



EXPLORATORY DATA ANALYSIS: TOPIC MODELING

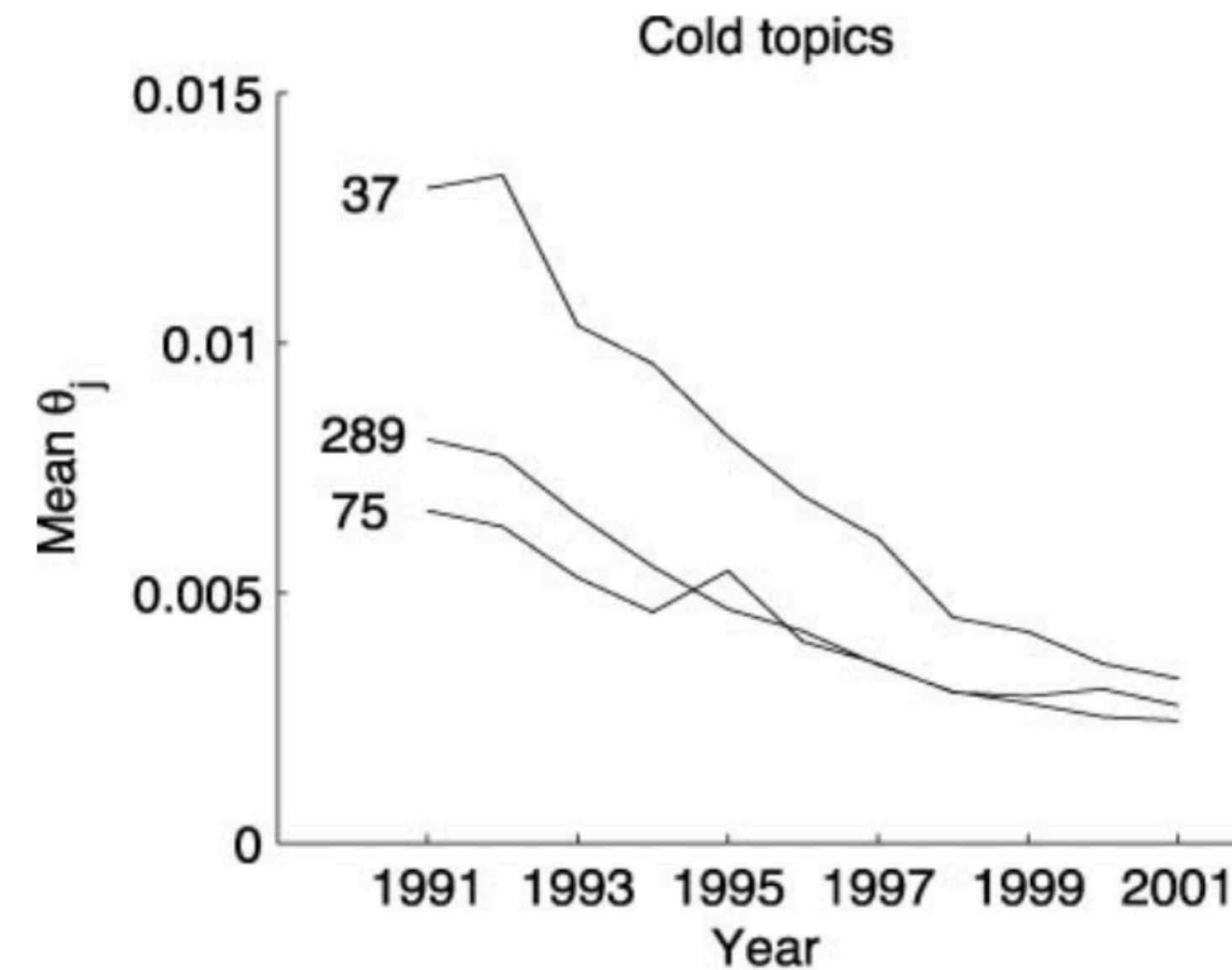
Sandeep Soni

02/08/2024

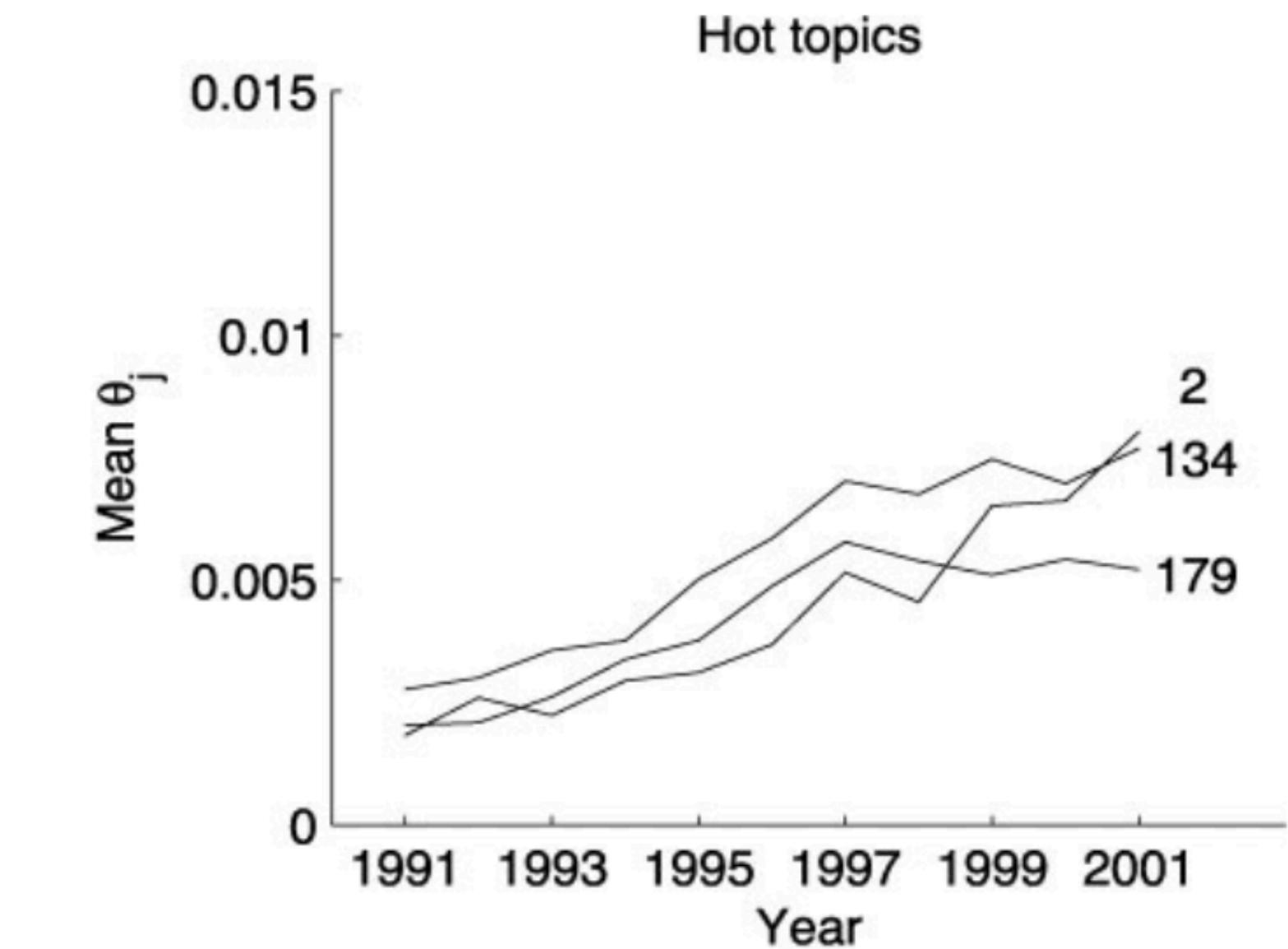
SOME POSSIBLE RESEARCH QUESTIONS

- What are some themes within a given document collection?
- What themes are popular, trending, or out-of-fashion?
- What are similar themes? What are dissimilar themes?

- What are the different research areas?
- Which ones are in vogue and which have fell out of fashion?

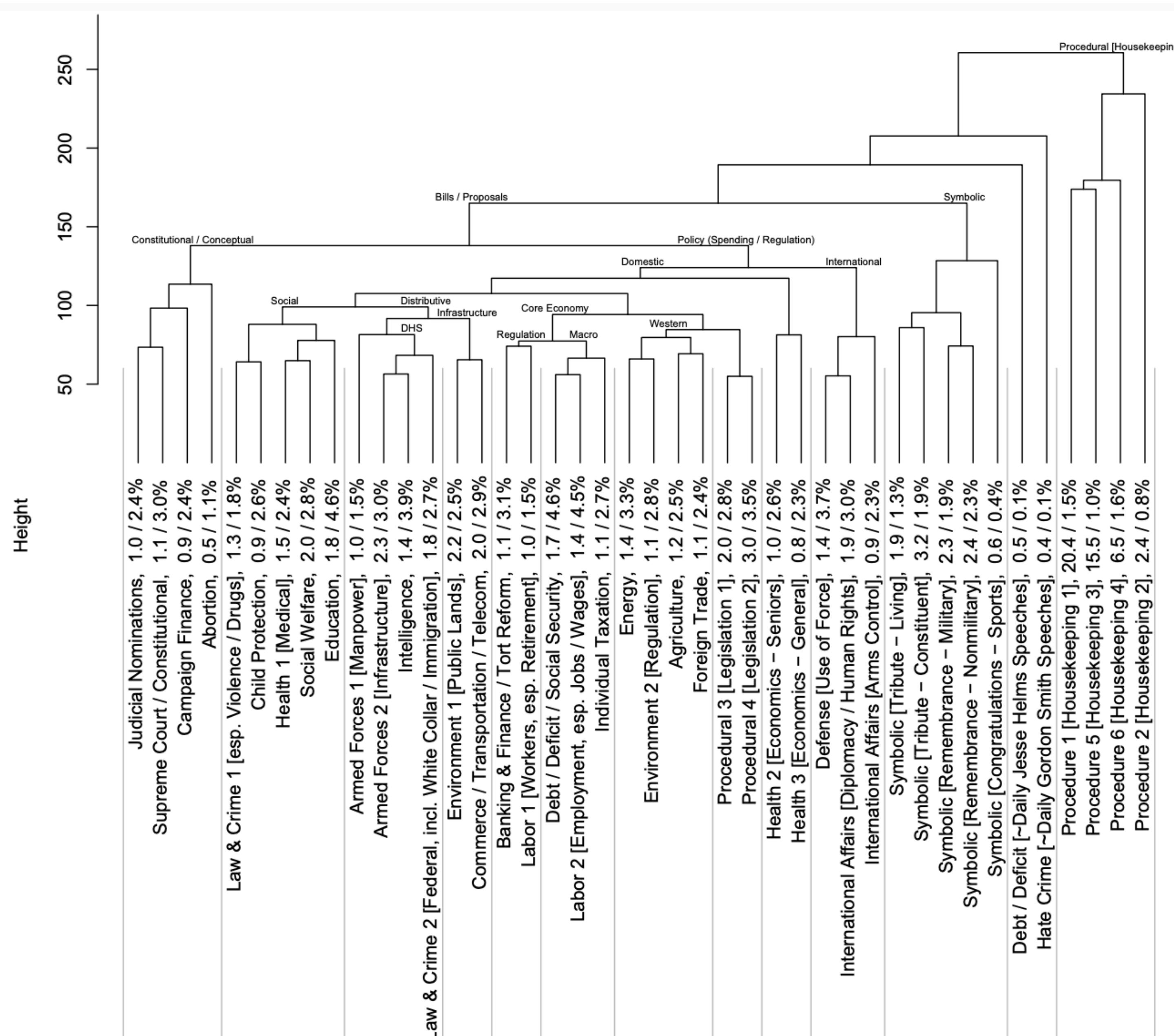


Topic 37	Topic 289	Topic 75
CDNA	KDA	ANTIBODY
AMINO	PROTEIN	ANTIBODIES
SEQUENCE	PURIFIED	MONOCLONAL
ACID	MOLECULAR	ANTIGEN
PROTEIN	MASS	IGG
ISOLATED	CHROMATOGRAPHY	MAB
ENCODING	POLYPEPTIDE	SPECIFIC
CLONED	GEL	EPITOPE
ACIDS	SDS	HUMAN
IDENTITY	BAND	MABS
CLONE	APPARENT	RECOGNIZED
EXPRESSED	LABELED	SERA

