Applications: Prediction

Matthew Gentzkow
Jesse M. Shapiro

Chicago Booth and NBER

Introduction

- Methods map high-dimensional x to low-dimensional z

Introduction

- Methods map high-dimensional x to low-dimensional z
- Three main uses of z

  1. Forecasting (e.g., what will inflation be next month?)
  2. Descriptive analysis (e.g., are there genes that predict risk aversion?)
  3. Input into subsequent causal analysis (e.g., as LHS var, RHS var, control, instrument, etc.)

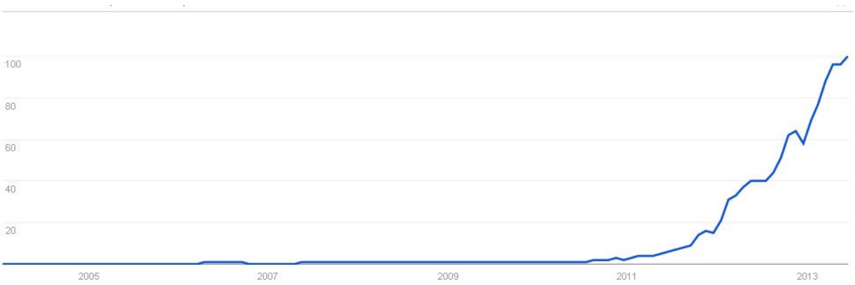
Outline

- Brief overview of data & applications
- Detailed discussion of text as data

Overview

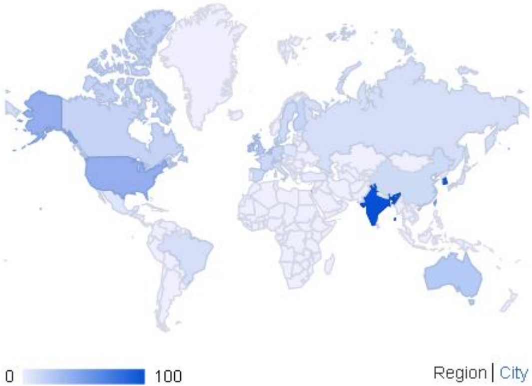
Google Searches

"Big Data"



Google Searches

"Big Data"



Google Searches: Applications

- Prediction
  - Google flu trends (Dukik et al. 2009)
  - Unemployment claims, retail sales, consumer confidence, etc. (Choi & Varian 2009, 2012)

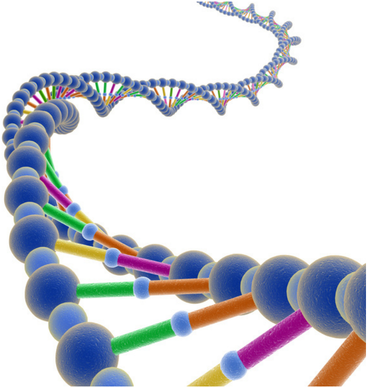
Google Searches: Applications

- Prediction
  - Google flu trends (Dukik et al. 2009)
  - Unemployment claims, retail sales, consumer confidence, etc. (Choi & Varian 2009, 2012)
- Descriptive
  - What searches predict "consumer confidence" (Varian 2013)

Google Searches: Applications

- Prediction
  - Google flu trends (Dukik et al. 2009)
  - Unemployment claims, retail sales, consumer confidence, etc. (Choi & Varian 2009, 2012)
- Descriptive
  - What searches predict "consumer confidence" (Varian 2013)
- Input to analysis
  - Saiz & Simonsohn (2013) → city-level corruption
  - Stephens-Davidowitz (2013) → racial animus & effect on voting for Obama

Genes



- Genetic data has been one of the main applications of high-dimensional methods
- LHS: Physical or behavioral outcome
- RHS: Single-nucleotide polymorphisms (SNPs)
- Typical dataset is $N \approx 10,000$ and $K \approx 2,500,000$

Genes: Applications

- Descriptive: look for genetic predictors of...
  - Risk aversion & social preferences (Cesarini et al. 2009)
  - Financial decision making (Cesarini et al. 2010)
  - Political preferences (Benjamin et al. 2012)
  - Self-employment (van der Loos et al. 2013)
  - Educational attainment, subjective well being (Rietveld et al. forthcoming)

Genes: Applications

- Descriptive: look for genetic predictors of...
  - Risk aversion & social preferences (Cesarini et al. 2009)
  - Financial decision making (Cesarini et al. 2010)
  - Political preferences (Benjamin et al. 2012)
  - Self-employment (van der Loos et al. 2013)
  - Educational attainment, subjective well being (Rietveld et al. forthcoming)
- Early reported associations have been shown to be spurious & non-replicable (Benjamin et al. 2012)

Medical Claims

- Big and high dimensional
    - 10 years of Medicare data on the order of 100 TB
    - ($patient \times doctor \times hospital \times treatment \times cost \ldots$)

Medical Claims

- Big and high dimensional
  - 10 years of Medicare data on the order of 100 TB
  - ($patient \times doctor \times hospital \times treatment \times cost \ldots$)
- Dimension reduction: How to collapse data into a single-dimensional index of "health" or "predicted spending"
  - Medicare "risk scores" based on ad hoc criteria
  - Johns Hopkins ACG system uses proprietary predictive model

Medical Claims: Applications

- Input into analysis
  - Numerous studies use Medicare risk scores as a control variable or independent variable of interest
  - Einav & Finkelstein (forthcoming) use risk score as mediator of health plan choice
  - Handel (2013) uses Johns Hopkins ACG score as measure of private information

Credit Scores

- Predicting default risk from consumer credit data is similar to predicting health risk from medical claims

Credit Scores

- Predicting default risk from consumer credit data is similar to predicting health risk from medical claims
- Forecasting
    - Large literature applies machine learning tools to improve forecasting of default risks (e.g., Khandani et al. 2010)

Credit Scores

- Predicting default risk from consumer credit data is similar to predicting health risk from medical claims
- Forecasting
    - Large literature applies machine learning tools to improve forecasting of default risks (e.g., Khandani et al. 2010)
- Input into analysis
    - Adams, Einav and Levin (2009) and Einav, Jenkins and Levin (2012) evaluate auto dealer's proprietary credit scoring algorithm
    - Rajan, Seru and Vig (forthcoming) look at market responses to using a limited set of variables in credit scoring

Online Purchases

- Amazon, Ebay, and other large Internet firms use purchase and browsing history to make recommendations, target advertising, etc.
- “Netflix Prize” contest for best algorithm to predict future user ratings based on past ratings

Congressional Roll Call Votes

- Poole & Rosenthal (1984, 1985, 1991, 2000, etc.) use factor analysis methods to project Roll Call votes into ideology scores
- Ask questions like
  - Are there multiple dimensions of ideology?
  - How has polarization changed over time?

Text as Data

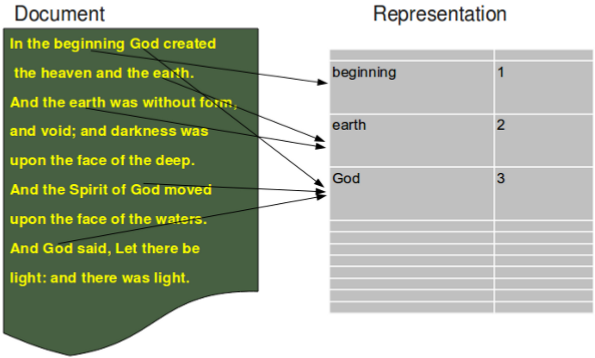
Sources

- News
- Books
- Web content
- Congressional speeches
- Corporate filings
- Twitter & Facebook

Less Obvious Sources

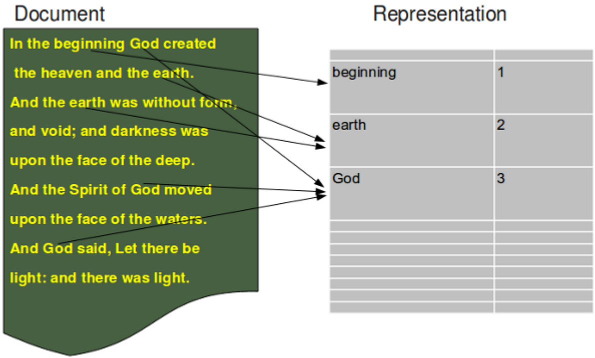
- Amazon and eBay listings
- Google search ads
- Medical records
- Central bank announcements

Bag of Words



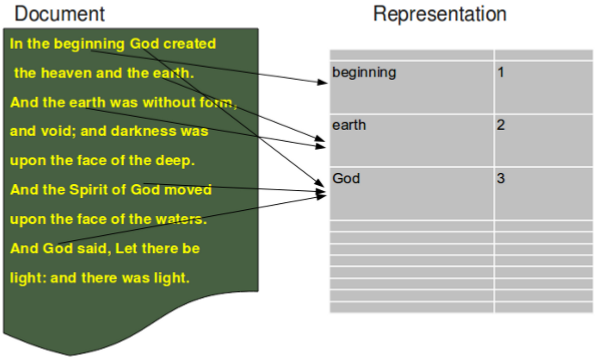
Document
In the beginning God created
the heaven and the earth.
And the earth was without form,
and void; and darkness was
upon the face of the deep.
And the Spirit of God moved
upon the face of the waters.
And God said, Let there be
light: and there was light.
Representation
beginning 1
earth 2
God 3

