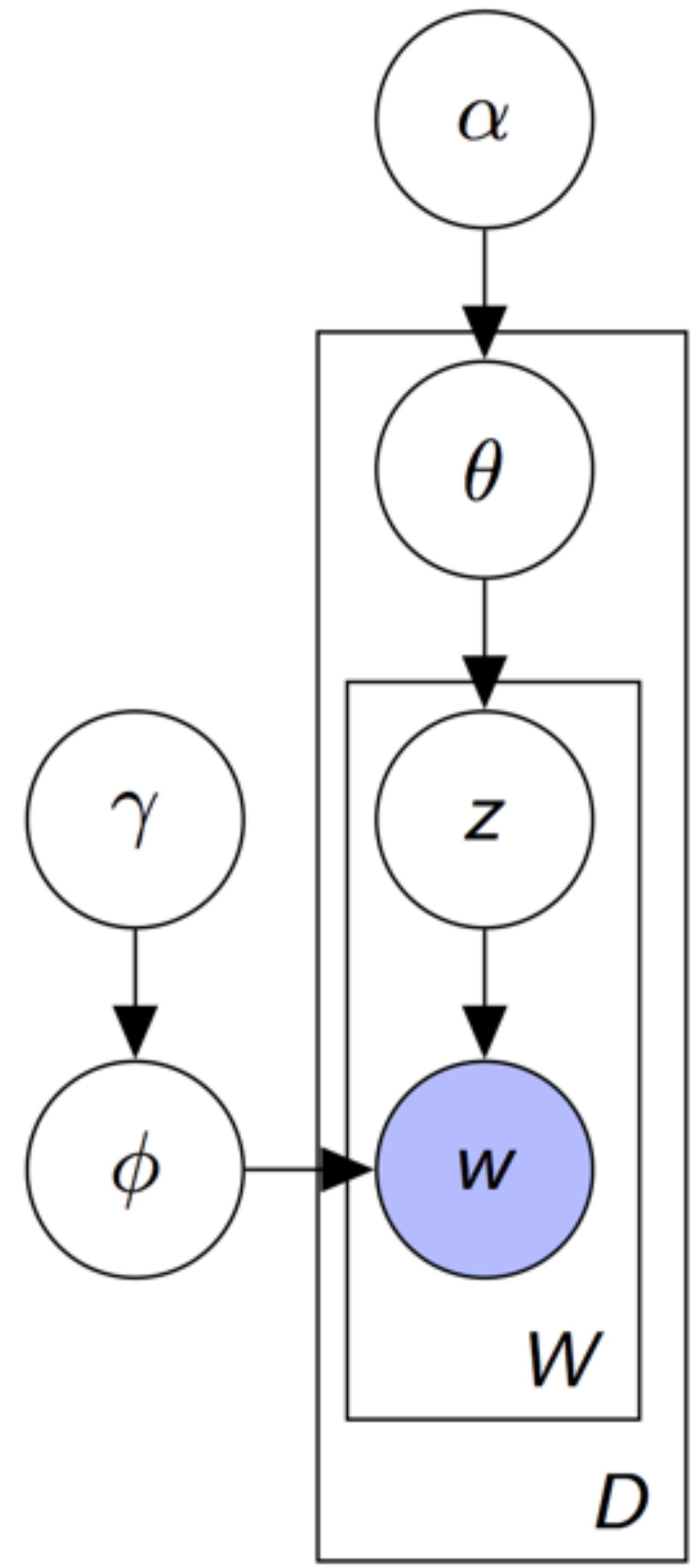
Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 11: Topic models

Feb 29, 2016

Topic models



Latent variables

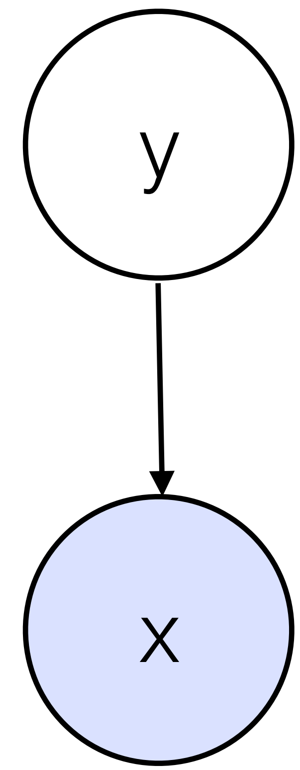
- A latent variable is one that's unobserved, either because:
    - we are predicting it (but have observed that variable for other data points)
    - it is unobservable

Latent variables

	observed variables	latent variables
email	text, date, sender	topic
novels	text, author, pub date	genre, topic
social network	nodes, friendship structure	communities
fitbit data	accelerometer output	steps, sleep patterns
legislators	voting behavior, speeches	political preference
netflix users	watching behavior, ratings	genre preference

Probabilistic graphical models

- Nodes represent variables (shaded = observed, clear = latent)
- Arrows indicate conditional relationships
- The probability of x here is dependent on y
- Simply a visual way of writing the joint probability:



$$P(x, y) = P(y)\, P(x \mid y)$$

Topic Models

- A probabilistic model for discovering hidden “topics” or “themes” (groups of terms that tend to occur together) in documents.
- Unsupervised (find *interesting structure* in the data)
- Clustering algorithm:

How to tokens cluster into topics?

Topic Models

- **Input**: set of documents, number of clusters to learn.
- **Output**:
  - topics
  - topic ratio in each document
  - topic distribution for each word in doc

{album, band, music}	{government, party, election}	{game, team, player}
album	government	game
band	party	team
music	election	player
song	state	win
release	political	play
{god, call, give}	**{company, market, business}**	**{math, number, function}**
god	company	math
call	market	number
give	business	function
man	year	code
time	product	set
{city, large, area}	**{math, energy, light}**	**{law, state, case}**
city	math	law
large	energy	state
area	light	case
station	field	court
include	star	legal

topic models cluster tokens into “topics”

… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."

topic models cluster tokens into “topics”

… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."

“Death”

topic models cluster tokens into “topics”

… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."

“Love”

topic models cluster tokens into “topics”

… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."

“Family”

topic models cluster tokens into “topics”

… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."

“Etc.”

tokens, not types

… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, Paris confronts him and, in the ensuing battle, Romeo kills Paris. Still believing Juliet to be dead, he drinks the poison. Juliet then awakens and, finding Romeo dead, stabs herself with his dagger. The feuding families and the Prince meet at the tomb to find all three dead. Friar Laurence recounts the story of the two "star-cross'd lovers". The families are reconciled by their children's deaths and agree to end their violent feud. The play ends with the Prince's elegy for the lovers: "For never was a story of more woe / Than this of Juliet and her Romeo."

"People"

A different *Paris* token might belong to a "Place" or "French" topic

Applications

A Topic Model of Literary Studies Journals

Overview | Topic | Article | Word | Bibliography | Word index | Settings | About

List | Grid | Years

click a column label to sort; click a row for more about a topic

topic	1889—2013	top words	proportion of corpus
1		see both own view role university further account critical particular	2.5%
2		other both two form same even each part experience process	2.6%
3		old beowulf english ic mid swa pe poet ond grendel	0.3%

http://www.rci.rutgers.edu/~ag978/quiet/

x = feature vector

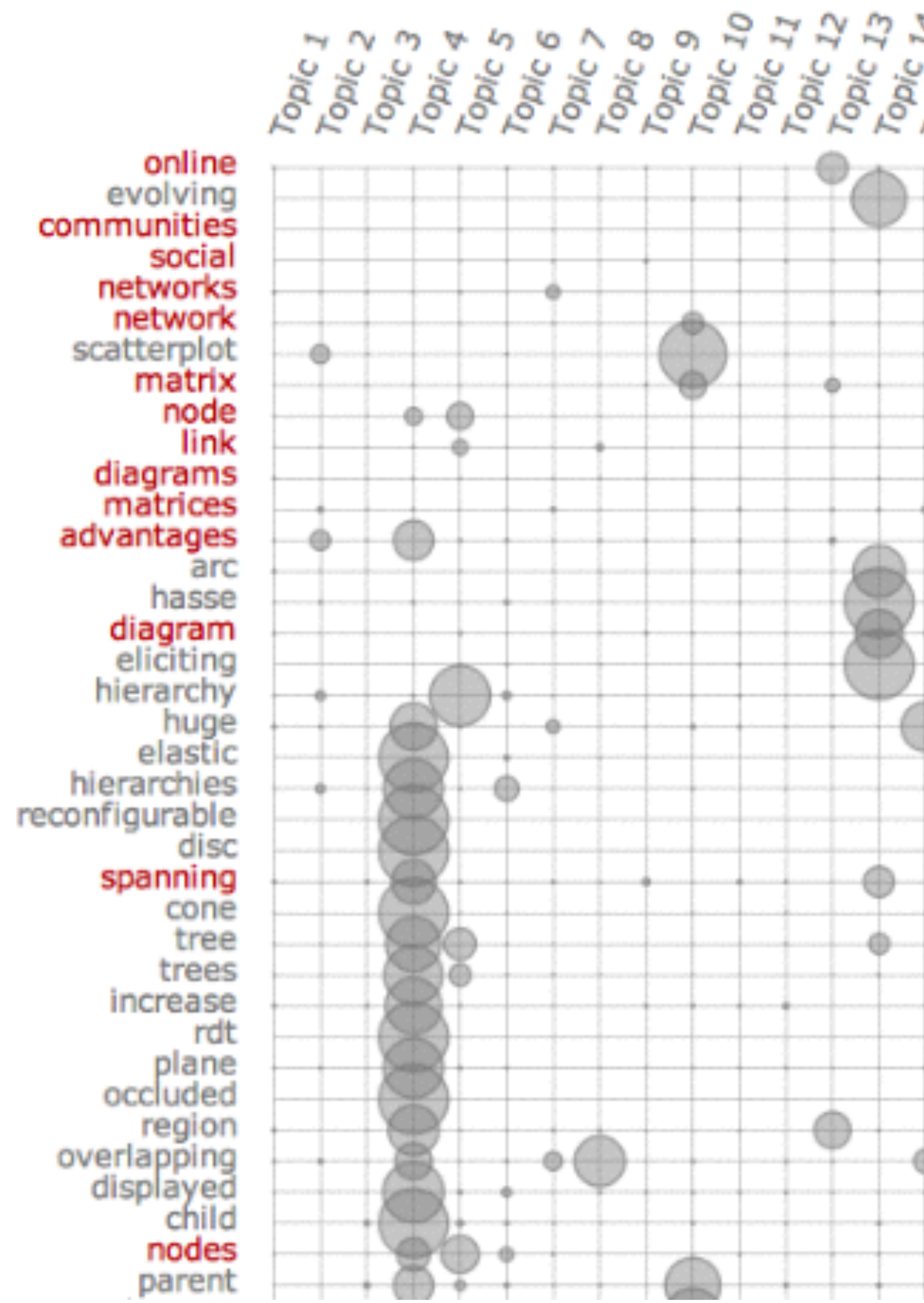
Feature	Value
follow clinton	0
follow trump	0
“republican” in profile	0
“democrat” in profile	0
“benghazi"	1
topic 1	0.55
topic 2	0.32
topic 3	0.13

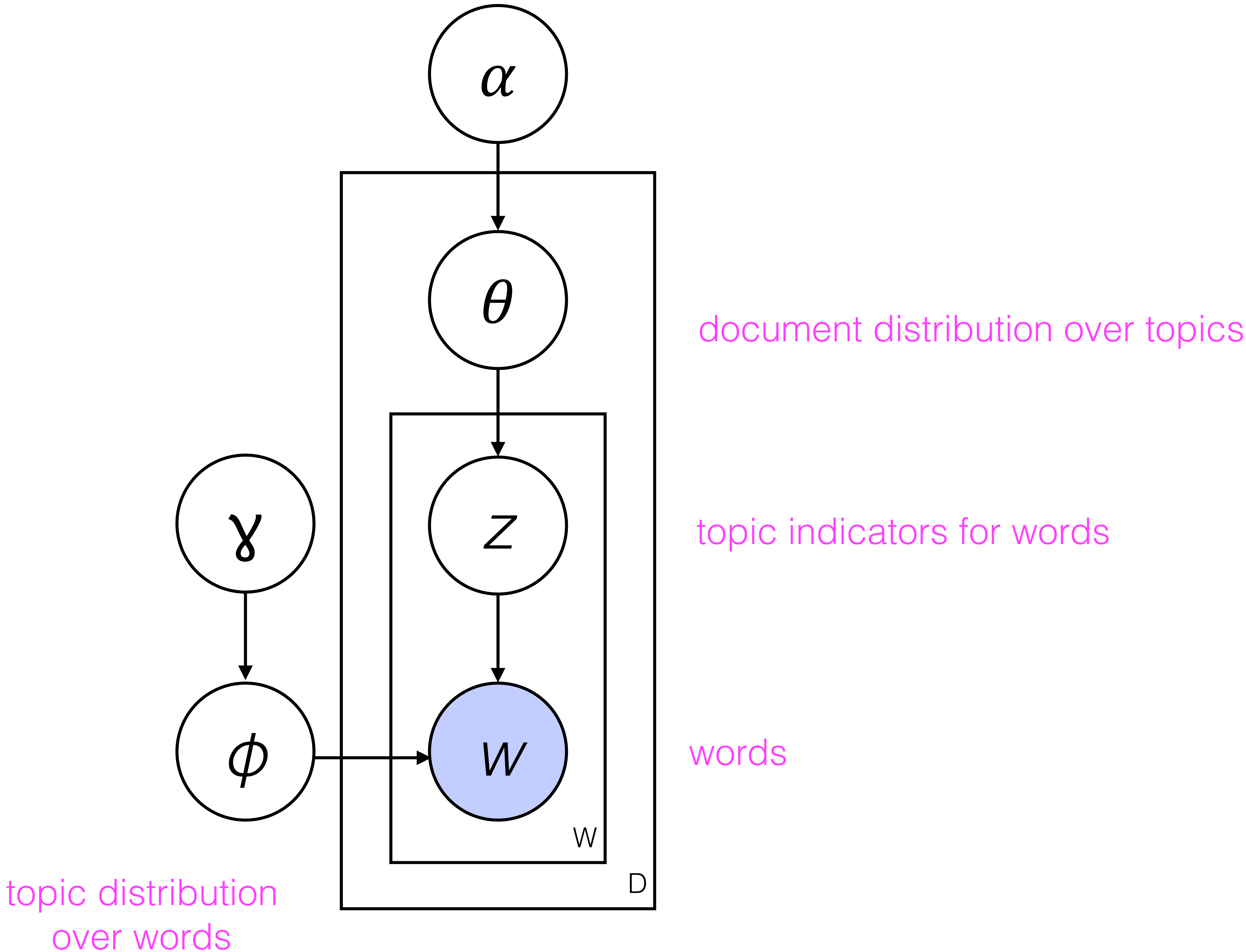
β = coefficients

Feature	β
follow clinton	-3.1
follow trump	6.8
“republican” in profile	7.9
“democrat” in profile	-3.0
“benghazi"	-1.7
topic 1	0.3
topic 2	-1.2
topic 3	5.7

Software

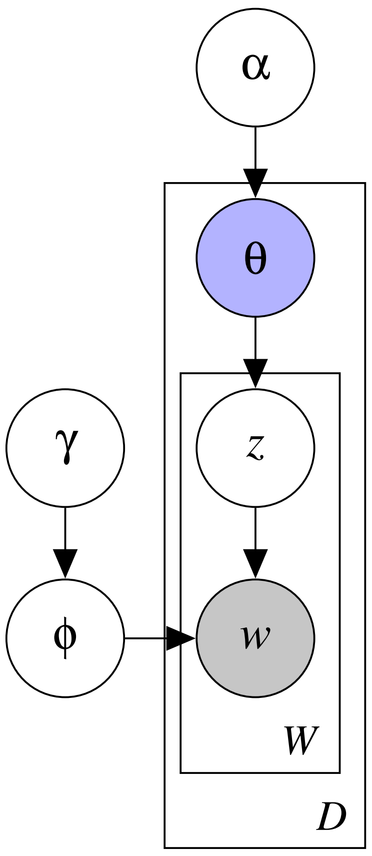
- Mallet
  http://mallet.cs.umass.edu/
- Gensim (python)
  https://radimrehurek.com/gensim/
- Visualization
  https://github.com/uwdata/termite-visualizations

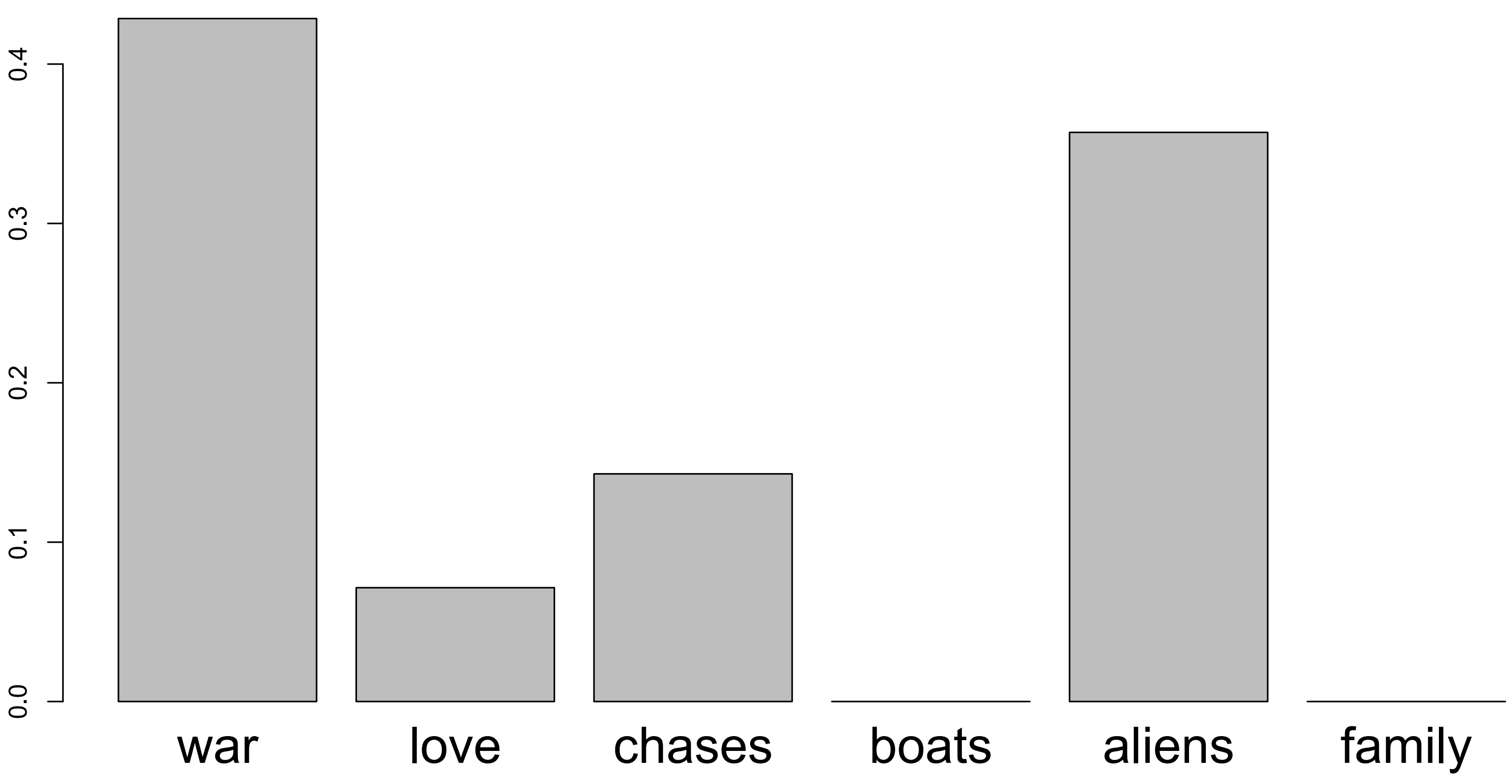



α
θ
document distribution over topics
γ
z
topic indicators for words
ϕ
w
words
W
D
topic distribution
over words