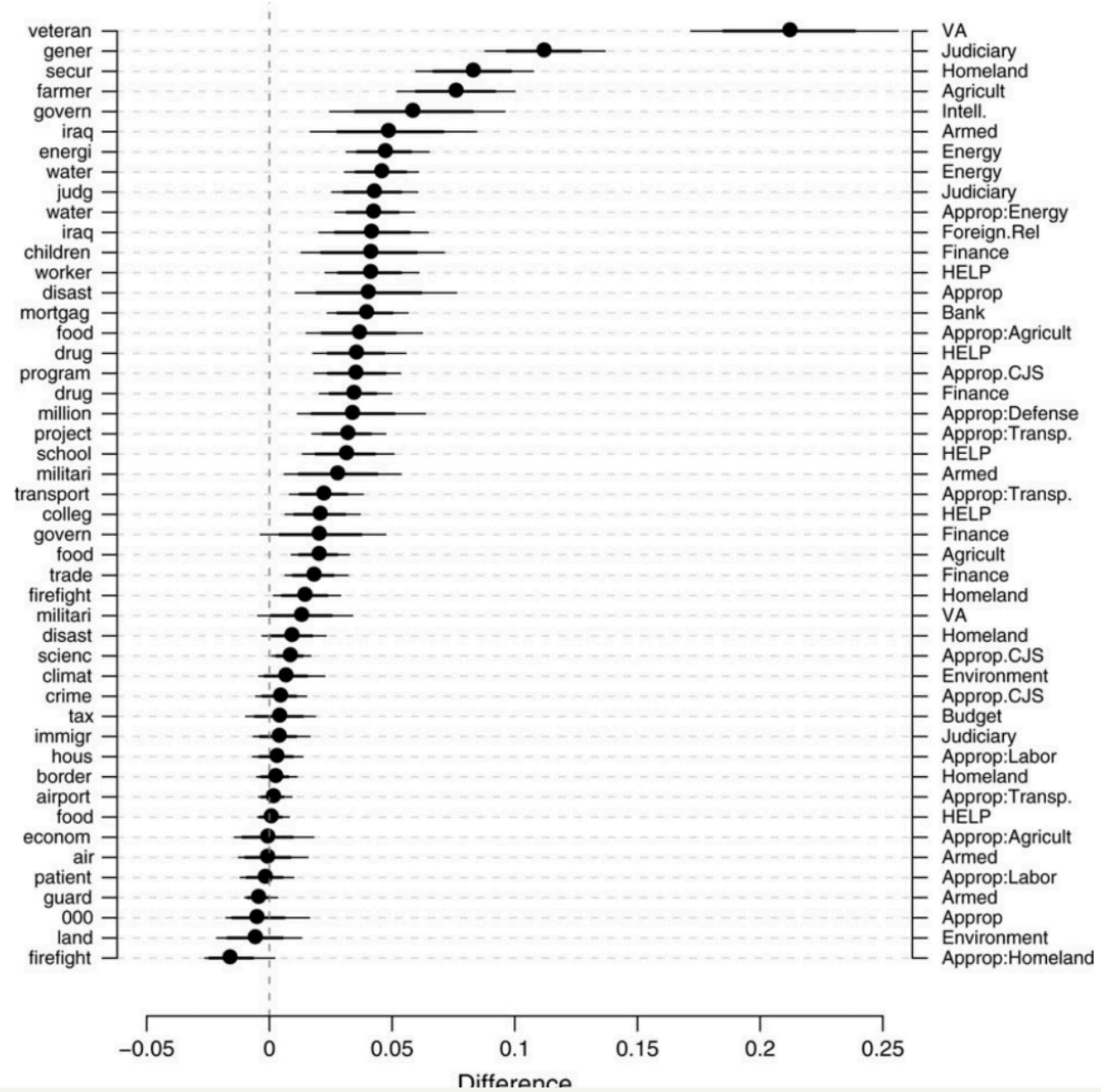


Topic 2	Topic 134	Topic 179
SPECIES	MICE	APOPTOSIS
GLOBAL	DEFICIENT	DEATH
CLIMATE	NORMAL	CELL
CO2	GENE	INDUCED
WATER	NULL	BCL
ENVIRONMENTAL	MOUSE	CELLS
YEARS	TYPE	APOPTOTIC
MARINE	HOMOZYGOUS	CASPASE
CARBON	ROLE	FAS
DIVERSITY	KNOCKOUT	SURVIVAL
OCEAN	DEVELOPMENT	PROGRAMMED
EXTINCTION	GENERATED	MEDIATED

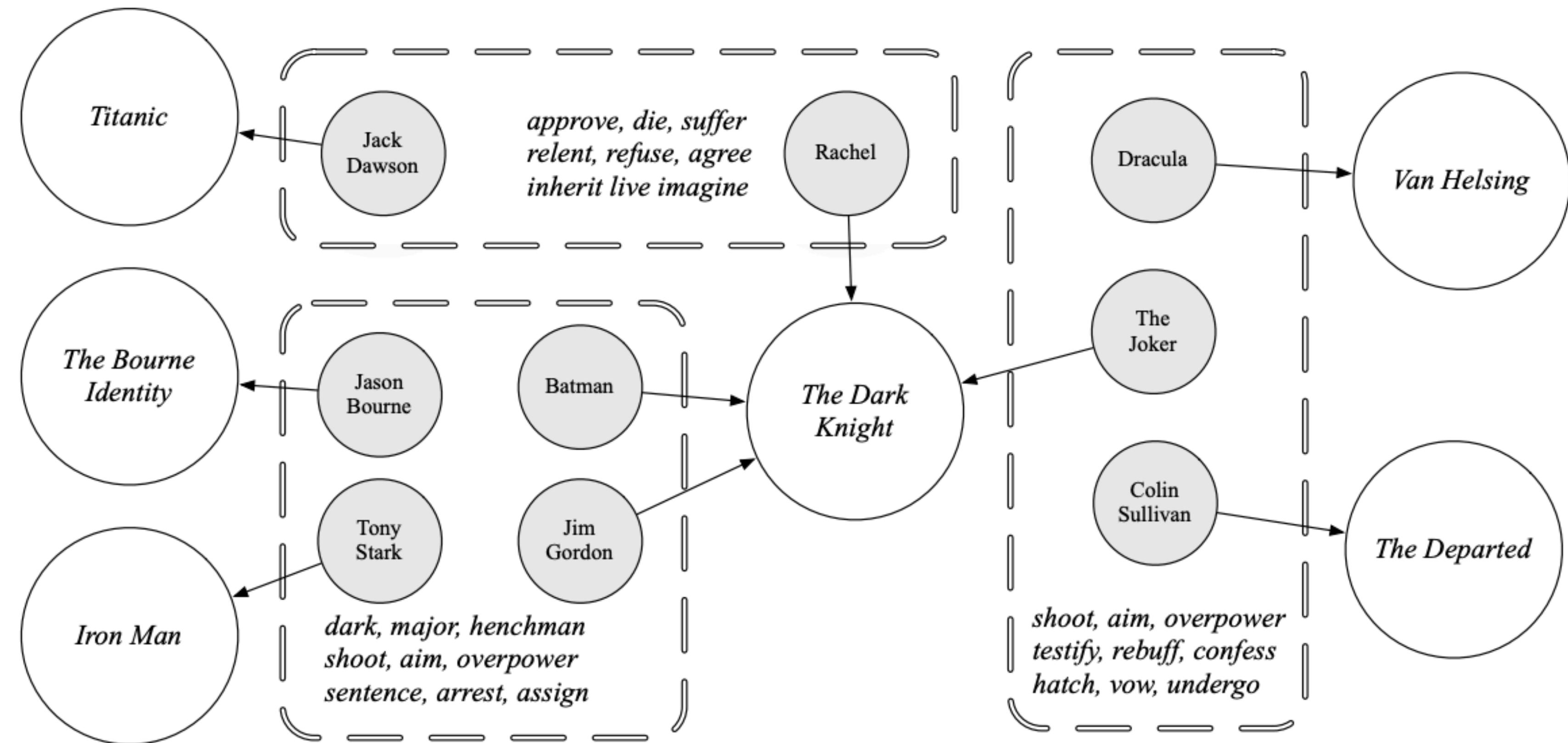
- Which topics are politicians paying attention?



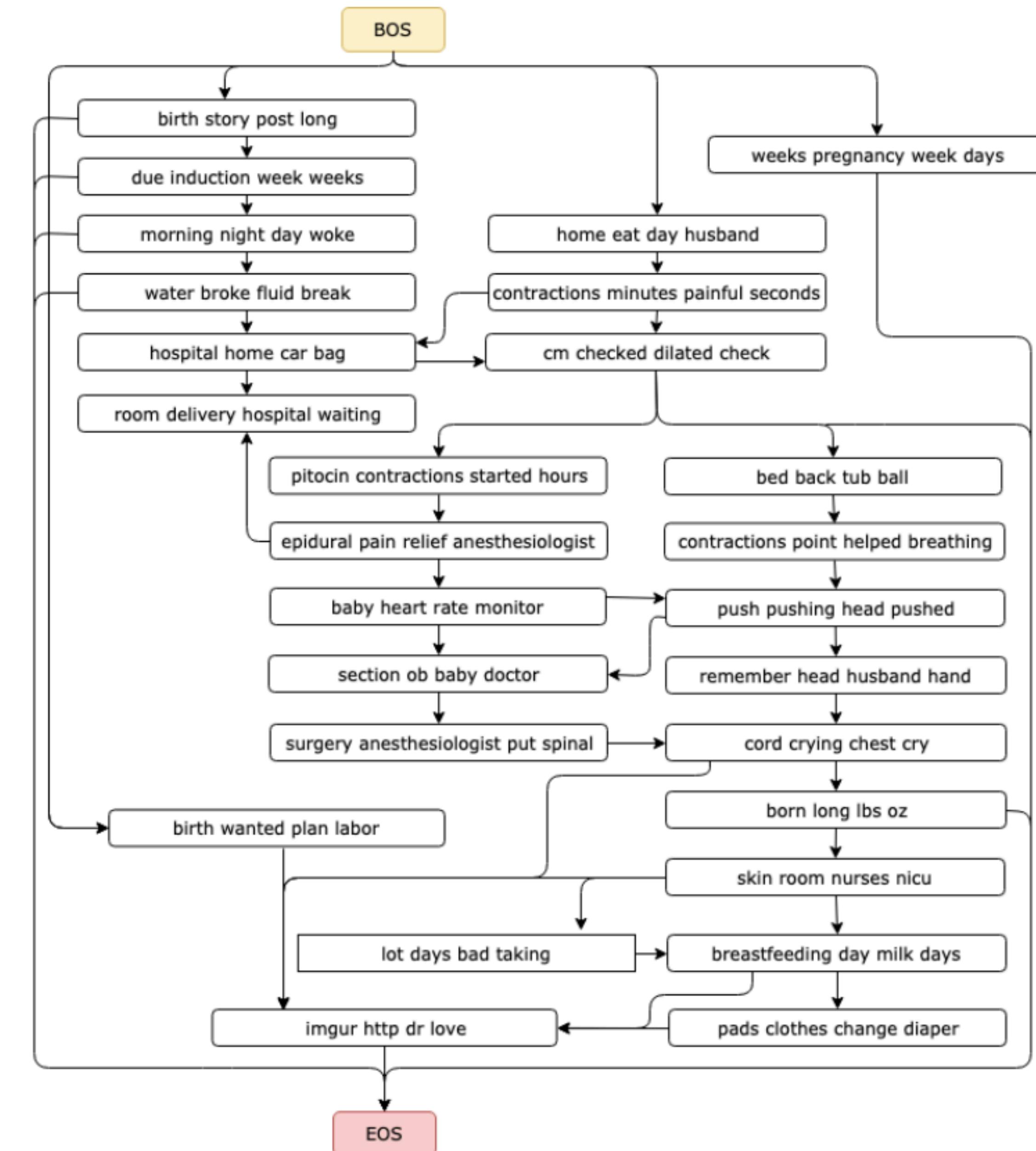
- Which topics are politicians paying attention?



- What are the different types of characters in movies?



- What are narrative sequences in birth stories?
- When do expecting mothers feel most vulnerable?



QUESTION FOR THE DAY

“How to find topics given a document collection?”

CLASS ACTIVITY I

We live in a time of extraordinary change -- change that's reshaping the way we live, the way we work, our planet, our place in the world. It's change that promises amazing medical breakthroughs, but also economic disruptions that strain working families. It promises education for girls in the most remote villages, but also connects terrorists plotting an ocean away. It's change that can broaden opportunity, or widen inequality. And whether we like it or not, the pace of this change will only accelerate.

CLASS ACTIVITY I

We live in a time of extraordinary change -- change that's reshaping the way we live, the way we work, our planet, our place in the world. It's change that promises amazing medical breakthroughs, but also **economic disruptions** that strain **working families**. It promises education for girls in the most remote villages, but also connects terrorists plotting an ocean away. It's change that can broaden **opportunity**, or widen **inequality**. And whether we like it or not, the pace of this change will only accelerate.

CLASS ACTIVITY I

We live in a time of extraordinary change -- change that's reshaping the way we live, the way we work, our **planet**, our **place** in the **world**. It's change that promises amazing medical breakthroughs, but also economic disruptions that strain working families. It promises education for girls in the most **remote villages**, but also connects terrorists plotting an **ocean** away. It's change that can broaden opportunity, or widen inequality. And whether we like it or not, the pace of this change will only accelerate.

