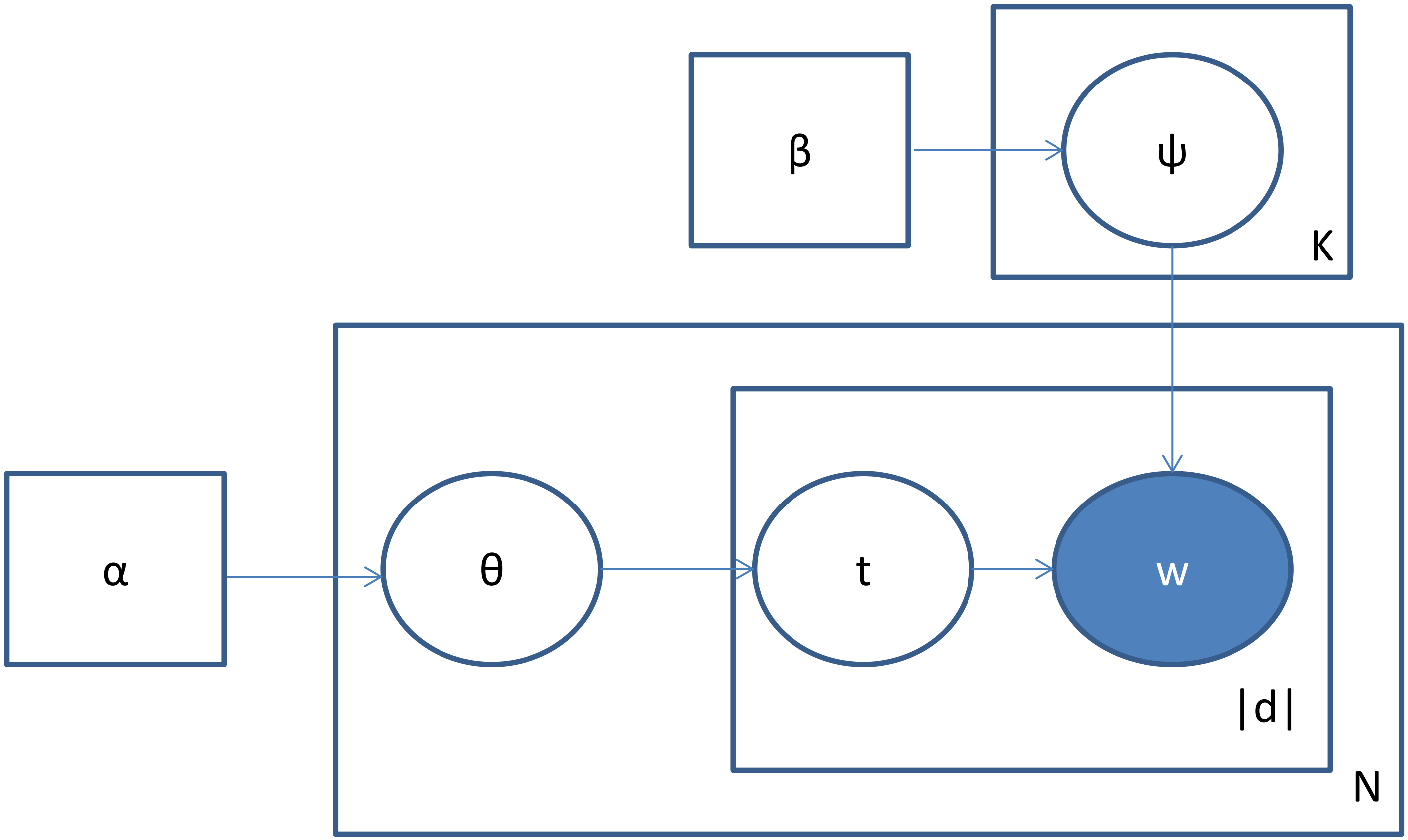
Befriending LDA

Nisheeth

LDA in plate notation



Generative model

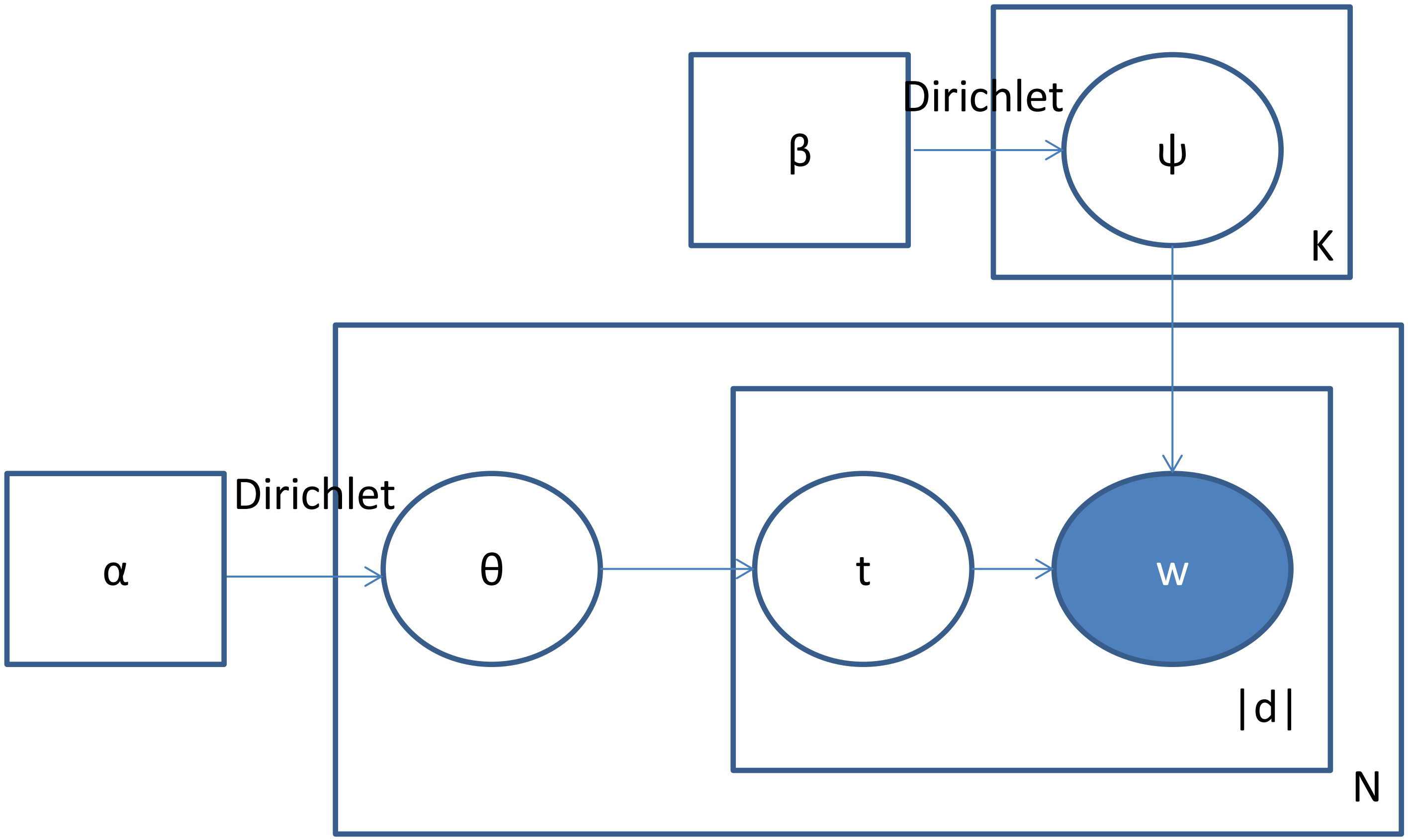
- For all documents
  - Generate $\theta \sim \text{Dirichlet}(\alpha)$
  - Generate all K $\psi \sim \text{Dirichlet}(\beta)$
- For all words in each document
  - Generate $t \sim \text{Multinomial}(\theta)$
  - Generate $w \sim \text{Multinomial}(\psi_t)$

