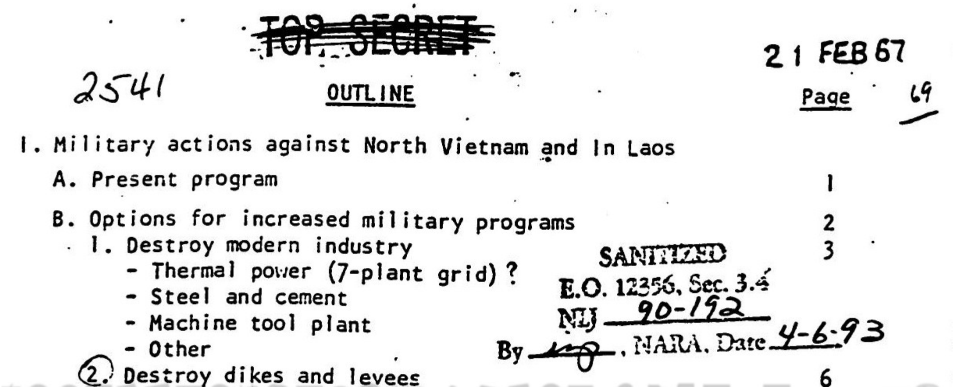
Statistical Topic Models for Computational Social Science

Hanna M. Wallach

University of Massachusetts Amherst

wallach@cs.umass.edu

Complex Social Processes



~~TOP SECRET~~

21 FEB 67

2541

OUTLINE

Page 69

I. Military actions against North Vietnam and in Laos

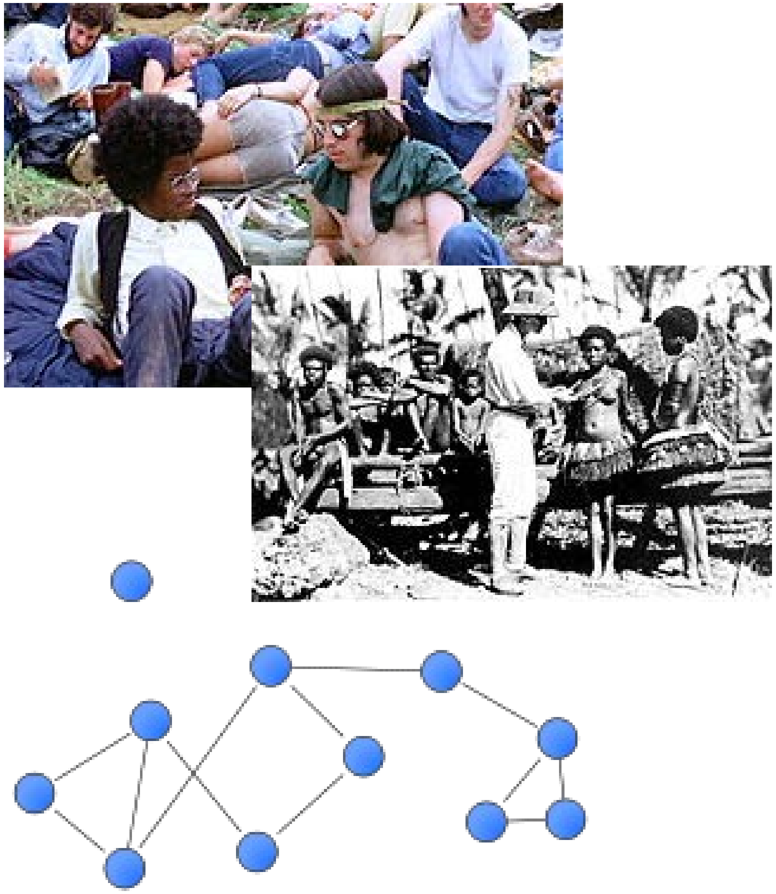
A. Present program 1

B. Options for increased military programs 2

1. Destroy modern industry 3
   - Thermal power (7-plant grid) ?
   - Steel and cement
   - Machine tool plant
   - Other
2. Destroy dikes and levees 6

SANITIZED
E.O. 12356, Sec. 3.4
NLJ 90-192
By , NARA, Date 4-6-93

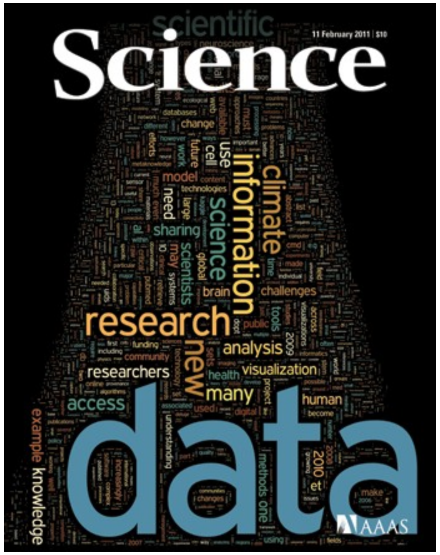
“Traditional” Social Science

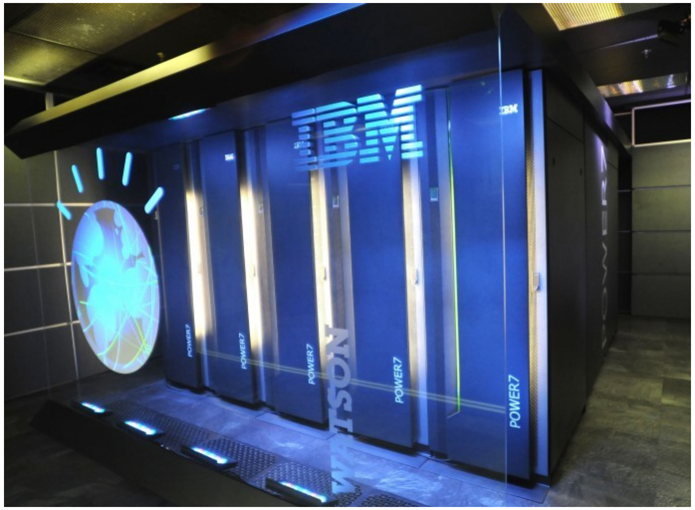


- Case studies
- Interviews
- Participant observation
- Survey research
- Social network analysis

⇒ Self-reports, one-time snapshots, small scale

The Computer “Revolution”





hadoop
Map Reduce

Google

The Unreasonable Effectiveness of Data

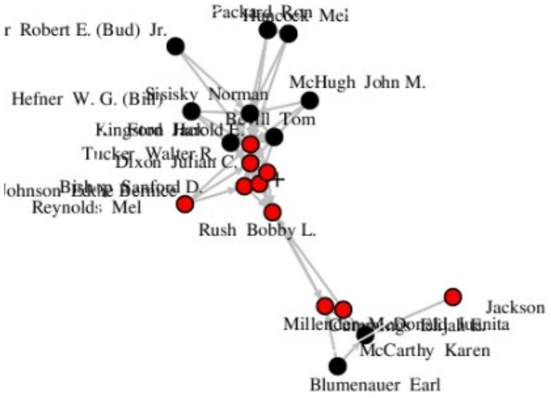
Computational Social Science

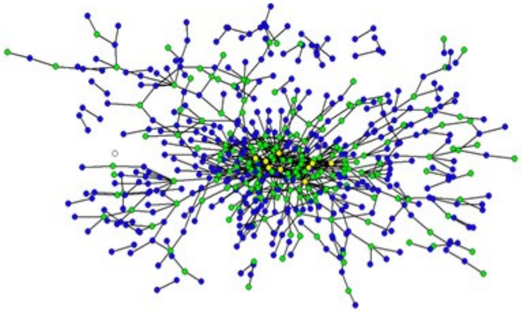
Candida Hofer

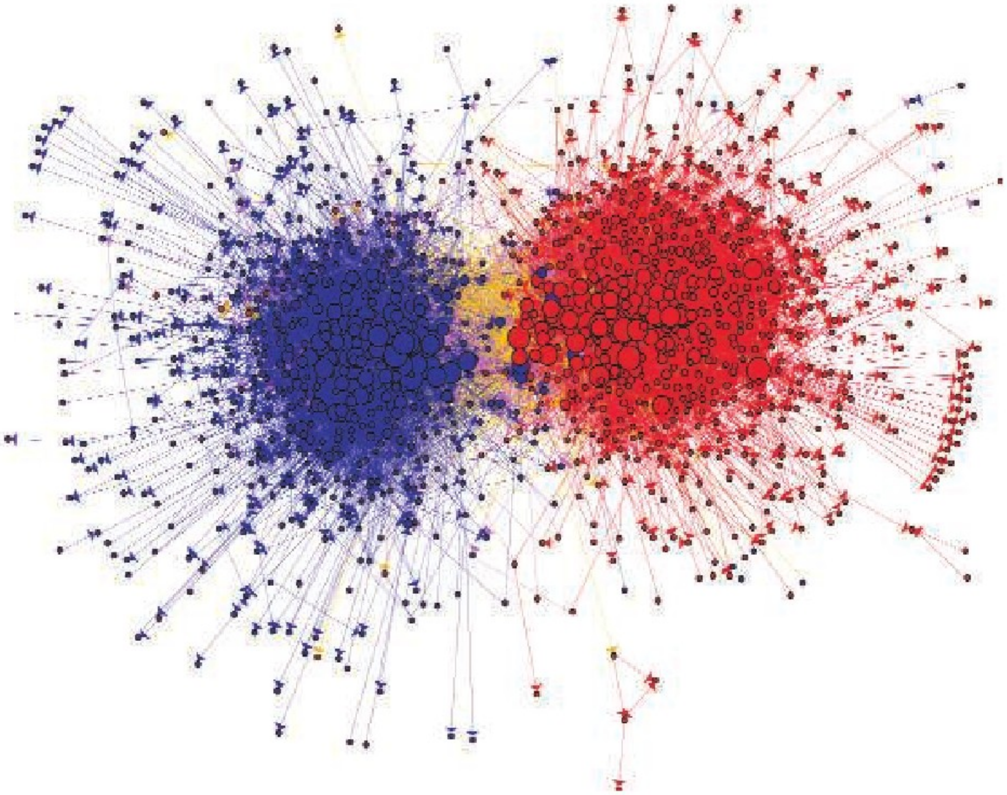
"A computational social science is emerging that leverages the capacity to collect and analyze data with an unprecedented breadth and depth and scale and may reveal patterns of individual and group behaviors.”

— Lazer et al., 2009

Structure vs. Content





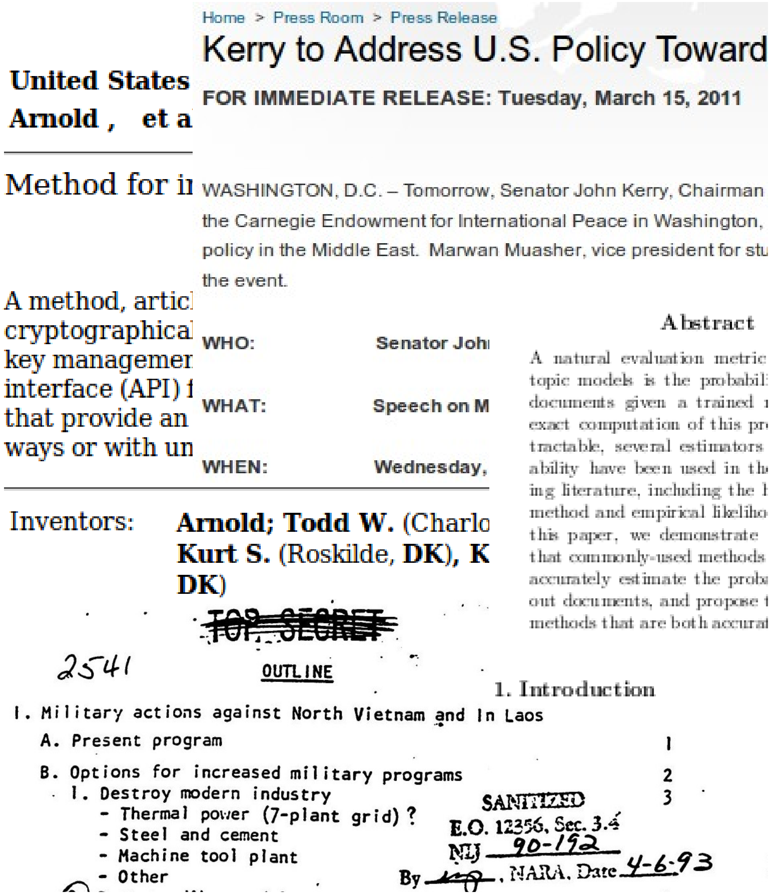


Products of Interactions

> "Scientific information is both the basic raw material for, and one of the principal products of, scientific research […] Scientists find out what other scientists are accomplishing through […] journals, books, abstracts and indexes, bibliographies, reviews."

— NSF Brochure, 1962

Text as Data



United States
Arnold , et a

Method for i

A method, artic
cryptographical
key managemer
interface (API) f
that provide an
ways or with un

Inventors: **Arnold; Todd W.** (Charlo
Kurt S. (Roskilde, **DK**), **K**
DK)

Home > Press Room > Press Release

Kerry to Address U.S. Policy Toward

FOR IMMEDIATE RELEASE: Tuesday, March 15, 2011

WASHINGTON, D.C. – Tomorrow, Senator John Kerry, Chairman
the Carnegie Endowment for International Peace in Washington,
policy in the Middle East. Marwan Muasher, vice president for stu
the event.

WHO:	**Senator Joh**
WHAT:	**Speech on M**
WHEN:	**Wednesday,**

Abstract

A natural evaluation metric
topic models is the probabil
documents given a trained i
exact computation of this pr
tractable, several estimators
ability have been used in th
ing literature, including the h
method and empirical likeliho
this paper, we demonstrate
that commonly-used methods
accurately estimate the prob
out documents, and propose t
methods that are both accura

1. Introduction

~~TOP SECRET~~

2541

OUTLINE

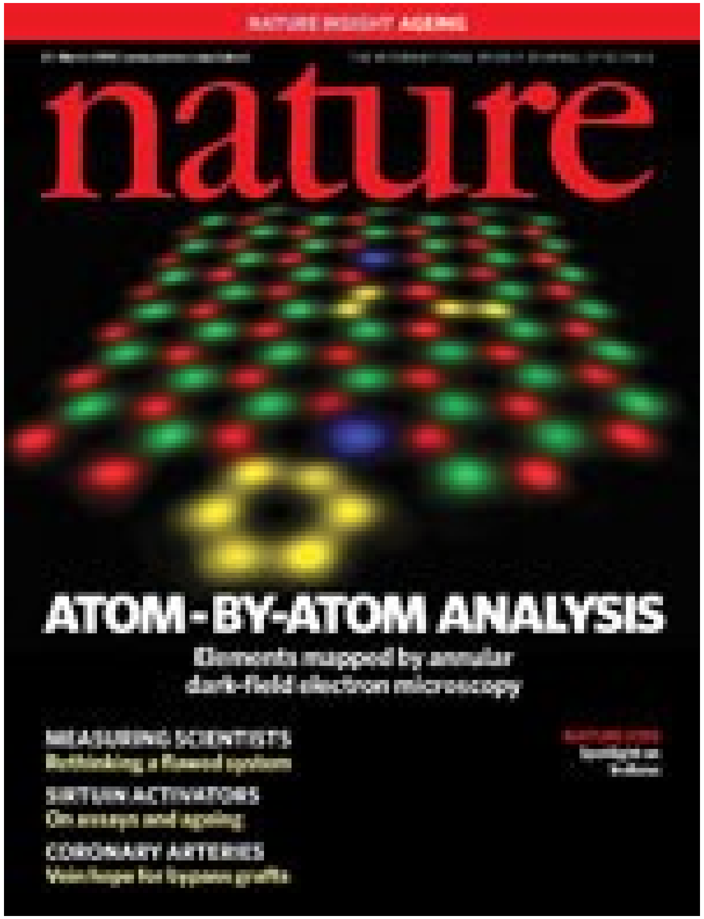
I. Military actions against North Vietnam and In Laos
A. Present program 1
B. Options for increased military programs 2
1. Destroy modern industry 3
- Thermal power (7-plant grid)?
- Steel and cement
- Machine tool plant
- Other

SANITIZED
E.O. 12356, Sec. 3.4
NLJ 90-192
By , NARA, Date 4-6-93

- Structured and formal: e.g., publications, patents, press releases
- Messy and unstructured: e.g., chat logs, OCRed documents, transcripts

⇒ Large scale, robust methods for analyzing text

Collaborate to Study Collaboration



> “There needs to be a greater focus on what these [interaction] data mean [...] This requires the input of social scientists, rather than just those more traditionally involved in data capture, such as computer scientists.”
>
> — Julia Lane, NSF, 24 March 2010

Different (But Overlapping) Roles

- **Social science:** specific models for specific applications, extensive post-analysis work
- **Computer science:** novel classes of models, mathematical and computational properties of models that extend across applications

This Talk

- Statistical topic models for text analysis
- “Off-the-shelf” topic models: priors, stop words
- Studying formerly-classified government documents

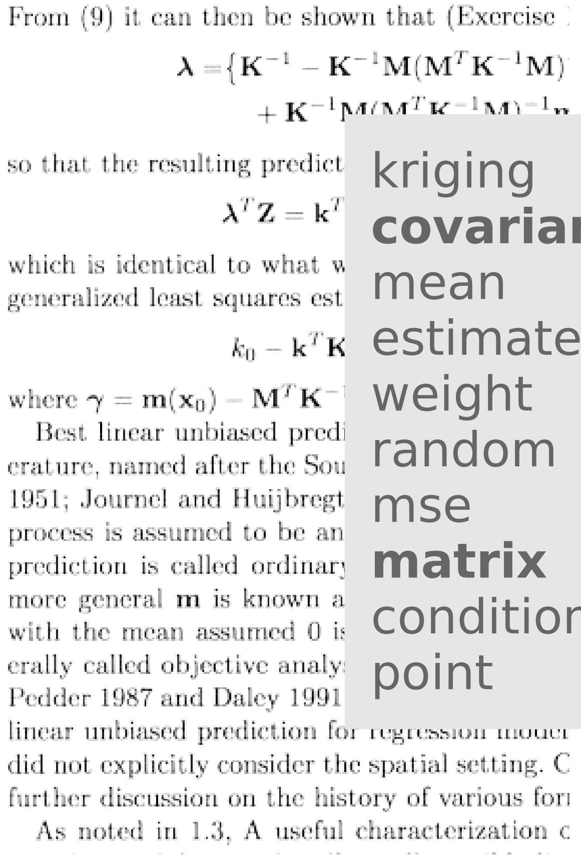
Statistical Modeling

- Modeling challenges:
  - Aggregating and representing large data sets
  - Handling data from sources with disparate emphases
  - Reasoning under uncertain information
  - Performing efficient inference
- Bayesian latent (hidden) variable models:
  - Powerful and flexible [Wallach et al. & Adams et al., AISTATS '10]
  - This talk: statistical topic models

Statistical Topic Modeling

- Three fundamental assumptions:
  - Documents have latent semantic structure (“topics”)
  - We can infer topics from word–document co-occurrences
  - Can simulate this inference algorithmically
- Given a data set, the goal is to
  - Learn the composition of the topics that best represent it
  - Learn which topics are used in each document

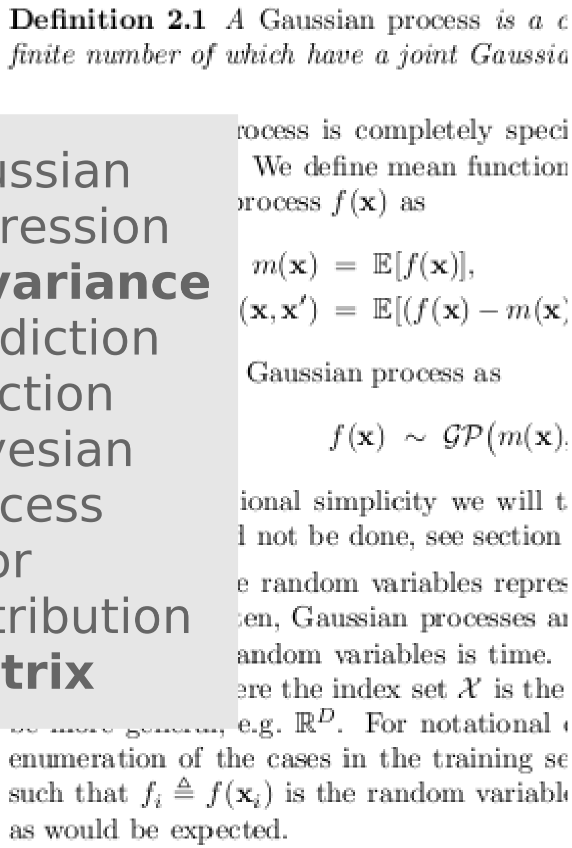
Why Topic Models?



From (9) it can then be shown that (Exercise
$\boldsymbol{\lambda} = \{\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{M}(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})$
$+ \mathbf{K}^{-1}\mathbf{M}(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})^{-1}\mathbf{m}$
so that the resulting predict
$\boldsymbol{\lambda}^T\mathbf{Z} = \mathbf{k}^T$
which is identical to what w
generalized least squares est
$k_0 - \mathbf{k}^T\mathbf{K}$
where $\boldsymbol{\gamma} = \mathbf{m}(\mathbf{x}_0) - \mathbf{M}^T\mathbf{K}^{-1}$
Best linear unbiased predi
erature, named after the Sou
1951; Journel and Huijbregt
process is assumed to be an
prediction is called ordinary
more general **m** is known a
with the mean assumed 0 is
erally called objective analy
Pedder 1987 and Daley 1991
linear unbiased prediction for regression model
did not explicitly consider the spatial setting. C
further discussion on the history of various for
As noted in 1.3, A useful characterization c

kriging
covariance
mean
estimate
weight
random
mse
matrix
conditional
point

vs.



Definition 2.1 *A* Gaussian process *is a c*
finite number of which have a joint Gaussia
rocess is completely speci
We define mean function
process $f(\mathbf{x})$ as
$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})],$
$(\mathbf{x},\mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x})$
Gaussian process as
$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}),$
ional simplicity we will t
l not be done, see section
e random variables repres
en, Gaussian processes ar
andom variables is time.
ere the index set $\mathcal{X}$ is the
be more general, e.g. $\mathbb{R}^D$. For notational
enumeration of the cases in the training se
such that $f_i \triangleq f(\mathbf{x}_i)$ is the random variable
as would be expected.

