Machine Learning for the Computational Humanities

David Bamman
Carnegie Mellon University

Oct 24, 2014

#mlch

Overview

- Classification
  - Probability
  - Independent (Logistic regression, Naive Bayes)
  - Structured (CRFs, HMMs)
- Clustering (hierarchical, K-means)
- Probabilistic graphical models (e.g., topic models)
- Representation learning

The big two

- Classification
  - Given a pre-defined set of categories, determine which category (or categories) apply to the text. Example: spam vs. not spam.
- Clustering
  - Learn coherent groups according to some notion of similarity.

Classification

- Supervised classification learns a mapping from an input to an output from training data

Application	Input	Output
Spam filtering	email	spam, not spam
Authorship attribution	document	author
Sentiment analysis	text	position, negative
Part of speech tagging	sentence	sequence of part of speech tags

Training data

Label	Input
Jane Austen	It is a truth universally acknowledged, that a single man in possession …
Jane Austen	Emma Woodhouse, handsome, clever, and rich, with a comfortable home …
Jane Austen	The family of Dashwood had long been settled in Sussex. Their estate…
Jane Austen	Sir Walter Elliot, of Kellynch Hall, in Somersetshire, was a man who, for …
Herman Melville	Call me Ishmael. Some years ago--never mind how long precisely…
Herman Melville	I am a rather elderly man. The nature of my avocations for the last thirty …
Mark Twain	You don't know about me without you have read a book by the name of…

What do you need?

Two steps to building and using a supervised classification model.

1. Train a model with data where you know the answers.

2. Use that model to predict data where you don't.

What do you need?

1. Data (emails, texts)

2. Labels for each data point (spam/not spam, which author it was written by)

3. A way of “featurizing" the data that’s conducive to discriminating the classes

4. To know that it works.

Recognizing a Classification Problem

- Can you formulate your question as a *choice* among some universe of possible classes?

- Can you create (or find) labeled data that marks that choice for a bunch of examples? Can ***you*** make that choice?

- Can you create features that might help in distinguishing those classes?

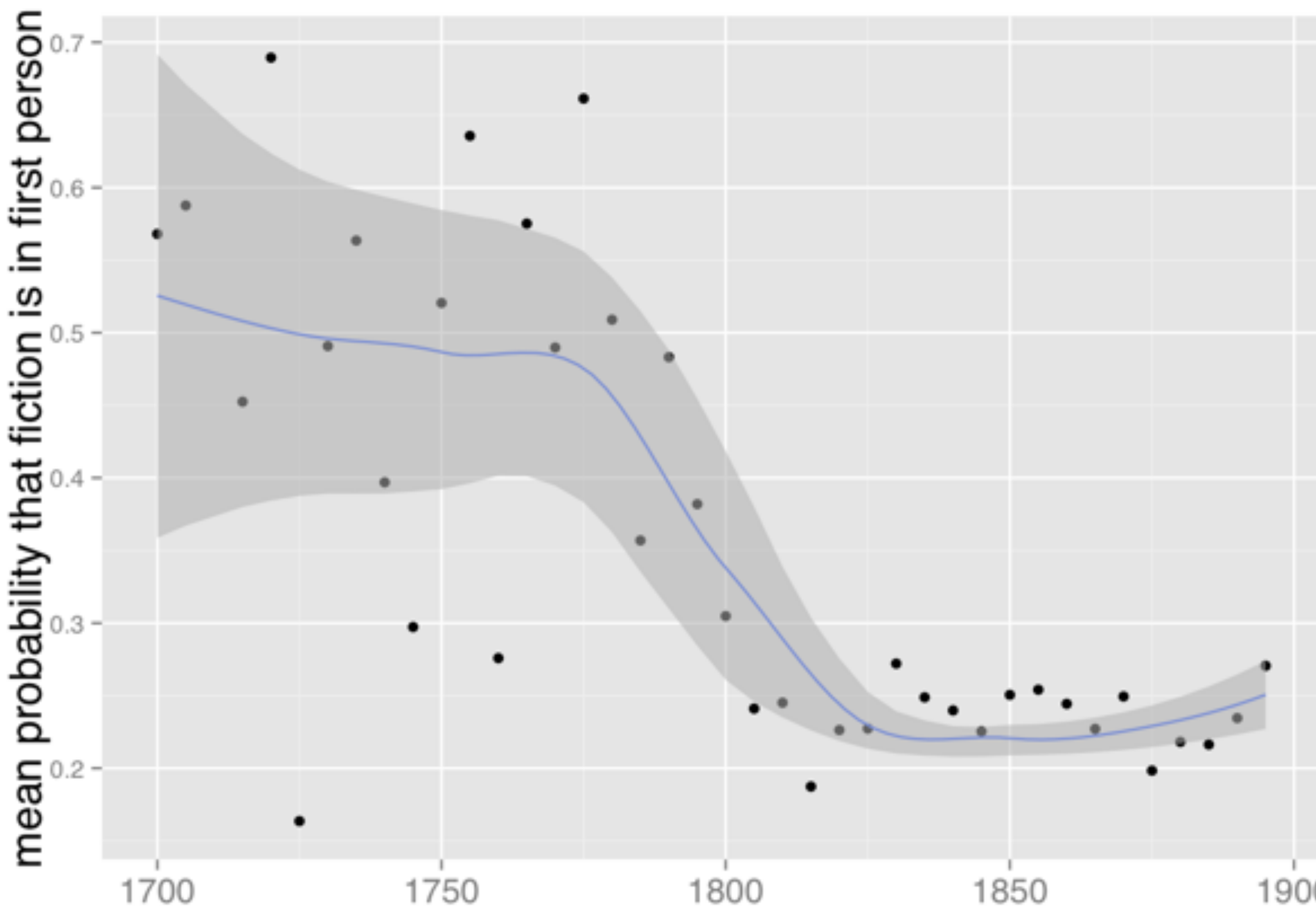
1. Those that belong to the emperor
2. Embalmed ones
3. Those that are trained
4. Suckling pigs
5. Mermaids (or Sirens)
6. Fabulous ones
7. Stray dogs
8. Those that are included in this classification
9. Those that tremble as if they were mad
10. Innumerable ones
11. Those drawn with a very fine camel hair brush
12. Et cetera
13. Those that have just broken the flower vase
14. Those that, at a distance, resemble flies

The “Celestial Emporium of Benevolent Knowledge” from Borges (1942)

{Tragedy, Comedy}

Point of view

Classifying 1st- vs. 3rd- person narration in 32K works of English-language fiction

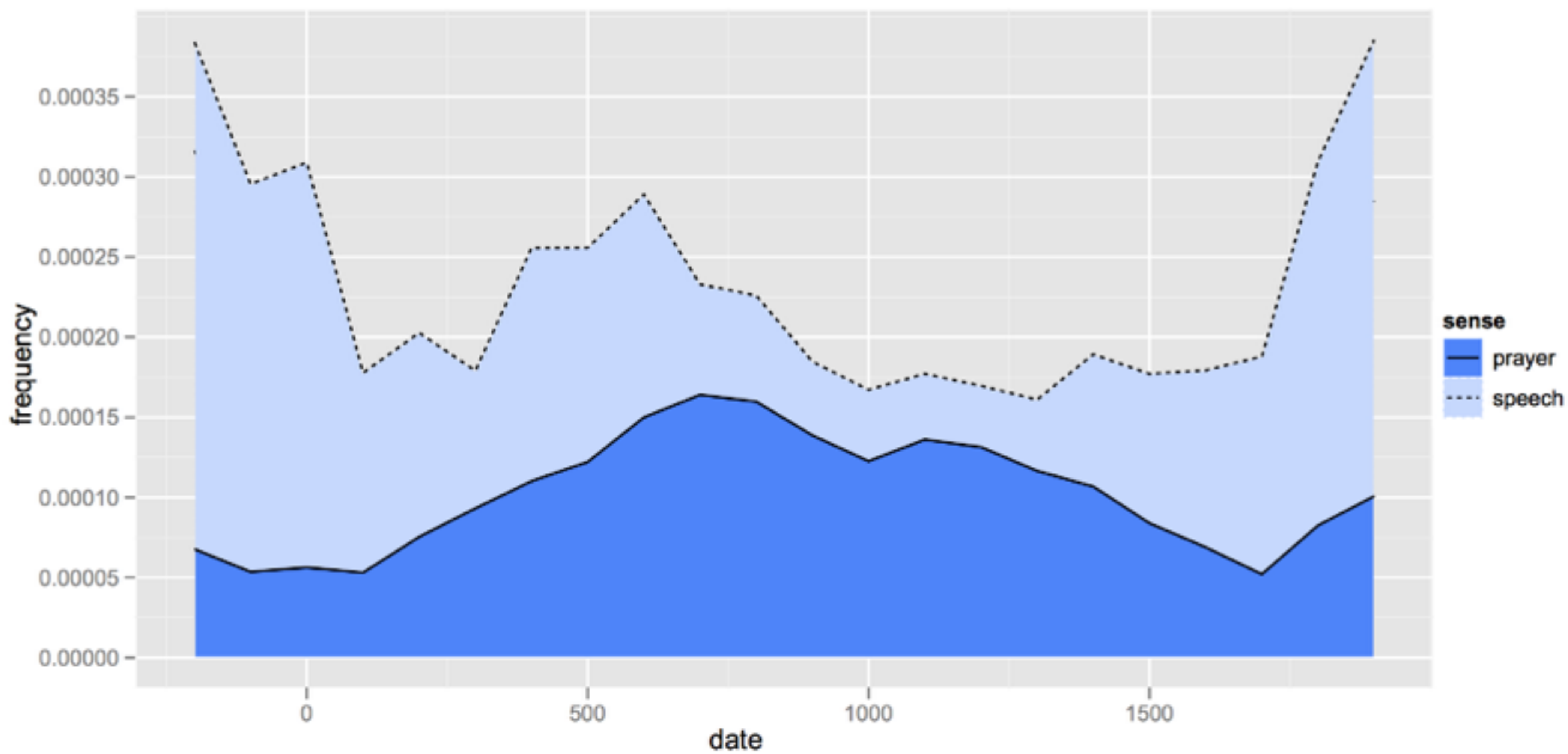


Ted Underwood, “Genre, gender and point of view” http://tedunderwood.com/2013/09/22/genre-gender-and-point-of-view/

#mlch

Word sense

Classifying Latin “oratio” as *speech* vs. *prayer.*



Bamman and Crane, “Measuring Historical Word Sense Variation (JCDL 2011)

#mlch

Recognizing a Classification Problem

- I want to find all of the texts that have allusions to *Paradise Lost*.
- I want to know when discussions of “electricity” changed from magical to scientific.
- I want to find all of the “love oaths” in Shakespeare.

Classification Algorithms

- Naive Bayes
- Logistic Regression
- Support Vector Machines (SVM)
- Decision Trees/Random Forests
- K-nearest neighbors
- Hidden Markov Models (HMM)
- Conditional Random Fields (CRF)
- Structural SVM

Probability

- Lots of methods in the digital humanities/machine learning are *probabilistic*:
    - clustering, topic models
    - classification

Probability distributions

Normal

Gamma

Poisson

Geometric

Exponential

Multinomial

Bernoulli

Beta

Binomial

Uniform

Dirichlet

Probability distributions

Normal

Gamma

Poisson

Geometric

Exponential

Multinomial

Bernoulli

Beta

Binomial

Uniform

Dirichlet

Random variable

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

$$X \in \{1, 2, 3, 4, 5, 6\}$$

$$X \in \{the, a, dog, cat, runs, to, store\}$$

$$P(X = x)$$

Probability that the random variable X takes the value x (e.g., 1)

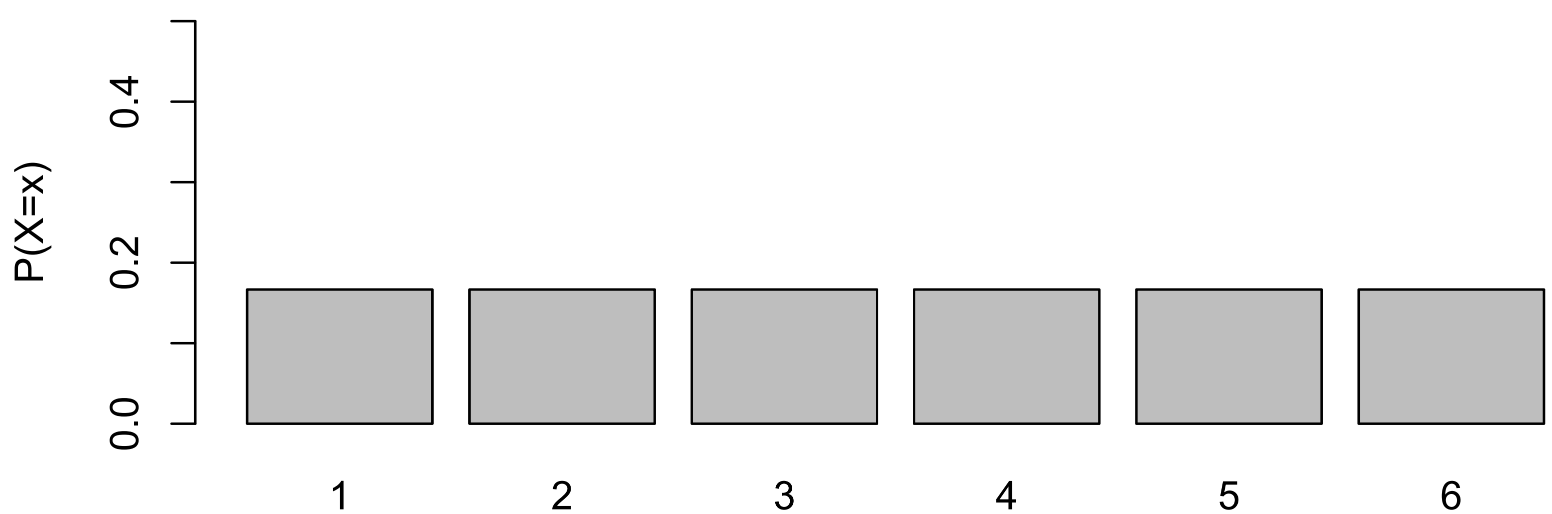
$$X \in \{1, 2, 3, 4, 5, 6\}$$

Two conditions:

1. Between 0 and 1: $0 \leq P(X = x) \leq 1$
2. Sum of all probabilities = 1 $\sum_x P(X = x) = 1$

Fair dice

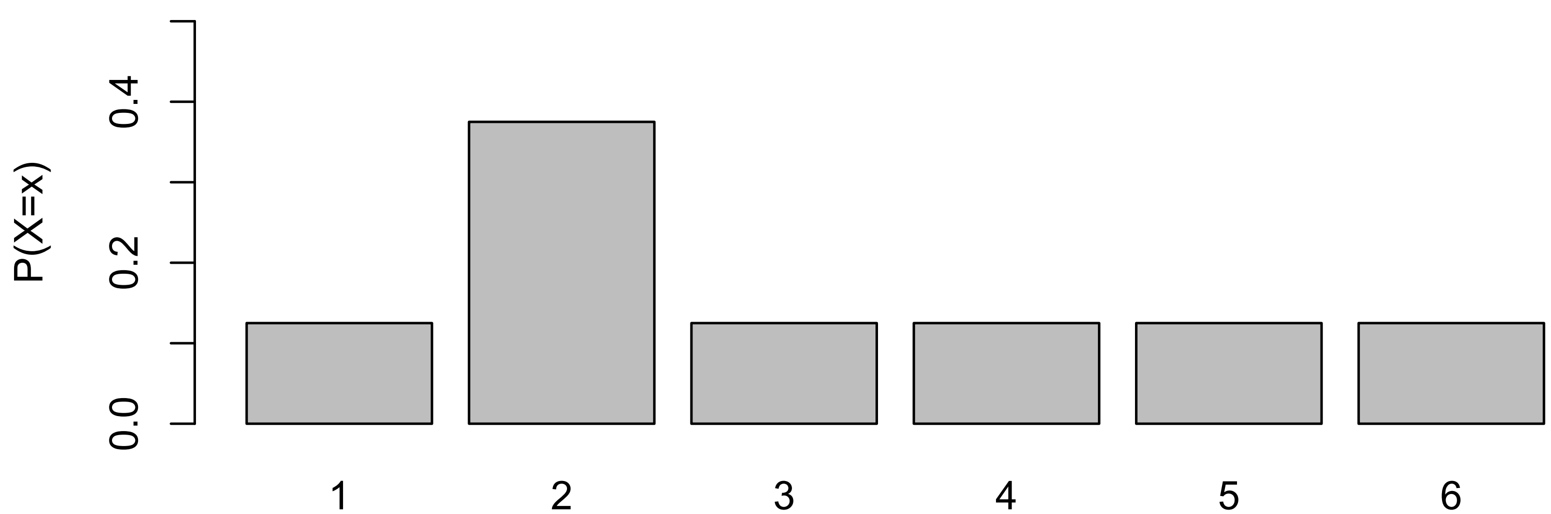
$$X \in \{1, 2, 3, 4, 5, 6\}$$



#mlch

Weighted dice

$$X \in \{1, 2, 3, 4, 5, 6\}$$



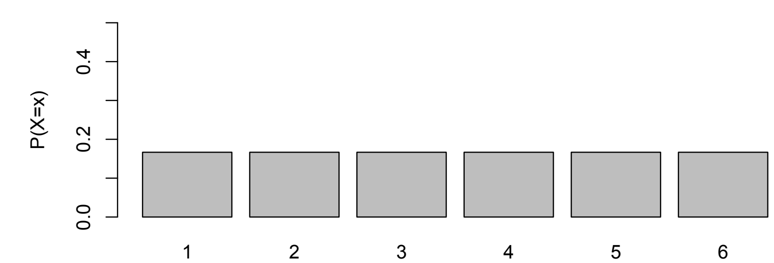
Parameter estimation

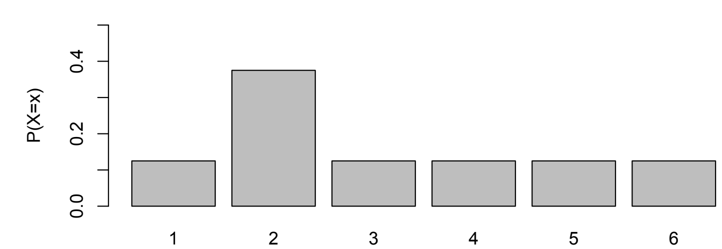
$$X \in \{1, 2, 3, 4, 5, 6\}$$

Data = 4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

We want to *estimate* the probability distribution that generated the data we see.

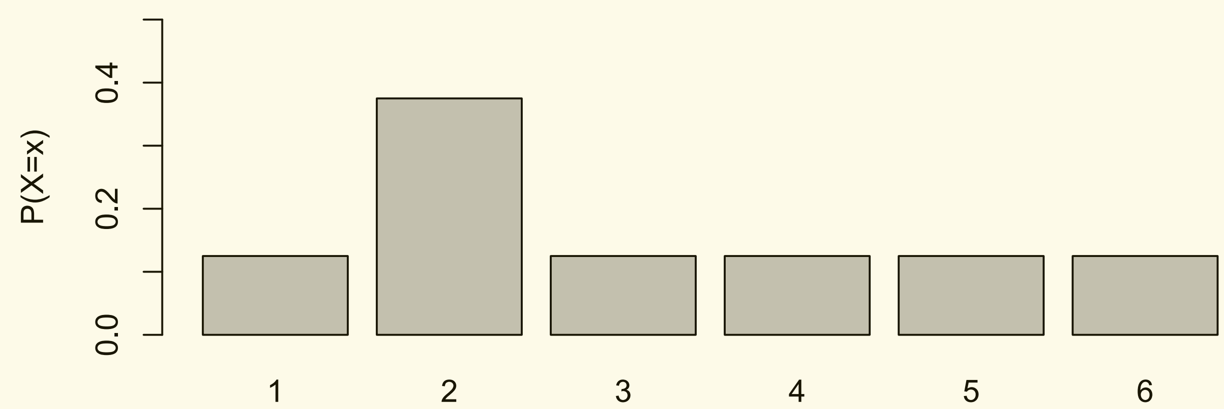
?

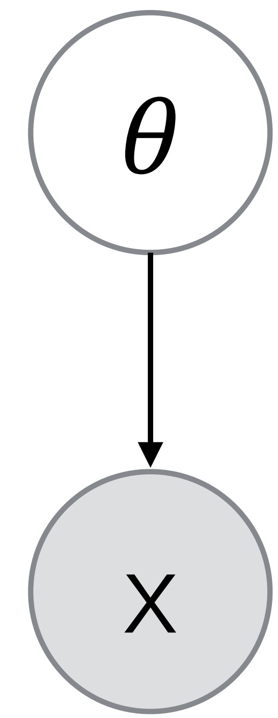




Generative story

$\theta =$



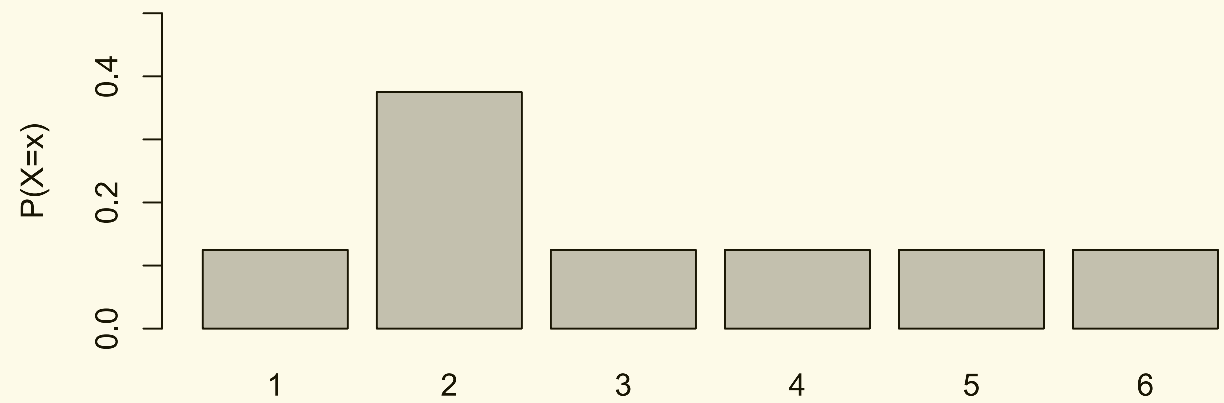


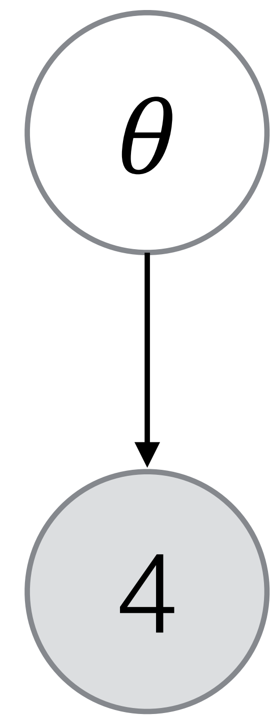
Data we see:

4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

Generative story

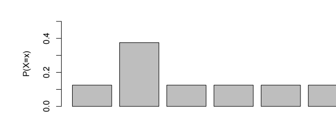
$\theta =$





Data we see:

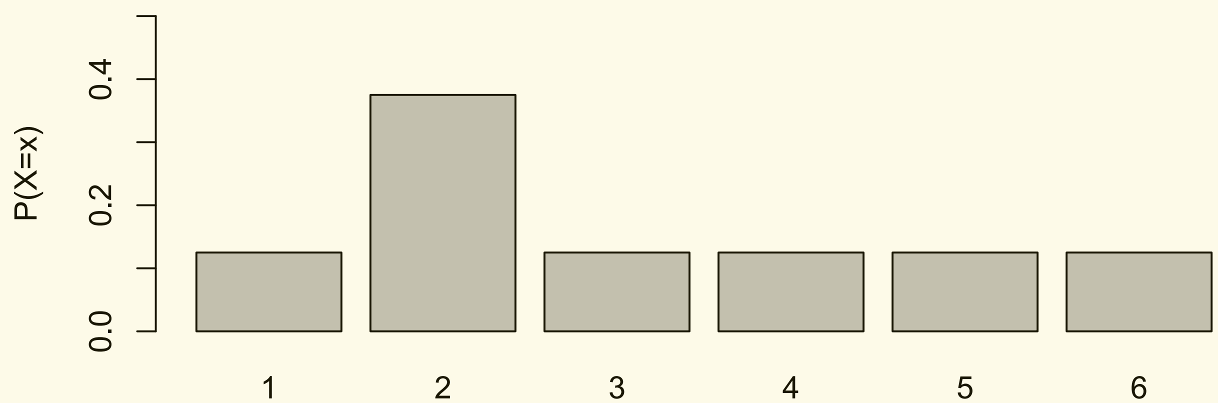
4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

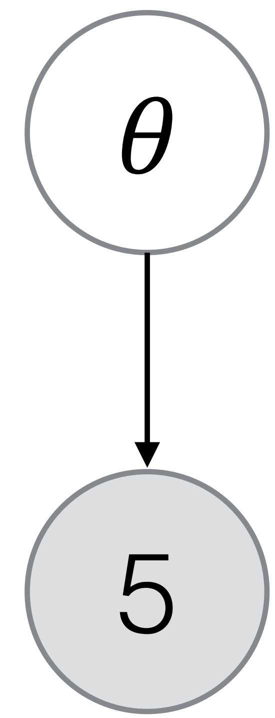
$P(X=4|\theta=$  $) = .125$

#mlch

Generative story

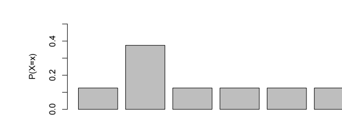
$\theta =$





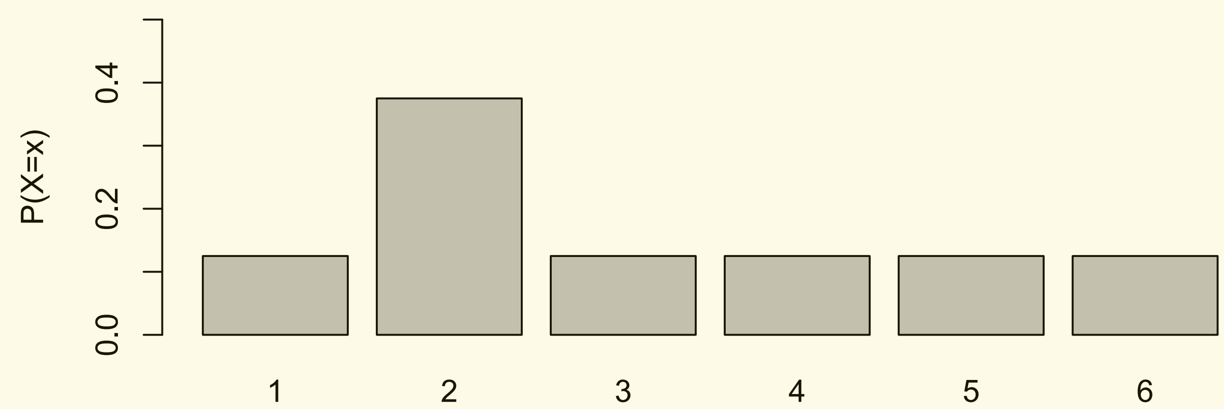
Data we see:

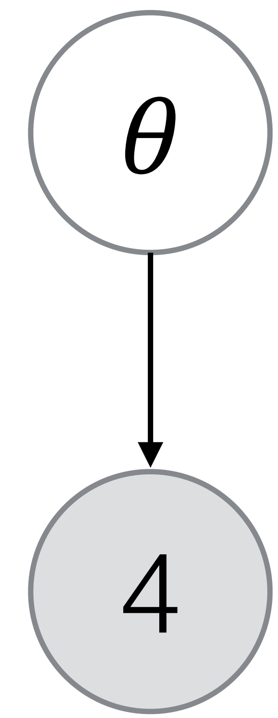
4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

$P(X=5|\theta=$  $) = .125$

Generative story

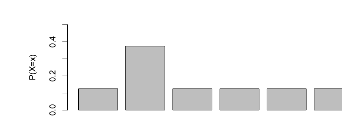
$\theta =$





Data we see:

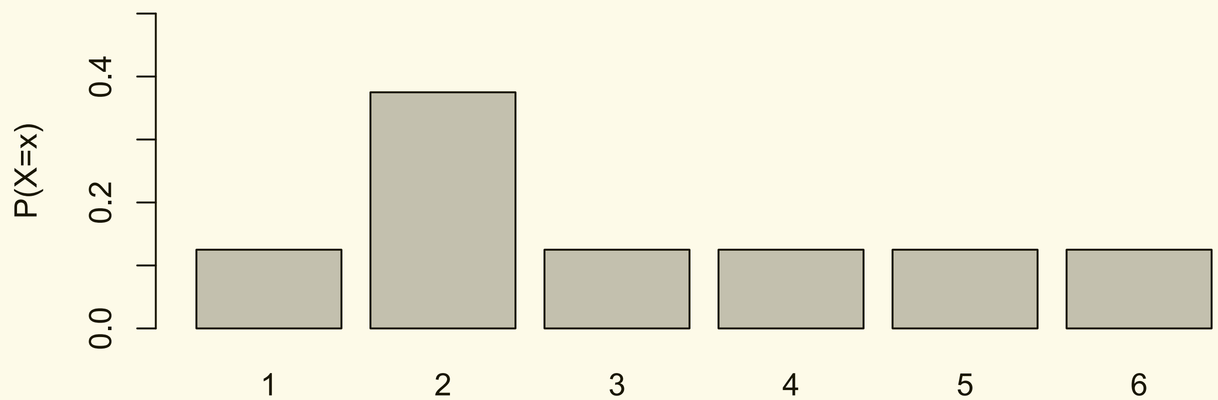
4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

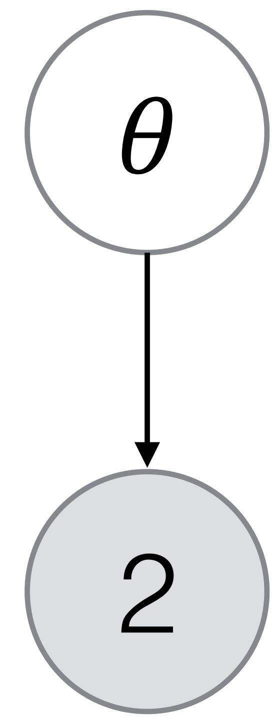
$P(X=4|\theta=$  $) = .125$

#mlch

Generative story

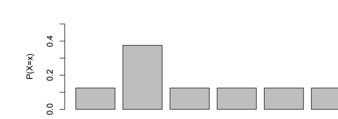
$\theta =$



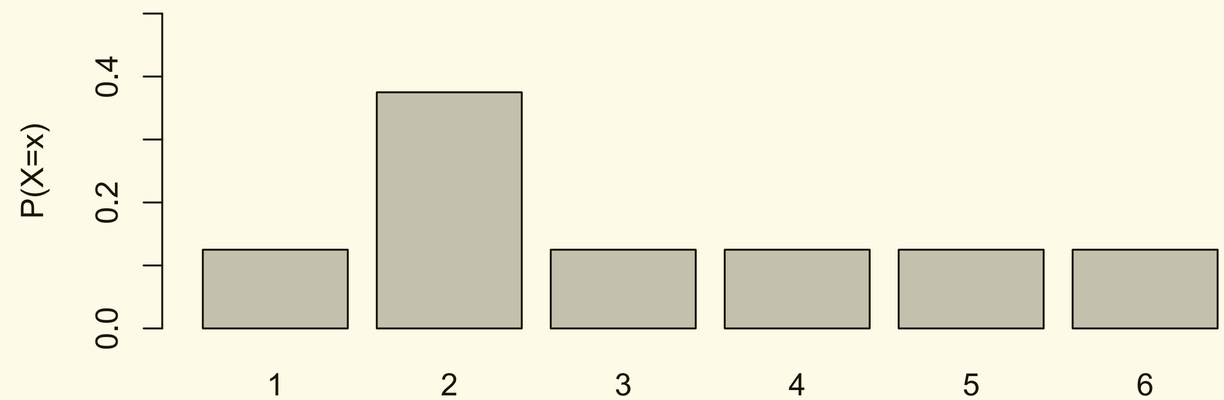


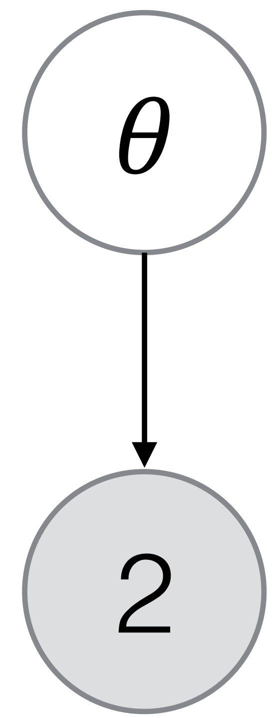
Data we see:

4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

$P(X=2|\theta=$  $) = .375$

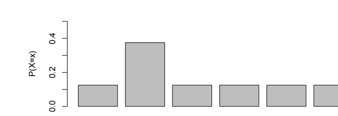
Generative story

$\theta =$ 



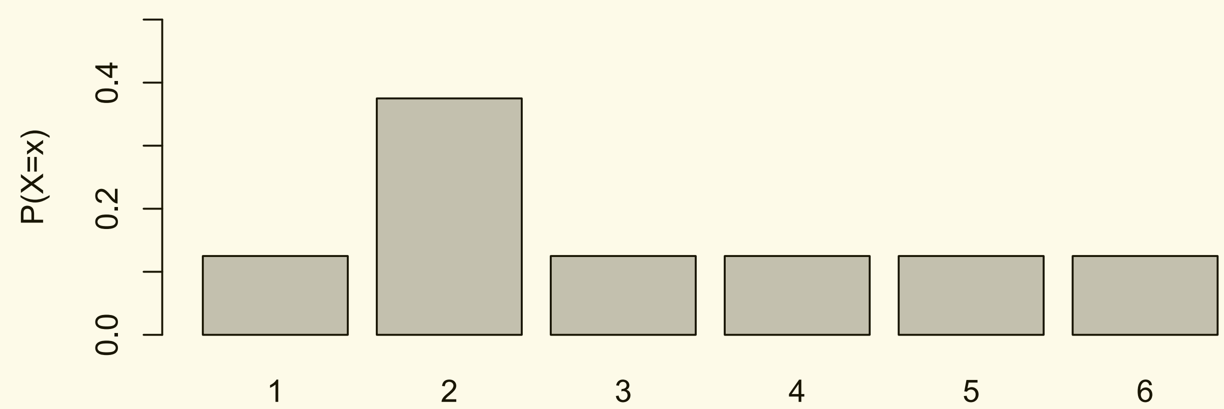
Data we see:

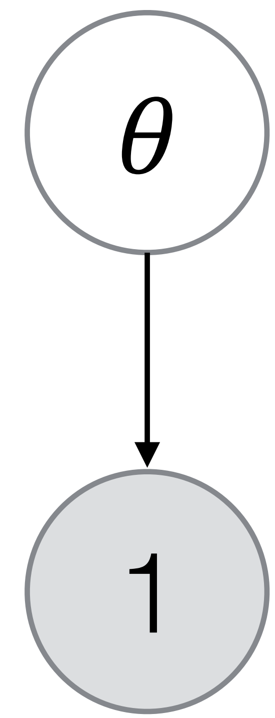
4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

$P(X=2|\theta=$  $) = .375$

#mlch

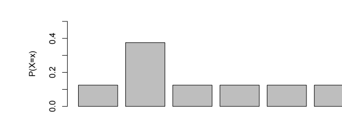
Generative story

$\theta =$ 



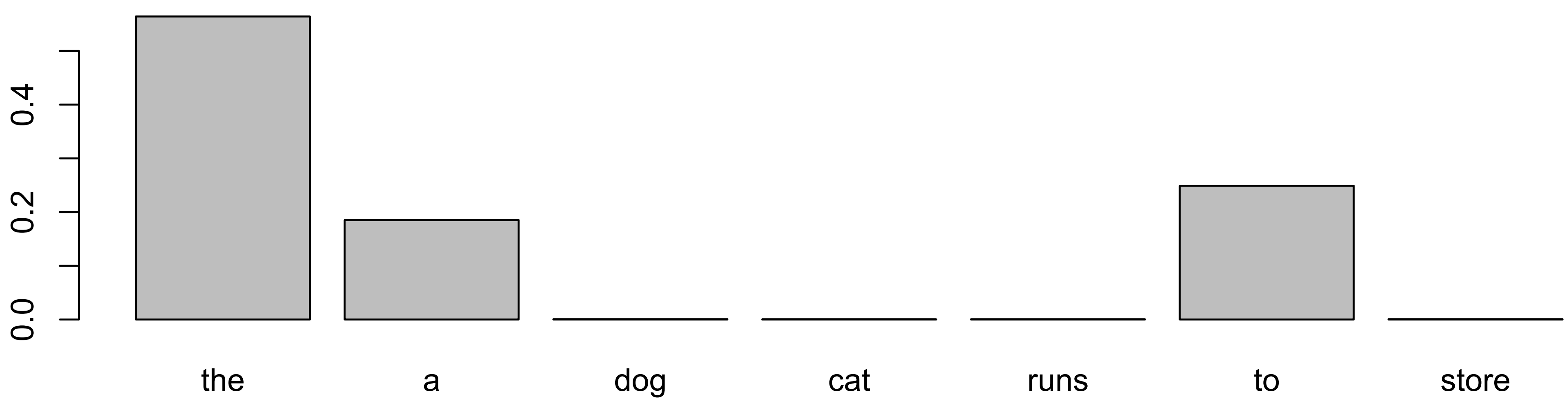
Data we see:

4,5,4,2,2,1,2,6,3,2,2,2,1,4,2

$P(X=1|\theta=$  $) = .125$

Unigram probability

$$X \in \{the, a, dog, cat, runs, to, store\}$$



How do we calculate this?

ined hopes of being admitted to a sight of the young ladies, of whose beauty he had heard much; but
ly the father. The ladies were somewhat more fortunate, for they had the advantage of ascertaining fro
window that he wore a blue coat, and rode a black horse. An invitation to dinner was soon afterwards
ched; and already had Mrs. Bennet planned the courses that were to do credit to her housekeeping, w
wer arrived which deferred it all. Mr. Bingley was obliged to be in town the following day, and, conseq
to accept the honour of their invitation, etc. Mrs. Bennet was quite disconcerted. She could not imagi
usiness he could have in town so soon after his arrival in Hertfordshire; and she began to fear that he
ays flying about from one place to another, and never settled at Netherfield as he ought to be. Lady Lu
d her fears a little by starting the idea of his being gone to London only to get a large party for the ball
soon followed that Mr. Bingley was to bring twelve ladies and seven gentlemen with him to the assemb
ls grieved over such a number of ladies, but were comforted the day before the ball by hearing, that i
ve he brought only six with him from London, his five sisters and a cousin. And when the party entered
bly room it consisted of only five … , the husband of the eldest, and
r young man. Mr. Bingley was g… d a pleasant countenance, and
cted manners. His sisters were f… ion. His brother-in-law, Mr. Hurs
looked the gentleman; but his f… of the room by his fine, tall per
ome features, noble mien, and t… on within five minutes after his
ce, of his having ten thousand a year. The gentlemen pronounced him to be a fine figure of a man, the
ed he was much handsomer than Mr. Bingley, and he was looked at with great admiration for about ha
g, till his manners gave a disgust which turned the tide of his popularity; for he was discovered to be p
bove his company, and above being pleased; and not all his large estate in Derbyshire could then sav
aving a most forbidding, disagreeable countenance, and being unworthy to be compared with his frien
gley had soon made himself acquainted with all the principal people in the room; he was lively and
rved, danced every dance, was angry that the ball closed so early, and talked of giving one himself at
field. Such amiable qualities must speak for themselves. What a contrast between him and his friend!
danced only once with Mrs. Hurst and once with Miss Bingley, declined being introduced to any other
ent the rest of the evening in walking about the room, speaking occasionally to one of his own party. H
ter was decided. He was the proudest, most disagreeable man in the world, and everybody hoped th
never come there again. Amongst the most violent against him was Mrs. Bennet, whose dislike of his

P(X="the") = 28/536 = .052

#mlch

Conditional Probability

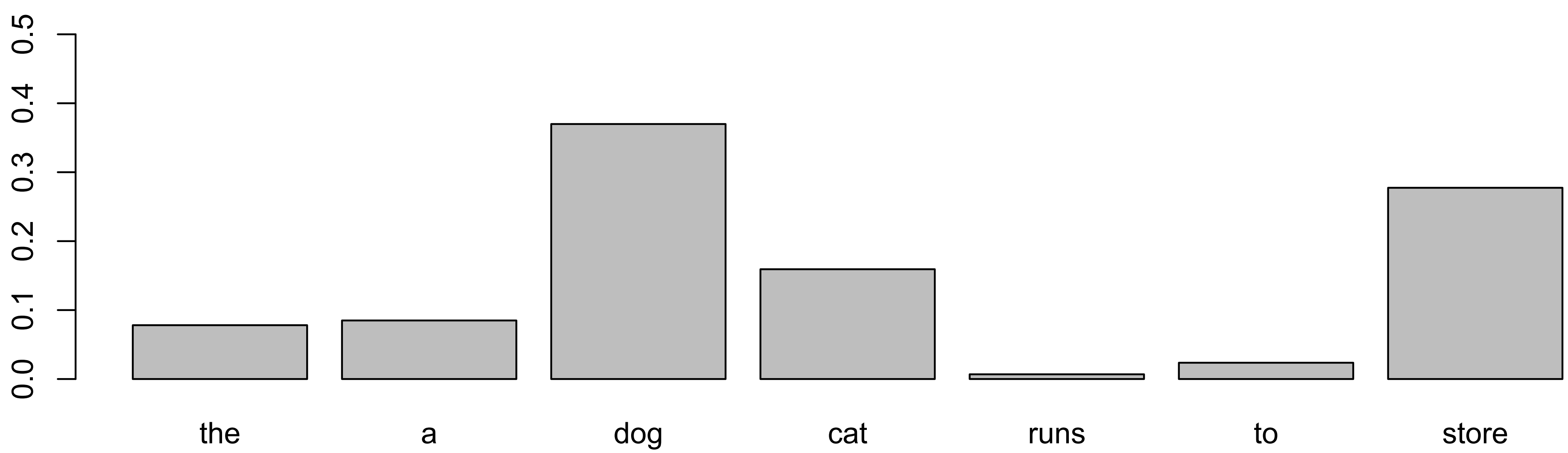
$$P(X = x|Y = y)$$

- Probability that one random variable takes a particular value *given* the fact that a different variable takes another

$$P(X_i = dog|X_{i-1} = the)$$

Conditional Probability

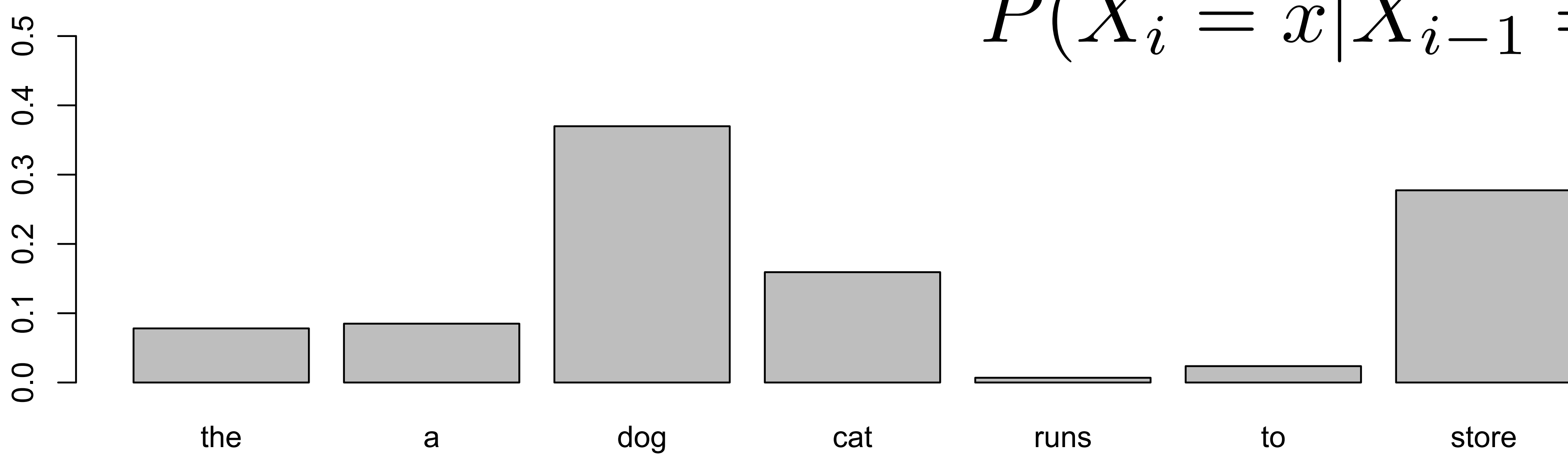
$$P(X_i = dog | X_{i-1} = the)$$



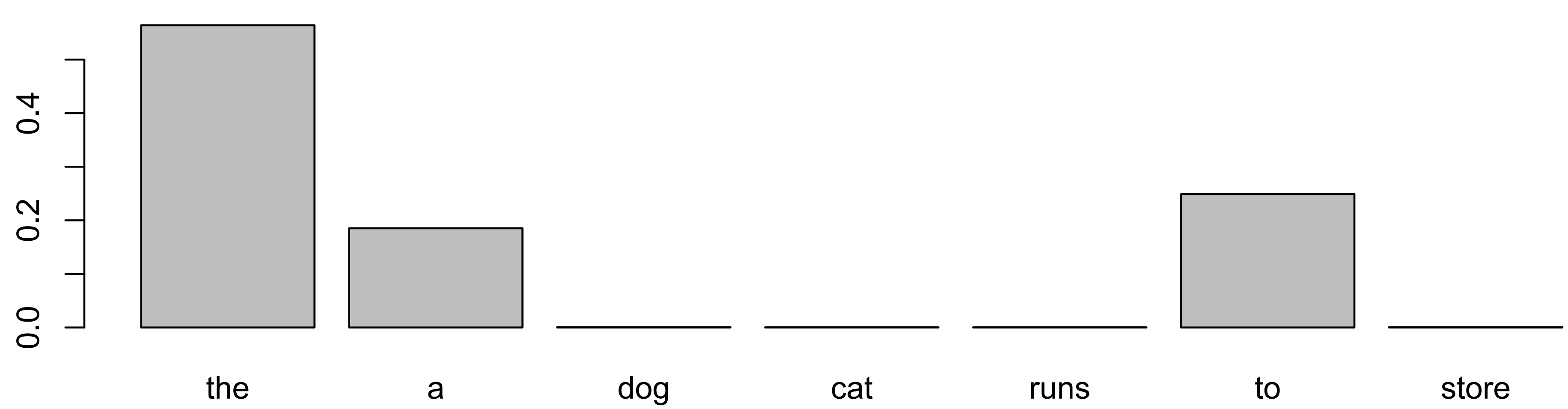
#mlch

Conditional Probability

$$P(X_i = x | X_{i-1} = the)$$



$$P(X_i = x)$$



#mlch

ined hopes of being admitted to a sight of the **young** ladies, of whose beauty he had heard much; bu
ly the **father**. The **ladies** were somewhat more fortunate, for they had the **advantage** of ascertaining f
window that he wore a blue coat, and rode a black horse. An invitation to dinner was soon afterwards
ched; and already had Mrs. Bennet planned the **courses** that were to do credit to her housekeeping, v
wer arrived which deferred it all. Mr. Bingley was obliged to be in town the **following** day, and,
quently, unable to accept the **honour** of their invitation, etc. Mrs. Bennet was quite disconcerted. She
gine what business he could have in town so soon after his arrival in Hertfordshire; and she began to
might be always flying about from one place to another, and never settled at Netherfield as he ought
ucas quieted her fears a little by starting the **idea** of his being gone to London only to get a large part
l; and a report soon followed that Mr. Bingley was to bring twelve ladies and seven gentlemen with him
sembly. The **girls** grieved over such a number of ladies, but were comforted the **day** before the **ball** by
g, that instead of twelve he brought only six with him from London--his five sisters and a cousin. And w
ty entered the **assembly** ro ... gley, his two sisters, the husb
eldest, and another young m ... manlike; he had a pleasant
nance, and easy, unaffected ... n air of decided fashion. His
-in-law, Mr. Hurst, merely loo ... soon drew the **attention** of th
y his fine, tall person, hands ... ch was in general circulation
ive minutes after his entranc ... **tlemen** pronounced him to b
ure of a man, the **ladies** declared he was much handsomer than Mr. Bingley, and he was looked at wi
dmiration for about half the evening, till his manners gave a disgust which turned the **tide** of his popu
was discovered to be proud; to be above his company, and above being pleased; and not all his large
in Derbyshire could then save him from having a most forbidding, disagreeable countenance, and bei
hy to be compared with his friend. Mr. Bingley had soon made himself acquainted with all the **princip**
in the **room**; he was lively and unreserved, danced every dance, was angry that the **ball** closed so e
ked of giving one himself at Netherfield. Such amiable qualities must speak for themselves. What a co
en him and his friend! Mr. Darcy danced only once with Mrs. Hurst and once with Miss Bingley, decline
ntroduced to any other lady, and spent the **rest** of the **evening** in walking about the **room**, speaking
onally to one of his own party. His character was decided. He was the **proudest**, most disagreeable m
rld, and everybody hoped that he would never come there again. Amongst the **most** violent against hi
ennet, whose dislike of his general behaviour was sharpened into particular resentment by his having

$P(X_i=\text{“room”} \mid X_{i-1}=\text{“the”}) = 2/28 = .071$

Conditional Probability

$$P(X = vampire) \text{ vs. } P(X = vampire|Y = horror)$$

$$P(X = manners|Y = austen) \text{ vs. } P(X = whale|Y = austen)$$

$$P(X = manners|Y = austen) \text{ vs. } P(X = manners|Y = dickens)$$

Our first classifier

"Mr. Collins was not a sensible man"

Austen		Dickens	
P(X=Mr. | Y=Austen)	0.0084	P(X=Mr. | Y=Dickens)	0.00421
P(X=Collins | Y=Austen)	0.00036	P(X=Collins | Y=Dickens)	0.000016
P(X=was | Y=Austen)	0.01475	P(X=was | Y=Dickens)	0.015043
P(X=not | Y=Austen)	0.01145	P(X=not | Y=Dickens)	0.00547
P(X=a | Y=Austen)	0.01591	P(X=a | Y=Dickens)	0.02156
P(X=sensible | Y=Austen)	0.00025	P(X=sensible | Y=Dickens)	0.00005
P(X=man | Y=Austen)	0.00121	P(X=man | Y=Dickens)	0.001707

Our first classifier

"Mr. Collins was not a sensible man"

P(X = "Mr. Collins was not a sensible man" | Y = Austen)

= P("Mr" | Austen) × P("Collins" | Austen) ×
P("was" | Austen) × P("not" | Austen) …
= 0.000000022507322 (**≈ 2.3 × 10^{-8}**)

P(X = "Mr. Collins was not a sensible man" | Y = Dickens)

P("Mr" | Dickens) × P("Collins" | Dickens) ×
P("was" | Dickens) × P("not" | Dickens) …
= 0.000000002078906 (**≈ 2.1 × 10^{-9}**)

Bayes' Rule

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

Bayes' Rule

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

Likelihood of the data
given that Y=y

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

Posterior belief that Y=y given that X=x

#mlch

Bayes' Rule

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

Likelihood of the data
given that Y=y

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

Posterior belief that Y=y given that X=x

Bayes' Rule

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

Likelihood of the data
given that Y=y

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

Posterior belief that Y=y given that X=x

Bayes' Rule

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

Likelihood of "Mr. Collins
was not a sensible man"
given that Y= Austen

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

Posterior belief that Y=Austen given that
X="Mr. Collins was not a sensible man"

This sum ranges over
y=Austen + y=Dickens
(so that it sums to 1)

#mlch

Naive Bayes Classifier

$$\frac{P(Y = Austen)P(X = \text{``}Mr...\text{''}|Y = Austen)}{P(Y = Austen)P(X = \text{``}Mr...\text{''}|Y = Austen) + P(Y = Dickens)P(X = \text{``}Mr...\text{''}|Y = Dickens)}$$

Let's say P(Y=Austen) = P(Y=Dickens) = 0.5
(i.e., both are equally likely a priori)

$$= \frac{0.5 \times (2.3 \times 10^{-8})}{0.5 \times (2.3 \times 10^{-8}) + 0.5 \times (2.1 \times 10^{-9})}$$

$$P(Y = Austen|X = \text{``}Mr...\text{''}) = 91.5\%$$

$$P(Y = Dickens|X = \text{``}Mr...\text{''}) = 8.5\%$$

Taxicab Problem

“A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

- 85% of the ca … lue.
- A witness ide … d the reliability of the witness u … ted on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

“Base rate fallacy”
Don’t ignore prior information!

What is the probability that the cab involved in the accident was Blue rather than Green knowing that this witness identified it as Blue?”

(Tversky & Kahneman 1981)

Prior Belief

- Now let's assume that Dickens published 1000 times more books than Austen.
  - P(Y= Austen) = 0.000999
  - P(Y = Dickens) = 0.999001

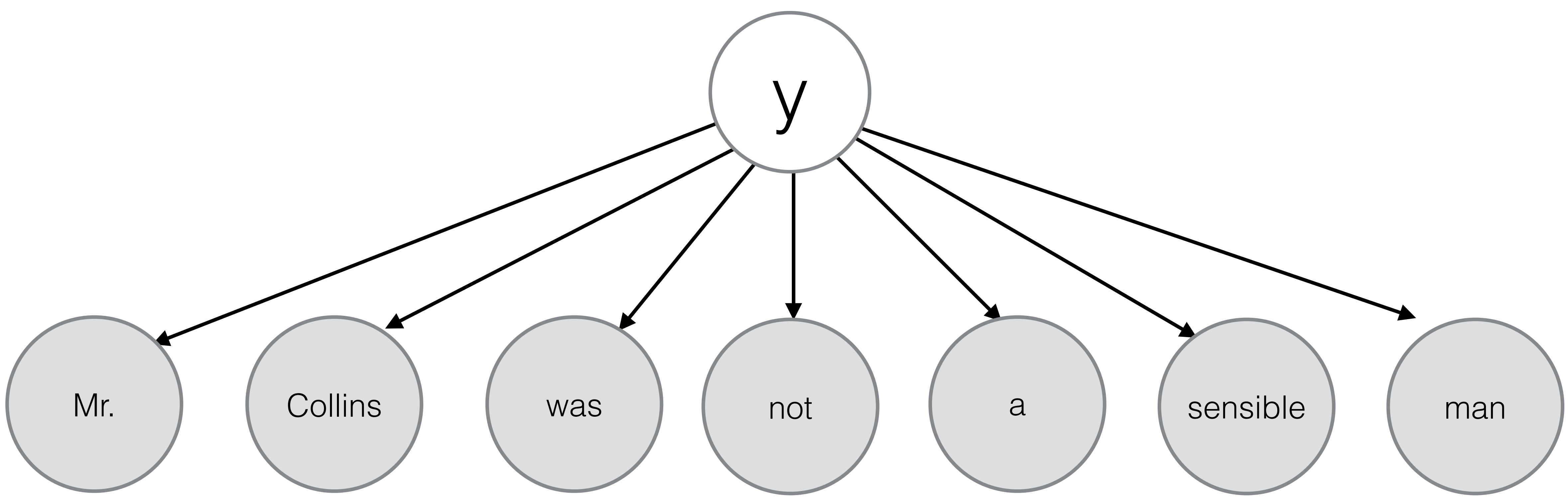
$$\frac{0.000999 \times (2.3 \times 10^{-8})}{0.000999 \times (2.3 \times 10^{-8}) + 0.999001 \times (2.1 \times 10^{-9})}$$

$$P(Y = Austen|X) = 0.011$$

$$P(Y = Dickens|X) = 0.989$$

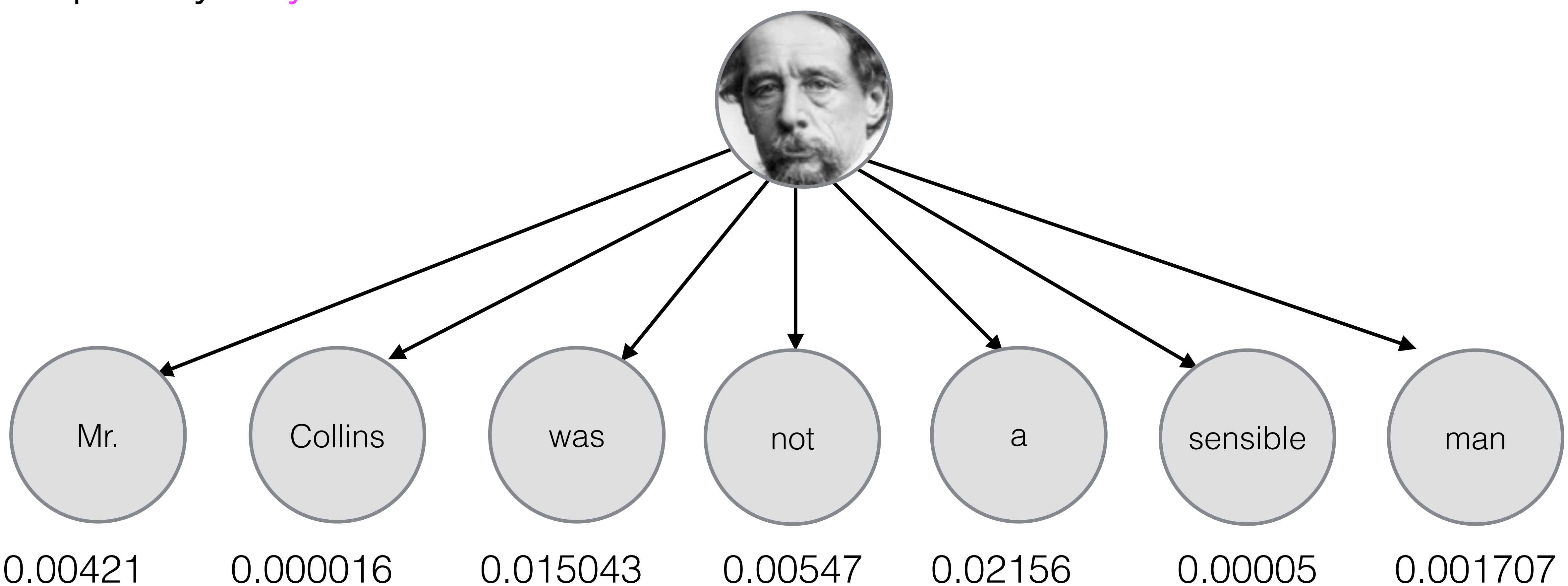
Naive Bayes

- Find the value of y (e.g., author) for which the probability of the x (e.g., the words) that we see is highest, along with the prior frequency of y.
- All x's are independent and contribute equally to finding the best y (the "naive" in Naive Bayes)



Naive Bayes

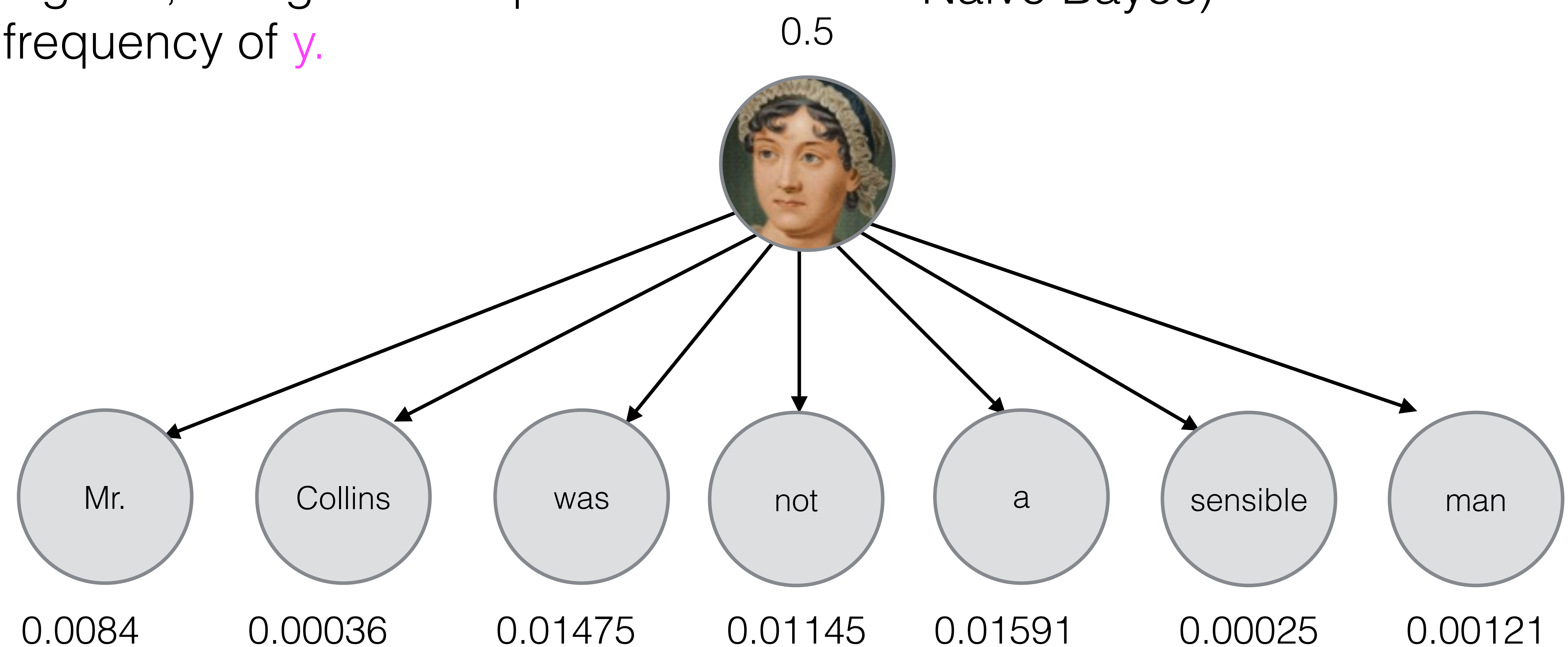
- Find the value of y (e.g., author) for which the probability of the x (e.g., the words) that we see is highest, along with the prior frequency of y.
- All x's are independent and contribute equally to finding the best y (the "naive" in Naive Bayes)



#mlch

Naive Bayes

- Find the value of y (e.g., author) for which the probability of the x (e.g., the words) that we see is highest, along with the prior frequency of y.
- All x's are independent and contribute equally to finding the best y (the "naive" in Naive Bayes)



Parameters

$P(X = x \mid Y= \text{Austen})$

value	prob
Mr.	0.0084
Collins	0.00036
was	0.01475
not	0.01145
a	0.01591
sensible	0.00025
man	0.00121
dog	0.003
chimney	0.004

…

$P(X = x \mid Y= \text{Dickens})$

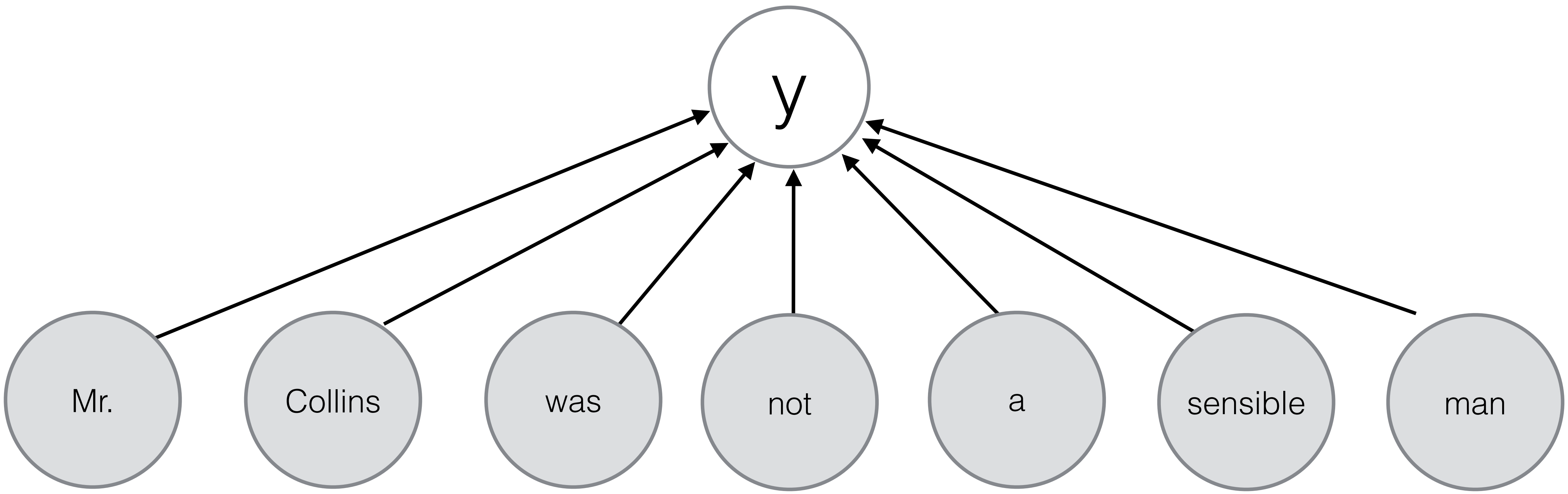
value	prob
Mr.	0.00421
Collins	0.000016
was	0.015043
not	0.00547
a	0.02156
sensible	0.00005
man	0.001707
dog	0.002
chimney	0.008

…

$P(Y = y)$

value	prob
Dickens	0.50
Austen	0.50

Logistic Regression



$$P(Y = y|X, \beta) = \frac{\exp\left(\sum_i^F \beta_{y,i} x_i\right)}{\sum_{y'} \exp\left(\sum_i^F \beta_{y',i} x_i\right)}$$

Logistic Regression

$\beta_{\text{Austen}} =$

i	feat	value
1	Mr.	1.4
2	Collins	15.7
3	was	0.01
4	a	-0.003
5	sensible	7.8
6	man	1.3
7	dog	-1.3
8	chimney	-10.3

x =

i	feat	value
1	Mr.	1
2	Collins	1
3	was	1
4	a	1
5	sensible	1
6	man	1
7	dog	0
8	chimney	0

$$P(Y = y | X, \beta) = \frac{\exp\left(\sum_i^F \beta_{y,i} x_i\right)}{\sum_{y'} \exp\left(\sum_i^F \beta_{y',i} x_i\right)}$$

Logistic Regression

- Find the value of β that maximizes P(Y=y | X = x, β) where we know the value of y given a particular x (i.e., in training data).

