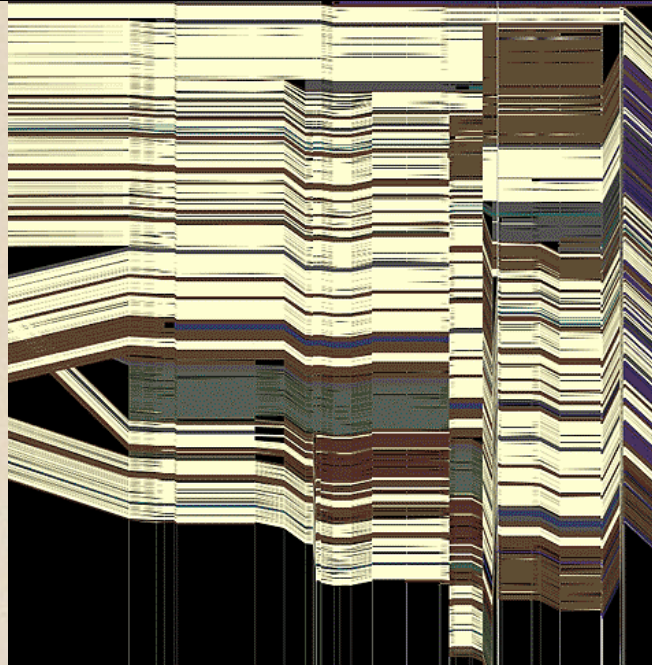
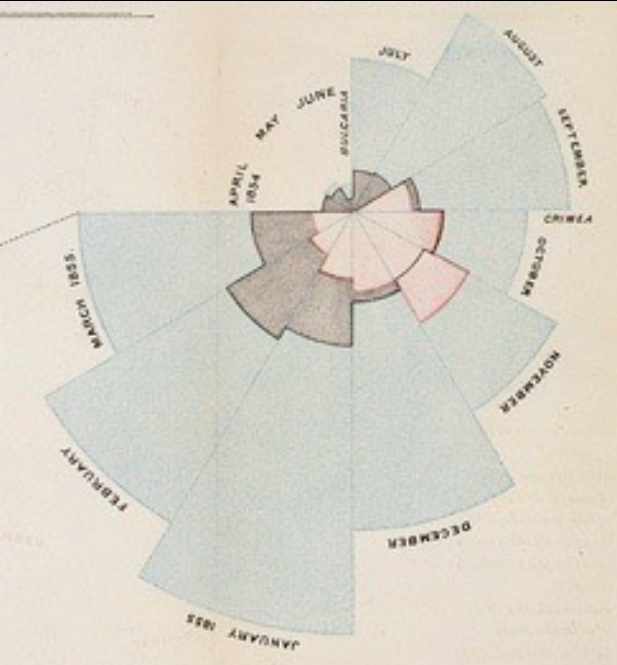


CSE 442 - Data Visualization

Text Visualization



Jeffrey Heer University of Washington

Why Visualize Text?

Why Visualize Text?

Understanding – get the “gist” of a document

Grouping – cluster for overview or classification

Comparison – compare document collections, or inspect evolution of collection over time

Correlation – compare patterns in text to those in other data, e.g., correlate with social network

Text as Data

Documents

Articles, books and novels

E-mails, web pages, blogs

Tags, comments

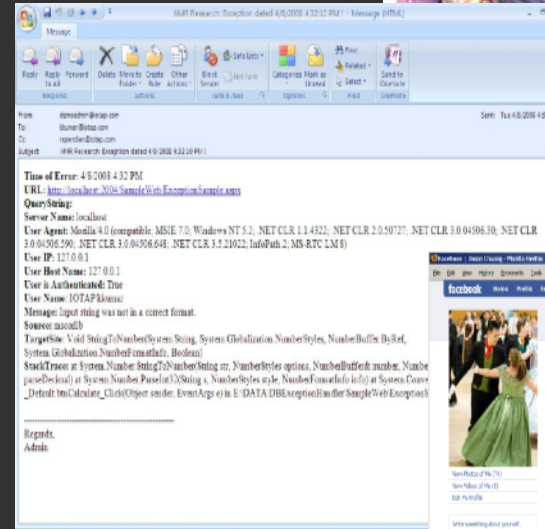
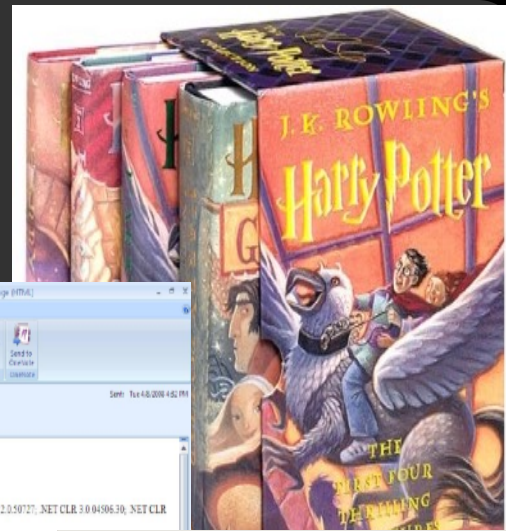
Computer programs, logs

Collections of Documents

Messages (e-mail, blogs, tags, comments)

Social networks (personal profiles)

Academic collaborations (publications)



Example:
Health Care Reform

Example: Health Care Reform

Recent History

Initiatives by President Clinton
Overhaul by President Obama

Text Data

News articles
Speech transcriptions
Legal documents

What questions might you want to answer?
What visualizations might help?

A Concrete Example

September 10, 2009

TEXT

Obama's Health Care Speech to Congress

Following is the prepared text of President Obama's speech to Congress on the need to overhaul health care in the United States, as released by the White House.

Madame Speaker, Vice President Biden, Members of Congress, and the American people:

When I spoke here last winter, this nation was facing the worst economic crisis since the Great Depression. We were losing an average of 700,000 jobs per month. Credit was frozen. And our financial system was on the verge of collapse.

As any American who is still looking for work or a way to pay their bills will tell you, we are by no means out of the woods. A full and vibrant recovery is many months away. And I will not let up until those Americans who seek jobs can find them; until those businesses that seek capital and credit can thrive; until all responsible homeowners can stay in their homes. That is our ultimate goal. But thanks to the bold and decisive action we have taken since January, I can stand here with confidence and say that we have pulled this economy back from the brink.

I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for our nation.

But we did not come here just to clean up crises. We came to build a future. So tonight, I return to speak to all of you

Word Tree: Word Sequences

Visualizations : Word Tree President Obama's Address to Congress on Health Care

Search Start End



Search **i will**

Back

Forward

Start End

Occurrence Order

Clicks Will Zoom

12 hits

i will

not

let up until those americans who seek jobs can find them - - (applause) - - until those

back down on the basic principle that if americans can't find affordable coverage , w

sign

a plan that adds one dime to our deficits - - either now or in the future

it if it adds one dime to the deficit , now or in the future , period .

make that same mistake with health care .

waste time with those who have made the calculation that it's better politics to kill thi

- - and i will not accept the status quo as a solution .

accept the status quo as a solution .

make sure that no government bureaucrat or insurance company bureaucrat gets between you and t

protect medicare .

continue to seek common ground in the weeks ahead .

be there to listen .

still believe

we can

act even when it's hard .

replace economy with civility , and gridlock with progress .

do great things , and that here and now we will meet history's test .

- - i still believe that we can act when it's hard .

that we can act when it's hard .

Gulfs of Evaluation

Many text visualizations do not represent the text directly. They represent the output of a **language model** (word counts, word sequences, etc.).

- Can you interpret the visualization? How well does it convey the properties of the model?
- Do you trust the model? How does the model enable us to reason about the text?

Text Visualization Challenges

High Dimensionality

Where possible use text to represent text...
... which terms are the most descriptive?

Context & Semantics

Provide relevant context to aid understanding.
Show (or provide access to) the source text.

Modeling Abstraction

Determine your analysis task.
Understand abstraction of your language models.
Match analysis task with appropriate tools and models.

Topics

Text as Data

Visualizing Document Content

Visualizing Conversation

Document Collections

Text as Data

Words as nominal data?

High dimensional (10,000+)

More than equality tests

Words have meanings and relations

- Correlations: *Hong Kong, Puget Sound, Bay Area*
- Order: *April, February, January, June, March, May*
- Membership: *Tennis, Running, Swimming, Hiking, Piano*
- Hierarchy, antonyms & synonyms, entities, ...

Text Processing Pipeline

1. Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#huskies, @UW, OMG!!!!!!!!!!*

Entities? *Washington State, O'Connor, U.S.A.*

Text Processing Pipeline

1. Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#huskies, @UW, OMG!!!!!!!!!!*

Entities? *Washington State, O'Connor, U.S.A.*

2. Stemming

Group together different forms of a word.

Porter stemmer? *visualization(s), visualize(s), visually -> visual*

Lemmatization? *goes, went, gone -> go*

Text Processing Pipeline

1. Tokenization

Segment text into terms.

Remove stop words? *a, an, the, of, to, be*

Numbers and symbols? *#huskies, @UW, OMG!!!!!!!!!!*

Entities? *Washington State, O'Connor, U.S.A.*

2. Stemming

Group together different forms of a word.

Porter stemmer? *visualization(s), visualize(s), visually -> visual*

Lemmatization? *goes, went, gone -> go*

3. Ordered list of terms

Bag of Words Model

Ignore ordering relationships within the text

A document \approx vector of term weights

- Each dimension corresponds to a term (10,000+)
- Each value represents the relevance

For example, simple term counts

Aggregate into a document-term matrix

- Document vector space model

Document-Term Matrix

Each document is a vector of term weights

Simplest weighting is to just count occurrences

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

Tag Clouds

Strengths

Can help with gisting and initial query formation.

Weaknesses

Sub-optimal visual encoding (size vs. position)

Inaccurate size encoding (long words are bigger)

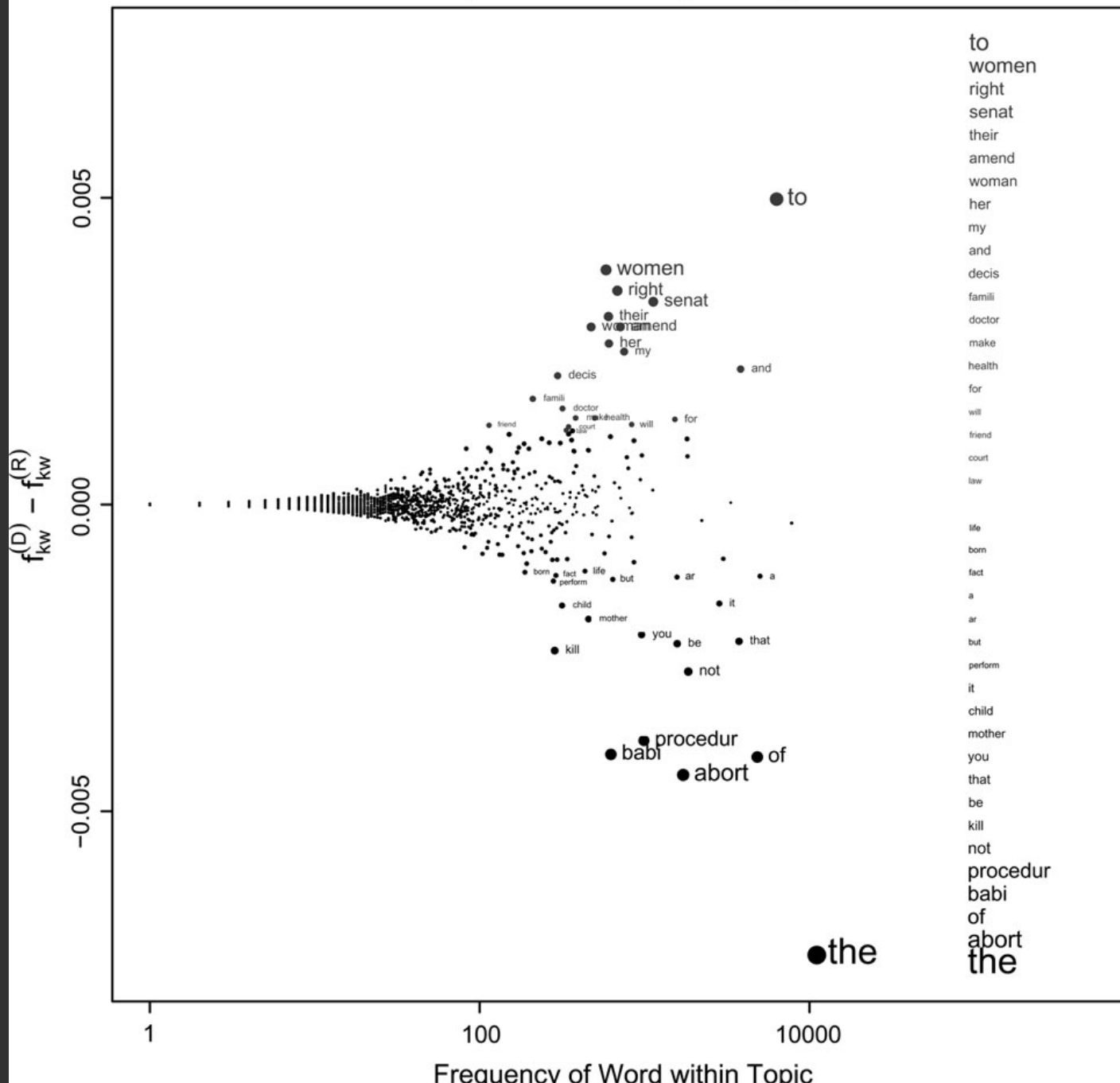
May not facilitate comparison (unstable layout)