LDA math – the Dirichlet distribution

- A *k*-dimensional Dirichlet random variable θ can take values in the (k-1)-simplex, and has the following probability density on this simplex:

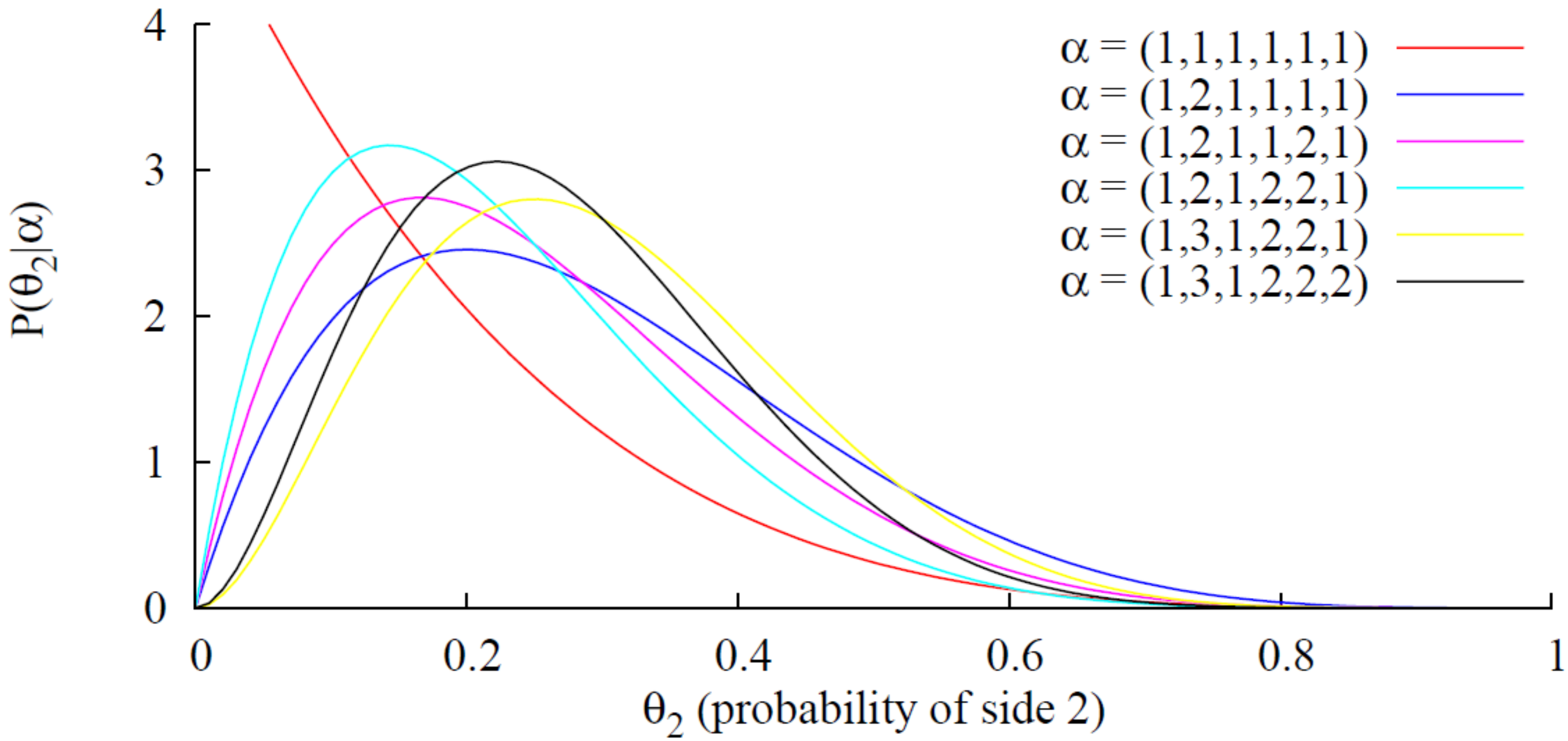
$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

- Easier to understand
  - Prior Dir(α_1, α_2)
  - Likelihood Multi(θ_1 , θ_2)
  - Outcome {n_1, n_2}
  - Posterior Dir($\alpha_1 + n_1$, $\alpha_2 + n_2$)
- Ignoring the normalization constant, what is the Dirichlet probability of a multinomial sample [0.1, 0.5, 0.4] with parameter 10
  - $(0.1)^9 (0.5)^9 (0.4)^9$ = 5e-16
- What would it be for parameter 0.2?
  - 22


β
Dirichlet
ψ
K
α
Dirichlet
θ
t
w
|d|
N

Dirichlet update – dice roll

- Data $d = (2, 5, 4, 2, 6)$

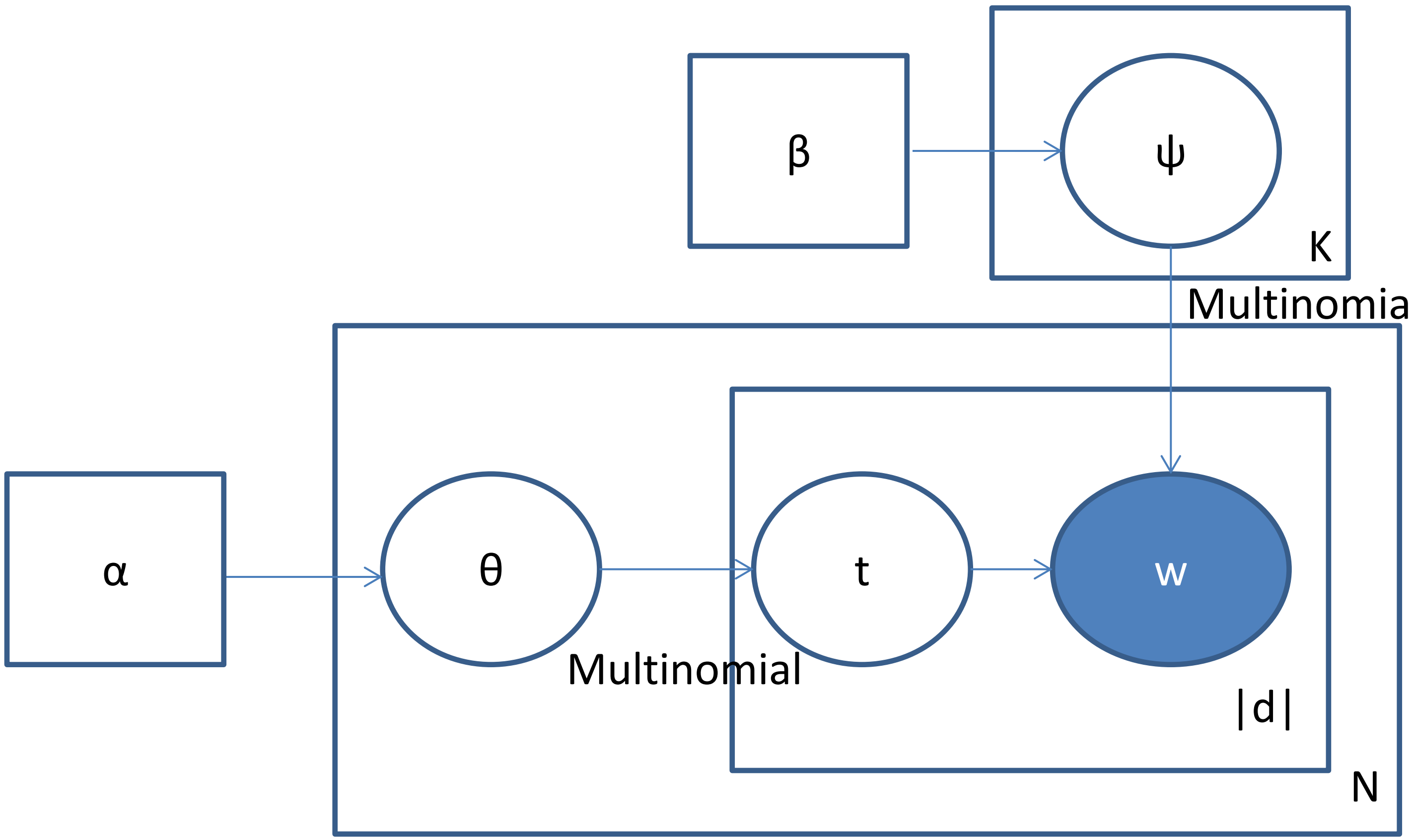


LDA math – the multinomial distribution

- For n independent trials that could yield exactly one of k possible results, the multinomial distribution gives the probability of seeing any particular combination of outcomes

$$p(\boldsymbol{x}, \boldsymbol{\gamma}) = \frac{n!}{x_1!\ x_2! \dots x_k!} {\gamma_1}^{x_1} {\gamma_2}^{x_2} \dots {\gamma_k}^{x_k}$$

