# Introduction To Machine Learning

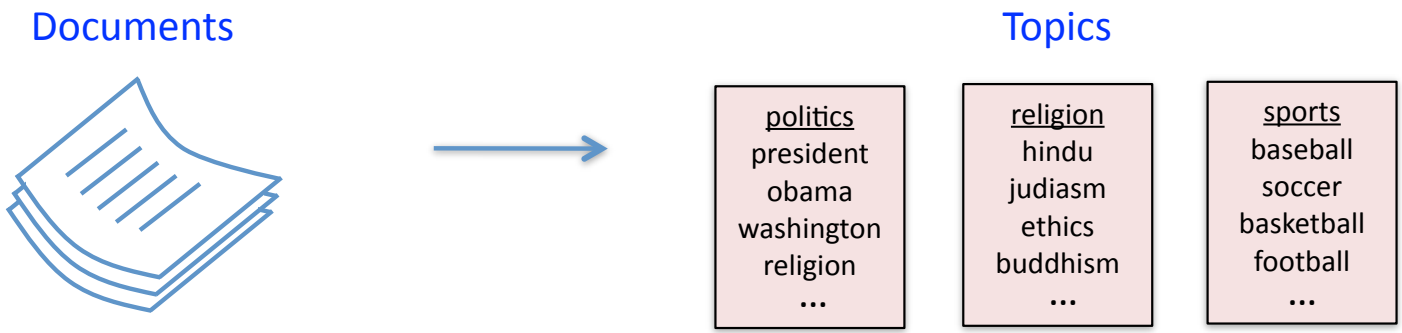
David Sontag

New York University

Lecture 22, April 19, 2016

# Latent Dirichlet allocation (LDA)

- **Topic models** are powerful tools for exploring large data sets and for making inferences about the content of documents




- Many applications in information retrieval, document summarization, and classification


- LDA is one of the simplest and most widely used topic models

# Generative model for a document in LDA

1. Sample the document's **topic distribution** $\theta$ (aka topic vector)

$$\theta \sim \text{Dirichlet}(\alpha_{1:T})$$

where the $\{\alpha_t\}_{t=1}^T$ are fixed hyperparameters. Thus $\theta$ is a distribution over $T$ topics with mean $\theta_t = \alpha_t / \sum_{t'} \alpha_{t'}$