Term frequency may not be meaningful

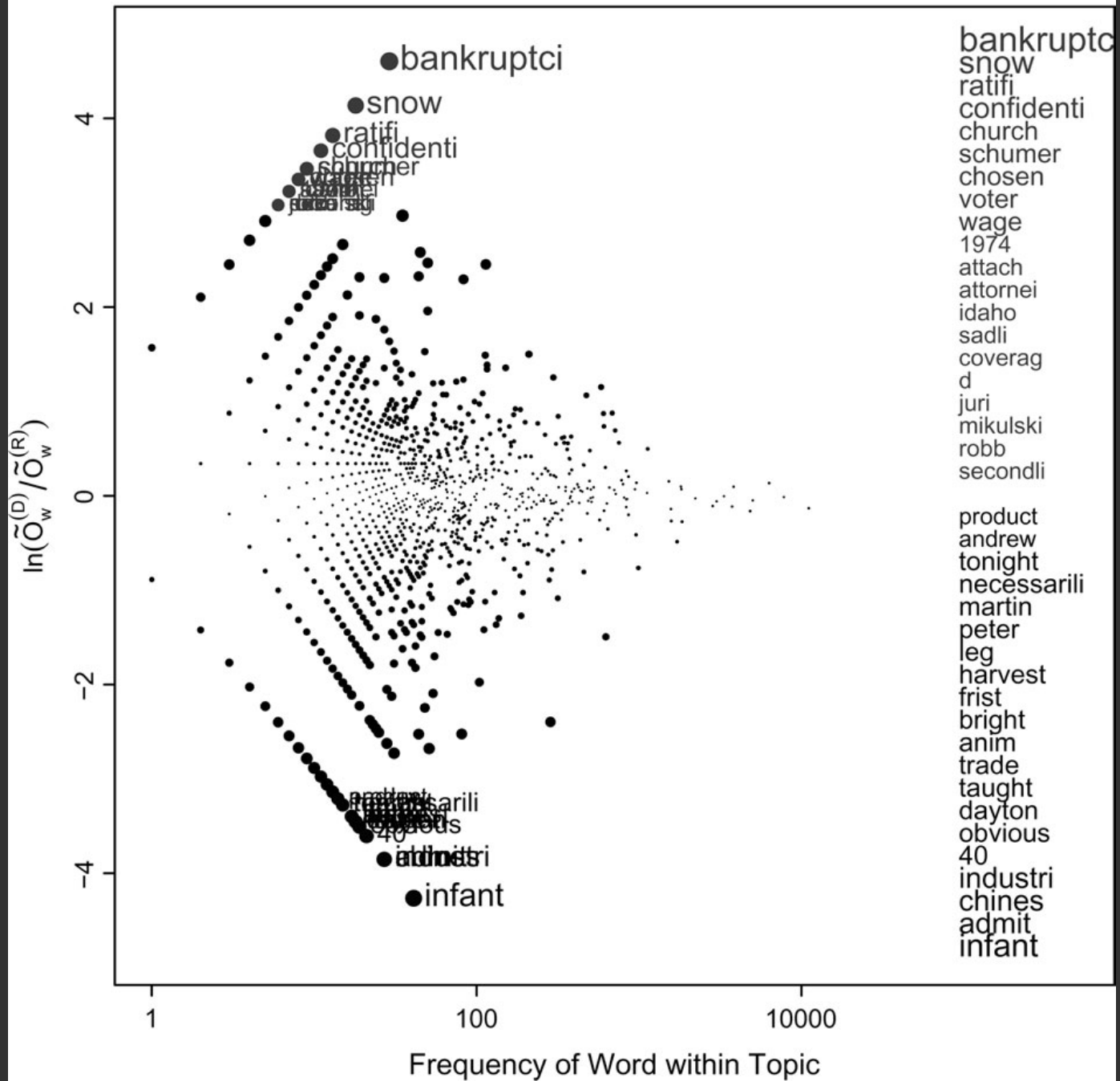
Does not show the structure of the text

**Given a text, what are the
best descriptive words?**

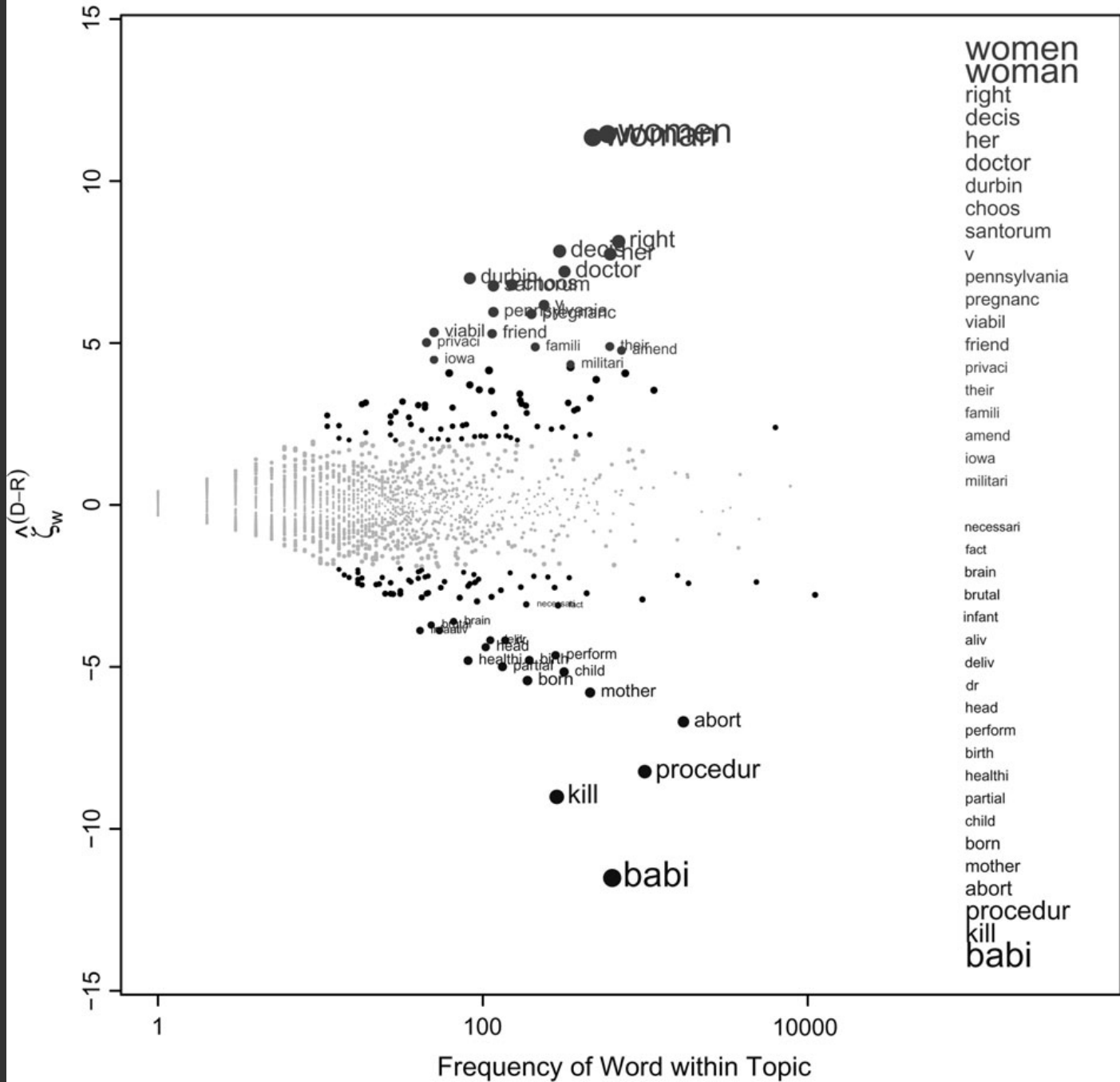
Partisan Words, 106th Congress, Abortion (Difference of Proportions)



Partisan Words, 106th Congress, Abortion
 (Log-Odds-Ratio, Smoothed Log-Odds-Ratio)



Partisan Words, 106th Congress, Abortion (Weighted Log-Odds-Ratio, Informative Dirichlet Prior)



Keyword Weighting

Term Frequency

$tf_{td} = \text{count}(t) \text{ in } d$

Can take log frequency: $\log(1 + tf_{td})$

Can normalize to show proportion: $tf_{td} / \sum_t tf_{td}$

Keyword Weighting

Term Frequency

$$tf_{td} = \text{count}(t) \text{ in } d$$

TF.IDF: Term Freq by Inverse Document Freq

$$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$$

$df_t = \# \text{ docs containing } t; N = \# \text{ of docs}$

Keyword Weighting

Term Frequency

$$tf_{td} = \text{count}(t) \text{ in } d$$

Require comparison
across full corpus!



TF.IDF: Term Freq by Inverse Document Freq

$$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_t)$$

df_t = # docs containing t ; N = # of docs

G²: Probability of different word frequency

$$E_1 = |d| \times (tf_{td} + tf_{t(C-d)}) / |C|$$

$$E_2 = |C-d| \times (tf_{td} + tf_{t(C-d)}) / |C|$$

$$G^2 = 2 \times (tf_{td} \log(tf_{td}/E_1) + tf_{t(C-d)} \log(tf_{t(C-d)}/E_2))$$

Limitations of Freq. Statistics

Typically focus on unigrams (single terms)

Often favors frequent (TF) or rare (IDF) terms

Not clear that these provide best description

A “bag of words” ignores additional information

Grammar / part-of-speech

Position within document

Recognizable entities

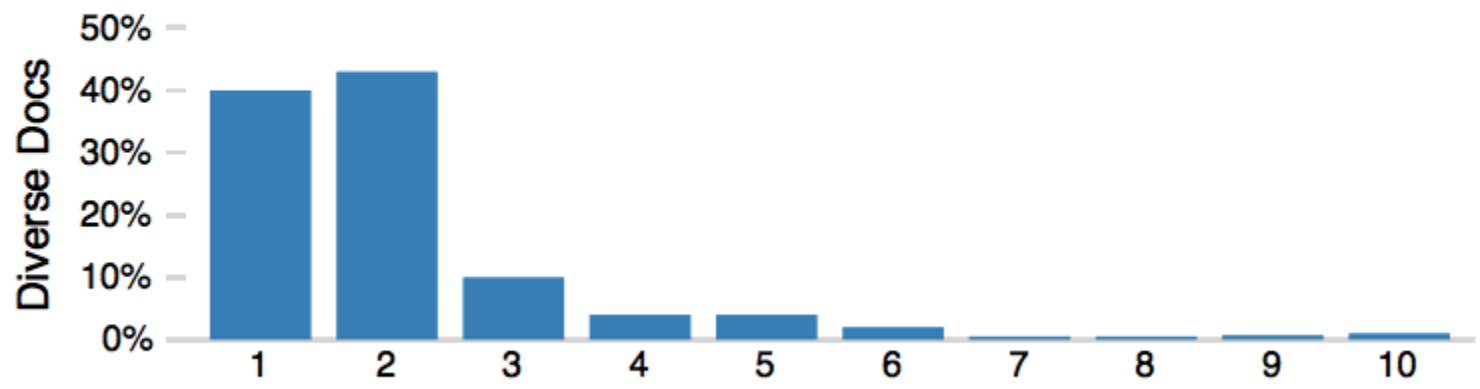
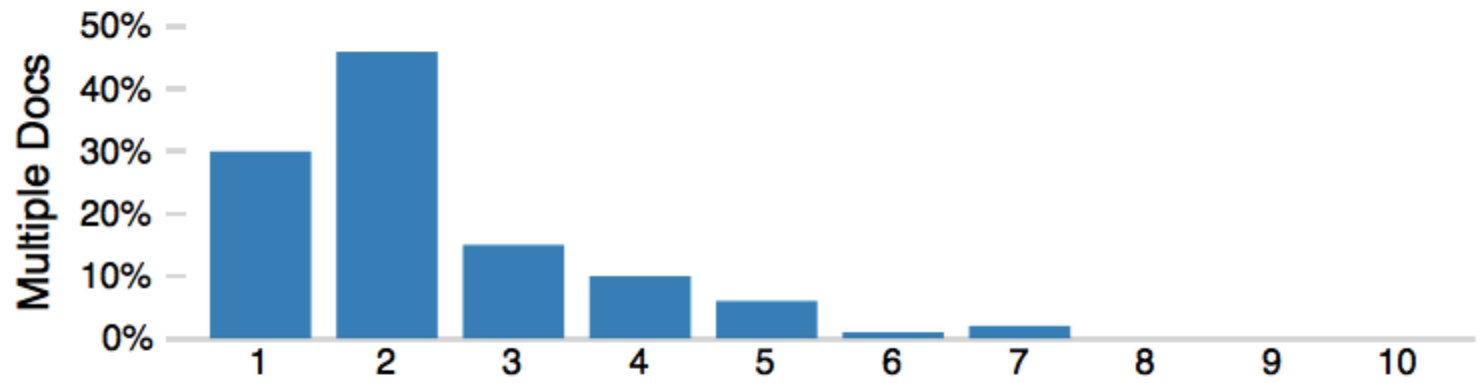
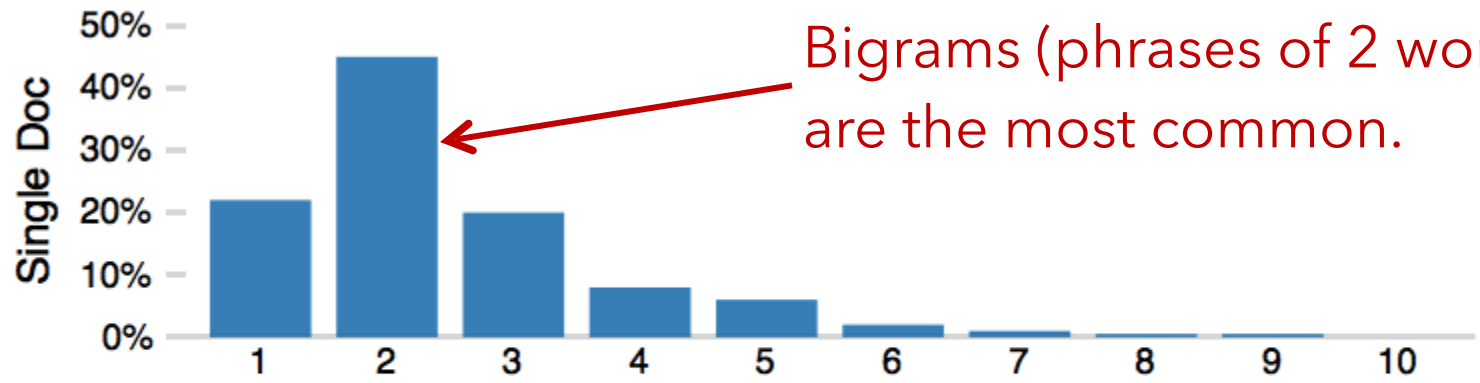
How do people describe text?

We asked 69 subjects (graduate students) to read and describe dissertation abstracts.

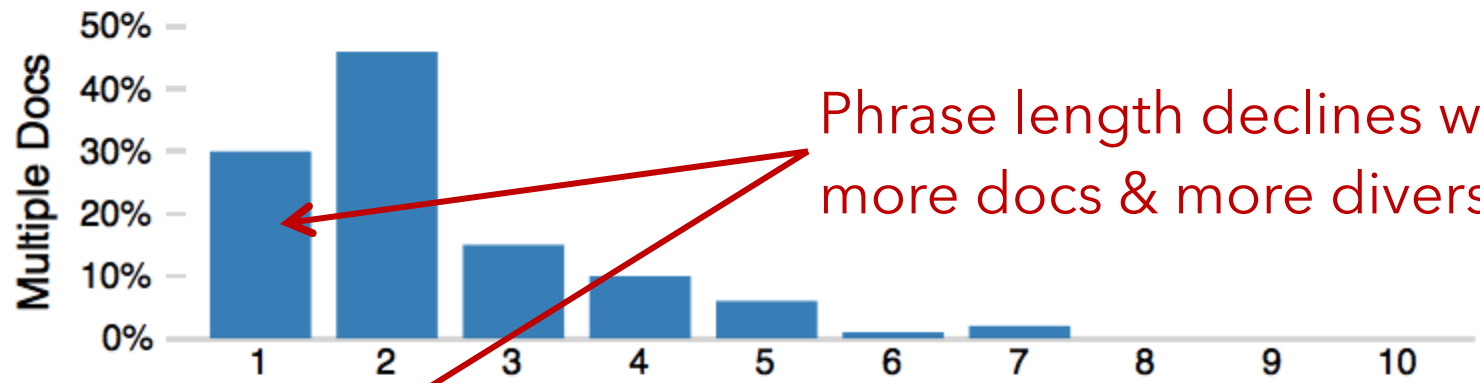
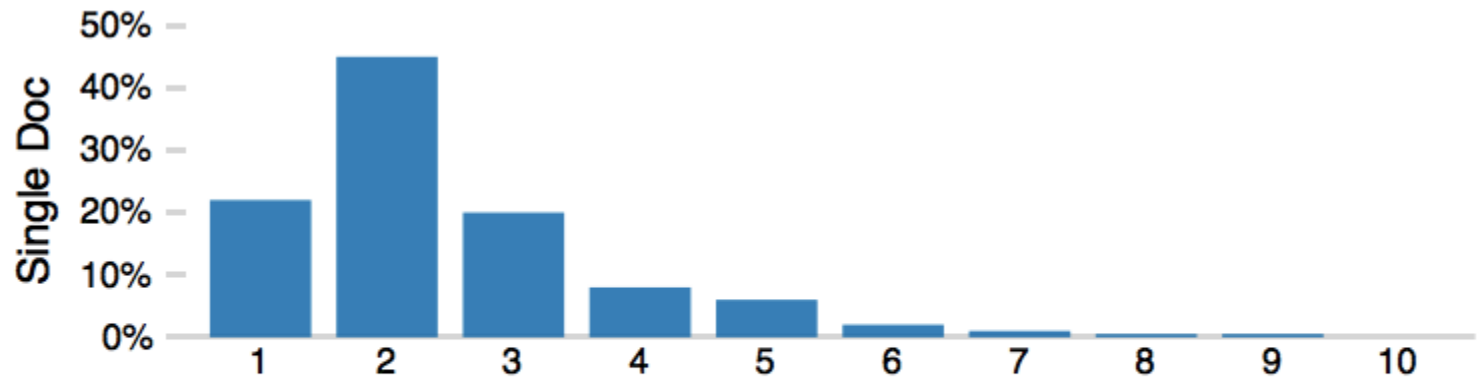
Students were given 3 documents in sequence; they then described the collection as a whole.

Students were matched to both *familiar* and *unfamiliar* topics; *topical diversity* within a collection was varied systematically.

