S&DS 265 / 365

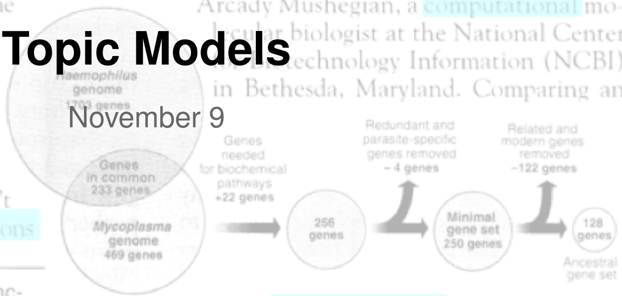
Introductory Machine Learning

Topic Models

November 9

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustai[illegible] just 250 genes, and that t[illegible] required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and [illegible]nced. "It may be a way of organizing [illegible]nced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for [illegible]echnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an esti-mate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Yale

Checkpoint

- Assignment 5 (Embeddings) due Thursday
- Assignment 6 (Topic models) posted Thursday
- Quiz 3 next Tuesday: LMs, embeddings, Bayes, topic models

Probabilistic modeling

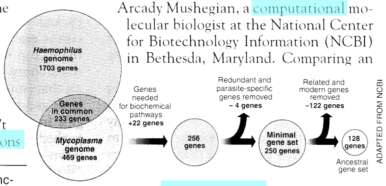
1. Data are assumed to be observed from a generative probabilistic process that includes hidden variables.
   - *In text, the hidden variables are the thematic structure.*
2. Infer the hidden structure using posterior inference
   - *What are the topics that describe this collection?*
3. Situate new data into the estimated model.
   - *How does a new document fit into the topic structure?*

Latent Dirichlet allocation (LDA)

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

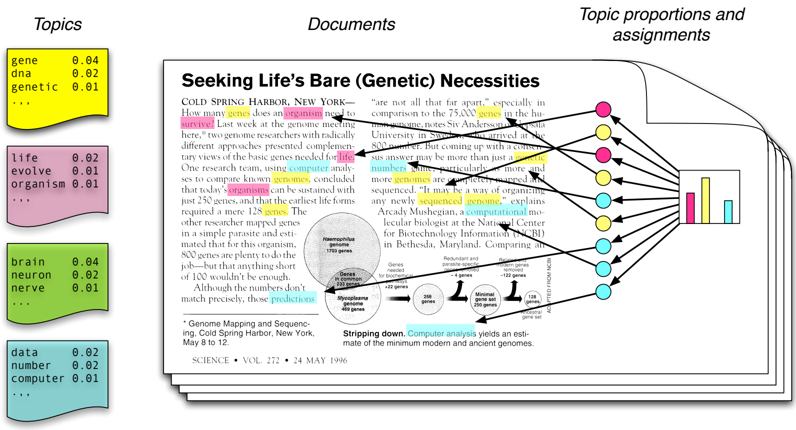


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

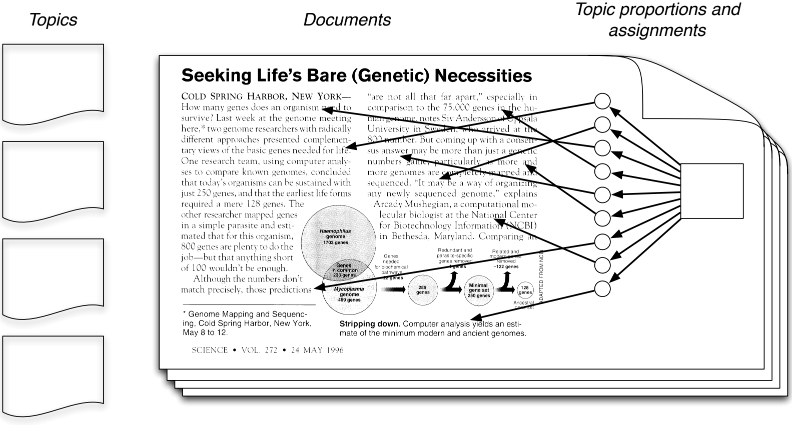
Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