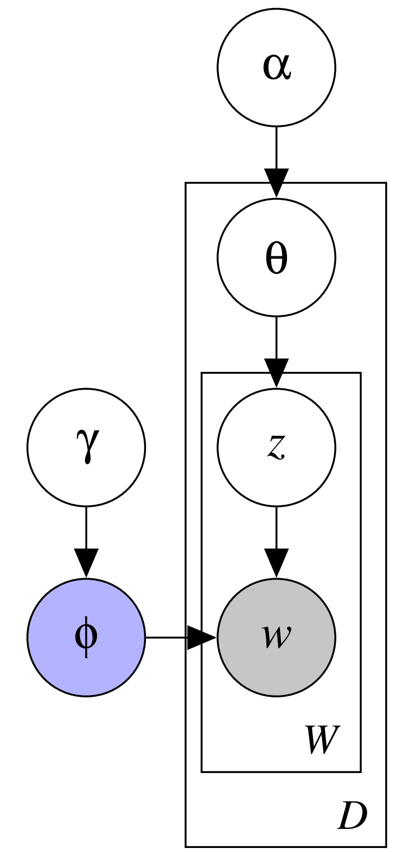
Topic Models

- A document has *distribution over topics*



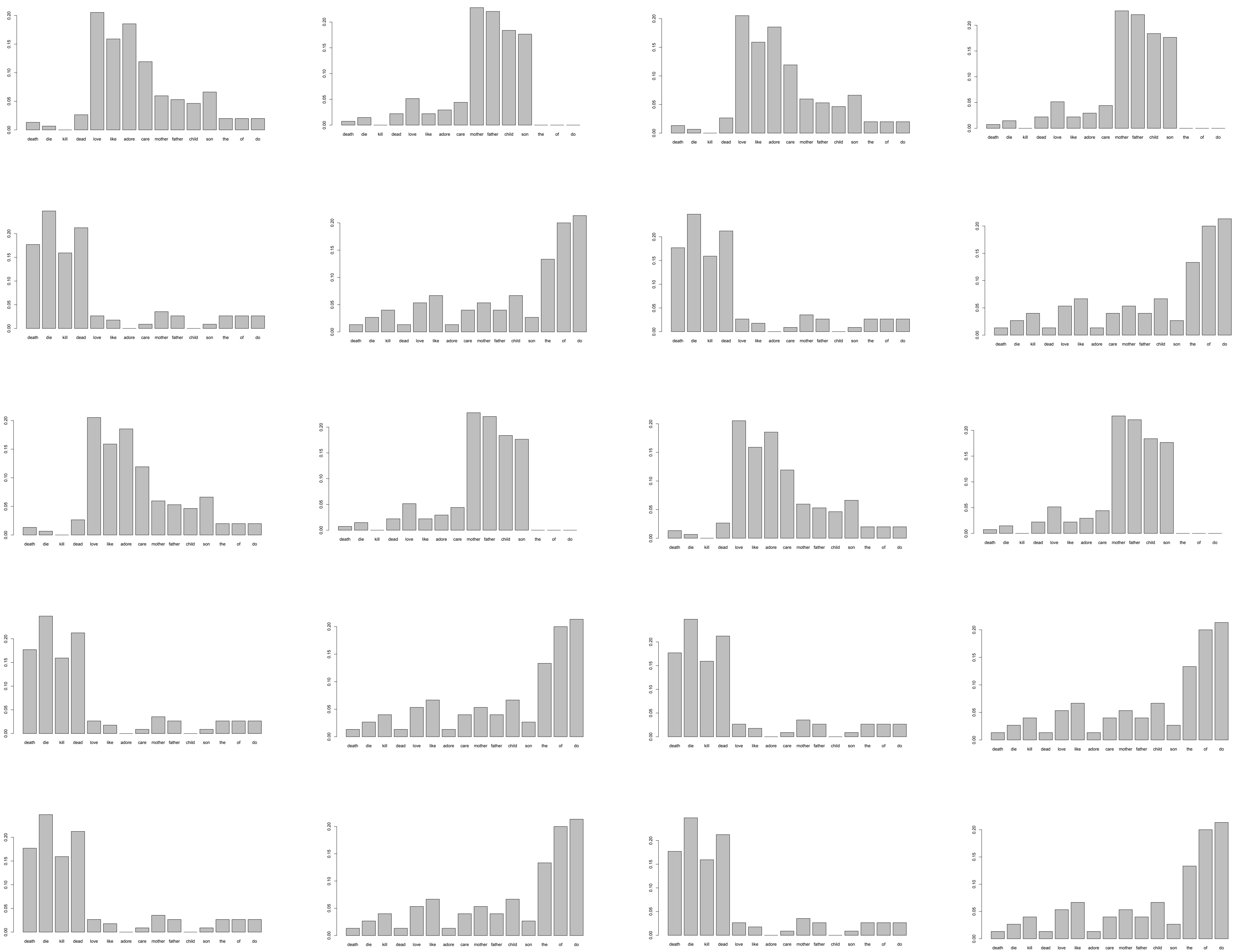


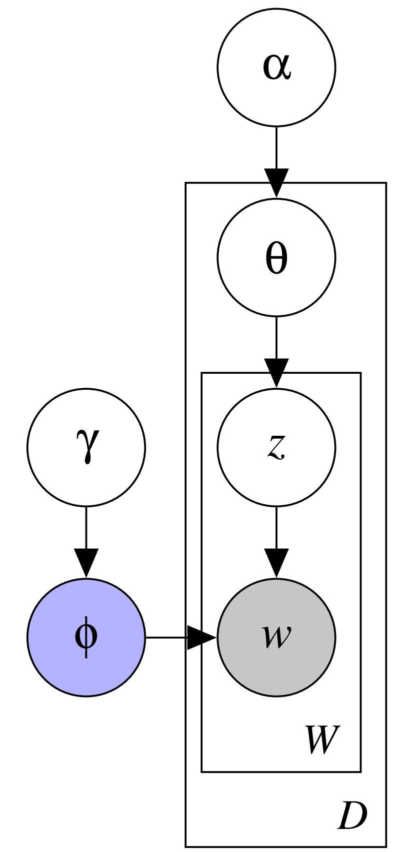
Topic Models



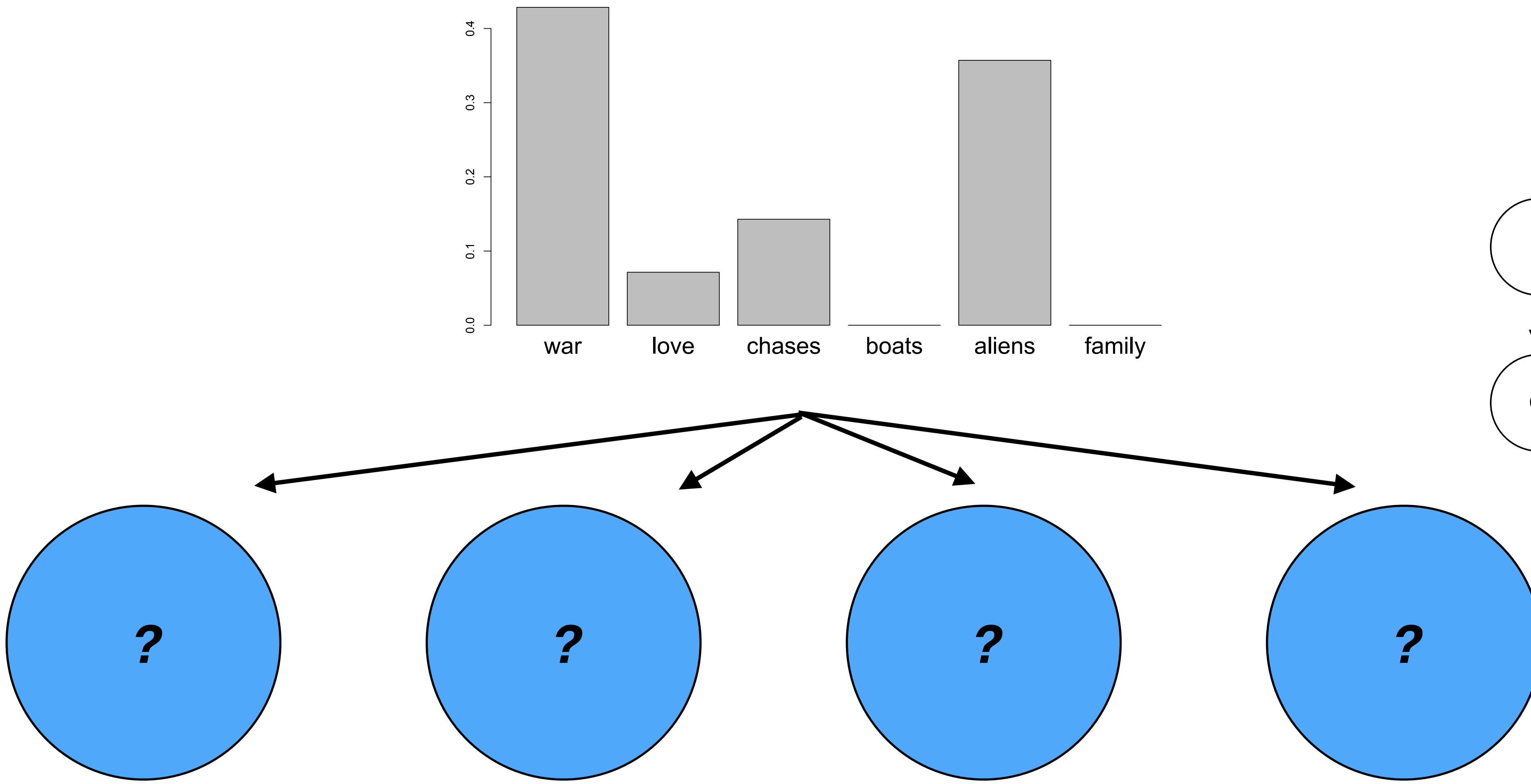
- A topic is a distribution over words

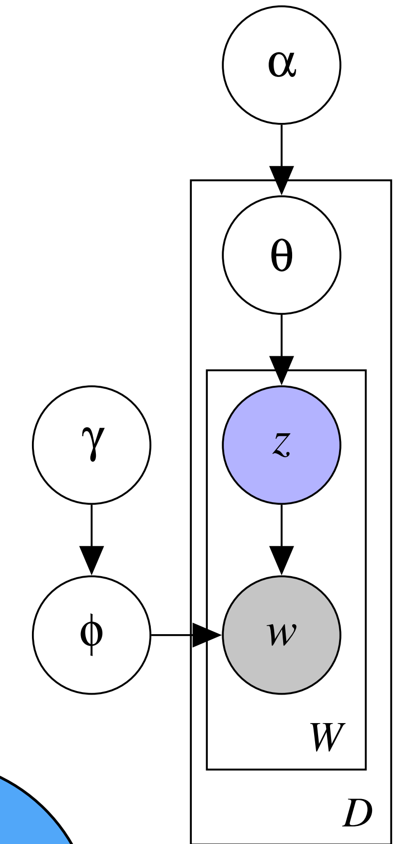
- e.g., P("adore" | topic = love) = .18


death die kill dead love like adore care mother father child son the of do

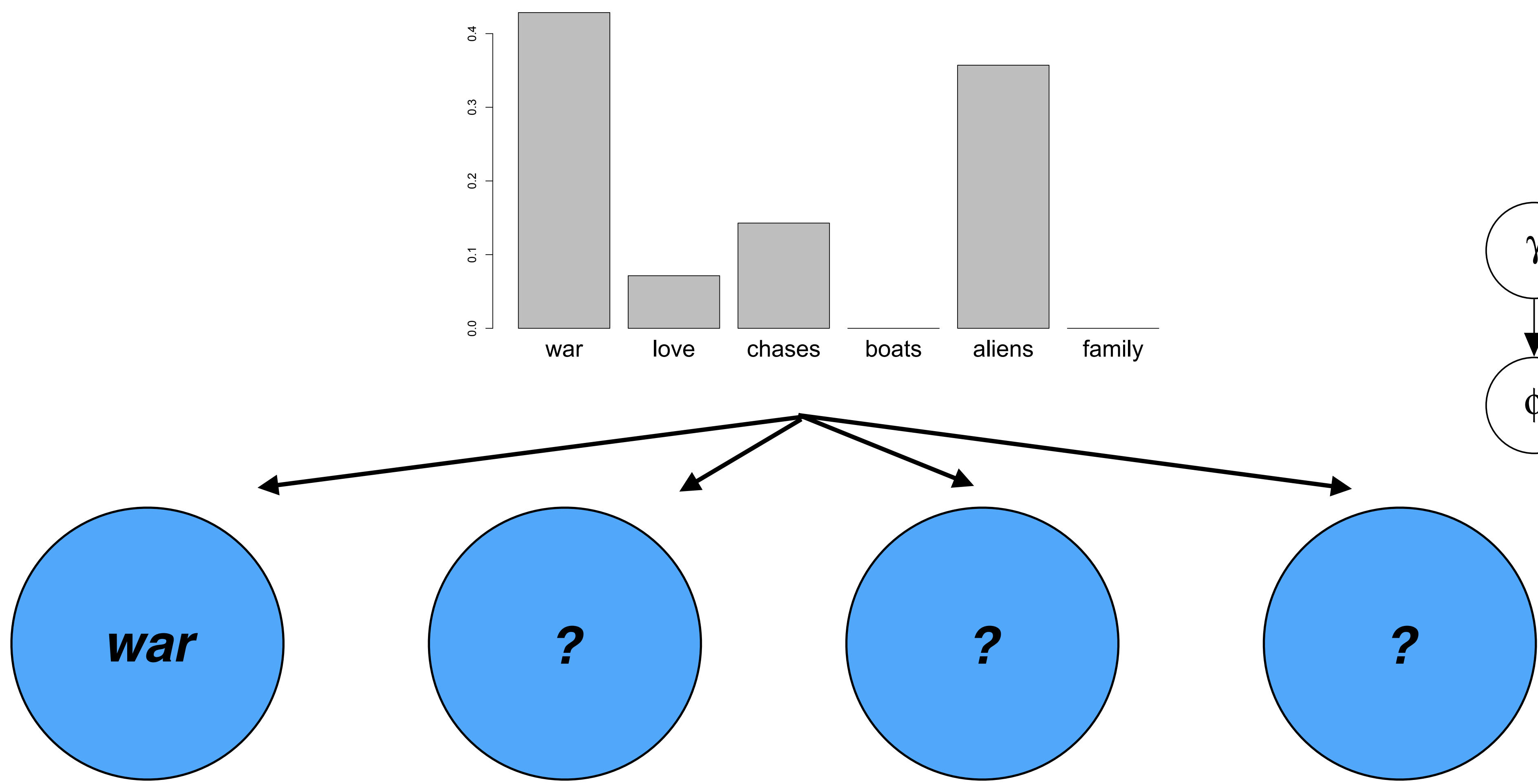

α
θ
γ
z
ϕ
w
W
D

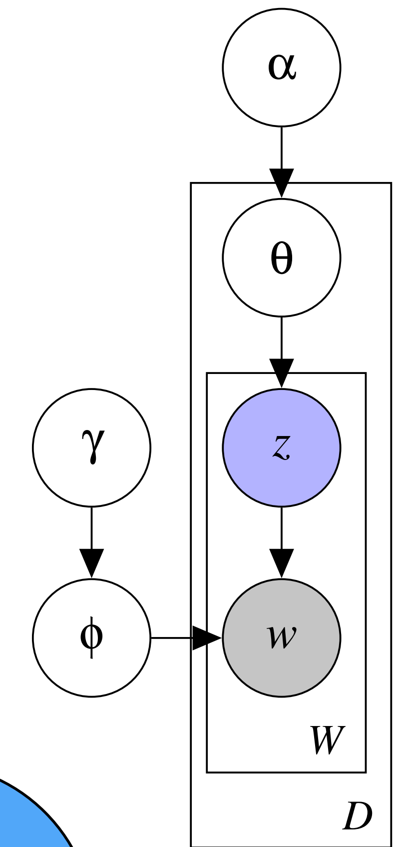
K=20



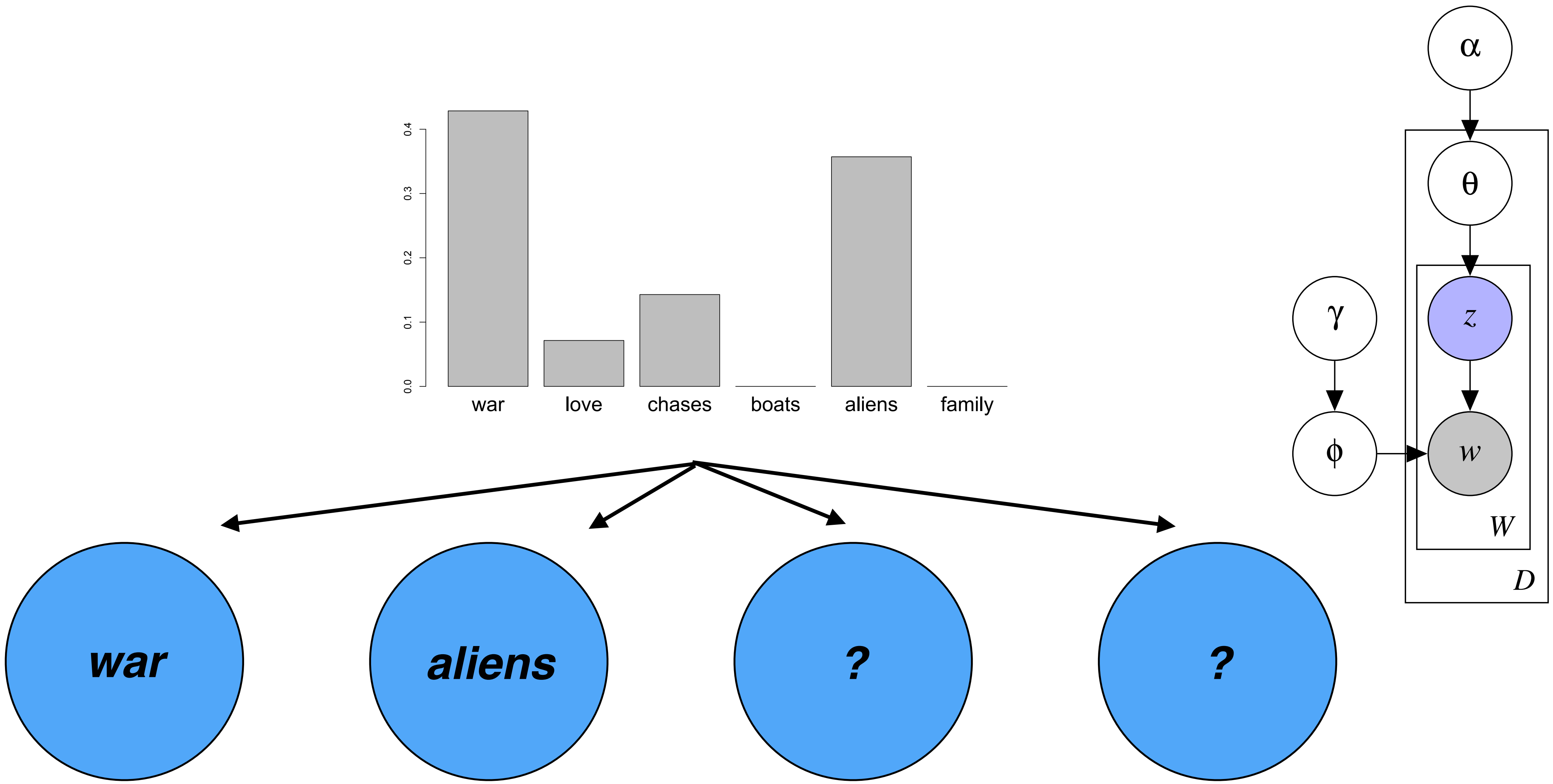


P(topic | topic distribution)

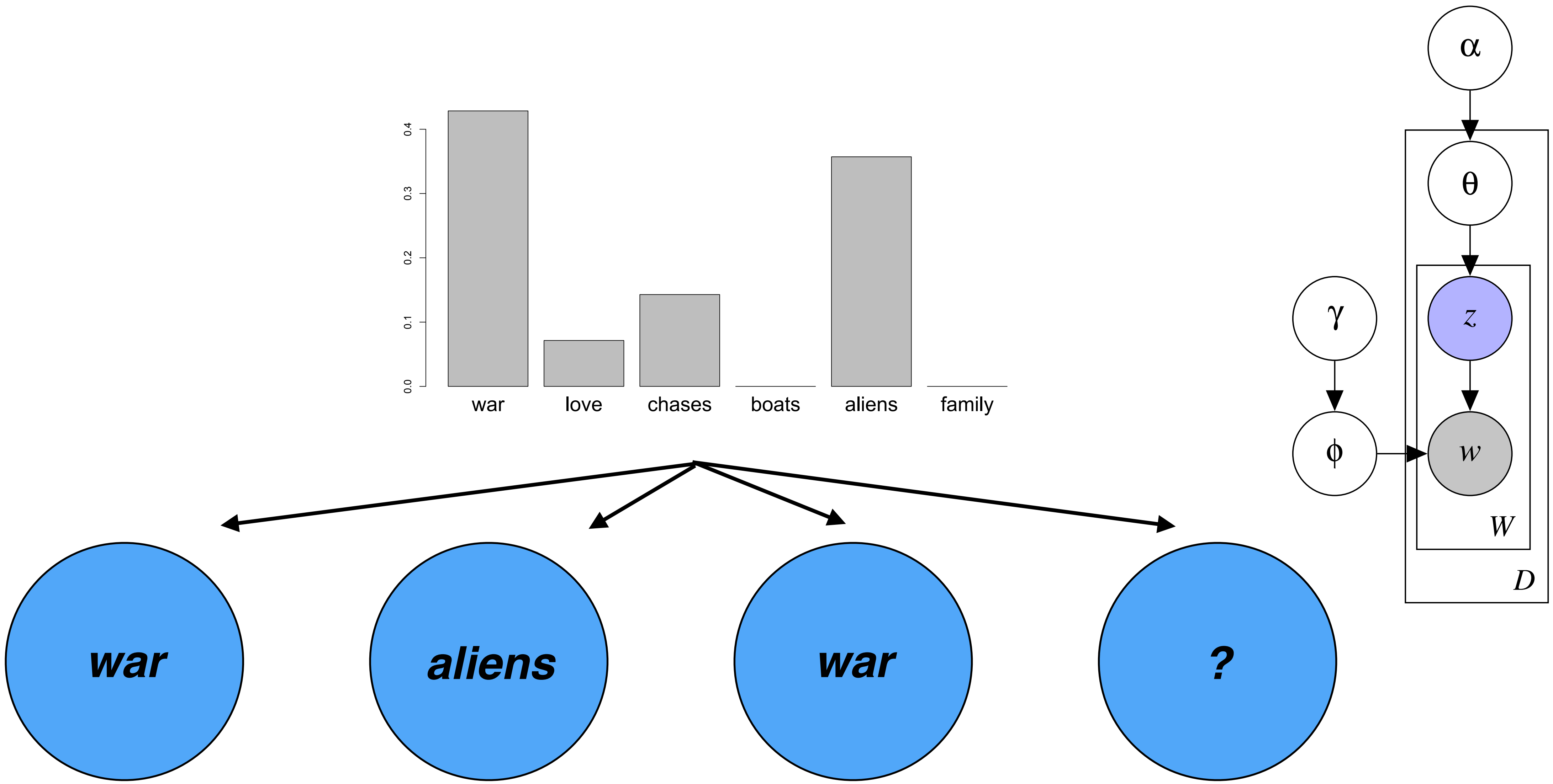




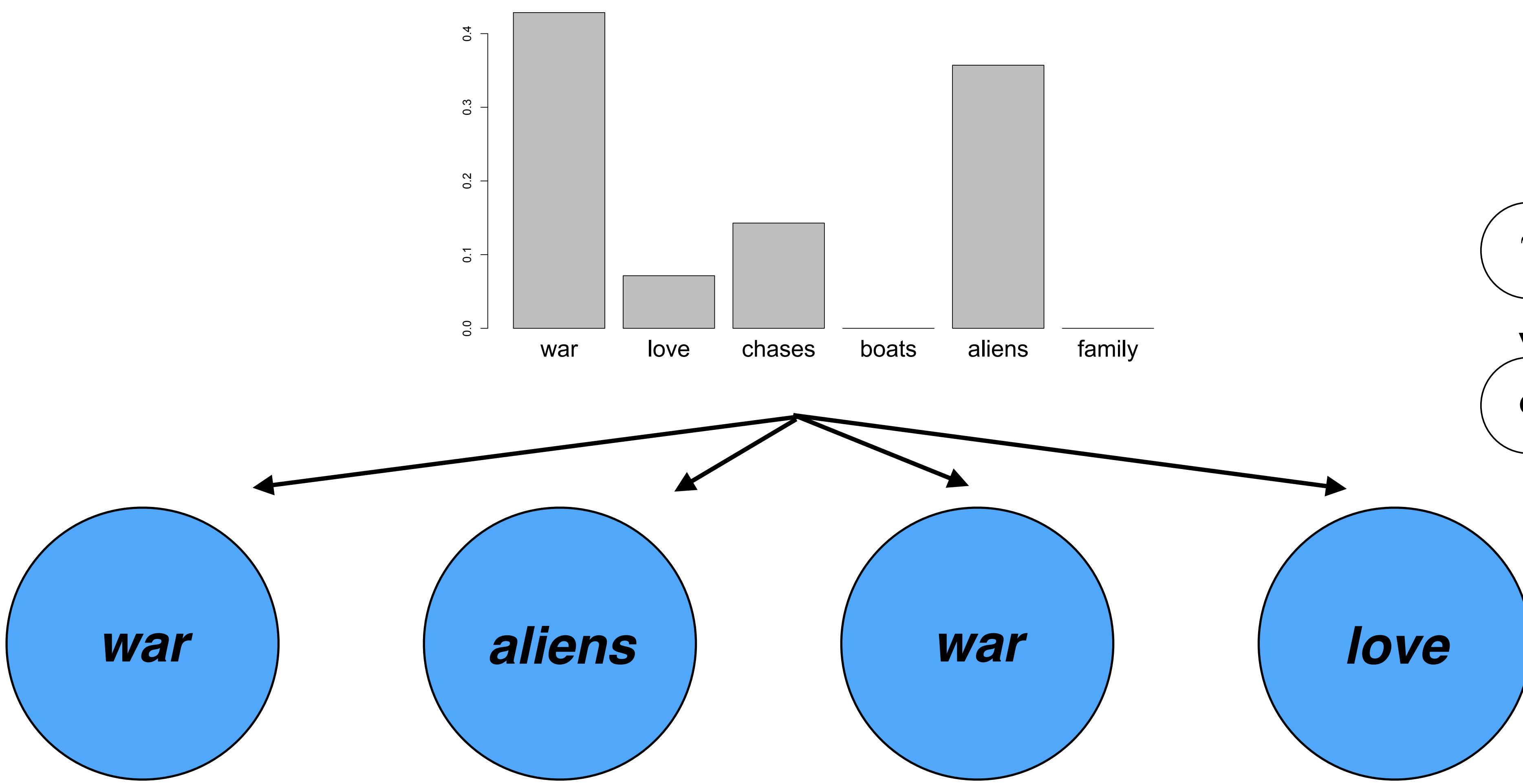
P(topic | topic distribution)