2. For $i = 1$ to $N$, sample the **topic** $z_i$ of the $i$'th word

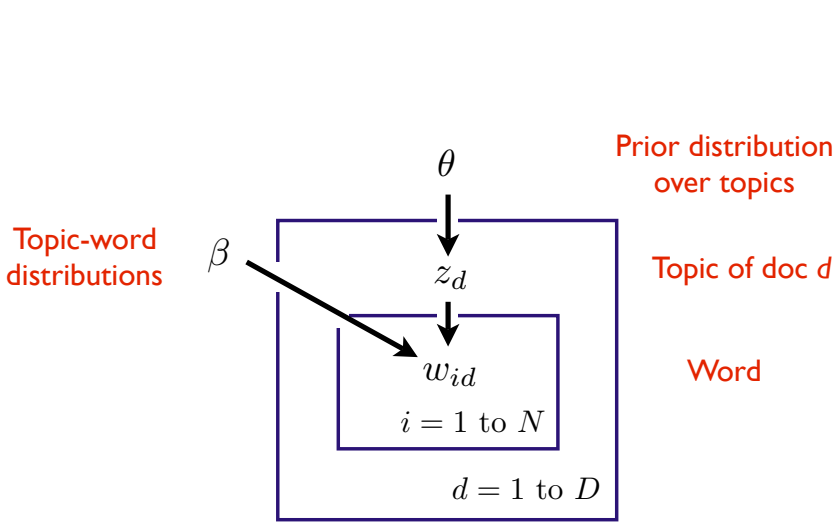
$$z_i | \theta \sim \theta$$

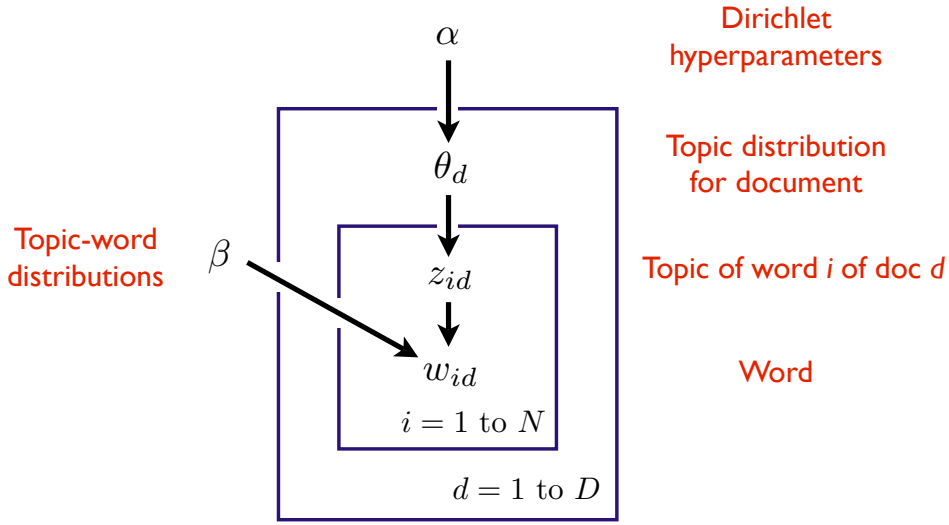
3. ... and then sample the actual **word** $w_i$ from the $z_i$'th topic

$$w_i | z_i \sim \beta_{z_i}$$

where $\{\beta_t\}_{t=1}^T$ are the *topics* (a fixed collection of distributions on words)

# Comparison of mixture and admixture models




- Model on left is a **mixture model**
  - Called *multinomial* naive Bayes (a word can appear multiple times)
  - Document is generated from a single topic
- Model on right (LDA) is an **admixture model**
  - Document is generated from a distribution over topics