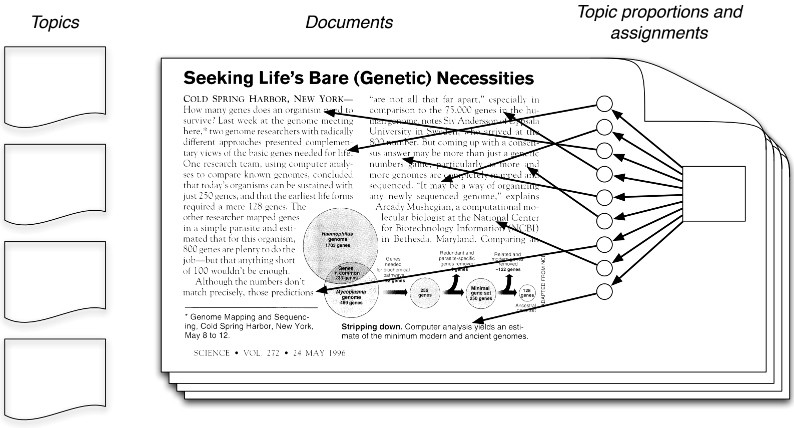
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



- In reality, we only observe the documents
- The other structure are **hidden variables**

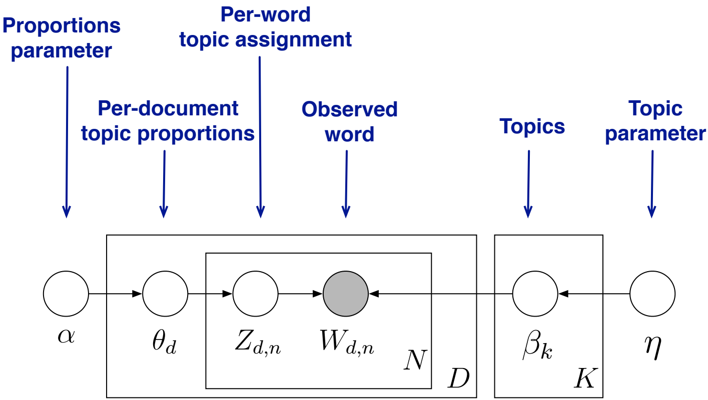
The posterior distribution



- Our goal is to **infer** the hidden variables
- I.e., compute their distribution conditioned on the documents

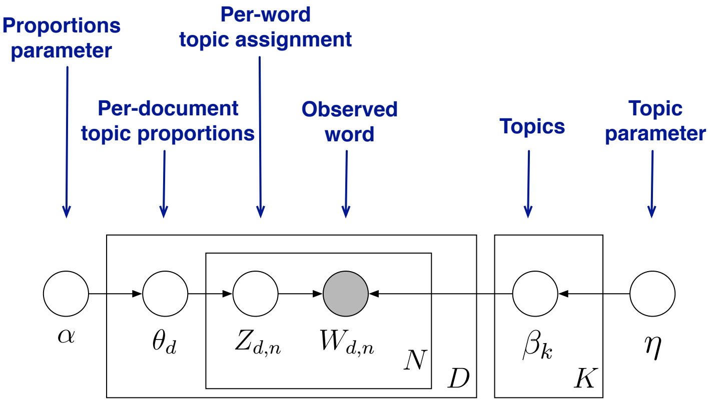
$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

LDA as a graphical model



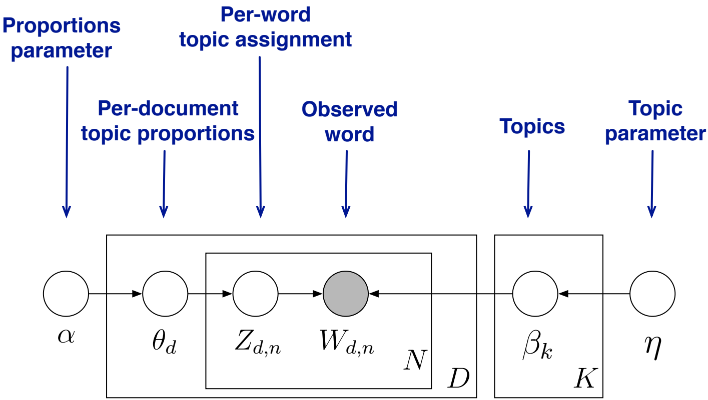
- Encodes our assumptions about the data
- Connects to algorithms for computing with data
- See *Pattern Recognition and Machine Learning* (Bishop, 2006).

