Text Models 2

Internet Analytics (COM-308)

Prof. Matthias Grossglauser
School of Computer and Communication Sciences

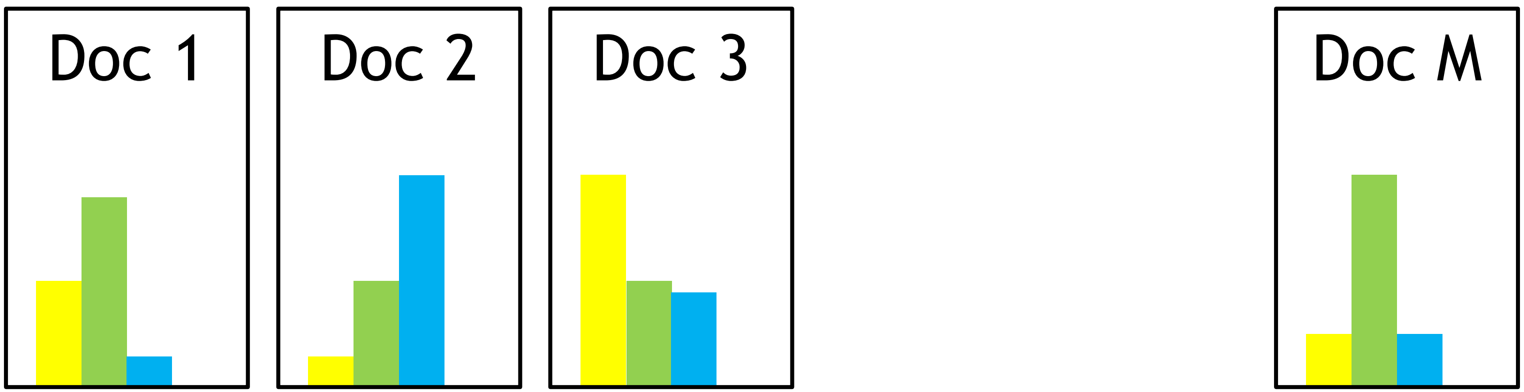
EPFL

Bayesian Networks: recap

- Defines a multivariate probability distribution
- Models direct causal influences
- In practice: as sparse as possible
- Conditional independence properties as graph (path) properties
- Inference:
  - Observe some variables (observables)
  - Obtain conditional distribution of some other variables of interest → estimate
  - Some variables we do not care about (latent)

Probabilistic topic models

- Unsupervised learning approach:
  - No topic labels given → topics are latent variables
- Each document has a topic distribution

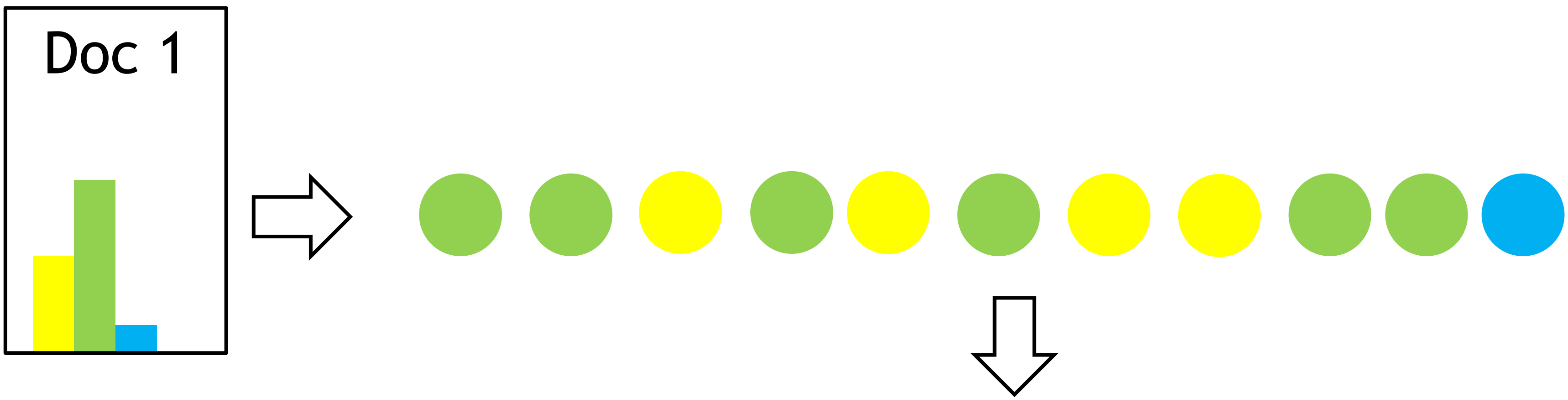


- Each topic has a word distribution

painting: 0.08	dna: 0.03	bieber: 0.05
exhibition: 0.05	gene: 0.02	arrest: 0.05
art: 0.04	transcript: 0.02	brawl: 0.02

Probabilistic topic models

- For each document, generate topic distribution
- For each term in document, generate topic
- For each term, generate word from that topic



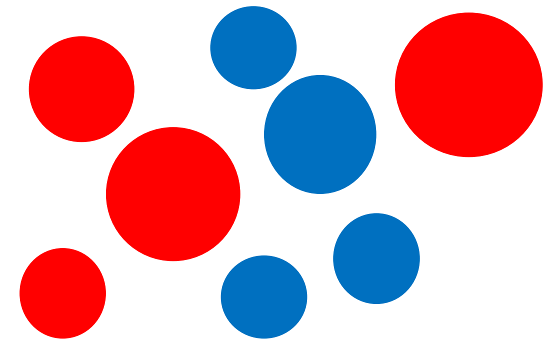
Lorem ipsum dolor sit amet consectetur adipisicing elit sed do eiusmod tempor incididunt ut labore et dolore magna aliqua Ut enim ad minim veniam quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur Excepteur sint occaecat cupidatat non proident sunt in culpa qui officia deserunt mollit anim id est laborum

