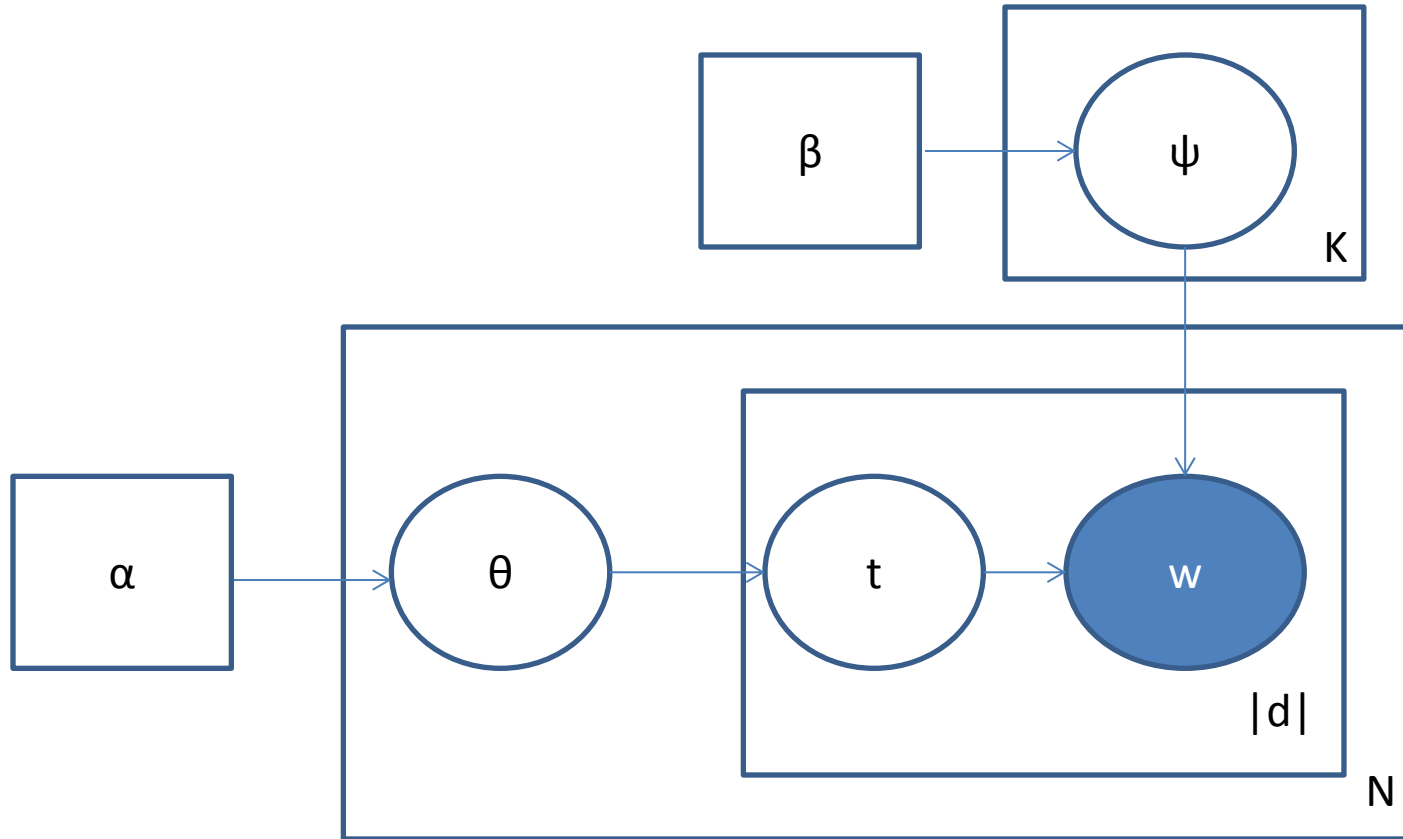


Befriending LDA

Nisheeth

LDA in plate notation



Generative model

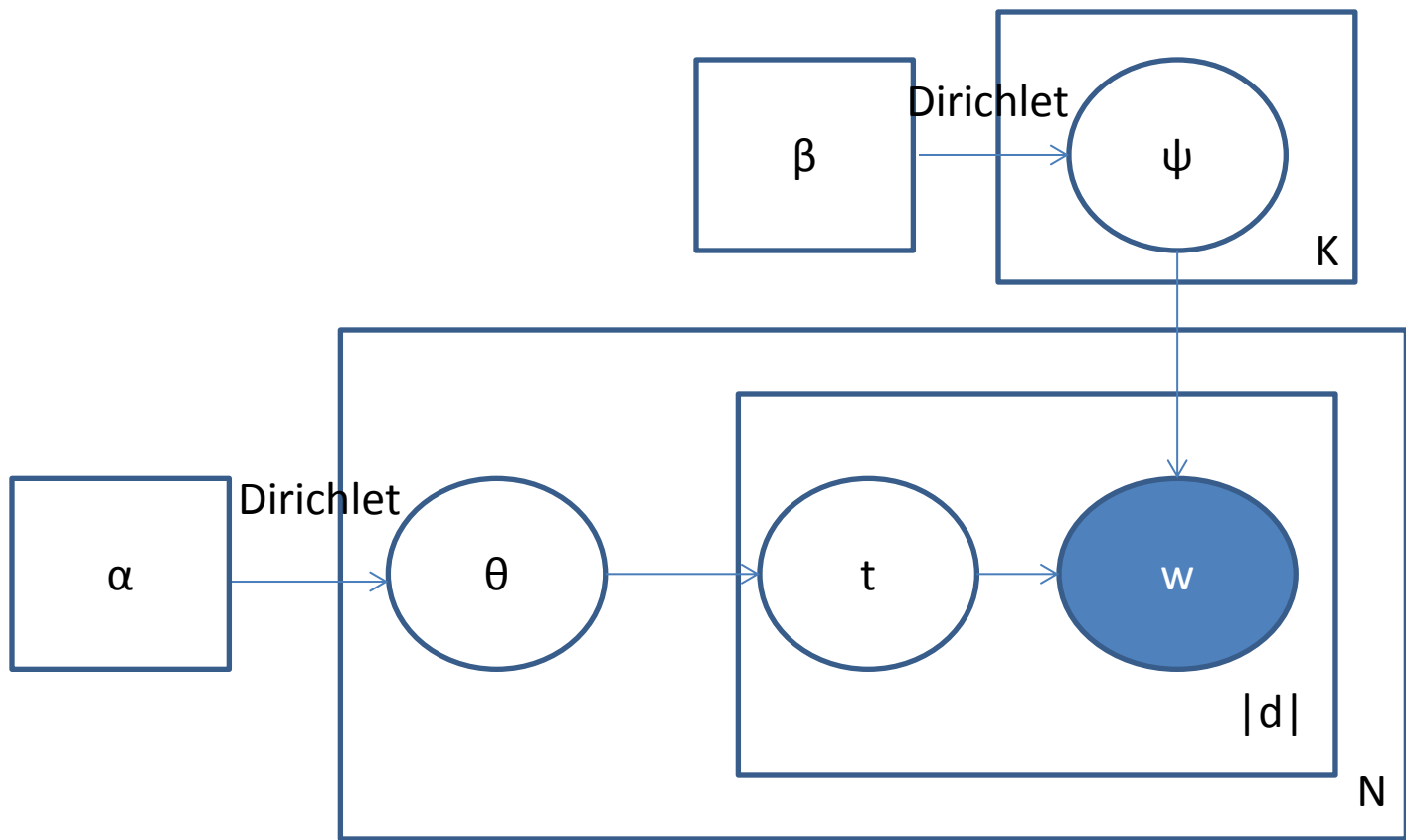
- For all documents
 - Generate $\theta \sim \text{Dirichlet}(\alpha)$
 - Generate all K $\psi \sim \text{Dirichlet}(\beta)$
- For all words in each document
 - Generate $t \sim \text{Multinomial}(\theta)$
 - Generate $w \sim \text{Multinomial}(\psi_t)$

