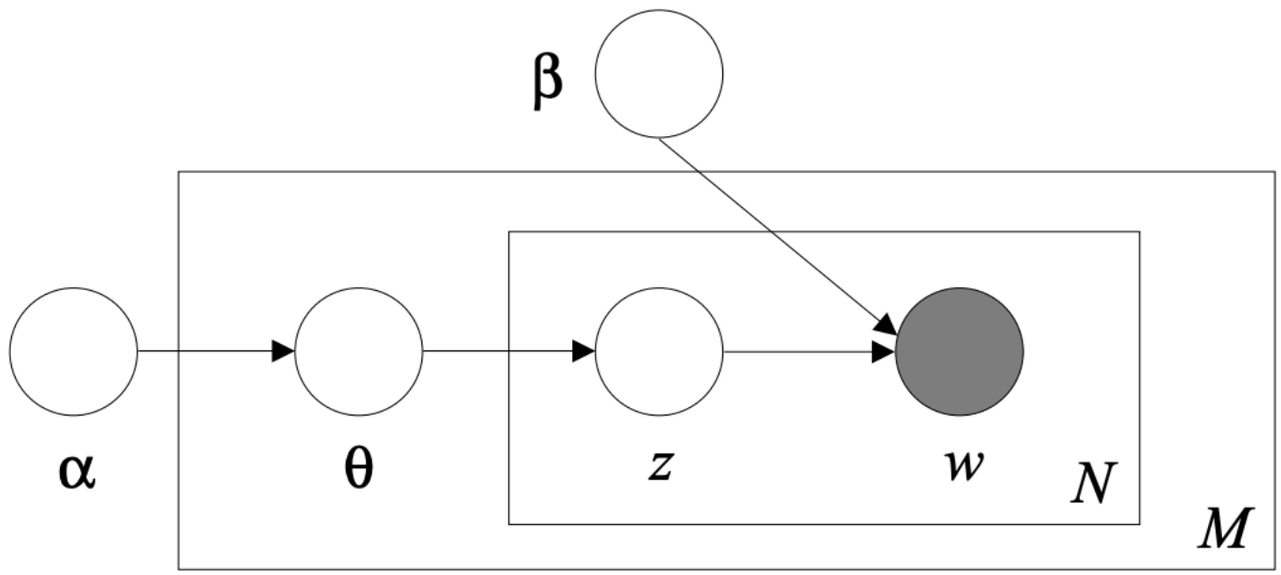
- De Finetti's theorem states that any set of exchangeable random variables has a representation as a mixture distribution.



LDA



- LDA improves pLSI to be a generative model by imposing Dirichlet priors on the model parameters.



$$p(\mathbf{w}) = \int_\theta \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n;\beta)p(z_n|\theta) \right) p(\theta;\alpha)d\theta,$$