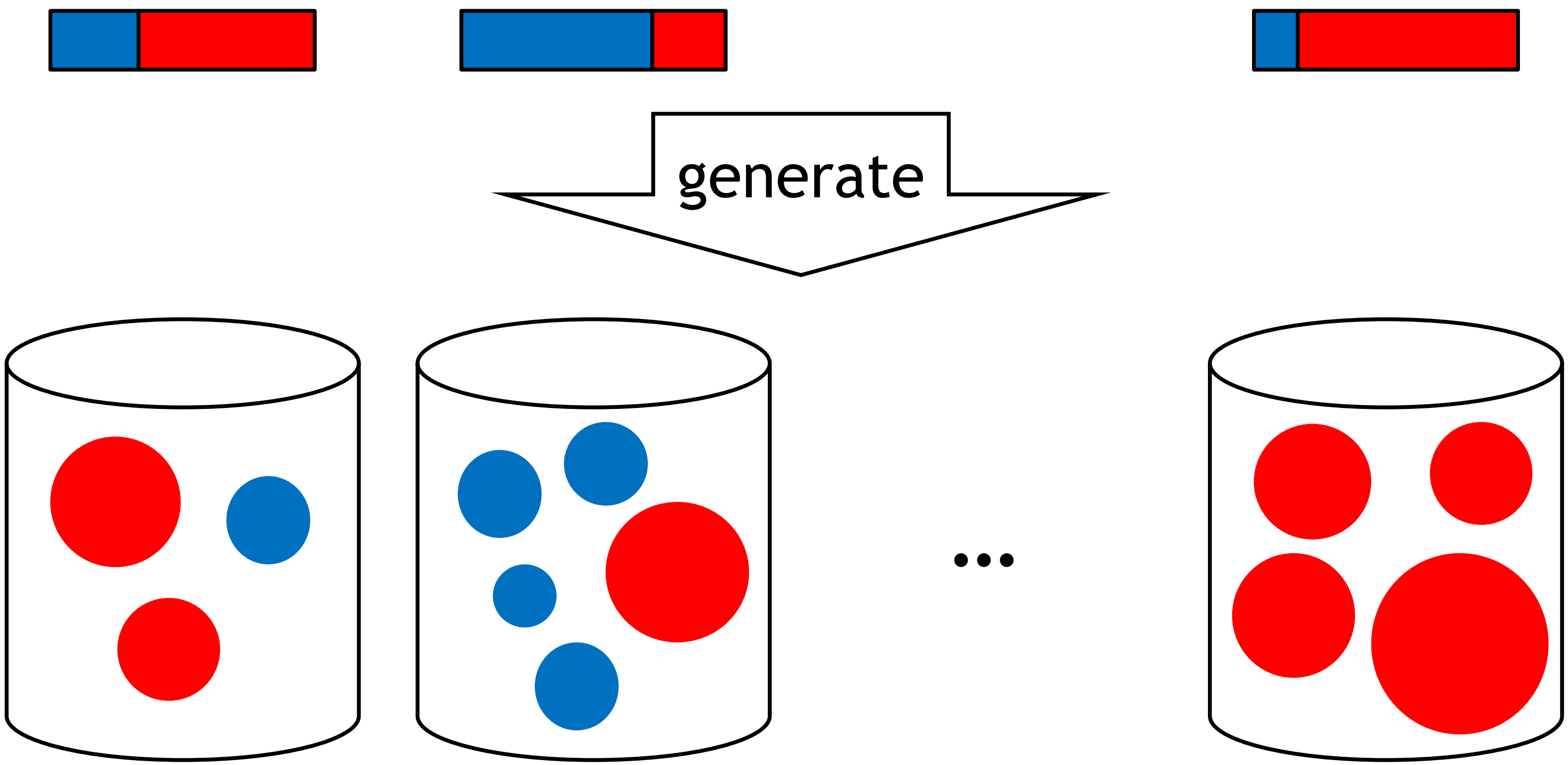
Approach 2: Probabilistic LSI (pLSI)

- Model:
  - M documents, each containing N words
  - Generating a word X:
    - $P(D, X) = P(X|D)P(D) =$
    - $= \sum_k P(X|Z = z_k, D)\, P(Z = z_k|D)P(D) =$
    - $= P(D) \sum_k P(X|Z = z_k)\, P(Z = z_k|D)$
  - D: document index
  - $P(X|Z)$: the distribution of words in each topic
    - E.g.: $P(''Bayes''|\text{topic=celebrity})$ is very low, $P(''bonus''|\text{topic=business})$ is higher
  - $P(Z|D)$: the distribution of topics for each document
    - E.g.: doc $D = 5$ is (80% celebrity, 20% business, 0% computer science)

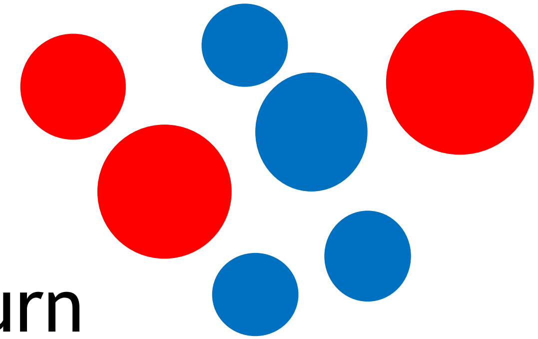
Urn with 2 classes of balls: generative

- Red balls and blue balls
- Size distribution: $\beta_{R,B} \sim N(\mu_{R,B}, \sigma^2)$
  - Red balls slightly larger on average

