Statistical Machine Learning Analysis of Debian Mailing Lists

Hanna M. Wallach

University of Massachusetts Amherst

wallach@cs.umass.edu

Introduction

- Contributor to Debian GNU/Linux & GNOME
- Co-leader of Debian Women & GNOME WSOP
- Workshop organizer for FLOSSPOLS gender study
- Assistant professor (Sept. 2010) UMass Amherst

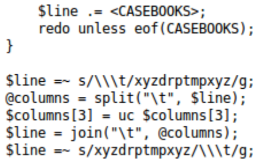
This Talk

- My research goal and methodology
- Document analysis and statistical topic modeling
- Analyzing Debian mailing lists:
  - Initial data sets
  - Preliminary results
- Future research directions:
  - Other statistical topic models

This Talk

- My research goal and methodology

My Research Goal

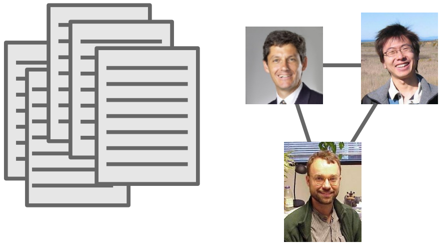


```
    $line .= <CASEBOOKS>;
    redo unless eof(CASEBOOKS);
}

$line =~ s/\\\t/xyzdrptmpxyz/g;
@columns = split("\t", $line);
$columns[3] = uc $columns[3];
$line = join("\t", @columns);
$line =~ s/xyzdrptmpxyz/\\\t/g;
```

$$\prod_t \frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \frac{\prod_w \Gamma(N_{w|t}+\beta)}{\Gamma(N_{.|t}+W\beta)}$$

+



→

To develop new statistical models and computational tools for representing and analyzing large quantities of complex communication and collaboration data in order to better enable social scientists and technologists to advance the study of scientific and technological and innovation.

FOSS Development Communities

- Considerable commercial, noncommercial, academic interest in FOSS development communities:
  - Complex technological, legal, social structures
  - Geographically distributed collaboration
- Organizational and social processes underlying collaborative FOSS development are largely unknown:
  - Area of study for social and computer scientists

Data: Products of Collaboration

"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

FOSS Collaboration Data

- Most FOSS collaboration data are publicly available:
  - Mailing lists, IRC channels
  - Commit messages, bug reports
  - Comments in source code, documentation
  - GPG keysigning records

⇒ Use these data to study organizational and social processes underlying FOSS development

Data Challenges

- Informal, messy, and often highly unstructured data:
  - Developers use different identifiers in different fora
  - IRC channels have multiple interleaved conversations
  - Mix of highly technical and “off-topic” discussion
  - Conversational style is often casual

⇒ Significant text analysis is required prior to developing models for answering social science questions

Approach: Statistical Models

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

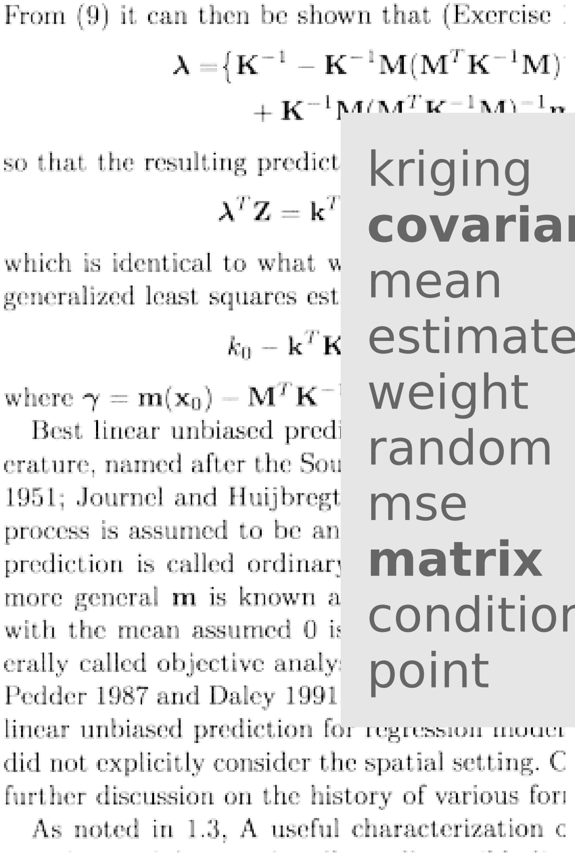
This Talk

- My research goal and methodology
- Document analysis and statistical topic modeling

Statistical Topic Modeling

- Three fundamental assumptions:
  - Documents have latent semantic structure (“topics”)
  - Can infer topics from word–document co-occurrences
  - Words are related to topics, topics to documents
- Given a data set, the goal is to
  - Learn the composition of the topics for that data set
  - Learn which topics are used in each document

Why Topic Models?



From (9) it can then be shown that (Exercise

$$\boldsymbol{\lambda} = \{\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{M}(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})$$

$$+ \mathbf{K}^{-1}\mathbf{M}(\mathbf{M}^T\mathbf{K}^{-1}\mathbf{M})^{-1}\mathbf{n}$$

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 $\boldsymbol{\gamma} = \mathbf{m}(\mathbf{x}_0) - \mathbf{M}^T\mathbf{K}^{-1}$

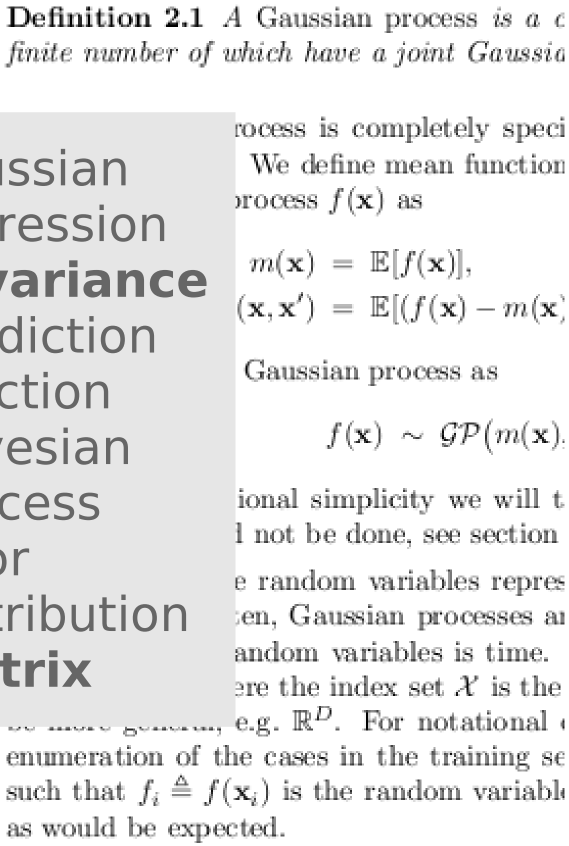
Best linear unbiased predi
erature, named after the Sou
1951; Journel and Huijbregt
process is assumed to be an
prediction is called ordinary
more general **m** is known a
with the mean assumed 0 is
erally called objective analy
Pedder 1987 and Daley 1991
linear unbiased prediction for regression model
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.

gaussian
regression
covariance
prediction
function
bayesian
process
prior
distribution
matrix