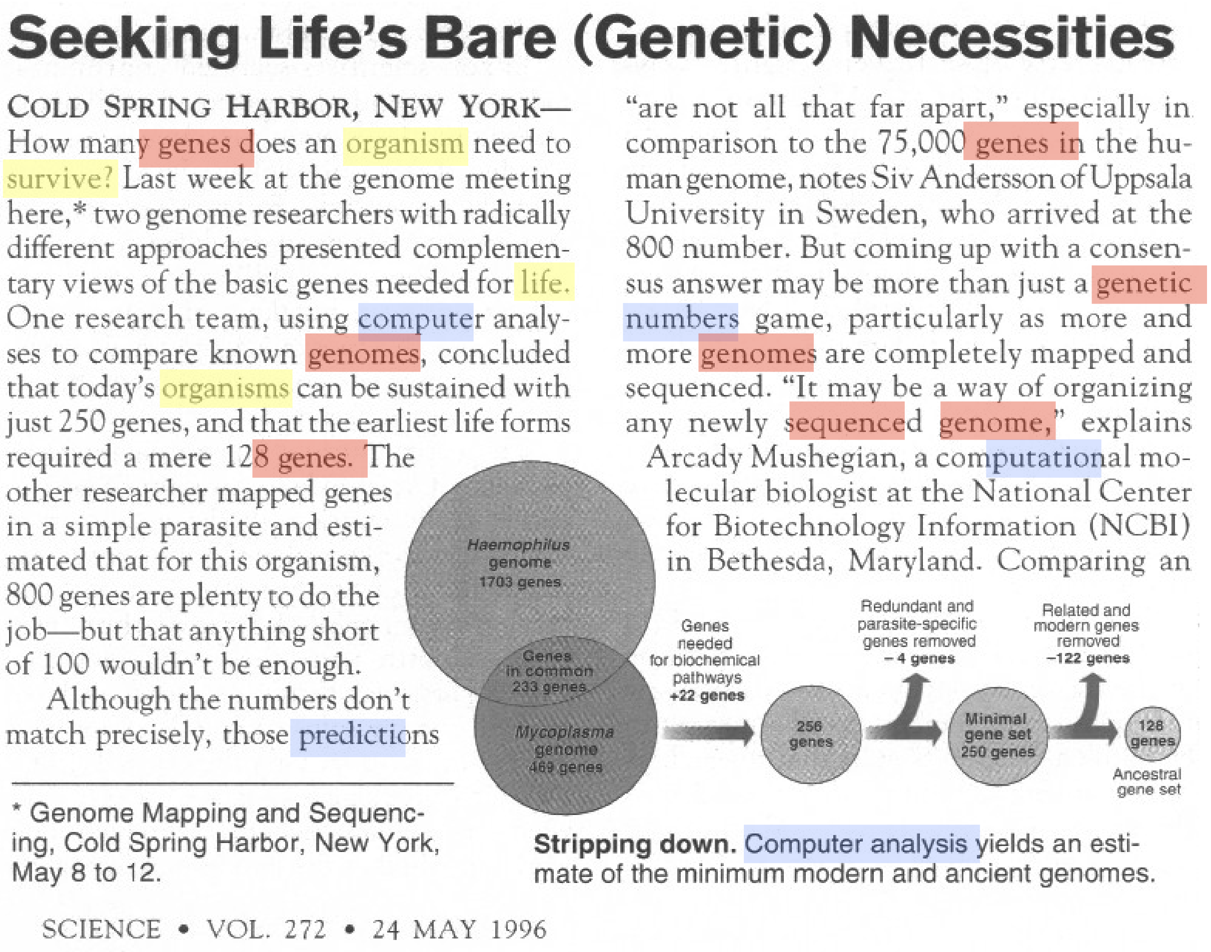
gaussian
regression
covariance
prediction
function
bayesian
process
prior
distribution
matrix

Topics and Words

probability

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
...	...	...	...

Documents and Topics



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.

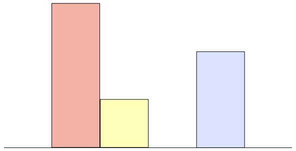
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set

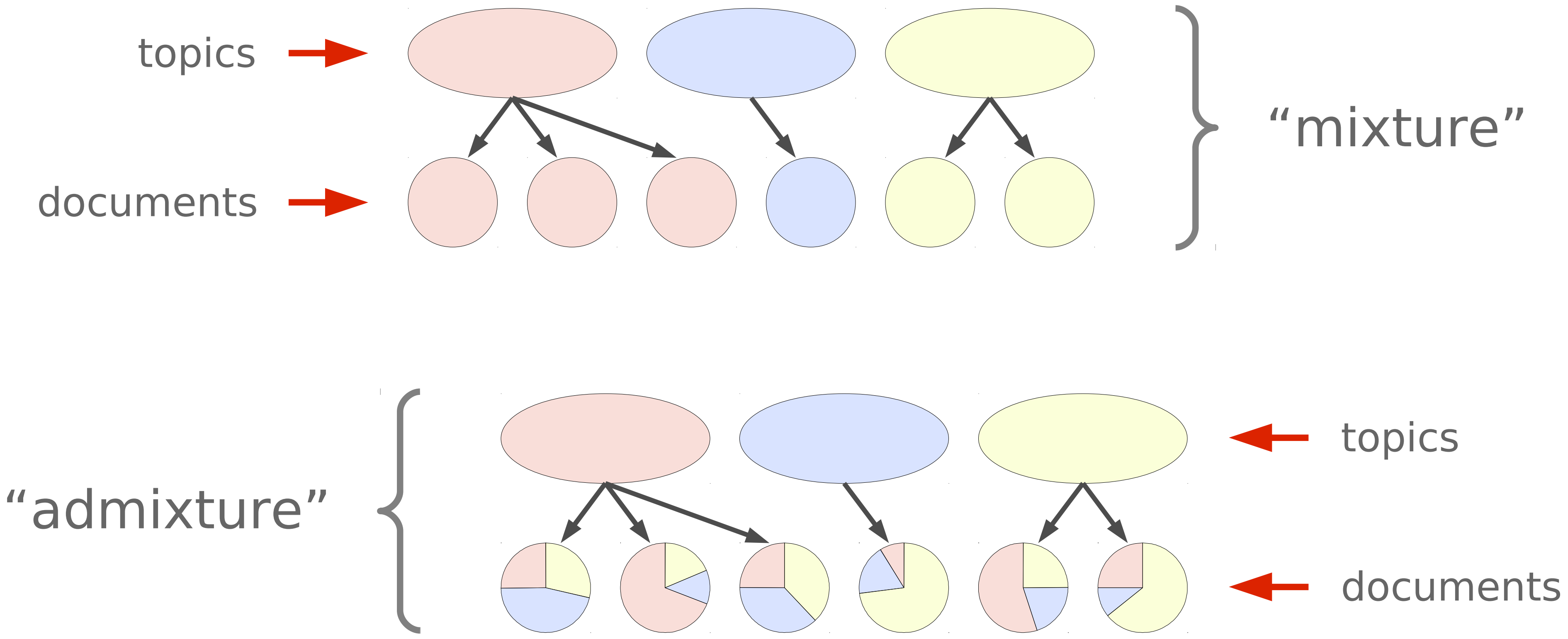
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996



Mixtures vs. Admixtures



topics

documents

“mixture”

“admixture”

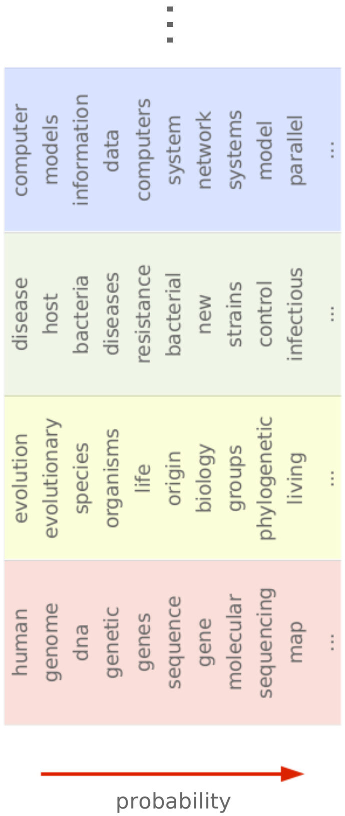
topics

documents

Generative Statistical Modeling

- Assume data was generated by a probabilistic model:
  - Model may have hidden structure (latent variables)
  - Model defines a joint distribution over all variables
  - Model parameters are unknown
- Infer hidden structure and model parameters from data
- Situate new data in estimated model

Generative Process



⋮

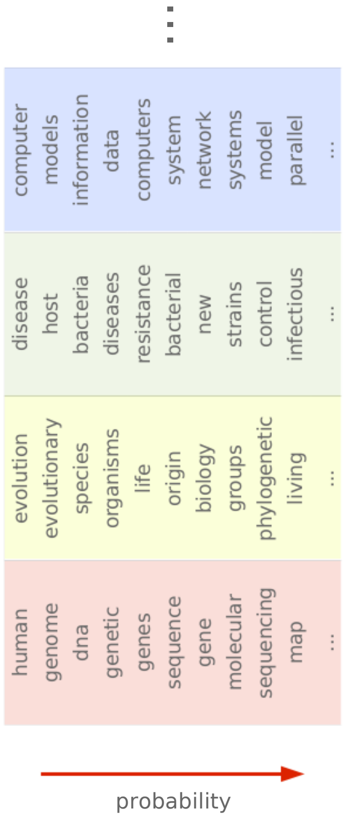
computer models information data computers system network systems model parallel …

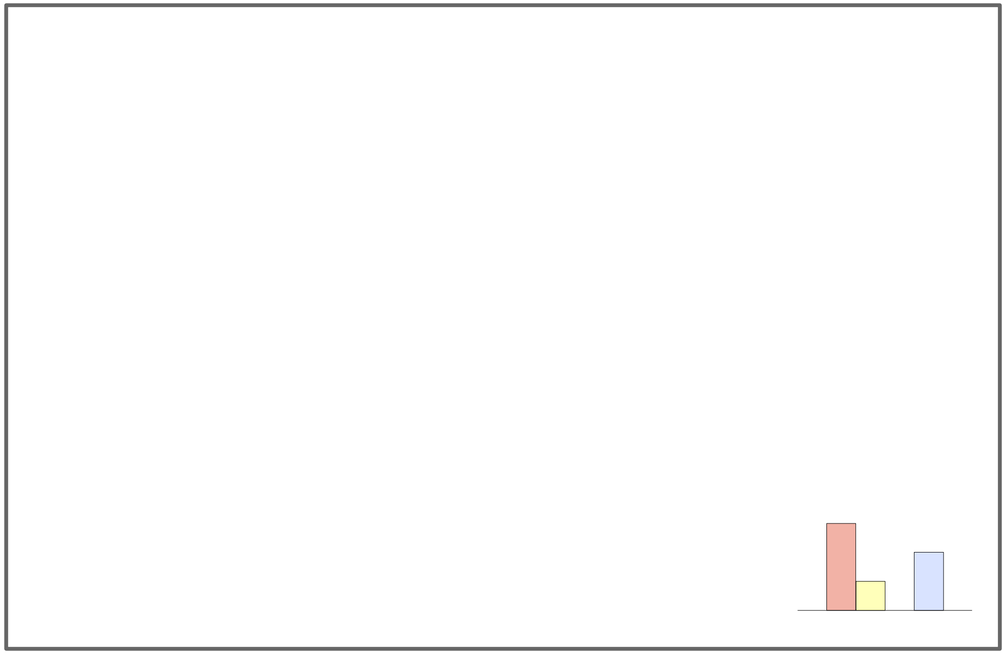
disease host bacteria diseases resistance bacterial new strains control infectious …

evolution evolutionary species organisms life origin biology groups phylogenetic living …

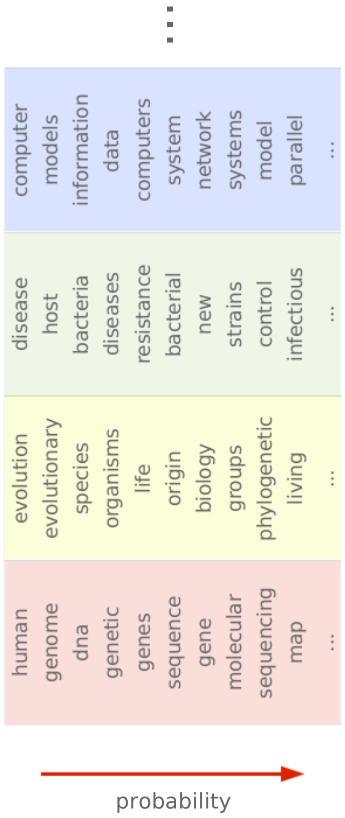
human genome dna genetic genes sequence gene molecular sequencing map …

Choose a Distribution Over Topics


⋮
computer models information data computers system network systems model parallel …
disease host bacteria diseases resistance bacterial new strains control infectious …
evolution evolutionary species organisms life origin biology groups phylogenetic living …
human genome dna genetic genes sequence gene molecular sequencing map …
probability



Choose a Topic



⋮

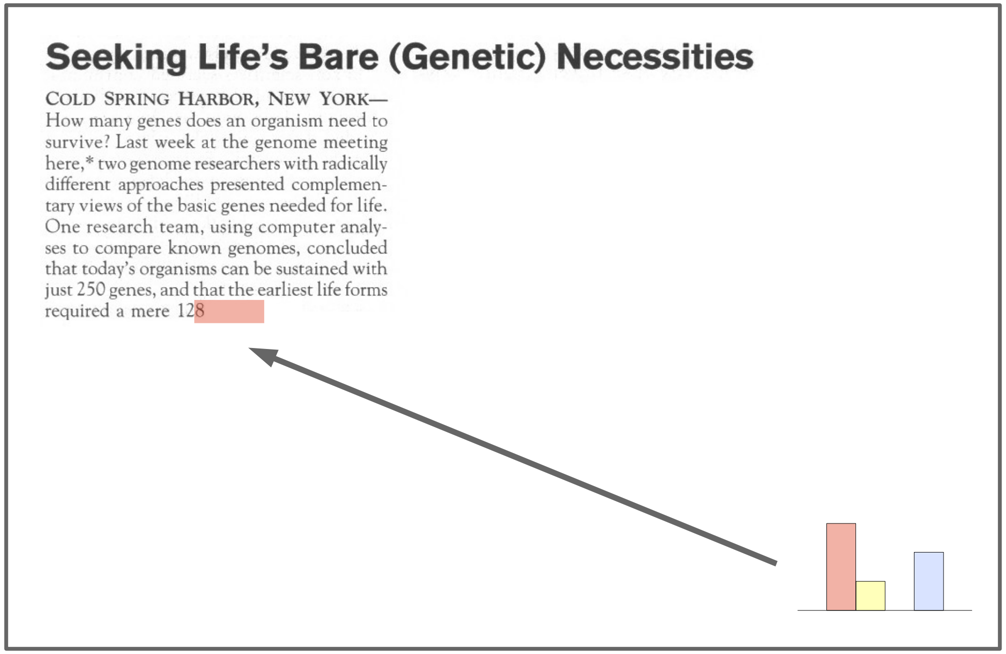
computer
models
information
data
computers
system
network
systems
model
parallel
...

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
...

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
...

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
...

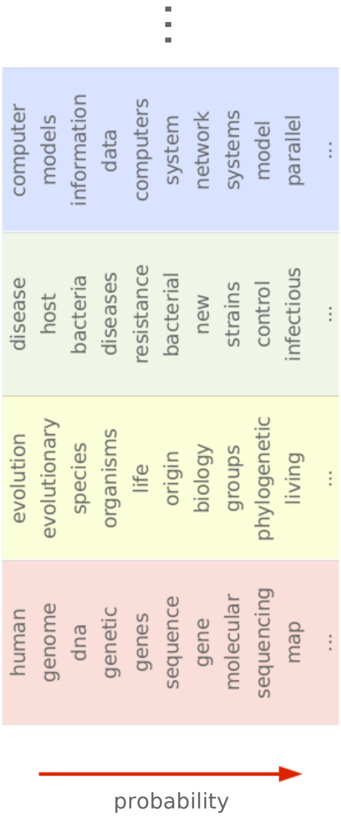
probability



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

Choose a Word



⋮

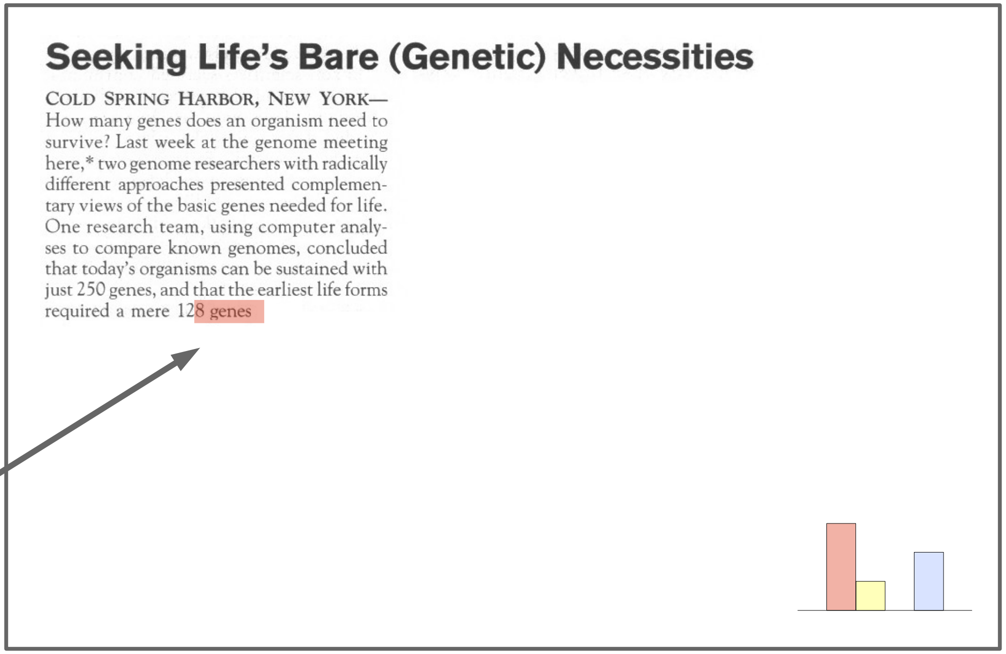
computer models information data computers system network systems model parallel …

disease host bacteria diseases resistance bacterial new strains control infectious …

evolution evolutionary species organisms life origin biology groups phylogenetic living …

human genome dna genetic genes sequence gene molecular sequencing map …

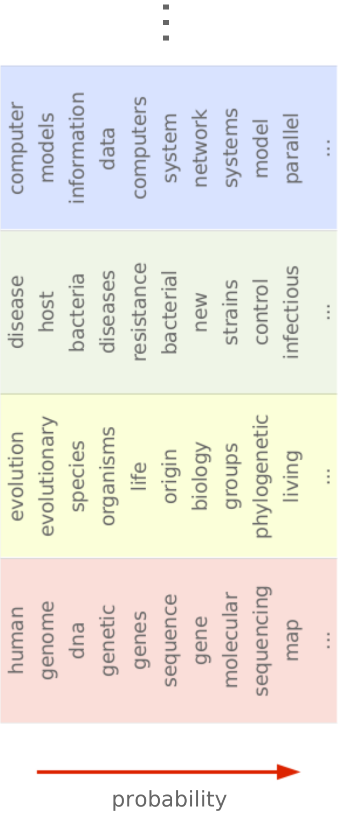
probability



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

… And So On



⋮

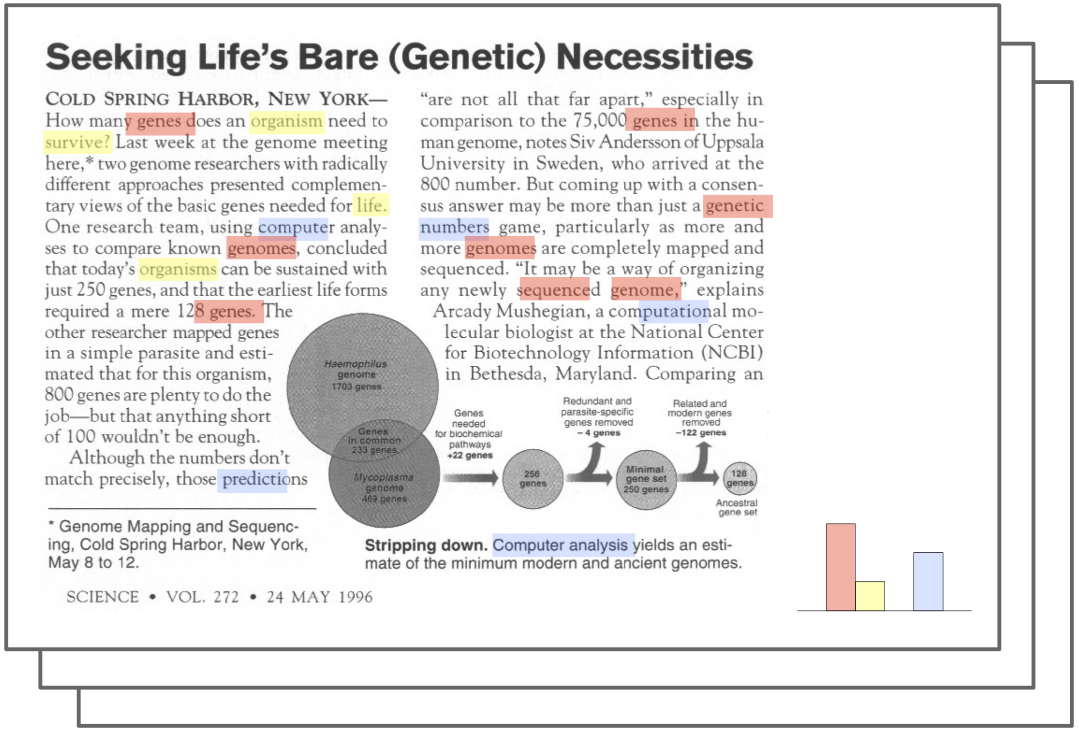
computer
models
information
data
computers
system
network
systems
model
parallel
…

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
…

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
…

human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
…

probability



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

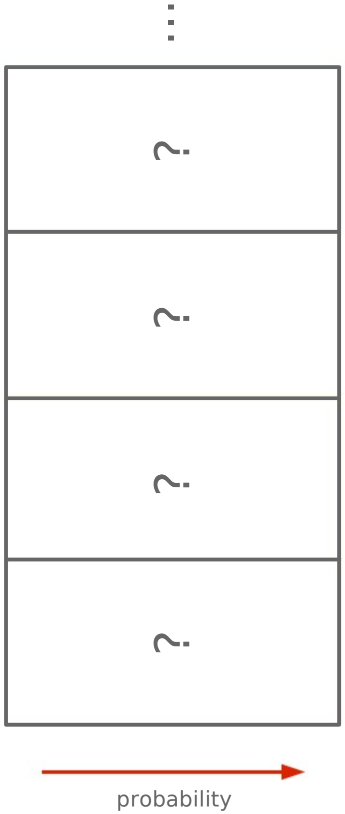
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set

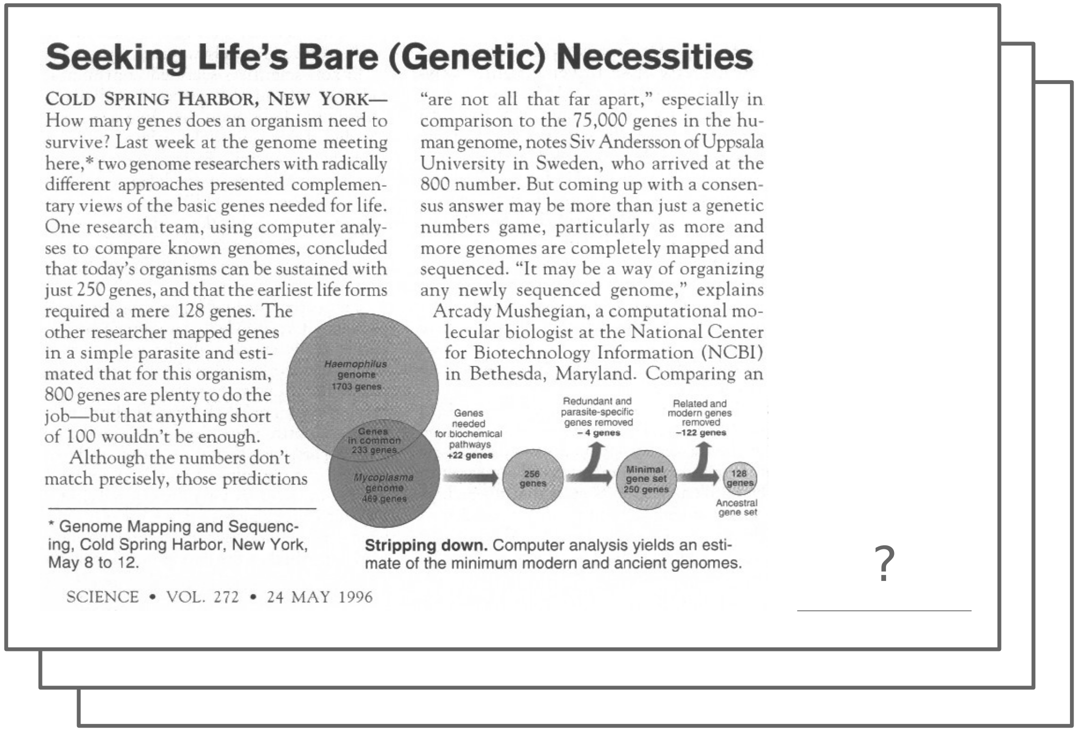
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Real Data: Statistical 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

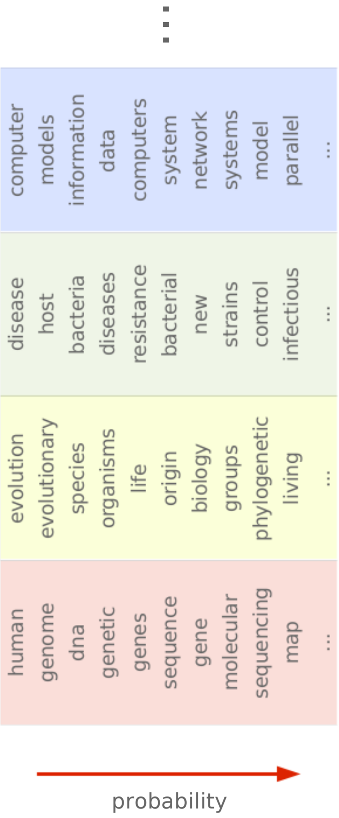
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

?

