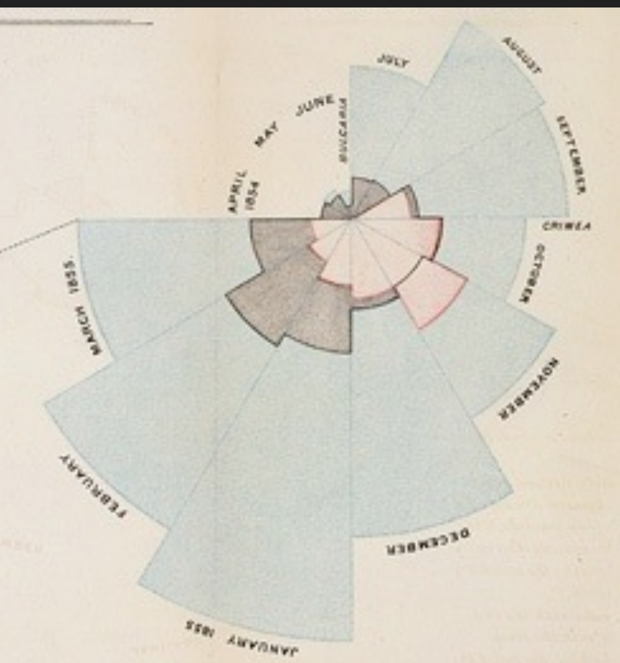
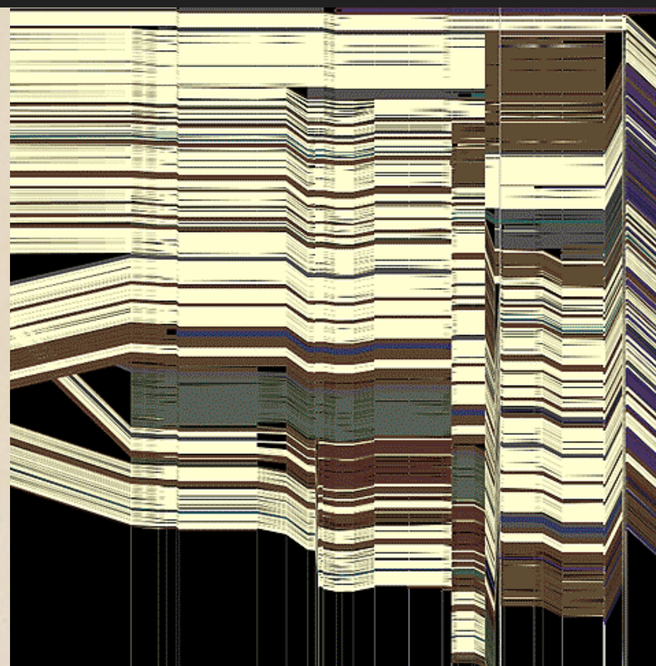
**CSE 412** - Intro to Data Visualization

# Text Visualization




Jane Hoffswell University of Washington

# Why Visualize Text?

# Why Visualize Text?

**Understanding** – get the "gist" of a document

**Grouping** – cluster for overview or classification

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

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

# Text 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.

# Example:
# Health Care Reform

# Example: Health Care Reform

**Background**

Initiatives by President Clinton (1993)
Overhaul by President Obama (2009)

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

# Obama on Health Care, 2009

**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]


Bill Clinton 1993


Barack Obama 2009

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

# Word Tree: Word Sequences

# Word Tree: Word Sequences

# 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 as Data

# Taxonomy of Data Types (?)

1D (sets and sequences)
Temporal
2D (maps)
3D (shapes)
nD (relational)
Trees (hierarchies)
Networks (graphs)

Are there others?

The eyes have it: A task by data type taxonomy for information visualization [Shneiderman 96]

# Unstructured Text

Words have meanings and relations

Correlations: *Hong Kong, Puget Sound, Bay Area*

Order: *January, February, March, April, May, June*

Membership: *Tennis, Running, Swimming, Hiking, Piano*

Hierarchy: *Person > Applicant > Job Candidate, Submitter*

Antonyms & synonyms

# WordNet: Structure, Relations

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: applicant [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss)

## Noun

- S: (n) **applicant**, applier (a person who requests or seeks something such as assistance or employment or admission)
  - ***direct hyponym*** / *full hyponym*
    - S: (n) aspirant, aspirer, hopeful, wannabe, wannabee (an ambitious and aspiring young person)
    - S: (n) bidder (someone who makes an offer)
    - S: (n) claimant (someone who claims a benefit or right or title)
    - S: (n) job candidate (an applicant who is being considered for a job)
    - S: (n) material (a person judged suitable for admission or employment)
    - S: (n) petitioner, suppliant, supplicant, requester (one praying humbly for something)
    - S: (n) possible (an applicant who might be suitable)
    - S: (n) probable (an applicant likely to be chosen)
    - S: (n) submitter (someone who submits something (as an application for a job or a manuscript for publication etc.) for the judgment of others)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*

# Text Processing Pipeline

## Tokenization

Segment text into terms.

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

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

Entities? *Washington State, Seattle, U.S.A*

# Text Processing Pipeline

## Tokenization

Segment text into terms.

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

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

Entities? *Washington State, Seattle, U.S.A*

## Stemming

Group together different forms of a word.

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

Lemmatization? *goes, went, gone* → go

# Text Processing Pipeline

## Tokenization

Segment text into terms.

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

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

Entities? *Washington State, Seattle, U.S.A*

## Stemming

Group together different forms of a word.

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

Lemmatization? *goes, went, gone* → go

## 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, e.g., term counts

Aggregate into a document-term matrix

Document vector space model

# Document-Term Matrix

Each document is a vector of term weights

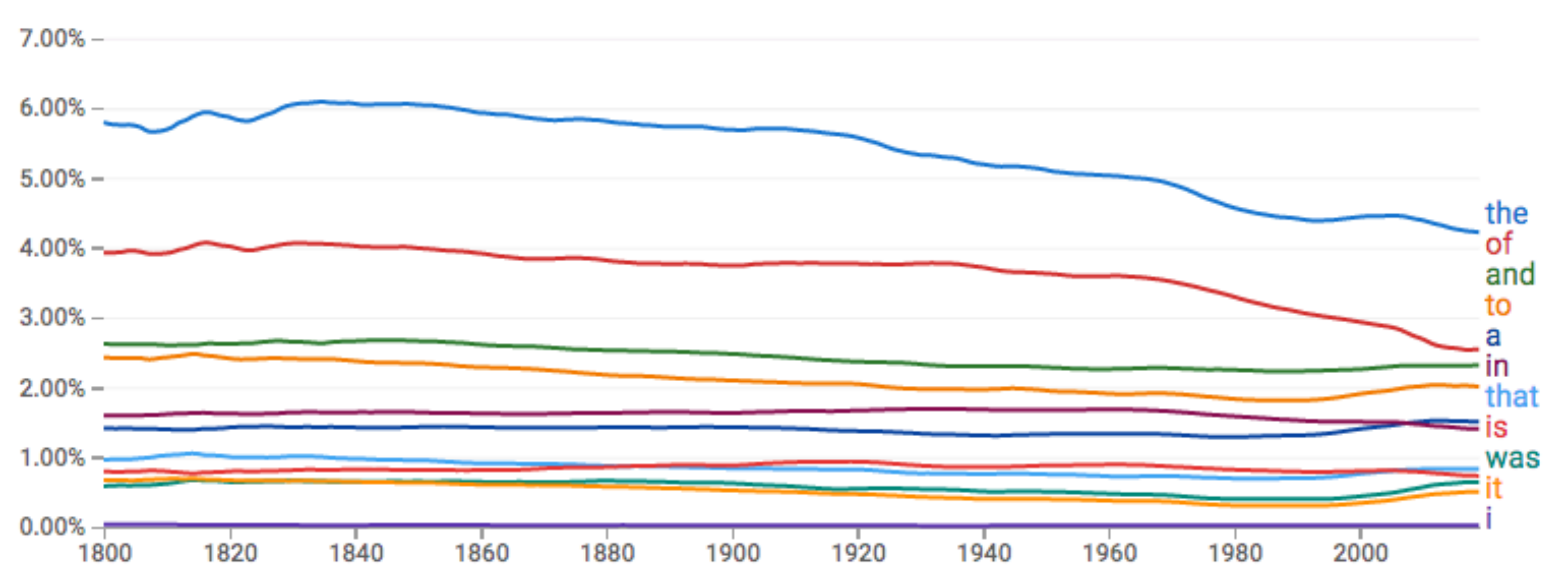
Simplest weighting is to just count occurrences

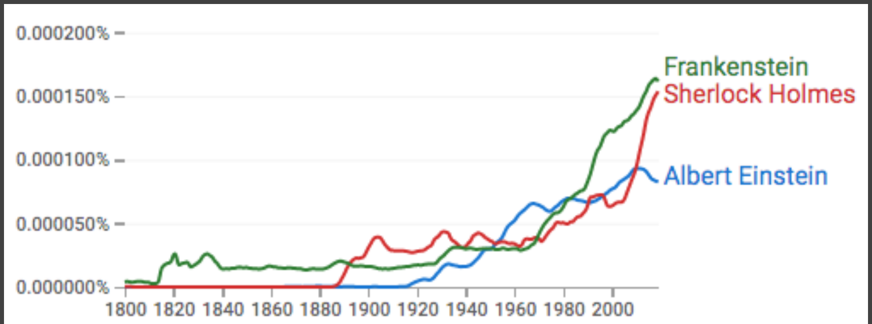
| | Antony and Cleopatra | Julius Caesar | The Tempest | Hamlet | Othello | Macbeth |
|---|---|---|---|---|---|---|
| Antony | 157 | 73 | 0 | 0 | 0 | 0 |
| Brutus | 4 | 157 | 0 | 1 | 0 | 0 |
| Caesar | 232 | 227 | 0 | 2 | 1 | 1 |
| Calpurnia | 0 | 10 | 0 | 0 | 0 | 0 |
| Cleopatra | 57 | 0 | 0 | 0 | 0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

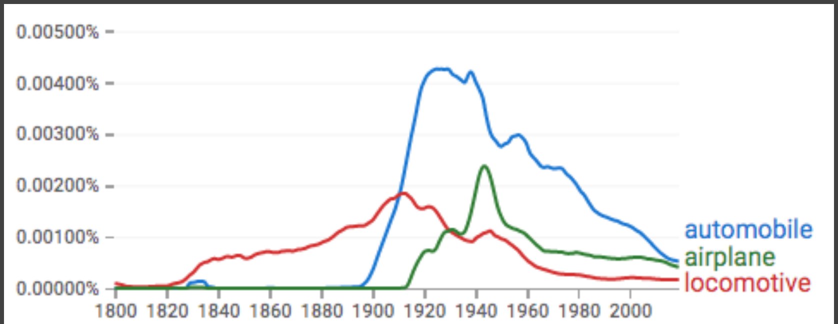
# WordCounts

[Harris '04]

