

# VISUALIZING TEXT

Petra Isenberg  
Anastasia Bezerianos

# HOUSEKEEPING

- TODAY: Your progress
- Mid-Feb: Assignment - website with your progress and peer-reviewing (check slack for date / instructions)
- Final presentations **25/02** (data providers will also be invited)

# RECAP

## STRUCTURED DATA



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

## UNSTRUCTURED DATA



(TODAY)

# VISUALIZING TEXT

incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

**TEXT IS DIFFERENT**

**COMMON**

**UNSTRUCTURED (MOSTLY)**

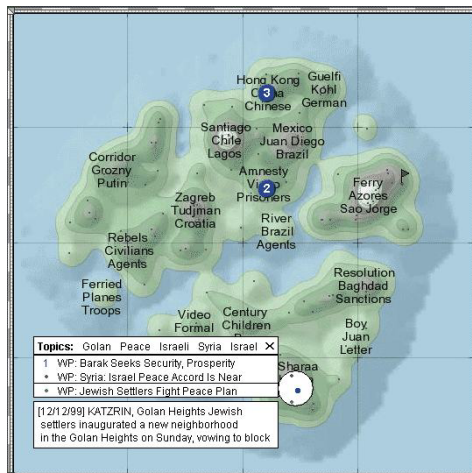
**HIGH-DIMENSIONAL (10,000+)**

**BIG!**

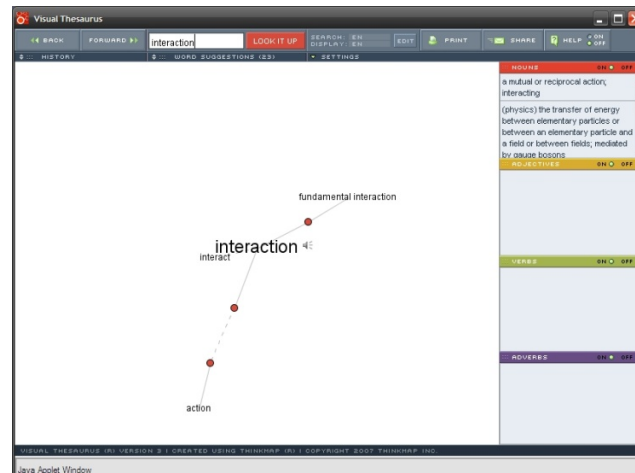
# WHY VISUALIZE TEXT?

# WHY

- To assist information retrieval
- To enable linguistic analysis
- To augment analytics on mixed data



Themescape



Visual Thesaurus



Thread Arcs

# WHY VISUALIZE TEXT? (TASKS)

UNDERSTANDING: GET THE “GIST” OF A DOCUMENT

GROUPING: CLUSTER FOR OVERVIEW OR CLASSIFICATION

COMPARE: COMPARE DOCUMENT COLLECTIONS, OR  
INSPECT EVOLUTION OF COLLECTION OVER TIME

CORRELATE: COMPARE PATTERNS IN TEXT TO THOSE IN  
OTHER DATA, E.G., CORRELATE WITH SOCIAL NETWORK



# WHAT IS TEXT DATA?

## DOCUMENTS

ARTICLES, BOOKS AND NOVELS  
COMPUTER PROGRAMS  
E-MAILS, WEB PAGES, BLOGS  
TAGS, COMMENTS

## COLLECTION OF DOCUMENTS

MESSAGES (E-MAIL, BLOGS, TAGS, COMMENTS)  
SOCIAL NETWORKS (PERSONAL PROFILES)  
ACADEMIC COLLABORATIONS (PUBLICATIONS)  
EVEN WHOLE LIBRARIES, WEBSITES, SOCIAL NETWORKS

# DIFFICULT DATA

## TOO MUCH DATA

- Millions of blog posts,
- Hundreds of thousands of news stories,
- 183 billion emails,
- ... per day

## NOISY DATA

- 70-72% of email is spam
- Text contains section headings, figure captions, and direct quotes
- ....

# ONCE YOU HAVE THE DATA...

Most meaning comes from our minds and common understanding.

“How much is that doggy in the window?”

- how much: social system of barter and trade (not the size of the dog)
- “doggy” implies childlike, plaintive, probably cannot do the purchasing on their own
- “in the window” implies behind a store window, not really inside a window, requires notion of window shopping

(Hearst, 2006)

# LANGUAGE IS AMBIGUOUS

- Words and phrases can have many meanings, determined by context and world knowledge.
- Interesting language is often figurative:
  - You are a couch potato.
  - They fought like cats and dogs.
  - Opportunity knocked on the door

# VISUAL CONSIDERATIONS

Supporters of Martin, who has been jailed without trial for more than two years, are calling on Prime Minister Stephen Harper to ask Mexican president Felipe Calderon to release Martin text is not preattentive under a section of the Mexican constitution that allows the government to expel undesirables from the country. Martin's supporters believe she has no chance of a fair trial in Mexico. Neither does Waage.

# VISUAL CONSIDERATIONS

Supporters of Martin, who has been jailed without trial for more than two years, are calling on Prime Minister Stephen Harper to ask Mexican president Felipe Calderon to release Martin **text is not preattentive** under a section of the Mexican constitution that allows the government to expel undesirables from the country. Martin's supporters believe she has no chance of a fair trial in Mexico. Neither does Waage.

# VISUAL CONSIDERATIONS



*Text readability is dependent on size, orientation, font, clutter...*

# VISUALIZING LANGUAGE IS ALSO EASY!

SO much data available for analysis

(Mostly) readily computer readable

Simple techniques can give instant summaries



# OUTLINE

TEXT AS DATA

VISUALIZING DOCUMENT CONTENT

EVOLVING DOCUMENTS

DOCUMENT COLLECTIONS

**TEXT AS DATA**

**Words are  
the basic  
unit of data.**

# WORD-LEVEL ATTRIBUTES

WORD LENGTH

PART OF SPEECH (NOUN, VERB, ADJECTIVE, ETC.)

FORMAT (*ITALIC*, UNDERLINE, ETC.)

LANGUAGE (ENGLISH? LATIN? JAPANESE?)

FREQUENCY / DIFFICULTY (IS IT COMMON?)

SENTIMENT (POSITIVE OR NEGATIVE CONNOTATION)

SYNONYMS / ANTONYMS / ETYMOLOGY (OTHER MEANINGS? ROOTS?)

ENTITIES (e.g. “Calgary”, “Obama”, “Telus” )

... AND MANY MORE

# AGGREGATION

- REPETITION,  
PLAGARISM,  
SHARED
- ENTITIES,
- AUTHOR STYLE
- 
- 
- 
- 

## COLLECTION

▲  
DOCUMENT

▲  
SECTION

▲  
PAGE

▲  
PARAGRAPH

▲  
SENTENCE

▲  
WORD

}] TENSE,  
SENTIMENT,  
SENTENCE  
LENGTH,  
READING LEVEL

# LINGUISTIC METHODS

- Word Counting
- Word Scoring
- Stemming
- Stop Word Removal
- Part of Speech Tagging
- Parsing
- Word Sense Disambiguation
- Named Entity Recognition
- Semantic Categorization
- Sentiment Analysis
- Topic Modeling (some caveats)

# NAMED ENTITY RECOGNITION

IDENTIFY AND CLASSIFY NAMED ENTITIES IN TEXT:

JOHN SMITH IS A PERSON

SOVIET UNION IS A COUNTRY

2500 UNIVERSITY DR IS AN  
ADDRESS

(555) 867-5309 IS A PHONE  
NUMBER

ENTITY RELATIONS: HOW DO THE ENTITIES RELATE?

DO THEY CO-OCCUR IN A DOCUMENT? IN A SENTENCE?

# TEXT PROCESSING PIPELINE

## TOKENIZATION: SEGMENT TEXT INTO TERMS

ENTITIES? "SAN FRANCISCO", "O'CONNOR", "U.S.A."

REMOVE STOP WORDS? "A", "AN", "THE", "TO", "BE"

N-GRAMS? CAN TAKE WORDS IN 2-WORD GROUPS (BI-GRAMS), 3-WORD (TRI-GRAMS), ETC.

## STEMMING: GROUP TOGETHER DIFFERENT FORMS

ROOTS: VISUALIZATION(S), VISUALIZE(S), VISUALLY → VISUAL

LEMMATIZATION: GOES, WENT, GONE → GO

FOR VISUALIZATION, SOMETIMES NEED TO REVERSE STEMMING FOR LABELS

SIMPLE SOLUTION: MAP FROM STEM TO THE MOST FREQUENT WORD

## RESULT: ORDERED STREAM OF TERMS



# TEXT PROCESSING PIPELINE

“The quick brown fox jumps over the lazy dog.”

TOKENIZE (N=1)

[The], [quick], [brown], [fox], [jumps], [over], [the], [lazy], [dog].

TOKENIZE (N=1), REMOVE STOPWORDS, STEM

[quick], [brown], [fox], [jump], [over], [lazy], [dog]

TOKENIZE (N=2)

[the quick], [quick brown], [brown fox], [fox jumps], [jumps over], [over the]...