Bag of Words



- Can apply to "N-grams" as well as single words

Bag of Words



- Can apply to "*N*-grams" as well as single words
- This seems crude, but it works remarkably well in practice, and gains to more sophisticated representations prove to be small

The Science and the Art

- General theme: Real research always combines automated dimension reduction techniques with "manual" steps based on priors

The Science and the Art

- General theme: Real research always combines automated dimension reduction techniques with "manual" steps based on priors
- E.g.,
  - Keep only words occurring more than X times
  - Drop very common "stopwords" like "the," "at," a"
  - "Stem" words to combine, e.g., "economics," "economic," "economically"
  - Drop HTML tags

The Science and the Art

- General theme: Real research always combines automated dimension reduction techniques with "manual" steps based on priors
- E.g.,
  - Keep only words occurring more than X times
  - Drop very common "stopwords" like "the," "at," a"
  - "Stem" words to combine, e.g., "economics," "economic," "economically"
  - Drop HTML tags
- In fact, 90% of text analysis in economics does not automated dimension reduction at all
  - Saiz & Simonsohn (2013) $\rightarrow$ city name + "corruption"
  - Baker, Bloom, and Davis (2013) $\rightarrow$ "economic" + "policy" + "uncertainty", etc.
  - Lucca & Trebbi (2011) $\rightarrow$ "hawkish/dovish," "loose/tight," + "Federal Open Market Committee"

Interfaces

- Bag of words assumes access to full text (at least for N-grams of interest)

Interfaces

- Bag of words assumes access to full text (at least for N-grams of interest)
- Many researchers, however, can only access text via *search* interfaces (e.g., Google page counts, news archives, etc.)

Interfaces

- Bag of words assumes access to full text (at least for *N*-grams of interest)
- Many researchers, however, can only access text via *search* interfaces (e.g., Google page counts, news archives, etc.)
    - This requires some external method of feature selection to narrow down vocabulary

Sentiment Analysis

Setup

- Outcome y_i
- Features x_i
- Data:

$$\underbrace{\{x_1, y_1\}, \{x_2, y_2\}, \{x_3, y_3\}, ..., \{x_N, y_N\}}_{\text{Training set}}, \underbrace{\{x_{N+1}, ?\}}_{\text{Target}}$$

Spam Filter

- Outcome $y_i \in \{spam, ham\}$
- Human coder classifies N cases as *spam* or *ham*
- Must decide whether to deliver the $(N+1)$ message or send it to the filter

Issues

- What features x_i do we use?
  - Counts of words?
  - Counts of characters?
  - Complete machine representation of e-mail?
- How do we avoid overfitting?
  - $>$1m words in English language
  - ASCII file with 100 printable characters has 95^{100} possible realizations

Applications

- Partisanship in the news media [TODAY]
  - Turn millions of words into an index of media slant or bias
- Sentiment in financial news [TODAY]
  - Classify news, chat room discussions, etc. as positive or negative
- Estimating causal effects [TOMORROW]
  - Turn a huge dataset into a low-dimensional control for endogeneity

Sentiment Analysis: *Partisanship in the News Media*

Overview

- Questions
  - How centrist are the news media?
  - What factors (owners, readers) predict how a newspaper portrays the news?
- Need measure of partisan orientation of news media
- Challenges
  - Training set: research assistants? surveys?
  - Dimensionality
    - Feature selection: words? phrases? images?
    - Parsimony: millions of possible words/phrases

Groseclose and Milyo (2005)

- Training set: US Congress
  - Assign members an ideology score y_i based on roll-call voting
- Dimensionality
  - Count frequency of citations to think tanks x_i in Congressional Record
  - Feature selection "by hand": use *ex ante* criterion to control dimensionality

Groseclose and Milyo (2005)

Groseclose and Milyo (2005)

Mr. DORGAN. Mr. President, I come to the floor to speak first about the Congressional Budget Office, which last week released its monthly budget projection. And I noticed that this projection, this estimate, received prominent coverage in the Washington Post and in other major daily newspapers around the country last week....

A study by a tax expert at the **Brookings Institution** says if you have a national sales tax, the rates would probably be over 30 percent, and then add the State and local taxes, and that would be on almost everything. So say you would like to buy a house and here is the price we have agreed on, and then have someone tell you, oh, yes, you have a 37-percent sales tax applied to that price, 30 percent Federal, 7 percent State and local.

Speaker	Think Tank	Count
Dorgan, Byron (D-ND)	Brookings Institution	1

Groseclose and Milyo (2005)

- Count references to think tanks in news media

Groseclose and Milyo (2005)

- Let y_i be ADA score of senator i
- Let x_{ijt} be indicator for senator i cites think tank j on occasion t
- Then

$$\Pr(x_{ijt} = 1) = \frac{\exp(\alpha_j + \beta_j y_i)}{\sum_{j'} \exp(\alpha_{j'} + \beta_{j'} y_i)}$$

- Assume same model applies to news media m but treat y_m as unknown
- Estimate α_j, β_j, y_m via joint maximum likelihood

Groseclose and Milyo (2005)

- Dimension of data $p = 50$
  - Start with 200 think tanks
  - Collapse all but top 44 into 6 groups
- Dimension of data $n = 535$
  - Less those who don't cite think tanks

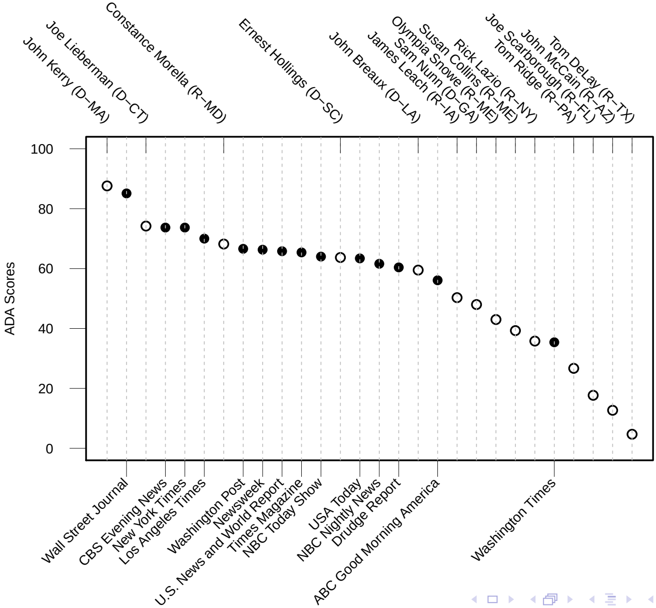
Groseclose and Milyo (2005)

Senator	Partisanship	News Outlet	Partisanship
John McCain	12.7	Fox News	
Arlen Specter	51.3	*USA Today*	
Joe Lieberman	74.2	*New York Times*	

Groseclose and Milyo (2005)

Senator	Partisanship	News Outlet	Partisanship
John McCain	12.7	Fox News	39.7
Arlen Specter	51.3	*USA Today*	63.4
Joe Lieberman	74.2	*New York Times*	73.7

Groseclose and Milyo (2005)



Gentzkow and Shapiro (2010)

- Text of 2005 Congressional Record
- Scripted pipeline:
  - Download text
  - Split up text into individual speeches
  - Identify speaker
  - Count *all* two-word/three-word phrases

Gentzkow and Shapiro (2010)

- Training set: US Congress
  - Assign members an ideology score y_i based on partisanship of constituents
- Dimensionality
  - Compute frequency table of phrase counts by party
  - Compute χ^2 statistic of independence
  - Identify 1000 phrases with highest χ^2

Example: Social Security

- Memo to Rep. candidates: "Never say '**privatization/private accounts**.' Instead say '**personalization/personal accounts**.' Two-thirds of America want to personalize Social Security while only one-third would privatize it. Why? Personalizing Social Security suggests ownership and control over your retirement savings, while privatizing it suggests a profit motive and winners and losers."

Example: Social Security

- Memo to Rep. candidates: "Never say '**privatization/private accounts**.' Instead say '**personalization/personal accounts**.' Two-thirds of America want to personalize Social Security while only one-third would privatize it. Why? Personalizing Social Security suggests ownership and control over your retirement savings, while privatizing it suggests a profit motive and winners and losers."
- Congress: "personal account" (48 D vs 184 R); "private account" (542 D vs 5 R)

Top Phrases

Republicans: 2-word	Republicans: 3-word	Democrats: 2-word	Democrats: 3-word
stem cell	embryonic stem cell	private accounts	veterans health care
natural gas	hate crimes legislation	trade agreement	congressional black caucus
death tax	adult stem cells	american people	va health care
illegal aliens	oil for food program	tax breaks	billion in tax cuts
class action	personal retirement accounts	trade deficit	credit card companies
war on terror	energy and natural resources	oil companies	security trust fund
embryonic stem	global war on terror	credit card	social security trust
tax relief	hate crimes law	nuclear option	privatize social security
illegal immigration	change hearts and minds	war in iraq	american free trade
date the time	global war on terrorism	middle class	central american free
boy scouts	class action fairness	african american	national wildlife refuge
hate crimes	committee on foreign relations	budget cuts	dependence on foreign oil
oil for food	deficit reduction bill	nuclear weapons	tax cuts for the wealthy
global war	boy scouts of america	checks and balances	vice president cheney
medical liability	repeal of the death tax	civil rights	arctic national wildlife
highway bill	highway trust fund	veterans health	bring our troops home
adult stem	action fairness act	cut medicaid	social security privatization
democratic leader	committee on commerce science	foreign oil	billion trade deficit
federal spending	cord blood stem	president plan	asian pacific american
tax increase	medical liability reform	gun violence	president bush took office