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

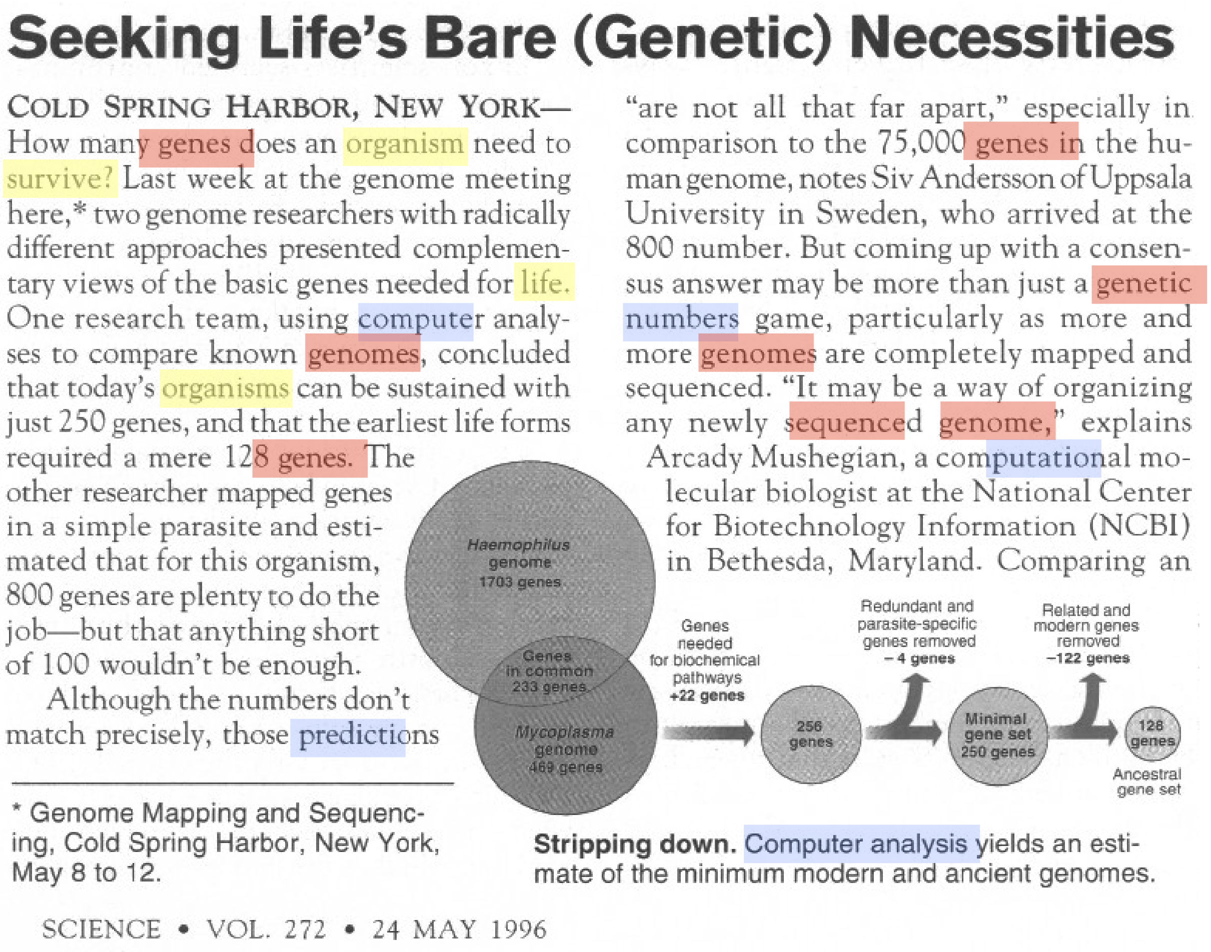
e random variables repres
ten, Gaussian processes a
andom variables is time.
ere the index set $\mathcal{X}$ is the
e.g. $\mathbb{R}^D$. For notational
enumeration of the cases in the training se
such that $f_i \triangleq f(\mathbf{x}_i)$ is the random variabl
as would be expected.

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.

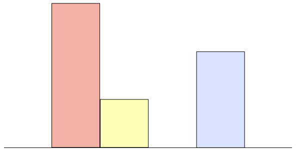
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

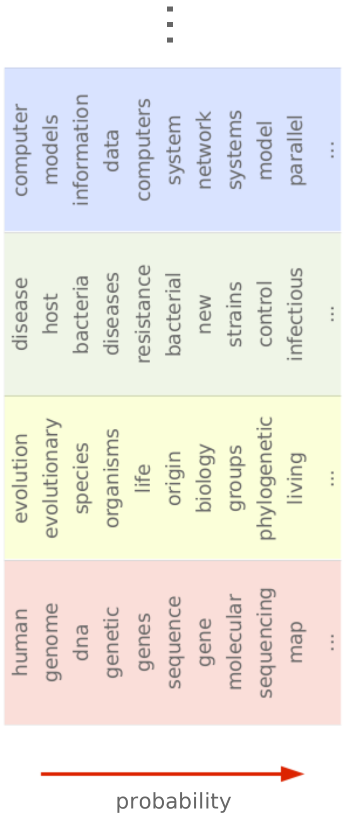
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996



Generative Process



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

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

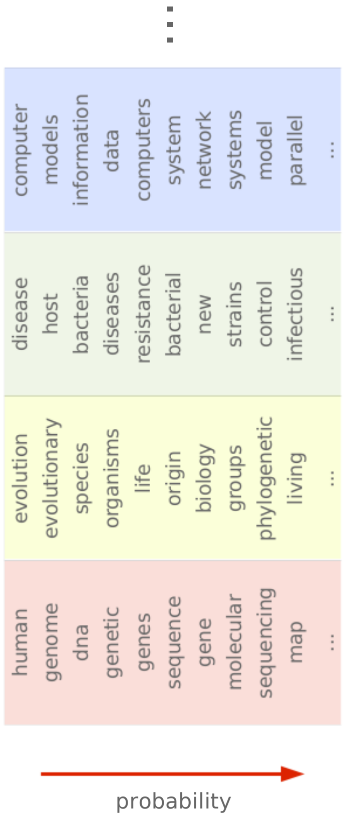
disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
...

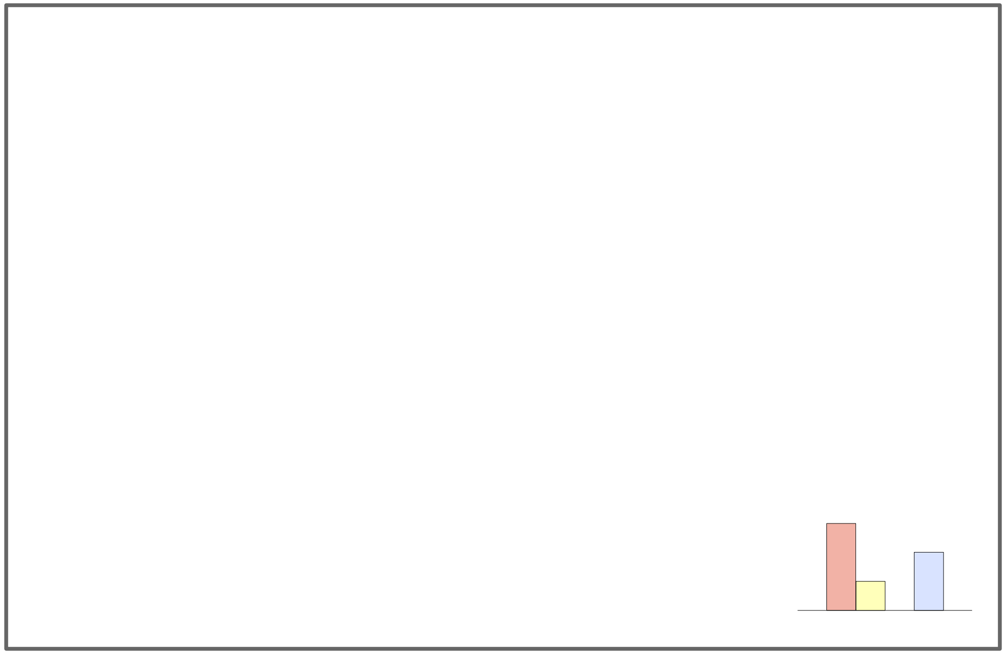
computer
models
information
data
computers
system
network
systems
model
parallel
...

