# Unit 3: Learning from other people

# 5. The structure in language

11/19/2020

### The structure in language

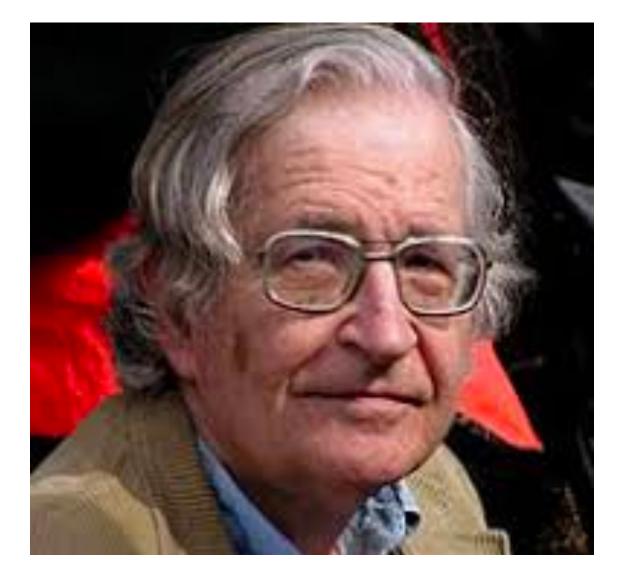
# 1. You can learn a lot form the co-occurrence structure of words in language

# 2. Latent semantic analysis and Topics models both use this structure to learn about the world

3. But some information is not (straightforwardly) in the co-occurrence structure of language

### How do you know so much without being told about it?





**Plato's Problem:** 

**Plato's Solution:** Knowledge is innate Plato (380 BC)

**Chomsky's Problem:** 

**Chomsky's Solution:** 

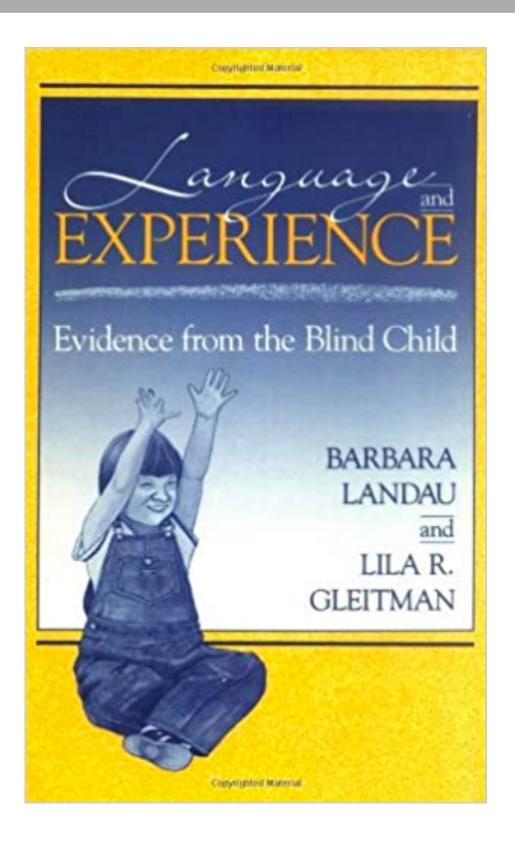
Chomsky (1986)

Even uneducated people seem to know a lot

Children seem to learn language from insufficient input

Universal grammar is innate

## Blind children know the meanings of sight words!





### Look up!



### Make it so mommy can't see the car



### Let me see the back of your pants



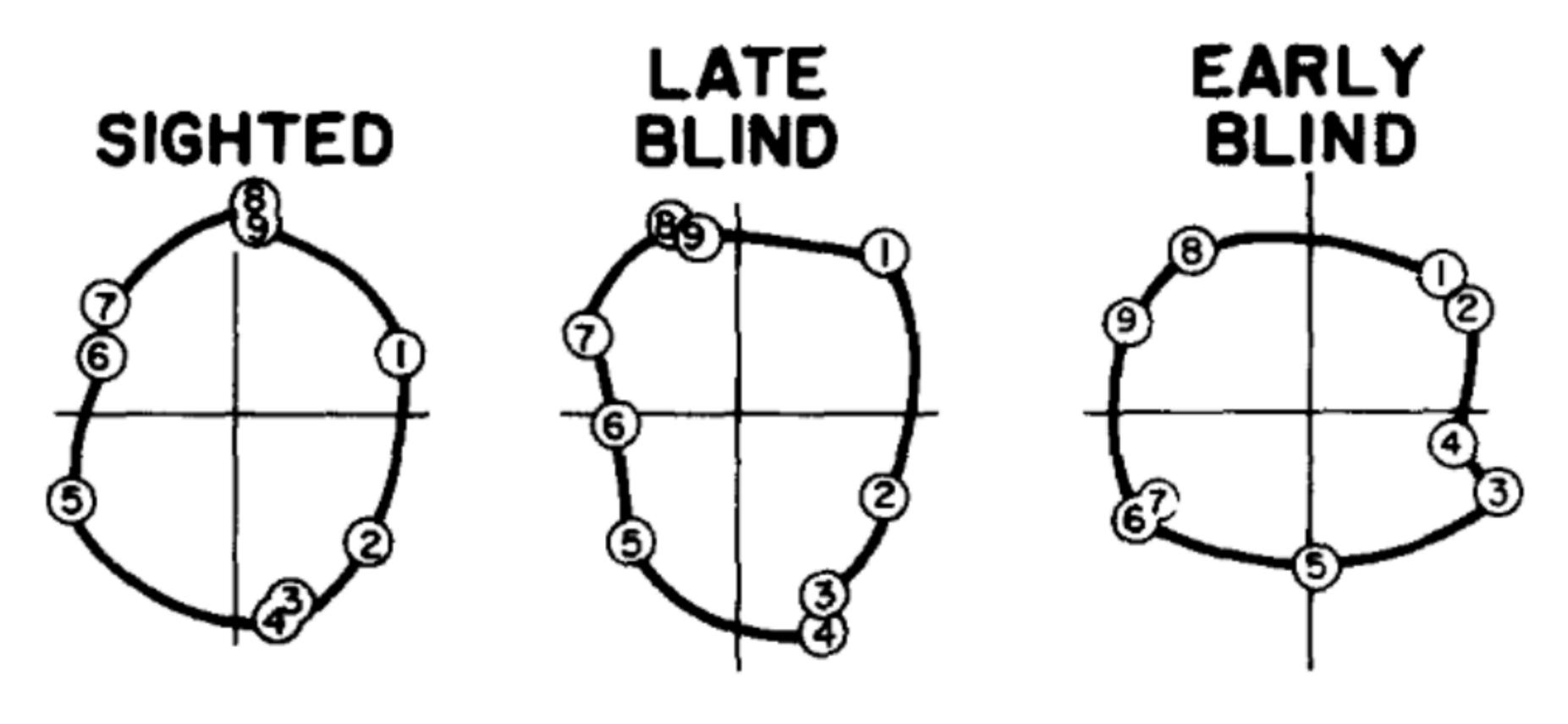


Let mommy see the car

Let mommy *touch* the car



### Blind adults color similarities look a lot like sighted adults



COLOR LEGEND:

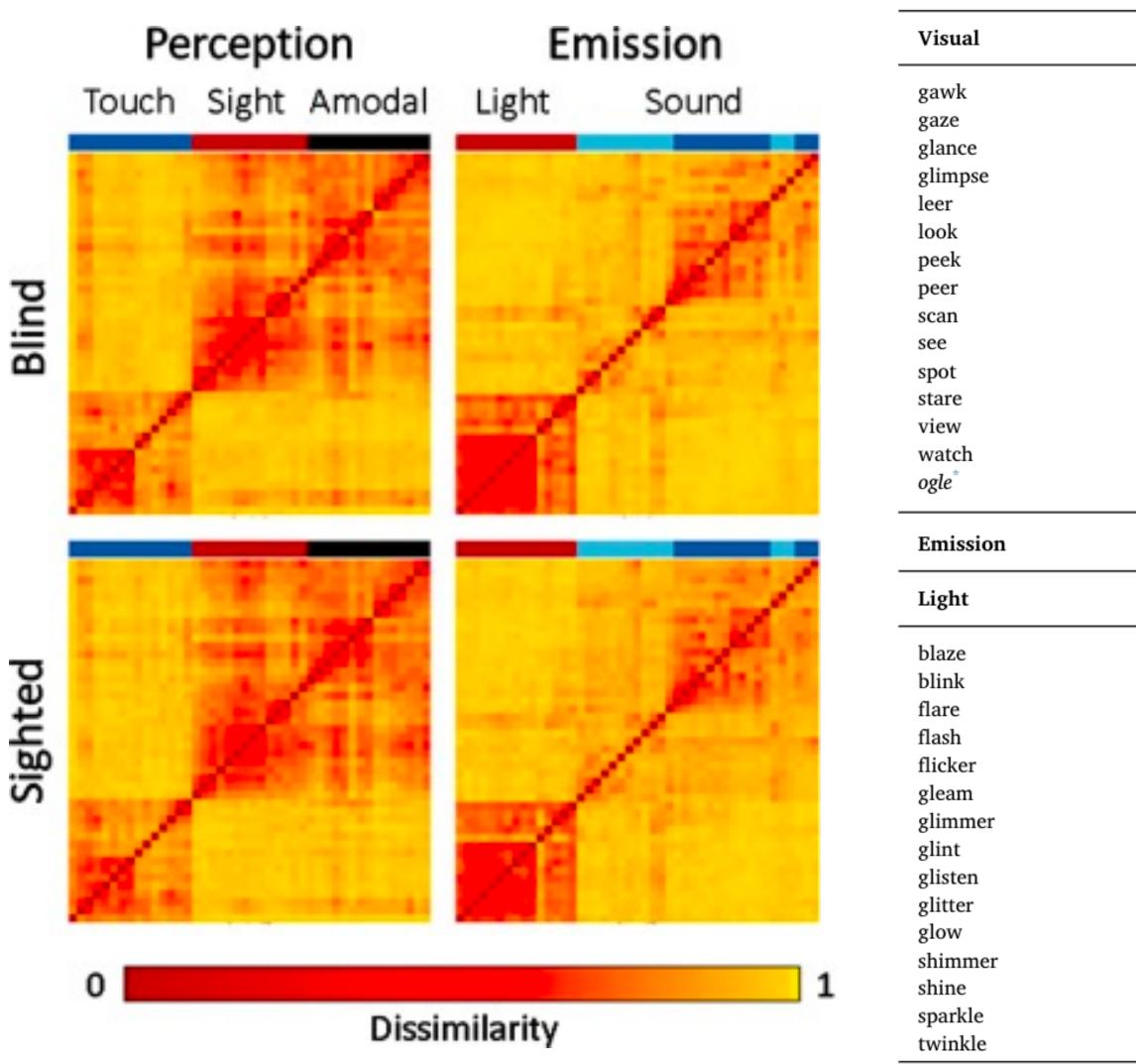
I. RED 2.ORANGE 3.GOLD 4. YELLOW 5.GREEN 6. TURQUISE 7. BLUE 8. PURPLE 9. VIOLET

Marmor (1978)



### Blind adults general perceptual verb similarities are like sighted adults'

. . . .



Touch	Amodal
caress dab feel grip nudge pat pet pinch prod pub scrape stroke tap	characterize classify discover examine identify investigate learn note notice perceive question recognize scrutinize
tickle touch	search study

Manner of Motion

Animate Sound	Inanimate Sound	_
bark	beep	bounce
bellow	boom	float
groan	buzz	glide
growl	chime	hobble
grumble	clang	roll
grunt	clank	saunter
howl	click	scurry
moan	crackle	skip
mutter	creak	slither
shout	crunch	spin
squawk	gurgle	strut
wail	hiss	trot
whimper	sizzle	twirl
whisper	squeak	twist
yelp	twang	waddle

### Bedny et al. (2019)



**Red** onions are sweeter than white ones

**Red** hair occurs naturally in one to two percent of the human population

Pittsburgh one of U.S. cities with highest number of gray days

Fall tips for a **green** spring lawn

Lake Tahoe stretches 22 miles long and 12 miles wide, with clear **blue** water that's more than 99 percent pure

# **Direct information:**

There is a relationship between e.g. red and hair

### **Indirect information:**

Red, white, gray, green, and blue are used in *similar contexts*.