The End Result...



⋮

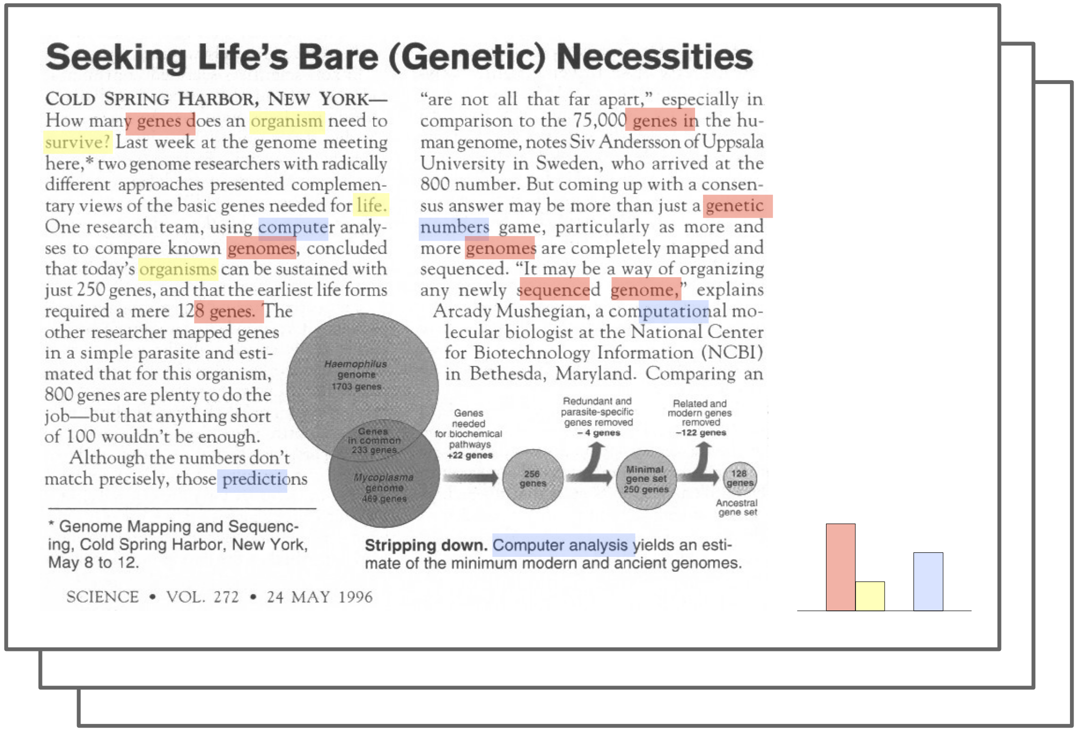
computer models information data computers system network systems model parallel ...

disease host bacteria diseases resistance bacterial new strains control infectious ...

evolution evolutionary species organisms life origin biology groups phylogenetic living ...

human genome dna genetic genes sequence gene molecular sequencing map ...

probability



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set

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

This Talk

- Statistical topic models for text analysis
- “Off-the-shelf” topic models: priors, stop words
- Studying formerly-classified government documents

The State of The Art

- Topic models are extremely appealing
- … but they're not always usable by non-experts
- Need to bridge this gap between producers and consumers of topic modeling technology:
  - Address problems/challenges faced by practitioners
  - Question unquestioned assumptions
  - Explore the interplay between theory and practice

“Off-the-Shelf” Topic Modeling

I want to model technology emergence by analyzing patent abstracts...

I have a statistical model that you can use...

“Off-the-Shelf” Topic Modeling

I want to model technology emergence by analyzing patent abstracts...

I have a statistical model that you can use...

a	a	the	the
field	the	of	**invention**
emission	**carbon**	a	of
an	and	to	to
electron	**gas**	and	**present**
...	...	...	...

“Off-the-Shelf” Topic Modeling?

Help! All my topics consist of “the, and of, to, a ...”

Now they all consist of “invention, present, thereof ...”

Wait, but how do I choose the right number of topics?

Preprocess your data to remove stop words...

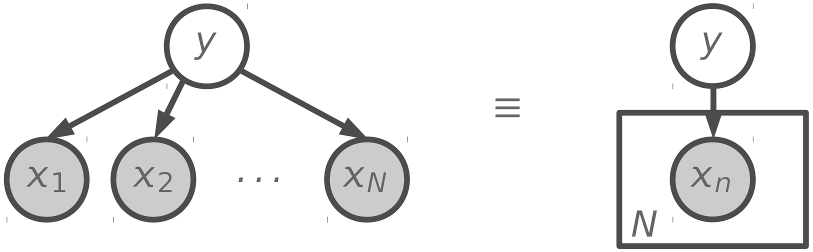
Make a domain-specific list of stop words...

Evaluate the probability of unseen data for different numbers...

Directed Graphical Models

$$P(y, x_1, \ldots, x_N) = P(y) \prod_{n=1}^{N} P(x_n \,|\, y)$$

- Nodes: random variables (latent or observed)
- Edges: probabilistic dependencies between variables
- Plates: “macros” that allow subgraphs to be replicated

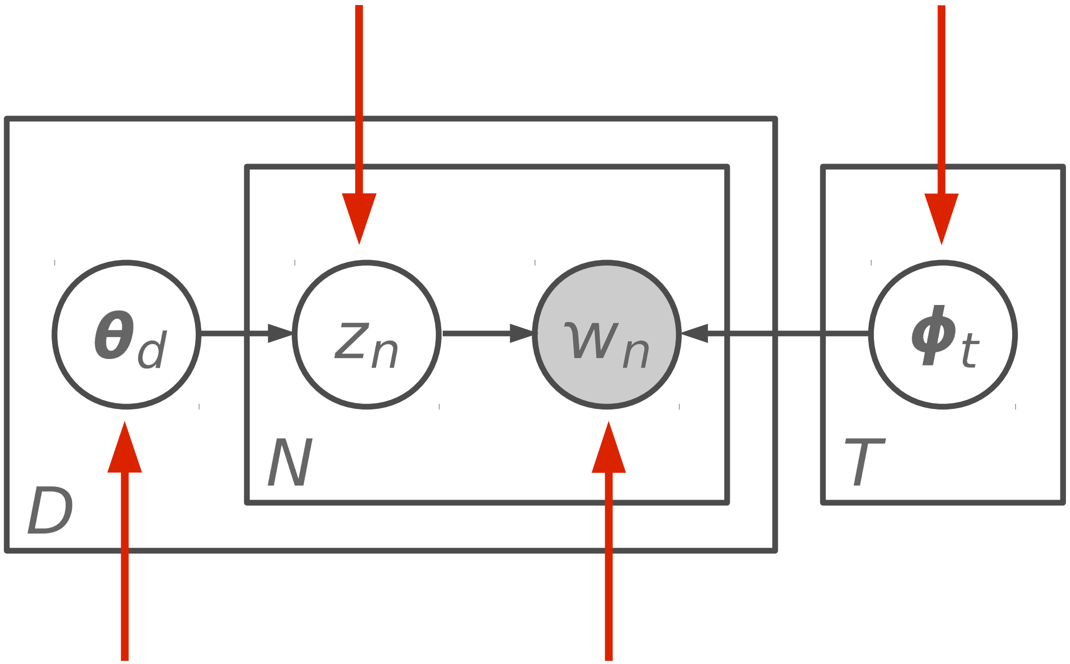


Statistical Topic Modeling

[Hofmann, '99]

topic
assignment

topics

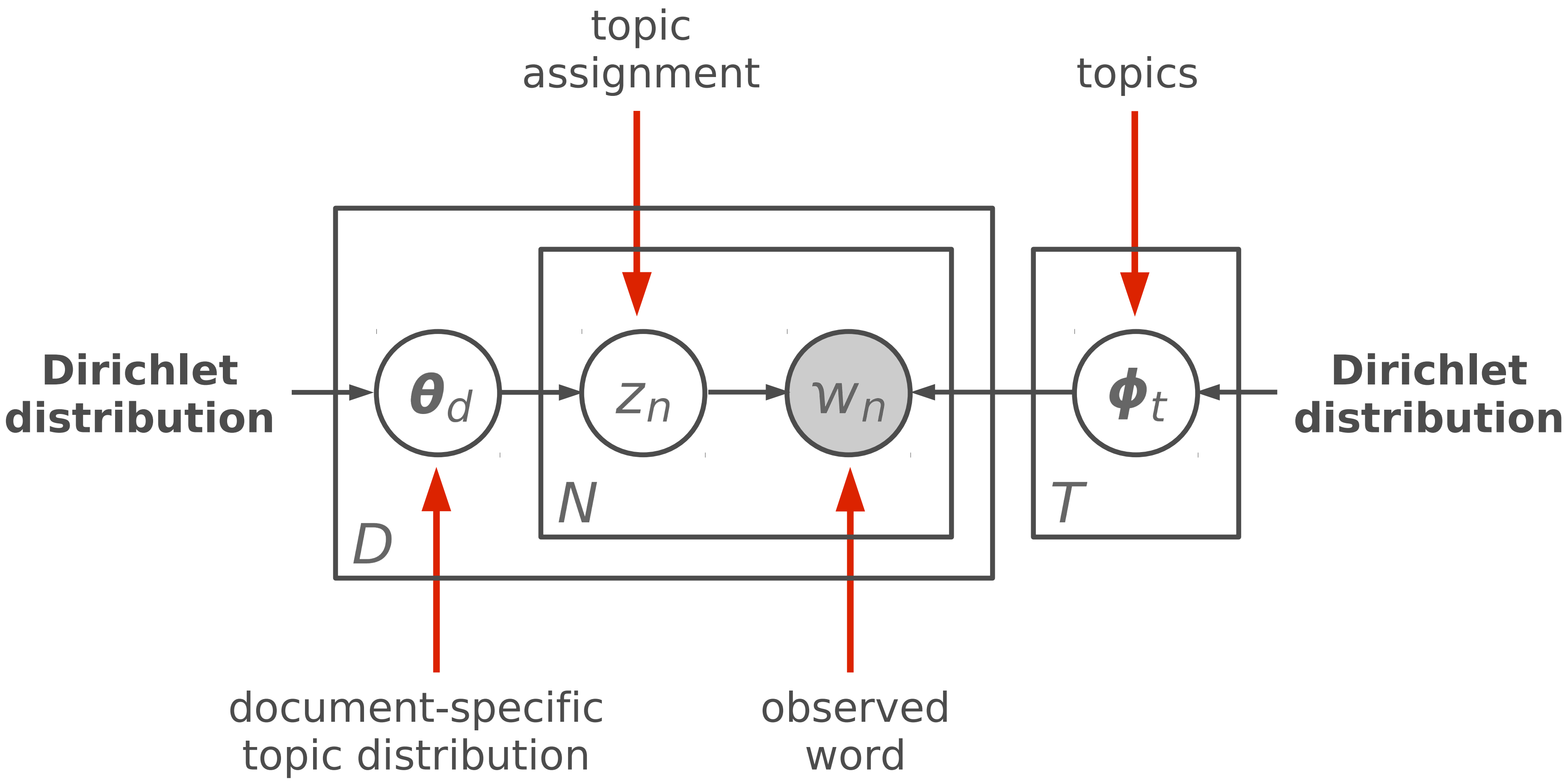


document-specific
topic distribution

observed
word

Latent Dirichlet Allocation (LDA)

[Blei, Ng & Jordan, '03]



topic assignment

topics

Dirichlet distribution

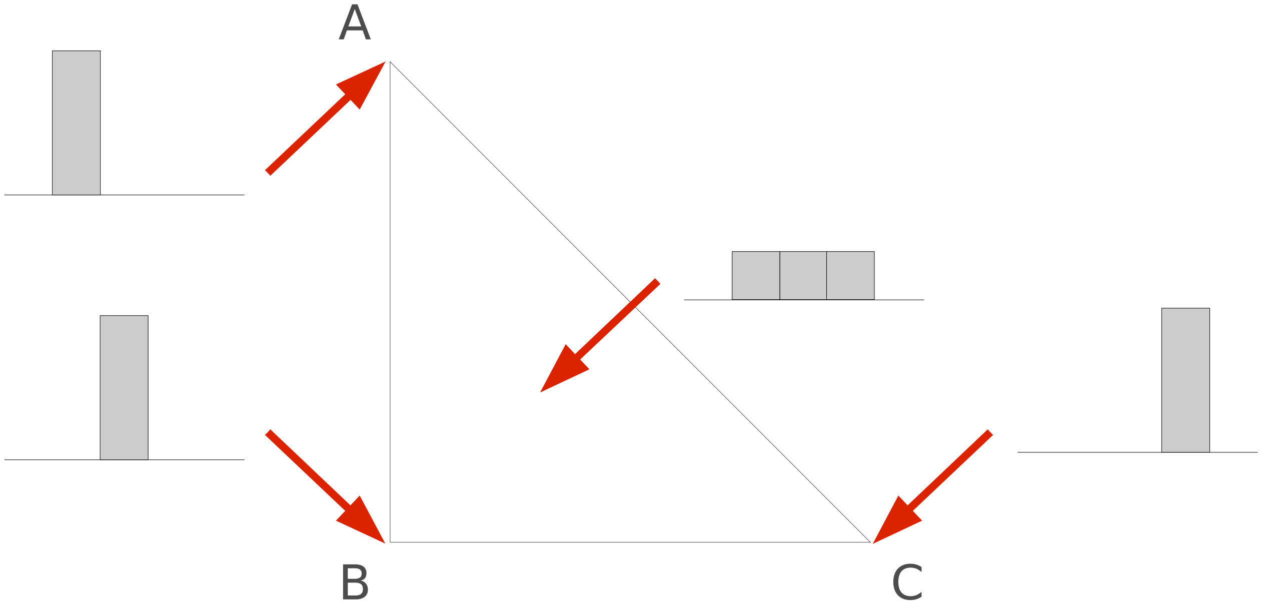
Dirichlet distribution

document-specific topic distribution

observed word

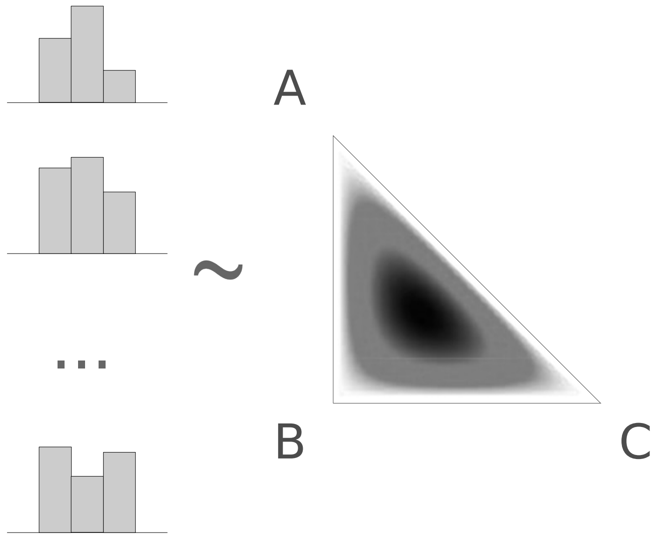
Discrete Probability Distributions

- 3-dimensional discrete probability distributions can be visually represented in 2-dimensional space:



Dirichlet Distribution

- Distribution over discrete probability distributions:

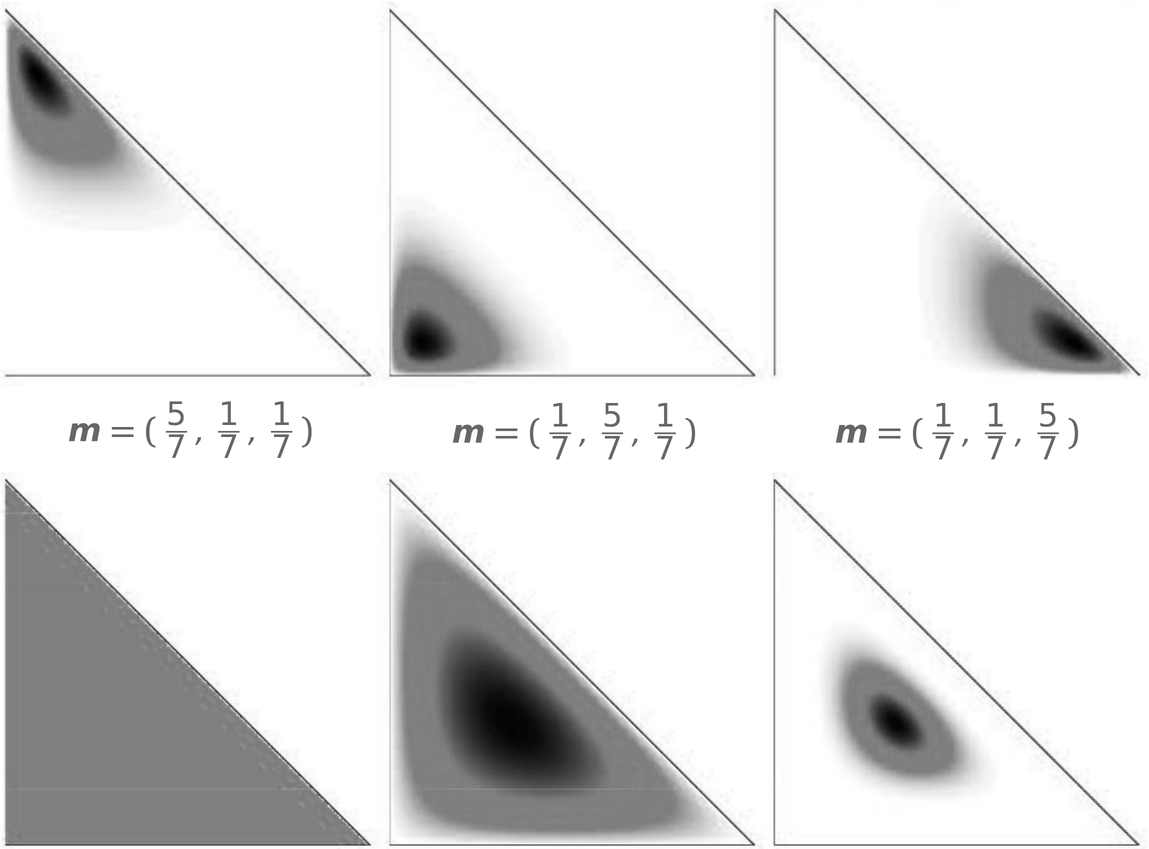


base measure (mean)

$$\equiv \quad \boldsymbol{p} \sim \mathrm{Dir}\,(\alpha \boldsymbol{m})$$

concentration parameter

Dirichlet Parameters



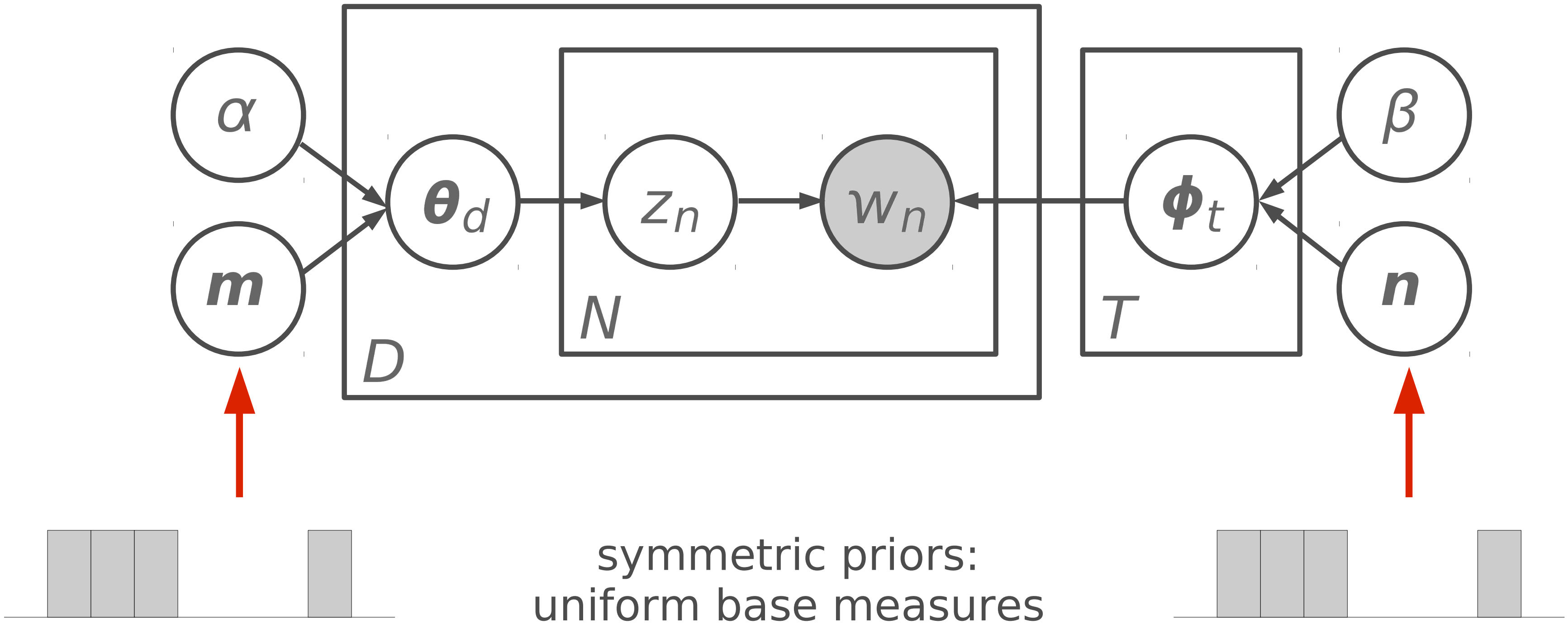
$\alpha = 14$

$\boldsymbol{m} = (\frac{5}{7}, \frac{1}{7}, \frac{1}{7})$ $\boldsymbol{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7})$ $\boldsymbol{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7})$