- Parameterized by **γ** and n
- Easier to understand
  - Tracks word frequencies
  - Given a vocabulary of 3 words A,B,C with normalized empirical frequencies [0.3, 0.4, 0.3] in a corpus and a document AABB
  - $p(document) = \frac{4!}{2!2!0!}\,(0.3)^2(0.4)^2 = 0.0864$
  - Given normalized empirical frequencies [0.1,0.1,0.8], what would the probability of the same document be?
  - Given normalized empirical frequencies [0.3, 0.4, 0.3] and a document A, what would its probability be?



p(w|t) is high when many words in a document show up as high frequency terms in the corresponding topic word distribution

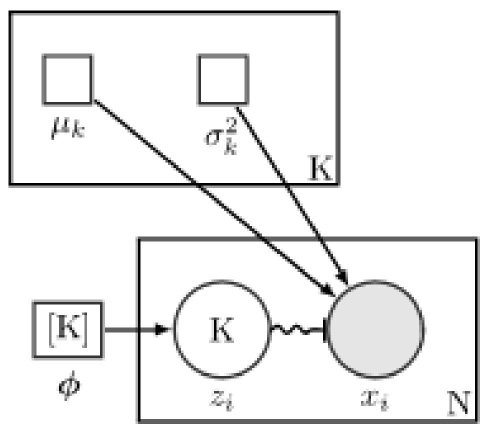
ψ is distribution of words in a topic

p(t|θ) is high when many words in a topic show up as high frequency terms in the document topic distribution

θ is distribution of topics in a document

Compare with Gaussian mixture model

- p(K|ϕ) is high when the ϕ value is high for the Kth label
- p(x|K) is high when x is statistically likely to be drawn from the Gaussian with the Kth summary statistics



LDA inference

- Latent variable inference

$$p(\theta, \boldsymbol{t} | \boldsymbol{w}, \alpha, \psi) = \frac{p(\theta, \boldsymbol{z}, \boldsymbol{w} | \alpha, \psi)}{p(\boldsymbol{w} | \alpha, \psi)}$$

- From the graphical model

$$p(\theta, \boldsymbol{z}, \boldsymbol{w} | \alpha, \psi) = p(\boldsymbol{w} | \boldsymbol{t}, \psi) p(\boldsymbol{t} | \theta) p(\theta | \alpha)$$

- What are these terms?
  1. $p(\boldsymbol{w} | \boldsymbol{t}, \psi) = \prod_{n=1}^{|d|} \psi_{t_n, w_n}$
  2. $p(\boldsymbol{t} | \theta) = \prod_{n=1}^{|d|} \theta_{t_n}$
  3. $p(\theta | \alpha) = C(\alpha) \sum_{i=1}^{K} {\theta_i}^{\alpha_i - 1}$

LDA intuition

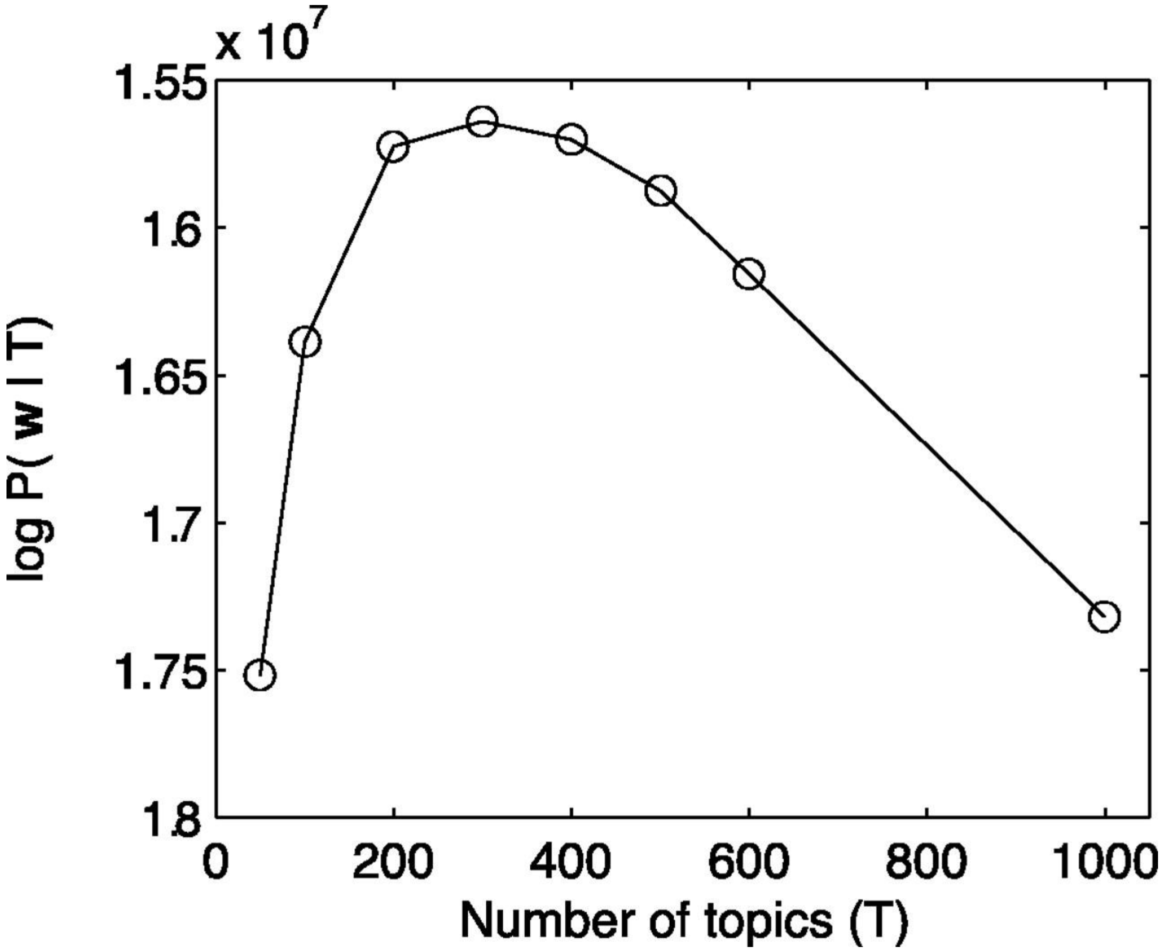
- Given the optimal denominator, the correct partitioning of the data into topics is determined by the numerator
- What does the numerator say about what constitutes a good topic partitioning?
  1. $p(\boldsymbol{w}|\boldsymbol{t},\psi)$ will have high values iff ψ is sparse
  2. $p(\boldsymbol{t}|\theta)$ will have high values iff θ is concentrated
  3. $p(\theta|\alpha)$ will have high values if α is small
- Implications
  1. Better to have non-overlapping topics
  2. Better to have fewer topics per document
  3. Better to be biased towards few topics in general
- Net upshot: make clusters with co-occurring terms

LDA inference

- From these building blocks we get the full numerator
- Denominator obtained by marginalizing over the latent variables
  - Involves an intractable integral
  - Have to use approximate inference methods
    - Variational EM
    - Gibbs sampling
- MLE inference is standard

$$\ell(\alpha, \beta) = \sum_{d=1}^{N} \log p(\mathbf{w}_d \mid \alpha, \beta)$$

Model selection



Document modeling

- Unlabeled data – our goal is density estimation.
- Compute the *perplexity* of a held-out test to evaluate the models – lower perplexity score indicates better generalization.