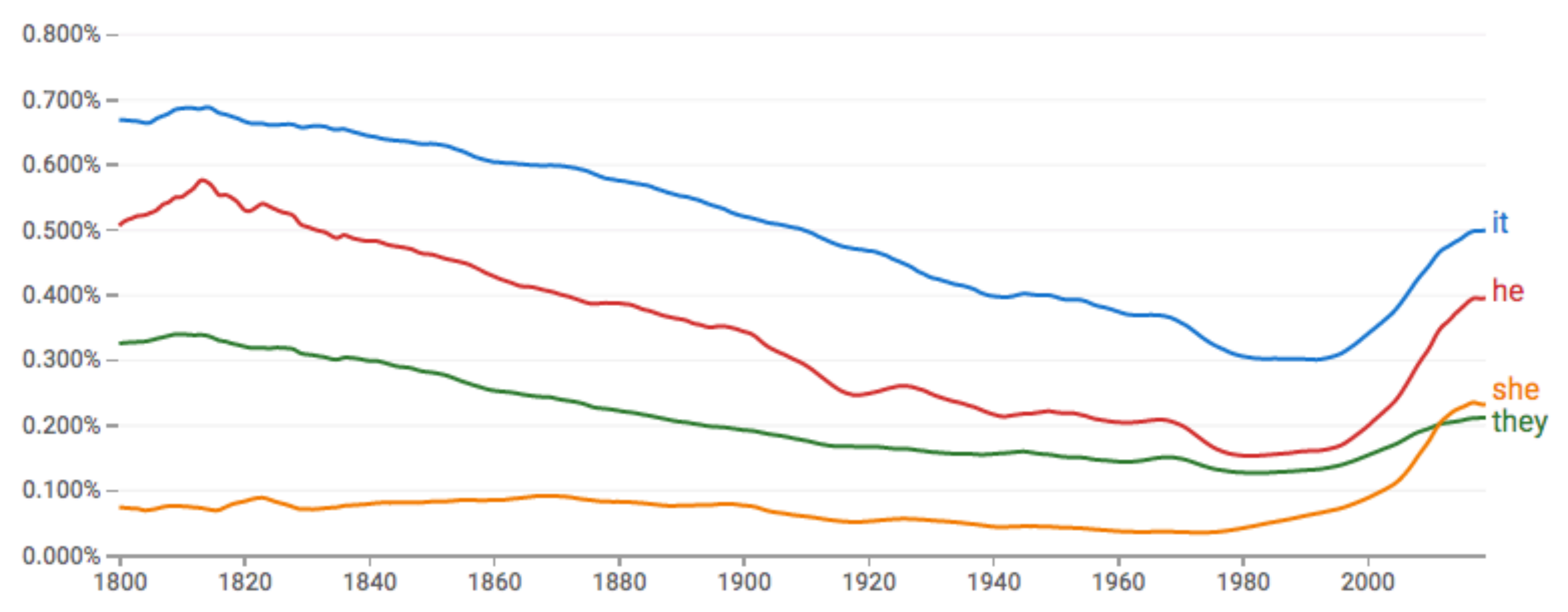
# Google Ngram Viewer



# Google Ngram Viewer

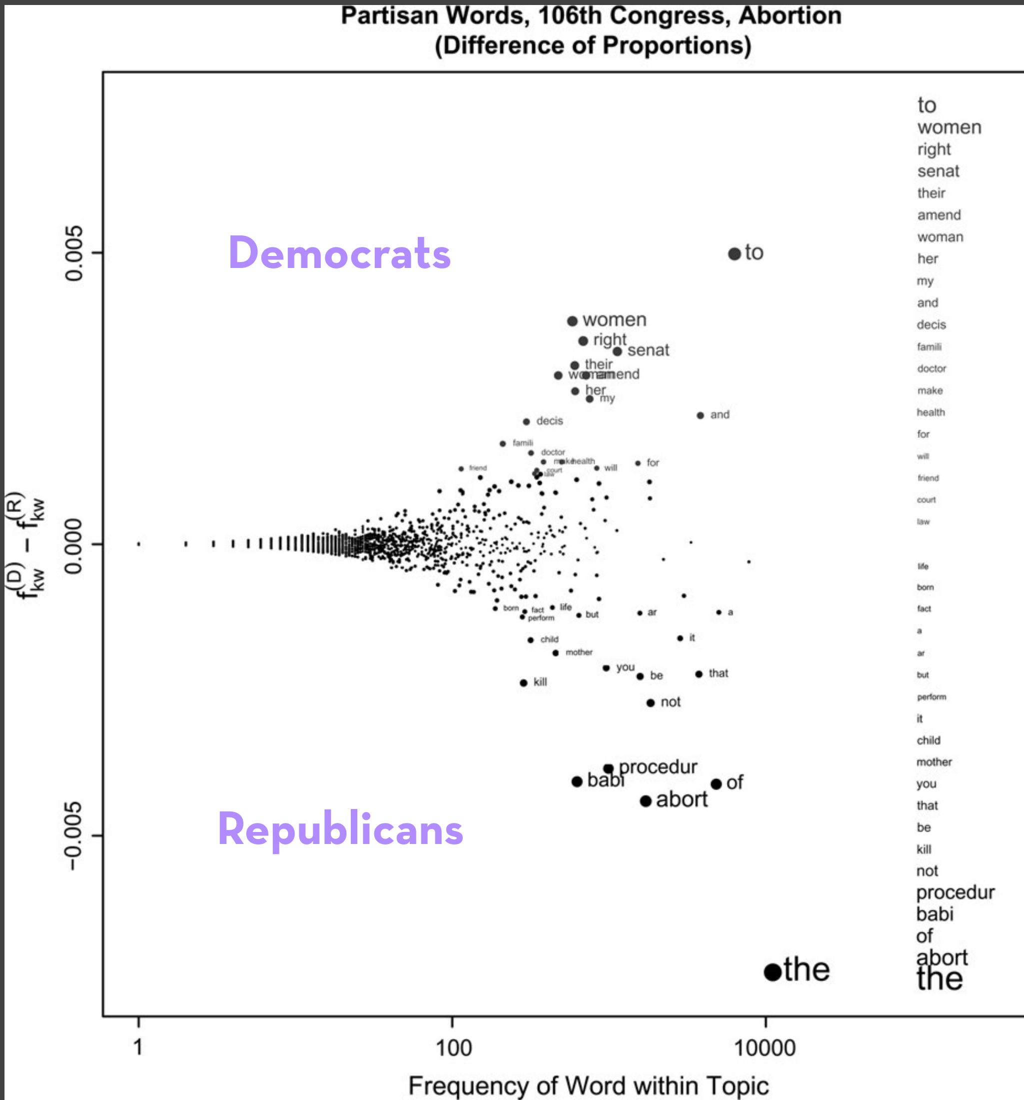


Given a text, what are the best descriptive words?

# Lexical Feature Selection [Monroe et al. '08]

Top 20 words labeled

Visualize proportion relative to the word frequency in overall document collection



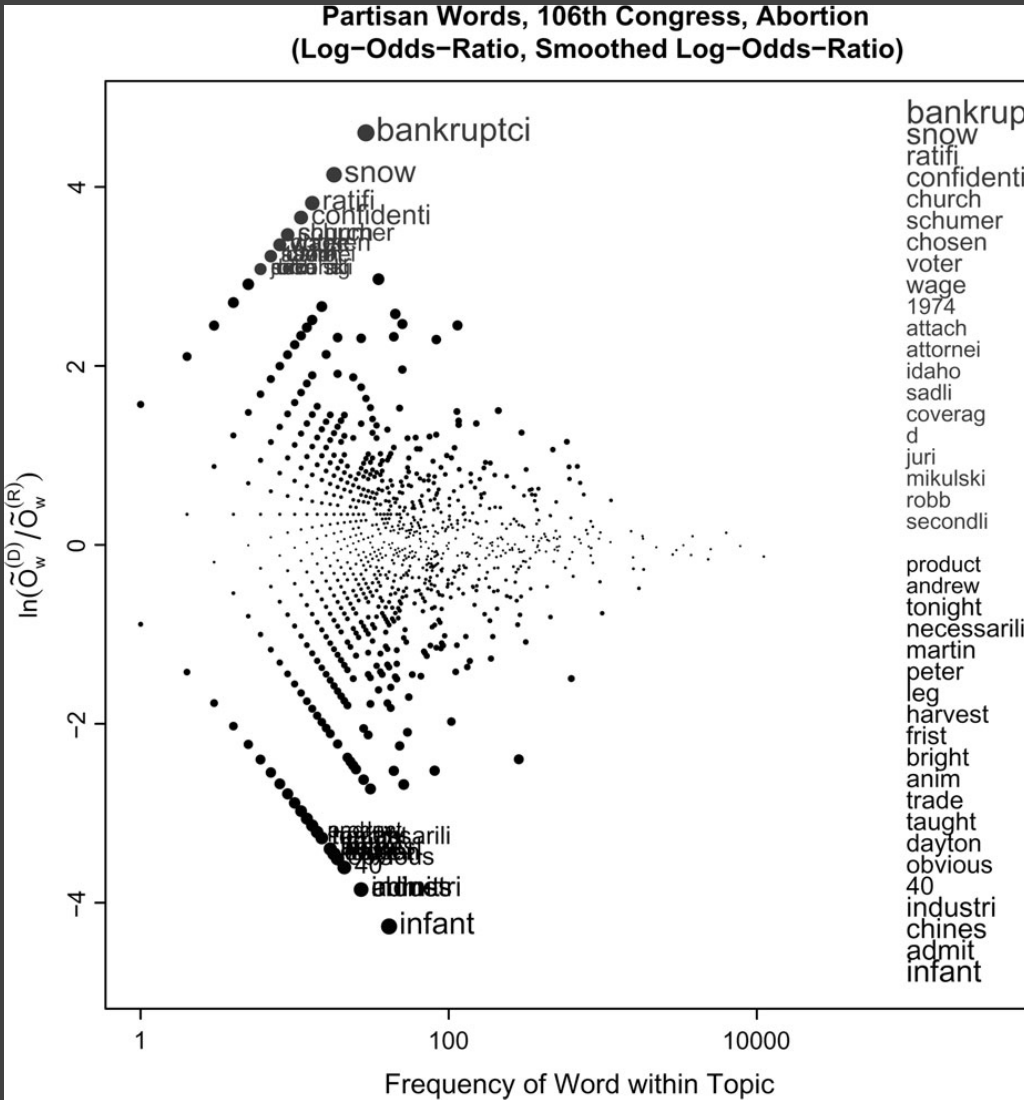
# Lexical Feature Selection [Monroe et al. '08]

Top 20 words labeled

Log-odds-ratio

Symmetric display between two parties

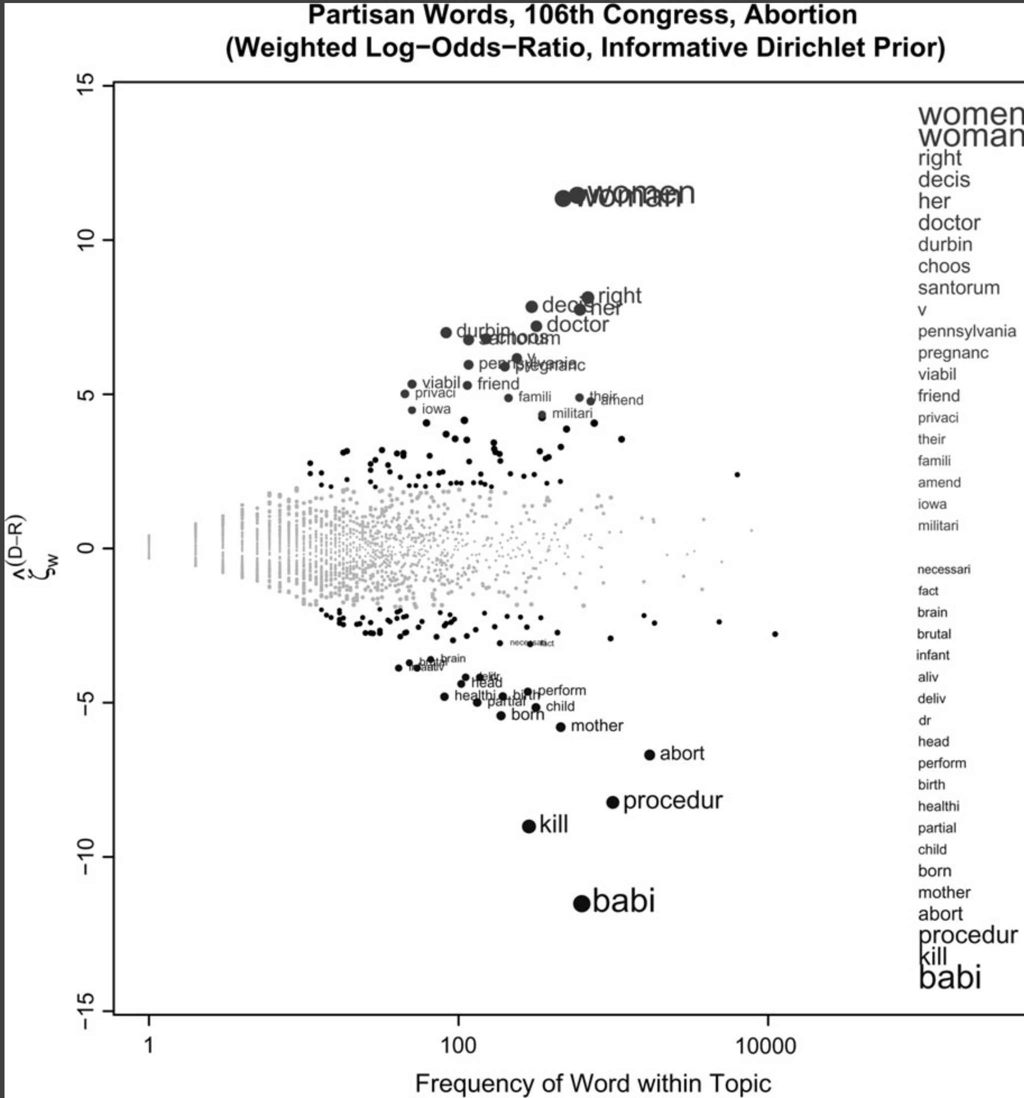
Words only spoken by a particular party (and not the other party)