- Likelihood: $$L(\beta) = \prod_{\{x,y\}} P(Y = y|X = x, \beta)$$

- Log Likelihood: $$\ell(\beta) = \sum_{\{x,y\}} \log P(Y = y|X = x, \beta)$$

Overfitting

- Memorizing patterns in the training data too well → perform worse on data you don't train on.

- e.g., if we see Collins only in Austen books in the training data, what happens if we see Collins in a new book we're predicting?

Regularization

- Penalize parameters that are very big (i.e., that are far away from 0).

i	feat	β
1	Mr.	1.4
2	Collins	18403.0
3	was	0.01
4	a	-0.003
5	sensible	7.8
6	man	1.3
7	dog	-1.3
8	chimney	-10.3

$$\arg\max_{\beta} \sum_{\{x,y\}} \log P(Y = y | X = x, \beta) - \lambda \sum_{j} \beta_j^2$$

Regularization

i	feat	β
1	Mr.	1.4
2	Collins	18403
3	was	0.01
4	a	-0.003
5	sensible	7.8
6	man	1.3
7	dog	-1.3
8	chimney	-10.3

i	feat	β
1	Mr.	1.1
2	Collins	13.8
3	was	0.005
4	a	-0.0007
5	sensible	6.9
6	man	0.9
7	dog	-0.7
8	chimney	-8.3

$$\arg\max_{\beta} \sum_{\{x,y\}} \log P(Y = y | X = x, \beta) - \lambda \sum_{j} \beta_j^2$$

#mlch

Regularization

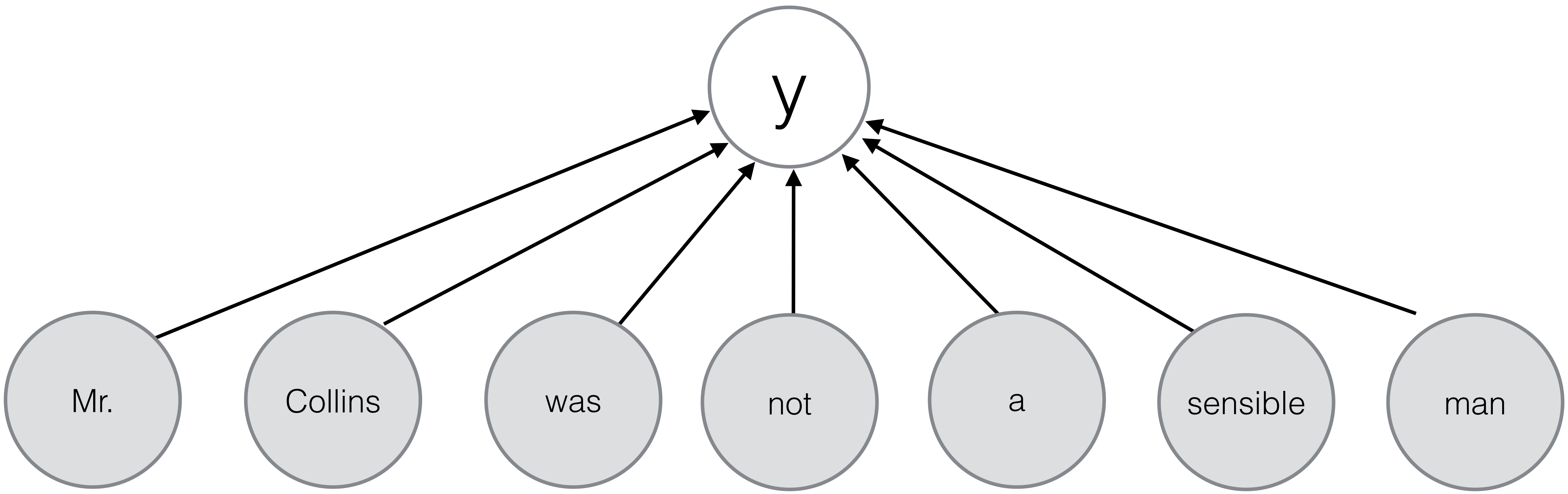
- L2 regularization encourages parameters to be close to 0.

$$\arg\max_{\beta} \sum_{\{x,y\}} \log P(Y = y|X = x, \beta) - \lambda \sum_{j} \beta_j^2$$

- L1 regularization also encourages them to be 0. (Sparsity)

$$\arg\max_{\beta} \sum_{\{x,y\}} \log P(Y = y|X = x, \beta) - \lambda \sum_{j} |\beta_j|$$

Logistic Regression

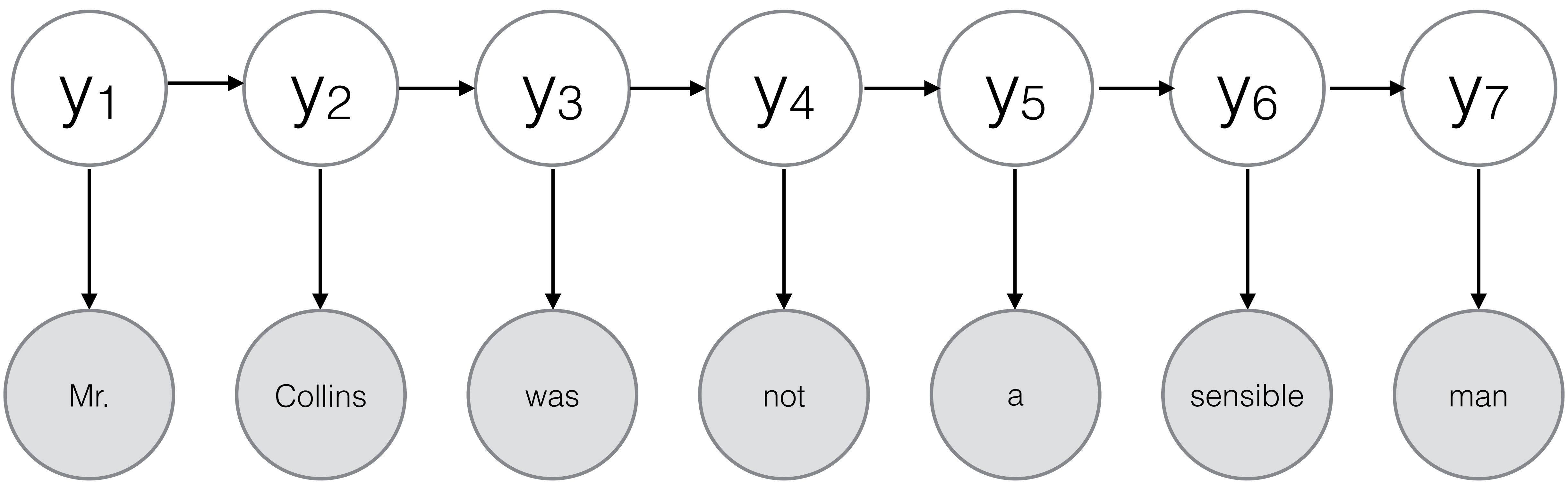


$$P(Y = y|X, \beta) = \frac{\exp\left(\sum_i^F \beta_{y,i} x_i\right)}{\sum_{y'} \exp\left(\sum_i^F \beta_{y',i} x_i\right)}$$

#mlch

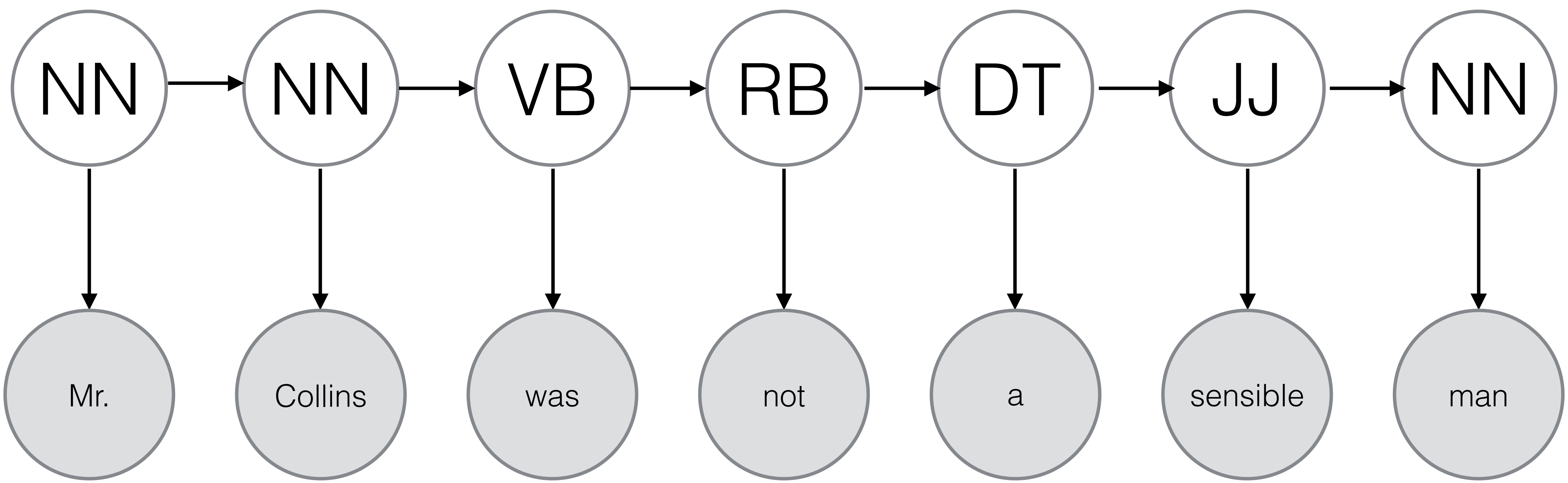
Hidden Markov Model

Generative model for predicting a sequence of variables.



Hidden Markov Model

Example: part of speech tagging



#mlch

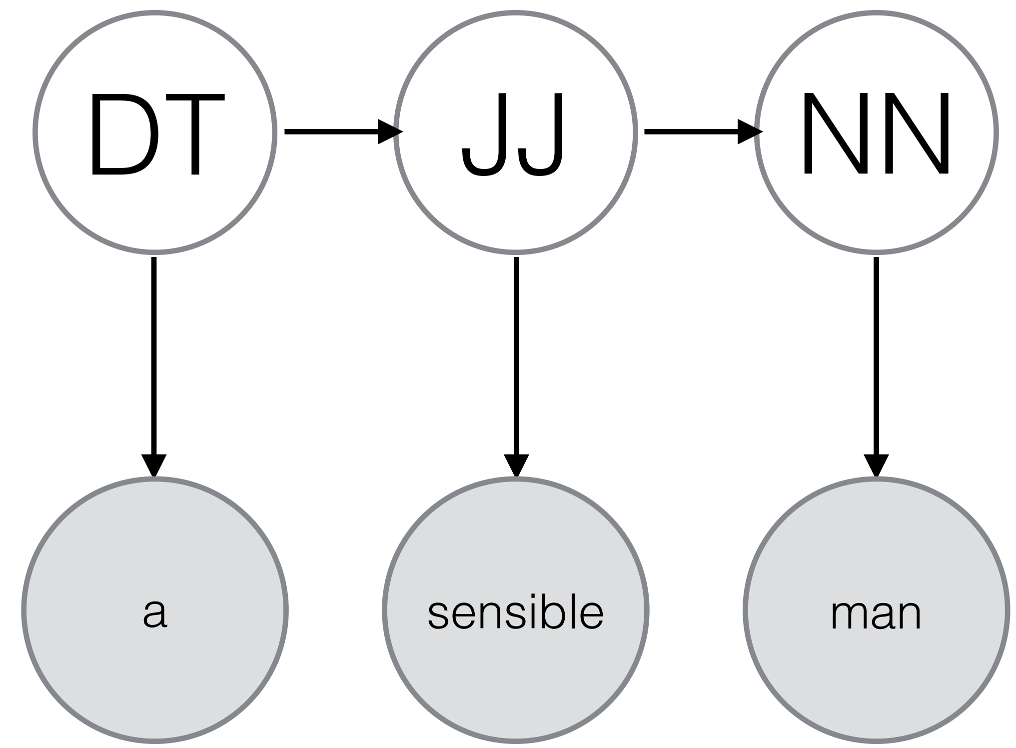
Hidden Markov Model

$P(X=x \mid y = DT)$

value	prob
a	0.37
the	0.33
an	0.17
sensible	0
dog	0

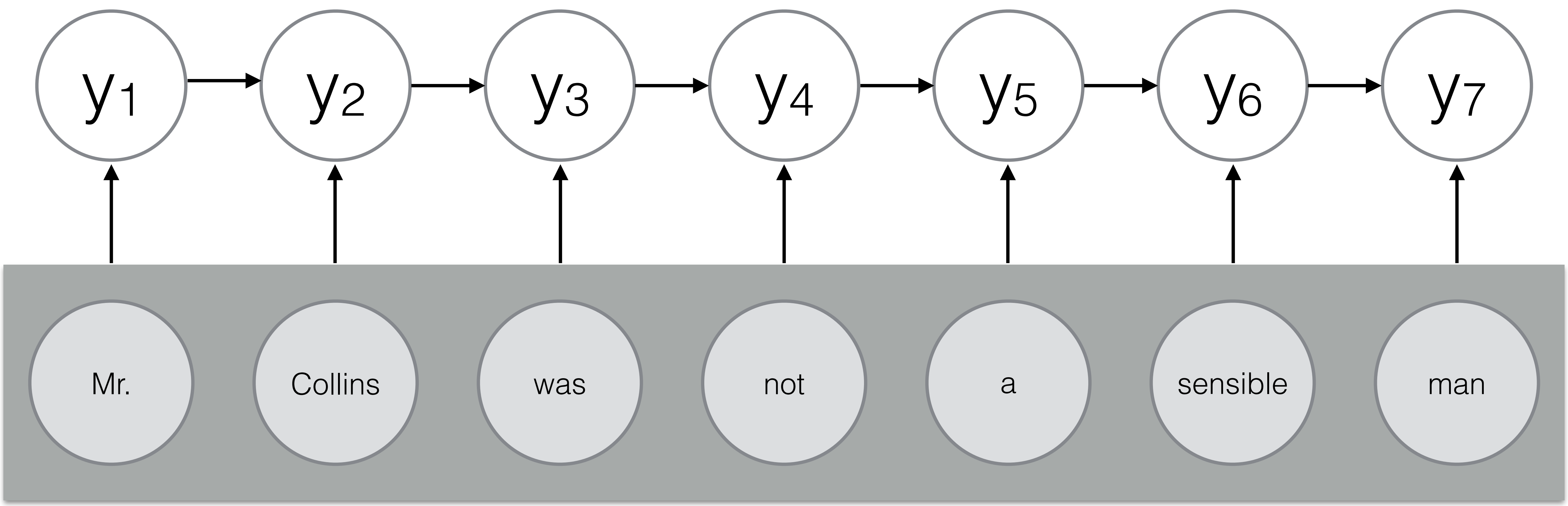
$P(Y_i = y \mid Y_{i-1} = DT)$

value	prob
NN	0.38
JJ	0.17
RB	0.15
DT	0



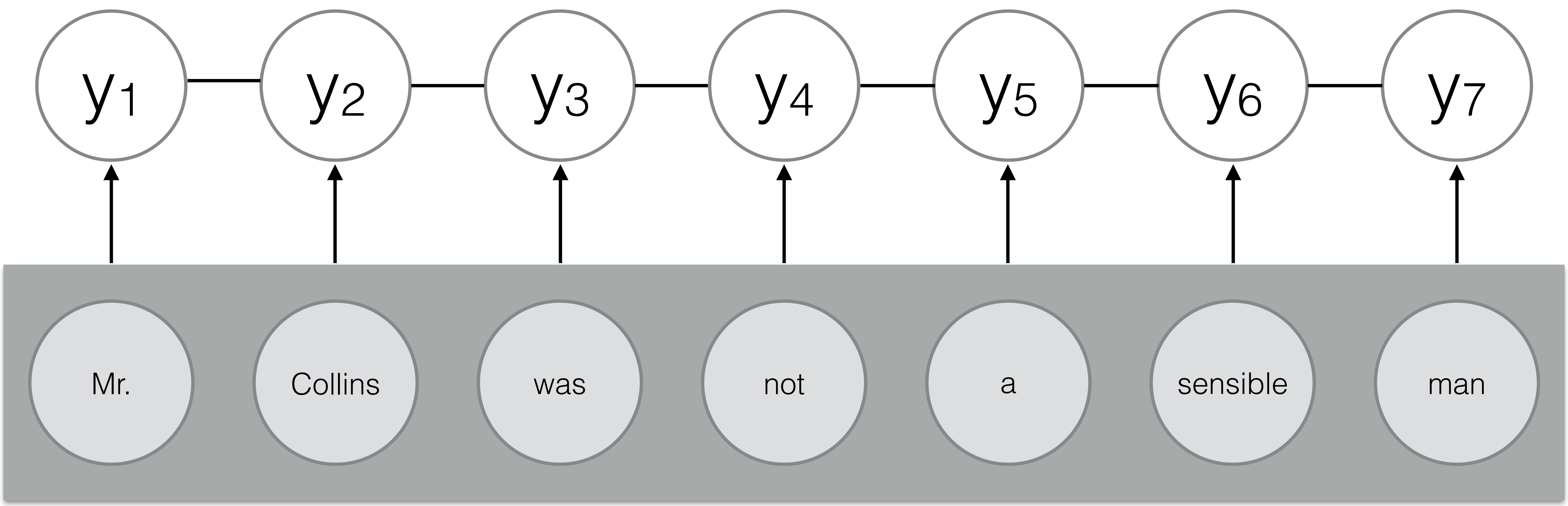
Maximum Entropy Markov Model

Discriminative model for predicting a sequence of variables.



Conditional Random Field

Discriminative model for predicting a sequence of variables.



#mlch

Rich features

Mr. Collins was not a sensible man

HMM

feature	val
word=Collins	1
word=the	0
word=a	0
word=not	0
word=sensible	0

MEMM/CRF

feature	val
word=Collins	1
word starts with capital letter	1
word is in list of known names	1
word ends in -ly	0

Try it yourself

- LightSide
  http://bit.ly/1hdKX0R

- (Google “LightSide Academic”)

Break!

Unsupervised Learning

- Unsupervised learning finds *interesting structure* in data.
    - clustering data into groups
    - discovering “factors”
    - discovering graph structure (6DFB)

Unsupervised Learning

- Matrix completion (e.g., user recommendations on Netflix, Amazon)

	Ann	Bob	Chris	David	Erik
Star Wars	5	5	4	5	3
Bridget Jones		4		4	1
Rocky	3		5		
Rambo		?		2	5

Unsupervised Learning

- Hierarchical clustering
- Flat clustering (K-means)
- Topic models

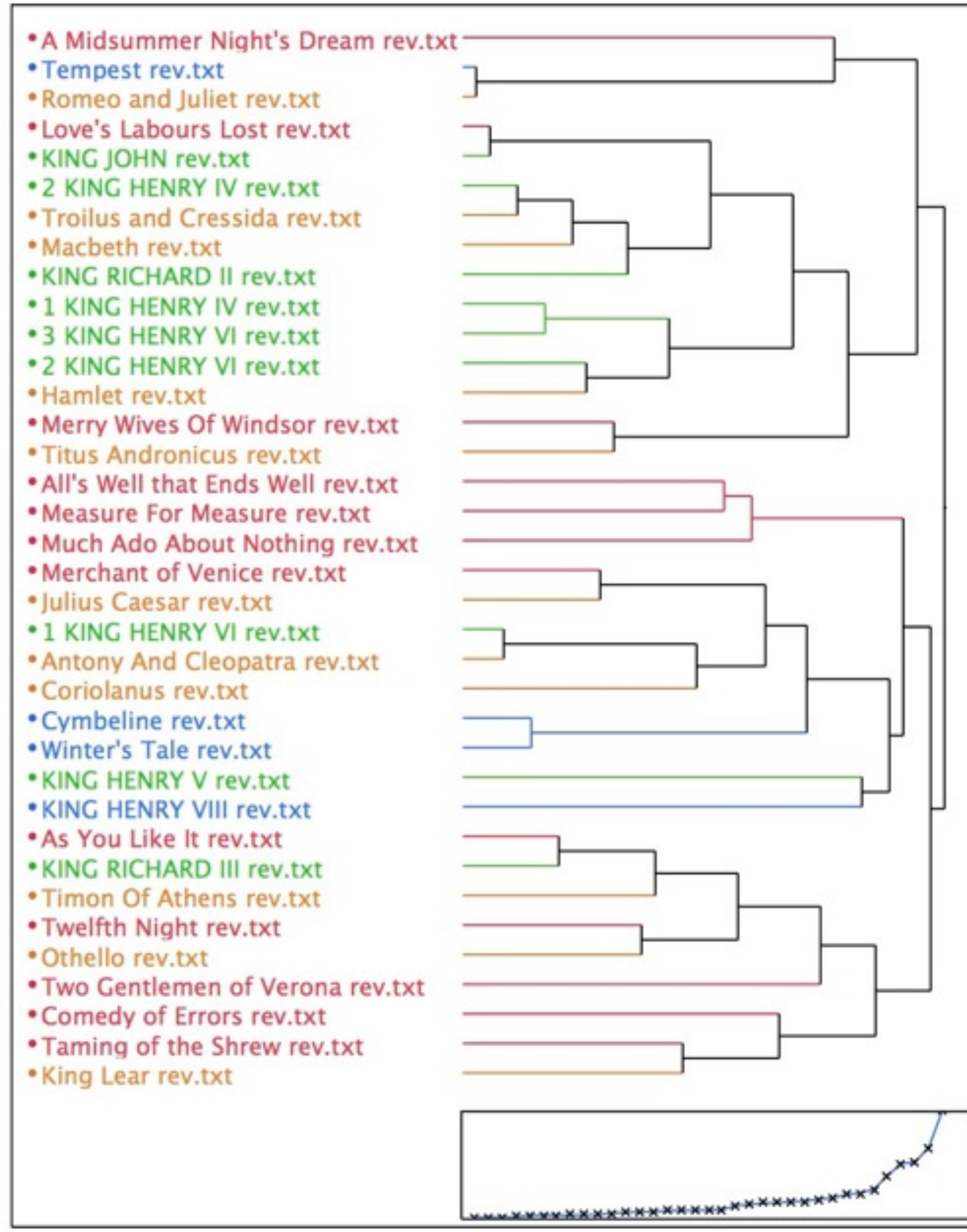
Hierarchical Clustering

- *Hierarchical* order among the elements being clustered
- Bottom-up = agglomerative clustering
- Top-down = divisive clustering

Dendrogram

Shakespeare's plays

Witmore (2009) http://winedarksea.org/?p=519



Bottom-up clustering

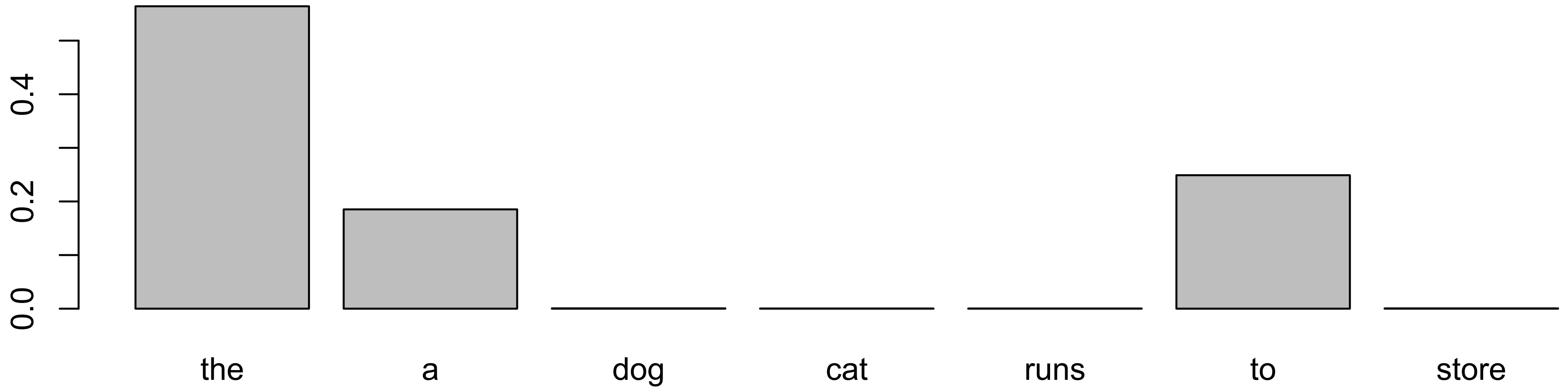
Given: a set $\mathcal{X} = \{x_1, \dots x_n\}$ of objects
a function sim: $\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$
for $i := 1$ **to** n **do**
$\quad c_i := \{x_i\}$ **end**
$C := \{c_1, \dots , c_n\}$
$j := n + 1$
while $C > 1$
$\quad (c_{n_1}, c_{n_2}) := \arg\max_{(c_u, c_v) \in C \times C} \mathrm{sim}(c_u, c_v)$
$\quad c_j = c_{n_1} \cup c_{n_2}$
$\quad C := C \setminus \{c_{n_1}, c_{n_2}\} \cup \{c_j\}$
$\quad j := j + 1$

Similarity

$$\mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$$

- What are you comparing?
- How do you quantify the similarity/difference of those things?

Probability



Unigram probability

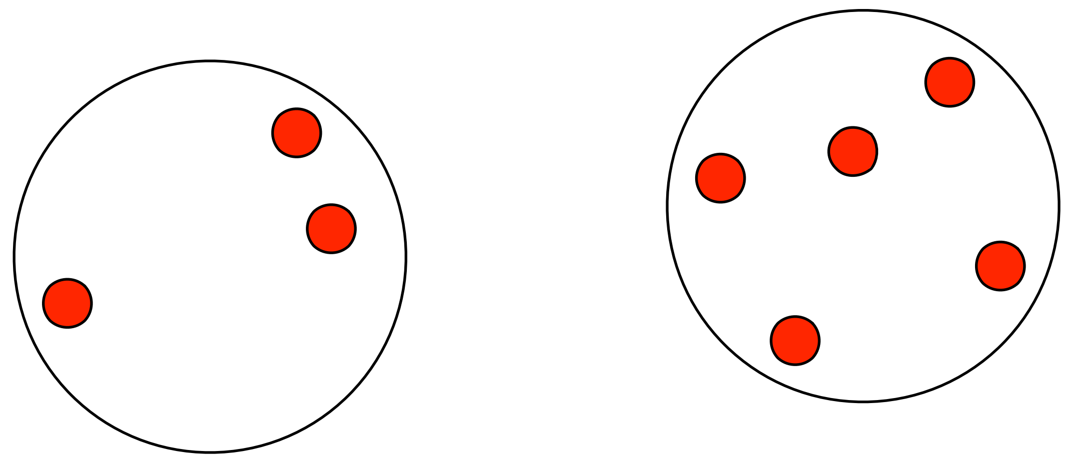
Similarity

$$\text{Euclidean} = \sqrt{\sum_{i}^{vocab} \left(P_i^{\text{Hamlet}} - P_i^{\text{Romeo}}\right)^2}$$

Cosine similarity, Jensen-Shannon divergence…

Cluster similarity

Cluster similarity



- Single link: two **most** similar elements
- Complete link: two **least** similar elements
- Group average: average of all members