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:

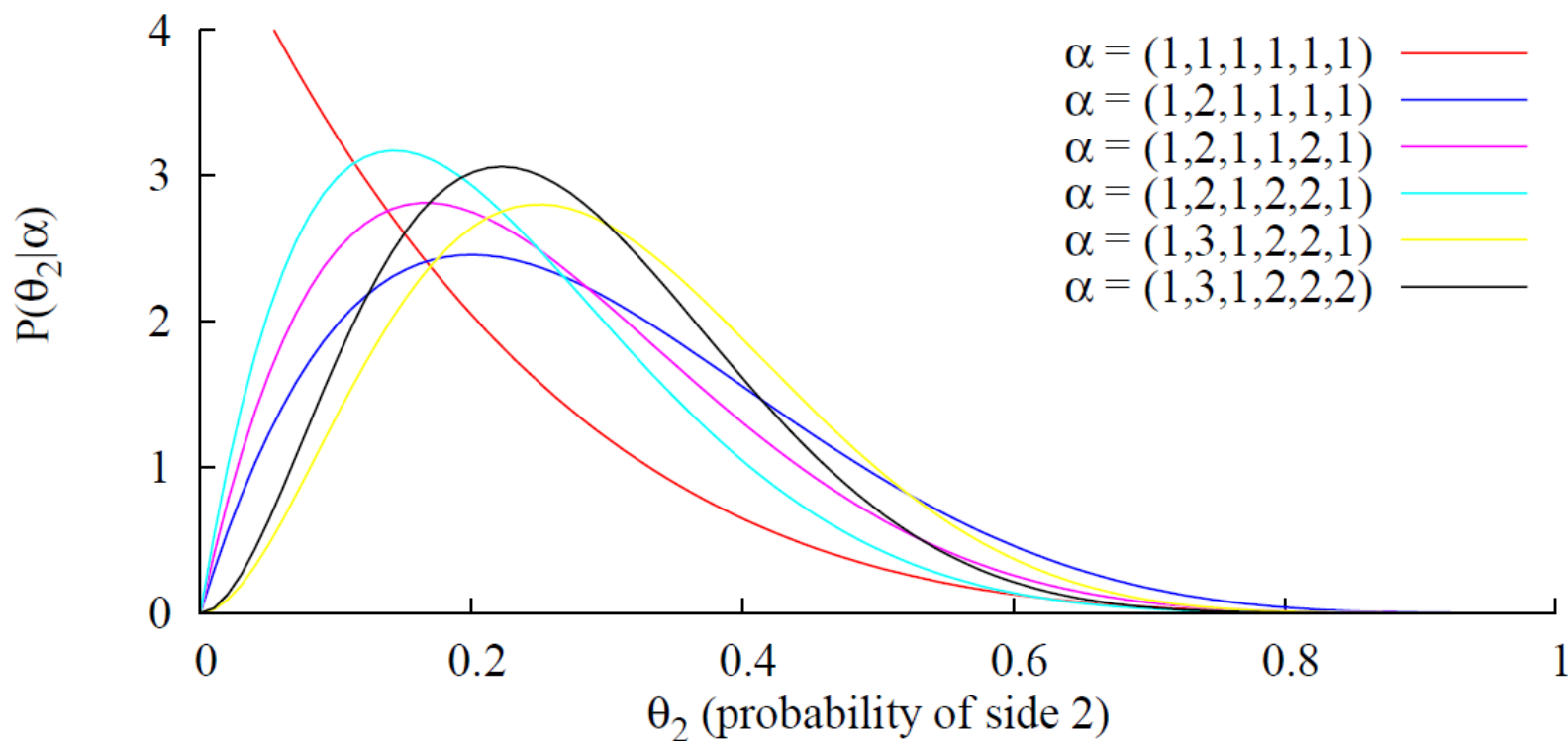
$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \dots \theta_k^{\alpha_k-1}$$

- Easier to understand
 - Prior $\text{Dir}(\alpha_1, \alpha_2)$
 - Likelihood $\text{Multi}(\theta_1, \theta_2)$
 - Outcome $\{n_1, n_2\}$
 - Posterior $\text{Dir}(\alpha_1 + n_1, \alpha_2 + n_2)$
- Ignoring the normalization constant, what is the Dirichlet probability of a multinomial sample $[0.1, 0.5, 0.4]$ with parameter 10
 - $(0.1)^9 (0.5)^9 (0.4)^9 = 5\text{e-}16$
- What would it be for parameter 0.2?
 - 22



Dirichlet update – dice roll

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

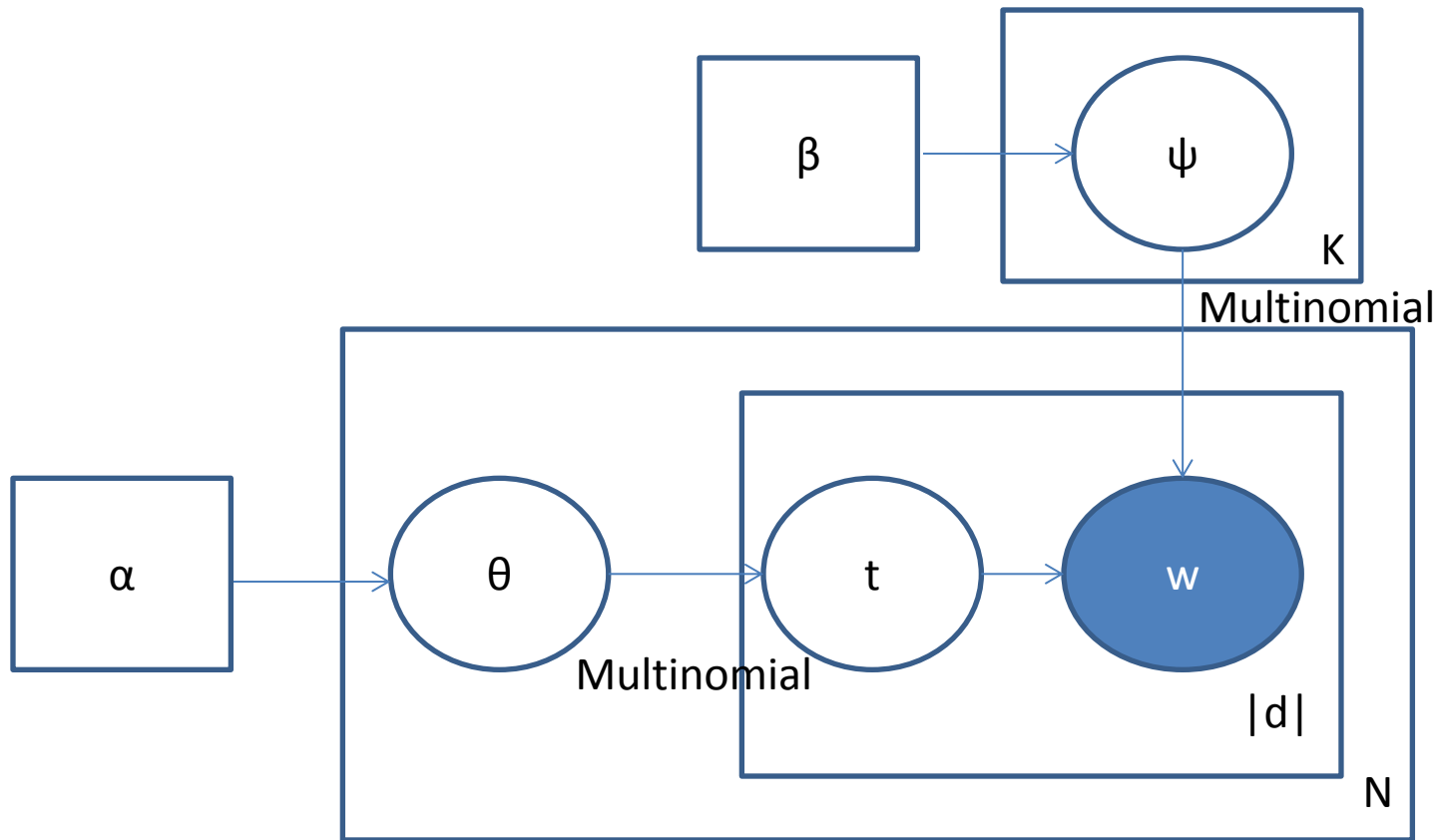


LDA math – the multinomial distribution

- 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

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

- Parameterized by $\boldsymbol{\gamma}$ 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
 - $p(\text{document}) = \frac{4!}{2!2!0!} (0.3)^2 (0.4)^2 = 0.0864$
 - 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