LDA as a graphical model



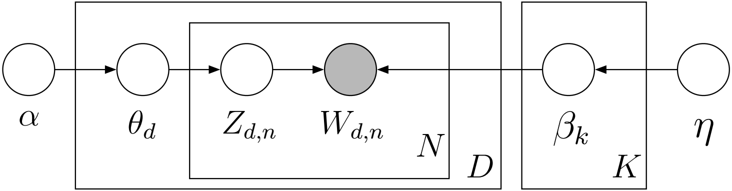
- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

LDA as a graphical model



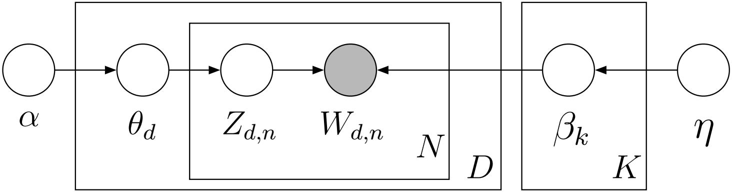
$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

LDA



- This joint defines a posterior.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, exploration, ...

Example inference



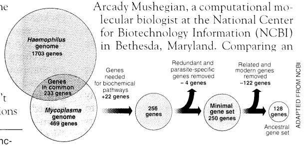
- **Data**: The OCR'ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

Example inference

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

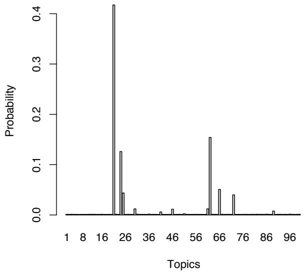
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996



Example inference

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

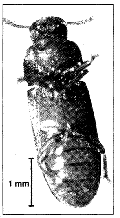
Example inference (II)

Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (*1*) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al*. (*2*) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (*3*).



Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

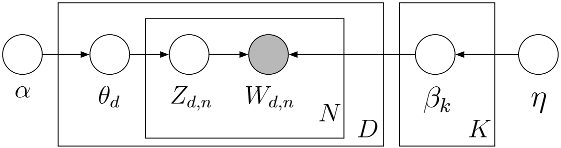
An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

Example inference (II)

problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

Posterior inference for LDA



- There is a large literature on approximating the posterior.
- We will focus on
  - Gibbs sampling
  - Mean-field variational methods (batch and online)


JOSIAH WILLARD GIBBS
BORN FEB. 11, 1839 — DIED APR. 28, 1903

Markov chain Monte Carlo

- Construct a **Markov chain** on the hidden variables, whose limiting distribution is the posterior.
- Collect **independent samples** from that distribution; approximate the posterior with them
- In **Gibbs sampling** the chain is defined by the conditional distribution of each hidden variable given observations and the current setting of the other hidden variables.

Toy example: 3 topics, 3 docs

w	z	w	z	w	z
meth	2	drug	3	inning	1
father	1	baseball	2	mother	3
divorce	3	hit	1	son	1
drug	1	inning	2	hit	2
illegal	1	steroids	1	baseball	3

Toy example: 3 topics, 3 docs

w	z	w	z	w	z
meth	1	drug	1	inning	1
father	2	baseball	2	mother	1
divorce	2	hit	2	son	2
drug	1	inning	3	hit	3
illegal	3	steroids	1	baseball	2

Toy example: 3 topics, 3 docs

w	z	w	z	w	z
meth	1	drug	1	inning	2
father	3	baseball	2	mother	3
divorce	3	hit	2	son	3
drug	1	inning	2	hit	2
illegal	1	steroids	1	baseball	2

