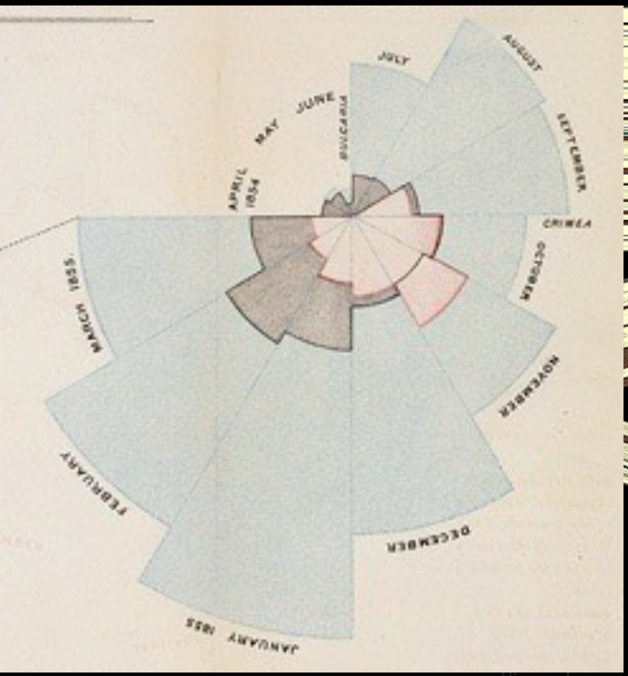
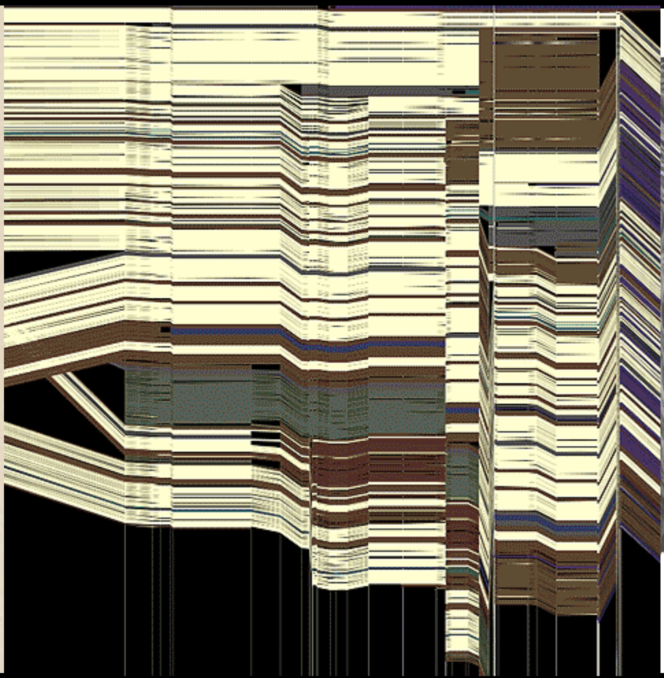
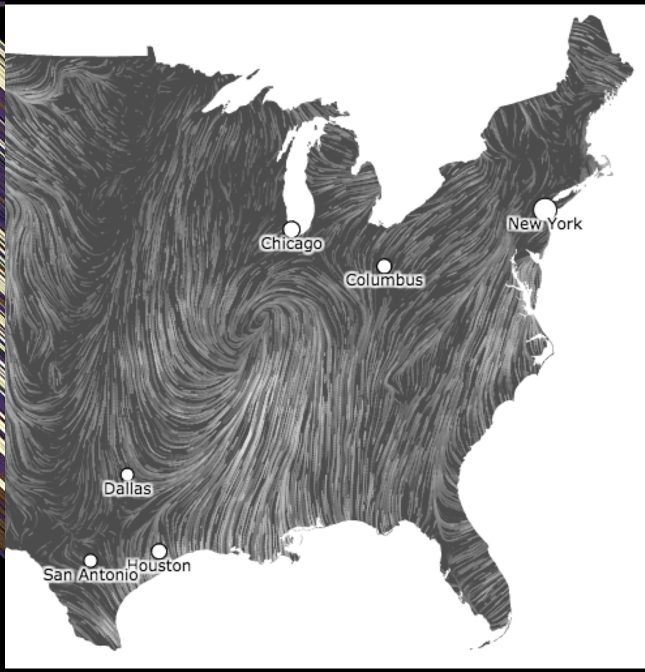
CSE 512 - Data Visualization

Text Visualization

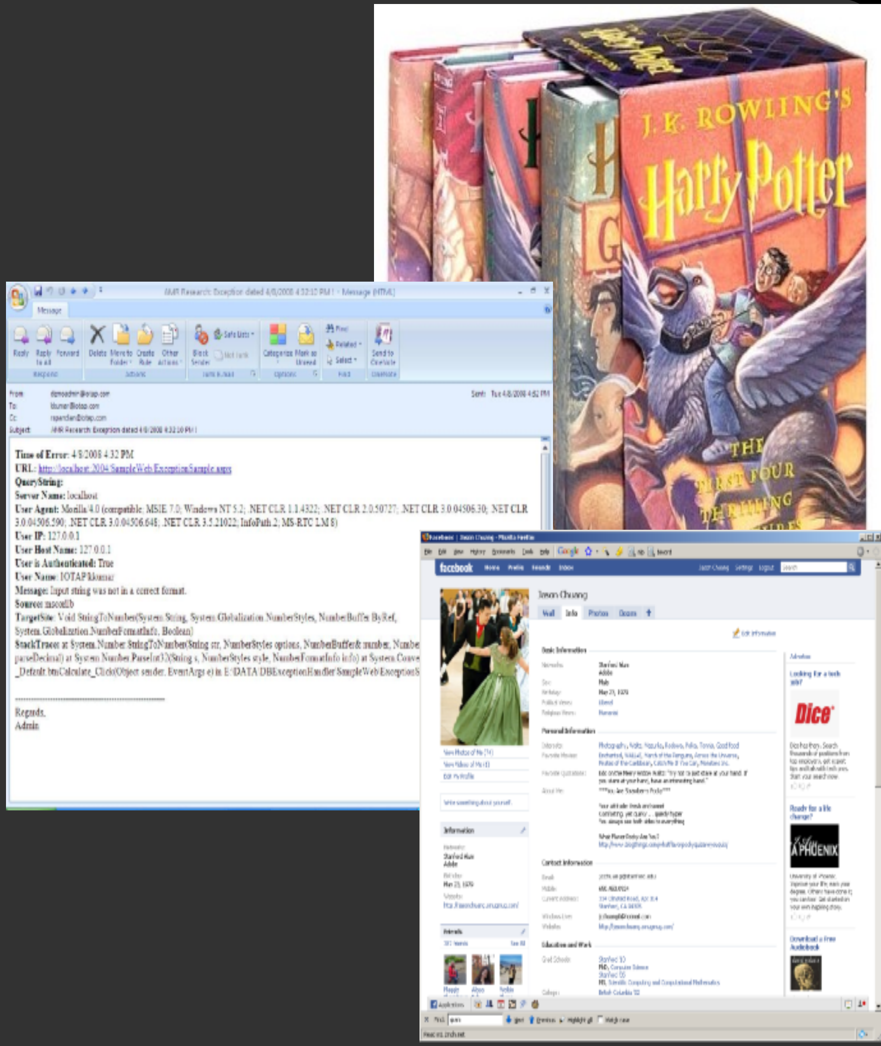






Jeffrey Heer University of Washington

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)

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

Example:
Health Care Reform

Example: Health Care Reform

Background

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 wer 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.

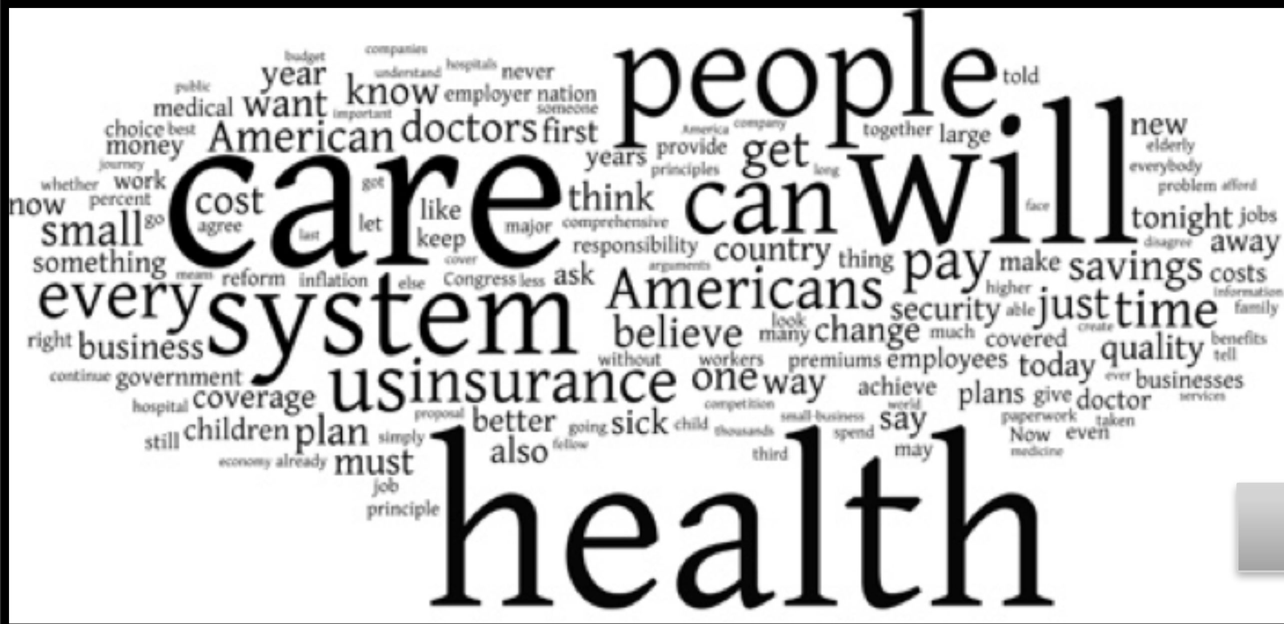
I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for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

Tag Clouds: Word Count

President Obama's Health Care Speech to Congress [NY Times]

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

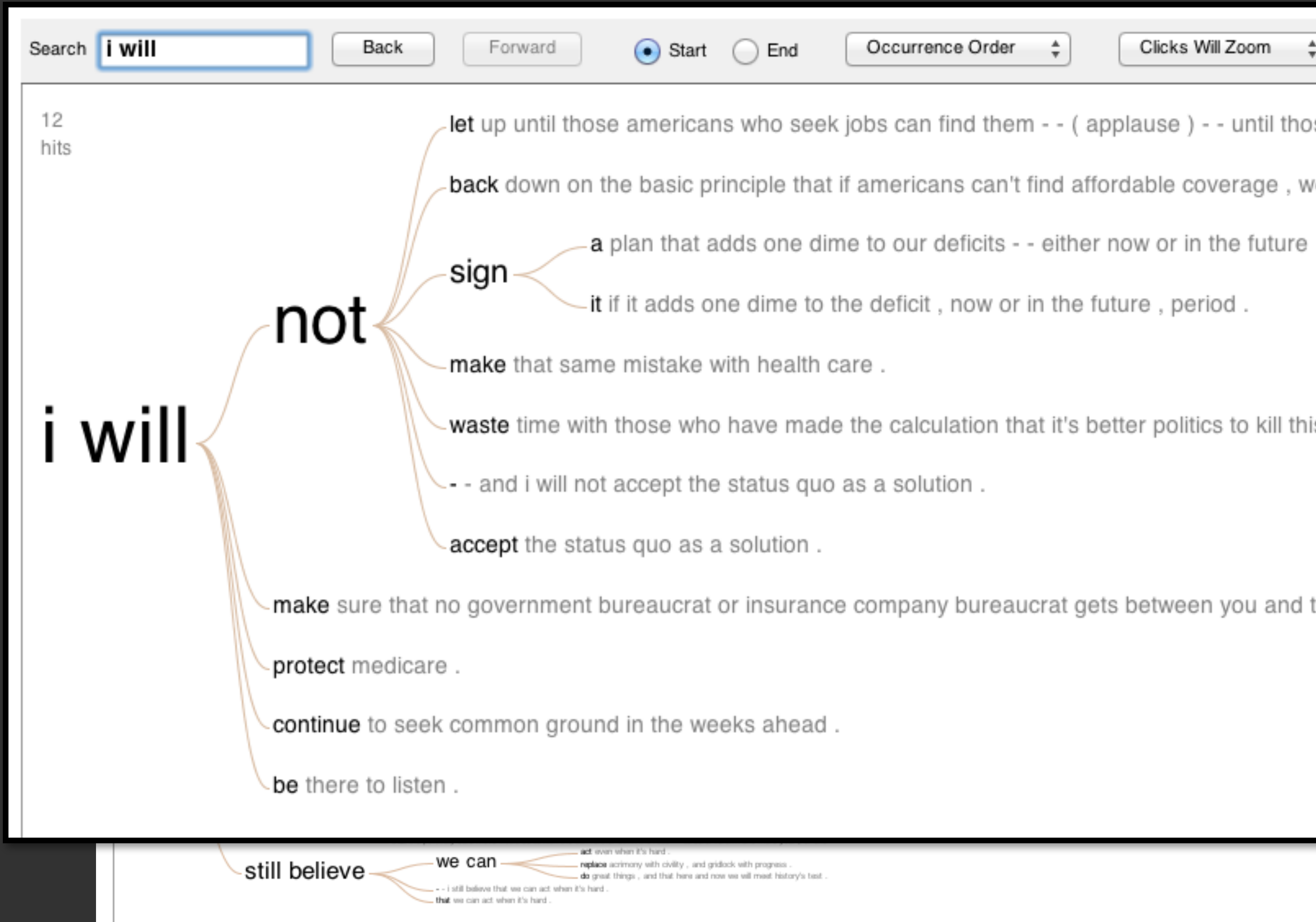


Bill Clinton 1993

Bill Clinton 1993 … Barack Obama 2009

economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

Word Tree: Word Sequences


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 tho
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 acrimony 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 ≈ 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

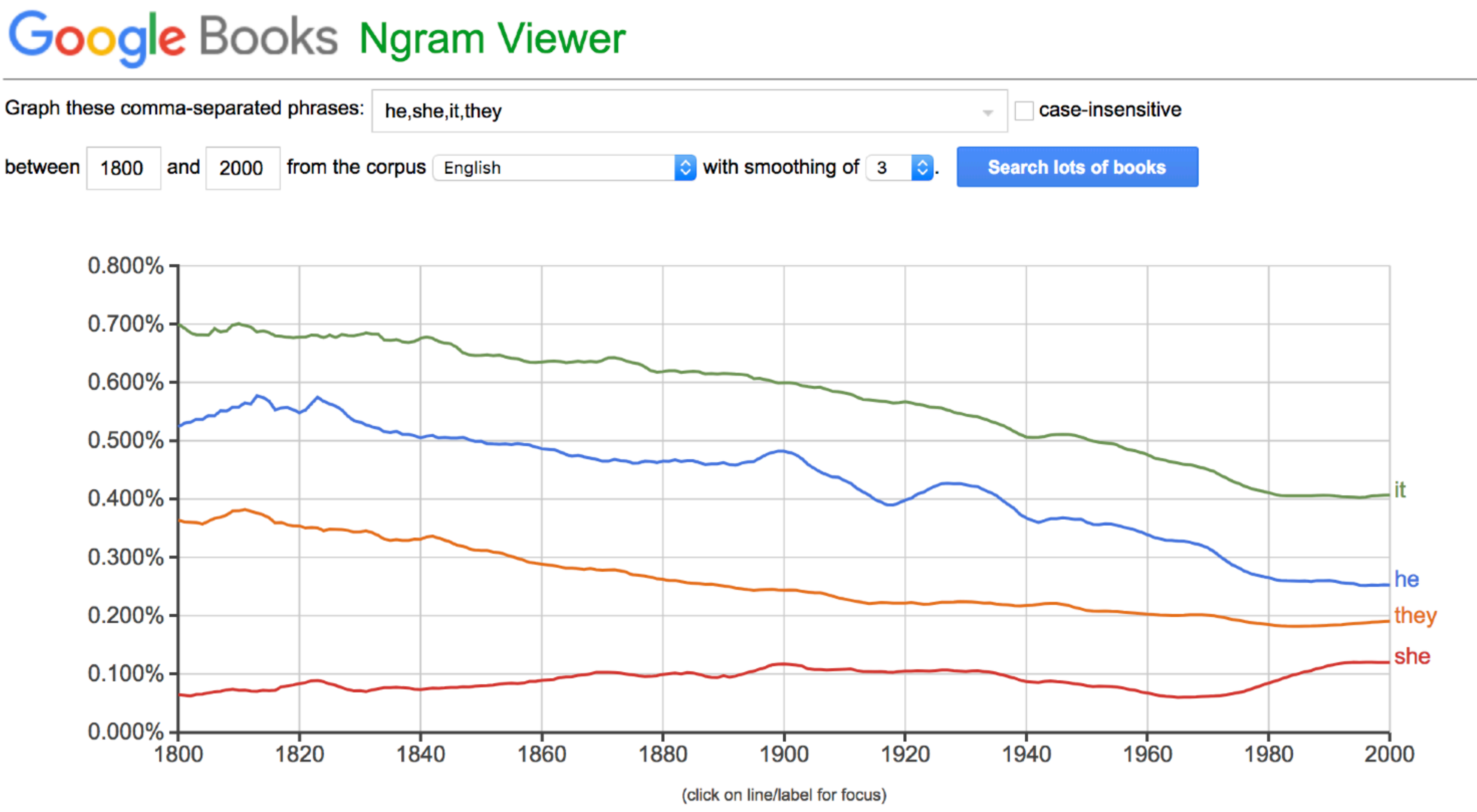
Simplest weighting is to just count occurrences

	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

WordCounts (Harris '04)

http://wordcount.org

https://books.google.com/ngrams/



https://books.google.com/ngrams/

Visualizations : Wordle of Sarah Palin RNC 9/3/2008 Speech

Creator: Anonymous
Tags:

Edit Language Font Layout Color

Data file: Sarah Palin speaks at the Republican National Convention, 9/3/2008 Data source: SFGate / AP This data set has not yet been rated

Tag Clouds

Strengths

Can help with gisting and initial query formation.

