Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 10: Latent variable models

Feb 24, 2016

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, ∞]

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

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

$P(X = x)$

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

	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

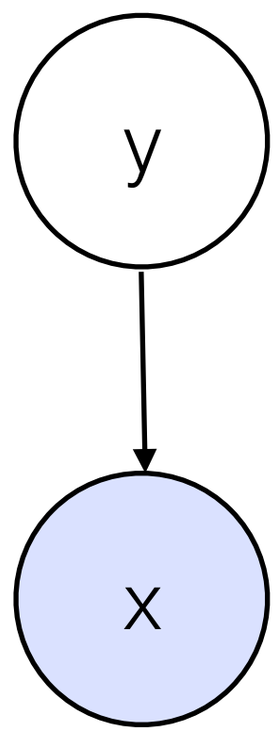
- A latent variable is one that's unobserved, either because:
  - we are predicting it (but have observed that variable for other data points)
  - it is unobservable

Latent variables

	observed variables	latent variables
email	text, date, sender	
novels		
social network		
fitbit data		
legislators		
netflix users		

Probabilistic graphical models

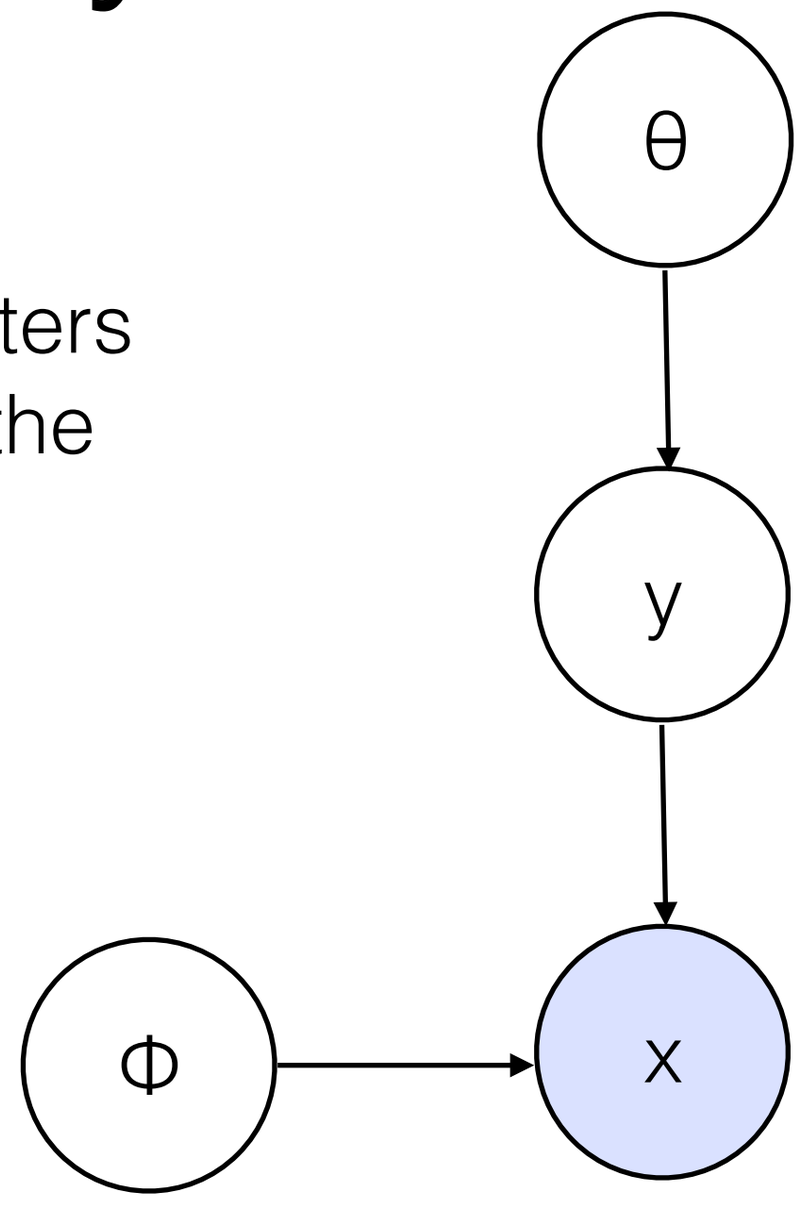
- Nodes represent variables (shaded = observed, clear = latent)
- Arrows indicate conditional relationships
- The probability of x here is dependent on y
- Simply a visual way of writing the joint probability:



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

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).



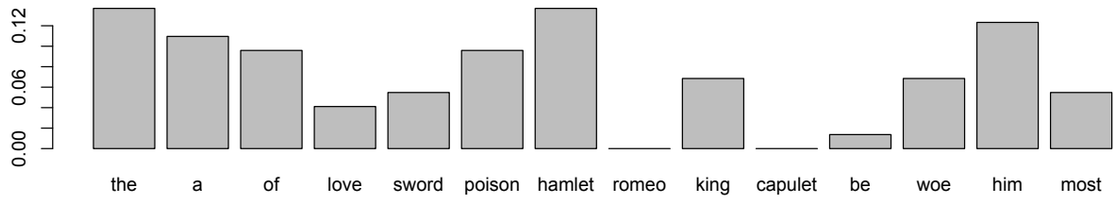
$$P(x, y \mid \theta, \phi) = P(y \mid \theta)\, P(x \mid y, \phi)$$

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

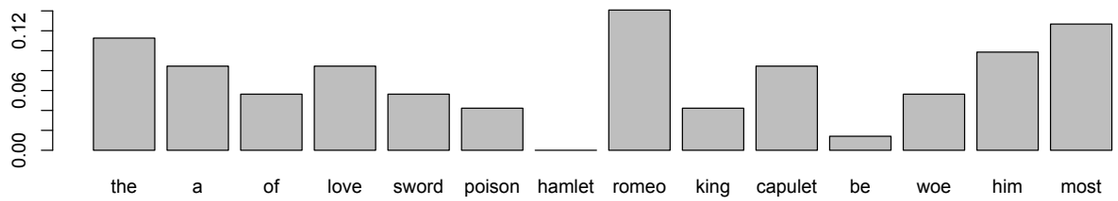
Look up the value of y in θ

θ

Φ_{austen}



$\Phi_{dickens}$

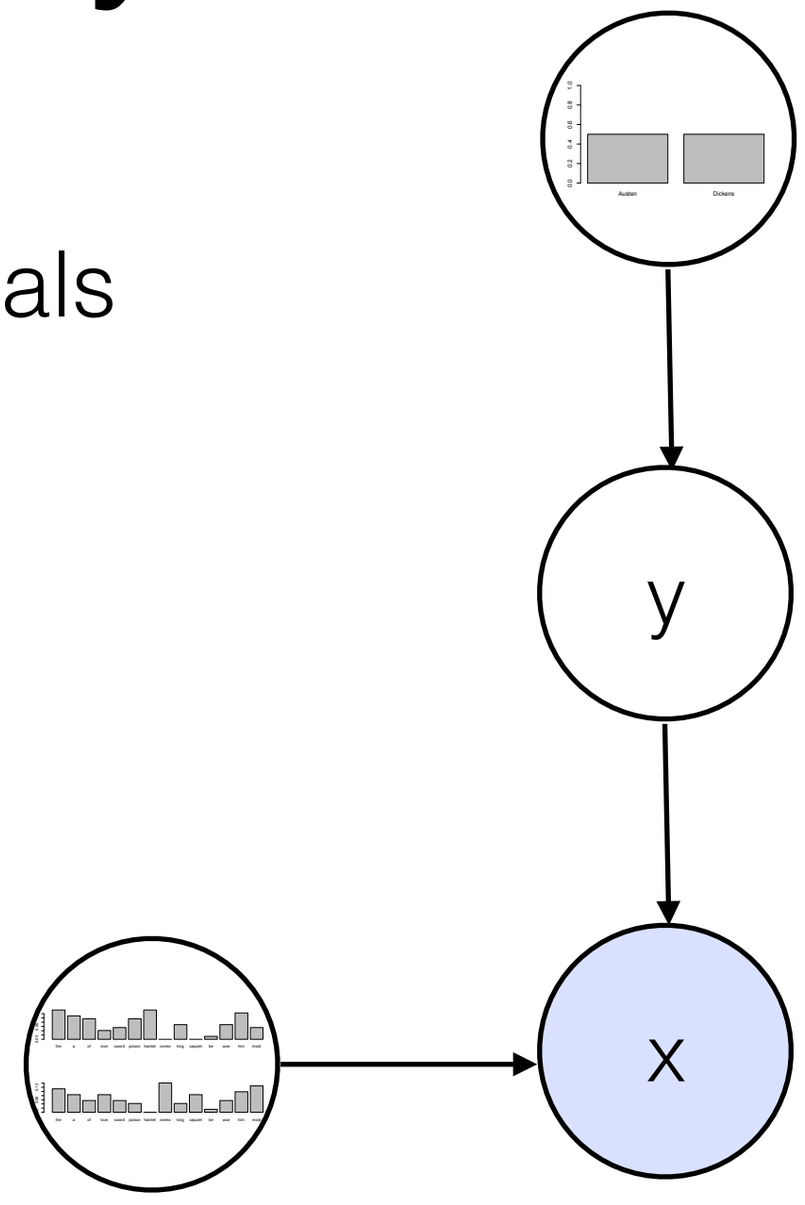


$$P(x = \text{love} \mid y = \text{Austen}, \phi) = 0.04$$

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



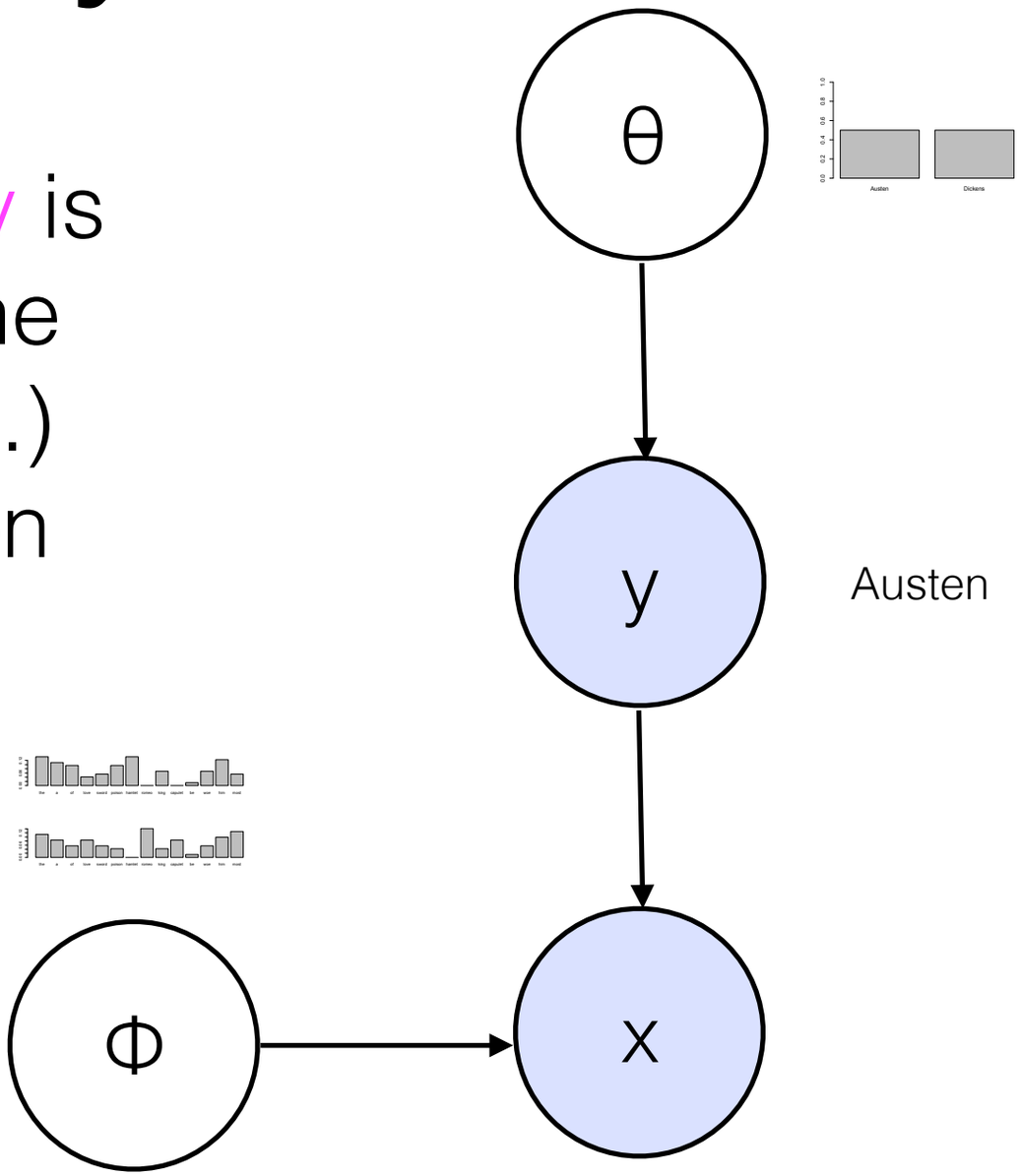
$$P(x, y \mid \theta, \phi) = P(y \mid \theta)\, P(x \mid y, \phi)$$