AGENDA

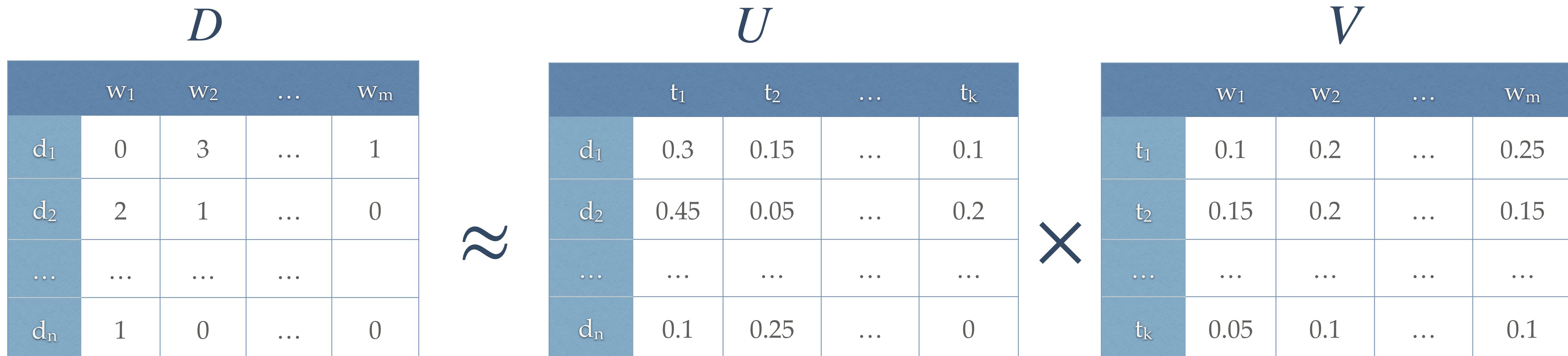
- Intro to topic models
- LDA
- Variations of LDA
- Evaluation

DOCUMENT-TERM MATRIX

- If this matrix is called D , then D_{ij} is the number of times the w_j appeared in d_i
- Every document is a distribution over word counts

	w_1	w_2	\dots	w_m
d_1	0	3	\dots	1
d_2	2	1	\dots	0
\dots	\dots	\dots	\dots	\dots
d_n	1	0	\dots	0

MATRIX DECOMPOSITION



- $D_{i,j} \approx \sum_{r=1}^k U_{i,r} \times V_{r,j}$
- Each document can be thought of as a mixture of K “topics”
- Each topic is a distribution over words

LATENT SEMANTIC ANALYSIS

	W ₁	W ₂	...	W _m
d ₁	0	3	...	1
d ₂	2	1	...	0
...	
d _n	1	0	...	0

≈

	t ₁	t ₂	...	t _k
d ₁	0.3	0.15	...	0.1
d ₂	0.45	0.05	...	0.2
...
d _n	0.1	0.25	...	0

×

	W ₁	W ₂	...	W _m
t ₁	0.1	0.2	...	0.25
t ₂	0.15	0.2	...	0.15
...
t _k	0.05	0.1	...	0.1

LATENT SEMANTIC ANALYSIS

	W ₁	W ₂	...	W _m
d ₁	0	3	...	1
d ₂	2	1	...	0
...	
d _n	1	0	...	0

≈

	t ₁	t ₂	...	t _k
d ₁	0.3	0.15	...	0.1
d ₂	0.45	0.05	...	0.2
...
d _n	0.1	0.25	...	0

×

	W ₁	W ₂	...	W _m
t ₁	0.1	0.2	...	0.25
t ₂	0.15	0.2	...	0.15
...
t _k	0.05	0.1	...	0.1

Can we do this probabilistically?

PROBABILITY BASICS

- Let's assume that we have a vocabulary $V = \{\text{"the"}, \text{"dog"}, \text{"runs"}\}$

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	
"the"	"runs"	15	
"dog"	"runs"	40	
"dog"	"the"	25	
"runs"	"the"	30	
"runs"	"dog"	45	

PROBABILITY BASICS

- $P(X, Y) = \frac{\text{Count}(X, Y)}{N}$
- N is the number of pairwise cooccurrences
- This is called the joint probability

X	Y	Count(X,Y)	P(X,Y)
“the”	“dog”	45	0.225
“the”	“runs”	15	0.075
“dog”	“runs”	40	0.2
“dog”	“the”	25	0.125
“runs”	“the”	30	0.15
“runs”	“dog”	45	0.225

MARGINALIZATION

- What is $P(X=\text{"dog"})$?

X	Y	Count(X,Y)	P(X,Y)
“the”	“dog”	45	0.225
“the”	“runs”	15	0.075
“dog”	“runs”	40	0.2
“dog”	“the”	25	0.125
“runs”	“the”	30	0.15
“runs”	“dog”	45	0.225

MARGINALIZATION

- $P(X) = \sum_y P(X, Y = y)$
- $P(X = \text{"dog"}) = P(X = \text{"dog"}, Y = \text{"runs"}) + P(X = \text{"dog"}, Y = \text{"the"})$
- $P(X = \text{"dog"}) = 0.2 + 0.125 = 0.325$

X	Y	Count(X,Y)	P(X,Y)
"dog"	"runs"	40	0.2
"dog"	"the"	25	0.125

CONDITIONAL PROBABILITY

- $P(X | Y = \text{"dog"})$

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	0.225
"the"	"runs"	15	0.075
"dog"	"runs"	40	0.2
"dog"	"the"	25	0.125
"runs"	"the"	30	0.15
"runs"	"dog"	45	0.225

CONDITIONAL PROBABILITY

- $P(X|Y=\text{"dog"})$
- This is a distribution
- Since it should be a probability distribution, we'll normalize it to sum up to 1

X	Y	Count(X,Y)	P(X,Y)
"the"	"dog"	45	0.225
"runs"	"dog"	45	0.225

X	Y	Count(X,Y)	P(X Y=\text{"dog"})
"the"	"dog"	45	0.5
"runs"	"dog"	45	0.5

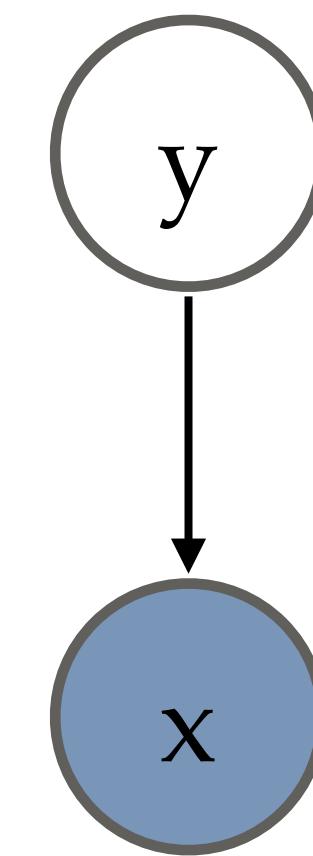
FACTORIZATION

- $P(X, Y) = P(X)P(Y|X)$
- $P(X, Y) = P(Y)P(X|Y)$

X	Y	Count(X,Y)	P(X,Y)
“the”	“dog”	45	0.225
“the”	“runs”	15	0.075
“dog”	“runs”	40	0.2
“dog”	“the”	25	0.125
“runs”	“the”	30	0.15
“runs”	“dog”	45	0.225

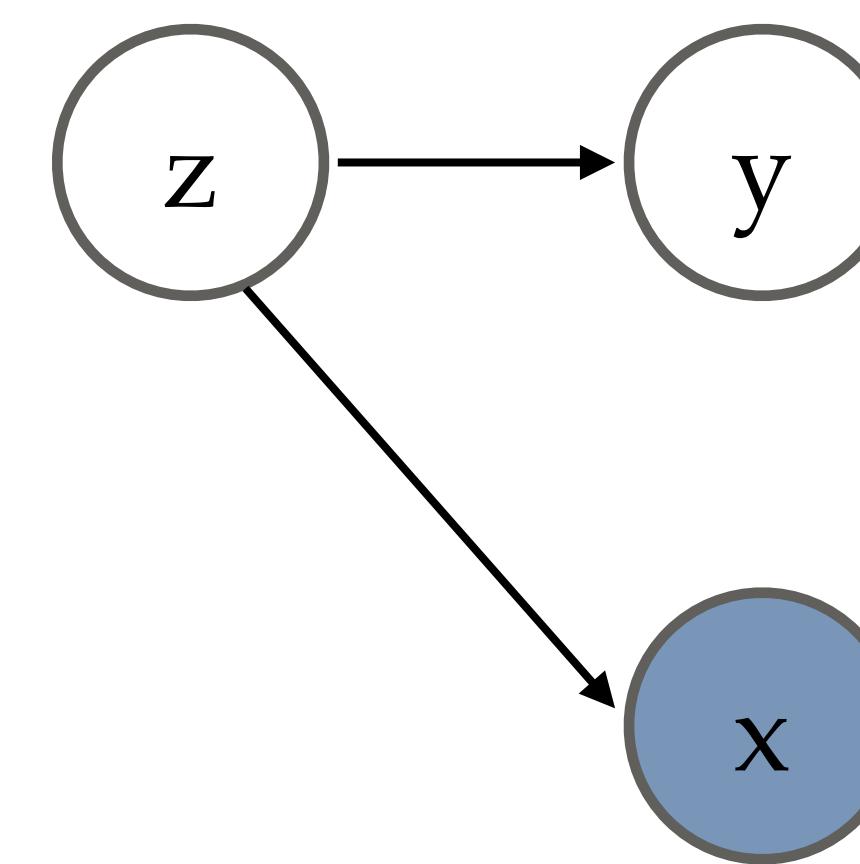
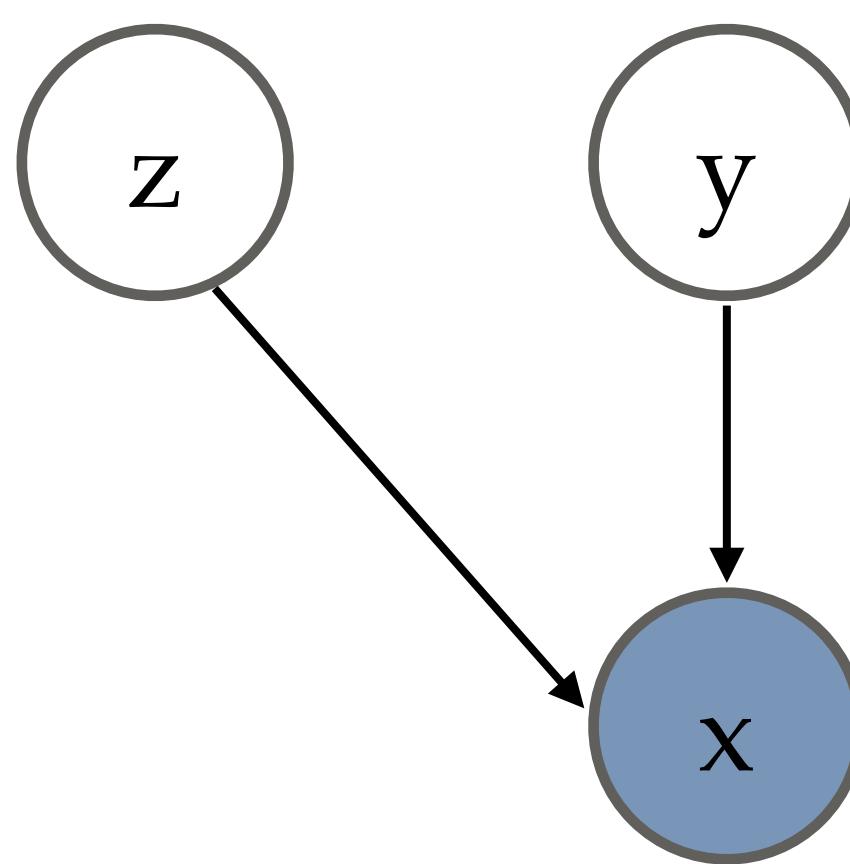
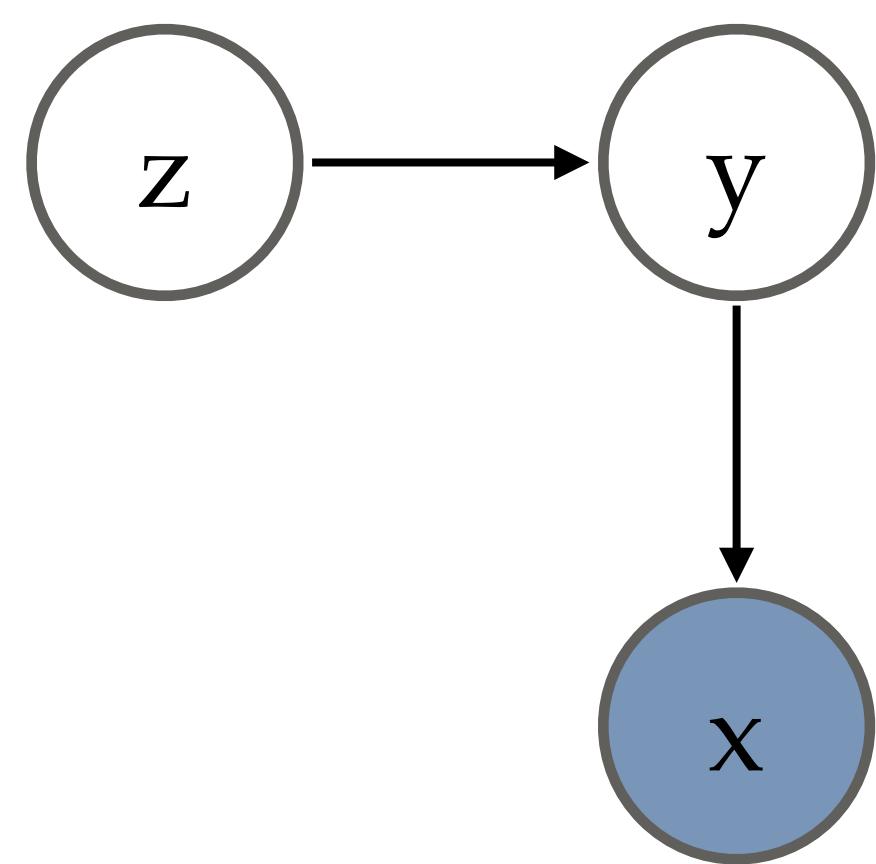
PROBABILISTIC GRAPHICAL MODELS