Weaknesses

Sub-optimal visual encoding (size vs. position)

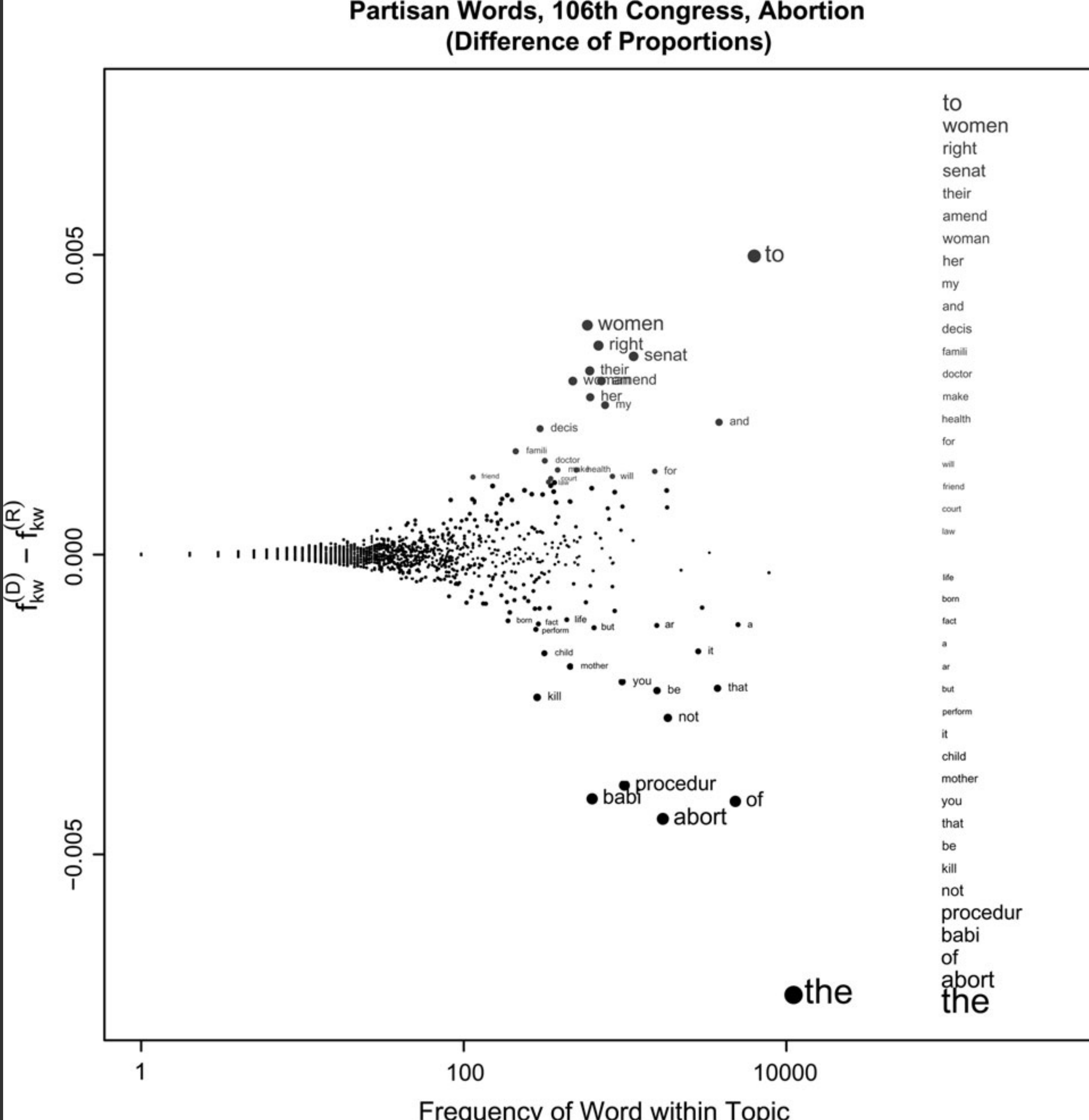
Inaccurate size encoding (long words are bigger)

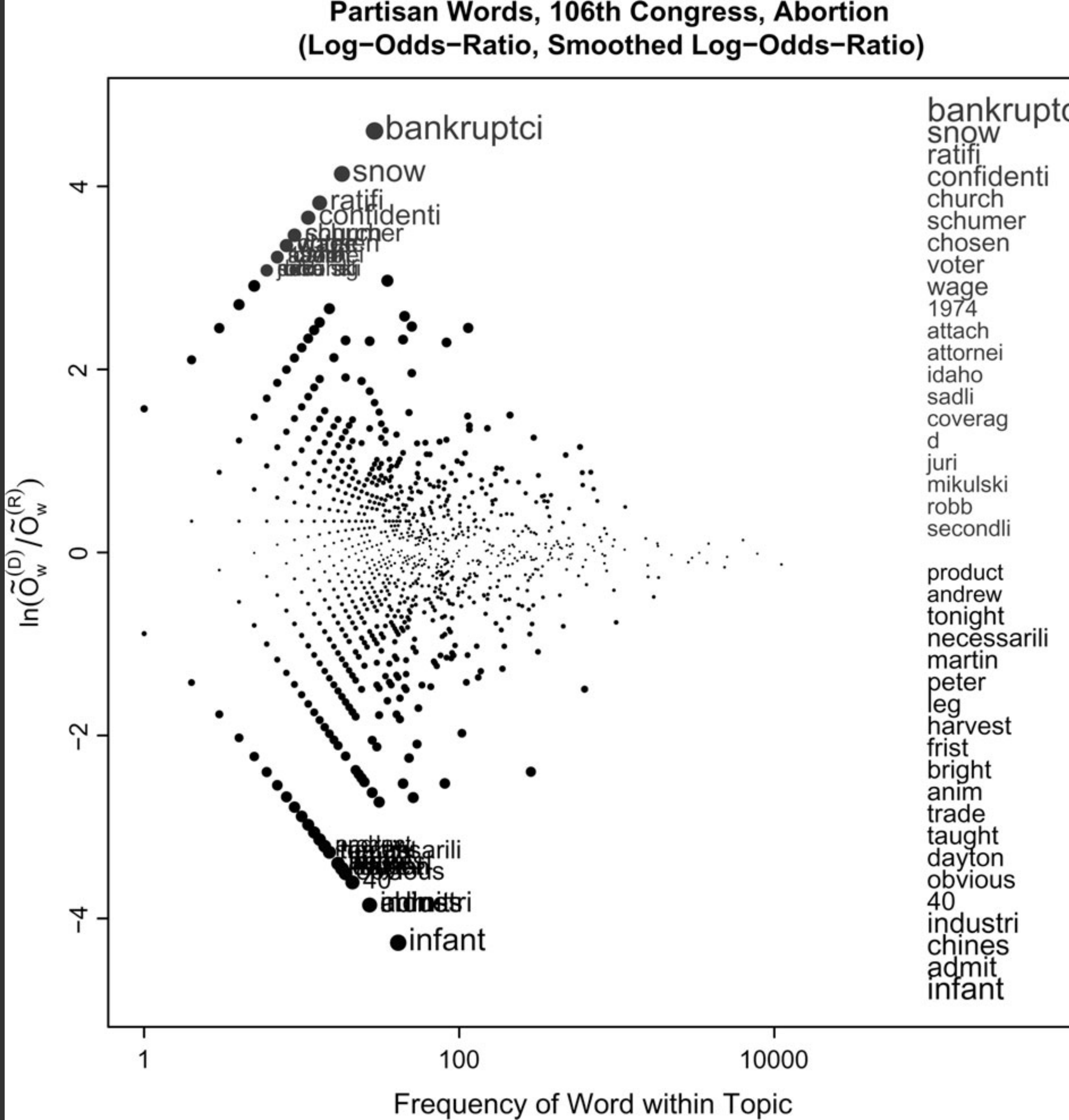
May not facilitate comparison (unstable layout)

Term frequency may not be meaningful

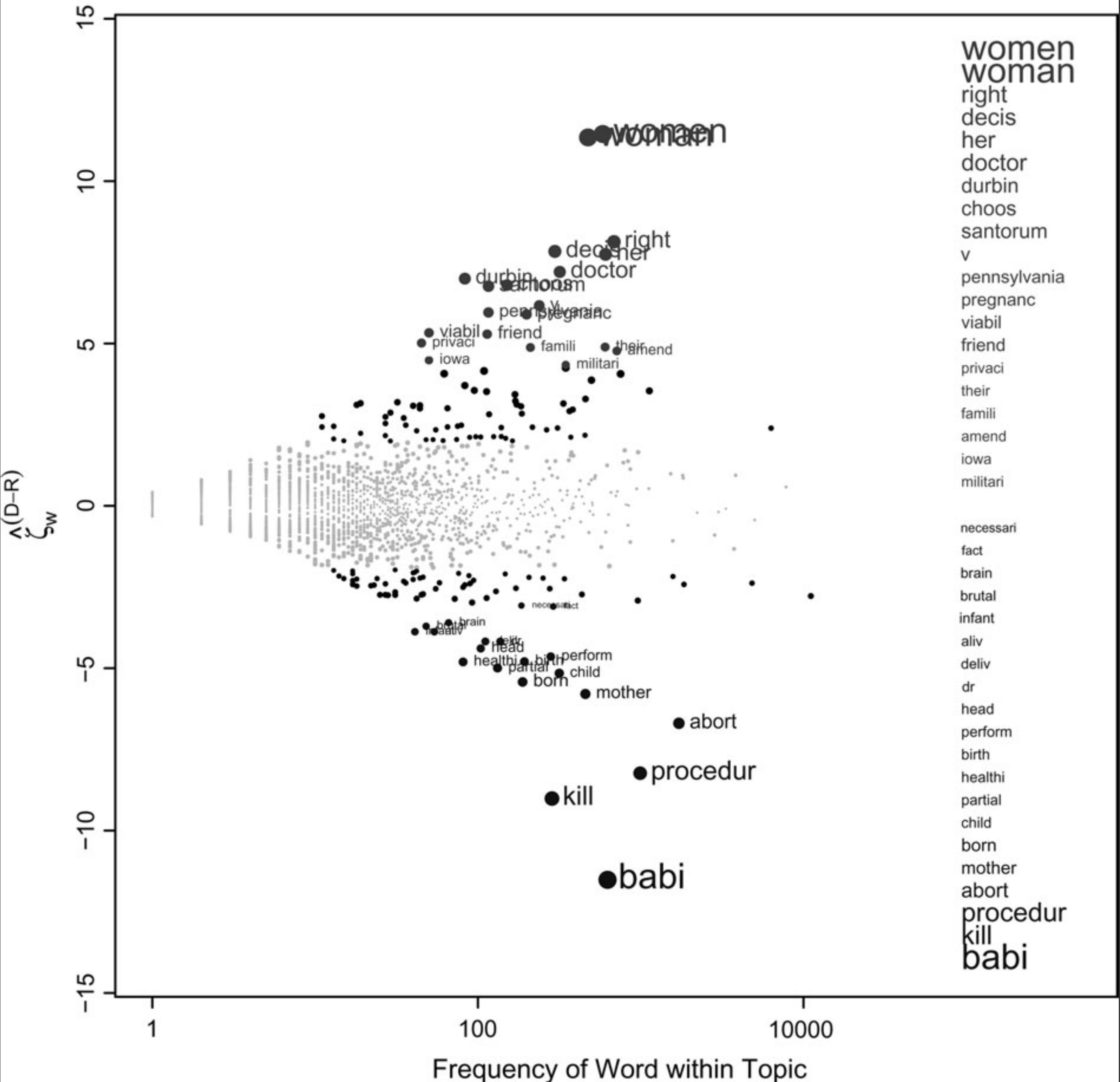
Does not show the structure of the text

Given a text, what are the best descriptive words?


Partisan Words, 106th Congress, Abortion
(Difference of Proportions)
$f_{kw}^{(D)} - f_{kw}^{(R)}$
0.005
0.000
-0.005
1
100
10000
Frequency of Word within Topic
to
women
right
senat
their
woman
amend
her
my
and
decis
famili
doctor
make
health
for
will
friend
court
law
life
born
fact
perform
but
ar
a
it
child
mother
you
be
that
kill
not
procedur
babi
of
abort
the


Partisan Words, 106th Congress, Abortion
(Log−Odds−Ratio, Smoothed Log−Odds−Ratio)
bankruptci
snow
ratifi
confidenti
infant
4
2
0
−2
−4
$\ln(\tilde{O}_w^{(D)}/\tilde{O}_w^{(R)})$
1
100
10000
Frequency of Word within Topic
bankruptc
snow
ratifi
confidenti
church
schumer
chosen
voter
wage
1974
attach
attornei
idaho
sadli
coverag
d
juri
mikulski
robb
secondli
product
andrew
tonight
necessarili
martin
peter
leg
harvest
frist
bright
anim
trade
taught
dayton
obvious
40
industri
chines
admit
infant

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



Keyword Weighting

Term Frequency

tf_{td} = count(t) in d

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

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

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 = # docs containing t; N = # of docs

Keyword Weighting

Require comparison across full corpus!

Term Frequency

tf_{td} = count(t) in d

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

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

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

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

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 information

- Grammar / part-of-speech
- Position within document
- Recognizable entities

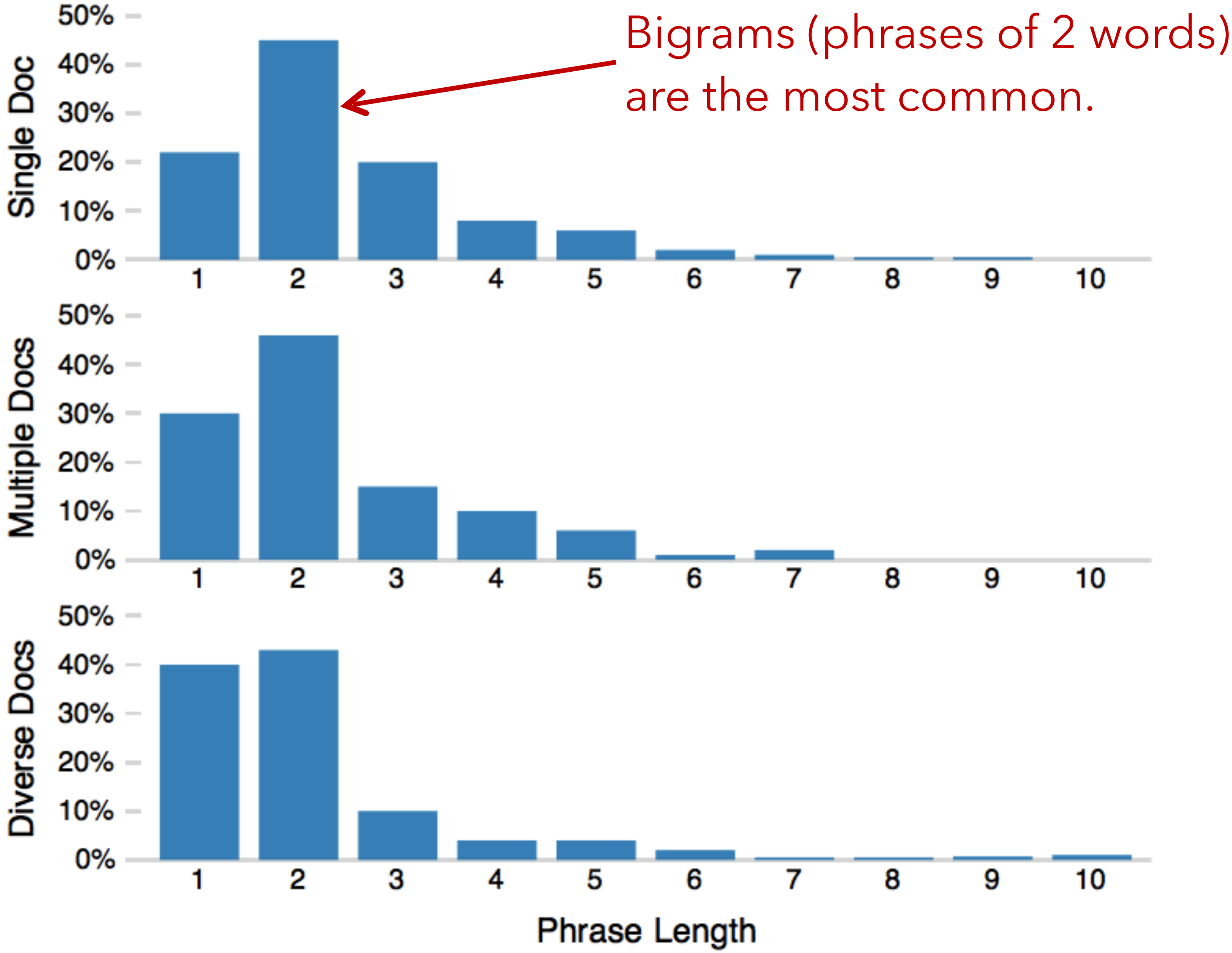
How do people describe text?

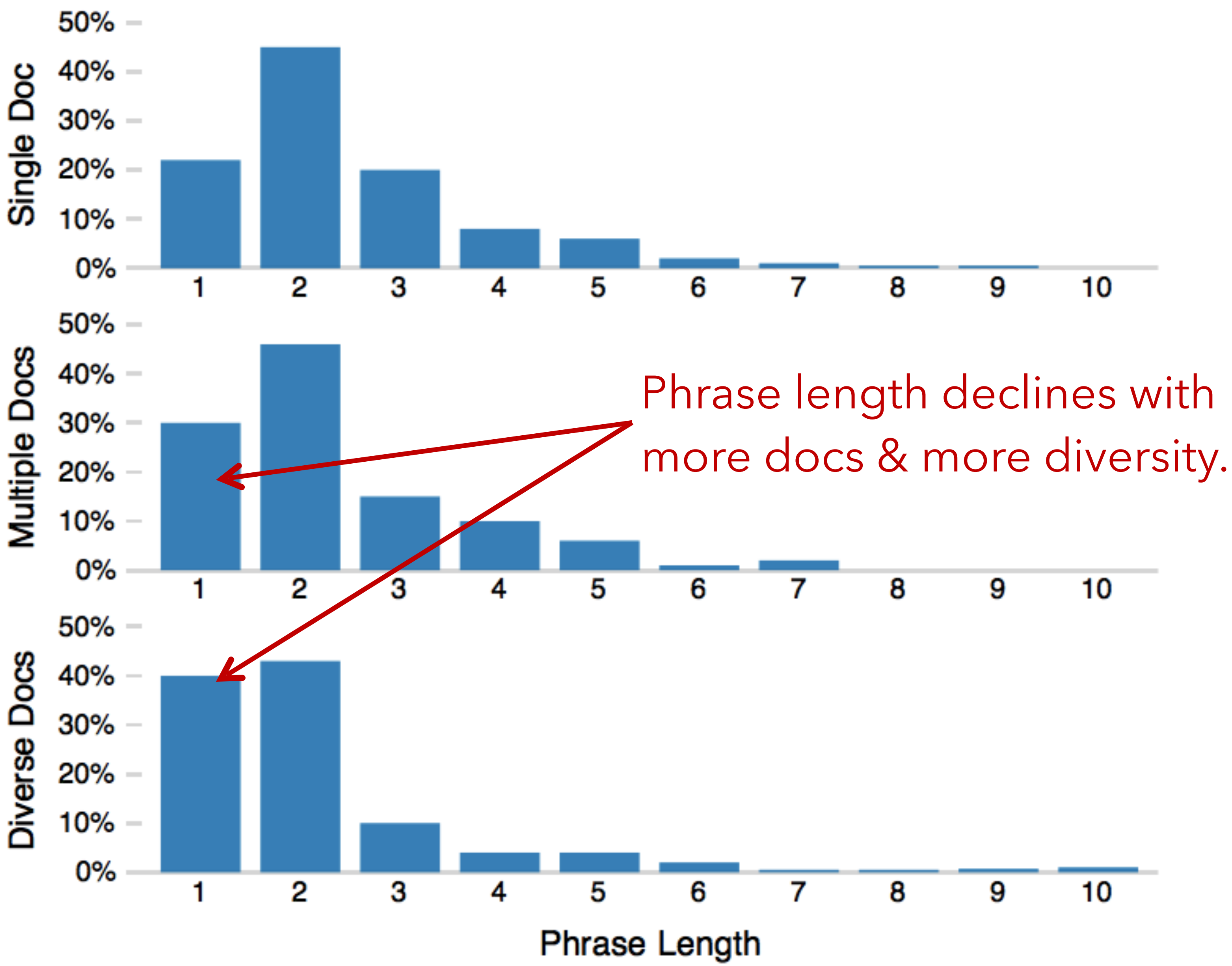
We asked 69 subjects (graduate students) to read and describe dissertation abstracts.

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]


Bigrams (phrases of 2 words) are the most common.
Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10
Phrase Length


Single Doc
Multiple Docs
Diverse Docs
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10
Phrase length declines with more docs & more diversity.
Phrase Length