$\boldsymbol{m} = \boldsymbol{u} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

$\alpha = 3$ $\alpha = 6$ $\alpha = 30$

Dirichlet Priors for LDA



symmetric priors:
uniform base measures

Dirichlet Priors for LDA

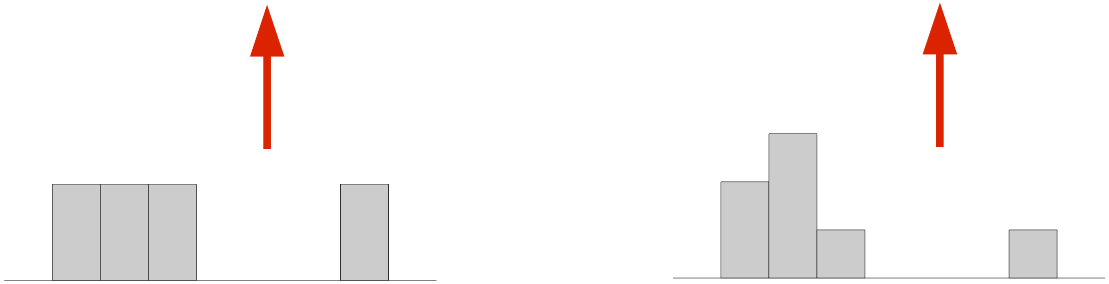
- Two scalar concentration parameters: α and β
- Concentration parameters are usually set heuristically
  - e.g., $\alpha = 50$ and $\beta = 0.01W$
- Some recent work on learning optimal values for the concentration parameters from data
- No rigorous study of the Dirichlet priors:
  - e.g., asymmetric vs. symmetric base measures
  - Effects of the base measures on the inferred topics

Symmetric → Asymmetric

[Wallach et al., '09]

- Use prior over $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$ as a running example
- Uniform base measure → nonuniform base measure

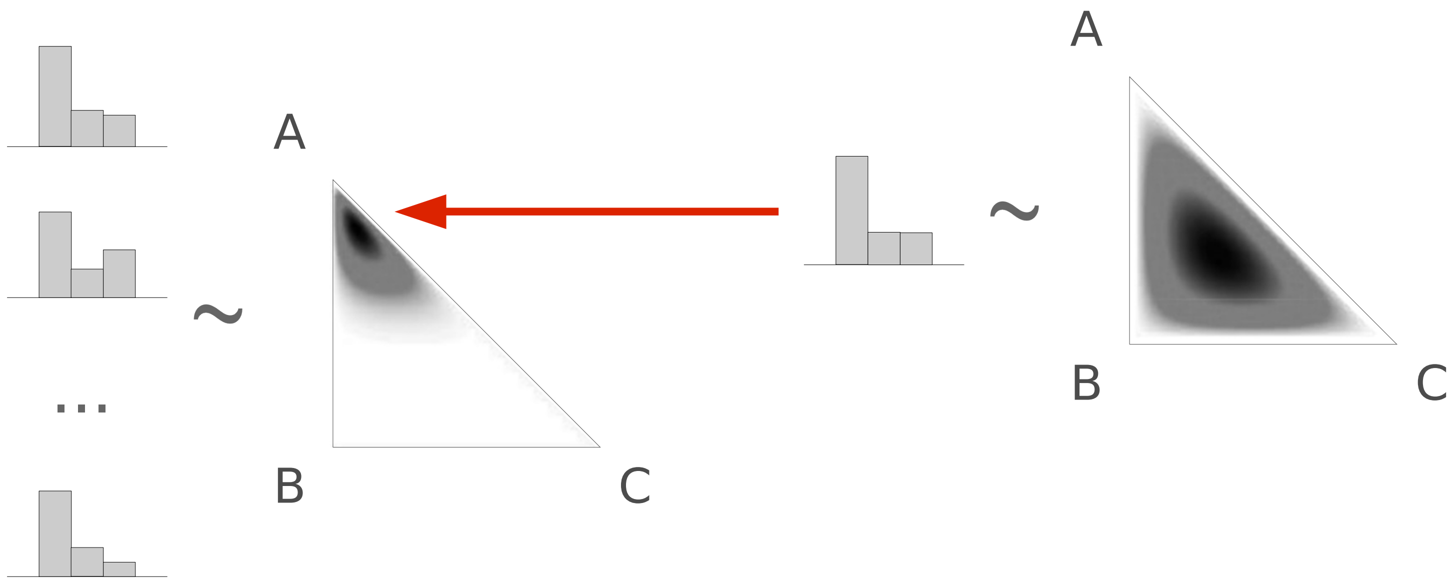
$\Theta \sim \mathrm{Dir}(\alpha \boldsymbol{m})$ $\qquad$ $\Theta \sim \mathrm{Dir}(\alpha \boldsymbol{m})$



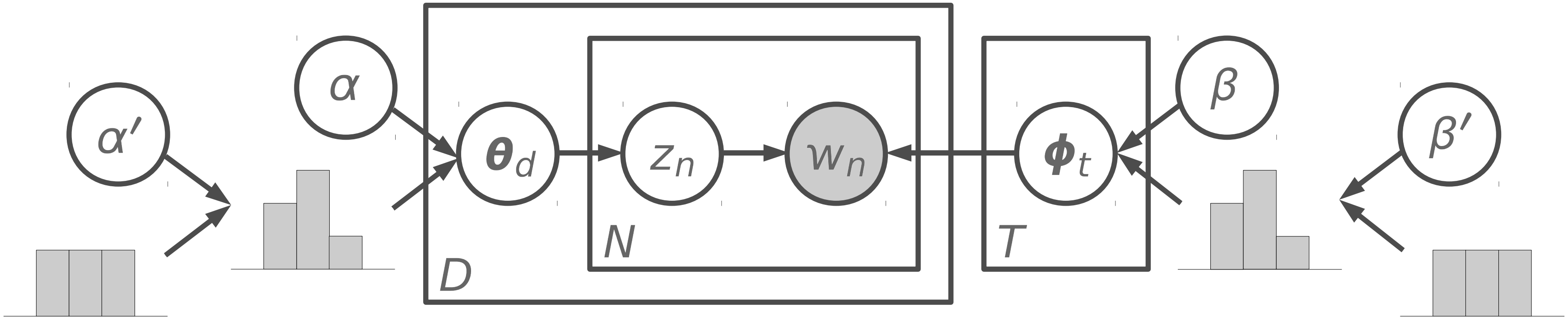
- Asymmetric prior: some topics more likely a priori

Hierarchical Asymmetric Dirichlet

- Which topics should be more probable a priori?
    - Draw ***m*** from a Dirichlet distribution:



Putting Everything Together



- Asymmetric hierarchical Dirichlet priors
- Integrate out Θ, Φ and base measures
- Learn $\boldsymbol{z}$ and concentration parameters from data

Data Sets

- Carbon nanotechnology patents:
  - Ultimate goal: track innovation and emergence
  - Fullerene and carbon nanotube patents
  - 1,016 abstracts (~100 words each)
  - 103,499 total words; 6,068 unique words
- 20 Newsgroups data (80,012 total words)
- New York Times articles (477,465 total words)

Inferred Topics

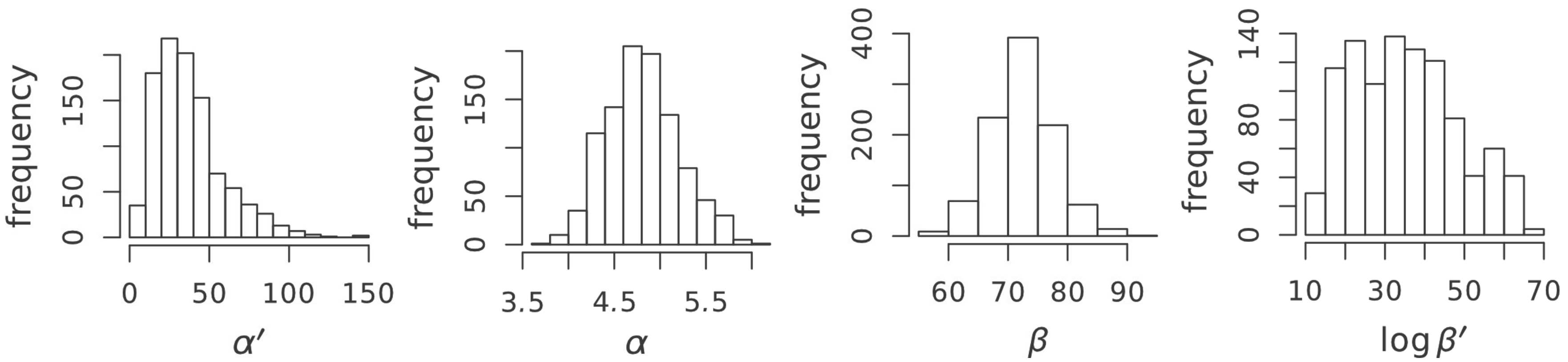
before →

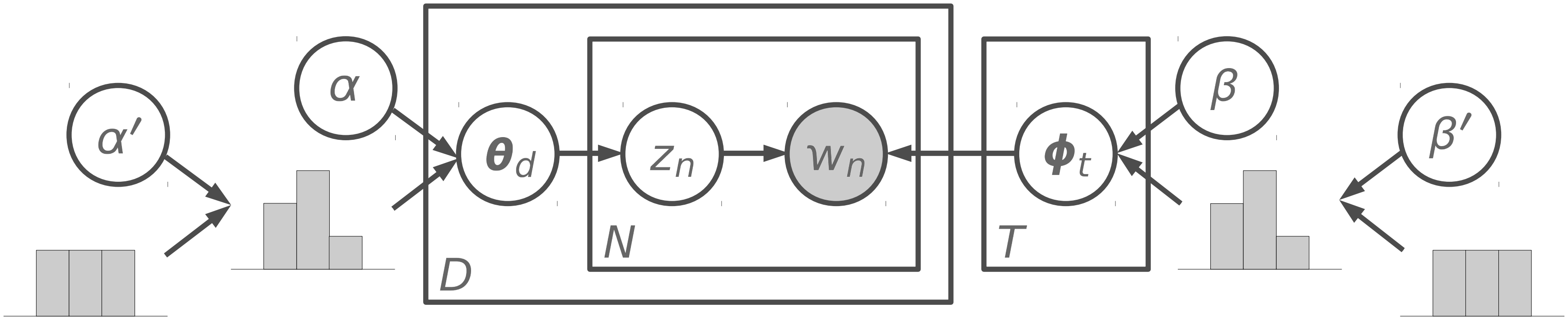
a	a	the	the
field	the	of	**invention**
emission	**carbon**	a	of
an	and	to	to
electron	**gas**	and	**present**
...	...	...	...

after →

the	**carbon**	**metal**	**composite**
a	**nanotubes**	**catalytic**	**polymer**
of	**nanotube**	**transition**	**matrix**
to	**catalyst**	**catalyst**	**weight**
and	**substrate**	from	**fiber**
...	...	...	...

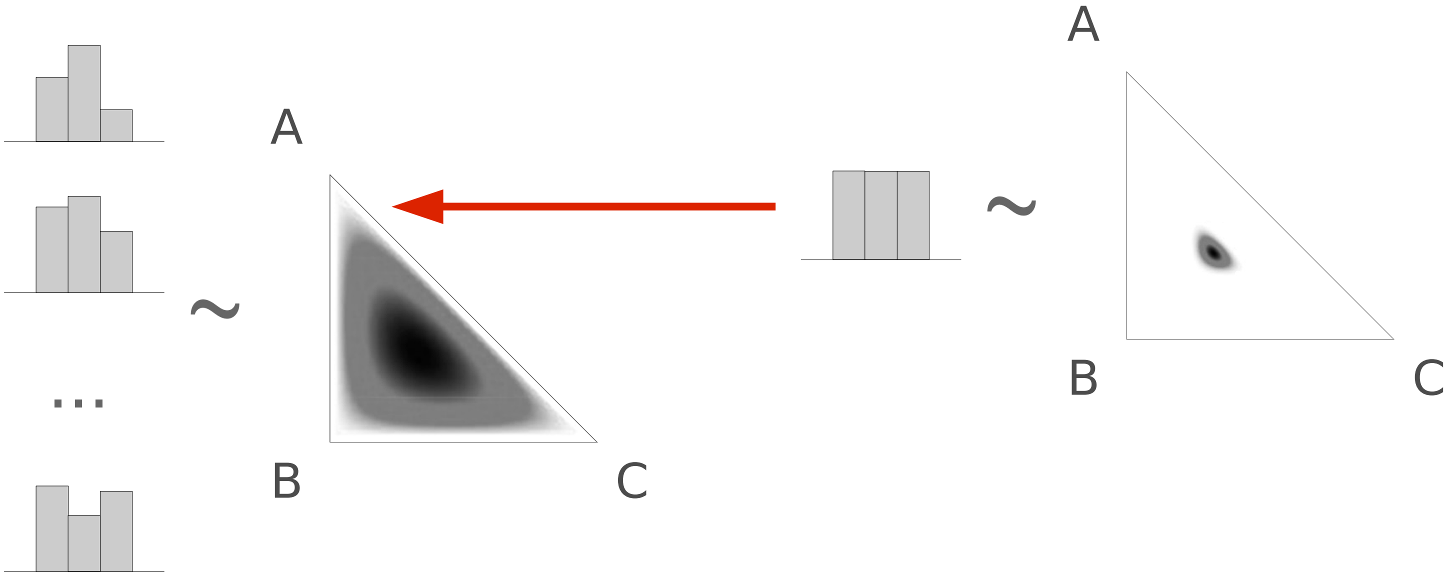
Sampled Concentration Parameters



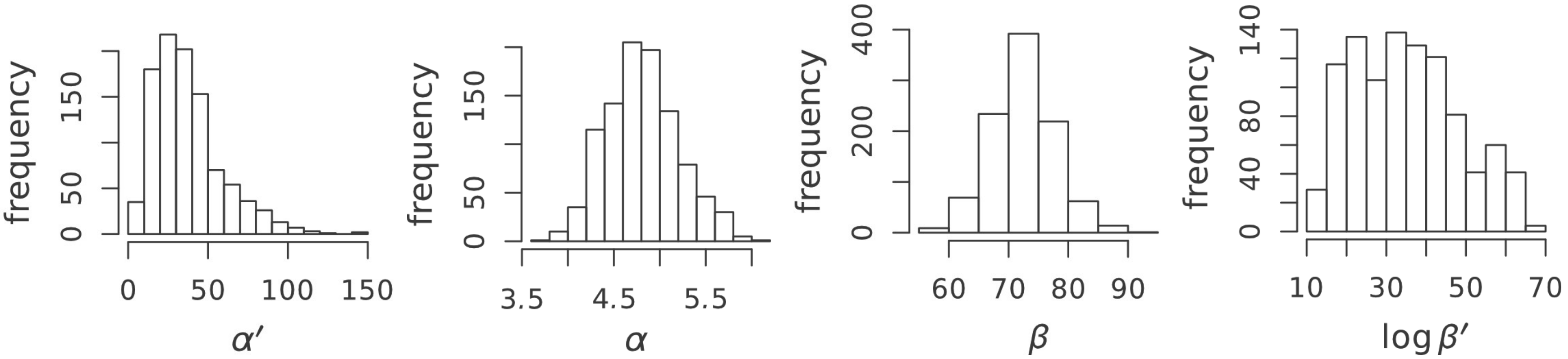


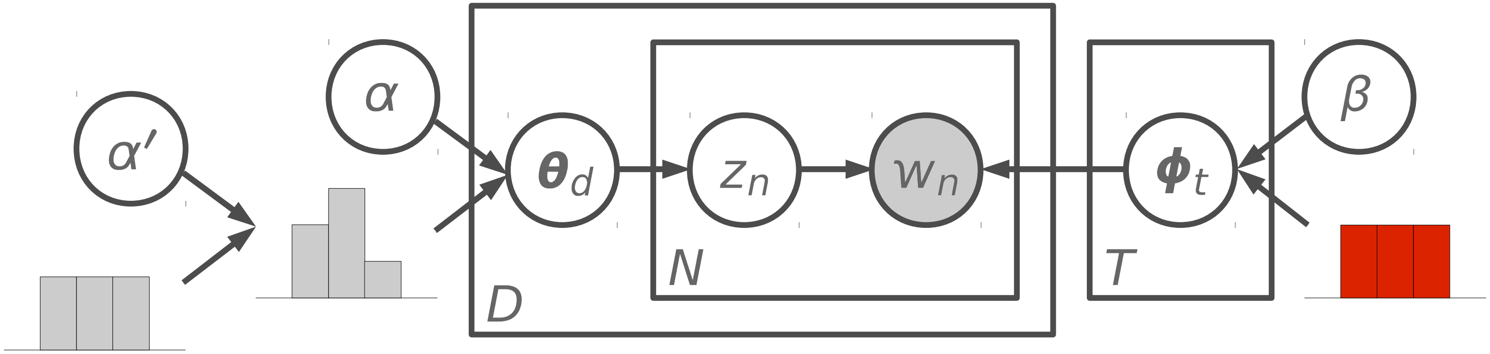
A Theoretical Observation...

- Symmetric Dirichlet is a special case of the hierarchical asymmetric Dirichlet (large concentration parameter)



Sampled Concentration Parameters





Intuition

- Topics should be distinct from each other:
  - Asymmetric prior over topics makes topics more similar to each other (and to corpus-wide word frequencies)
  - Want a symmetric prior to preserve topic “distinctness”
- Still have to account for power-law word usage:
  - Asymmetric prior over document-specific topic distributions means some topics (e.g., “the, a, of, to ...”) can be used more often than others in all documents

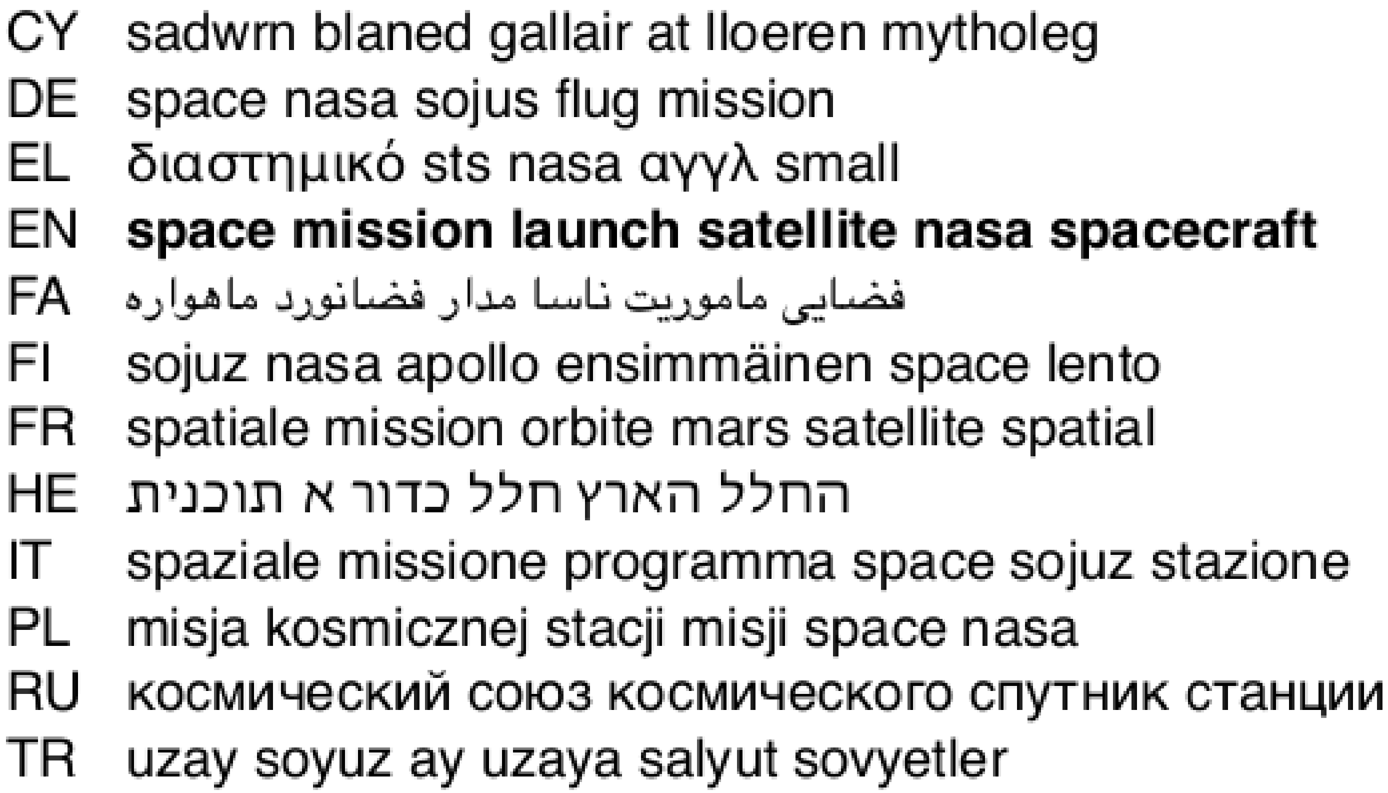
“Off-the-Shelf” Topic Modeling

I can model technology emergence by analyzing patent abstracts!

Great! Let me know if you need any more help!