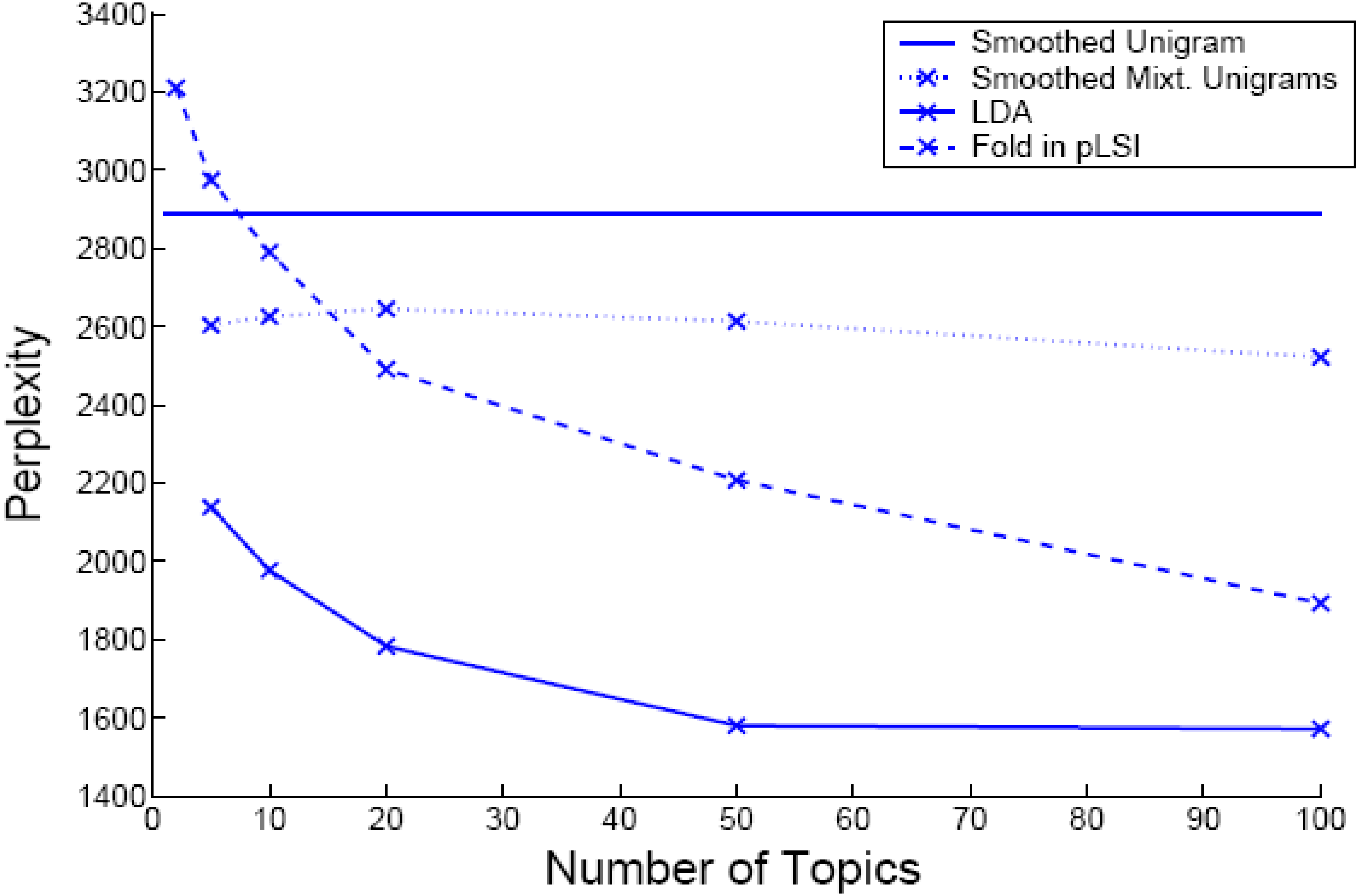
$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(\mathbf{w}_d)}{\sum_{d=1}^{M} N_d}\right\}$$

.

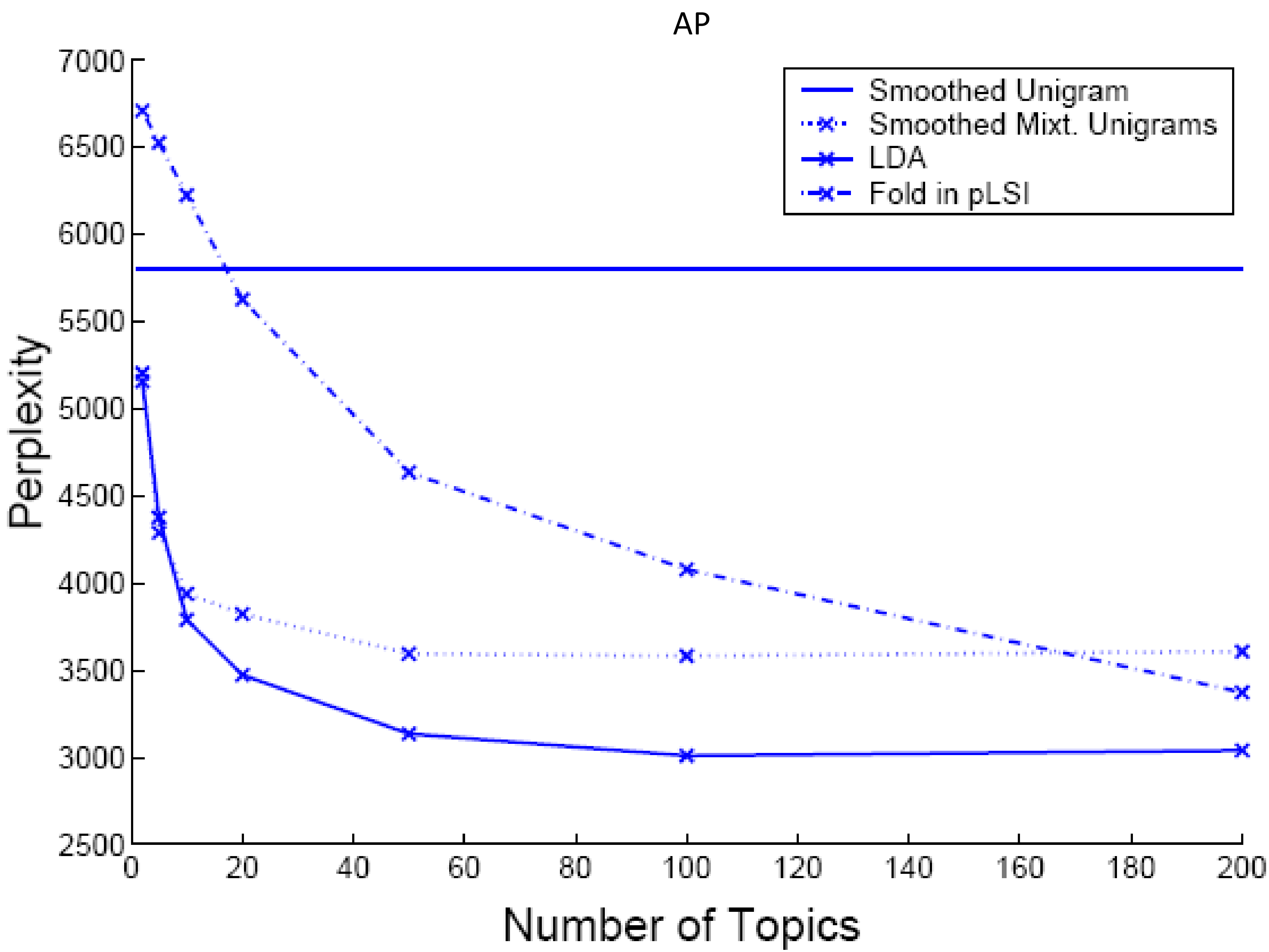
Document Modeling – cont.
data used

- C. Elegans Community abstracts
  - 5,225 abstracts
  - 28,414 unique terms
- TREC AP corpus (subset)
  - 16,333 newswire articles
  - 23,075 unique terms
- Held-out data – 10%
- Removed terms – 50 stop words, words appearing once (AP)

nematode



Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
Number of Topics
3400
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
0
10
20
30
40
50
60
70
80
90
100


