Communicating with Data – Text

Dr. Ab Mosca (they/them)

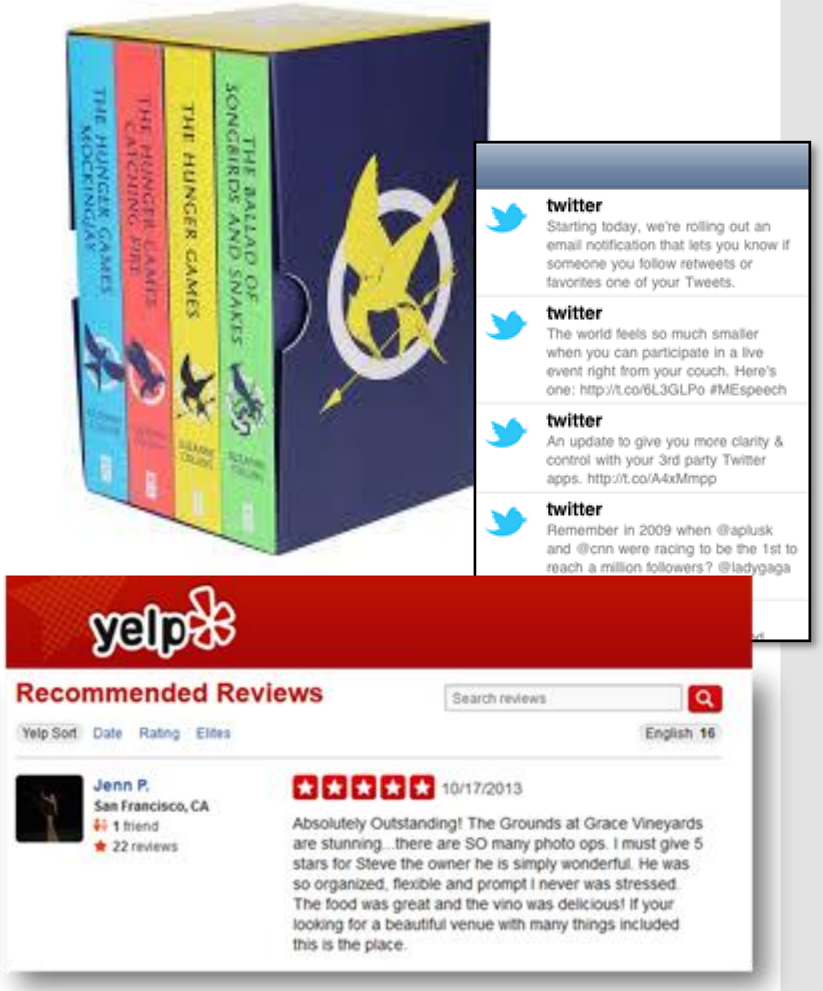
Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

- What is text data?
- Why visualize text?
- Techniques
- Lab

What is text data?

- Documents
  - Articles, books and novels
  - E-mails, web pages, blogs
- Text snippets
  - Tweets, SMS messages
  - Tags, comments, profiles
- And more…
  - Computer programs, logs
  - Collections of documents
  - This slide!


THE HUNGER GAMES MOCKINGJAY
THE HUNGER GAMES CATCHING FIRE
THE HUNGER GAMES
THE BALLAD OF SONGBIRDS AND SNAKES

twitter
Starting today, we're rolling out an email notification that lets you know if someone you follow retweets or favorites one of your Tweets.

twitter
The world feels so much smaller when you can participate in a live event right from your couch. Here's one: http://t.co/6L3GLPo #MEspeech

twitter
An update to give you more clarity & control with your 3rd party Twitter apps. http://t.co/A4xMmpp

twitter
Remember in 2009 when @aplusk and @cnn were racing to be the 1st to reach a million followers? @ladygaga

yelp

Recommended Reviews

Search reviews

Yelp Sort Date Rating Elites English 16

Jenn P.
San Francisco, CA
1 friend
22 reviews

★★★★★ 10/17/2013

Absolutely Outstanding! The Grounds at Grace Vineyards are stunning...there are SO many photo ops. I must give 5 stars for Steve the owner he is simply wonderful. He was so organized, flexible and prompt I never was stressed. The food was great and the vino was delicious! If your looking for a beautiful venue with many things included this is the place.