Contexts for e.g. blue and green are more similar than blue and red



### Start with a term x document matrix

Term

	<b>d1</b>	d2	<b>d</b> 3	<b>d4</b>	d5	<b>d6</b>
rock	2	1	0	2	0	1
granite	1	0	1	0	0	0
marble	1	2	0	0	0	0
music	0	0	0	1	2	0
song	0	0	0	1	0	2
band	0	0	0	0	1	0

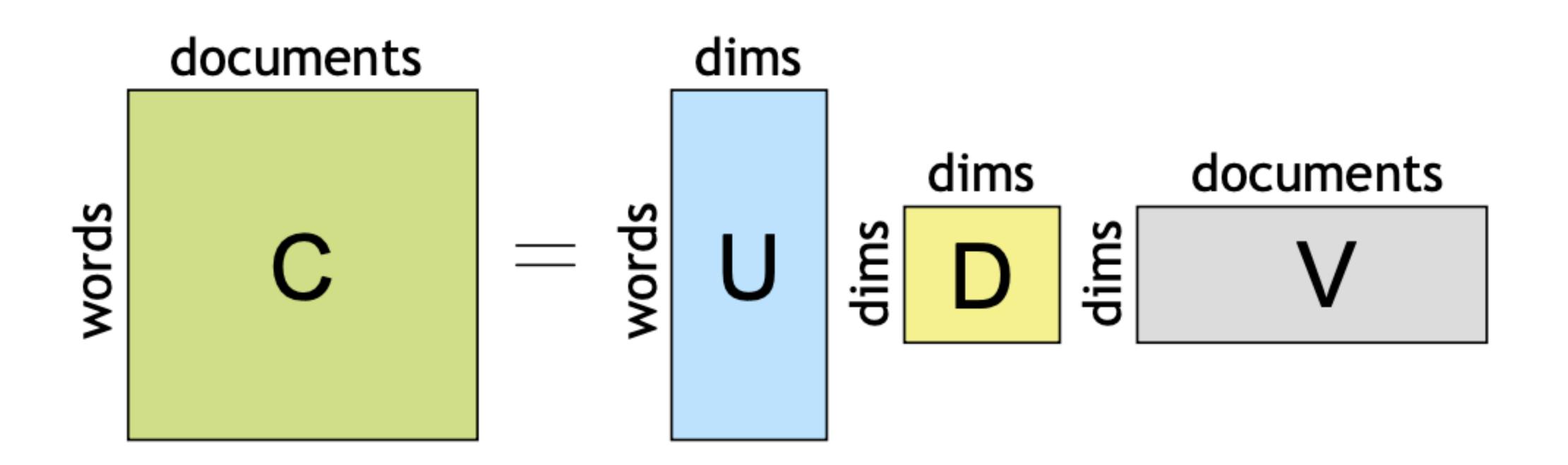
Words that occur in similar documents are probably similar

### Document

Adapted from Ho Tu Bao



## Latent semantic analysis: Finding the hidden structure in documents



Instead of representing words by co-occurrence in documents,

Semantics are a low-dimensional compression of documents

# we want to represent their co-occurrence in semantic dimensions

Adapted from Ho Tu Bao





### An example of LSA (Landauer, Foltz, & Laham, 1998)

### **Technical Memo Titles**

c1: *Human* machine *interface* for ABC *computer* applications c2: A survey of user opinion of computer system response time c3: The EPS user interface management system c4: System and human system engineering testing of EPS c5: Relation of *user* perceived *response time* to error measurement

m1: The generation of random, binary, ordered trees m2: The intersection graph of paths in trees m3: Graph minors IV: Widths of trees and well-quasi-ordering m4: *Graph minors*: A survey

### An example of LSA (Landauer, Foltz, & Laham, 1998)

	<b>c</b> 1	c 2	c 3	<b>c 4</b>	c 5	m 1	m2	<b>m3</b>	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

r(human, user) = -.38

r(human, minors) = -.29

### Decomposing the matrix

C

	<b>c</b> 1	c 2	c 3	c 4	c 5	m1	m2	<b>m3</b>	<b>m4</b>
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

D

U

$\begin{array}{c} 0.22\\ 0.20\\ 0.24\\ 0.40\\ 0.64\\ 0.27\\ 0.27\\ 0.27\\ 0.30\\ 0.21\\ 0.01\\ 0.04 \end{array}$	-0.11 -0.07 0.04 0.06 -0.17 0.11 0.11 -0.14 0.27 0.49 0.62	$\begin{array}{c} 0.29\\ 0.14\\ -0.16\\ -0.34\\ 0.36\\ -0.43\\ -0.43\\ 0.33\\ -0.18\\ 0.23\\ 0.22\\ \end{array}$	$\begin{array}{c} -0.41 \\ -0.55 \\ -0.59 \\ 0.10 \\ 0.33 \\ 0.07 \\ 0.07 \\ 0.19 \\ -0.03 \\ 0.03 \\ 0.00 \end{array}$	-0.11 0.28 -0.11 0.33 -0.16 0.08 0.08 0.11 -0.54 0.59 -0.07	-0.34 0.50 -0.25 0.38 -0.21 -0.17 -0.17 0.27 0.08 -0.39 0.11	$\begin{array}{c} 0.52 \\ -0.07 \\ -0.30 \\ 0.00 \\ -0.17 \\ 0.28 \\ 0.28 \\ 0.03 \\ -0.47 \\ -0.29 \\ 0.16 \end{array}$	-0.06 -0.01 0.06 0.00 0.03 -0.02 -0.02 -0.02 -0.04 0.25 -0.68	-0.41 -0.11 0.49 0.01 0.27 -0.05 -0.05 -0.05 -0.17 -0.58 -0.23 0.23	3.34	2.54	2.35	1.64	1.50
0.04 0.03	0.62 0.45	0.22 0.14	0.00 -0.01	-0.07 -0.30	0.11 0.28	0.16 0.34	-0.68 0.68	0.23 0.18					

				0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02
				-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62
				0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25
				-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01
1.31				0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15
1.51	0.85			-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00
		0.56	0.26	0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25
			0.36	-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45
				-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52

V

### A low-dimensional reconstruction using the first 2 dimensions

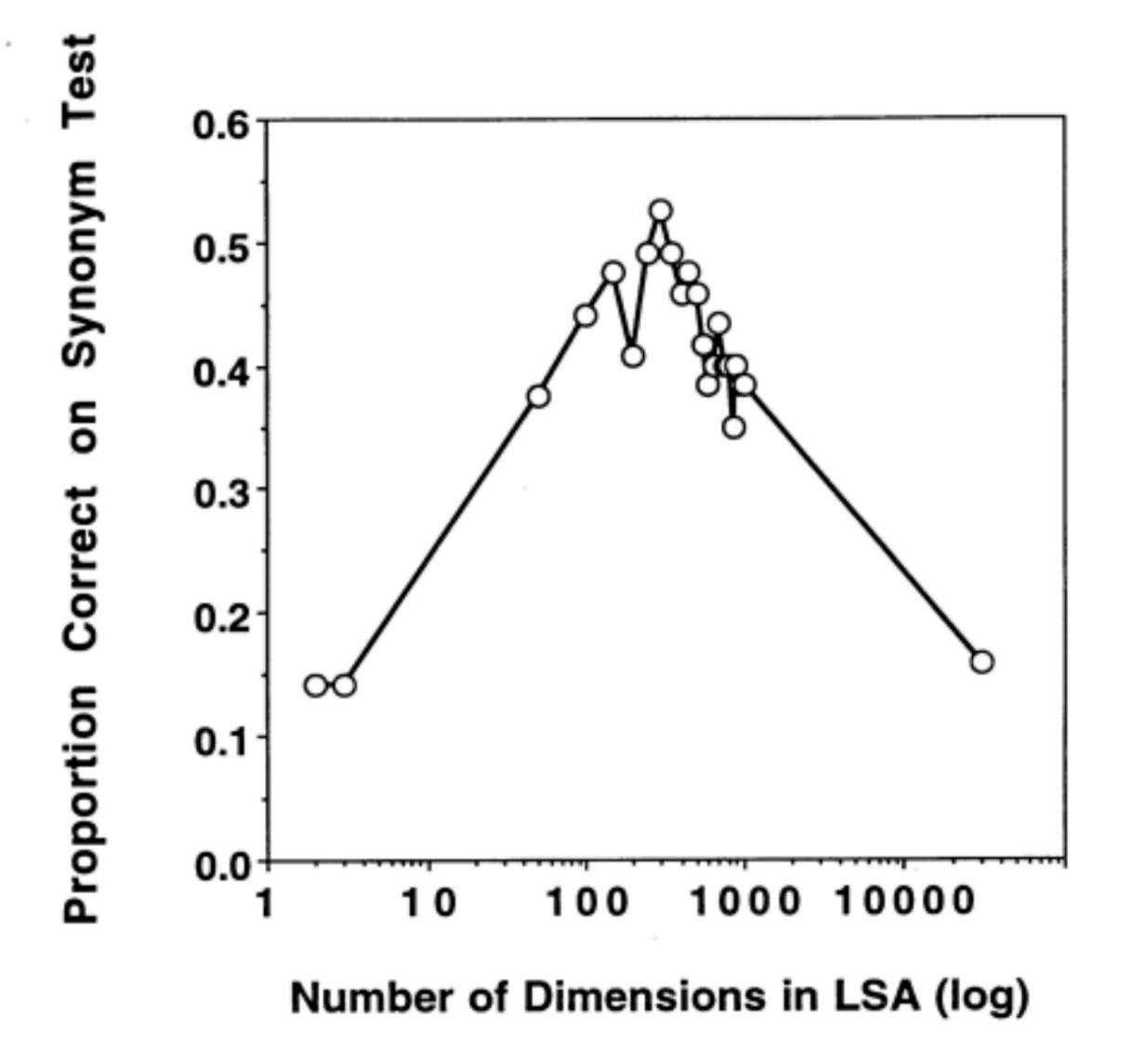
	c 1	c 2	c3	c 4	c 5	m 1	m2	m3	m4	, 1,,			
human	1	0	0	1	0	0	0	0	0				
interface	1	0	1	0	0	0	0	0	0			r(hu	man, user) = .
computer	1	1	0	0	0	0	0	0	0			i (i i u i	nan, user) – .
user	0	1	1	0	1	0	0	0	0				
system	0	1	I	2	0	0	0	0	0		rl	huma	n minarc) -
response	0	1	0	0	1	0	0	0	0			IIUIIId	n, minors) = -
time F D S	0 0	$1 \\ 0$	0	0		0	0	0 0	0				
EPS	0	1			0	0	0	0	0				
survey trees	0	$\stackrel{1}{0}$	0	0	0	1	1	1	0				
graph	0 0	0 0	õ	Ŏ	ŏ	$\hat{0}$	1	1	1				
minors	0	0	0	0	0	0	0	1	1				
	<b>c</b> 1	(	c2	c3		c4		c5	m1	m2	m3	m4	
human	0.16	5	0.40	0	.38	0.4	7	0.18	-0.0	5 -0.12	-0.16	-0.09	
interface	0.14	1	0.37	0	.33	0.4	0	0.16	-0.0	3 -0.07	-0.10	-0.04	
computer	0.15	5	0.51	0	.36	0.4	1	0.24	0.0	2 0.06	0.09	0.12	
user	0.26		0.84		.61	0.7		0.39	0.0		0.12	0.19	
system	0.45		1.23		.05	1.2		0.56			-0.21	-0.05	
response	0.16		0.58		.38	0.4		0.28	0.0		0.19	0.22	
time	0.16		0.58		.38	0.4		0.28	0.0		0.19	0.22	
EPS	0.22		0.55		.51	0.6		0.24	-0.0			-0.11	
survey	0.10		0.53		.23	0.2		0.27	0.1		0.44	0.42	
trees	-0.06 -0.06		0.23 0.34		.14 .15	-0.2 -0.3		0.14 0.20	0.2 0.3		0.77 0.98	0.66 0.85	
graph	-0.04		0.34		.15	-0.3		0.20	0.3		0.98	0.85	7
minors	-0.04	т	0.23	-0	.10	-0.2	, <b>1</b>	0.15	0.2	2 0.50	0.71	0.02	