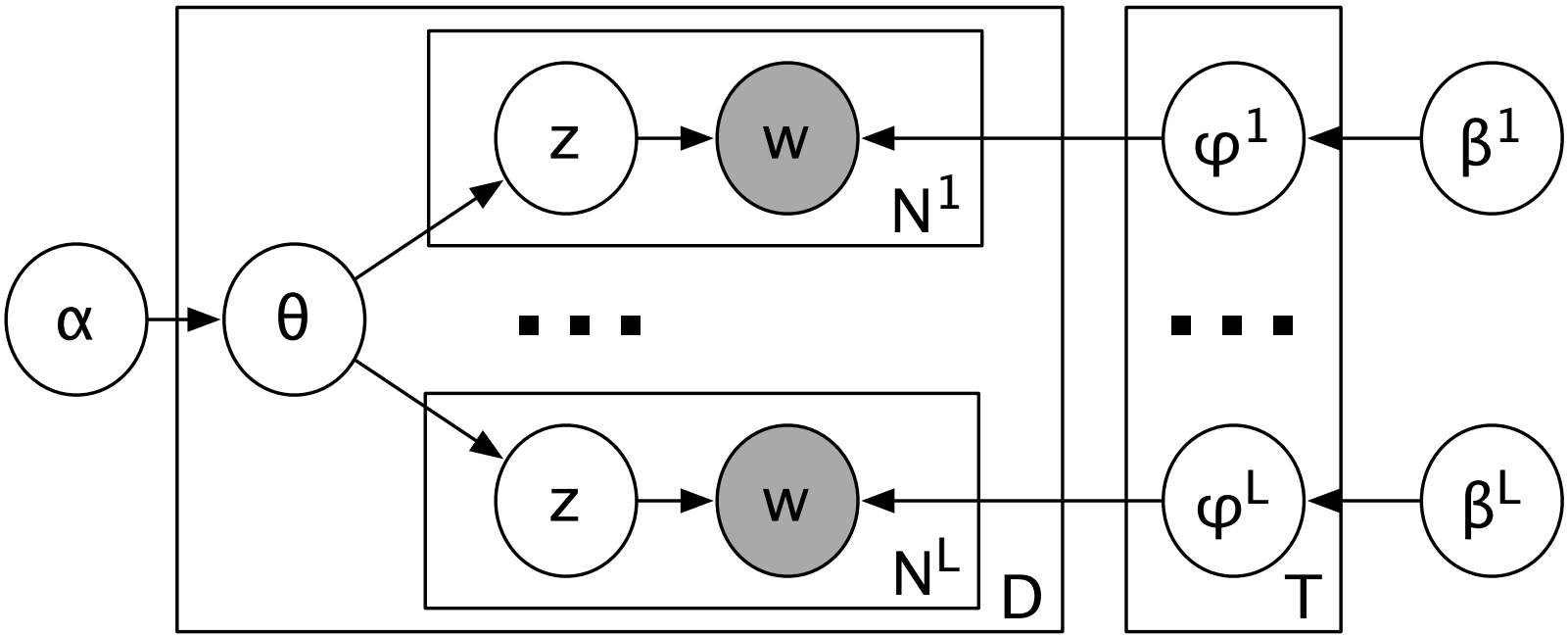
# Two steps

- Can typically separate out these two uses of topic models:
  1. *Learn* the model parameters $(\alpha, \beta)$
  2. Use model to make *inferences* about a single document
- Step 1 is when topic discovery happens. Since the topic assignments $z$ are never observed, one can use EM to do this
- Exact inference is intractable: approximate inference (typically Gibbs sampling) is used

# Polylingual topic models (Mimno et al., EMNLP '09)

- Goal: topic models that are aligned across languages
- Training data: corpora with multiple documents in each language
  - EuroParl corpus of parliamentary proceedings (11 western languages; exact translations)
  - Wikipedia articles (12 languages; not exact translations)
- How to do this?

# Polylingual topic models (Mimno et al., EMNLP '09)



# Learned topics

| | |
|---|---|
| DA | centralbank europæiske ecb s lån centralbanks |
| DE | zentralbank ezb bank europäischen investitionsbank darlehen |
| EL | τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες |
| EN | **bank central ecb banks european monetary** |
| ES | banco central europeo bce bancos centrales |
| FI | keskuspankin ekp n euroopan keskuspankki eip |
| FR | banque centrale bce européenne banques monétaire |
| IT | banca centrale bce europea banche prestiti |
| NL | bank centrale ecb europese banken leningen |
| PT | banco central europeu bce bancos empréstimos |
| SV | centralbanken europeiska ecb centralbankens s lån |

# Learned topics

| | |
|---|---|
| DA | børn familie udnyttelse børns børnene seksuel |
| DE | kinder kindern familie ausbeutung familien eltern |
| EL | παιδιά παιδιών οικογένεια οικογένειας γονείς παιδικής |
| EN | **children family child sexual families exploitation** |
| ES | niños familia hijos sexual infantil menores |
| FI | lasten lapsia lapset perheen lapsen lapsiin |
| FR | enfants famille enfant parents exploitation familles |
| IT | bambini famiglia figli minori sessuale sfruttamento |
| NL | kinderen kind gezin seksuele ouders familie |
| PT | crianças família filhos sexual criança infantil |
| SV | barn barnen familjen sexuellt familj utnyttjande |

# Discussion

- How would you use this?
- How could you extend this?

# Author-topic model (Rosen-Zvi et al., UAI '04)

- Goal: topic models that take into consideration author *interests*
- Training data: corpora with label for who wrote each document
  - Papers from NIPS conference from 1987 to 1999
  - Twitter posts from US politicians
- Why do this?
- How to do this?