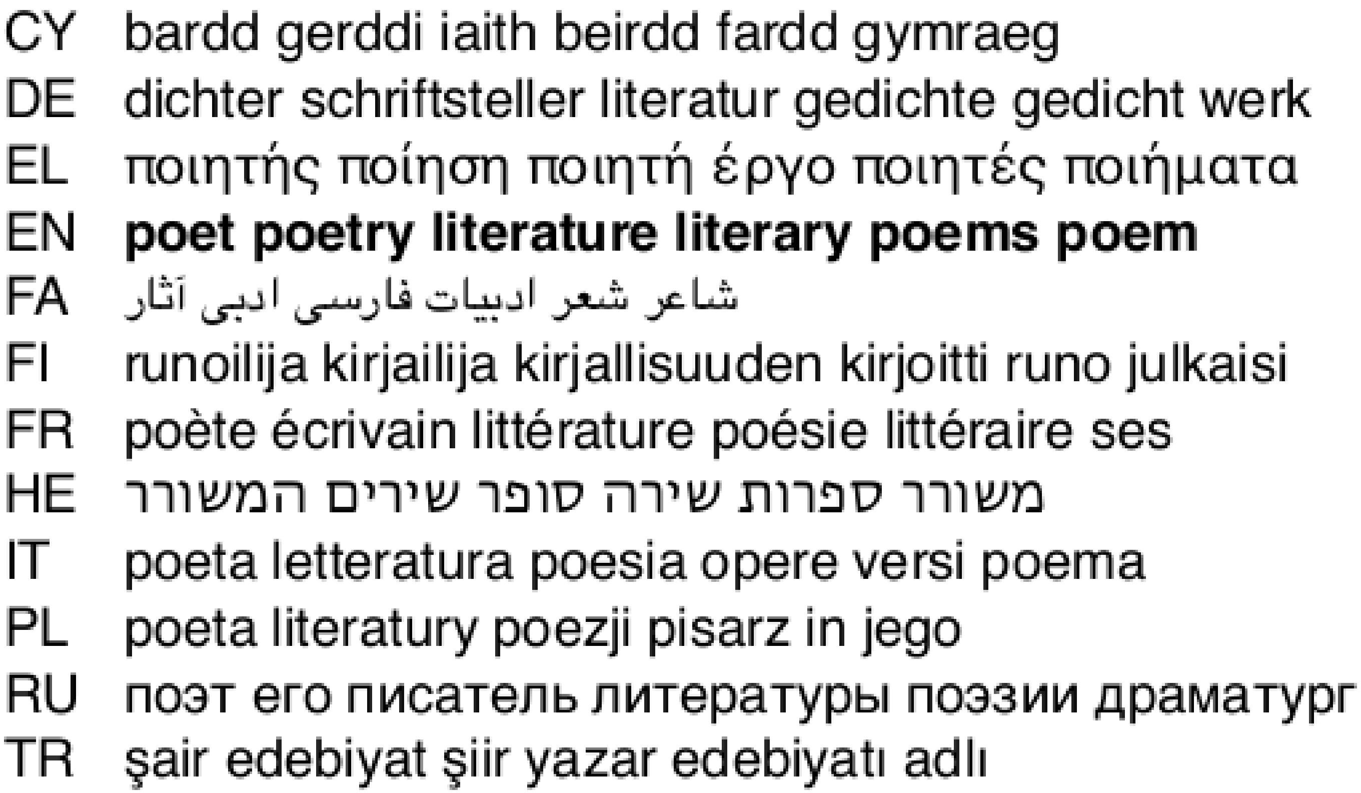
the	**carbon**	**metal**	**composite**
a	**nanotubes**	**catalytic**	**polymer**
of	**nanotube**	**transition**	**matrix**
to	**catalyst**	**catalyst**	**weight**
and	**substrate**	from	**fiber**
...	...	...	...

Polylingual Topics



CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضايي ماموريت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

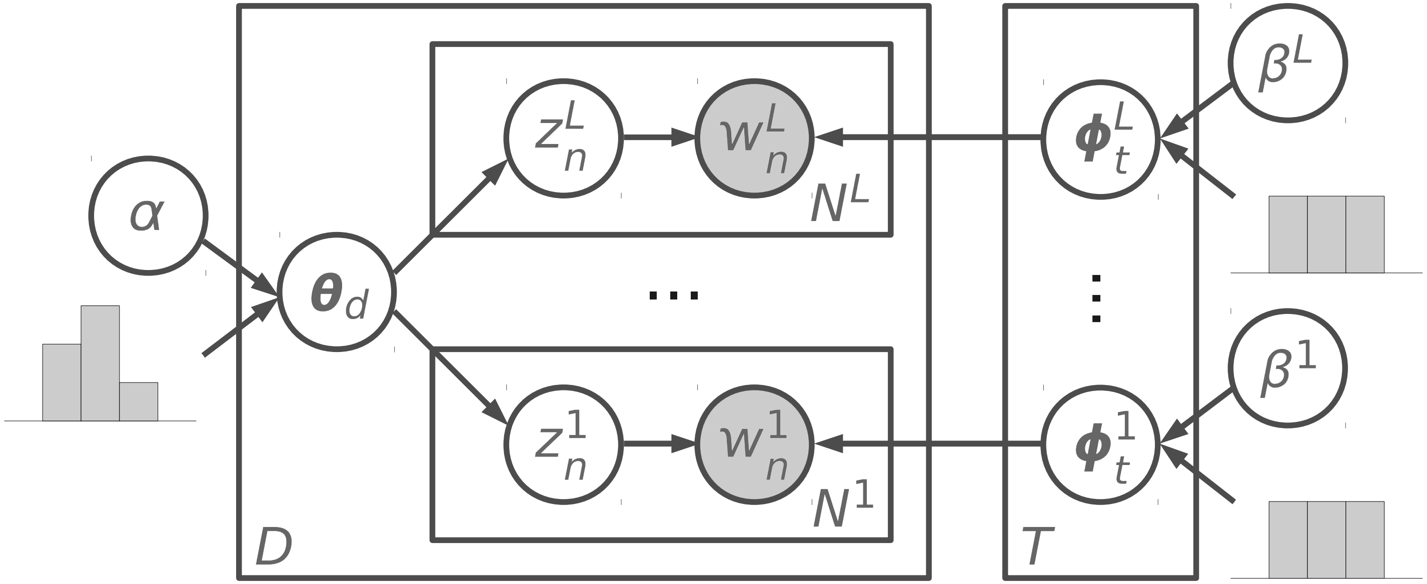
Polylingual Topics



CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses
HE	משורר ספרות שירה סופר שירים המשורר
IT	poeta letteratura poesia opere versi poema
PL	poeta literatury poezji pisarz in jego
RU	поэт его писатель литературы поэзии драматург
TR	şair edebiyat şiir yazar edebiyatı adlı

Polylingual Topic Model

[Mimno et al., '09]



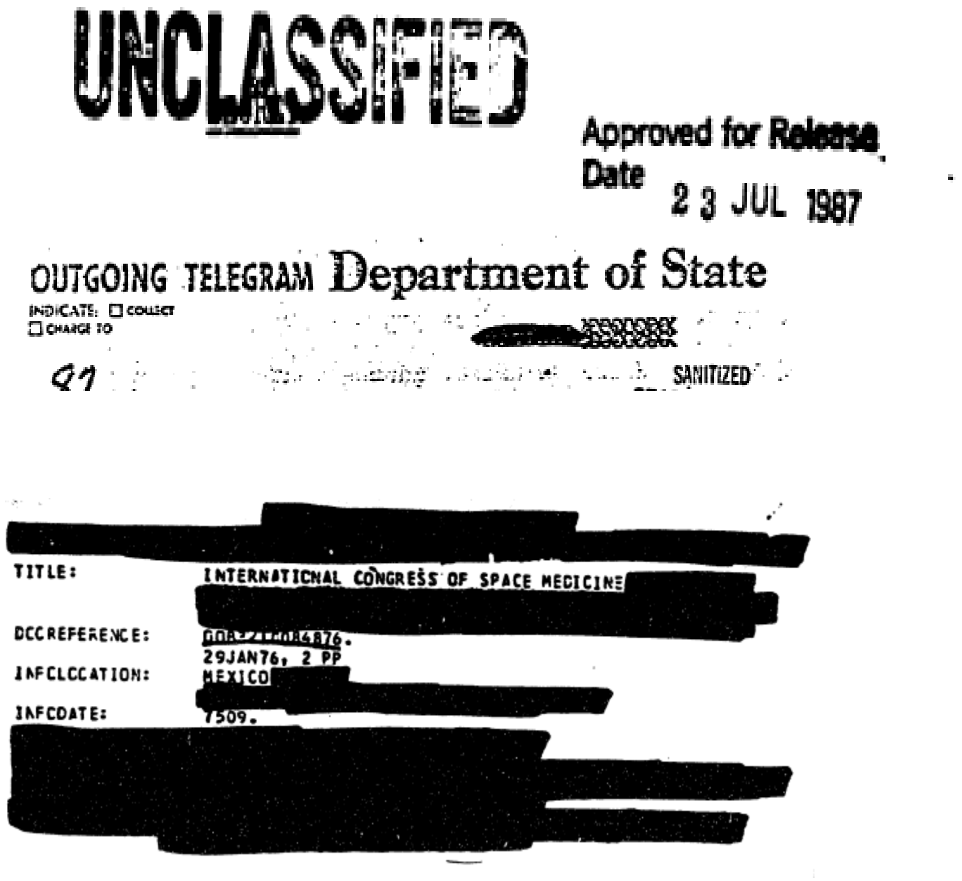
“tuple” of
aligned documents

language-specific
Dirichlet parameters

This Talk

- Statistical topic models for text analysis
- “Off-the-shelf” topic models: priors, stop words
- Studying formerly-classified government documents

In 2009 Alone...



UNCLASSIFIED

Approved for Release
Date 2 3 JUL 1987

OUTGOING TELEGRAM Department of State

INDICATE: □ COLLECT
□ CHARGE TO

[redacted]

SANITIZED

[redacted]

TITLE: INTERNATICNAL CONGRESS OF SPACE MECICINE [redacted]

DCCREFERENCE: [illegible]4876.
29JAN76, 2 PP

INFCLCCATION: MEXICO [redacted]

INFCDATE: 7509.

[redacted]

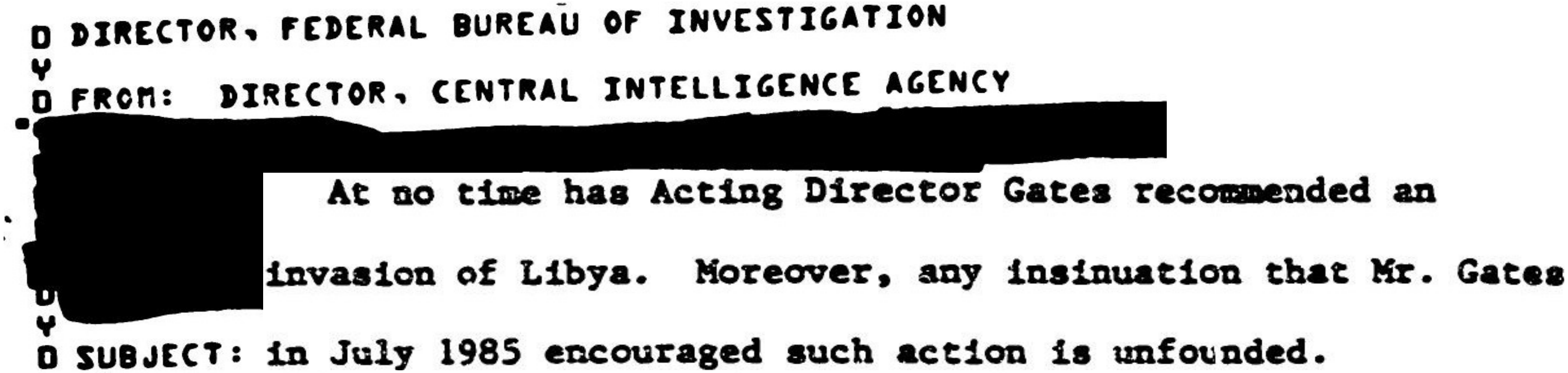
- 52 million pages reviewed for declassification
- 29 million pages declassified
- $8.8 billion spent on administration of the US government classification system

How Sensitive?

> “After a 14-year legal battle by a California history professor, the FBI has released a new cache of material from a 300-page dossier on the late rock star John Lennon, and has agreed to pay $204,000 to cover legal fees incurred in his efforts to open the file. For all the years of challenge, however, the file contains little, if any, new information about Lennon, though it does present some bizarre details, like a description of an antiwar activist trying to train a parrot to speak profanities.”

— NYT, 25 September 2007

A Problematic Trade-off



D DIRECTOR, FEDERAL BUREAU OF INVESTIGATION
Y
D FROM: DIRECTOR, CENTRAL INTELLIGENCE AGENCY

[redacted]

At no time has Acting Director Gates recommended an invasion of Libya. Moreover, any insinuation that Mr. Gates

Y
D SUBJECT: in July 1985 encouraged such action is unfounded.

- The more data kept secret, the less secure the data:
  - More people need to have access to the data
  - More storage space is required

What We Are *NOT* Studying...

guardian.co.uk

News | Sport | Comment | Culture | Business | Money

News › World news › Afghanistan: The war logs

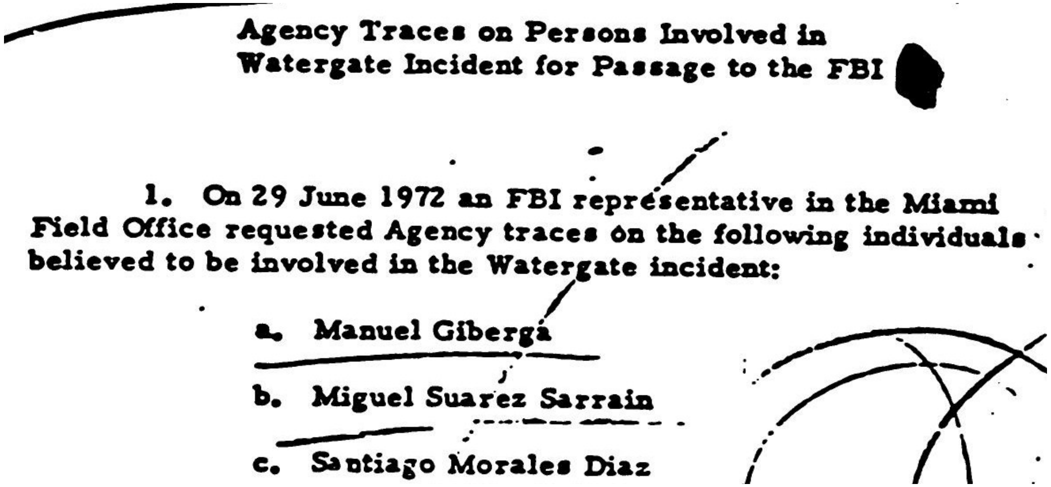
AFGHANISTAN
THE WAR LOGS

Exploring Declassified Documents

- Declassification goals:
  - Recommend documents for human review
  - Match documents with human reviewers' expertise
- Transparency research goals:
  - High-level characterization of the data
  - Finding specific, known information of interest
  - Finding “interesting” or “unexpected” information

Declassified Documents: DDRS

- ~88,000 formerly-classified government documents
- Created and declassified between 1926 and 2005



Agency Traces on Persons Involved in
Watergate Incident for Passage to the FBI

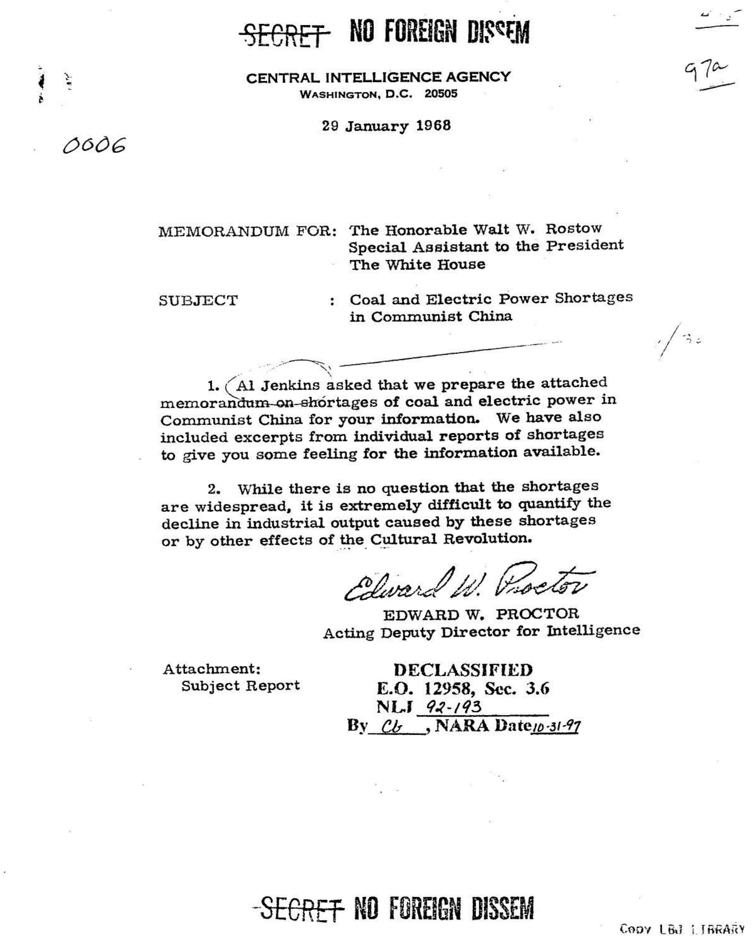
1. On 29 June 1972 an FBI representative in the Miami Field Office requested Agency traces on the following individuals believed to be involved in the Watergate incident:

a. Manuel Giberga

b. Miguel Suarez Sarrain

c. Santiago Morales Diaz

Available Information



~~SECRET~~ NO FOREIGN DISSEM

97a

CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505

29 January 1968

0006

MEMORANDUM FOR: The Honorable Walt W. Rostow
Special Assistant to the President
The White House

SUBJECT : Coal and Electric Power Shortages
in Communist China

1. Al Jenkins asked that we prepare the attached memorandum on shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

Edward W. Proctor
EDWARD W. PROCTOR
Acting Deputy Director for Intelligence

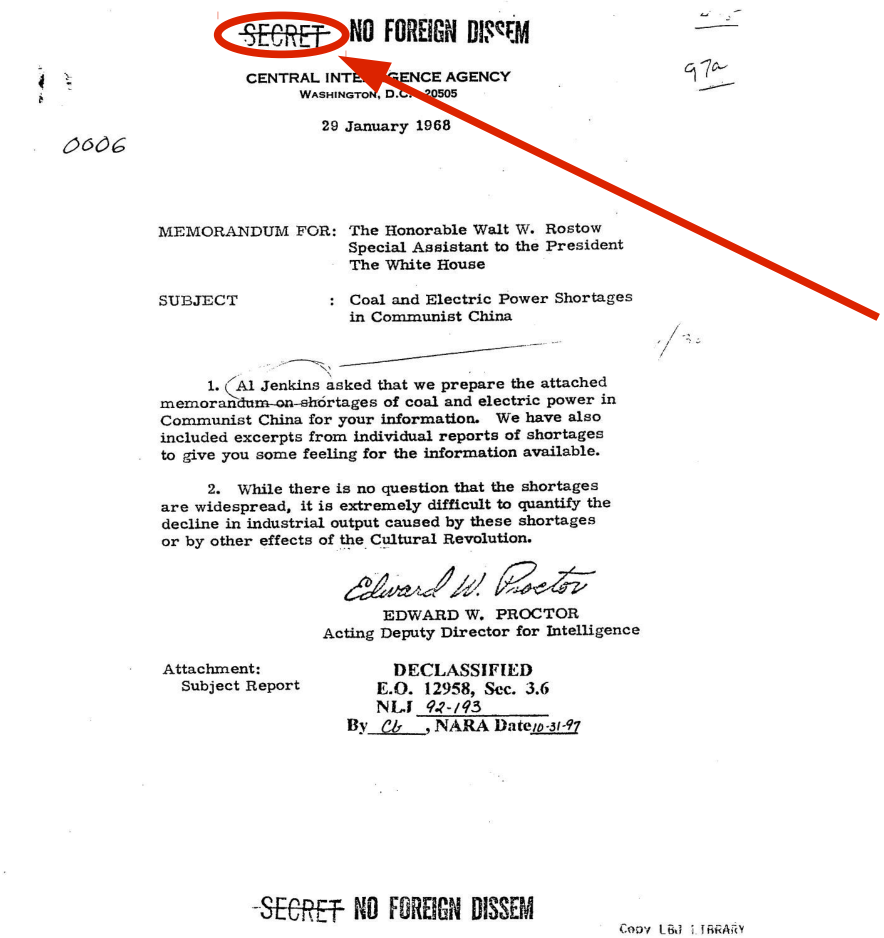
Attachment:
Subject Report

DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb , NARA Date 10-31-97

~~SECRET~~ NO FOREIGN DISSEM

- Sanitized?
- Classification level
- Issuer
- Creation date
- Document type
- Declassification date

Available Information



~~SECRET~~ NO FOREIGN DISSEM

97a

CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505

29 January 1968

0006

MEMORANDUM FOR: The Honorable Walt W. Rostow
Special Assistant to the President
The White House

SUBJECT : Coal and Electric Power Shortages in Communist China

1. Al Jenkins asked that we prepare the attached ~~memorandum on~~ shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

Edward W. Proctor
EDWARD W. PROCTOR
Acting Deputy Director for Intelligence

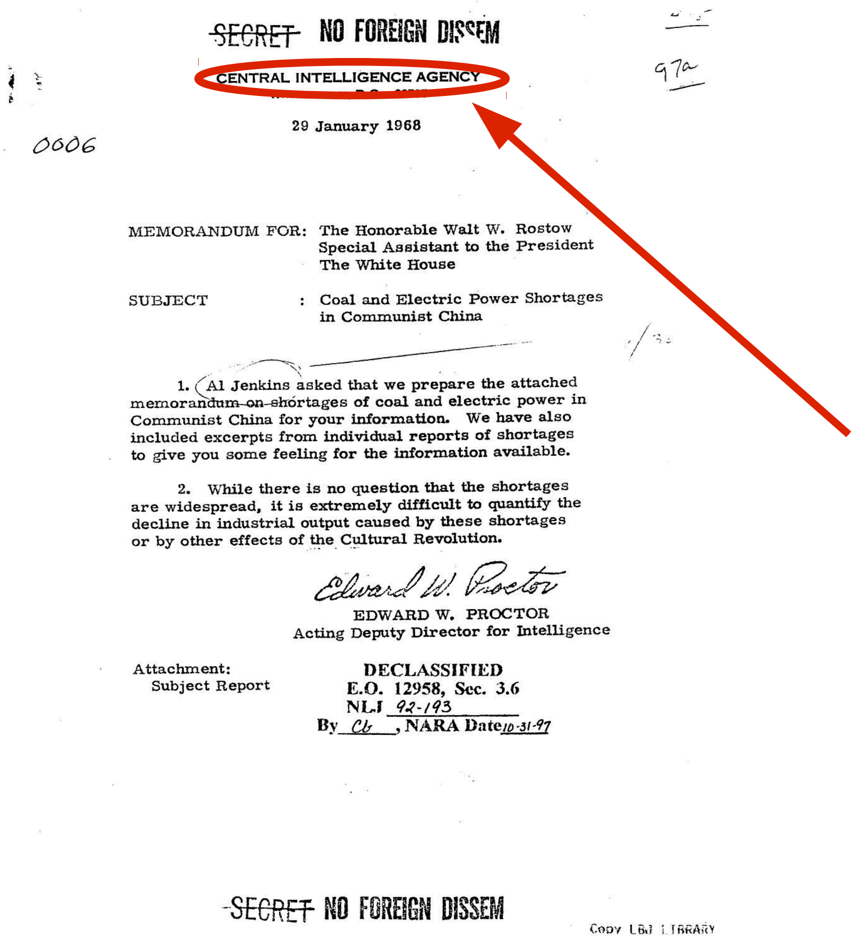
Attachment:
Subject Report

DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb , NARA Date 10-31-97

~~SECRET~~ NO FOREIGN DISSEM

- Sanitized?
- Classification level
- Issuer
- Creation date
- Document type
- Declassification date

Available Information



~~SECRET~~ NO FOREIGN DISSEM

97a

CENTRAL INTELLIGENCE AGENCY

0006

29 January 1968

MEMORANDUM FOR: The Honorable Walt W. Rostow
Special Assistant to the President
The White House

SUBJECT : Coal and Electric Power Shortages in Communist China

1. Al Jenkins asked that we prepare the attached ~~memorandum on~~ shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