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

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

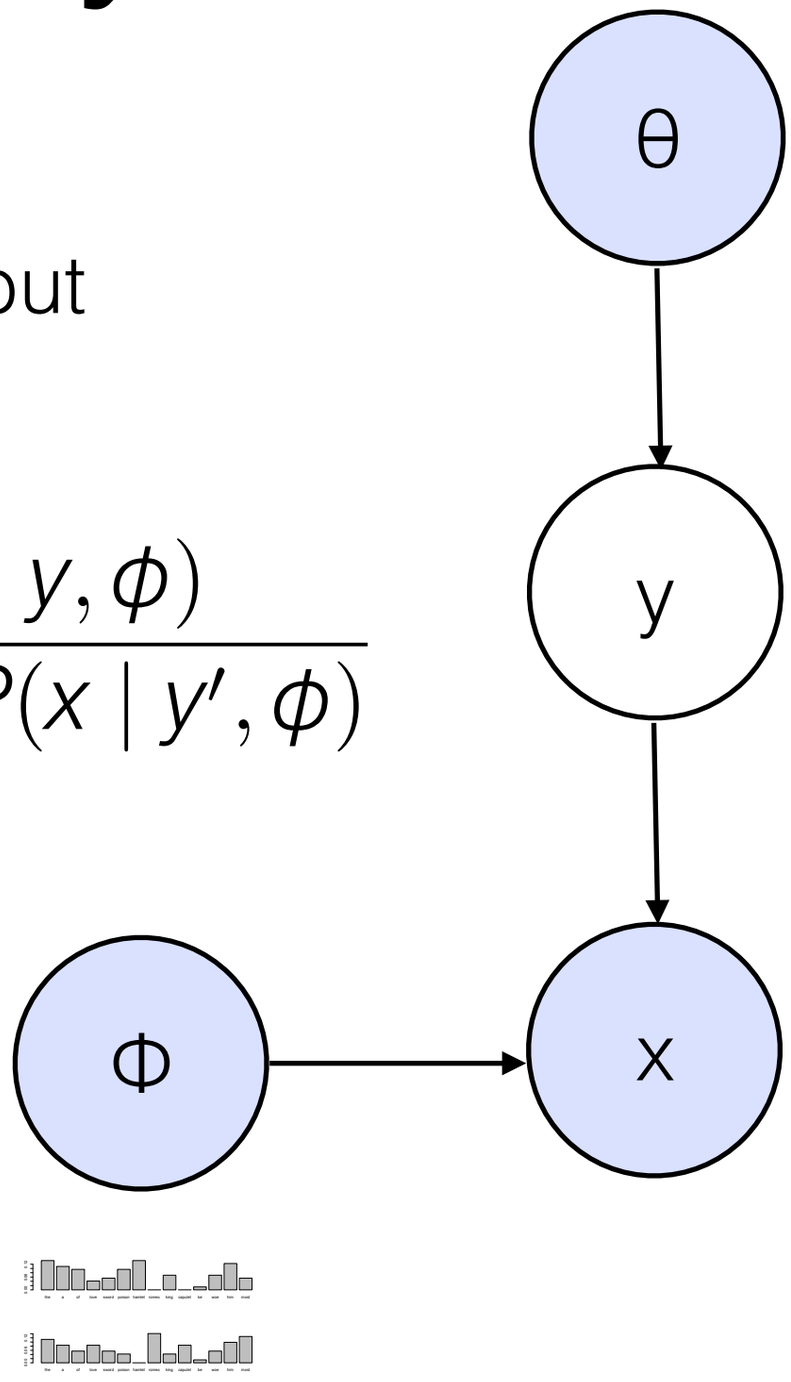
The number of times “love” appears in Austen texts divided by the total number of words in Austen texts

Naive Bayes

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

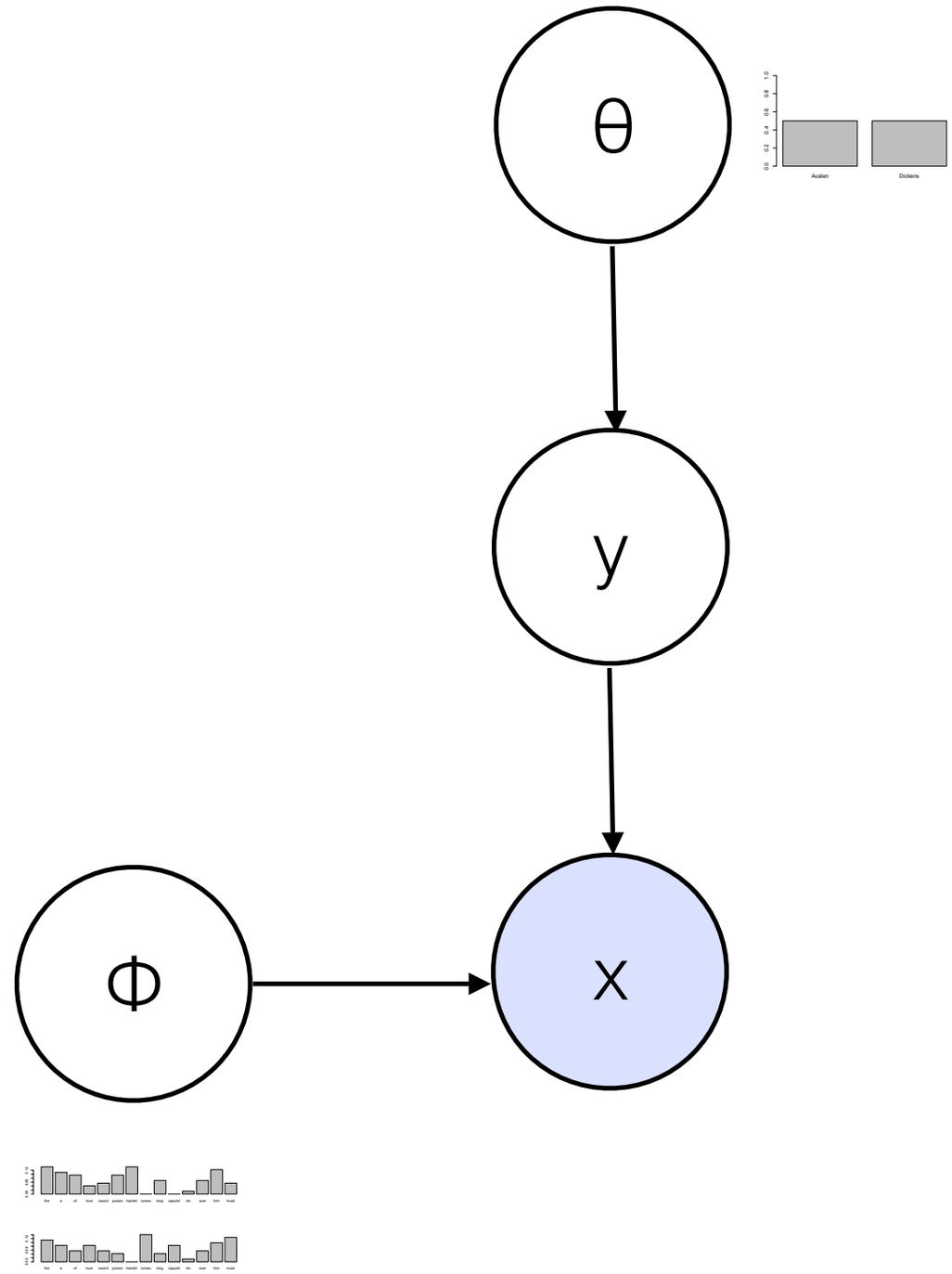
$$P(y \mid x, \theta, \phi) = \frac{P(y \mid \theta)P(x \mid y, \phi)}{\sum_{y' \in \mathcal{Y}} P(y' \mid \theta)P(x \mid y', \phi)}$$

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



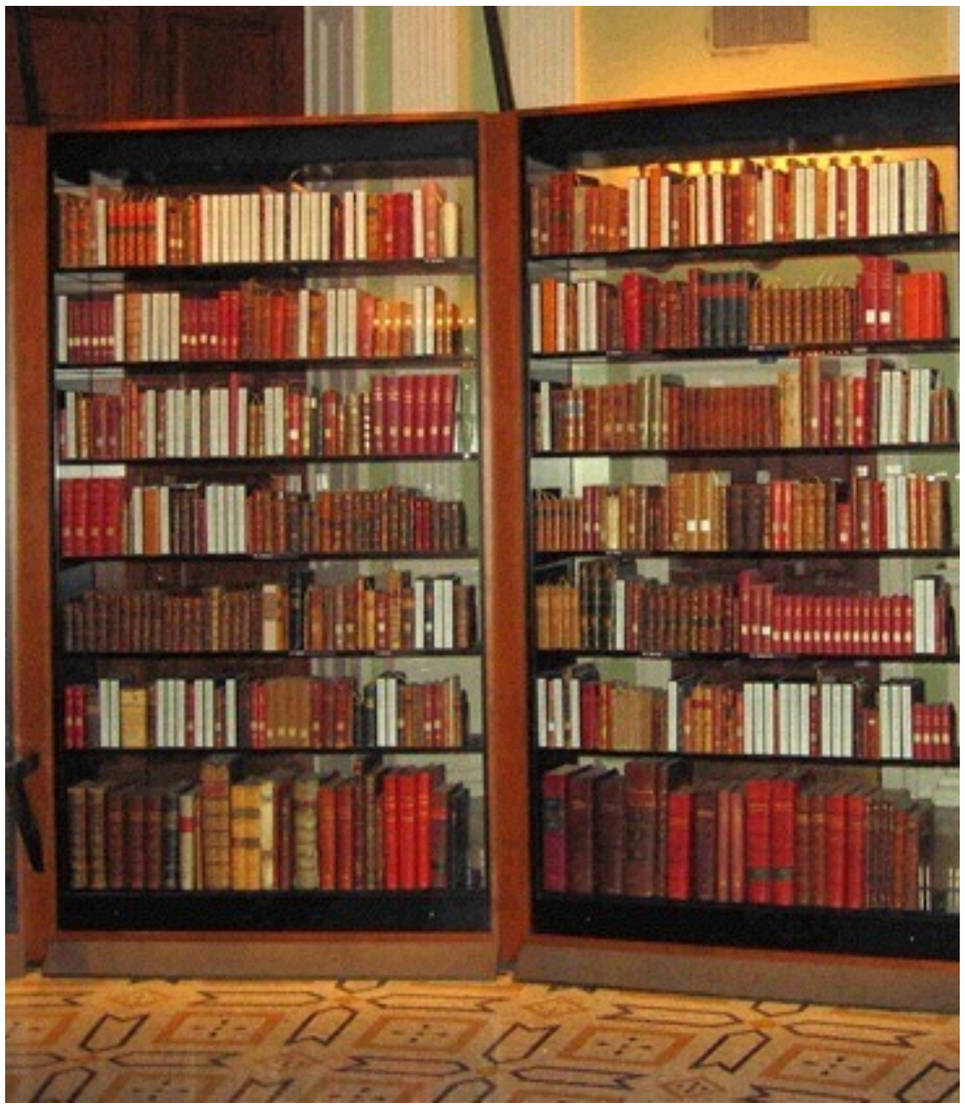
Unsupervised Naive Bayes

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



Structure

- Unsupervised learning finds *structure* in data.