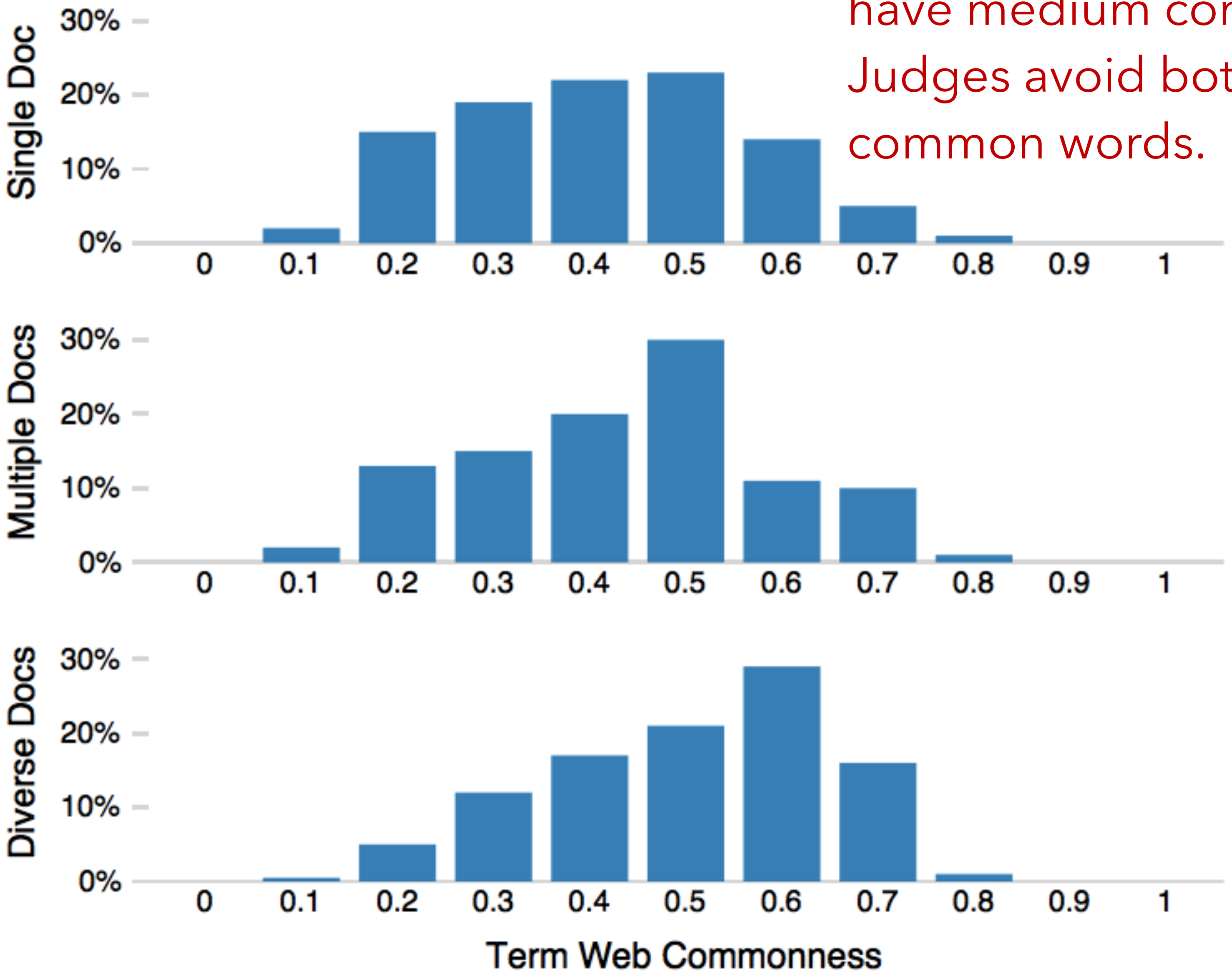
Term Commonness

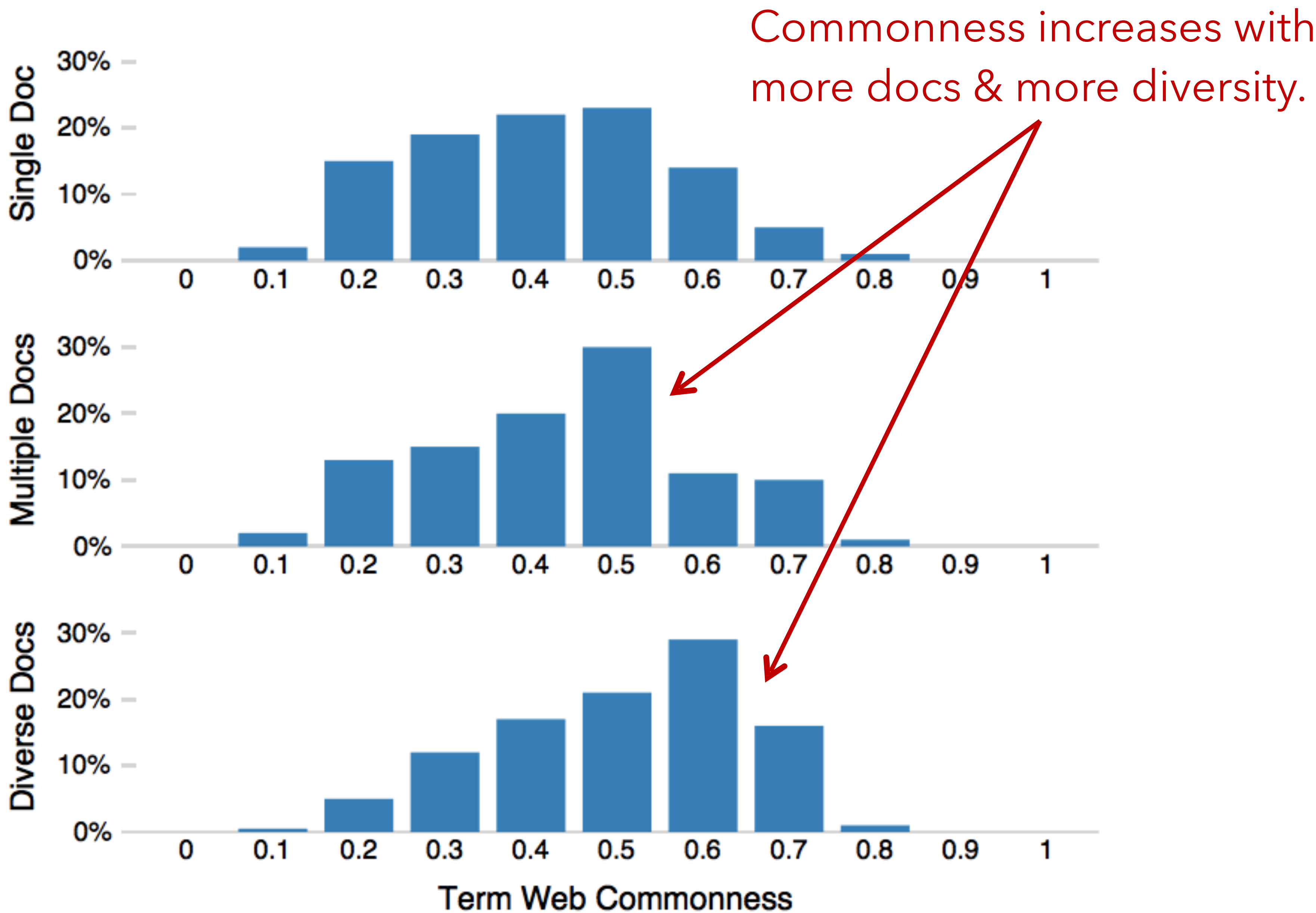
$$\log(tf_w) \ / \ \log(tf_{the})$$

The normalized term frequency relative to the most frequent n-gram, i.e., the word "the".

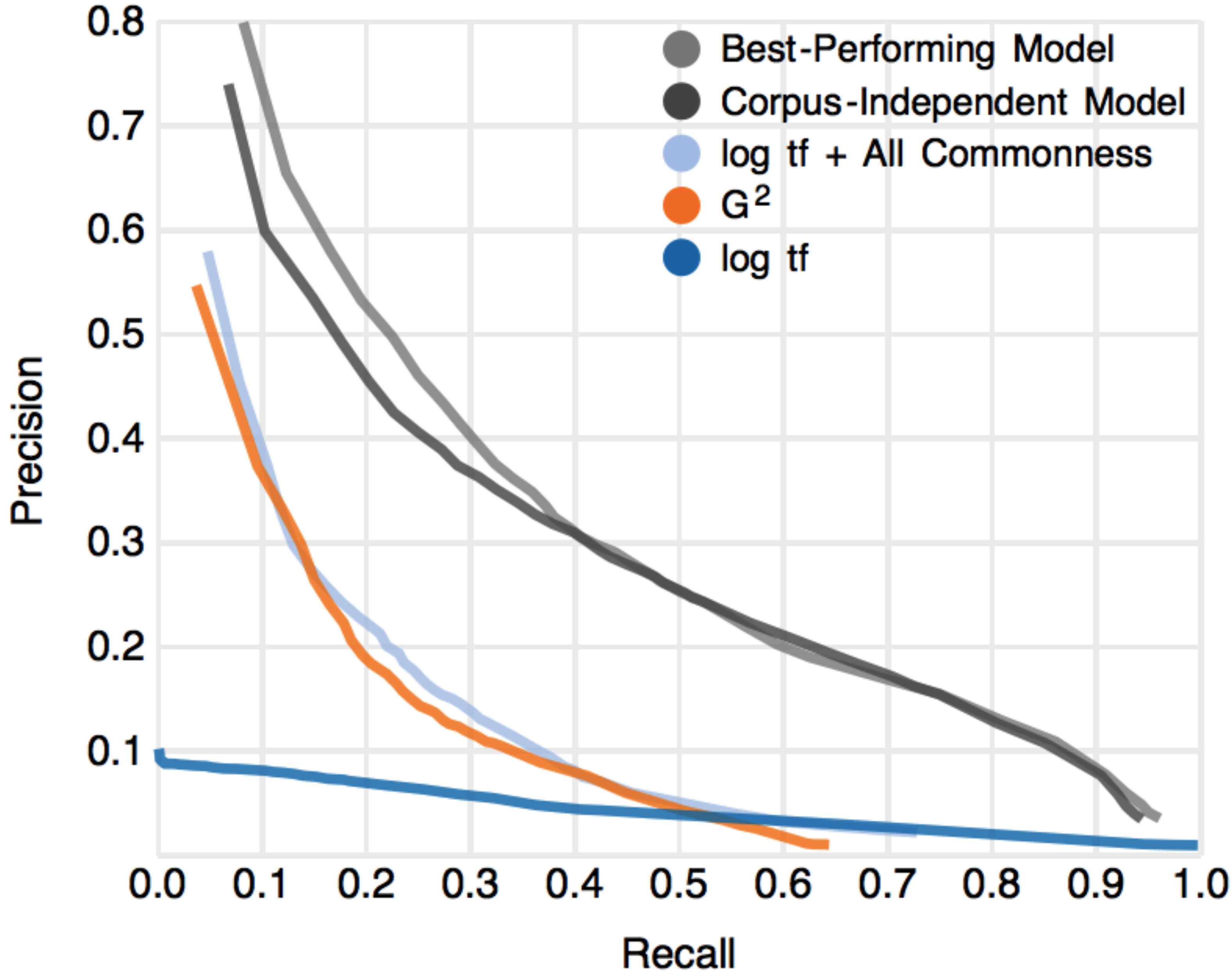
Measured across a corpus or across the entire English language (using Google n-grams)

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




Commonness increases with more docs & more diversity.
Single Doc
Multiple Docs
Diverse Docs
30%
20%
10%
0%
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Term Web Commonness

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.

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.

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 moreabout how the process works and watch the action unfold, click above to see the video story.

G^2

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

Regression Model

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 pilots
sea-based fighter

Yelp Review Spotlight (Yatani 2011)

Yelp Review Spotlight (Yatani 2011)

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

Administrivia

Final Project Schedule

Proposal	***Thur, May 10***
Milestone	Mon, May 21 (reviews 5/22, 5/24)
Final Paper	Wed, May 30
Poster & Demo	Thur, May 31 (11:45am-2pm)

Logistics

Final project description posted online

Work in groups of up to 4 people

Start thinking about project topics!

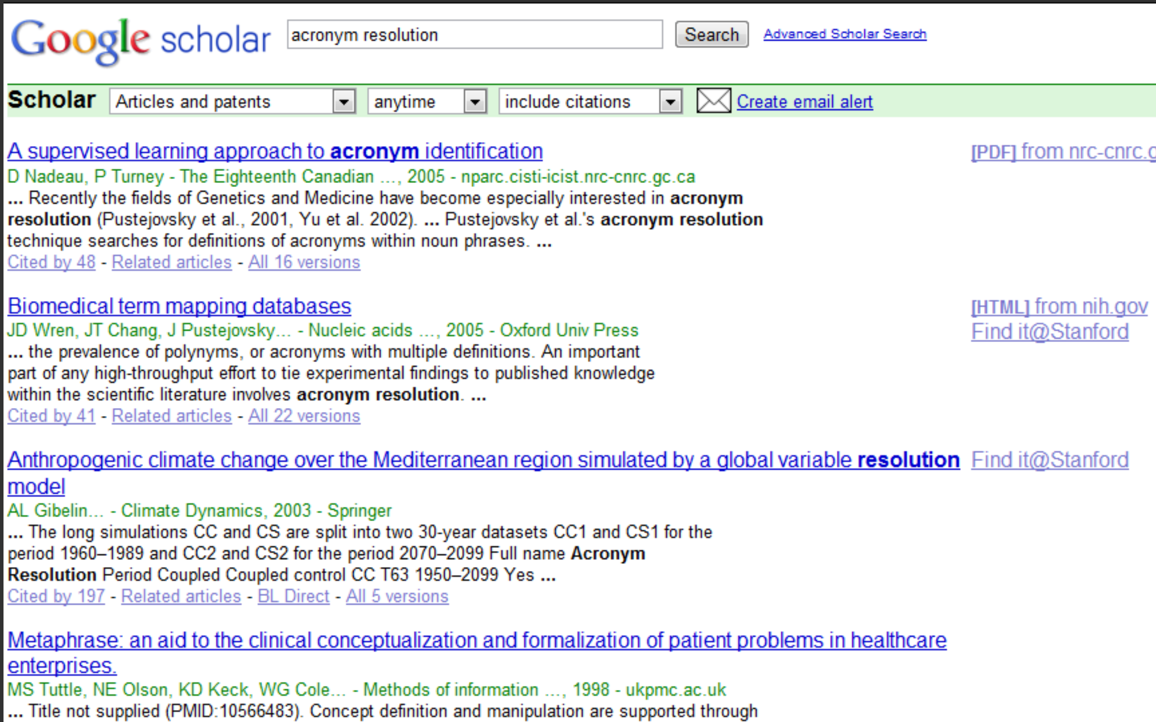
Document Content

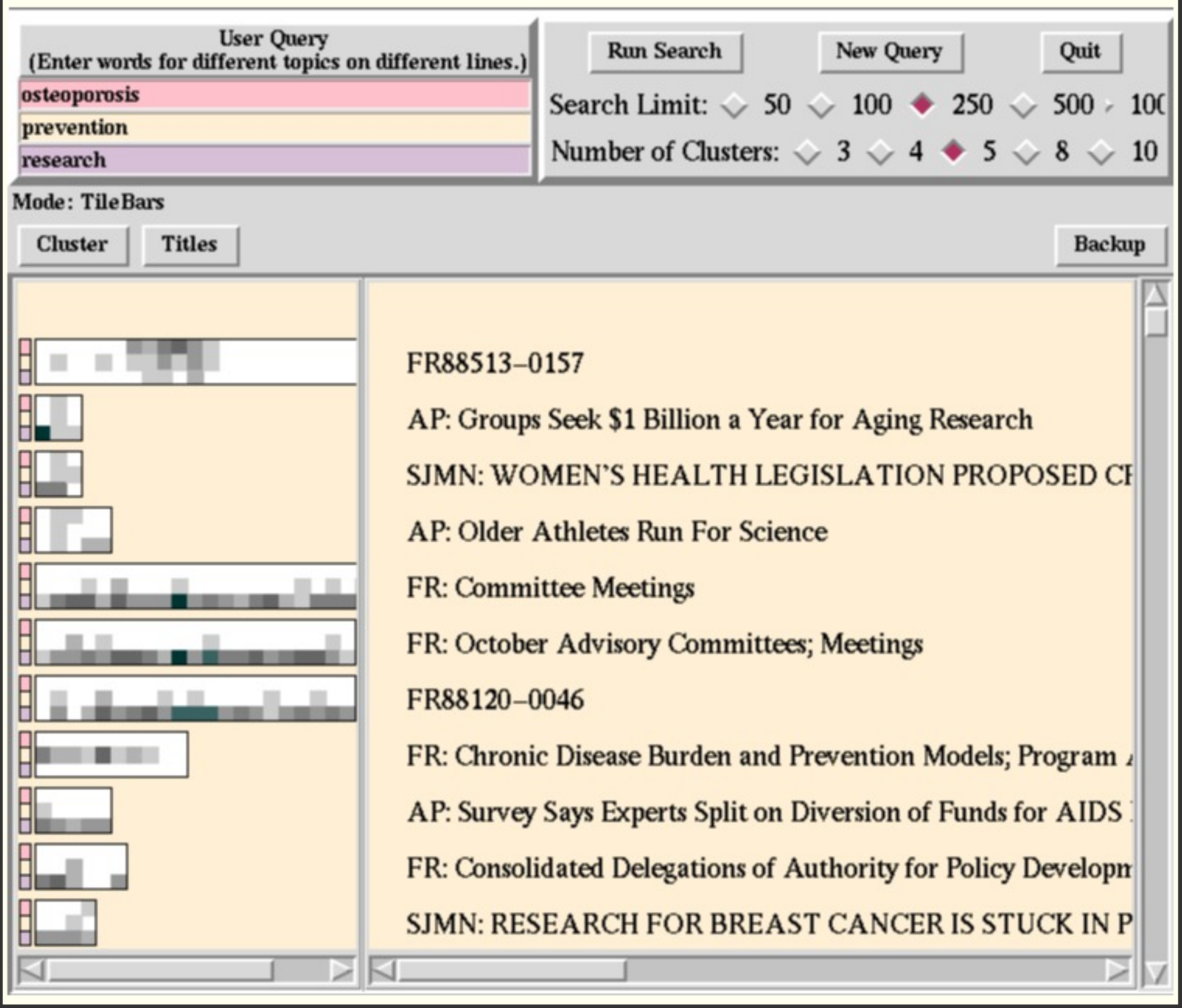
Information Retrieval

Search for documents

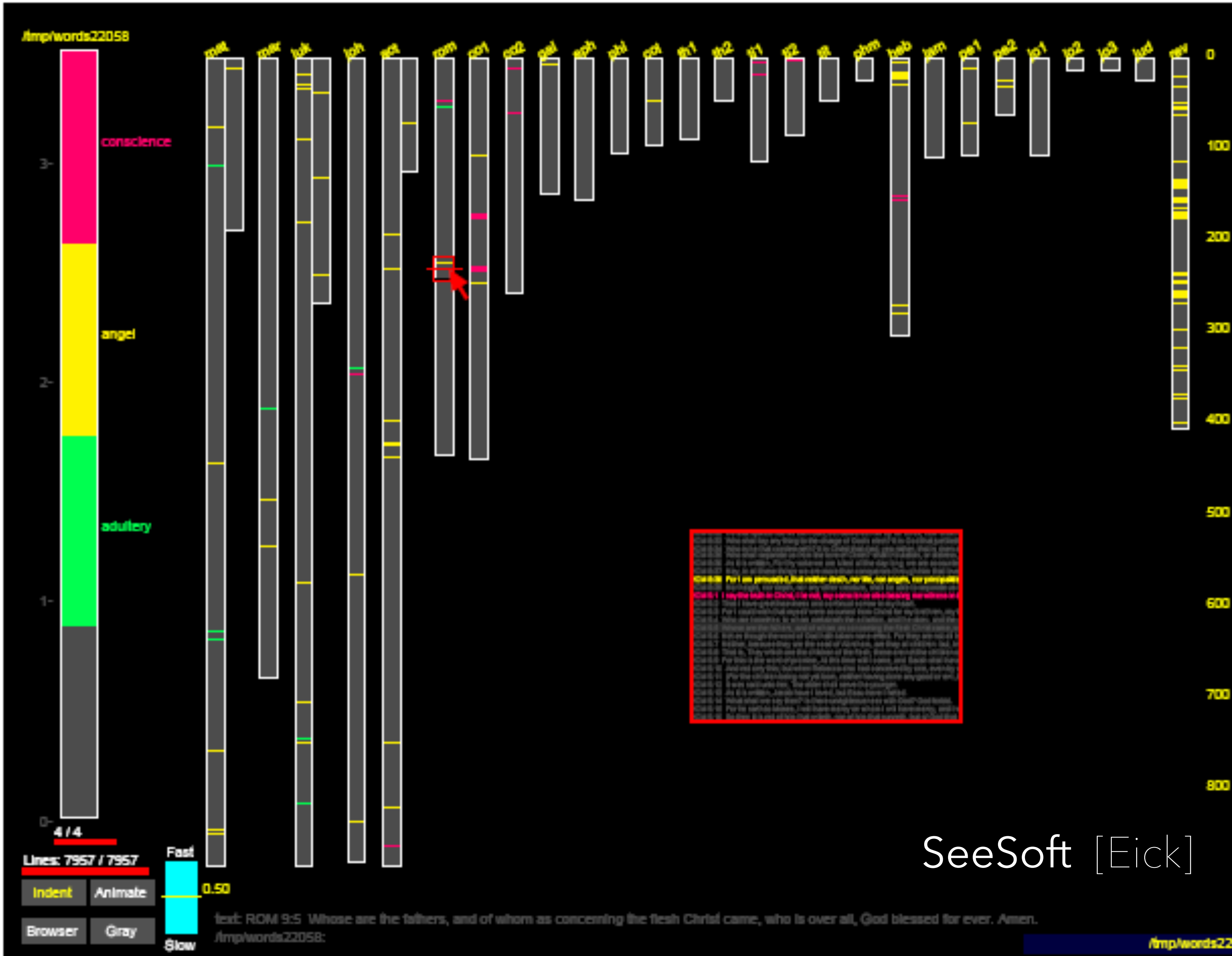
Match query string with documents

Visualization to **contextualize results**

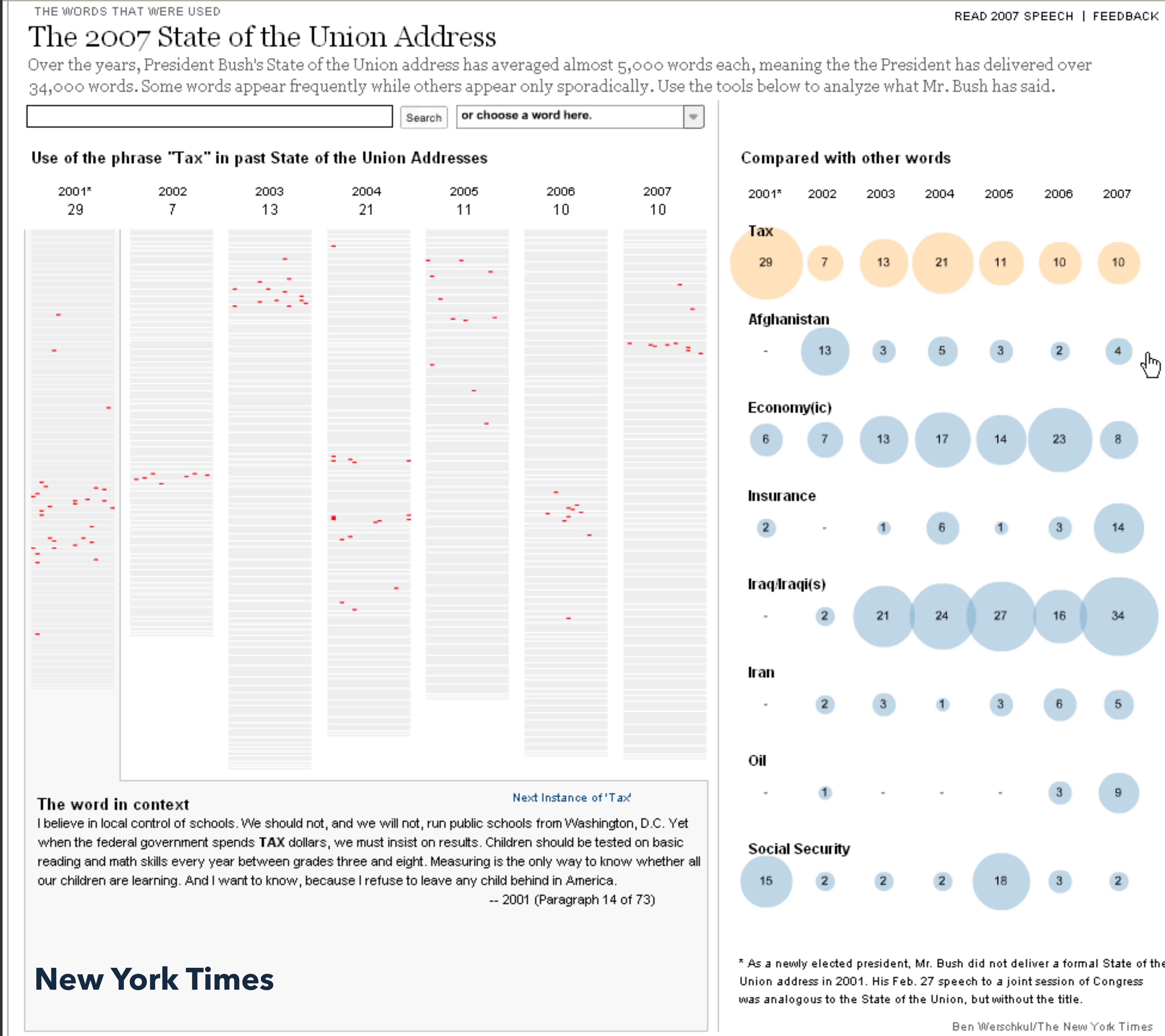




TileBars [Hearst]