C

**('** 



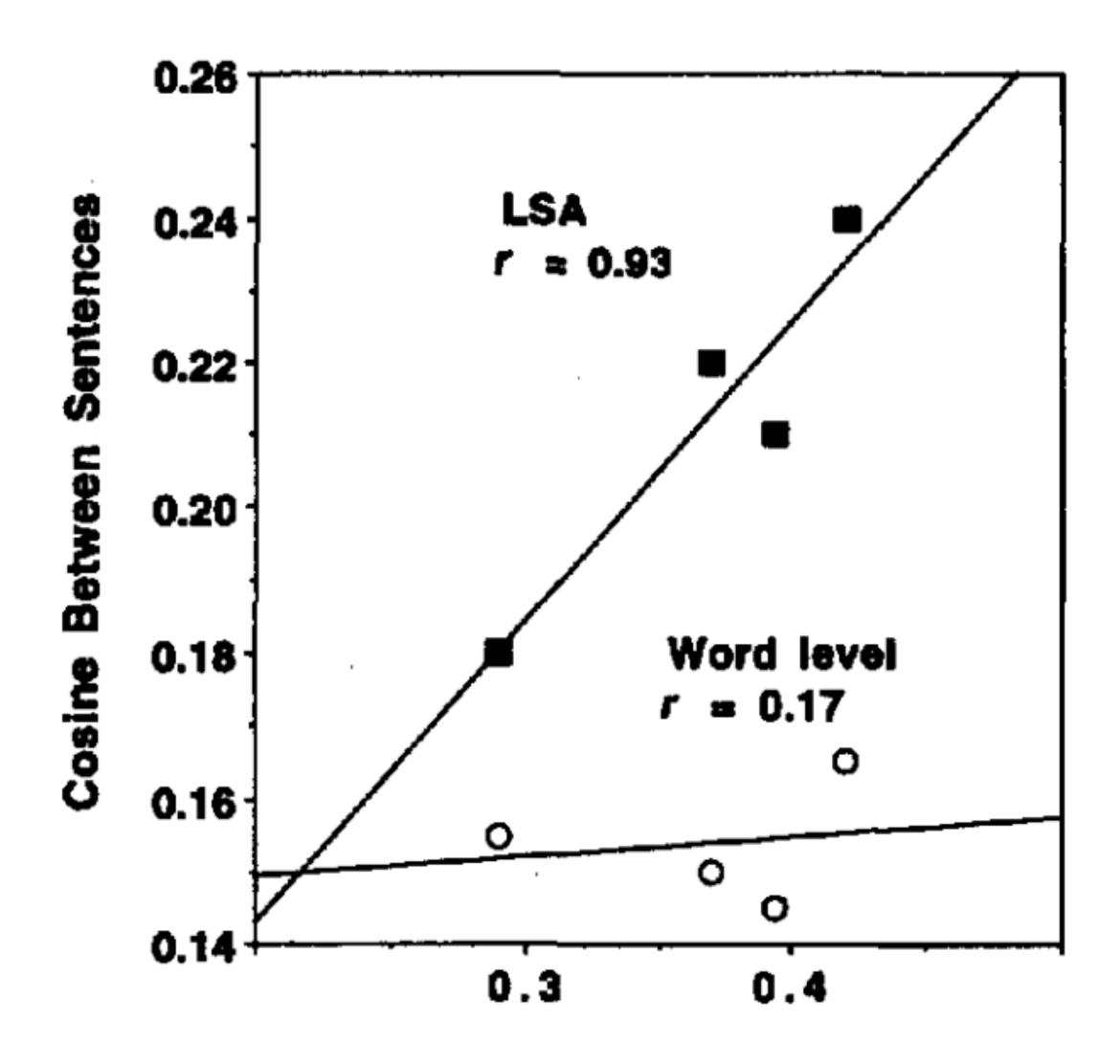
## Latent semantic analysis predicts human similarity judgments



# LSA can pass the synonym part of the TOEFL!

# Chance is 25% Applicants to US Universities on average get 52.7%

### Latent semantic analysis predicts sentence comprehensibility



Comprehension (%)

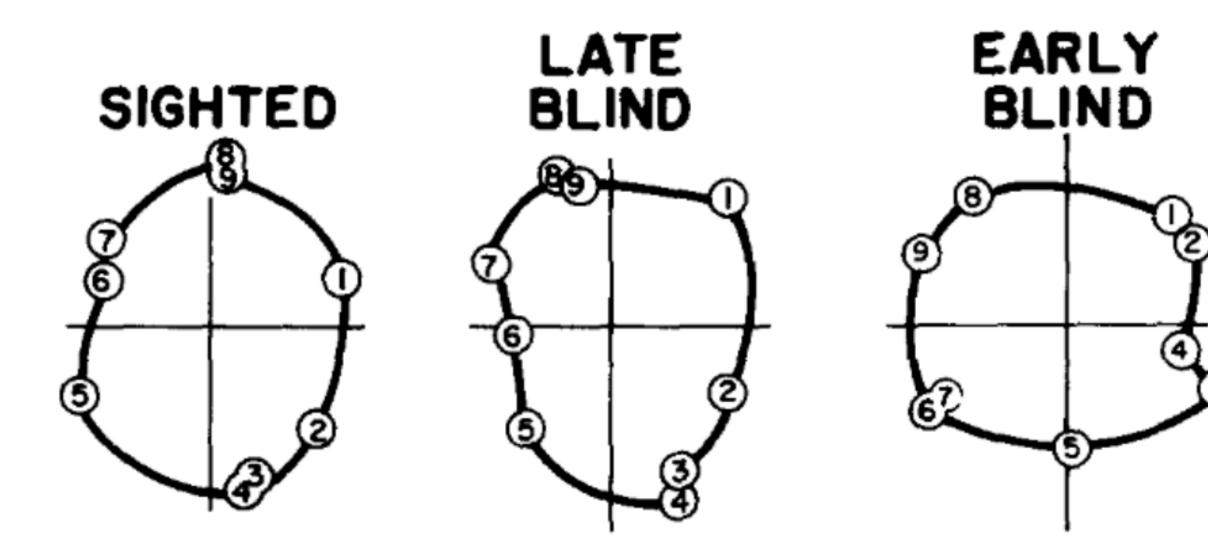
Mammals have very specialized teeth. There are four types of teeth in mammals: incisors, canines, premolars, and molars. The number and shape of each of these types of teeth are related to the kind of food the mammal eats.

Another physical trait of mammals is that they can eat many different kinds of food because they have very specialized teeth. This trait also helps them to live in different kinds of environments





### **Can LSA recover perceptual similarities?**



COLOR LEGEND:

I. RED 2.ORANGE 3.GOLD 4. YELLOW 5.GREEN 6. TURQUISE 7. BLUE 8. PURPLE 9. VIOLET

Visual	Touch	Amodal
gawk	caress	characterize
gaze	dab	classify
glance	feel	discover
glimpse	grip	examine
leer	nudge	identify
look	pat	investigate
peek	pet	learn
peer	pinch	note
scan	prod	notice
see	pub	perceive
spot	scrape	question
stare	stroke	recognize
view	tap	scrutinize
watch	tickle	search
ogle <sup>*</sup>	touch	study

### Emission

Manner of Motion

Light	Animate Sound	Inanimate Sound	
blaze	bark	beep	bounce
blink	bellow	boom	float
flare	groan	buzz	glide
flash	growl	chime	hobble
flicker	grumble	clang	roll
gleam	grunt	clank	saunter
glimmer	howl	click	scurry
glint	moan	crackle	skip
glisten	mutter	creak	slither
glitter	shout	crunch	spin
glow	squawk	gurgle	strut
shimmer	wail	hiss	trot
shine	whimper	sizzle	twirl
sparkle	whisper	squeak	twist
twinkle	yelp	twang	waddle

- - - - - -

## https://bit.ly/test-lsa

### But what are these underlying dimensions?

	c 1	c 2	c3	c 4	c 5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

D

U

$\begin{array}{c} 0.22 \\ 0.20 \\ 0.24 \\ 0.40 \\ 0.64 \\ 0.27 \\ 0.27 \\ 0.30 \\ 0.21 \\ 0.01 \\ 0.04 \end{array}$	-0.11 -0.07 0.04 0.06 -0.17 0.11 0.11 -0.14 0.27 0.49	$\begin{array}{c} 0.29\\ 0.14\\ -0.16\\ -0.34\\ 0.36\\ -0.43\\ -0.43\\ 0.33\\ -0.18\\ 0.23\\ 0.23\\ \end{array}$	-0.41 -0.55 -0.59 0.10 0.33 0.07 0.07 0.07 0.19 -0.03 0.03	-0.11 0.28 -0.11 0.33 -0.16 0.08 0.08 0.11 -0.54 0.59	-0.34 0.50 -0.25 0.38 -0.21 -0.17 -0.17 0.27 0.08 -0.39	$\begin{array}{c} 0.52 \\ -0.07 \\ -0.30 \\ 0.00 \\ -0.17 \\ 0.28 \\ 0.28 \\ 0.03 \\ -0.47 \\ -0.29 \\ 0.16 \end{array}$	$\begin{array}{c} -0.06\\ -0.01\\ 0.06\\ 0.00\\ 0.03\\ -0.02\\ -0.02\\ -0.02\\ -0.02\\ -0.04\\ 0.25\\ 0.68\end{array}$	-0.41 -0.11 0.49 0.01 0.27 -0.05 -0.05 -0.17 -0.58 -0.23	3.34	2.54	2.35	1.64	1.50
0.04	0.49 0.62	0.23 0.22	0.03 0.00	0.59 -0.07	-0.39 0.11	-0.29 0.16	0.25 -0.68	-0.23 0.23					
0.03	0.45	0.14	-0.01	-0.30	0.28	0.34	0.68	0.18					

				0.20	0.61	0.46	0.54	0.28	0.00	0.01	0.02
				-0.06	0.17	-0.13	-0.23	0.11	0.19	0.44	0.62
				0.11	-0.50	0.21	0.57	-0.51	0.10	0.19	0.25
				-0.95	-0.03	0.04	0.27	0.15	0.02	0.02	0.01
1.31				0.05	-0.21	0.38	-0.21	0.33	0.39	0.35	0.15
1.51	0.85			-0.08	-0.26	0.72	-0.37	0.03	-0.30	-0.21	0.00
		0.56	0.26	0.18	-0.43	-0.24	0.26	0.67	-0.34	-0.15	0.25
			0.36	-0.01	0.05	0.01	-0.02	-0.06	0.45	-0.76	0.45
				-0.06	0.24	0.02	-0.08	-0.26	-0.62	0.02	0.52

### Topic Models (Latent Dirichlet Allocation) - Steyvers & Griffiths (2007)

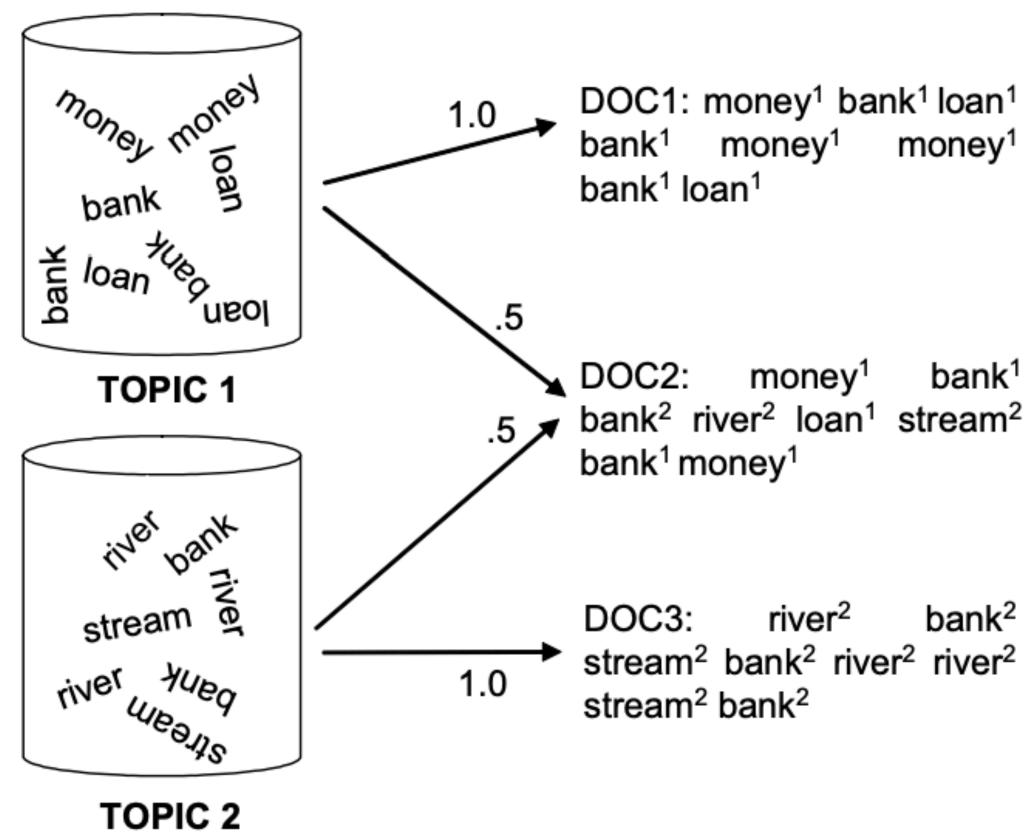
### **Idea:** A generative model for how words appear in documents

Each document is created by picking a set of topics, each with some weight.