Top Phrases: Social Security

Republicans: 2-word	Republicans: 3-word	Democrats: 2-word	Democrats: 3-word
stem cell	embryonic stem cell	**private accounts**	veterans health care
natural gas	hate crimes legislation	trade agreement	congressional black caucus
death tax	adult stem cells	american people	va health care
illegal aliens	oil for food program	tax breaks	billion in tax cuts
class action	**personal retirement accounts**	trade deficit	credit card companies
war on terror	energy and natural resources	oil companies	**security trust fund**
embryonic stem	global war on terror	credit card	**social security trust**
tax relief	hate crimes law	nuclear option	**privatize social security**
illegal immigration	change hearts and minds	war in iraq	american free trade
date the time	global war on terrorism	middle class	central american free
boy scouts	class action fairness	african american	national wildlife refuge
hate crimes	committee on foreign relations	budget cuts	dependence on foreign oil
oil for food	deficit reduction bill	nuclear weapons	tax cuts for the wealthy
global war	boy scouts of america	checks and balances	vice president cheney
medical liability	repeal of the death tax	civil rights	arctic national wildlife
highway bill	highway trust fund	veterans health	bring our troops home
adult stem	action fairness act	cut medicaid	**social security privatization**
democratic leader	committee on commerce science	foreign oil	billion trade deficit
federal spending	cord blood stem	president plan	asian pacific american
tax increase	medical liability reform	gun violence	president bush took office

Other R: **personal accounts; social security reform; social security system**

Other D: **privatization plan; security trust; security trust fund; social security trust; privatize social security; social security privatization; privatization of social security; cut social security**

Top Phrases: Foreign Policy

Republicans: 2-word	Republicans: 3-word	Democrats: 2-word	Democrats: 3-word
stem cell	embryonic stem cell	private accounts	**veterans health care**
natural gas	hate crimes legislation	trade agreement	congressional black caucus
death tax	adult stem cells	american people	**va health care**
illegal aliens	**oil for food program**	tax breaks	billion in tax cuts
class action	personal retirement accounts	trade deficit	credit card companies
war on terror	energy and natural resources	oil companies	security trust fund
embryonic stem	**global war on terror**	credit card	social security trust
tax relief	hate crimes law	nuclear option	privatize social security
illegal immigration	**change hearts and minds**	**war in iraq**	american free trade
date the time	**global war on terrorism**	middle class	central american free
boy scouts	class action fairness	african american	national wildlife refuge
hate crimes	committee on foreign relations	budget cuts	**dependence on foreign oil**
oil for food	deficit reduction bill	**nuclear weapons**	tax cuts for the wealthy
global war	boy scouts of america	checks and balances	vice president cheney
medical liability	repeal of the death tax	civil rights	arctic national wildlife
highway bill	highway trust fund	veterans health	**bring our troops home**
adult stem	action fairness act	cut medicaid	social security privatization
democratic leader	committee on commerce science	**foreign oil**	billion trade deficit
federal spending	cord blood stem	president plan	asian pacific american
tax increase	medical liability reform	gun violence	president bush took office

Other R: **saddam hussein, war on terrorism, iraqi people**

Other D: **funding for veterans health; war in iraq and afghanistan; improvised explosive device**

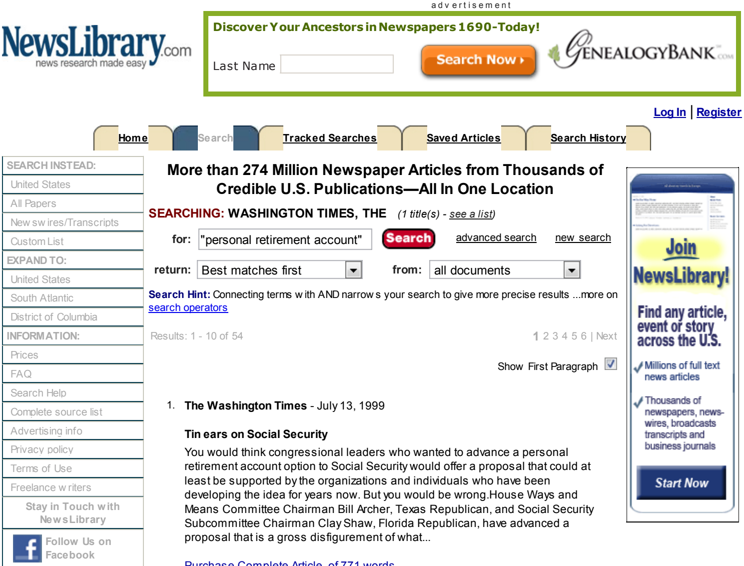
Top Phrases: Fiscal Policy

Republicans: 2-word	Republicans: 3-word	Democrats: 2-word	Democrats: 3-word
stem cell	embryonic stem cell	private accounts	veterans health care
natural gas	hate crimes legislation	trade agreement	congressional black caucus
death tax	adult stem cells	american people	va health care
illegal aliens	oil for food program	**tax breaks**	**billion in tax cuts**
class action	personal retirement accounts	trade deficit	credit card companies
war on terror	energy and natural resources	oil companies	security trust fund
embryonic stem	global war on terror	credit card	social security trust
tax relief	hate crimes law	nuclear option	privatize social security
illegal immigration	change hearts and minds	war in iraq	american free trade
date the time	global war on terrorism	middle class	central american free
boy scouts	class action fairness	african american	national wildlife refuge
hate crimes	committee on foreign relations	**budget cuts**	dependence on foreign oil
oil for food	**deficit reduction bill**	nuclear weapons	**tax cuts for the wealthy**
global war	boy scouts of america	checks and balances	vice president cheney
medical liability	**repeal of the death tax**	civil rights	arctic national wildlife
highway bill	highway trust fund	veterans health	bring our troops home
adult stem	action fairness act	**cut medicaid**	social security privatization
democratic leader	committee on commerce science	foreign oil	billion trade deficit
federal spending	cord blood stem	president plan	asian pacific american
tax increase	medical liability reform	gun violence	president bush took office

Other R: **raise taxes; percent growth; increase taxes; growth rate; government spending; raising taxes; death tax repeal; million jobs created; percent growth rate**

Other D: **estate tax; budget deficit; bill cuts; medicaid cuts; cut funding; spending cuts; pay for tax cuts; cut student loans; cut food stamps; cut social security; billion in tax breaks**

Obtain Phrase Counts from Newspapers



NewsLibrary.com news research made easy

Log In | Register

Home | Search | Tracked Searches | Saved Articles | Search History

SEARCH INSTEAD:
- United States
- All Papers
- Newswires/Transcripts
- Custom List

EXPAND TO:
- United States
- South Atlantic
- District of Columbia

INFORMATION:
- Prices
- FAQ
- Search Help
- Complete source list
- Advertising info
- Privacy policy
- Terms of Use
- Freelance writers

Stay in Touch with NewsLibrary

Follow Us on Facebook

More than 274 Million Newspaper Articles from Thousands of Credible U.S. Publications—All In One Location

SEARCHING: WASHINGTON TIMES, THE *(1 title(s) - see a list)*

for: "personal retirement account" **Search** advanced search new search

return: Best matches first **from:** all documents

Search Hint: Connecting terms with AND narrows your search to give more precise results ...more on search operators

Results: 1 - 10 of 54 — 1 2 3 4 5 6 | Next

Show First Paragraph ☑

1. **The Washington Times** - July 13, 1999

Tin ears on Social Security

You would think congressional leaders who wanted to advance a personal retirement account option to Social Security would offer a proposal that could at least be supported by the organizations and individuals who have been developing the idea for years now. But you would be wrong.House Ways and Means Committee Chairman Bill Archer, Texas Republican, and Social Security Subcommittee Chairman Clay Shaw, Florida Republican, have advanced a proposal that is a gross disfigurement of what...

Purchase Complete Article, of 771 words

Obtain Phrase Counts from Newspapers

All databases   News & Newspapers databases   Preferences   English   Help

ProQuest

ProQuest Newsstand

"personal retirement account" AND pub(washington times)

Full text   Modify search   Tips

Suggested subjects Hide   Powered by ProQuest® Smart Search

Washington Times (Company/Org)   Washington Times (Company/Org) AND Newspapers

Washington Times (Company/Org) AND Moon, Sun Myung (Person)   Washington Times (Company/Org) AND Washington DC (Place)

53 Results *   Search within   Create alert   Create RSS feed   Save search

Select 1-20   Brief view | Detailed view

1 [Tin ears on Social Security: [2 Edition 1]]

Ferrara, Peter. **Washington Times** [Washington, D.C] 13 July 1999: A17.

...who wanted to advance a personal retirement account option to Social Security
...worker's wages into a personal retirement account for the worker. These payments
...and proposed a sound personal retirement account plan. He would have granted

Citation/Abstract   Full text

2 [Washington's financial miscreants sucker us]

Hurt, Charles. **Washington Times** [Washington, D.C] 05 Dec 2012: A.6.

...billions out of our personal retirement account to fund an obscene lavishness

Citation/Abstract   Full text

3 [Brickbats blur Bush proposal for Social Security ; Plan backers say 'truth' obscured]

Lambro, Donald. **Washington Times** [Washington, D.C] 06 Feb 2005: A03.

... Mr. Bush's personal retirement account (PRA) plan has been attacked by

Citation/Abstract   Full text

4 [Touching Social Security's hot third rail: [2 Edition]]

Lambro, Donald. **Washington Times** [Washington, D.C] 12 Dec 1996: A.17.

