

## Example learned topics and document model

Train on 160K documents, subset of TREC AP corpus; use variational EM, with 100 topics: compute  $\gamma$  and  $\phi_n$  for test document

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

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 Center's 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.

## Example topics learned from corpus

- learned topics reveal hidden, implicit semantic categories in the corpus
- in many cases, can represent documents with  $10^2$  topics instead of  $10^5$  words
- especially important for short documents, e.g., emails – topics overlap when words don't

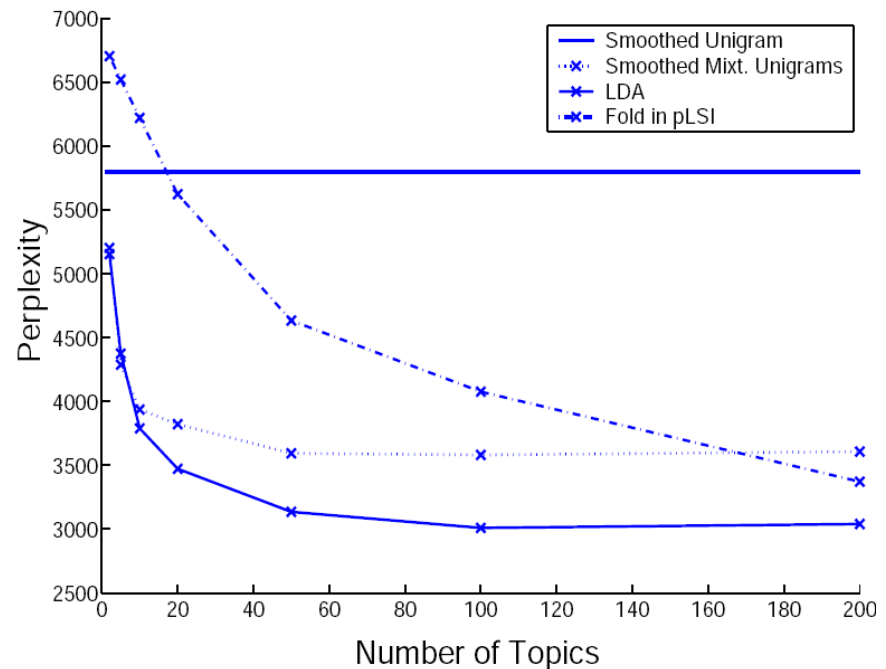
<b>FIELD</b>	SCIENCE	BALL	JOB
MAGNETIC	STUDY	GAME	WORK
MAGNET	SCIENTISTS	TEAM	JOBS
WIRE	SCIENTIFIC	FOOTBALL	CAREER
NEEDLE	KNOWLEDGE	BASEBALL	EXPERIENCE
CURRENT	WORK	PLAYERS	EMPLOYMENT
COIL	RESEARCH	PLAY	OPPORTUNITIES
POLES	CHEMISTRY	<b>FIELD</b>	WORKING
IRON	TECHNOLOGY	PLAYER	TRAINING
COMPASS	MANY	BASKETBALL	SKILLS
LINES	MATHEMATICS	COACH	CAREERS
CORE	BIOLOGY	PLAYED	POSITIONS
ELECTRIC	<b>FIELD</b>	PLAYING	FIND
DIRECTION	PHYSICS	HIT	POSITION
FORCE	LABORATORY	TENNIS	<b>FIELD</b>
MAGNETS	STUDIES	TEAMS	OCCUPATIONS
BE	WORLD	GAMES	REQUIRE
MAGNETISM	SCIENTIST	SPORTS	OPPORTUNITY
POLE	STUDYING	BAT	EARN
INDUCED	SCIENCES	TERRY	ABLE

## How to evaluate model?

Standard topic model results entail showing some suggestive groupings of words into topics; quantitative evaluation not easy

One popular approach in language models: measure **perplexity** of test documents

$$\text{perplexity}(D_{\text{test}}) = \exp \left( - \frac{\sum_{d=1}^M \log p(\mathbf{w}_d)}{\sum_{d=1}^M N_d} \right)$$



Can also consider document classification, representing document using its posterior Dirichlet parameters  $\phi(\mathbf{w})$ : can obtain advantage when only small proportion of dataset labeled