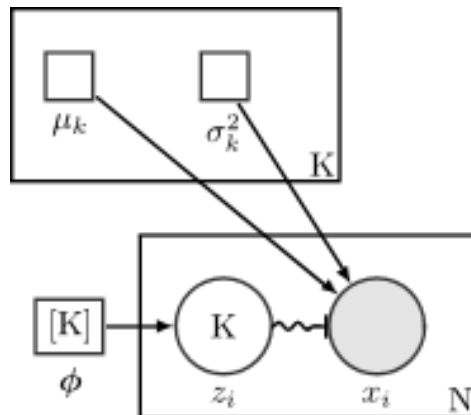
ψ is distribution of words in a topic

$p(t|\theta)$ 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

- $p(K | \phi)$ is high when the ϕ value is high for the K^{th} label
- $p(x | K)$ is high when x is statistically likely to be drawn from the Gaussian with the K th summary statistics



LDA inference

- Latent variable inference

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

- From the graphical model

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

- What are these terms?

$$1. \quad p(\mathbf{w} | \mathbf{t}, \psi) = \prod_{n=1}^{|\mathbf{d}|} \psi_{t_n, w_n}$$

$$2. \quad p(\mathbf{t} | \theta) = \prod_{n=1}^{|\mathbf{d}|} \theta_{t_n}$$

$$3. \quad p(\theta | \alpha) = C(\alpha) \sum_{i=1}^K \theta_i^{\alpha_i - 1}$$

LDA intuition

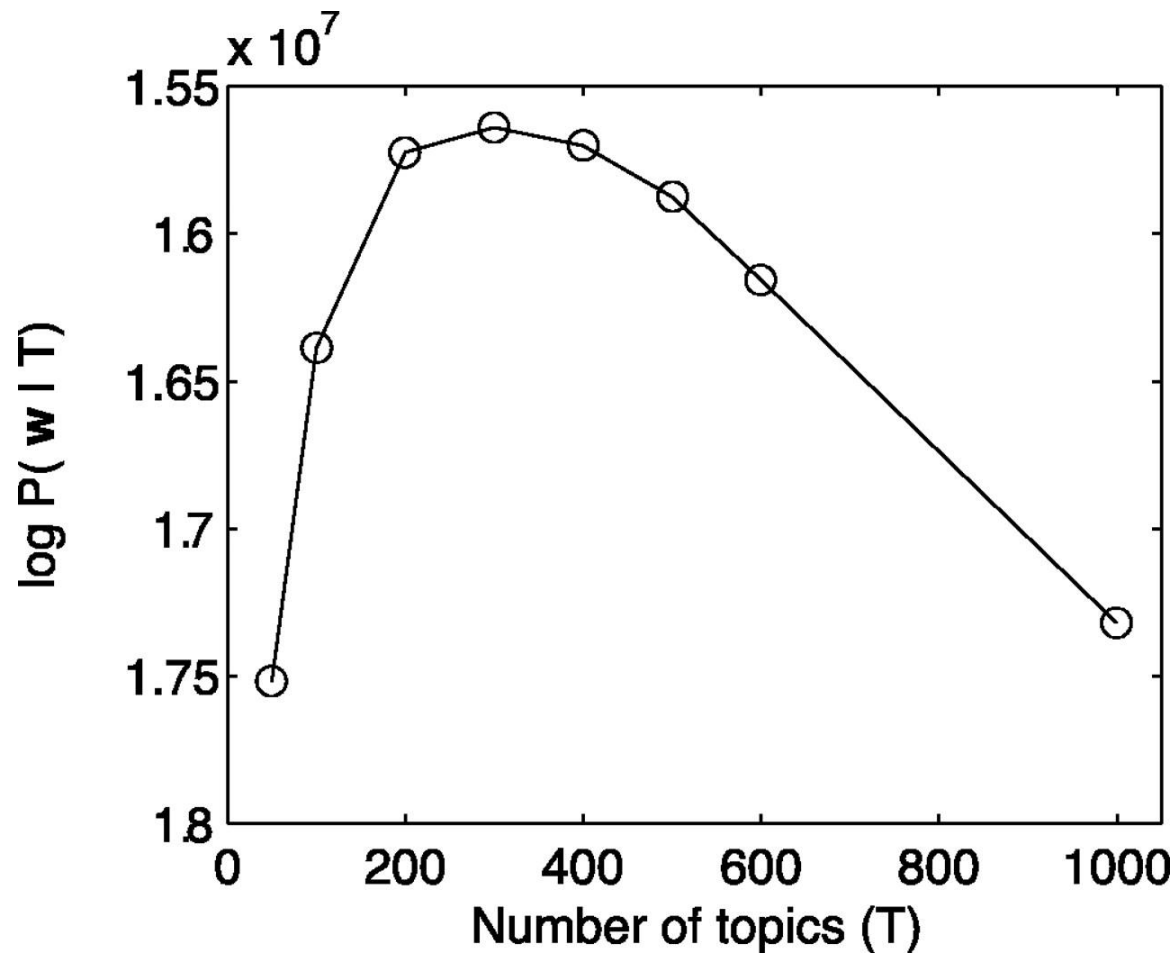
- Given the optimal denominator, the correct partitioning of the data into topics is determined by the numerator
- What does the numerator say about what constitutes a good topic partitioning?
 1. $p(\mathbf{w}|\mathbf{t}, \psi)$ will have high values iff ψ is sparse
 2. $p(\mathbf{t}|\theta)$ will have 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.