Dirichlet Distribution

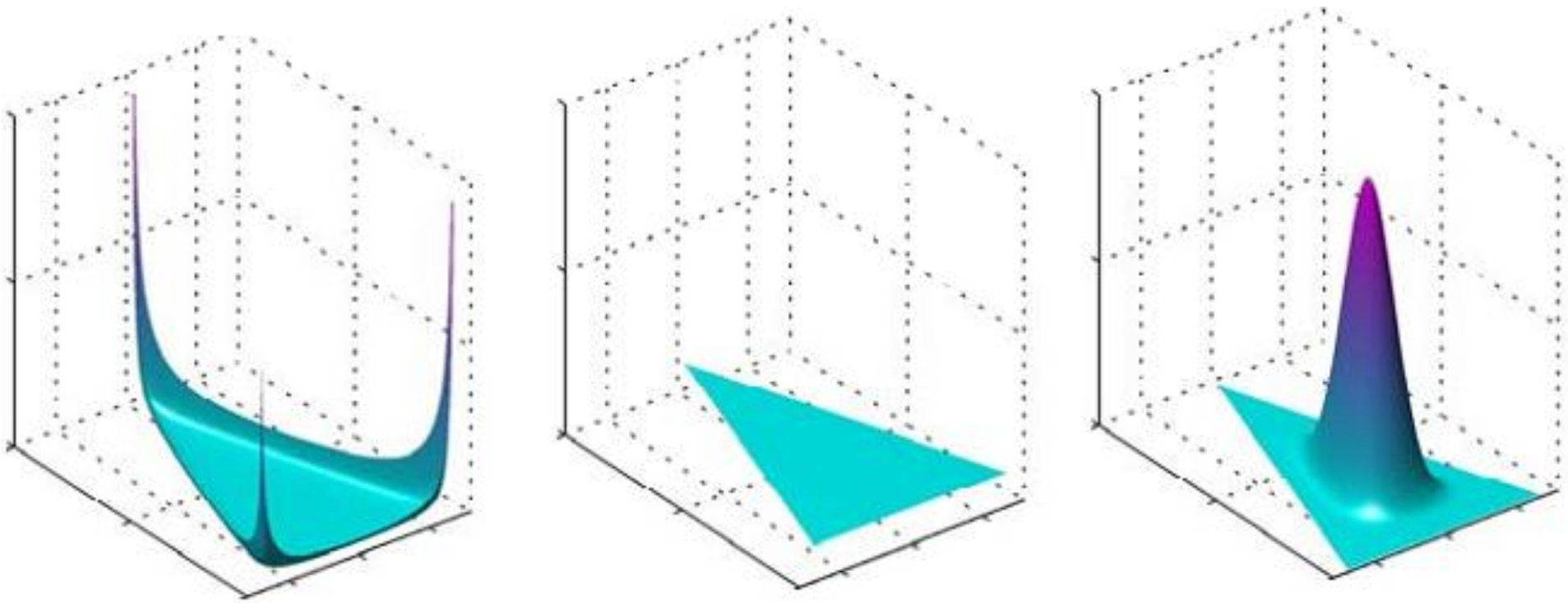
- The topic mixture proportions for each document are drawn from some distribution

- The space of all of these multinomials has a nice geometric interpretation as (k-1)-simplex

- The distribution is defined over a (k-1)-simplex. It takes k non-negative arguments which sum to one. A natural distribution to use over multinomial distribution