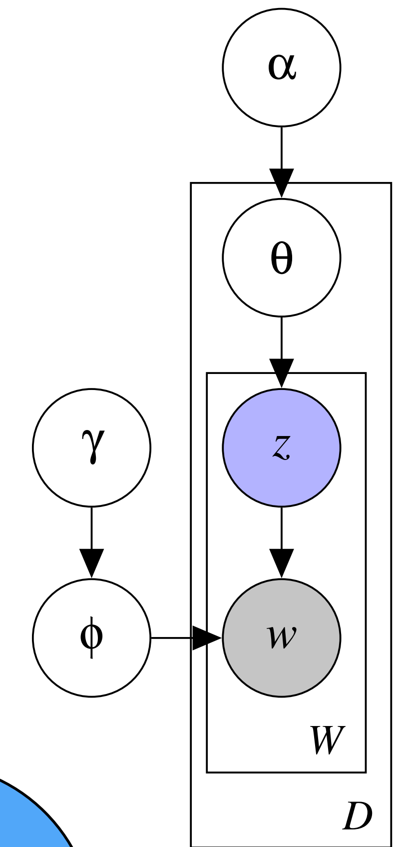


P(topic | topic distribution)

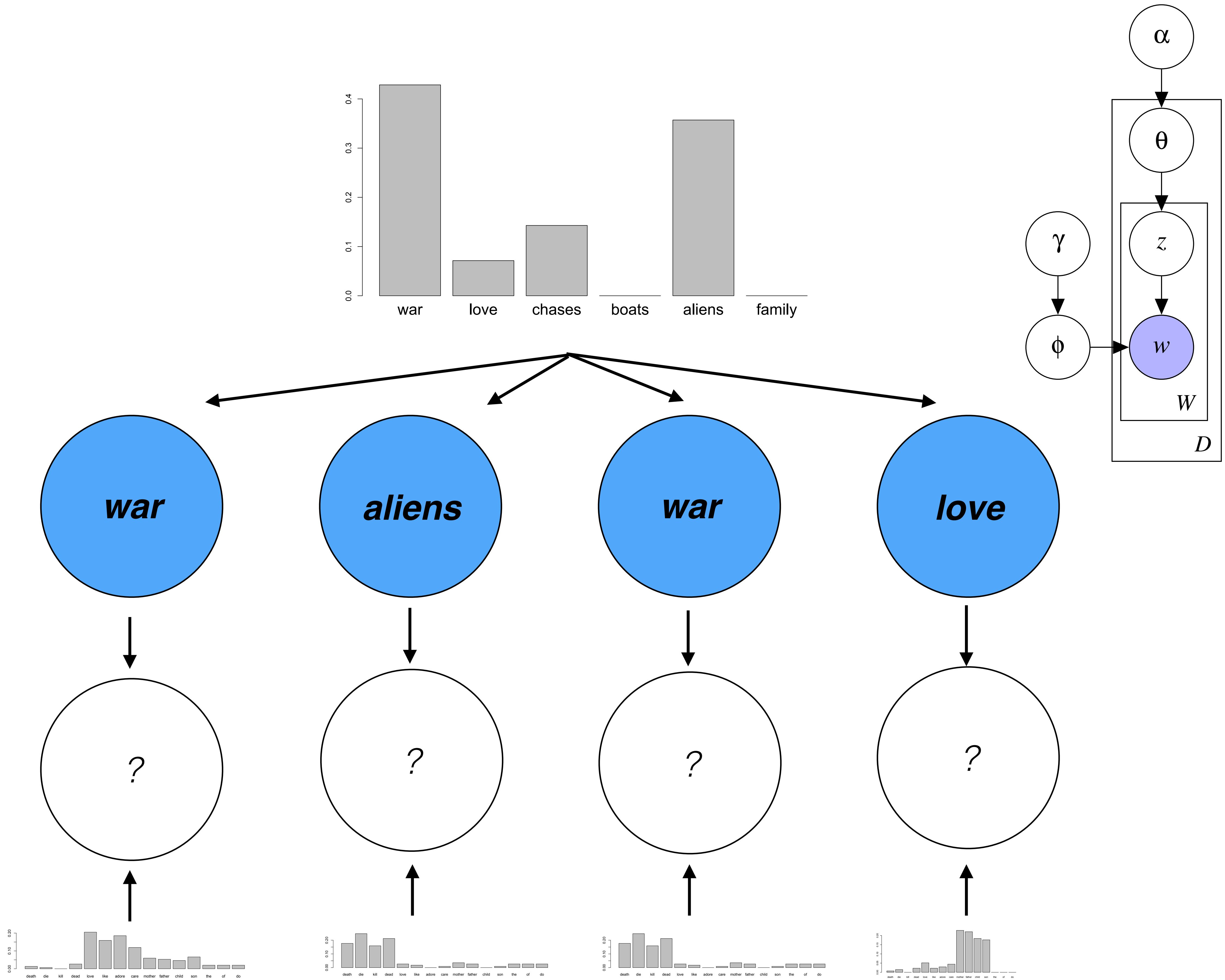


P(topic | topic distribution)

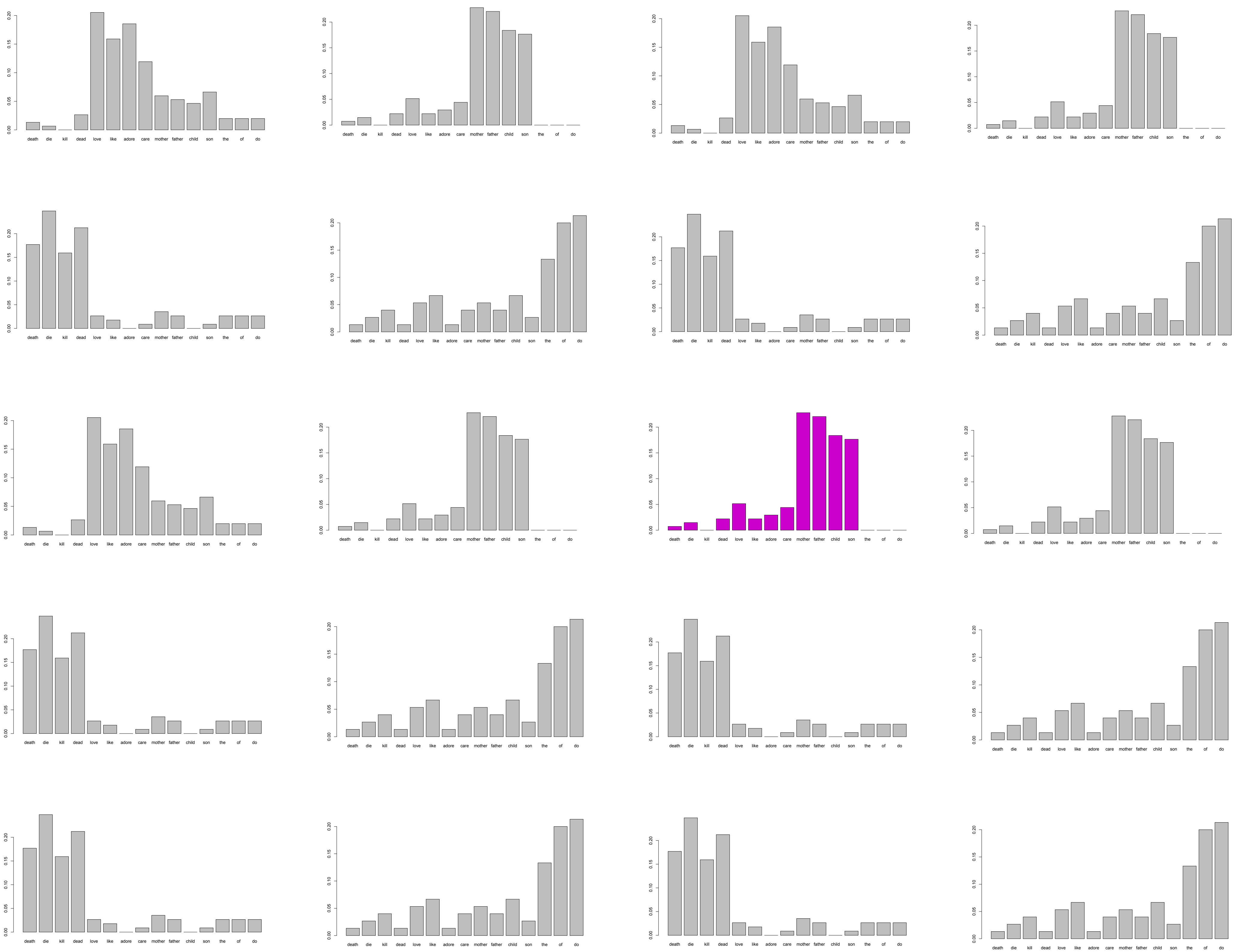


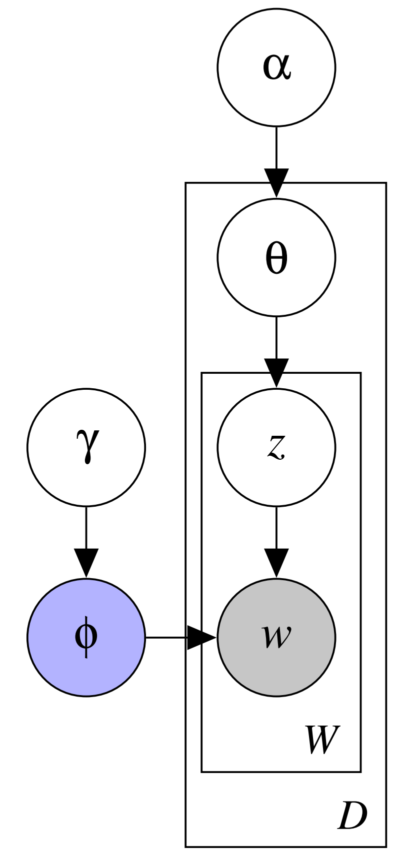


P(topic | topic distribution)



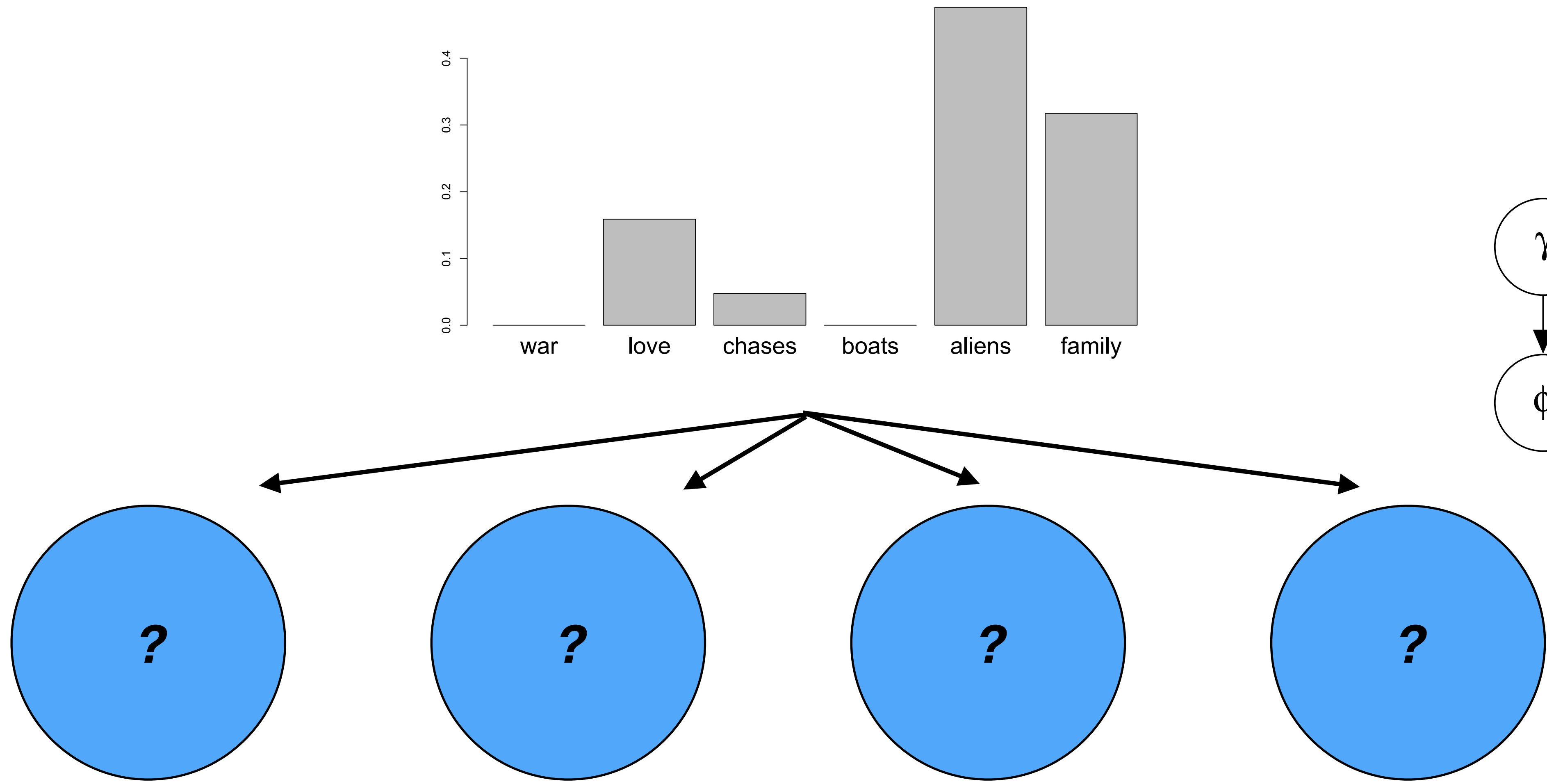
0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
α
θ
γ
z
ϕ
w
W
D
war
aliens
war
love
?
?
?
?

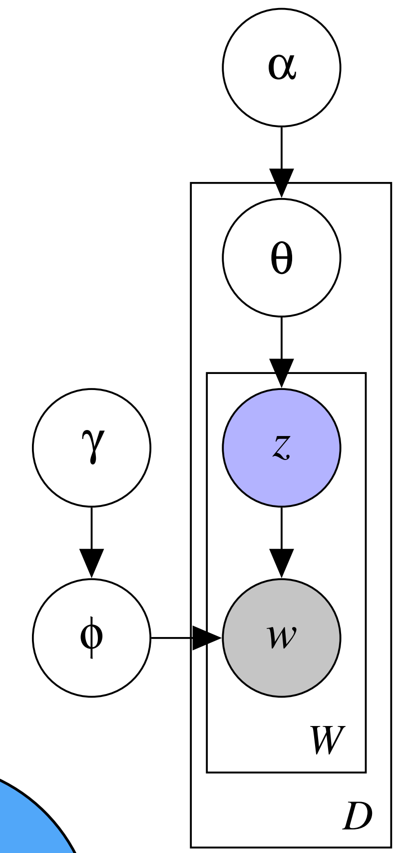

0.00 0.05 0.10 0.15 0.20
death die kill dead love like adore care mother father child son the of do


α
θ
γ
z
ϕ
w
W
D

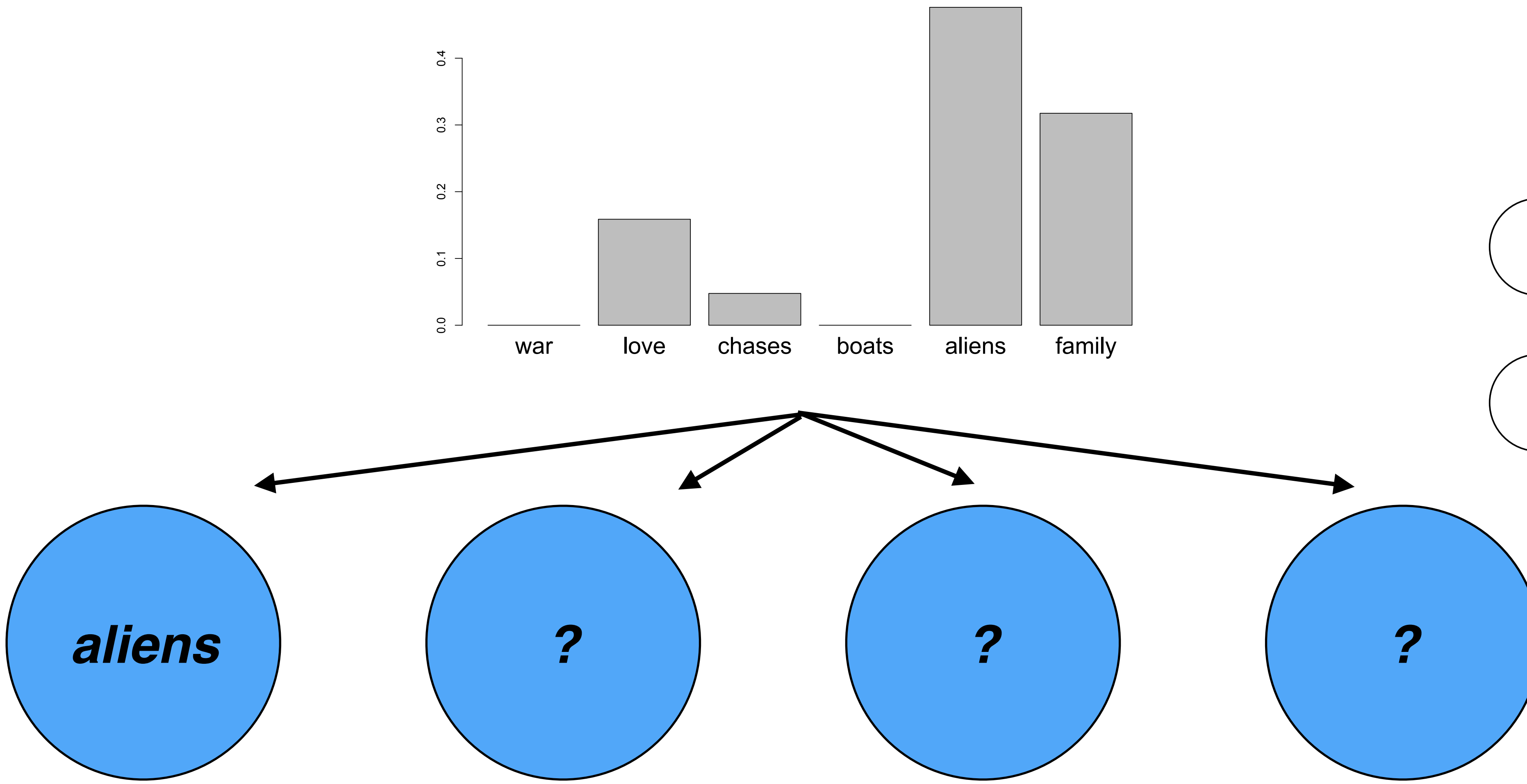
K=20

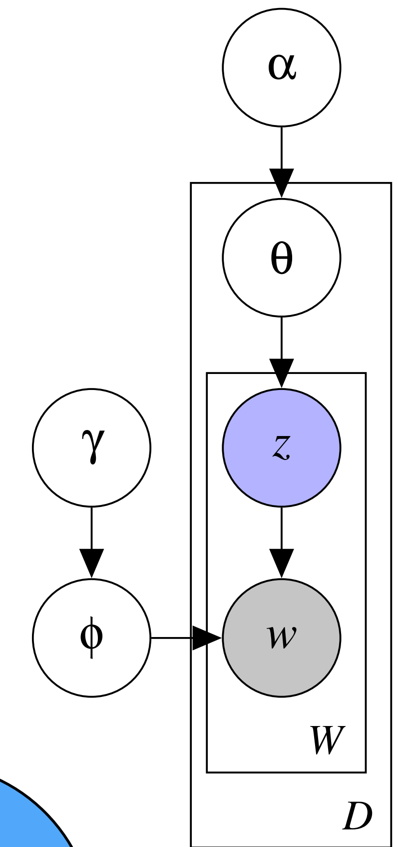
0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
α
θ
γ
z
ϕ
w
W
D
war
aliens
war
love
“fights”
“alien”
“kills”
“marries”