Edward W. Proctor
EDWARD W. PROCTOR
Acting Deputy Director for Intelligence

Attachment:
Subject Report

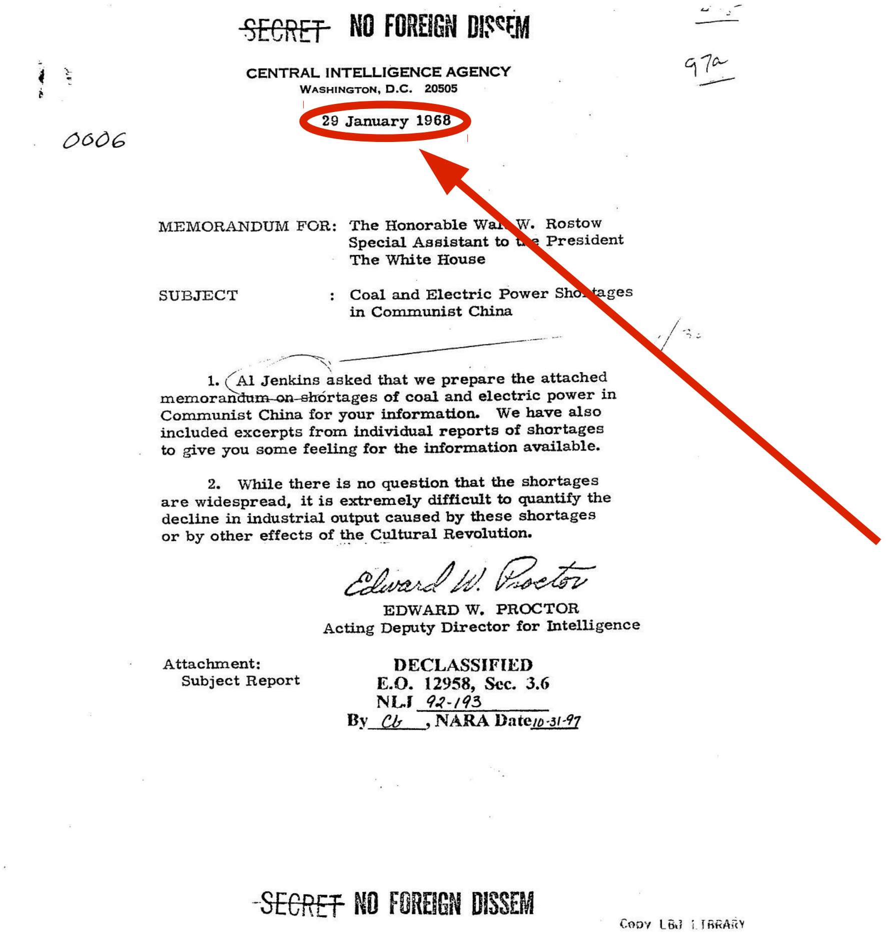
DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb , NARA Date 10-31-97

~~SECRET~~ NO FOREIGN DISSEM

Copy LBJ LIBRARY

- Sanitized?
- Classification level
- Issuer
- Creation date
- Document type
- Declassification date

Available Information



~~SECRET~~ NO FOREIGN DISSEM

97a

CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505

0006

29 January 1968

MEMORANDUM FOR: The Honorable Walt W. Rostow
Special Assistant to the President
The White House

SUBJECT : Coal and Electric Power Shortages
in Communist China

1. Al Jenkins asked that we prepare the attached memoran~~dum on~~ shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

Edward W. Proctor
EDWARD W. PROCTOR
Acting Deputy Director for Intelligence

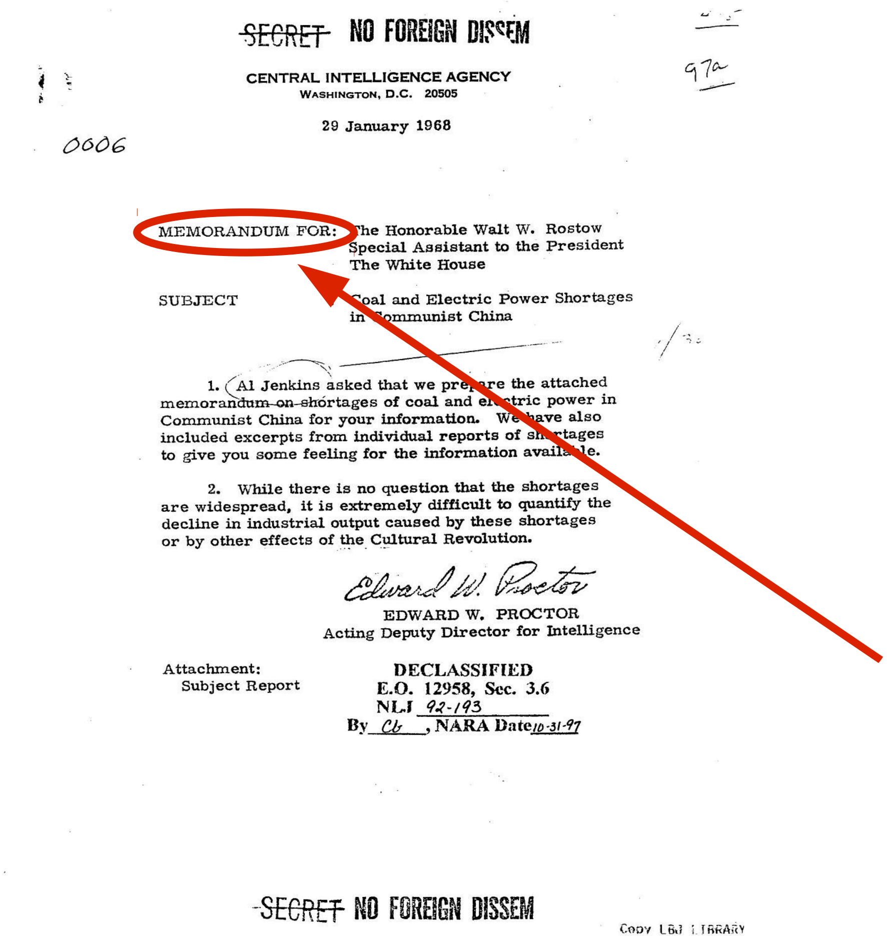
Attachment:
Subject Report

DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb, NARA Date 10-31-97

~~SECRET~~ NO FOREIGN DISSEM

- Sanitized?
- Classification level
- Issuer
- Creation date
- Document type
- Declassification date

Available Information



~~SECRET~~ NO FOREIGN DISSEM

97a

CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505

29 January 1968

0006

MEMORANDUM FOR: The Honorable Walt W. Rostow
Special Assistant to the President
The White House

SUBJECT Coal and Electric Power Shortages
in Communist China

1. Al Jenkins asked that we prepare the attached memorandum ~~on~~ shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

Edward W. Proctor
EDWARD W. PROCTOR
Acting Deputy Director for Intelligence

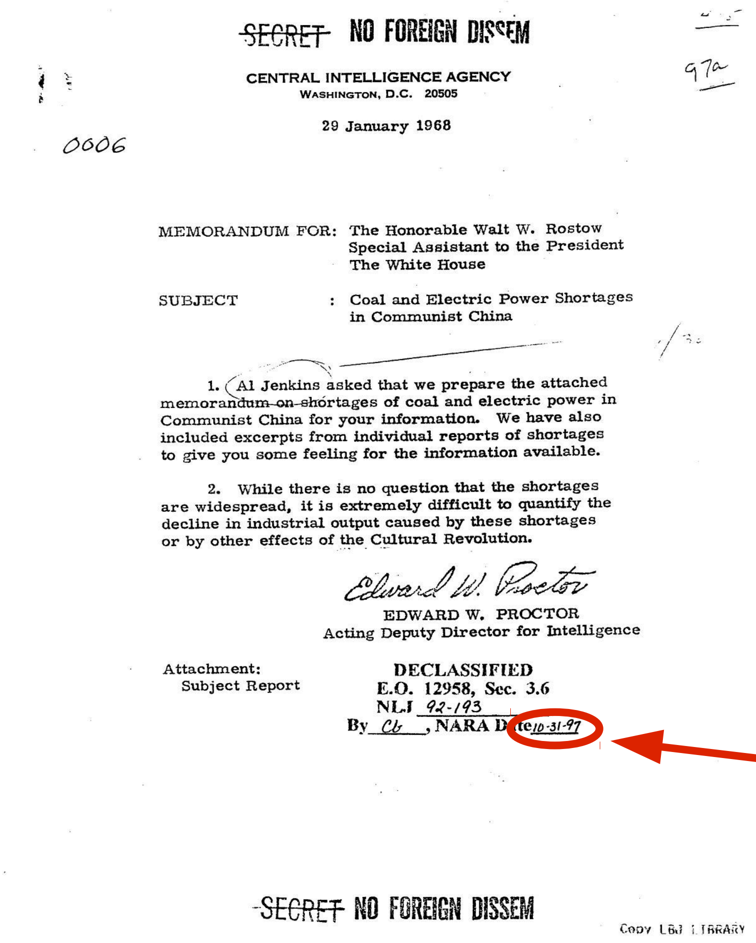
Attachment:
Subject Report

DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb, NARA Date 10-31-97

~~SECRET~~ NO FOREIGN DISSEM

- Sanitized?
- Classification level
- Issuer
- Creation date
- Document type
- Declassification date

Available Information



~~SECRET~~ NO FOREIGN DISSEM

97a

CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505

29 January 1968

0006

MEMORANDUM FOR: The Honorable Walt W. Rostow
Special Assistant to the President
The White House

SUBJECT : Coal and Electric Power Shortages
in Communist China

1. Al Jenkins asked that we prepare the attached memorandum on shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

Edward W. Proctor
EDWARD W. PROCTOR
Acting Deputy Director for Intelligence

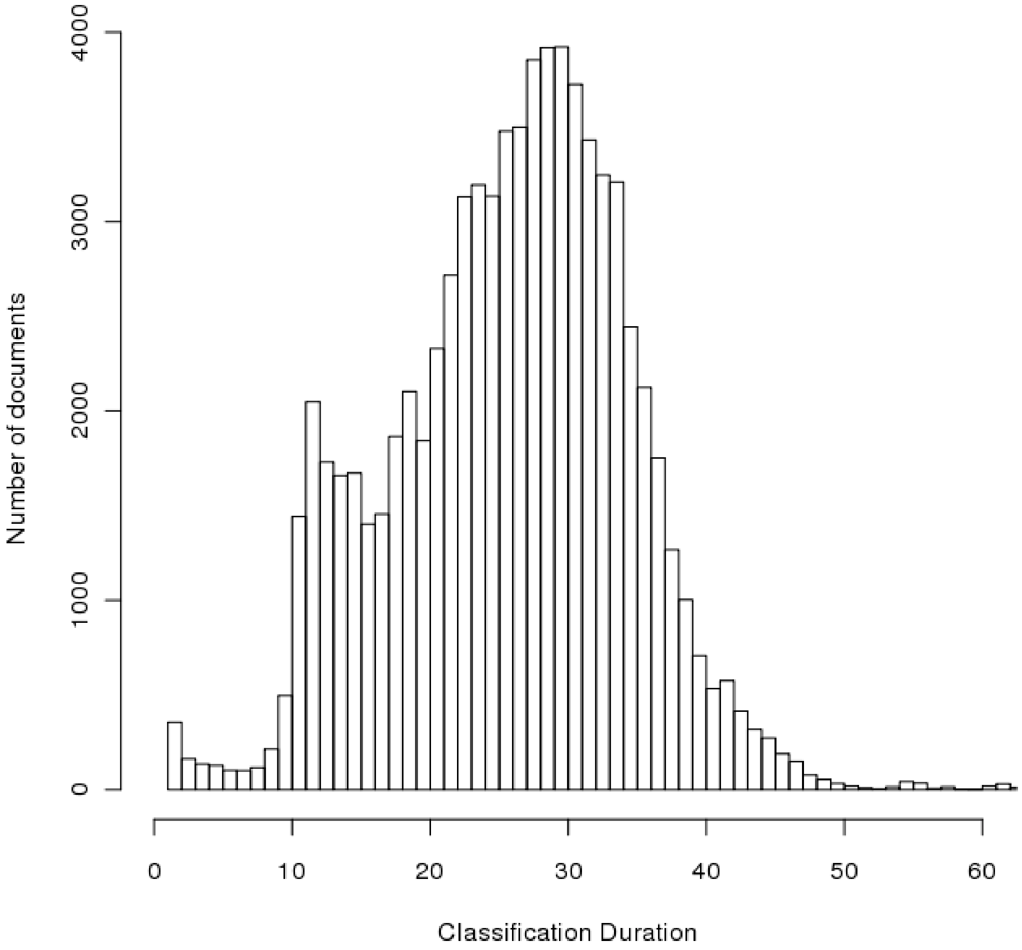
Attachment:
Subject Report

DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb, NARA Date 10-31-97

~~SECRET~~ NO FOREIGN DISSEM

- Sanitized?
- Classification level
- Issuer
- Creation date
- Type
- Declassification date

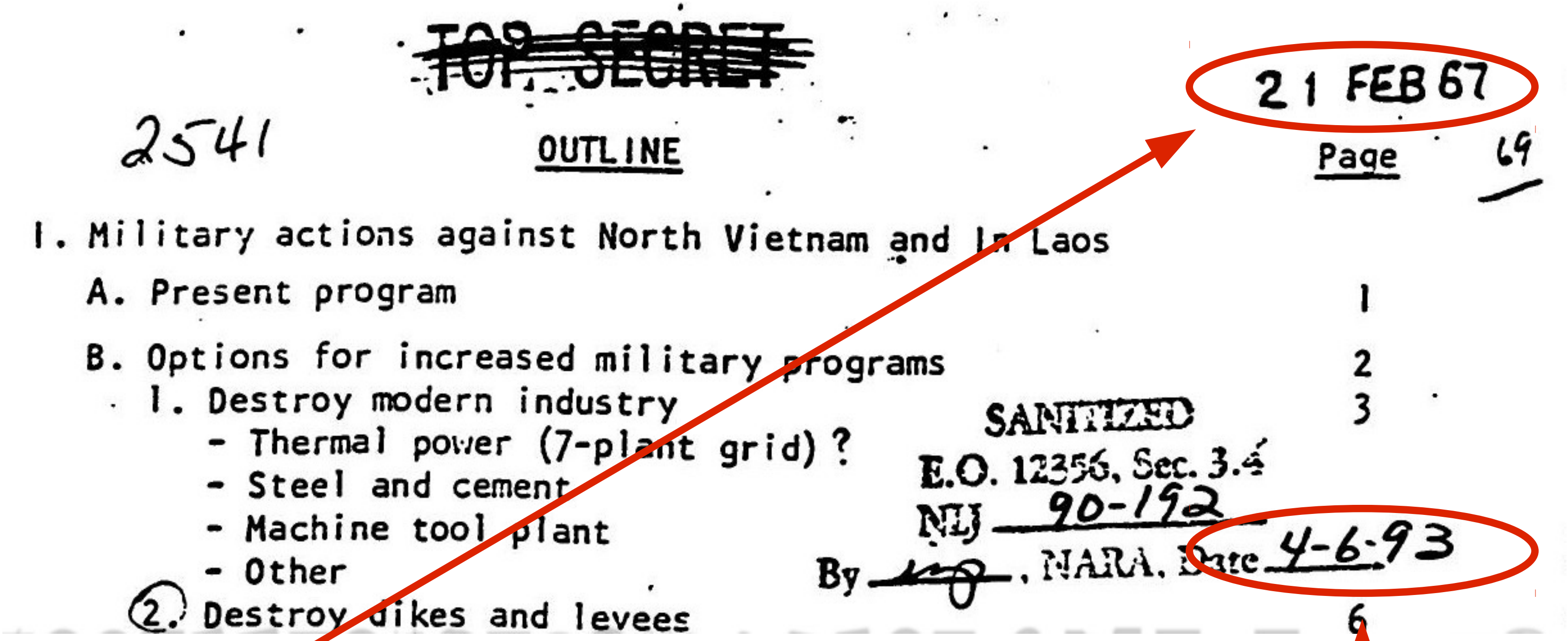
Declassification Durations



Survival Analysis

- Statistical methods for evaluating “time until death”:
  - Biology/medicine: organism death
  - Engineering: component failure
  - Social science: event durations (e.g., parolee recidivism)
- Goal: model effect on survival time of covariates, e.g.,
  - Vaccine treatments
  - Temperature differences
  - Job placement or education programs

Document “Survival”



~~TOP SECRET~~

2541

21 FEB 67

OUTLINE

Page 69

I. Military actions against North Vietnam and In Laos

A. Present program 1

B. Options for increased military programs 2

1. Destroy modern industry 3

- Thermal power (7-plant grid) ?
- Steel and cement
- Machine tool plant
- Other

2. Destroy dikes and levees 6

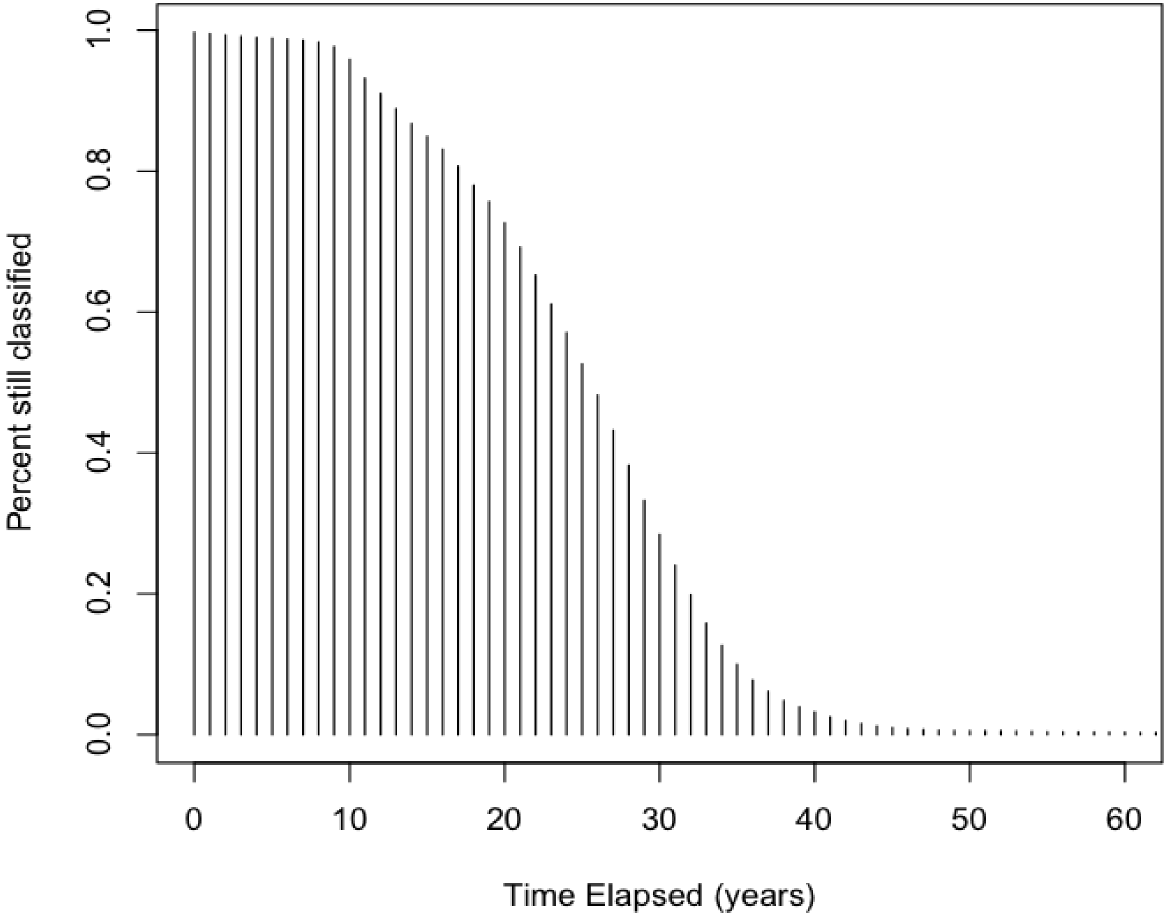
SANITIZED
E.O. 12356, Sec. 3.4
NLJ 90-192
By , NARA, Date 4-6-93

creation date
2/21/67

26 years

declassification date
4/6/93

Survival Distribution of Documents



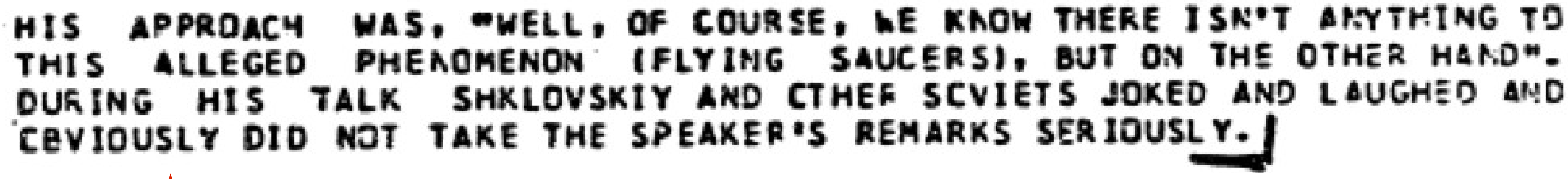
Accelerated Failure Time Models

- Survival analysis with covariates $\boldsymbol{x}_i$
- Linear models for the log of the “duration”:

$$\log(t_i) = \boldsymbol{x}_i{}^\top \boldsymbol{\beta} + \epsilon_i$$

- Parametric: a probability distribution is specified
  - e.g., Weibull, log-normal, gamma, log-logistic...
- Can make predictions for unseen data

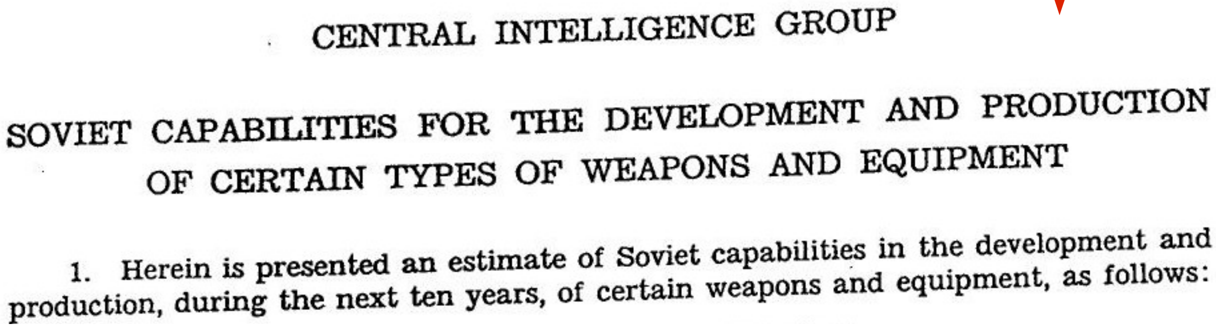
Classification and Content



HIS APPROACH WAS, "WELL, OF COURSE, WE KNOW THERE ISN'T ANYTHING TO THIS ALLEGED PHENOMENON (FLYING SAUCERS), BUT ON THE OTHER HAND". DURING HIS TALK SHKLOVSKIY AND OTHER SOVIETS JOKED AND LAUGHED AND OBVIOUSLY DID NOT TAKE THE SPEAKER'S REMARKS SERIOUSLY.

