Latent Dirichlet Allocation (LDA)

Also Known As

Topic Modeling

The Domain: Natural Language Text

Collection of documents

Each document consists of a set of word *tokens* drawn (with replacement) from a set of word *types*

e.g., "The big dog ate the small dog."

Goal

construct models of domain via unsupervised learning

i.e., learning structure of domain

What Does It Mean To Understand The Structure Of A Domain?

- Obtain a compact representation of each document
- Obtain a generative model that produces observed documents with high probability (and others with lower probability)

Two Contrasting Approaches To Modeling Environments Of Words And Text

Latent Semantic Analysis (LSA)

- mathematical model
- a bit hacky

Topic Model (LDA)

- probabilistic model
- principled -> has produced many extensions and embellishments

LSA

The set up

D documents
W distinct words
F = WxD coocurrence matrix
f_{wd} = frequency of word w in document d

LSA: Transforming The Co-occurence Matrix

Relative entropy of a word across documents

$$H_w = -\frac{\sum_{d=1}^{D} \frac{f_{wd}}{f_{w\cdot}} \log\{\frac{f_{wd}}{f_{w\cdot}}\}}{\log D}$$

f_{wd}/f_w: P(d|w)

H_w = value in [0, 1]
0=word appears in only 1 doc
1=word spread across all documents

Specificity: (1-H_w)

0 = word tells you nothing about the document;
1= word tells you a lot about the document

LSA: Transforming The Co-occurence Matrix

G = WxD normalized coocurrence matrix

$$g_{wd} = \log\{f_{wd} + 1\}(1 - H_w)$$

log transform common for word freq analysis

+1 ensures no log(0)

weighted by specificity

Representation of word i: row i of G

problem: high dimensional representation

problem: doesn’t capture similarity structure of documents

LSA: Representing A Word

Dimensionality reduction via SVD

$G = \quad M_1 \quad M_2 \quad M_3$

[WxD] = [WxR] [RxR] [RxD]

if R = min(W,D) reconstruction is perfect

if R < min(W,D) least squares reconstruction, i.e., capture whatever structure there is in matrix with a reduced number of parameters

Reduced representation of word i: row i of (M_1M_2)

Reduced representation of document j: column j of (M_2M_3)

Can used reduced representation to determine semantic relationships

What’s the advantage of a reduced representation?

LSA Versus Topic Model

The reduced representations in LSA are vectors whose elements (features)

- can be negative
- are completely unconstrained

If we wish to operate in a currency of probability, tthen the elements

- must be nonnegative
- must sum to 1

Terminology

- LSA = LSI = latent semantic indexing
- pLSI = probabilistic latent semantic indexing
- LDA

topic model

pLSI (Hoffman, 1999)

Probabilistic model of language production

Generative model

Select a document with probability P(D)

Select a (latent) topic with probability P(Z|D)

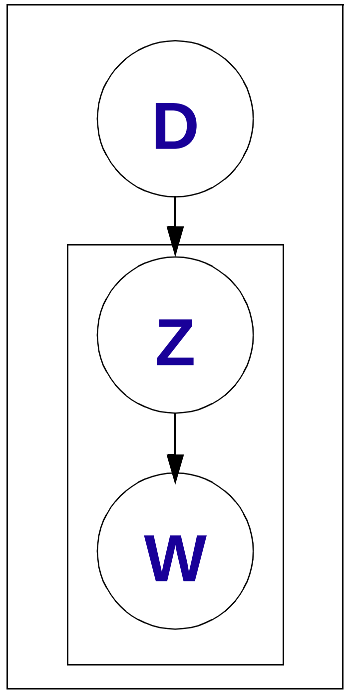
Generate a word with probability P(W|Z)

Produce pair <d_i, w_i> on draw i

P(D, W, Z) = P(D) P(Z|D) P(W|Z)

P(D, W) = Σ_z P(D) P(z|D) P(W|z)

P(W | D) = Σ_z P(z|D) P(W|z)



Inferring Latent Variable

P(Z|D,W)