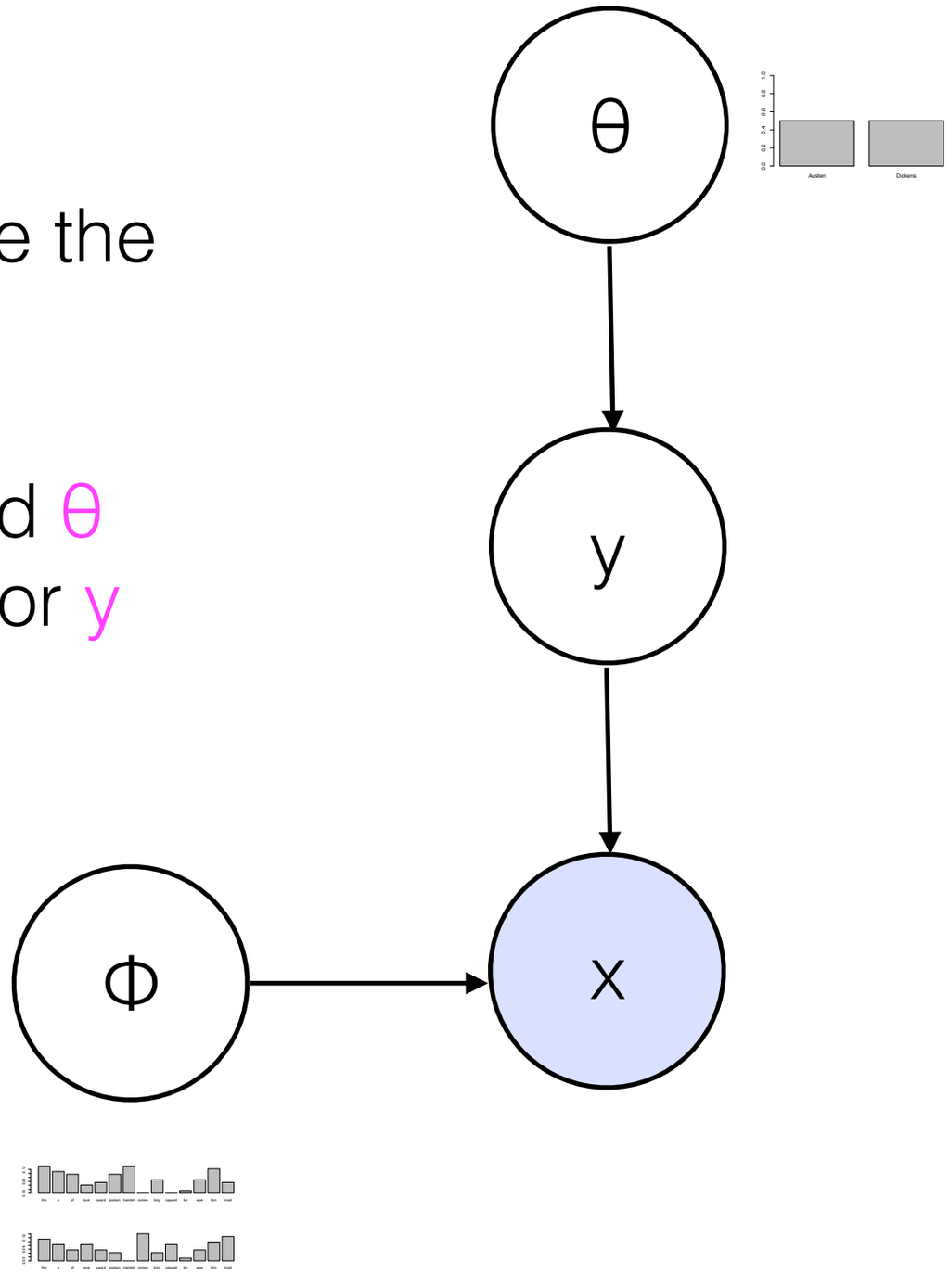
Unsupervised Naive Bayes

- 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

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

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

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

	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

$$P(Y = y) = \sum_{x \in \mathcal{X}} P(X = x, Y = y)$$

$$= \sum_{x \in \mathcal{X}} P(X = x) P(Y = y \mid X = x)$$

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

	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

$$= \sum_{x \in \mathcal{X}} P(X = x) P(Y = y \mid X = x)$$

$$P(Y = \text{ice cream}) = \underbrace{\frac{3}{7}\frac{3}{3}}_{\text{hot}} + \underbrace{\frac{2}{7}\frac{1}{2}}_{\text{cloudy}} + \underbrace{\frac{2}{7}\frac{0}{3}}_{\text{rainy}} = \frac{4}{7}$$

Inference

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

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

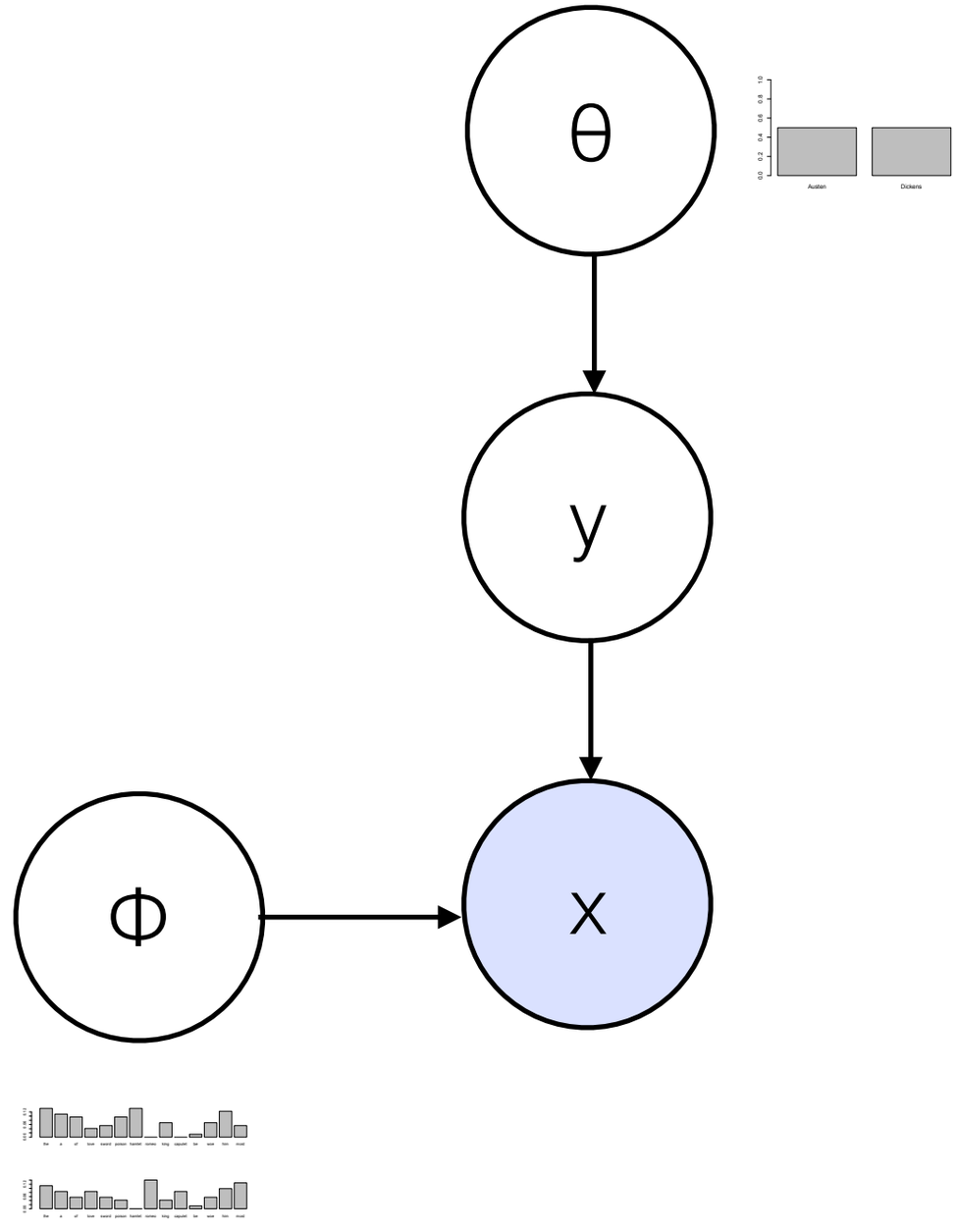
this sum in the log makes this likelihood hard to optimize

Inference

Lots of standard inference techniques we can use

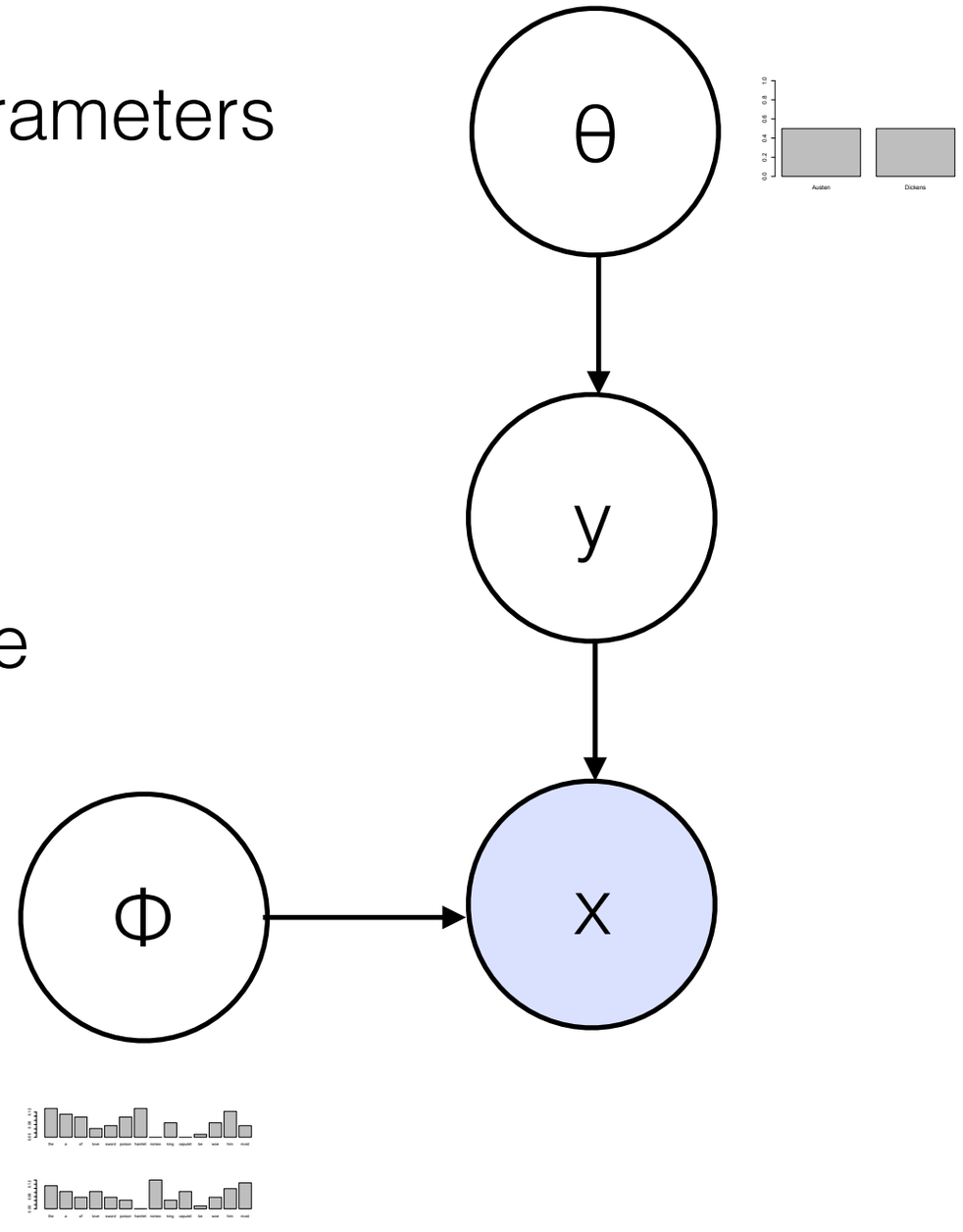
- Expectation Maximization
- Markov chain Monte Carlo (Gibbs sampling, Metropolis Hastings, etc.)
- Variational methods
- Spectral methods (Anandkumar et al. 2012, Arora et al. 2013)