AP
Smoothed Unigram
Smoothed Mixt. Unigrams
LDA
Fold in pLSI
Perplexity
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
0
20
40
60
80
100
120
140
160
180
200
Number of Topics

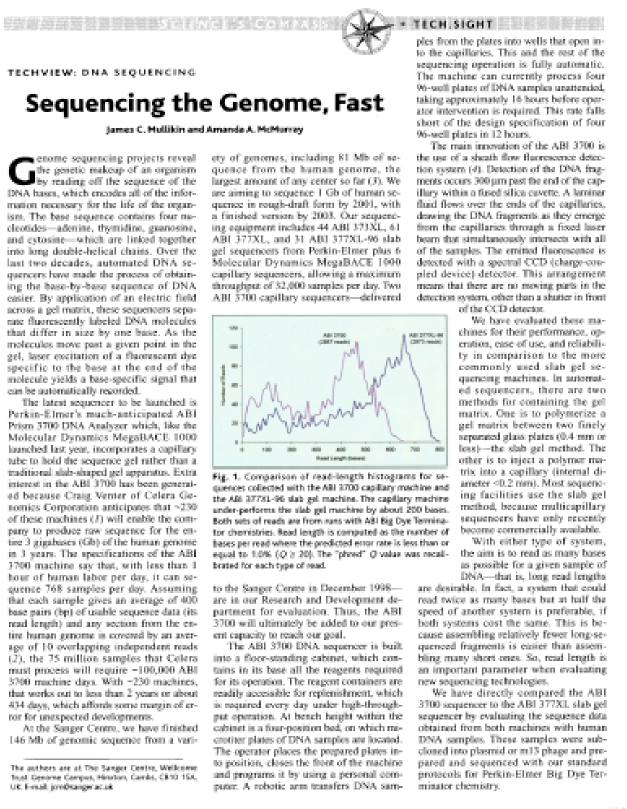
What can you get from it?

"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

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.

Topic membership

Original article


SCIENCE'S COMPASS • TECH.SIGHT

TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurray

Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine under-performs the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistries. Read length is computed as the number of bases per read where the predicted error rate is less than or equal to 1.0% ($Q \geq 20$). The "phred" Q value was recalibrated for each type of read.

www.sciencemag.org SCIENCE VOL 283 19 MARCH 1999 1867

Most likely words from top topics

sequence	devices	data
genome	device	information
genes	materials	network
sequences	current	web
human	high	computer
gene	gate	language
dna	light	networks
sequencing	silicon	time
chromosome	material	software
regions	technology	system
analysis	electrical	words
data	fiber	algorithm
genomic	power	number
number	based	internet

Document similarity

$$d_{ij} = \mathrm{E}\left[\sum_{k=1}^{K}(\sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}})^2 \mid \mathbf{w}_i, \mathbf{w}_j\right]$$

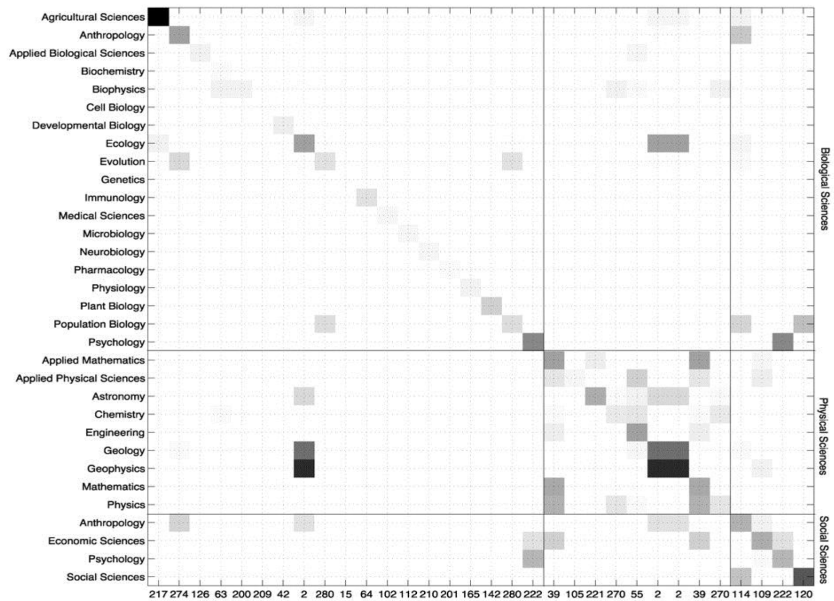
Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Topic similarity



217 INSECT MYB PHEROMONE LENS LARVAE	274 SPECIES PHYLOGENETIC EVOLUTION EVOLUTIONARY SEQUENCES	126 GENE VECTOR VECTORS EXPRESSION TRANSFER	63 STRUCTURE ANGSTROM CRYSTAL RESIDUES STRUCTURES	200 FOLDING NATIVE PROTEIN STATE ENERGY	209 NUCLEAR NUCLEUS LOCALIZATION CYTOPLASM EXPORT
42 NEURAL DEVELOPMENT DORSAL EMBRYOS VENTRAL	2 SPECIES GLOBAL CLIMATE CO2 WATER	280 SPECIES SELECTION EVOLUTION GENETIC POPULATIONS	15 CHROMOSOME REGION CHROMOSOMES KB MAP	64 CELLS CELL ANTIGEN LYMPHOCYTES CD4	102 TUMOR CANCER TUMORS HUMAN CELLS
112 HOST BACTERIAL BACTERIA STRAINS SALMONELLA	210 SYNAPTIC NEURONS POSTSYNAPTIC HIPPOCAMPAL SYNAPSES	201 RESISTANCE RESISTANT DRUG DRUGS SENSITIVE	165 CHANNEL CHANNELS VOLTAGE CURRENT CURRENTS	142 PLANTS PLANT ARABIDOPSIS TOBACCO LEAVES	222 CORTEX BRAIN SUBJECTS TASK AREAS
39 THEORY TIME SPACE GIVEN PROBLEM	105 HAIR MECHANICAL MB SENSORY EAR	221 LARGE SCALE DENSITY OBSERVED OBSERVATIONS	270 TIME SPECTROSCOPY NMR SPECTRA TRANSFER	55 FORCE SURFACE MOLECULES SOLUTION SURFACES	114 POPULATION POPULATIONS GENETIC DIVERSITY ISOLATES
		109 RESEARCH NEW INFORMATION UNDERSTANDING PAPER	120 AGE OLD AGING LIFE YOUNG		

Thomas L. Griffiths, and Mark Steyvers PNAS 2004;101:5228-5235

PNAS

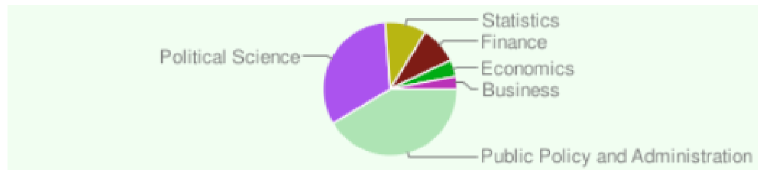
Document tagging, relevance scoring

Tax Innovation in the States: Capitalizing on Political Opportunity

Frances Stokes Berry;William D. Berry. *American Journal of Political Science* (1992), pp. 715-742

Journal Disciplines:

- **Political Science**

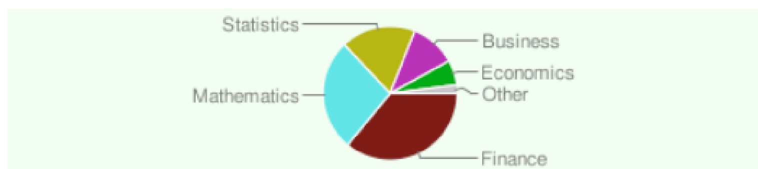


Chaos and Nonlinear Forecastability in Economics and Finance

Blake LeBaron. *Philosophical Transactions: Physical Sciences and Engineering* (1994), pp. 397-404