1975 to 1989

1946 to 2003

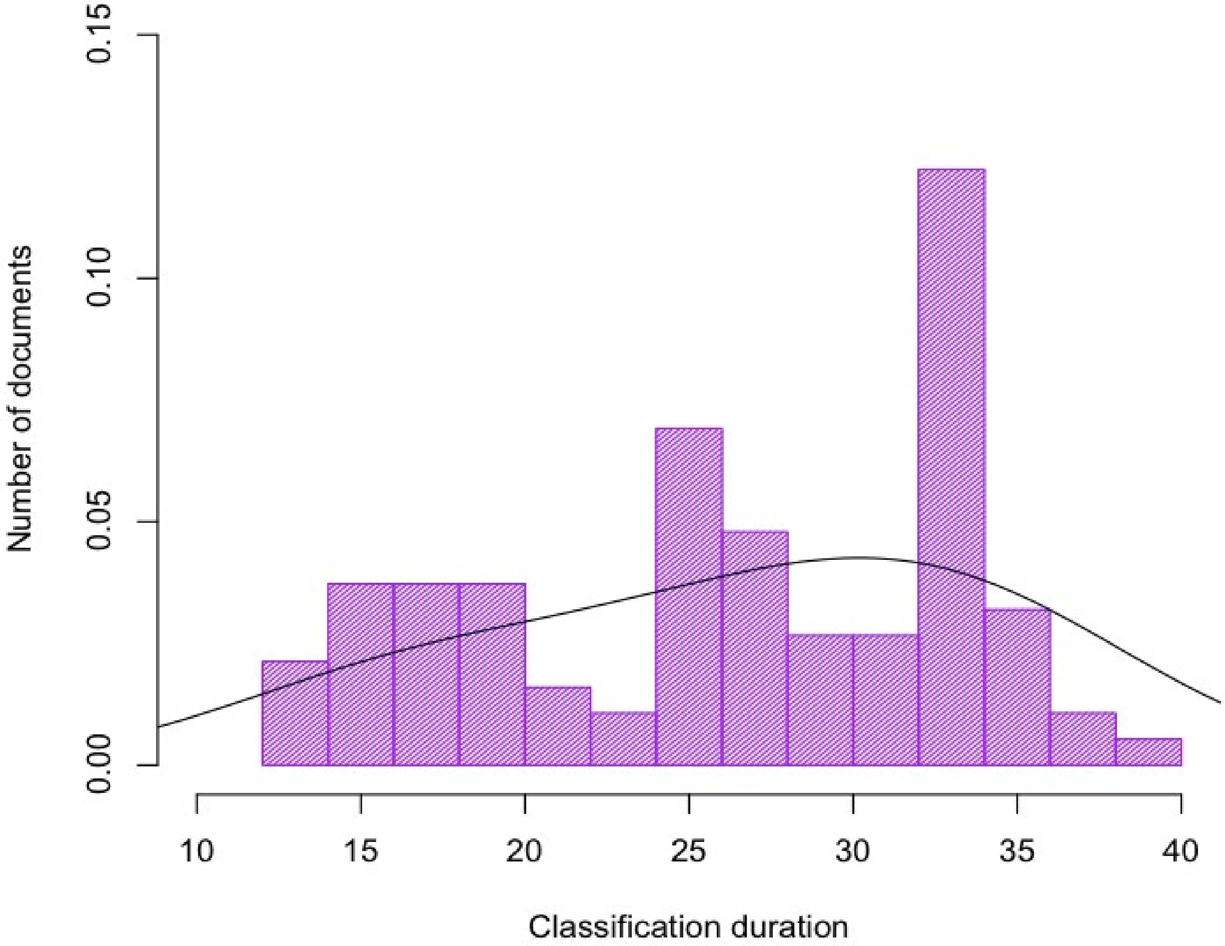


CENTRAL INTELLIGENCE GROUP

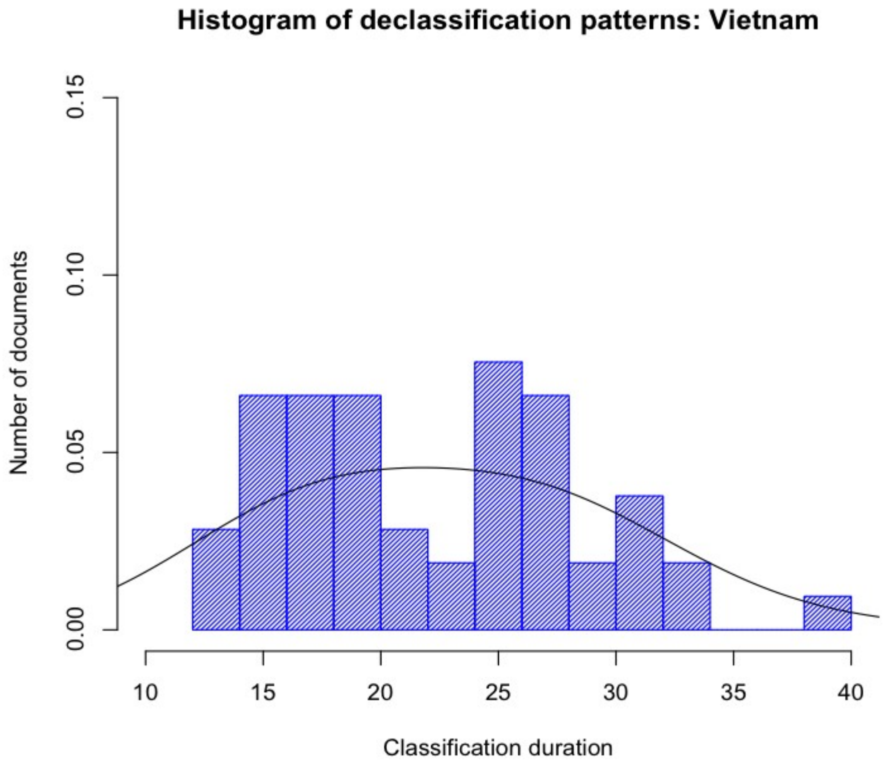
SOVIET CAPABILITIES FOR THE DEVELOPMENT AND PRODUCTION OF CERTAIN TYPES OF WEAPONS AND EQUIPMENT

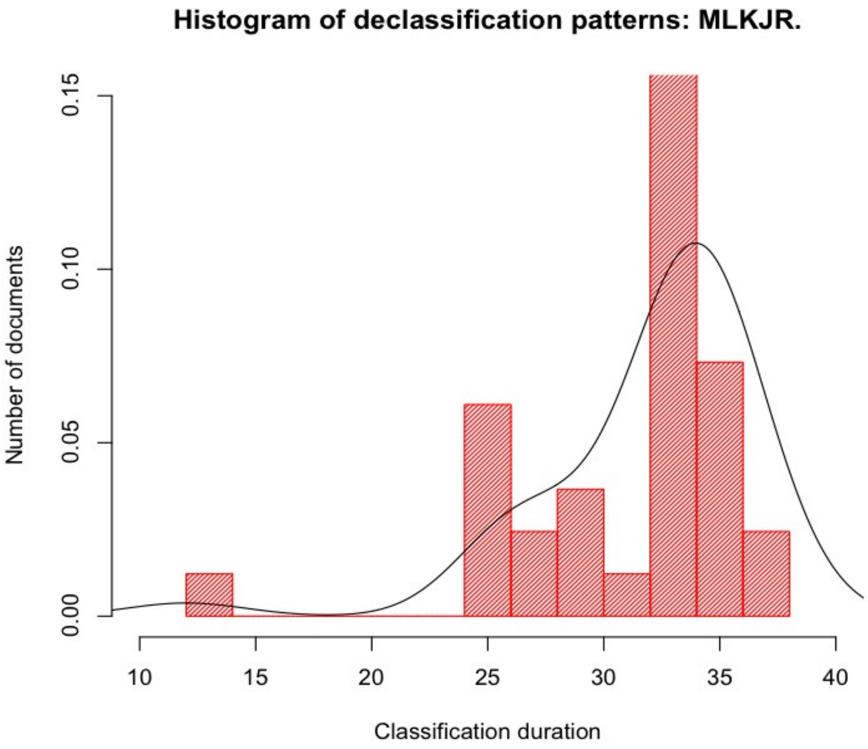
1. Herein is presented an estimate of Soviet capabilities in the development and production, during the next ten years, of certain weapons and equipment, as follows:

Classification and Content: 1960s

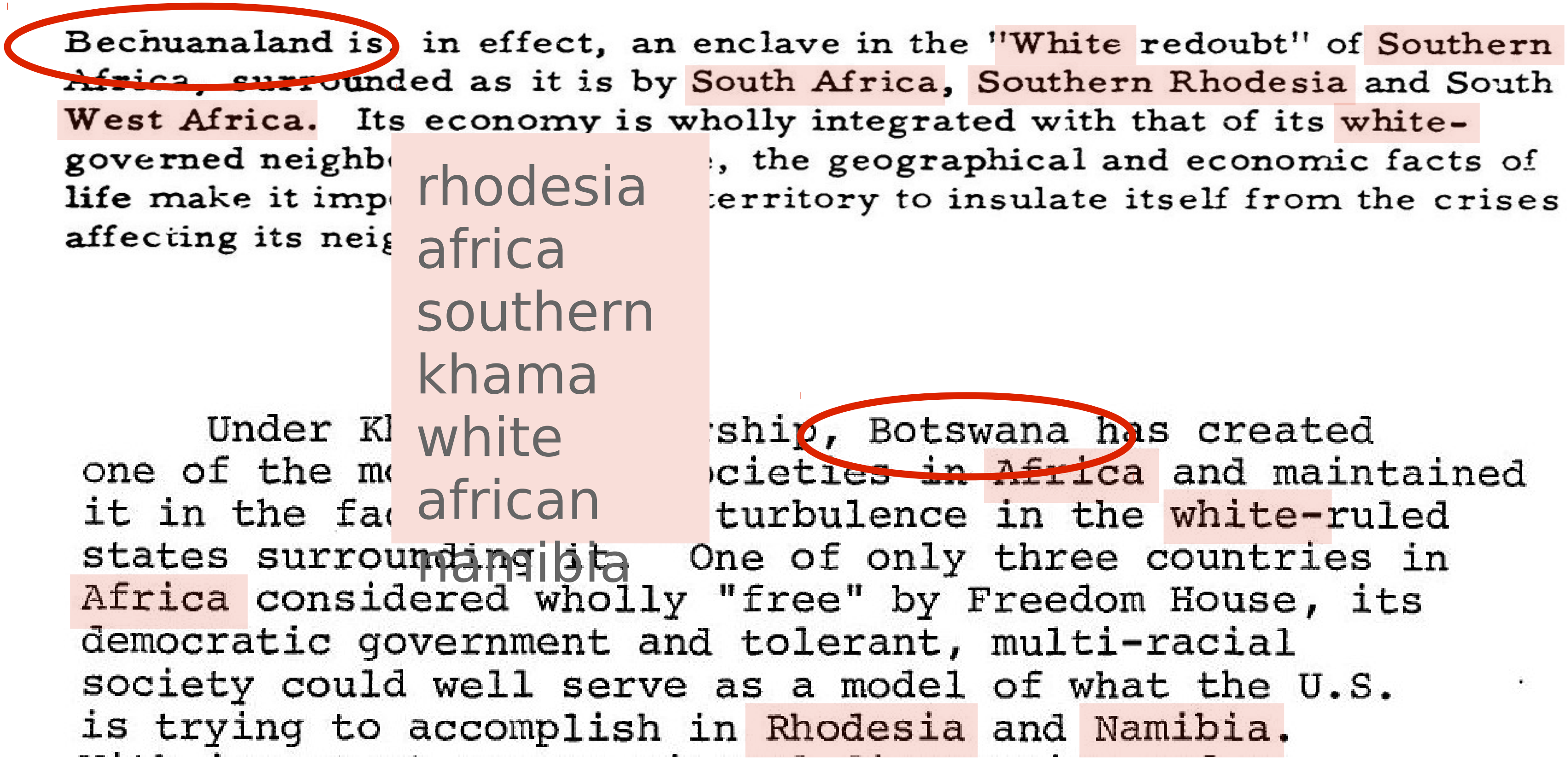


Classification and Content: 1960s


Histogram of declassification patterns: Vietnam


Histogram of declassification patterns: MLKJR.

Word Frequencies?

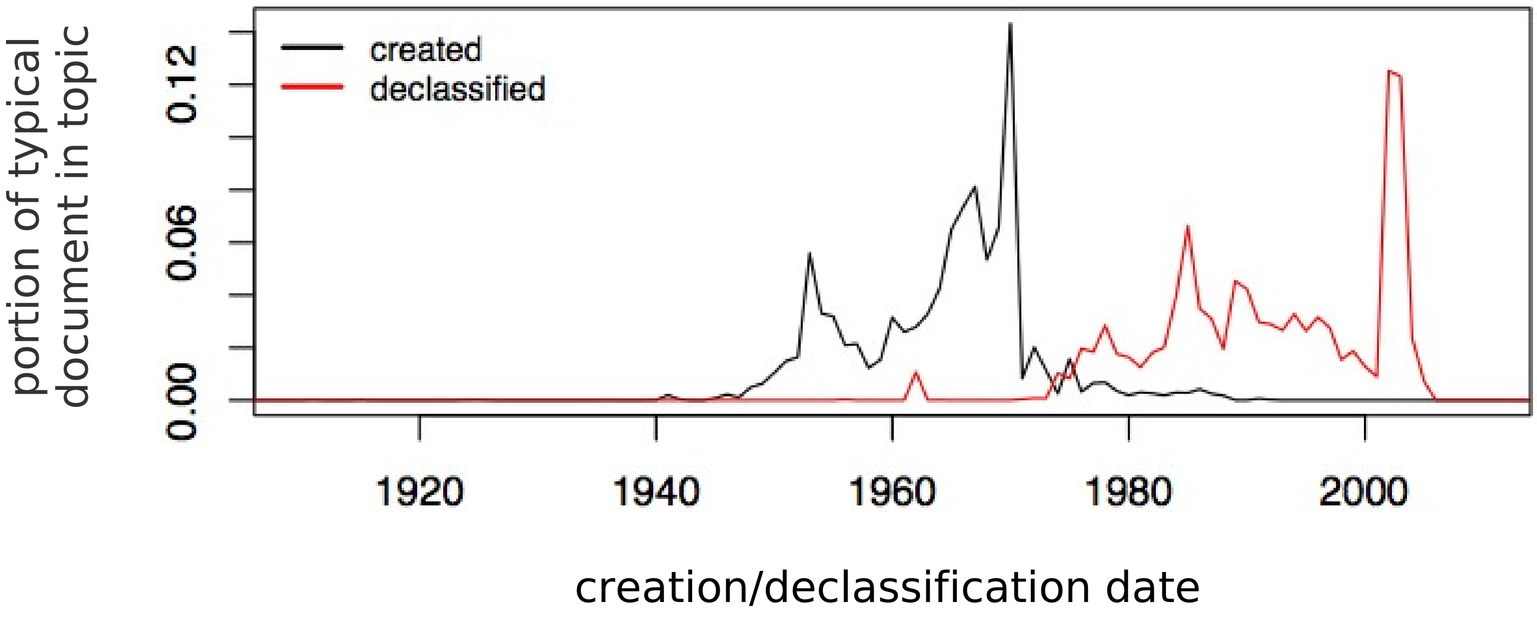


Bechuanaland is, in effect, an enclave in the "White redoubt" of Southern Africa, surrounded as it is by South Africa, Southern Rhodesia and South West Africa. Its economy is wholly integrated with that of its white-governed neighb ,, the geographical and economic facts of life make it imp territory to insulate itself from the crises affecting its neig

rhodesia
africa
southern
khama
white
african
namibia

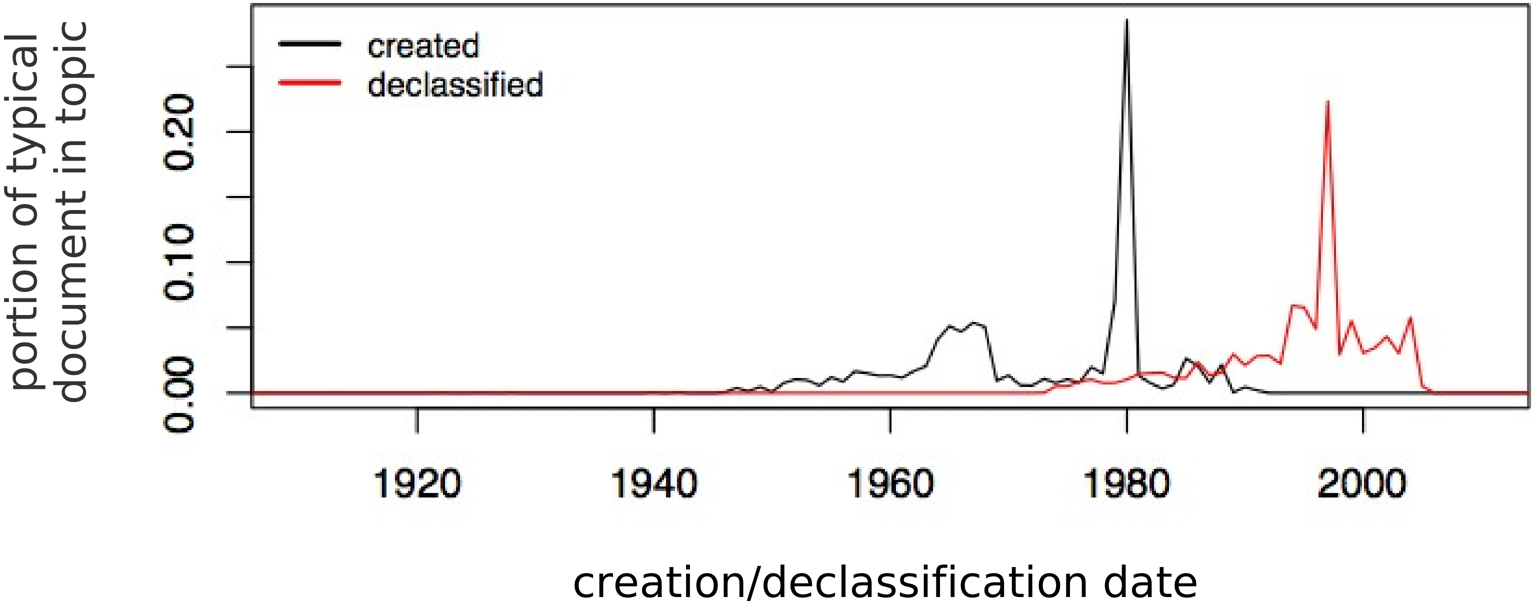
Under Kh rship, Botswana has created one of the mo cieties in Africa and maintained it in the fac turbulence in the white-ruled states surrounding it. One of only three countries in Africa considered wholly "free" by Freedom House, its democratic government and tolerant, multi-racial society could well serve as a model of what the U.S. is trying to accomplish in Rhodesia and Namibia.

Topics in Declassified Documents



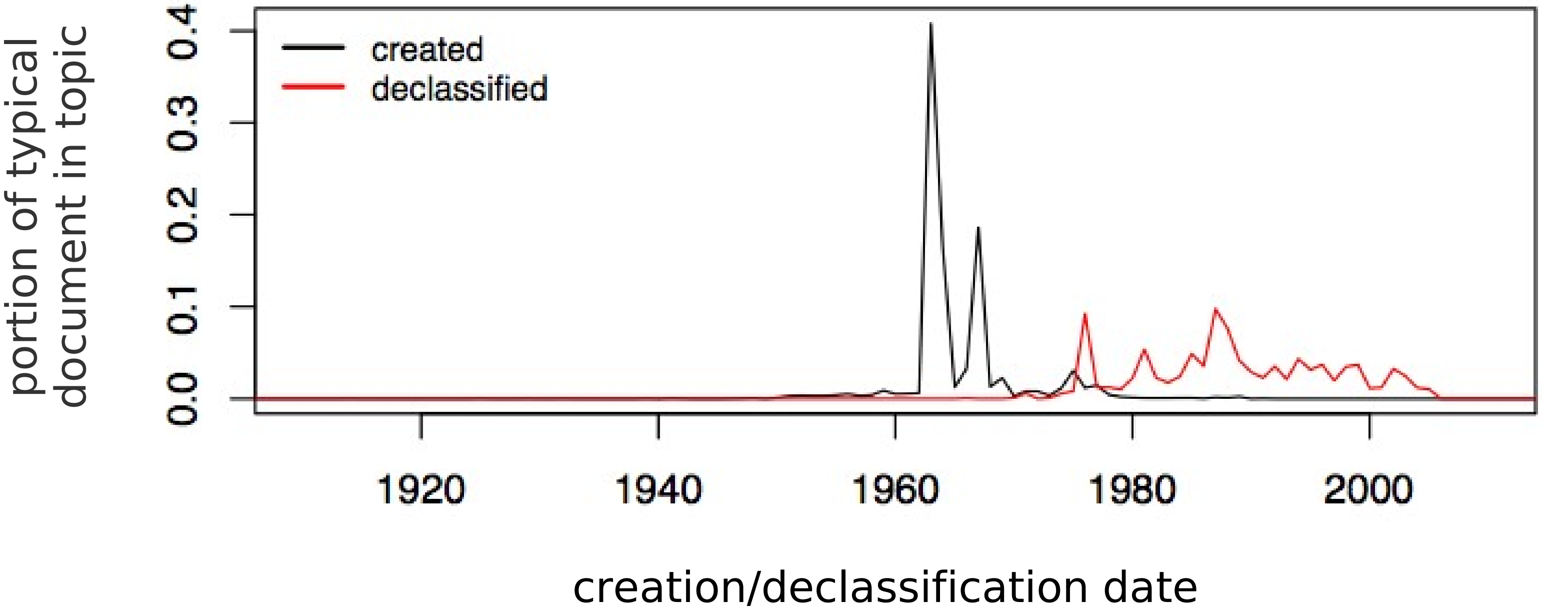
corps, service, volunteers, men, volunteer, age, draft, selective, calls, young, manpower, year, army, deferments, induction, armed, freedom, ...

Topics in Declassified Documents



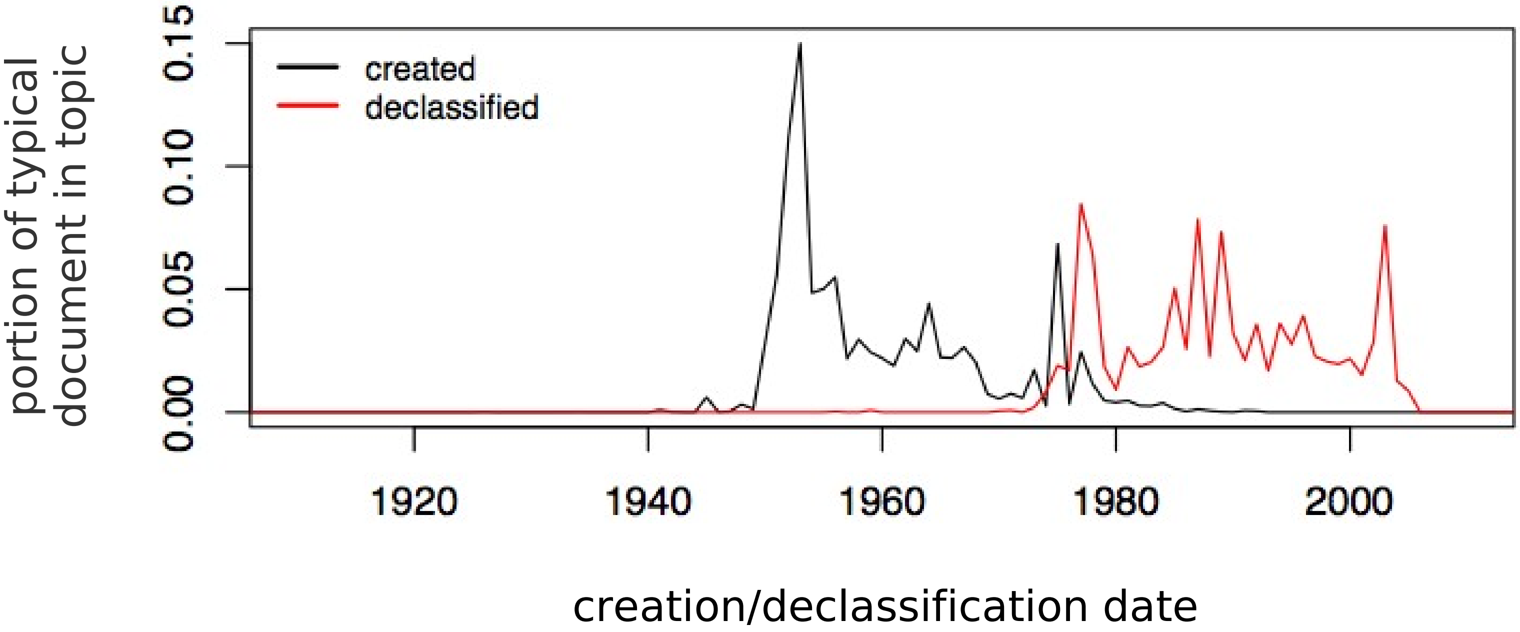
package, hostages, release, hostage, khomeini, packages, ghotbzadeh, held, released, banisadr, revolutionary, debriefing, scenario, family, date, ...