Expectation Maximization



Expectation Maximization

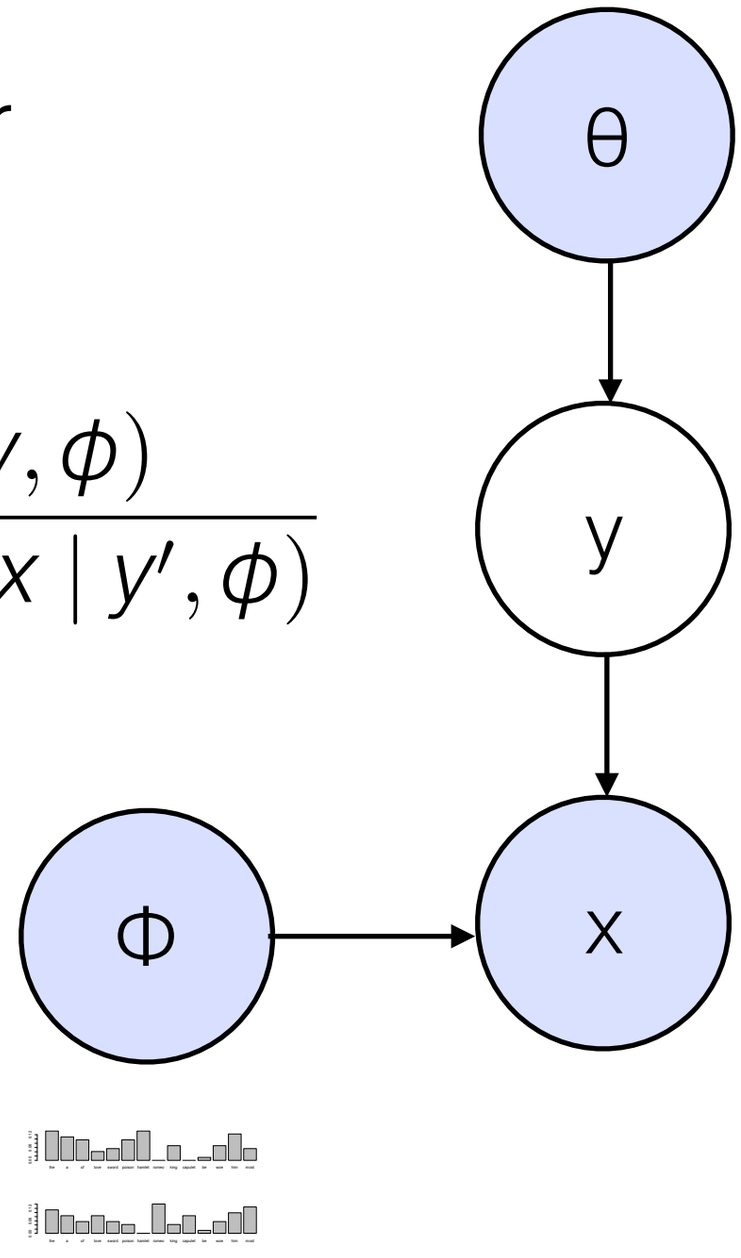
- Start out with random values for the parameters
- Iterate until convergence:
  - Calculate expected values for latent variables y
  - Use those expected values to update parameters Φ and θ



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

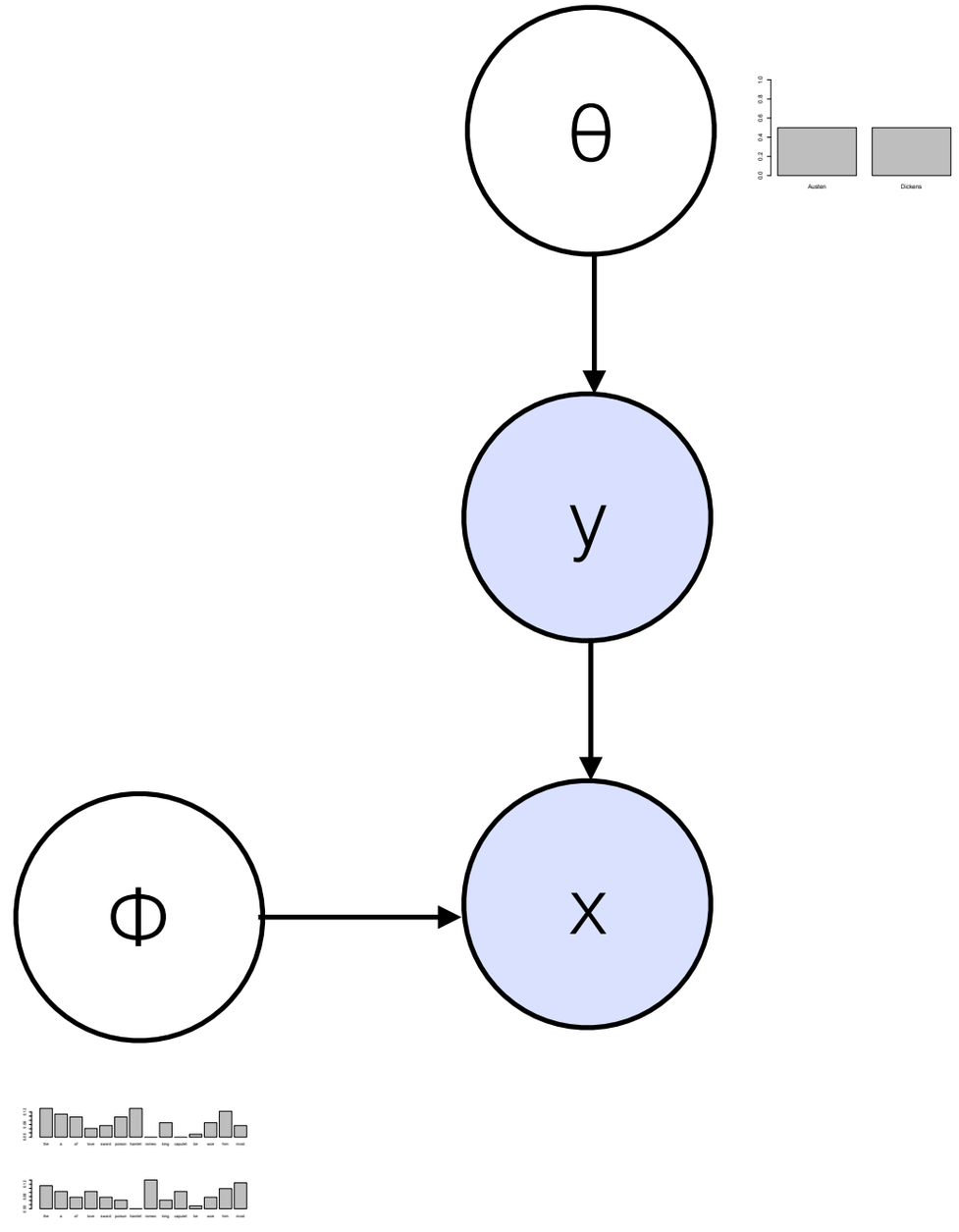
Expectation Maximization

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				
	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

i

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

i \ k	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

1 Given: a set $\mathcal{X} = \{\vec{x}_1, \ldots, \vec{x}_n\} \subseteq \mathbb{R}^m$
2 a distance measure $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$
3 a function for computing the mean $\mu : \mathcal{P}(\mathbb{R}) \to \mathbb{R}^m$
4 Select k initial centers $\vec{f}_1, \ldots, \vec{f}_k$
5 **while** stopping criterion is not true **do**
6 **for** all clusters c_j **do**
7 $c_j = \{\vec{x}_i \mid \forall \vec{f}_l \, d(\vec{x}_i, \vec{f}_j) \leq d(\vec{x}_i, \vec{f}_l)\}$
8 **end**
9 **for** all means $\vec{f}_j$ **do**
10 $\vec{f}_j = \mu(c_j)$
11 **end**
12 **end**