- All nodes are random variables; dark nodes are observed and the others are latent
- Arrows indicate conditional relationships
- Visual specification of how the joint distribution factors

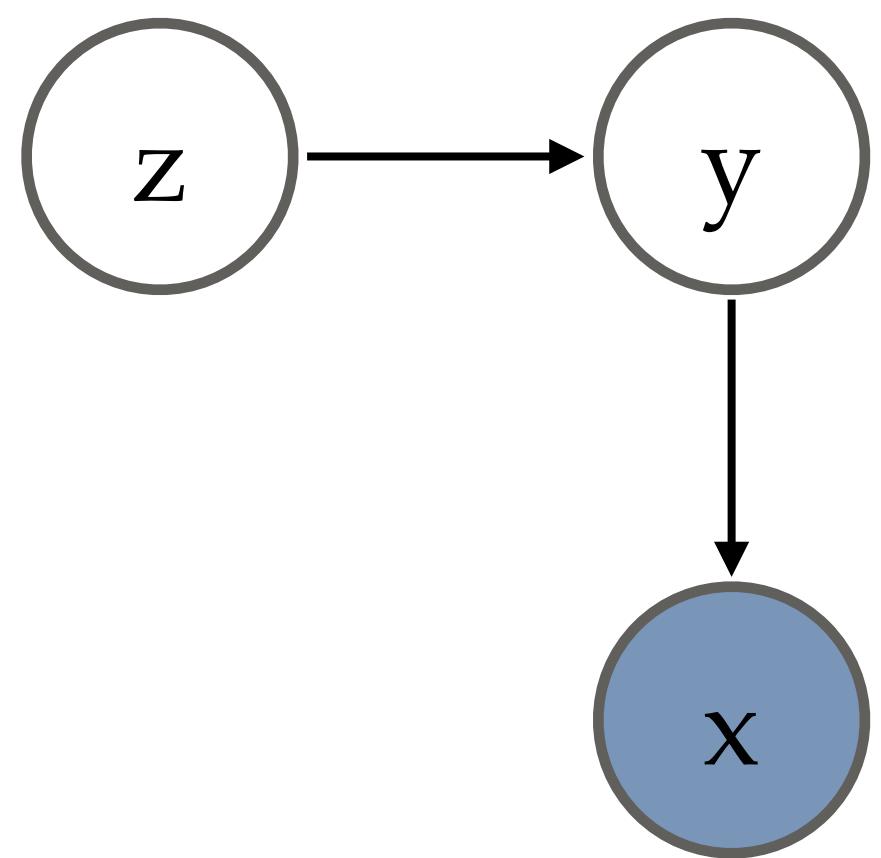


$$P(x, y) = P(y)P(x|y)$$

CLASS ACTIVITY II

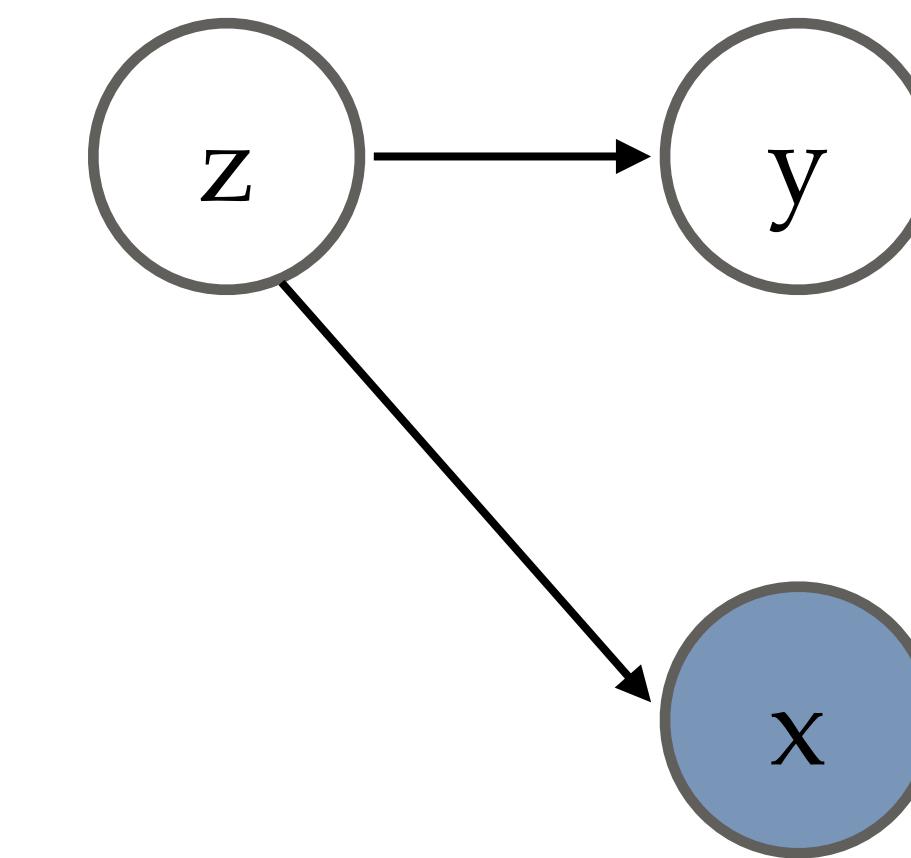


CLASS ACTIVITY II



$$P(x, y, z) = P(z)P(y | z)P(x | y)$$

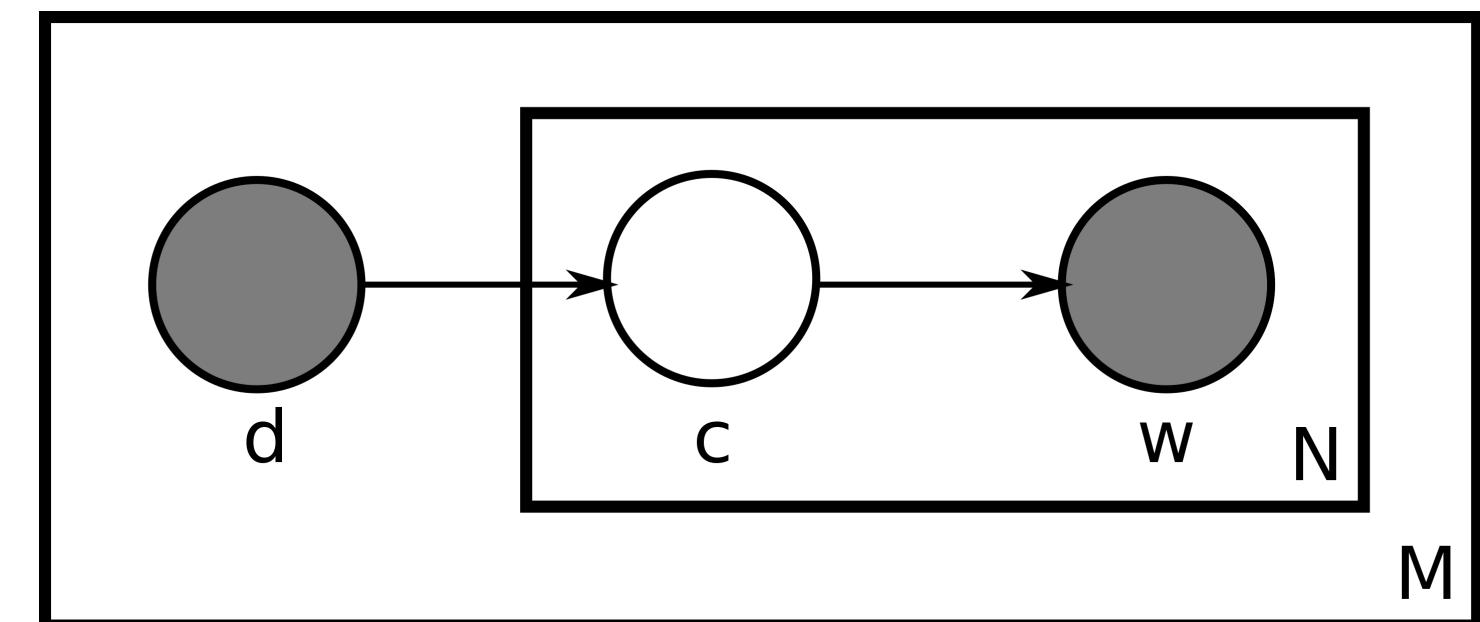
$$P(x, y, z) = P(z)P(y)P(x | y, z)$$



$$P(x, y, z) = P(z)P(y | z)P(x | z)$$

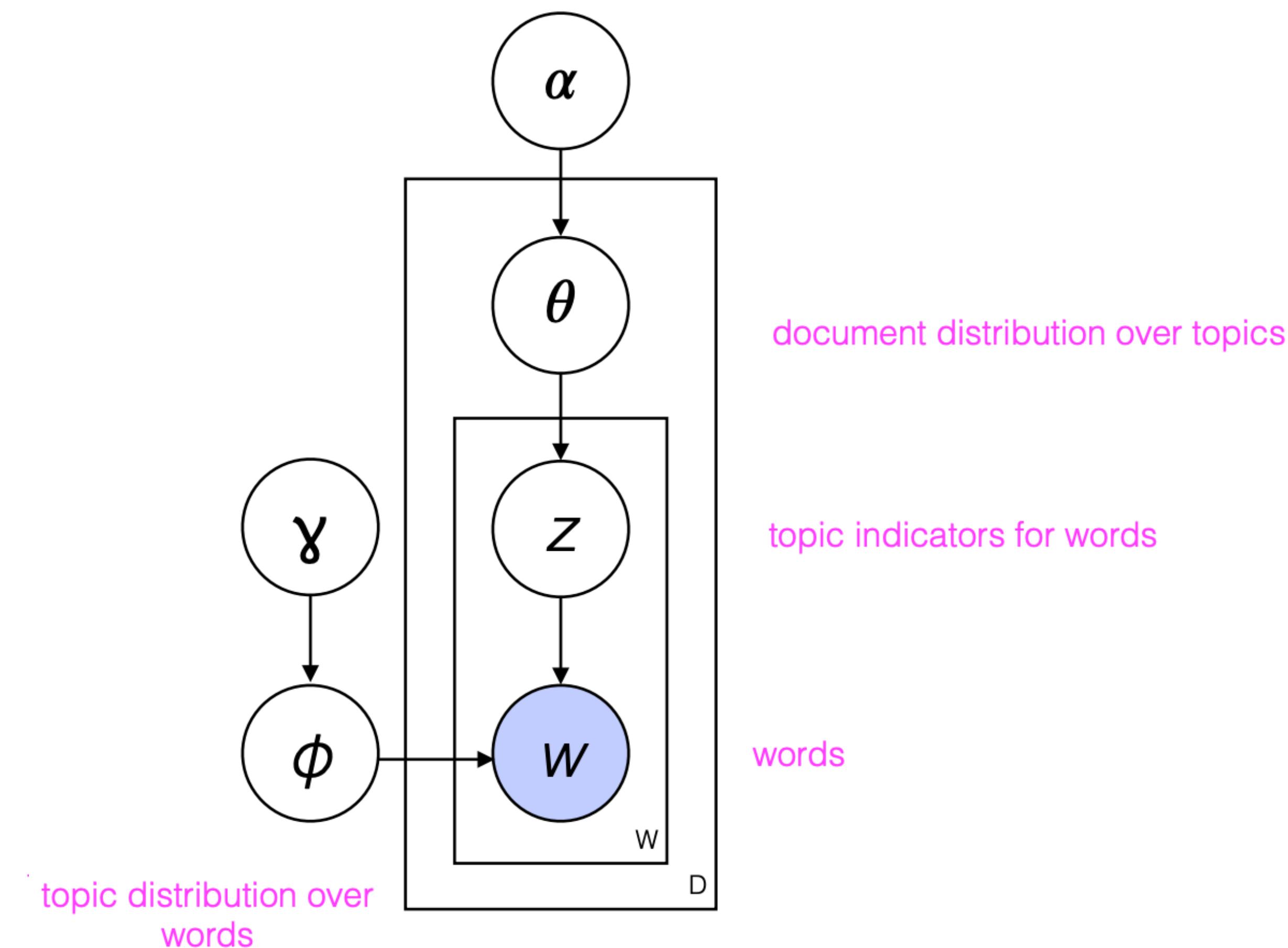
PROBABILISTIC LATENT SEMANTIC ANALYSIS

- There are M documents and N words in the vocabulary.
- Blocks surrounding the circle means repeating structure
- We observe document indices and the words within each document
- c is the topic assigned to every word