⋮

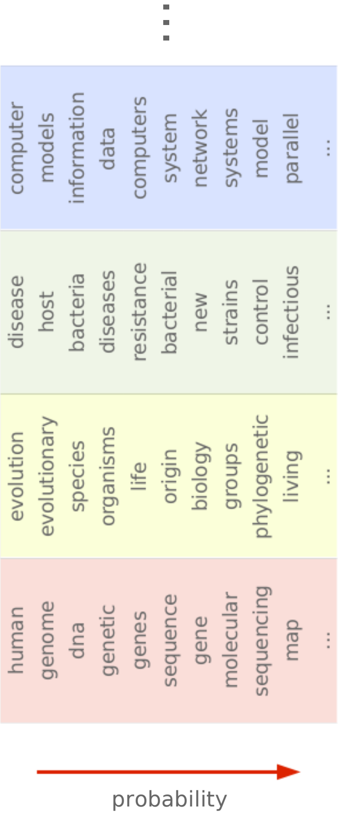
probability

Choose a Topic Distribution


⋮
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



⋮

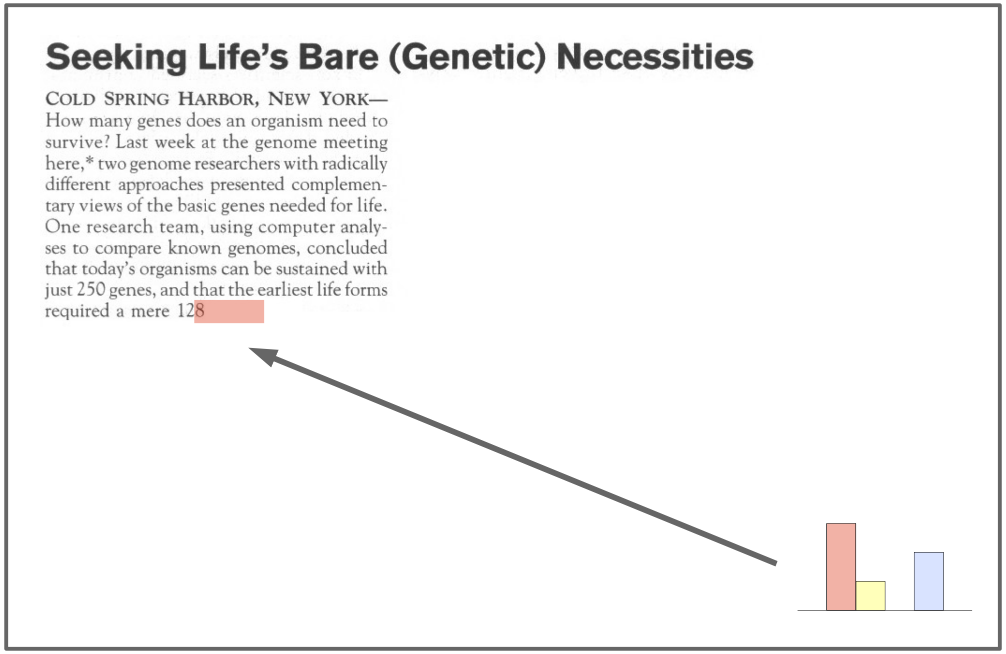
computer models information data computers system network systems model parallel …

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 …

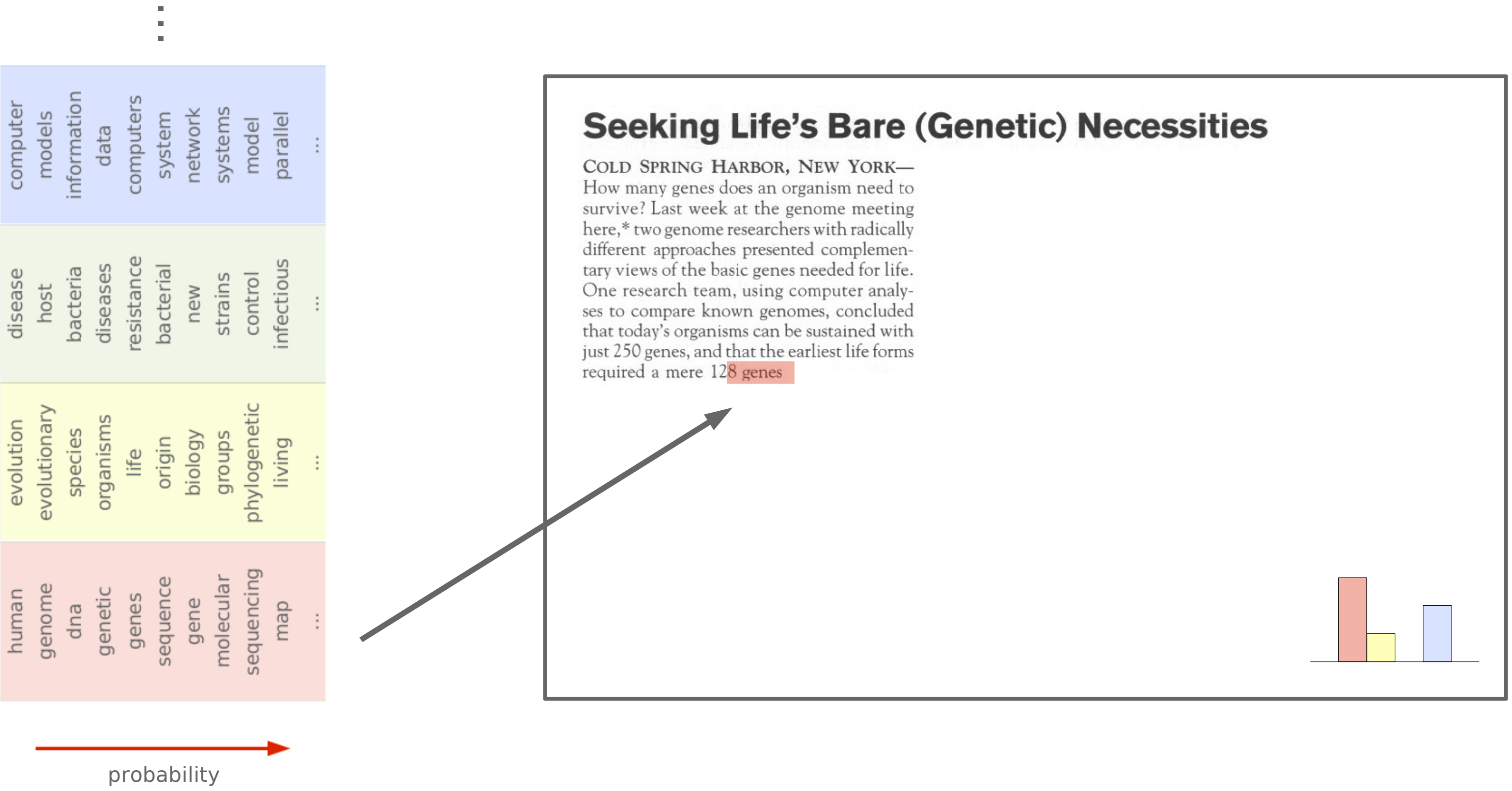
probability



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
...

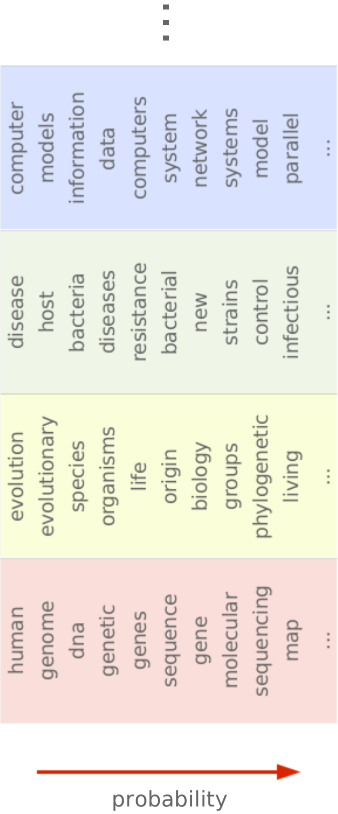
human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
...