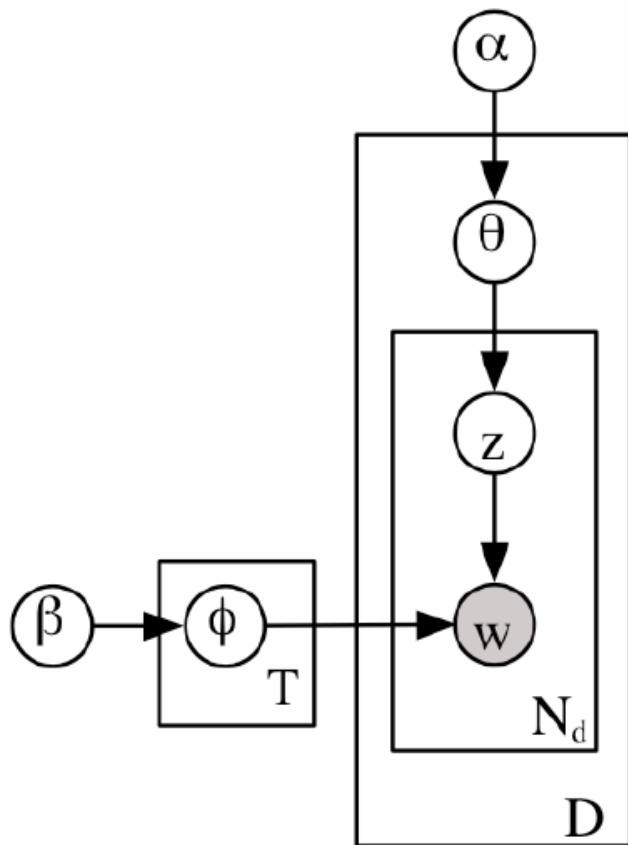
# Author-Recipient-Topic model (McCallum et al., 2007)

extend LDA: analyze roles and relationships between people by analyzing email words wrt topic distributions

## Latent Dirichlet Allocation

(LDA)

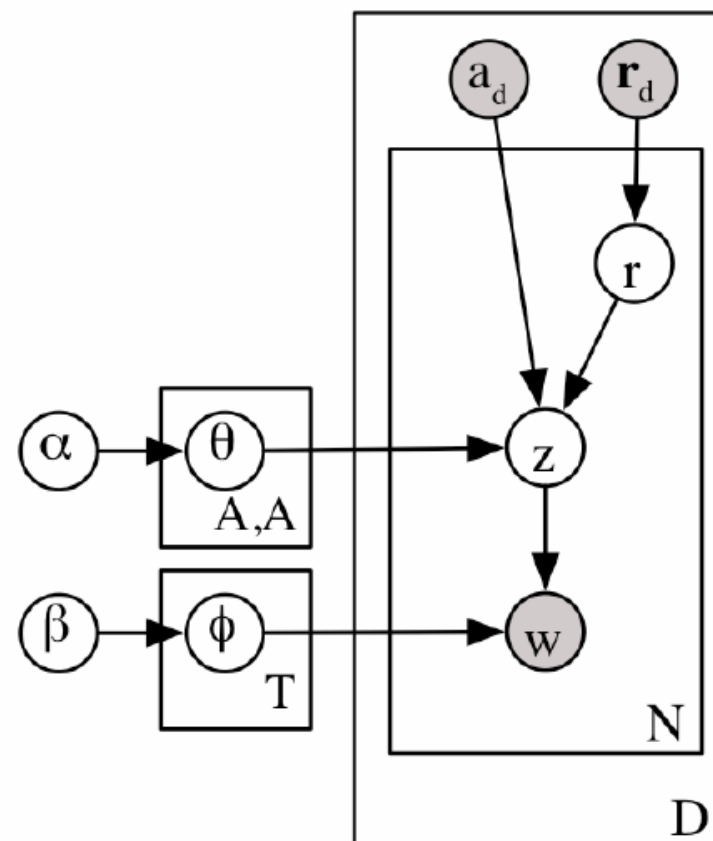
[Blei, Ng, Jordan, 2003]



## Author-Recipient Topic

(ART)

[McCallum, Corrada, Wang, 2004]



## Inference in Author-Recipient-Topic model

models message content, and directed social network in which messages are sent

generative process, for each message  $d$ :

1. observe author  $a_d$  and set of recipients  $\mathbf{r}_d$
2. for each word in message  $d$ 
  - (a) pick recipient  $r$  from  $\mathbf{r}_d$
  - (b) pick topic from author-recipient pair-specific multinomial  $\theta_{a_d,r}$
  - (c) pick word  $w$  from topic-specific multinomial  $\phi_z$