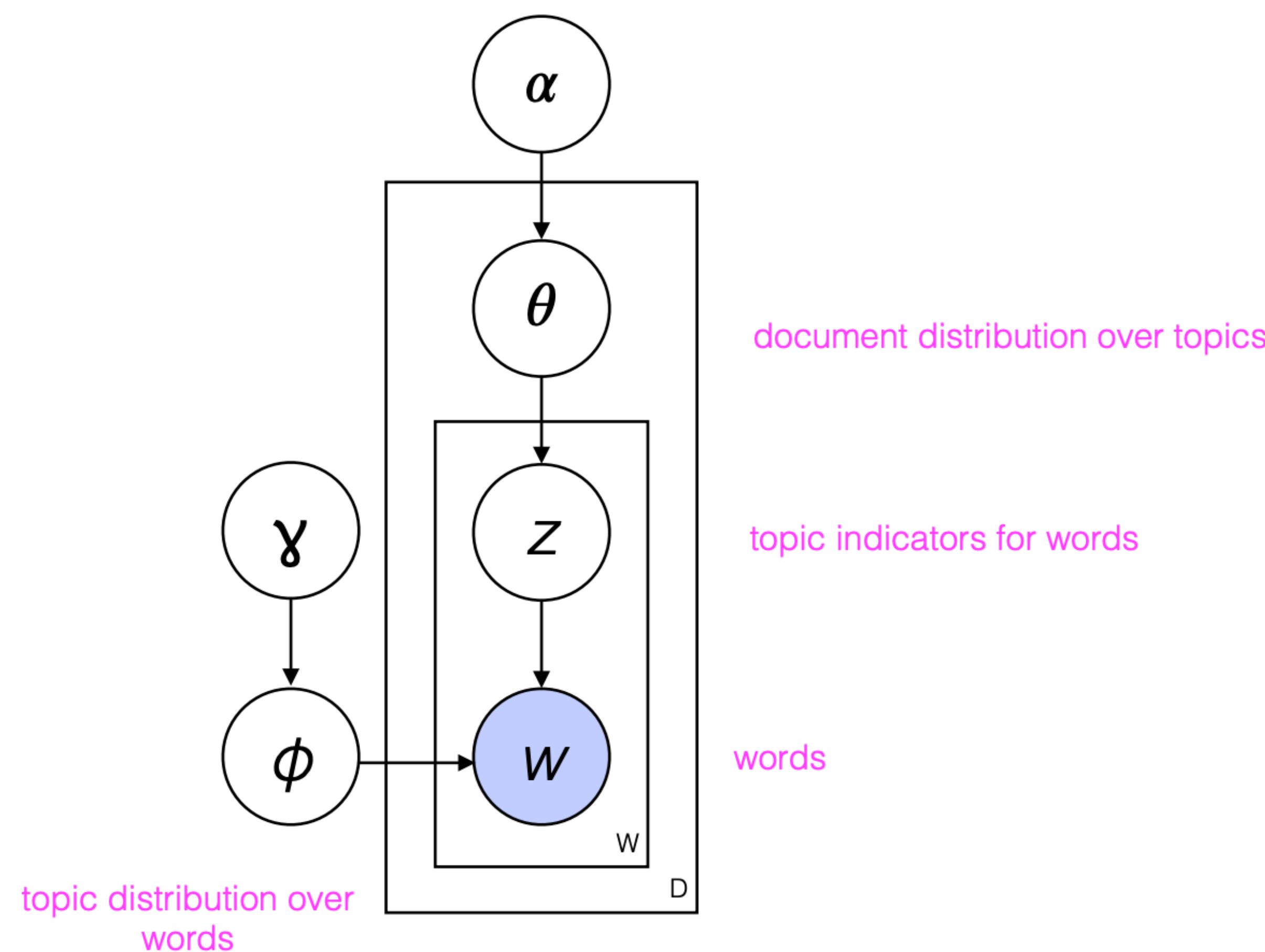


- Generative story:
 - Generate doc indices
 - For every doc index
 - generate the topic
 - Generate word for the topic

LATENT DIRICHLET ALLOCATION

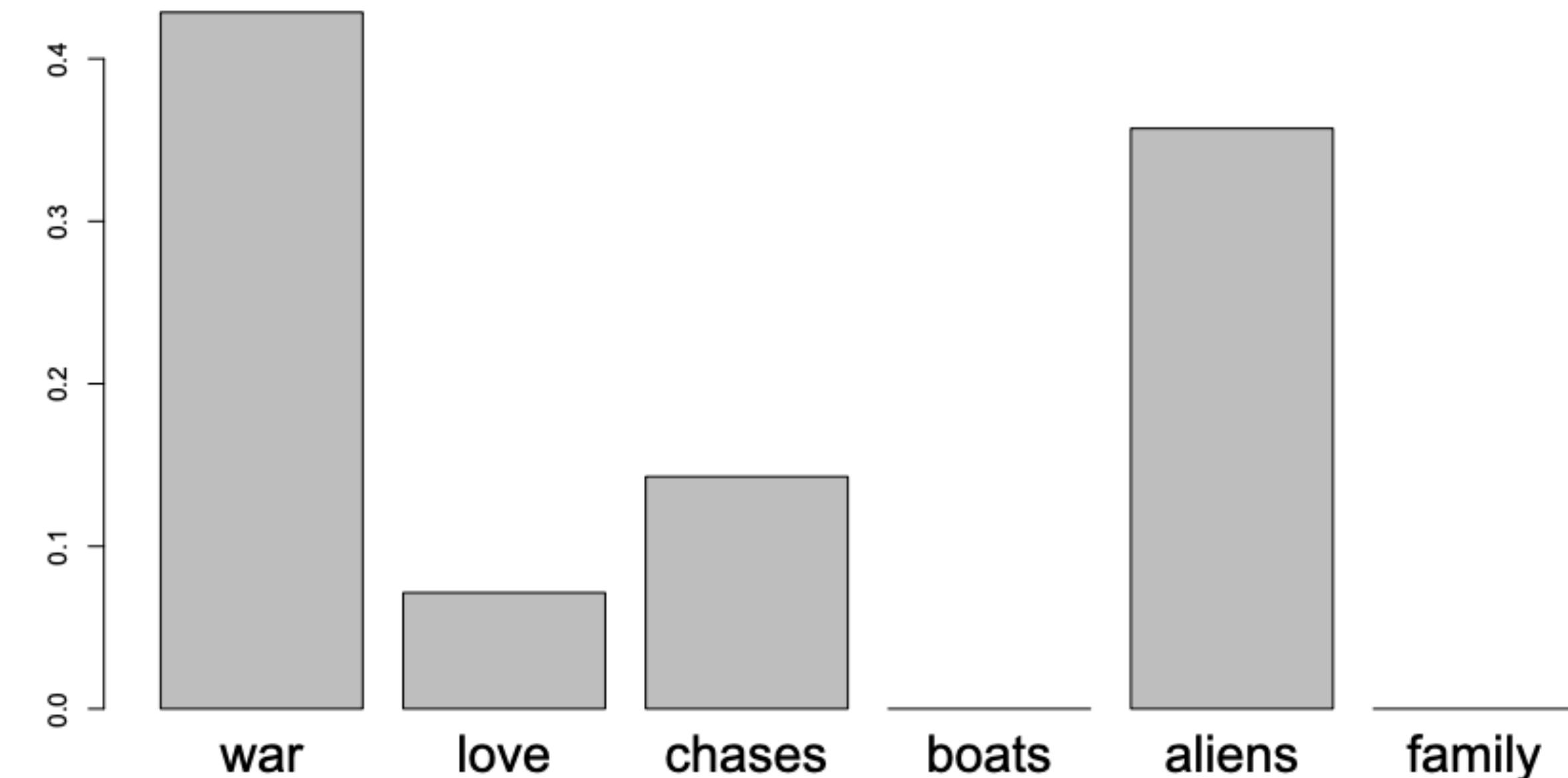


TOPIC MODELS

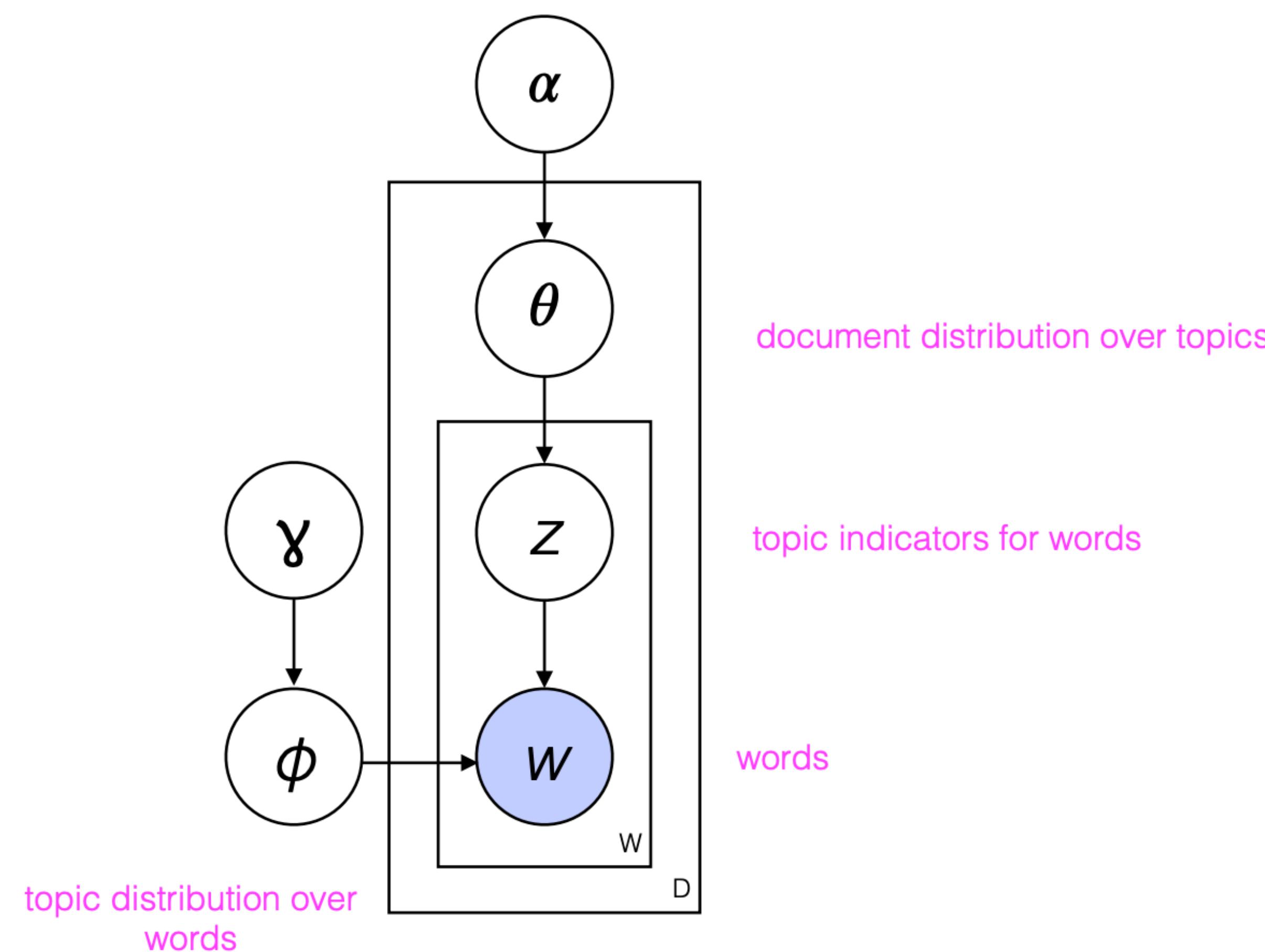


TOPIC MODELS

- A document is conceptualized as a distribution over topics
- θ gives this distribution

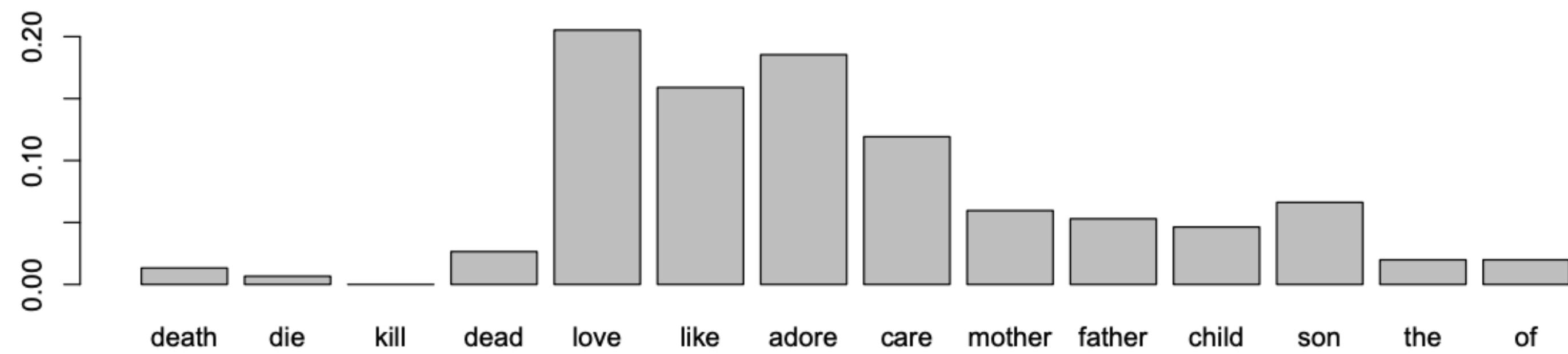


TOPIC MODELS



TOPIC MODELS

- A topic is a distribution over words
- ϕ gives this distribution
- $P(\text{"like"} | \text{topic}=\text{love})=0.15$