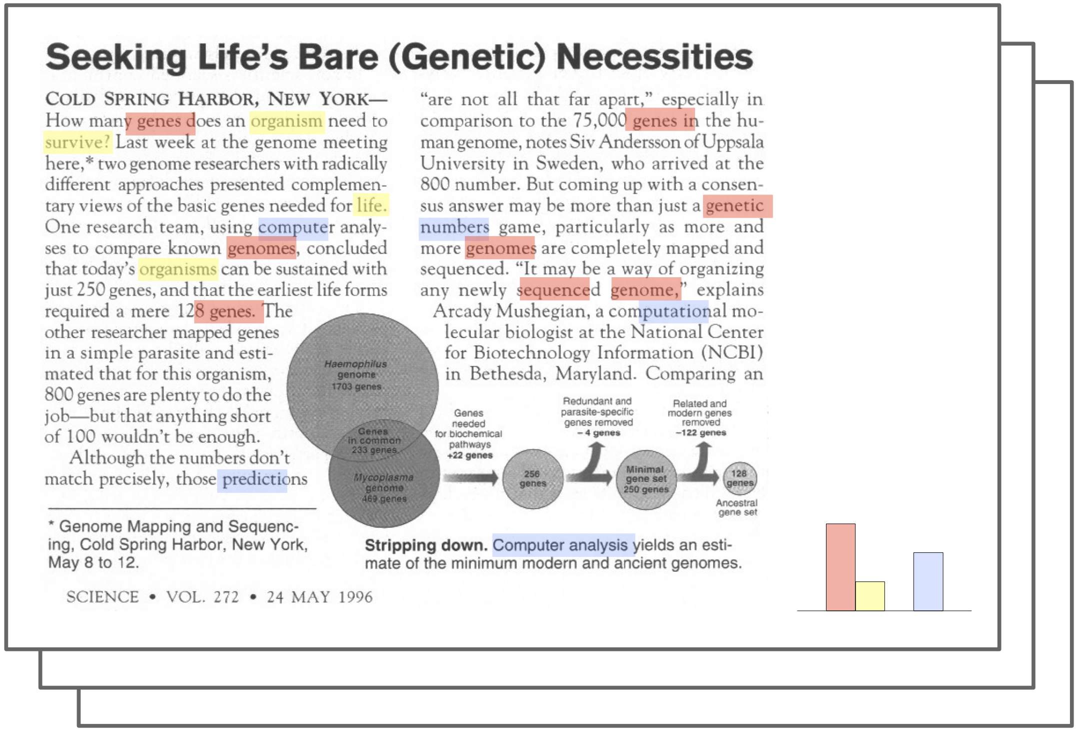
probability

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





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

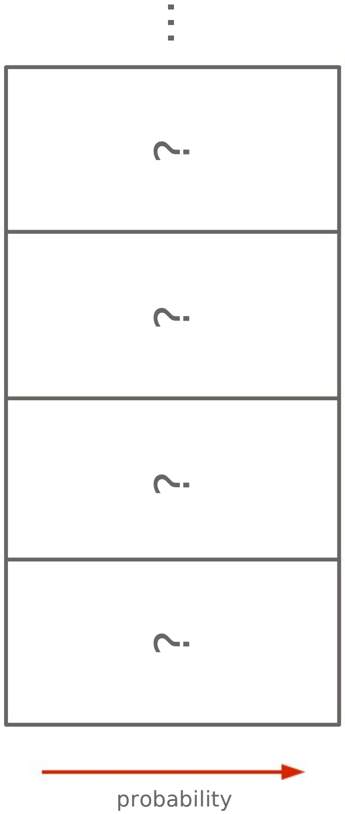
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed – 4 genes
Minimal gene set 250 genes
Related and modern genes removed –122 genes
128 genes
Ancestral gene set

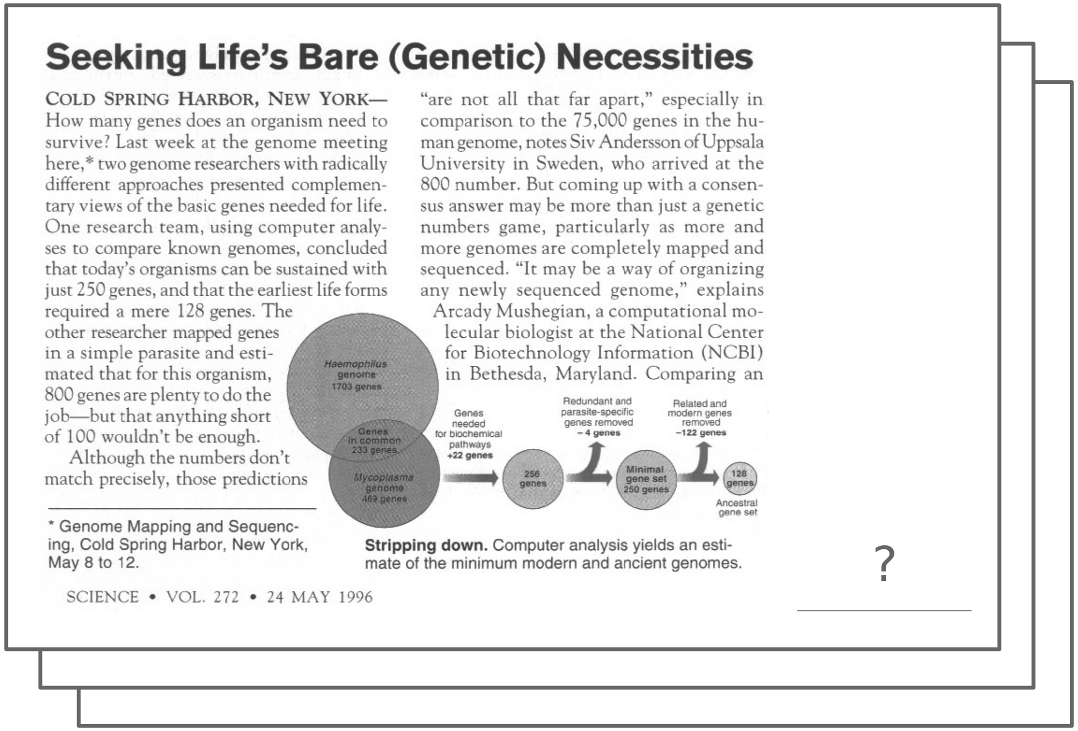
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* 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

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.

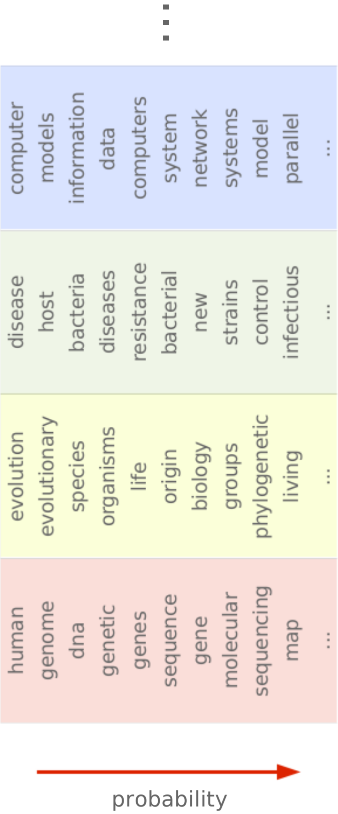
SCIENCE • VOL. 272 • 24 MAY 1996

?

Statistical Inference

- Randomly guess which topic “generated” each word:
- Given a set of guesses, can estimate the distributions
  - Initially the distributions will be random
- Repeatedly refine the guess for each word:
  - Probability of guessing topic t for word w in document d is proportional to # of times topic t has been guessed for other words in document d and # of times topic t has been guessed for all other occurrences of word w

The End Result...