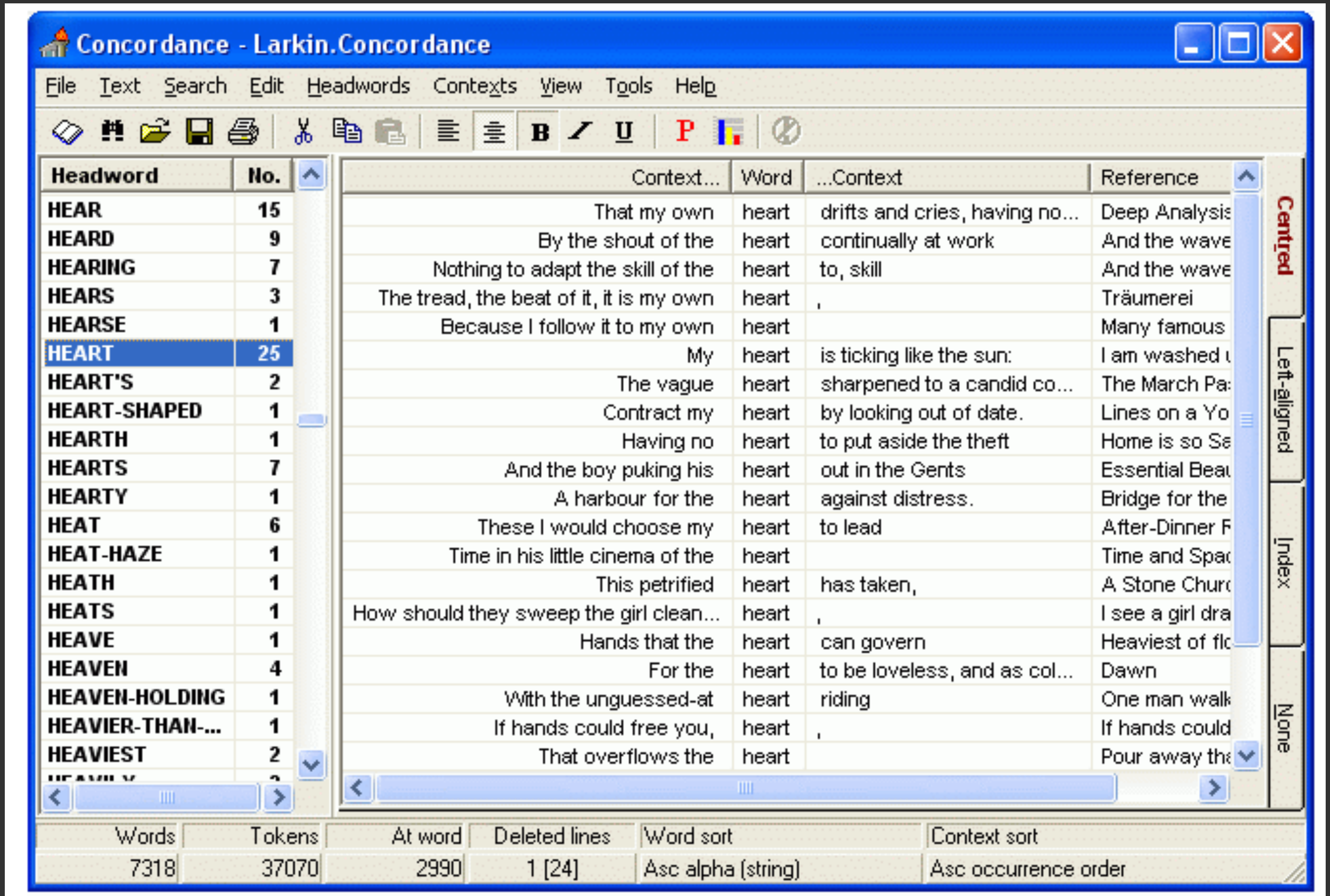
SeeSoft [Eick]



New York Times

Concordance

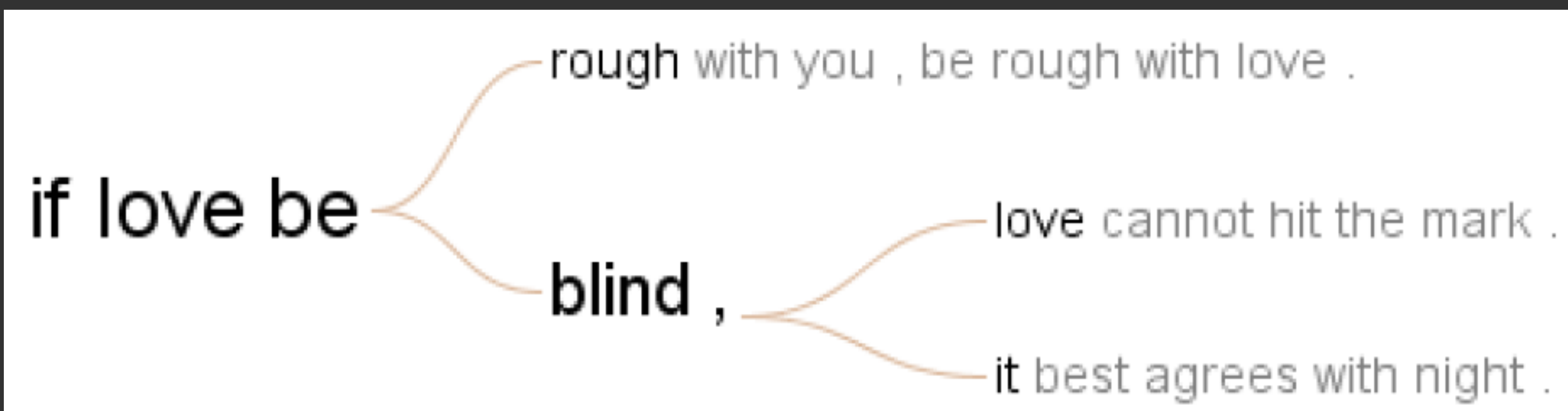
What is the common local context of a term?



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

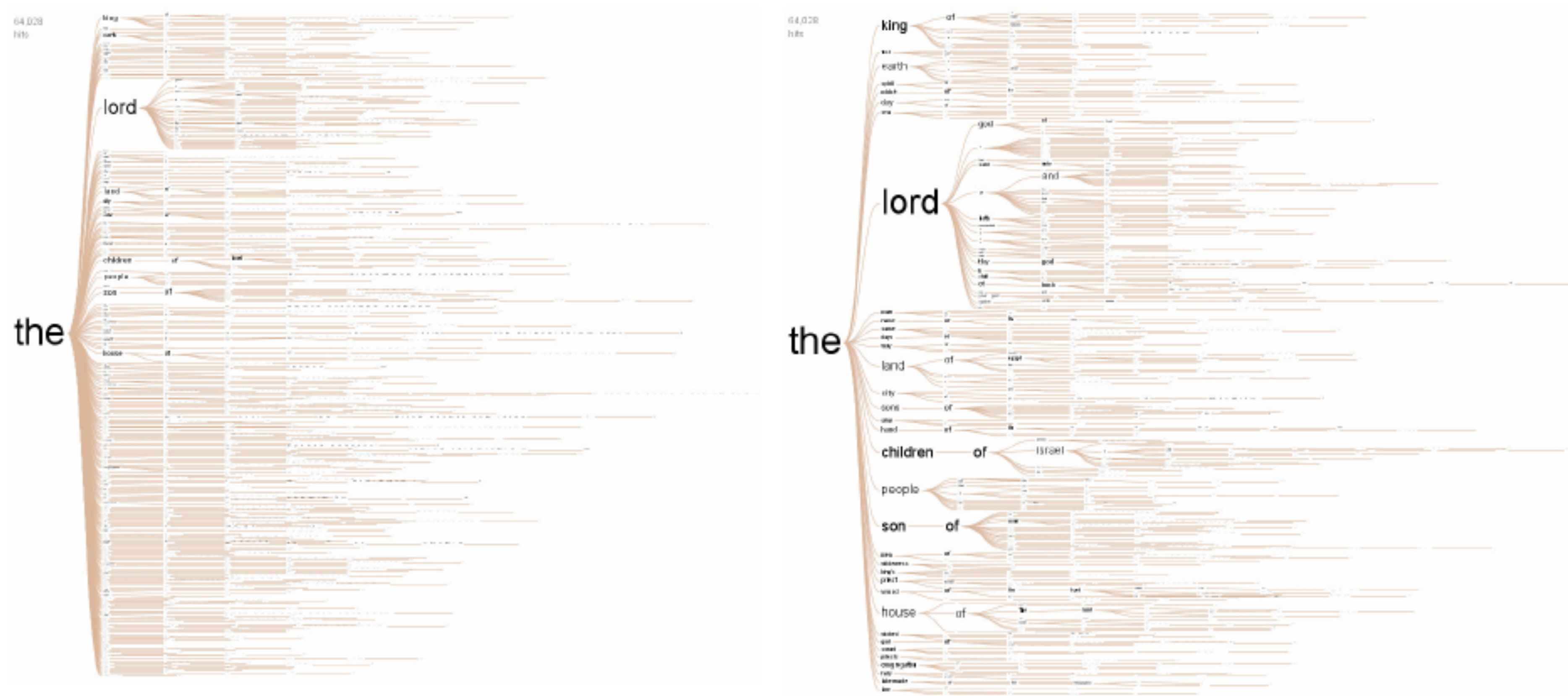
if love be blind , love cannot hit the mark .

if love be blind , it best agrees with night .

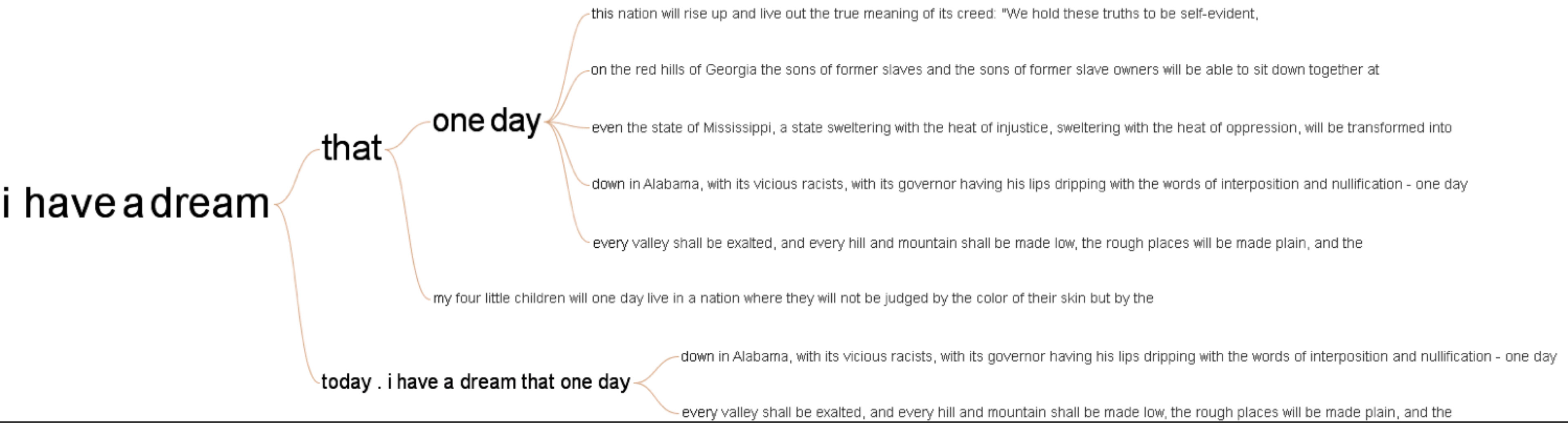


Word Tree [Wattenberg et al.]

Filter Infrequent Runs



Recurrent Themes in Speeches



Visualizations : Word tree / Alberto Gonzales

Creator: Martin Wattenberg

Tags:

explore
- visualizations
- data sets
- comments
- topic hubs

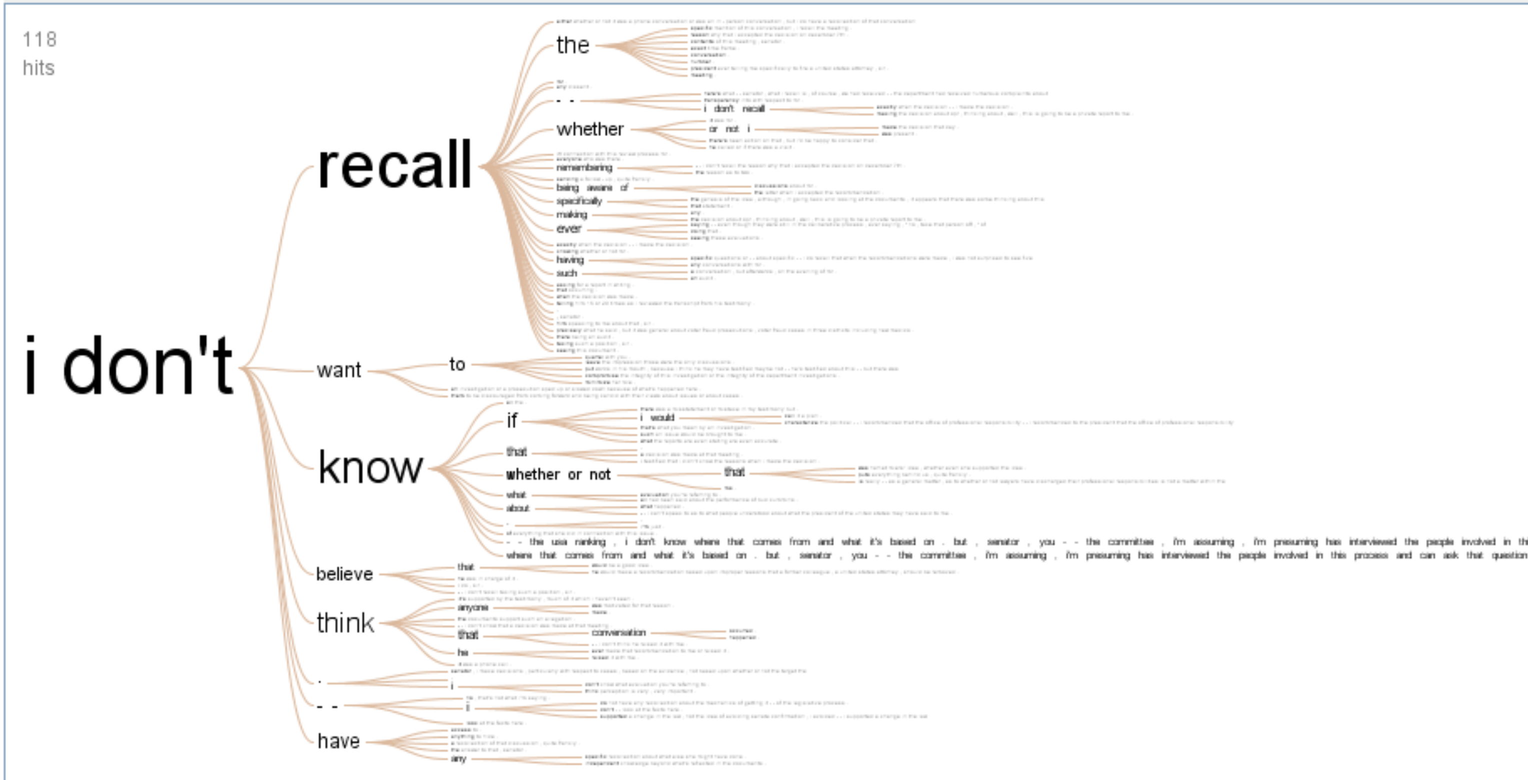
participate
- create visualization
- upload data set
- create topic hub
- register

learn more
- quick start
- visualization types
- data format & style
- about Many Eyes
- FAQ
- blog

contact Us
- contact
- report a bug

legal
- terms of use

Search i don't Back Forward Start End Occurrence Order Clicks Will Zoom



Data file: Word in testimony from Gonzales, 4/19/2007 Data source: CQ Transcript Wire via the Washington Post This data set has not yet been rated

full image share this watch this add to topic hub rate this

Comments (4)

currently showing

This visualization has 4 positive and 0 negative

Popular Dataset Tags

2007 2008 bible blog books census crime education eharmony election energy food health inauguration internet ireland literature lyrics media music network obama people politics population president prices religion

Glimpses of Structure…

Concordances show local, repeated structure

But what about other types of patterns?

Lexical: <A> at <B>

Syntactic: <Noun> <Verb> <Object>

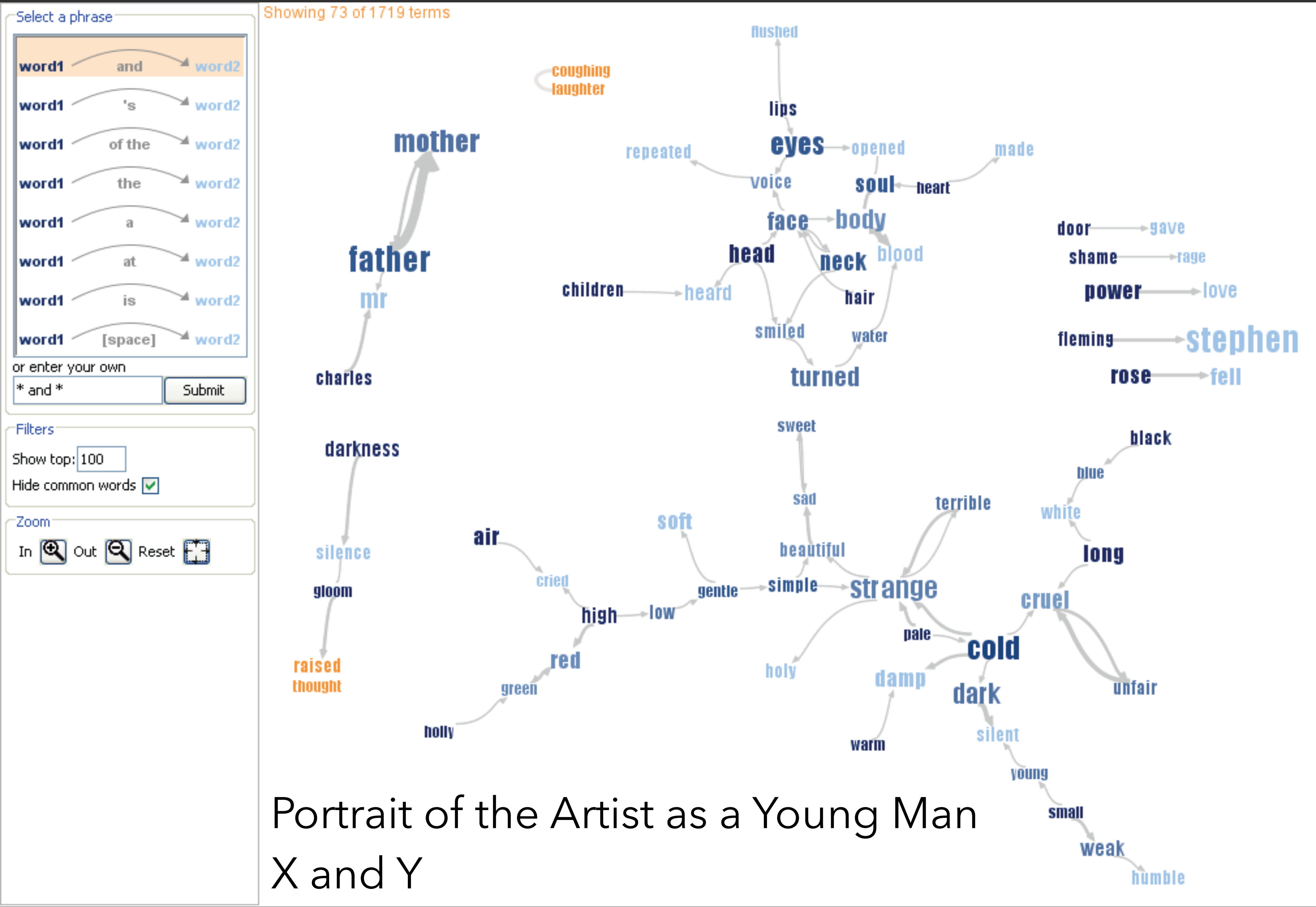
Phrase Nets [van Ham et al.]