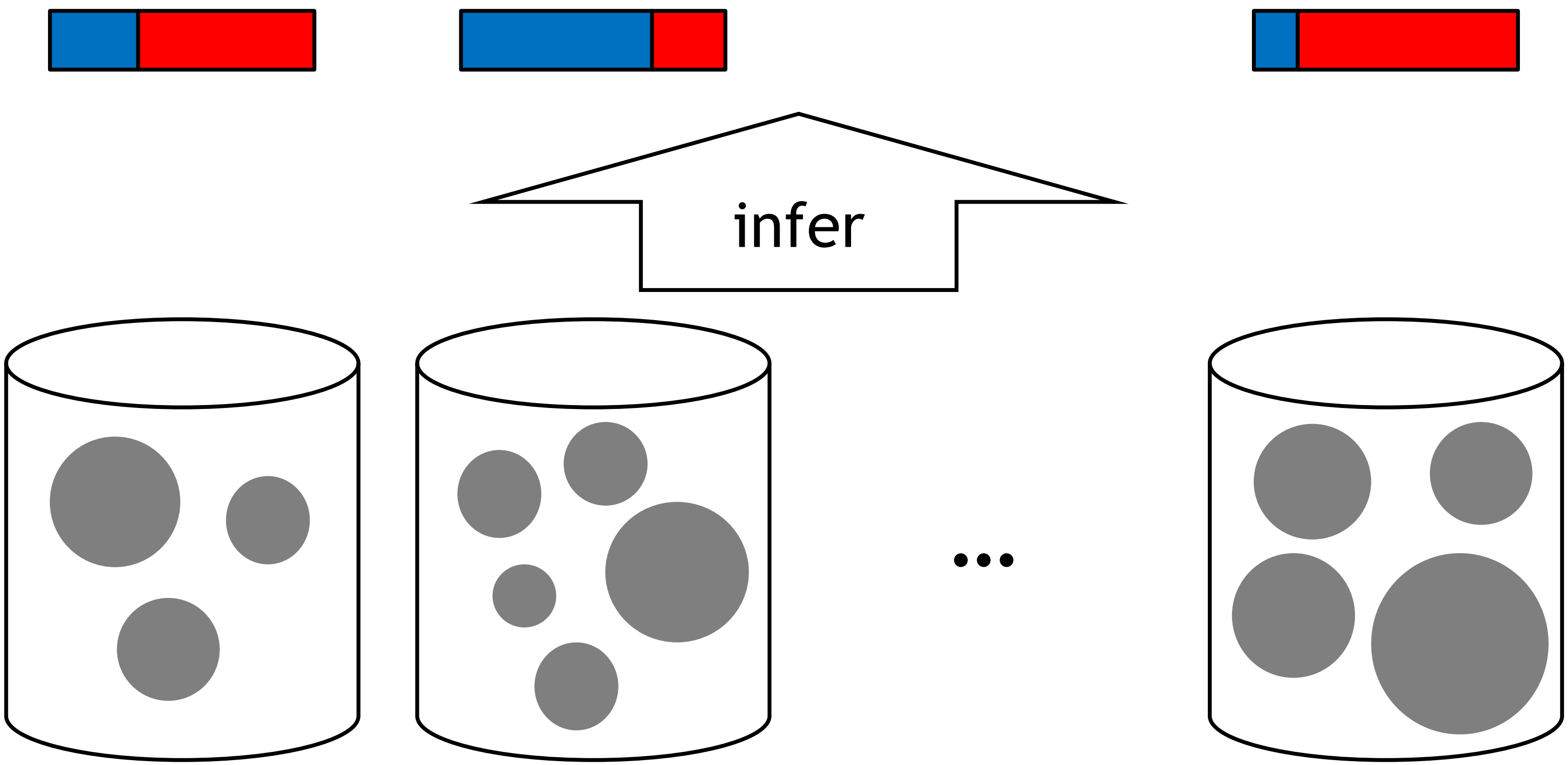




Urn with 2 classes of balls: inference

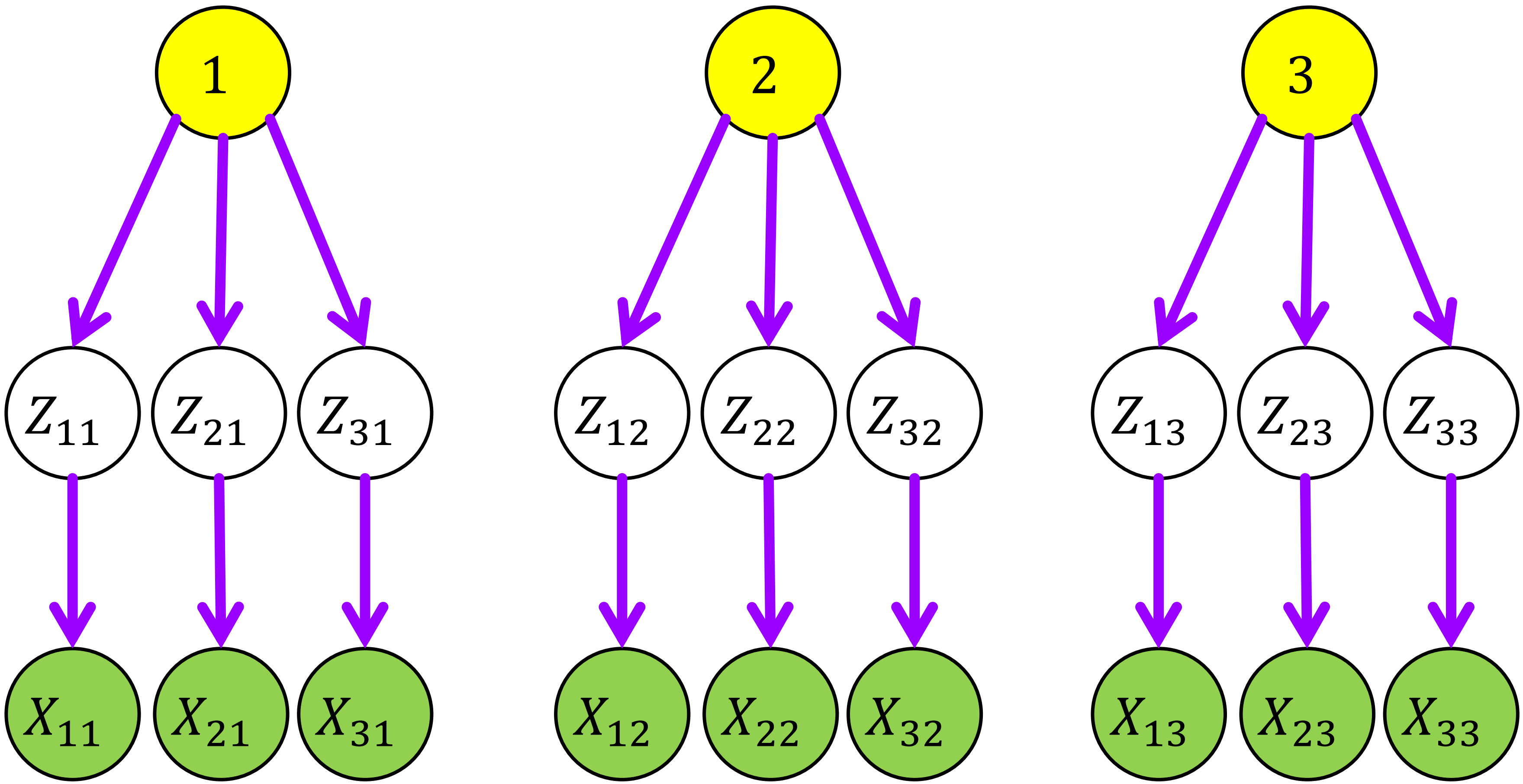


- We see sizes of balls in each urn
- But not color!
- Estimate fractions of red/blue per urn