⋮

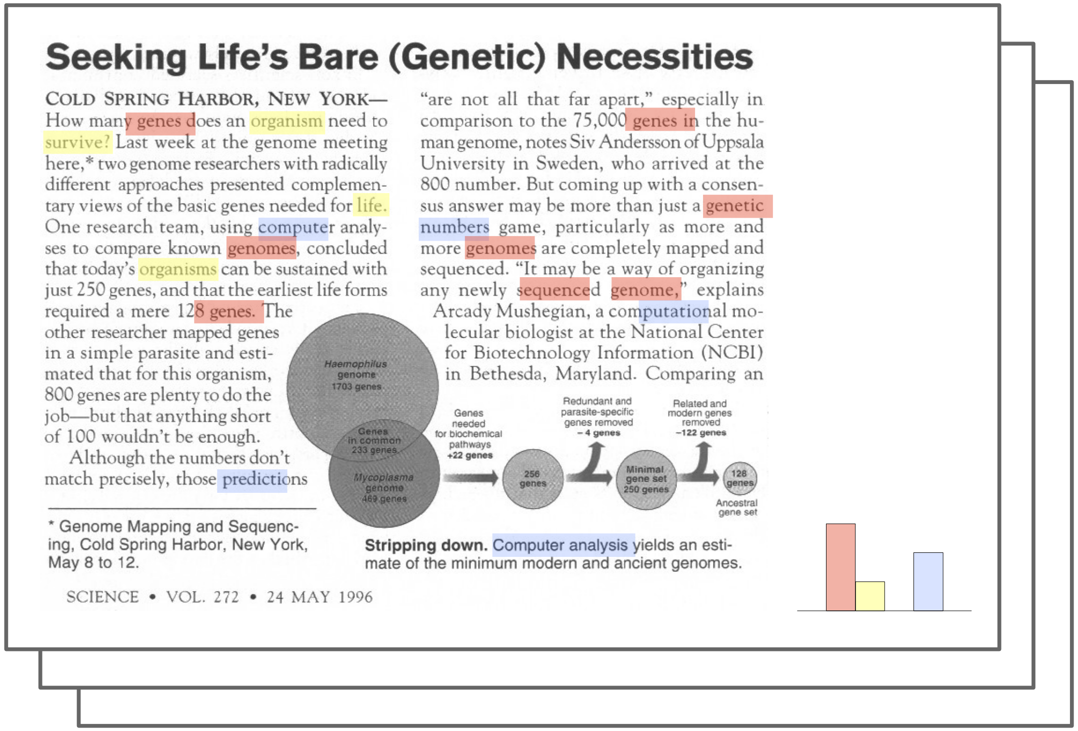
computer models information data computers system network systems model parallel ...

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

- My research goal and methodology
- Document analysis and statistical topic modeling
- Analyzing Debian mailing lists:
    - Initial data sets

This Talk

- My research goal and methodology
- Document analysis and statistical topic modeling
- Analyzing Debian mailing lists:
  - Initial data sets
  - Preliminary results

Initial Data Sets

- Quoted text and signatures stripped
- debian-project:
  - 19,347 messages
  - 1225797 words (max. 7,916 per message)
- debian-women:
  - 4,124 messages
  - 228,076 words (max. 1,524 per message)

100 Topics

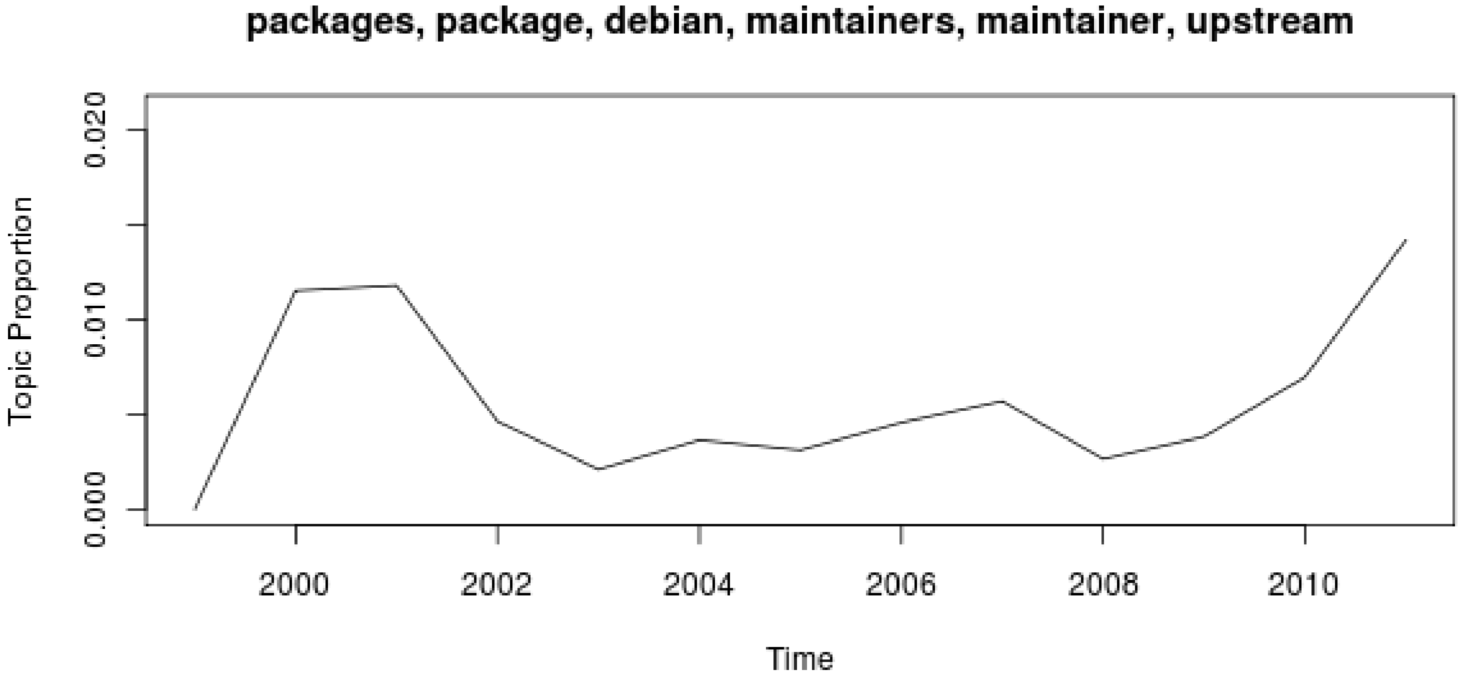
d-project →

package	ubuntu	nm	ftp-master
packages	debian	process	queue
install	patches	applicant	packages
apt-get	derivatives	dam	upload
apt	lts	fd	team
...	...	...	...