For every word in the document, it is chosen from one of the topics according to their weight

Model from Blei, Ng, & Jordan (2003)

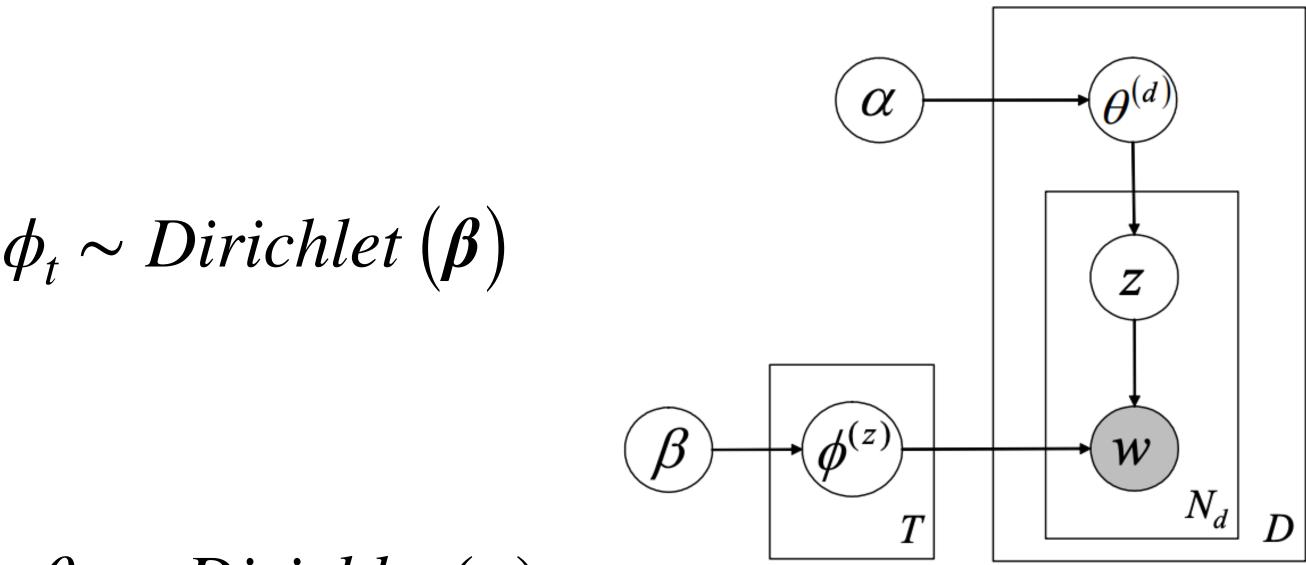
### PROBABILISTIC GENERATIVE PROCESS





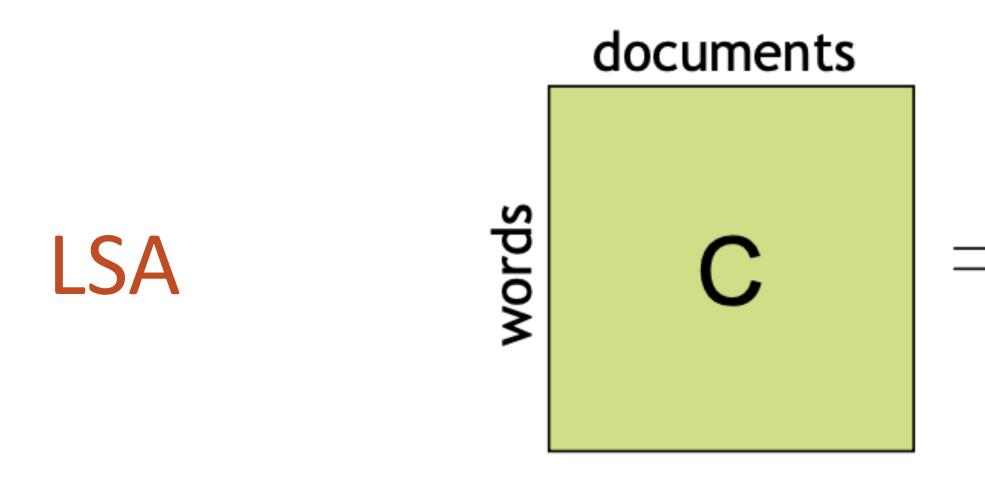
- For each **topic** 1...*t*:
- Draw a multinomial over words  $\phi_t \sim Dirichlet(\beta)$
- For each **document** 1...*d* :
- Draw a multinomial over topics  $\theta_d \sim Dirichlet(\alpha)$
- For each word *W*<sub>*d*,*n*</sub>:
  - Draw a topic  $Z_{d,n} \sim \text{Multinomial}(\theta_d)$  with  $Z_{d,n} \in [1...t]$
  - Draw a word  $W_{d,n} \sim \text{Multinomial} \left( \beta_{Z_d,n} \right)$