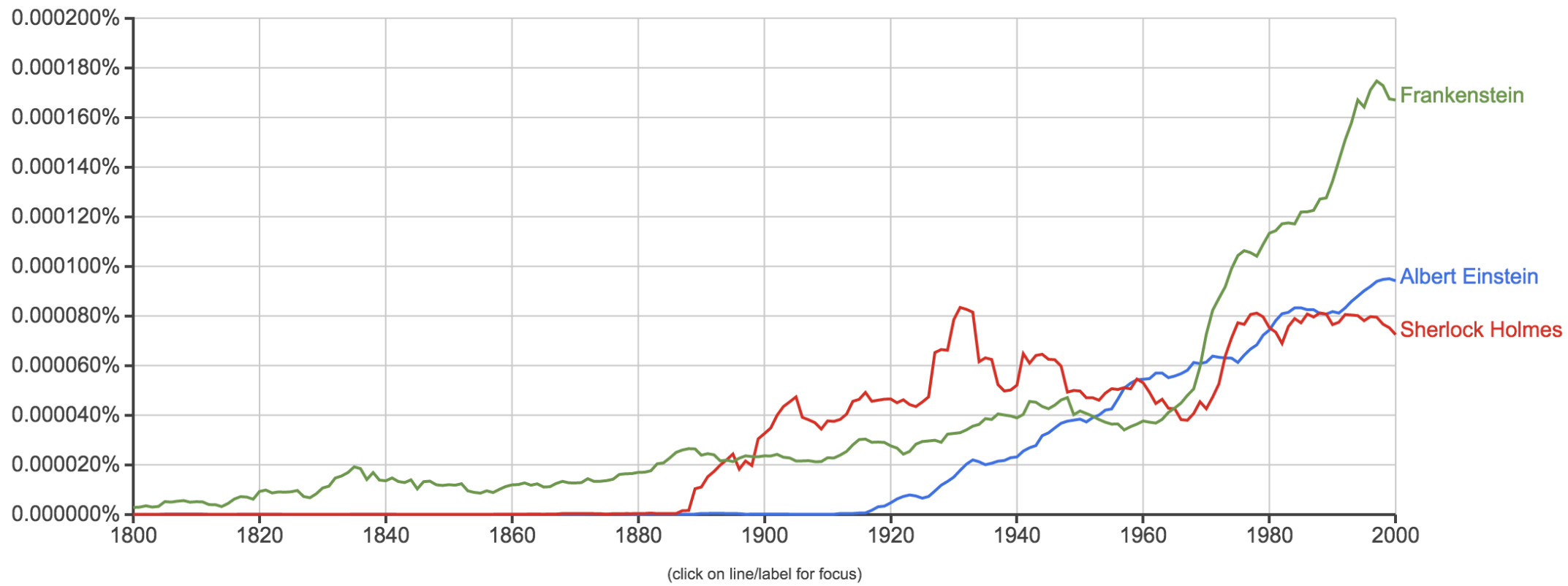
TOKENIZE (N=5)

[the quick brown fox jumps], [quick brown fox jumps over], [brown fox jumps over ...

# Google Books Ngram Viewer

Graph these comma-separated phrases:   case-insensitive

between  and  from the corpus  with smoothing of  [Search lots of books](#)



# NLTK (NATURAL LANGUAGE TOOLKIT)

Tokenize and tag some text:

```
>>> import nltk
>>> sentence = """At eight o'clock on Thursday morning
... Arthur didn't feel very good."""
>>> tokens = nltk.word_tokenize(sentence)
>>> tokens
['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',
'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']
>>> tagged = nltk.pos_tag(tokens)
>>> tagged[0:6]
[('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'), ('on', 'IN'),
('Thursday', 'NNP'), ('morning', 'NN')]
```

NLTK.org  
Python

Identify named entities:

```
>>> entities = nltk.chunk.ne_chunk(tagged)
>>> entities
Tree('S', [
  (('At', 'IN'), ('eight', 'CD'), ("o'clock", 'JJ'),
  ('on', 'IN'), ('Thursday', 'NNP'), ('morning', 'NN')),
  Tree('PERSON', [
    ('Arthur', 'NNP')]),
  ('did', 'VBD'), ("n't", 'RB'), ('feel', 'VB'),
  ('very', 'RB'), ('good', 'JJ'), ('.', '.')])
```

# VISUALIZING DOCUMENT CONTENT

# TAG CLOUDS

WORD COUNT

additional air **analysis** analysts annotation applications approach asked author  
average based build **chart** citizen **clustering** collaborative collection  
**comments** commentspace community complete condition contributions  
crowd crowdsourcing **data** datasets design different discussion evidence example  
experiment experts **explanations** explore features figure  
filtering **generated** group help hypotheses hypothesis identify including indicating  
information interactive interface knowledge **links** members microtasks multiple novice number oae  
observations organize **participants** phases pp proceedings process produced  
prompt **provide quality** questions rate redundant requires **responses results** score  
sense share showing similar site **social source** specific state strategies study support  
systems **tags tasks tools** understanding used **users views**  
**visualization** web work **workers**

<http://tagcrowd.com/>

THESIS WESLEY WILLETT



# WHAT'S PROBLEMS DO YOU SEE WITH TAG CLOUDS?

additional air **analysis** analysts annotation applications approach asked author  
average based build chart citizen **clustering** collaborative collection  
**comments** commentspace community complete condition contributions  
crowd crowdsourcing **data** datasets design different discussion evidence example  
experiment experts **explanations** explore features figure  
filtering **generated** group help hypotheses hypothesis identify including indicating  
information interactive interface knowledge **links** members microtasks multiple novice number oæn  
observations organize **participants** phases pp proceedings process produced  
prompt **provide quality** questions rate redundant requires responses results score  
sense share showing similar site social source specific state strategies study support  
systems tags tasks tools understanding used **users** views  
**visualization** web work **workers**



# 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)

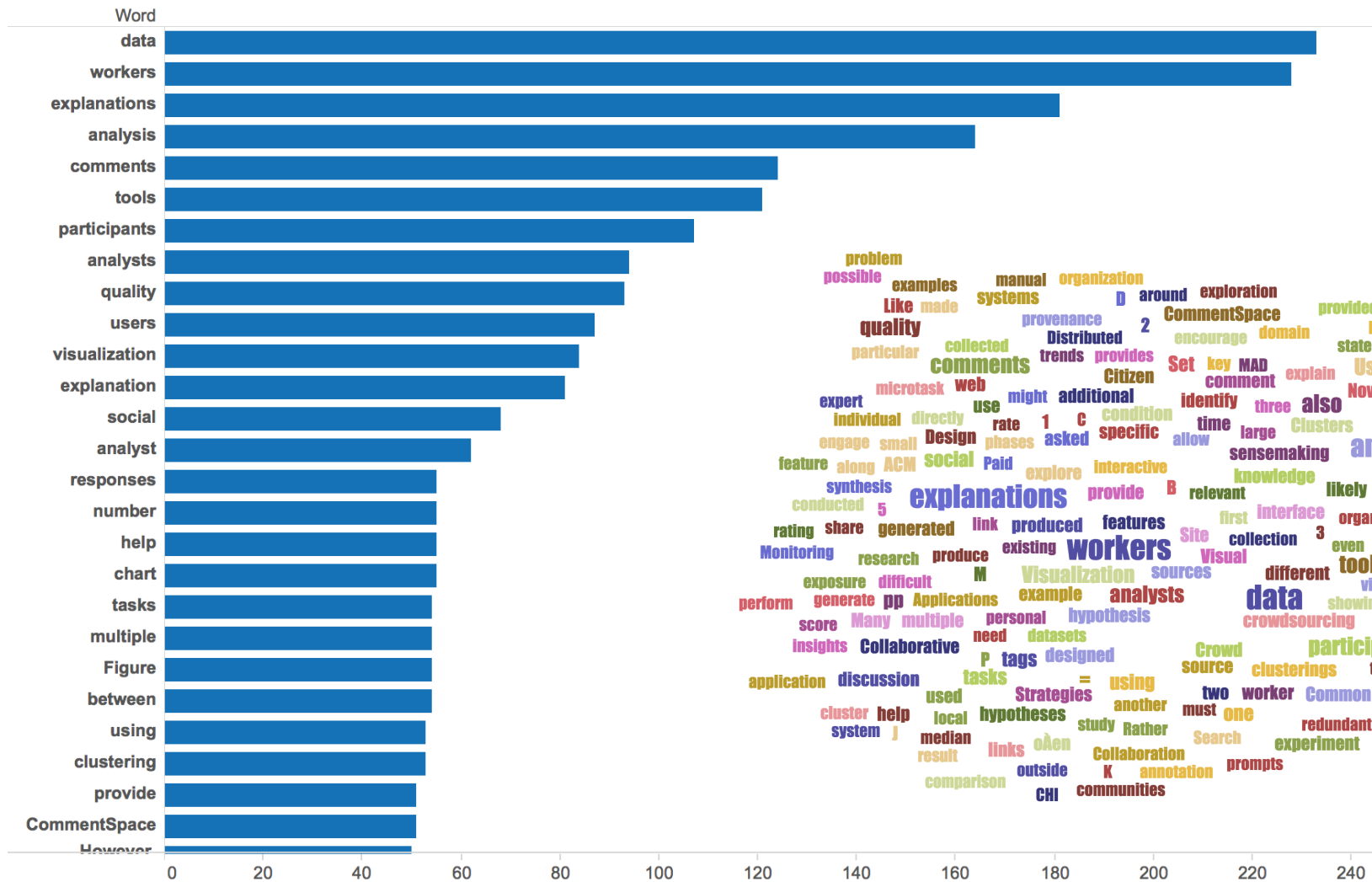
- ORDER USUALLY MEANINGLESS (USUALLY ALPHABETICAL OR RANDOM)

TERM FREQUENCY MAY NOT BE MEANINGFUL

DOES NOT SHOW THE STRUCTURE OF THE TEXT



# WORD COUNTS



# WORDCOUNT

WORDCOUNT

◀ PREVIOUS WORD

NEXT WORD ▶

the of and to ain that it is was i for on you he be with as by a have are this no but had his they from she which we in there he were one do over all it has could will what if on when no in she about the in said the be you are the

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50

CURRENT WORD

FIND WORD:

BY RANK:

REQUESTED WORD: THE

RANK: 1

ARCHIVE

COUNT

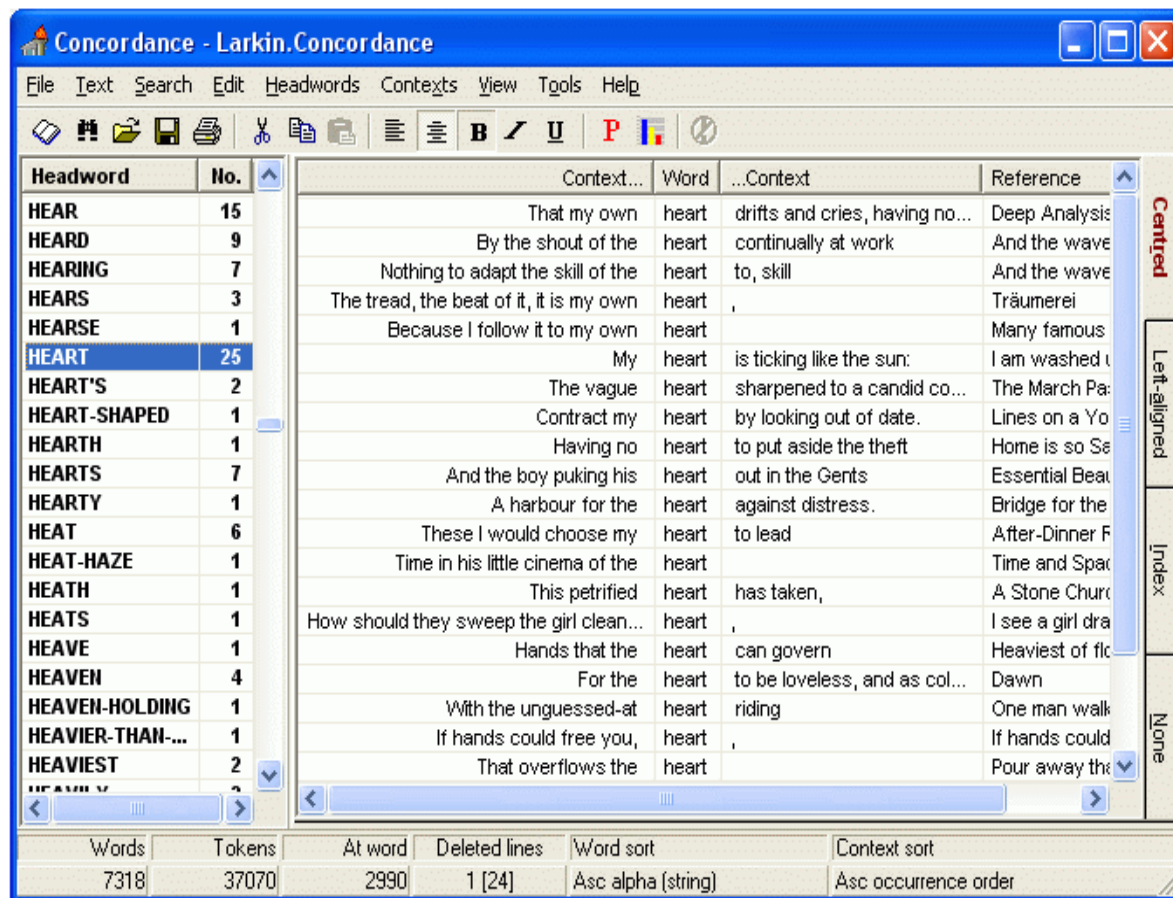


JONATHAN HARRIS

<http://wordcount.org>

# CONCORDANCE

WHAT IS THE COMMON LOCAL CONTEXT OF A TERM?



The screenshot shows the Larkin Concordance software interface. The main window displays a list of words and their occurrences in a text corpus. The word 'HEART' is highlighted in blue, indicating it is the current term being viewed. The interface includes a menu bar (File, Text, Search, Edit, Headwords, Contexts, View, Tools, Help), a toolbar with various icons, and a status bar at the bottom showing statistics: Words (7318), Tokens (37070), At word (2990), Deleted lines (1 [24]), Word sort (Asc alpha [string]), and 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
<b>HEART</b>	<b>25</b>	My	heart	is ticking like the sun:	I am washed t
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 Spar
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 th

# WORD TREES

- cats are better than dogs
- cats eat kibble
- cats are better than hamsters
- cats are awesome
- cats are people too
- cats eat mice
- cats meowing
- cats in the cradle
- cats eat mice
- cats in the cradle lyrics
- cats eat kibble
- cats for adoption
- cats are family
- cats eat mice
- cats are better than kittens
- cats are evil
- cats are weird
- cats eat mice



WATTENBERG & VIÉGAS 2008

# love the

## lord

### thy god

#### with all

- thine heart , and with all thy soul ,
  - and with all thy might .
  - that thou mayest live .
- thy heart , and with all thy soul , and with all thy

mind  
strength , a

#### and

- keep his charge , and his statutes , and his judgments , and his commandments , always .
- to walk ever in his ways ; then shalt thou add three cities more for thee , beside these three : 19
- that thou mayest obey his voice , and that thou mayest cleave unto him : for he is thy life , and t

#### ,

to walk in his ways , and to keep his commandments and his statutes and his judgments , that thou mayest liv

#### and to

- serve him with all your heart and with all your soul , 11 : 14 that i will give you the rain of your lan
- walk in all his ways , and to keep his commandments , and to cleave unto him , and to serve him

#### ,

to walk in all his ways , and to cleave unto him ; 11 : 23 then will the lord drive out all these nations from

with all your heart and with all your soul .

### your god

- all ye his saints : for the lord preserveth the faithful , and plentifully rewardeth the proud doer .
- hate evil : he preserveth the souls of his saints ; he delivereth them out of the hand of the wicked .
- because he hath heard my voice and my supplications .

name of the lord , to be his servants , every one that keepeth the sabbath from polluting it , and taketh hold of my covenant  
 good , and establish judgment in the gate : it may be that the lord god of hosts will be gracious unto the remnant of joseph  
 evil ; who pluck off their skin from off them , and their flesh from off their bones ; 3 : 3 who also eat the  
 truth and peace .

other ; or else he will hold to the one , and despise the other . ye cannot serve god and mammon .

6 : 25 therefore i say unto  
16 : 14 and the pharisees

### uppermost

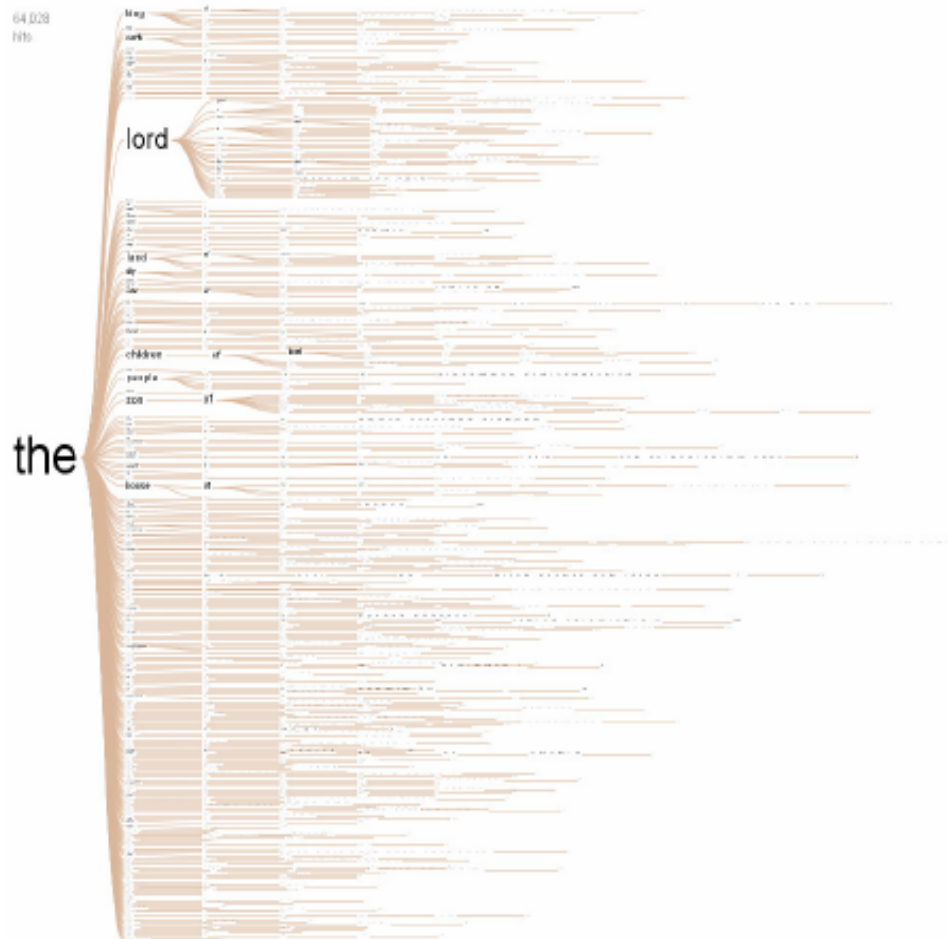
- rooms at feasts , and the chief seats in the synagogues , 23 : 7 and greetings in the markets , and to be called of
- seats in the synagogues , and greetings in the markets .

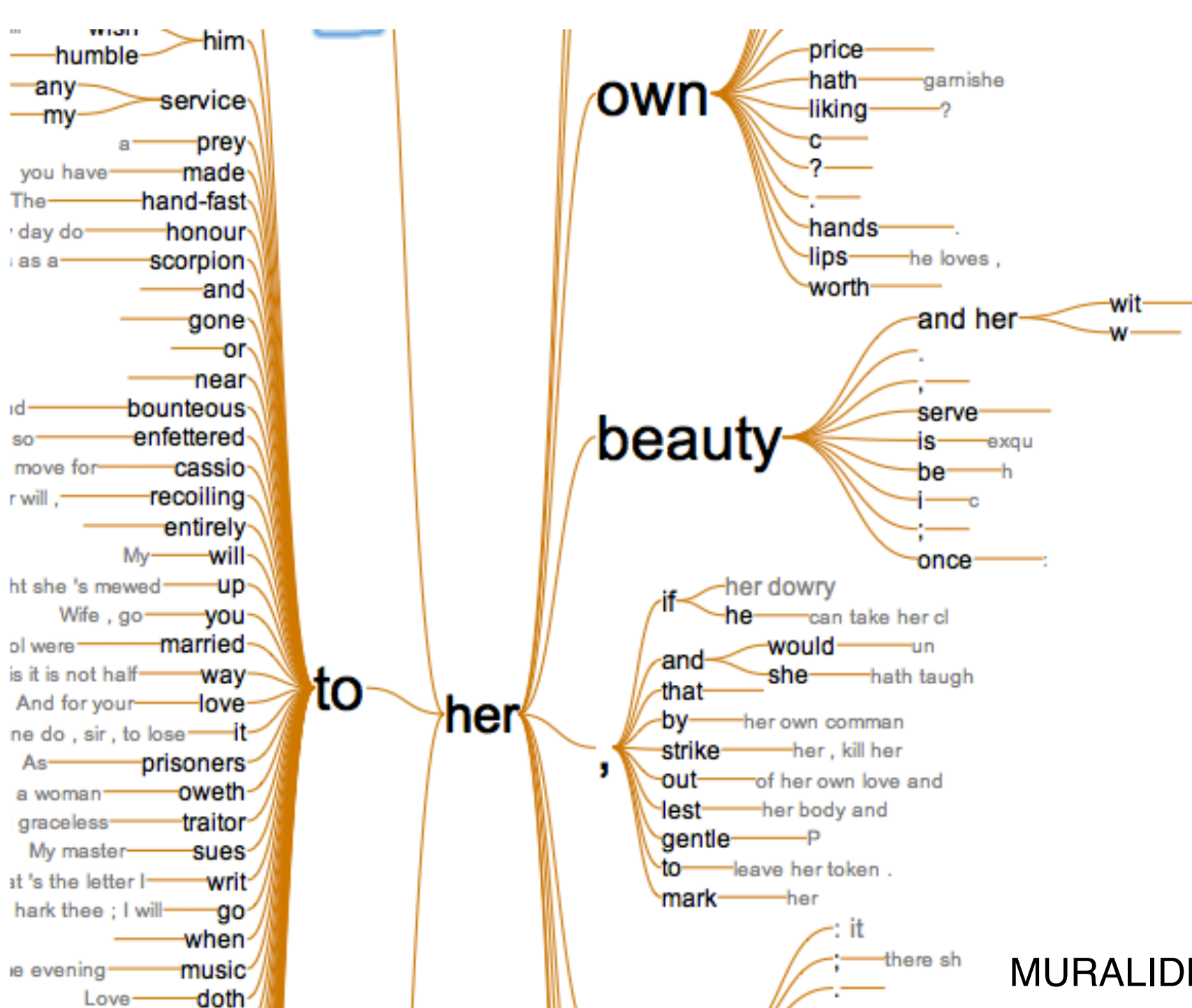
### father

- ; and as the father gave me commandment , even so i do .
- hath bestowed upon us , that we should be called the sons of god : therefore the world knoweth us not , because it knew him

brotherhood .  
 world , the love of the father is not in him .  
 brethren .  
 children of god , when we love god , and keep his commandments .

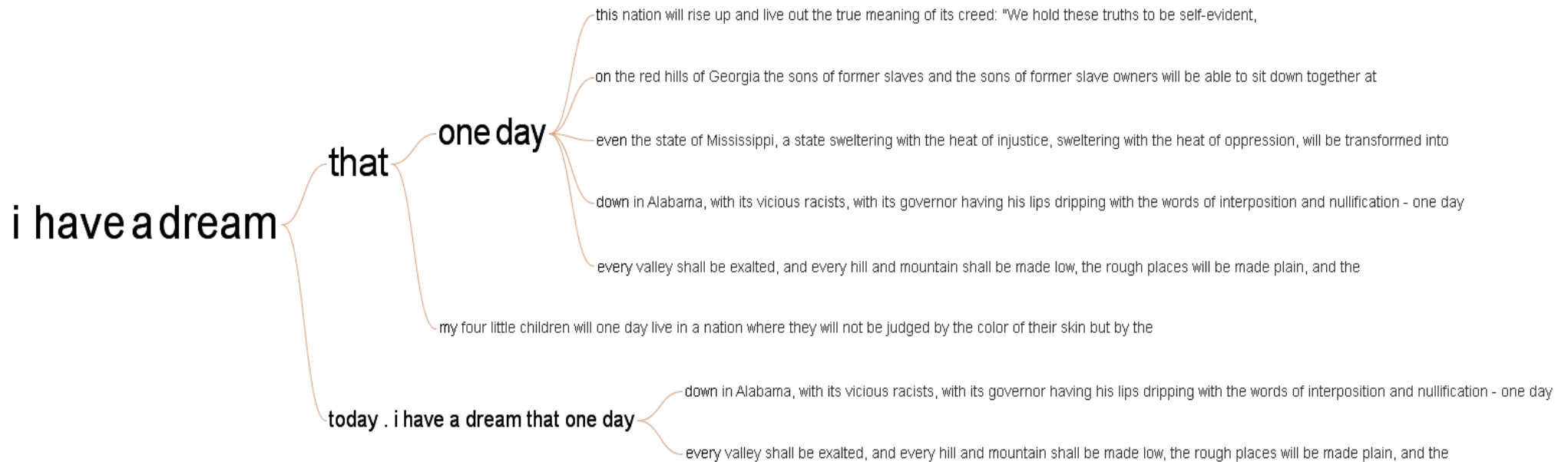
# FILTER INFREQUENT RUNS





**WORDSEER**  
**MURALIDHARAN & HEARST**

# RECURRENT THEMES IN SPEECH





# GLIMPSES OF STRUCTURE

CONCORDANCES SHOW LOCAL, REPEATED  
STRUCTURE

BUT WHAT ABOUT OTHER TYPES OF PATTERNS?

FOR EXAMPLE

LEXICAL:        <A> at <B>

SYNTACTIC:    <Noun> <Verb> <Object>

# PHRASE NETS

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 EXTRACTED PATTERNS IN A NODE-LINK VIEW

OCCURRENCES = NODE SIZE

PATTERN POSITION = EDGE DIRECTION

van Ham et al

Select a phrase

word1	and	word2
word1	's	word2
word1	of the	word2
word1	the	word2
word1	a	word2
word1	at	word2
word1	is	word2
word1	[space]	word2

or enter your own  
\* and \*

Filters

Show top:

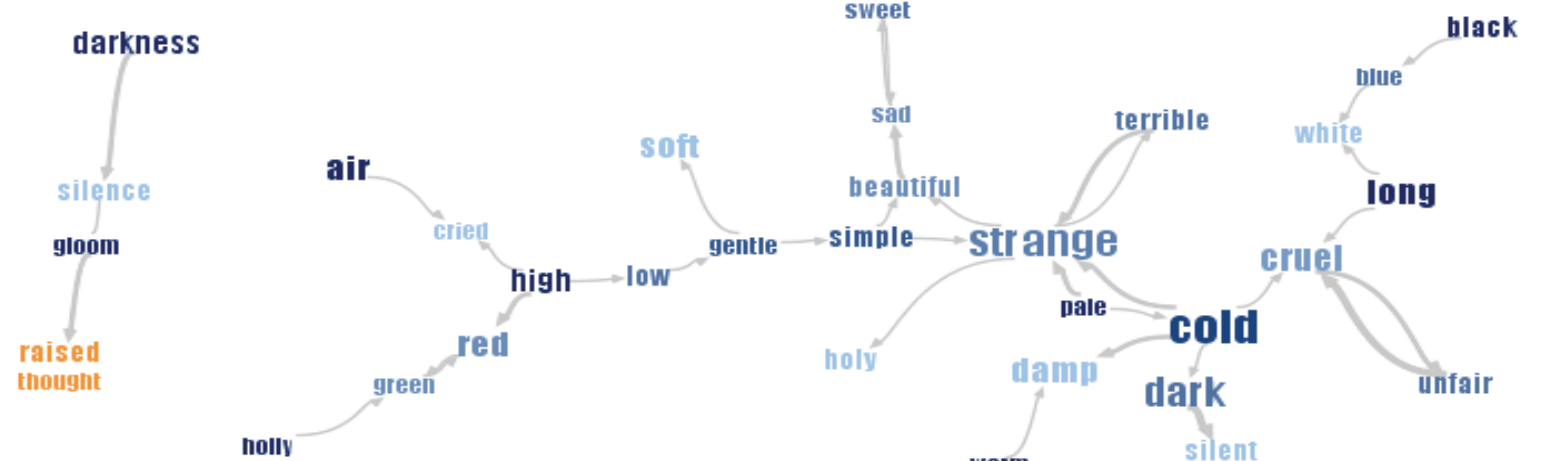
Hide common words

Zoom

In  Out  Reset

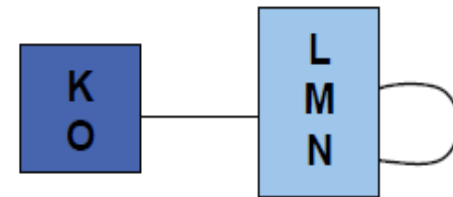
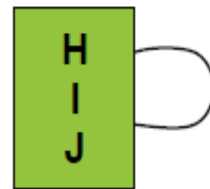
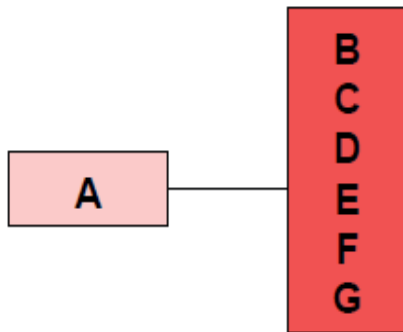
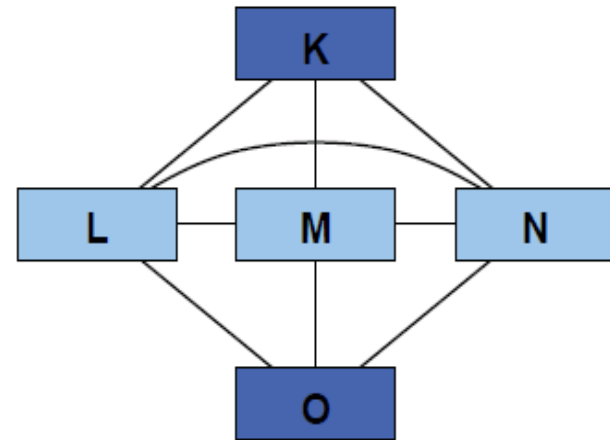
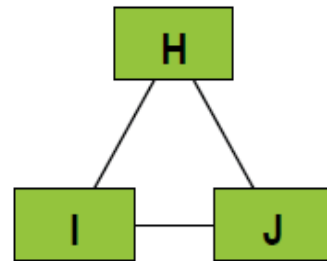
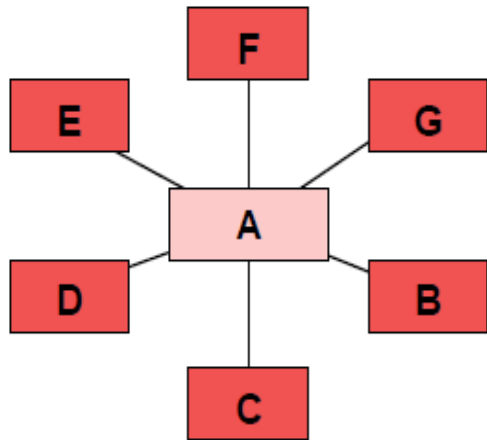
Showing 73 of 1719 terms

X and Y



# PORTRAIT OF THE ARTIST AS A YOUNG MAN JAMES JOYCE

# NODE GROUPING



(a)

(b)

(c)

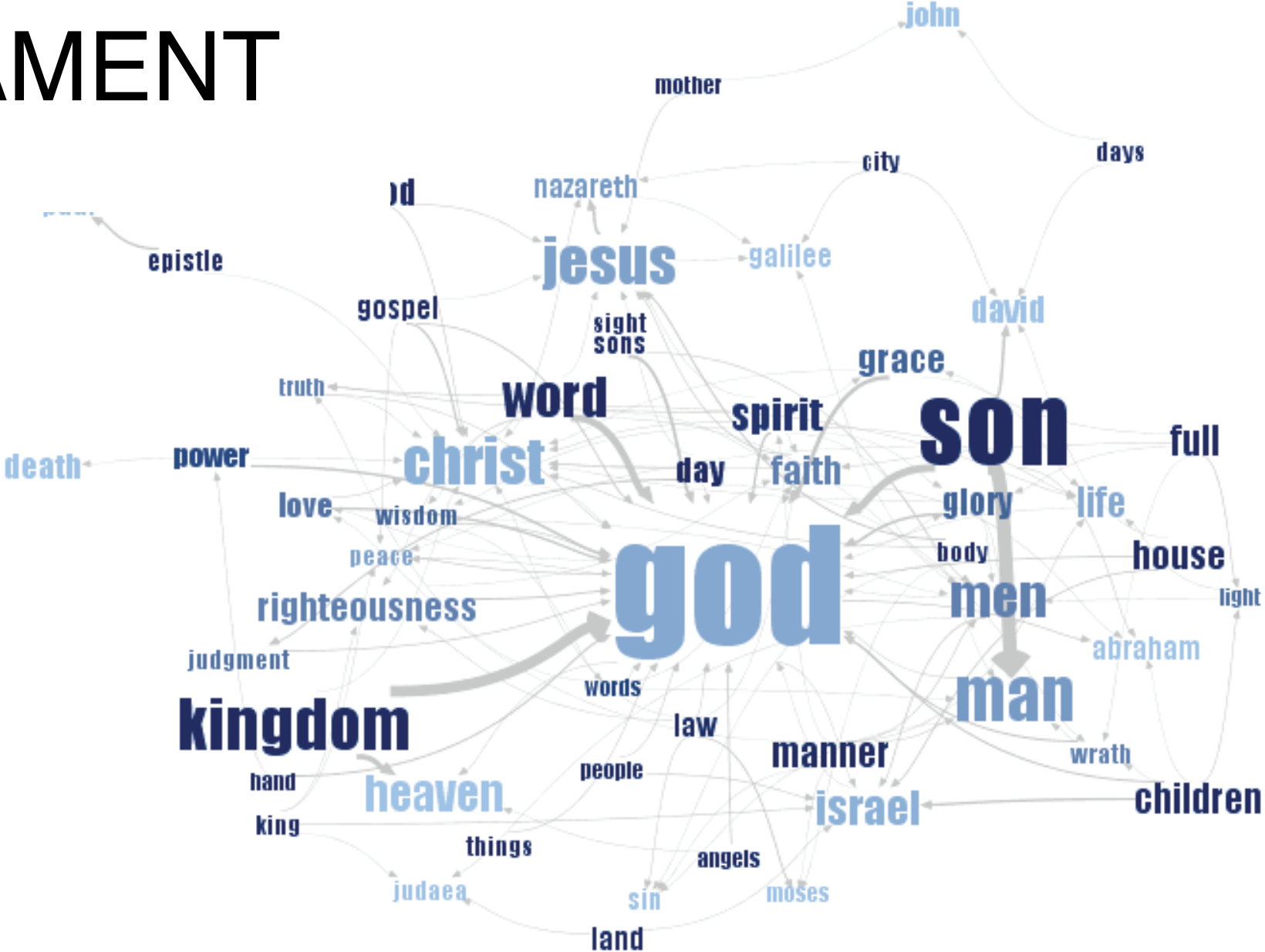






# NEW TESTAMENT

{X} of {Y}





# VISUALIZING DOCUMENT COLLECTIONS

# Analysts: GOP may regret gridlock over Scalia replacement

# Update: Uber driver arrested in Michigan rampage that killed 6


# Boris Johnson backs EU exit: London mayor confirms support for Brexit

## 'A multifaceted catastrophe': Turkey has 'so alienated everyone'

## Blasts rock Syrian city of Homs, killing at least 32

## Palestinians struggle to define those who attack Israelis

**Canada, USA renew rivalry in CONCACAF final**



Sportsnet's James Sharman met with coach John Herdman and members of the Canadian women's soccer team, who are looking to beat the USA in Sunday's CONCACAF final. headshot Gavin Day February 20, 2016, 8:08 PM. headshot Gavin Day February 20, 2016, 8:08 PM.

Feb 20 17:47 | 587 related articles | Sportsnet.ca

## North Korea peace talks offer before last nuclear test

## Malaysia, south-east Asia nations warned of terror attacks

# Samsung, LG unveil new devices in bid for smartphone recovery

## Raceline Radio Program Guide: February 21, 2016

## Canada, USA renew rivalry in CONCACAF final

## 'Deadpool' dominates again with \$55 million in 2nd week

## Judge blocks attempt to halt deposition of Bill Cosby's wife

## LG Unveils the LG G5, Its First Modular Smartphone [Video]

**LG G5 vs LG V10: first look**

EPA asks Volkswagen to make electric cars in the US

What if San Bernardino suspect hid an Android rootkit?

Fire Extinguisher: Rules for Using One

HECO To Bring Local Climate Plan, Study, Policy

ZTE and China

New York

London Dies: Possible, Big Users With Budget Tablet

Choking apparatus of outfit that is just spilling from the ready and low

Disputed climbing: Police arrest man after climbing rockery

Has our report? Climbing dog

One dead, one injured, in avalanche near Golden

La Luche staff students return to school this week

Did it? After Robert Pickton pens book from prison

Has our report? Climbing dog

One dead, one injured, in avalanche near Golden

Headings to watch for in the Canadian CP&I&A

Others

## Chan wins Four Continents figure-skating championship

Years later, ex-Raptor Vince Carter's still soaring

SPRING TRAINING Blue Jays' focus at 2016 camp is on 2017

Canadian women earn historic 19-10 rugby victory over New Zealand

Miller puts an end to Canucks' losing streak

Leafs get set for a busy draft with Matthias trade

Water powers Ducks to 5-2 win over Flames

Blair's Giants' playoff push

Canucks' Miller

How player Patrick says like Jays' playoff push

Signs climbing: surgeons risk developing high blood pressure and heart disease

HPV cases drop since vaccinations started

Cherry

Asian stocks rebound in anticipation of G20 meeting

Headings to watch for in the Canadian CP&I&A

Others

Scientists at Brock studying Zika to see if Canadian mosquitoes can spread the virus

How Syrian refugees arriving in Canada became 'extras' in their own stories

One dead, another injured, in avalanche near Golden

La Luche staff students return to school this week

Did it? After Robert Pickton pens book from prison

Has our report? Climbing dog

One dead, one injured, in avalanche near Golden

Headings to watch for in the Canadian CP&I&A

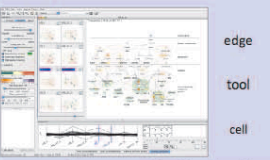
Others

# DOCUMENT CARDS

## SMALL MULTIPLES FOR DOCUMENTS

**1** Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context

systems biologist context Interaction graph  
graph model dataset figure



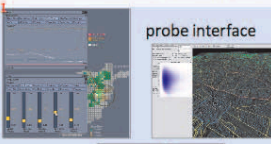
edge  
tool  
cell

gene  
layout algorithm  
process node cerebral

Aaron Barsky, Tamara Munzner, Jennifer Gardy, and Robert Kincaid

**2** Multi-Focused Geospatial Analysis Using Probes

probe interface

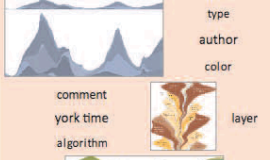


participant figure  
type scale  
window region-of-interest  
local region  
data  
application

Thomas Butkiewicz, Wenwen Dou, Zachary Wartell, William Ribarsky, and Remco Chang

**3** Stacked Graphs: Geometry & Aesthetics

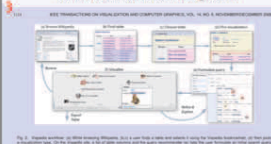
question visualization paper



type author color  
comment york time layer  
algorithm trend  
namevoyager  
people graphic time sery system  
legibility design issue layout method

Lee Byron and Martin Wattenberg

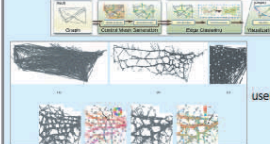
**4** Vispedia: Interactive Visual Exploration of Wikipedia Data via Search-Based Integration



Bryan Chan, Leslie Wu, Justin Talbot, Mike Cammarano, and Pat Hanrahan

**5** Geometry-Based Edge Clustering for Graph Visualization

edge bundle technique polyline segment  
large graph control mesh straight line  
road map mesh edge pattern transfer function

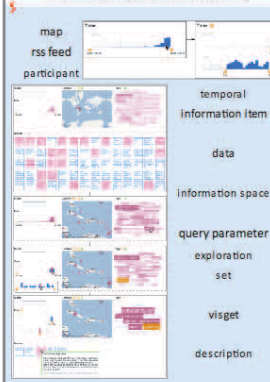


color and opacity enhancement  
node position control point graph layout  
result method visual clutter  
general graph primary direction  
user

Weiwei Cui, Hong Zhou, Student, Huamin Qu, Pak Chung Wong, and Xiaoming Li

**6** VisGts: Coordinated Visualizations for Web-based Information Exploration and Discovery

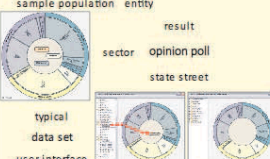
map rss feed participant  
temporal information item  
data information space  
query parameter exploration set  
visget  
description



Marian Dörk, Sheelagh Carpendale, Christopher Collins, and Carey Williamson

**7** Who Votes For What? A Visual Query Language for Opinion Data

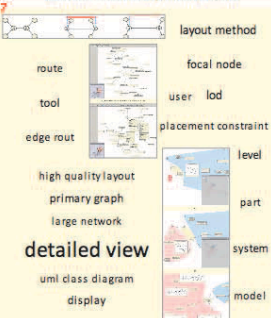
user study report attribute paper  
sample population entity result  
sector opinion poll state street  
typical data set user interface  
visual query language visualization task  
design participant poll data ring system  
data point



Geoffrey M. Draper, and Richard F. Riesenfeld

**8** Exploration of Networks Using Overview+Detail with Constraint-based Cooperative Layout

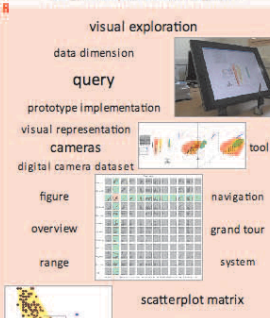
layout method route focal node  
tool user lod  
edge rout placement constraint  
high quality layout primary graph large network  
detailed view uml class diagram display  
model  
layout technique cluster position structure  
focus node



Tim Dwyer, Kim Marriott, Falk Schreiber, Peter J. Stuckey, Michael Woodward, and Michael Wybrow

**9** Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation

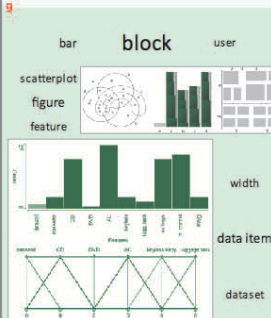
visual exploration data dimension query  
prototype implementation visual representation cameras  
digital camera dataset figure navigation  
overview grand tour system  
range scatterplot matrix user  
operation method order



Niklas Elmquist, Pierre Dragicevic, and Jean-Daniel Fekete

**10** Interactive Visual Analysis of Set-Typed Data

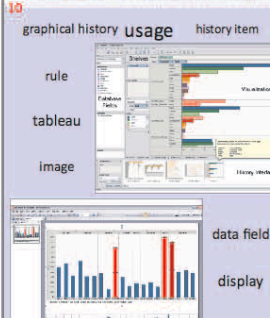
bar block user  
scatterplot figure feature  
width data item dataset  
data record histogram washing agent set-typed data view



Wolfgang Freiler, Kresimir Metković, Computer Society, and Helwig Hauser

**11** Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation

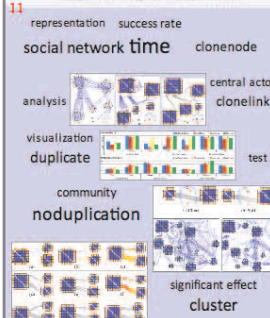
graphical history usage history item  
rule tableau Image  
data field display approach  
event history interface history tool



Jeffrey Heer, Jock D. Mackinlay, Chris Stolte, and Maneesh Agrawala

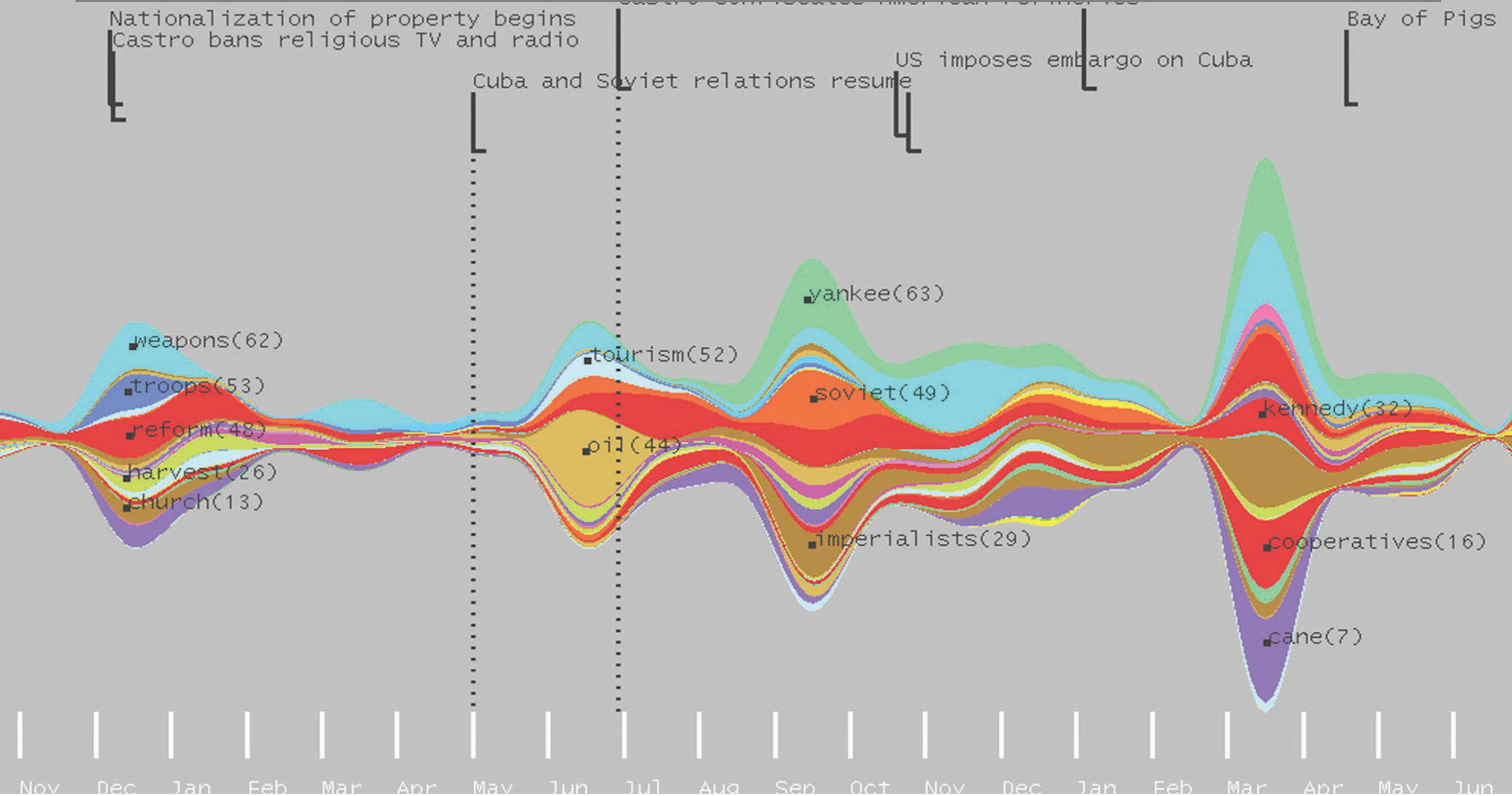
**12** Improving the Readability of Clustered Social Networks using Node Duplication

representation success rate social network time clonemode  
analysis central actor clonelink  
visualize duplicate test  
community noduplication significant effect cluster  
duplication link participant splitlink readability

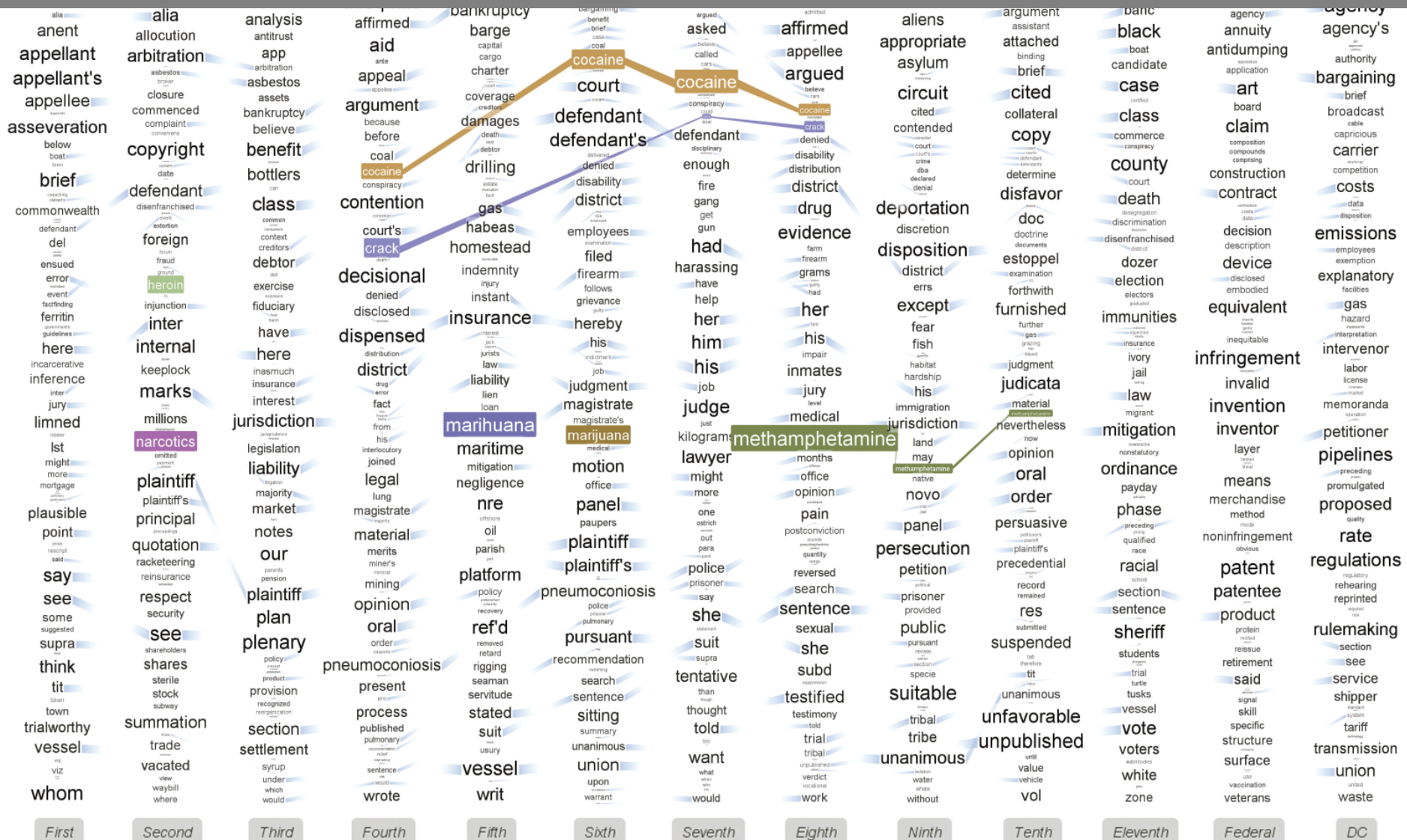


Nathalie Henry, Anastasia Bezerianos, and Jean-Daniel Fekete

# THEMERIVER HAVRE ET AL 1999



# PARALLEL TAG CLOUDS



# SUPPORTING SEARCH

The screenshot displays the TileBars search interface. At the top, a 'User Query' box contains the text 'osteoporosis', 'prevention', and 'research' on separate lines. To the right of the query box are three buttons: 'Run Search', 'New Query', and 'Quit'. Below these buttons are two rows of search parameters: 'Search Limit' with options 50, 100, 250, 500, and 1000; and 'Number of Clusters' with options 3, 4, 5, 8, and 10. The '250' and '5' options are selected. Below the search parameters, the mode is set to 'TileBars'. There are two tabs, 'Cluster' and 'Titles', and a 'Backup' button. The main display area is split into two panes. The left pane shows a vertical list of 12 horizontal bar charts, each representing a different cluster. The right pane shows a list of search results, including document IDs and titles such as 'FR88513-0157', 'AP: Groups Seek \$1 Billion a Year for Aging Research', and 'SJMN: WOMEN'S HEALTH LEGISLATION PROPOSED CF'.

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

FR88513-0157  
AP: Groups Seek \$1 Billion a Year for Aging Research  
SJMN: WOMEN'S HEALTH LEGISLATION PROPOSED CF  
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

**TileBars Hearst 1999**

/tmp/words22058



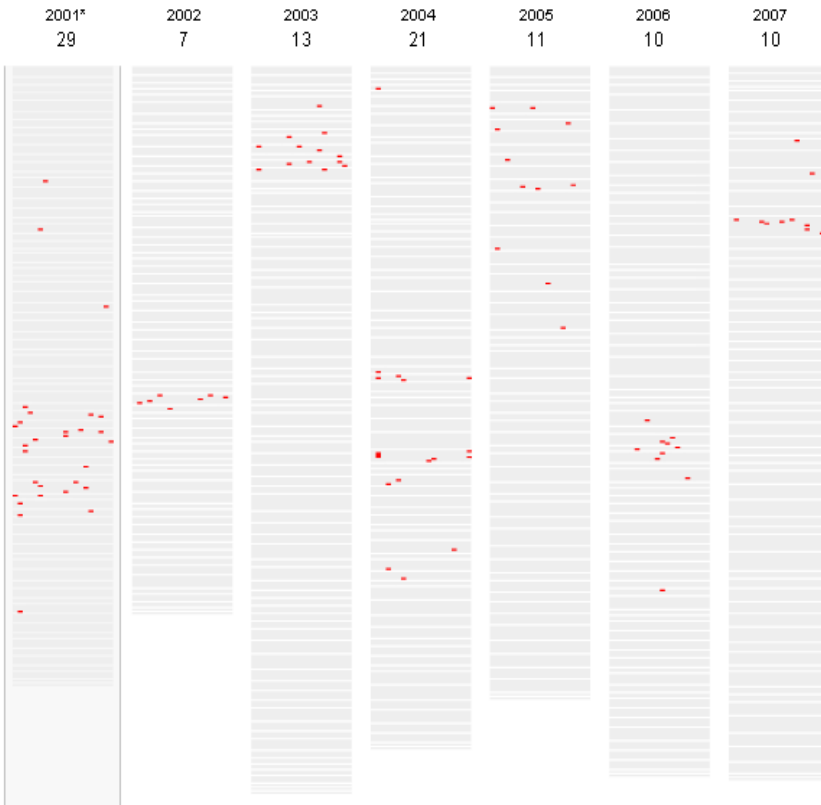
SeeSoft Eick 19

# 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.

  or [choose a word here.](#)

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



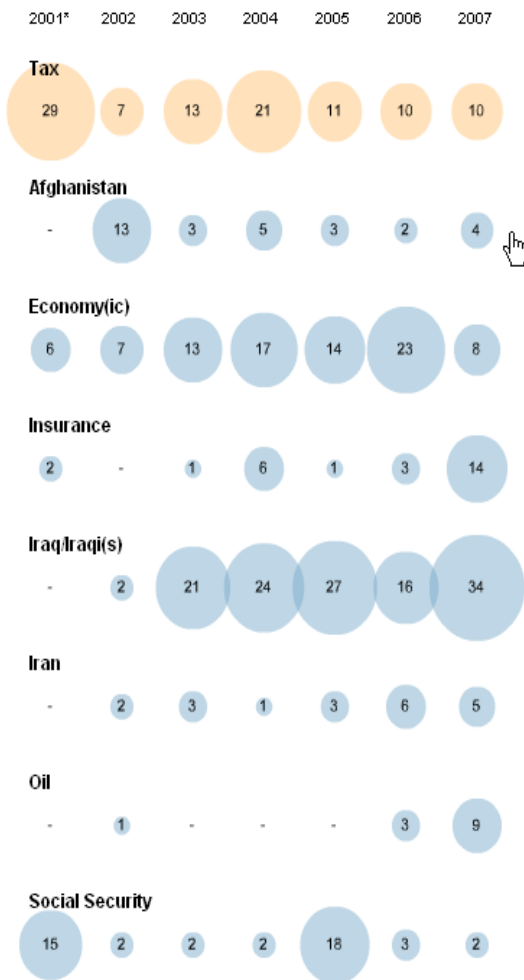
### The word in context

[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)

## 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.



List View

Edit View Bookmarks Lists Options Export

year Add all Clear

author Add all Clear

concept Add all Clear

1995  
1996  
1997  
1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011

Keim, D.A.  
Delke, D.  
Schneidewind, J.  
Dayal, U.  
Hao, M.C.  
Mansmann, F.  
North, S.  
Panse, C.  
Sips, M.  
Bak, P.  
Janetzko, H.  
Rohrdanz, C.  
Schrock, T.  
Stoffel, A.  
Albuquerque, G.  
Ankerst, M.  
Berchtold, S.  
Danon, G.  
Deussen, O.  
Eisemann, M.  
Gruse, T.  
Haug, L. E.  
Heilmann, R.  
Hsu, M.  
Jenny, M.  
Ladisch, J.  
Last, M.  
Magnork, M.  
Marsch, D.

insight  
text  
pixel  
distortion  
document  
geographic  
hierarchy  
interaction  
parallel coordinates  
case study  
clustering  
color  
evaluation  
network  
time series  
treemap  
animation  
business  
cluster  
financial  
geospatial  
high-dimensional da...  
overview  
radial  
security  
toolkit  
visual analytics  
zooming  
aesthetics

Document Cluster View

Edit View Bookmarks Export Options

Highlight viewed documents

Filters

All Filters

Group by Filters

Undo Filters

Hide Unfiltered

Clusters

Text Seed (20)


Freq Words Unique Words

All Documents

- visual, animation, trends: 21
- visualization, users, tables:
- visualizations, used, transfo
- visualization, use, classifica
- visual, insight, genes: 19
- visualization, treemaps, layc
- visualizing, set, use: 56
- visualizing, users, spaces: 1

animation, trends, causality  
transform, quality, studied  
insight, genes, expression  
tables, database, interfaces  
classification, geographic, statistics  
treemaps, coloring, hierarchically  
network, graph, social  
graphs, edge, algorithm  
dimensions, coordinates, parallel  
spaces, internet, search  
interact, understand, cognition  
diverse, environments, toolkit  
text, features, topic  
video, explorer, stories  
3d, displays, navigation  
analytics, anomalies, detect  
history, mining, patterns  
querying, series, temporal  
state, displayed, explored

JIGSAW

**A**  **CENDARI**  
COLLABORATIVE EUROPEAN DIGITAL ARCHIVE INFRASTRUCTURE

Home Browse About Issue Report Survey Search anthi

**B** Resources

- My Projects:
  - Green Cadres
  - WW1
    - Notes (1)
      - Green Cadres Notes
    - Documents (144)
    - Entities (7)
      - Event (1)
      - Organization (0)
      - Person (3)
      - Publication (0)
      - Artifact (0)
      - Place (5)
      - Tag (3)

**C** New Save Import Help

Note 5: Green Cadres Notes

Entities (12) Status (Open) Assigned Users

Green Cadres Notes

Note Description [ Read Only ] --- click here for Edit mode

In 1918, as privations and social unrest began to undermine the Austro-Hungarian war effort on the home front, a specific kind of revolt gripped the countryside in a number of regions of the empire. The so-called **Green Cadres** or **Green Brigades** were groups of armed deserters, supplemented by the local poor peasantry, who hid themselves in forested areas, staging raids on livestock and crops, attacking the local gendarmerie and military, and (in some instances) articulating social revolutionary programs. Reports on these irregular armed bands abounded in the final year of the year in many regions of both **Austria and Hungary**, but they were concentrated in **Croatia-Slavonia** (current Croatia and **Serbia**) and southern **Moravia** (current Czech republic). The **Green Cadres** represented a specifically rural form of unrest—largely unhitched from **nationalist** and party political agendas—reflecting the widespread sense of apocalyptic collapse among the rural population of Austria-Hungary.

The historical research on the Green Cadres is scant and preponderantly concentrated on the region of Croatia-Slavonia, where the Cadres were most numerous and their actions most ambitious. Communist-era **Yugoslav** scholarship treated the Green Cadres as proto-Bolsheviks, overemphasizing the prevalence of **Leninist** ideas among them. Indeed, research has revealed that soldiers returning from Russian imprisonment in 1918 played leading roles in mass desertions, mutinies, and the propagation of social-revolutionist ideas. But scholars have not identified the specific mechanisms by which former POWs became Green Cadres or how the Russian experience was reinterpreted in rural Austro-Hungarian contexts. More importantly, a comparative study of the cadres in various regions is missing because of the challenges of finding, organizing, and interpreting sources that are now fragmented in various national archival research 'siloes'.

This project seeks to open up comparative vistas on the problem of the Green Cadres. Among the possible questions it seeks to answer are: 1. How did the far-flung groups identified as Green Cadres compare to each other in terms of actions and aims; 2. Why did the Cadres appear in the places that they did? 3. What were the social, political, and **cultural** factors that facilitated the formation or concentration of Cadres in specific locales? 4. What kind of **deserters** made up the bulk of the Cadres—deserters from the front, replacement regiments, or allotted leave after returning from **Russian internment**? 5. What played a bigger role in the formation of Green Cadres: social revolutionary influences from Russian imprisonment or disillusionment with the war effort?

**D** Visualizations


Most Common Person **FRAPET, Guillaume**

Most Common Place **Nantes** 128 docs

Most Recent **Date: 1711/1/29** 1711-1-29

Oldest **Date: 1669/6/5** 1669-6-5

Most Common Place **Nantes** 128 docs



# CENDARI NOTE-TAKING ENVIRONMENT 2015

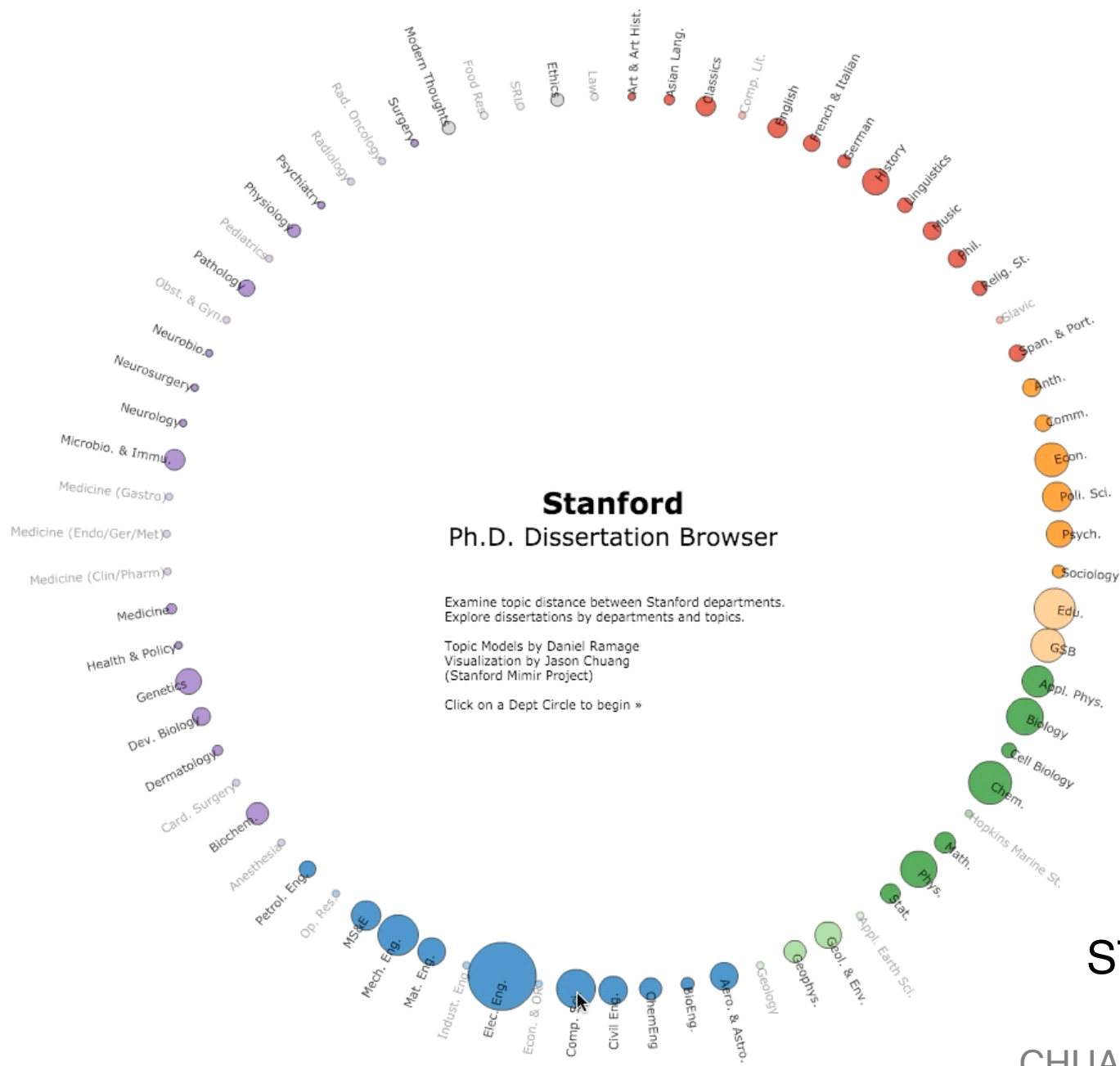
# DOCUMENT SIMILARITY & CLUSTERING

COMPUTE SIMILARITY BETWEEN DOCUMENTS  
BASED ON THE WORDS THEY SHARE

- TF-IDF (TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY) IS COMMON

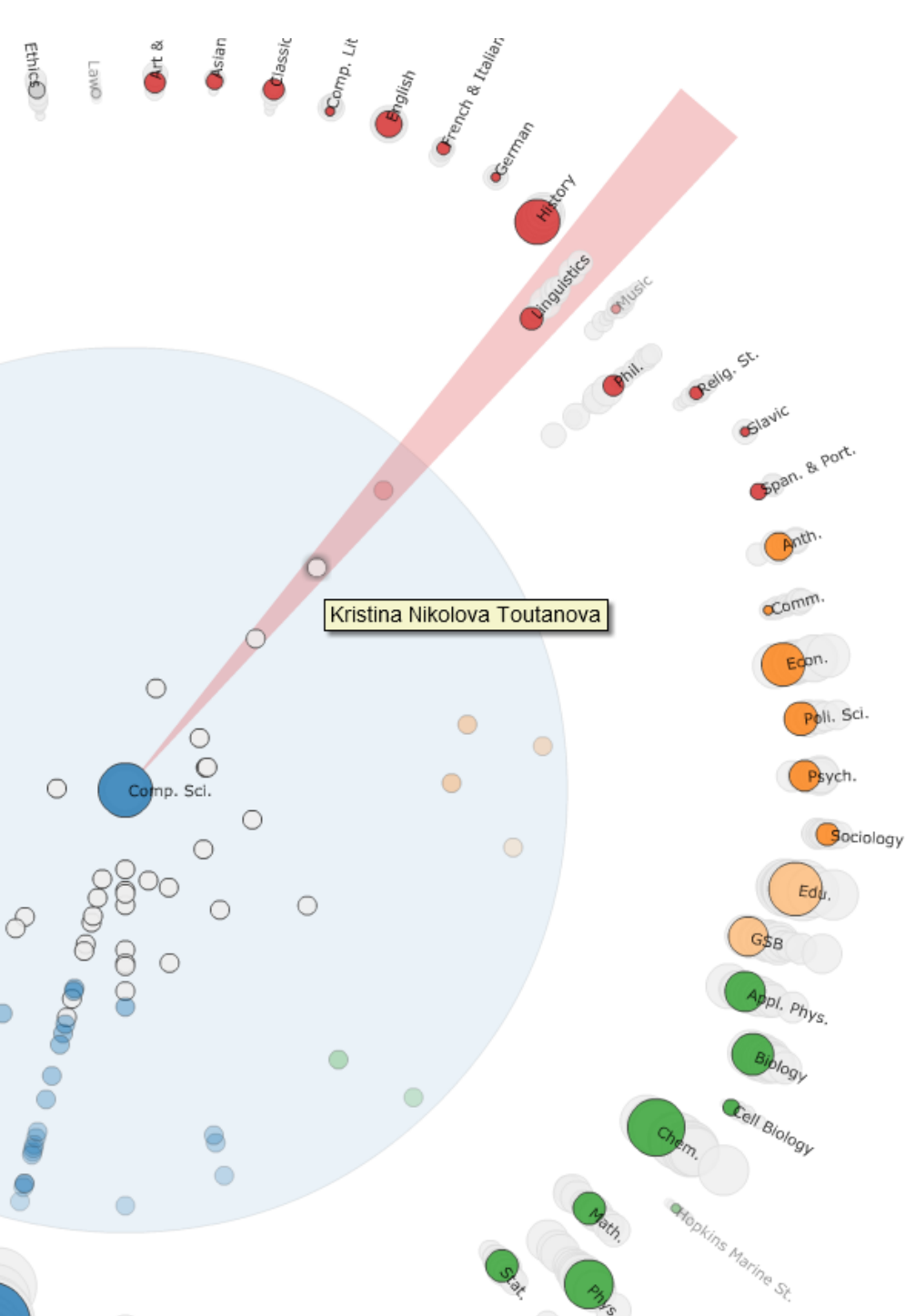
TOPIC MODELING APPROACHES

- ASSUME DOCUMENTS ARE A MIXTURE OF TOPICS
  - TOPICS ARE (ROUGHLY) A SET OF CO-OCCURRING TERMS
  - LATENT SEMANTIC ANALYSIS (LSA): REDUCE TERM MATRIX
- 
- MANY, MANY APPROACHES EXIST



# STANFORD DISSERTATION BROWSER

CHUANG, RAMAGE, MANNING & HEER  
2012



**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**

CHUANG, RAMAGE, MANNING & HEER 2012

# WARNING

OFTEN, TEXT VISUALIZATIONS DO NOT REPRESENT TEXT DIRECTLY, BUT THEY REPRESENT A MODEL  
WORD COUNTS, WORD SEQUENCES, CLUSTERS, ETC.

ASK:

CAN YOU INTERPRET THE VISUALIZATION?

DOES THE MODEL ACCURATELY REPRESENT THE ORIGINAL TEXT?

# LESSONS FOR TEXT VISUALIZATION

SHOW SOURCE TEXT (OR PROVIDE ACCESS TO IT)

WHERE POSSIBLE, USE VISUALIZATION AS INDEX INTO DOCUMENTS

GROUP DOCUMENTS IN MEANINGFUL WAYS

WILL VIEWERS UNDERSTAND THE CLUSTERS?

WHERE POSSIBLE USE TEXT TO REPRESENT TEXT

# HUNDREDS OF TOOLS & TECHNIQUES FOR TEXT AT <http://textvis.lnu.se/>

The screenshot shows the 'Text Visualization Browser' website. The browser's address bar displays 'textvis.lnu.se'. The page title is 'Text Visualization Browser' with the subtitle 'A Visual Survey of Text Visualization Techniques' and 'Provided by ISOVIS group'. Navigation links for 'About', 'Add entry', and 'Other surveys' are visible. On the left, there is a sidebar with 'Techniques displayed: 272', a search box, a time filter slider set to 1976-2016, and a grid of 'Analytic Tasks' icons. The main content area is a grid of 28 thumbnail images representing different text visualization techniques. A tooltip over one thumbnail reads 'Visual Plagiarism Analysis Tool (2015)'. A 'Display a menu' button is located in the bottom right corner.



**QUESTIONS?**