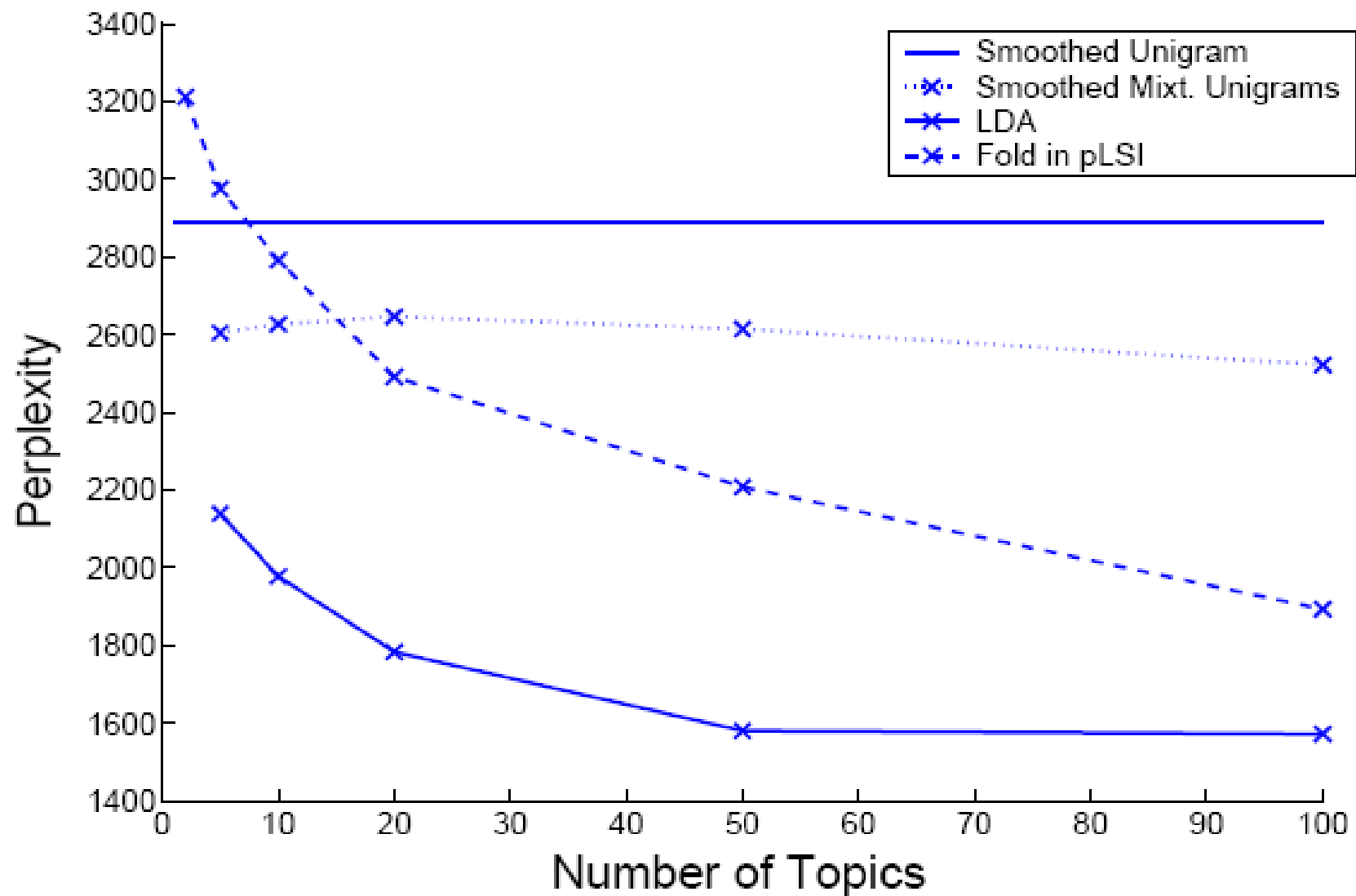
$$perplexity(D_{test}) = \exp \left\{ - \frac{\sum_{d=1}^M \log p(\mathbf{w}_d)}{\sum_{d=1}^M N_d} \right\}$$

Document Modeling – cont.

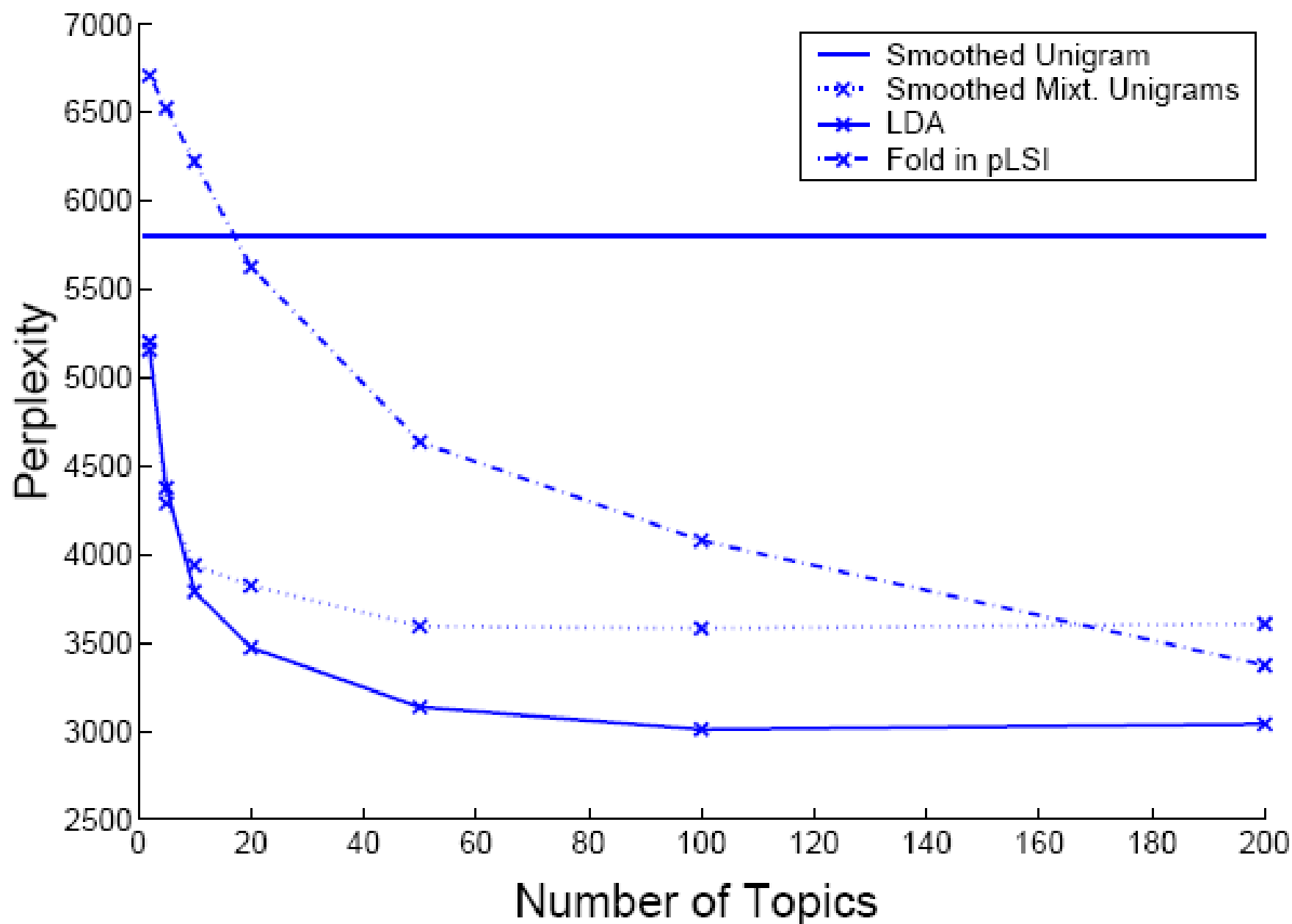
data used

- 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



AP



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

Most likely words from top topics



TECHVIEW: DNA SEQUENCING

Sequencing the Genome, Fast

James C. Mullikin and Amanda A. McMurtry

Genome sequencing projects reveal the genetic makeup of an organism by reading off the sequence of the DNA bases, which encodes all of the information necessary for the life of the organism. The base sequence contains four nucleotides—adenine, thymine, guanine, and cytosine—which are linked together into long double-helical chains. Over the last two decades, automated DNA sequencers have made the process of obtaining the base-by-base sequence of DNA easier. By application of an electric field across a gel matrix, these sequencers separate fluorescently labeled DNA molecules that differ in size by one base. As the molecules move past a given point in the gel, laser excitation of a fluorescent dye specific to the base at the end of the molecule yields a base-specific signal that can be automatically recorded.

The latest sequencer to be launched is Perkin-Elmer's much-anticipated ABI Prism 3700 DNA Analyzer which, like the Molecular Dynamics MegabACE 1000 launched last year, incorporates a capillary tube to hold the sequence gel rather than a traditional slab-shaped gel apparatus. Extra interest in the ABI 3700 has been generated because Craig Venter of Celera Genomics Corporation anticipates that ~230 of these machines (1) will enable the company to produce raw sequence for the entire 3 gigabases (Gb) of the human genome in 3 years. The specifications of the ABI 3700 machine say that, with less than 1 hour of human labor per day, it can sequence 768 samples per day. Assuming that each sample gives an average of 400 base pairs (bp) of usable sequence data (its read length) and any section from the entire human genome is covered by an average of 10 overlapping independent reads (2), the 75 million samples that Celera must process will require ~100,000 ABI 3700 machine days. With ~230 machines, that works out to less than 2 years or about 434 days, which affords some margin of error for unexpected developments.