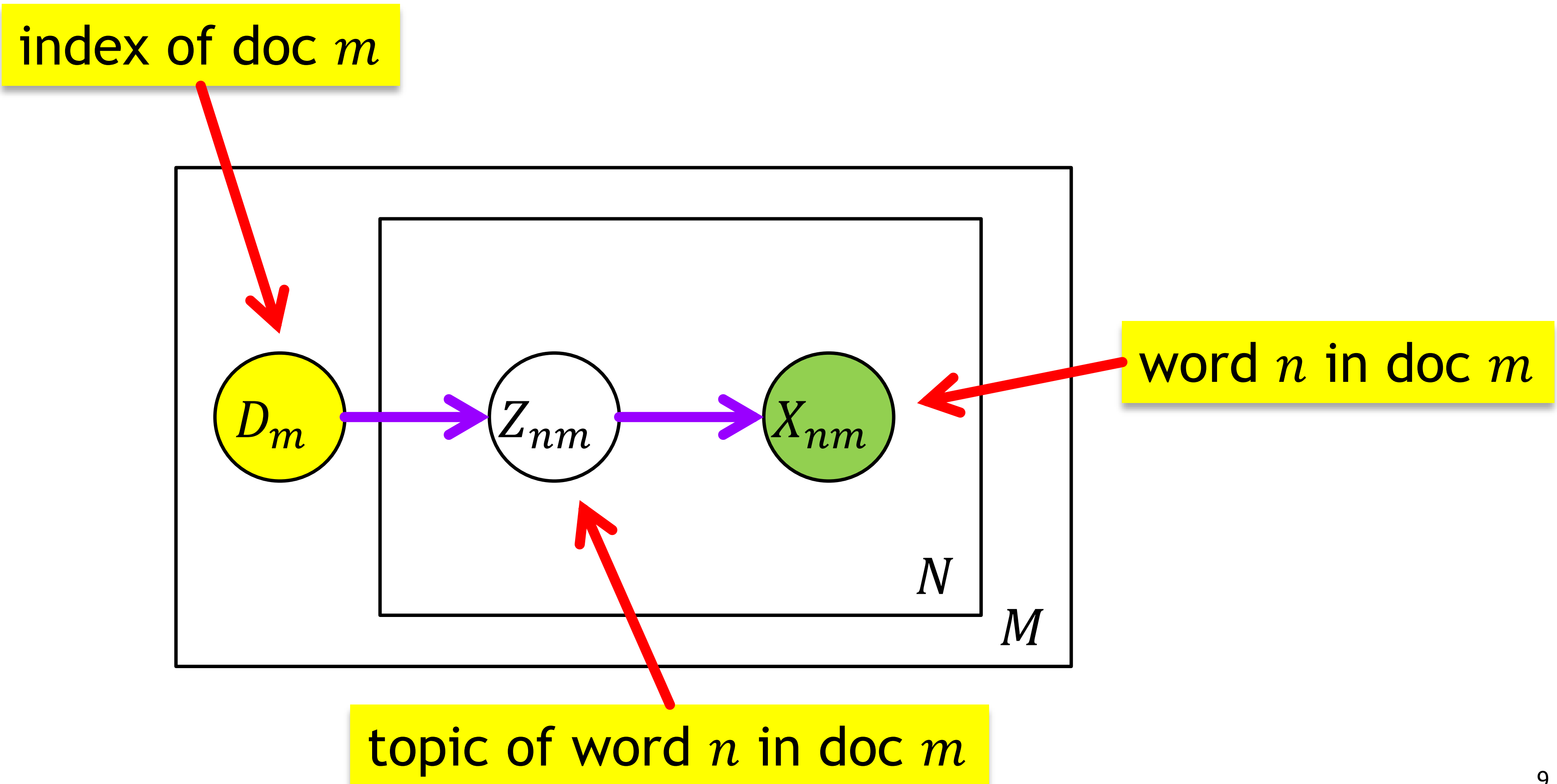
pLSI: topic mixture model

- Document: has topic distribution
- Word: has topic



pLSI: plate notation

- Document: has topic distribution
- Word: has topic

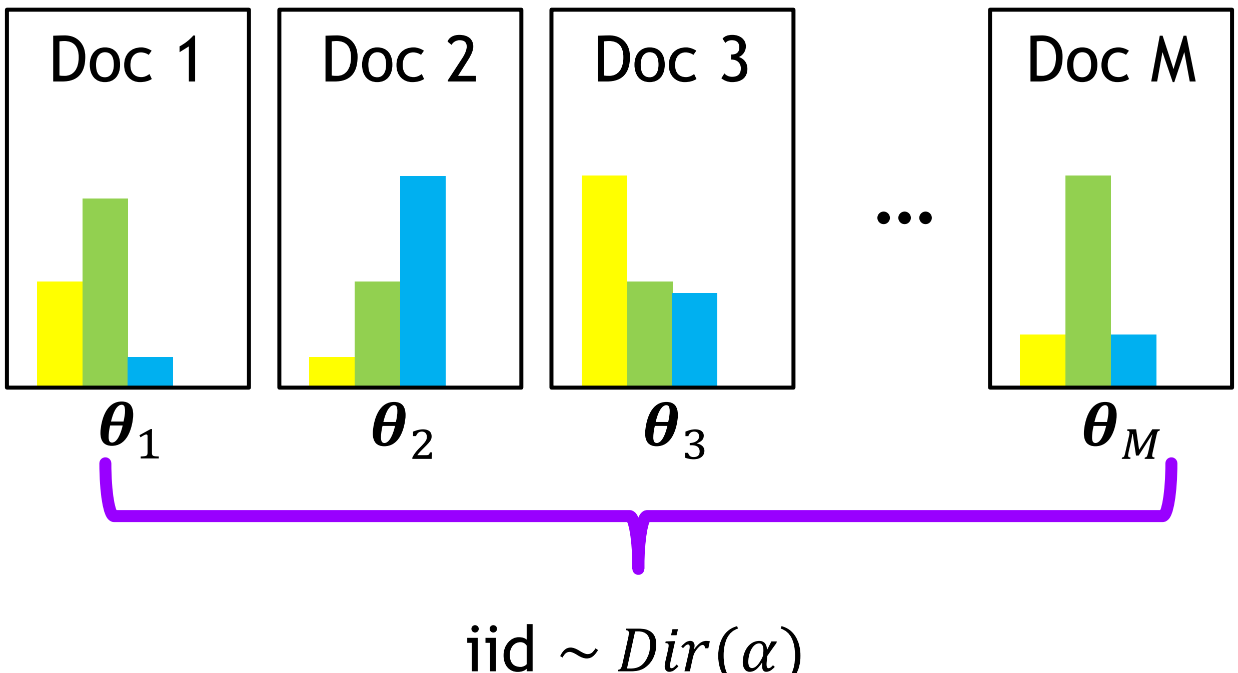


Critique of pLSI

- Flexible:
  - Each document is a mixture of topics
  - Each word can be generated by different topics
- Number of parameters:
  - $P(X|Z)$: KV params (K: # topics, V: size of vocabulary)
  - $P(Z|D)$: KM params (M: # documents)
- Linear growth in # documents M
  - Danger of overfitting!
- Inference:
  - Original paper: EM-algorithm to identify $P(Z|D)$ for each doc, and $P(X|Z)$ for each topic
- Not a fully generative model:
  - Can produce words for existing doc, but not new doc

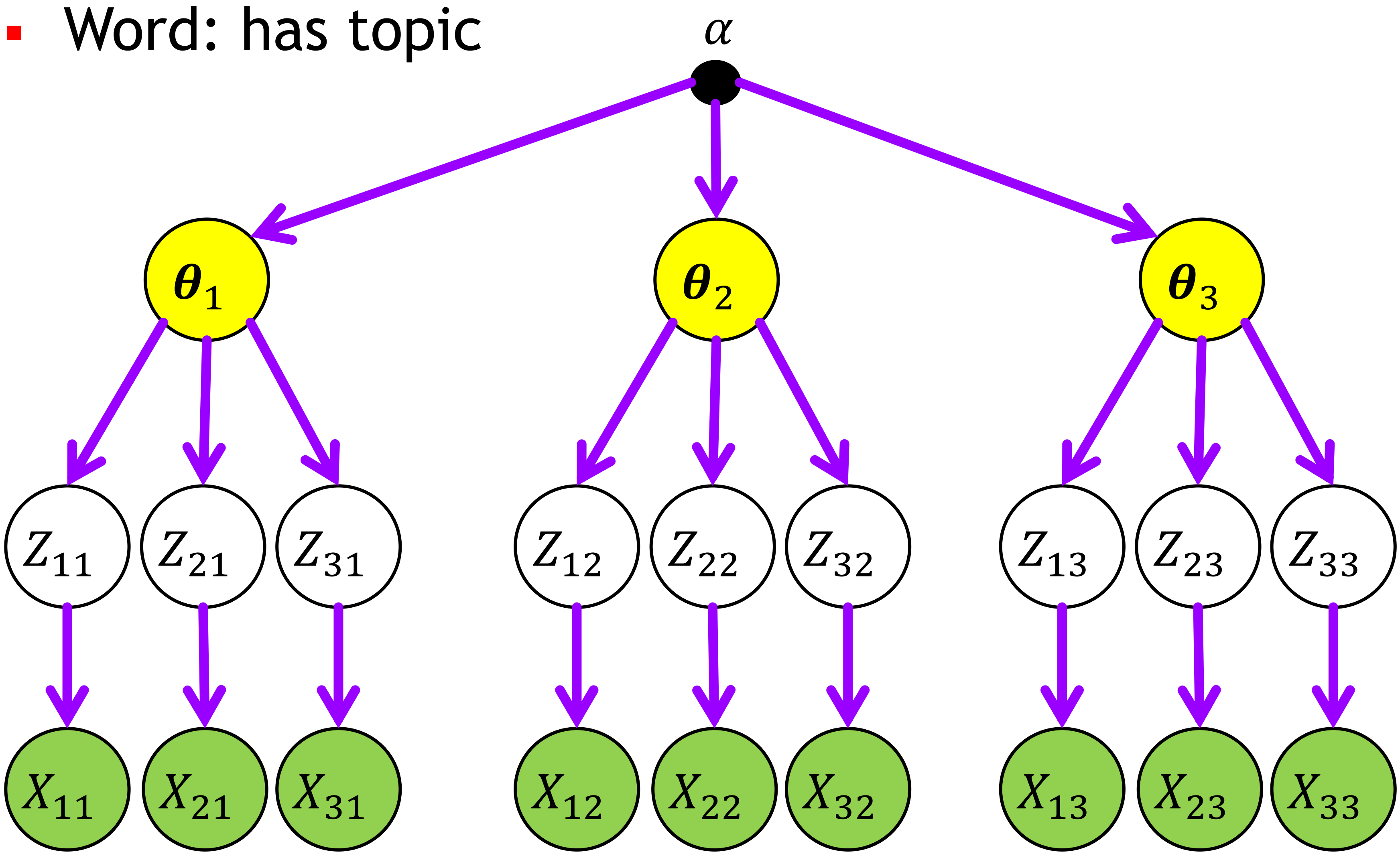
Approach 3: Latent Dirichlet Allocation (LDA)