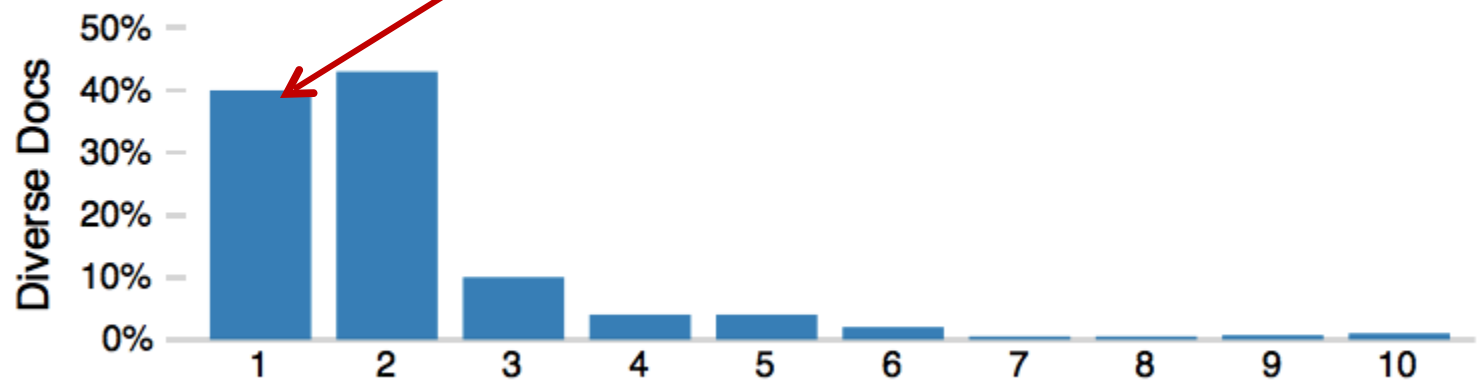
[Chuang, Manning & Heer, 2012]



Phrase Length



Phrase length declines with more docs & more diversity.



Phrase Length

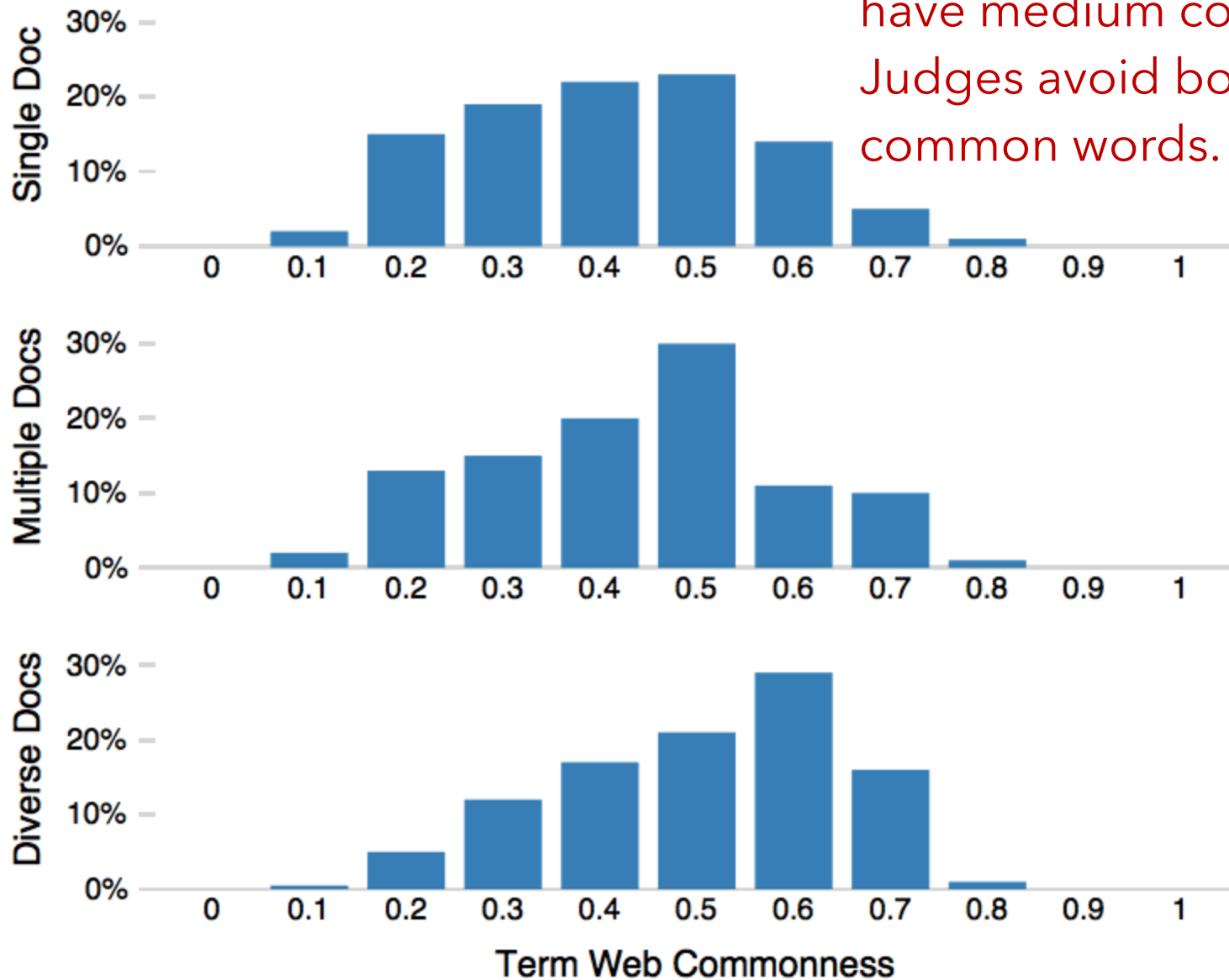
Term Commonness

$$\log(\text{tf}_w) / \log(\text{tf}_{\text{the}})$$

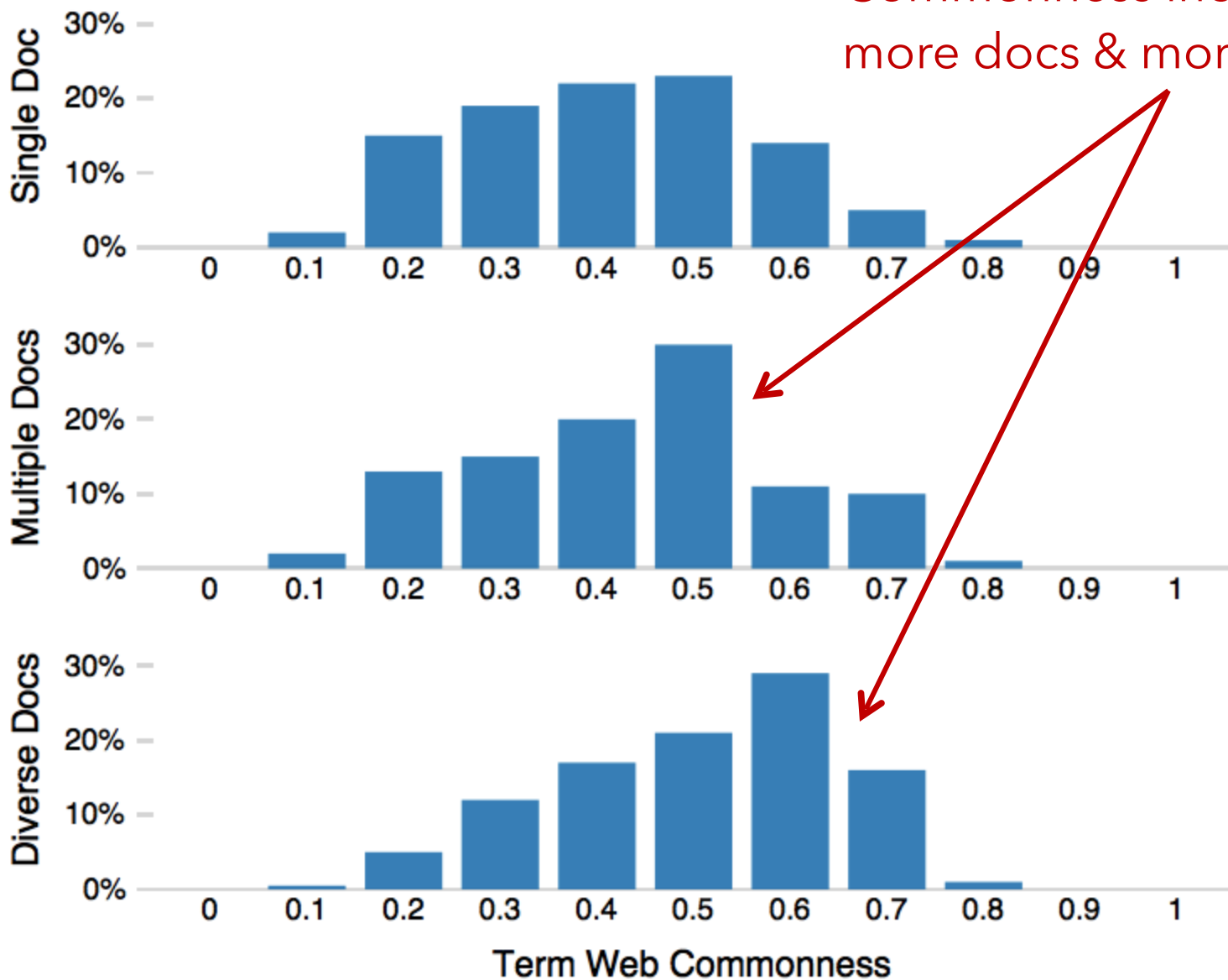
The normalized term frequency relative to the most frequent n-gram, e.g., the word “the”.

Measured across a corpus or across the entire English language (using Google n-grams)

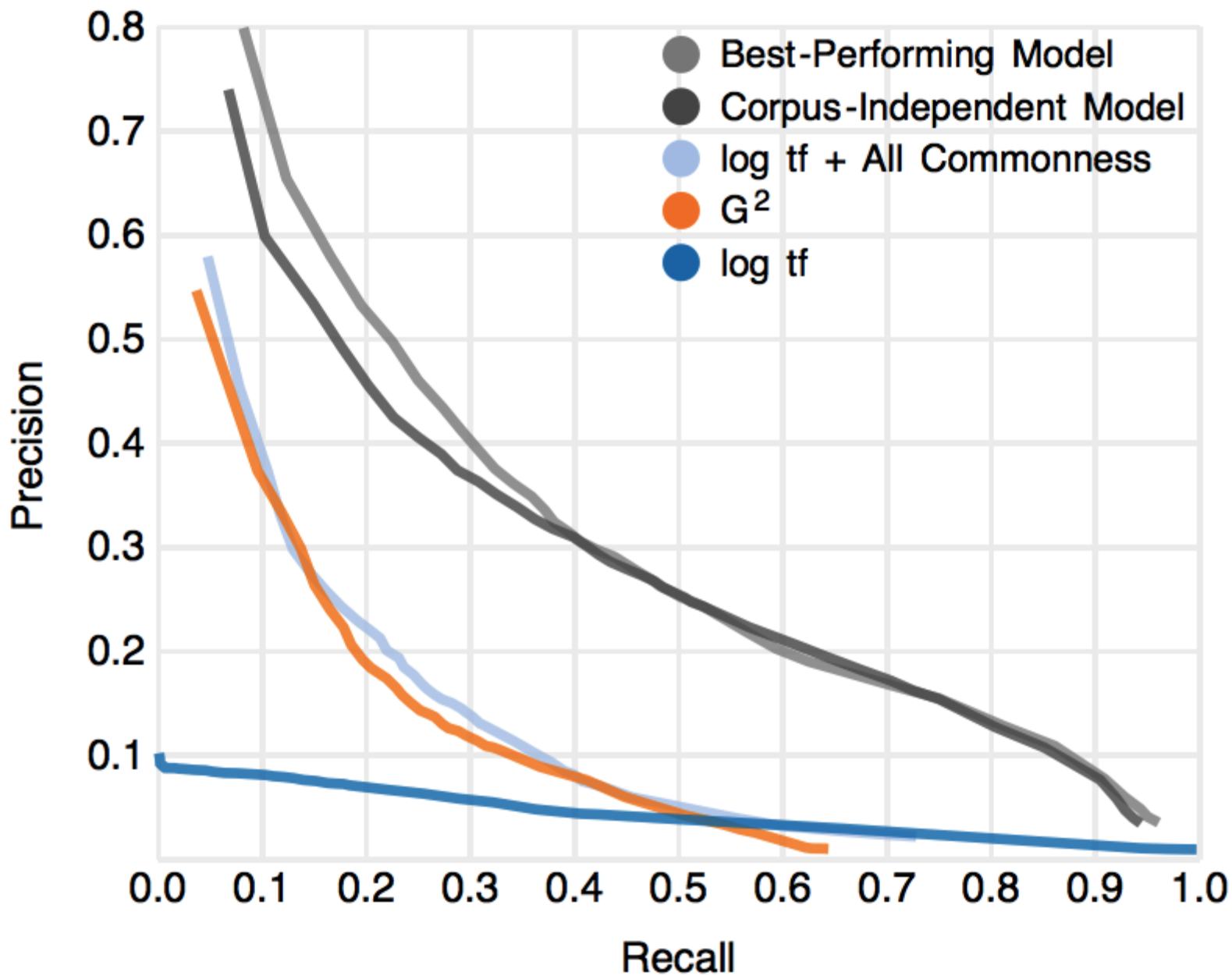
Selected descriptive terms have medium commonness. Judges avoid both rare and common words.



Commonness increases with more docs & more diversity.



Scoring Terms with Freq, Grammar & Position





A fighter jet rain check

Story and video by [Chamila Jayaweera](#)

Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.

"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

To find out more about how the process works and watch the action unfold, click above to see the video story.



CHAMILA JAYAWEERA/BOEING

The Water Check team rolls in a large metal frame, which they affectionately call their "spray tree," over a Super Hornet inside a St. Louis hangar.



G²

Regression Model

fighter

F/A

Hornet

Super

Boeing

-18

rain

St.

jet

Louis

15-inch-per-hour

douse

hangar

water-tight

Check

Baxter

sea-based

aircraft

Rich

seep

click

Navy

sure

Water

moisture

watch

enormous

stay

want

Super Hornet

F/A -18

fighter jet

Boeing engineers

special process

rain check

electronics suite

Program experts

simulated rain

ultimate customers

enormous hangar

water-tight integrity

Rich Baxter

15-inch-per-hour rate

video story

aircraft

U.S. Navy fighter pilots

Super Hornet final assembly manager

U.S.
Navy fighter
fighter pilot
sea-based fighter

Yelp Review Spotlight (Yatani 2011)

'09 amazing around baked bar bass best chef delicious eat

elite e

hawaii

night

expe

sake

tabl

b) best sf
baked sea bass best sushi sure in striped bass
other person
fresh fish slow service sushi bar
sushi chef baked mussel more hour
only thing
long wait long time sushi restaurant good food
long line hawaiian roll reasonable price
baked mango
small place delicious everything

Mentioned 63 times

possess sage of the halos wisdom , and know in advance sushi zone only accepts cash and the waits will be long and arduous .

yes , its a long wait , learn the master of zen if you want to eat here .

Tips: Descriptive Phrases

Understand the limitations of your language model.

Bag of words:

- Easy to compute

- Single words

- Loss of word ordering

Select appropriate model and visualization

- Generate longer, more meaningful phrases

- Adjective-noun word pairs for reviews

- Show keyphrases within source text

Document Content

Information Retrieval

Search for documents

Match query string with documents

Visualization to **contextualize results**

The screenshot shows a Google Scholar search interface. The search bar contains the text 'acronym resolution'. Below the search bar, there are several filters: 'Articles and patents', 'anytime', and 'include citations'. There is also a 'Create email alert' button. The search results are listed below, each with a title, author information, a brief abstract, and citation information.