Aim: calculate posterior distribution of topic and recipient assignments given words –  $P(\mathbf{z}, \mathbf{r} | \mathbf{w}) = P(\mathbf{w}, \mathbf{z}, \mathbf{r}) / \sum_{\mathbf{z}, \mathbf{r}} P(\mathbf{w}, \mathbf{z}, \mathbf{r})$

can compute joint, by integrating out unknown  $\phi$  and  $\theta$  distributions (taking advantage of conjugate Dirichlet priors), but denominator cannot be calculated directly

instead use Gibbs sampling (see tutorial)

## Enron email corpus

250K email messages, 147 people, 23K unique words

Date: Wed, 11 Apr 2001 06:56:00 -0700 (PDT)  
From: debra.perlingiere@enron.com  
To: steve.hooser@enron.com  
Subject: Enron/TransAltaContract dated Jan 1, 2001

Please see below. Katalin Kiss of TransAlta has requested an electronic copy of our final draft? Are you OK with this? If so, the only version I have is the original draft without revisions.

DP

Debra Perlingiere  
Enron North America Corp.  
Legal Department  
1400 Smith Street, EB 3885  
Houston, Texas 77002  
dperlin@enron.com

## Topics and prominent sender/receivers

Top words  
within topic :

<b>Topic 17 “Document Review”</b>		<b>Topic 27 “Time Scheduling”</b>		<b>Topic 45 “Sports Pool”</b>	
attached	0.0742	day	0.0419	game	0.0170
agreement	0.0493	friday	0.0418	draft	0.0156
review	0.0340	morning	0.0369	week	0.0135
questions	0.0257	monday	0.0282	team	0.0135
draft	0.0245	office	0.0282	eric	0.0130
letter	0.0239	wednesday	0.0267	make	0.0125
comments	0.0207	tuesday	0.0261	free	0.0107
copy	0.0165	time	0.0218	year	0.0106
revised	0.0161	good	0.0214	pick	0.0097
document	0.0156	thursday	0.0191	phillip	0.0095

Top  
author-recipients  
exhibiting this  
topic

G.Nemec	0.0737	J.Dasovich	0.0340	E.Bass	0.3050
B.Tycholiz		R.Shapiro		M.Lenhart	
G.Nemec	0.0551	J.Dasovich	0.0289	E.Bass	0.0780
M.Whitt		J.Steffes		P.Love	
B.Tycholiz	0.0325	C.Clair	0.0175	M.Motley	0.0522
G.Nemec		M.Taylor		M.Grigsby	

## ART: Learns roles

ART implicitly finds roles of individuals

Topic 34 “Operations”		Topic 37 “Power Market”		Topic 41 “Government Relations”		Topic 42 “Wireless”	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck	0.2158	J.Dasovich	0.1231	J.Dasovich	0.3338	R.Haylett	0.1432
L.Kitchen		J.Steffes		R.Shapiro		T.Geaccone	
S.Beck	0.0826	J.Dasovich	0.1133	J.Dasovich	0.2440	T.Geaccone	0.0737
J.Lavorato		R.Shapiro		J.Steffes		R.Haylett	
S.Beck	0.0530	M.Taylor	0.0218	J.Dasovich	0.1394	R.Haylett	0.0420
S.White		E.Sager		R.Sanders		D.Fossum	

**Beck = “Chief Operations Officer”**

**Dasovich = “Government Relations Executive”**

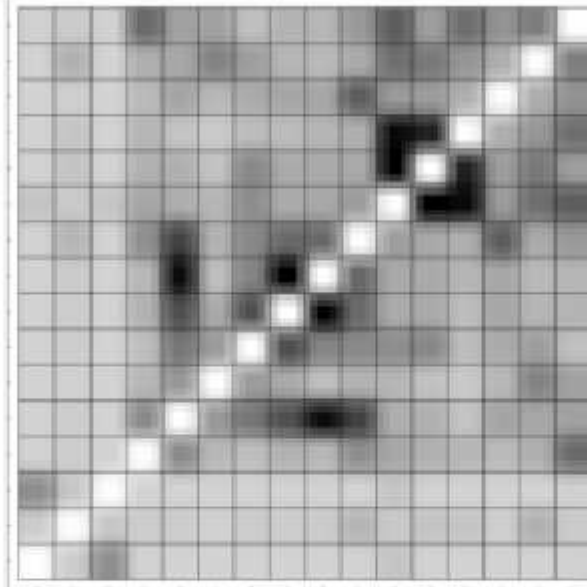
**Shapiro = “Vice Presidency of Regulatory Affairs”**