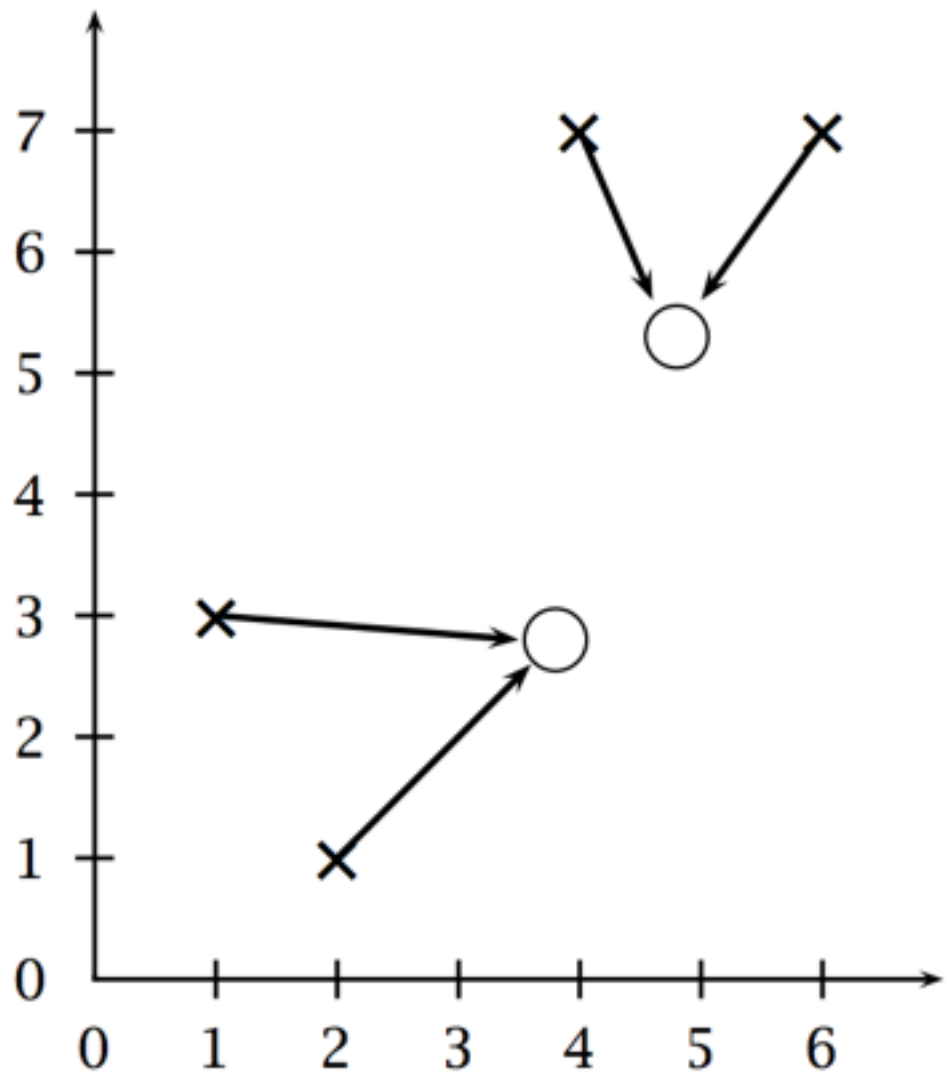
Flat Clustering

- Partitions the data into a set of *K* clusters

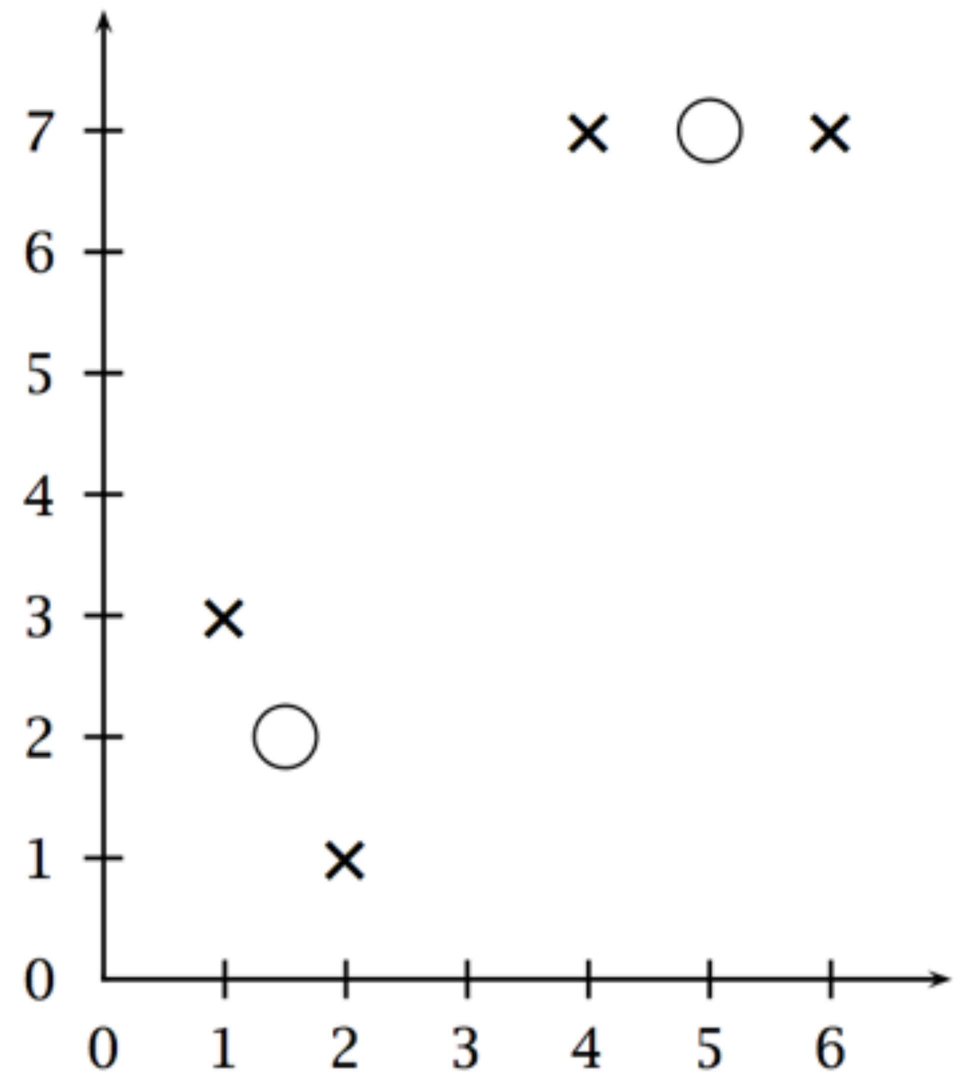
K-means

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

K-means



assignment

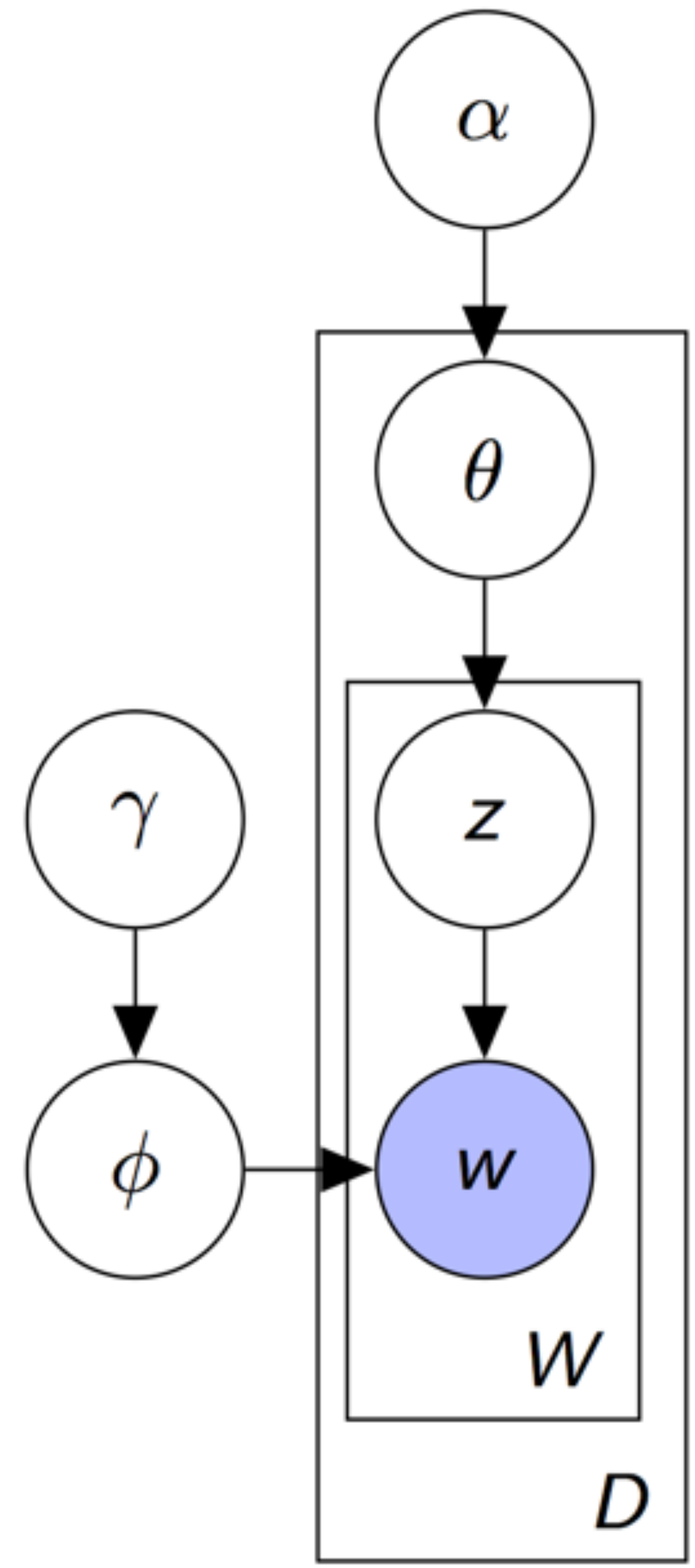


recomputation of means

Try it yourself

- Shakespeare + English stoplist
  http://bit.ly/1hdKX0R
- http://lexos.wheatoncollege.edu

Topic Models



Topic Models

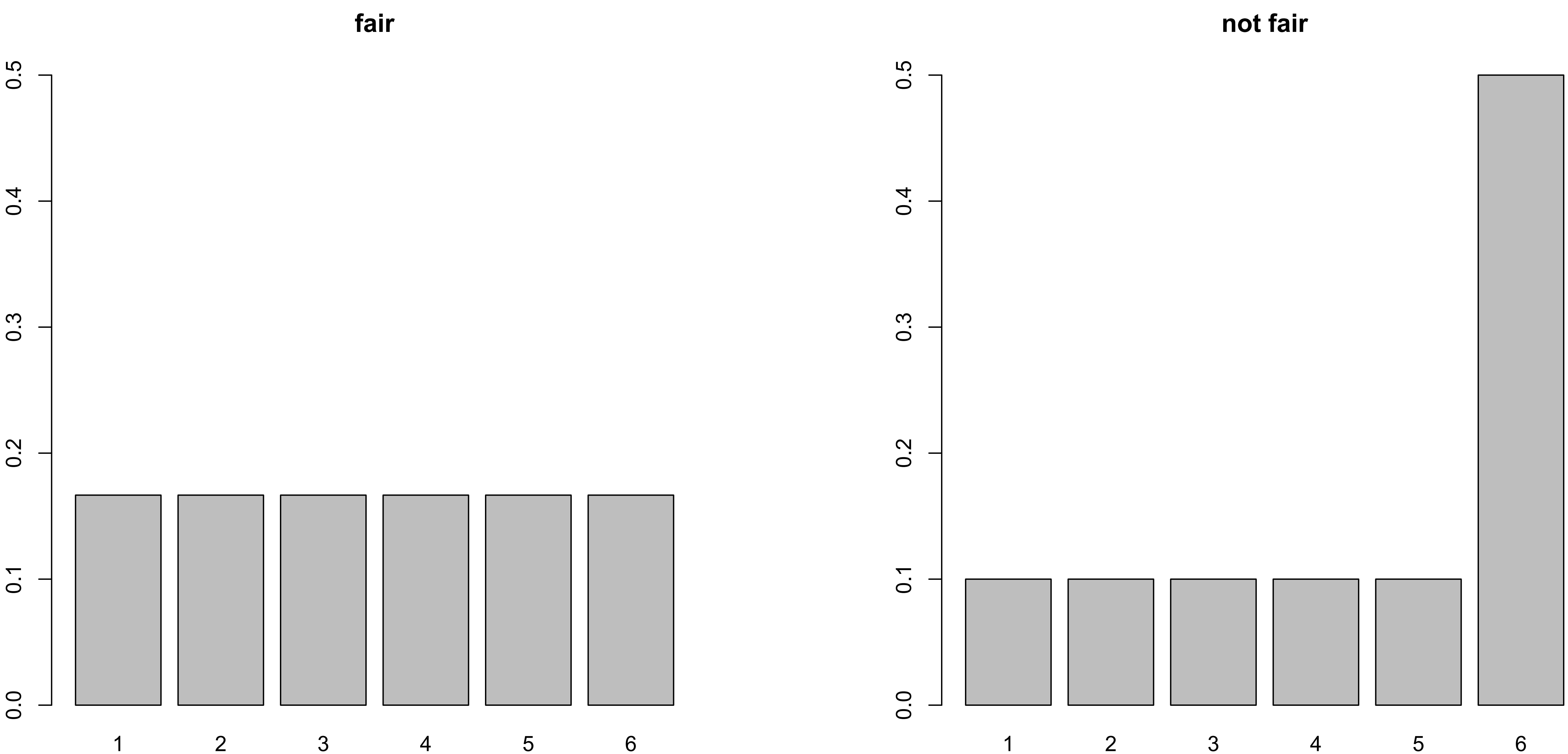
- A probabilistic model for discovering hidden “topics” or “themes” (groups of terms that tend to occur together) in documents.
- Unsupervised (find *interesting structure* in the data)
- Clustering algorithm

Topic Models

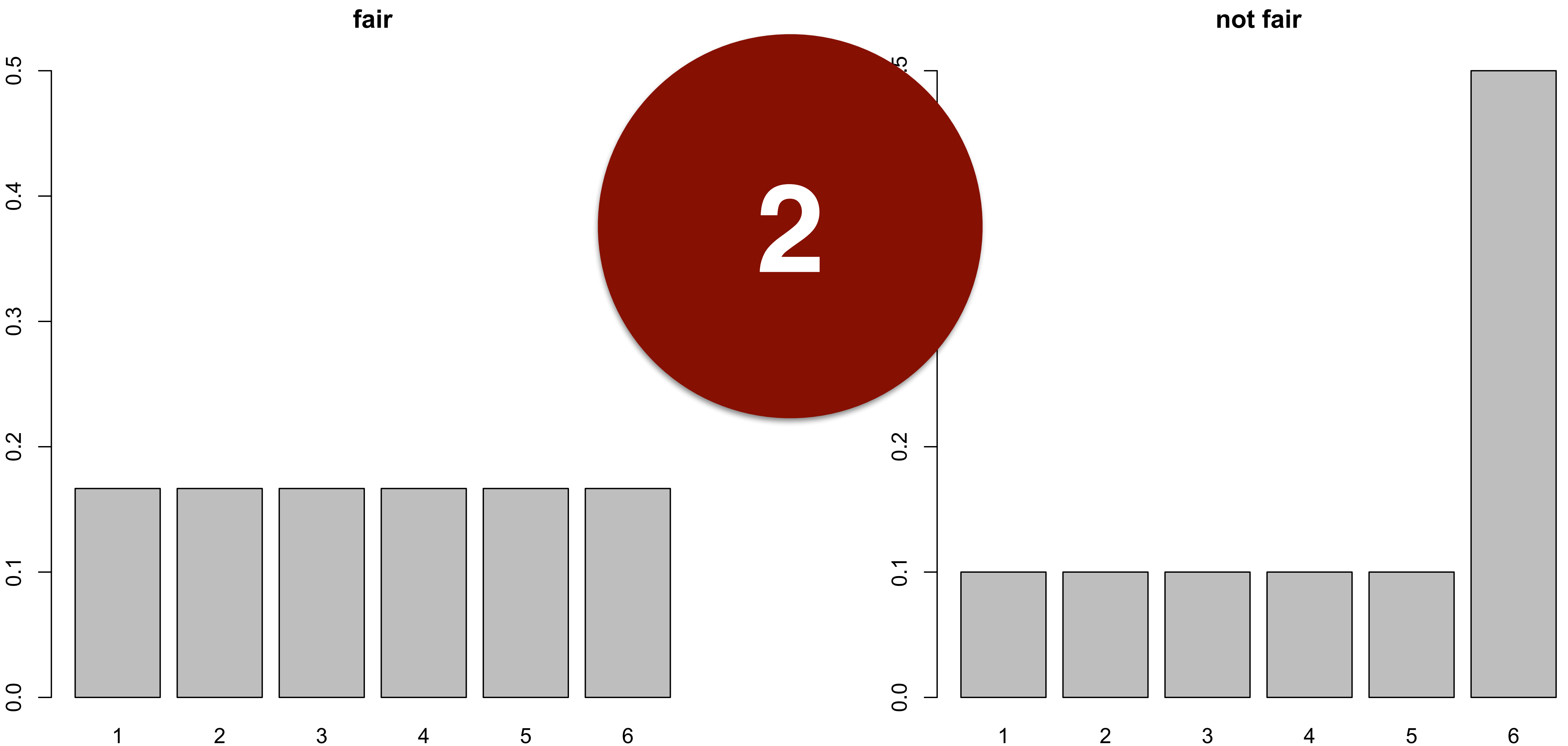
- **Input**: set of documents, number of clusters to learn.
- **Output**:
    - topics
    - topic ratio in each document
    - topic distribution for each word in doc