### **Topic Models (Latent Dirichlet Allocation) - Steyvers & Griffiths (2007)**

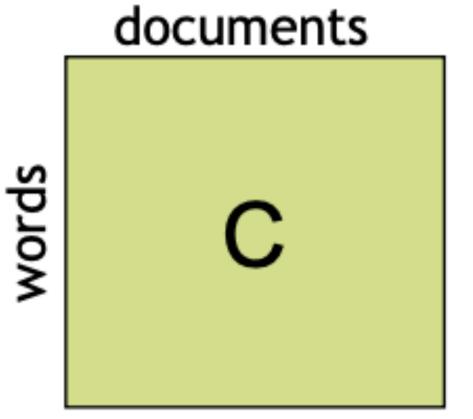




### **Comparing LSA and Topic Models**

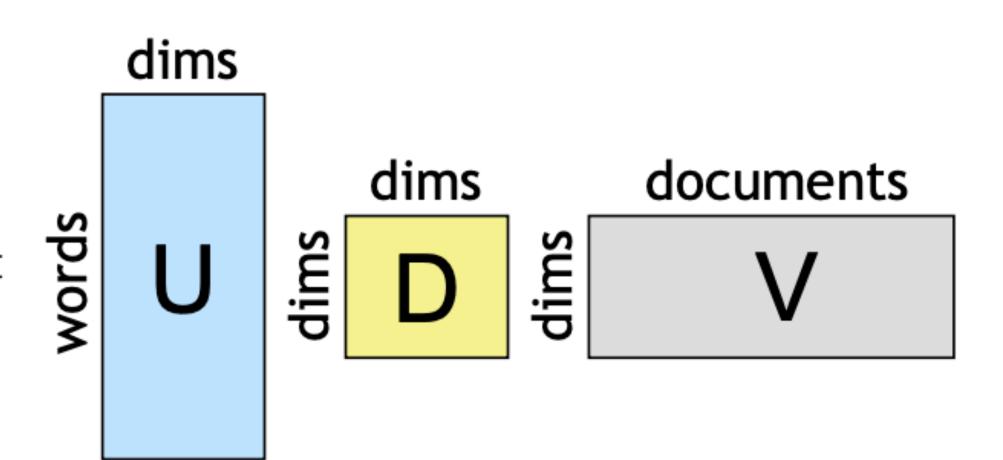


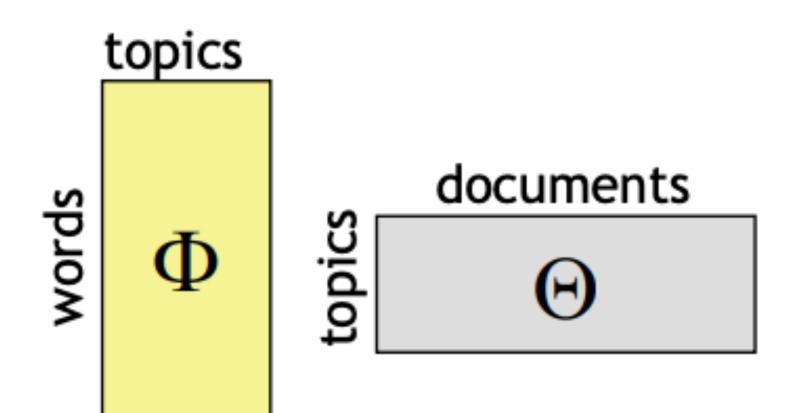
# **Topic Model**



Normalized cooccurrence matrix

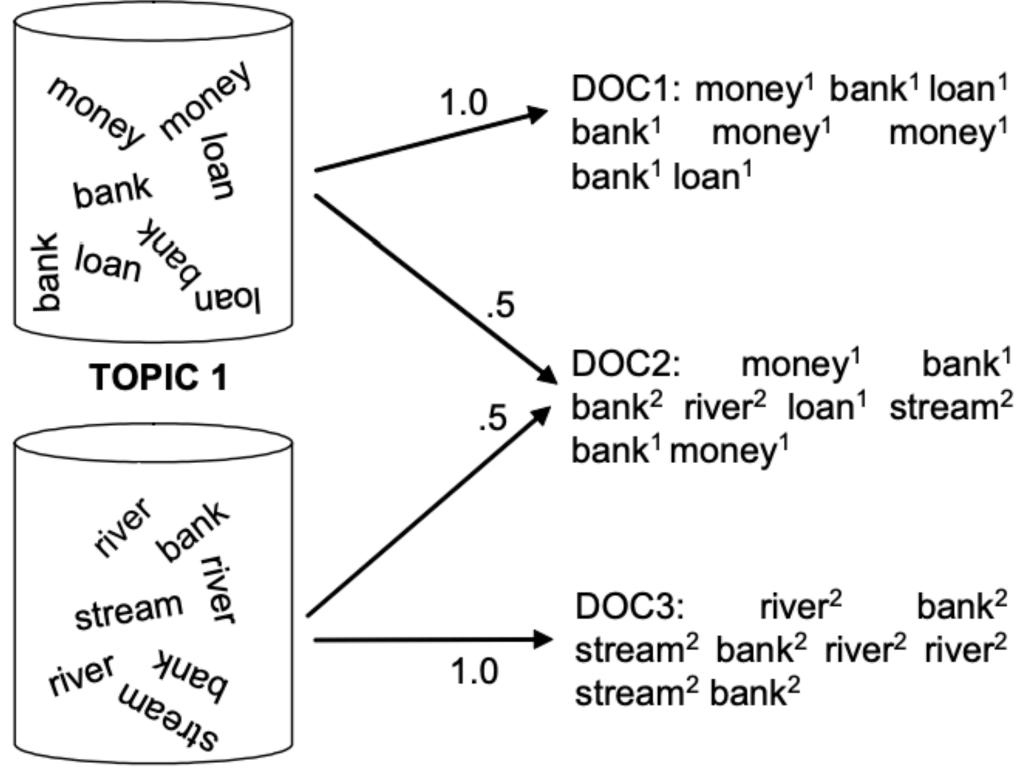






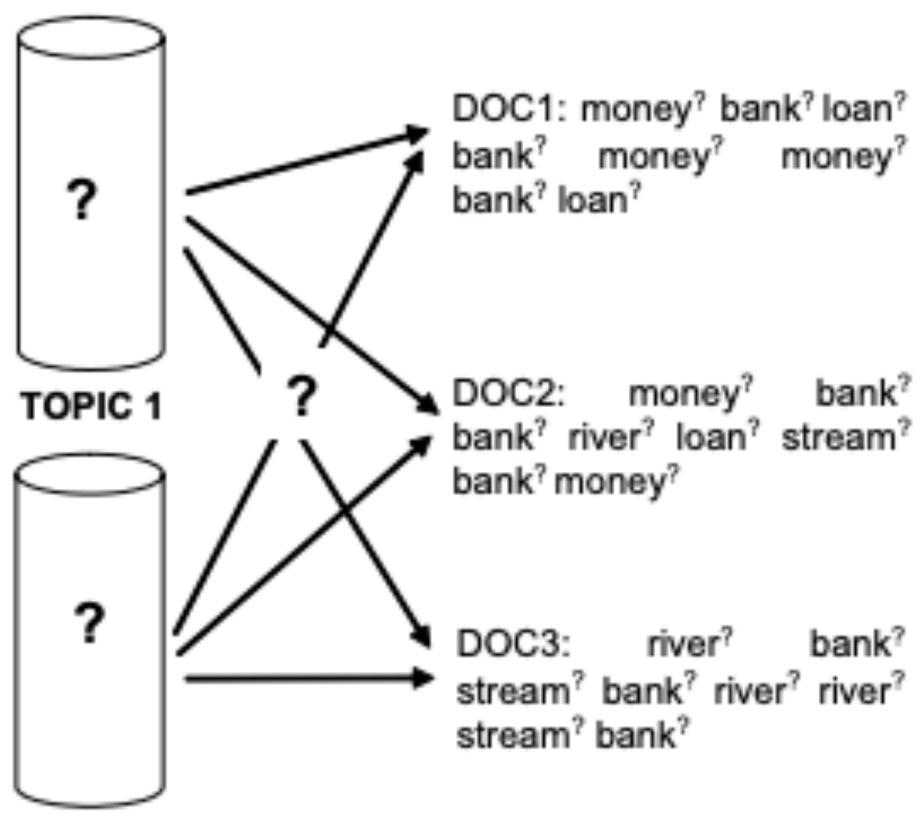
### Inferring topics by Bayesian inference using sampling

### PROBABILISTIC GENERATIVE PROCESS



**TOPIC 2** 

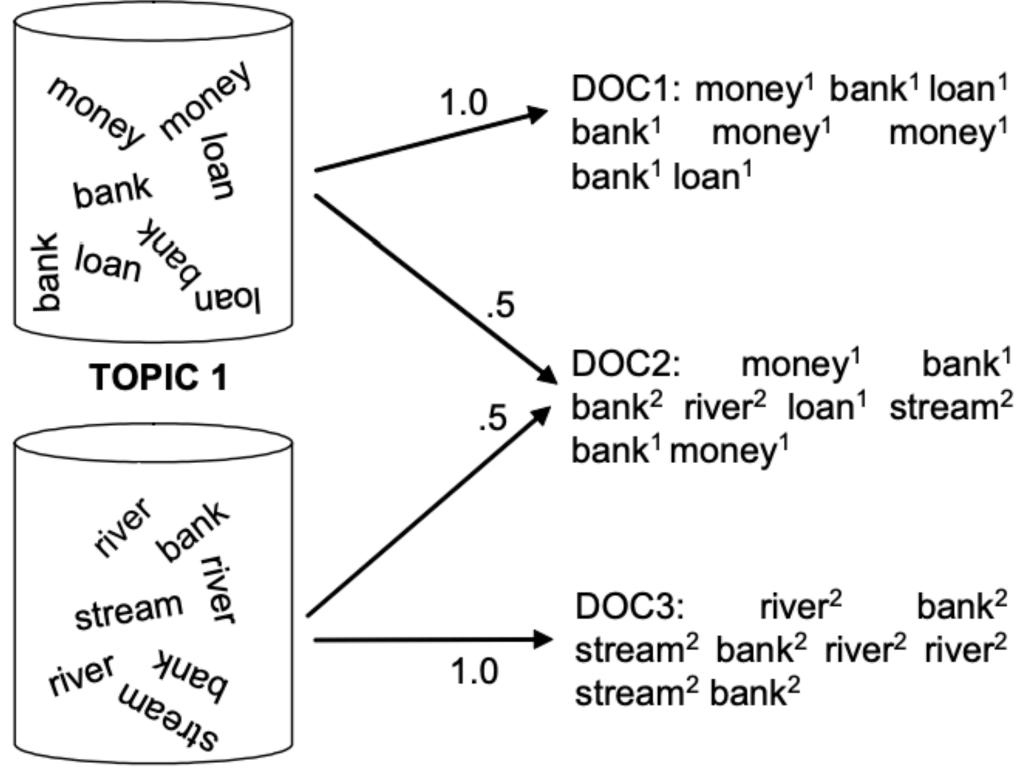
### STATISTICAL INFERENCE



TOPIC 2

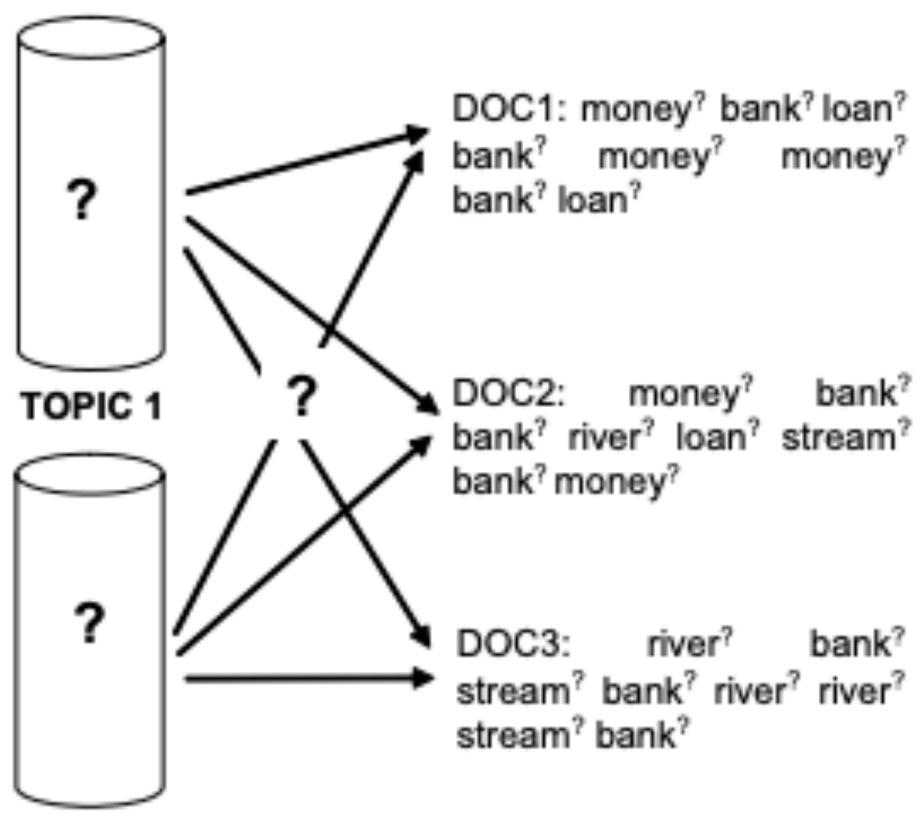
### Inferring topics by Bayesian inference using sampling

### PROBABILISTIC GENERATIVE PROCESS



**TOPIC 2** 

### STATISTICAL INFERENCE



TOPIC 2

### Example topics form the Touchstone Applied Sciences Association (TASA) corpus

prob.

.202

.099

.096

.073

.048

.048

.030

.029

.027

.027

.017

.017

.016

.015

.011

.009

### Topic 247

### Topic 5

prob.	word
.069	RED
.060	BLUE
.027	GREEN
.026	YELLOW
.023	WHITE
.019	COLOR
.016	BRIGHT
.016	COLORS
.014	ORANGE
.012	BROWN
.012	PINK
.011	LOOK
.009	BLACK
.009	PURPLE
.008	CROSS
.008	COLORED

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

### Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011



### Topics models can resolve polysemy

Bix beiderbecke, at age<sup>060</sup> fifteen<sup>207</sup>, sat<sup>174</sup> on the slope<sup>071</sup> of a bluff<sup>055</sup> overlooking<sup>027</sup> the mississippi<sup>137</sup> river<sup>137</sup>. He was listening<sup>077</sup> to music<sup>077</sup> coming<sup>009</sup> from a passing<sup>043</sup> riverboat. The music<sup>077</sup> had already captured<sup>006</sup> his heart<sup>157</sup> as well as his ear<sup>119</sup>. It was jazz<sup>077</sup>. Bix beiderbecke had already had music<sup>077</sup> lessons<sup>077</sup>. He showed<sup>002</sup> promise<sup>134</sup> on the piano<sup>077</sup>, and his parents<sup>035</sup> hoped<sup>268</sup> he might consider<sup>118</sup> becoming a concert<sup>077</sup> pianist<sup>077</sup>. But bix was interested<sup>268</sup> in another kind<sup>050</sup> of music<sup>077</sup>. He wanted<sup>268</sup> to play<sup>077</sup> the cornet. And he wanted<sup>268</sup> to play<sup>077</sup> jazz<sup>077</sup>...

word prob. LITER/ MUSIC .090 DANCE .034 SONG .033 PO .030 PLAY SING .026 SINGING .026 BAND .026 PLAYED LITE .023 WR SANG .022 SONGS .021 D DANCING .020 PIANO .017 PLAYING .016 RHYTHM .015 SHAKES WR ALBERT .013 MUSICAL .013

Topic 77

word	prob.	word	prob.
ATURE	.031	PLAY	.136
POEM	.028	BALL	.129
OETRY	.027	GAME	.065
POET	.020	PLAYING	.042
PLAYS	.019	HIT	.032
POEMS	.019	PLAYED	.031
PLAY	.015	BASEBALL	.027
ERARY	.013	GAMES	.025
RITERS	.013	BAT	.019
DRAMA	.012	RUN	.019
WROTE	.012	THROW	.016
POETS	.011	BALLS	.015
VRITER	.011	TENNIS	.011
SPEARE	.010	HOME	.010
RITTEN	.009	CATCH	.010
STAGE	.009	FIELD	.010

Topic 166

## Topics models can resolve polysemy

There is a simple<sup>050</sup> reason<sup>106</sup> why there are so few periods<sup>078</sup> of really great theater<sup>082</sup> in our whole western<sup>046</sup> world. Too many things<sup>300</sup> have to come right at the very same time. The dramatists must have the right actors<sup>082</sup>, the actors<sup>082</sup> must have the right playhouses, the playhouses must have the right audiences<sup>082</sup>. We must remember<sup>288</sup> that plays<sup>082</sup> exist<sup>143</sup> to be performed<sup>077</sup>, not merely<sup>050</sup> to be read<sup>254</sup>. (even when you read<sup>254</sup> a play<sup>082</sup> to yourself, try<sup>288</sup> to perform<sup>062</sup> it, to put<sup>174</sup> it on a stage<sup>078</sup>, as you go along.) as soon<sup>028</sup> as a play<sup>082</sup> has to be performed<sup>082</sup>, then some kind<sup>126</sup> of theatrical<sup>082</sup>...

