Module #9d:
Visualization Text & Conversations

Visualization Techniques

1. Text as Data
2. Visualizing Document Content
3. Evolving Documents
4. Visualizing Conversation
5. Document Collections

4. CONVERSATIONS

Visualizing Conversation

Many dimensions to consider:

Who (senders, receivers)

What (the content of communication) When (temporal patterns)

Interesting cross-products:

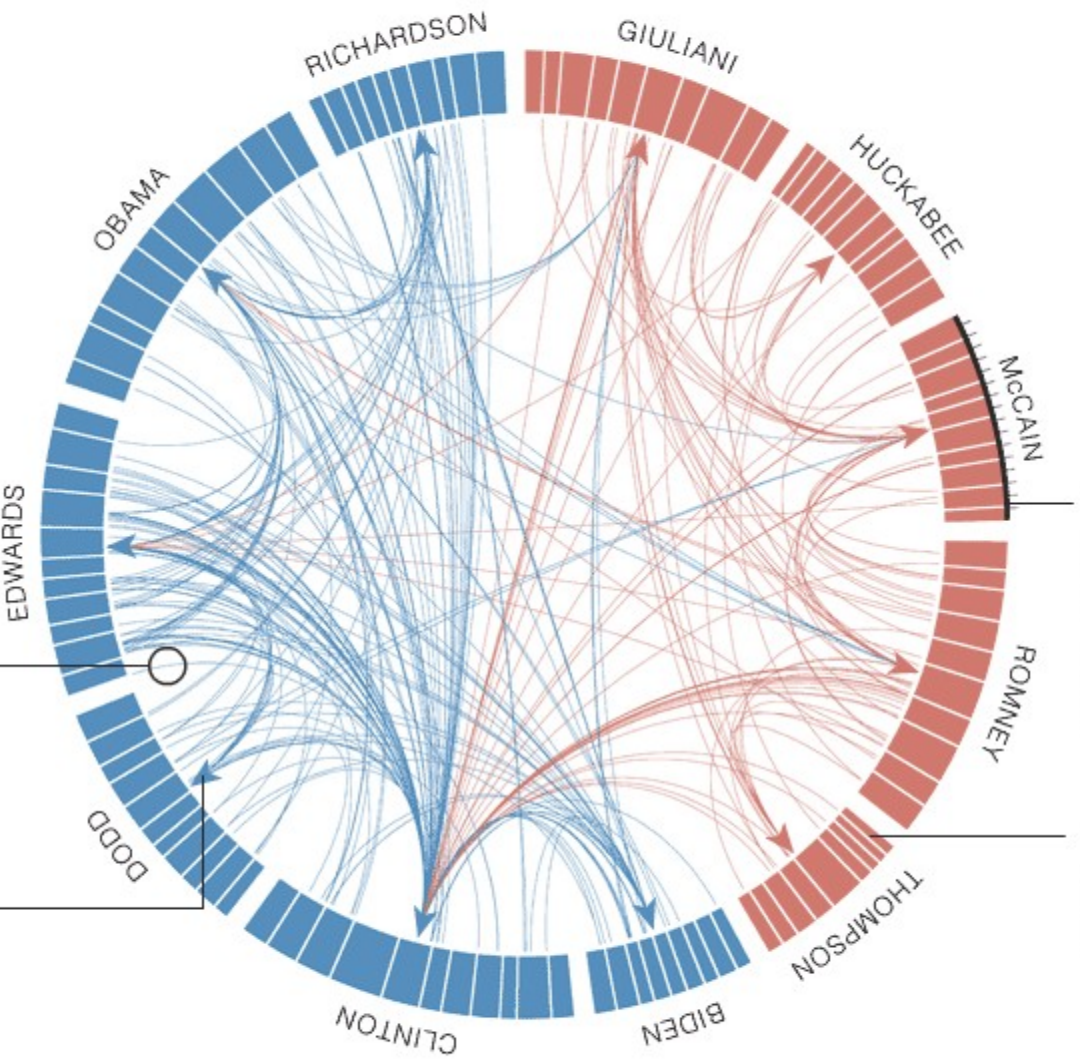
What x When -> Topic “Zeitgeist” Who x Who -> Social network

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.

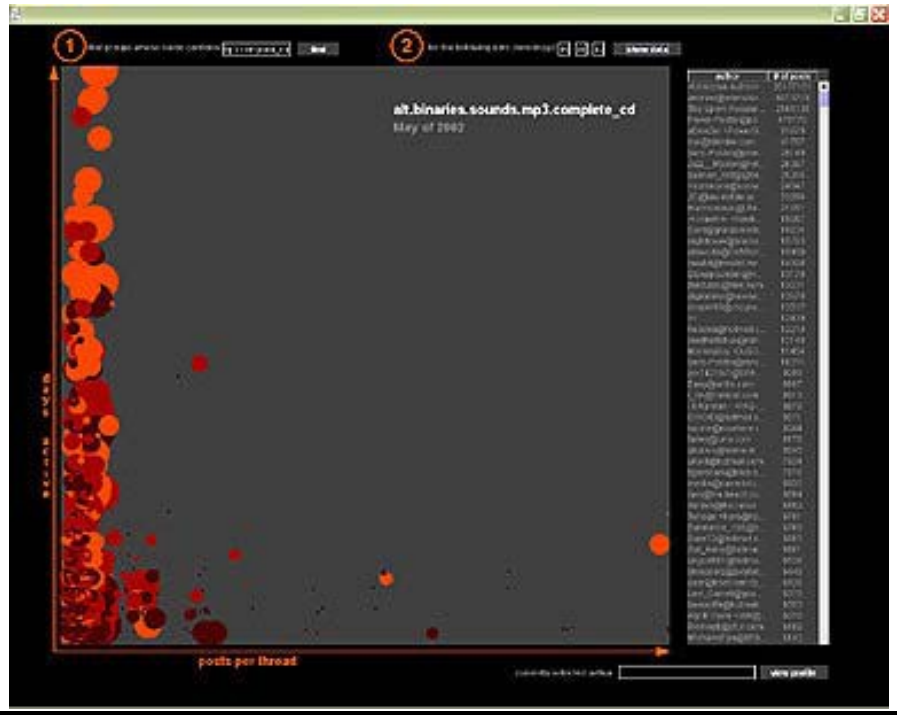
Each slice represents one debate, arranged clockwise from the first to the final debate.

Source: Debate transcripts

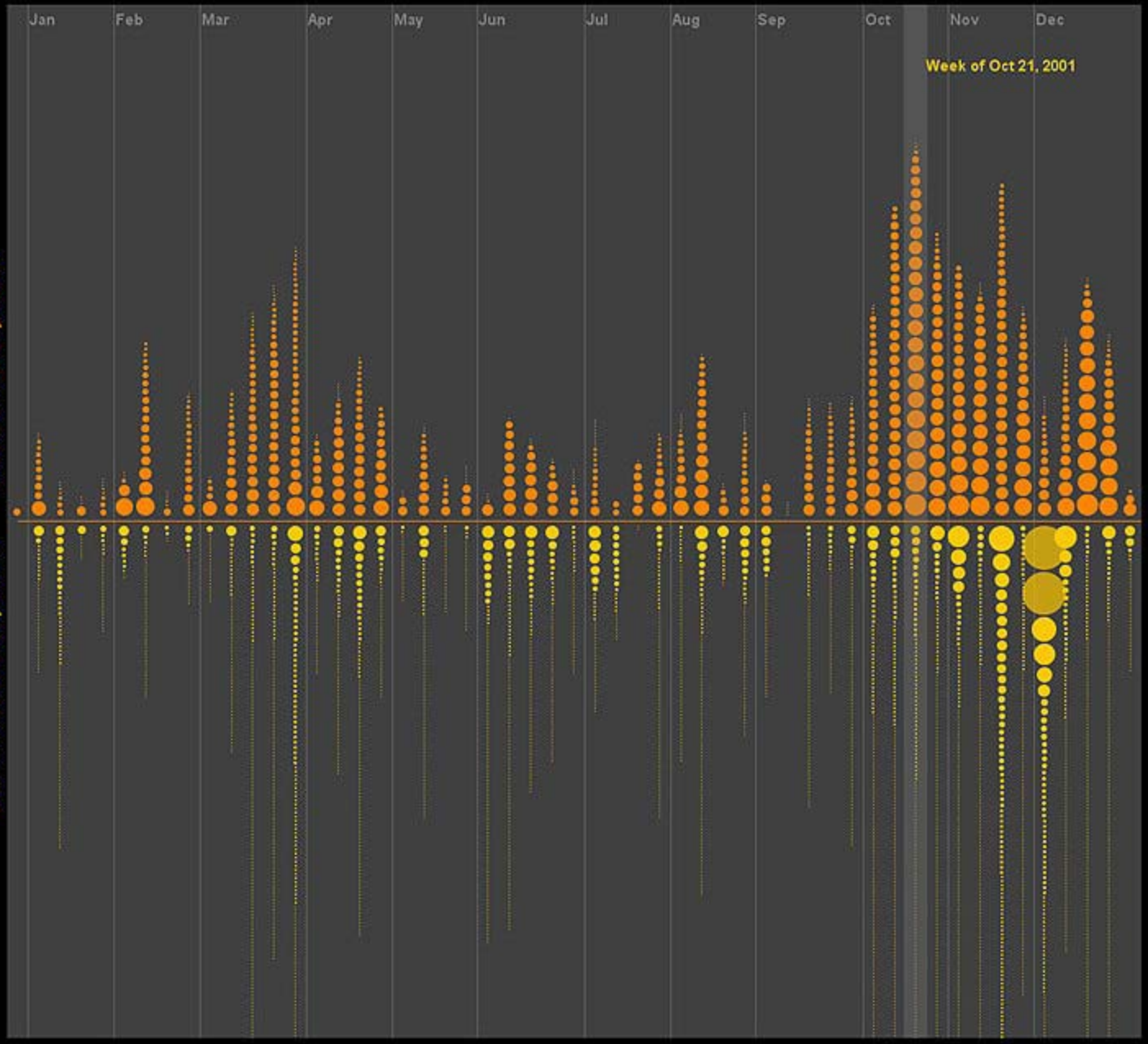
Jonathan Corum and Farhana Hossain/The New York Times

Usenet Visualization [Viegas & Smith]

Show correspondence patterns in text forums Initiate vs. reply; size and duration of discussion



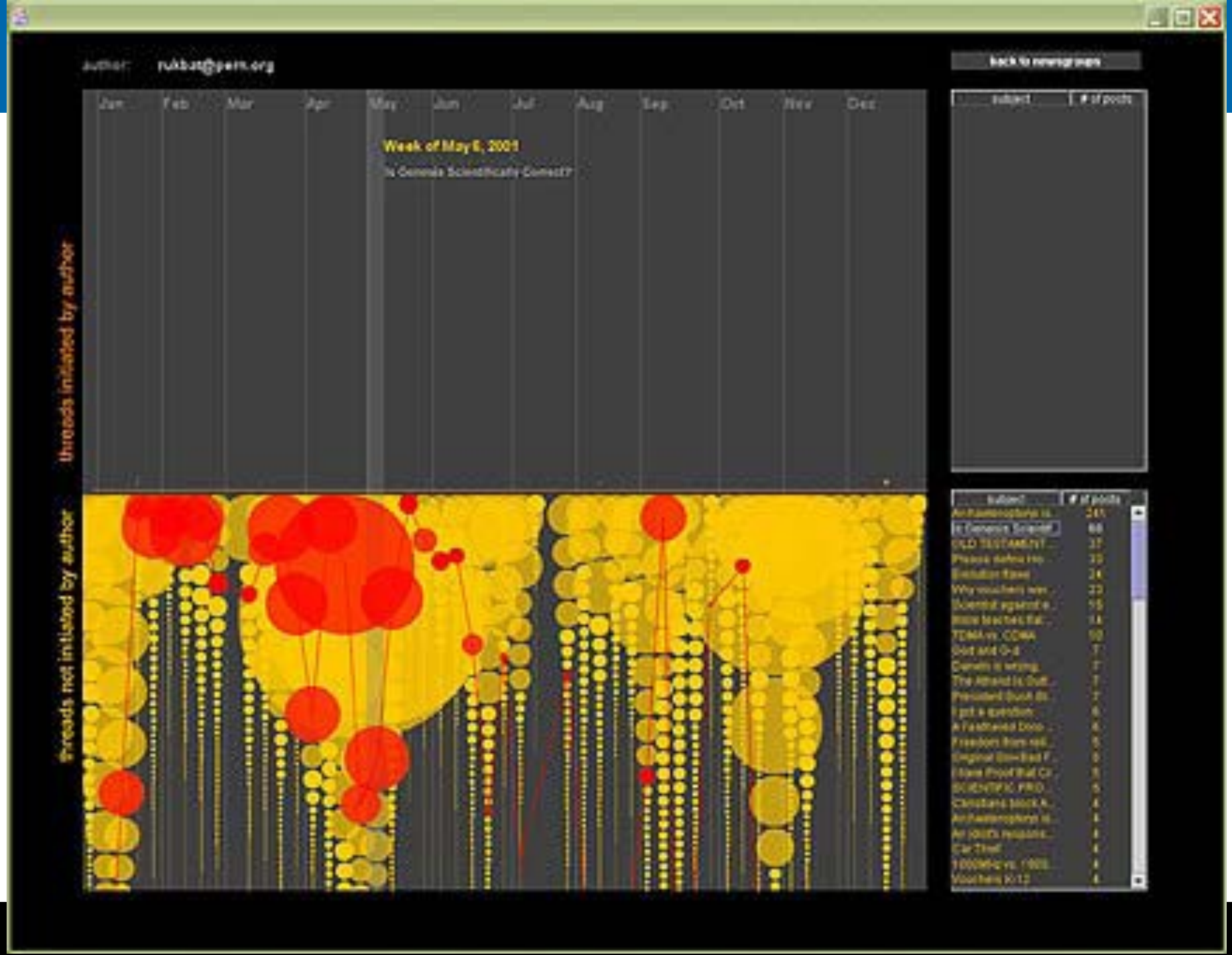
author: jillyb@mail.com



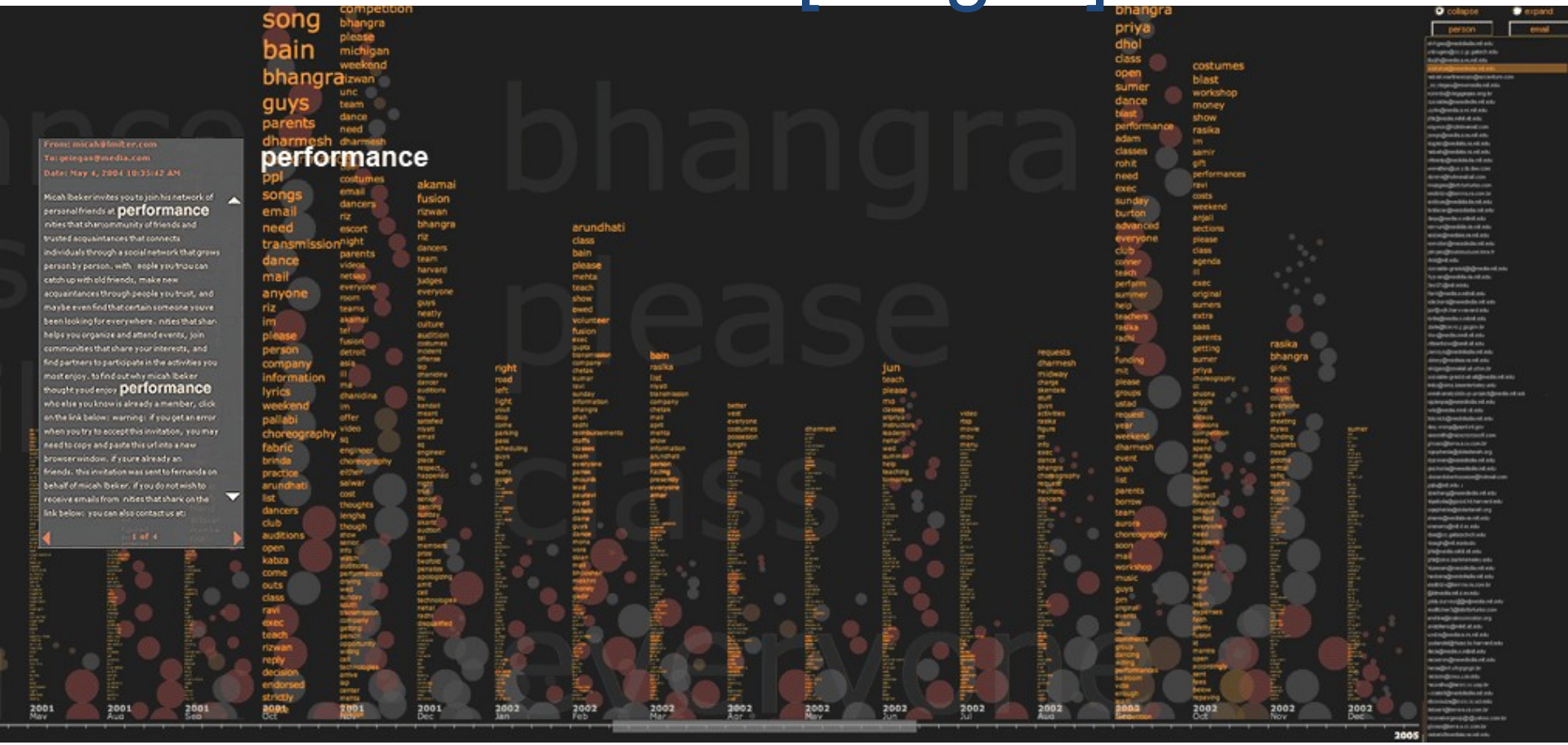
back to newsgroups

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

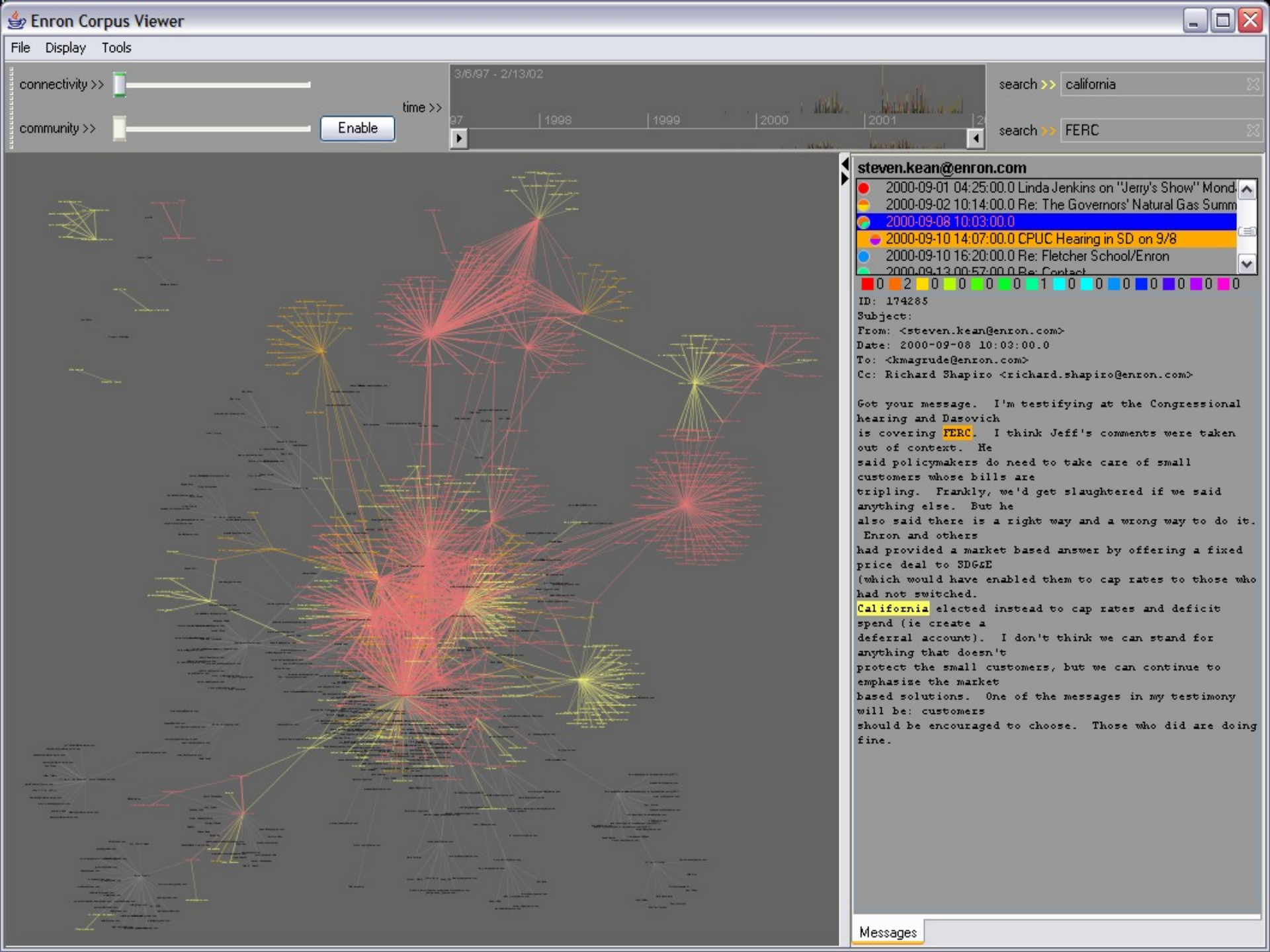
subject	# of posts
Sunday Twofer ASF :)	9
Chopsticks/A Jilly fake	8
Oh no! Trouble in Discworld! ...	7
WET.....your thirst ! ASF ...	6
A pretty for you...Reposted fro...	5
Saturday Spooker ASF	5
Sample Previous install Upgr...	4
Tennessee weather tonite	4
WET - Well I am not smiling!	4
Somethin' mushy <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
Is Gemma Scientifically Correct?
subject # of posts
threads initiated by author
threads not initiated by author

Themail [Viegas]



One person over time, TF.IDF weighted terms


Enron Corpus Viewer
File
Display
Tools
connectivity >>
community >>
Enable
time >>
3/6/97 - 2/13/02
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 0 1 0 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 Corpus Viewer

File Display Tools

connectivity >>

community >>

Enable

time >>

1/20/01 - 6/27/01

Feb Mar Apr May Jun

search >> california

search >> ferc

Washington Lobbyist

?

Jeff Dasovich

Susan Mara

Tim Belden

Tom Hoatson

joe.hartsoe@enron.com

Robert Frank

Linda Robertson

mary.hain@enr

miyung.buster@enron.com

james.steffes@enron.com

paul.kaufman@enron.com

Ray Alvarez

Richard Shapiro

Dan Leff

mark.palmer@enron.com

mona.petrochko@enron.com

ann.schmidt@enron.com

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.

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

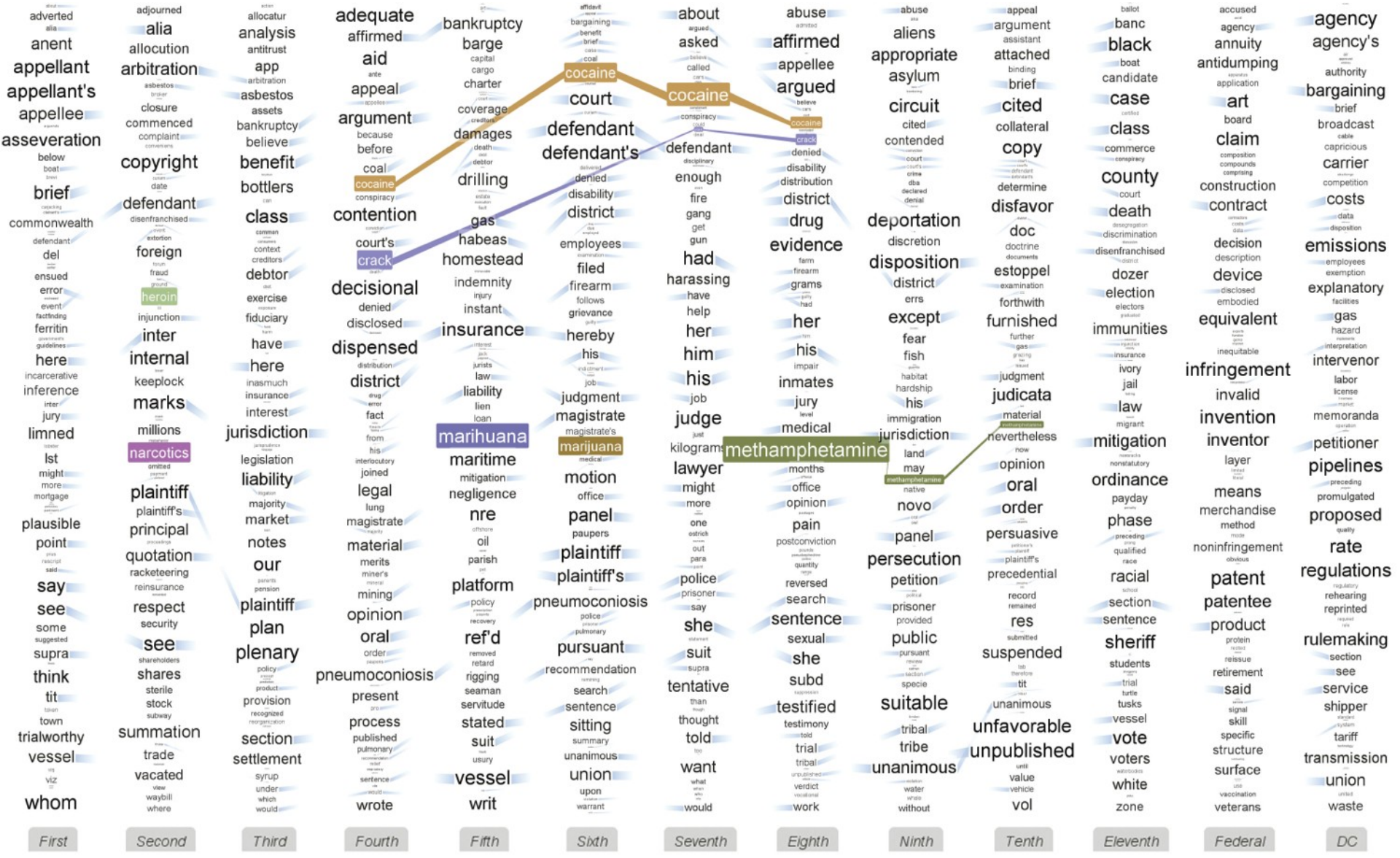
from four western governors -- those from Arizona, North Dakota, Utah and Wyoming -- saying that since FERC has acted, there is no need for Congress to pursue price control legislation.

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

5. DOCUMENT COLLECTIONS

Parallel Tag Clouds [Collins et al.]

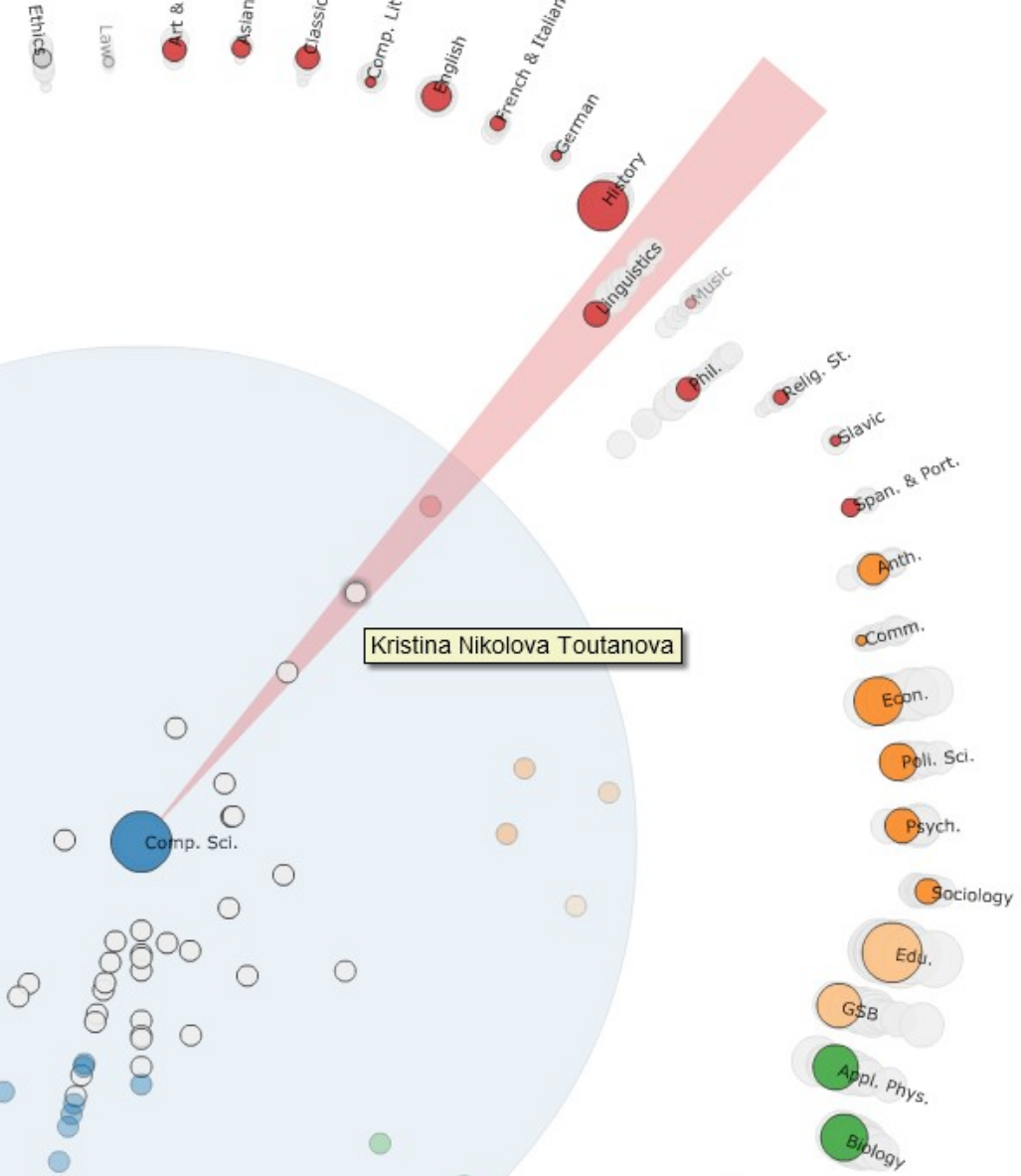


Similarity & Clustering

- 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)

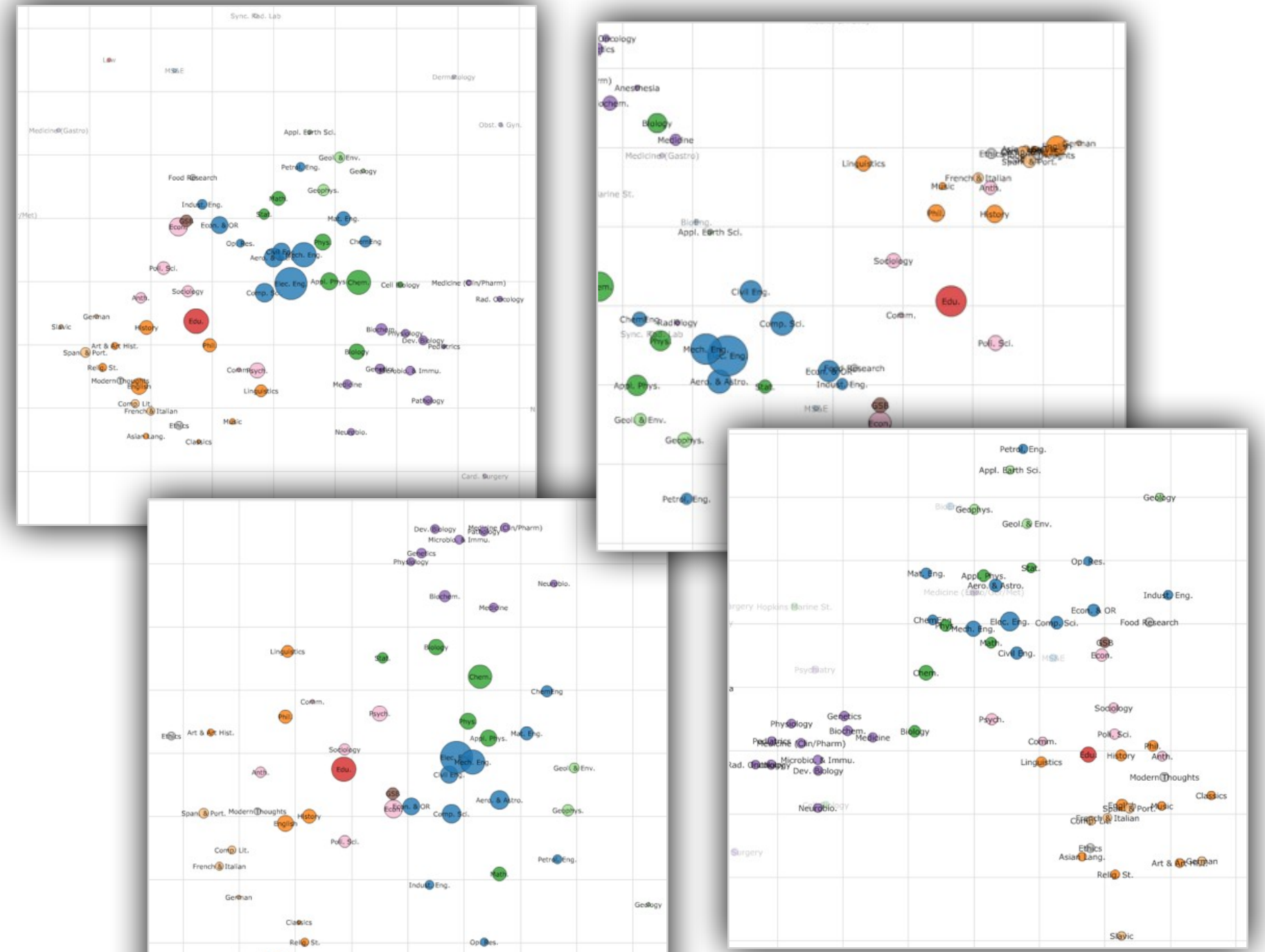
Keywords: Syntactic, Semantic, Tree kernels, Parsing

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.

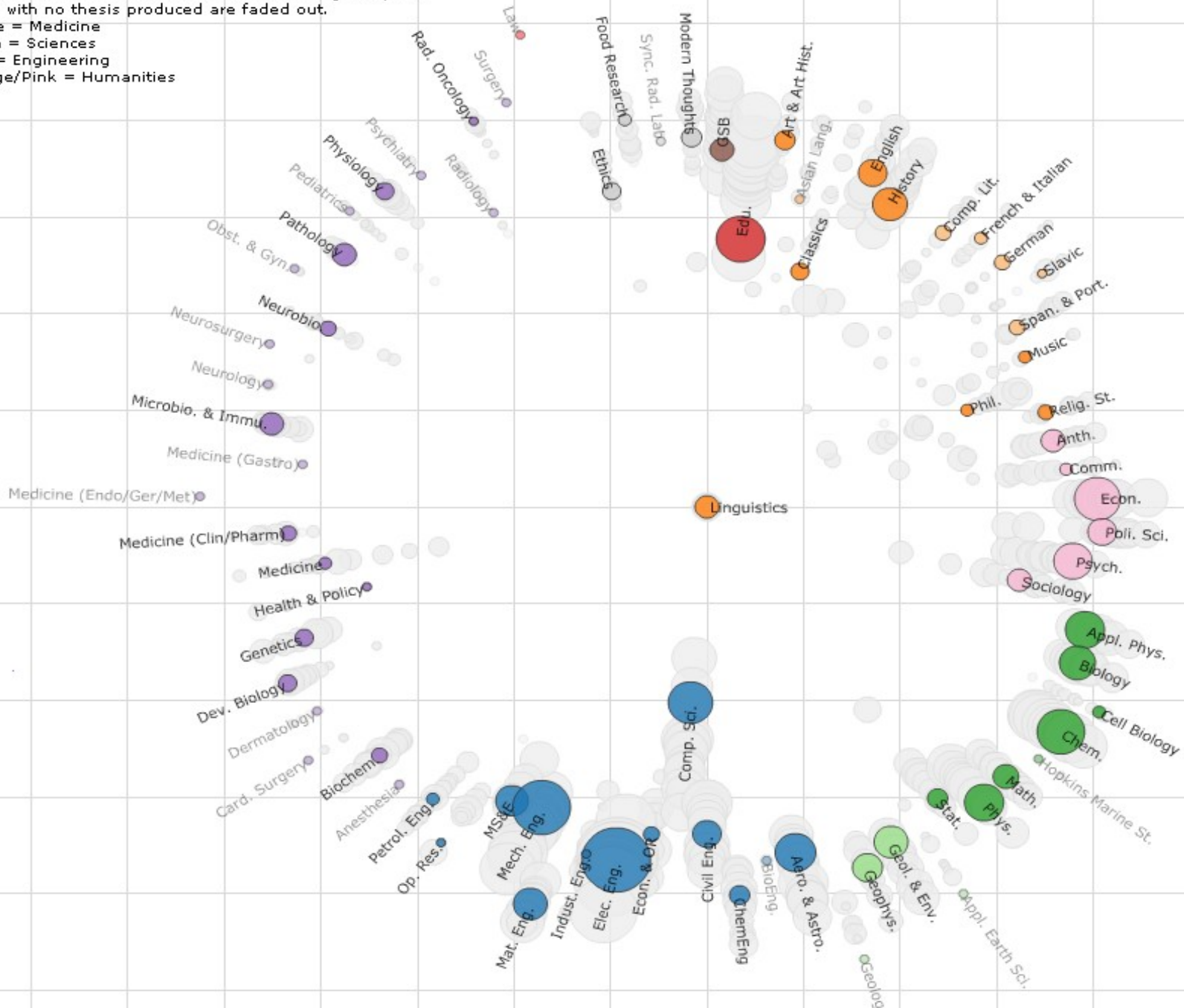
Stanford Dissertation Browser

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

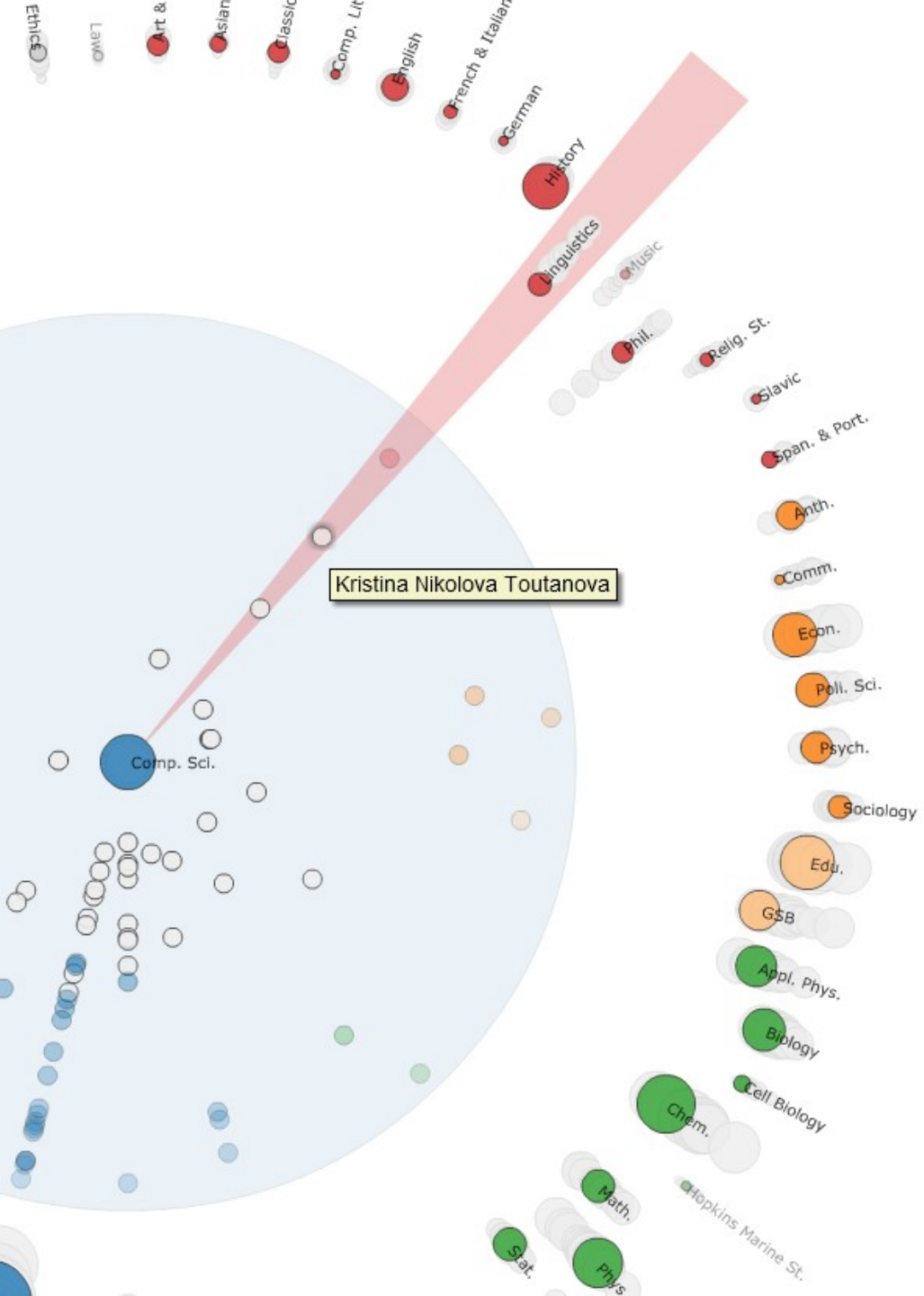

Sync. Rad. Lab
Dermatology
Medicine (Gastro)
Obst. & Gyn.
Appl. Earth Sci.
Geol. & Env.
Petrol. Eng.
Geology
Food Research
Geophys.
Math.
Stat.
Indust. Eng.
Econ.
Econ. & OR
Mat. Eng.
Optics
Phys.
Chem. Eng.
Aero. & Astro.
Civil Eng.
Mech. Eng.
Pol. Sci.
Elec. Eng.
Appl. Phys.
Chem.
Cell Biology
Medicine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

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



Oh, the humanities!



Effective statistical models for syntactic and semantic disambiguation

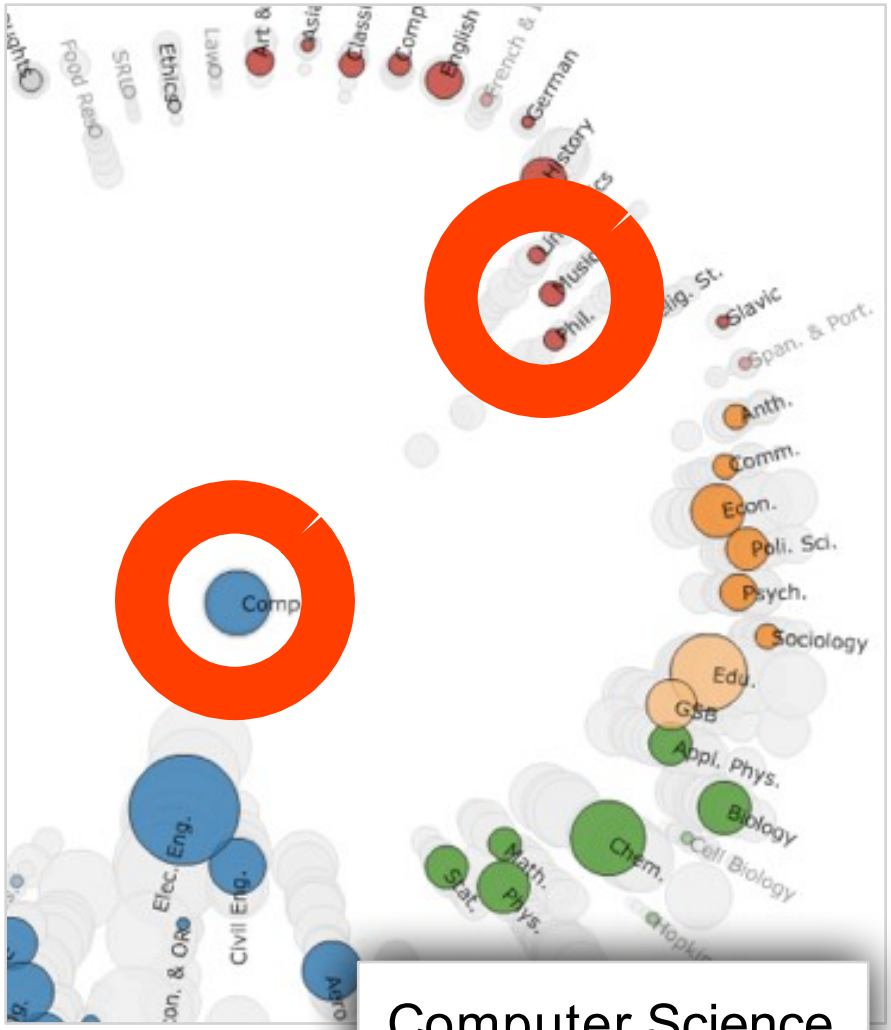
Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

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.



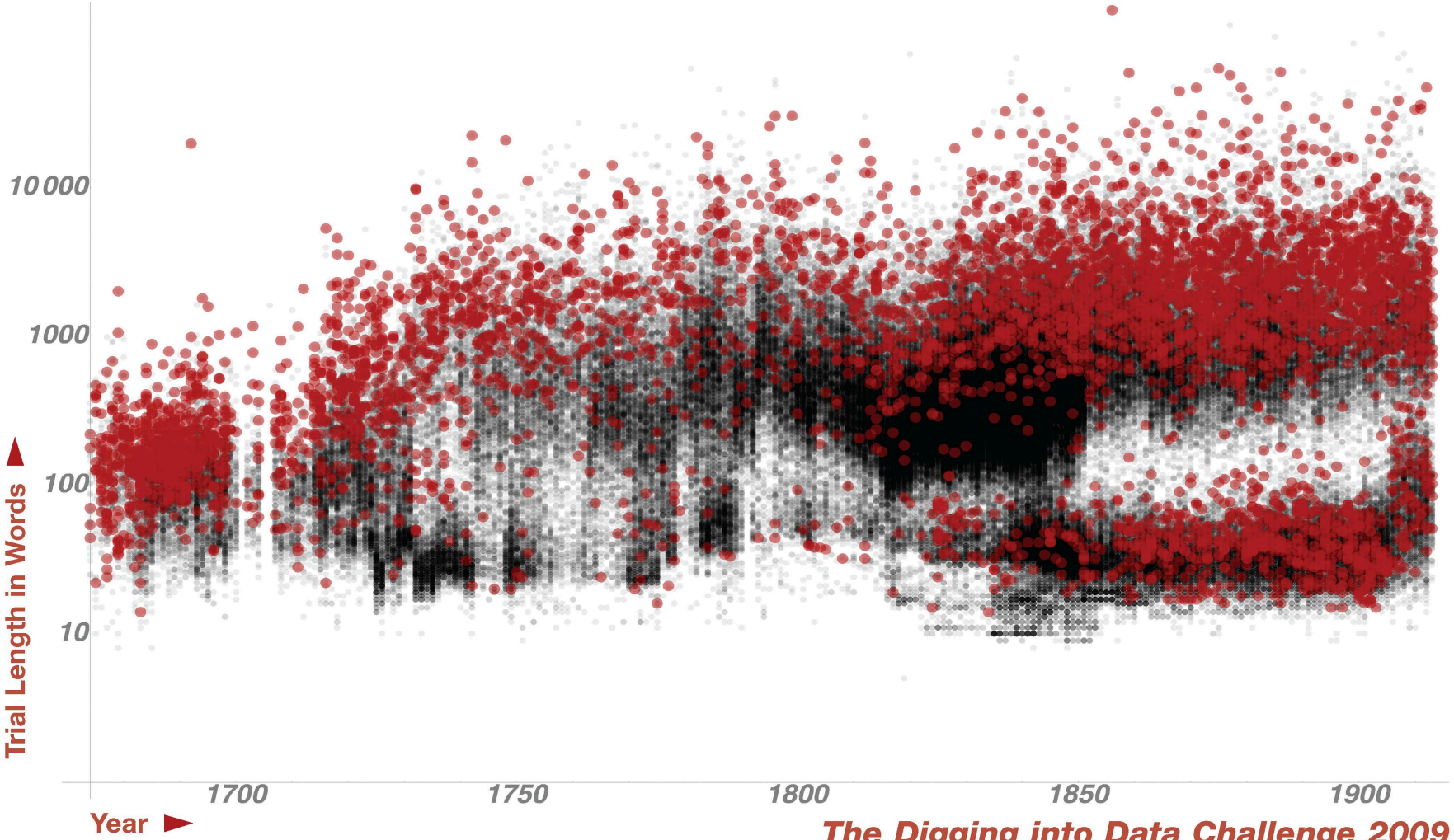
Computer Science

Music

“Word Borrowing” via Labeled LDA

Data Mining With Criminal Intent

Cyril Briquet • Dan Cohen • Frederick Gibbs • Tim Hitchcock • Jamie McLaughlin • Geoffrey Rockwell
Joerg Sander • Robert Shoemaker • John Simpson • Stéfan Sinclair • Sean Takats • William J. Turkel

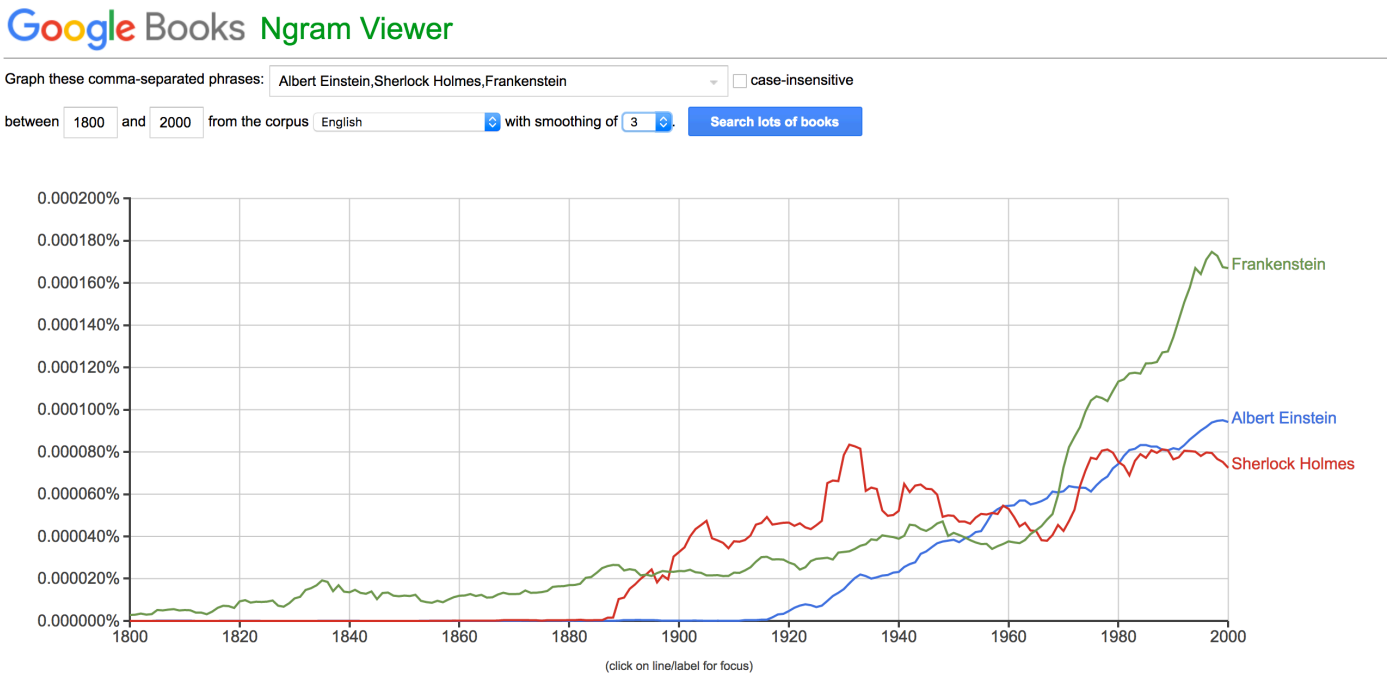


Length in words of trials involving 'killing' (red) versus all other trials (black) from the Old Bailey Proceedings, plotted with a logarithmic scale for the x axis. 197,745 Trials.

The Digging into Data Challenge 2009

NEH • JISC • SSHRC

- google


Google Books Ngram Viewer
Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein
case-insensitive
between 1800 and 2000 from the corpus English with smoothing of 3
Search lots of books
0.000200%
0.000180%
0.000160%
0.000140%
0.000120%
0.000100%
0.000080%
0.000060%
0.000040%
0.000020%
0.000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
Frankenstein
Albert Einstein
Sherlock Holmes
(click on line/label for focus)

Conclusion

- Discuss challenges of visualizing unstructured data and large collection of text and documents
- Explain different techniques for representing words and concepts in a document
- Talk about the benefits and limitations of word clouds
- Discuss
    1. Text as Data
    2. Visualizing Document Content
    3. Evolving Documents
    4. Visualizing Conversation
    5. Document Collections