Google scholar acronym resolution Search [Advanced Scholar Search](#)

Scholar Articles and patents anytime include citations [Create email alert](#)

[A supervised learning approach to acronym identification](#) [\[PDF\] from nrc-cnrc.ca](#)
D Nadeau, P Turney - *The Eighteenth Canadian ...*, 2005 - nparc.cisti-icist.nrc-cnrc.gc.ca
... Recently the fields of Genetics and Medicine have become especially interested in **acronym resolution** (Pustejovsky et al., 2001, Yu et al. 2002). ... Pustejovsky et al.'s **acronym resolution** technique searches for definitions of acronyms within noun phrases. ...
[Cited by 48](#) - [Related articles](#) - [All 16 versions](#)

[Biomedical term mapping databases](#) [\[HTML\] from nih.gov](#)
JD Wren, JT Chang, J Pustejovsky... - *Nucleic acids ...*, 2005 - Oxford Univ Press
... the prevalence of polynoms, or acronyms with multiple definitions. An important part of any high-throughput effort to tie experimental findings to published knowledge within the scientific literature involves **acronym resolution**. ...
[Cited by 41](#) - [Related articles](#) - [All 22 versions](#)

[Anthropogenic climate change over the Mediterranean region simulated by a global variable resolution model](#) [Find it@Stanford](#)
AL Gibelin... - *Climate Dynamics*, 2003 - Springer
... The long simulations CC and CS are split into two 30-year datasets CC1 and CS1 for the period 1960–1989 and CC2 and CS2 for the period 2070–2099 Full name **Acronym Resolution** Period Coupled Coupled control CC T63 1950–2099 Yes ...
[Cited by 197](#) - [Related articles](#) - [BL Direct](#) - [All 5 versions](#)

[Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises.](#)
MS Tuttle, NE Olson, KD Keck, WG Cole... - *Methods of information ...*, 1998 - ukpmc.ac.uk
... Title not supplied (PMID:10566483). Concept definition and manipulation are supported through

User Query
(Enter words for different topics on different lines.)

osteoporosis
prevention
research

Run Search New Query Quit

Search Limit: ◇ 50 ◇ 100 ◆ 250 ◇ 500 ◇ 1000

Number of Clusters: ◇ 3 ◇ 4 ◆ 5 ◇ 8 ◇ 10

Mode: TileBars

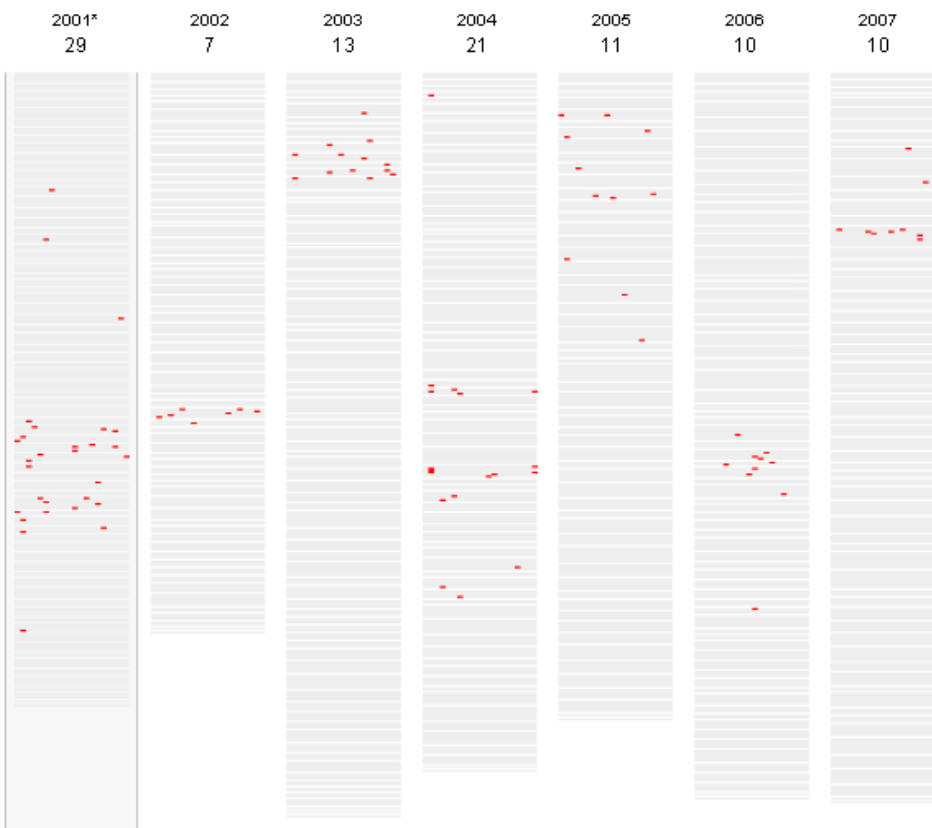
Cluster Titles Backup

Cluster	Titles
	FR88513-0157
	AP: Groups Seek \$1 Billion a Year for Aging Research
	SJMN: WOMEN'S HEALTH LEGISLATION PROPOSED C...
	AP: Older Athletes Run For Science
	FR: Committee Meetings
	FR: October Advisory Committees; Meetings
	FR88120-0046
	FR: Chronic Disease Burden and Prevention Models; Program
	AP: Survey Says Experts Split on Diversion of Funds for AIDS
	FR: Consolidated Delegations of Authority for Policy Developm...
	SJMN: RESEARCH FOR BREAST CANCER IS STUCK IN P...

The 2007 State of the Union Address

Over the years, President Bush's State of the Union address has averaged almost 5,000 words each, meaning the the President has delivered over 34,000 words. Some words appear frequently while others appear only sporadically. Use the tools below to analyze what Mr. Bush has said.

Use of the phrase "Tax" in past State of the Union Addresses



The word in context

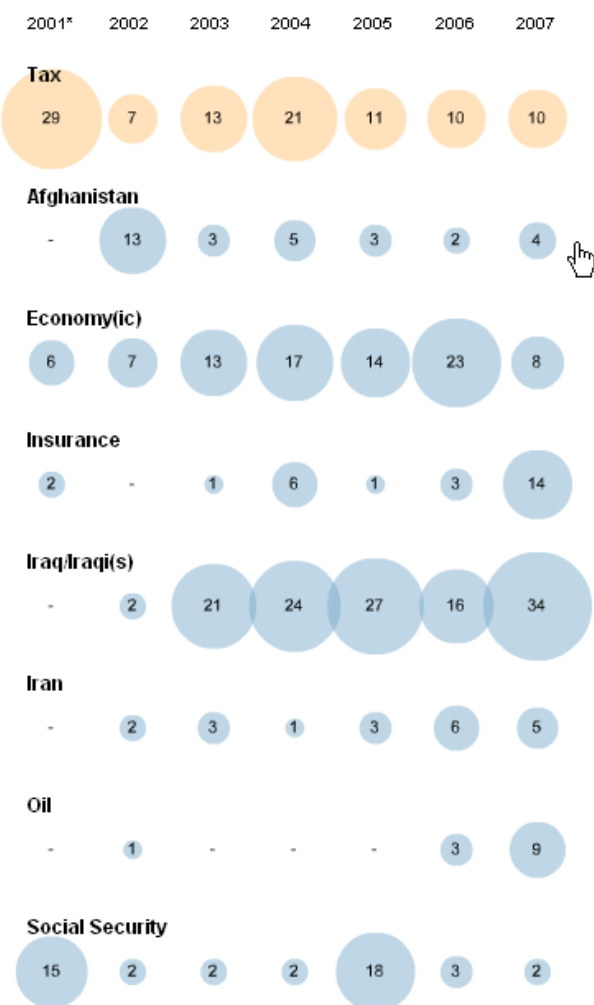
[Next Instance of 'Tax'](#)

I believe in local control of schools. We should not, and we will not, run public schools from Washington, D.C. Yet when the federal government spends **TAX** dollars, we must insist on results. Children should be tested on basic reading and math skills every year between grades three and eight. Measuring is the only way to know whether all our children are learning. And I want to know, because I refuse to leave any child behind in America.

-- 2001 (Paragraph 14 of 73)

New York Times

Compared with other words



* As a newly elected president, Mr. Bush did not deliver a formal State of the Union address in 2001. His Feb. 27 speech to a joint session of Congress was analogous to the State of the Union, but without the title.

Concordance

What is the common local context of a term?

The screenshot shows the 'Concordance - Larkin.Concordance' window. The interface includes a menu bar (File, Text, Search, Edit, Headwords, Contexts, View, Tools, Help), a toolbar with icons for file operations and text formatting, and a main display area with a table of concordance results. The table has columns for 'Headword', 'No.', 'Context...', 'Word', '...Context', and 'Reference'. The 'HEART' headword is selected, showing 25 occurrences. The status bar at the bottom provides summary statistics: Words: 7318, Tokens: 37070, At word: 2990, Deleted lines: 1 [24], Word sort: Asc alpha (string), Context sort: Asc occurrence order.

Headword	No.	Context...	Word	...Context	Reference
HEAR	15	That my own	heart	drifts and cries, having no...	Deep Analysis
HEARD	9	By the shout of the	heart	continually at work	And the wave
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	,	Träumerei
HEARSE	1	Because I follow it to my own	heart		Many famous
HEART	25	My	heart	is ticking like the sun:	I am washed u
HEART'S	2	The vague	heart	sharpened to a candid co...	The March Pa
HEART-SHAPED	1	Contract my	heart	by looking out of date.	Lines on a Yo
HEARTH	1	Having no	heart	to put aside the theft	Home is so Se
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Bea
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the
HEAT	6	These I would choose my	heart	to lead	After-Dinner F
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spac
HEATH	1	This petrified	heart	has taken,	A Stone Churc
HEATS	1	How should they sweep the girl clean...	heart	,	I see a girl dra
HEAVE	1	Hands that the	heart	can govern	Heaviest of flc
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessed-at	heart	riding	One man walk
HEAVIER-THAN...	1	If hands could free you,	heart	,	If hands could
HEAVIEST	2	That overflows the	heart		Pour away the

if love be rough with you , be rough with love .

if love be blind , love cannot hit the mark .

if love be blind , it best agrees with night .

if love be

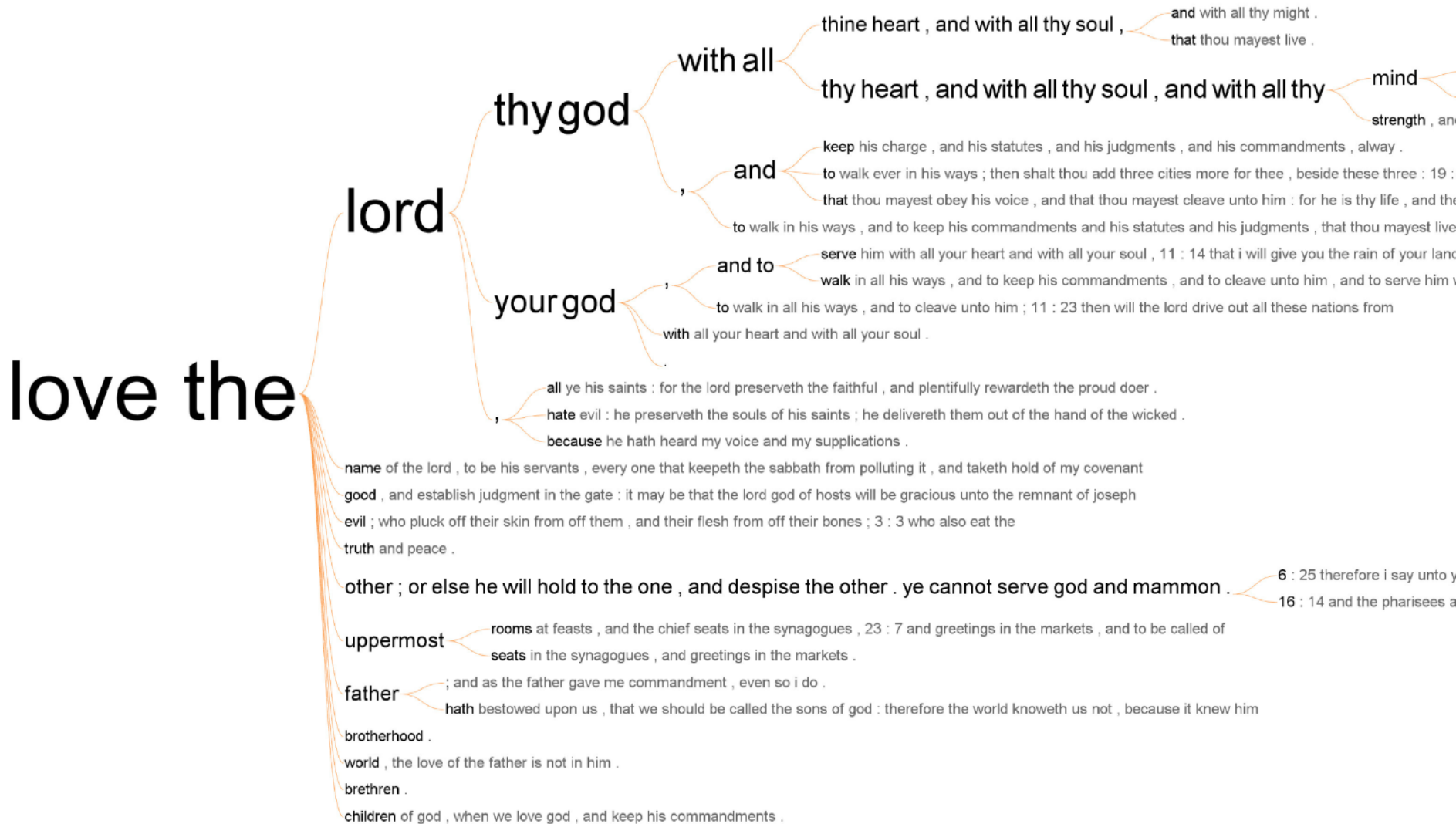
rough with you , be rough with love .

blind ,

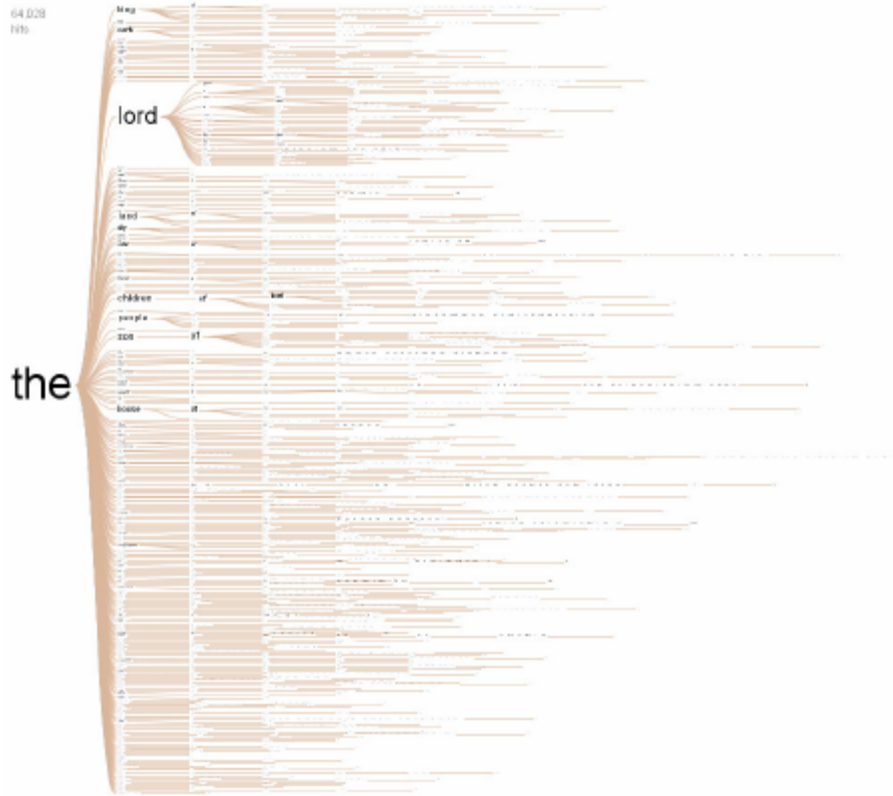
love cannot hit the mark .

it best agrees with night .

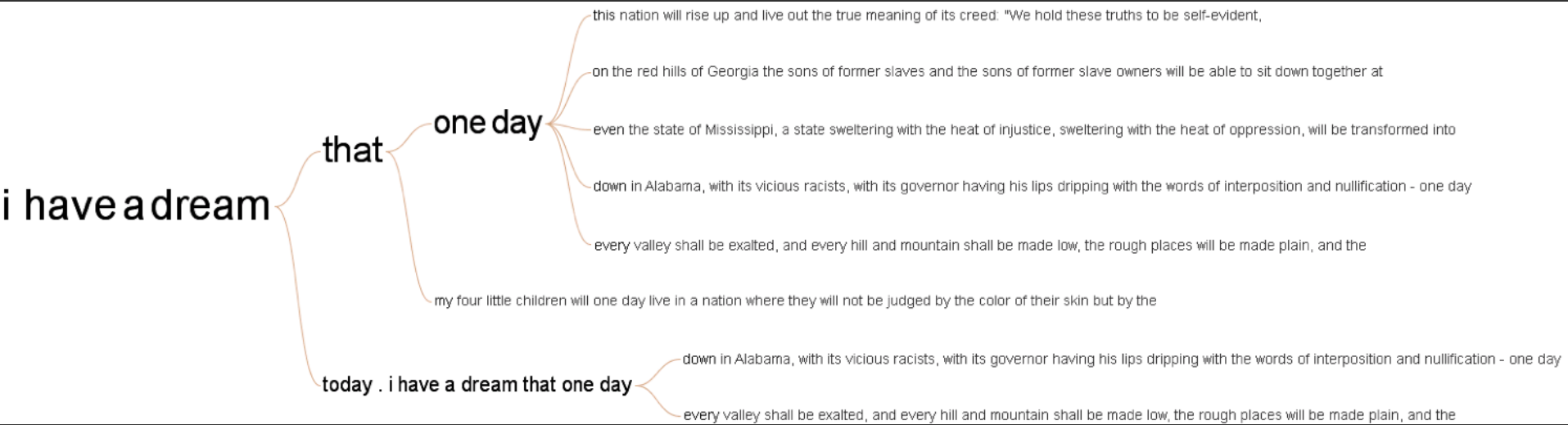
Word Tree [Wattenberg et al.]



