Natural Language Processing

Info 159/259

Lecture 24: Latent variable models (April 20, 2023)

David Bamman, UC Berkeley

Random variable

- A variable that can take values within a fixed set (discrete) or within some range (continuous).

event	event space
dice throw	{1, 2, 3, 4, 5, 6}
the next word I say	{the, a, dog, runs, to, store}
author of a text	{Austen, Dickens}
height of a skyscraper	$[0, \infty]$

Note this includes both data (X) and labels we're predicting (Y) — they can all be thought of as random variables

Joint probability

weather	hot	cloudy	rainy	hot	hot	cloudy	rainy
ice cream	1	0	0	1	1	1	0

$$P(X = \text{hot}, Y = \text{ice cream})$$

The probability of multiple things happening at the same time.

Chain Rule of Probability

$$P(X, Y) = P(X)P(Y \mid X)$$

Joint probability

weather	hot	cloudy	rainy	hot	hot	cloudy	rainy
ice cream	1	0	0	1	1	1	0

$$P(X, Y) = P(X)P(Y \mid X)$$

	hot	cloudy	rainy
$P(X = x)$	3/7 = 0.42	2/7 = 0.29	2/7 = 0.29
$P(Y = \text{ice cream} \mid X = x)$	3/3 = 1.0	1/2 = 0.50	0.2 = 0.0

$$P(X = \text{hot}, Y = \text{ice cream}) = 0.42$$

Latent variables

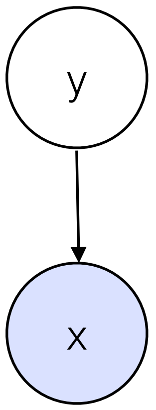
- A latent variable is one that’s unobserved, either because:
    - we are predicting it (but have observed that variable for other data points)
    - it is unobservable

Latent variables

	observed variables	latent variables
email	text, date, sender	topic, urgency
novels	text, author, date	genre, happy ending, archetypes
netflix users	viewing data	preferences

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:



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

Classification

P(y = dickens | x = “it was the best of times”)

Bayes' Rule

Prior belief that Y = positive
(before you see any data)

Likelihood of "it was the best of times"
given that Y= dickens

$$P(Y=y|X=x) = \frac{P(Y=y)P(X=x|Y=y)}{\sum_y P(Y=y)P(X=x|Y=y)}$$

Posterior belief that Y=dickens given that X="it was the best of times"

This sum ranges over
y=dickens + y=austen
(so that it sums to 1)

Independence Assumption

it was the best of times

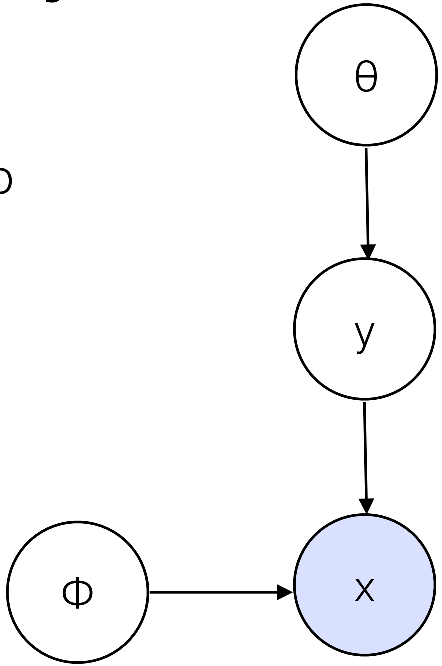
x_1 x_2 x_3 x_4 x_5 x_6

We will assume the features are independent:

$$P(x_1, \ldots, x_N \mid y) = \prod_{i}^{N} P(x_i \mid y)$$

Naive Bayes

- To fully specify Naive Bayes, we need to add the implicit parameters θ (the prior distribution) and Φ (the distribution of x given y).

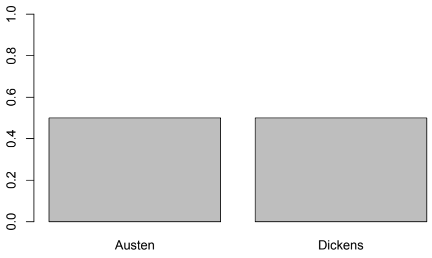


$$P(x, y \mid \theta, \phi) = P(y \mid \theta)\, P(x \mid y, \phi)$$

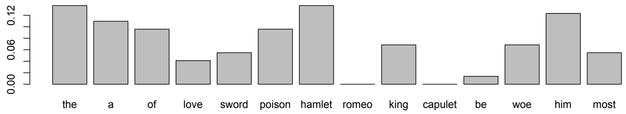
$$P(y = \text{Austen} \mid \theta) = 0.5$$

Look up the value of y in θ

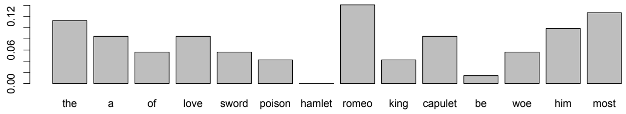
θ



Φ_{austen}



Φ_{dickens}

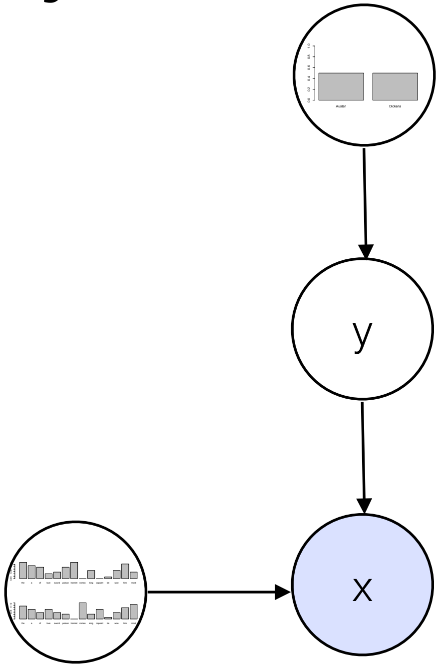


$$P(x = \text{love} \mid y = \text{Austen}, \phi) = 0.04$$

Look up the value of x in the Φ indexed by y

Naive Bayes

- We can plug these multinomials in to make this more clear



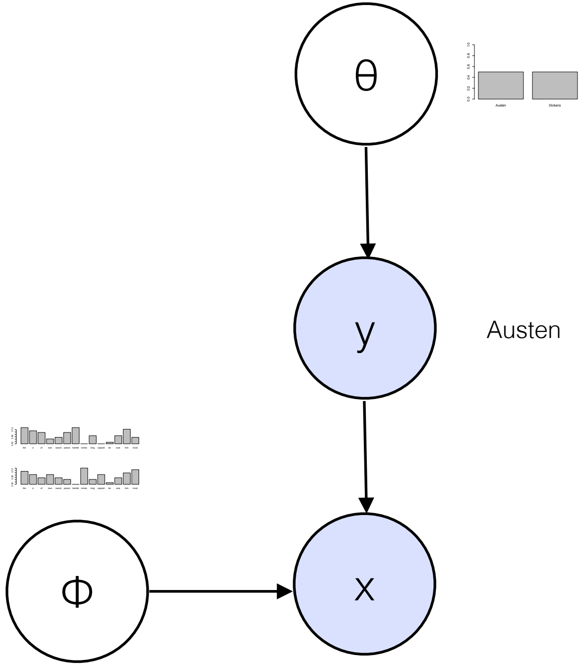
$$P(x, y \mid \theta, \phi) = P(y \mid \theta)\; P(x \mid y, \phi)$$

Naive Bayes

- When we train Naive Bayes, y is observed, and we estimate the parameters θ and Φ with (e.g.) maximum likelihood estimation

$$\theta_i = \frac{count(i)}{N}$$

$$\phi_{y,i} = \frac{count(y,i)}{N_y}$$



Naive Bayes MLE

$$\theta_i = \frac{count(i)}{N}$$

The number of Austen texts divided by the total number of texts

$$\phi_{y,i} = \frac{count(y,i)}{N_y}$$

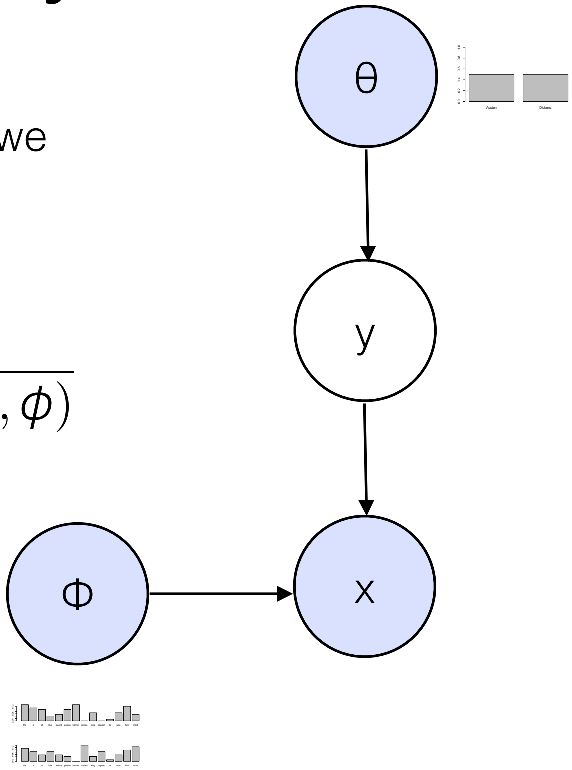
The number of times “love” appears in Austen texts divided by the total number of words in Austen texts

Naive Bayes

- When we predict, y is no longer observed (we are predicting it), but Φ and θ are.

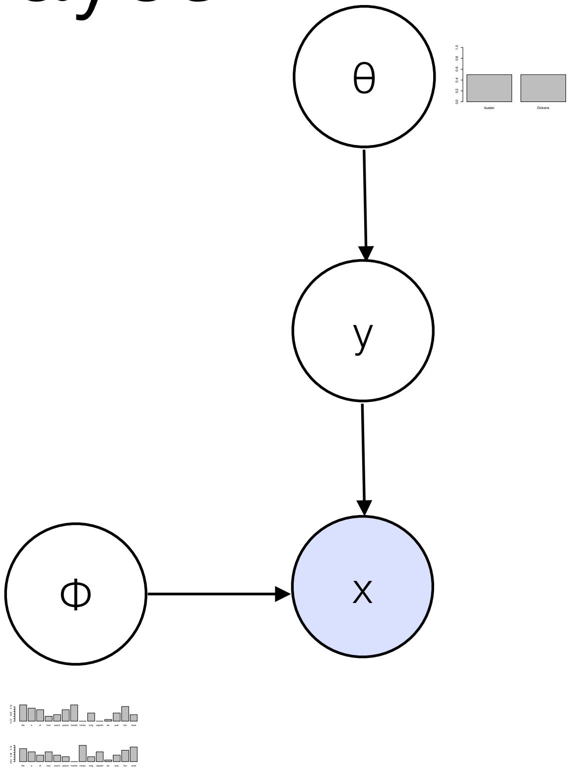
$$P(y \mid x, \theta, \phi) = \frac{P(y \mid \theta)P(x \mid y, \phi)}{\sum_{y' \in \mathcal{Y}} P(y' \mid \theta)P(x \mid y', \phi)}$$

- We calculate the posterior probability of y using Bayes' rule



Unsupervised Naive Bayes

- Same model structure
- Same conditional relationships
- No observed labels y
- Why would we do this??



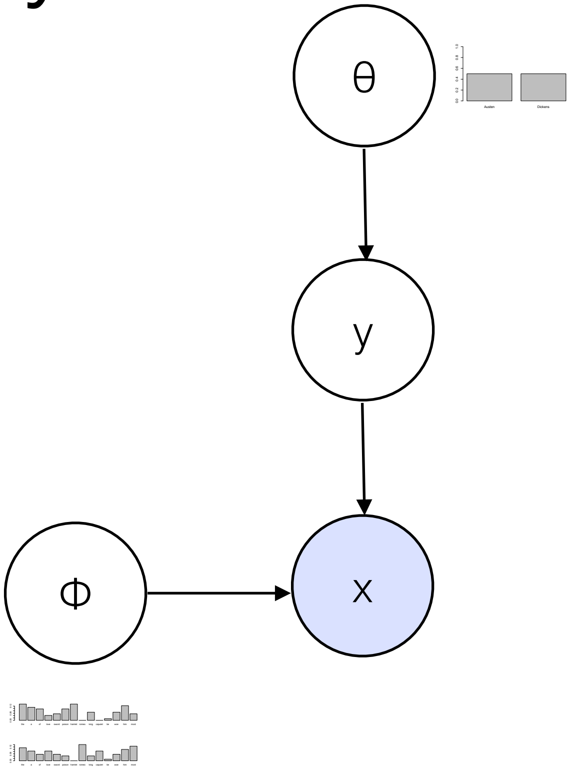
Structure

- Unsupervised learning finds *structure* in data.