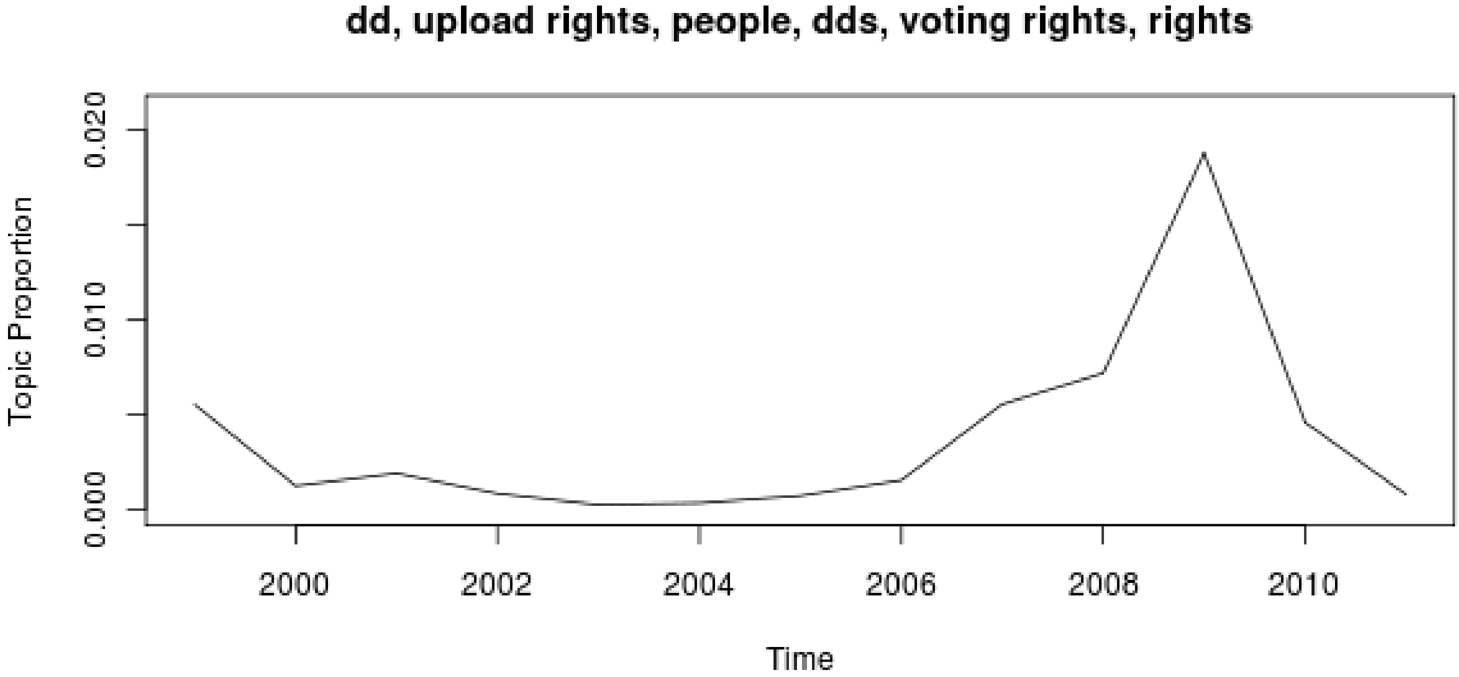
d-women →

women	website	post	nm
men	page	culture	debian
female	site	response	process
male	work	posts	dd
man	d-w	behavior	packages
...	...	...	...