...Security taxes into their own personal retirement account. Mr. Forbes'

Citation/Abstract   Full text

Example: Social Security

- "House GOP offers plan for Social Security; Bush's **private accounts** would be scaled back" (*Washington Post*, 6/23/05)
- "GOP backs use of Social Security surplus ; Finds funding for **personal accounts**" (*Washington Times*, 6/23/05)

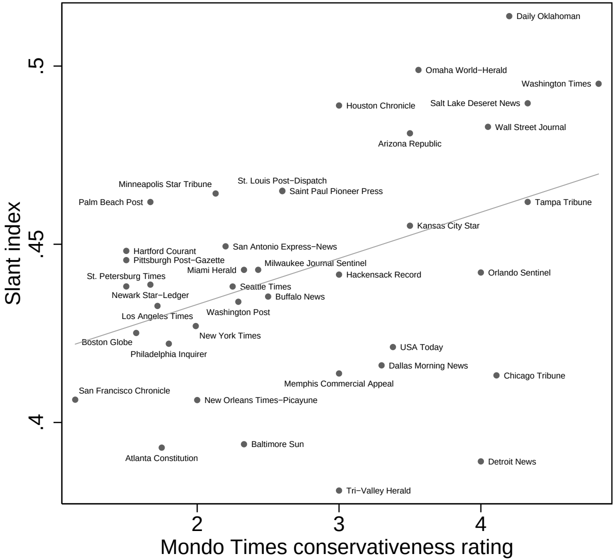
Linear Model of Phrase Frequency

- Let y_i be Republican vote share in senator i's state
- Let x_{ij} be share of senator $i's$ speech going to phrase j

$$\mathrm{E}\left(x_{ij}|y_i\right) = \alpha_j + \beta_j y_i$$

- Estimate via least squares
  - Procedure called *marginal regression*
- Apply same model to newspapers to infer y_m

Validation



Slant index
.5
.45
.4
2
3
4
Mondo Times conservativeness rating
Daily Oklahoman
Omaha World−Herald
Washington Times
Houston Chronicle
Salt Lake Deseret News
Wall Street Journal
Arizona Republic
Minneapolis Star Tribune
St. Louis Post−Dispatch
Saint Paul Pioneer Press
Palm Beach Post
Tampa Tribune
Kansas City Star
Hartford Courant
San Antonio Express−News
Pittsburgh Post−Gazette
Milwaukee Journal Sentinel
Miami Herald
St. Petersburg Times
Hackensack Record
Orlando Sentinel
Seattle Times
Newark Star−Ledger
Buffalo News
Washington Post
Los Angeles Times
New York Times
Boston Globe
USA Today
Philadelphia Inquirer
Dallas Morning News
Chicago Tribune
Memphis Commercial Appeal
San Francisco Chronicle
New Orleans Times−Picayune
Baltimore Sun
Atlanta Constitution
Detroit News
Tri−Valley Herald

How to Validate

- Newspaper rankings consistent with external sources
- Phrases make sense
- Sensitivity analysis
  - Change scores y_i (ADA, NOMINATE)
  - Change set of phrases
- Check agreement across sources
- Go look at the newspapers

Audit

TABLE A.I

AUDIT OF SEARCH RESULTS[a]

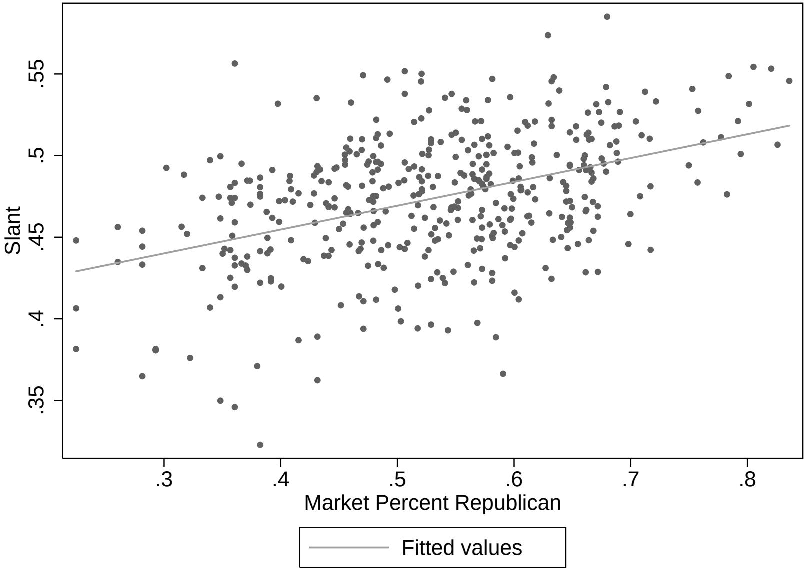
Phrase	Total Hits	Share of Hits in Quotes	Share of Hits That Are: AP Wire Stories	Other Wire Stories	Letters to the Editor	Maybe Opinion	Clearly Opinion	Independently Produced News
Global war on terrorism	2064	16%	3%	4%	1%	2%	10%	80%
Malpractice insurance	2190	5%	0%	0%	1%	3%	12%	84%
Universal health care	1523	9%	1%	0%	7%	8%	28%	56%
Assault weapons	1411	9%	3%	12%	4%	1%	25%	56%
Child support enforcement	1054	3%	0%	0%	1%	2%	11%	86%
Public broadcasting	3375	8%	1%	0%	2%	4%	22%	71%
Death tax	595	36%	0%	0%	2%	5%	46%	47%
Average (hit weighted)		10%	1%	2%	3%	3%	19%	71%

[a] Authors' calculations based on ProQuest and NewsLibrary data base searches. See Appendix A for details.

Economic Hypotheses

- Now that we have a measure, we can use it to model newspaper ideology
- Possible drivers
  - Consumer ideology
  - Owner ideology
  - Influence of incumbent politicians

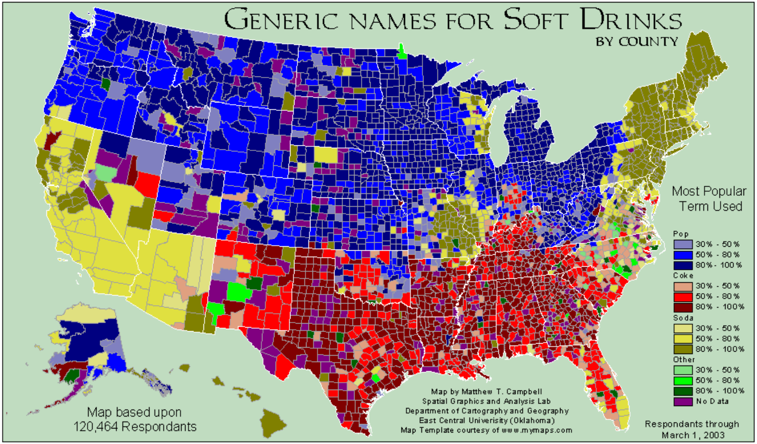
Role of Consumer Ideology



Possible Confounds

- Reverse causality
- Slant is proxying for other newspaper attributes (e.g. emphasis on sports vs. business)
- Slant is proxying for other market attributes (e.g. geography)

Soda vs. Pop



GENERIC NAMES FOR SOFT DRINKS
BY COUNTY
Most Popular Term Used
Pop
30% - 50%
50% - 80%
80% - 100%
Coke
30% - 50%
50% - 80%
80% - 100%
Soda
30% - 50%
50% - 80%
80% - 100%
Other
30% - 50%
50% - 80%
80% - 100%
No Data
Map based upon
120,464 Respondants
Map by Matthew T. Campbell
Spatial Graphics and Analysis Lab
Department of Cartography and Geography
East Central University (Oklahoma)
Map Template courtesy of www.mymaps.com
Respondants through
March 1, 2003

Solutions

- Control carefully for geography when relating slant to other variables
- Incorporate geography into predictive model
    - Predict the component of congressperson ideology that is orthogonal to Census division

Broader Lesson

- Can use predictive modeling as an aid to social science
- But
  - You get out what you put in
  - Consider possible sources of bias and misspecification

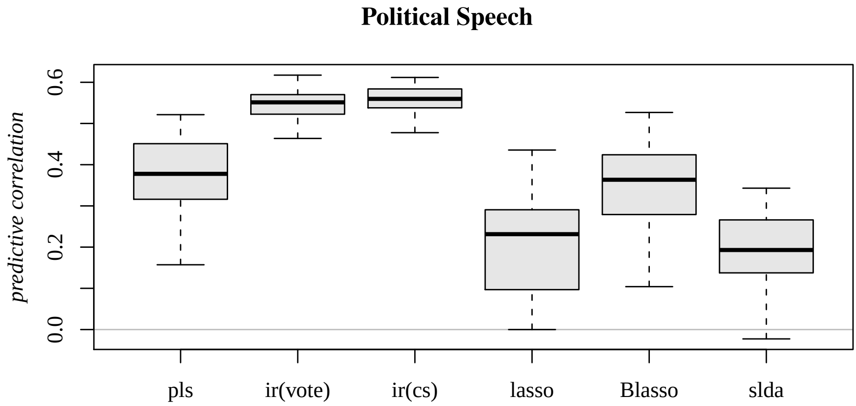
Taddy (2013)

- Two main limitations of Gentzkow & Shapiro (2010)
  - Feature selection separate from model estimation
  - Linear model doesn't exploit multinomial structure of data

Taddy (2013)