Unsupervised Naive Bayes

- The only variables we observe are the data x
- But we still want to estimate Φ and θ and learn posterior probabilities for y
- y here is still a choice among K alternatives:

$$\mathcal{Y} = \{1, 2, \ldots K\}$$



Inference

- We want to estimate the best values of the parameters Φ and θ and infer the most likely values for latent variables y

Inference

- Guiding principle: we want to maximize the likelihood of the observed data

$$P(x \mid \phi, \theta) = \sum_{y \in \mathcal{Y}} P(x, y \mid \phi, \theta)$$

$$P(x \mid \phi, \theta) = \sum_{y \in \mathcal{Y}} P(x \mid y, \phi)\, P(y \mid \theta)$$

Inference

$$\ell(\phi, \theta) = \sum_{i=1}^{N} \log P(x \mid \phi, \theta)$$

$$\ell(\phi, \theta) = \sum_{i=1}^{N} \log \sum_{y \in \mathcal{Y}} P(x \mid y, \phi)\, P(y \mid \theta)$$

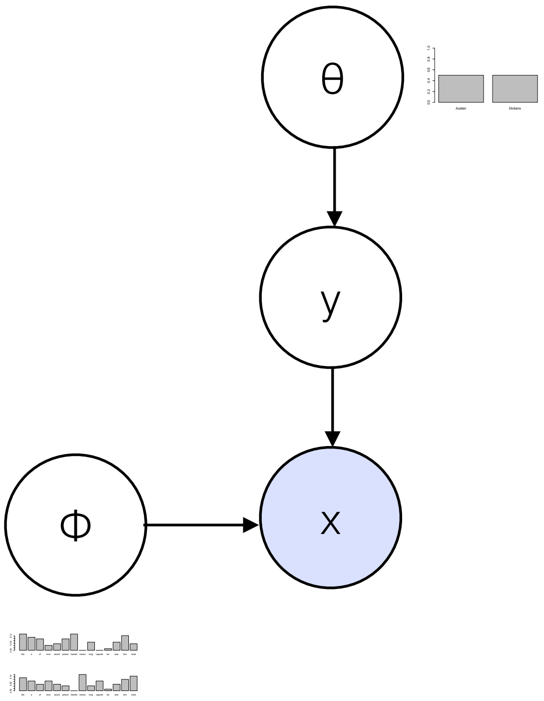
this sum in the log makes this
likelihood hard to optimize

Inference

Lots of standard inference techniques we can use

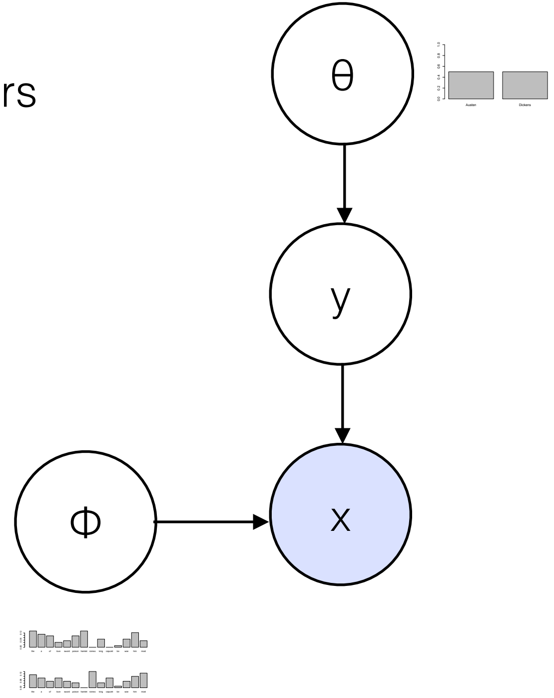
- Expectation Maximization
- Markov chain Monte Carlo (Gibbs sampling, Metropolis Hastings, etc.)
- Variational methods
- Spectral methods (Anandkumar et al. 2012, Arora et al. 2013)

Expectation Maximization



Expectation Maximization

- Start out with random values for the parameters
- Iterate until convergence:
    - Calculate expected values for latent variables y
    - Use those expected values to update parameters Φ and θ

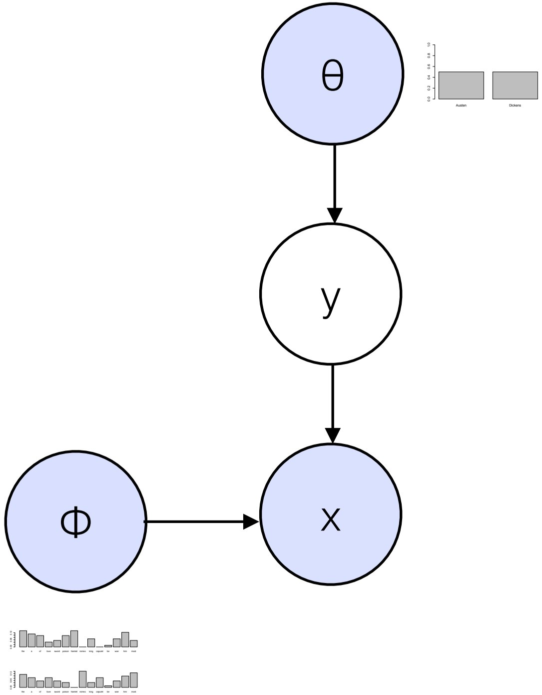


Expectation Maximization

1. Calculate expected values for latent variables

$$P(y \mid x, \theta, \phi) = \frac{P(y \mid \theta)P(x \mid y, \phi)}{\sum_{y' \in \mathcal{Y}} P(y' \mid \theta)P(x \mid y', \phi)}$$

1	2	3	4	5
0.10	0.50	0.25	0.07	0.08



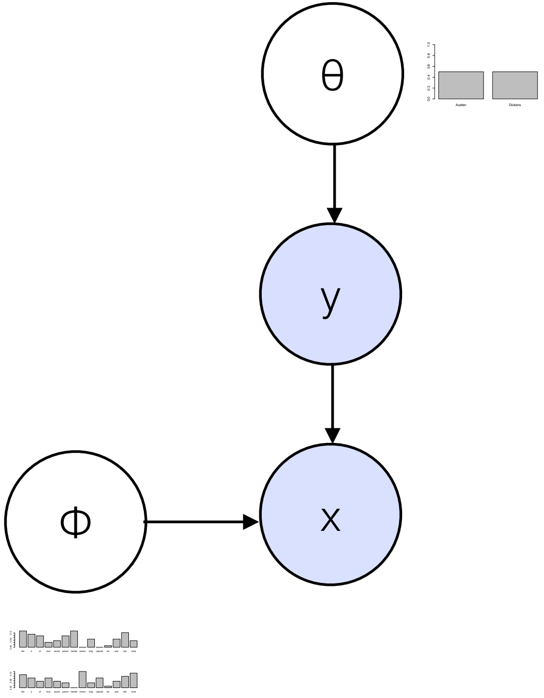
Expectation Maximization

Expected values for 10 data points, with K=5

	1	2	3	4	5
y1	0.35	0.03	0.12	0.27	0.23
y2	0.39	0.08	0.31	0.03	0.19
y3	0.05	0.36	0.22	0.1	0.27
y4	0.31	0.14	0.05	0.28	0.22
y5	0.65	0.05	0.17	0.07	0.06
y6	0.11	0.04	0.34	0.27	0.24
y7	0.07	0.07	0.45	0.02	0.39
y8	0.14	0.54	0.03	0.11	0.18
y9	0.51	0.06	0.09	0.29	0.05
y10	0.01	0.23	0.08	0.14	0.54

Expectation Maximization

2. Use those expected values to maximize parameters



Expectation Maximization

2. Use those expected values to maximize parameters

$$\theta_k = \frac{1}{N}\sum_{i=1}^{N} r_{i,k}$$

$r_{i,k}$ is proportion of the count we attribute to k

k

i

	1	2	3	4	5
y1	0.35	0.03	0.12	0.27	0.23
y2	0.39	0.08	0.31	0.03	0.19
y3	0.05	0.36	0.22	0.1	0.27
y4	0.31	0.14	0.05	0.28	0.22
y5	0.65	0.05	0.17	0.07	0.06
y6	0.11	0.04	0.34	0.27	0.24
y7	0.07	0.07	0.45	0.02	0.39
y8	0.14	0.54	0.03	0.11	0.18
y9	0.51	0.06	0.09	0.29	0.05
y10	0.01	0.23	0.08	0.14	0.54
avg	0.259	0.160	0.186	0.158	0.237

Expectation Maximization

2. Use those expected values to maximize parameters

$$\phi_{k,w} = \frac{\sum_{i=1}^{N} r_{i,k}\, count(i,w)}{\sum_{i=1}^{N} r_{i,k}\, N_i}$$

$r_{i,k}$ is proportion of the count we attribute to k

count(i,w) = count of word w in document i

N_i is the total word count in document i

k

i	1	2	3	4	5
y1	0.35	0.03	0.12	0.27	0.23
y2	0.39	0.08	0.31	0.03	0.19
y3	0.05	0.36	0.22	0.1	0.27
y4	0.31	0.14	0.05	0.28	0.22
y5	0.65	0.05	0.17	0.07	0.06
y6	0.11	0.04	0.34	0.27	0.24
y7	0.07	0.07	0.45	0.02	0.39
y8	0.14	0.54	0.03	0.11	0.18
y9	0.51	0.06	0.09	0.29	0.05
y10	0.01	0.23	0.08	0.14	0.54

Expectation Maximization

In general, EM involves iterating between two steps:

E-step: calculate the posterior probability of latent y

$$Q(y) = P(y \mid x_i, \theta)$$

M-step: find the values of parameters θ that maximize:

$$\theta = \arg\max_{\theta} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} Q(y) \log \frac{P(x_i, y \mid \theta)}{Q(y)}$$

Expectation Maximization

- Start out with random values for the parameters
- Iterate until convergence:
    - Calculate expected values for latent variables
    - Use those expected values to maximize parameter values

K-means

Given: a set $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_n\} \subseteq \mathbb{R}^m$
a distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$
a function for computing the mean $\mu : \mathcal{P}(\mathbb{R}) \to \mathbb{R}^m$
Select k initial centers $\vec{f}_1, \ldots, \vec{f}_k$
while stopping criterion is not true **do**
for all clusters c_j **do**
$c_j = \{\vec{x}_i \mid \forall \vec{f}_l \; d(\vec{x}_i, \vec{f}_j) \leq d(\vec{x}_i, \vec{f}_l)\}$
end
for all means $\vec{f}_j$ **do**
$\vec{f}_j = \mu(c_j)$
end
end

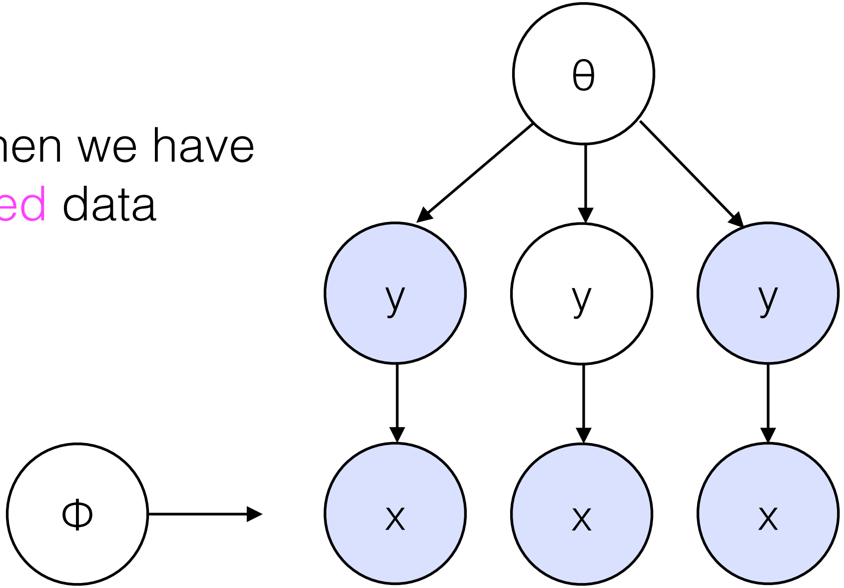
Expectation Maximization

Expectation maximization yields a soft clustering (where a given data point can have fractional membership in multiple clusters.

K-means is an approximation to this: instead of allowing fractional membership, each data point is placed into its single most likely cluster. Also known as "hard EM"

Semi-supervised

EM is useful for when we have partially labeled data



Semi-supervised

How would the presence of *some* supervised labels change your calculating of the E and M steps?

1. Calculate expected values for latent variables

2. Use those expected values to maximize parameters

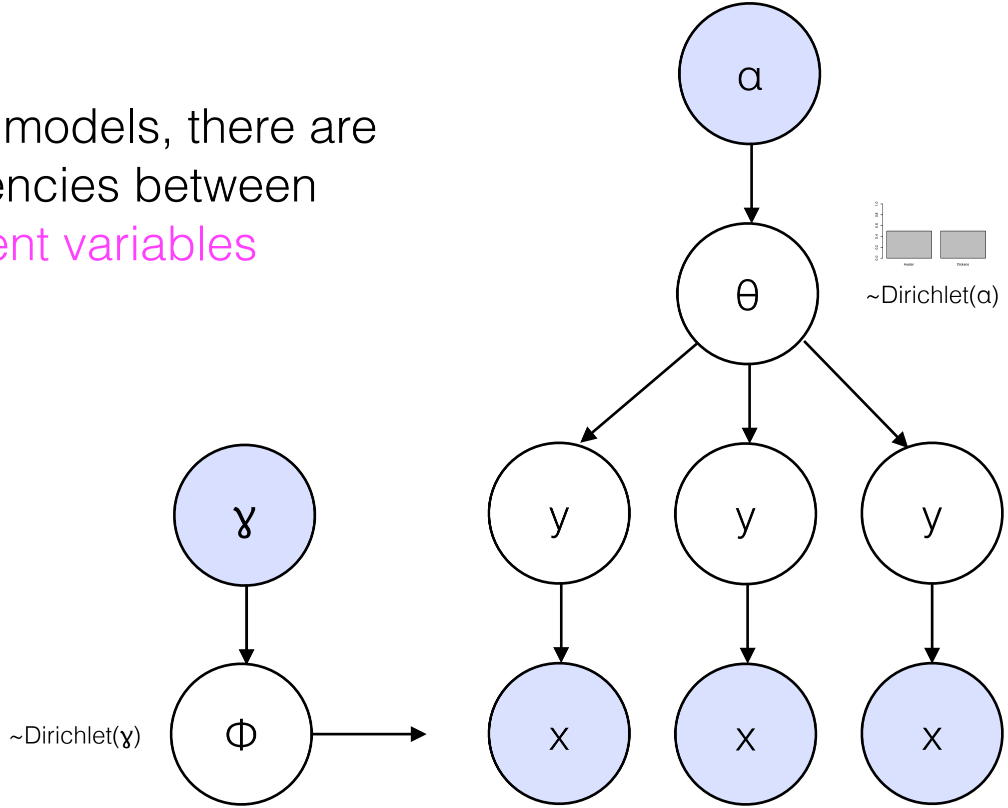
$$P(y \mid x, \theta, \phi) = \frac{P(y \mid \theta)P(x \mid y, \phi)}{\sum_{y' \in \mathcal{Y}} P(y' \mid \theta)P(x \mid y', \phi)}$$

what's this value for an observed label?

$$\theta_k = \frac{1}{N}\sum_{i=1}^{N} r_{i,k}$$

what's $r_{i,k}$ for a data point with observed label?