Journal Disciplines:

- **Mathematics**
- **Biological Sciences**
- **General Science**

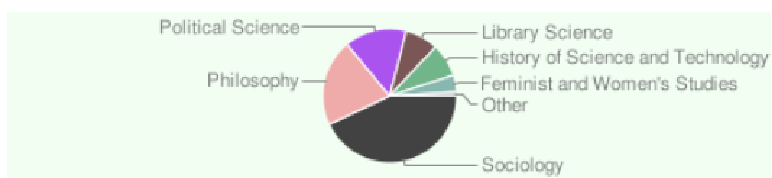


Reply: Theory Is Not a Social Dilemma

Gerald Marwell;Pamela Oliver. *Social Psychology Quarterly* (1994), pp. 373

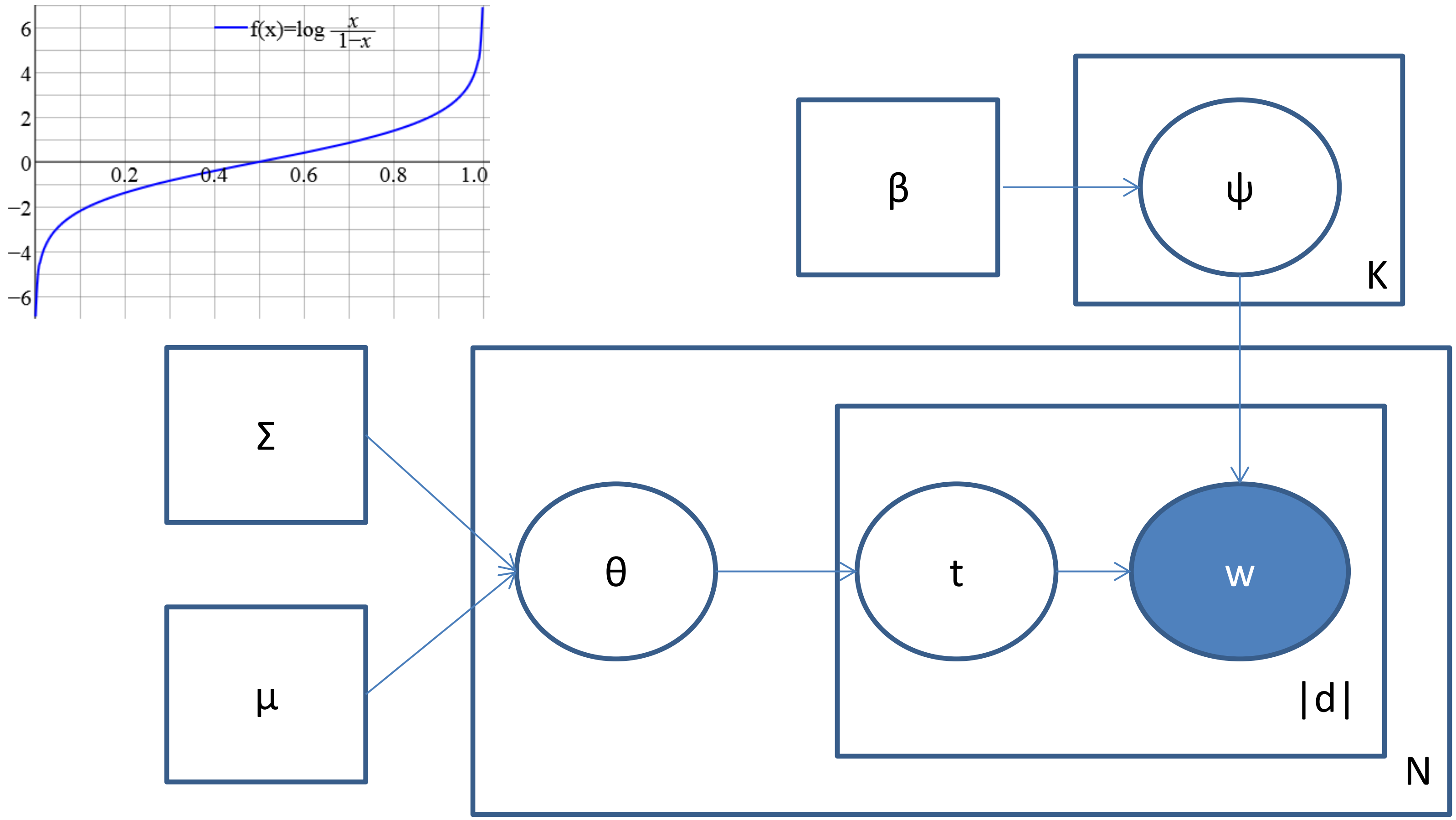
Journal Disciplines:

- **Psychology**
- **Sociology**



Extension: correlated topic models

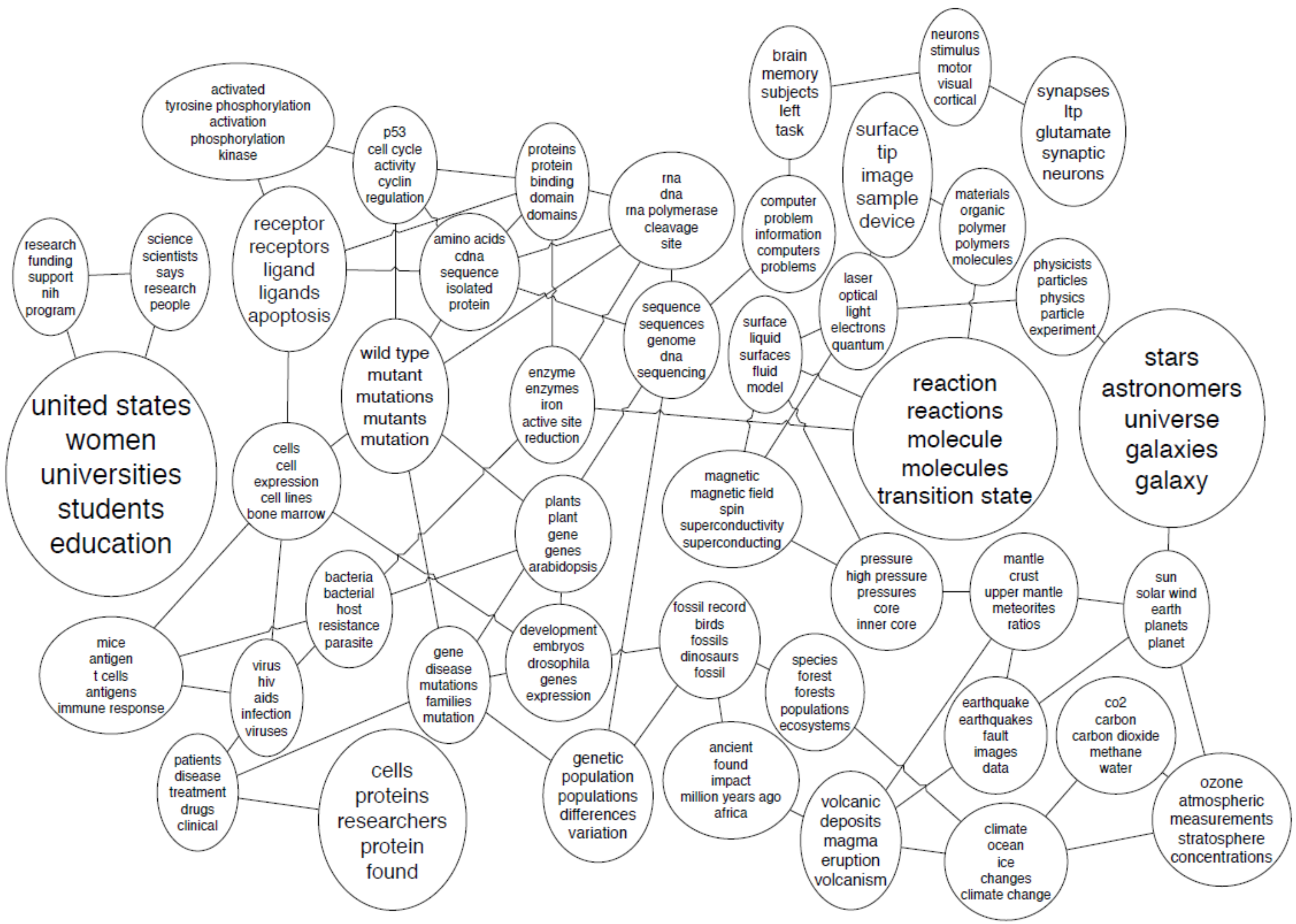
Logit plot



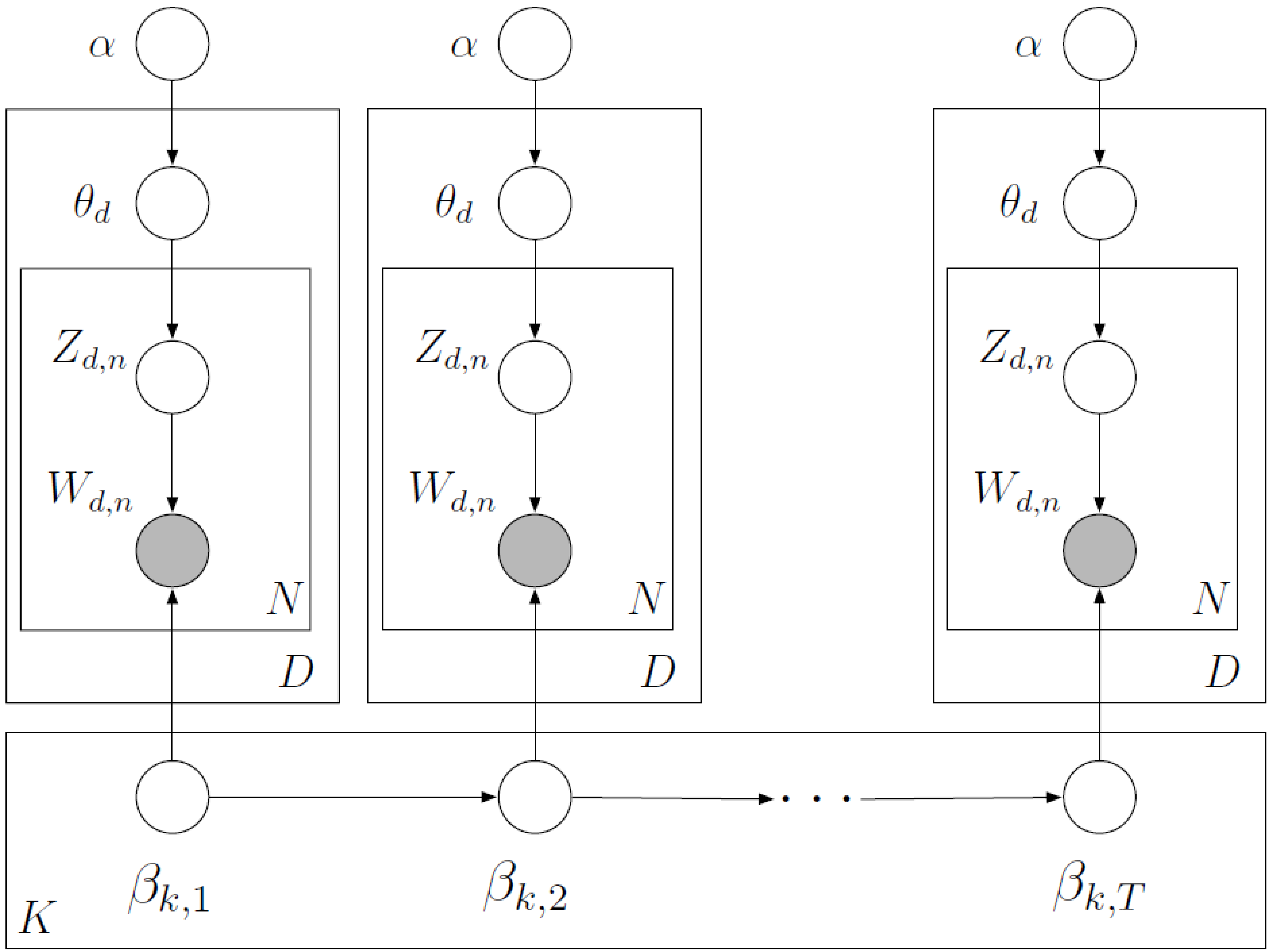
$$\boldsymbol{x} \sim N(\mu, \Sigma)$$

$$\theta \propto \exp(x_i)$$

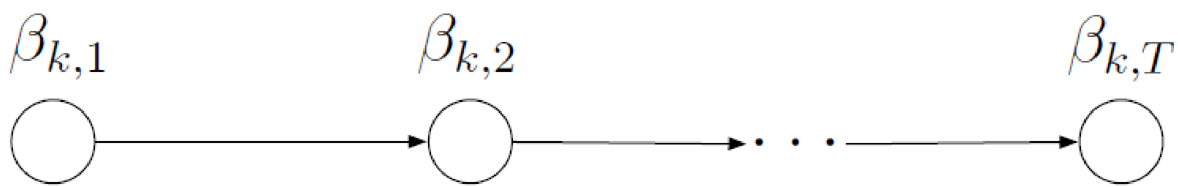
Topic hierarchies



Extension: dynamic topic modeling



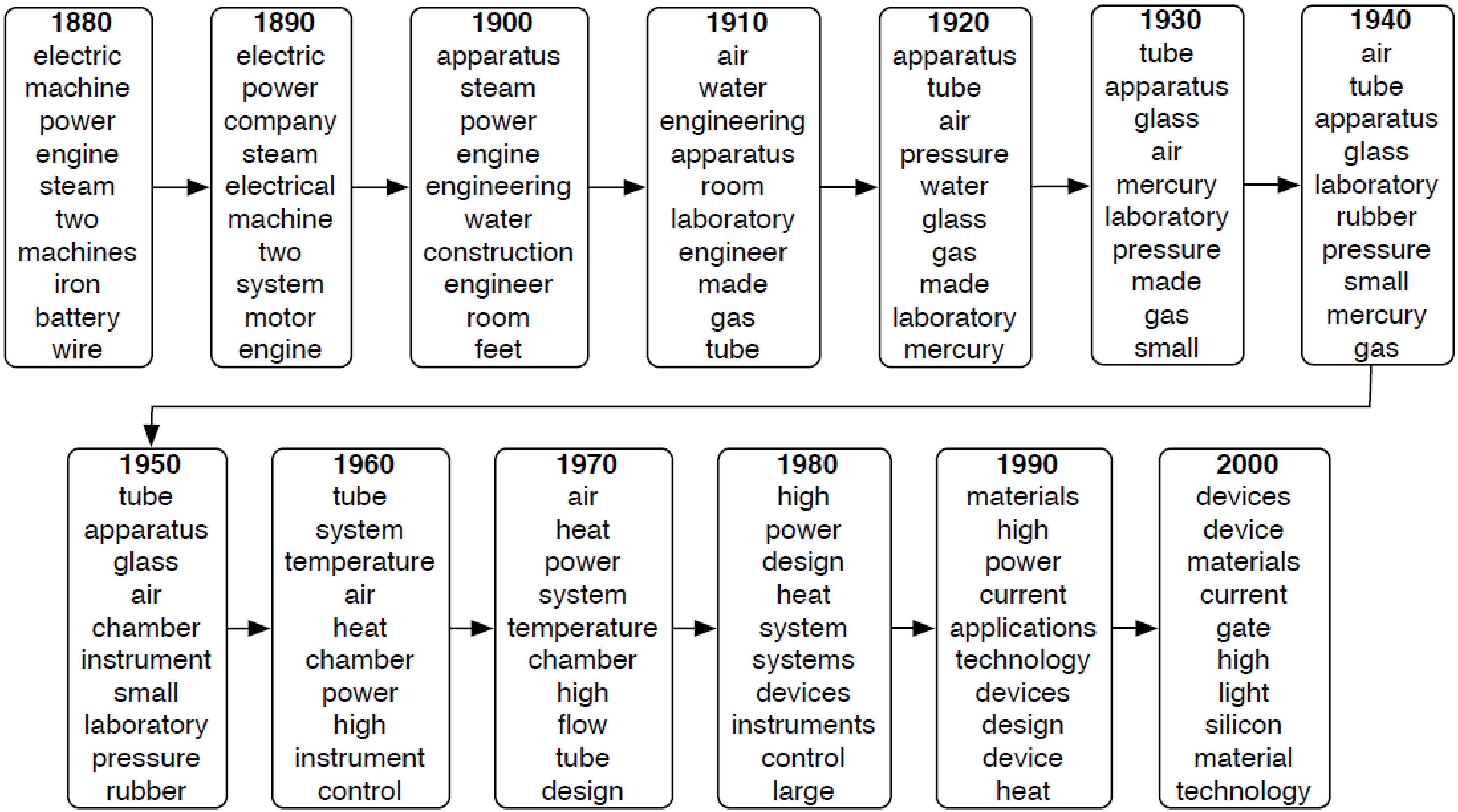
Time-drifting topic distributions



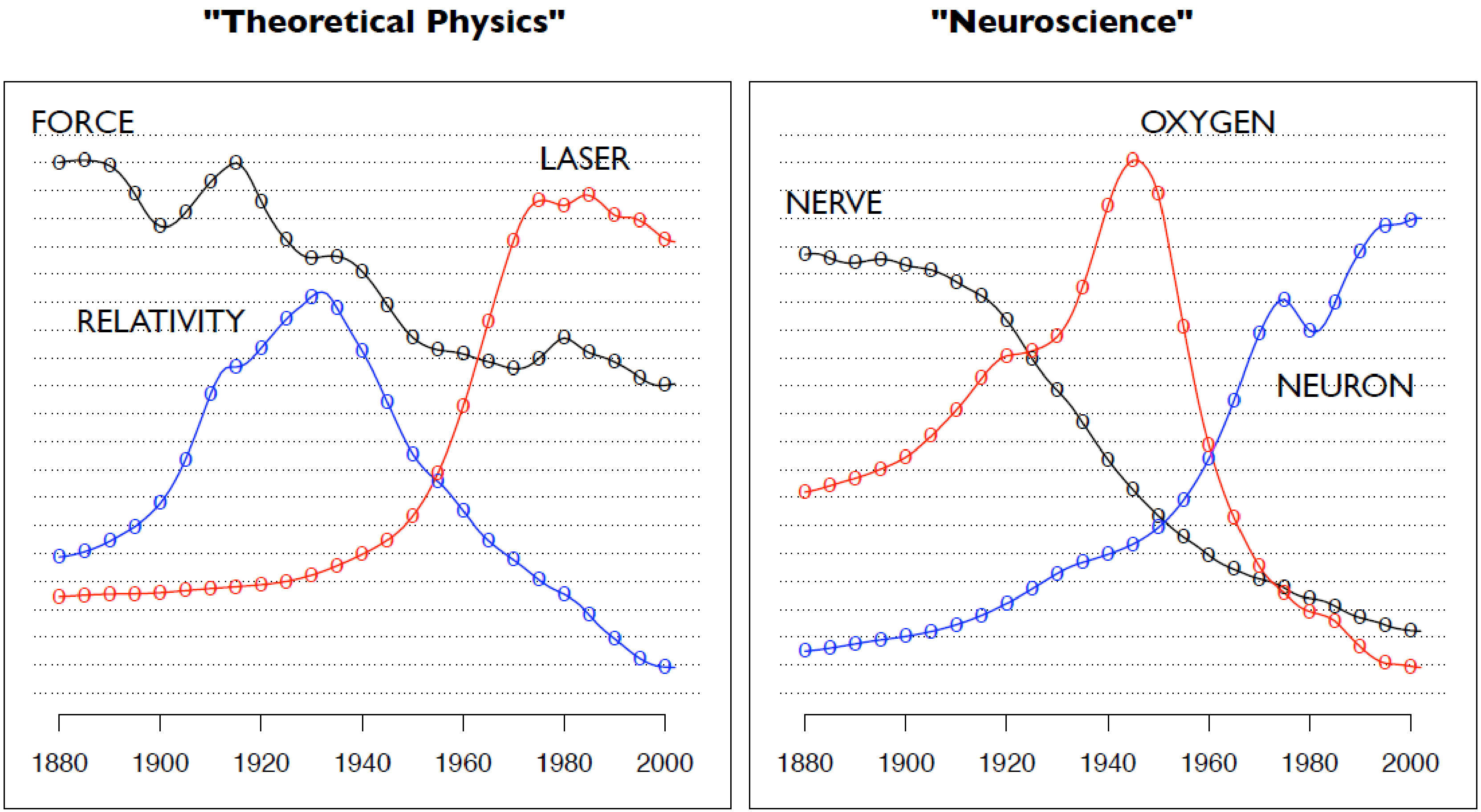
- Use a logistic normal distribution to model topics evolving over time (Aitchison, 1980)
- A state-space model on the natural parameter of the topic multinomial (West and Harrison, 1997)

$$\begin{aligned} \beta_{t,k} \mid \beta_{t-1,k} &\sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2) \\ p(w \mid \beta_{t,k}) &\propto \exp\{\beta_{t,k}\} \end{aligned}$$