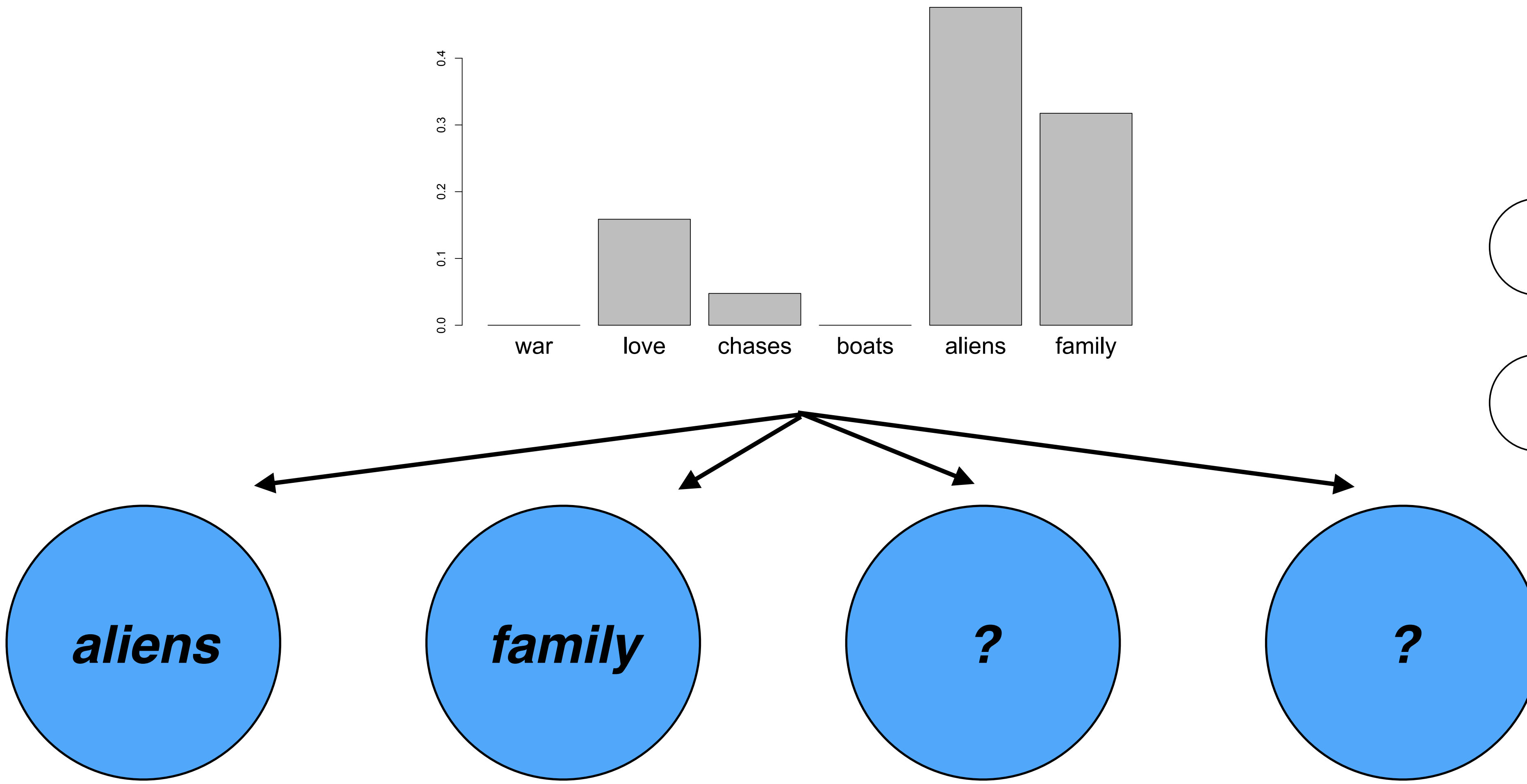


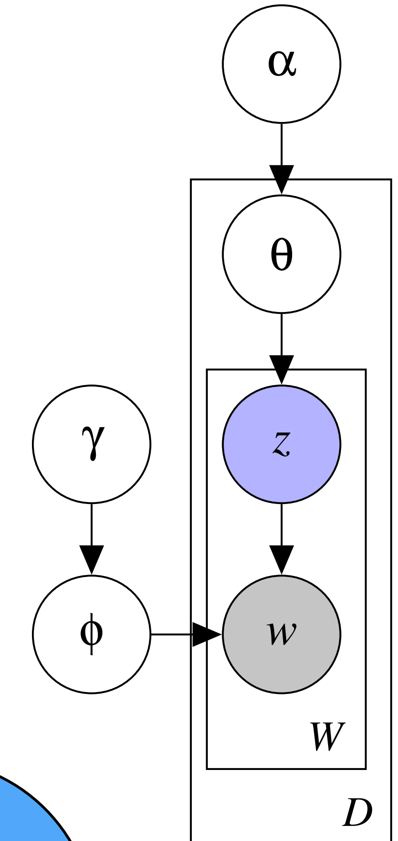
P(topic | topic distribution)



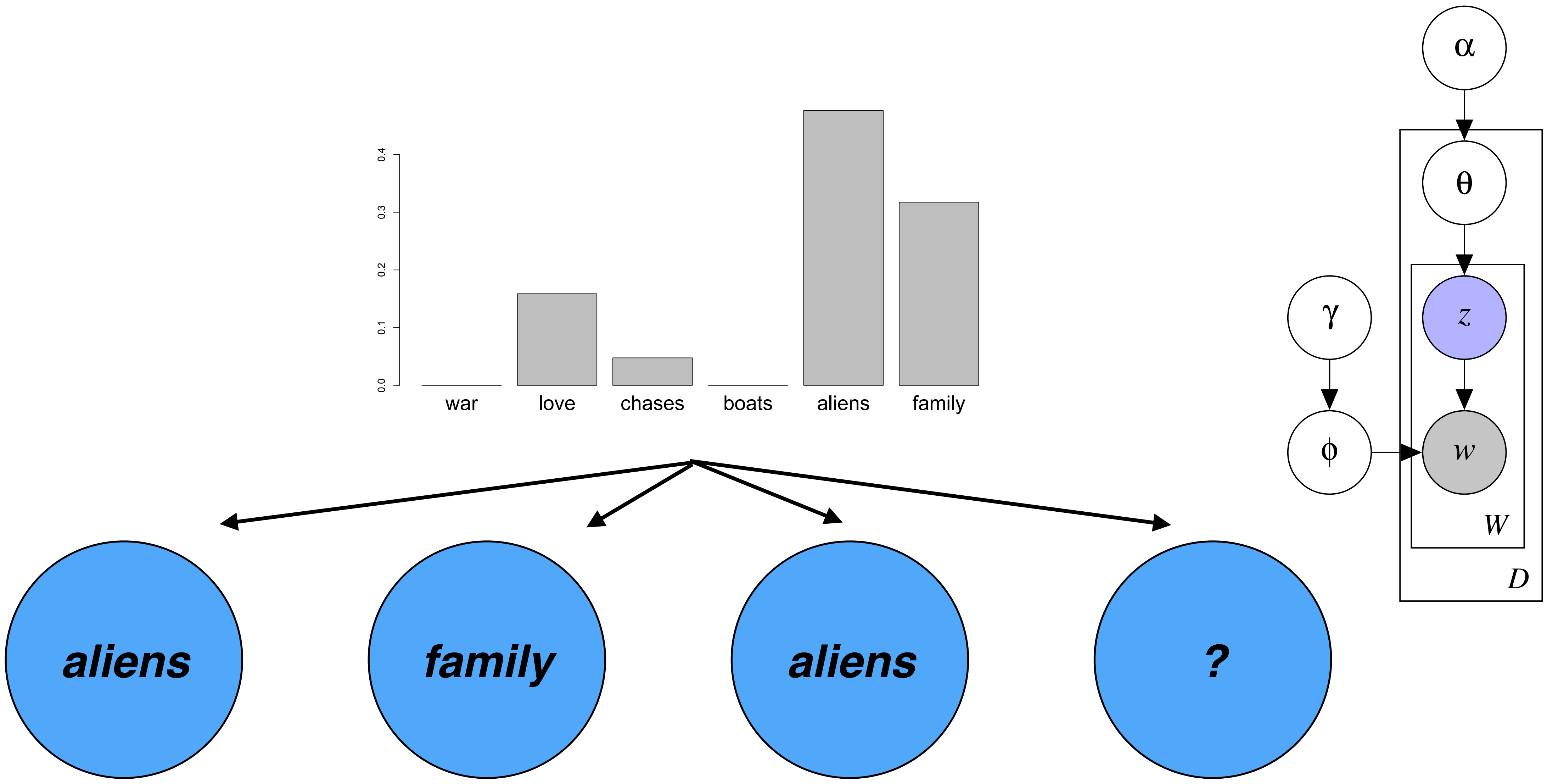


P(topic | topic distribution)

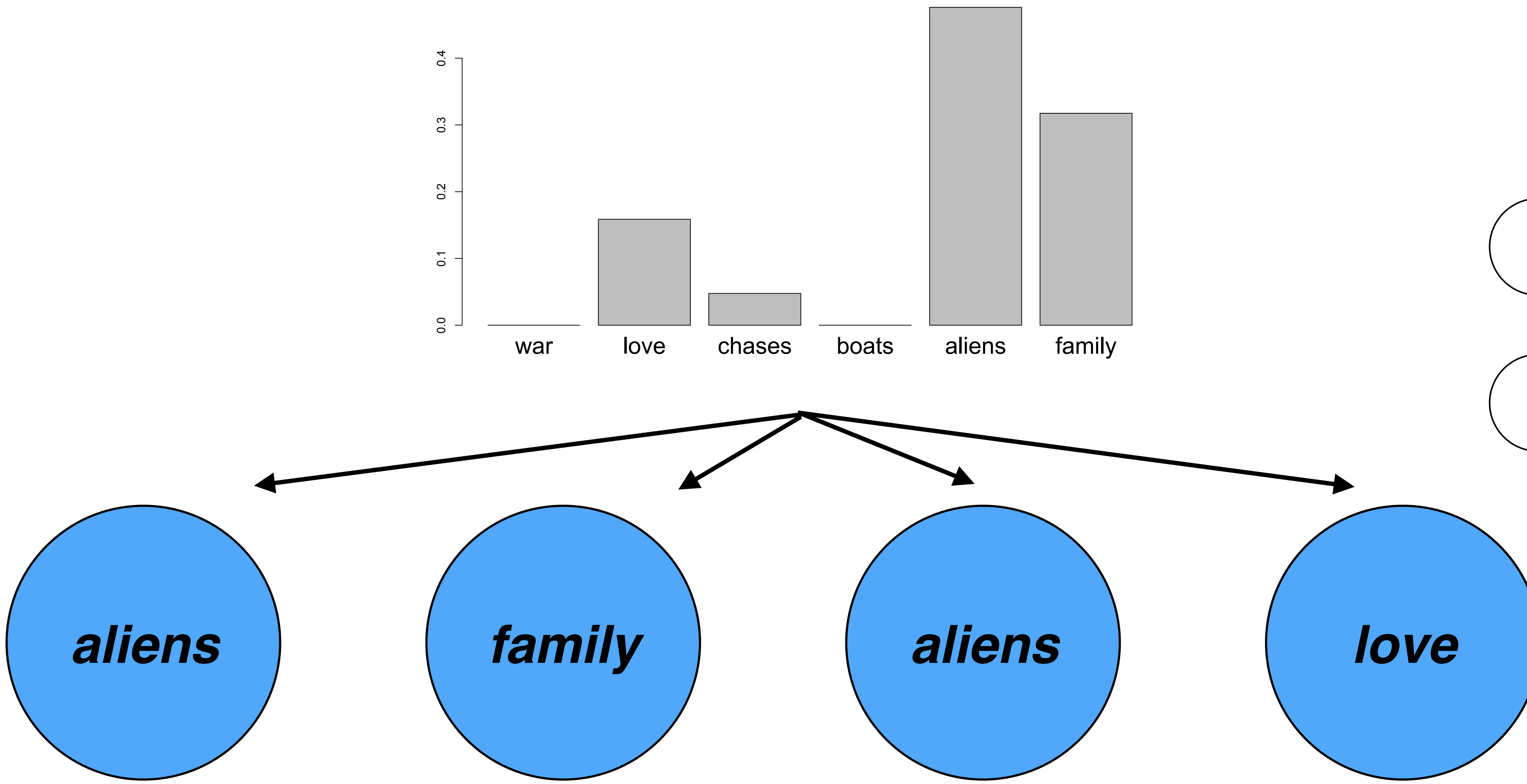


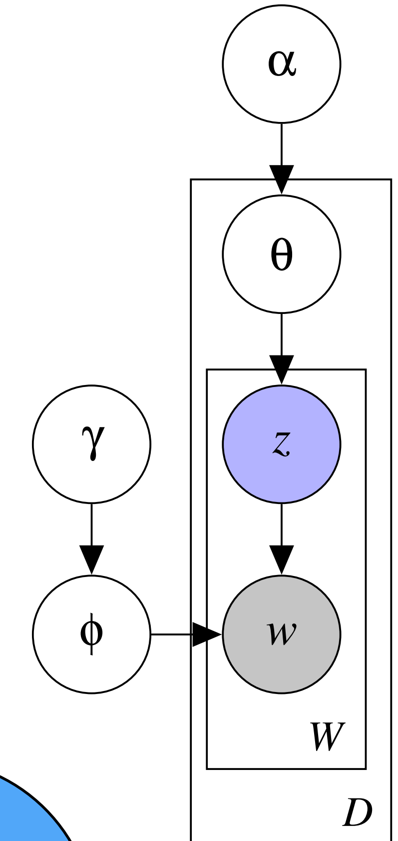


P(topic | topic distribution)

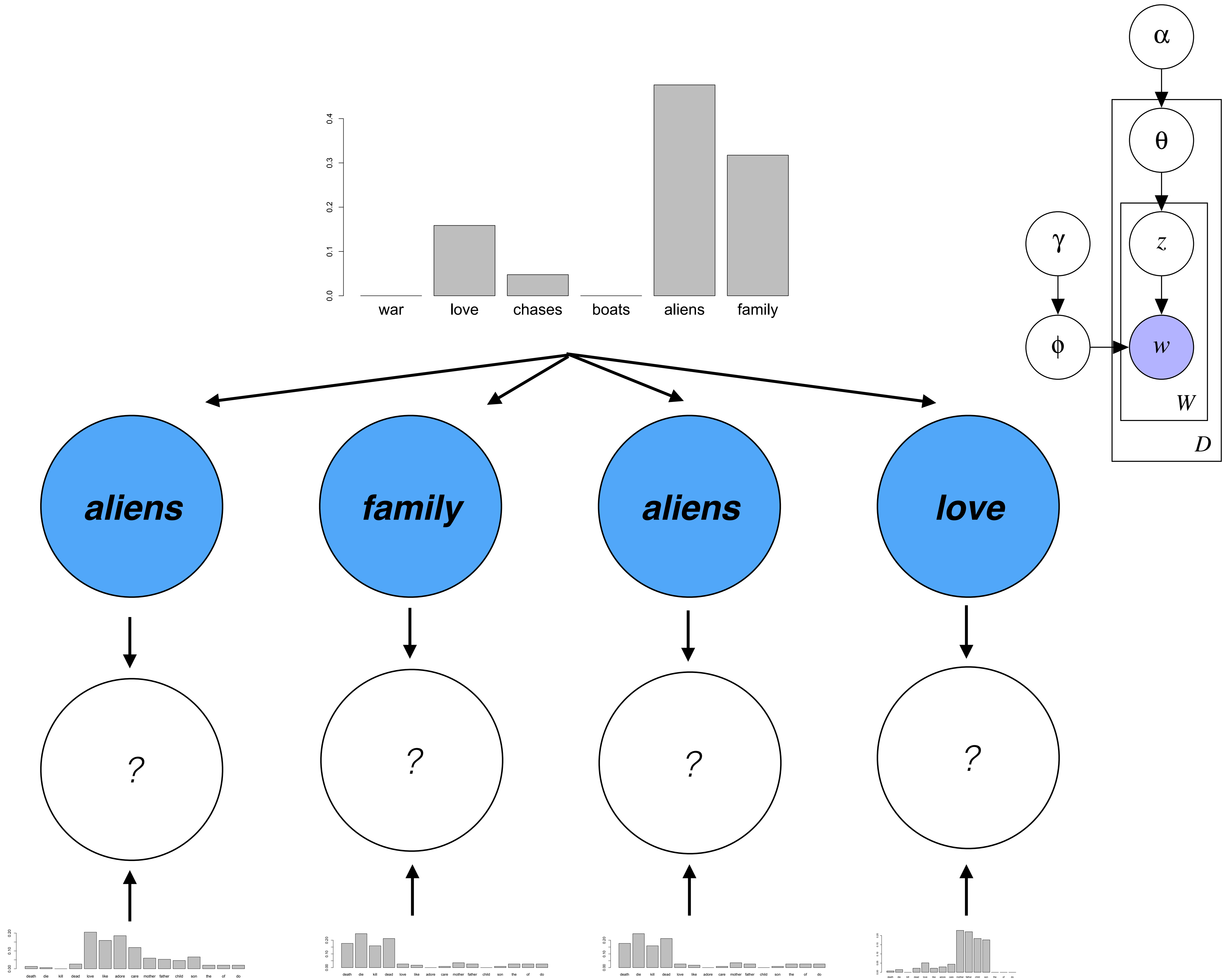


P(topic | topic distribution)

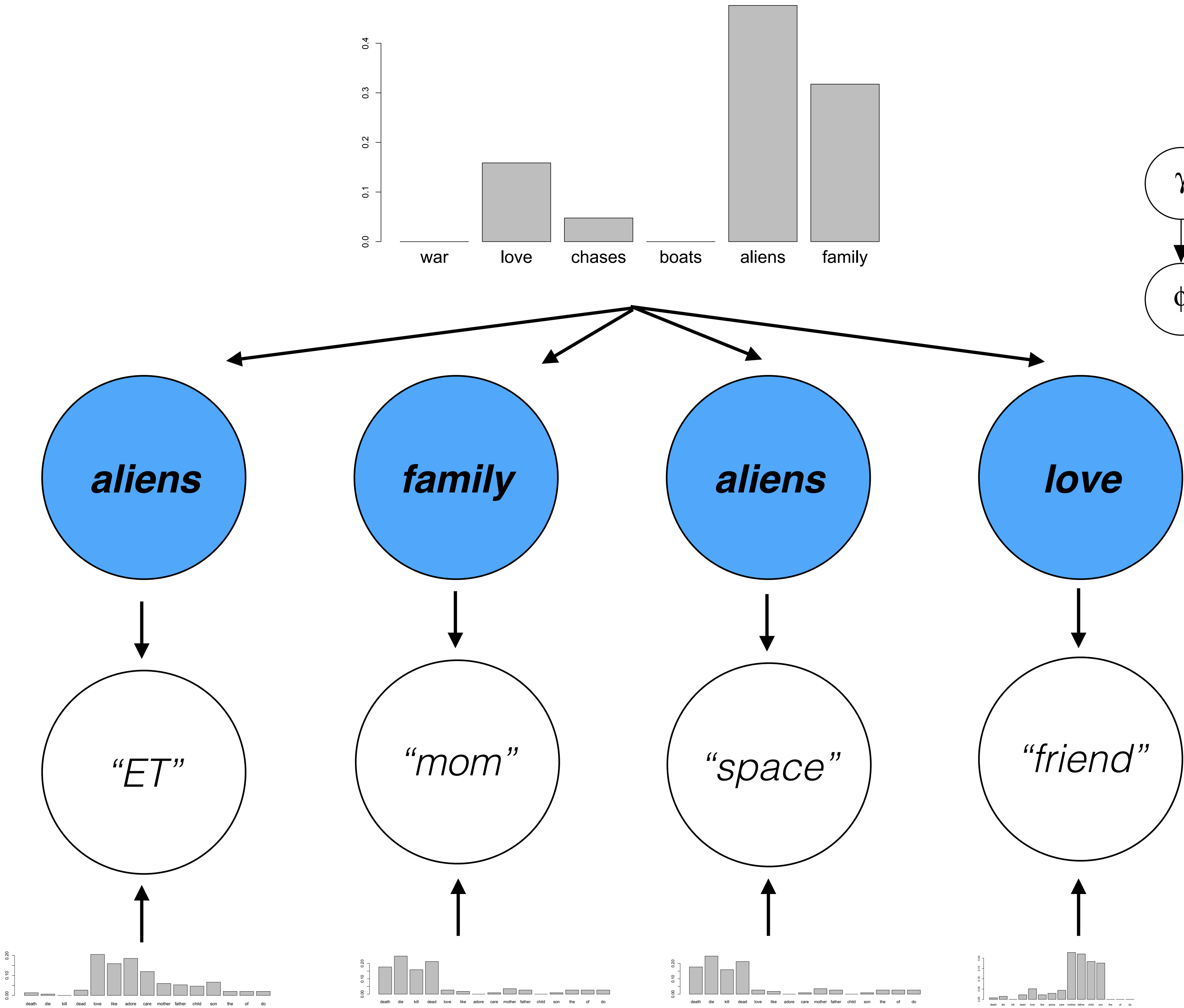




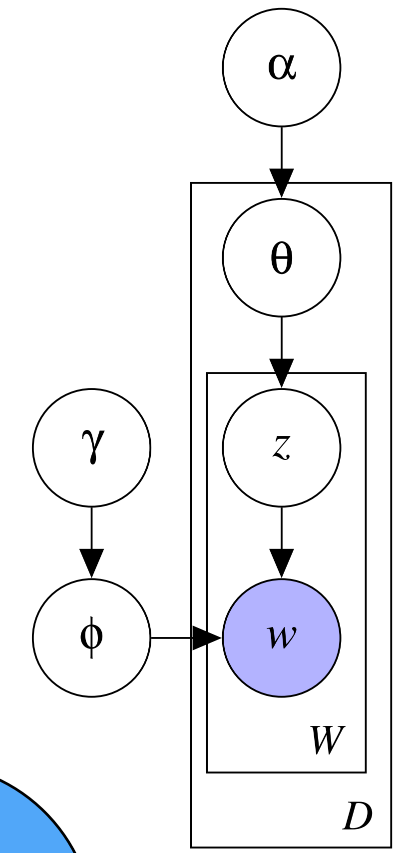
P(topic | topic distribution)



0.4
0.3
0.2
0.1
0.0
war
love
chases
boats
aliens
family
aliens
family
aliens
love
?
?
?
?
α
θ
γ
z
φ
w
W
D



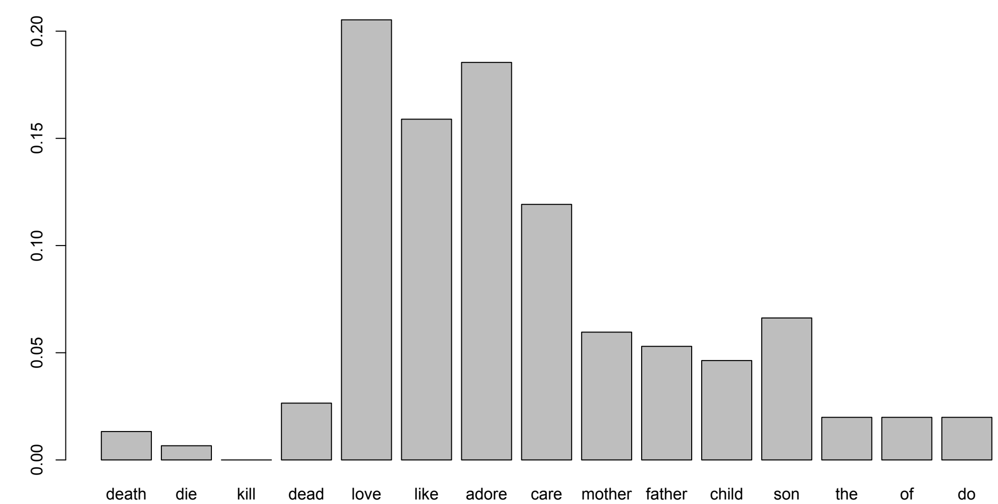
0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
aliens
family
aliens
love
“ET”
“mom”
“space”
“friend”