Overview

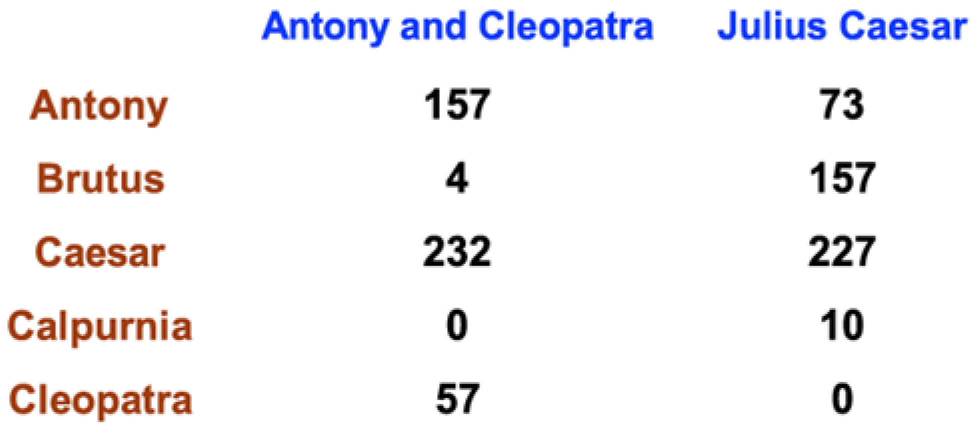
- **Question:** what are some **characteristics** of text data?
- **Answer:**
  - Often high dimensional (over **228,000** words in OED)
  - Packed with meaning and relationships:
    - **Correlations**: Hong Kong, San Francisco, Bay Area
    - **Order**: April, February, January, June, March, May
    - **Membership**: Tennis, Running, Swimming, Hiking, Piano
    - **Hierarchy**, antonyms & synonyms, entities, …

Why visualize text data?

- **Understand** – read a document
- **Summarize** – get the “gist” of a document
- **Cluster** – group together similar contents
- **Quantify** – convert to numerical measures
- **Correlate** – compare patterns in text to those in other data, e.g., test scores with conversations on social media

“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
- For example, simple term counts
- Aggregate into a document-term matrix



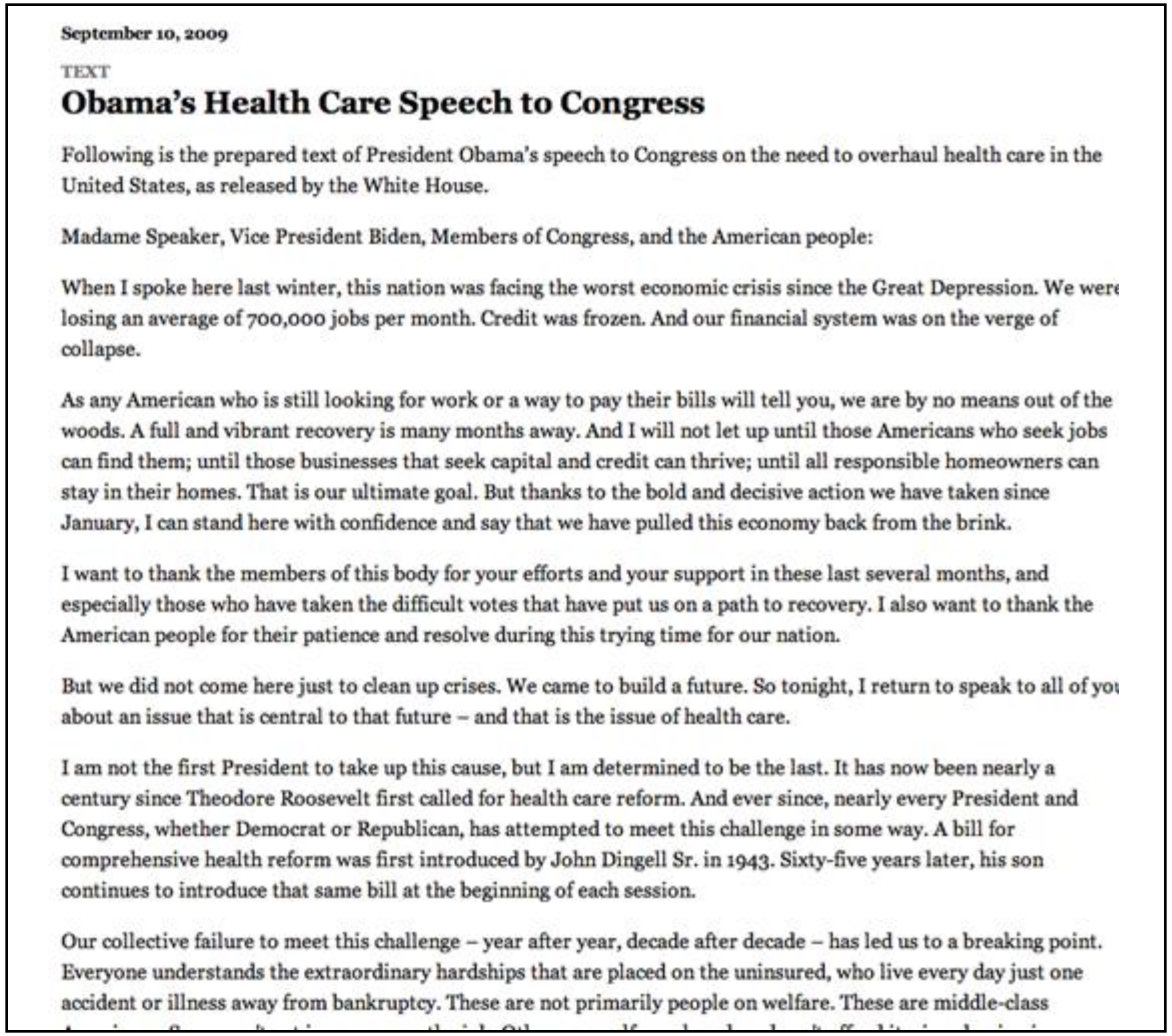
	Antony and Cleopatra	Julius Caesar
Antony	157	73
Brutus	4	157
Caesar	232	227
Calpurnia	0	10
Cleopatra	57	0