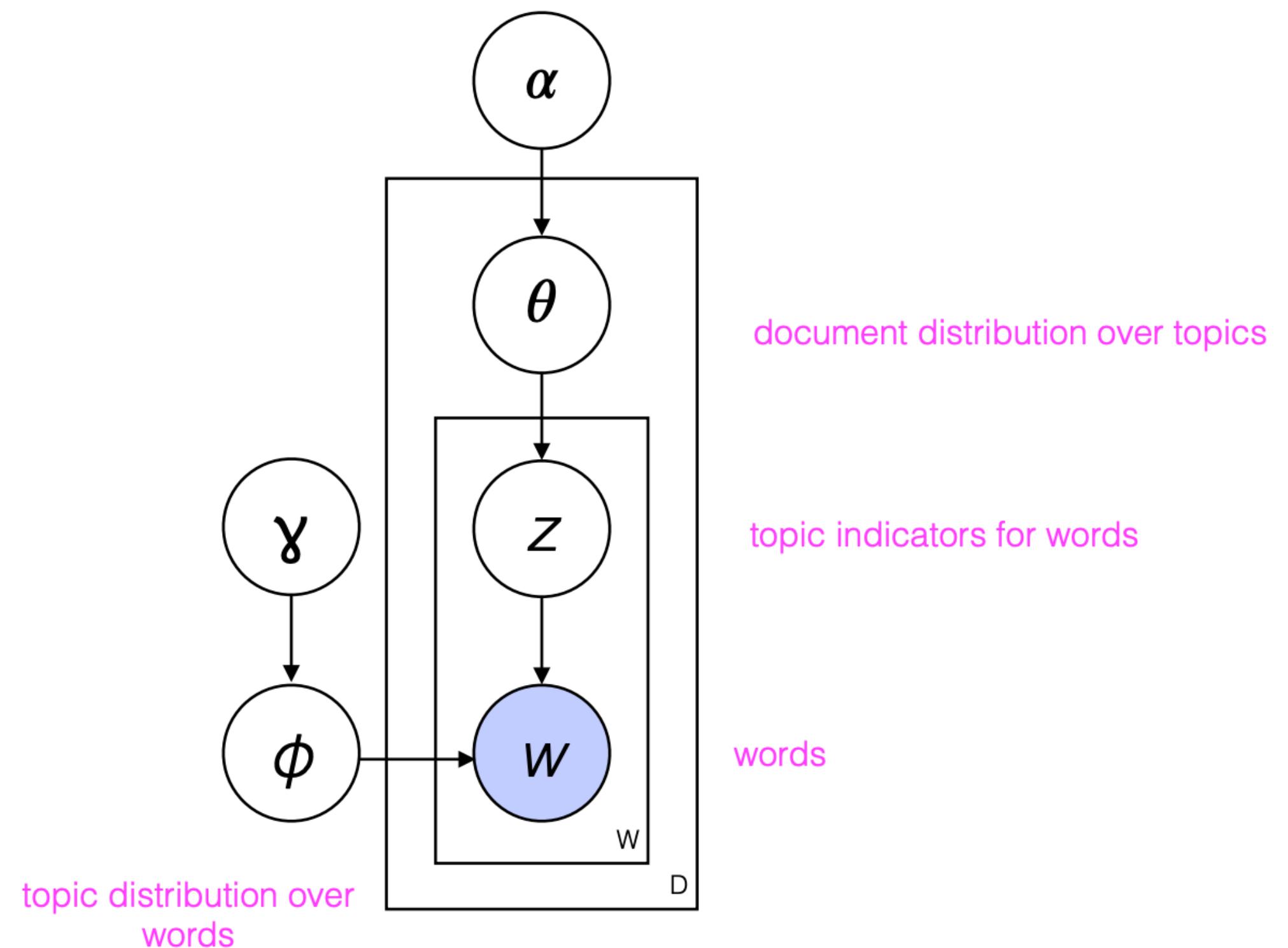


TOPIC MODELS

- For every document, you will have a different topic distribution
- For every topic, you will have a different word distribution

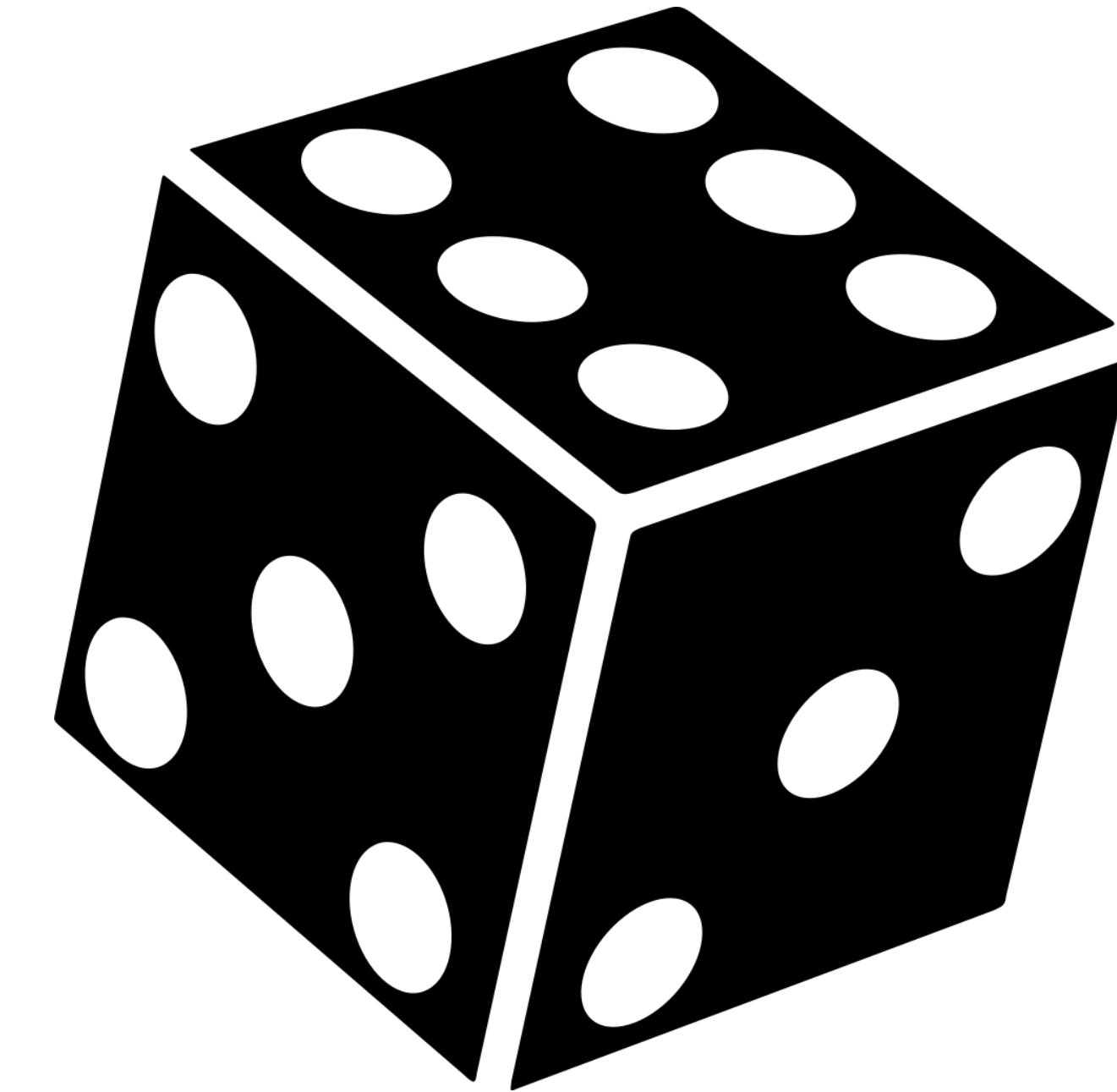
GENERATIVE STORY

- For every document d :
 - Generate topic distribution θ_d
 - For every word position in d :
 - Draw topic $z_{w,d}$
 - Draw word w from $\phi_{z_{w,d}}$



GENERATIVE STORY

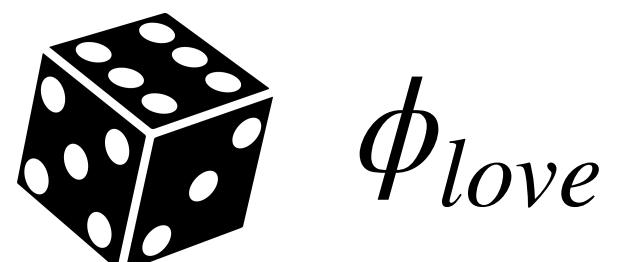
- For every document d :
 - Generate topic distribution θ_d
 - For every word position in d :
 - Draw topic $z_{w,d}$
 - Draw word w from $\phi_{z_{w,d}}$



GENERATE FIRST WORD

$\theta_d = \{“love”=0.5,$
 $“war”=0.2,$
 $“aliens” = 0.3\}$

- $z_{w,d} = “love”$
- $w = \text{like}$

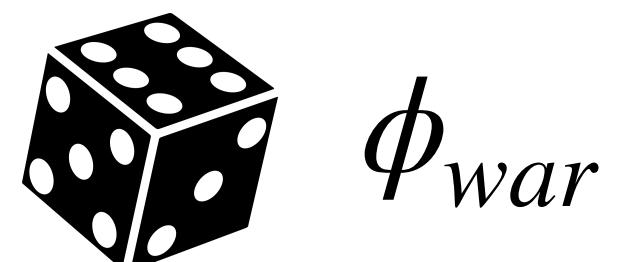


like

GENERATE NEXT WORD

$\theta_d = \{“love”=0.5,$
 $“war”=0.2,$
 $“aliens” = 0.3\}$

- $z_{w,d} = “war”$
- $w = \text{peace}$

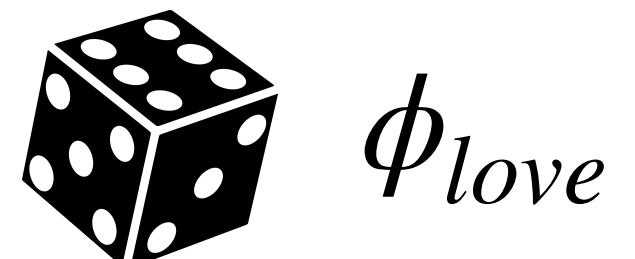


like peace

GENERATE NEXT WORD

$\theta_d = \{“love”=0.5,$
 $“war”=0.2,$
 $“aliens” = 0.3\}$

- $z_{w,d} = “love”$
- $w = \text{love}$

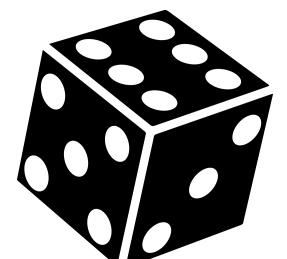


like peace love

GENERATE NEXT WORD

$\theta_d = \{“love”=0.5,$
 $“war”=0.2,$
 $“aliens” = 0.3\}$

- $z_{w,d} = “aliens”$
- $w = \text{people}$



ϕ_{aliens}

like peace love people

AND SO ON

like peace love people ...

SOFTWARE

SOFTWARE

- Mallet (<https://mimno.github.io/Mallet/>)

SOFTWARE

- Mallet (<https://mimno.github.io/Mallet/>)
- Gensim (<https://radimrehurek.com/gensim/>)

SOFTWARE

- Mallet (<https://mimno.github.io/Mallet/>)
- Gensim (<https://radimrehurek.com/gensim/>)
- BertTopic (<https://maartengr.github.io/BERTopic/index.html>)

INFERENCE

INFERENCE

- How do we estimate the distributions?