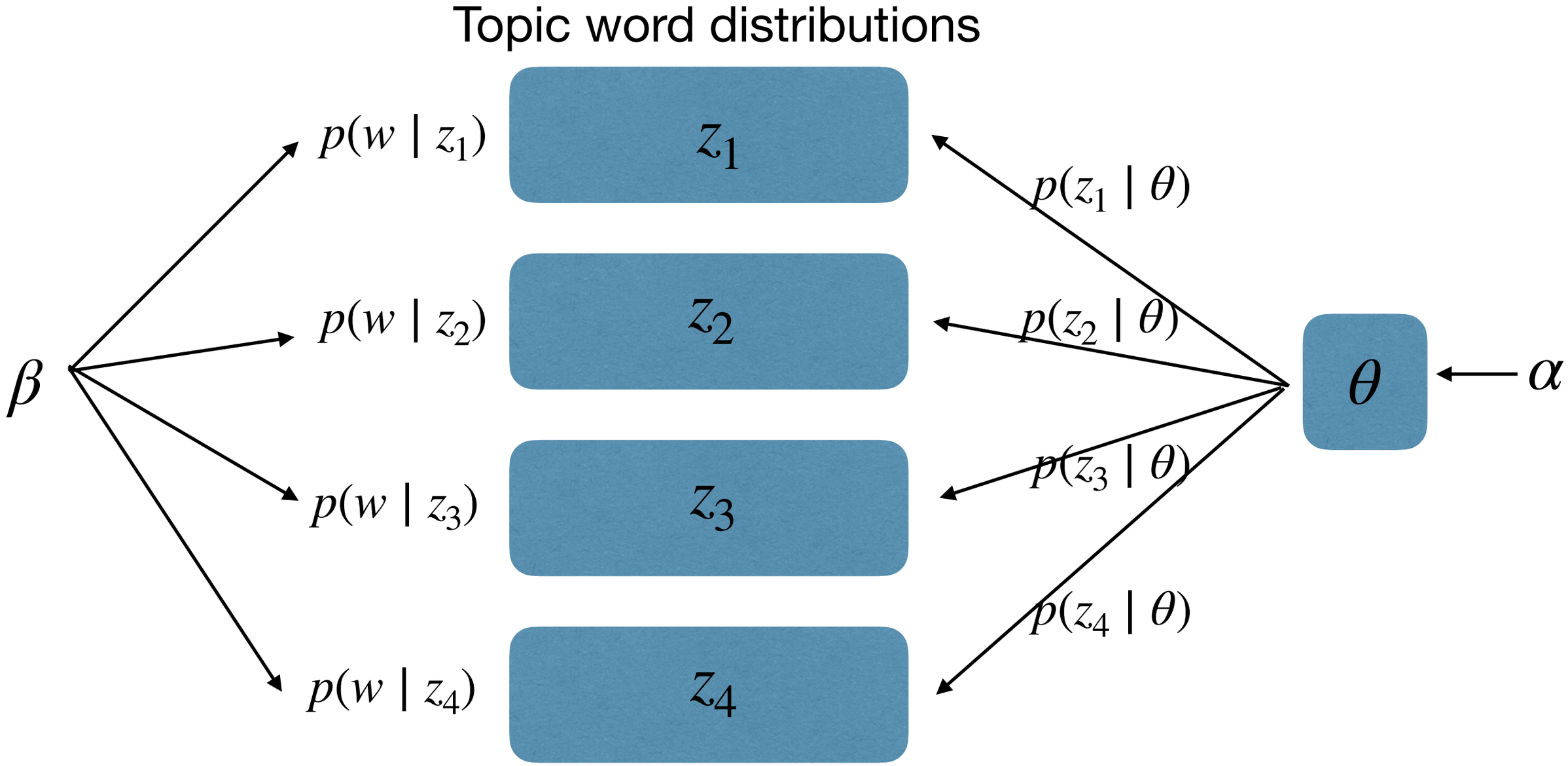
- Explain probability distribution of probability distribution

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1},$$



$\{\alpha_k\} = 0.1$ $\{\alpha_k\} = 1$ $\{\alpha_k\} = 10$

$$p(\mathbf{w}) = \int_{\theta} \left(\prod_{n=1}^{N} \sum_{z_n=1}^{k} p(w_n|z_n;\beta)p(z_n|\theta) \right) p(\theta;\alpha)d\theta,$$