At the Sanger Centre, we have finished 146 Mb of genomic sequence from a vari-

ety of genomes, including 81 Mb of sequence from the human genome, the largest amount of any center so far (3). We are aiming to sequence 1 Gb of human sequence in rough-draft form by 2001, with a finished version by 2003. Our sequencing equipment includes 44 ABI 373XL 61 ABI 377XL, and 31 ABI 377XL-96 slab gel sequencers from Perkin-Elmer plus 6 Molecular Dynamics MegabACE 1000 capillary sequencers, allowing a maximum throughput of 32,000 samples per day. Two ABI 3700 capillary sequencers—delivered

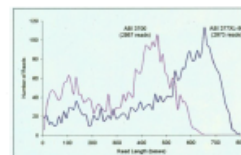


Fig. 1. Comparison of read-length histograms for sequences collected with the ABI 3700 capillary machine and the ABI 377XL-96 slab gel machine. The capillary machine underperforms the slab gel machine by about 200 bases. Both sets of reads are from runs with ABI Big Dye Terminator chemistry. Read length 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 recalculated for each type of read.

to the Sanger Centre in December 1998—are in our Research and Development department for evaluation. Thus, the ABI 3700 will ultimately be added to our present capacity to reach our goal.

The ABI 3700 DNA sequencer is built into a floor-standing cabinet, which contains in its base all the reagents required for its operation. The reagent containers are readily accessible for replenishment, which is required every day under high-throughput operation. At bench height within the cabinet is a four-position bed, on which microtiter plates of DNA samples are located. The operator places the prepared plates into position, closes the front of the machine and programs it by using a personal computer. A robotic arm transfers DNA sam-

ples from the plates into wells that open into the capillaries. This and the rest of the sequencing operation is fully automatic. The machine can currently process four 96-well plates of DNA samples unattended, taking approximately 16 hours before operator intervention is required. This rate falls short of the design specification of four 96-well plates in 12 hours.

The main innovation of the ABI 3700 is the use of a sheath flow fluorescence detection system (4). Detection of the DNA fragments occurs 300 μ m past the end of the capillary within a fused silica cuvette. A laminar fluid flows over the ends of the capillaries, drawing the DNA fragments as they emerge from the capillaries through a fused laser beam that simultaneously intersects with all of the samples. The emitted fluorescence is detected with a spectral CCD (charge-coupled device) detector. This arrangement means that there are no moving parts in the detection system, other than a shutter in front of the CCD detector.

We have evaluated these machines for their performance, operation, ease of use, and reliability in comparison to the more commonly used slab gel sequencing machines. In automated sequencers, there are two methods for containing the gel matrix. One is to polymerize a gel matrix between two finely separated glass plates (0.4 mm or less)—the slab gel method. The other is to inject a polymer matrix into a capillary (internal diameter <0.2 mm). Most sequencing facilities use the slab gel method, because multicapillary sequencers have only recently become commercially available.

With either type of system, the aim is to read as many bases as possible for a given sample of DNA—that is, long read lengths are desirable. In fact, a system that could read twice as many bases but at half the speed of another system is preferable, if both systems cost the same. This is because assembling relatively fewer long-sequenced fragments is easier than assembling many short ones. So, read length is an important parameter when evaluating new sequencing technologies.

We have directly compared the ABI 3700 sequencer to the ABI 377XL slab gel sequencer by evaluating the sequence data obtained from both machines with human DNA samples. These samples were subcloned into plasmid or m13 phage and prepared and sequenced with our standard protocols for Perkin-Elmer Big Dye Terminator chemistry.

sequence
genome
genes
sequences
human
gene
dna
sequencing
chromosome
regions
analysis
data
genomic
number

devices
device
materials
current
high
gate
light
silicon
material
technology
electrical
fiber
power
based

data
information
network
web
computer
language
networks
time
software
system
words
algorithm
number
internet

The authors are at The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs, CB9 0TA, UK. E-mail: jcm@sanger.ac.uk

Document similarity

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

**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

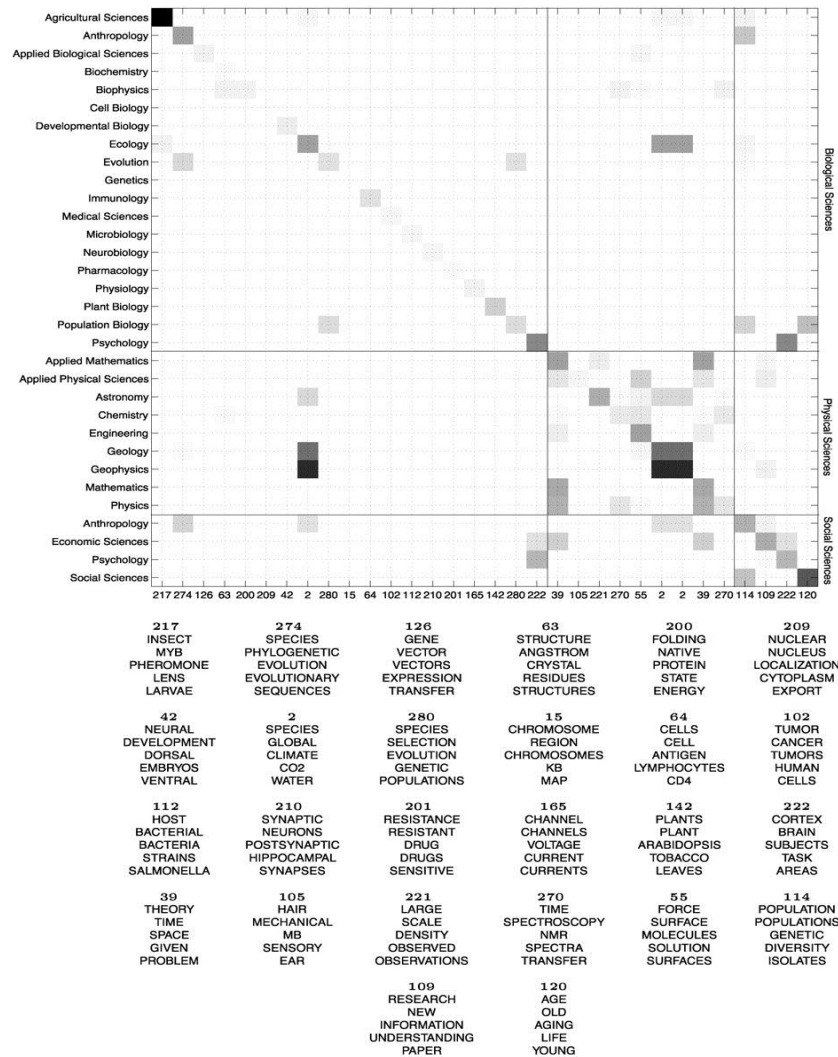
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure

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



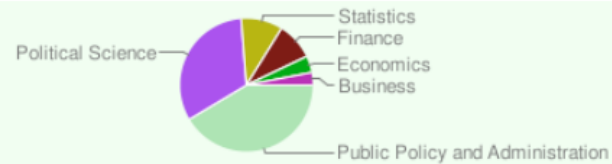
Document tagging, relevance scoring

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**

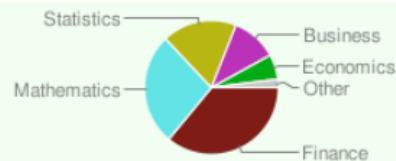


Chaos and Nonlinear Forecastability in Economics and Finance

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

Journal Disciplines:

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

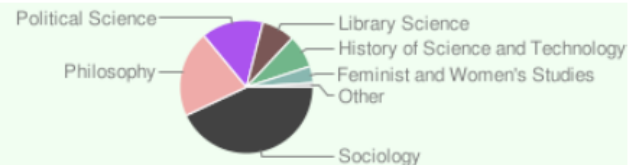


Reply: Theory Is Not a Social Dilemma

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

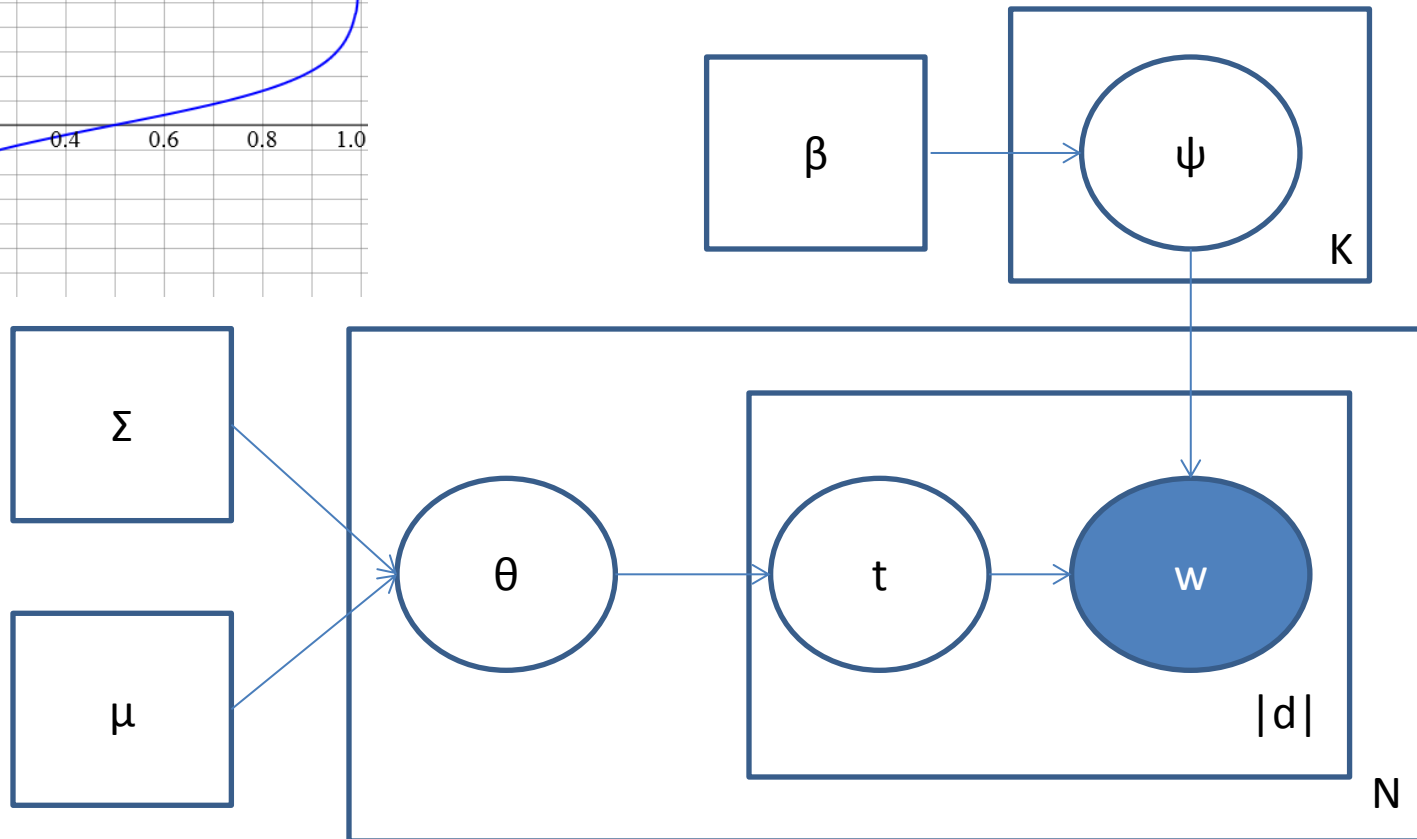
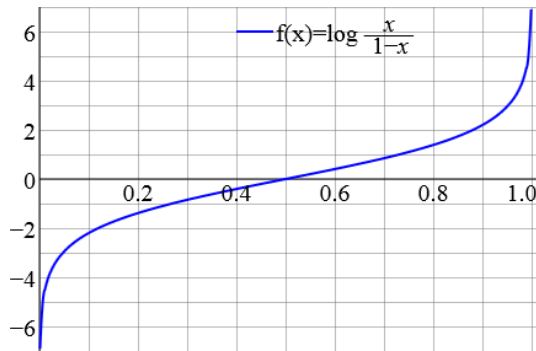
Journal Disciplines:

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



Extension: correlated topic models

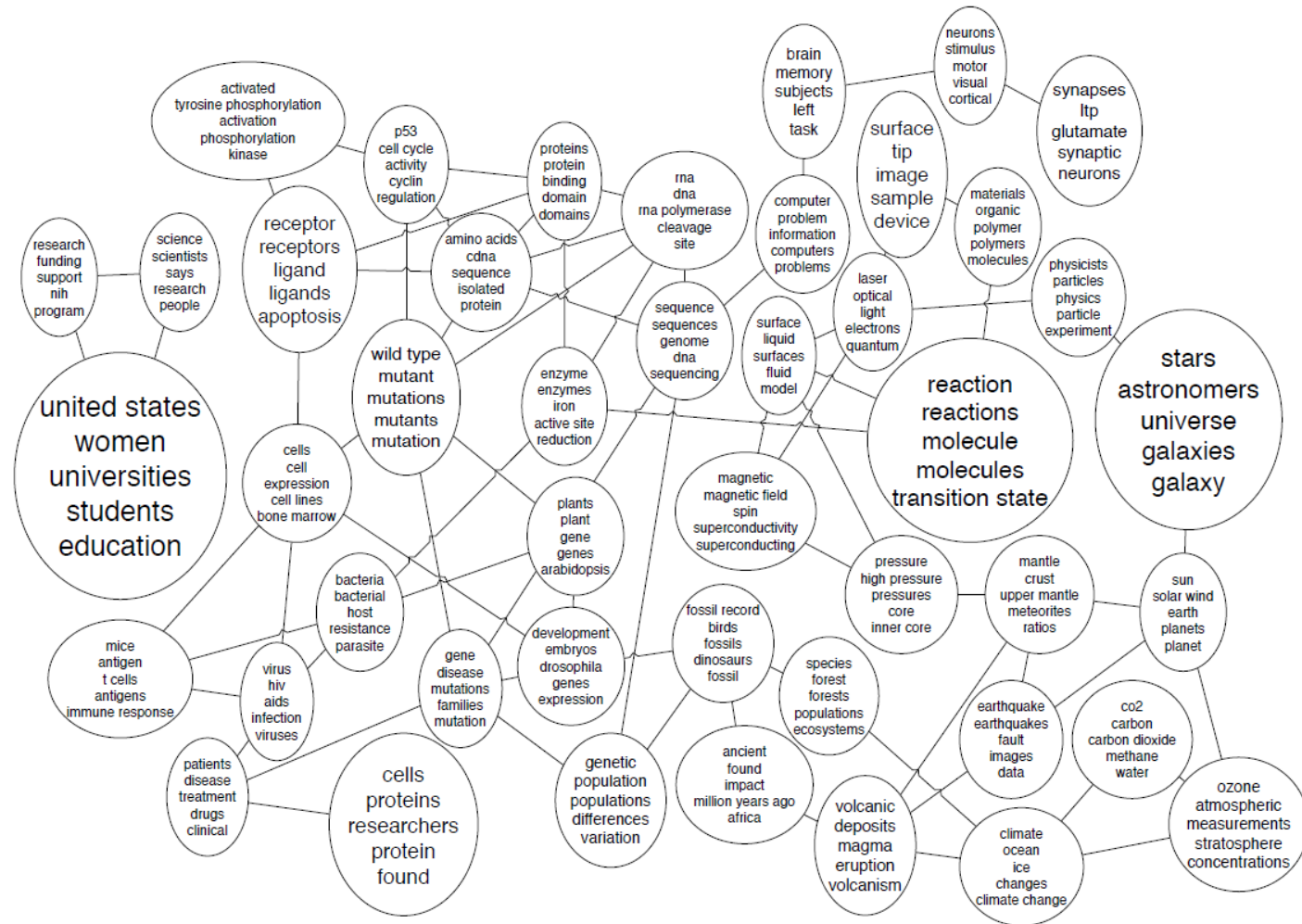
Logit plot



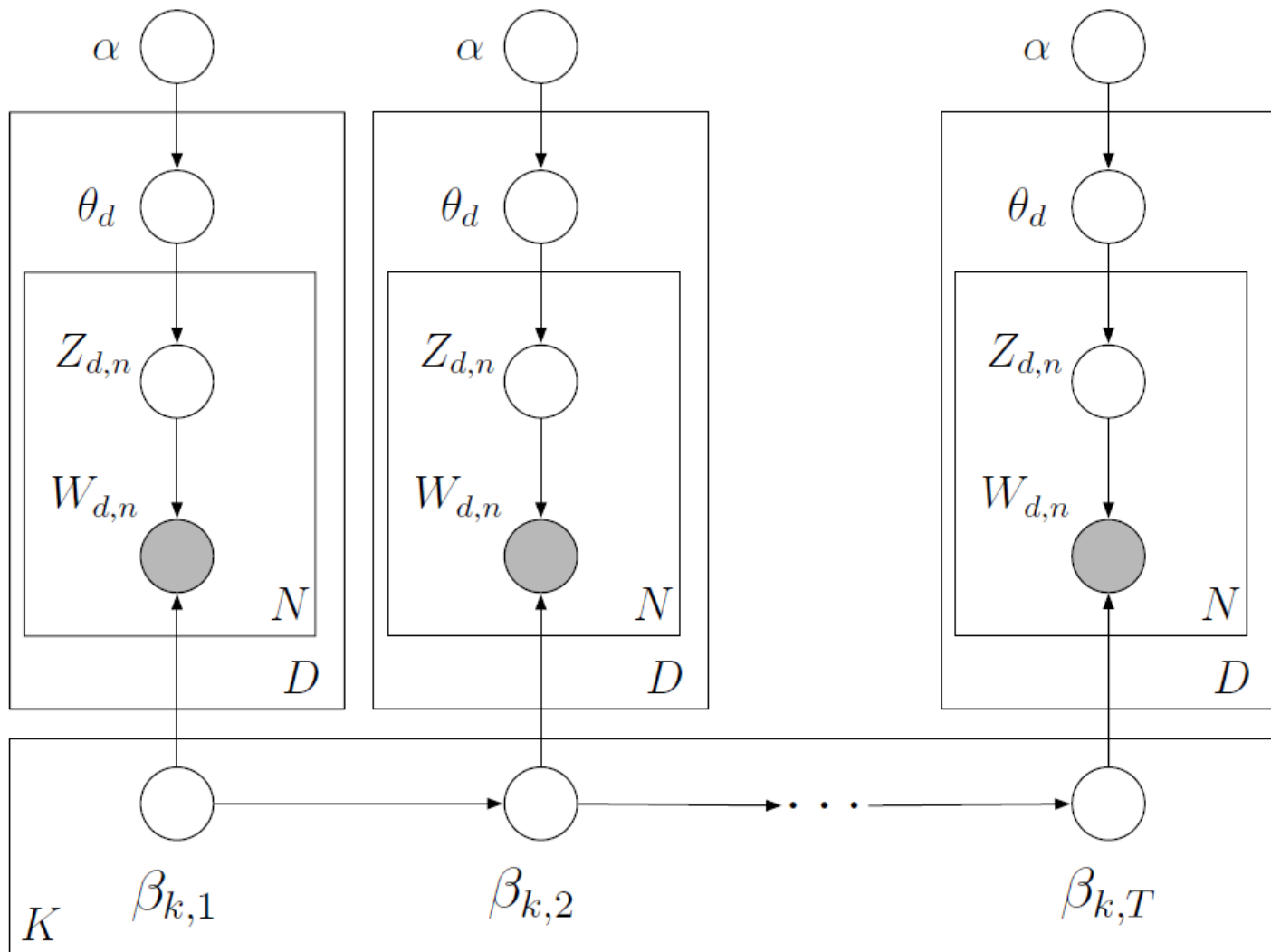
$$\mathbf{x} \sim N(\mu, \Sigma)$$

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

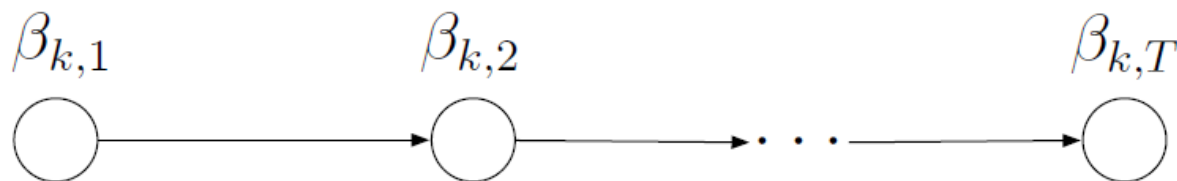
Topic hierarchies



Extension: dynamic topic modeling



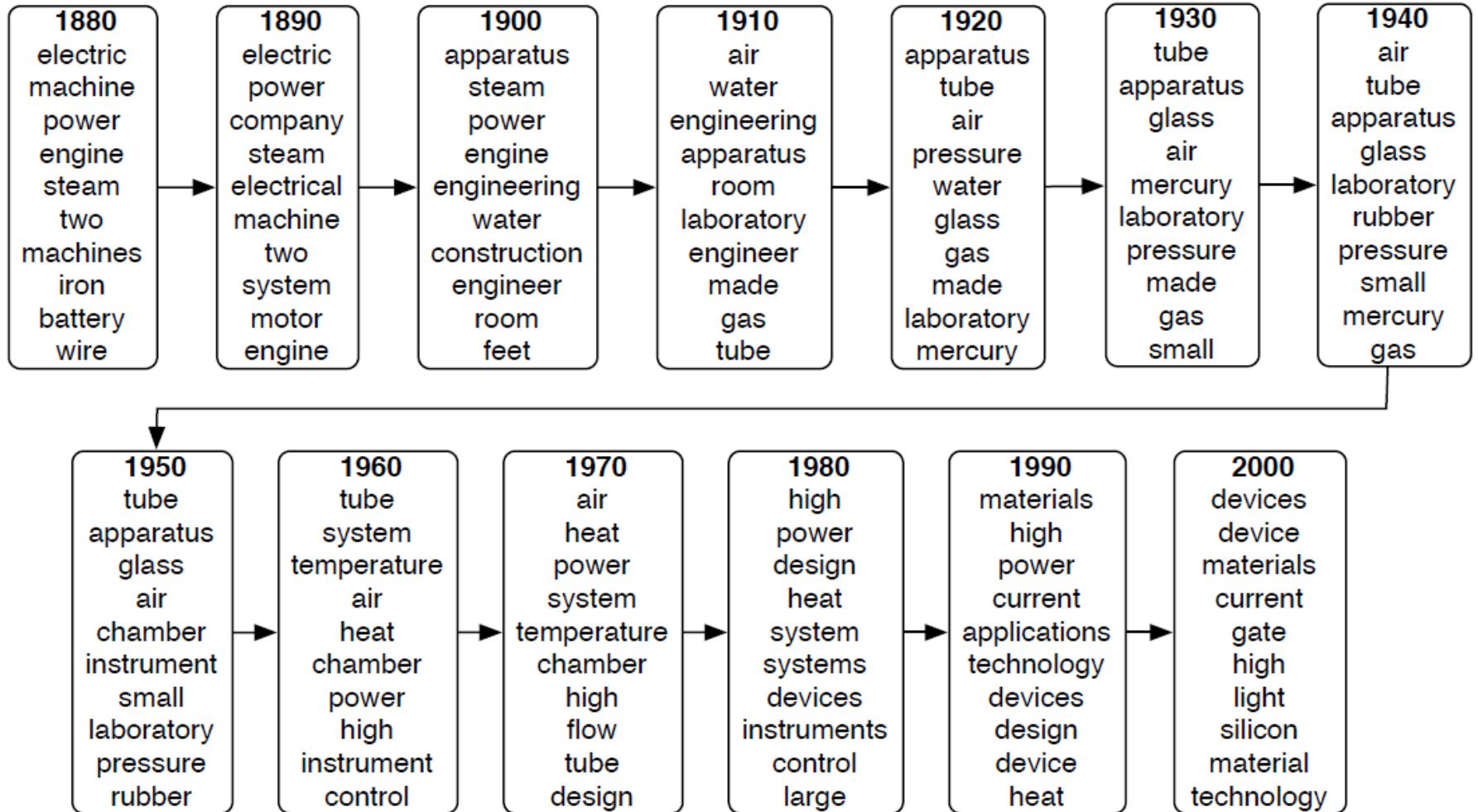
Time-drifting topic distributions



- 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)