# Lexical Feature Selection [Monroe et al. '08]

Top 20 words labeled

Leverage word priors: expected distribution of words (across many Senate topics)



# 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

# Bag of Words Model: Word or Tag Clouds

# 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 overview 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

# Size: Perceptual Biases [Alexander et al. '18]

| | Factor agreement | | | | | |
|---|---|---|---|---|---|---|
| Factor | agree | | neutral | | disagree | |
| word length | hello<br>sam | bigger font,<br>longer word | hello<br>world | same length | hello<br>goodbye | bigger font,<br>shorter word |
| word height | help<br>corn | bigger font,<br>taller word | plot<br>flop | same<br>"raw height" | corn<br>help | bigger font,<br>shorter word |
| word width | joyful<br>letter | bigger font,<br>wider word | litter<br>fillet | same<br>"raw width" | little<br>hummed | bigger font,<br>narrower word |

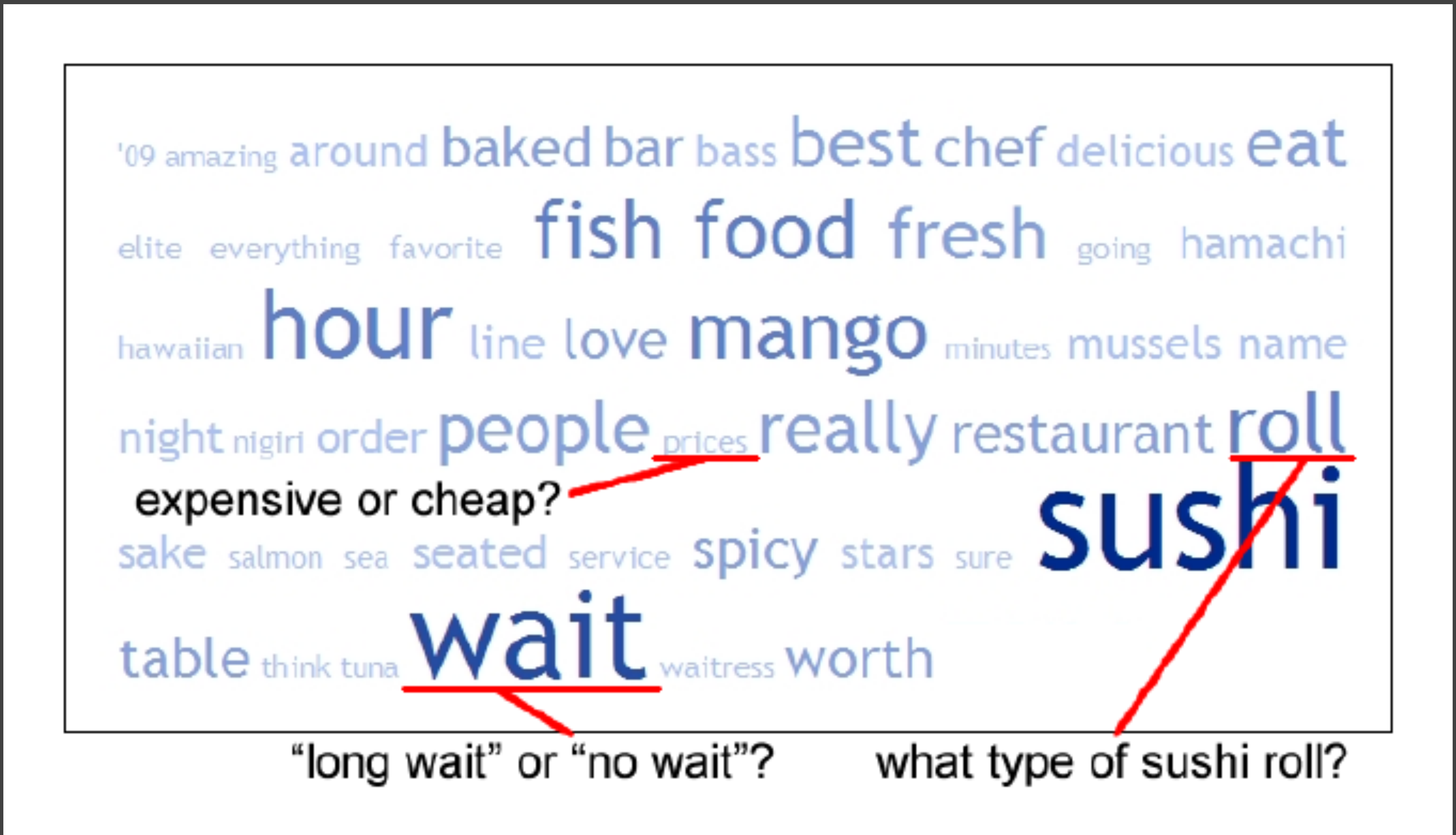
# Size: Perceptual Biases [Alexander et al. '18]

| Label | E/P | Effect of Δ font size | Primary bias factor | Effect of bias factor agreement | Additional factor | Accuracy at min Δ font size | | | Notes |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | agree | neutral | disagree | |
| len1 | P | ✓ | word length† | ✓ | - | 0.860 | 0.879 | 0.753 | Word length biases perception of font size |
| len2 | P | ✓ | word length† | ✓ | base font size‡ | 0.861 | 0.816 | 0.734 | We see a greater bias at larger base font (30 px versus 20 px) |
| len3 | P | ✓ | word length† | ✓ | base font size† | 0.825 | 0.838 | 0.642 | Tested wider variety of baseline font sizes |
| len4 | E | ✓ | word length† | ✓ | - | 0.992 | 0.942 | 0.867 | Bias still present with English words and denser word clouds |
| height1 | P | ✓ | word height† | ✓ | - | 0.974 | 0.909 | 0.684 | Character heights bias perception of font size |
| height2 | P | ✓ | word height† | ✓ | - | 0.929 | 0.810 | 0.529 | Proportional difference in font size seems to matter more than absolute difference |
| height3 | P | ✓ | word height† | ✓ | - | 0.937 | 0.795 | 0.525 | Bias still present when word clouds use sans serif font |
| height4 | P | ✓ | word height† | ✓ | base font size† | 0.931 | 0.790 | 0.479 | We see a greater bias at larger base font (30 px versus 20 px) |
| height5 | P | ✓ | word height† | ✓ | base font size‡ | 0.963 | 0.854 | 0.489 | Accuracy hits ceiling between 20-25% size difference |
| width1 | E | ✓ | word width† | ✓ | - | 0.975 | - | 0.909 | Bias present when length is held constant and width varies |
| width2 | E | × | word length† | × | - | 0.982 | - | 0.982 | No bias when width is held constant and length varies |
| box1 | E | ✓ | word width† | × | - | 0.914 | 0.932 | 0.908 | No bias with corrected-width rectangular bounding boxes |
| big1 | P | ✓ | word length† | ✓ | number of near misses | 0.888 | 0.826 | 0.658 | Tested using "pick the biggest word" task |
| big2 | P | ✓ | word length† | ✓ | number of near misses | 0.811 | - | 0.562 | Tested wider variety of length differences |

# Size: Perceptual Biases [Alexander et al. '18]

# Yelp Review Spotlight

[Yatani et al. '11]



# Yelp Review Spotlight

[Yatani et al. '11]

# Descriptive Phrases

Understand the limitations of your language model.

Bag of words: (1) easy to compute, (2) single words, (3) loss of order

Select appropriate model and visualization

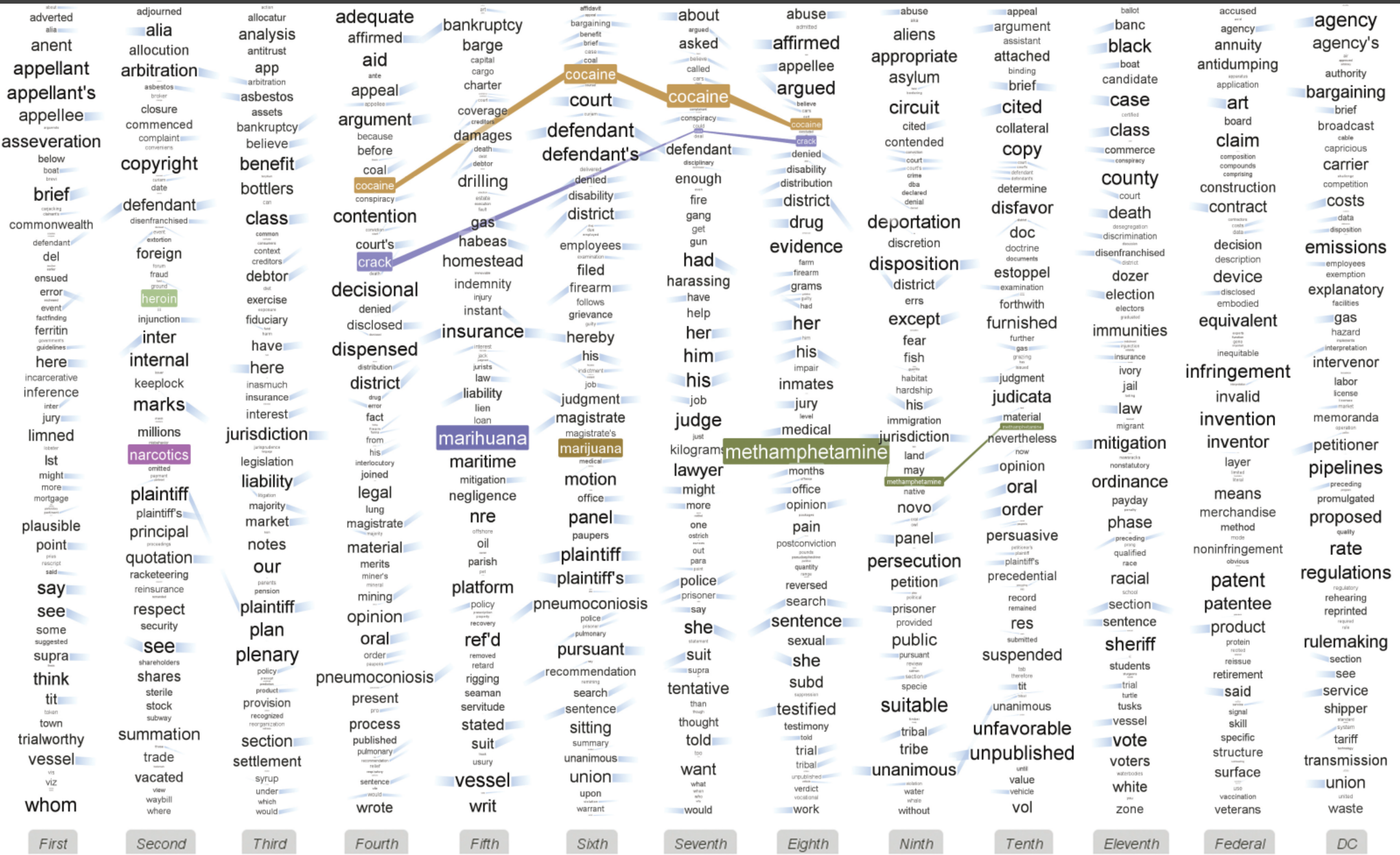
Generate longer, more meaningful phrases

Adjective-noun word pairs for reviews

Show keyphrases within source text

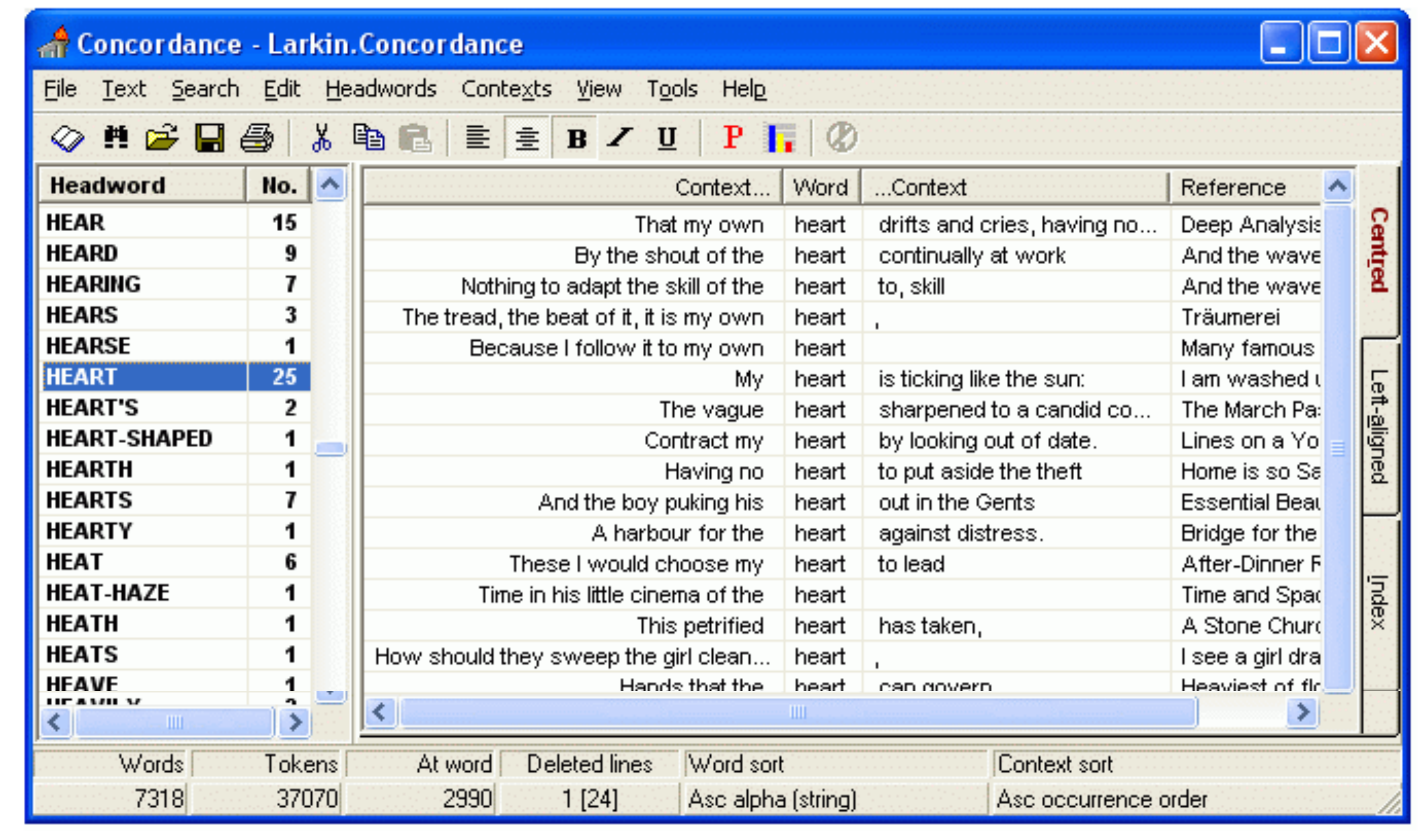
# Parallel Tag Clouds

[Collins et al. '09]



# Context and Structure

# Concordance



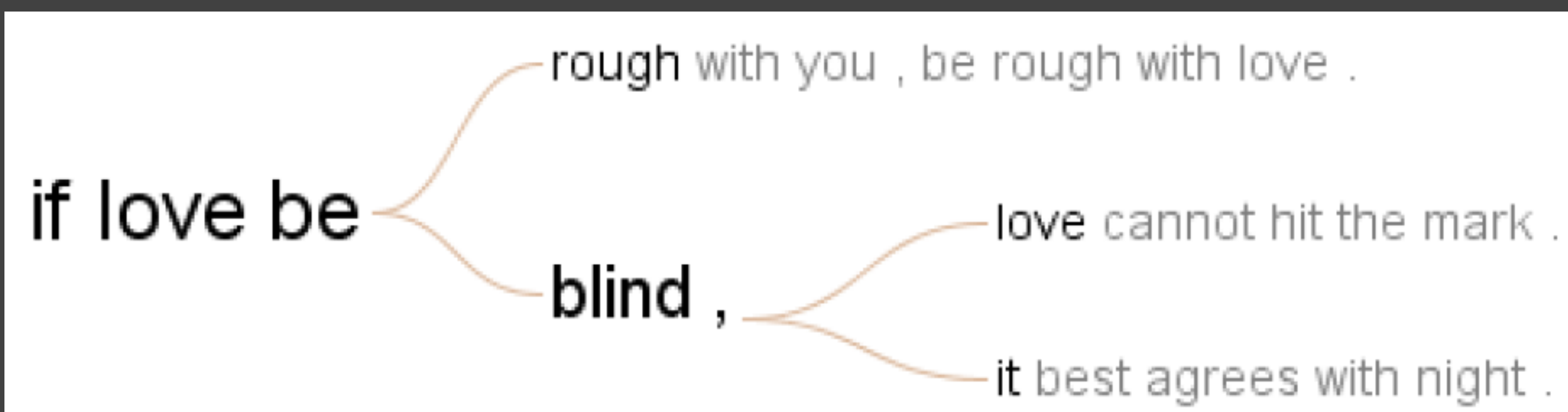
# Context & Structure

[Wattenberg et al. '08]

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 .

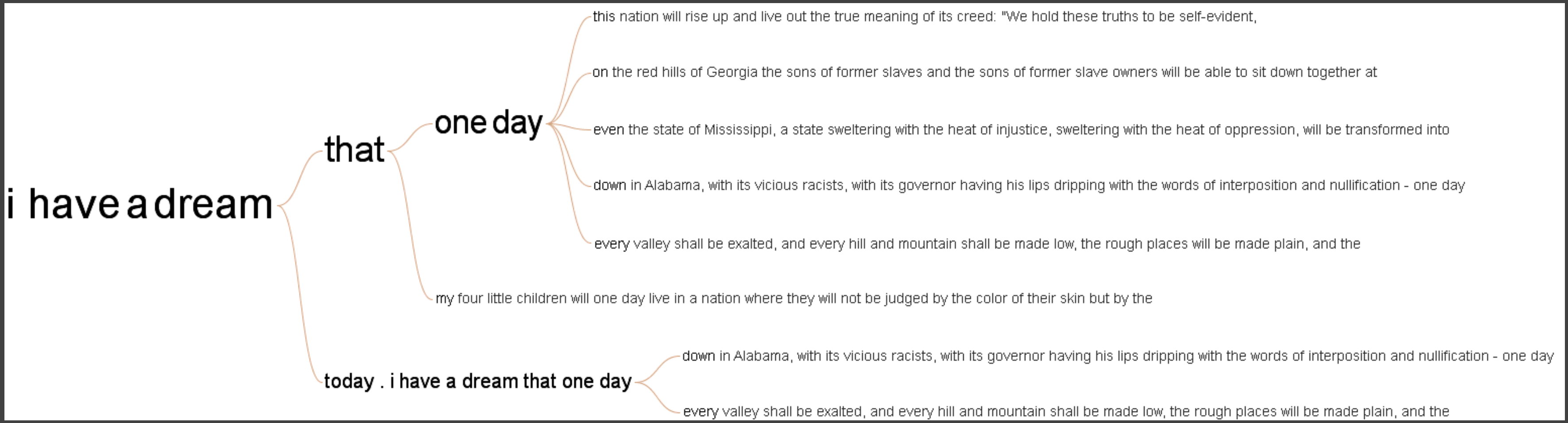


# Word Tree

[Wattenberg et al. '08]

Recurrent themes in speech structure

Visualization of all occurrences of "I have a dream" in Martin Luther King's historic speech:

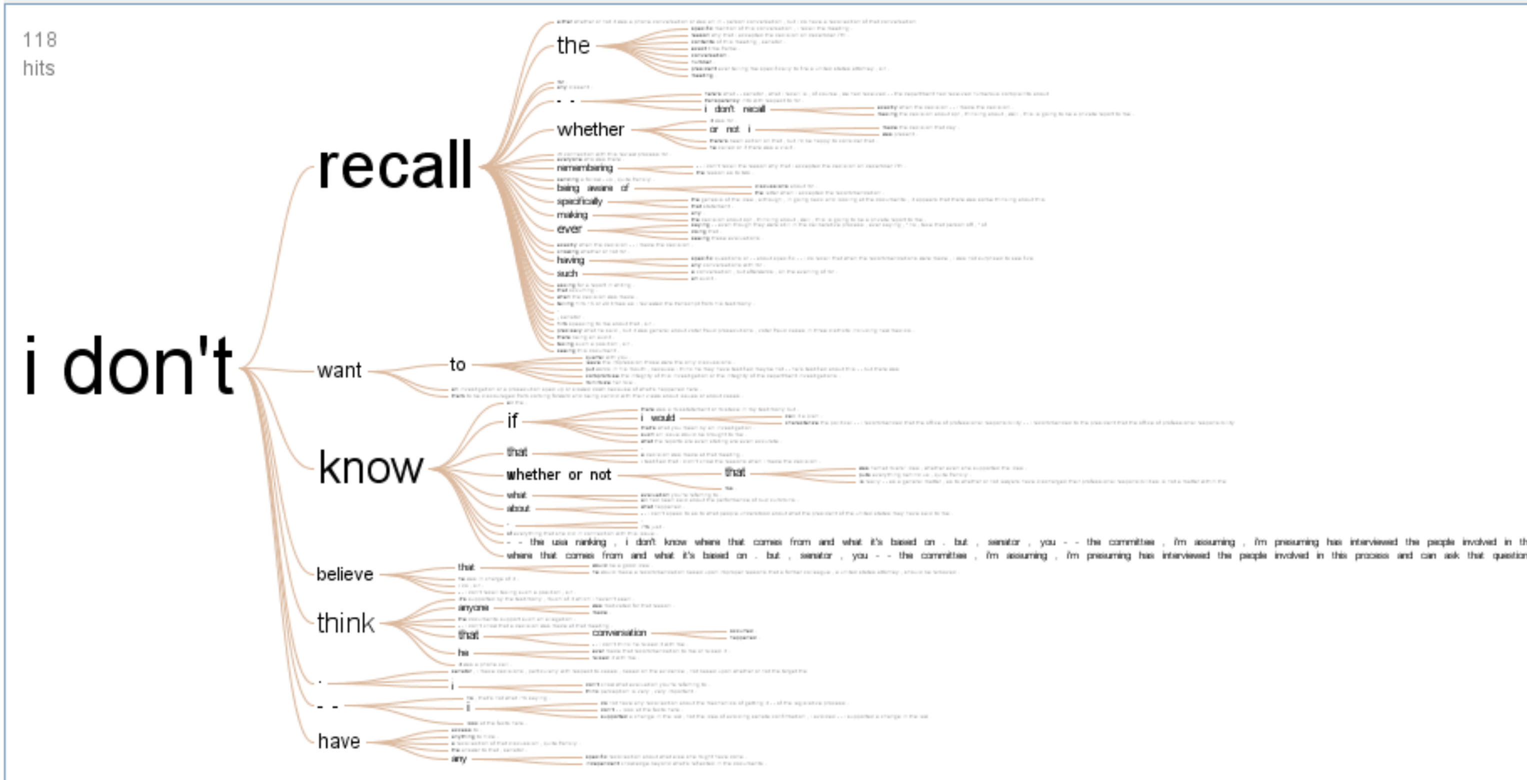


# Visualizations : Word tree / Alberto Gonzales

**Creator:** Martin Wattenberg

**Tags:**

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

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

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. '09]

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

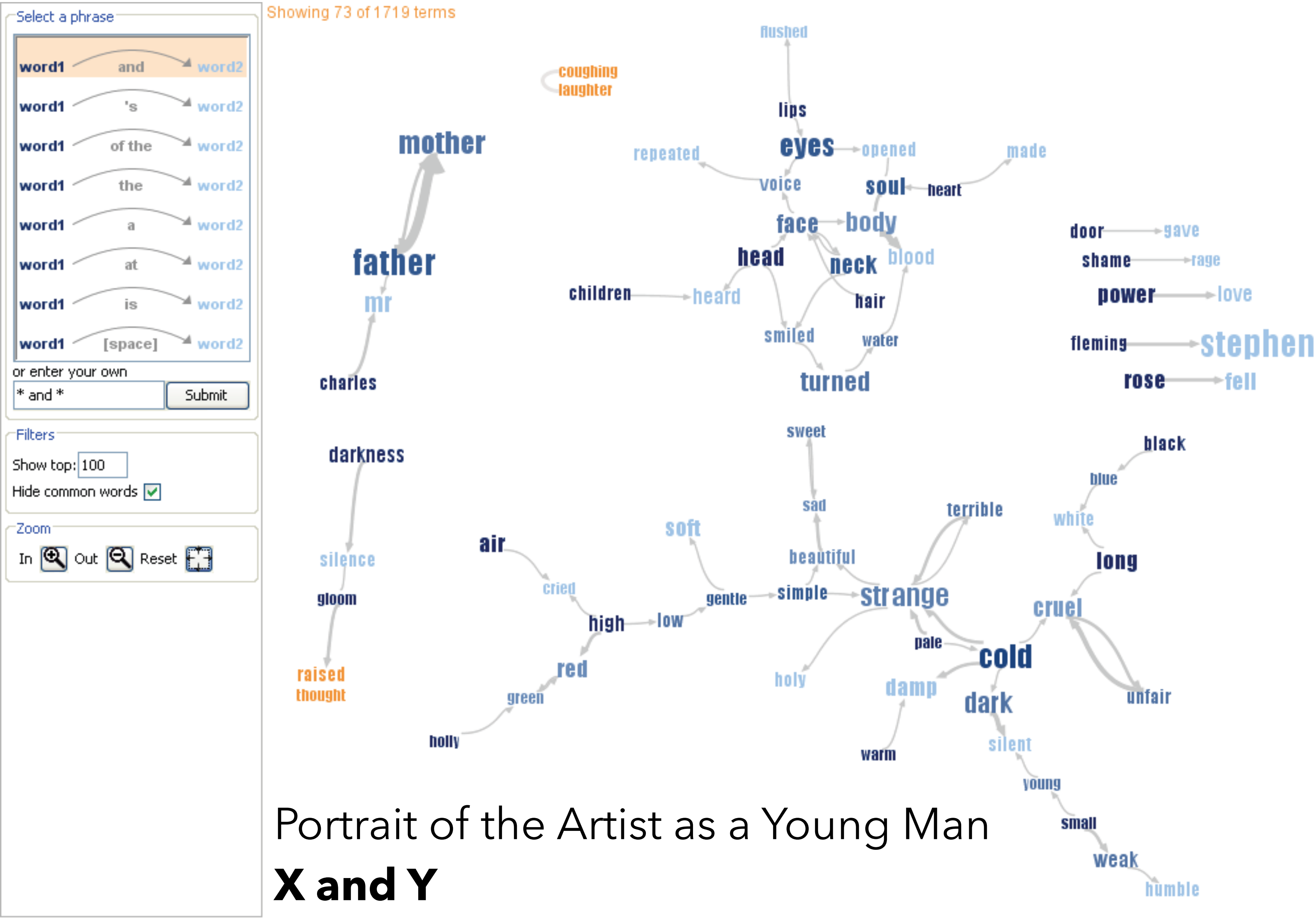
"A and B", "A at B", "A of B", etc.

Could be output of regexp or parser.

Visualize patterns in a node-link view:

Occurrences → Node size

Pattern position → Edge direction

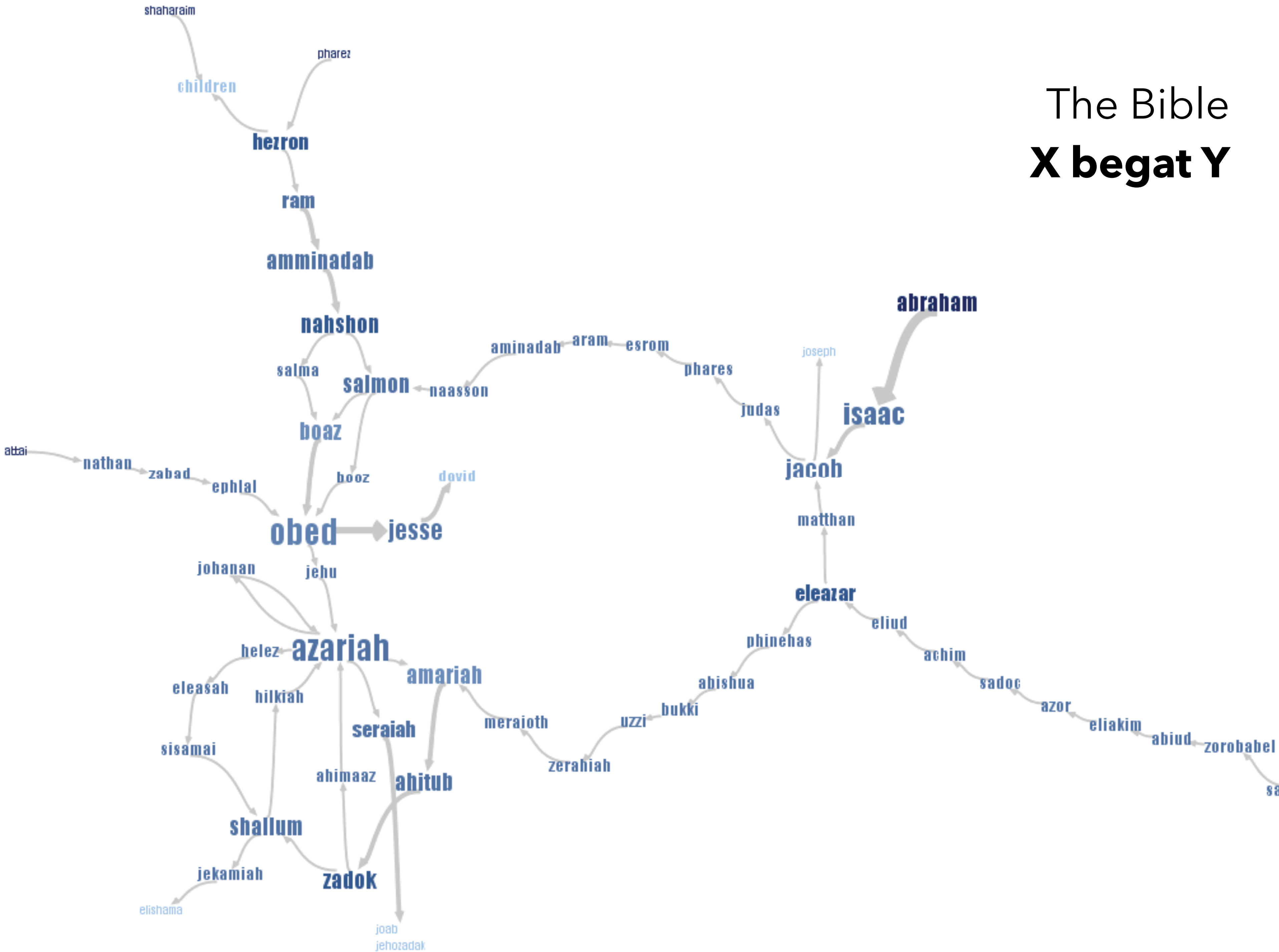


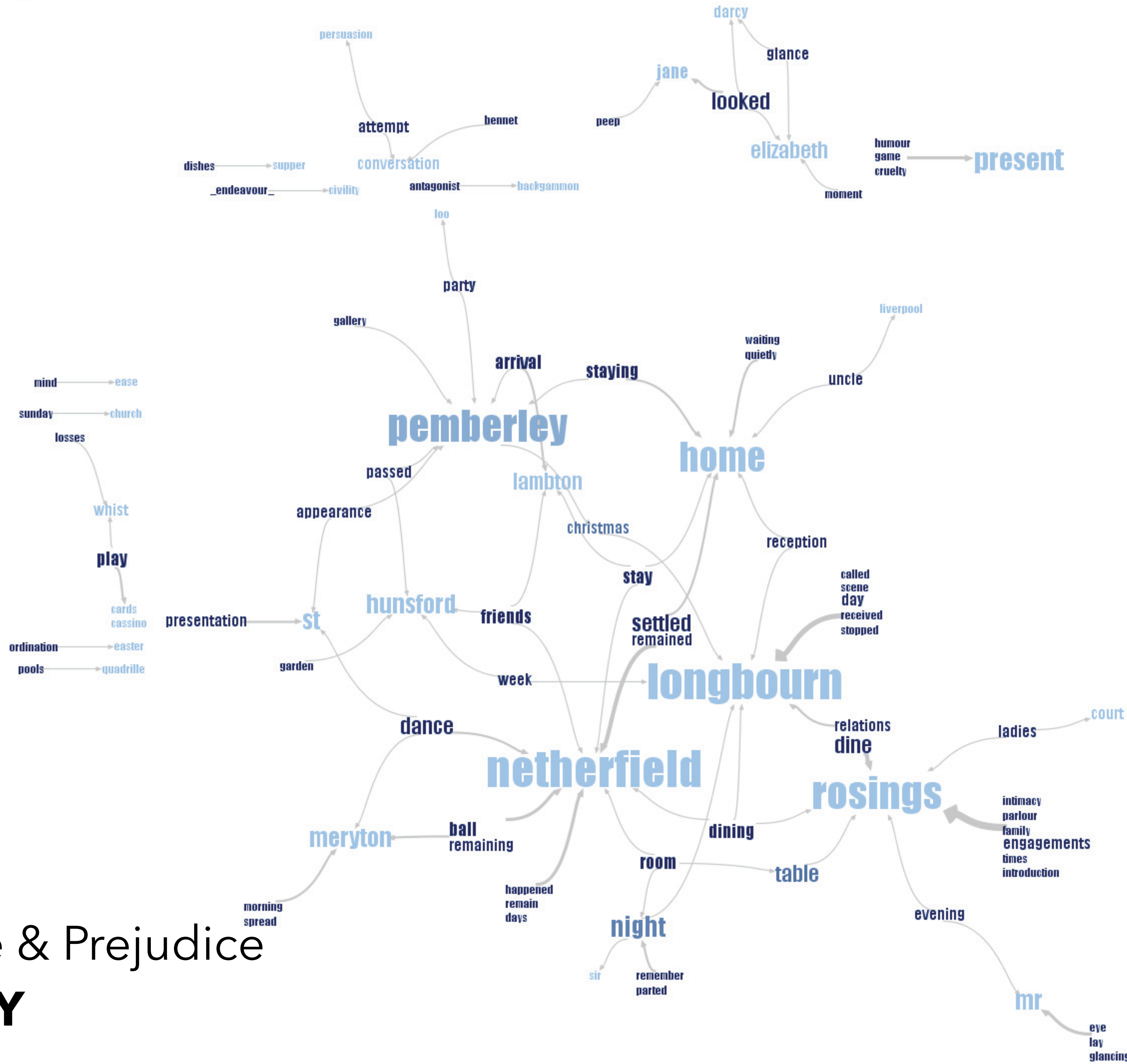

# Portrait of the Artist as a Young Man

## X and Y

The Bible

**X begat Y**



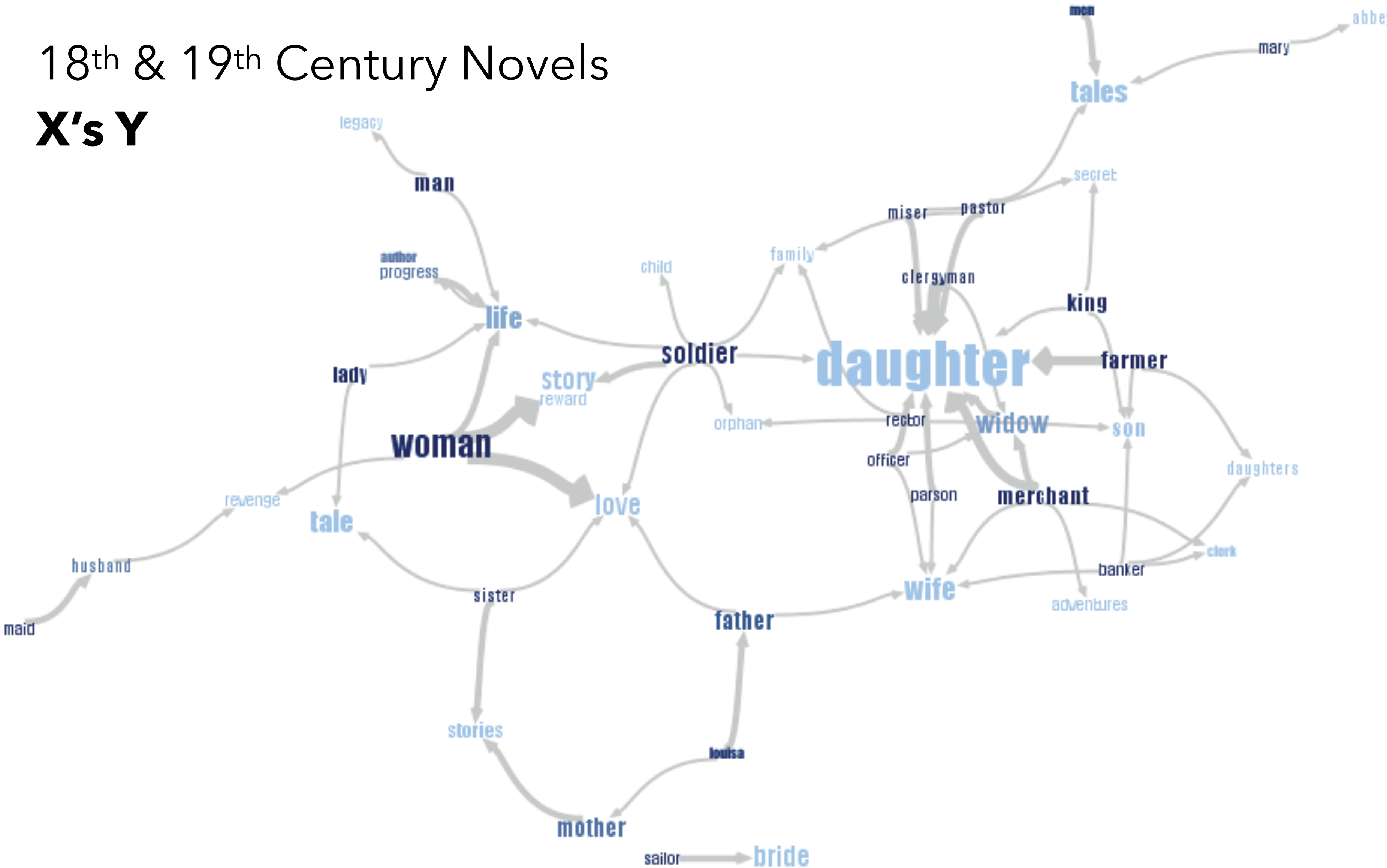



Pride & Prejudice

**X at Y**

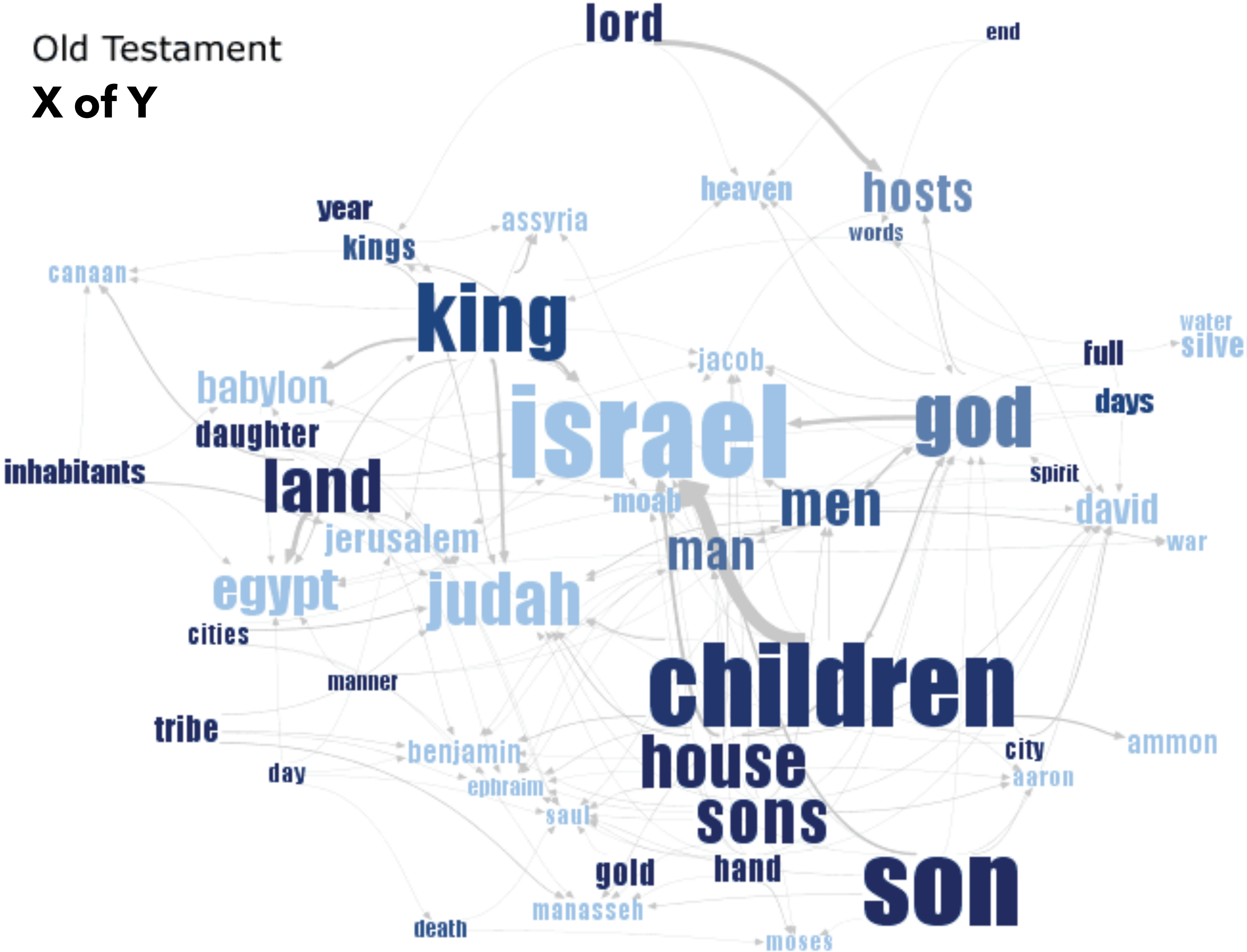
# 18th & 19th Century Novels

**X's Y**



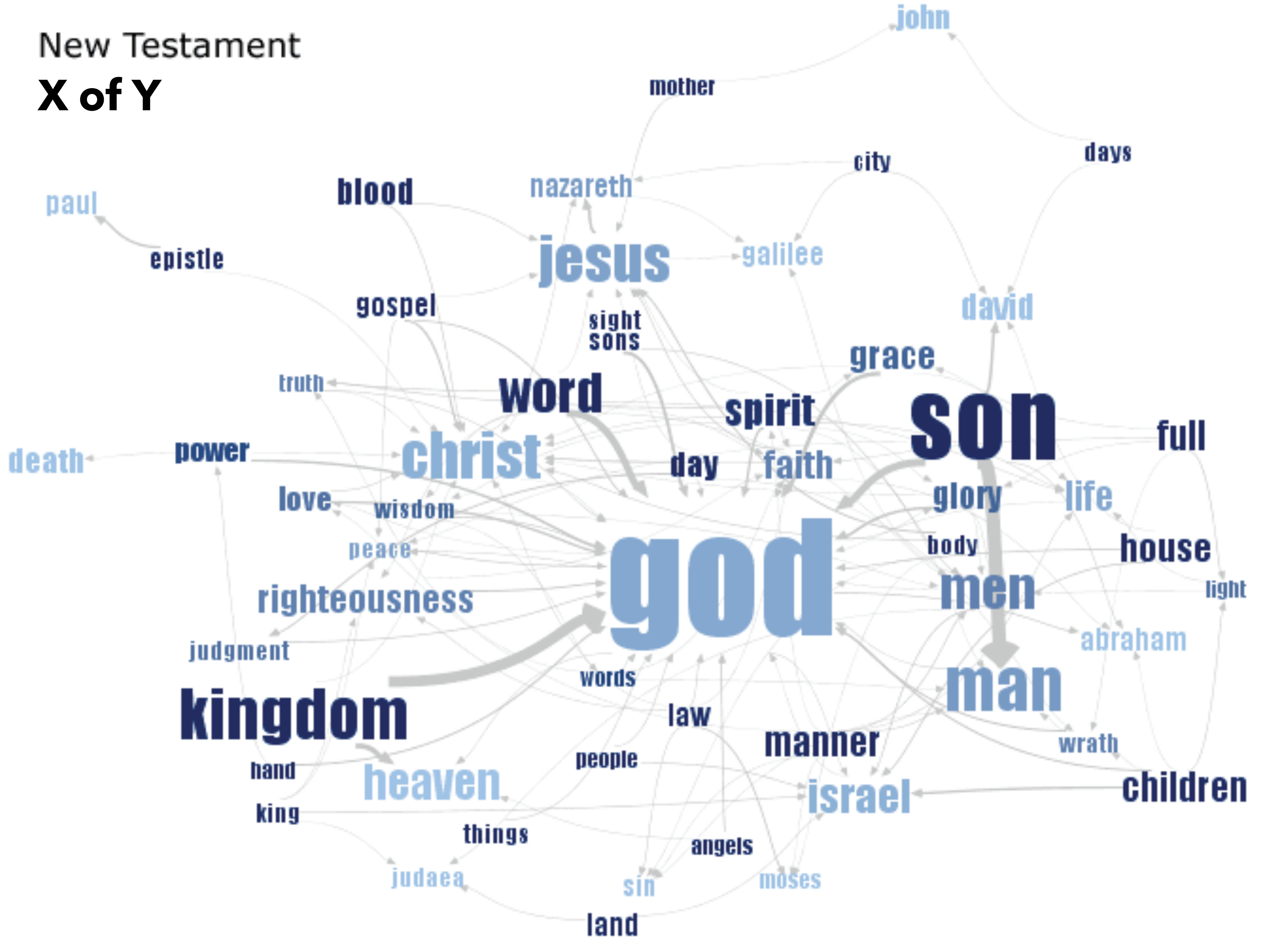
Old Testament

**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

# Document Collections

# Named Entity Recognition

**Label named entities in text:**

John Smith -> PERSON

Soviet Union -> COUNTRY

353 Serra St -> ADDRESS

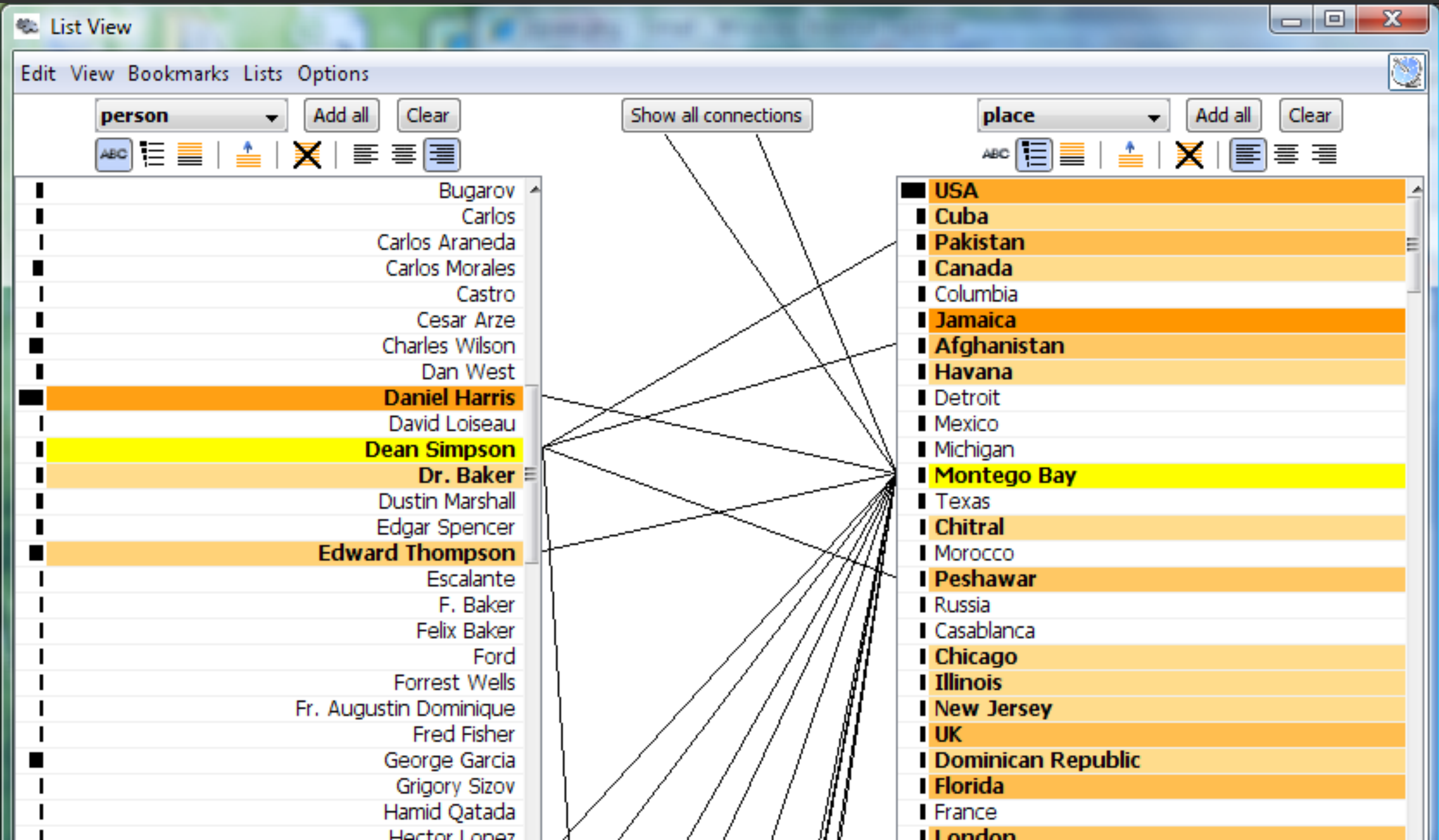
(555) 721-4312 -> PHONE NUMBER

Entity relations: how do the entities relate?

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

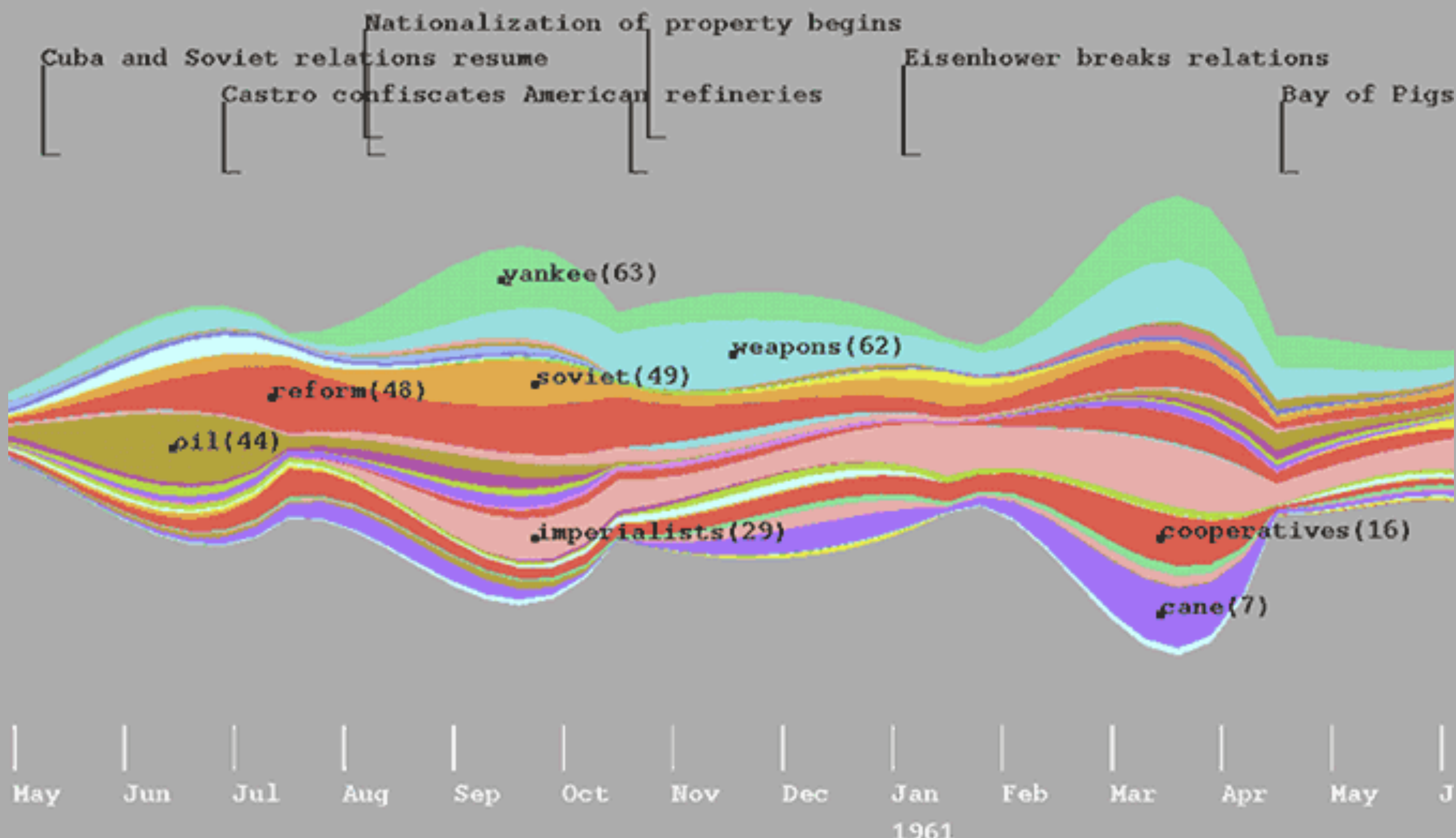
# Entity Relationships

[Görg et al. '07]



# Theme River

[Havre et al. '00]



# Similarity & Clustering

**Compute vector distance among docs**

Similarity measure can be used to cluster

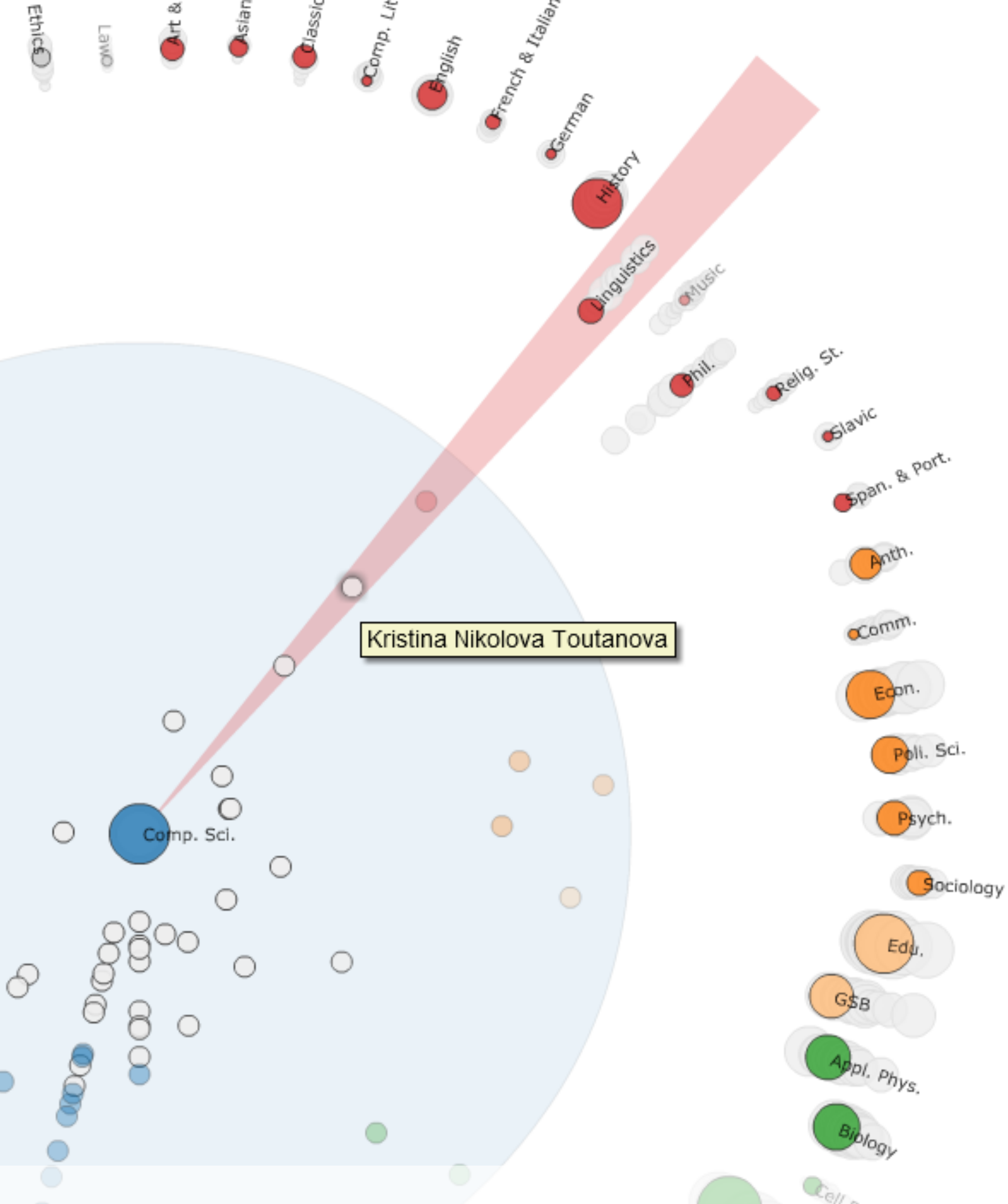
**Topic modeling**

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:

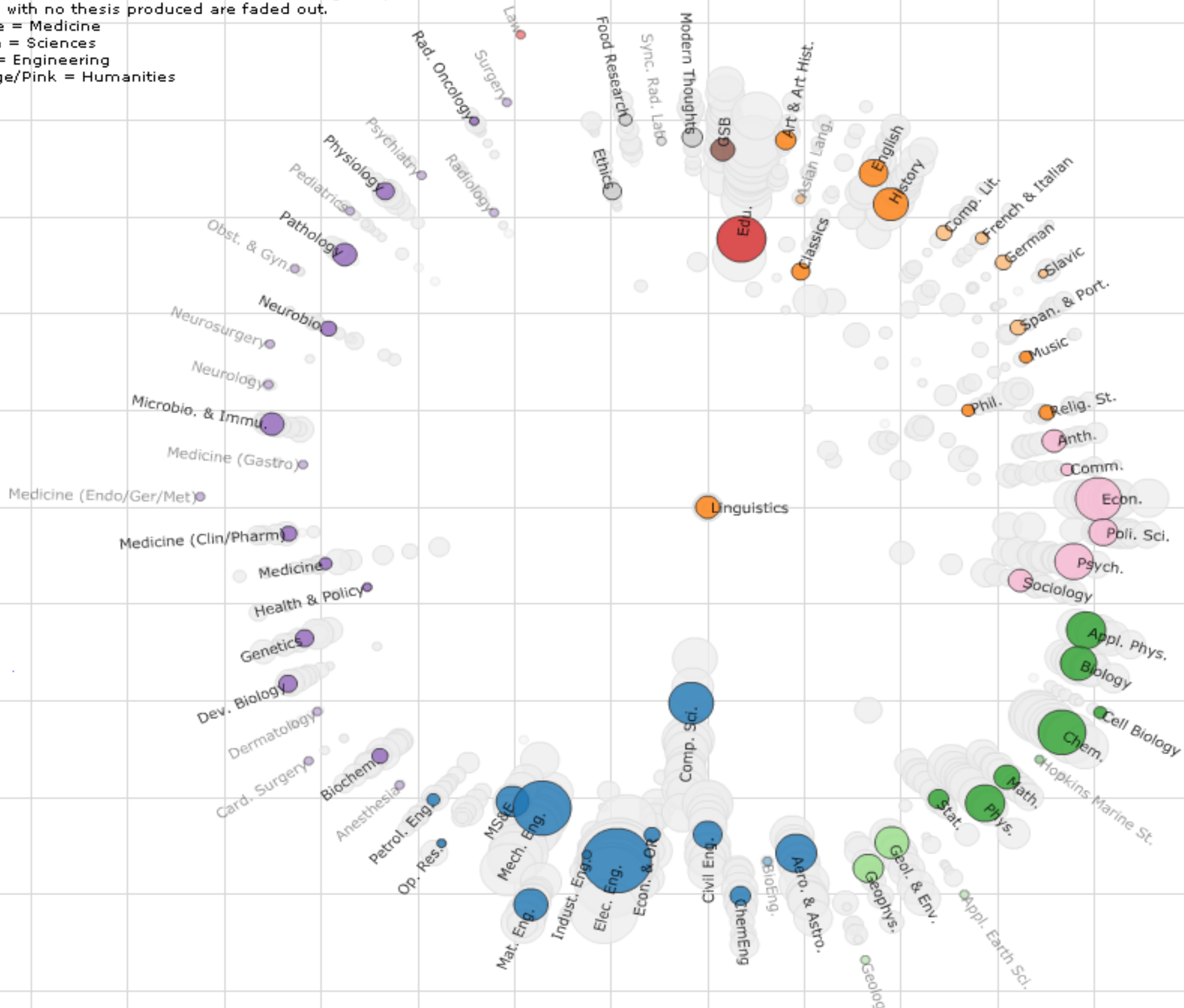
This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing 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

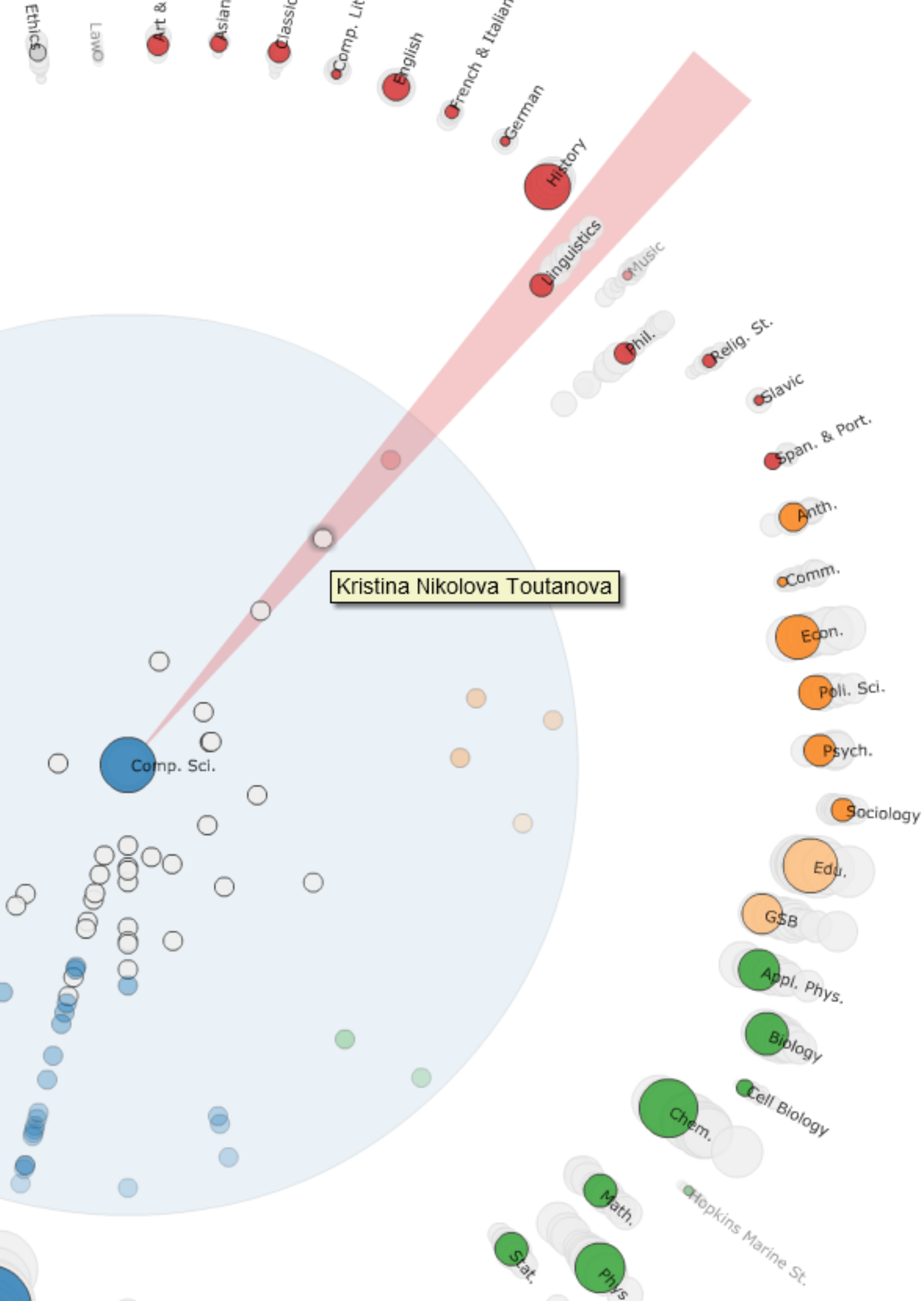
Jason Chuang, Dan Ramage, Christopher Manning, Jeffrey Heer

# 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**

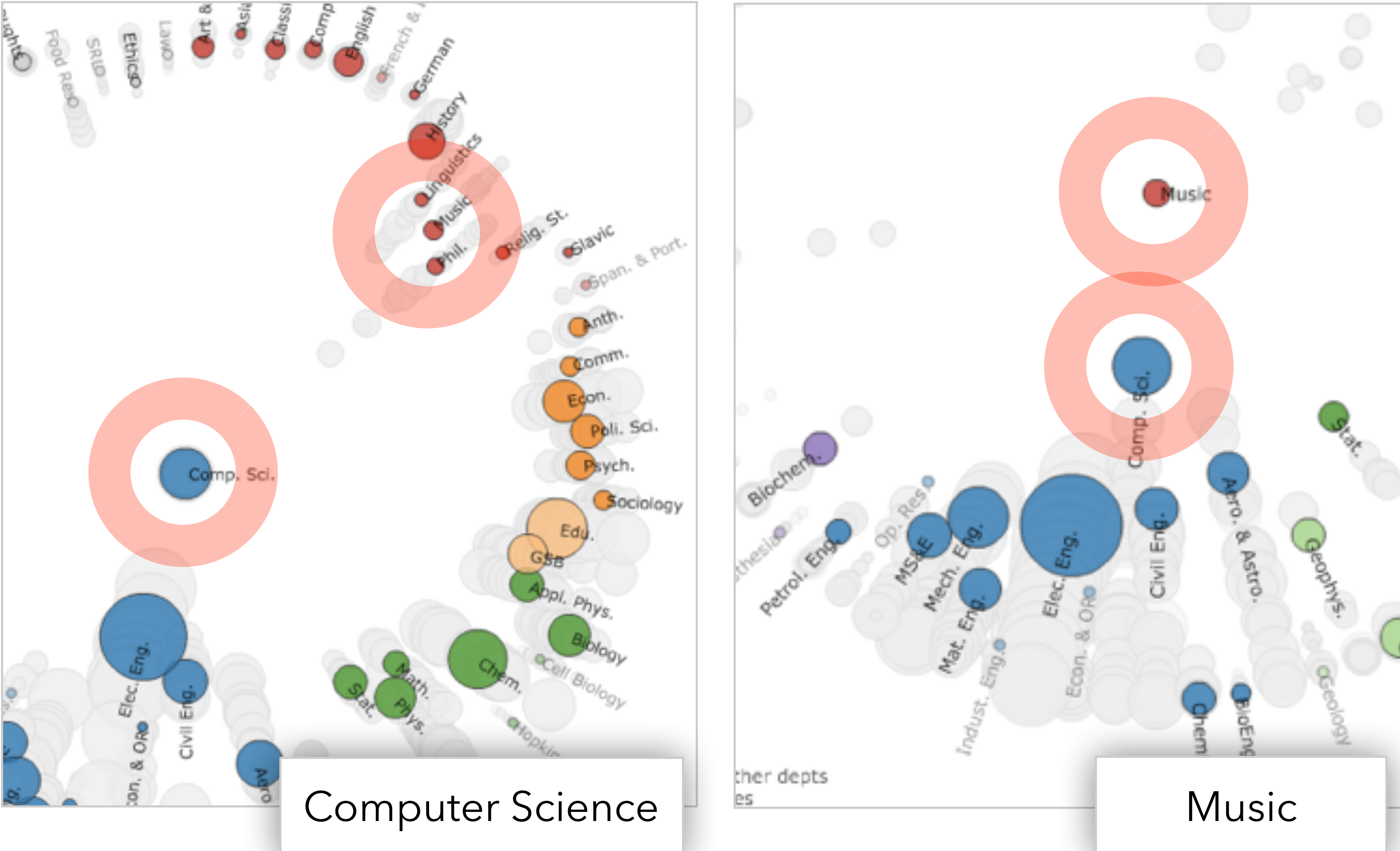
Student: Kristina Nikolova Toutanova
Advisor: Christopher D. Manning

Computer Science (2005)

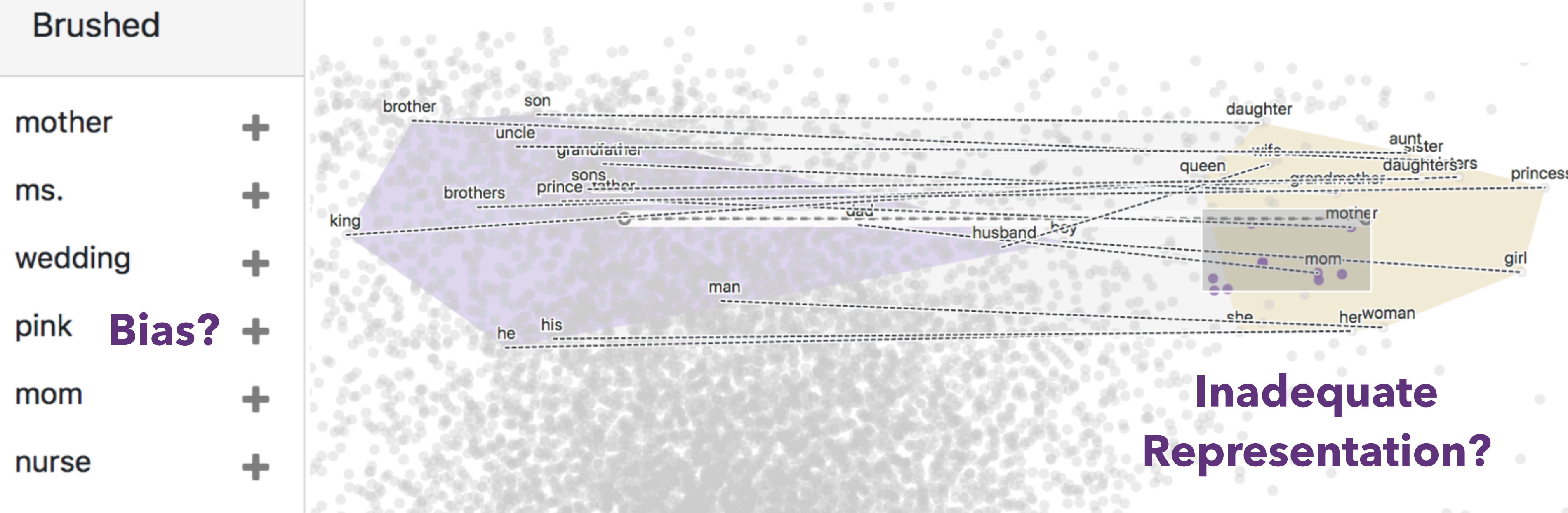
Keywords: Syntactic, Semantic, Tree kernels, Parsing

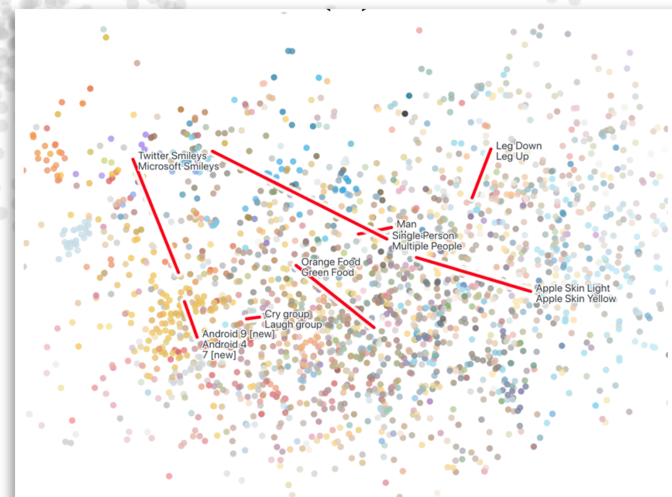

Abstract:

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.




# “Word Borrowing” via Labeled LDA




# Latent Space Cartography

## Visual Analysis of Vector Space Embeddings

Yang Liu, Eunice Jun, Qisheng Li (CSE 512, Spring '18)

# 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*