Continuous mixture of unigram

Mixture of unigram

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right) d\theta.$$

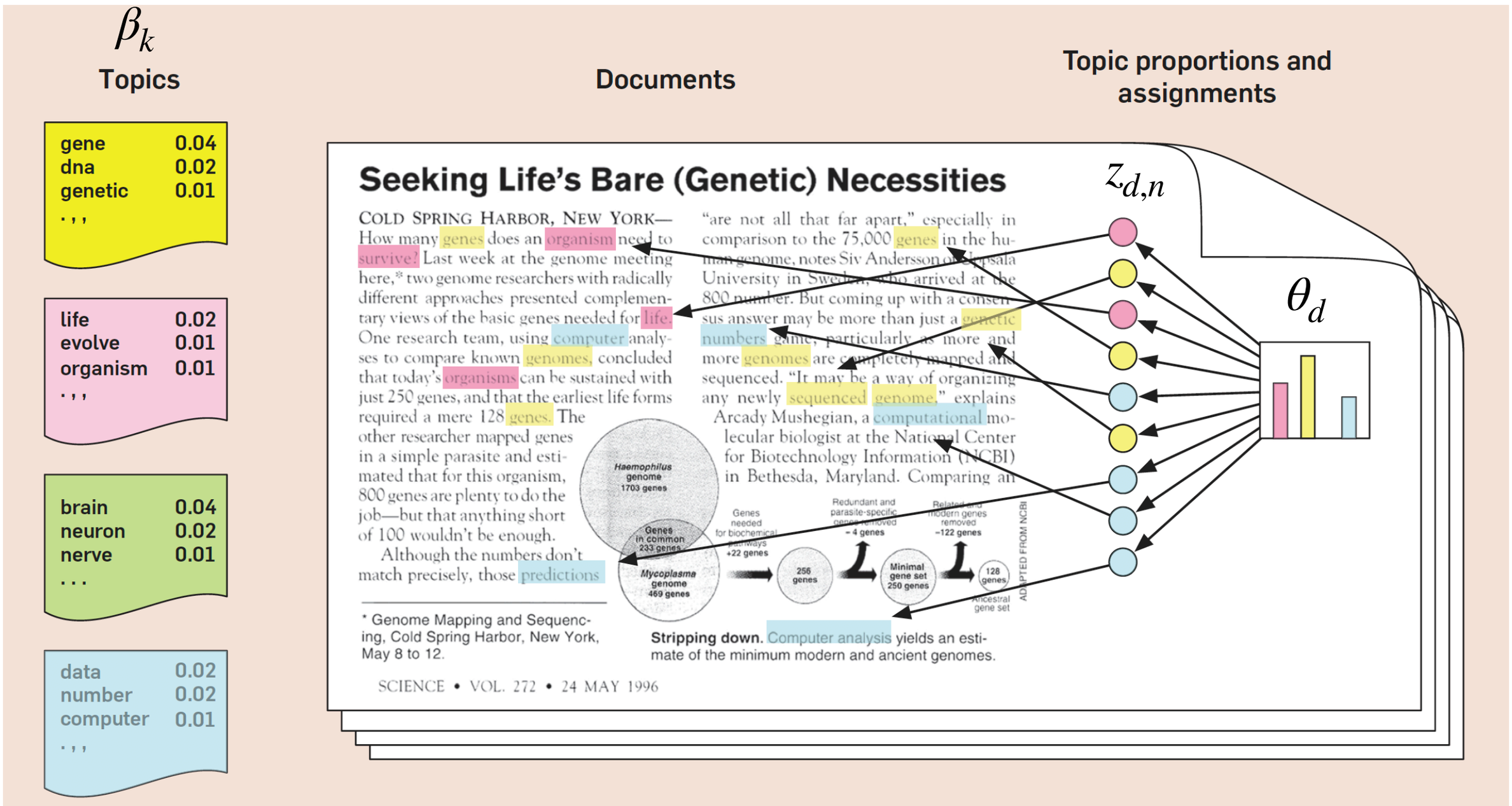
- Integral over all possibilities because document topic distribution is not fixed anymore.

Marginal likelihood

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right) d\theta_d.$$

- Our goal is to infer or estimate the hidden variables, i.e. computing their distribution conditioned on the documents.

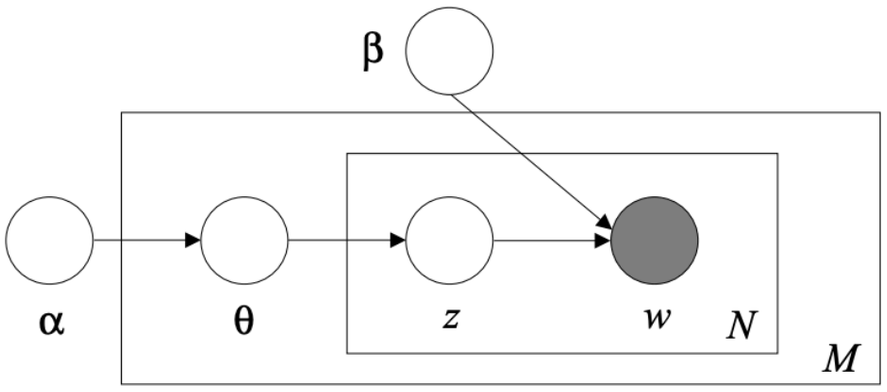
$$P(topics, proportions, assignments \mid documents) = P(\beta, z, \theta \mid w)$$