Extensions

Modeling richer assumptions

- Correlated topic model
- Dynamic topic model

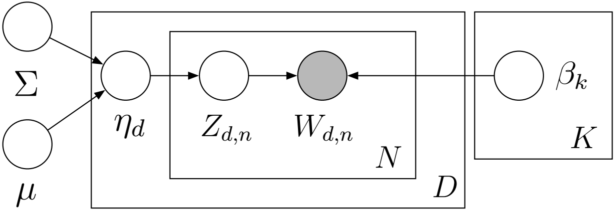
Shortcoming of the Dirichlet

- Dirichlet for topic proportions:

$$p(\theta \mid \alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

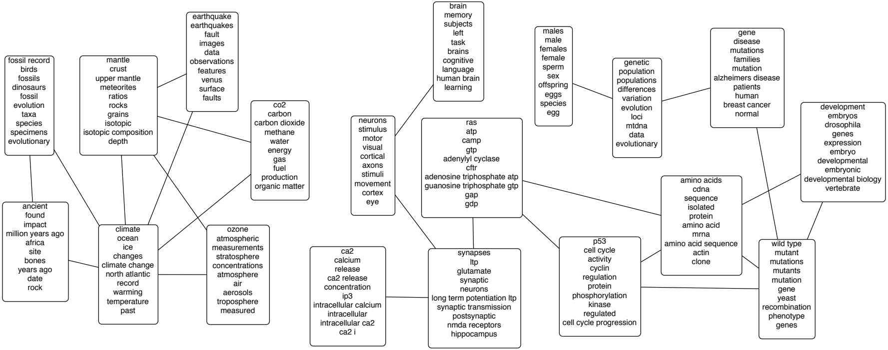
- Near independence of components makes it an unrealistic model
  - An article about *fossil fuels* is more likely to also be about *geology* than about *genetics*

Correlated topic model

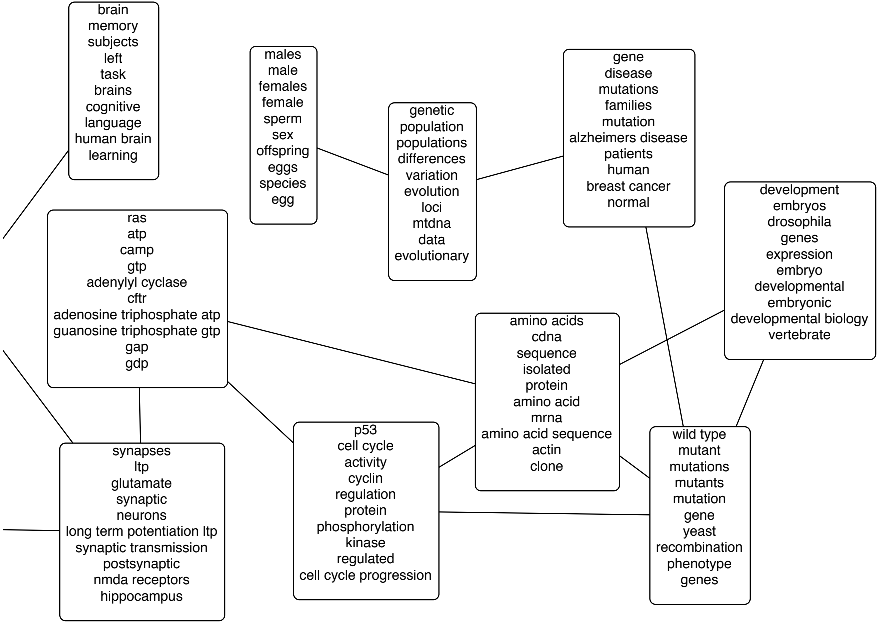


- Draw topic proportions from a logistic normal.
- Useful for:
  - providing a "map" of topics and how they are related;
  - better prediction via correlated topics.
- Sacrifice conjugacy: Posterior over θ does not have same form

Topic graphs



Topic graphs



Modeling Evolution of Topics

- In LDA, document order doesn't matter
- The topics should *evolve* over time
  - "Cleaning Birds" (1883)
  - "Interspecific Brood Parasitism in Blackbirds (Icterinae): A Phylogenetic Perspective" (1992)
- Many document collections have such dynamics: emails, query logs, news articles, etc.