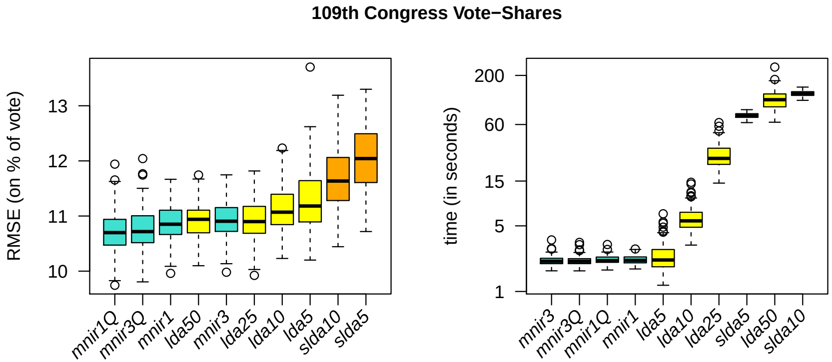
- Let y_i be Republican vote share in senator i's state
- Let x_{ijt} be an indicator for senator i says phrase j at occasion t
- Then
$$\Pr\left(x_{ij}=1\right)=\frac{\exp\left(\alpha_j+\beta_j y_i\right)}{\sum_{j'}\exp\left(\alpha_{j'}+\beta_{j'} y_i\right)}$$
- Estimate via maximum likelihood
  - Uses log penalty for regularization
  - Uses novel algorithm for maximization
- Penalty imposes sparsity in β_js
  - Means we can use a very large number of phrases p
  - Can think of this as a way to maximize performance subject to a phrase "budget"

Taddy (2013)



- Note: LDA = latent Dirichlet allocation (stay tuned)

Taddy (2013)



109th Congress Vote−Shares
RMSE (on % of vote)
13
12
11
10
mnir1Q
mnir3Q
mnir1
lda50
mnir3
lda25
lda10
lda5
slda10
slda5
time (in seconds)
200
60
15
5
1
mnir3
mnir3Q
mnir1Q
mnir1
lda5
lda10
lda25
slda5
lda50
slda10

Sentiment Analysis: *Financial News*

Overview

- Questions
  - What explains time-series/cross-section of equity returns?
  - Is there information beyond what is reflected in quantitative fundamentals (e.g. earnings)?

Tetlock (2007)

- Data: counts of words in WSJ “Abreast of the Market” column

Tetlock (2007)

- Features: Counts of words in each of 77 “Harvard-IV General Inquirer” categories
  - Weak
  - Positive
  - Negative
  - Active
  - Passive
  - etc.

List of entries in tag category:

Weak

List shows first 100 entries. Total number of entries in this category:

755

Entries for this category are shown with all tags assigned and sense definitions:

ABANDON

H4Lvd Negativ Ngtv Weak Fail IAV AffLoss AffTot SUPV

ABANDONMENT

H4 Negativ Weak Fail Noun

ABDICATE

H4 Negativ Weak Submit Passive Finish IAV SUPV

Tetlock (2007)

- Regressors
  - Weak words
  - Negative words
  - First principal component (“pessimism”)

Table II

Predicting Dow Jones Returns Using Negative Sentiment

The table data come from CRSP, NYSE, and the General Inquirer program. This table shows OLS estimates of the coefficient γ_1 in equation (1). Each coefficient measures the impact of a one-standard deviation increase in negative investor sentiment on returns in basis points (one basis point equals a daily return of 0.01%). The regression is based on 3,709 observations from January 1, 1984, to September 17, 1999. I use Newey and West (1987) standard errors that are robust to heteroskedasticity and autocorrelation up to five lags. Bold denotes significance at the 5% level; italics and bold denotes significance at the 1% level.

	Regressand: Dow Jones Returns		
News Measure	Pessimism	Negative	Weak
$BdNws_{t-1}$	***−8.1***	***−4.4***	***−6.0***
$BdNws_{t-2}$	0.4	3.6	2.0
$BdNws_{t-3}$	0.5	−2.4	−1.2
$BdNws_{t-4}$	**4.7**	***4.4***	***6.3***
$BdNws_{t-5}$	1.2	2.9	**3.6**
$\chi^2(5)$ [*Joint*]	***20.0***	***20.8***	***26.5***
p-value	0.001	0.001	0.000
Sum of 2 to 5	**6.8**	***9.5***	***10.7***
$\chi^2(1)$ [*Reversal*]	**4.05**	***8.35***	***10.1***
p-value	0.044	0.004	0.002

Antweiler and Frank (2004)

- Data: Message board contents on Yahoo! Finance and Raging Bull

```
---------------------
FROM YF
COMP ETYS
MGID 13639
NAME CaptainLihai
LINK 1
DATE 2000/01/25 04:11
SKIP
TITL ETYS will surprise all pt II
SKIP
TEXT ETYS will surprise all when it drops to below 15$ a pop, and even then
TEXT it will be too expensive.
TEXT
TEXT If the DOJ report is real, there will definately be a backlash against
TEXT the stock. Watch your asses. Get out while you can.
---------------------
FROM YF
COMP IBM
MGID 43653
NAME plainfielder
LINK 1
DATE 2000/03/29 11:39
SKIP
TITL BUY ON DIPS - This is the opportunity
SKIP
TEXT to make $$$ when IBM will be going up again following this profit taking
TEXT bout by Abbey Cohen and her brokerage firm.
TEXT
TEXT IBM shall go up again after today.
-----------------
```

Antweiler and Frank (2004)

- Count words
- Create training set of 1000 messages hand-coded as buy, sell, hold
- Compute "naive Bayes classification:" posterior guess assuming words are independent

Table I
Naive Bayes Classification Accuracy within Sample and Overall Classification Distribution

The first percentage column shows the actual shares of 1,000 hand-coded messages that were classified as buy (B), hold (H), or sell (S). The buy-hold-sell matrix entries show the in-sample prediction accuracy of the classification algorithm with respect to the learned samples, which were classified by the authors (Us).

Classified:		By Algorithm		
by Us	%	Buy	Hold	Sell
Buy	25.2	18.1	7.1	0.0
Hold	69.3	3.4	65.9	0.0
Sell	5.5	0.2	1.2	4.1
1,000 messages[a]		21.7	74.2	4.1
All messages[b]		20.0	78.8	1.3

[a]These are the 1,000 messages contained in the training data set.
[b]This line provides summary statistics for the out-of-sample classification of all 1,559,621 messages.

Antweiler and Frank (2004)

- Small amount of predictability in returns
- Messages predict volatility
- Disagreement (variable recommendations) predicts volume

Other Examples

- Li (2010): Uses naive Bayes to measure sentiment of forward-looking statements in 10Ks/10Qs
- Hanley and Hoberg (2012): Use cosine distance to measure revisions to IPO prospectuses

Topic Models

Factor Models

- "Unsupervised" methods (factor analysis, PCA) project high-dimensional data into low-dimensional measures, preserving as much variation as possible.

Factor Models

- "Unsupervised" methods (factor analysis, PCA) project high-dimensional data into low-dimensional measures, preserving as much variation as possible.
- E.g.,
    - Congressional roll call votes $\rightarrow$"Common space" scores
    - Survey responses $\rightarrow$ "Big 5" personality traits

Factor Models

- “Unsupervised” methods (factor analysis, PCA) project high-dimensional data into low-dimensional measures, preserving as much variation as possible.
- E.g.,
  - Congressional roll call votes $\rightarrow$”Common space” scores
  - Survey responses $\rightarrow$ “Big 5” personality traits
- Low dimensional measures are then inputs into subsequent analysis

Factor Models

- "Unsupervised" methods (factor analysis, PCA) project high-dimensional data into low-dimensional measures, preserving as much variation as possible.
- E.g.,
  - Congressional roll call votes $\rightarrow$"Common space" scores
  - Survey responses $\rightarrow$ "Big 5" personality traits
- Low dimensional measures are then inputs into subsequent analysis
- E.g.,
  - How has polarization in Congress changed over time? (Poole & Rosenthal 1984)
  - How does personality correlate with job performance (Tett et al. 1991)

Topic Models

- Topic models extend these methods to multinomial data such as text
- Relevant to measuring, e.g.,
  - What people talk about on social networks
  - What products share similar descriptions on Amazon / EBay
  - What "stories" are in the news today
  - What are economists studying

Purpose

- As with other unsupervised methods, topic models are of most interest to social scientists as an input into subsequent analysis

Purpose

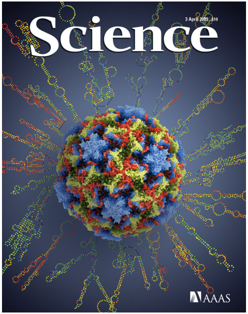
- As with other unsupervised methods, topic models are of most interest to social scientists as an input into subsequent analysis
- E.g.,
  - Do discussions of particular topics on Twitter predict stock movements?
  - Which products are close substitutes on EBay?
  - Is media slant driven by what you talk about or how you talk about it?
  - How has the distribution of topics in economics changed over time?