ropie //		100002		10/10 100
word	prob.	word	prob.	word prob.
MUSIC	.090	LITERATURE	.031	PLAY .136
DANCE	.034	POEM	.028	BALL .129
SONG	.033	POETRY	.027	GAME .065
PLAY	.030	POET	.020	PLAYING .042
SING	.026	PLAYS	.019	HIT .032
SINGING	.026	POEMS	.019	PLAYED .031
BAND	.026	PLAY	.015	BASEBALL .027
PLAYED	.023	LITERARY	.013	GAMES .025
SANG	.022	WRITERS	.013	BAT .019
SONGS	.021	DRAMA	.012	RUN .019
DANCING	.020	WROTE	.012	THROW .016
PIANO	.017	POETS	.011	BALLS .015
PLAYING	.016	WRITER	.011	TENNIS .011
RHYTHM	.015	SHAKESPEARE	.010	HOME .010
ALBERT	.013	WRITTEN	.009	CATCH .010
MUSICAL	.013	STAGE	.009	FIELD .010

Topic 77

To	pic	16	6
			-

### Topics models can resolve polysemy

Jim<sup>296</sup> has a game<sup>166</sup> book<sup>254</sup>. Jim<sup>296</sup> reads<sup>254</sup> the book<sup>254</sup>. Jim<sup>296</sup> sees<sup>081</sup> a game<sup>166</sup> for one. Jim<sup>296</sup> plays<sup>166</sup> the game<sup>166</sup>. Jim<sup>296</sup> likes<sup>081</sup> the game<sup>166</sup> for one. The game<sup>166</sup> book<sup>254</sup> helps<sup>081</sup> jim<sup>296</sup>. Don<sup>180</sup> comes<sup>040</sup> into the house<sup>038</sup>. Don<sup>180</sup> and jim<sup>296</sup> read<sup>254</sup> the game<sup>166</sup> book<sup>254</sup>. The boys<sup>020</sup> see a game<sup>166</sup> for two. The two boys<sup>020</sup> play<sup>166</sup> the game<sup>166</sup>. The boys<sup>020</sup> play<sup>166</sup> the game<sup>166</sup> for two. The boys<sup>020</sup> like the game<sup>166</sup>. Meg<sup>282</sup> comes<sup>040</sup> into the house<sup>282</sup>. Meg<sup>282</sup> and don<sup>180</sup> and jim<sup>296</sup> read<sup>254</sup> the book<sup>254</sup>. They see a game<sup>166</sup> for three. Meg<sup>282</sup> and don<sup>180</sup> and jim<sup>296</sup> play<sup>166</sup> the game<sup>166</sup>. They play<sup>166</sup>...

Topic 77

Topic 82

word	prob.	word	prob.		word	prob.
MUSIC	.090	LITERATURE	.031	1	PLAY	.136
DANCE	.034	POEM	.028		BALL	.129
SONG	.033	POETRY	.027		GAME	.065
PLAY	.030	POET	.020		PLAYING	.042
SING	.026	PLAYS	.019		HIT	.032
SINGING	.026	POEMS	.019		PLAYED	.031
BAND	.026	PLAY	.015		BASEBALL	.027
PLAYED	.023	LITERARY	.013		GAMES	.025
SANG	.022	WRITERS	.013		BAT	.019
SONGS	.021	DRAMA	.012		RUN	.019
DANCING	.020	WROTE	.012		THROW	.016
PIANO	.017	POETS	.011		BALLS	.015
PLAYING	.016	WRITER	.011		TENNIS	.011
RHYTHM	.015	SHAKESPEARE	.010		HOME	.010
ALBERT	.013	WRITTEN	.009		CATCH	.010
MUSICAL	.013	STAGE	.009		FIELD	.010

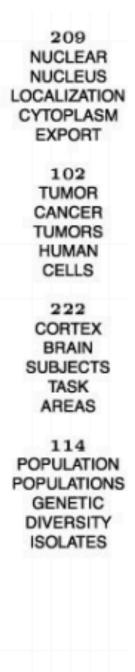
### Training topic models on twitter

https://ermoore.shinyapps.io/twittr/

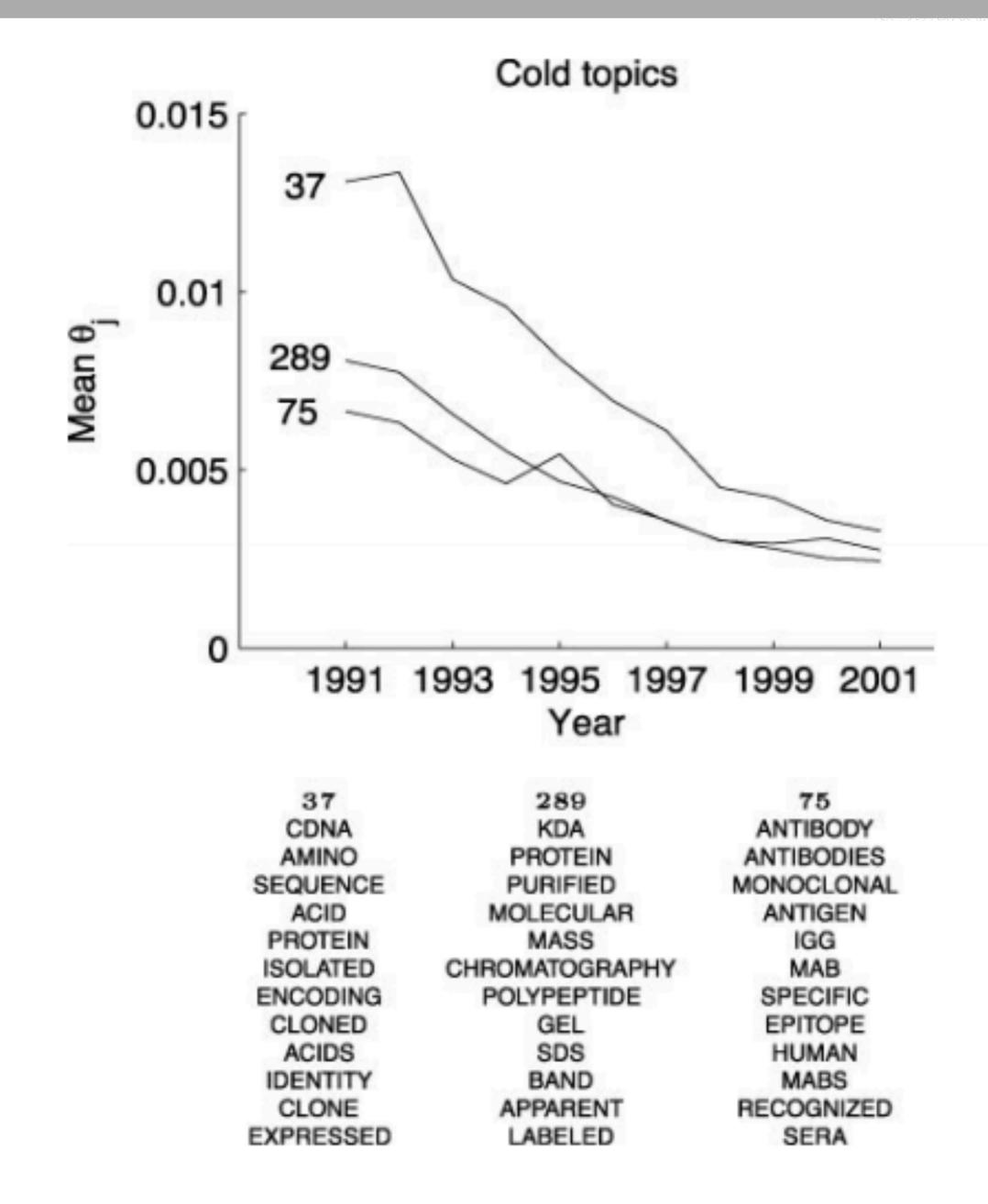
### Finding scientific topics (Griffiths & Steyvers, 2004)

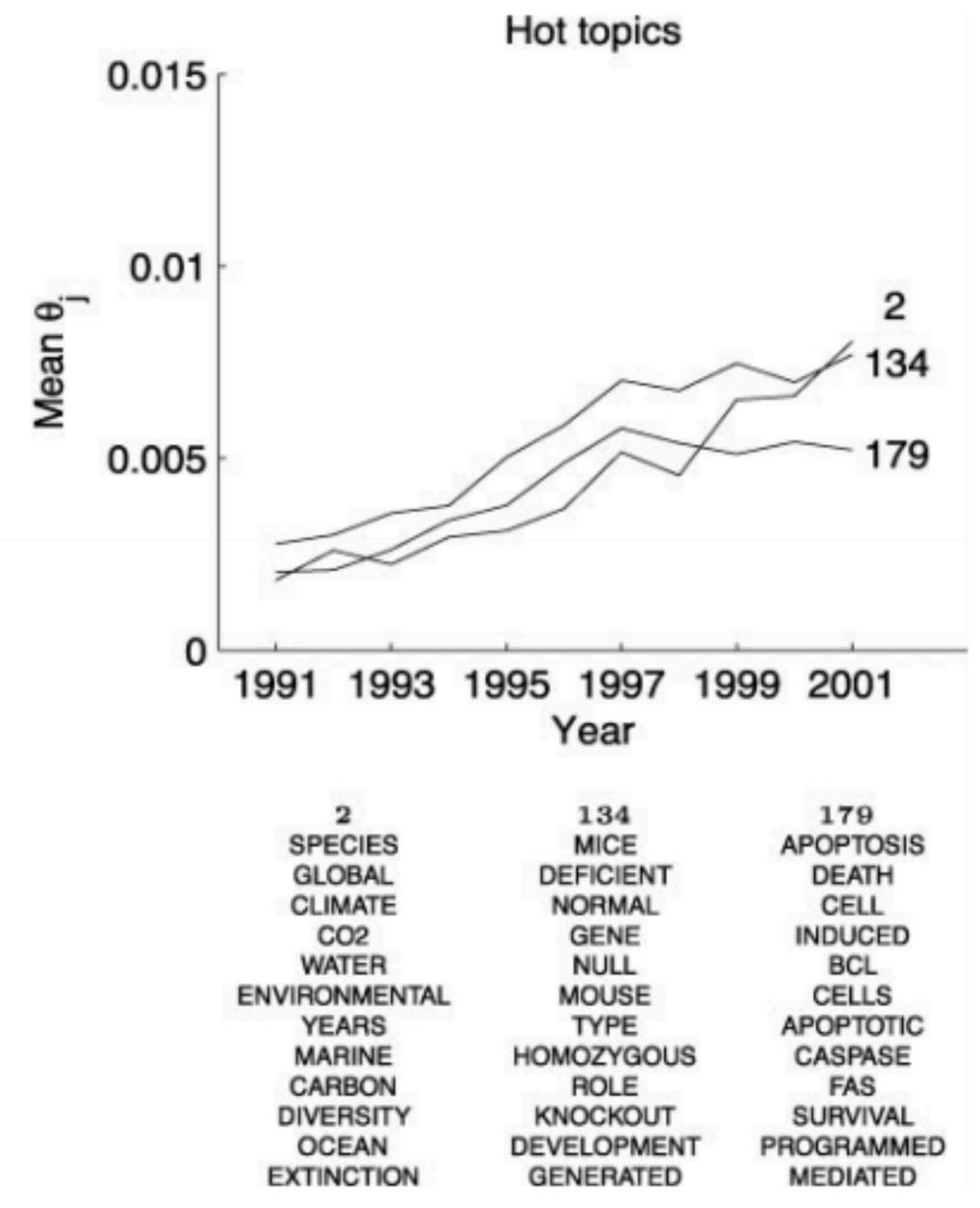
Agricultural Sciences Anthropology Applied Biological Sciences Biochemistry Biophysics Cell Biology Developmental Biology Ecology Evolution Genetics Immunology Medical Sciences Microbiology Neurobiology Pharmacology Physiology Plant Biology Population Biology Psychology Applied Mathematics **Applied Physical Sciences** Astronomy Chemistry Engineering Geology Geophysics Mathematics Physics Anthropology Economic Sciences Psychology Social Sciences 217 274 126 63 200 209 42 2 280 15 64 102 112 210 201 165 142 280 222 39 105 221 270 55 2 2 39 270 114 109 222 120