Science 1893 $\Rightarrow$ *Science* 1976

Administration	Limnology	Astronomy	Psychology
association	water	observatory	mind
meeting	lake	observations	nature
american	sea	stars	say
committee	waters	time	science
congress	lakes	made	psychology
members	gulf	astronomical	work
held	great	comet	knowledge
international	depth	star	truth
meetings	river	observed	religion
section	stream	telescope	human

Administration	Limnology	Astronomy	Psychology
house	water	mass	human
congress	concentrations	radio	attempts
science	mercury	objects	theory
bill	fish	astronomy	learning
nsf	samples	xray	ideas
president	soil	stars	new
budget	lake	astronomical	memory
office	ppm	sources	psychology
committee	concentration	observations	behavior
new	waters	observatory	complex

Time series topic model

1. Allow topics evolve between time slices

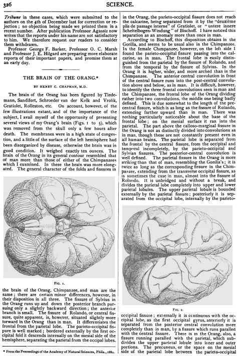
3. For each document in the current time slice:

   a. Select a distribution over topics;

   b. Generate the words from the resulting topic mixture.

Time-corrected document similarity

The Brain of the Orang (1880)



326 SCIENCE.

Tribune in these cases, which were submitted to the authors on the 4th of December last for correction or rejection; no objection being made we printed them in a recent number. After publication Professor Agassiz now writes that the reports under his name are not satisfactory to him. We therefore request our readers to consider them withdrawn.

Professor George F. Barker, Professor O. C. Marsh and Professor J. E. Hilgard are preparing more elaborate reports of their important papers, and promise them at an early day.

THE BRAIN OF THE ORANG.*

BY HENRY C. CHAPMAN, M.D.

The brain of the Orang has been figured by Tiedemann, Sandifort, Schroeder van der Kolk and Vrolik, Gratiolet, Rolleston, etc. On account, however, of the few illustrations extant, and of the importance of the subject, I avail myself of the opportunity of presenting several views of my Orang's brain (Figs. 1 to 5), which was removed from the skull only a few hours after death. The membranes were in a high state of congestion, and a little of the surface of the left hemisphere had been disorganized by disease, otherwise the brain was in good condition. It weighed exactly ten ounces. The brain of the Orang in its general contour resembled that of man more than those of either of the Chimpanzees which I examined. In these the brain was more elongated. The general character of the folds and fissures in

FIG. 1.

the brain of the Orang, Chimpanzee, and man are the same; there are certain minor differences, however, in their disposition in all three. The fissure of Sylvius in the Orang runs up and down the posterior branch pursuing only a slightly backward direction; the anterior branch is small. The fissure of Rolando, or central fissure, quite apparent, is, however, situated slightly more forward in the Orang than in man. It differentiates the frontal from the parietal lobe. The parieto-occipital fispure is well marked; bordered externally by the first occipital fold it descends internally on the mesial side of the hemisphere, separating the parietal from the occipal lobes.

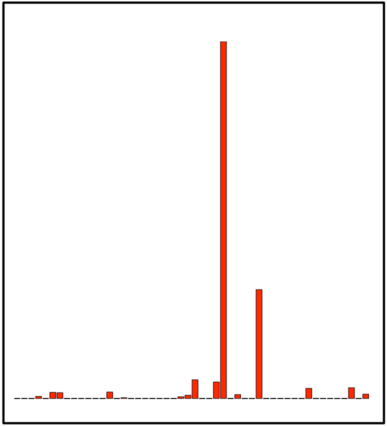
* From the Proceedings of the Academy of Natural Sciences, Phila., 1880.

in the Orang, the parieto-occipital fissure does not reach the calcarine, being separated from it by the "deuxième plis de passage interne" of Gratiolet, or "untere innere Scheitelbogen-Windung" of Bischoff. I have noticed this separation as an anomaly more than once in man.