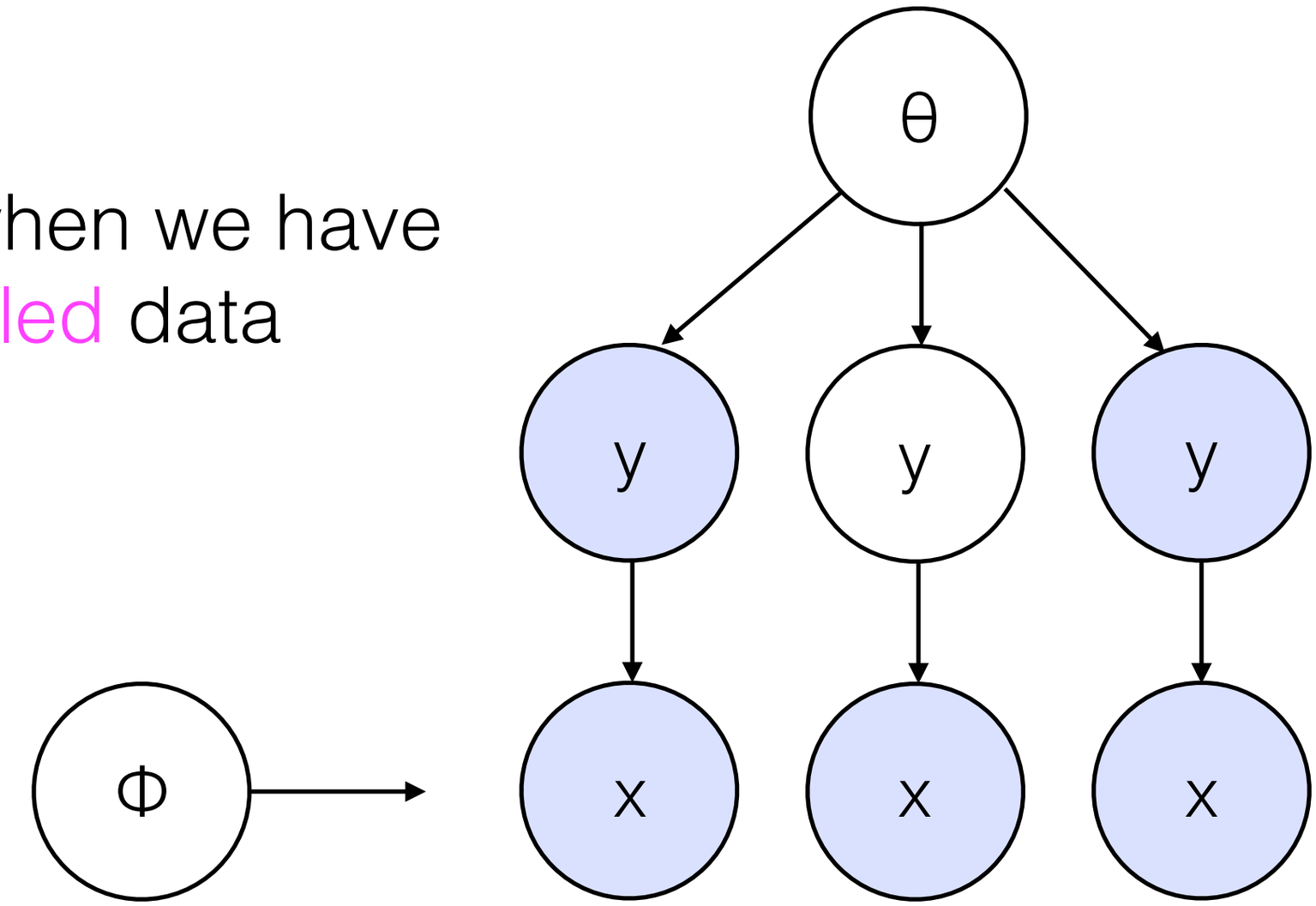
Expectation Maximization

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

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

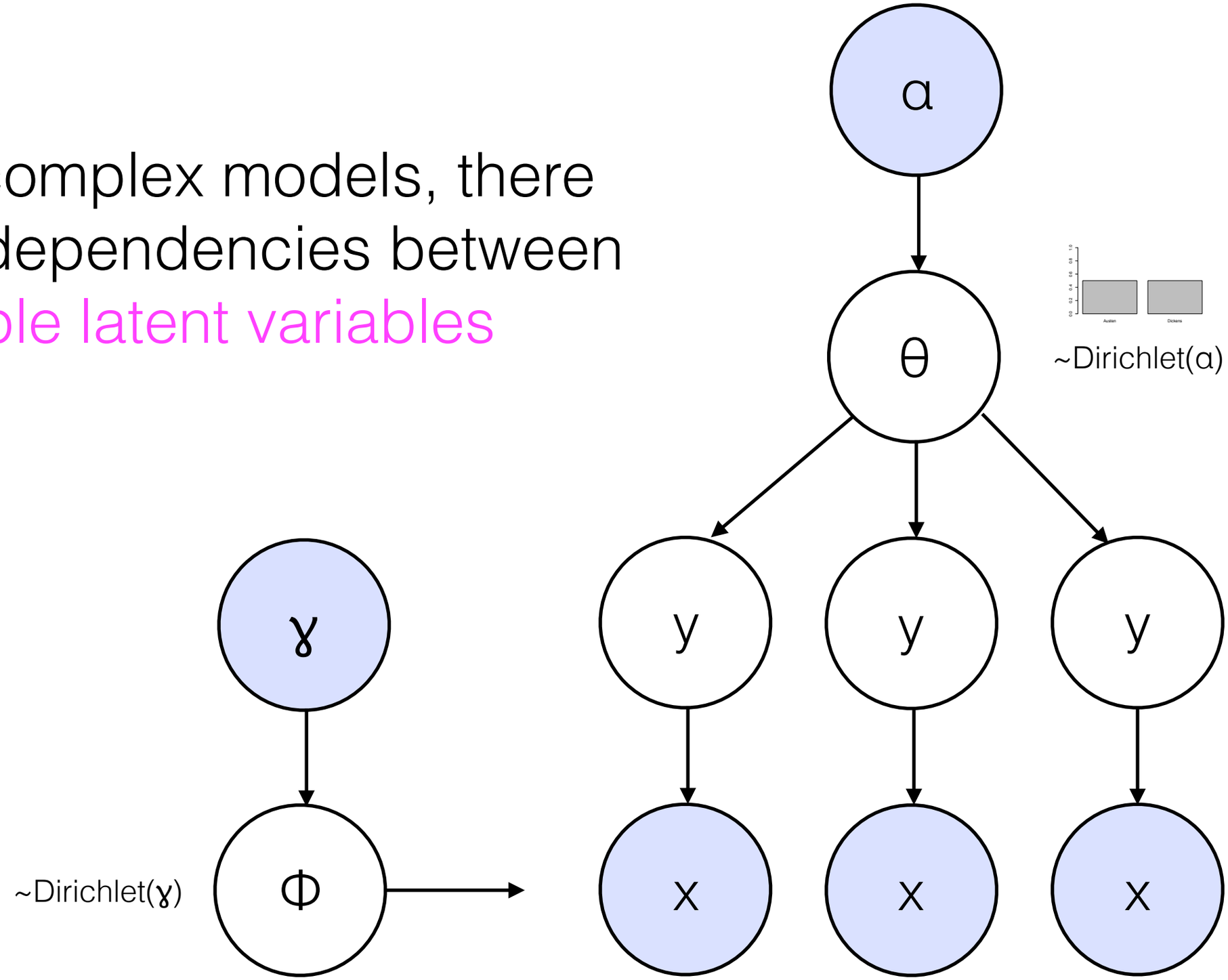
what's this value for an observed label?

2. Use those expected values to maximize parameters

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

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

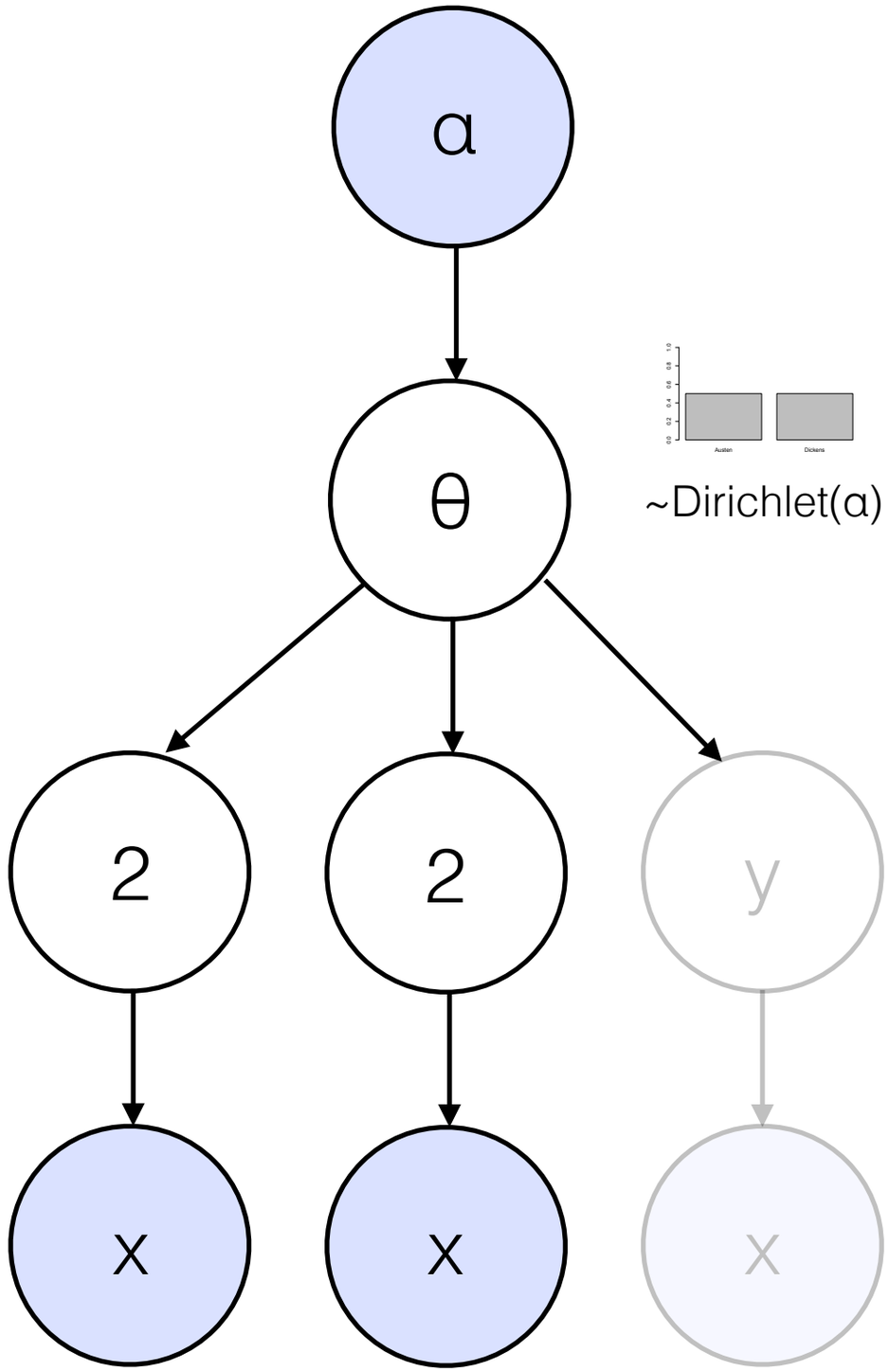
In more complex models, there are often dependencies between multiple latent variables



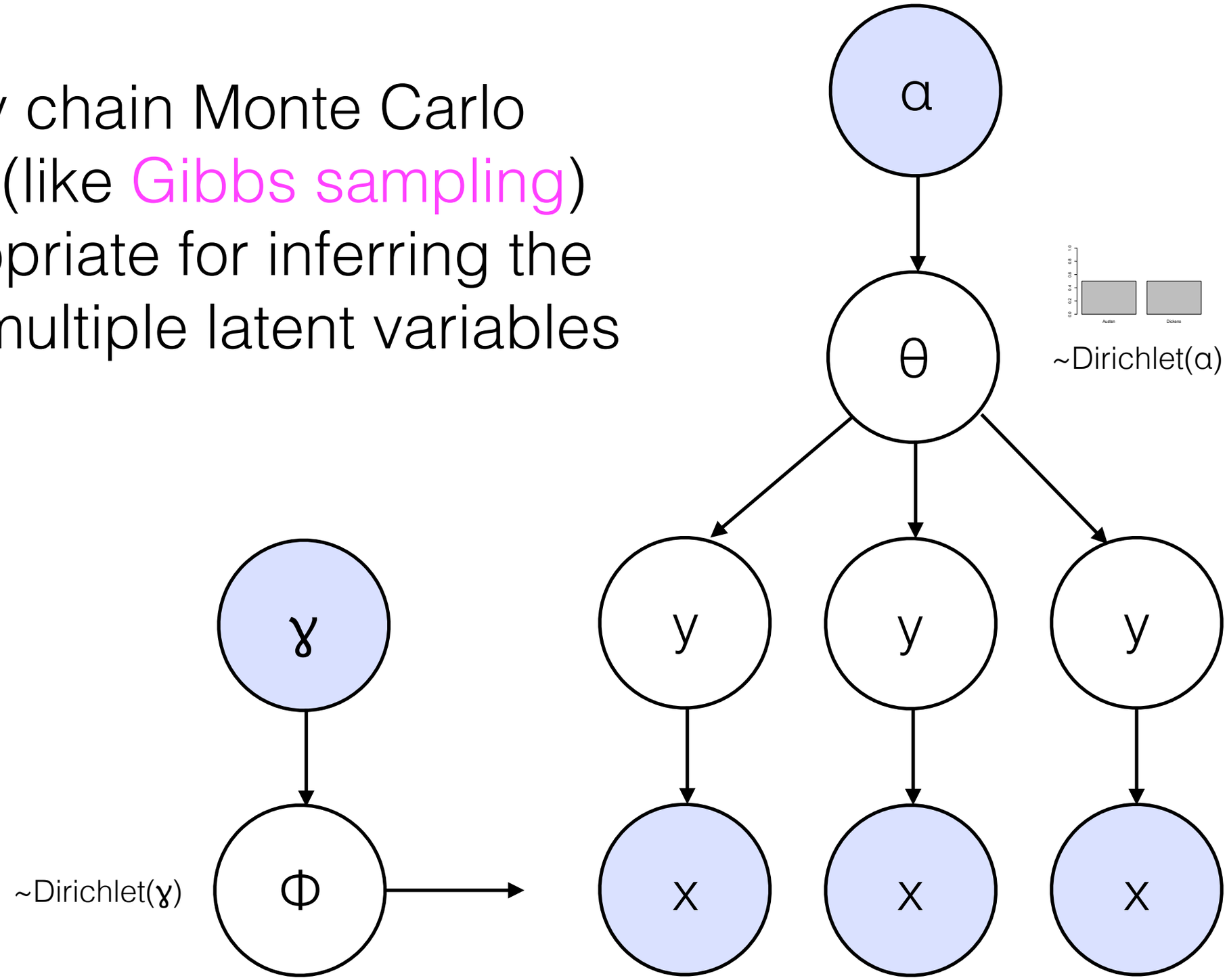
In more complex models, there are often dependencies between multiple latent variables