Purpose

- As with other unsupervised methods, topic models are of most interest to social scientists as an input into subsequent analysis
- E.g.,
  - Do discussions of particular topics on Twitter predict stock movements?
  - Which products are close substitutes on EBay?
  - Is media slant driven by what you talk about or how you talk about it?
  - How has the distribution of topics in economics changed over time?
- A fair critique of topic modeling literature is that it hasn't progressed much beyond the measurement stage

Topic Models: *Blei & Lafferty (2006)*

Input

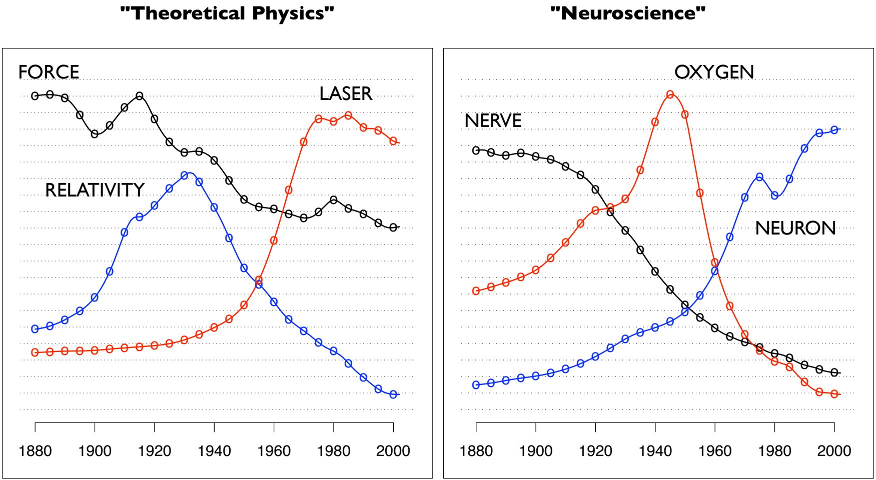


- OCR text of *Science* 1880-2002 (from JSTOR)
    - Count words used 25 or more times (after stemming and removing stopwords)
    - Vocabulary: 15,955 words
    - Total documents: 30,000 articles

Output

1	2	3	4
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

Output


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

Model: LDA

- Latent Dirichlet Allocation (LDA) introduced by Blei et al. (2003) as an extension of factor models to discrete data

Model: LDA

- Setup
  - Documents $i \in \{1, ..., n\}$
  - Words $j \in \{1, ..., p\}$
  - Data $\mathbf{x}_i$ is $(1 \times p)$ vector of word counts for document i

Model: LDA

- Setup
  - Documents $i \in \{1, ..., n\}$
  - Words $j \in \{1, ..., p\}$
  - Data $\mathbf{x}_i$ is $(1 \times p)$ vector of word counts for document i
- Factor model
  - θ_{ik} is value of k-th **factor** for document i
  - $\boldsymbol{\beta}_k$ is $(1 \times p)$ vector of **loadings** for factor k

$$E(\mathbf{x}_i) = \boldsymbol{\beta}_1 \theta_{i1} + ... \boldsymbol{\beta}_K \theta_{iK}$$

Model: LDA

- Setup
  - Documents $i \in \{1, ..., n\}$
  - Words $j \in \{1, ..., p\}$
  - Data $\mathbf{x}_i$ is $(1 \times p)$ vector of word counts for document i
- Factor model
  - θ_{ik} is value of k-th **factor** for document i
  - $\boldsymbol{\beta_k}$ is $(1 \times p)$ vector of **loadings** for factor k

$$E(\mathbf{x}_i) = \boldsymbol{\beta}_1 \theta_{i1} + ...\boldsymbol{\beta}_K \theta_{iK}$$

- LDA
  - θ_{ik} is weight on k-th **topic** for document i
  - $\boldsymbol{\beta_k}$ is $(1 \times p)$ vector of **word probabilities** for topic k

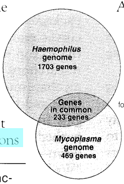
$$\mathbf{x}_i \sim Multinomial(\boldsymbol{\beta}_1 \theta_{i1} + ...\boldsymbol{\beta}_K \theta_{iK})$$

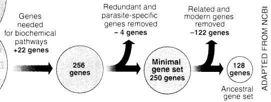
Model: LDA

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



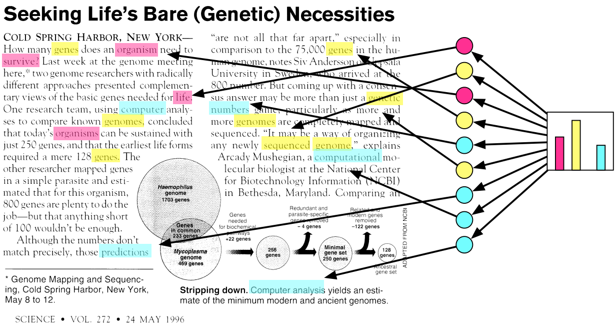


Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Model: 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.

Model: LDA

- For each document i...
  - Draw topic proportions θ_i from Dirichlet distribution with parameter α
  - For each word j...
    - Draw a topic assignment $k \sim Multinomial(\theta_i)$
    - Draw word $x_{ij} \sim Multinomial\left(\beta_k\right)$

Model: Dynamic

- One limitation of LDA is it assumes documents are exchangeable; in many settings of interest, topics evolve systematically over time

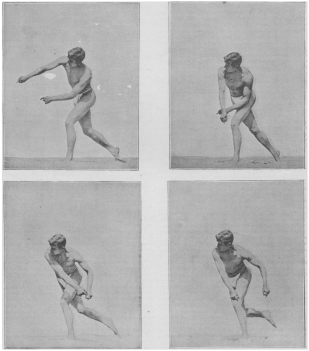
Model: Dynamic

"Instantaneous Photography" (1890)

Prussia, who has taken thousands of pictures of flying birds, running horses, jumping men, etc., all admirable for their perfect ''technique,'' and for the great artistic tact and scientific skill with which the moments of exposure had been chosen. In these pictures the characteristic positions peculiar to different motions are well presented. Many of them at first appear absolutely unnatural, because the eye has never been able to observe them.

These pictures produced rich and important material for the study of motion, but Mr. Anschuetz succeeded in making his experiments more valuable by obtaining whole series of pictures giving the different phases of motion. He made it his object to get of one period of motion, for instance, of the step of a walking man, as many views as possible in equal intervals o time, and he succeeded admirably in his undertaking. He was able to observe in this manner even the fastest motion, for instance, the hurdle-jump of a racing horse, which occupies only seventy-two one-hundredths of a second, and in this short time made twenty-four pictures of the different positions in equal intervals. A dozen pictures showing the different phases of position assumed by an athlete in throwing a javelin, reproduced from instantaneous photographs taken by Mr. Anschuetz, are given on this and the preceding pages.

Mr. Anschuetz next constructed an apparatus which he called the electric tachyscope, in which be was financially assisted by the German Government. In this instrument the series o.



INSTANTANEOUS PHOTOGRAPHS OF AN ATHLETE THROWING A JAVELIN.

"Infrared Reflectance in Leaf-Sitting Neotropical Frogs" (1977)

North American frogs so examined [*Bufo debilis*, *B. boreas* (2), *B. coniferus*; *Rana pipiens* (2), *R. palmipes*, *R. catesbeiana*; *Hyla cinerea*, *H. squirella*, *H. euphorbiacea*, *H. chaneque*, and *H. cyanomma*] absorb infrared light and stand out sharply against foliage (Fig. 1).

Cott (*3*), using black and white infrared film, found that the Australian tree-frog *Hyla coerulea* (=*Litoria caerulea*) reflects infrared light. *Litoria caerulea*, *A. moreletii*, and *A.* (=*Pachymedusa*) *dacnicolor* all contain a newly discovered red pigment in unusual melanosomes (*4*). Both *fleischmanni* and *prosoblepon* groups of *Centrolenella* contain a purple pigment in their chromatophores (*5*). Whether these two skin pigments are identical, or play any role in infrared reflectance, has not been determined.

There are two likely functions for infrared reflectance in leaf-sitting frogs. (i) Although the near-infrared is not heat (*6*), photons of these wavelengths will lose energy as heat if they are absorbed by the skin. Thus, the ability to reflect infrared may play a physiological role in thermoregulation by preventing excessive heat gain. (ii) Infrared reflectance may conceal frogs from predators with infrared receptors (*3*). Little research has been done on near-infrared sensitivity, and supportive evidence is sparse. Both the eyes of birds and the pit organs of snakes may act as near-infrared light receptors. In pigeons and chickens, the sensitivity maxima of the eyes are shifted toward longer wavelengths than those of humans (*7*), and the tawny owl responds to infrared light (900 nm) (*8*). Visual sensitivity extending just into the near-infrared would allow birds to see most green frogs on green leaves, although centrolenids and phyllomedusines would remain camouflaged. Boid and crotaline pit organs are usually interpreted as thermal detectors, adaptations for nocturnal predation on warm-blooded prey (*9*). In diurnal snakes, however, these receptors may be used to detect frogs that act as infrared sinks among leaves that are reflecting light of these wavelengths. The facial pits of crotaline snakes are directionally sensitive and may allow infrared depth perception (*10*). Many species of birds and snakes are known to eat frogs and forage in their diurnal retreats. Predation by birds and snakes may have selected for infrared cryptic coloration in tropical leaf-sitting frogs.

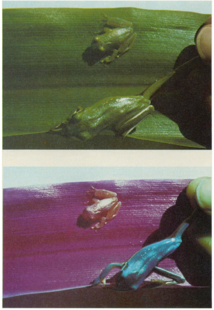


Fig. 1. A comparison of the color characteristics of a hylid and a centrolenid frog in a conventional (top) and an infrared (bottom) color photograph. Although both frogs match the green leaf in light ranges visible to man, only *Centrolenella fleischmanni* (top frog) reflects near-infrared light. This allows it to blend with foliage both in the visible and near-infrared ranges of light, unlike *Hyla cinerea* (bottom frog), which absorbs infrared and is distinguished from the leaf surface in an infrared photograph.