Filter Infrequent Runs



Recurrent Themes in Speeches



Glimpses of Structure...

Concordances show local, repeated structure

But what about other types of patterns?

Lexical: <A> at

Syntactic: <Noun> <Verb> <Object>

Phrase Nets [van Ham et al.]

Look for specific **linking patterns** in the text:

'A and B', 'A at B', 'A of B', etc

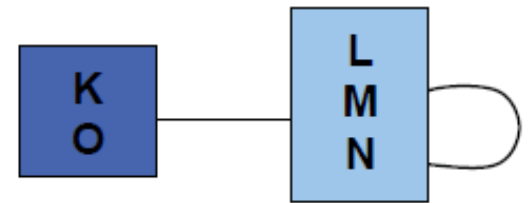
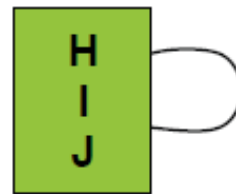
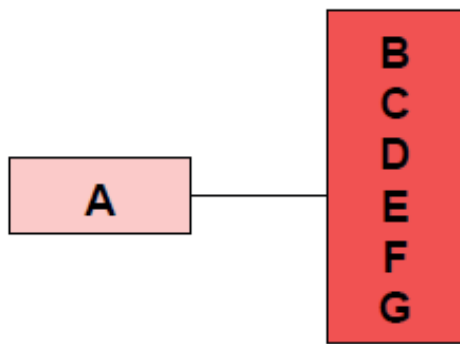
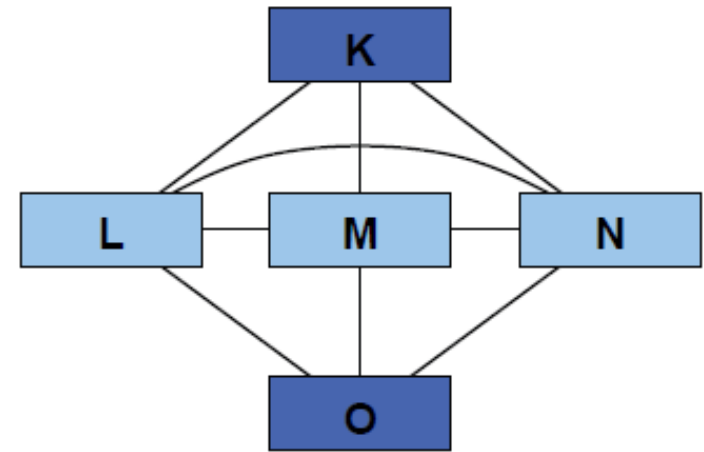
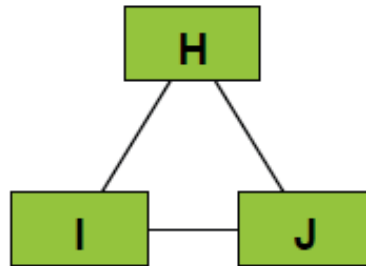
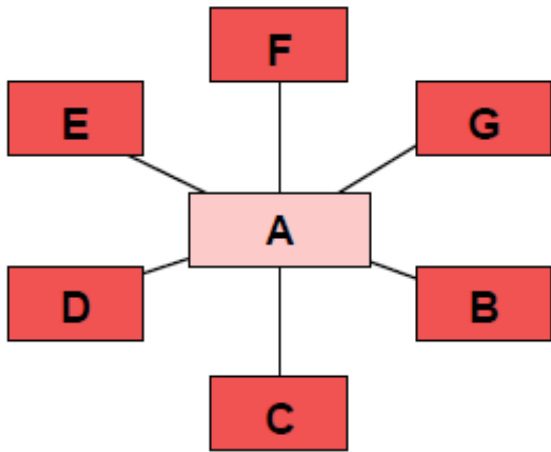
Could be output of regexp or parser.

Visualize patterns in a node-link view

Occurrences -> Node size

Pattern position -> Edge direction

Node Grouping

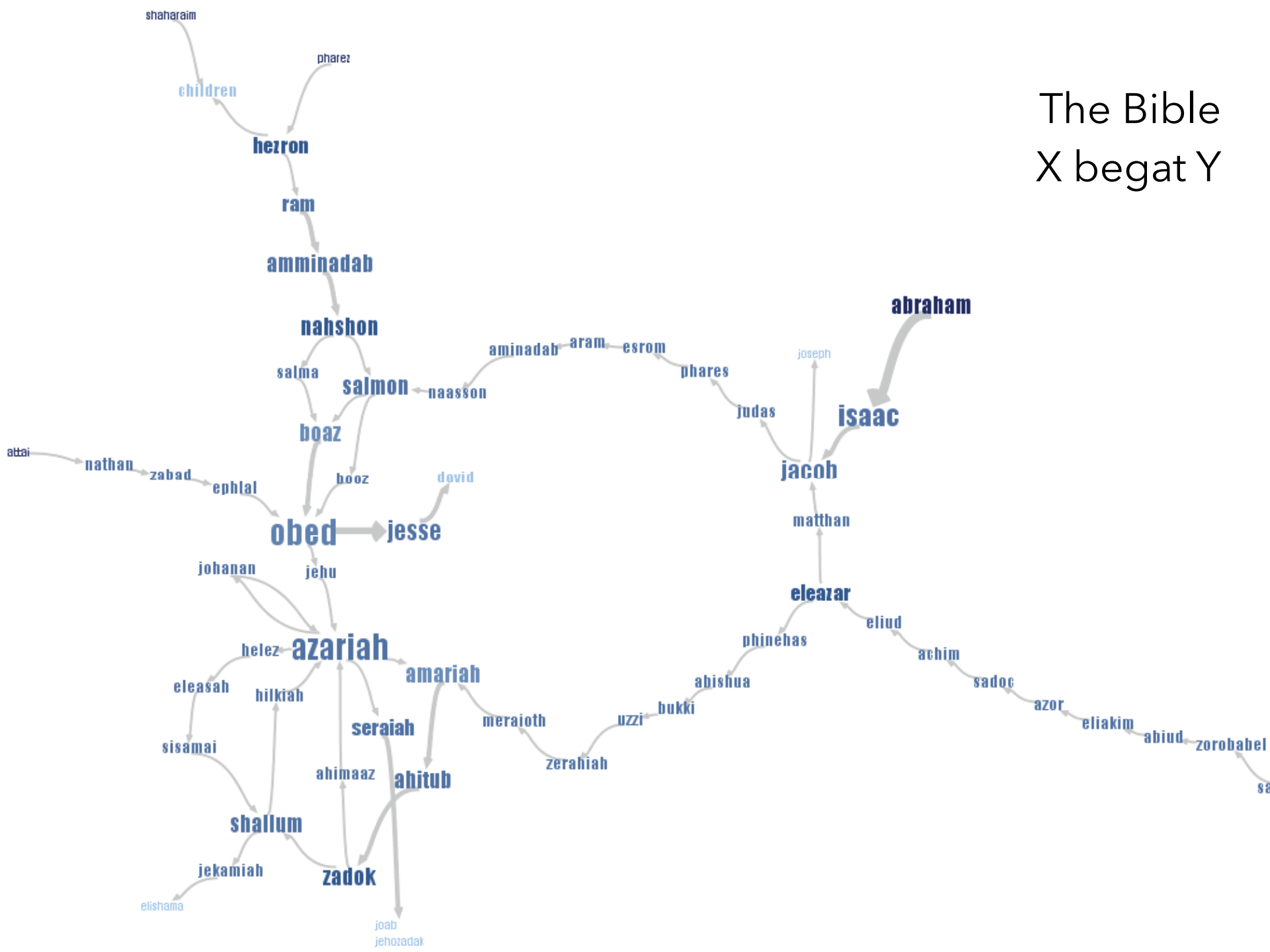


(a)

(b)

(c)

The Bible X begat Y



Document Content

Understand Your Analysis Task

Visually: Word position, browsing, brush & link

Semantically: Word sequence, hierarchy, clustering

Both: Spatial layout reflects semantic relationships

The Role of Interaction

Language model supports visual analysis cycles

Allow modifications to the model: custom patterns for expressing contextual or domain knowledge

Conversations

Visualizing Conversation

Many dimensions to consider:

Who (senders, receivers)

What (the content of communication)

When (temporal patterns)

Interesting cross-products:

What x When -> Topic "Zeitgeist"

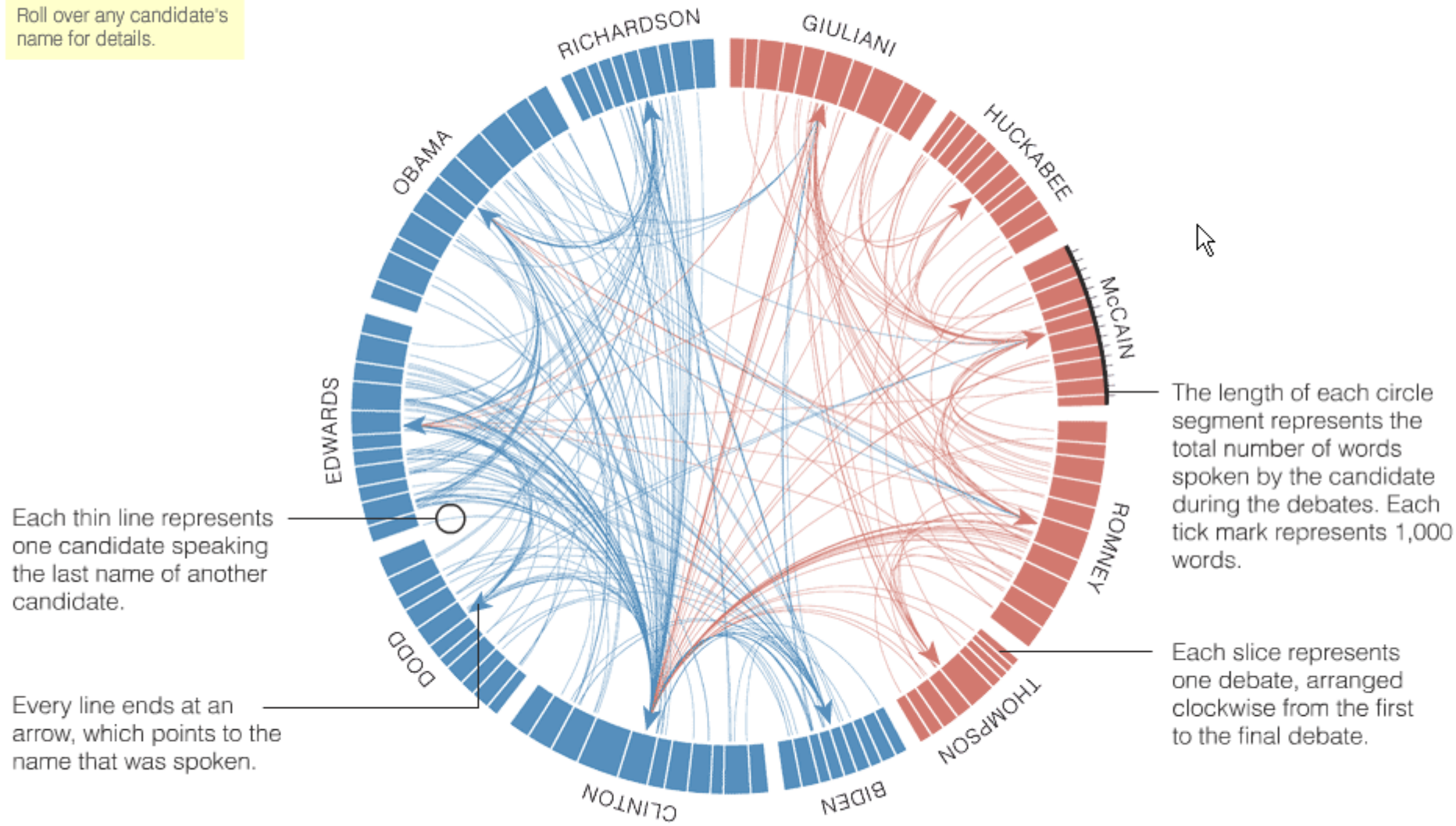
Who x Who -> Social network

Who x Who x What x When -> Information flow

Naming Names

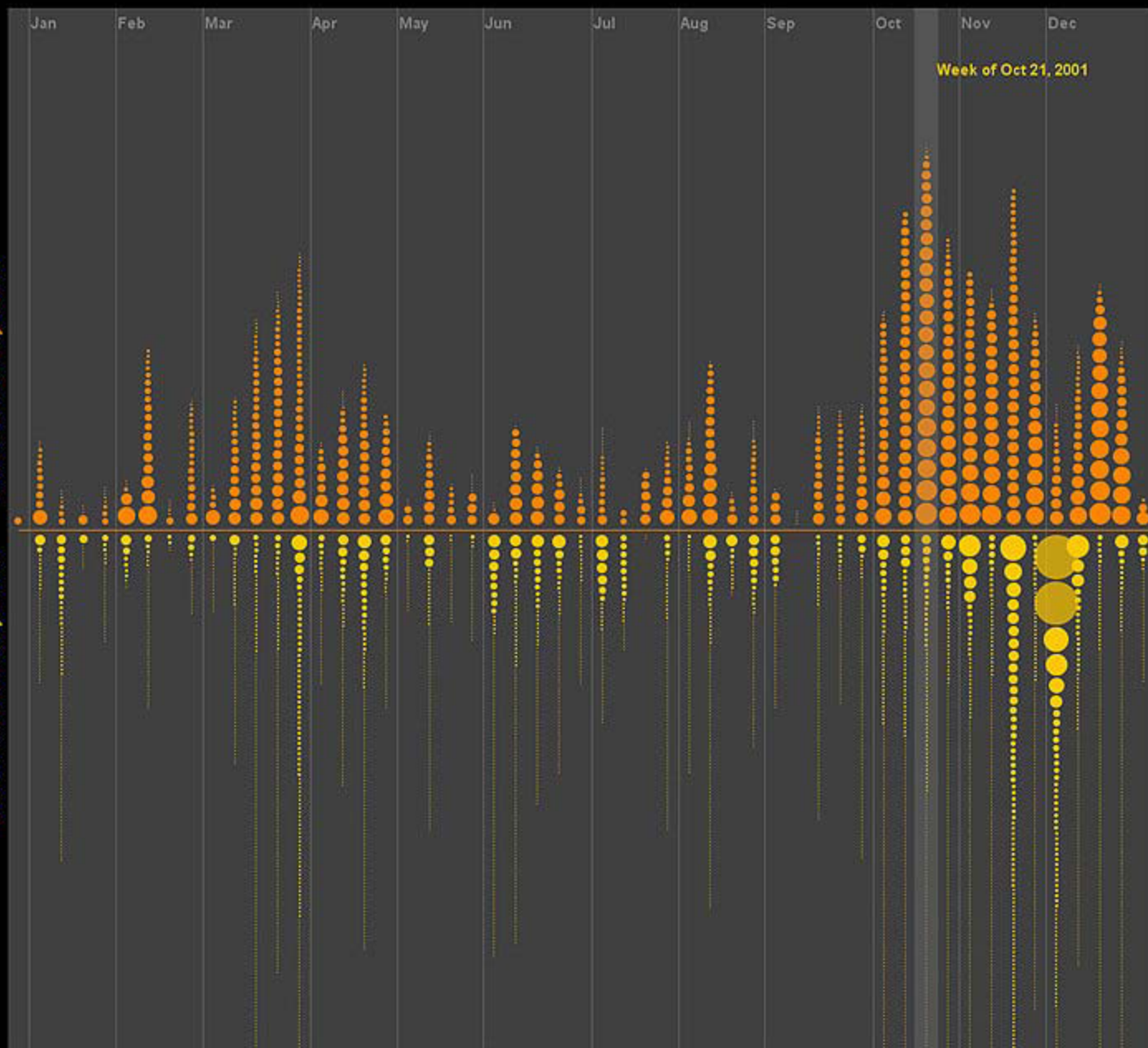
Names used by major presidential candidates in the series of Democratic and Republican debates leading up to the Iowa caucuses.