	1	217	274	126	63	200
		INSECT	SPECIES	GENE	STRUCTURE	FOLDING
		MYB	PHYLOGENETIC	VECTOR	ANGSTROM	NATIVE
		PHEROMONE	EVOLUTION	VECTORS	CRYSTAL	PROTEIN
	[	LENS	SEQUENCES	TRANSFER	RESIDUES	STATE
A CONTRACTOR OF A CONTRACTOR OFTA CONTRACTOR O		42	2	280	15	64
	1	NEURAL	SPECIES	SPECIES	CHROMOSOME	CELLS
		DEVELOPMENT	GLOBAL	SELECTION	REGION	CELL
		DORSAL EMBRYOS	CLIMATE CO2	EVOLUTION GENETIC	CHROMOSOMES KB	ANTIGEN LYMPHOCYTES
	-	VENTRAL	WATER	POPULATIONS	MAP	CD4
	00	112	210	201	165	142
	Biological	HOST	SYNAPTIC	RESISTANCE	CHANNEL	PLANTS
1 1 1 1 1 1 1	ě.	BACTERIAL BACTERIA	NEURONS POSTSYNAPTIC	DRUG	CHANNELS VOLTAGE	PLANT ARABIDOPSIS
	8	STRAINS	HIPPOCAMPAL	DRUGS	CURRENT	TOBACCO
ada in the standard and		SALMONELLA	SYNAPSES	SENSITIVE	CURRENTS	LEAVES
	Sciences					
	8	39	105	221	270	55
	8	THEORY	HAIR	LARGE	TIME	FORCE
		SPACE	MECHANICAL MB	SCALE DENSITY	SPECTROSCOPY	SURFACE MOLECULES
		GIVEN	SENSORY	OBSERVED	SPECTRA	SOLUTION
		PROBLEM	EAR	OBSERVATIONS	TRANSFER	SURFACES
					1111	
				109	120	
	1			RESEARCH	AGE	
mine internet internet				INFORMATION	OLD	
and the second s				UNDERSTANDING	LIFE	
				PAPER	YOUNG	
	1					
- in the in the inter-						
	P					
	Physical					
	2					
	Ő.					
	Science					
	- Ceg					
The second se						
	1					
	0					
	Social					
and a strength of the strength						
	Scie					
	loes					
1 1 1 1 1 1 1	60					

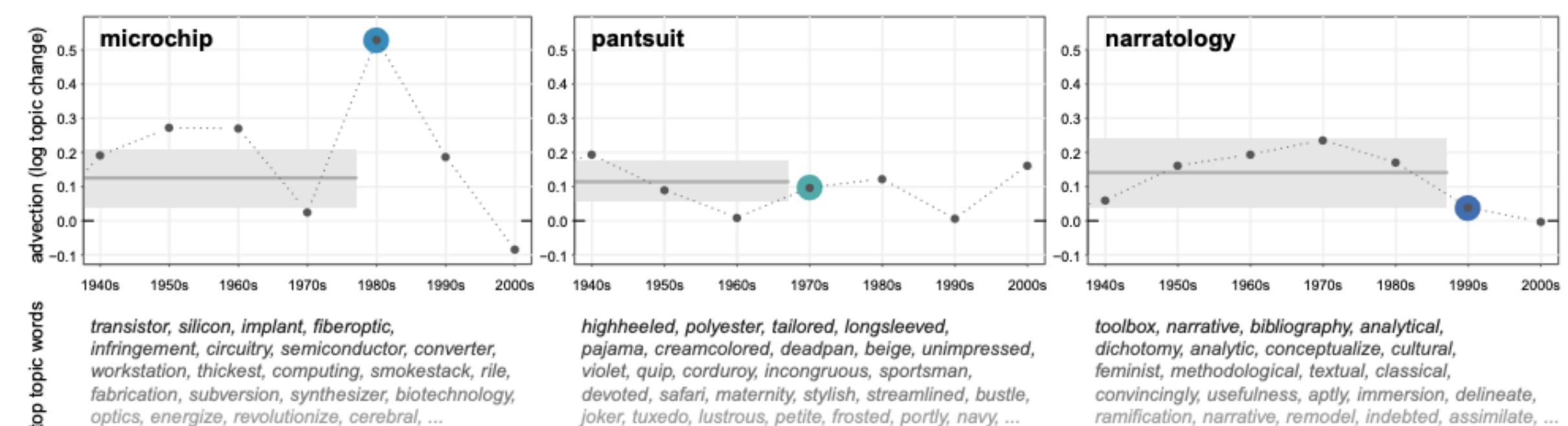


### Finding scientific topics (Griffiths & Steyvers, 2004)





## When do words come into language? (Karjus, Blythe, Kirby, Smith, 2020)



transistor, silicon, implant, fiberoptic, infringement, circuitry, semiconductor, converter, workstation, thickest, computing, smokestack, rile, fabrication, subversion, synthesizer, biotechnology, optics, energize, revolutionize, cerebral, ...

highheeled, polyester, tailored, longsleeved, pajama, creamcolored, deadpan, beige, unimpressed, violet, quip, corduroy, incongruous, sportsman, devoted, safari, maternity, stylish, streamlined, bustle, joker, tuxedo, lustrous, petite, frosted, portly, navy, ...

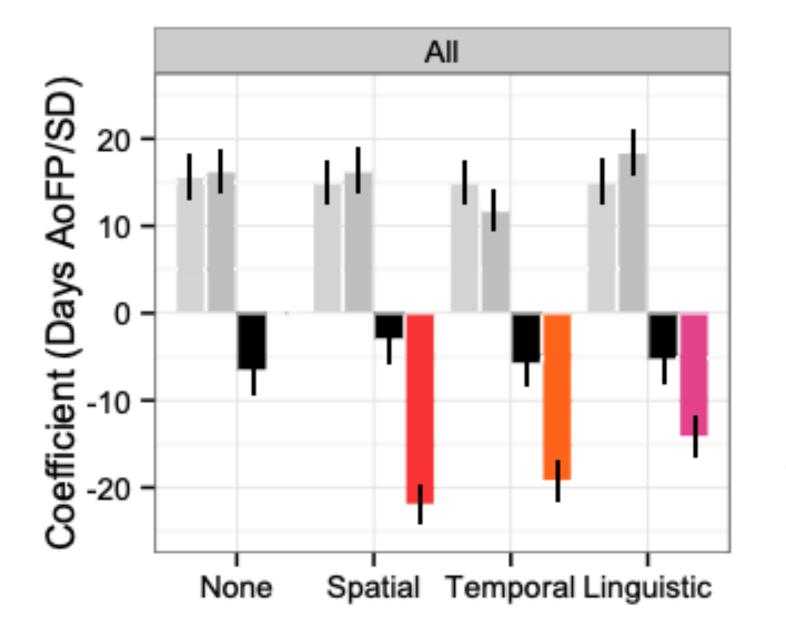
### 58% of words are coined when their topic is trending

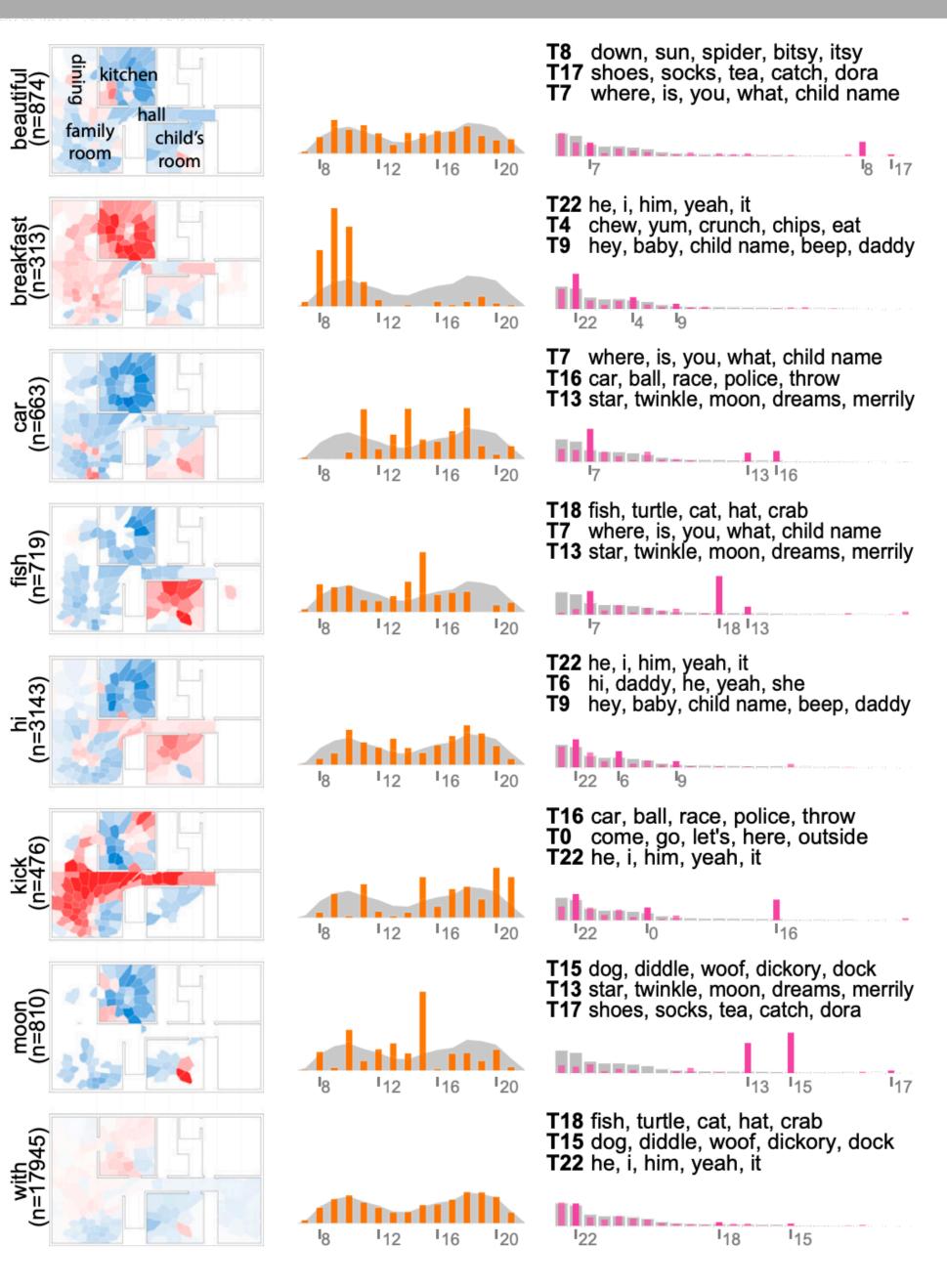
toolbox, narrative, bibliography, analytical, dichotomy, analytic, conceptualize, cultural, feminist, methodological, textual, classical, convincingly, usefulness, aptly, immersion, delineate, ramification, narrative, remodel, indebted, assimilate, ...