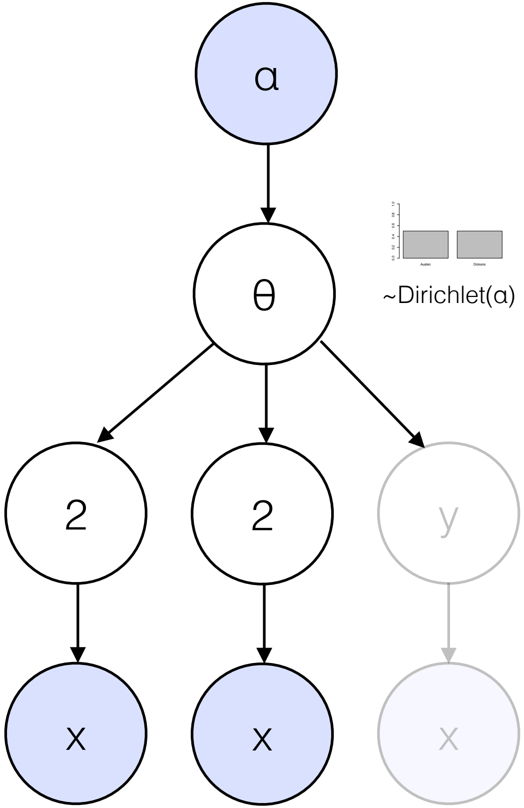
In more complex models, there are often dependencies between multiple latent variables



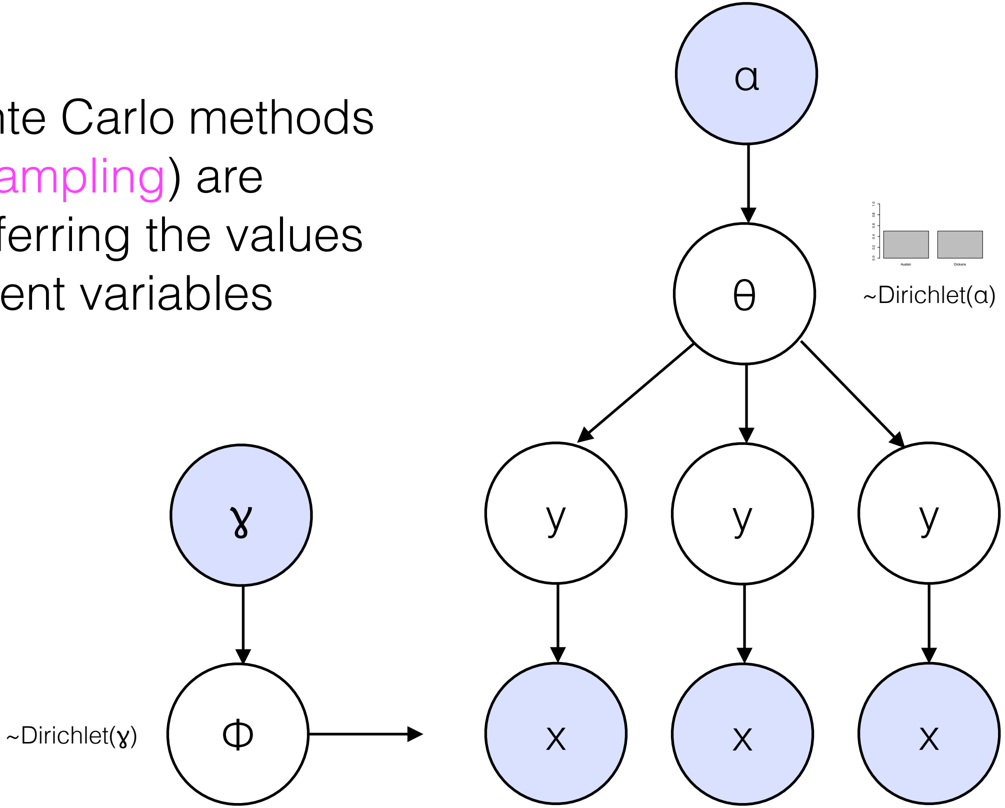
In more complex models, there are often dependencies between multiple latent variables

Here's an example: if you don't know the value of θ, but you believe y_1 and y_2 = 2, then your best estimate of θ will favor 2, making $P(y_3 = 2)$ high

the y's are dependent on each other

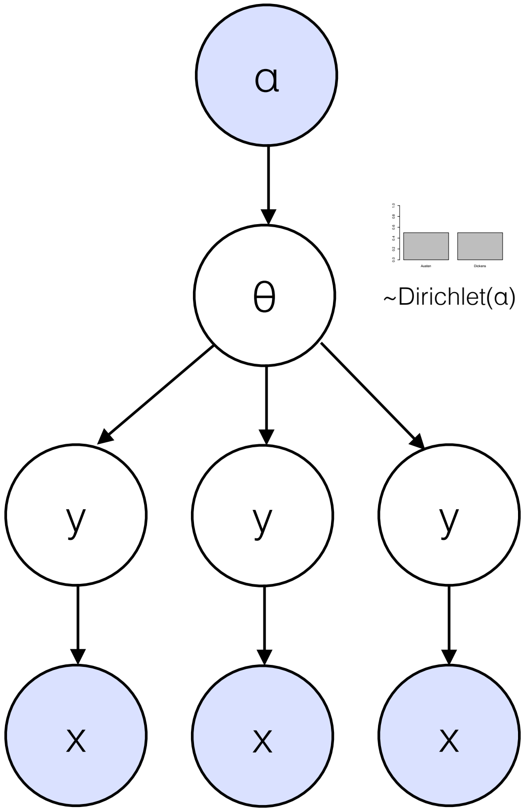


Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables



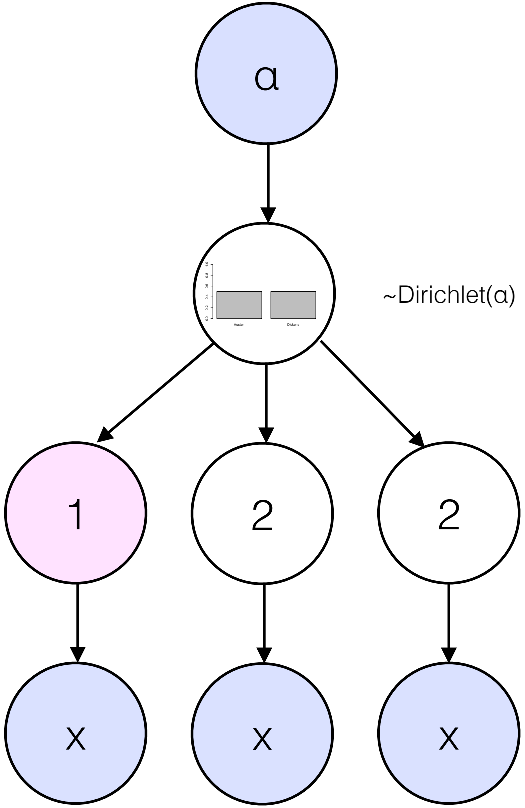
Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

The idea is very simple: start out with random guesses for all variables



Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

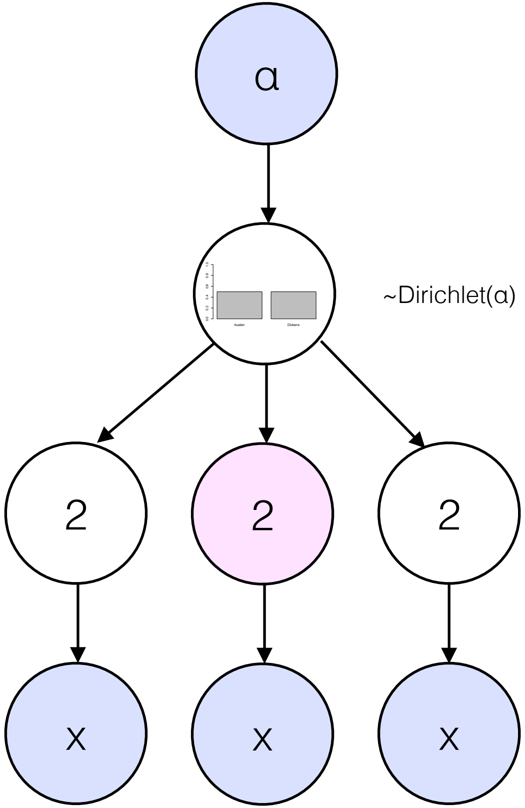
Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else

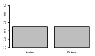


$P(y \mid \theta =$  $, x) \propto P(y \mid \theta =$  $)\, P(x \mid y)$

Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

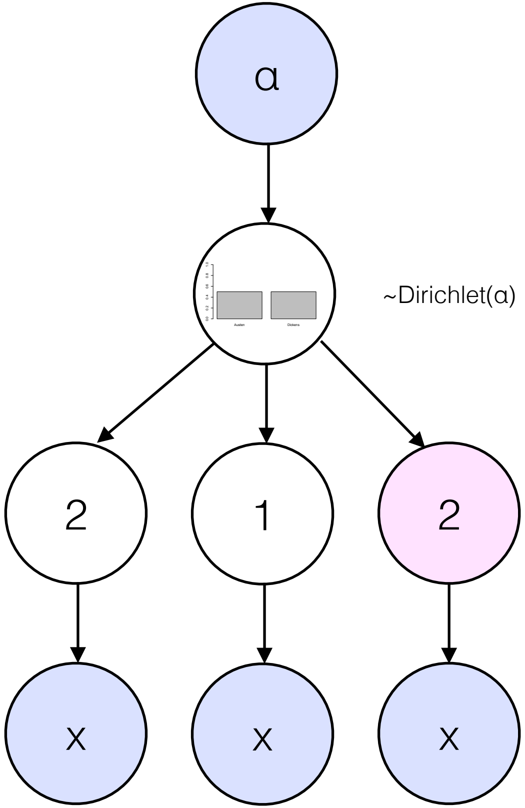
Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else



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

Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else

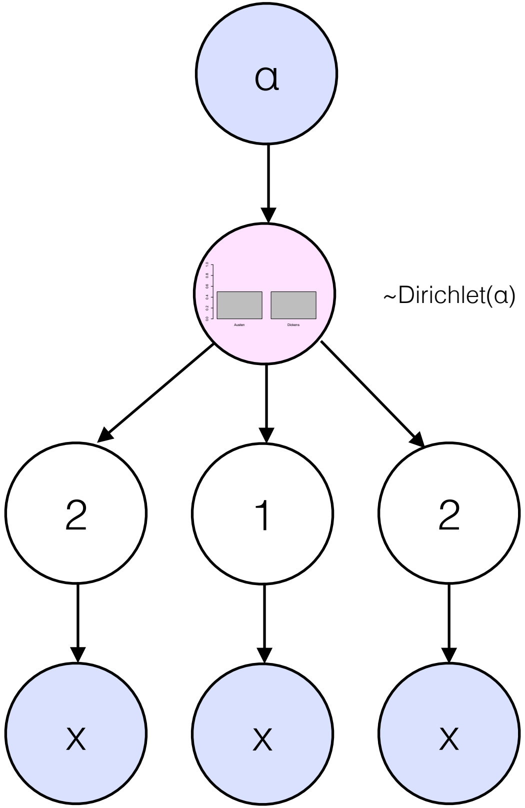


$P(y \mid \theta =$  $, x) \propto P(y \mid \theta =$  $)\, P(x \mid y)$

Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

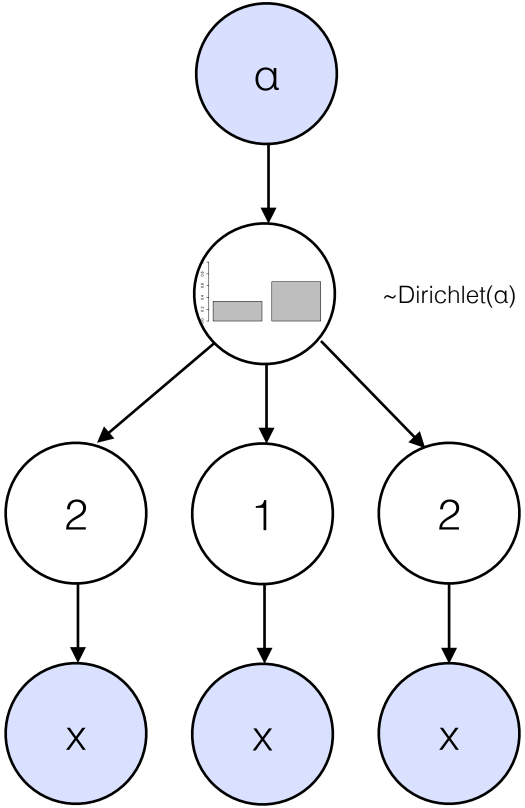
Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else

$$P(\theta \mid \alpha, y) \propto P(\theta \mid \alpha) \prod_{i=1}^{N} P(y_i \mid \theta)$$



Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

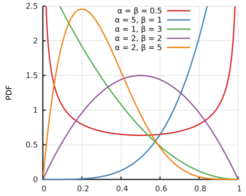
Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else

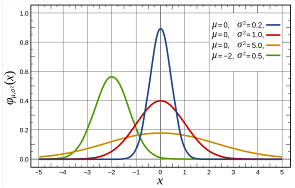


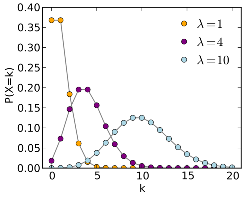
Graphical models

- Graphical models articulate the relationship between variables
- Lots of standard inference techniques are available; the art is in defining the structure of the model:
    - what the variables are
    - what parametric form they take
    - what's observed and what's latent
    - what the relationship is between the variables

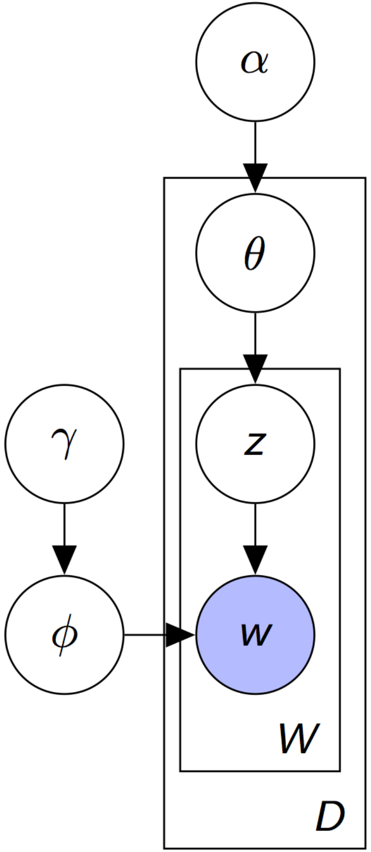
Beta	[0,1]	position in time bounded series	real
Bernoulli	0 or 1	presence of feature	binary
Normal	$(-\infty, \infty)$	age, height	real
Multinomial	count data	word counts	discrete
Poisson	$\{0, 1, 2, \ldots, \infty\}$	number of children	discrete







Topic models

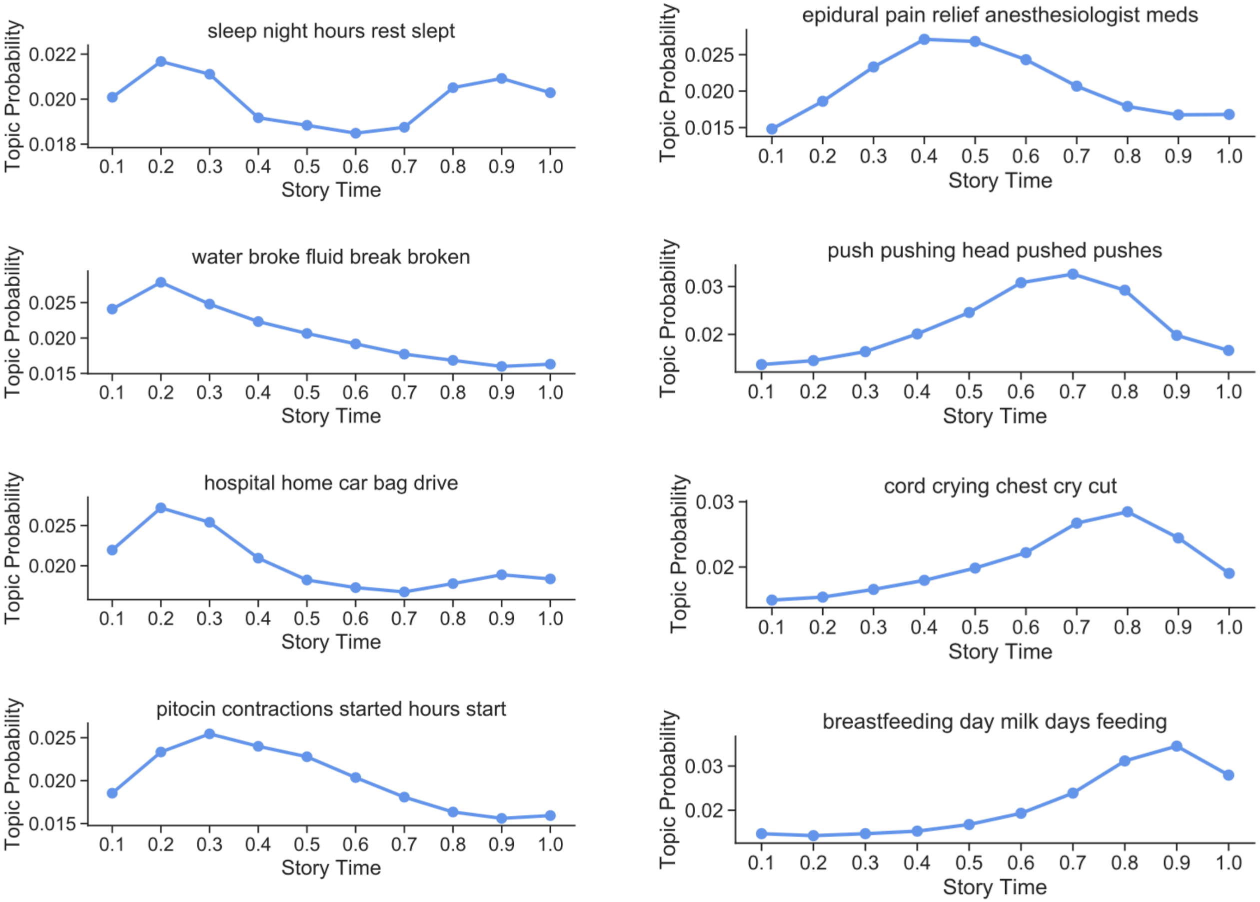


Clustering

- Clustering is designed to learn structure in the data:
    - Hierarchical structure between data points
    - Natural partitions between data points

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, clustering tokens into topics



Antoniak et al. 2019, “Narrative Paths and Negotiation of Power in Birth Stories”

The Ten Most Significant Topics in the *National Anti-Slavery Standard*

Topic	Label	PMI	Keywords
T49	places	1.54	ohio, philadelphia, mass, office, york, miller, penn, standard, thomas, free
T32	miscellaneous ads	1.48	table, york, duty, free, street, fair, ad, cotton, good, cent
T91	shopping	1.10	street, philadelphia, books, goods, hand, prices, store, cases, assortment, attention
T46	ads for dry goods	0.87	cents, corn, flour, wheat, american, advance, made, paper, white, sales
T16	abolition	0.87	slavery, anti, abolitionists, american, society, abolition, pro, slave, liberty, garrison
T7	organizing	0.52	friends, aid, fair, money, work, make, means, committee, time, funds
T2	time	0.38	time, made, found, left, place, day, return, received, immediately, told
T62	war and expansion	0.37	texas, mexico, war, states, united, annexation, california, mexican, government, country
T42	formal organizing	0.36	society, meeting, friends, held, annual, county, anti, present, members, meetings
T97	slavery	0.24	slave, slaves, slavery, free, master, negroes, states, property, slaveholders, emancipation

The Ten Most Significant Topics in the *National Anti-Slavery Standard* While Child Was Editor

Topic	Label	PMI	Keywords
T70	cooking	0.88	water, put, half, sugar, pound, cold, milk, salt, add, butter
T26	foreign relations	0.63	united, government, states, american, cuba, foreign, british, treaty, trade, president
T49	places	0.63	ohio, philadelphia, mass, office, york, miller, penn, standard, thomas, free
T40	correspondence	0.53	letter, office, post, letters, received, written, send, addressed, department, general
T42	formal organizing	0.49	society, meeting, friends, held, annual, county, anti, present, members, meetings
T14	Massachusetts	0.45	boston, mass, rev, john, wm, george, salem, charles, samuel, esq
T25	travel and accidents	0.44	fire, railroad, city, train, boston, cars, company, york, road, accident
T35	federal government	0.40	house, congress, district, petition, representatives, adams, legislature, petitions, people
T9	violence and crime	0.39	house, man, shot, negro, murder, mob, night, city, arrested, men
T5	state government	0.38	state, law, laws, act, states, citizens, person, persons, united, legislature

Klein 2020, “Dimensions of Scale: Invisible Labor, Editorial Work, and the Future of Quantitative Literary Studies”

A Topic Model of Literary Studies Journals

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%

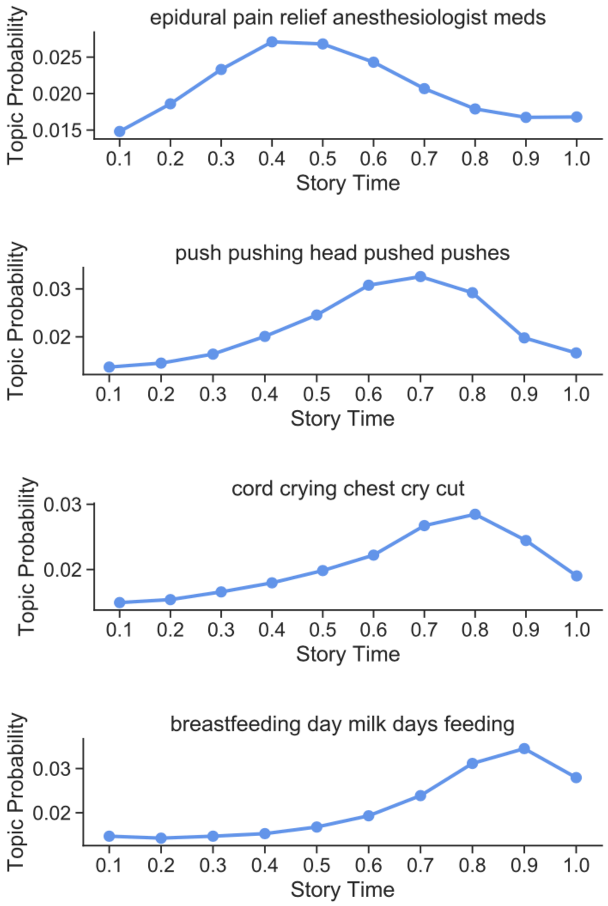
Goldstone and Underwood (2014),
The Quiet Transformations of Literary Studies

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

Applications



Antoniak et al. 2019, “Narrative Paths and Negotiation of Power in Birth Stories”

x = feature vector

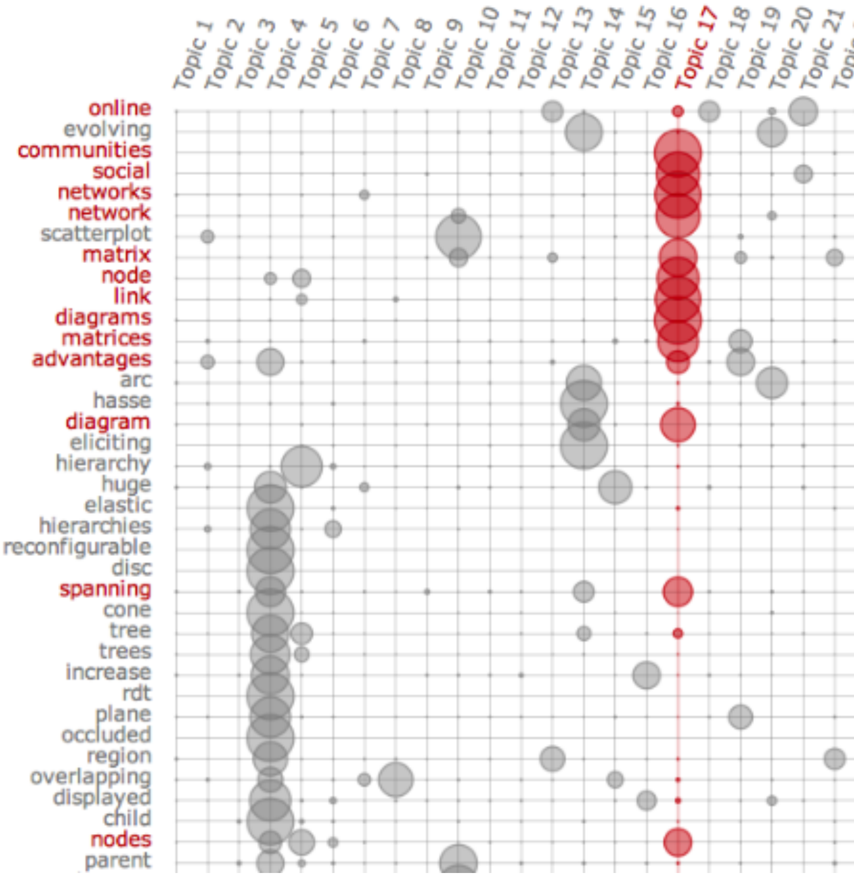
Feature	Value
contains “love”	0
contains “castle”	0
contains “dagger”	0
contains “run”	0
contains “the”	1
topic 1	0.55
topic 2	0.32
topic 3	0.13

β = coefficients

Feature	β
contains “love”	-3.1
contains “castle”	6.8
contains “dagger”	7.9
contains “run”	-3.0
contains “the”	-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



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."

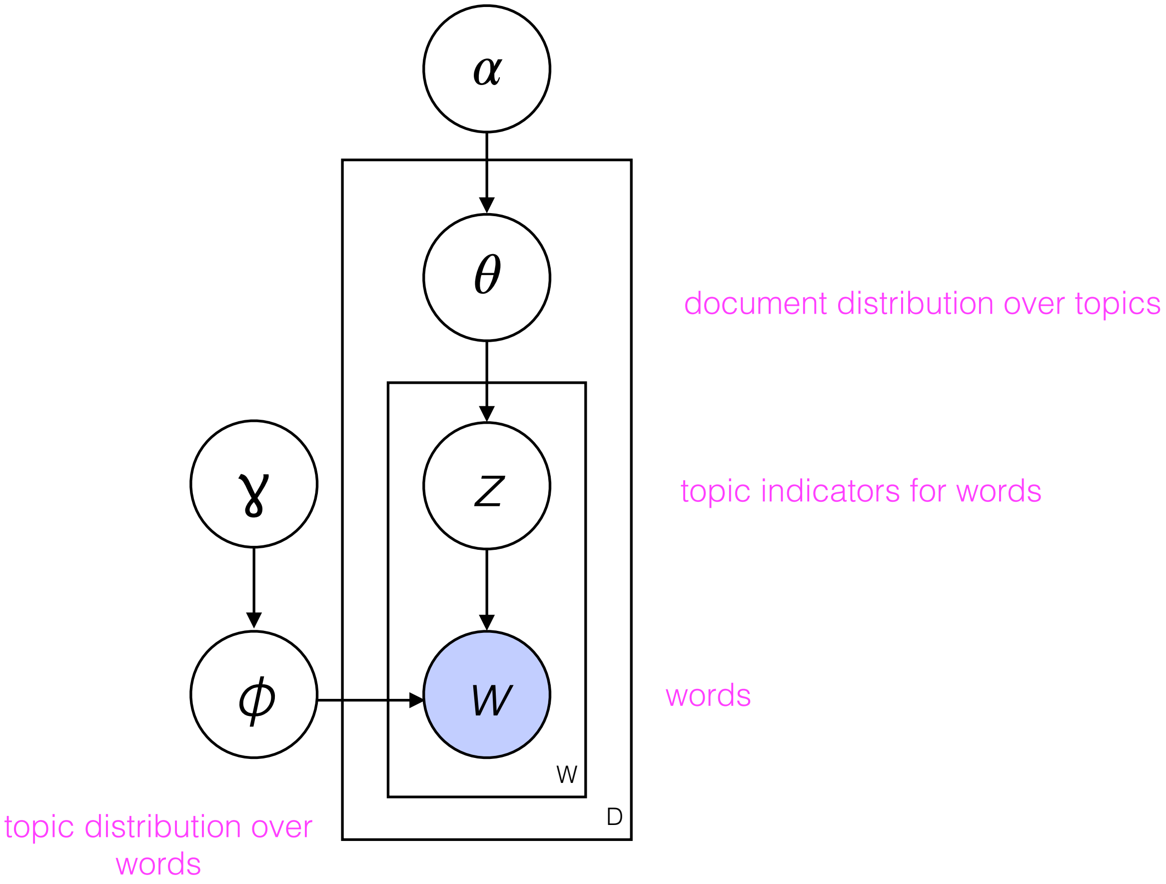
“Etc.”

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."

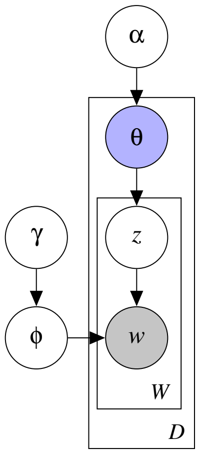
“People”

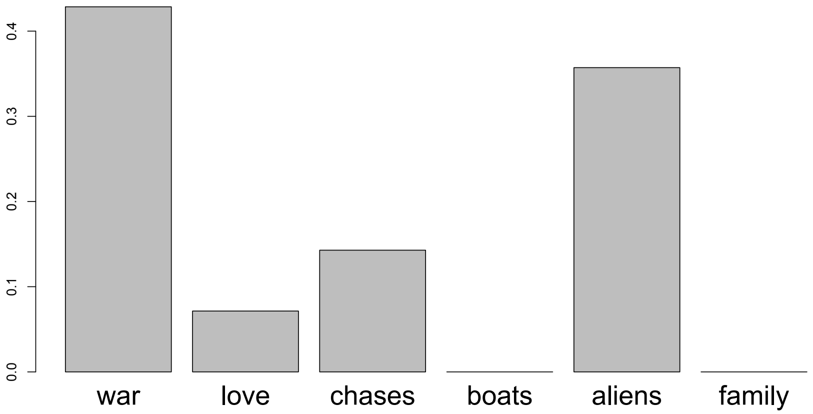
A different *Paris* token might belong to a “Place” or “French” topic


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

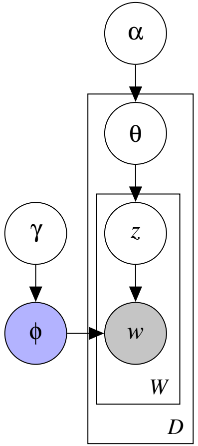
Topic Models

- A document has *distribution over topics*

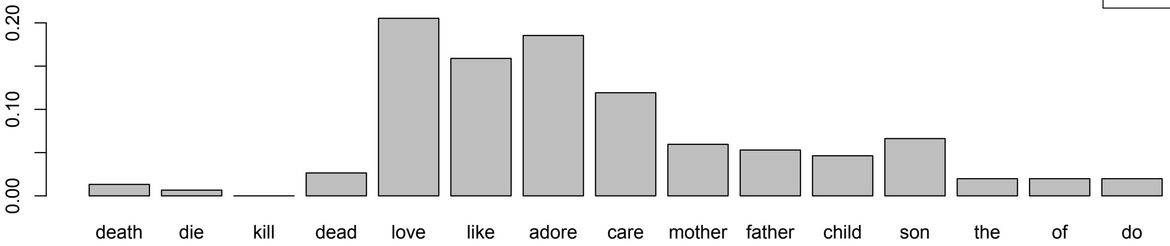




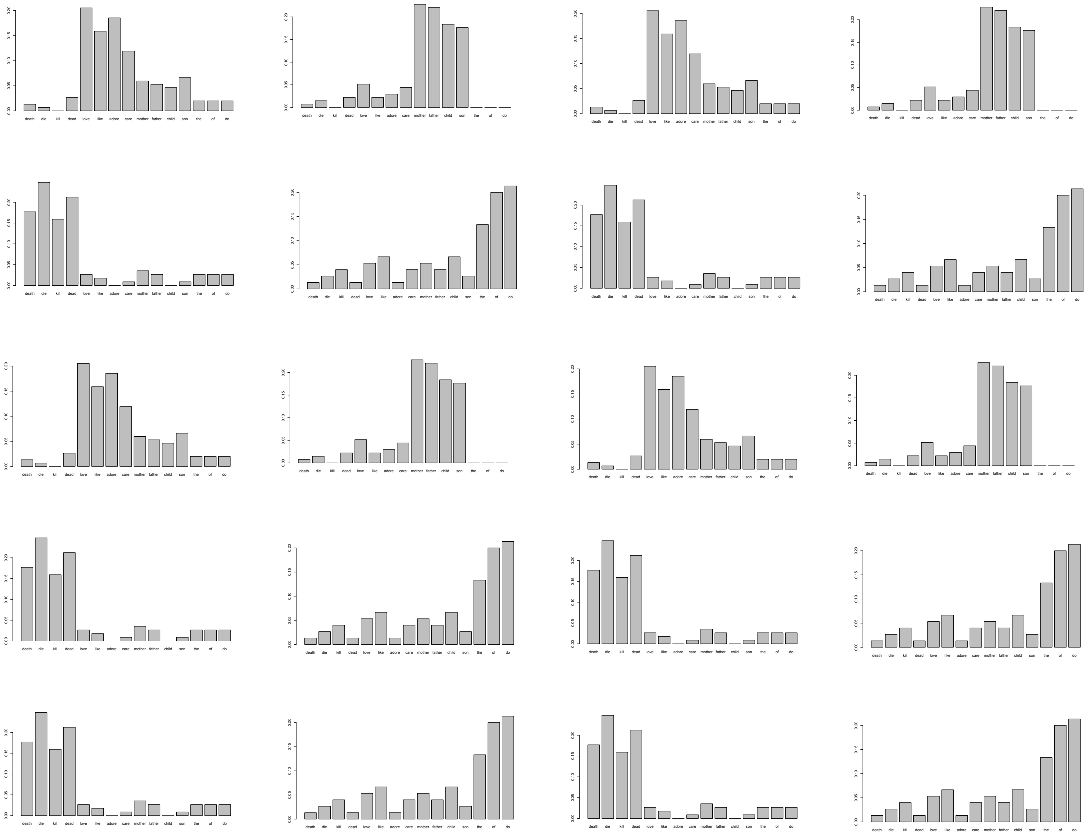
Topic Models

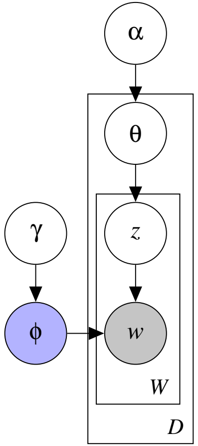


- A topic is a distribution over words

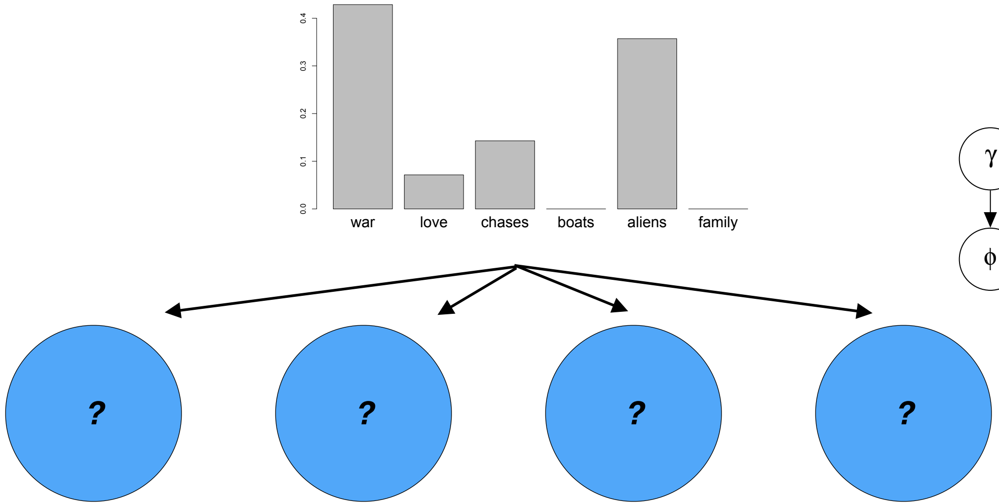


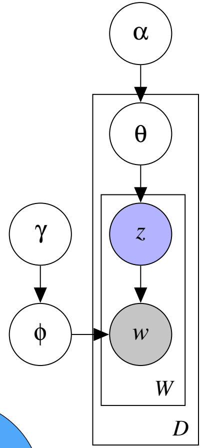
- e.g., P("adore" | topic = love) = .18


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

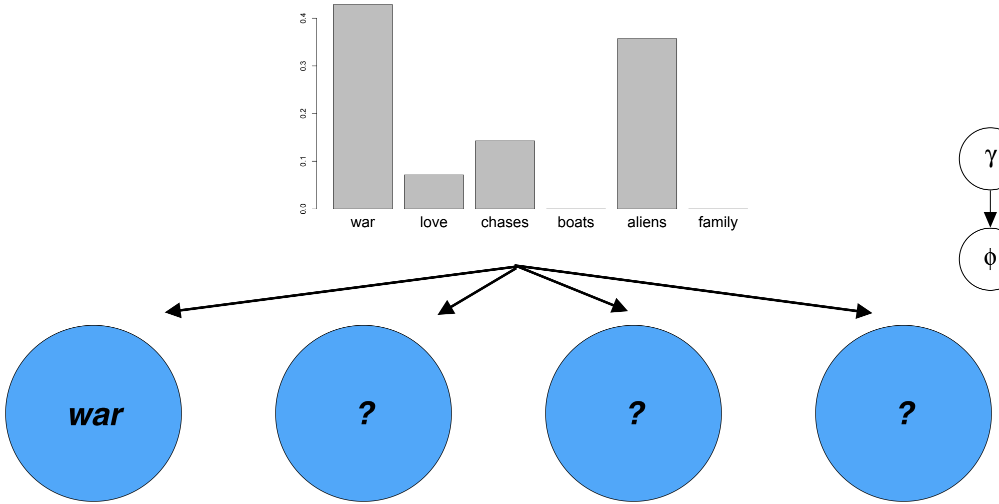

α
θ
γ
z
ϕ
w
W
D

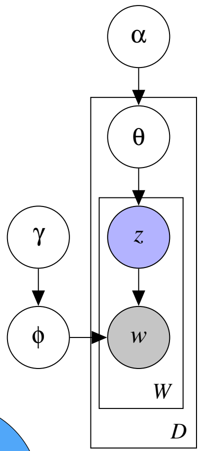
K=20



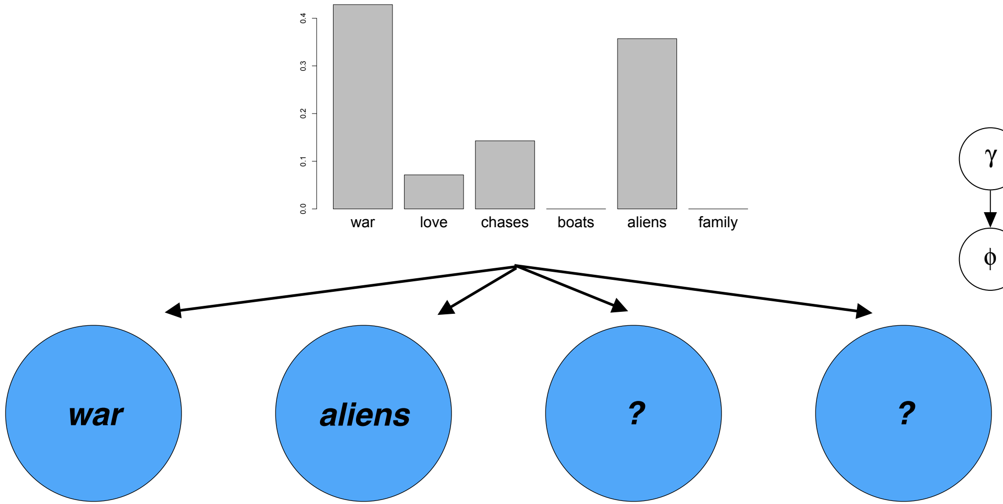


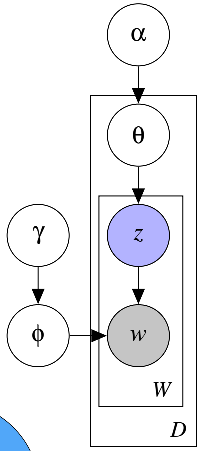
P(topic | topic distribution)



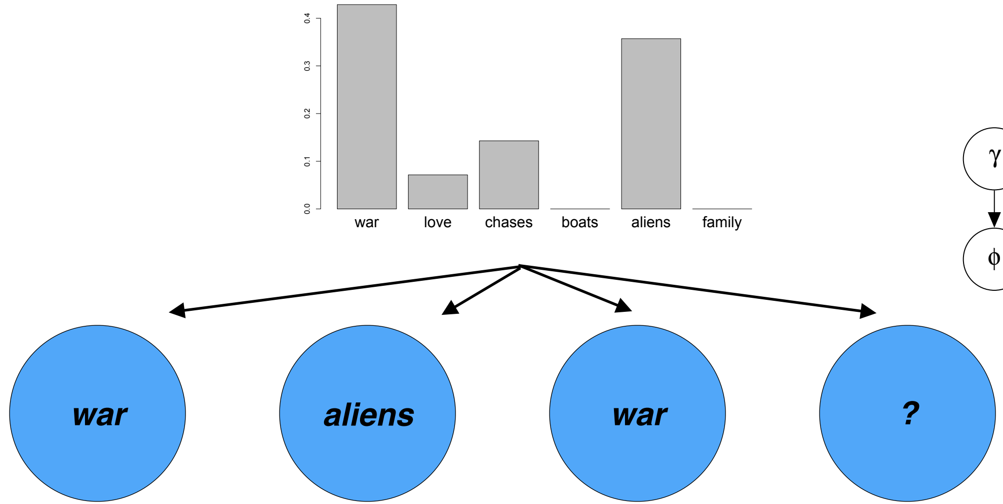


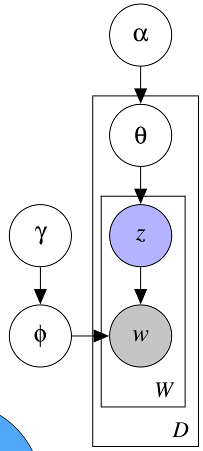
P(topic | topic distribution)



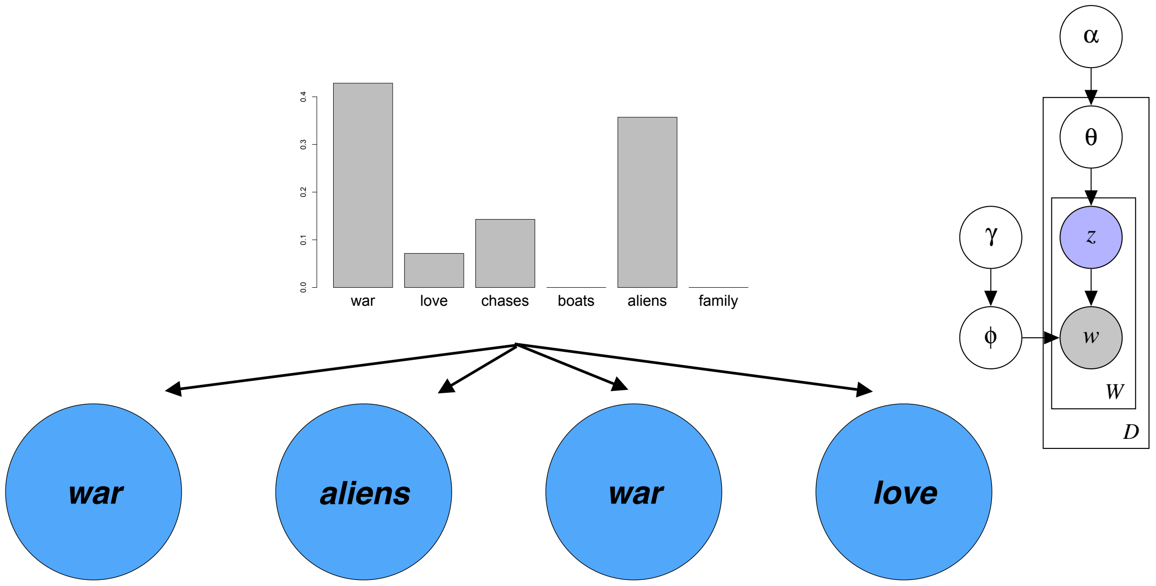


P(topic | topic distribution)

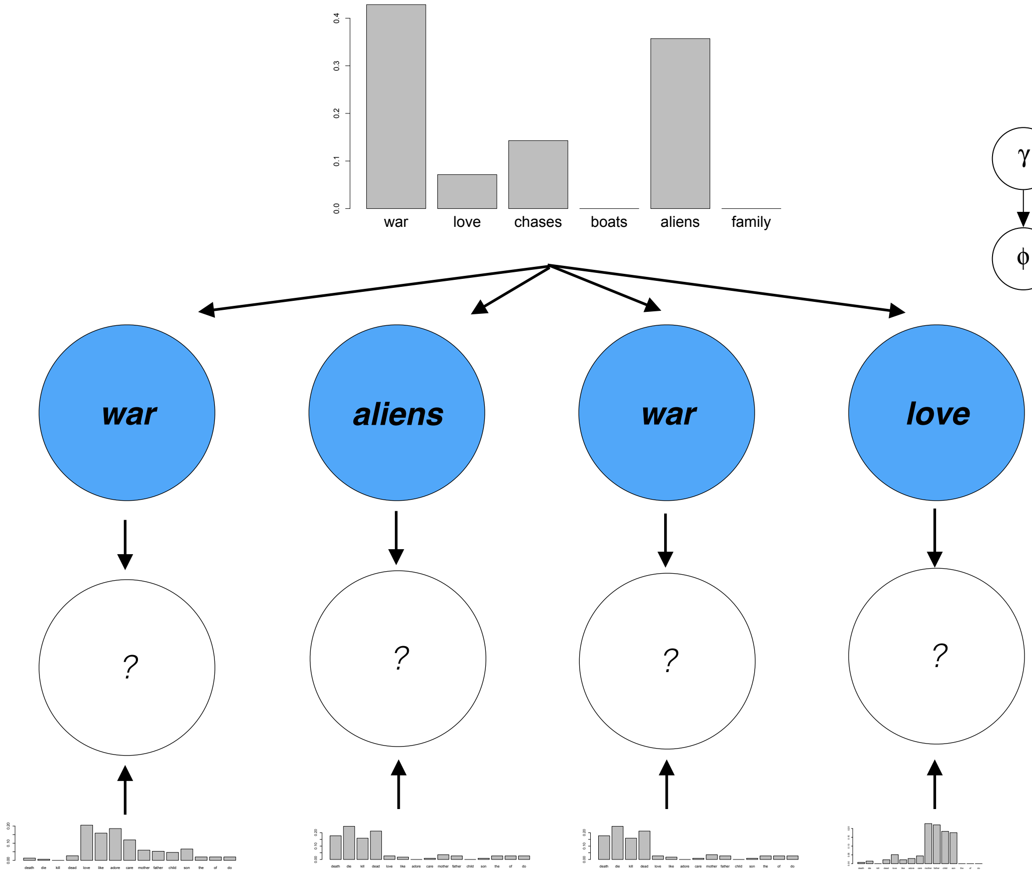


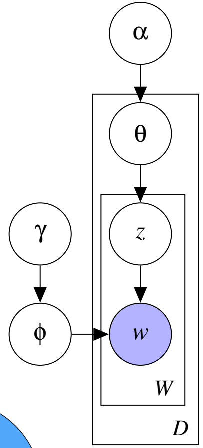


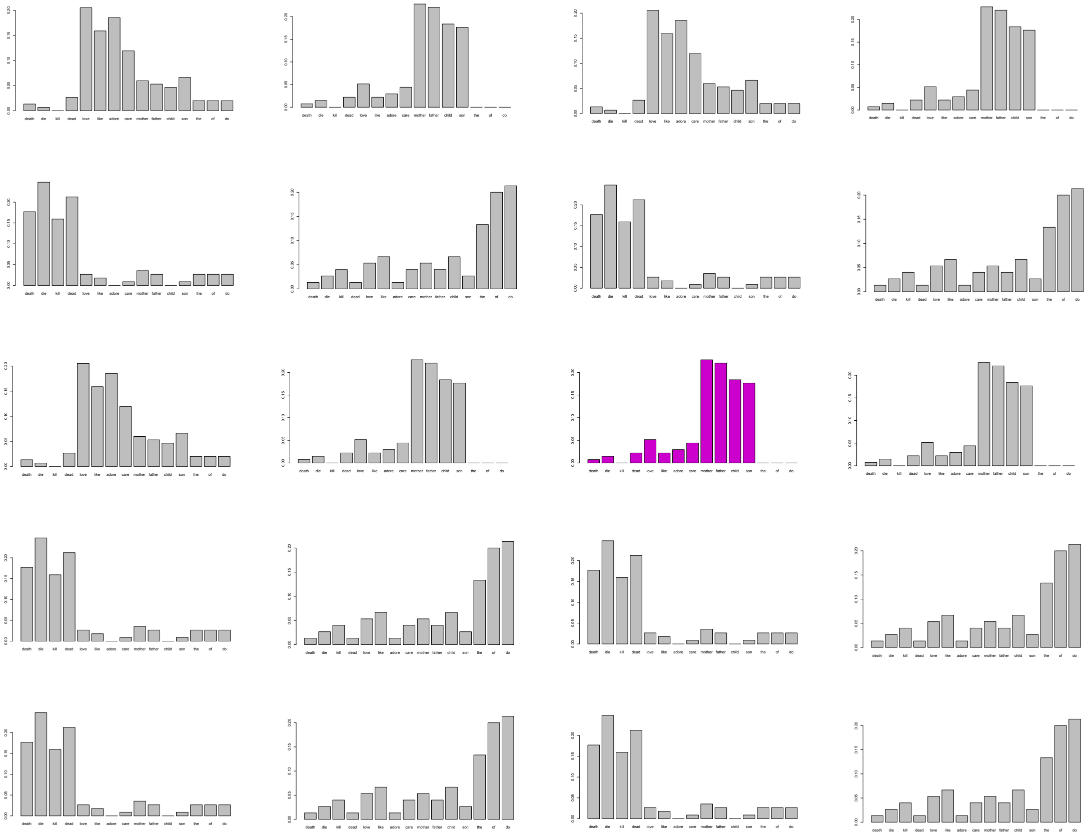
P(topic | topic distribution)

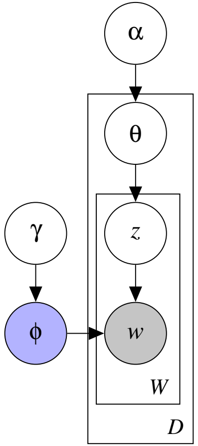


P(topic | topic distribution)

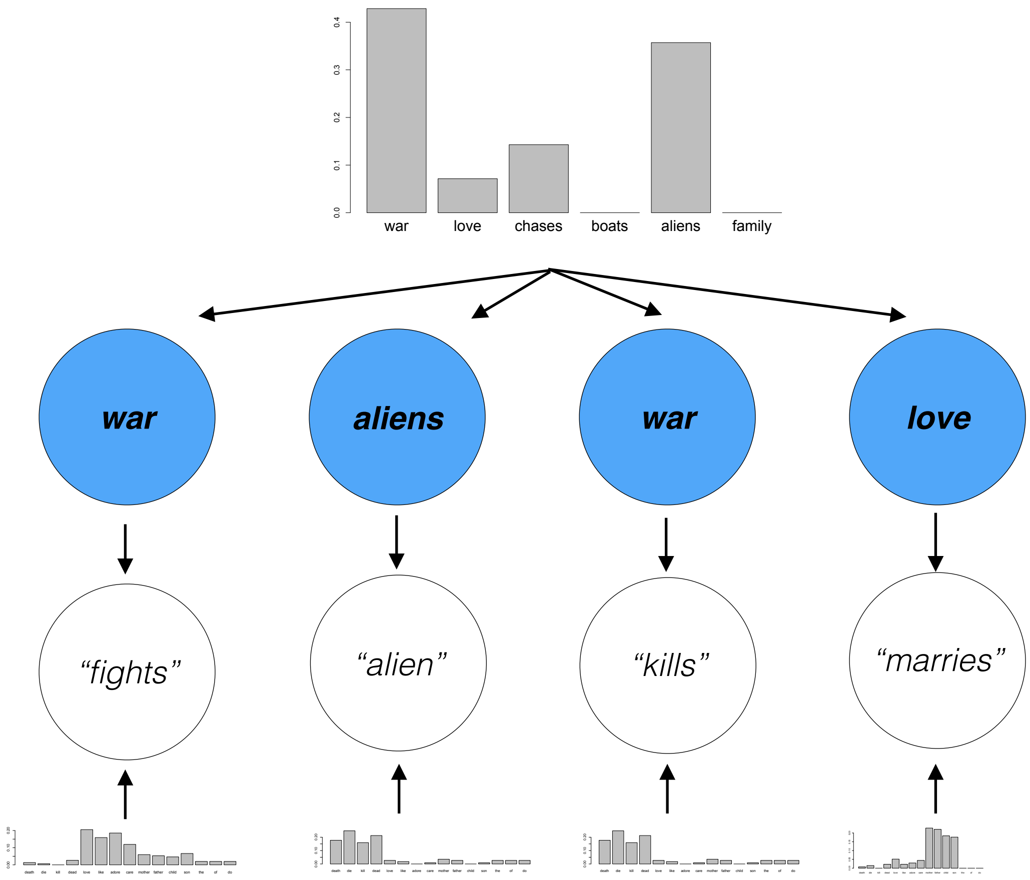

0.4
0.3
0.2
0.1
0.0
war
love
chases
boats
aliens
family
war
aliens
war
love
?
?
?
?

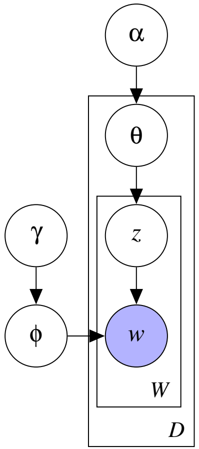

α
θ
γ
z
ϕ
w
W
D

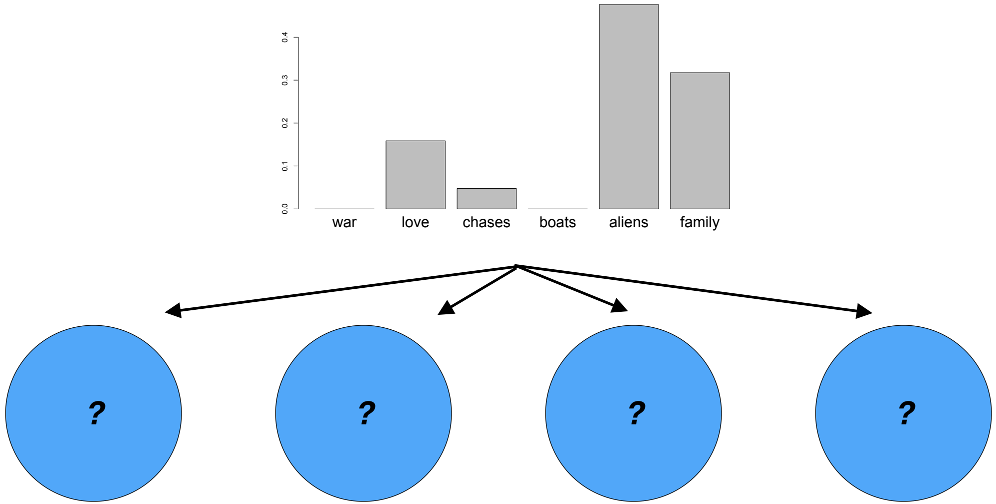


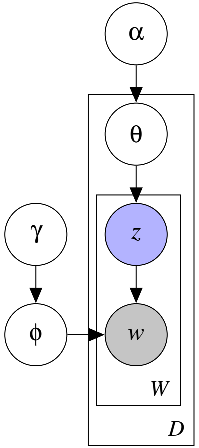


K=20

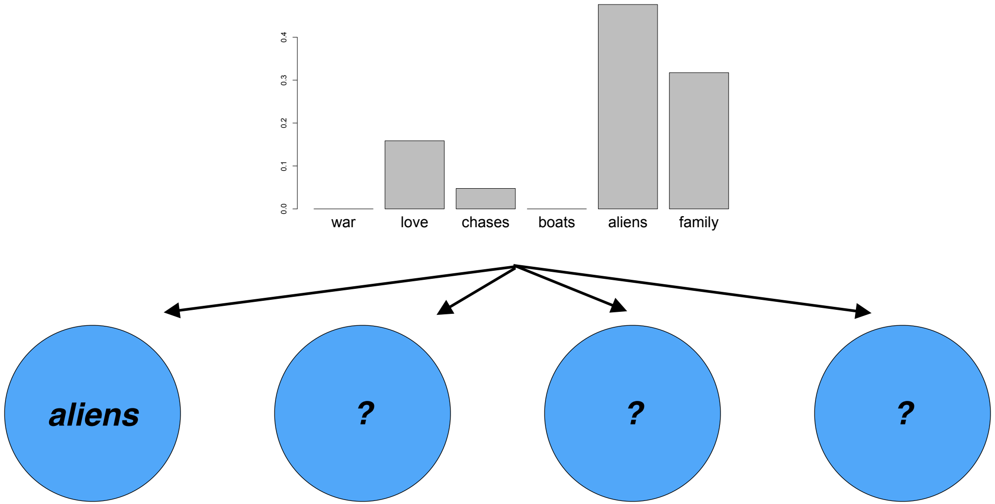

0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
war
aliens
war
love
“fights”
“alien”
“kills”
“marries”

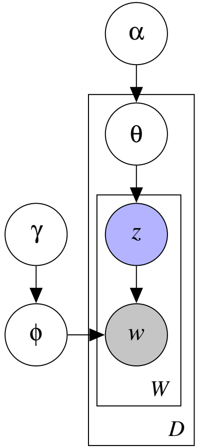

α
θ
γ
z
ϕ
w
W
D



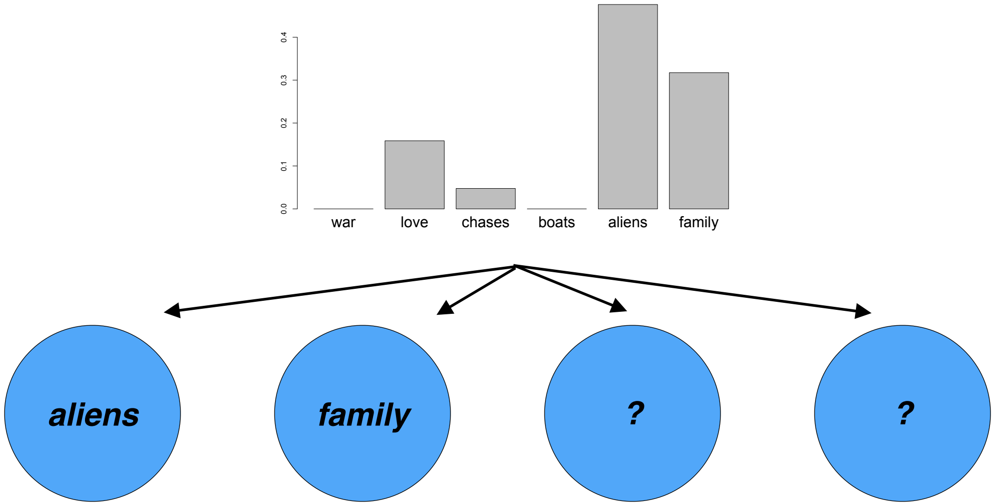


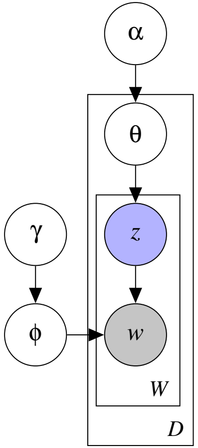
P(topic | topic distribution)



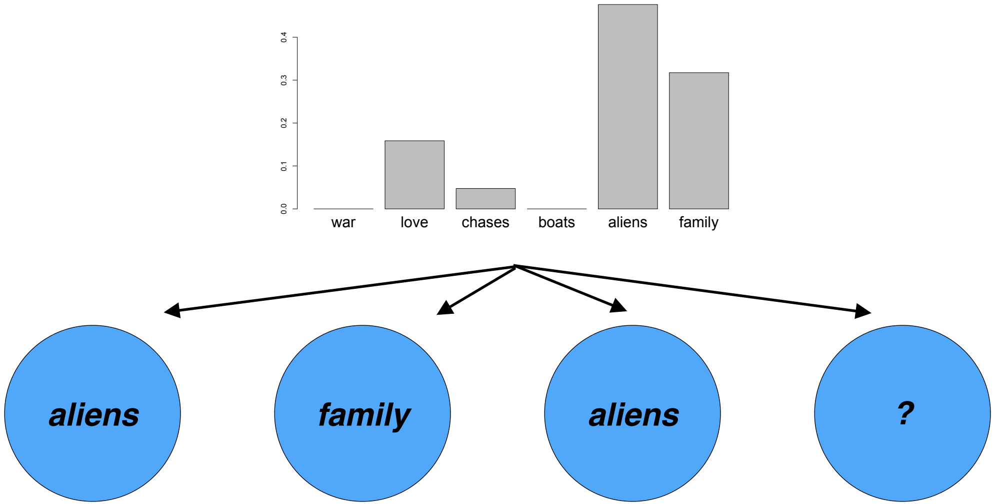


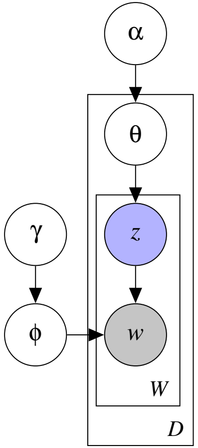
P(topic | topic distribution)



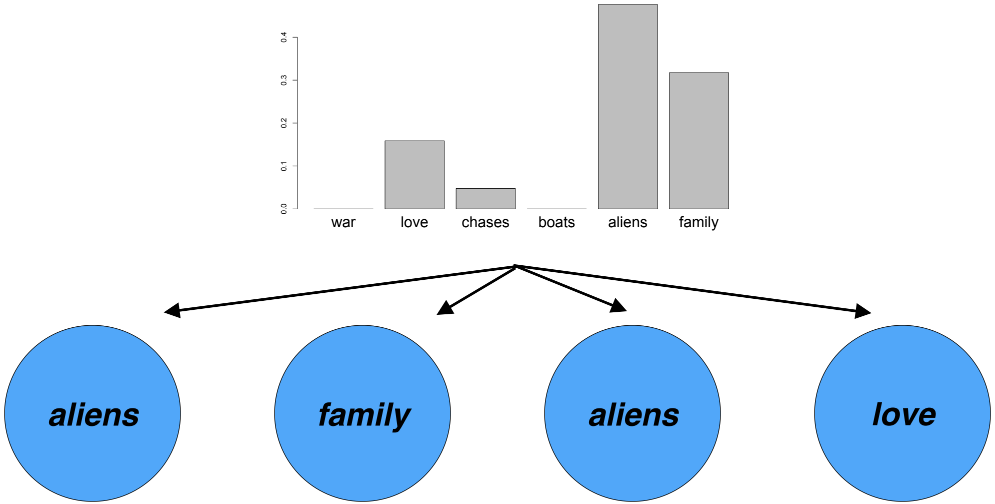


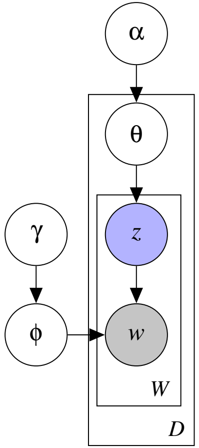
P(topic | topic distribution)



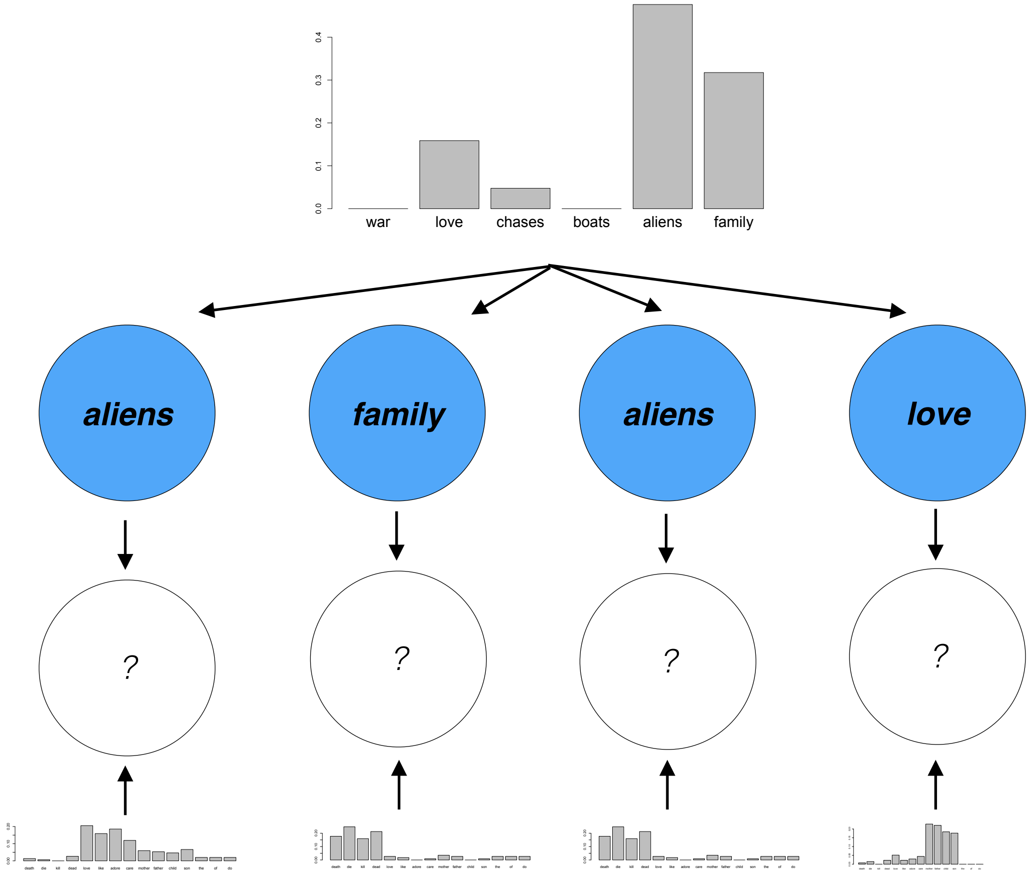


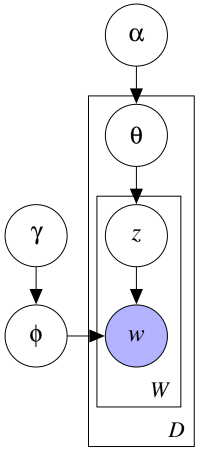
P(topic | topic distribution)

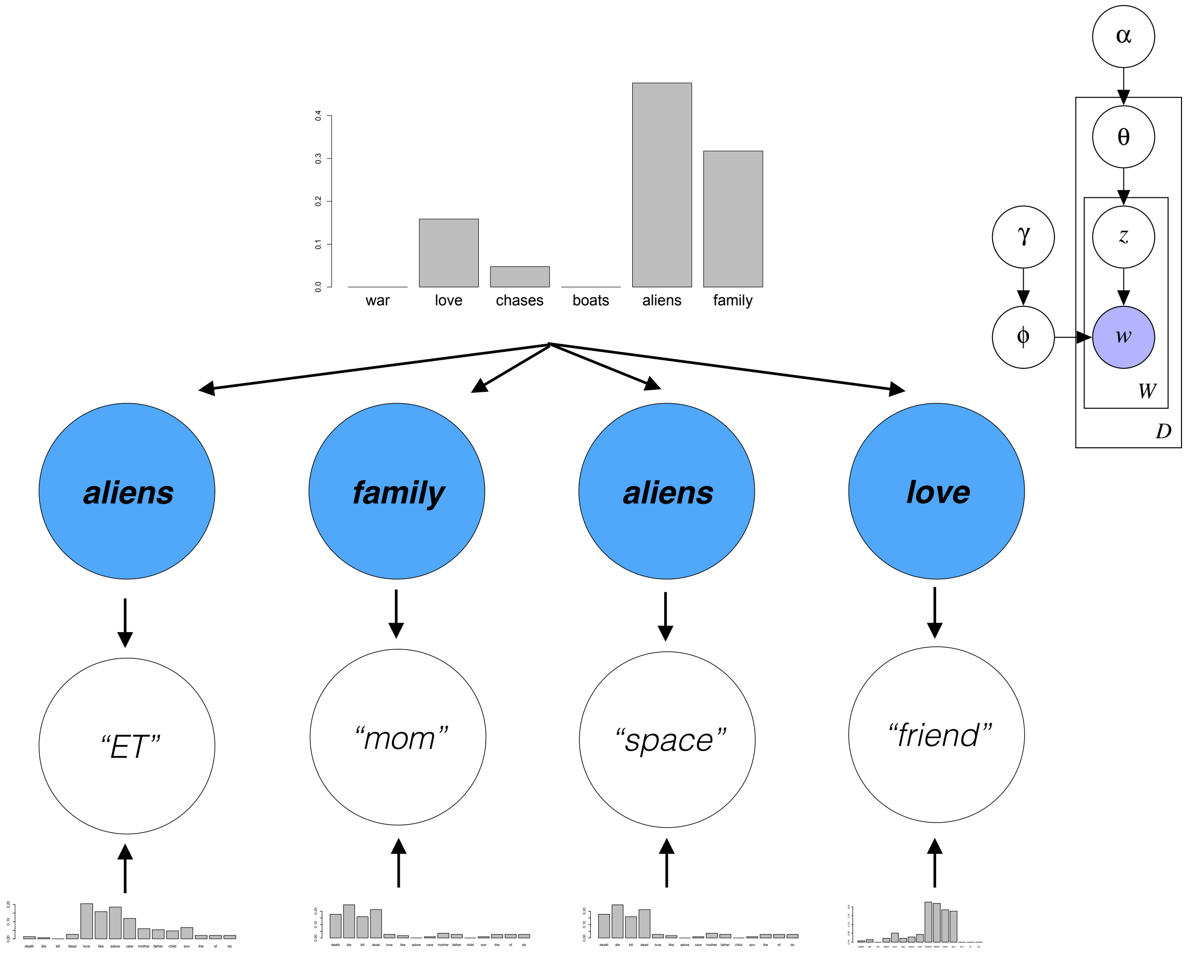




P(topic | topic distribution)

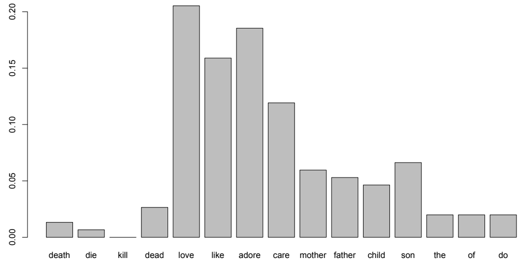

0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
aliens
family
aliens
love
?
?
?
?

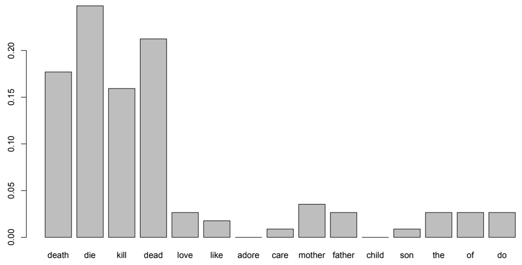

α
θ
z
γ
ϕ
w
W
D


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"

Inferred Topics

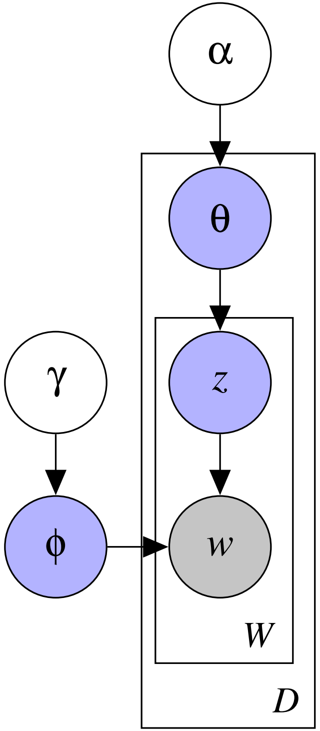
{album, band, music}	{government, party, election}	{game, team, player}
album	government	game
band	party	team
music	election	player
song	state	win
release	political	play
{god, call, give}	**{company, market, business}**	**{math, number, function}**
god	company	math
call	market	number
give	business	function
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



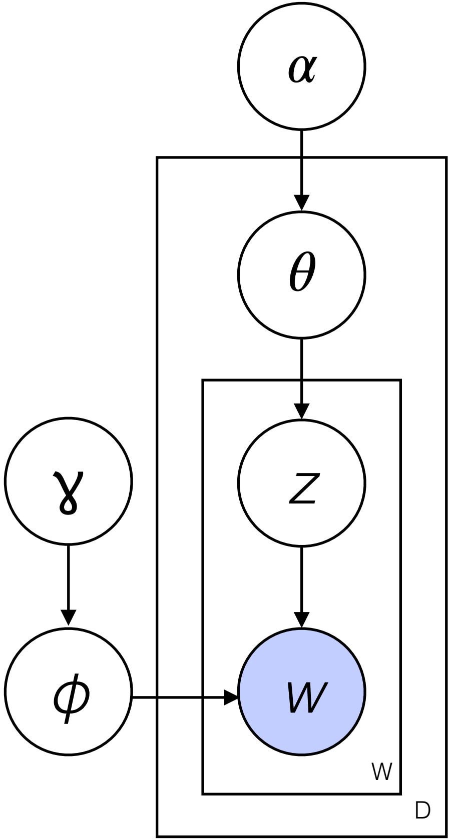


Inference

- 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?



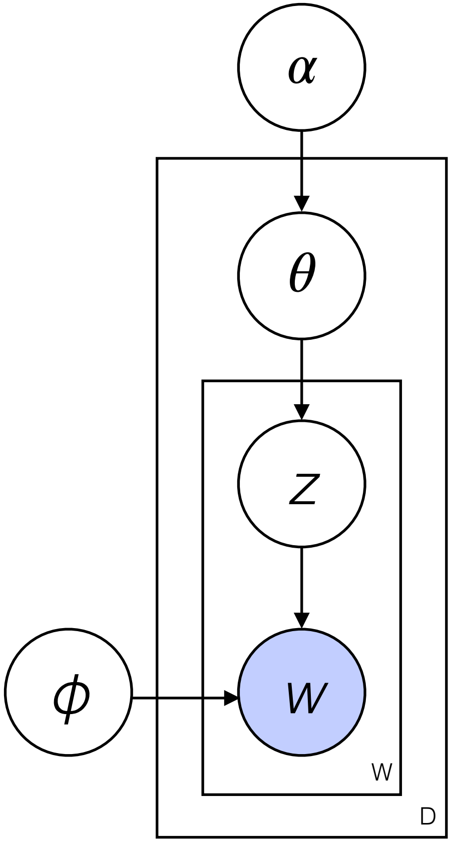
Find the parameters that maximize the likelihood of the data!

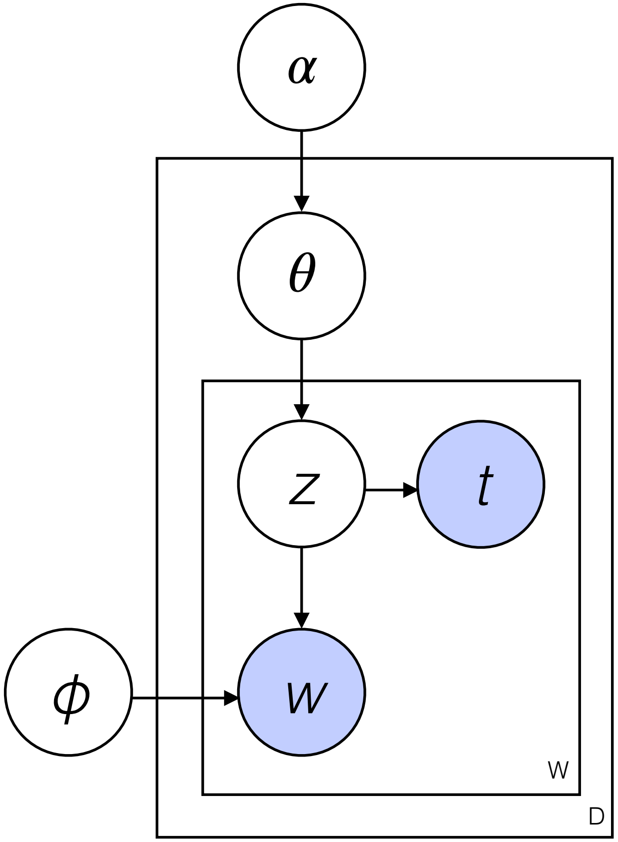


Assumptions

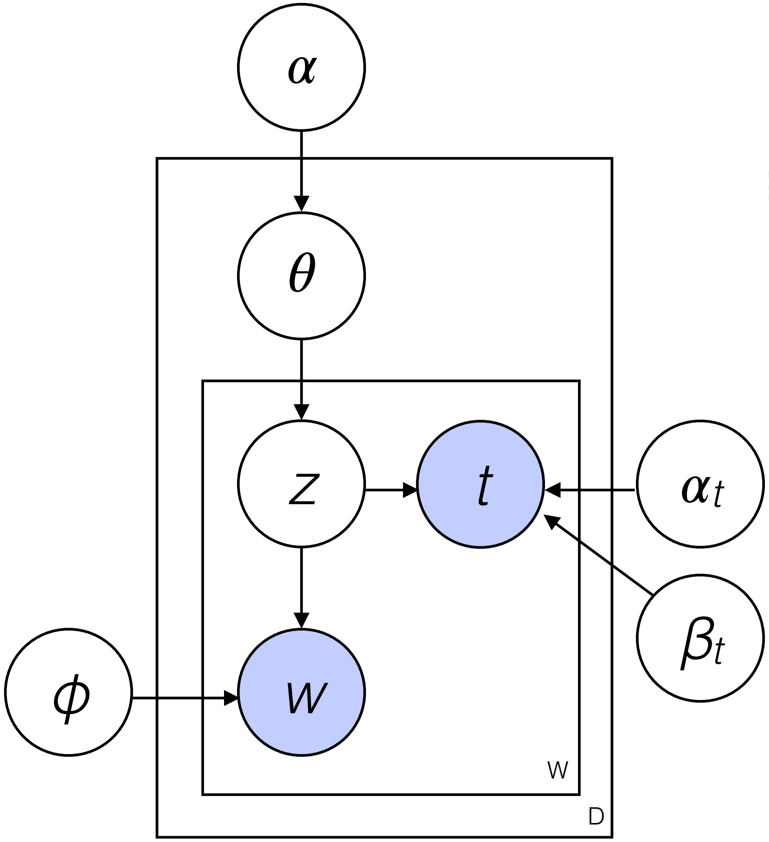
- Every word has one topic
- Every document has one topic distribution
- No sequential information (topics for words are independent of each other given the set of topics for a document)
- Topics don't have arbitrary correlations (Dirichlet prior)
- Words don't have arbitrary correlations (Dirichlet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

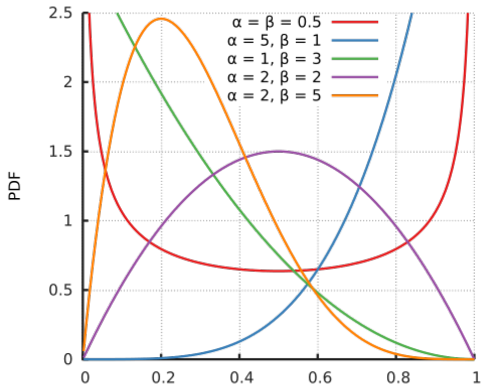
What if you want to encode other assumptions or reason over other observations?


α
θ
z
w
ϕ
W
D



α
θ
z
t
ϕ
w
W
D

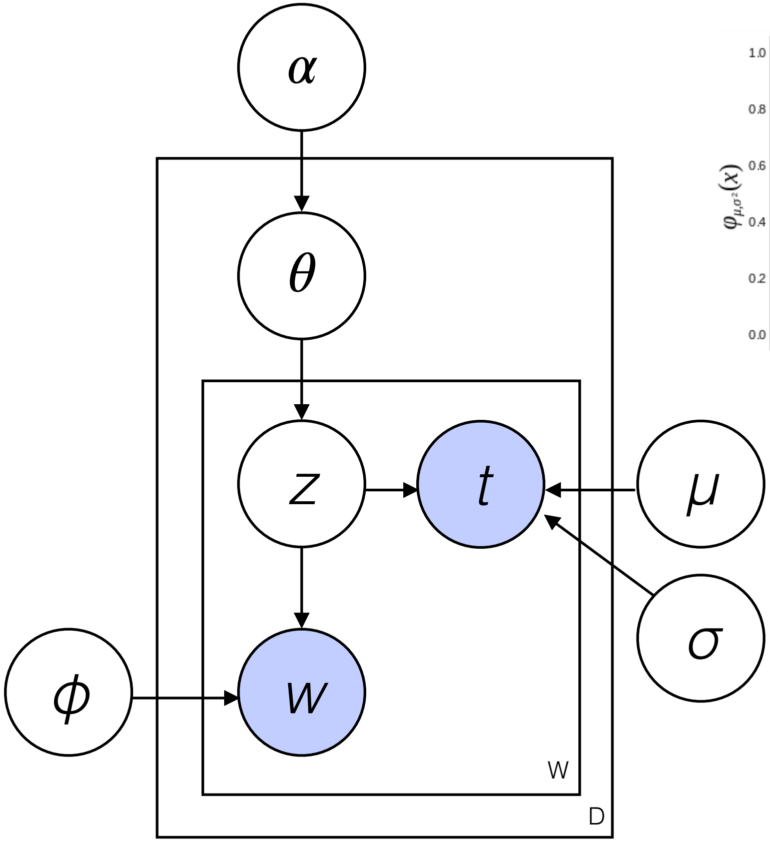


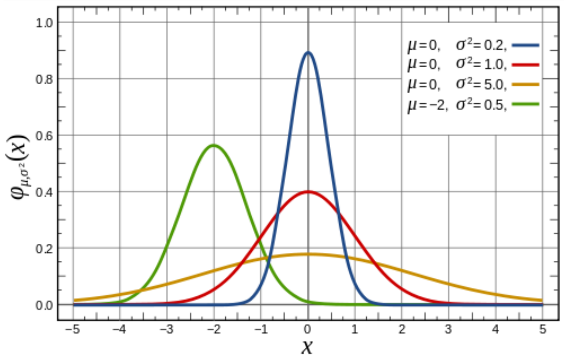


Time is drawn from a Beta distribution

[0,1]

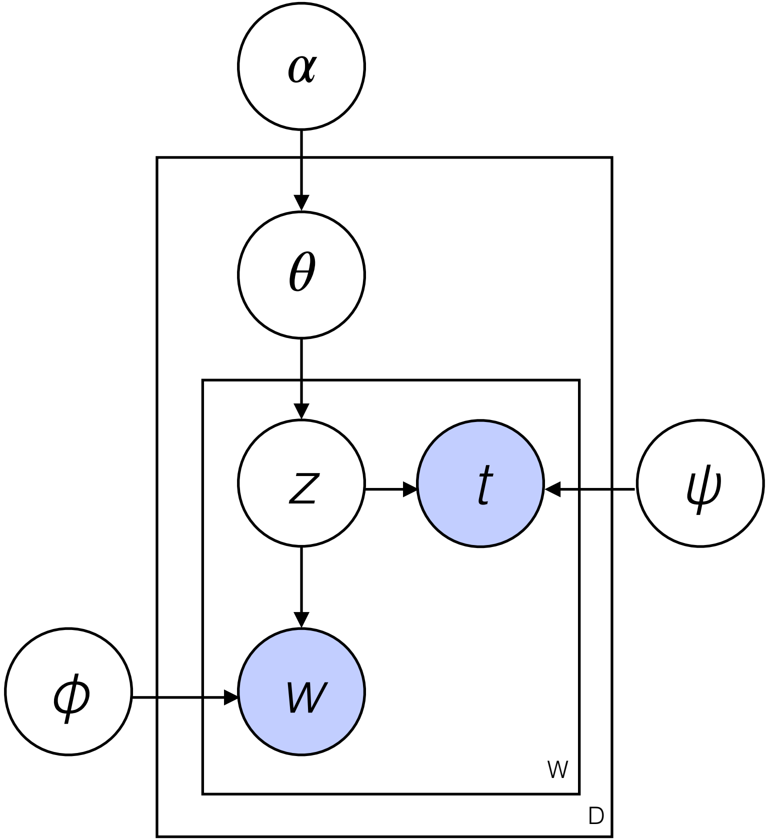
(Wang and McCallum 2006)

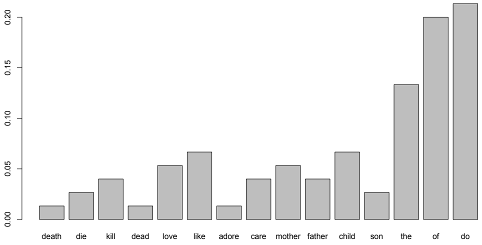




Time is drawn from a Normal distribution

[-∞, ∞]





Time is drawn from a Multinomial distribution

[1, … , K]

Deep Latent Variable Models

- Making models "deep" involves changing the parameterization of the underlying probability distribution

- Multinomial ~ Dirichlet → FFNN, RNNLM, Transformer

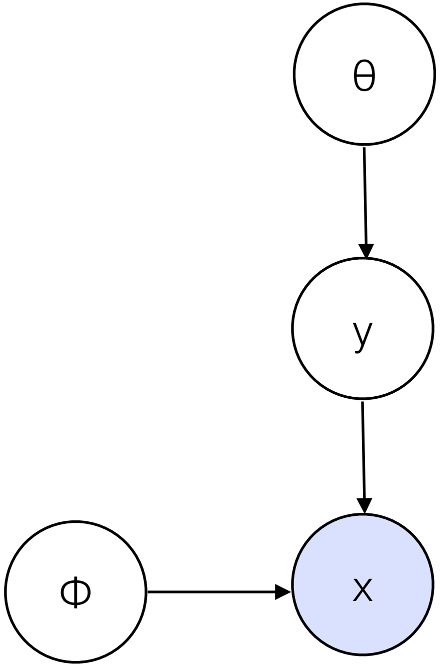
- Allows for more flexible parameterization, alternative independence assumptions, and richer models of context.

Deep Latent Variable Models

Unsupervised "Naive" Bayes (each x_i is independent from the others)

$$P(x_1, \ldots, x_N \mid y; \phi) = \prod_{i}^{N} P(x_i \mid y)$$

$$= \prod_{i}^{N} \phi_{x_i}^{y}$$

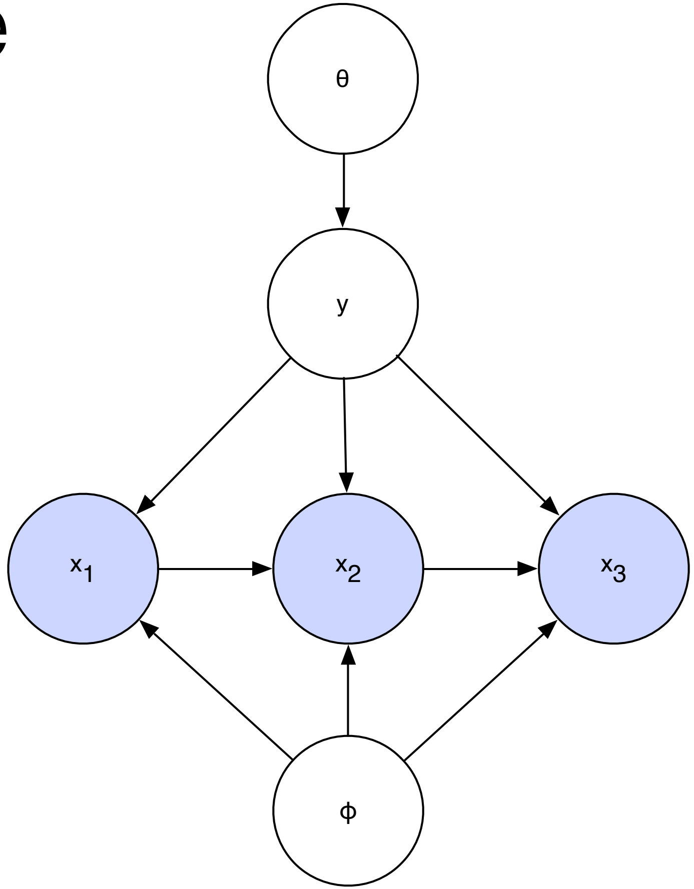


Deep Latent Variable Models

Deep version (each x_i depends on the other tokens)

$$P(x_i \mid y, \phi) = \text{RNNLM}\left(x_{1:i}; \phi^y\right)$$

ϕ^y here is a vector that gets passed into each RNN time step (e.g., concatenated with x)



Latent variable models

- See Kim et al. 2019 for more!