# Author-topic model (Rosen-Zvi et al., UAI '04)

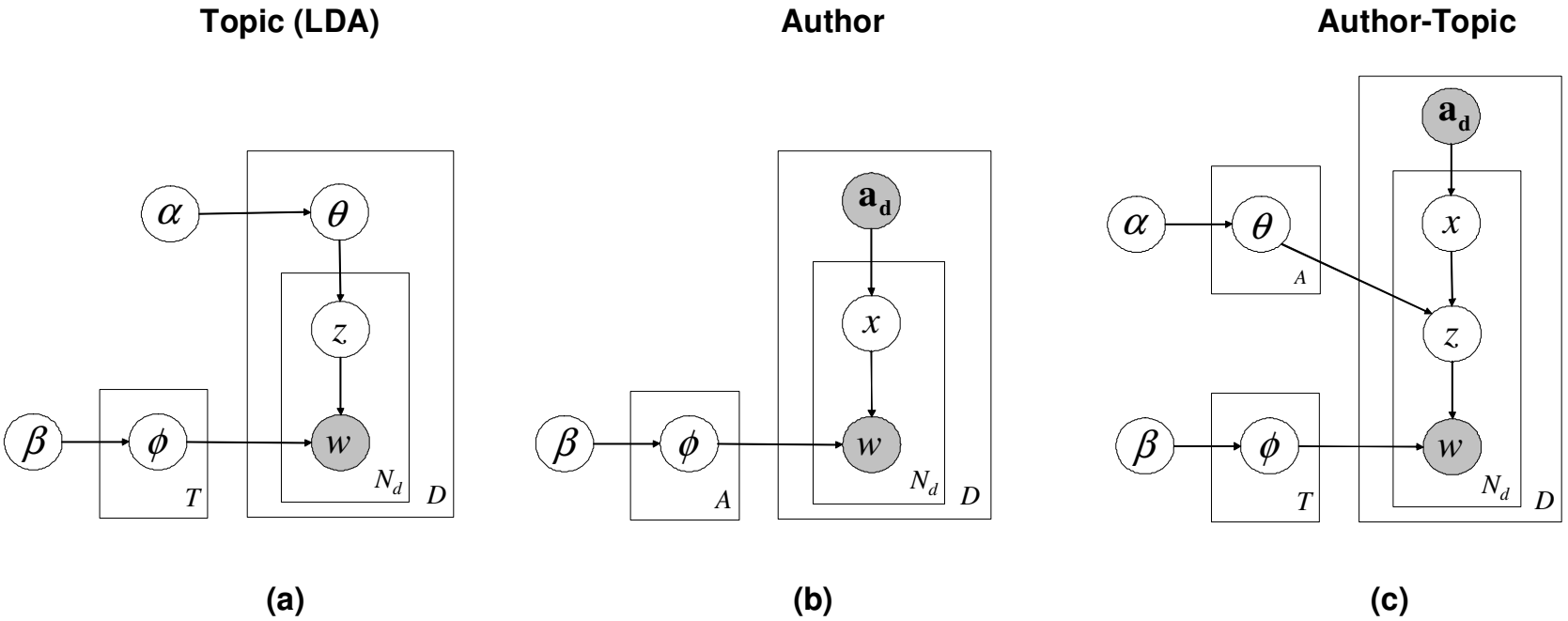



Figure 1: Generative models for documents. (a) Latent Dirichlet Allocation (LDA; Blei et al., 2003), a topic model. (b) An author model. (c) The author-topic model.