- Drawback of pLSI:
  - Learns a separate $P(Z|D)$ for each document
- LDA model: one additional level
  - α: prior on topic distribution
  - β: hyperparameter on word distribution per topic



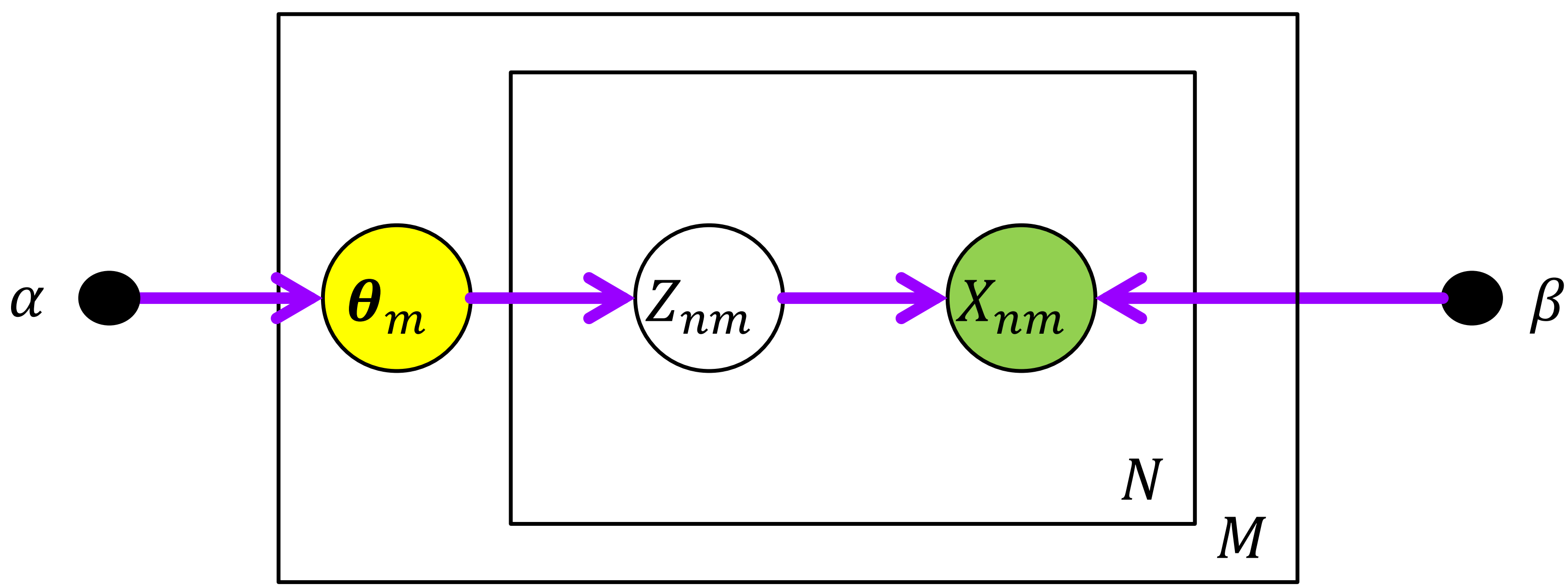
LDA

- Document: has topic distribution
- Word: has topic



LDA: plate notation

- Drawback of pLSI:
  - Learns a separate $P(Z|D)$ for each document
- LDA model: one additional level
  - α: prior on topic distribution
  - β: hyperparameter on word distribution per topic
    - $K \times V$ topic matrix