$P(D, W, Z) = P(D)\, P(Z|D)\, P(W|Z)$

$P(D, W) = \Sigma_z\, P(D)\, P(z|D)\, P(W|z)$

$P(Z|D,W) = P(D, W, Z) \,/\, P(D, W)$

$= P(Z|D)\, P(W|Z) \,/\, [\Sigma_z\, P(z|D)\, P(W|z)]$

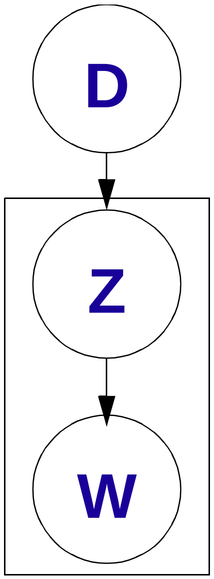


Plate Notation

Way of representing

- **multiple documents**
- **multiple words per document**

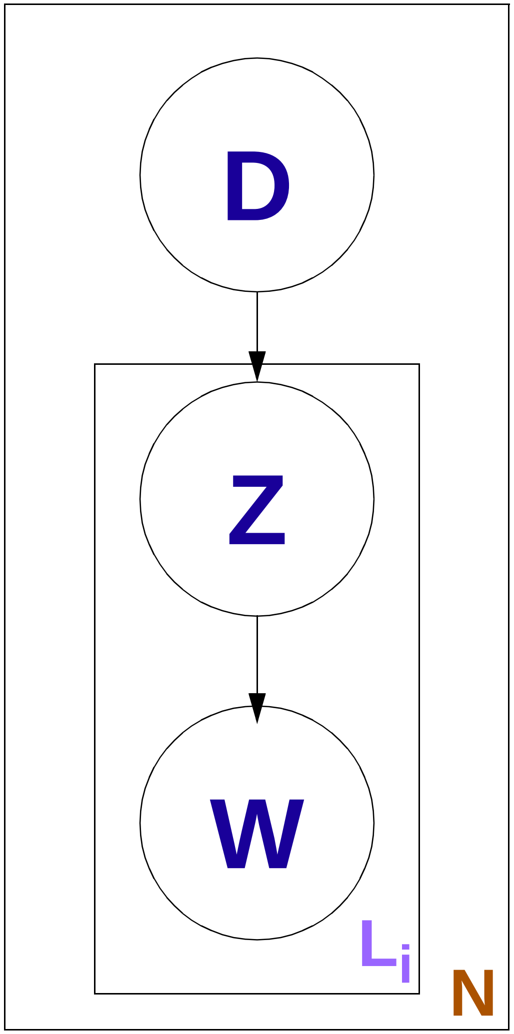
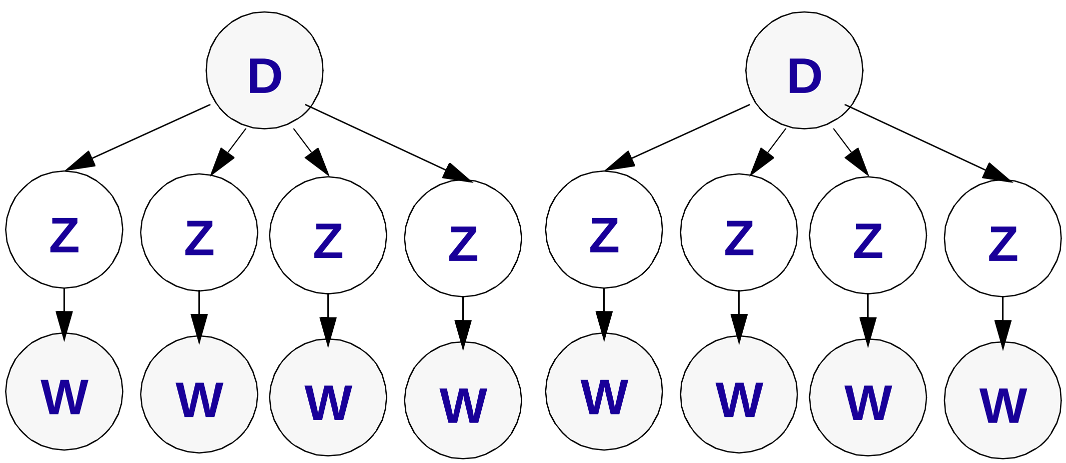
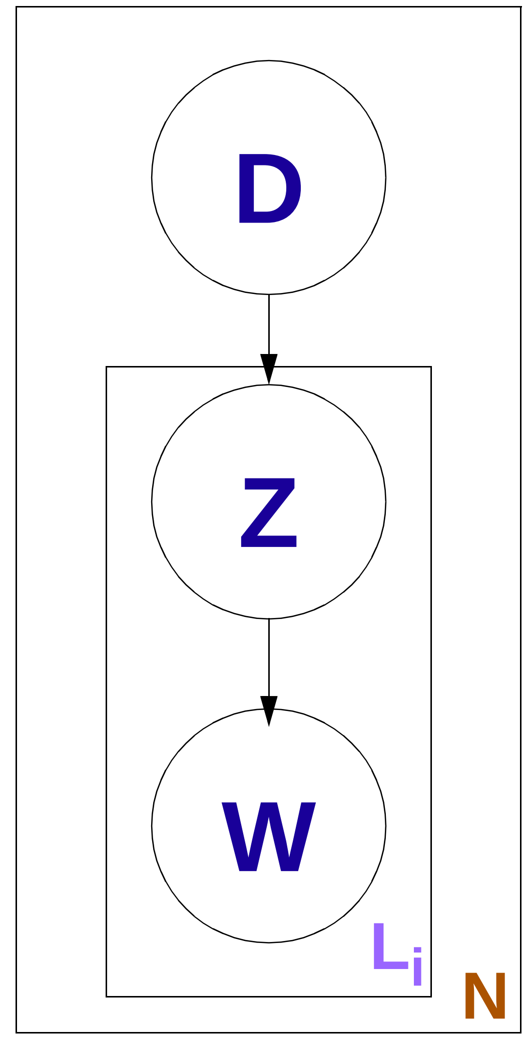


Plate Notation

Way of representing

- **multiple documents**
- **multiple words per document**





Translating Notation

	Barber	Typical Topic Modeling Notation
total # documents	N	N
total # topics	K	T
total # word types	D (dictionary)	W
index over documents	n	i: index over document-word pairs $\{w_i, d_i\}$
index over words in document	w	
index over words in dictionary	i	
topic assignment	z^n_w	z_i: topic of word-document pair i
distribution over topics	$\{\pi^n_k\}$	$\{\theta^{d_i}_j\}$
distribution over words	$\{\theta^k_i\}$	$\{\phi^j_{w_i}\}$
index over topics	k	j

Two Approaches To Learning Conditional Probabilities

$P(Z=j \mid D=d_i)$ or $\theta_j^{d_i}$

$P(W=w_i \mid Z=j)$ or $\phi_{w_i}^{j}$

Hoffmann (1999)

Search for the single best θ and ϕ via gradient descent in cross entropy (difference between distribution) of data and model

$-\sum_{w,d} n(d,w) \log P(d,w)$

Griffiths & Steyvers (2002, 2005); Blei, Ng, & Jordan (2003)

Treat θ and ϕ as random variables.

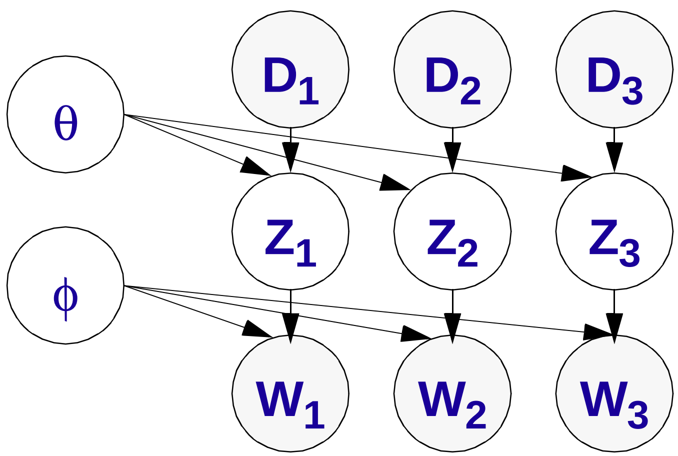
Treating θ And ϕ As Random Variables

Can marginalize over uncertainty, i.e.,

$$P(Z|D) = \int_{\theta} P(Z|D,\theta)P(\theta)$$

$$P(W|Z) = \int_{\phi} P(W|Z,\phi)P(\phi)$$

Model



Treating θ And ϕ As Random Variables

The two conditional distributions are defined over *discrete alternatives.*

$P(Z=j \mid D=d_i)$ or $\theta_j^{d_i}$

$P(W=w_i \mid Z=j)$ or $\phi_{w_i}^{j}$

If *n* alternatives, distribution can be represented by multinomial with *n–1* degrees of freedom.