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

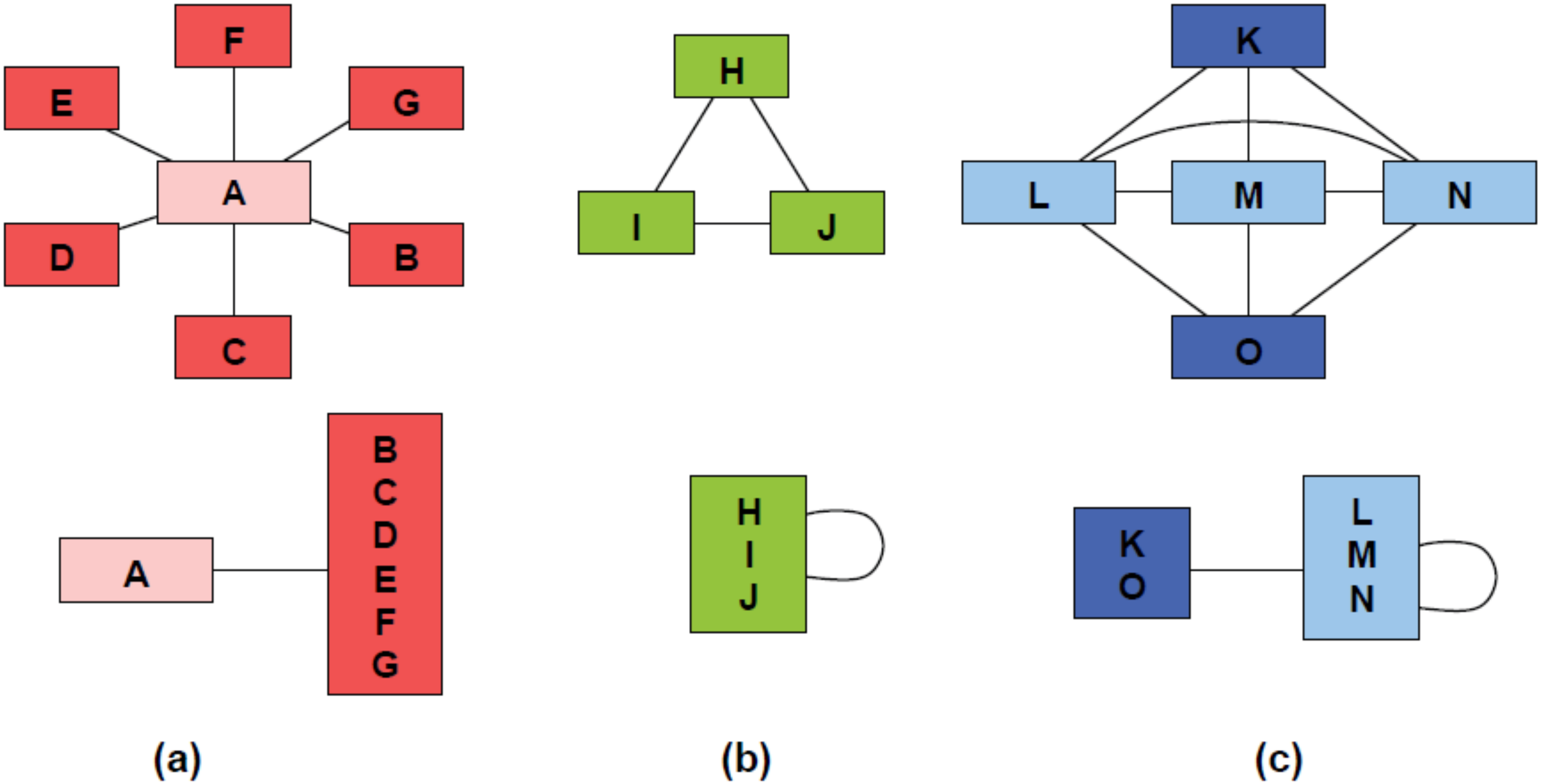
- Occurrences -> Node size
- Pattern position -> Edge direction



Portrait of the Artist as a Young Man

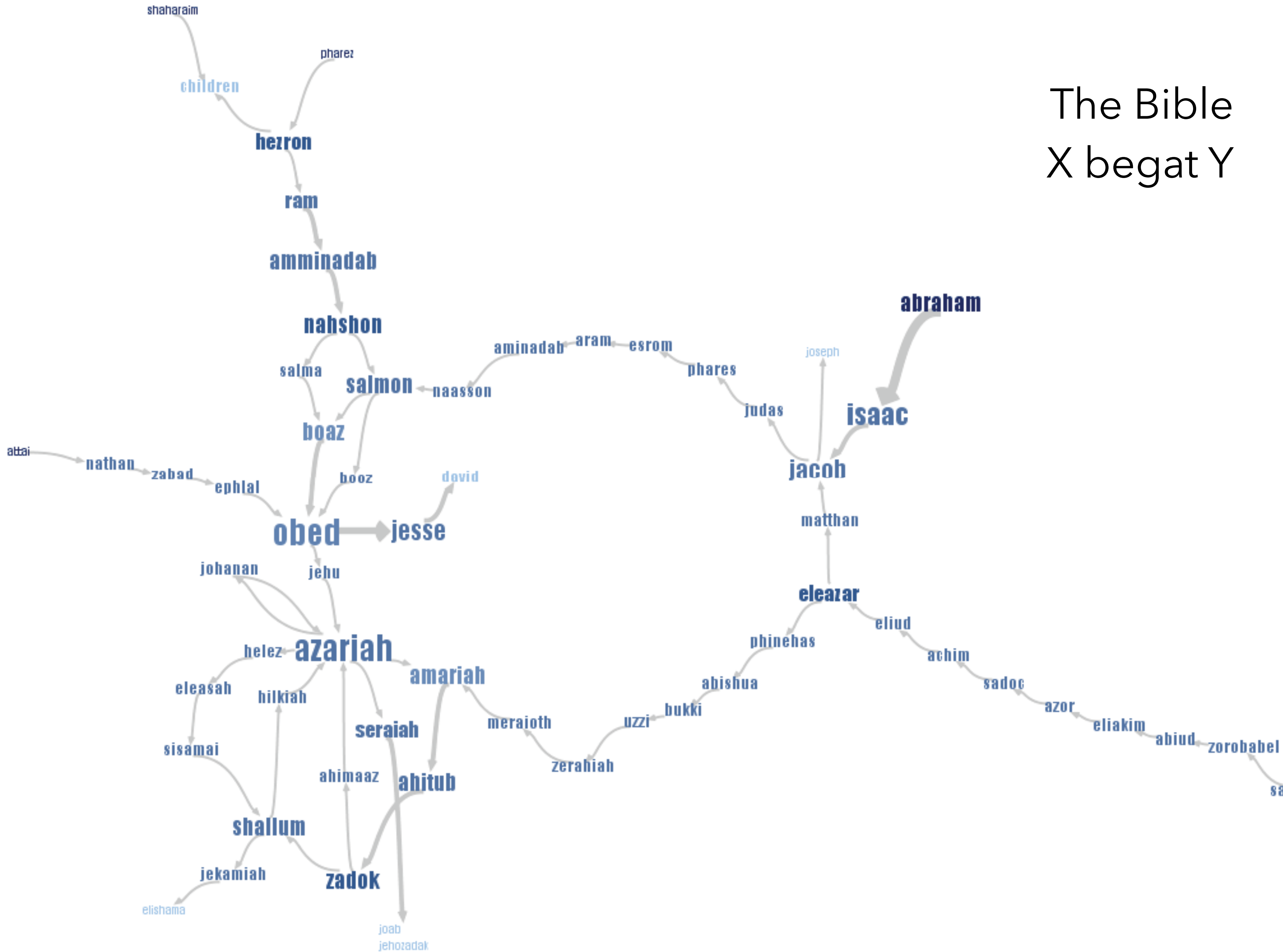
X and Y

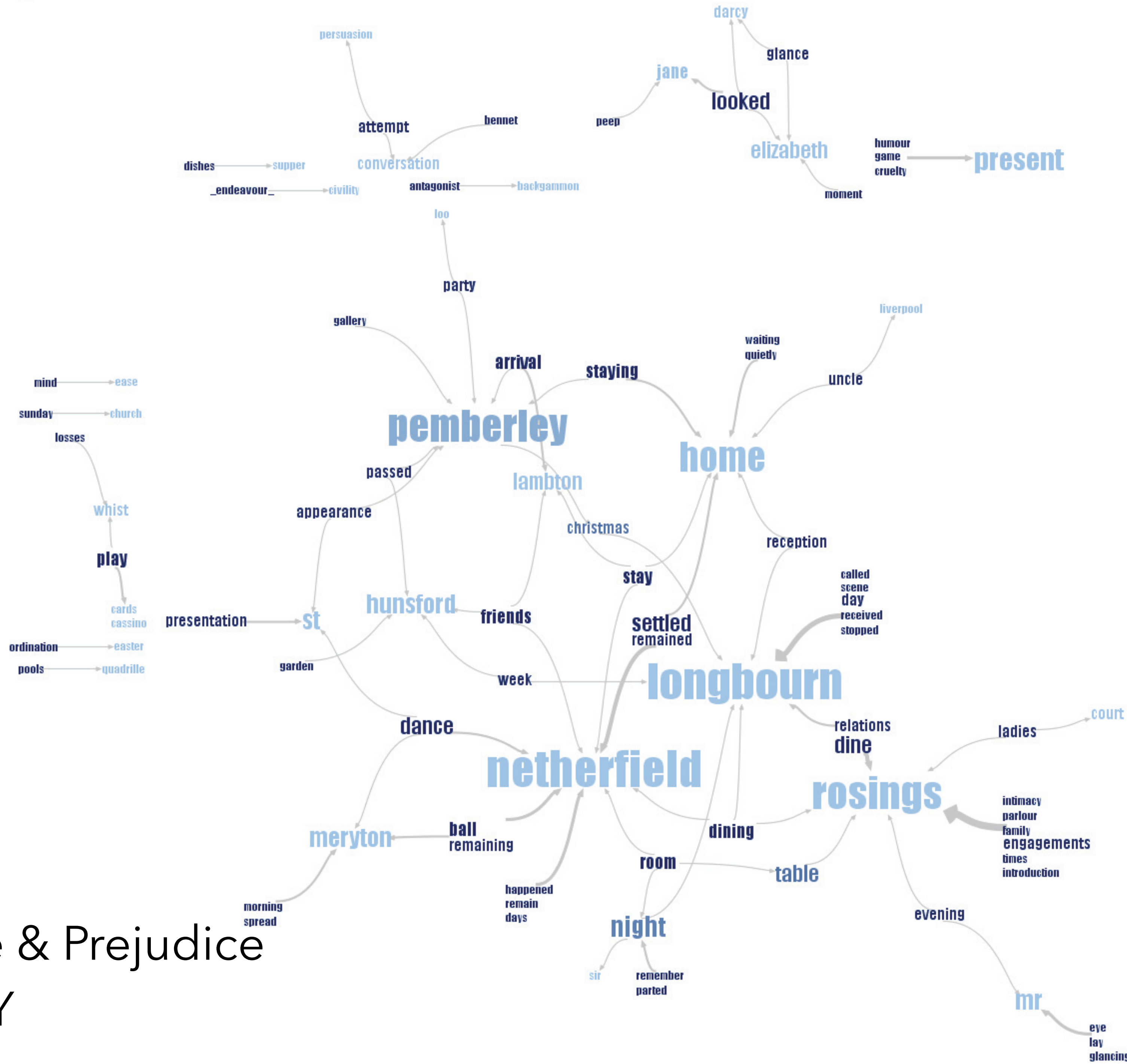
Node Grouping



The Bible

X begat Y

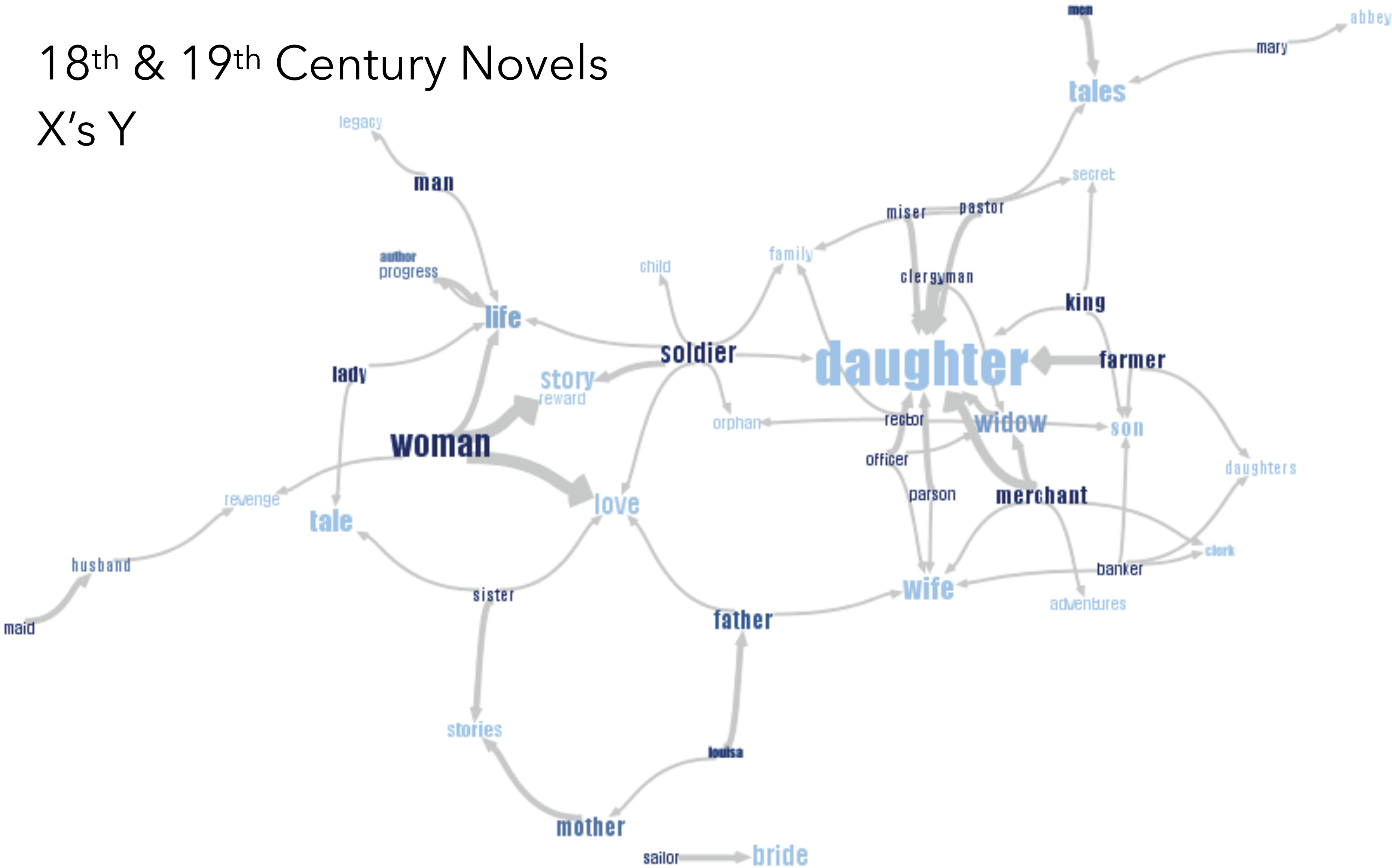




Pride & Prejudice

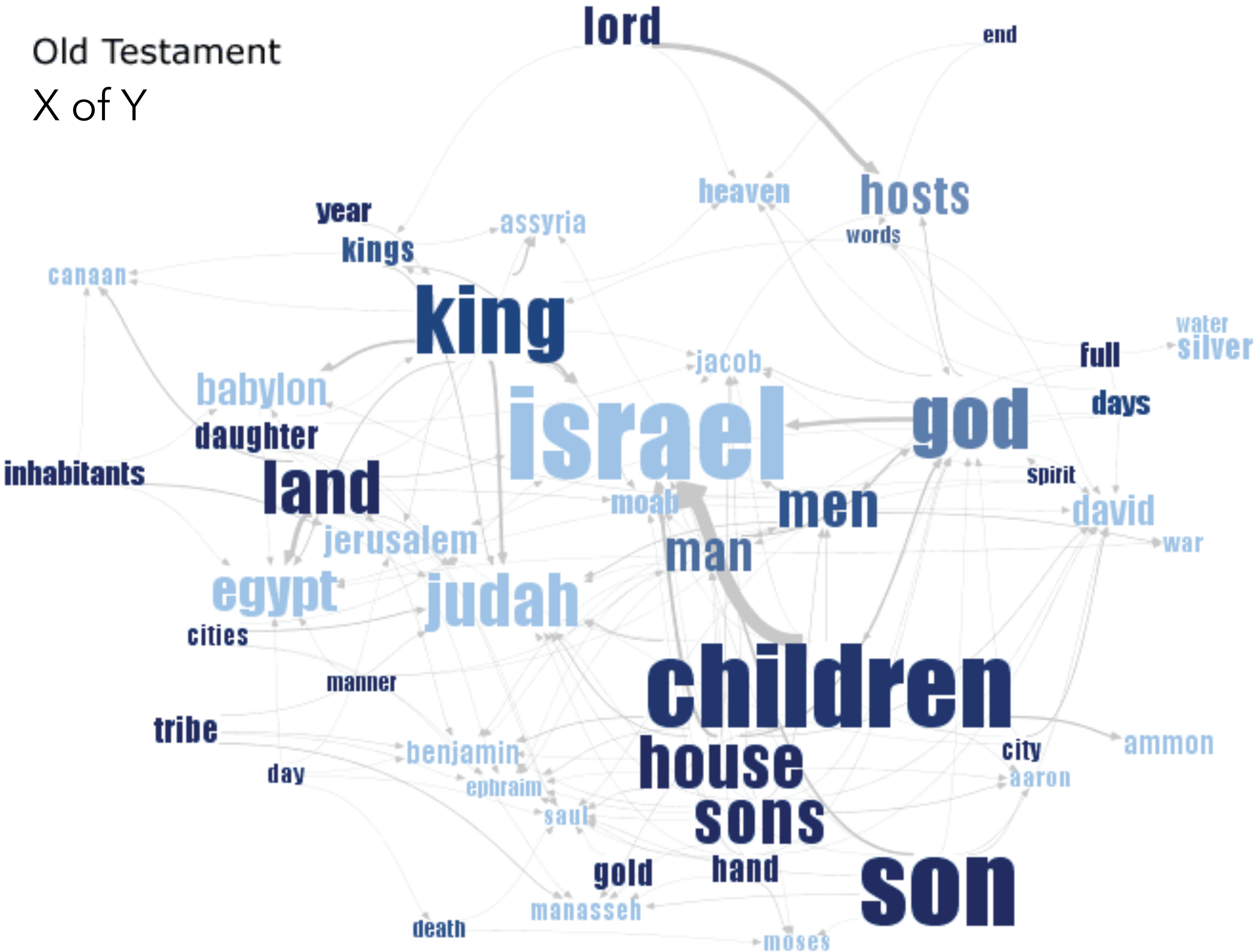
X at Y

18th & 19th Century Novels
X's Y



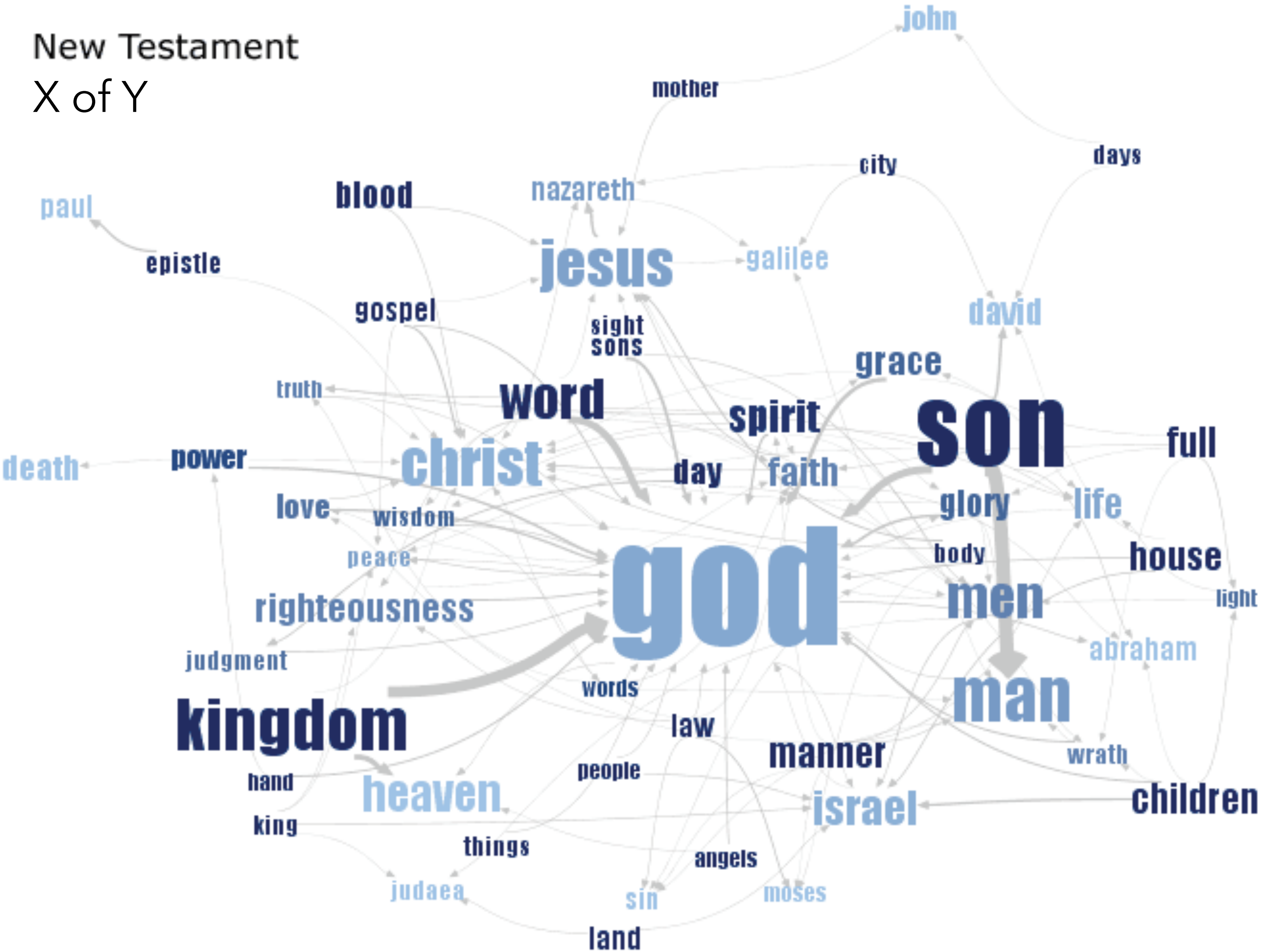
Old Testament

X of Y



New Testament

X of 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"