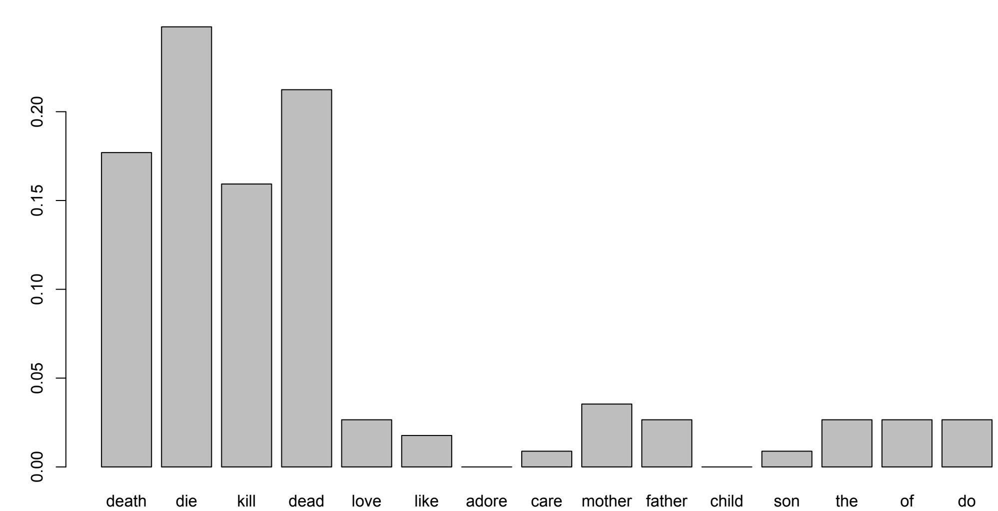


α
θ
γ
z
ϕ
w
W
D

Inferred Topics

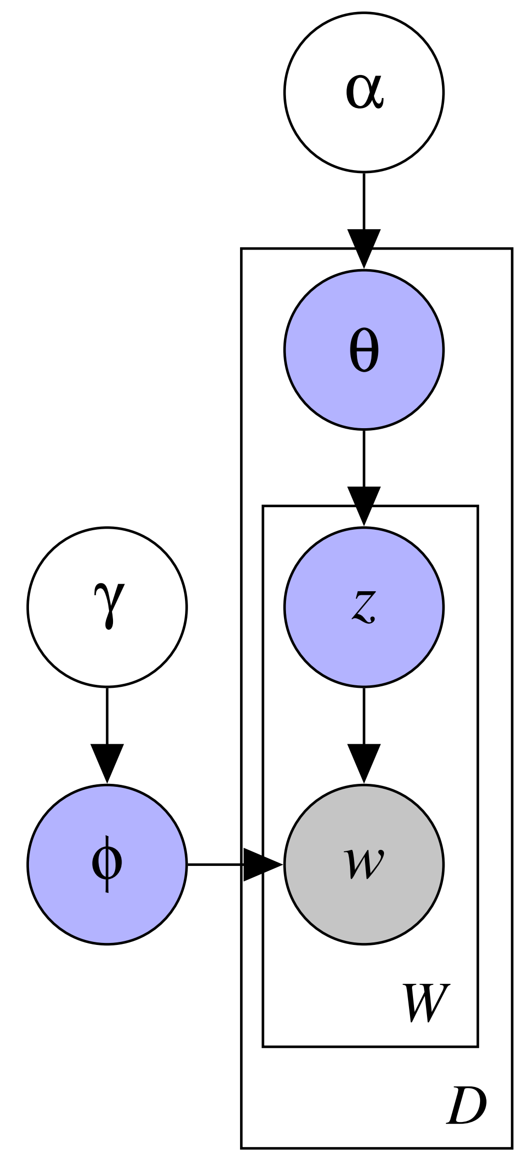
{album, band, music}	{government, party, election}	{game, team, player}
album	government	game
band	party	team
music	election	player
song	state	win
release	political	play
{god, call, give}	**{company, market, business}**	**{math, number, function}**
god	company	math
call	market	number
give	business	function
man	year	code
time	product	set
{city, large, area}	**{math, energy, light}**	**{law, state, case}**
city	math	law
large	energy	state
area	light	case
station	field	court
include	star	legal





Inference

- What are the topic distributions for each document?
- What are the topic assignments for each word in a document?
- What are the word distributions for each topic?



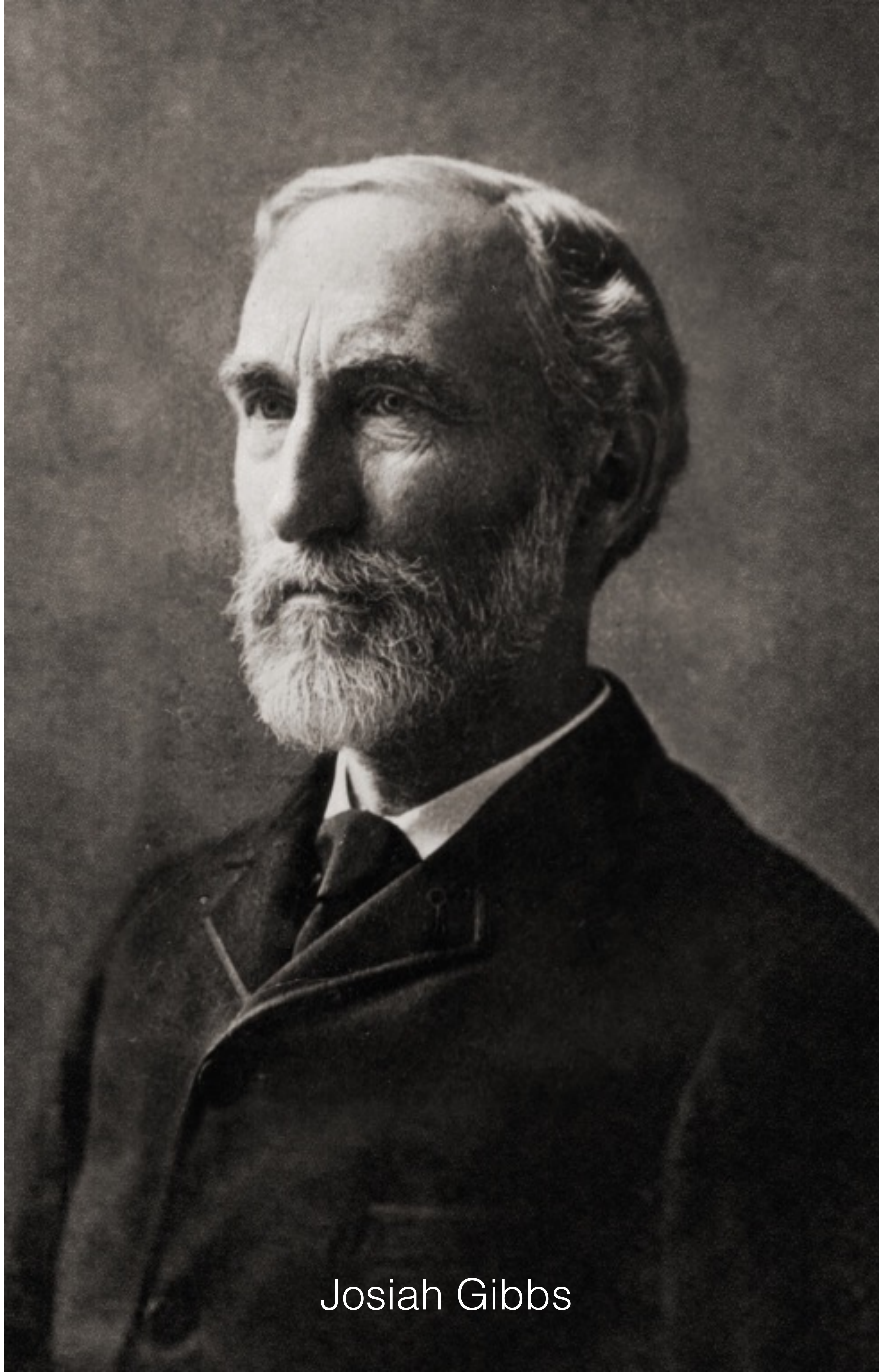
Find the parameters that maximize the likelihood of the data!

Inference

- Markov chain Monte Carlo (Gibbs sampling, Metropolis Hastings, etc.)
- Variational methods
- Spectral methods (Anandkumar et al. 2012, Arora et al. 2013)

Gibbs Sampling

- Markov chain Monte Carlo method for approximating the joint distribution of a set of variables (Geman and Geman 1984; Metropolis et al. 1953; Hastings et al. 1970)

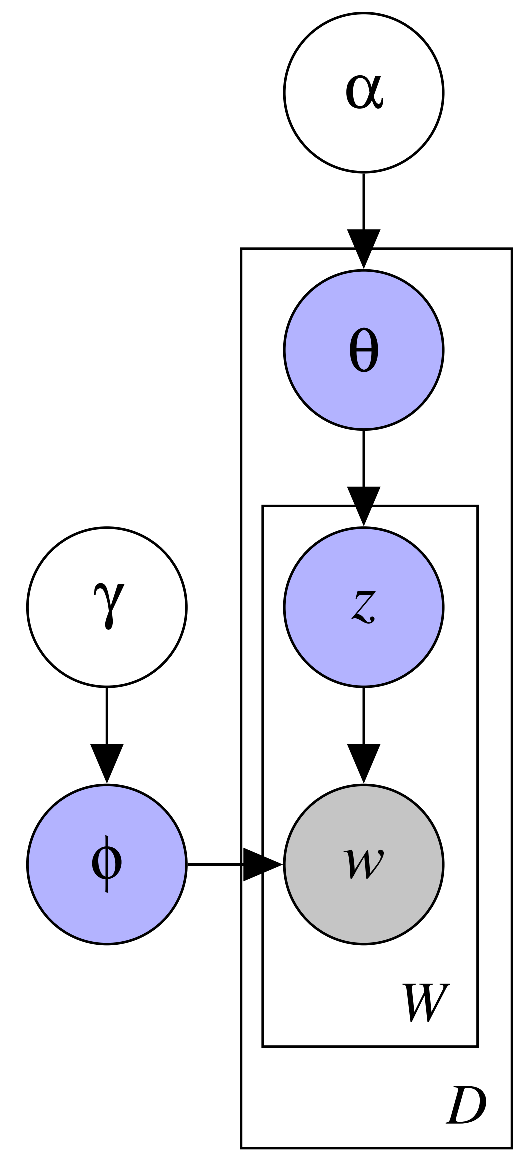


Gibbs Sampling

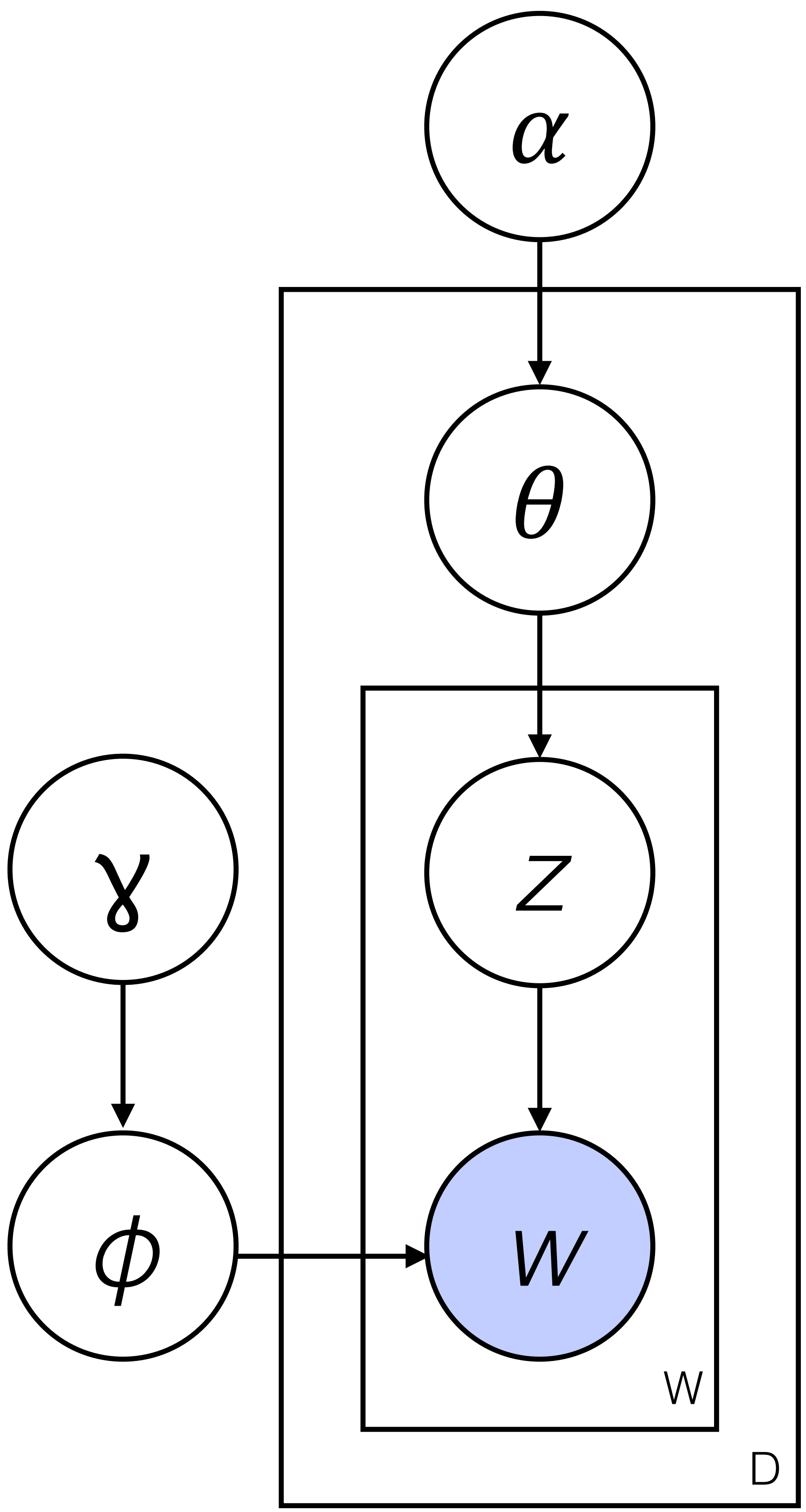
1. Start with some initial value for all the variables

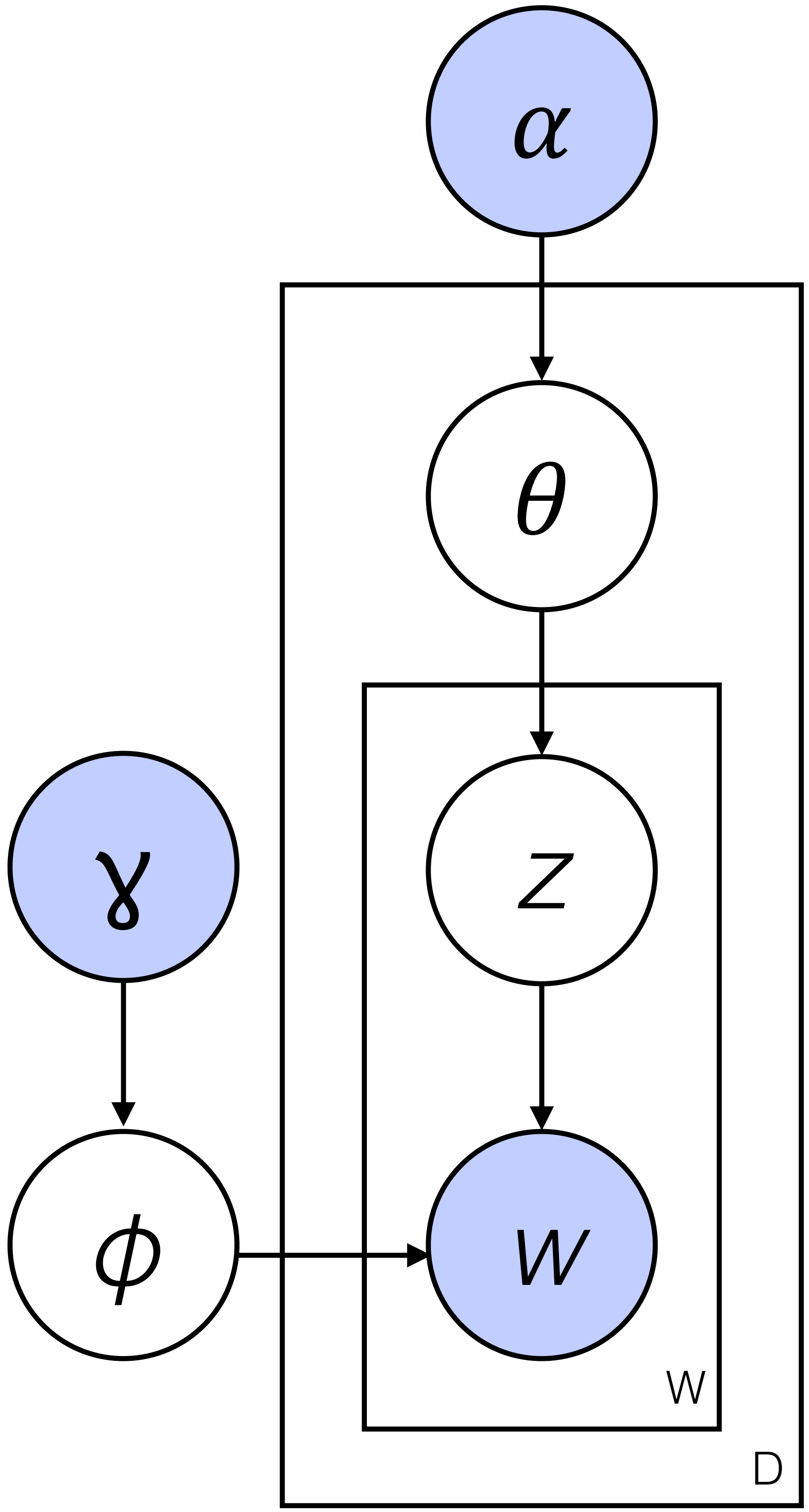
2. Sample a value for a variable conditioned on all of the other variables around it (using Bayes' theorem)

$$P(\theta|X) = \frac{P(\theta)P(X|\theta)}{\sum_\theta P(\theta)P(X|\theta)}$$



Inference



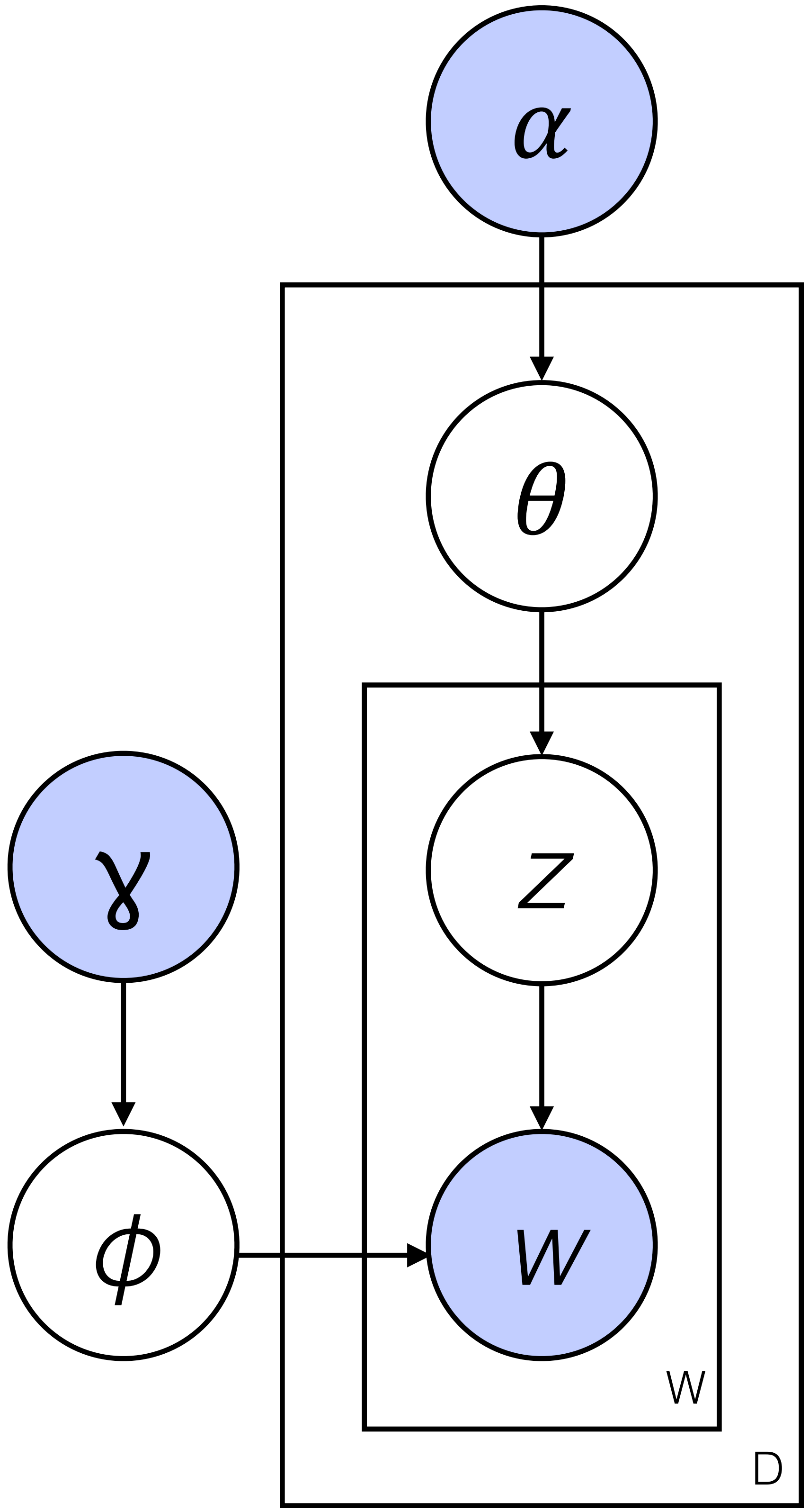


Inference

$$P(\theta_d \mid \alpha, \mathbf{z_d})$$

$$\propto P(\theta_d \mid \alpha) \prod_i P(z_i \mid \theta_d)$$

$$\propto \mathrm{Dir}(\theta \mid \alpha) \prod_i \mathrm{Cat}(z_i \mid \theta)$$