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

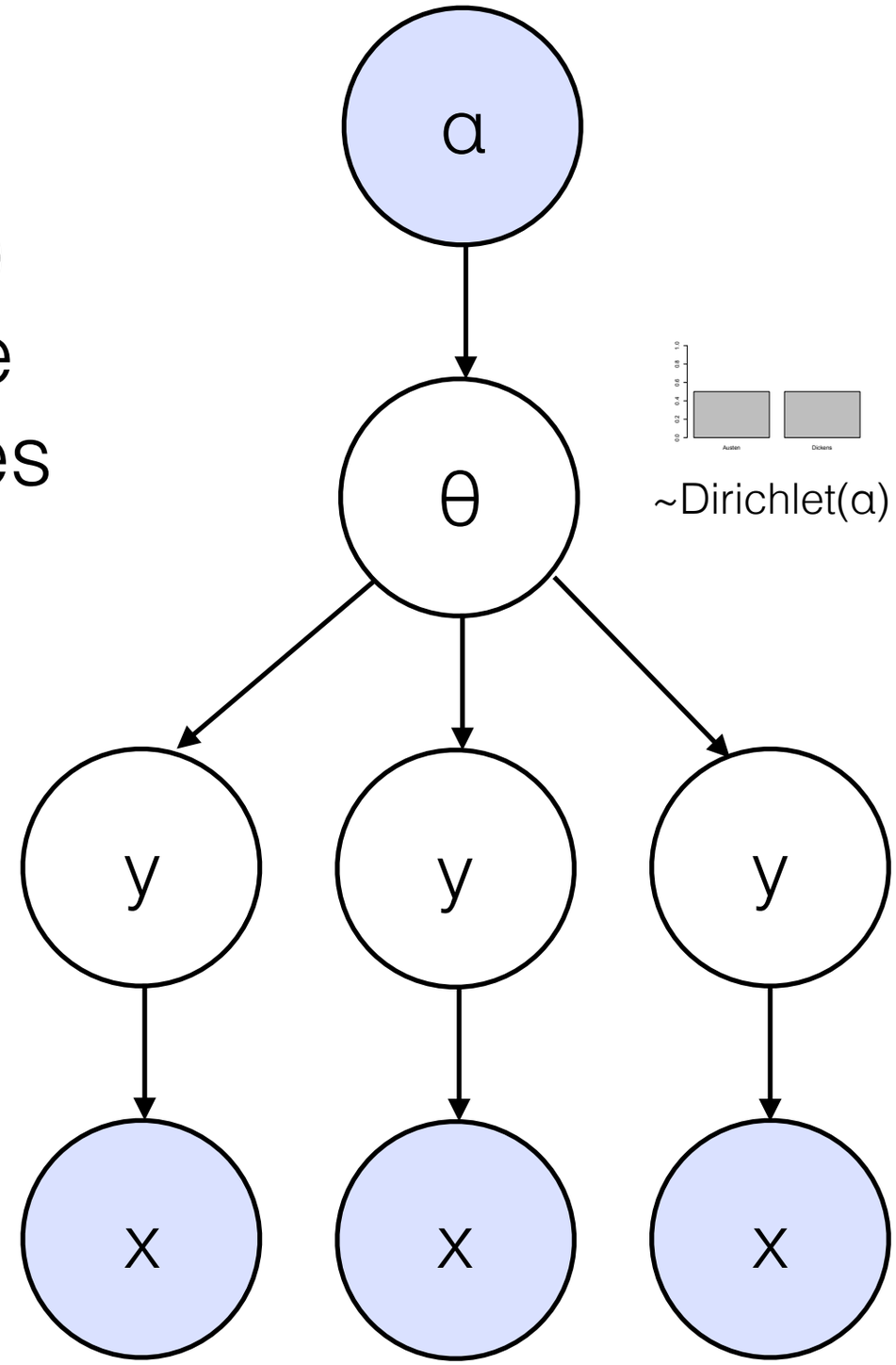


Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

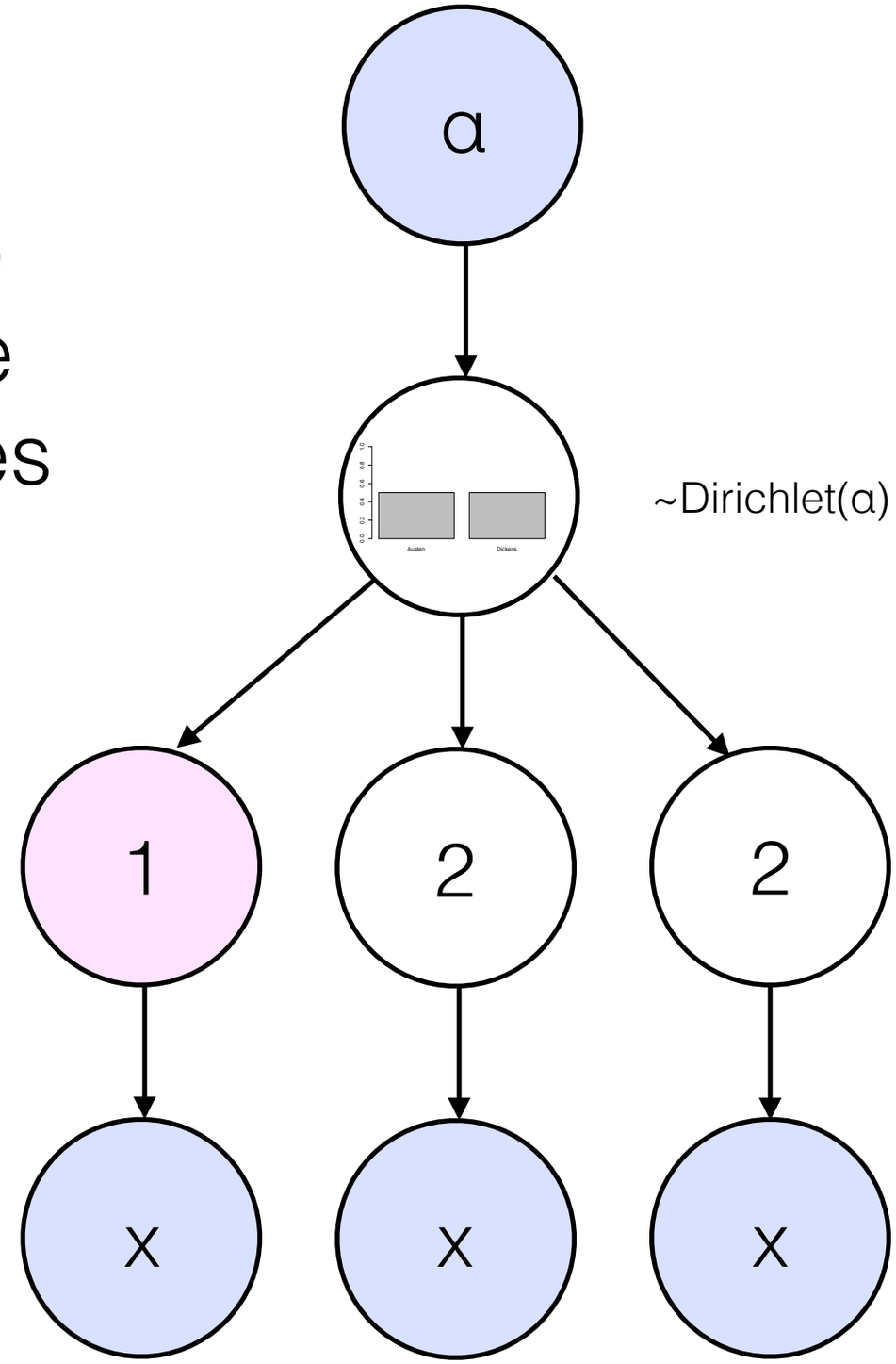


Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

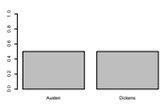
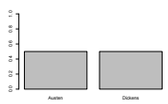
The idea is very simple: start out with random guesses for all variables



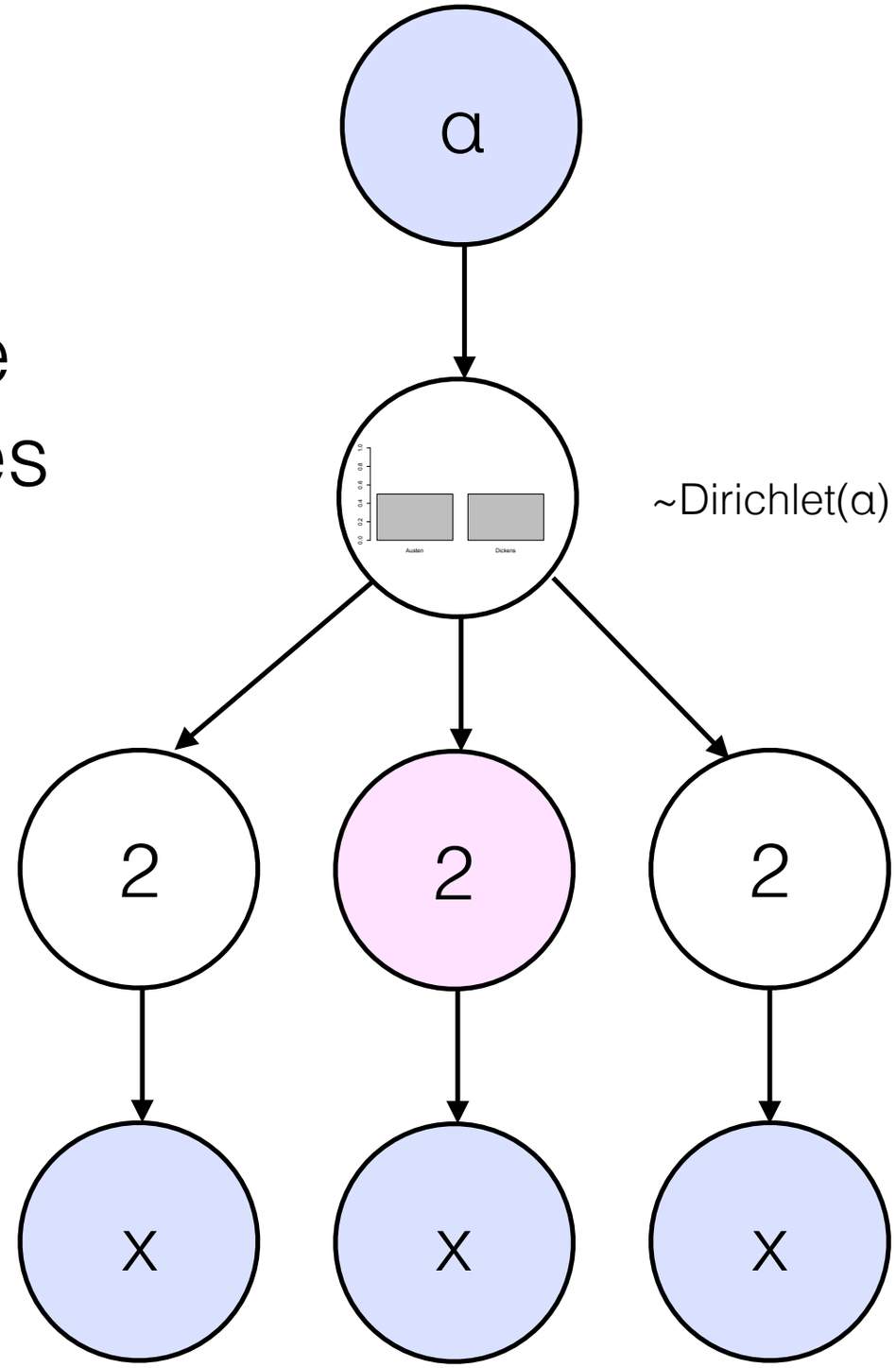
Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables



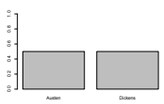
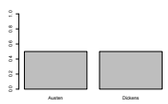
Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else