Roll over any candidate's name for details.



threads initiated by author

threads not initiated by author



subject	# of posts
Wednesday Spooker ASF	21
WET #3 Anyone for breakfast	20
Sunny Side Up ASF)	18
Saturday Ensemble and WET	18
Oh no! Watch out! ASF	18
Thursday Combo-Post WET #	16
The Yellow Rose Inn... A gift to	16
WET #1 JBP The First Time	16
We Love the Earth ASF	15
Monday Spooker "The Sight"	15
C'mon!!!!	14
Theberge "Le Vent Se Lève"	14
Holiday Tog #3)	13
Spooker du Jour)	13
Beginning ASF Short and ...	13
Second Try A Kalie for Suzy...	12
Come On a Safari With Me.....	11
Tuesday Spooker ASF	11
Curses, Foiled Again..... ASF ...	10
Halloween Togs Take Two)	9
Beauty of the Fury Jim Warren...	9
I thought I saw.....? ASF	7
Wednesday Evening at the Con...	4
Second Try A Kalie for Suzy...	2
Frank Was A Monster ASF...	1

subject	# of posts
Sunday Twofer ASF)	9
Chopsticks!A Jilly fake	8
Oh no! Trouble in Discworld!	7
WET... your thirst! ASF	6
A pretty for you... Reposted fro...	5
Saturday Spooker ASF	5
Sample Previous install Upgr...	4
Tennessee weather tonite	4
WET - Well I am not smiling!	4
Somethin' mushy <asf>	3
Getting seasonal with workin...	3
A Haunted House)	3
do you wonder what debi's be...	3
Question: Ethics of posters in	3
For Jerry	3
Olu's Tribe - slightly rated	3
WET - Glass Bottles	3
Peace Train<ASF>	2
Arrival at Stewart Island !!	2
WET 195 Wrap-up	2
Cat O'Lantern	2
I Put a Spell on You (Happy H...	2
Goodbye to Summer - A Timel...	2
Two Pumpkins In A Strange B...	2
Still Heading South !!	2
WET- Frank Sinatra - The Man...	2
WET Autumn	2
Purple Martin ASF	2
Opposites Attract ...	2
Time	2

author: rubat@pam.org

[back to message](#)

Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

Week of May 6, 2001

In Genes Scientifically Correct?

subject # of posts

threads initiated by author

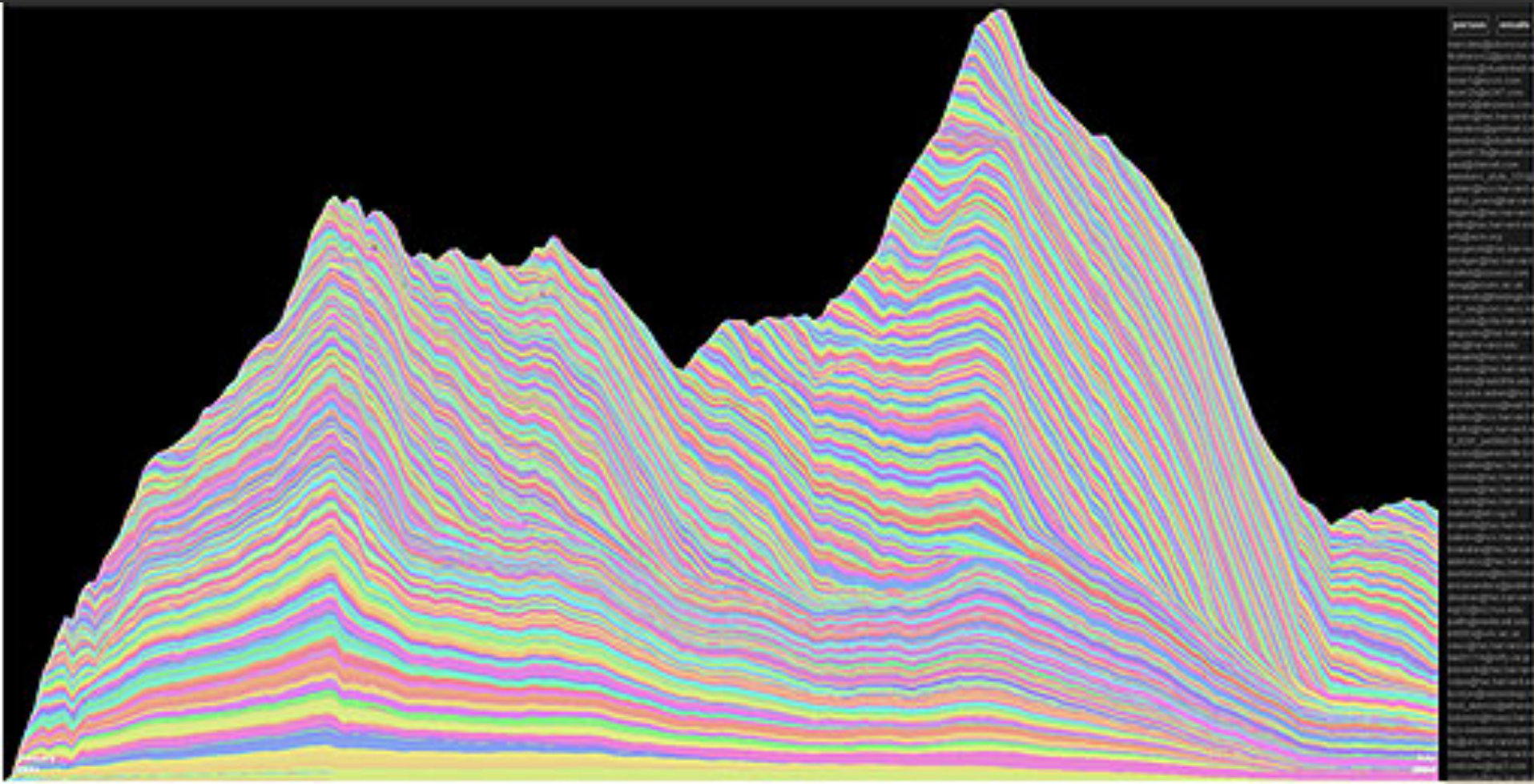
threads not initiated by author



subject	# of posts
Anti-Humanity is	21
Et Tu, Henry, Scientist	16
OLD TESTAMENT	17
Please define Ho	10
Evolution fails	14
Why wouldn't we?	10
Scientist against a	15
How to teach Ho?	14
TDM vs. CDMA	10
Dot and O-2	7
Genes is wrong	7
The Abroad is Out	7
President Bush is	7
I got a question	6
A Faithful Deed	6
Freedom from Ho	5
Original Sin - Bad F	5
Evolution is not C	5
SCIENTIFIC PRO	5
Christians teach a	4
Anti-Humanity is	4
All post's remain	4
Car Theft	4
1000Ho vs. 1000	4
Voicemail 1012	4

Email Mountain [Viegas]

Conversation by person over time (who x when).



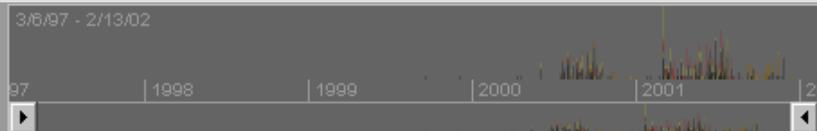


Enron E-Mail Corpus

[Heer]

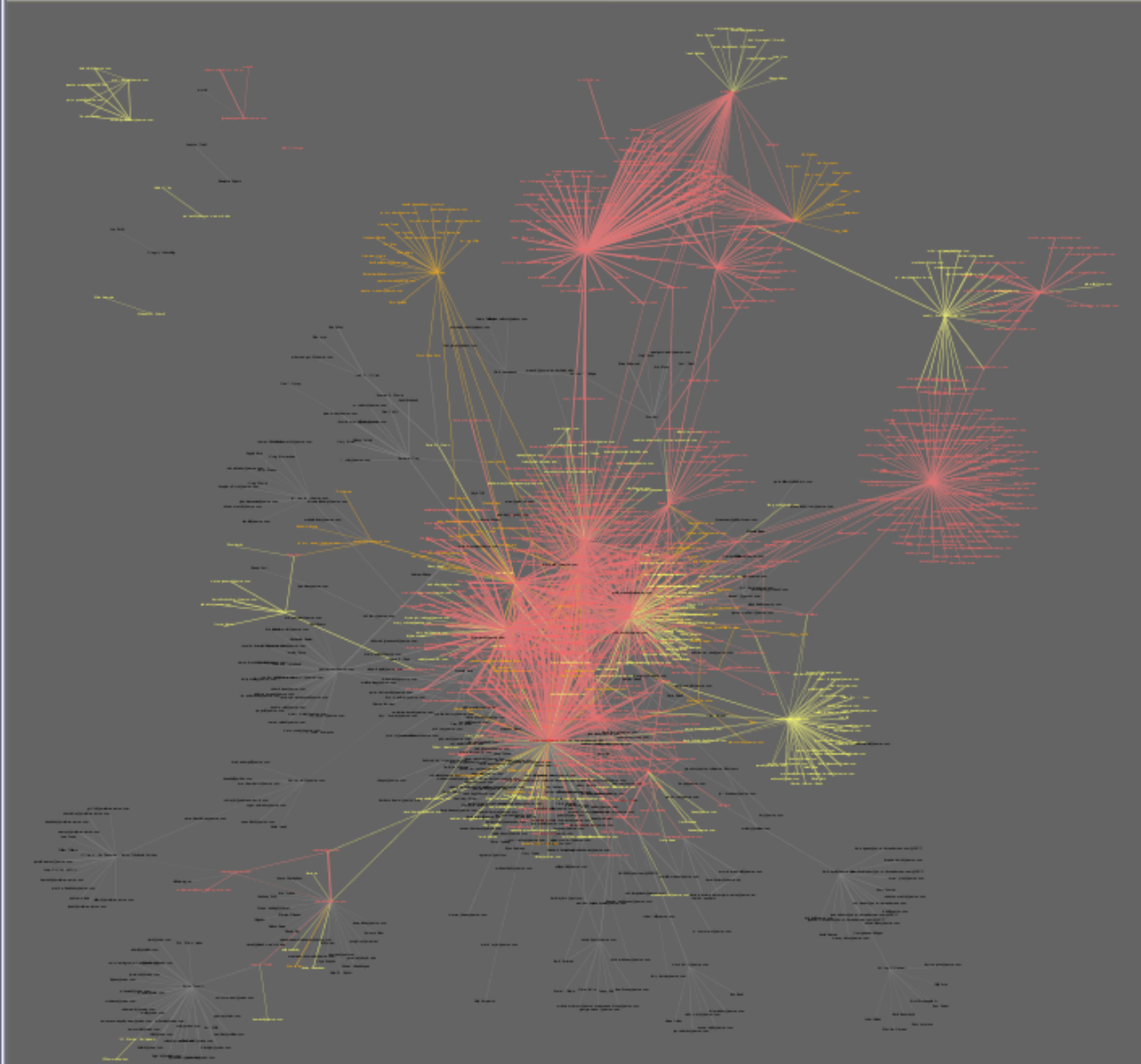
connectivity >>

community >>



search >>

search >>



steven.kean@enron.com

- 2000-09-01 04:25:00.0 Linda Jenkins on "Jerry's Show" Mond
- 2000-09-02 10:14:00.0 Re: The Governors' Natural Gas Summ
- 2000-09-08 10:03:00.0
- 2000-09-10 14:07:00.0 CPUC Hearing in SD on 9/8
- 2000-09-10 16:20:00.0 Re: Fletcher School/Enron
- 2000-09-13 00:57:00.0 Re: Contact

0 2 0 0 0 0 1 0 0 0 0 0 0 0

ID: 174285

Subject:

From: <steven.kean@enron.com>

Date: 2000-09-08 10:03:00.0

To: <kmagrude@enron.com>

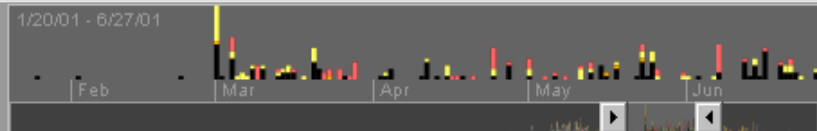
Cc: Richard Shapiro <richard.shapiro@enron.com>

Got your message. I'm testifying at the Congressional hearing and Dasovich is covering FERC. I think Jeff's comments were taken out of context. He said policymakers do need to take care of small customers whose bills are tripling. Frankly, we'd get slaughtered if we said anything else. But he also said there is a right way and a wrong way to do it. Enron and others had provided a market based answer by offering a fixed price deal to SDG&E (which would have enabled them to cap rates to those who had not switched. California elected instead to cap rates and deficit spend (ie create a deferral account). I don't think we can stand for anything that doesn't protect the small customers, but we can continue to emphasize the market based solutions. One of the messages in my testimony will be: customers should be encouraged to choose. Those who did are doing fine.

Messages

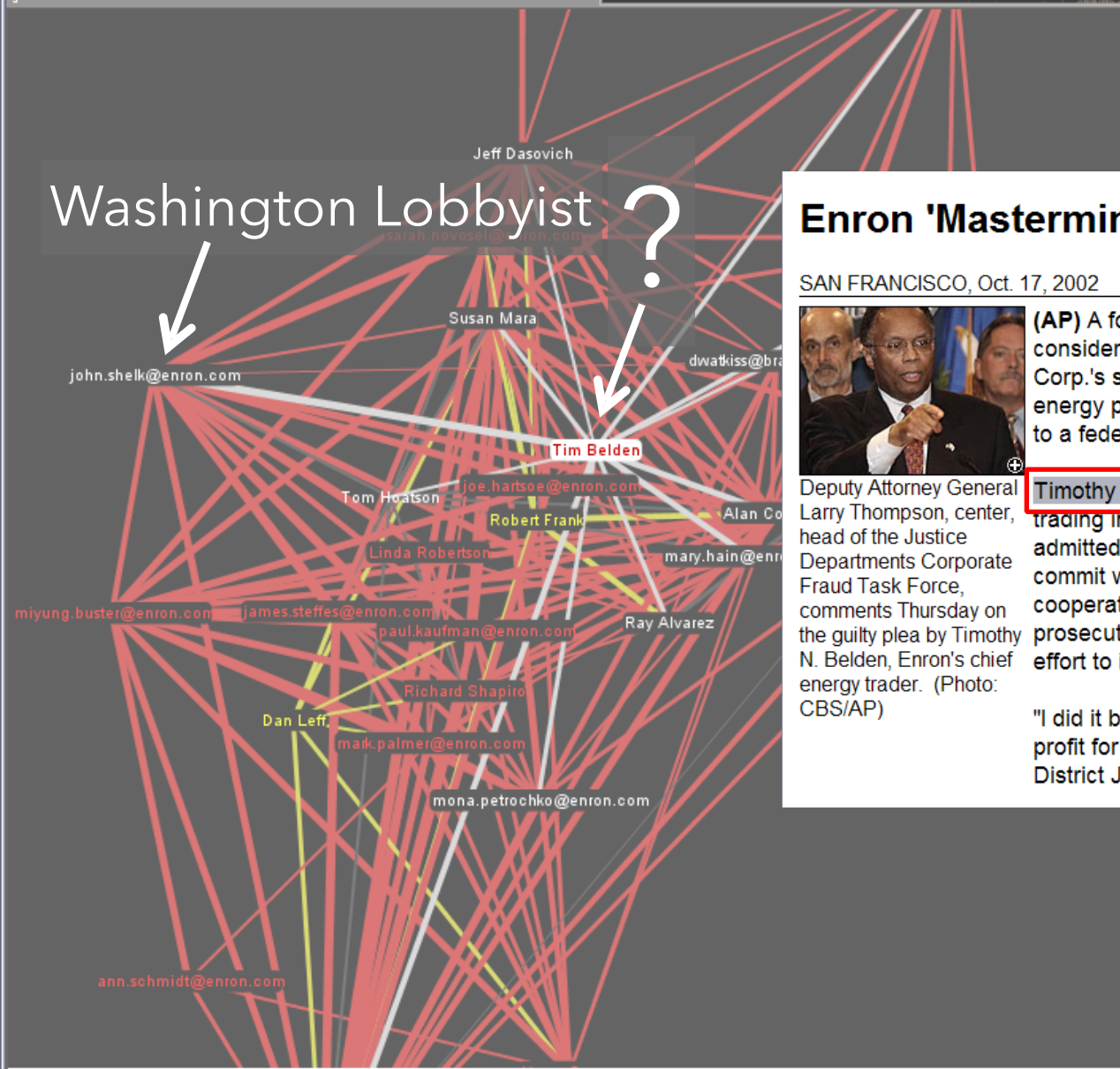
connectivity >>

community >>



search >>

search >>



Washington Lobbyist ?