{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

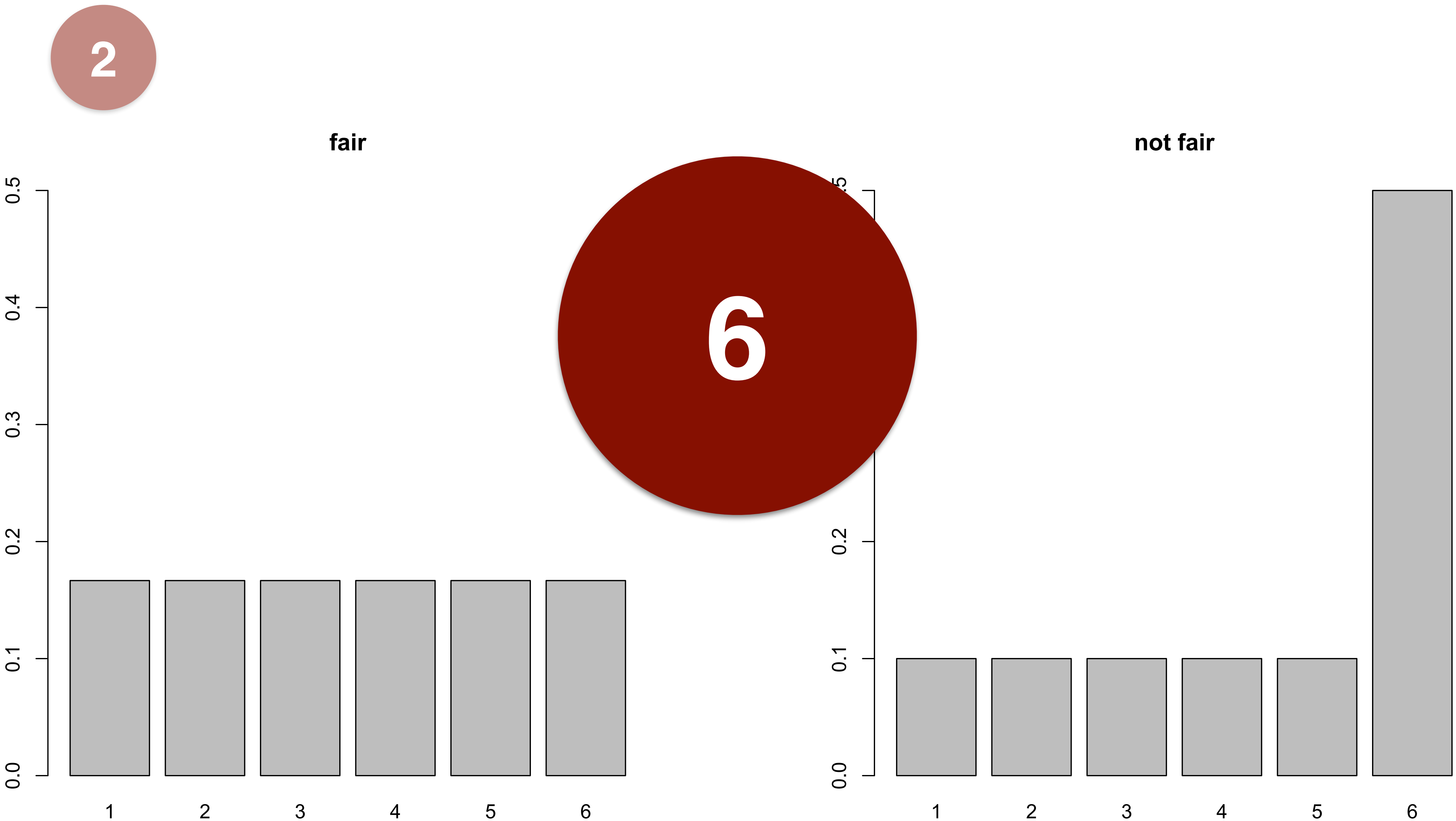
Probability



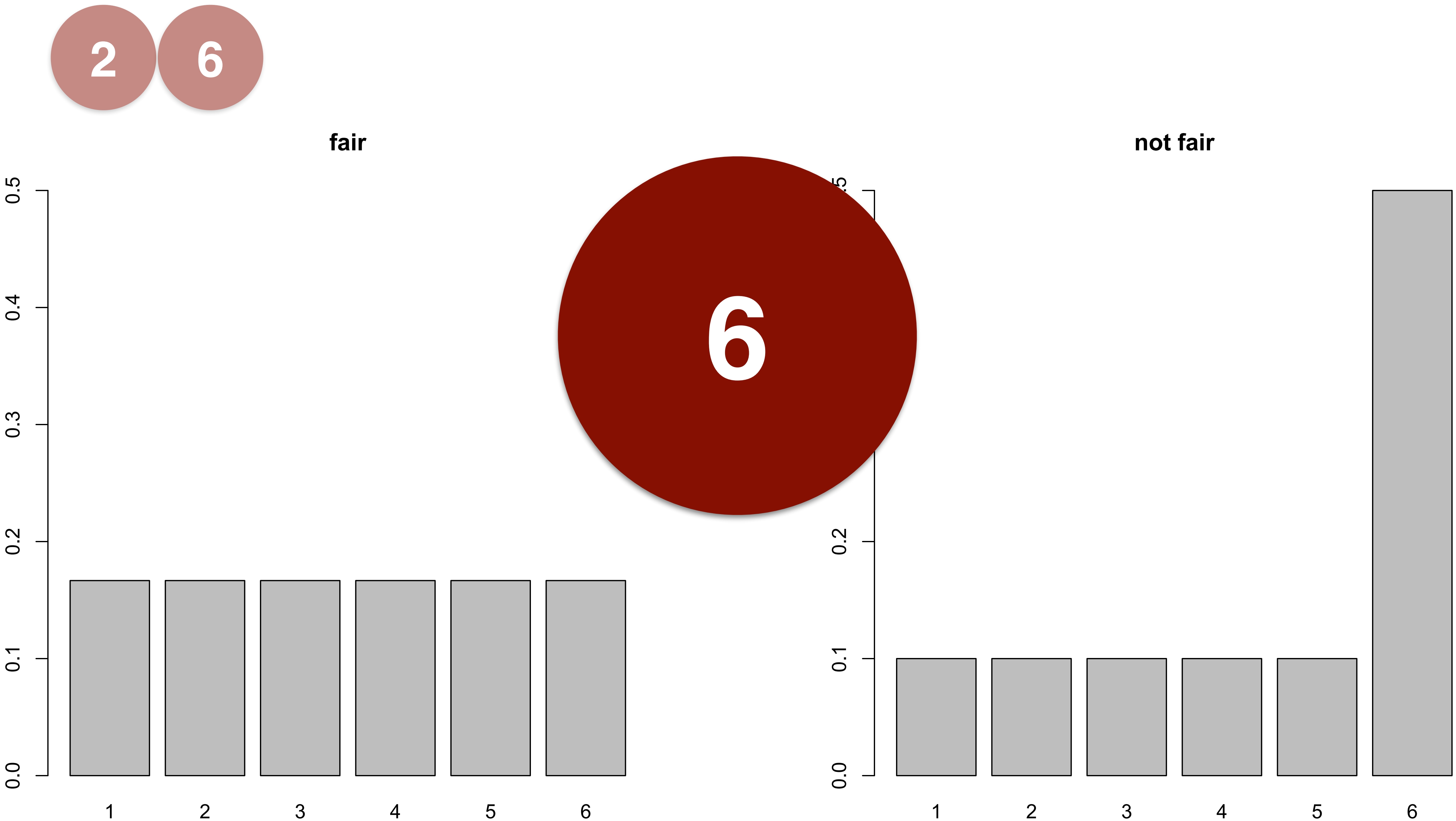
Probability



Probability

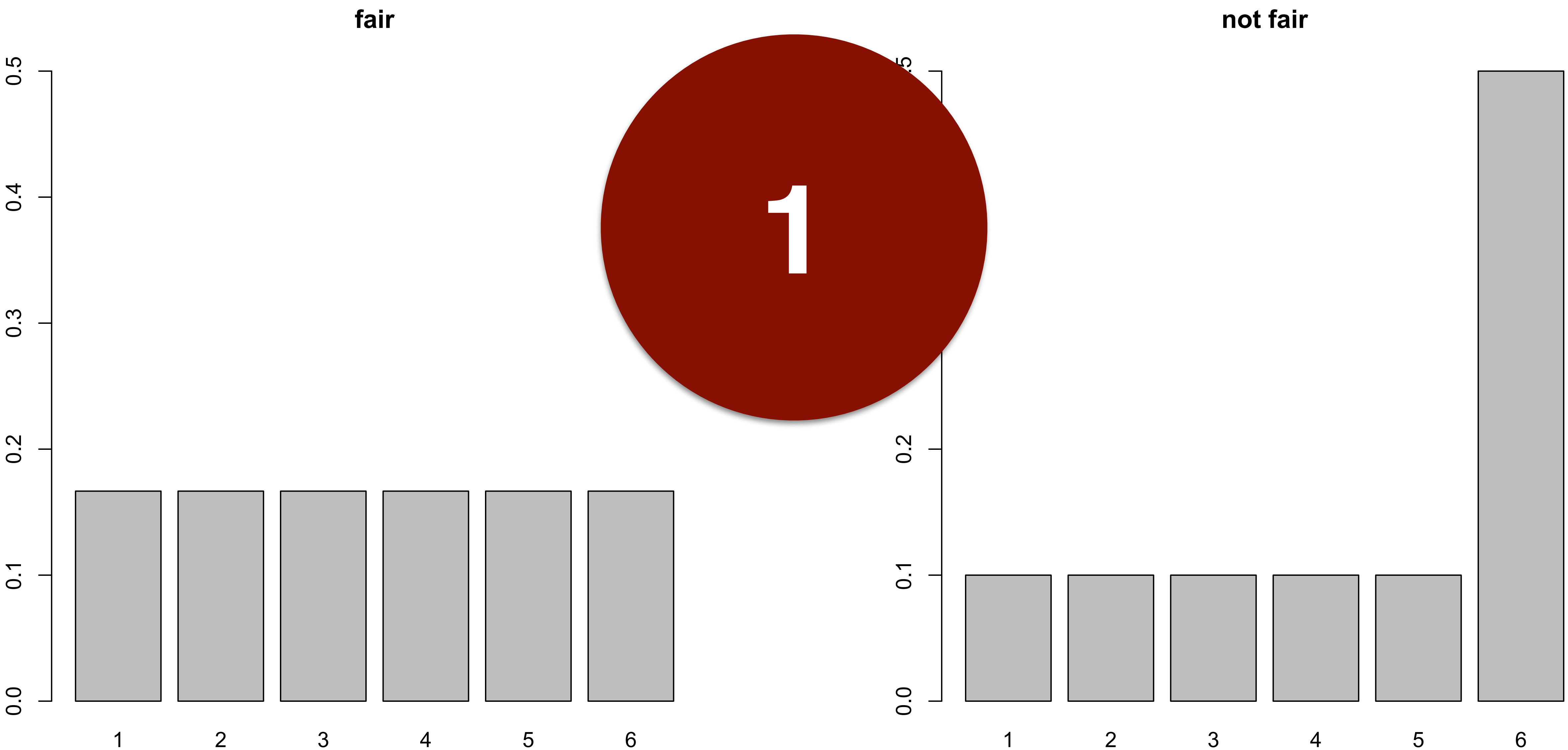


Probability



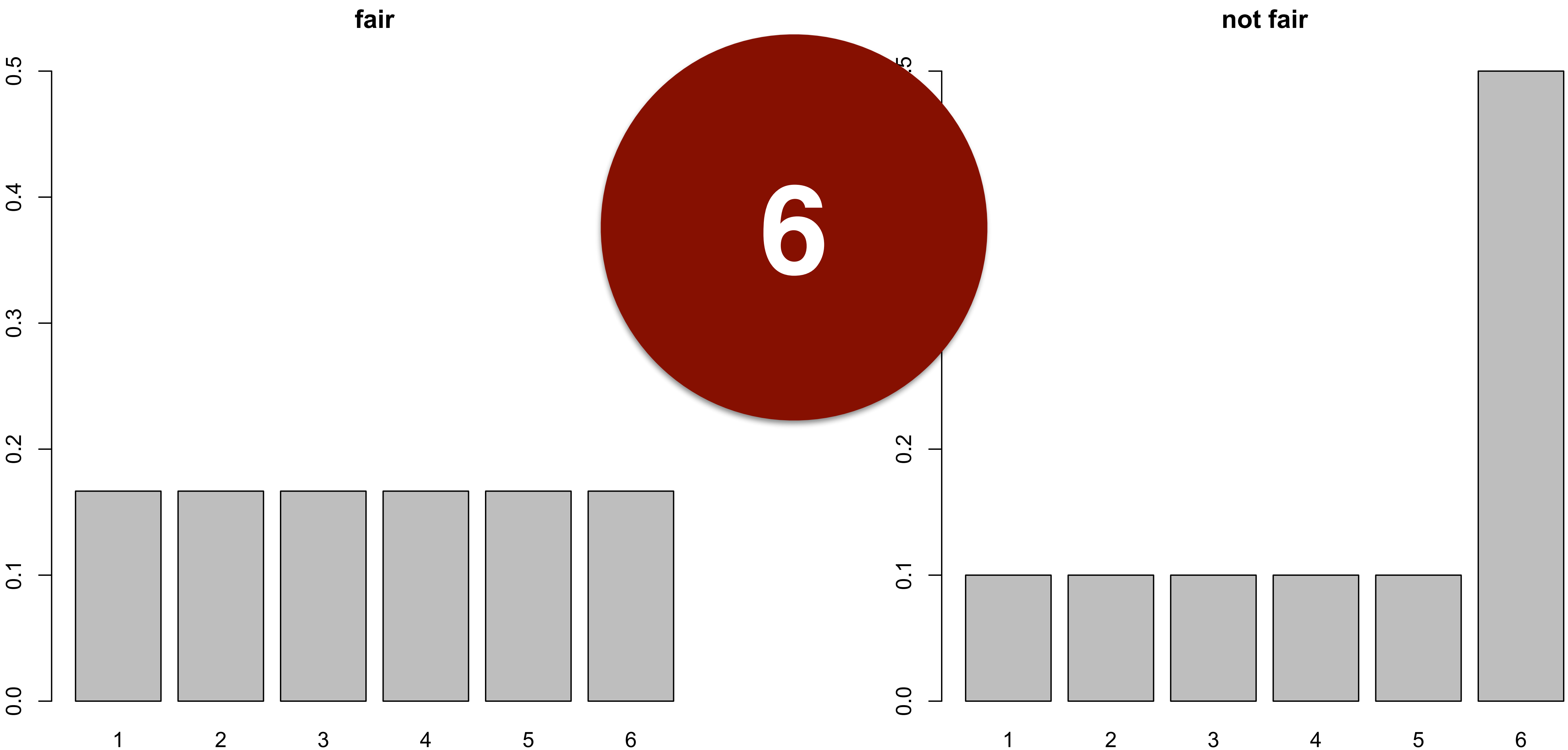
Probability

2 6 6

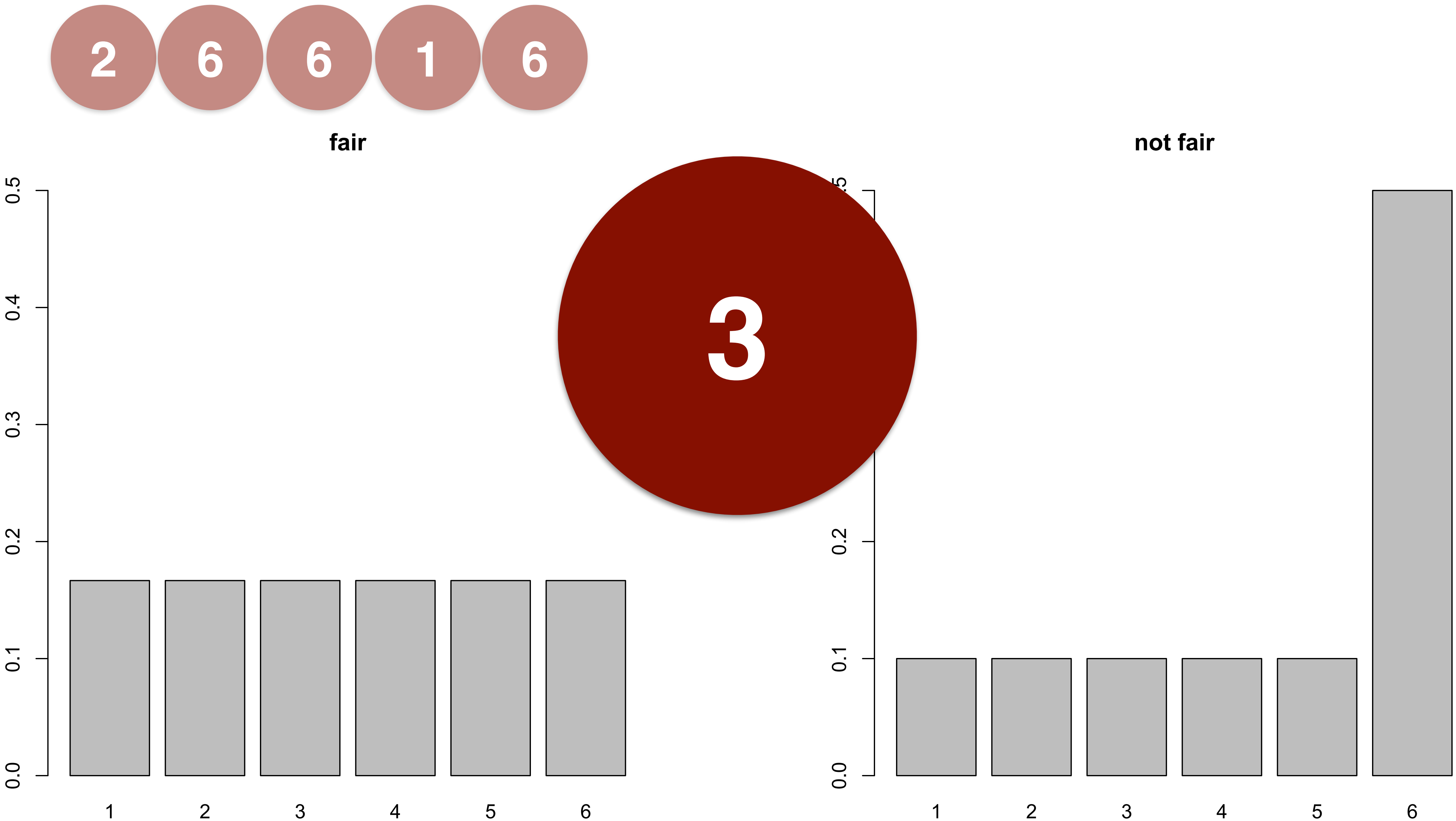


Probability

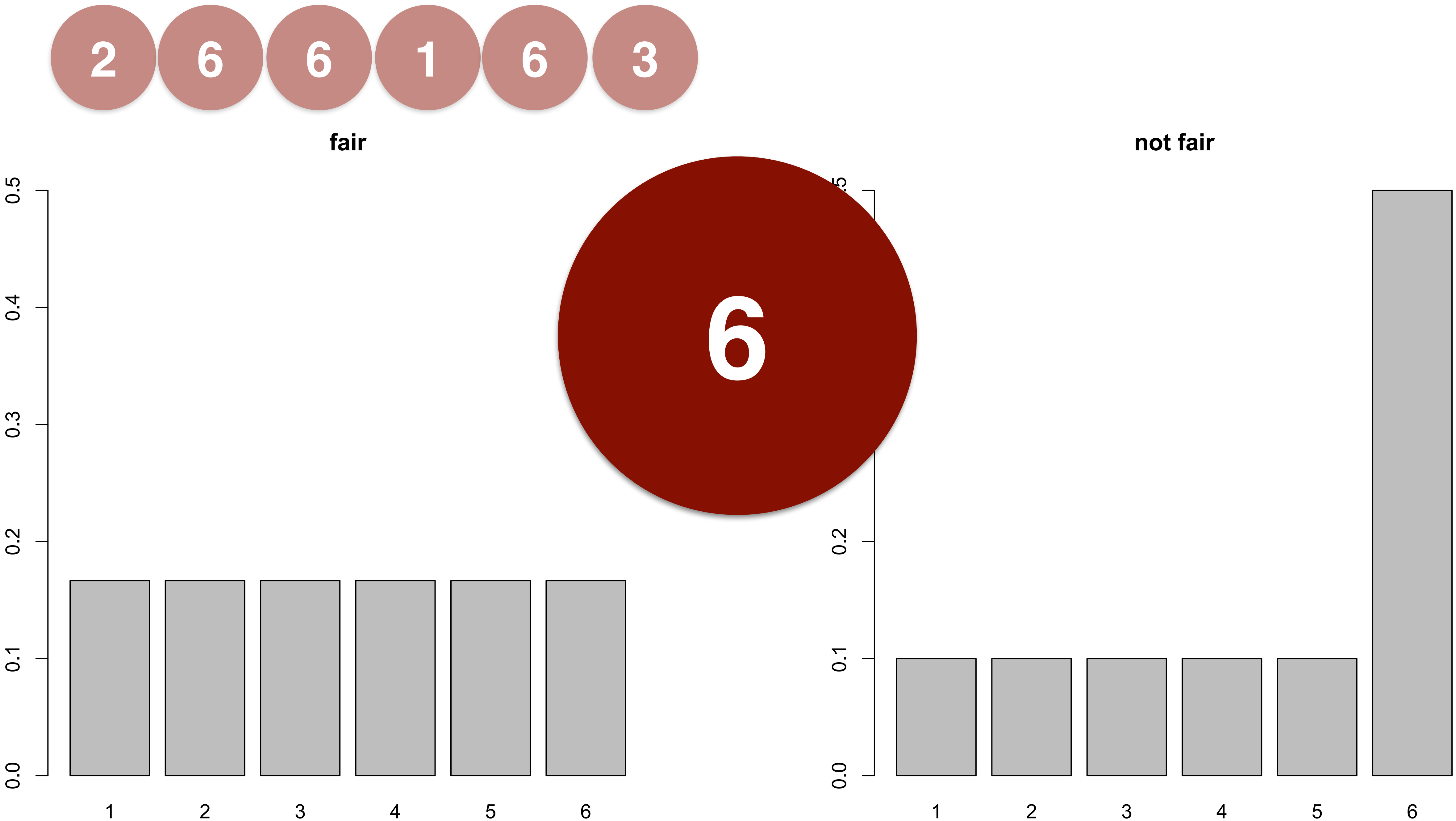
2 6 6 1



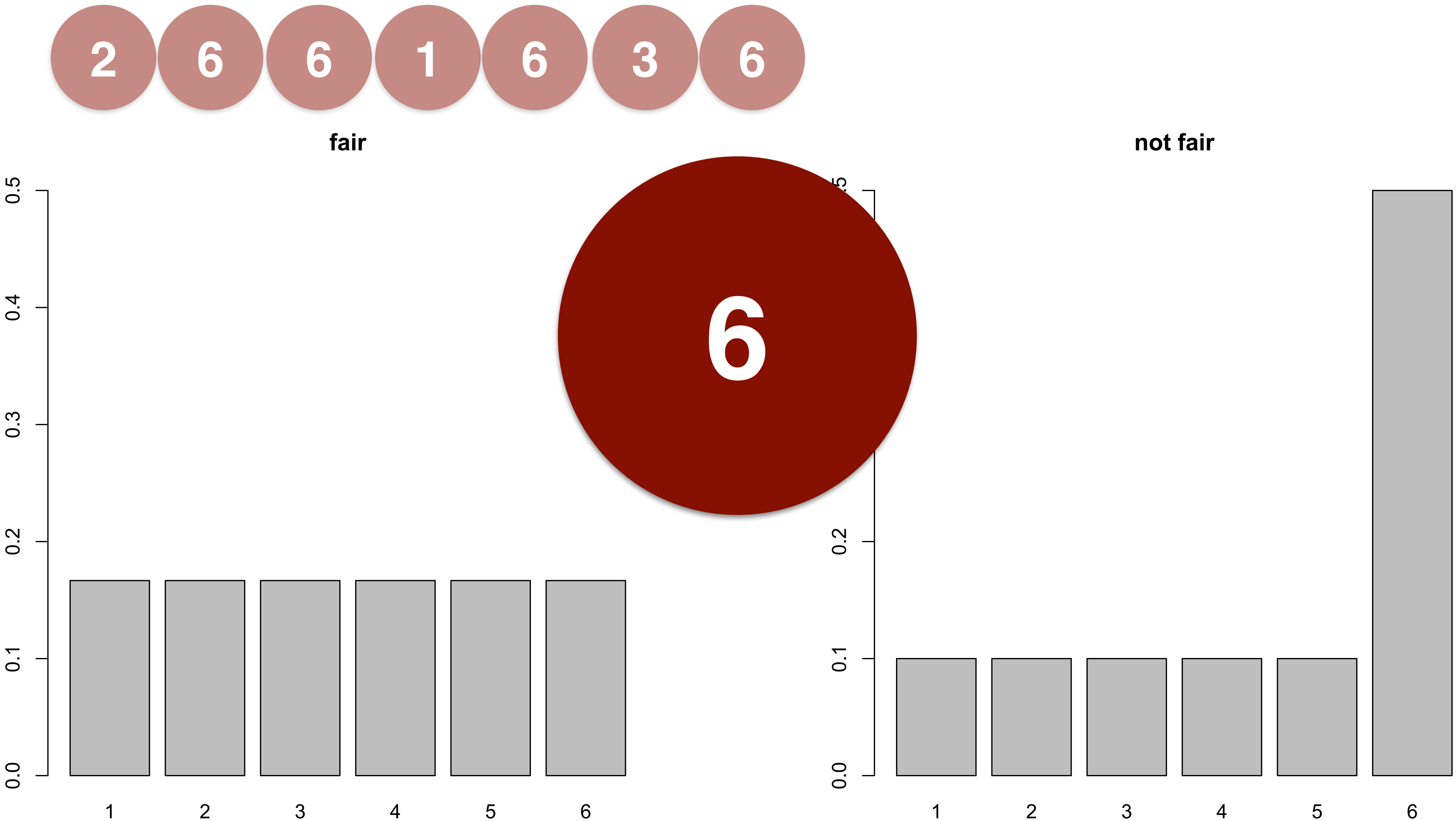
Probability



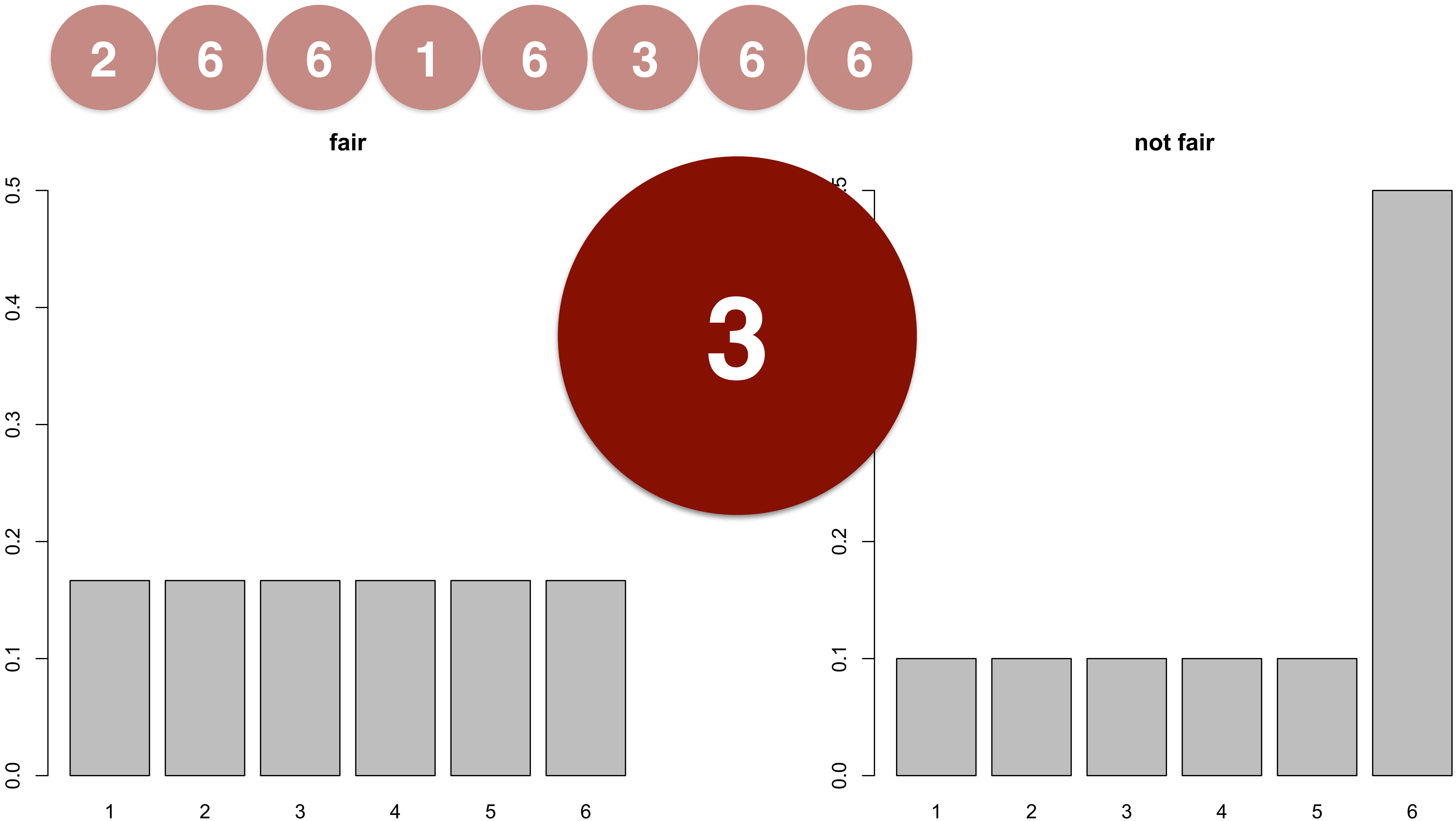
Probability



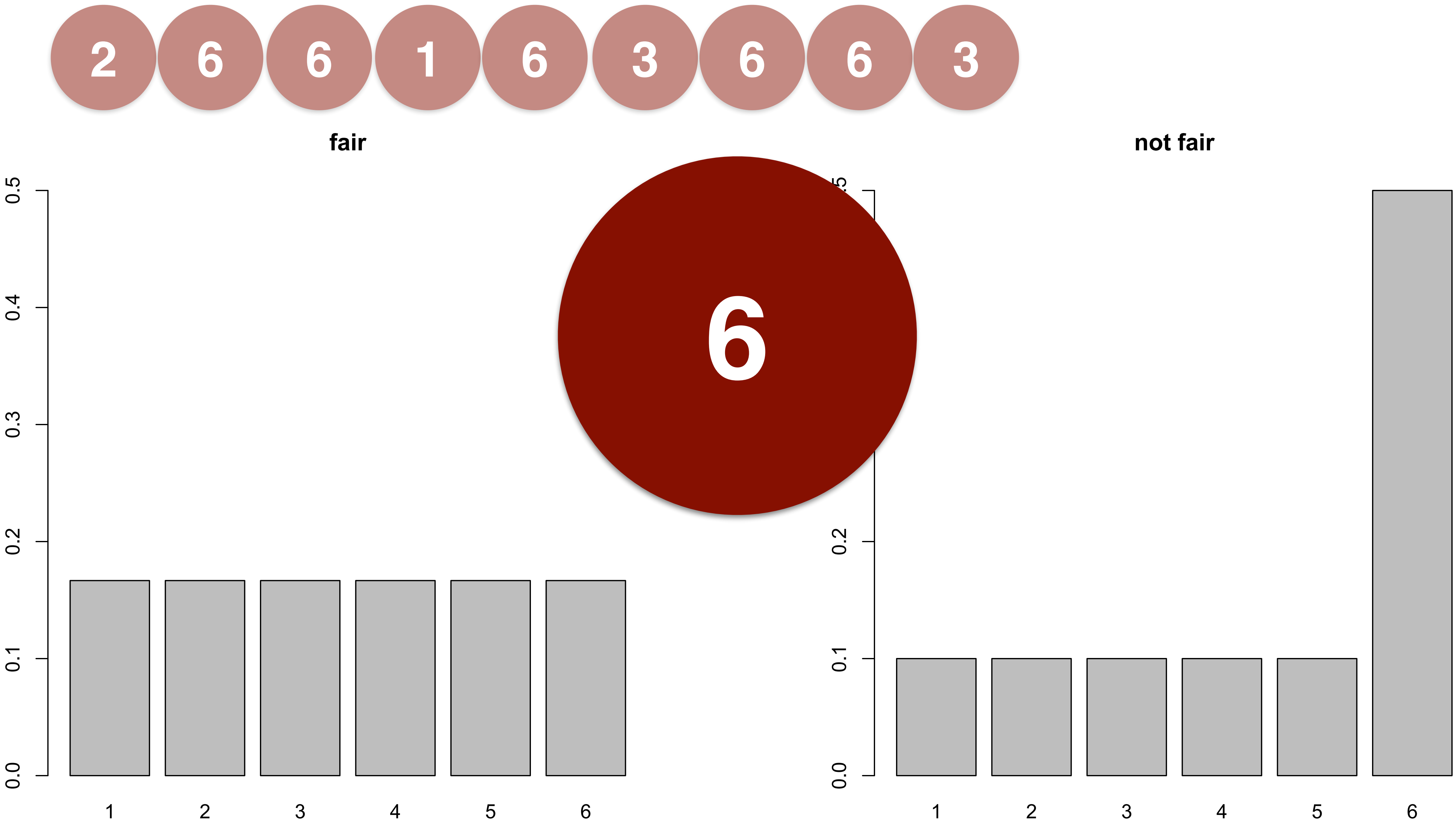
Probability



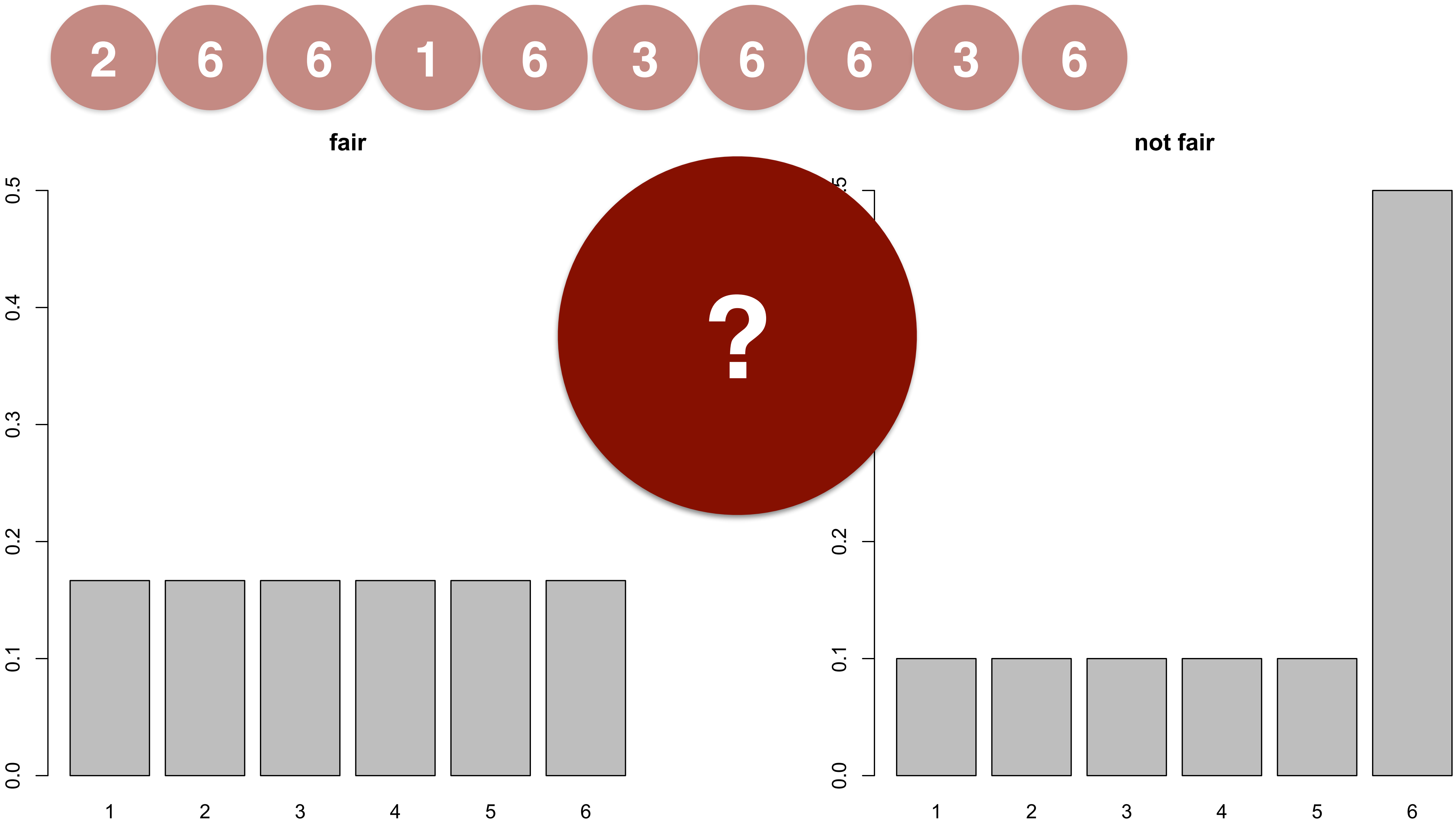
Probability



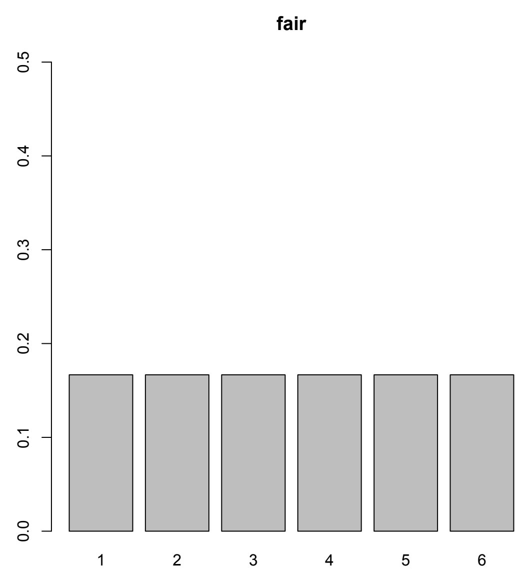
Probability



Probability

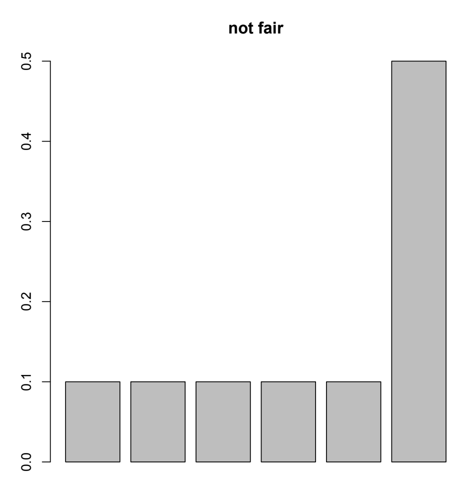


1. Data "Likelihood"

P(2 6 6 | )

$= .17 \times .17 \times .17$

$= 0.004913$

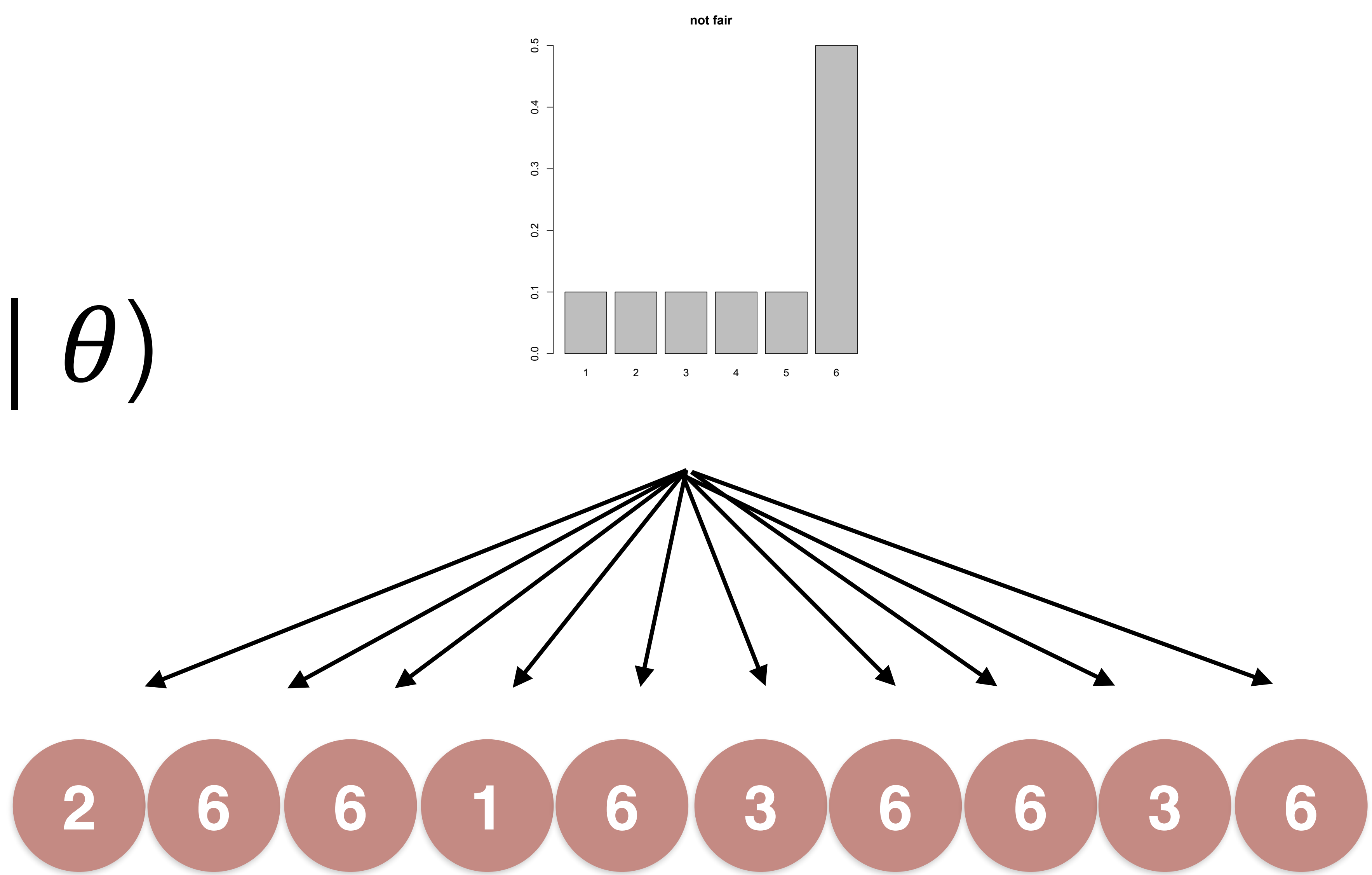
P(2 6 6 | )