# Most likely author for a topic

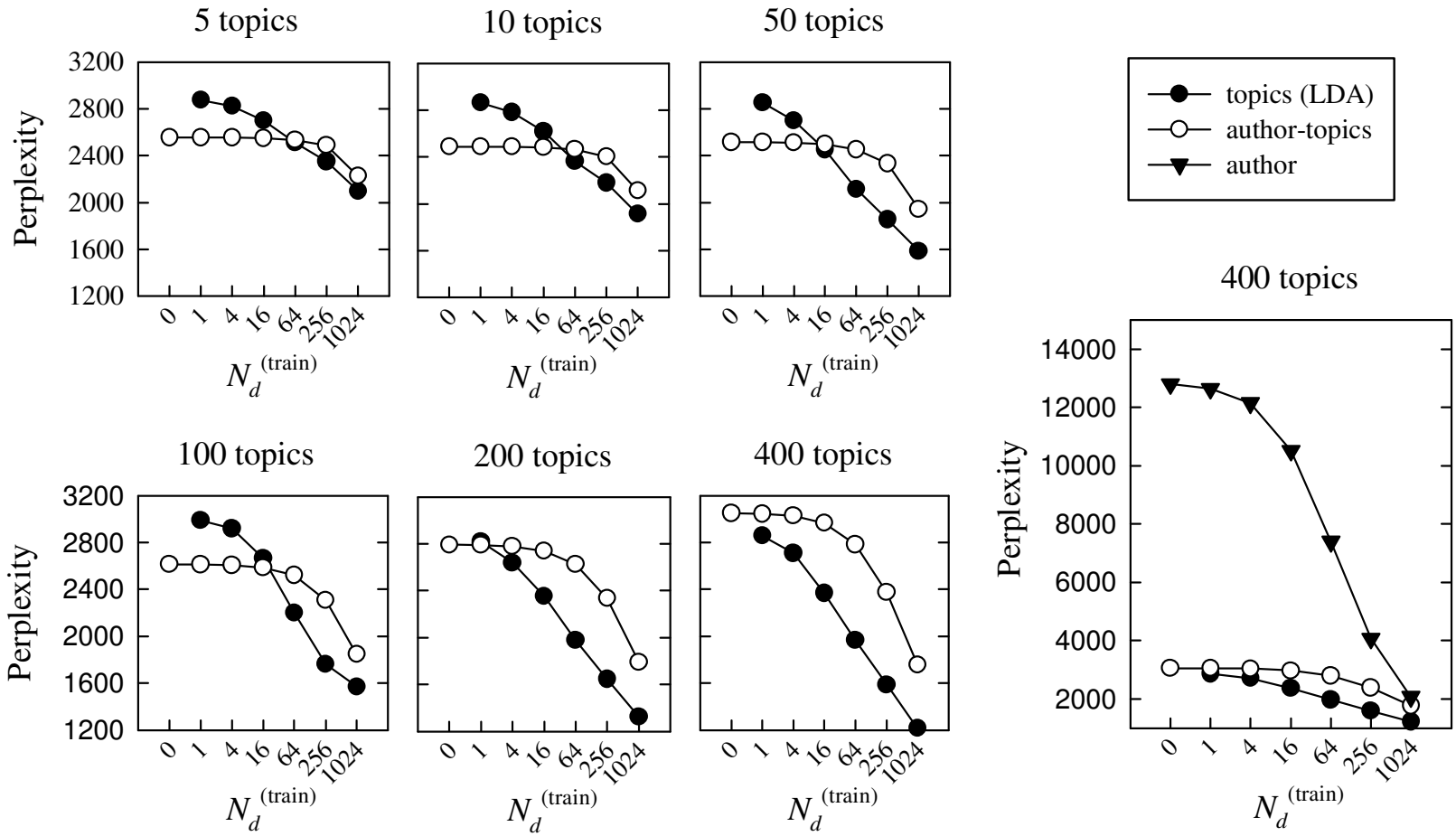
| TOPIC 31 | |
|---|---|
| **WORD** | **PROB.** |
| SPEECH | 0.0823 |
| RECOGNITION | 0.0497 |
| HMM | 0.0234 |
| SPEAKER | 0.0226 |
| CONTEXT | 0.0224 |
| WORD | 0.0166 |
| SYSTEM | 0.0151 |
| ACOUSTIC | 0.0134 |
| PHONEME | 0.0131 |
| CONTINUOUS | 0.0129 |
| **AUTHOR** | **PROB.** |
| Waibel_A | 0.0936 |
| Makhoul_J | 0.0238 |
| De-Mori_R | 0.0225 |
| Bourlard_H | 0.0216 |
| Cole_R | 0.0200 |
| Rigoll_G | 0.0191 |
| Hochberg_M | 0.0176 |
| Franco_H | 0.0163 |
| Abrash_V | 0.0157 |
| Movellan_J | 0.0149 |