Example: health care reform

- Recent history
  - Initiatives by President Clinton
  - Overhaul by President Obama
- Text data
  - News articles
  - Speech transcriptions
  - Legal documents

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

A concrete example



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 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.

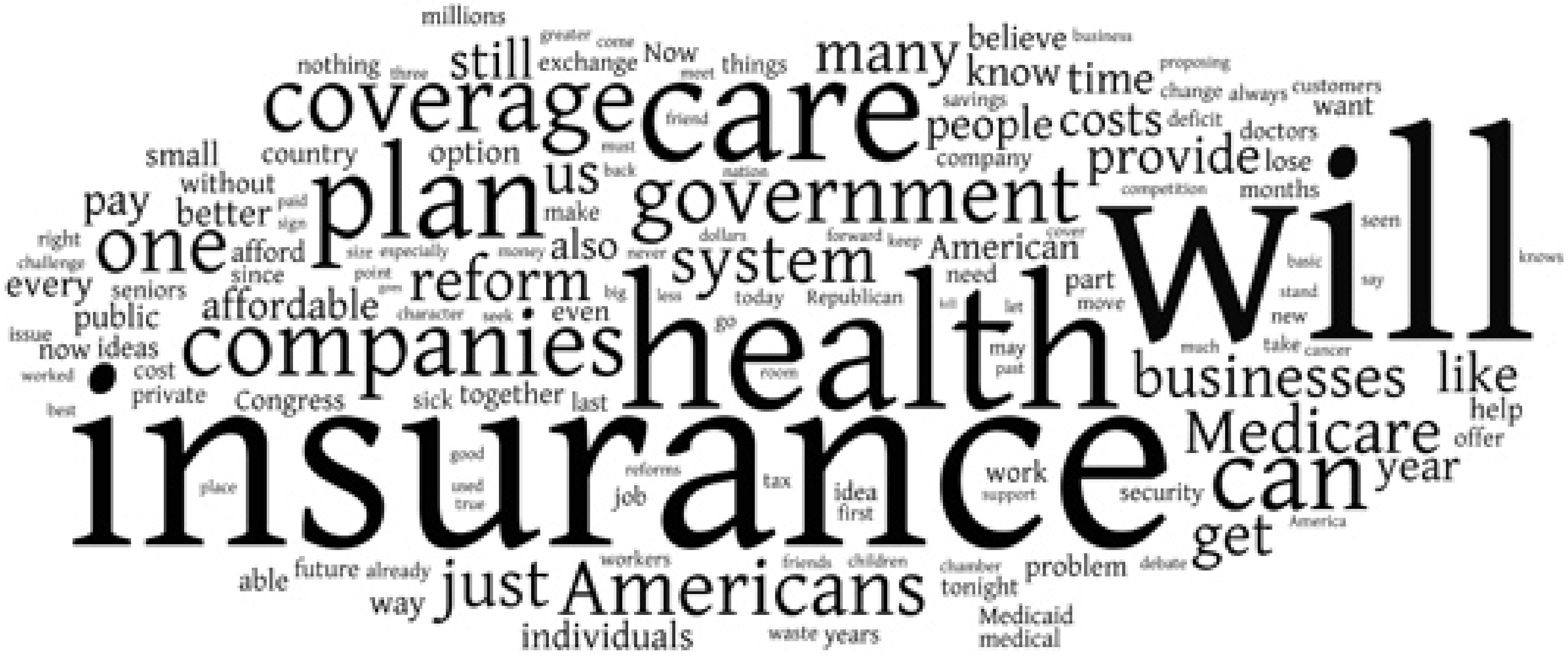
I want to thank the members of this body for your efforts and your support in these last several months, and especially those who have taken the difficult votes that have put us on a path to recovery. I also want to thank the American people for their patience and resolve during this trying time for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

I am not the first President to take up this cause, but I am determined to be the last. It has now been nearly a century since Theodore Roosevelt first called for health care reform. And ever since, nearly every President and Congress, whether Democrat or Republican, has attempted to meet this challenge in some way. A bill for comprehensive health reform was first introduced by John Dingell Sr. in 1943. Sixty-five years later, his son continues to introduce that same bill at the beginning of each session.