PATRICIA A. SCHWALM*
PRISCILLA H. STARRETT
Department of Biological Sciences, Allan Hancock Foundation, University of Southern California, Los Angeles 90007
ROY W. MCDIARMID
Department of Biology, University of South Florida, Tampa 33620

References and Notes

1. Kodak Infrared Ektachrome film has a sensitivity extending to about 900 nm [*Kodak Publ. N-17* (1974)]. The three emulsion layers are sensitized to green, red, and infrared light rather than to blue, green, and red. By placing a yellow filter over the camera lens, blue light is excluded, resulting in a shift of colors in the photograph. Green objects appear blue, red objects appear green, and infrared appears red.
2. H. L. Gibson, W. R. Buckley, K. E. Whitmore, *J. Biol. Photogr. Assoc.* **33**, 1 (1965).
3. H. B. Cott, *Adaptive Coloration in Animals* (Methuen, London, 1940), pp. 9–11.
4. J. T. Bagnara and W. Ferris, *Copeia*, **3**, 592 (1975).
5. P. H. Starrett and J. M. Savage, *Bull. South. Calif. Acad. Sci.* **72**, 57 (1973).
6. Radiation in these wavelengths (700 to 900 nm) can be emitted by very hot objects, such as the sun or lamp filaments, but it is not heat per se. Similarly, the near-infrared can be reflected or emitted as luminescence by objects that are not hot themselves [*Kodak Publ. M-28* (1972)].
7. R. Granit, *Sensory Mechanisms of the Retina* (Oxford Univ. Press, London, 1947), pp. 294–296.
8. F. L. Vanderplank, *Proc. Zool. Soc. London* **1934**, 505 (1934).
9. P. H. Hartline, in *Handbook of Sensory Physiology*, vol. 3, part 3, *Electroreceptors and Other Specialized Receptors in Lower Vertebrates*, A. Fessard, Ed. (Springer-Verlag, New York, 1974), pp. 297–312.

Model: Dynamic

- Divide text into sequential slices (e.g., by year)
- Assume each slice's documents drawn from LDA model
- Allow word distribution within topics β and distribution over topics α to evolve via markov process

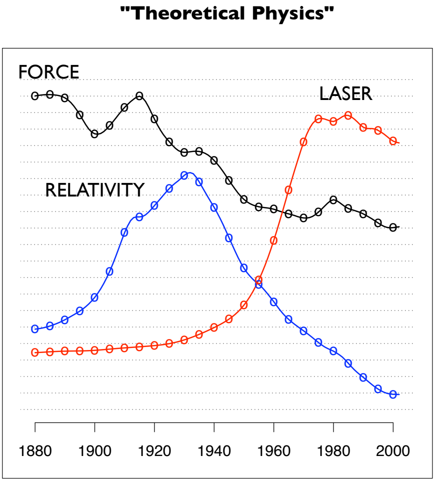
Estimation

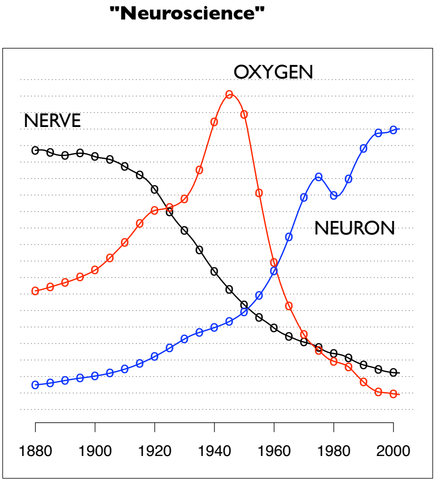
- Bayesian inference intractable using standard methods (e.g., Gibbs sampling)
  - Blei (2006) $\rightarrow$ variational inference
  - Taddy R package $\rightarrow$ MAP estimation
  - Current favorite $\rightarrow$ Stochastic gradient descent
- Main estimates are for 20 topic model

Results: LDA

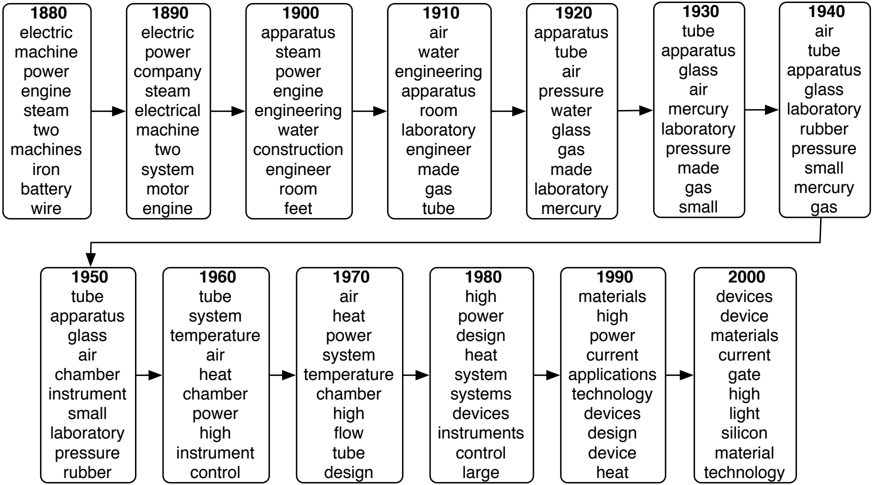
1	2	3	4
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Results: Dynamic



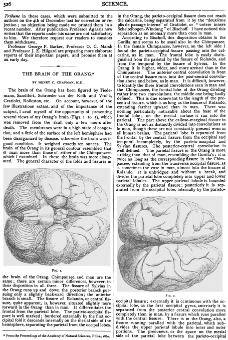


Results: Dynamic



Results: Dynamic

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.

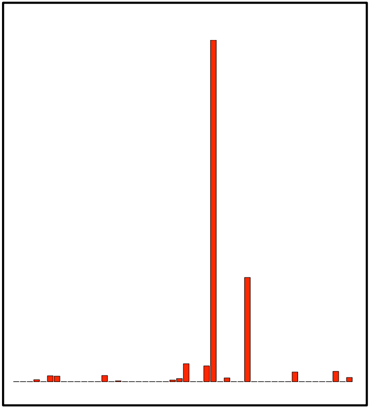
* From the Proceedings of the Academy of Natural Sciences, Phila., 1880.

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-

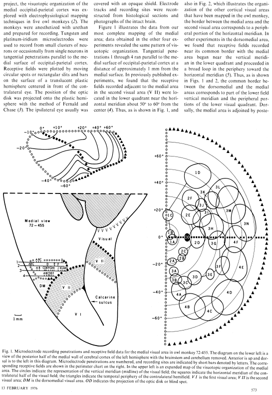
Fig. 2.

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



Results: Dynamic

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

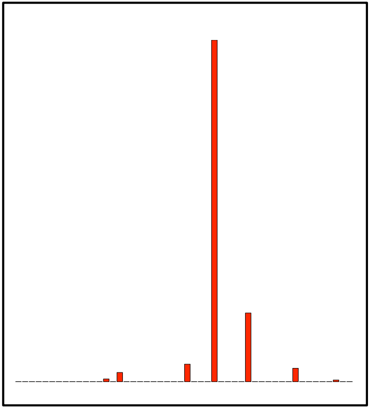


project, the visuotopic organization of the medial occipital-parietal cortex was explored with electrophysiological mapping techniques in five owl monkeys (*2*). The monkeys were anesthetized with urethan and prepared for recording. Tungsten and platinum-iridium microelectrodes were used to record from 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-

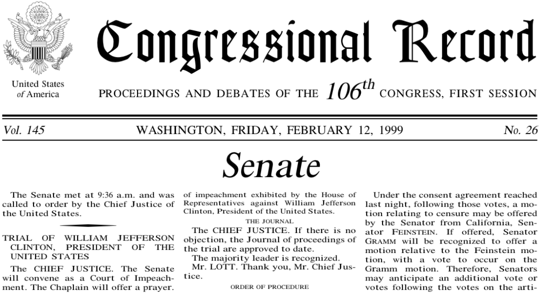
Fig. 1. Microelectrode recording penetrations and receptive field data for the medial visual area in owl monkey 72-455. The diagram on the lower left is a view of the posterior half of the medial wall of cerebral cortex of the left hemisphere with the brainstem and cerebellum removed. Anterior is up 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



Topic Models: *Quinn et al. (2010)*

Input



United States
of America

Congressional Record

PROCEEDINGS AND DEBATES OF THE *106th* CONGRESS, FIRST SESSION

Vol. 145 WASHINGTON, FRIDAY, FEBRUARY 12, 1999 *No. 26*

Senate

The Senate met at 9:36 a.m. and was called to order by the Chief Justice of the United States.

TRIAL OF WILLIAM JEFFERSON CLINTON, PRESIDENT OF THE UNITED STATES

The CHIEF JUSTICE. The Senate will convene as a Court of Impeachment. The Chaplain will offer a prayer. of impeachment exhibited by the House of Representatives against William Jefferson Clinton, President of the United States.

THE JOURNAL

The CHIEF JUSTICE. If there is no objection, the Journal of proceedings of the trial are approved to date.

The majority leader is recognized.

Mr. LOTT. Thank you, Mr. Chief Justice.