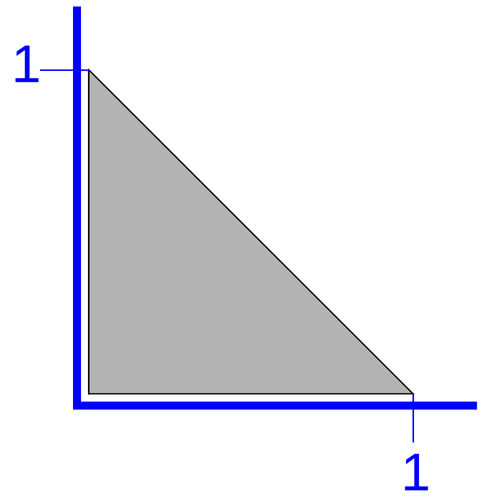
To represent θ and ϕ as random variables, need to encode a distribution over distributions...

Dirichlet Distribution

- **represents probability distribution over multinomial distributions**

  You can think of the uncertainty space over *n* probabilities constrained such that P(x) = 0 if $(\Sigma_i x_i) \neq 1$ or if $x_i < 0$...

  ...or the representational space over *n–1* probabilities constrained such that P(x)=0 if $(\Sigma_i x_i) > 1$ or if $x_i < 0$.



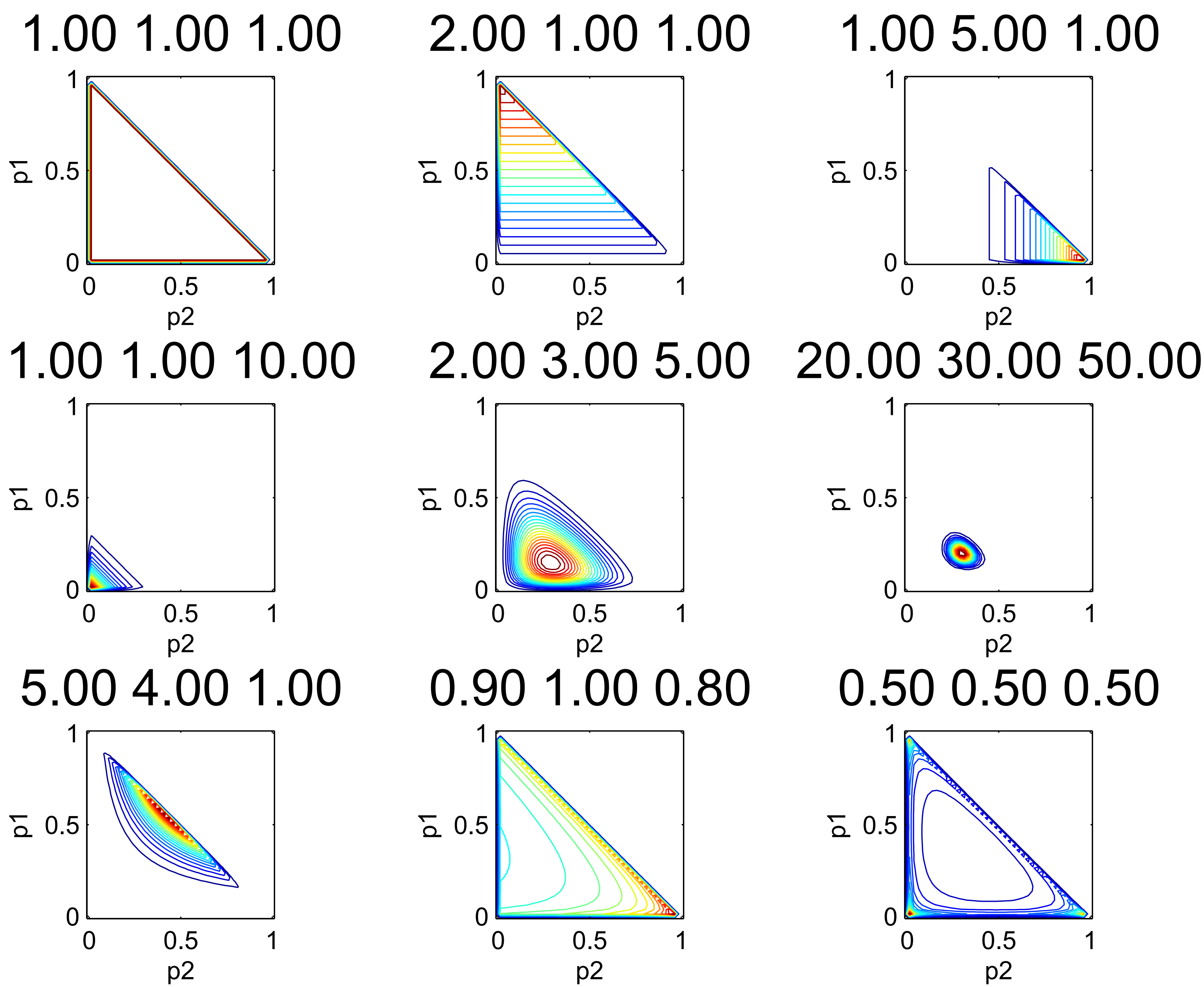
- **generalization of beta distribution**

- **for multinomial RV with *n* alternatives, Dirichlet has *n* parameters.**

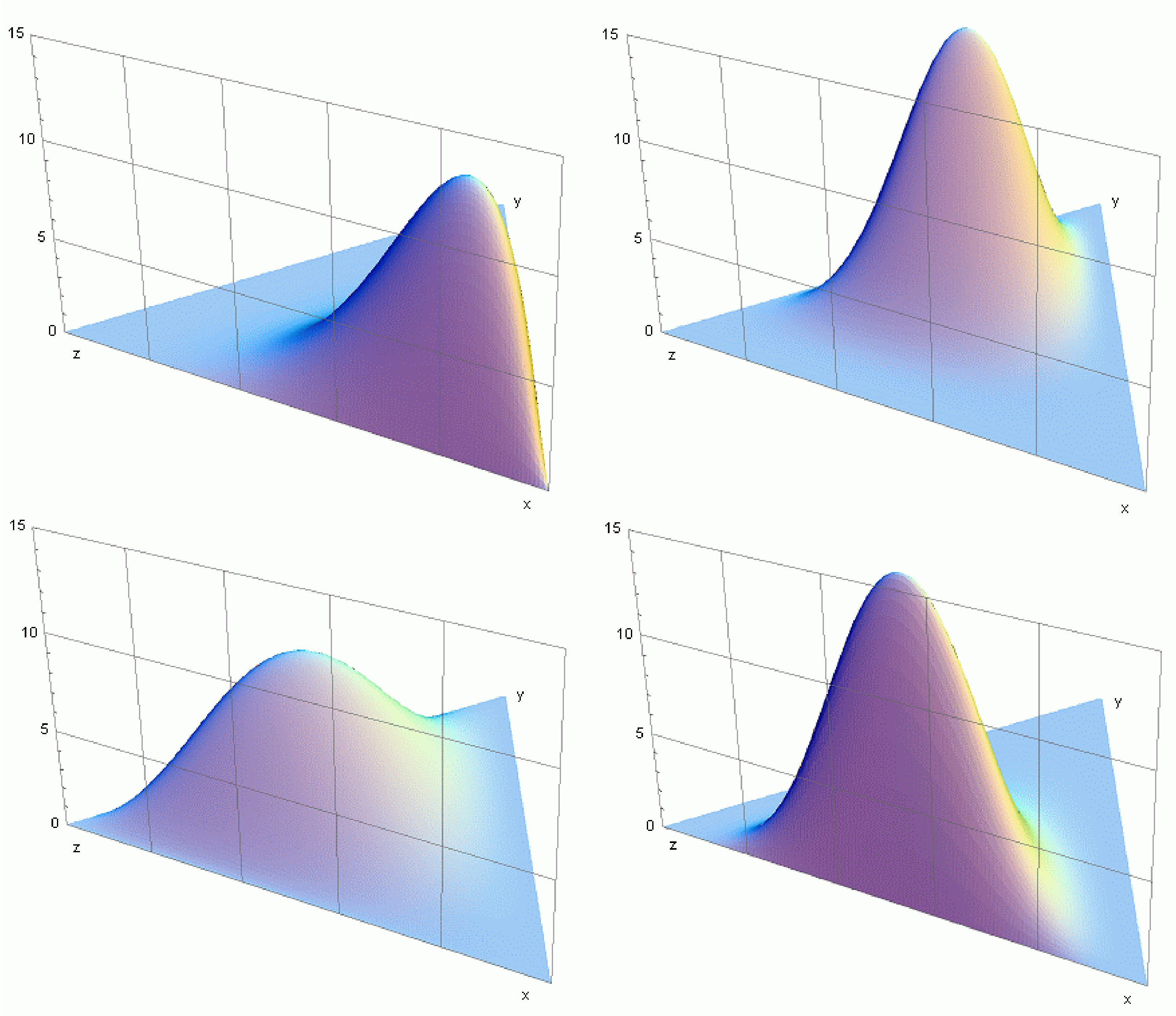
  Each parameter is a count of the number of occurrences.

  Why *n* and not *n–1* since there are *n–1* degrees of freedom?

Dirichlet Distribution (n=3)



Dirichlet Distribution (n=3)



Dirichlet Is Conjugate Prior Of Multinomial

Simple example

$\phi \sim \text{Dirichlet}(1, 3, 4)$

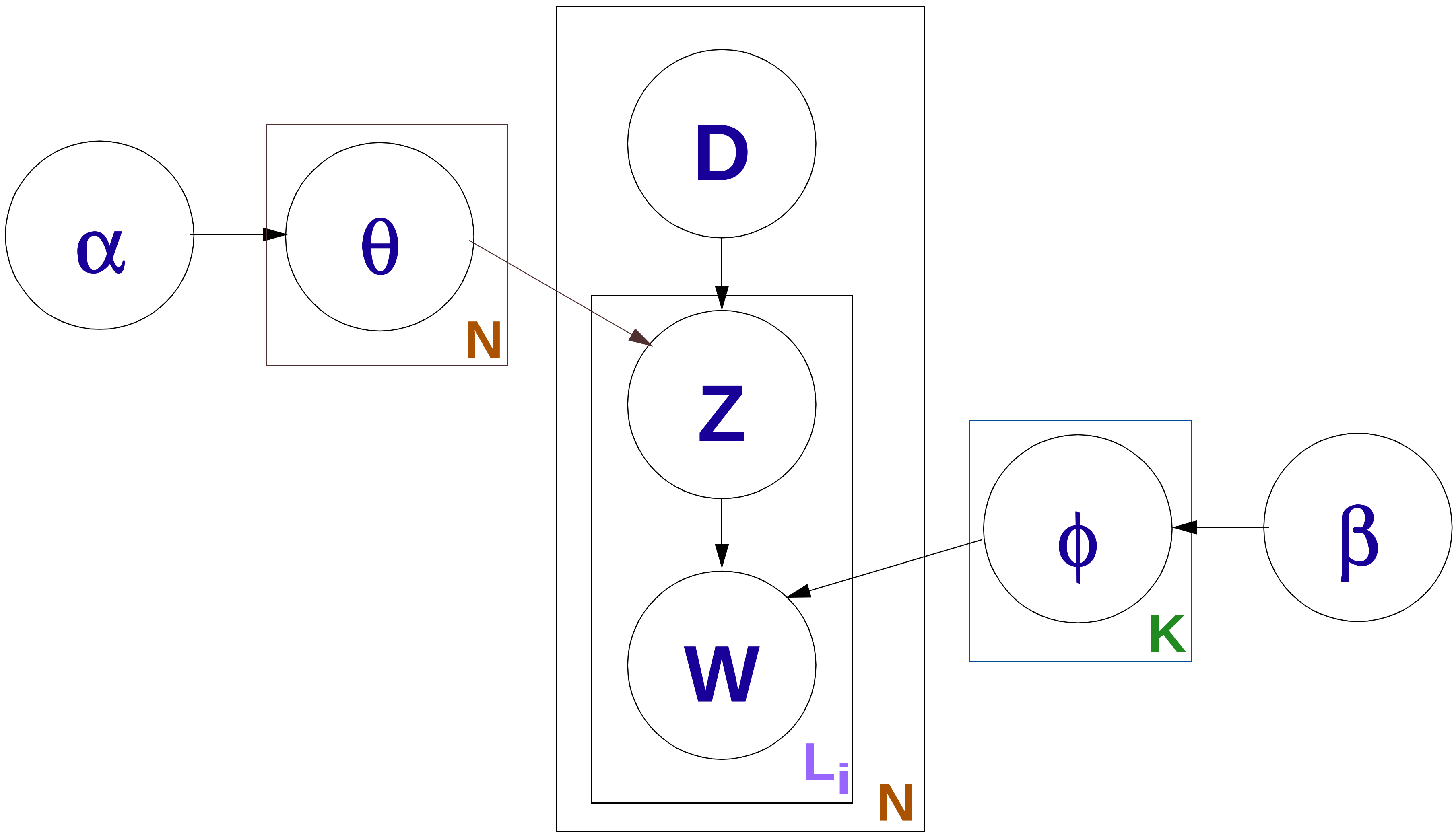
$O = \{w1, w1, w2, w3, w2, w1\}$

$\phi \mid O \sim \text{Dirichlet}(4, 5, 5)$

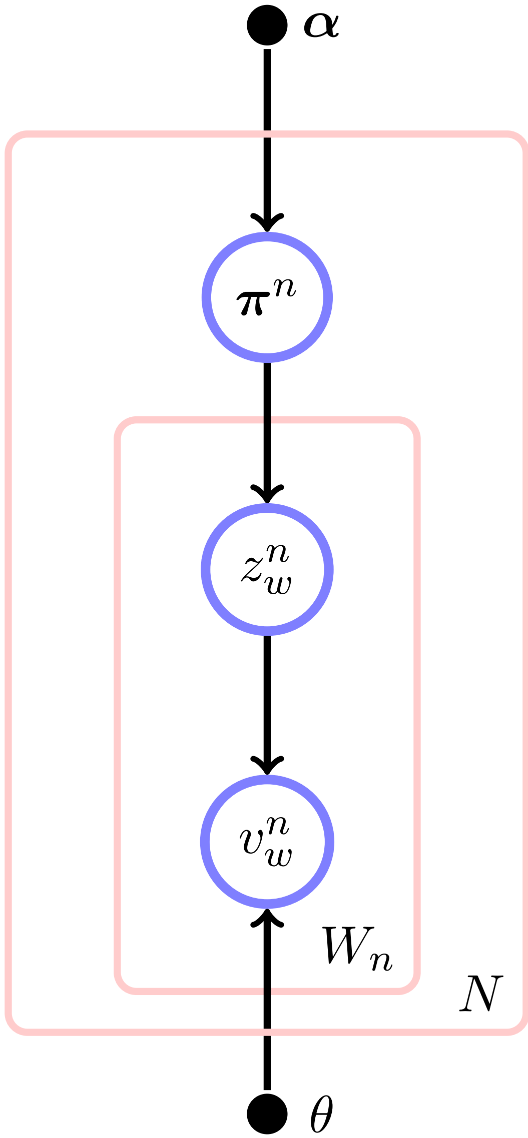
Weak assumption about prior

$\phi \sim \text{Dirichlet}(\beta, \beta, \beta)$

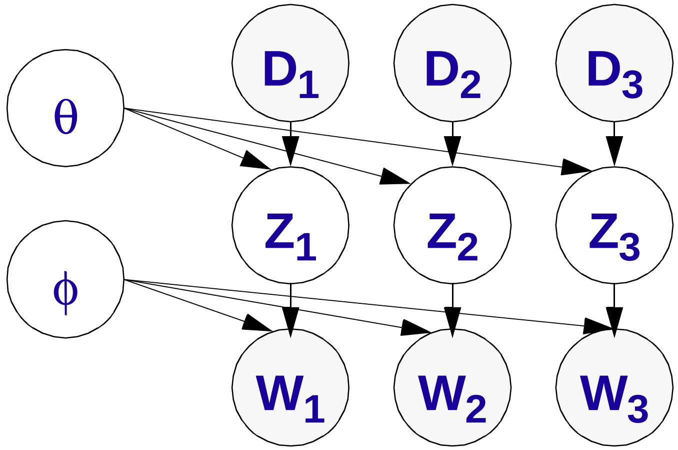
Full Model



Barber Figure



Collapsed Gibbs Sampling Approach



1. Define Dirichlet priors on θ^{d_i} and ϕ^j

2. Perform sampling over latent variables Z, integrating out or collapsing over θ and ϕ

$$P(Z_i|Z_{-i}, D, W) \sim \int_{\theta,\phi} P(W|Z,\phi)P(Z|D,\theta)P(\phi)P(\theta)d\phi d\theta$$

This can be done analytically due to Dirichlet-Multinomial relationship

Note: no explicit representation of posterior $P(\theta, \phi| Z, D, W)$

Collapsed Gibbs Sampling

$$P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(\cdot)} + W\beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + T\alpha}$$

i: index over
word-doc pairs

Ignore α and β for the moment

First term: proportion of topic j draws in which w_i picked

Second term: proportion of words in document d_i assigned to topic j

This formula integrates out the Dirichlet uncertainty over the multinomial probabilities!

What are α and β?

Effectively, they function as smoothing parameters

Large values -> more smoothing

Detailed Procedure For Sampling From P(Z|D,W)

1. **Randomly assign each <d_i, w_i> pair a z_i value.**

2. **For each i, resample according to equation on previous slide (one *iteration*)**

3. **Repeat for a burn in of, say, 1000 iterations**

4. **Use current assignment as a sample and estimate**

   P(Z|D)

   P(W|Z)

Typically with Gibbs sampling, the results of multiple chains (restarts) are used. Why wouldn't that work here?