INFERENCE

- How do we estimate the distributions?
 - Markov chain Monte Carlo (Metropolis Hastings, Gibbs sampling)

INFERENCE

- How do we estimate the distributions?
 - Markov chain Monte Carlo (Metropolis Hastings, Gibbs sampling)
 - Variational methods

INFERENCE

- How do we estimate the distributions?
 - Markov chain Monte Carlo (Metropolis Hastings, Gibbs sampling)
 - Variational methods
 - Hybrid methods

THINGS TO REMEMBER

THINGS TO REMEMBER

- Priors matter [Wallach et. al. 2009]

THINGS TO REMEMBER

- Priors matter [Wallach et. al. 2009]
- Preprocessing matters [Schofield et. al. 2016, 2017]

THINGS TO REMEMBER

- Priors matter [Wallach et. al. 2009]
- Preprocessing matters [Schofield et. al. 2016, 2017]
- Inference techniques matter

LDA ASSUMPTIONS

- A word gets a single topic assigned
- A document gets a single distribution of topics
- No sequential info. used
- No other structure assumed except word and document identities

SOME POPULAR TOPIC MODELS

- Probabilistic graphical models offer much flexibility in modeling the generation of documents

HIERARCHICAL TOPIC MODELING

- Model the data generating process as a nested Chinese restaurant process

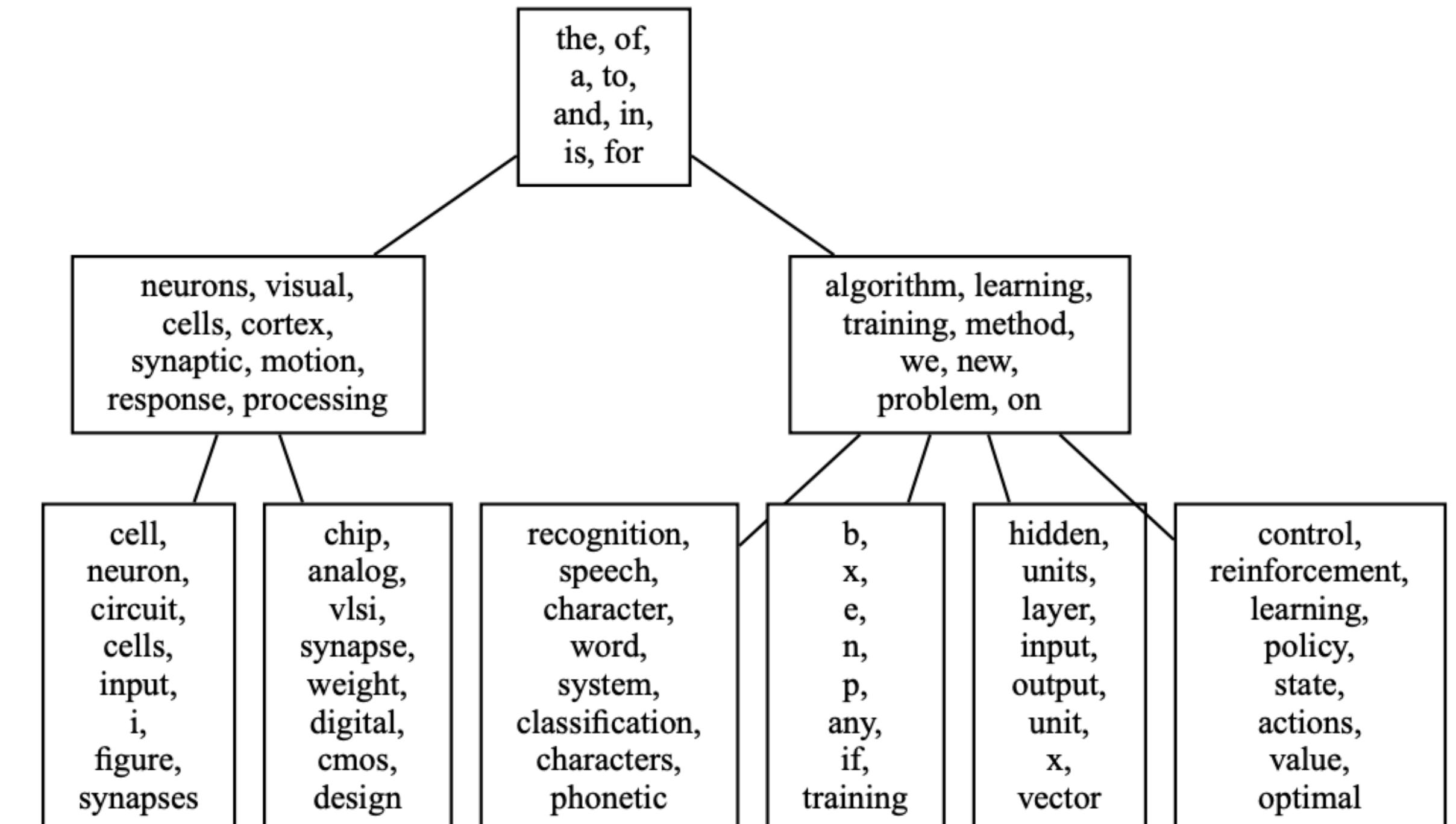


Figure 5: A topic hierarchy estimated from 1717 abstracts from NIPS01 through NIPS12. Each node contains the top eight words from its corresponding topic distribution.

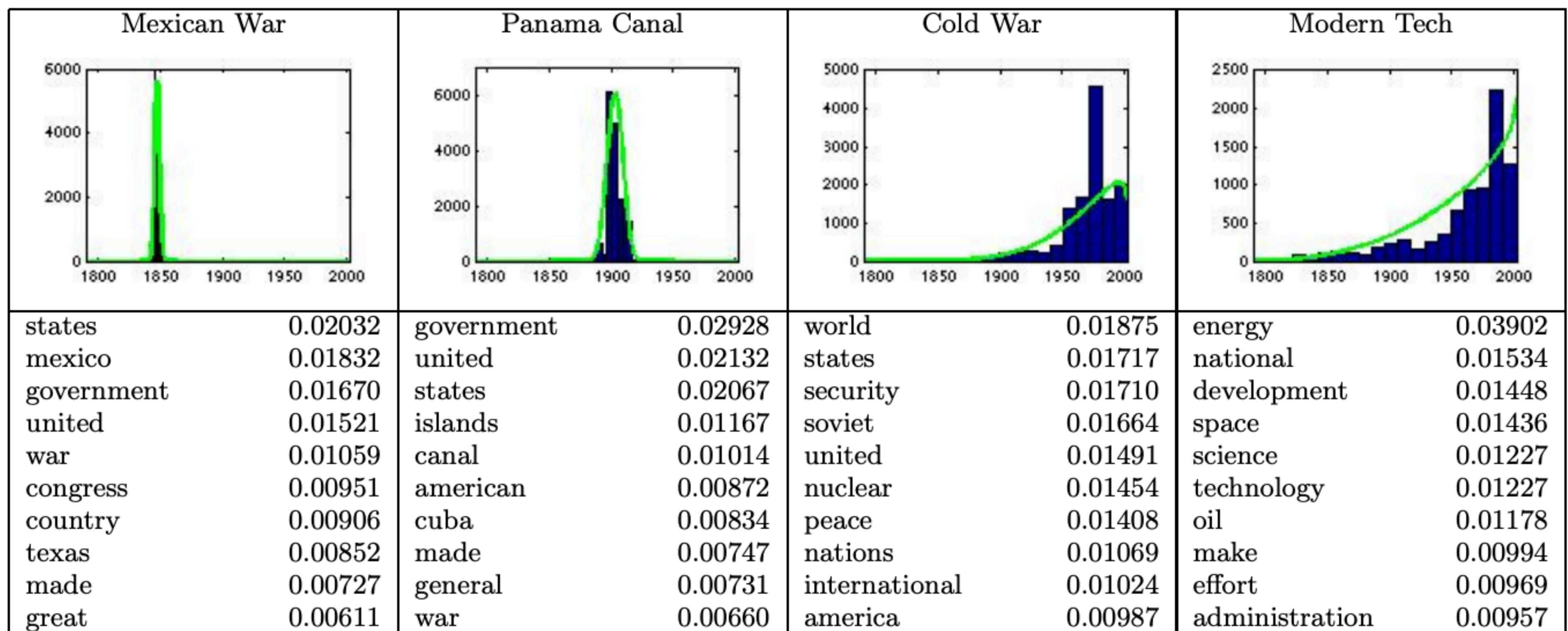
GEOGRAPHICAL TOPIC MODEL

- Model text and author geolocation

	“basketball”	“popular music”	“daily life”	“emoticons”	“chit chat”
	PISTONS KOBE LAKERS game DUKE NBA CAVS STUCKEY JETS KNICKS	album music beats artist video #LAKERS ITUNES tour produced vol	tonight shop weekend getting going chilling ready discount waiting iam	:) haha :d :(;) :p xd :/ hahaha hahah	lol smh jk yea wyd coo ima wassup somethin jp
Boston		CELTICS victory BOSTON CHARLOTTE	playing daughter PEARL alive war comp	BOSTON	;p gna loveee <i>ese</i> exam suttin sippin
N. California		THUNDER KINGS GIANTS pimp trees clap	SIMON dl mountain seee	6am OAKLAND	<i>pues</i> hella koo SAN fckn hella flirt hut iono OAKLAND
New York		NETS KNICKS	BRONX	iam cab	oww wasssup nm
Los Angeles		#KOBE #LAKERS AUSTIN	#LAKERS load HOLLYWOOD imm MICKEY TUPAC	omw tacos hr HOLLYWOOD	af <i>papi</i> raining th bomb coo HOLLYWOOD wyd coo af <i>nada</i> tacos messin fasho bomb
Lake Erie		CAVS CLEVELAND OHIO BUCKS od COLUMBUS	premiere prod joint TORONTO onto designer CANADA village burr	stink CHIPOTLE tipsy	foul <i>WIZ</i> salty excuses lames officer lastnight ;d blvd BIEBER hve OHIO

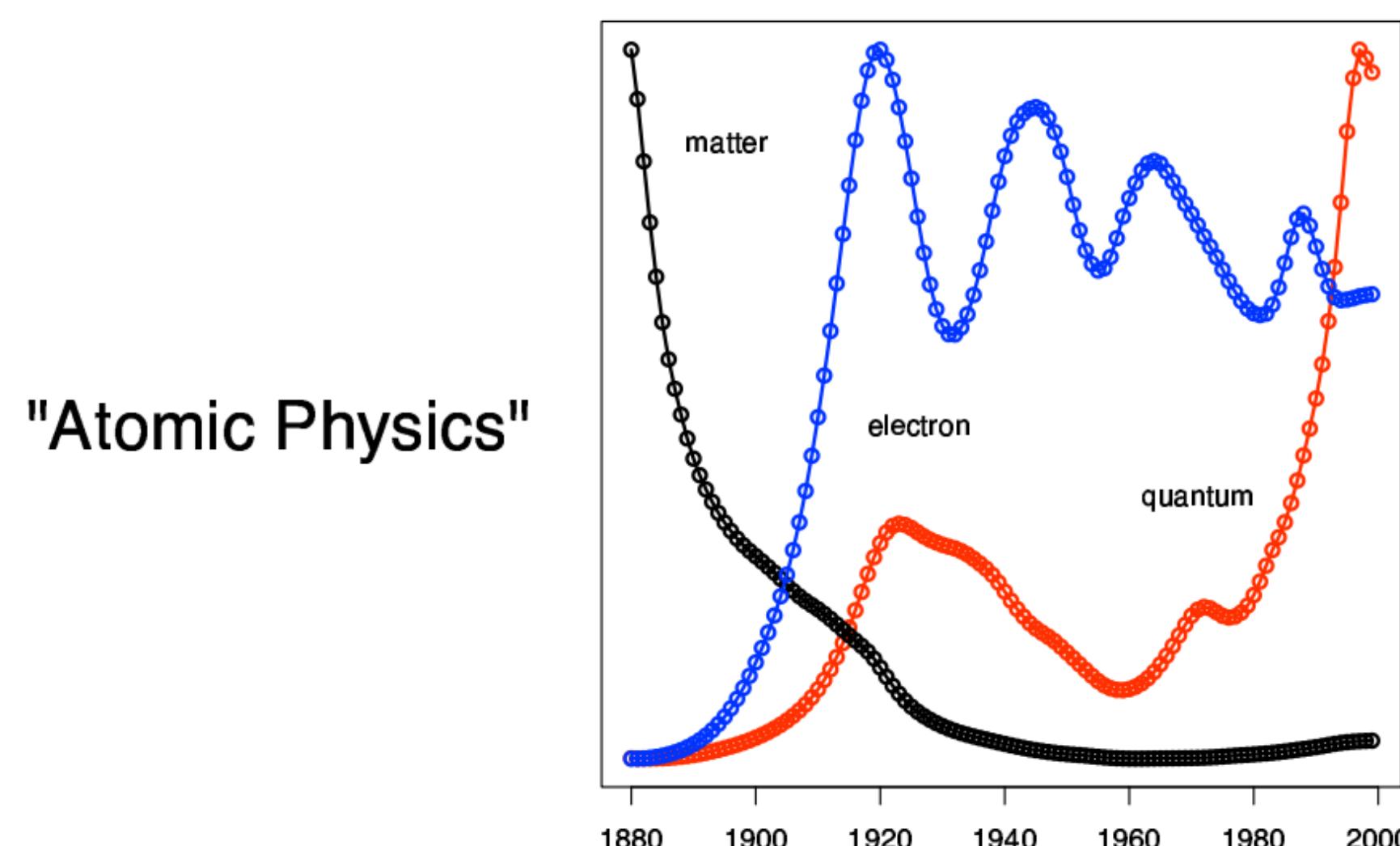
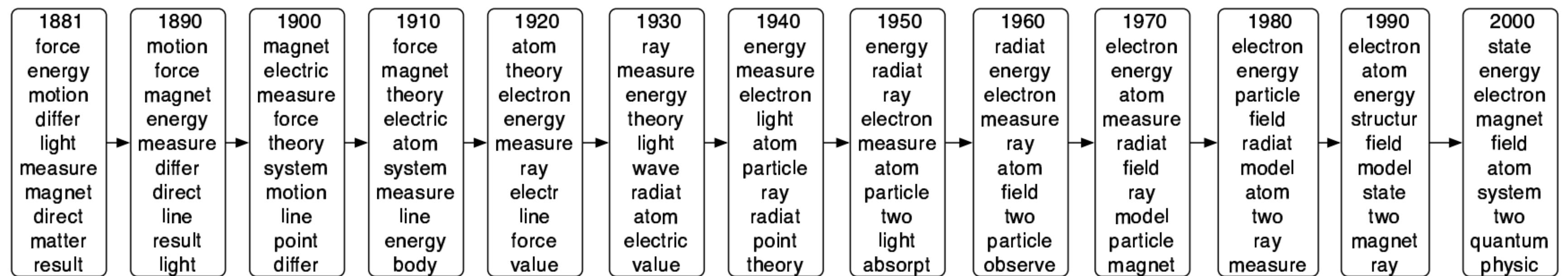
TOPICS OVER TIME