ORDER OF PROCEDURE

Under the consent agreement reached last night, following those votes, a motion relating to censure may be offered by the Senator from California, Senator FEINSTEIN. If offered, Senator GRAMM will be recognized to offer a motion relative to the Feinstein motion, with a vote to occur on the Gramm motion. Therefore, Senators may anticipate an additional vote or votes following the votes on the arti-

- Full text of speeches in US Senate 1995-2004
    - Count words appearing in 0.5% or more of speeches (after stemming)
    - Vocabulary: $3,807$ words
    - Total documents: $118,065$ speeches

Model

- Like Blei & Lafferty (2006), except
  1. Each document is in exactly one topic
  2. Dynamic distribution of topics, but topics themselves are static

Model

- Blei & Lafferty (2006)
  - $\mathbf{x}_i \sim Multinomial\left(\boldsymbol{\beta}_1 \theta_{i1} + ...\boldsymbol{\beta}_K \theta_{iK}\right)$
  - $\boldsymbol{\theta_i} \sim F\left(\alpha\right)$
  - β and α both evolve over time
- Quinn et al. (2010)
  - $\mathbf{x}_i \sim Multinomial\left(\beta_{k(i)}\right)$
  - $Pr\left(k\left(i\right) = j\right) = \alpha_j$
  - α evolves over time; β constant

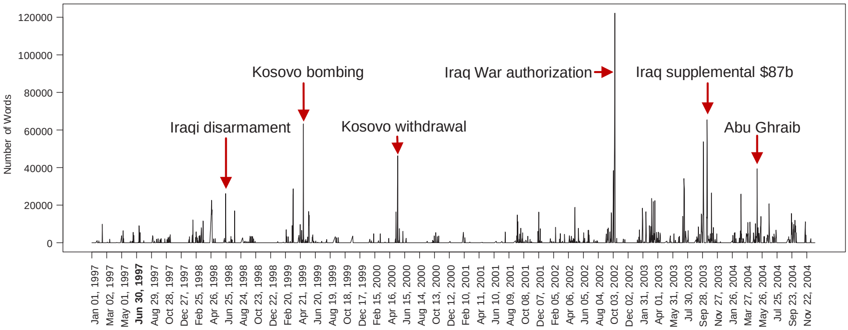
Estimation

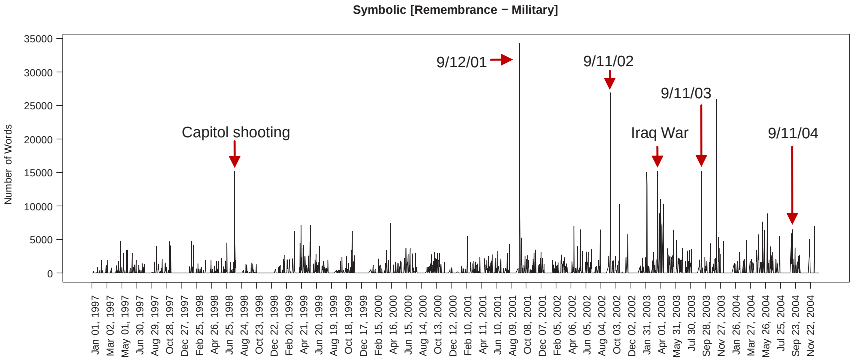
- Estimate using ECM algorithm
- Main estimates are for 42 topic model (chosen based on “substantive and conceptual” criteria)

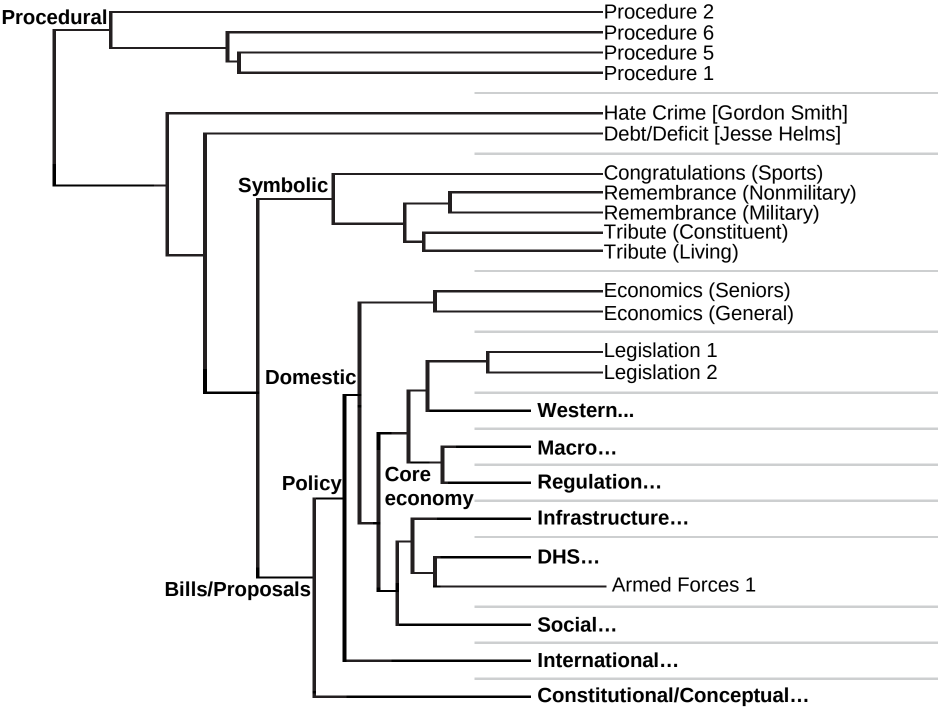
Table 3 Topic Keywords for 42-Topic Model

Topic (Short Label)	Keys
1. Judicial Nominations	*nomine, confirm, nomin, circuit, hear, court, judg, judici, case, vacanc*
2. Constitutional	*case, court, attornei, supreme, justic, nomin, judg, m, decis, constitut*
3. Campaign Finance	*campaign, candid, elect, monei, contribut, polit, soft, ad, parti, limit*
4. Abortion	*procedur, abort, babi, thi, life, doctor, human, ban, decis, or*
5. Crime 1 [Violent]	*enforc, act, crime, gun, law, victim, violenc, abus, prevent, juvenil*
6. Child Protection	*gun, tobacco, smoke, kid, show, firearm, crime, kill, law, school*
7. Health 1 [Medical]	*diseas, cancer, research, health, prevent, patient, treatment, devic, food*
8. Social Welfare	*care, health, act, home, hospit, support, children, educ, student, nurs*
9. Education	*school, teacher, educ, student, children, test, local, learn, district, class*
10. Military 1 [Manpower]	*veteran, va, forc, militari, care, reserv, serv, men, guard, member*
11. Military 2 [Infrastructure]	*appropri, defens, forc, report, request, confer, guard, depart, fund, project*
12. Intelligence	*intellig, homeland, commiss, depart, agenc, director, secur, base, defens*
13. Crime 2 [Federal]	*act, inform, enforc, record, law, court, section, crimin, internet, investig*
14. Environment 1 [Public Lands]	*land, water, park, act, river, natur, wildlif, area, conserv, forest*
15. Commercial Infrastructure	*small, busi, act, highwai, transport, internet, loan, credit, local, capit*
16. Banking / Finance	*bankruptci, bank, credit, case, ir, compani, file, card, financi, lawyer*
17. Labor 1 [Workers]	*worker, social, retir, benefit, plan, act, employ, pension, small, employe*

Defense [Use of Force]




Symbolic [Remembrance − Military]
Number of Words
35000
30000
25000
20000
15000
10000
5000
0
9/12/01
9/11/02
9/11/03
Capitol shooting
Iraq War
9/11/04
Jan 01, 1997
Mar 02, 1997
May 01, 1997
Jun 30, 1997
Aug 29, 1997
Oct 28, 1997
Dec 27, 1997
Feb 25, 1998
Apr 26, 1998
Jun 25, 1998
Aug 24, 1998
Oct 23, 1998
Dec 22, 1998
Feb 20, 1999
Apr 21, 1999
Jun 20, 1999
Aug 19, 1999
Oct 18, 1999
Dec 17, 1999
Feb 15, 2000
Apr 16, 2000
Jun 15, 2000
Aug 14, 2000
Oct 13, 2000
Dec 12, 2000
Feb 10, 2001
Apr 11, 2001
Jun 10, 2001
Aug 09, 2001
Oct 08, 2001
Dec 07, 2001
Feb 05, 2002
Apr 06, 2002
Jun 05, 2002
Aug 04, 2002
Oct 03, 2002
Dec 02, 2002
Jan 31, 2003
Apr 01, 2003
May 31, 2003
Jul 30, 2003
Sep 28, 2003
Nov 27, 2003
Jan 26, 2004
Mar 27, 2004
May 26, 2004
Jul 25, 2004
Sep 23, 2004
Nov 22, 2004


Procedural
Procedure 2
Procedure 6
Procedure 5
Procedure 1
Hate Crime [Gordon Smith]
Debt/Deficit [Jesse Helms]
Symbolic
Congratulations (Sports)
Remembrance (Nonmilitary)
Remembrance (Military)
Tribute (Constituent)
Tribute (Living)
Economics (Seniors)
Economics (General)
Domestic
Legislation 1
Legislation 2
Western...
Macro…
Core economy
Regulation…
Policy
Infrastructure…
DHS…
Armed Forces 1
Bills/Proposals
Social…
International…
Constitutional/Conceptual…

Conclusion

Conclusion

- Today
  - Prediction with high-dimensional data
  - Applications
- Tomorrow
  - Estimating treatment effects with high-dimensional data
  - Application