| TOPIC 61 | |
|---|---|
| **WORD** | **PROB.** |
| BAYESIAN | 0.0450 |
| GAUSSIAN | 0.0364 |
| POSTERIOR | 0.0355 |
| PRIOR | 0.0345 |
| DISTRIBUTION | 0.0259 |
| PARAMETERS | 0.0199 |
| EVIDENCE | 0.0127 |
| SAMPLING | 0.0117 |
| COVARIANCE | 0.0117 |
| LOG | 0.0112 |
| **AUTHOR** | **PROB.** |
| Bishop_C | 0.0563 |
| Williams_C | 0.0497 |
| Barber_D | 0.0368 |
| MacKay_D | 0.0323 |
| Tipping_M | 0.0216 |
| Rasmussen_C | 0.0215 |
| Opper_M | 0.0204 |
| Attias_H | 0.0155 |
| Sollich_P | 0.0143 |
| Schottky_B | 0.0128 |

| TOPIC 71 | |
|---|---|
| **WORD** | **PROB.** |
| MODEL | 0.4963 |
| MODELS | 0.1445 |
| MODELING | 0.0218 |
| PARAMETERS | 0.0205 |
| BASED | 0.0116 |
| PROPOSED | 0.0103 |
| OBSERVED | 0.0100 |
| SIMILAR | 0.0083 |
| ACCOUNT | 0.0069 |
| PARAMETER | 0.0068 |
| **AUTHOR** | **PROB.** |
| Omohundro_S | 0.0088 |
| Zemel_R | 0.0084 |
| Ghahramani_Z | 0.0076 |
| Jordan_M | 0.0075 |
| Sejnowski_T | 0.0071 |
| Atkeson_C | 0.0070 |
| Bower_J | 0.0066 |
| Bengio_Y | 0.0062 |
| Revow_M | 0.0059 |
| Williams_C | 0.0054 |

| TOPIC 100 | |
|---|---|
| **WORD** | **PROB.** |
| HINTON | 0.0329 |
| VISIBLE | 0.0124 |
| PROCEDURE | 0.0120 |
| DAYAN | 0.0114 |
| UNIVERSITY | 0.0114 |
| SINGLE | 0.0111 |
| GENERATIVE | 0.0109 |
| COST | 0.0106 |
| WEIGHTS | 0.0105 |
| PARAMETERS | 0.0096 |
| **AUTHOR** | **PROB.** |
| Hinton_G | 0.2202 |
| Zemel_R | 0.0545 |
| Dayan_P | 0.0340 |
| Becker_S | 0.0266 |
| Jordan_M | 0.0190 |
| Mozer_M | 0.0150 |
| Williams_C | 0.0099 |
| de-Sa_V | 0.0087 |
| Schraudolph_N | 0.0078 |
| Schmidhuber_J | 0.0056 |

# Perplexity as a function of number of observed words




$$\text{perplexity}(\mathbf{w}_{test,d} \mid \mathbf{w}_{train,d}, \mathbf{a}_d) = \exp\left[-\frac{\ln p(\mathbf{w}_{test,d}|\mathbf{w}_{train,d},\mathbf{a}_d)}{N_{test,d}}\right]$$

# Supervised Topic Models

- The inferred $\theta$ or $\mathbf{z}$ can be used as features in many prediction tasks.
- Performance can be improved by jointly training the representation and the predictor.
- Hence, supervised LDA: