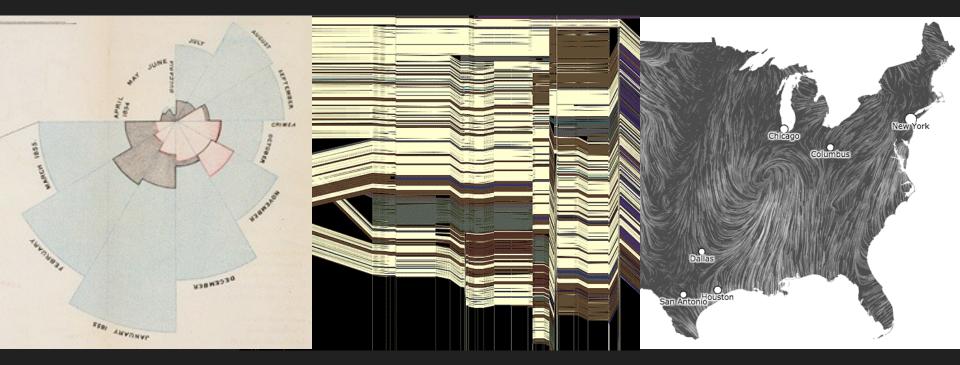
## CSE 412 - Intro to Data Visualization Text Visualization



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

### **Obama's Health Care Speech to Congress**

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

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

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

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

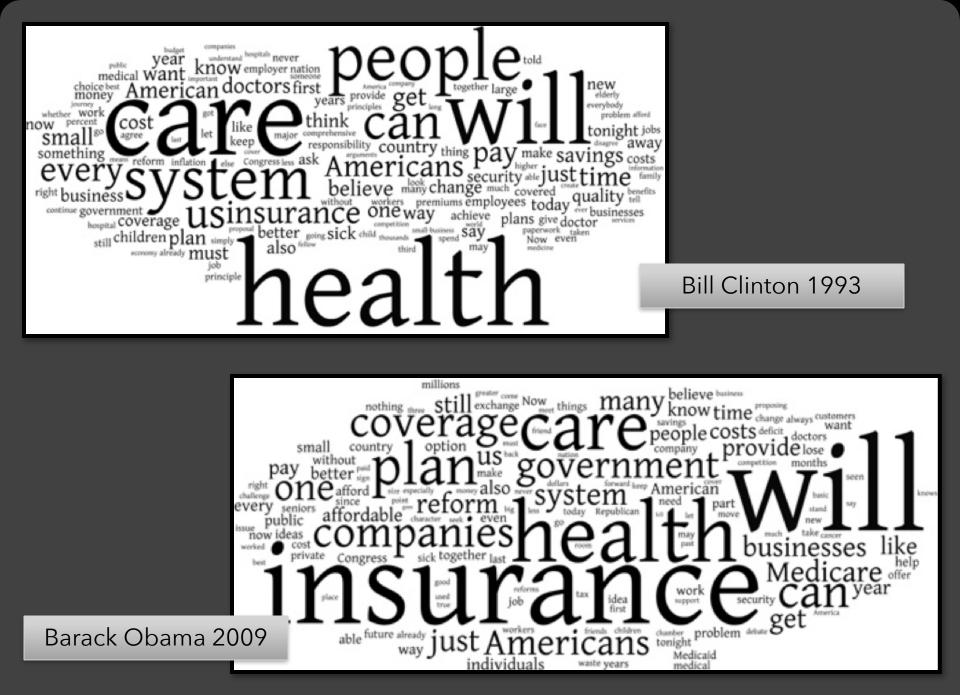
I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for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

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

## Tag Clouds: Word Count

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

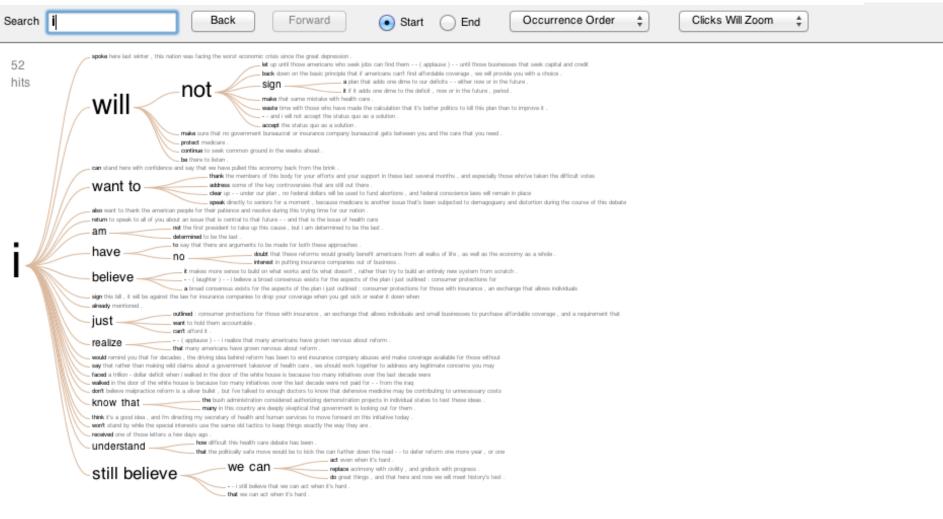




economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

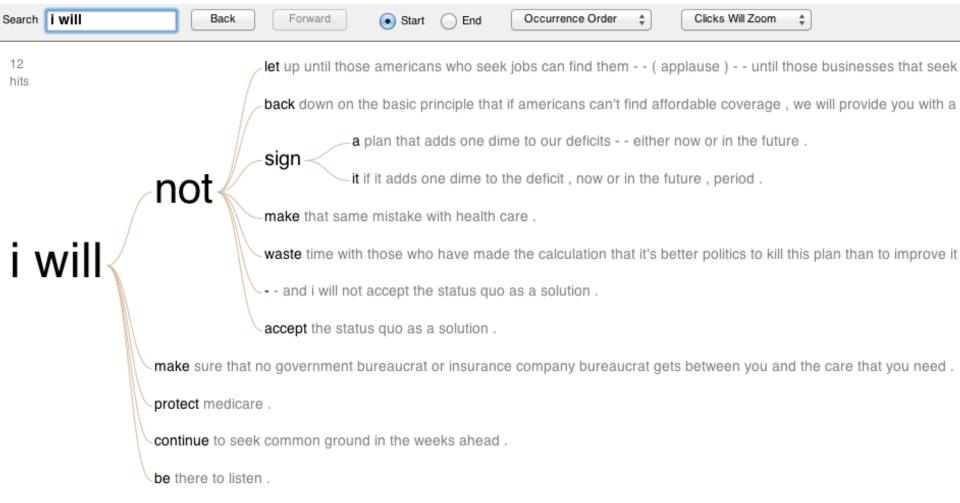
### Word Tree: Word Sequences

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



### Word Tree: Word Sequences

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



## **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) V Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations Display options for sense: (gloss)