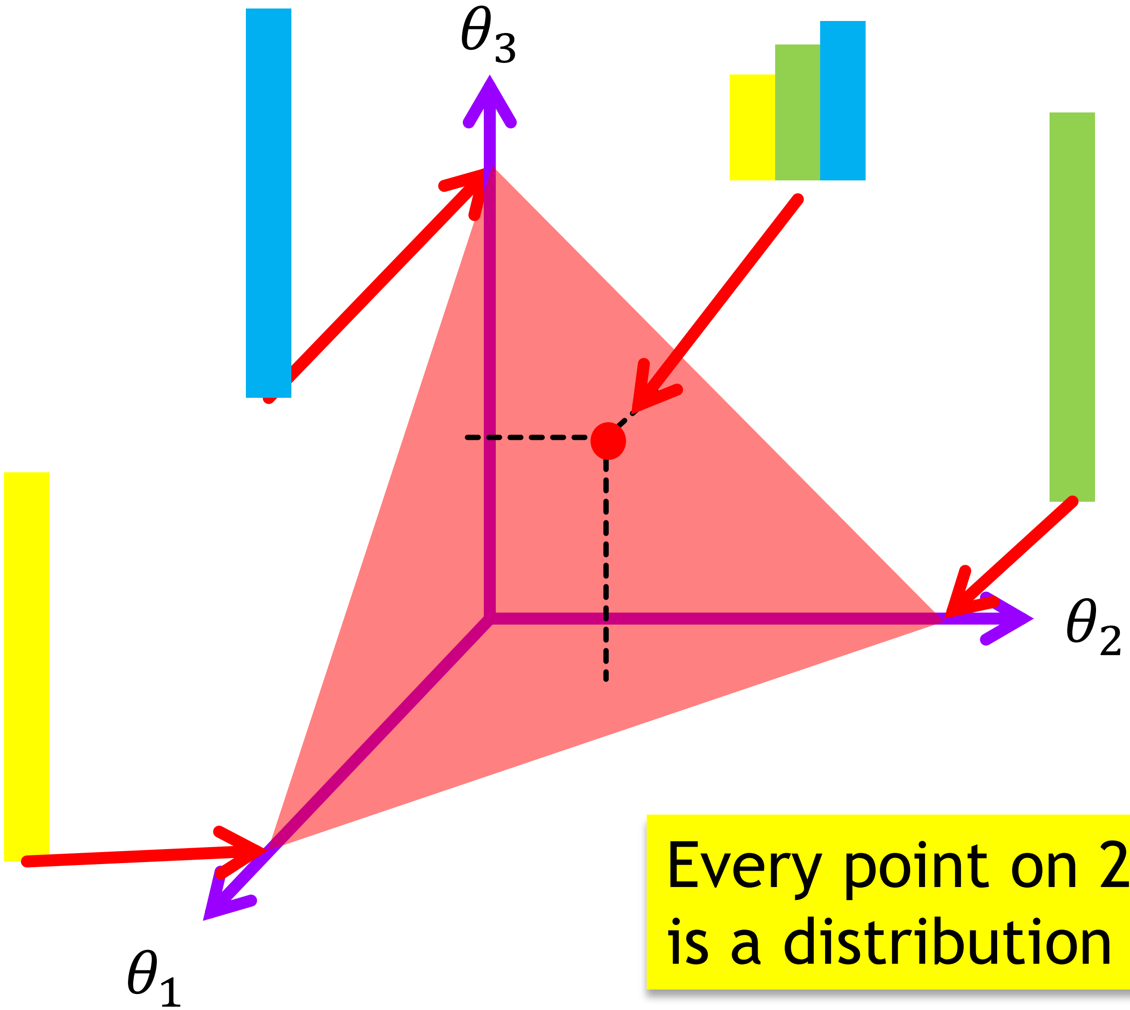


LDA

- Conceptual difference to pLSI:
  - pLSI: $P(Z|D)$ is separate for each document
    - A parameter to be learned for each document
  - LDA: topic distribution θ is itself sampled from $\mathrm{Dir}(\alpha)$
    - A latent (hidden) variable
- Dirichlet distribution: a distribution over distributions
  - $P(\boldsymbol{\theta}_m; \alpha) \propto \theta_{1m}^{\alpha_1 - 1} \theta_{2m}^{\alpha_2 - 1} \dots \theta_{Km}^{\alpha_K - 1}$
  - $\alpha_k > 0$
  - Normalized s.t. $\sum_k \theta_{km} = 1$ and $\theta_{km} > 0$, i.e., $\boldsymbol{\theta}_m$ can be viewed as a probability mass function
  - Lies on the $(K-1)$-simplex

Dirichlet: a distribution of distributions

- Every point on the ($K-1$ dim simplex) is a distribution over K values

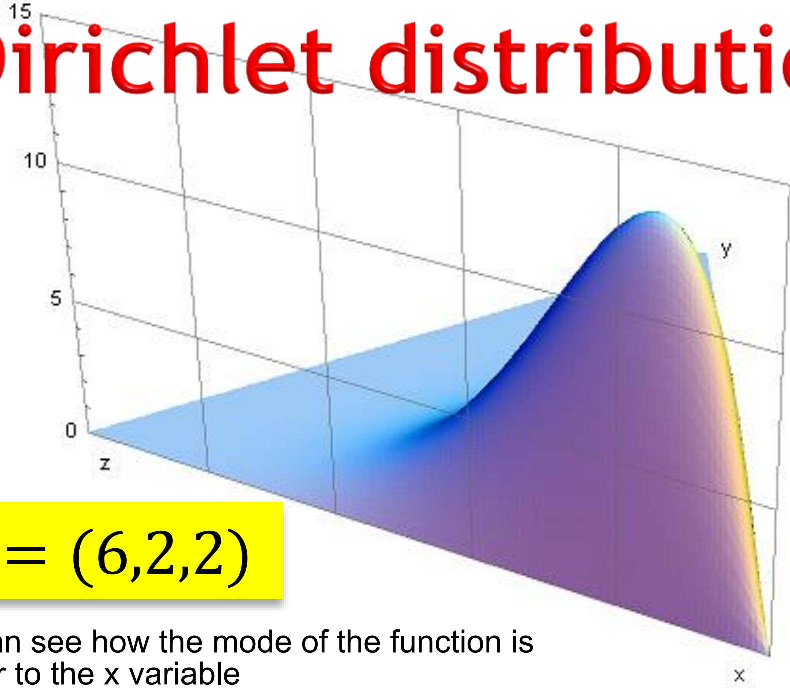


Every point on 2-simplex
is a distribution over 3 topics

Dirichlet: a distribution of distributions

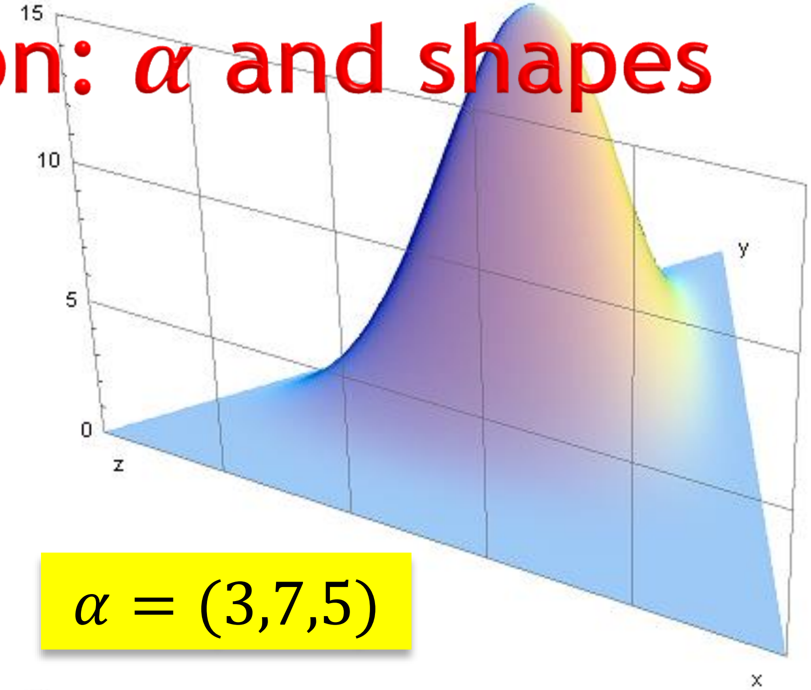
- Dirichlet distribution: $\theta \sim \mathrm{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_K)$
  - $\theta_1, \ldots, \theta_K > 0$
  - $\theta_1 + \cdots + \theta_K = 1$ → a realization can be viewed as a distribution over $(1, \ldots, K)$
  - $P(\theta; \alpha) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_K^{\alpha_K - 1}$
- Parameter α determines mode:
  - $E[\theta_i] = \frac{\alpha_i}{\sum \alpha_i}$