$= .1 \times .5 \times .5$

$= 0.025$

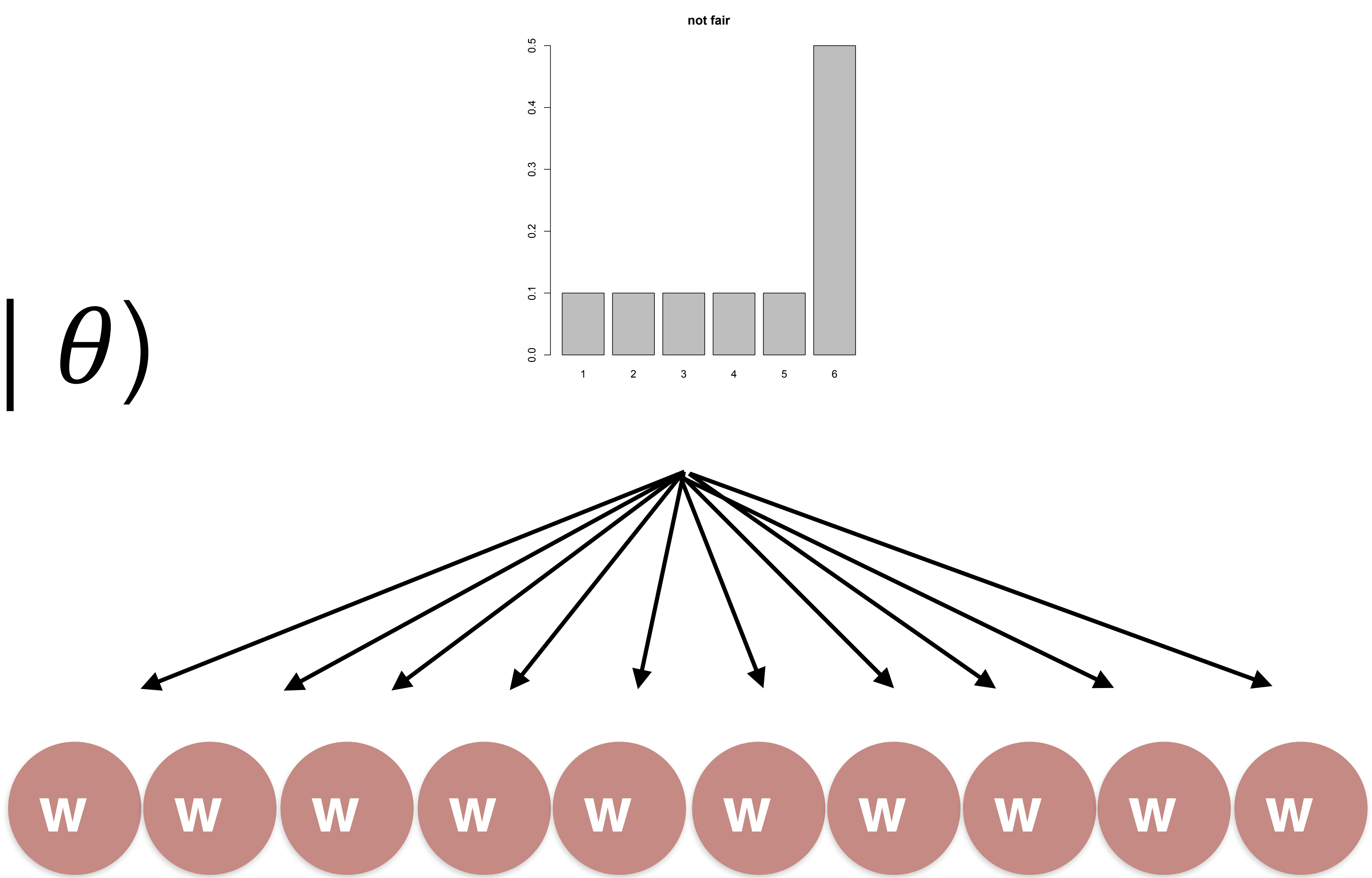
2. Conditional Probability

$P(w \mid \theta)$



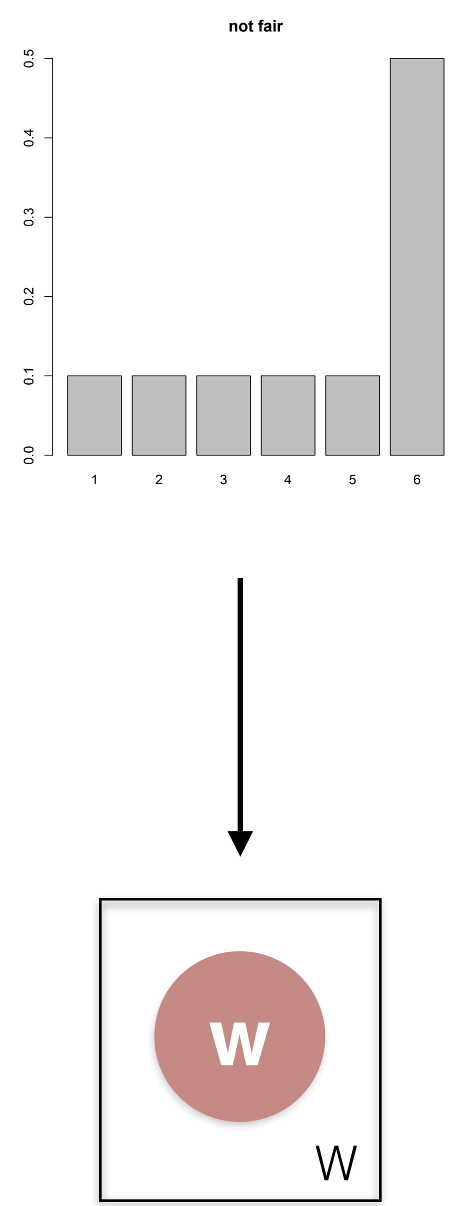
2. Conditional Probability

$P(w \mid \theta)$



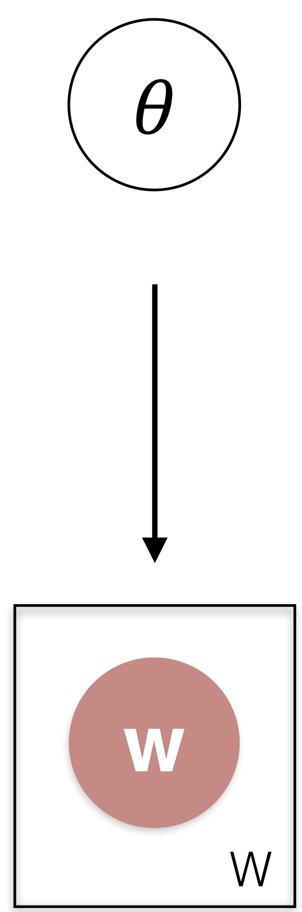
2. Conditional Probability

$P(w \mid \theta)$



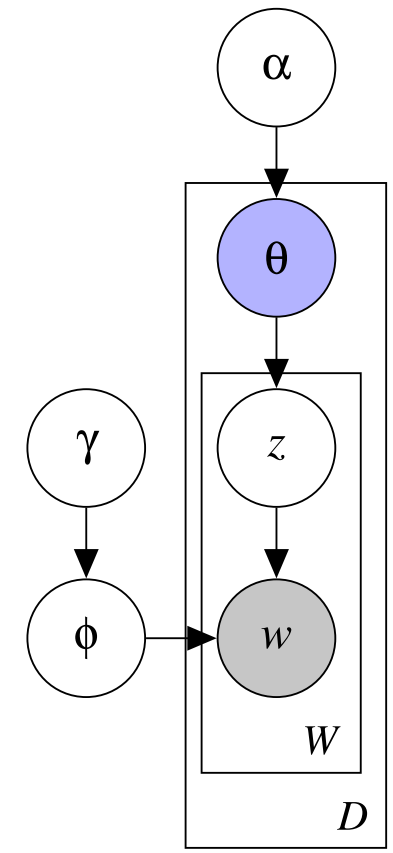
2. Conditional Probability

$P(w \mid \theta)$

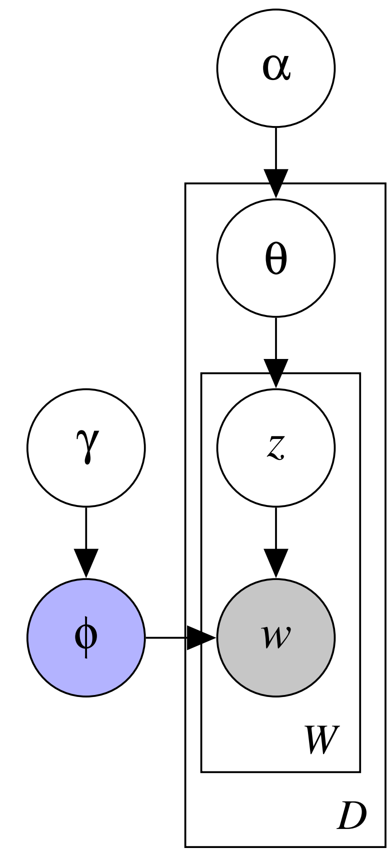


Topic Models

- A document has *distribution over topics*

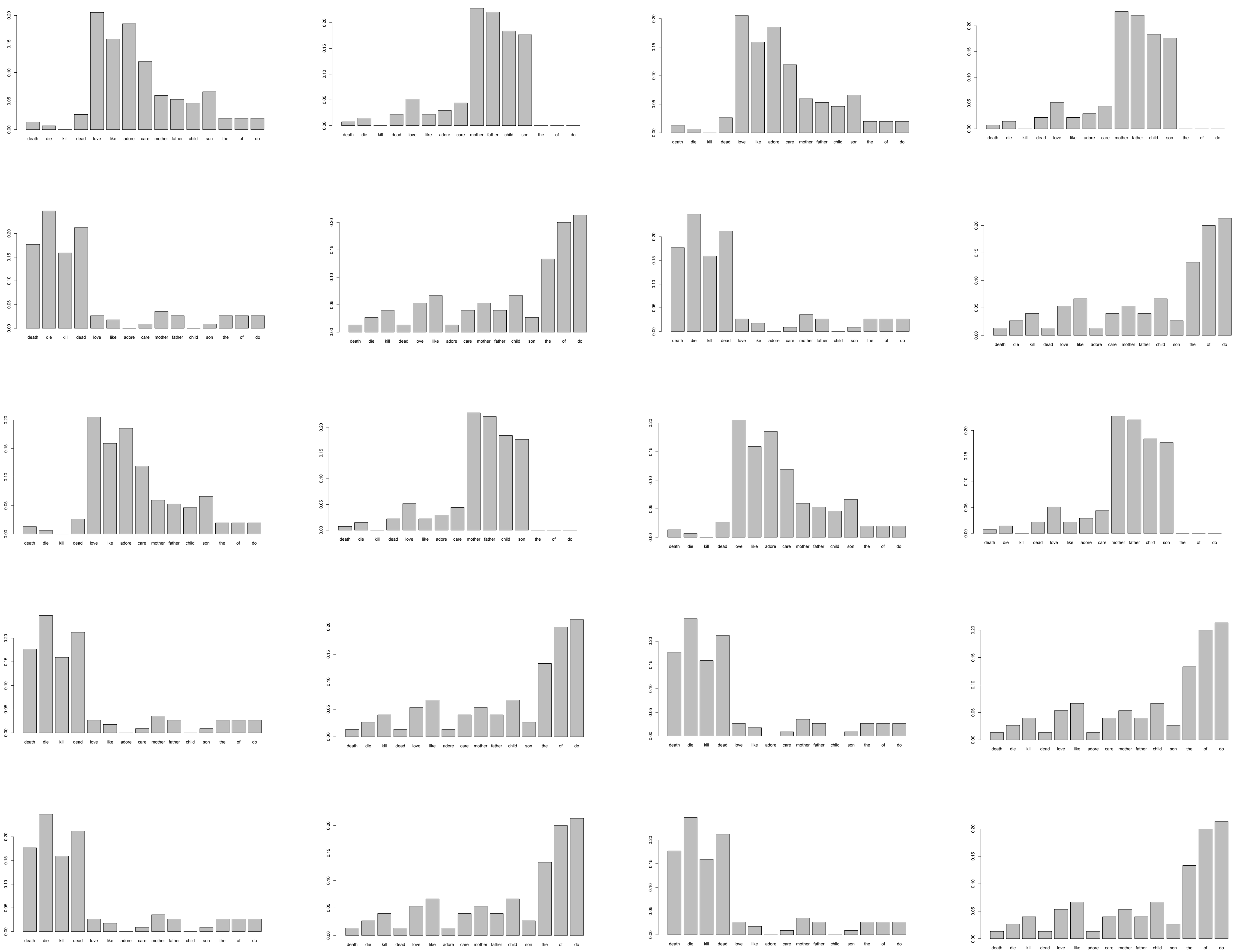


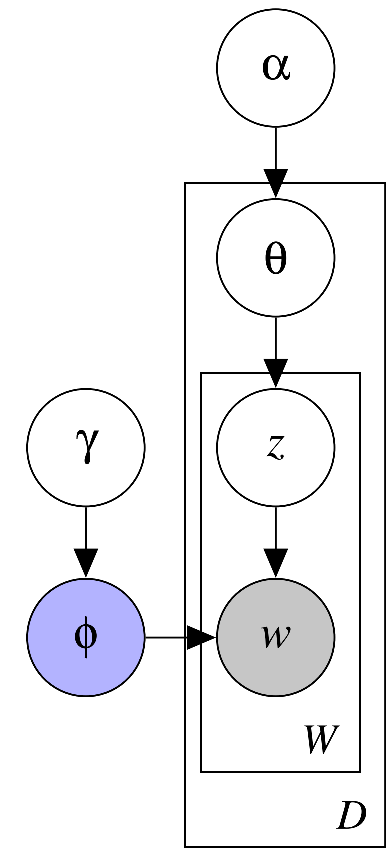
Topic Models



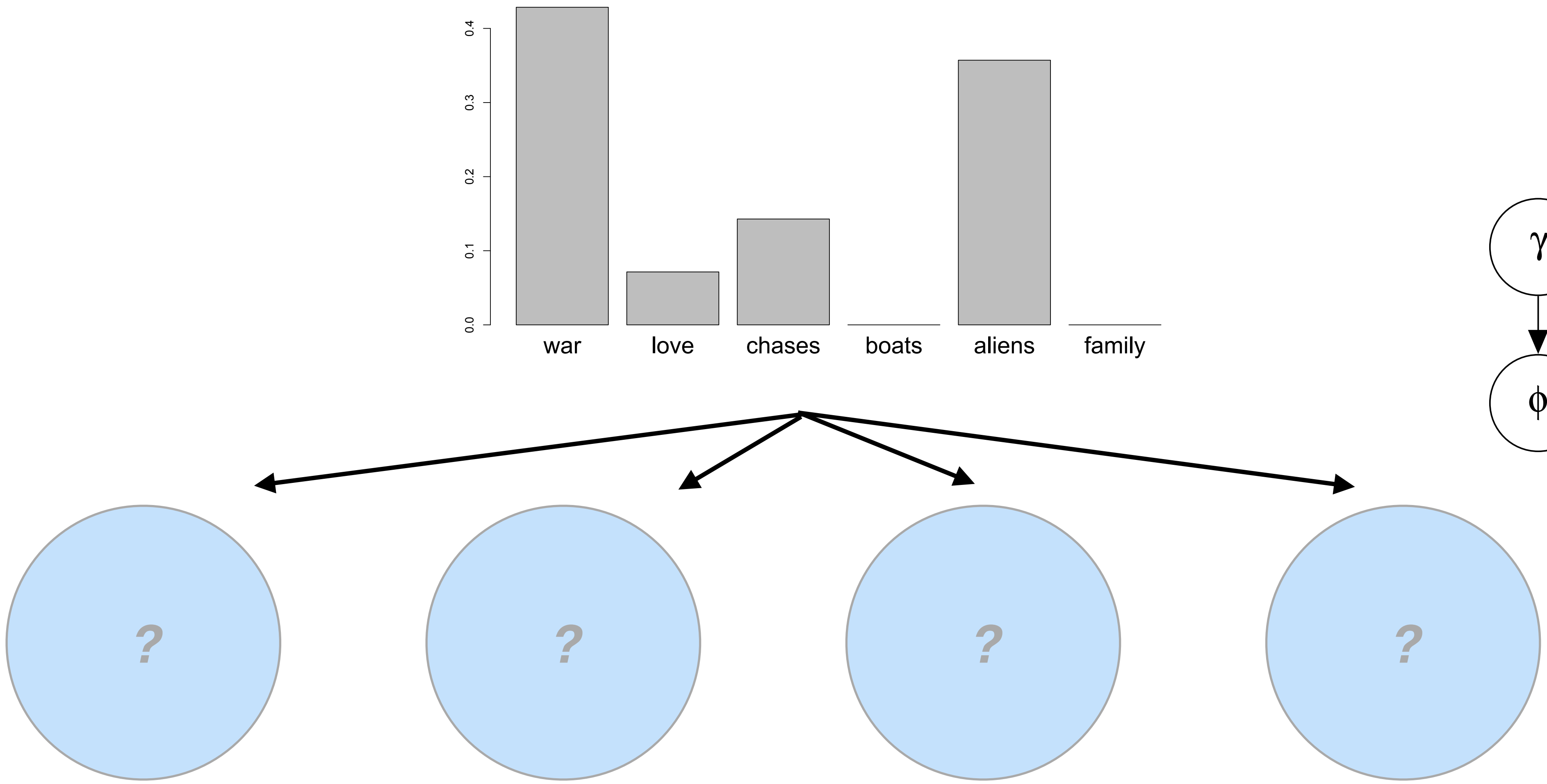
- A topic is a distribution over words

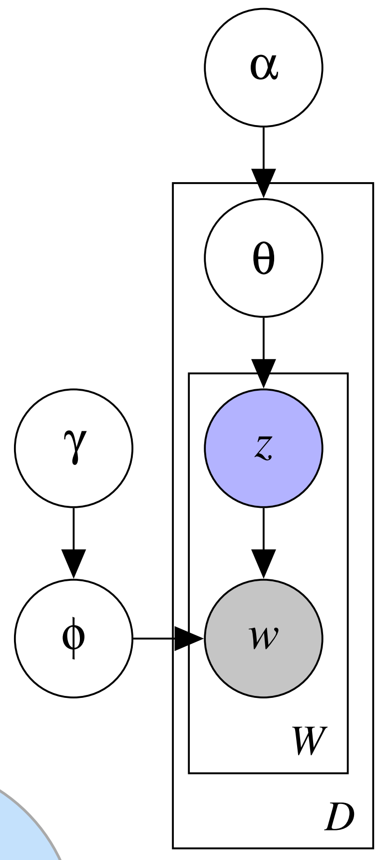
- e.g., P("adore" | topic = love) = .18



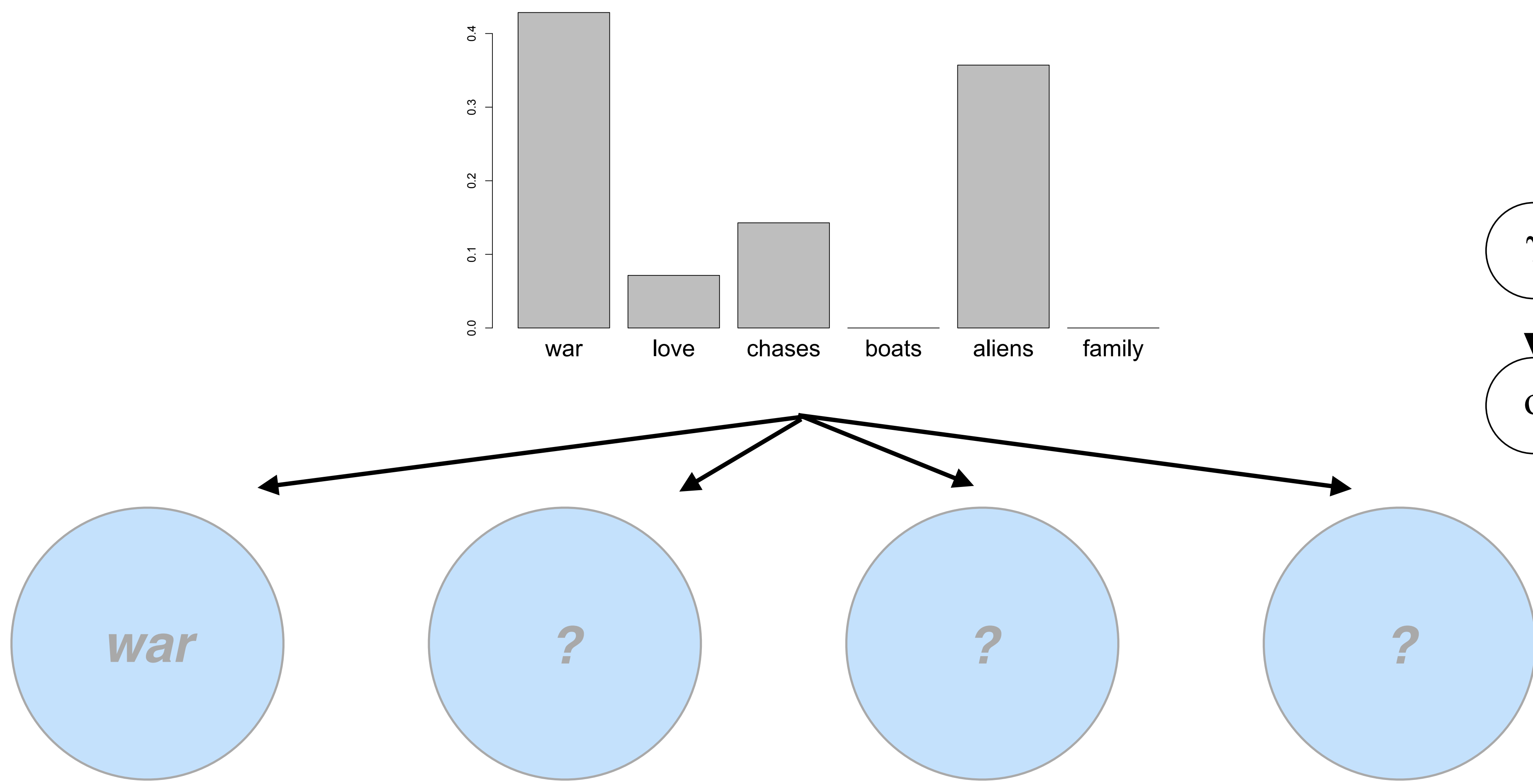


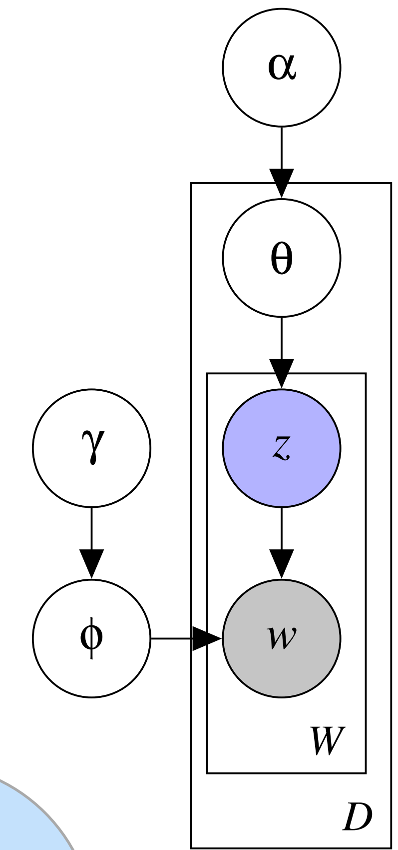
K=20



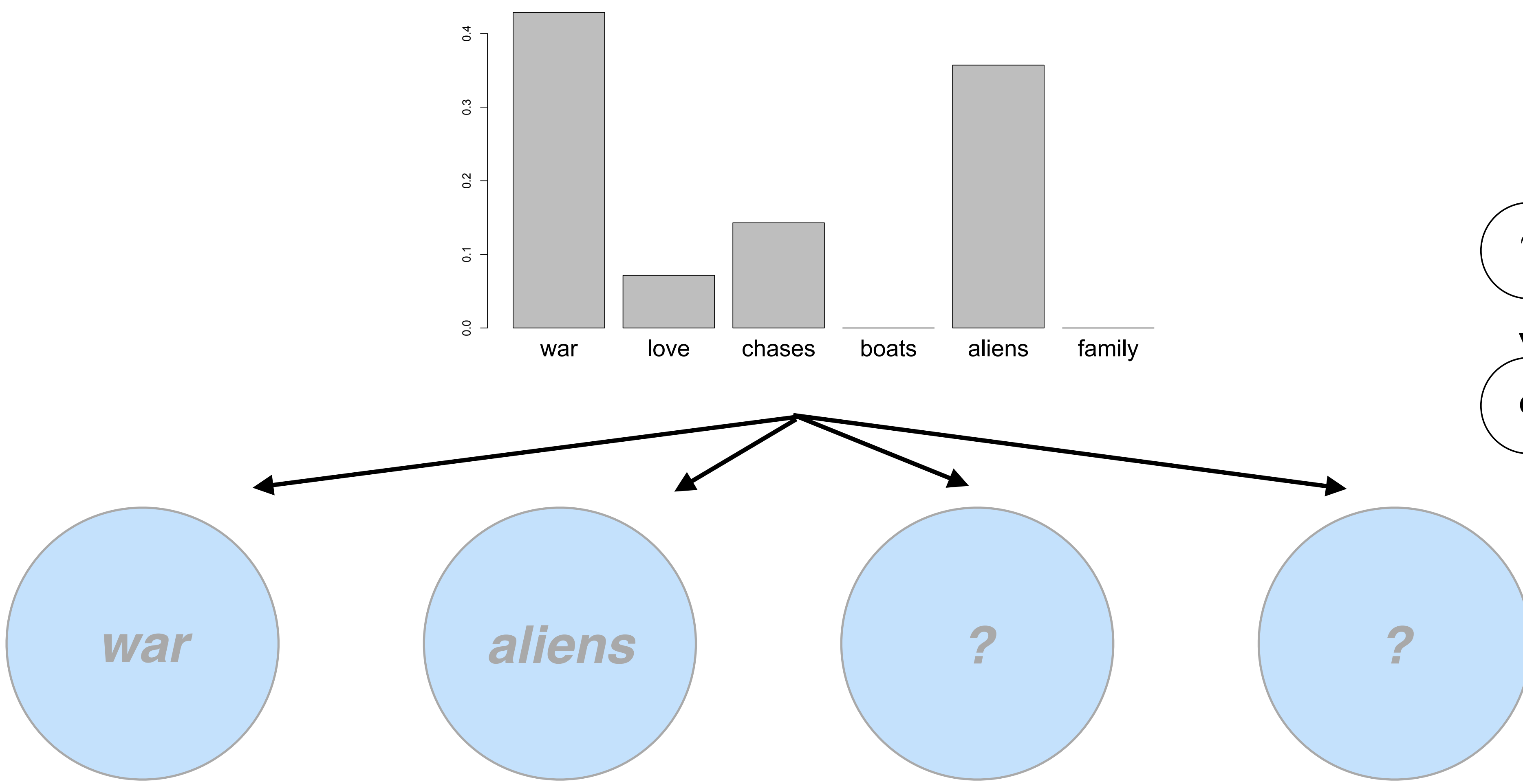


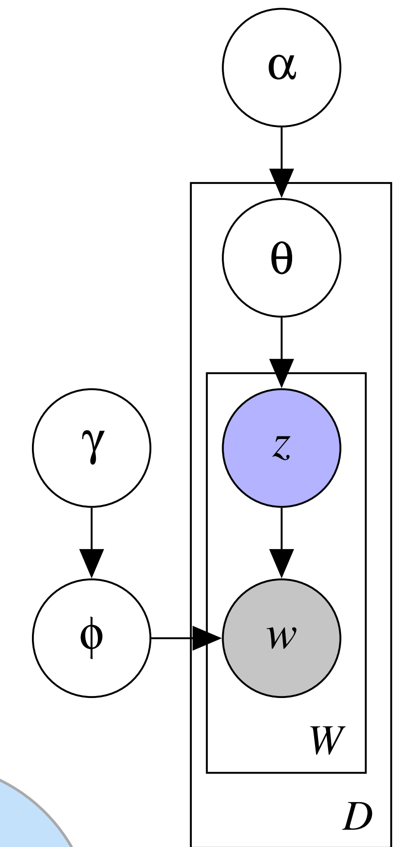
P(topic | topic distribution)



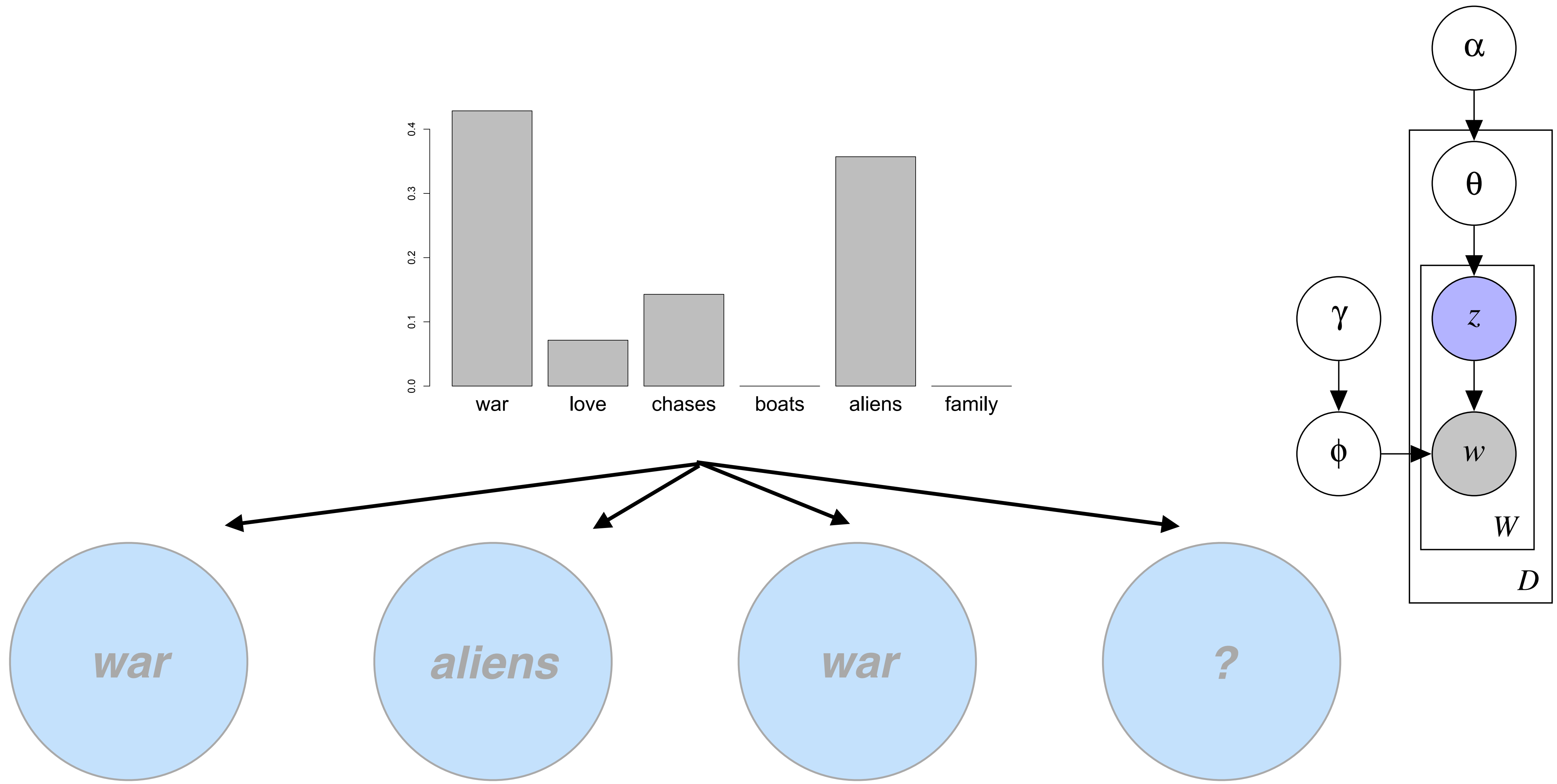


P(topic | topic distribution)

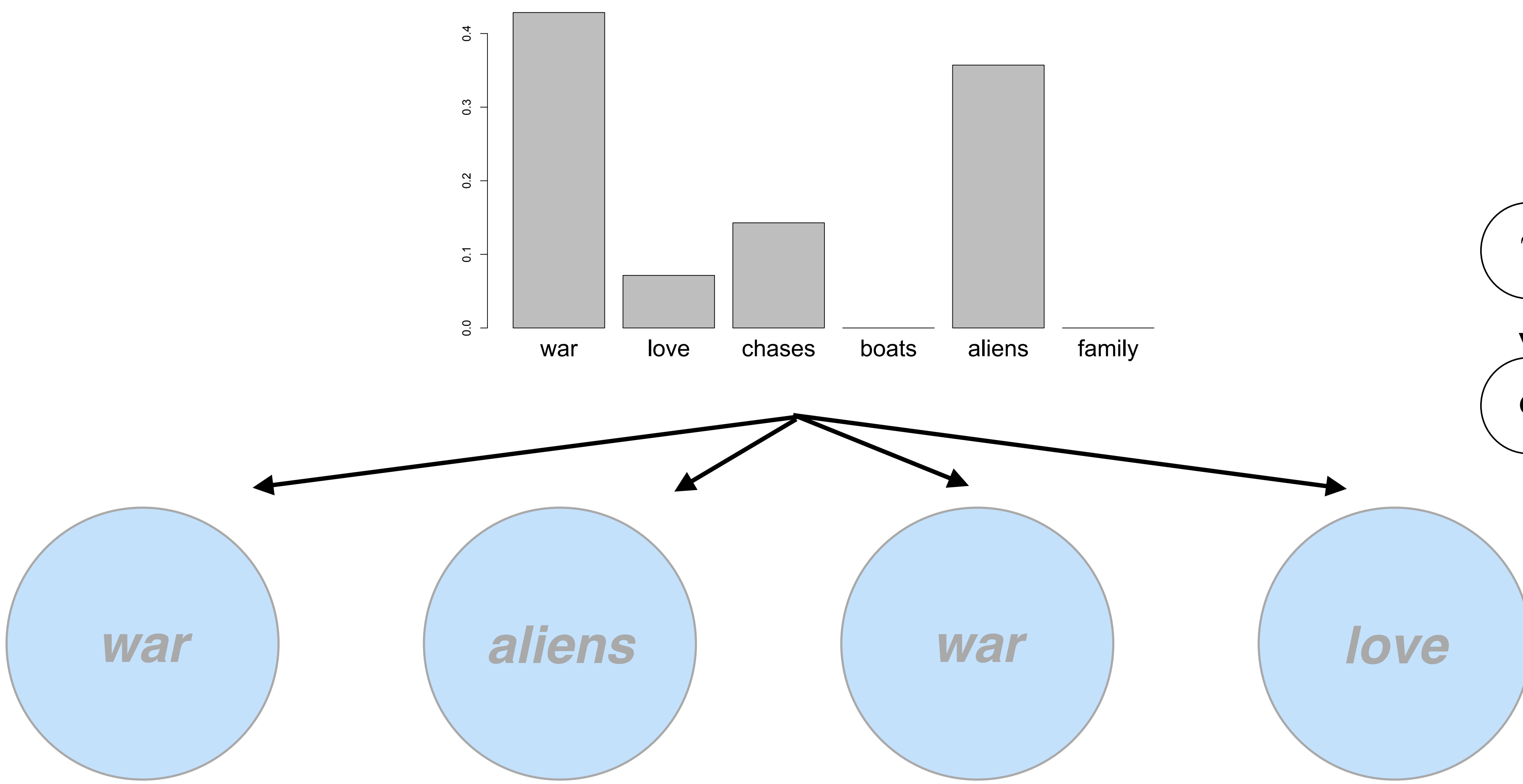


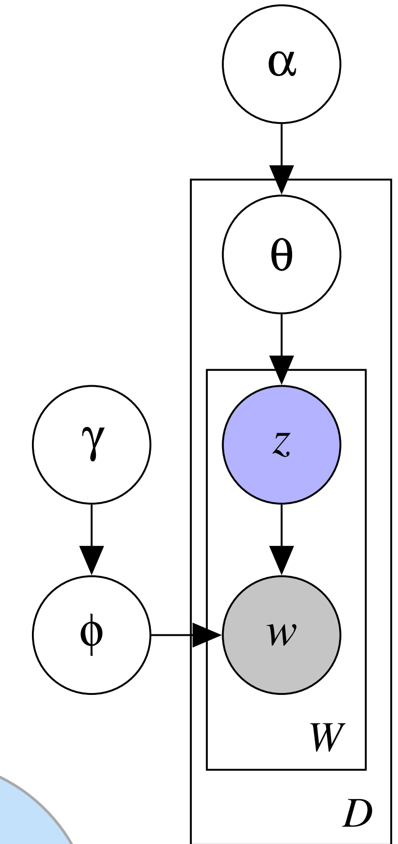


P(topic | topic distribution)



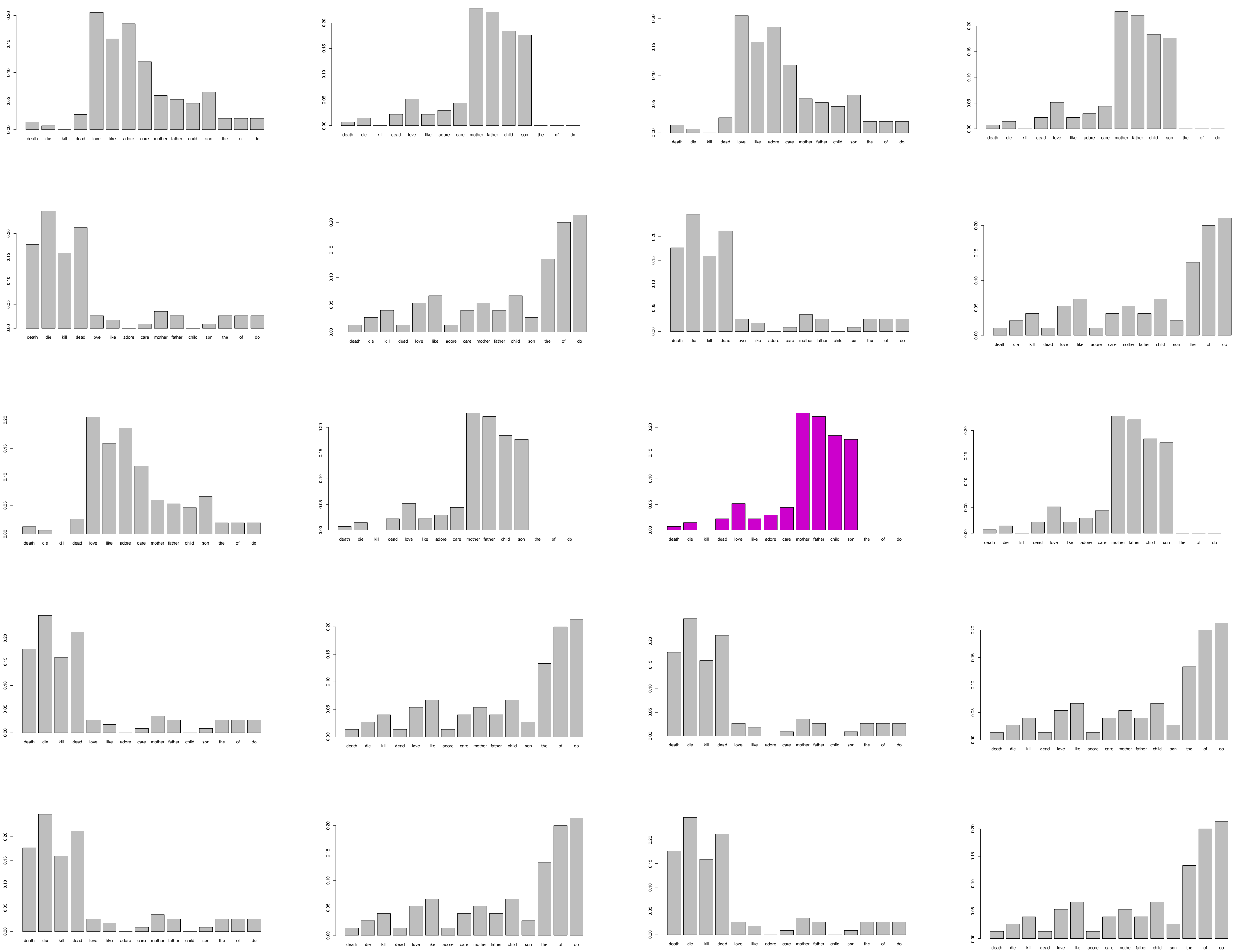
P(topic | topic distribution)