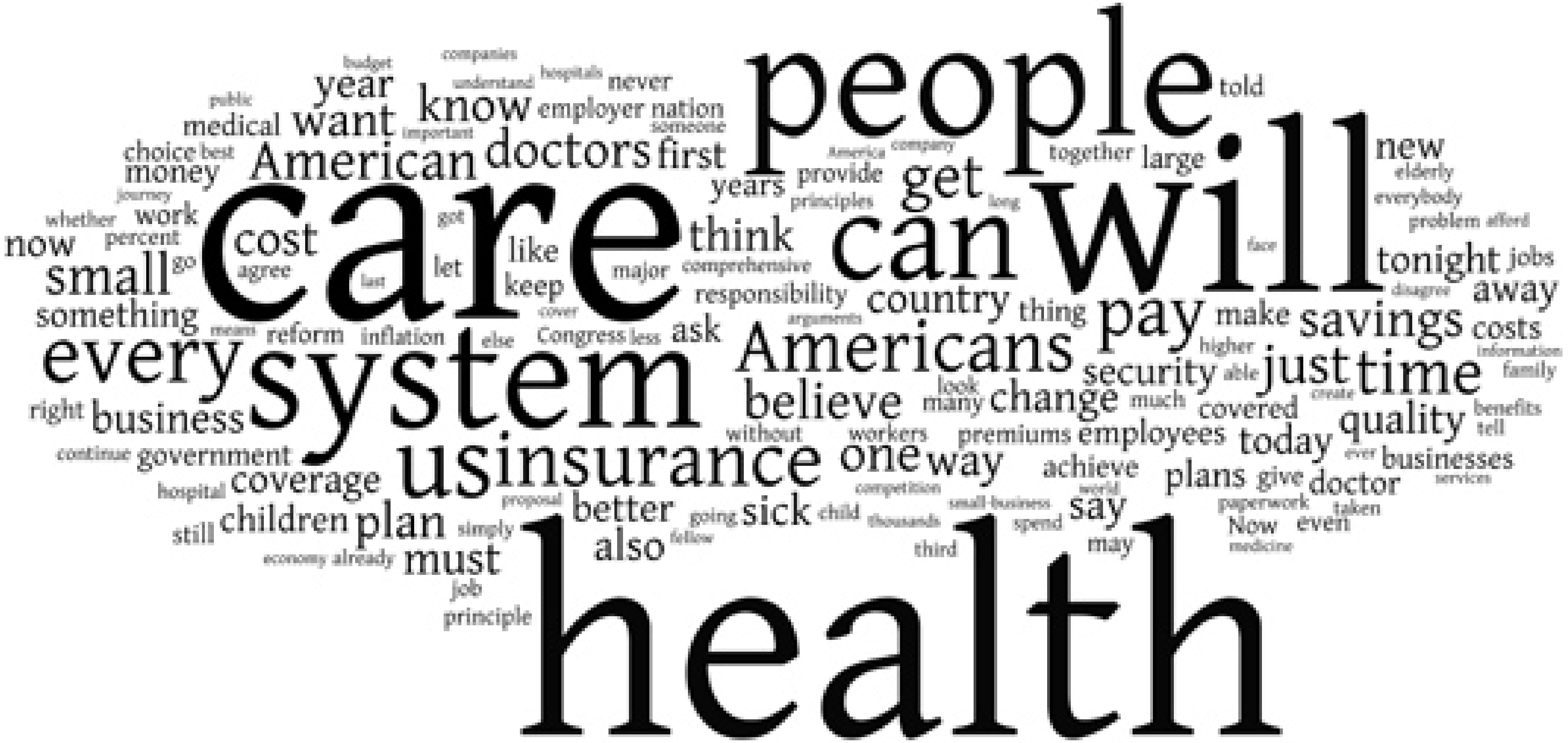
Our collective failure to meet this challenge – year after year, decade after decade – has led us to a breaking point. Everyone understands the extraordinary hardships that are placed on the uninsured, who live every day just one accident or illness away from bankruptcy. These are not primarily people on welfare. These are middle-class

