

# **Text Visualization**

*Maneesh Agrawala*

**CS 448B: Visualization  
Winter 2020**

1

# **Announcements**

2

# Final project

---

## New visualization research or data analysis project

- **Research:** Pose problem, Implement creative solution
- **Data analysis:** Analyze dataset in depth & make a visual explainer

## Deliverables

- **Research:** Implementation of solution
- **Data analysis/explainer:** Article with multiple interactive visualizations
- 6-8 page paper

## Schedule

- Project proposal: **Wed 2/19**
- Design review and feedback: **3/9 and 3/11**
- Final presentation: **3/16 (7-9pm) Location: TBD**
- Final code and writeup: **3/18 11:59pm**

## Grading

- Groups of **up to 3 people**, graded individually
- Clearly report responsibilities of each member

3

# Design Feedback (Week 10)

---

## Signup for a 10 min slot

<https://docs.google.com/spreadsheets/d/1BiXmbQHrC3-chPT6kKS51Q-2p9XhbiM3Qct0N847yPM/edit?usp=sharing>

- M 3/9 4-6pm
- T 3/10 7-8pm (SCPD only)
- W 3/11 4-6pm

**Plan to give a 5 min presentation (mostly demo) of work so far. We will give oral feedback.**

4

# Final Presentation

---

**M Mar 16 7-10pm, Location TBD**

- Short presentation (5 min, mostly demo)
- Make sure there is time for questions

5

# Text Visualization

6

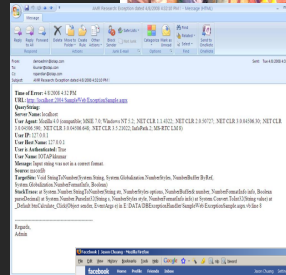
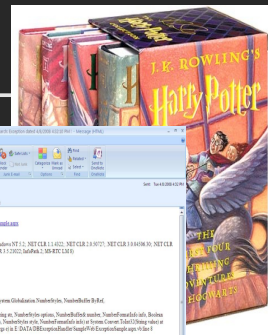
# Text as data

## 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)



7

# Why visualize text?

8

## Why Visualize Text?

---

**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

9

## Example: Health Care Reform

---

### Background

Initiatives by President Clinton

Overhaul by President Obama

### Text data

News articles

Speech transcriptions

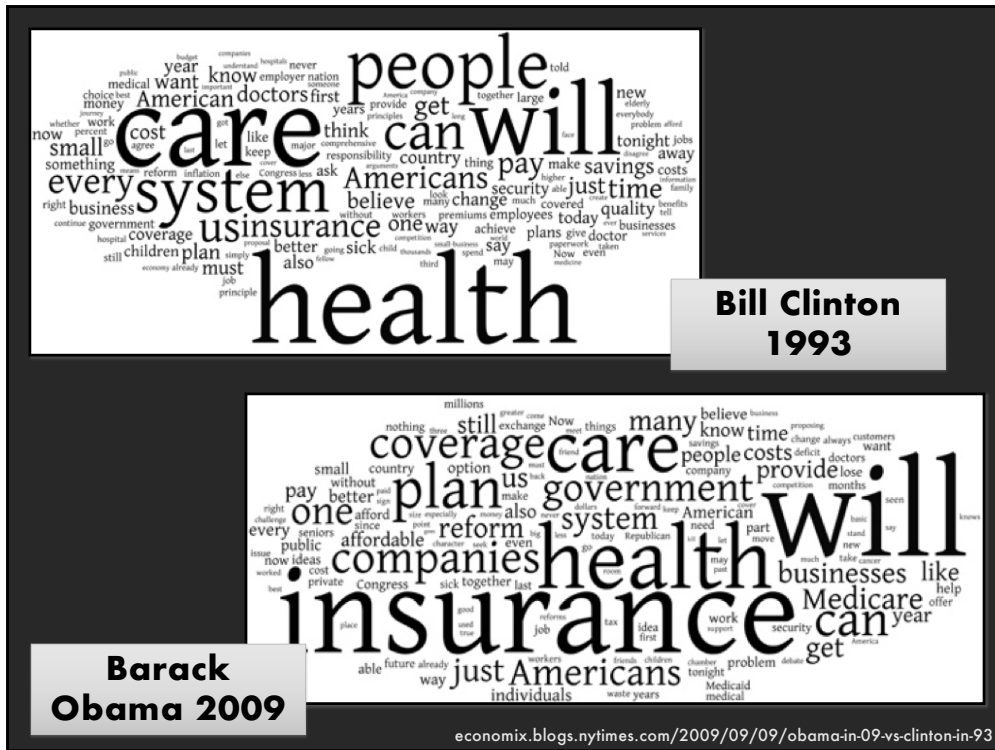
Legal documents

**What questions might you want to answer?**

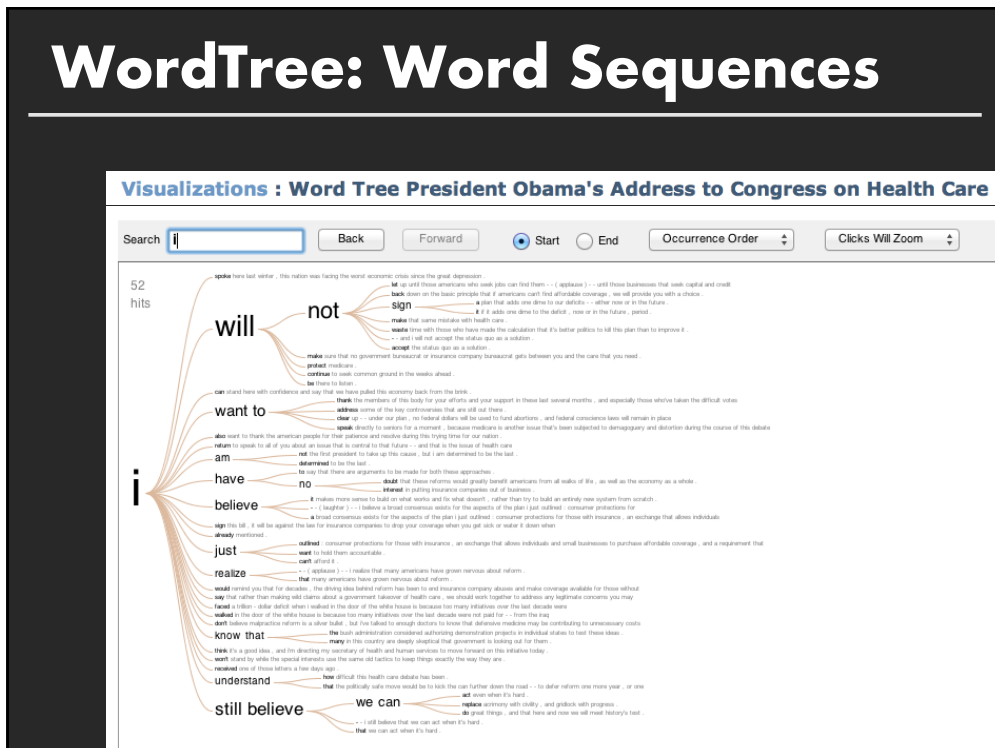
**What visualizations might help?**

10

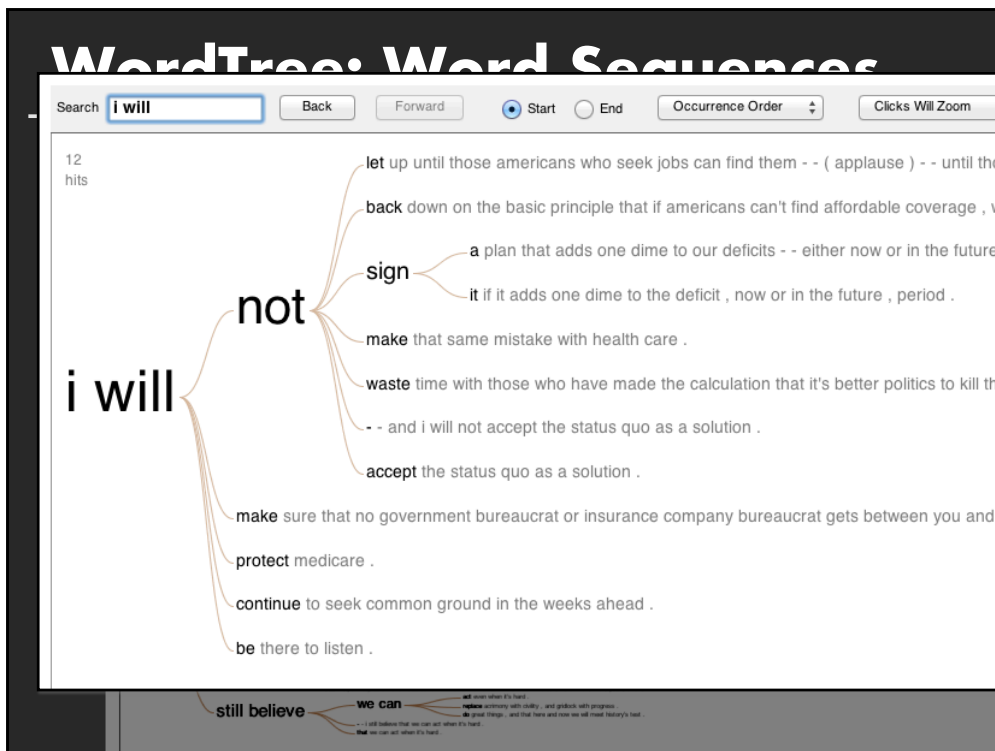




13



14



15

## Gulf of Evaluation

Many (most?) text visualizations do not represent 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?

16



# 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

17

# Topics

---

**Text as Data**

**Visualizing Document Content**

**Visualizing Conversation**

**Document Collections**

19

# Text as Data

20

## Words as nominal data?

---

High dimensional (10,000+)

### More than equality tests

- Correlations: *Hong Kong, San Francisco, Bay Area*
- Order: *April, February, January, June, March, May*
- Membership: *Tennis, Running, Swimming, Hiking, Piano*
- Hierarchy, antonyms & synonyms, entities, ...

**Words have meanings and relations**

21

# Text Processing Pipeline

---

## Tokenization

Segment text into terms.

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

Numbers and symbols? *#cardinal, @Staanford, OMG!!!!!!!*

Entities? *Palo Alto, O'Connor, U.S.A.*

22

# Text Processing Pipeline

---

## Tokenization

Segment text into terms.

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

Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!*

Entities? *Palo Alto, O'Connor, U.S.A.*

## Stemming

Group together different forms of a word.

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

Lemmatization? *goes, went, gone -> go*

23

# Text Processing Pipeline

---

## Tokenization

Segment text into terms.

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

Numbers and symbols? *#cardinal, @Stanford, OMG!!!!!!!!*

Entities? *Palo Alto, O'Connor, 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

24

# The Bag of Words Model

---

**Ignore ordering relationships within the text**

**A document  $\approx$  vector of term weights**

Each term corresponds to a dimension (10,000+)

Each value represents the relevance

- For example, simple term counts

**Aggregate into a document  $\times$  term matrix**

Document vector space model

25

# Document x 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

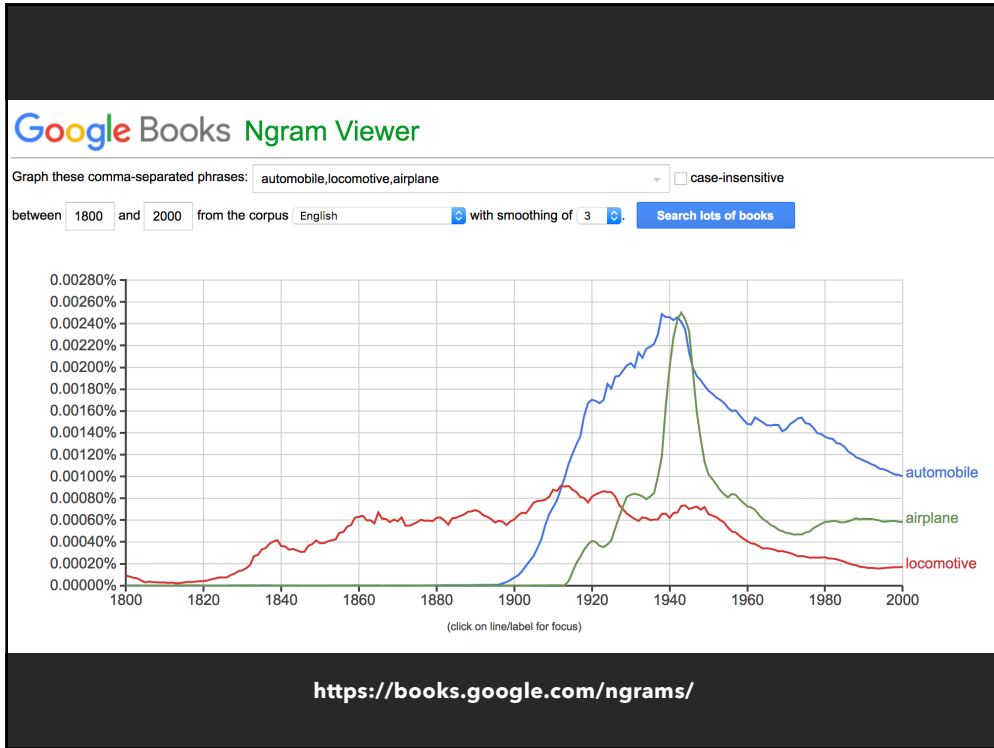
26

# WordCount (Harris 2004)

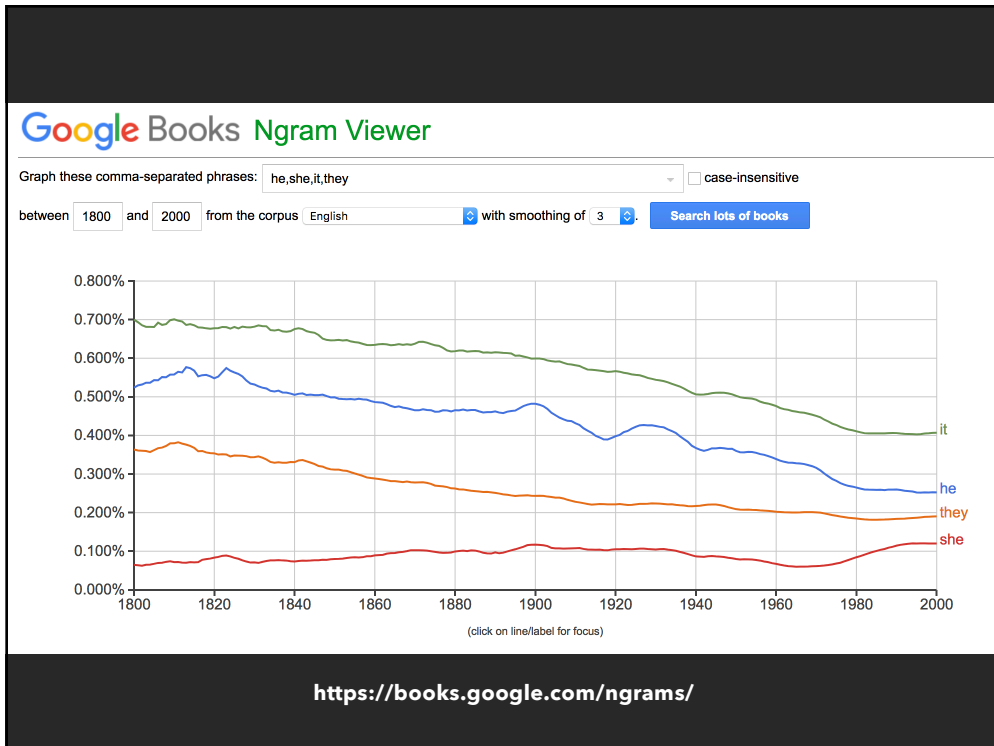
The screenshot shows the WordCount website interface. At the top right, there is a button labeled "WORDCOUNT". Below it, there are navigation links for "PREVIOUS WORD" and "NEXT WORD". The main display area shows a horizontal bar chart representing the frequency distribution of words. The word "the" is highlighted in large black text, with a red "1" below it. Other words are shown in smaller, lighter text: "of" (rank 2), "and" (rank 3), "to" (rank 4), "ain" (rank 5), "that" (rank 6), "is" (rank 7), "was" (rank 8), "for" (rank 9), "on" (rank 10), "you" (rank 11), "help" (rank 12), "with" (rank 13), "by" (rank 14), "the" (rank 15), "at" (rank 16), "the" (rank 17), "on" (rank 18), "the" (rank 19), "the" (rank 20), "the" (rank 21), "the" (rank 22), "the" (rank 23), "the" (rank 24), "the" (rank 25). Below the chart, there is a "CURRENT WORD" label. At the bottom, there are input fields for "FIND WORD:", "BY RANK:", and "REQUESTED WORD: THE". The "REQUESTED WORD" field is filled with "THE" and "RANK: 1" is displayed below it. On the right side, there is text that says "86800 WORDS IN ARCHIVE" and "ABOUT WORDCOUNT".

<http://wordcount.org>

27



28



29



**Given a text, what are the best descriptive words?**

32

## **Keyword Weighting**

---

### **Term Frequency**

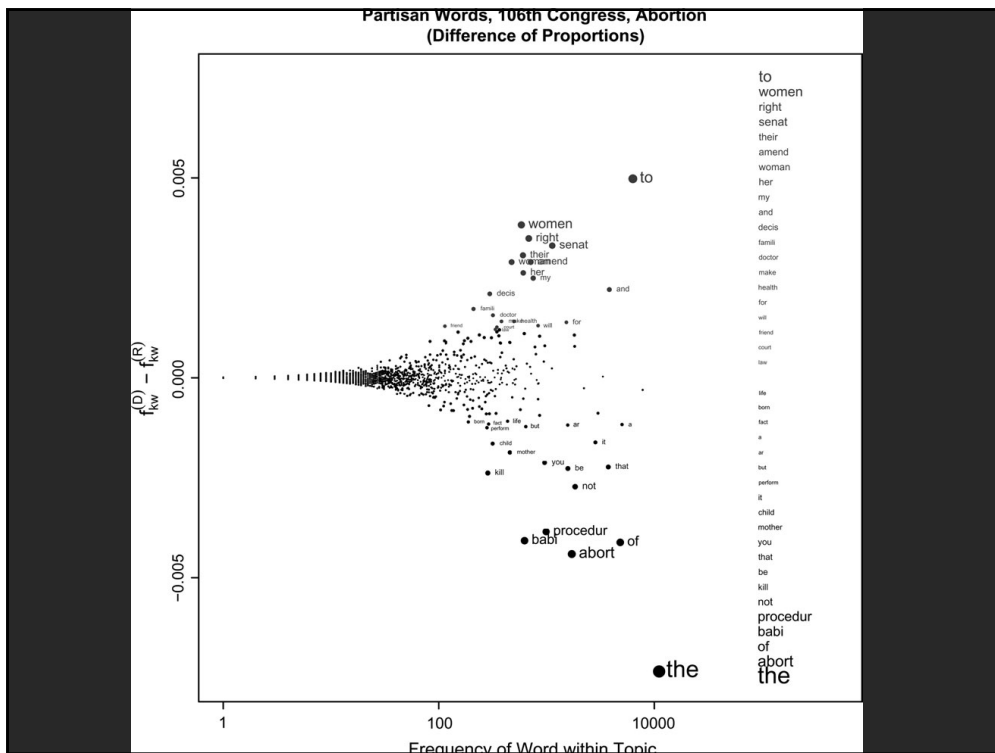
$tf_{td} = \text{count}(t) \text{ in } d$

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

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

33





34

## Keyword Weighting

### Term Frequency

$$tf_{td} = \text{count}(t) \text{ in } d$$

### TF.IDF: Term Freq by Inverse Document Freq

$$tf.idf_{td} = \log(1 + tf_{td}) \times \log(N/df_i)$$

$$df_i = \# \text{ docs containing } t; \quad N = \# \text{ of docs}$$

35

## Limitations of Frequency Statistics

---

Typically focus on unigrams (single terms)

Often favors frequent (TF) or rare (IDF) terms

Not clear that these provide best description

**“Bag of words” ignores additional info.**

Grammar / part-of-speech

Position within document

Recognizable entities

41

## How do people describe text?

---

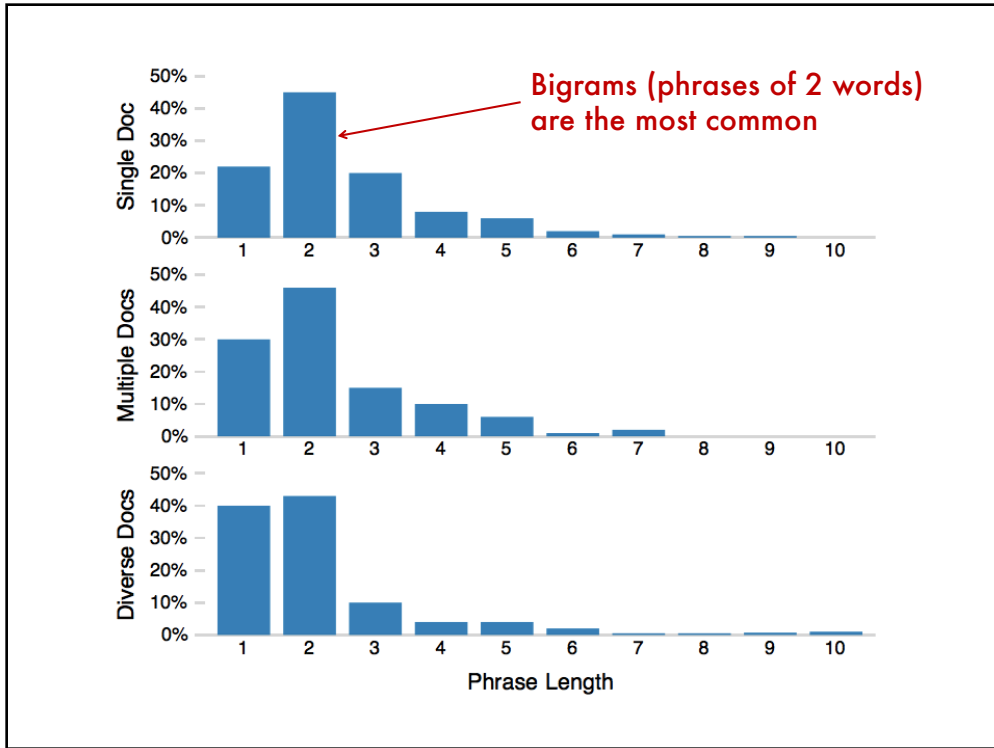
Asked 69 graduate students to read and describe dissertation abstracts

Each given 3 documents in sequence; summarized each using keyphrases, then summarized the 3 together as a whole using keyphrases

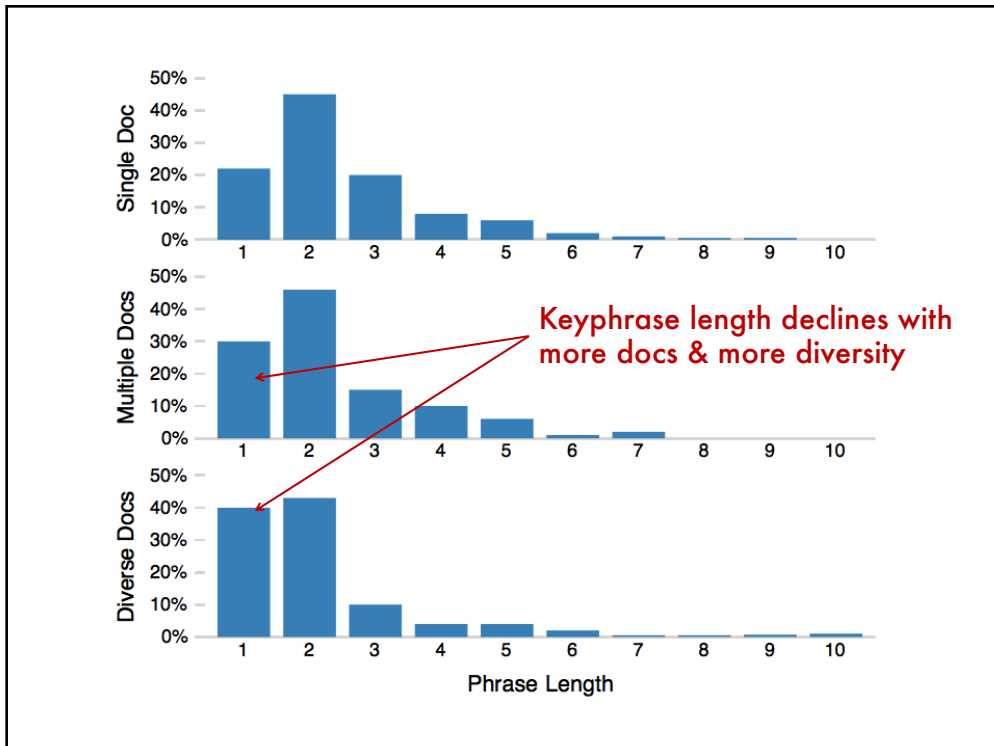
Were matched to both *familiar* and *unfamiliar* topics; *topical diversity* within a collection was varied systematically

[Chuang 2012]

42



43



44

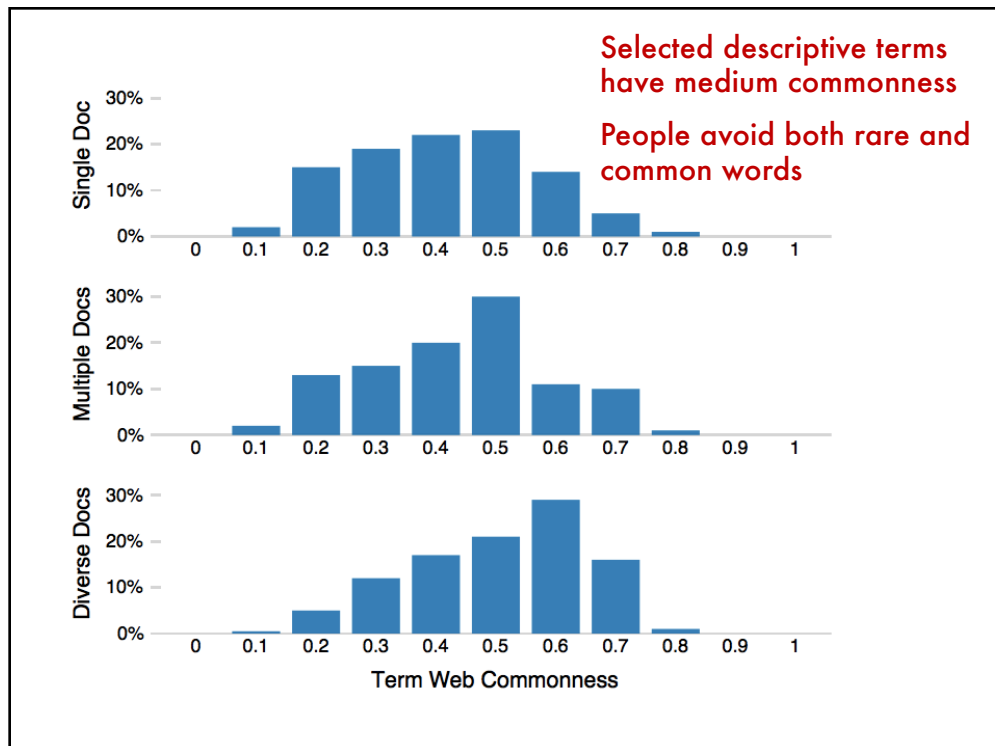
## Term Commonness

$$\log(\text{tf}_w) / \log(\text{tf}_{\text{the}})$$

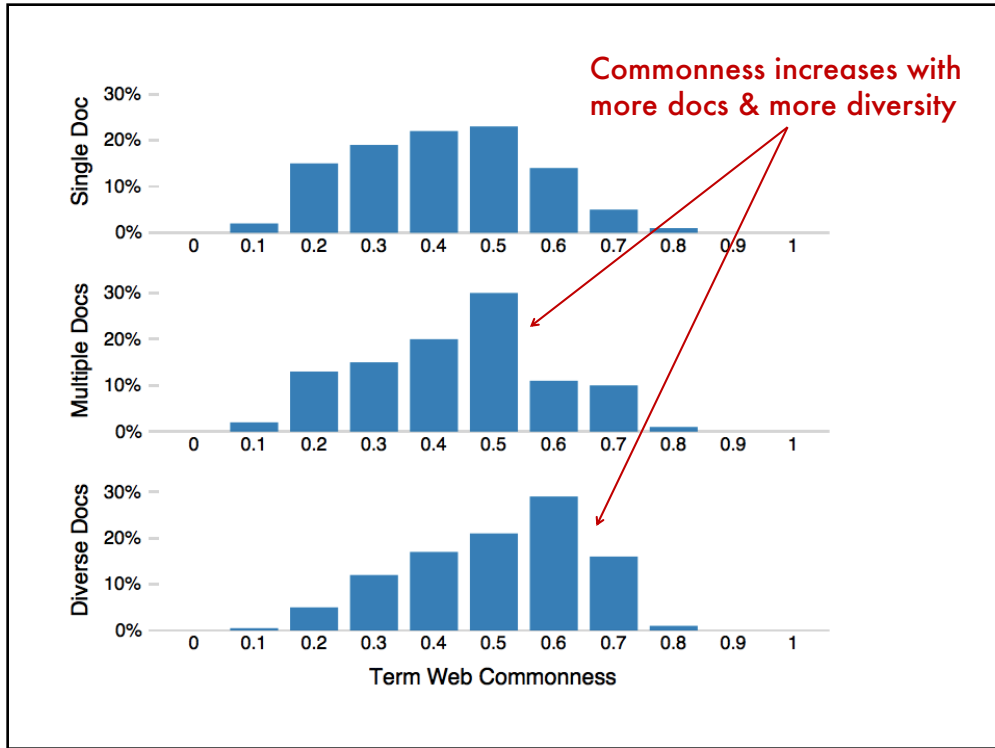
The normalized term frequency relative to the most frequent n-gram, e.g., the word "the".

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

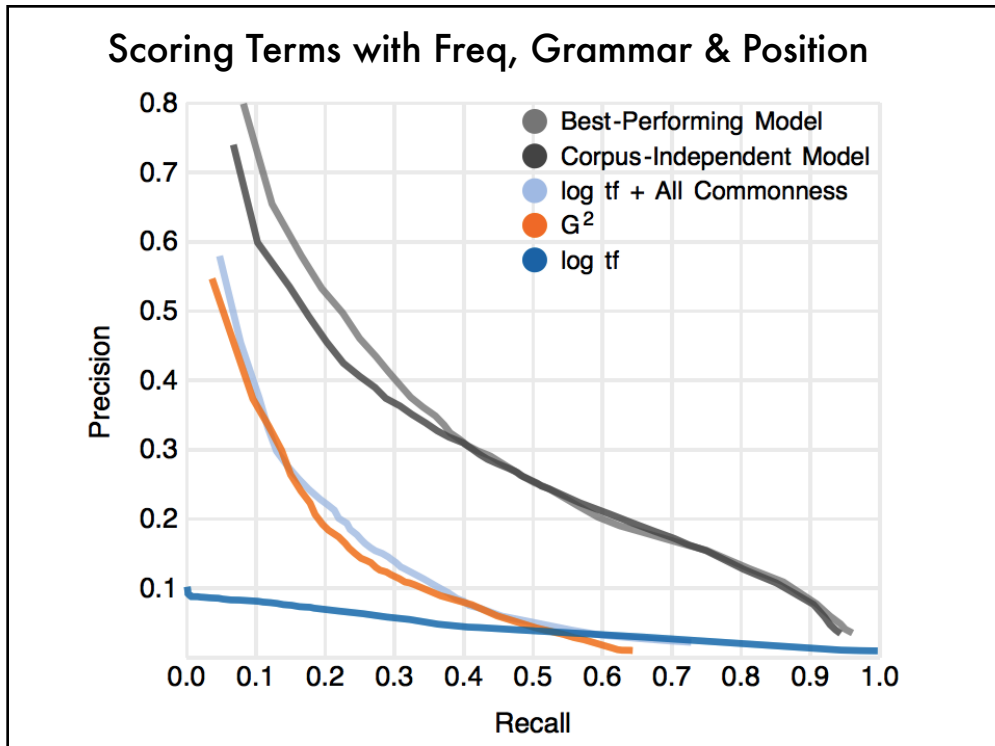
45



46



47



48

A fighter jet rain check

http://www.boeing.com/Features/2010/09/bds\_feat\_morewater

### A fighter jet rain check

Story and video by [Chamila Jayaweera](#)


Have you ever thought about what it takes to make sure that sea-based fighter jets stay dry?

When it comes to the F/A-18 Super Hornet, Boeing engineers in St. Louis use a special process called the Water Check Test to rule out areas where moisture could seep into the aircraft and its electronics suite.

Program experts douse the jet with simulated rain at a 15-inch-per-hour rate for about 20 minutes inside an enormous hangar in St. Louis.


"Our ultimate customers are U.S. Navy fighter pilots, and we want to ensure their safety in flight and on the ground, and water-tight integrity of the aircraft also helps increase their effectiveness," said Boeing's Rich Baxter, F/A-18 Super Hornet final assembly manager.

To find out more about how the process works and watch the action unfold, click above to see the video story.



CHAMILA JAYAWEERA/BOEING

The Water Check team rolls in a large metal frame, which they affectionately call their "spray tree," over a Super Hornet inside a St. Louis hangar.



49

$G^2$	Regression Model
fighter	Super Hornet
F/A	F/A -18
Hornet	fighter jet
Super	Boeing engineers
Boeing	special process
-18	rain check
rain	electronics suite
St.	Program experts
jet	simulated rain
Louis	ultimate customers
15-inch-per-hour	enormous hangar
douse	water-tight integrity
hangar	Rich Baxter
water-tight	15-inch-per-hour rate
Check	video story
Baxter	aircraft
sea-based	U.S. Navy fighter pilots
aircraft	Super Hornet final assembly manager
Rich	
seep	
click	
Navy	
sure	
Water	
moisture	
watch	
enormous	
stay	
water	

50



## Tips: Descriptive Keyphrases

---

### Understand the limitations of your language model

#### Bag of words:

- Easy to compute
- Single words
- Loss of word ordering

### Select appropriate model and visualization

- Generate longer, more meaningful phrases
- Adjective-noun word pairs for reviews
- Show keyphrases within source text

53

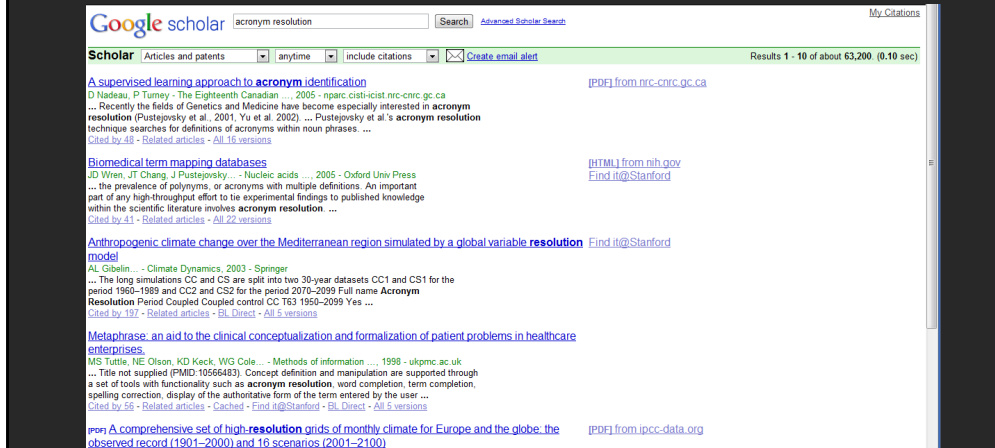
## Visualizing Document Content

54

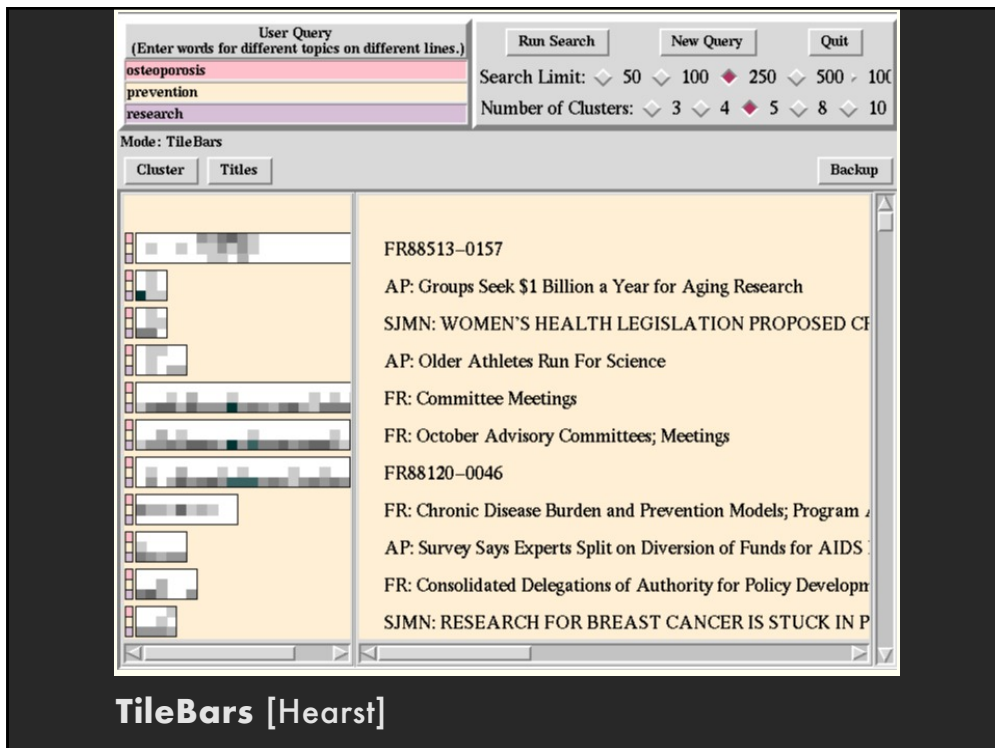


# Information Retrieval

Search for documents  
Match query string with documents  
Visualization to **contextualize results**

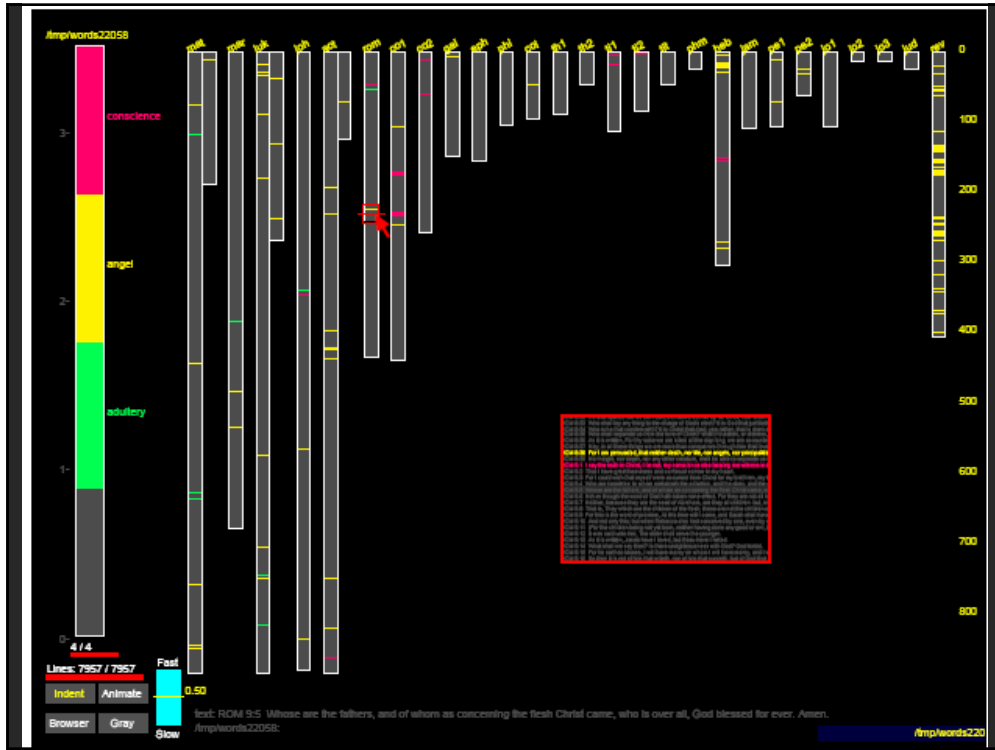


55

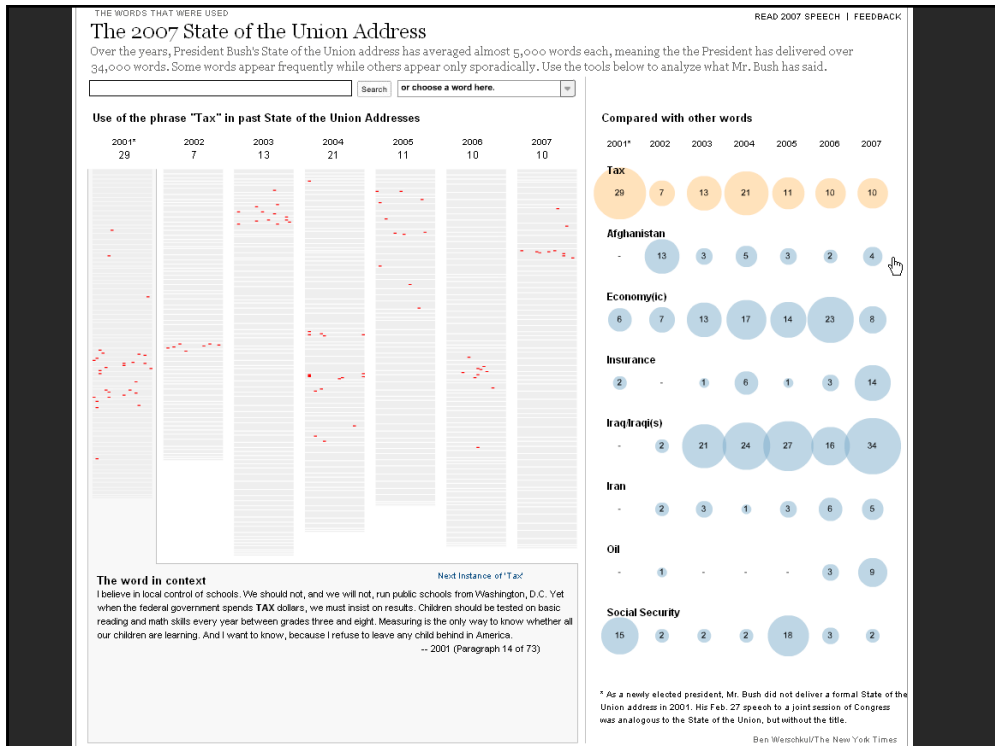


TileBars [Hearst]

56



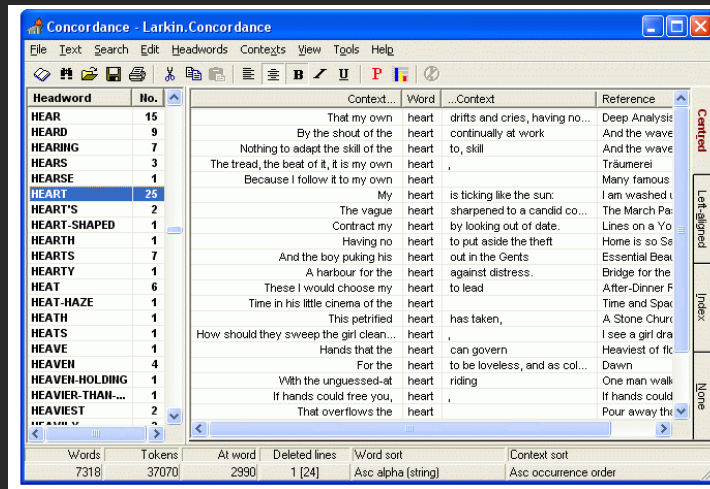
57



58

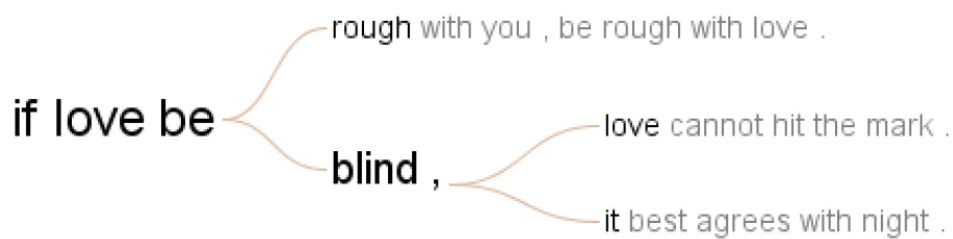
# Concordance

What is the common local context of a term?



61

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 .



63

# WordTree



64

# Filter infrequent runs



65



## 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>

68

## Phrase Nets [van Ham 2009]

---

**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**

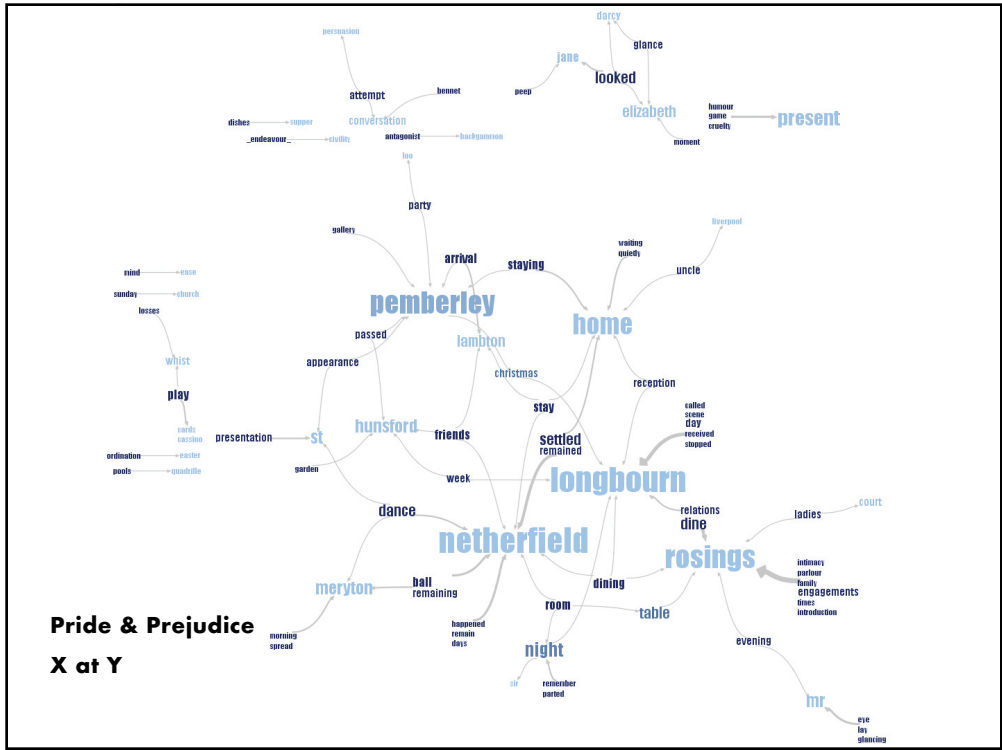
Occurrences → Node size

Pattern position → Edge direction

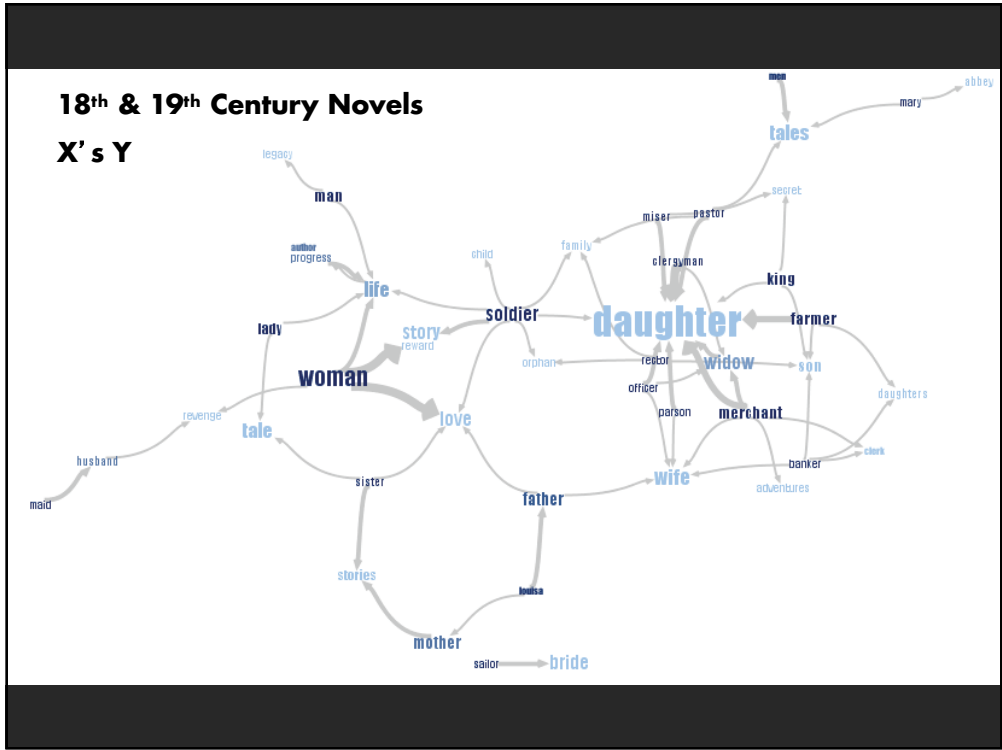
Darker color → higher ratio of out-edges to in-edges

69





73



76





# Visualizing Conversation

89

## Visualizing Conversation

---

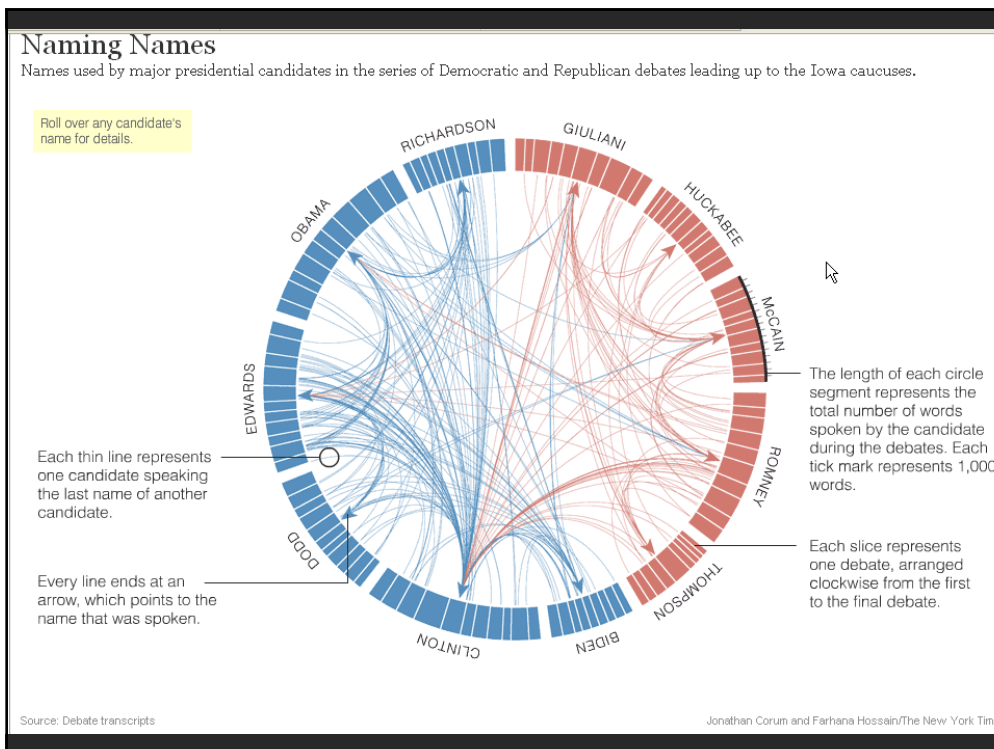
### Many dimensions to consider:

- Who (senders, receivers)
- What (the content of communication)
- When (temporal patterns)

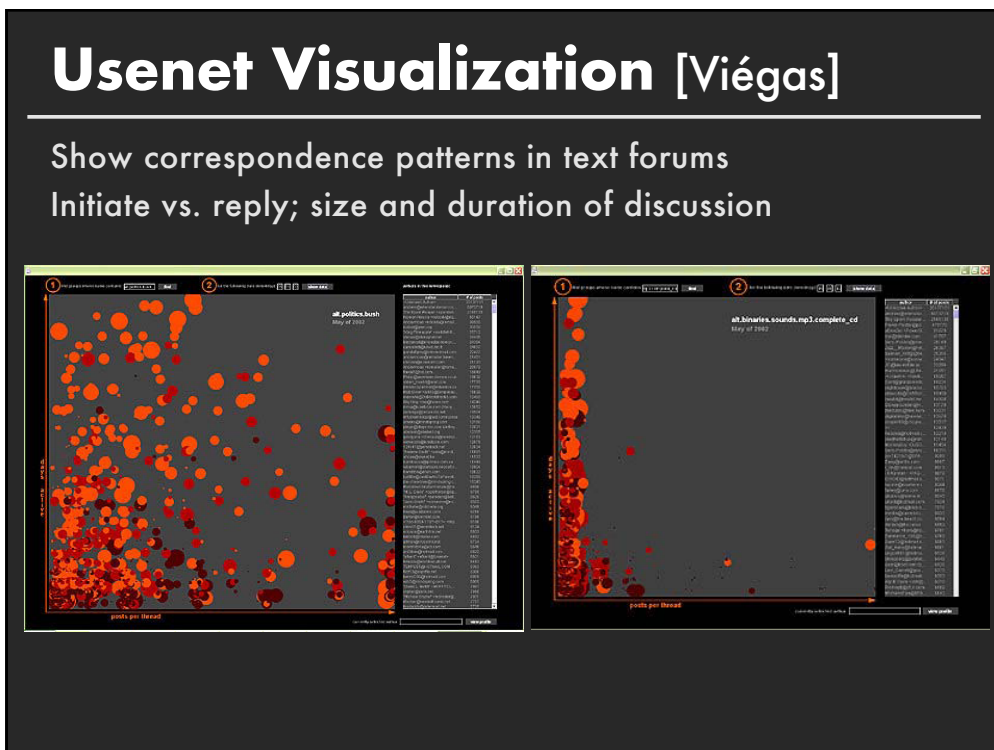
### Interesting cross-products:

- What x When → Topic “Zeitgeist”
- Who x Who → Social network
- Who x Who x What x When → Information flow

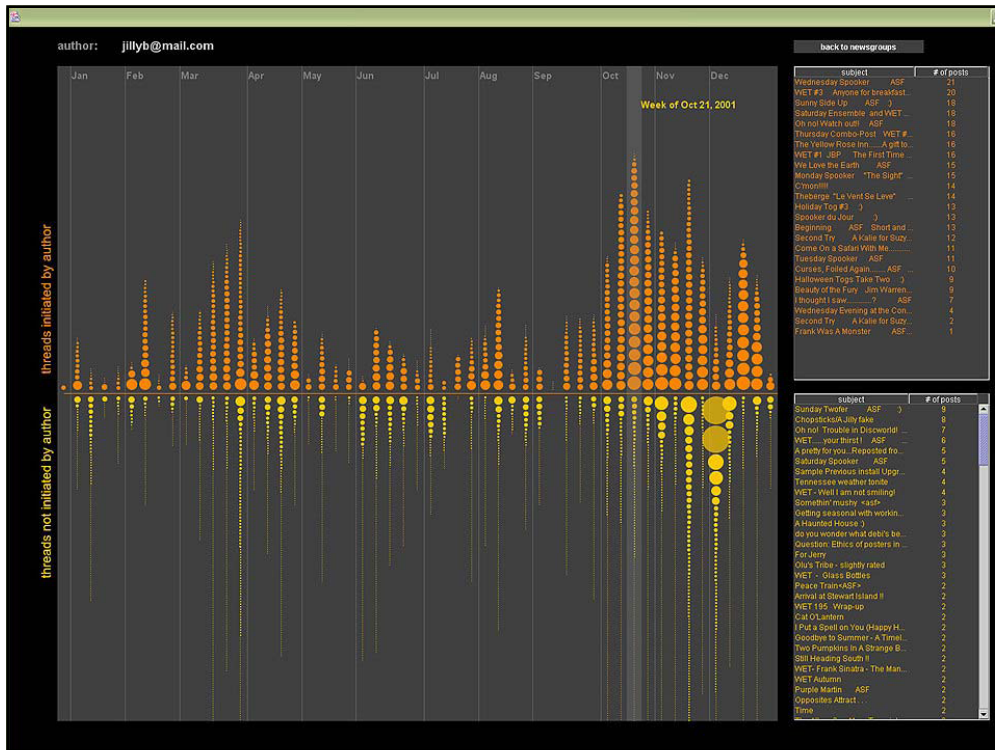
90



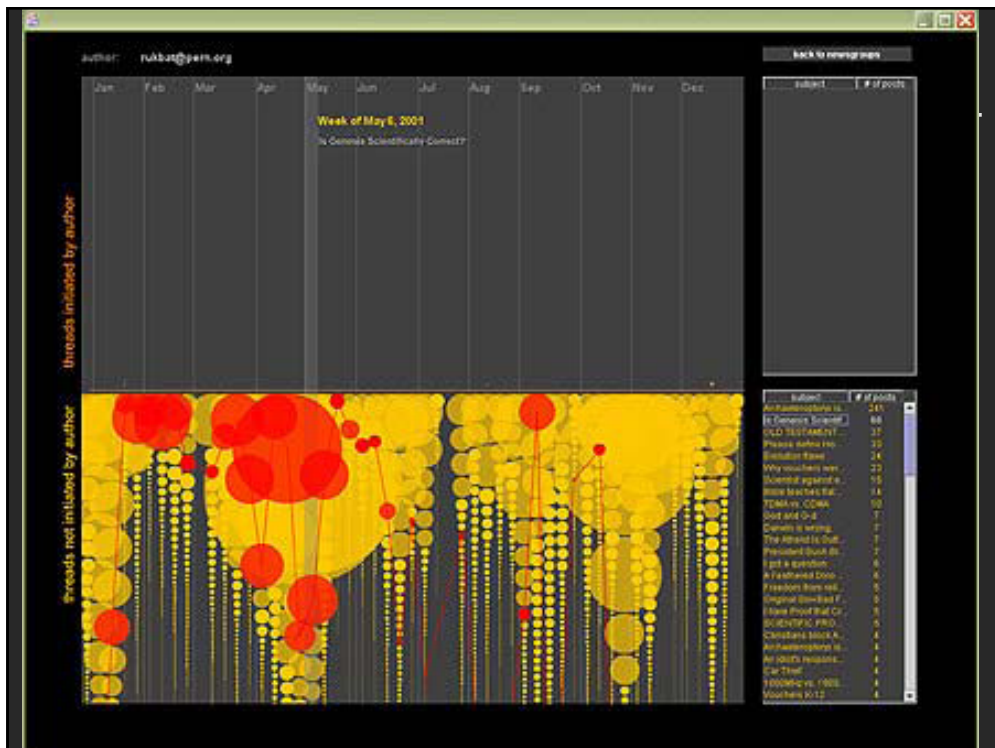
91



94

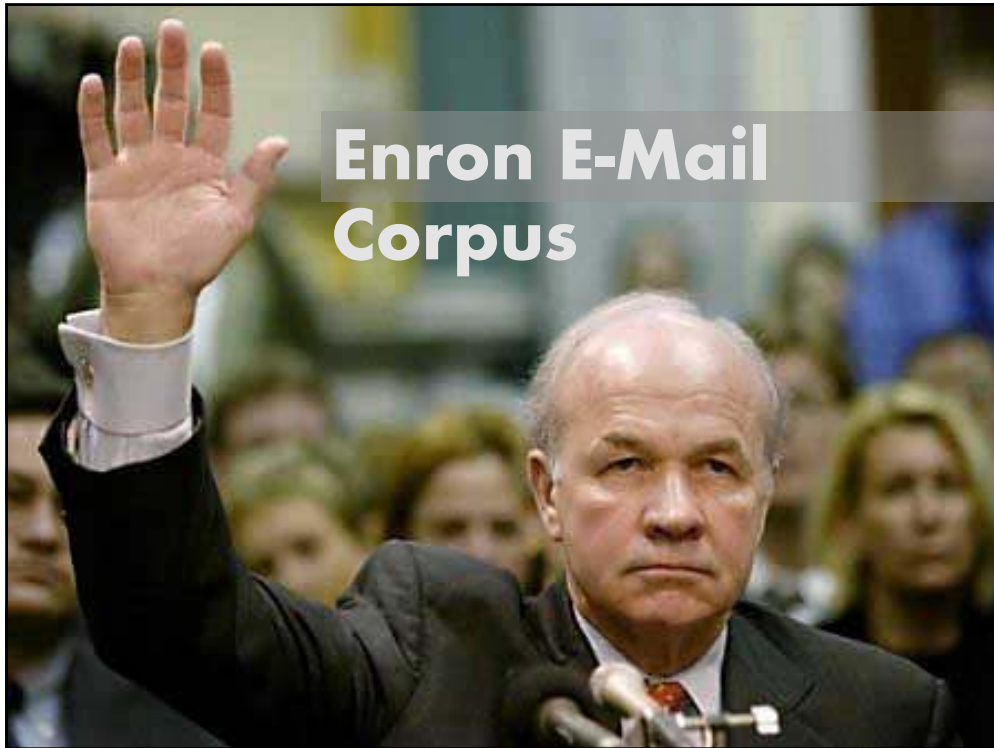


95

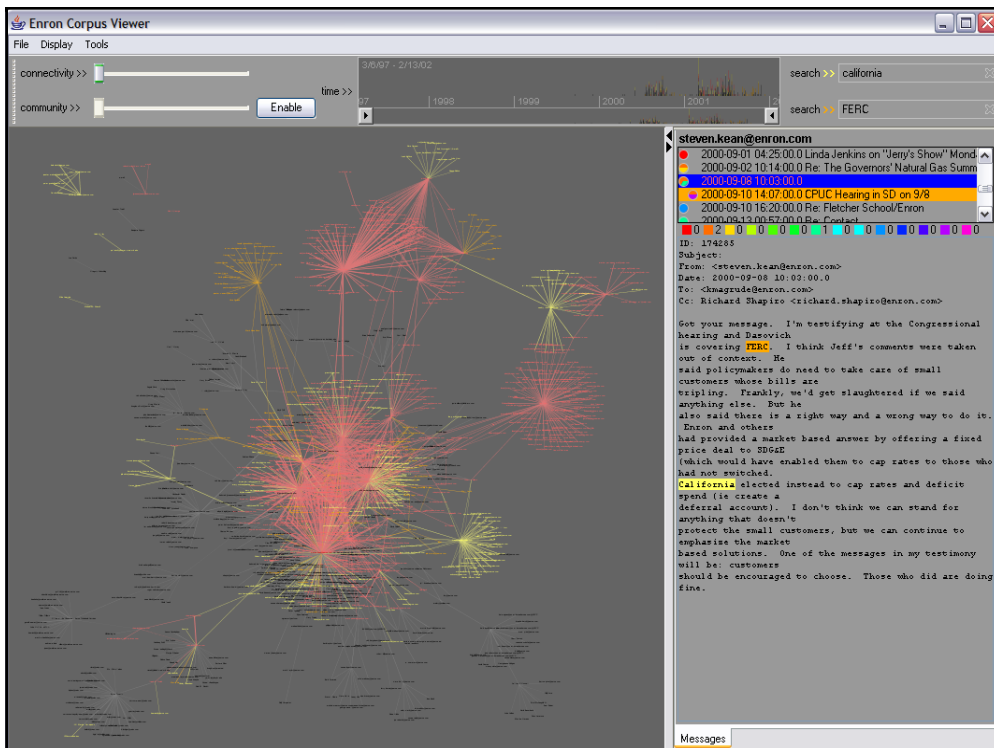


96





99



100

The screenshot displays the 'Enron Corpus Viewer' interface. At the top, there are controls for 'connectivity' and 'community', a date range of '1/20/01 - 6/27/01', and search fields for 'california' and 'ferc'. The main area features a network graph with nodes representing individuals and edges representing email links. A large text overlay reads 'Washington Lobby?' with a question mark. A news article snippet is visible, titled 'Enron 'Mastermind' Pleads Guilty', dated 'SAN FRANCISCO, Oct. 17, 2002'. The article mentions Timothy Belden, a former top energy trader, and Deputy Attorney General Larry Thompson. A quote from Belden is also present: 'I did it because I was trying to maximize profit for Enron.' The interface also shows a list of search results on the right and a 'Messages' section at the bottom.

101

A dark gray rectangular area with the text 'Visualizing Document Collections' centered in a large, white, sans-serif font.

102

# Named Entity Recognition

## Identify and classify named entities in text:

John Smith → PERSON

Soviet Union → COUNTRY

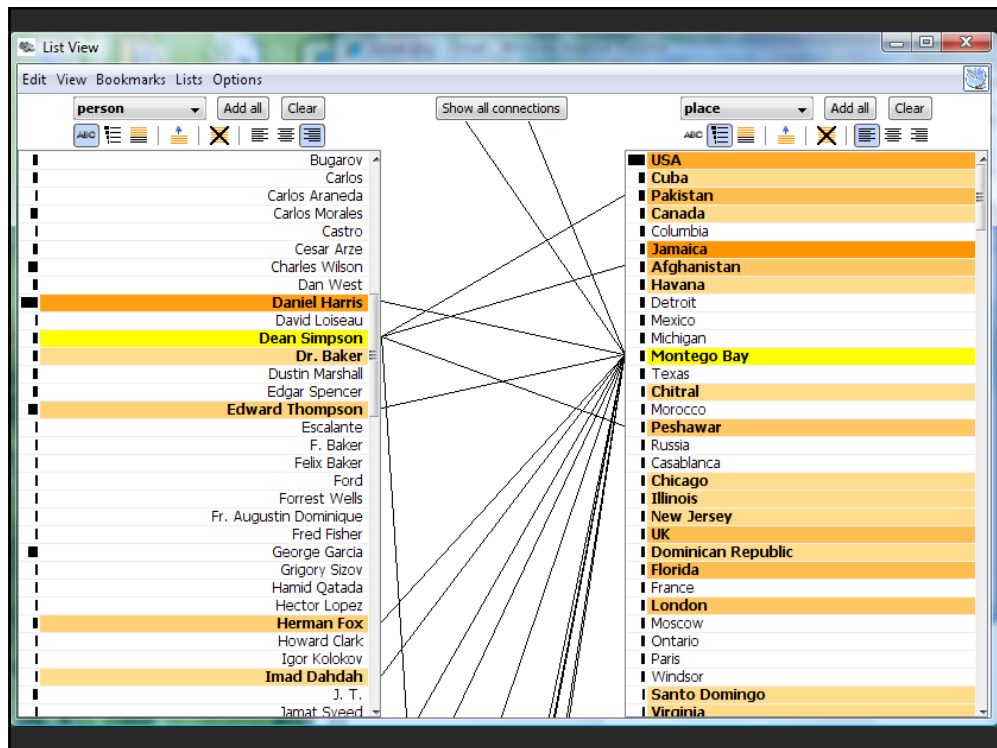
353 Serra St → ADDRESS

(555) 721-4312 → PHONE NUMBER

## Entity relations: how do the entities relate?

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

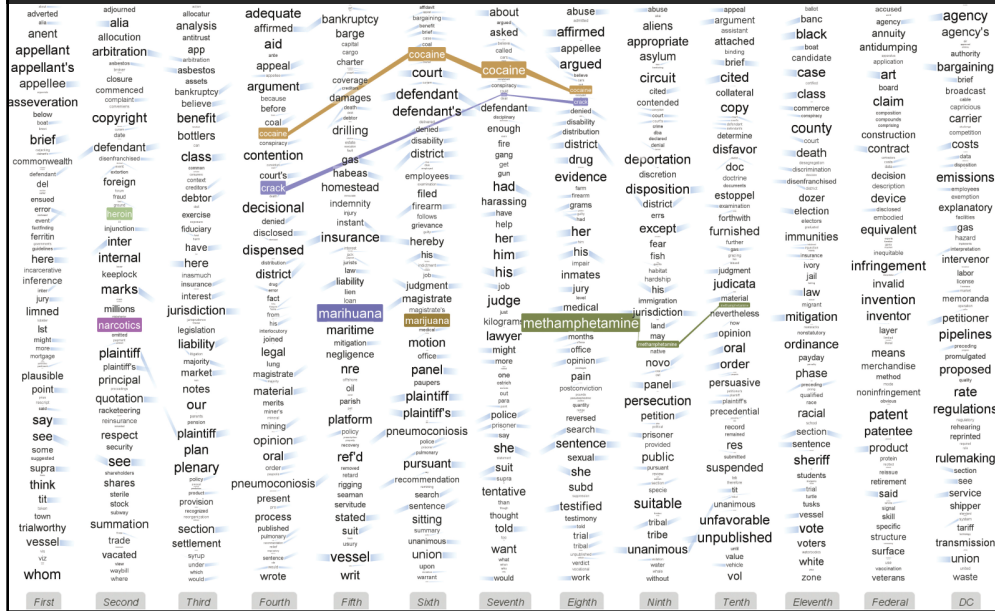
105



106



# Parallel Tag Clouds [Collins 09]



107

# Topic modeling

## Topic modeling approaches

Assume documents are a mixture of topics

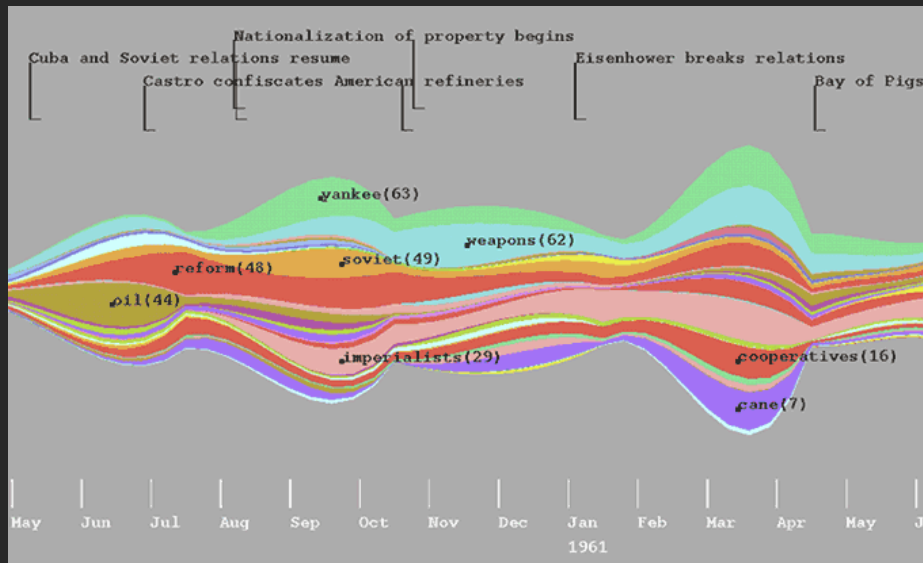
Topics are (roughly) a set of co-occurring terms

Latent Semantic Analysis (LSA): reduce term matrix

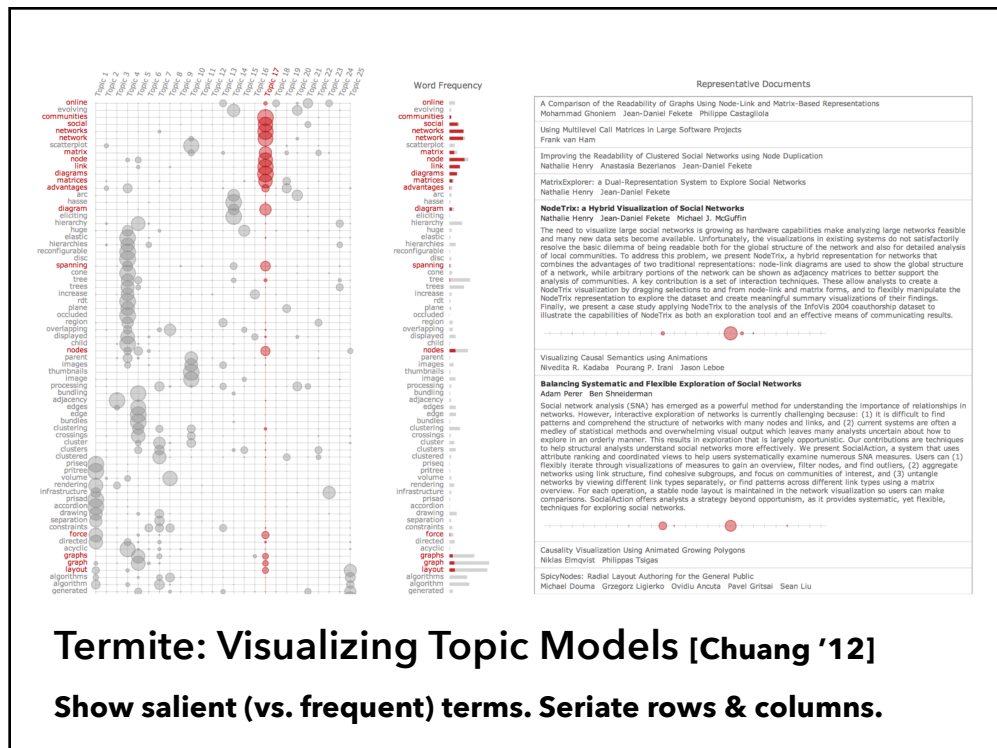
Latent Dirichlet Allocation (LDA): statistical model

111

# ThemeRiver (Havre et al 99)

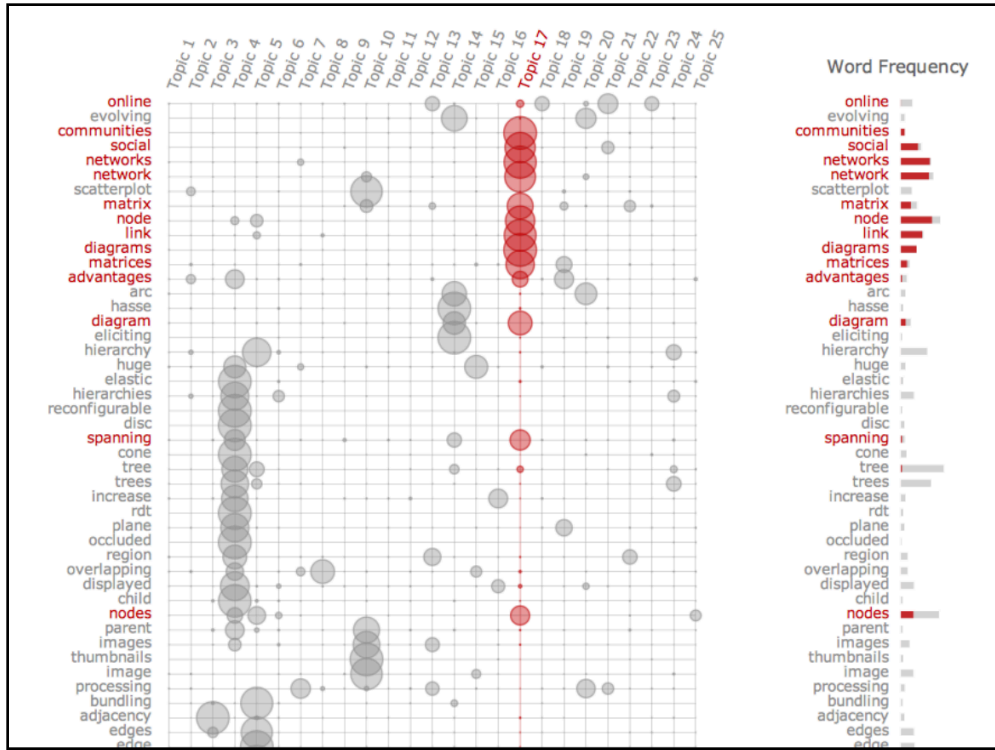


112

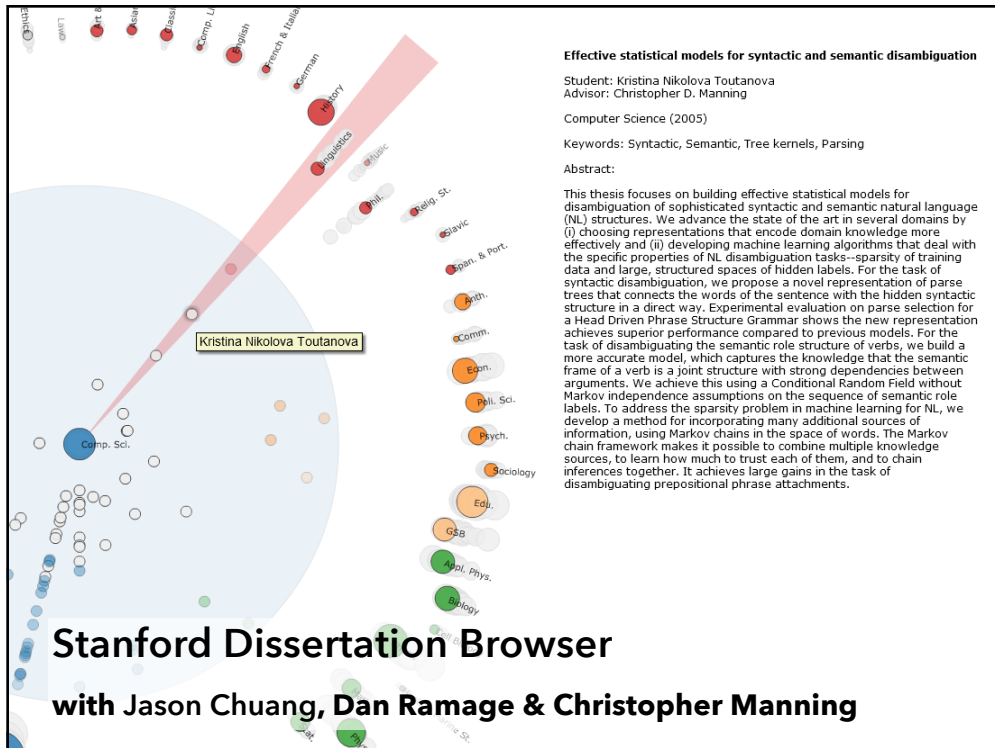


120

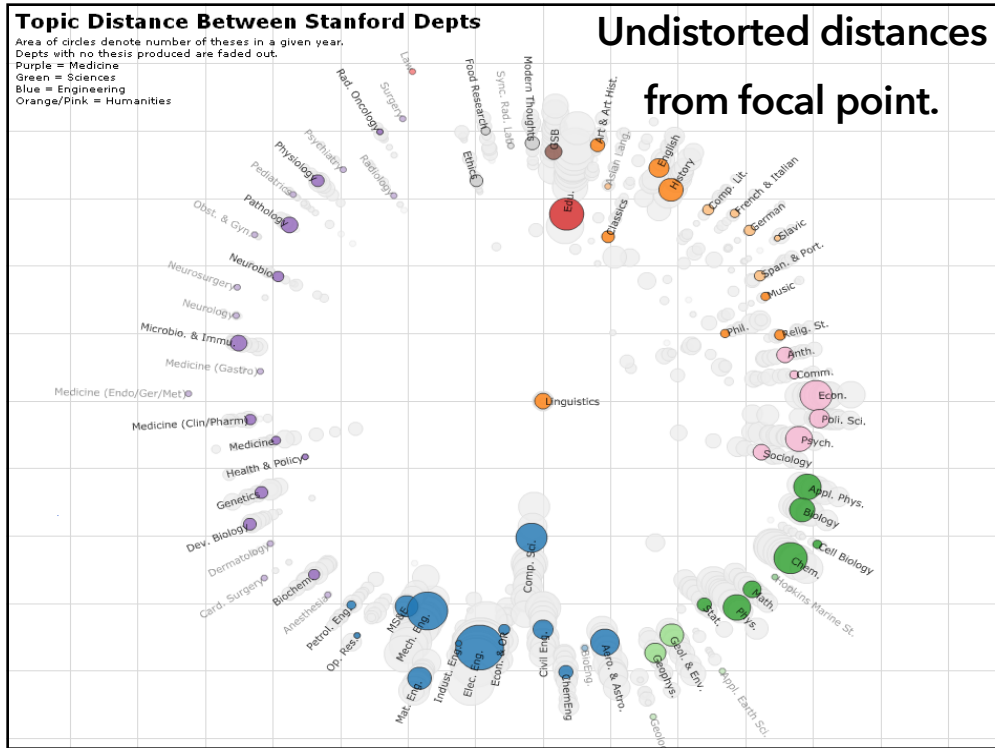
**Termite: Visualizing Topic Models [Chuang '12]**  
**Show salient (vs. frequent) terms. Seriate rows & columns.**



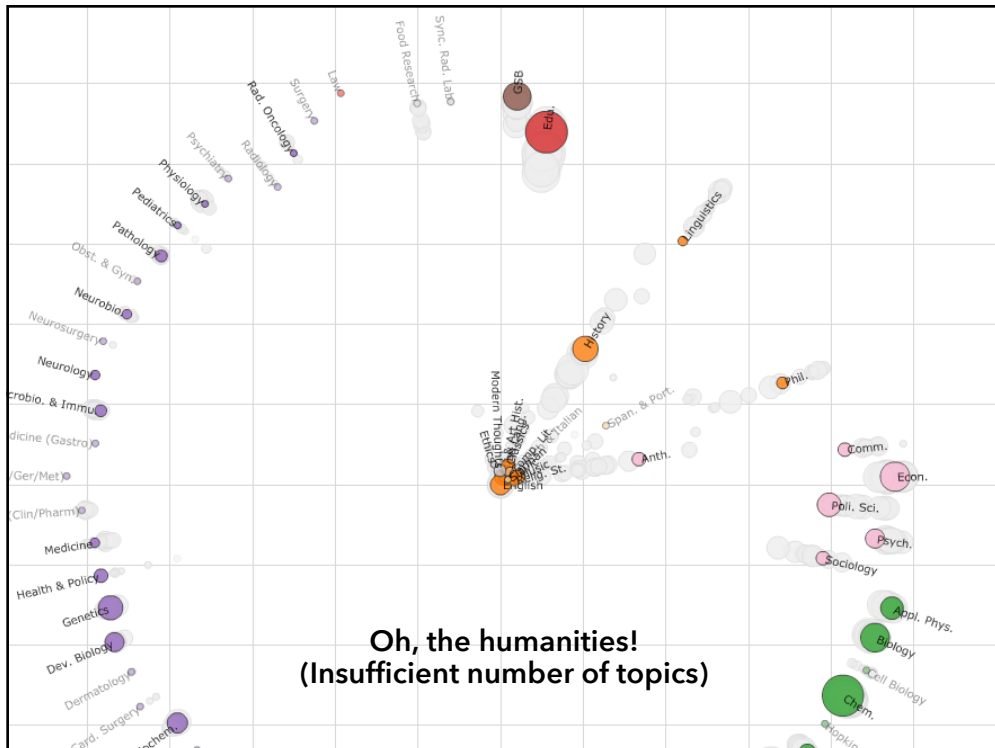
121



122



124



125

# 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 & models.

**Currently:** from bag-of-words to *vector space embeddings*