**Steffes = “Vice President of Government Affairs”**

# Discovering role similarity

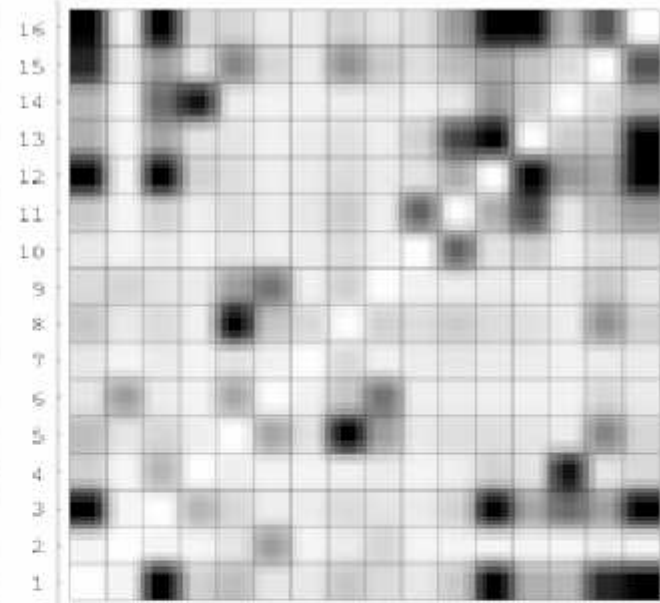
## Traditional SNA

```
16 : teb.lokey
15 : steven.harris
4 : kimberly.watson
13 : paul.y'barbo
12 : bill.rapp
11 : kevin.hyatt
10 : drew.fossum
9 : tracy.geaconne
8 : danny.mccarty
7 : shelley.corman
6 : rod.hayslett
5 : stanley.horton
4 : lynn.blair
3 : paul.thomas
2 : larry.campbell
: joe.stepenovitch
```



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

## ART



1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

**connection strength (A,B) =**

**Similarity in recipients they sent email to**

**Similarity in authored topics, conditioned on recipient**

reflects jobs: Blair ('gas pipeline logistics')  $\approx$  Watson ('pipeline facility planning'); Geaconne ('executive assistant') vs. McCarty ('vice-president')

## Dynamic Topic Models (Blei & Lafferty, 2006)

imagine topics evolve over time, so order of documents important; assume data divided by time-slice (e.g., year)

both Dirichlet distributions (over document topic proportions, and topic word proportions) replaced by simple dynamic model