Temporal changes



Trends


"Theoretical Physics"
FORCE
LASER
RELATIVITY
1880
1900
1920
1940
1960
1980
2000
"Neuroscience"
OXYGEN
NERVE
NEURON
1880
1900
1920
1940
1960
1980
2000

Other uses

Corr–LDA:

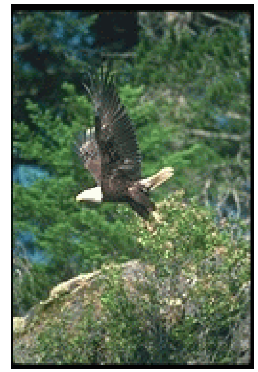
TREE, LIGHT, SUNSET, WATER, SKY

GM–Mixture:

CLOSE–UP, TREE, PEOPLE, MUSHROOMS, LICHEN

GM–LDA:

WATER, SKY, TREE, PEOPLE, GRASS



Corr–LDA:

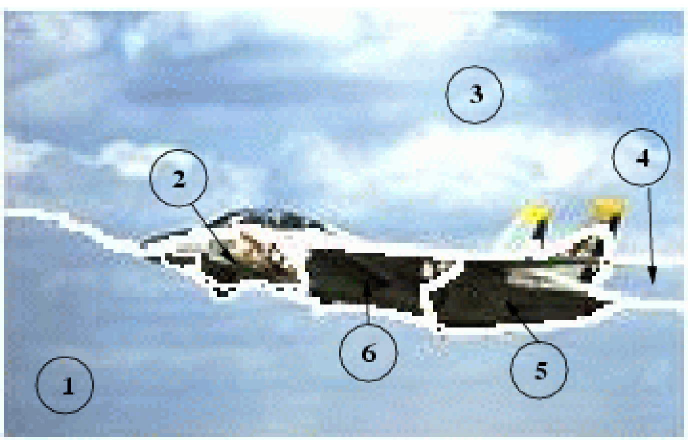
TREE, WATER, GRASS, FLOWERS, BIRDS

GM–Mixture:

TREE, WATER, GRASS, SKY, FIELD

GM–LDA:

WATER, SKY, TREE, PEOPLE, GRASS



Corr–LDA:	**GM–LDA:**
1. PEOPLE, TREE	1. HOTEL, WATER
2. SKY, JET	2. PLANE, JET
3. SKY, CLOUDS	3. TUNDRA, PENGUIN
4. SKY, MOUNTAIN	4. PLANE, JET
5. PLANE, JET	5. WATER, SKY
6. PLANE, JET	6. BOATS, WATER