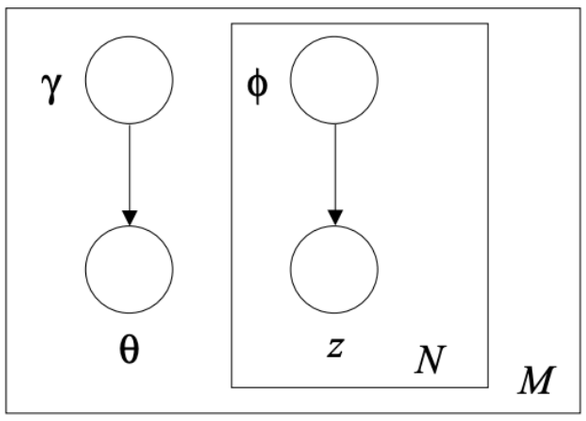


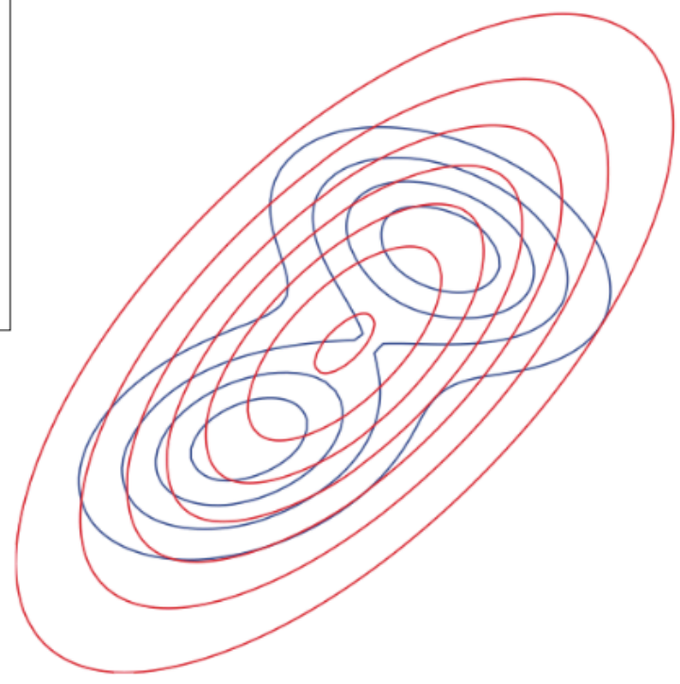
Variation Inference

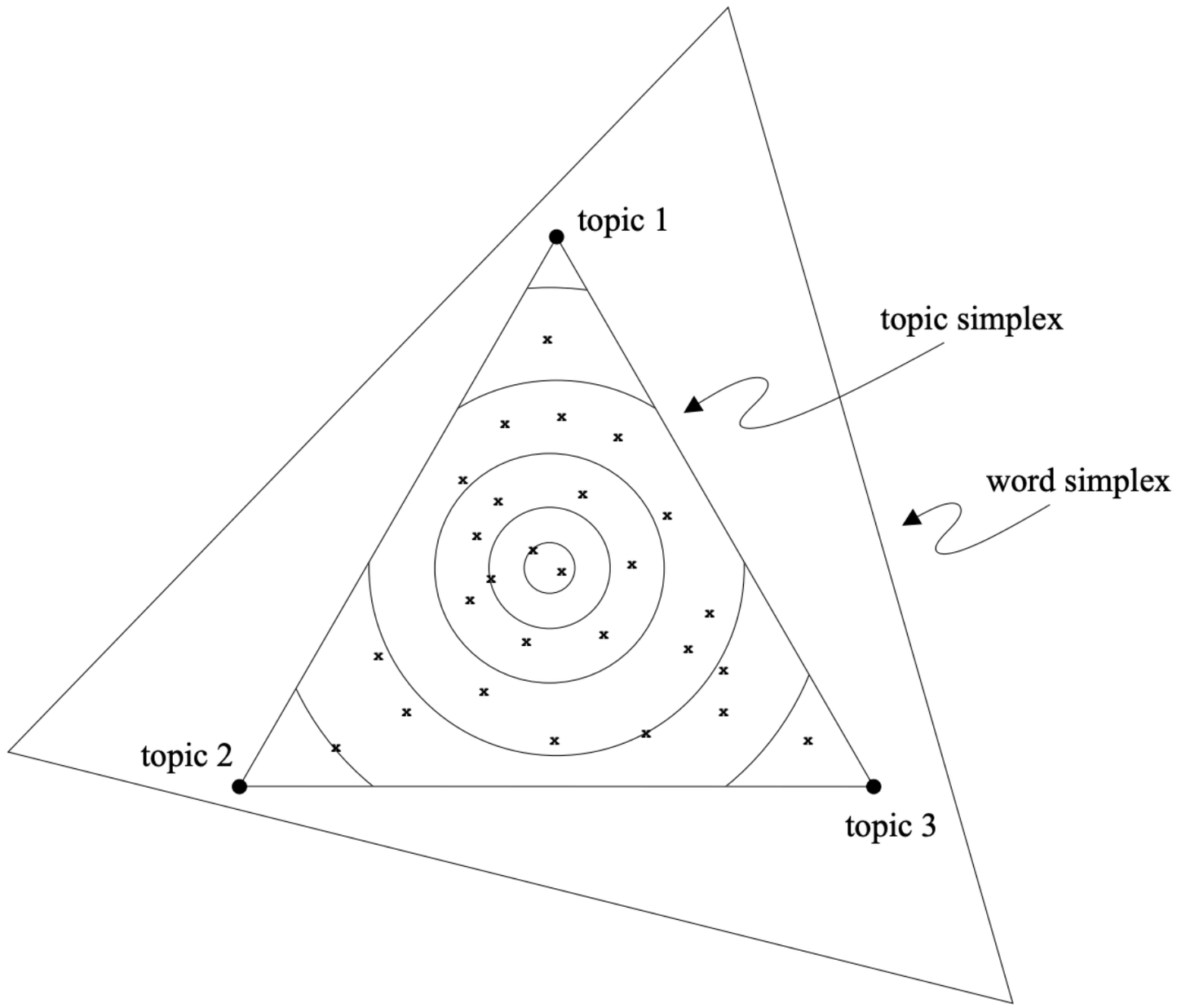
- The key inferential problem that we need to solve in order to use LDA is that of computing the posterior distribution of the hidden variables given a document