According to Bischoff, this disposition obtains in the Gorilla, and seems to be usual also in the Chimpanzee. In the female Chimpanzee, however, on the left side I found the parieto-occipital fissure passing into the calcarine, as in man. The frontal lobe is easily distinguished from the parietal by the fissure of Rolando, and from the temporal by the fissure of Sylvius. In the Orang it is higher, wider, and more arched than in the Chimpanzee. The anterior central convolution in front of the central fissure runs into the post-central convolution above and below, as in man. It is difficult, however, to identify the three frontal convolutions seen in man and the Chimpanzee, the frontal lobe of the Orang dividing rather into two convolutions, the middle one being badly defined. This is due somewhat to the length of the precentral fissure, which is as long as the fissure of Rolando, extending farther upward than in man. There was nothing particularly noticeable about the base of the frontal lobe; on the mesial surface it ran into the parietal. The part above the calloso-marginal fissure in the Orang is not as distinctly divided into convolutions as in man, though these are not constantly present even in all human brains. The parietal lobe is separated from the frontal by the central fissure, from the occipital and temporal incompletely, by the parieto-occipital and Sylvian fissures. The posterior-central convolution is well defined. The parietal fissure in the Orang is more striking than that of man, resembling the Gorilla's; it is twice as long as the corresponding fissure in the Chimpanzee, extending from the transverse occipital fissure, as is sometimes the case in man, almost into the fissure of Rolando. It is unbridged and without a break, and divides the parietal lobe completely into upper and lower parietal lobules. The upper parietal lobule is bounded externally by the parietal fissure; posteriorly it is separated from the occipital lobe, internally by the parieto-

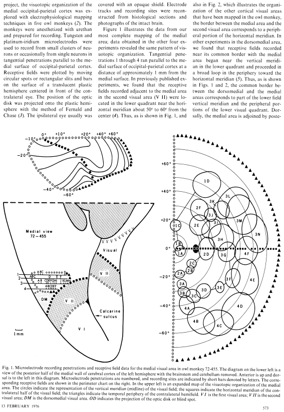
FIG. 2.

occipital fissure; externally it is continuous with the occipital lobe, as the first occipital gyrus, anteriorly it is separated from the posterior central convolution more completely than in man, by a fissure which runs paralled with the central fissure. There is in the Orang, also, a fissure running paralled with the parietal, which subdivides the upper parietal lobule into inner and outer portions. The precuneus, or the space on the mesial side of the parietal lobe between the parieto-occipital



Time-corrected document similarity

Representation of the Visual Field on the Medial Wall of Occipital-Parietal Cortex in the Owl Monkey (1976)

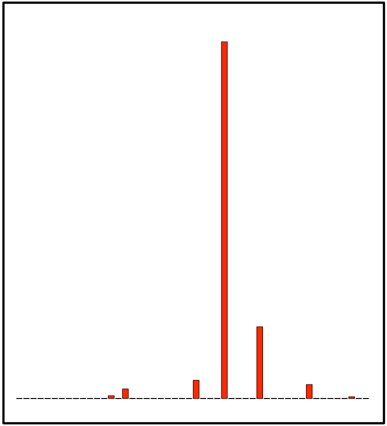


project, the visuotopic organization of the medial occipital-parietal cortex was explored with electrophysiological mapping techniques in five owl monkeys (*2*). The monkeys were anesthetized with urethan and prepared for recording. Tungsten and platinum-iridium microelectrodes were used to record from small clusters of neurons or occasionally from single neurons in tangential penetrations parallel to the medial surface of occipital-parietal cortex. Receptive fields were plotted by moving circular spots or rectangular slits and bars on the surface of a translucent plastic hemisphere centered in front of the contralateral eye. The position of the optic disk was projected onto the plastic hemisphere with the method of Fernald and Chase (*3*). The ipsilateral eye usually was covered with an opaque shield. Electrode tracks and recording sites were reconstructed from histological sections and photographs of the intact brain.