Inference

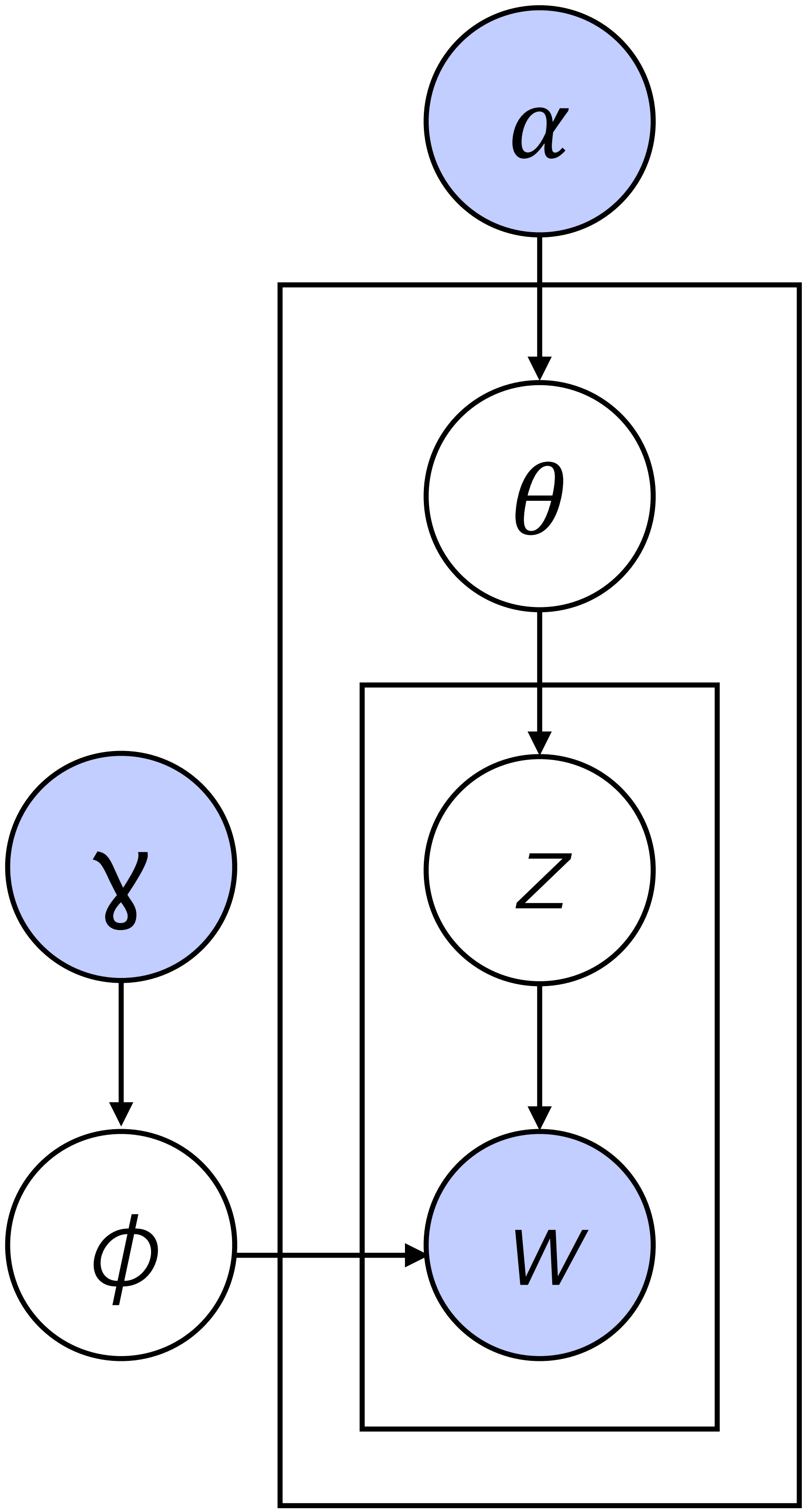
$$P(z \mid \theta_d, w, \phi)$$

$$\propto P(z \mid \theta_d) P(w \mid z, \phi)$$

$$\propto \mathrm{Cat}(z \mid \theta_d) \mathrm{Cat}(w \mid z, \phi)$$

$$\propto \theta_d^z \times \phi_z^w$$

Sampling



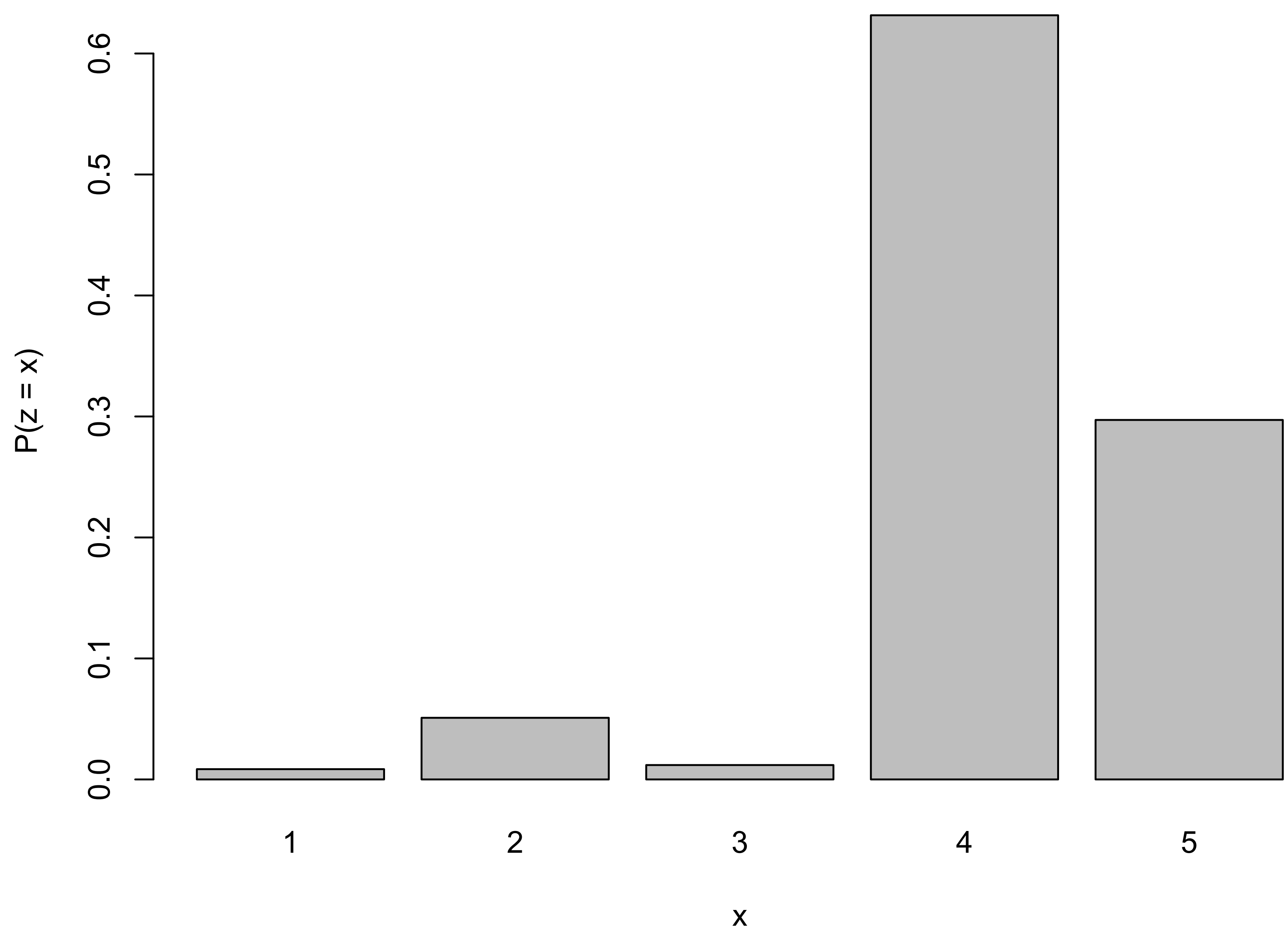
	$P(z\|\theta)$	$P(w\|z)$	$P(z\|\theta)$ $P(w\|z)$	norm
z=1	0.100	0.010	0.001	0.019
z=2	0.200	0.030	0.006	0.112
z=3	0.070	0.020	0.001	0.026
z=4	0.130	0.080	0.010	0.193
z=5	0.500	0.070	0.035	0.651

Aside: sampling?

Sampling from a Multinomial

Probability mass function (PMF)

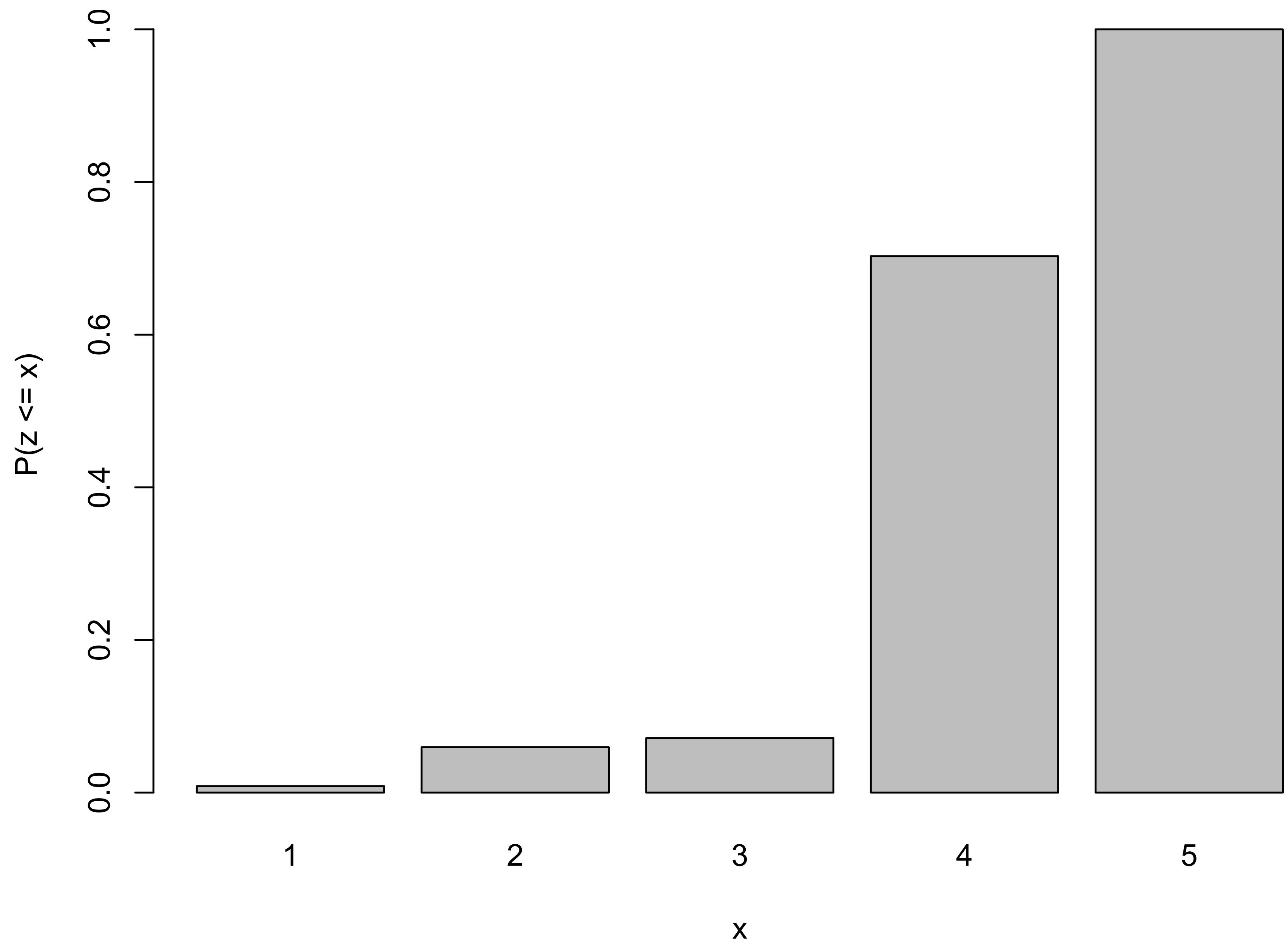
P(z = x) exactly



Sampling from a Multinomial

Cumulative density function (CDF)

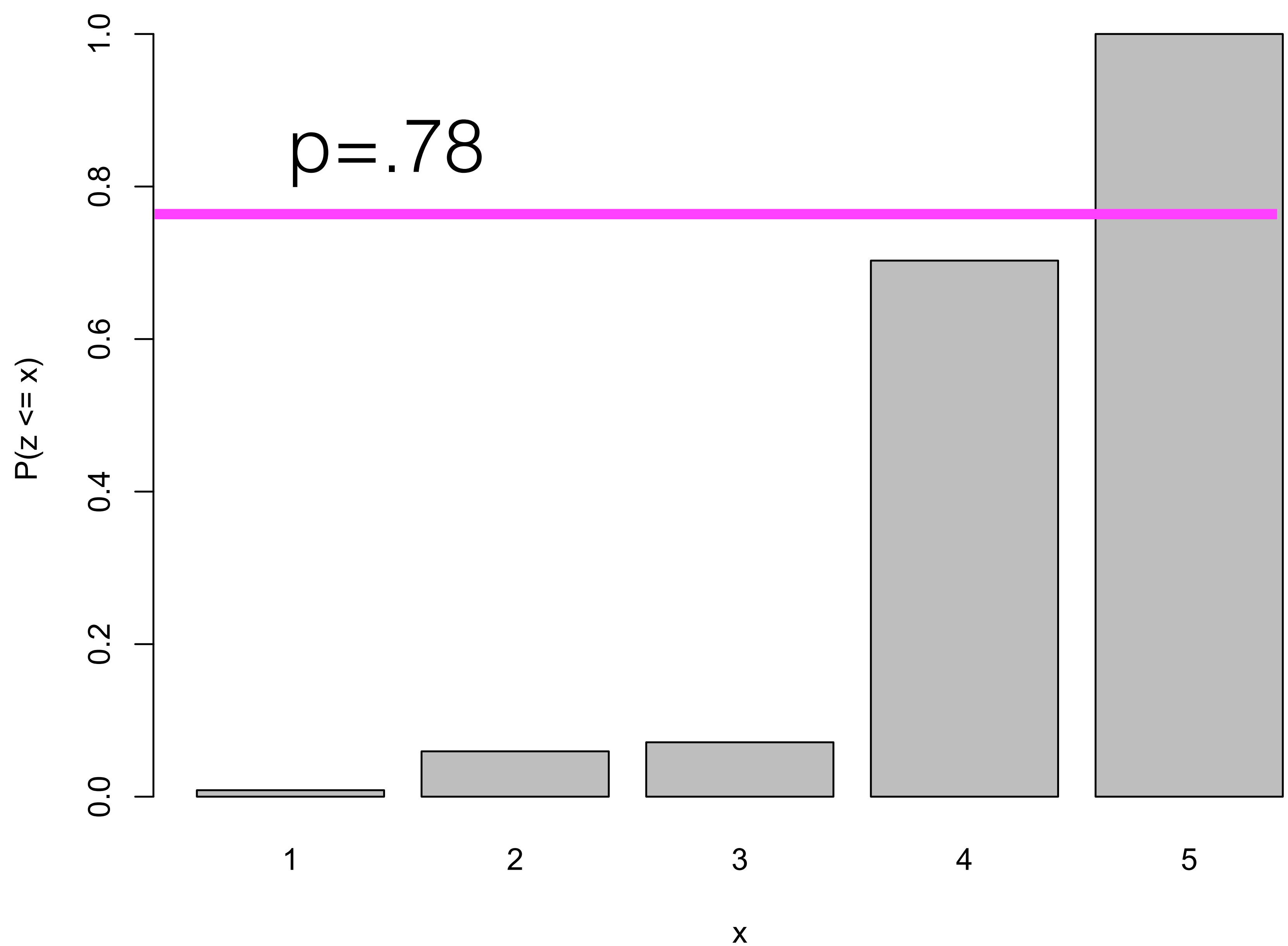
$P(z \leq x)$



Sampling from a Multinomial

Sample *p* uniformly in [0,1]

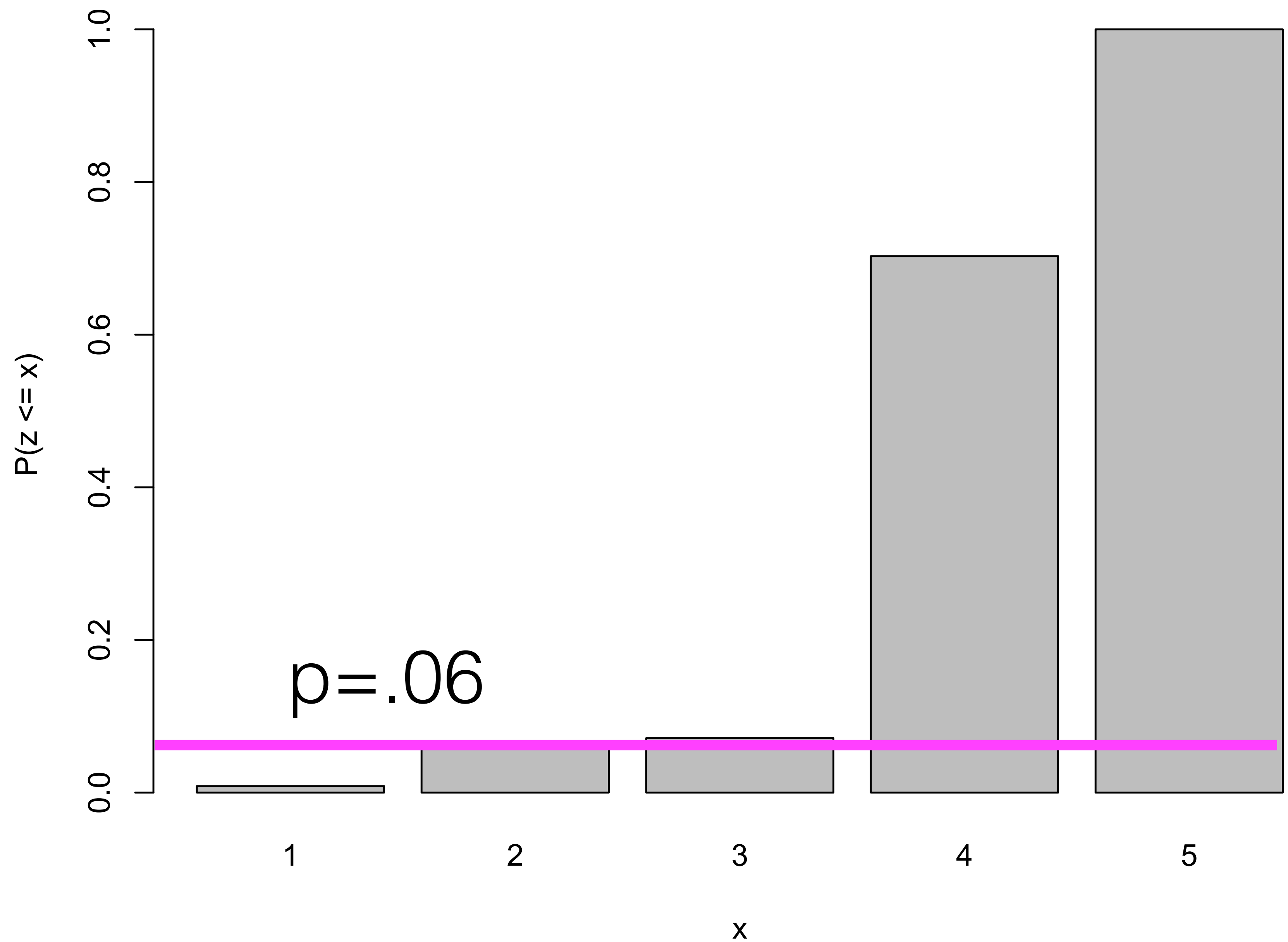
Find the point $CDF^{-1}(p)$



Sampling from a Multinomial

Sample *p* uniformly in [0,1]

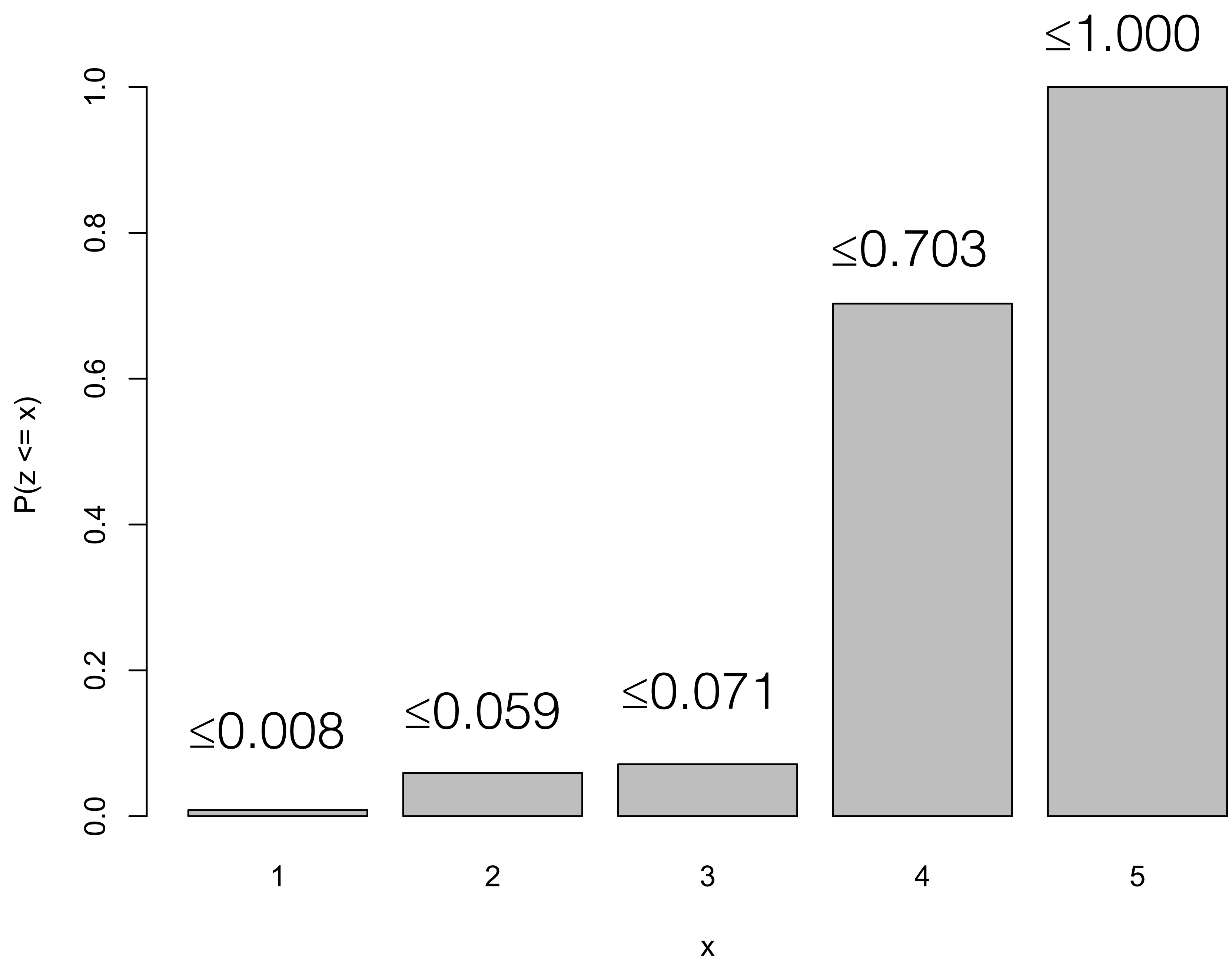
Find the point $CDF^{-1}(p)$

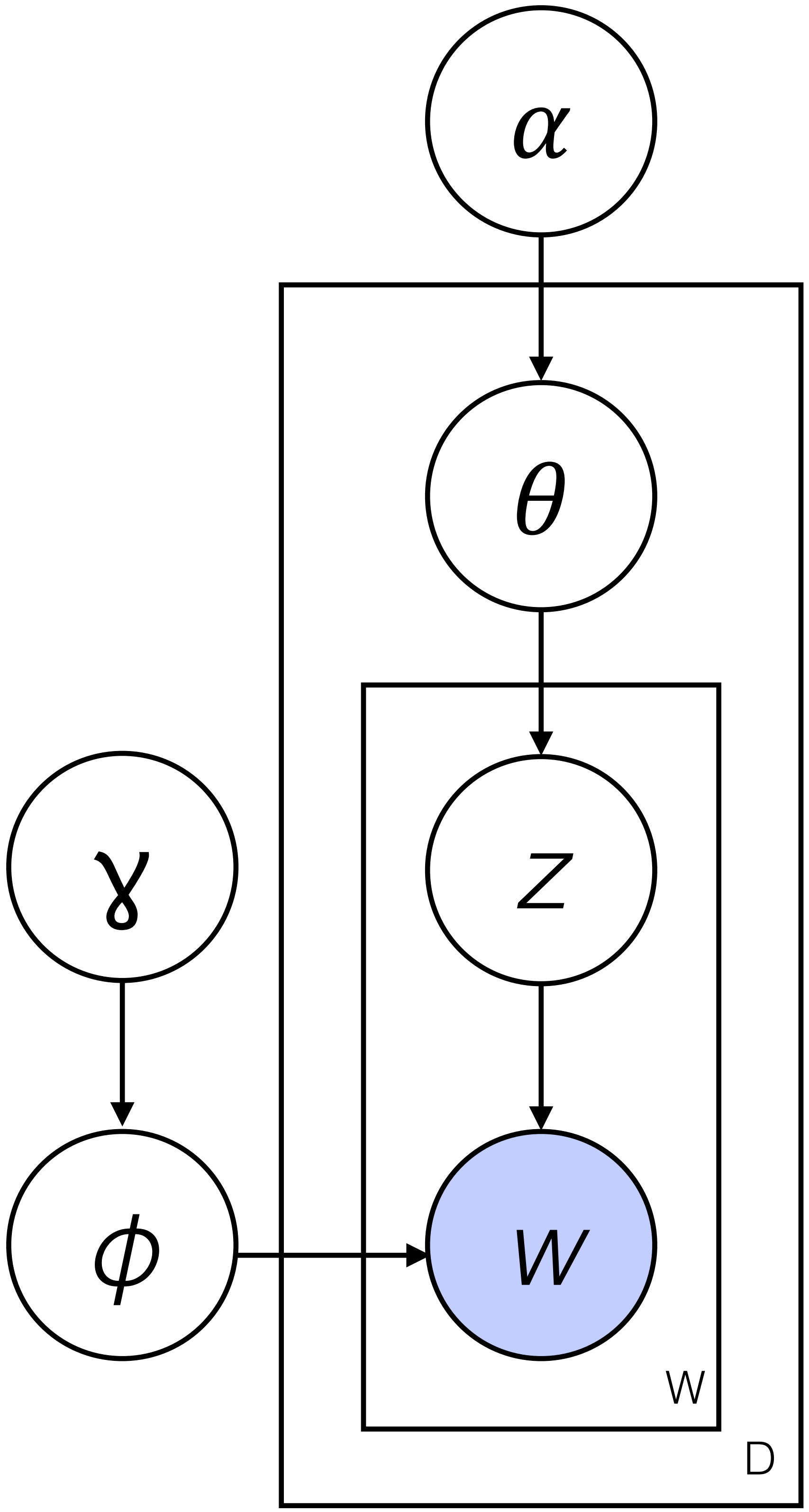


Sampling from a Multinomial

Sample *p* uniformly in [0,1]