- Unfortunately, this distribution is intractable to compute in general







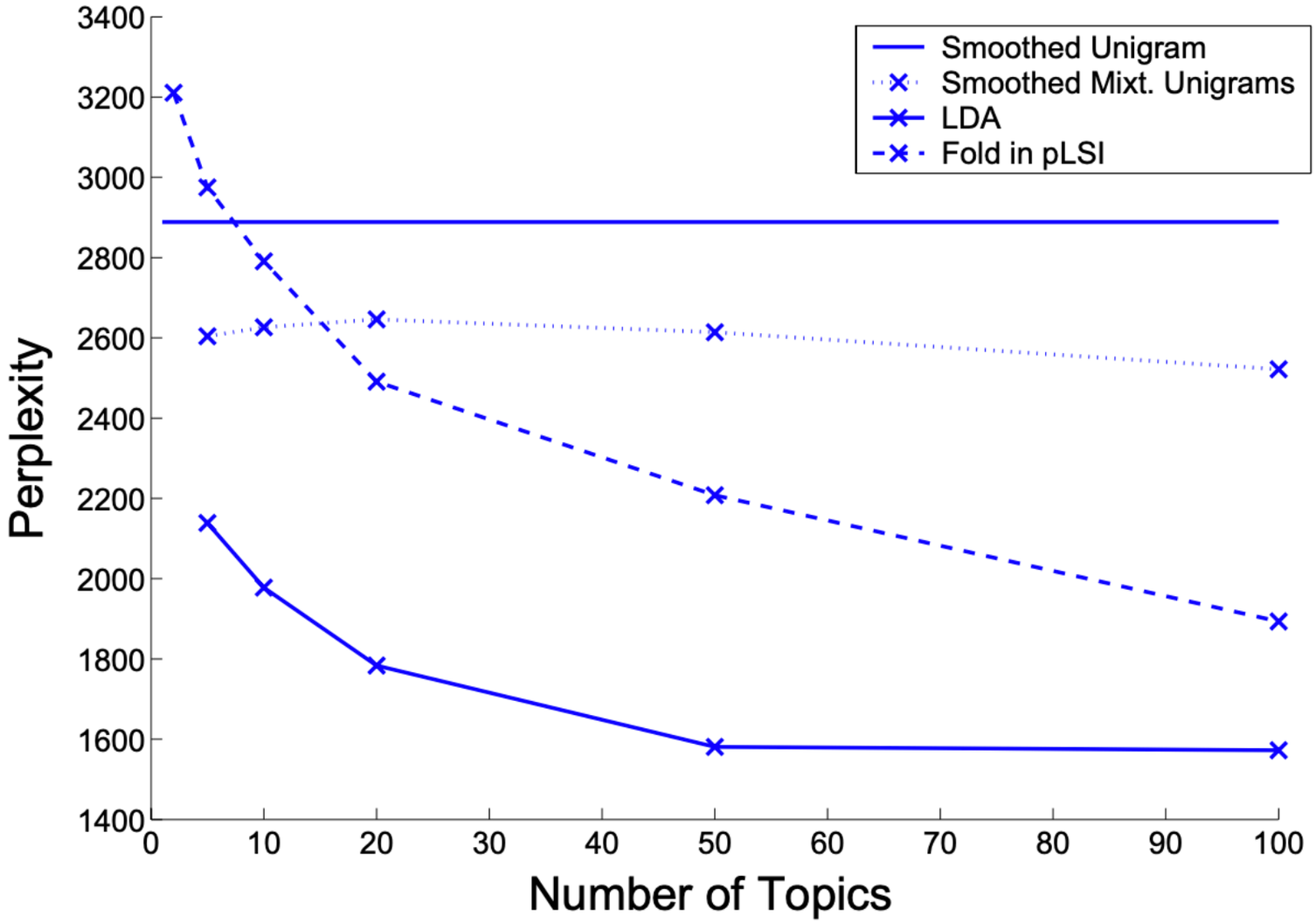


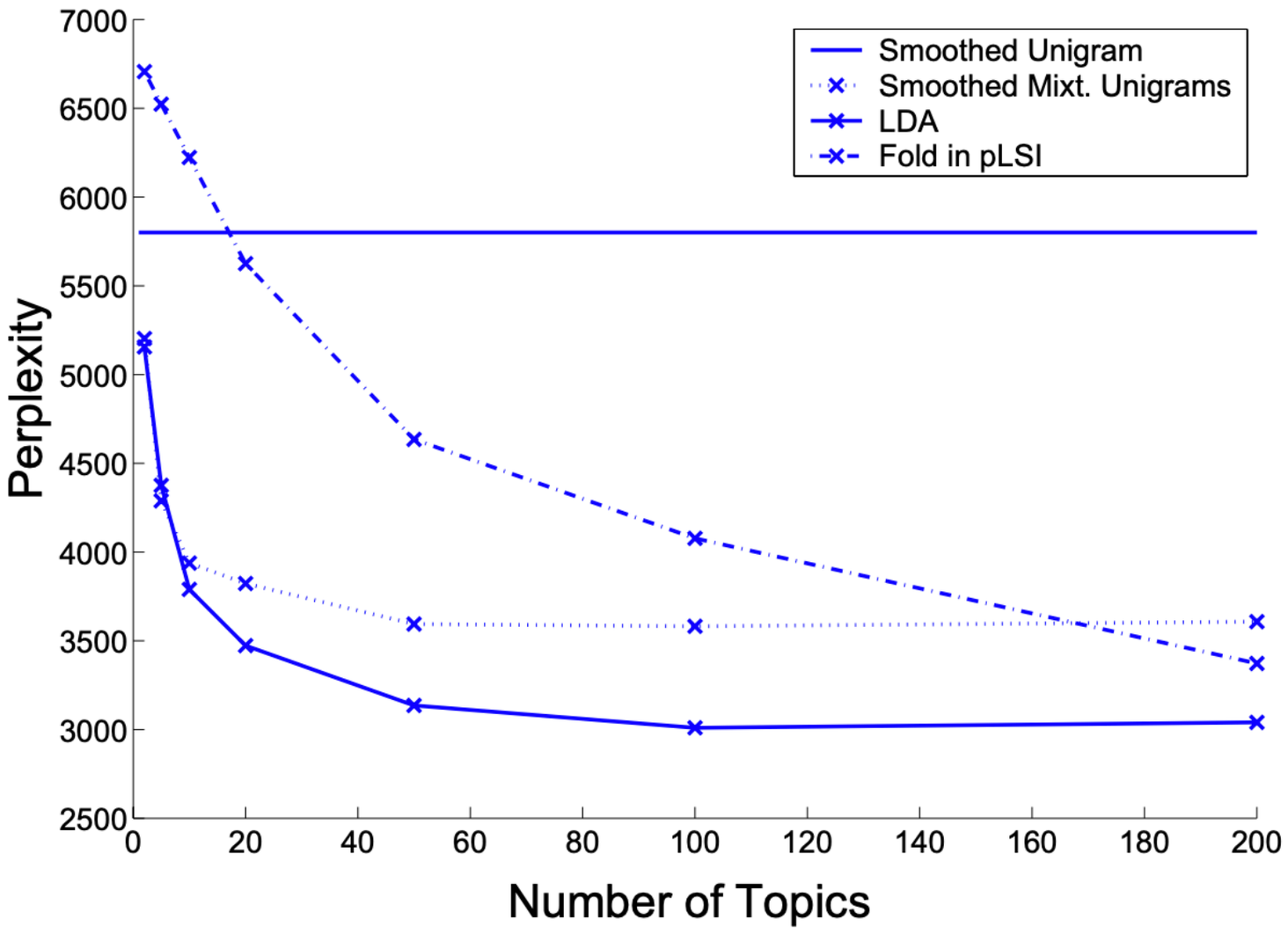
The topic simplex for three topics embedded in the word simplex for three words. The corners of the word simplex correspond to the three distributions where each word (respectively) has probability one. The three points of the topic simplex correspond to three different distributions over words. The mixture of unigrams places each document at one of the corners of the topic simplex. The pLSI model induces an empirical distribution on the topic simplex denoted by x. LDA places a smooth distribution on the topic simplex denoted by the contour lines.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.


Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
0
10
20
30
40
50
60
70
80
90
100
Number of Topics


Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
0
20
40
60
80
100
120
140
160
180
200
Number of Topics

Mixture model parameters: $\Lambda = (\{p(w \mid z)\}, \{p(z)\})$

Likelihood: $p(C \mid \Lambda)$

ML estimate: $\hat{\Lambda} = \arg\max_{\Lambda} p(C \mid \Lambda)$

Mixture model parameters: $\Lambda = (\{p(w \mid z)\}, \{p(z \mid d)\})$

Likelihood: $p(C \mid \Lambda)$

ML estimate: $\hat{\Lambda} = \arg\max_{\Lambda} p(C \mid \Lambda)$