Dirichlet distribution: α and shapes

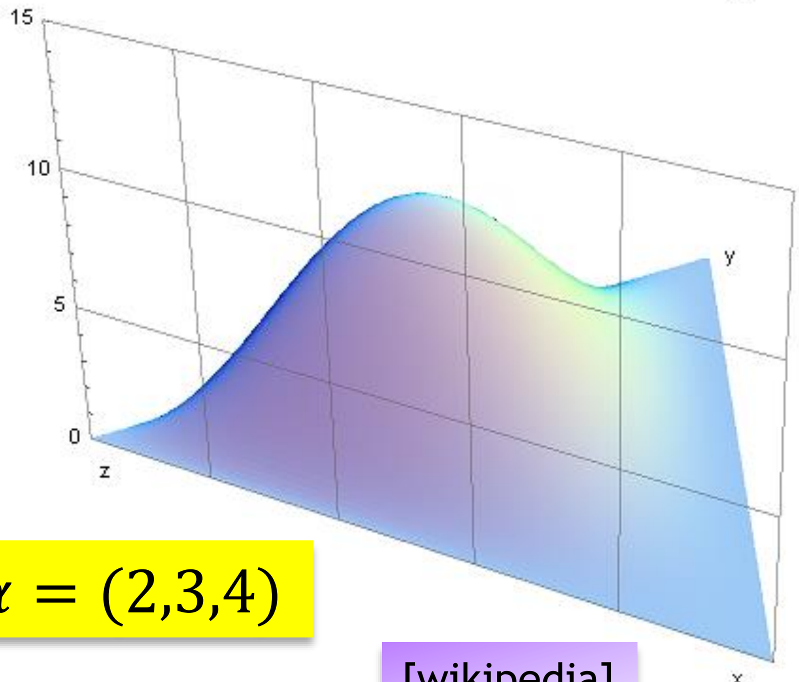


$\alpha = (6,2,2)$

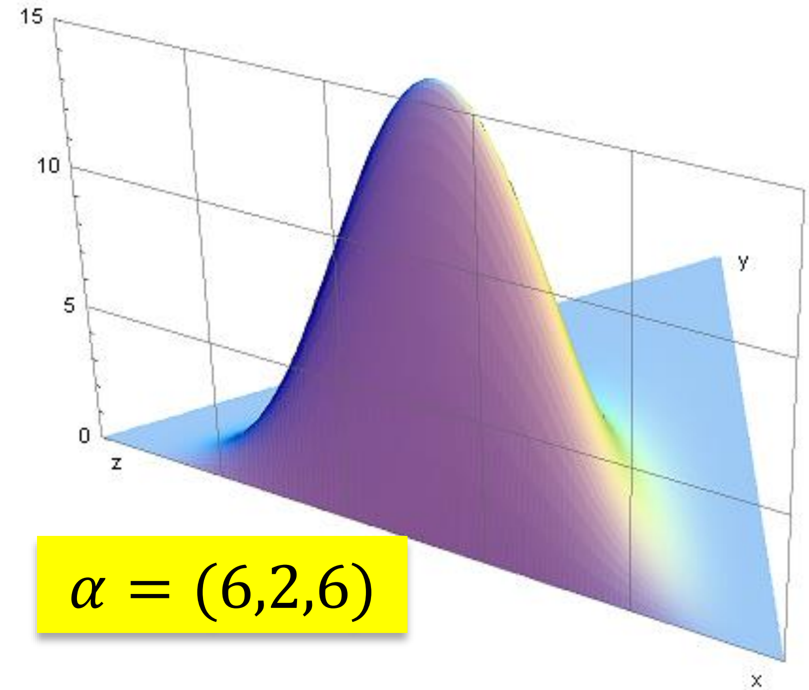
we can see how the mode of the function is
closer to the x variable



$\alpha = (3,7,5)$



$\alpha = (2,3,4)$

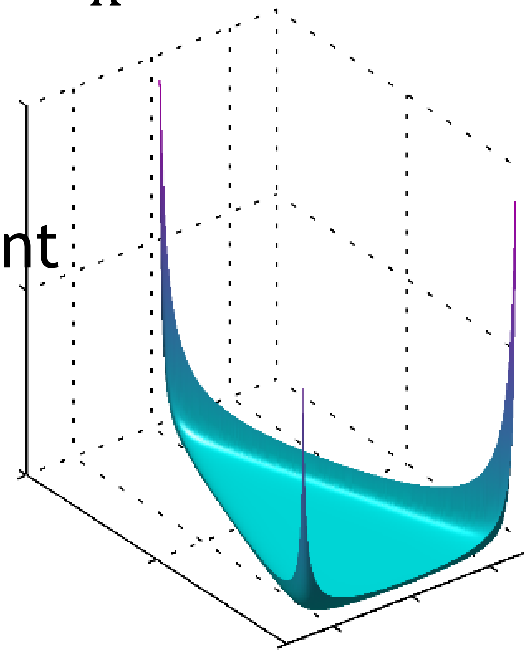


$\alpha = (6,2,6)$

[wikipedia]

Dirichlet distribution: samples

- Symmetric Dirichlet: $\alpha_1 = \alpha_2 = \cdots = \alpha_K = \alpha$
- α controls uniformity/sparsity of topic vectors
  - $0 < \alpha < 1$ → tendency to map document to small set of dominant topics (small posterior in the middle + "peaks" (modes) in the corners)



  - $\alpha \cong 1$ → tendency towards uniform topic posterior
- In particular:
  - $\alpha = 1$: $\theta \sim$ uniform(simplex)
  - $\alpha = \infty$: $\theta =$ constant =uniform topic distribution

LDA: summary

- Parameters:
  - α: K params
  - β: KV params
  - Number of parameters independent of M! Overfitting naturally controlled in reasonably large corpus
- LDA:
  - Currently best-performing approach for unsupervised soft document clustering
  - General idea applied to many related scenarios (e.g., dynamic topic models)
- Inference:
  - MCMC (Gibbs sampling), or Variational Bayesian inference (generalization of EM)
  - Active area of research

LDA examples from data

- Corpus:
  - TREC-AP: Associated Press newswire stories
  - Approx. 16k documents, with 23k word vocabulary (after stemming etc.)
  - Up to 100 topics
- Ref: [D. Blei, A. Ng, M. Jordan: Latent Dirichlet Allocation, JMLR, 2003]

LDA examples from data: top β_{k*}

"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

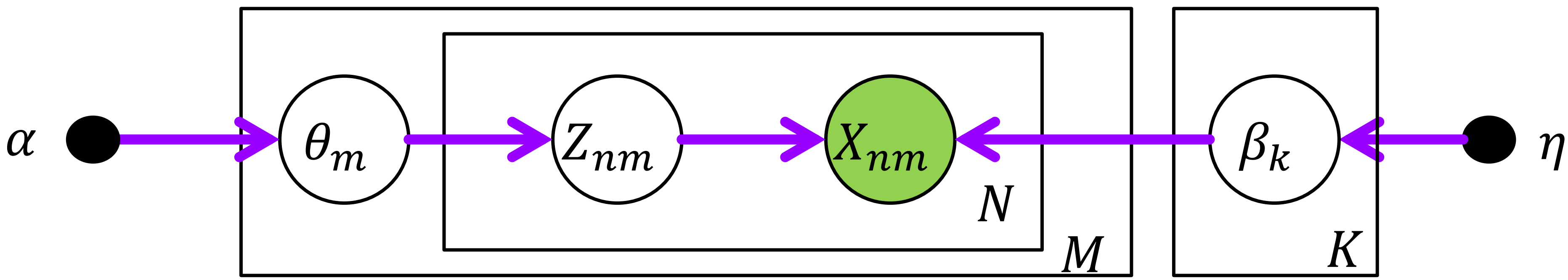
LDA examples from data: posterior Z_{nm}

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.