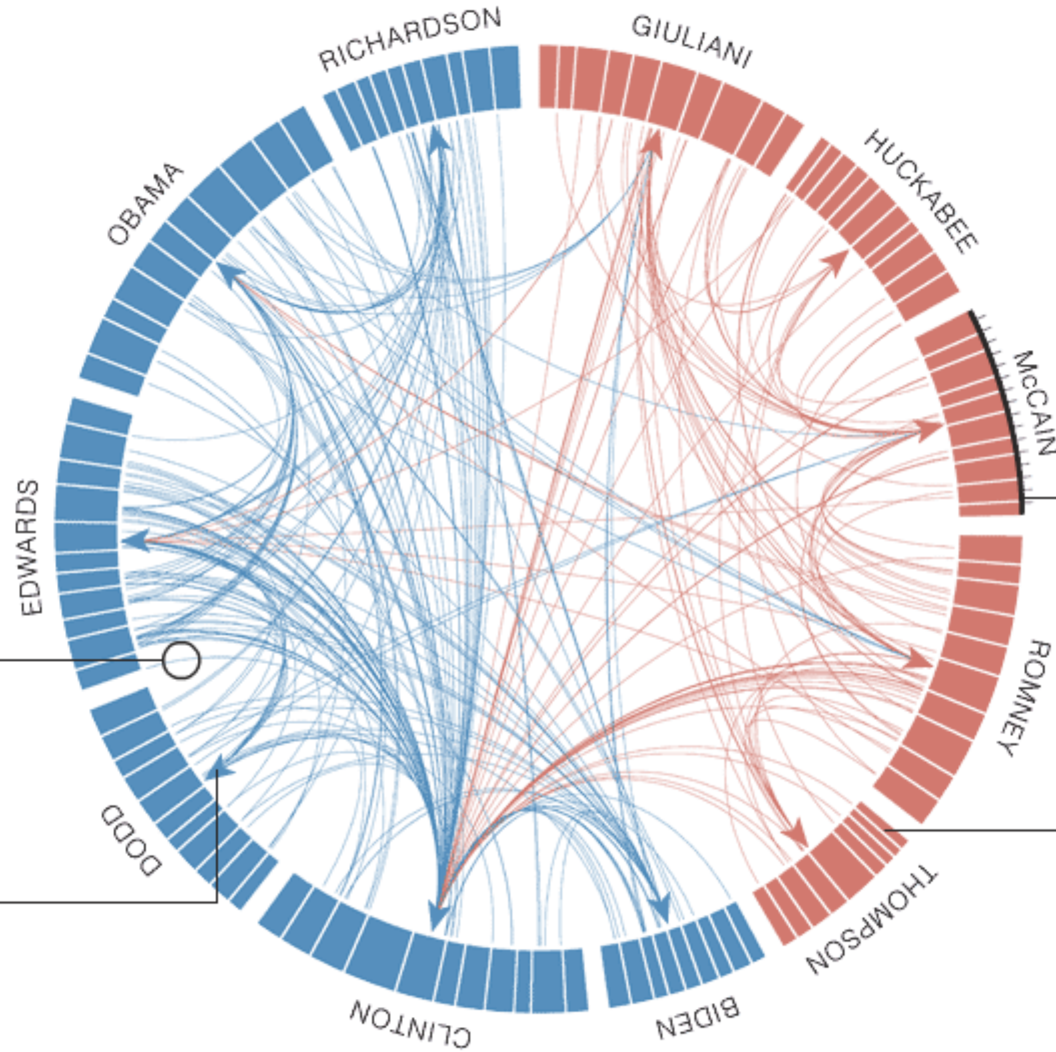
Who x Who -> Social network

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.



Each thin line represents one candidate speaking the last name of another candidate.

Every line ends at an arrow, which points to the name that was spoken.

The length of each circle segment represents the total number of words spoken by the candidate during the debates. Each tick mark represents 1,000 words.

Each slice represents one debate, arranged clockwise from the first to the final debate.

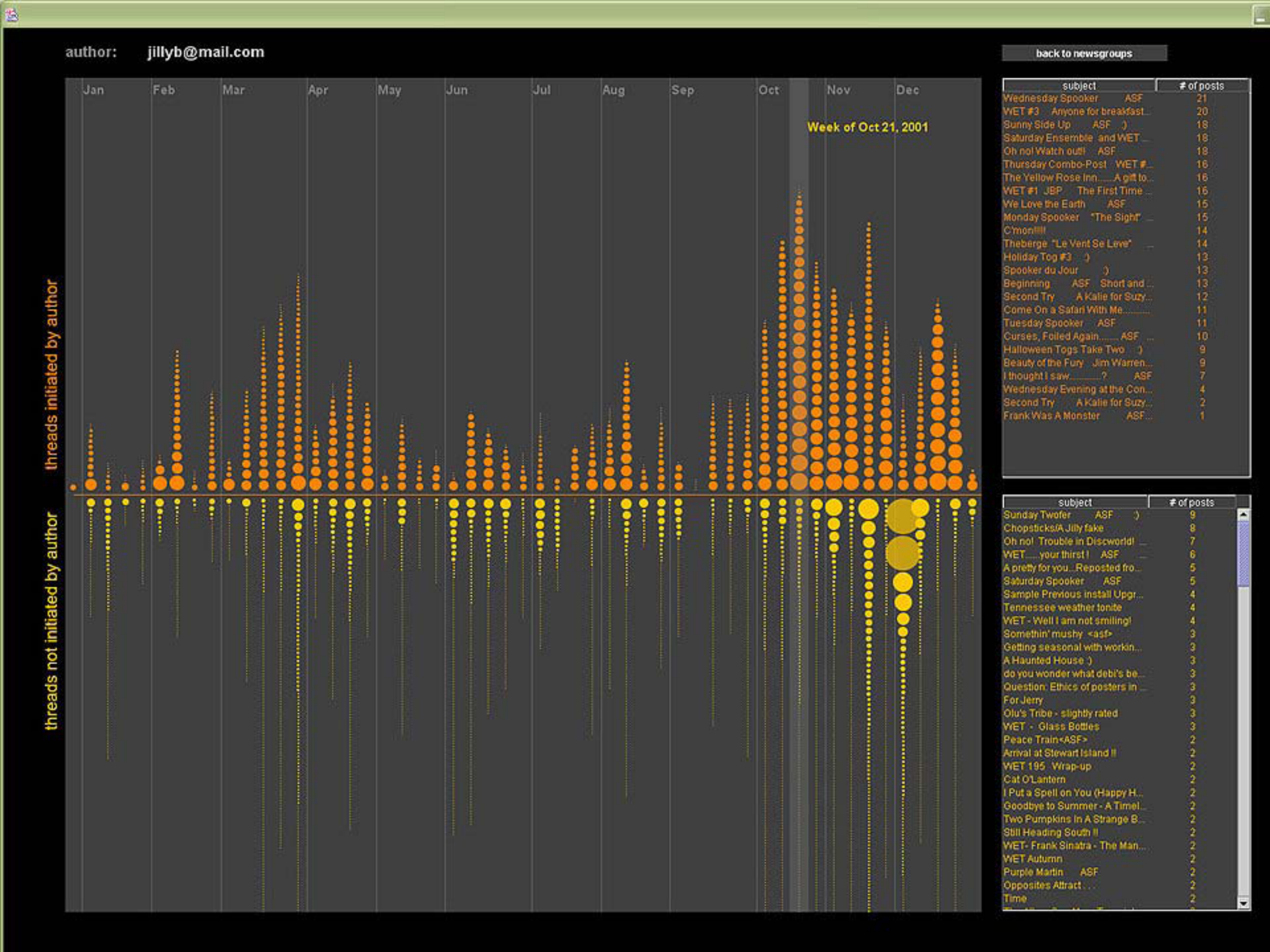
Source: Debate transcripts

Jonathan Corum and Farhana Hossain/The New York Times

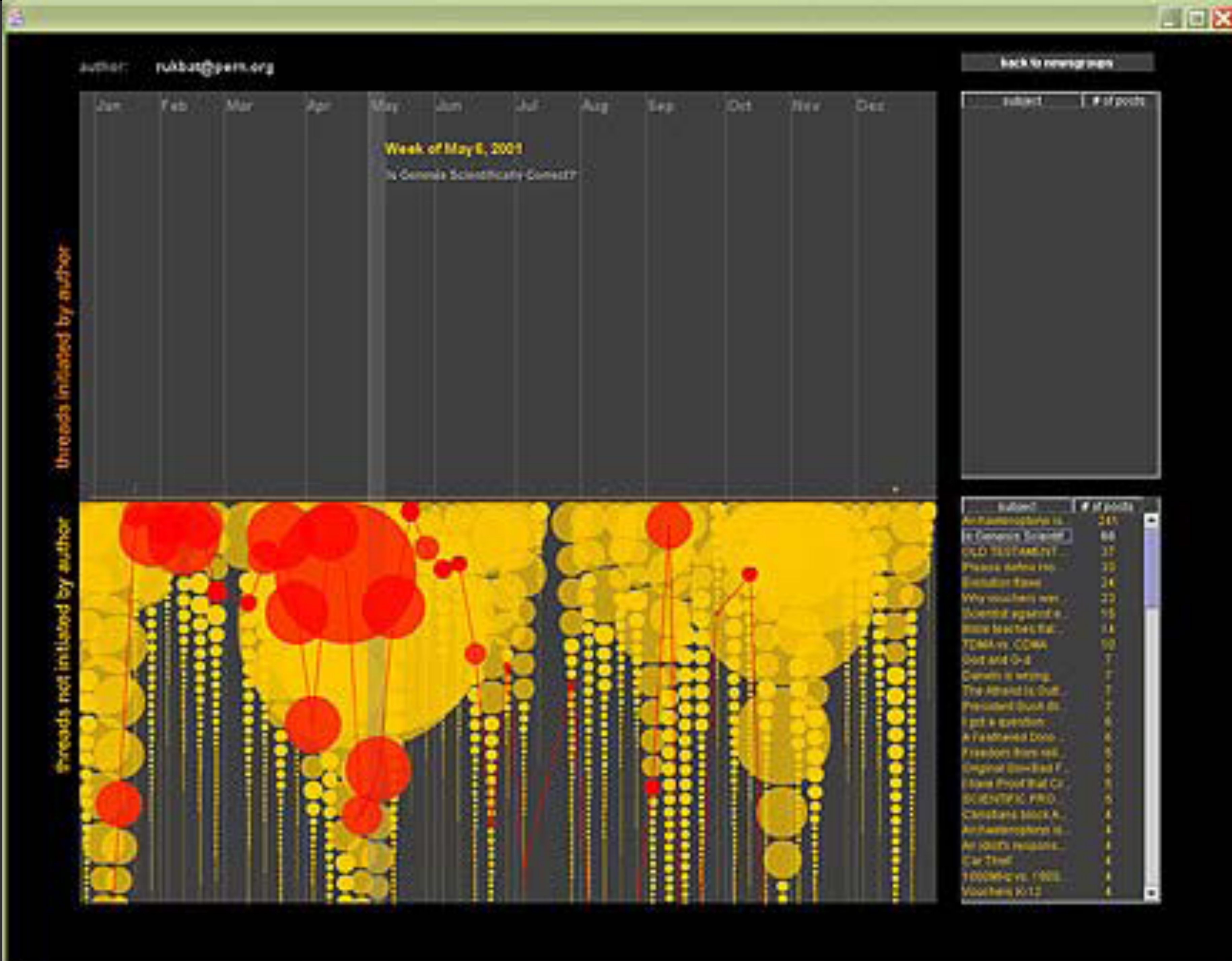
Usenet Visualization [Viegas & Smith]

Show correspondence patterns in text forums

Initiate vs. reply; size and duration of discussion



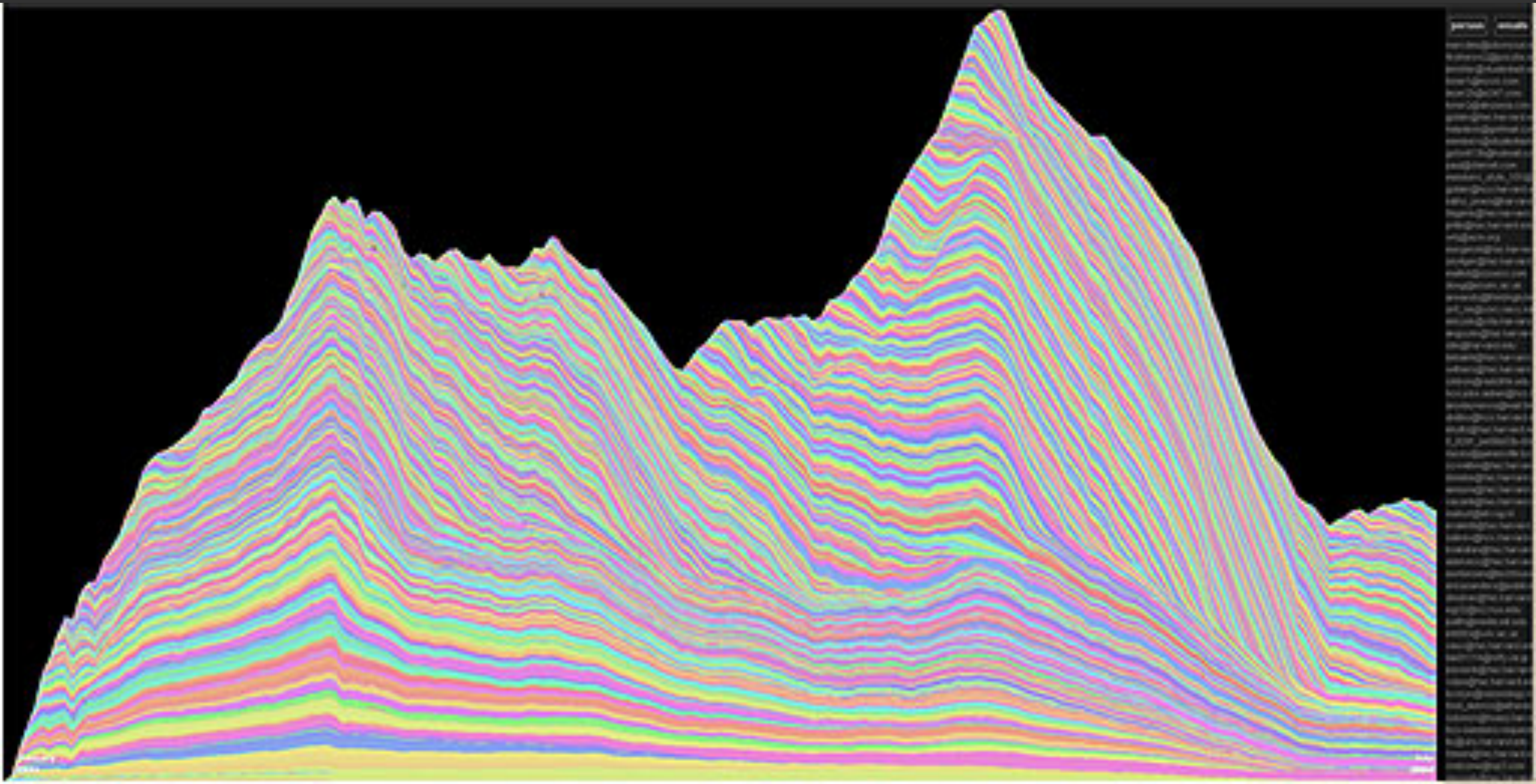
author: jillyb@mail.com
back to newsgroups
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Week of Oct 21, 2001
threads initiated by author
threads not initiated by author
subject
# of posts
Wednesday Spooker ASF 21
WET #3 Anyone for breakfast... 20
Sunny Side Up ASF :) 18
Saturday Ensemble and WET... 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 Leve" ... 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



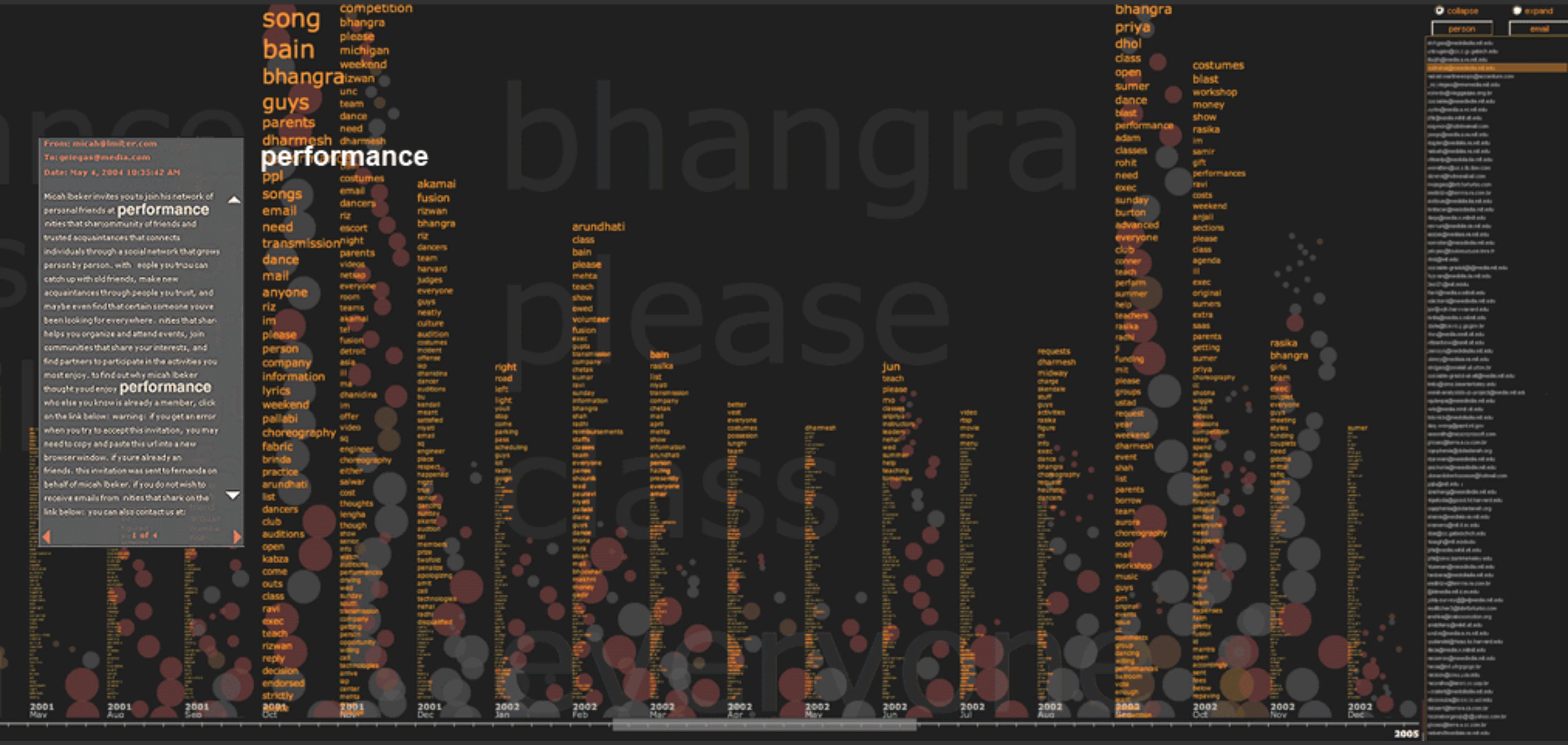
author:
rukbat@pern.org
back to newsgroups
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Week of May 6, 2001
threads initiated by author
threads not initiated by author
subject
# of posts
TDMA vs. CDMA

Email Mountain [Viegas]

Conversation by person over time (who x when).