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

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

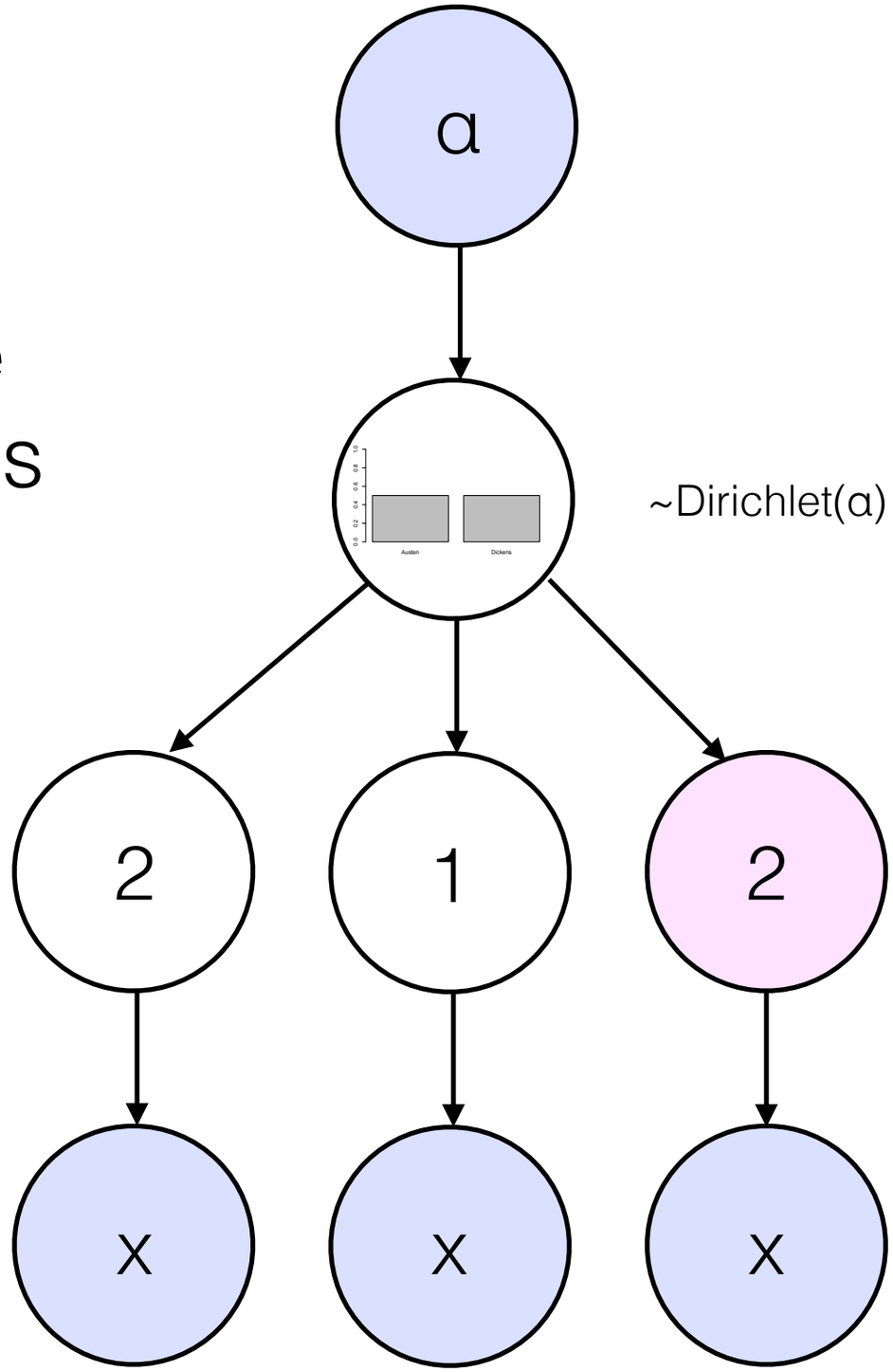


Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else

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

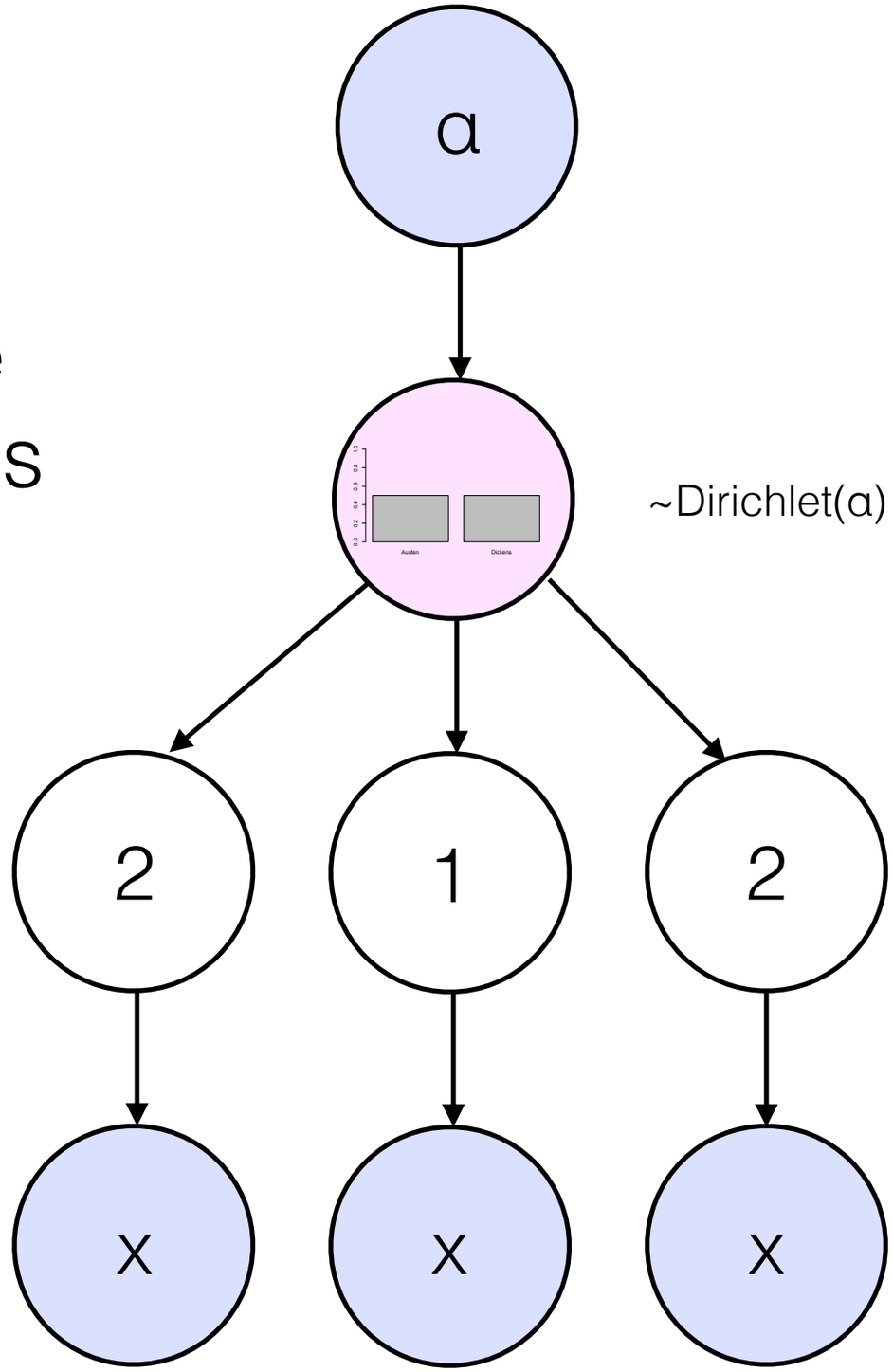
Markov chain Monte Carlo methods (like Gibbs sampling) are appropriate for inferring the values of multiple latent variables

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



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

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

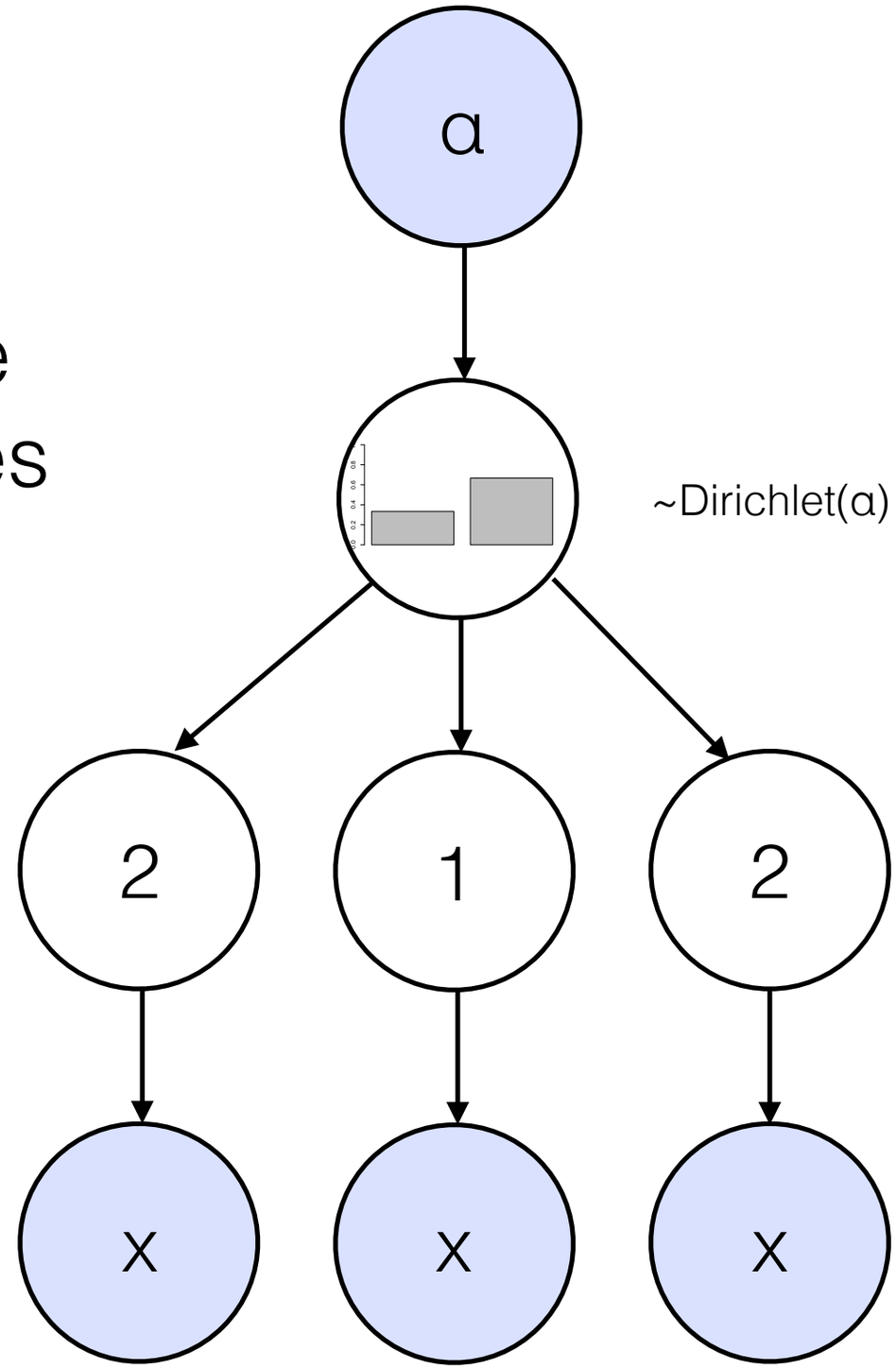


Then, iterate through each variable and sample a new value for it conditioned on the current samples of everything else