New York Times: Obama 2009



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

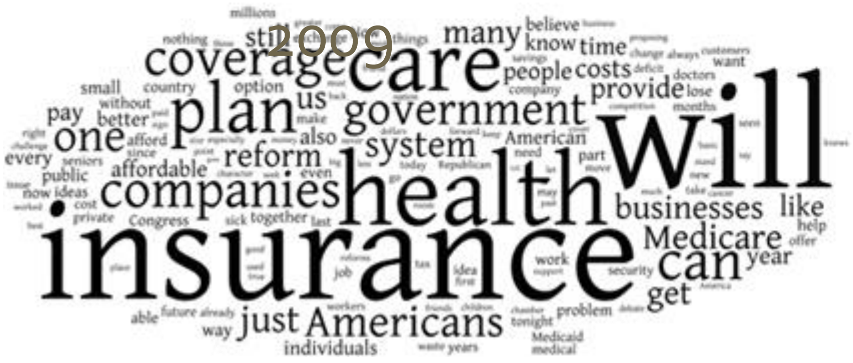
New York Times: Clinton 1993



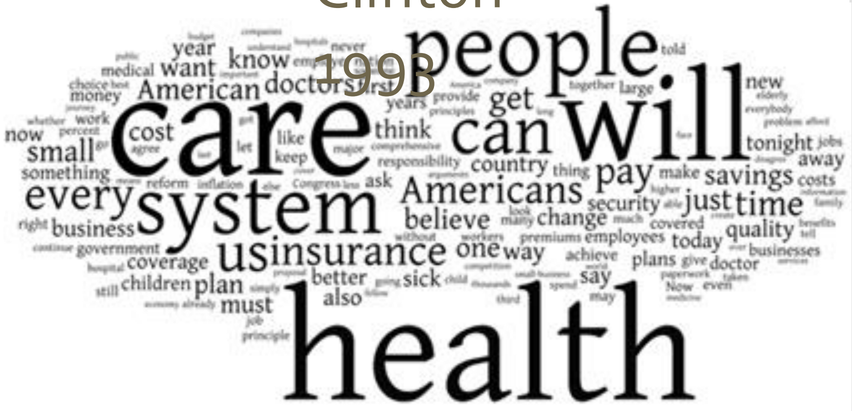
economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

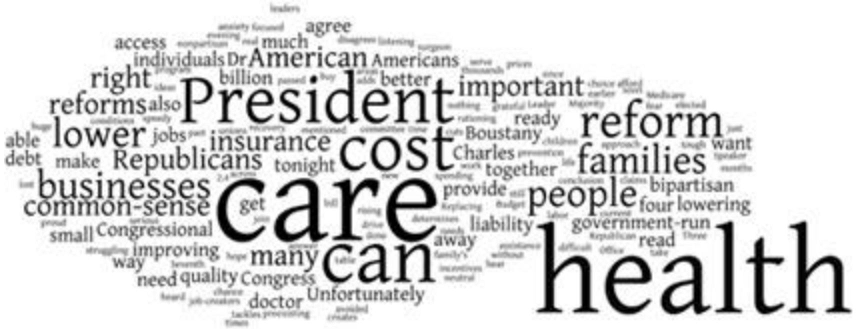
Comparison

Obama 2009



Clinton 1993





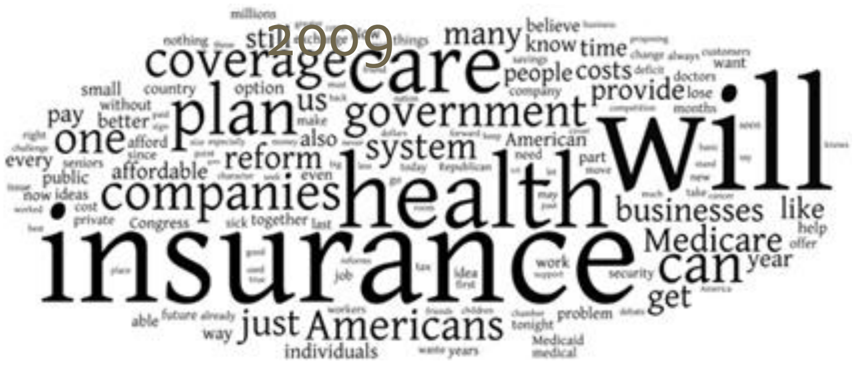
Rep. Charles Boustany of Louisiana 2009

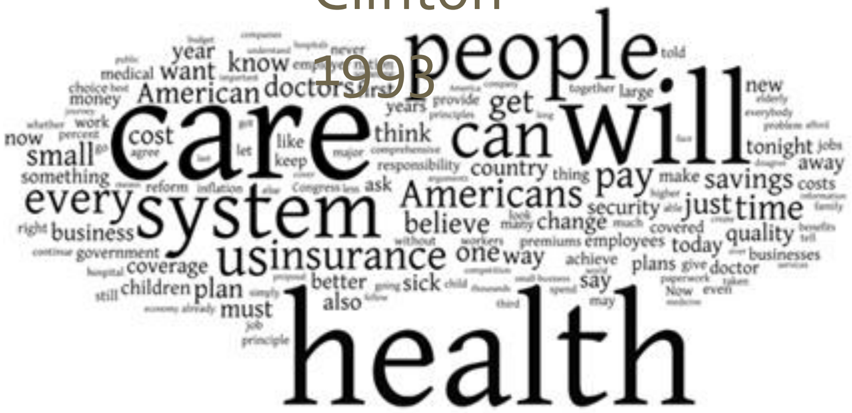
economix.blogs.nytimes.com/2009/09/09/obama-in-09-vs-clinton-in-93

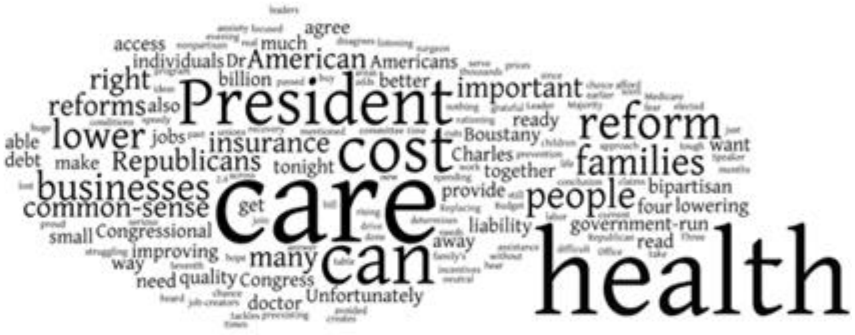
Word clouds

- Strengths
  - Familiar to many people
  - Can help with “gisting” and initial query formation
- Weaknesses
  - Does not show the structure of the text
  - Sub-optimal visual encoding (position is not meaningful)
  - Inaccurate size encoding (long words are bigger)
  - May not facilitate comparison (unstable layout)
  - Term frequency **may not be meaningful**

Flashback


Obama 2009


Clinton 1993



Rep. Charles Boustany of Louisiana 2009

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

Weighting words

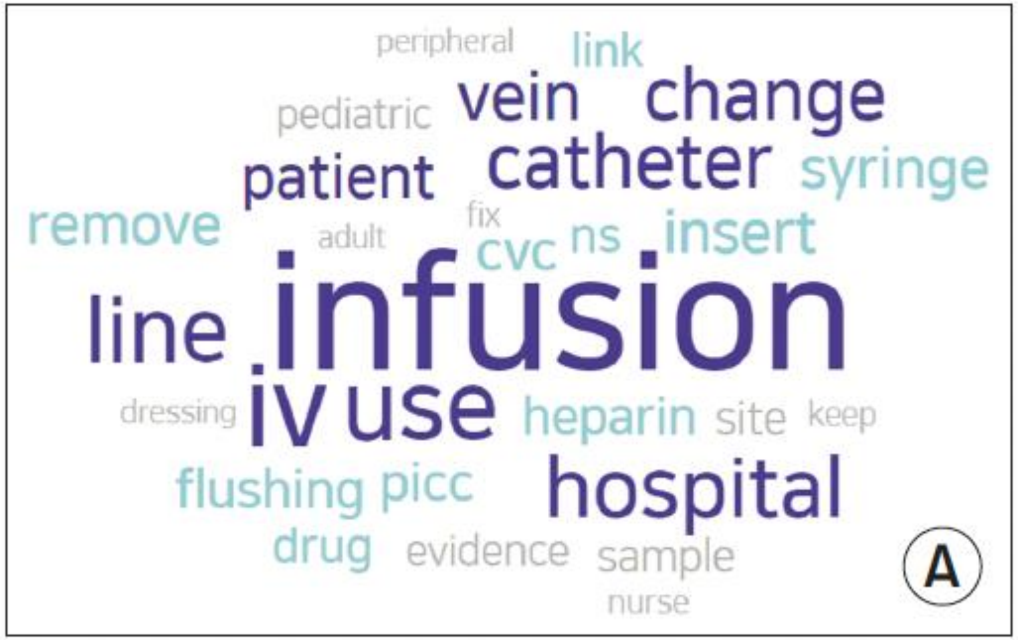
- Term Frequency

$$tf_{td} = \text{\# of times term } t \text{ appears in document } d$$

- TF-IDF: Term Frequency by Inverse Document Frequency

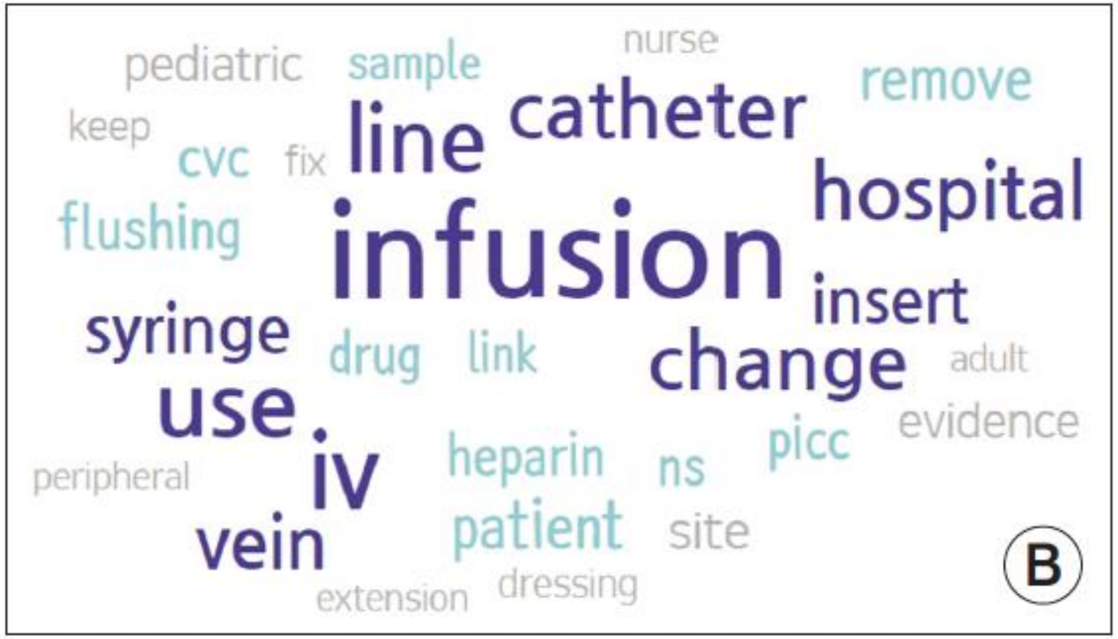
$$tf\text{-}idf_{td} = \frac{\text{\# of times term } t \text{ appears in document } d}{\text{\# of times term } t \text{ appears in all documents}}$$

Frequency example



Keyword Network Analysis of Infusion Nursing from Posts on the Q&A Board in the Intravenous Nurses Caf.

TF-IDF example

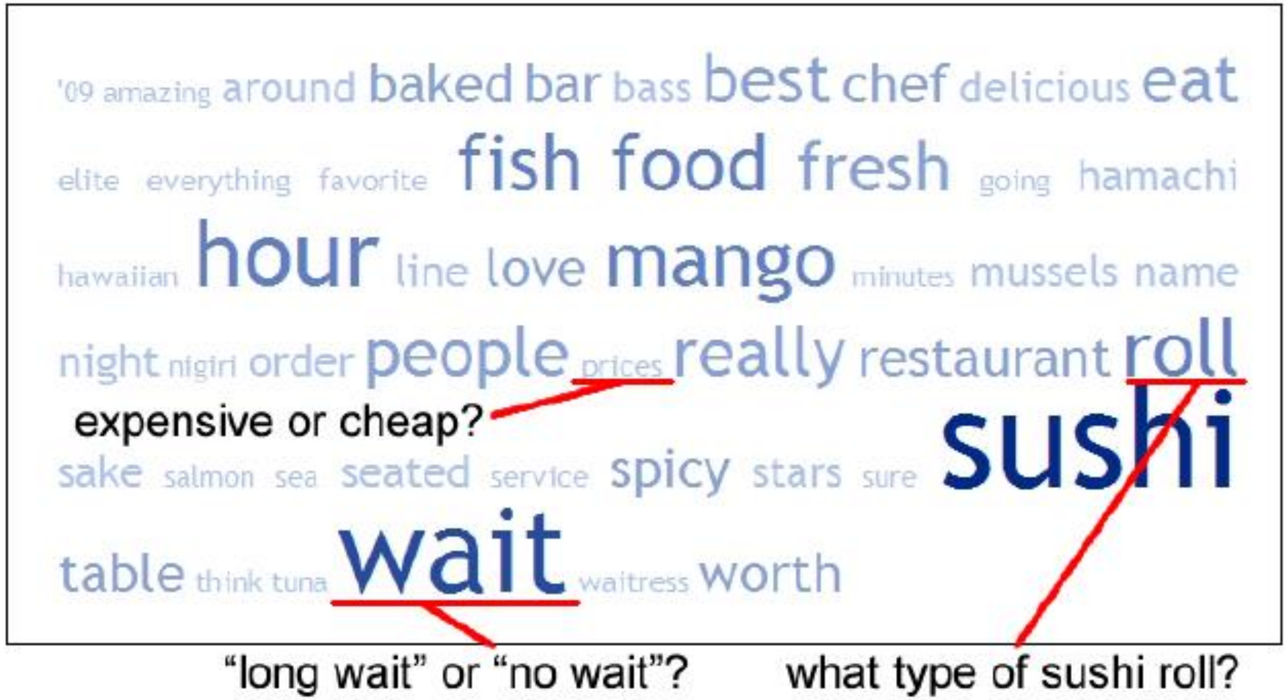


Keyword Network Analysis of Infusion Nursing from Posts on the Q&A Board in the Intravenous Nurses Caf.

Limitations of frequency statistics

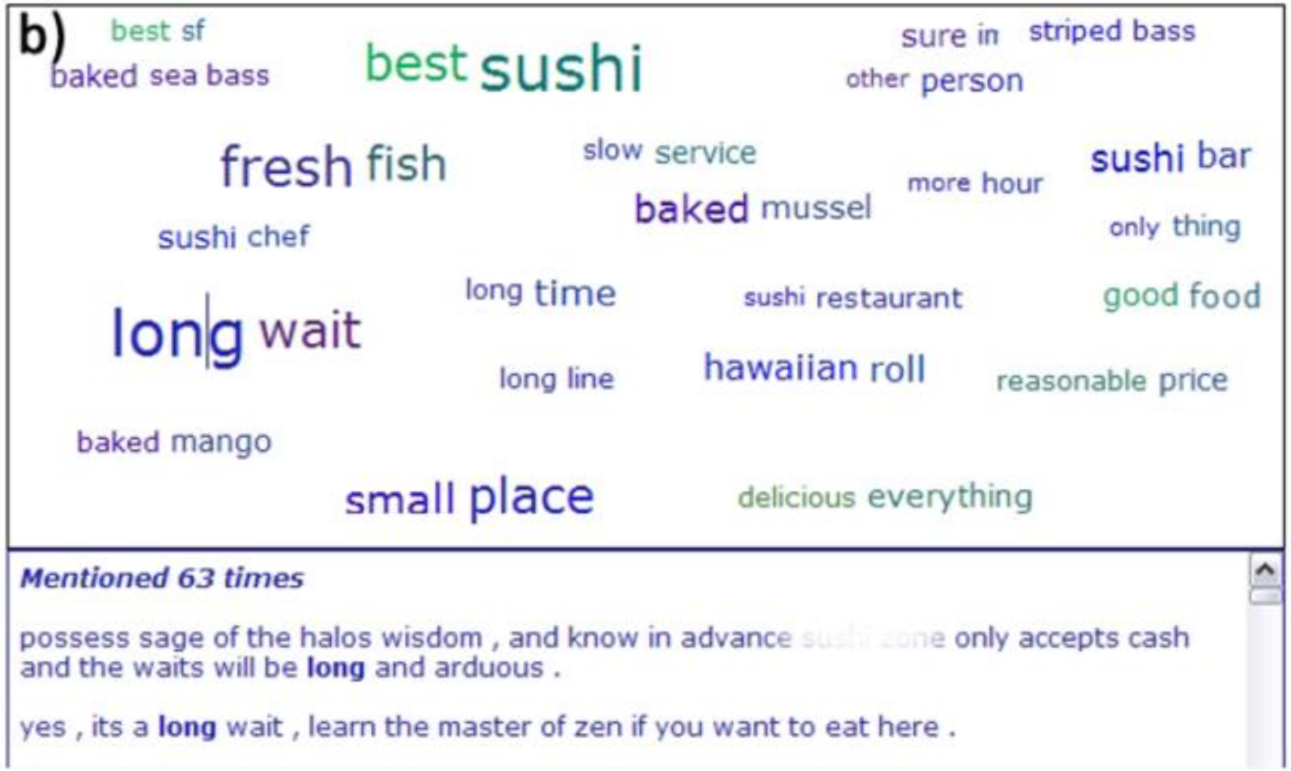
- Often favors frequent (TF) or rare (IDF) terms
  - Still not clear that these provide best description
- A “bag of words” ignores additional information
  - Grammar / part-of-speech
  - Position within document
  - Recognizable entities
- Typically focus on unigrams (single terms)

Example: Yelp reviews


'09 amazing around baked bar bass best chef delicious eat
elite everything favorite fish food fresh going hamachi
hawaiian hour line love mango minutes mussels name
night nigiri order people prices really restaurant roll
expensive or cheap?
sake salmon sea seated service spicy stars sure sushi
table think tuna wait waitress worth
“long wait” or “no wait”?
what type of sushi roll?

Yatani 2011

Example: Yelp reviews



Mentioned 63 times

possess sage of the halos wisdom , and know in advance sushi zone only accepts cash and the waits will be **long** and arduous .

yes , its a **long** wait , learn the master of zen if you want to eat here .

Yatani 2011

Tips: descriptive key-phrases

- 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 key-phrases in context

Discussion

What are some other ways we might **visualize text data?**

Text lab

- Find instructions for today’s lab on the course website