Find the point $CDF^{-1}(p)$

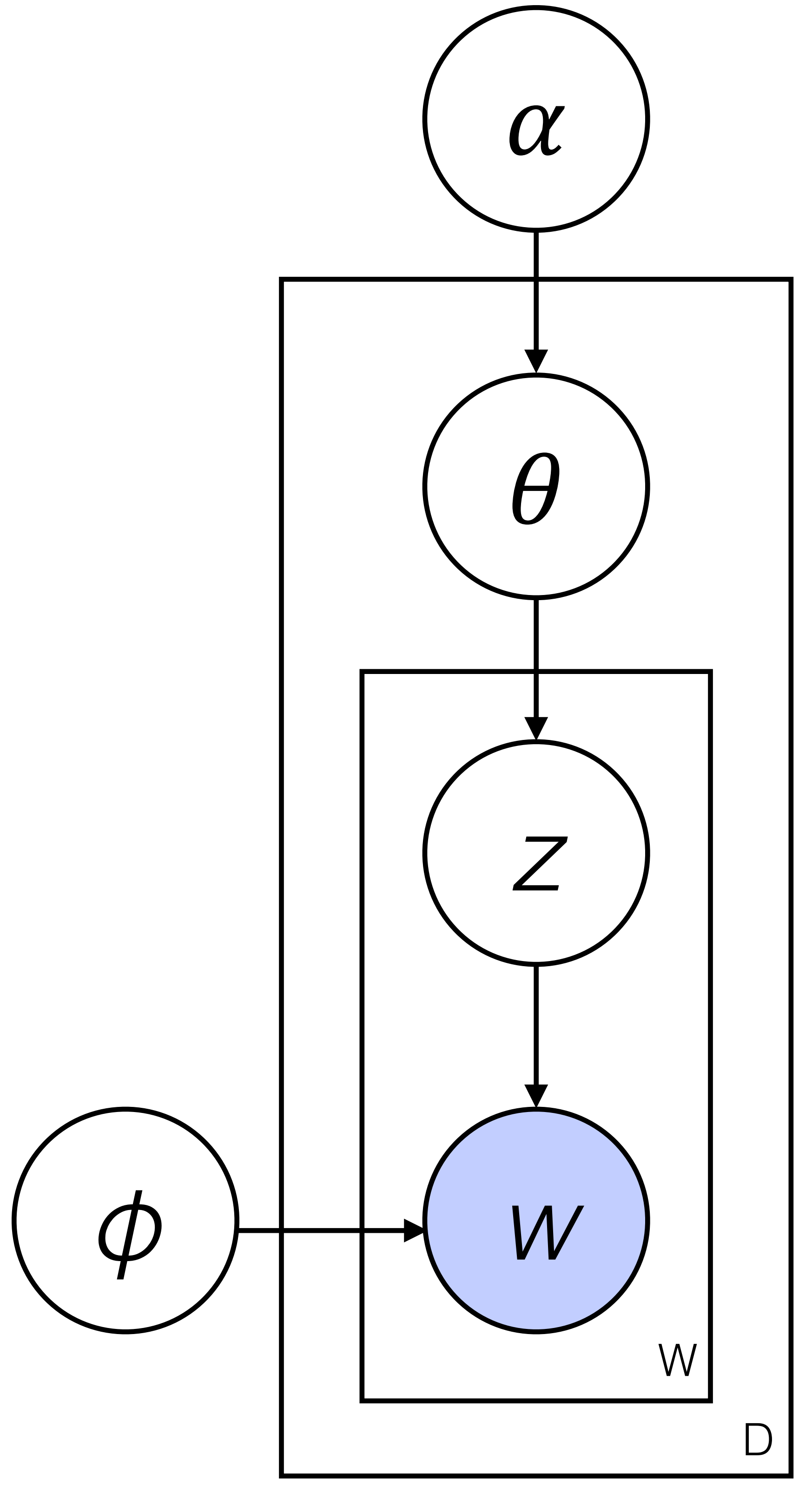


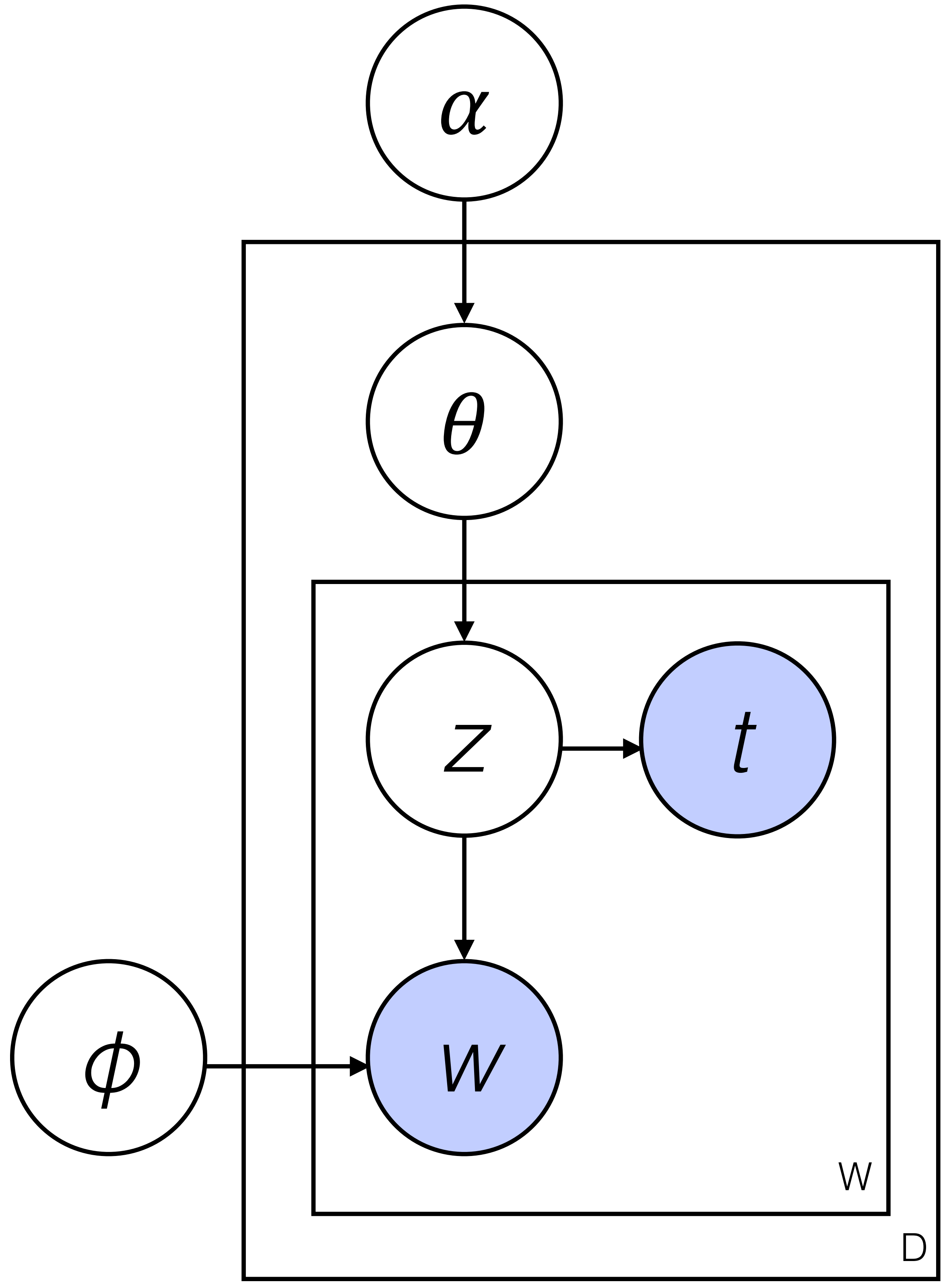


Assumptions

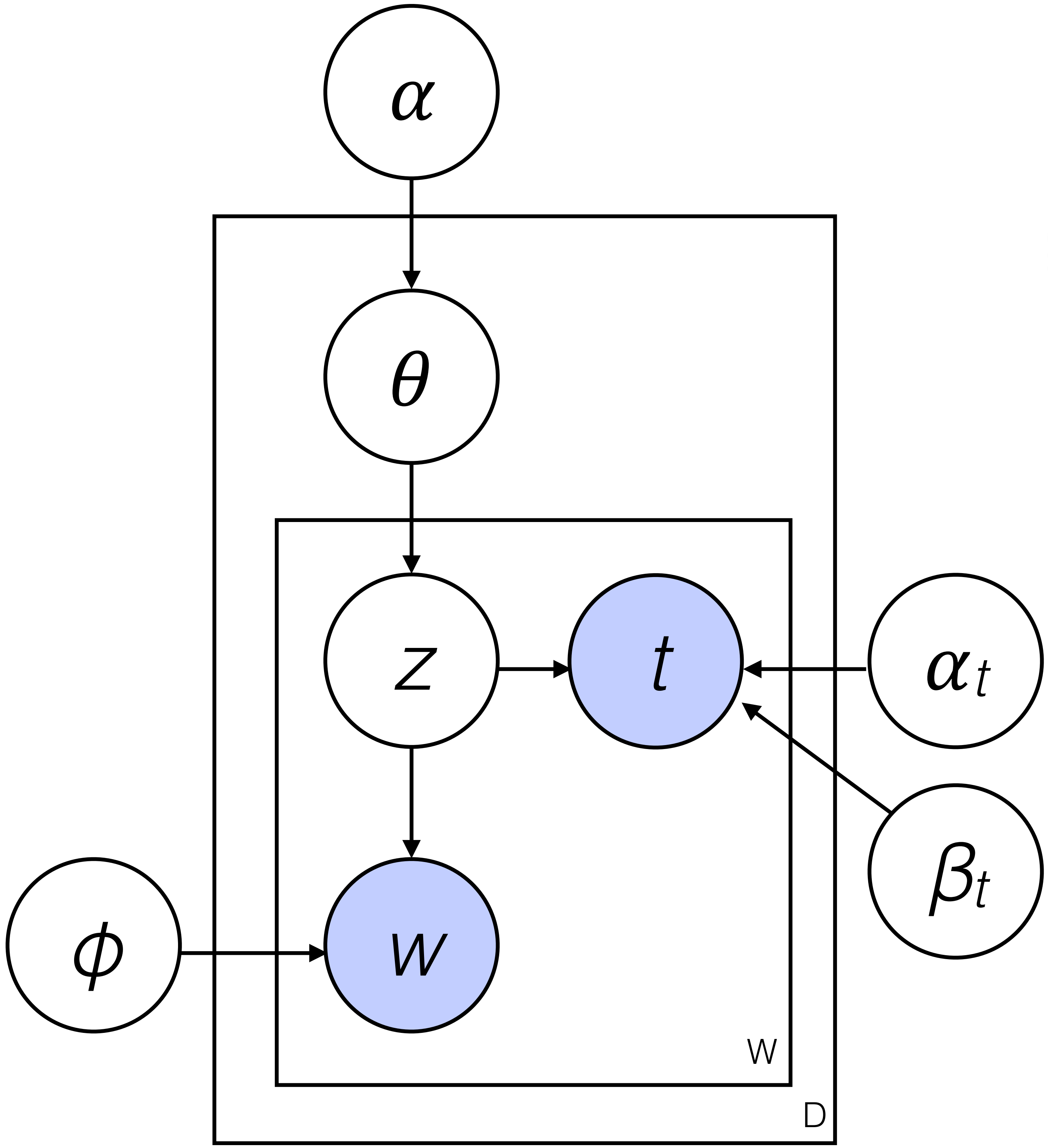
- Every word has one topic
- Every document has one topic distribution
- No sequential information (topics for words are independent of each other given the set of topics for a document)
- Topics don't have arbitrary correlations (Dirichlet prior)
- Words don't have arbitrary correlations (Dirichlet prior)
- The only information you learn from are the identities of **words** and how they are divided into **documents**.

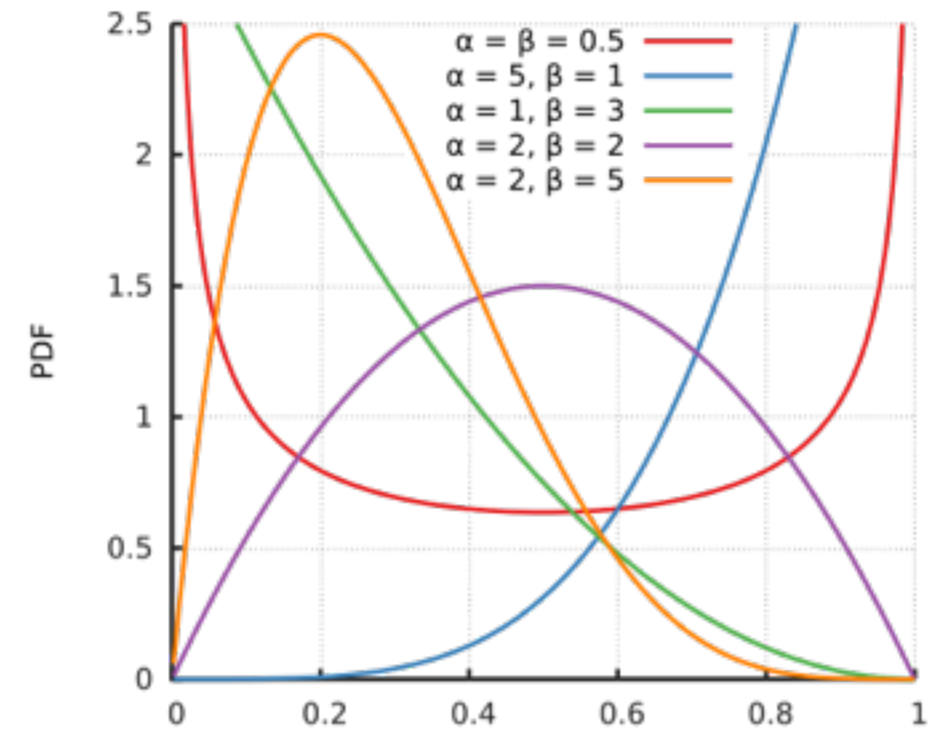
What if you want to encode other assumptions or reason over other observations?


α
θ
z
ϕ
w
W
D



α
θ
z
t
ϕ
w
W
D

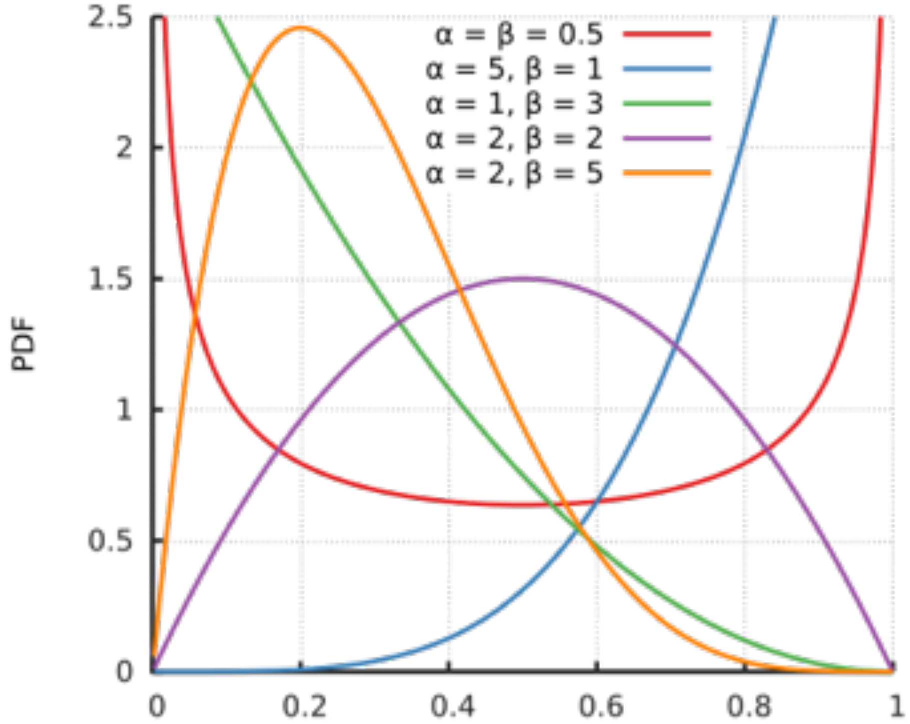


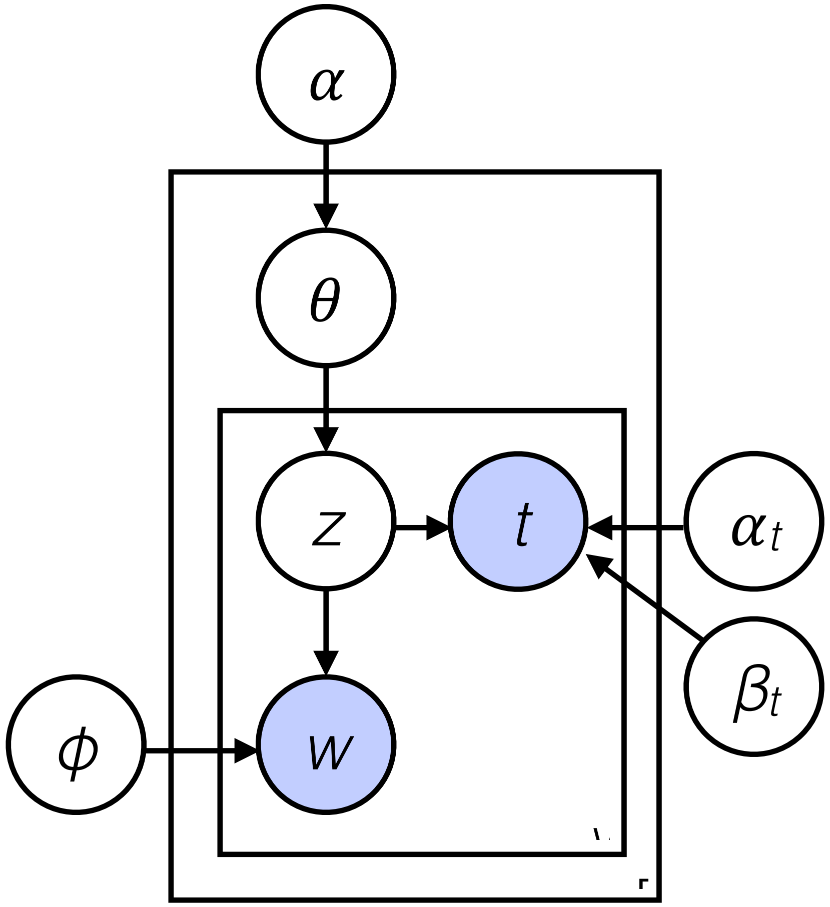


Time is drawn from a Beta distribution

[0,1]

(Wang and McCallum 2006)