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

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

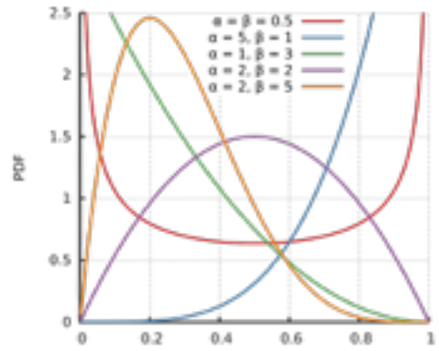
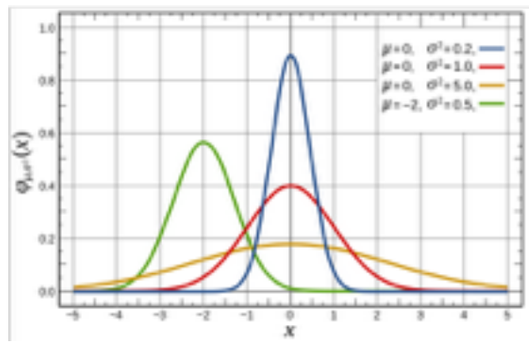
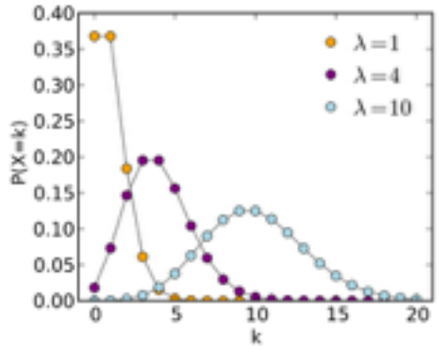


Graphical models

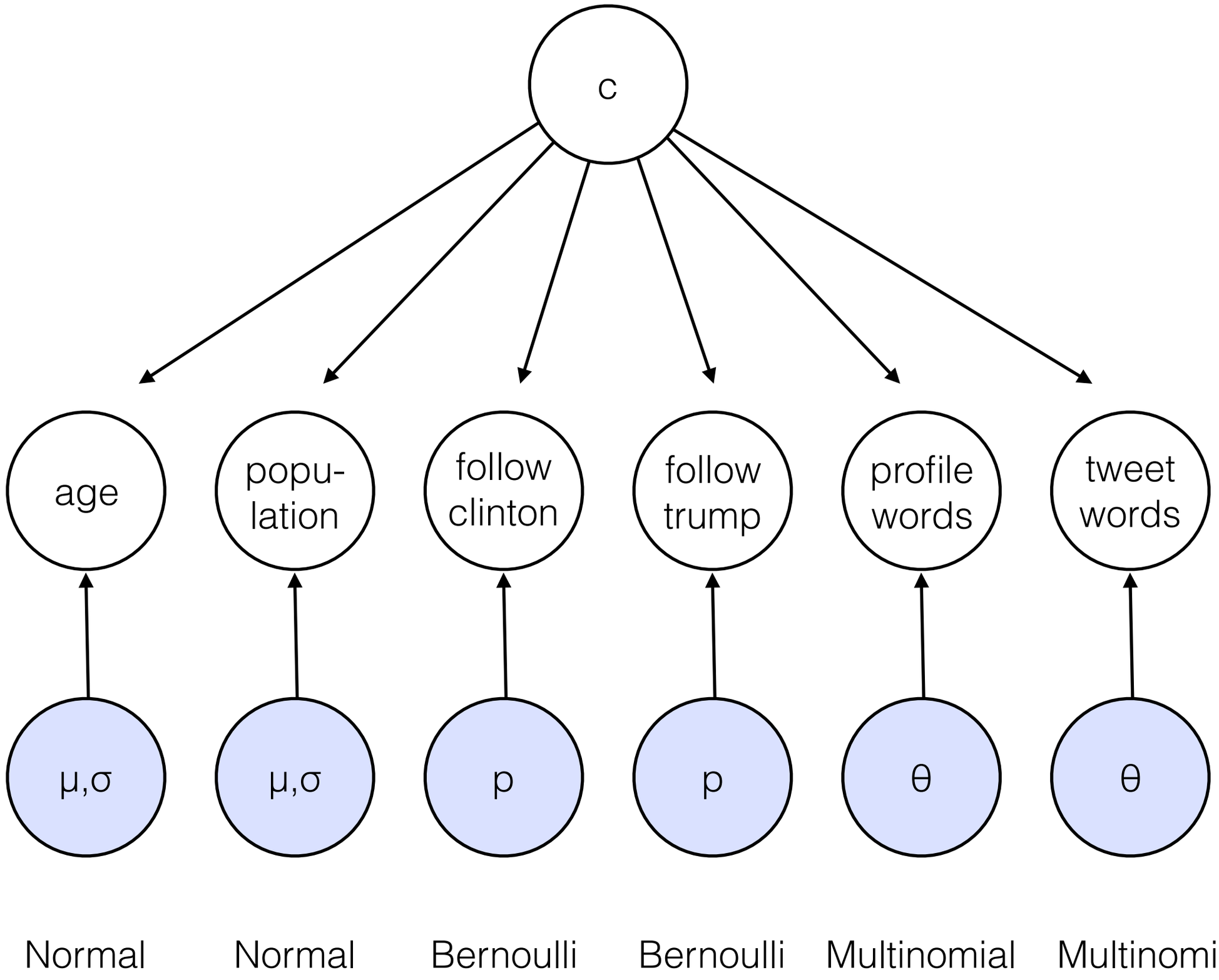
- Graphical models articulate the relationship between variables

- 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	

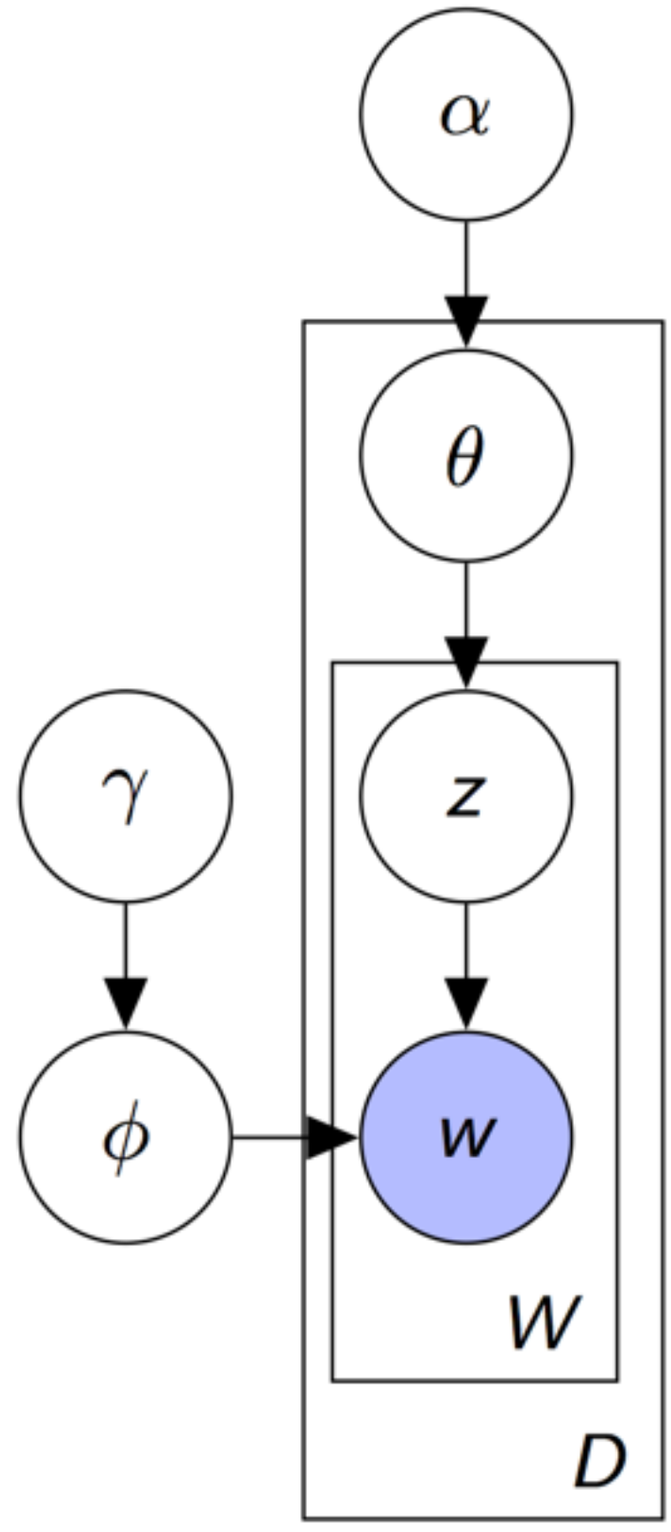
Feature	Value	Distribution?
follow clinton	0	
follow trump	0	
age	24	
word counts in profile	Berkeley, liberal, runner	
word counts in profile	the, election, a, data, movies	
population size of your city	116,000	


c
age
popu-
lation
follow
clinton
follow
trump
profile
words
tweet
words
μ,σ
μ,σ
p
p
θ
θ
Normal
Normal
Bernoulli
Bernoulli
Multinomial
Multinomial

Topic models

David Mimno, “Topic models without the randomness”

NLP seminar
tomorrow (2/25) 4pm, South Hall room 205



Rao et al. (2010)

FEATURE	*Description/Example*
SIMLEYS	A list of emoticons compiled from the Wikipedia.
OMG	Abbreviation for 'Oh My God'
ELLIPSES	'....'
POSSESIVE BIGRAMS	E.g. my_XXX, our_XXX
REPATED ALPHABETS	E.g. niceeeeee, noooo waaaay
SELF	E.g., I_xxx, Im_xxx
LAUGH	E.g. LOL, ROTFL, LMFAO, haha, hehe
SHOUT	Text in ALLCAPS
EXASPERATION	E.g. Ugh, mmmm, hmmm, ahh, grrr
AGREEMENT	E.g. yea, yeah, ohya
HONORIFICS	E.g. dude, man, bro, sir
AFFECTION	E.g. xoxo
EXCITEMENT	A string of exclamation symbols (!!!!!)
SINGLE EXCLAIM	A single exclamation at the end of the tweet
PUZZLED PUNCT	A combination of any number of ? and ! (!?!!??!)

Democrat		Republican	
my_youthful	1	*my_zionist*	1
my_yoga	1	*my_yuengling*	1
my_vegetarianism	1	*my_weapons*	1
my_upscale	1	*my_walmart*	1
my_tofurkey	1	*my_trucker*	1
my_synagogue	1	*my_patroit*	1
my_lakers	0.93	*my_lsu*	1
my_gays	0.8	*my_blackeberry*	1
my_feminist	0.67	*my_redneck*	0.89
my_sushi	0.6	*my_marine*	0.82
my_marathon	-10	my_partner	-0.29
my_trailer	-11	my_atheism	-1
my_liberty	-11.5	my_sushi	-1.5
my_information	-12.5	my_netflix	-2.2
my_teleprompter	-13	my_passport	-2.43
my_warrior	-14	my_manager	-3.67
my_property	-19	my_bicycle	-4
my_lines	-19	my_android	-6
my_guns	-19.67	my_medicare	-14
my_bishop	-33	my_nigga	-17

Above 30		Below 30	
my_zzzzzzz	1	*my_zunehd*	1
my_work	1	*my_yuppie*	1
my_epidural	1	*my_sorors*	0.94
my_daughters	0.98	*my_rents*	0.93
my_grandkids	0.95	*my_classes*	0.90
my_retirement	0.92	*my_xbox*	0.87
my_hubbys	0.91	*my_greek*	0.79
my_workouts	0.9	*my_biceps*	0.75
my_teenage	0.88	*my_homies*	0.70
my_inlaws	0.86	*my_uniform*	0.56
my_bestfriend	-17	my_memoir	-21
my_internship	-18.17	my_daughter	-24.70
my_dorm	-18.75	my_youngest	-24.71
my_cuzzo	-19	my_tribe	-29
my_bby	-26	my_nelson	-36
my_boi	-30	my_oldest	-39
my_dudes	-34	my_2yo	-39
my_roomate	-37	my_kiddos	-45
my_formspring	-42	my_daughters	-56
my_hw	-51	my_prayer	-62

Disfluency/Agreement	*#female/#male*
oh	2.3
ah	2.1
hmm	1.6
ugh	1.6
grrr	1.3
yeah, yea, ...	0.8

Feature	*#female/#male*
Emoticons	3.5
Elipses	1.5
Character repetition	1.4
Repeated exclamation	2.0
Puzzled punctuation	1.8
OMG	4.0