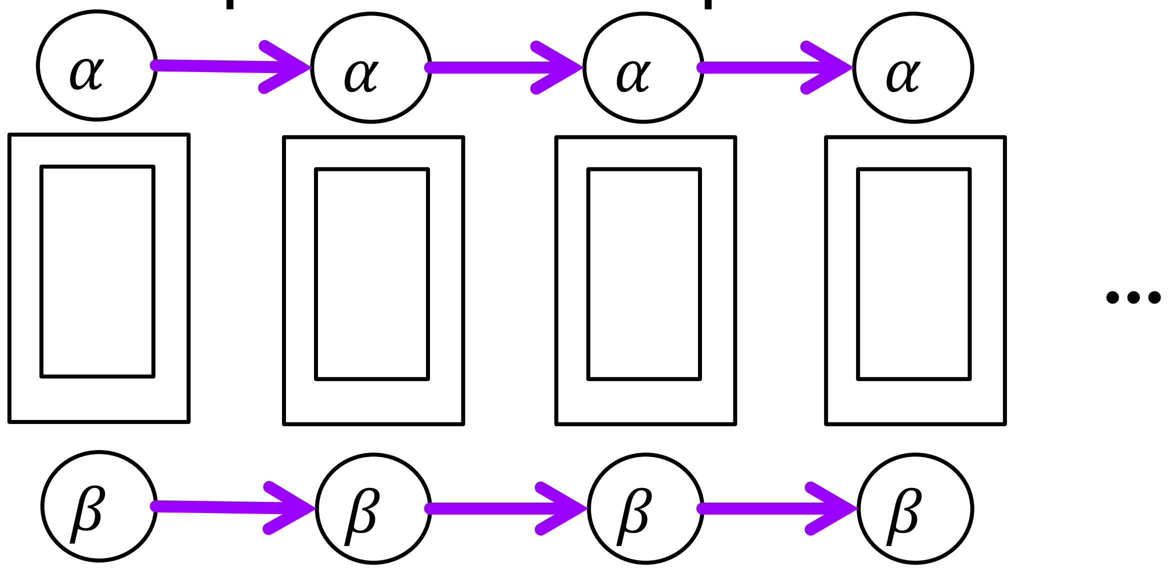
[D. Blei, A. Ng, M. Jordan: Latent Dirichlet Allocation, JMLR, 2003]

LDA variants and extensions

- Dirichlet prior on topic-word distribution

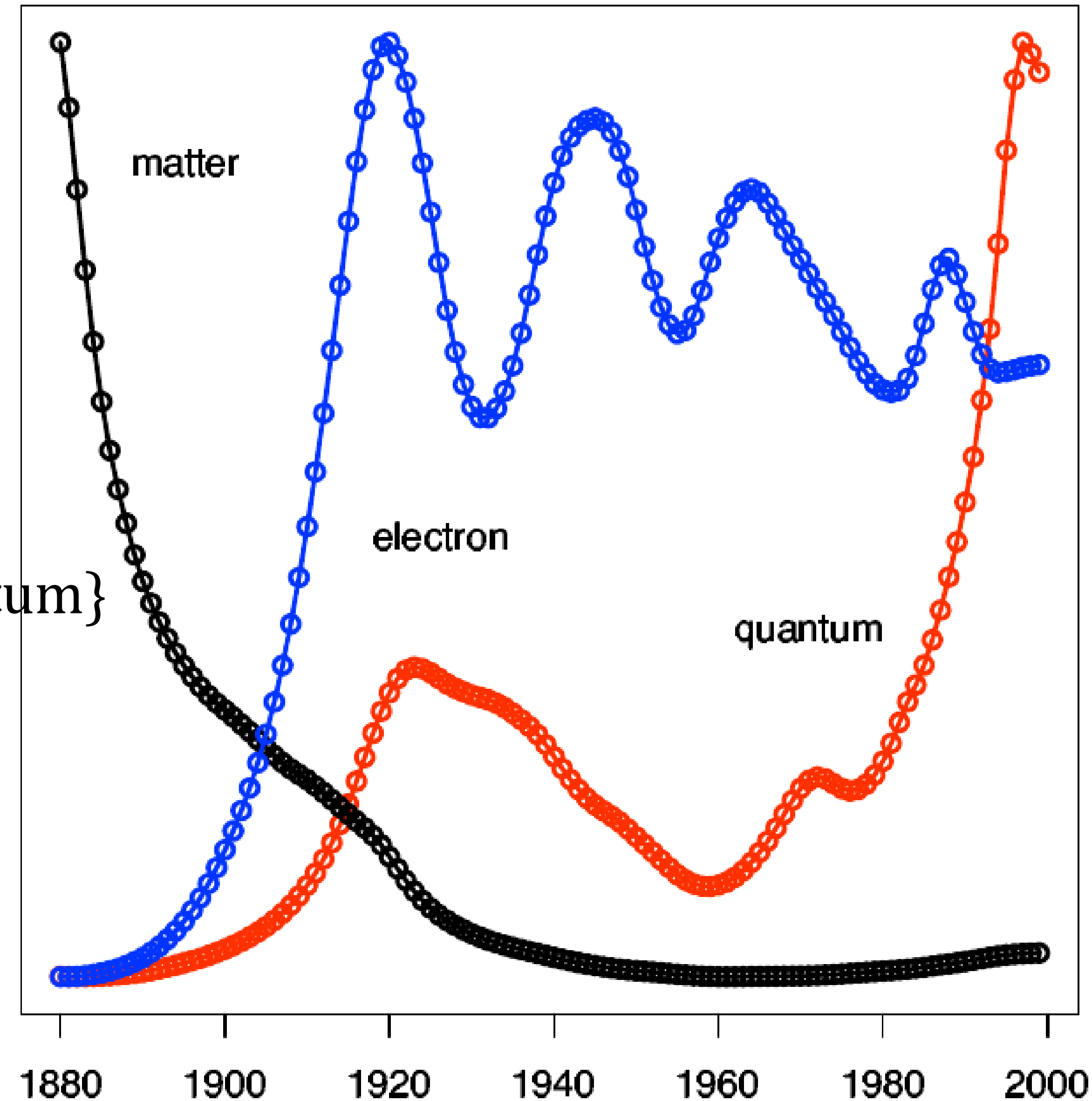


- Dynamic topic models: topics over time



LDA variants and extensions

Evolution of β_{kx} over time, for k =electrical eng., $x \in \{\text{matter}, \text{electron}, \text{quantum}\}$



[D. Blei, J. Lafferty: Dynamic Topic Models, CMU/CS Tech Report, 2006]

Summary

- Topic models:
  - Find clusters of documents that are similar in meaning
  - Main challenges: synonymy + homonymy; topics not sharp
- Applications:
  - Classification; automatic organization of corpus
  - Retrieval: θ as doc descriptor instead of TF-IDF etc.
- LSI: SVD of term frequency matrix
  - Topics are "kept apart" by orthogonality
- pLSI: Each doc is a mixture of topics
- LDA: Topic distribution (mix coeff) is itself a latent variable

References

- [D. Koller, N. Friedman: Probabilistic Graphical Models, MIT Press, 2009]
- [D. Blei, A. Ng, M. Jordan: Latent Dirichlet Allocation, JMLR, 2003]
- [Ch. D. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval, Cambridge, 2008]
- [C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006]