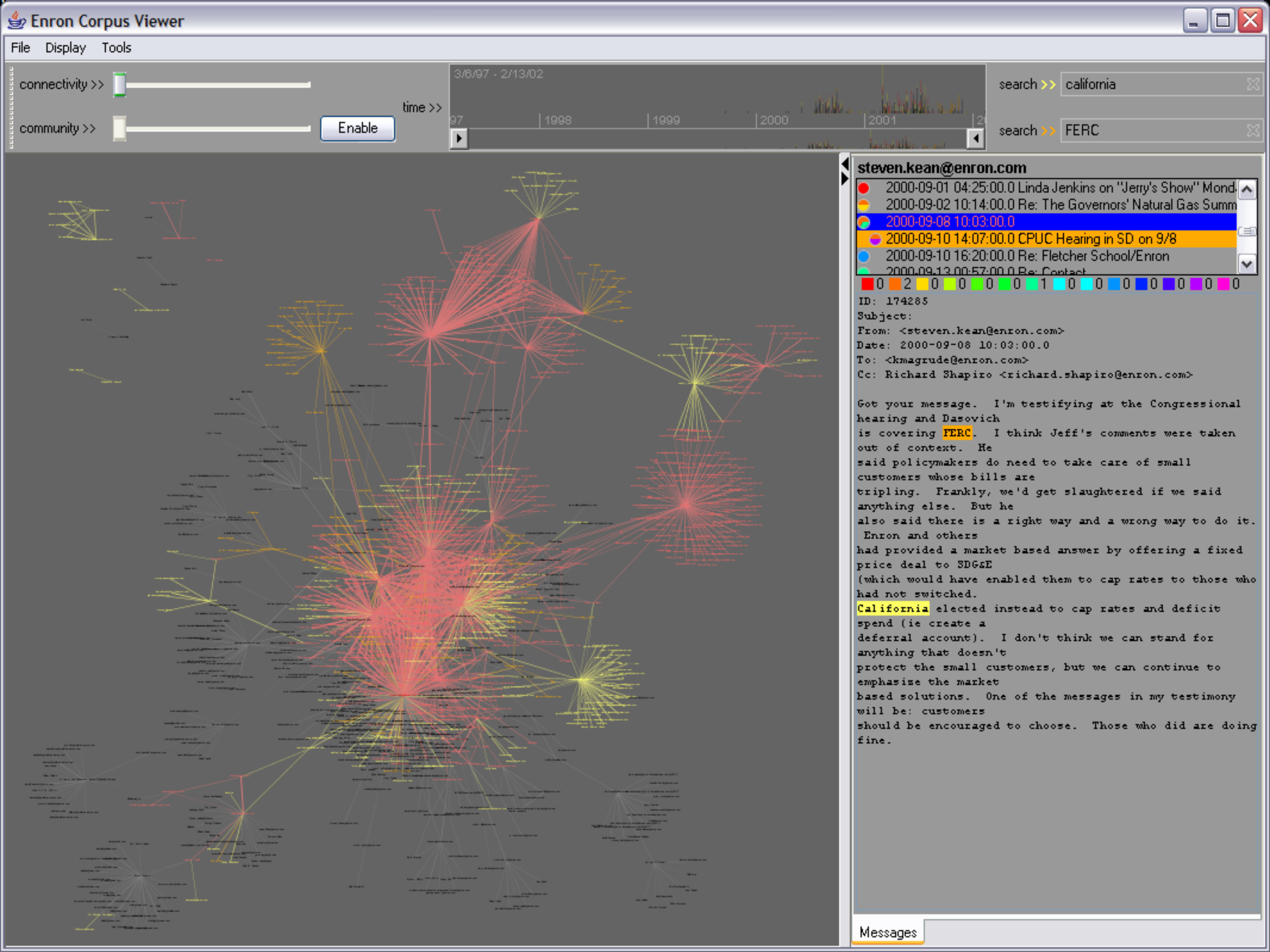


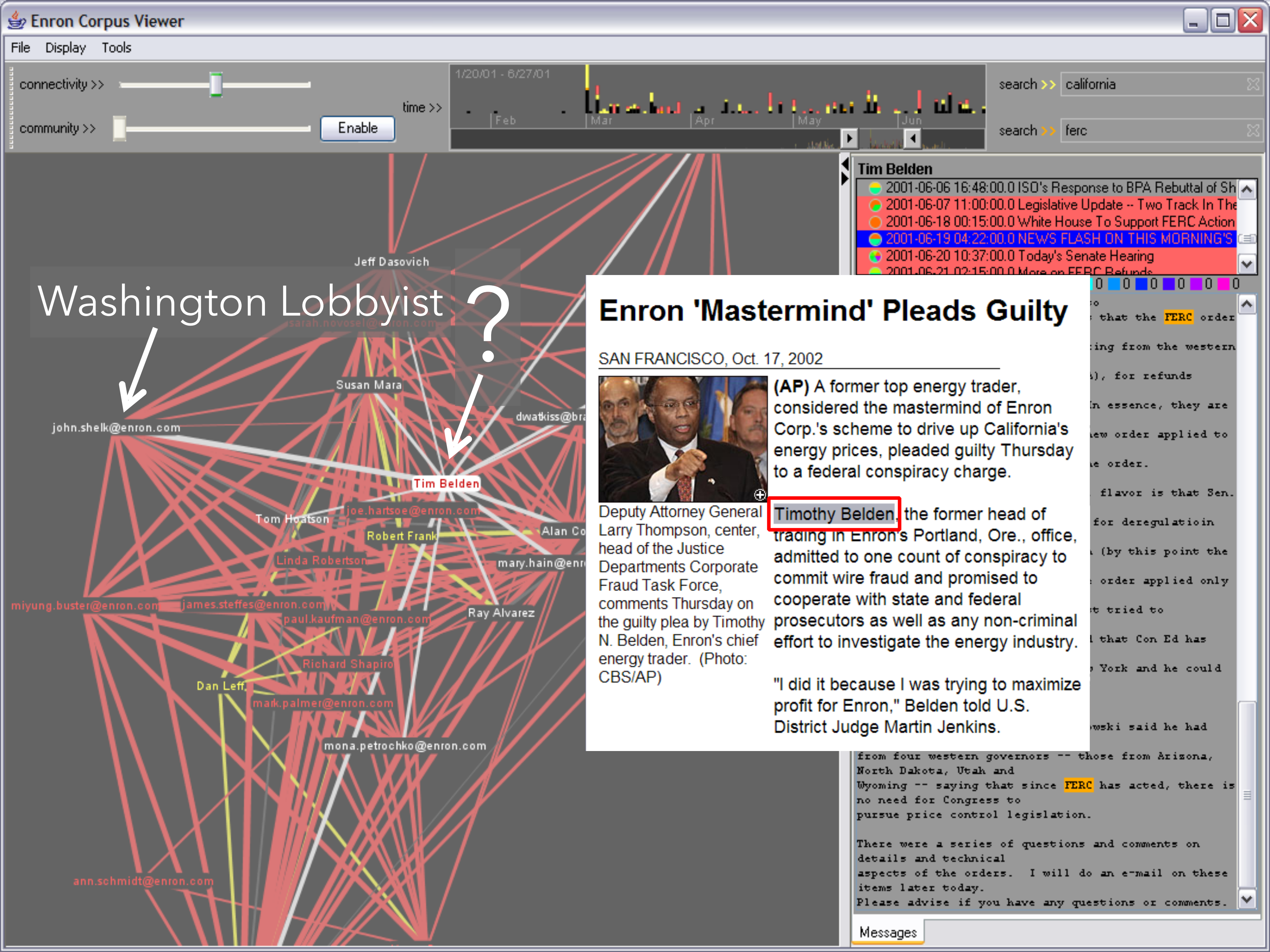
Themail [Viegas]



One person over time, TF.IDF weighted terms


Enron E-Mail Corpus
[Heer]


Enron Corpus Viewer
File
Display
Tools
connectivity >>
community >>
Enable
time >>
3/6/97 - 2/13/02
97
1998
1999
2000
2001
search >>
california
search >>
FERC
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
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



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.

Document Collections

Named Entity Recognition

Label named entities in text:

John Smith -> PERSON

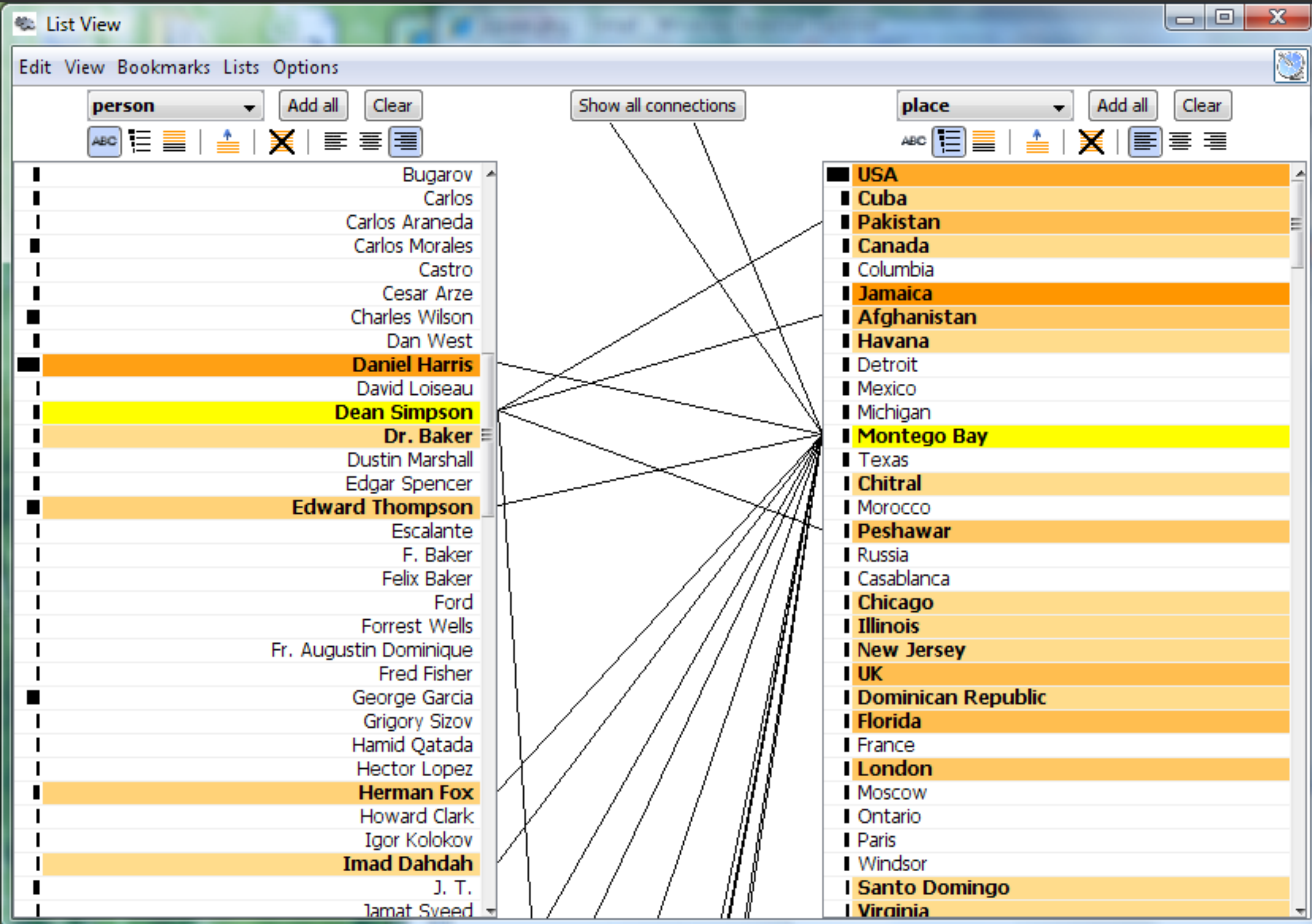
Soviet Union -> COUNTRY

353 Serra St -> ADDRESS

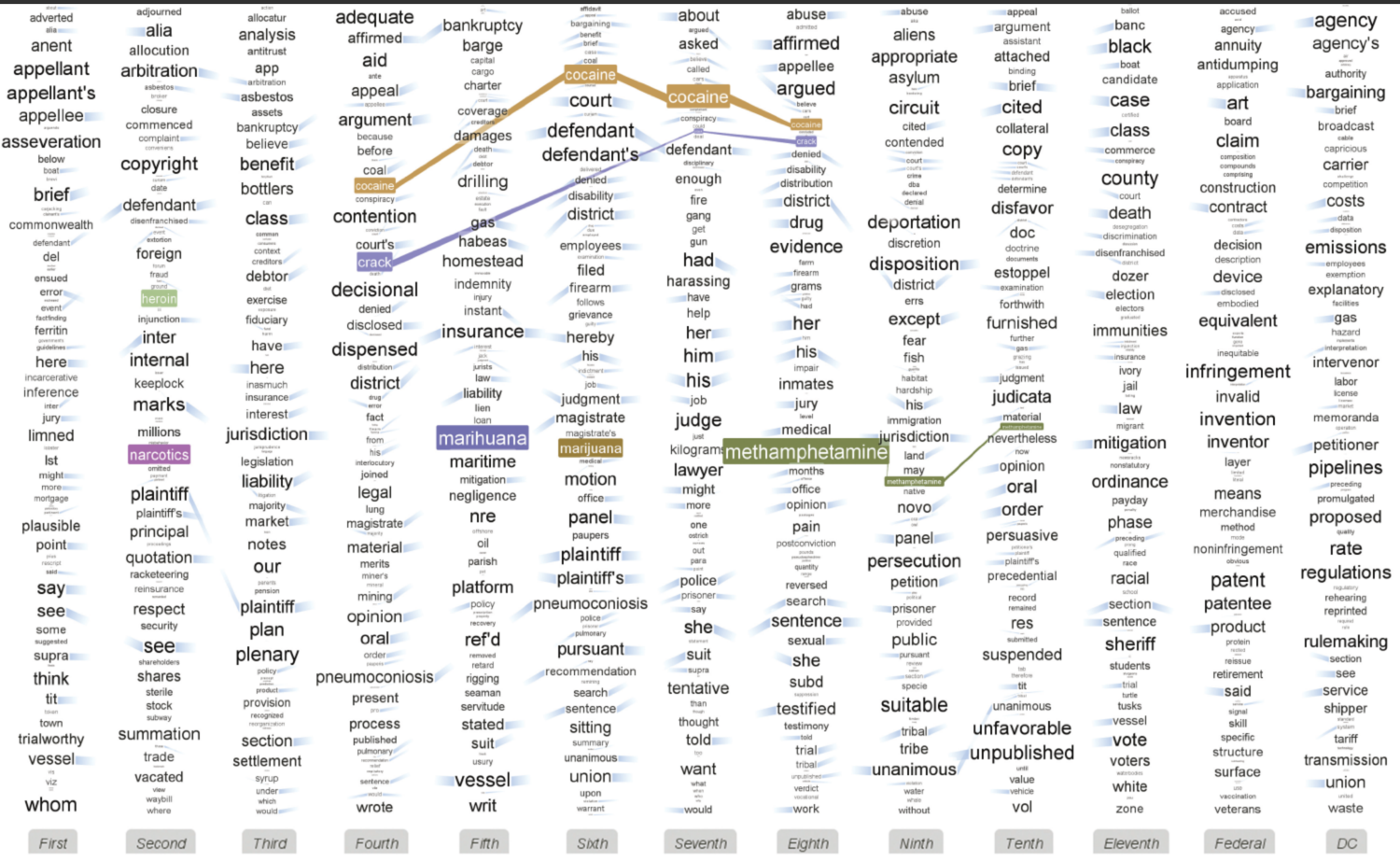
(555) 721-4312 -> PHONE NUMBER

Entity relations: how do the entities relate?

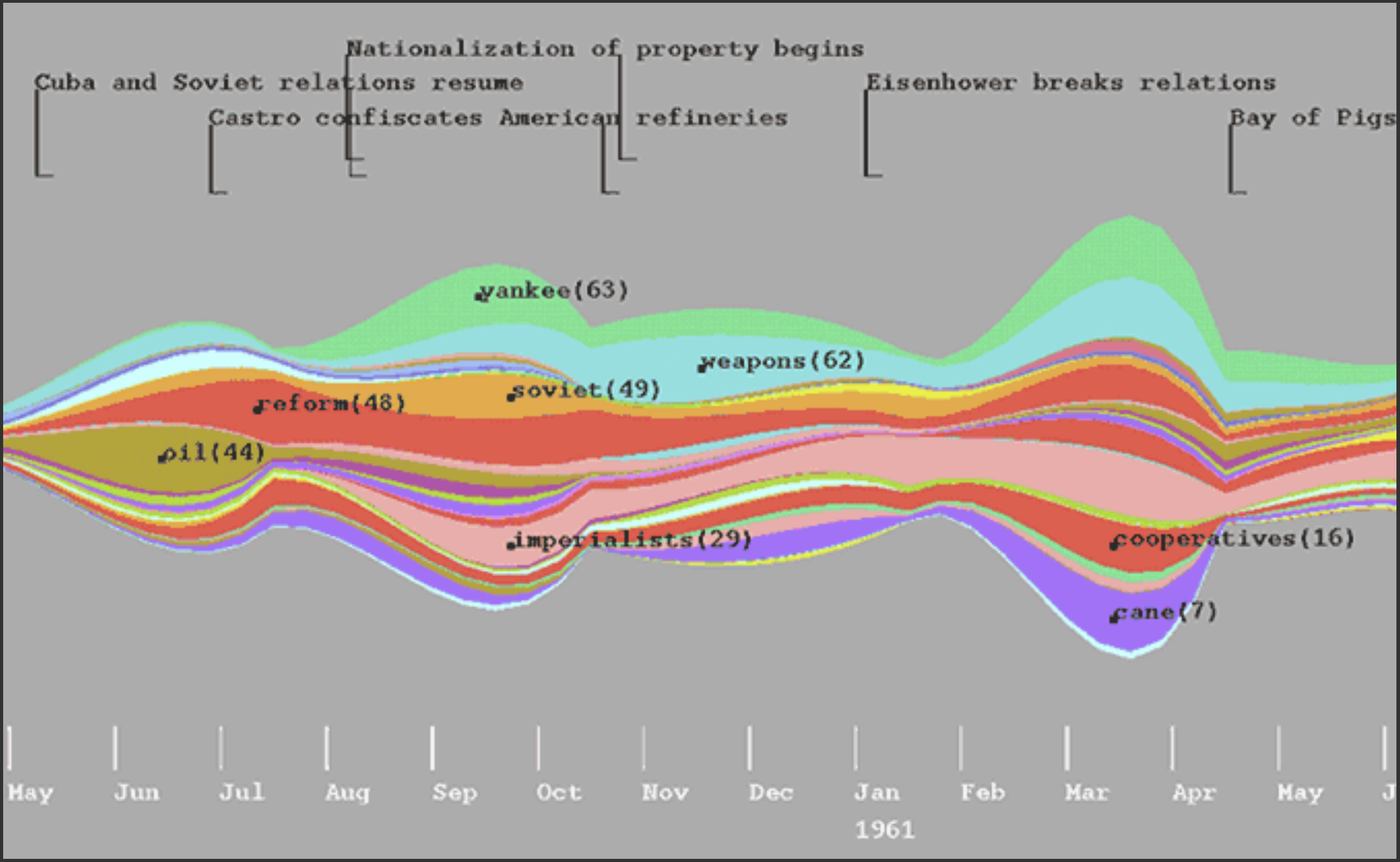
Simple approach: do they co-occur in a small window of text?


List View
Edit View Bookmarks Lists Options
person
Add all
Clear
Show all connections
place
Add all
Clear
Bugarov
Carlos
Carlos Araneda
Carlos Morales
Castro
Cesar Arze
Charles Wilson
Dan West
Daniel Harris
David Loiseau
Dean Simpson
Dr. Baker
Dustin Marshall
Edgar Spencer
Edward Thompson
Escalante
F. Baker
Felix Baker
Ford
Forrest Wells
Fr. Augustin Dominique
Fred Fisher
George Garcia
Grigory Sizov
Hamid Qatada
Hector Lopez
Herman Fox
Howard Clark
Igor Kolokov
Imad Dahdah
J. T.
Jamat Syeed
USA
Cuba
Pakistan
Canada
Columbia
Jamaica
Afghanistan
Havana
Detroit
Mexico
Michigan
Montego Bay
Texas
Chitral
Morocco
Peshawar
Russia
Casablanca
Chicago
Illinois
New Jersey
UK
Dominican Republic
Florida
France
London
Moscow
Ontario
Paris
Windsor
Santo Domingo
Virginia

Parallel Tag Clouds [Collins et al.]