Results

Arts	Budgets	Children	Education
new	million	children	school
film	tax	women	students
show	program	people	schools
music	budget	child	education
movie	billion	years	teachers
play	federal	families	high
musical	year	work	public
best	spending	parents	teacher
actor	new	says	bennett
first	state	family	manigat
york	plan	welfare	namphy
opera	money	men	state
theater	programs	percent	president
actress	government	care	elementary
love	congress	life	haiti

(a)

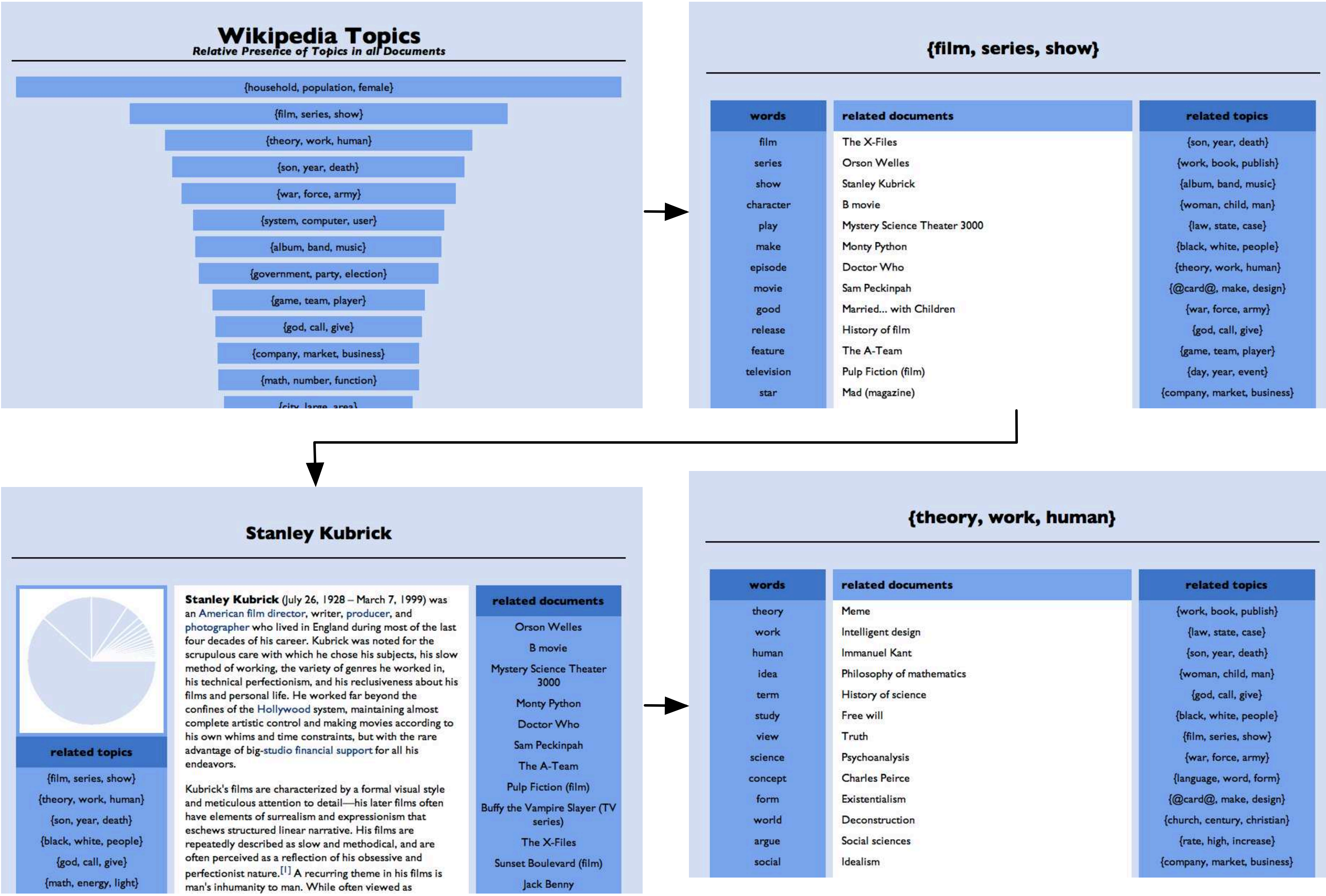
The William Randolph Hearst Foundation will give $ 1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services, Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Centers share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

(b)

Results

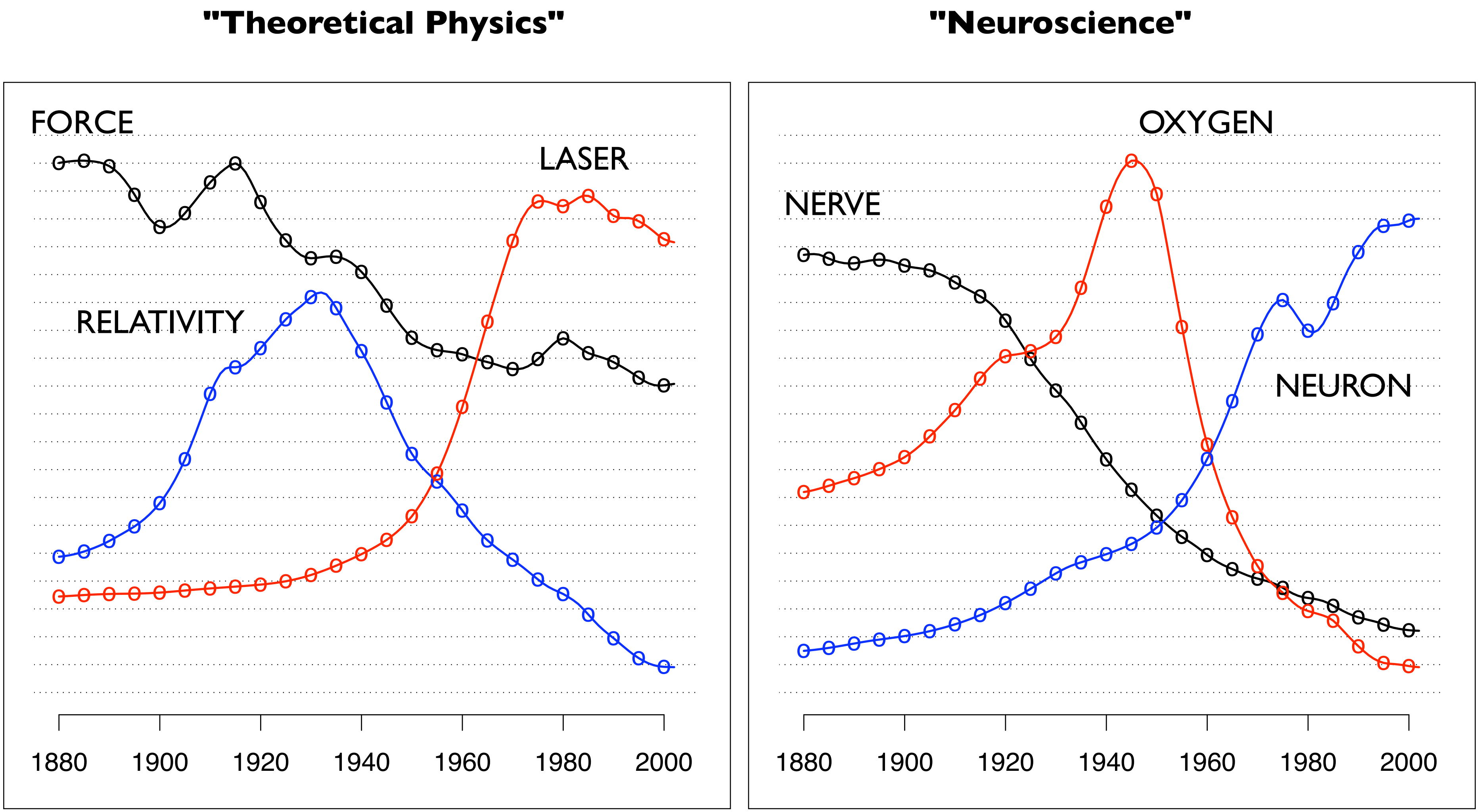
FEEL	MUSIC	BALL	SCIENCE	WORKERS	**FORCE**
FEELINGS	**PLAY**	GAME	STUDY	**WORK**	FORCES
FEELING	DANCE	TEAM	SCIENTISTS	LABOR	MOTION
ANGRY	PLAYS	**PLAY**	SCIENTIFIC	JOBS	BODY
WAY	STAGE	BASEBALL	KNOWLEDGE	WORKING	GRAVITY
THINK	PLAYED	FOOTBALL	**WORK**	WORKER	MASS
SHOW	BAND	PLAYERS	CHEMISTRY	WAGES	PULL
FEELS	AUDIENCE	GAMES	RESEARCH	FACTORY	NEWTON
PEOPLE	MUSICAL	PLAYING	BIOLOGY	JOB	OBJECT
FRIENDS	DANCING	**FIELD**	MATHEMATICS	WAGE	LAW
THINGS	RHYTHM	PLAYED	LABORATORY	SKILLED	DIRECTION
MIGHT	PLAYING	PLAYER	STUDYING	PAID	MOVING
HELP	THEATER	COACH	SCIENTIST	CONDITIONS	REST
HAPPY	DRUM	BASKETBALL	PHYSICS	PAY	FALL
FELT	ACTORS	SPORTS	**FIELD**	**FORCE**	ACTING
LOVE	**SHOW**	HIT	STUDIES	MANY	MOMENTUM
ANGER	BALLET	BAT	UNDERSTAND	HOURS	DISTANCE
BEING	ACTOR	TENNIS	STUDIED	EMPLOYMENT	GRAVITATIONAL
WAYS	DRAMA	TEAMS	SCIENCES	EMPLOYED	PUSH
FEAR	SONG	SOCCER	MANY	EMPLOYERS	VELOCITY

Results



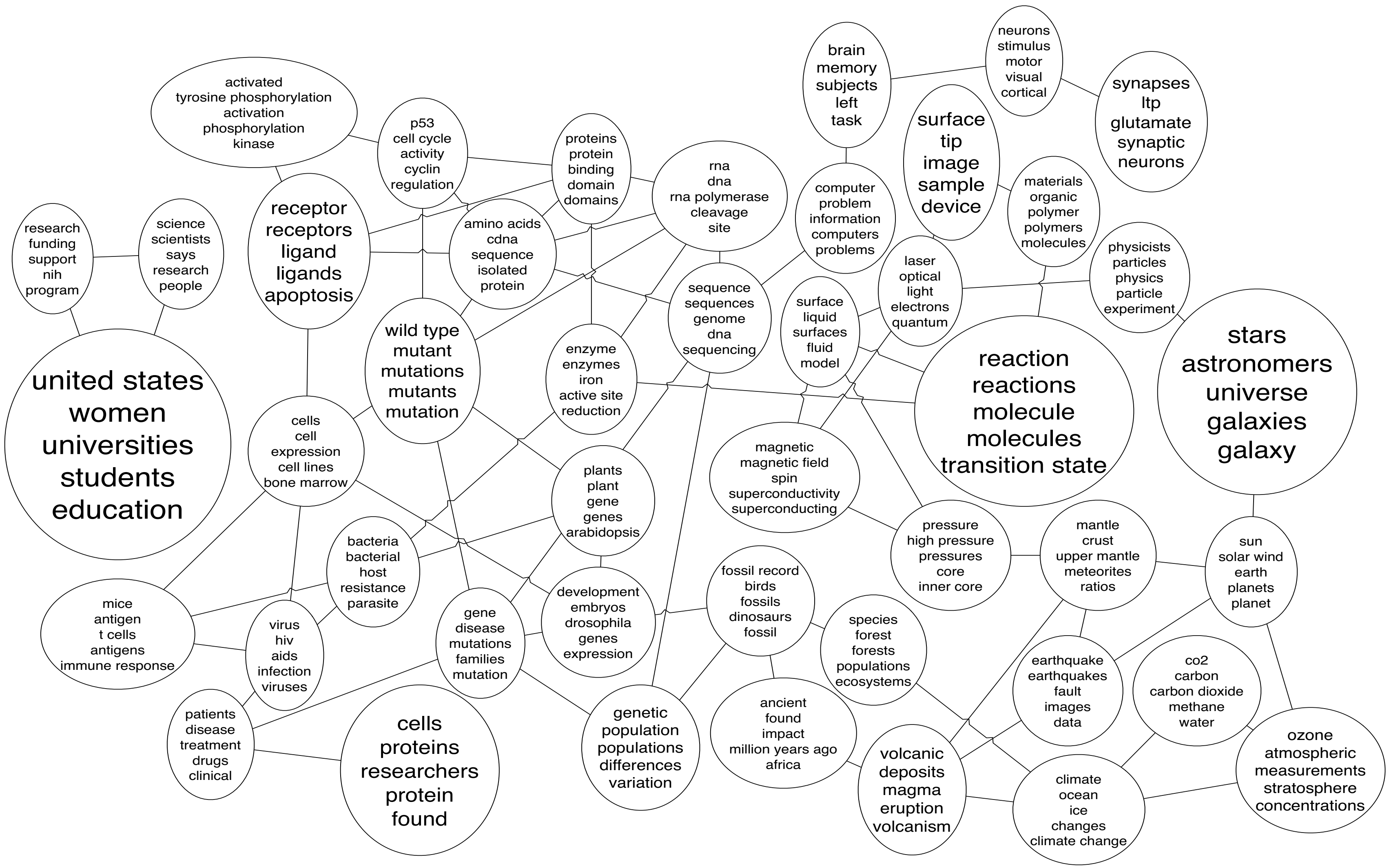
(This and following slides from David Blei tutorial)

Results



How might these graphs have been obtained?