Figure 1 illustrates the data from our most complete mapping of the medial area; data obtained in the other four experiments revealed the same pattern of visuotopic organization. Tangential penetrations 1 through 4 ran parallel to the medial surface of occipital-parietal cortex at a distance of approximately 1 mm from the medial surface. In previously published experiments, we found that the receptive fields recorded adjacent to the medial area in the second visual area (V II) were located in the lower quadrant near the horizontal meridian about 50° to 60° from the center (*4*). Thus, as is shown in Fig. 1, and also in Fig. 2, which illustrates the organization of the other cortical visual areas that have been mapped in the owl monkey, the border between the medial area and the second visual area corresponds to a peripheral portion of the horizontal meridian. In other experiments in the dorsomedial area, we found that receptive fields recorded near its common border with the medial area began near the vertical meridian in the lower quadrant and proceeded in a broad loop in the periphery toward the horizontal meridian (*5*). Thus, as is shown in Figs. 1 and 2, the common border between the dorsomedial and the medial areas corresponds to part of the lower field vertical meridian and the peripheral portions of the lower visual quadrant. Dorsally, the medial area is adjoined by poste-

Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-455. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up and dorsal is to the left in this diagram. Microelectrode penetrations are numbered, and recording sites are indicated by short bars denoted by letters. The corresponding receptive fields are shown in the perimeter chart on the right. In the upper left is an expanded map of the visuotopic organization of the medial area. The circles indicate the representation of the vertical meridian (midline) of the visual field; the squares indicate the horizontal meridian of the contralateral half of the visual field; the triangles indicate the temporal periphery of the contralateral hemifield. *V I* is the first visual area; *V II* is the second visual area; *DM* is the dorsomedial visual area. *OD* indicates the projection of the optic disk or blind spot.

13 FEBRUARY 1976 573



Exploring the UN General Debates with Dynamic Topic Models

Luke Lefebure Follow
Oct 17, 2018 · 11 min read

Credit: Vladislav Klapin on Unsplash

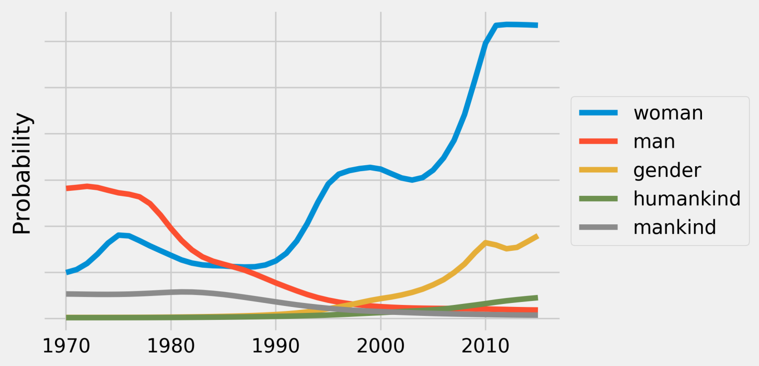
https://towardsdatascience.com/

exploring-the-un-general-debates-with-dynamic-topic-models-72dc0e307696

Human rights

	1970	1980	1990	2000	2010
0	right	right	right	right	right
1	human	human	human	human	human
2	people	people	freedom	law	law
3	international	freedom	people	democracy	woman
4	principle	international	democracy	respect	freedom
5	justice	political	respect	international	respect
6	freedom	principle	law	people	people
7	law	respect	international	freedom	democracy
8	state	justice	principle	principle	rule
9	must	social	state	must	international