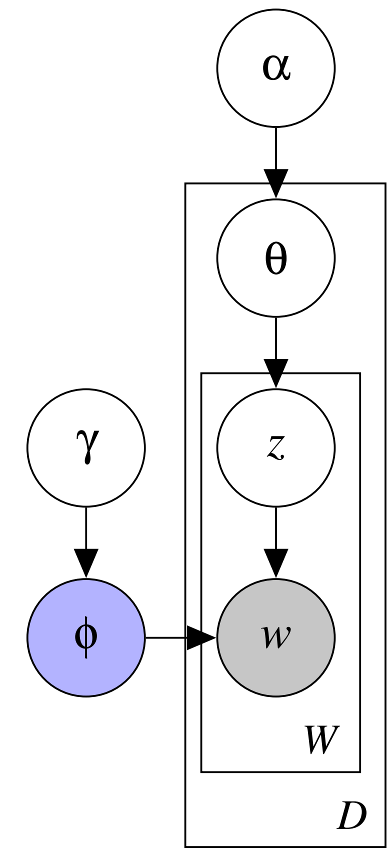




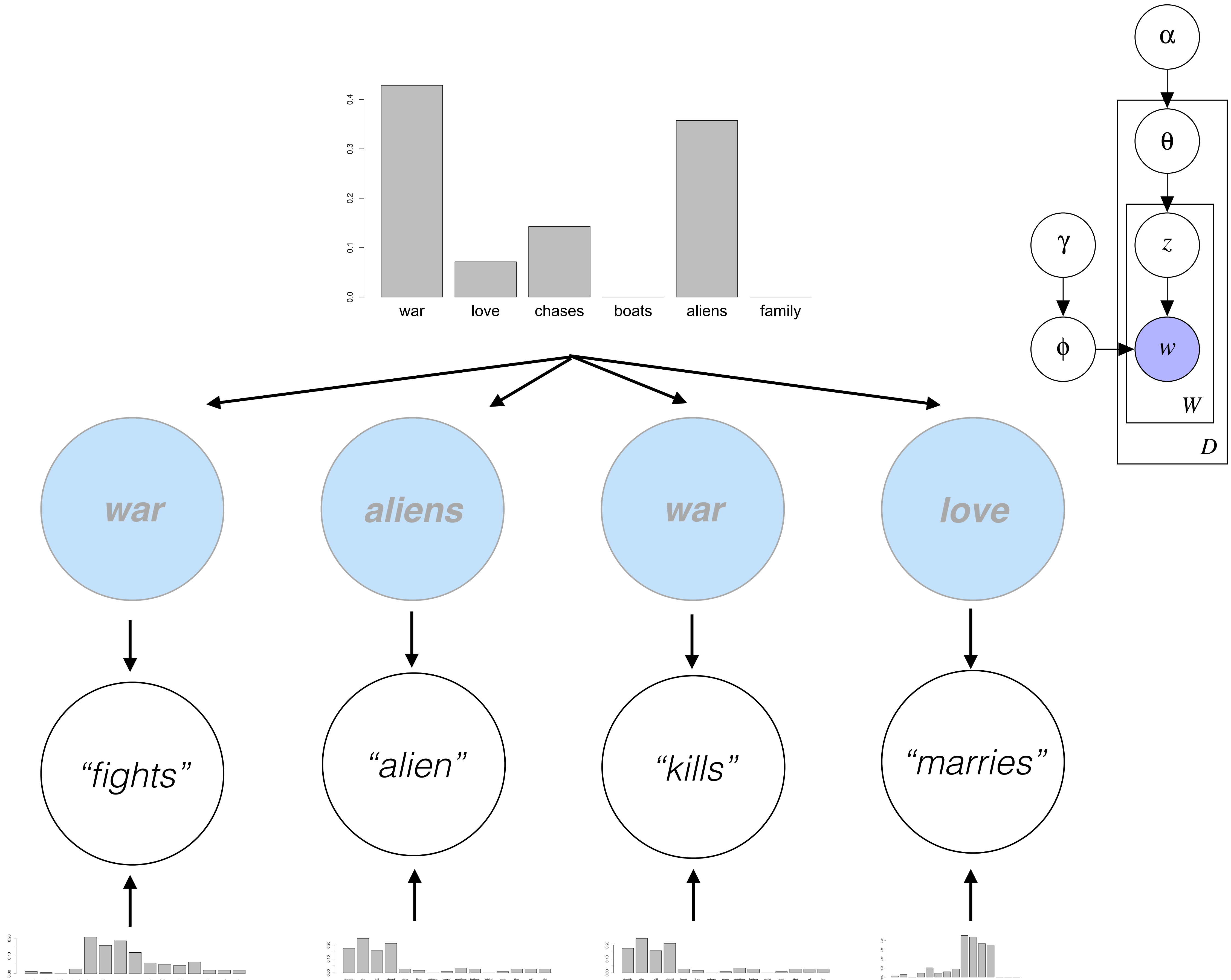
P(topic | topic distribution)

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

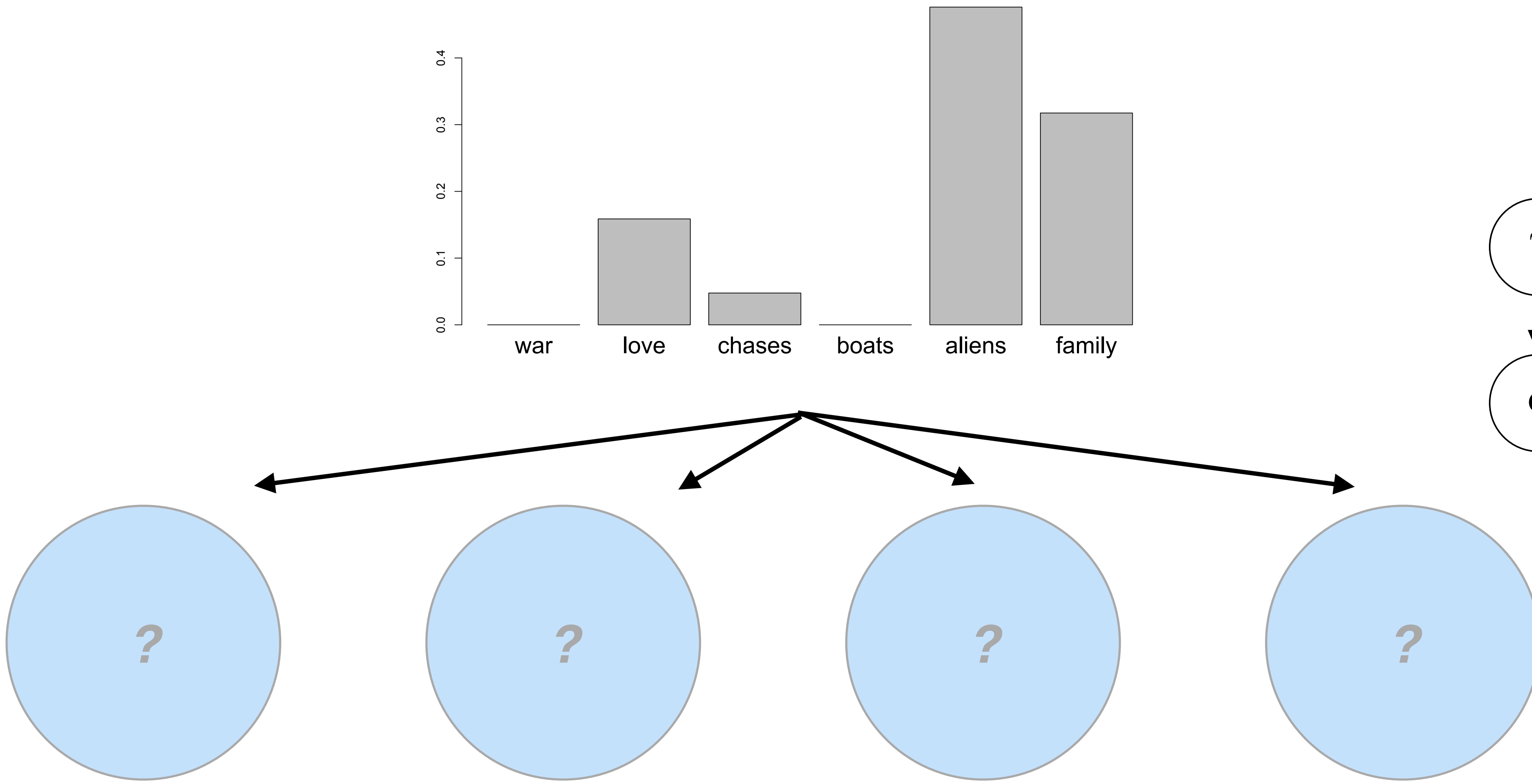

death die kill dead love like adore care mother father child son the of do

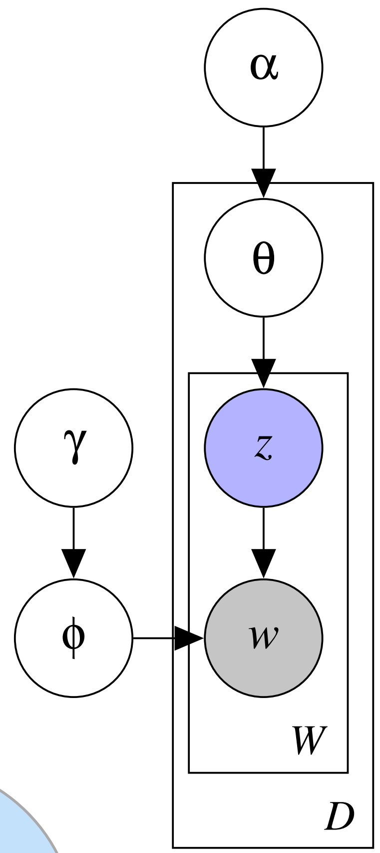

α
θ
γ
z
ϕ
w
W
D

K=20

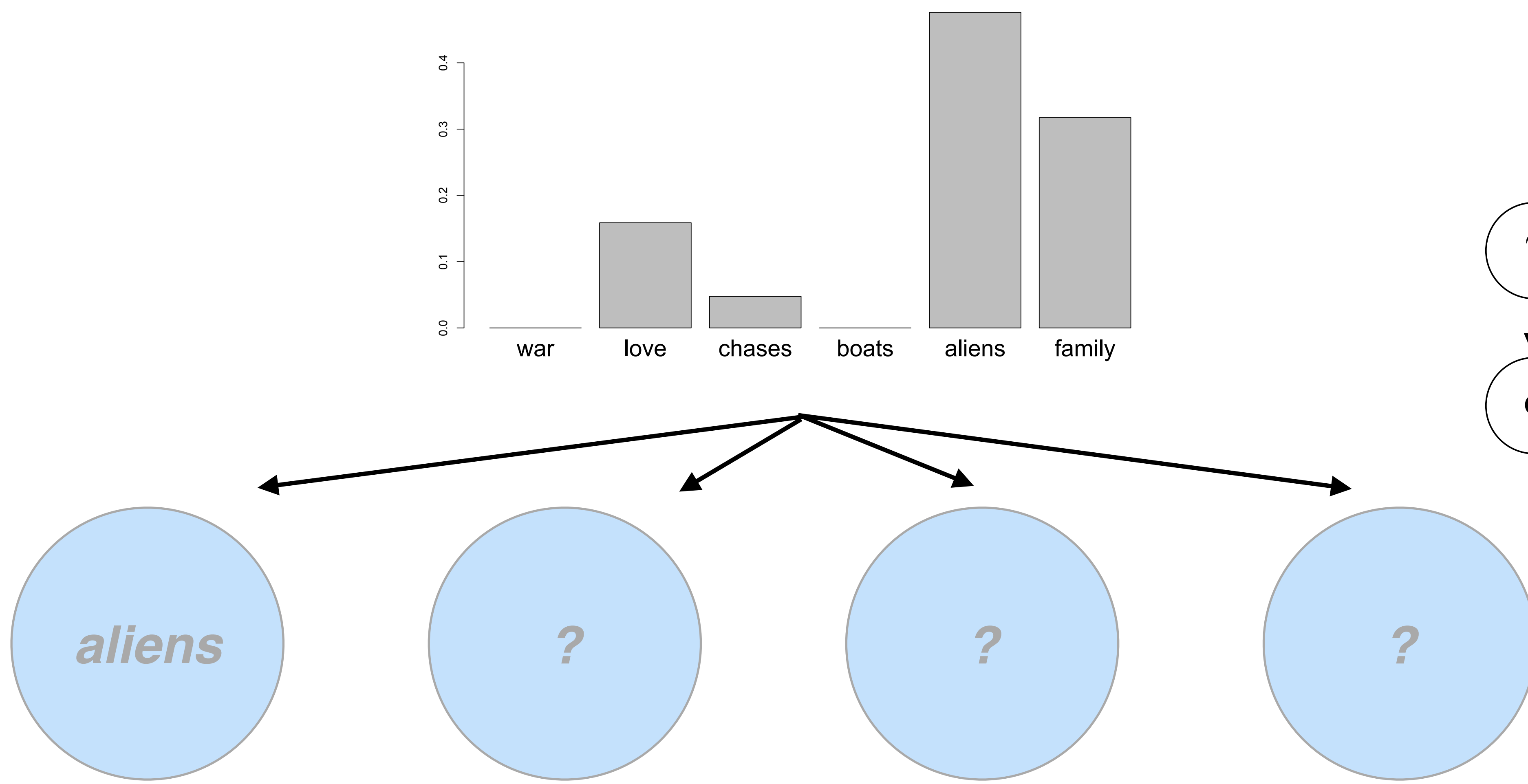


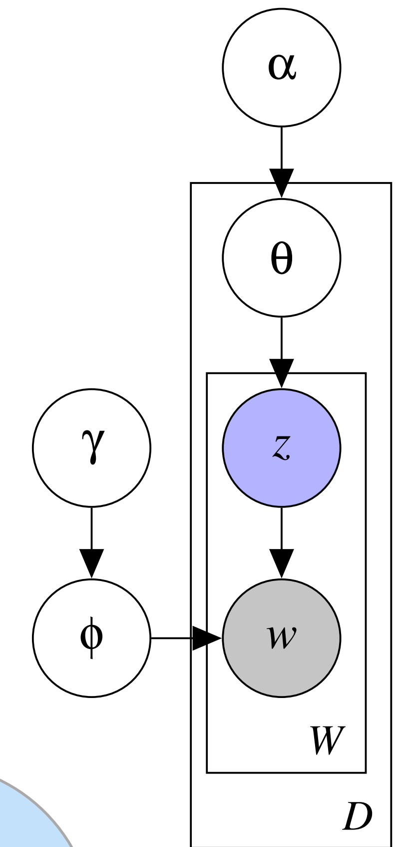
0.0
0.1
0.2
0.3
0.4
war
love
chases
boats
aliens
family
α
θ
γ
z
ϕ
w
W
D
war
aliens
war
love
"fights"
"alien"
"kills"
"marries"



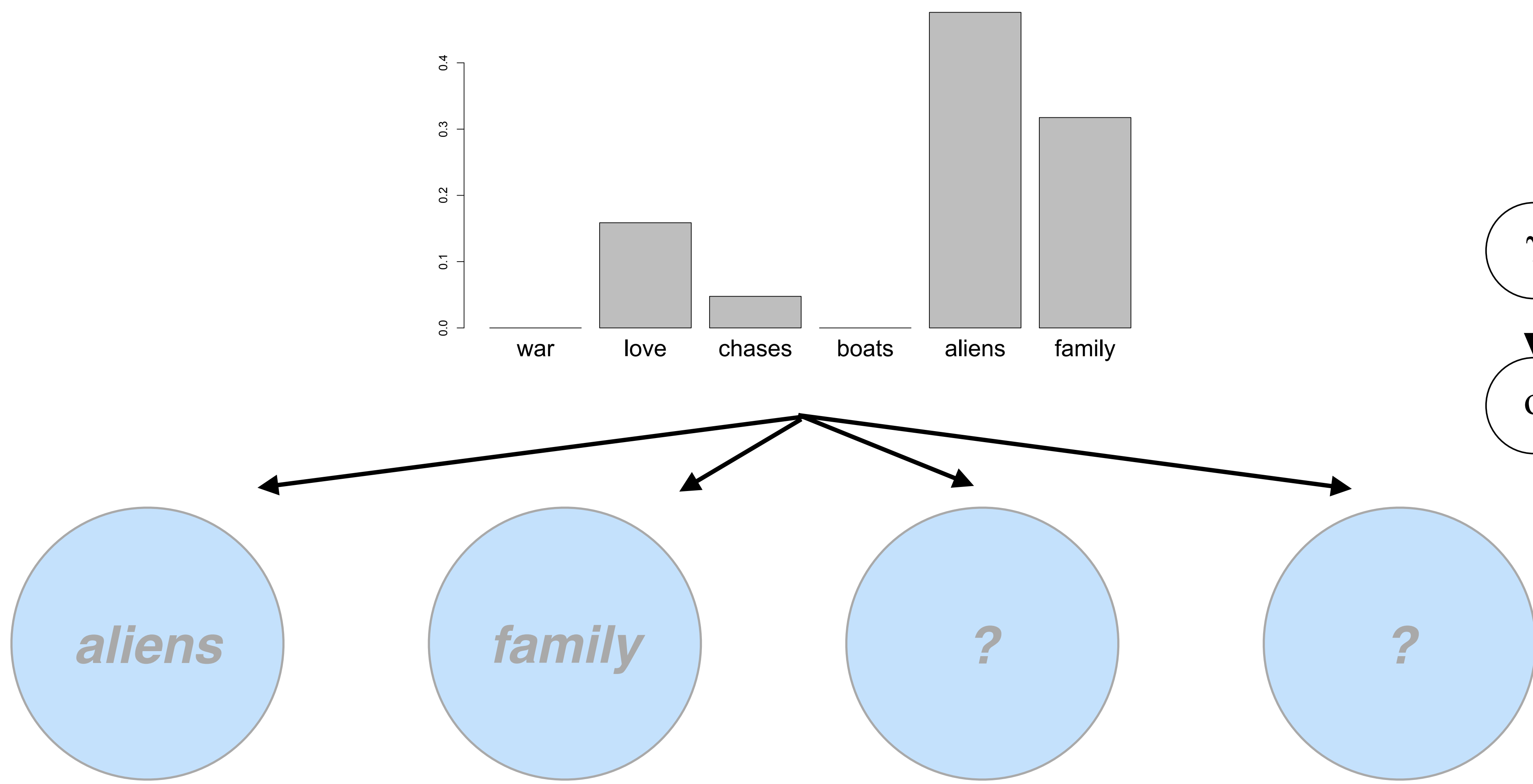


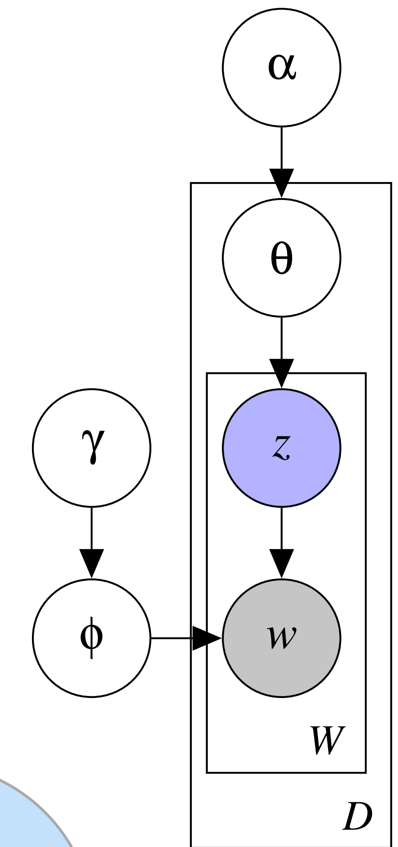
P(topic | topic distribution)



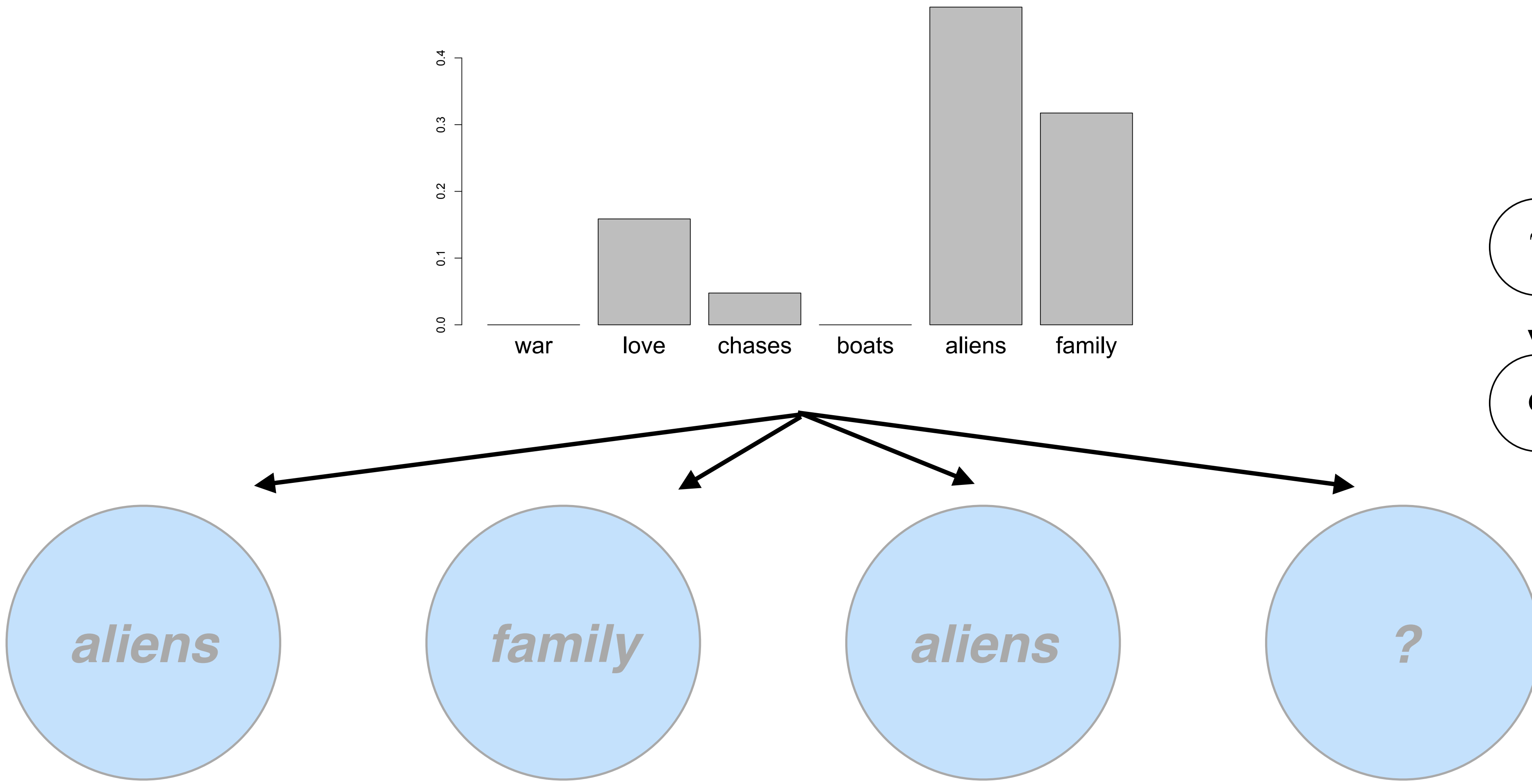


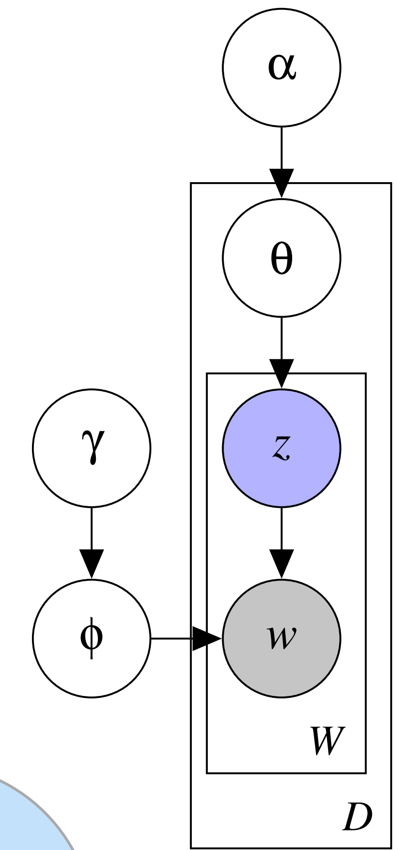
P(topic | topic distribution)



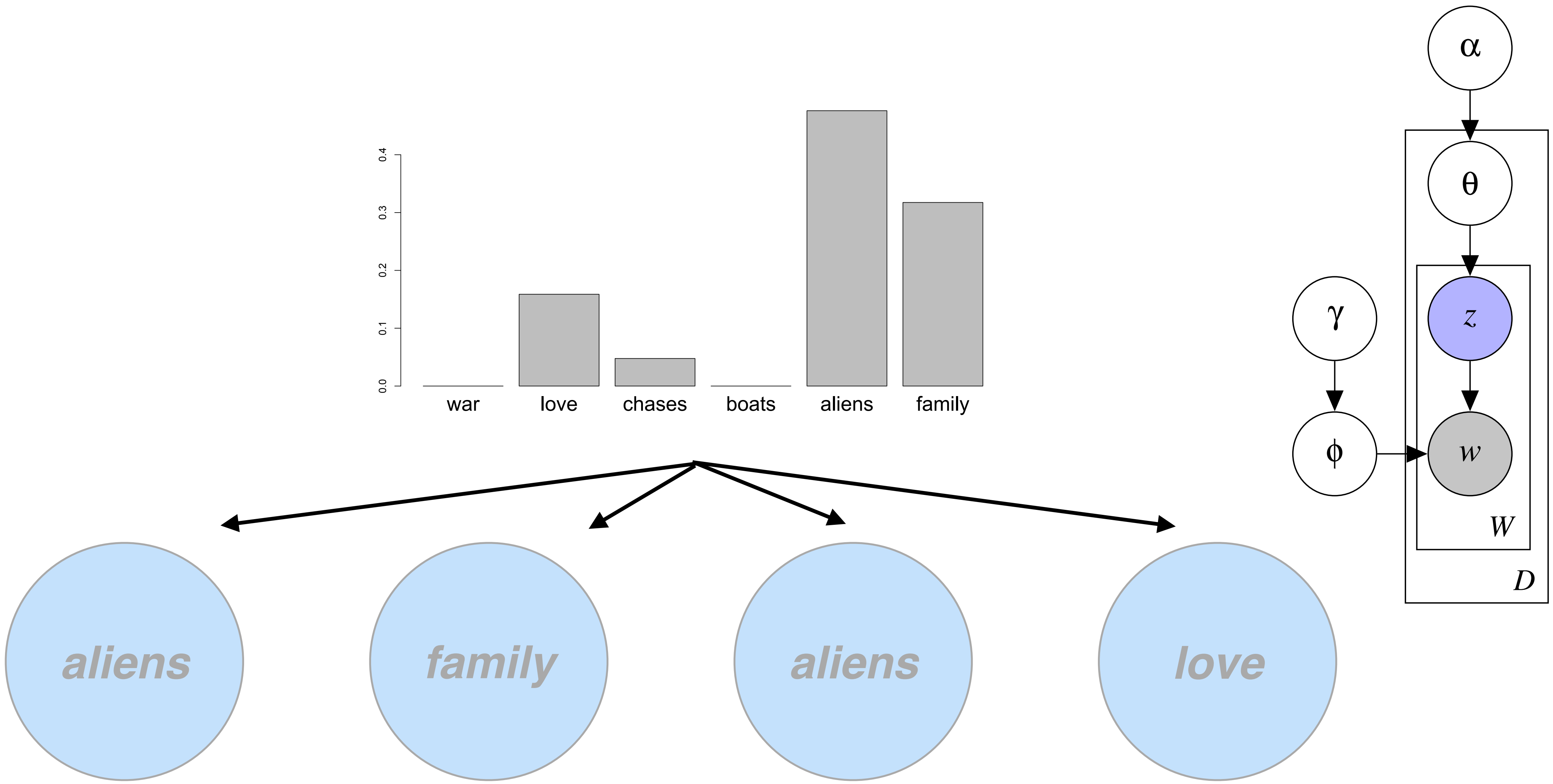


P(topic | topic distribution)



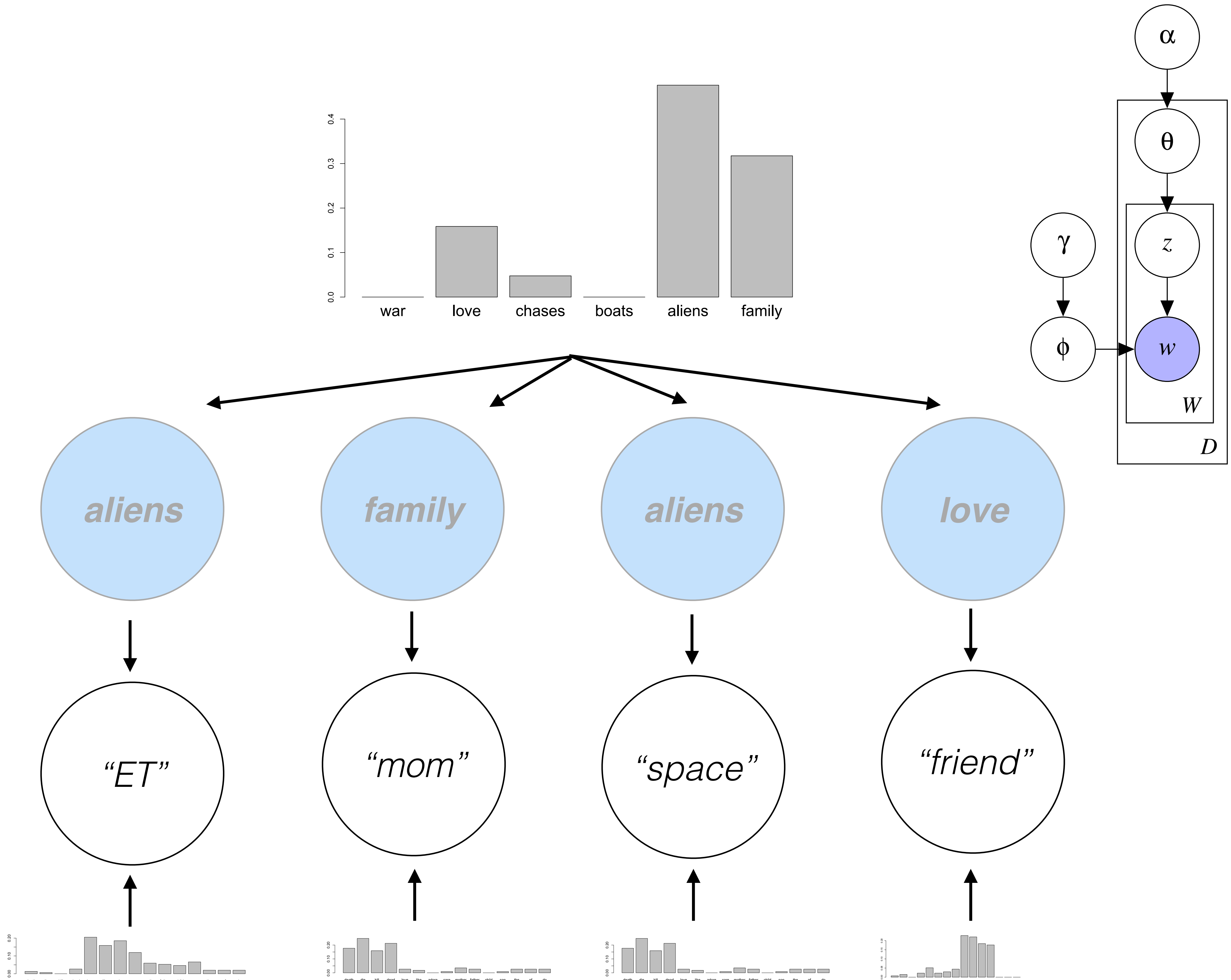


P(topic | topic distribution)



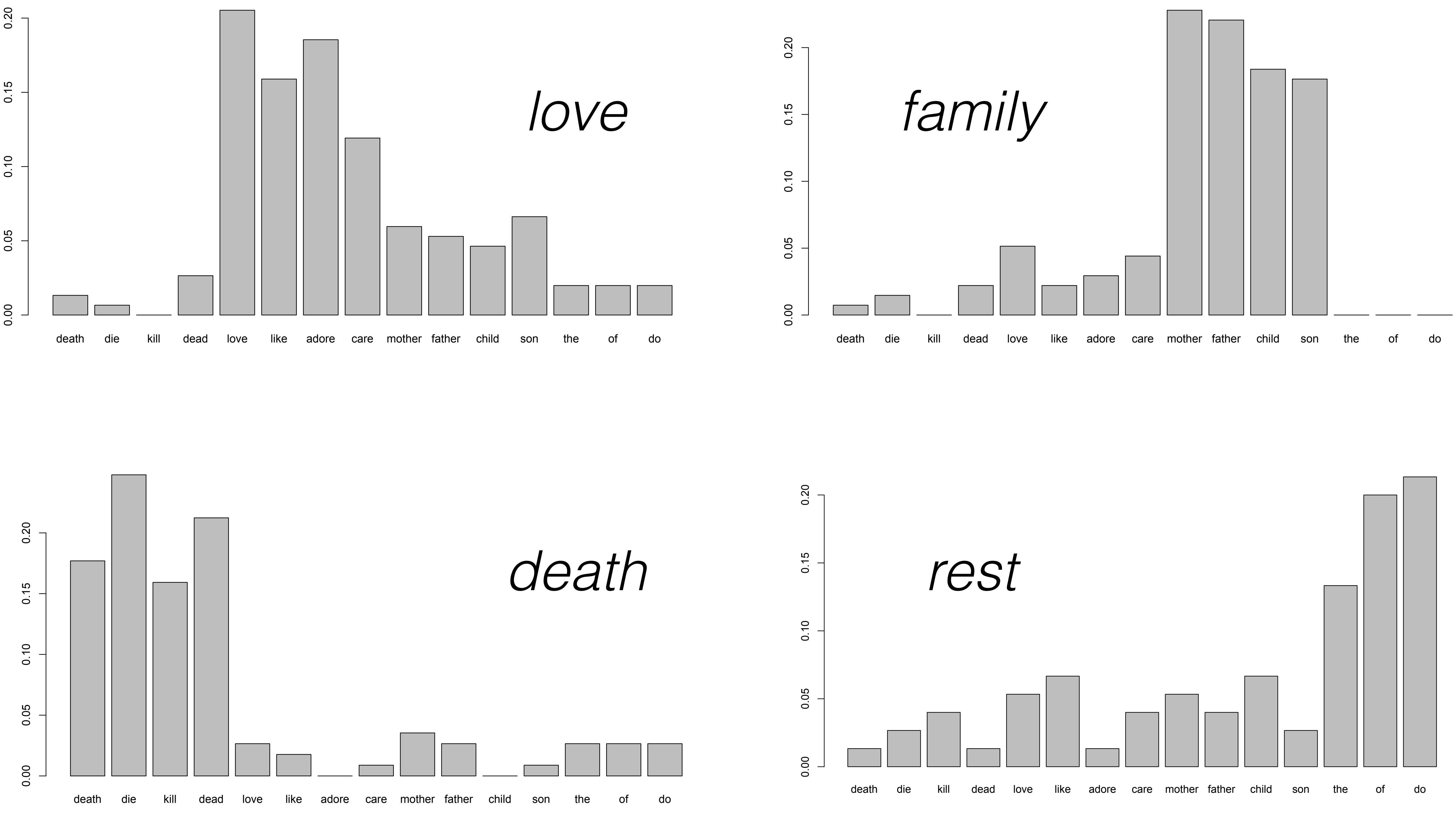
P(topic | topic distribution)

α
θ
γ
z
ϕ
w
W
D
war
love
chases
boats
aliens
family
aliens
family
aliens
love
?
?
?
?