Topics in Declassified Documents



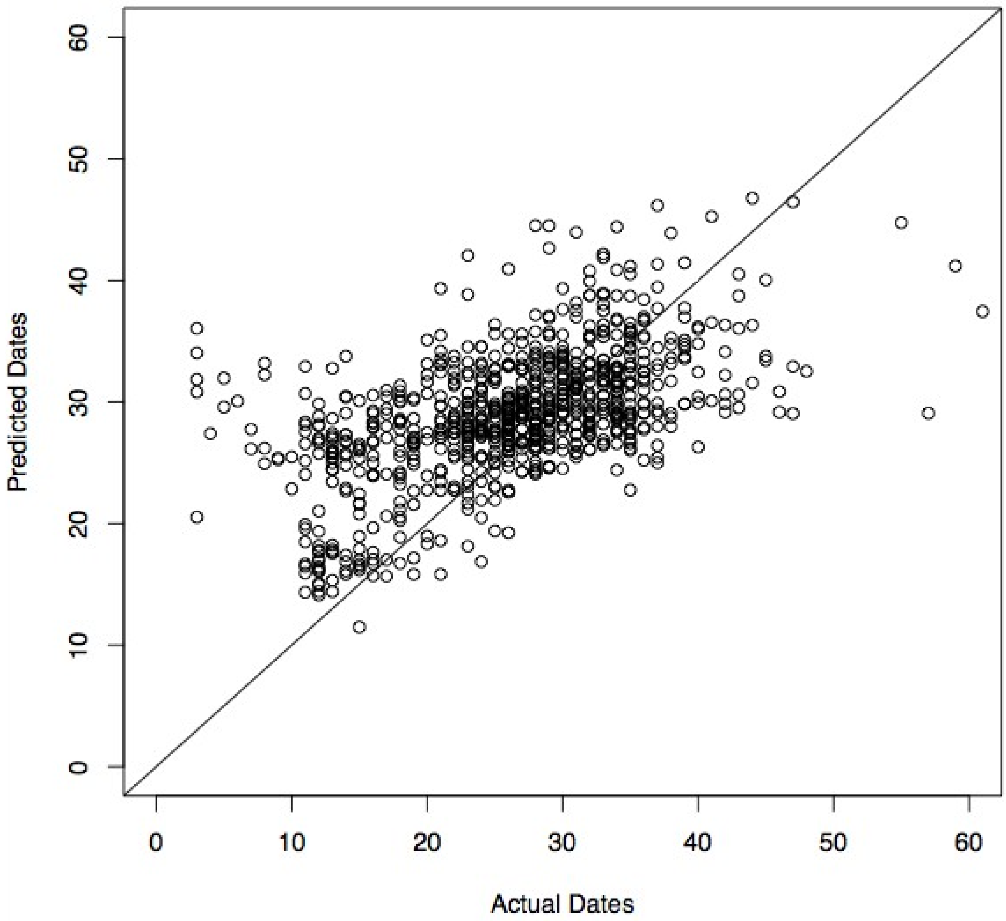
oswald, dallas, assassination, kennedy, texas, fbi, orleans, advised,
lee, president, bureau, started, harvey, john, information, ruby, november, ...

Topics in Declassified Documents

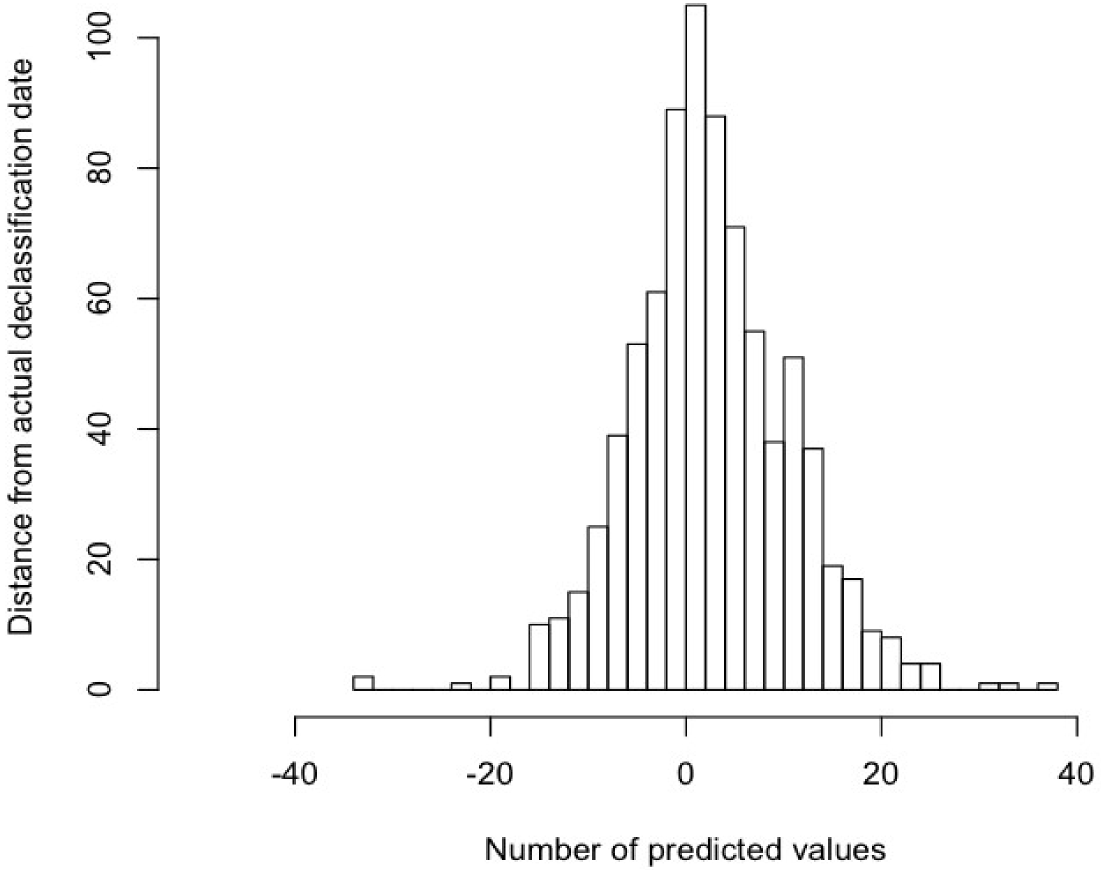


artichoke, subject, drugs, techniques, work, interrogation, writer, drug, lsd, effects, hypnosis, methods, medical, physical, subjects, human, ...

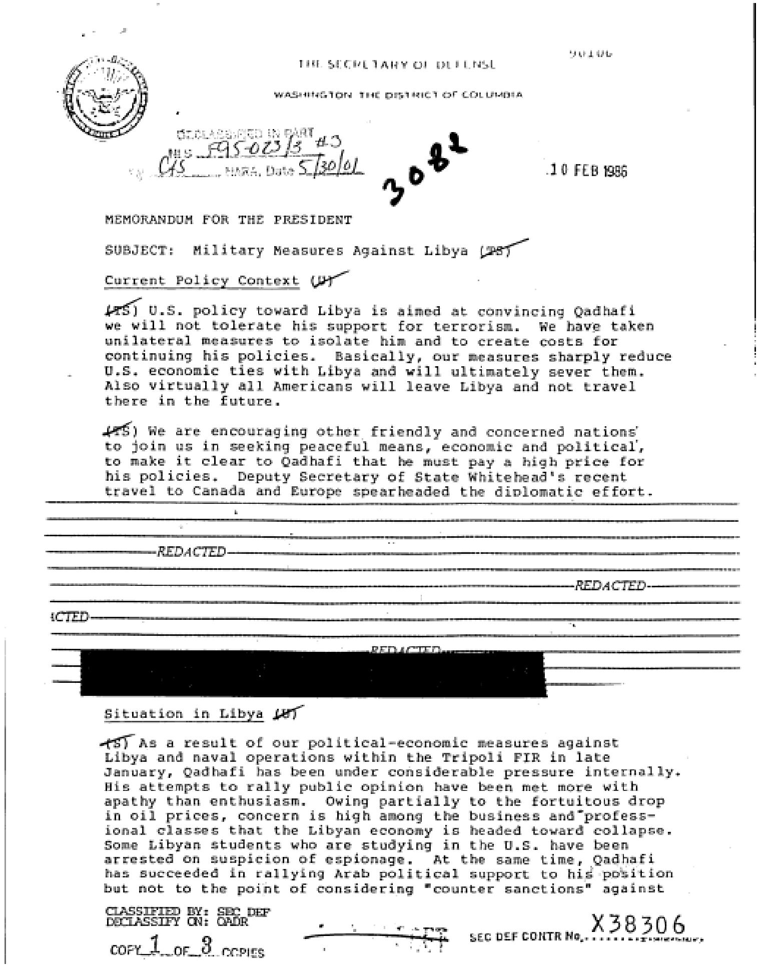
Predicting Duration Using Topics



Predicted Duration - Actual Duration



Jointly Modeling Text and Duration



90106

THE SECRETARY OF DEFENSE

WASHINGTON THE DISTRICT OF COLUMBIA

DECLASSIFIED IN PART
NLS F95-023/3 #3
By CAS NARA, Date 5/30/01

3082

10 FEB 1986

MEMORANDUM FOR THE PRESIDENT

SUBJECT: Military Measures Against Libya (~~TS~~)

Current Policy Context (~~U~~)

(~~TS~~) U.S. policy toward Libya is aimed at convincing Qadhafi we will not tolerate his support for terrorism. We have taken unilateral measures to isolate him and to create costs for continuing his policies. Basically, our measures sharply reduce U.S. economic ties with Libya and will ultimately sever them. Also virtually all Americans will leave Libya and not travel there in the future.

(~~TS~~) We are encouraging other friendly and concerned nations to join us in seeking peaceful means, economic and political, to make it clear to Qadhafi that he must pay a high price for his policies. Deputy Secretary of State Whitehead's recent travel to Canada and Europe spearheaded the diplomatic effort.

REDACTED

REDACTED

CTED

[redacted]

Situation in Libya (~~U~~)

(~~S~~) As a result of our political-economic measures against Libya and naval operations within the Tripoli FIR in late January, Qadhafi has been under considerable pressure internally. His attempts to rally public opinion have been met more with apathy than enthusiasm. Owing partially to the fortuitous drop in oil prices, concern is high among the business and professional classes that the Libyan economy is headed toward collapse. Some Libyan students who are studying in the U.S. have been arrested on suspicion of espionage. At the same time, Qadhafi has succeeded in rallying Arab political support to his position but not to the point of considering "counter sanctions" against

CLASSIFIED BY: SEC DEF
DECLASSIFY ON: OADR

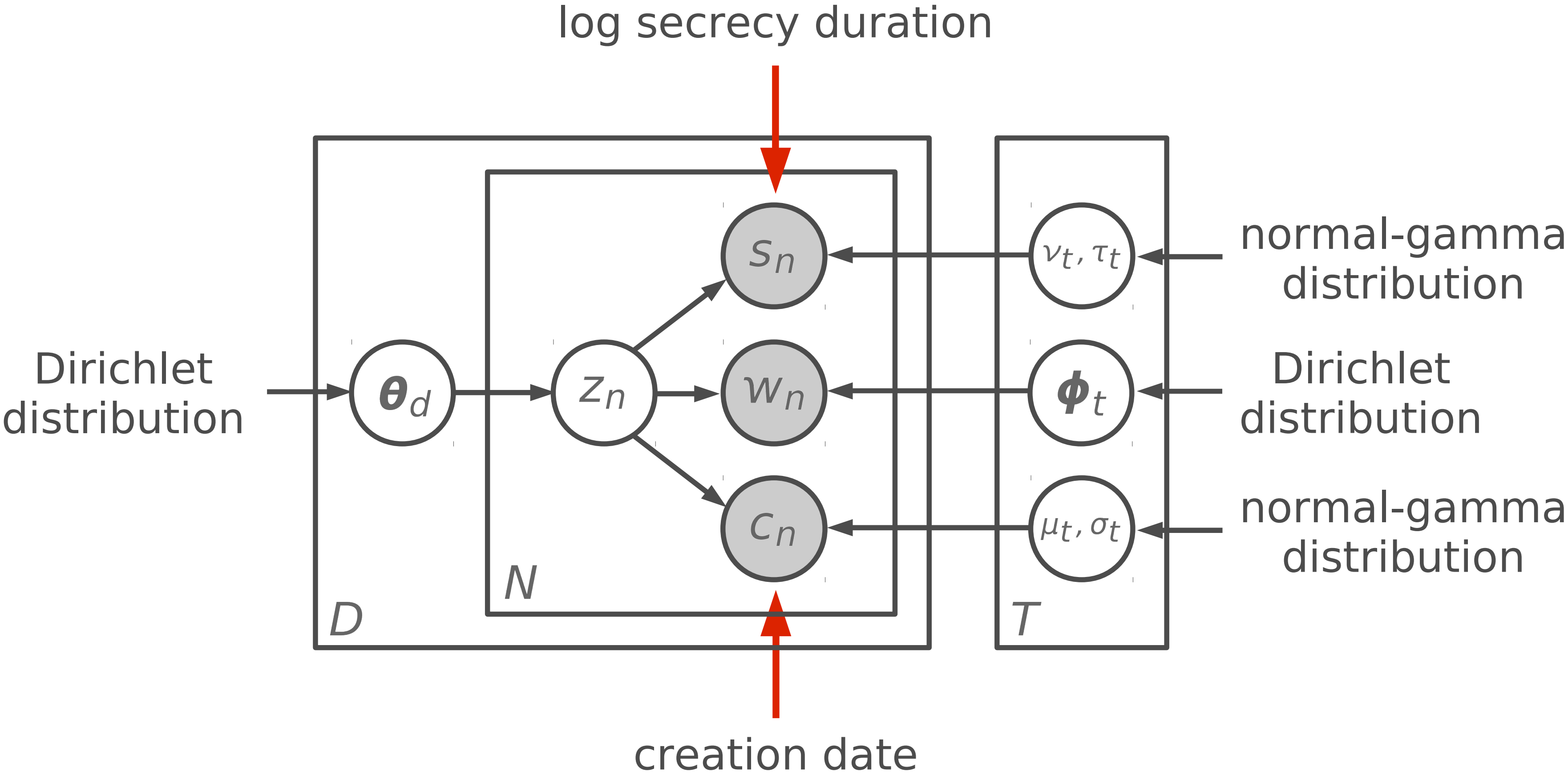
COPY 1 OF 3 COPIES

SEC DEF CONTR No. X38306

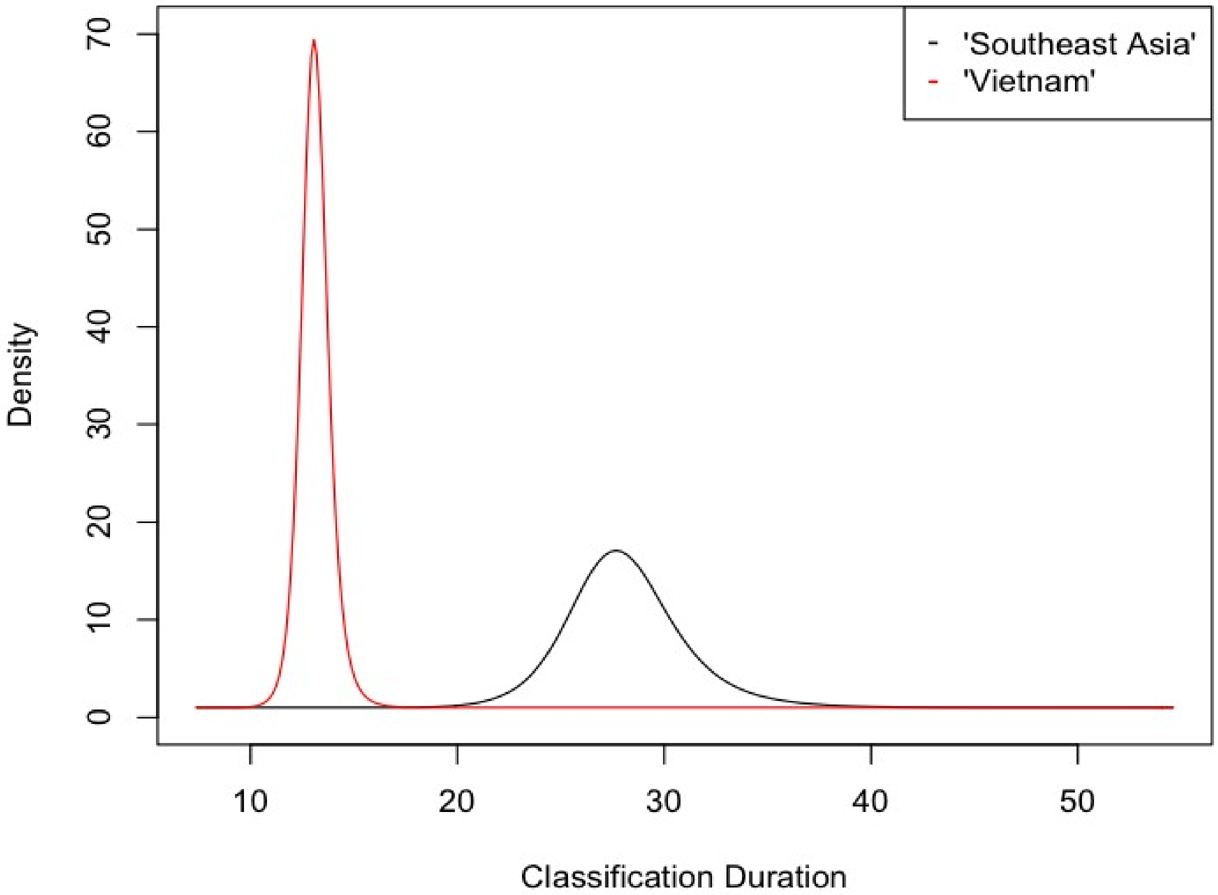
- Topics provide information about classification durations
- Goal: incorporate durations into the generative model
- Infer latent topics using both textual and temporal information

Jointly Modeling Text and Duration

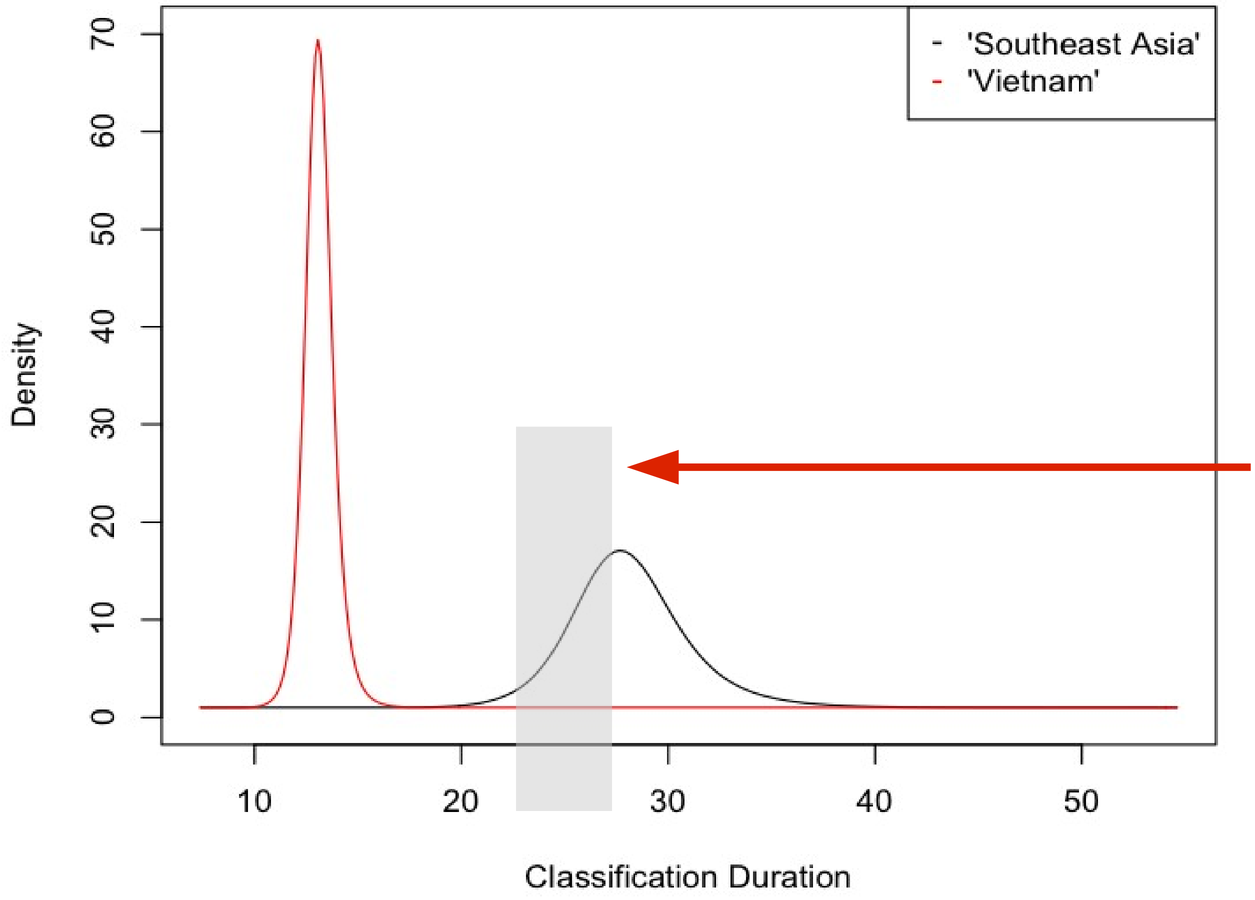
[Shorey et al., '11]



Topic-Specific Duration Distributions



Topic-Specific Duration Distributions



means of all LDA topics relating to Vietnam & Laos

What's Next?

- Predict durations directly from the generative model
    - Mixture vs. admixture topics
    - Supervised topic modeling
    - Unseen content
- Subject matter experts
- Analysis and prediction of redactions

Thanks!

Acknowledgements: B. Desmarais, A. McCallum, D. Mimno, R. Shorey

wallach@cs.umass.edu
http://www.cs.umass.edu/~wallach/