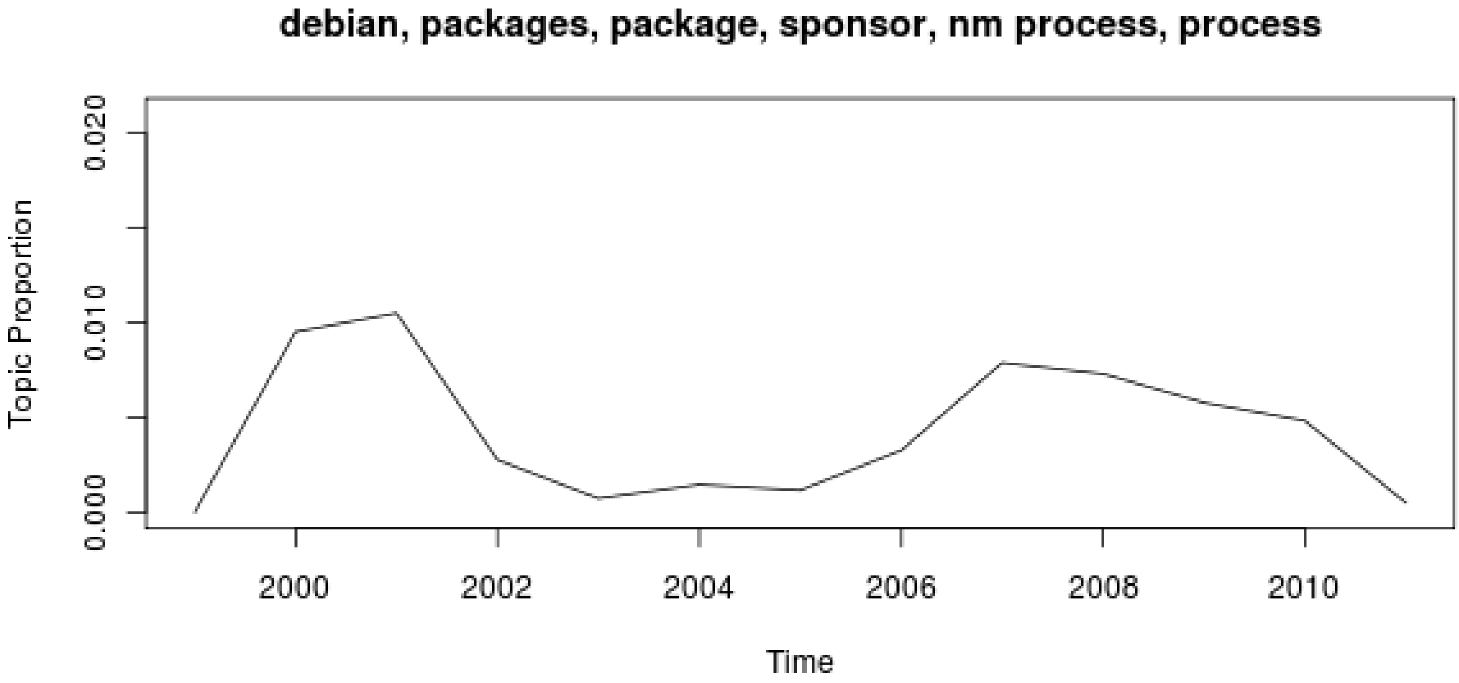
Topic Usage Over Time



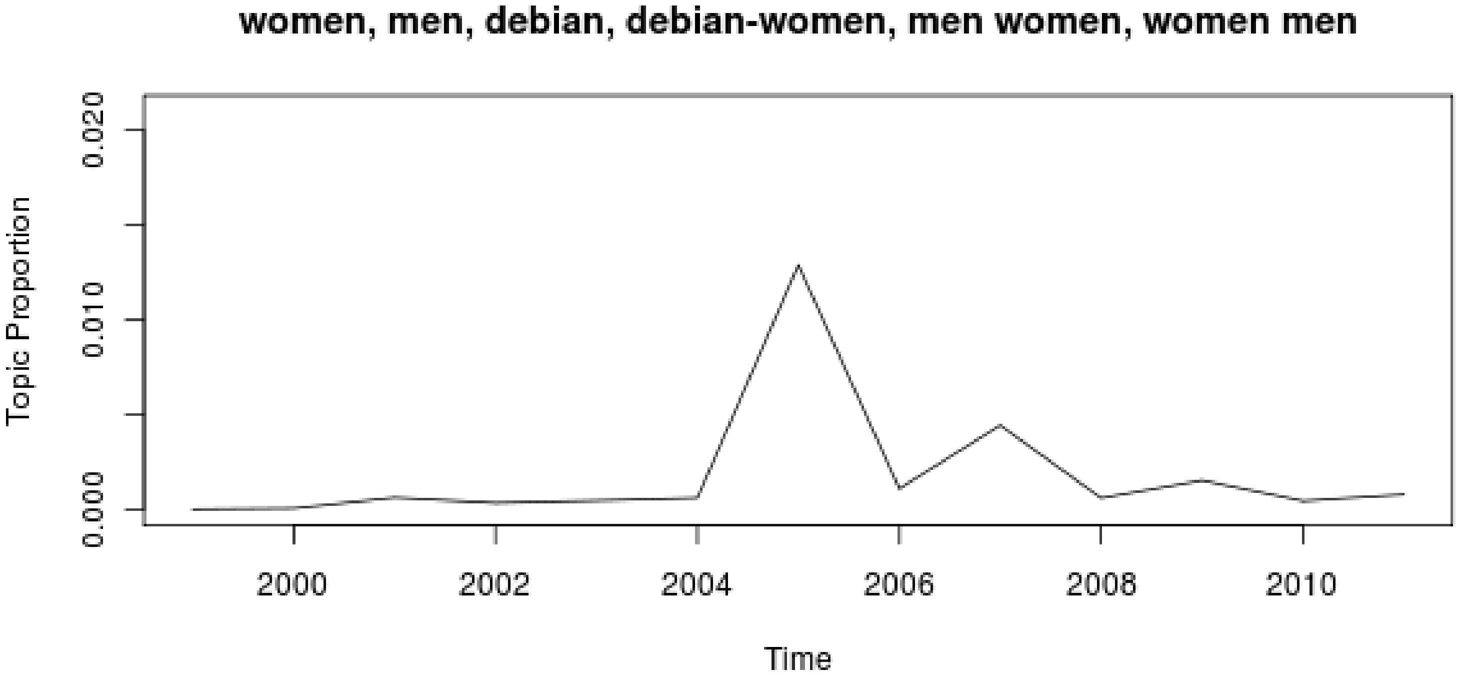
Topic Usage Over Time



Topic Usage Over Time



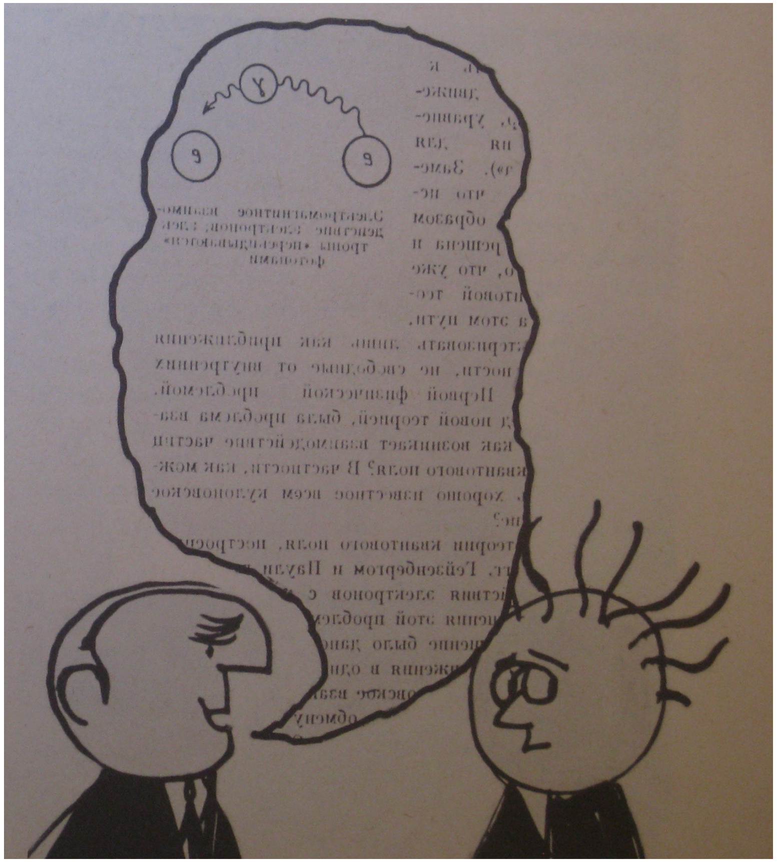
Topic Usage Over Time



This Talk

- My research goal and methodology
- Document analysis and statistical topic modeling
- Analyzing Debian mailing lists:
  - Initial data sets
  - Preliminary results
- Future research directions:
  - Other statistical topic models

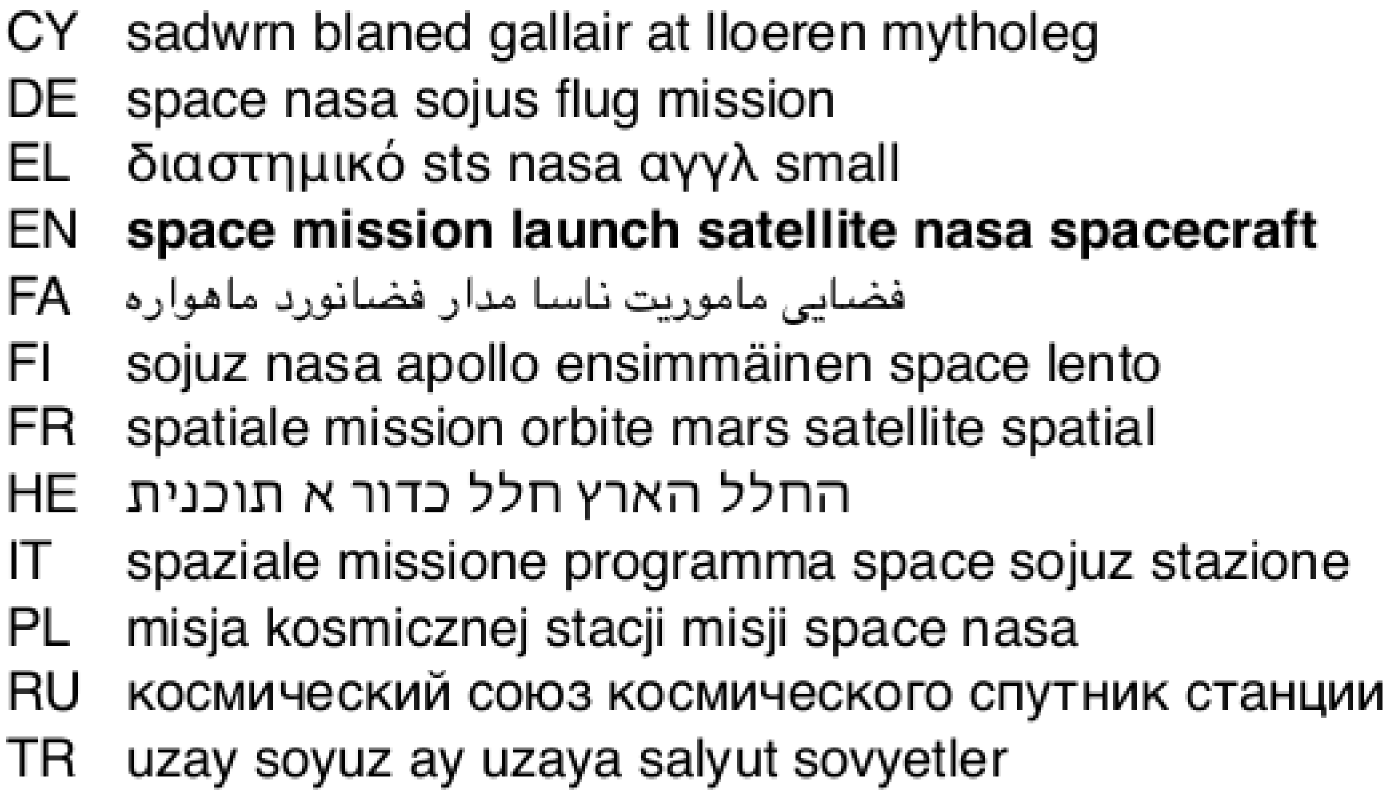
Cross-language Analysis



“He may know one language backwards and forward, but he can't communicate with a scientist who only knows another: a graphic illustration of the need for translation of foreign scientific documents.”