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

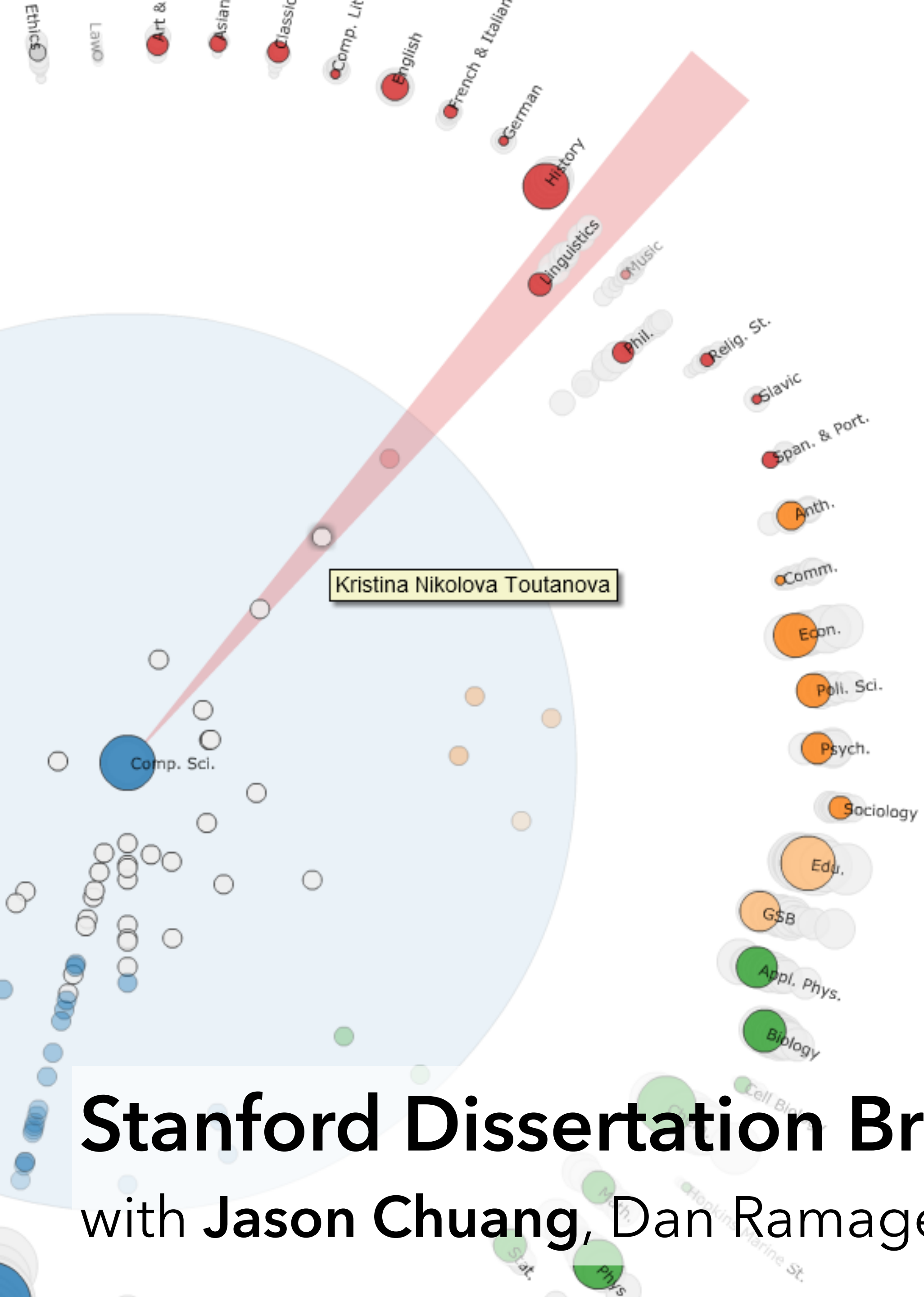
Topic modeling

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

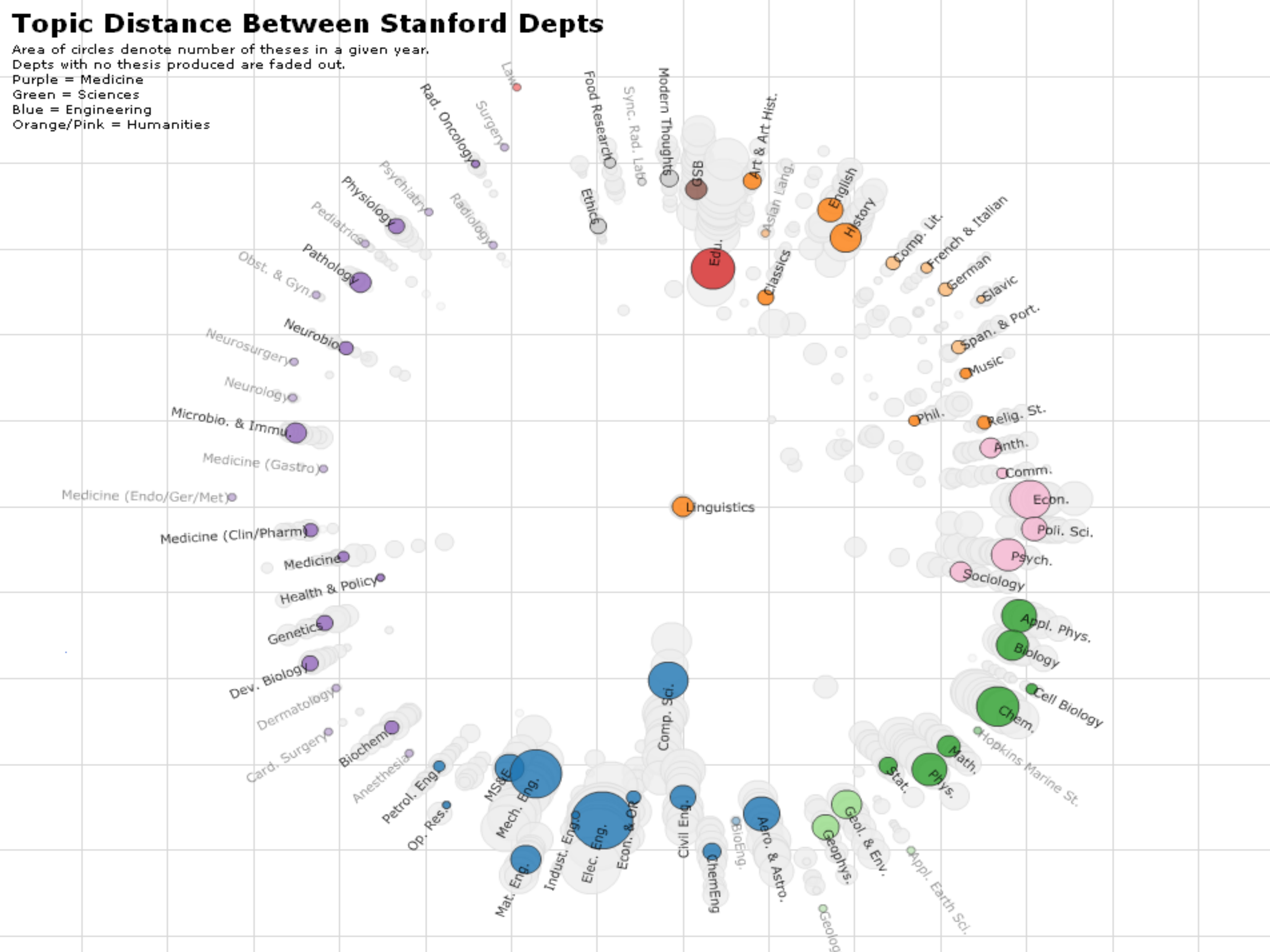
Abstract:

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.

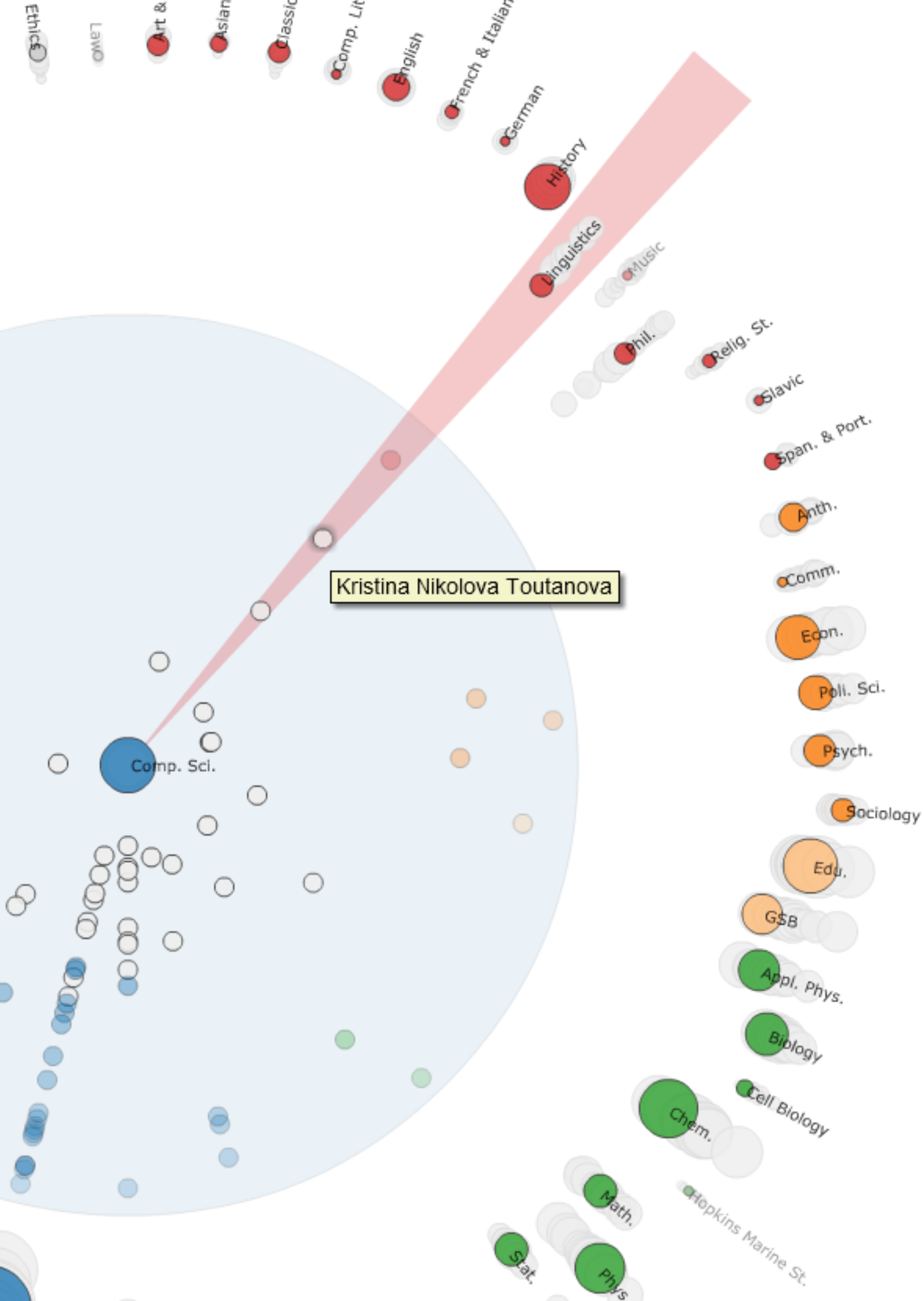
Stanford Dissertation Browser

with **Jason Chuang**, Dan Ramage & Christopher Manning

Law
Dermatology
Medicine (Gastro)
Obst. & Gyn.
Appl. Earth Sci.
Geol. & Env.
Petrol. Eng.
Geology
Food Research
Geophys.
Indust. Eng.
Math.
Stat.
Mat. Eng.
Econ. & OR
Op. Res.
Phys.
Chem. Eng.
Poli. Sci.
Elec. Eng.
Appl. Phys.
Chem.
Cell Biology
Medicine (Clin/Pharm)
Rad. Oncology
Sociology
Comp. Sci.
Anth.
German
Slavic
History
Edu.
Art & Art Hist.
Span. & Port.
Phil.
Biochem.
Physiology
Dev. Biology
Pediatrics
Relig. St.
Modern Thoughts
English
Comm.
Psych.
Genetics
Microbio. & Immu.
Medicine
Linguistics
Comp. Lit.
French & Italian
Pathology
Ethics
Music
Asian Lang.
Classics
Neurobio.
Card. Surgery
Oncology
Anesthesia
Biology
Medicine (Gastro)
Linguistics
French & Italian
Music
Anth.
Phil.
History
Bio. Eng.
Appl. Earth Sci.
Sociology
Civil Eng.
Edu.
Comm.
Chem. Eng.
Radiology
Comp. Sci.
Poli. Sci.
Mech. Eng.
Food Research
Aero. & Astro.
Indust. Eng.
Appl. Phys.
Stat.
MS&E
GSB
Geol. & Env.
Econ.
Geophys.
Petrol. Eng.
Dev. Biology
Medicine (Clin/Pharm)
Microbio. & Immu.
Genetics
Physiology
Neurobio.
Biochem.
Medicine
Linguistics
Stat.
Biology
Chem.
Chem. Eng
Comm.
Phil.
Psych.
Phys.
Appl. Phys.
Mat. Eng.
Ethics
Art & Art Hist.
Sociology
Edu.
Anth.
Mech. Eng.
Civil Eng.
Geol. & Env.
Econ. & OR
Aero. & Astro.
Span. & Port.
Modern Thoughts
History
Comp. Sci.
Geophys.
English
Poli. Sci.
Comp. Lit.
Petrol. Eng.
French & Italian
Math.
Indust. Eng.
German
Geology
Classics
Relig. St.
Op. Res.
Asian Lang.
Petrol. Eng.
Appl. Earth Sci.
Geology
Geophys.
Geol. & Env.
Op. Res.
Stat.
Mat. Eng.
Appl. Phys.
Aero. & Astro.
Indust. Eng.
Econ. & OR
Food Research
Mech. Eng.
Elec. Eng.
Comp. Sci.
GSB
Math.
Civil Eng.
Econ.
Chem.
Psychiatry
Sociology
Genetics
Physiology
Psych.
Biochem.
Medicine
Biology
Comm.
Poli. Sci.
Phil.
Edu.
History
Anth.
Linguistics
Microbio. & Immu.
Dev. Biology
Modern Thoughts
Classics
Neurobio.
French & Italian
Ethics
Asian Lang.
Art & Art Hist.
Relig. St.
Slavic


Topic Distance Between Stanford Depts
Area of circles denote number of theses in a given year.
Depts with no thesis produced are faded out.
Purple = Medicine
Green = Sciences
Blue = Engineering
Orange/Pink = Humanities
Law
Surgery
Rad. Oncology
Food Research
Sync. Rad. Lab
Modern Thought
GSB
Art & Art Hist.
Asian Lang.
English
History
Comp. Lit.
French & Italian
German
Slavic
Span. & Port.
Music
Phil.
Relig. St.
Anth.
Comm.
Econ.
Poli. Sci.
Psych.
Sociology
Appl. Phys.
Biology
Cell Biology
Chem.
Hopkins Marine St.
Math.
Stat.
Phys.
Geol. & Env.
Geophys.
Appl. Earth Sci.
Geology
Aero. & Astro.
BioEng.
ChemEng
Civil Eng.
Comp. Sci.
Econ. & OR
Elec. Eng.
Indust. Eng.
Mat. Eng.
Mech. Eng.
MS&E
Op. Res.
Petrol. Eng
Anesthesia
Biochem
Card. Surgery
Dermatology
Dev. Biology
Genetics
Health & Policy
Medicine
Medicine (Clin/Pharm)
Medicine (Endo/Ger/Met)
Medicine (Gastro)
Microbio. & Immu.
Neurology
Neurosurgery
Neurobio.
Obst. & Gyn.
Pathology
Pediatrics
Physiology
Psychiatry
Radiology
Ethics
Edu.
Classics
Linguistics

Oh, the humanities!



Effective statistical models for syntactic and semantic disambiguation

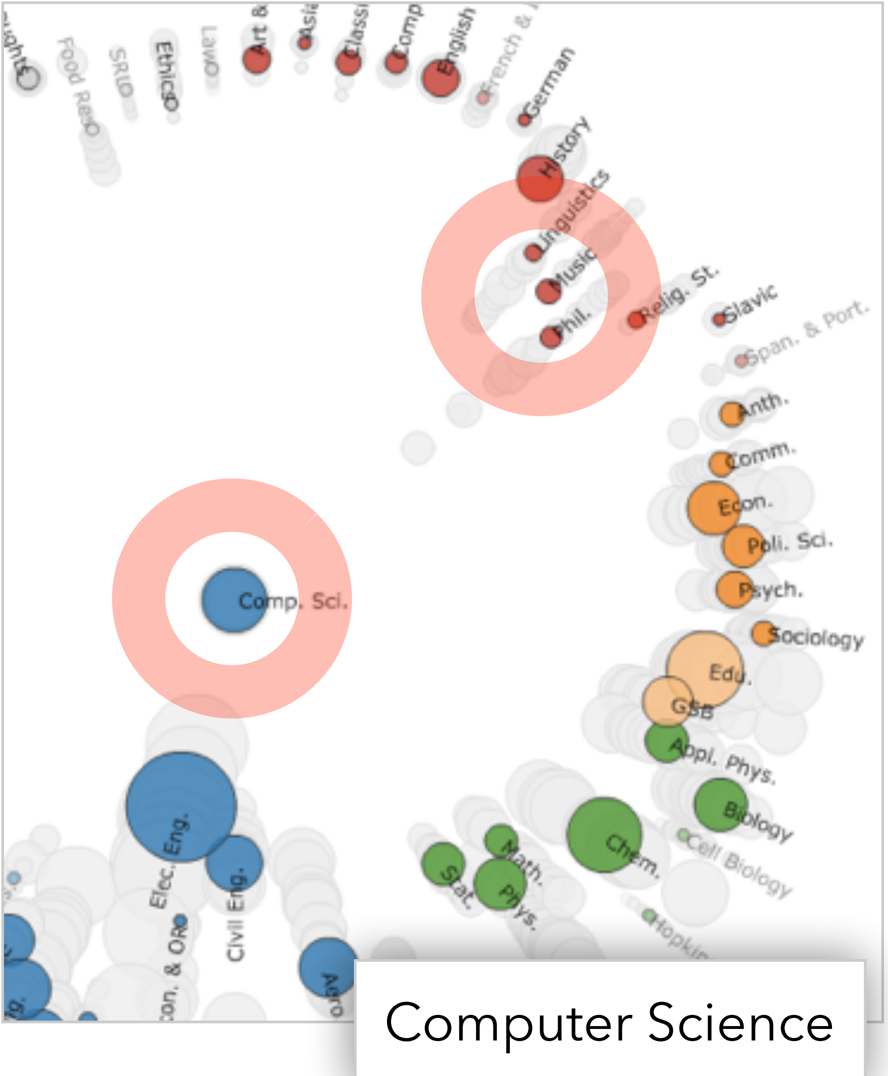
Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

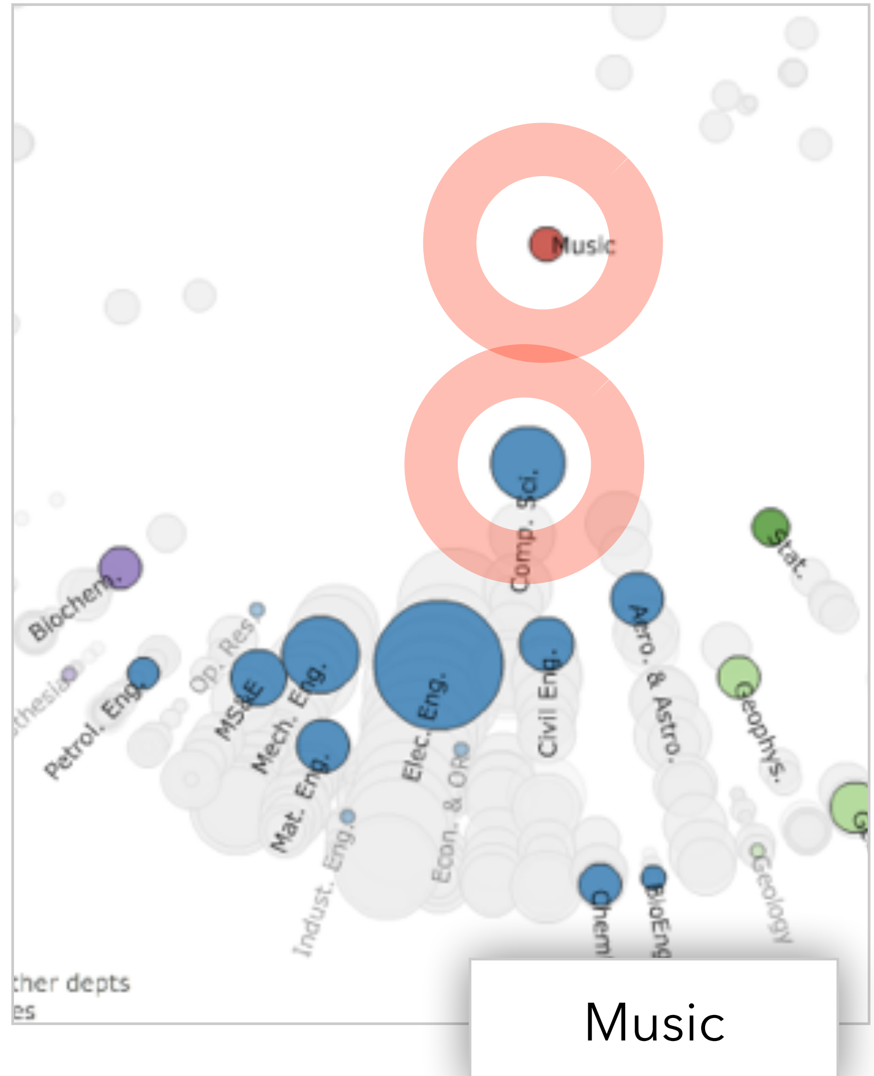
Computer Science (2005)

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.





"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*