#### Noun

- <u>S:</u> (n) applicant, <u>applier</u> (a person who requests or seeks something such as assistance or employment or admission)
  - direct hyponym / full hyponym
    - <u>S:</u> (n) <u>aspirant</u>, <u>aspirer</u>, <u>hopeful</u>, <u>wannabe</u>, <u>wannabee</u> (an ambitious and aspiring young person)
    - <u>S:</u> (n) <u>bidder</u> (someone who makes an offer)
    - <u>S:</u> (n) <u>claimant</u> (someone who claims a benefit or right or title)
    - <u>S:</u> (n) job candidate (an applicant who is being considered for a job)
    - <u>S:</u> (n) <u>material</u> (a person judged suitable for admission or employment)
    - <u>S:</u> (n) petitioner, suppliant, supplicant, requester (one praying humbly for something)
    - <u>S:</u> (n) possible (an applicant who might be suitable)
    - <u>S:</u> (n) probable (an applicant likely to be chosen)
    - <u>S:</u> (n) <u>submitter</u> (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

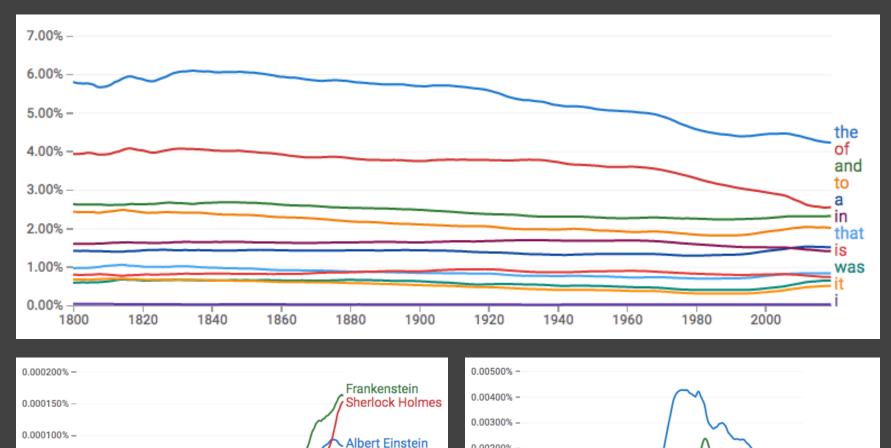
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

### WordCounts

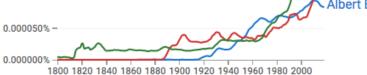


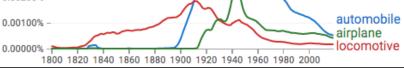
		WORDCOUNT
PREVIOUS WORD		NEXT WORD
the <sub>o</sub> fandtoair	] <b>țhat</b> itiswasi foronyouhebewithasyydavauratisodaulaatistaataa	n a Maringeren a serie de la marine de la mari
CURRENT WORD		
FIND WORD: BY RANK:	REQUESTED WORD: THE RANK: 1	86800 WORDS IN ARCHIVE ABOUT WORDCOUNT

### **Google Ngram Viewer**

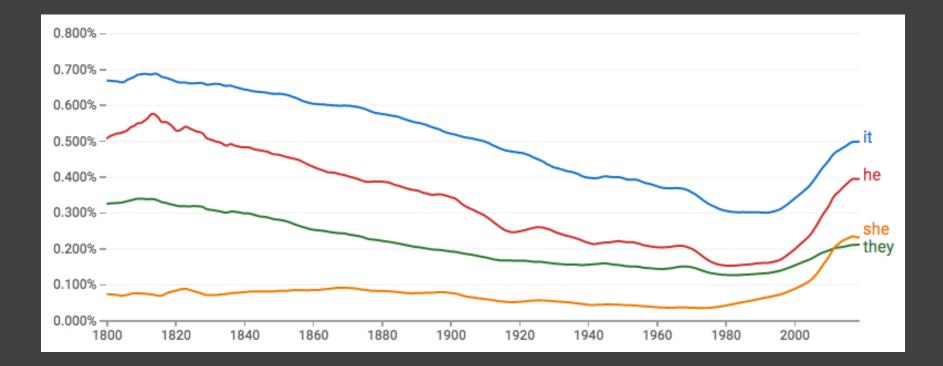


0.00200% -





### Google Ngram Viewer

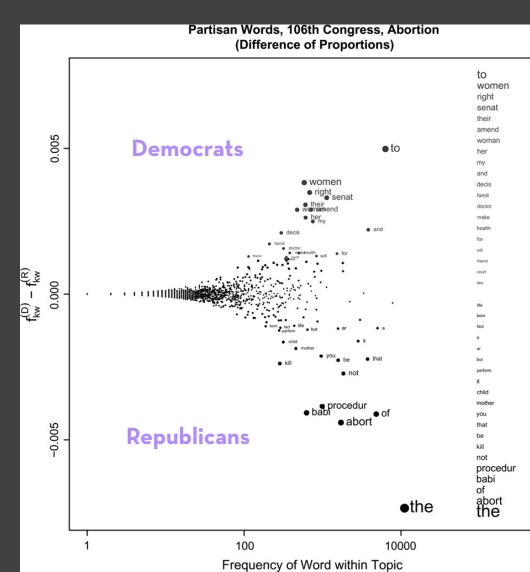


Given a text, what are the best descriptive words?

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

Top 20 words labeled

Visualize proportion relative to the word frequency in overall document collection



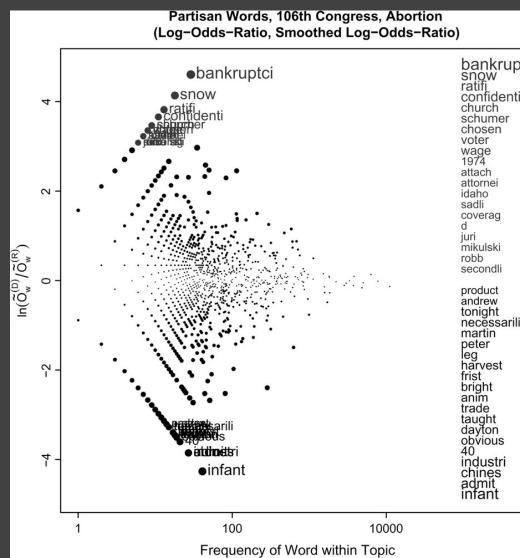
### Lexical Feature Selection [Monroe et al. '08]

Top 20 words labeled

Log-odds-ratio

Symmetric display between two parties

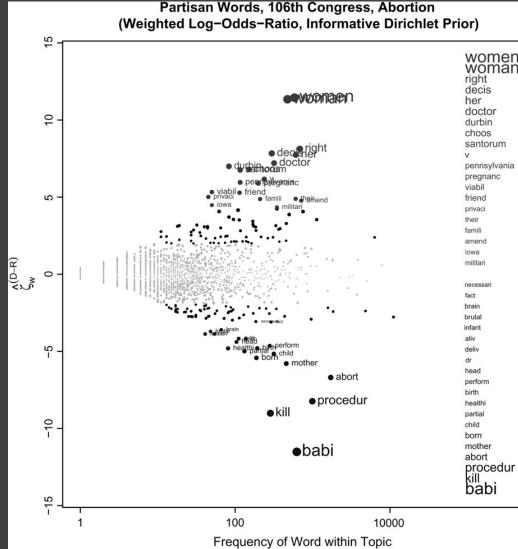
Words only spoken by a particular party (and not the other party)



### 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 숮 Inis 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		neu	utral	disagree		
word length	hello sam	bigger font, longer word	hello world	same length	hello goodbye	bigger font, shorter word	
word height	help corn	bigger font, taller word	plot flop	same "raw height"	corn <sub>help</sub>	bigger font, shorter word	
word width joyfu		bigger font, wider word	litter	same "raw width"	little	bigger font, narrower word	

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

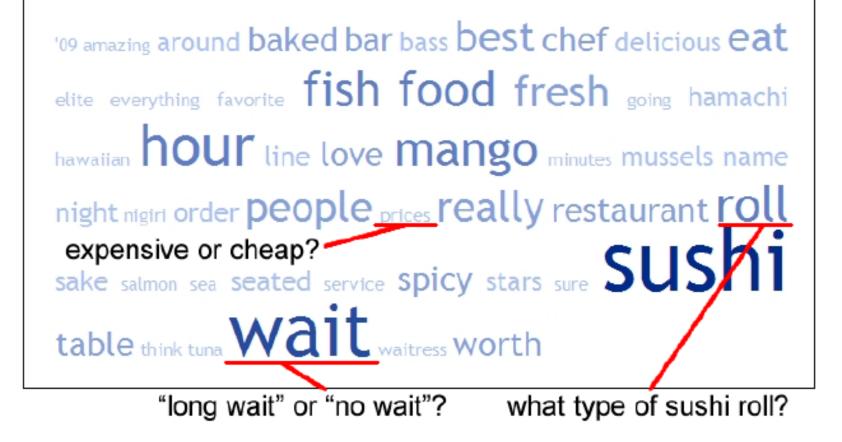
Label E/	E/P	Effect of	Primary bias factor	Effect of bias factor agreement	Additional factor	Accuracy at min $\Delta$ font size			Notes
		$\Delta$ font size				agree	neutral	disagree	
len1	Р	$\checkmark$	word length $^{\dagger}$	V	-	0.860	0.879	0.753	Word length biases percep- tion of font size
len2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

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



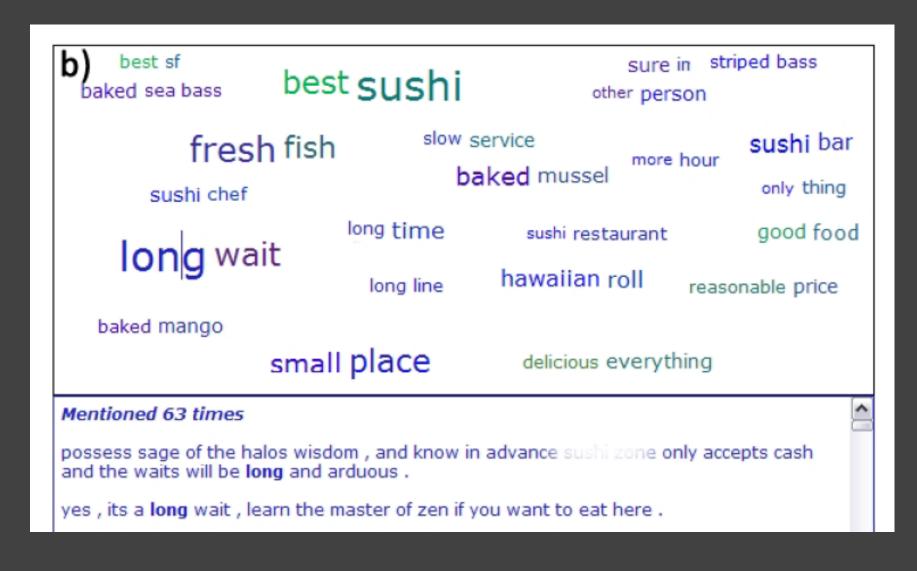
## Yelp Review Spotlight

#### [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]

adverted	adjourned	allocatur	adequate	bankruptcy	bargaining	about	abuse	abuse	appeal	ballot	accused	agency
anent	alia	analysis	affirmed		benefit	asked	affirmed	aliens	argument		agency	agency's
	allocution	antitrust	aid	barge capital	case	believe	-	appropriate	attached	black	antidumping	agency o
appellant	arbitration	app	ante	cargo	cocaine	called	appellee	asylum	binding	candidate	application	authority
appellant's	asbestos	asbestos	appeal	charter	court	cocaine	argued	have been been been been been been been be	brief	case		bargaining
appellee	closure	assets	argument	coverage	Dariam.	conspiracy	Deleve	circuit	cited	certified	board	brief
asseveration	commenced complaint	bankruptcy	because	damages	defendant	could deal	crack	cited	collateral	class	claim	broadcast
below	conveniens	believe	before	death	defendant's	defendant	denied	contended	сору	commerce	composition	capricious
boat	copyright	benefit	coal	debtor	detwared	enough	disability	courts	defendant.	county	compounds comprising	carrier
brief	date	bottlers	cocaine	drilling	disability	0101	distribution	dba declared	determine	court	construction	competition
ca(acking chinarts	defendant	class	contention	execution feet	district	fire gang	district	denial	disfavor	death	contract	COStS
commonwealth	disenfranchised	Class	CONTENTION	gas	dia data data data data data data data d	get	drug	deportation	doc	desegregation	costs data	disposition
defendant	foreign	context	court's	habeas	employees	gun	evidence	discretion	doctrine	disenfranchised	decision	emissions
del	fraud	debtor	crack	homestead	filed	had	farm	disposition	estoppel	district	description	employees exemption
ensued	ground	dist.	decisional	indemnity	firearm	harassing	grams	district	examination	dozer election	device	explanatory
event	heroin	exercise	denied	instant	follows grievance	have	had	errs	forthwith	electors	embodied	facilities
factfinding	injunction	fiduciary	disclosed	insurance	guilty.	her	her	except	furnished	immunities	equivalent	hazard
guidelines	inter	have	dispensed	interest	hereby		his	fear	further	injustion staty	inequitable	interpretation
here	internal	here	distribution	jurists	his	him	impair	fish	greeing her	insurance		intervenor
incarcerative	keeplock	inasmuch	district	law	job	his	inmates	habitat hardship	judgment	jail	infringement	labol
inference	marks	insurance	drug	liability	judgment	job	jury	his	judicata	law	invalid	license market
jury	144	interest	fact	loan	magistrate	judge	level	immigration	material	migrant	invention	memoranda
limned	millions	jurisdiction	from	marihuana	magistrate's	just	ethamphetan	jurisdiction	nevertheless	mitigation	inventor	petitioner
lst	narcotics	legislation	interlocutory	maritime	medical	lawyer	months	may	opinion	nonstatutory	layer	pipelines
might	pagmant	liability	joined	mitigation	motion	might	office	methamphetamine	oral	ordinance	limited Recal	
mortgage	plaintiff	majority	legal	negligence	office	more	opinion	novo	order	payday	means	promulgated
plausible	plaintiff's	market	magistrate	nre	panel	one	pain	Cont Cont	order	phase	merchandise method	proposed
point	principal	notes	material	offshore	paupers	ostrich	postconviction	panel	persuasive	preceding	mode	rate
prius rescript	quotation	our	merits	parish	plaintiff	out para	pounds press/particle guantity	persecution	plaintiff's	qualified race	noninfringement	
said	racketeering	parents	miner's	platform	plaintiff's	police	reversed	petition	precedential	racial	patent	regulations
say	reinsurance	plaintiff	mining	policy	pneumoconiosis	prisoner	search	political	record	section	patentee	rehearing
see	respect security		opinion	recovery	police	say	sentence	prisoner	remained	sentence		reprinted
SOME suggested		plan	oral	ref'd	pulmonary	she	sexual	public	submitted	sheriff	product	rulemaking
supra	shareholders	plenary	order	removed	pursuant	suit	she	pursuant	suspended	4	reissue	section
think	shares	policy	pneumoconiosis		recommendation	supra	subd	section	tab therefore	students	retirement	see
tit	sterile	product	present	seaman	search	tentative	sappression	specie	tit	turtle	said	service
token	stock	recognized	pro	servitude	sentence	than thought	testified	suitable	unanimous	tusks	signal	shipper
town trialworthy	summation	section	process	stated	sitting	told	testimony	tribal	unfavorable	vote	skill specific	tariff
vessel	trade		pulmonary	suit	unanimous	too	trial	tribe	unpublished		structure	transmission
vis	vacated	settlement	relative relative	usury	union	want	tribal	unanimous	value	voters waterbodies	surface	-
viz =	view	under	would	vessel	upon	what who	verdict	water	vehicle	white	vaccination	union
whom	waybill where	would	wrote	writ	warrant	would	work	without	vol	zone	veterans	waste
First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth	Ninth	Tenth	Eleventh	Federal	DC

# **Context and Structure**

## Concordance

者 Concordance - Larkin.Concordance 📃 🔲 🔀										
<u>F</u> ile <u>T</u> ext <u>S</u> earch <u>E</u> dit <u>H</u> eadwords Conte <u>x</u> ts <u>V</u> iew T <u>o</u> ols Hel <u>p</u>										
Headword	No. 🔺	Context	Word	Context	Reference					
HEAR	15	That my own	heart	drifts and cries, having no	Deep Analysis	Ce				
HEARD	9	By the shout of the	heart	continually at work	And the wave	Centred				
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave	8				
HEARS	3	The tread, the beat of it, it is my own	heart		Träumerei					
HEARSE	1	Because I follow it to my own	heart		Many famous					
HEART	25	My	heart	is ticking like the sun:	lam washed u	6				
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	di Da: Yo Se				
HEARTH	1	Having no	heart	to put aside the theft	Home is so Sa	led				
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Beau					
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the	r F				
HEAT	6	These I would choose my	heart	to lead	After-Dinner F					
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spac	Index				
HEATH	1	This petrified	heart	has taken,	A Stone Churc	×				
HEATS	1	How should they sweep the girl clean	heart	1	l see a girl dra					
HEAVE	1 -	Hands that the	heart	can dovern	Heaviest of flo					
<	>	<	Ш							
Words Tokens At word Deleted lines Word sort Context sort										
7318 37070 2990 1 [24] Asc alpha (string) Asc occurrence order										

## Context & Structure

if love be

[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 .

blind

rough with you , be rough with love .

love cannot hit the mark .

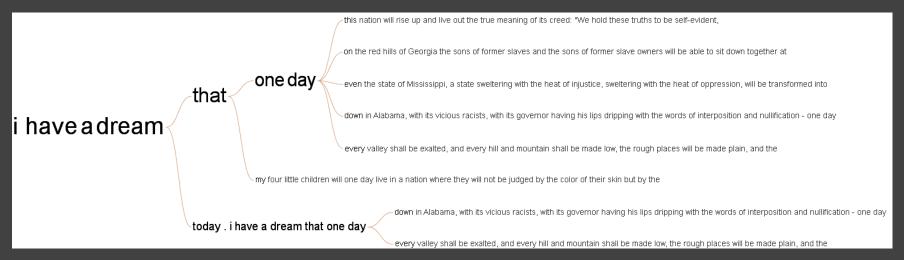
it best agrees with night .

## 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

currently showing

Tags:

many eyes

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

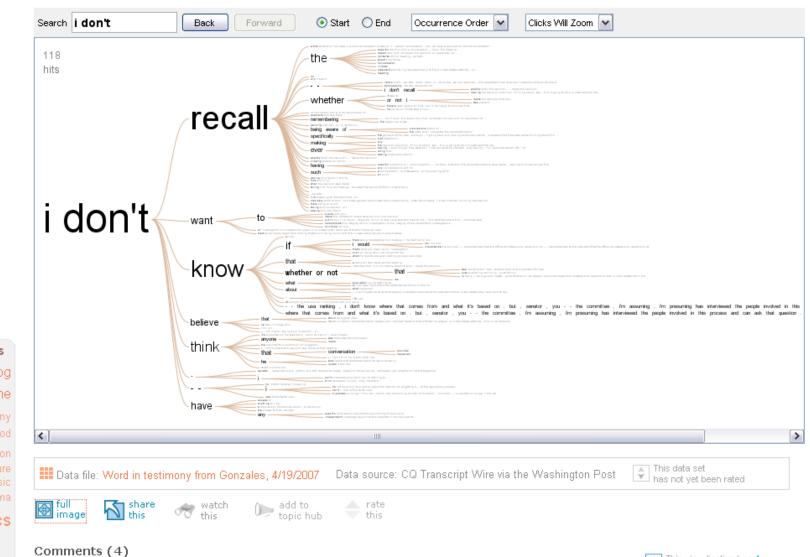
contact Us contact report a bug

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



#### This visualization has 4 positive and 0 negative

## Glimpses of Structure...

Concordances show local, repeated structure But what about other types of patterns?

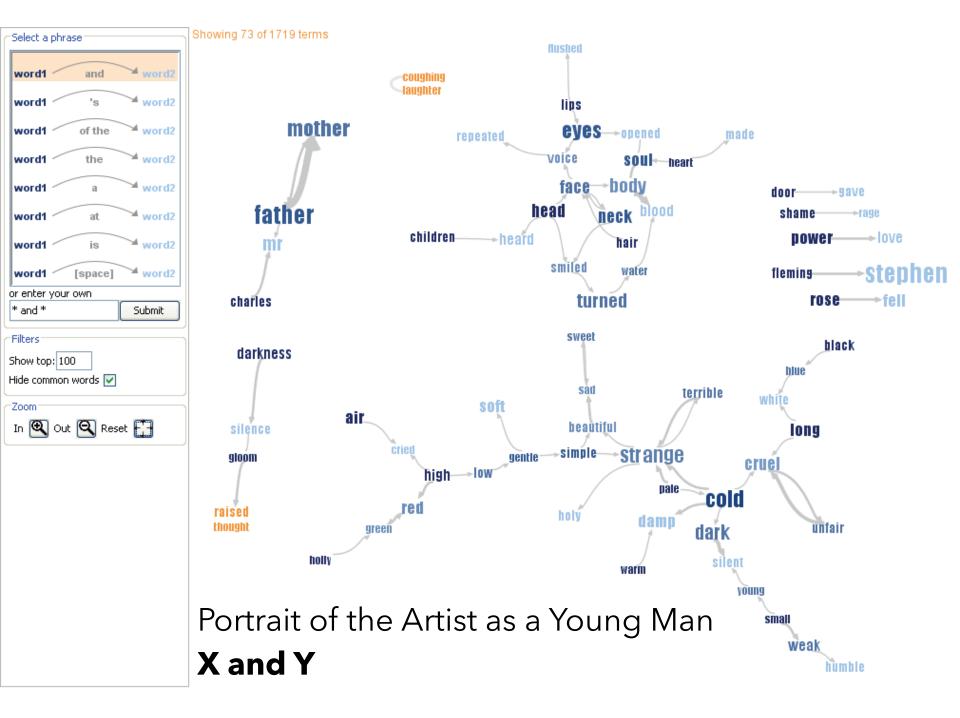
Lexical:<A> at <B>Syntactic:<Noun> <Verb> <Object>

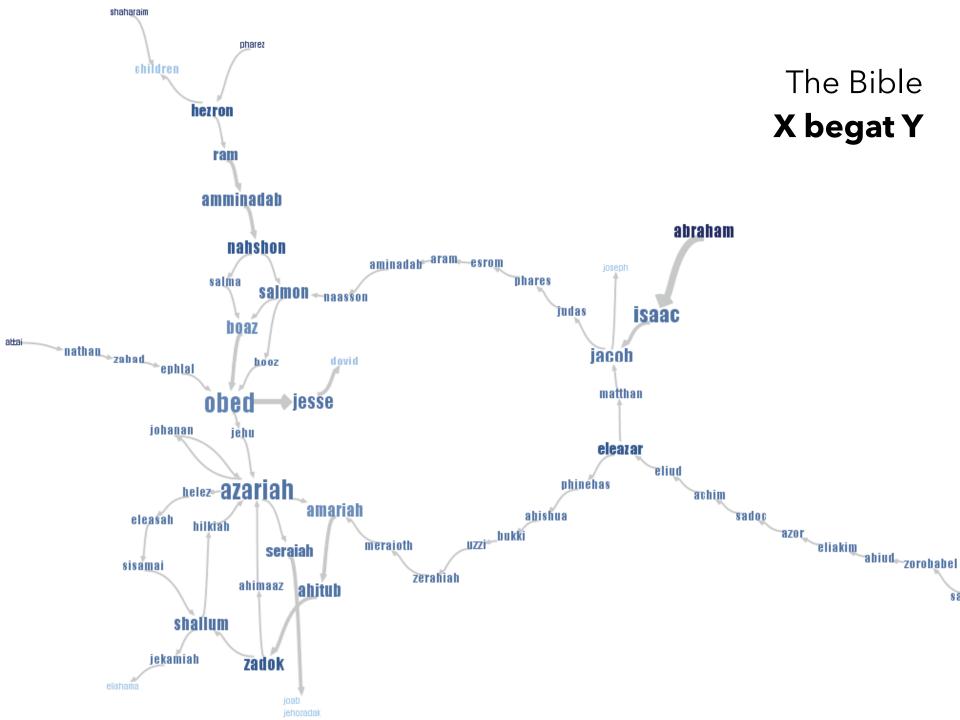
## **Phrase Nets**

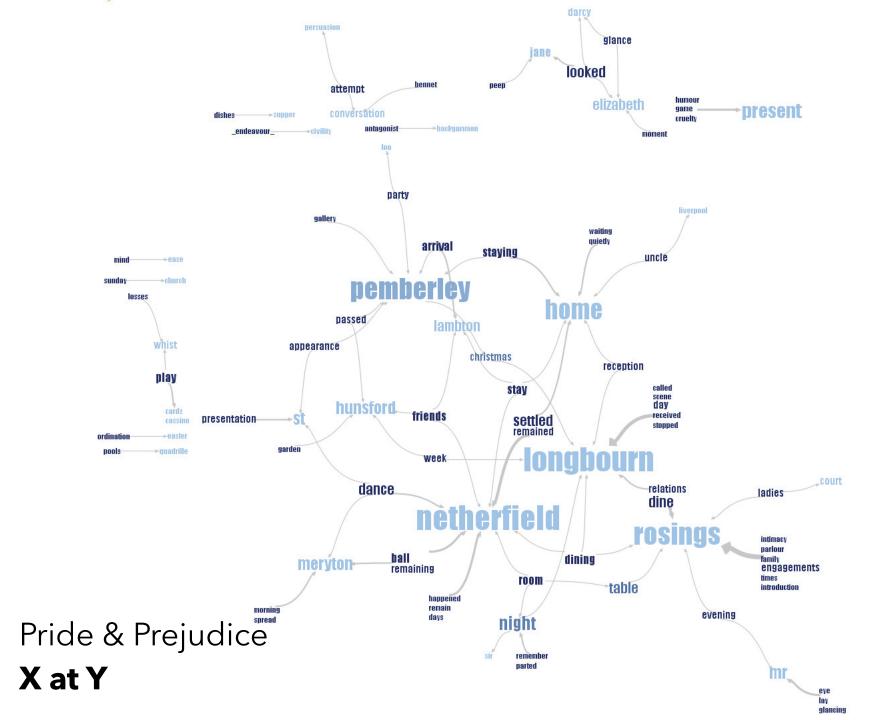
[van Ham et al. '09]

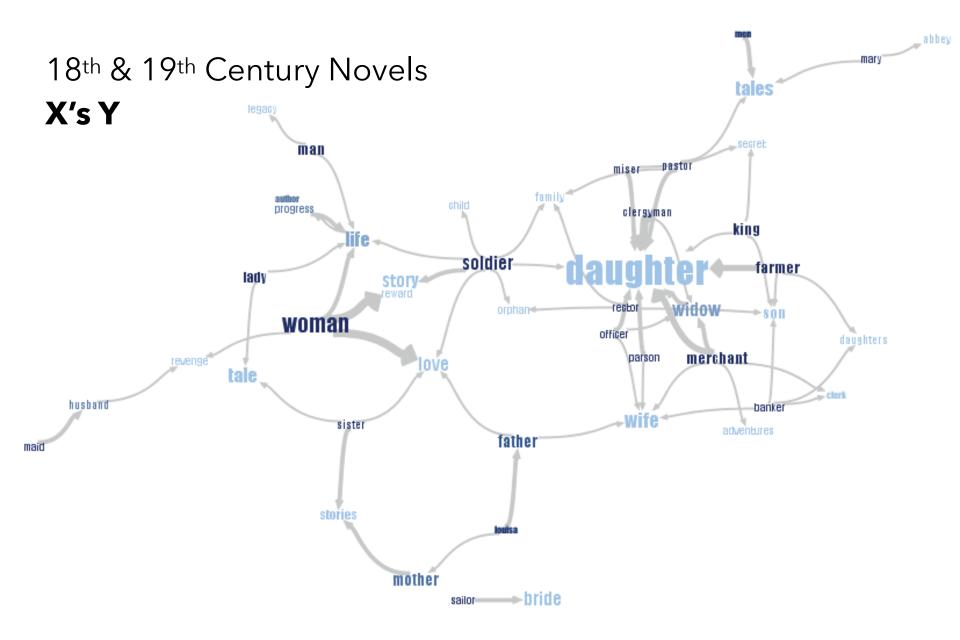
Look for specific **linking patterns** in the text: "A and B", "A at B", "A of B", etc. Could be output of regexp or parser.

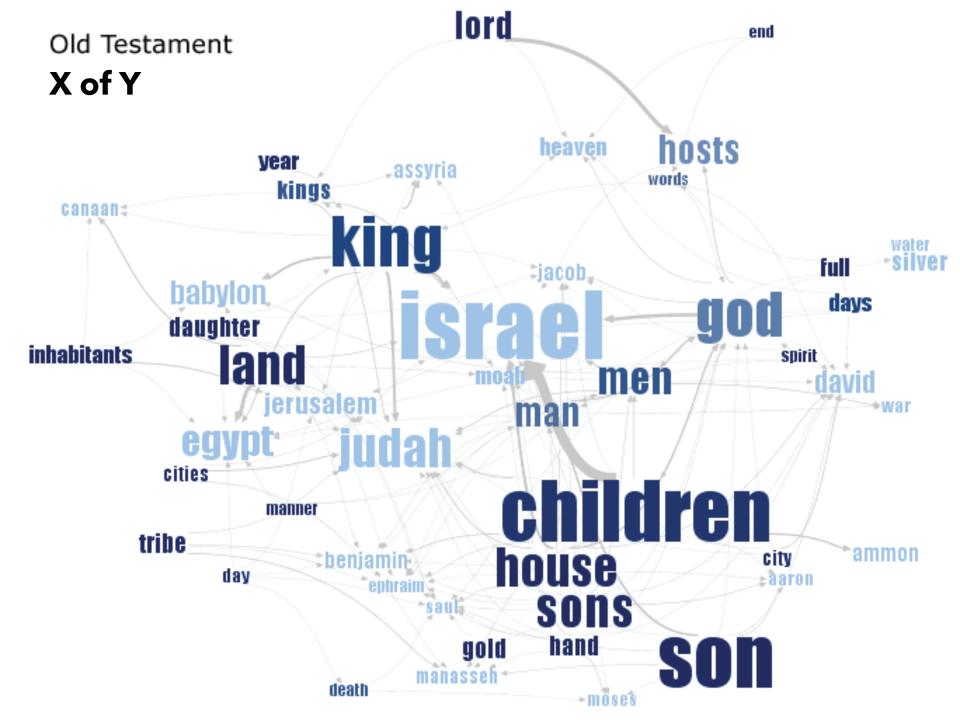
Visualize patterns in a node-link view: Occurrences → Node size Pattern position → Edge direction

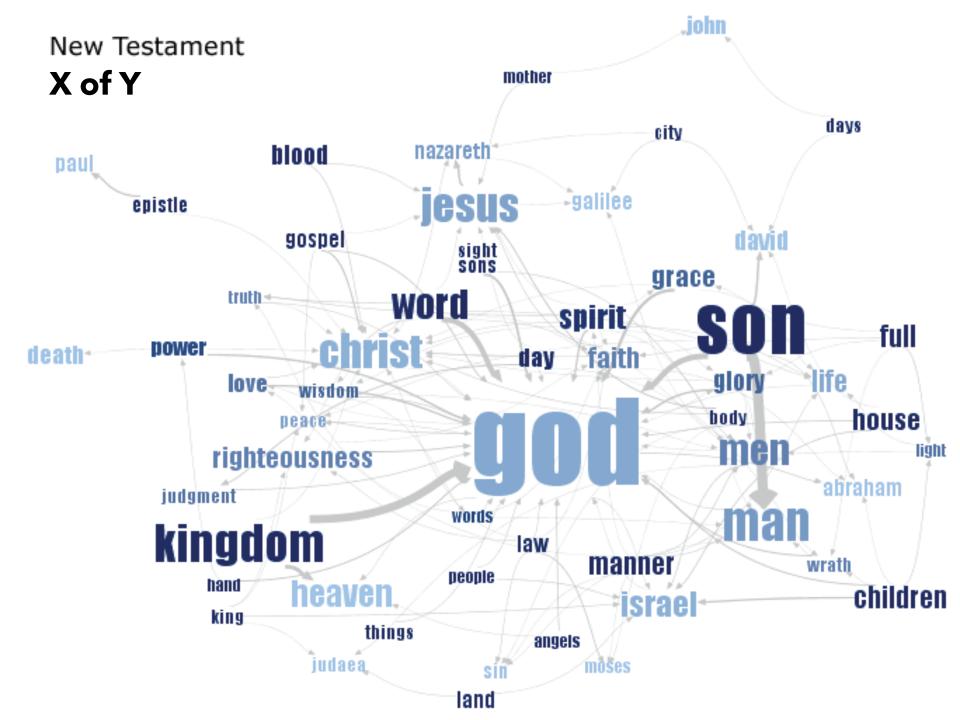












## Document Content

#### **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?

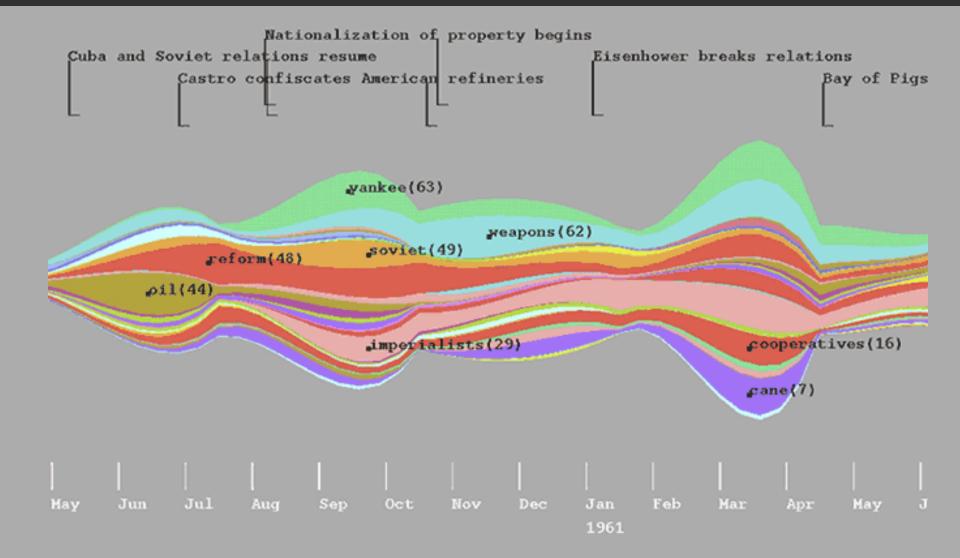
## **Entity Relationships**

#### [Görg et al. '07]

🕸 List View	sough from which there are							
Edit View Bookmarks Lists Options								
person 👻 Add all Clear	Show all connections	place   Add all Clear						
🔤 🗄 📄   📥   🗙   🎫 🚍		ABC 🔚 📃   📤   🗙   🔳 🚍 🗏						
Bugarov 4		USA						
Carlos		Cuba						
I Carlos Araneda		Pakistan 🗉						
Carlos Morales		Canada						
Castro		Columbia –						
Cesar Arze		Jamaica						
Charles Wilson		Afghanistan						
Dan West		Havana						
Daniel Harris		Detroit						
David Loiseau		Mexico						
Dean Simpson		Michigan						
Dr. Baker		Montego Bay						
Dustin Marshall		Texas						
Edgar Spencer		Chitral						
Edward Thompson	T THE	Morocco						
Escalante Escalante	I. //// <b>I</b> T	Peshawar						
F. Baker	////#	Russia						
I Felix Baker		Casablanca						
I Ford	////////////////////////////////////	Chicago						
Forrest Wells	////#	I Illinois						
Fr. Augustin Dominique	/ / / / <b>//</b>	New Jersey						
Fred Fisher		UK						
George Garcia	////////////////////////////////////	Dominican Republic						
Grigory Sizov	/////////////////////////////////////	l Florida						
I Hamid Qatada	////////////////////////////////////	France						
Hector Lopez	X / / / H	London						

## **Theme River**

#### [Havre et al. '00]

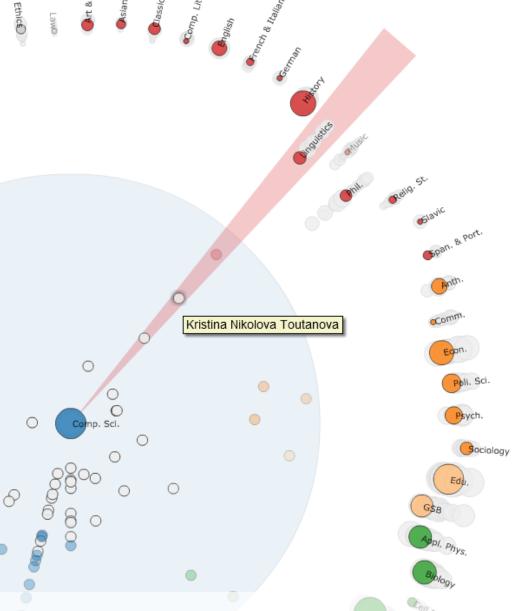


## Similarity & Clustering

**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)

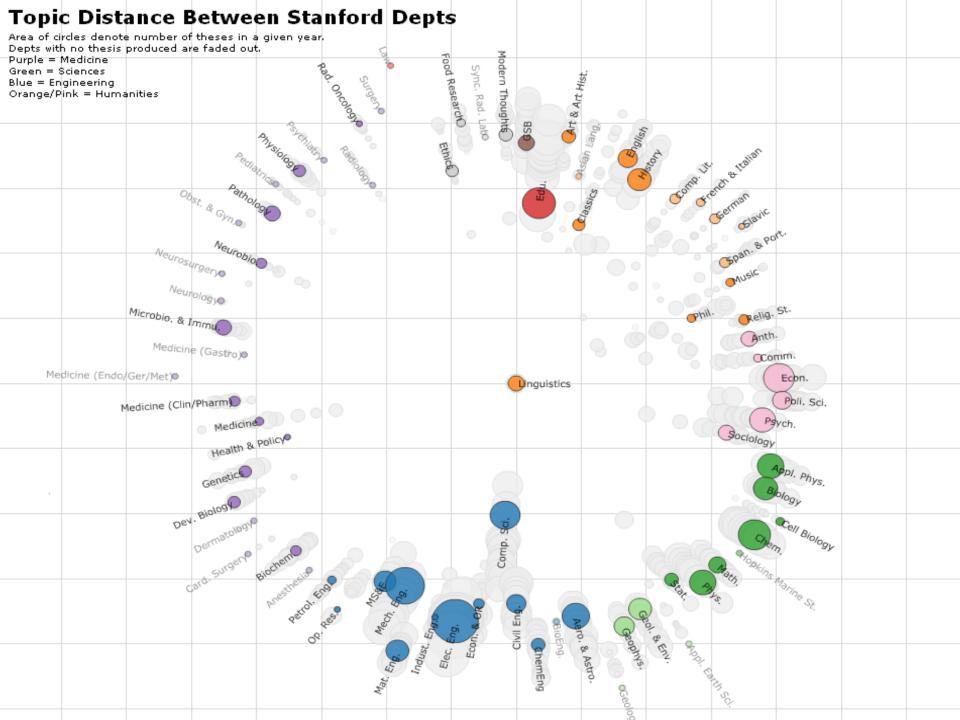
Keywords: Syntactic, Semantic, Tree kernels, Parsing

Abstract:

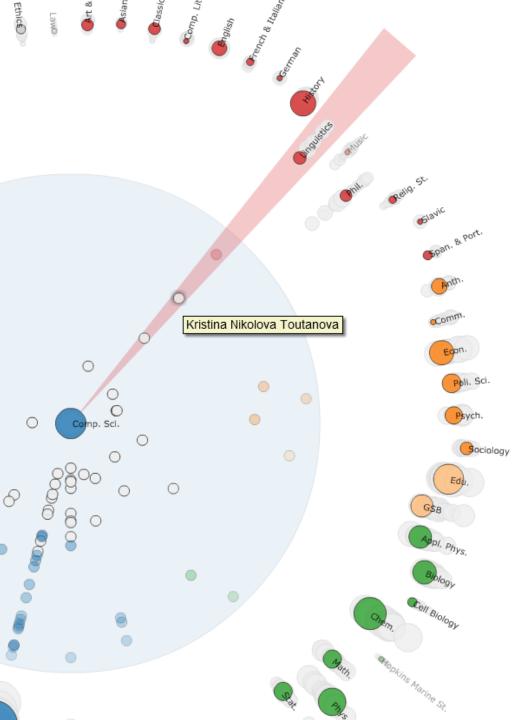
This thesis focuses on building effective statistical models for disambiguation of sophisticated syntactic and semantic natural language (NL) structures. We advance the state of the art in several domains by (i) choosing representations that encode domain knowledge more effectively and (ii) developing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







#### Effective statistical models for syntactic and semantic disambiguation

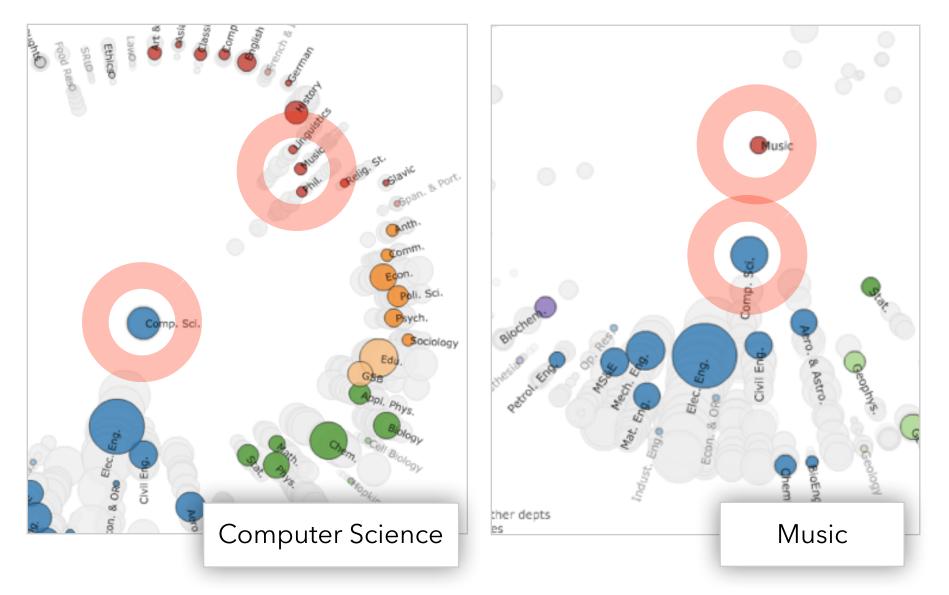
Student: Kristina Nikolova Toutanova Advisor: Christopher D. Manning

Computer Science (2005)

Keywords: Syntactic, Semantic, Tree kernels, Parsing

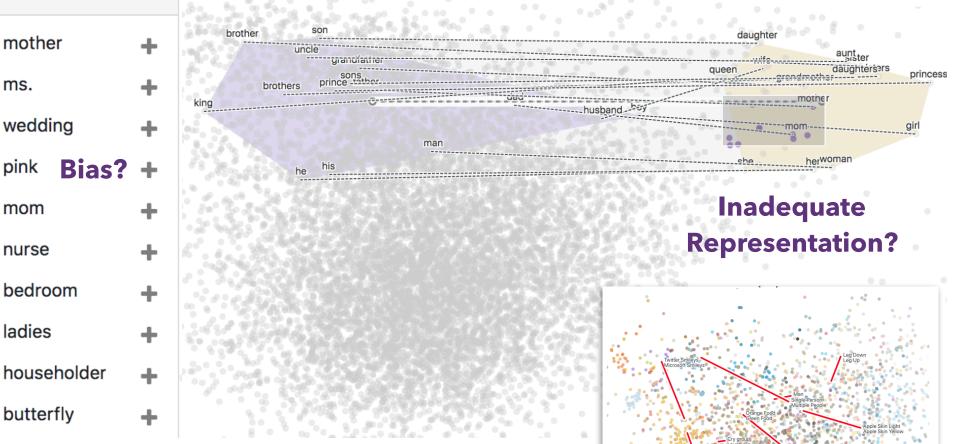
#### Abstract:

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