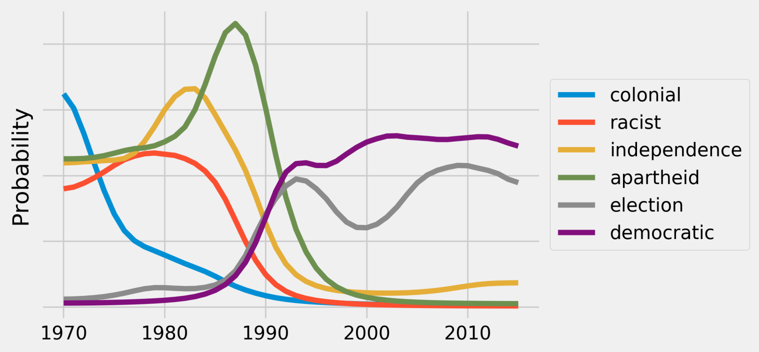
Human rights



Apartheid

	1970	1980	1990	2000	2010
0	africa	africa	africa	african	african
1	african	south	south	peace	country
2	south	african	african	africa	government
3	colonial	namibia	apartheid	country	africa
4	people	people	people	government	people
5	regime	regime	government	community	peace
6	southern	independence	country	democratic	political
7	government	apartheid	namibia	international	community
8	territory	racist	community	republic	democratic
9	apartheid	southern	process	effort	national

Apartheid

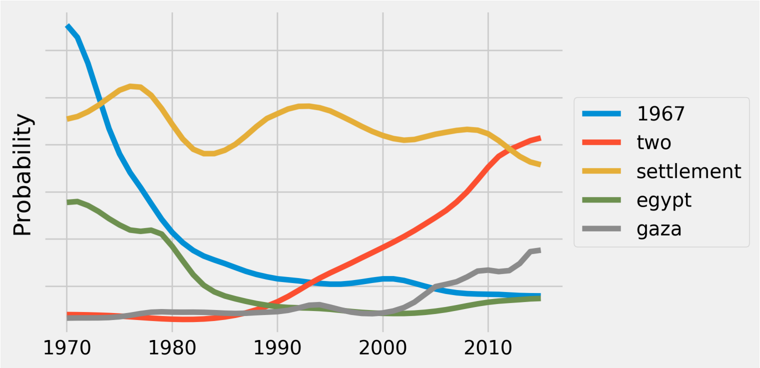


Probability
1970
1980
1990
2000
2010
colonial
racist
independence
apartheid
election
democratic

Arab-Israeli conflict

	1970	1980	1990	2000	2010
0	arab	palestinian	peace	peace	peace
1	israel	israel	east	east	state
2	east	right	middle	resolution	palestinian
3	middle	east	arab	palestinian	solution
4	peace	people	people	middle	east
5	territory	arab	palestinian	security	international
6	resolution	middle	international	people	israel
7	1967	peace	israel	israel	middle
8	palestinian	territory	kuwait	state	arab
9	israeli	state	right	international	people

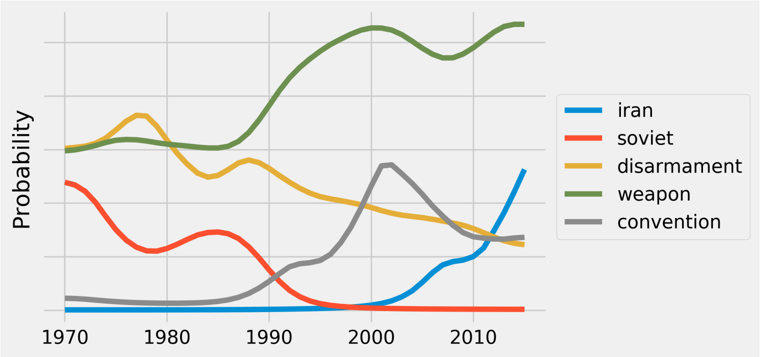
Arab-Israeli conflict



Nuclear arms

	1970	1980	1990	2000	2010
0	nuclear	nuclear	nuclear	weapon	nuclear
1	disarmament	disarmament	weapon	nuclear	weapon
2	weapon	weapon	disarmament	convention	non
3	soviet	arm	treaty	arm	proliferation
4	arm	state	arm	treaty	arm
5	treaty	race	state	disarmament	treaty
6	union	military	chemical	proliferation	international
7	agreement	treaty	agreement	international	disarmament
8	power	soviet	proliferation	non	convention
9	state	power	soviet	destruction	state

Nuclear arms



Tutorials

https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

https://towardsdatascience.com/

latent-dirichlet-allocation-intuition-math-implementation-and-visualisation-63ccb616e094

Summary

- Topic models automatically extract “semantic themes” from large document collections
- Use mixtures and latent variables
- Estimating Bayesian posterior done with Gibbs sampling
- Many extensions are possible