- Tim Belden**
- 2001-06-06 16:48:00.0 ISO's Response to BPA Rebuttal of Sh
 - 2001-06-07 11:00:00.0 Legislative Update -- Two Track In The
 - 2001-06-18 00:15:00.0 White House To Support FERC Action
 - 2001-06-19 04:22:00.0 NEWS FLASH ON THIS MORNING'S
 - 2001-06-20 10:37:00.0 Today's Senate Hearing
 - 2001-06-21 02:15:00.0 More on FERC Refunds

Enron 'Mastermind' Pleads Guilty

SAN FRANCISCO, Oct. 17, 2002



Deputy Attorney General Larry Thompson, center, head of the Justice Departments Corporate Fraud Task Force, comments Thursday on the guilty plea by Timothy N. Belden, Enron's chief energy trader. (Photo: CBS/AP)

(AP) A former top energy trader, considered the mastermind of Enron Corp.'s scheme to drive up California's energy prices, pleaded guilty Thursday to a federal conspiracy charge.

Timothy Belden, the former head of trading in Enron's Portland, Ore., office, admitted to one count of conspiracy to commit wire fraud and promised to cooperate with state and federal prosecutors as well as any non-criminal effort to investigate the energy industry.

"I did it because I was trying to maximize profit for Enron," Belden told U.S. District Judge Martin Jenkins.

from four western governors -- those from Arizona, North Dakota, Utah and Wyoming -- saying that since FERC has acted, there is no need for Congress to pursue price control legislation.

There were a series of questions and comments on details and technical aspects of the orders. I will do an e-mail on these items later today. Please advise if you have any questions or comments.

Messages

Document Collections

Named Entity Recognition

Label named entities in text:

John Smith -> PERSON

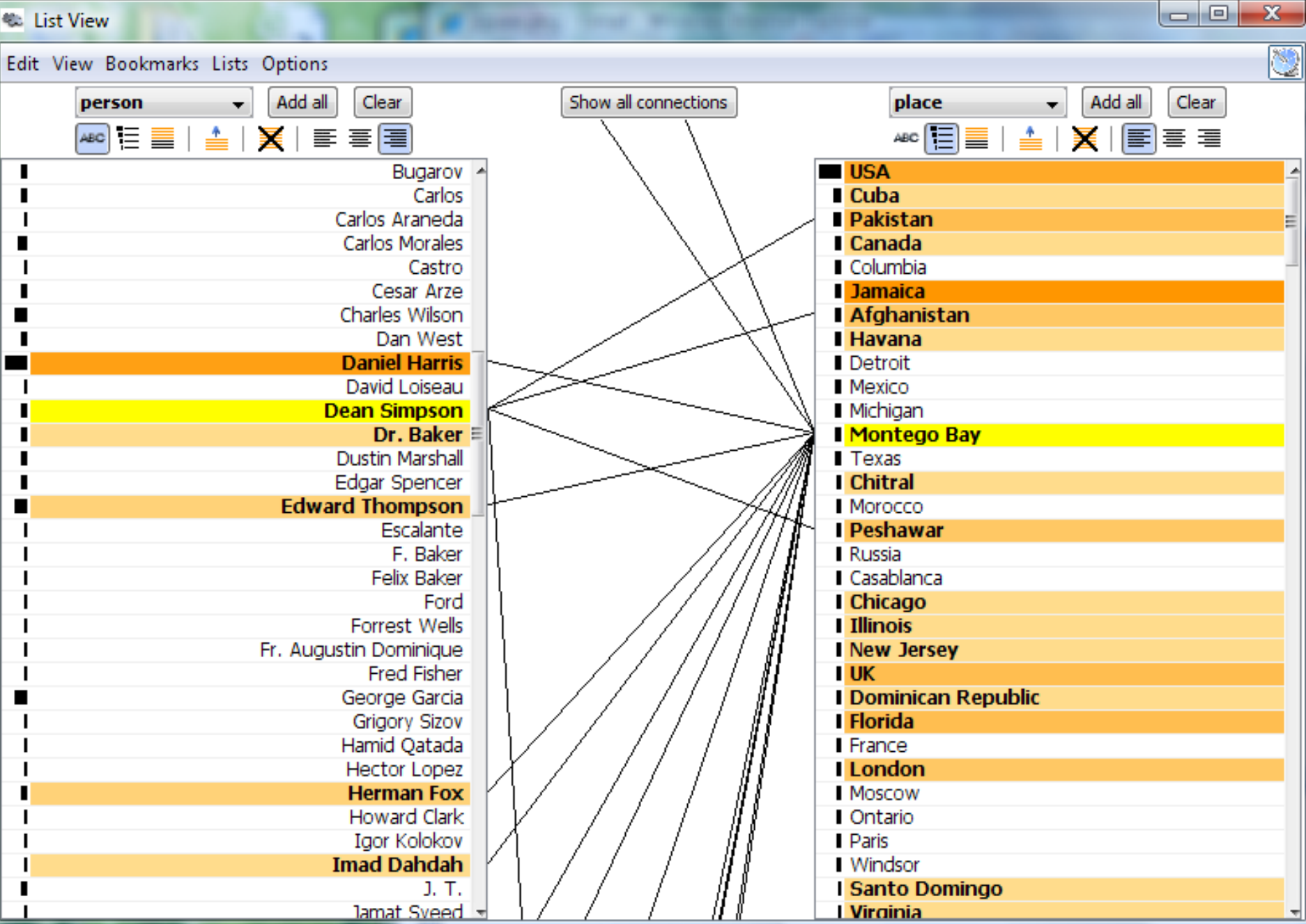
Soviet Union -> COUNTRY

353 Serra St -> ADDRESS

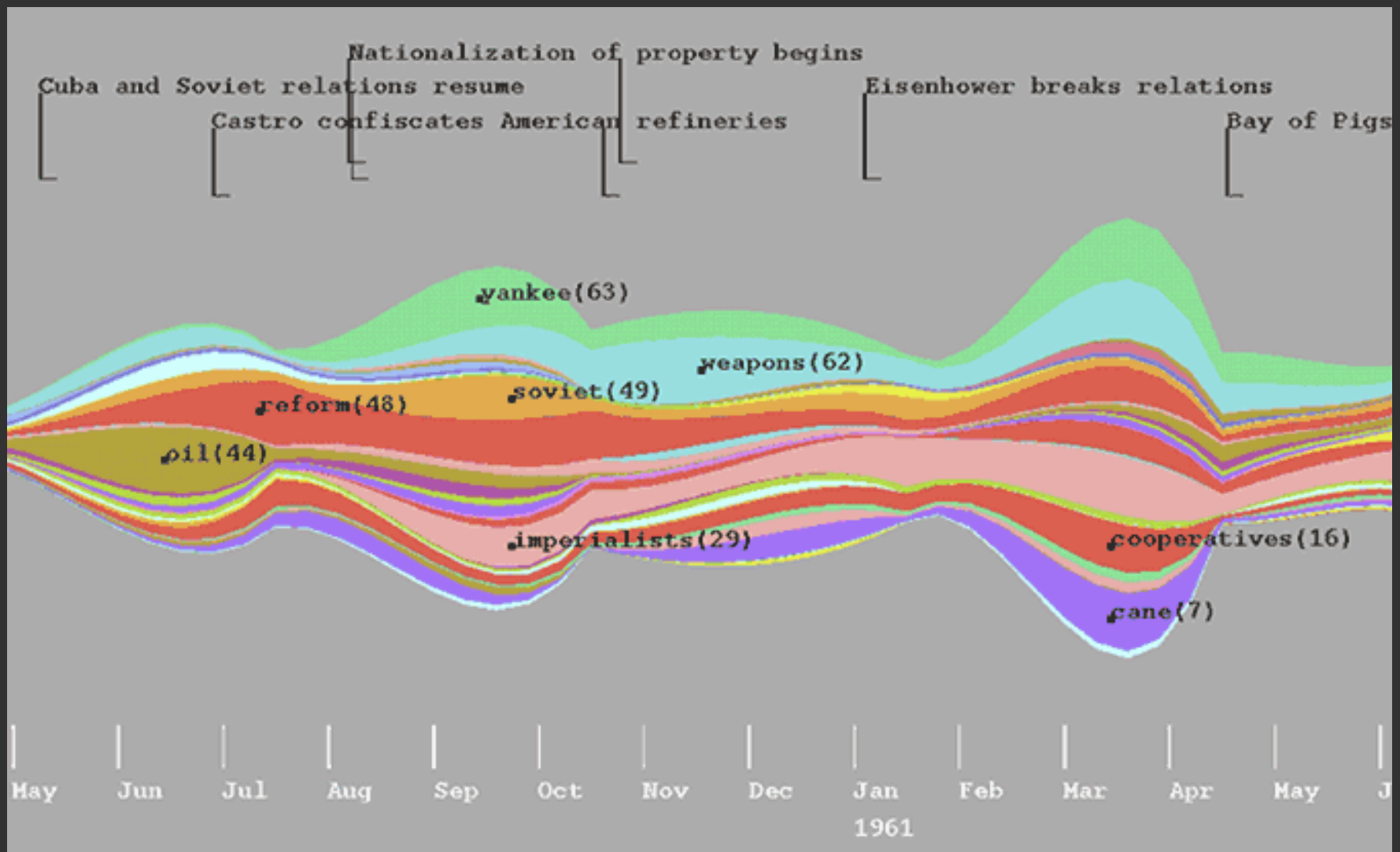
(555) 721-4312 -> PHONE NUMBER

Entity relations: how do the entities relate?

Simple approach: do they co-occur in a small window of text?



Theme River [Havre et al.]