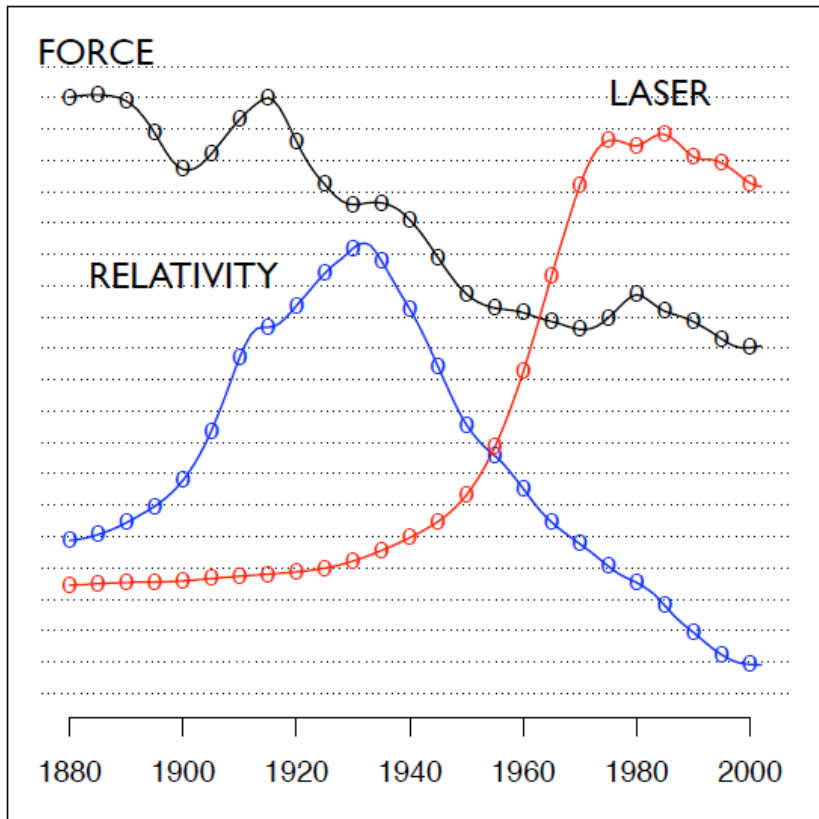
$$\begin{aligned}\beta_{t,k} \mid \beta_{t-1,k} &\sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2) \\ p(w \mid \beta_{t,k}) &\propto \exp\{\beta_{t,k}\}\end{aligned}$$

Temporal changes

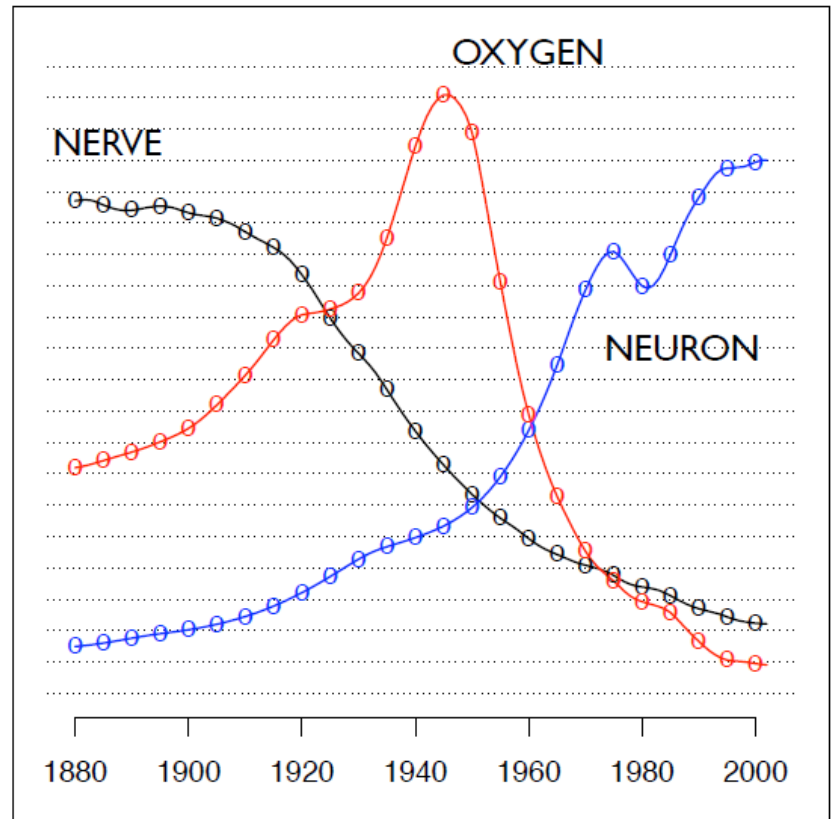


Trends

"Theoretical Physics"



"Neuroscience"



Other uses



Corr-LDA:

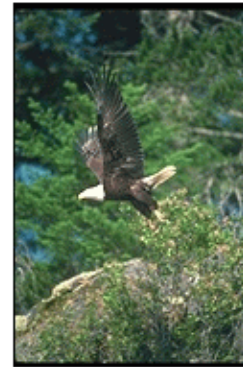
TREE, LIGHT, SUNSET, WATER, SKY

GM-Mixture:

CLOSE-UP, TREE, PEOPLE, MUSHROOMS, LICHEN

GM-LDA:

WATER, SKY, TREE, PEOPLE, GRASS



Corr-LDA:

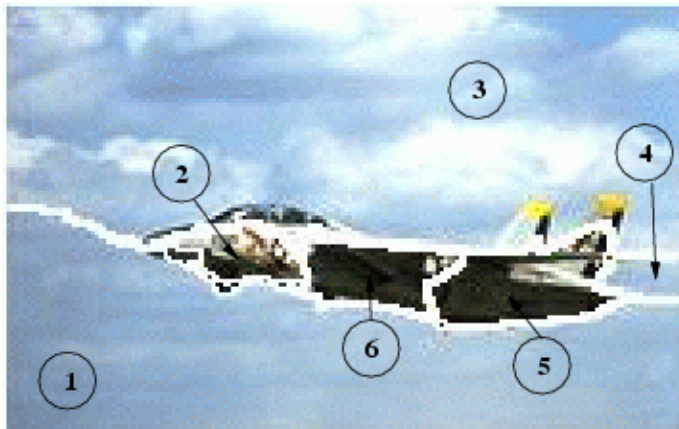
TREE, WATER, GRASS, FLOWERS, BIRDS

GM-Mixture:

TREE, WATER, GRASS, SKY, FIELD

GM-LDA:

WATER, SKY, TREE, PEOPLE, GRASS



Corr-LDA:

1. PEOPLE, TREE

2. SKY, JET

3. SKY, CLOUDS

4. SKY, MOUNTAIN

5. PLANE, JET

6. PLANE, JET

GM-LDA:

1. HOTEL, WATER

2. PLANE, JET

3. TUNDRA, PENGUIN

4. PLANE, JET

5. WATER, SKY

6. BOATS, WATER