## Predicting the birth of words (Roy et al. 2015)







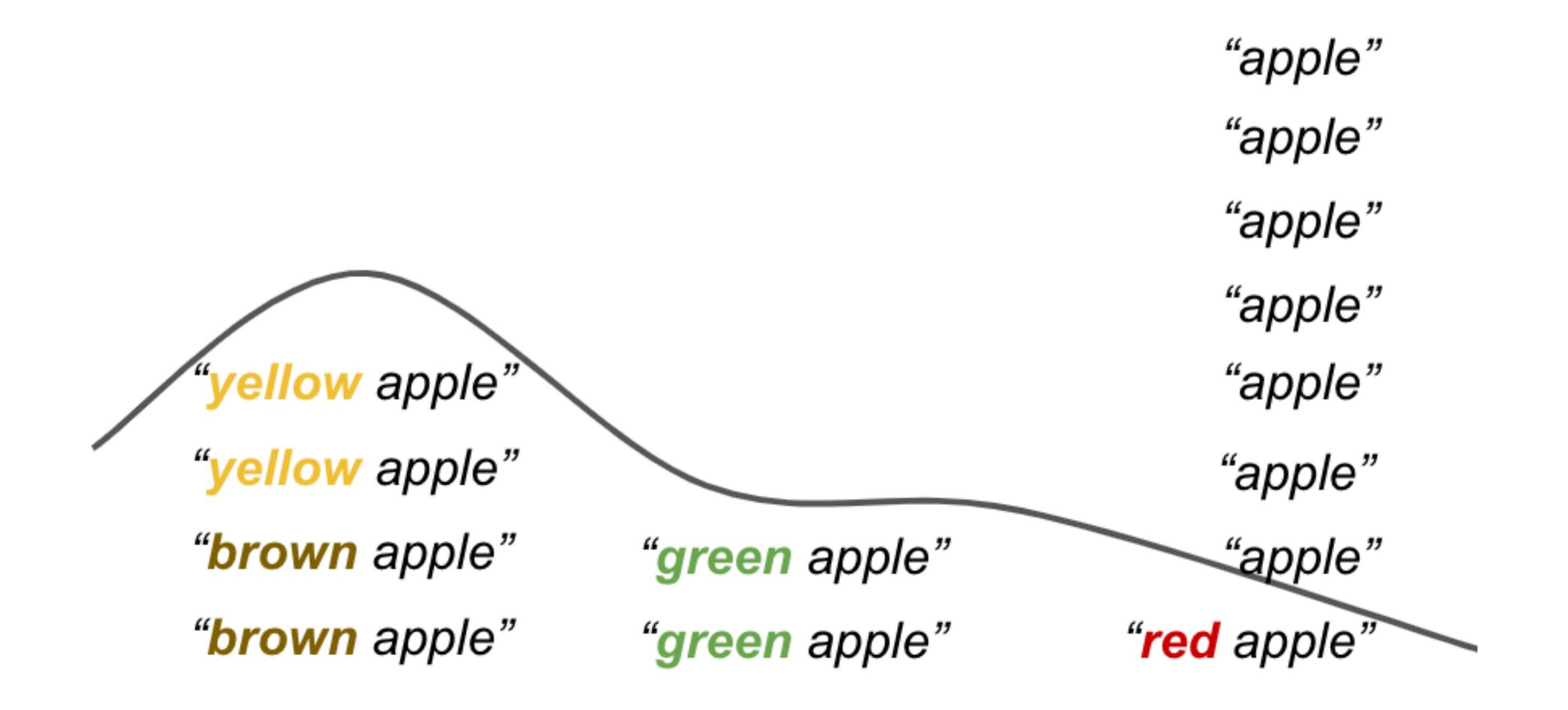
### What's missing from language?



### What's missing from language?



### We don't narrate the world! We talk about atypical things.



### What does speech to children look like? (Bergey, Morris, & Yurovsky, 2019)

### don't put a **plastic cup** in there.

l like your **scaly skin**.

### are **bananas yellow**?

oh, look, are you giving him some nice **curly hair**?

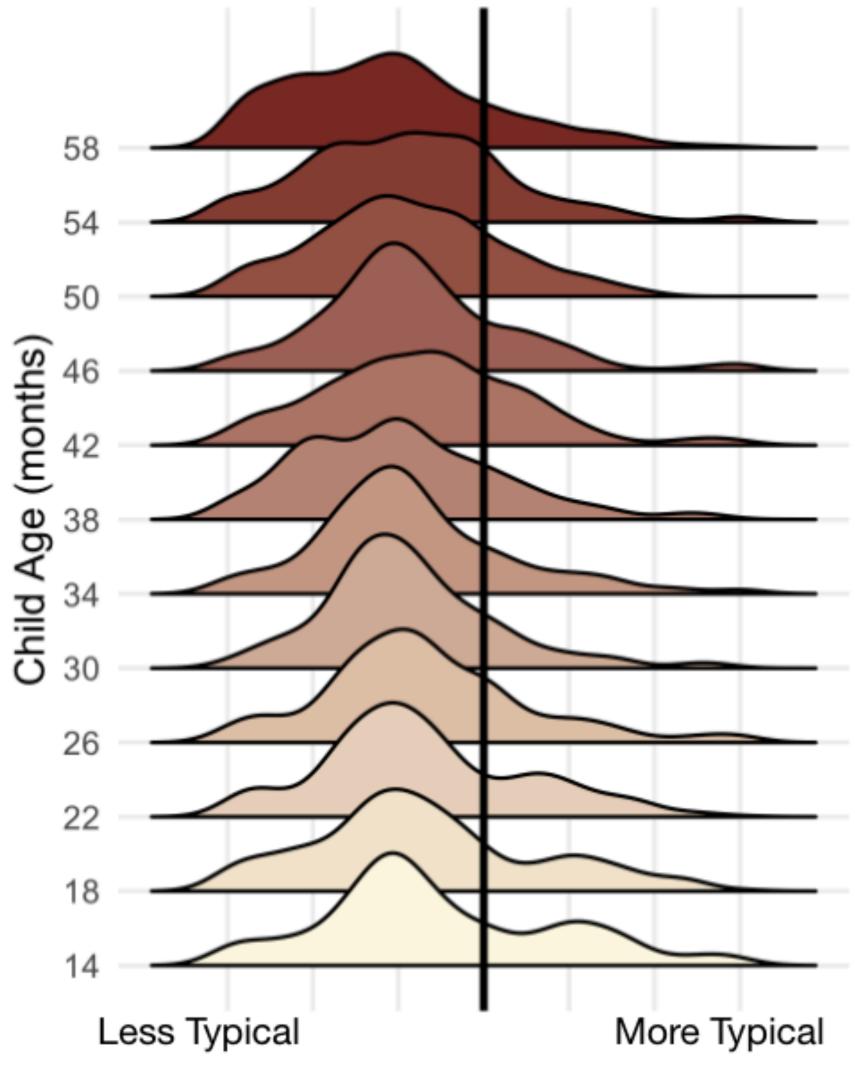
here I'll put it on because your **hands** are **chalky**.

How common is it for an [apple] to be a [yellow apple]?

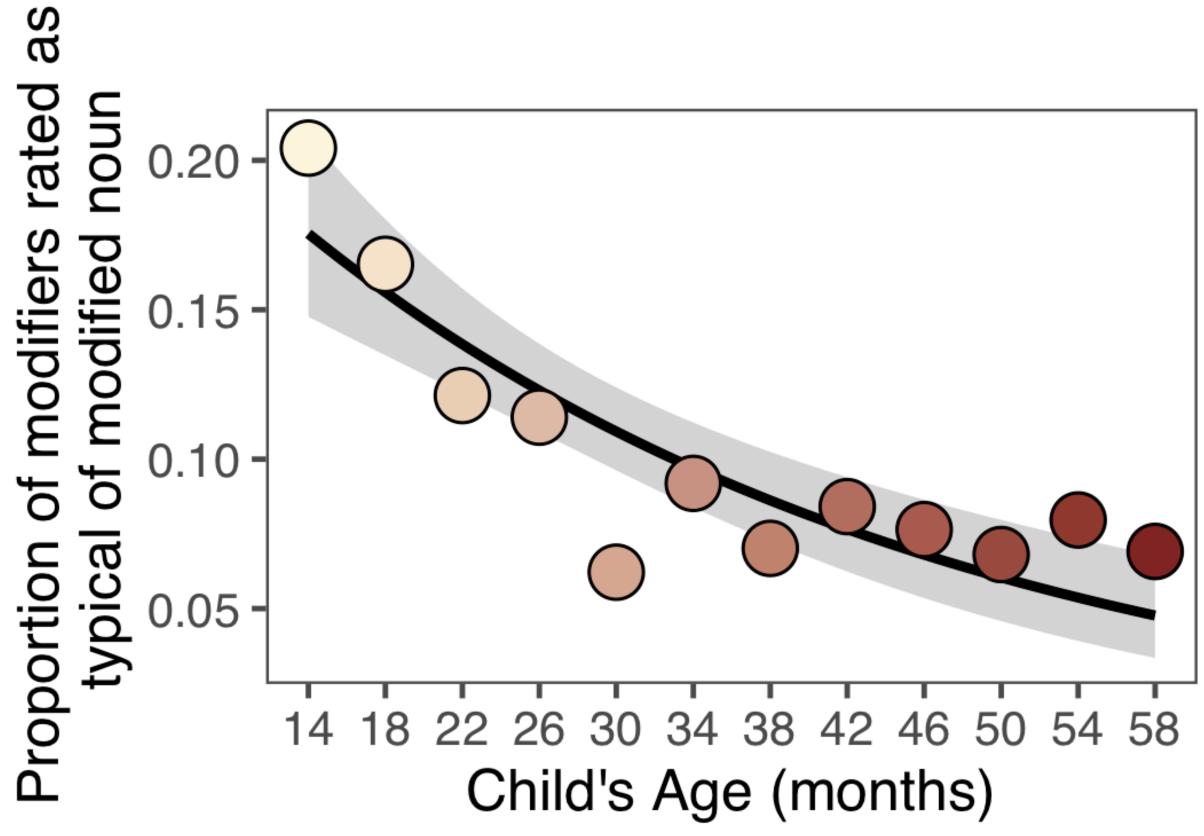
never	rarely	sometimes	about half the time	often	almost always	always
(1)	(2)	(3)	(4)	(5)	(6)	(7)



### Children hear more about atypical colors

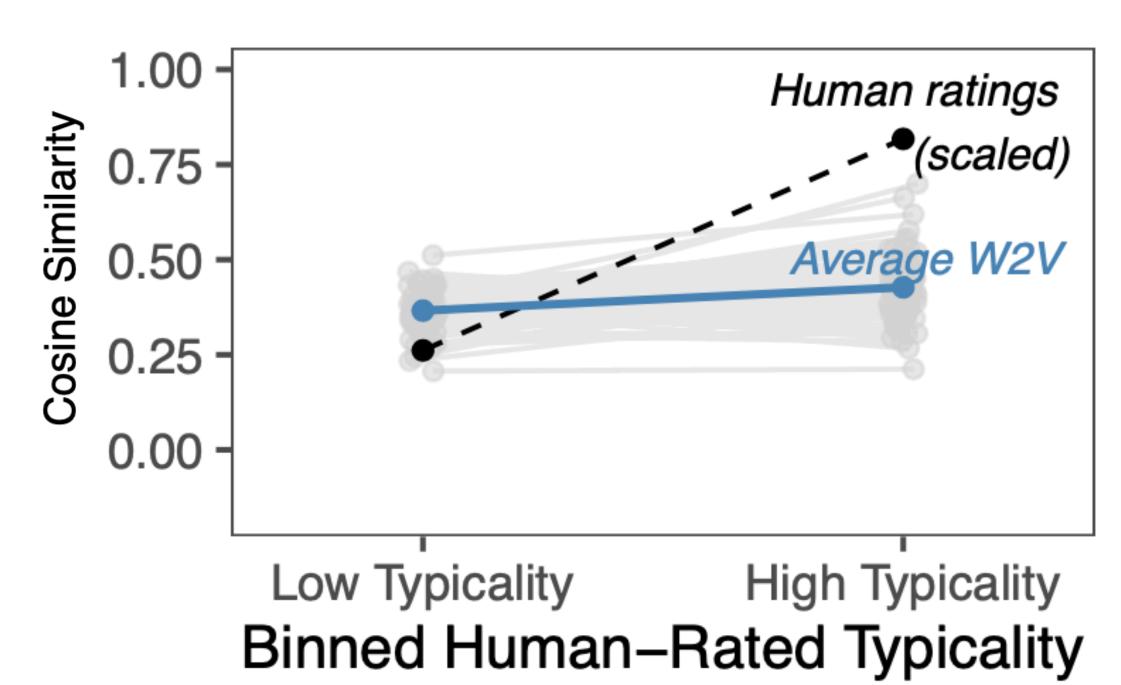


Typicality of adjective-noun pairs



### Semantic space models can't recover typical semantics

noun	typical adjective	atypical adjective
puzzle	flat	giant
apple	red	brown
bird	outside	purple
elephant	fat	pink
whale	wet	red
frog	green	purple





### The structure in language

# 1. You can learn a lot form the co-occurrence structure of words in language

# 2. Latent semantic analysis and Topics models both use this structure to learn about the world

3. But some information is not (straightforwardly) in the co-occurrence structure of language