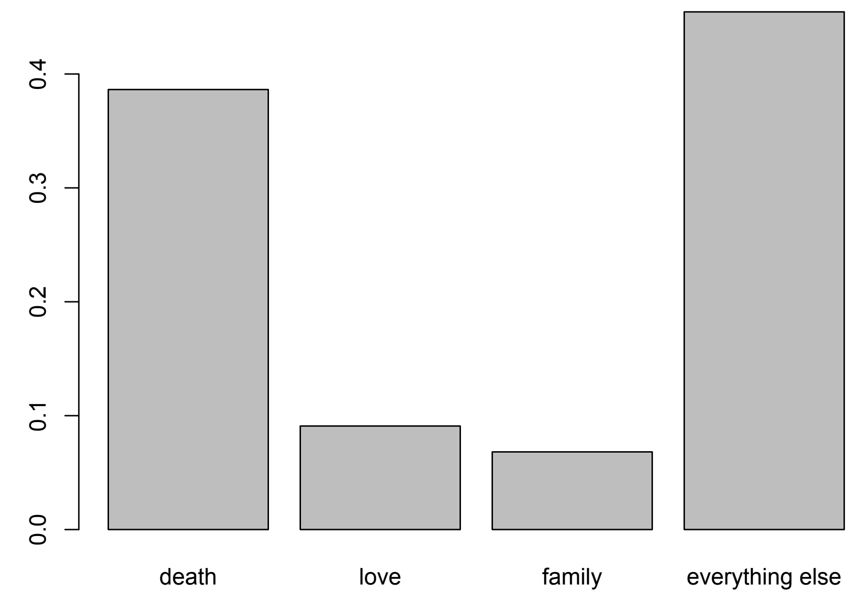


0.4
0.3
0.2
0.1
0.0
war
love
chases
boats
aliens
family
α
θ
γ
z
ϕ
w
W
D
aliens
family
aliens
love
"ET"
"mom"
"space"
"friend"

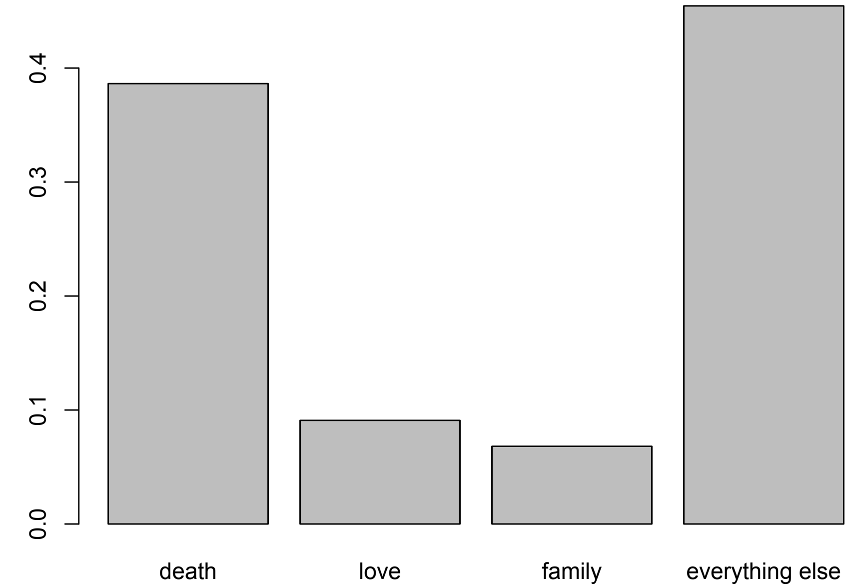
Romeo and Juliet



… 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."

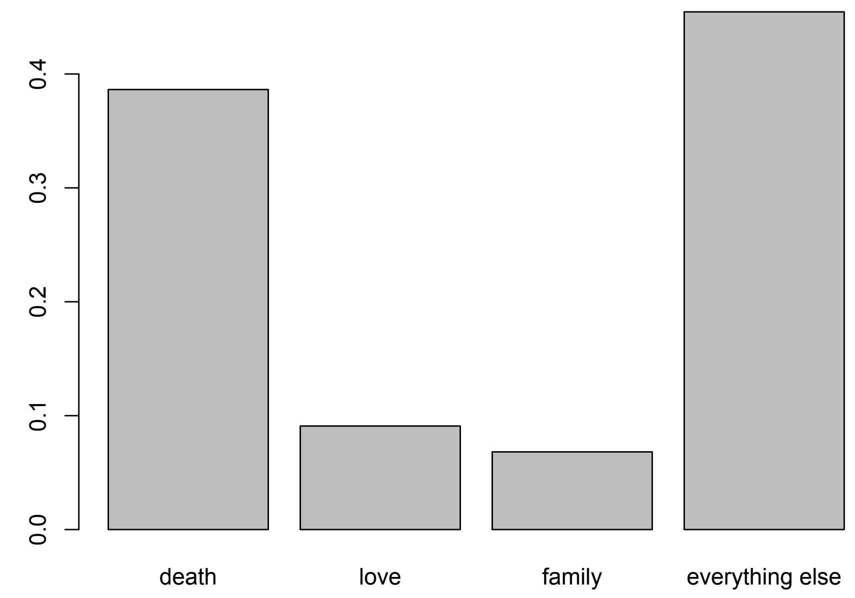


- DEATH
- LOVE
- FAMILY
- (EVERYTHING ELSE)



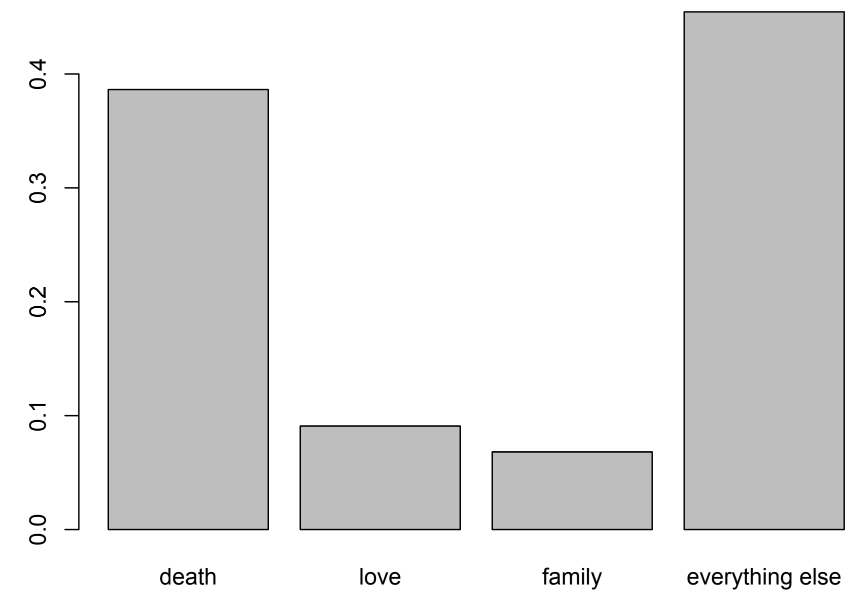
… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, 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
- LOVE
- FAMILY
- (EVERYTHING ELSE)



… 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
- LOVE
- FAMILY
- (EVERYTHING ELSE)

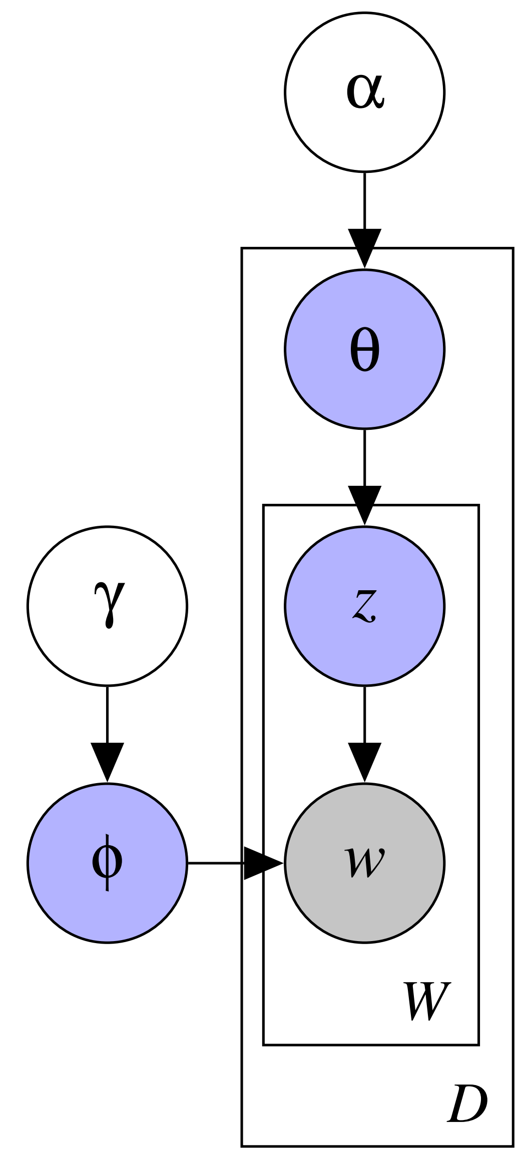


… The messenger, however, does not reach Romeo and, instead, Romeo learns of Juliet's apparent death from his servant Balthasar. Heartbroken, Romeo buys poison from an apothecary and goes to the Capulet crypt. He encounters Paris who has come to mourn Juliet privately. Believing Romeo to be a vandal, 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
- LOVE
- FAMILY
- (EVERYTHING ELSE)

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?

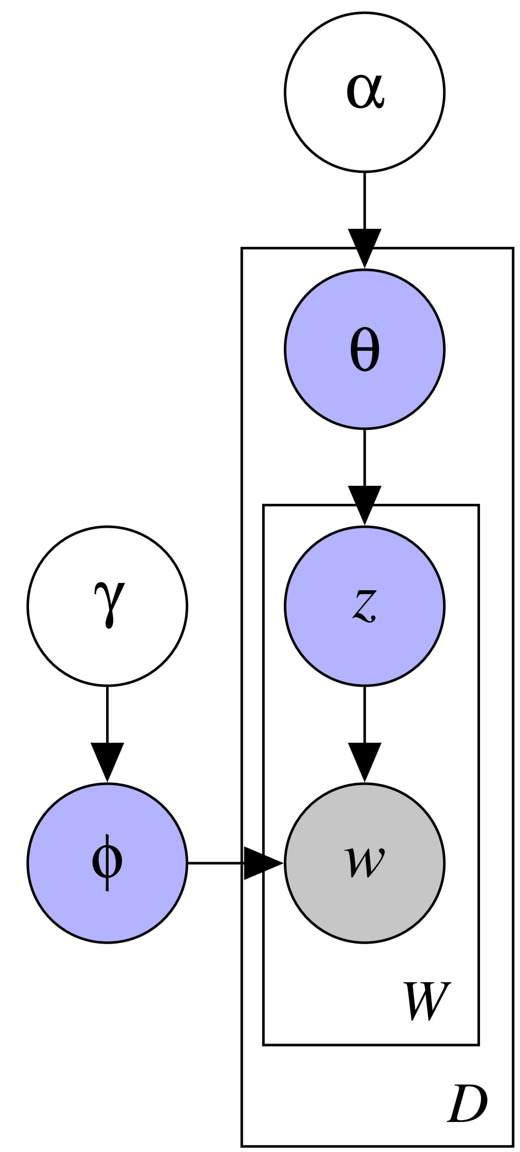


Find the parameters that maximize the likelihood of the data!

Gibbs Sampling

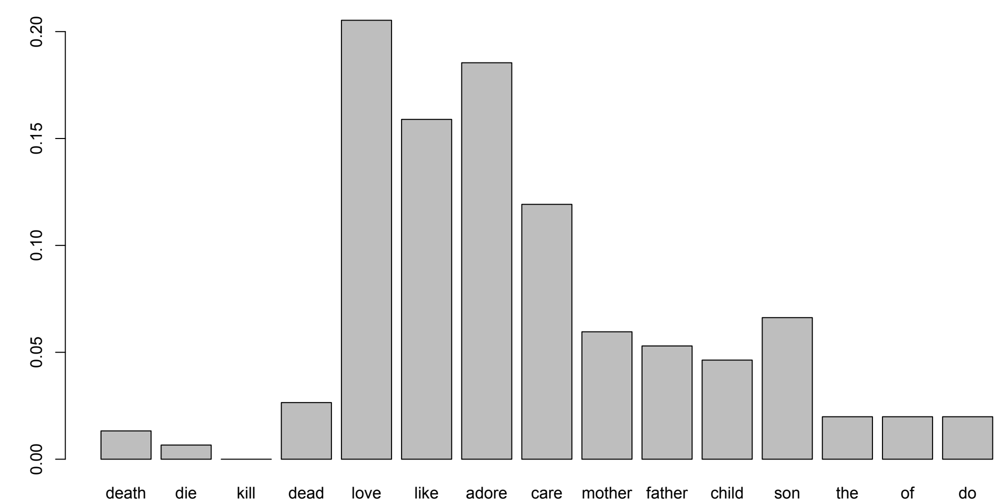
- 1. Start with some initial value for all the variables

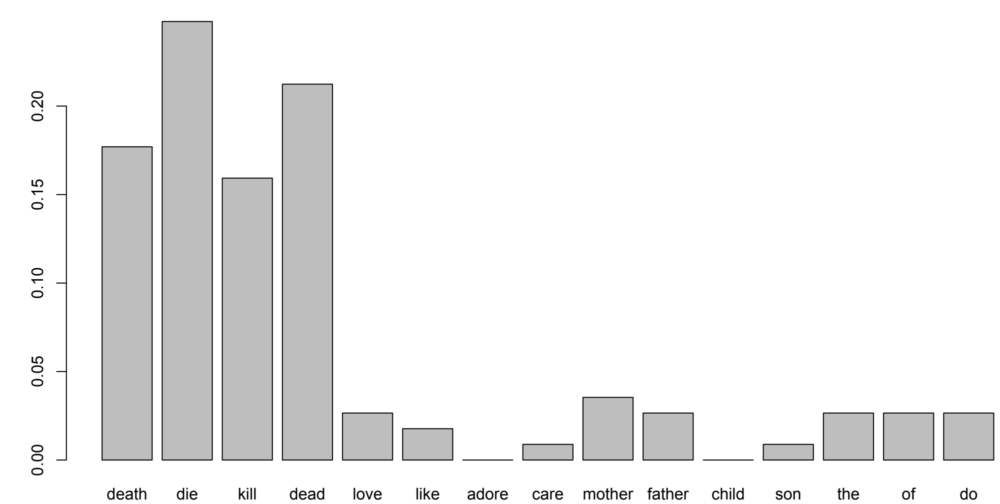
- 2. Sample a value for a variable conditioned on all of the other variables around it (using Bayes' theorem)



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





Examples

- Mining the Dispatch
  http://dsl.richmond.edu/dispatch/

- Wikipedia Topics
  http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-list.html

- Quiet Transformations of Literary Studies
  http://www.rci.rutgers.edu/~ag978/quiet/

Try it yourself

- book summaries, movie summaries, PMLA , Classical Quarterly, Renaissance Quarterly, Shakespeare + English stoplist http://bit.ly/1hdKX0R

- Topic Modeling Tool https://code.google.com/p/topic-modeling-tool/

Representation Learning

Assume we've trained a logistic regression classifier to predict whether a tweet was written by a person who lives in Chicago.

$\beta_{Chicago}$

i	feat	value
1	I	0.004
2	live	0.0013
3	in	-0.001
4	New York	-13.7
5	Chicago	8.7
6	Boston	-10.8
7	Pittsburgh	-5.7
8	snow	2.7

Representation Learning

"I live in Chicago"

$\beta_{Chicago}$ =

i	feat	value
1	I	0.004
2	live	0.0013
3	in	-0.01
4	New York	-13.7
5	Chicago	8.7
6	Boston	-10.8
7	Pittsburgh	-5.7
8	snow	2.7

x =

i	feat	value
1	I	1
2	live	1
3	in	1
4	New York	0
5	Chicago	1
6	Boston	0
7	Pittsburgh	0
8	snow	0

Representation Learning

"I live in Chicagoland"

$\beta_{\text{Chicago}} =$

i	feat	value
1	I	0.004
2	live	0.0013
3	in	-0.01
4	New York	-13.7
5	Chicago	8.7
6	Boston	-10.8
7	Pittsburgh	-5.7
8	snow	2.7

$x =$

i	feat	value
1	I	1
2	live	1
3	in	1
4	New York	0
5	Chicago	0
6	Boston	0
7	Pittsburgh	0
8	snow	0

Representation Learning

- Learn alternate representations for inputs (and sometimes outputs) aside from their raw (atomic) values.

- For words, this generally means representations that encode some measure of similarity.

    - Hard word clusters (e.g., Brown clusters
    - Low-dimensional "embeddings" ($w \in \mathbb{R}^K$)

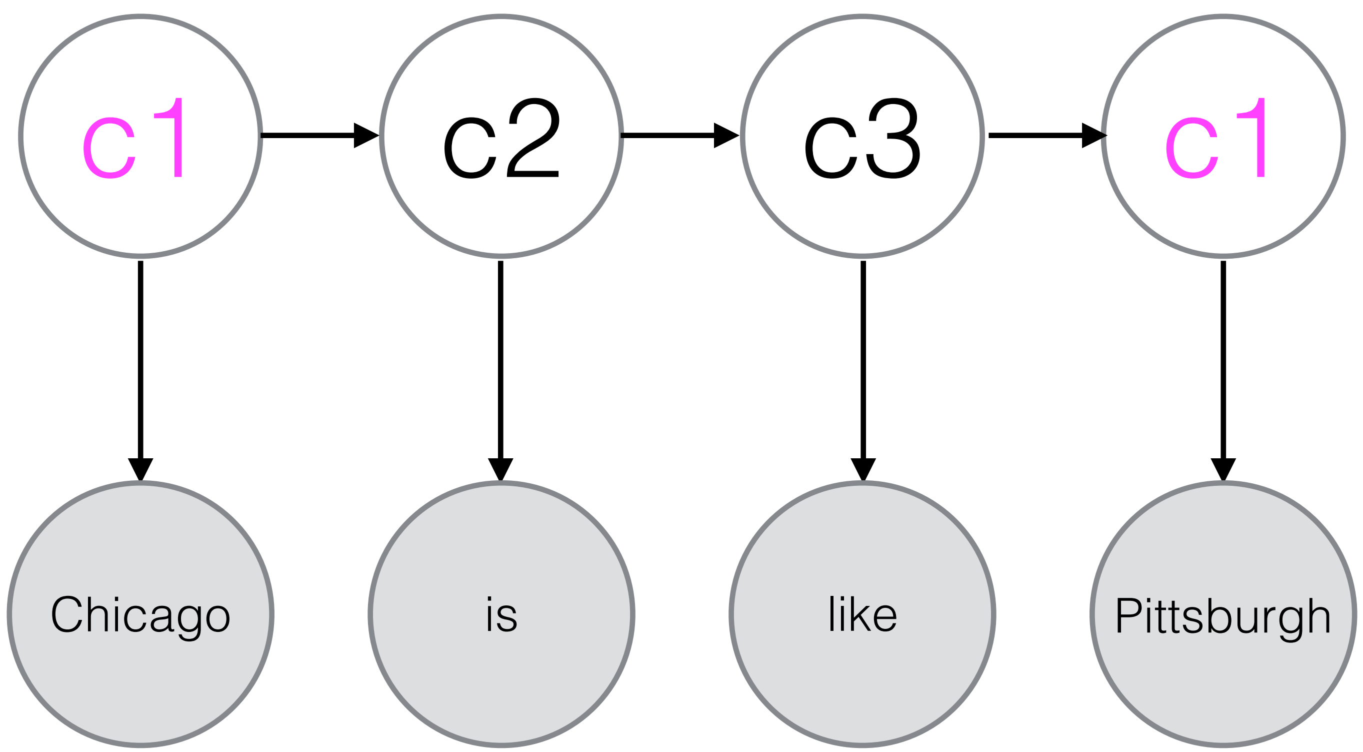
Representation Learning

“You shall know a word by the company it keeps” (Firth 1957)

- my boy’s wicked smart
- my boy’s hella smart
- my boy’s very smart
- my boy’s extremely smart
- my boy’s ________ smart

Brown clustering

Unsupervised HMM, where each word type belongs to one class.



Brown et al. (1992), “Class-Based n-gram Models of Natural Language” (Computational Linguistics)

#mlch

Brown clustering

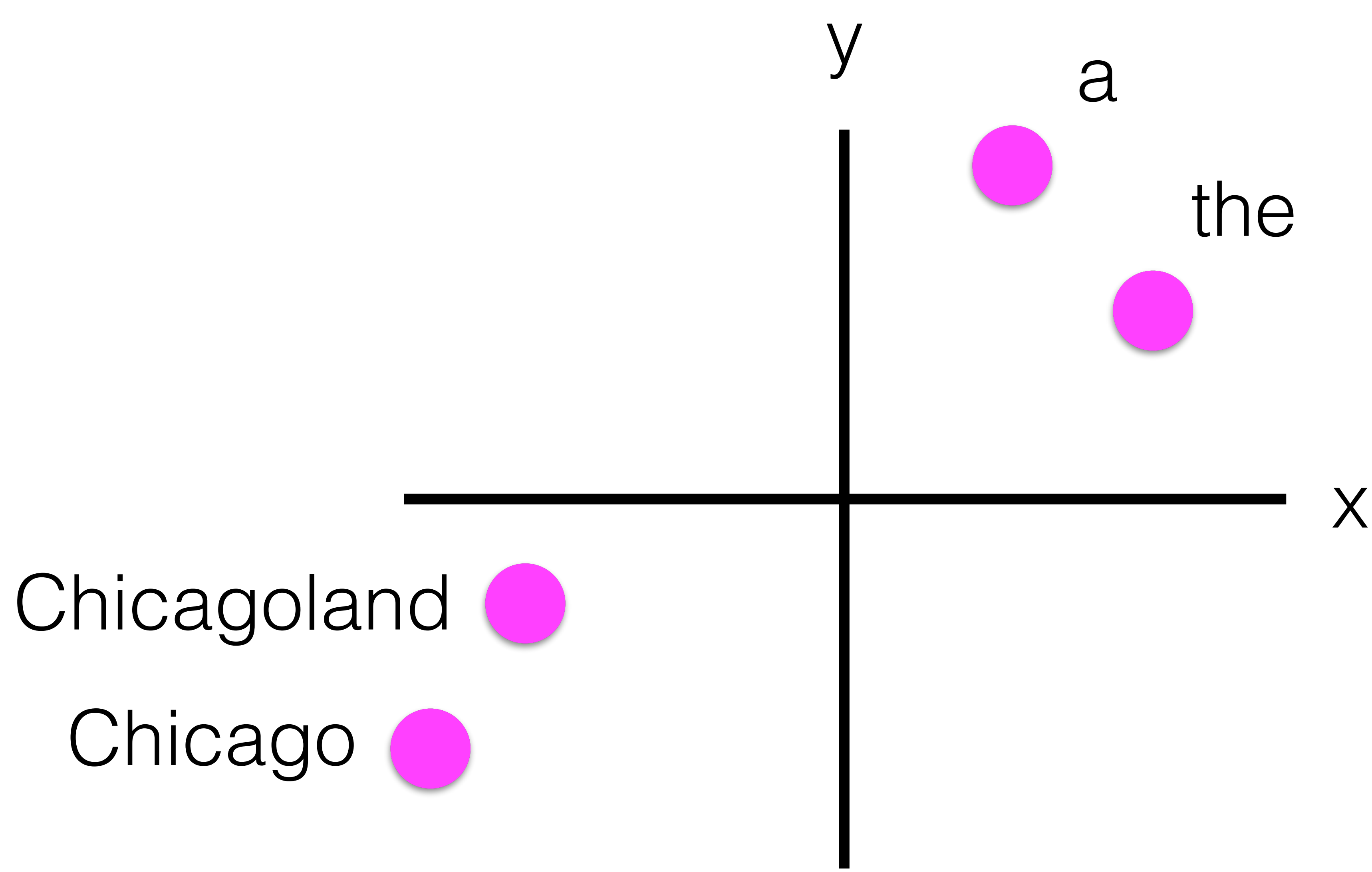
- Demo: 1000 clusters learned from 56M tweets
- http://www.ark.cs.cmu.edu/TweetNLP/cluster_viewer.html
- Code: https://github.com/percyliang/brown-cluster

Embeddings

- Represent each word in your vocabulary as a vector of *K* numbers

	x	y
the	2.1	2.5
a	1.5	3.7
Chicago	-3.0	-3.4
Chicagoland	-2.6	-0.5

Embeddings



Embeddings

- Basic intuition: use a K-dimensional embedding for a word in a sentence to predict all of the words around it; find the value of the embedding to maximize your predictive accuracy.

3.1
1.7

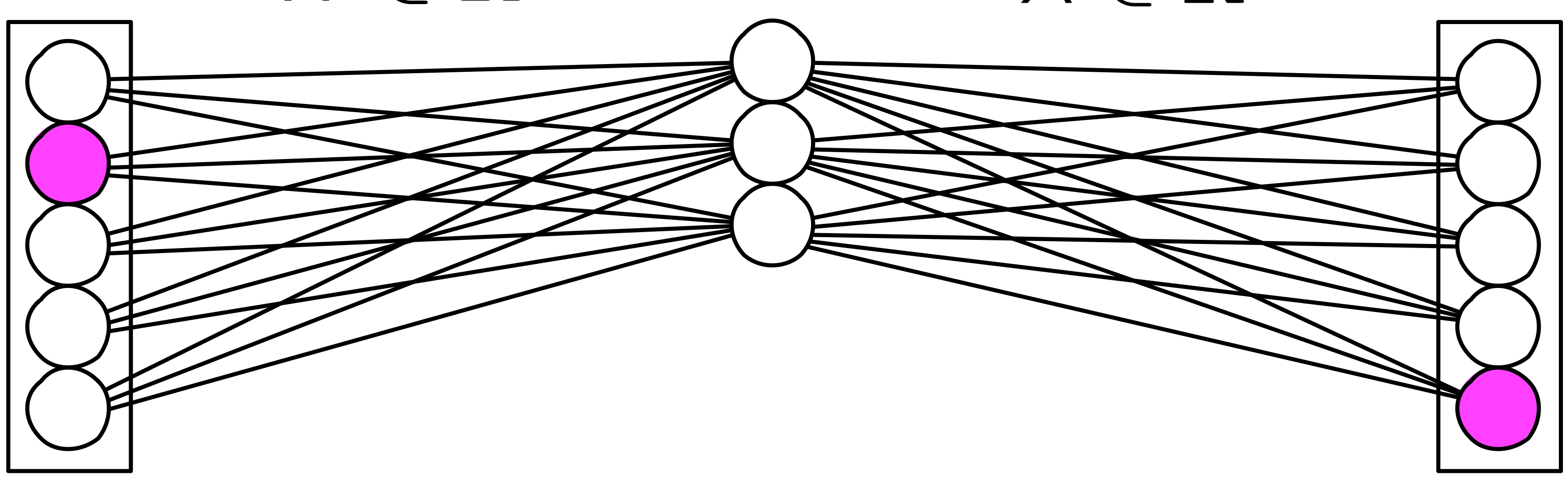
Let's go to the ______ to buy some eggs.

Skip-Gram Embeddings

$$h_i = x^\top W_i$$

$$P(word = w | x, h, X) = \frac{\exp(h^\top X_w)}{\sum_v^V \exp(h^\top X_v)}$$

$W \in \mathbb{R}^{F \times K}$ $X \in \mathbb{R}^{K \times V}$



"store" "buy"

Mikolov et al. (2013), "Efficient Estimation of Word Representations in Vector Space," ICLR.

Embeddings

- Demo: http://radimrehurek.com/2014/02/word2vec-tutorial/#app
- Code: https://code.google.com/p/word2vec/

Word Representations

What do you do with word representations?

Word Representations

brown:169	brown:170	brown:171
Mr.	Chicago	New York
Mrs.	Chicagoland	NYC
	Chitown	NY

Word Representations

“I live in Chicago”

i	feat	value
1	I	1
2	live	1
3	in	1
4	New York	0
5	**Chicago**	**1**
6	Boston	0
7	Pittsburgh	0
8	snow	0
9	**brown:170**	**1**

“I live in Chicagoland”

i	feat	value
1	I	1
2	live	1
3	in	1
4	New York	0
5	**Chicago**	**0**
6	Boston	0
7	Pittsburgh	0
8	snow	0
9	**brown:170**	**1**

NLP and beyond

subject object

syntactic parsing

Ben: Have you importun'd him ?

VB PRP VBN PRP

POS tagging

have you importune he ?

lemmatization

Montague Romeo

coreference resolution

#mlch

NLP and beyond

subject object

Ben: Have you importun'd him ? 90%

VB PRP VBN PRP 98%

have you importune he ? 98%

Montague Romeo 70%

NLP toolkits

- Tokenization, part of speech tagging, syntactic parsing, named entity recognition, coreference resolution.
- CoreNLP
  http://nlp.stanford.edu/software/corenlp.shtml
- BookNLP
  https://github.com/dbamman/book-nlp
- NLTK
  http://www.nltk.org

Thanks!

- David Bamman
  dbamman@cs.cmu.edu