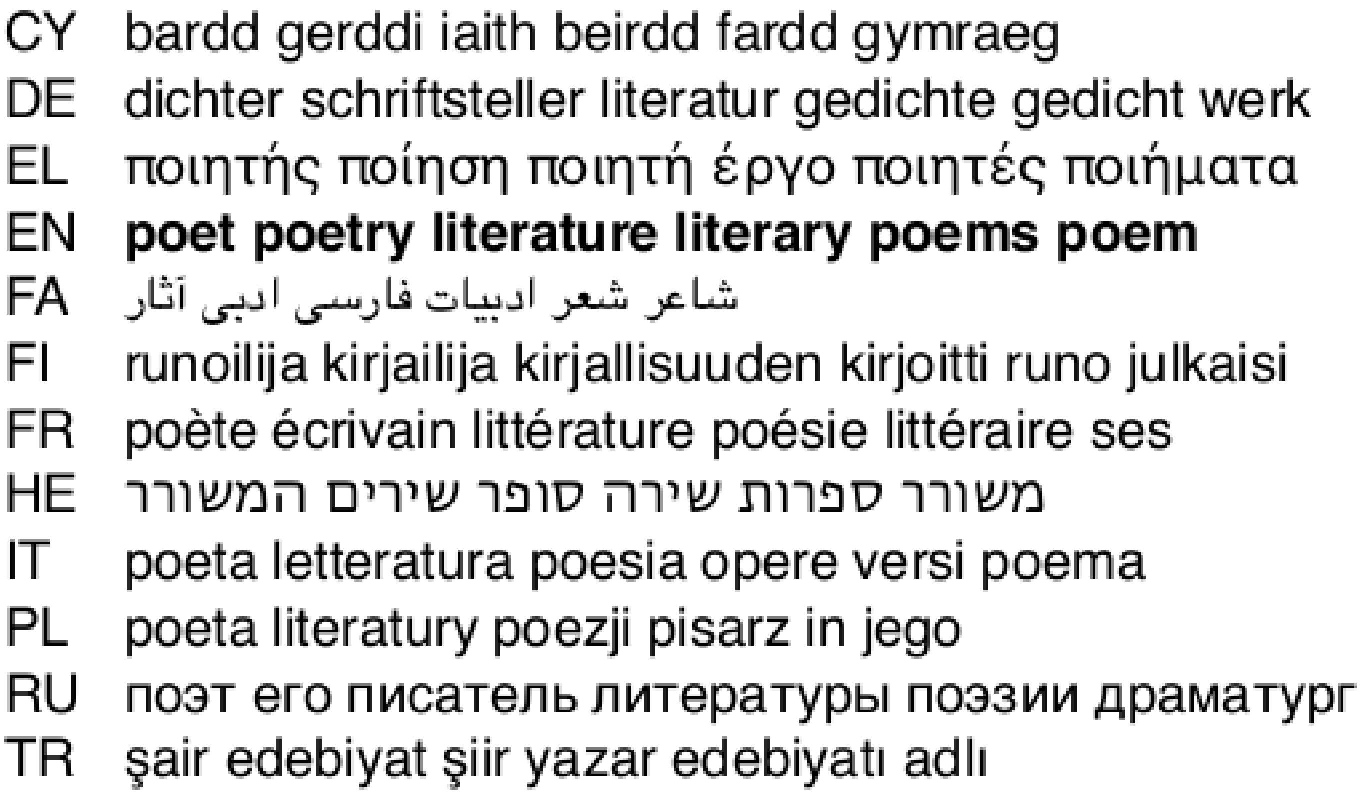
— NSF Brochure, 1962

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

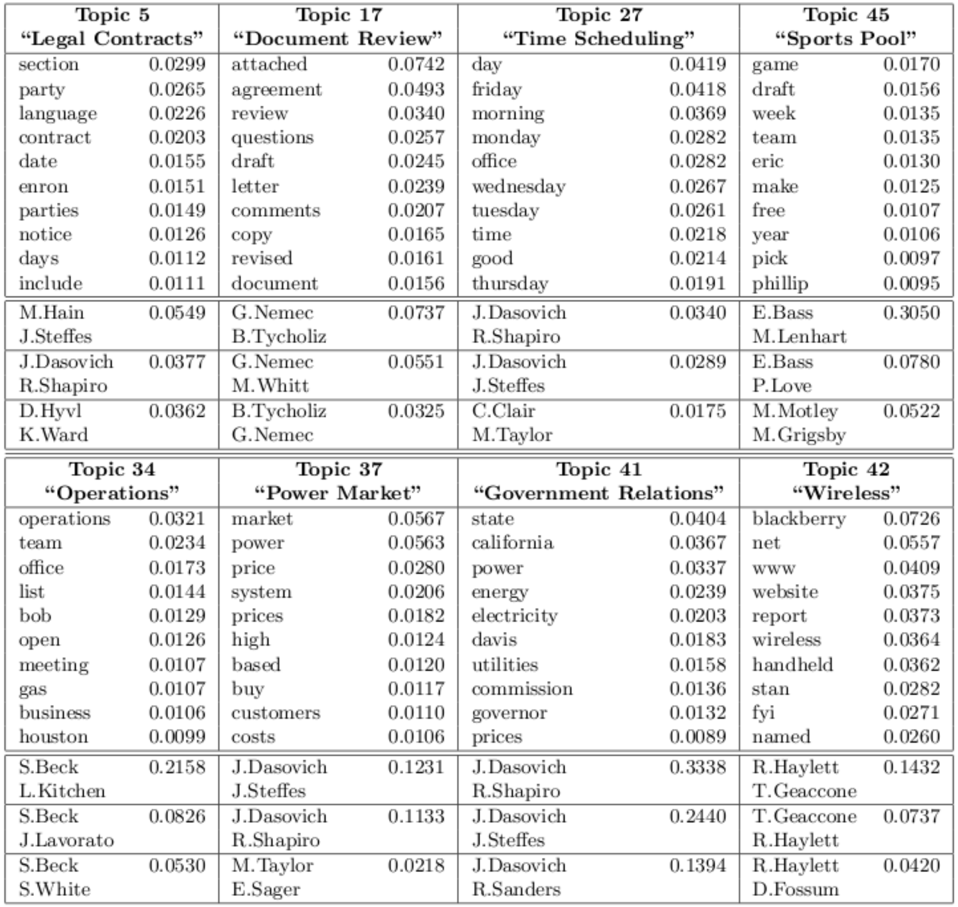
Aligned Corpora

- Fully parallel corpora: direct translations
  - Expensive to produce, relatively rare
- Partially parallel corpora: few parallel “glue” tuples
  - < 25% is sufficient to obtain aligned topics
- Can we use documentation (nearly-direct translations) as glue tuples for simultaneously analyzing the content of mailing lists in multiple languages?

Analyzing Groups and Topics

- Simultaneously find groups of people and topics
- Do people who work on similar parts of Debian talk about similar things on Debian mailing lists?
- Can we automatically discover groups of people from mailing lists without any prior knowledge?
  - Discovery of groups is guided by topics
  - Discovery of topics is guided by groups

Groups and Topics



Topic 5 “Legal Contracts”		Topic 17 “Document Review”		Topic 27 “Time Scheduling”		Topic 45 “Sports Pool”	
section	0.0299	attached	0.0742	day	0.0419	game	0.0170
party	0.0265	agreement	0.0493	friday	0.0418	draft	0.0156
language	0.0226	review	0.0340	morning	0.0369	week	0.0135
contract	0.0203	questions	0.0257	monday	0.0282	team	0.0135
date	0.0155	draft	0.0245	office	0.0282	eric	0.0130
enron	0.0151	letter	0.0239	wednesday	0.0267	make	0.0125
parties	0.0149	comments	0.0207	tuesday	0.0261	free	0.0107
notice	0.0126	copy	0.0165	time	0.0218	year	0.0106
days	0.0112	revised	0.0161	good	0.0214	pick	0.0097
include	0.0111	document	0.0156	thursday	0.0191	phillip	0.0095
M.Hain J.Steffes	0.0549	G.Nemec B.Tycholiz	0.0737	J.Dasovich R.Shapiro	0.0340	E.Bass M.Lenhart	0.3050
J.Dasovich R.Shapiro	0.0377	G.Nemec M.Whitt	0.0551	J.Dasovich J.Steffes	0.0289	E.Bass P.Love	0.0780
D.Hyvl K.Ward	0.0362	B.Tycholiz G.Nemec	0.0325	C.Clair M.Taylor	0.0175	M.Motley M.Grigsby	0.0522

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

Thanks!

Acknowledgements: Dafydd Harries, Benjamin Mako Hill, Andrew McCallum, David Mimno, David Nusinow