### 1. Draw topics

$$\beta_t | \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$$

### 2. $\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \gamma^2 I)$

### 3. for each document

#### (a) $\theta \sim \mathcal{N}(\alpha_t, a^2 I)$

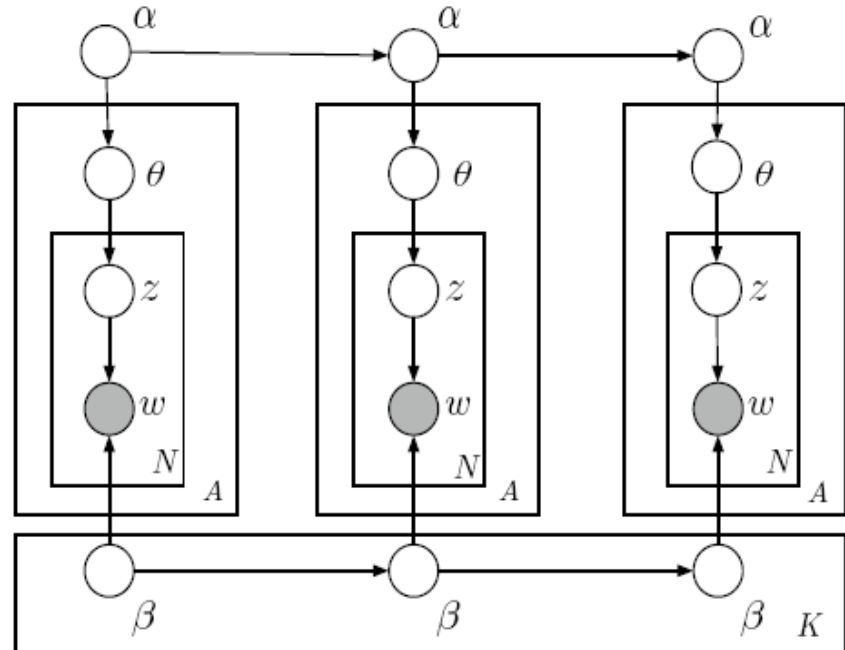
#### (b) for each word

##### i. $Z \sim$

$$\text{Mult}(\text{softmax}(\theta))$$

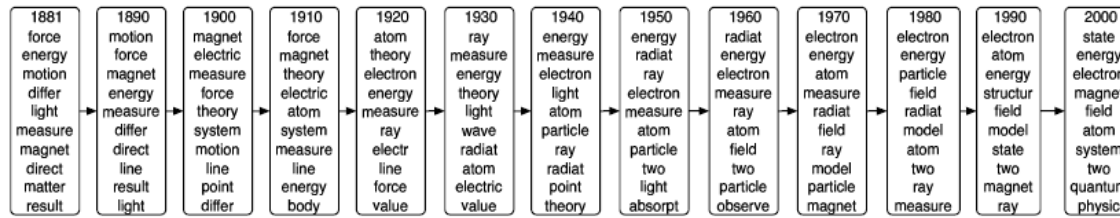
##### ii. $W_{tdn} \sim$

$$\text{Mult}(\text{softmax}(\beta_t z))$$

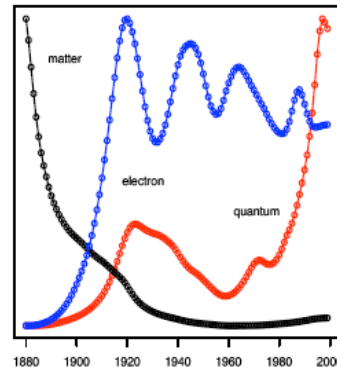


# Dynamic Topic Models: Results

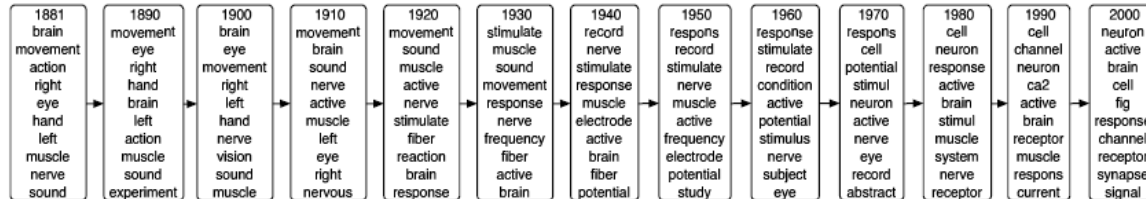
Science corpus: 30K articles, 1881-1999, 250/yr; 16K vocabulary; 20-topic dynamic model; trained using Kalman filter variational approximation



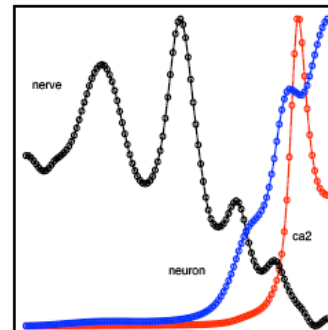
"Atomic Physics"



- 1881 On Matter as a form of Energy
- 1892 Non-Euclidean Geometry
- 1900 On Kathode Rays and Some Related Phenomena
- 1917 "Keep Your Eye on the Ball"
- 1920 The Arrangement of Atoms in Some Common Metals
- 1933 Studies in Nuclear Physics
- 1943 Aristotle, Newton, Einstein. II
- 1950 Instrumentation for Radioactivity
- 1965 Lasers
- 1975 Particle Physics: Evidence for Magnetic Monopole Obtained
- 1985 Fermilab Tests its Antiproton Factory
- 1999 Quantum Computing with Electrons Floating on Liquid Helium



"Neuroscience"



- 1887 Mental Science
- 1900 Hemianopsia in Migraine
- 1912 A Defence of the "New Phrenology"
- 1921 The Synchronal Flashing of Fireflies
- 1932 Myoesthesia and Imageless Thought
- 1943 Acetylcholine and the Physiology of the Nervous System
- 1952 Brain Waves and Unit Discharge in Cerebral Cortex
- 1963 Errorless Discrimination Learning in the Pigeon
- 1974 Temporal Summation of Light by a Vertebrate Visual Receptor
- 1983 Hysteresis in the Force-Calcium Relation in Muscle
- 1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

## Infinite Topic Models

So far all the topic models require specification of the number of topics

now consider infinite version, where the number of topics is potentially infinite

non-intuitive, yet fundamental idea underlying nonparametric Bayesian statistics

represent only as many topics as needed for a given dataset

examples of infinite models: Gaussian processes, Dirichlet process mixture models