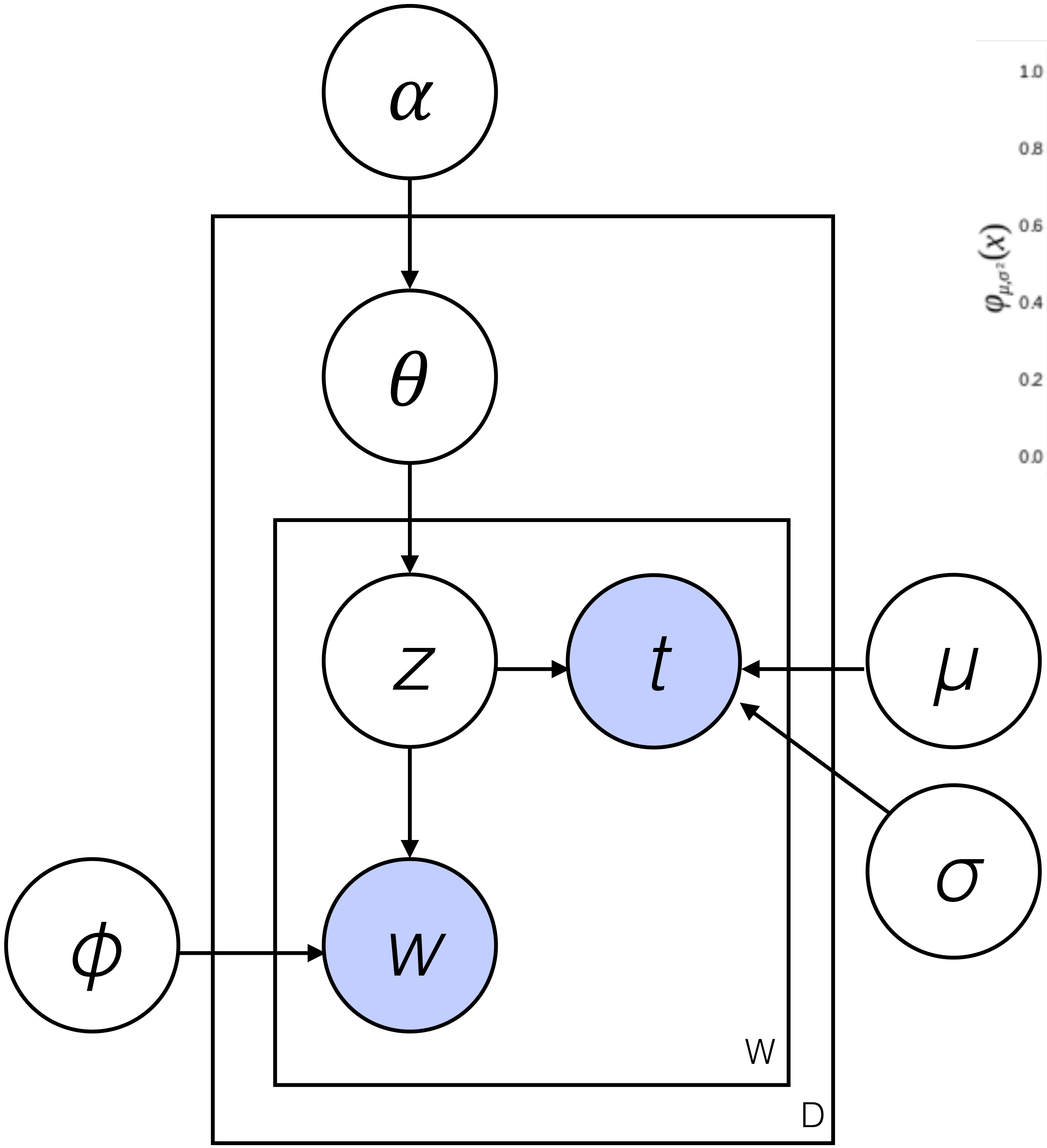


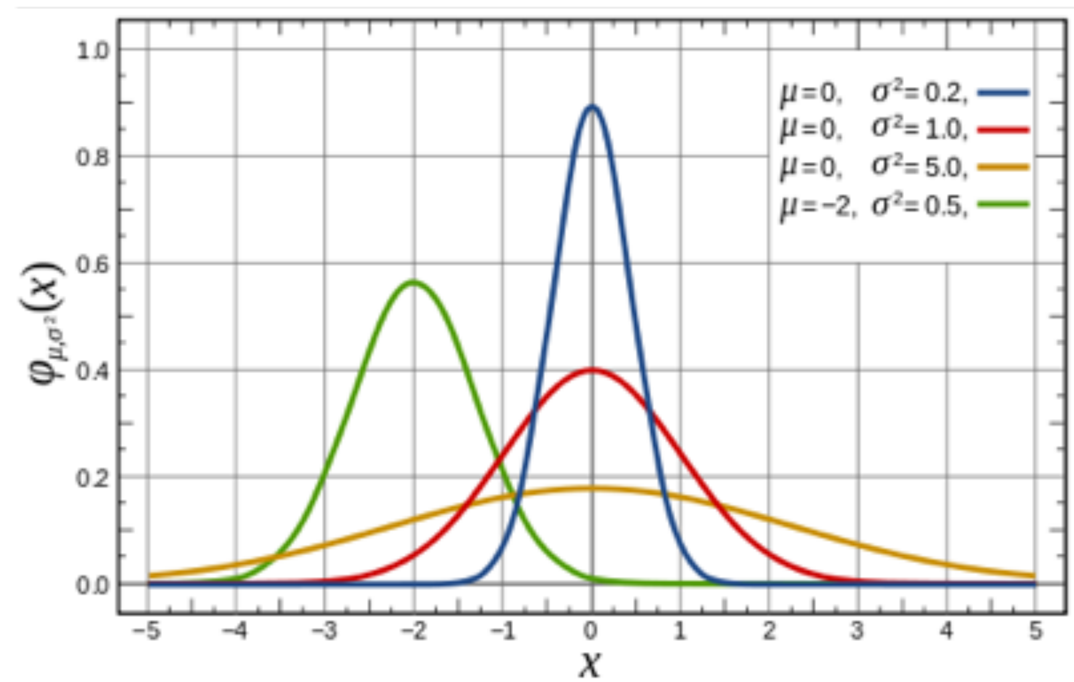
$$P(z \mid \theta, w, t, \phi, \alpha_t, \beta_t)$$

$$\propto P(z \mid \theta_d) P(w \mid z, \phi) P(t \mid z, \alpha, \beta)$$

$$\propto \mathrm{Cat}(z \mid \theta_d) \mathrm{Cat}(w \mid z, \phi) \mathrm{Beta}(t \mid \alpha_t, \beta_t)$$

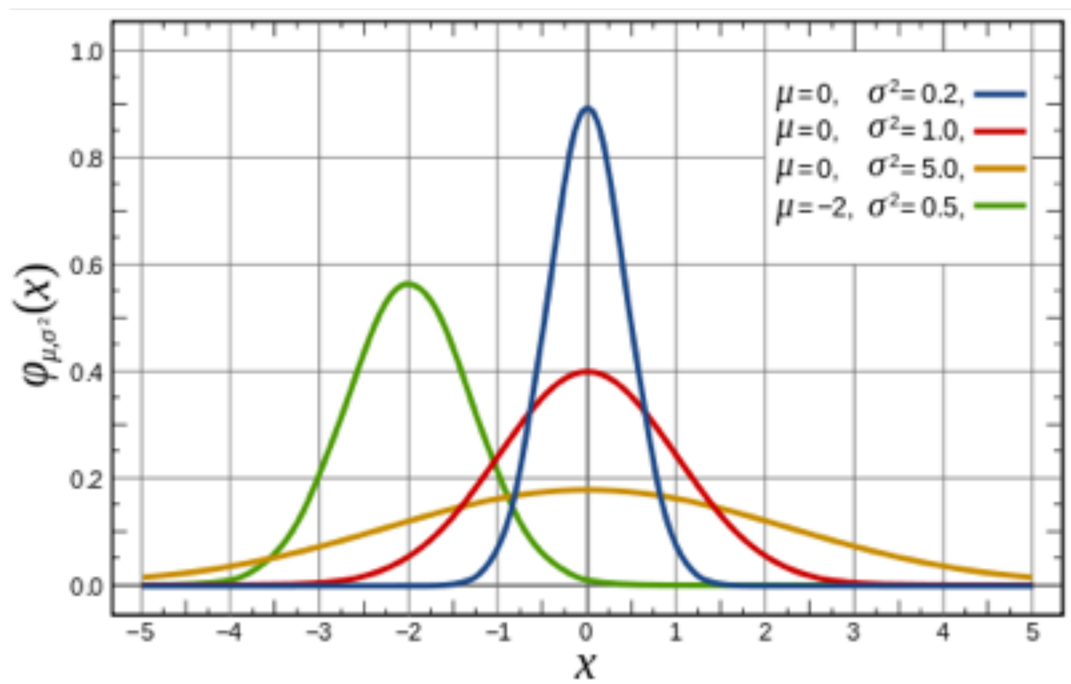
$$\propto \theta_d^z \times \phi_z^w \times \frac{t^{\alpha_t - 1}(1-t)^{\beta_t - 1}}{B(\alpha_t, \beta_t)}$$

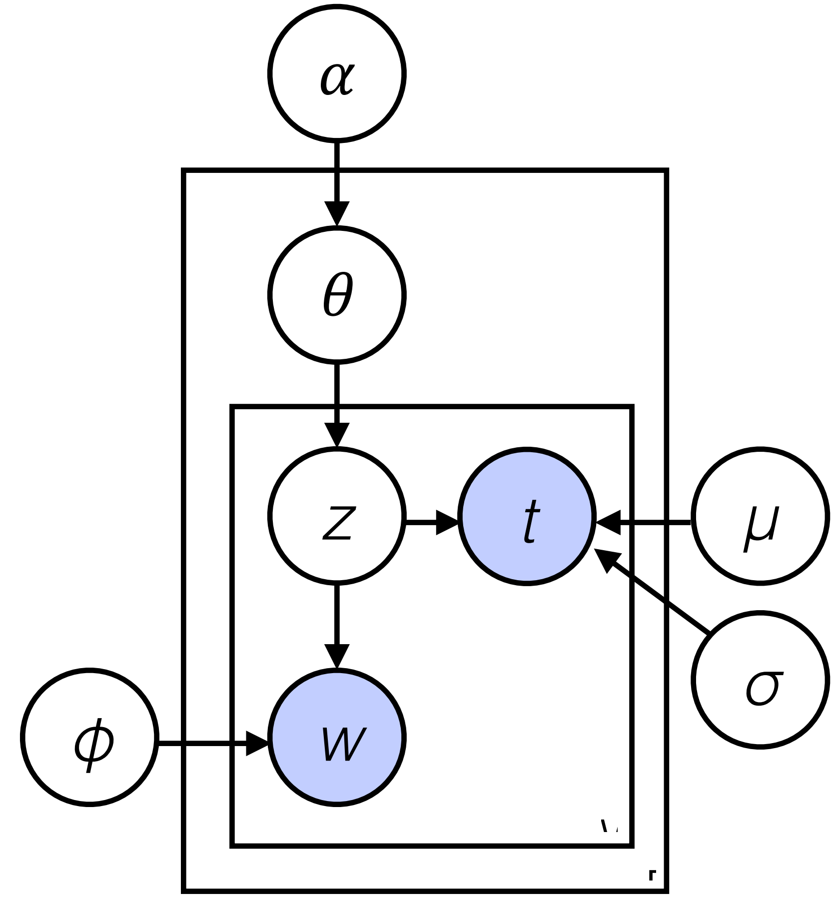




Time is drawn from a Normal distribution

[-∞, ∞]



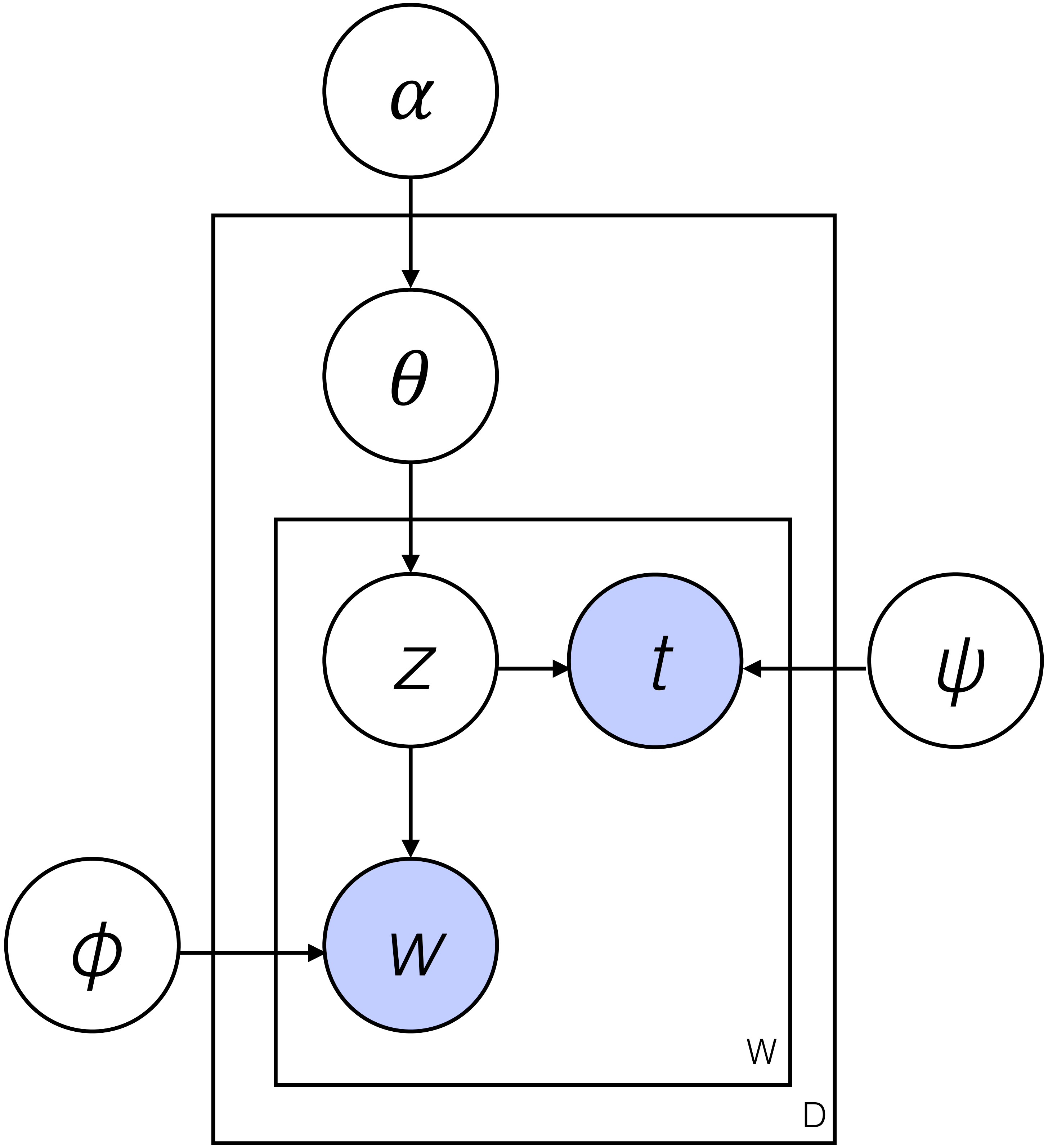


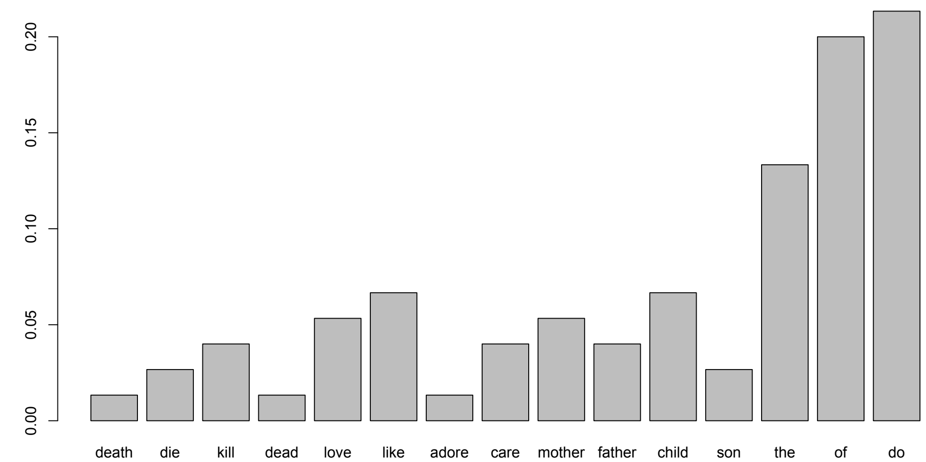
$$P(z \mid \theta, w, t, \phi, \mu, \sigma)$$

$$\propto P(z \mid \theta_d)P(w \mid z, \phi), P(t \mid z, \mu_z, \sigma_z)$$

$$\propto \mathrm{Cat}(z \mid \theta_d)\mathrm{Cat}(w \mid z, \phi)\mathrm{Norm}(t \mid \mu_z, \sigma_z)$$

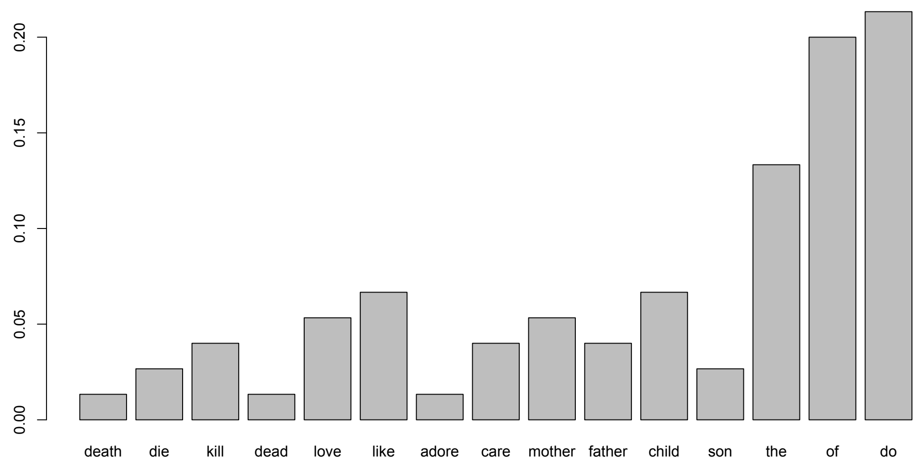
$$\propto \theta_d^z \times \phi_z^w \times \frac{1}{\sigma_z\sqrt{2\pi}} \exp\left(-\frac{(t-\mu_z)^2}{2\sigma_z^2}\right)$$

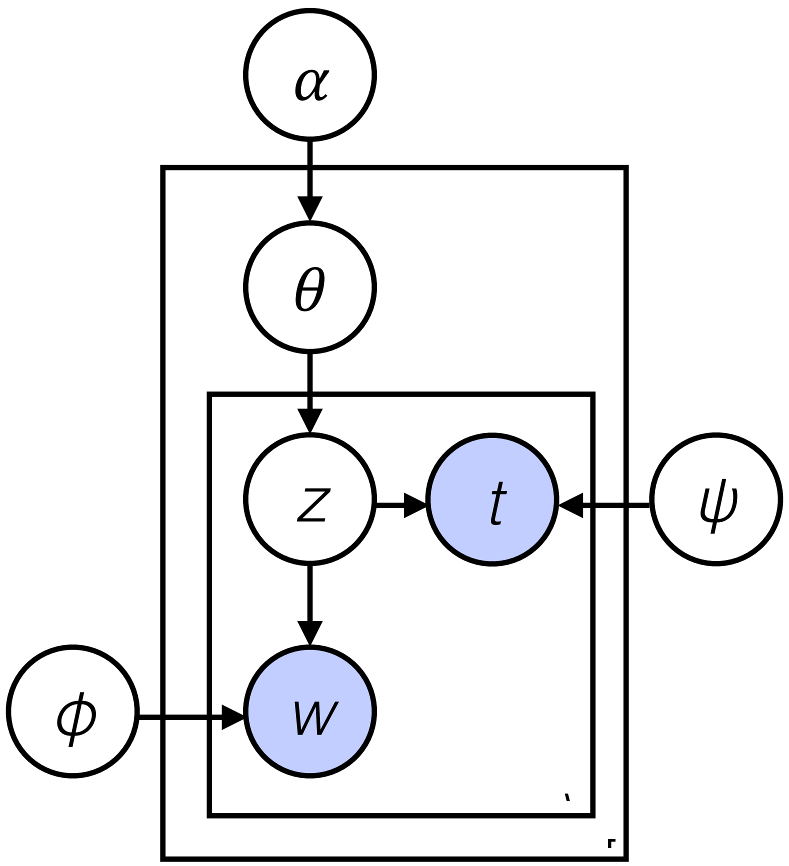




Time is drawn from a Multinomial distribution

[1, … , K]





$$P(z \mid \theta, w, \phi, t, \psi)$$

$$\propto P(z \mid \theta_d) P(w \mid z, \phi) P(t \mid z, \psi)$$

$$\propto \mathrm{Cat}(z \mid \theta_d) \mathrm{Cat}(w \mid w, \phi) \mathrm{Cat}(t \mid z, \psi)$$

$$\propto \theta_d^z \times \phi_z^w \times \psi_z^t$$

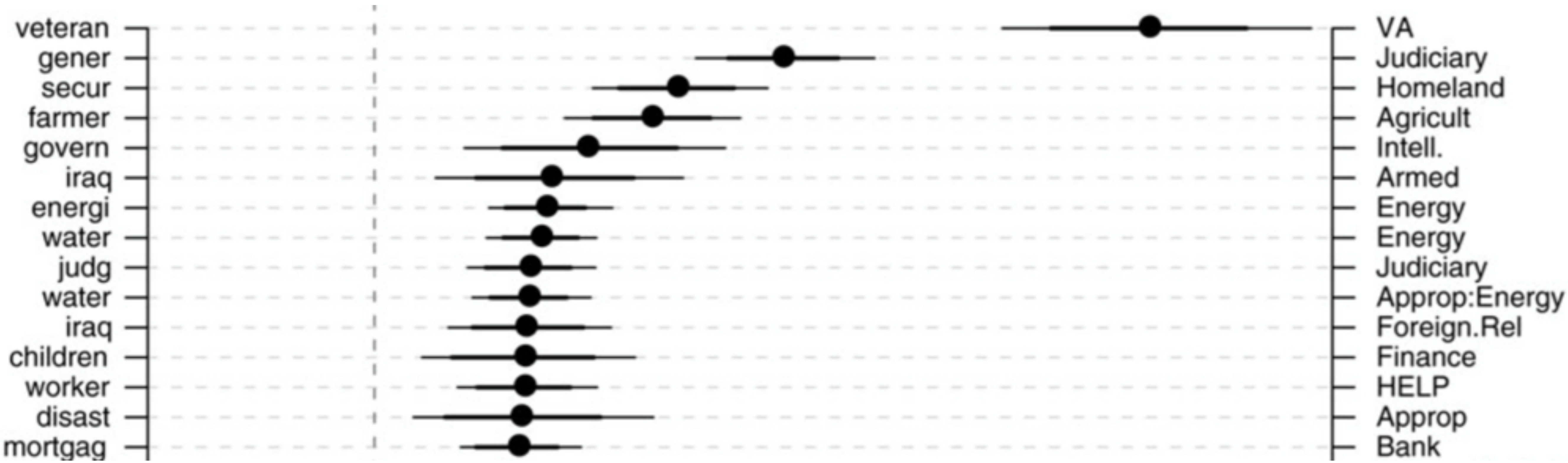
A Topic Model of Literary Studies Journals

Overview Topic Article Word Bibliography Word index Settings About

List Grid Years *click a column label to sort; click a row for more about a topic*

topic	1889—2013	top words	proportion of corpus
1		see both own view role university further account critical particular	2.5%
2		other both two form same even each part experience process	2.6%
3		old beowulf english ic mid swa pe poet ond grendel	0.3%

Goldstone and Underwood (2014),
The Quiet Transformations of Literary Studies



Grimmer (2010), A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases