# Applied Natural Language Processing

Info 256

Lecture 6: Topic models (Sept. 13, 2023)

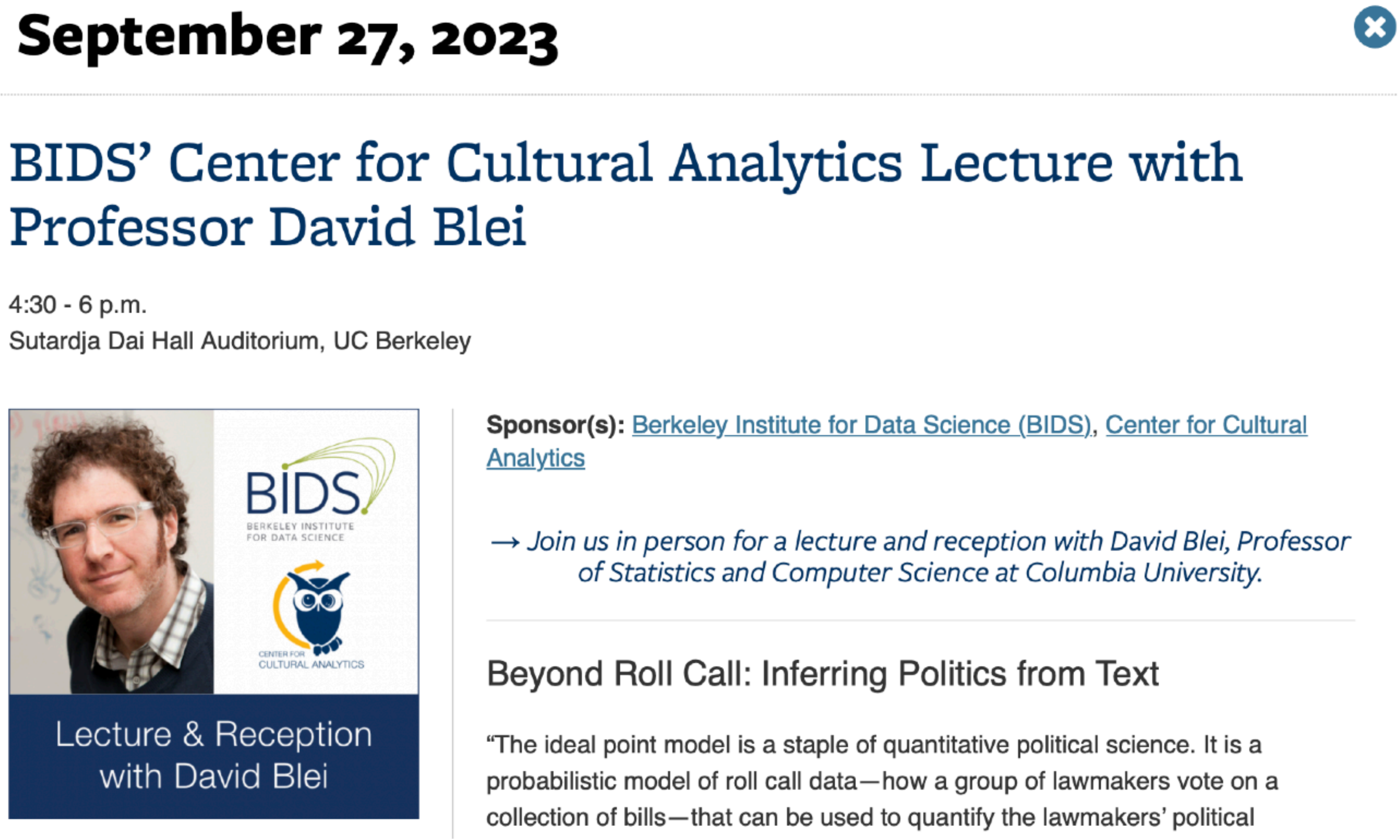
David Bamman, UC Berkeley

# 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



# September 27, 2023

## BIDS' Center for Cultural Analytics Lecture with Professor David Blei

4:30 - 6 p.m.
Sutardja Dai Hall Auditorium, UC Berkeley

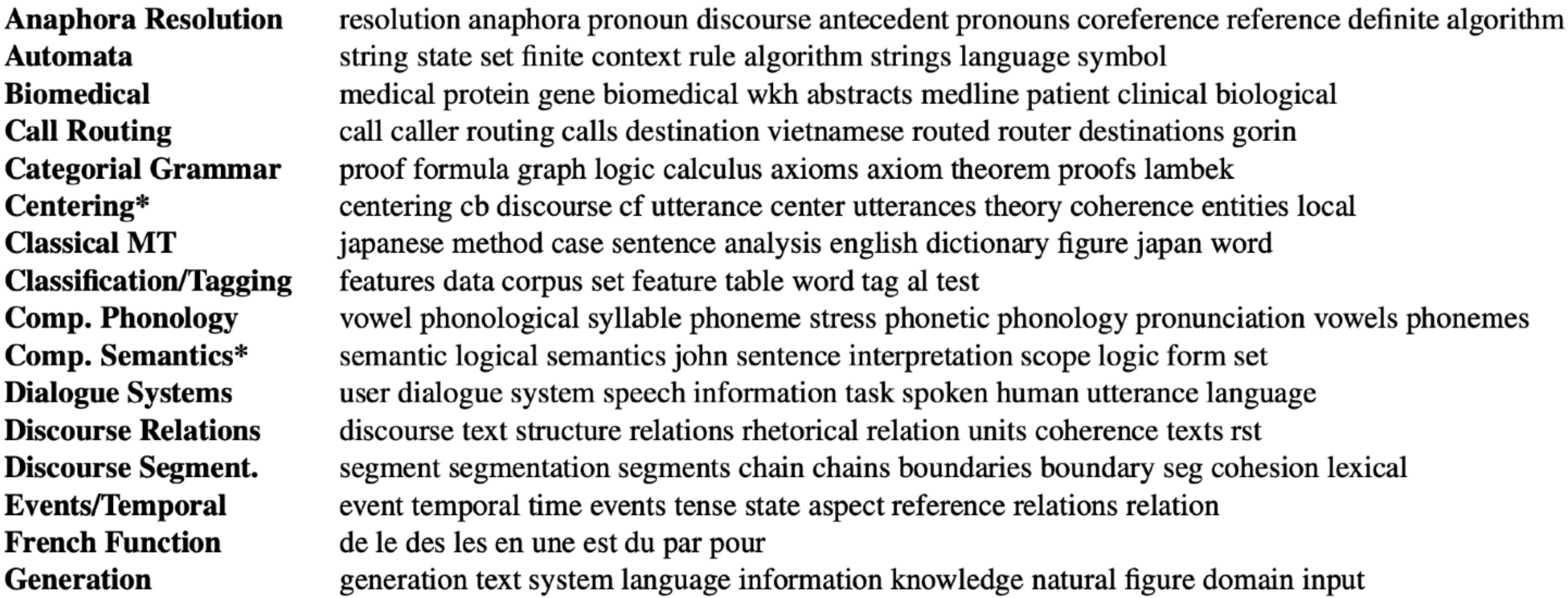
BIDS
BERKELEY INSTITUTE FOR DATA SCIENCE
CENTER FOR CULTURAL ANALYTICS
Lecture & Reception with David Blei

**Sponsor(s):** Berkeley Institute for Data Science (BIDS), Center for Cultural Analytics

*→ Join us in person for a lecture and reception with David Blei, Professor of Statistics and Computer Science at Columbia University.*

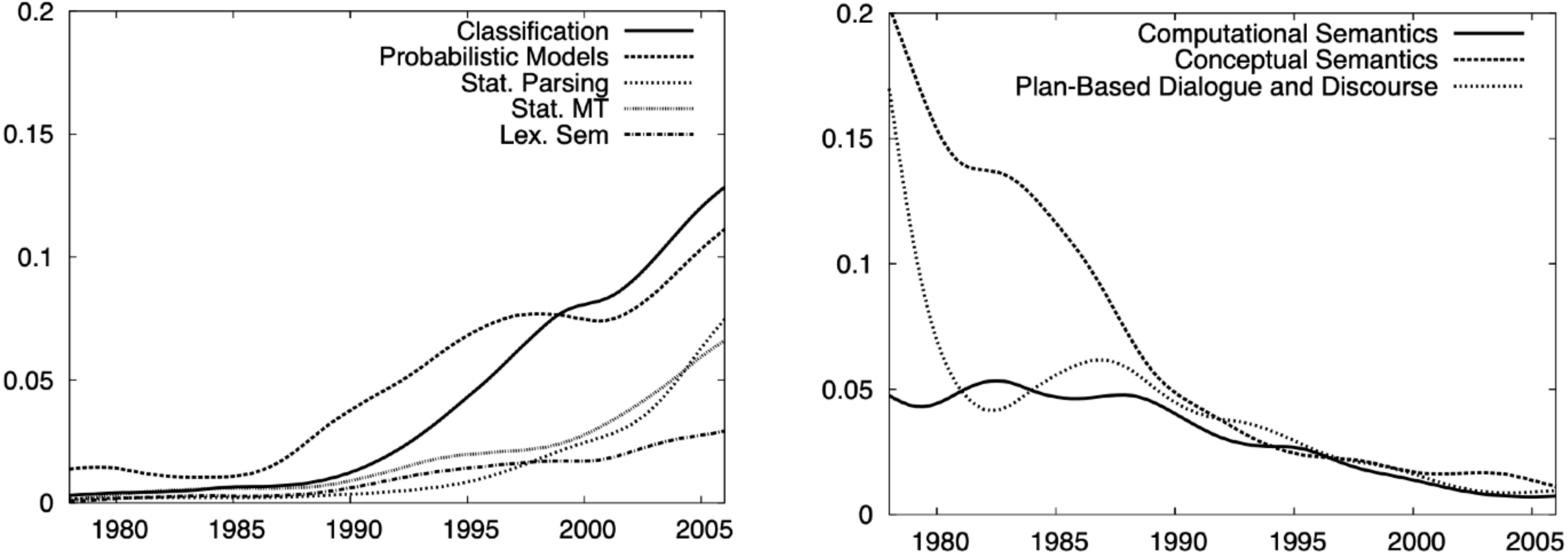
### Beyond Roll Call: Inferring Politics from Text

"The ideal point model is a staple of quantitative political science. It is a probabilistic model of roll call data—how a group of lawmakers vote on a collection of bills—that can be used to quantify the lawmakers' political

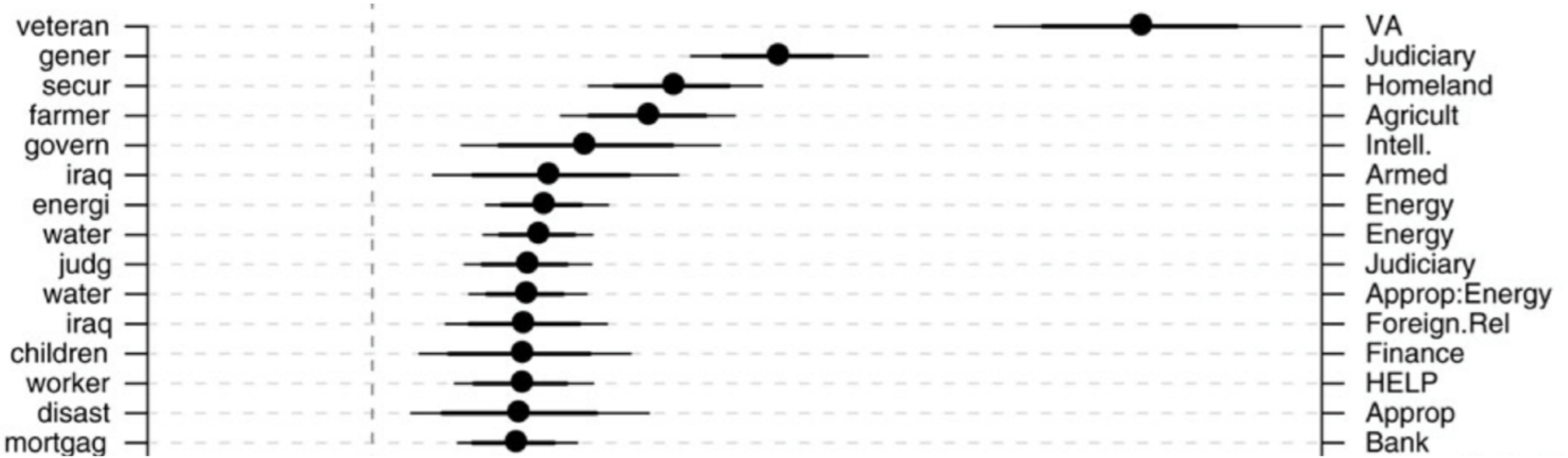


| | |
|---|---|
| **Anaphora Resolution** | resolution anaphora pronoun discourse antecedent pronouns coreference reference definite algorithm |
| **Automata** | string state set finite context rule algorithm strings language symbol |
| **Biomedical** | medical protein gene biomedical wkh abstracts medline patient clinical biological |
| **Call Routing** | call caller routing calls destination vietnamese routed router destinations gorin |
| **Categorial Grammar** | proof formula graph logic calculus axioms axiom theorem proofs lambek |
| **Centering*** | centering cb discourse cf utterance center utterances theory coherence entities local |
| **Classical MT** | japanese method case sentence analysis english dictionary figure japan word |
| **Classification/Tagging** | features data corpus set feature table word tag al test |
| **Comp. Phonology** | vowel phonological syllable phoneme stress phonetic phonology pronunciation vowels phonemes |
| **Comp. Semantics*** | semantic logical semantics john sentence interpretation scope logic form set |
| **Dialogue Systems** | user dialogue system speech information task spoken human utterance language |
| **Discourse Relations** | discourse text structure relations rhetorical relation units coherence texts rst |
| **Discourse Segment.** | segment segmentation segments chain chains boundaries boundary seg cohesion lexical |
| **Events/Temporal** | event temporal time events tense state aspect reference relations relation |
| **French Function** | de le des les en une est du par pour |
| **Generation** | generation text system language information knowledge natural figure domain input |

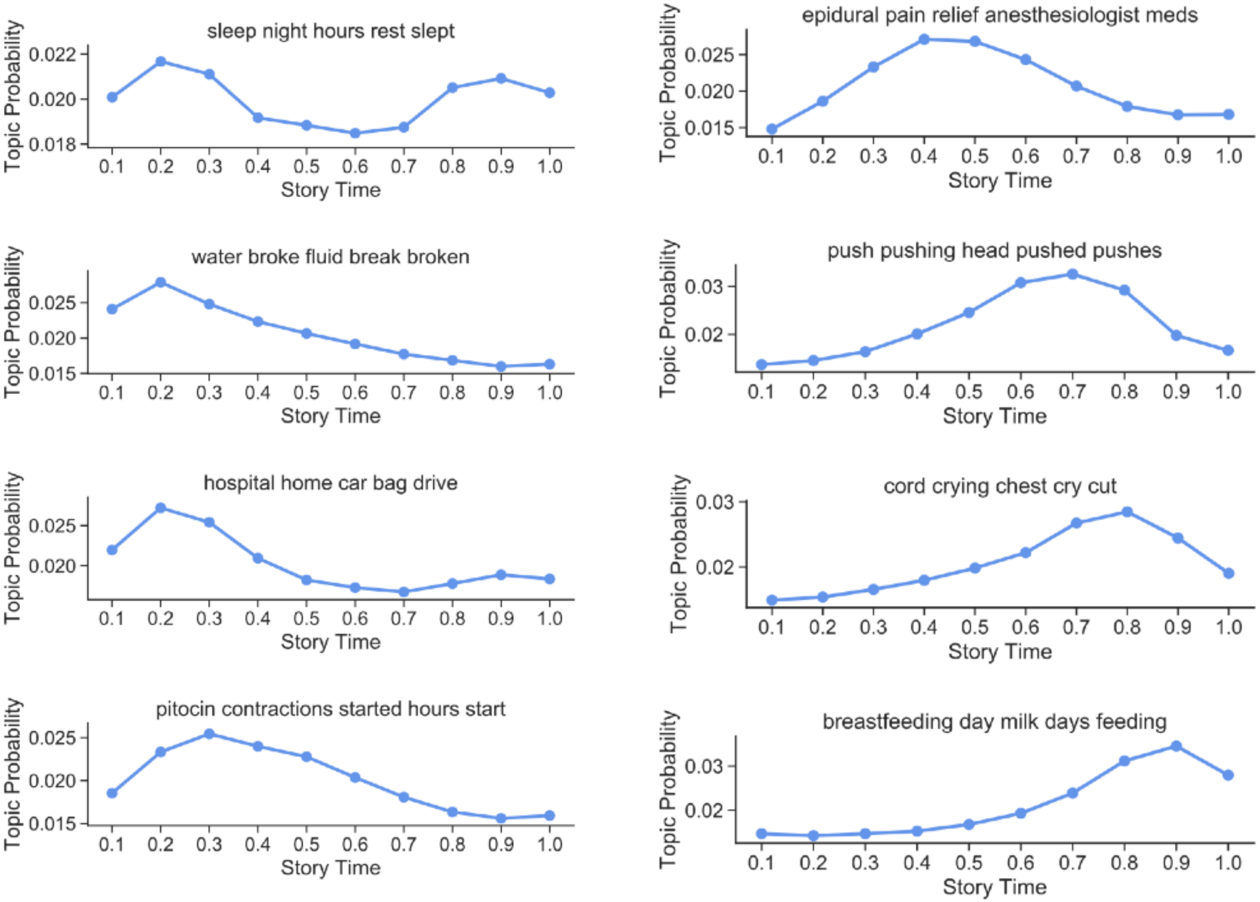
Hall et al. 2008, “Studying the History of Ideas Using Topic Models”



Hall et al. 2008, “Studying the History of Ideas Using Topic Models”



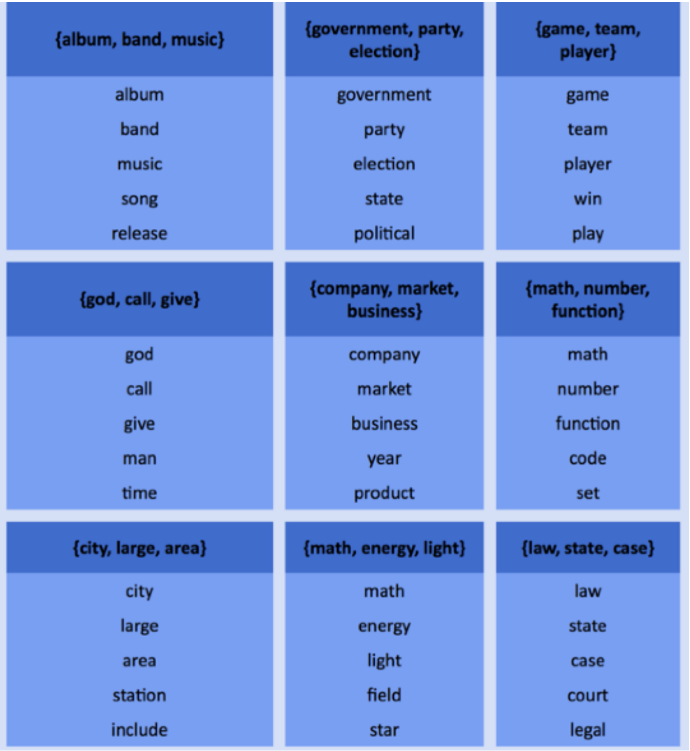
Grimmer (2010), A Bayesian Hierarchical Topic Model for Political Texts:
Measuring Expressed Agendas in Senate Press Releases



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

# 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 |

U.S. Navy Naval Aviator Lieutenant Pete "Maverick" Mitchell and his Radar Intercept Officer (RIO) Lieutenant Junior Grade Nick "Goose" Bradshaw, stationed in the Indian Ocean aboard USS Enterprise, fly the F-14A Tomcat. During an interception with two hostile MiG-28s, Maverick missile-locks on one, while the other hostile locks onto Maverick's wingman, Cougar. Maverick drives it off, but Cougar is so shaken that Maverick defies orders to land and shepherds him back to the carrier. Cougar resigns his commission.

More than 30 years after graduating from Top Gun, United States Navy Captain Pete "Maverick" Mitchell is a decorated test pilot whose repeated insubordination has kept him from flag rank. When Rear Admiral Chester "Hammer" Cain plans to cancel Maverick's hypersonic "Darkstar" scramjet program, Maverick unilaterally changes the target speed for that day's test from Mach 9 to the final contract specification of Mach 10. However, the prototype is destroyed when he cannot resist pushing beyond Mach 10.

Stereotypical Barbie ("Barbie") and fellow dolls reside in Barbieland; a matriarchal society with different variations of Barbies, Kens, and a group of discontinued models, who are treated like outcasts due to their unconventional traits. While the Kens spend their days playing at the beach, considering it as their profession, the Barbies hold prestigious jobs such as doctors, lawyers, and politicians. Beach Ken ("Ken") is only happy when he is with Barbie and seeks a closer relationship, but Barbie rebuffs him in favor of other activities and female friendships.

# 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.”

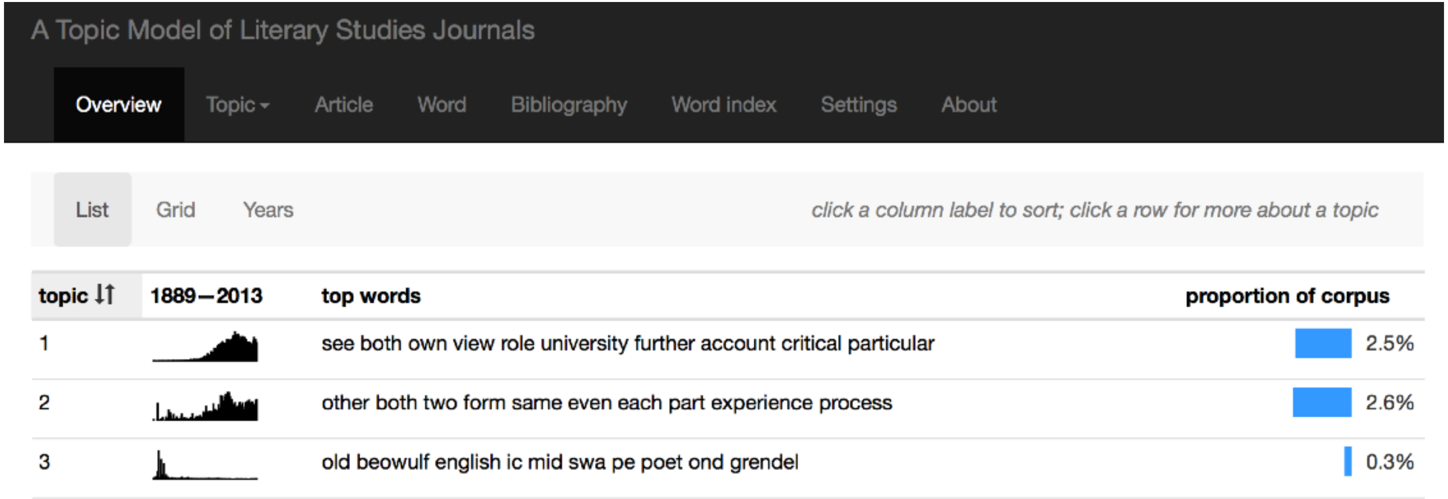
# tokens, not types

… 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



*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 |
|---|---|
| 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 |

$\beta$ = coefficients

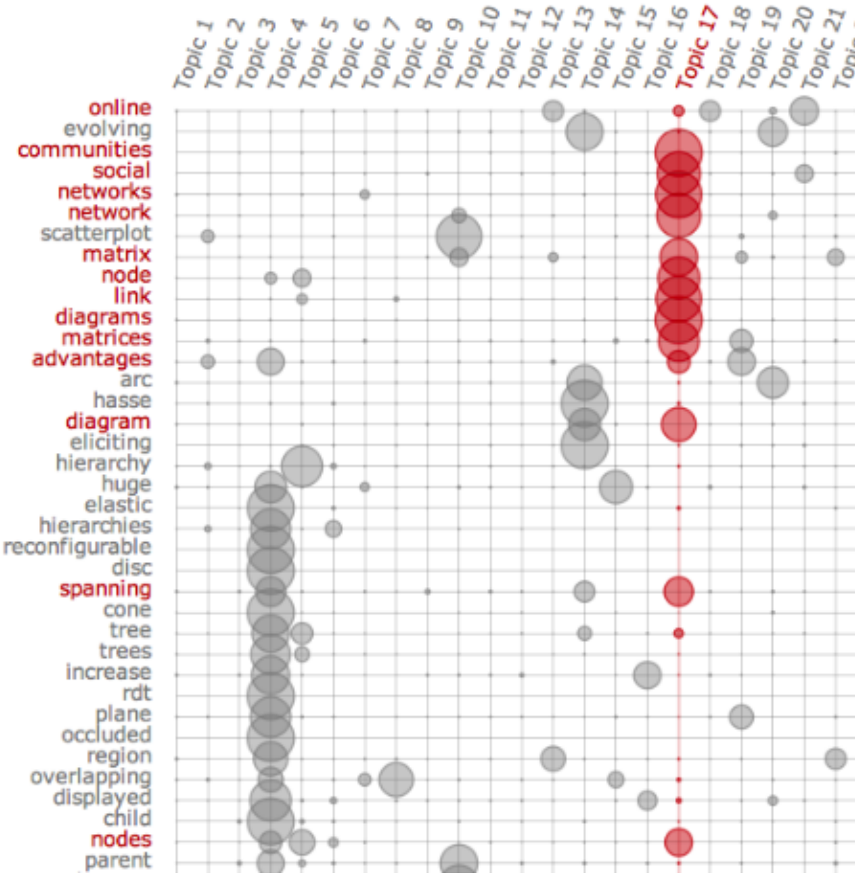
| Feature | $\beta$ |
|---|---|
| contains “love” | -3.1 |
| contains “castle” | 6.8 |
| contains “dagger” | 7.9 |
| contains “run” | -3.0 |
| contains “the” | -1.7 |
| topic 1 | 0.3 |
| topic 2 | -1.2 |
| topic 3 | 5.7 |

# Software

- Mallet
  http://mallet.cs.umass.edu/

- Gensim (python)
  https://radimrehurek.com/gensim/

- Visualization
  https://github.com/uwdata/termite-visualizations



# 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

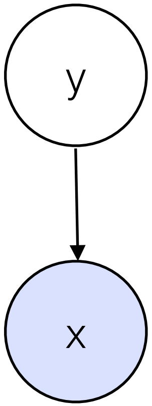
# Probabilistic graphical models

- Nodes represent variables (shaded = observed, clear = latent)

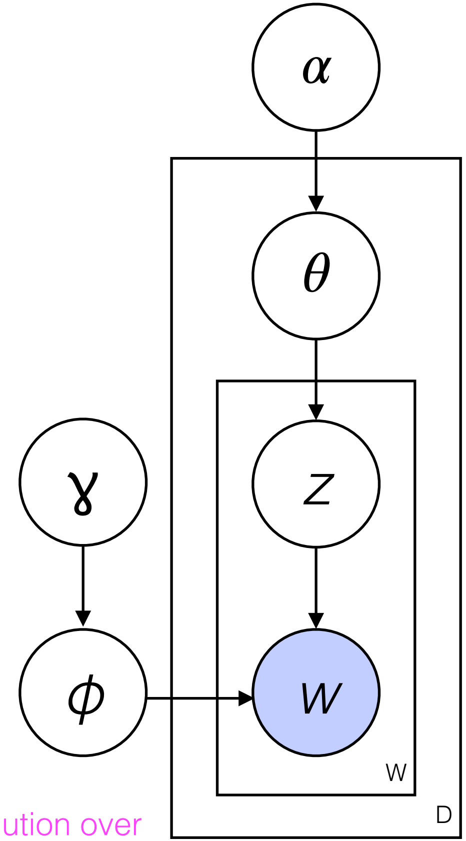
- Arrows indicate conditional relationships

- The probability of x here is dependent on y

- Simply a visual way of writing the joint probability:



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



document distribution over topics

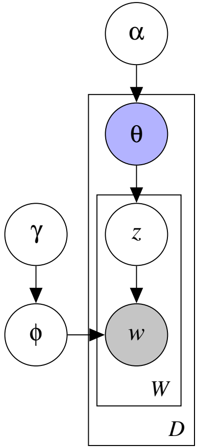
topic indicators for words

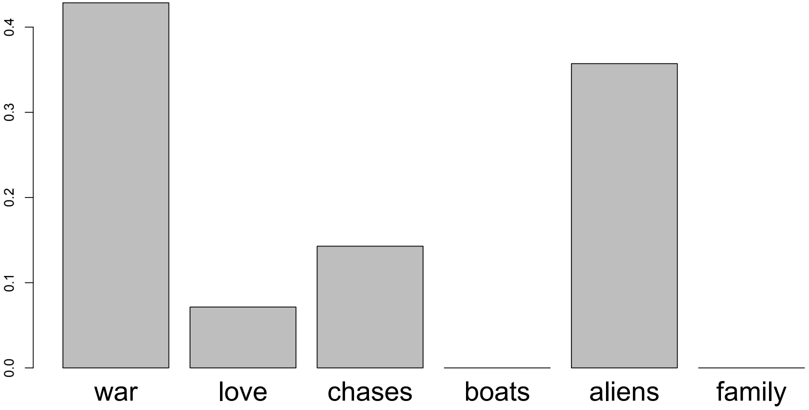
words

topic distribution over words

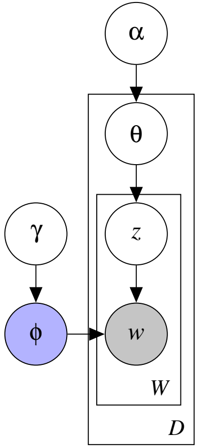
# Topic Models

- A document has *distribution over topics*

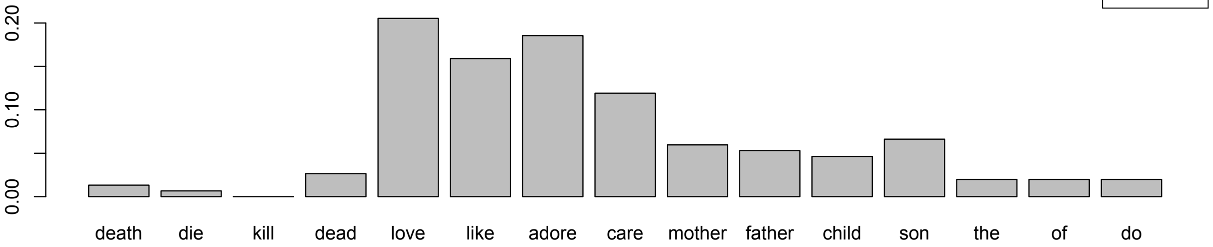




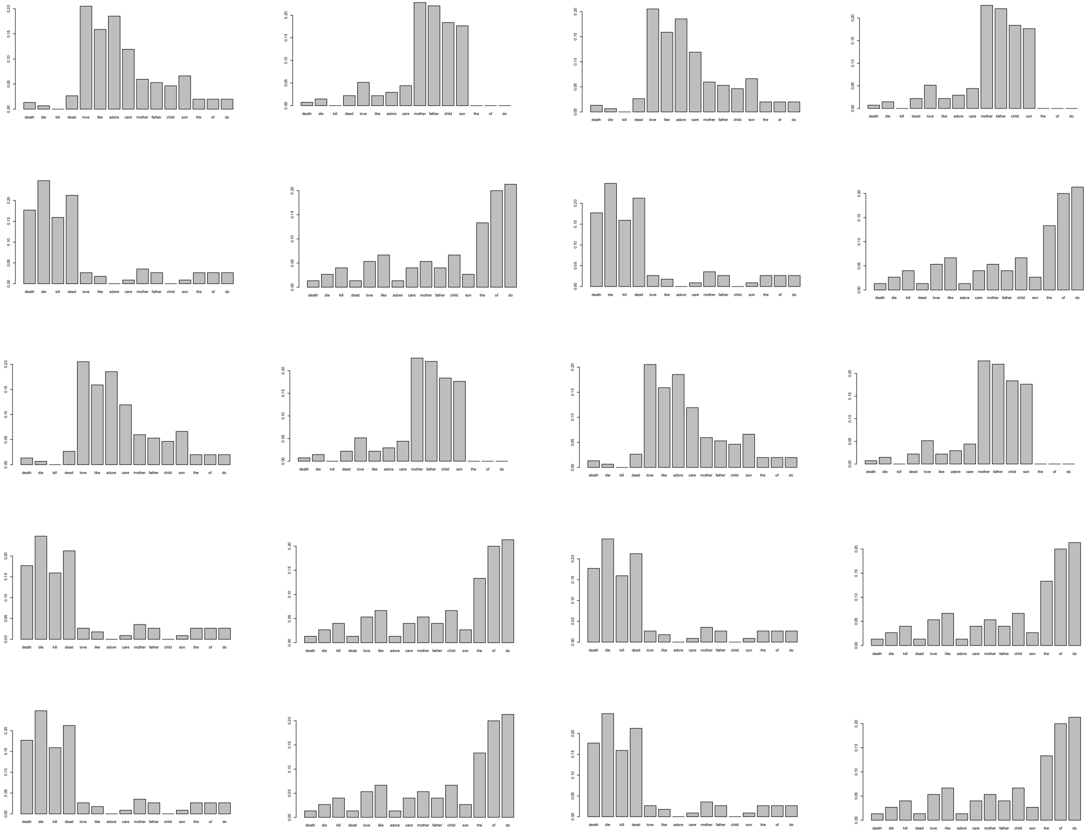
# Topic Models

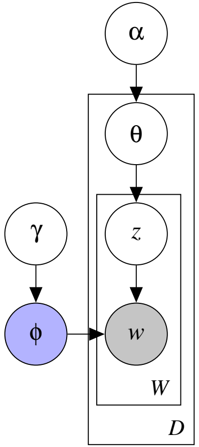


- A topic is a distribution over words

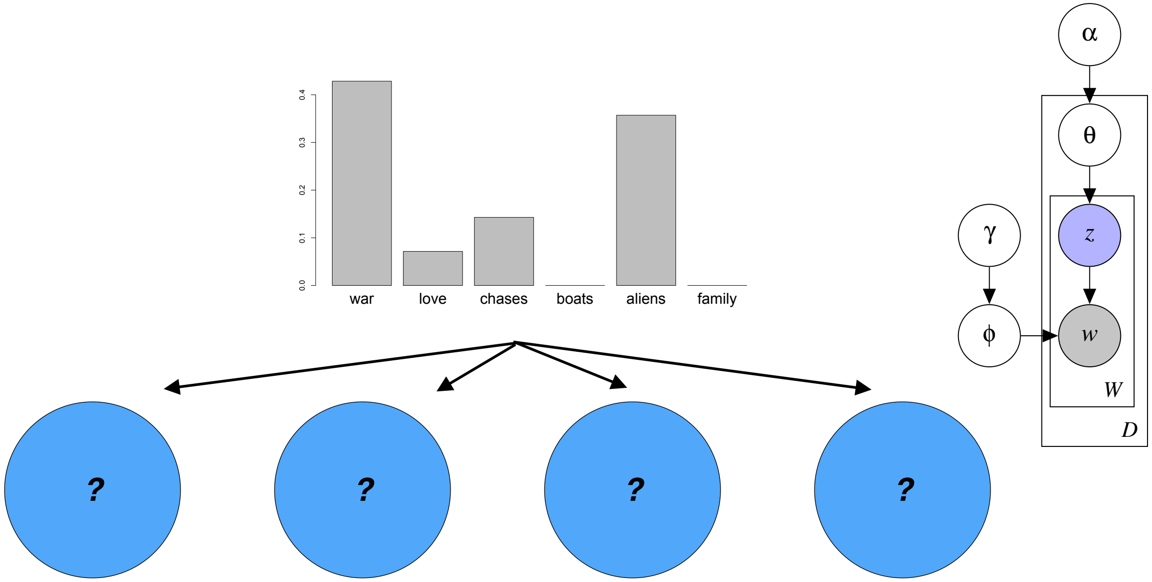


- e.g., P("adore" | topic = love) = .18

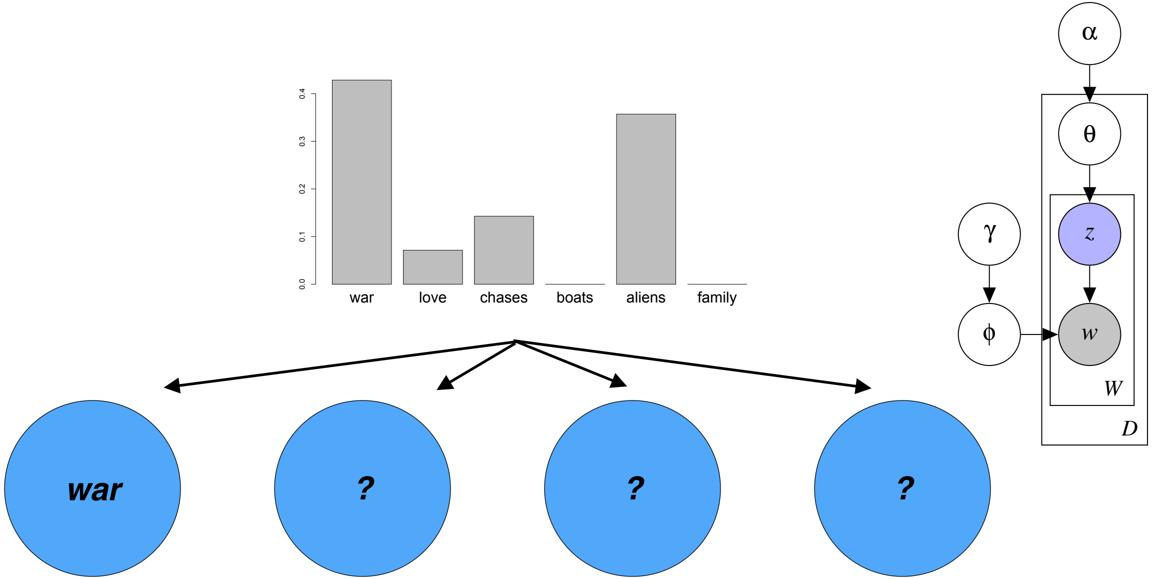

0.00
0.05
0.10
0.15
0.20
death
die
kill
dead
love
like
adore
care
mother
father
child
son
the
of
do


α
θ
γ
z
ϕ
w
W
D

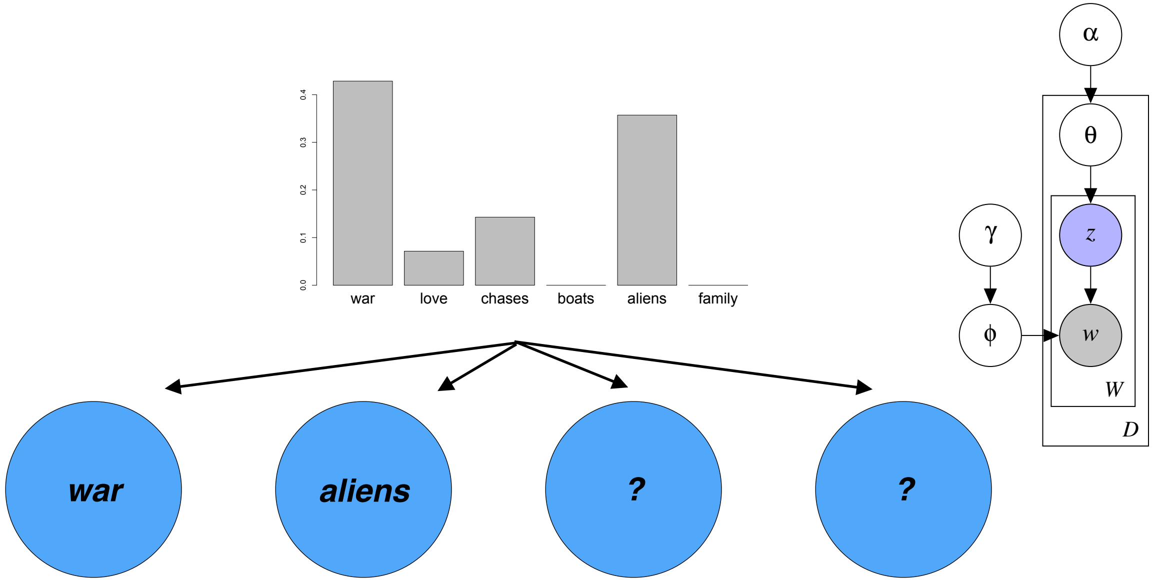
K=20



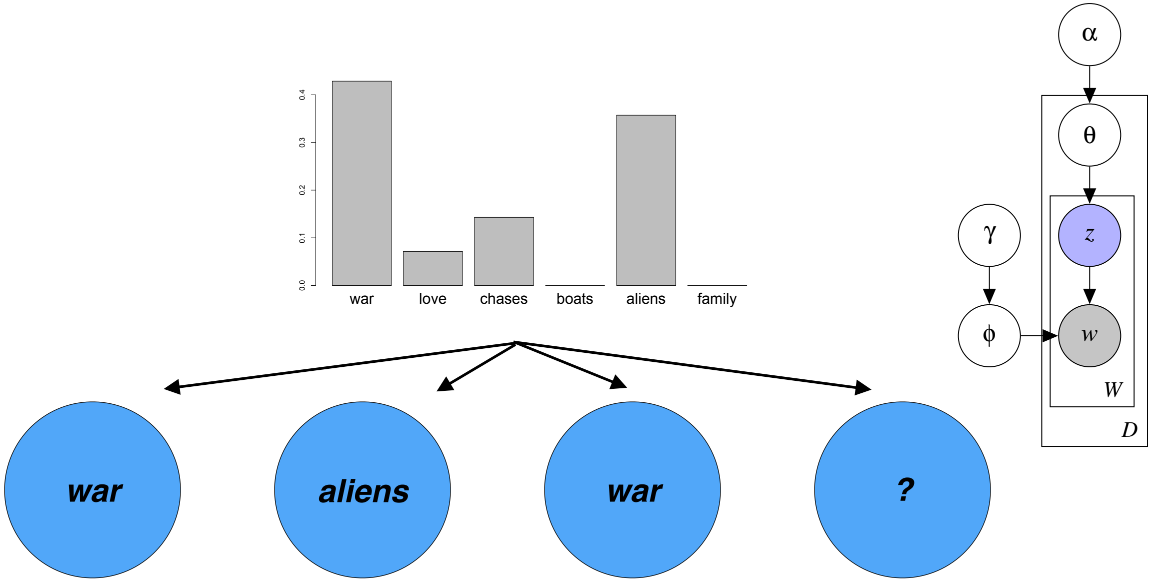
P(topic | topic distribution)



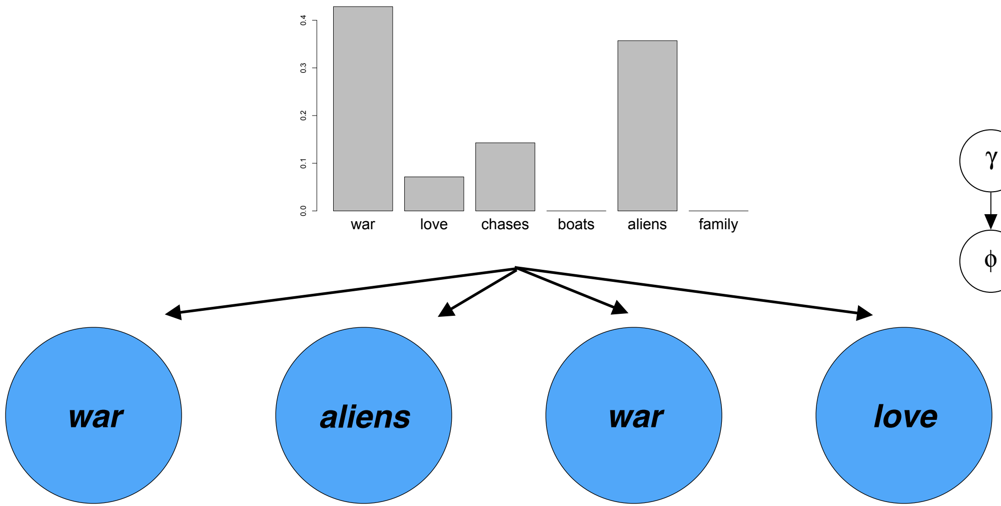
P(topic | topic distribution)

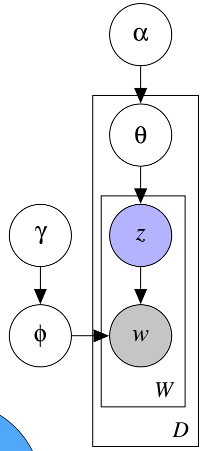


P(topic | topic distribution)

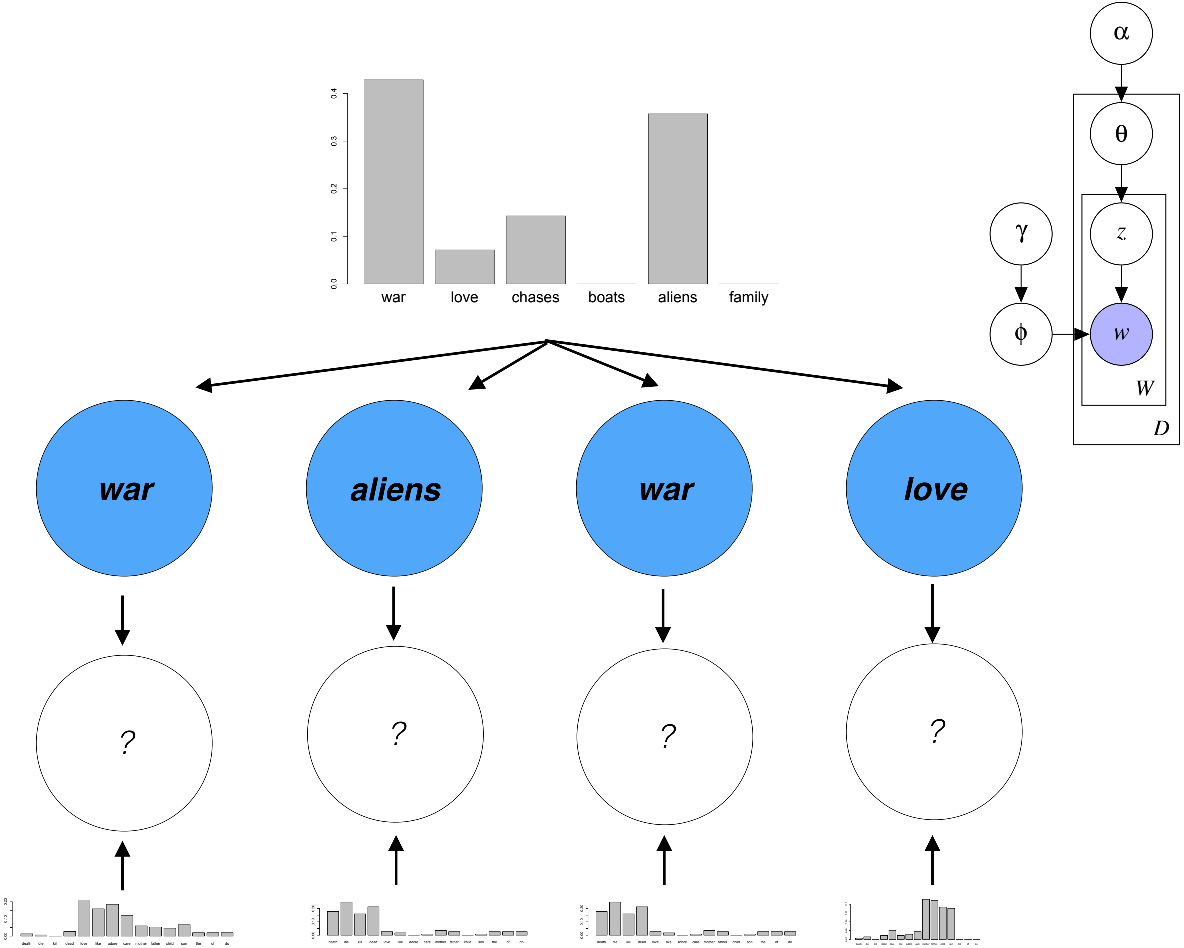


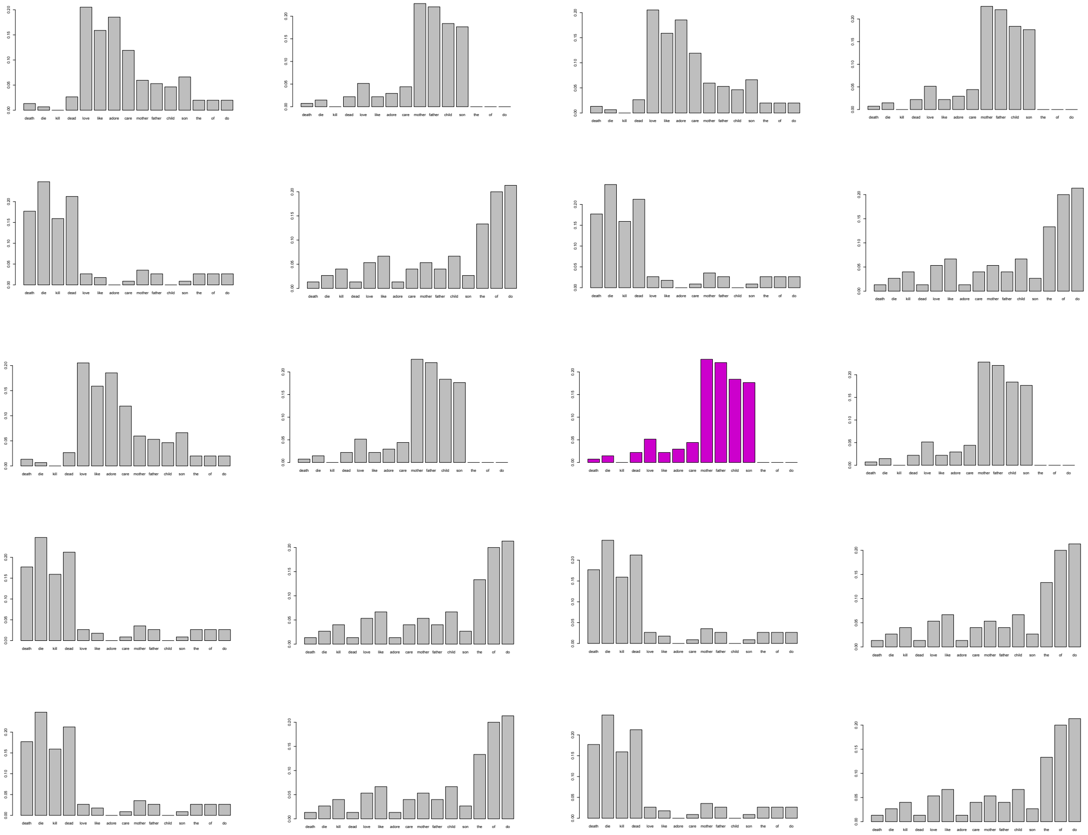
P(topic | topic distribution)

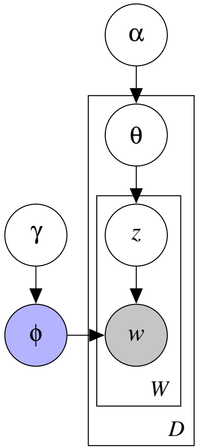




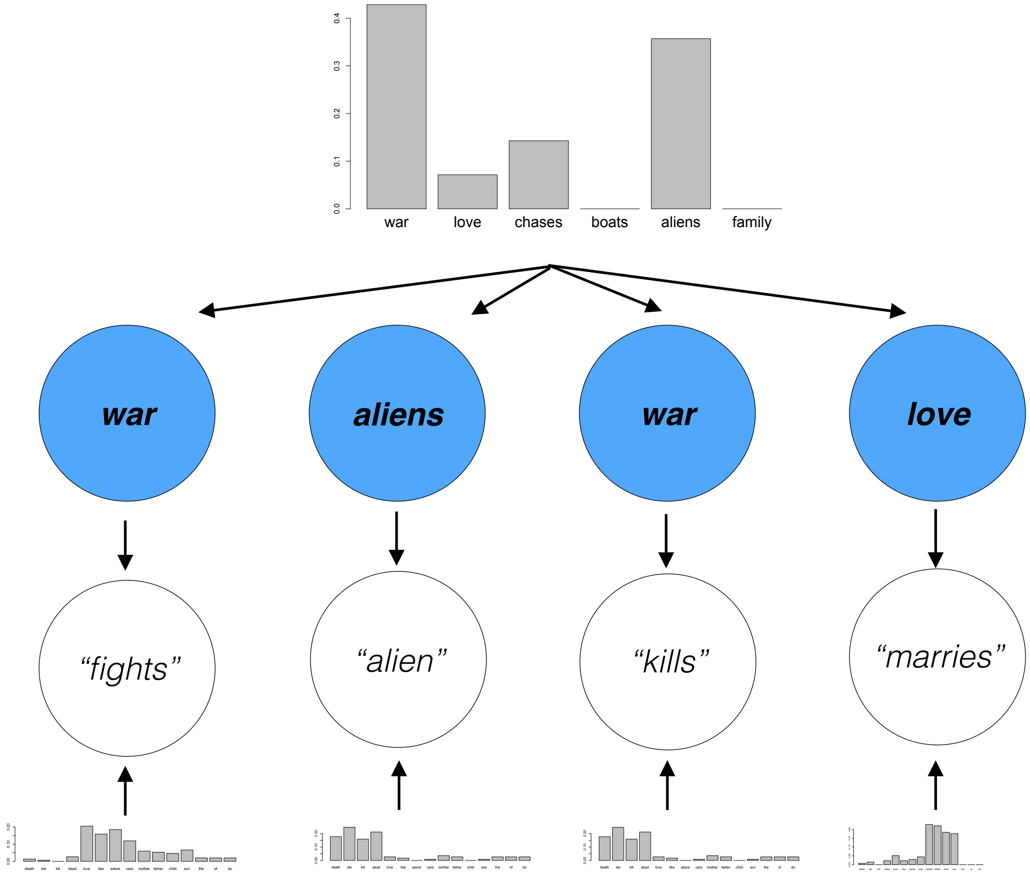
P(topic | topic distribution)

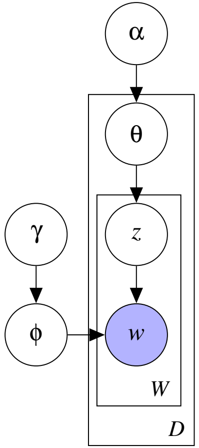

0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
war
aliens
war
love
?
?
?
?
death
die
kill
dead
love
like
adore
care
mother
father
child
son
the
of
do
α
θ
γ
z
ϕ
w
W
D

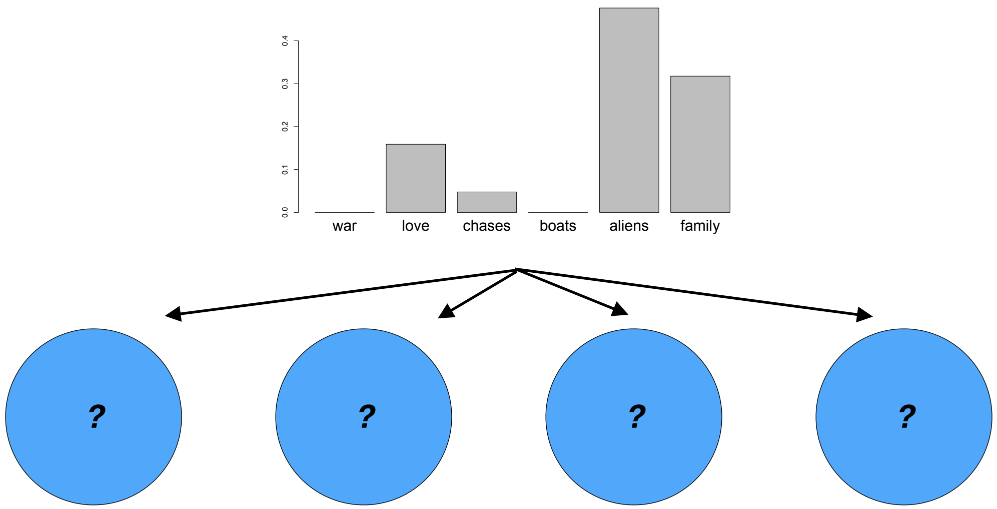

0.00
0.05
0.10
0.15
0.20
death
die
kill
dead
love
like
adore
care
mother
father
child
son
the
of
do

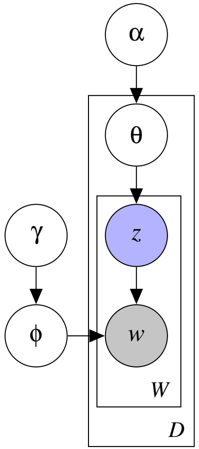

α
θ
γ
z
ϕ
w
W
D

K=20

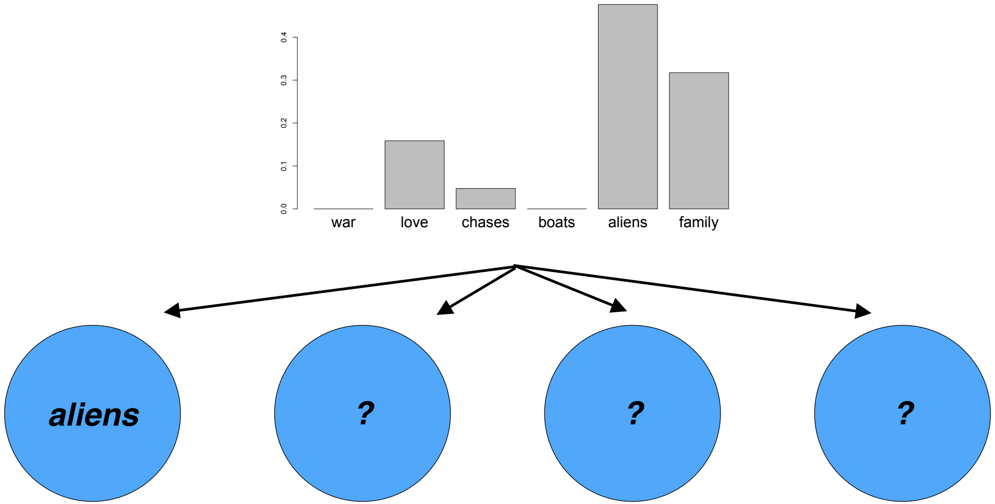

0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
war
aliens
war
love
“fights”
“alien”
“kills”
“marries”
0.00
0.10
0.20
death
die
kill
dead
love
like
adore
care
mother
father
child
son
the
of
do

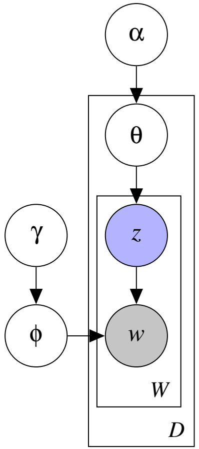

α
θ
γ
z
ϕ
w
W
D



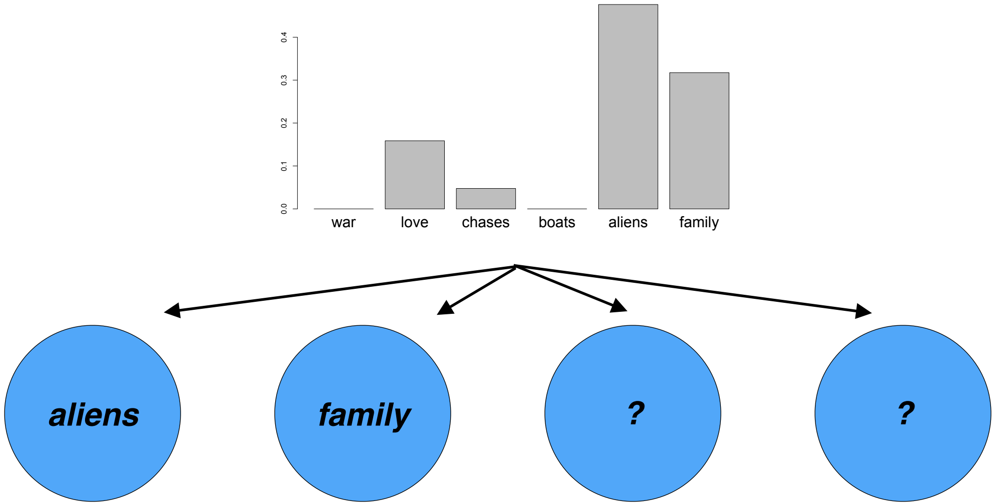


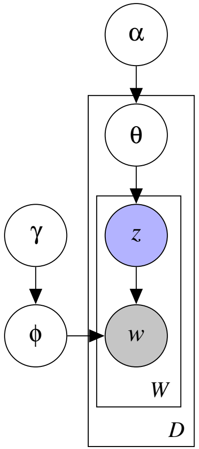
P(topic | topic distribution)



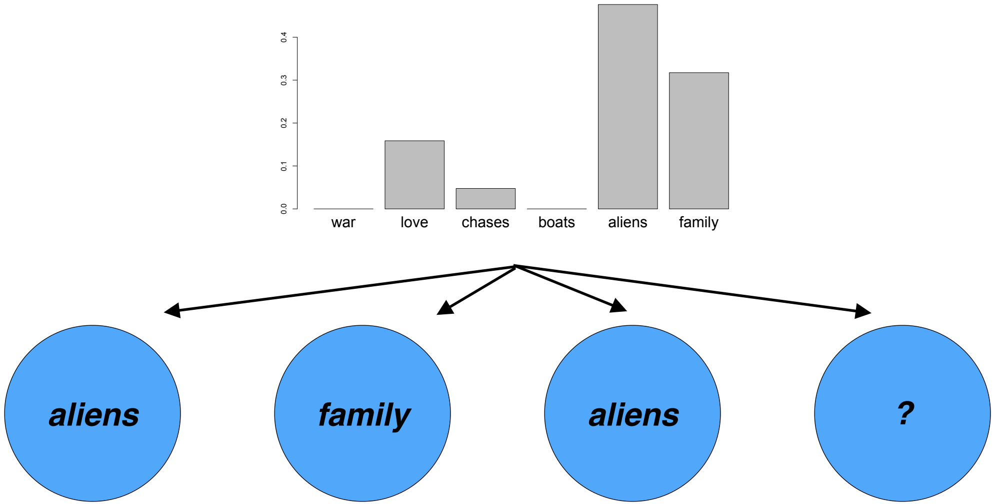


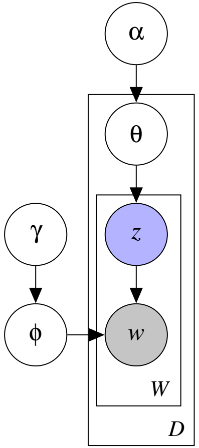
P(topic | topic distribution)



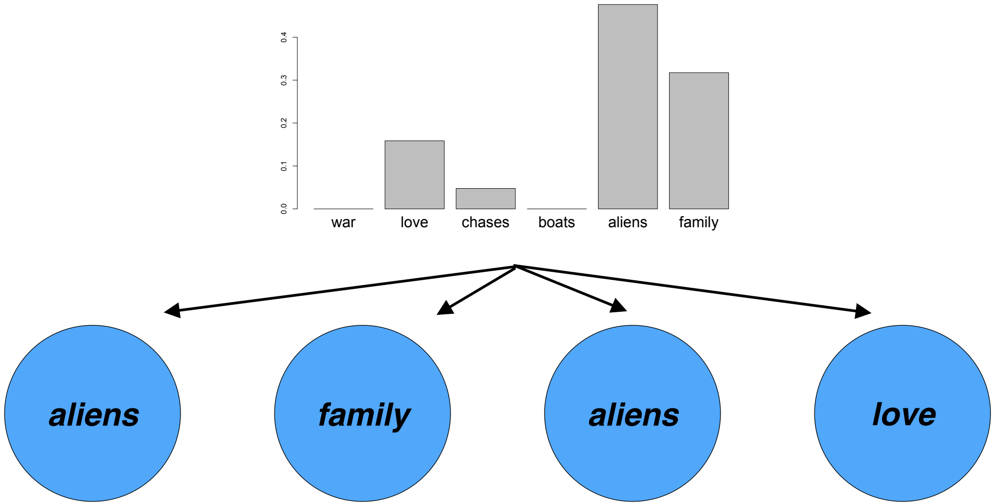


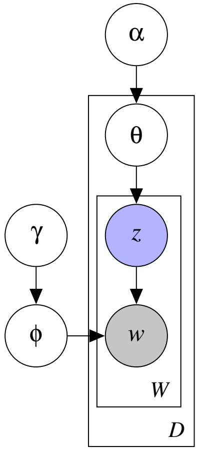
P(topic | topic distribution)



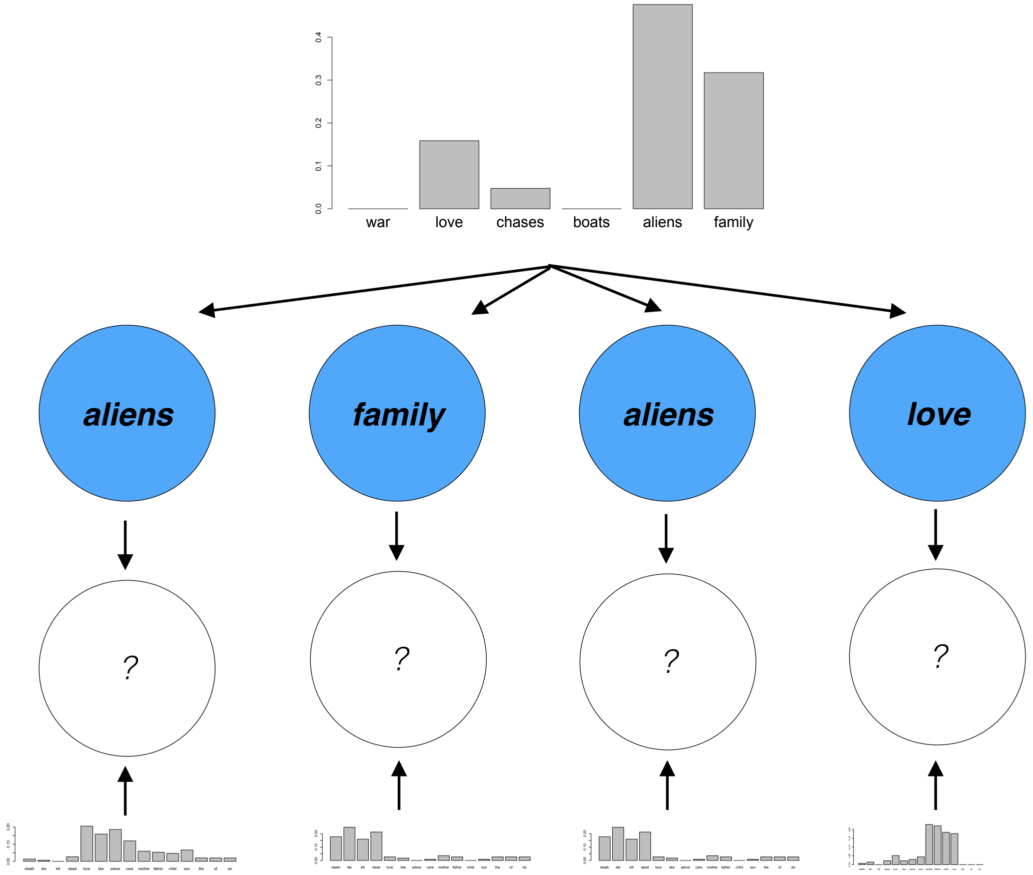


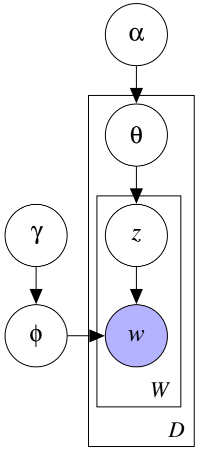
P(topic | topic distribution)

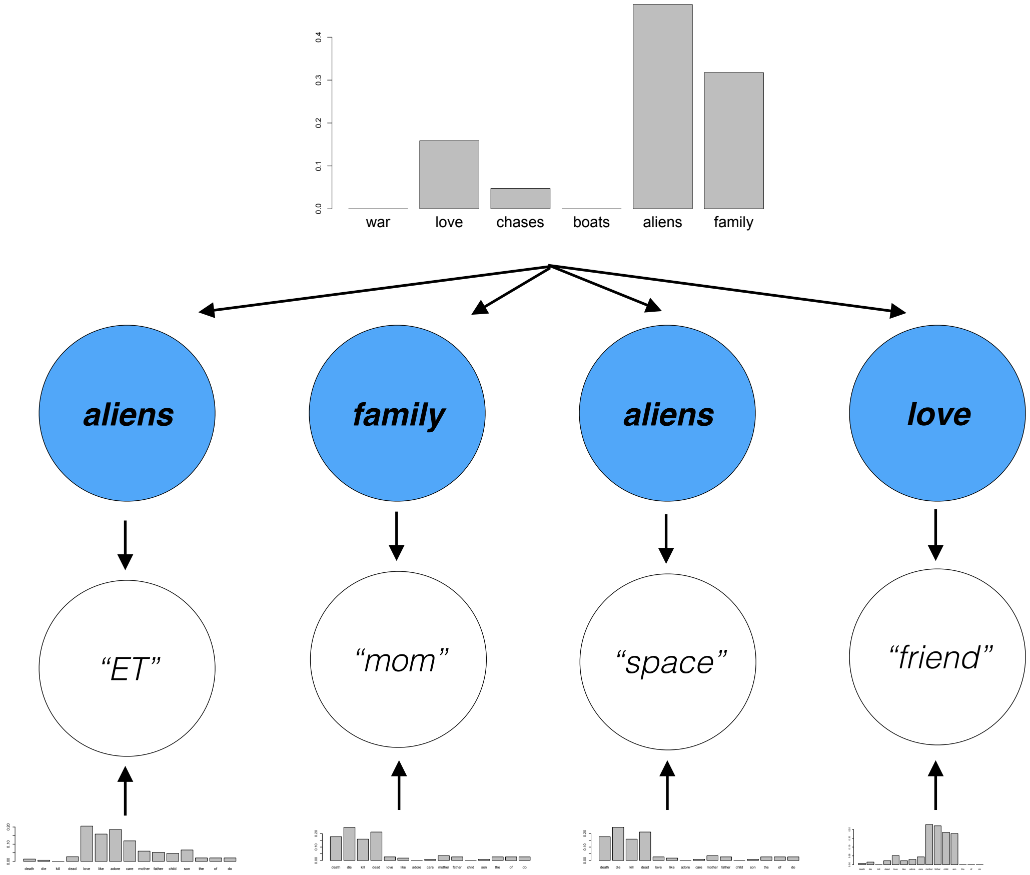


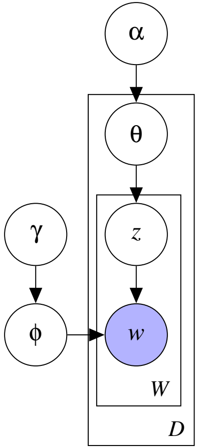


P(topic | topic distribution)


0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
aliens
family
aliens
love
?
?
?
?
0.00
0.10
0.20
death die kill dead love like adore care mother father child son the of do
0.00 0.10 0.20
death die kill dead love like adore care mother father child son the of do
0.00 0.10 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


0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
aliens
family
aliens
love
“ET”
“mom”
“space”
“friend”
0.00
0.10
0.20
death
die
kill
dead
love
like
adore
care
mother
father
child
son
the
of
do

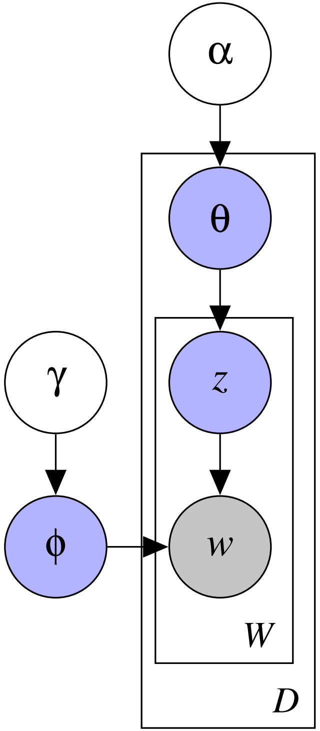

α
θ
γ
z
ϕ
w
W
D

# Inference

- What are the topic distributions for each document?

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

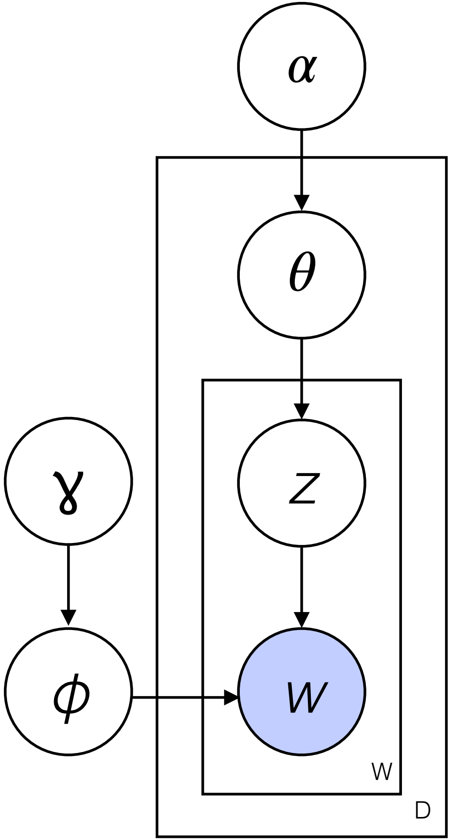
- What are the word distributions for each topic?



**Find the parameters that maximize the likelihood of the data!**

# Inference

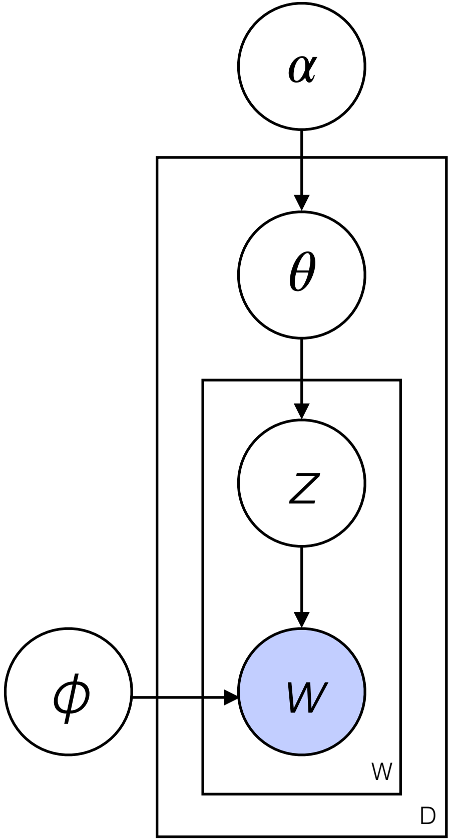
- Markov chain Monte Carlo (Gibbs sampling, Metropolis Hastings, etc.)
- Variational methods
- Spectral methods (Anandkumar et al. 2012, Arora et al. 2013)

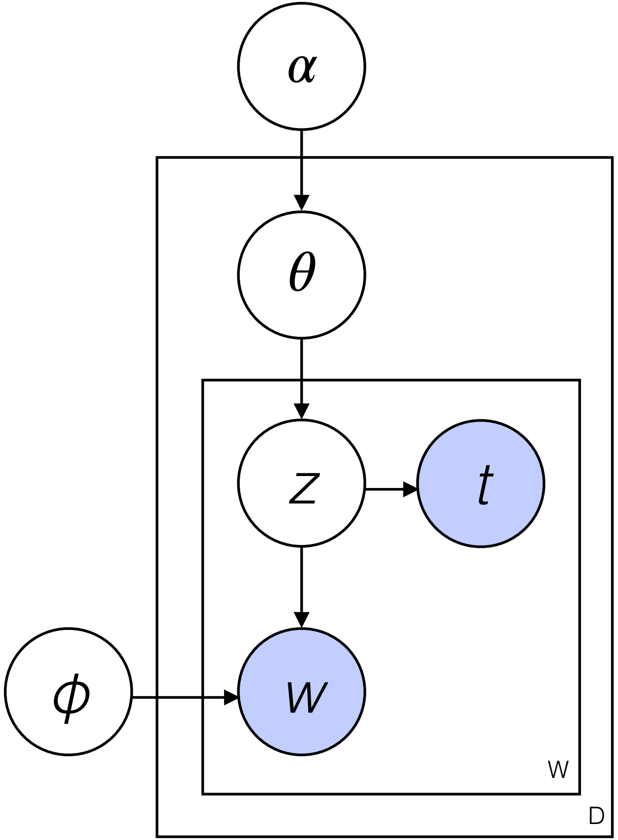


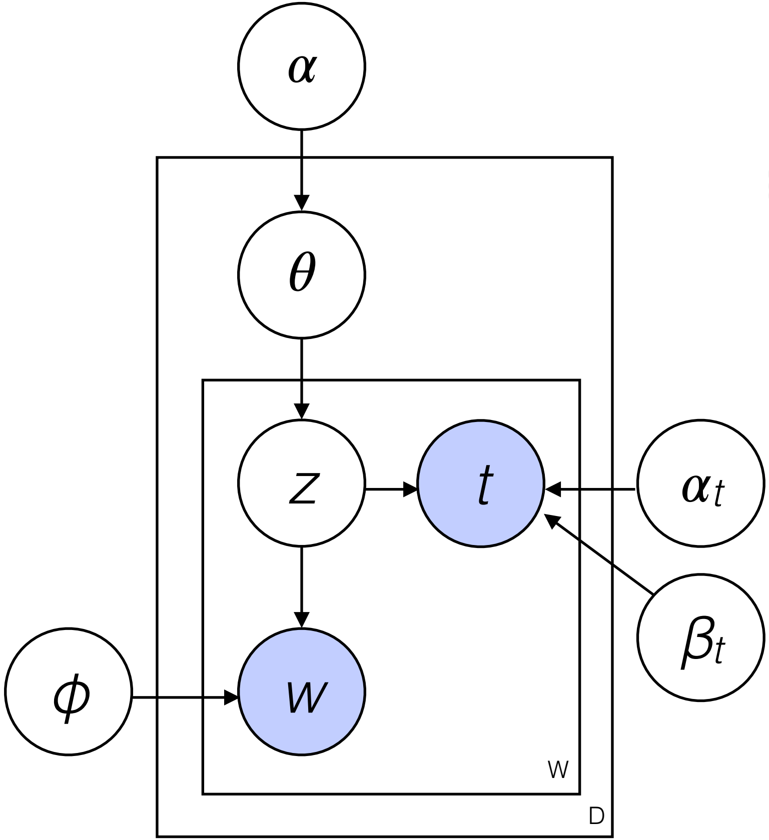
# Assumptions

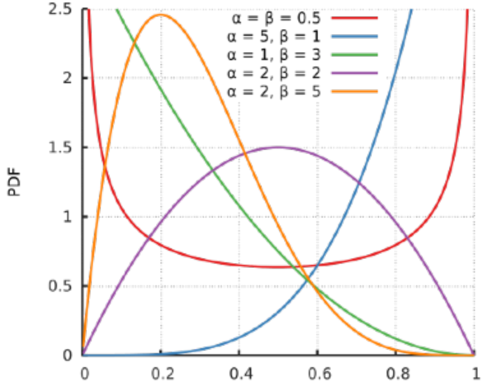
- Every word has one topic
- Every document has one topic distribution
- No sequential information (topics for words are independent of each other given the set of topics for a document)
- Topics don’t have arbitrary correlations (Dirichlet prior)
- Words don’t have arbitrary correlations (Dirichlet 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?


$\alpha$
$\theta$
$z$
$w$
$\phi$
W
D


$\alpha$
$\theta$
$z$
$t$
$\phi$
$w$
W
D

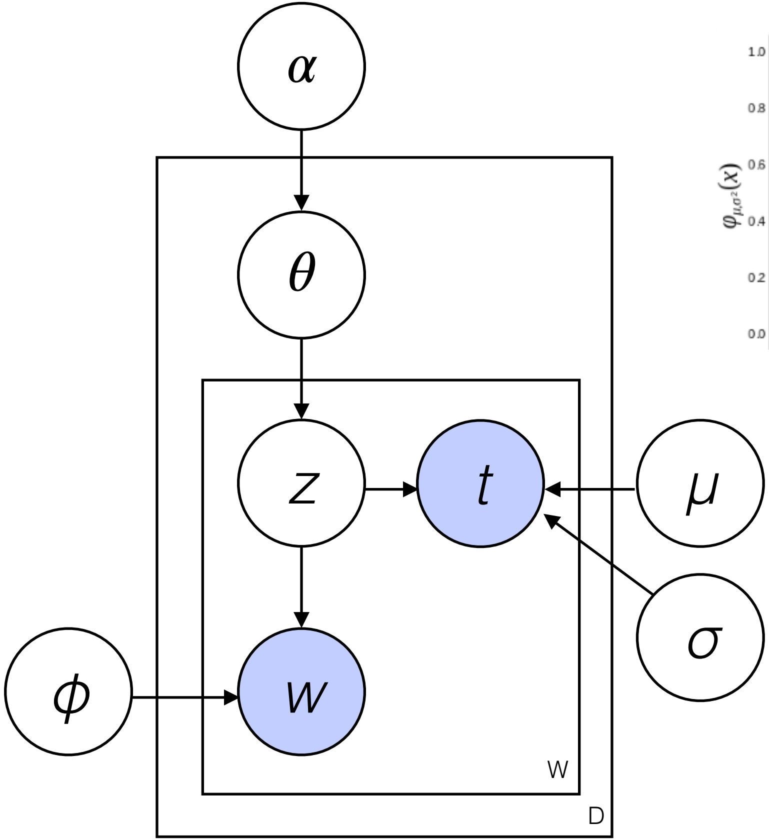


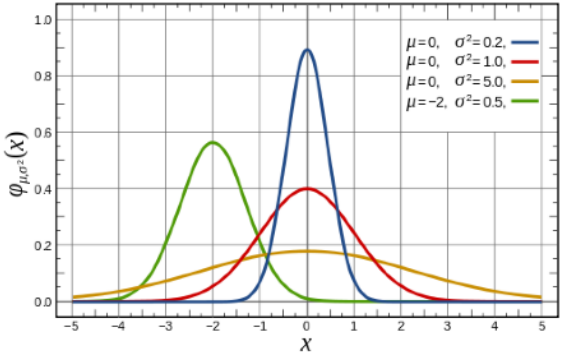


Time is drawn from a Beta distribution

[0,1]

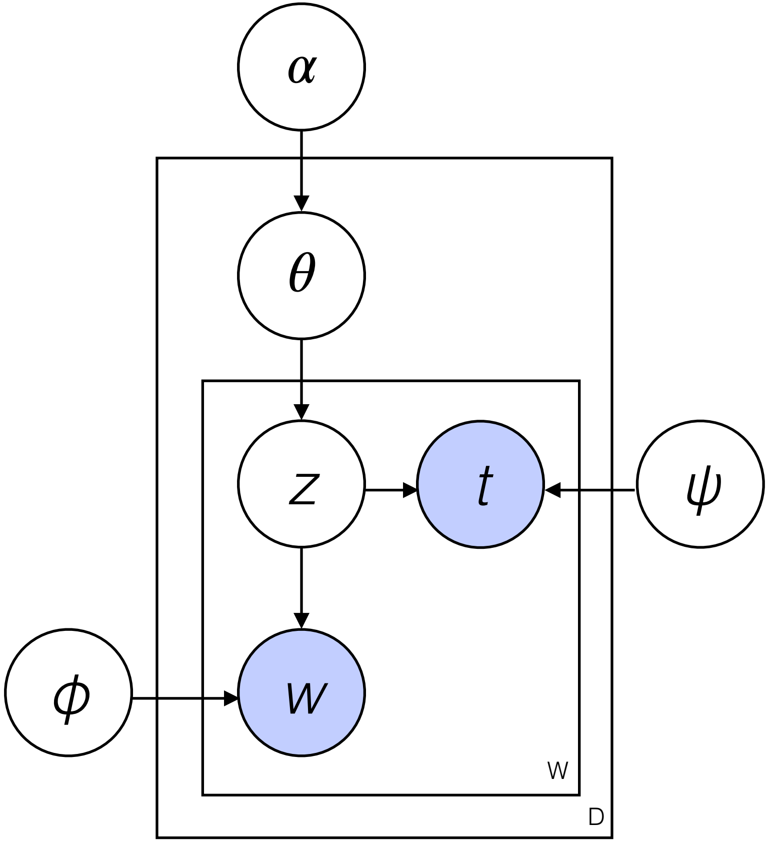
(Wang and McCallum 2006)

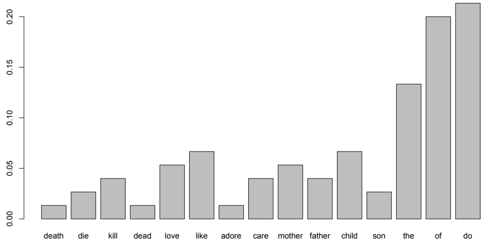




Time is drawn from a Normal distribution

[-∞, ∞]





Time is drawn from a Multinomial distribution

[1, … , K]

# Activity

```
4.topics/TopicModeling
```