Results



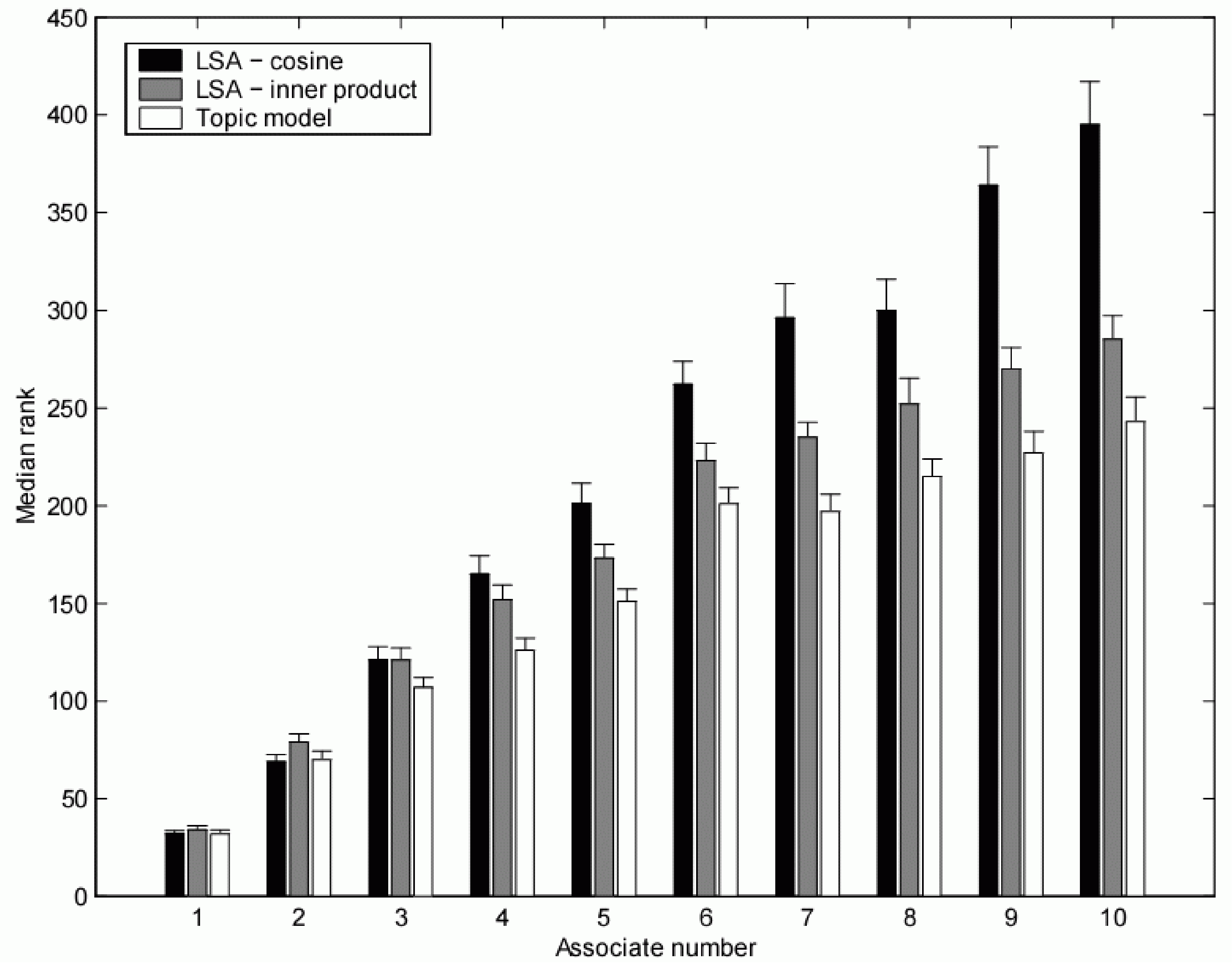
How might this graph have been obtained?

Predicting word association norms

“the” -> ?

“dog” -> ?

Median Rank of k'th Associate



Note: multiple resamples can be used here

Combining Syntax and Semantics

LSA and Topic Model are “bag o’ words” models

Model sequential structure with 3d order HMM

- hidden state is category of word; 50 states
- 1 state for start or end of a sentence
- 48 states for document-independent words (syntax)
- 1 state for document-dependent words (semantics)
- Semantics generated by topic model

Multinomial distributions (most probable words in state)

"syntax"				"semantics"	
HE	ON	BE	SAID	MAP	DOCTOR
YOU	AT	MAKE	ASKED	NORTH	PATIENT
THEY	INTO	GET	THOUGHT	EARTH	HEALTH
I	FROM	HAVE	TOLD	SOUTH	HOSPITAL
SHE	WITH	GO	SAYS	POLE	MEDICAL
WE	THROUGH	TAKE	MEANS	MAPS	CARE
IT	OVER	DO	CALLED	EQUATOR	PATIENTS
PEOPLE	AROUND	FIND	CRIED	WEST	NURSE
EVERYONE	AGAINST	USE	SHOWS	LINES	DOCTORS
OTHERS	ACROSS	SEE	ANSWERED	EAST	MEDICINE
SCIENTISTS	UPON	HELP	TELLS	AUSTRALIA	NURSING
SOMEONE	TOWARD	KEEP	REPLIED	GLOBE	TREATMENT
WHO	UNDER	GIVE	SHOUTED	POLES	NURSES
NOBODY	ALONG	LOOK	EXPLAINED	HEMISPHERE	PHYSICIAN
ONE	NEAR	COME	LAUGHED	LATITUDE	HOSPITALS
SOMETHING	BEHIND	WORK	MEANT	PLACES	DR
ANYONE	OFF	MOVE	WROTE	LAND	SICK
EVERYBODY	ABOVE	LIVE	SHOWED	WORLD	ASSISTANT
SOME	DOWN	EAT	BELIEVED	COMPASS	EMERGENCY
THEN	BEFORE	BECOME	WHISPERED	CONTINENTS	PRACTICE