Similarity & Clustering

Compute vector distance among docs

For TF.IDF, typically cosine distance

Similarity measure can be used to cluster

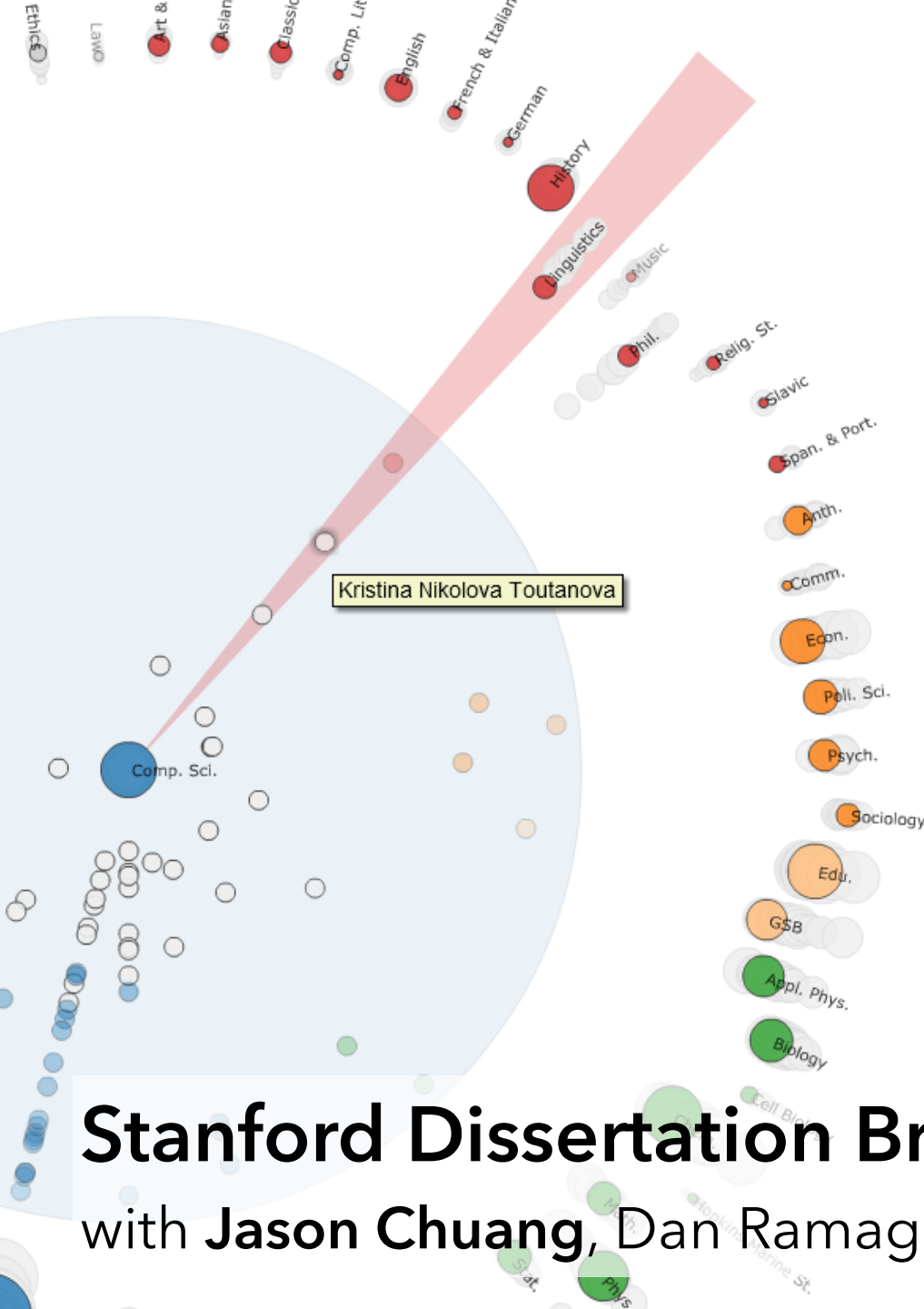
Topic modeling

Assume documents are a mixture of topics

Topics are (roughly) a set of co-occurring terms

Latent Semantic Analysis (LSA): reduce term matrix

Latent Dirichlet Allocation (LDA): statistical model



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

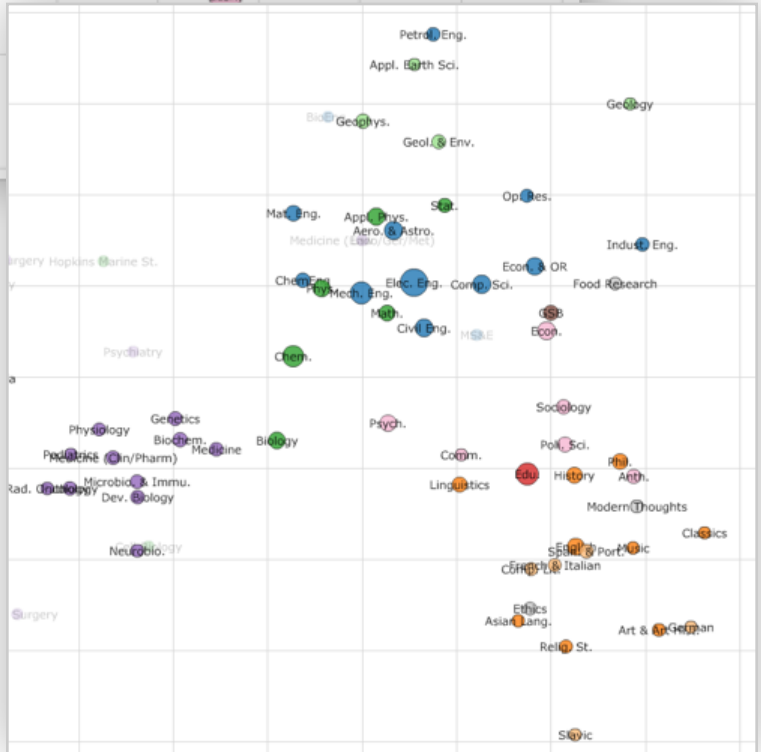
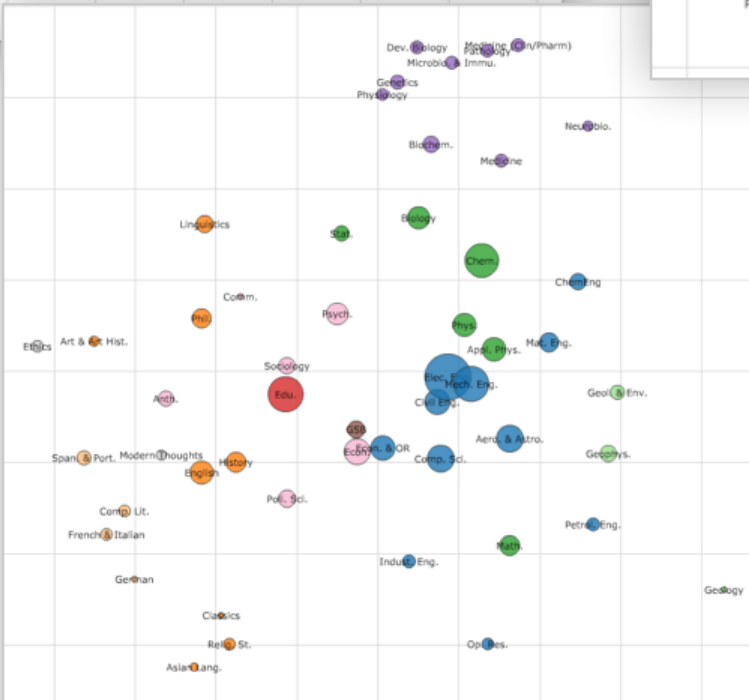
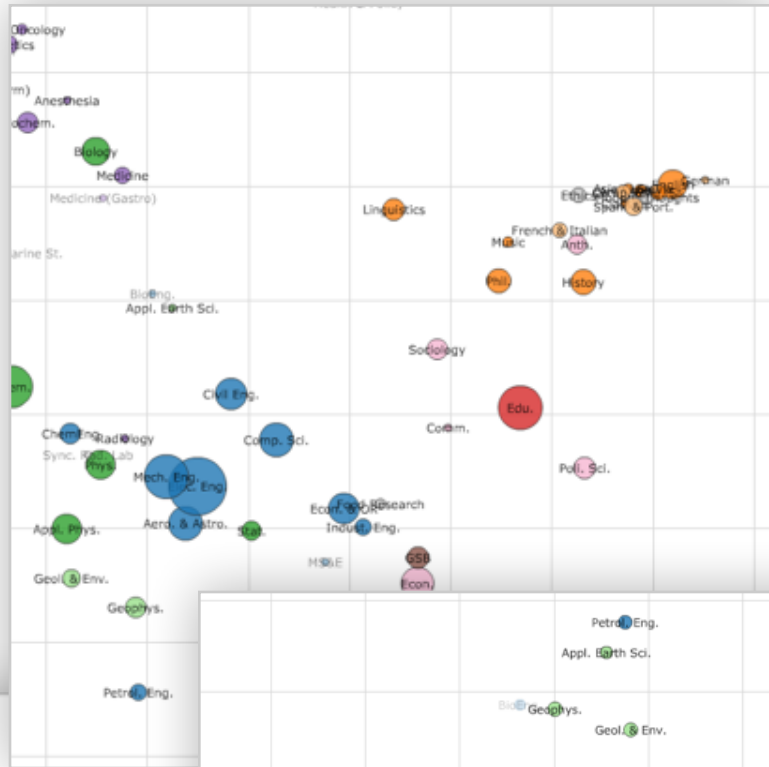
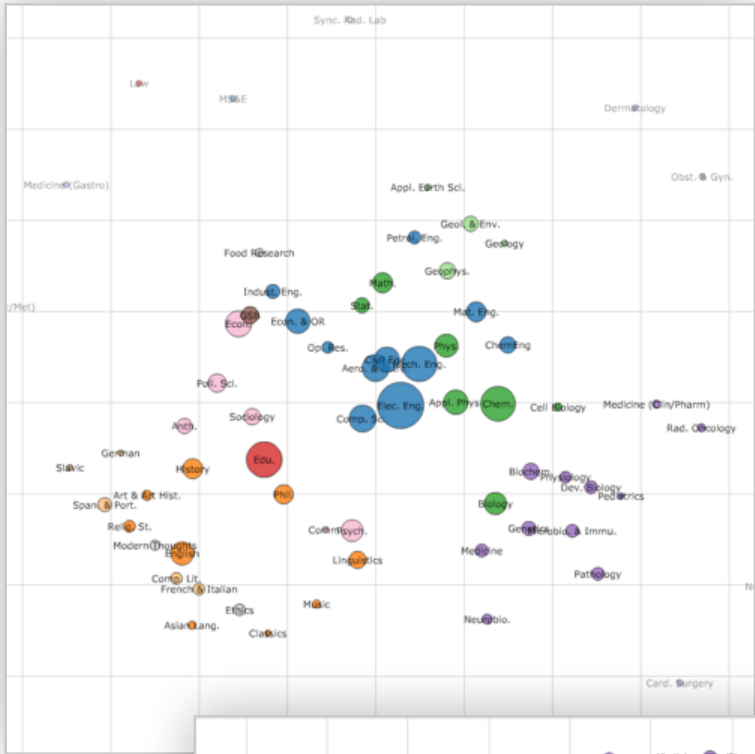
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.

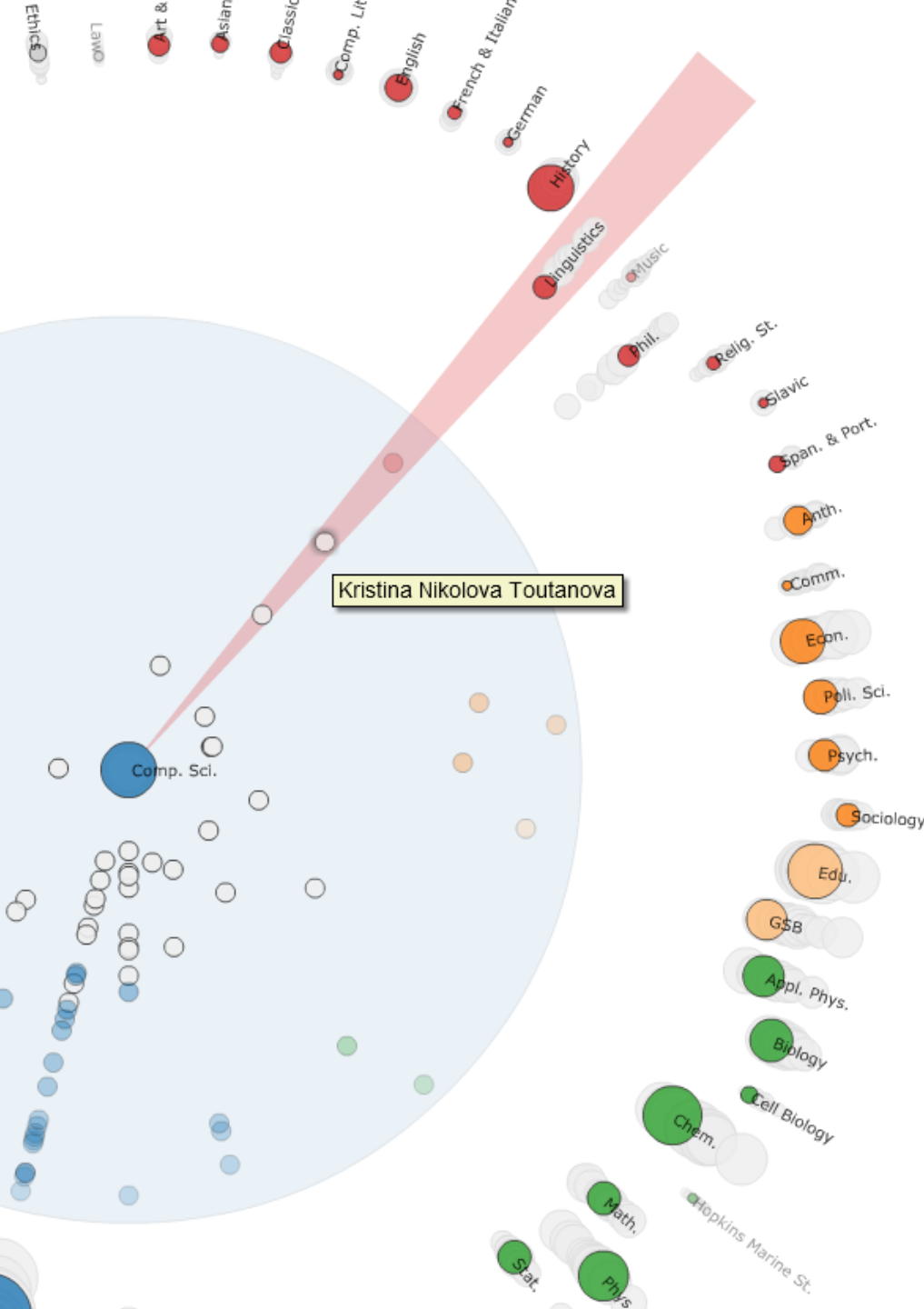
Stanford Dissertation Browser

with Jason Chuang, Dan Ramage & Christopher Manning





Oh, the humanities!



Effective statistical models for syntactic and semantic disambiguation

Student: Kristina Nikolova Toutanova

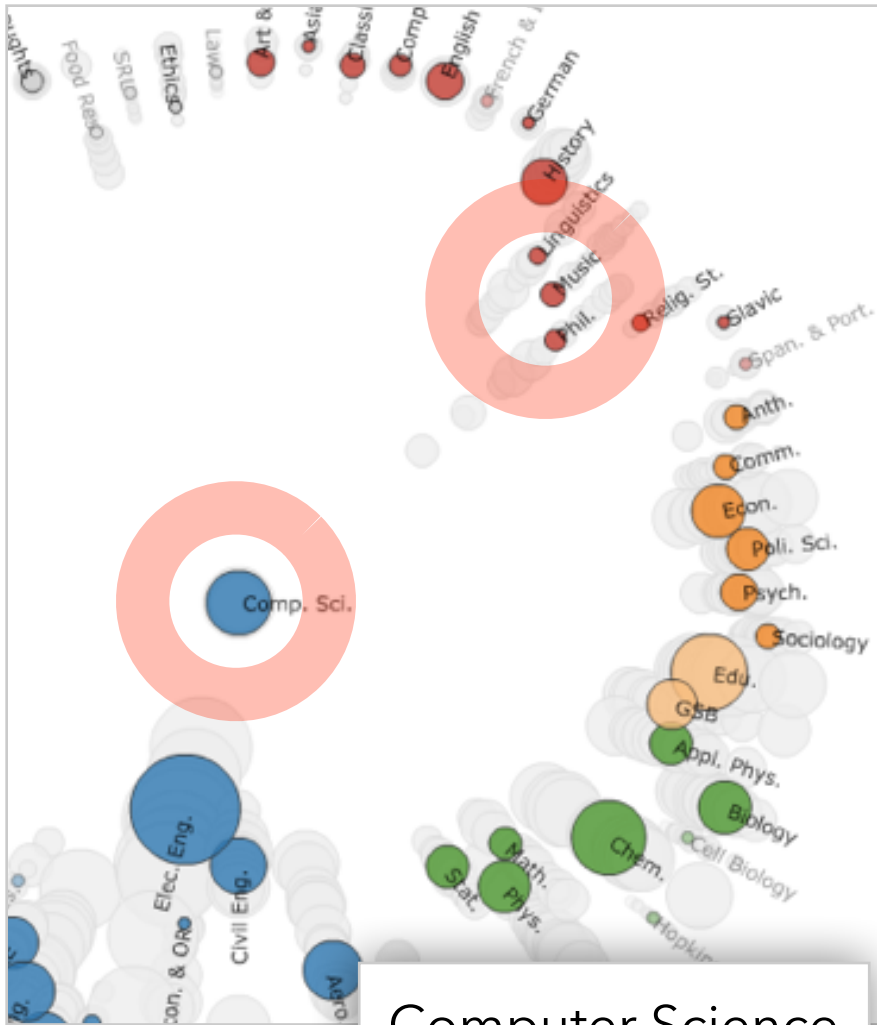
Advisor: Christopher D. Manning

Computer Science (2005)

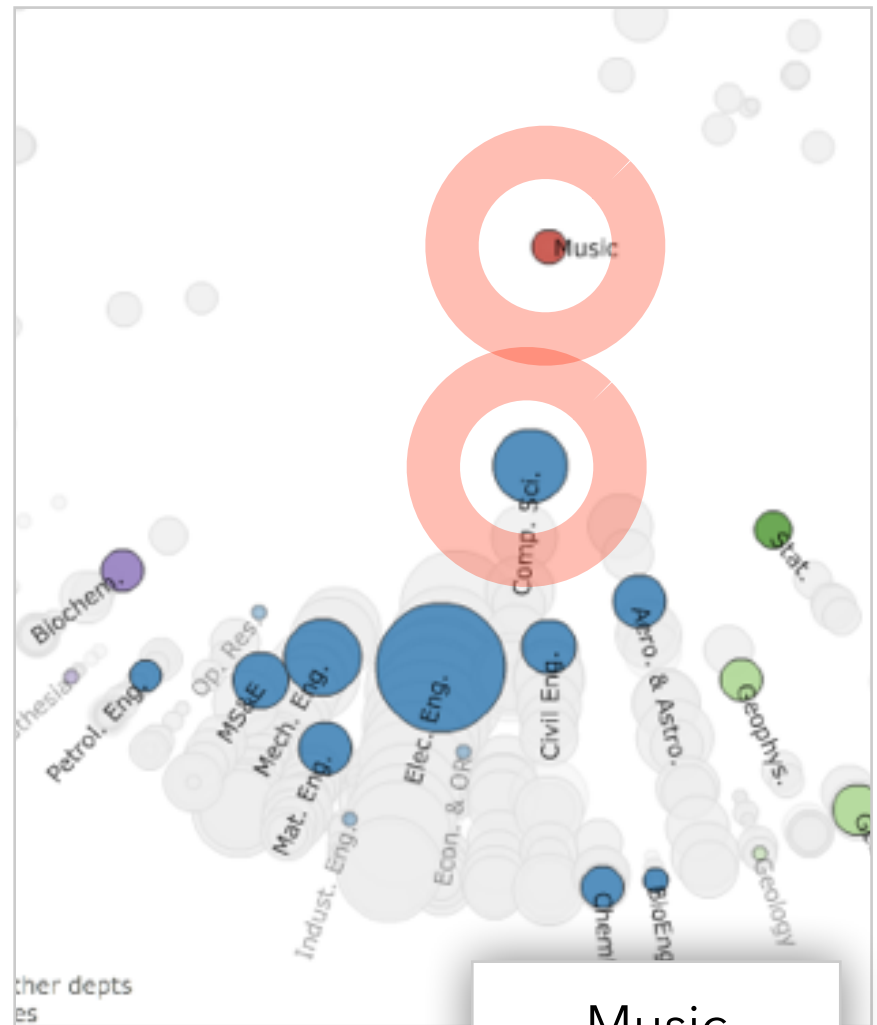
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing machine learning algorithms that deal with the specific properties of NL disambiguation tasks--sparsity of training data and large, structured spaces of hidden labels. For the task of syntactic disambiguation, we propose a novel representation of parse trees that connects the words of the sentence with the hidden syntactic structure in a direct way. Experimental evaluation on parse selection for a Head Driven Phrase Structure Grammar shows the new representation achieves superior performance compared to previous models. For the task of disambiguating the semantic role structure of verbs, we build a more accurate model, which captures the knowledge that the semantic frame of a verb is a joint structure with strong dependencies between arguments. We achieve this using a Conditional Random Field without Markov independence assumptions on the sequence of semantic role labels. To address the sparsity problem in machine learning for NL, we develop a method for incorporating many additional sources of information, using Markov chains in the space of words. The Markov chain framework makes it possible to combine multiple knowledge sources, to learn how much to trust each of them, and to chain inferences together. It achieves large gains in the task of disambiguating prepositional phrase attachments.



Computer Science



Music

“Word Borrowing” via Labeled LDA

Summary

High Dimensionality

Where possible use text to represent text...
... which terms are the most descriptive?

Context & Semantics

Provide relevant context to aid understanding.
Show (or provide access to) the source text.

Modeling Abstraction

Understand abstraction of your language models.
Match analysis task with appropriate tools and models.

Currently: from bag-of-words to *vector space embeddings*