- Text and time
- Time is discrete



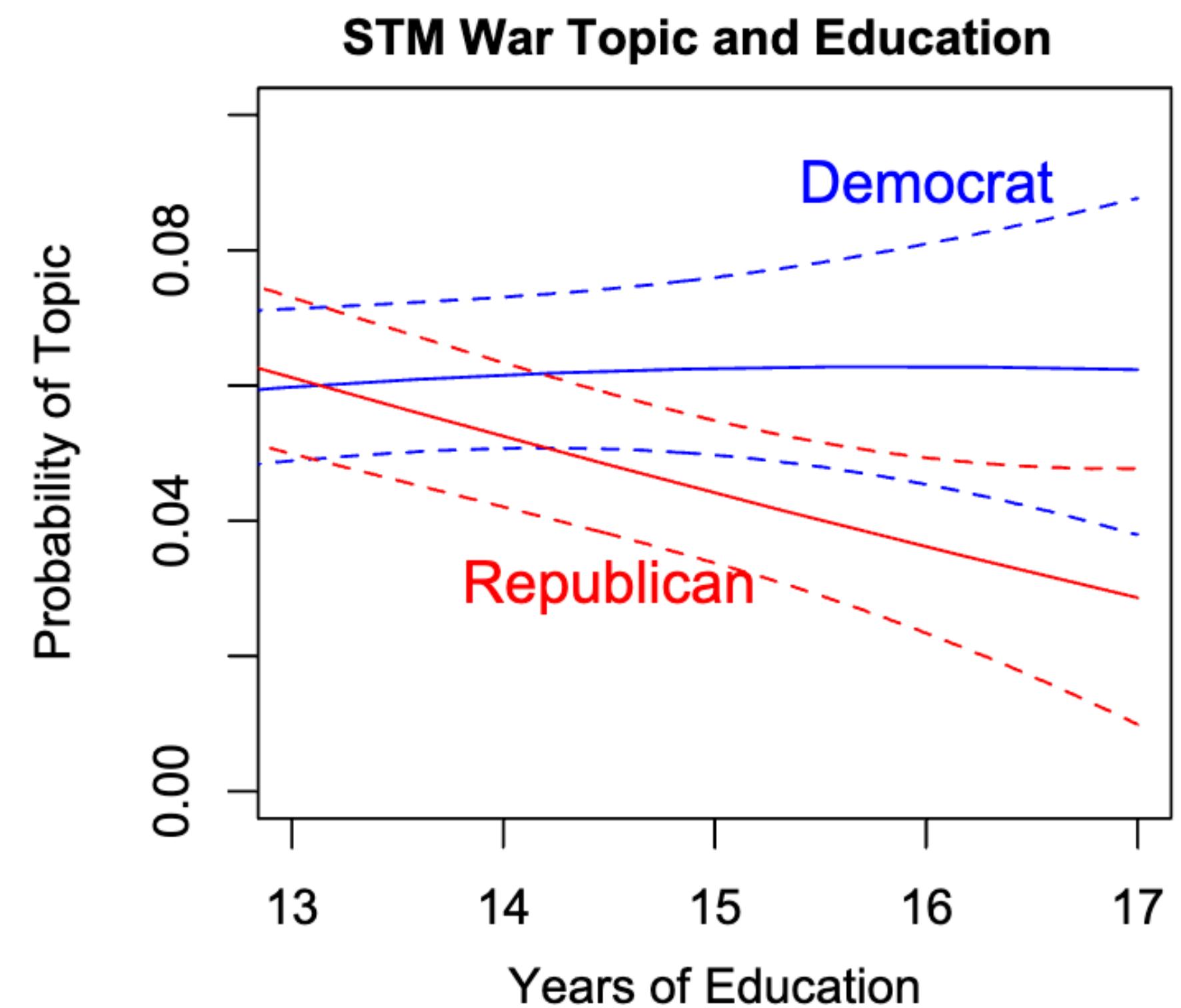
DYNAMIC TOPIC MODELS

- Text and time
- Time is continuous



STRUCTURAL TOPIC MODELS

- Model the text and all the associated metadata together
- Useful to do causal analysis



EVALUATION

EVALUATION

Human evaluation

Word Intrusion

1 / 10	floppy	alphabet	computer	processor	memory	disk
2 / 10	molecule	education	study	university	school	student
3 / 10	linguistics	actor	film	comedy	director	movie
4 / 10	islands	island	bird	coast	portuguese	mainland

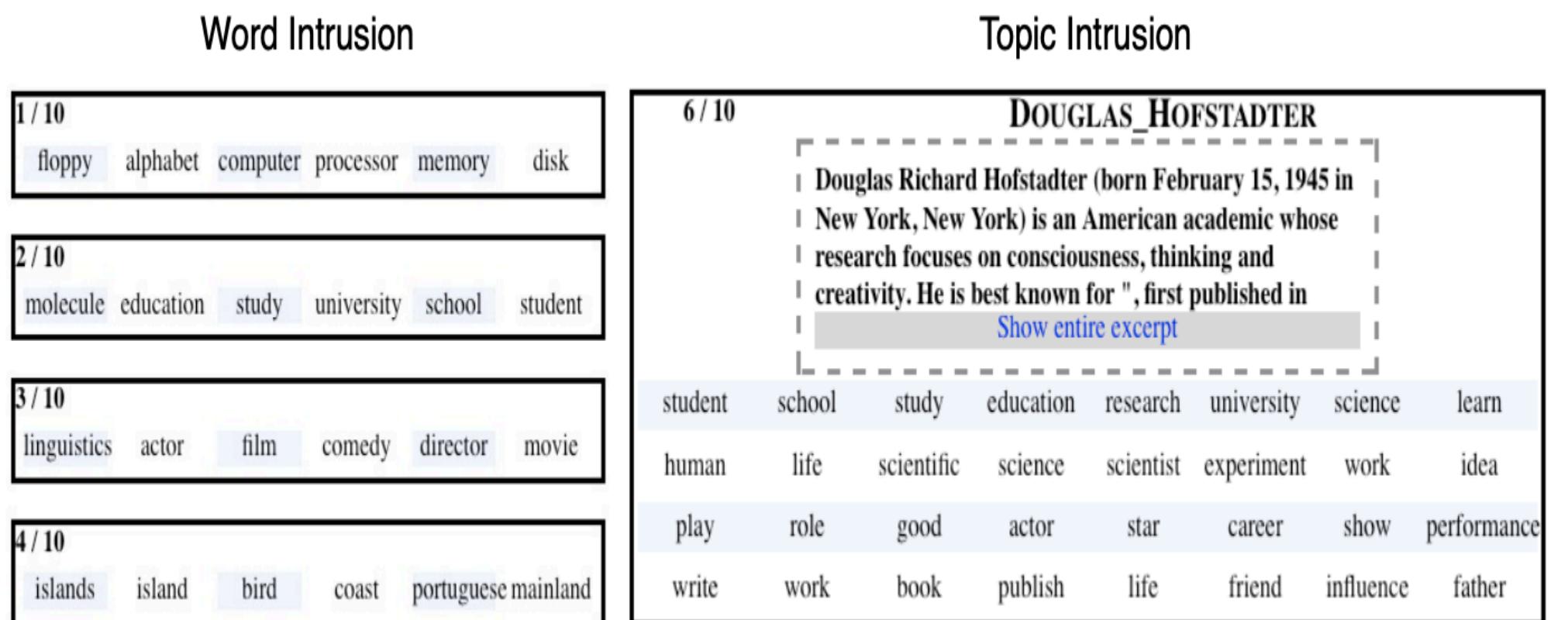
Topic Intrusion

6 / 10	DOUGLAS_HOFSTADTER															
Douglas Richard Hofstadter (born February 15, 1945 in New York, New York) is an American academic whose research focuses on consciousness, thinking and creativity. He is best known for ", first published in																
Show entire excerpt																
student school study education research university science learn																
human life scientific science scientist experiment work idea																
play role good actor star career show performance																
write work book publish life friend influence father																

Chang et. al. 2009

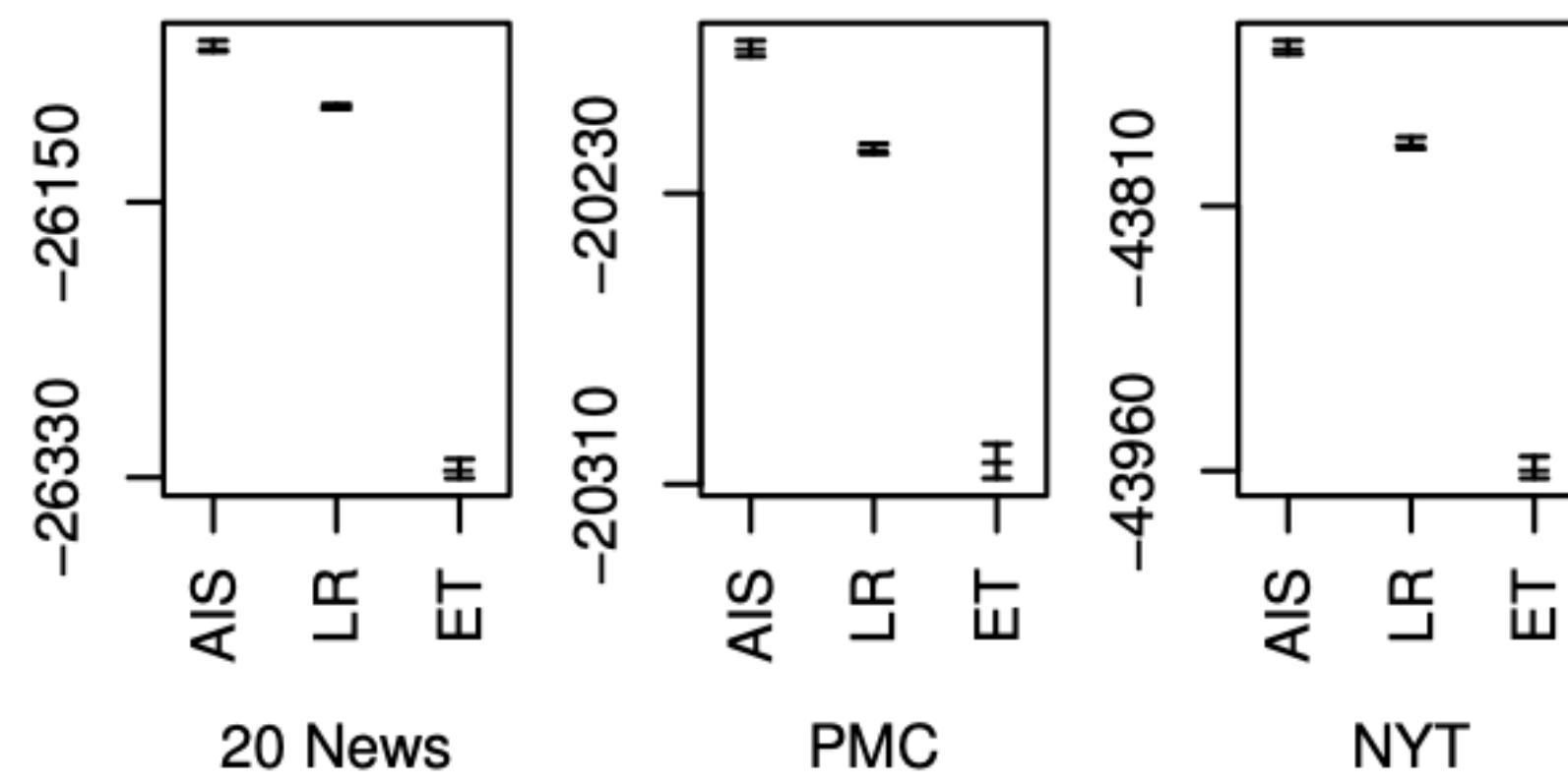
EVALUATION

Human evaluation



Chang et. al. 2009

Evaluating generation ability



Wallach et. al. 2009

IN CLASS

- Topic modeling demo